# Geek.Head()

## So You Want to Be a Data Analyst?

### Job Prospects for Data Analysts in the United States

Matt Blaylock | Monicah Cloud | Marvin Fontaine | Tina Rowland

*Create a write-up summarizing your major findings. This should include a heading for each "question" you asked of your data, and under each heading, a short description of what you found and any relevant plots.*

The objective of this project, based on the provided dataset, is to determine where the job prospects for data analysts in the United States are most promising.

This data set contains 2253 job listings for data analyst positions. The features include job title, salary estimation, job description, rating, company name, and location.

Using Jupyter Notebook, we imported the libraries and read in the CSV files where we instantly noticed that the first column is an unnamed column with no valuable data so we removed the unnamed column.

We then checked for missing values.

The dataset showed that there were no null values, but after observing the dataset head and tail we saw that instead of null value there were -1, so we replaced -1 with np.nan so we can calculate the missing values in the dataset for better observation.

We noticed that the columns Easy Apply and Competitors had more than 50% of its values missing, so we dropped those columns.

In Company Column and Job Title columns we saw that data have a suffix showing either the company department or the rating of company. We eliminated everything after the suffix and only kept the Company Name, because we already have a seperate Rating Column, and the Job Title column with only the job titles.

Looking at the Salary Estimate column, we saw that there were a lot of special characters and text. We only needed the estimated salary (in integer format). We began by removing the text (Glassdoor est, $, K, and -) from the column. Then we split Salary Estimate into two columns, Min Salary and Max Salary.

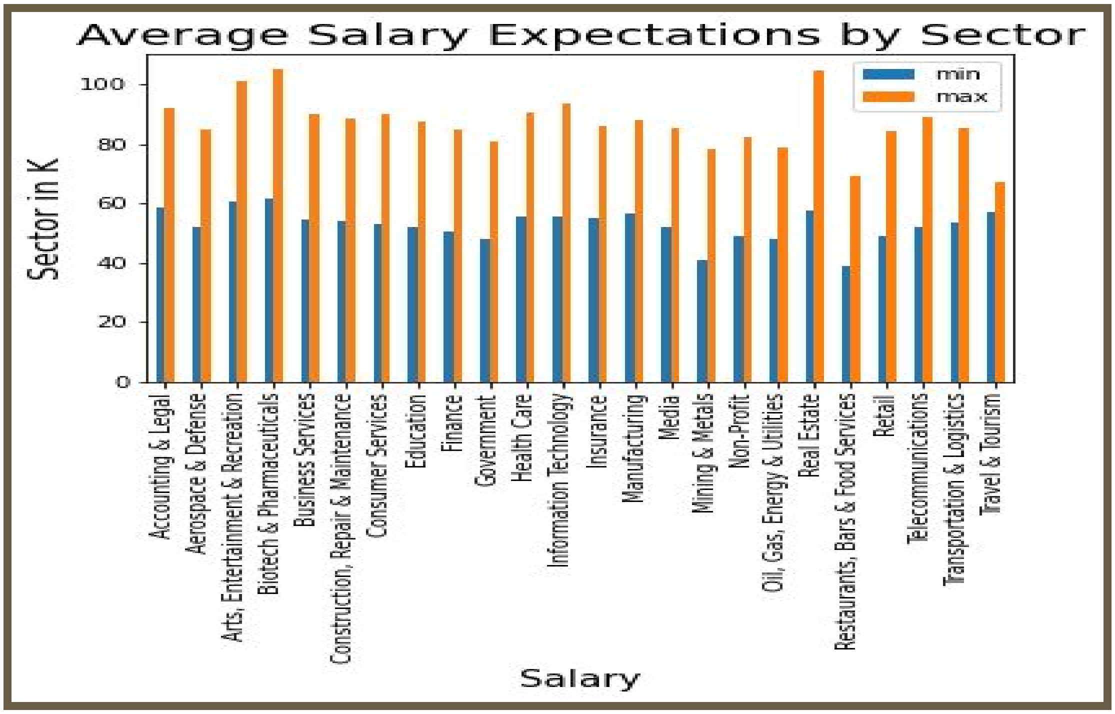
We removed the Revenue column because we didn't have a need for it for the questions we were answering.

