

Analysis of IT Jobs in Pakistan using Data Science



Session (2020-2024)

Program

Bachelor of Computer Science

Submitted By:

Mian Hammad Mustafa COMSC-F20-023

Supervised By:

Mr. Waqas Ahmed

Mr. Muhammad Abid

Assistant Professor of Computer Science

DEPARTMENT OF COMPUTER SCIENCE

GOVERNMENT POSTGRADUATE COLLEGE MANSEHR

Analysis of IT Jobs in Pakistan using Data Science



A report submitted to
Department of Computer Science
Government Postgraduate College Mansehra, HED, KP
As a partial fulfillment of requirements
For the award of the degree of
Bachelors of Computer Sciences.

DEPARTMENT OF COMPUTER SCIENCE
GOVERNMENT POSTGRADUATE COLLEGE MANSEHRA, HED, KP

FINAL APPROVAL

This is to certify that we have read the thesis titled " **Analysis of IT Jobs in Pakistan using Data Science**" submitted by the following students of BSCS of Government Postgraduate College, Mansehra.

| S.No | Name | Roll No |
|------|---------------------|---------------|
| 1. | Mian Hammad Mustafa | COMSC-F20-023 |

We judge that this thesis is of sufficient standard to warrant its acceptance by the Department of Computer Science Government Postgraduate College, Mansehra for the award of BS degree in computer science from Hazara university, Mansehra.

External Examiner

Dr. Zafar Mehmood

Professor & HOD of IT Department

Hazara University, Battagram Campus

Internal Examiner

Mr. Muhammad Abid

Assistant professor of computer science,

Government Postgraduate College, Mansehra.

Supervisor

Mr. Muhammad Abid

Assistant professor of computer science

Government Postgraduate College , Mansehra.

Head of Department

Mr. Muhammad Abid

Assistant professor of computer science

Government Postgraduate College Mansehra.

DEDICATION

We dedicate this thesis and degree to our parents, who have always been there for us, providing unwavering financial and moral support, and fulfilling our every need. We extend our dedication to our respected teachers, who have supported us throughout our academic journey. Additionally, we dedicate this achievement to our beloved family and friends, as well as those teachers who have consistently motivated, supported, and encouraged us in every aspect of our lives.

DECLARATION

We solemnly declare that this project, in its entirety and its individual components, has not been plagiarized from any source. We affirm that the software and accompanying report were developed solely through our personal efforts, with sincere guidance from our supervisor and teachers. We take full responsibility for any consequences that may arise if any part of this system is proven to be copied or reproduced from another source.

Name: Mian Hammad Mustafa

Signature: _____

ACKNOWLEDGEMENT

We express our heartfelt gratitude and joyous appreciation to the Almighty Allah for granting us the opportunity to benefit from His abundant blessings. Our deepest thanks go to our dear parents, whose unwavering prayers provided the foundation for the successful completion of our project. We are immensely grateful to our esteemed and kind-hearted supervisor, Mr. Waqas Ahmed, whose brilliant idea, dedicated supervision, and unwavering support contributed significantly to our project's success. We extend our sincere thanks to all teachers of BS Computer Science for their invaluable support. Lastly, we wish to express our utmost respect and gratitude to Mr. Waqas Ahmed, Mr. Muhammad Abid and all other concerned teachers for their moral support and genuine well-wishes throughout the entire process of our work.

PROJECT IN BRIEF

| | |
|-------------------------|---|
| Project Title | Analysis of IT Jobs in Pakistan using Data Science |
| Organization | Government Postgraduates College Manshra |
| Undertaken By | Mian Hammad Mustafa |
| Supervised By | Mr. Waqas Ahmed (7 th semester) and Mr. Muhammad Abid (8 th semester) |
| Starting Month | September 2023 |
| Ending Month | June 2024 |
| Software Used | Python and Power BI |
| Environment Used | Jupiter Lab |
| System Used | Intel(R) Core(TM) i3-3110M CPU @ 2.40GHz 2.40 GHz |

PREFACE

This report will present end to end lifecycle of data Science project “*Analysis of IT Jobs in Pakistan using Data Science*”. To easily understand the project, I have divided it into different chapters.

- Chapter One** **Introduction of Project:** Provides an introduction to the project.
- Chapter Two** **Literature review:** Presents an overview of Research studies conducted on this topic in the past.
- Chapter Three** **Data Collection:** All about different ways and Sources to get Job data along with the organization of data Files into folders and combining data into one Excel File.
- Chapter Four** **Data Preparation 1:** This chapter is about data cleaning, exploring data (EDA), and phase 1 of data preprocessing (Feature Engineering) which involves making new features this is the data preparation stage before Filtering IT jobs (Separating IT jobs from Non-IT jobs).
- Chapter Five** **Filtration of IT jobs from Non-IT jobs:** This chapter is about criteria to define IT Jobs, Separating IT jobs from Non-IT jobs (Segmentation of IT jobs), and all ways to do keyword research and text analysis to correctly filter IT jobs.
- Chapter Six** **Data Preparation 2:** This chapter is about data preparation phase 2 in which some new features are created and finally prepared data for analysis and reporting.

- Chapter Seven** **Data Visualization:** This chapter is about Effective data visualization in Python and making a dynamic Power BI dashboard.
- Chapter Eight** **Results and Discussion:** This chapter is about the Results and Findings gained by my Research study, and discussion on different aspects of Pakistan's IT Jobs market also highlighted what challenges (Weak Areas) we face in the IT sector.
- References** Included the list of books and manuals and other material referenced during the project.

ABSTRACT

In today's digital era, the Information Technology industry is crucial in driving economic growth and job creation. This research study aims to provide a comprehensive and in-depth analysis of IT jobs in Pakistan. This research was conducted as part of my final year project in data science, particularly focusing on data analytics and business intelligence utilizing LinkedIn Jobs data scraped from September 26, 2023, to November 26, 2023, involving the analysis of 3734 IT jobs data, using python programming and JupyterLab environment along with Microsoft Power BI as Business Intelligence tool. The analysis focuses on various key features, including job trends, skills demand, geographical distributions, company-specific patterns, and industry-based distributions. Additionally, competition analysis is conducted to understand the intensity of competition in the job market, contributing to a comprehensive understanding of the landscape. The results and findings are presented through an interactive and dynamic online dashboard. This research is valuable for job seekers, employers, policymakers, and academia, aiding in informed decision-making regarding recruitment strategies, skill development initiatives, and workforce planning in the Pakistani IT sector. The findings not only highlight key trends but also identify challenges that, with the implementation of sound policies, can be transformed into opportunities for further growth and development of the IT industry in Pakistan.

Contents

| | |
|--|----|
| CHAPTER NO 1 | 1 |
| Introduction of Project | 1 |
| 1.1 Project Objectives: | 2 |
| 1.2 Background: | 3 |
| 1.3 Problem Statement: | 3 |
| 1.4 Research Questions: | 4 |
| 1.5 Key Features of Project: | 5 |
| 1.6 Scope of Project: | 6 |
| 1.7 Feasibility Study:..... | 7 |
| 1.8 Methodology: | 8 |
| 1.9 Tools and Technologies: | 9 |
| CHAPTER NO 2 | 10 |
| Literature Review | 10 |
| 2.1 Introduction of chapter:..... | 11 |
| 2.2 Exploring Industrial Demand Trends in Pakistan Software Industry Using Online Job Portal Data [1]: | 11 |
| 2.2.1 Objective:..... | 11 |
| 2.2.2 Methodology:..... | 11 |
| 2.2.3 Findings of Research Paper: | 12 |
| 2.2.4 Relevance of the Research:..... | 13 |
| 2.3 Labor Market Analysis Using Big Data the Case of a Pakistani Online Job Portal [2] | 14 |
| 2.3.1 Objective:..... | 14 |
| 2.3.2 Methodology:..... | 15 |
| 2.3.3 Findings of Research | 15 |
| 2.3.4 Relevance: | 16 |
| 2.4 Skills Set Required for Web Developers in Pakistan [3] | 16 |
| 2.4.1 Objective:..... | 16 |
| 2.4.2 Methodology:..... | 17 |
| 2.4.3 Findings of Research Study:..... | 17 |

| | |
|--|----|
| 2.4.4 Relevance to My Research | 18 |
| 2.5 DIGITAL PAKISTAN: OPPORTUNITIES & CHALLENGES [4]:..... | 18 |
| 2.5.1 The Objective: | 18 |
| 2.5.2 Methodology..... | 19 |
| 2.5.3 Key Findings: | 19 |
| 2.5.4 Relevance: | 19 |
| 2.6 Summary and Identified Gaps in Previous Research Studies | 20 |
| CHAPTER NO 3 | 21 |
| Data Collection and Organization | 21 |
| 3.1 Data Collection..... | 22 |
| 3.2 Concepts Related to Data Collection | 22 |
| 3.2.1 What is Data Collection? | 22 |
| 3.2.2 Role of data strategy in effective data collection | 22 |
| 3.2.3 Data Research | 23 |
| 3.2.4 Questions Arising in Data Research..... | 23 |
| 3.3 Data Collection Methods..... | 23 |
| 3.3.1 Three Ways to Collect Data: | 23 |
| 3.3.2 Data Scraping Without APIs | 24 |
| 3.3.3 Python for Scraping: Python Frameworks..... | 24 |
| 3.3.4 Limitations of Software Tools Used for Data Scraping | 24 |
| 3.3.5 Why I Chose LinkedIn Jobs for Analysis..... | 25 |
| 3.4 Problems Faced in Getting Jobs Data Using Data Partnership and LinkedIn APIs | 25 |
| 3.4.1 LinkedIn Economic Graph | 25 |
| 3.4.2 Development Data Partnership..... | 26 |
| 3.4.3 Challenges Encountered: | 27 |
| 3.4.4 LinkedIn APIs..... | 28 |
| 3.4.5 Third-Party Platforms | 29 |
| 3.4.6 General Issues with Data Partnerships and APIs: | 29 |
| 3.4.7 Reasons for Choosing Python for LinkedIn Job Data Scraping:..... | 30 |
| 3.5 Scraping Jobs from LinkedIn | 30 |

| | |
|---|-----------|
| 3.5.1 Characteristics: | 30 |
| 3.6 Two Phases of Scraping Jobs from LinkedIn..... | 31 |
| 3.6.1 Scraping Job Links | 31 |
| 3.6.2 Scraping Job Details and Descriptions | 32 |
| 3.7 Problems Faced in Scraping Jobs..... | 34 |
| 3.8 Data Organization | 35 |
| 3.8.1 What is Data Organization and Management?..... | 35 |
| 3.8.2 Why data Organization is Important?..... | 35 |
| 3.9 Process of organizing LinkedIn Jobs dataset | 36 |
| 3.9.1 Structure of Folders:..... | 36 |
| 3.9.2 Concatenation vs. Merging of Files:..... | 37 |
| 3.10 Concatenating Files into One File:..... | 37 |
| 3.10.1 Concatenation of Files Concept:..... | 37 |
| 3.10.2 Concatenating Different Files Separated into a Daily File..... | 37 |
| 3.10.2 Concatenating Daily Files into Weekly Files | 38 |
| 3.10.3 Concatenating Weekly Files into One File..... | 38 |
| CHAPTER NO 4 | 40 |
| Data Preparation 1 | 40 |
| 4.1 Introduction of chapter:..... | 41 |
| 4.2 Basic Concepts related to Data preparation: | 41 |
| 4.2.1 Data preparation and Data Preprocessing:..... | 41 |
| 4.2.2 EDA and Importance of EDA | 42 |
| 4.2.3 Data Cleaning: | 42 |
| 4.2.4 Feature Engineering and its Importance | 43 |
| 4.3 The process of Data Cleaning | 44 |
| 4.3.1 Identifying Null or missing Values in the Jobs details dataset:..... | 44 |
| 4.3.2 Identifying Duplicates: | 45 |
| 4.3.3 Removing Duplicates and Missing Values:..... | 45 |
| 4.3.4 Removing duplicates and null values after merging of jobs details dataset and Jobs description dataset: | 45 |
| 4.3.5 Abnormal Descriptions:..... | 46 |

| | |
|--|----|
| 4.4 EDA Phase 1: | 46 |
| 4.4.1 How EDA is performed? | 46 |
| 4.4.2 Summary of exploring each Feature one by one in table | 47 |
| 4.2.3 Exploring Job Title: | 48 |
| 4.2.4 Exploring Job Info Feature: | 48 |
| 4.2.5 Exploring Job Type Feature: | 49 |
| 4.2.6 Exploring Employee Feature: | 50 |
| 4.3 The Process of Feature Engineering: | 50 |
| 4.3.1 The Summarized Overview of Feature Engineering Phase 1 | 50 |
| 4.3.2 Extracting Company Feature: | 51 |
| 4.3.3 Extracting Location Feature: | 51 |
| 4.3.4 Extracting job posting time feature: | 52 |
| 4.3.5 Extracting the number of applicants feature: | 52 |
| | 53 |
| 4.3.6 Extracting Type of Employment Feature: | 53 |
| 4.3.7 Extracting Experience Level Feature: | 54 |
| 4.3.8 Extracting Remote/On-Site (Work Arrangement) Feature: | 55 |
| | 56 |
| 4.3.9 Extracting Company Employee Size and Company Industry Features: | 57 |
| 4.4 Feature Engineering phase 2: | 58 |
| 4.4.1 Exploring Location Feature: | 58 |
| 4.4.2 Extracting City, Province/Region, and Country Feature: | 59 |
| 4.4.3 Handling Multiple Styles or names of Cities to clarify a strong City analysis: | 60 |
| CHAPTER NO 5 | 61 |
| Filtration of IT jobs from Non-IT jobs | 61 |
| 5.1 Introduction of chapter: | 62 |
| 5.2 Basic Terminologies: | 62 |
| 5.2.1 Categorizing and Filtering IT and Non-IT Jobs: | 62 |
| 5.2.2 Keyword-Based Text Classification and Segmentation of IT and Non-IT Jobs: | 62 |
| | 62 |
| 5.2.3 Keyword Filtering: | 62 |

| | |
|--|----|
| 5.2.4 Keyword Matching: | 62 |
| 5.2.5 Text Classification:..... | 63 |
| 5.2.6 Text Analysis:..... | 63 |
| 5.2.7 Why Categorization is Important: IT and Non-IT Jobs..... | 63 |
| 5.3 Pre-Planning to Filter IT Jobs | 64 |
| 5.3.1 Criteria for IT Jobs: | 64 |
| 4.3.2 What about Digital Marketing, Copywriting, Graphic Design, and other Tech Jobs:..... | 64 |
| 4.3.3 Importance of Job Title and Job Description in Skills Analysis: | 65 |
| 4.3.4 Total Number of Job Titles:..... | 65 |
| 4.3.5 Why it is Difficult to Analyze Job Titles:..... | 65 |
| 4.3.6 Null Values and Incomplete Descriptions in Job Descriptions: | 65 |
| 4.3.7 Keywords Research for IT Jobs..... | 66 |
| 4.3.8 Keyword Analysis: | 66 |
| 4.3.9 Why Keyword Analysis?..... | 67 |
| 4.4 Process of IT Jobs Filtering:..... | 67 |
| 4.4.1 Phase 1: Selection of Proper Keywords Based on user Defined Testing Cases: | 67 |
| 4.4.2 Phase 2: Apply Keywords or Sets of Keywords to Filter Jobs..... | 70 |
| 4.4.3 Filtering IT Jobs Based on Job Titles: | 70 |
| 4.4.4 Example of Segmenting app development jobs: | 71 |
| 4.5 Concatenating Filtered IT Jobs in Filtered_Files-concatenated Folder: | 72 |
| 4.5.1 Drop Duplicates from Concatenated Filtered File: Data Cleaning Phase 2: | 72 |
| 4.6 EDA at the end of Segmentation..... | 73 |
| CHAPTER NO 6 | 74 |
| Data Preparation 2 | 74 |
| 6.1 Introduction to Data Preparation Phase 2:..... | 75 |
| 6.2 Why Categorization of IT Jobs is Needed: | 75 |
| 6.3 Process of Grouping IT Jobs | 76 |
| 6.4 Process of Categorizing IT Jobs into IT Fields and Sub-Fields..... | 77 |
| 6.4.1 Identification of IT Fields:..... | 77 |

| | |
|---|-----------|
| 6.4.2 Building IT Field Group Feature: | 78 |
| 6.4.3 Identification of IT Sub-Fields: | 78 |
| 6.4.4 Building IT Sub-Field Feature:..... | 80 |
| 6.5 Extraction of Tools/Platforms Feature: | 81 |
| 6.5.1 Extracting Tools/Platforms based on Job Titles: | 81 |
| 6.5.2 Extracting Tools/Platforms_2 based on Job Descriptions:..... | 82 |
| 6.6 Working on Extracting Competition Analysis Features: | 82 |
| 6.6.1 Exploring Numbers of Applicants Feature and Job Posting Time: | 82 |
| 6.7 Normalizing the Number of Applicants and Job Posting Time Feature | 84 |
| 6.7.1 Conversion of the Number of Applicants feature into numerical feature: | 84 |
| 6.7.2 Addressing Null Values in the Number of Applicants feature: | 84 |
| 6.7.3 Extracting Job Posting Time in only Numerical Form and Conversion into equivalent hours:..... | 84 |
| 6.8 Building Competition Analysis Features | 85 |
| 6.8.1 Building Feature 'Time to Number of Applicants Ratio': | 85 |
| 6.8.2 Using "Time to Number of Applicants Ratio" for Competition Analysis:..... | 85 |
| 6.8.3 Building a New Feature: Applicants to Hour Ratio | 85 |
| 6.8.4 How Applicants to Hour Ratio Can Be Utilized for Competition Analysis..... | 86 |
| 6.8.5 Which is better? Applicants to Hour Ratio OR Time to Number of Applicants Ratio: | 86 |
| 6.9 Exploring Other Possibilities to Make New Features | 86 |
| 6.9.1 Extraction of Education Requirement Status: | 87 |
| 6.9.2 Why Job Description and Company Info Feature is Not Utilized Much: | 88 |
| CHAPTER NO 7 | 89 |
| Data Visualization and Reporting | 89 |
| 7.1 Data Visualization: | 90 |
| 7.2 Concepts of data visualization: | 90 |
| 7.2.1 Data visualization and its importance:..... | 90 |
| 7.2.2 Data Visualization in Python:..... | 90 |
| 7.2.3 Types of Charts and Graphs: | 90 |
| 7.2.4 Numerical Analysis vs. Categorical Analysis: | 90 |

| | |
|--|-----------|
| 7.3 Data Visualization in Python Project Structure: | 91 |
| 7.3.1 Data Visualization in Skills Analysis: | 91 |
| 7.3.2 Data Visualization in Company Analysis:..... | 91 |
| 7.3.3 Data Visualization in Geography Analysis: | 91 |
| 7.3.4 Data Visualization in Job Nature Analysis:..... | 92 |
| 7.3.5 Data Visualization in Industry Analysis:..... | 93 |
| 7.3.6 Preview of Horizontal Charts, Pie Charts, and Subplots in Python | 93 |
| 7.4 Data Visualization and Dashboard Building in Power BI..... | 94 |
| 7.5 Structure of Power BI Dashboard | 95 |
| 7.5.1 IT Jobs Trends Analysis: | 96 |
| 7.5.2 Skills Analysis based on IT Fields: | 96 |
| 7.5.3 Skills Analysis based on IT Sub-Fields:..... | 96 |
| 7.5.4 Skills Analysis Based on IT Tools/Platforms:..... | 96 |
| 7.5.6 City-Based Geography Analysis: | 96 |
| 7.5.7 Province/Region-based Geography Analysis: | 96 |
| 7.5.8 Country-based Geography Analysis:..... | 97 |
| 7.5.9 Employment Type Analysis: | 97 |
| 7.5.10 Onsite/Remote Analysis: | 97 |
| 7.5.11 Seniority Level (Experience Level) Analysis:..... | 97 |
| 7.5.12 Company Employee Size Analysis: | 97 |
| 7.5.13 Company Industry-based Analysis:..... | 97 |
| 7.5.14 Preview of IT Jobs Trends and Skills Analysis Dashboards: | 97 |
| CHAPTER NO 8 | 99 |
| Results and Discussion | 99 |
| 8.1 Introduction Results and Discussion:..... | 100 |
| 8.2 Skills Analysis:..... | 100 |
| 8.2.1 Skills Analysis based on IT Fields: | 100 |
| 8.2.2 Skills Analysis based on IT Sub-Fields:..... | 102 |
| 8.2.3 Skills Analysis based on IT Tools / Platforms: | 105 |
| 8.3 Company Analysis: | 108 |

| | |
|--|-----|
| 8.4 Geography Analysis: | 111 |
| 8.4.1 Geography Analysis based on City: | 112 |
| 8.4.2 Geography Analysis based on Regions or Provinces: | 114 |
| 8.4.3 Geography Analysis based on Country: | 116 |
| 8.5 Job Dynamics Analysis: | 118 |
| 8.5.1 Job Dynamics Analysis based on Work Arrangement: | 118 |
| 8.5.2 Job Dynamics Analysis based on Type of Employment: | 119 |
| 8.5.3 Job Dynamics Analysis Based on Experience Level..... | 121 |
| 8.6 Industry Preference Analysis: | 122 |
| 8.6.1 Industry Preference Analysis Based on Company Employee Size: | 122 |
| 8.6.2 Industry Preference Analysis Based on Company Employee Size: | 124 |
| 8.7 Analysis of Education/Degree Specification: | 126 |

List of Figures

| | |
|---|----|
| Figure 3.1:LinkedIn Economic Graph | 26 |
| Figure 3.2:Development Data Partnership..... | 26 |
| Figure 3.3: Data Partner Vs Development partner..... | 27 |
| Figure 3.4:Different LinkedIn APIs..... | 28 |
| Figure 3.5:LinkedIn Job page | 31 |
| Figure 3.6:Job ID and Job Links scrapped..... | 31 |
| Figure 3.7:Job details features Extracted | 32 |
| Figure 3.8: Job Description..... | 33 |
| Figure 3.9: About Company | 34 |
| Figure 3.10: Daily Folder Structure | 36 |
| Figure 3.11: Weekly Folder Structure | 36 |
| Figure 3.12: Concatenation into one daily File..... | 37 |
| Figure 3.13: Concatenation of daily Files into weekly Files | 38 |
| Figure 3.14: Concatenation of Weekly Files into 1 Excel File..... | 38 |
| Figure 4.1: Data preparation in data analytics project life cycle | 41 |
| Figure 4.2:Null values in Jobs details features | 44 |
| Figure 4.3:Null values before and after removing duplicates and 273 Null rows | 46 |
| Figure 4.4:Unique values in Job Title..... | 48 |
| Figure 4.5:Top 15 Job Titles..... | 48 |
| Figure 4.6:Unique values in Job Info feature..... | 49 |
| Figure 4.7:Unique values in Job Type Feature | 49 |
| Figure 4.8:Unique values in Employee Feature..... | 50 |
| Figure 4.9:Tree of Secondary Features to be formed by Primary Features | 51 |
| Figure 4.10:Extraction of Company in Job Info String | 51 |
| Figure 4.11:Extraction of Location in Job Info String..... | 52 |
| Figure 4.12:Extraction of Job Posting Time in Job Info..... | 52 |
| Figure 4.13:Extraction of Number of applicants from Job Info Feature | 53 |

| | |
|---|----|
| Figure 4.14: Extraction of Type of Employment from Job Type Feature | 53 |
| Figure 4.15: Extraction of Experience Level from Job Type | 54 |
| Figure 4.16: How Remote onsite is present in two features and Row of change | 56 |
| Figure 4.17: Extraction of Remote/Onsite Feature from two features Job Info and Job Type | 56 |
| Figure 4.18: Extraction of Company Employee Size and Company Industry Feature from Employee Feature | 57 |
| Figure 4.19: Components of Location to be extracted and abnormalities in City Problem | 59 |
| Figure 4.20: Tree of Extracting Location Features | 59 |
| Figure 5.1: Testing 'IT' keyword by user defined function test case 1 | 68 |
| Figure 5.2: Non related Job Titles under 'IT' Keyword | 69 |
| Figure 5.3: List of Droppable Titles | 69 |
| Figure 5.4: Preview of some Finalized List of keywords | 70 |
| Figure 5.5: Droppable Titles while splitting 'App' Jobs..... | 71 |
| Figure 5.6: Filtered Excel and Jupiter Labs Files | 72 |
| Figure 6.1: Grouping of Job Titles into 'Job Title2' Feature | 77 |
| Figure 6.2: Formation of 'IT Field Group' Feature | 78 |
| Figure 6.3: Formation of 'IT Sub-Field Group' Feature..... | 80 |
| Figure 6.4: List of Tools/Platforms..... | 81 |
| Figure 6.5: Extraction of 'Tools and Platforms' Feature from Job titles | 81 |
| Figure 6.6: Extraction of 'Tools/Platforms2' from Job descriptions | 82 |
| Figure 6.7: Unique values in number of applicants Feature | 83 |
| Figure 6.8: Unique values in Job Posting Time Feature..... | 83 |
| Figure 6.9: Converting Number of applicants into numerical | 84 |
| Figure 6.10: Statistical Distribution of 'Number of applicants' Feature | 84 |
| Figure 6.11: Converting job posting time into hours numerical | 85 |
| Figure 6.12: Education and Degree related keywords | 87 |
| Figure 6.13: Value Count of Education Requirement Status 'Yes' or 'No' | 88 |
| Figure 7.1: Horizontal bar chart to show value count of IT fields..... | 93 |

| | |
|--|-----|
| Figure 7.2: Pie chart to show percentage distribution..... | 93 |
| Figure 7.3: Dashboard of Sub-Plots to show detailed Distribution of jobs with respect to other features..... | 94 |
| Figure 7.4: Power BI dashboard for overall Jobs Trends Analysis page 1..... | 98 |
| Figure 7.5: Power BI dashboard for Skills Analysis based on IT Fields..... | 98 |
| Figure 8.1: Distribution of Jobs across IT Fields..... | 101 |
| Figure 8.2: Competition by IT Fields..... | 102 |
| Figure 8.3: Distribution of Jobs across IT Sub-Fields..... | 103 |
| Figure 8.4: Competition by IT Sub-Fields..... | 104 |
| Figure 8.5: Distribution of Jobs across Tools/platforms..... | 106 |
| Figure 8.6: Competition by Tools/platforms..... | 107 |
| Figure 8.7: Distribution of Jobs across Top 20 companies..... | 109 |
| Figure 8.8: Competition by Top twenty companies..... | 110 |
| Figure 8.9: Top twenty companies by competition score..... | 111 |
| Figure 4.10: Distribution of Jobs across top twenty cities..... | 112 |
| Figure 8.11: Competition by Top twenty companies..... | 114 |
| Figure 8.12: Distribution of Jobs across Provinces in Pakistan..... | 115 |
| Figure 8.13: Competition by Provinces in Pakistan..... | 116 |
| Figure 8.14: Distribution of Jobs across Countries..... | 117 |
| Figure 8.15: Competition by Countries..... | 117 |
| Figure 8.16: Distribution of Jobs across Work arrangement (Remote/Onsite)..... | 118 |
| Figure 8.17: Competition across work arrangement..... | 119 |
| Figure 8.18: Distribution of Jobs across Type of Employment..... | 120 |
| Figure 8.19: Competition by Type of Employment..... | 120 |
| Figure 8.20: Distribution of Jobs across Experience Levels..... | 121 |
| Figure 8.21: Competition by Experience level..... | 122 |
| Figure 8.22: Distribution of Jobs across Company Employee Size..... | 123 |
| Figure 8.23: Competition by Company Employee Size..... | 124 |
| Figure 8.24: Distribution of Jobs across Industries..... | 125 |

| | |
|---|-----|
| Figure 8.25: Competition by Industries | 125 |
| Figure 8.26: Distribution of Jobs across Education/Degree Status..... | 127 |
| Figure 8.27: Competition of Jobs across Education/Degree Status..... | 128 |

List of Table

| | |
|---|----|
| Table 3.1: 4 Primary Features Extracted..... | 32 |
| Table 3.2: Total count of Jobs scrapped | 39 |
| Table 4.1: stats of data cleaning before merging two data setts | 45 |
| Table 4.2: Stats of Data Cleaning after merging two data sets | 45 |
| Table 4.3: Summarized EDA of 4 Primary Features | 47 |
| Table 4.4: Secondary Features Derived by Primary Features..... | 50 |
| Table 4.5: Types of Employment..... | 53 |
| Table 4.6: Experience Levels..... | 55 |
| Table 4.7: Four categories of Remote/Onsite | 57 |
| Table 4.8: Same Cities with different Styled names..... | 60 |
| Table 4.9: Repeated words in location to be removed..... | 60 |
| Table 5.1: EDA at the End of Segmentation of IT jobs..... | 73 |
| Table 6.1: IT Fields..... | 77 |
| Table 6.2: IT Sub-Fields | 79 |
| Table 6.3: Exploring unique and missing values in 'Number of applicants 'feature..... | 82 |
| Table 6.4: Exploring unique and missing values in 'Job posting Time 'feature..... | 83 |

CHAPTER NO 1
Introduction of Project

1.1 Project Objectives:

1. **To Conduct a Comprehensive Analysis of IT Jobs in Pakistan:** This objective aims to analyze various aspects of IT jobs in Pakistan, including job trends, skills demand, geographical distributions, company-specific patterns, and industry-based distributions.
2. **To Utilize Data Analytics and Business Intelligence Tools:** This objective focuses on employing modern data analytics techniques and business intelligence tools, such as Python programming, JupyterLab, and Microsoft Power BI, to analyze and visualize LinkedIn Jobs data effectively.
3. **To Explore Key Features of IT Job Market:** This objective seeks to explore key features of the IT job market in Pakistan, including trends over time, geographical variations, and industry-specific dynamics, to provide a comprehensive understanding of the landscape.
4. **To Present Findings Through an Interactive Dashboard:** This objective involves presenting the research findings through an interactive and dynamic online dashboard, allowing stakeholders to explore and interact with the data visually.
5. **To Provide Insights for Stakeholders:** This objective aims to provide valuable insights for various stakeholders, including job seekers, employers, policymakers, and academia, to inform decision-making regarding recruitment strategies, skill development initiatives, and workforce planning in the Pakistani IT sector.
6. **To Identify Challenges and Opportunities:** This objective involves identifying challenges and opportunities within the Pakistani IT job market, highlighting areas for improvement and growth that can be addressed through sound policies and initiatives.
7. **To Contribute to the Development of the IT Industry in Pakistan:** This objective seeks to contribute to the sustainable growth and development of the

IT industry in Pakistan by providing actionable insights and recommendations based on the research findings.

1.2 Background:

The Information Technology (IT) industry is crucial for economic growth and innovation worldwide. In Pakistan, the IT sector has great potential to create jobs and drive progress. However, there's a lack of detailed information on the IT job market, making it hard for job seekers, employers, policymakers, and educators to make informed decisions.

Online job platforms like LinkedIn provide valuable data on job postings and trends. This project uses LinkedIn data from September 26, 2023, to November 26, 2023, to analyze IT jobs in Pakistan. The study looks at job trends, skills demand, job locations, company hiring patterns, and competition in the job market.

Using Python for data analysis and Microsoft Power BI for visualization, the research presents findings through an interactive online dashboard. These insights will help job seekers match their skills with market demand, assist employers in recruitment, guide policymakers in workforce planning, and support educators in developing relevant courses.

The goal is to provide a clear understanding of the IT job market in Pakistan, turning challenges into opportunities for growth and development.

1.3 Problem Statement:

In Pakistan's rapidly evolving digital landscape, the Information Technology (IT) industry plays a pivotal role in driving economic growth and fostering employment opportunities. However, despite its significance, there remains a need for a comprehensive understanding of the dynamics within the IT job market. Existing studies often lack detailed insights into job trends, skills demand, geographical distributions, company-specific patterns, and industry-based distributions, which are essential for stakeholders to make informed decisions regarding recruitment strategies, skill development initiatives, and workforce planning.

Furthermore, while various data analytics and business intelligence tools are available, there is a gap in research that utilizes these tools to analyze real-time job market data, particularly in the context of Pakistan. This gap limits the ability of stakeholders, including job seekers, employers, policymakers, and academia, to access timely and actionable insights that can drive strategic decision-making and contribute to the sustainable growth and development of the IT industry.

Therefore, the primary problem addressed by this research is the lack of a comprehensive and data-driven analysis of IT jobs in Pakistan, utilizing modern data analytics techniques and business intelligence tools. By addressing this gap, this study aims to provide valuable insights that can inform stakeholders and empower them to navigate the dynamic landscape of the Pakistani IT job market effectively.

1.4 Research Questions:

- What are the prevailing job trends within the Information Technology (IT) sector in Pakistan?
- What are the key skills in demand among employers in the Pakistani IT job market?
- How do geographical distributions of IT jobs vary across different regions of Pakistan?
- What are the patterns observed in terms of company-specific hiring practices within the Pakistani IT industry?
- How are IT job opportunities distributed across various sectors and industries in Pakistan?
- What is the intensity of competition among job seekers in the Pakistani IT job market?
- How can data analytics and business intelligence tools be effectively utilized to analyze and visualize real-time job market data?

- What are the implications of the findings for stakeholders, including job seekers, employers, policymakers, and academia?
- How can the insights gained from the analysis contribute to informed decision-making regarding recruitment strategies, skill development initiatives, and workforce planning in the Pakistani IT sector?
- What are the potential challenges and opportunities identified through the research, and how can they be addressed to foster the growth and development of the IT industry in Pakistan?

1.5 Key Features of Project:

1. Job Trends Analysis

This section focuses on analyzing trends in the IT job market in Pakistan over a specified period. It examines fluctuations in job postings, identifies emerging job roles, and investigates patterns in job demand and supply dynamics.

2. Skills Demand Assessment

This section assesses the skills in demand among employers in the Pakistani IT job market. It identifies the most sought-after technical and soft skills, evaluates skill gaps, and explores trends in skill requirements across different job roles and industries.

3. Geographical Distribution Analysis

This section examines the geographical distribution of IT jobs across various regions of Pakistan. It analyzes the concentration of job opportunities in different cities, explores regional disparities in job availability, and identifies emerging IT hubs.

4. Company-Specific Patterns Examination

This section investigates patterns in hiring practices among IT companies operating in Pakistan. It examines recruitment trends, identifies top hiring companies, and explores variations in hiring preferences and strategies across different organizations.

5. Industry-Based Distribution Analysis

This section analyzes the distribution of IT jobs across various sectors and industries in Pakistan. It explores job opportunities in sectors such as software development, telecommunications, finance, and healthcare, providing insights into industry-specific trends and dynamics.

