Project Report
on
Job Description Analysis Using Big Data Tools

TEAM MEMBERS:

- Dhanalakshmi Kannur Munirathnam [kannu1d]
- Vyom Patel [patel5v]
- Prathyusha Davanapalli [davan1p]

ITC686 Big Data Analytics 22453873

May2nd, Friday

Table of Contents

1. Introduction
2. Problem Statement and Motivation
3. Objectives
4. Dataset Description
5. Tools and Technologies Used
6. Data Preprocessing
7. Exploratory Data Analysis (EDA)
8. Hypotheses
9. Analytical Queries
10. Predictive Modeling
11. Evaluation and Findings
12. Challenges Faced
13. Conclusion
14. Future Scope
15. References
16. Appendix

1. Introduction

This project explores **large-scale job description data** to uncover patterns in hiring trends, top skills, qualifications, and salary expectations across different geographies and companies. The dataset, sized at approximately **1.5 GB**, includes over **13,000 job postings** with 23 fields such as skills, experience, salary, and job role.

As students preparing to enter the job market, we selected this topic to better understand current industry expectations and the most in-demand skills and qualifications. Analyzing real-world data allows us to gain actionable insights into what employers are seeking in various sectors and regions.

To handle this volume of semi-structured data, we used **Apache Spark with PySpark** on **Google Colab**, enabling scalable data preprocessing, transformation, and analysis. Spark's distributed framework ensures performance and efficiency when working with large datasets.

This project not only strengthened our understanding of big data tools and practices but also provided us with meaningful knowledge of the job market landscape. It bridges the gap between academic learning and practical workforce readiness.

2. Problem Statement and Motivation

The labor market is evolving rapidly with technological shifts and skill demand changes. This project aims to bridge the gap between job market expectations and student preparedness by analyzing real-world job listings.

3. Objectives

- Identify high-demand job titles, skills, and qualifications.
- Use scalable technologies (PySpark) for data querying and visualization.
- Explore the feasibility of salary prediction using machine learning models.

4. Dataset Description

The dataset used in this project is sourced from Kaggle and consists of over 13,000 job postings totaling approximately 1.5 GB in size. It represents a wide variety of job roles across multiple countries, collected from various job portals. The dataset captures both structured and semi-structured data, making it highly suitable for big data processing and exploratory analysis using PySpark.

Dataset Schema Overview

The schema includes 23 attributes, each capturing different aspects of the job listing. Below is a breakdown of the most critical features:

```
|-- Job Id: long (nullable = true)
|-- Experience: string (nullable = true)
|-- Qualifications: string (nullable = true)
|-- Salary Range: string (nullable = true)
|-- location: string (nullable = true)
|-- Country: string (nullable = true)
|-- latitude: double (nullable = true)
|-- longitude: double (nullable = true)
|-- Work Type: string (nullable = true)
|-- Company Size: integer (nullable = true)
|-- Job Posting Date: date (nullable = true)
|-- Preference: string (nullable = true)
|-- Contact Person: string (nullable = true)
|-- Contact: string (nullable = true)
|-- Job Title: string (nullable = true)
|-- Role: string (nullable = true)
|-- Job Portal: string (nullable = true)
|-- Job Description: string (nullable = true)
|-- Benefits: string (nullable = true)
|-- skills: string (nullable = true)
|-- Responsibilities: string (nullable = true)
|-- Company: string (nullable = true)
|-- Company Profile: string (nullable = true)
```

5. Tools and Technologies Used

To analyze a 1.5 GB job description dataset, we used the following tools:

- **Apache Spark with PySpark:**
  Used for scalable data processing and querying. PySpark enabled efficient handling of large volumes of semi-structured data through DataFrame operations and SQL-like queries.

- **Google Colab:**
  A cloud-based environment used to run PySpark and collaborate

in real time. It provided necessary computing power without local setup.

- **Scikit-learn:**
  Employed for implementing regression models (Linear, Lasso, Ridge, Decision Tree, Random Forest) and evaluating them using MSE, RMSE, and R^2 scores.

- **Matplotlib & Seaborn:**
  Python libraries used for creating visualizations of EDA results, skill distributions, salary trends, and model evaluations.

6. Data Preprocessing

To ensure the dataset was analysis-ready, the following preprocessing steps were applied:

1. **Null Value Handling:**
   Checked for and removed records with missing values in critical fields such as *Job Title*, *Job Description*, and *Skills*.

2. **Data Standardization:**
   Trimmed whitespaces and converted text fields (e.g., *Job Title*, *Company Name*) to lowercase to ensure consistency.

3. **Duplicate Removal:**
   Identified and eliminated duplicate records using *Job ID* and full-row comparison to avoid redundancy.

4. **Type Conversion:**
   Transformed columns like *Experience*, *Salary Range*, and *Job Posting Date* into appropriate numeric or date formats.

5. **Multi-Value Field Processing:**
   Split multi-valued fields (e.g., *Skills*) for token-level analysis and frequency computation.

These steps ensure data consistency and enable accurate querying, visualization, and model training.

7. Exploratory Data Analysis (EDA)

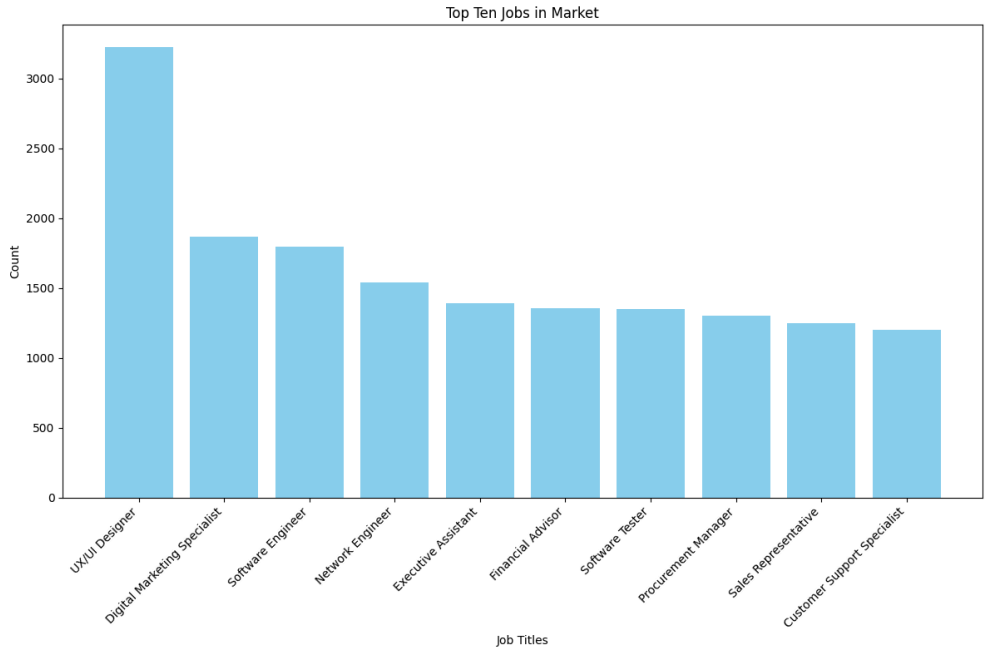
Exploratory Data Analysis (EDA)

EDA was conducted to gain insights into job roles, skill demand, qualifications, salary trends, geographic distributions, and temporal patterns. Key visualizations and statistical summaries were used to support our findings.

1. Job Role Frequency

- The top recurring job titles were:
  - **UX/UI Designer**
  - **Software Engineer**
  - **Digital Marketing Specialist**

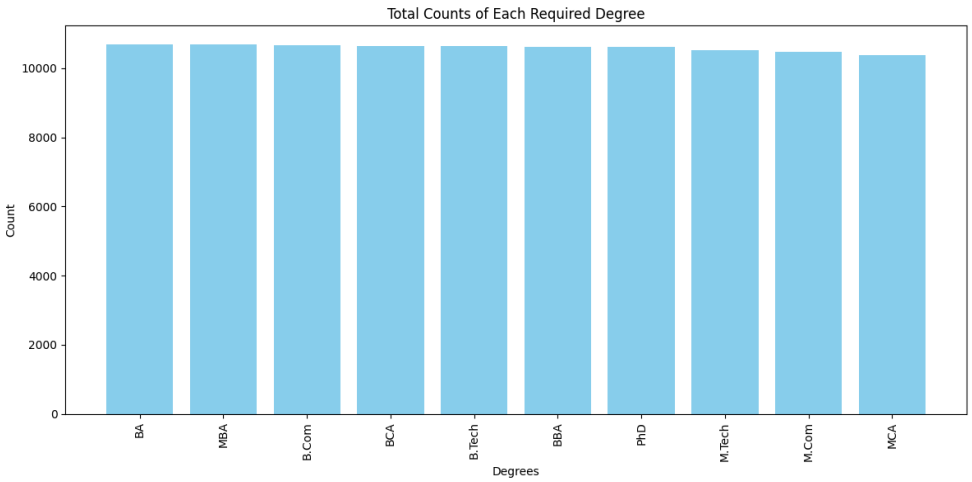
- These roles appeared across multiple companies and countries, indicating global demand.



2. Most Requested Qualifications

- The most in-demand qualifications were:
    - **BA (Bachelor of Arts)**
    - **MBA (Master of Business Administration)**
    - **B.Tech (Bachelor of Technology)**

- Job posts often listed combinations of degrees, reflecting varying education expectations by role.

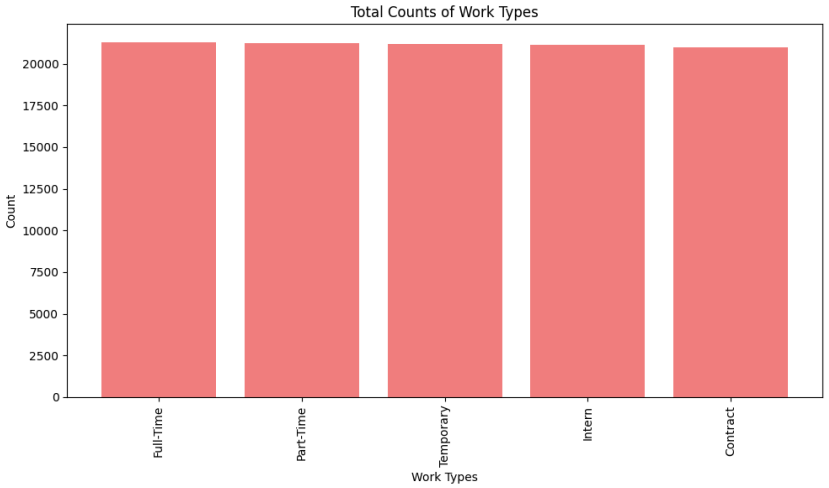


4. Job Type Distribution

The dataset showed a balanced representation of:

- **Full-time** roles (dominant category)
- **Internships**
- **Contract-based** positions

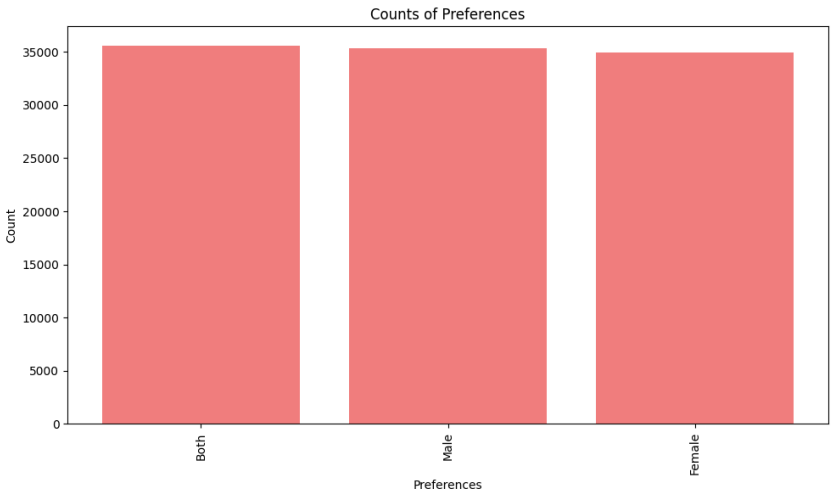
This diversity allowed for comparison of qualifications and salaries across employment types.



5. Gender Preference

- Most postings selected **“Both”** under gender preference, suggesting inclusive hiring trends.

- A small number specified *Male* or *Female* for particular roles (e.g., field agents, customer service).