We have Successfully cleaned the data.

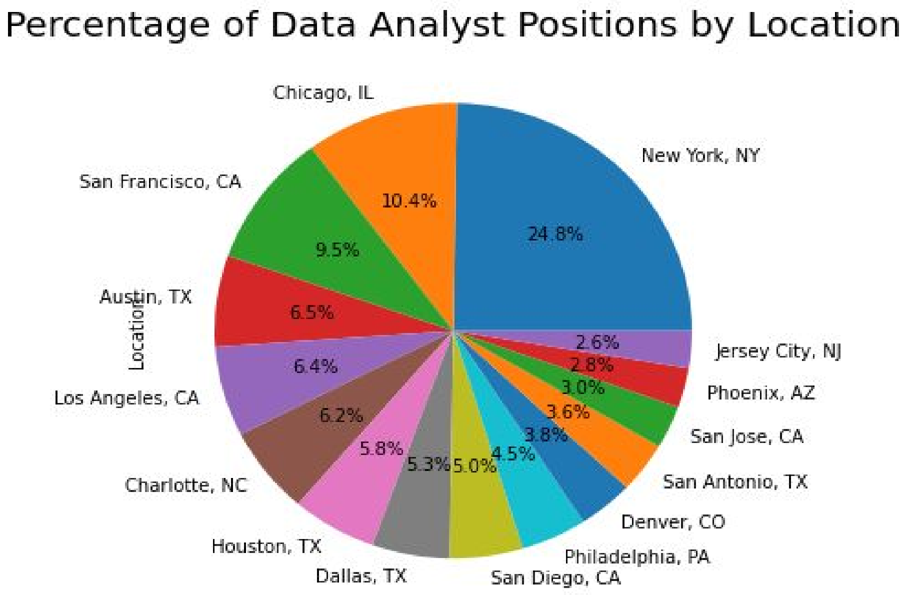
To meet the objectives of the project, the research questions detailed in the following section were developed.

## Research Questions

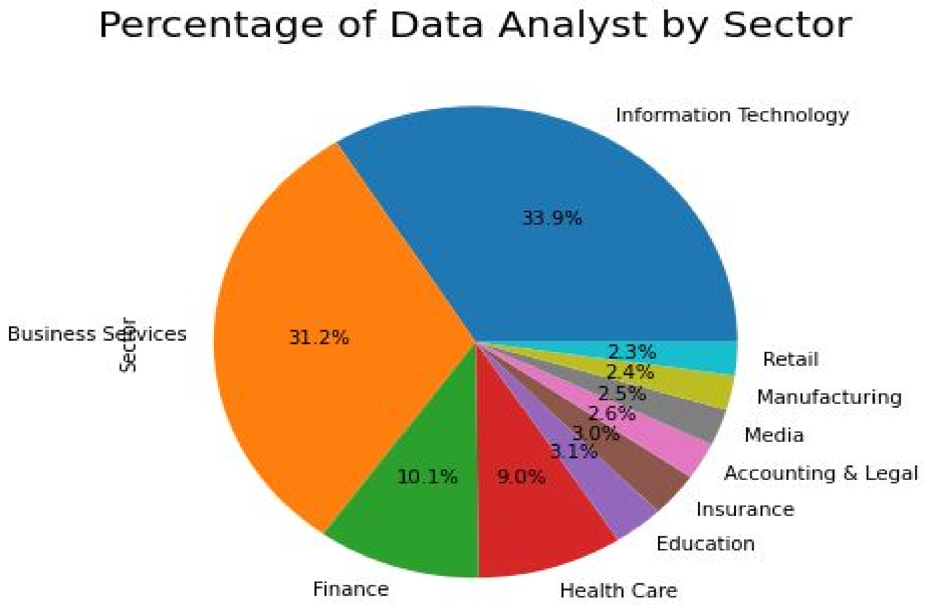
### Which employment sectors have the best salary expectations for Data Analysts?



