DATA SCIENCE SALARY ANALYSIS

Data Science Project

DECEMBER 10, 2021

UNIVERSITY OF KARACHI

Department of Computer Science (UBIT)

TABLE OF CONTENTS

DATA SCIENCE SALARY ANALYSIS

DATA SCIENCE PROJECT

BY ALI AHMED, NASRULLAH, MUNIR NIZAM AND BILAL SHEIKH

SUBMITTED TO DR. TAHSIN AHMED JILANI (COURSE SUPERVISOR)

INTRODUCTION

Our aim is to provide proper average salaries for a person who is going to work for a certain company, certain industry and certain location, certain job type and etc. to help him better negotiate his or her salary when he or she is recruited. And also we wanted to explore how much various companies in various locations are paying their employees for a certain job in data science related field. During this data exploratory analysis, we discovered many surprising results like what type of company is more likely to pay higher salary in a certain location which further mentioned in the report.

THE DATA

The data we are using for our model is from Glassdoor for data science related jobs in the USA. It was scrapped from Glassdoor using the keyword 'Data Science', and have scrapped a thousand jobs posting related to data science field/industry. The data includes the expected salary of the employee, the type of the job, the location of the job, the company type, the industry sector, the age of the company and the description of the job including what skills were repeated and emphasized in most of the jobs posting such as Python, Spark, R-studio, AWS and Excel.

DATA MINING

Mining the data required extracting numerical values from the text, separating quantitative values from the data. The qualitative data was categorized. Jobs were simplified such as a junior level job and a senior level job, and the jobs were further simplified to more categories. The categories include Data Engineer, Data Scientist, Machine Learning Engineer (MLE), Data Analyst, etc. Created columns for different job skills mentioned in the job description. Basically, we simplified the data so that it is usable for our model.

EXPLORATORY DATA ANALYSIS (EDA)

After the data cleaning was done, we started our exploration of the data. For exploring the data we used various Python libraries such as Matplot, Seaborn, Pandas, etc.

In our EDA, we found Machine Learning Engineers were the most paid of all, after the director, even the manager. Senior Data Scientist were also well payed off.

JOB SIMPLIFICATION

After the simplification of the jobs, we discovered from the jobs description that:

- 279 Data Scientists were required
- 119 Data Engineers were required
- 102 Data Analysts were needed
- 22 Machine Learning engineers were needed
- 22 of the were managerial posts
- 14 of them were asking for a director
- 184 postings could not be distinguished

SENIORITY AND THEIR SALARIES

Further in our data exploratory analysis, five hundred twenty postings of the jobs title did not have any mention of the seniority or vice-versa, two of them mentioned jobs at a junior level, and others remaining two hundred twenty were of senior level jobs. Average annual salaries are mentioned below in the table:

Jobs	Senior	Junior
Analyst	79K	57K
Data Engineer	125K	—
Data Scientist	140K	107K
MLE	142K	-
Director	169K	-
Manager	85K	-

RATINGS OF THE COMPANNIES

Most of the companies were between ratings 3 to rating 5.

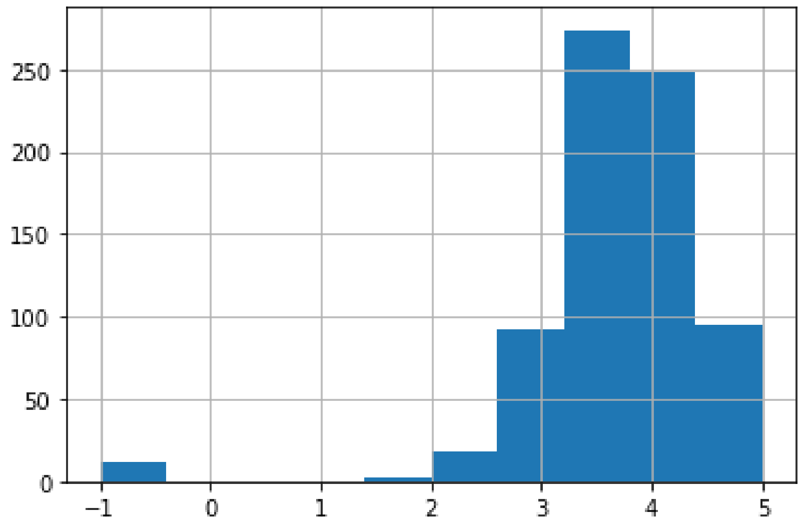

Figure 1

AVERAGE SALARIES

Looking up the average salaries, approximately hundred seventy-five jobs were paying annual 100K, and the distribution was between 50K to 200K annual salary. Surprisingly, Managers were being paid much less than a Data Engineers, Data Scientists and Machine Learning Engineer. And Machine Learning Engineers were the most paid of all of them.