6. Competition Analysis

This section assesses the intensity of competition in the Pakistani IT job market. It examines factors such as the number of applicants per job opening, competition ratios for different job roles, and trends in job seeker behavior to understand the competitive landscape and its implications for job seekers and employers.

1.6 Scope of Project:

Following are Beneficiaries or Stakeholders of the Project, Stakeholders are those people who can benefit from my Project.

1. Job Seekers

The primary beneficiaries of this project are job seekers in the IT industry. By analyzing job trends and identifying the most sought-after skills, the project provides valuable insights that can help job seekers align their career goals with market demand. The data on geographical distributions of IT jobs also enables job seekers to make informed decisions about where to focus their job search, whether considering relocation or remote work opportunities.

2. Employers

For employers, the project offers critical information to refine recruitment strategies. Understanding the latest job trends and skill shortages helps employers to better target their recruitment efforts and attract the right talent. Additionally, insights into company-specific hiring patterns allow employers to benchmark their practices against industry standards and competitors, leading to more effective and competitive hiring processes.

3. Policymakers

Policymakers stand to gain significantly from this research as it provides a detailed analysis of the IT job market, aiding in workforce planning and development. By aligning education and training programs with industry needs, policymakers can ensure that the workforce is equipped with the necessary skills. Furthermore, understanding job market dynamics helps in creating policies that support the growth of the IT sector, thereby boosting overall economic development.

4. Academic Institutions

Academic institutions can utilize the findings of this project to enhance their curriculum and training programs. By tailoring courses to meet current and future demands of the IT job market, educational institutions can better prepare students for successful careers. The project also facilitates stronger collaborations between academia and industry, promoting internships, research partnerships, and job placements that benefit both students and employers.

5. Industry Analysts and Researchers

Lastly, industry analysts and researchers benefit from the comprehensive data and insights provided by this project. The findings serve as a foundation for further research and market analysis, contributing to a deeper understanding of the IT job market in Pakistan. Analysts can track changes and trends over time, using this information to make predictions and guide strategic decisions, ultimately fostering a more robust and dynamic IT sector.

1.7 Feasibility Study:

The primary objectives of the feasibility study were to:

- Evaluate the availability and accessibility of relevant data sources for conducting the research.
- Assess the feasibility of employing data scraping techniques to gather LinkedIn Jobs data.
- Determine the suitability of data analytics tools, including Python programming and Microsoft Power BI, for analyzing and visualizing the collected data.

- Identify potential challenges and limitations associated with the research methodology.
- Determine the potential value and impact of the research findings for stakeholders in the Pakistani IT sector.

1.8 Methodology:

The following are Techniques used in Data Science project with a focus on data analysis and Business Intelligence

Data Collection (Phase I)

To initiate the project, relevant job market data will be gathered from prominent online job portals such as LinkedIn.

Data Preprocessing (Phase II)

Following data collection, the collected dataset will undergo meticulous cleaning and preprocessing procedures to ensure accuracy and consistency. This process involves handling missing values, standardizing data formats, and eliminating duplicates to prepare a refined dataset for subsequent analysis.

Exploratory Data Analysis (EDA) (Phase III)

Exploratory Data Analysis (EDA) will be conducted to unveil underlying patterns, trends, and relationships within the dataset. Utilizing statistical techniques and visualizations, insights will be extracted to guide further analysis.

Feature Engineering (Phase IV)

Feature Engineering will be implemented to enhance the predictive power of the dataset. This involves creating new features or transforming existing variables to augment the dataset's depth and efficacy. Techniques such as extracting key information from job descriptions or calculating additional metrics will be employed to enrich the dataset.

Insights and Reporting (Phase V)

The culmination of the analysis phase involves interpreting the results derived from the models and applications developed. Actionable insights will be generated, and comprehensive reports and dashboards will be created to present findings to stakeholders. This facilitates informed decision-making processes for job seekers, employers, and recruiters alike.

1.9 Tools and Technologies:

Data Collection:

- Web scraping tools such as Python Libraries BeautifulSoup or Selenium for collecting job data from online portals.

Data Cleaning and Preprocessing:

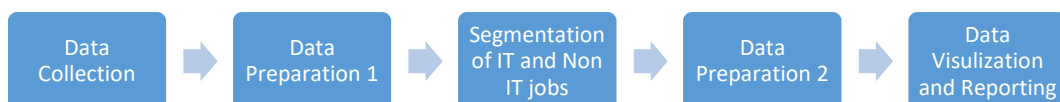
- Python with libraries like Pandas and NumPy for data manipulation.
- Regular expressions (Regex) for pattern matching and text processing.

Data Analysis and Visualization:

- Python libraries including NumPy, Pandas, Matplotlib, and Seaborn for data analysis and visualization.
- Business Intelligence (BI) tools like Power BI and Tableau for creating interactive dashboards and reports.

Cloud Computing and Storage:

- Utilization of BI platforms like NoviPro for dashboard sharing and collaborative analysis.



CHAPTER NO 2
Literature Review

2.1 Introduction of chapter:

The information technology (IT) sector in Pakistan has witnessed significant growth in recent years, contributing substantially to the country's economy. Understanding the dynamics of the IT job market is crucial for aligning educational programs, informing policy decisions, and guiding job seekers. This literature review explores existing research on IT Jobs market analysis using online job portal data, with a focus on the Pakistani context. By leveraging Jobs data and data science techniques, this study aims to unlock insights into Pakistan's IT job market, building on the foundation laid by prior research.

Here I discussed Four Existing studies that are somehow relevant to my Project. Each Research study is discussed with objectives, methodology, findings and how it relevant to my research, also discusses how my research fills the Gap of Research by Previous Studies.

2.2 Exploring Industrial Demand Trends in Pakistan Software Industry

Using Online Job Portal Data [1]:

2.2.1 Objective:

The research paper explores the demand trends in the Pakistani software industry by analyzing job advertisements on Rozee. pk, a popular online job portal.

2.2.2 Methodology:

i. Data Extraction:

Data was collected from Rozee. pk, focusing on software and web development job ads, using the "Web Scraper" Chrome extension. A total of 492 job postings were extracted.

ii. Data Preprocessing:

The raw data was cleaned to remove inconsistencies and missing information, and categorized by job roles, programming languages, qualifications, city, age limits, and salaries. Job titles and descriptions were standardized to facilitate accurate analysis.

iii. Data Analysis:

Quantitative analysis, including frequency distribution and comparative analysis, was performed on the preprocessed data. The results were visualized using pie charts, bar charts, column charts, and geographic heat maps.

iv. **Results Interpretation:**

The analysis provided insights into job market trends, highlighting in-demand roles and skills for students and job seekers. Employers, policymakers, and student counseling organizations can use these insights for planning, decision-making, and career guidance.

2.2.3 Findings of Research Paper:

The study analyzes job ads on Rozee.pk to reveal demand trends in Pakistan's software industry. Key findings include:

i. **Job Roles:**

Web developers are the most in-demand, constituting 39% of postings, followed by software developers at 29%. Other roles include Android developers (8%), iOS developers (5%), game developers (3%), and graphics designers (3%).

ii. **Programming Languages:**

PHP is the most demanded programming language at 30%, followed by ASP.NET at 17% and Java at 16%. JavaScript accounts for 13% and C for 7% of the demand.

iii. **Salaries:**

Information security professionals earn the highest salaries, ranging from PKR 250,000 to 500,000, while software architects earn between PKR 100,000 and 300,000. These highly paid roles are not the most in-demand.

iv. **Gender Preferences:**

Most job postings (85%) do not specify a gender preference, indicating minimal gender discrimination. Only 13% prefer male candidates, and 2% prefer female candidates.

v. **Age and Qualification:**

The typical age range for job applicants is 18 to 50 years. Most job postings (84%) require a bachelor's degree, with 6% requiring a master's degree, 6% requiring intermediate/A-level, 1% requiring matriculation/O-level, and 3% other qualifications.

vi. **Geographical Distribution:**

Lahore accounts for 39% of job postings, followed by Karachi with 28% and Islamabad with 15%. Other locations include unspecified or virtual jobs.

vii. **Summary of Key Findings:**

Web and software developers are highly sought after. PHP, JavaScript, ASP.NET, and Java are the top programming languages.

viii. **Salary Insights:**

Information security professionals and software architects are the highest-paid roles. However, these roles are less in demand compared to others.

ix. **Gender Equality:**

Most job postings do not specify gender preference, suggesting minimal gender discrimination in job opportunities.

x. **Age and Educational Requirements:**

A significant majority of job postings prefer candidates with a bachelor's degree, and the typical age range for applicants is 18 to 50 years.

xi. **City-wise Job Distribution:**

Major cities like Lahore, Karachi, and Islamabad dominate the job market in the software industry.

2.2.4 Relevance of the Research:

Both my research and the study "Exploring Industrial Demand Trends in Pakistan's Software Industry Using Online Job Portal Data" offer valuable insights into the IT job market, focusing on job roles, programming languages, geographic analysis, and targeting job seekers and employers.

i. Job Roles

My research analyzed job roles across all IT fields, encompassing 6 major fields and 22 subfields, with a total of 1,794 titles. In contrast, the research paper "Exploring Industrial Demand Trends in Pakistan's Software Industry Using Online Job Portal Data" focuses specifically on software engineering job roles. This broader scope in my study provides a comprehensive understanding of the entire IT job market.

ii. Programming Languages

I conducted an extensive analysis of a wide range of programming languages, frameworks, and tools. The other research paper, however, explored a more limited selection of programming languages. This wider analysis in my research offers deeper insights into the skill sets demanded across the IT industry.

iii. Geographic Analysis

My research includes a detailed city analysis covering 49 cities and all regions/provinces of Pakistan. The research paper on software industry trends focused on job data from a few major cities. This comprehensive geographic analysis in my study helps in understanding regional job market trends and demand variations.

iv. Targeting Job Seekers and Employers

Both studies aim to provide valuable insights for job seekers and employers. By understanding the demand for various IT roles, programming languages, and regional job trends, both pieces of research offer essential information to aid in recruitment strategies and career planning, benefiting stakeholders in the IT sector.

2.3 Labor Market Analysis Using Big Data the Case of a Pakistani Online Job Portal [2]

2.3.1 Objective:

The primary objective of this research paper is to provide new descriptive insights about labor market conditions and the supply and demand of skills in Pakistan by utilizing data from an online job portal, Rozee. pk. The study aims to address the mismatch between the skills produced by the education and training systems and those demanded by the private sector, particularly in the context of educated youth facing high unemployment rates despite a well-educated labor force.

2.3.2 Methodology:

The research analyzes data from 412,000 jobs posted on the Rozee.pk platform, utilizing four types of datasets to provide a comprehensive view of labor market dynamics. The jobseekers' data is based on resumes created on Rozee. pk and includes demographics, education, professional experience, skills possessed, and current and desired salaries. Employers' data is derived from profiles created on Rozee. pk. The job posting archive contains detailed information such as job titles, descriptions, qualifications, and salary ranges. Transactions data includes records of job applications, noting who applied for which postings and the date and time, as well as incomplete information on shortlisted candidates, which is voluntarily updated by employers. Spanning from 2012 to 2019, this dataset enables detailed analyses of job market trends, the matching process between job seekers and employers, the demand and supply of skills, and other labor market dynamics.

Unlike traditional labor force surveys that have a significant time lag, online job portal data is updated in real time, offering immediate insights. This data includes rich text information, providing detailed descriptions of job titles, qualifications, skills, and experiences required for job postings, as well as information from job seekers' resumes. Additionally, the data encompasses the actual processes of job matching, offering insights into which qualifications and skills are more likely to lead to successful job matches. The linked data between employers and job seekers allows for a comprehensive analysis of both the demand and supply sides of the labor market.

2.3.3 Findings of Research

The study reveals a mismatch between the supply of highly educated workers and the demand for specialized skills in certain industries, such as information and communications technology (ICT), which lack workers with the necessary expertise. It underscores the importance of exact skill matches, noting that job applicants whose qualifications precisely align with job requirements are more likely to be shortlisted, while both underqualified and overqualified candidates have lower chances. The job portal data primarily represents the high-skill segment of the labor market, with postings offering higher salaries than the national average, and job seekers being younger and better educated. Industry-specific trends show higher job market

tightness in sectors like ICT, indicating more available jobs relative to job seekers. The study also explores gender preferences in job ads and highlights the high demand for specific skills, such as programming, across various industries. These findings emphasize the need to align education and training programs with the private sector's skill demands to address the challenges faced by the educated young labor force in Pakistan.

2.3.4 Relevance:

My research on IT jobs in Pakistan aligns closely with the study "Labor Market Analysis Using Big Data: The Case of a Pakistani Online Job Portal," as both address significant labor market dynamics and skill mismatches. Both studies identify a mismatch between the qualifications of highly educated workers and the specialized skills demanded by industries such as ICT, emphasizing the need for precise skill matches for successful job placements. While my research provides a comprehensive analysis of IT jobs using LinkedIn data, encompassing job trends, skills demand, geographical distributions, and industry-based patterns, the referenced study focuses on data from Rozee. pk, highlighting job market tightness, salary trends, and gender preferences in job ads. Both studies aim to inform job seekers, employers, policymakers, and academia, aiding in the development of recruitment strategies, skill development initiatives, and workforce planning. By integrating the findings from both studies, stakeholders can gain a richer, more detailed understanding of the IT job market in Pakistan, supporting economic growth and job creation in the IT industry.

2.4 Skills Set Required for Web Developers in Pakistan [3]

2.4.1 Objective:

The objective of the research paper "Skills Set Required for Web Developers in Pakistan" is to identify and analyze the current demand for skills in the Pakistani web development industry. This includes understanding the specific job roles, programming languages, technical skills, and other qualifications that are in demand. The findings aim to assist various stakeholders such as student counseling groups, job seekers, researchers, industry experts, curriculum developers, government planners, and decision-makers in aligning their efforts with industry demands. By analyzing job advertisements, the study seeks to determine the specific skills required for web development jobs in Pakistan, providing valuable insights for stakeholders to

make informed decisions. Additionally, it aims to understand market trends in the Pakistani web sector and their impact on the economy and IT job market, facilitating better job matching by aligning job seekers' skills with market demand.

2.4.2 Methodology:

The methodology of the research paper "Exploring Industrial Demand Trends in Pakistan Software Industry Using Online Job Portal Data" involves several stages to gather, process, and analyze data on the skill requirements for web developers in Pakistan. Data was collected from three major online job portals—Rozee.pk, Mustaqbil.com, and Indeed. com—up to October 25th, 2022, resulting in a sample of 151 job postings. Detailed information, including company details, job specifics, and technical requirements, was extracted from each advertisement. The data was then cleaned, normalized, and categorized to ensure consistency and relevance. Analytical techniques such as frequency, trend, and comparative analyses were employed to assess the demand for various skills, programming languages, tools, and frameworks. The findings were visualized using charts and graphs, and interpreted to provide insights into the current skill demand trends in Pakistan's software industry, particularly for web developers.

2.4.3 Findings of Research Study:

The research paper "Skills Set Required for Web Developers in Pakistan " examines the current demand for web development skills in Pakistan based on job postings from major online portals. The findings highlight that JavaScript, PHP, HTML/CSS, and Python are the most demanded programming languages. React.js and Angular are preferred front-end frameworks, while Laravel is popular for back-end development. MySQL is the most commonly required database, with MongoDB gaining traction. Essential tools include Git and Docker, reflecting modern development and deployment practices. Job postings show significant opportunities for both entry-level and experienced developers, with communication and problem-solving skills highly valued. There is a growing demand for full-stack developers and remote work options are increasingly mentioned. Most job postings require a bachelor's degree in computer science or a related field, with certifications providing an added advantage. These findings

provide valuable insights for web developers, educators, and industry stakeholders to align their skills and strategies with market demands in Pakistan.

2.4.4 Relevance to My Research

My research focuses on the entire IT job market, encompassing various IT fields, including web development, which is a significant segment of the overall market. The findings of the paper "Skills Set Required for Web Developers in Pakistan " are relevant as they provide detailed insights into the web development segment, which complements my broader analysis. This research highlights key programming languages, frameworks, tools, and skills demanded specifically for web development, aligning with the more extensive range of IT roles, technologies, and trends I analyzed. Both studies aim to inform job seekers, educators, and industry stakeholders about market demands, enhancing the overall understanding of the IT job landscape in Pakistan.

2.5 DIGITAL PAKISTAN: OPPORTUNITIES & CHALLENGES [4]:

2.5.1 The Objective:

The objectives of the study are to identify and analyze barriers hindering the computerization process in Pakistan, focusing on bureaucratic, political, educational, and social factors. Through primary data analysis using structured questionnaires and statistical tools like correlation, regression analysis, and t-tests, the study aims to provide empirical evidence to policymakers. It also seeks to investigate the significant impact of various variables on shaping IT in Pakistan, highlighting inconsistencies in IT policy, negative administrative attitudes, cumbersome procedures, and weak implementation. Additionally, the study examines the influence of unstable political environments, insufficient infrastructure, and alignment issues on IT adoption in Pakistan, while also highlighting positive indicators such as government incentives and growing private sector interest. These objectives collectively aim to offer a comprehensive understanding of the opportunities and challenges in IT adoption in Pakistan, providing valuable insights for policymakers and stakeholders in the country.

2.5.2 Methodology

The methodology used in the study involved both secondary and primary data collection methods, utilizing literature surveys and questionnaires. A pilot study was conducted to optimize constructs and develop a structured questionnaire. Data analysis employed descriptive and inferential statistical tools, including correlation, regression analysis, and t-test. The study emphasized the importance of IT education in Pakistan, highlighting issues such as outdated curricula and fraudulent practices in computer training institutions. It also addressed the significance of quality IT professionals and challenges related to brain drain and lack of qualified trainers. Hypotheses were formulated based on literature review and empirical data to investigate barriers to computerization in Pakistan. Overall, this methodology facilitated a comprehensive analysis of opportunities and challenges in IT adoption, focusing on education, IT professional quality, and institutional practices.

2.5.3 Key Findings:

- i. The major challenges in IT adoption in Pakistan include inconsistent IT policy, negative administrative attitudes, cumbersome procedures, weak implementation, lack of IT knowledge, unstable political environment, inadequate physical and legal infrastructure, and shortage of IT professionals.
- ii. Negative administrative attitudes can affect IT adoption by creating resistance to change within organizations, hindering progress, and slowing down the implementation of IT initiatives.
- iii. Lack of IT knowledge is a challenge in Pakistan because it can impede the effective implementation of IT initiatives, lead to operational inefficiencies, create skill mismatches, hinder necessary training and support for employees, and impact strategic decision-making processes.

2.5.4 Relevance:

Understanding the major challenges in IT adoption in Pakistan is relevant to my research as it provides valuable insights into the obstacles hindering technological advancement in the country. By addressing issues such as negative administrative attitudes and the lack of IT

knowledge among bureaucratic personnel, my research can contribute to identifying solutions to enhance IT adoption and innovation. Additionally, highlighting the importance of building a digitally literate workforce aligns with the objectives of my study, which aims to analyze IT jobs and skill demands in Pakistan. By addressing these challenges, my research can help bridge the gap between the demand for skilled IT professionals and the existing knowledge base, thereby facilitating smoother IT adoption and fostering technological growth in Pakistan.

2.6 Summary and Identified Gaps in Previous Research Studies

The four research studies reviewed provide valuable insights into various aspects of Pakistan's IT job market but also highlight several gaps that my research addresses. "Exploring Industrial Demand Trends in Pakistan Software Industry Using Online Job Portal Data" focuses narrowly on job roles and programming languages within the software sector based on Rozee.pk data, offering a limited geographical scope and an absence of interactive data tools. "Labor Market Analysis Using Big Data: The Case of a Pakistani Online Job Portal" presents a broader labor market overview using Rozee.pk data, identifying skill mismatches but lacking real-time insights and comprehensive regional analysis. "Skills Set Required for Web Developers in Pakistan" examines specific skills for web developers through job postings on three portals, offering detailed but segmented insights restricted to web development. Lastly, "Digital Pakistan: Opportunities & Challenges" identifies systemic barriers to IT adoption, such as inconsistent policies and inadequate IT knowledge, but does not delve deeply into job market trends or skill demands. These studies lack a holistic and interactive approach, comprehensive regional coverage, and in-depth competition analysis. My research fills these gaps by leveraging LinkedIn data to provide a detailed analysis of IT jobs across 6 major fields and 22 subfields, covering 49 cities, with dynamic online dashboards and competition insights, offering a more extensive, real-time, and actionable understanding of Pakistan's IT job market.

CHAPTER NO 3
Data Collection and Organization

3.1 Data Collection

This chapter outlines the methodologies for gathering data on IT jobs in Pakistan, covering data research, strategy formulation, and scraping techniques. It also discusses organizing datasets into folders, merging them into a single file, and tallying scraped jobs. Additionally, it addresses challenges encountered during the process, including technical hurdles and data quality issues, providing valuable insights for researchers and analysts navigating the complexities of data collection in the Pakistani IT job market.

3.2 Concepts Related to Data Collection

3.2.1 What is Data Collection?

Data collection represents the initial practical phase of the data science or data analysis project life cycle. It serves as the foundational step where relevant data is gathered from various sources to form the basis for subsequent analysis and decision-making. Effective data collection involves identifying pertinent data sources, devising strategies for data acquisition, and employing appropriate methods such as web scraping, surveys, or partnerships with data providers. This phase is crucial as the quality and comprehensiveness of the collected data significantly impact the accuracy and reliability of the ensuing analyses and insights. Therefore, meticulous planning, execution, and validation are essential during the data collection stage to ensure the success of the overall project.

3.2.2 Role of data strategy in effective data collection

Data collection is the foundational phase of any data science or data analysis project, and an effective data strategy is crucial for its success. A well-defined data strategy guides the selection of data sources, methodologies, and tools, ensuring the systematic acquisition of high-quality data relevant to the project's objectives. It helps mitigate risks, optimize resource allocation, and adapt to evolving data requirements, ultimately enhancing the likelihood of achieving meaningful insights and actionable outcomes.

3.2.3 Data Research

The initial step in data collection was identifying which job portals to use for gathering job data. This involved evaluating the most popular and relevant job portals for IT jobs in Pakistan. Additionally, it involved determining the types of datasets required and searching for proper datasets.

3.2.4 Questions Arising in Data Research

- i. Which job portals have the most comprehensive and reliable data for IT jobs in Pakistan?
- ii. What specific information is necessary to analyze the IT job market effectively?
- iii. How can we ensure the accuracy and completeness of the datasets?
- iv. What legal and ethical considerations must be taken into account when collecting data from job portals?
- v. Are there any existing partnerships or collaborations that can facilitate data access?
- vi. What are the most effective methods for integrating and standardizing data from multiple sources?

3.3 Data Collection Methods

3.3.1 Three Ways to Collect Data:

- i. **Data Partnership:** Collaborating with job portals to obtain direct access to their job listings databases.
- ii. **Data Scraping Using LinkedIn APIs:** Leveraging LinkedIn's official APIs to extract data systematically and legally.
- iii. **Data Scraping Without APIs:** Implementing alternative scraping methods to collect data where APIs were not available.

3.3.2 Data Scraping Without APIs

When data partnerships and APIs were not feasible, data scraping without APIs was employed using both automated tools and programming techniques.

- i. **Using Automated Tools:** Automated tools were utilized for scraping data from job portals. These tools varied in functionality and cost.

Types of Automated Tools:

- **Paid Tools or Credit Tools:** Tools like Octoparse, Mozenda, and ParseHub which require a subscription or credits to use.
 - **Free Tools:** Open-source or free tools such as WebHarvy and OutWit Hub.
- ii. **Using Programming Languages and Frameworks:** For more customized and flexible data scraping, programming languages, particularly Python, were used.

3.3.3 Python for Scraping: Python Frameworks

Python offers several powerful frameworks for web scraping, which were instrumental in collecting the data.

- i. **Beautiful Soup:** A Python library used for parsing HTML and XML documents. It creates a parse tree for web pages that can be used to extract data easily.
- ii. **Selenium:** A web testing library that can be used for web scraping. It automates browser interaction and is particularly useful for scraping dynamic content.
- iii. **Scrapy:** An open-source and collaborative web crawling framework for Python. It provides a comprehensive suite for building and running web spiders.

3.3.4 Limitations of Software Tools Used for Data Scraping

While automated tools and programming frameworks are powerful, they come with limitations that must be acknowledged.

- i. **Accuracy and Completeness:** Some tools may miss or incorrectly scrape data, leading to gaps or errors in the dataset.

- ii. **Legal and Ethical Considerations:** Scraping data without proper authorization can breach terms of service of websites and raise ethical concerns.
- iii. **Resource Intensive:** Scraping, particularly without APIs, can be resource-intensive and require significant computational power and time.
- iv. **Maintenance:** Web scraping scripts and tools need constant maintenance to adapt to changes in website structures.

3.3.5 Why I Chose LinkedIn Jobs for Analysis

- i. **Largest Job Portal Used in Pakistan:** LinkedIn is one of the largest job portals globally and is extensively used in Pakistan. Its widespread adoption among professionals and companies makes it a crucial platform for analyzing IT jobs in the country.
- ii. **Comprehensive Job Listings:** LinkedIn offers a vast array of job listings, including numerous opportunities in the IT sector. Its extensive database provides a comprehensive view of the job market, making it an ideal source for research.
- iii. **Detailed Job Information:** Job postings on LinkedIn typically include detailed information such as job titles, descriptions, required skills, qualifications, and company information. This richness of detail is essential for conducting a thorough analysis of IT jobs.

3.4 Problems Faced in Getting Jobs Data Using Data Partnership and LinkedIn APIs

I explored different options for data partnerships to obtain a substantial amount of LinkedIn job data. Despite the promising potential of these partnerships, several challenges arose.

3.4.1 LinkedIn Economic Graph

The LinkedIn Economic Graph is a research program that provides access to LinkedIn's vast datasets for academic and policy research. However, when I tried to research this option, I

found that it is currently closed to new researchers. This closure significantly limits the ability to access detailed job market data directly from LinkedIn through this program.

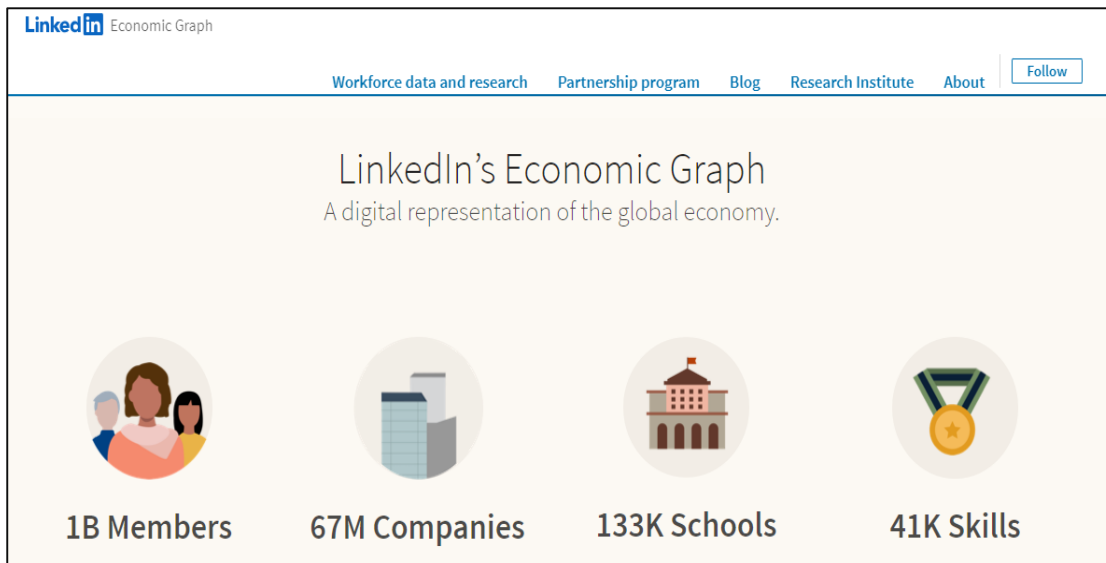


Figure 3.1: LinkedIn Economic Graph

3.4.2 Development Data Partnership

Another promising avenue was the Development Data Partnership, which involves LinkedIn partnering with major development organizations such as the World Bank and the World Economic Forum. These partnerships aim to leverage data for social good, providing valuable insights and data for large-scale research projects.



Figure 3.2: Development Data Partnership

| HOW TO ENGAGE | |
|--|--|
| <p>DATA PARTNER</p> <p>The Partnership unlocks public good opportunities from proprietary data in a secure, responsible manner. Through partnership, Data Partners can open markets in emerging economies, discover new data methods that inform future products, and increase staff skills through collaboration and secondment opportunities.</p> | <p>DEVELOPMENT PARTNER</p> <p>The Partnership is open to donors and entities engaged in international development work. Members have access to the Data Partnership Portal and are invited to participate in exchanges and training activities.</p> |

Figure 3.3: Data Partner Vs Development partner

3.4.3 Challenges Encountered:

- i. **Access Restrictions:** I attempted to access data through this partnership by reaching out via emails and LinkedIn messages. However, I discovered that this partnership is inaccessible to individual college students.



- ii. **Institutional Requirements:** The Development Data Partnership requires collaborations to be established for universities or large institutions with formal

partnerships with LinkedIn. As an individual student, I lacked the institutional backing necessary to qualify for access under this program.

- iii. **Administrative Barriers:** Even if my university had such partnerships, gaining the necessary permissions and approvals to utilize these resources can be a lengthy and bureaucratic process, often involving multiple levels of administrative review.

3.4.4 LinkedIn APIs

In addition to exploring data partnerships, I also investigated the use of LinkedIn APIs to obtain job data. However, after extensive research and effort, I discovered that LinkedIn does not provide APIs specifically for accessing job listing data. This limitation severely restricts the ability to collect job data from LinkedIn using official APIs directly.

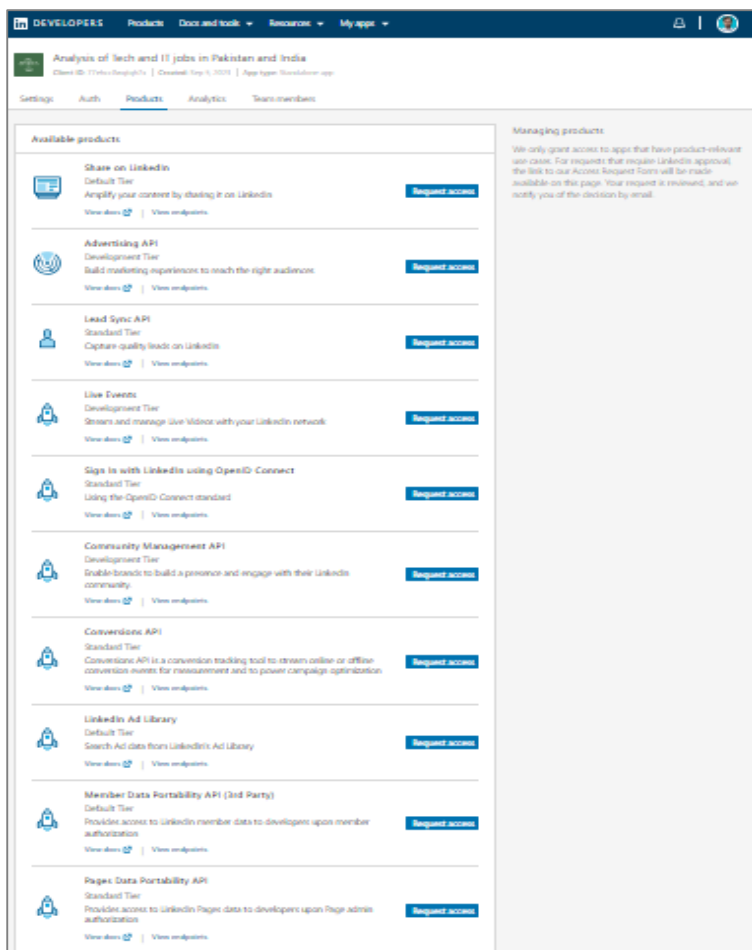


Figure 3.4: Different LinkedIn APIs

3.4.5 Third-Party Platforms

Although LinkedIn itself does not offer APIs for job data, there are third-party platforms that provide APIs to scrape LinkedIn data. However, these third-party services come with their own set of challenges:

- i. **Cost:** Third-party scraping services are typically not free and often require a subscription or payment, which can be a significant barrier for students or individual researchers.
- ii. **Reliability and Legitimacy:** Relying on third-party services can introduce concerns about the reliability and legitimacy of the data collected, as well as potential legal and ethical issues related to data scraping.

3.4.6 General Issues with Data Partnerships and APIs:

- i. **Negotiation and Approval Processes:** Establishing data partnerships typically requires extensive negotiations and formal agreements, which can be time-consuming and complex, especially for students or individual researchers.
- ii. **Data Privacy and Security:** Ensuring compliance with data privacy laws and maintaining the security of sensitive information is a significant concern, which can further complicate the partnership process.
- iii. **API Access Restrictions:** LinkedIn APIs are subject to strict access controls and rate limits, which can restrict the volume of data that can be collected. Access to these APIs often requires approval from LinkedIn and is typically granted to partners with specific, predefined use cases.
- iv. **Technical and Implementation Challenges:** Integrating LinkedIn APIs requires technical expertise and can be complex, particularly if the data needs to be combined with other sources for a comprehensive analysis.
- v. **Cost Implications:** Data partnerships and API access can involve significant costs, including subscription fees or data access charges, which may not be feasible for individual researchers or students.

3.4.7 Reasons for Choosing Python for LinkedIn Job Data Scraping:

- i. **Exhaustive Research:** Explored APIs and automated tools but found limitations and challenges.
- ii. **Focus on Python Libraries:** Shifted attention to Python frameworks due to flexibility and customization.
- iii. **Successful Scraping:** Achieved successful scraping with BeautifulSoup for HTML parsing and Selenium for dynamic content.
- iv. **Flexibility and Effectiveness:** Python provided the flexibility to customize scraping processes and effectively handle LinkedIn's dynamic content.
- v. **Scalability and Support:** Python frameworks offer scalability for large-scale scraping and benefit from a robust community and extensive documentation for support.

3.5 Scraping Jobs from LinkedIn

3.5.1 Characteristics:

- i. **All Jobs Filters:** To effectively scrape job data from LinkedIn, various filters available on the LinkedIn job search page are utilized. These filters can include job title, location, industry, experience level, and more. Applying these filters helps in narrowing down the search to relevant job postings, making the scraping process more efficient and targeted.
- ii. **Every Day Scraping after 24 Hours:** To ensure the data is up-to-date and captures the latest job postings, the scraping process is scheduled to run every 24 hours. This daily scraping routine helps maintain a current dataset, reflecting the most recent trends and opportunities in the IT job market.
- iii. **Only 40 Pages of Jobs Is Accessible:** LinkedIn imposes limitations on the number of job listing pages that can be accessed. Typically, only the first 40 pages are accessible, which may restrict the total number of job postings that can be scraped at any given time. This limitation necessitates regular and frequent scraping to gather a comprehensive dataset.

