Data Analyst Job Postings [Pay, Skills, Benefits]

About Dataset

Data Set

The DataSet we used in this Project contains 27 Columns, and 29,738 Rows.

Agenda

To find the Relationship between Programming Languages, Locations, Companies, Salaries, and other Columns, then using Machine Learning model such as Linear Regression, Polynomial Regression and Random Forest Regression for understanding the key findings and insights, portraying the results at each step.

1

Importing various Libraries

2

Data Cleaning

3

Statistical Analysis

4

Exploratory Data Analysis

5

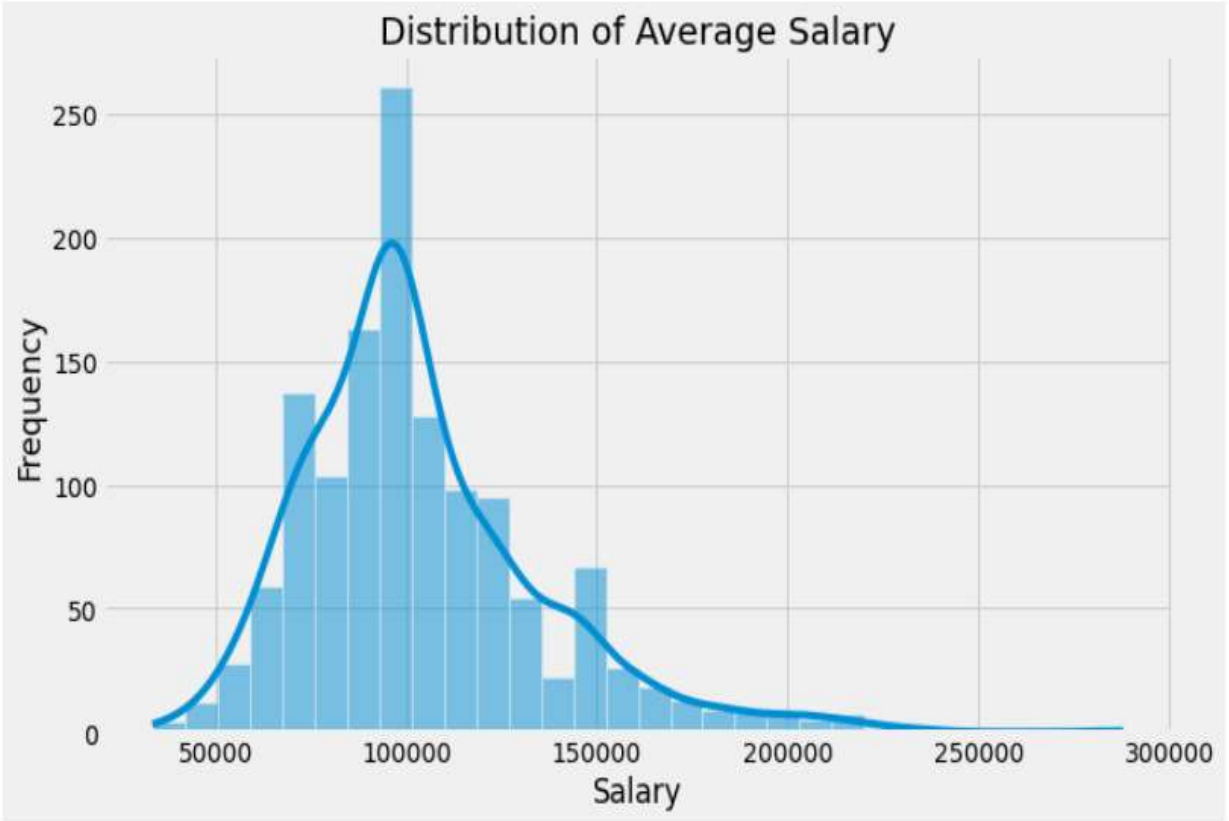
Data Visualizations

6

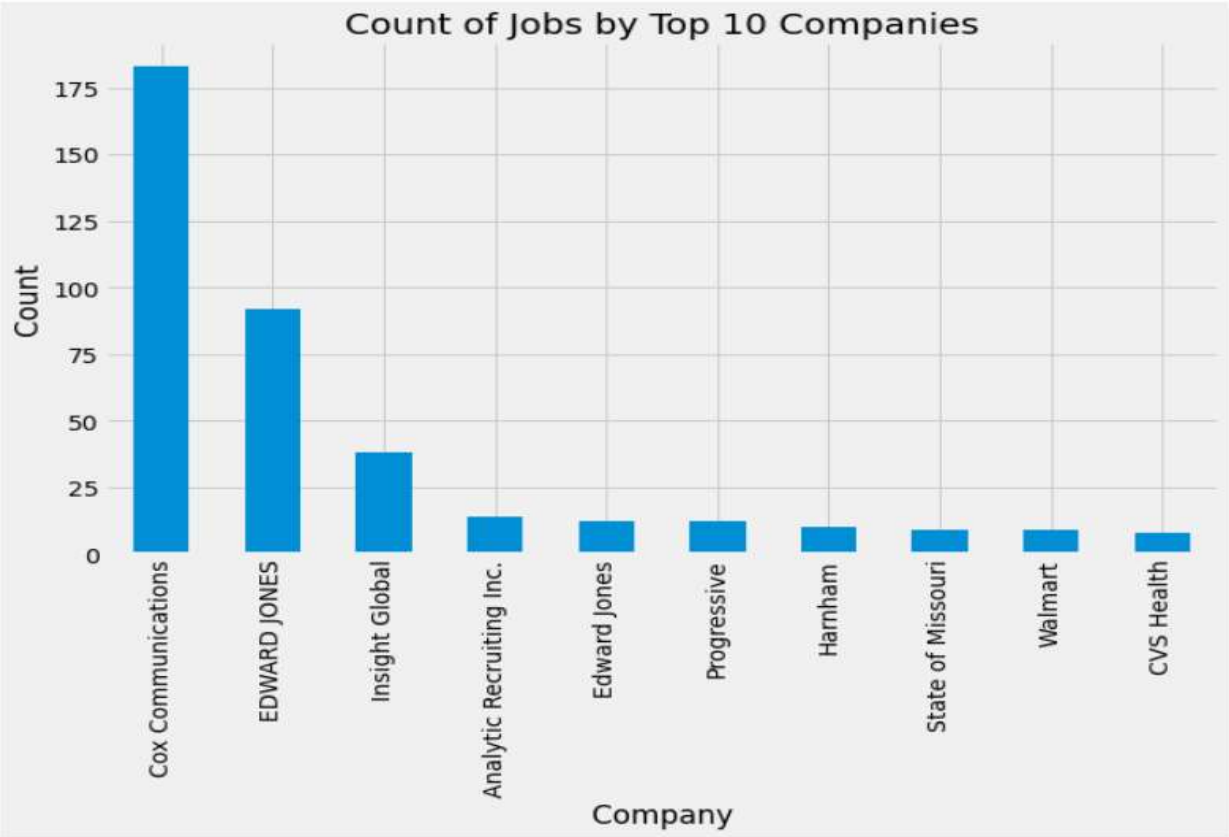
Machine Learning Model [Linear, Polynomial Regression, and Random Forest Regression

Exploratory Data Analysis (EDA)

