SC1015

Mini Project

Team 5 | A135
Mohamed Muhsin Mohammad Zubair Rahman
Darrell Ma Wei Ze
Lim Weijun

Table of contents

Diamonds

- Valuable due to unique physical & chemical properties
- Prices range from **$300 to $20,000**
- Consumers feel overwhelmed by the wide range of diamond prices

How can we predict the price of a diamond based on its characteristics?

Setting the Stage

Data Cleaning

Basic Visualisation

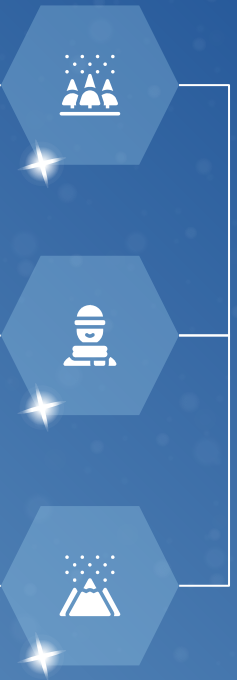
Exploratory Data Analysis(EDA)

Data Cleaning

Removed an insignificant column named “unnamed”

Capitalised all variable names for ease

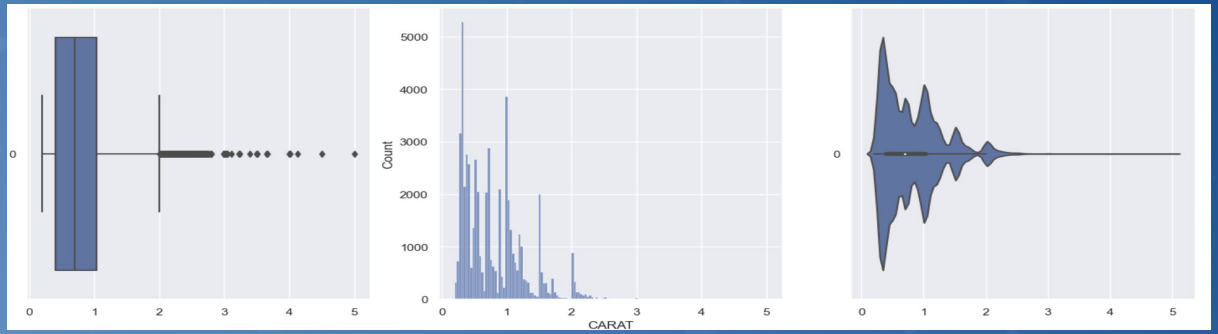
Check for missing values in dataset

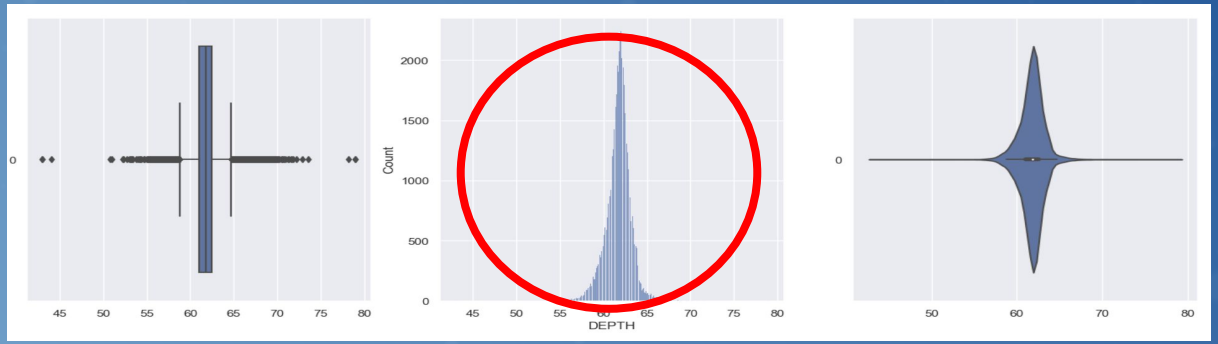


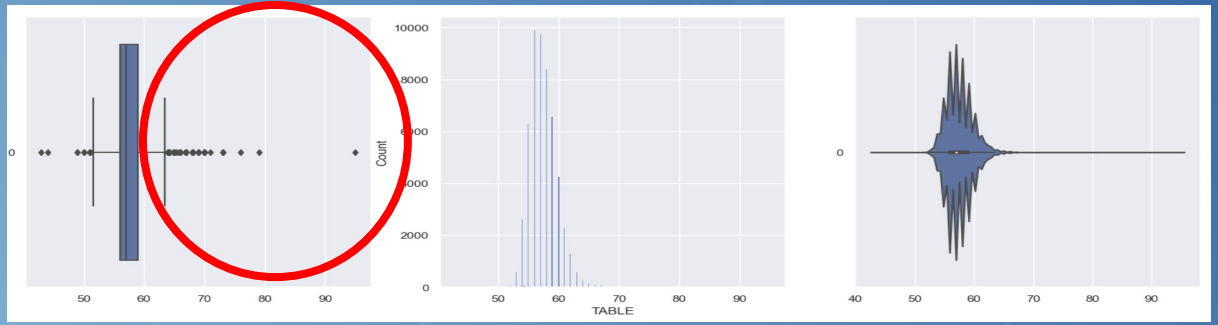
Basic Visualisation

Numerical Variables(Carat, depth and table)

- Used boxplots, histograms and violin plots to visualise the numerical variables


0
0 1 2 3 4 5
Count
5000
4000
3000
2000
1000
0
0 1 2 3 4 5
CARAT
0
0 1 2 3 4 5


0
45 50 55 60 65 70 75 80
Count
2000
1500
1000
500
0
45 50 55 60 65 70 75 80
DEPTH
0
50 60 70 80


0
50 60 70 80 90
Count
10000
8000
6000
4000
2000
0
50 60 70 80 90
TABLE
0
40 50 60 70 80 90