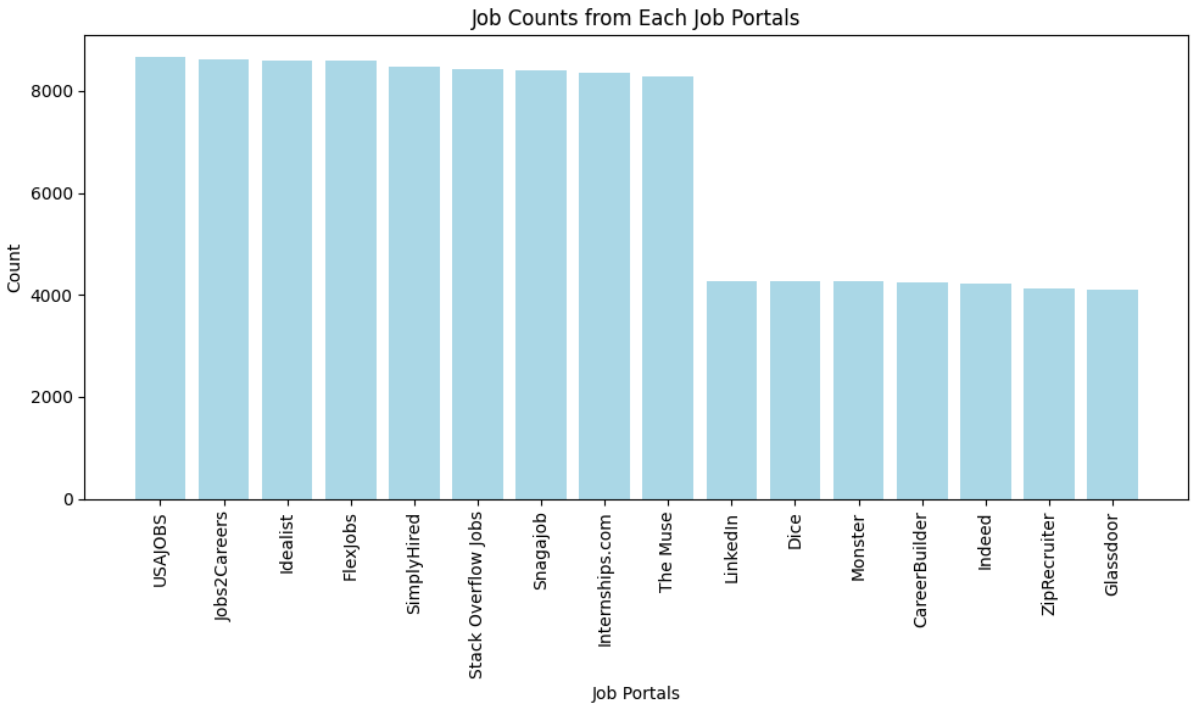


6. Top Job Portals by Job Count

- **USAJOBS, Jobs2Careers, Idealist, FlexJobs were the top portals, each contributing over 8,000 job listings.**

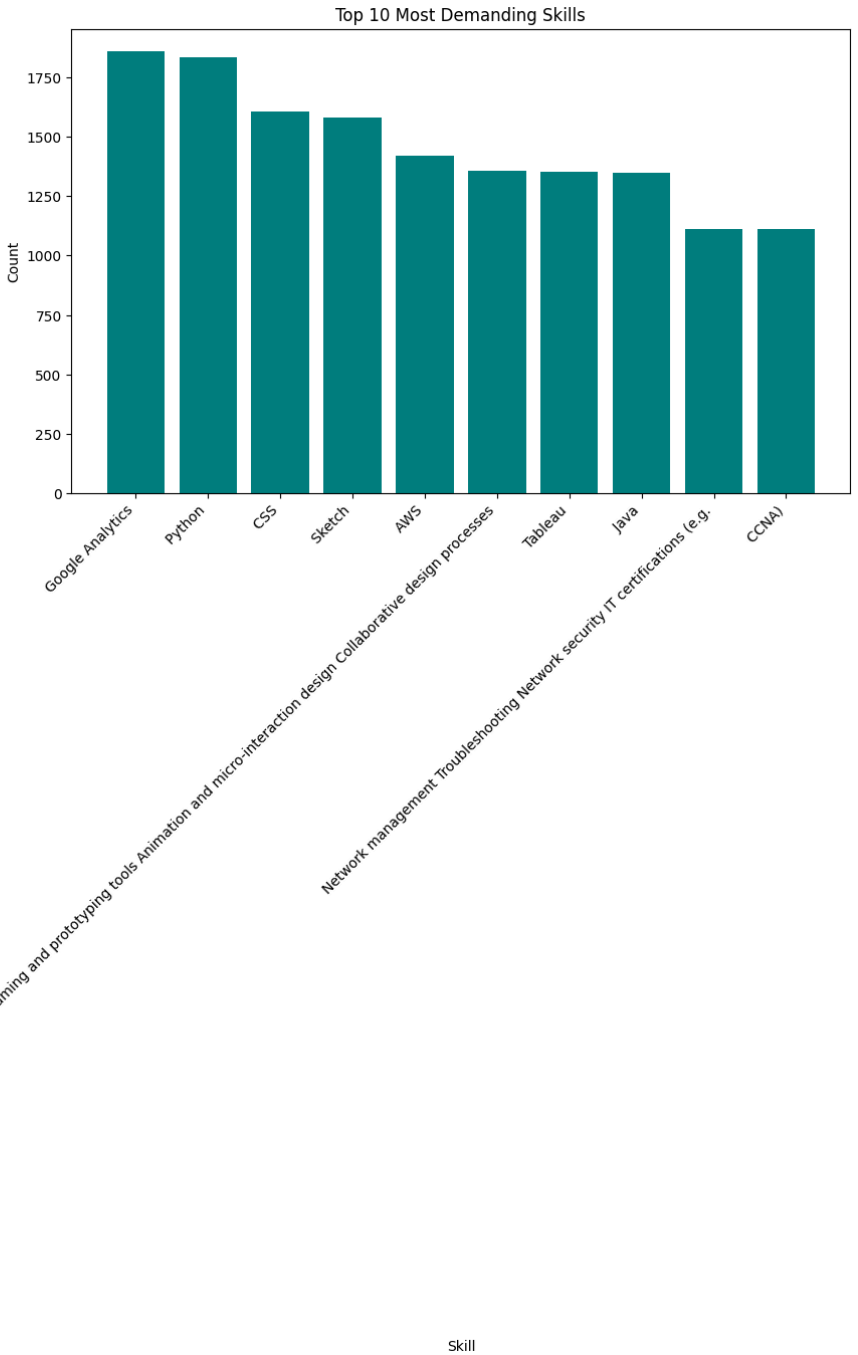
- **Portals like Glassdoor, ZipRecruiter, Indeed, and Monster had significantly fewer listings (around 4,000).**

- ***Insight:*** **Government and career-focused platforms (like USAJOBS) dominate in volume, whereas commercial platforms are more balanced.**



7. Top 10 Most Demanding Skills

- **Google Analytics and Python led the list, followed by:**
    - **CSS, Sketch, AWS, Tableau, Java, and CCNA.**
- ***Insight:*** **A strong demand for analytical, design, and cloud computing skills was observed across job roles.**

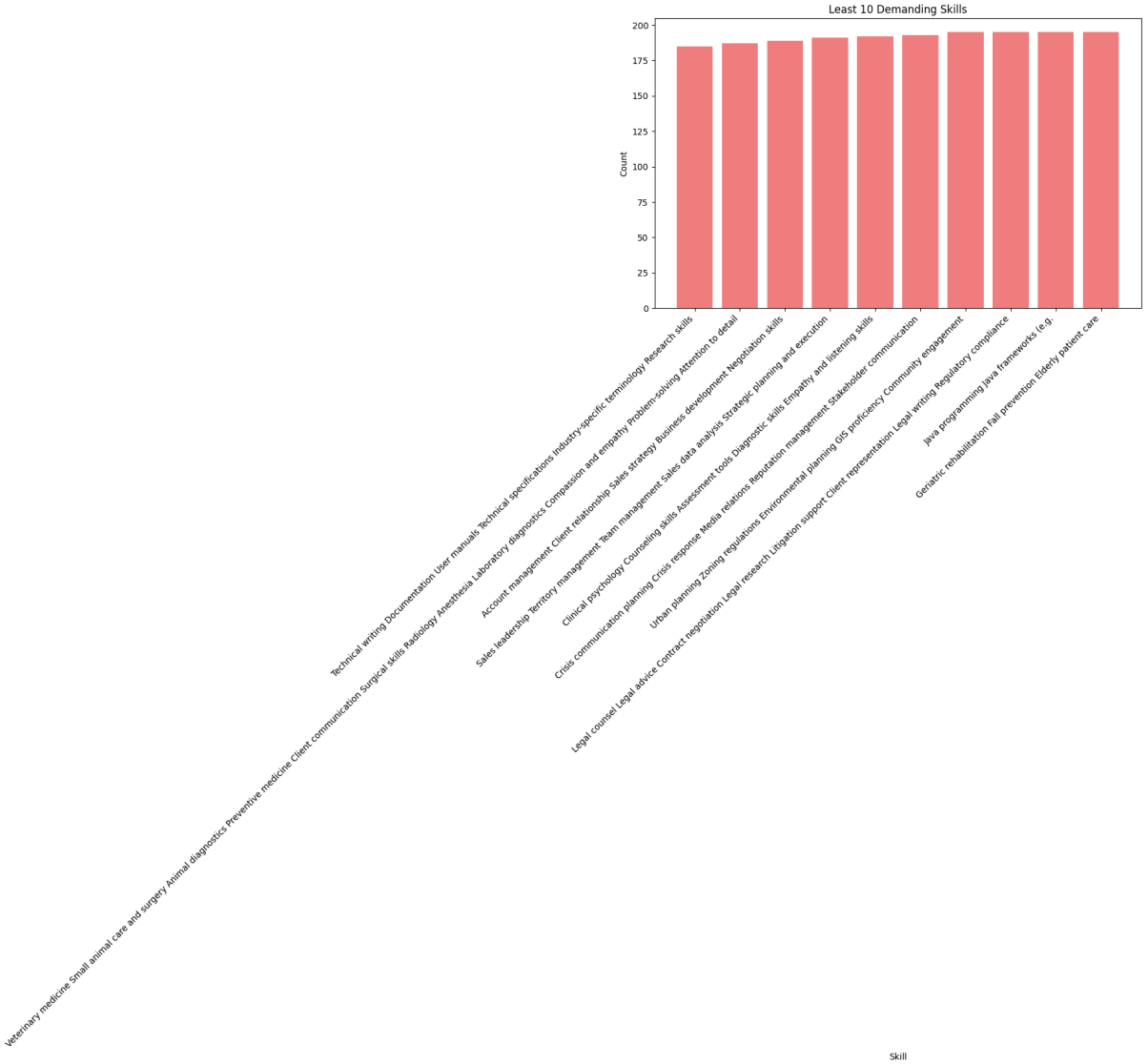


---

8. Least 10 Demanding Skills

- **Rarely requested skills included:**
    - **Veterinary care, Court presentation, Legal research, Civic communication, Geriatric rehabilitation, etc.**

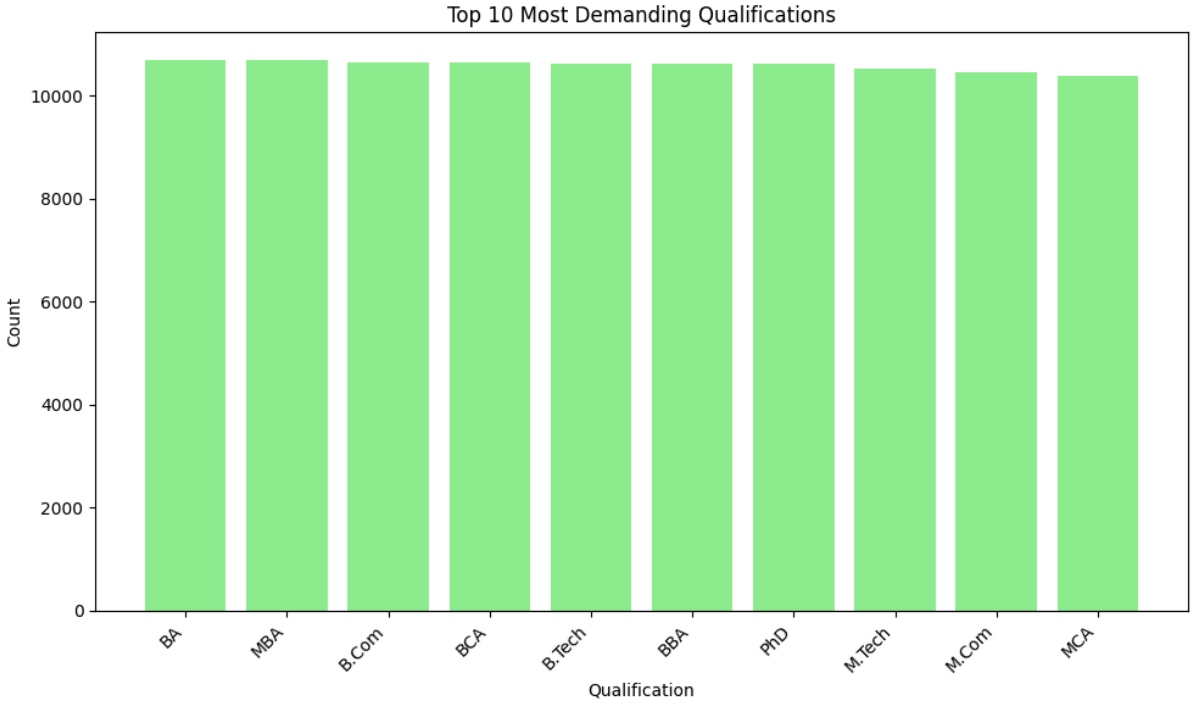
- ***Insight:*** **These skills are highly domain-specific and tied to niche professions, hence the low frequency.**