Average Salaries, Data Science Field

Jobs	Average Salary
Data Analyst	65.85K
Data Engineer	105.40K
Data Scientist	117.56K
Director	168.60K
Manager	84.02K
MLE	126.43K
NA	84.85K

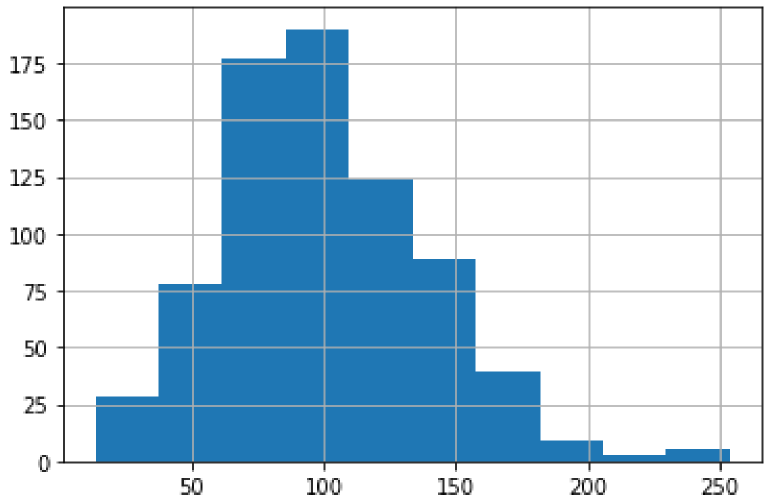

Figure 2

AGES OF THE COMPANY

Looking at distribution, majority of them belong to the bin of zero to twenty-five age category, nearly four hundred companies. Coming to the bin two, nearly hundred and fifty of them were of age between twenty five to fifty, further reading the distribution nearly 75 of them belonged to age bin of sixty to seventy, fifty of them were of age seventy to hundred and ten. And the some of them were of age hundred and thirty to hundred and sixty and some outliers belonged to the category of nearly two hundred and fifty years old.

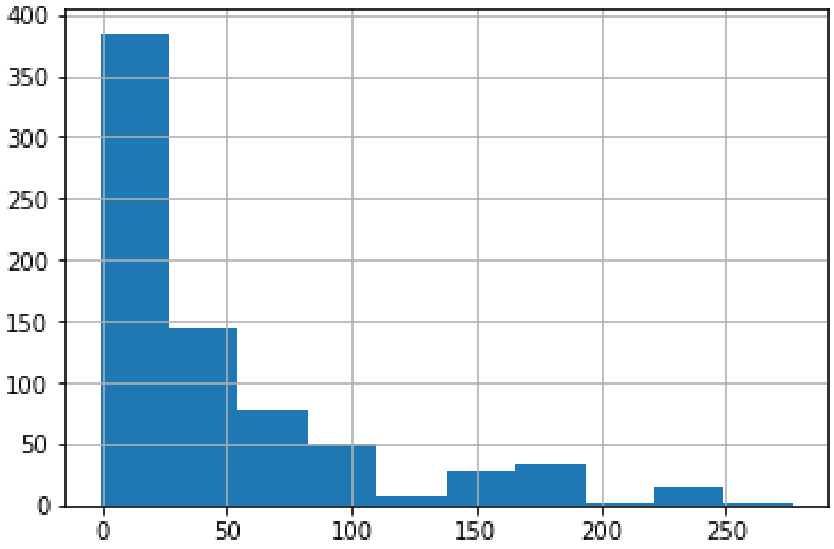

Figure 3

DESCRIPTION LENGTH OF THE JOB

Majority of the jobs (more than hundred and seventy-five) mentioned approximately more than four thousand characters in their job description, second majority of the jobs (a little lower than the hundred and seventy) had less than approximately three thousands characters, thirdly most of them had nearly more than five thousand characters in the description. And approximately sixty of the jobs were crossing six thousand characters in description length. Further, approximately twenty five jobs had more than seven thousand and eight thousand description length and less than twenty five jobs had less than seventeen characters in description length and a few touched ten thousand characters for their description length.

