Vivek A

Click here to view the github

Salary Visualization &Prediction

Data Science project

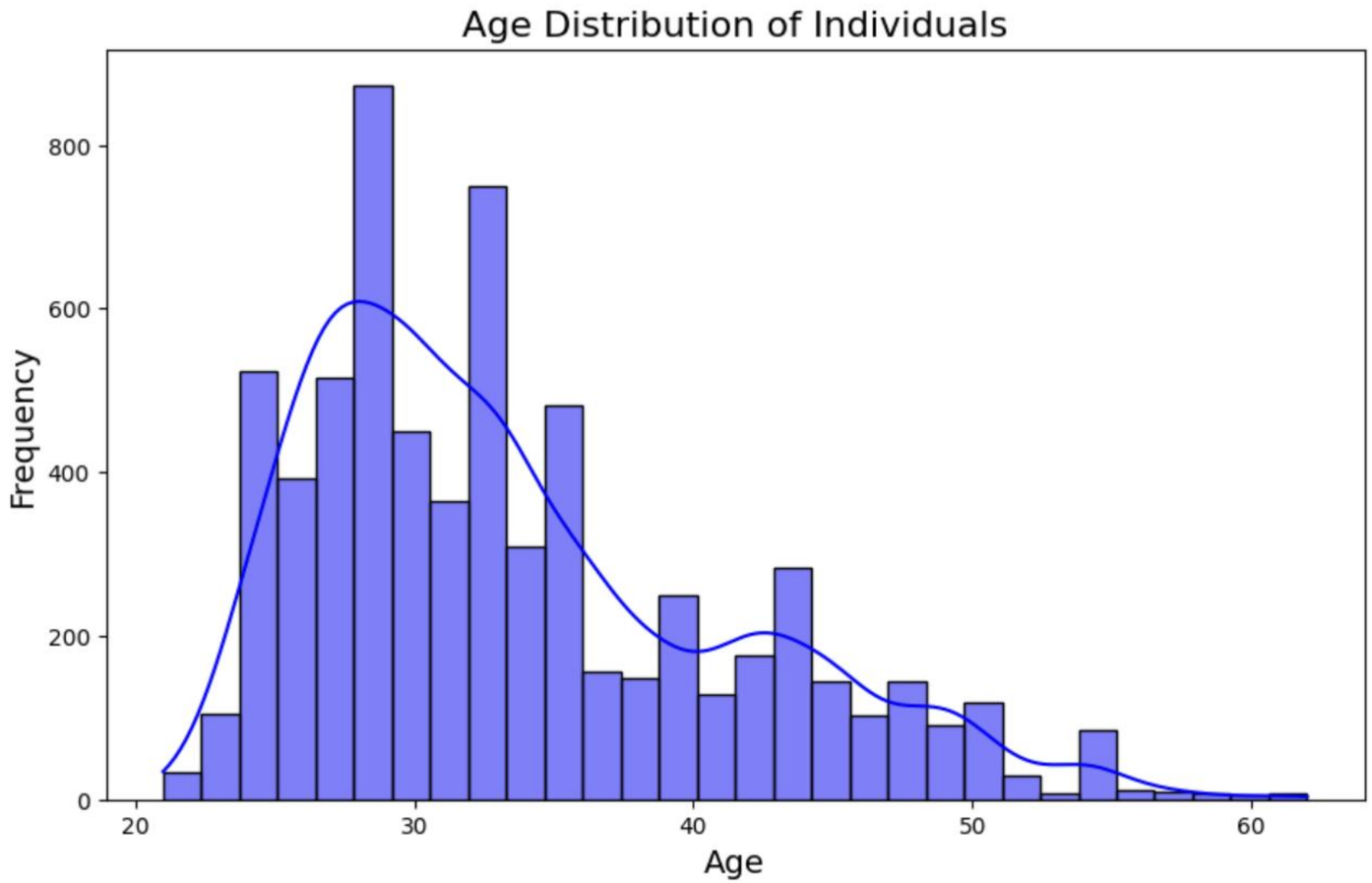
Project Objective

The primary objective of this project is to analyze and predict salary data using a variety of data science techniques. The goal is to provide a comprehensive analysis of the salary dataset, build predictive models to estimate salaries, and present the findings in a clear and actionable manner. This will help in understanding salary determinants and making informed decisions based on data-driven insights.

