Abby Liu

Final Project Report: Data Analyst Job Salary Prediction

Problem Statement

The COVID-19 pandemic has severely affected the global economy and financial markets, which has caused a rise in the number of unemployed workers across the globe. In a brutal job market like this, it becomes harder for people to find a good job with a salary proportionate to their market values. On one hand, employers want to empower their teams and attract more high-quality job candidates by setting competitive salary levels; on the other hand, job seekers want to understand their current worth in the job market, find out if they are being paid fairly, and explore ways to increase their pays. Based on my observations, a lot of jobs posted on the job search platform don't include information about salary.

This salary prediction model aims to provide a means to help estimate/predict the salary ranges of data analyst jobs based on information about sector, location, company size, revenue, skills required, and company rating. It will help answering the following questions:

- For job seekers: What limits or expectations can job applicants have about salary when looking for Data Analyst jobs?
- For employers: How to set a competitive and reasonable salary to attract and retain talent?

Data

Our job salary dataset was downloaded from Kaggle, and contains more than 2000 job listings scraped from Glassdoor for data analyst positions, with features such as salary estimate, location, company rating, job description, and more.

Data Wrangling

The raw dataset from Kaggle contains 2253 rows and 16 columns. After converting some null data like -1, '-1', -1.0, 'Unknown' and 'Unknown / Non-Applicable' to NaN, I've performed some data cleaning and data transforming on the following features:

- Removed some useless features like 'Unnamed: 0'.
- Removed columns ('Easy Apply' and 'Competitors') that contain more than 50% of missing values
- Transformed the feature 'Founded' to 'Years Founded'.
- Split 'Salary Estimate' to 'Min_Salary' and 'Max_Salary'.
- Cleaned the column 'Company Name'.
- Transformed the feature 'Job Title' to 'Seniority'.
- Removed city names from 'Location' and transformed 'Headquarters' to 'HQ_Same' to indicate whether the headquarter is at the same location as the company..
- Added features 'Python', 'SQL', 'R', 'Excel', 'SAS', and 'Tableau' since those are the most popular technical skills that we extracted from 'Job Description'.

I've also dropped 164 rows that contained more than 50% of missing values and 1 row that had missing salary information. Outlier data points were looked at individually to determine if the number was an incorrect entry or legitimate measurement, and the former had been corrected.

At this point, we have 2088 rows and 20 columns left in our cleaned data.

Exploratory Data Analysis

Some EDA was done on the cleaned dataset, which consists of 4 numeric features('Rating', 'Years Founded', 'Min_Salary', and 'Max_Salary') and 16 categorical features('Company Name','Location', 'Size','Type of ownership', 'Industry', 'Sector', ‘job description’, 'Revenue','Seniority', 'HQ_Same', 'SQL', 'Excel', 'Python', 'Tableau', 'R', and 'SAS').

a) Skills required

I first looked at the distribution of my data, and I quickly noticed that out of the 6 skills I listed above, SQL, Excel, and Python were the top three most popular skills mentioned in the job description for data analyst positions.

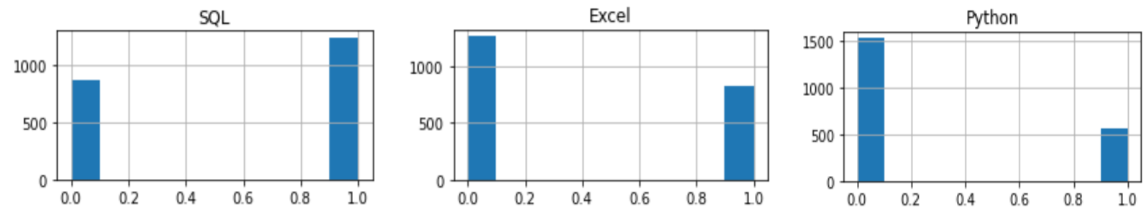


Figure 1: Histograms of the top three skills

We can see that SQL is the only skill that is required by more than 50% of the jobs; moreover, 39% of the jobs require experience in Excel, followed by Python (27%).

