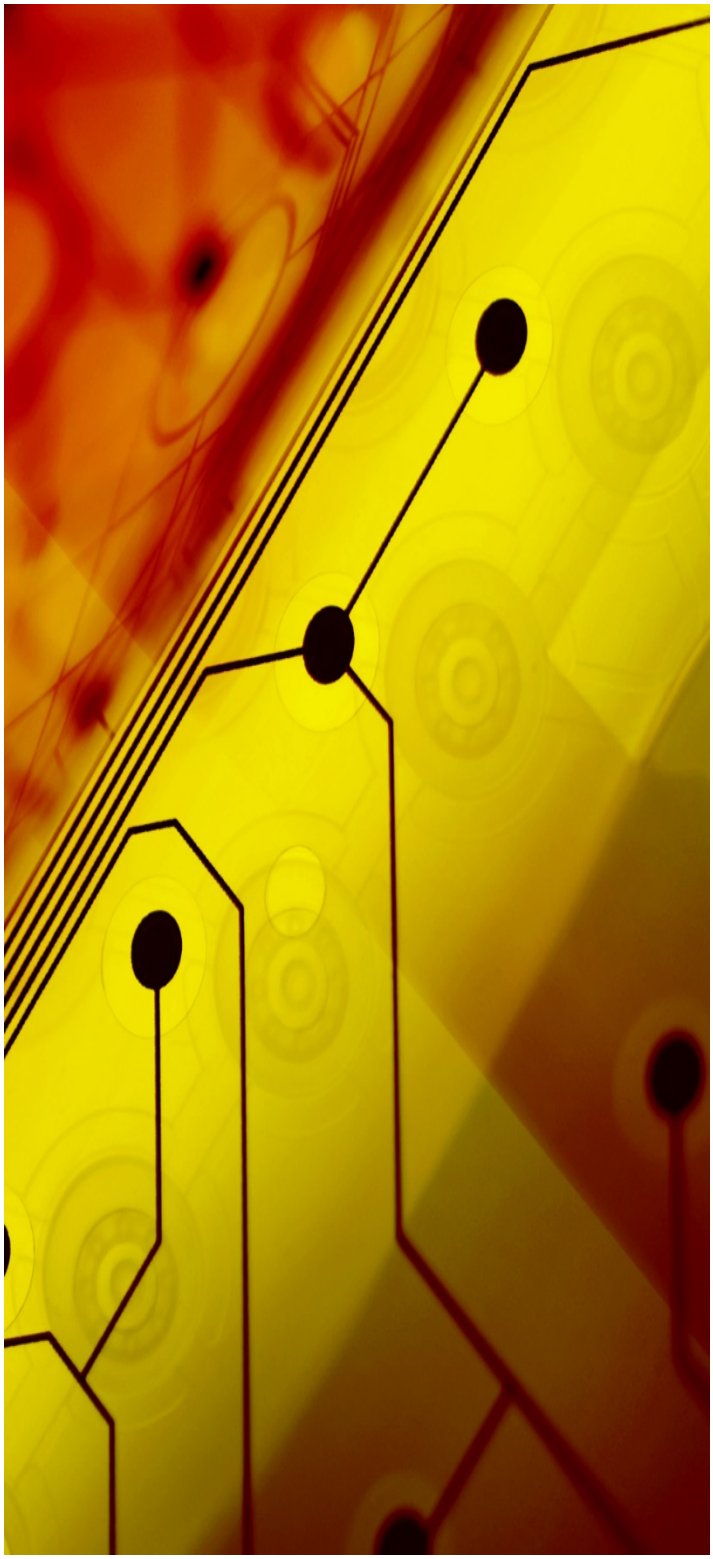


A First Job in Data Science

Leigh Ann Kudloff

Nataly Valenzuela Mullen

Data Science Tools 1

Winter Quarter 2021

March 2021

Introduction

Like most Data Science students, concern about job opportunities in the near future looms heavily at the halfway point of this degree program. In one year, employment is necessary to utilize skills learned, pay off student loans, and enter the world of data science. This project focuses on the world of job hunting for the first time in Data Science. The goal is to explore job descriptions to categorize skill sets, look for patterns, and prepare tools for the job hunt later this year. Through the exploration of trends in data science job postings, the concepts learned from Data Science Tools 1 and other classes will be applied and hopefully some ideal jobs will emerge.

GitHub Repository

The Jupyter Notebook is shared on GitHub at this site:
https://github.com/natvalenz/TOOL1_FINAL_PROJECT
Here is the binder link: # TOOL1_FINAL_PROJECT
[![Binder](https://mybinder.org/badge_logo.svg)](https://mybinder.org/v2/gh/natvalenz/TOOL1_FINAL_PROJECT/master?filepath=FinalProject.ipynb)

Dataset and Motivation

How the Dataset Was Collected

The dataset is comprised of eight datasets obtained from Kaggle pertaining to Data Science jobs posted since January 2020. The majority of these datasets contain Glassdoor job postings with a few datasets from Indeed and LinkedIn included as well. The original dataset included over 30,000 job postings. Through the data cleaning process, duplicates were removed resulting in just over 27,000 job listings. The postings include only jobs in the United States. The job titles considered included: Data Scientist, Data Engineer, Data Analyst, and many others. Cost of living data for 50 US cities obtained from this website: https://www.numbeo.com/cost-of-living/region_rankings.jsp?title=2020&region=021

Description of the Dataset

After merging the eight datasets, the final dataset had 18 columns or attributes. The 18 columns include: Job Title, Job Description, Company Name, Location, City, State, Remote, Size, Founded, Industry, Rating, SalaryMin, SalaryMax, Salary Estimate, Hourly, Junior, Senior, and DF. Using NLP, additional columns for temporary work (temp), and part time work were added, and additional columns were added for remote, junior, and senior. After the job descriptions were processed using NLP, another column was created

for the newer clean version of the job descriptions and job titles. In the search for finding the "best" job, additional data regarding cost of living was merged with the final dataset. In this merge of data frames, these additional attributes were added to the data: Col_Ind, State_Ind, Rank, Cost of Living Index, and Local Purchasing Power Index. At one time, there were about 30 columns in the dataset. Several were duplicates of other columns but had a reason for data cleaning or data processing. In the end, duplicate columns have been removed and the data frame has 25 columns.

Task Definition and Research Questions

The Research Questions

- What are the skills required for different jobs in data science?
- What is the difference between a Data Scientist, a Data Engineer, and a Data Analyst?
- How are these jobs different based on job requirements, skills, and expectations?
- Where are these jobs congregated?
- How do the salaries for these jobs compare with the cost of living in these locales?
- What should recent graduates in data science know about searching for jobs in data science?

The Problem to Solve

The real-world problem addressed in this project is finding the "perfect" data science job in the "ideal" city at the end of a degree program. The exploration includes skills required, terms used to describe the positions and responsibilities, and determining the best location for the new job considering cost of living and other lifestyle factors. In the end, this project will help prepare data science students for the job hunt with information about data science positions, an understanding of the expectations and requirements for those positions, and tools for success.

Analysis Input and Output

One part of the project uses NLP to compare job titles and job descriptions to find similarities and differences of responsibilities. The working hypothesis is there is not much difference between job descriptions for a variety of job titles in data science, and the goal is to uncover any differences that exist. Another piece of this project includes locating the "hubs" of data science jobs and determining the cost of living in those "hubs" in order to find the "perfect" job. This project will also explore other aspects of job hunting relevant to recent data science graduates.

Literature Review

Work Done on This Topic

There are many articles about job hunting for a first position as a data scientist, and each contains advice that could be useful for a new graduate, including visibility, internships, networking, conferences, posting projects on GitHub, creating a portfolio of work, increasing skill sets, obtaining certificates, and building relationships with recruiters. Some of these articles include both specific and general advice:
https://www.dataquest.io/blog/how-to-find-an-entry-level-job-in-data-science/
https://www.kdnuggets.com/2017/03/get-data-science-job-guide.html
https://towardsdatascience.com/3-strategies-to-guarantee-a-data-science-job-with-no-experience-68d85b345f21
https://www.springboard.com/library/data-science/how-to-become-without-experience/
https://medium.com/jovianml/how-to-land-your-first-data-science-or-machine-learning-job-ffdcd17c8b61

Another perspective to consider is that of the company hiring the new data science graduate. The job poster needs to be clear about what the job entails, the skills and tools necessary, and the work environment. Several sites provide tips for the job poster:
https://www.indeed.com/hire/job-description/data-scientist?aceid=&gclid=Cj0KCQiA-OeBBhDiARIsADyBcE7ev3QyxsrdfJB2leiOqoCw9RVWXfRnFPtZrv_Hx4BzMnW4qdWwkvsaArebEALw_wcB

There is some confusion about different job titles in the field of data science with roles often having overlapping responsibilities. It is helpful for both the job hunter and the job poster to be clear about these titles and roles. This article clarifies the differences:
https://towardsdatascience.com/10-different-data-science-job-titles-and-what-they-mean-d385fc3c58ae This article also explains different data science job titles relevant for 2021:
https://www.mygreatlearning.com/blog/different-data-science-jobs-roles-industry/

One goal of this project is to clear up the confusion about the various job titles and these overlapping responsibilities. The working hypothesis is many job descriptions do not adequately describe the responsibilities and day to day work of people in data science. While it is the responsibility of the job hunter to network and make connections to learn more about work environments and job responsibilities, it is also the responsibility of companies hiring data scientists to be clear about the positions they advertise/offer. This article outlines different data science titles and the corresponding salaries.
https://www.dataquest.io/blog/career-guide-data-science-options/ It also offers advice for beginning data scientists regarding freelancing and internships.

In addition to the confusion about job titles in data science and the overlapping responsibilities, some job descriptions include buzzwords and are not clear about other job specifics. This article addresses this issue, outlines the process for searching for a job in data science (including a test or a project), and highlights the characteristics of good companies to work for: https://www.kdnuggets.com/2019/04/recognize-good-data-scientist-job-from-bad.html Company ratings matter; working on a data science team is also important.

Consideration needs to be given to a diverse work force as well as the impact of women in the field of data science. While women may enter the field of data science, many do not stay. Research provides some reasons for this trend:
https://www.bcg.com/en-us/publications/2020/what-keeps-women-out-data-science
https://www.betterbuys.com/bi/women-in-data-science/

This recent graduate in data science did a similar project and came to conclusions about next steps based on her research. https://towardsdatascience.com/how-to-identify-the-most-requested-skills-on-the-data-science-job-market-with-data-science-726845ca9638 The work was focused in London and not the United States but definitely provides guidelines for continued skill development and job searching next steps.

Of course, the US Bureau of Labor Statistics publishes data about data science occupations. This site provides an overview and includes information from 2019 related to this project:
https://www.bls.gov/oes/current/oes152098.htm

Novel Work Compared to Previous Work

This project is a first step for students considering a job hunt after graduation (or even before graduation) and how to prepare for the process. The research is different than other work because consideration is given to cost of living as compared to salary information. This project also considers the impact of remote work due to COVID and the desire for opportunities to work from home. Remote work is a strong possibility which makes cost of living compared to job location less important because the new data scientist may be able to live anywhere. Junior positions and internships are strong considerations as a starting point. Similarly, part time work or temporary work are considerations to gain skills/experiences and add to portfolios of work. The industry of the data science job is also important to this research.