### What are some of the better locations for job prospects?

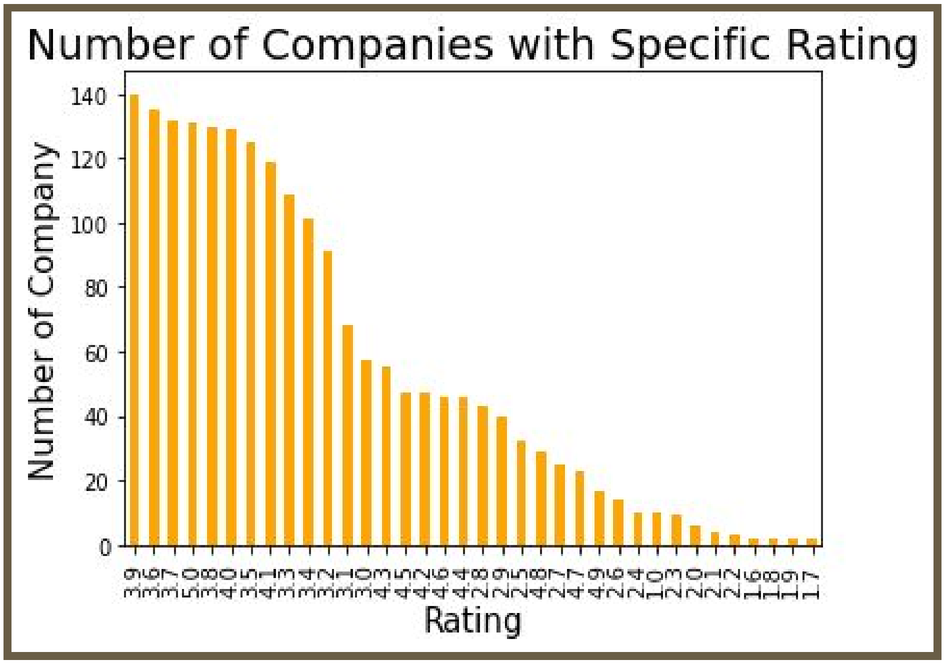


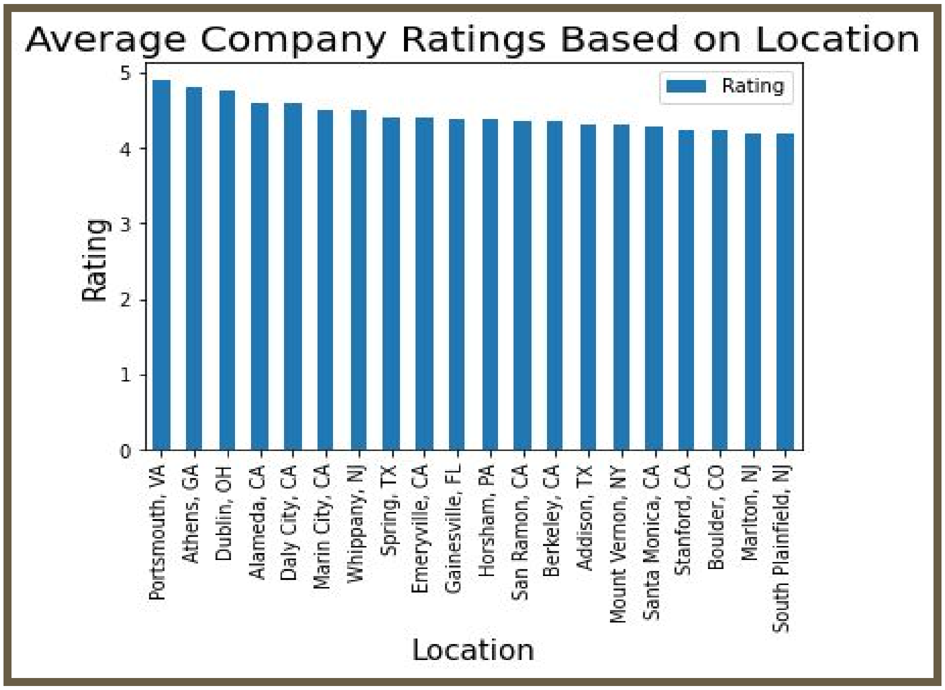
This pie chart shows the percentage of Data Analyst positions by location. New York had the highest number of percentage hiring of 25% followed by San Francisco and Chicago respectively.



