Glassdoor Salary Listings
Exploratory Analysis

Becket D. Johnson, bjohnson21@bellarmine.edu
Sal M. Pusateri, spusateri@bellarmine.edu
Chris B. Douglas, cdouglas@bellarmine.edu

I. INTRODUCTION

Our data set, linked here—Salary Prediction, eda_data.csv—is a set of data science positions and salaries taken from Glassdoor listings in 2017. We decided to work with it due to its relevance to understanding some of our potential future careers. Through analyzing this data, we hope to see which skills have the strongest positive correlation with salary, the distribution of entry-level and junior positions to senior ones, how data scientists and analysts rate their jobs per industry, and more. These questions should lead to a better picture of what the data science field is looking like for us, as students who may enter it in the near future.

II. DATA SET DESCRIPTION

Originally, our set contained 742 samples with 33 columns with various data types. For various reasons, certain columns and many rows were removed for being mostly unclean or otherwise redundant or not useful. We have a list of major changes made from when we were cleaning, with the below as documentation:

- Variables to remove:
  - Salary Estimate—redundant with min and max salary variables included
  - Job Description—unimportant, long string info about job
  - Company_name—company_txt has the same info, but without a formatting issue of having the rating attached
  - Competitors—comp_num provides a number rather than list of names
  - desc_len—describes length of job description, also not needed
  - job title—job_simp is more useful for having a few categories of job to sort by vs dozens of titles with nonstandard formatting
  - Headquarters is also not useful, as salary data corresponds to job location and not company HQ's location
- Issues to clean
  - Age has impossible ages and some nulls, cut out those rows outside a realistic range
  - Across data set, the value -1 represents NaN; reformat to NaN for isnull().sum() count

After cleaning, we ended up with 312 rows and 25 columns of data.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)* *Pre-Cleaning*
Rating	Ratio, Float64	1.48
Location	Nominal, Object	0
Size	Ordinal, Object	0
Founded	Interval, Float64	6.74
Type of Ownership	Nominal, Object	0
Industry	Nominal, Object	0
Sector	Nominal, Object	0
Revenue	Ratio, Object	0
Hourly	Ordinal, Float64	0
Employer_provided	Ordinal, Float64	0
Min_salary	Ratio, Float64	0

Max_salary	Ratio, Float64	0
Avg_salary	Ratio, Float64	0
Company_txt	Nominal, Object	0
Job_state	Nominal, Object	0
Same_state	Ordinal, Float64	0
Age	Ratio, Float64	58.22
Python_yn	Ordinal, Float64	0
R_yn	Ordinal, Float64	0
Spark	Ordinal, Float64	0
AWS	Ordinal, Float64	0
Excel	Ordinal, Float64	0
Job_simp	Ratio, Object	0
Seniority	Nominal, Object	0
Num_comp	Ratio, Float64	0

III. Data Set Summary Statistics

The following seven numerical variables were summarized for being useful ratios that show key insights of the data science job market across 310 job postings. Of note are the summary statistics, which provide a general shape of the job market, and the heatmap's avg_salary row. As we can see, the average data science salary in 2017 from this dataset is $91,502, with about $36,000 of standard deviation. In other words, 2/3rds of data science salaries could be expected to fall between $55,500 and $127,500 in that year, assuming this set resembles a normal distribution. The other variables worth highlighting are (year) Founded and Rating. Taken together, most companies (as in, 25th percentile or greater) have above a 6.6/10 rating and are well-established 20–55-year-old businesses. This paints a relatively nice picture of working conditions for the industry, albeit a vague one.

Turning to the heatmap, it feels most important to mention that across every cell, very few have meaningful correlation coefficients. The hourly row is worth explaining; it implies that having an hourly pay will usually lead to lower salaries and are less likely to require Python. On a related noted, Python as a needed skill positively correlates with the salary range of a job posting; in fact, it is the 3rd strongest set of positive correlations, behind only the salary variables understandably being nearly perfectly correlated to each other, and employer_name and hourly's irrelevant relationship. Other than Spark having a similar but weaker correlation pattern to Python's, no other notable correlations are observed besides the outlier of age and founded being perfectly correlated. In fact, age has essentially zero correlation to salary, which is shocking; usually middle aged workers are at the peak of their salary range, but in this industry they are equal earners to those younger and older than them.

