Data Scientist Salaries

Carson Brower, cbrower@bellarmine.edu
Garrett Ward, gward2@bellarmine.edu

ABSTRACT
For this Final Project, our group used a Data Scientist Salary dataset from Kaggle. We used this dataset because of the question we had, can we predict a salary of a Data Scientist based on variables that are in connection and are closely related to what affects salaries. Throughout this document our data will be shown with graphs, plots, and a correlation map of our variables. The variables, along with the use of linear and logistic regression will allow for the most accurate results to allow us to produce a conclusion of our findings.

I. INTRODUCTION

To start off, with the use of Kaggle, we found a dataset that was interesting to us and consisted of a topic we thought would be good to present to the class since it is related to the class. The data set we chose was data relating to data scientist salaries. Within the data set there are variables such as job title, salary estimate, rating, location, size of company, company age, different software is, and a few more. After looking at the data set in its entirety we thought this was a good data set to do our final project on. From the dataset, the classification target, or the question we had, can we predict what salary a data scientist is going to have? Using the variables that closely relate and using the data for graphs, plots, matrix, and testing will allow us to produce the most accurate results and allow us to find our conclusion and solutions to the classification target of the data.

II. BACKGROUND

Data scientists are analytical people, they use data to understand and explain the things that happen around them and help organizations make better decisions. Careers in data are growing rapidly and is it becoming a more popular and needed career choice. Within their job they often develop predictive models for theorizing and forecasting. Data scientist find patterns and trends to uncover insights as well as create algorithms and data models to forecast outcomes. The use of machine learning helps improve the quality of the data and allows for communication of recommendations to others. For the data set we used itself, it contains job postings from Glassdoor from 2017. It can be used to analyze current trends based on job positions, company size, etc. This data set can be used to identify which factors most affect data science salaries, determine which states and cities offer higher pay, and predict what data science job will pay based on job description.

III. EXPLORATORY ANALYSIS

This data set has 732 entries with thirty-two columns. From our data there were no missing values or missing data, however there were a few miscues within one of our variables that we had to fix and fill in ourselves. We had a wide range of variable types in our data from int, float, and even object. When we first started working with our data, we noticed an age column and thought it meant the person’s age, but it meant the age of the company and that required us to use that in a unique way than originally. Something that stood out to us was the skew to the right on salary distribution where majority of the data scientist made anywhere between $50,000 and $100,000. Something that was surprising to us was the fact that company age really did not have an influence on salary. Our EDA graphs, plots, and more are shown below.

