Data Set Title
Exploratory Analysis

Keegan Henderson, khenderson5@bellarmine.edu
Eric Schneider, eschneider@bellarmine.edu

I. INTRODUCTION

Our dataset is made to describe the different salaries for data science job openings with various companies around the United States. The dataset contains information about the minimum salary, maximum salary, average salary, job description, age of the company in years, though for the purposes of the project, we only use twenty-five of the forty-two columns. The full dataset can be found at https://www.kaggle.com/nikhilbhathi/data-scientist-salary-us-glassdoor

II. DATA SET DESCRIPTION

This dataset contains 742 rows with 42 columns of various data types, though only 25 of the will be used for the project. A complete listing of these variables can be seen in table 1.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Revenue	*Int64 (Ratio)*	27%
Size	*Int64 (Ratio)*	1%
Age	*Int64 (Ratio)*	7%
Rating	*Int64 (Ordinal)*	1%
Founding	*Obj (Ordinal)*	7%
Upper Salary	*Int64 (Ratio)*	0%
Lower Salary	*Int64 (Ratio)*	0%
Avg Salary(K)	*Int64 (Ratio)*	0%
Salary Spread	*Int64 (Ratio)*	0%
Job Title	*Object (Nominal)*	0%
Company Name	*Object (Nominal)*	0%
Location of Headquarters	*Object (Nominal)*	0%
Location of Job	*Object (Nominal)*	0%
Form of Ownership	*Object (Nominal)*	0%
Industry Type	*Object (Nominal)*	0%
Job Sector	*Object (Nominal)*	0%
Hourly Wage	*Bool (Nominal)*	0%
Python Experience	*Bool (Nominal)*	0%
Spark Experience	*Bool (Nominal)*	0%
Excel Experience	*Bool (Nominal)*	0%
Sql Experience	*Bool (Nominal)*	0%
Sas Experience	*Bool (Nominal)*	0%
Tableau Experience	*Bool (Nominal)*	0%
Pytorch Experience	*Bool (Nominal)*	0%
Google Analytics Certificate	*Bool (Nominal)*	0%

III. Data Set Summary Statistics

The most drastic differences between the variables would be in the revenue and size columns where the standard deviation is equal to 3761 million dollars and 3741 people respectively. However, when the IQR of both columns is examined, it can be seen that the size column actually has the true largest spread, with an IQR of 4800 people. The revenue column only has an IQR of 1.5 billion dollars or 1500 million dollars, meaning that the mean and standard deviation are both being skewed by outliers, most likely rightward outliers since the mean is larger than the median. The remaining integer columns are shown below.

Table 2: Summary Statistics for data_scientists_salary_2021 (name of dataset)

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25^{th}*	*50^{th}*	*75^{th}*	*Max*
Rating	*742*	*3.2*	*0.620*	*1*	*3*	*3*	*4*	*5*

Size	*742*	*2747.034*	*3741.615*	*51*	*201*	*1001*	*5001*	*10000*
Revenue	*742*	*2637.63*	*3761.47*	*1*	*500*	*500*	*2000*	*10000*
Upper Salary	*742*	*128.214*	*45.128*	*16*	*96.0*	*124.0*	*155.0*	*306.0*
Lower Salary	*742*	*74.754*	*30.94*	*15*	*52*	*69.5*	*91.0*	*202.0*
Avg Salary(K)	*742*	*101.48*	*37.48*	*15.5*	*73.5*	*97.5*	*122.5*	*254.0*
Age	*742*	*50.75*	*52.24*	*2*	*14.0*	*33.0*	*60.0*	*277.0*
Salary Spread	*742*	*53.45*	*19.2*	*1*	*41.0*	*53.0*	*65.0*	*143.0*

Table 3: Proportions for XXX (n=yyy)