Table 2: Summary Statistics for Glassdoor Salary Data

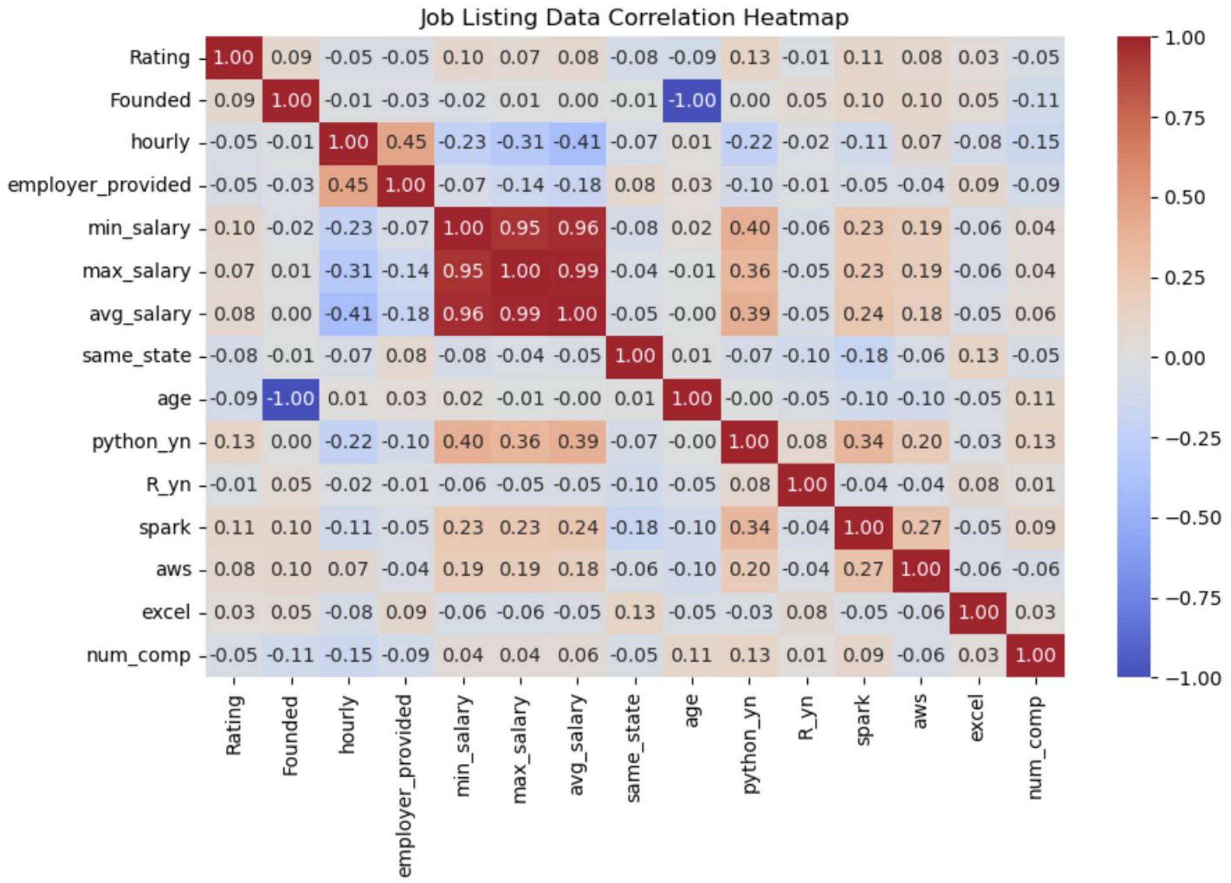
Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Rating	*310*	*3.605*	*0.560*	*1.9*	*3.3*	*3.6.*	*3.9*	*4.7*
Founded	*310*	*1981.035*	*16.025*	*1942*	*1968*	*1984*	*1996*	*2002*
Min_salary	*310*	*67.610*	*26.481*	*15*	*49*	*61*	*81.750*	*176*
Max_salary	*310*	*117.568*	*42.595*	*16*	*86*	*113.500*	*142*	*289*
Avg_salary	*310*	*91.502*	*35.820*	*13.500*	*67*	*87.500*	*112.875*	*232.500*
Age	*310*	*38.965*	*16.026*	*18*	*24*	*36*	*52*	*78*
Num_comp	*310*	*1.268*	*1.431*	*0*	*0*	*0*	*3*	*4*

There should be a table for **EACH** categorical variable.