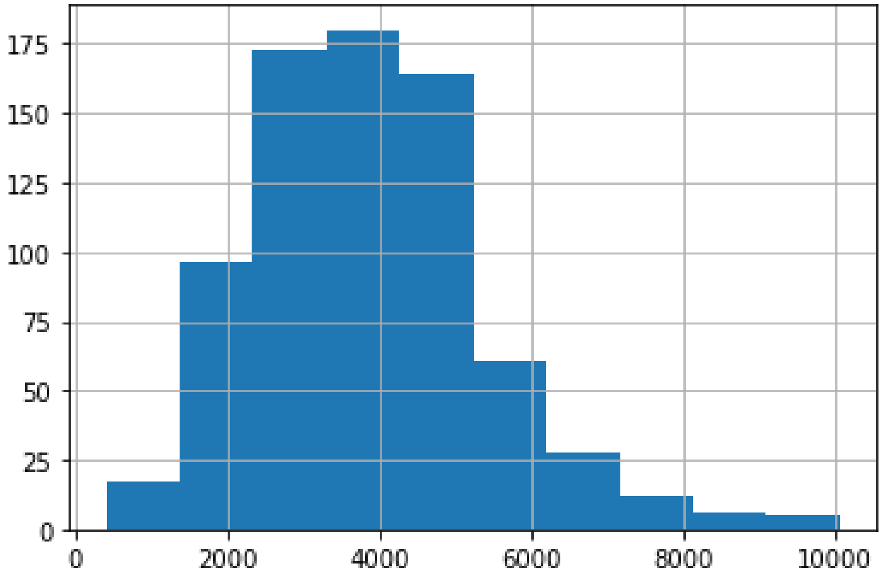

Figure 4

BOX PLOTS

Checking out the box plot readings for average salary, age of the company and ratings, we have the following distribution:

The box plot of the age reading shows median ages are below fifty. The box plot for the average salaries show median salary is a bit lower than hundred thousand. And the box plot reading for the rating is between four and three.

