Big Data Project: Job Description Analysis

ITC 686, Spring 2025.

Exploring big data techniques to analyze global job market trends using PySpark.

Dhanalakshmi - kannu1d

Vyom Patel -patel5v

Prathyusha – davan1p

Project Overview

Goal

Analyze job descriptions to identify key skills, roles, experience levels, and geographic trends.

Why It Matters

Job postings provide insights into evolving market demands and large-scale hiring patterns.

Dataset Description

Source & Size

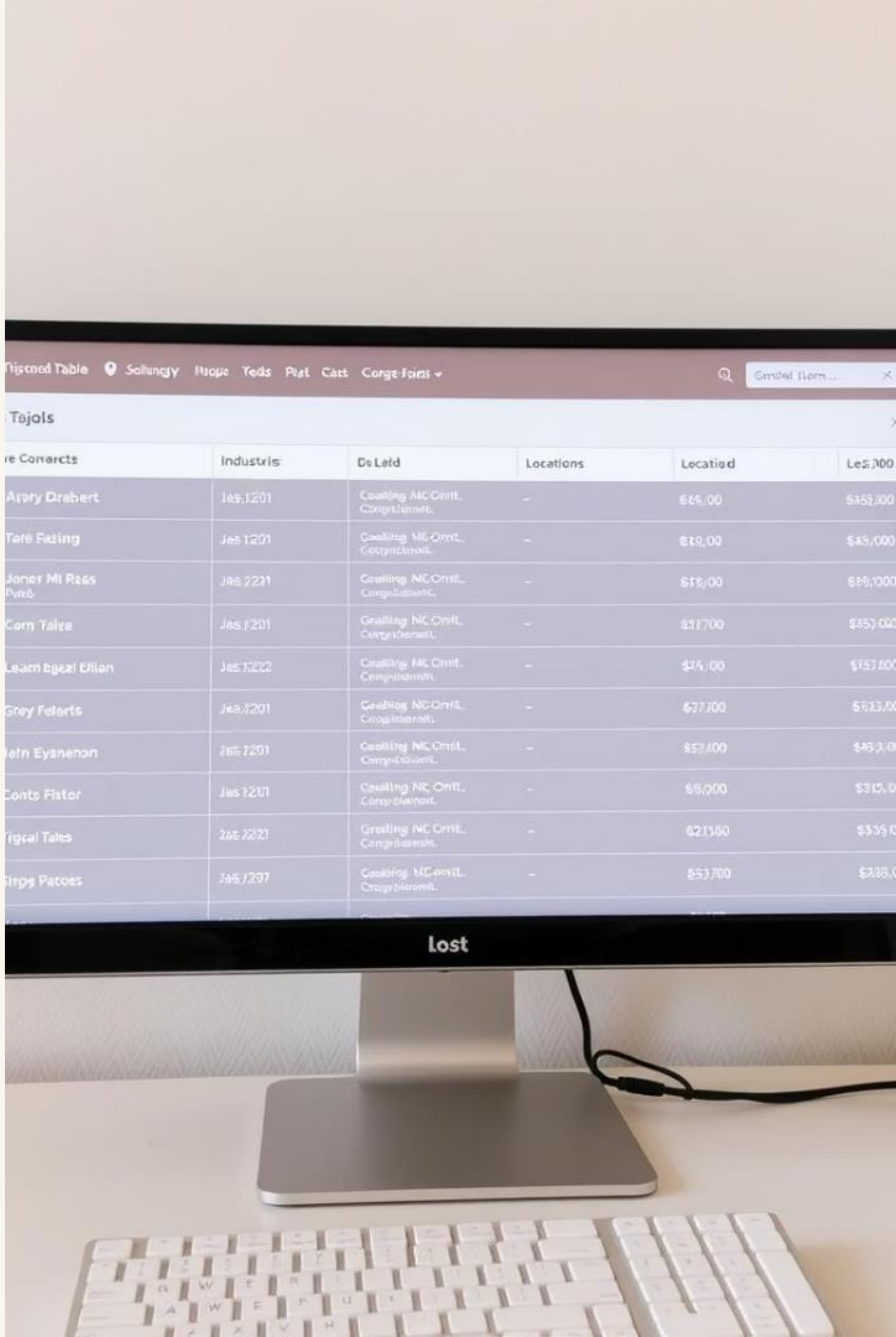
Kaggle dataset with 13,000+ rows, 23 fields, 1.5GB in size.

Key Fields

- Job Id, Title, Description
- Skills, Experience, Salary
- Location, Industry, Company Info

Scope

Global jobs across various sectors, capturing rich hiring data.



Tools and Environment

Platform & Language

Apache Spark with PySpark enables distributed large dataset processing.

Execution Environment

Google Colab configured for PySpark ensures flexible cloud computation.

Benefits

- Fast parallel processing
- APIs for structured & unstructured data
- Handles big data efficiently

Data Cleaning and Preparation

Cleaning Steps

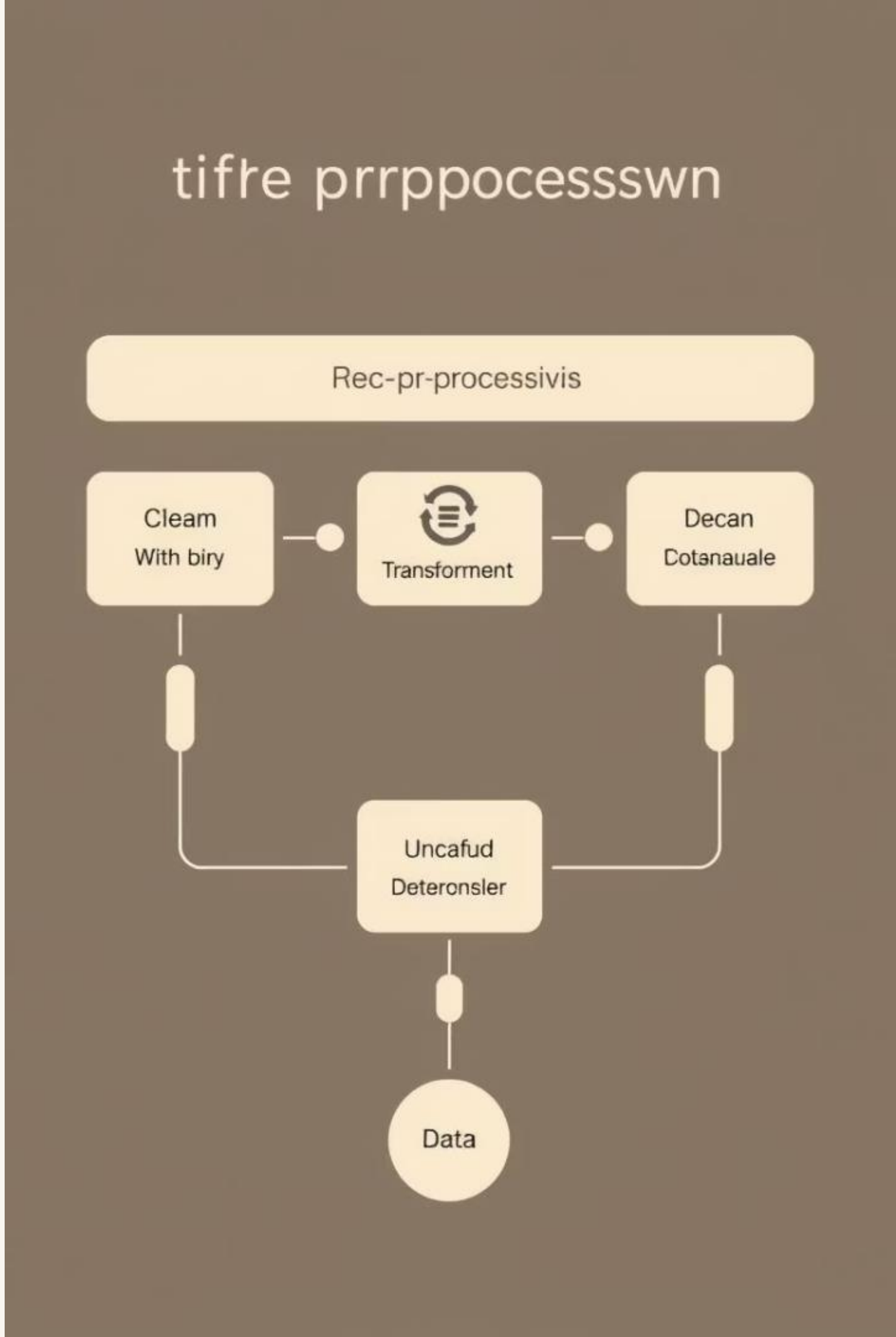
Dropped rows missing critical data, lowered case, trimmed whitespace.

