Vivek A

Click here to view the github

Diamond price prediction

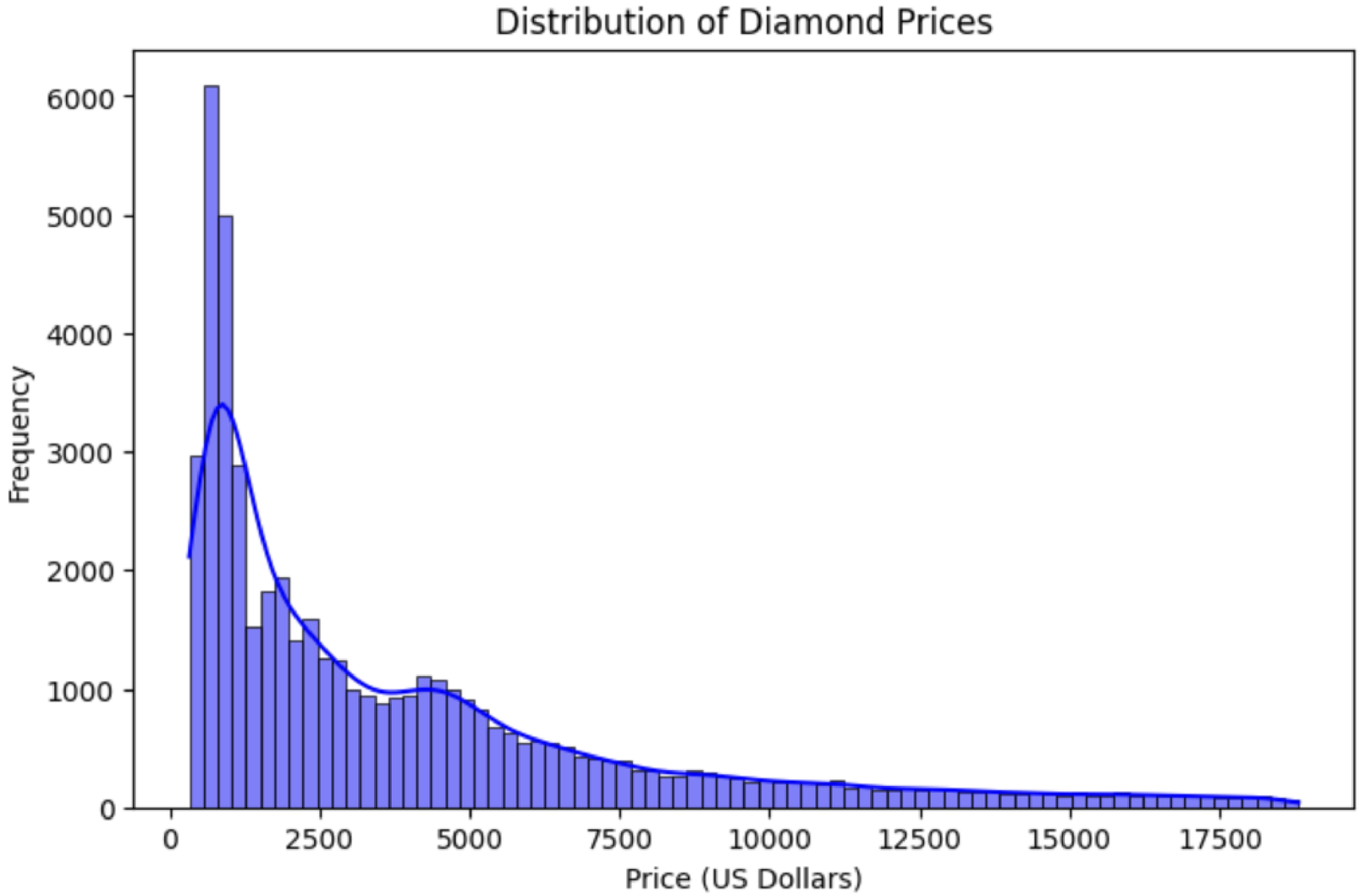
Data Science project

Project Objective

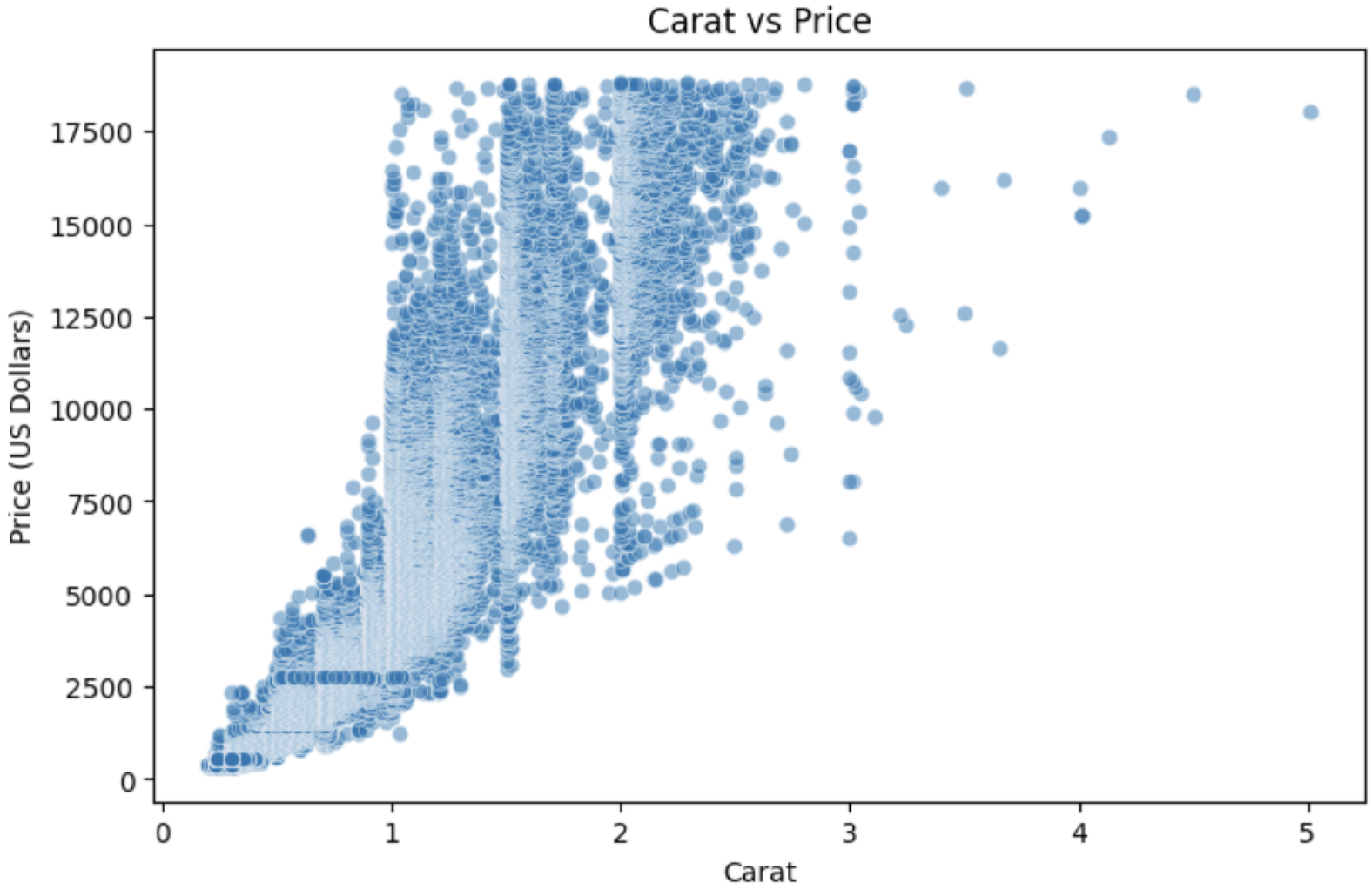
This project aims to predict the price of diamonds based on several physical and qualitative attributes using a machine learning model. The dataset includes various features that impact the value of a diamond, such as carat (weight), cut (quality), color (from J to D), clarity (from I1 to IF), and physical dimensions like length (x), width (y), depth (z), depth percentage, and table width. By analyzing these features, the model will predict the price of the diamond, which ranges from 326 to 18,823.