Table 3: Proportions for Categorical Variables:

Please read: several categorical variables had far too many categories to fit into tables; please see the notebook (part 2d) instead for each category's value count. I'm aware of and regret that the fact that this will not include a % ratio for the proportions, but as there are 310 entries, dividing each value's frequency in a value count by 3 will give a fairly decent estimate. Sorry, and thank you for working with us here.

Table 4: Correlation Table



Job Listing Data Correlation Heatmap

	Rating	Founded	hourly	employer_provided	min_salary	max_salary	avg_salary	same_state	age	python_yn	R_yn	spark	aws	excel	num_comp
Rating	1.00	0.09	-0.05	-0.05	0.10	0.07	0.08	-0.08	-0.09	0.13	-0.01	0.11	0.08	0.03	-0.05
Founded	0.09	1.00	-0.01	-0.03	-0.02	0.01	0.00	-0.01	-1.00	0.00	0.05	0.10	0.10	0.05	-0.11
hourly	-0.05	-0.01	1.00	0.45	-0.23	-0.31	-0.41	-0.07	0.01	-0.22	-0.02	-0.11	0.07	-0.08	-0.15
employer_provided	-0.05	-0.03	0.45	1.00	-0.07	-0.14	-0.18	0.08	0.03	-0.10	-0.01	-0.05	-0.04	0.09	-0.09
min_salary	0.10	-0.02	-0.23	-0.07	1.00	0.95	0.96	-0.08	0.02	0.40	-0.06	0.23	0.19	-0.06	0.04
max_salary	0.07	0.01	-0.31	-0.14	0.95	1.00	0.99	-0.04	-0.01	0.36	-0.05	0.23	0.19	-0.06	0.04
avg_salary	0.08	0.00	-0.41	-0.18	0.96	0.99	1.00	-0.05	-0.00	0.39	-0.05	0.24	0.18	-0.05	0.06
same_state	-0.08	-0.01	-0.07	0.08	-0.08	-0.04	-0.05	1.00	0.01	-0.07	-0.10	-0.18	-0.06	0.13	-0.05
age	-0.09	-1.00	0.01	0.03	0.02	-0.01	-0.00	0.01	1.00	-0.00	-0.05	-0.10	-0.10	-0.05	0.11
python_yn	0.13	0.00	-0.22	-0.10	0.40	0.36	0.39	-0.07	-0.00	1.00	0.08	0.34	0.20	-0.03	0.13
R_yn	-0.01	0.05	-0.02	-0.01	-0.06	-0.05	-0.05	-0.10	-0.05	0.08	1.00	-0.04	-0.04	0.08	0.01
spark	0.11	0.10	-0.11	-0.05	0.23	0.23	0.24	-0.18	-0.10	0.34	-0.04	1.00	0.27	-0.05	0.09
aws	0.08	0.10	0.07	-0.04	0.19	0.19	0.18	-0.06	-0.10	0.20	-0.04	0.27	1.00	-0.06	-0.06
excel	0.03	0.05	-0.08	0.09	-0.06	-0.06	-0.05	0.13	-0.05	-0.03	0.08	-0.05	-0.06	1.00	0.03
num_comp	-0.05	-0.11	-0.15	-0.09	0.04	0.04	0.06	-0.05	0.11	0.13	0.01	0.09	-0.06	0.03	1.00

IV. DATA SET GRAPHICAL EXPLORATION

The univariate distributions largely held little insight that could not be gleaned from prior knowledge of the industry or guessed easily; the average age is weighted to those between 18-46, the vast majority of jobs are salaried and not hourly, the majority of jobs had applicants list Python as a skill, and so on. This pattern repeats for the univariate barplots as well. Considering the heatmap had relatively little correlation across most bivariate comparisons, expectations were lower for the bivariate statistics. The pairplot, even when sorted for the most correlated data, didn't offer much to see.

A. Distributions

No notable insights here, but the following graphs provide decent visual perspective on the job market.