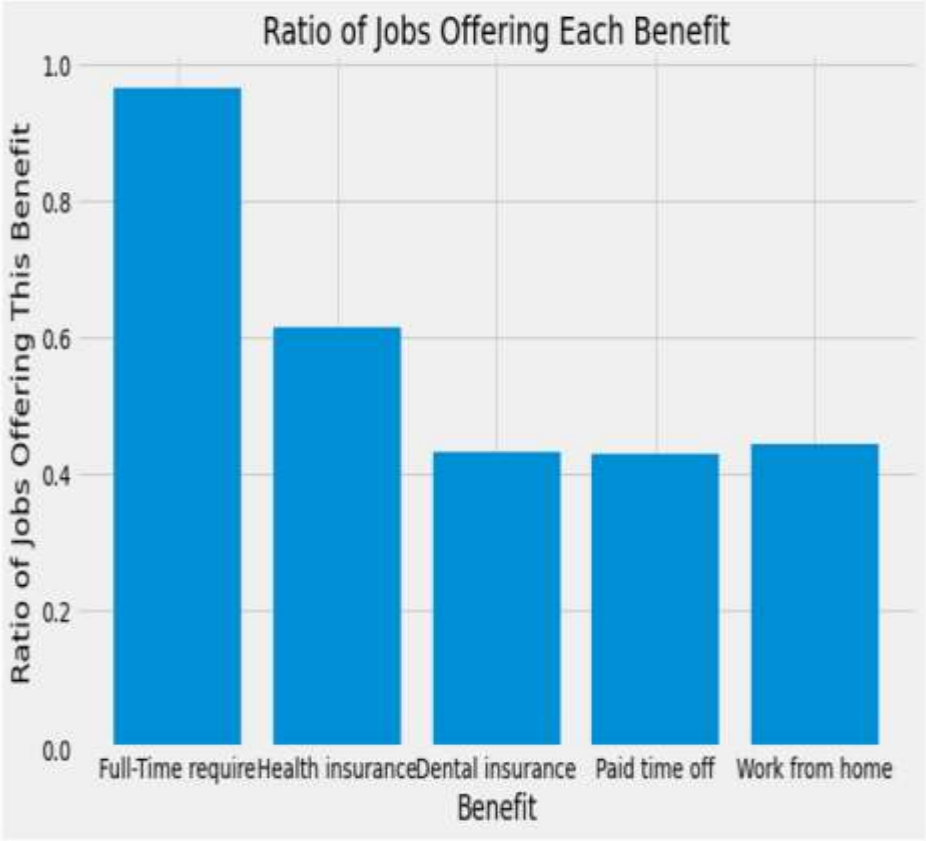
- Descriptive analysis of the data.
- Several factors that may affect salaries: companies, locations, languages, and benefit offerings such as:
- 1] Correlation between company size and salary.
- 2] Correlation between location and salary
- 3] Correlation between Language and salary