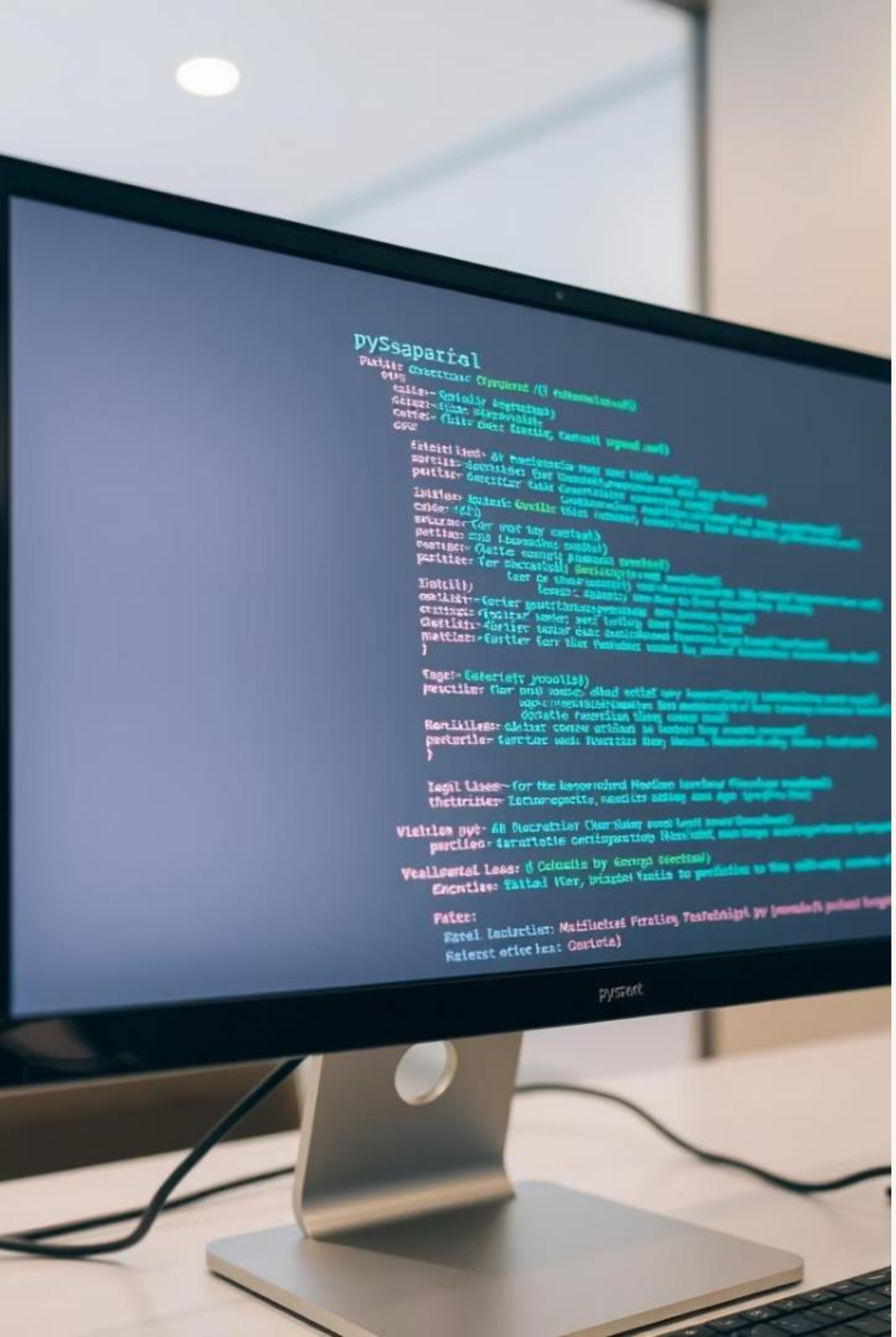
Deduplication & Features

Removed duplicates by Job Id; created average experience column.

Text Processing

Tokenized skills field for frequency and trend analysis.





Query Structure

1

Simple Queries [5]

Count, distinct, null checks for initial data integrity.

2

Moderate Queries [7]

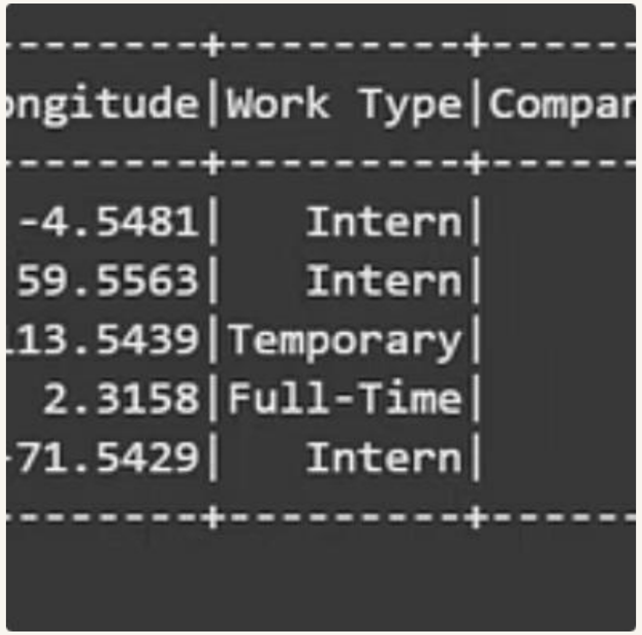
Filtering and grouping by locations and job titles.

3

Complex Queries [8]

Skill extraction, joins, trends in experience and salary.

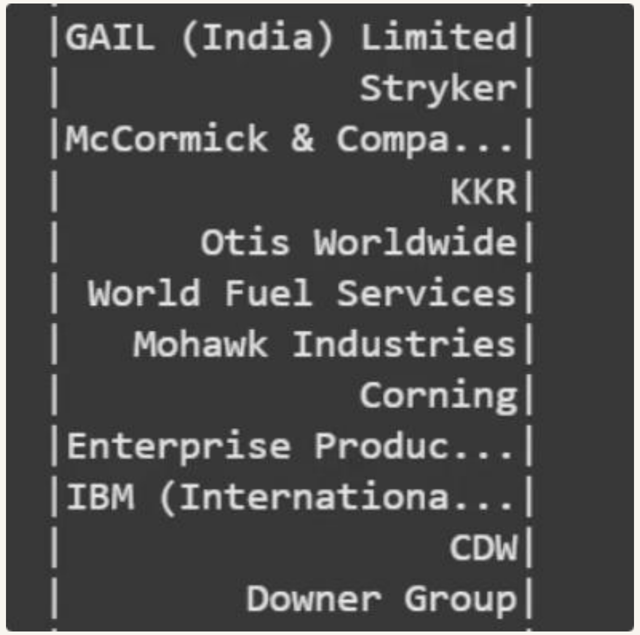
Simple Queries Insights



```
--------+---------+-----
ngitude|Work Type|Compar
--------+---------+-----
 -4.5481|   Intern|
 59.5563|   Intern|
 13.5439|Temporary|
  2.3158|Full-Time|
 71.5429|   Intern|
--------+---------+-----
```

Total Records Verified

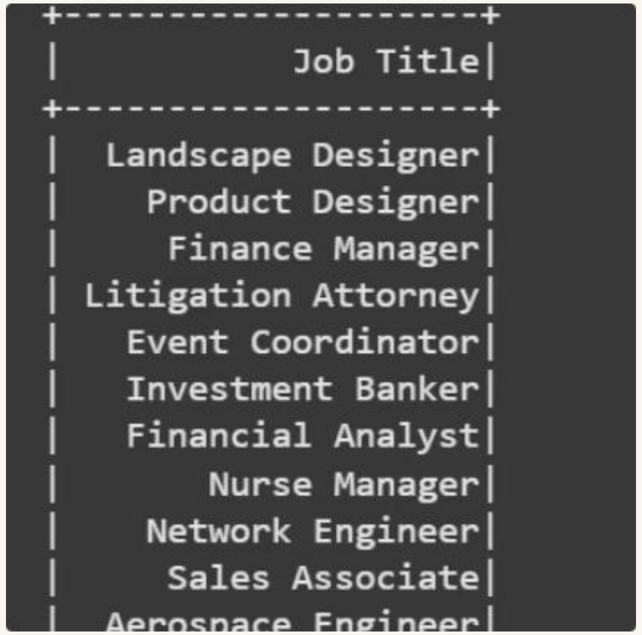
Confirmed that the total number of records matches the expected dataset size.



```
|GAIL (India) Limited|
|             Stryker|
|McCormick & Compa...|
|                 KKR|
|     Otis Worldwide|
| World Fuel Services|
|   Mohawk Industries|
|             Corning|
|Enterprise Produc...|
|IBM (Internationa...|
|                 CDW|
|        Downer Group|
```

Unique Companies

Calculated the total count of distinct company names in the dataset.

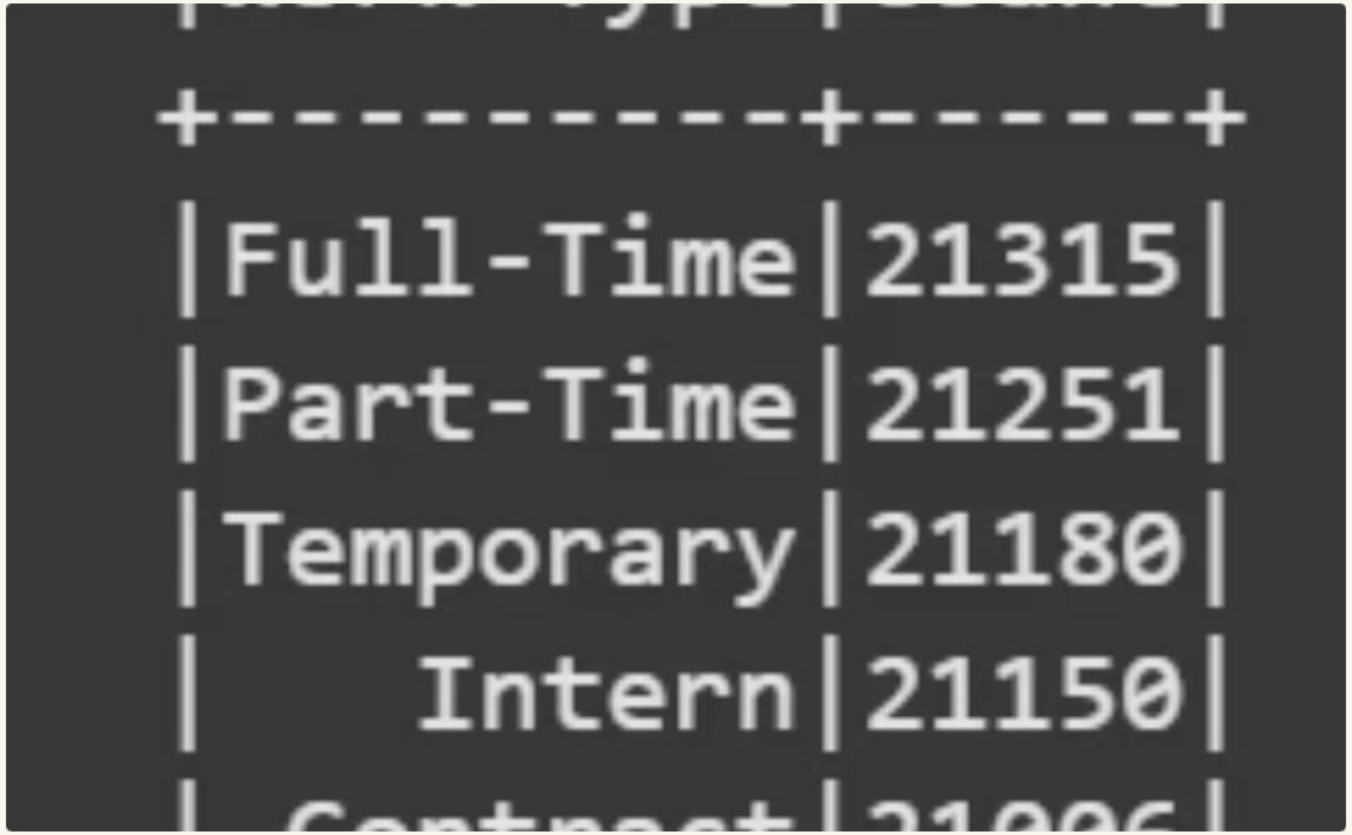


```
+--------------------+
|           Job Title|
+--------------------+
|  Landscape Designer|
|    Product Designer|
|     Finance Manager|
| Litigation Attorney|
|   Event Coordinator|
|    Investment Banker|
|   Financial Analyst|
|       Nurse Manager|
|    Network Engineer|
|     Sales Associate|
|  Aerospace Engineer|
```