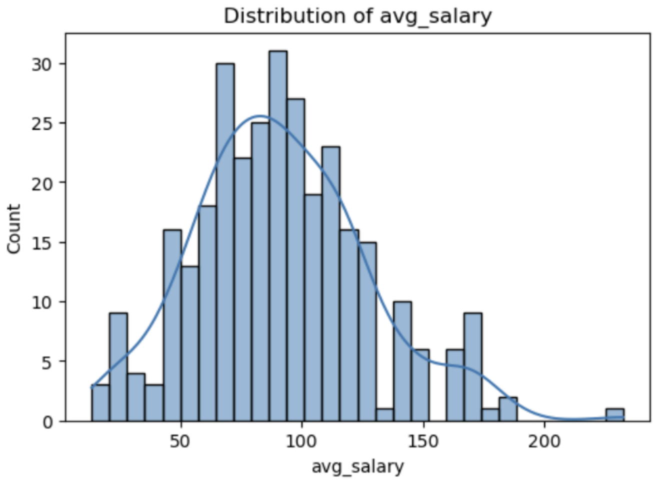


Figure 1: Unsorted Distribution of Average Salary (Histogram with KDE)

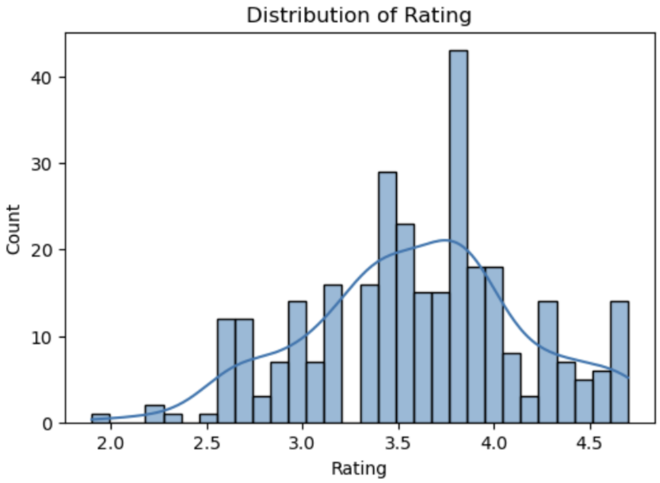


Figure 2: Unsorted Distribution of Company Rating (Histogram with KDE)

B. ScatterPlots / Pairwise Plots (continuous variables)

Given how weak the correlation was between any 2 variables, a pair plot seemed most appropriate to hone in on potential insights in the strongest areas of correlation: salary and Python skills. This was used to guide further exploration in section IV D, rather than to get anything out of itself. The struggle for conclusions from the data analysis will be a major theme of the summary.

