



The University of Texas at San Antonio™

College for Business
DA_6223_Spring2021

Best location to find job for data scientists

Web scrapping by python

By: Narges Shahmohammadi Heydari

introduction

Main idea :

- Which locations are the best to find a job for data scientists

Data for search:

- Companies
 - their salary and rating
- The states and cities which have the most offering jobs
- Offering remote jobs



<https://www.indeed.com/>

Web Scrapping

Some Details of scrapping

- For the best result, I Chose 100 pages of indeed.com and there are 10/15 block in each page.
- So I used **Pagination** to have a complete search. But I figure out that not all of blocks have the rating or salary.



Pacific Northwest National Laboratory 4.1 ★
Seattle, WA +1 location

- This position is for a Scientist in Data Science and Applied or Mathematical Statistics who will provide scientific and technical research within the National...

4 days ago · Save job · More...

REMOTE Safety Data Scientist I

Raland Compliance Partners

Remote

\$50 - \$95 an hour

➤ Easily apply ⚡ Responsive employer 🕒 Urgently hiring

- Analyze and interpret data from multiple sources (clinical trials, safety databases, literature, etc).
- Ability to interpret, summarize and present clinical data...

9 days ago · Save job



Data Scientist

VertMarkets 3.7 ★

Cranberry Township, PA 16066

➤ Easily apply

- The Data Scientist will also be responsible for applying these insights, as well as market data they have discovered, to make recommendations on ways to...

17 days ago · Save job

Be the first to see new data scientist jobs in usa

Email address

Activate

By creating a job alert, you agree to our [Terms](#). You can change your consent settings at any time by unsubscribing or as detailed in our terms.

I constructed the list of page links ,
iterate and scrapping of each
blocks

```
In [2]: # creating the list of urls
path_links = ['https://www.indeed.com/q-data-scientist-l-usa-jobs.html']
stirng= 'https://www.indeed.com/jobs?q=data+scientist&l=usa&start='
for i in range(1,100):
    path_links.append(stirng+str(i*10))
```

```
In [ ]: city_col = []
rating_col = []
salary_col = []
title_col = []
company_col = []
n=0
for url in path_links:

    # url = "https://www.indeed.com/jobs?q=data+scientist&l=usa"
    response = requests.get(url)
    page_content = BeautifulSoup(response.content)
    job_div = page_content.find_all('div', attrs ={'class': 'jobsearch-SerpJobCard'})
```

Data Preprocessing

Small part of my data frame

It needed lots of cleaning...

```
In [3]: df['CITY'] = df['city'].str.split(', ').str.get(0).str.title()# seperating city and state, create new
df.head(10)
```

Out[3]:

	title	company	city	rating	salary	CITY
0	Data Scientist	United Medical Credit	Remote	NaN	105,000–125,000 a year	Remote
1	Data Scientist Apprentice	IBM	United States	3.9	NaN	United States
2	Data Scientist - Entry Level	Numerdox	Sacramento, CA	NaN	NaN	Sacramento
3	Data Classification Specialist - Contract	Idiomatic	Remote	NaN	\$20 an hour	Remote
4	Data Scientist, Analytics, Intern	Facebook	Remote	4.2	NaN	Remote
5	Data Scientist	ForMotiv	Remote	NaN	75,000–120,000 a year	Remote
6	REMOTE Data Scientist	Insight Global	Remote	4.0	60–72 an hour	Remote

Some details of cleaning data frame

- Stripping non-digit characters
- Split the min max salaries
- Change hourly to annually
- Average of yearly salaries
- Rating: Change to float
- Sperate city and states from location