Insights from EDA

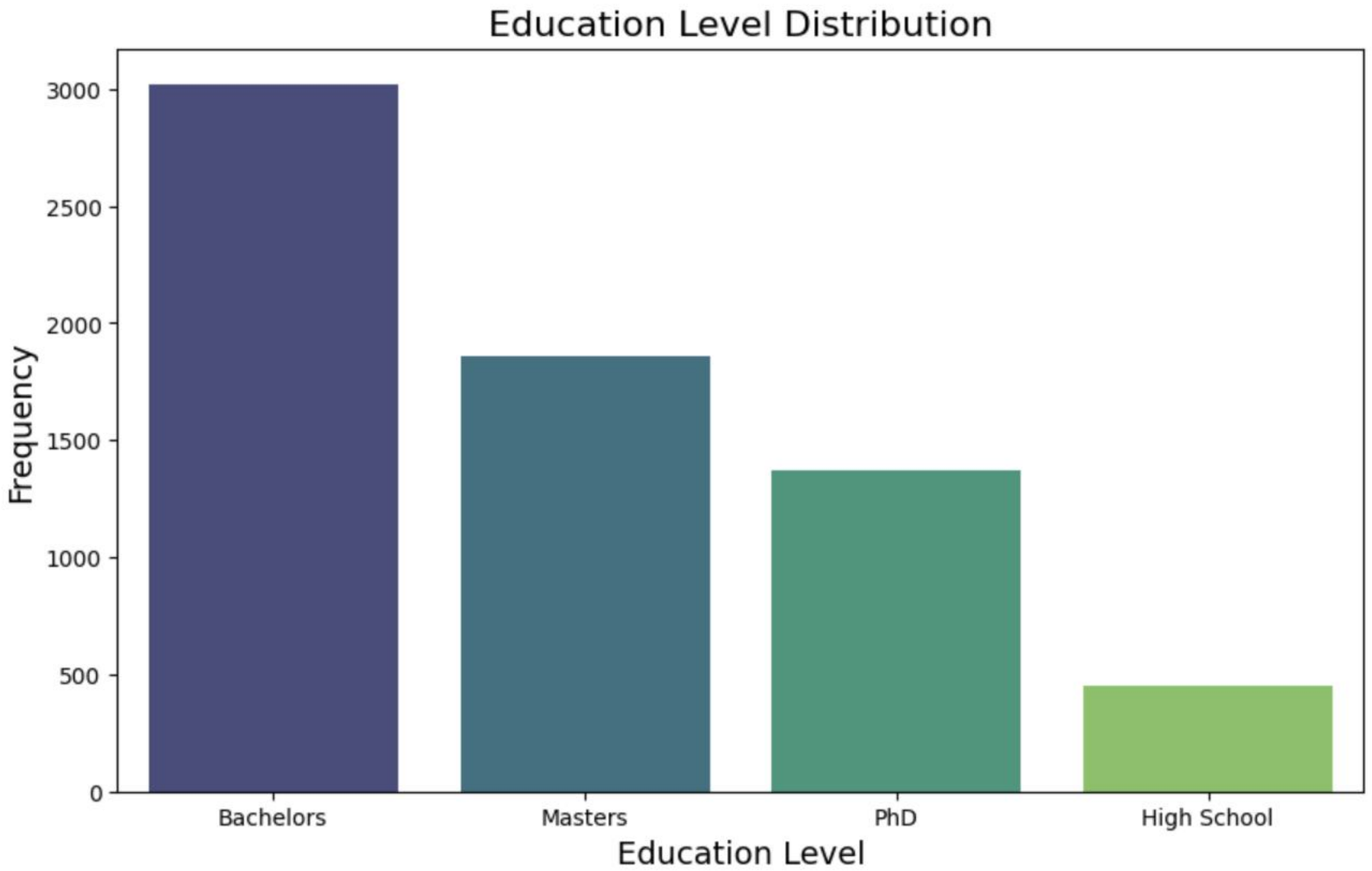
Age distribution

- Majority of employees are 25-35. Means the employees are energetic and can be suitable for long run