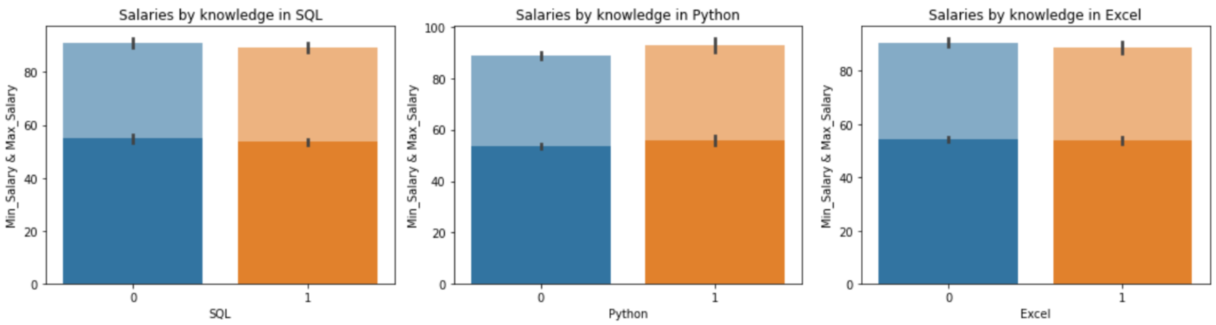


Figure 2: Bar plots that show the differences of salaries when a skill is required vs. salaries when a skill is not required.

However, as it's shown in figure 2, to investigate the relationship between skills and salary, if we grouped our data by a specific skill and tried to compare their minimum salary and maximum salary, we would see that having a specific skill mentioned in the job descriptions does not have a direct impact on salary.

