**Glassdoor: Estimating data science salaries**

## Introduction

1. **Problem Statement**

Salary expectations are always a key factor when looking for new job opportunities and, if you are considering breaking into data science, here is the great news: data scientists, as well as other data-related jobs, command some of the highest salaries in the tech industry.

According to the 2020 LinkedIn U.S. Emerging Jobs Report, data science has experienced continued growth on a tremendous scale in recent years. Across nearly every industry, organizations are competing for data science professionals to leverage their data and drive smart business decisions. However, since the supply of data professionals has not yet caught up with the demand, the competition between companies in hiring these professionals is fierce compared to in other tech sectors. As a result, employers are willing to pay top salaries to talented data scientists.

Notwithstanding the profitable honeymoon this field is experiencing, it is important to note that the data science salaries can range considerably across professionals. Employers take into account several factors when deciding how much to pay, including experience, skills, job title, or company size.

In this report, I will provide an overview of the model that predicts data science salaries in 2022. I will also discuss the different criteria that influence salary based on the data from Glassdoor, a job portal. To narrow our subject matter, I will only focus on the salary landscape in the U.S.

Unfortunately, many of the job postings available today do not offer information on the available salary range per open position, which is why I decided to collect, and model said information from Glassdoor.

1. **Background**

Salary is often the most guarded piece of information held by companies, but Glassdoor lifted the veil of secrecy in 2007 by making it possible for users to report the amounts of money they earned. The most important reason users post information considered private, such as their salaries, is because Glassdoor allows them to do so anonymously. At Glassdoor, our goal is to help match the right people with the right employers by leveraging workplace transparency, including offering more information around pay to help ensure pay expectations meet pay realities. - To better match the right people with the right companies, Glassdoor includes salary estimates in job listings. As of Sep 15, 2022, 30081 salaries were submitted anonymously to Glassdoor by Data Scientist employees.

**Glassdoor estimated salaries**. When employers do not provide salary information. Salary estimates generated by Glassdoor use salary data from millions of employees and third-party sources using machine learning algorithms. Salary estimates factor in recent user-generated salary reports for similar job titles at the company, its competitors, and other employers for a specific location.

**Employer provided salaries** - When employers provide salary information

**Shortcomings of Glassdoor data science model:**

* Salary estimates in job listings display a range for annual base or hourly pay and are specific to job title, company, and location. They do not factor in variable compensation, such as bonuses, commissions, tips, stock, benefits, or other components. They also do not factor in years of experience.
* Glassdoor generates salary estimates using salary data from current and former employees, applying emphasis on salaries submitted over the past two years. The algorithm takes into account inflation trends.

1. **Goal**

This project aims to provide a salary range for a data scientist in the U.S. with Mean Absolute Error of +-$20K. This is backed by cost of living data for top 40 cities of the U.S.

## Datasets

Tweaked the web scraper GitHub repo (below) to scrape 5578 job postings from glassdoor.com.

[scraping-glassdoor-selenium/glassdoor scraping.ipynb at master · arapfaik/scraping-glassdoor-selenium · GitHub](https://github.com/arapfaik/scraping-glassdoor-selenium/blob/master/glassdoor%20scraping.ipynb)

The web scraping was done on the 1st of September 2022 for 39 cities. With each job, I got the following:

1. Job title
2. Salary Estimate
3. Job Description
4. Rating
5. Company Name
6. Location
7. Company Size
8. Company Founded Date
9. Type of Ownership
10. Industry
11. Sector
12. Revenue

First 100 Glassdoor job postings for following cities are accessed through the search page: [Glassdoor Salary Search](https://www.glassdoor.com/Salaries/index.htm)

1. San Francisco, CA
2. New York City, NY
3. Raleigh, North Carolina
4. Philadelphia, PA
5. Dallas, TX
6. Miami, FL
7. Atlanta, GA
8. Detroit, MI
9. Cleveland, OH
10. Denver, CO
11. Orlando, FL
12. St. Louis, MO
13. Charlotte, NC
14. Salt Lake City, UT
15. Columbus, OH
16. Las Vegas, NV
17. Kansas City, Mo
18. Indianapolis, IN
19. Cincinnati, OH
20. Phoenix, AZ
21. Portland, OR
22. Houston, TX
23. Seattle, WA
24. Austin, TX
25. Boston, MA
26. Washington D.C.
27. Arlington, VA
28. Los Angeles, CA
29. Chicago, IL
30. Minneapolis, MN
31. Baltimore, MD
32. San Diego, CA
33. Irvine, CA
34. Tampa, Fl
35. Nashville, TN
36. Santa Barbara, CA
37. Oakland, CA
38. Fremont, CA
39. Honolulu, HI

I selected the top 39 cities (above) based on cost of living index data. Cost of Living Index was accessed through Numbeo: [Cost of Living Index 2023 (numbeo.com)](https://www.numbeo.com/cost-of-living/rankings.jsp). This data was also sourced as of 1st of September 2022.

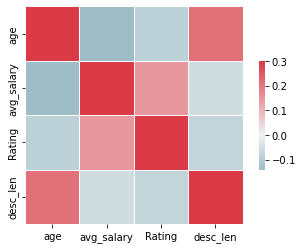
## Data Cleaning and Data Wrangling

After scraping the data, I needed to clean it up so that it was usable for our model. I made the following changes and created the following variables:

1. Added Cost of Living Index per city to the dataframe from data sourced from Numbeo
2. Removed 2090 duplicate job listings from the initial 5578 job listings. After this step, dataset had 3488 job listings
3. Removed 139 rows with no salary information. This is around 4% of the total number of rows
4. Made columns for employer provided salary, employer estimated salary and hourly wages
5. Made a new column for:
   1. company state
   2. simplified job title and Seniority
   3. job description length
6. Transformed founded date into age of company and replaced null values with median value of the column
7. Replaced null values in rating with median value
8. Made columns for if different skills were listed in the job description:
9. Python
10. SQL
11. Reviewed size information per company and simplified data into 8 categories. Used Ordinal coding for encoding
12. Reviewed revenue information per company and simplified data into 10 categories. Used Ordinal coding for encoding
13. Reviewed type of ownership information per company and simplified said category from 12 categories to represent 3 relevant business categories
14. Reviewed sector information per company and simplified said category from 26 categories to represent 10 sector categories
15. **Creating zones:** Mapped city location by geographical location in the United States and created a "Geograpic Map" category for modelling.
16. Standardized numerical data