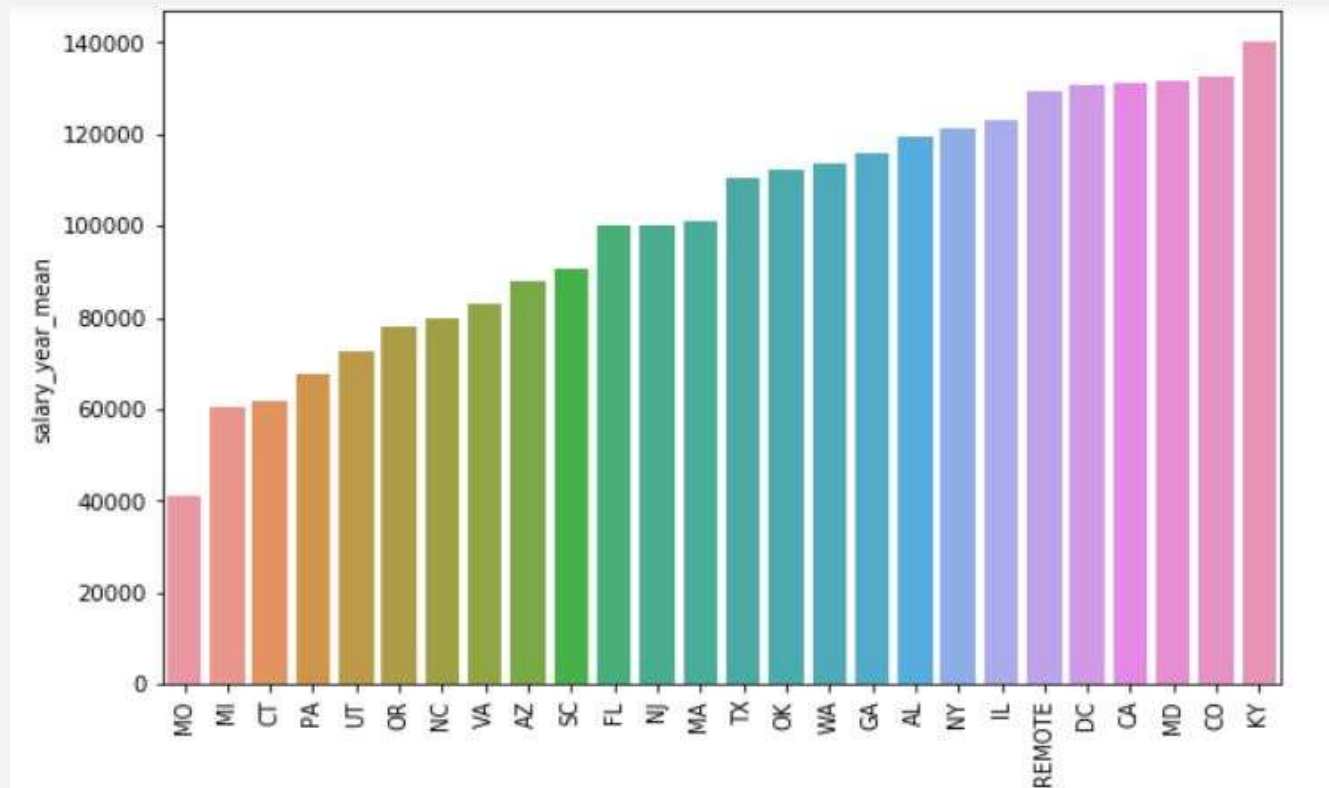
Exploratory Data Analysis

States and highest salaries

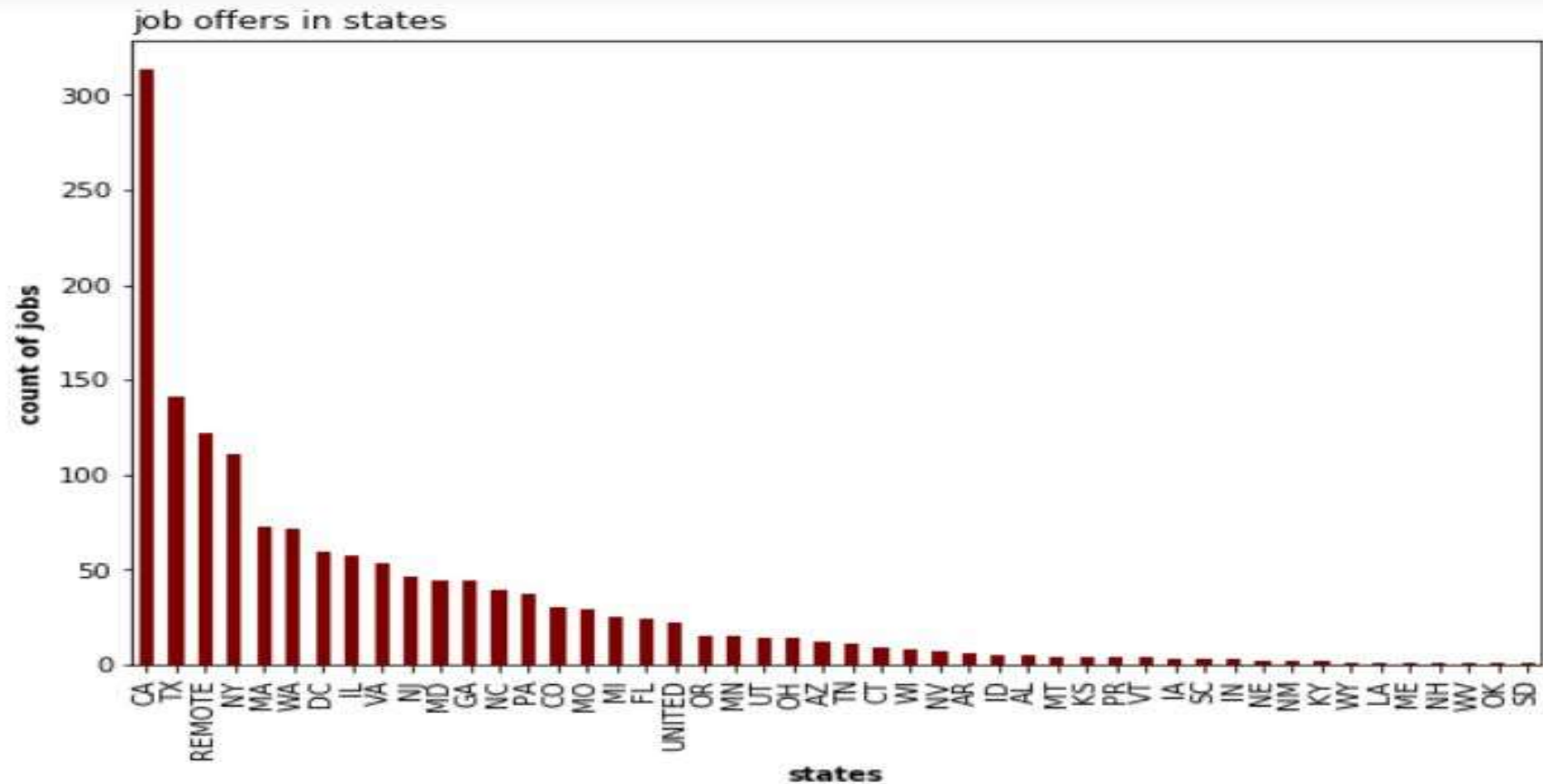
- I got average for salary

Kentucky has the highest and then Colorado, Maryland, California and Washington, D.C are the best in this part.