Table 1: Data Description

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 732 entries, 0 to 731
Data columns (total 33 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         732 non-null    int64
 1   Job Title          732 non-null    object
 2   Salary Estimate    732 non-null    object
 3   Job Description    732 non-null    object
 4   Rating             732 non-null    float64
 5   Company Name       732 non-null    object
 6   Location           732 non-null    object
 7   Headquarters       732 non-null    object
 8   Size               732 non-null    object
 9   Founded            732 non-null    int64
 10  Type of ownership  732 non-null    object
 11  Industry           732 non-null    object
 12  Sector             732 non-null    object
 13  Revenue            732 non-null    object
 14  Competitors        732 non-null    object
 15  hourly             732 non-null    int64
 16  employer_provided  732 non-null    int64
 17  min_salary         732 non-null    int64
 18  max_salary         732 non-null    int64
 19  avg_salary         732 non-null    float64
 20  company_txt        732 non-null    object
 21  job_state          732 non-null    object
 22  same_state         732 non-null    int64
 23  age                732 non-null    int64
 24  python_yn          732 non-null    int64
 25  R_yn               732 non-null    int64
 26  spark              732 non-null    int64
 27  aws                732 non-null    int64
 28  excel              732 non-null    int64
 29  job_simp           732 non-null    object
 30  seniority          732 non-null    object
 31  desc_len           732 non-null    int64
 32  num_comp           732 non-null    int64
dtypes: float64(2), int64(15), object(16)
memory usage: 188.8+ KB
```

Table 2: Data Sample

	Unnamed: 0	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	...	age
0	0	Data Scientist	$53K-91K (Glassdoor est.)	Data Scientist\nLocation: Albuquerque, NM\nEdu...	3.8	Tecolote Research\n3.8	Albuquerque, NM	Goleta, CA	501 to 1000 employees	1973	...	47