- iv. **Scraping Regularly Job Links:** Regularly scraping job links ensures that all new job postings are captured as they appear. This involves extracting the URLs of job postings, which are then used in subsequent phases to scrape detailed job information and descriptions.

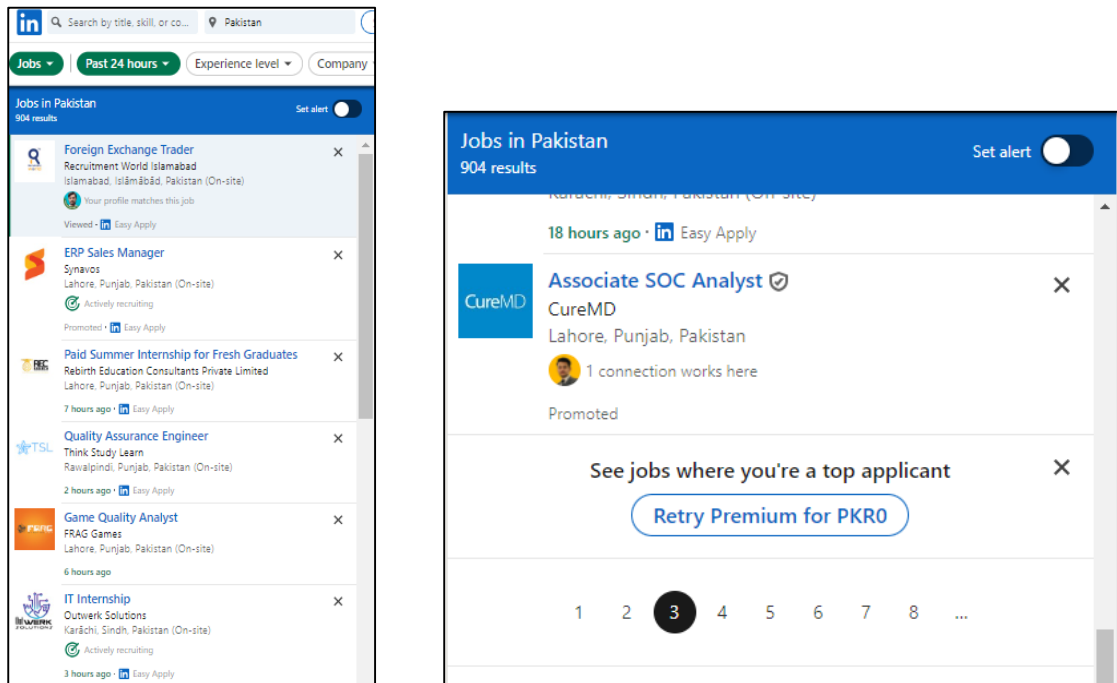


Figure 3.5: LinkedIn Job page

3.6 Two Phases of Scraping Jobs from LinkedIn

3.6.1 Scraping Job Links

Why Scraping Job Links First?

Scraping job links first allows for the initial collection of URLs for job postings. This step is crucial because it lays the foundation for the next phase, where detailed job information is extracted. By collecting job links first, it ensures that only relevant and new job postings are processed in detail.

| Job Title | Job ID | Job Link |
|------------------------------|------------|--|
| Digital Marketing Executive | 3730677619 | https://www.linkedin.com/jobs/view/3730677619/?eBP=CwEAAAGK8SUWgff9BdgNvUZbjMxW9qQbSMuNCMtpeI4bGOPUw_-sta3jG2ctzi9bBmbzkHD80jmtJK7Z5 |
| DM Technical Performance | 3731408240 | https://www.linkedin.com/jobs/view/3731408240/?eBP=CwEAAAGK8SUWgbb88kFASOP110muzcy9FAGplqMVvBjOYU6cQAM48KOFZME4XtPSCG8ReG484weAsyYkqy |
| Estimation Engineer (MV Swit | 3727206659 | https://www.linkedin.com/jobs/view/3727206659/?eBP=CwEAAAGK8SUWgY5zeuZWYdlper8aAKThzOmU6u77qMq-amzgcPI3_Th3CRfZSTP6zCWA3cM_YaDnNCQhGzX |

Figure 3.6: Job ID and Job Links scrapped

Here the Job ID is extracted from the job link, Every Job has a unique ID.

3.6.2 Scraping Job Details and Descriptions

How Job Details and Descriptions Were Scraped by Job Links at Any Time?

Using the previously scraped job links, detailed job information and descriptions are scraped.

This phase involves visiting each job link and extracting the necessary data.

This phase is divided into two sub-phases:

i. Scraping Job Details

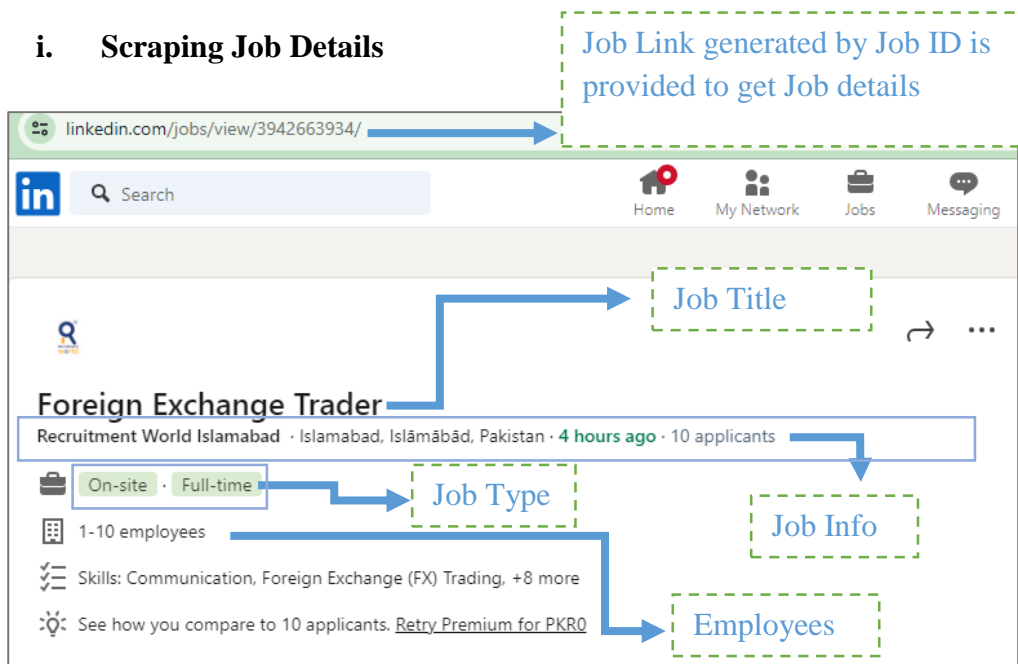


Figure 3.7: Job details features Extracted

Four main features are scraped from the job postings:

Table 3.1: 4 Primary Features Extracted

| | Features Names | A short description of the feature |
|---|----------------|--|
| 1 | Job Title | A short introduction to the job |
| 2 | Job Info | contains the company name, location, remote/onsite status, time passed to job posting, and number of applicants. |

| | | |
|---|-----------|---|
| 3 | Job Type | Contains information about the nature of jobs such as Type of Employment, Experience Level, and remote /onsite status |
| 4 | Employees | Contains extra information about the company such as the Company's Employee Size and Company's Industry |

ii. Scraping Job Descriptions:

Two main features are scraped:

- **Job Descriptions:** Detailed descriptions of the job responsibilities, requirements, and qualifications.

| |
|--|
| <p>About the job</p> <p>Job Description: The Business Analyst (BA) is responsible for evaluating business processes, anticipating requirements, uncovering areas for improvement, and developing and implementing solutions.</p> <p>Role and Responsibilities: The role involves requirement elicitation, analyzing business needs, translating requirements, designing wireframes, defining process flows, and technical document writing (SRS, BRDs, User Manuals).</p> <p>Requirements:</p> <ul style="list-style-type: none"> • We are seeking a Business Analyst with a BSCS/BSIT degree • Experience in content writing or business analysis. • Sharp fresh graduates with excellent English communication skills are also welcome. • Ability to thrive in a dynamic and evolving environment. |
|--|

Figure 3.8: Job Description

- **Company Info:** Extra Information about the company such as Company Industry, Company Employee Size, and Number of Employees On LinkedIn

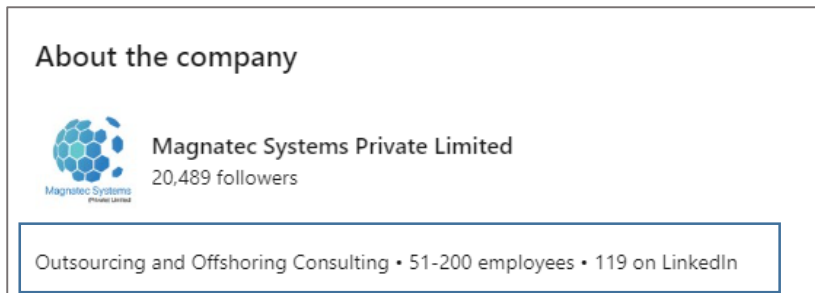


Figure 3.9: About Company

3.7 Problems Faced in Scraping Jobs

- i. **Errors:** Encountering errors during scraping is common, such as HTTP errors, missing elements on web pages, or unexpected page layouts. Handling these errors gracefully is essential to ensure the scraping process continues smoothly without interruptions.
- ii. **Network Problems:** Network issues, including slow internet connections or intermittent connectivity, can disrupt the scraping process and lead to incomplete or inconsistent data collection. Implementing robust error-handling mechanisms can mitigate the impact of network problems.
- iii. **Taking Huge Time:** Scraping a large volume of job data can be time-consuming, especially when processing multiple pages of search results or scraping detailed job descriptions. Optimizing scraping scripts and utilizing parallel processing techniques can help reduce the time required for data collection.
- iv. **Careful Process:** Scraping jobs from LinkedIn requires careful attention to detail to avoid violating website terms of service or triggering anti-scraping measures. Adhering to ethical scraping practices and monitoring scraping activities closely is essential to prevent potential repercussions.
- v. **Management of Scraped Files:** Managing the files generated during scraping, including organizing, storing, and backing up the scraped data, can become challenging, especially when dealing with large datasets. Developing a systematic approach to file management can streamline the data collection process.

- vi. **Files Naming:** Naming scraped files in a consistent and meaningful manner is crucial for easy identification and retrieval of data. Establishing naming conventions that include relevant information such as date, source, and content can facilitate efficient file management.
- vii. **Scraping One-Day Jobs in Parts:** Scraping job postings that are available for only one day requires careful planning to ensure all relevant postings are captured within the limited timeframe. Implementing scheduling mechanisms to scrape periodically throughout the day and prioritizing real-time data collection can help address this challenge.

3.8 Data Organization

3.8.1 What is Data Organization and Management?

Data organization and management refer to the processes of structuring, storing, and handling data in a systematic and efficient manner. It involves organizing data in a way that facilitates easy access, retrieval, analysis, and utilization.

3.8.2 Why data Organization is Important?

Effective data organization and management are crucial for several reasons:

- i. **Efficiency:** Well-organized data enables quicker access and retrieval, reducing the time and effort required for data processing tasks.
- ii. **Accuracy:** Properly managed data is less prone to errors and inconsistencies, ensuring the reliability and integrity of information.
- iii. **Decision-Making:** Organized data provides a clear and comprehensive view of information, enabling informed decision-making and strategic planning.
- iv. **Compliance:** Proper data management practices ensure adherence to regulatory requirements and data protection laws, reducing the risk of legal and compliance issues.
- v. **Scalability:** A structured approach to data organization allows for scalability, accommodating the growth of data volumes and complexity over time.

3.9 Process of organizing LinkedIn Jobs dataset

In organizing the LinkedIn jobs dataset, I opted for a files-based system as it aligned well with my requirements. I established a folder structure where each folder represents a category or aspect of the dataset. Within the main directory, subfolders were created to categorize data based on attributes such as job title, company name, location, and posting date. Within each subfolder, individual files were organized to contain specific job listings, with each file named according to a standardized convention that includes relevant information such as job title or posting date. This hierarchical organization facilitated easy navigation and retrieval of data, allowing for efficient analysis and utilization of the LinkedIn jobs dataset.

3.9.1 Structure of Folders:

i. Daily Folders



Figure 3.10: Daily Folder Structure

ii. Weekly Folders:



Figure 3.11: Weekly Folder Structure

3.9.2 Concatenation vs. Merging of Files:

Concatenation: Concatenation is the process of stacking datasets with the same columns vertically, resulting in an increased number of rows. This is typically used when combining multiple datasets that contain the same types of data but from different sources or time periods.

Merging: Merging, on the other hand, involves combining two datasets with different features horizontally, resulting in an increased number of columns. This process is used to mix datasets based on a common key or index, effectively integrating different types of information into a single, cohesive dataset.

3.10 Concatenating Files into One File:

3.10.1 Concatenation of Files Concept:

Now, these distributed files need to be consolidated into a single file to begin the data analysis process.

3.10.2 Concatenating Different Files Separated into a Daily File

To concatenate different files separated into a daily file, each daily file containing data for a specific day is sequentially appended to create a consolidated file. This process involves opening each daily file, reading its contents, and appending them to a new file in the desired order. The resulting concatenated file contains all the data from the individual daily files, arranged chronologically.

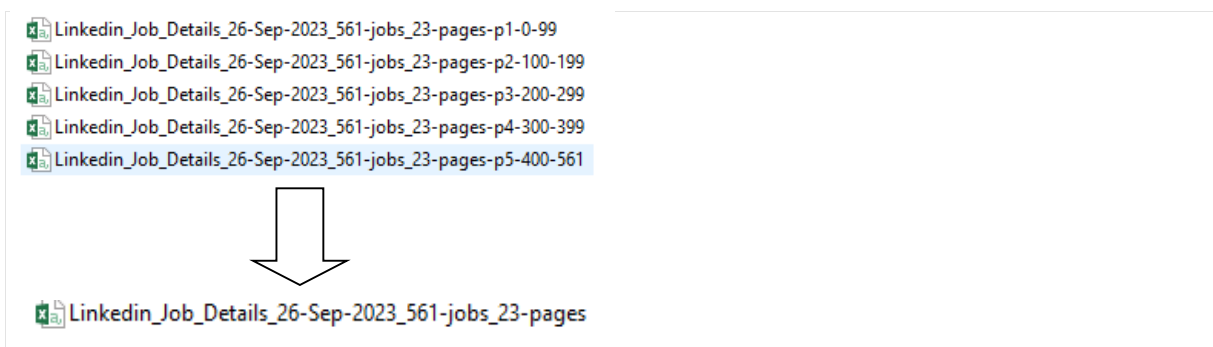


Figure 3.12: Concatenation into one daily File

3.10.2 Concatenating Daily Files into Weekly Files

Similarly, to concatenate daily files into weekly files, the daily files for each week are combined into a single weekly file. This involves iterating through the daily files for a given week, reading their contents, and appending them to create a new file representing the entire week. Each weekly file thus contains the aggregated data for a specific week, facilitating higher-level analysis and reporting.

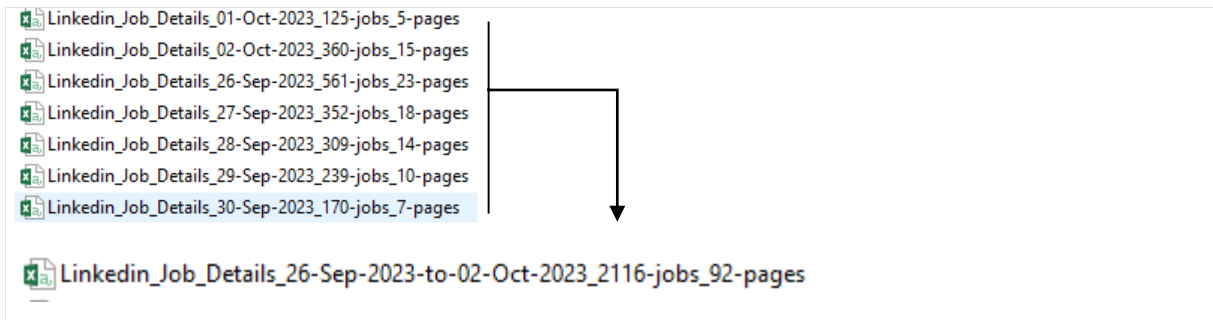


Figure 3.13: Concatenation of daily Files into weekly Files

3.10.3 Concatenating Weekly Files into One File

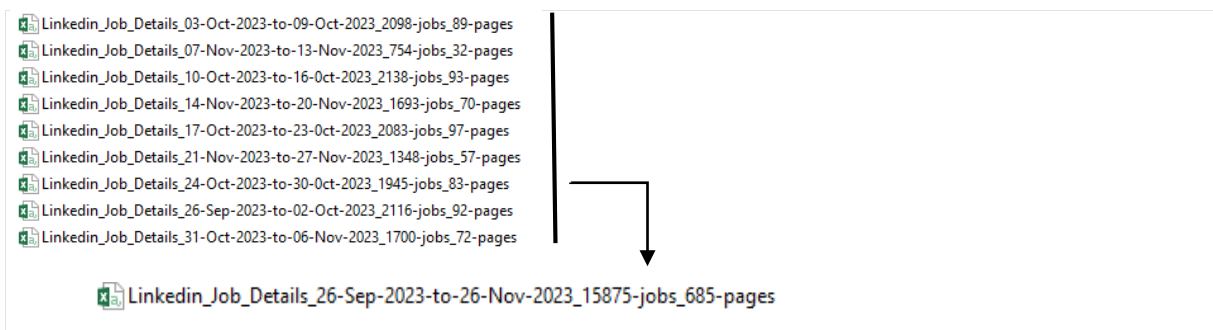
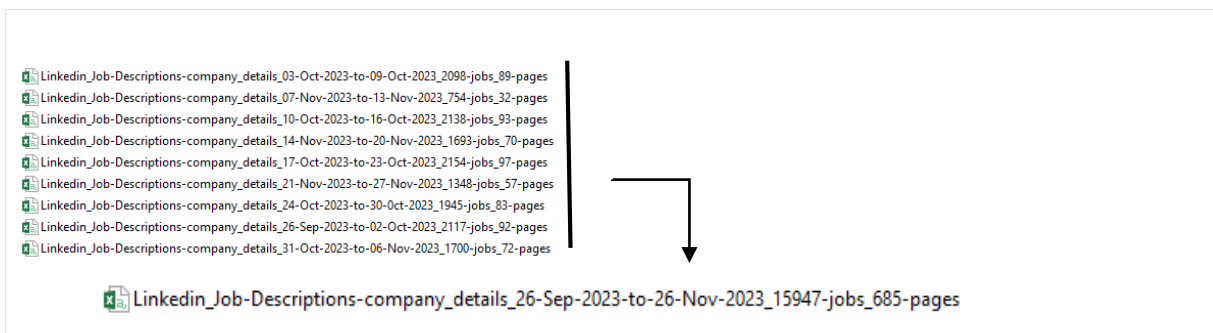


Figure 3.14: Concatenation of Weekly Files into 1 Excel File



Finally, to concatenate weekly files into one file, the weekly files are merged to form a comprehensive dataset spanning multiple weeks. This process entails opening each weekly file,

reading its contents, and appending them to create a single consolidated file containing data from all the weeks. The resulting concatenated file represents the entire dataset, providing a holistic view of the aggregated information over the entire period.

Table 3.2: Total count of Jobs scrapped

| | |
|--|-------|
| Total count of Jobs in Jobs details dataset from 26 September 2023 to 26 November 2023 | 15875 |
| Total count of Jobs in Jobs Descriptions dataset from 26 September 2023b to 26 November 2023 | 15947 |

CHAPTER NO 4
Data Preparation 1

4.1 Introduction of chapter:

This chapter covers the essential steps and techniques, focusing on cleaning, preprocessing, and feature engineering of jobs data to ensure it is accurate, complete, and suitable for analysis. We lay the foundation for robust and insightful data analysis through data cleaning, EDA, and feature engineering.

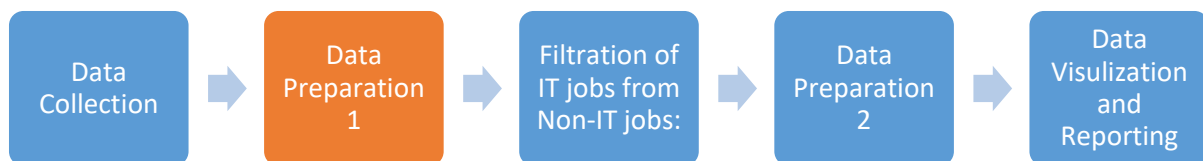


Figure 4.1: Data preparation in data analytics project life cycle

This is phase 1 of data preparation because it is focused on the data preparation before separating IT and non-IT jobs, so all processes or techniques applied in this chapter utilized All jobs data that contain IT and Also Non-IT jobs scrapped from LinkedIn.

4.2 Basic Concepts related to Data preparation:

4.2.1 Data preparation and Data Preprocessing:

- **Data Preparation** is the process of collecting, cleaning, and consolidating raw data into a form suitable for analysis. This step involves gathering data from various sources, handling missing values, correcting inconsistencies, and ensuring that the data format aligns with the analysis requirements. Proper data preparation ensures that the data is accurate, complete, and ready for subsequent data processing and analysis stages.
- **Data Preprocessing** is a crucial step in the data analysis workflow, involving transforming raw data into a clean and usable format. This includes normalization, scaling, encoding categorical variables, handling outliers, and partitioning data into

training and testing sets. Effective data preprocessing improves the data quality, enhances the performance of machine learning models, and ensures that the results are reliable and meaningful.

4.2.2 EDA and Importance of EDA

Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often using visual methods. EDA involves examining the data's distribution, identifying patterns, spotting anomalies, testing hypotheses, and checking assumptions. This preliminary analysis helps understand the data's underlying structure, informs subsequent modeling choices, and guides further data preprocessing steps to ensure robust analytical outcomes.

Why EDA?

- i. **Understanding Data Distribution:** EDA helps business analysts understand the underlying distribution of data, revealing patterns, trends, and relationships that can inform strategic decisions.
- ii. **Data Quality Assessment:** By assessing data quality, EDA identifies missing values, inconsistencies, and errors, ensuring that the data used in business intelligence systems is reliable and accurate.
- iii. **Guiding Feature Engineering:** Insights from EDA inform the feature engineering process by highlighting important variables and their interactions, which can be crucial for building predictive models that support business decision-making.

4.2.3 Data Cleaning:

Data Cleaning is the process of identifying and correcting errors, inconsistencies, and inaccuracies in a dataset to enhance its quality and reliability. This involves tasks such as removing duplicate records, filling in or imputing missing values, correcting data entry errors, and eliminating outliers. Data cleaning ensures that the dataset is accurate, complete, and suitable for analysis, which is crucial for deriving valid and reliable insights from the data.

4.2.4 Feature Engineering and its Importance

Feature Engineering is the process of transforming raw data into meaningful and useful metrics that can provide deeper insights and support decision-making. This involves creating new variables or modifying existing ones to highlight important trends, relationships, and patterns in the data. Techniques include aggregating data, creating ratios, extracting date and time components, and generating categorical features from continuous variables. The goal is to enhance the dataset's analytical value, making it more relevant and actionable for business intelligence and reporting.

Why Feature Engineering?

- i. **Improving Model Performance:** Feature engineering enhances the performance of predictive models, which can lead to more accurate forecasts and better decision support systems.
- ii. **Capturing Domain Knowledge:** It allows the incorporation of domain-specific knowledge into the analysis by creating features that reflect important aspects of the business context.
- iii. **Reducing Complexity:** By selecting and constructing the most relevant features, feature engineering simplifies models, making them more robust and easier to interpret for business stakeholders.
- iv. **Enhancing Interpretability:** Well-engineered features improve the interpretability of models, helping business users understand how different factors contribute to outcomes, which is crucial for gaining trust and actionable insights.
- v. **Handling Various Data Types:** Feature engineering techniques can handle diverse data types (numerical, categorical, time-series), ensuring that the data is in a suitable format for analysis and business intelligence reporting.

4.3 The process of Data Cleaning

After the collection of raw data, there are often missing values and duplicate records that need to be addressed to ensure data quality. This phase of cleaning data is divided into two essential steps:

As there are two datasets

- i. Jobs details Dataset
- ii. Jobs descriptions Dataset

Before merging job details and Job descriptions, Data Cleaning is applied to the job details dataset.

- i. Null or missing values due to
- ii. scrapping errors
- iii. Jobs links scrapped but jobs are not found, may be due to error or closing of job

4.3.1 Identifying Null or missing Values in the Jobs details dataset:

The total number of missing values for each column in the Data Frame, reveals that the 'Job Title', 'Job Info', 'Job Type', and 'employees' columns have 258, 258, 272, and 273 missing values respectively, while the 'Job Link' and 'Job ID' columns have no missing values. These null or missing values in the four features of the Jobs details dataset indicate that the highest null value is 273, so the total number of null rows is 273.

| | |
|-----------|-----|
| Job Title | 258 |
| Job Info | 258 |
| Job Type | 272 |
| Job Link | 0 |
| employees | 273 |
| Job ID | 0 |

Figure 4.2: Null values in Jobs details features

4.3.2 Identifying Duplicates:

As Job ID is unique for each job, it is used as the Primary key based on which 3865 duplicates were identified. The total number of duplicate rows is based on the 'Job ID' column, revealing that there are 3,865 duplicates in the dataset.

```
duplicates_count
```

```
3865
```

4.3.3 Removing Duplicates and Missing Values:

There are 15875 rows of jobs scrapped, In the first step, 3835 duplicate rows are removed then 273 rows having null values are removed a total of 4138 jobs or rows are removed, and finally, we have 11737 jobs.

Table 4.1: stats of data cleaning before merging two data sets

| | |
|---|-------|
| Total Number of Jobs scrapped | 15875 |
| Number of Duplicates Jobs | 3865 |
| Number of Rows containing Null values | 273 |
| Number of jobs preserved after removing missing values and Duplicates | 11737 |

4.3.4 Removing duplicates and null values after merging of jobs details dataset and Jobs description dataset:

Table 4.2: Stats of Data Cleaning after merging two data sets

| | |
|---|-------|
| Total Number of Jobs or rows after merging two datasets: LinkedIn Job details and LinkedIn Job descriptions | 27644 |
| Number of Duplicates Jobs | 15634 |
| Number of Null rows removed | 273 |
| Number of jobs preserved after removing missing values and Duplicates | 11737 |

| | | | |
|----------------|------|----------------|------|
| Job Title | 258 | Job Title | 0 |
| Job Info | 258 | Job Info | 0 |
| Job Type | 272 | Job Type | 0 |
| Job Link_x | 0 | Job Link_x | 0 |
| employees | 273 | employees | 0 |
| Job ID | 0 | Job ID | 0 |
| Job Link_y | 0 | Job Link_y | 0 |
| Job Desception | 284 | Job Desception | 31 |
| Company Info | 3185 | Company Info | 2916 |

Figure 4.3: Null values before and after removing duplicates and 273 Null rows

Why there are still null values in Job descriptions and Company Info:

There is a total of 273 rows or jobs removed based on 4 Jobs details Features, there are still 31 missing values in Jobs descriptions and 2916 in the Company Info feature, so here in the Jobs description feature there are still issues of uncompleted Jobs description.

4.3.5 Abnormal Descriptions:

A description that has less than 20 characters is considered an Abnormal or Uncompleted description, so there are 1586 abnormal descriptions.

As there are 31 null values in the job description, so total jobs that have abnormal descriptions is 1617

There are a total of 1617 rows that cannot be removed due to inconsistency in the job description because due to these features of the job details dataset such as Job title, Job Info, Job Type, and Employee are affected as these features are important.

4.4 EDA Phase 1:

4.4.1 How EDA is performed?

In this process, I explored every feature one by one.

- i. Exploring Unique Values and number of Unique Values
- ii. Exploring Value counts or Top 20 /Top 30 value counts
- iii. Exploring inconsistent values

- iv. Exploring how different features can be extracted from parent features

4.4.2 Summary of exploring each Feature one by one in table

So I explored 4 categorical features, Details are as follows:

Table 4.3: Summarized EDA of 4 Primary Features

| | Features Names | Data Type | Type of feature | A short description of the feature | Usage | Unique values |
|---|-----------------------|------------------|------------------------|---|--|----------------------|
| 1 | Job Title | object | Single-valued | A short introduction to the job | Skills analysis | 6442 |
| 2 | Job Info | object | Multi-valued | contains the company name, location, remote/onsite status, time passed to job posting, and number of applicants. | Company, Geography, and Competition Analysis | 10209 |
| 3 | Job Type | object | Multi-valued | Contains information about the nature of jobs such as Type of Employment, Experience Level, and remote /onsite status | Job nature analysis | 129 |
| 4 | employees | object | Multi-valued | Contains extra information about the company such as the Company's Employee Size and Company's Industry | Industry analysis | 975 |

4.2.3 Exploring Job Title:

Job title is a single-valued feature that can be used for skills analysis, there are a total of 6442 unique Job titles out of 11737 jobs, which contain a diverse range of titles from IT and non-IT Jobs.

Here below **first 10 unique values:**

```
array(['LinkedIn Data Extractor', 'It Management Intern',
      'SQL Database Developer', 'Business Intelligence Developer',
      'Web Developer - Internship',
      'Research Analyst for Business Subjects (Office Based + Night Shift)',
      'Performance Marketing Intern', 'Sr. SDET Engineer',
      'Associate Engineer', 'Fresh Graduate Engineers',
```

Figure 4.4: Unique values in Job Title

Top 15 Job Titles:

| | |
|---------------------------------|-----|
| Job Title | |
| Sales Executive | 142 |
| Business Development Executive | 135 |
| Sales Specialist | 76 |
| Business Development Manager | 71 |
| Business Development Specialist | 63 |
| Graphic Designer | 58 |
| Web Back-End Developer | 57 |
| Software Engineer | 56 |
| Customer Service Representative | 55 |
| Salesperson | 50 |
| Full Stack Engineer | 47 |
| Senior Software Engineer | 45 |
| Sales And Marketing Specialist | 40 |
| Wordpress Developer | 38 |
| Call Center Representative | 37 |
| PHP Developer | 36 |

Figure 4.5: Top 15 Job Titles

As there is a wide range of job titles that cannot be efficient for in-depth Skills analysis, so that is why jobs need to be filtered based on job title and job description, and then new features such as Field, Sub-Field, and Tools/Platforms were formed.

4.2.4 Exploring Job Info Feature:

It is multi-valued feature that contains the company name, location, remote/onsite status, time passed to job posting, and number of applicants. It has 10902 unique values. This feature is used in Company analysis, Geography analysis, and Competition analysis, so it is very important feature because Geography analysis, Company analysis and Competition analysis are the main features of my data analytics project.

First 10 Unique values in the Job Info Feature:

```
array(['EcomFleet · Pakistan (Remote) \n1 week ago',
      'Digicon Valley · Lahore, Punjab, Pakistan (On-site) \n2 weeks ago',
      'Murkez Technologies · Lahore, Punjab, Pakistan (On-site) \n1 week ago',
      'DevNatives · Islamabad, Islāmābād, Pakistan (Hybrid) \n1 week ago\n\n · 51 applicants',
      'ItCreators · Lahore, Punjab, Pakistan (On-site) \n1 week ago\n\n · 51 applicants',
      'TechoMatrix · Karāchi, Sindh, Pakistan (On-site) \n2 weeks ago',
      '8th Loop · Karāchi, Sindh, Pakistan (On-site) \n1 week ago',
      'Afiniti · Peshawar, Khyber Pakhtunkhwa, Pakistan (On-site) Reposted \n1 week ago',
      'Zones IT Solutions · Islamabad, Islāmābād, Pakistan (On-site) \n2 weeks ago\n\n · 312 applicants',
      'HUAWEI Pakistan · Islāmābād, Pakistan (On-site) \n2 weeks ago\n\n · 2,048 applicants'],
      dtype=object)
```

Figure 4.6: Unique values in Job Info feature

4.2.5 Exploring Job Type Feature:

It is a multi-valued feature that contains information about the nature of jobs such as Type of Employment, Experience Level, and remote /onsite status it can be used for Job nature analysis, it contains 129 unique values.

First 30 and last 10 Unique values in the Job Type Feature:

```
array(['Full-time', 'Full-time · Mid-Senior level',
      'Internship · Internship', 'Full-time · Associate', 'Internship',
      'Full-time · Entry level', 'Part-time', 'Full-time · Internship',
      'Contract · Mid-Senior level', 'Contract', 'Contract · Associate',
      '$150/month - $300/month · Full-time', 'Full-time · Director',
      'Part-time · Associate', 'Mr Ahsan is hiring for this job',
      'Part-time · Executive', 'Contract · Executive',
      'Part-time · Mid-Senior level', 'Contract · Entry level',
      'Temporary · Mid-Senior level', 'Executive',
      'Full-time · Executive', 'Sean Sato is hiring for this job',
      'Part-time · Internship', 'Internship · Mid-Senior level',
      '$12,000/yr - $15,000/yr · Full-time · Mid-Senior level',
      '$200,000/yr · Full-time', 'Part-time · Entry level',
      'dr.ry. L. is hiring for this job',
      '10,001+ employees · Tobacco Manufacturing'], dtype=object)

array(['$50,000/month - $70,000/month + Commission\n\n \nOn-site\n\n \nFull-time',
      'On-site\n\n \nInternship\n\n \nDirector',
      'Hybrid\n\n \nFull-time\n\n \nExecutive',
      '₹180,000/yr - ₹360,000/yr\n\n \nRemote\n\n \nFull-time\n\n \nEntry level',
      'On-site\n\n \nTemporary\n\n \nInternship',
      '$12,000/yr - $18,000/yr\n\n \nRemote\n\n \nFull-time\n\n \nMid-Senior level',
      'Contract\n\n \nEntry level',
      'Hybrid\n\n \nFull-time\n\n \nInternship',
      'On-site\n\n \nTemporary\n\n \nMid-Senior level',
      'Hybrid\n\n \nInternship\n\n \nMid-Senior level'], dtype=object)
```

Figure 4.7: Unique values in Job Type Feature

As there are some inconsistent values (highlighted) in the Job Type feature.

4.2.6 Exploring Employee Feature:

This is also a multi-valued feature that Contains extra information about the company such as Company’s Employee Size and Company’s Industry, can be used for industry analysis, and contains 975 unique values.