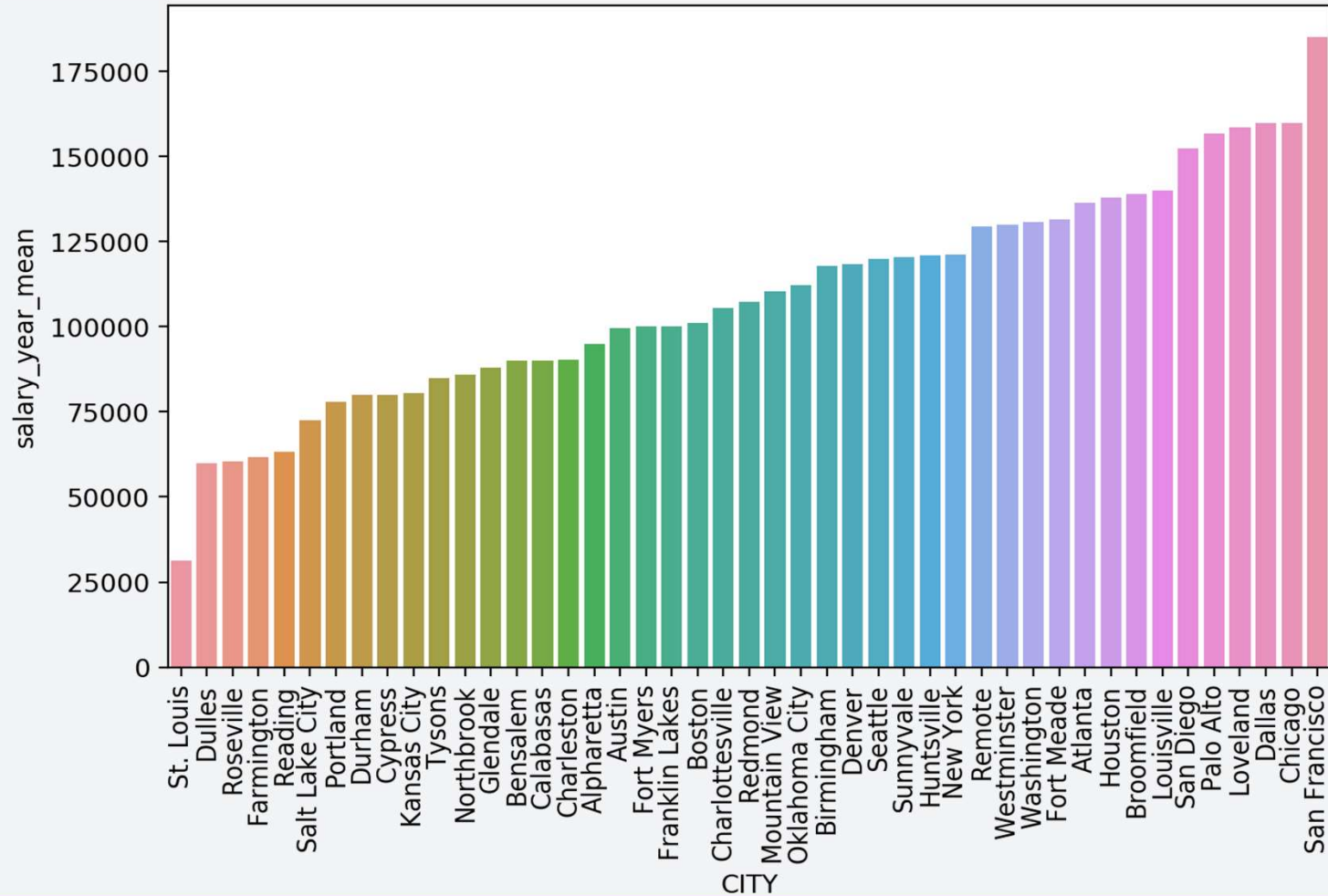
Notice to REMOTE !