Category	*Frequency*	*Proportion (%)*
Job Title (Data Scientist)	*313*	*42%*
Company Name (Takeda Pharmaceuticals)	*14*	*2%*
Headquarters Location (New York, NY)	*52*	*7%*
Job Location (New York, NY)	*55*	*7%*
Type of Ownership (Company – Private)	*410*	*55%*
Industry Type (Biotech and Pharmaceuticals)	*112*	*15%*
Job Sector (Information Technology)	*180*	*24%*
Founding Year (2010)	*32*	*5%*
Hourly Pay (False)	*718*	*96*
Python Experience (True)	*392*	*53%*
Spark Experience (False)	*575*	*77%*
Excel Experience (True)	*388*	*52%*
Sql Experience (True)	*380*	*51%*
Sas Experience (False)	*676*	*91.1%*
Tableau Experience (False)	*594*	*80%*
Pytorch Experience (False)	*703*	*94.7%*
Google Analytics Certificate (False)	*728*	*98%*

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn't make sense)

Table 4: Correlation Table/Tables

There are some very interesting correlations in this data set, or rather, lack of correlation. Nearly every relationship in the dataset of continuous variables have weak correlations, if any to begin with. Despite What would be assumed, the revenue of a company does not seem to influence the salary of a data science position. In fact, it is more likely that the company size influences the salary. However, there are a few places where the correlation is strong. Older companies tend to be larger ones. They also tend to earn more revenue. Other than variables that are naturally linked with each other such as upper and lower salary, there are no other significant correlations in the dataset.

IV. DATA SET GRAPHICAL EXPLORATION

The main focus of the dataset is the salaries of the numerous positions which can be held in data science. After all, that is the name of the overall dataset. However, the goal of this analysis is to show what else changes the salaries other than the type of job position. Certain computer language efficiencies as well as location and company rating have been known to lead to higher paying jobs, but is that actually the case. Through the following figures