Future work will include research into women in the realm of data science and what needs to be done to make the workplace inclusive and desirable. This work might include the perspective of males in the field data science and the perspective of companies that want to attract more women to the world of data science. Women definitely add to the diversity of the work force and the strength of data science teams.

Data Cleaning

Cleaning Process and Conversion Activity

Each of the eight datasets went through a somewhat similar process of data cleaning.

1. Column headings were made consistent as much as possible based on data available. Some column headings were dropped because they were not relevant or necessary to this project. These included column names like: Headquarters, Sector, Revenue, Competitors, Type of Ownership, Number of Applicants, Length of Job Description. Most of the datasets include the following column headings when the data was available: Job Title, Job Description, Company Name, Location, City, State, Remote, Size, Founded, Industry, Rating, SalaryMin, SalaryMax, and Salary Estimate. One dataset included Jr/Sr which was included due to usefulness for this project. Some column headings had similar meanings but headings were changed for consistency between all datasets. The columns were also re-ordered for consistency between datasets.
2. Unnecessary white space was removed.
3. Some original datasets had the rating as part of the company name. The rating was removed and separated into a separate column.
4. One dataset had a column for "years founded", and this was changed to align with the other datasets. The column now reflects the year the company was founded. Subtraction was used to accomplish this task. The data type was checked and now all founded entries are floating values.
5. Bins were created for company sizes, and now each company has a rating for company size, including: very small, small, medium-small, medium, medium-large, large, and very large.
6. Some datasets had industry names separated with underscores. These underscores were removed for consistency and to prepare for later uses.
7. While some datasets included a column for location while others included columns for city and state. To improve consistency and prepare for later use, locations were split into city and state, and datasets with city and state had a column for location added.
8. The location issue above also led to the creation of a "remote" column as some job listings were clearly remote with no specific location necessary.
9. While working on the location columns, data was examined to ensure all jobs were located in the United States. Even if a company has a headquarters in another country, effort was made to ensure there were positions inside the US. Only US jobs are included.
10. Salary information differed between datasets. Sometimes salary information provided was a range and other times as an estimate. The ranges were split apart to create columns for SalaryMin and SalaryMax. A column for Salary Estimate is provided when information is available.

11. Some of the salary information included a dollar sign ($) in front of the amount and a (K) attached at the end to signify thousands of dollars. These symbols were removed so that float values could be used for statistical purposes.
12. While checking for duplicates in the combined dataset, it was discovered that several (51) job postings include an hourly wage instead of actual salary information. These jobs are indicated in a column labeled “Hourly” in order to note the difference.
13. Some datasets had a (-1) or the word “unknown” used to signify missing values. These missing values were all changed to NaN.
14. Duplicates were removed from each individual dataset. Then effort was made to remove duplicates across the full datasets.

After the eight datasets were combined to create the full data set, other aspects of the data had to be cleaned in order to employ NLP (Natural Language Processing).

1. Job descriptions were searched for words like “remote”, “part time”, “temporary”, “junior”, or “senior” so data could be included in columns with those headings to represent other types of jobs that might be of interest to someone new to data science.
2. Several job postings were dropped because the job descriptions were duplicated, even though salary information was different in order for the focus to be on words used in job descriptions.
3. Contractions and stop words were removed from job descriptions.
4. Excess punctuation and e-mail addresses were also removed from job descriptions
5. Remaining words were lemmatized. Attempts were made to use tf_idf with limited success.
6. Initial word clouds were created to check process. At that point in the process words common to all job hunting were dominating the lists of words from the job descriptions. Therefore, over 250 words were removed so the focus could be on terms specific to data science. Some terms common to data science were also specifically addressed. For example, “artificial intelligence” was coded to be “ai”, and “big data” was merged into one word: “bigdata”.
7. Job titles varied greatly. In addition to removing blank spaces and searching for key words, the job titles were grouped into similar categories based on job titles such as: Data Scientist, Data Analyst, Data Engineer, Business Analyst, and Analyst. After grouping there were about 30 job titles. Job titles involving “sales” or “environmental scientist” were ultimately dropped. Other titles dropped included “Workforce Analyst”, “Medical Scientist”, and “Clinical Scientist”. The final dataset includes the following 28 job titles: 'Data Engineer', 'Data Scientist', 'Business Analyst', 'Analyst', 'AI', 'Data Analyst', 'Intern', 'Consultant', 'Business Intelligence', 'Cyber Security' ,'Researcher', 'Scientist', 'Data Modeler' ,'Data Architect', 'Research Scientist', 'Machine Learning', 'Developer', 'Applied Scientist', 'Big Data', 'Computer Scientist', 'Engineer' ,'Product Manager', 'Systems Engineer', 'Research Engineer',

'Software Engineer', 'Applications Engineer', 'Cloud Engineer', and 'Full Stack Engineer'.

In order to compare the salary data from the job listings with cost of living data, data about cost of living had to be merged with the job description data.

1. The cost of living data started with 75 cities from North America, and the Canadian cities were removed, resulting in data for 57 US cities.
2. The following columns from the cost of living data were added to the dataset: “Rank”, “Cost of Living Index”, “Local Purchasing Power Index.” Additionally, two columns were created to help match this data to the dataset. “State_Ind” helped match by state, and “Col_Ind” served as a temporary heading to store city information.
3. After combining the data frames, many of the cities in the job data frame matched with cities in the cost of living data. However, there were 9133 job postings that did not match. The cities from those postings were grouped by metropolitan area and matched with the cities from the cost of living data to create “cityKeys” or groups of suburbs connected to the major metropolitan areas. In this process, the city of the original data frame was preserved and the newer column called “Col_Ind” became the label for the major metropolitan area nearby, if one exists. The column called “State_Ind” helped with this, especially when major metropolitan areas cross state borders.
4. To create some of the visualizations included in this project additional data cleaning had to be completed. For example, to create the graphs comparing salaries and industries, punctuation and stop words had to be removed from the industry data.

Unusual Discoveries

- When checking for duplicates, one discovery was that SalaryMax had hourly wages and not salaries.
- Several job postings only listed a state in the location column which made it difficult to split into city and state. So missing values were assigned to the city column. Postal abbreviations were not used either; so postal codes had to be added.
- Another unusual discovery had to do with the number of duplicates within each dataset. Similarly, there were duplicates between the datasets that needed to be removed.
- The dataset contains 2391 postings for remote work, 393 postings for part time work, 1911 postings for temporary work, 7239 senior positions, and 1338 junior positions.
- The data was really “dirty.” Throughout the project, adjustments had to be made and more cleaning steps had to be implemented. Dirty data provided challenges for each part of this project and consumed about 100 hours of time to address different types of “dirty” data.

Missing Values

- Some datasets labeled missing values with a (-1) or the word "unknown". To be consistent, these were changed to NaN.
- Each dataset had different missing information. The decision was made to keep all of the data because of the information each dataset can provide. All datasets have job titles and job descriptions as this information is critical to the project. Further, some of the datasets have other information vital to the project even if these datasets are missing other important information.
- Time was spent locating companies that were listed only as "United States" or had missing values for city or state. Now if the information in the "City" column is missing, the job posting is a remote job. This precipitated the creation of the remote column.
- Several job postings were dropped because there was duplication within the original dataset. After the full data set was created, time was spent searching for duplication between datasets and within the combined dataset. In the end, there were 18 additional job postings dropped due to duplication.

New Feature/Attribute Creation

- Each of the eight datasets had different features created or omitted. Combined together, the new columns created included "Remote" and "Jr/Sr." Later columns were created for temporary work, part time work, remote work, junior positions, and senior positions.
- As explained above, some datasets had columns created for "City" and/or "State" based on the information provided in a column called "Location." After merging the data with data about cost of living, two new attributes dealing with location were created to help with the merge: "Col_Ind" and "State_Ind."
- When combining all of the eight datasets, a column was created to label the original dataset so that searching might be easier if necessary. This column was also helpful in identifying duplicates that needed to be removed.
- Another column/attribute was added when examining the data for duplicates. Job postings that have an hourly wage instead of a salary were discovered and a column was created to flag that data because the entries are based on dollars per hour instead of thousands of dollars per year.
- After applying NLP, new columns for the cleaned/modified job titles and job descriptions were created.

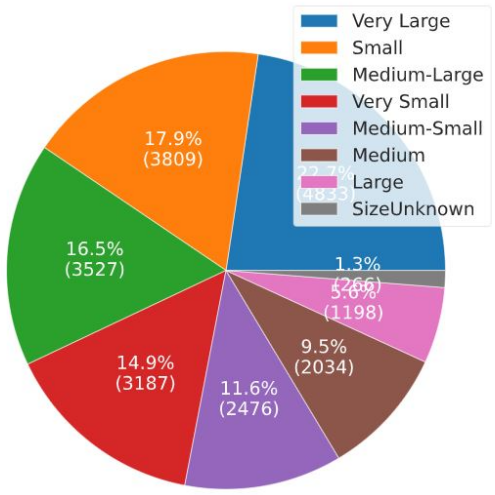
Data Summary Statistics and Interpretation

Pandas Profiler provides data summary statistics on the entire combined dataset. Salary means were calculated. Company ratings were compared. Data visualizations were created to illustrate these statistics.