Basic Visualisation

```
CARAT     1889
DEPTH     2545
TABLE      605
dtype: int64
```

- Table - least number of outliers (605)
- Depth - most number of outliers (2545)

```
CARAT     1.116705
DEPTH    -0.082187
TABLE     0.796836
dtype: float64
```

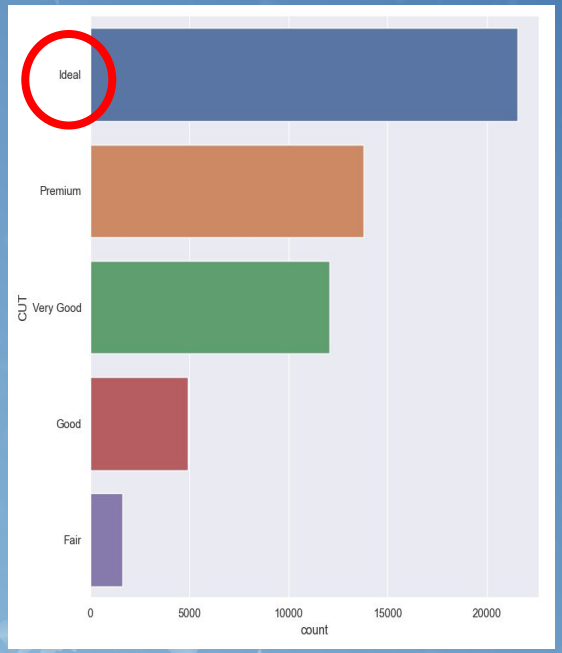
- Depth - minimum skewness (-0.08)
- Carat - highly skewed (1.12)

Basic Visualisation

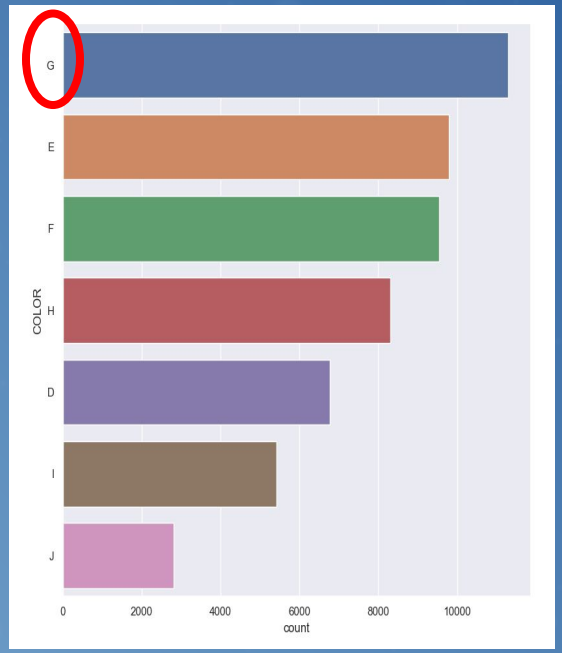
Categorical Variables(Cut, color and clarity)