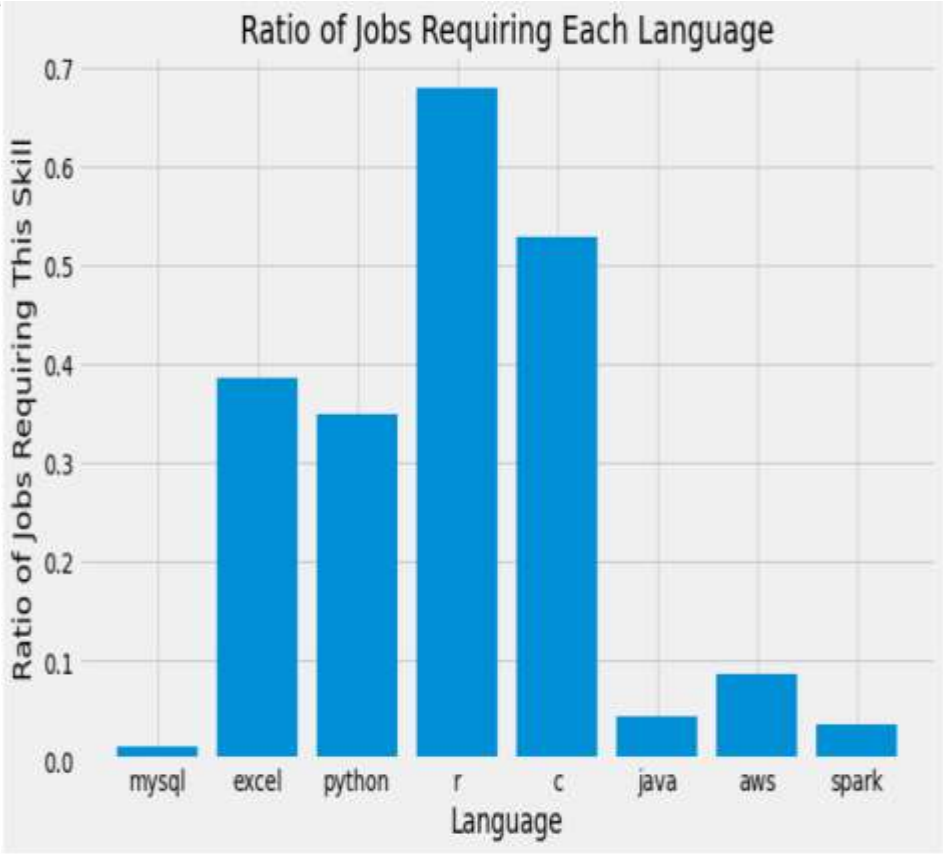


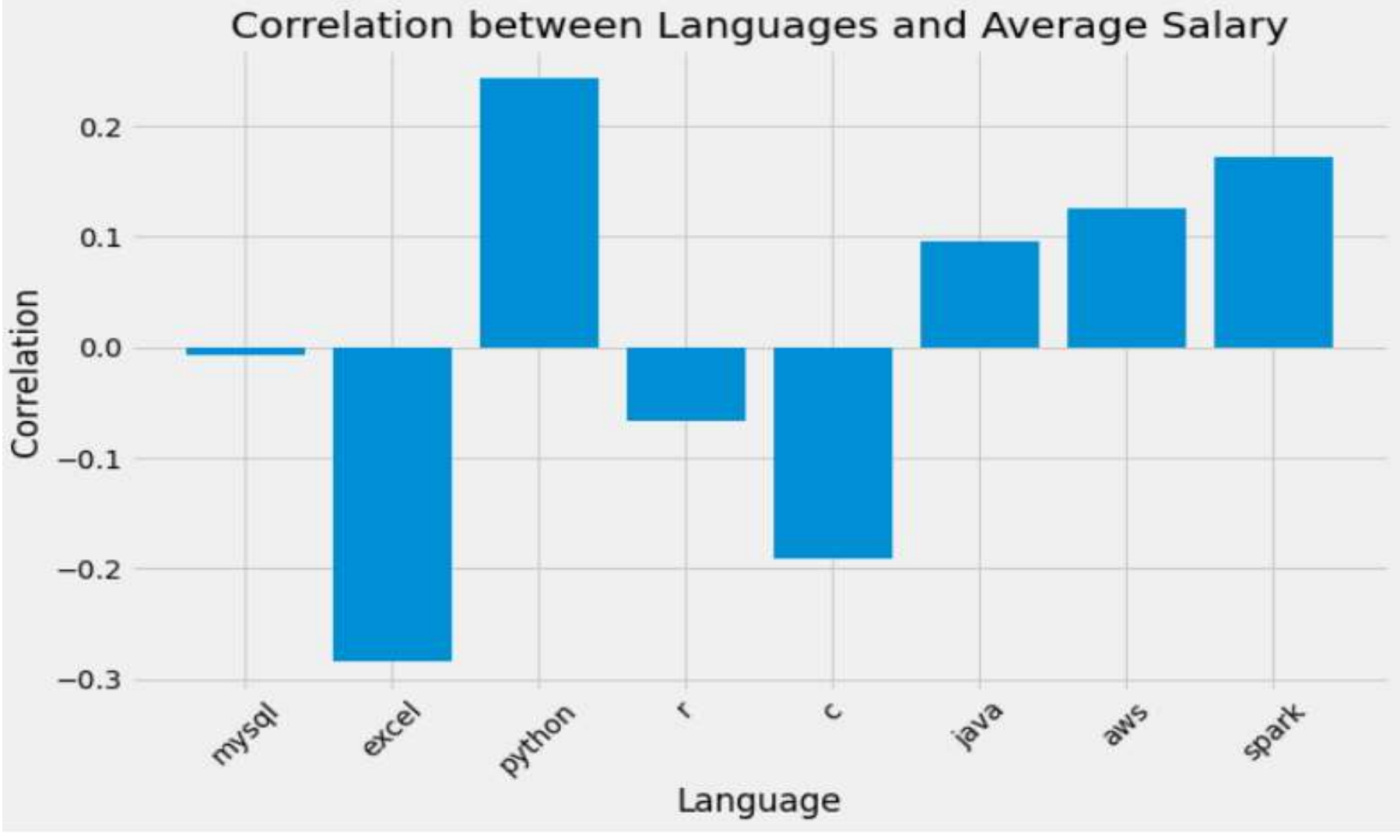
Data Visualizations