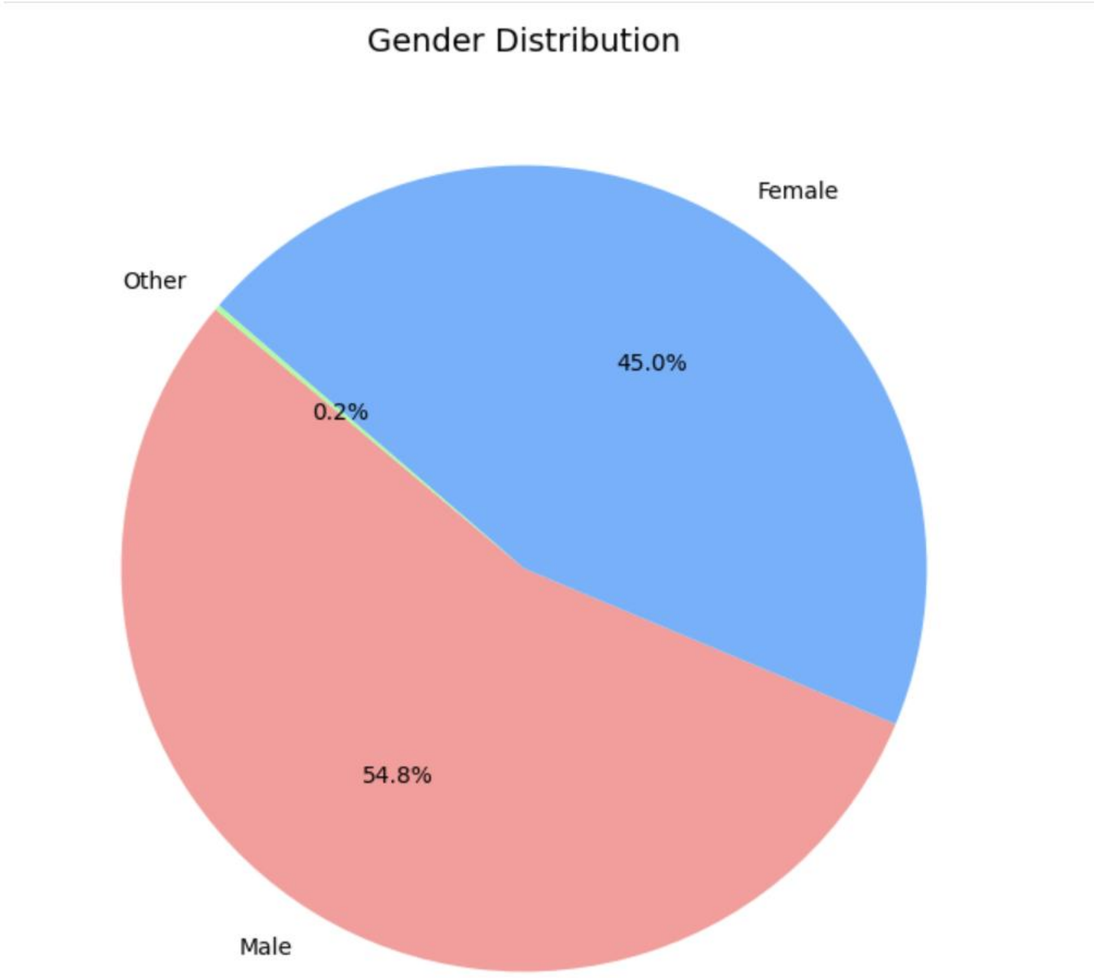
Education Level Distribution

- Most of the employees have bachelors degree in the company. High school is the lowest