- Used count plots to visualise the categorical variables

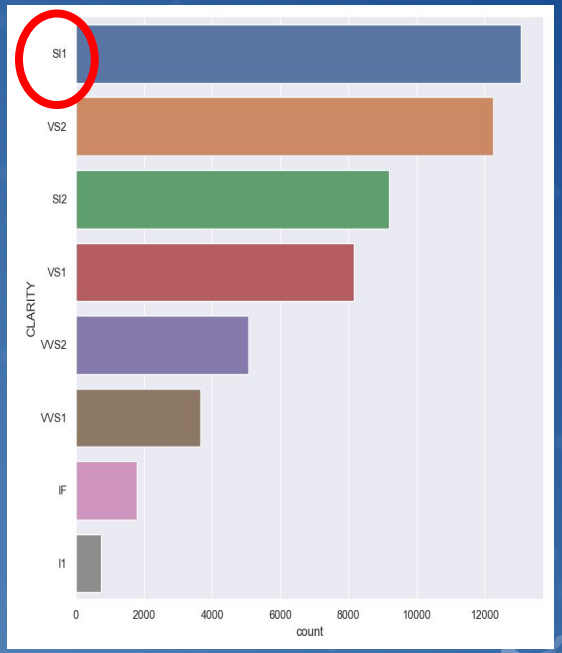
Basic Visualisation



Cut

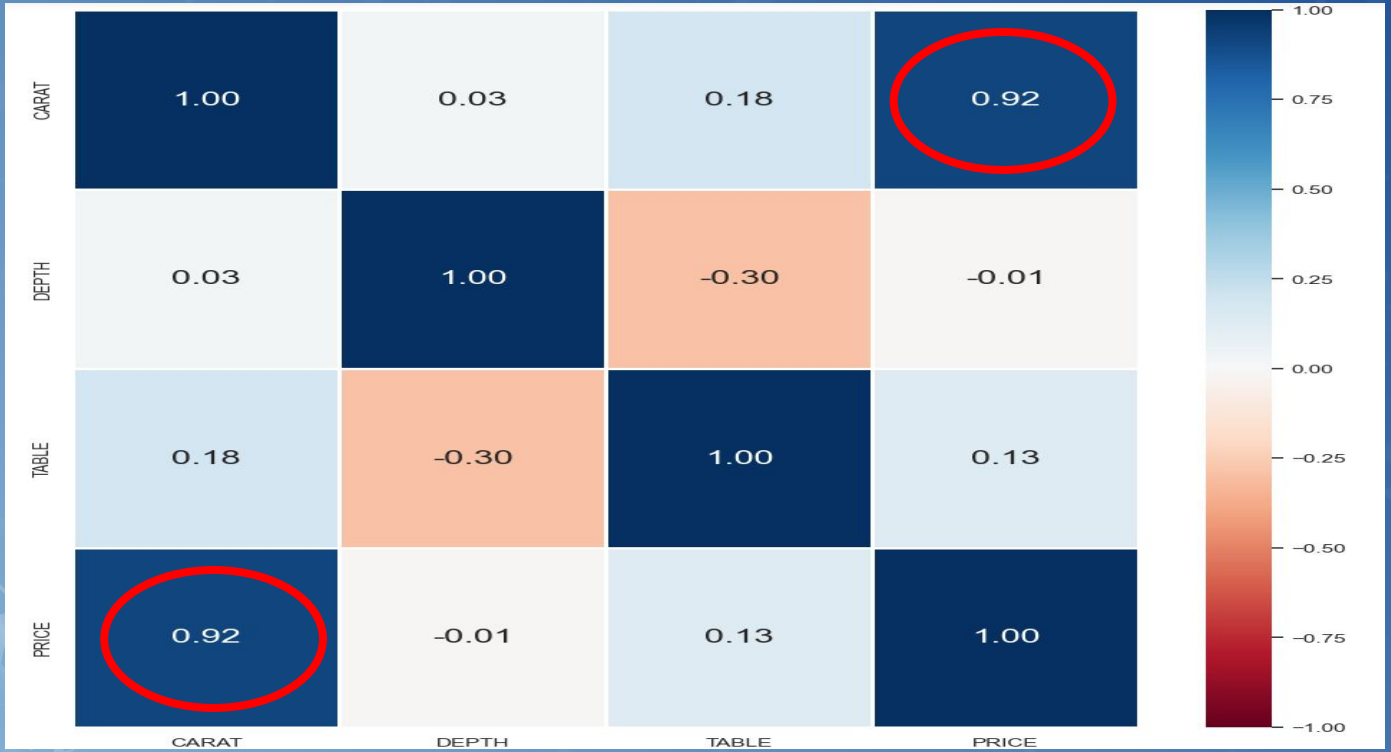


Color

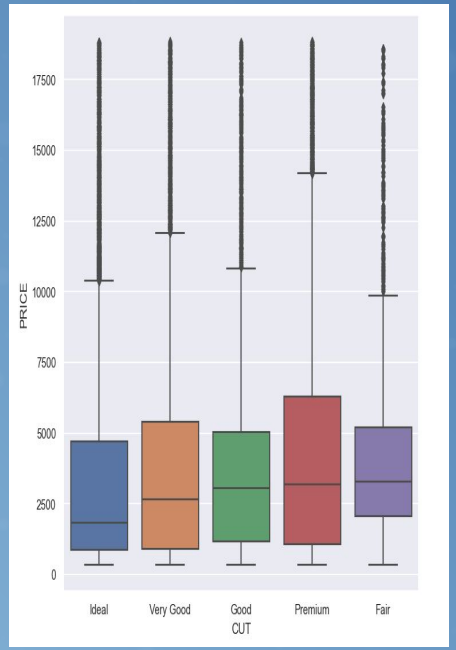


Clarity

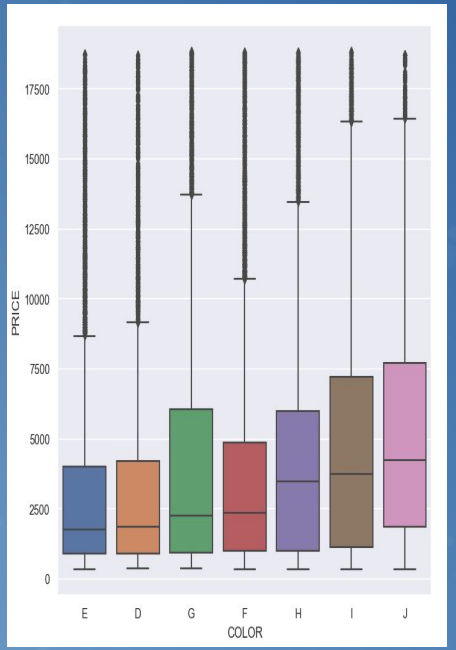
Exploratory Data Analysis(Numerical)



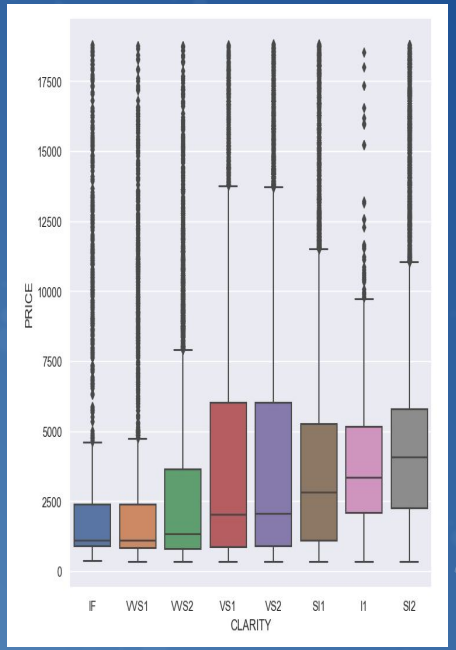
Exploratory Data Analysis(Categorical)