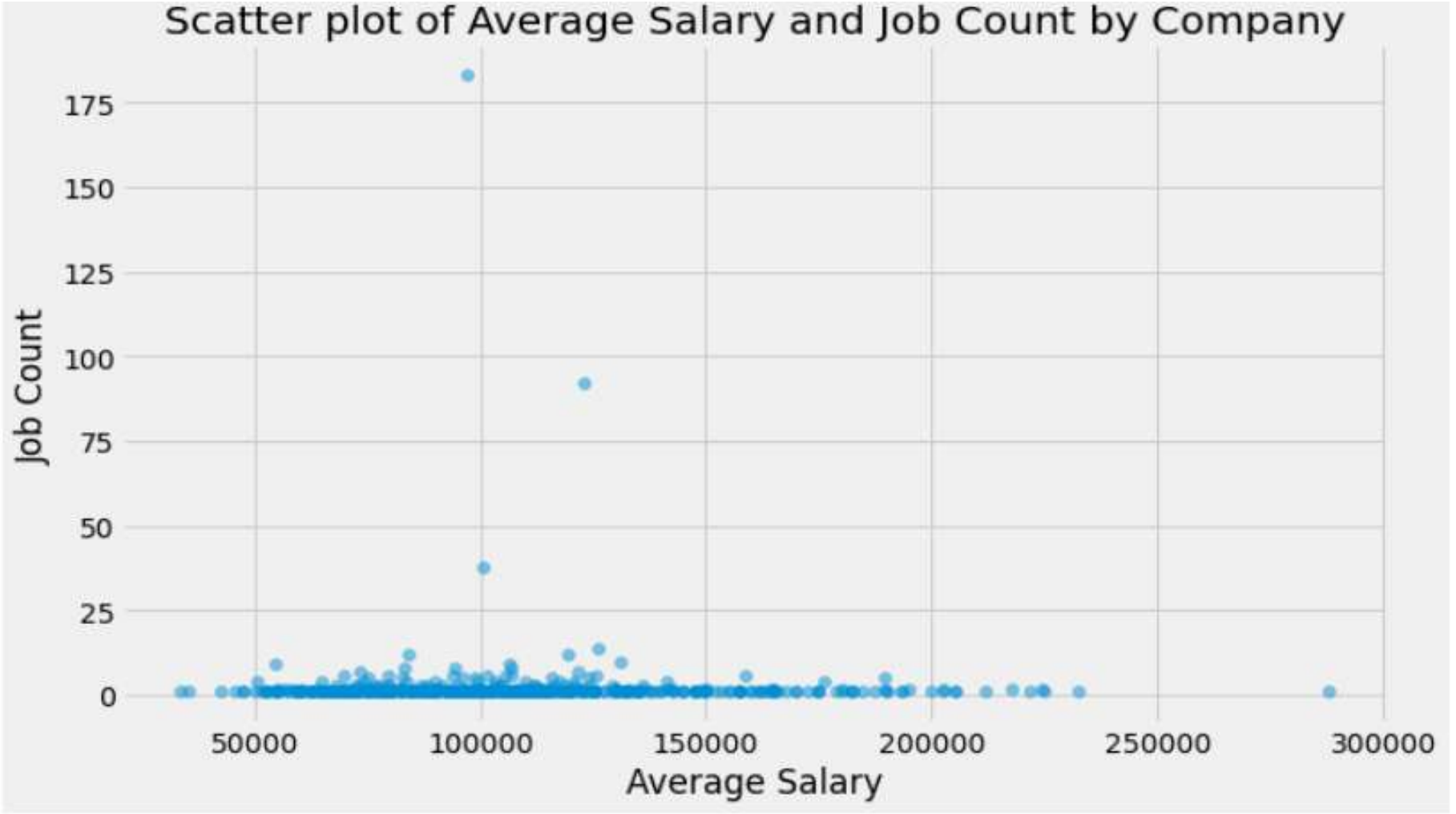


This Visualization Shows the number of Data Jobs frequency within Top 10 Companies.