First 10 Unique Values in Employee Feature:

```
array(['11-50 employees', '51-200 employees',
      'Mobeen Naqvi is hiring for this job',
      '1,001-5,000 employees · Software Development',
      '501-1,000 employees · IT Services and IT Consulting',
      '501-1,000 employees · Telecommunications',
      '1,001-5,000 employees · Banking', '201-500 employees',
      '10,001+ employees', '51-200 employees · Human Resources Services'],
      dtype=object)
```

Figure 4.8: Unique values in Employee Feature

4.3 The Process of Feature Engineering:

4.3.1 The Summarized Overview of Feature Engineering Phase 1

In this process of Feature Engineering, many new features are created from Three Primary Features.

Table 4.4: Secondary Features Derived by Primary Features

| Parent Primary Features | Derived/ Secondary Features or Features that are derived from Primary Features |
|-------------------------|--|
| Job Info | Company, Location, Number of applicants, Job posting Time and Onsite/Remote |
| Job Type | Type of Employment, Experience Level, and Onsite/ Remote |
| Employee | Company Employee Size and Company Industry |

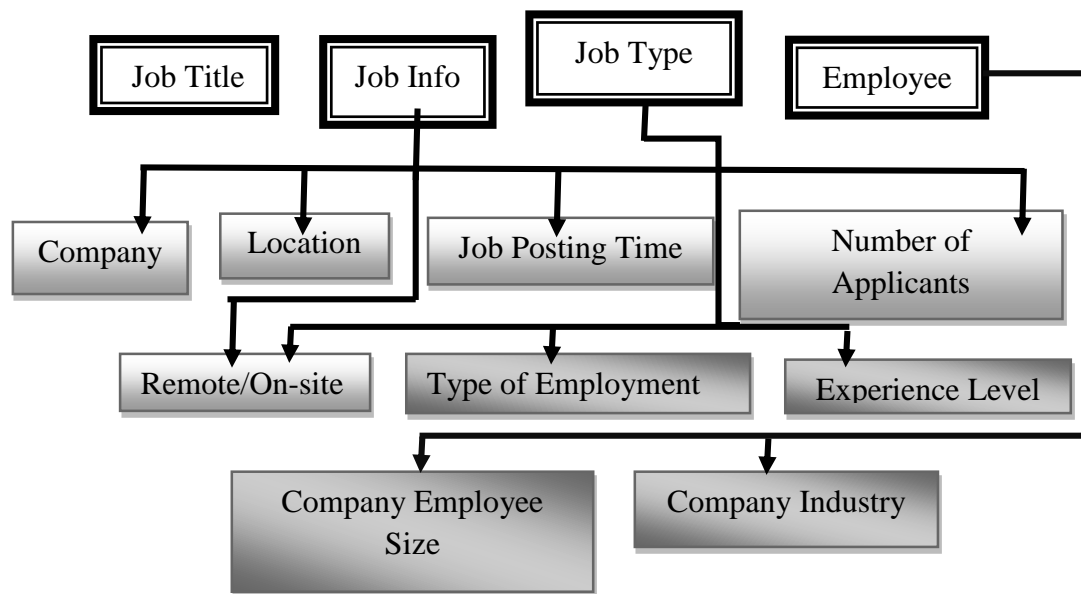


Figure 4.9: Tree of Secondary Features to be formed by Primary Features

4.3.2 Extracting Company Feature:

The **company** feature is derived from the **Job Info** feature, where it represents the first element of the **Job Info** string, separated by a dot (.).

```
'DevNatives · Islamabad, Islāmābād, Pakistan (Hybrid) \n1 week ago\n\n · 51 applicants'
```



Company Name

Figure 4.10: Extraction of Company in Job Info String

4.3.3 Extracting Location Feature:

The **location** feature is derived from the **Job Info** feature, where it represents the second element of the **Job Info** string. This element is separated by a dot (·) and open parenthesis “(“ or “\n”.

It contains City Region and Country name.

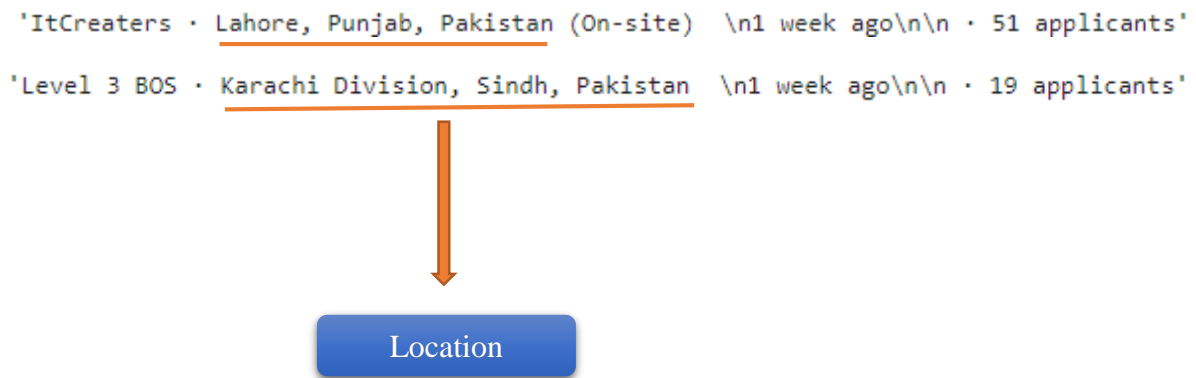


Figure 4.11: Extraction of Location in Job Info String

4.3.4 Extracting job posting time feature:

The job posting time feature is derived from the **Job Info** feature, which it represents the third or fourth element of the **Job Info** string. This element is separated by a backslash N “\n”.

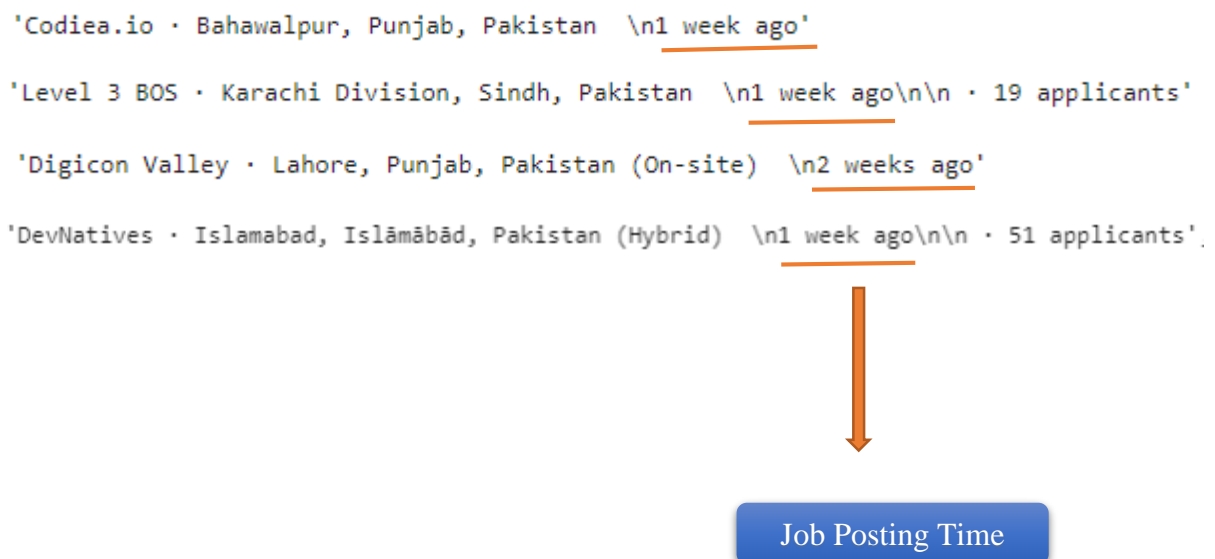


Figure 4.12: Extraction of Job Posting Time in Job Info

4.3.5 Extracting the number of applicants feature:

The Number of Applicants (n_applicant) feature is derived from the **Job Info** feature, which represents the last (4th or 5th) element of the **Job Info** string. This element is separated by a dot (.).

' applicants' is stripped from dot dot-separated element to store only numbers.

In case if number of applicants is missing then the Null value (np. nan) is stored instead of this.

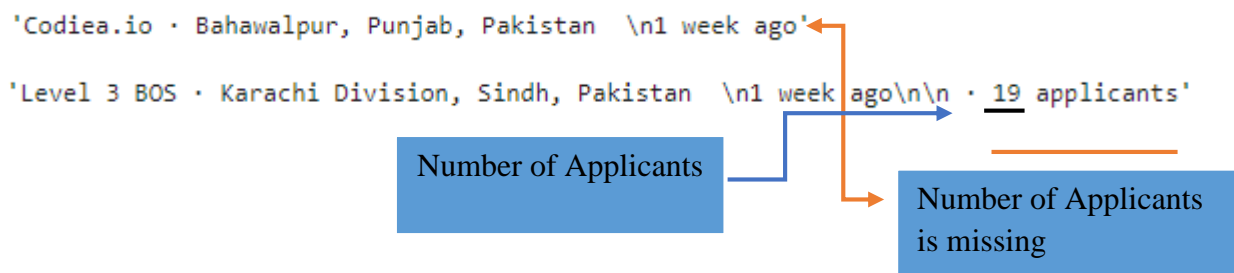


Figure 4.13: Extraction of Number of applicants from Job Info Feature

4.3.6 Extracting Type of Employment Feature:

The Type of Employment or Employment Type (created as Job-Type) feature is derived from the **Job Type** feature, which represents the sometimes 1st or sometimes 2nd position element of the **Job Type** string separated by a dot(.) or backslash N “\n”.

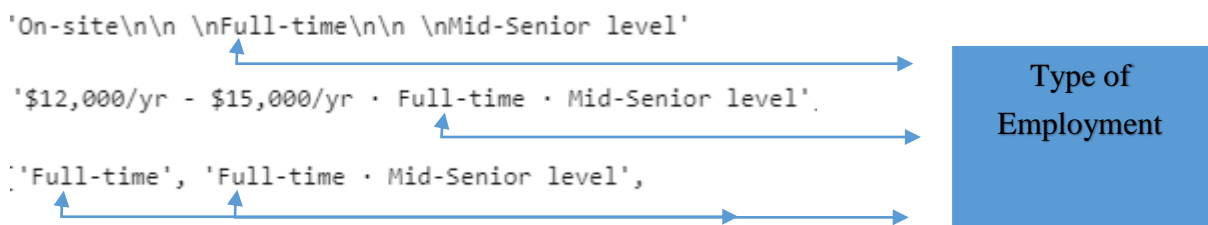


Figure 4.14: Extraction of Type of Employment from Job Type Feature

There are 6-7 categories(options) of Type of Employment on LinkedIn, in case of a null value (type of employment is not mentioned) is filled with the option of 'Not specified'.

Table 4.5: Types of Employment

| Type of Employment option | description |
|---------------------------|--|
| Full-time | Employment with a standard workweek (typically 35-40 hours) including benefits and job security. |
| Part-time | Employment with fewer hours than full-time, often offering flexibility but fewer benefits. |
| Contract | Employment for a fixed term or specific project, usually without traditional benefits. |

| | |
|---------------|---|
| Temporary | Short-term employment to meet immediate needs, often through staffing agencies, with limited job security and benefits. |
| Volunteer | Unpaid work is done to support a cause or organization, providing no monetary compensation. |
| Internship | Temporary work, often for students, to gain practical experience, sometimes unpaid or with a stipend. |
| Other | Not in 6 options |
| Not specified | If the type of employment is not mentioned then it is categorized as Not specified |

4.3.7 Extracting Experience Level Feature:

The Experience Level feature is derived from the **Job Type** feature, which represents the sometimes 1st sometimes 2nd, or sometimes 3rd position element but mostly the 2nd, or 3rd position element of the **Job Type** string is separated by a dot(.) or backslash N “\n”, In most Job Type strings experience level is not mentioned so it is labeled as “Not specified”.

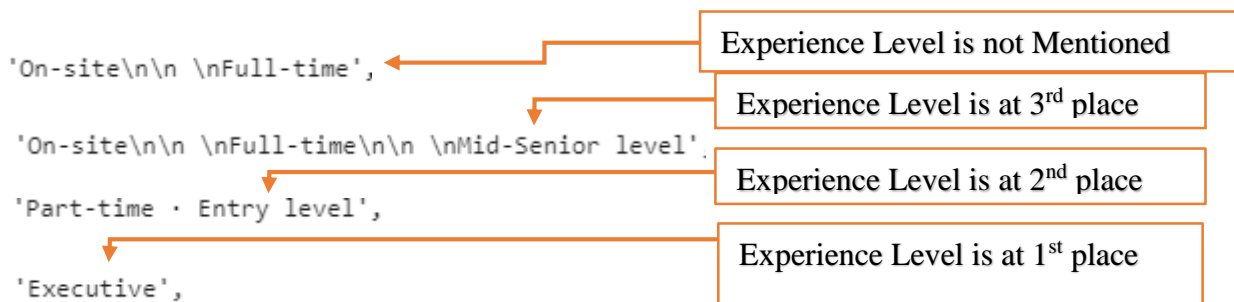


Figure 4.15: Extraction of Experience Level from Job Type

There are 6-7 categories(options) of Experience Level on LinkedIn, in case of a null value (Experience Level is not mentioned) is filled with the option of 'Not specified'.

Table 4.6: Experience Levels

| Experience Levels | Description |
|--------------------------|---|
| Internship | A position for gaining practical experience, typically for students or recent graduates, often with minimal prior experience. |
| Entry level | A starting position requires little to no professional experience, typically for recent graduates or those new to a field. |
| Associate | A role for employees with some experience, usually requiring basic professional skills and knowledge in a specific area. |
| Mid-Senior level | A position for individuals with significant experience and expertise, often involving leadership and more complex responsibilities. |
| Director | A senior management role responsible for overseeing departments or functions, requiring substantial experience and strategic planning skills. |
| Executive | A top-level management position with overall responsibility for company strategy and operations, typically requiring extensive experience and leadership abilities. |
| Not specified | If the Experience Level is not mentioned then it is categorized as Not specified. |

4.3.8 Extracting Remote/On-Site (Work Arrangement) Feature:

The Remote/On-Site (created as Remote/Onsite) feature is derived from the **Job Info and Job Type** features,

It is placed at the 3rd position element in the **Job Info** strings enclosed by small parenthesis “()”. In the **Job Type** string, it is placed at 1st position separated by two backslash N “\n\n”.

It is observed that the pattern of Onsite/Remote existence in two features is changed on the 9th October file at the 119th row.

So by investigating the data frame, it is observed that the 3438th row which is the 3997th index is a row of change in the File ‘Linkedin_Job_Details_26-Sep-2023-to-26-Nov-2023_11737-jobs_17-features_685-pages_Feature-Eng-processed_v4.csv’ because below this index Remote/Onsite is located in **Job Info** feature and above this index It is located in **Job Type** Feature.

Where Remote/onsite is not mentioned it is labeled as ‘Not specified’

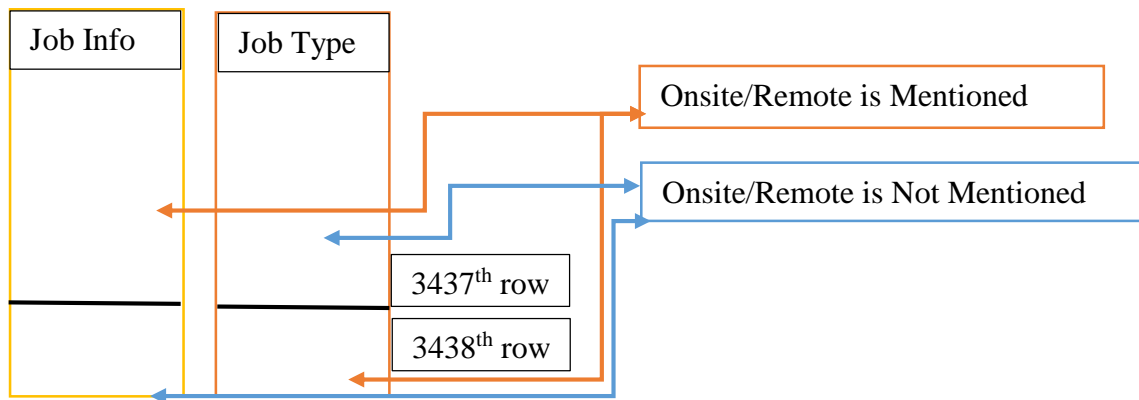


Figure 4.16: How Remote onsite is present in two features and Row of change

In strings of Job Info the feature of Remote/Onsite look like this:

```
'DevNatives · Islamabad, Islāmābād, Pakistan (Hybrid) \n1 week ago\n\n · 51 applicants',
```

↑ Onsite/Remote is at 3rd position in Job Info

In strings of Job Type the feature of Remote/Onsite look like this:

```
'Remote\n\n \nInternship\n\n \nMid-Senior level'
```

↑ Onsite/Remote is at 1st position in Job Type

Figure 4.17: Extraction of Remote/Onsite Feature from two features Job Info and Job Type

There are 3-4 categories(options) of **Remote/On-Site** on LinkedIn, in case of a null value (Remote/On-Site is not mentioned) is replaced with the option of 'Not specified'.

Table 4.7: Four categories of Remote/Onsite

| Remote/On-Site Options | description |
|------------------------|---|
| On-site | Work performed at a designated physical location, such as an office or facility |
| Remote | Work completed from any location outside the traditional office, typically from home or a co-working space. |
| Hybrid | A work arrangement combining both onsite and remote work, allowing flexibility in the work location. |
| Not specified | If the work arrangement is not mentioned, it is categorized as Not specified. |

4.3.9 Extracting Company Employee Size and Company Industry Features:

The **Company Employee Size** feature (created as `Comp_N_Employees`) and **Company Industry** feature (created as `Comp_Industry`) are derived from the **Employees** feature, both features give us some advanced analytics about jobs such as the Number of Employees in the company (Company Employee size) posting job and industry of the company.

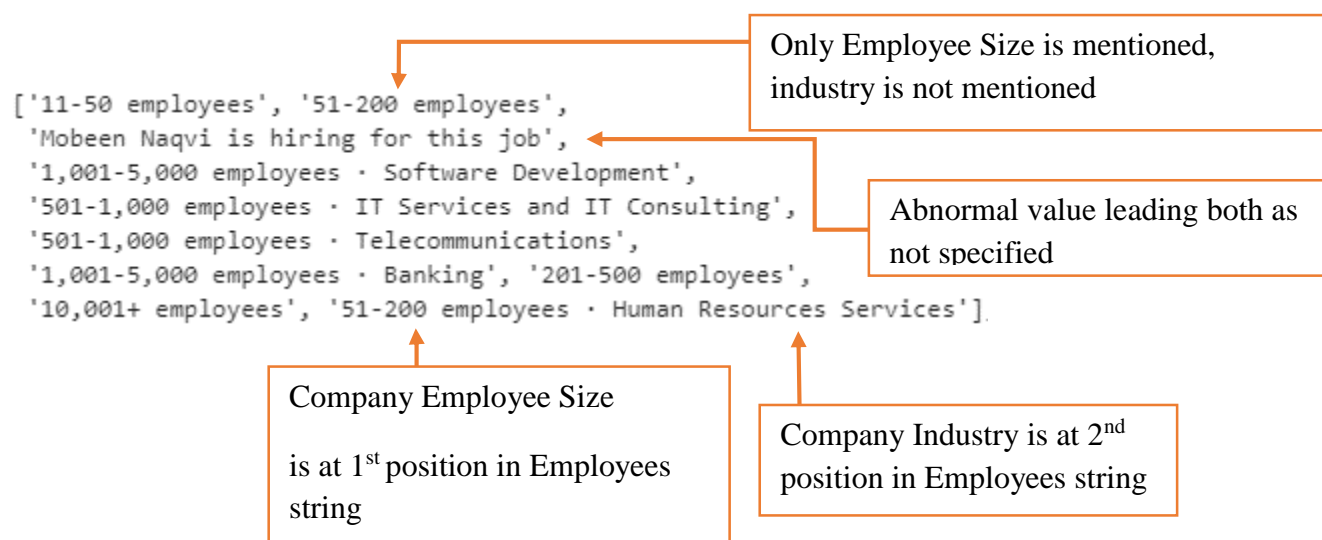


Figure 4.18: Extraction of Company Employee Size and Company Industry Feature from Employee Feature

Company Employee Size is placed at the 1st position element in the **Employees** strings, separated by a dot”.”.

On the other side Company Industry is mostly not mentioned, placed at the 2nd position element in the **Employees** strings, also separated by a dot”.”.

In case if Company Employee Size or Company Industry is not mentioned then it is labeled as ‘Not specified’.

4.4 Feature Engineering phase 2:

In 1st phase of Feature Engineering some secondary Features are extracted from primary Features such as Job Info, Job Type and Employees but in 2nd phase of Feature Engineering I am working on exploring Location feature (which is extracted from Job Info in 1st phase of Feature engineering) , I will explore how different meaningful features such as City, Province/region and country can be formed , these features which I am going to extract in Feature Engineering phase 2 are highly contributing in Geography analysis which is very important aspect of Analysis of IT jobs in Pakistan.

4.4.1 Exploring Location Feature:

Location is secondary feature extracted from Job Info, it is also multi valued feature contains information such as City, Province or Region and Country.

- **Unique values in location feature:**

There are 1163 unique values in location feature, abnormalities are also existing here.

Somewhere location is single value like 1st string, mostly it has 3 values like 2nd string.

```
[ ' Pakistan ', ' Lahore, Punjab, Pakistan ',
  ' Islamabad, Islāmābād, Pakistan ', ...,
  ' Sakrand, Sindh, Pakistan \n1 week ago\n\n ',
  ' Sheikhpura District, Punjab, Pakistan \n1 week ago\n\n ',
  ' Islamabad, Punjab, Pakistan \n2 weeks ago\n\n '], dtype=object)
```

Location with only
Country

4th value is garbage which
needs to be removed

- **Top 6 locations:**

| | |
|--|------|
| Lahore, Punjab, Pakistan | 1052 |
| Karāchi, Sindh, Pakistan | 624 |
| Islamabad, Islāmābād, Pakistan | 518 |
| Lahore, Punjab, Pakistan \n2 weeks ago | 326 |
| Karachi Division, Sindh, Pakistan | 242 |
| Pakistan | 236 |

Cities like Karachi and Islamabad are in Multiple Styles needs to be fixed in only one style

Figure 4.19: Components of Location to be extracted and abnormalities in City Problem

4.4.2 Extracting City, Province/Region, and Country Feature:

In this process, three main Geography Features City, Province/Region, and Country are extracted from the Location feature, the location is a secondary feature that is extracted from Job Info so that is why these three features are called ternary features.

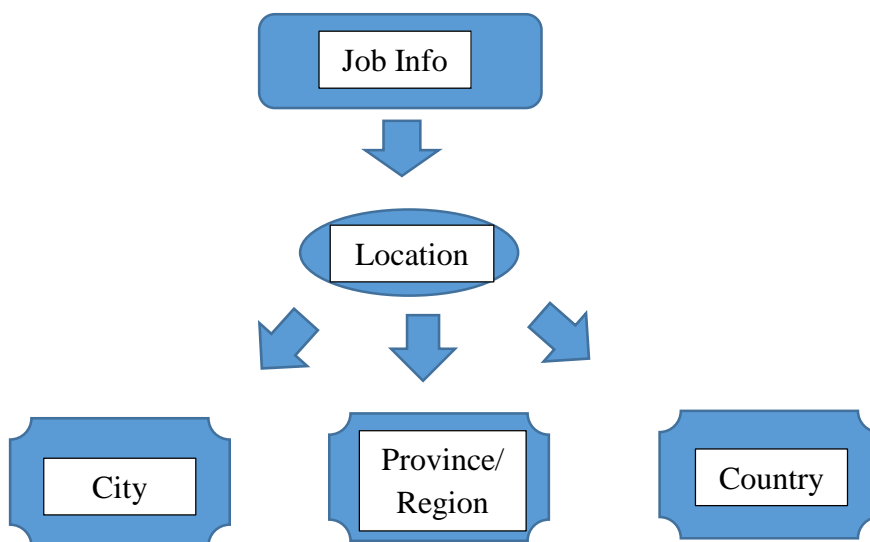
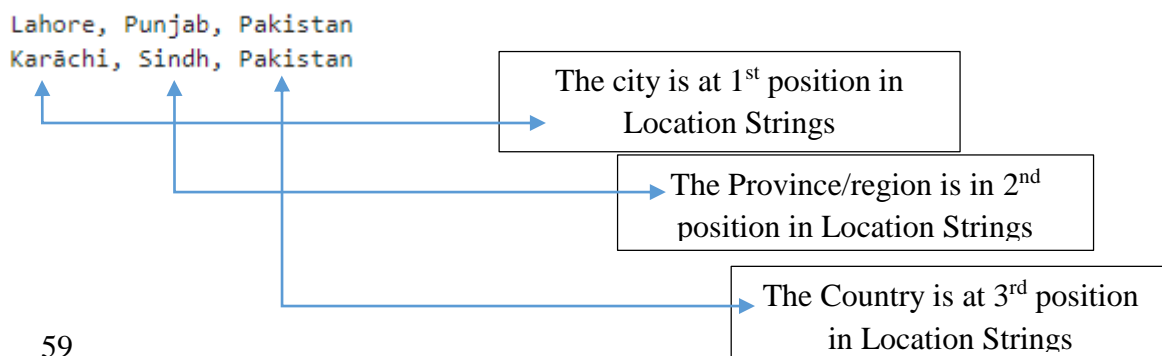


Figure 4.20: Tree of Extracting Location Features

City is located at 1st, Province or Region is placed at 2nd and country is placed at 3rd Position in Location feature and all are separated by comma “,” as shown below in output:



4.4.3 Handling Multiple Styles or names of Cities to clarify a strong City analysis:

Some cities like Islamabad, and Karachi have different names or styles so these are replaced with single names or styles as shown below in the table:

Table 4.8: Same Cities with different Styled names

| Name of Cities Before Replacement | Name of Cities after Replacement | Name of Cities Before Replacement | Name of Cities after Replacement |
|--|---|--|---|
| Islāmābād | Islamabad | Korangi | Karachi |
| Karāchi | Karachi | Johar Town | Karachi |
| Abbottābād | Abbottabad | Malir | Karachi |
| New Karachi Town | Karachi | Gulshan Town | Karachi |

As some words create the complexity of Repeatability are removed in location Features which are:

Table 4.9: Repeated words in location to be removed

| | | |
|-------------|-------------|-----------|
| ' District' | ' Division' | ' Tehsil' |
|-------------|-------------|-----------|

CHAPTER NO 5

Filtration of IT jobs from Non-IT jobs

5.1 Introduction of chapter:

This chapter focuses on the filtration or segmentation of IT and non-IT jobs, a critical step in my project on the analysis of IT jobs. Accurate separation of IT from non-IT roles is essential for a comprehensive understanding of the IT job market. By employing keyword-based text classification and filtering techniques, we can effectively categorize job listings, ensuring that IT roles are accurately identified and analyzed.

5.2 Basic Terminologies:

5.2.1 Categorizing and Filtering IT and Non-IT Jobs:

Categorizing and filtering IT and non-IT jobs involve the systematic organization of job listings based on their nature and requirements. IT jobs are typically those that involve technology, software development, data analysis, and other computer-related tasks. Non-IT jobs, on the other hand, encompass a wide range of occupations that do not primarily focus on technology.

5.2.2 Keyword-Based Text Classification and Segmentation of IT and Non-IT Jobs:

This approach uses specific keywords to classify and segment job postings into IT and non-IT categories. By analyzing the presence and frequency of certain keywords within job descriptions, user-defined functions can determine the likely category of a job listing. This method relies on natural language processing (NLP) techniques to scan and interpret the text, enabling accurate classification based on the content of the job postings.

5.2.3 Keyword Filtering:

Keyword filtering is the process of identifying and isolating relevant job listings by using predefined keywords. This method involves scanning job titles or descriptions for specific terms that are indicative of IT or non-IT roles. Keywords can include job roles, required skills, tools, programming languages, and other relevant terminology that help in distinguishing between IT and non-IT jobs.

5.2.4 Keyword Matching:

Keyword matching involves comparing the keywords found in job descriptions or job titles with a predefined list of keywords associated with IT or non-IT jobs. This process helps in

determining the category of a job by checking the presence of specific keywords. If a Job title or job description contains IT-related keywords, it is classified as an IT job, and vice versa.

5.2.5 Text Classification:

Text classification is a broader technique that involves categorizing text into predefined categories based on its content. In the context of job segmentation, text classification algorithms are used to analyze job descriptions and classify them as IT or non-IT based on the overall context and presence of relevant keywords. Machine learning models can be trained to improve the accuracy of this classification process.

5.2.6 Text Analysis:

Text analysis, or text mining, is a comprehensive process that involves extracting meaningful information from text data using various techniques, including natural language processing (NLP). In job classification, text analysis goes beyond simple keyword matching by understanding the context, semantics, and structure of the text. It can identify patterns, categorize content, and derive insights from job descriptions or job titles to improve classification or filtering accuracy.

5.2.7 Why Categorization is Important: IT and Non-IT Jobs

- i. **Diversified and complex dataset:** Identifying which jobs are IT and which are not is crucial for my project, which focuses on the analysis of IT jobs. With 6,442 job titles spread across 11,737 job listings, the diversity makes it challenging to conduct an efficient analysis. Effective categorization of IT jobs is necessary to streamline the process, ensuring accurate and meaningful insights into the IT job market.
- ii. **Identification of IT and Non-IT Jobs:** Identifying IT and non-IT jobs is important because there are no clear rules to separate them, causing a lot of confusion. The first step in categorizing these jobs is to set specific criteria for what makes a job an IT job. IT jobs usually involve tasks like computing, programming, software development, and managing information. However, tech jobs can be confusing because not all tech jobs are IT jobs. Mixing IT and non-IT roles can lead to poor analysis and misunderstandings about job trends. By clearly defining what qualifies as an IT job—such as the need for

certain degrees, technical skills, and specific fields like Software Engineering, AI, Cyber Security, and IT Support—I can accurately categorize IT jobs and ensure precise analysis.

5.3 Pre-Planning to Filter IT Jobs

5.3.1 Criteria for IT Jobs:

- i. **Not any Clear Definition of IT and Tech Jobs:** There is often no universally accepted definition that separates IT jobs from other technology-related roles, making categorization challenging.
- ii. **Use of Computer, Programming:** Jobs that primarily involve the use of computers, programming, software development, and other technical tasks are typically classified as IT jobs.
- iii. **CS, IT, AI, or any Computing Degree:** Positions requiring degrees in Computer Science (CS), Information Technology (IT), Artificial Intelligence (AI), or related fields are generally considered IT jobs.
- iv. **Major Fields of IT:** IT encompasses various specialized fields, including Software Engineering, Artificial Intelligence, Computer Networking, Cyber Security, Metaverse and Block chain, and IT Support.
- v. **Tech Fields vs IT Fields:** While all IT fields are tech fields, not all tech fields are IT fields. IT is specifically focused on computing and information technology, whereas tech fields can include broader areas like digital marketing and graphic design.

4.3.2 What about Digital Marketing, Copywriting, Graphic Design, and other Tech Jobs:

Roles like digital marketing, copywriting, and graphic design, although tech-related, are not traditionally considered IT jobs as they do not primarily involve computing and programming tasks.

4.3.3 Importance of Job Title and Job Description in Skills Analysis:

Job titles and descriptions provide critical information about the skills required for a position. They help in identifying the competencies needed and the nature of the job, making them essential for skills analysis and job matching.

4.3.4 Total Number of Job Titles:

The vast number of job titles, with 6,442 distinct titles, and their variations can make classification challenging. Understanding and standardizing these titles is necessary for accurate job categorization.

4.3.5 Why it is Difficult to Analyze Job Titles:

- **Job Titles with Different Names for Similar Roles:** Roles such as software engineer and software developer may have different titles but often entail similar responsibilities, complicating the analysis.
- **Extra Words in Titles:** Titles like "Senior Software Developer" include additional descriptors that can make it harder to standardize job titles.
- **Lengthy and Complex Titles:** Some job titles are excessively long or complex, making it difficult to categorize them accurately.
- **Irrelevant Titles:** Titles like "Software Seller" might include the word "software" but refer to a role that is not related to software development.
- **Generalized vs. Specified Titles:** Generalized titles like "Software Engineer" versus more specified ones like "Java Developer" highlight the need for nuanced classification.

4.3.6 Null Values and Incomplete Descriptions in Job Descriptions:

Incomplete job descriptions and null values can hinder the accurate classification of job postings. Missing information about required skills, responsibilities, or job context can lead to incorrect categorization.

4.3.7 Keywords Research for IT Jobs

Why a Huge List of Keywords: A comprehensive list of keywords is necessary to capture the diverse and evolving nature of IT jobs. It ensures that the classification system can accurately identify a wide range of roles and responsibilities within the IT sector.

Research on Keywords Based On:

- i. **Name of Skills:** Identifying keywords related to specific skills required for IT jobs, such as programming languages (e.g., Python, Java) and technical skills (e.g., system analysis, network configuration).
- ii. **Name of Tools:** Including keywords for tools commonly used in IT roles, like software (e.g., Git, Docker) and platforms (e.g., AWS, Azure).
- iii. **Name of Frameworks:** Keywords related to popular frameworks in IT, such as Angular, React, and Django, which are essential for certain development roles.
- iv. **Name of Computer and Tech-Related Keywords:** General tech-related keywords that help in identifying IT jobs, including terms like "software development," "network security," and "database management."

4.3.8 Keyword Analysis:

After conducting keyword research, the next step is keyword analysis. Keyword analysis is a type of text analysis that involves identifying and evaluating the occurrence of specific keywords within job descriptions and job titles to determine how many IT jobs and non-IT jobs are associated with each keyword. This process helps in understanding the relevance and frequency of particular terms, which in turn aids in accurately categorizing job postings.

During keyword analysis, job descriptions and job titles are scanned for predefined keywords that indicate IT or non-IT roles. The analysis records how many times each keyword appears and in which type of job listing. This information is used to refine the classification criteria, ensuring that the keywords effectively differentiate between IT and non-IT jobs. By analyzing keyword patterns, this process enhances the accuracy of job segmentation and helps maintain the integrity of job categorization efforts.