Cut



Color



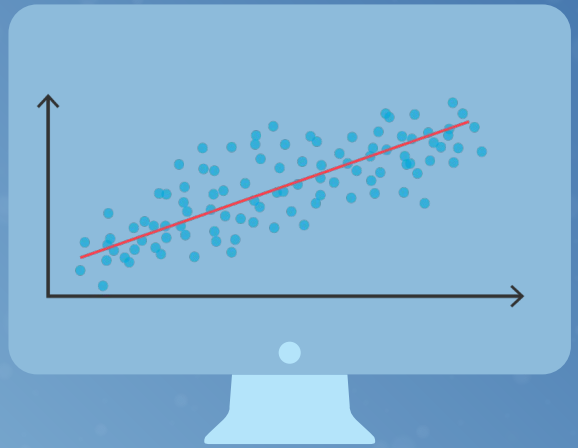
Clarity

Core Analysis

Machine Learning

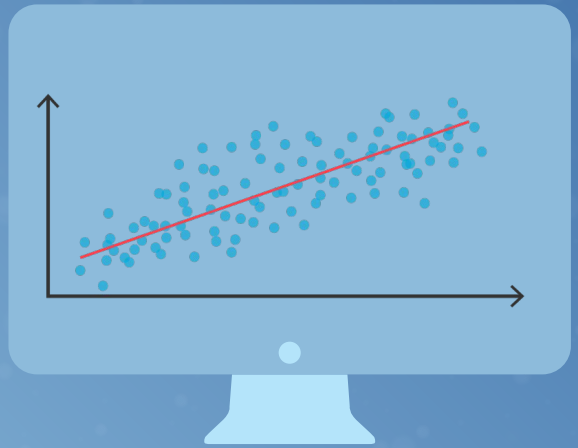
Application

Other techniques used

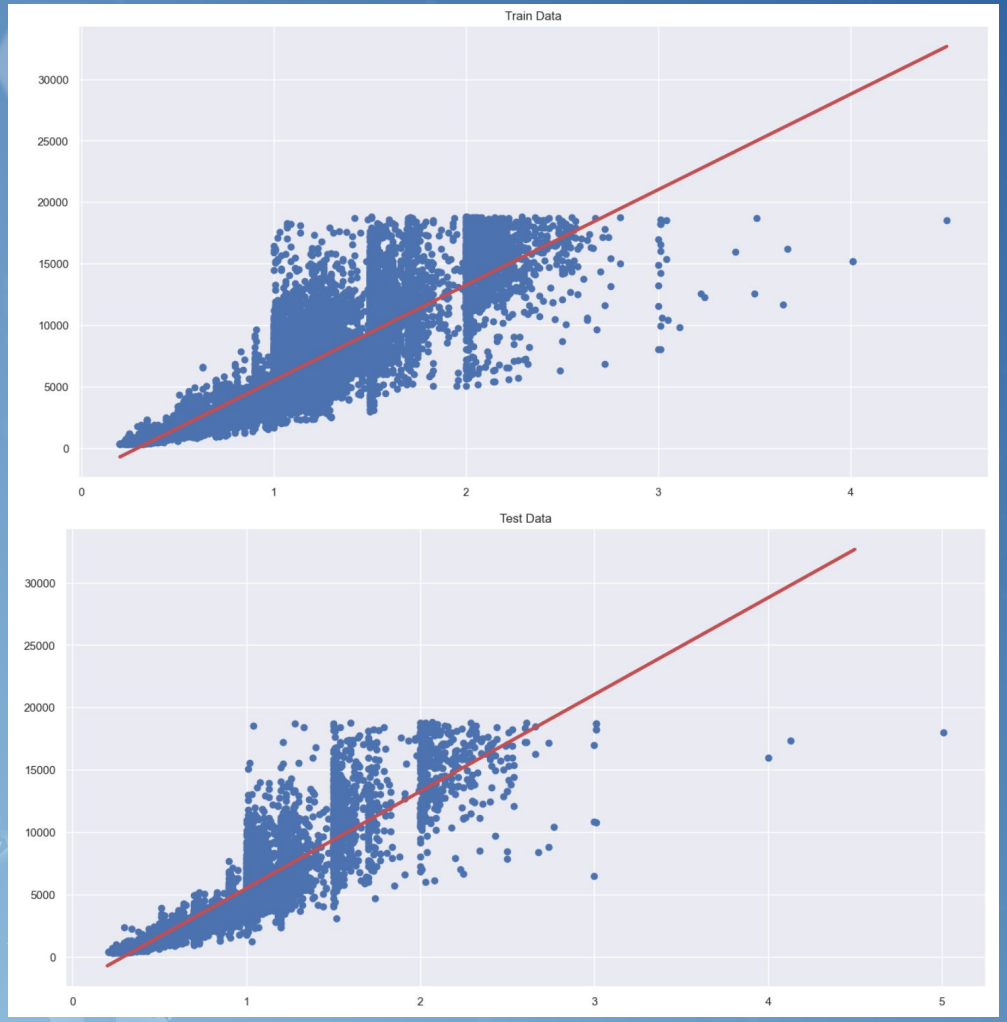


Regression Model

Using different regression models
to predict price



Linear Regression Model



Linear Regression CARAT Train Data

Explained Variance (R^2) :
0.8486179377482752
Mean Squared Error (MSE) :
2398890.660974258
Root Mean Squared Error (RMSE) :
1548.8352594689527