Insights from EDA

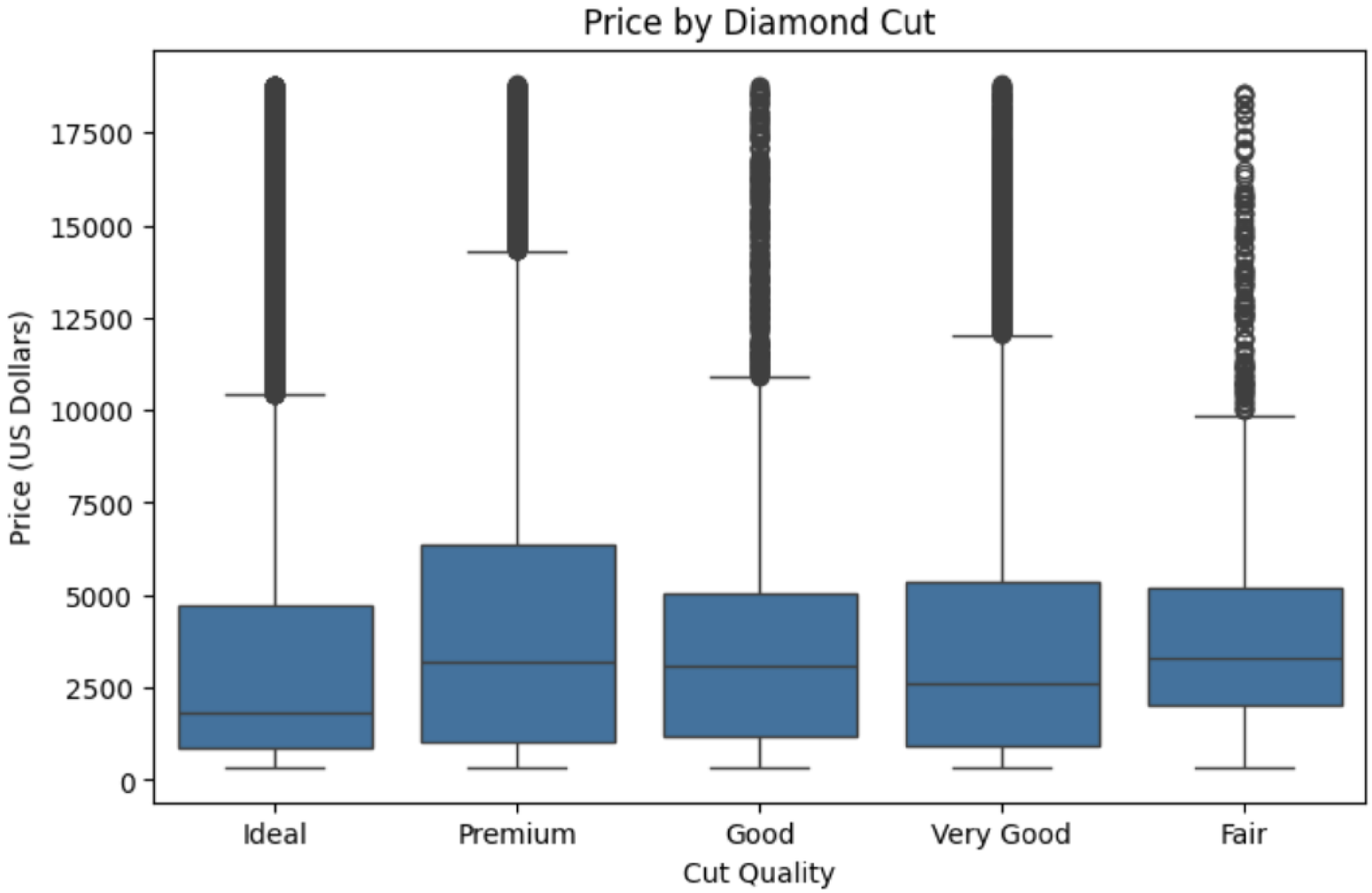
Price distribution



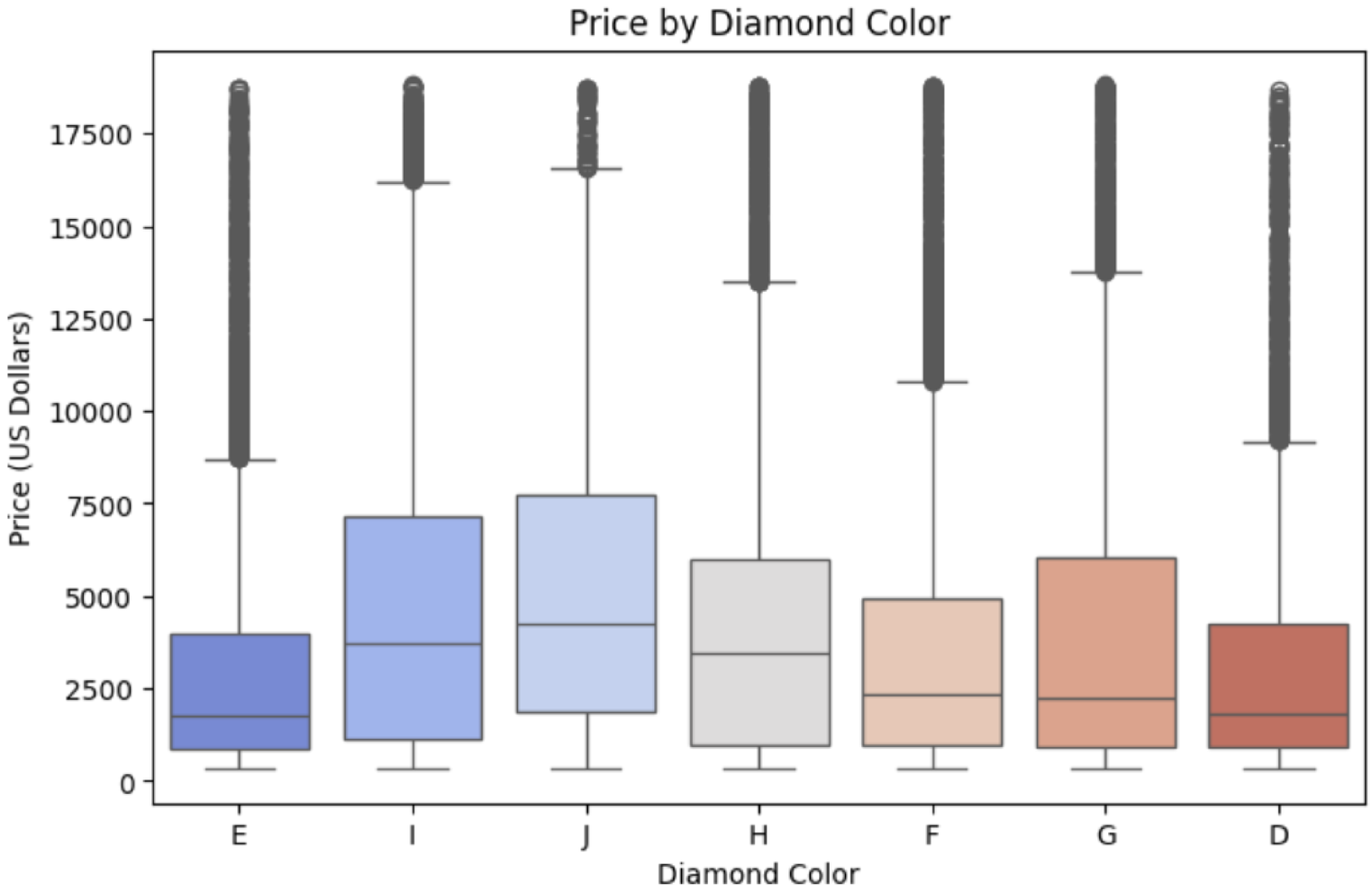
Carat with price



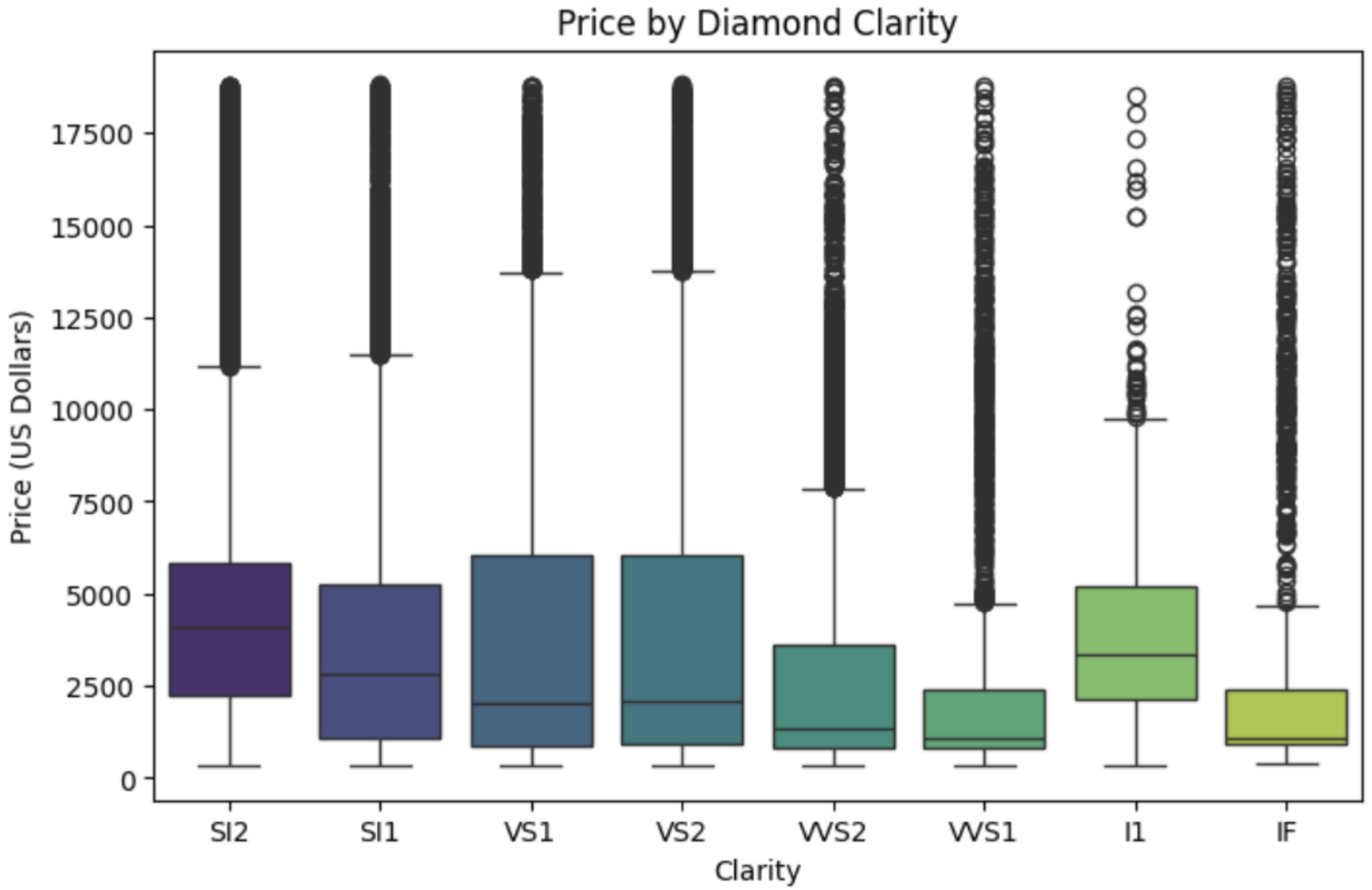
Price by Cut


Price by Diamond Cut
Price (US Dollars)
17500
15000
12500
10000
7500
5000
2500
0
Ideal
Premium
Good
Very Good
Fair
Cut Quality

Price by Color



Price according to Clarity



Conclusions from visualization

1. Distribution of Diamond Prices Observation: The distribution of diamond prices is highly right-skewed, indicating that most diamonds in the dataset are priced at the lower end (below $5,000), with a few very expensive diamonds priced much higher. Finding: The market likely has a higher demand for more affordable diamonds, and fewer diamonds are priced in the luxury category.