Distinct Job Titles

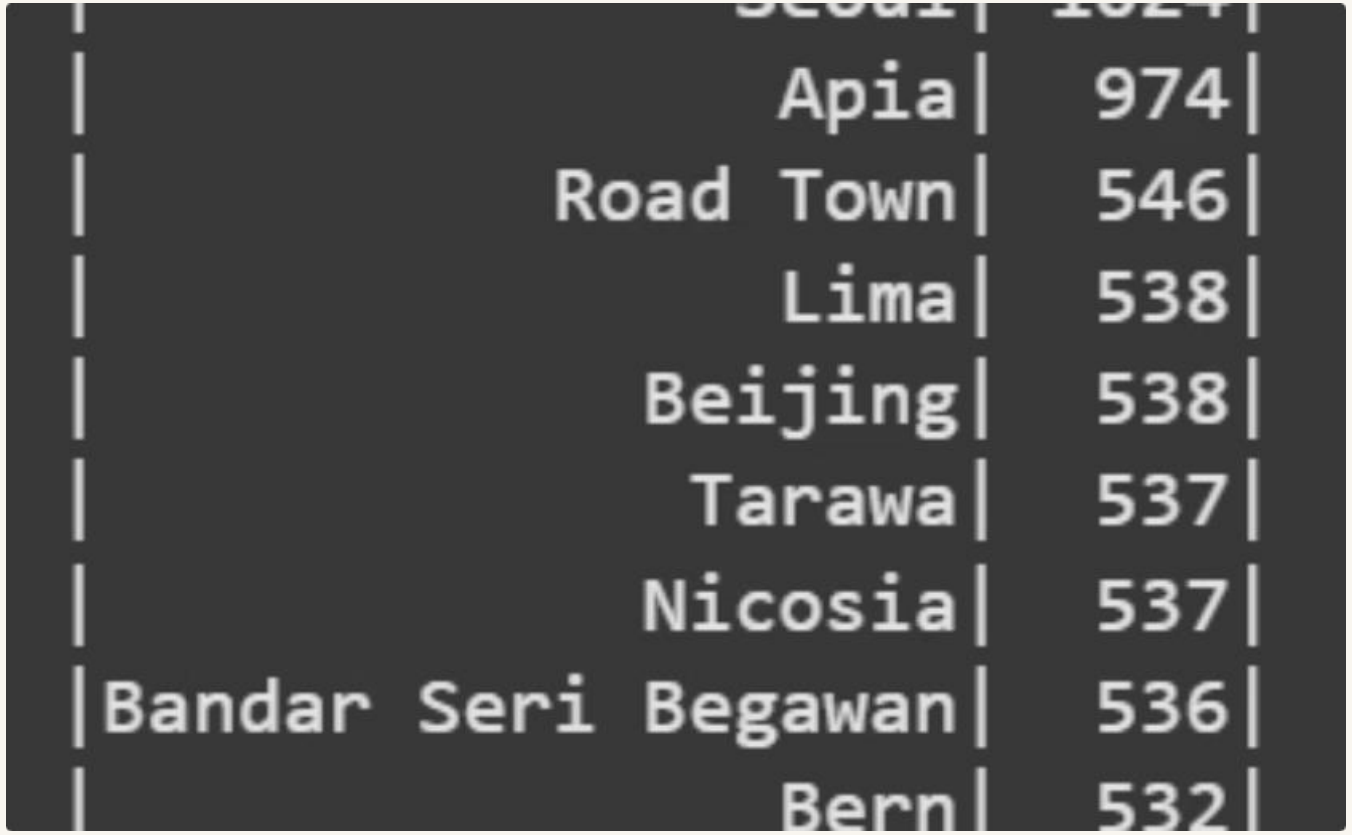
Determined the number of unique job titles present across all records.

Moderate Queries Insights



```
|work_type|count|
+---------+-----+
|Full-Time|21315|
|Part-Time|21251|
|Temporary|21180|
|   Intern|21150|
| Contract|21006|
```

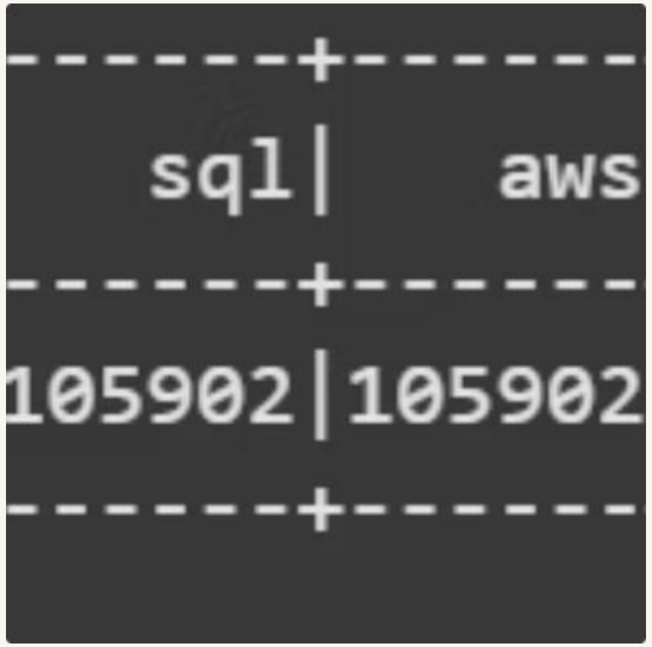
Job Counts by Different Work Types



```
|               Seoul| 1024|
|                Apia|  974|
|           Road Town|  546|
|                Lima|  538|
|             Beijing|  538|
|              Tarawa|  537|
|             Nicosia|  537|
|Bandar Seri Begawan|  536|
|                Bern|  532|
```

Number of Job Postings by Location

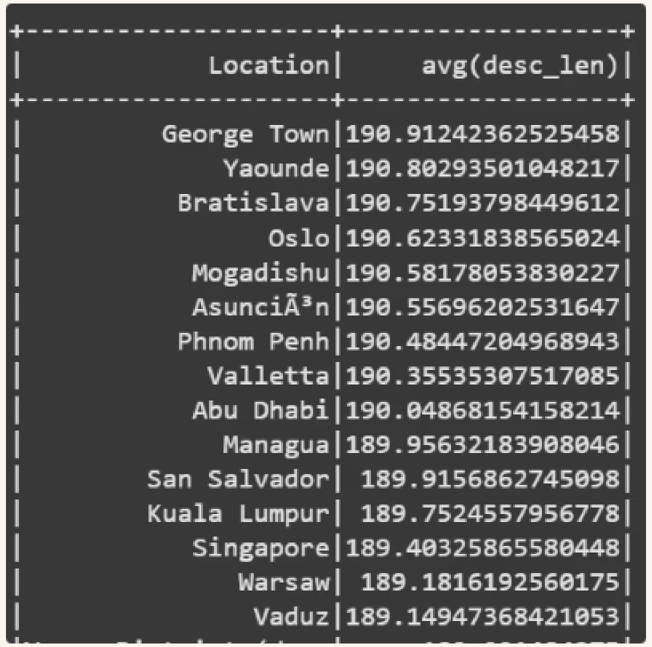
Complex Queries Insights



```
------+-------
   sql|    aws
------+-------
105902|105902
------+-------
```

Popular Skills Extraction

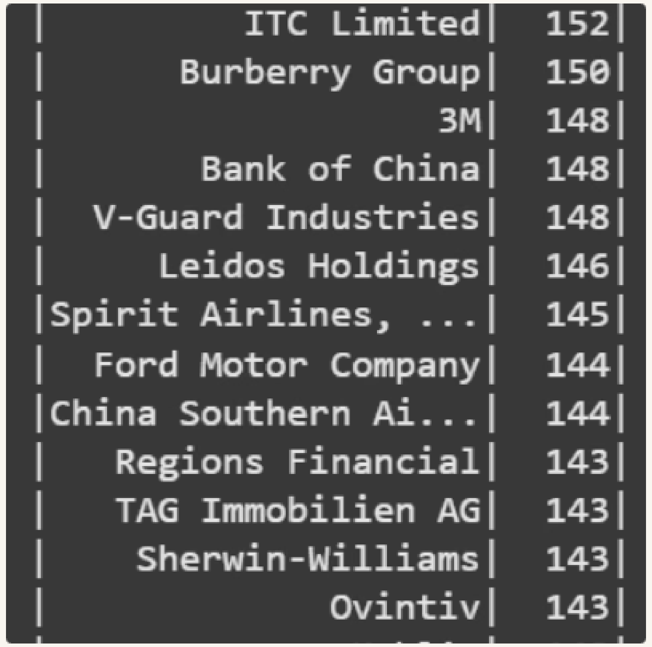
Extracted the most common skills from job descriptions.



```
+--------------------+------------------+
|            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|
```

Average Experience by City

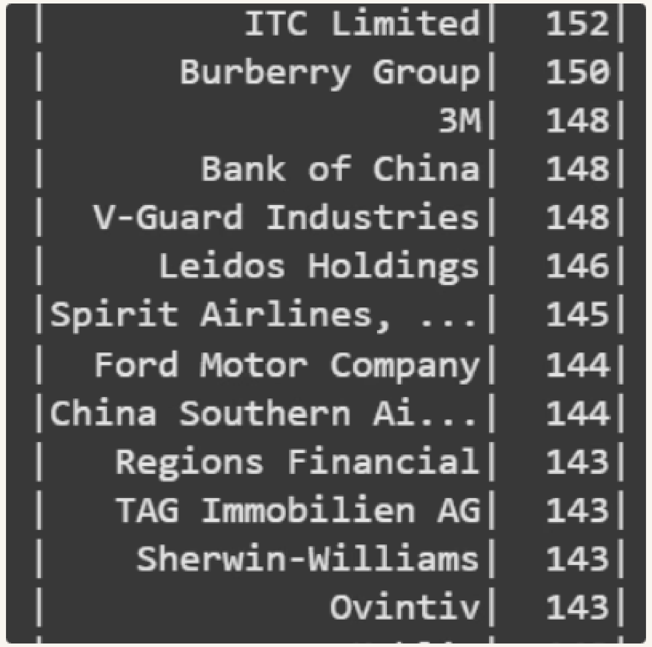
Calculated average experience levels for jobs in different cities.



```
|         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|
```

Hiring Trends Analysis

Examined monthly trends in job postings to identify patterns.