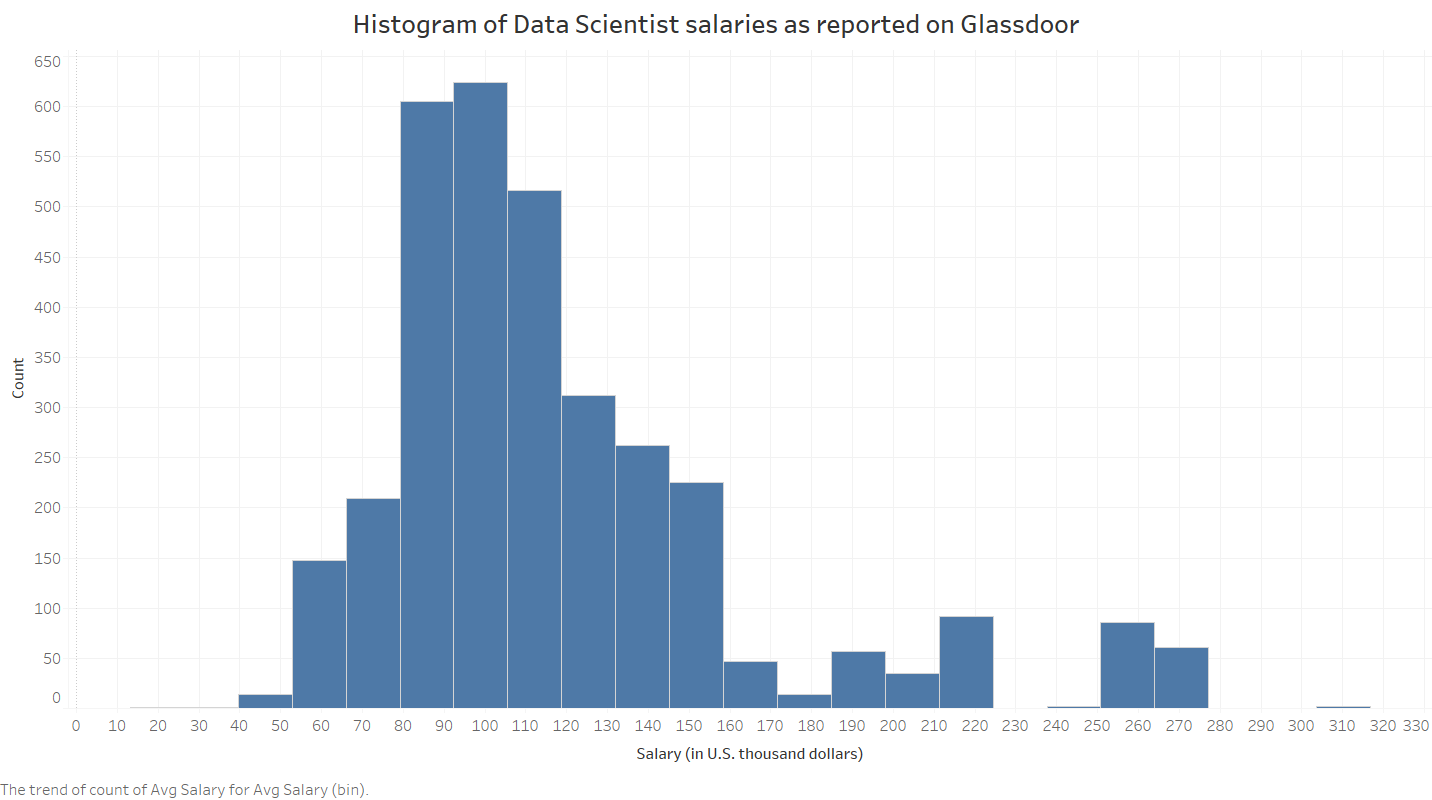
## Exploratory Data Analysis and Initial Findings

I looked at correlation of the variables. None of the variables had correlation greater than 0.2. This shows that independent variables were not correlated.



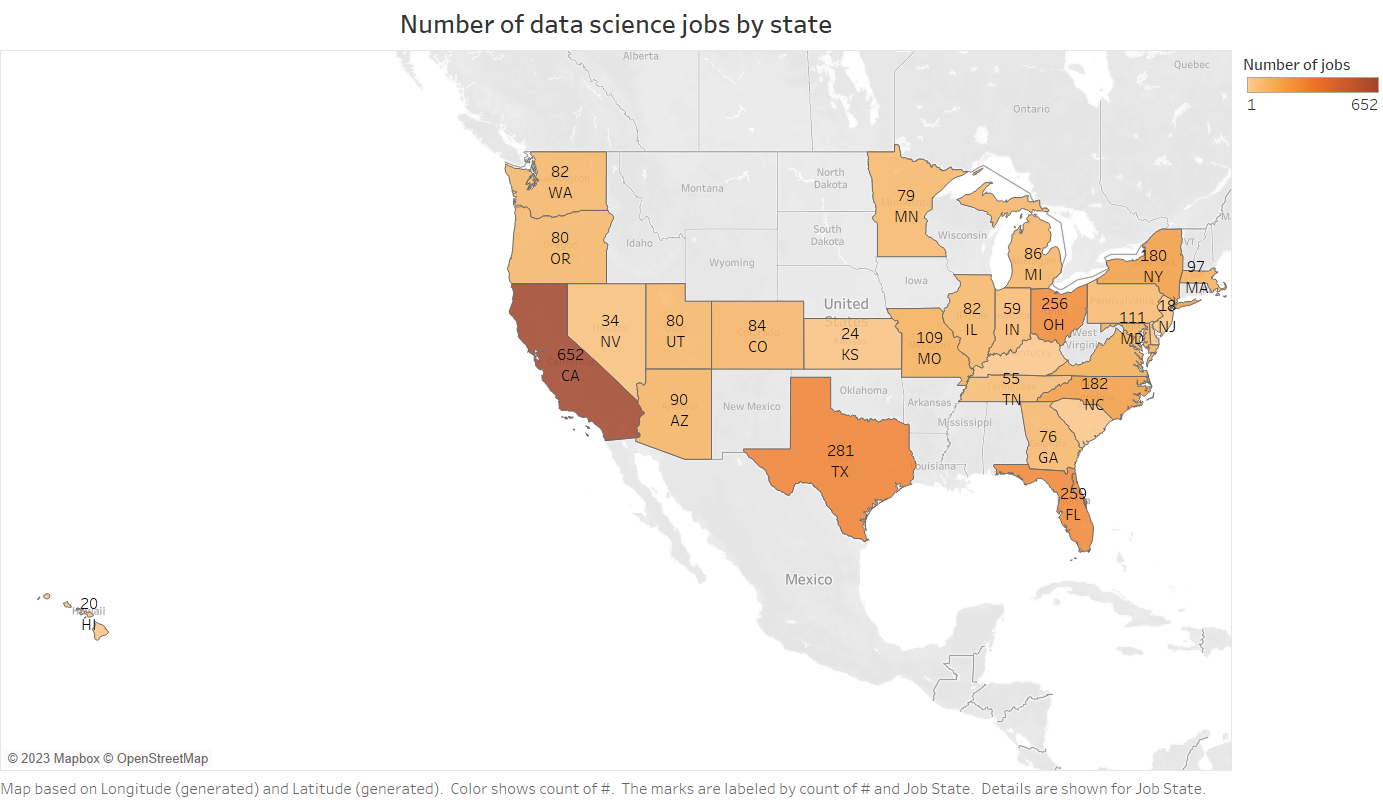
Although, my scrapper code searched only for 'data scientist' roles on Glassdoor, data got pulled for many other types of roles like data analyst, ML engineer, Data Engineer, and Manager level positions.