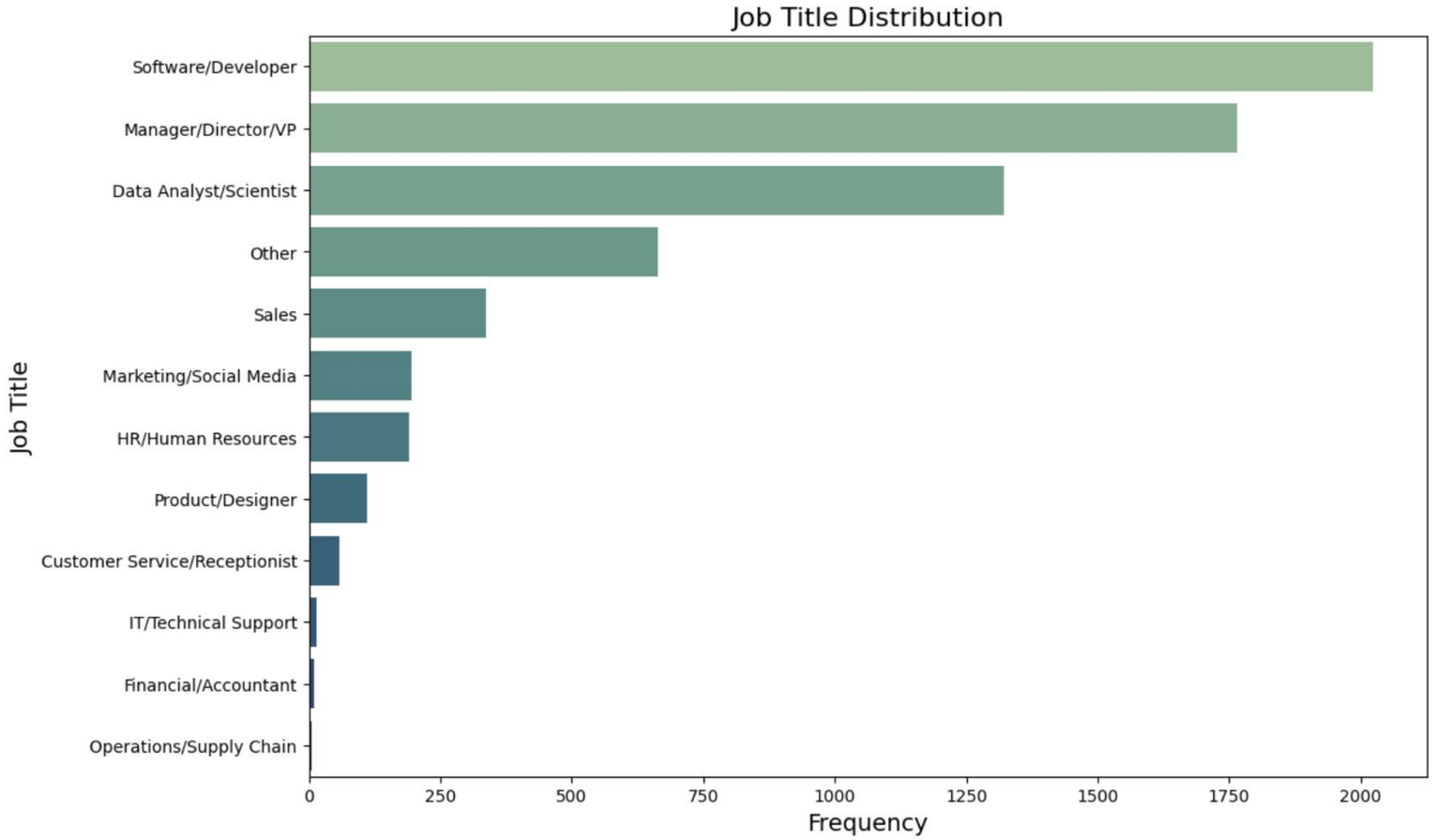
Gender Distribution

- Majority of the employees are males with 54 percent.