9.Top 10 Most Demanding Qualifications

- **Common qualifications included:**
    - **BA, MBA, B.Com, BCA, B.Tech, BBA, PhD, M.Tech, M.Com, MCA**

- ***Insight:*** **Both undergraduate and postgraduate degrees are in high demand, with preference for business, technology, and computer science disciplines.**



10. Least Demanding Job Titles

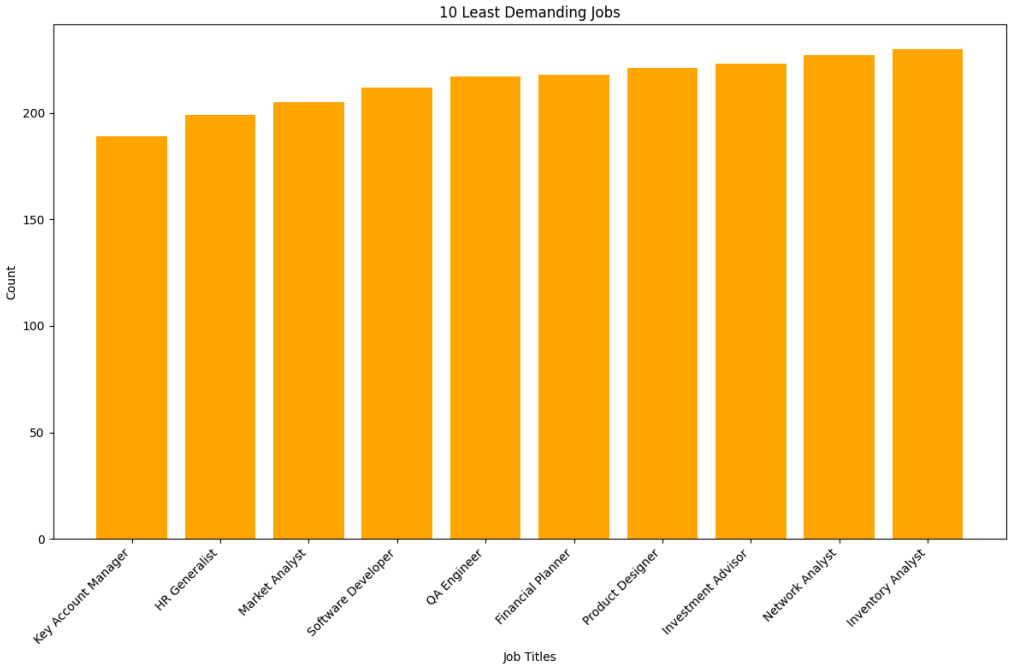
- **Titles like Key Account Manager, HR Generalist, and Market Analyst were among the least frequent.**

- ***Insight:*** **These jobs may either be niche roles or less advertised in the dataset, indicating limited openings or hiring focus.**

Company-Level Analysis

- Top hiring companies based on number of listings included:
    - **Infosys**
    - **HDFC**
    - **Procter & Gamble**
    - **Arrow Electronics**

- These companies also appeared in the list of top salary providers.

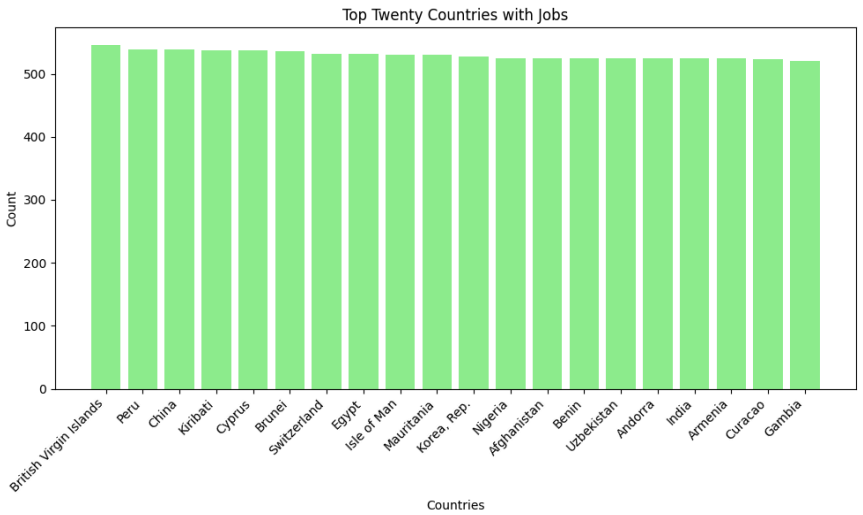


11. Country and Location Trends

Countries with the highest number of job listings:

- **India**
- **Peru**
- **British Virgin Islands**

Within countries, metro cities had a concentration of roles. *India* showed diverse city-wise demand.

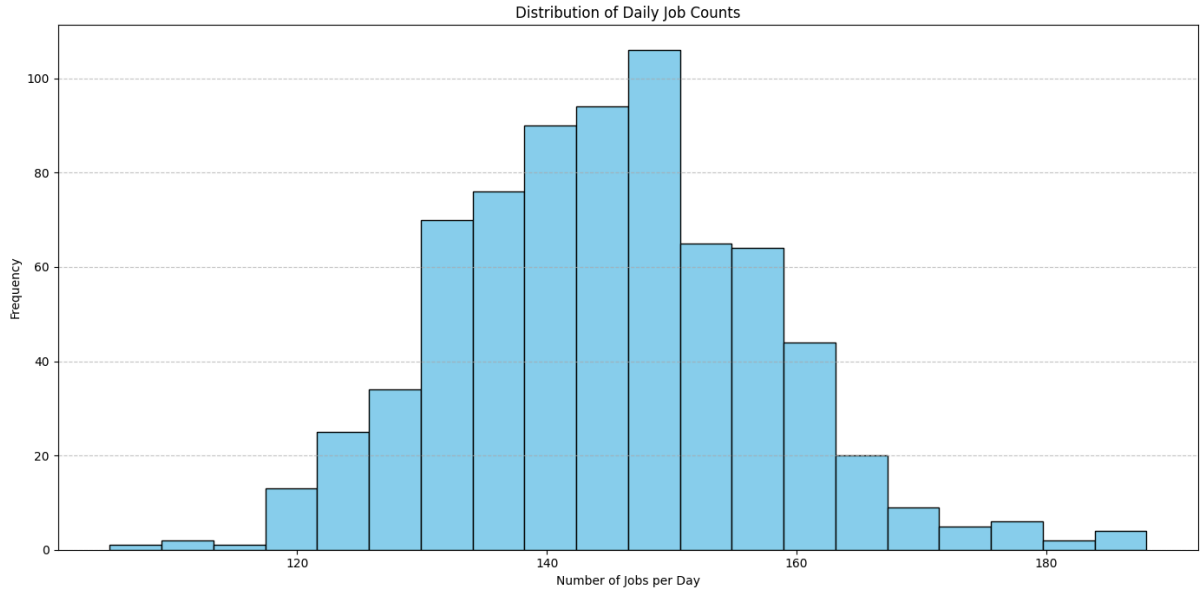


---

8. Temporal Patterns

12.Daily Job Count Distribution

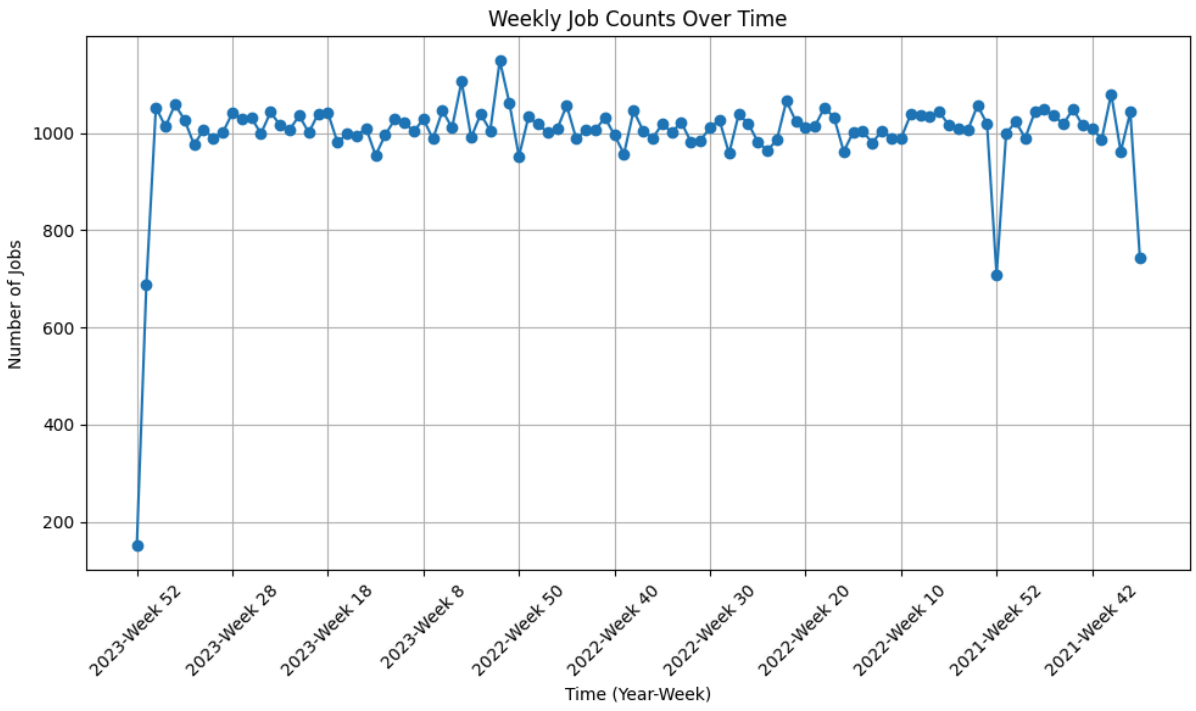
The histogram displays how many job postings occurred per day. Most days recorded between **130 to 160 jobs**, with a bell-shaped distribution indicating consistent daily posting behavior.



---

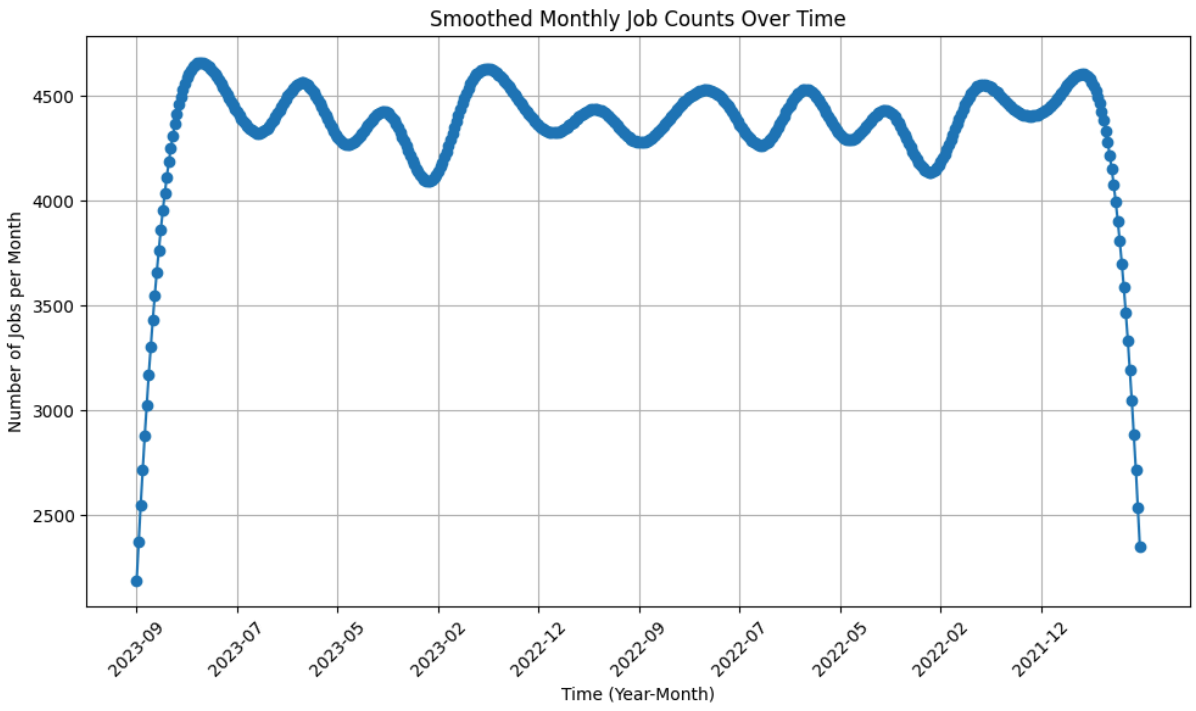
13. Weekly Job Posting Trends

This line plot shows weekly job counts over time. The dataset appears stable with weekly postings averaging around **1000 jobs**, but with occasional dips likely due to missing or incomplete data during holidays or system downtimes.