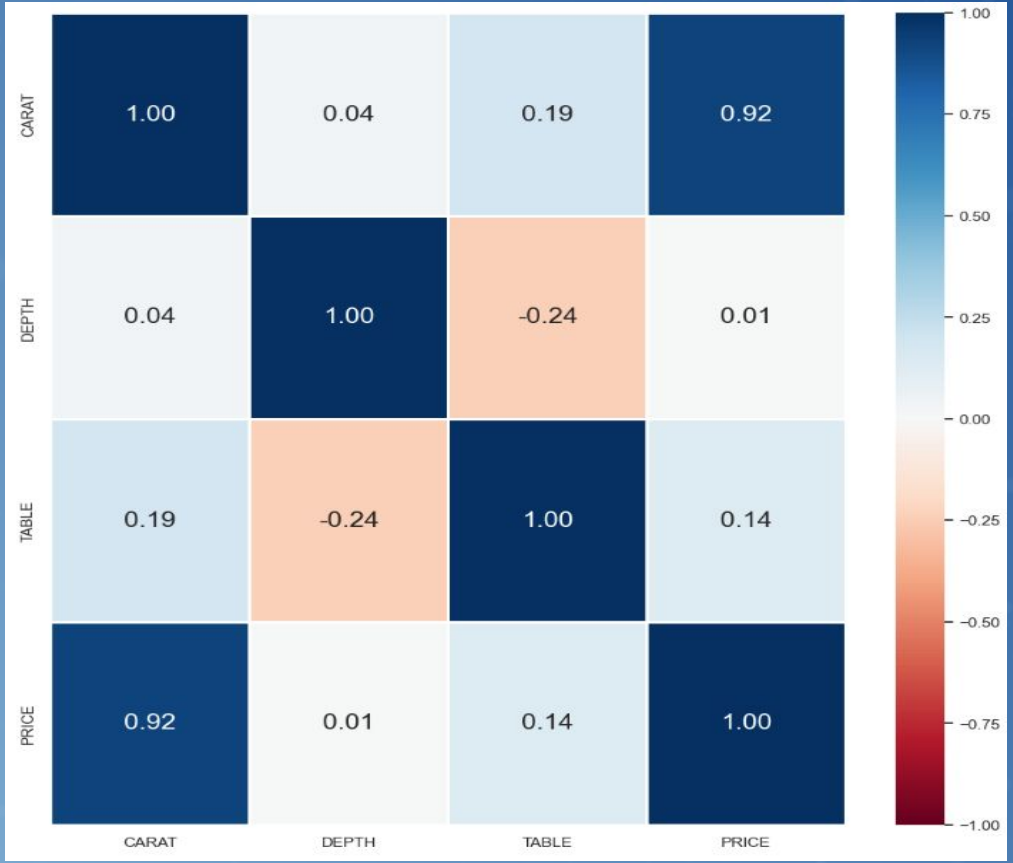
Linear Regression CARAT Test Data

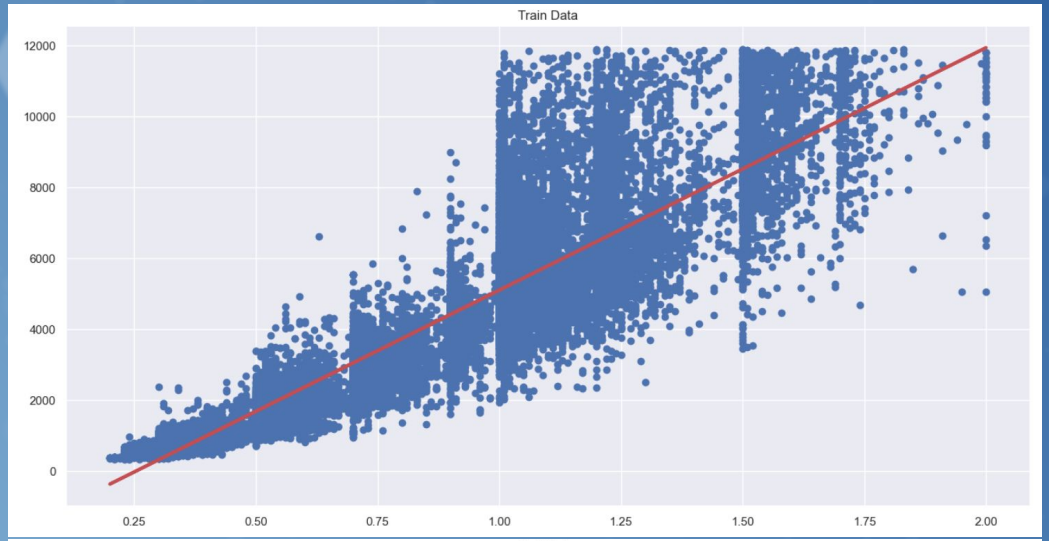
Explained Variance (R^2) :
0.8521152579253212
Mean Squared Error (MSE) :
2393681.009910226
Root Mean Squared Error (RMSE) :
1547.152549010674

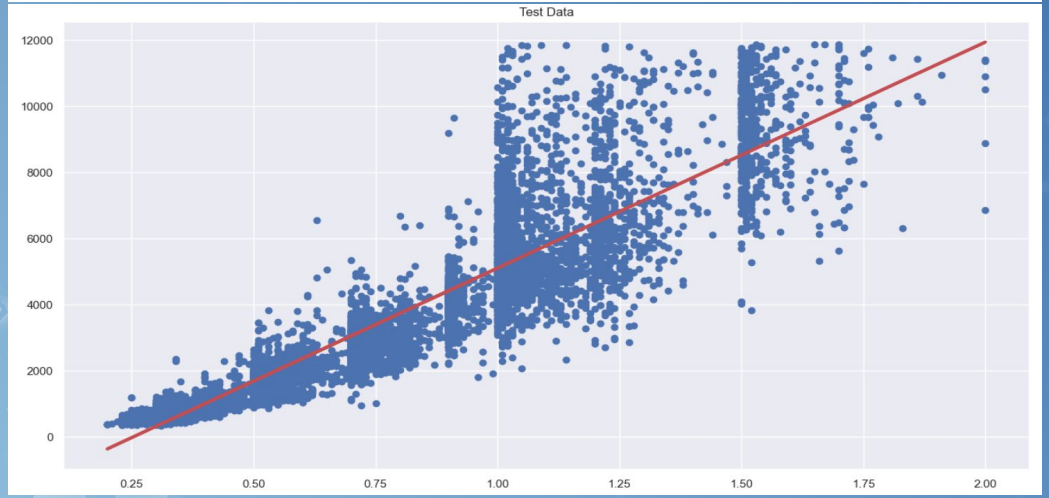
Removing Outliers

- Checking if removing outliers will improve results

Removing Outliers







Linear Regression CARAT Train Data

Explained Variance (R^2) :
0.852749870478203
Mean Squared Error (MSE) :
1117453.6236266212
Root Mean Squared Error (RMSE) :
1057.0967900938026

Linear Regression CARAT Test Data

Explained Variance (R^2) :
0.8474300745340442
Mean Squared Error (MSE) :
1134893.7347619676
Root Mean Squared Error (RMSE) :
1065.313913718378

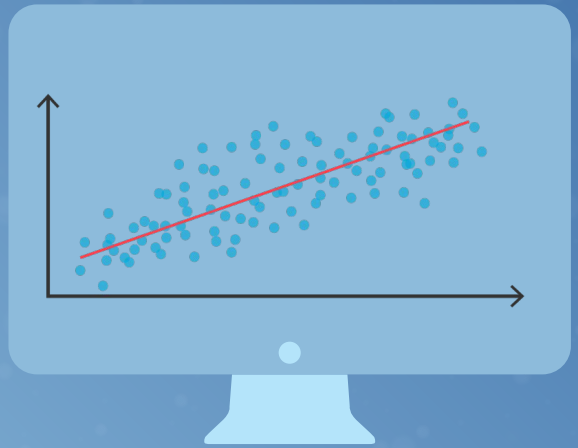
Comparing the Data

Linear Regression with outliers

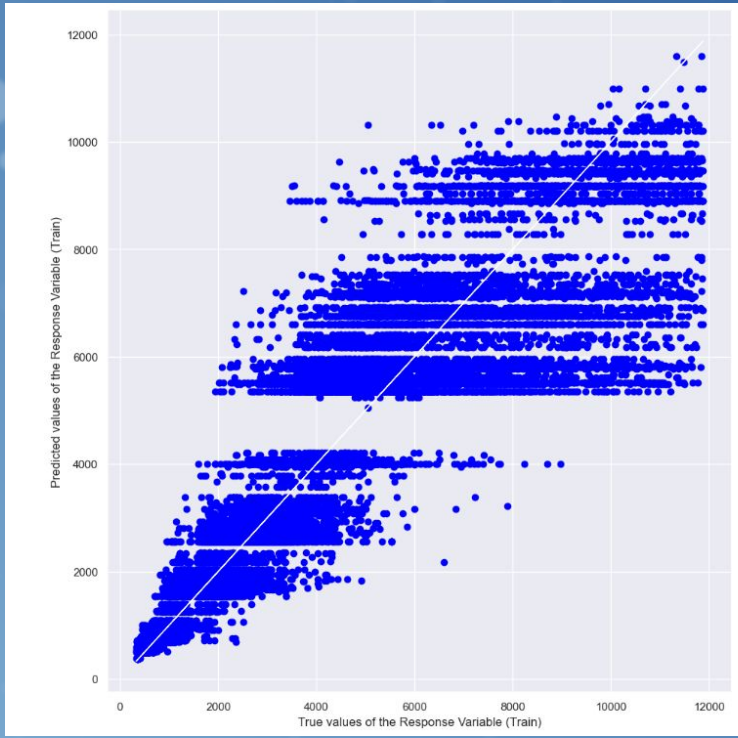
Explained Variance (R^2) :
0.8521152579253212
Mean Squared Error (MSE) :
2393681.009910226
Root Mean Squared Error (RMSE) :
1547.152549010674

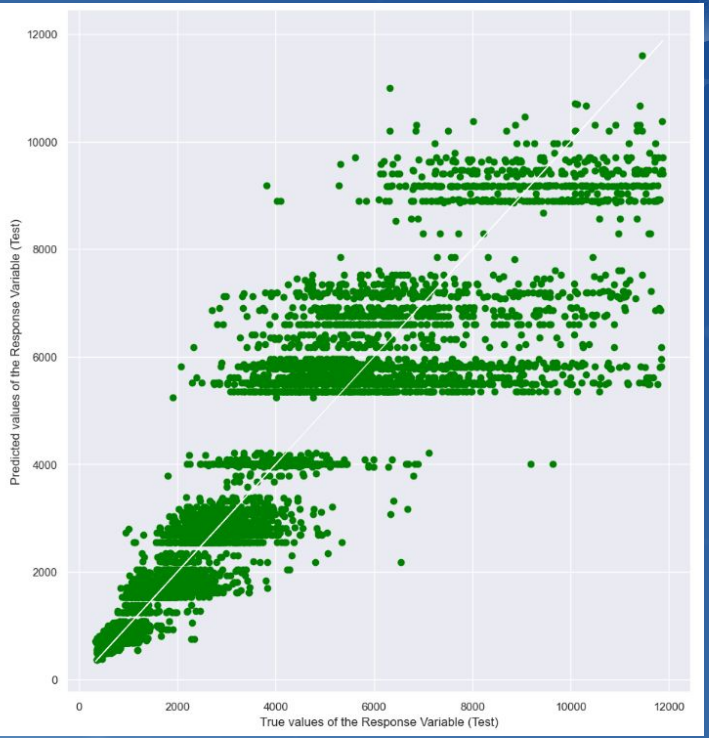
Linear Regression without outliers

Explained Variance (R^2) :
0.8474300745340442
Mean Squared Error (MSE) :
1134893.7347619676
Root Mean Squared Error (RMSE) :
1065.313913718378



