Job Market Analysis Report

For Project 1 Part A

Clinton K J[a]

[a]*Dept. of Information Science University of Arizona*

November, 2023

Abstract

This project report is dedicated to analyzing and visualizing the job market in the USA based on data sourced from over 4000 job postings on Glassdoor. The data collection method involved the use of both BeautifulSoup and Selenium in combination with Glassdoor API.

1. Introduction

The Job market analysis project consist of comprehensive analysis of open job position in USA. This report goes through the data analysis and visualization of 4000 samples. This analysis is specifically focused on the jobs related to machine learning and data science.

1.1. Problem Description

Analysis of job market data to explore the open positions related to the jobs in the field of data or machine learning in a particular region.

2. Process of Data Collection

This section will explain the methods used for the collection of Data

The two major data collection tools used are Selenium and Beautiful Soup. The job websites such as Glassdoor won't allow the collection of data directly from their website, so Glassdoor API was used in combination with Beautiful Soup to extract the data from website. But the challenge faced while collection is that the data related to skills are not collectable from the website. So to solve the problem Selenium has been used where every Job posting are clicked manually and made an algorithm for collecting the relevant skill from Qualifications in Job Posting.

The Algorithm for the collection of skills involves a text file that consists of all the relevant skills needed for particular jobs such as Python, Machine learning, SQL, etc and comparing the list of skills with the qualification list of job postings and adding the same in the collected data as an array.

3. Data wrangling

This section explains the data preprocessing and data cleaning for analysis purposes.

This analysis faced a major challenge while examining the collected data. After a thorough review, it became apparent that most of the job postings were duplicated within certain locations and moreover, data from one location was also posted under different other locations. The actual data collected for this report is over 20000 but after wrangling it became 4000 . Using the Pandas library drop_duplicates function all of the duplicates are deleted from data. Also, entries with missing values are deleted using dropna function from the Pandas library.

4. Market Data Visualization

The data visualization part consists of most of the possible visualization and analysis of Job Market Data. Given the constraints of space and the relevance of this particular project, it's crucial to highlight that the major part of visualization concentrates on the top 10 variables. These Includes titles, locations, skills, and other factors, providing a comprehensive overview of the Job Market

4.1. Titles of Postions

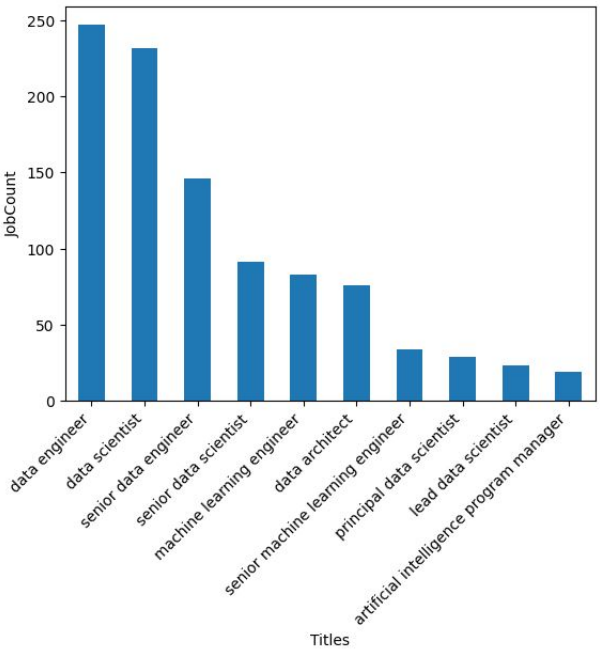
The fig. 1(a) shows the distribution of jobs with respect to Title.

The graph clearly shows that both Data Scientists and Data Engineer have the most number of positions available across the USA.