14.Smoothed Monthly Job Trends

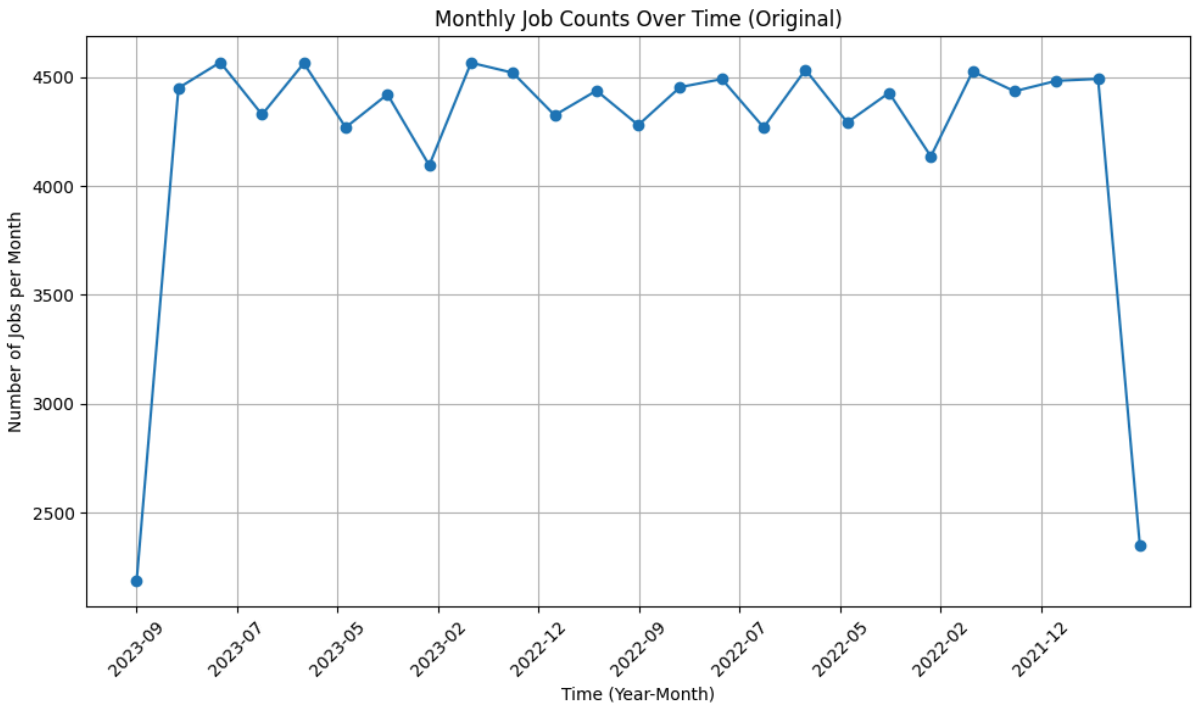
Using a rolling average, the smoothed monthly trend line highlights consistent hiring activity over time. Peaks and troughs reflect industry hiring cycles, with **periodic spikes** suggesting seasonal recruitment periods.



---

15. Original Monthly Job Counts

This chart represents raw monthly job count data. It aligns with the smoothed trend, affirming the data integrity and revealing **stable hiring volumes** across months, especially in **2022**, which shows the highest activity.



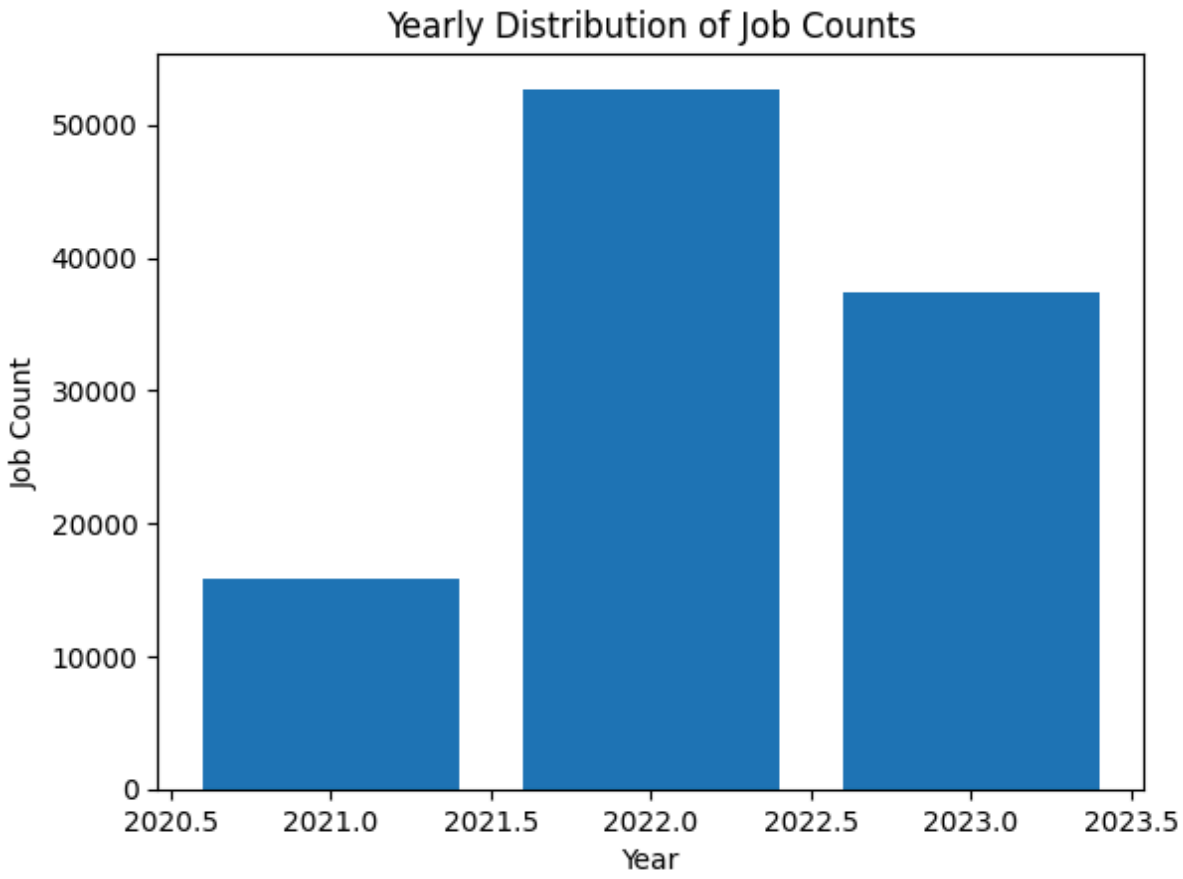
---

16. Yearly Job Distribution

Bar chart comparing job counts across years:

- **2022** had the most job postings (~52,000)
- **2023** saw a slight decline (~38,000)
- **2021** had fewer postings (~16,000)

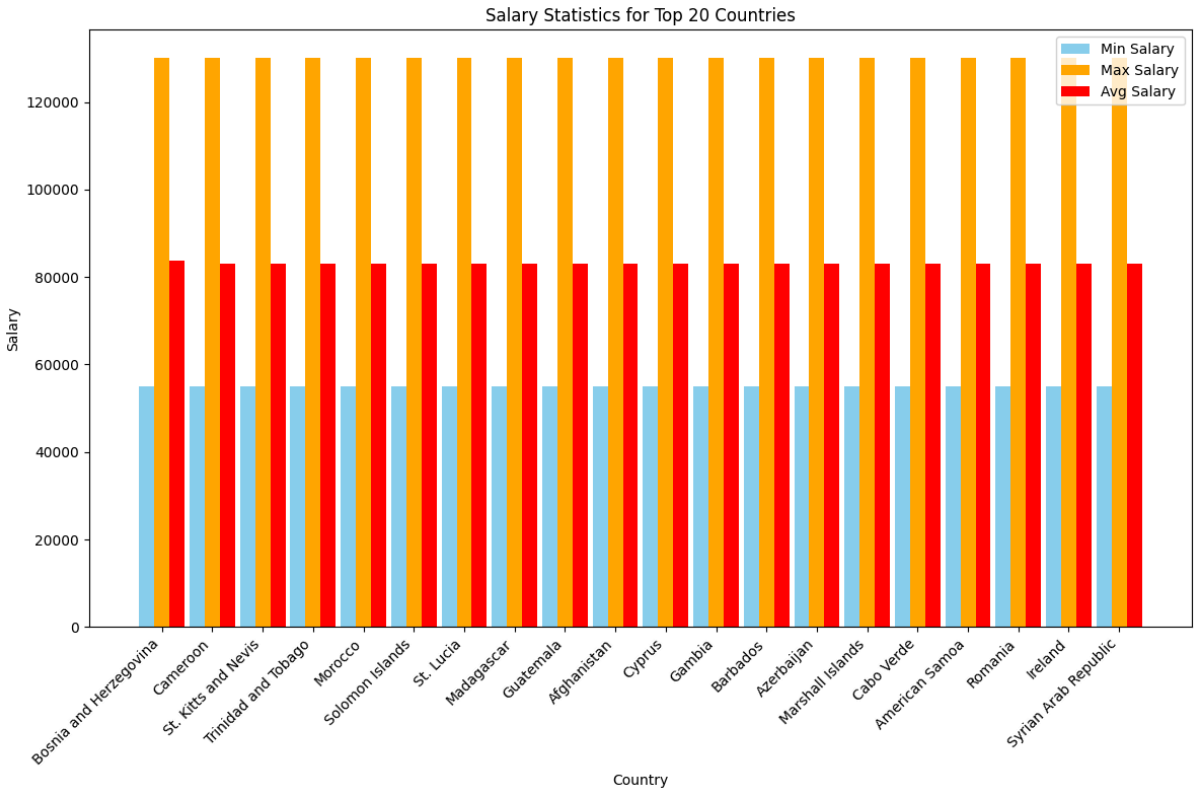
This confirms a **post-pandemic hiring boom in 2022** followed by normalization in 2023.