This allows us to understand something interesting things about the analytics filed as a whole.

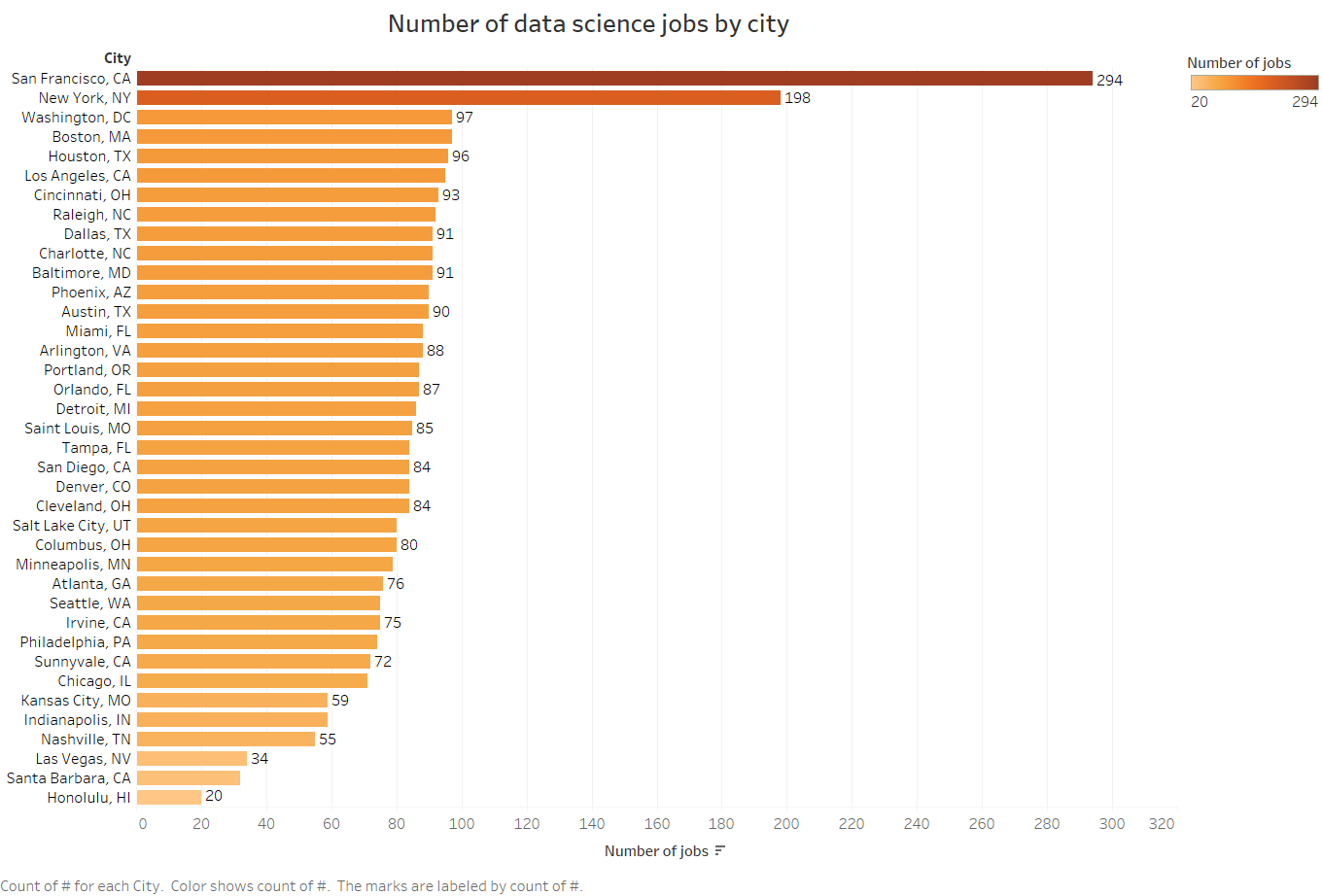


The histogram of salaries as reported on Glassdoor show a mean salary of $119,312.71 across the continental United States, with a standard deviation of $46,602.00.

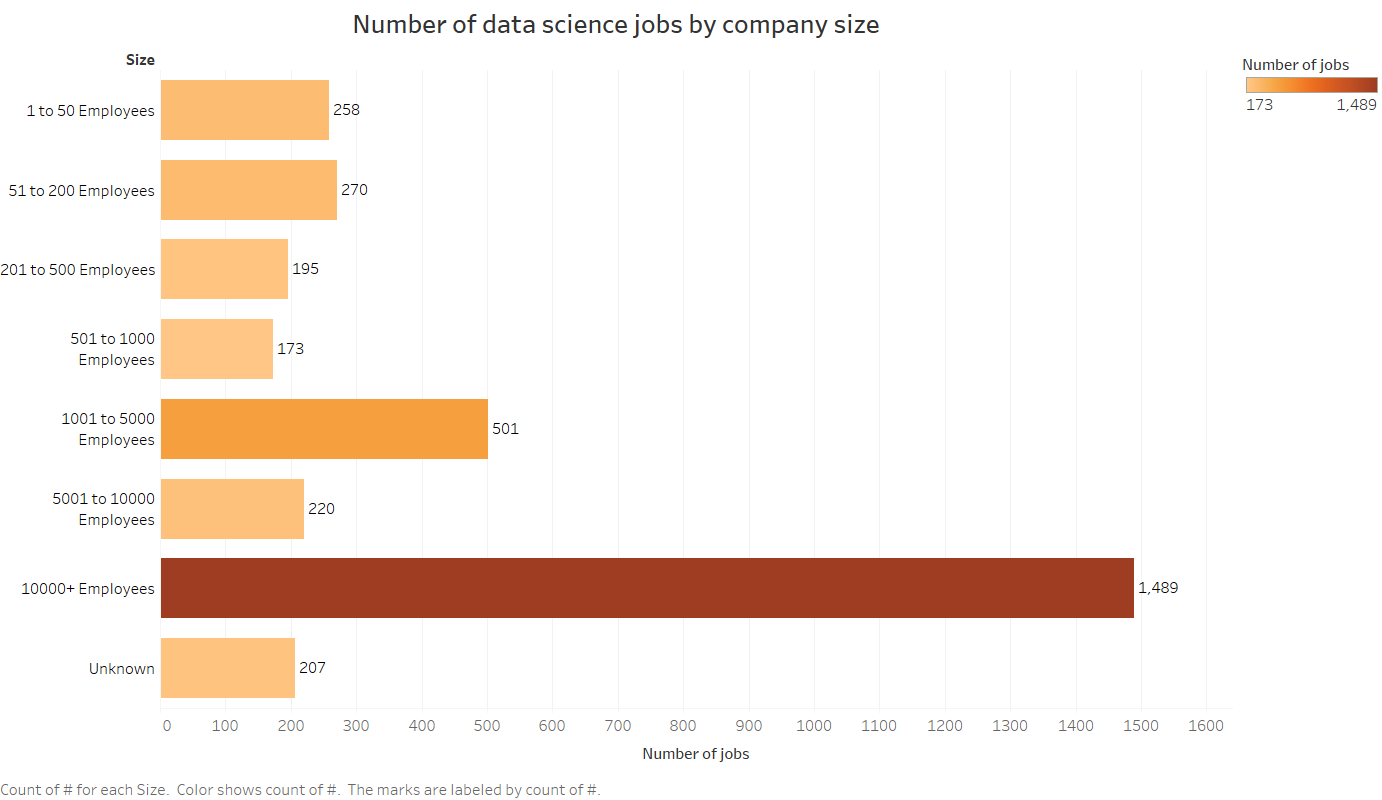
1. **Location**: First, let’s look at where these positions are being hired.
   1. By state: As we would expect California leads this category with a little around 20% of the total job postings. Other states with the largest cities in the country also top this list like Texas (8.48%), Florida (7.82%), and Ohio (7.73%).



