Foundations of Information
Project Part-A

Job Market Analysis

By

Abhay Kumara Sri Krishna Nandiraju

1. Data Collection Method:

The data regarding job postings was collected from the famous job search website 'Indeed' using a technique called web scraping. Web scraping is a widely used technique for acquiring useful information from websites or the internet. In this project, information about data scientist roles in the USA was obtained from the 'Indeed' website.

Various Python libraries like pandas, beautiful-soup, matplotlib, and cloudscraper have been used in this project. Pandas was used to perform data cleaning and exploratory data analysis on the collected data. Matplotlib was used to visualize the cleaned data and obtain insights from it. Initially, the requests library was used to send a get request to the Indeed website but this was not successful because of the 'Cloudflare' protection it had. After researching, it was found that cloudscraper can be used to successfully send get requests and obtain information from the website. Cloudscraper is a Python module that can efficiently bypass Cloudflare's security. Hence, cloudscraper was used. Beautiful-Soup was used to parse the HTML content received from the get request and extract useful information from it.

Data is collected from the first 100 pages of the Indeed website with the keyword 'Data Scientist'. To collect the data the pagination in the URL(https://www.indeed.com/jobs?q=data+scientist&l=United+States&sort=date &filter=0&start={i*10}) is varied from 0 to 990. Here the variable 'i' in the URL is varied from 0 to 990 to scrape the job postings data. The title regarding a particular job was retrieved from the 'h2' HTML element with the class 'jobTitle'. The company name was collected from the 'span' HTML element's attribute 'data-testid' with the value 'company-name'. The location of the company was obtained from the 'div' HTML element's attribute 'data-testid' with the value 'text-location'. The details regarding the salary and type of position were collected from the 'div' HTML element with its attribute 'id' as 'salaryInfoAndJobType'. The job description details were collected from the 'div' HTML element's attribute 'id' with the value 'jobDescriptionText'.

Initially, the data collection process involved extracting information on job title, company, location, type of positions, and description. This information was stored in a pandas data frame. Later, this information was analyzed and salaries were

extracted and a new salary column was added to the data frame. This data frame had 1830 entries or jobs but had some rows with null values in the salary column. These rows with null values cannot be used for analysis as they result in errors. So, rows with null values were deleted from the data frame and it had 692 rows finally. In this data frame, the salaries were in the string data type but for analysis, they had to be in the float data type. Hence, salaries were converted into float data type values. City locations in the data frame were converted to state locations and were encoded using the two-letter abbreviations of the states for better analysis. Some of the entries in the location column have 'Remote' as they are remote jobs.

Secondly, job descriptions were processed and analyzed, and relevant skills regarding the job were made available as a Python list data structure. Then, a new column called 'Identified_Skills' is added to the data frame, and the relevant skills list is added to the respective row. Finally, the cleaned data frame has columns Title, Company, Location, Type of Positions, Job Description, Salary(in $), and Indentified_Skills. This data frame is stored as a CSV file.

```
[327…  final_job_list_df = pd.read_csv('final_job_list.csv')
       final_job_list_df
```

[327…

	Title	Company	Location	Type of Positions	Job Description	Salary	Identified_Skills
0	Staff Applied Scientist, Marketplace	ThredUp Inc.	CA	Full-time	\nAbout thredUP thredUP is transforming resale...	190000.0	['python', 'machine learning', 'aws']
1	Prompt Engineer for Generative AI (chatbot and...	Vicarious Talent Agency	WA	Full-time	We are Vicarious, a talent agency that represe...	70000.0	['python', 'machine learning', 'deep learning'...
2	Data Systems Analyst/Architect	General Dynamics Information Technology	Remote	Full-time	Clearance Level None Category Data Science Loc...	104000.0	['machine learning', 'aws', 'azure']
3	AIOps Principle Data Scientist	CVS Health	CT	Full-time	\nBring your heart to CVS Health. Every one of...	140000.0	['python', 'machine learning', 'aws', 'gcp']
4	Sr. Data Scientist	Altak Group	Remote	Full-time	Job Summary:\nWe are seeking an experienced Da...	136000.0	['python', 'machine learning', 'tableau', 'aws']
...	...	...	...	...	...	...	...
687	Senior Staff AI Data Engineer	Recruiting From Scratch	TX	Full-time	\n\n\nWho is \nRecruiting from Scratch \n: \n\...	160000.0	['python', 'pandas', 'machine learning', 'aws'...
688	Senior Staff AI Data Engineer	Recruiting From Scratch	CO	Full-time	\n\n\nWho is \nRecruiting from Scratch \n: \n\...	160000.0	['python', 'pandas', 'machine learning', 'aws'...
689	Senior Staff AI Data Engineer	Recruiting From Scratch	HI	Full-time	\n\n\nWho is \nRecruiting from Scratch \n: \n\...	160000.0	['python', 'pandas', 'machine learning', 'aws'...
690	Senior Staff AI Data Engineer	Recruiting From Scratch	CA	Full-time	\n\n\nWho is \nRecruiting from Scratch \n: \n\...	160000.0	['python', 'pandas', 'machine learning', 'aws'...
691	Senior Staff AI Data Engineer	Recruiting From Scratch	CA	Full-time	\n\n\nWho is \nRecruiting from Scratch \n: \n\...	160000.0	['python', 'pandas', 'machine learning', 'aws'...