19.Top 20 Countries with Highest Average Salaries

These countries showed consistently high average salaries, with minimal variation between min and max salary ranges. Examples include:

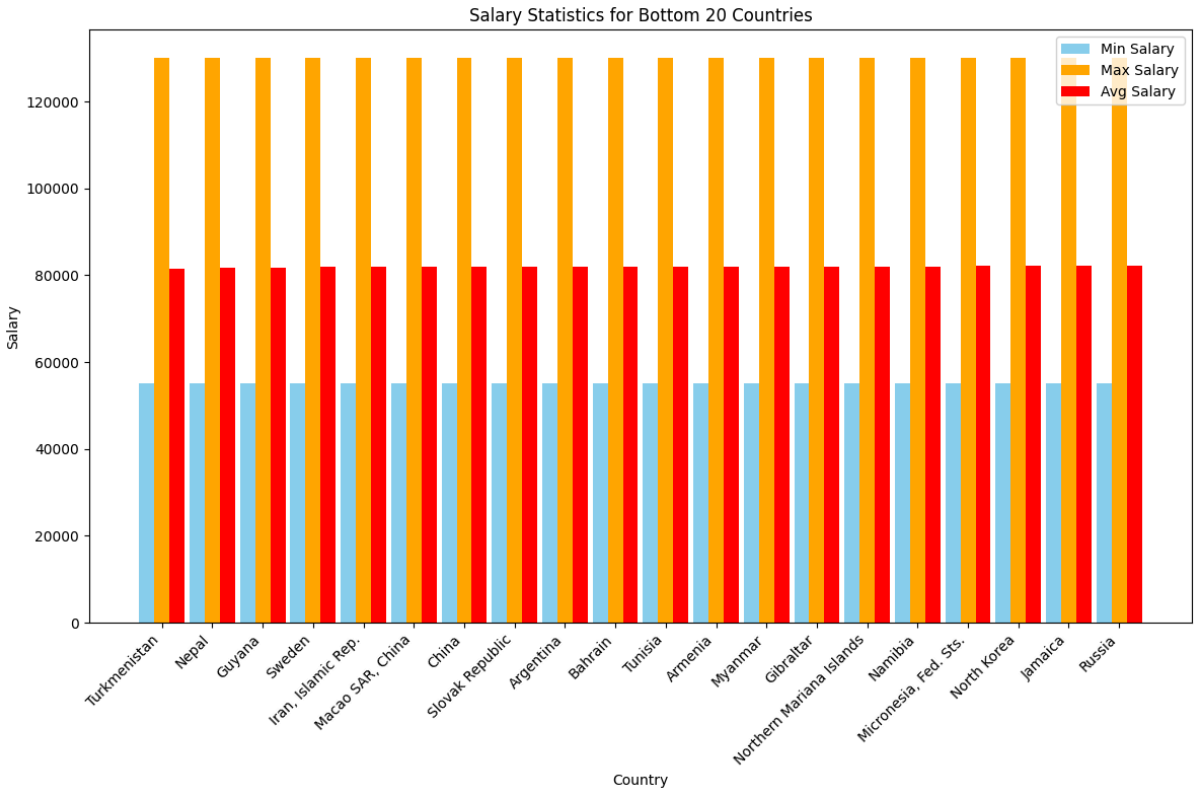
- **Bosnia and Herzegovina**
- **Cameroon**
- **Trinidad and Tobago**
- **Ireland**
- **Syrian Arab Republic**



20. Bottom 20 Countries with Lowest Average Salaries

Despite similar max salaries as high-income countries, these nations had much lower average salaries, indicating fewer high-paying job roles:

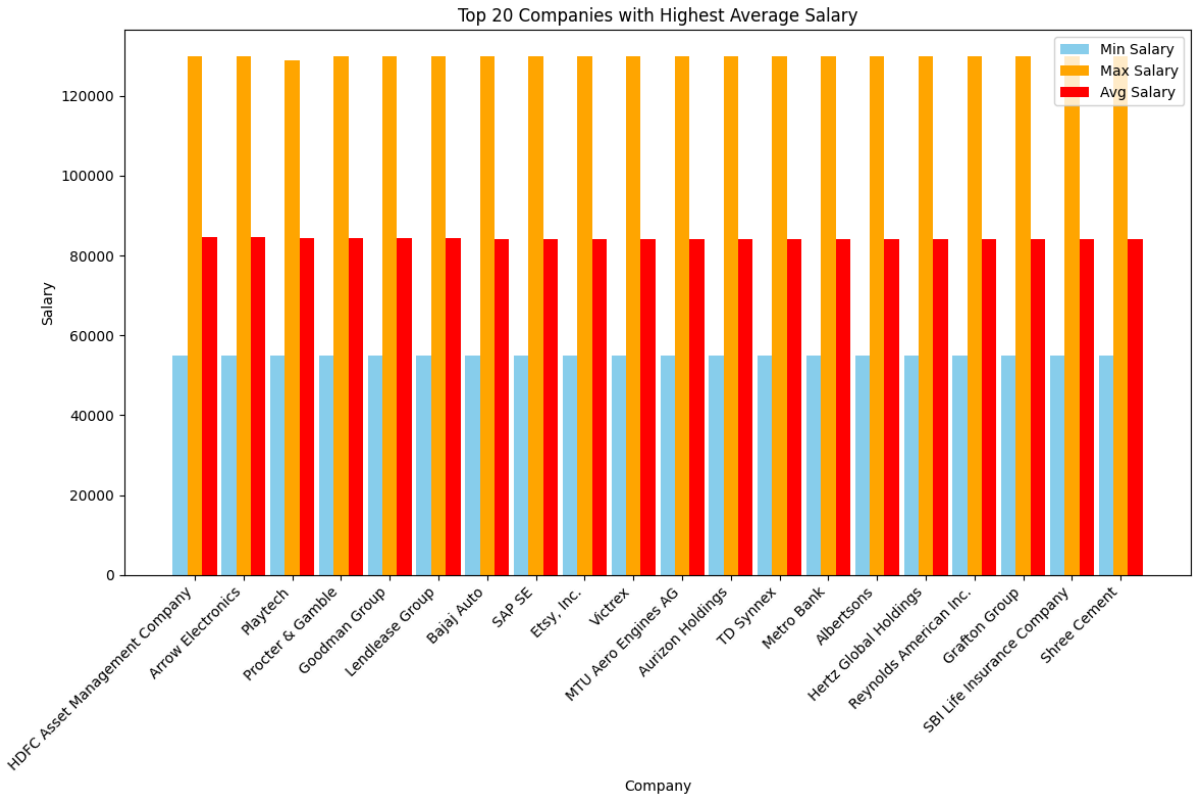
- **Turkmenistan**
- **Nepal**
- **Iran**
- **Gibraltar**
- **North Korea**



21. Top 20 Companies with Highest Average Salaries

These companies offered attractive average pay, suggesting premium job roles or competitive markets:

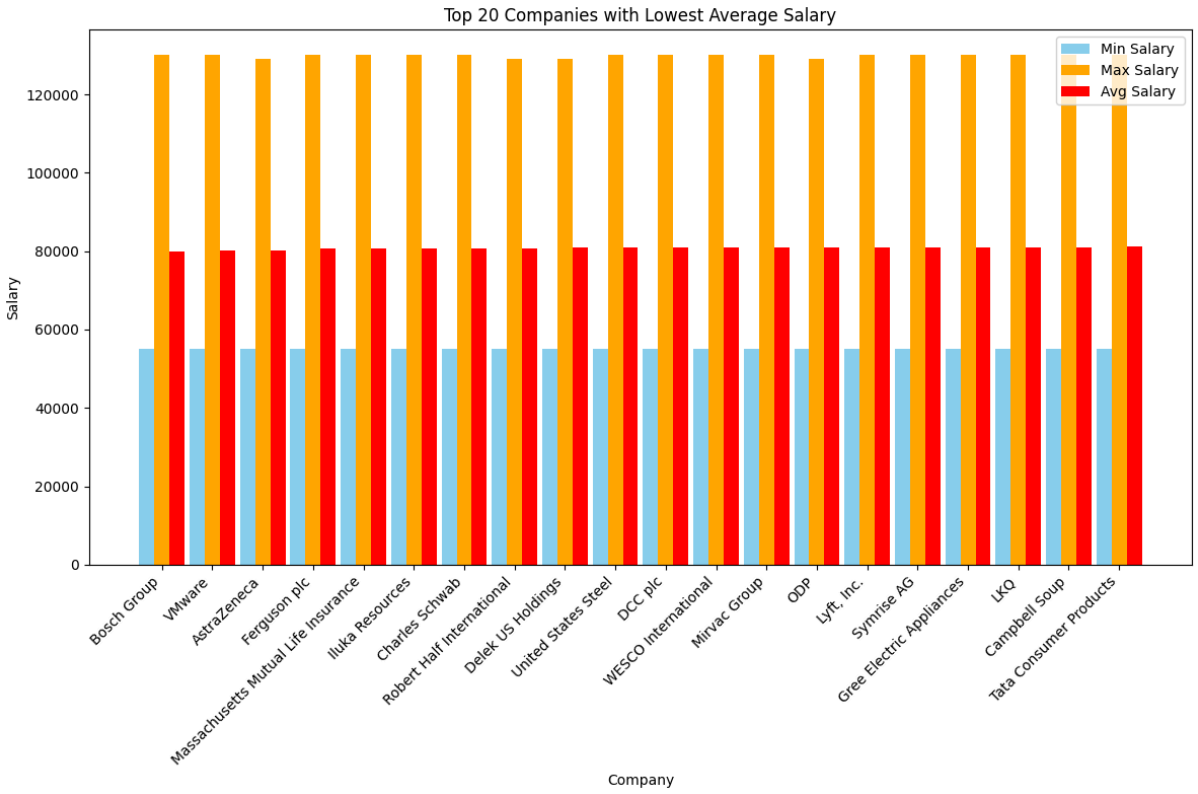
- **HDFC Asset Management Company**
- **Arrow Electronics**
- **SAP SE**
- **TD Synex**
- **Shree Cement**



22.Top 20 Companies with Lowest Average Salaries

Organizations here may target entry-level roles or operate in low-paying sectors:

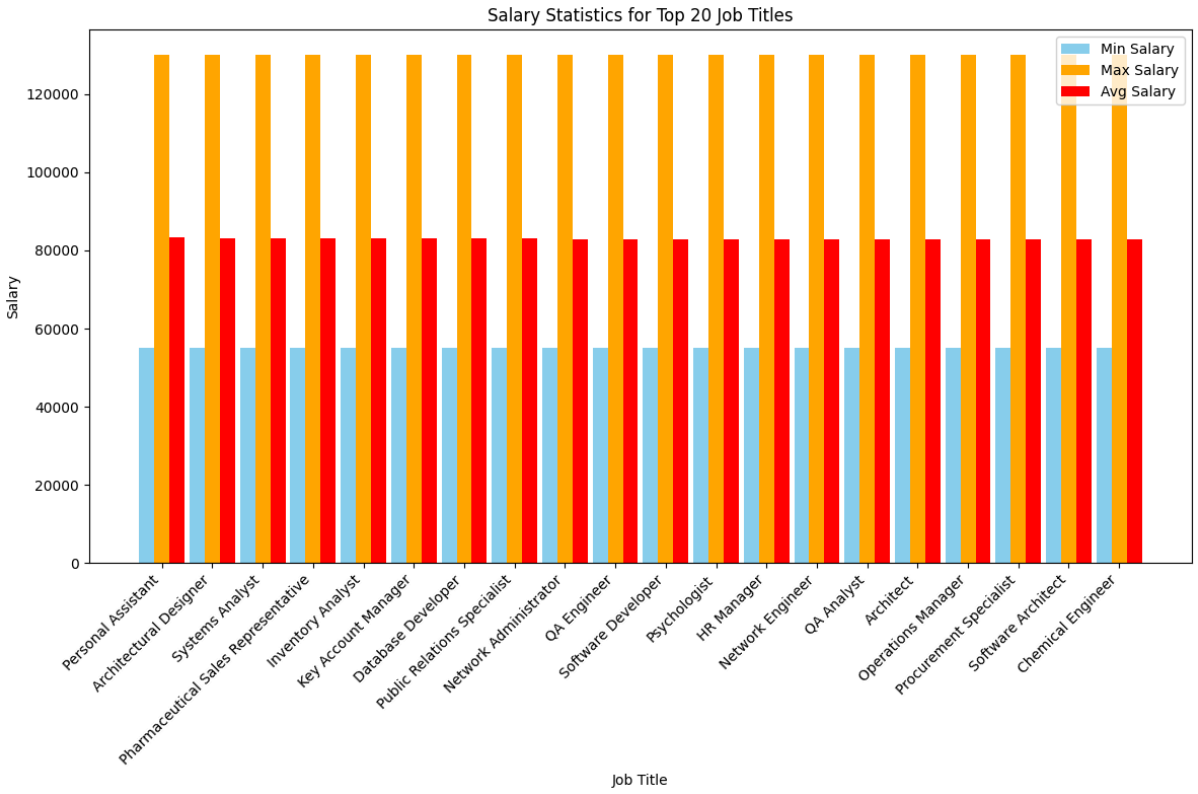
- **Bosch Group**
- **VMware**
- **AstraZeneca**
- **Campbell Soup**
- **Tata Consumer Products**



23.Top 20 Job Titles with Highest Average Salaries

These roles are typically technical or executive-level positions:

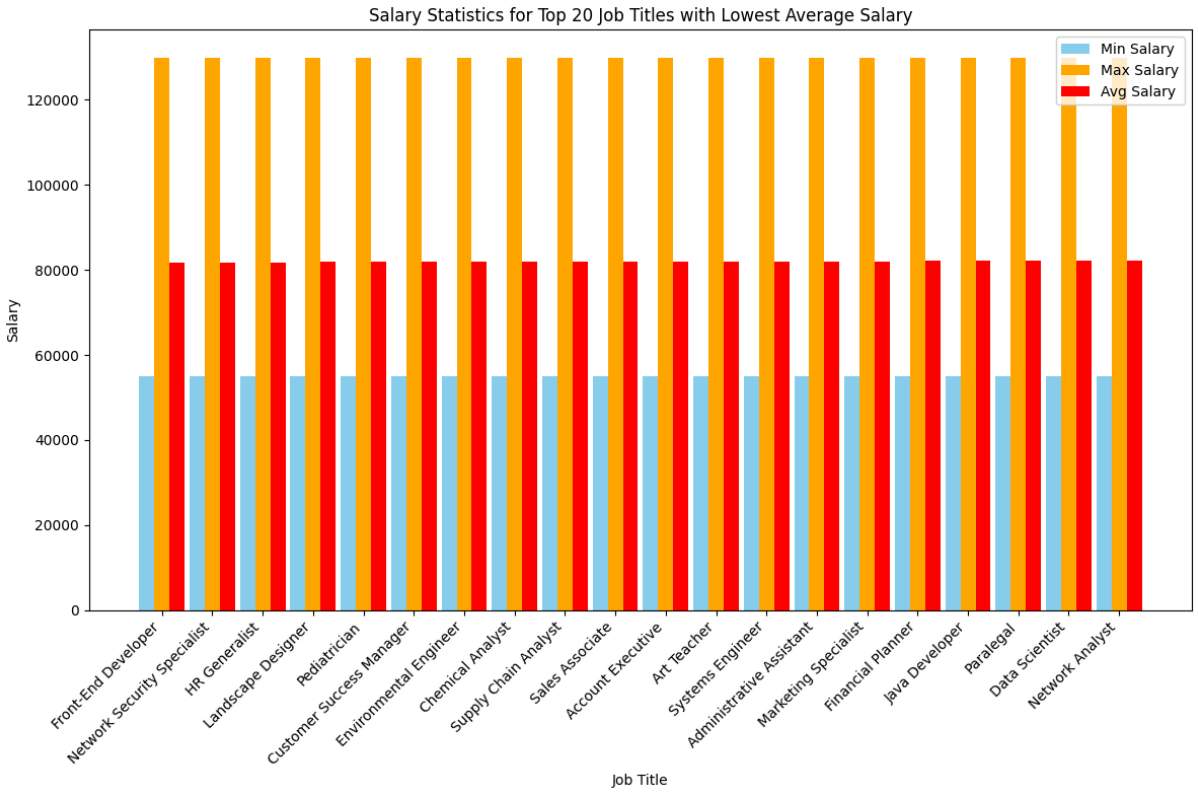
- **Software Architect**
- **Procurement Specialist**
- **QA Analyst**
- **Database Developer**
- **Chemical Engineer**



24.Top 20 Job Titles with Lowest Average Salaries

These roles may be entry-level, support-oriented, or in non-tech domains:

- **Landscape Designer**
- **Customer Success Manager**
- **HR Generalist**
- **Art Teacher**
- **Administrative Assistant**



8. Analytical Queries [20]

Simple Queries [5]

1. **Display Sample Data**:

   Retrieves the top 5 rows to preview the dataset contents.

Qualifications	location	Country	latitude	longitude	Work Type	Company Size	Job Posting Date	Preference	Job Title
M.Tech	Douglas	Isle of Man	54.2361	-4.5481	Intern	26801	1.6507584E9	Female	Digital Marketing...
BCA	Ashgabat	Turkmenistan	38.9697	59.5563	Intern	100340	1.671408E9	Female	Web Developer
PhD	Macao	Macao SAR, China	22.1987	113.5439	Temporary	84525	1.6631136E9	Male	Operations Manager
PhD	Porto-Novo	Benin	9.3077	2.3158	Full-Time	129896	1.6772832E9	Female	Network Engineer
MBA	Santiago	Chile	-35.6751	-71.5429	Intern	53944	1.6654464E9	Female	Event Manager

2. **Unique Job Titles**:
   Lists all distinct job titles found in the dataset.

Job Title
Landscape Designer
Product Designer
Finance Manager
Litigation Attorney
Event Coordinator
Investment Banker
Financial Analyst
Nurse Manager
Network Engineer
Sales Associate
Aerospace Engineer
Digital Marketing...
Civil Engineer
Procurement Manager
Substance Abuse C...
Architectural Des...
Dental Hygienist
Account Executive
Registered Nurse
QA Analyst

3. **Total Job Postings**:
   Counts the total number of job entries in the dataset.

Qualifications	location	Country	latitude	longitude	Work Type	Company Size	Job Posting Date	Preference	Job Title	Role	Job Portal	Job Description
105902	105902	105902	105902	105902	105902	105902	105902	105902	105902	105902	105902	105902

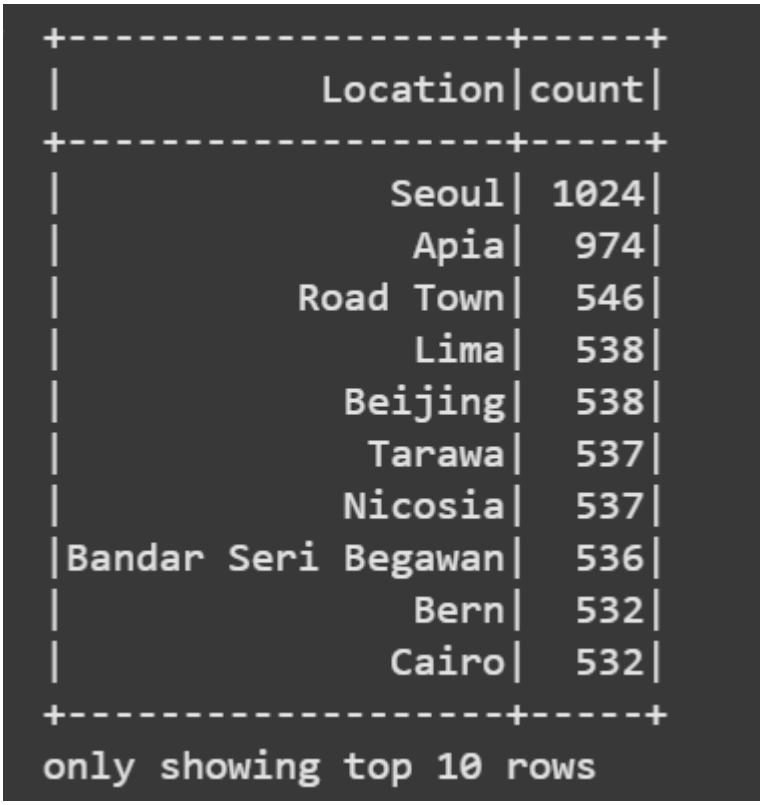
4. **Null Value Analysis**:
   Counts how many null values are present in each column.

```
+--------------------+
|             Company|
+--------------------+
|      Carrier Global|
|             E.ON SE|
|   Ping An Insurance|
|Deutsche Lufthans...|
|GAIL (India) Limited|
|             Stryker|
|McCormick & Compa...|
|                 KKR|
|      Otis Worldwide|
| World Fuel Services|
|   Mohawk Industries|
|             Corning|
|Enterprise Produc...|
|IBM (Internationa...|
|                 CDW|
|        Downer Group|
|The Kraft Heinz C...|
|   AmerisourceBergen|
|           Nordstrom|
|Fortescue Metals ...|
+--------------------+
only showing top 20 rows
```

5. **Distinct Companies**:
   Extracts the list of unique company names posting jobs.

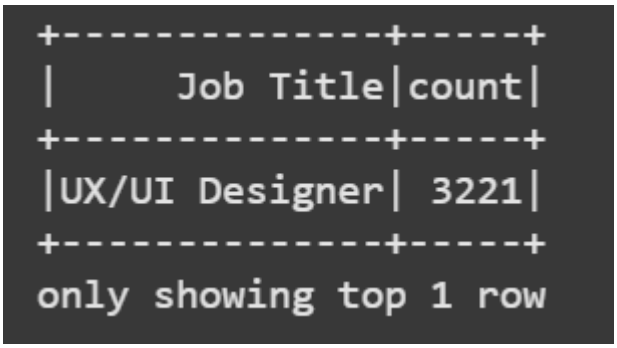
Moderate Queries

6. **Jobs per Location**:
   Groups job listings by location to identify hiring hotspots.



```
+------------------+-----+
|          Location|count|
+------------------+-----+
|             Seoul| 1024|
|              Apia|  974|
|         Road Town|  546|
|              Lima|  538|
|           Beijing|  538|
|            Tarawa|  537|
|           Nicosia|  537|
|Bandar Seri Begawan|  536|
|              Bern|  532|
|             Cairo|  532|
+------------------+-----+
only showing top 10 rows
```

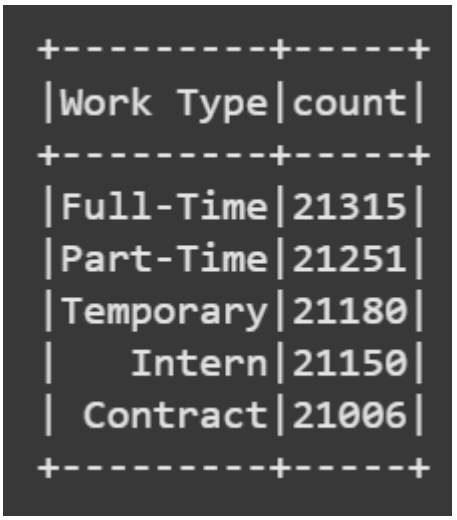
7. **Most Common Job Title**:
   Finds the most frequently occurring job title in the data.



```
+--------------+-----+
|     Job Title|count|
+--------------+-----+
|UX/UI Designer| 3221|
+--------------+-----+
only showing top 1 row
```

8. **Jobs by Work Type**:
   Shows how job postings are distributed across work types (e.g.,

full-time, remote).



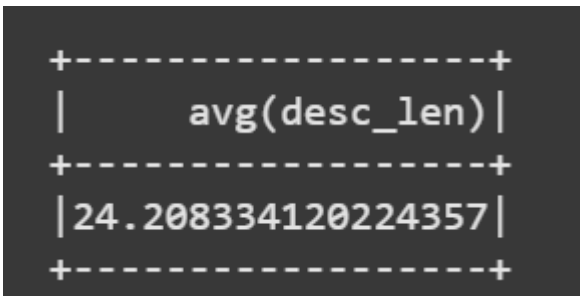
Work Type	count
Full-Time	21315
Part-Time	21251
Temporary	21180
Intern	21150
Contract	21006

Average Job Description Word Count:
Uses string length and space replacement to estimate the average word count of job descriptions:

python
Copy code

```
length("Job Description") -
length(regexp_replace("Job Description", " ", "")) +
```

9. Average word count in job description



avg(desc_len)
24.208334120224357

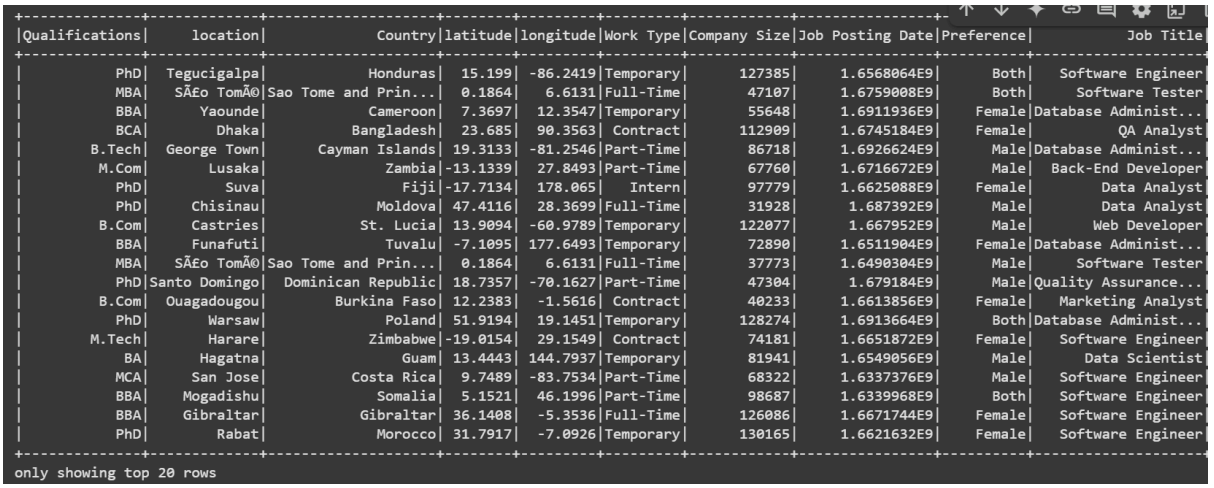
10. **Filter by Job Title ('Data Scientist')**:

Extracts job listings where the title includes "data scientist".