This pie chart showed the percentage Data Analyst by Sector. The Information Technology led the sector with the highest hiring of Data Analysts followed closely by the Business Service and the financial industry.

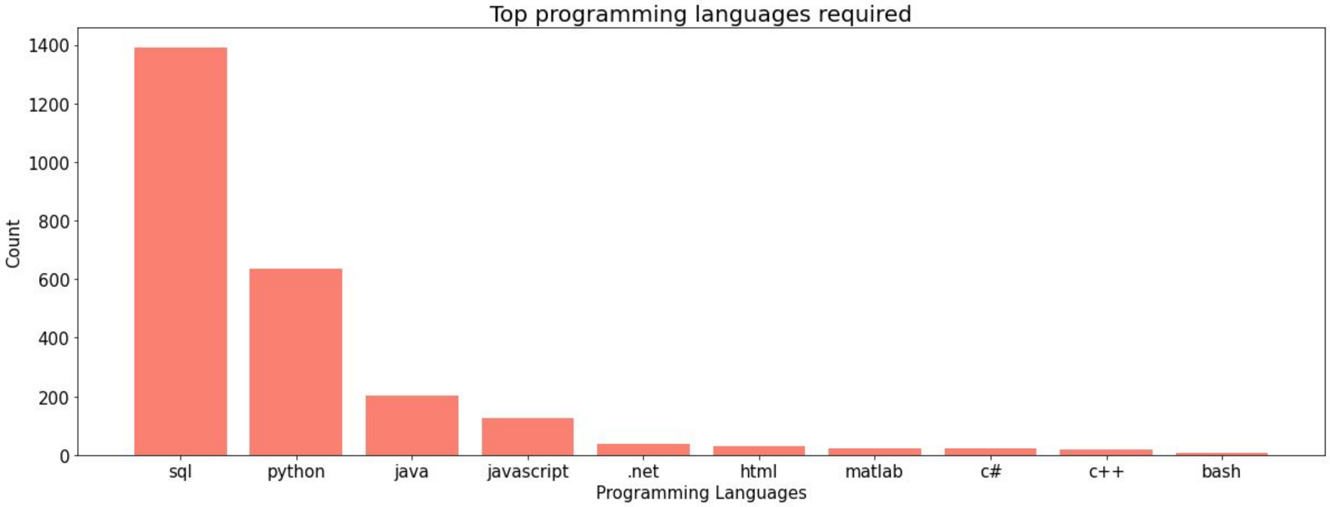
**What are the ratings of the companies that hire Data Analysts?**





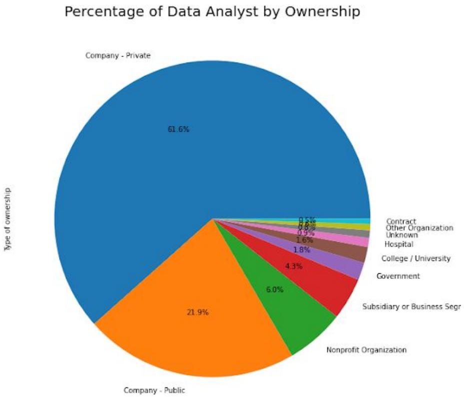
The two charts show the number of companies with specific ratings and average ratings based on Location. There were 139-140 companies with 3.9 star ratings but noted that Companies in smaller cities had higher ratings than those in large cities.  Companies in Portsmouth-Virginia, Athens-Georgia and Dublin-Ohio were the top three highest rated companies.