2. Relationship Between Carat and Price Observation: There is a clear positive relationship between carat and price. As the carat (weight) of the diamond increases, the price tends to rise, although the increase is not strictly linear. Finding: Carat is a strong determinant of diamond price, with larger diamonds generally commanding higher prices. However, the price variation for diamonds of similar carat weights suggests other factors (like cut, color, and clarity) also influence pricing.

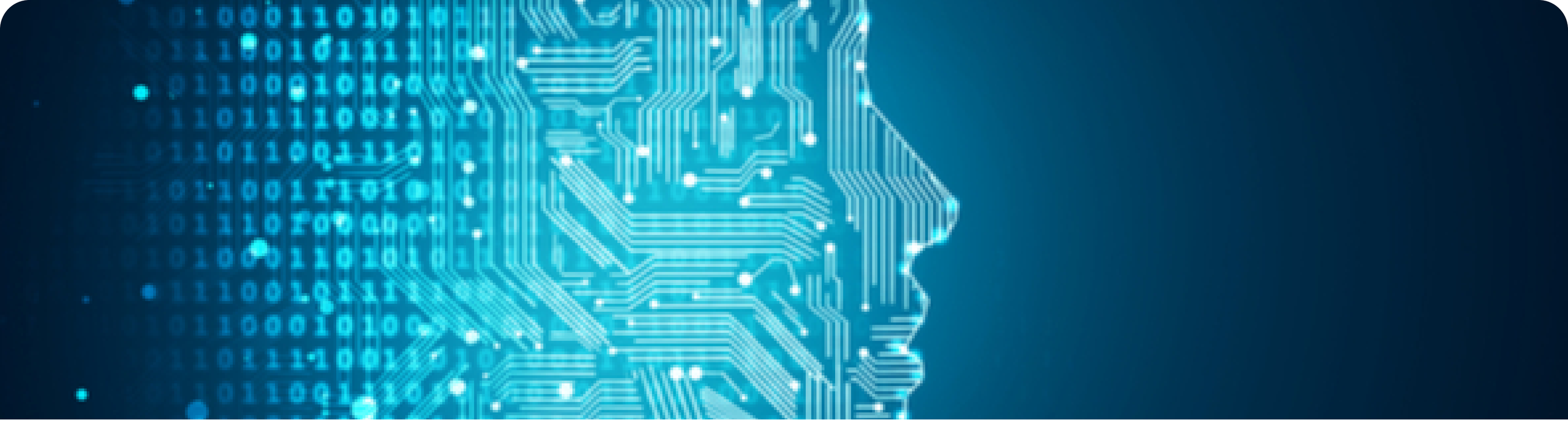
Conclusions from visualization

3. Boxplot of Price by Cut Observation: The median price of diamonds increases with the quality of the cut, from Fair to Ideal. However, the price variance within each cut category can be significant. Finding: Diamonds with better cuts (like Premium and Ideal) generally fetch higher prices, indicating that cut quality plays an important role in price determination. Still, within each cut category, prices can vary widely depending on other factors.

4. Price by Color Observation: Diamonds with better color grades (closer to D) tend to have higher median prices, though the price difference is not as pronounced as with cut quality. Some overlap in price distributions occurs across different color grades. Finding: While diamond color impacts price, the effect is less substantial compared to other factors like carat or cut. This suggests that consumers may not place as much emphasis on color as they do on other attributes.

Conclusions from visualization

5. Price by Clarity Observation: Diamonds with higher clarity grades (like IF and VVS1) tend to be priced higher, but the overall variation within each clarity category is considerable. Finding: Clarity affects diamond pricing, with clearer diamonds (fewer inclusions) generally costing more. However, the wide price variance within clarity categories suggests that clarity alone may not be the sole driver of price differences.