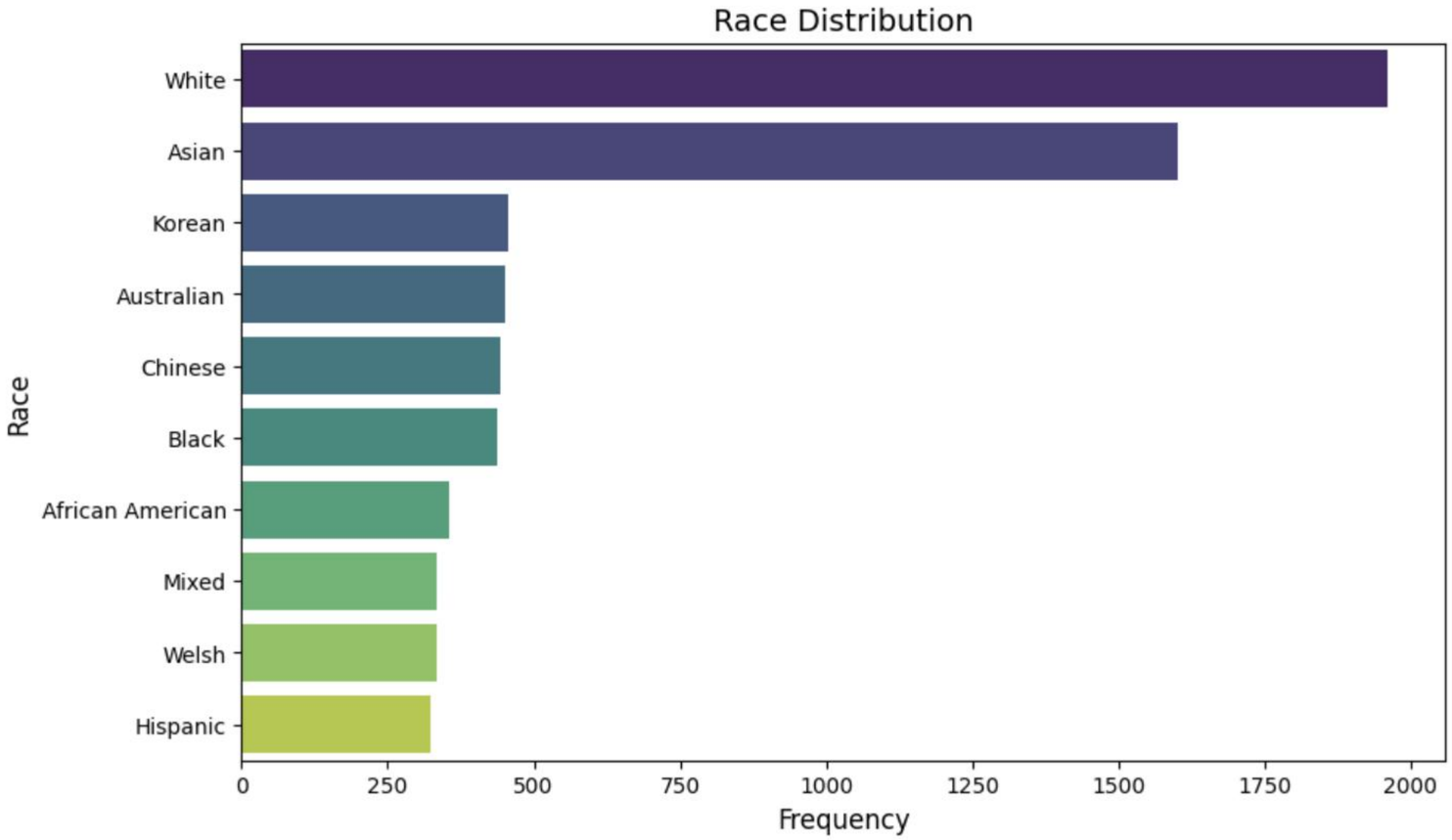
Job Title Distribution

- Based on this observation, one could hypothesize that roles such as Software Developer, Data Analyst/Scientist, and Manager/ Director are in higher demand compared to other job titles. This also suggests that positions in Finance/Accounting, Operations/ Supply Chain Management, and Customer Service are less in demand and may offer lower compensation.



Race Distribution

- This graph help us to know about the racial distribution in the dataset. From the graph, it is clear that most of the employees are either White or Asian, followed by Korean, Chinese, Australian and Black.

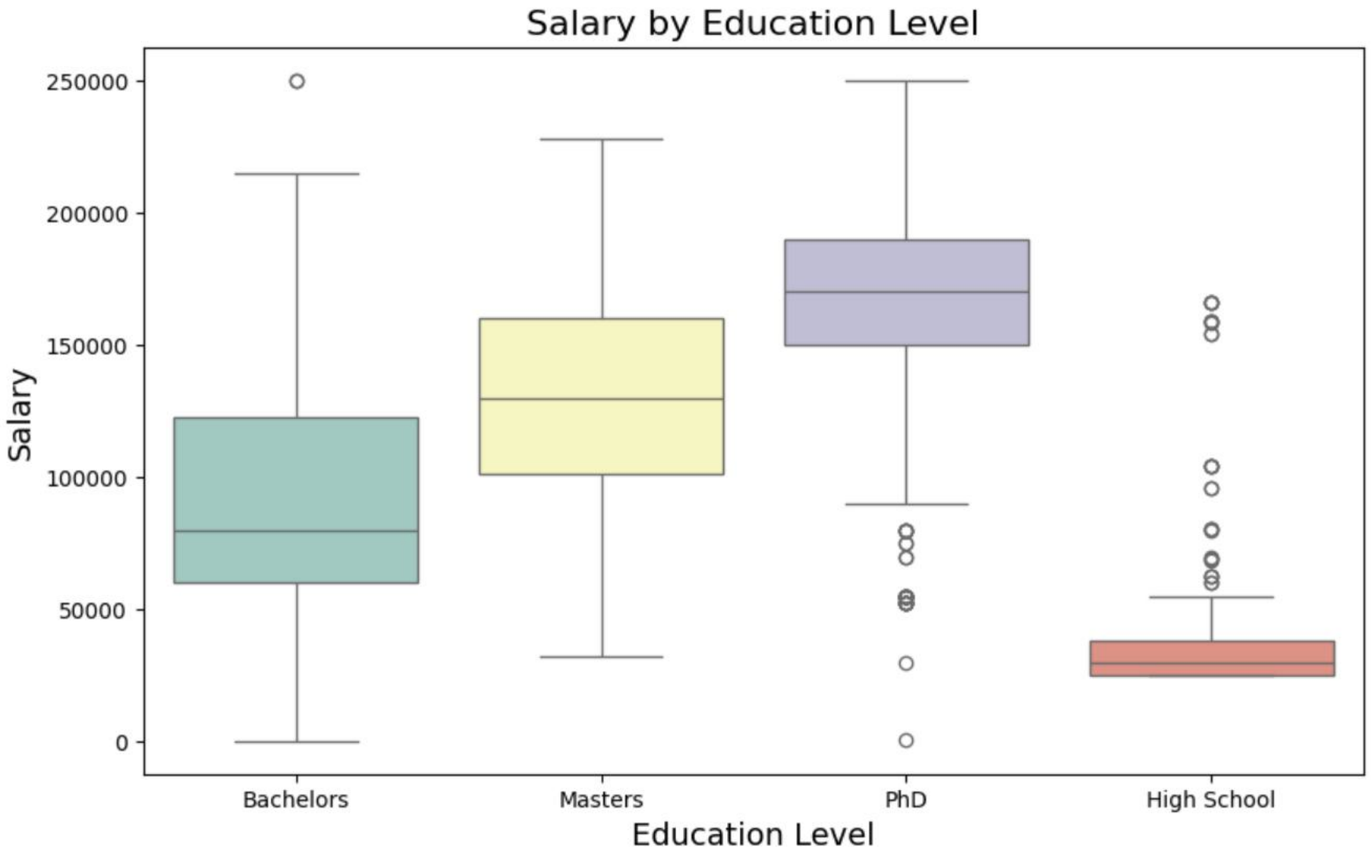


Salary by Gender

- In the boxplot the employees from Other gender has quite high salary as compared to Makes and Females.
- median salary of others is quite high maybe because of the less data availble for that section

Salary by Gender

- PhD holders have the highest median salary, followed by those with Master's and Bachelor's degrees, while employees without a degree have the lowest median salary. Bachelor's degree holders have salaries centered around 50,000 $$, while employees without a degree have salaries mainly between 40,000 and $45,000.

