Job Market Analysis for Data and Software Roles Using Web Scraping and APIs

Nishkarsh Mittal
University of Southern California
Email: nishkars@usc.edu
USC ID: *3625374778*

1 Project Name and Team Members

Project Title: Job Market Analysis for Data and Software Roles Using Web Scraping and APIs

Team Members:

- Nishkarsh Mittal (nishkars@usc.edu, USC ID: *3625374778*)

2 Short Description

This project analyzes trends in the technology job market by collecting and examining job postings related to data science, machine learning, and software engineering roles. Using web scraping and public APIs, job listings were gathered from multiple online job platforms and processed to extract structured information such as job roles, required skills, locations, and salary ranges. The cleaned data was analyzed and visualized to identify regional demand patterns, popular job roles, and skill requirements. The project aims to provide insights into hiring trends and the geographic distribution of technology jobs.

3 Data

3.1 Data Sources

The data for this project was collected using a combination of public APIs and web scraping techniques. The following sources were used:

1. **RemoteOK**: Data was collected using the RemoteOK public API, which provides structured JSON data for remote job postings, including job titles, companies, locations, descriptions, and salary information.
2. **Remotive**: Data was collected using the Remotive public API, which offers structured access to remote job postings across multiple categories, including software development and data-related roles.

Both sources were selected due to their accessibility, structured data format, and reproducibility.

3.2 Number of Data Samples

After combining data from all sources and removing duplicates, approximately 112 raw job postings were collected.

4 Data Cleaning, Analysis, and Visualization

4.1 Data Cleaning Process

The data cleaning process involved the following steps:

- Text normalization to remove extra whitespace, HTML tags, and inconsistent formatting from job titles, company names, locations, and descriptions.
- Duplicate job postings were removed using unique job URLs.
- Location strings were standardized and categorized into regional buckets such as West Coast, East Coast, Central, Remote or Unspecified, and Other.
- Job titles were mapped to broader role categories, including Data Scientist, Machine Learning Engineer, Data Analyst, Software Engineer, Data Engineer, and Other.
- Required skills were extracted from job titles and descriptions using keyword matching.
- Salary information was parsed and normalized to estimated yearly salary values in USD, where possible.

4.2 Data Analysis Process

The cleaned dataset was analyzed to uncover meaningful trends:

- Distribution of job roles to identify high-demand positions.
- Geographic analysis to determine regions with the highest concentration of job postings.
- Skill frequency analysis to identify the most in-demand technical skills.
- Salary analysis by role and region to examine compensation trends.

4.3 Data Visualization

Jupyter Notebooks were used to create visualizations, including bar charts for regional distribution, job role frequency, and skill demand. All visualizations are fully reproducible by running the provided Python scripts and notebooks.

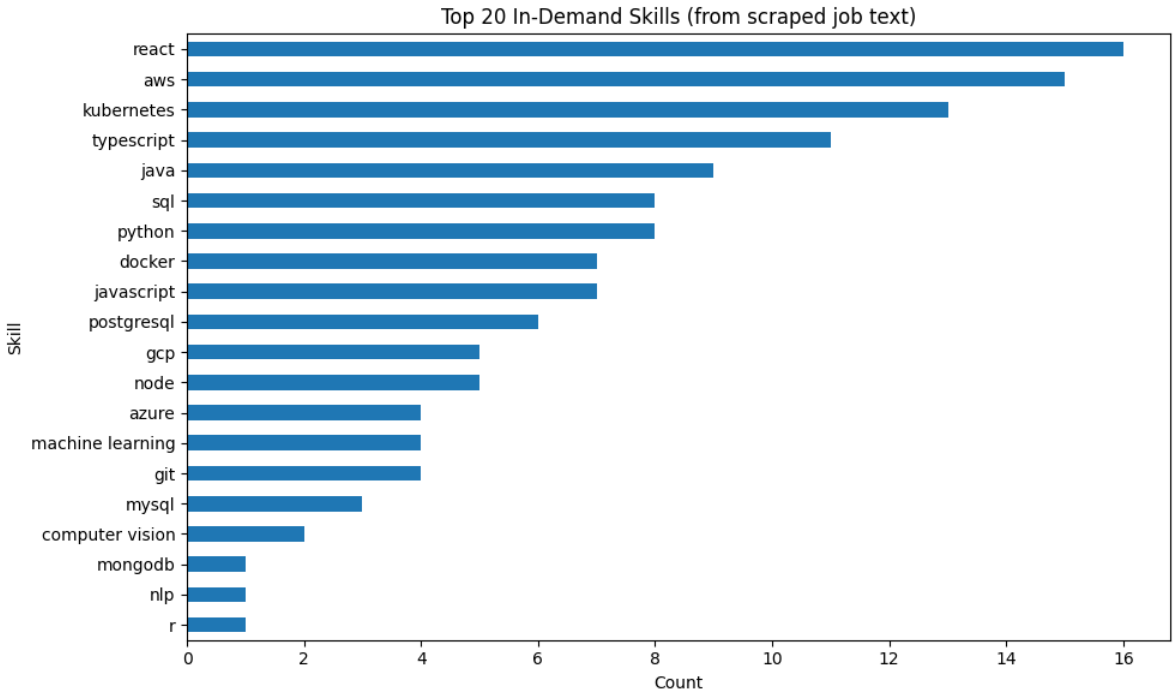

Figure 1: Top 20 in-demand skills extracted from job titles and descriptions.

4.4 Key Visualizations

4.5 Hypothesis and Conclusions

Initial Hypothesis: Remote and coastal regions have a higher concentration of data and software-related job opportunities, and technical skills such as Python, SQL, and cloud technologies are consistently in high demand.

Conclusions: The analysis supports the initial hypothesis. A significant portion of job postings were classified as remote or located in major coastal regions such as the West Coast and East Coast. Data science, machine learning, and software engineering roles dominated the dataset. Skills such as React, cloud platforms, and machine learning frameworks appeared most frequently across job postings. These results indicate sustained demand for data-driven and software-focused skill sets across diverse geographic regions.

5 Changes from Original Proposal

The original proposal planned to collect data from multiple job portals, including WeWorkRemotely and Himalayas. During implementation, these platforms employed client-side rendering and anti-scraping mechanisms that prevented the reliable extraction of data using standard HTTP requests. To address this challenge, the project pivoted to using RemoteOK and Remotive, both of which provide stable and reproducible APIs. This adjustment ensured data quality, reproducibility, and successful completion of the project objectives.

6 Future Work

Potential future extensions of this project include incorporating additional job platforms with official APIs, applying natural language processing techniques for more in-depth job description analysis, performing time series analysis to study hiring trends over time, and developing an interactive dashboard for real-time exploration of job market trends.

7 File Name, Format, and Page Limit

The final report is submitted as `final_report.pdf`. The document is provided in PDF format and adheres to the required page limit of two to five pages.