b) Location

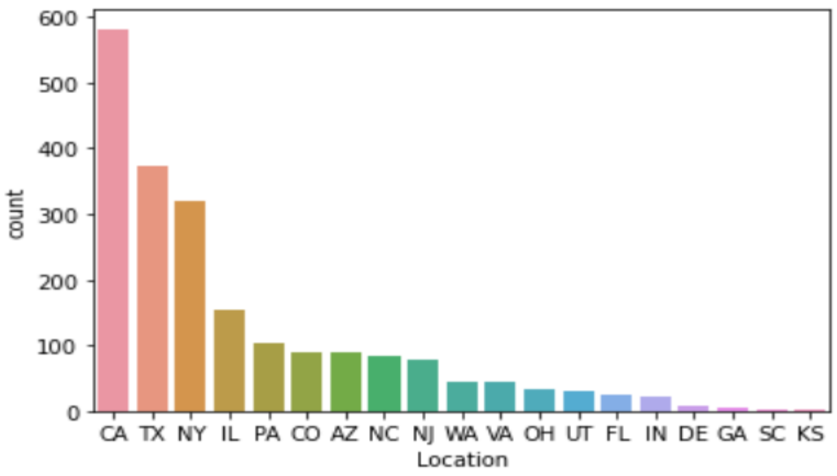


Figure3: Countplot for location

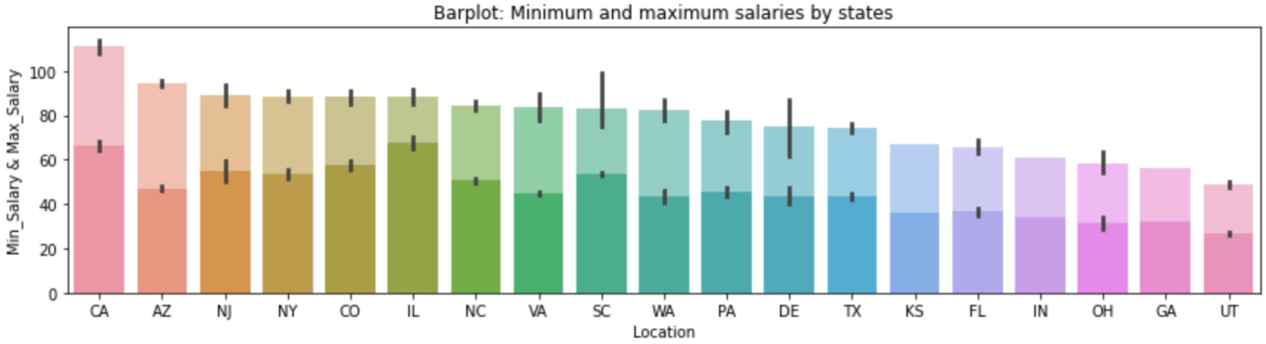


Figure 4: Barplot of mean minimum & maximum salary of different states

Figure 3 shows that more than 550 job postings are from CA, which indicates that there is a great demand for data analysts in California. Also, from figure 4, we can see that California not only has the highest amount of job postings, but it also has the highest mean maximum salary and the second highest minimum salary. The job market for data analysts in California seems to be the greatest. This was also confirmed in the subsequent modeling section when I was looking at feature importances. Besides the findings about CA, it's worth mentioning that Illinois has the highest mean minimum salary, and Utah has both the lowest mean minimum salary and the lowest mean maximum salary among all states in the United States.