1	1	Healthcare Data Scientist	$63K-112K (Glassdoor est.)	What You Will Do:\n\nI. General Summary\n\nThe...	3.4	University of Maryland Medical System\n3.4	Linthicum, MD	Baltimore, MD	10000+ employees	1984	...	36
2	2	Data Scientist	$80K-90K (Glassdoor est.)	KnowBe4, Inc. is a high growth information sec...	4.8	KnowBe4\n4.8	Clearwater, FL	Clearwater, FL	501 to 1000 employees	2010	...	10
3	3	Data Scientist	$56K-97K (Glassdoor est.)	*Organization and Job ID**\nJob ID: 310709\n\n...	3.8	PNNL\n3.8	Richland, WA	Richland, WA	1001 to 5000 employees	1965	...	55
4	4	Data Scientist	$86K-143K (Glassdoor est.)	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions\n2.9	New York, NY	New York, NY	51 to 200 employees	1998	...	22

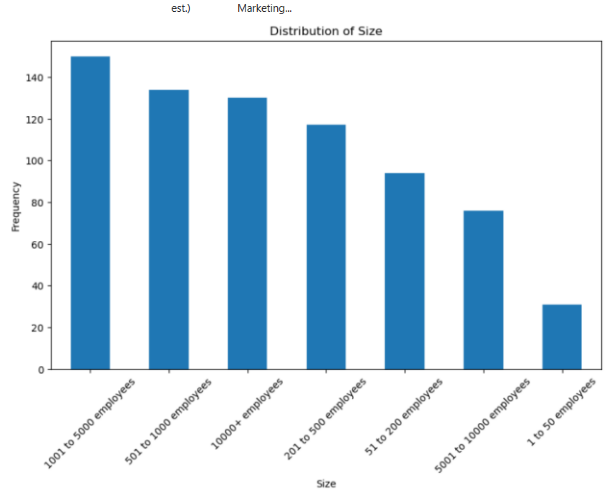


Figure 1: Size of Company Distribution

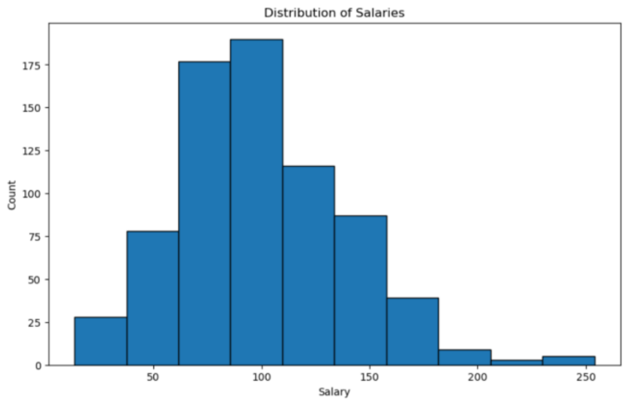


Figure 2: Distribution of Salaries

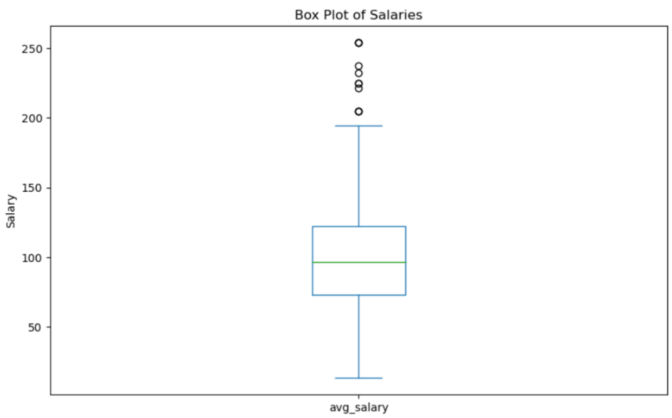


Figure 3: Box Plot of Salaries

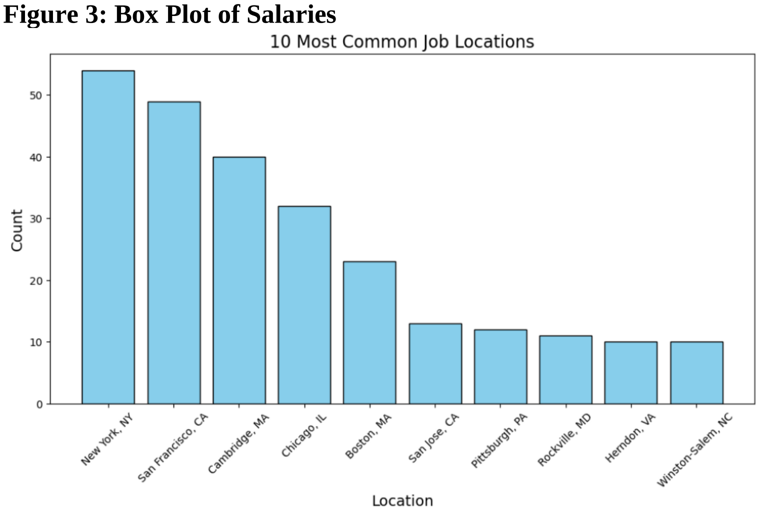


Figure 4: Top Job Locations

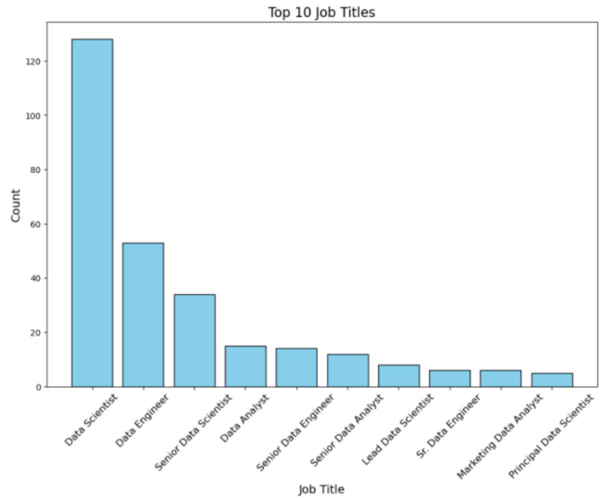


Figure 5: Top Job Titles

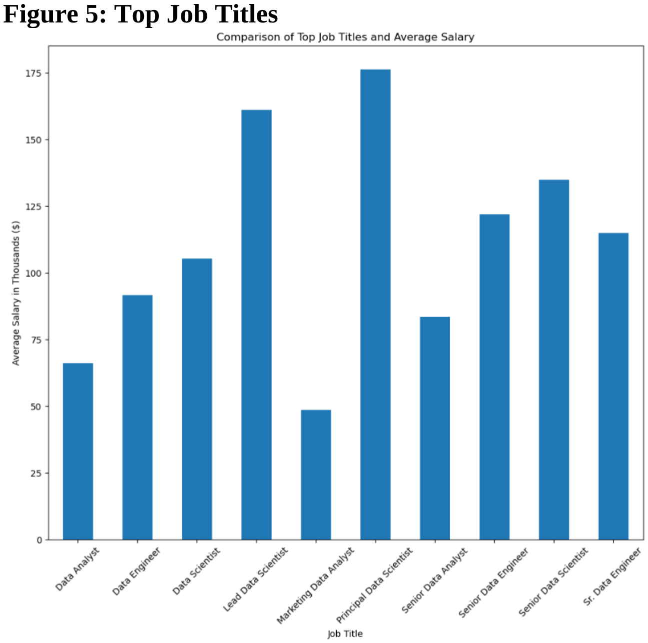


Figure 6: Job Titles and Average Salary

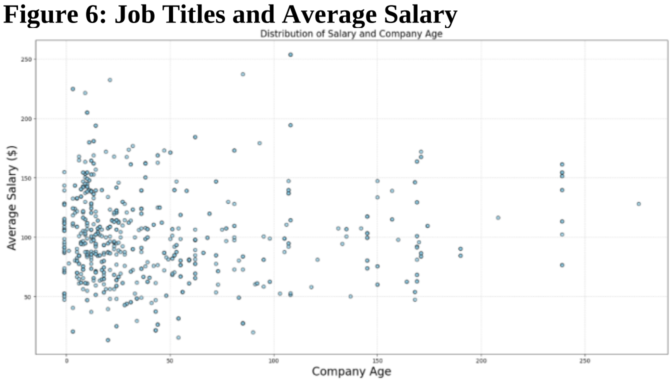


Figure 7: Scatterplot of Company Age and Average Salary

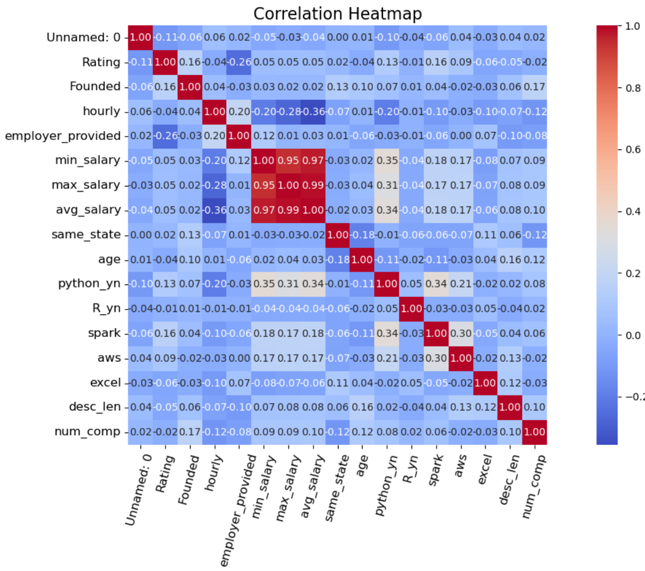


Figure 8: Correlation Heatmap of all Variables

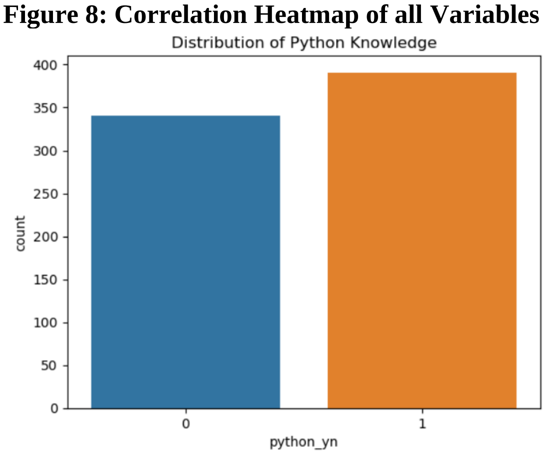


Figure 9: Python Knowledge

IV. METHODS

A. Data Preparation

To start off we got a sample of our data and realized no missing values were present, so we did not have to drop anything. We did have to fix a variable, but it was a simple task. After that we did EDA and gathered graphs, plots, and heatmap to give is and insight of the variables and our data that was present. The correlation heatmap gave us the greatest insight to see what variables directly correlated with each other. After we finished with our EDA, we then went into our experiment with logistic and linear regression which will be explained and shown below.

B. Experimental Design

