PROSPECT OF DATA RELATED JOBS IN UNITED STATES

Name:LakshmiManaswitha Chimakurthi
DA5020,Semester-1

With more and more companies using big data, the demand for data analytic specialists, sometimes called data scientists, data analysts, business analysts who know how to manage the tsunami of information, spot patterns within it and draw conclusions and insights—is nearing a frenzy. The main theme in this project is to analyze the salaries of the data related jobs across the United States. The job salaries depend on various factors such as technical skills, previous work experience, qualifying educational degree.

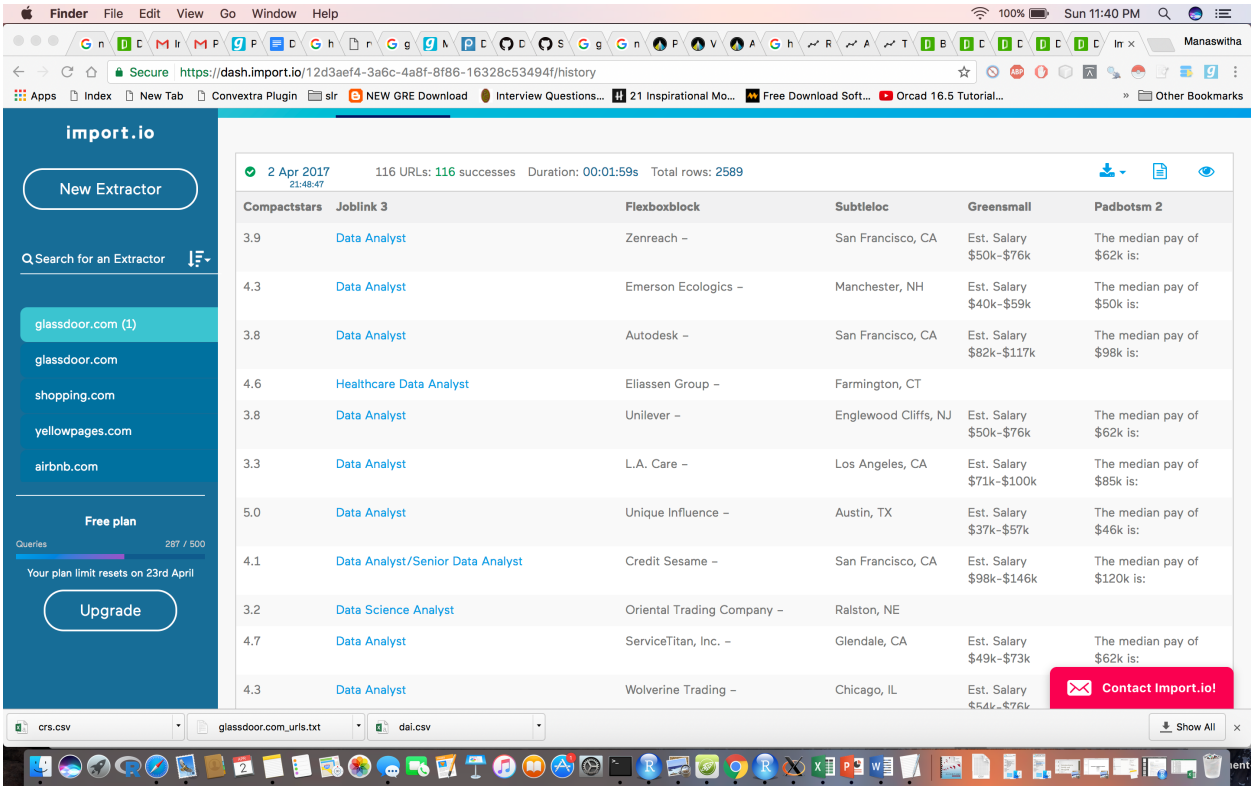
DATA COLLECTION

Data Source: www.glassdoor.com.

Glassdoor.com is popular among many data scientist as the go to site for finding employment and company scouting. Unsurprisingly, they are a popular site for data science practitioners to scrape. Prospective students and job searchers will be able to find insight on best paying career paths as well as desirable companies according to company reviews.