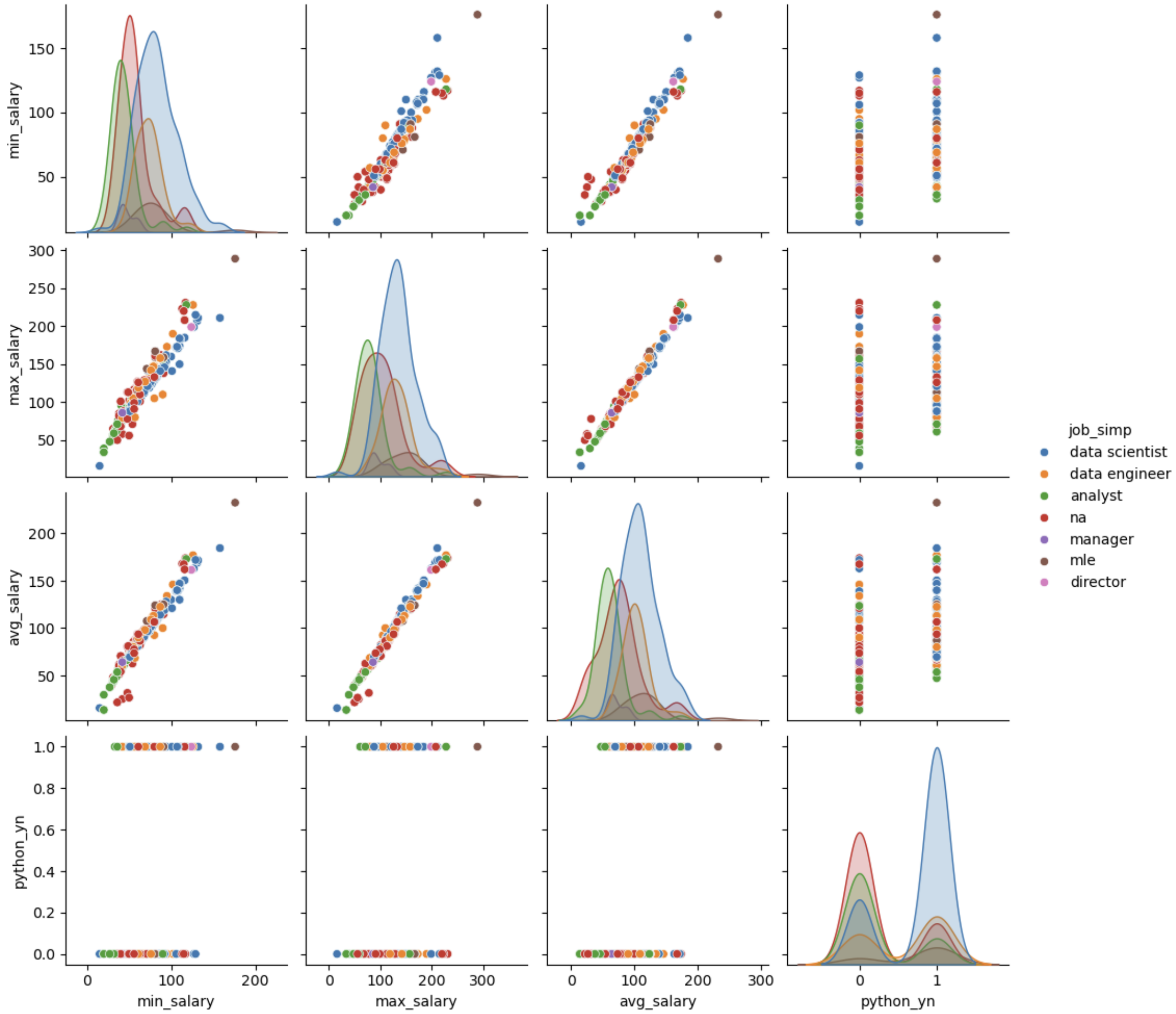


Figure 1: 4x4 Grid of Salary Statistics and Python experience, with Hue by Job Title (pair plot)

C. Barcharts (categorical variables)

Same as section A: few insights but some useful graphics were generated.

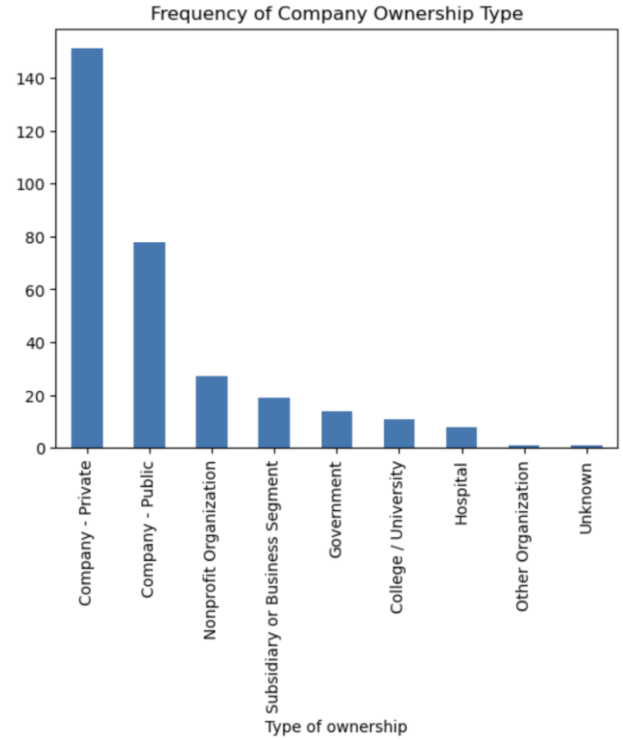


Figure 1: Descending Order Frequency of Type of Company (bar plot)