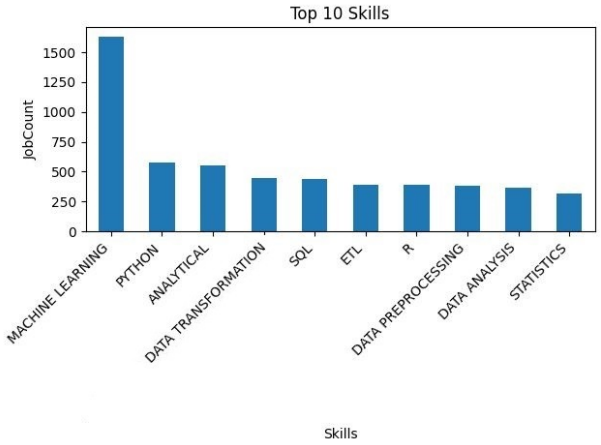
4.2. Required Skills

The fig. 1(b) shows the distribution of jobs with respect to Skills.

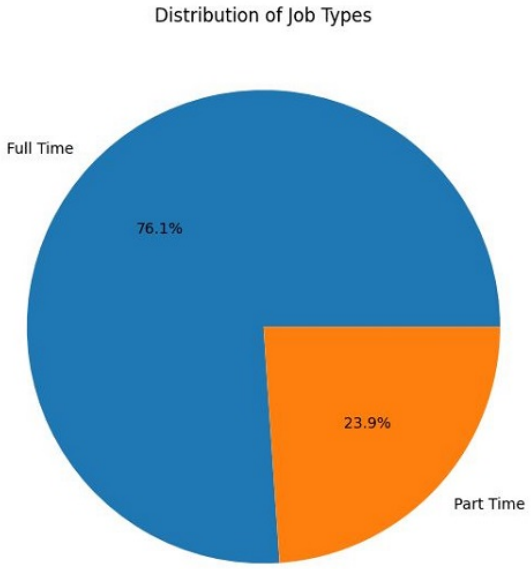
The bar graph illustrates that a significant portion of jobs in the field of data science require mandatory skills in machine learning.