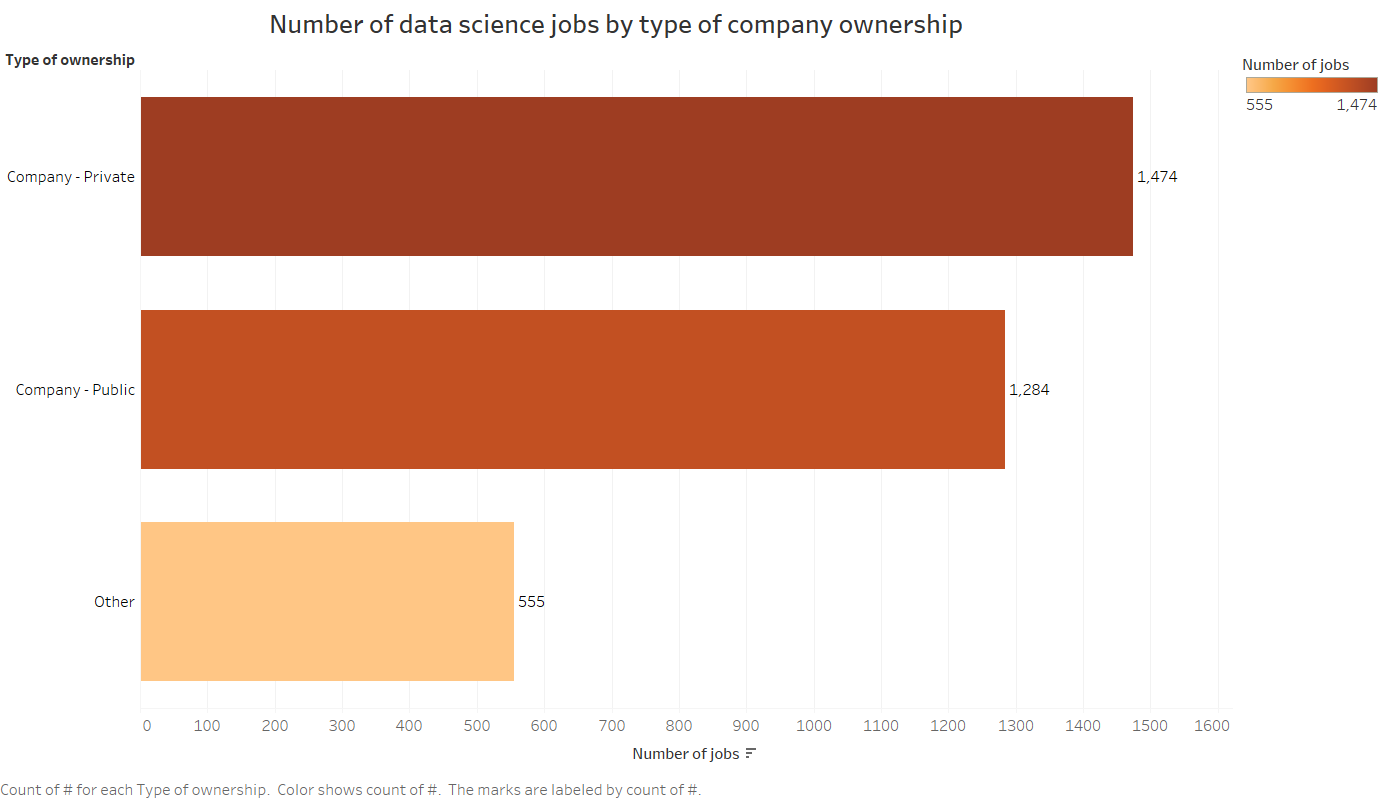
* 1. By City: If you are looking for a data scientist position, your best bet is somewhere near one of the largest cities like San Francisco, New York City, Boston, Washington D.C. Top 10 cities have around 40% of the total data science job.