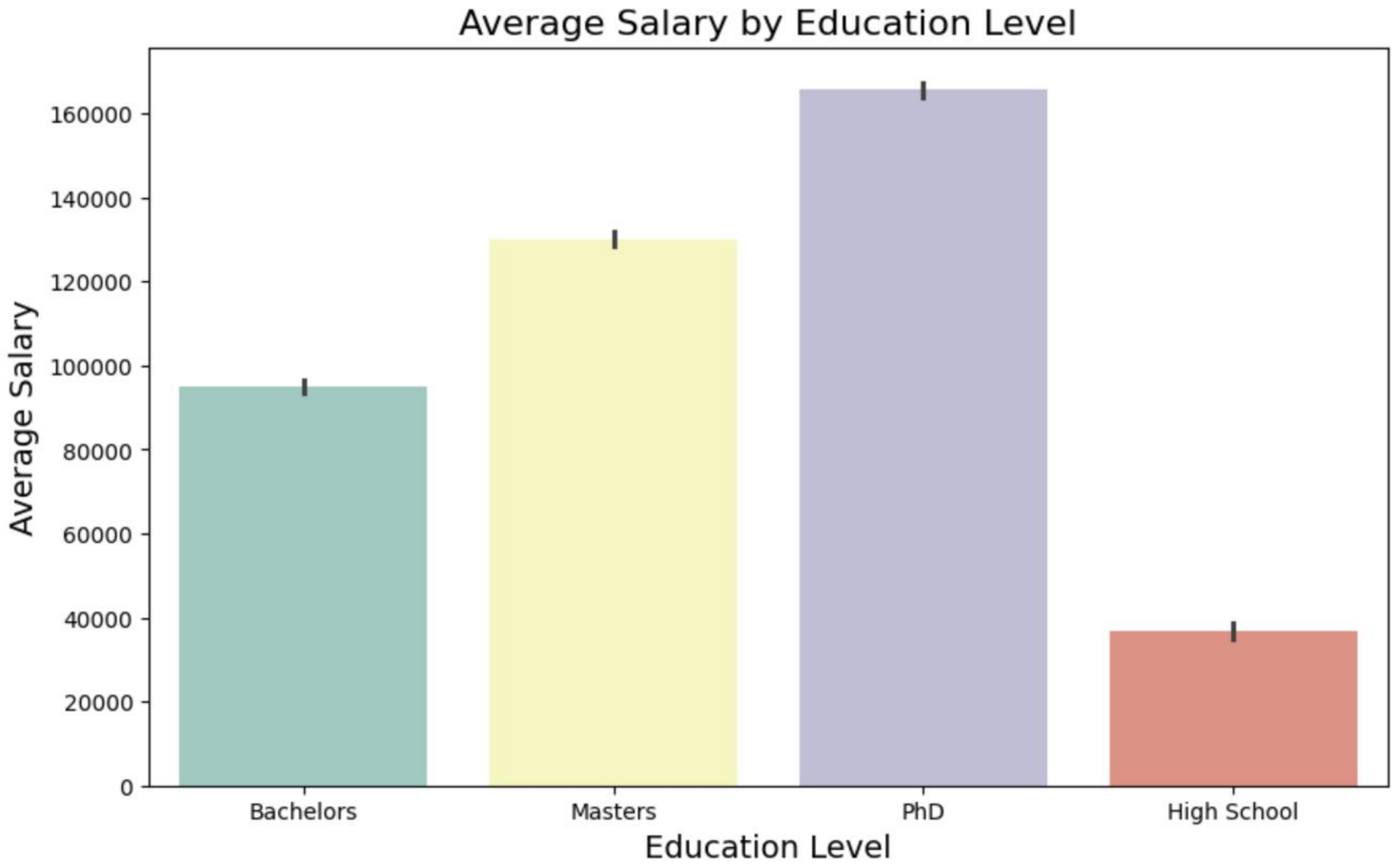


Salary by Experience

- From this scaaterplot, it is clear that on the whole, the salary of the employees is increasing with the years of experience. However, on closer look we can see that similar experience have different salaries. This is because the salary is also dependent on other factors like job title, age, gender education level as discussed earlier.