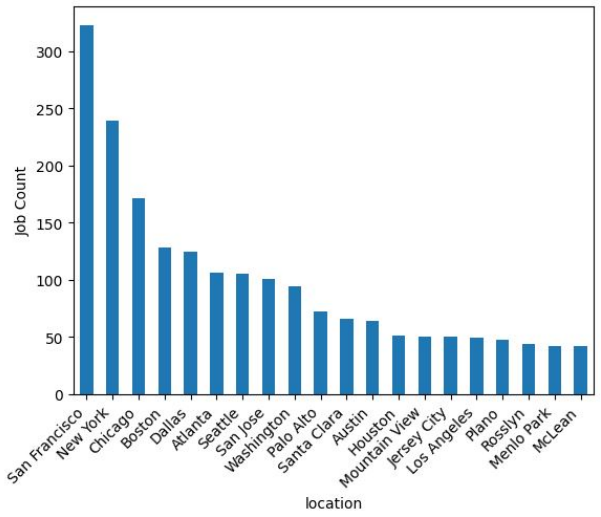
(a) Job Distribution by Title



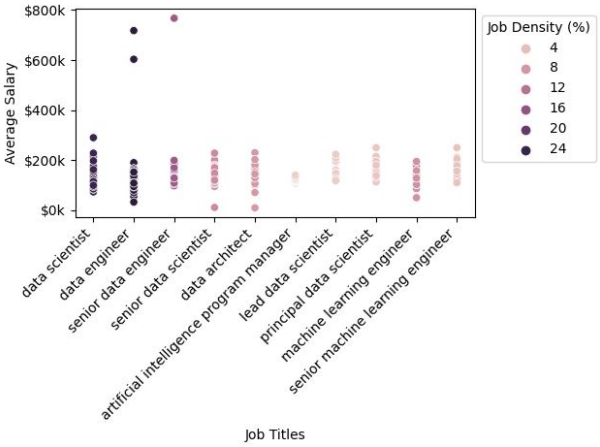
(b) Job Distribution by Skills



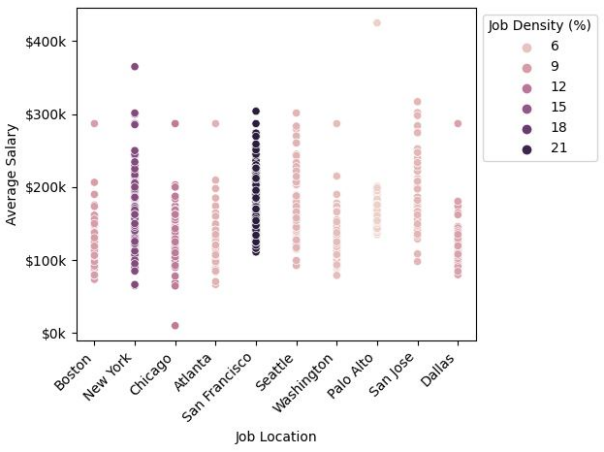
(c) Job Distribution by Job Type (By percentage)



(d) Job Distribution by Location



(e) Salary Distribution by Title and Job density (Darker the dot higher the Job density)



(f) Salary Distribution by Location and Job density (Darker the dot higher the Job density)

Figure 1: Data visualization of Job market Analysis

4.3. Type of Positions

The fig. 1(c) shows the distribution of jobs with respect to Job Type.

The Pie chart depicts the percentage of full-time jobs and part-time jobs in which most of the jobs in the collected data are full-time (76%) compared to part-time.

4.4. Locations where Jobs offered

The fig. 1(d) shows the distribution of jobs concerning Location.

The bar chart illustrates the top locations where data science-related jobs are available. Cities such as San Francisco, New York and Chicago stand out as prominent hubs, offering most of the jobs in this domain.

4.5. Multivariate graph with Job Titles, Salary and job density

The fig. 1(e) shows the Avg. Salary is based on their Job Titles and job density.

The multivariate scatter plot visually represents the average salary of different job titles, with colors indicating the density of jobs associated with each title in comparison to others. For instance, the darker color assigned to the Data Engineer position signifies a concentration of 24% of jobs, highlighting its prevalence compared to other roles. Conversely, lighter colors represent lower job concentrations, such as the Lead Data Scientist position, which comprises only 4% of the total jobs.This graph provides a comprehensive overview of job salaries for specific roles. Notably, it aligns with the "Job Distribution by Title" graph, highlighting a correlation between job count and title concerning job density.

4.6. Multivariate graph with Location, Salary and job density

The fig. 1(f) shows the Avg. Salary is based on Job location and job density.

The multivariate scatter plot provides a visual representation of the average salary across various job locations, using colors to signify the density of jobs for each location relative to others. For example, the darker color associated with San Francisco indicates a concentration of 21% of jobs, emphasizing its prevalence compared to other locations. On the contrary, lighter colors represent lower job concentrations, as seen with Palo Alto, which comprises only 6% of the total jobs. This graph correlates with the "Job Distribution by Location" graph, highlighting the relationship between job count and location in terms of job density.

5. Ideal Job

The Machine Learning Engineer role at Adobe in San Jose stands out as my ideal job, identified as both prevalent and intriguing in the gathered data. This position requires a skill set encompassing data analysis, problem-solving, machine learning, PyTorch, Python, and data transformation.

6. Conclusion

The Job Market Analysis project successfully implemented a data collection methodology utilizing a prominent job posting company. The gathered data underwent thorough preprocessing, followed by comprehensive analysis and visualization, resulting in a clear and informative representation of the entire dataset.

7. References

- Web Scraping with Python Beautiful Soup (Website Link)
- Multivariate Scatter plot Seaborn Documentation (Offical Documentation Link)
- Matplotlib : Plotting in Python (Website Link)
- Selenium Documentation (Documentation Link)
- Pandas Data Wrangling (Pdf Link)
- Glassdoor Api (Documentation Link)