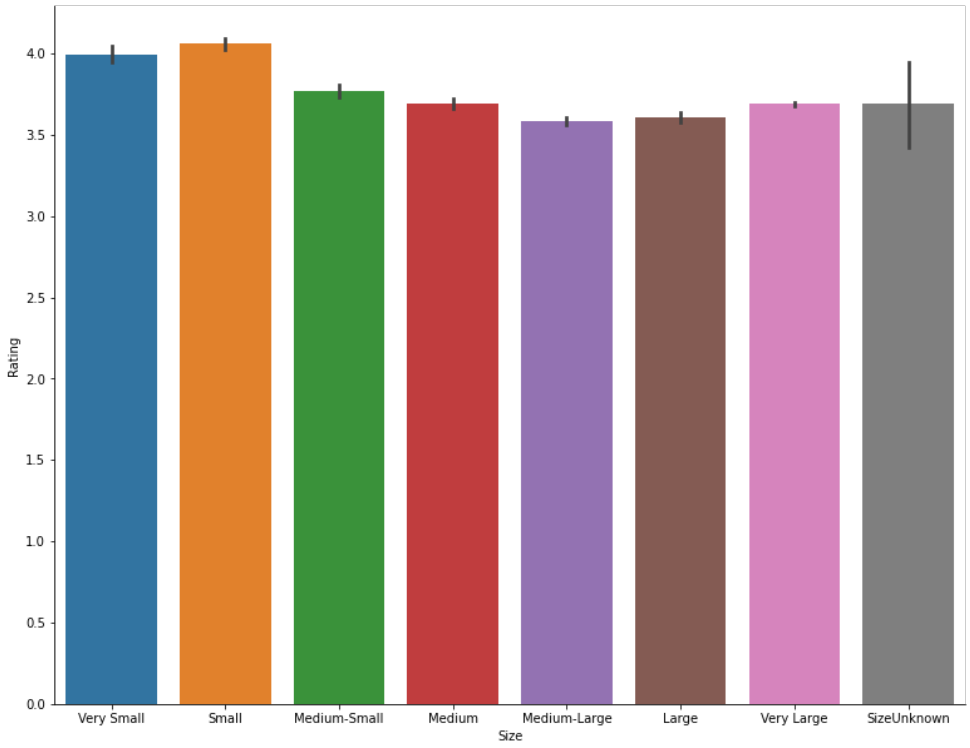
- The data set started with 8152 unique job titles which were combined into 28 job titles to examine in this project.
- Frequency of cities is highlighted in the Pandas Profiling Report with a bar graph which was used to help with comparing the job description data with the cost of living information.
- All but two US states have data science jobs in this data set. Alaska and Montana do not have data science jobs included in this data set.
- Many job descriptions include company ratings from current/previous employees. The minimum rating is 1, and the maximum rating is 5. The mean rating for all companies in the data set with ratings is 3.789.
- Salaries range from $10,000 to $200,000 with a mean minimum salary of $74,632 and a mean salary maximum of $120,483.
- The cost of living maximum is set at 100 and is based upon the cost of living in New York City. The cost of living minimum in this dataset is 52.38 in El Paso, TX. The mean cost of living index in this dataset is 77.55.
- There is a correlation between company size and junior/senior positions. The interpretation of this is that larger companies have teams of data scientists with a range of experiences. This creates an opportunity for both junior and senior roles.
- Similarly, the age of the company correlates with junior/senior positions. This probably occurs because the company has had time to develop positions and responsibilities for data scientists with different amounts of experience.
- The dataset includes 2391 remote positions, 393 part time jobs, 1911 temporary jobs, 7239 roles marked as "senior", and 1338 roles marked as "junior." The number of remote positions could be a result of COVID protocols and could potentially increase.
- Early in the project, the decision was made to eliminate "Sector" from the data bases and instead choose "Industry." In retrospect, this decision may have been a mistake. Graphs comparing salaries and industry are included in this project but are a bit cumbersome. If "Sector" had been chosen, the graphs may have been more useful.
- Cost of Living Index is a relative indicator of consumer goods prices, including groceries, restaurants, transportation and utilities. Cost of Living Index does not include accommodation expenses such as rent or mortgage. If a city has a Cost of Living Index of 120, it means Numbeo has estimated it is 20% more expensive than New York (excluding rent).
- Local Purchasing Power shows relative purchasing power in buying goods and services in a given city for the average net salary in that city. If domestic purchasing power is 40, this means that the inhabitants of that city with an average salary can afford to buy on an average 60% less goods and services than New York City residents with an average salary.

Visualizations

Size of Companies included in Dataset by Category

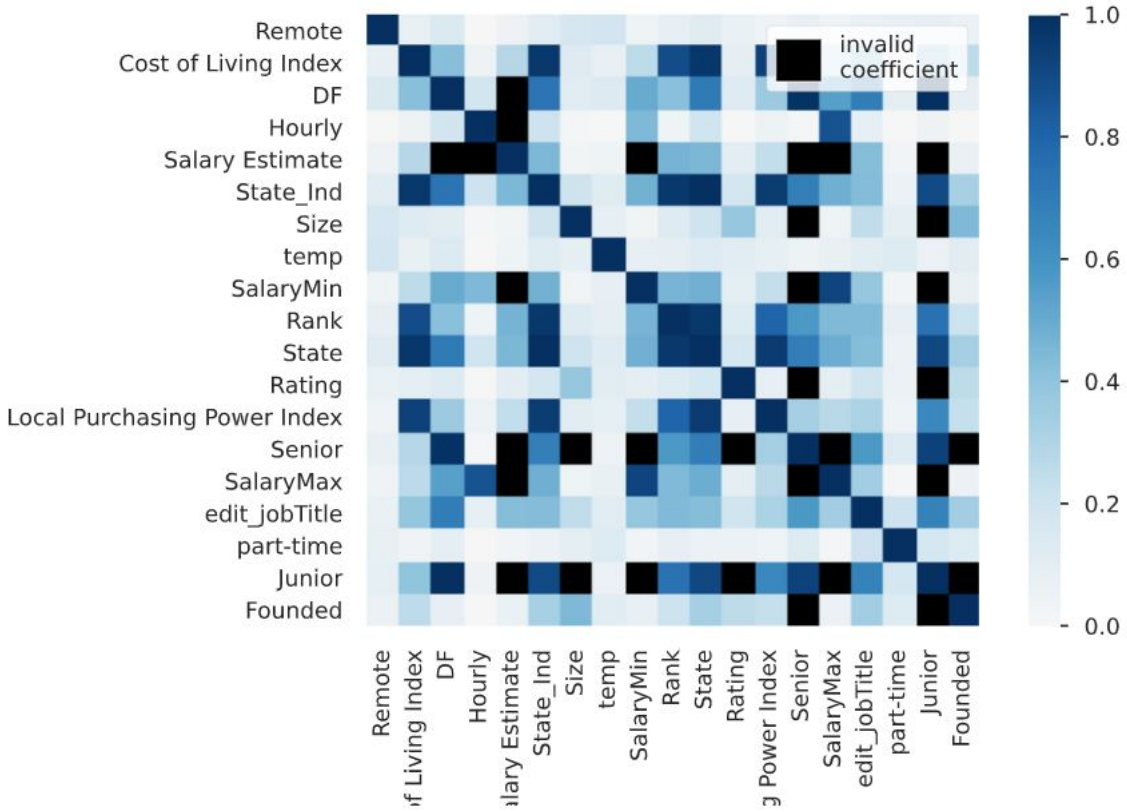


Size of Company Compared to Company Rating



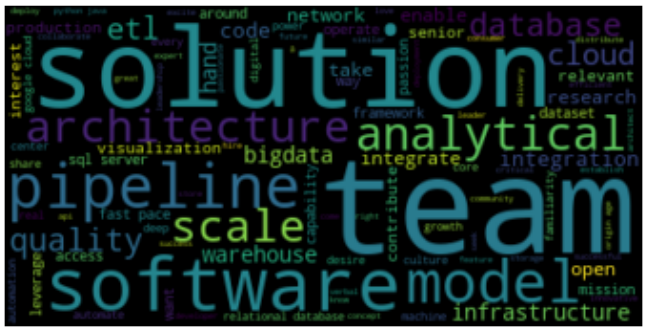
One observation from the graph above is that smaller companies receive higher ratings from employees and former employees. A resulting interpretation might be smaller companies can be more personal to employees and address employee needs.

Heat Map to Show Correlation of Some Attributes

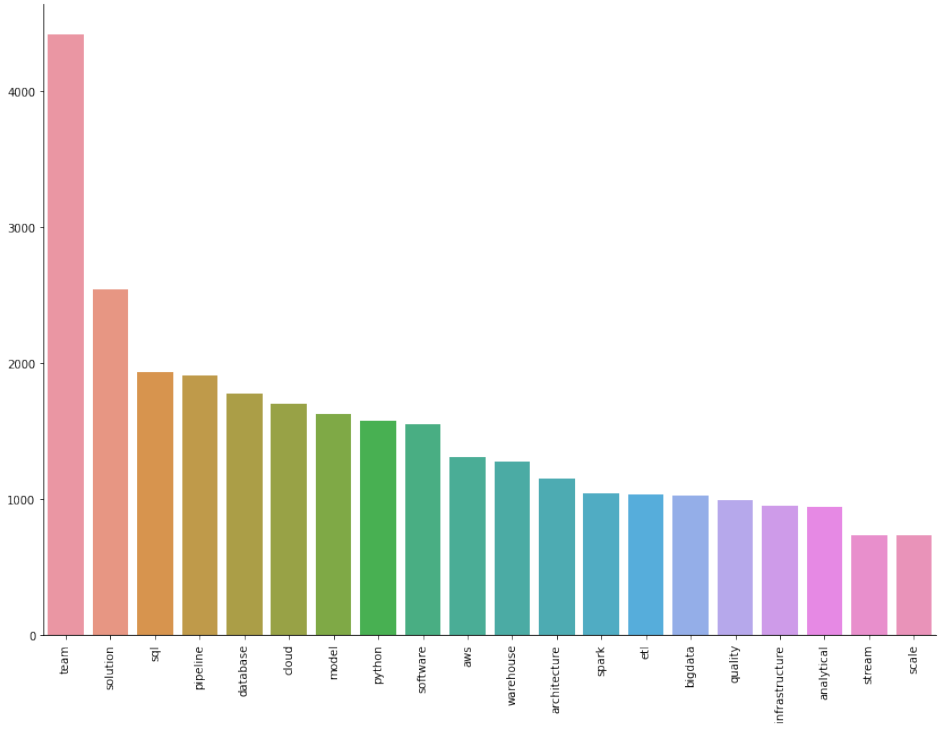


This heat map provided some insight to help make correlations in the data set. One conclusion noted was the relationship between company size and the junior/senior positions. This was interpreted that larger companies can create positions for a variety of data scientists with different levels of experience. Further, another correlation was noted. Older companies also have the ability to have junior and senior positions probably due to the length of time of existence and the resulting need for a variety of data science experiences. A few other correlations were obvious and expected based upon cost of living with locations and junior/senior positions with salaries. (The labels on the horizontal axis were truncated in the copying of the graph. They are clearly labeled on the left axis.)

Word Cloud for Data Engineer:

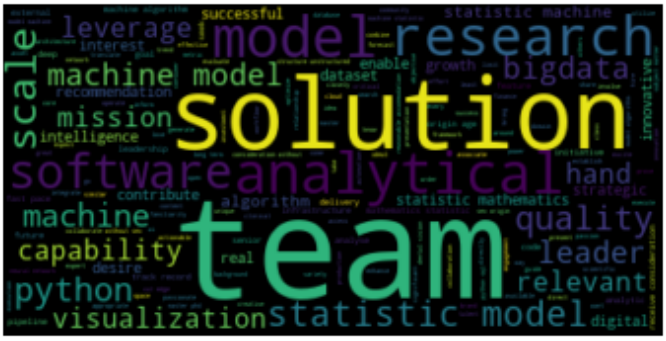


Frequency of Top 20 Words in Data Engineer Job Description:

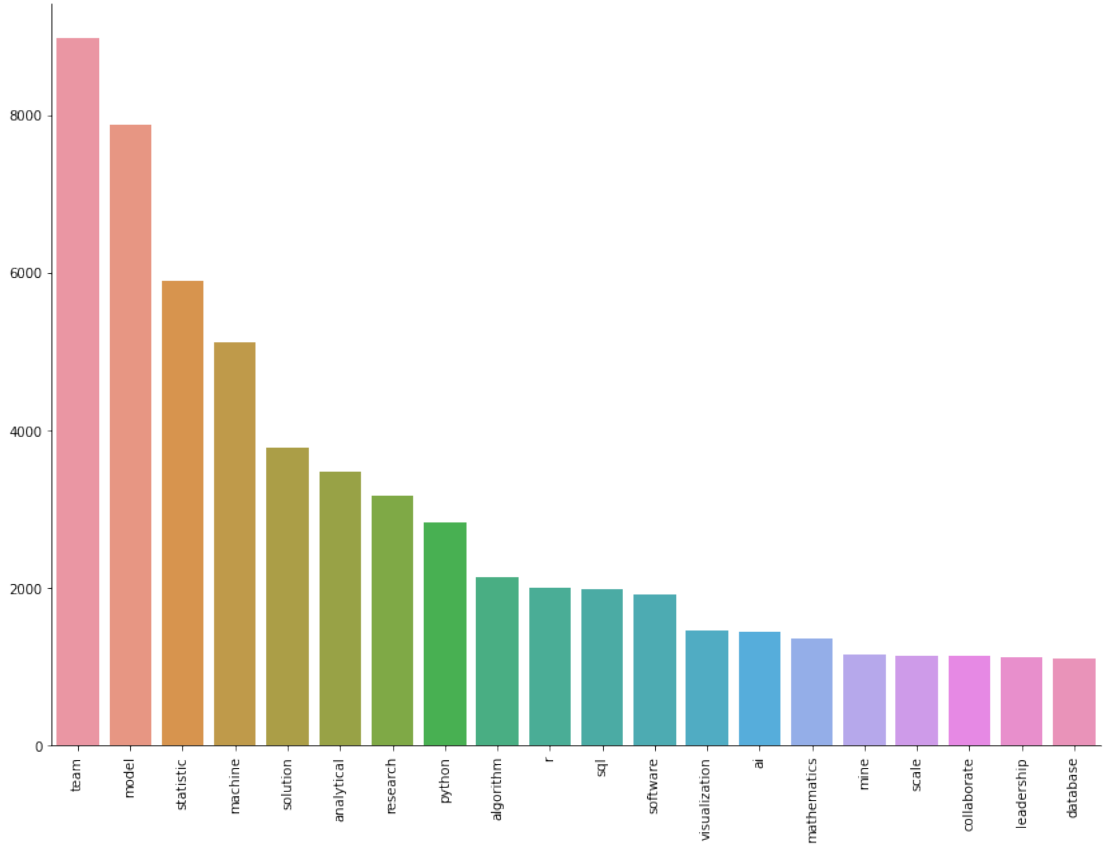


From the word cloud and the bar graph above, the words team and solution appear most frequently in the data engineer job descriptions implying that data engineers work on teams to arrive at solutions. Other terms appearing frequently illustrate tools used by data engineers (sql, python, software, cloud, aws, spark) and skills necessary for success.

Word Cloud for Data Scientist:



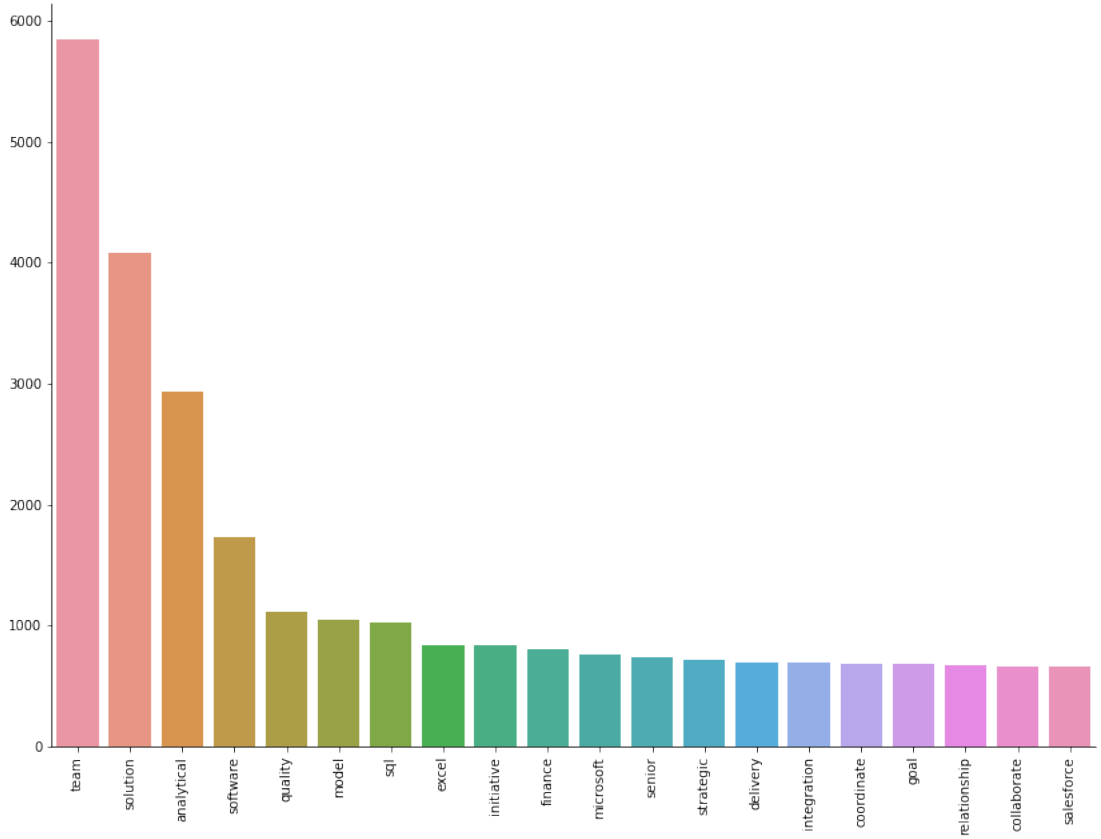
Frequency of Top 20 Words in Data Scientist Job Description:



From the word cloud and the bar graph above, the words team, model, and statistic appear most frequently in the data scientist job descriptions implying that data scientists work on teams to model statistics. Other terms appearing frequently illustrate tools used by data scientists (python, r, sql, software, ai, algorithm, mathematics) and skills necessary for success (collaborate, leadership, analytical, research).

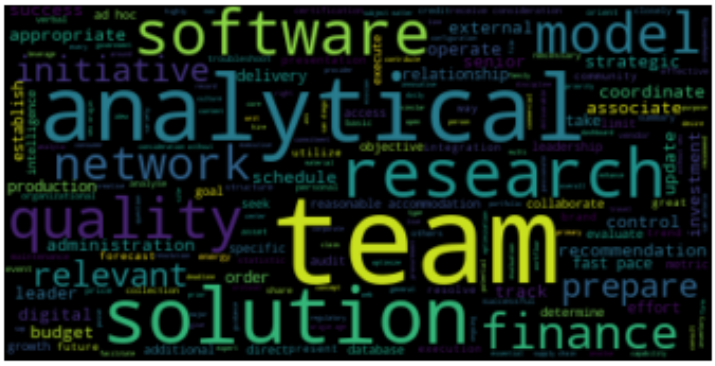
Word Cloud for Business Analyst:

Frequency of Top 20 Words in Business Analyst Job Description:

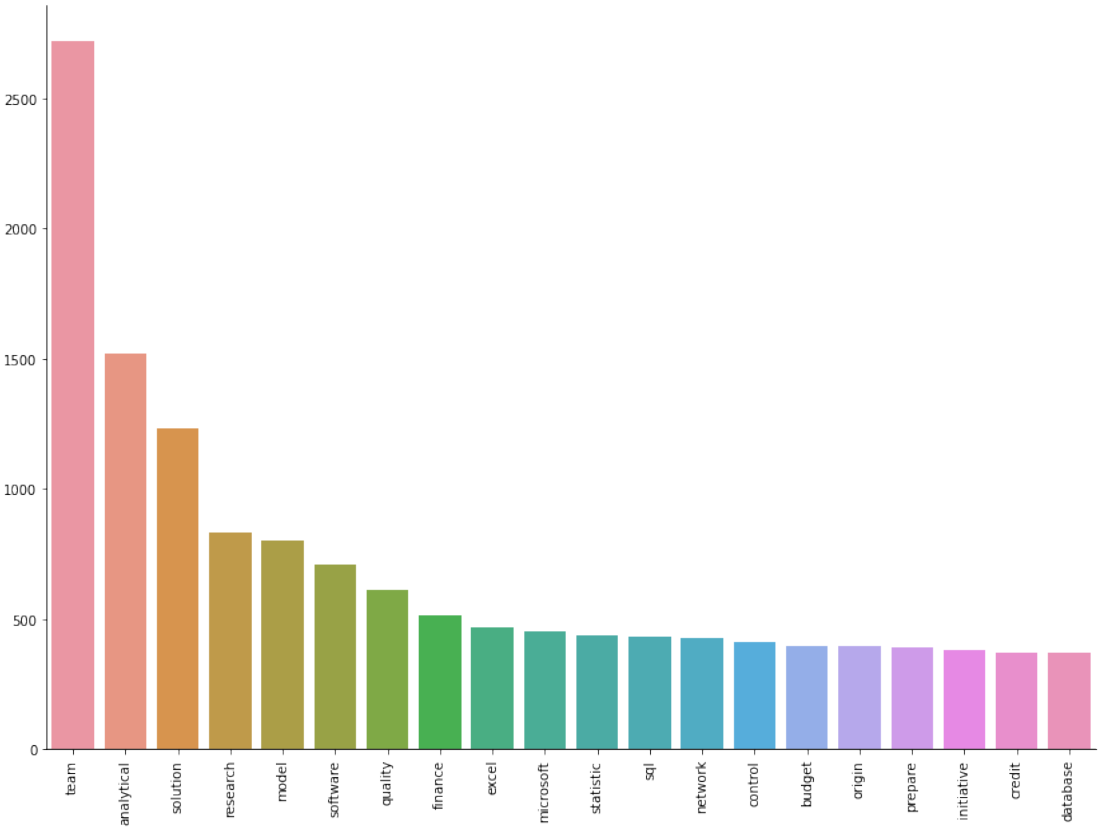


While many of the same terms appear in the Business Analyst job descriptions (team, solution, analytical, software) as the data engineer and data scientist job descriptions, there are new terms related to business included (excel, finance, Microsoft) and new skills required (salesforce, integration, strategic, delivery).