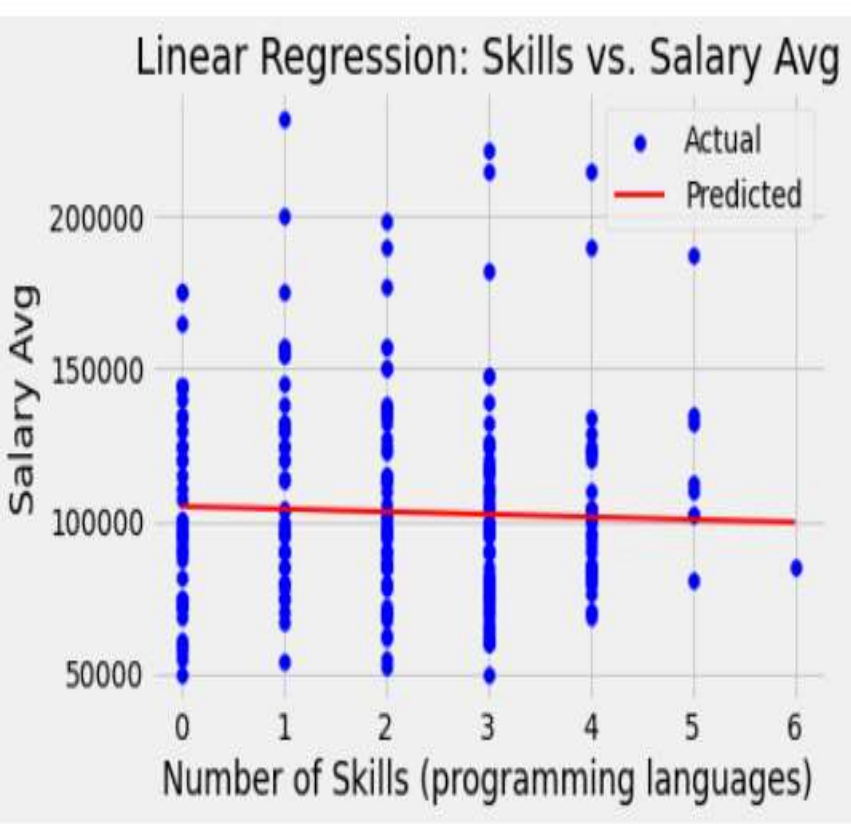

Ratio of Jobs Offering Each Benefit
Ratio of Jobs Offering This Benefit
1.0
0.8
0.6
0.4
0.2
0.0
Full-Time require
Health insurance
Dental insurance
Paid time off
Work from home
Benefit