692 rows × 7 columns

Fig-1: Image of the cleaned data used for analysis

2. Market Data Visualization:

a. Distribution of Job Titles:

The bar graph shown below depicts the distribution of the top twenty job titles arranged in descending order according to the frequency of occurrence of the respective job title. This bar graph provides details about the most common job positions available in the current market relating to the data science domain.

It is observed that there is a huge demand for the role of Senior Staff AI Data Engineer with more than 250 job vacancies followed by the Data Scientist role with around 50 job openings. The position of Senior Data Analyst stood in third place with approximately 15 job postings followed by the Machine Learning Engineer role with around 10 job openings. Jobs like Solutions Engineer, Lead Data Scientist, AI Engineer, and Sr. Data Scientist had an awfully small number of vacancies.

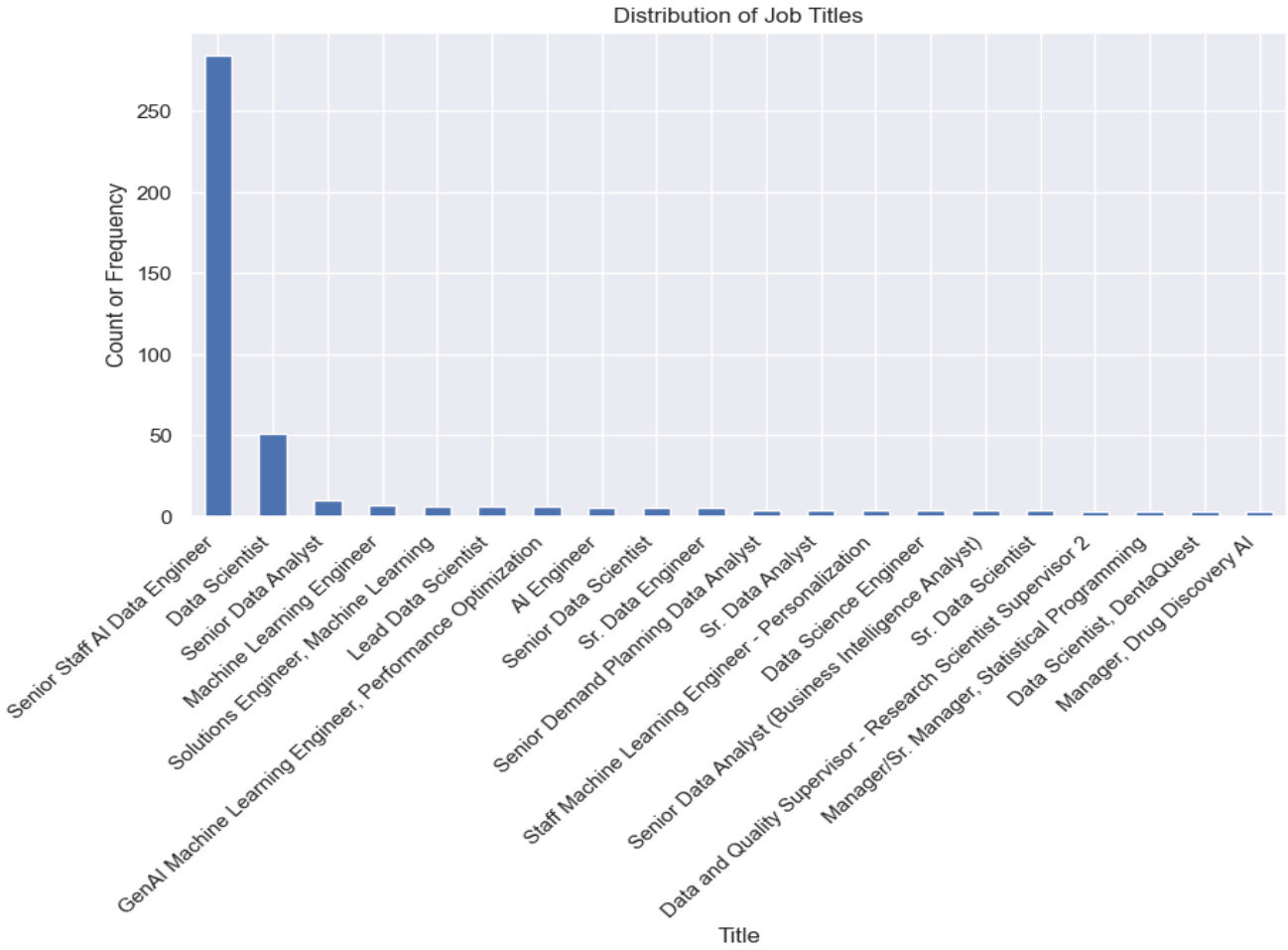

Fig-2: Distribution of Job Titles

b. Geographic Distribution of Jobs:

The below bar graph illustrates the geographic distribution of jobs arranged in descending order based on the number of jobs in a certain location. For clear visualization, the top twenty locations were picked out of the fifty available locations.

The city with the highest number of job openings in the data science domain is none other than California(CA). It has over 250 data science-related jobs and the new emerging trend called Remote jobs stands in second place with close to 50 jobs in the market. With little to no variation, New York(NY) comes next having around 50 job openings followed by Texas(TX) with over 45 job postings in the field of data science. States like Colorado(CO), Virginia(VA), and Maryland(MD) have more than 40 data science-related jobs followed by Ohio(OH), Arizona(AZ), and North Carolina(NC) with less than 25 jobs in each state.