4.3.9 Why Keyword Analysis?

- i. **To Counter Any Incorrectness in Jobs Filtering:** Keyword analysis helps in refining the filtering process to minimize errors. It ensures that job listings are accurately segmented by continuously updating and validating the list of keywords against job descriptions.
- ii. **To Counter Mixing of Non-IT Jobs in IT Jobs:** Mixing non-IT jobs with IT jobs can lead to misleading analysis and inefficiencies. For example, a job title like "Marketing Specialist" might include terms related to digital tools but is not an IT role. Through careful keyword analysis, such misclassifications can be identified and corrected, ensuring that only relevant IT jobs are categorized as such.
- iii. **Keyword Matching is Case Sensitive:** Keyword matching can be case-sensitive, leading to potential issues. For instance, the keyword "IT" (referring to Information Technology) can be misclassified if matched with "it" (a common pronoun). Case sensitivity must be managed to avoid such errors, ensuring that keywords are recognized correctly regardless of capitalization.

4.4 Process of IT Jobs Filtering:

4.4.1 Phase 1: Selection of Proper Keywords Based on user Defined Testing Cases:

Phase 1 focuses on selecting precise keywords via user-defined testing cases, including filtering job listings based on criteria like the presence of key terms such as "IT." It involves rigorous checks for relevance and completeness in job descriptions and identifies conflicting job titles containing non-IT-related keywords. Two tests are conducted to prepare an accurate keyword list with droppable titles.

Test 1: Filtering Each Keyword Using User-Defined Functions

To filter IT jobs using the keyword "IT," I applied a two-step process to a dataset of job listings. First, I filtered the job titles for the presence of the keyword "IT," resulting in 99 job listings with 88 unique titles. I then checked the job descriptions for completeness and relevance, finding 12 abnormal descriptions (either incomplete or null). In the second step, I further

filtered these job listings by checking for the presence of IT-related keywords in their descriptions, which included terms like "software," "network," and "developer." This process identified 85 relevant job listings with 78 unique titles. The analysis showed that approximately 98% of the jobs were correctly identified as IT-related based on their descriptions. However, 10 job titles did not match, with 8 due to abnormal descriptions and 2 not matching the IT keywords in their descriptions. So this process repeated on Every keyword belongs to IT jobs.

Example: A job titled " IT & Network Administrator" might be included in `filtered_df_title` but if its description lacks detailed IT-related keywords, it might be displayed as the below output shows how 1st test works.

```
It is nearly 80.8080808080808 % varified by Jobs desc that It belongs to IT and Tech Jobs
```

```
list of 27 job titles which present in filtered_df_title but not present in filtered_df_desc:
['IT Sales Executive', 'IT & Network Administrator', 'Assistant Manager IT', 'Project Manager - IT Services', 'IT Project
Manager', 'Lead Generation Specialist - IT Sector', 'IT Technical Analyst (SAP) - HR', 'IT Service Desk Engineer (Night S
hift)', 'Executive IT/Cyber Security', 'Sr. Executive Admin & IT-Quetta-(H-225-02)', 'Senior Associate - IT Auditor (CIS
A)', 'International Sales Specialist & Lead Generation (IT Services)', 'IT On/Offboarding Coordinator', 'Business Develop
ment Manager (IT Sales-LinkedIn)', 'Senior Manager REIT (Real Estate investment Trust)', 'IT Business Development Executi
ve', 'Media & IT Assistant', 'IT Sales Representative', 'IT Support Assistant (Fresh Graduates)', 'Executive IT - OMC',
'Executive Secretary to CEO for HR, Admin, and IT related matters', 'Manager- IT Business Analyst-Logistics', 'IT Help De
sk Support Specialist', 'Deputy IT Field Infrastructure and Operations Officer', 'IT Business Partner', 'IT HelpDesk Repr
esentative', 'Digitization/IT Analyst']
```

```
List of 8 titles that are not cleared as IT and Tech Job due to abnoramal desc:
['IT Sales Executive', 'IT Project Manager', 'IT Technical Analyst (SAP) - HR', 'Senior Associate - IT Auditor (CISA)',
'International Sales Specialist & Lead Generation (IT Services)', 'IT On/Offboarding Coordinator', 'Business Development
Manager (IT Sales-LinkedIn)', 'Digitization/IT Analyst']
```

```
List of 19 titles that are not cleared as IT and Tech Job due to not matching of II_keywords in desc:
['IT & Network Administrator', 'Assistant Manager IT', 'Project Manager - IT Services', 'Lead Generation Specialist - IT
Sector', 'IT Service Desk Engineer (Night Shift)', 'Executive IT/Cyber Security', 'Sr. Executive Admin & IT-Quetta-(H-225
-02)', 'Senior Manager REIT (Real Estate investment Trust)', 'IT Business Development Executive', 'Media & IT Assistant',
'IT Sales Representative', 'IT Support Assistant (Fresh Graduates)', 'Executive IT - OMC', 'Executive Secretary to CEO fo
r HR, Admin, and IT related matters', 'Manager- IT Business Analyst-Logistics', 'IT Help Desk Support Specialist', 'Deput
y IT Field Infrastructure and Operations Officer', 'IT Business Partner', 'IT HelpDesk Representative']
```

Figure 5.1: Testing 'IT' keyword by user defined function test case 1

Test 2: Checking Each Job Title Against Non-Related Keywords

The process begins by identifying job titles that potentially contain non-IT related terms, using a predefined list of keywords like "Sales" or "Product." Employing list comprehension, each job title in the dataset is examined for the presence of any of these keywords, resulting in a list of potentially conflicting titles. Subsequently, among these titles, those that are still deemed to be IT-related despite containing non-IT-related keywords are removed from consideration. For

example, the title 'Manager- IT Business Analyst-Logistics' is recognized as an IT job and is thus excluded from the list of titles to be dropped. The remaining titles in the droppable list are then confirmed as non-IT jobs. After generating the droppable titles list, each keyword is applied to filter job listings, and the identified droppable titles are subsequently removed from consideration. This iterative approach helps refine the dataset, ensuring that only relevant IT job titles are retained for accurate analysis.

```
['SEO Executive (ON-SITE)',
 'IT Sales Executive',
 'Lead Generation Specialist - IT Sector',
 'Head of Sales Operations - IT Industry',
 'Senior SEO Executive - NIGHT SHIFT - ONSITE JOB ROLE',
 'International Sales Specialist & Lead Generation (IT Services)',
 'Business Development Manager (IT Sales-LinkedIn)',
 'IT Business Development Executive',
 'IT Sales Representative',
 'IT Technical Sales Executive (L2)',
 'Manager- IT Business Analyst-Logistics',
 'IT Business Partner',
 'Sales Development Representative-IT/Software',
 'IT Business Analyst']
```

Figure 5.2: Non related Job Titles under 'IT' Keyword

This is the list of Droppable Titles associated with the “IT keyword”, It is copied from confusing Titles but a title: 'Manager- IT Business Analyst-Logistics' is removed from droppable titles because by scanning confusing titles it is considered an IT job

```
['SEO Executive (ON-SITE)',
 'IT Sales Executive',
 'Lead Generation Specialist - IT Sector',
 'Head of Sales Operations - IT Industry',
 'Senior SEO Executive - NIGHT SHIFT - ONSITE JOB ROLE',
 'International Sales Specialist & Lead Generation (IT Services)',
 'Business Development Manager (IT Sales-LinkedIn)',
 'IT Business Development Executive',
 'IT Sales Representative',
 'IT Technical Sales Executive (L2)',
 'IT Business Partner',
 'Sales Development Representative-IT/Software',
 'IT Business Analyst']
```

Figure 5.3: List of Droppable Titles

Prepare a Finalized List of Keywords for Splitting:

Based on the above tests, a finalized list of keywords is created to accurately split and filter IT jobs. The list below provides a glimpse of the keywords I have finalized; the complete list contains nearly 100 keywords.

```
python, ios, android, desktop, Flutter, C++, UI/UX,html,css, scrapper, Game,unreal_engine dev,Unicity,.net, AI, desktop, testing, SQA, DevOps, Cloud, AWS, SQL, oracle, Data analyst, Business Analyst, Power BI,tabulea, database, analyst, ML, DL, CV, Network, Web3/Blockchain, Security, MS Office,
```

Figure 5.4: Preview of some Finalized List of keywords

4.4.2 Phase 2: Apply Keywords or Sets of Keywords to Filter Jobs

After finalizing the keywords related to IT jobs in the keyword analysis, the next step is to apply these keywords or sets of keywords associated with specific skills. This process is repeated for each keyword or set of keywords, and in each iteration, an Excel file is generated to organize the filtered job listings.

There are two primary methods for applying keywords to segment IT jobs from non-IT jobs:

- i. filtering based on job descriptions
- ii. filtering based on job titles.

Initially, I employed filtering based on job descriptions to identify IT-related roles. This approach involves scanning job descriptions for specific IT keywords or sets of keywords associated with relevant skills. However, as the project progressed, I found that filtering based on job titles offered a more efficient and accurate method for categorizing IT jobs.

4.4.3 Filtering IT Jobs Based on Job Titles:

- i. **Search Keywords in Job Titles:** Utilize finalized IT-related keywords to filter job titles for relevant skills and terms.
- ii. **Remove Irrelevant Titles:** Eliminate titles identified in the droppable titles list to retain only pertinent IT job titles.
- iii. **Save Filtered Listings:** Organize filtered job listings into Excel files within the designated Filtered_Files folder.

- iv. **Iterate the Process:** Repeat the filtering process for each keyword or set of keywords, ensuring comprehensive segmentation of IT jobs.

4.4.4 Example of Segmenting app development jobs:

Using the keyword "App," I found 166 job listings with 112 unique job titles containing that term. After removing 14 irrelevant titles identified in the droppable list during keyword analysis of 'app', the dataset was refined to 143 jobs with 98 unique job titles remaining. The filtered data frame has been saved as an Excel file named "k-App_26_sep_to_26_nov_143_jobs_98_titles_v5_Feature_Eng_Processed.csv" in the **Filtered_Files** folder.

```
titles_to_drop=['Telesales Representative - Appointment Setter',
'Business Application Manager',
'Appointment Setter',
'Administrative Assistant (Temporary Appointment), GS-5, Islamabad, Pakistan, # 26003',
'Creative Video Editor for Apps Promotions & Ad Campaigns.',
'Applied LLM Researcher',
'Phone Messaging Apps translation Urdu (Pakistan) Language Specialist',
'Senior Business Developer - Web/App Design and Development (Upwork)',
'Temporary Appointment: Programme Specialist (DRR), NOC, Islamabad # 126561',
'Appointment Setter/Fronter',
'Senior Apps Marketing Executive',
'RE-ADVERTISED: FAST-TRACK: (Temporary Appointment), Humanitarian Programme Specialist, Islamabad, Pakistan, P-3 (Closing
'Temporary Appointment: Social & Behavior Change Officer (Polio), NOB, Islamabad, Pakistan # 124305 (364 days)',
'RE-ADVERTISED: FAST-TRACK: (Temporary Appointment), Humanitarian Programme Specialist, Islamabad, Pakistan, P-3']
```

Figure 5.5: Droppable Titles while splitting 'App' Jobs

Below are the Jupiter Lab files used for filtering jobs and Excel files containing filtered IT jobs:

```
k-Android-ios-Mobile-Flutter-Xamarin-React Native-Swift-Kotlin-Ionic_26_sep_to_26_nov_229_jobs_102_titles_v5_Feature_Eng_Processed
k-App_26_sep_to_26_nov_143_jobs_98_titles_v5_Feature_Eng_Processed
k-Back_26_sep_to_26_nov_230_jobs_81_titles_v5_Feature_Eng_Processed
k-C++-C#-dotnet-.net_26_sep_to_26_nov_189_jobs_99_titles_v5_Feature_Eng_Processed
k-Cloud-AWS-Azure-GCP-IBM-Salesforce-Oracle_26_sep_to_26_nov_195_jobs_138_titles_v5_Feature_Eng_Processed
k-Desktop_26_sep_to_26_nov_10_jobs_6_titles_v5_Feature_Eng_Processed
k-Devops_26_sep_to_26_nov_80_jobs_31_titles_v5_Feature_Eng_Processed
k-Front_26_sep_to_26_nov_114_jobs_66_titles_v5_Feature_Eng_Processed
k-Full Stack_26_sep_to_26_nov_257_jobs_122_titles_v5_Feature_Eng_Processed
k-game-Unreal-Unity_26_sep_to_26_nov_103_jobs_41_titles_v5_Feature_Eng_Processed
k-IT_26_sep_to_26_nov_86_jobs_75_titles_v5_Feature_Eng_Processed
k-JavaScript-js-Node-React-Angular-Vue-Express-Java-jQuery-HTML-CSS-and-Java_26_sep_to_26_nov_292_jobs_203_titles_v5_Feature_Eng_Proce...
k-MERN_26_sep_to_26_nov_70_jobs_42_titles_v5_Feature_Eng_Processed
k-ML-DL-BI-AI_26_sep_to_26_nov_180_jobs_88_titles_v5_Feature_Eng_Processed
k-Network-Cisco-Internet_26_sep_to_26_nov_55_jobs_48_titles_v5_Feature_Eng_Processed
k-php_26_sep_to_26_nov_207_jobs_105_titles_v5_Feature_Eng_Processed
k-Python_26_sep_to_26_nov_77_jobs_41_titles_v5_Feature_Eng_Processed
k-Scrap_26_sep_to_26_nov_9_jobs_7_titles_v5_Feature_Eng_Processed
k-security-Cyber-Threat-Surveillance-infosec_26_sep_to_26_nov_19_jobs_18_titles_v5_Feature_Eng_Processed
k-Shopify_26_sep_to_26_nov_64_jobs_29_titles_v5_Feature_Eng_Processed
k-SQA-testing-Automation-Quality Assurance-Scrum-SDET_26_sep_to_26_nov_151_jobs_88_titles_v5_Feature_Eng_Processed
k-UI-UX_26_sep_to_26_nov_82_jobs_54_titles_v5_Feature_Eng_Processed
k-Web_26_sep_to_26_nov_229_jobs_84_titles_v5_Feature_Eng_Processed
k-Web3-Blockchain-crypto_26_sep_to_26_nov_13_jobs_12_titles_v5_Feature_Eng_Processed
k-WordPress_26_sep_to_26_nov_102_jobs_56_titles_v5_Feature_Eng_Processed
software_engineer_26_sep_to_26_nov_735_jobs_304_titles_v5_Feature_Eng_Processed
```


- filtering IT jobs_26_sep_to_26_nov_v5_k-AI-BI-ML-DL.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-Android-ios-Flutter-React Native+.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-app.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-back.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-Cloud-salesforce-oracle.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-data-Artificial-intelligence-Machine-Learning.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-desktop.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-Devops.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-dotnet-c++-c#.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-front.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-Full Stack.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-game-unity.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-IT.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-javascript-and-java.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-MERN_MEAN.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-Network.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-php.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-python.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-scrap.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-security_cyber.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-shopify.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-testing_SQA_scrum.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-UI-UX.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-web.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-web3-block-chain-crypto.ipynb
- filtering IT jobs_26_sep_to_26_nov_v5_k-wordpress.ipynb

Figure 5.6: Filtered Excel and Jupiter Labs Files

4.5 Concatenating Filtered IT Jobs in Filtered_Files-concatenated Folder:

In this phase, the filtered IT job listings from individual Excel files within the **Filtered_Files** folder are concatenated into a single consolidated file. This consolidation simplifies the subsequent analysis process by providing a unified dataset.

A consolidated CSV file was generated which has 4690 jobs with 1856 unique titles as shown by the name of the concatenated CSV file.

 IT_roles_Linkedin_Job_Details_and_Job_Descriptions_26-sep-to-26-nov_4690-jobs-1856_titles_v5_Feature_Eng_Processed.csv

4.5.1 Drop Duplicates from Concatenated Filtered File: Data Cleaning Phase 2:

Following the concatenation process, it was observed that there were 956 duplicate entries identified based on Job IDs. These duplicates were subsequently removed during the data-cleaning phase. After the removal of duplicates. The refined dataset is saved as an Excel file named "IT_roles_Linkedin_Job_Details_and_Job_Descriptions_26-sep-to-26-nov_3734-jobs_1856-titles_20-features_Not-duplicated_v5_Feature_Eng_Processed.csv", stored within the "Filtered_Files-concatenated" folder. It contains a total of 3734 IT jobs with 1856 unique

job titles and 20 features. This file serves as the consolidated dataset for further analysis and insights extraction.

4.6 EDA at the end of Segmentation

In this exploratory data analysis (EDA) phase, the concatenated filtered dataset is subjected to further analysis to gain deeper insights into its characteristics. Details on the number of unique values and value counts for each feature provide a comprehensive understanding of the dataset's attributes. The following table shows some observations conducted in EDA after the segmentation of IT jobs.

Table 5.1: EDA at the End of Segmentation of IT jobs

| | |
|--|------|
| How many Jobs or number of rows? | 3734 |
| How many features or columns? | 20 |
| How many Unique job titles? | 1856 |
| How many Unique companies? | 1420 |
| How many Unique Cities? | 49 |
| How many Unique provinces/Regions? | 17 |
| How many Unique countries? | 9 |
| How many Unique categories are in the company employee size feature? | 17 |
| How many Unique categories are in the company Industry feature? | 62 |

CHAPTER NO 6
Data Preparation 2

6.1 Introduction to Data Preparation Phase 2:

After filtering the IT jobs, we have separated the data, resulting in 3,734 IT job postings. In this section, we utilize this filtered IT jobs data to create more efficient features. This includes categorizing jobs into IT Fields, Sub-Fields, and Tools, and developing metrics for competition analysis, such as the applicants-to-hour ratio. These enhancements ensure our data is well-prepared for detailed analysis, allowing us to generate insightful dashboards and reports on the IT job market in Pakistan.

6.2 Why Categorization of IT Jobs is Needed:

In the realm of Information Technology (IT), the landscape of job titles can be overwhelming and often lacks clarity regarding specific skill demands. To address the varied and diversified demands reflected in the dataset of 3734 IT jobs, categorization into fields, sub-fields, and tools is imperative. Here's why:

i. Huge Job Titles:

- With 1856 unique job titles, the dataset highlights the vast array of roles within the IT industry. From "Software Engineer" to "Data Scientist" to "Network Administrator," the job titles encompass a broad spectrum of responsibilities and skill requirements.
- Categorization helps in simplifying this complexity by grouping similar job titles under overarching fields and sub-fields. For example, roles such as "Software Developer" and "Web Developer" can be categorized under the field of "Software Engineering."

ii. Lack of Clarification:

- Many job titles lack clarity regarding the specific skills and expertise needed for the role. For instance, a job title like "IT Specialist" could encompass a wide range of responsibilities, from technical support to system administration to cybersecurity.
- By categorizing jobs into fields and sub-fields, the ambiguity surrounding job titles is reduced. Job seekers can better understand the skill requirements of each category, enabling them to tailor their applications accordingly.

iii. Necessary for Skills Analysis:

- Analyzing the skills and competencies demanded by IT jobs is crucial for workforce planning and talent development. However, with such diverse job titles, conducting a skills analysis becomes challenging.
- Categorization allows for a systematic assessment of skill demands within each field and sub-field. This analysis can reveal trends in skill requirements, such as the prevalence of programming languages in software engineering roles or the importance of cybersecurity skills in network security positions.

6.3 Process of Grouping IT Jobs

The process of grouping IT jobs involves several steps aimed at reducing the number of job titles and enhancing clarity regarding skill demands:

- i. **Minimizing Job Titles:** Similar job titles are consolidated into a single job title to streamline the recruitment process and eliminate redundancy. This consolidation is based on similarities in responsibilities, required skills, and qualifications.
- ii. **Utilizing Filtered Data for Keyword Application:** IT jobs filtered data is employed to apply a specific keyword or set of keywords representing a particular skill or competency. This allows for a more precise matching of job roles with the skills possessed by potential candidates.
- iii. **Grouping into Job Title_2 Feature:** Job titles are further categorized into a secondary feature, Job Title_2, based on specific keywords associated with each role.

Example:

This process filters job titles containing keywords related to app development (e.g., 'App', 'Android', 'iOS', 'Mobile', 'Flutter', 'Xamarin', 'React Native', 'Swift', 'Kotlin', 'Ionic'), and assigns 'App Developer' to the 'Job Title_2' column for these entries, resulting in a refined categorization of job roles.

| | Job Title | Job Title_2 |
|---|-------------------------------------|---------------|
| 0 | iOS Intern | App Developer |
| 1 | Trainee iOS Developer | App Developer |
| 2 | Android Developer | App Developer |
| 3 | MERN and React Native Developer | App Developer |
| 4 | Mobile Application Developer | App Developer |
| 5 | iOS Developer | App Developer |
| 6 | Sr Flutter Developers | App Developer |
| 7 | Unity Android Mobile Game Developer | App Developer |
| 8 | Junior iOS Developer | App Developer |
| 9 | iOS Developer | App Developer |

Figure 6.1: Grouping of Job Titles into 'Job Title2' Feature

6.4 Process of Categorizing IT Jobs into IT Fields and Sub-Fields

Categorization of IT jobs is based on the identification of IT fields, sub-fields, and tools, each serving a distinct purpose in organizing job roles:

6.4.1 Identification of IT Fields:

IT fields encompass broad areas of specialization within the industry, such as Software Engineering, Artificial Intelligence (AI), Networking, Cybersecurity, etc. These fields serve as the overarching categories for organizing IT jobs.

Table 6.1: IT Fields

| Name of IT Field | A short description |
|--------------------------------------|---|
| Software Development | Focuses on creating and maintaining software applications. |
| Artificial Intelligence/Data Science | Utilizes algorithms and data analysis to build intelligent systems. |
| IT Support/IT Management | Utilizes algorithms and data analysis to build intelligent systems. |

| | |
|------------------------------|---|
| Cyber security | Protects systems and data from cyber threats. |
| Web3.0/Block chain/Metaverse | Develops decentralized technologies and virtual environments. |
| Computer Networking | Designs and maintains network infrastructures for data communication. |

6.4.2 Building IT Field Group Feature:

The IT field group feature is derived from job titles, Job Title_2, and keyword matching.

Example:

This process assigns 'Artificial Intelligence/Data Science' to the 'Field Group' column for job titles classified as 'AI/Data Science/Machine Learning Engineer' or 'Data Mining Engineer' in the 'Job Title_2' column, resulting in a clear categorization of these roles under the broader AI/Data Science field.

| | Job Title | Job Title_2 | Field Group |
|------|--|---|--------------------------------------|
| 928 | LinkedIn Data Extractor | Data Mining Engineer | Artificial Intelligence/Data Science |
| 929 | SQL Database Developer | AI/Data Science/Machine Learning Engineer | Artificial Intelligence/Data Science |
| 930 | Business Intelligence Developer | AI/Data Science/Machine Learning Engineer | Artificial Intelligence/Data Science |
| 931 | Research Analyst for Business Subjects (Office... | AI/Data Science/Machine Learning Engineer | Artificial Intelligence/Data Science |
| 932 | Business System Analyst | AI/Data Science/Machine Learning Engineer | Artificial Intelligence/Data Science |
| ... | ... | ... | ... |
| 2607 | AI Super Engineer, Trilogy (Remote) - \$200,000... | AI/Data Science/Machine Learning Engineer | Artificial Intelligence/Data Science |
| 2611 | AI Super Engineer, Trilogy (Remote) - \$200,000... | AI/Data Science/Machine Learning Engineer | Artificial Intelligence/Data Science |
| 2613 | ML Engineer | AI/Data Science/Machine Learning Engineer | Artificial Intelligence/Data Science |
| 2615 | AI Engineer | AI/Data Science/Machine Learning Engineer | Artificial Intelligence/Data Science |
| 2619 | Senior ML Engineer | AI/Data Science/Machine Learning Engineer | Artificial Intelligence/Data Science |

Figure 6.2: Formation of 'IT Field Group' Feature

6.4.3 Identification of IT Sub-Fields:

Sub-fields represent specialized domains within each IT field, providing further granularity in categorizing job roles. For example, within Software Engineering, sub-fields may include Web Development, Mobile App Development, Front End Development, etc.

Table 6.2: IT Sub-Fields

| | IT Sub-Field | IT Field |
|----|--------------------------------------|--------------------------------------|
| 1 | Mobile App Development | Software Development |
| 2 | Game Development | Software Development |
| 3 | Software Testing | Software Development |
| 4 | Cloud Development | Software Development |
| 5 | Cybersecurity | Software Development |
| 6 | Back-End Development | Software Development |
| 7 | Web Development | Software Development |
| 8 | Desktop Development | Software Development |
| 9 | Full-Stack Development | Software Development |
| 10 | DevOps Development | Software Development |
| 11 | Software Development | Software Development |
| 12 | Front-End Development | Software Development |
| 13 | Data Engineering | Artificial Intelligence/Data Science |
| 14 | Data Analytics/Business Intelligence | Artificial Intelligence/Data Science |
| 15 | Machine Learning/AI | Artificial Intelligence/Data Science |
| 16 | Data Science | Artificial Intelligence/Data Science |

| | | |
|----|-----------------------------|--------------------------------------|
| 17 | Computer Vision | Artificial Intelligence/Data Science |
| 18 | Natural Language Processing | Artificial Intelligence/Data Science |
| 19 | Data Mining/Data Annotation | Artificial Intelligence/Data Science |
| 20 | IT Support/IT Management | IT Support/IT Management |
| 21 | Web3.0/Blockchain/Metaverse | Web3.0/Blockchain/Metaverse |
| 22 | Computer Networking | Computer Networking |

There are a total of 22 sub-fields defined: 12 within Software Development, 7 within Data Science/AI, and one sub-field each for Cybersecurity, IT Support/IT Management, and Computer Networking.

6.4.4 Building IT Sub-Field Feature:

Similar to field identification, sub-field features are extracted using a combination of field groups, job titles, Job Title_2, and keyword matching. This helps in categorizing job roles into specific domains of expertise.

Example:

This process assigns 'Data Analytics/Business Intelligence' to the 'Sub-Field Group' column for job titles within the 'Artificial Intelligence/Data Science' field that contain keywords such as 'Analyst', 'Analysis', 'Analytics', 'Business Intelligence', 'Tableau', 'Power BI', 'Power BI', 'Excel', or 'BI', ensuring that relevant job roles are categorized under the appropriate sub-field.

| | Job Title | Field Group | Sub-Field Group |
|-----|---|--------------------------------------|--------------------------------------|
| 930 | Business Intelligence Developer | Artificial Intelligence/Data Science | Data Analytics/Business Intelligence |
| 931 | Research Analyst for Business Subjects (Office... | Artificial Intelligence/Data Science | Data Analytics/Business Intelligence |
| 932 | Business System Analyst | Artificial Intelligence/Data Science | Data Analytics/Business Intelligence |
| 933 | Research Analyst - Finance | Artificial Intelligence/Data Science | Data Analytics/Business Intelligence |
| 937 | Data Analyst | Artificial Intelligence/Data Science | Data Analytics/Business Intelligence |

Figure 6.3: Formation of 'IT Sub-Field Group' Feature

6.5 Extraction of Tools/Platforms Feature:

Tools refer to the technologies, platforms, and software applications commonly used within IT roles. Identifying and categorizing these tools allows for a more comprehensive understanding of skill requirements and job responsibilities.

Below is a list of 89 Tools/Platforms to search these keywords in the Job Title or Job description.

```
Set_of_Tools_Platforms=[
"Python", "Java", "JavaScript", "C#", "C++", "Ruby", "Swift", "Kotlin", "TypeScript", "PHP", "Rust", "HTML", "CSS", "Unity",
"React", "Angular", "Vue", "Node", "Django", "Flask", ".NET", "dotnet", "Express", "MERN", "MEAN"
"Laravel", "Bootstrap", "SQL"
"MySQL", "PostgreSQL", "MongoDB", "SQL Server", "Oracle",
"Docker", "Kubernetes", "Jenkins", "Git", "GitHub",
"AWS", "Amazon Web Services", "Azure", "Google Cloud Platform", "GCP"
"TensorFlow", "PyTorch", "scikit-learn", "Keras", "Spark", "Hadoop", "FastAPI",
"Android", "React Native", "Flutter", "Xamarin",
"Slack", "Microsoft Teams",
"VS Code", "IntelliJ IDEA", "Eclipse",
"Word Press", "Word-Press", "WordPress", "Word_Press", "Joomla", "Drupal", "Wix", "Squarespace", "Shopify", "Magento", "Webflow", "Ghost", "Strill
"R Programming", "Jupyter Notebook", "Pandas", "NumPy", "Matplotlib", "Seaborn", "Plotly", "Tableau", "Power BI", "Excel", "Google Colab", "Goo
"MS Office", "MS Dynamics", "Unreal", "jQuery", "Salesforce", "IBM", "Ionic", "ios"]
```

Figure 6.4: List of Tools/Platforms

6.5.1 Extracting Tools/Platforms based on Job Titles:

This process defines a list of tools and platforms, then assigns a value from this list to a new 'Tools/Platforms' column in the Data Frame based on whether a job title contains any of these keywords, thereby identifying relevant technologies associated with each job role.

| | Job Title | Tools/Platforms |
|------|--|-----------------|
| 0 | iOS Intern | ios |
| 1 | Trainee iOS Developer | ios |
| 2 | Android Developer | Android |
| 3 | MERN and React Native Developer | React |
| 4 | Mobile Application Developer | None |
| ... | ... | ... |
| 3729 | Pre-Sales Software | None |
| 3730 | Lead Software Engineer, Trilogy (Remote) - \$10... | None |
| 3731 | QA Engineer | None |
| 3732 | Principal Software Engineer, IgniteTech (Remot... | None |
| 3733 | Software Development Support Engineer (L2 & L3) | None |

Figure 6.5: Extraction of 'Tools and Platforms' Feature from Job titles

The 'None' value in the 'Tools/Platforms' column indicates that no keyword from the defined list of tools and platforms was found in the job title, signifying that the job title does not explicitly mention any of the specified technologies.

6.5.2 Extracting Tools/Platforms_2 based on Job Descriptions:

This process defines a function to search for keywords from a list of tools and platforms within job descriptions, and then assigns the first matching keyword to a new 'Tools/Platforms_2' column in the Data Frame, with 'None' indicating that no specified tools or platforms were found in the job description or jobs description is null or inconsistent.

| | Job Title | Tools/Platforms_2 |
|------|--|-------------------|
| 0 | iOS Intern | Swift |
| 1 | Trainee iOS Developer | Swift |
| 2 | Android Developer | None |
| 3 | MERN and React Native Developer | Java |
| 4 | Mobile Application Developer | Java |
| ... | ... | ... |
| 3729 | Pre-Sales Software | Excel |
| 3730 | Lead Software Engineer, Trilogy (Remote) - \$10... | Unity |
| 3731 | QA Engineer | Java |
| 3732 | Principal Software Engineer, IgniteTech (Remot... | C++ |
| 3733 | Software Development Support Engineer (L2 & L3) | None |

Figure 6.6: Extraction of 'Tools/Platforms2' from Job descriptions

6.6 Working on Extracting Competition Analysis Features:

6.6.1 Exploring Numbers of Applicants Feature and Job Posting Time:

This involves analyzing the distribution and characteristics of the number of applicants for each job posting to understand applicant behavior and job popularity.

Table 6.3: Exploring unique and missing values in 'Number of applicants' feature

| | |
|---|----------------|
| Number of Unique Values in Number of Applicants Feature | 396 |
| The data type of the Number of Applicants Feature | Object /String |

| | |
|--|------|
| Missing values in the Number of Applicants Feature | 1489 |
|--|------|

Table 6.4: Exploring unique and missing values in 'Job posting Time 'feature

| | |
|---|----------------|
| Number of Unique Values in Job Posting Time Feature | 40 |
| The data type of the Job Posting Time Feature | Object /String |
| Missing values in the Job Posting Time Feature | 0 |

Below are the first 30 unique values of the "Number of Applicants" feature, this feature needs to be converted to float to ensure accurate numerical analysis.

```
array([nan, '51', '50', '132', '24', '33', '298', '247', '335', '926',
       '95', '346', '83', '52', '36', '79', '68', '59', '72', '71', '10',
       '7', '27', '25', '18', '35', '28', '117', '118', '546'],
```

Figure 6.7: Unique values in number of applicants Feature

Below are the unique values of the "Job Posting Time " feature, this feature also needs to be converted to float to ensure accurate numerical analysis.

```
array(['2 weeks ago', '1 week ago', '6 days ago', '2 hours ago',
       '1 day ago', '4 hours ago', '5 days ago', '2 days ago',
       '19 hours ago', '4 days ago', '12 hours ago', '3 days ago',
       '22 hours ago', '17 hours ago', '14 hours ago', '11 hours ago',
       '16 hours ago', '20 hours ago', '3 hours ago', '10 hours ago',
       '15 hours ago', '1 month ago', '3 weeks ago', '32 minutes ago',
       '5 hours ago', '7 hours ago', '13 hours ago', '18 hours ago',
       '9 hours ago', '21 hours ago', '8 hours ago', '6 hours ago',
       '23 hours ago', '1 hour ago', '49 minutes ago', '44 minutes ago',
       '55 minutes ago',
       'https://anitechnologies.net/ · Islamabad, Islāmābād, Pakistan Reposted ',
       '13 minutes ago', '14 minutes ago'], dtype=object)
```

Figure 6.8: Unique values in Job Posting Time Feature

6.7 Normalizing the Number of Applicants and Job Posting Time Feature

6.7.1 Conversion of the Number of Applicants feature into numerical feature:

As explored previously the Number of applicants has values in the form of a string that represents it as a categorical feature. So it is necessary to convert the Number of Applicants feature into the numerical feature. So that is why the number of applicants is converted to Float.

```
[nan, '51', '50', '132', '24', '33']
```



```
nan, 5.100e+01, 5.000e+01, 1.320e+02, 2.400e+01, 3.300e+01,
```

Figure 6.9: *Converting Number of applicants into numerical*

Below is the statistical Description of the Number of Applicants Feature.

```
count    2245.000000
mean     101.856570
std      190.079714
min       0.000000
25%      17.000000
50%      42.000000
75%     109.000000
max     3052.000000
Name: n_applicant, dtype: float64
```

Figure 6.10: *Statistical Distribution of 'Number of applicants' Feature*

6.7.2 Addressing Null Values in the Number of Applicants feature:

Missing values in the Number of Applicants feature are addressed by filling it with the median value which is 42.0.

6.7.3 Extracting Job Posting Time in only Numerical Form and Conversion into equivalent hours:

This process converts 'job posting time' strings from weeks, days, minutes, and months into equivalent hours using a regular expression, applies this conversion to the entire 'job posting time' column in the Data Frame, and then removes the strings ' hours ago' and ' hour ago' to isolate the numerical values and convert into Float.