Decision Tree Regression Model



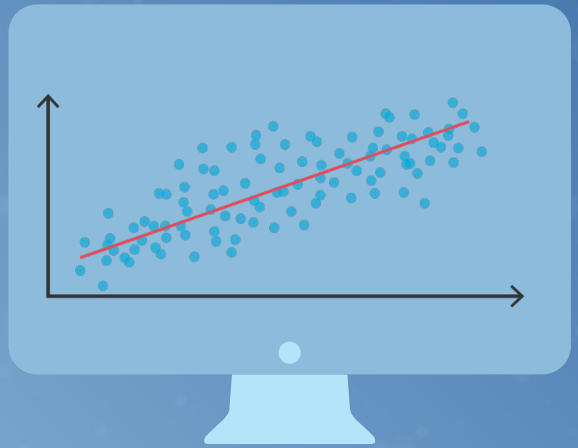


Decision Tree Regression Train Data

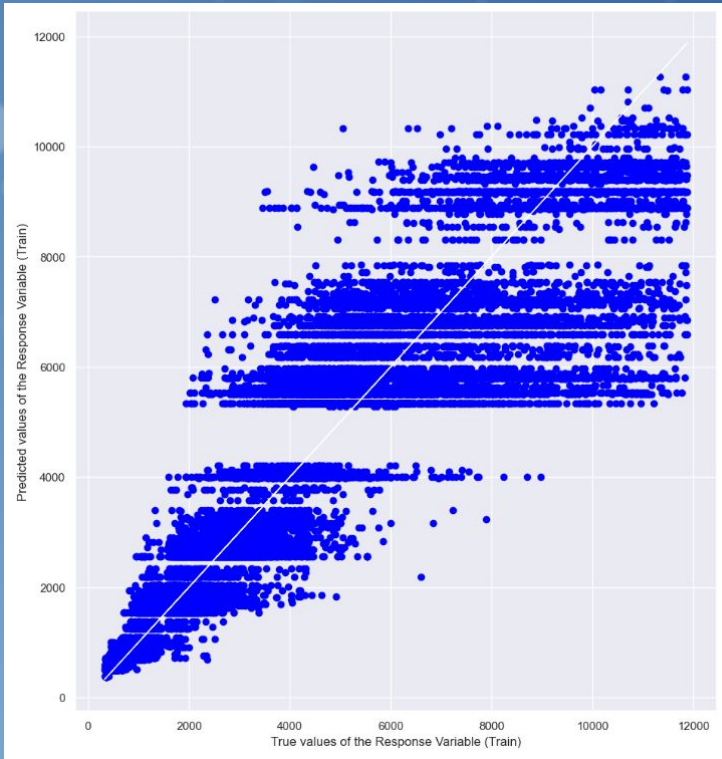
Explained Variance (R^2) : 0.8697151492326016
Mean Squared Error (MSE) : 988707.3041394651
Root Mean Squared Error (RMSE) : 994.3376208006339

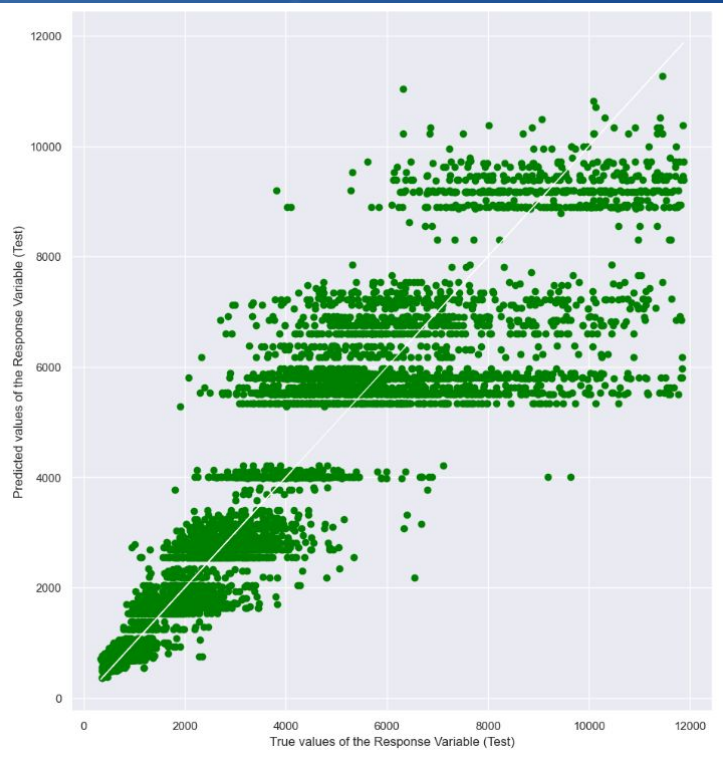
Decision Tree Regression Test Data

Explained Variance (R^2) : 0.8621940177554098
Mean Squared Error (MSE) on Test Set : 1025071.9162670241
Root Mean Squared Error (RMSE) on Test Set : 1012.4583528555751



Random Forest Regression Model





Random Forest Regression Train Data

Explained Variance (R^2) : 0.8696767461579404
Mean Squared Error (MSE) : 988998.7378725141
Root Mean Squared Error (RMSE) : 994.48415667245

Random Forest Regression Test Data

Explained Variance (R^2) : 0.8621856708870331
Mean Squared Error (MSE) on Test Set : 1025134.0045756906
Root Mean Squared Error (RMSE) on Test Set : 1012.4890145456842