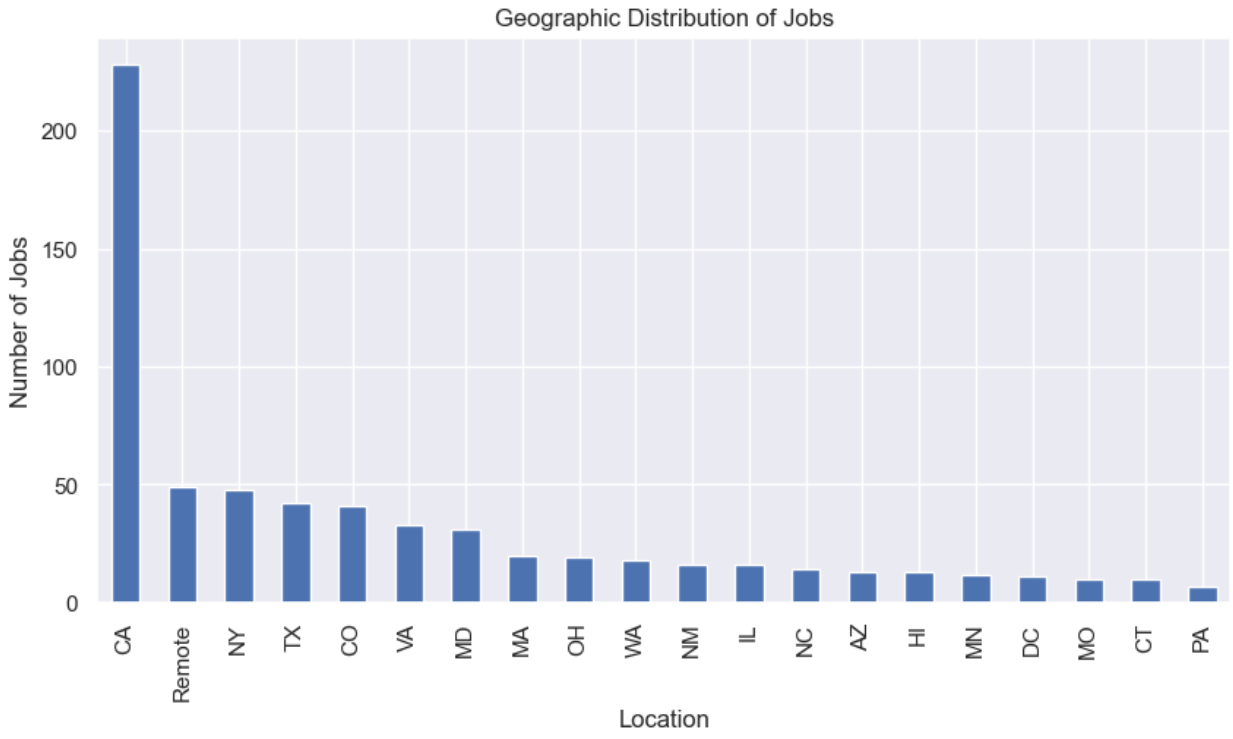

Fig-3: Geographic Distribution of Jobs

c. Pie Chart of Type of Positions:

The pie chart given below illustrates the percentage of job types like full-time and part-time. Surprisingly, all the jobs in the cleaned data were found to be Full-time job roles. In the raw data, there were three Part-time jobs and seven contract based jobs but these were filtered out as their salaries were not mentioned in the website. Hence, the final or the cleaned data has only full-time job roles.