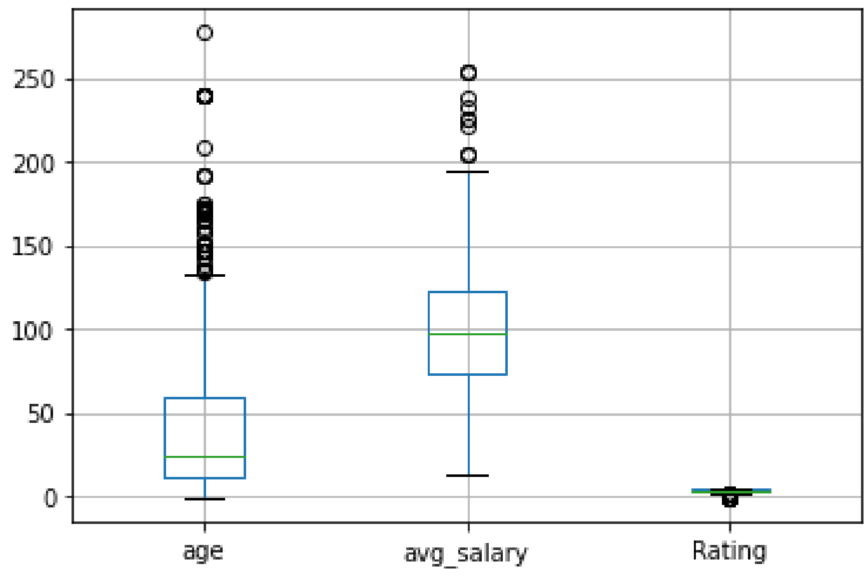

Figure 5

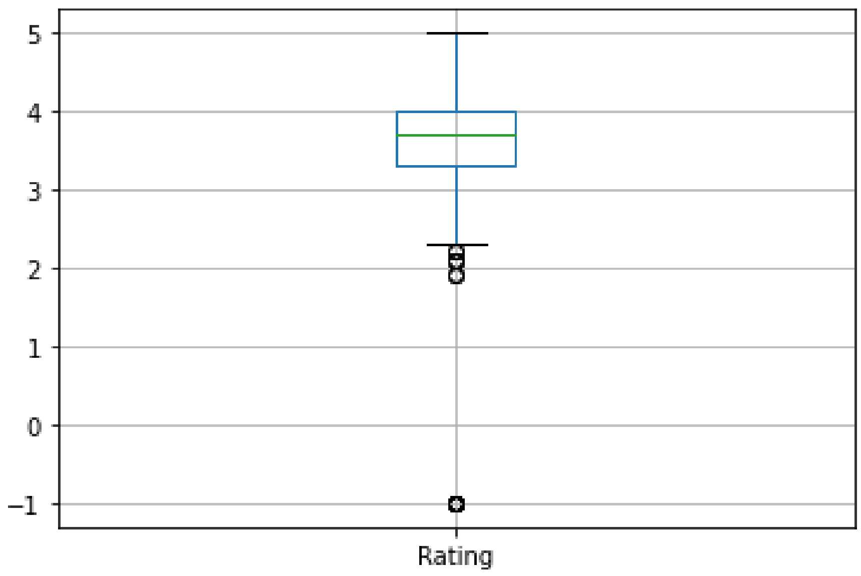

Figure 6

CORRELATION

Looking at the correlation between companies' age, average salary, Rating and the description length provided by a company in its job posting. Age has the highest correlation with description length.

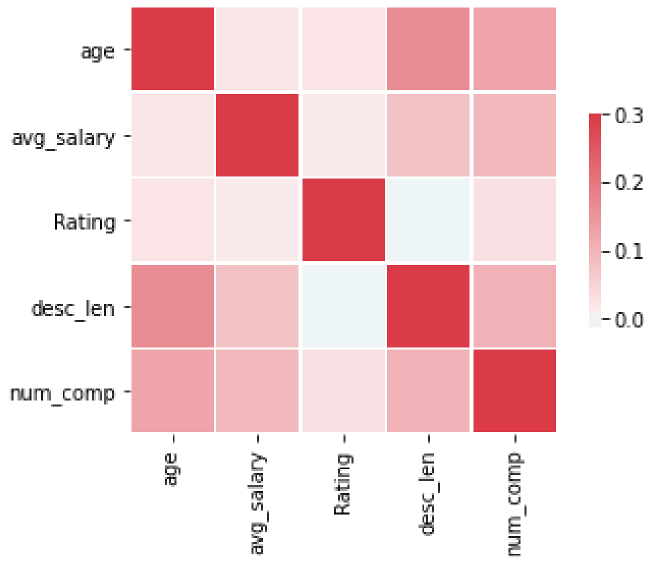

Figure 7

GRAPHS

SIZE OF THE COMPANY

From the graph, we can clearly see that majority of the companies are big companies which include number of employees from five hundred to ten thousand plus.