Qualifications	location	Country	latitude	longitude	Work Type	Company Size	Job Posting Date	Preference	Job Title
BA	Hagatna	Guam	13.4443	144.7937	Temporary	81941	1.6549056E9	Male	Data Scientist
BA	Douglas	Isle of Man	54.2361	-4.5481	Full-Time	120658	1.6789248E9	Both	Data Scientist
PhD	San Jose	Costa Rica	9.7489	-83.7534	Contract	89309	1.641168E9	Male	Data Scientist
BBA	Muscat	Oman	21.4735	55.9754	Full-Time	89311	1.6910208E9	Female	Data Scientist
BCA	Yaren District (d...	Nauru	-0.5228	166.9315	Contract	121396	1.6728768E9	Female	Data Scientist
BA	Road Town	British Virgin Is...	18.4207	-64.6399	Contract	91597	1.642464E9	Male	Data Scientist
M.Com	Macao	Macao SAR, China	22.1987	113.5439	Full-Time	81987	1.6550784E9	Female	Data Scientist
M.Tech	New Delhi	India	20.5937	78.9629	Full-Time	43424	1.6669152E9	Both	Data Scientist
BA	Apia	Samoa	-13.759	-172.1046	Temporary	31009	1.6745184E9	Male	Data Scientist
M.Com	Yerevan	Armenia	40.0691	45.0382	Part-Time	118085	1.6651008E9	Both	Data Scientist
M.Tech	Windhoek	Namibia	-22.9576	18.4904	Contract	127027	1.6578432E9	Both	Data Scientist
BCA	Dodoma	Tanzania	-6.369	34.8888	Full-Time	28385	1.6663968E9	Female	Data Scientist
M.Tech	Funafuti	Tuvalu	-7.1095	177.6493	Contract	104386	1.674864E9	Both	Data Scientist
MBA	Banjul	Gambia	13.4432	-15.3101	Temporary	69333	1.6470432E9	Both	Data Scientist
M.Com	Brussels	Belgium	50.5039	4.4699	Part-Time	63881	1.6939584E9	Both	Data Scientist
B.Tech	Majuro	Marshall Islands	7.1315	171.1845	Contract	120637	1.679184E9	Female	Data Scientist
B.Tech	Seoul	Korea, Rep.	35.9078	127.7669	Full-Time	109920	1.693008E9	Female	Data Scientist
B.Tech	Moroni	Comoros	-11.6455	43.3333	Temporary	79956	1.6511904E9	Male	Data Scientist
M.Tech	Victoria	Seychelles	-4.6796	55.492	Full-Time	25455	1.6768512E9	Female	Data Scientist
BA	Athens	Greece	39.0742	21.8243	Intern	57726	1.6508448E9	Male	Data Scientist

11. **Jobs Requiring Python**:

Filters job posts where Python is mentioned in the skills section.

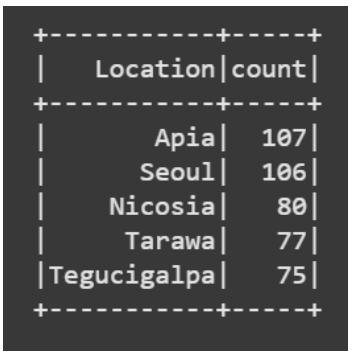


Qualifications	location	Country	latitude	longitude	Work Type	Company Size	Job Posting Date	Preference	Job Title
PhD	Tegucigalpa	Honduras	15.199	-86.2419	Temporary	127385	1.6568064E9	Both	Software Engineer
MBA	SÃ£o TomÃ©	Sao Tome and Prin...	0.1864	6.6131	Full-Time	47107	1.6759008E9	Both	Software Tester
BBA	Yaounde	Cameroon	7.3697	12.3547	Temporary	55648	1.6911936E9	Female	Database Administ...
BCA	Dhaka	Bangladesh	23.685	90.3563	Contract	112909	1.6745184E9	Female	QA Analyst
B.Tech	George Town	Cayman Islands	19.3133	-81.2546	Part-Time	86718	1.6926624E9	Male	Database Administ...
M.Com	Lusaka	Zambia	-13.1339	27.8493	Part-Time	67760	1.6716672E9	Male	Back-End Developer
PhD	Suva	Fiji	-17.7134	178.065	Intern	97779	1.6625088E9	Female	Data Analyst
PhD	Chisinau	Moldova	47.4116	28.3699	Full-Time	31928	1.687392E9	Male	Data Analyst
B.Com	Castries	St. Lucia	13.9094	-60.9789	Temporary	122077	1.667952E9	Male	Web Developer
BBA	Funafuti	Tuvalu	-7.1095	177.6493	Temporary	72890	1.6511904E9	Female	Database Administ...
MBA	SÃ£o TomÃ©	Sao Tome and Prin...	0.1864	6.6131	Full-Time	37773	1.6490304E9	Male	Software Tester
PhD	Santo Domingo	Dominican Republic	18.7357	-70.1627	Part-Time	47304	1.679184E9	Male	Quality Assurance...
B.Com	Ouagadougou	Burkina Faso	12.2383	-1.5616	Contract	40233	1.6613856E9	Female	Marketing Analyst
PhD	Warsaw	Poland	51.9194	19.1451	Temporary	128274	1.6913664E9	Both	Database Administ...
M.Tech	Harare	Zimbabwe	-19.0154	29.1549	Contract	74181	1.6651872E9	Female	Software Engineer
BA	Hagatna	Guam	13.4443	144.7937	Temporary	81941	1.6549056E9	Male	Data Scientist
MCA	San Jose	Costa Rica	9.7489	-83.7534	Part-Time	68322	1.6337376E9	Male	Software Engineer
BBA	Mogadishu	Somalia	5.1521	46.1996	Part-Time	98687	1.6339968E9	Both	Software Engineer
BBA	Gibraltar	Gibraltar	36.1408	-5.3536	Full-Time	126086	1.6671744E9	Female	Software Engineer
PhD	Rabat	Morocco	31.7917	-7.0926	Temporary	130165	1.6621632E9	Female	Software Engineer

only showing top 20 rows

12. **Top 5 Locations Hiring for IT Roles**:

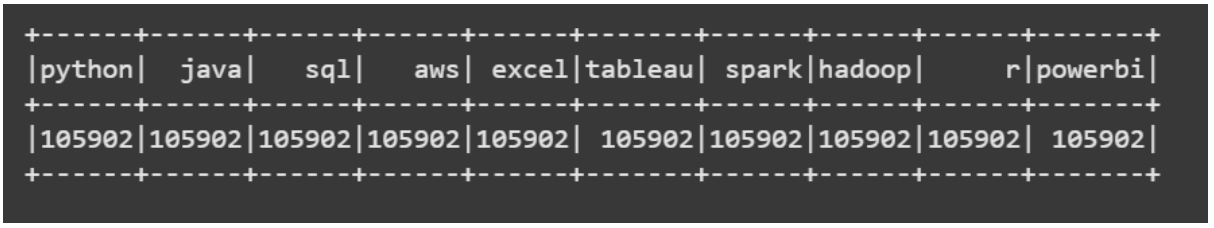
Filters for "IT" in job titles and groups by location to rank hiring cities.



Location	count
Apia	107
Seoul	106
Nicosia	80
Tarawa	77
Tegucigalpa	75

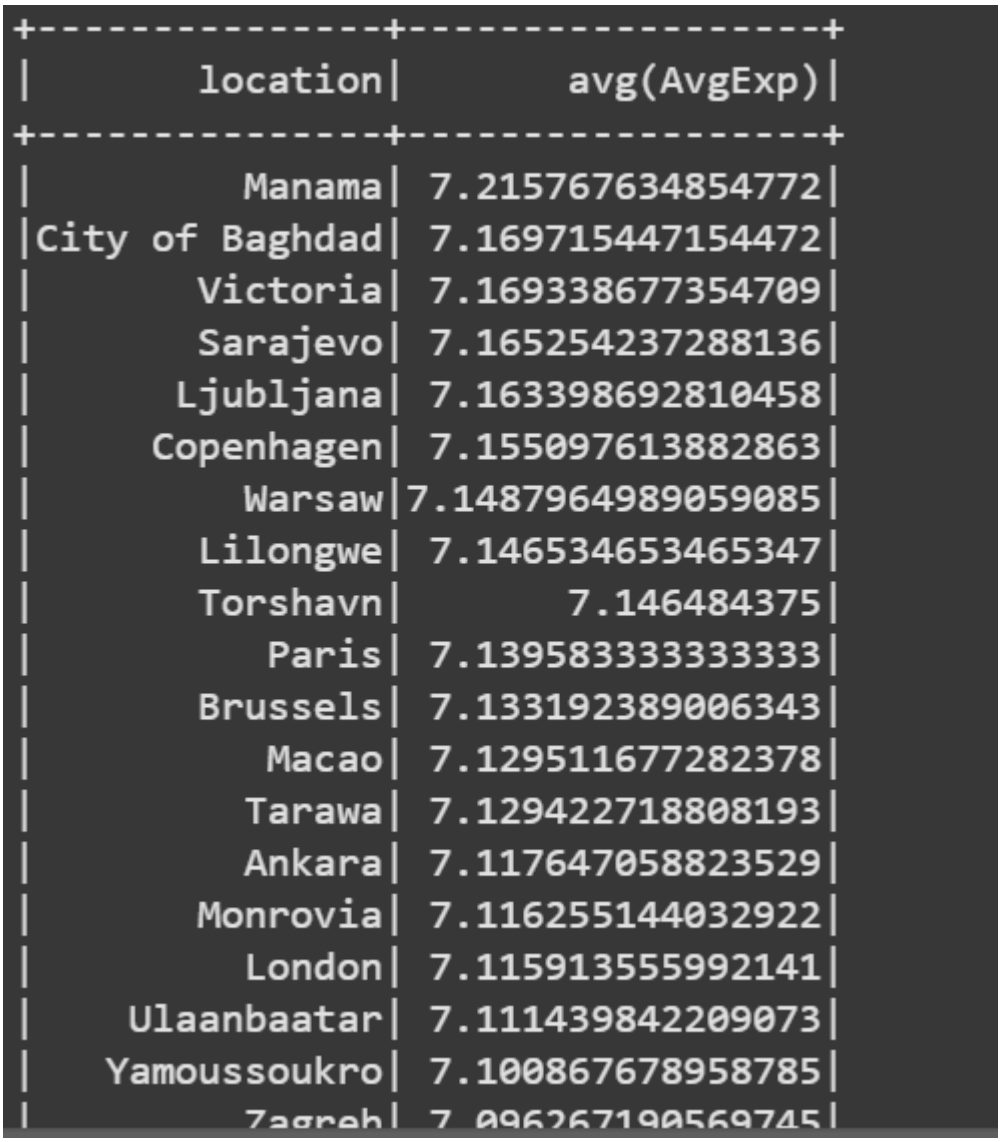
Complex Queries

13. **Tech Skills Frequency**:
    Counts how often popular tech skills (e.g., Python, Java, SQL, etc.) appear across all job listings.



python	java	sql	aws	excel	tableau	spark	hadoop	r	powerbi
105902	105902	105902	105902	105902	105902	105902	105902	105902	105902

14. **Join with Industry-Sector Mapping**:
    Joins the main dataset with a dummy mapping of industries to sectors to enable hierarchical analysis.



location	avg(AvgExp)
Manama	7.215767634854772
City of Baghdad	7.169715447154472
Victoria	7.169338677354709
Sarajevo	7.165254237288136
Ljubljana	7.163398692810458
Copenhagen	7.155097613882863
Warsaw	7.1487964989059085
Lilongwe	7.146534653465347
Torshavn	7.146484375
Paris	7.139583333333333
Brussels	7.133192389006343
Macao	7.129511677282378
Tarawa	7.129422718808193
Ankara	7.117647058823529
Monrovia	7.116255144032922
London	7.115913555992141
Ulaanbaatar	7.111439842209073
Yamoussoukro	7.100867678958785
Zagreb	7.096267190569745

15. **Avg. Description Length by Location**:

Computes average job description length grouped by location, revealing where jobs are described in more detail.

Location	avg(desc_len)
George Town	190.91242362525458
Yaounde	190.80293501048217
Bratislava	190.75193798449612
Oslo	190.62331838565024
Mogadishu	190.58178053830227
AsunciÃ³n	190.55696202531647
Phnom Penh	190.48447204968943
Valletta	190.35535307517085
Abu Dhabi	190.04868154158214
Managua	189.95632183908046
San Salvador	189.9156862745098
Kuala Lumpur	189.7524557956778
Singapore	189.40325865580448
Warsaw	189.1816192560175
Vaduz	189.14947368421053
Yaren District (d...	189.021484375
Podgorica	188.9958762886598
Praia	188.95841995841997
Pristina	188.95689655172413
Roseau	188.73752711496746

16. **Companies with High Volume Postings**:
Identifies companies that posted more than 50 jobs, indicating active recruiters.