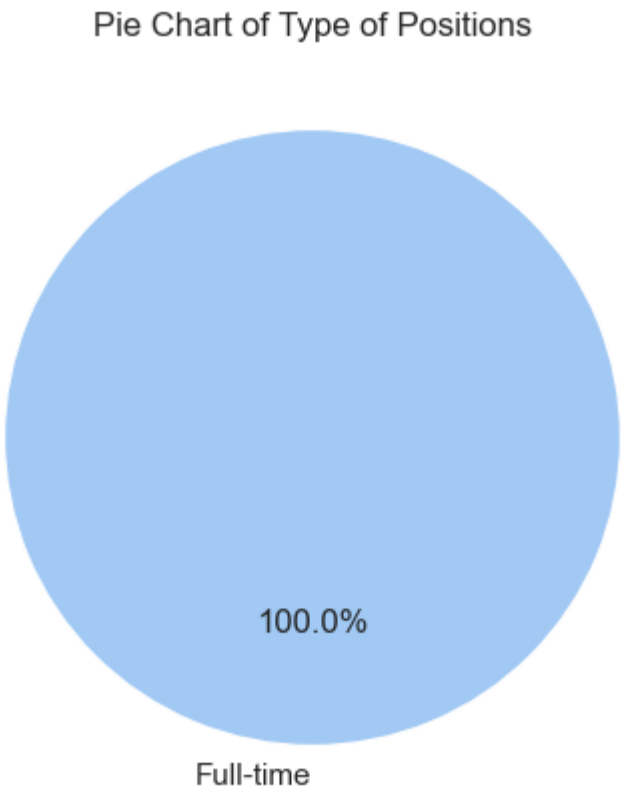

Fig-4: Pie Chart of Type of Positions

d. Location vs Salary:

The below given scatter plot displays various salaries at different locations in the US. Here each dot represents a job posting. From the scatter plot it can be seen that high salaries are paid in states like California, New York, Virginia, Maryland, Washington. Moreover, some of the remote jobs pay over $150,000 which is on par with the onsite full-time jobs.