Word Cloud for Analyst:



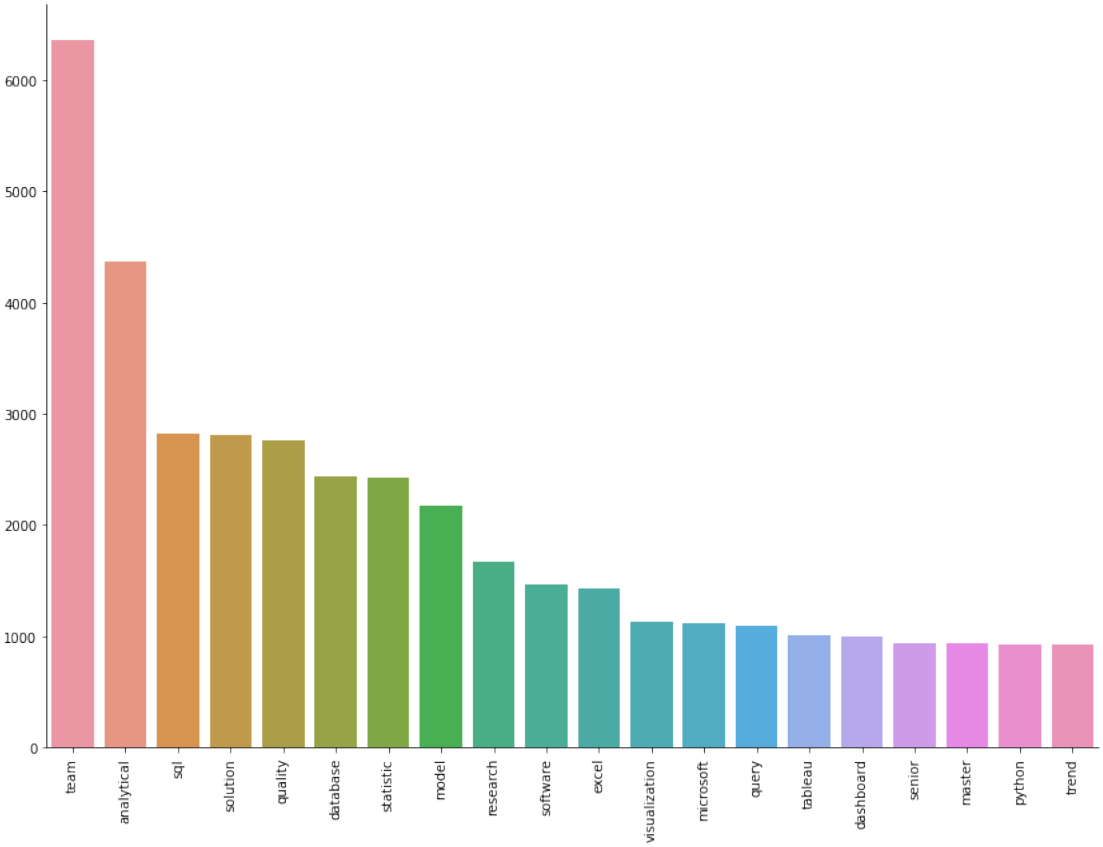
Frequency of Top 20 Words in Analyst Job Description:



The words used in an Analyst job description seem to be similar to the three categories of job descriptions outlined previously. Like data engineers and data scientists, analysts work on teams and find analytical solutions using models. Like business analysts, analysts employ excel and Microsoft on topics regarding finance and budgets. This category appears to be a "catch all" for the other job descriptions. However new terms also rose to the top in frequency, including: credit, initiative, control, and origin.

Word Cloud for Data Analyst:

Frequency of Top 20 Words in Analyst Job Description:



Many of the same words seen in the previous four job descriptions also appear in the Data Analyst job description (team, analytical, solution). Similarly, many of the tools used by Data Analysts have appeared in previous job descriptions (sql, database, statistic, visualization, python). There are a few new words that appear in this group of top 20: query, trend, tableau, master.

Observations about These Job Descriptions:

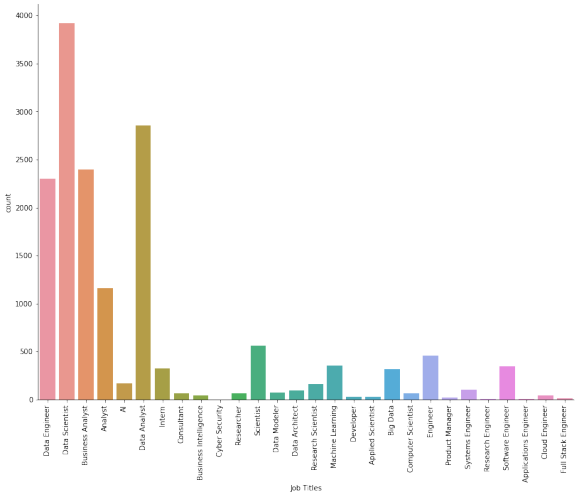
All of these job descriptions have much in common. Clearly people in these roles work on teams finding analytical solutions using a variety of tools and employing many necessary skills. The hypothesis for this project was that most of these job descriptions would be similar and there would be tremendous overlap of terms in the five categories of job descriptions.

While the hypothesis appears to hold, clearly there are a few differences which someone looking for a first job in the field might want to know. As the name implies, Business Analysts work specifically in a world of business with a focus on sales, budgets, and finance. Analysts seem to also have a slight business focus. It appears that different jobs may use different tools. For example, all use python and sql, but tableau appeared only in the job description for a data analyst.

Through this process, awareness about new tools developed as they were unfamiliar to someone halfway through a data science program. Terms like “etl”, “spark”, and “warehouse” were researched to ensure inclusion in these lists.

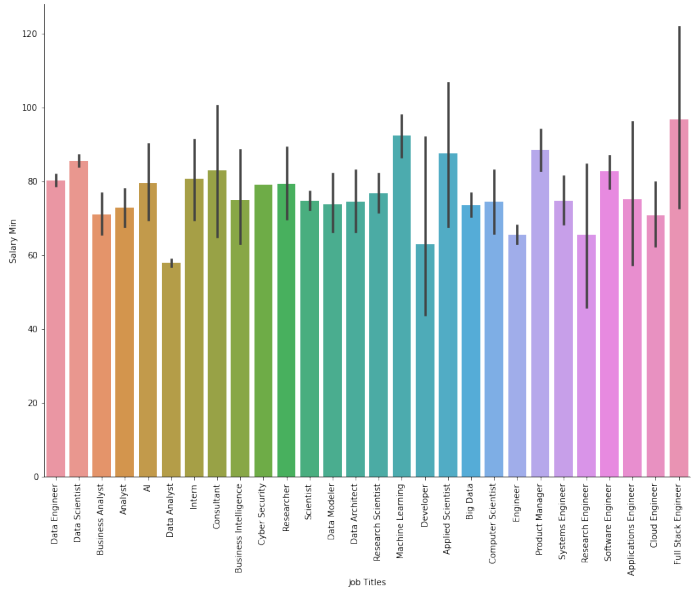
All the relevant data was included, and no real outliers were detected. As stated previously, terms related to data science were included and words related to general job searching were excluded. Some data that could be considered outliers includes some data for hourly wages for jobs such as data science tutors. This data is included and noted in the column labeled “Hourly.” As mentioned previously, job postings with job titles including “Medical Scientist,” “Sales,” “Environmental Scientist,” “Clinical Scientist,” and “Workforce Analyst” were dropped because the jobs were not considered to be in the realm of data science. In effect they could be considered outliers as these jobs are related but really on the borders of data science.

Frequency of Job Descriptions in Data Used for This Project:



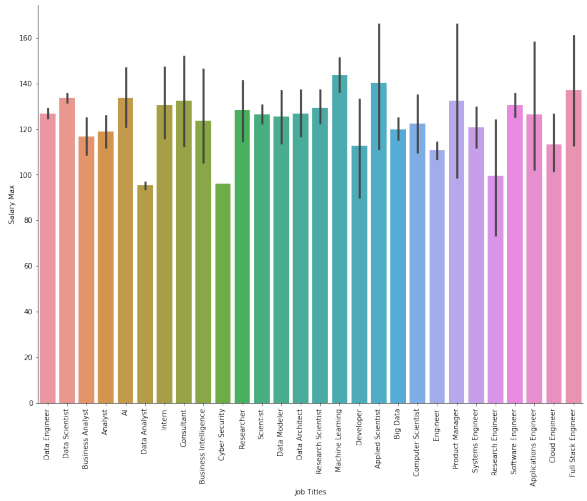
Salary Visualizations

Job Title and Salary Minimums



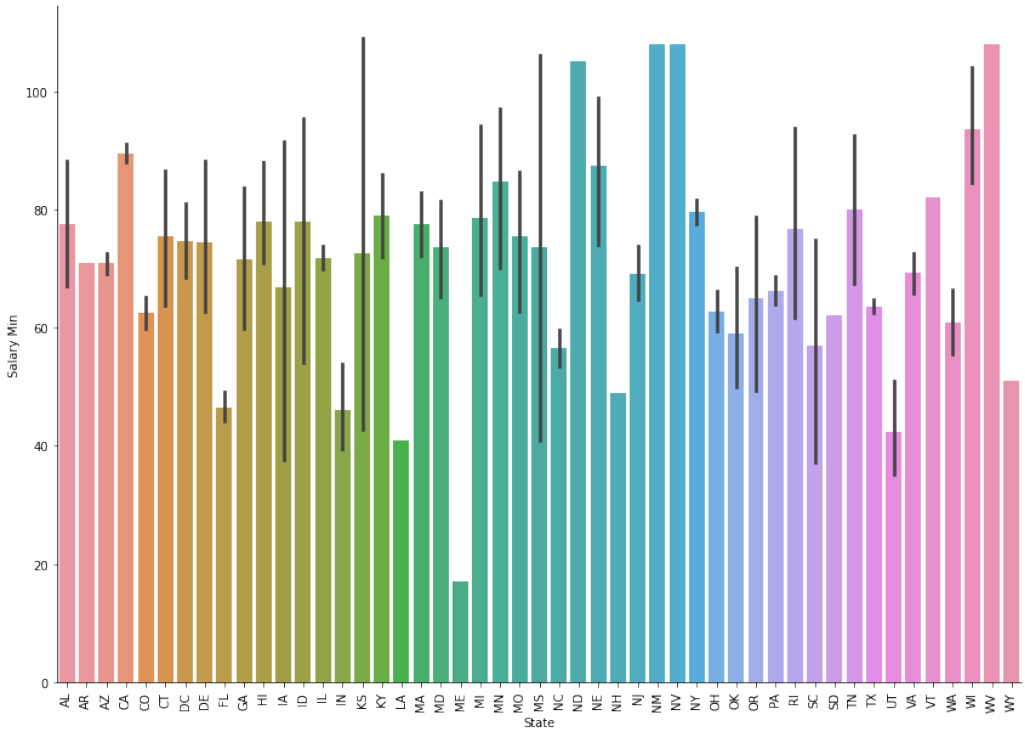
Full Stack Engineers have the potential for the highest minimum salaries while Data Analysts have the potential for the lowest minimum salaries.