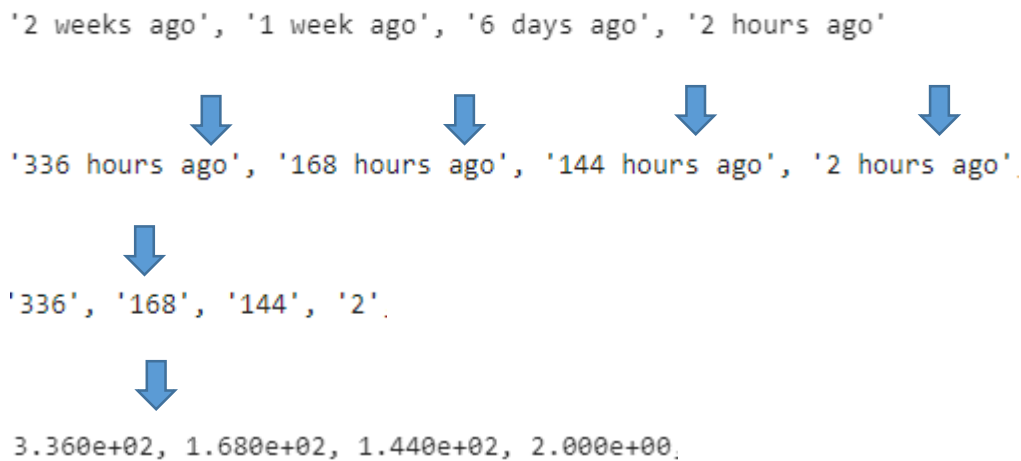


Figure 6.11: Converting job posting time into hours numerical

6.8 Building Competition Analysis Features

6.8.1 Building Feature 'Time to Number of Applicants Ratio':

This process creates a new column, 'Time to Number of applicants Ratio', by dividing the 'job posting time' by the 'n_applicant' (Number of Applicants) for each row where 'n_applicant' is not zero, and assigning a value of zero where 'n_applicant' is zero.

Equation 6.1: Time to Number of applicants Ratio

Time to Number of Applicants Ratio = Job Posting Time / Number of Applicants

6.8.2 Using "Time to Number of Applicants Ratio" for Competition Analysis:

This ratio indicates how many hours a job posting has been live per applicant.

- **Less Ratio, More Competition:** A lower ratio means many applicants quickly, indicating high competition for the job.
- **More Ratio, Less Competition:** A higher ratio means fewer applicants over time, indicating lower competition for the job.

6.8.3 Building a New Feature: Applicants to Hour Ratio

This Process Creates a new column, 'Applicants to Hour Ratio', by dividing the 'n_applicant' (Number of Applicants) by the 'job posting time'.

- **Mathematical formula:**

Equation 6.2: Applicants to Hour ratio

Applicants to Hour Ratio = Number of Applicants / Job Posting Time

6.8.4 How Applicants to Hour Ratio Can Be Utilized for Competition Analysis

- **More Ratio, More Competition:** A higher ratio means more applicants per hour, indicating higher competition for the job.
- **Less Ratio, Less Competition:** A lower ratio means fewer applicants per hour, indicating lower competition for the job.
- **Direct Relation:** This ratio has a direct relationship with competition: more applicants per hour signifies higher competition.
- **Insights:**
 - **For Specific Jobs:** Determine the number of applicants per hour for a specific job or a set of jobs.
 - **Efficiency:** The applicants-to-hour ratio is an efficient metric for analyzing the competitiveness of job postings.

6.8.5 Which is better? Applicants to Hour Ratio OR Time to Number of Applicants Ratio:

The applicants-to-hour ratio is a better metric for competition analysis due to its ability to provide real-time, precise, and easily interpretable insights into the competitiveness of job postings, compared to the more historical and averaged perspective offered by the Time-to-number of Applicants Ratio. I utilized this feature for competition analysis later on by building dashboards and reporting tools, which enabled dynamic and insightful visualizations of job market trends and competition levels.

6.9 Exploring Other Possibilities to Make New Features

Exploring other possibilities to enrich feature sets involves considering aspects beyond the immediate dataset. For instance, while job descriptions and company info contain potentially

valuable insights, certain attributes such as the number of employees and company industry already captured through the extraction of company size and industry. Nonetheless, job descriptions offer a wealth of untapped potential, including detailed insights into required skills, educational prerequisites, qualifications, and offered benefits. However, extracting meaningful features from job descriptions poses significant challenges due to their unstructured nature, often compounded by incomplete or null values. Leveraging advanced Natural Language Processing (NLP) techniques becomes essential in unraveling the complexities inherent in job descriptions, enabling the extraction of valuable insights despite the inherent difficulties.

6.9.1 Extraction of Education Requirement Status:

This process checks if specific education-related keywords are present in the job descriptions of the Data Frame and assign 'Yes' to the 'Education Requirement Status' column if any of the keywords are found, and 'No' if none of the keywords are found.

```
Edu_keywords
```

```
['Education',  
'Degree',  
'bachelor',  
'master',  
'phd',  
'Computer Science',  
'Computer Engineering',  
'Information Technology',  
'Software Engineering',  
'Computer Information Systems',  
'Information Systems',  
'Computer Applications',  
'Computer Programming',  
'Computer Networks and Security',  
'Data Science',  
'Artificial Intelligence',  
'Bioinformatics',  
'Business Information Systems',  
'Computer Games Development',  
'Human-Computer Interaction',  
'Computational Mathematics',  
'Multimedia Computing',  
'Mobile Computing',  
'Web Development',  
'Cloud Computing']
```

Figure 6.12: Education and Degree related keywords

The below process involves counting the occurrences of 'Yes' and 'No' values in the 'Education Requirement Status' column of the Data Frame, revealing that 2596 job postings specify education requirements while 1127 do not.

```
Education Requirement Status
Yes      2596
No       1127
Name: count, dtype: int64
```

Figure 6.13: Value Count of Education Requirement Status 'Yes' or No'

6.9.2 Why Job Description and Company Info Feature is Not Utilized Much:

There are several reasons why job descriptions and company information features are not extensively utilized for skills analysis:

- i. **Null Values:** Many job postings may have missing (null) values in these fields, leading to incomplete data that is difficult to analyze accurately.
- ii. **Incomplete Descriptions:** Some job descriptions may be very short, with fewer than 20 characters, providing insufficient information for meaningful analysis.
- iii. **Unstructured Text:** Job descriptions and company information often contain unstructured text, making it challenging to extract useful data without advanced natural language processing (NLP) techniques.
- iv. **Relevance:** Compared to other structured data like job titles, locations, and applicant numbers, job descriptions, and company information might offer less immediate insight for skills analysis, as they are more variable and less consistent in describing skills and requirements.

CHAPTER NO 7
Data Visualization and Reporting

7.1 Data Visualization:

The last phase of the data analytics project is Data Visualization and Reporting. In this section, we explore various methods to visualize data, demonstrating how effective visualization can enhance the understanding of complex datasets. This part of the project covers the techniques and tools used for data visualization, specifically focusing on Python and Power BI. By leveraging these tools, we can create insightful and interactive visualizations that support the findings and conclusions of the project, making the data more accessible and actionable for stakeholders.

7.2 Concepts of data visualization:

7.2.1 Data visualization and its importance:

- **Definition:** Data visualization is the graphical representation of information and data. It uses visual elements like charts, graphs, and maps to help viewers understand trends, patterns, and insights in data.
- **Importance:** It enhances data comprehension, aids decision-making, and communicates complex information effectively.

7.2.2 Data Visualization in Python:

- **Libraries:** Python offers various libraries like Matplotlib, Seaborn, and Plotly for data visualization.

7.2.3 Types of Charts and Graphs:

Python supports various charts and graphs, including horizontal and vertical bar charts, pie charts, and count plots. Sub-plots are suitable for different data types and analysis purposes.

7.2.4 Numerical Analysis vs. Categorical Analysis:

- **Numerical Analysis:** Involves quantitative data analysis, dealing with numbers and measurements.
- **Categorical Analysis:** Focuses on qualitative data analysis, dealing with categories or labels.

7.3 Data Visualization in Python Project Structure:

Data visualization in Python is completed in Five Stages which are explained below:

7.3.1 Data Visualization in Skills Analysis:

- **Visualizing Field Group:** This process creates a horizontal bar chart to display the count of jobs by "Field Group," followed by a pie chart illustrating the percentage distribution of employment across these field groups. It then generates a series of subplots for each field group to show detailed distributions, including the top 20 companies, top 10 cities, remote/onsite status, top 10 industries, types of employment, and experience levels.
- **Visualizing Sub-Field Group:** This process starts by creating a horizontal bar chart to display the count of jobs by "Sub-Field Group," followed by a pie chart illustrating the percentage distribution of jobs across these sub-field groups. It then generates a series of subplots for each sub-field group to show detailed distributions, including the top 20 companies, top 10 cities, remote/onsite status, top 10 industries, types of employment, and experience levels.

7.3.2 Data Visualization in Company Analysis:

- **Visualizing Company Data with Bar and Pie Charts:** This process first creates a horizontal bar chart to display the top 20 companies by number of jobs, followed by pie charts showing the percentage of these top 20 companies out of the total number of companies and jobs. It then generates detailed subplots for each of these top 20 companies, illustrating distributions of field groups, cities, remote/onsite status, industries, types of employment, and experience levels.

7.3.3 Data Visualization in Geography Analysis:

- **Visualizing City Feature with Bar and Pie Charts:** This process starts with a horizontal bar chart showing the top 20 cities by job count, then displays two pie charts representing the top 10 cities' percentage out of all 57 cities and by number of jobs. It concludes with detailed subplots for each of the top 20 cities, illustrating distributions

of field groups, companies, remote/onsite status, industries, types of employment, and experience levels.

- **Visualizing Region/Province Data with Bar and Pie Charts:** This process begins with a horizontal bar chart depicting the job counts across different regions/provinces, followed by a pie chart that shows the percentage distribution of jobs among the regions/provinces. It then provides detailed subplots for the top 20 regions/provinces, displaying distributions of field groups, companies, remote/onsite status, industries, types of employment, and experience levels.
- **Visualizing Country Data with Pie and Bar Charts:** This process involves creating a pie chart that displays the percentage distribution of jobs across different countries, followed by a horizontal bar chart to illustrate the job counts in each country.

7.3.4 Data Visualization in Job Nature Analysis:

- **Visualizing Remote/Onsite Job Data with Count Plots and Pie Chart:** This process includes generating multiple visualizations, starting with count plots to display the distribution of jobs based on Remote/Onsite status, Job Type, and Experience Level, followed by a pie chart showing the percentage breakdown of Remote/Onsite jobs, and detailed bar plots for each Remote/Onsite category across various job-related attributes.
- **Visualizing Type of Employment Data with Count Plots and Pie Charts:** This process involves creating visualizations starting with count plots to show the distribution of jobs based on Type of Employment, both overall and segmented by Remote/Onsite status and Experience Level, followed by a pie chart illustrating the percentage distribution of Types of Employment, and detailed bar plots for each employment type category across various job-related attributes.
- **Visualizing Experience Level Data with Count Plots and Pie Charts:** This process involves generating count plots to display the distribution of jobs by Experience Level, including breakdowns by Job Type and Remote/Onsite status, followed by a pie chart illustrating the percentage distribution of jobs across different Experience Levels.

7.3.5 Data Visualization in Industry Analysis:

- Visualizing the Number of Employees (Company Employee Size) in the Company:**
 This process involves creating a horizontal bar plot to show the top 20 companies by the number of employees, followed by a pie chart depicting the percentage distribution of jobs across different company sizes based on the number of employees.
- Visualizing Company Industry Distribution:** This process involves creating a pie chart to show the percentage distribution of jobs across different industries, followed by a horizontal bar plot to depict the top 20 sectors by the number of jobs.

7.3.6 Preview of Horizontal Charts, Pie Charts, and Subplots in Python

- A horizontal bar chart to display the count of jobs by "Field Group":**

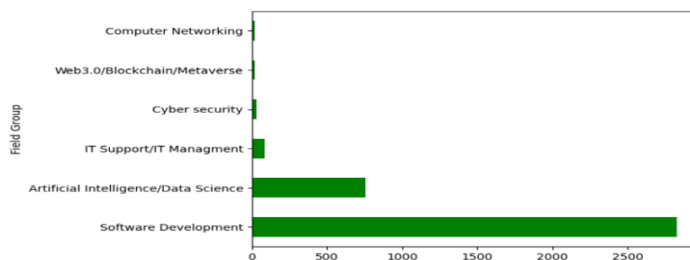


Figure 7.1: Horizontal bar chart to show value count of IT fields

- A pie chart illustrating the percentage distribution of IT jobs across 6 IT Fields.**

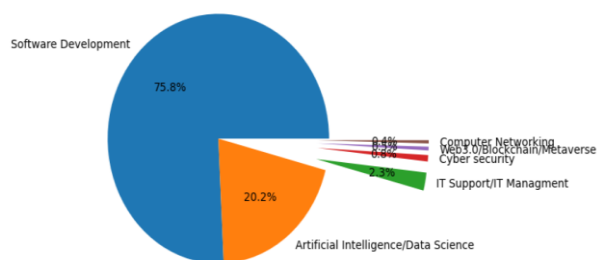


Figure 7.2: Pie chart to show percentage distribution

- Subplots for the Software Development Field showcasing detailed distributions:**
- Top 20 companies, Top 10 cities, Remote/Onsite status, Top 10 industries, Types of employment, and Experience levels

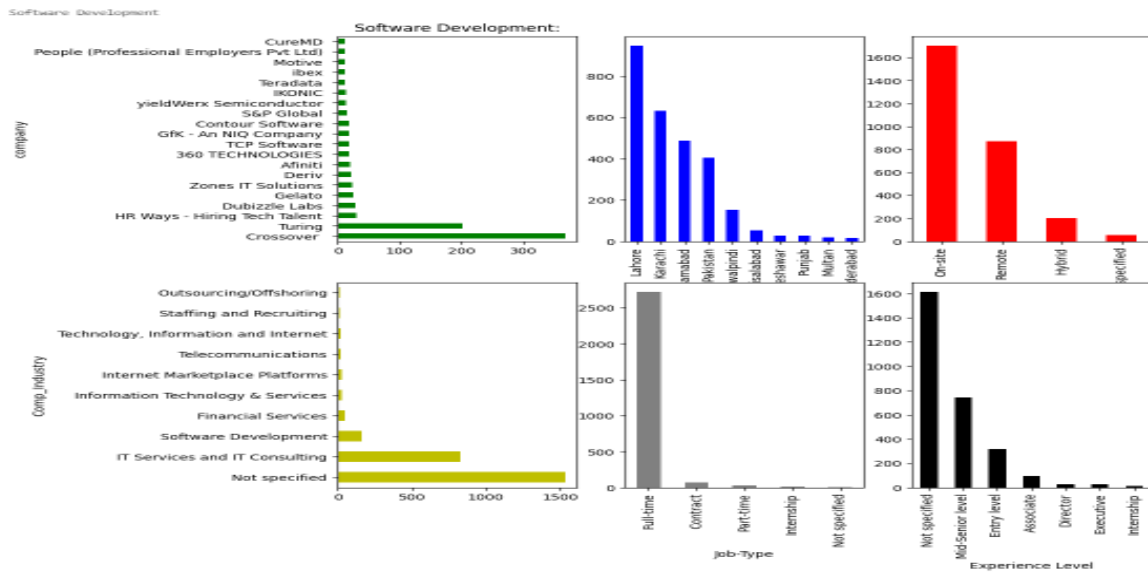
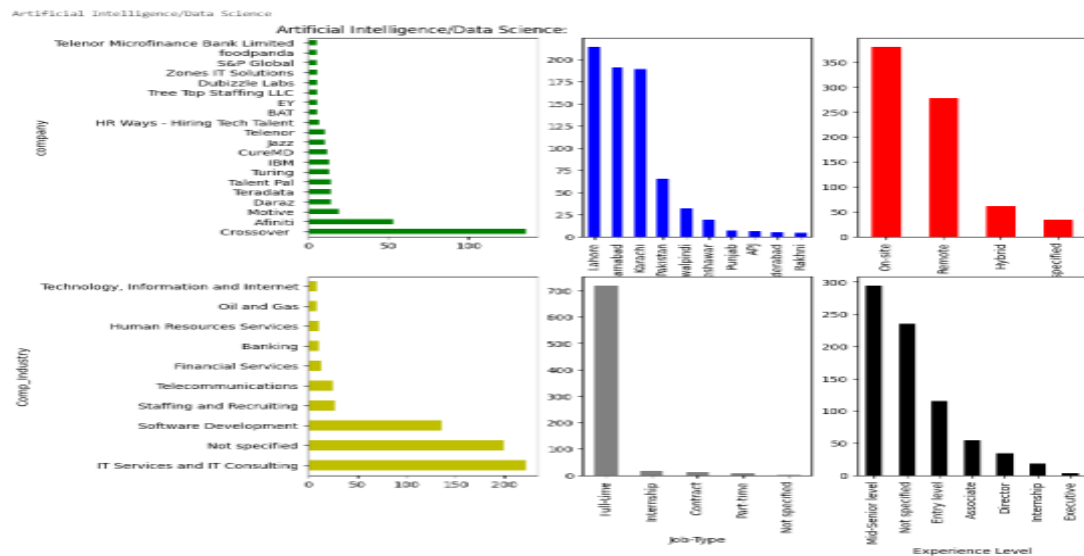


Figure 7.3: Dashboard of Sub-Plots to show detailed Distribution of jobs with respect to other features

- **Subplots for AI Fields** showcasing detailed distributions: Top 20 companies, Top 10 cities, Remote/Onsite status, Top 10 industries, Types of employment, and Experience level.



7.4 Data Visualization and Dashboard Building in Power BI

In today's data-driven world, the ability to effectively analyze and visualize data is crucial for making informed decisions. Power BI, a powerful business intelligence tool developed by

Microsoft, offers a comprehensive platform for creating interactive and insightful dashboards. This thesis explores the significance of Power BI in data visualization and dashboard building, focusing on its various features and capabilities.

Why Power BI: Power BI stands out as a preferred choice for data visualization due to its user-friendly interface, robust functionality, and seamless integration with other Microsoft products. Its ability to connect to multiple data sources, perform complex data transformations, and generate interactive reports makes it an indispensable tool for businesses seeking to harness the power of their data.

Dashboard Building: Dashboard building in Power BI involves the creation of visualizations that provide a comprehensive overview of key performance indicators (KPIs) and trends within an organization. These dashboards can be customized to suit specific business needs, allowing users to monitor performance, identify areas for improvement, and make data-driven decisions.

KPIs: Key Performance Indicators (KPIs) are essential metrics used to evaluate the success of an organization or specific activities within it. In Power BI, KPIs can be visualized using various charts and graphs, providing stakeholders with real-time insights into performance against predefined targets or benchmarks.

Types of Charts: Power BI offers a wide range of chart types to visualize data effectively. These include row charts, column charts, pie charts, and more. Each chart type has its advantages and is selected based on the nature of the data and the insights required. Additionally, Power BI allows for the implementation of filters to drill down into specific data subsets, enhancing the overall dashboard experience.

7.5 Structure of Power BI Dashboard

The Power BI dashboard is structured to provide comprehensive insights into various aspects of IT job trends, encompassing skills analysis, company analysis, geography analysis, and employment-related metrics. Each section of the dashboard focuses on specific dimensions of IT job trends, allowing stakeholders to gain actionable insights and make informed decisions.

7.5.1 IT Jobs Trends Analysis:

This page offers an overarching analysis of IT job trends, covering skills, geography, and job nature. It provides a holistic view of the IT job market landscape, highlighting key trends and patterns.

7.5.2 Skills Analysis based on IT Fields:

Focused on six major IT fields, this section delves into the number of jobs and competition within each field. It helps identify the demand for specific IT skills and the level of competition in the market.

7.5.3 Skills Analysis based on IT Sub-Fields:

Zooming in further, this page examines 22 IT sub-fields, providing insights into job trends and competition at a more granular level. It helps stakeholders understand the nuances of different IT specializations.

7.5.4 Skills Analysis Based on IT Tools/Platforms:

This section explores job trends and competition related to 47 IT tools and platforms. It sheds light on the popularity and demand for specific technologies within the IT job market.

7.5.5 Company Analysis:

Focused on 1420 IT companies, this page analyzes the number of jobs and competition among different companies. It helps stakeholders identify key players in the IT industry and assess their market presence.

7.5.6 City-Based Geography Analysis:

Examining job trends across 49 cities, this section provides insights into the distribution of IT jobs and competition at the city level. It helps stakeholders understand regional variations in the IT job market.

7.5.7 Province/Region-based Geography Analysis:

This page explores job trends and competition across 17 provinces/regions. It highlights regional disparities in the IT job market landscape, enabling stakeholders to tailor their strategies accordingly.

7.5.8 Country-based Geography Analysis:

Focusing on nine countries, this section provides a comparative analysis of IT job trends and competition at the national level. It helps stakeholders identify lucrative markets and potential expansion opportunities.

7.5.9 Employment Type Analysis:

Examining six types of employment, this page analyzes job trends and competition across different employment arrangements. It provides insights into the prevalence of full-time, part-time, contract, and freelance positions in the IT industry.

7.5.10 Onsite/Remote Analysis:

This section analyzes job trends and competition based on onsite/remote work arrangements across four categories. It sheds light on the growing trend of remote work and its impact on the IT job market.

7.5.11 Seniority Level (Experience Level) Analysis:

Focused on seven categories of seniority levels, this page examines job trends and competition across different career stages. It helps stakeholders understand the distribution of job opportunities based on experience and expertise.

7.5.12 Company Employee Size Analysis:

Examining job trends and competition across nine categories of company employee size, this section provides insights into the hiring practices of companies based on their workforce size.

7.5.13 Company Industry-based Analysis:

Focused on 62 categories of company industries, this page analyzes job trends and competition across various sectors. It helps stakeholders identify emerging industries and assess their potential for growth in the IT job market.

7.5.14 Preview of IT Jobs Trends and Skills Analysis Dashboards:

The dashboards presented below provide a glimpse into the IT Jobs Trends Analysis on page 1 and Skills Analysis on page 2, offering insights into the overall landscape of IT job trends and the specific skill sets in demand within the industry.

- IT Jobs Trends Analysis Dashboard
- IT Field-based Skills Analysis Dashboard

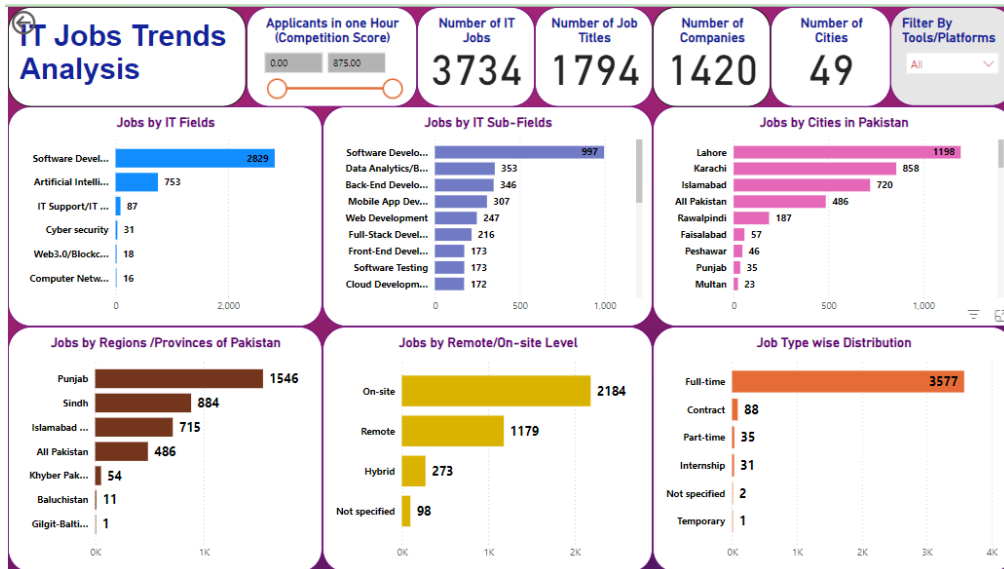


Figure 7.4: Power BI dashboard for overall Jobs Trends Analysis page 1

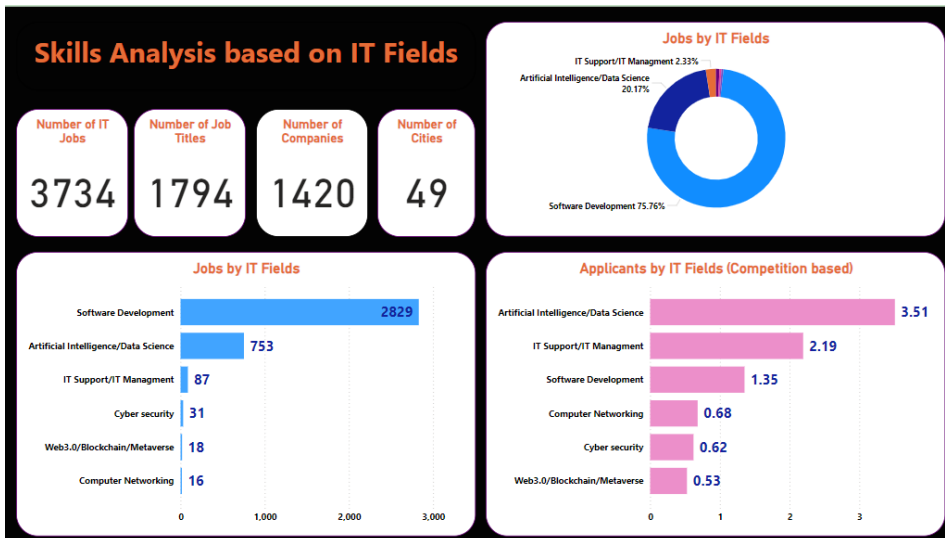


Figure 7.5: Power BI dashboard for Skills Analysis based on IT Fields

CHAPTER NO 8
Results and Discussion

8.1 Introduction Results and Discussion:

This chapter presents a detailed analysis of the IT job market in Pakistan, highlighting key findings and trends. Through this comprehensive analysis, all research questions posed in this study are addressed. The visualization of data, accompanied by detailed descriptions and analyses, supports the findings of the project.

The chapter examines jobs from multiple perspectives, including skills demand, geographical distribution, company-specific patterns, industry-based distribution, and competition analysis.

Overall, the "Results and Discussion" chapter provides strategic insights for job seekers, educators, and policymakers to enhance the IT job market in Pakistan, answering all the research questions and supporting the project's findings with detailed analysis and visualization.

8.2 Skills Analysis:

A key question in the IT job market is identifying which fields, programming languages, and tools have the highest demand, as well as understanding the competition for these jobs. This section provides an in-depth skills demand analysis divided into three parts: 6 IT Fields, 22 IT Sub-Fields, and 47 Tools/Programming Languages.

8.2.1 Skills Analysis based on IT Fields:

Jobs by IT Fields

The job market in IT is heavily dominated by software development, which accounts for 75.76% of the total job openings with 2829 positions available. This high demand underscores the critical role of software developers in the tech industry, as they are essential for creating, maintaining, and improving software systems and applications.

Artificial Intelligence (AI) and Data Science also represent a significant portion of the job market, with 753 job openings making up 20.17%. These fields are growing rapidly due to their big data, machine learning, and predictive analytics applications.

In contrast, IT Support/IT Management, Cyber Security, Web 3.0/Blockchain/Metaverse, and Computer Networking collectively hold a minor market share, with job openings ranging from 87 to 16, indicating more specialized and niche opportunities.

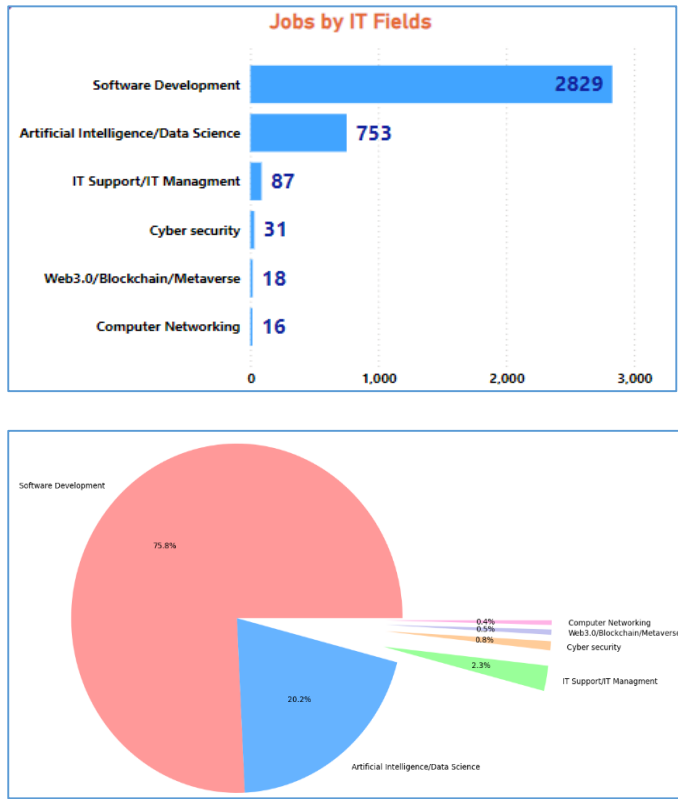


Figure 8.1: Distribution of Jobs across IT Fields

Applicants by IT Fields (Competition Score: Number of Applicants Per Hour)

The competition for jobs varies significantly across different IT fields. Artificial Intelligence and Data Science are the most competitive, with 3.51 applicants per job. This high level of competition reflects the popularity and attractiveness of these fields, driven by the increasing importance of data-driven decision-making in businesses. IT Support/IT Management also faces high competition, with 2.19 applicants per job, suggesting a saturated job market. In contrast, Software Development has a more balanced scenario with 1.35 applicants per job, indicating ample opportunities relative to the number of job seekers. Fields like Cyber Security, Web3.0/Blockchain/Metaverse, and Computer Networking have the lowest competition, with

less than one applicant per job, highlighting these as areas with potential for easier entry for those with the right skills.

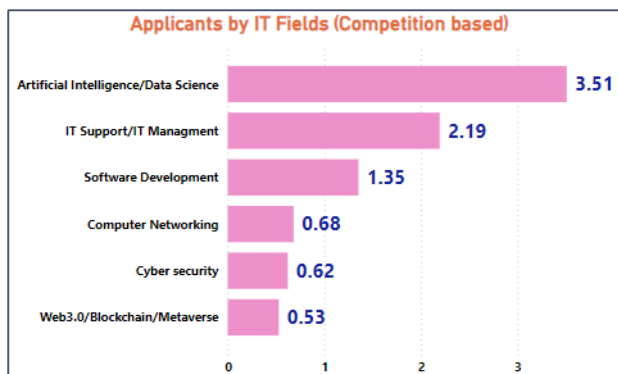


Figure 8.2: Competition by IT Fields

Combined Strategic Insights

Combining the insights from job availability and competition levels provides a strategic overview for job seekers in the IT field. Software Development offers the most opportunities with moderate competition, making it a lucrative option for those with the necessary skills and qualifications. Artificial Intelligence and Data Science, while also offering a significant number of jobs, are highly competitive, requiring job seekers to have advanced expertise to stand out. On the other hand, IT Support/IT Management presents a challenging landscape due to limited job availability and high competition. For those seeking less crowded job markets, specialized fields like Cyber Security, Web 3.0/Blockchain/Metaverse, and Computer Networking offer fewer job openings but also significantly lower competition, presenting opportunities for individuals with niche skills to secure positions more easily. By aligning their career strategies with these market dynamics, job seekers can optimize their chances of success in the IT industry.

8.2.2 Skills Analysis based on IT Sub-Fields:

Jobs by IT Sub-Fields:

The job market in IT is diverse, with Software Development leading significantly with 997 job openings, accounting for 26.75% of the total IT jobs. This is followed by Data Analytics/Business Intelligence and Back-End Development, which offer 353 (9.45%) and 346

(9.27%) jobs respectively. Other prominent sub-fields include Mobile App Development (307 jobs), Web Development (247 jobs), and Full-Stack Development (216 jobs). While fields like Cyber Security (31 jobs), Web 3.0/Blockchain/Metaverse (18 jobs), and Natural Language Processing (2 jobs) have fewer openings, they still represent critical, specialized areas within the IT landscape.

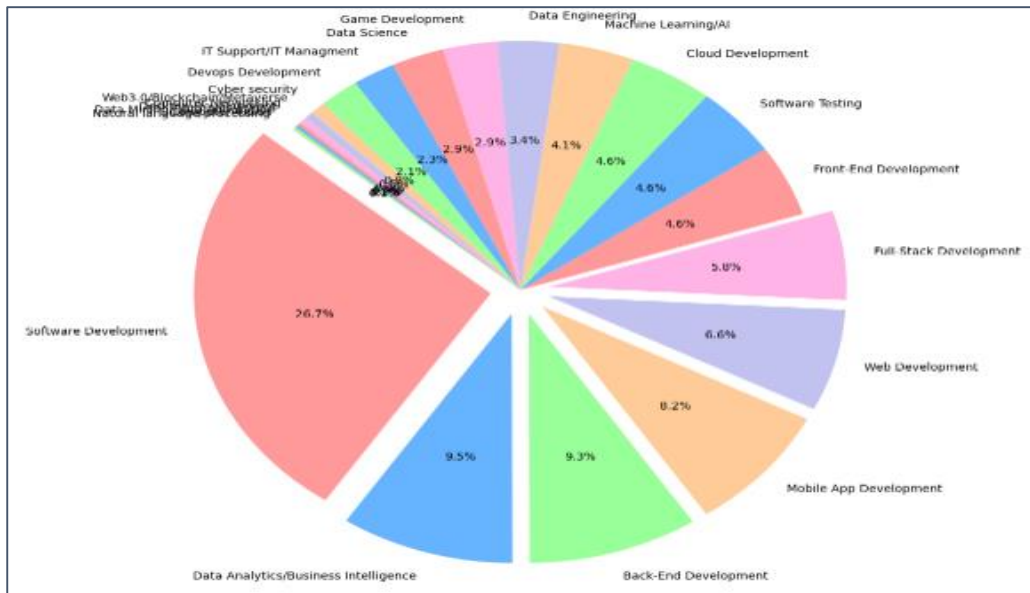
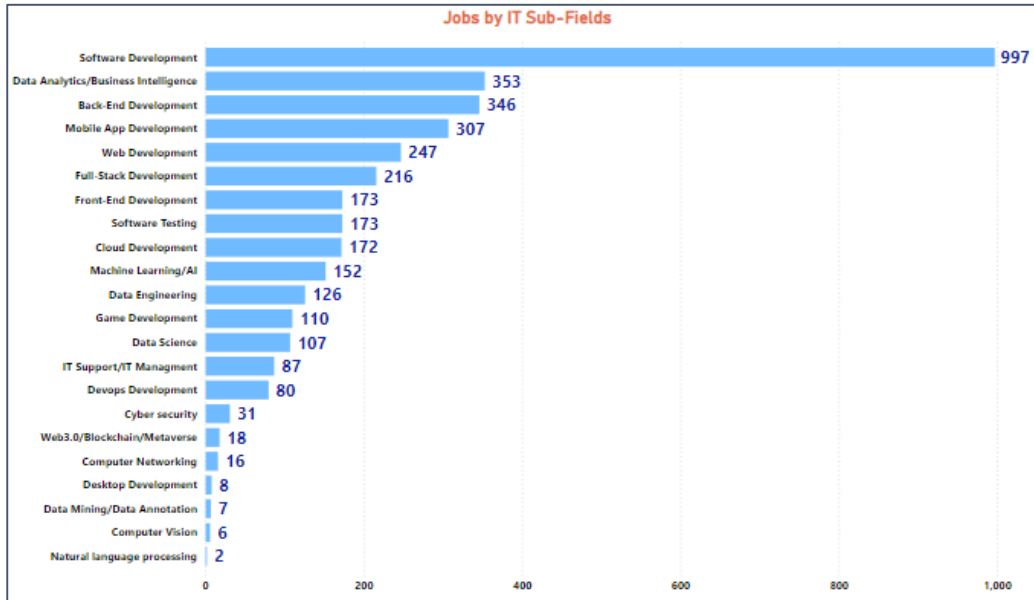


Figure 8.3: Distribution of Jobs across IT Sub-Fields

Applicants by IT Sub-Fields (Competition Score: Number of Applicants Per Hour):

Competition for jobs varies widely across IT sub-fields. Computer Vision and Data Science face the highest competition, with 9.18 and 8.93 applicants per job, respectively. Data Analytics/Business Intelligence and Natural Language Processing also experience high competition ratios of 3.97 and 3.84 applicants per job. Meanwhile, Software Development, despite its large number of openings, has a moderate competition level with 1.71 applicants per job. Sub-fields such as IT Support/IT Management and Front-End Development also see relatively high competition with 2.19 and 2.68 applicants per job, respectively. Lower competition is observed in specialized areas like Web 3.0/Blockchain/Metaverse (0.53 applicants per job) and Cyber Security (0.62 applicants per job).