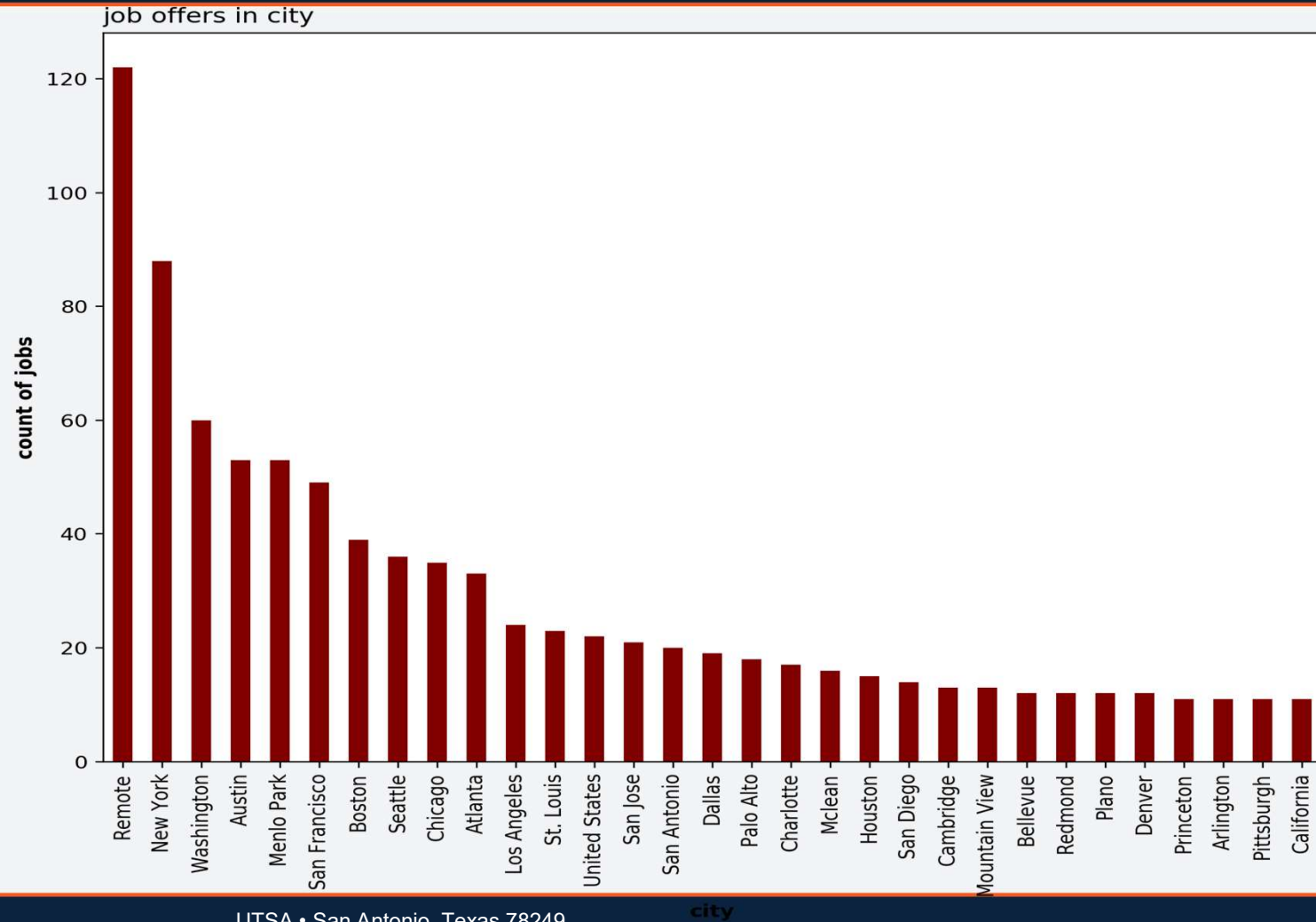
Lets look which states have the most job offer...