```
# Select features and target variable
X = df[['Rating', 'min_salary','max_salary','Size', 'Founded', 'Industry', 'Sector', 'Revenue','seniority', 'Type of owne
y = df['avg_salary']

# Convert categorical variables to numeric using one-hot encoding
X = pd.get_dummies(X, drop_first=True)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create and train the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

mse = mean_squared_error(y_pred, y_test)
mae = mean_absolute_error(y_pred, y_test)
r2_score1=r2_score(y_pred, y_test)
print("Mean Squared Error", mse)
print("Mean Absolute Error", mae)
print("R2 Score", r2_score1)
```

Listed above is our experimental design on Python for our train and test split for the Linear Regression prediction model.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# Select features and target variable
X = df[['Rating', 'min_salary', 'max_salary', 'Size', 'Founded', 'Industry',
        'Sector', 'Revenue', 'seniority', 'Type of ownership', 'age', 'job_simp']]
y = df['python_yn']

# Convert categorical variables to numeric using one-hot encoding
X = pd.get_dummies(X, drop_first=True)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create and train the logistic regression model
model = LogisticRegression(max_iter=1000)  # Increase max_iter if necessary
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)

# Print evaluation metrics
print(f"Accuracy: {accuracy:.2f}")
print("Confusion Matrix:")
print(conf_matrix)
print("Classification Report:")
print(class_report)
```

Listed above is the experimental design we used using Python for our train and test split for the Logistic Regression prediction model.