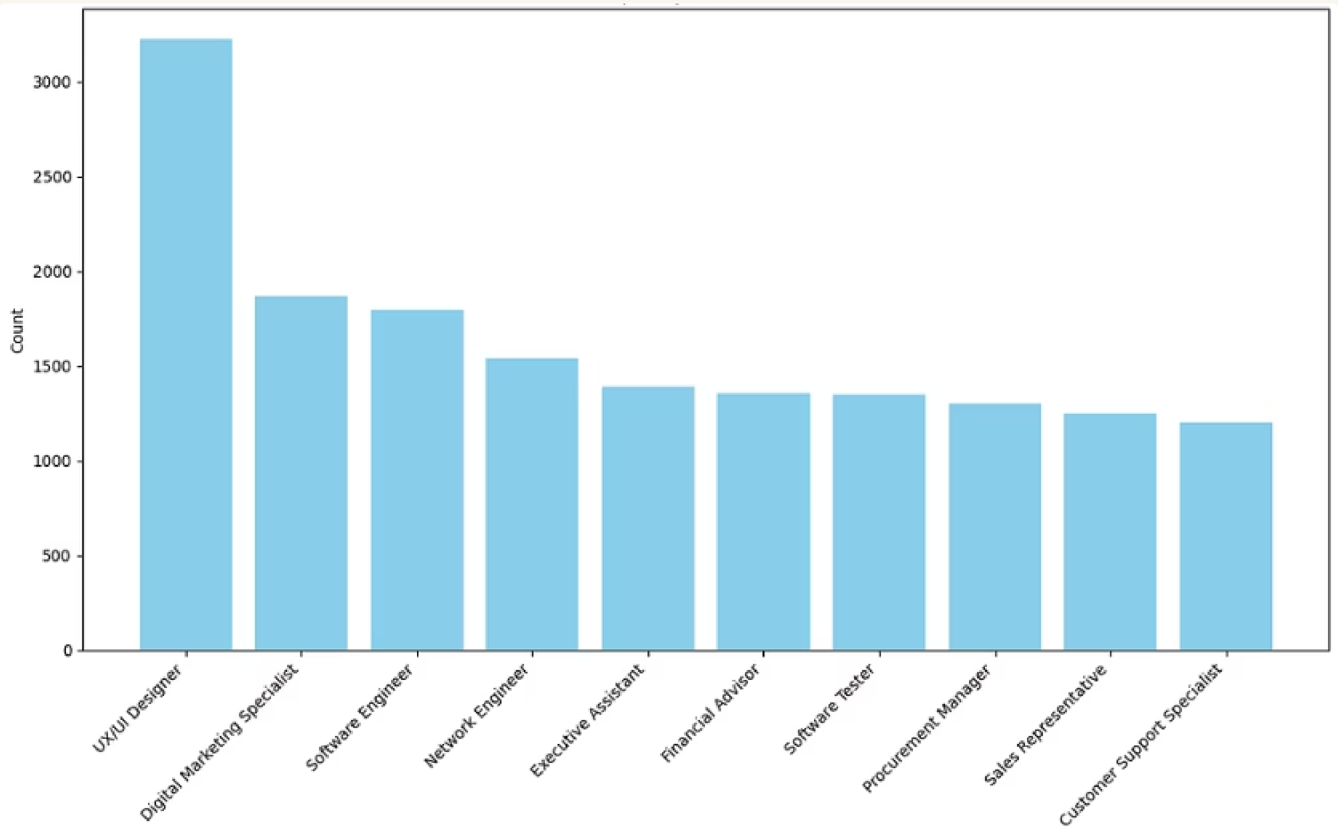
```
|         ITC Limited|   152|
|      Burberry Group|   150|
|                  3M|   148|
|       Bank of China|   148|
|  V-Guard Industries|   148|
|     Leidos Holdings|   146|
|Spirit Airlines, ...|   145|
|  Ford Motor Company|   144|
|China Southern Ai...|   144|
|   Regions Financial|   143|
|    TAG Immobilien AG|   143|
|    Sherwin-Williams|   143|
|             Ovintiv|   143|
```

Skills Frequency Breakdown

Exploded skills text fields to analyze frequency of each skill.