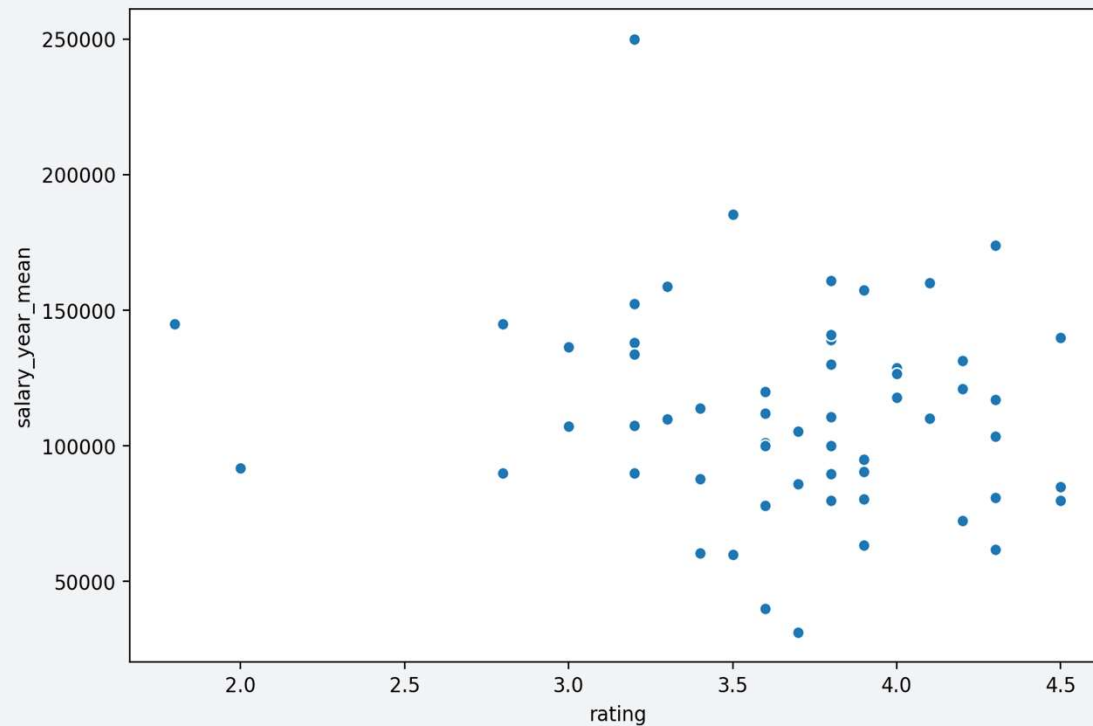
Cities and highest salaries



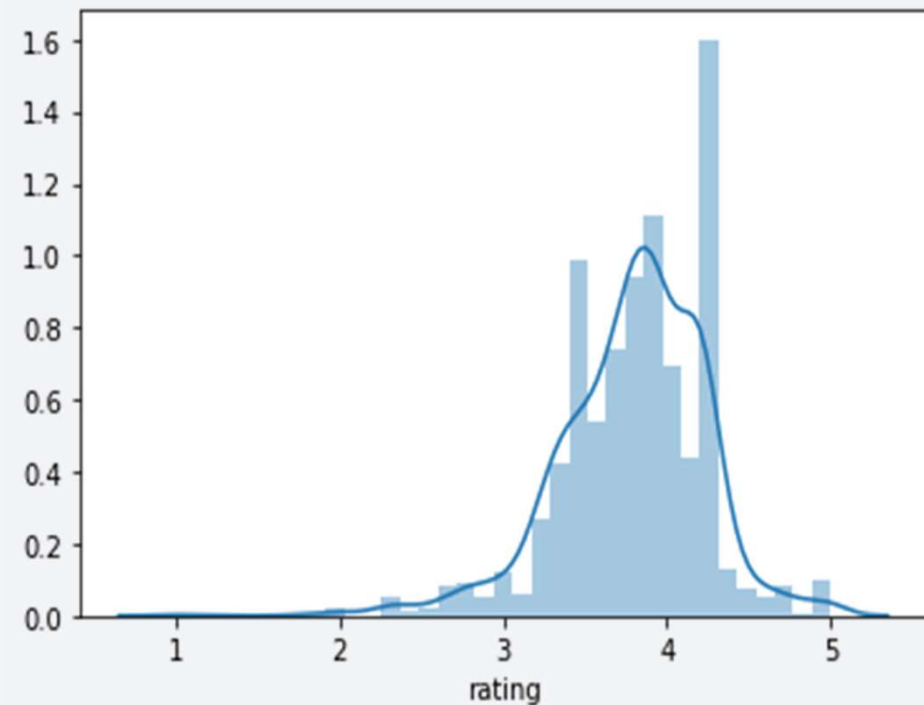
which
cities have
the most