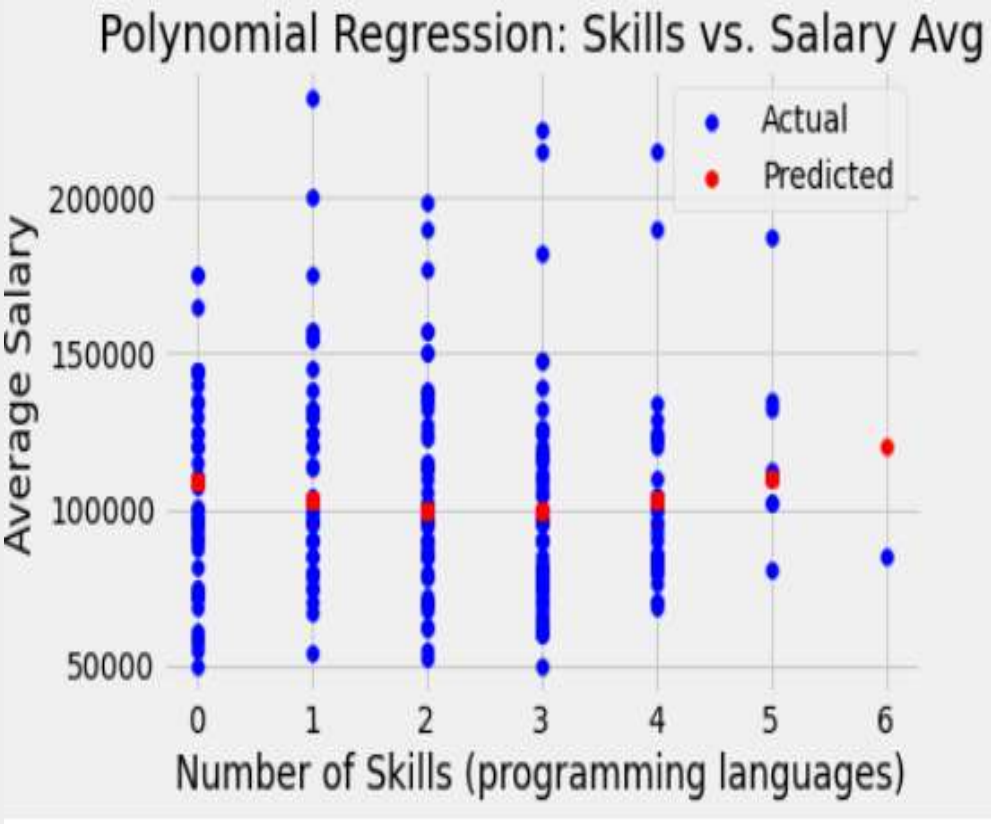

Ratio of Jobs Requiring Each Language
Ratio of Jobs Requiring This Skill
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
mysql
excel
python
r
c
java
aws
spark
Language


Correlation between Languages and Average Salary
0.2
0.1
0.0
−0.1
−0.2
−0.3
Correlation
mysql
excel
python
r
c
java
aws
spark
Language


Scatter plot of Average Salary and Job Count by Company
Job Count
175
150
125
100
75
50
25
0
50000
100000
150000
200000
250000
300000
Average Salary

Evaluating the Machine Learning Model





Predictions: [Using Linear Regression]

After performing the analysis we want to complete a prediction program that is able to predict the possible salary of a position based on the input characteristic variables.

We will first build the model using some of the variables; here and then bring in all of them to see what the difference is. Here is a program to make predictions on a person's salary based on the programming languages he knows, such as 'mysql', 'excel', 'python', ‘r’, ‘c’, ‘java’, ‘aws’, ‘spark’. The input variable is an array with 8 binary entries (0 or 1) each corresponds to whether he knows the corresponding language or not, to assess the average salary for the mentioned skills.

```
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.linear_model import LinearRegression

X = df[['mysql', 'excel', 'python', "r", "c", "java", "aws", "spark"]]
y = df['salary_avg']

# Perform one-hot encoding for the skills columns
encoder = ColumnTransformer(transformers=[('encoder', OneHotEncoder(), [0, 1, 2, 3, 4, 5, 6, 7])], remainder='passthrough')
X_encoded = encoder.fit_transform(X)

model = LinearRegression()
model.fit(X_encoded, y)

# Make predictions on new data
new_skills = [[1, 0, 1, 0, 0, 1, 1, 1]]  # Example new skills to predict salary for (mysql: 1, excel: 0, python: 1, etc.)
new_skills_encoded = encoder.transform(new_skills)
predictions = model.predict(new_skills_encoded)

print("Your estimated salary is:", predictions)
```

```
Your estimated salary is: [144806.35158705]
```

Conclusion

The variables we studied included the company rating, location (city), benefits such as vacation or insurance availability, and required skills. The primary hypothesis was that the number of coding skills mastered by an individual significantly predicts annual salary. This hypothesis stemmed from the belief that having multiple programming skills can demonstrate a wider range of abilities, making a candidate more attractive to employers and improving their bargaining power in salary negotiations.

Our analyses involved examining the correlations between individual factors and salaries, followed by the development of predictive models. We found that the size of a company does not significantly correlate with wages. Location, on the other hand, was found to have a statistically significant impact on average wages, but the explanatory power was relatively low. Among the programming languages, .Python showed a positive correlation with salary, while Excel and C showed negative correlations. Given these findings, we can reject the null hypothesis that the number of coding skills mastered, the location of work, and the company one works for have no significant impact on their annual salary.. Instead, our data supports the alternative hypothesis, indicating that specific coding skills, location, and certain employee benefits do impact the salary of a data analyst.

.However, it is worth noting that the number of coding skills mastered did not show a strong correlation with salary levels., suggesting that the types of skills may be more important than the quantity. Finally, there were several limitations to our project. First, our dataset might not be comprehensive enough to capture all the factors that affect salaries. Second, the linear regression models we used might not be able to effectively capture the complex, possibly non-linear relationships between the variables and salaries. Despite these limitations, this work contributes to a better understanding of the factors influencing data analysts' salaries, which can inform individuals and organizations about what to prioritize in career development and recruitment processes.