Figure 1: Percent of positions requiring Python Experience

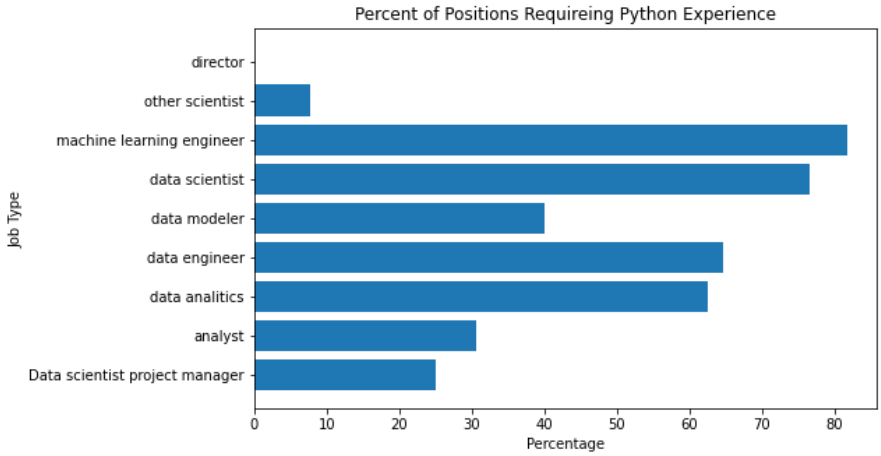


Figure 2: Average and Spread of Salaries of Data Science Careers

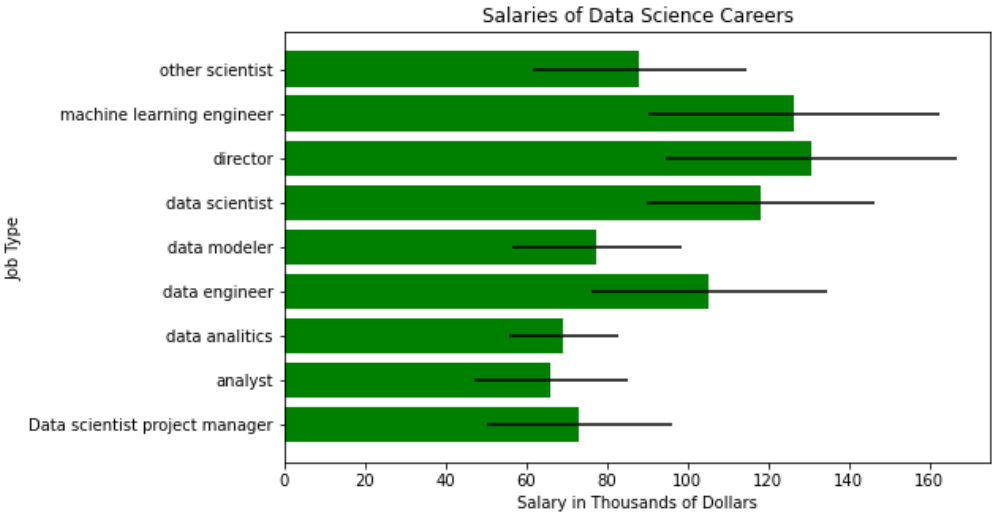


Figure 3: Average Salaries and count for Company Ratings

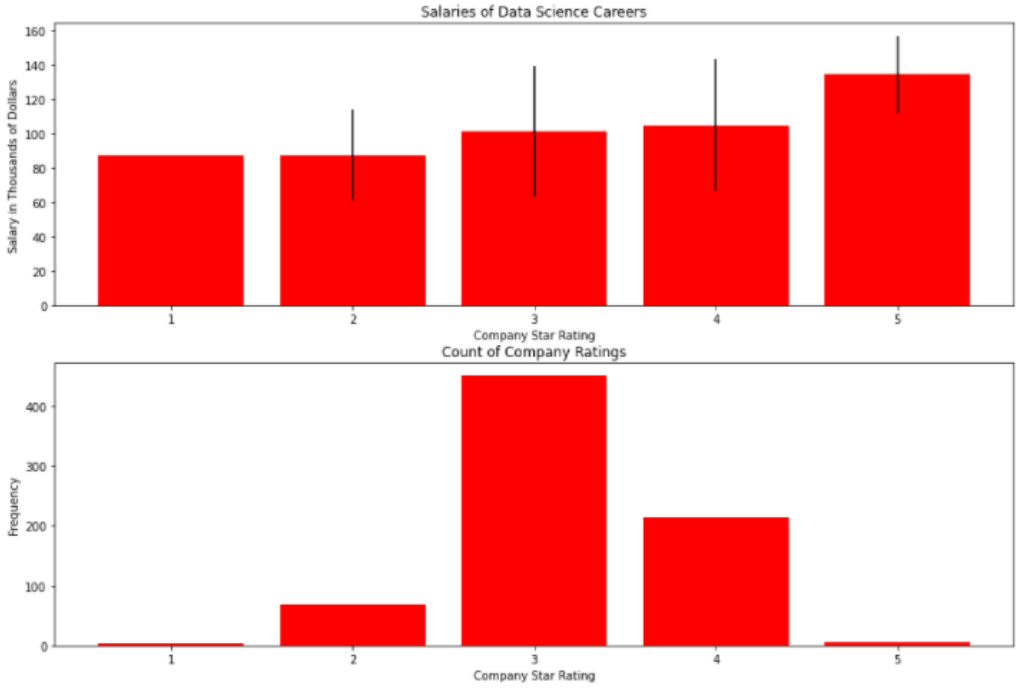