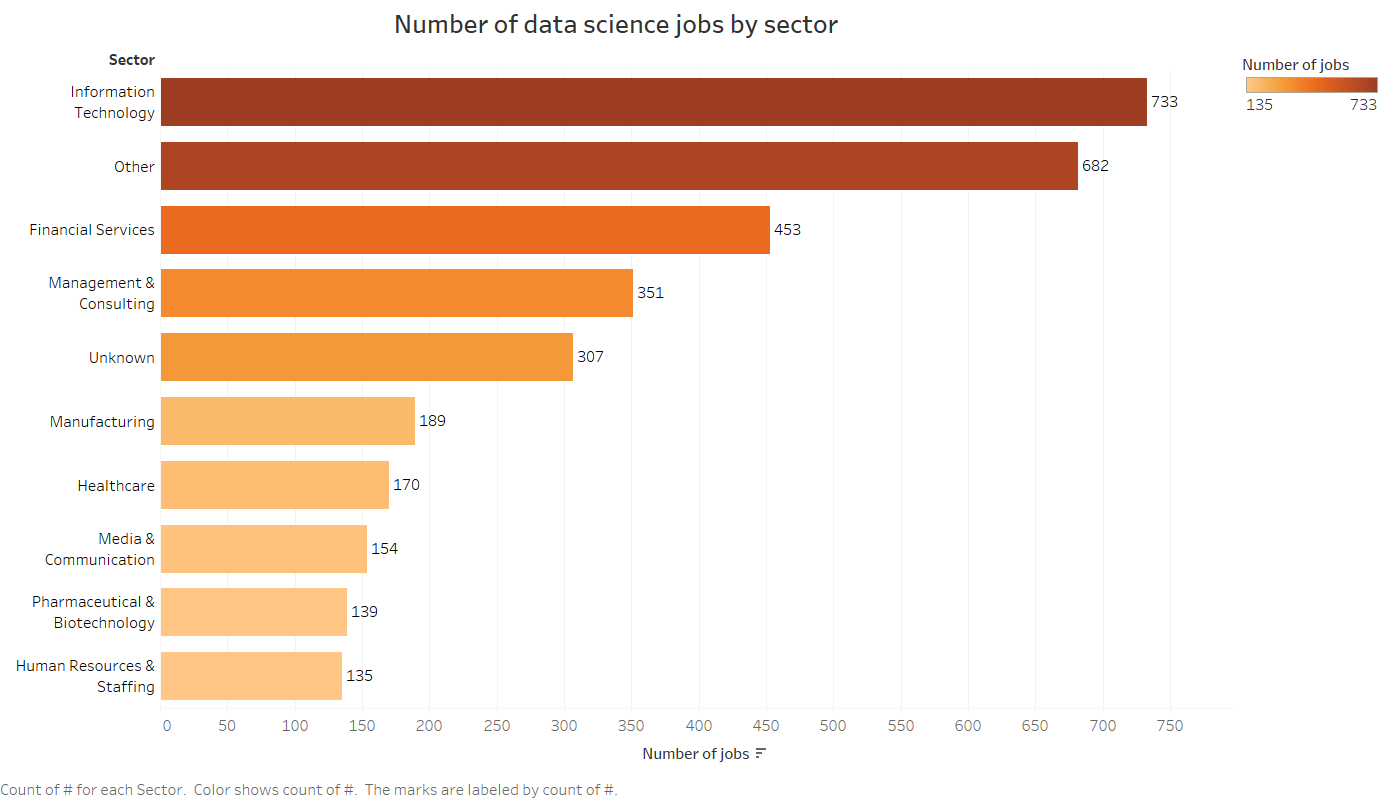
1. **Who’s hiring**: Let’s look at who is hiring.
   1. By size: From this sample, looks like large sized companies are hiring most of the data scientists.



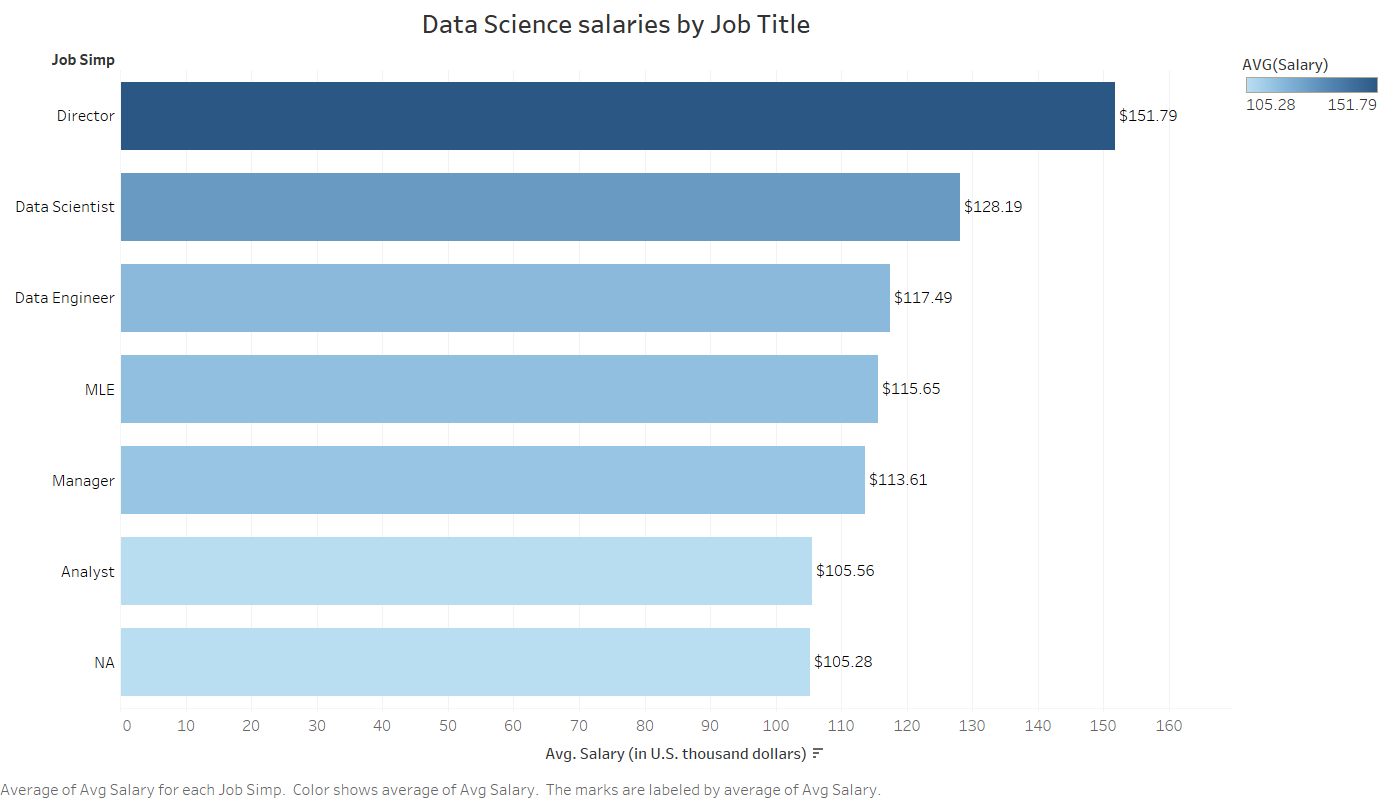
* 1. By type of ownership: Private companies are taking the lead here with almost half of all data science jobs being hired by private organizations.