Predictive Models

Linear Regression

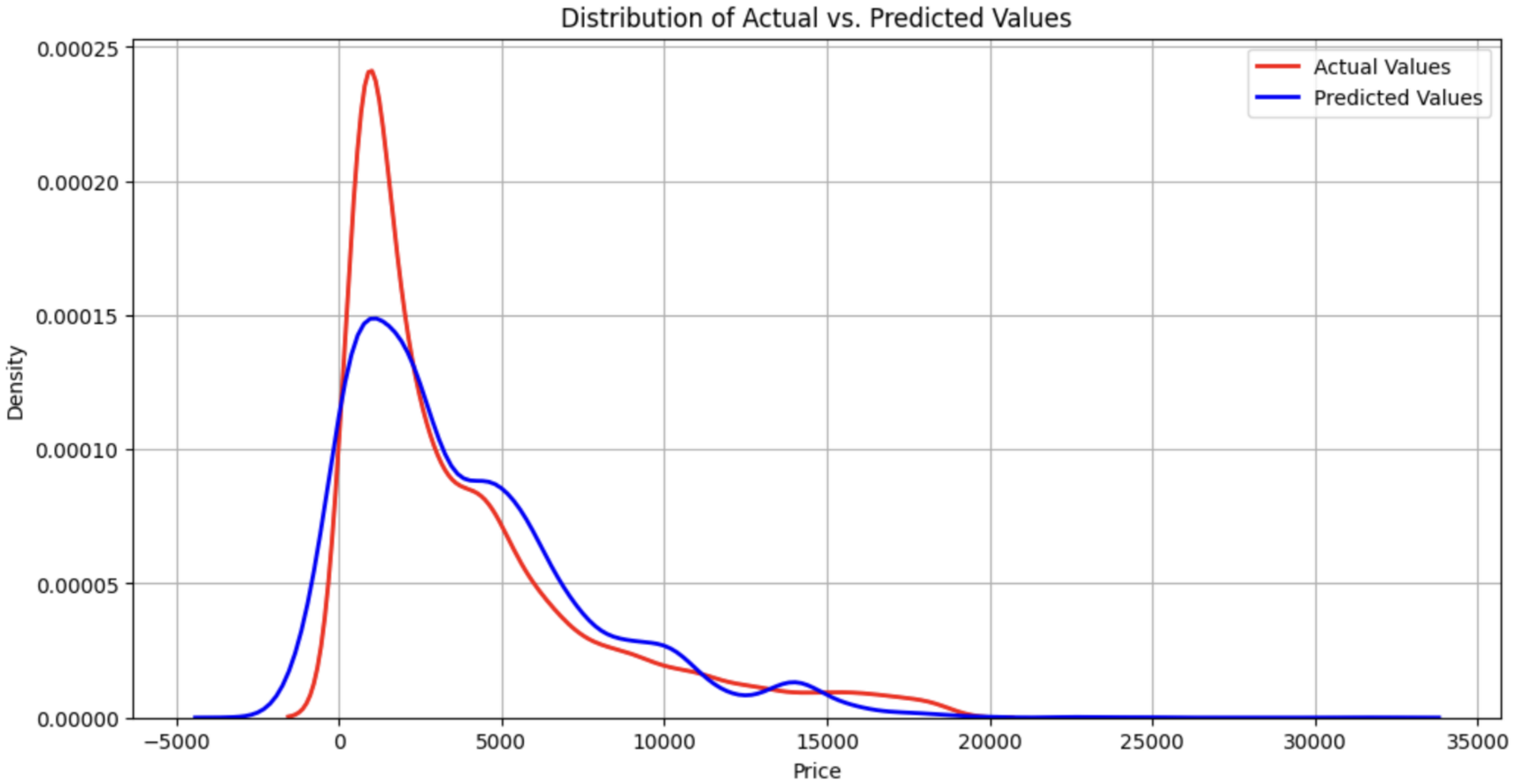
```
# Model evaluation
print('Train RMSE:', mean_squared_error(y_train, y_pred_train, squared=False))
print('Test RMSE:', mean_squared_error(y_test, y_pred_test, squared=False))
print('R² Score:', r2_score(y_test, y_pred_test))
```

```
Train RMSE: 1217.2167246581892
Test RMSE: 1220.6192606249244
R² Score: 0.9051412400569623
```