job offer



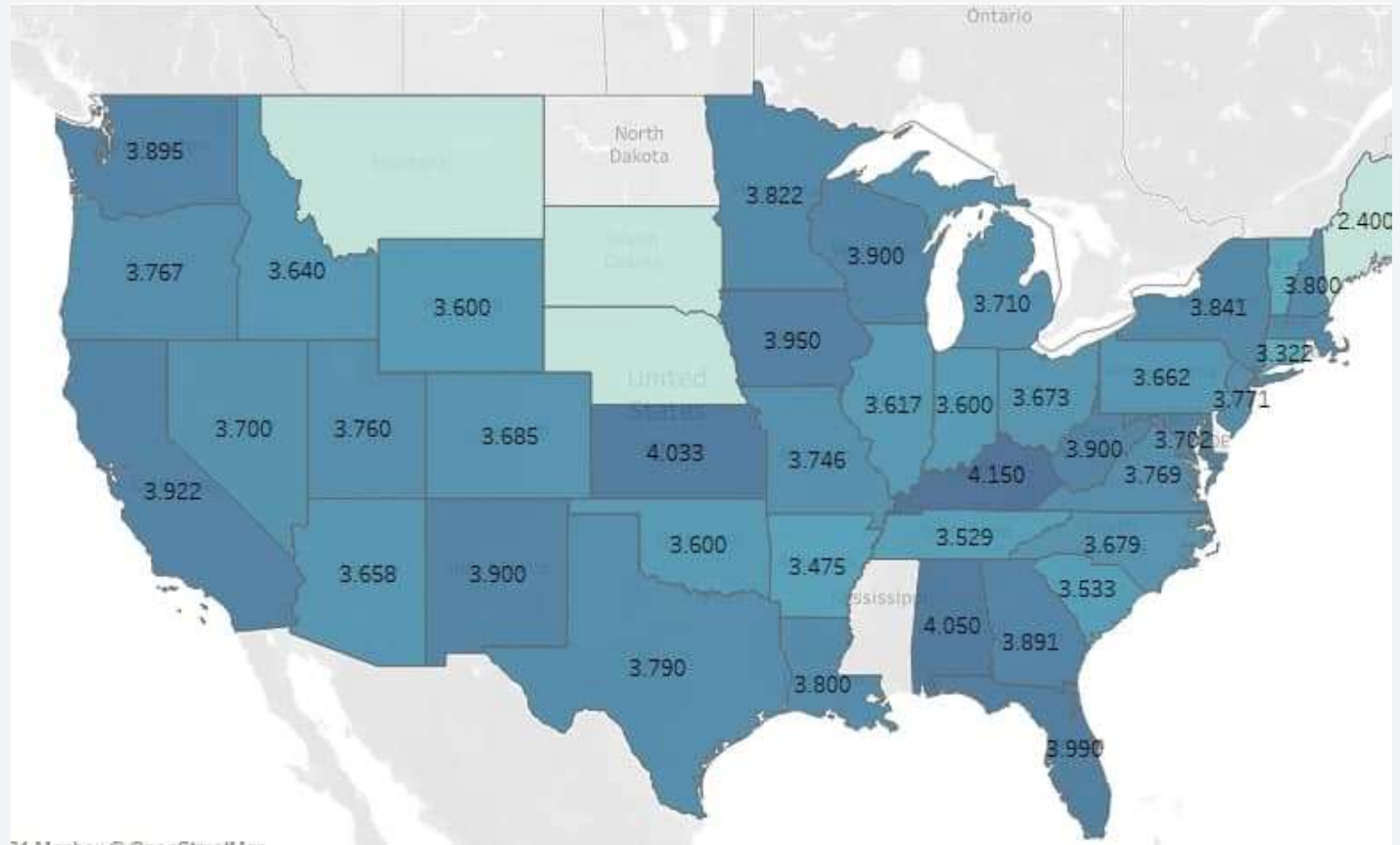
**It seems there is no relationship
between salary and rating of company!**