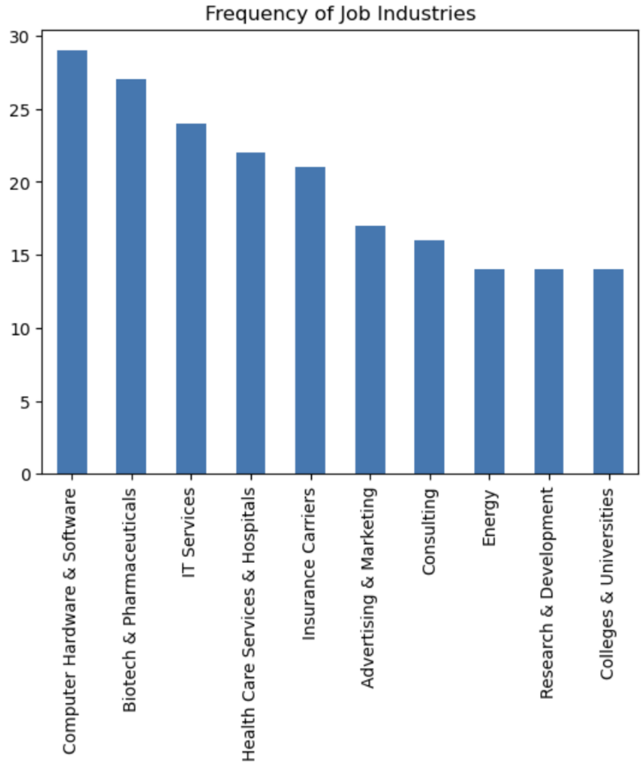


Figure 2: Descending Order of Most Common Industries for Job Listings (bar plot)

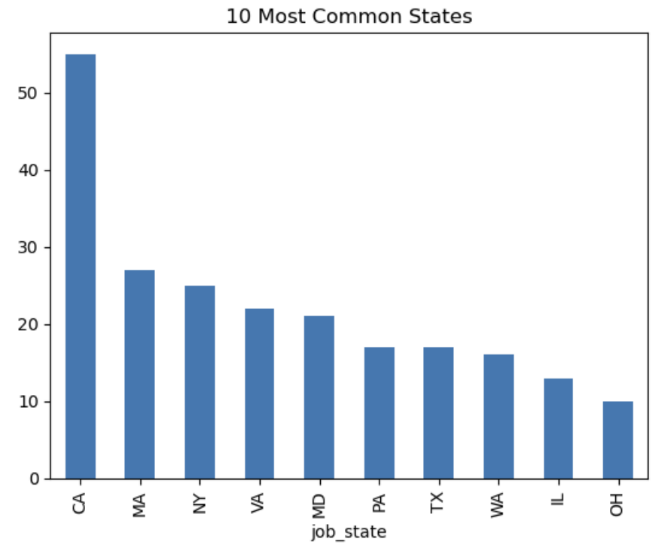


Figure 3: Top 10 Most Common States for Location of Jobs (bar plot)

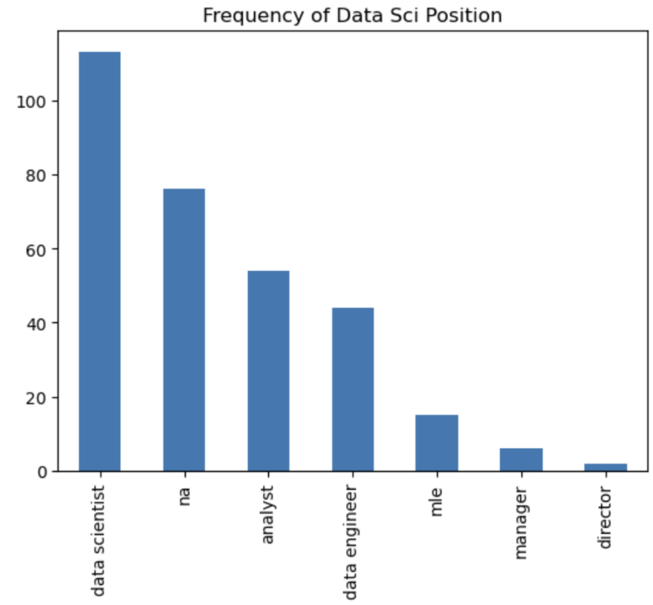


Figure 4: Proportion of Data Science Positions Across Job Postings (bar plot)

D. Other Plots

Choosing what to explore with this dataset has been difficult; a lot of our early hypotheses on what variables strongly explained others have been disproven by the heatmap; correlation is not causation, but no correlation means that none of these factors are really causing any other (barring what has been discussed already). So for this final section of plots, the focus here is finding practical information that other data scientists interacting with the job market could use in their decision making and future planning. This emphasizes the variables related to salary, rating, and possession of skills; the following graphs reflect this focus.