Furthermore, a large number of high paying jobs are concentrated in heavily populated states like California, New York, Texas, Virginia. The jobs in California, Virginia, Maryland and New York are spread over a wide range of salaries. The salaries in California, New York and Virginia are spread over a range of $20,000 to $270,000 and the highest salary being $270,000 is paid in California. In contrast, the salaries in states like Indiana, Illinois and Ohio are concentrated in a small range of $50,000 to $100,000. There are only a handful of data science related jobs in remaining states like Arizona, Georgia, Arkansas and Oregon etc.,

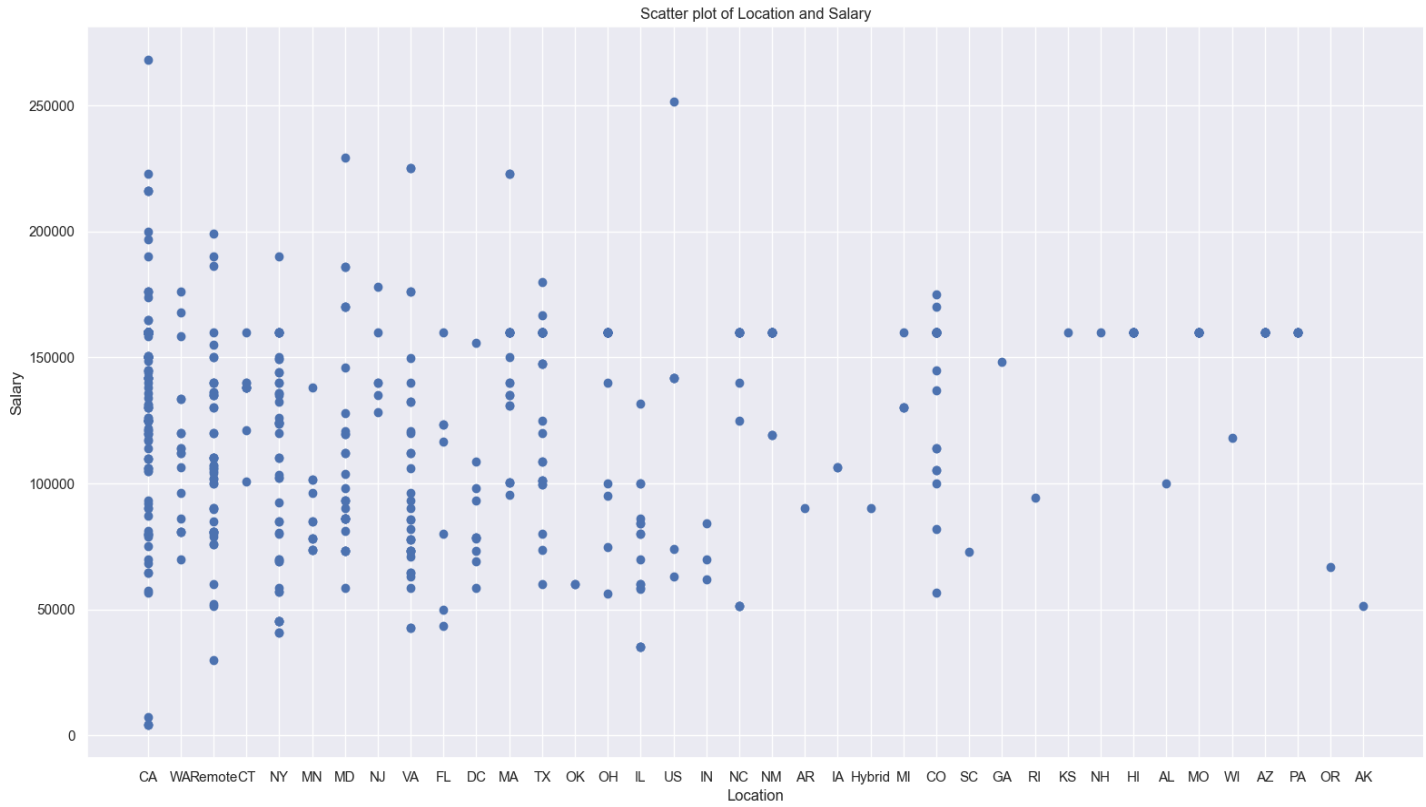

Fig-5:Scatter Plot of Location and Salary

e. Key Skills:

The bar graph given below illustrates the number of times a particular technology or skill was mentioned in the job postings. The key skills required for these positions arranged in descending order are Python, Machine Learning, SQL, AWS, Azure, Gcp, docker, ETL, Git, Kafka, Airflow, Nosql, Pandas, Pyspark, Tableau, Deep Learning, Tensorflow, PyTorch, Hadoop, Numpy, Apache Spark, MySQL, PostgreSQL, Spark ML.