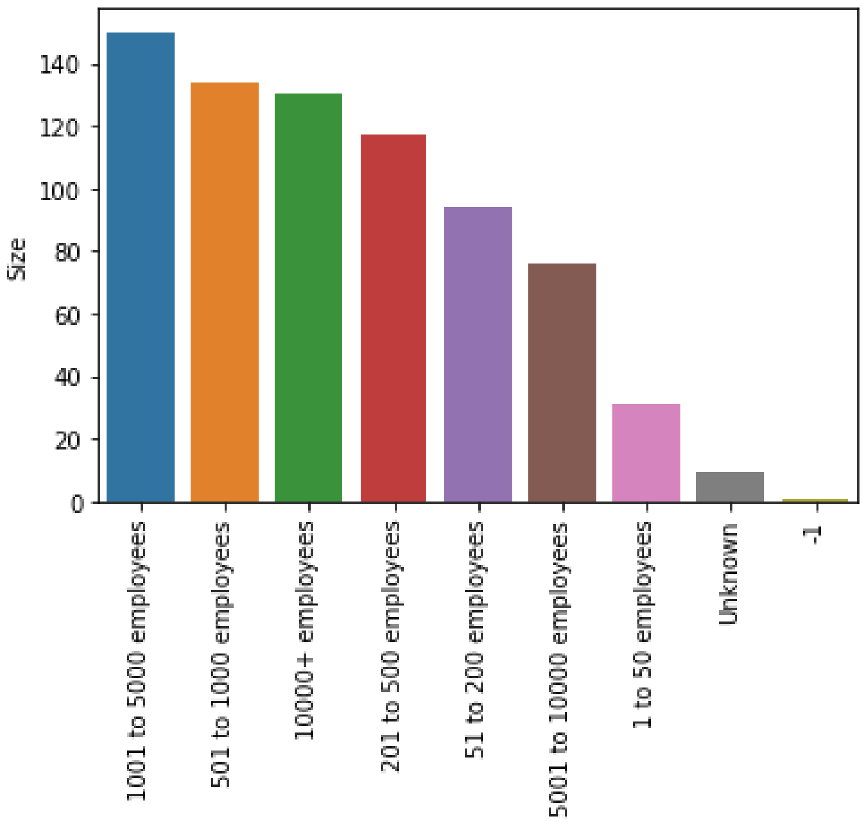

Figure 8

TYPE OF COMPANY

Our dataset includes total of ten types of companies

1. Private Company
   - More than four hundred companies
2. Public Company
   - Less than two hundred
3. Nonprofit Organization
   - A little less than fifty
4. Subsidiary or Business Segment
   - Less than fifty

Below categories fall below fifty in quantity of jobs

5. Hospital
6. Government
7. University or College

8. Other Organizations
9. School District
10. Unknown

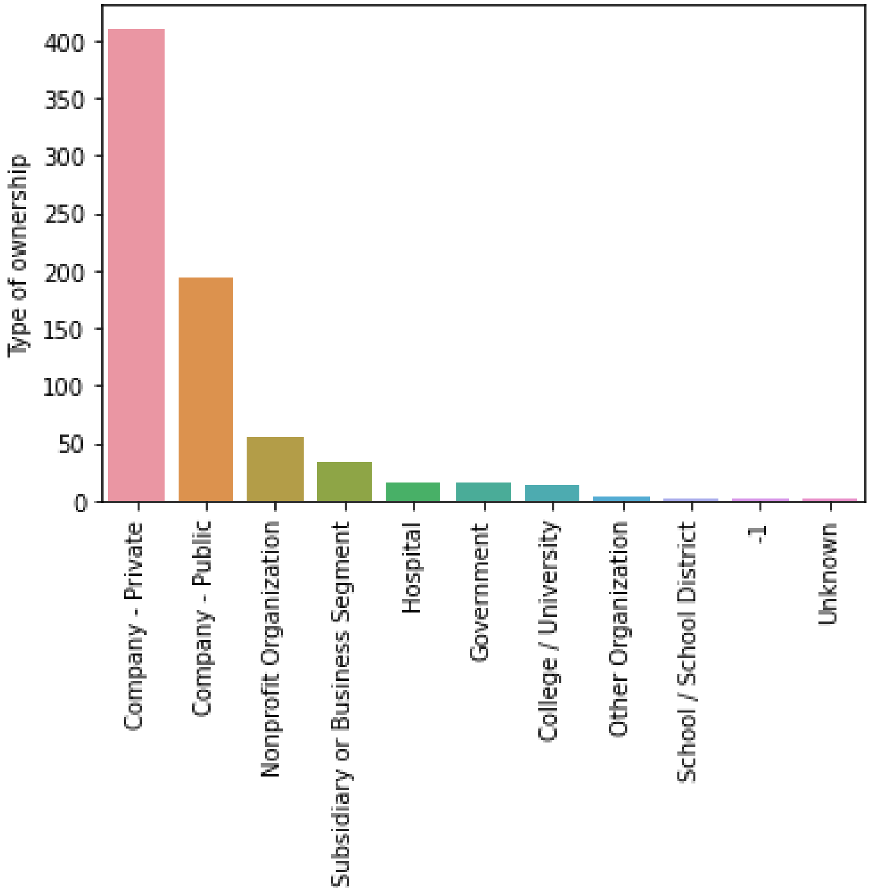

Figure 9

SECTOR

Analyzing further the twenty five sectors that we mined from the data, we discovered that nearly more than one hundred and seventy jobs were from information technology sector, surprisingly and maybe due to the current situation of the world Biotech and Pharmaceuticals stand second with more than a hundred jobs. Business Services sector comes third in this list with a less than a hundred jobs, followed by Insurance with less than seventy five jobs, then comes health care sector with a little less than fifty jobs. Surprisingly Finance sector is behind with less than fifty job requirements. Further down comes Manufacturing, Aerospace and Defense and Education with less than fifty jobs. And other sectors such as Retail, Oil Gas, Energy and Utilities, Government, Unknown, Non-profit, Travel and Tourism, Real Estate, Transportation and Logistics, Telecommunications, Media, Arts, Entertainment and Recreation, Consumer Services, Construction, Repair and Maintenance, Mining and Metals, agriculture and

Forestry and Accounting and Legal all of them fall below with less than twenty five job requirements.