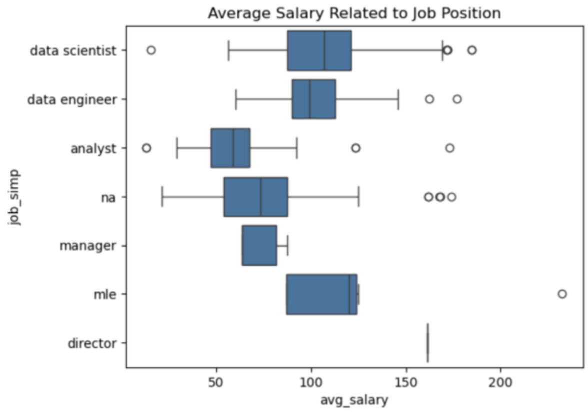


Figure 1: Relation between Average Salary and Job Position (box plot)

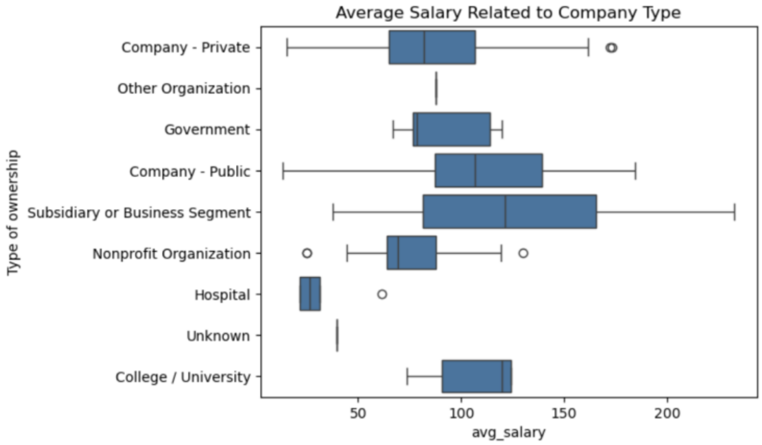


Figure 2: Relation between Average Salary and Job Position (box plot)

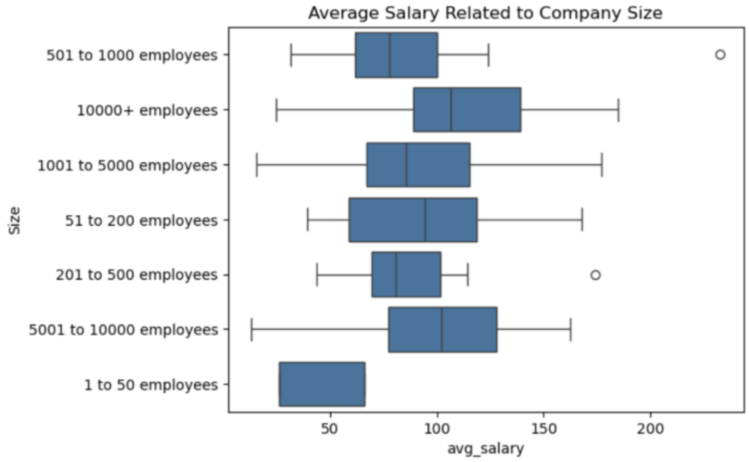


Figure 3: Relation between Average Salary and Company Size (box plot)

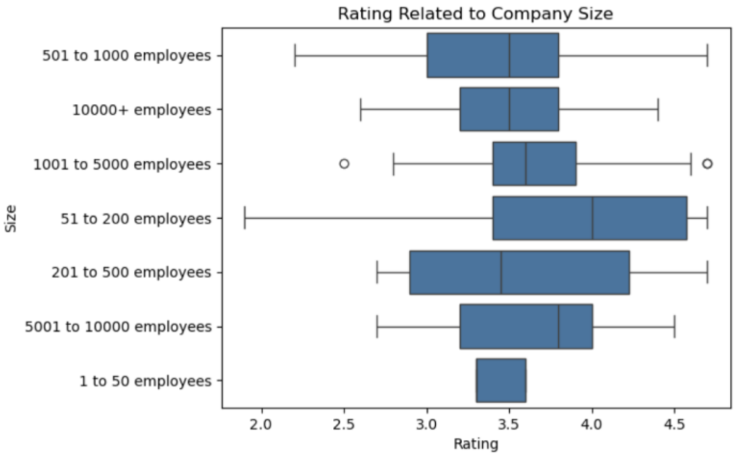


Figure 4: Relation between Rating and Company Size (box plot)