c) Sector

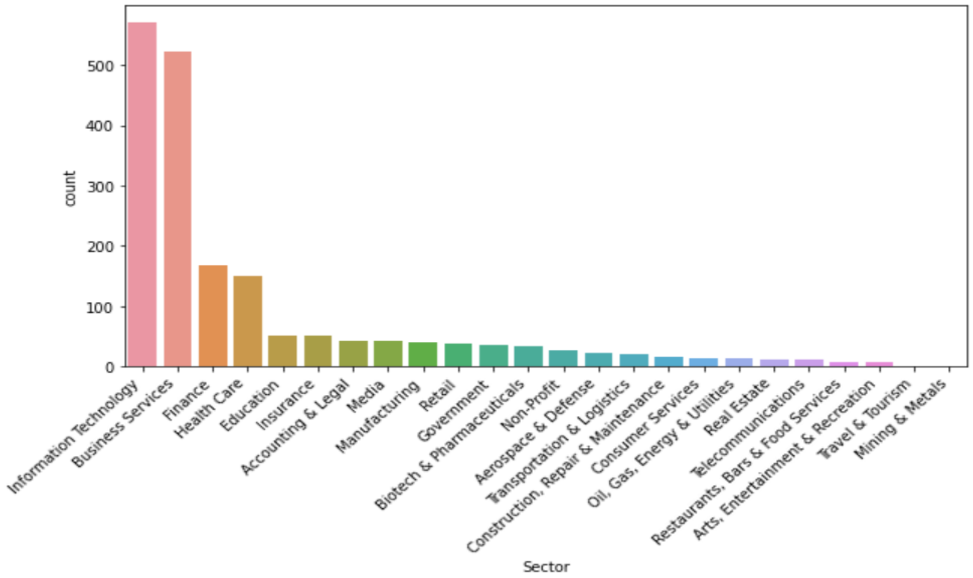


Figure 5: Countplot for sector

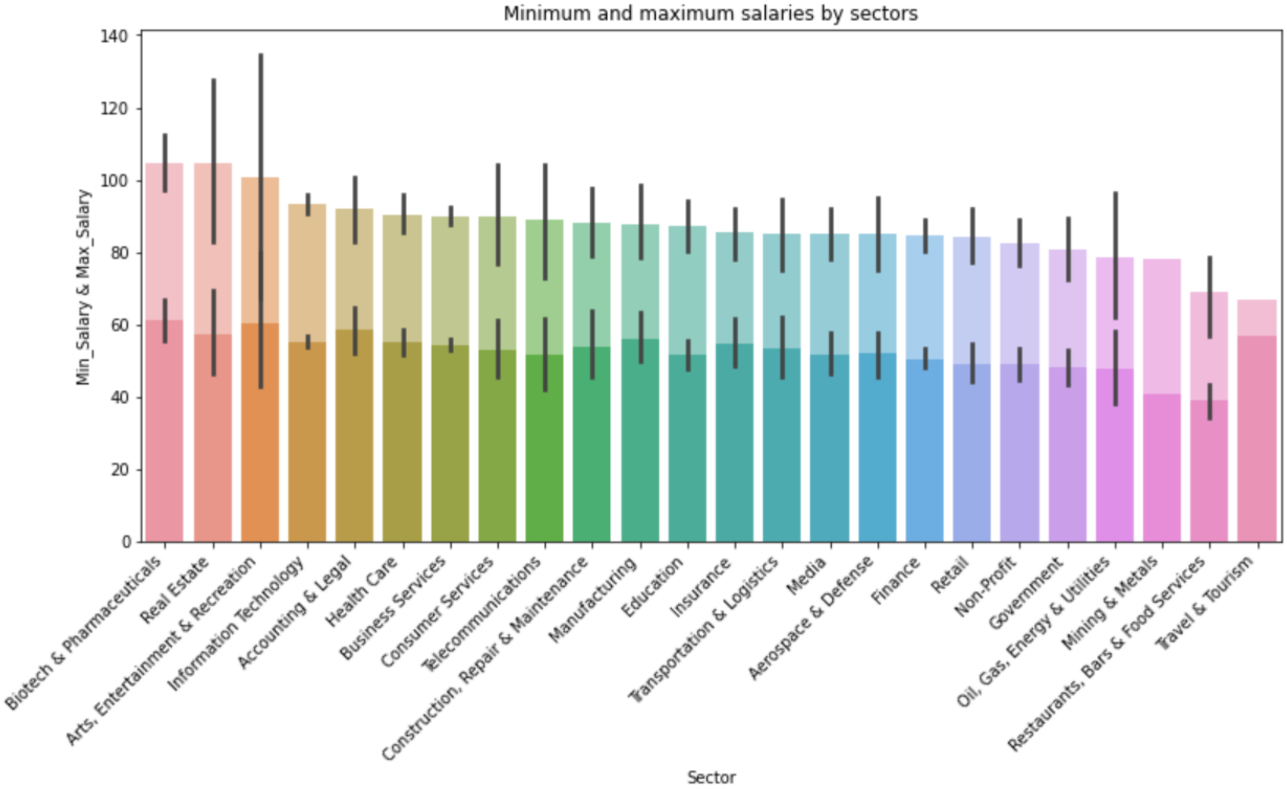


Figure 6: Barplot of mean minimum & maximum salary of different sectors

The countplot for sector(figure 5) shows us that IT is the number 1 sector/industry hiring data analysts. Some other top sectors that hire data analysts include business services, finance, and health care. As for salary, it does not vary too much by sectors compared to locations, but both the biotech & pharmaceuticals sector and real estate sector offer great pay($60K-$105K) to data analysts. Since there’s only one observation in the travel & tourism sector, the result may be biased, I’ll not make a conclusion based one this one. However, we can still see that the restaurant and food sector has the lowest pay($42K-$70K) to data analysts.

d) Seniority and Salary

To determine if senior positions offer better salaries, I grouped the data by seniority and compared the mean minimum salary and mean maximum salary of each group.