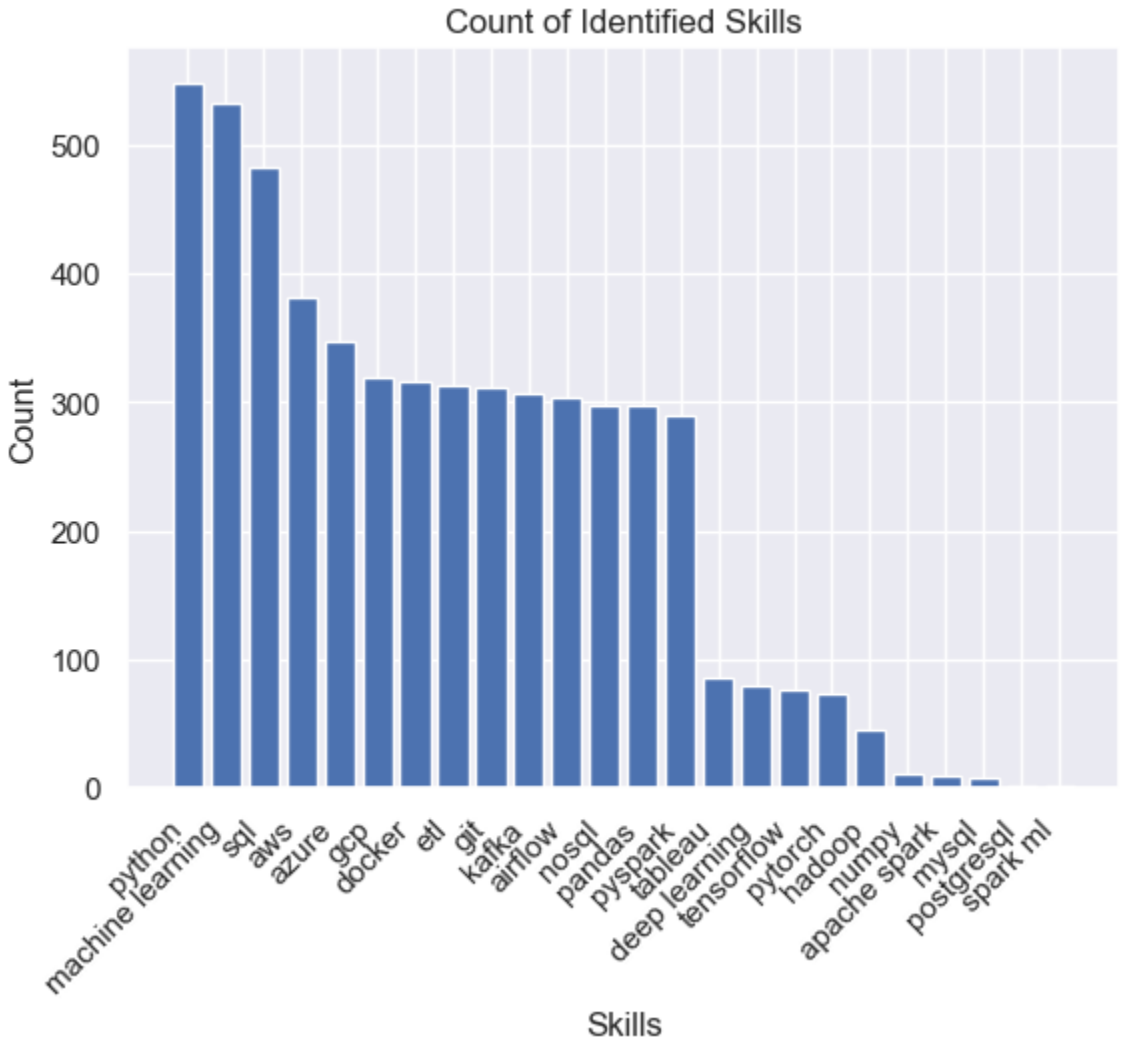

Fig-6: Count of Identified Skills

3. My Ideal Job:

My ideal job would be a Data Scientist as I have been working with technologies and tools like Python, Machine Learning, Statistics and Deep Learning from the past 14 months. The key skills required for this role are Python, Machine Learning, SQL, Statistics, Deep Learning, Pandas, familiarity with a cloud platform, Hadoop, Spark, Git, Data Visualization and Communication skills.