- Linear regression model showing high level of accuracy of 90.51 %

- Linear regression model has 90.51% accuracy

Linear Regression - Actual vs predicted



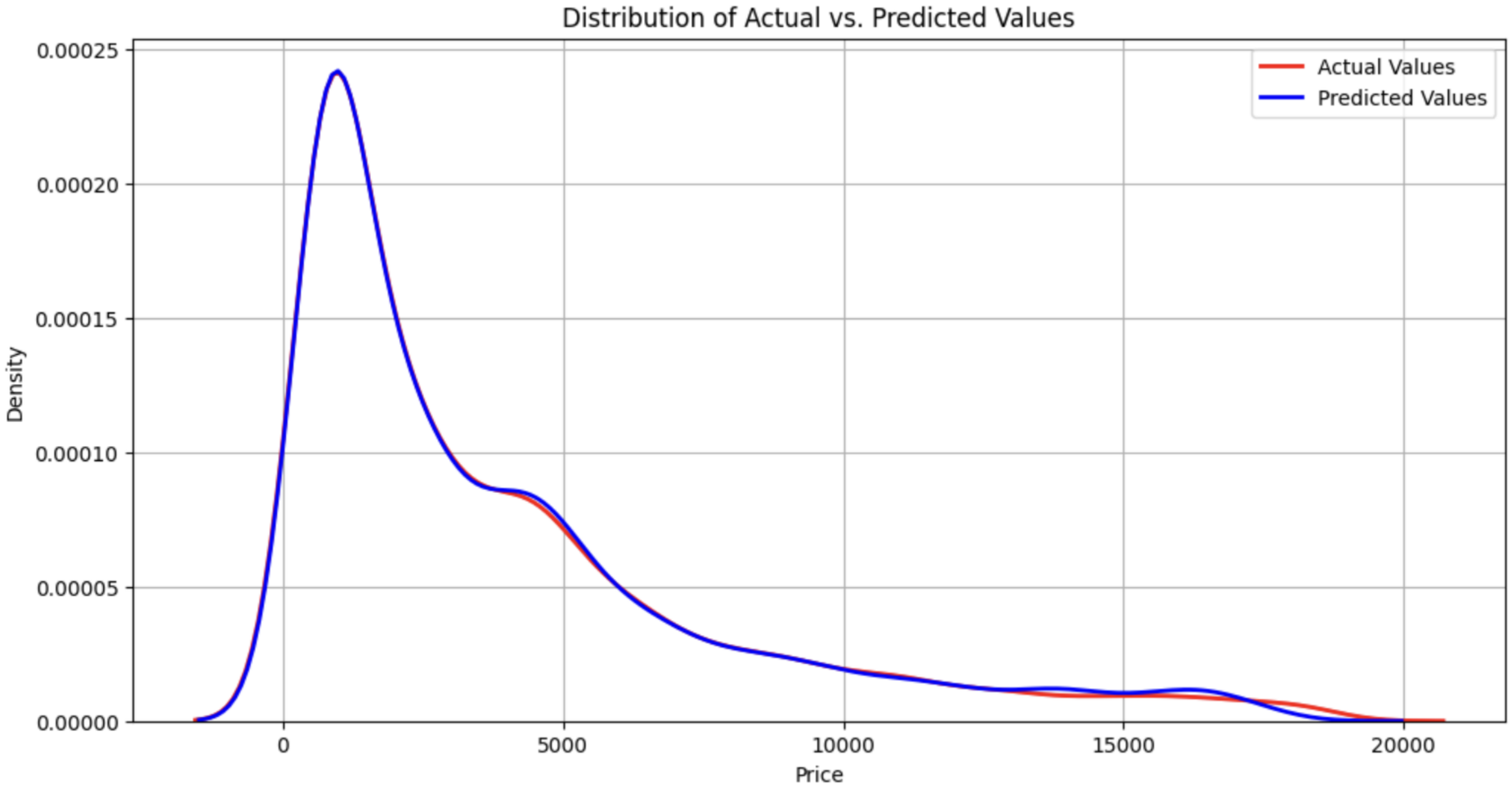
Random Forest

```
# 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: 202.9190184417962
Test RMSE: 549.9768662314119
R² Score: 0.980742236454181 8
```

- Random Forest model has 98.07% accuracy

Random Forest - Actual vs Predicted



Thanks

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