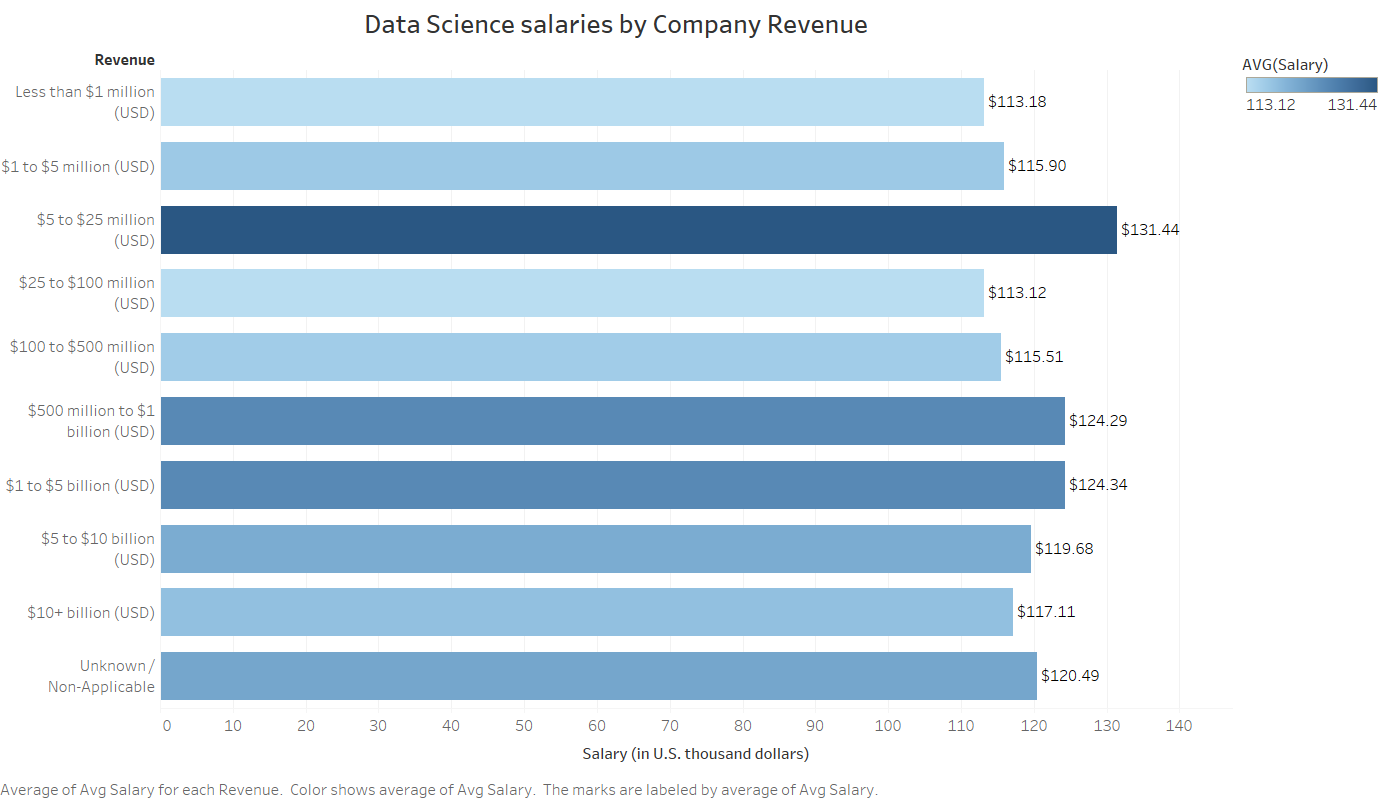


1. **Sector:** Companies in IT, financial services, management & consulting, and manufacturing top the list of the most popular jobs by sector. If you have a hybrid skillset where you have experience with one of these domains, it can greatly help you create an opportunity for yourself. It should be noted that these can be slightly skewed by the type of companies that would post their jobs online. Small companies are likely underrepresented here because they are less likely to post their jobs on websites like Glassdoor or LinkedIn.

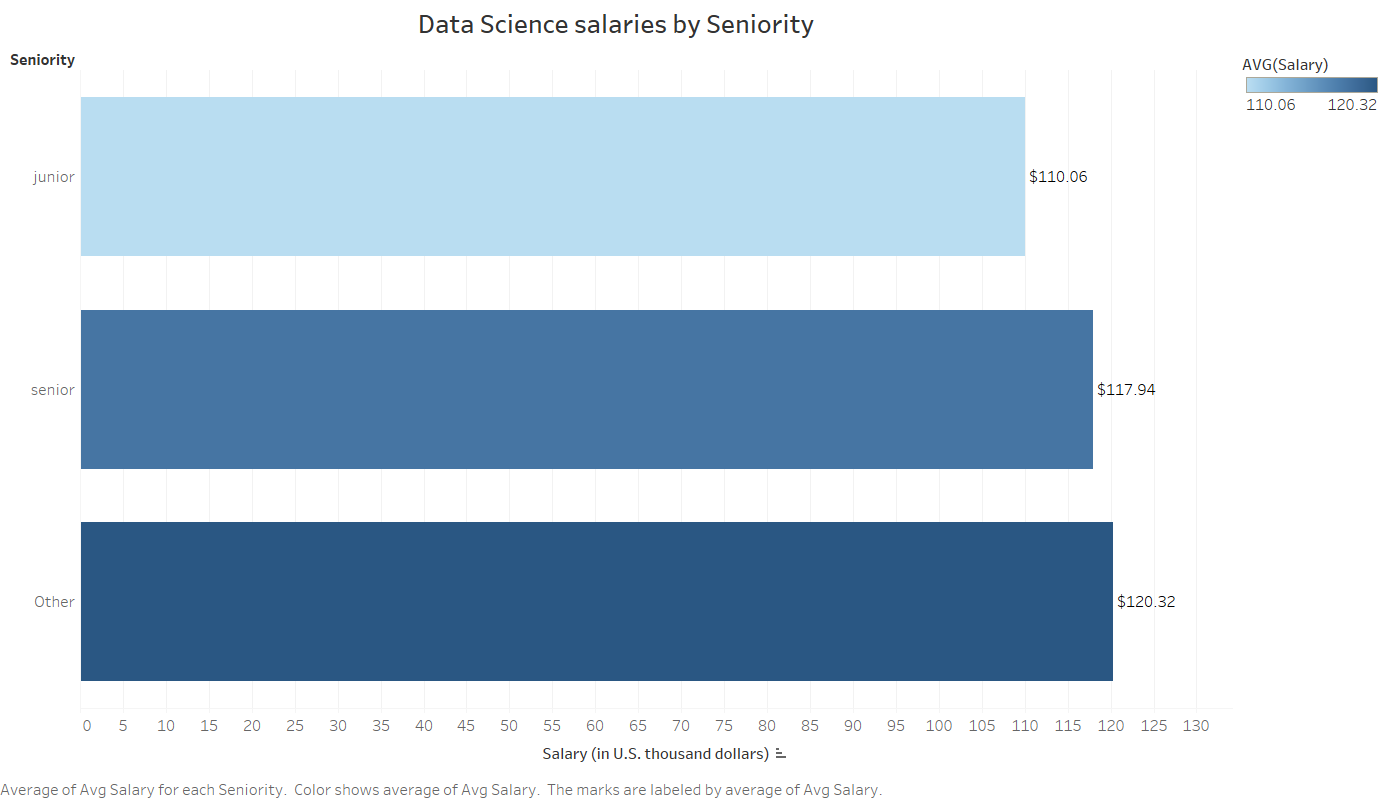


1. **Programming skills**: In terms of programming skills, 296 jobs mentioned python and 195 jobs mentioned SQL in their job requirement. As I could not get complete job requirements during scraping, I can not make a strong conclusion about the requirement of programming skills in data science roles. I need to collect more data to explore this aspect.
2. **Expected Salaries**: I also looked at expected salaries across the job types, company revenue and locations. From the estimates on Glassdoor, it looks like Directors make the most followed by data scientists and then data engineers. What was surprising is MLEs were expected to make significantly less than the data scientists and data engineers. Also, managers were making less money than the data scientists and data engineers they were managing.