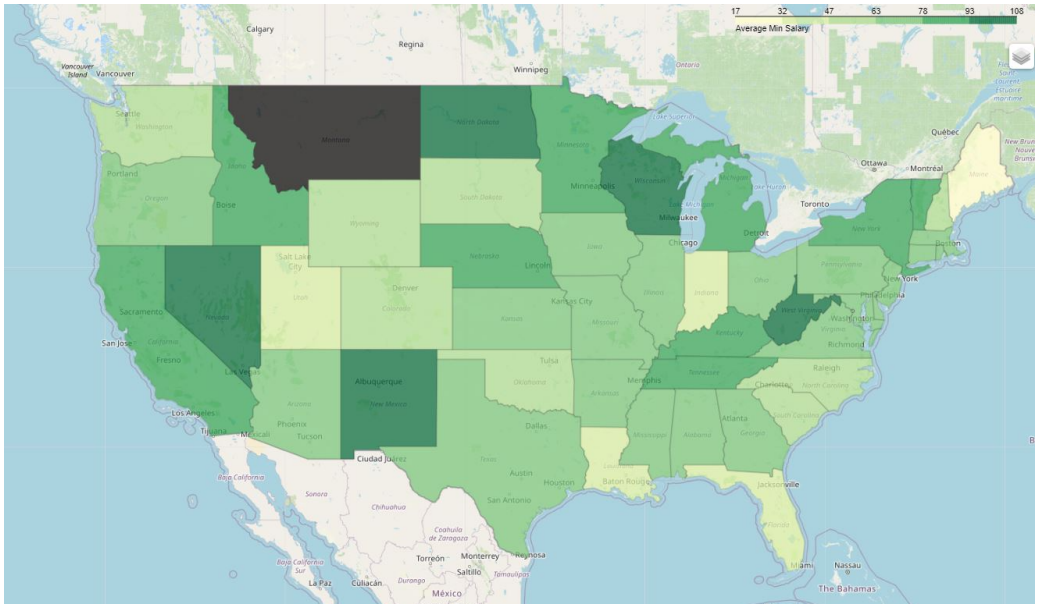
Job Title and Salary Maximums



More positions have the potential for higher salary maximums, including applied scientists, product managers, application engineers, full stack engineers, and machine learning positions. Surprisingly, data analysts, research engineer, and cyber security positions have lower maximum salaries. Clearly title has some effect on salary.

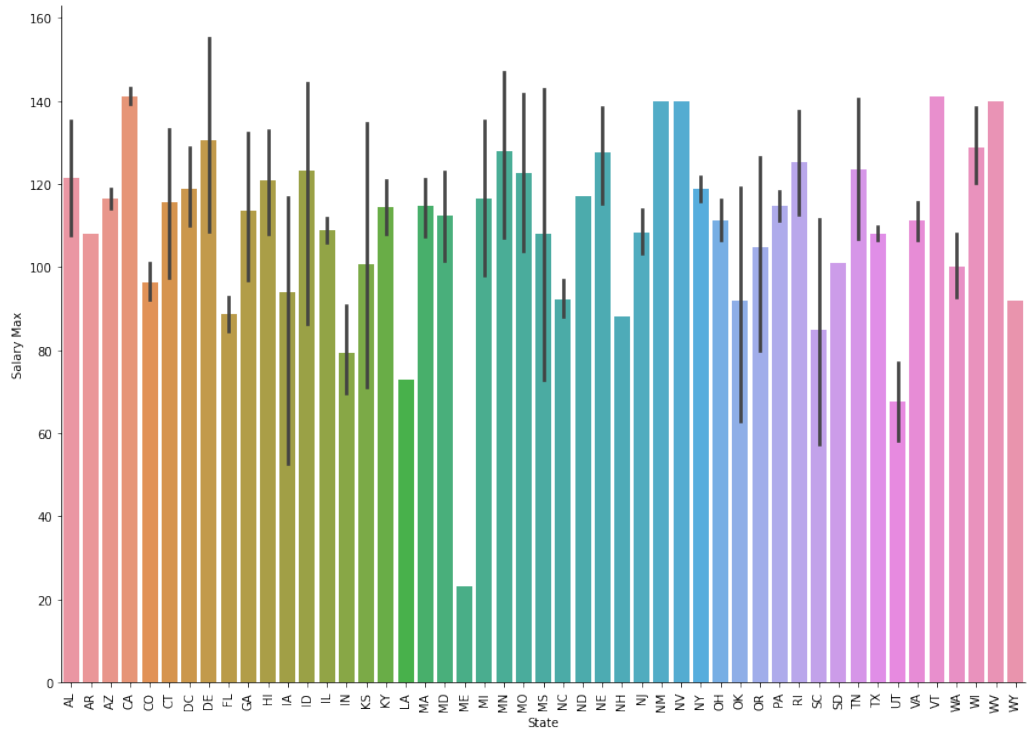
Salary Minimums by State

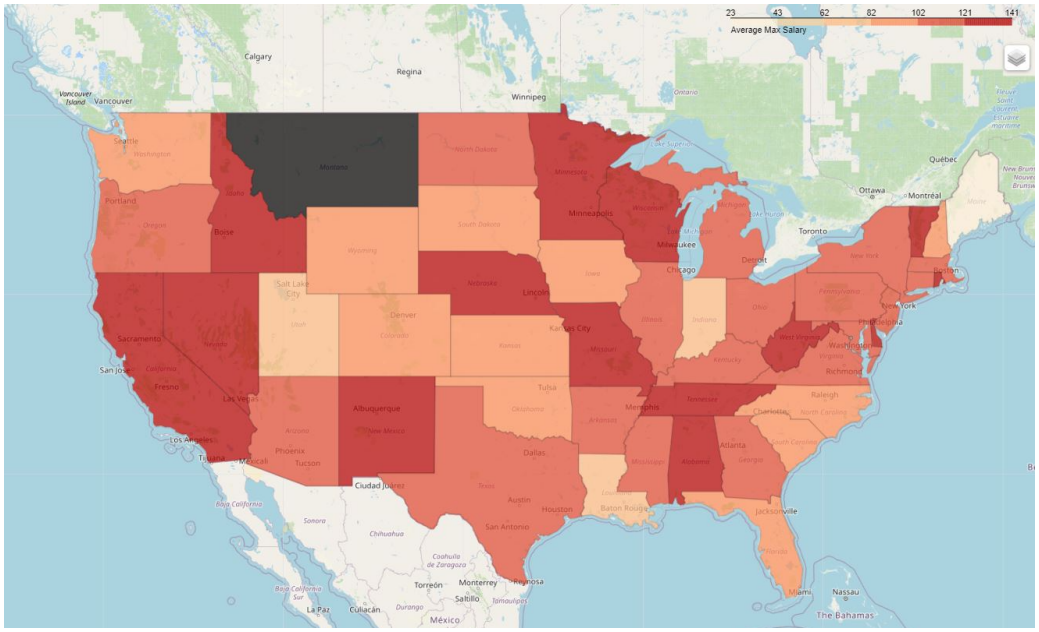




Some states that have surprisingly higher minimum salaries include: New Mexico, Nevada, North Dakota, Wisconsin, and West Virginia. Maine has a very low minimum salary. Other states with relatively low minimum salaries are Florida, Utah, Louisiana, and Indiana.

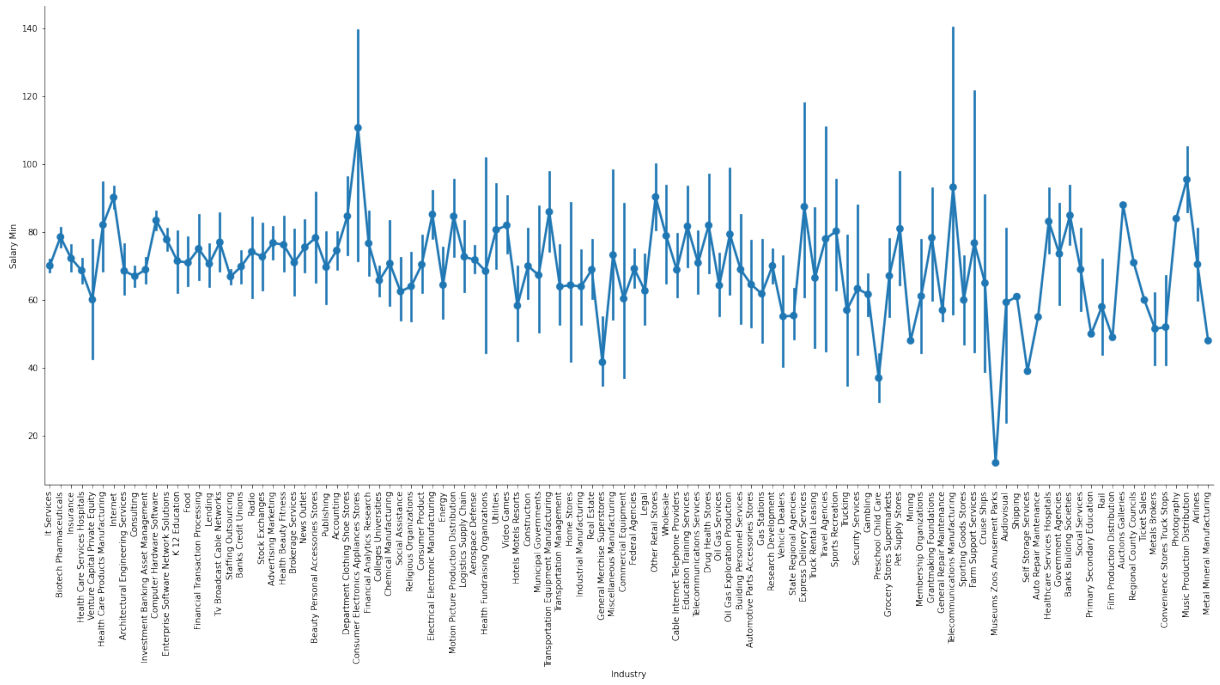
Salary Maximums by State



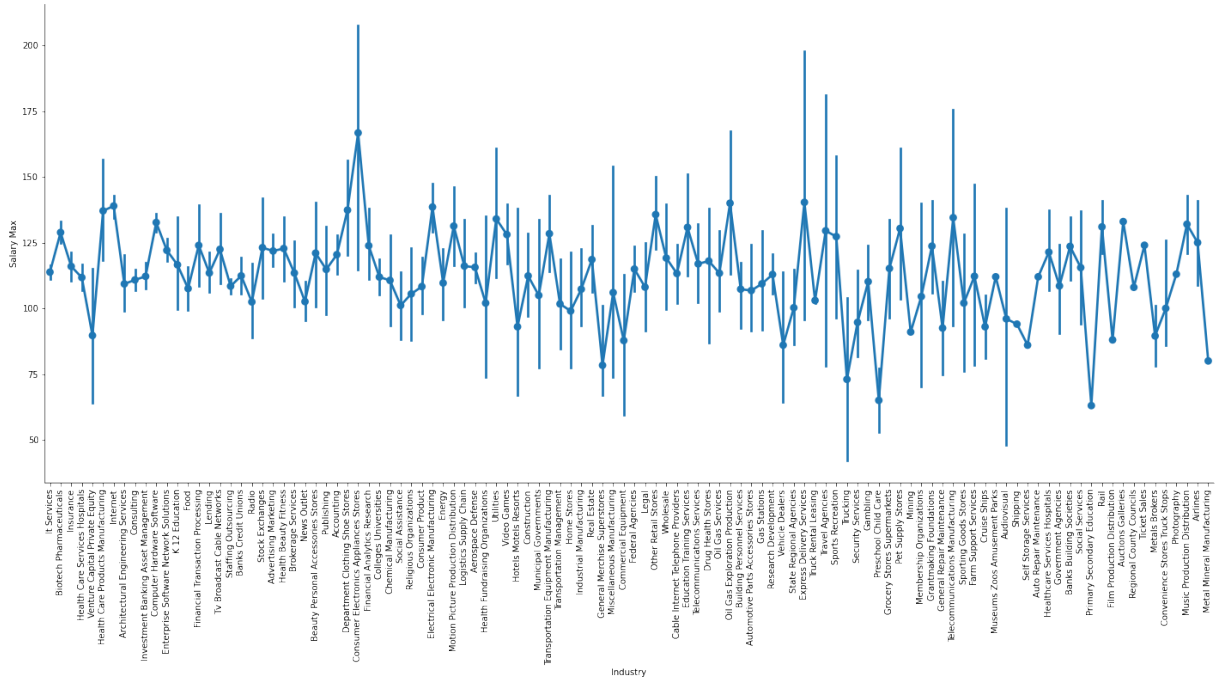


States with higher maximum salaries include: California, Nevada, Idaho, New Mexico, Nebraska, Minnesota, Wisconsin, Missouri, Tennessee, West Virginia, Vermont, Alabama, and Rhode Island. Maine has a really low maximum salary. Utah, Indiana, and Louisiana have relatively low maximum salaries.