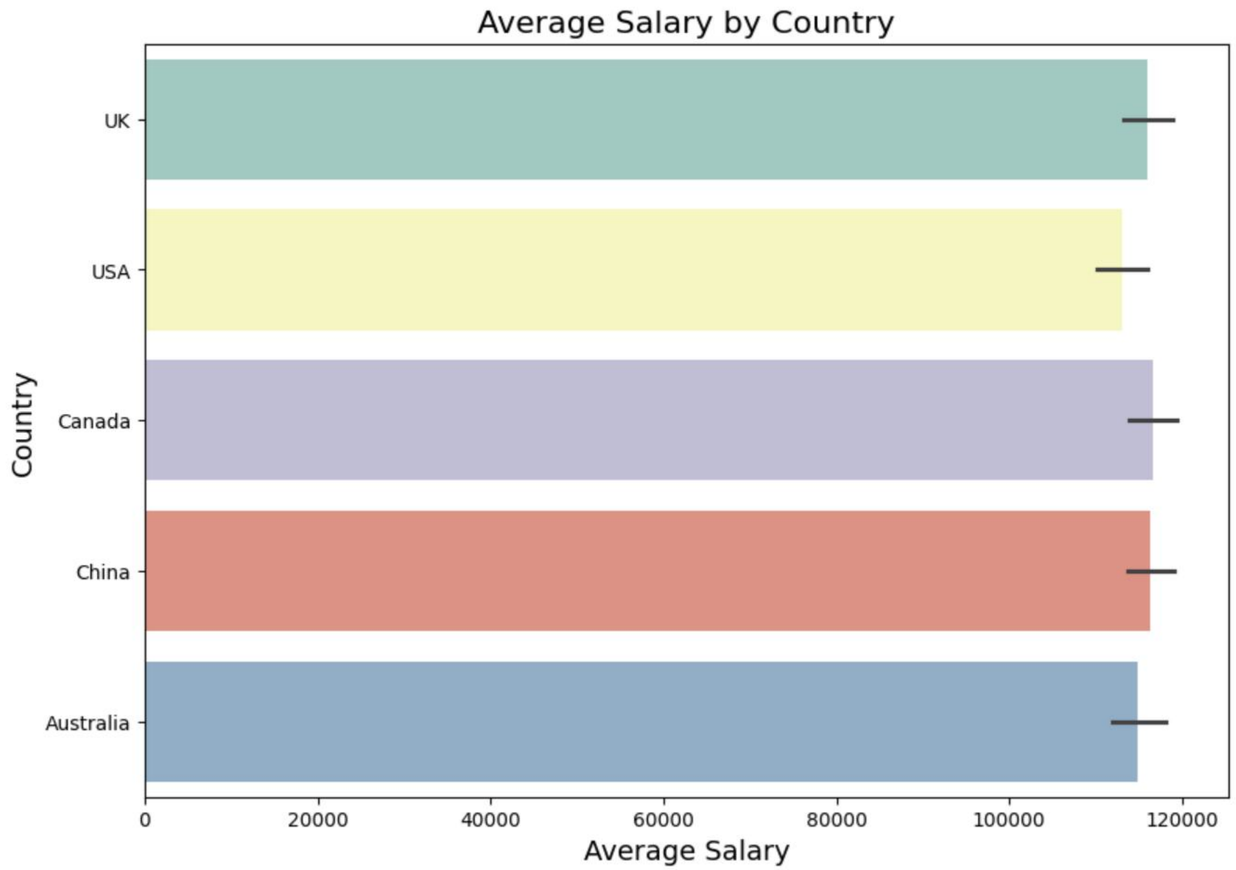
Average Salary by Education level

- It is not a surprise that the phd level of education boost the salary level of employees. Higher the education, Higher the salary



Average Salary by Country

- The distribution of salary is almost same for all of the employees from different country



Predictive Models

Linear Regression

- Linear regression model has 69 % accuracy

```
y_pred = model.predict(X_test)
```

```
# Evaluating the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse:.2f}")
print(f"R^2 Score: {r2:.2f}")
```

```
Mean Squared Error: 892338766.37
R^2 Score: 0.69
```

Decision Tree

- Decision Tree model has 66 % accuracy

```
# Fit Grid Search
grid_search.fit(X_train, y_train)

# Best parameters and best score
print(f"Best Parameters: {grid_search.best_params_}")
print(f"Best Accuracy: {grid_search.best_score_:.2f}")
```

```
Best Parameters: {'max_depth': None, 'max_features': None,
Best Accuracy: 0.66
```

Random Forest

- Random Forest has 94.65 % accuracy

```
# Evaluate the model
mse_rf = mean_squared_error(y_test, y_pred_rf)
r2_rf = r2_score(y_test, y_pred_rf)

print(f"Random Forest MSE: {mse_rf}")
print(f"Random Forest R² Score: {r2_rf}")
```

```
Random Forest MSE: 152328491.86770138
Random Forest R² Score: 0.9465787589477788
```