Choosing the Best Model

Linear Regression CARAT Test Data

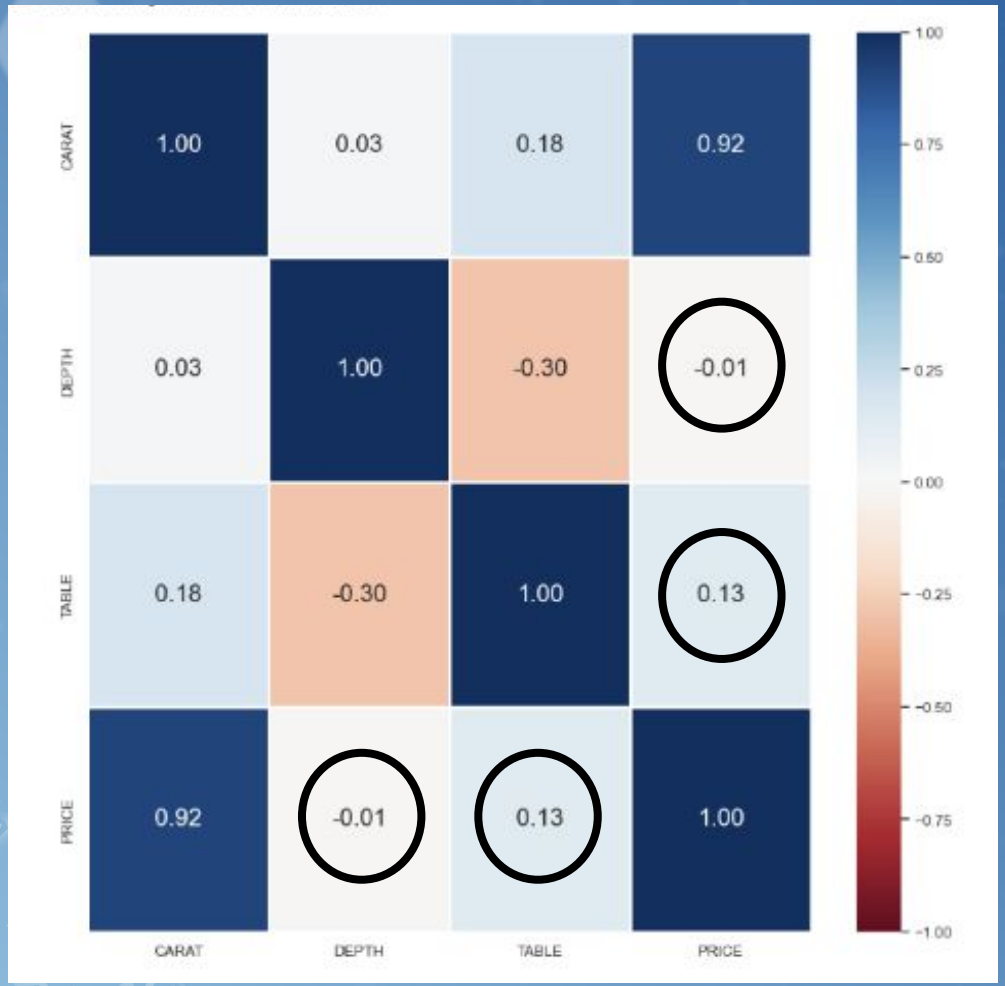
Explained Variance (R^2) : 0.8474300745340442
Mean Squared Error (MSE) : 1134893.7347619676
Root Mean Squared Error (RMSE) : 1065.313913718378

Decision Tree Regression Test Data

Explained Variance (R^2) : 0.8621940177554098
Mean Squared Error (MSE) on Test Set : 1025071.9162670241
Root Mean Squared Error (RMSE) on Test Set : 1012.4583528555751

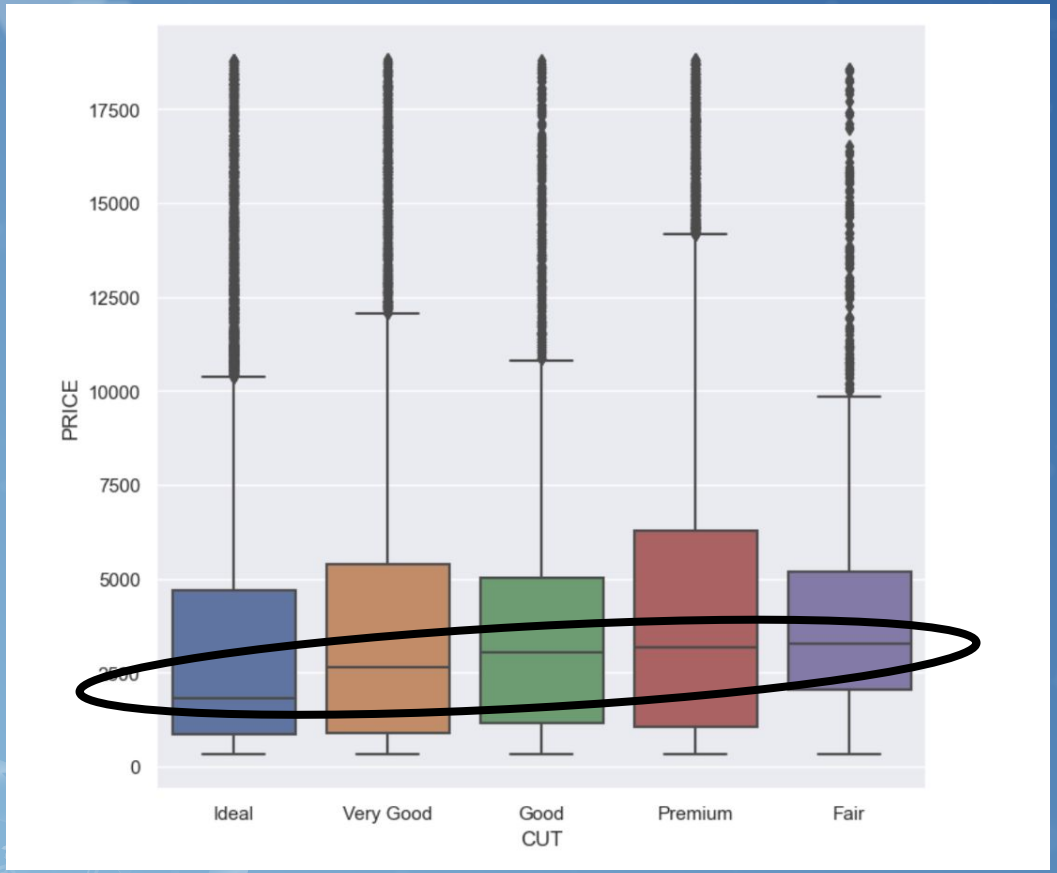
Random Forest Regression Test Data

Explained Variance (R^2) : 0.8621856708870331
Mean Squared Error (MSE) on Test Set : 1025134.0045756906
Root Mean Squared Error (RMSE) on Test Set : 1012.4890145456842



New Insights

- Depth and Table very low correlation with price
- These 2 variable are very insignificant and do not any impact on the prices of diamond



New Insights

- Cut has not much variation with price
- The median prices of each level does increases but not significantly

Conclusion

Carat

- High positive correlation
- Decision Tree Regression Model is the best model to use to predict price.

Color & Clarity

- High variation with prices
- Color J and Clarity SI2 have the highest median

THANK YOU!