Salary Minimums by Industry



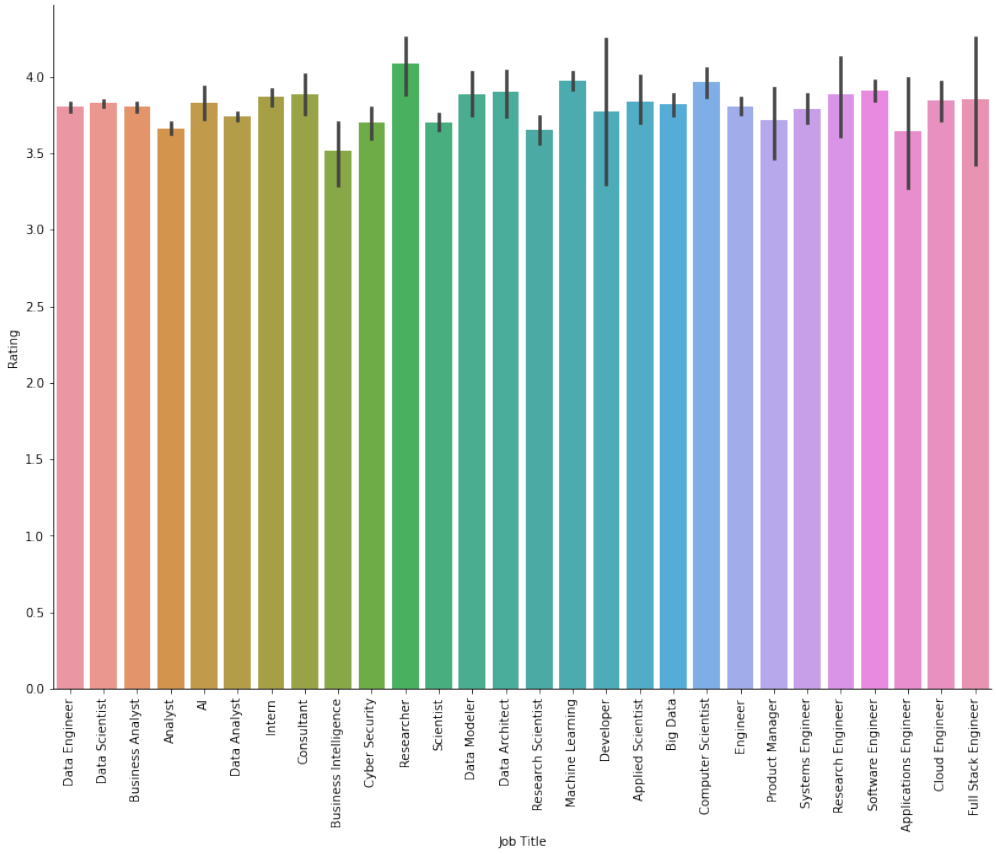
Salary Maximums by Industry



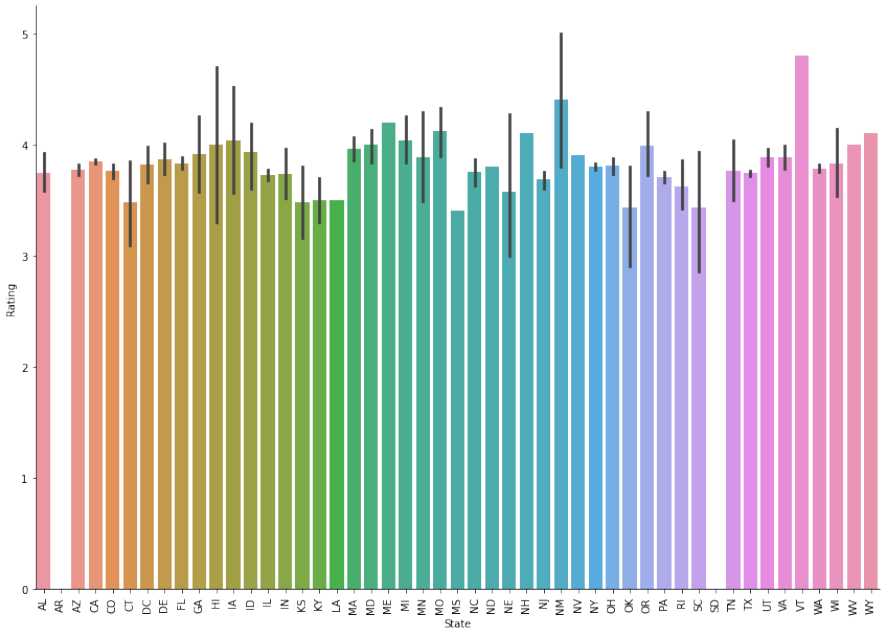
One consideration for data science graduates when searching for a first job is the industry in which that first job focuses. This is significant because an employee often stays in that industry for an entire career because of expertise in that industry. So the graphs above compare industry to salary.