C. *Tools Used*

The following tools were used for this analysis: Python v3.11.11 running the Anaconda 22.9.0 environment for HP ENVY Laptop computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 2.2.3, NumPy 2.1.0, Matplotlib 3.9.0, Seaborn 0.13.0, SKLearn 1.3.0, and Patsy 1.0.1. These were the tools we used because they are what we had and have been using for the whole semester. We also used these because it is what we are most familiar with and gave us the most accurate results and allowed us to find our solutions.

V. RESULTS

A. *Classification Measures/ Accuracy measure*

```
Mean Squared Error 8.320104034776497
Mean Absolute Error 1.8164291576070575
R2 Score 0.9950812025846161
```

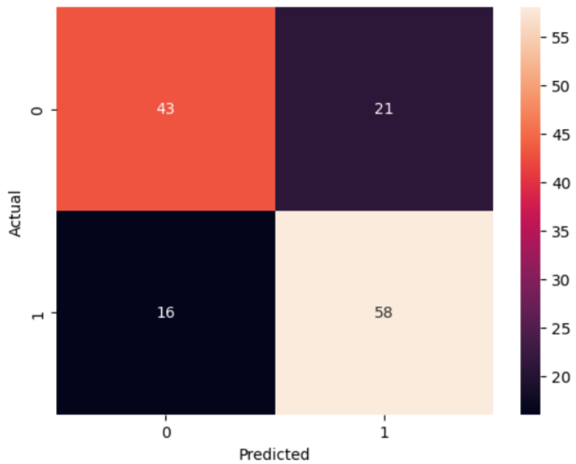
Results from Linear Regression Model

```
Accuracy: 0.73
Confusion Matrix:
[[43 21]
 [16 58]]
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.67      0.70        64
           1       0.73      0.78      0.76        74

    accuracy                           0.73       138
   macro avg       0.73      0.73      0.73       138
weighted avg       0.73      0.73      0.73       138
```

Results from Logistic Regression

0.7318840579710145

Accuracy Score



Confusion Matrix Model

B. Discussion of Results

After going through our results, we got clear information regarding the linear and logistic regression tests we ran. When looking at the linear regression, a mean square error of 8.32 is low which emphasizes that the model's results are similar or close to the actual results. With an R Square of 0.995 it is high, which is a good sign which means the model is fitting the data in a correct way. When looking at Logistic, we got an accurate score of approximately 73% which suggests that majority of the time the model correctly predicted the class of 0 or 1. An F1-Score of 0.70 and .76 indicates a decent balance oof the precision and recall of 0 and 1.

C. Problems Encountered

When looking at problems we encountered the first thing was we had invalid entries in the company size column and had to fix the entries so that our data was not messed up and allowed us to have accurate results. Overall, this was a simple task to fix but was a setback. Another problem we had was a problematic entry for a company which was like the first problem we had. The last problem we encountered was output errors when running the linear regression and we had to redo the code multiple times.

D. Limitations of Implementation

The main limitation to the dataset and the model is how the model predicts a lot of false positives. Our dataset is also limited due to the sample size and the number of variables available for testing. If the dataset had more variety and more data would allow for the model to have better and more accurate results. Between the logistic and linear regression models the linear had better and more accurate results and gave better insight than the logistic regression.

E. Improvements/Future Work

Areas to improve and to allow for better use for future work are things like resampling techniques for over sampling for more results. As well as adding more variables and collecting more data for more validation from both models. As well as managing outliers and evaluating more data, all these things will allow for better results for future work and improvements for the models.

VI. CONCLUSION

After going through this whole project, it was a good project, and it allowed us to learn a lot and use our skills. We were able to gain insight into EDA and logistic and linear regression and how the models work. After getting a data set, we were unfamiliar with and developing some EDA and training and evaluating our data we were able to develop a conclusion. Overall, we can conclude that for the most part and majority of the time we are able to predict a Data Scientist Salary. Using the variables provided from the data as well as using the models we can also see that there were many variables that are related to the salary. The model was good, but it most defiantly can still use some work and improvements. Lastly, this was a great project that allowed us to learn more, and it allowed us to look at variables and use models to answer our question. We were able to with a good model that yes, we can predict a data scientist salary.

REFERENCES

https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor