Vivek A

Click here to view the github

Medical Charges prediction

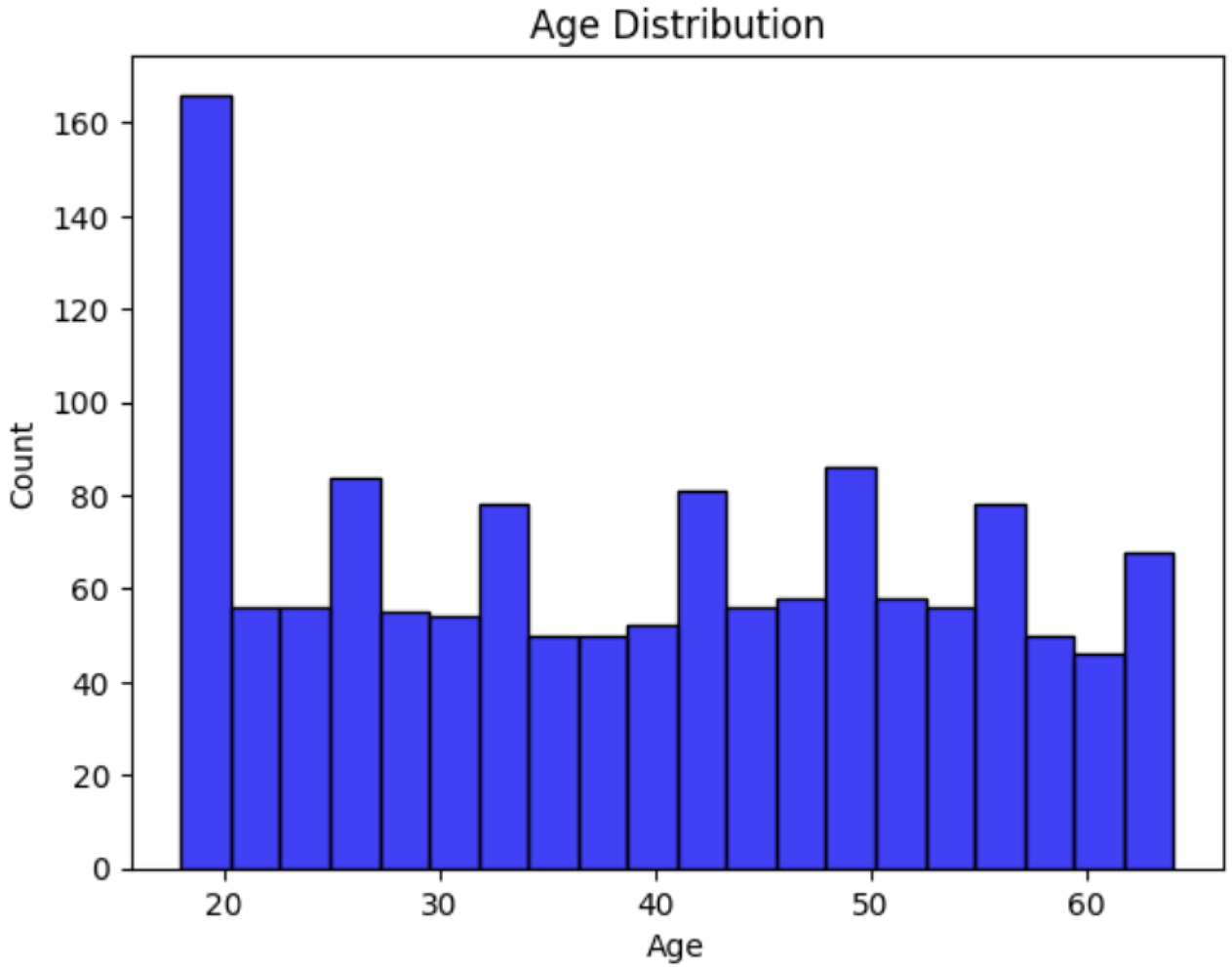
Data Science project

Project Objective

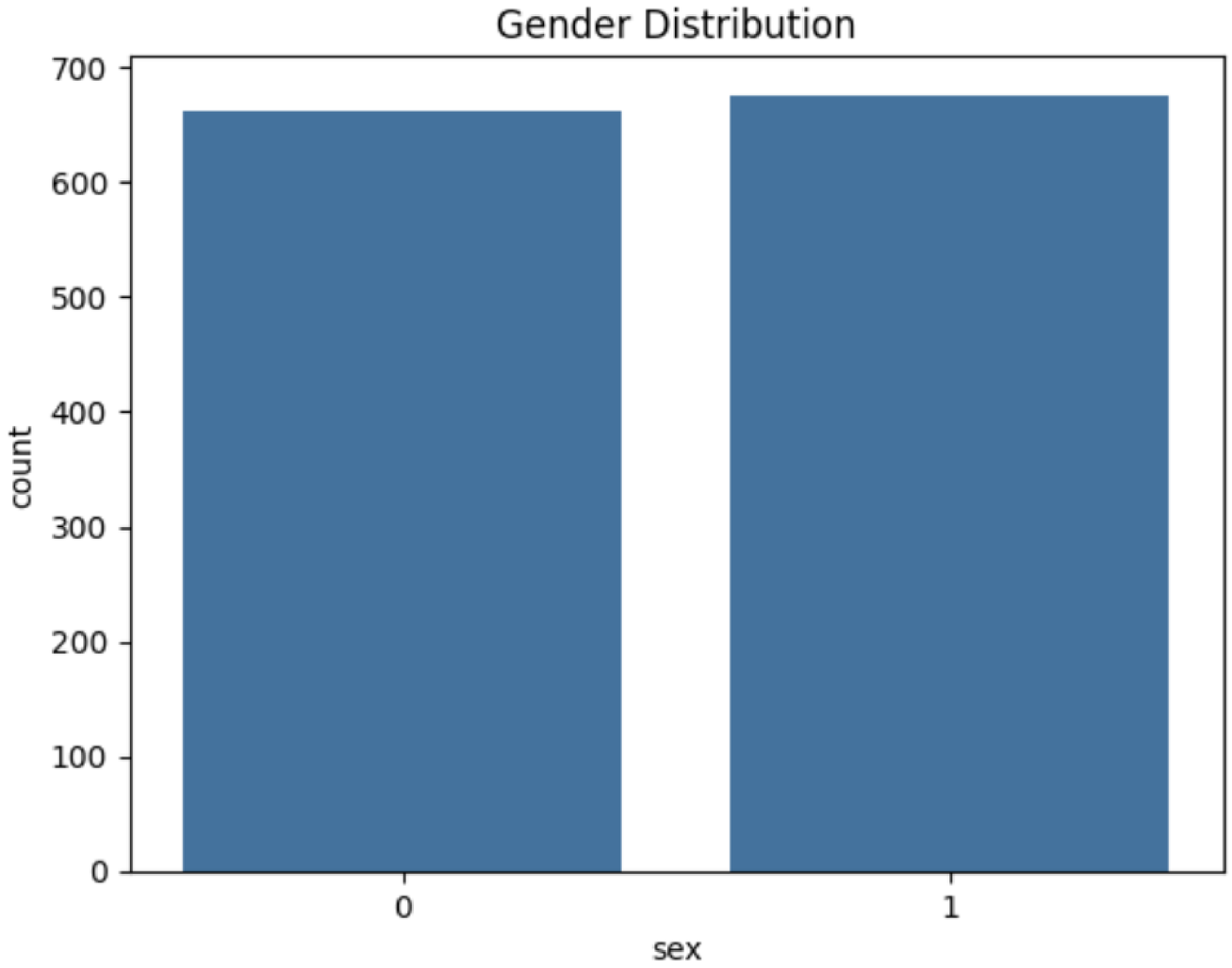
Objective: The goal of this project is to predict individual medical expenses based on various patient attributes using the Insurance dataset. The dataset includes 1338 observations and 7 variables, namely age, body mass index (BMI), number of children covered, smoking status, residential region, and the target variable, medical charges.

Insights from EDA

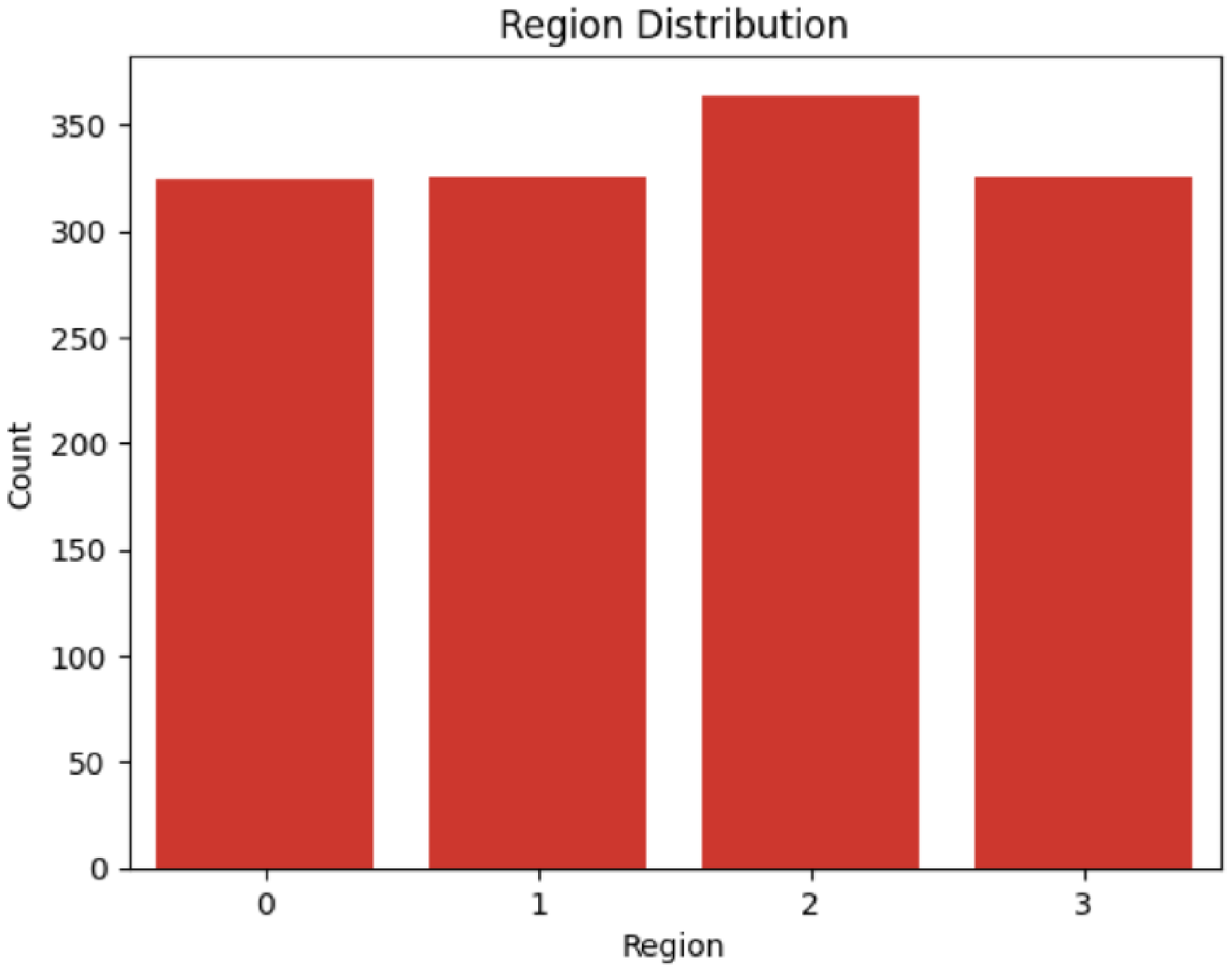
Age distribution



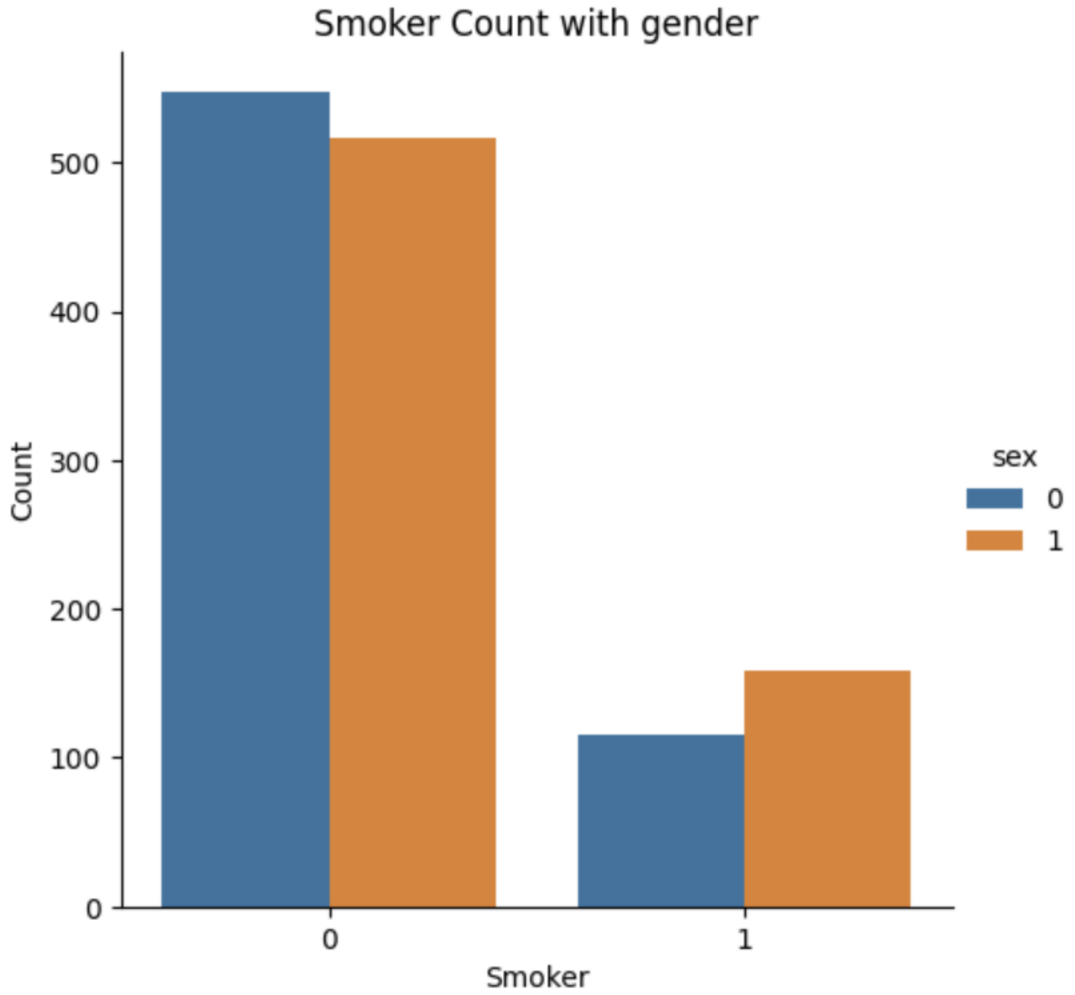
Gender Distibution



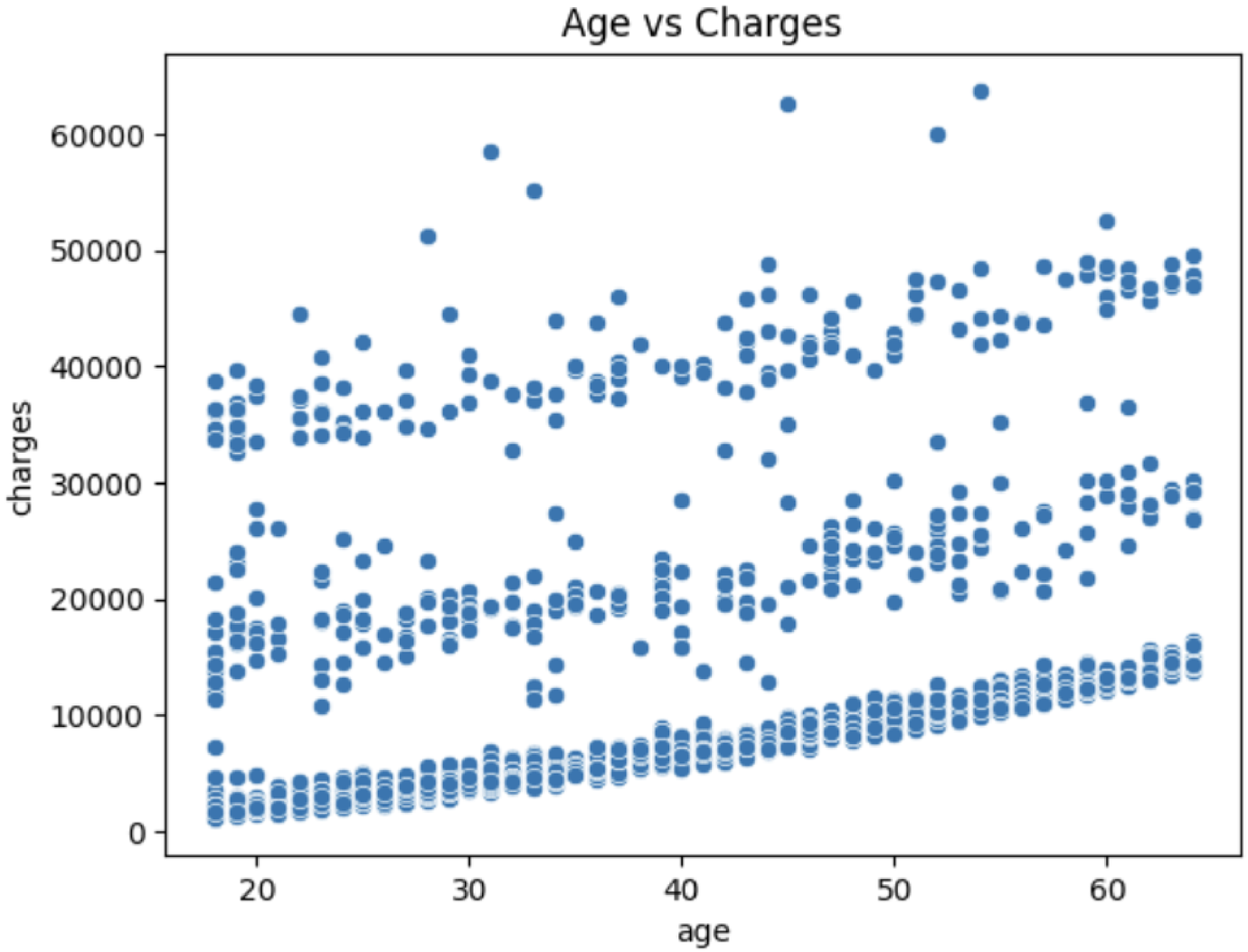
Region Distribution



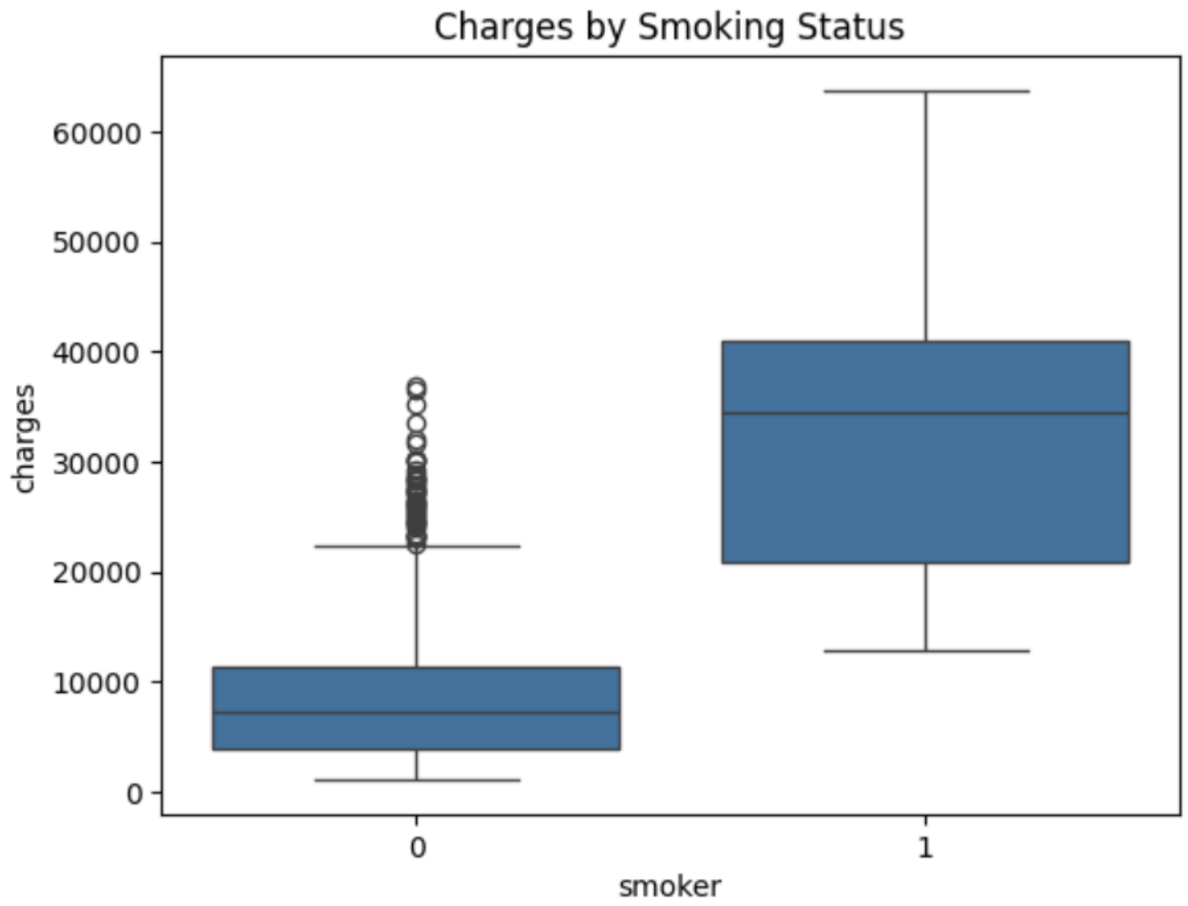
Smoker Count of Male & Female



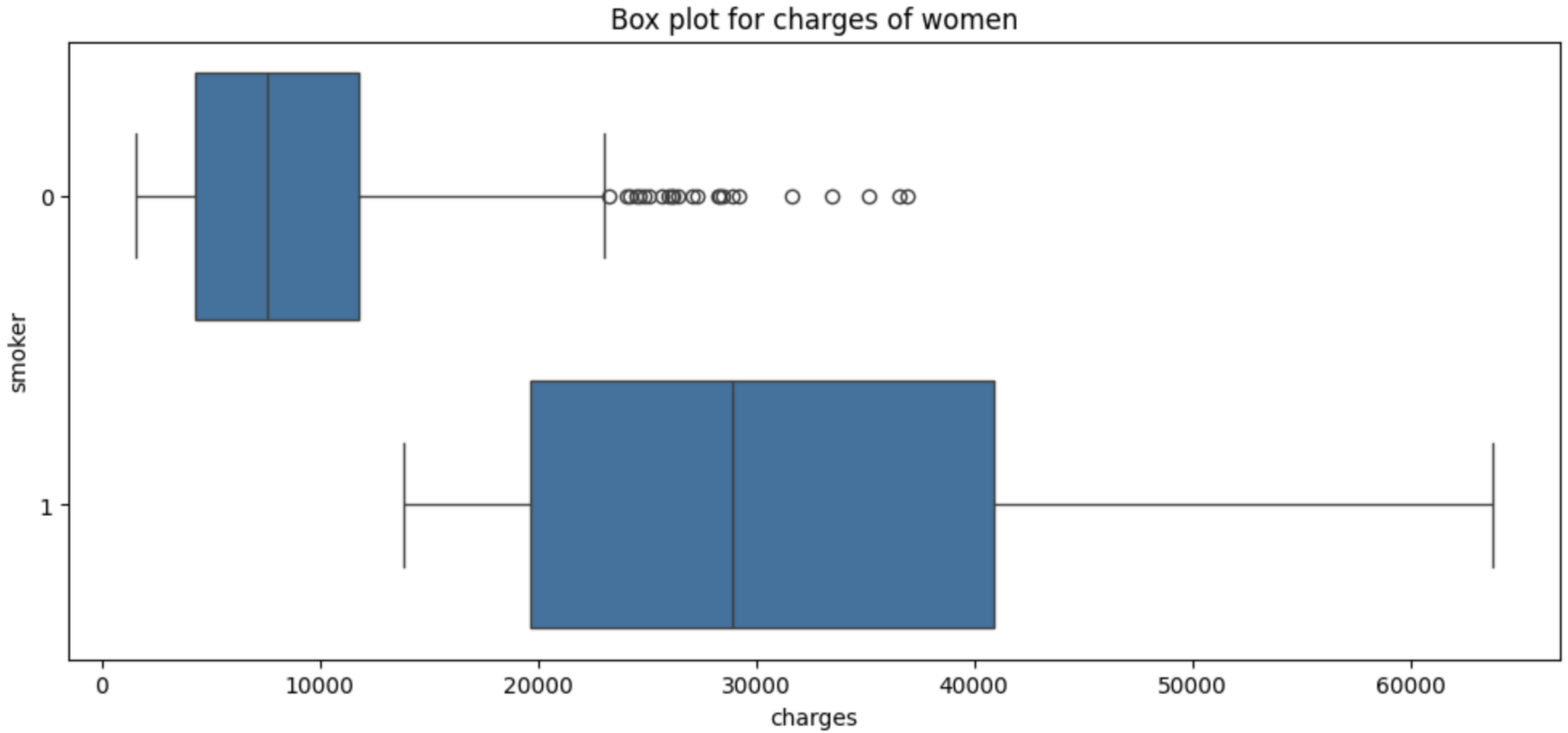
Charge variance accordance with age



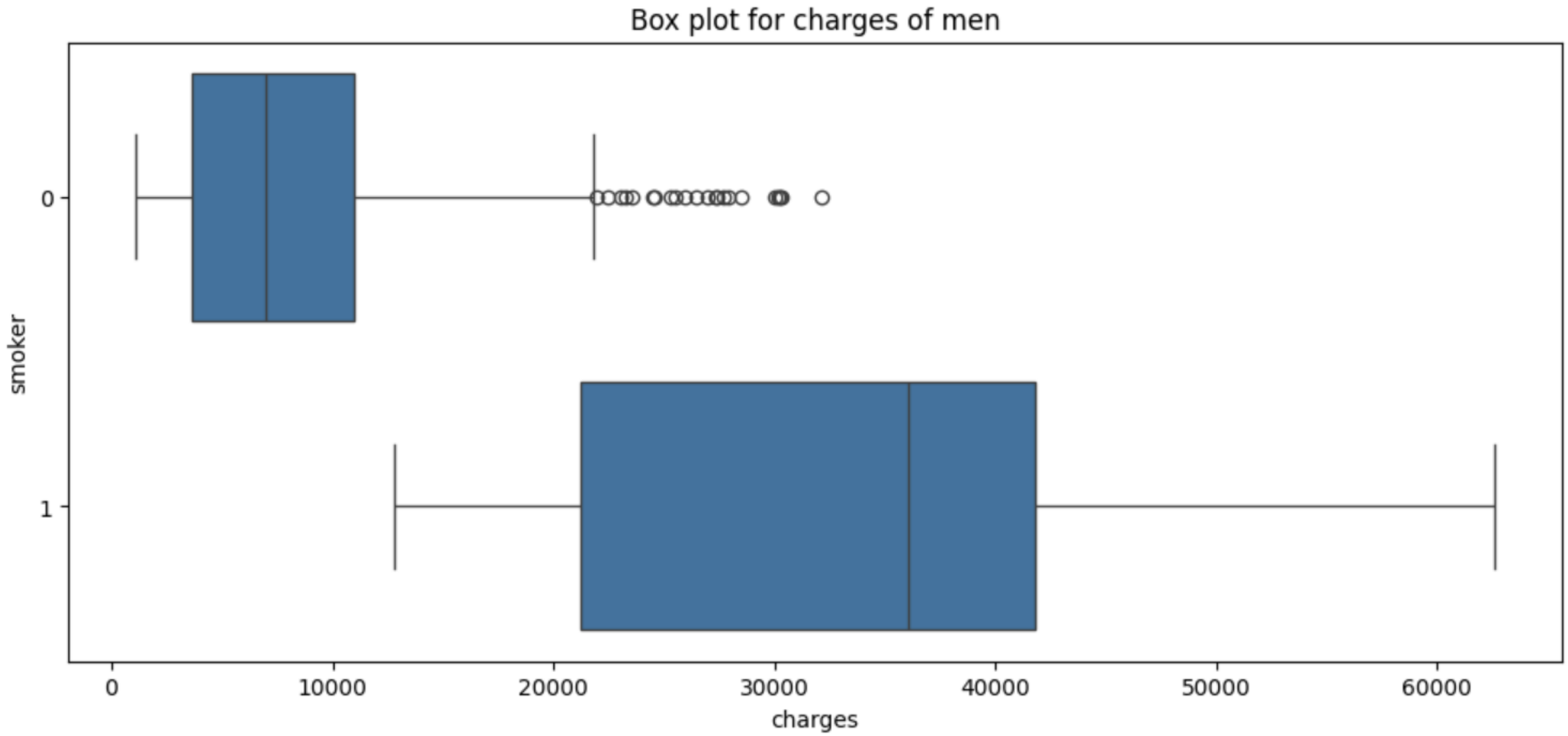
Charge variance - Smokers



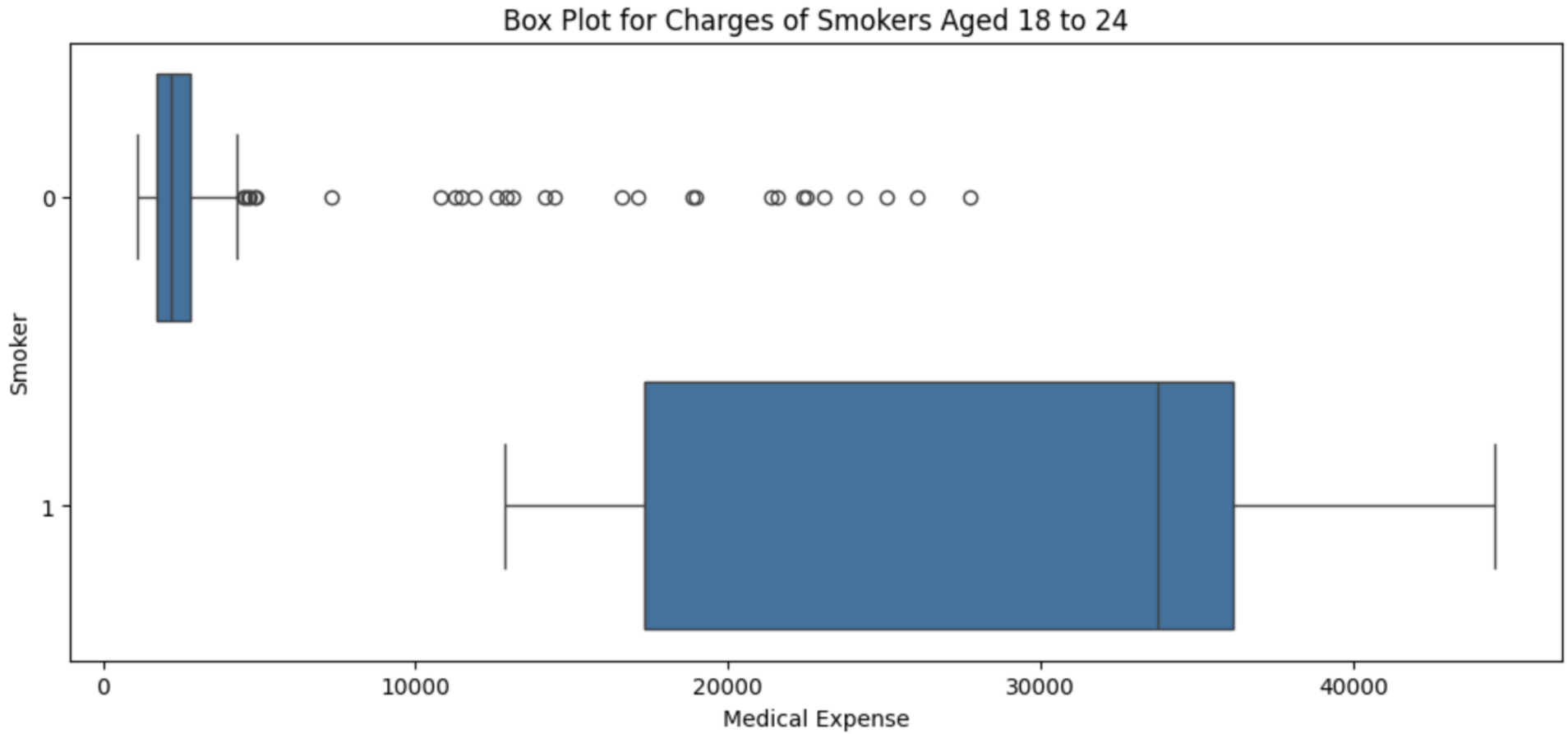
Charge variance - Women



Charge variance - Men

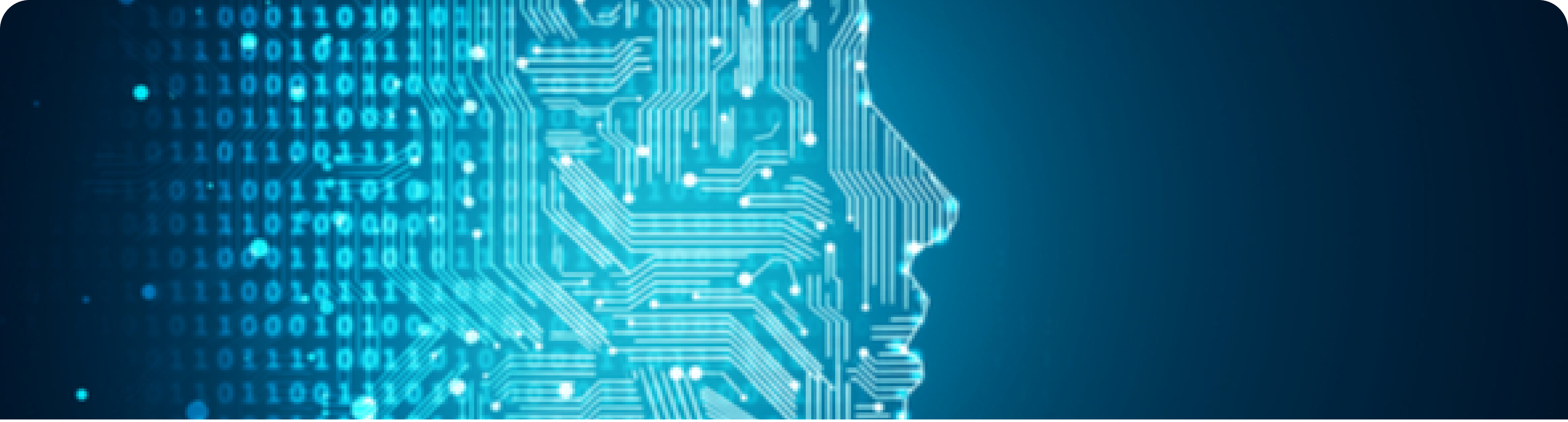


Young Smokers medical expenses



Conclusions from visualization

1. The data has good balance of distribution in age, gender and region
2. The charges are in a steady but not steep inclination with age
3. Smokers have much bigger charges than non smokers
4. From the data, young men tends to be recorded very less charges compared to smoking individuals



Predictive Models

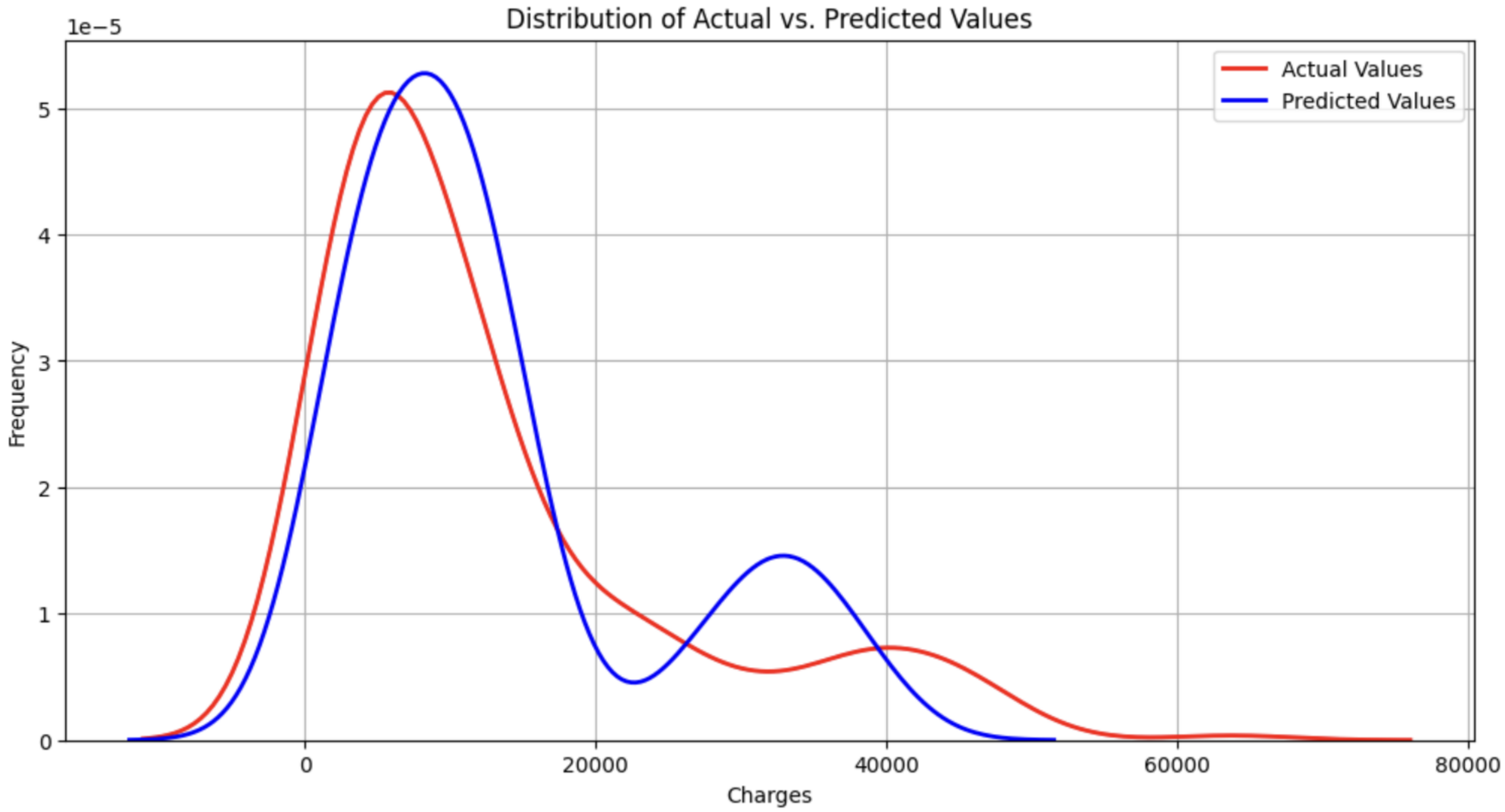
Linear Regression

```
# Model evaluation
print('Train RMSE:', mean_squared_error(y_train, y_pred_train, squared=False))
print('Test RMSE:', mean_squared_error(y_test, y_pred_test, squared=False))
print('R² Score:', r2_score(y_test, y_pred_test))
```

```
Train RMSE: 6105.789320191615
Test RMSE: 5799.587091438356
R² Score: 0.7833463107364539
```

- Linear regression model has 78.33 % accuracy

Linear Regression - Actual vs predicted



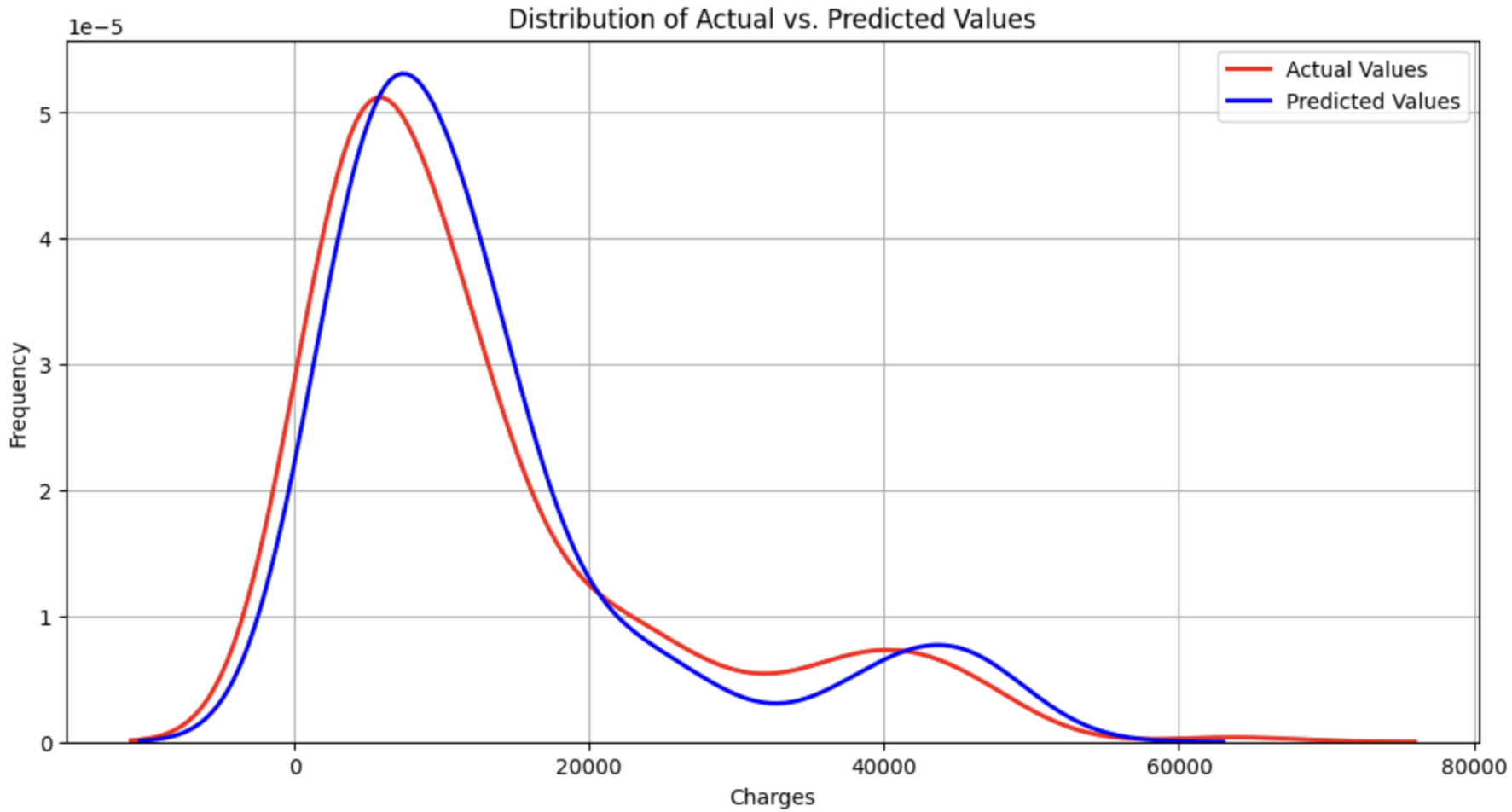
Random Forest

```python
# Evaluate the model
print(f'Training MSE: {mean_squared_error(y_train, y_pred_train)}')
print(f'Training R^2: {r2_score(y_train, y_pred_train)}')
print(f'Test MSE: {mean_squared_error(y_test, y_pred_test)}')
print(f'Test R^2: {r2_score(y_test, y_pred_test)}')
```

```
Training MSE: 3708028.280447489
Training R^2: 0.974309324253187 7
Test MSE: 21073365.415079337
Test R^2: 0.8642606273649586
```

- Random Forest model has 86.42 % accuracy

Random Forest - Actual vs Predicted



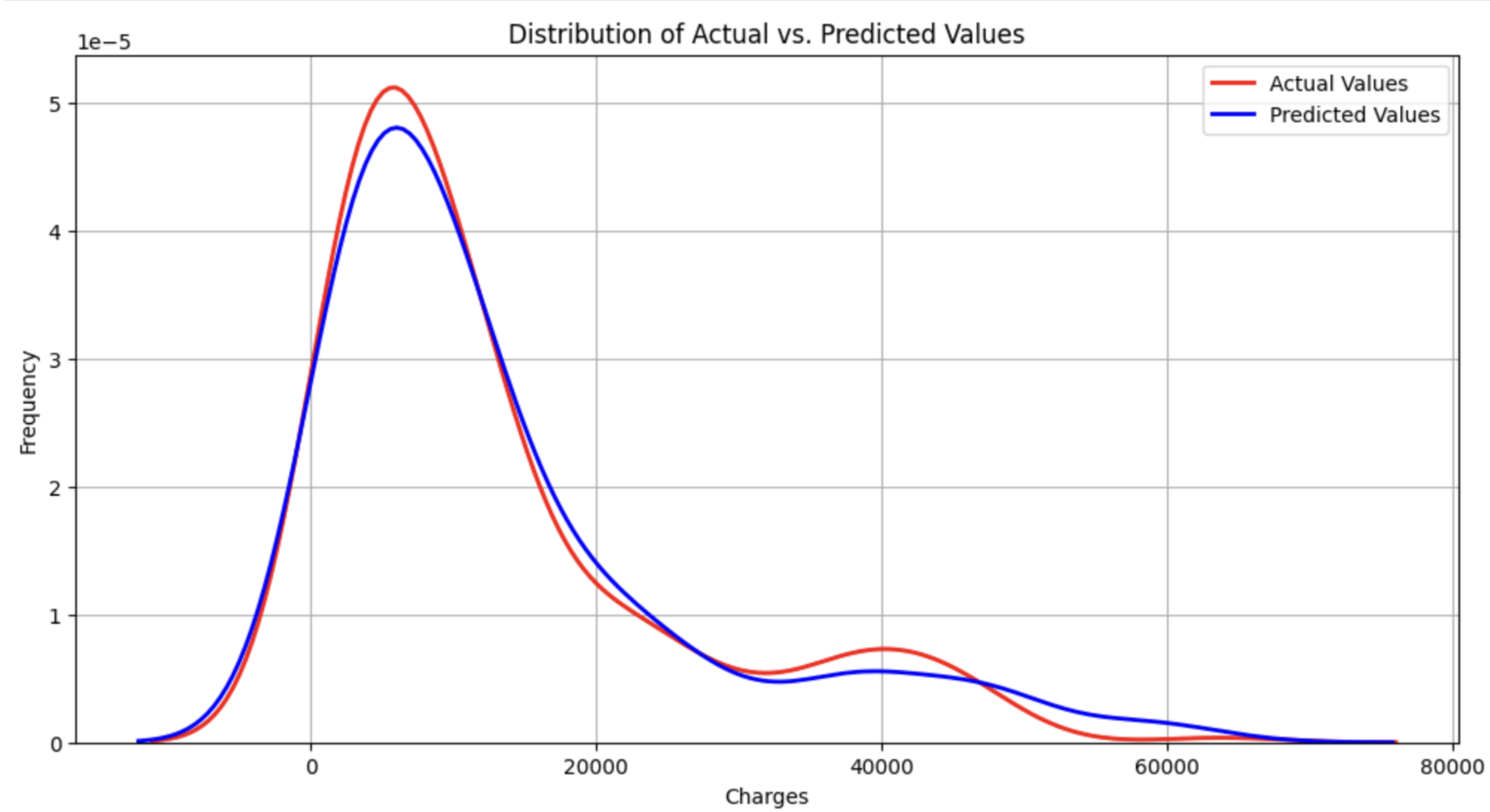
Decision Tree

```
# Evaluate the model
print(f'Training MSE: {mean_squared_error(y_train, y_pred_train)}')
print(f'Training R^2: {r2_score(y_train, y_pred_train)}')
print(f'Test MSE: {mean_squared_error(y_test, y_pred_test)}')
print(f'Test R^2: {r2_score(y_test, y_pred_test)}')
```

```
Training MSE: 244239.5543823394
Training R^2: 0.9983078124756305
Test MSE: 49003243.60682007
Test R^2: 0.6843565603663775
```