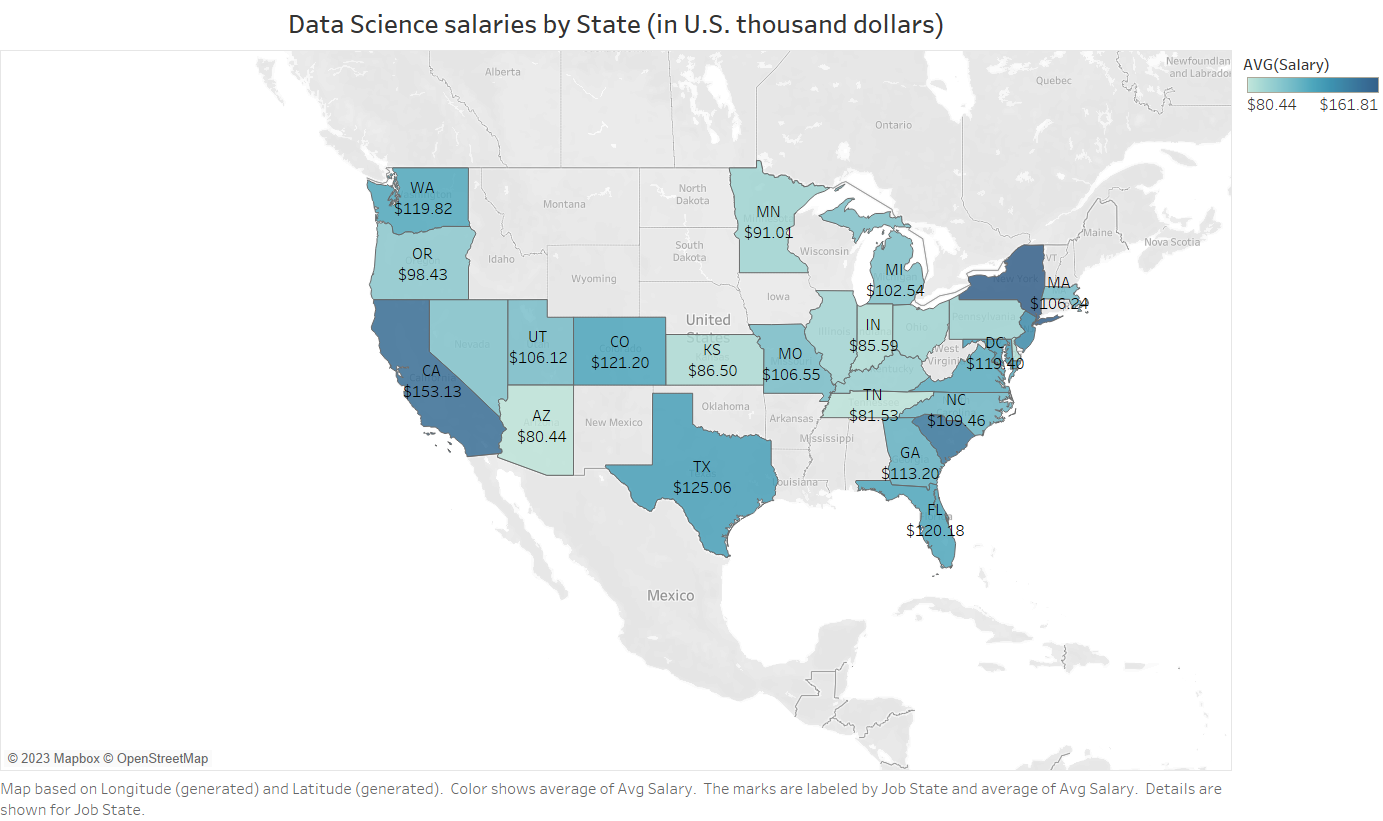


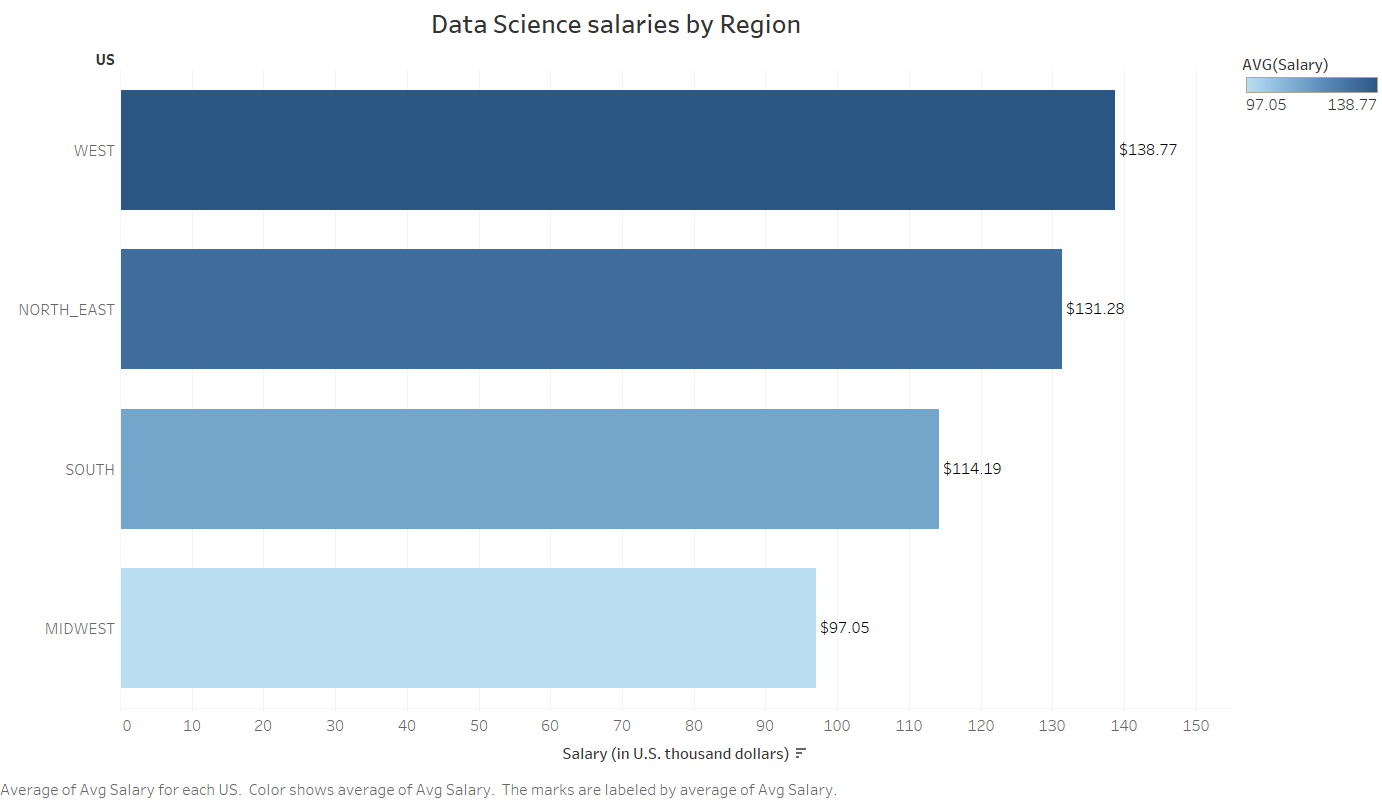
Salary per company revenue show that working for a $5 to $25 million company definitely does pay a bit higher compared to other companies, while Unknown/ Non-Applicable revenue refers to Non-Profit, Education, and Government salary ranges.