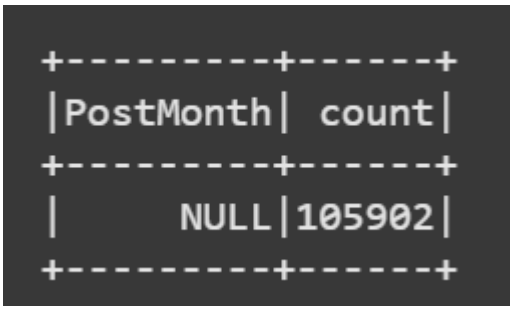
Company	count
Fifth Third Bancorp	165
Royal Mail	155
ITC Limited	152
Burberry Group	150
3M	148
Bank of China	148
V-Guard Industries	148
Leidos Holdings	146
Spirit Airlines, ...	145
Ford Motor Company	144
China Southern Ai...	144
Regions Financial	143
TAG Immobilien AG	143
Sherwin-Williams	143
Ovintiv	143
Kohl's	142
Corning	141
Equinix	141
GlaxoSmithKline	141
Huntington Ingall...	140

17. **Avg. Experience by Location**:

Parses min and max experience, averages them per row, and groups by location to compute city-wise job experience expectations.

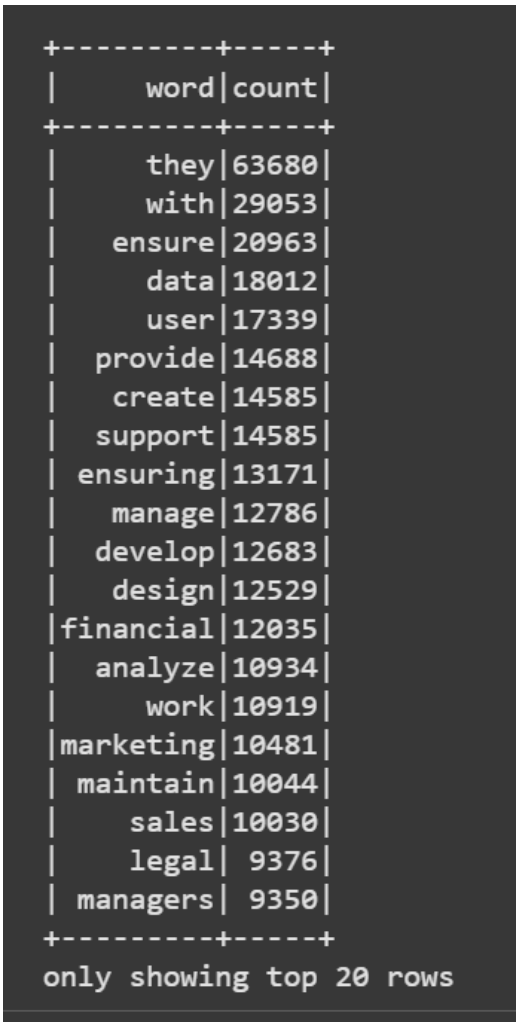
location	avg(AvgExp)
Manama	7.215767634854772
City of Baghdad	7.169715447154472
Victoria	7.169338677354709
Sarajevo	7.165254237288136
Ljubljana	7.163398692810458
Copenhagen	7.155097613882863
Warsaw	7.1487964989059085
Lilongwe	7.146534653465347
Torshavn	7.146484375
Paris	7.139583333333333
Brussels	7.133192389006343
Macao	7.129511677282378
Tarawa	7.129422718808193
Ankara	7.117647058823529
Monrovia	7.116255144032922
London	7.115913555992141
Ulaanbaatar	7.111439842209073
Yamoussoukro	7.100867678958785
Zagreb	7.096267190569745
Belgrade	7.095435684647303

18. **Monthly Posting Trends**:
Extracts the month from job posting dates and visualizes posting frequency per month.



PostMonth	count
NULL	105902

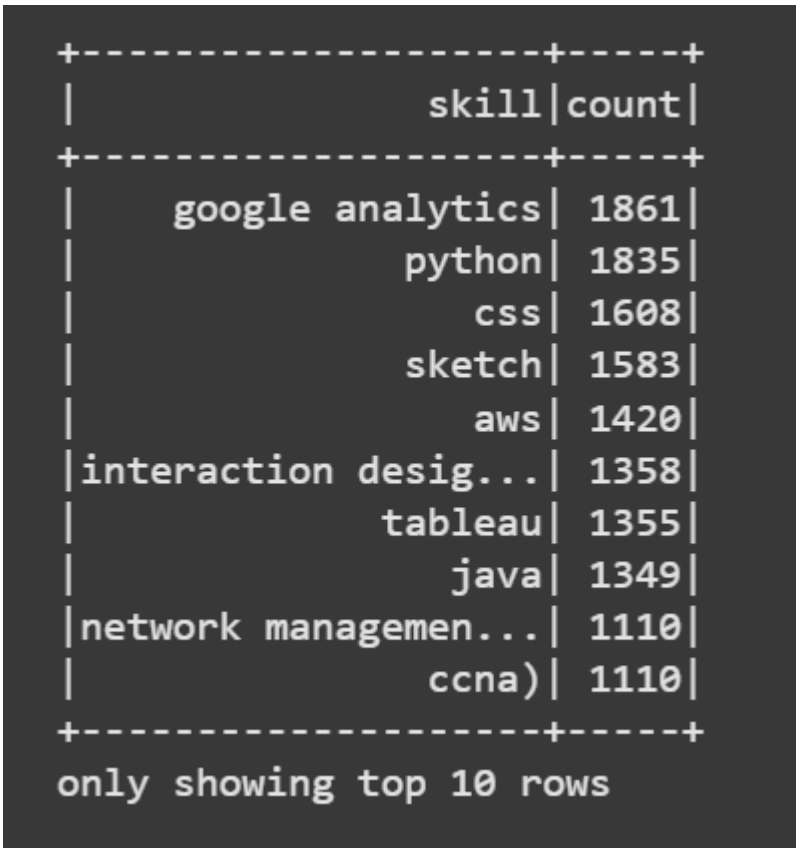
19. **Top Words in Job Descriptions**:
Tokenizes and counts most frequent words (excluding short terms) used in job descriptions, aiding in keyword trends.



word	count
they	63680
with	29053
ensure	20963
data	18012
user	17339
provide	14688
create	14585
support	14585
ensuring	13171
manage	12786
develop	12683
design	12529
financial	12035
analyze	10934
work	10919
marketing	10481
maintain	10044
sales	10030
legal	9376
managers	9350

only showing top 20 rows

20. **Top Skills Used in Listings**:
    Splits the skills column into individual tokens, then counts and ranks the most mentioned skills.



```
+--------------------+-----+
|               skill|count|
+--------------------+-----+
|    google analytics| 1861|
|              python| 1835|
|                 css| 1608|
|              sketch| 1583|
|                 aws| 1420|
|interaction desig...| 1358|
|             tableau| 1355|
|                java| 1349|
|network managemen...| 1110|
|               ccna)| 1110|
+--------------------+-----+
only showing top 10 rows
```

9. Predictive Modeling

To estimate job salaries based on available attributes, we implemented several regression models using **Scikit-learn** and **XGBoost**. The goal was to explore the feasibility of salary prediction using cleaned and transformed features such as experience, job role, and qualifications.

We used the following models:

1. **Linear Regression**
   A basic regression model that assumes a linear relationship between features and target (salary). It served as a baseline for

comparison.

```
Fitting 5 folds for each of 1 candidates, totalling 5 fits
Mean Squared Error: 77168302.79771788
R-squared (R2) Score: -0.21289376403654253
```

2. **Lasso Regression**
   A regularized version of linear regression that adds L1 penalty to reduce overfitting by shrinking less relevant feature weights to zero.

```
Fitting 5 folds for each of 3 candidates, totalling 15 fits
Mean Squared Error: 109946969.95651309
R-squared (R2) Score: -0.7113923795940452
```

3. **Ridge Regression**
   Similar to Lasso but uses L2 regularization to penalize large coefficients, improving generalization on unseen data.

```
Fitting 5 folds for each of 3 candidates, totalling 15 fits
Mean Squared Error: 77168311.79869996
R-squared (R2) Score: -0.21285008597413585
```

4. **Decision Tree Regressor**
   A non-linear model that splits data into subsets based on feature thresholds. It can capture complex patterns but risks overfitting without pruning.

```
Fitting 5 folds for each of 36 candidates, totalling 180 fits
Mean Squared Error: 92550547.29618277
R-squared (R2) Score: -0.14638357318663475
```

5. **Random Forest Regressor**
   An ensemble of decision trees that averages predictions to reduce variance and improve robustness. It performed best among all

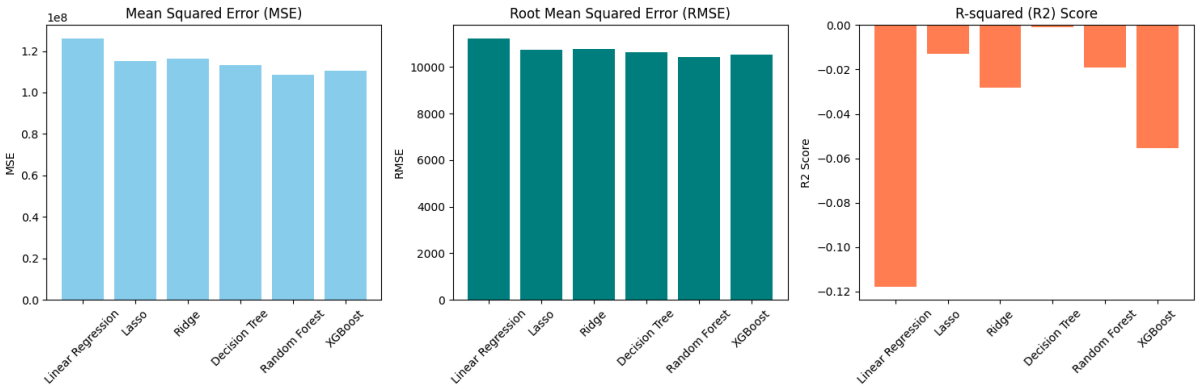
models in terms of RMSE.

```
Fitting 5 folds for each of 16 candidates, totalling 80 fits
Mean Squared Error: 74750951.24415363
R-squared (R2) Score: -0.040530196427054244
```

6. **XGBoost Regressor**
   A gradient-boosting technique that builds trees sequentially, each correcting the previous one. Though powerful, it underperformed in our case due to limited feature richness.

```
Fitting 5 folds for each of 32 candidates, totalling 160 fits
Mean Squared Error: 81316184.0
R-squared (R2) Score: -0.23985642194747925
```

Each model was trained on a feature-engineered dataset, and performance was evaluated using **MSE (Mean Squared Error)**, **RMSE (Root Mean Squared Error)**, and **R² (R-squared)** score. While **Random Forest** showed the lowest error, **all models had low R² scores**, indicating that salary prediction remains difficult with the given dataset due to the absence of key influencing features such as company level, benefits, or detailed job responsibilities.



10. Evaluation and Findings

Model performance was assessed using standard regression metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 (coefficient of determination).

- The Random Forest Regressor achieved the lowest RMSE, indicating better prediction accuracy compared to others
- However, all models had low R^2 scores, showing weak explanatory power and highlighting limitations in the feature set.
- Results suggest that while the models could capture partial patterns, predicting salary from the current features alone is not sufficient for high accuracy.

11. Challenges Faced

- **Inconsistent Formatting**: Fields like salary, experience, and qualifications had varied formats (e.g., ranges, free-text).

- **Missing Values**: Several rows had incomplete or null fields in key columns, requiring preprocessing decisions like row removal or imputation.

- **Unstructured Data**: Many features were in free-text format (e.g., job description, skills), which required tokenization and cleaning.

- **Limited Predictive Features**: Absence of company-level metrics, benefits quantification, or detailed role expectations made salary modeling difficult.

- **Resource Constraints**: Large data size and Spark initialization in Colab occasionally caused slowdowns or timeouts.

12. Conclusion

This project successfully demonstrated the use of **big data tools like PySpark** to process and analyze a large semi-structured dataset of job descriptions.
We identified trends in job roles, skill demand, qualifications, and salary patterns across different locations and companies. While **EDA revealed valuable hiring insights**, predictive modeling for salary estimation showed limited success due to data complexity and missing context. Overall, the project enhanced our understanding of real-world job market trends and strengthened our skills in distributed data processing and exploratory analysis.

13. Future Scope

- **Add More Features**: Incorporate external company-level data (size, revenue, Glassdoor ratings) to improve salary predictions.
- **Natural Language Processing (NLP)**: Apply NLP techniques to extract deeper patterns from job descriptions and responsibilities.
- **Classification Models**: Explore classification tasks such as predicting job category, seniority level, or required skill clusters.
- **Interactive Dashboard**: Build a web-based dashboard using Streamlit or Dash for real-time filtering and visualization of job data.

14. References

- Rana, R. (2023). *Job Description Dataset*. Kaggle. https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset

- Apache Software Foundation. (2023). *Apache Spark Documentation*. https://spark.apache.org/docs/latest/

- Scikit-learn Developers. (2023). *Scikit-learn: Machine Learning in Python*. https://scikit-learn.org/stable/

- Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD.

- Google. (2023). *Google Colaboratory*. https://colab.research.google.com/