- Decision Tree model has 68.43 % accuracy

Decision Tree - Actual vs Predicted



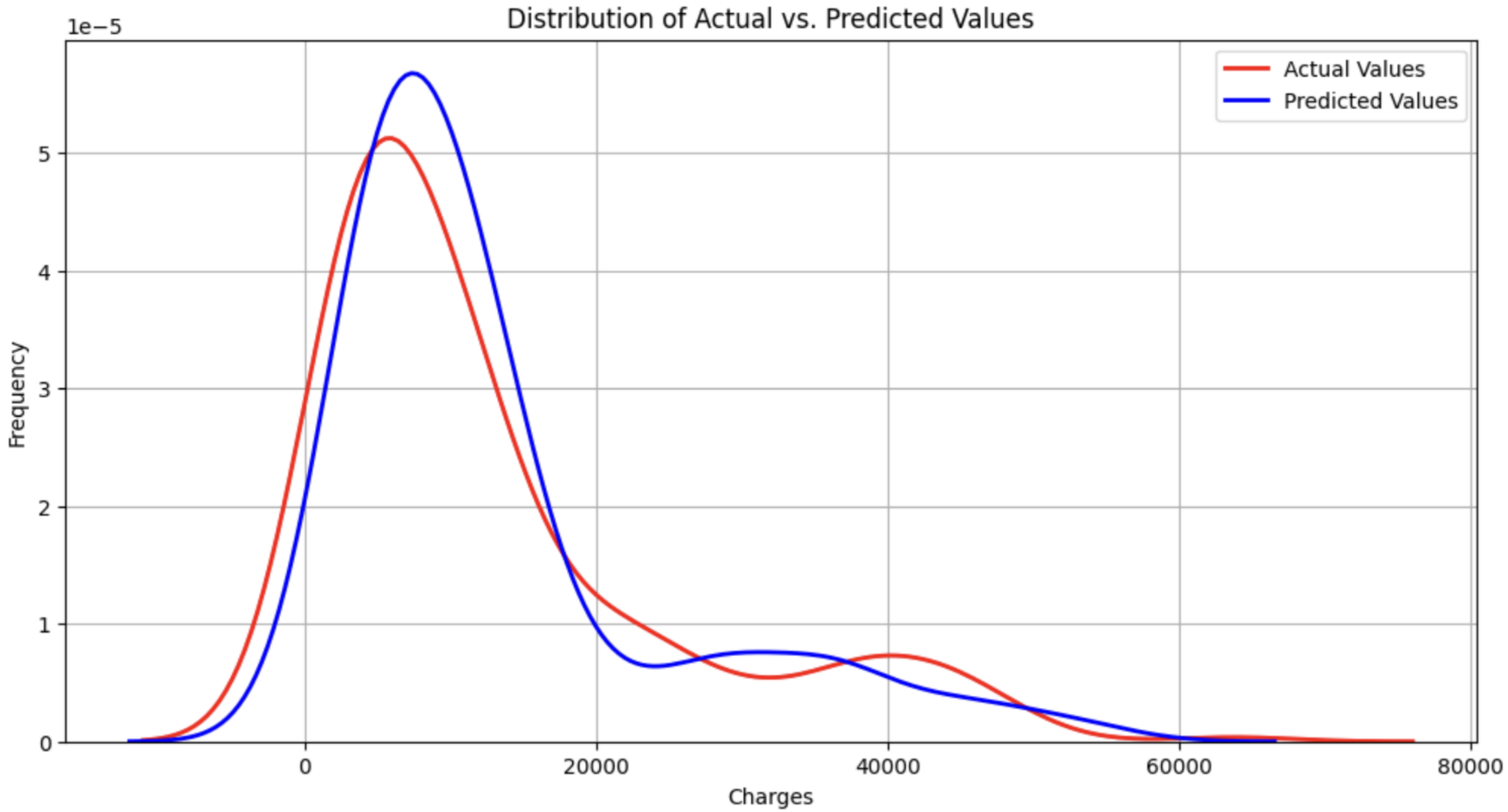
Polynomial Regression

```
# Evaluate the model
print(f'Training MSE: {mean_squared_error(y_train, y_pred_train)}')
print(f'Training R^2: {r2_score(y_train, y_pred_train)}')
print(f'Test MSE: {mean_squared_error(y_test, y_pred_test)}')
print(f'Test R^2: {r2_score(y_test, y_pred_test)}')
```

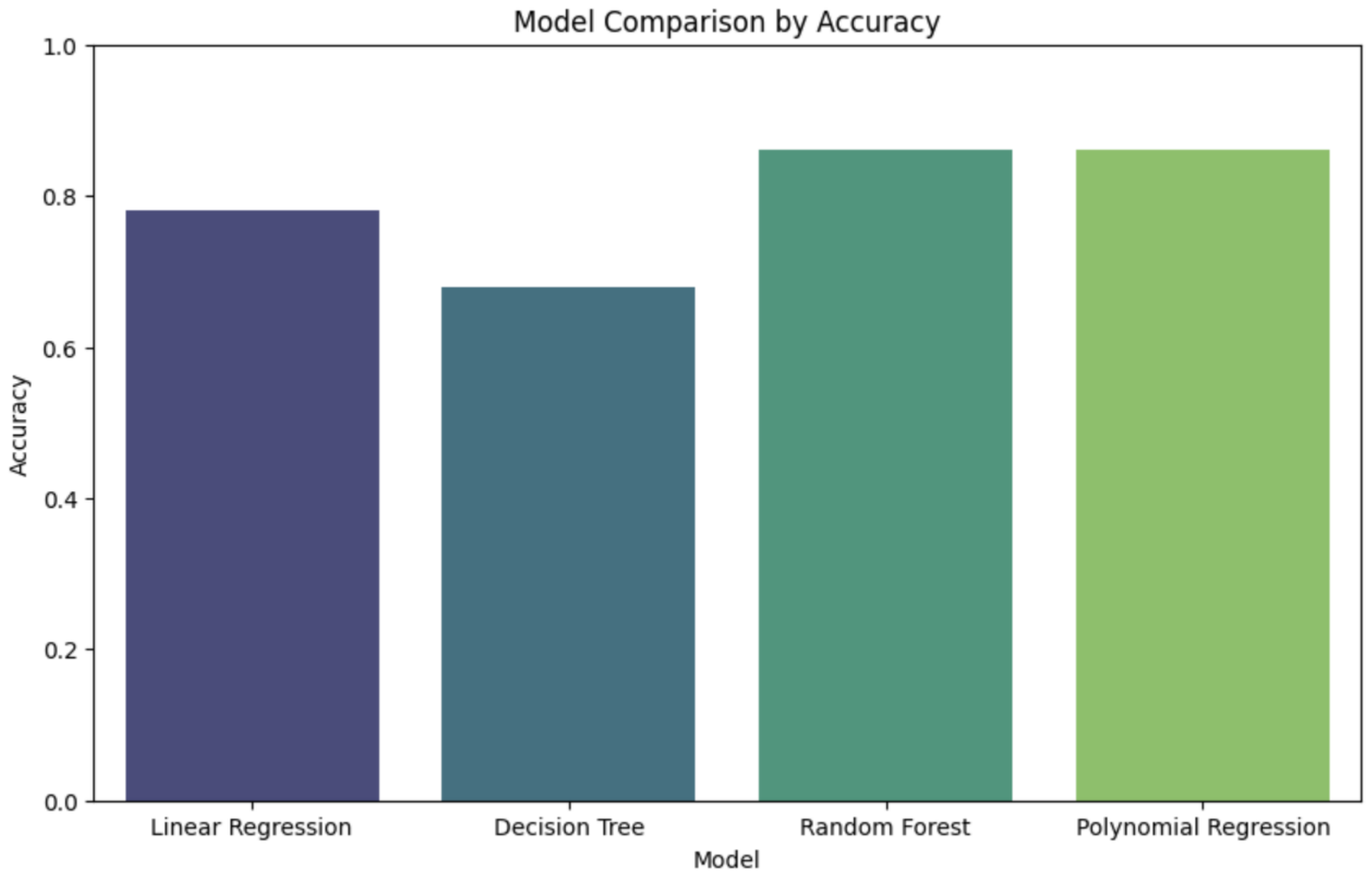
```
Training MSE: 23468410.545391146
Training R^2: 0.837401637739494
Test MSE: 20970715.271709584
Test R^2: 0.864921825317324 3
```

- Polynomial Regression model has 86.49 % accuracy

Decision Tree - Actual vs Predicted



Comparing model accuracy



Thanks

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