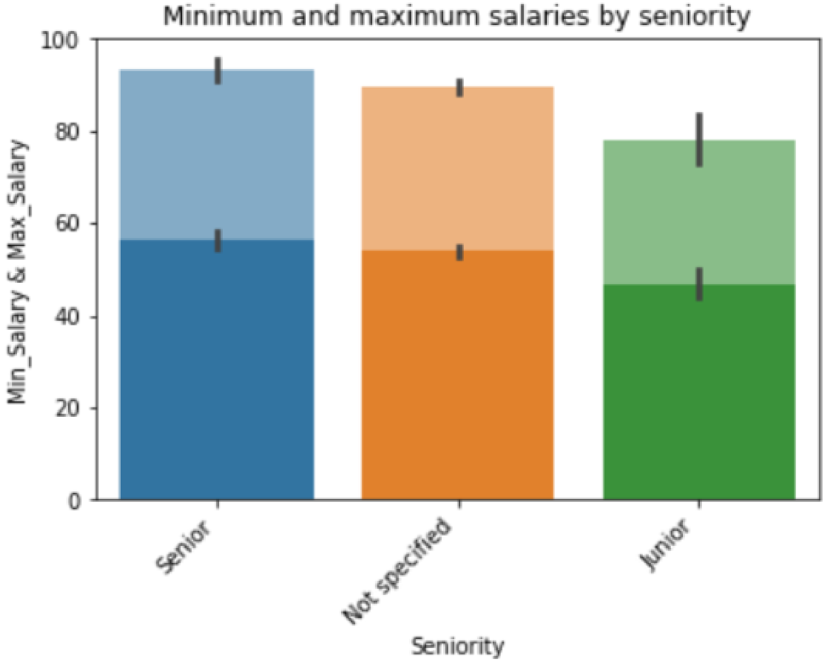
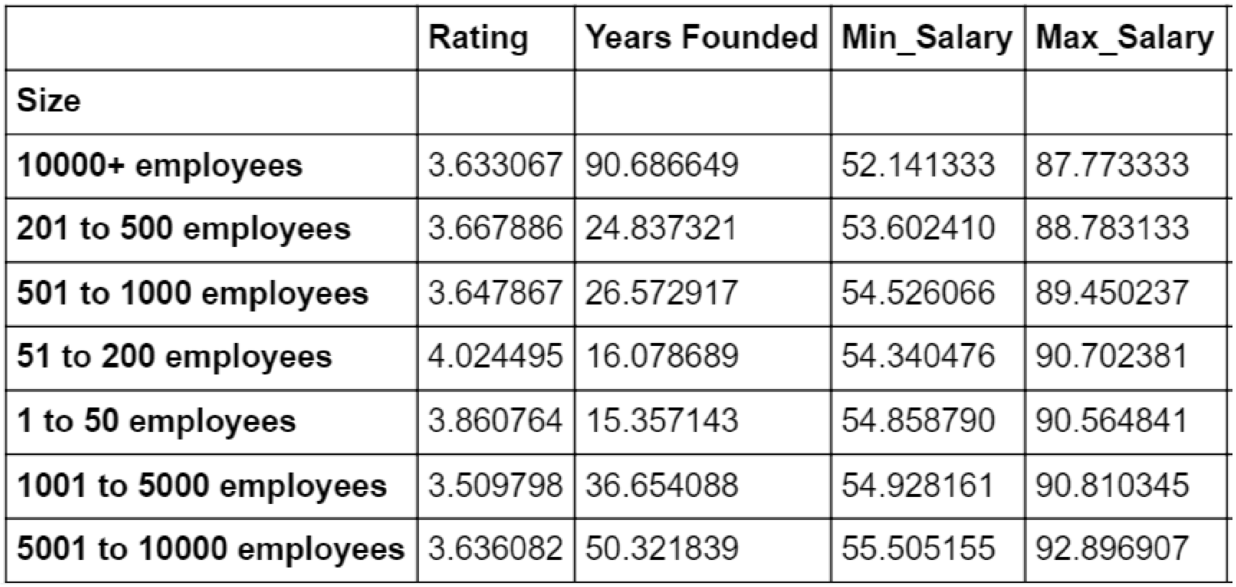


Figure 7: Mean min_salary and mean max_salary in each seniority group.

According to figure 7, not surprisingly, senior data analysts($58K-$92K) are paid slightly more than those whose titles did not specify seniority, and they are paid a lot more than junior data analysts($48K-78K).

e) Size & Years Founded

I was interested in seeing if bigger companies pay more than smaller companies, so I grouped the data by size and compared the mean minimum salary and the mean maximum salary in different size groups as shown in figure 8.