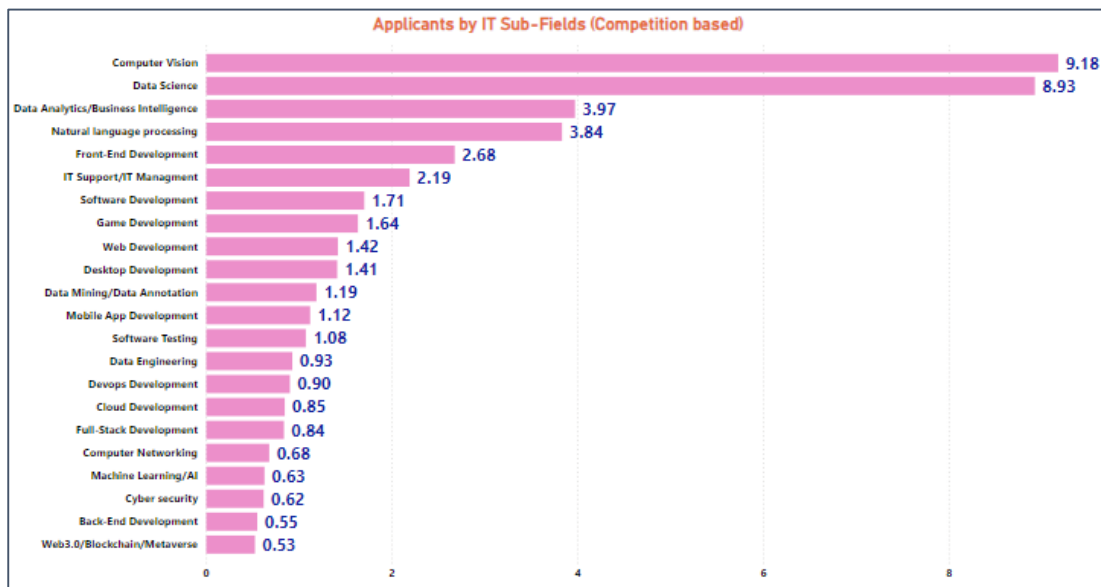


Figure 8.4: Competition by IT Sub-Fields

Combined Strategic Insights:

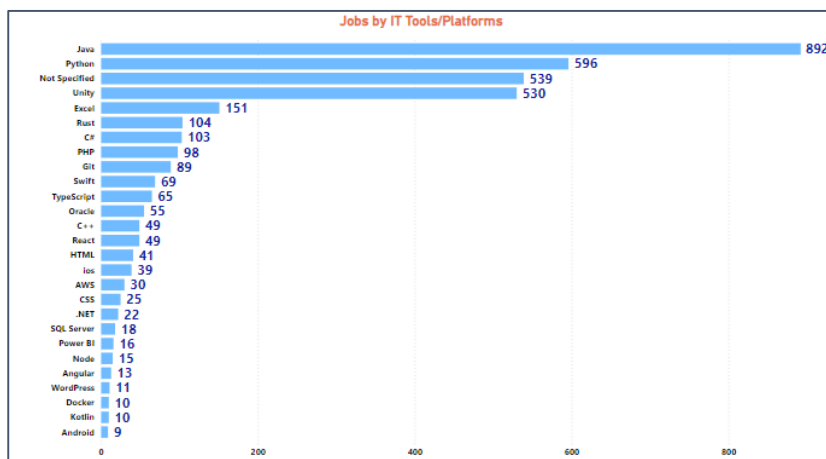
The analysis of job availability and competition within IT sub-fields reveal strategic insights for job seekers. Software Development, with its ample job opportunities and moderate competition, is an attractive field for those equipped with the necessary skills. Data Analytics/Business Intelligence and Back-End Development also present numerous

opportunities, though with varying levels of competition. Conversely, sub-fields like Computer Vision and Data Science, despite offering fewer positions, are highly competitive, suggesting a need for advanced qualifications and expertise to succeed. Specialized fields such as Web3.0/Blockchain/Metaverse and Cyber Security, although having fewer job openings, present less competition, providing advantageous opportunities for individuals with specific skill sets. Job seekers should consider these dynamics to align their career strategies with market demands effectively.

8.2.3 Skills Analysis based on IT Tools / Platforms:

IT Jobs by Tools/Platforms:

The analysis of IT jobs by tools and platforms reveals that Java is the most in-demand skill, accounting for 23.89% of job openings. Python follows with 15.96%, and 'Not Specified' roles make up 14.43%. Unity is also significant with 14.19% of job openings. Excel, Rust, and C# also feature prominently, highlighting the diverse technical requirements across the industry. Other notable tools include PHP, Git, Swift, Typescript, Oracle, and C++, each contributing a smaller but still significant share of job openings. These findings emphasize the importance of proficiency in various programming languages, software tools, and platforms to meet the demands of the job market.



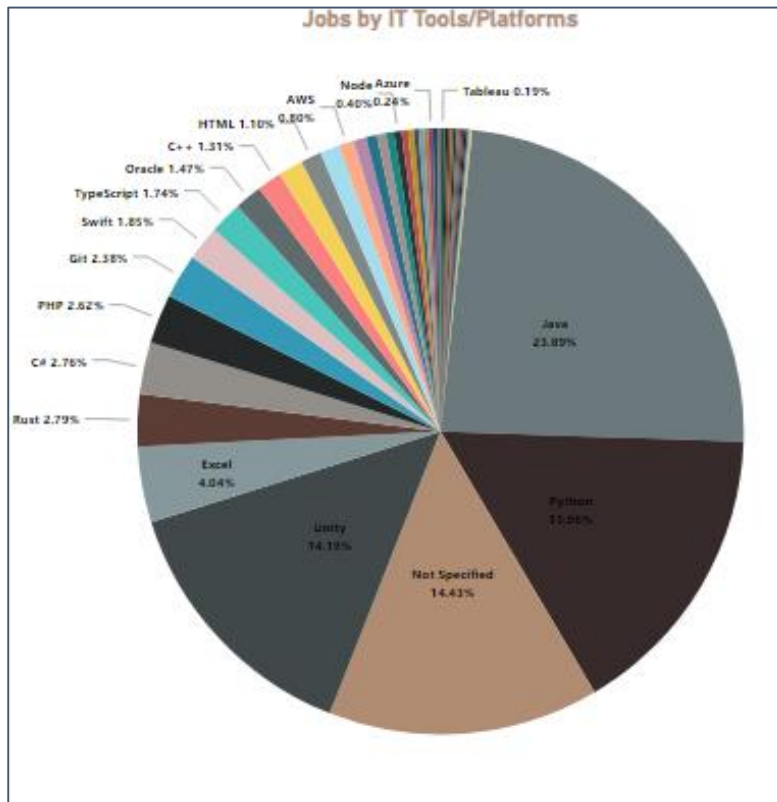
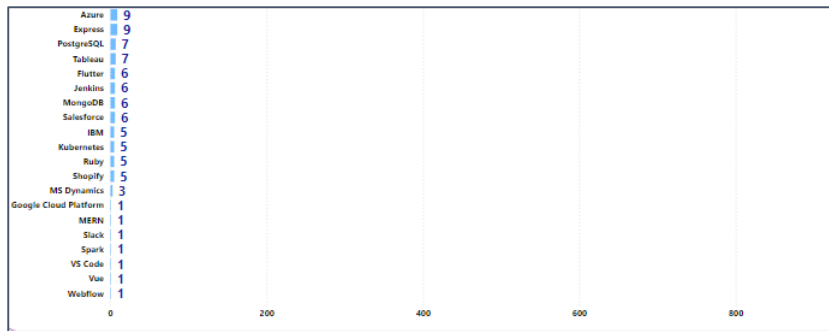


Figure 8.5: Distribution of Jobs across Tools/platforms

Applicants by Tools/Platforms (Competition Score: Number of Applicants Per Hour):

When evaluating the competition for jobs across these tools and platforms, Rust stands out with a high competition score of 11.26, indicating intense competition for positions requiring Rust skills. Power BI follows with a score of 3.99, and 'Not Specified' roles see a score of 2.92. Ruby and React also experience significant competition, with scores of 2.74 and 2.62, respectively. Tools like Excel and Python have moderate competition levels, with scores of 2.48 and 1.91. Unity, IBM, and Kotlin show lower but notable competition, with scores around 1.5 to 1.35.

Java, despite having the most job openings, sees a relatively low competition score of 0.95. Other tools such as AWS, HTML, and Swift have even lower competition, ranging between 0.87 and 0.81, indicating better opportunities for job seekers with these skills.

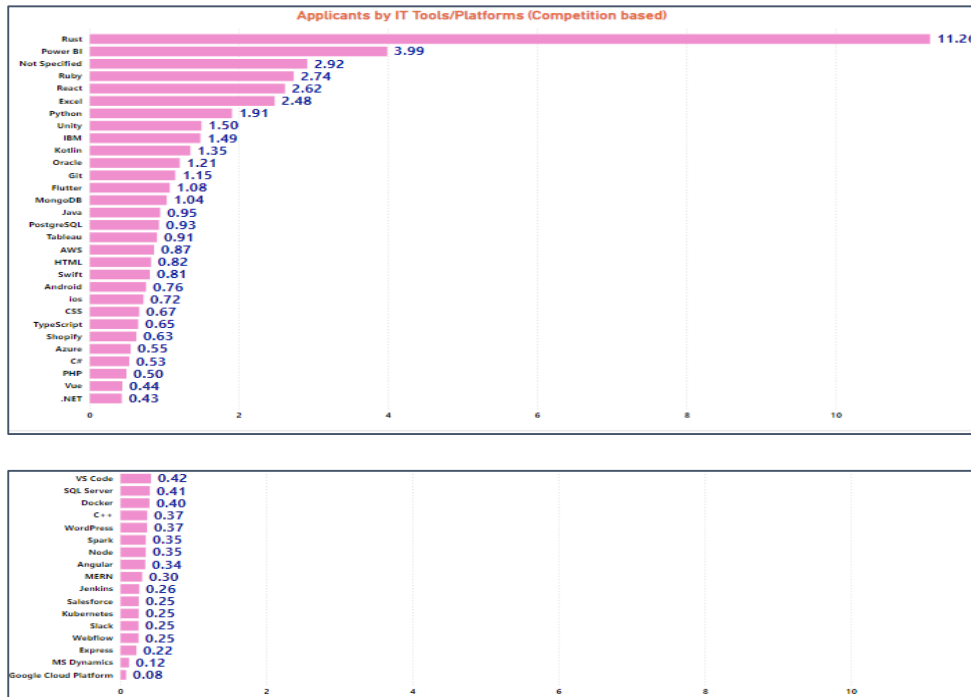


Figure 8.6: Competition by Tools/platforms

Combined Strategic Insights

The combined analysis of job availability and competition provides strategic insights for job seekers. Java, with the highest number of job openings and a low competition score, offers significant opportunities for those proficient in this language. Similarly, Python, Unity, and C# also present good prospects, balancing a substantial number of job openings with manageable competition. Conversely, while tools like Rust and Power BI have fewer job openings, they experience very high competition, suggesting that job seekers need to possess exceptional skills to stand out. Tools such as React and Ruby also require a competitive edge due to their higher applicant scores. On the other hand, tools like AWS, HTML, and Swift, with their lower competition scores, provide advantageous opportunities for individuals to secure positions more easily. By understanding these dynamics, job seekers can tailor their skill development and job search strategies to align with market demands and competition levels effectively.

Additionally, staying updated with emerging tools and technologies can further enhance job prospects in this competitive landscape.

8.3 Company Analysis:

After understanding which skills are in demand, the next crucial aspect to examine is which companies are at the forefront of the IT job market. This section delves into company analysis, identifying the top employers, their hiring trends, and the competitive landscape.

IT Jobs by Companies

In the analysis of IT jobs among the top 20 companies, Crossover leads with 506 job openings, representing 13.55% of the total among these companies. Turing follows with 215 jobs (5.76%), and Afiniti holds a smaller share at 74 jobs (1.98%). Other notable companies include HR Ways - Hiring Tech Talent with 39 jobs (1.04%), Dubizzle Labs with 35 jobs (0.94%), Zones IT Solutions with 31 jobs (0.83%), and Motive also with 31 jobs (0.83%). Companies like TCP Software, Gelato, CureMD, and Teradata also have significant job offerings, each contributing 27 jobs (0.72%) to the job market. These top 20 companies account for only 1.4% of the total 1420 companies but hold a significant 32.8% of the total jobs, highlighting their dominant presence in the job market. The remaining 67.2% of jobs are distributed among the other companies.

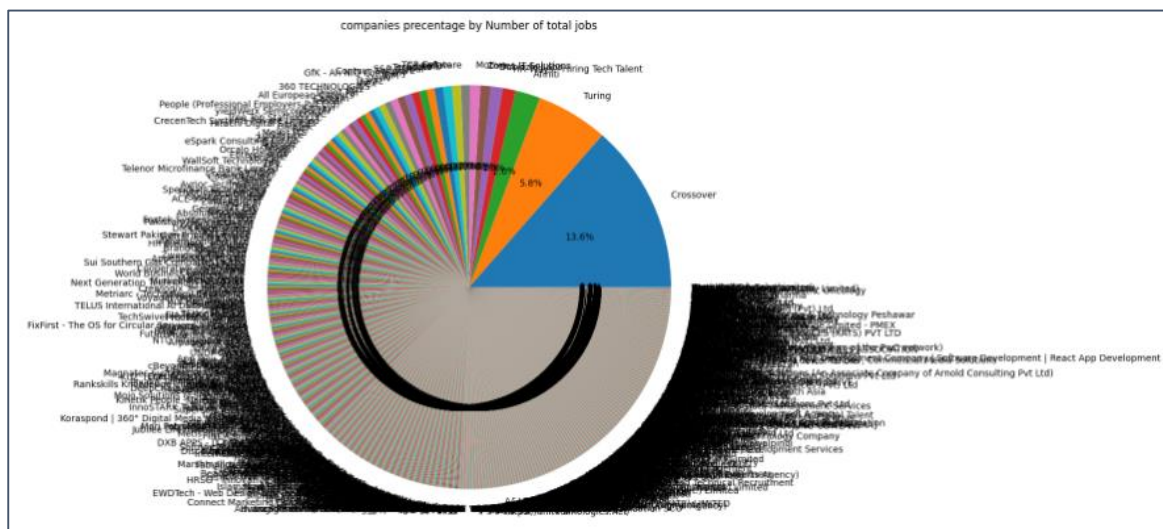
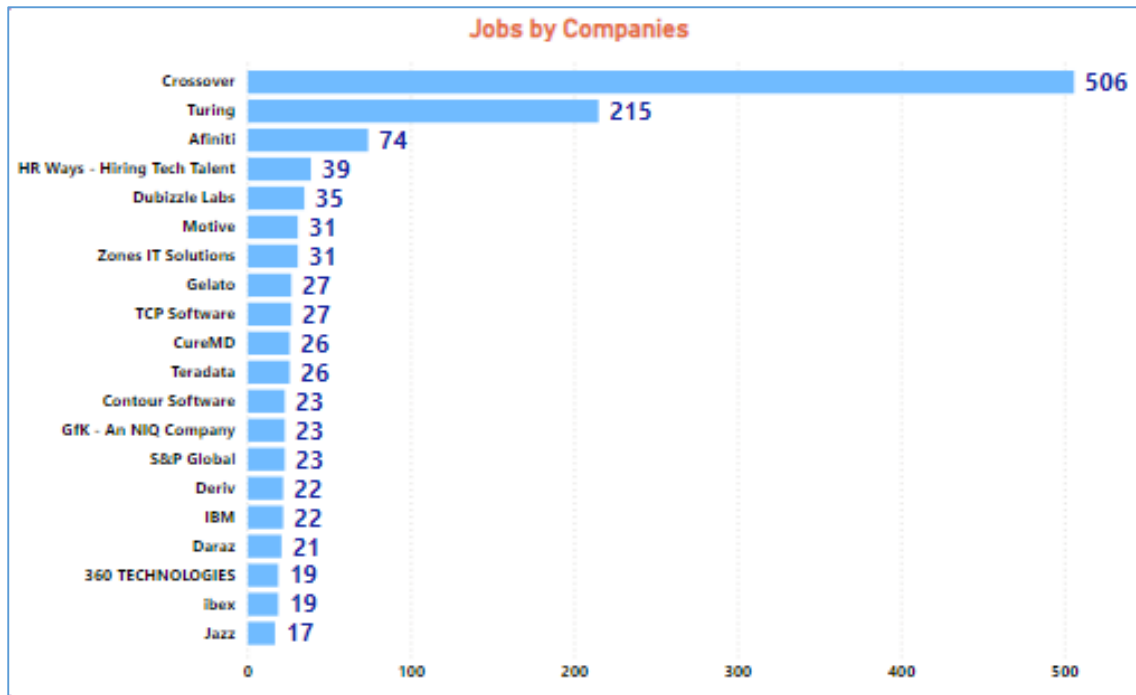


Figure 8.7: Distribution of Jobs across Top 20 companies

Applicants by Companies (Competition Score: Number of Applicants Per Hour):

The competition for jobs across these top 20 companies shows significant variation. Motive experiences the highest competition score with 32.34 applicants per hour, indicating intense competition for positions at this company. Zones IT Solutions follows with a score of 8.78, and GfK - An NIQ Company sees a score of 3.47. TCP Software and Daraz also experience notable competition, with scores of 3.41 and 3.12, respectively. Crossover, despite having the most job

openings, has a moderate competition score of 2.84, suggesting it remains a popular choice among applicants. CureMD, Dubizzle Labs, and ibex show varying levels of competition, highlighting the competitive landscape in the job market across these companies.

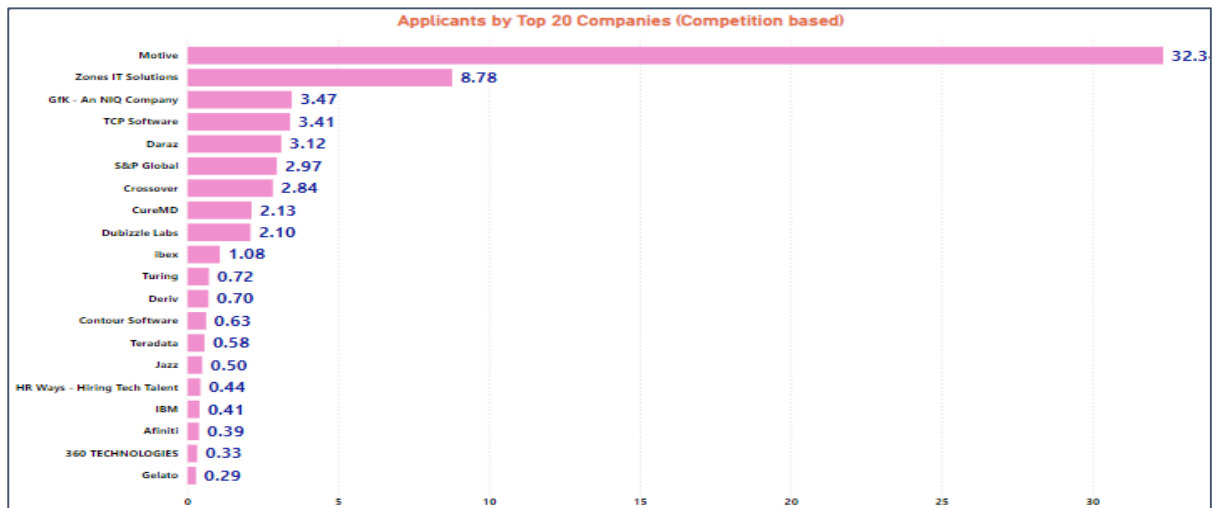


Figure 8.8: Competition by Top twenty companies

Overall Competition in the Job Market:

The broader job market also presents significant competition. WebFX tops the list with a staggering 90.71 applicants per hour, followed by Tree Top Staffing LLC with 58.88, and Motive with 32.34. Other high-competition companies include BrainCX (25.85), UN Women (22.62), and Goodwork (22.15). While Motive appears both in the top 20 companies and high-competition lists, others like Zones IT Solutions also show significant competition with 8.78 applicants per hour. Understanding these competition dynamics across all companies, not just the top 20, can provide job seekers with a comprehensive view of the market, helping them make informed decisions about where to apply based on both job availability and competitive pressure.

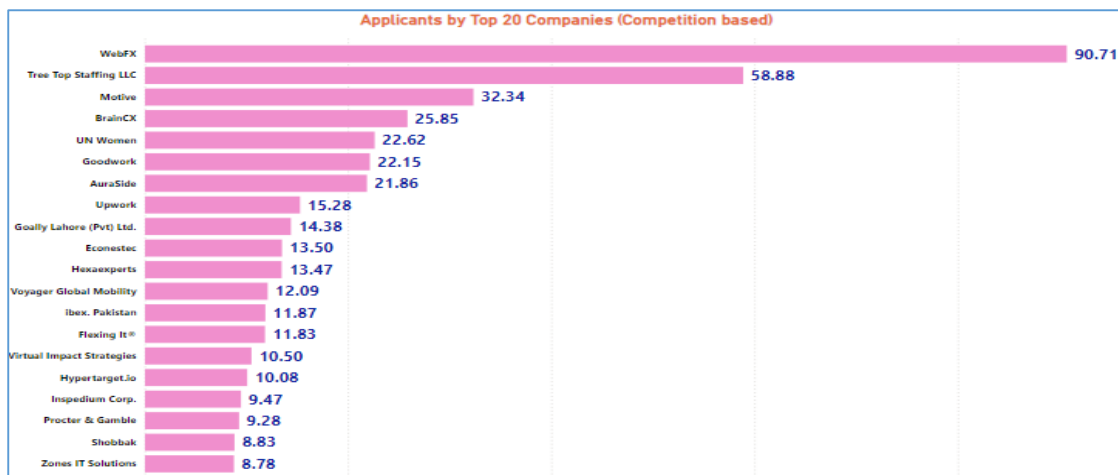


Figure 8.9: Top twenty companies by competition score

Combined Strategic Insights:

Combining job availability and competition data provides strategic insights for job seekers. Crossover, with 506 jobs and a moderate competition score of 2.84, offers significant opportunities for applicants. Similarly, Turing (215 jobs) and Afiniti (74 jobs) present reasonable prospects due to their balanced job openings and lower competition scores. In contrast, Motive, despite having fewer job openings (31), faces very high competition with 32.34 applicants per hour, suggesting that candidates need exceptional skills and qualifications to succeed there. Zones IT Solutions (31 jobs) and GfK - An NIQ Company (23 jobs) also show high competition, emphasizing the need for job seekers to stand out. Companies like TCP Software (27 jobs) and Daraz (21 jobs), with moderate competition scores, provide viable opportunities for candidates with relevant skills. These top 20 companies, holding 32.8% of the total jobs, play a crucial role in the job market, and understanding these dynamics helps job seekers prioritize their applications and focus on companies that align with their skills and competitive advantage, ultimately enhancing their chances of securing employment.

8.4 Geography Analysis:

Understanding the geographical distribution of IT job opportunities is crucial for job seekers and policymakers alike. This section addresses important questions about which cities, provinces, and regions, including foreign countries, post the most jobs in Pakistan. We can identify the top cities and provincial hubs for IT jobs by analyzing the geographic data.

Additionally, this analysis includes an assessment of job competition across these regions, providing insights into the competitive landscape.

8.4.1 Geography Analysis based on City:

Jobs by Top 20 Cities:

In the analysis of IT job distribution among the top 20 cities, Lahore leads with 1,198 job openings, representing 32.08% of the total jobs in these cities. Karachi follows with 858 jobs (22.98%), and Islamabad holds 720 jobs (19.28%). Other notable cities include Pakistan as a general location (these are not city-specific) jobs with 486 jobs (13.02%), Rawalpindi with 187 jobs (5.01%), and Faisalabad with 57 jobs (1.53%). Smaller cities such as Peshawar (46 jobs, 1.23%), Punjab (All cities of Panjab) (35 jobs, 0.94%), Multan (23 jobs, 0.62%), and Hyderabad (21 jobs, 0.56%) also contribute to the job market. These top 20 cities account for a significant portion of IT jobs, emphasizing their importance as hubs for employment opportunities in the IT sector.

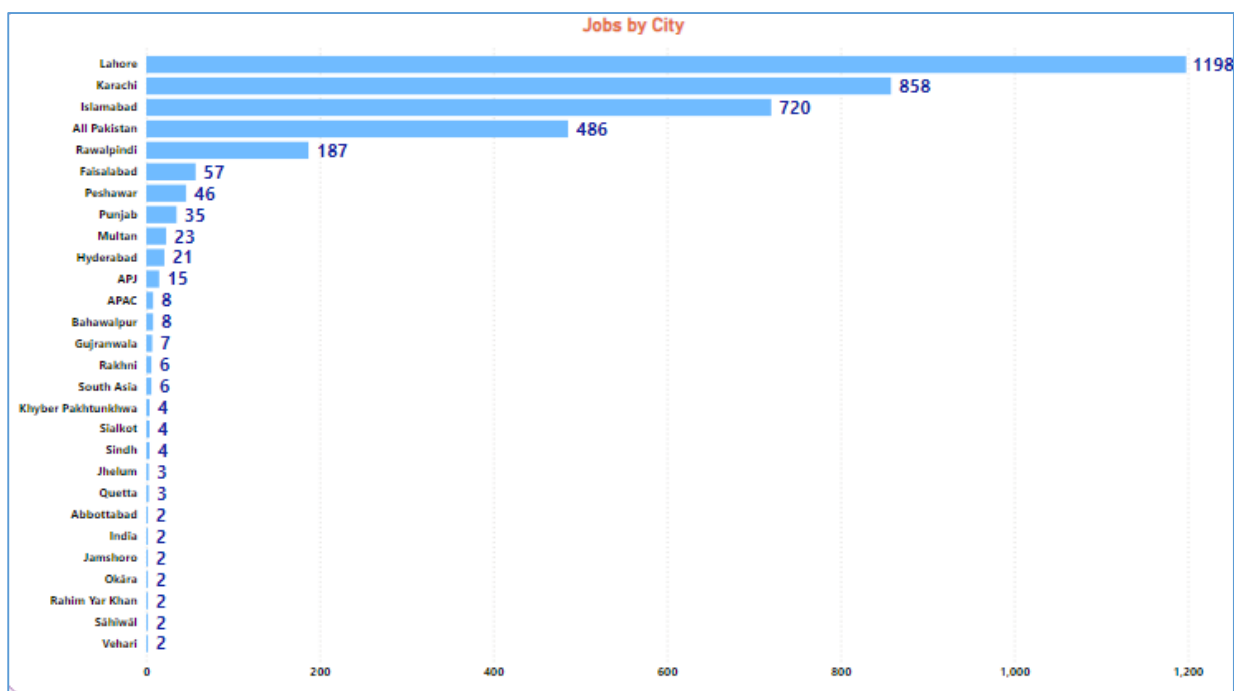
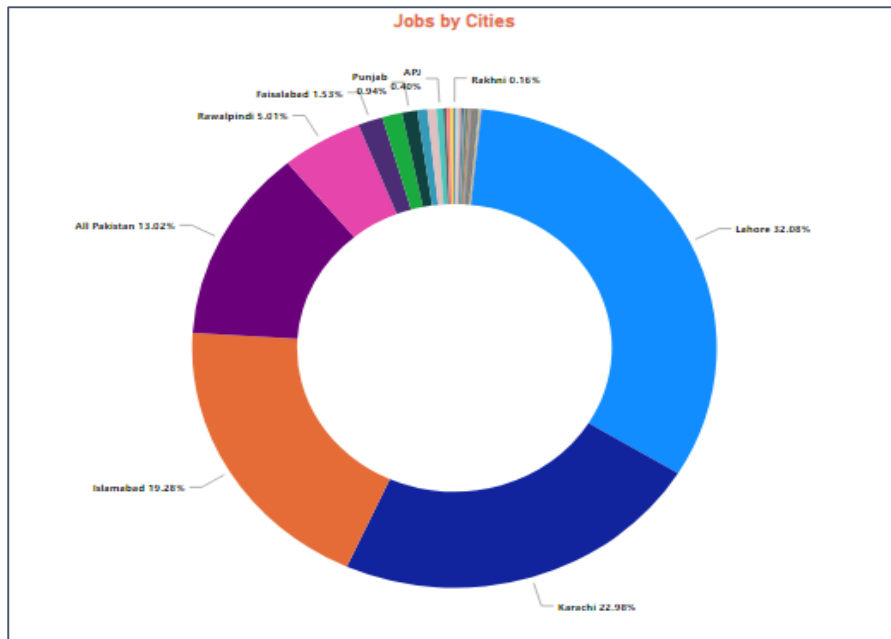


Figure 4.10: Distribution of Jobs across top twenty cities



Applicants by Cities (Competition Score: Number of Applicants Per Hour)

The competition for IT jobs varies significantly across the top 20 cities. Hyderabad experiences the highest competition score with 4.23 applicants per hour, indicating intense competition for positions in this city. The general location "All Pakistan" follows with a score of 3.51, and Rawalpindi has a score of 1.75. Other notable cities include Islamabad with a score of 1.62, Karachi with 1.25, and Lahore with 1.08. Cities like Jamshoro (0.67), Okara (0.56), Jhelum (0.55), and Multan (0.38) show varying levels of competition, highlighting the competitive landscape for job seekers across different locations.

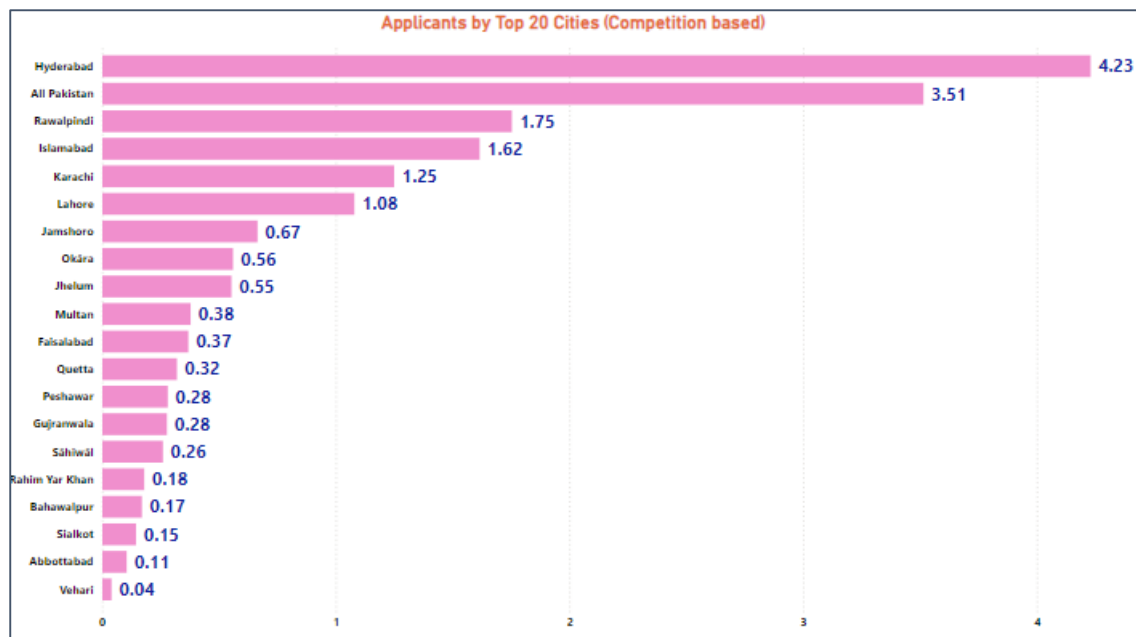


Figure 8.11: Competition by Top twenty companies

Combined Strategic Insights:

Combining job availability and competition data provides strategic insights for job seekers. Lahore, with 1,198 jobs and a moderate competition score of 1.08, offers significant opportunities for applicants. Similarly, Karachi (858 jobs) and Islamabad (720 jobs) present substantial prospects due to their large number of job openings and moderate competition scores. In contrast, Hyderabad, despite having fewer job openings (21), faces very high competition with 4.23 applicants per hour, suggesting that candidates need exceptional skills and qualifications to succeed there. The general location "All Pakistan" also shows high competition with a score of 3.51, emphasizing the need for job seekers to stand out in these areas.

It is noteworthy that the top 10 cities in Pakistan, out of a total of 49 cities, hold a staggering 97.2% of the total job openings, leaving the remaining 39 cities with just 2.8% of the jobs. This surprising disparity underscores the concentration of IT job opportunities in major urban centers, making them critical areas for job seekers to target.