Company Rating Visualizations

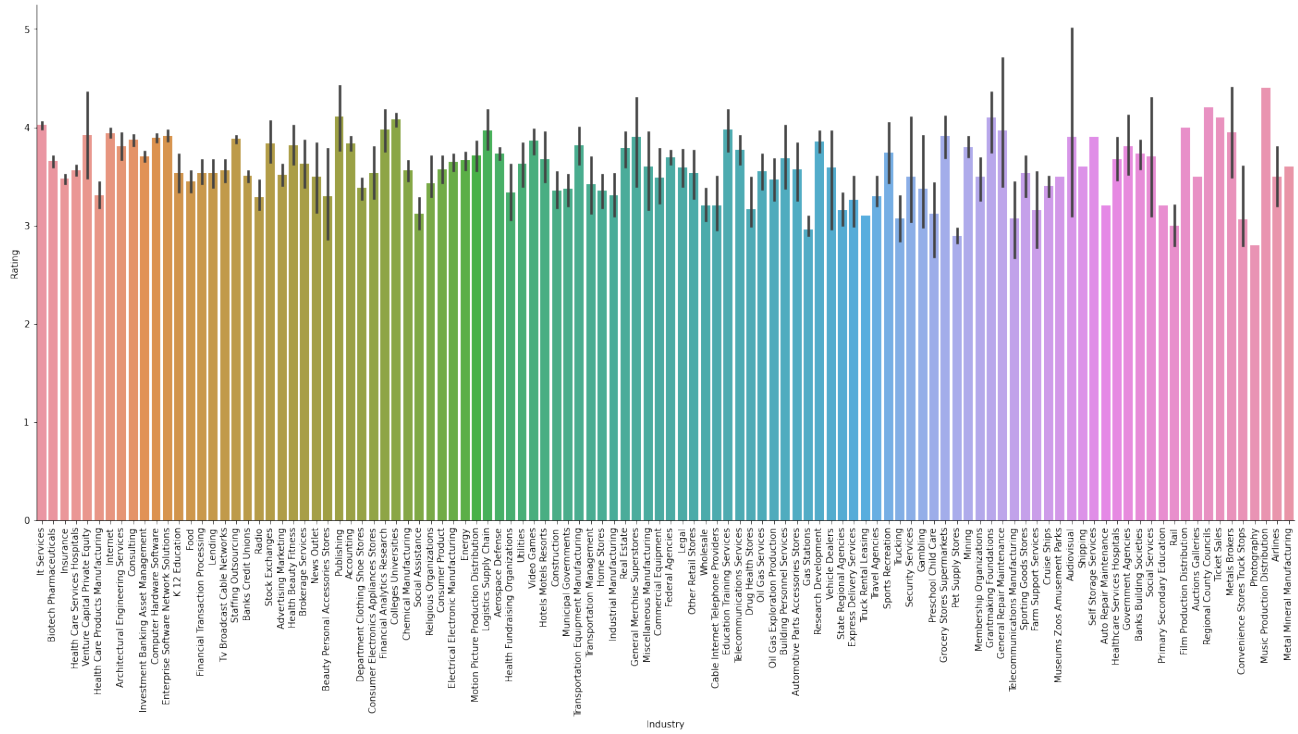
Company Ratings Compared to Job Titles

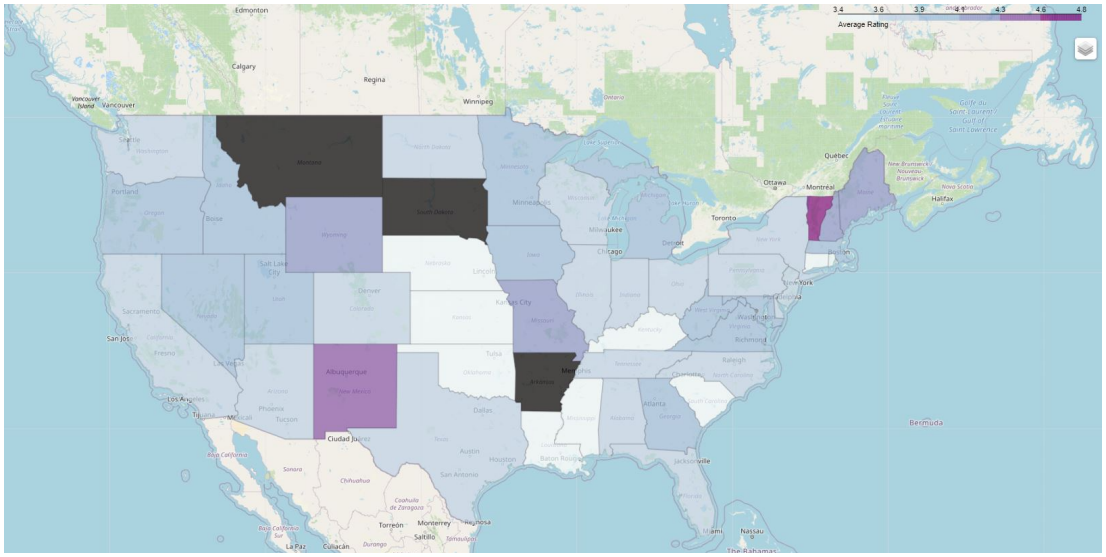


Company Ratings by State



Company Ratings by Industry



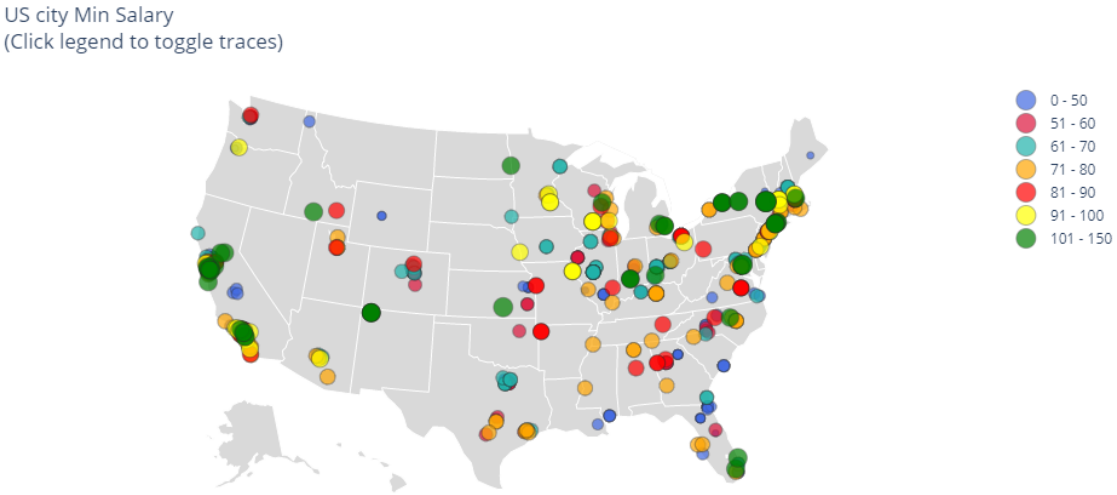


Using employee ratings of the companies employing data scientists, the choropleth map above illustrates the ratings by state. New Mexico and Vermont have companies with the best ratings. No data is included from Montana, Alaska, South Dakota, and Arkansas. Company ratings can be a significant factor to a new employee and someone new to data science. This might also be a significant factor to women in data science.

City Visualizations

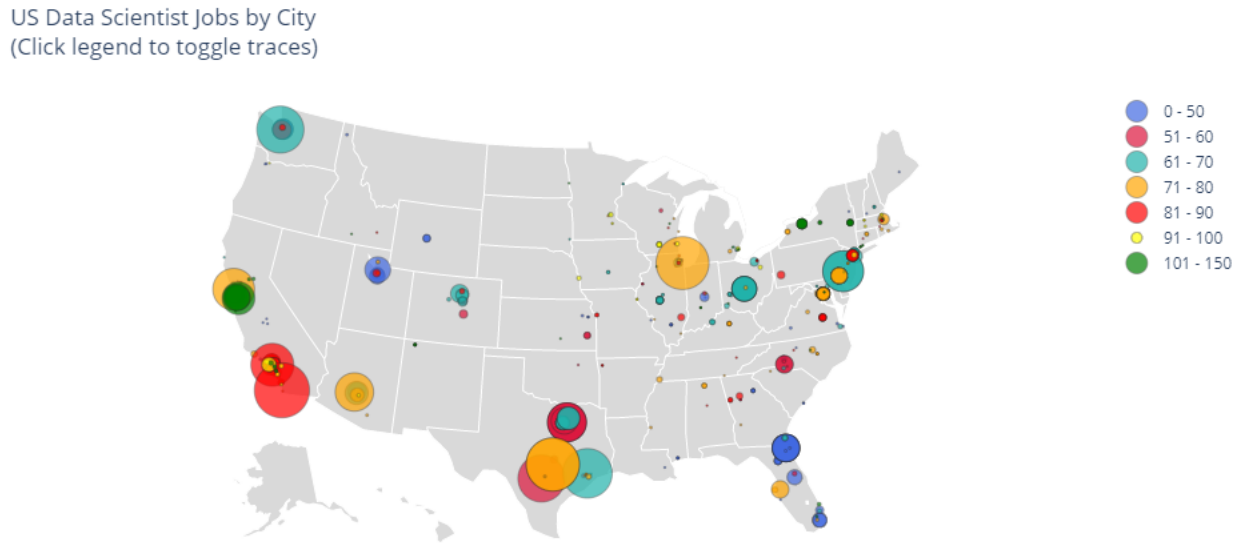
Salary Minimums by City

Min Salary by City (Color represents minimum salaries times 1000.)



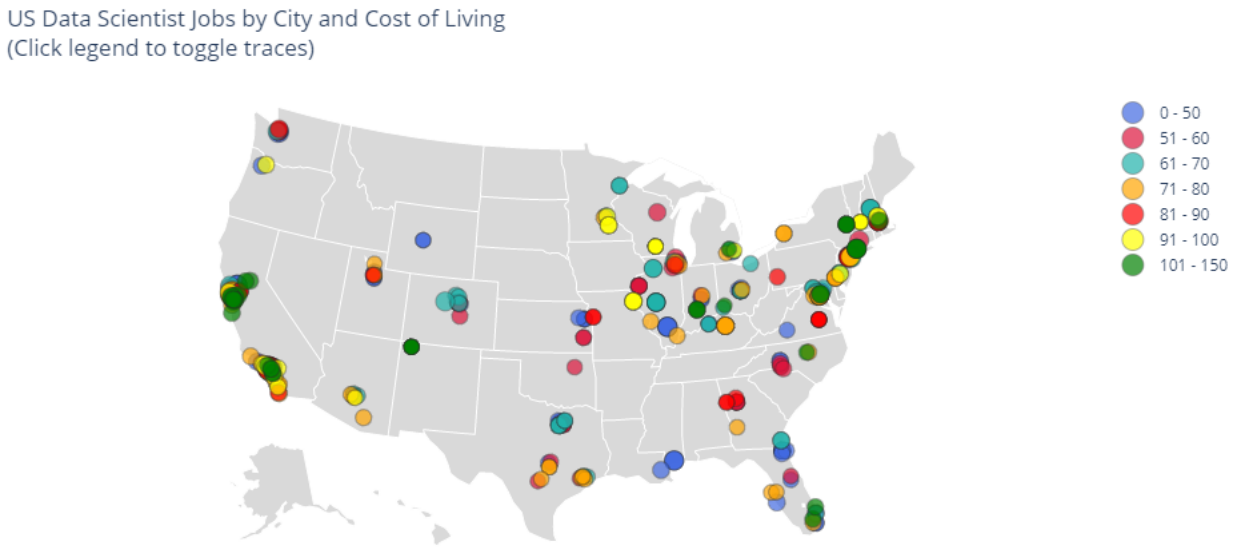
Salary Minimums and Job Count by City

Minimum Salary by City and Job Count (Color represents minimum salary times 1000, and size of bubble represents number of jobs in that location.)



Salary Minimums by City Compared to Cost of Living

Minimum Salary by City Compared to Cost of Living Index (Color represents minimum salary times 1000, and size of bubble represents Cost of Living Index.)



These graphs illustrate minimum salaries in large metropolitan areas of the United States. When interactive (in the Jupyter Notebook) hovering over a city provides additional information about the salaries, the cost of living, and other information relevant to a new data science job seeker.

One purpose of this project was to find the "ideal" job in the "ideal" city comparing salary to cost of living. The graph above does not differentiate either ideal based on these criteria. The second graph above illustrates pretty effectively by size of circle where data science jobs are located in major metropolitan areas and includes salary information as well. The first graph of city visualizations portrays salary information across the United States by city. Again, the interactive version on Jupyter Notebook is more informative.

Conclusions

As suspected, there is great overlap in language used in job titles and job descriptions in data science. The word clouds and bar graphs of the top 20 words in each job title's job descriptions show much overlap in responsibilities and tools used. While searching for the "ideal" job in the "ideal" city, the pattern that emerges is that most of the jobs are located along the US coastlines. At first New Mexico appeared to be "ideal" because jobs there have high salaries and the state has a low cost of living; yet there are few jobs in the state. Another conclusion reached from this project is remote work opportunities are increasing

because of the pandemic. With remote work, the job seeker can choose where to live so alignment between salary and cost of living is less important. The job seeker must consider many factors and make trade-offs based on the desired criteria. While this project was a great introduction to job hunting for a data science student beginning the process, there is definitely more to explore and learn.

Sources

Aakash, N.S. (2021). How to Land Your First Data Science or Machine Learning Job. Retrieved from: https://medium.com/jovianml/how-to-land-your-first-data-science-or-machine-learning-job-ffdcd17c8b61

Better Buys. (2019). Why We Need Women in Data Science. Retrieved from: https://www.betterbuys.com/bi/women-in-data-science/

Chatterjee, M. (2020). Top 9 Job Roles in the World of Data Science 2021. Retrieved from: https://www.mygreatlearning.com/blog/different-data-science-jobs-roles-industry/

Custer, C. (2018). How to Find an Entry-Level Job in Data Science. Retrieved from: https://www.dataquest.io/blog/how-to-find-an-entry-level-job-in-data-science/

Custer, C. (2019). Starting Your Career in Data Science: What Are Your Options? Retrieved from: https://www.dataquest.io/blog/career-guide-data-science-options/

Duranton, S., Erlebach, J., Brege, C., Danziger, J., Gallego, A., Pauly, M. (2020). What's Keeping Women Out of Data Science? Retrieved from: https://www.bcg.com/en-us/publications/2020/what-keeps-women-out-data-science

Indeed for Employers. (2021). How to Write a Data Science Job Description. Retrieved from: https://www.indeed.com/hire/job-description/data-scientist?aceid=&gclid=Cj0KCQiA-OeBBhDiARIsADyBcE7ev3QyxsrdfJB2leiOqoCw9RVWXfRnFPtZrv_Hx4BzMnW4qdWwkvsaArebEALw_wcB

Jung, H. (2019). 3 Strategies to Guarantee a Data Science Job with No Experience. Retrieved from: https://towardsdatascience.com/3-strategies-to-guarantee-a-data-science-job-with-no-experience-68d85b345f21

Matthews, K. (2019). How to Recognize a Good Data Scientist Job From a Bad One. Retrieved from: https://www.kdnuggets.com/2019/04/recognize-good-data-scientist-job-from-bad.html

Metwalli, S. (2020). 10 Different Data Science Job Titles and What They Mean. Retrieved from: https://towardsdatascience.com/10-different-data-science-job-titles-and-what-they-mean-d385fc3c58ae

Nolli, R. (2020). How to Identify the Most Requested Skills on the Data Science Job Market, with Data Science. Retrieved from: https://towardsdatascience.com/how-to-identify-the-most-requested-skills-on-the-data-science-job-market-with-data-science-726845ca9638

Adamovic, M. (2020). Northern America: Cost of Living Index by City 2020. Retrieved from: https://www.numbeo.com/common/about.jsp

Rohrer, B. (2017). How to Get a Data Science Job: A Ridiculously Specific Guide. Retrieved from: https://www.kdnuggets.com/2017/03/get-data-science-job-guide.html

Springboard. (2019). How to Become a Data Scientist with No Experience. Retrieved from: https://www.springboard.com/library/data-science/how-to-become-without-experience/

U.S. Bureau of Labor Statistics. (2019). Data Scientists and Mathematical Science Occupations, All Other. Retrieved from: https://www.bls.gov/oes/current/oes152098.htm