Senior data scientists can expect to get 10K higher than the junior roles.

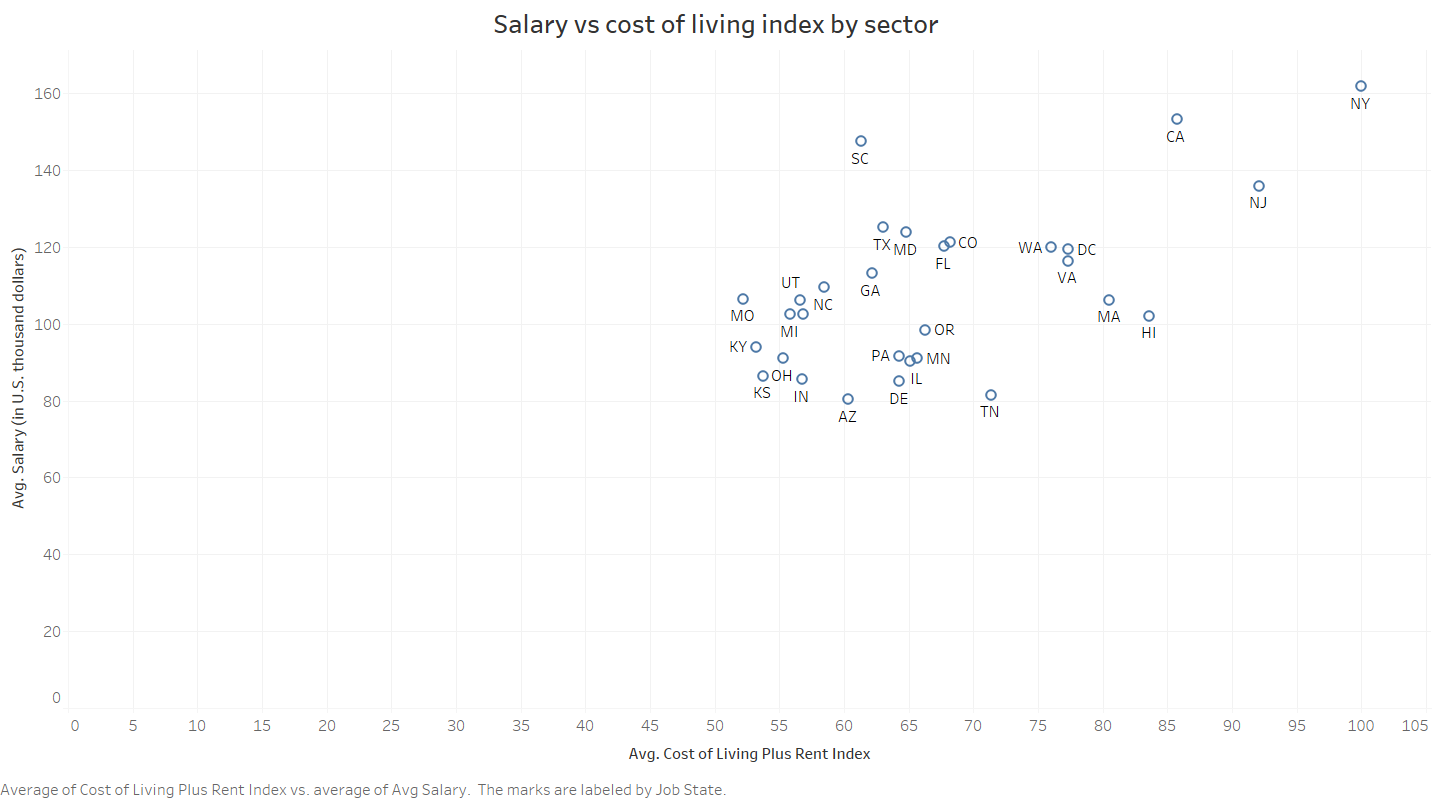
1. **Data scientists’ salary by state:** It appears NY and California lead this list.





Geographical location was mapped onto the city in which the salary was posted, with a higher representation in the West Coast due to the number of cities associated with the Tech Industry in California, which also contributes to the increase in Salary range for that area.

Hawaii and Massachusetts offered low salaries even though they have a very high cost of living. If someone wants to get most bang of their buck, they will consider positions in South Carolina, Texas, California, Maryland, or North Carolina.



## Model Building

First, I transformed the categorical variables into dummy variables. I also split the data into train and tests sets with a test size of 20%.

I tried four different models and evaluated them using Mean Absolute Error. I chose MAE because it is relatively easy to interpret, and outliers aren’t particularly bad in for this type of model.

I tried three different models:

* Multiple Linear Regression – Baseline for the model
* Lasso Regression – Because of the sparse data from the many categorical variables, I thought a normalized regression like lasso would be effective.
* Random Forest – Again, with the sparsity associated with the data, I thought that this would be a good fit. I tuned following hyperparameters for this model:

|  |  |
| --- | --- |
| **Hyperparameter** | **Value** |
| n\_estimators | 10 to 300 |
| Criterion | Sqaured\_error, MAE |
| max\_features | auto, sqrt, log2 |

* XGBoost – I thought this would remove any overfitting. I tuned following hyperparameters for this model:

|  |  |
| --- | --- |
| **Hyperparameter** | **Value** |
| n\_estimators | 10 to 300 |
| Max\_depth | 3 to 5 |
| Min\_child\_weight | 0 to 5 |

## Model performance

The Random Forest model far outperformed the other approaches on the test and validation sets.

Linear Regression: MAE = 23.39

Lasso Regression: MAE = 23.26

Random Forest: MAE = 16.86

XGBoost: MAE = 32.27

## Future Work

1. To add more sources for salary like H1B data to make salary data more reliable
2. To add more observations for programming skills, qualifications and years of experience required for data science jobs
3. To get stock information along with salary as in tech sector stocks are a major portion of compensation. This will give a better picture of the data science compensation.
4. To productionize this model