8.4.2 Geography Analysis based on Regions or Provinces:

IT Jobs by Regions/Provinces

In the analysis of IT job distribution across various regions and provinces in Pakistan, the top regions emerge as follows: Punjab leads with 1,546 job openings, constituting 41.82% of the total. Sindh follows with 884 jobs (23.91%), and Islamabad Capital Territory has 715 jobs (19.34%). Pakistan, as a general location, accounts for 486 jobs (13.15%). Khyber Pakhtunkhwa offers 54 jobs (1.46%), while Baluchistan has 11 jobs (0.30%), and Gilgit-Baltistan offers 1 job (0.03%). These regions collectively represent the primary hubs for IT job opportunities in Pakistan.

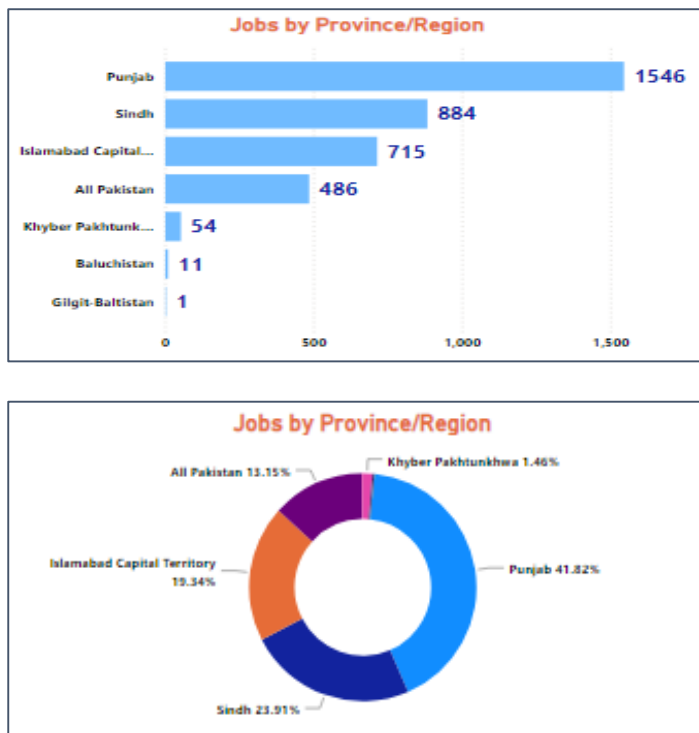


Figure 8.12: Distribution of Jobs across Provinces in Pakistan

Applicants by Regions/Provinces (Competition Score: Number of Applicants Per Hour):

The competition for IT jobs varies significantly across these regions: Baluchistan experiences the highest competition score with 5.71 applicants per hour, followed by Pakistan with a score of 3.51. Islamabad Capital Territory has a score of 1.62, Sindh and Punjab have scores of 1.32 and 1.09, respectively. Khyber Pakhtunkhwa faces relatively lower competition with a score of 0.30, while Gilgit-Baltistan has a competition score of 0.20. These scores underscore the competitive landscape for job seekers in different regions, with some areas presenting notably higher challenges than others.

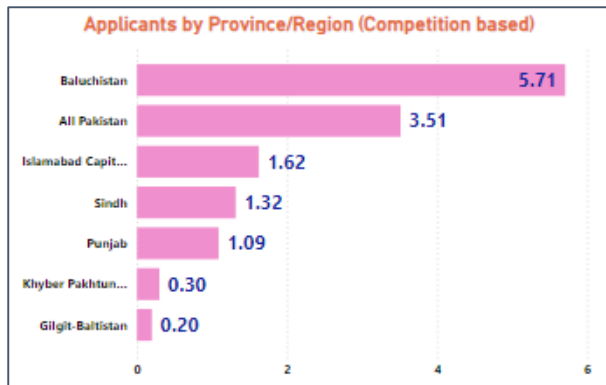


Figure 8.13: Competition by Provinces in Pakistan

Combined Strategic Insights:

Combining job availability and competition data provides valuable strategic insights for job seekers. Punjab, with 1,546 jobs and a moderate competition score of 1.09, offers significant opportunities for applicants. Similarly, Sindh (884 jobs) and Islamabad Capital Territory (715 jobs) present substantial prospects due to their large number of job openings and moderate competition scores. In contrast, Baluchistan, despite having only 11 job openings, faces extremely high competition with a score of 5.71, suggesting that candidates need exceptional skills and qualifications to succeed there. Understanding these dynamics helps job seekers prioritize their applications and focus on regions that align with their skills and competitive advantage, ultimately enhancing their chances of securing employment.

8.4.3 Geography Analysis based on Country:

IT Jobs by Country:

In the analysis of IT job distribution across countries, Pakistan leads significantly with 3,697 job openings, representing 99.01% of the total among these countries. Other regions such as Asia-Pacific Japan (APJ), Asia-Pacific (APAC), Middle East and North Africa (MENA), South Asia, India, Uganda, Turkey, and Austria have relatively fewer job openings. APJ, representing a regional area, has the highest count after Pakistan, with 15 job openings (0.40%), followed by APAC with 8 jobs (0.21%). South Asia, India, MENA, Uganda, Turkey, and Austria each contribute minimally, with job counts ranging from 1 to 6.



Figure 8.14: Distribution of Jobs across Countries

Applicants by Country (Competition Score: Number of Applicants Per Hour):

The competition for IT jobs varies significantly across these countries. APJ, a regional area, experiences the highest competition score with 55.29 applicants per hour, indicating intense competition for positions in this region. South Asia follows with a score of 6.93, and APAC has a score of 3.40. Pakistan, despite having the most job openings, still faces notable competition with a score of 1.57. India, Turkey, Austria, MENA, and Uganda also experience varying levels of competition, emphasizing the competitive landscape in the job market across these regions.

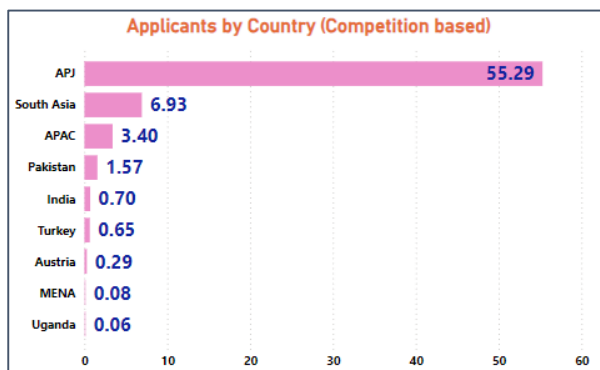


Figure 8.15: Competition by Countries

Combined Strategic Insights:

Combining the job availability and competition data provides strategic insights for job seekers. Pakistan, with its significant number of job openings and moderate competition score, offers substantial opportunities for applicants. APJ, despite representing a regional area and having fewer job openings, faces extremely high competition, suggesting that candidates need exceptional skills and qualifications to succeed there. Other regions like APAC, South Asia,

India, Turkey, Austria, MENA, and Uganda also present varying levels of competition, highlighting the importance for job seekers to understand regional dynamics and tailor their applications accordingly to maximize their chances of securing employment.

8.5 Job Dynamics Analysis:

Job dynamics analysis is divided into three parts: Work Arrangement Analysis, Type of Employment Analysis, and Experience Level Analysis. The major questions addressed in this section include which types of jobs are in higher demand, such as Full-Time vs. Part-Time and Remote vs. Onsite positions. Additionally, the analysis examines the competition within these job categories, providing a comprehensive overview of the employment landscape.

8.5.1 Job Dynamics Analysis based on Work Arrangement:

IT Jobs by Work Arrangement(Remote/Onsite):

In the analysis of IT job distribution based on work arrangement, on-site positions lead significantly with 2,184 job openings, constituting 58.49% of the total. Remote positions follow closely with 1,179 jobs (31.57%), while hybrid roles account for 273 jobs (7.31%). A smaller portion of job postings, 98 in total, does not specify the work arrangement.

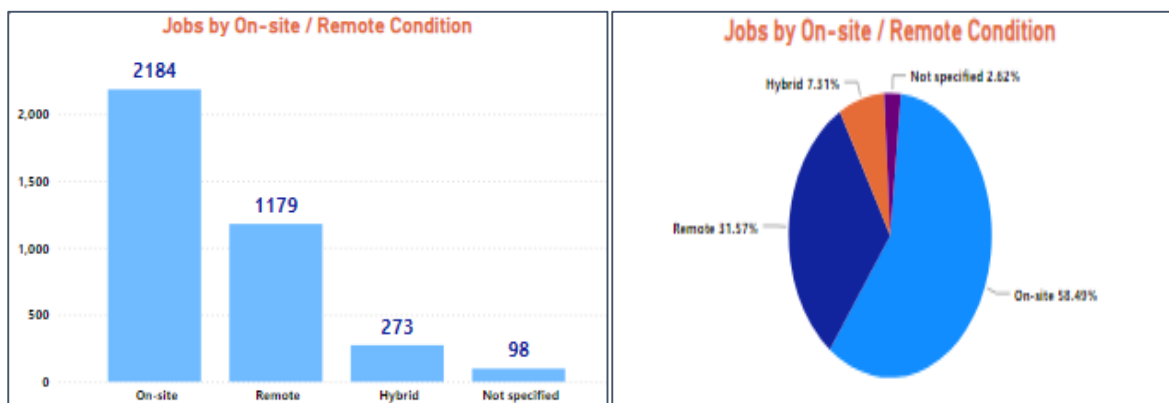


Figure 8.16: Distribution of Jobs across Work arrangement (Remote/Onsite)

Applicants by Work Arrangement (Competition Score: Number of Applicants Per Hour):

The competition for IT jobs varies across different work arrangements. Remote positions experience the highest competition score with 3.58 applicants per hour, indicating intense competition for these roles. Not specified work arrangements follow with a score of 2.63,

suggesting significant interest despite the lack of clarity. Hybrid roles have a competition score of 1.08, while on-site positions have a relatively lower competition score of 0.88 despite having the most job openings.

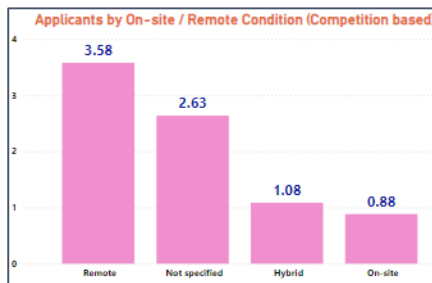


Figure 8.17: Competition across work arrangement

Combined Strategic Insights:

Combining job availability and competition data provides strategic insights for job seekers. Despite being the most abundant, on-site positions present a relatively lower level of competition, making them attractive options for applicants seeking traditional work environments. Remote roles, although highly sought-after, face intense competition, highlighting the need for candidates to showcase exceptional skills and qualifications. Hybrid roles offer a balanced mix of remote and on-site work, presenting viable opportunities with moderate competition. Understanding these dynamics can help job seekers tailor their applications to align with their preferred work arrangements and maximize their chances of securing employment.

8.5.2 Job Dynamics Analysis based on Type of Employment:

IT Jobs by Type of Employment:

In the analysis of IT job distribution based on the type of employment, full-time positions dominate significantly with 3,577 job openings, constituting 95.80% of the total. Contract roles follow with a much smaller share of 88 jobs (2.36%), while part-time positions account for 35 jobs (0.94%). Internship opportunities are slightly less common, with 31 postings (0.83%), and only 2 job postings do not specify the type of employment. Temporary roles represent the smallest portion, with just 1 job posting.

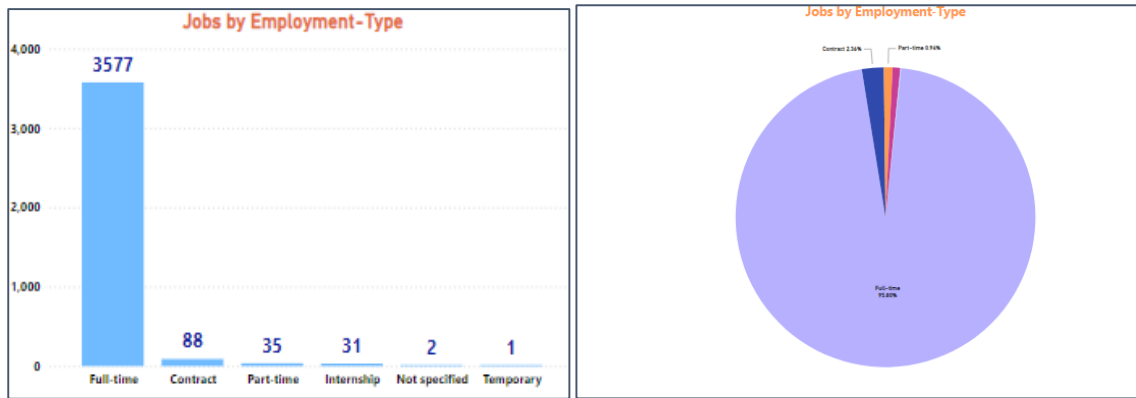


Figure 8.18: Distribution of Jobs across Type of Employment

Applicants by Type of Employment (Competition Score: Number of Applicants Per Hour):

The competition for IT jobs varies across different types of employment. Full-time positions experience a moderate competition score of 1.83 applicants per hour, indicating reasonable demand for these roles. Contract roles have a slightly lower competition score of 1.15, while part-time positions face a competition score of 0.69. Internship opportunities also have a moderate competition score of 0.56. Job postings that do not specify the type of employment face minimal competition with a score of 0.18, and temporary roles have a score of 0.15.

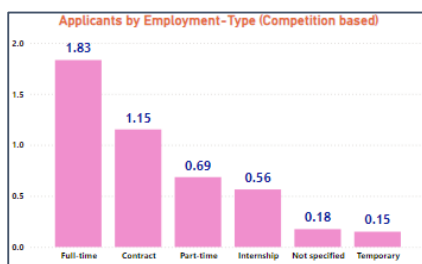


Figure 8.19: Competition by Type of Employment

Combined Strategic Insights:

Combining job availability and competition data provides strategic insights for job seekers. Full-time positions, being the most abundant, offer substantial opportunities with moderate competition, making them attractive options for applicants seeking stable employment. Contract roles, although less common, also present viable opportunities with reasonable competition. Part-time positions offer a more flexible work arrangement, attracting candidates seeking reduced hours, while internships provide valuable learning experiences with moderate

competition. Understanding these dynamics can help job seekers tailor their applications to align with their preferred types of employment and maximize their chances of securing employment jobs.

8.5.3 Job Dynamics Analysis Based on Experience Level

IT Jobs by Experience Level:

In the analysis of IT job distribution based on experience level, roles with unspecified experience requirements dominate significantly, with 1,919 job openings, constituting 51.39% of the total. Mid-senior level positions follow with 1,089 jobs (29.16%), while entry-level positions account for 443 jobs (11.86%). Associate roles represent a smaller portion, with 160 postings (4.28%), followed by director-level positions with 61 jobs (1.63%). Internship opportunities are slightly less common, with 33 postings (0.88%), and executive-level roles represent the smallest portion, with 29 job postings (0.78%).

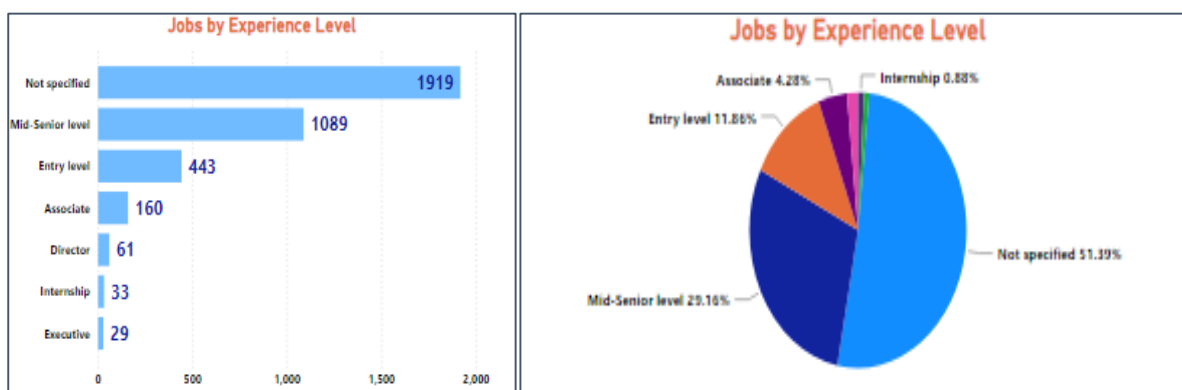


Figure 8.20: Distribution of Jobs across Experience Levels

Applicants by Experience Level (Competition Score: Number of Applicants Per Hour):

The competition for IT jobs varies across different experience levels. Director-level positions experience the highest competition score with 11.05 applicants per hour, indicating intense competition for these roles. Mid-senior level positions follow with a competition score of 2.45, while associate roles have a score of 1.76. Entry-level positions face a competition score of 1.62, and job postings with unspecified experience requirements have a score of 1.21. Internship opportunities also have moderate competition with a score of 0.68, while executive-level roles have a lower competition score of 0.42.

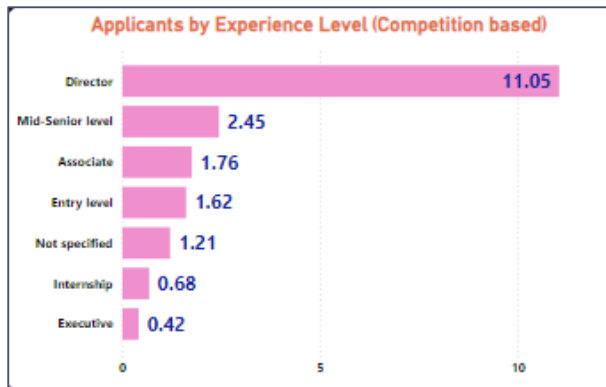


Figure 8.21: Competition by Experience level

Combined Strategic Insights:

Combining job availability and competition data provides strategic insights for job seekers. Roles with unspecified experience requirements offer substantial opportunities with moderate competition, making them attractive options for applicants with diverse backgrounds. Mid-senior level positions, although abundant, face relatively higher competition, suggesting the need for applicants to showcase relevant experience and skills. Entry-level positions offer opportunities for candidates starting their careers, while associate and director-level roles provide growth opportunities for experienced professionals. Internship opportunities offer valuable learning experiences with moderate competition, while executive-level roles represent more specialized positions with lower competition. Understanding these dynamics can help job seekers tailor their applications to align with their experience levels and maximize their chances of securing employment.

8.6 Industry Preference Analysis:

Industry preference analysis is divided into two main parts: Company Employee Size and Company Industry. This section addresses the key question of which industries provide more IT jobs. By examining both the size of companies and the specific industries they operate in, we can identify where the majority of IT job opportunities are concentrated.

8.6.1 Industry Preference Analysis Based on Company Employee Size:

IT Jobs by Company Employee Size:

In the analysis of IT job distribution based on company employee size, various trends emerge. Companies with 11-50 employees offer the highest number of job openings, totaling 875 positions, indicating a substantial demand for IT professionals in small to medium-sized enterprises. This is followed by companies with 51-200 employees, contributing 625 job openings, highlighting the robust recruitment activities within larger corporations. The mid-sized sector, comprising companies with 1,001-5,000 employees, also presents significant opportunities, with 608 job openings, reflecting a balanced demand across different scales of organizations.

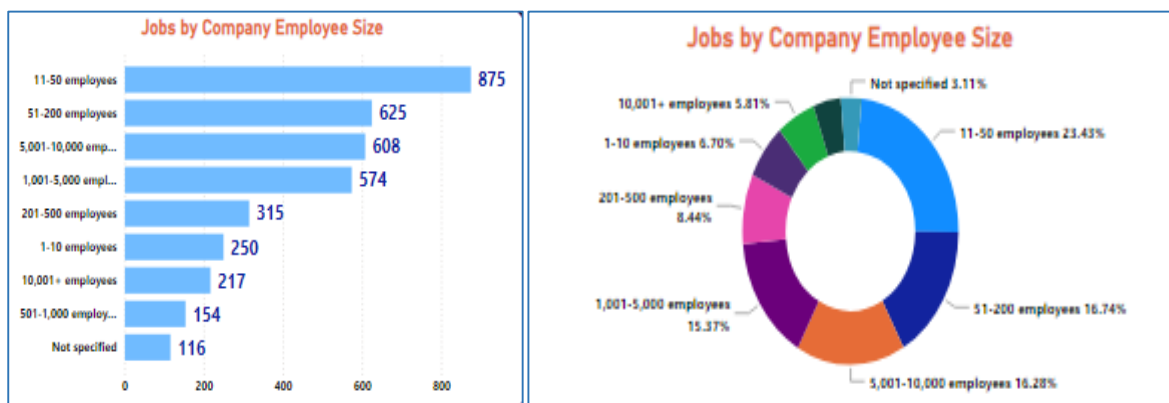


Figure 8.22: Distribution of Jobs across Company Employee Size

Applicants by Company Employee Size (Competition Score: Number of Applicants Per Hour):

The competition for IT jobs varies across different company employee sizes. Companies with 1-10 employees experience the highest competition score with 4.02 applicants per hour, indicating intense competition for positions in small companies. Companies with 501-1,000 employees follow closely with a score of 3.61, while companies with 1,001-5,000 employees have a score of 2.71. Larger companies with 5,001-10,000 employees and 10,001+ employees face competition scores of 2.65 and 2.03, respectively. Companies with 201-500 employees, 51-200 employees, and 11-50 employees also experience varying levels of competition.

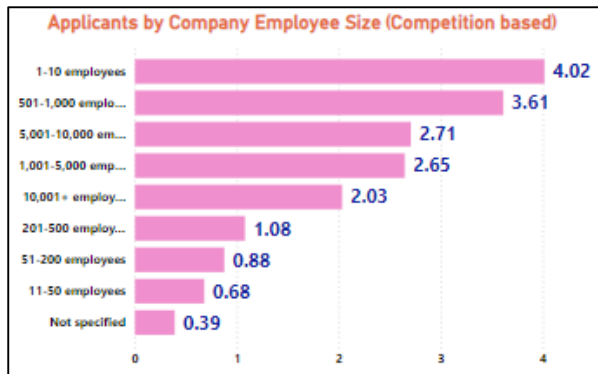


Figure 8.23: Competition by Company Employee Size

Combined Strategic Insights:

Combining job availability and competition data provides strategic insights for job seekers. Positions in small companies (1-10 employees) offer numerous opportunities but face intense competition, suggesting the need for candidates to stand out. Conversely, positions in larger companies (501-1,000 employees, 1,001-5,000 employees, 5,001-10,000 employees, and 10,001+ employees) offer competitive opportunities with moderate competition. Understanding these dynamics can help job seekers tailor their applications to align with their preferences regarding company employee size and maximize their chances of securing employment.

8.6.2 Industry Preference Analysis Based on Company Employee Size:

IT jobs by Top 20 Industries:

Among the top twenty industries out of 62 analyzed, the category "Not specified" leads with 1,802 job openings, comprising approximately 48.26% of the total among these industries. This broad category indicates a significant portion of job postings lacking specific industry categorization. Following closely, "IT Services and IT Consulting" represent a substantial segment, with 1,072 job openings, accounting for around 28.71% of the total. "Software Development" also emerges as a prominent category, with 307 job openings, making up approximately 8.22% of the total. These findings suggest a diverse landscape of job opportunities across various sectors within the IT industry.

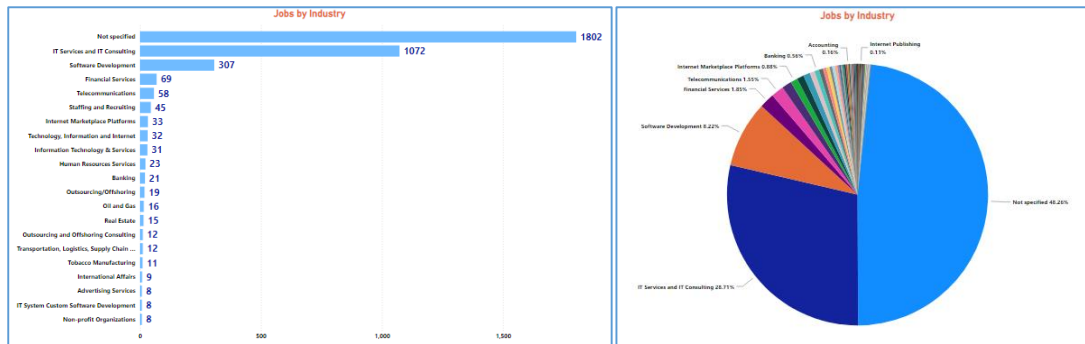


Figure 8.24: Distribution of Jobs across Industries

Applicants' Interest Across Industries(Competition Score: Number of Applicants Per Hour):

The competition for IT jobs varies significantly across different industries. Industries such as "Staffing and Recruiting" experience the highest competition score with 19.07 applicants per hour, indicating intense competition for positions in this sector. Similarly, "Strategic Management Services" and "Advertising Services" also face considerable competition, with scores of 12.09 and 11.85, respectively. Conversely, industries like "IT Services and IT Consulting" and "Software Development" also have notable competition, with scores of 2.19 and 4.82, respectively. These scores shed light on the competitive landscape for job seekers across different industries within the IT sector.

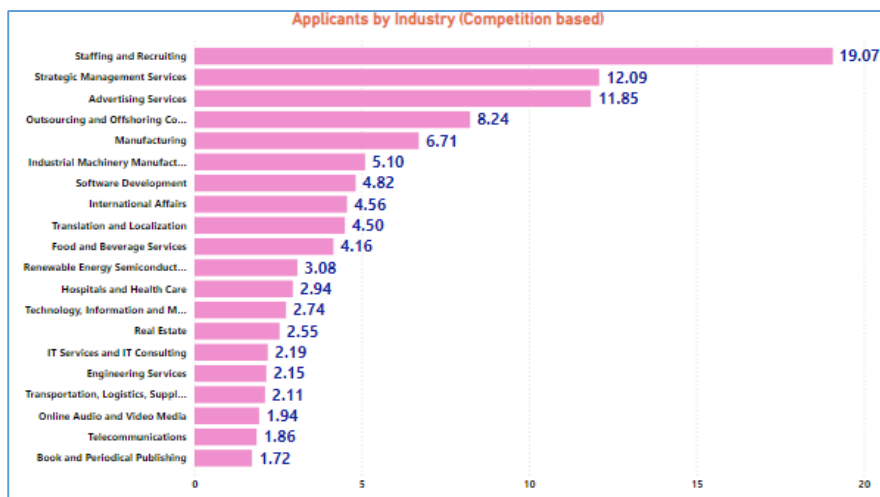


Figure 8.25: Competition by Industries

Strategic Insights and Recommendations:

Understanding the distribution of job openings and the level of competition across various industries provides valuable strategic insights for both job seekers and employers. Industries with high job openings but low competition present attractive opportunities for job seekers to explore. Conversely, industries with intense competition may require candidates to demonstrate exceptional skills and qualifications to stand out. Employers can leverage these insights to refine their recruitment strategies and attract top talent. Overall, understanding these dynamics allows stakeholders to make informed decisions and optimize their approaches to navigating the IT job market effectively.

8.7 Analysis of Education/Degree Specification:

This section analyzes the education requirements specified in job descriptions. Jobs are marked as 'Yes' if any education-related keywords are found in the job description. The job is marked as 'No' if no education-related keywords are present. Jobs without mentioned educational requirements are considered to be purely experience-based. This analysis helps in understanding the importance of educational qualifications in the IT job market and highlights the roles where practical experience is prioritized over formal education.

IT Jobs by Educational/Degree Requirements:

In the analysis of IT job openings based on educational requirements, the data reveals that most positions require specific academic qualifications. Of the total job postings, 2,596 (approximately 69.73%) specify an educational requirement. In contrast, 1,127 job openings (around 30.27%) do not mandate any particular academic qualifications. This indicates that while there is a substantial demand for formally educated candidates, many opportunities remain accessible to individuals without strict educational prerequisites.

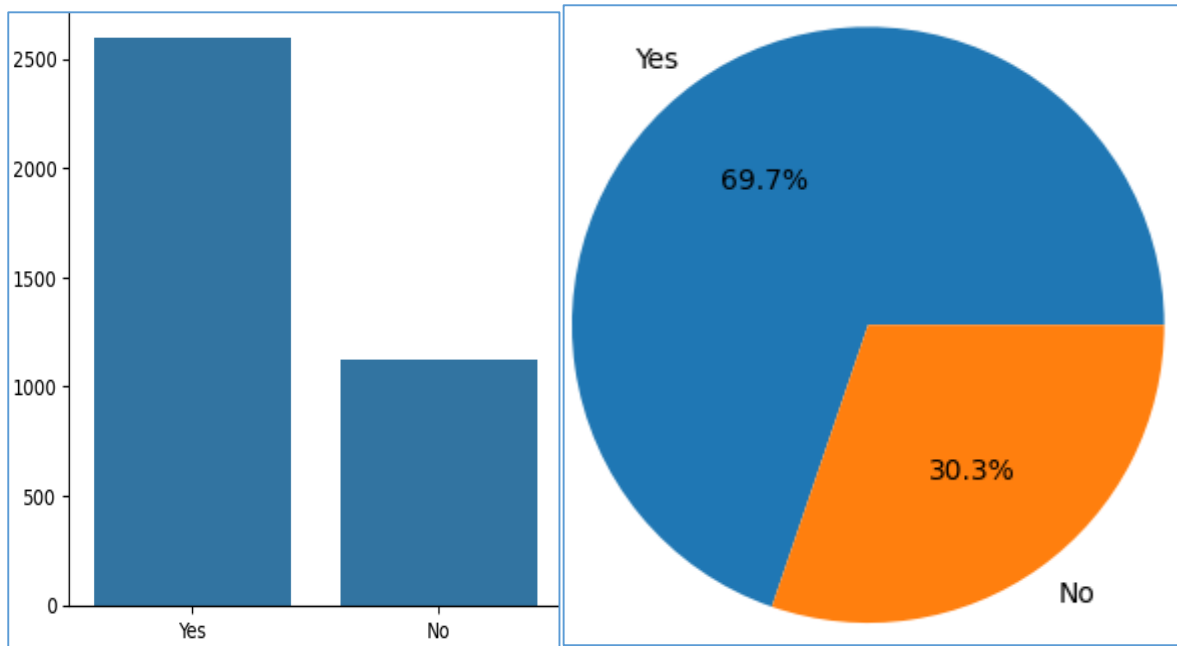


Figure 8.26: Distribution of Jobs across Education/Degree Status

Applicants by Educational Requirements (Competition Score: Number of Applicants Per Hour):

The competition for IT jobs also varies based on whether educational qualifications are specified. Jobs that do not require specific educational credentials experience a higher competition score, with 2.07 applicants per hour, indicating a more competitive environment. On the other hand, positions that do require educational qualifications have a slightly lower competition score of 1.67 applicants per hour.

This suggests that while there is a high demand for jobs without educational prerequisites, the requirement of educational credentials slightly moderates the competition for those roles.

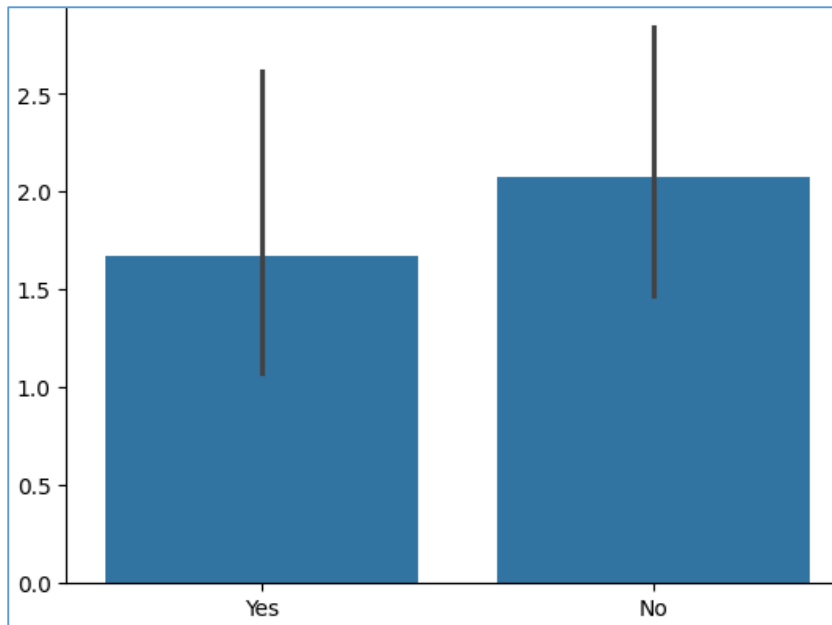


Figure 8.27: Competition of Jobs across Education/Degree Status

Strategic Insights and Recommendations:

Combining the job availability and competition data provides strategic insights for job seekers and employers alike. For job seekers, understanding that positions without specific educational requirements are more competitive may encourage them to enhance their qualifications to gain an edge. Conversely, candidates with relevant educational credentials might find slightly less competition for roles that match their qualifications. Employers can use these insights to tailor their recruitment strategies, either by emphasizing the educational requirements to filter candidates or by considering a wider range of applicants for roles that do not necessitate specific qualifications. Overall, these dynamics highlight the importance of educational credentials in the IT job market and their impact on job competition.

REFERENCES

References of Research papers, articles and books:

- [1] M. Bilal, N. Malik, M. Khalid and M. I. Lali, "Exploring industrial demand trend's in Pakistan software industry using online job portal data.," *University of Sindh Journal of Information and Communication Technology*, pp. 17-24, 2017.
- [2] Norihiko Matsuda, Tutan Ahmed, Shinsaku Nomura, "Labor market analysis using big data: The case of a Pakistani online job portal.," *World Bank Policy Research Working Paper*, (9063), p. 42, 2019.
- [3] N. F. S. N. a. B. A. Bilal Raza, "Skills Set Required for Web Developers in Pakistan.," *Pakistan Journal of Engineering and Technology*, 6(1), pp. 86-91, 2023.
- [4] B. S. a. A. N. Ghulam Muhammad Kundi, "Digital Pakistan: opportunities & challenges.," *Jistem Journal of Information Systems and Technology Management*, vol. 5, pp. 365-390, 2008.
- [5] "LinkedIn's Economic Graph," [Online]. Available: <https://economicgraph.linkedin.com/>.
- [6] "Data for Impact ,A Partnership for Economic Opportunity," [Online]. Available: <https://economicgraph.linkedin.com/data-for-impact>.
- [7] A. t. Analyst. [Online]. Available: <https://github.com/AlexTheAnalyst/PythonYouTubeSeries/>.
- [8] [Online]. Available: <https://www.scrapingbee.com/blog/selenium-python/>.
- [9] [Online]. Available: <https://www.scrapingdog.com/blog/scrape-linkedin-jobs/>.
- [10] Z. u. H. usmani, Data analytics and data visulization.