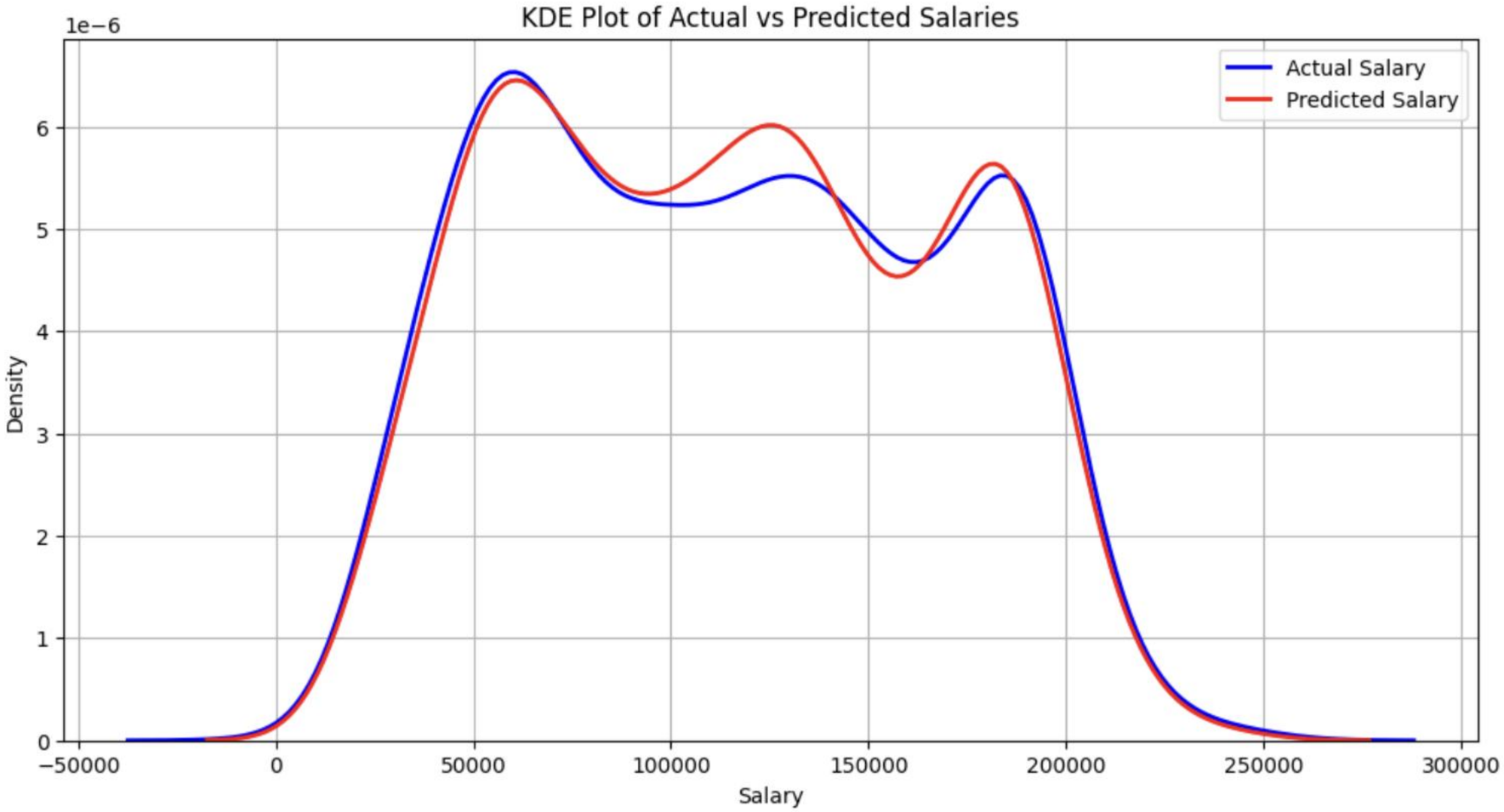
Actual vs predicted values

- Random Forest shown higher accuracy among all 3 models with 94 percent accuracy

	Actual Salary	Predicted Salary
1881	150000.0	149243.645833
2627	75969.0	77403.760000
496	100000.0	100000.000000
5968	60000.0	60000.000000
4104	80000.0	81020.833333
1039	190000.0	194670.833333
4844	72000.0	72654.000000
6611	55000.0	56630.238095
1084	195000.0	196402.297619
5525	130000.0	131661.369048

- Random Forest shown higher accuracy among all 3 models with 94 percent accuracy

Actual vs predicted values



Thanks

Thank you everyone who been with me with this journey.
Check the github repository for more info about the project