The data is scraped from Glassdoor.com using a scraping tool IMPORT.IO
which takes the url for the respective webpage to be scraped and makes a csv table format output of it.
After feeding the data url to import.io it scrapes data from the webpage. It also has an additional feature to eliminate the unnecessary columns. The data is now imported into R studio in the form of a csv format.

The main attributes of the data include
1.Company Rating
2.Job Title
3.Company Name
4.Company Location.
5.Estimated Salary.



Compactstars	Joblink 3	Flexboxblock	Subtleloc	Greensmall	Padbotsm 2
3.9	Data Analyst	Zenreach –	San Francisco, CA	Est. Salary $50k-$76k	The median pay of $62k is:
4.3	Data Analyst	Emerson Ecologics –	Manchester, NH	Est. Salary $40k-$59k	The median pay of $50k is:
3.8	Data Analyst	Autodesk –	San Francisco, CA	Est. Salary $82k-$117k	The median pay of $98k is:
4.6	Healthcare Data Analyst	Eliassen Group –	Farmington, CT		
3.8	Data Analyst	Unilever –	Englewood Cliffs, NJ	Est. Salary $50k-$76k	The median pay of $62k is:
3.3	Data Analyst	L.A. Care –	Los Angeles, CA	Est. Salary $71k-$100k	The median pay of $85k is:
5.0	Data Analyst	Unique Influence –	Austin, TX	Est. Salary $37k-$57k	The median pay of $46k is:
4.1	Data Analyst/Senior Data Analyst	Credit Sesame –	San Francisco, CA	Est. Salary $98k-$146k	The median pay of $120k is:
3.2	Data Science Analyst	Oriental Trading Company –	Ralston, NE		
4.7	Data Analyst	ServiceTitan, Inc. –	Glendale, CA	Est. Salary $49k-$73k	The median pay of $62k is:
4.3	Data Analyst	Wolverine Trading –	Chicago, IL	Est. Salary $54k-$76k	

A snapshot of scraping data using import.io

DATA CLEANING

The data cleaning involves the separation of company city and company state.
Conversion of Estimated Salary range into Median Salary
Removing the rows which do not contain Salary
Removal of unneeded information.

CLEANED DATA SAMPLE

	JobTitle	companyname	state	Rating	parse
1	Data Analyst	Zenreach –	CA	3.9	62
2	Data Analyst	Emerson Ecologics –	NH	4.3	50
3	Data Analyst	Autodesk –	CA	3.8	98
4	Data Analyst	Unilever –	NJ	3.8	62
5	Data Analyst	L.A. Care –	CA	3.3	85
6	Data Analyst	Unique Influence –	TX	5.0	46
7	Data Analyst/Senior Data Analyst	Credit Sesame –	CA	4.1	120
8	Data Analyst	ServiceTitan, Inc. –	CA	4.7	62
9	Data Analyst	Wolverine Trading –	IL	4.3	64
10	Data Analyst	Bankrate.com –	FL	3.3	51
11	Data Analyst	Iron Mountain –	MA	3.1	58
12	Data Analyst	Grand Circle –	MA	2.0	63
13	Data Analyst	The Oakleaf Group, LLC –	MD	3.4	68
14	Data Analyst	Grubhub –	IL	3.3	77
15	Data Analytics Analyst	Tradeweb –	NY	3.8	90
16	Data Analyst Transplant Administration	Houston Methodist –	TX	3.9	50
17	Data Analyst	Shutterstock –	NY	3.4	69
18	Data Analyst	Nytec –	CA	4.2	74

DATA STORAGE

The data collected in the form of csv from import.io undergoes a series of cleaning steps and is then stored in a nonrelational data bae namely **MONGODB.**MongoDB is a NoSQL document oriented database. It can be used to store large semi-structured data as an object. A database consists of 0 to many collections. A collection represents a collection of similar real world projects. An object or a document consists of named fields and values. The values of the fields may in turn consist of objects or a collection of objects.

The storage process is done in R using mongolite package.

The major functions used here are insert, export, find, aggregate, count.

```
28
29
30 mongo_data<-mongo("cbn")
31 #inserting data into mongo db
32 mongo_data$insert(jobdata)
33 #count the results
34 mongo_data$count
35 mongo_data$export(file("cbn.txt"))
36 avgsal<-mongo_data$aggregate('[{"$group":{"_id":"$state", "averagepay": {"$avg":"$parse"}}}]')
37 mostjobs<-mongo_data$aggregate('[{"$group":{"_id":"$state", "numberofjobs": {"$sum":1}}}]')
38 mostcompany<-mongo_data$aggregate('[{"$group":{"_id":"$companyname", "pay": {"$max":"$parse"}}}]')
39
40
41
```

A snap shot of mongodb queries in R

The major technologies used here are

1.IMPORT.IO for data scraping from GLASSDOOR

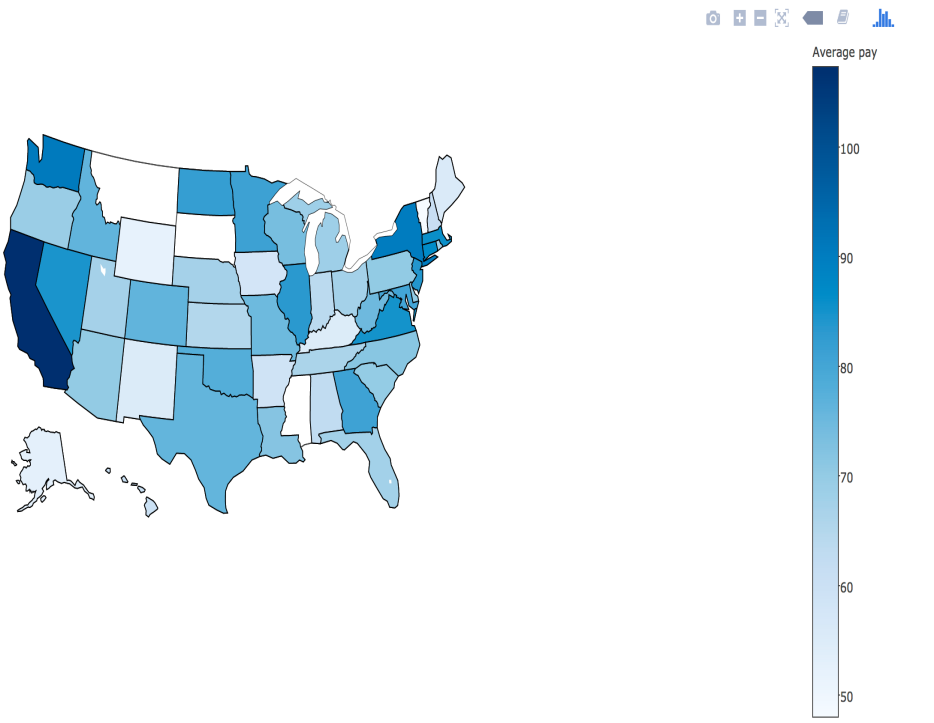
2.MONGODB for Data Storage

3.R ggplot2, plotly for Data Visualization.

DATA VISUALIZATION

The below US map shows the average pay for data related jobs for each state in US. The darker the colour the highest salary the state pays for the data related jobs. The below map is plotted using a package called plotly.

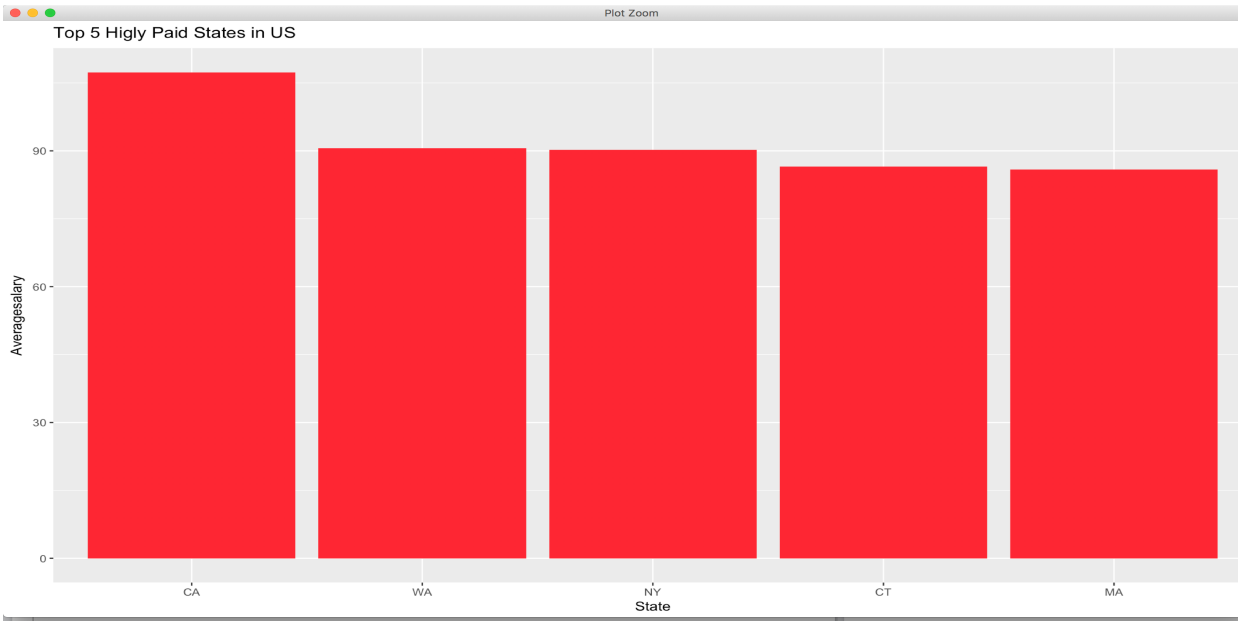
AVERAGE SALARY FOR DATA RELATED JOBS IN US



Digging on deeper the top 5 most paid states for data related jobs are shown in the below graph.

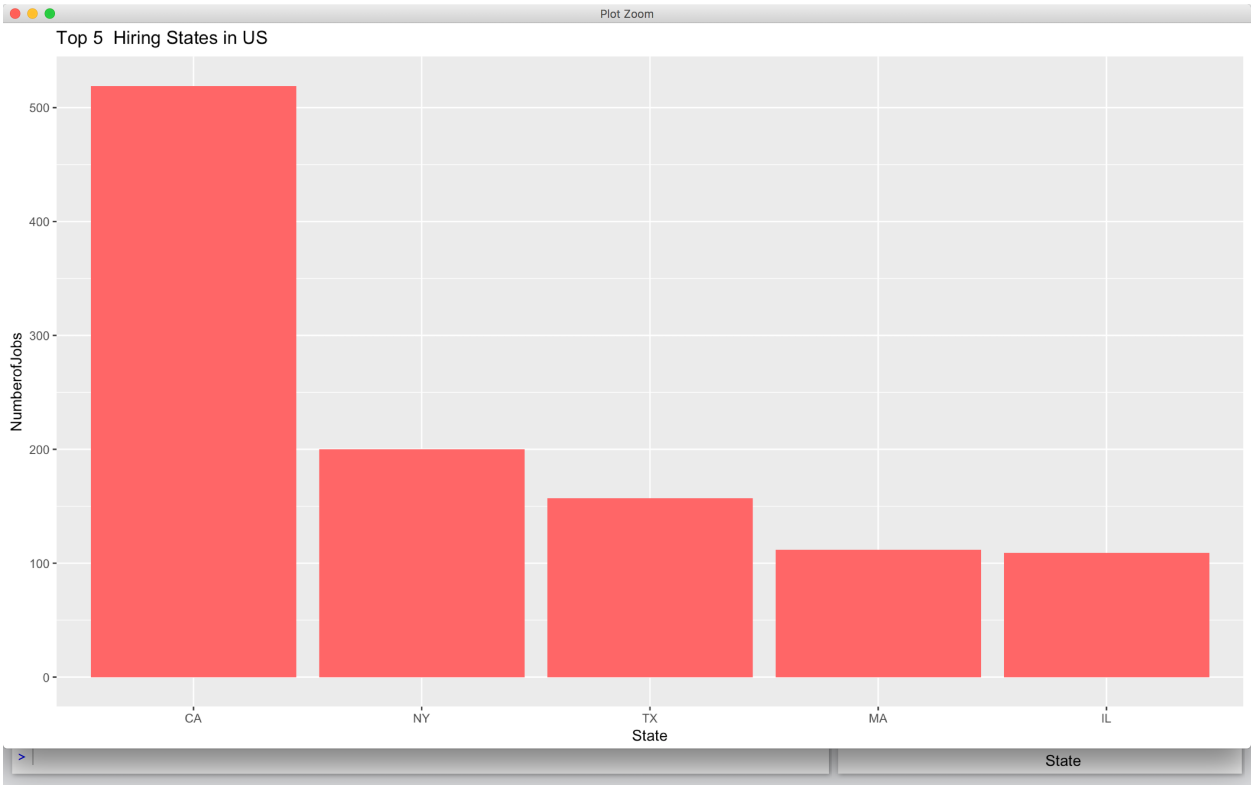
CALIFORNIA, WASHINGTON, NEWYORK, CONNECTICUT, MASSACHUSETS

Offers higher salaries compared to other states.

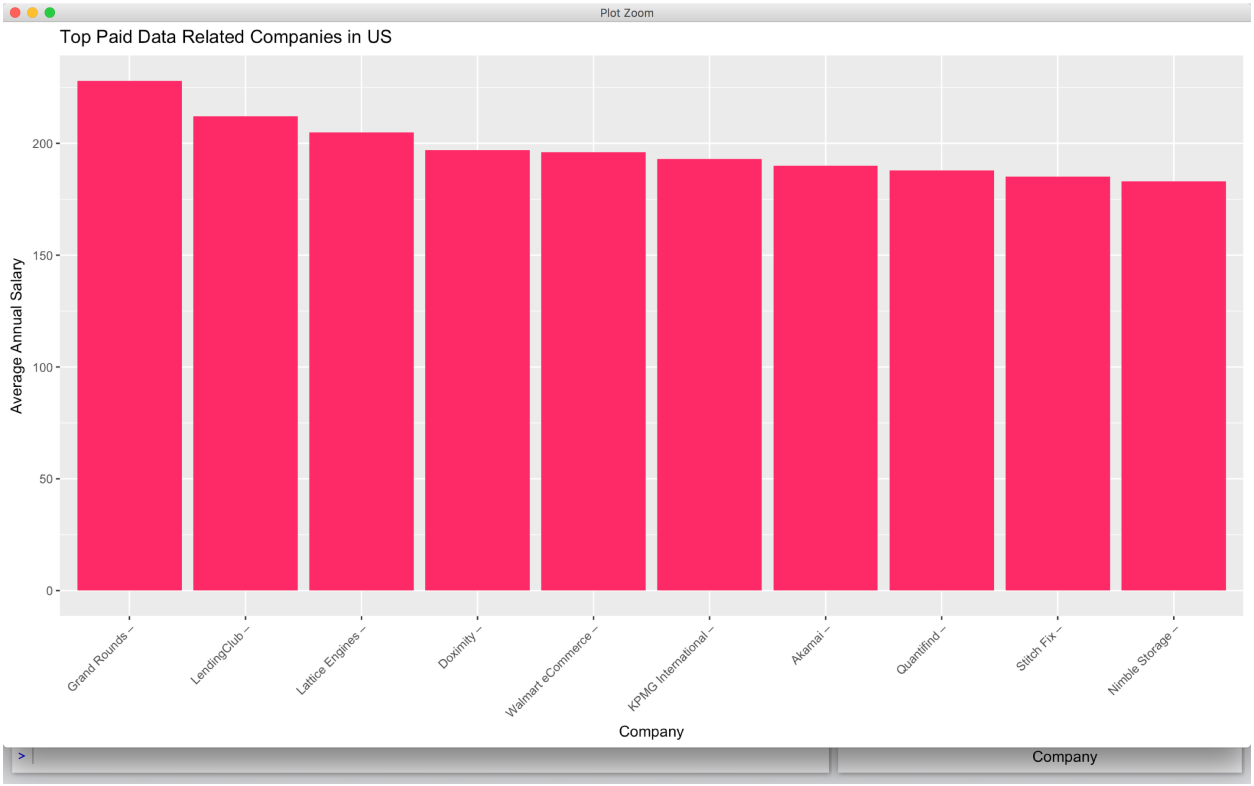


The below graph tells about the top5 US states which hires most data related jobs.
CALIFORNIA, NEWYORK, TEXAS, MASSACHUSETS, ILLINOIS
are the top 5 hiring States in US.

TOP 5 HIRING STATES IN UNITED STATES



TOP PAID COMPANIES IN UNITED STATES FOR DATA RELATED JOBS.



The above graph tells about the most highly paid companies in US for the data related jobs.