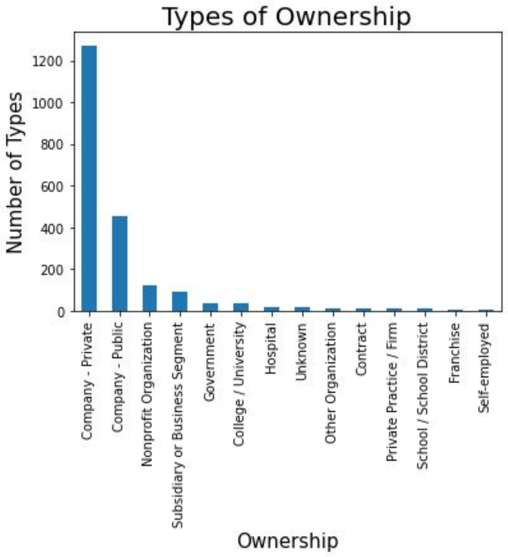
## Additional Insights



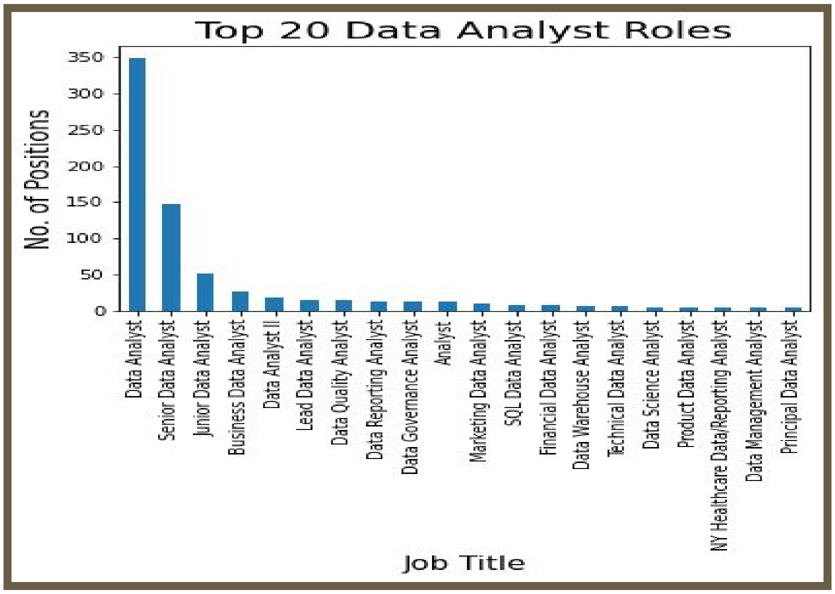
In this bar graph we were able to see which programming languages are most required as a data analyst and here we can see that SQL is in high demand followed by Python, and then Java.

This gave great insight and answers to a question not included in the proposal.





This bar graph shows the types of ownerships like private and public and government as such. We can see that the Private companies are seeking more data analyst jobs than any other type of ownership by 61.6%. Both graphs show the same type of data.



This shows the top 20 most Popular Jobs of Data Analyst roles. This could be because a data analyst is almost always an entry level role which is probably why it has a great number of positions for this role.

## Limitations

The study data included in this project was based on survey responses from Glassdoor. Limiting the participation to one source could skew the results. It's assumed that the data collected was collected via self-reported survey responses. As such, there was the potential for bias. For example, social desirability bias could lead to respondents to respond in a way to make the results appear better than they actually are. Also, with a static data set, the results only reflect the outcome at the time the dataset was published.

## Conclusion

This project set out to determine where the job prospects for data analysts in the United States were most promising. The project uncovered that the most in demand skills for data analysts are SQL, python, and java. It was also shown that the private sector had the highest job availability for data analysts. When it comes to the job sectors reviewed, IT, business services, and finance lead all sectors.

Finally, after analyzing the available data, a definitive recommendation can not be given. But, based on the totality of the review, an individual should be able to determine the specific place, company, or sector to focus his or her job search efforts.

## **Post Mortem**

If we could go back, analyzing whether or not there was a direct correlation between job rating and job size would've been very valuable information. Sometimes we can get caught up on the name and size of the company, and fail to realize that those attributes don't always necessarily correlate when it comes to how much you enjoy your job and work environment.