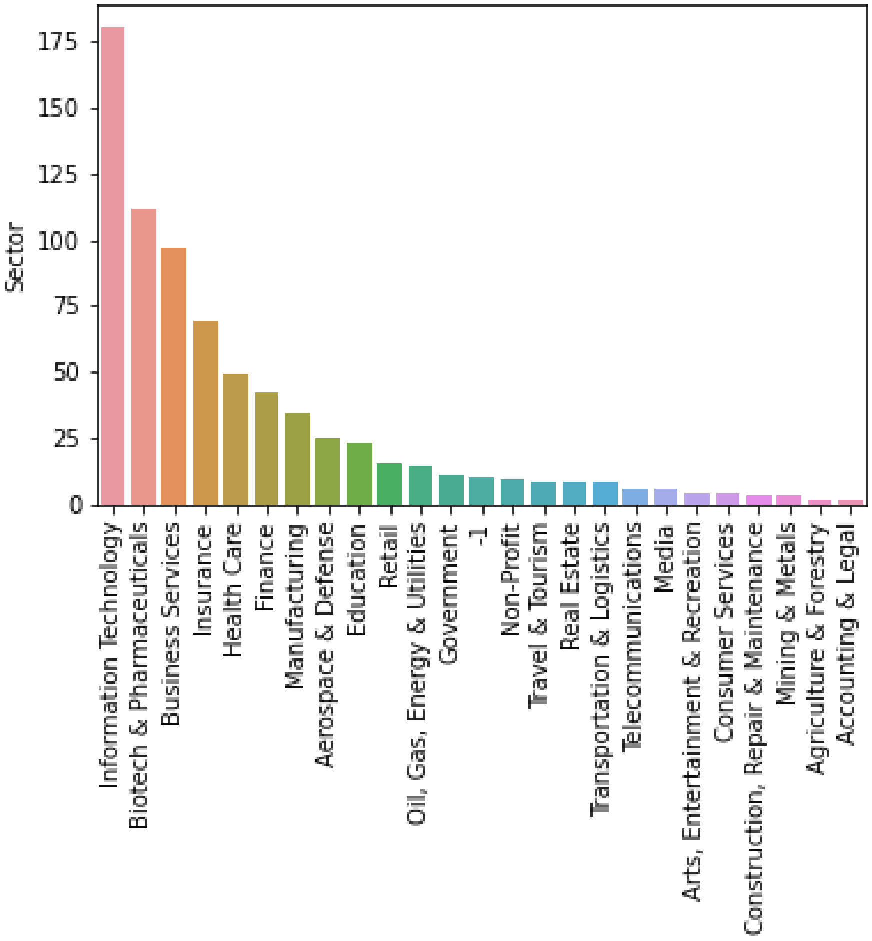

Figure 10

REVENUE

Looking at exploration of the revenue of the companies that require data science related jobs, we have nearly hundred and twenty five companies that have revenue of more than ten billion US dollars. Less than hundred and twenty five companies with a revenue of hundred to five hundred million US dollars are also requiring data science jobs which is interesting that smaller companies are looking more data science jobs too.

On third comes again big companies with one to two billion turnover looking for more than fifty data science jobs, followed by smaller companies with turnover of less than one billion and more than five hundred million that require more than fifty jobs in data science related jobs.

Smaller companies with fifty to hundred million turnover looking for approximately fifty data scientist jobs, which is surprising though. Companies with twenty five to fifty million turnover requiring less than 50 jobs and surprisingly, BIG companies with two billion to five billion are looking for much less jobs than other bigger and smaller companies, followed by the next category of smaller companies that are looking for less than fifty jobs. More surprisingly huge companies with turnover of five billion to ten billion are looking for much less jobs which is approximately less than twenty five jobs. Companies with five to ten million, one to five million turnover and less than 1 million turnover also required less than twenty five jobs.