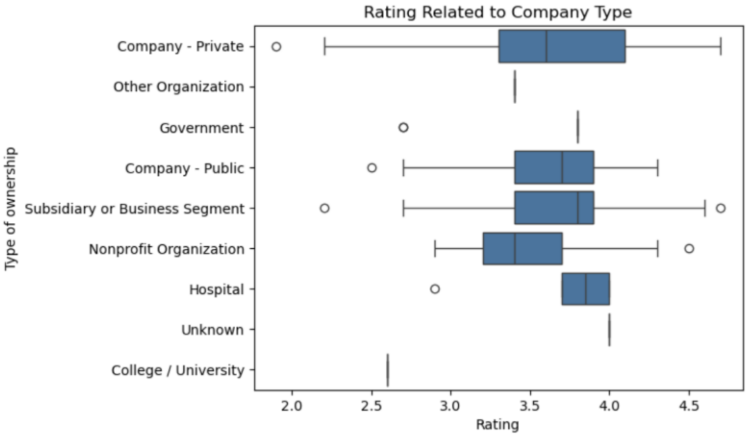


Figure 5: Relation between Rating and Type of Company (box plot)

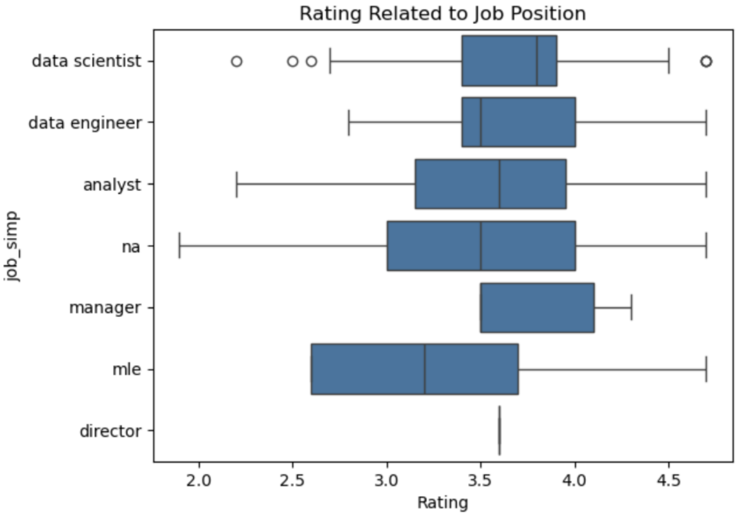


Figure 6: Relation between Rating and Job Position (box plot)

V. SUMMARY OF FINDINGS

If there was one thing to emphasize about the data, it would be this. Do not assume any strong correlations exist between most aspects of a data science job and its employee besides very simple facts like seniority, larger company size, and Python skills leading to higher salaries. It feels underwhelming to say after so much statistical work that there aren't many insights here, but there simply aren't. Data scientists and engineers tend to do the best in measures of company rating and salary, and smaller and very large companies are in a similar position. These jobs concentrate in wealthier states with large urban centers like Massachusetts, New York, and Boston. It feels obvious to say, but these are the most defined characteristics of the data set. This analysis lends a very detailed view of the job market during this time, but it is only an outside view; any ‘why’s are vague and few in number.

Focusing on what could be done differently, what comes to mind first is seeing if this dataset holds up across time and across other job platforms. This is only 300 postings from one website in 2017, but what about years after? Data should always be relevant, and this analysis will become less so as we move away from that period. In other words, the data needs to be broader, timelier, and larger in volume. Then it could be more certain if these low correlations across the board are that way, or merely in our sample. As we learn more statistical methods and ways to use them in software packages, perhaps better methods and new angles to analysis the data would reveal information that we as inexperienced students are missing. Other than these things, we are unsure how else to glean better results.