	Rating	Years Founded	Min_Salary	Max_Salary
Size				
10000+ employees	3.633067	90.686649	52.141333	87.773333
201 to 500 employees	3.667886	24.837321	53.602410	88.783133
501 to 1000 employees	3.647867	26.572917	54.526066	89.450237
51 to 200 employees	4.024495	16.078689	54.340476	90.702381
1 to 50 employees	3.860764	15.357143	54.858790	90.564841
1001 to 5000 employees	3.509798	36.654088	54.928161	90.810345
5001 to 10000 employees	3.636082	50.321839	55.505155	92.896907

Figure 8: Table of data grouped by Size.

Figure 8 suggests that there doesn't seem to be a linear correlation between company size and salary. Bigger companies don’t necessarily pay more to data analysts. Another interesting finding is that there seems to be a positive linear correlation between Size and Years Founded. The older the companies, the bigger they are. This finding became helpful when I was doing imputation for missing values later. I imputed the missing values in the ‘YearsFounded’ column using the mean values of YearsFounded in different size groups, respectively.

To prepare for the preprocessing and training step, we've dropped the features 'Job Description','Company Name' and 'Industry', and we saved the rest to a new data file.

Modeling

a) Preprocessing and Training

Before I started building my models, here were some data preprocessing steps I did:

- Split the data into training and testing sets (Note: we have two dependent variables: Min_Salary and Max_Salary)
- Impute missing numerical values using their corresponding medians
- Scaled the numerical data to zero mean and unit variance

- Impute missing categorical data using their most frequent values
- Create dummy or indicator features for categorical variables

b) Model Selection

After these preprocessing steps, I was ready to dive into the modeling section. Since my goal was to predict salaries, I had a regression problem. Here I had tried out some models: Linear Regression (without regularization), Lasso Regression, Random Forest Regressor, and Support Vector Regression (SVR).

Evaluating the performance of a model by training and testing on the same dataset could lead to overfitting. To prevent that, I used the Cross-Validation technique under the 5-fold CV approach. The metrics I focused on when building and comparing my models were R-squared and mean absolute values(MAE).

I carried out the grid search CV for hyperparameter tuning for all models separately except for the linear regression model, then I evaluated the r-squared score with the optimized hyperparameters. I then had three models with their best parameters for me to compare:

- Lasso Regression Model with alpha = 0.05
- Random Forest Regressor Model with max_depth = 10, max_features = sqrt, and n_estimators = 780
- Support Vector Regression Model with C = 21 and epsilon = 5

I compared these three models by their MAE scores. The result indicates that the Lasso, Random Forest, and SVR models have about the same cross-validation mean absolute errors for the training set, but my Lasso Regression model performs slightly better than the others when predicting on the test set. Reviewing the predictions of my model compared to the actual outcomes in my test dataset, the Lasso model gives a MAE of 11.72 for the prediction on minimum salary, and a MAE of 37.47 on the prediction on maximum salary.

Therefore, I've selected the Lasso model as the final model for job salary prediction. When predicting minimum salary, it's expected to be off by $11K; and when predicting maximum salary, it's expected to be off by $37K.

It's worth mentioning that both our Lasso and Random Forest models show that the locations CA and IL have a big positive impact on data analyst salary, and the locations UT ad OH have negative impact on data analyst salary. Moreover, the features YearsFounded and Rating are also shown to be important features by our Random Forest model.

Takeaways

Based on my findings, when determining the expected salary for a data analyst position, location, seniority, sector, rating of the company, and the number of years the company has been founded may be the most important features to be considered, especially location.

Future work

My data was scraped from Glassdoor, and the salary ranges for jobs in this dataset were not the real salary that the employers offered. Instead, they were more likely to be salary ranges estimated by Glassdoor, and that's an important reason why the predictions of my model are not very accurate compared to the actual outcomes. However, the process that I followed could be replicated if we have a better dataset that contains real salary information.

The job market for data analysts seems to be the best in California, it has the highest avenge pay and the biggest demand. Therefore, we can further look more into the CA job market specifically and find out which cities have the best job opportunities for data analyst job seekers.

For the skill features('SQL', 'Excel', 'Python', 'Tableau', 'R', and 'SAS'), we can't see a clear pattern of their relationships with salaries, and we can do some hypothesis tests to investigate deeper and confirm our observation.