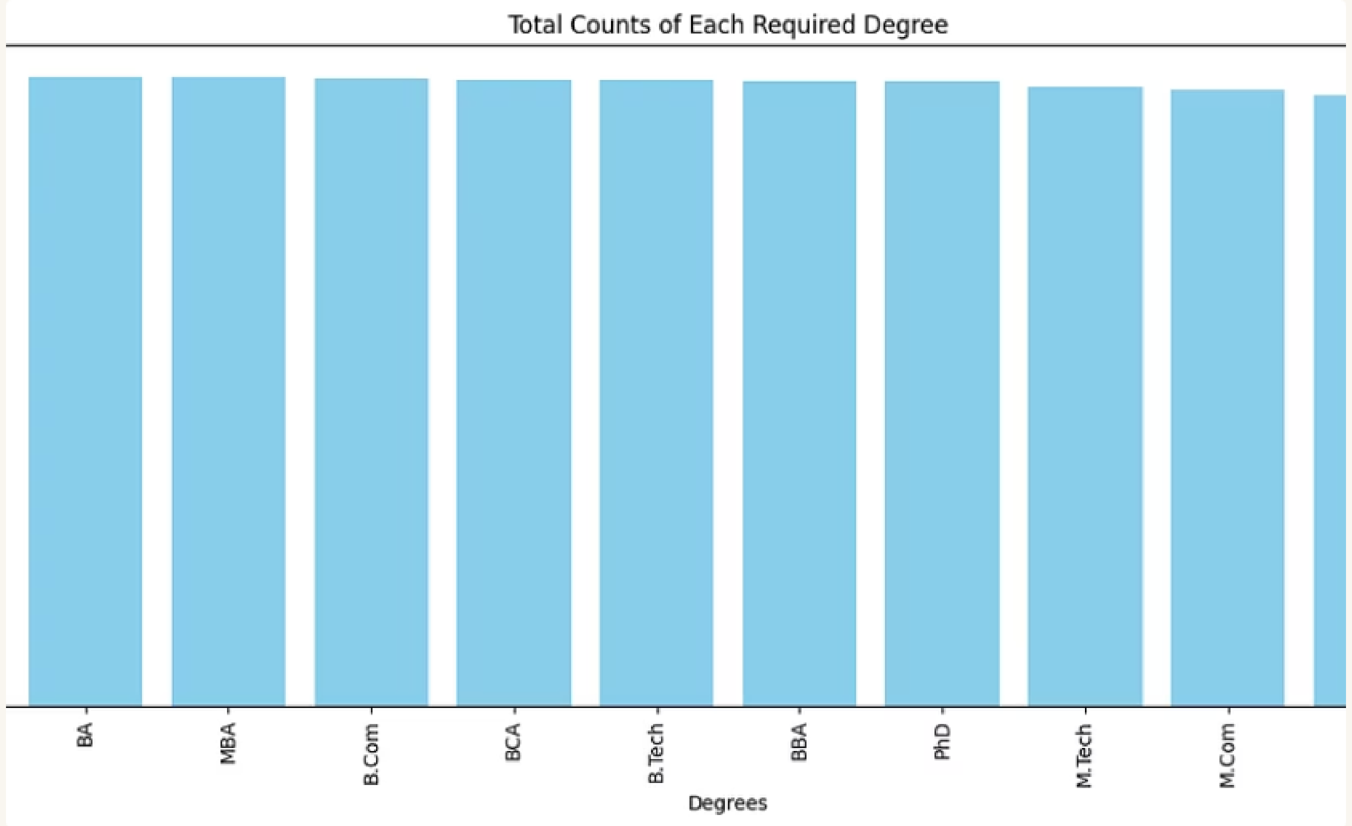
Made with GAMMA

Exploratory Data Analysis: Top Job Titles



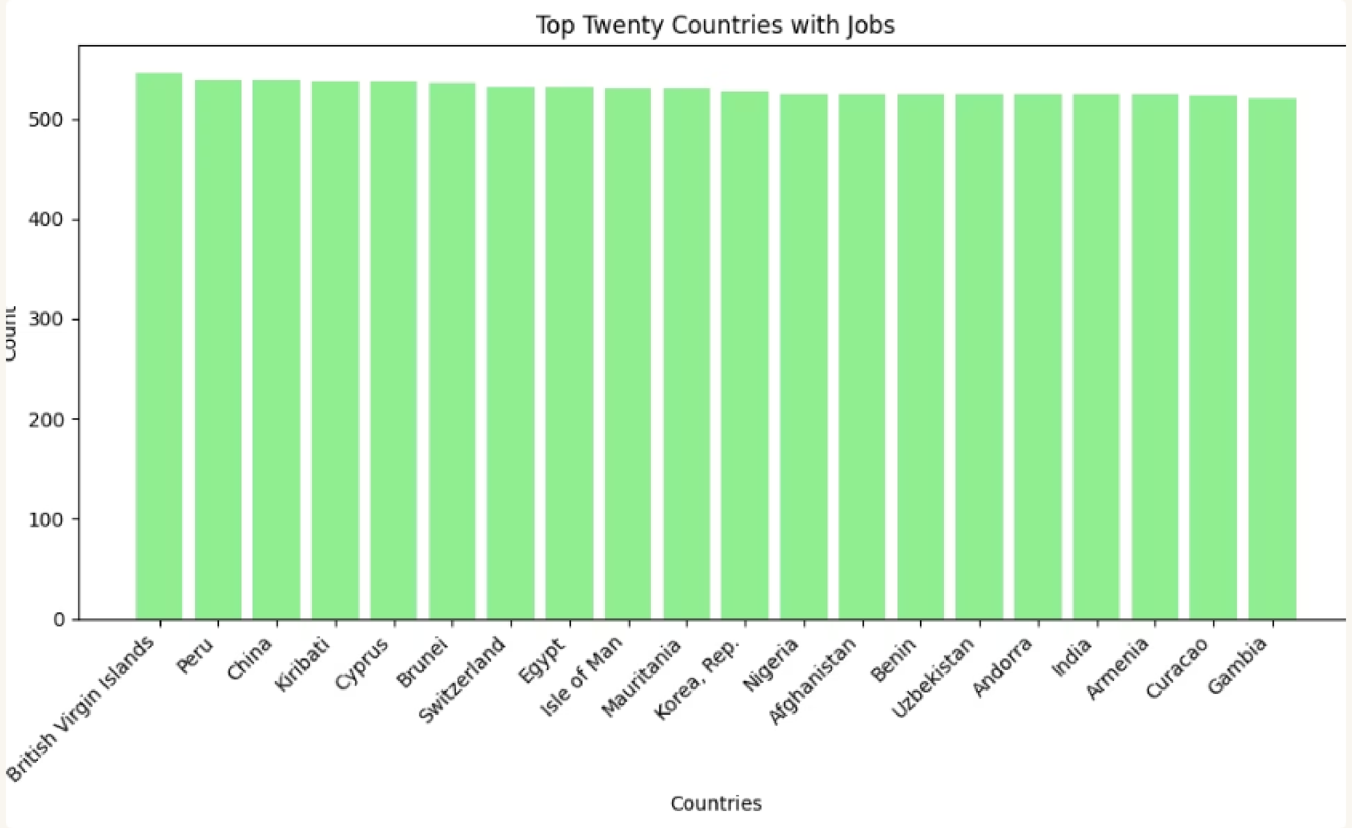
Top Ten Jobs in Market

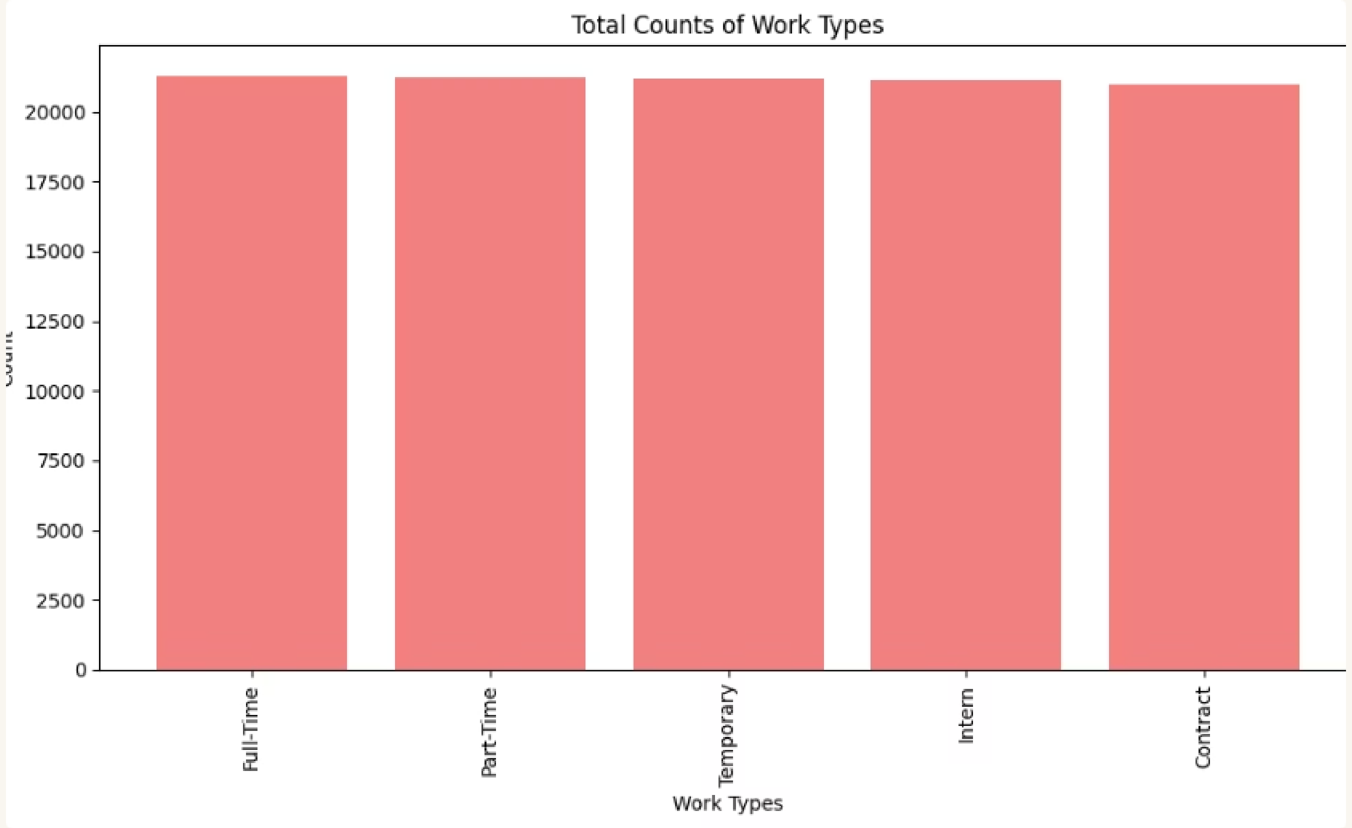
UX/UI Designer is the most in-demand role, followed by Digital Marketing Specialist and Software Engineer.



Total Counts of Each Required Degree

Degrees like BA, MBA, B.Com, and B.Tech are most frequently requested.



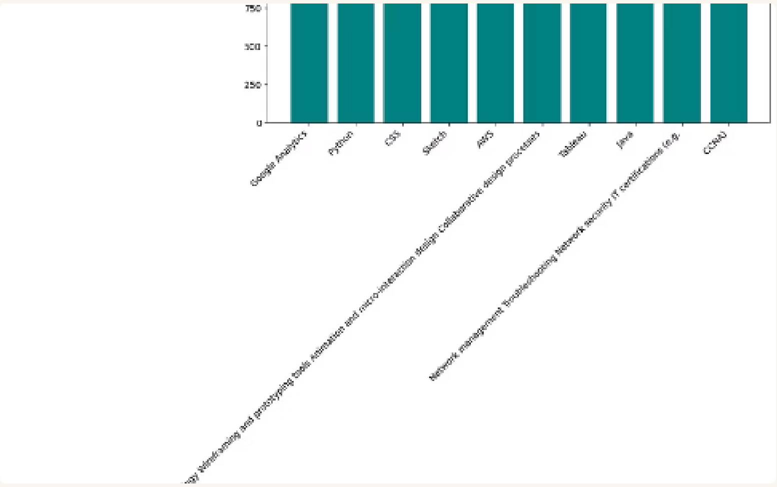


Top Twenty Countries with Jobs

Countries like the British Virgin Islands, Peru, China, and India are among the top 20 with the most listings.

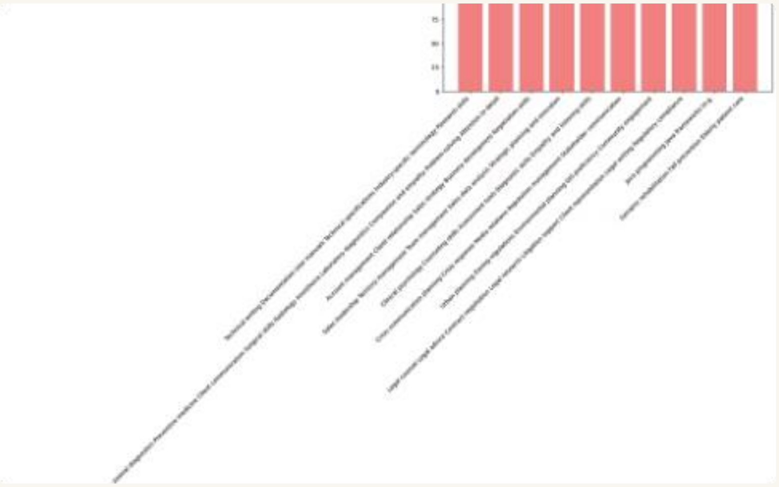
Total Counts of Work Types

Job types are fairly evenly distributed across Full-Time, Part-Time, Intern, and Contract roles.



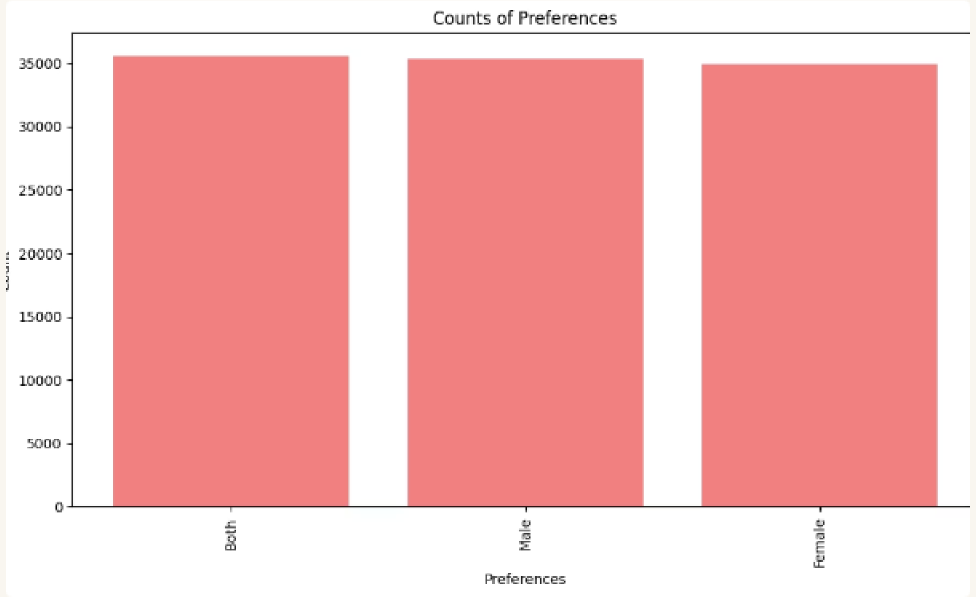
Top 10 Most Demanding Skills

Google Analytics, Python, CSS, AWS, and Tableau top the list of high-demand skills.



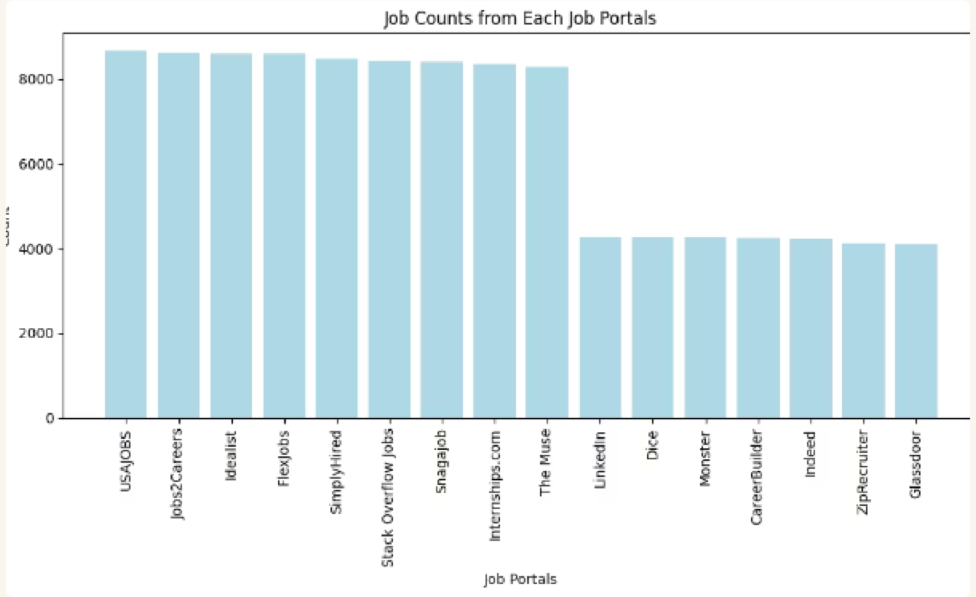
Least 10 Demanding Skills

Skills like Veterinary practices, Civic communication, and Legal documentation appear least often.



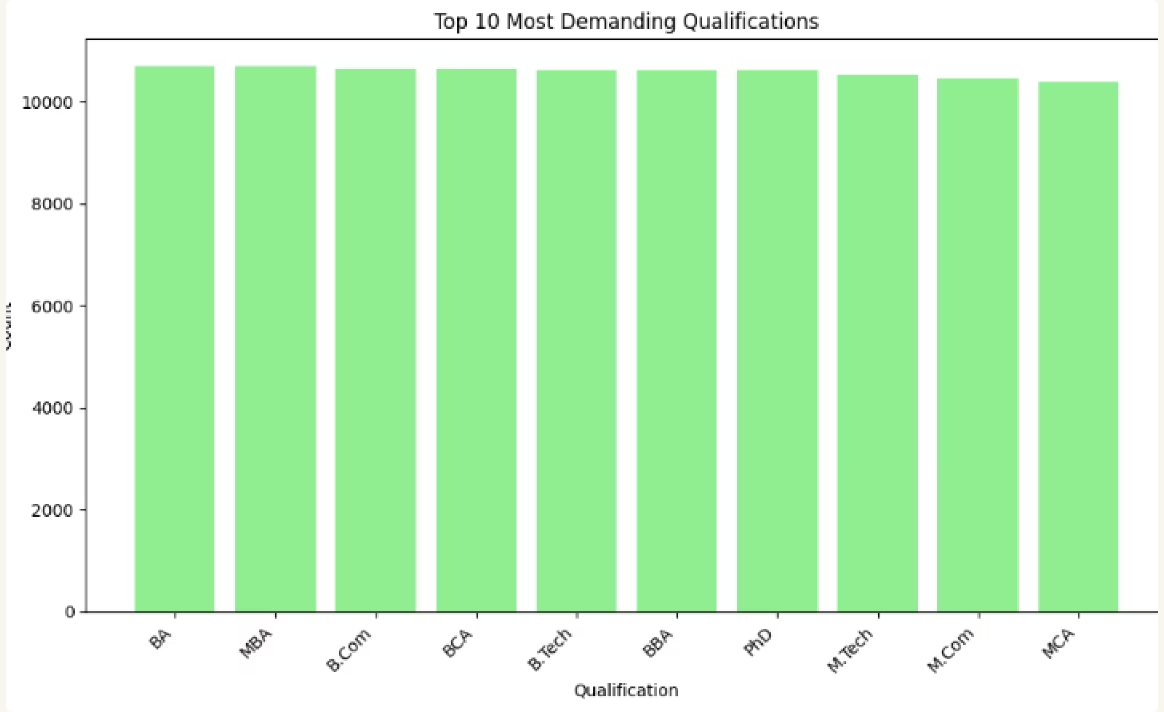
Counts of Preferences

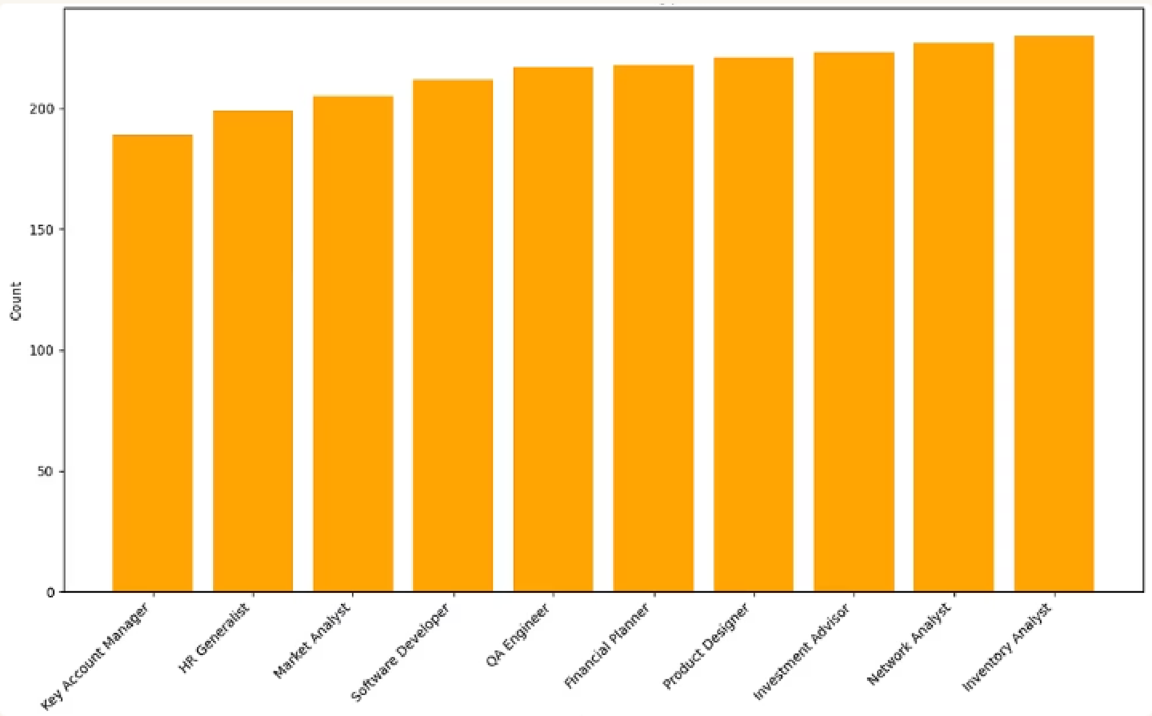
Most employers selected "Both" for gender preference, indicating inclusivity.



Job Counts from Each Job Portal

USAJOBS, Jobs2Careers, and Idealist are the top platforms with the highest job postings.



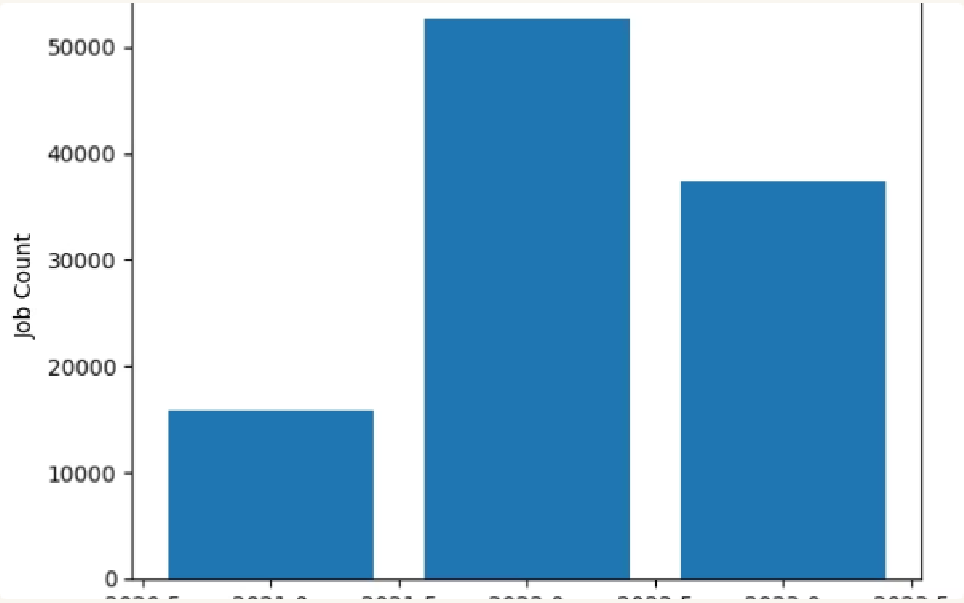


Top 10 Most Demanding Qualifications

Degrees like BA, MBA, and B.Tech are in high demand.

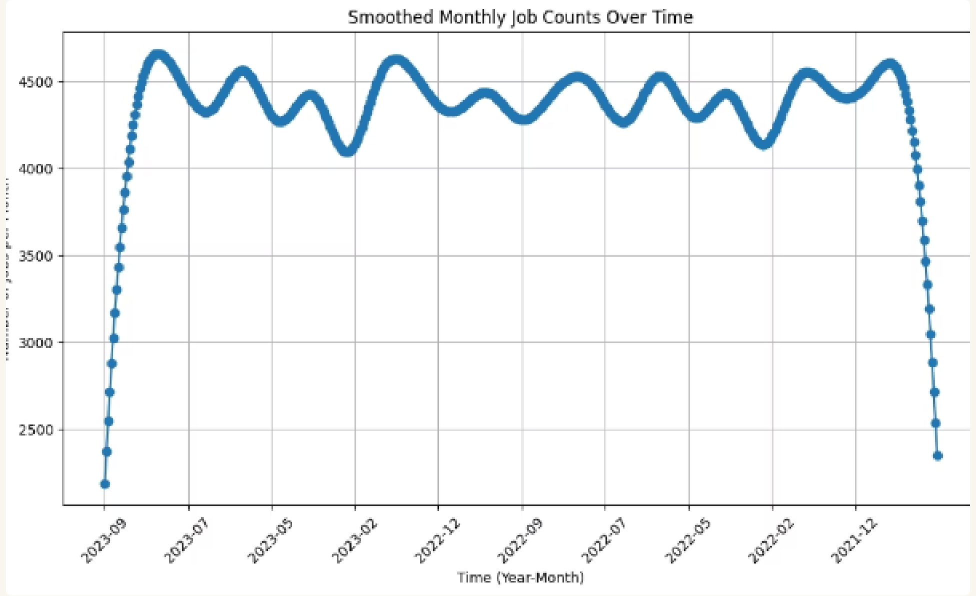
10 Least Demanding Jobs

Roles like QA Engineer and Network Analyst are least posted.



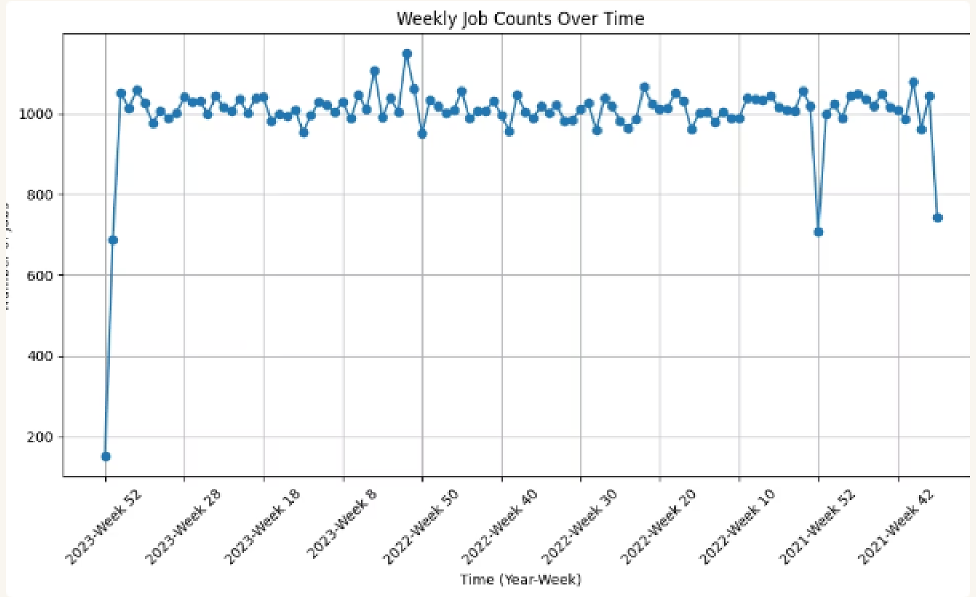
Yearly Job Count Distribution

2022 had the highest job count, followed by a drop in 2023.



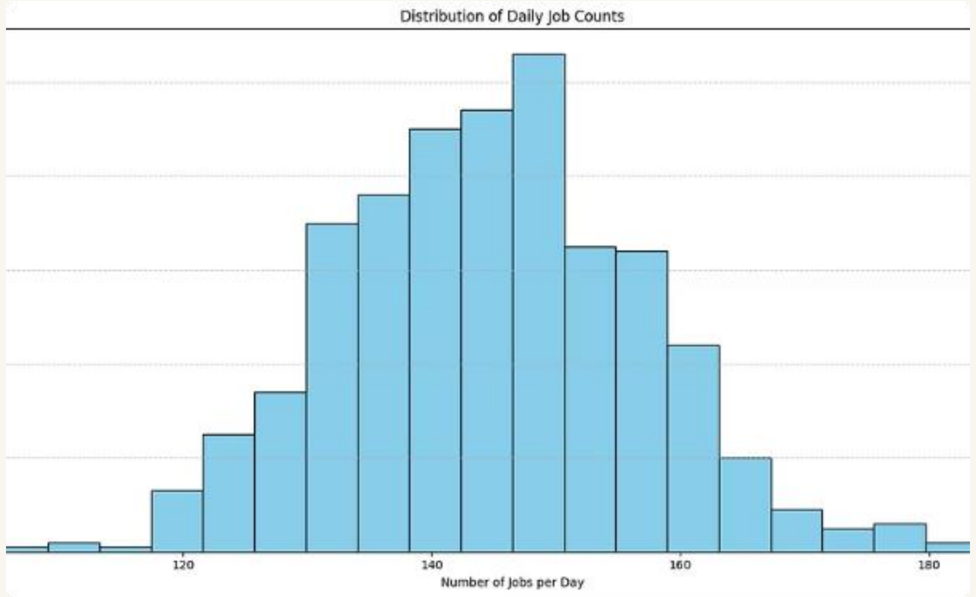
Smoothed Monthly Job Counts

Monthly patterns show repeated hiring cycles.



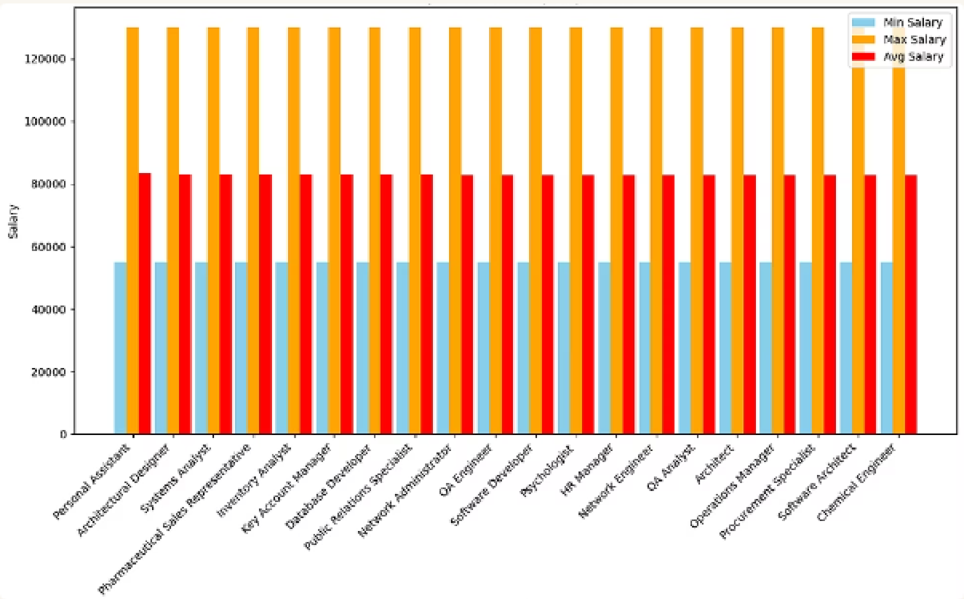
Weekly Job Counts Over Time

Weekly job counts remain mostly consistent.



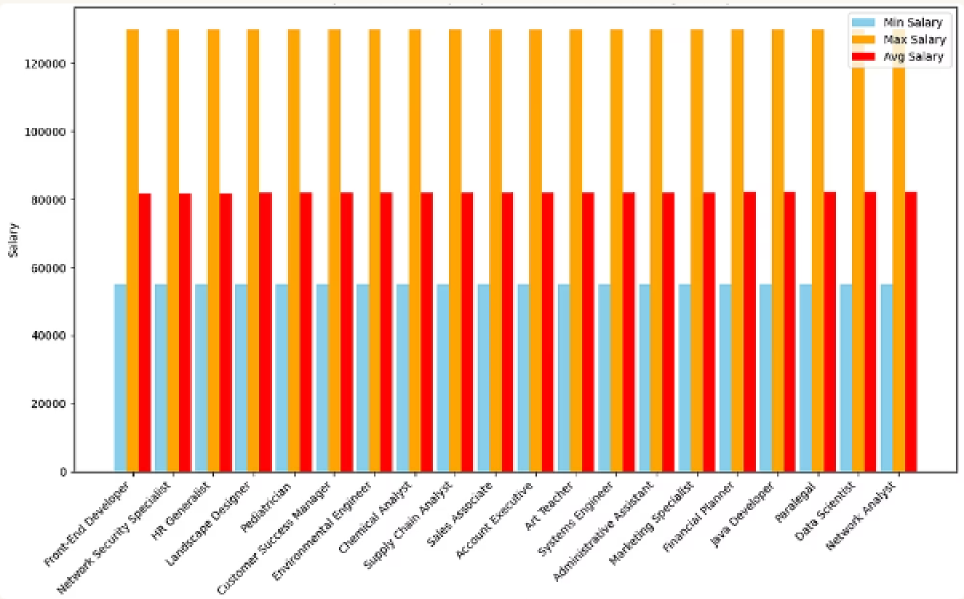
Distribution of Daily Job Counts

Most days have 135–155 job listings.



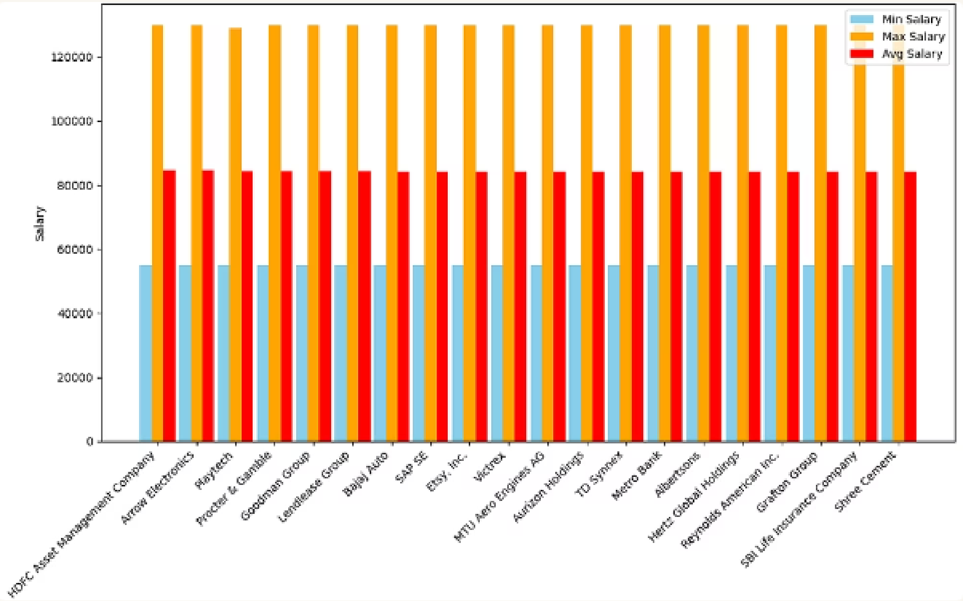
Jobs with Highest Average Salary

Software Architect, Operations Manager, and QA Analyst are among the highest average-paid roles.



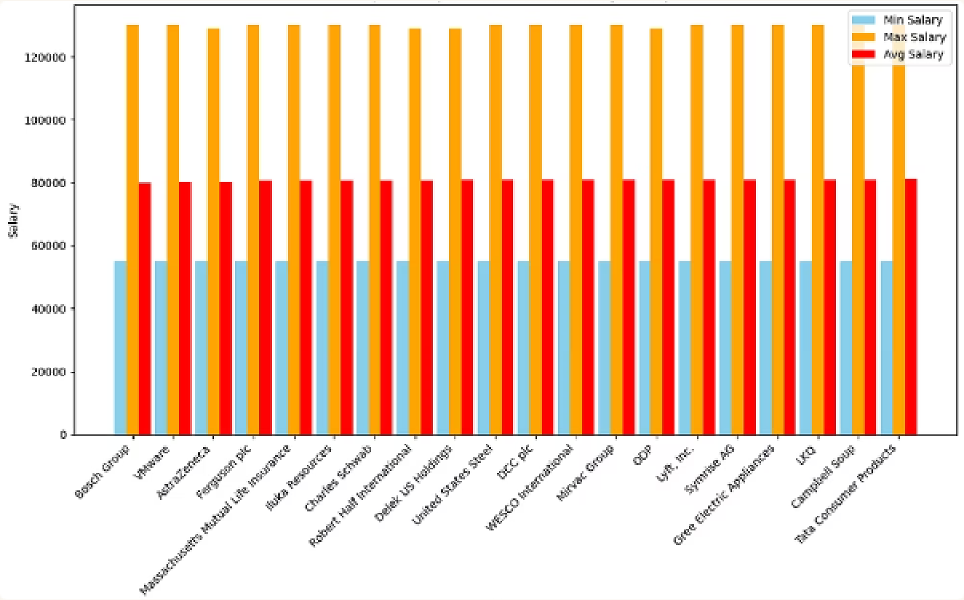
Jobs with Lowest Average Salary

Roles like Customer Success Manager and HR Generalist are among the lowest-paying on average.



Top 20 Companies with Highest Average Salary

HDFC, Arrow Electronics, and Procter & Gamble rank among the top payers.oles.



Top 20 Companies with Lowest Average Salary

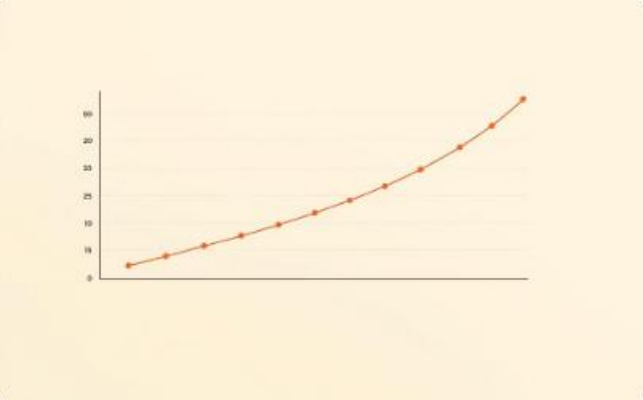
Companies like Bosch Group and Lyft offer the lowest average salaries in the dataset.

Predictive Modeling – Salary Estimation

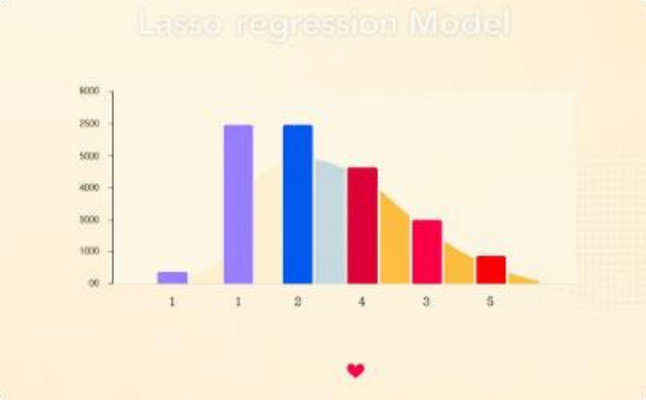
Objective:

Develop a regression model to accurately predict salaries based on various factors including experience, job title, qualifications, and other relevant features.

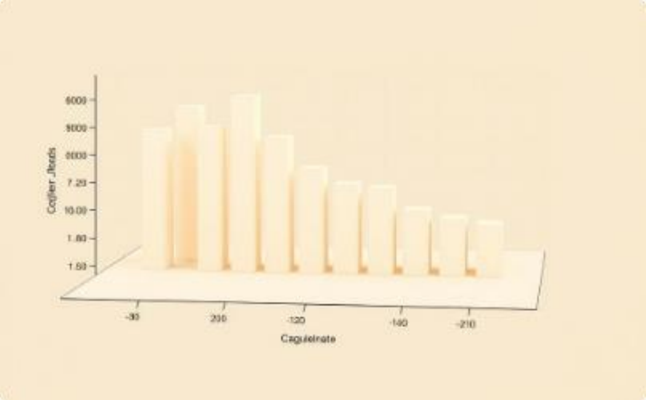
Models Implemented



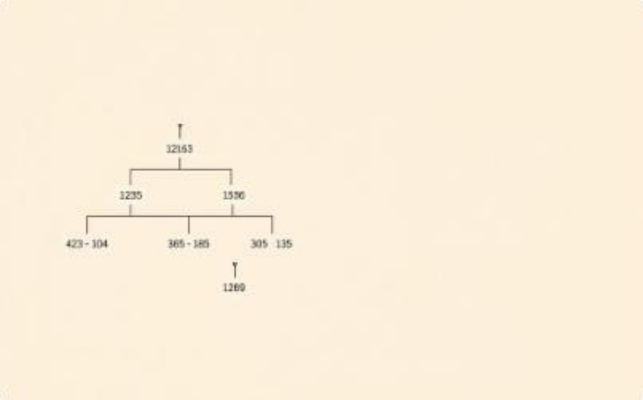
Linear Regression



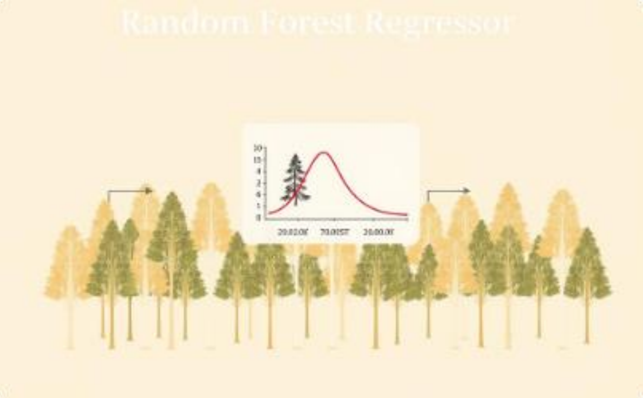
Lasso Regression



Ridge Regression

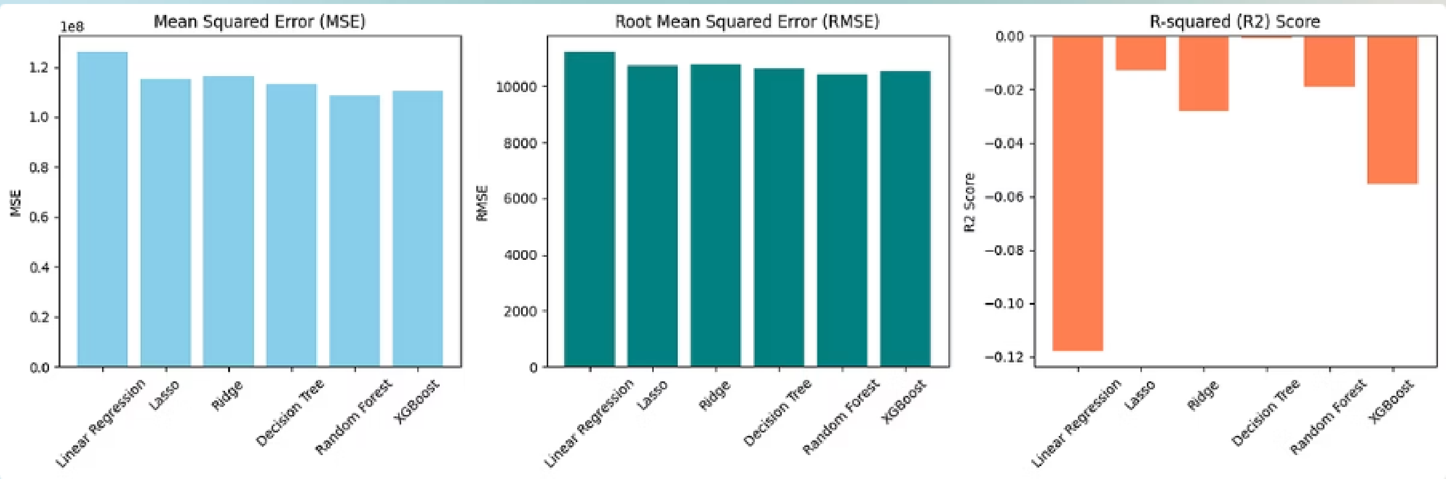


Decision Tree Regressor



Random Forest Regressor

XGBoost Regressor



Model Evaluation:

Metrics Used:

- **MSE (Mean Squared Error)**: Measures average squared difference between predicted and actual values.
- **RMSE (Root Mean Squared Error)**: Gives a better idea of average error magnitude.
- **R^2 Score (Coefficient of Determination)**: Measures goodness-of-fit (closer to 1 is better).

Observation: Random Forest and Decision Tree performed slightly better than linear models, but R^2 values are close to 0 or negative, suggesting poor predictive power overall.

Conclusion & Key Takeaways

- Analyzed large-scale job data using **PySpark**, focusing on roles, skills, locations, and salaries.
- Found **UX/UI Designer**, **Software Engineer**, and **Digital Marketer** to be top-demand roles.
- **BA, MBA, B.Tech** were the most commonly requested qualifications.
- Salary distribution showed minimal variation across countries and companies.
- Implemented regression models (**Linear, Random Forest, XGBoost**) to predict salary.
- **Random Forest performed best**, but overall **R^2 scores were low**, indicating weak predictive power.
- Project demonstrates how **big data tools** and **ML models** can be applied to real-world job market insights.