Figure 4: Change in Average Salary based on Company Age

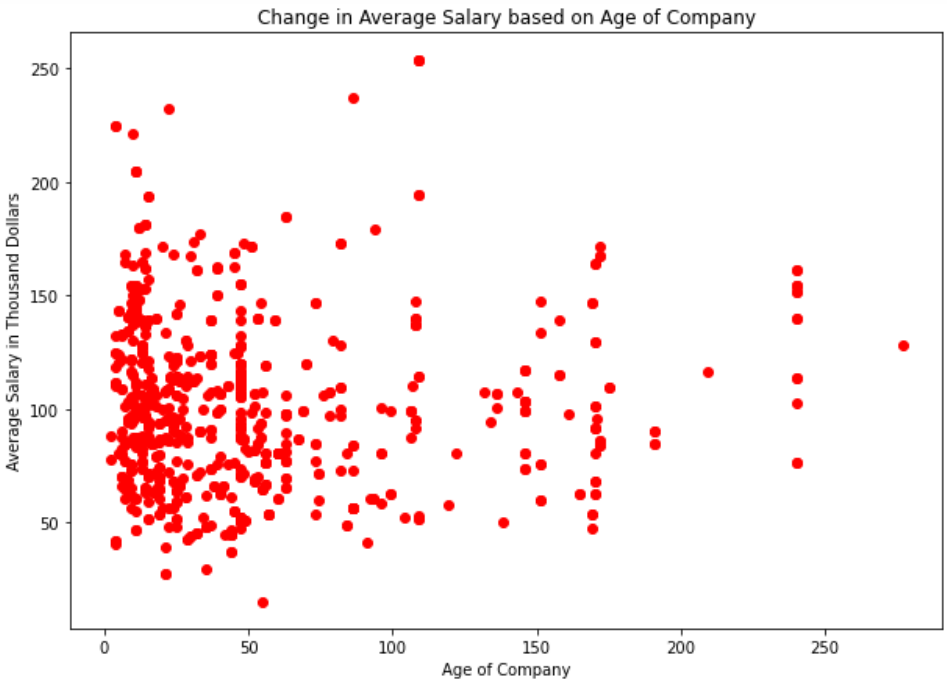


Figure 5: Company Size and Average Salaries based on the State which the Job position resides in.

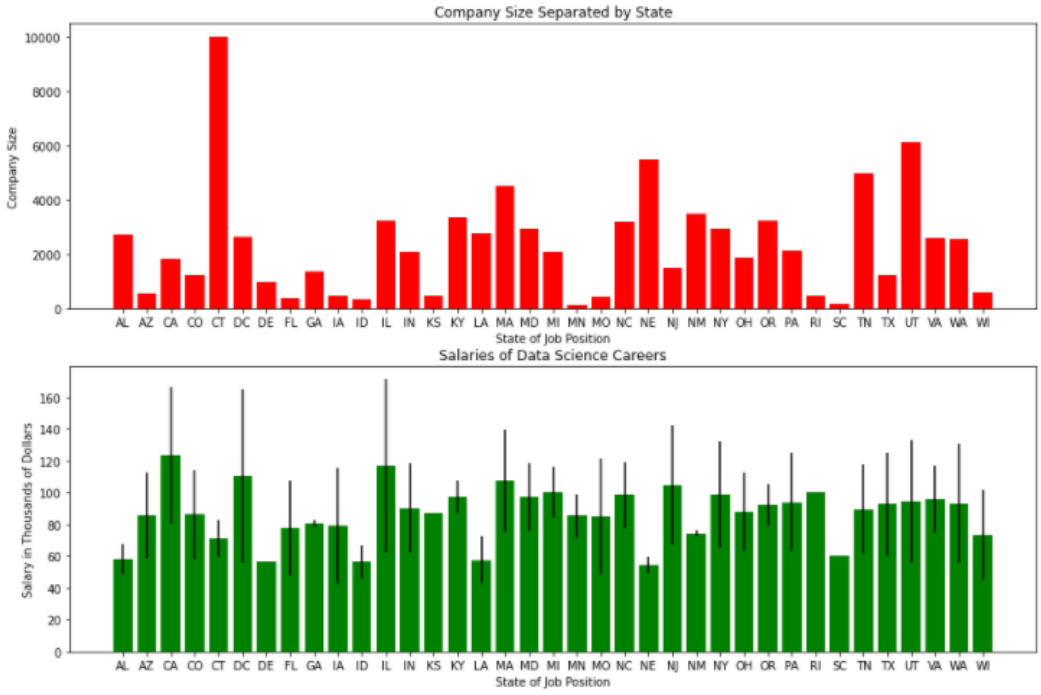
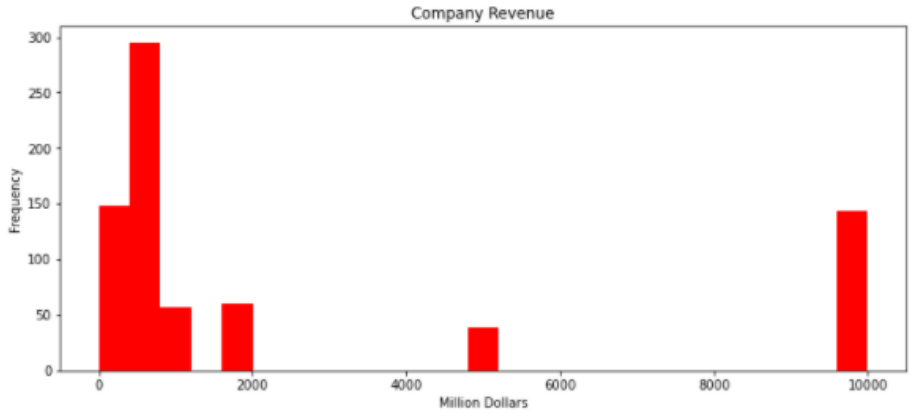


Figure 6: Histogram showing Frequency of Company Revenue



V. SUMMARY OF FINDINGS

Finish up with a paragraph or two of summarizing your findings about this data set.

Within this data set we observe the percentage of a job that requires experience in python. This allows for us to easily compare (visually) which jobs are more python intensive. Showing us that although they have very similar job titles their job description might have different tools they utilize. A couple interesting points of interest include the Director position and also the top machine learning engineer position. The director position requires no Python experience, which makes sense due to the heavy concentration of business and managing not programming. The Machine Learning Engineer is extremely high (#1) due to Python being a great tool for machine learning, which explains why it's so heavily required. This requirement might also explain why positions such as data scientists and machine learning engineers earn higher salaries.

Within this graph of figure 2 we can see the average salary (in thousands of dollars) for each job type within the data set. This gives us a visual representation of how much each job makes in reference to all the others, rather than just looking at the numbers. Some interesting points within the data are how data scientist project managers make significantly less than the actual data scientists, and also how the more you make the larger the spread is. Possibly due to the difference in skill, and that they are delegating these tasks out rather than being able to perform them. The second interesting point we can see is that the higher paying the job is, the higher the pay variability. This can be attributed to attributes such as seniority and placement on team (team worker, team lead, and so on). You can also look back to figure one and observe that those jobs requiring python experience are also the jobs that are higher paying.

In figure 3 we are showing how much the average data science Career makes at each company, based on their star rating. In general, we see an upwards trend from one star all the way to five-star companies. However, we do need to note that there a few one- and five-star companies so data could be skewed.
Figure four it depicts a spread of average salary for the companies age (each dot being one salary). The viewer of this graph can see that there is a heavy concentration of salaries from 0 to about 75-year-old companies. However, even as the companies age their salaries do not grow linearly with them; they stay roughly the same.

In figure five we see the average company size from each state, and also the average salary for each state. This allows us to see how these two variables are related to one another. In this case Location seems to be a confounding variable in the correlation between average salary and company size. Another confounding variable would be the revenue of the company. In figure six the graph represents the frequency of companies making a certain amount of revenue (in millions of dollars). We can see that there are a large number of companies making between 0 and 2 billion dollars. However, there are a number of outliers

making over 5 billion and another group making around 10 billion. Because of the fact that the majority of companies make less than one billion dollars, the graph can be used to explain the low correlation between salary and revenue. Even though some companies make much more revenue than others, they still pay their data scientists relatively the same salary.