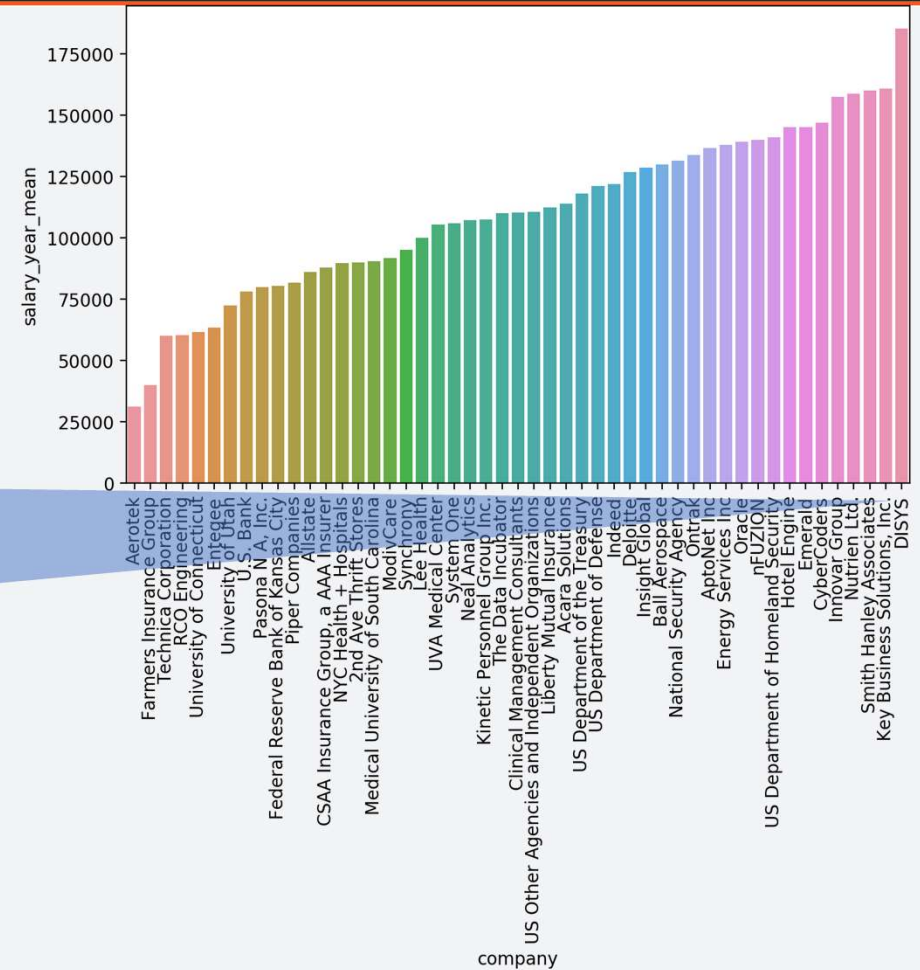
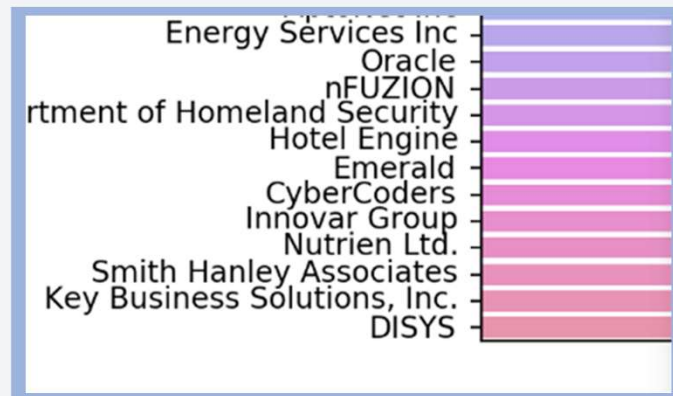
Distribution of rating



A brief look at map of rating



Top 10 companies by average of salary



Conclusion

- Although, California is one of the top 5 states which offers jobs with high salary, but it seems that it is just San Francisco where you can get very high salary. However as I mentioned about relationship between salary and rating, California's rating is not very high.
- Beside all of these cities and state, position of remote jobs are notable. In salary for example, it has a good level and it has also the most amount of job offers; which I believe is due to COVID-19.
- Texas as a second state in offering job for data scientists, is one of the important location:
 1. **Austin** which is the fourth city in the USA that is offering job
 2. **Dallas** which is third city in the country that has high salary in average.

Thank you