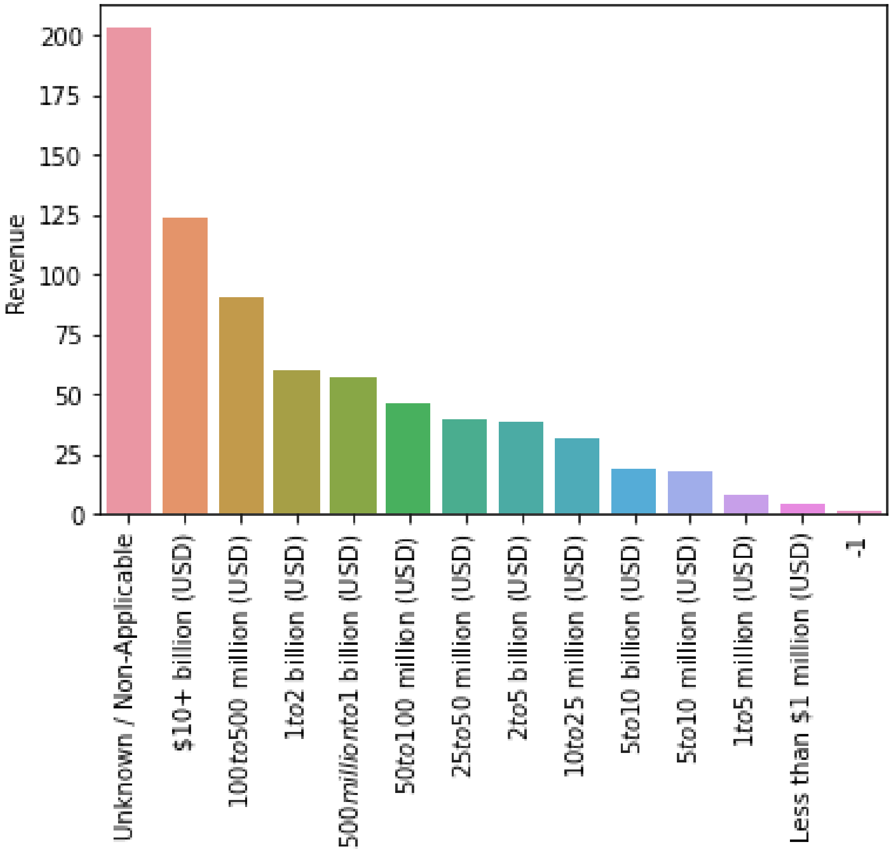

Figure 11

STATE

Further exploration was in the states that required jobs, California was on top with 152 jobs posting, second was Massachusetts with 103 jobs postings, New York was third with 72 jobs, Virginia was fourth with 41 jobs, Illinois had 41, Maryland had 35, Pennsylvania with 33, Texas had 28 jobs, Washington had 21 jobs, North Carolina had 21, New Jersey had 17, Florida had 16, Ohio had 14, and others state had significantly lower rates of jobs.

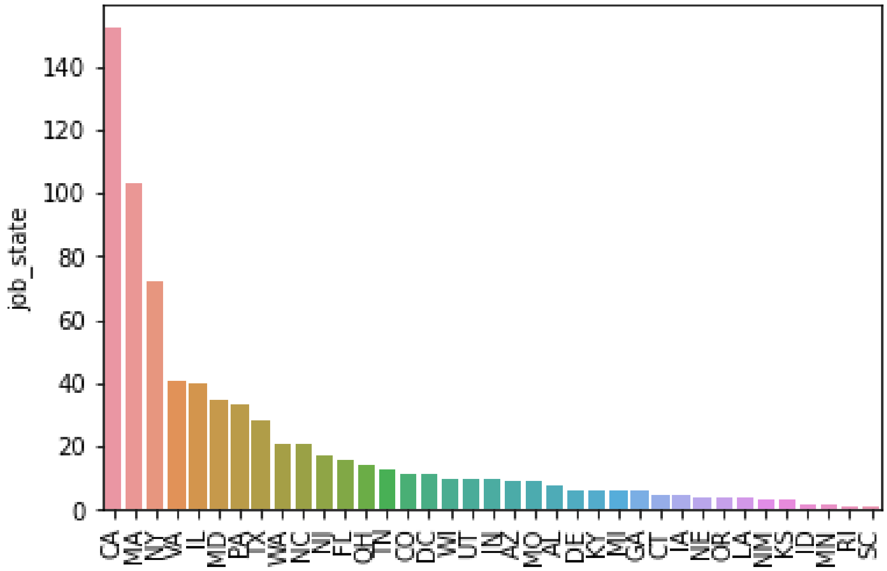

Figure 12

AVERAGE SALARIES BY STATES

According to our EDA, Washington DC tops the list with USD 149K per annum, followed by California with USD 142K per annum. More interestingly Utah is on the third with USD 140K per annum. And the lowest of all is Arizona with USD 69K per annum.

State	Avg Salary
DC	149.000000
CA	142.522059
UT	140.500000
MO	127.666667
IL	117.233333
NC	117.000000
NY	115.250000
MA	113.750000
WI	113.500000
PA	113.333333
MD	109.115385
CO	108.666667
VA	108.416667
NJ	106.875000
MI	106.625000

```
OH      105.285714
TX      100.730769
WA      99.764706
OR      98.500000
FL      97.357143
TN      96.000000
IN      84.500000
KY      84.000000
CT      84.000000
GA      81.333333
NM      74.333333
AZ      69.500000
```

NLTK (NATURAL LANGUAGE TOOLKIT)

Here we have created a word cloud using NLTK to see which words are most mentioned in the jobs description. And here is what we got. It looks like ‘Data’, ‘Team’, ‘Machine Learning’, ‘Data Scientist’, ‘Support’ and more words were the most stressed in a job description.

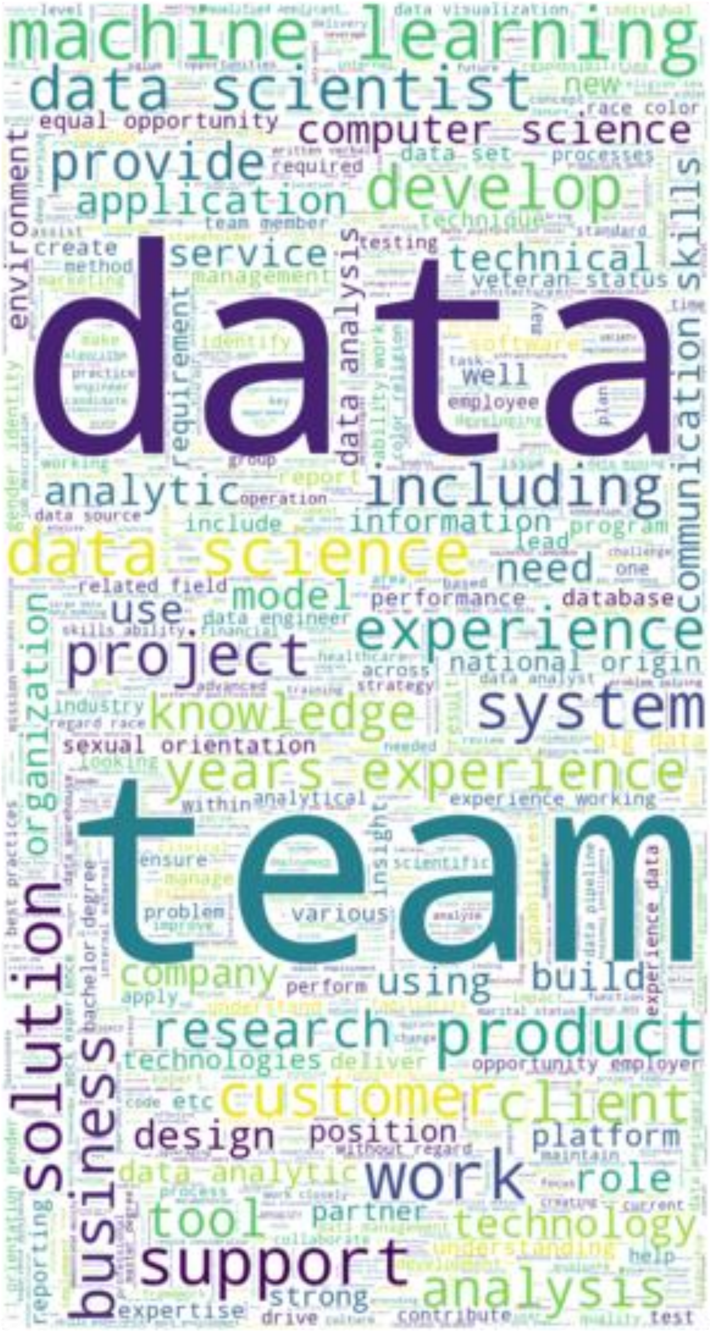

Figure 13

REFERENCES

Sakarya, Omar. 2019. "Selenium Tutorial: Scraping Glassdoor.com." *Towards Data Science.* 10 14. https://towardsdatascience.com/selenium-tutorial-scraping-glassdoor-com-in-10-minutes-3d0915c6d905.

(Sakarya 2019)