DIAMOND PRICE PREDICTION

ANJALI YADAV
DHANUSH GOWDA
KARTHIK M K

ATHARVA AGRAWAL
HARSHAD BANDI
YESHPAL SINGH

18-NOVEMBER-2022

INDEX

PROBLEM STATEMENT

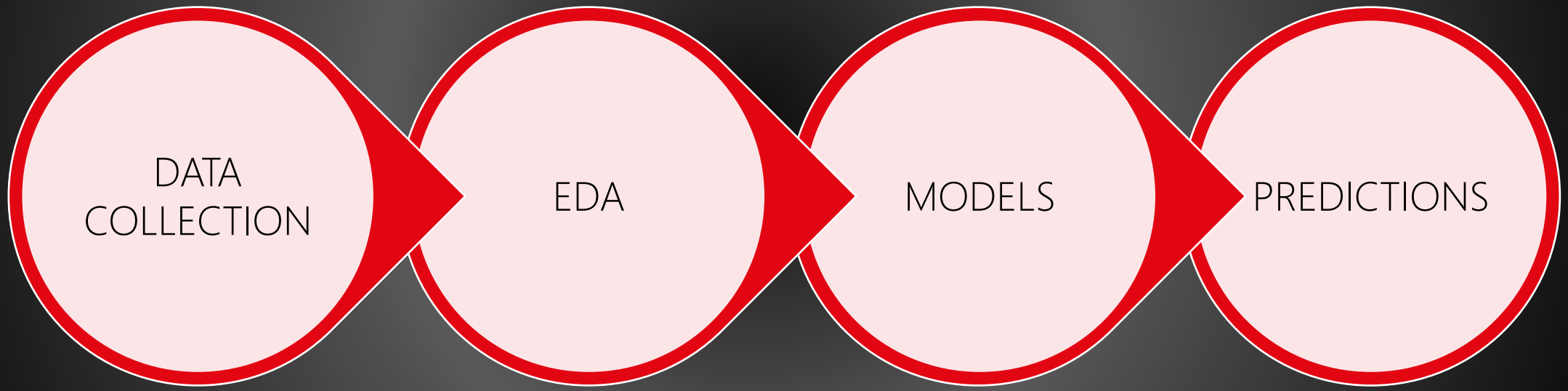
We have datapoints for 53,940 across 10 features of diamonds. In this supervised model we have been provided with the target column which is the Price of the diamond. The goal is to predict the price of Diamond using different Regression Algorithms.

FEATURES

CARAT	• Unit of weight for Diamond
CUT	• The cut type of the Diamond, it determines the shine
COLOR	• Hue of a Diamond based on the GIA's color scale
CLARITY	• Visual appearance of Diamond in qualitative metrics

DEPTH	• The value of how deep or shallow the Diamond is
TABLE	• The flat surface on very top of the stone
LENGTH	• Length of the Diamond
WIDTH	• Width of Diamond
HEIGHT	• Height of Diamond

METHODOLOGY



DATA COLLECTION

The Diamond price prediction data chosen for this project is taken from Kaggle. We used pandas read_csv() function to load the data to the notebook. This dataset comes with both categorical and numerical data which has been cleaned and processed to build the model.

Source : Data Source

EXPLORATORY DATA ANALYSIS

1. Missing value treatment : We used isnull() function on our dataset to find the missing values.
2. Outlier treatment : We checked the outliers in the data through boxplot.
3. Linearity : We checked the data linearity through Pairplot.
4. Normality : We used Displot to find the normality of target variable.
5. Count : We used value_counts() to get the count of attributes
6. Numerical data conversion : We used LabelEncoder to convert the categorical columns to numerical data.

MODEL BUILDING

- We divided the data as train and test set using train_test_split() function.
- We imported the required libraries for the models used
- Models built were:
  - Linear Regression
  - OLS
  - Decision Tree
  - Random Forest
- Models were evaluated using r2_score

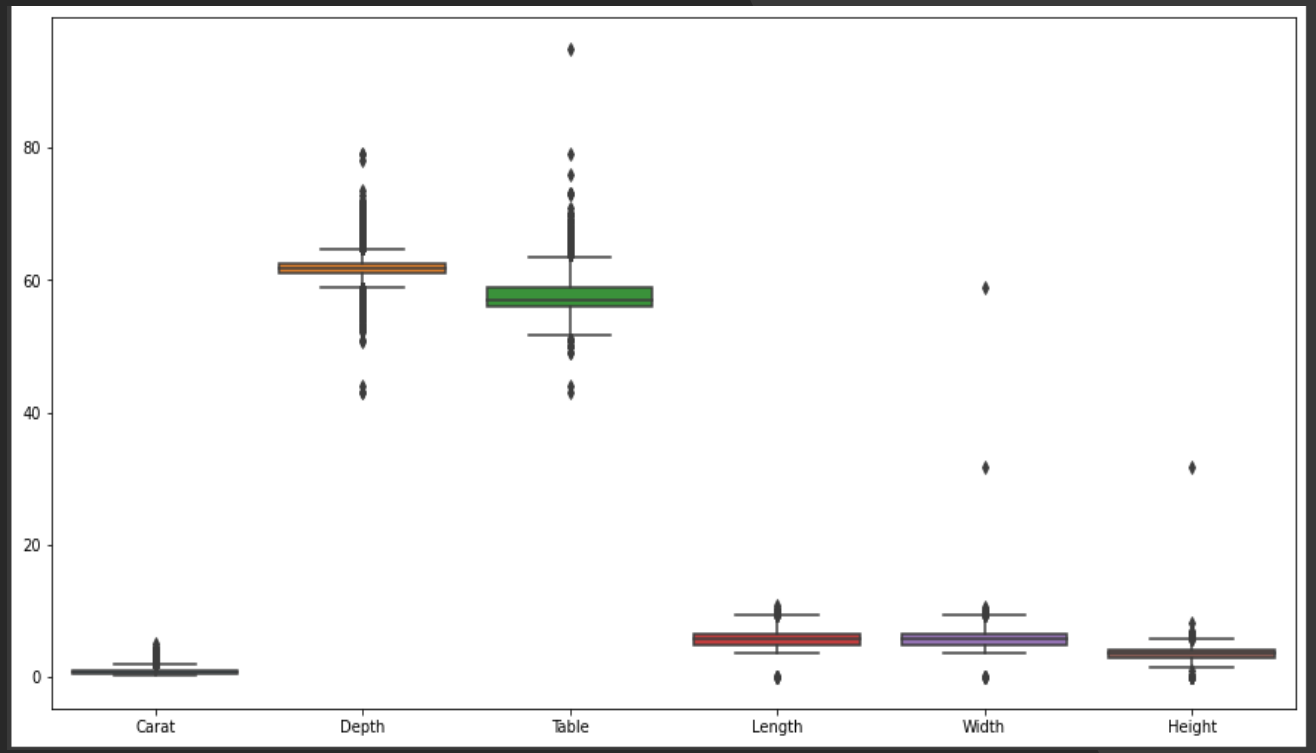
PREDICTIONS

- We have tabulated the result metrics of each model
- We have compared first 10 points of target column price and compared it with model predictions.

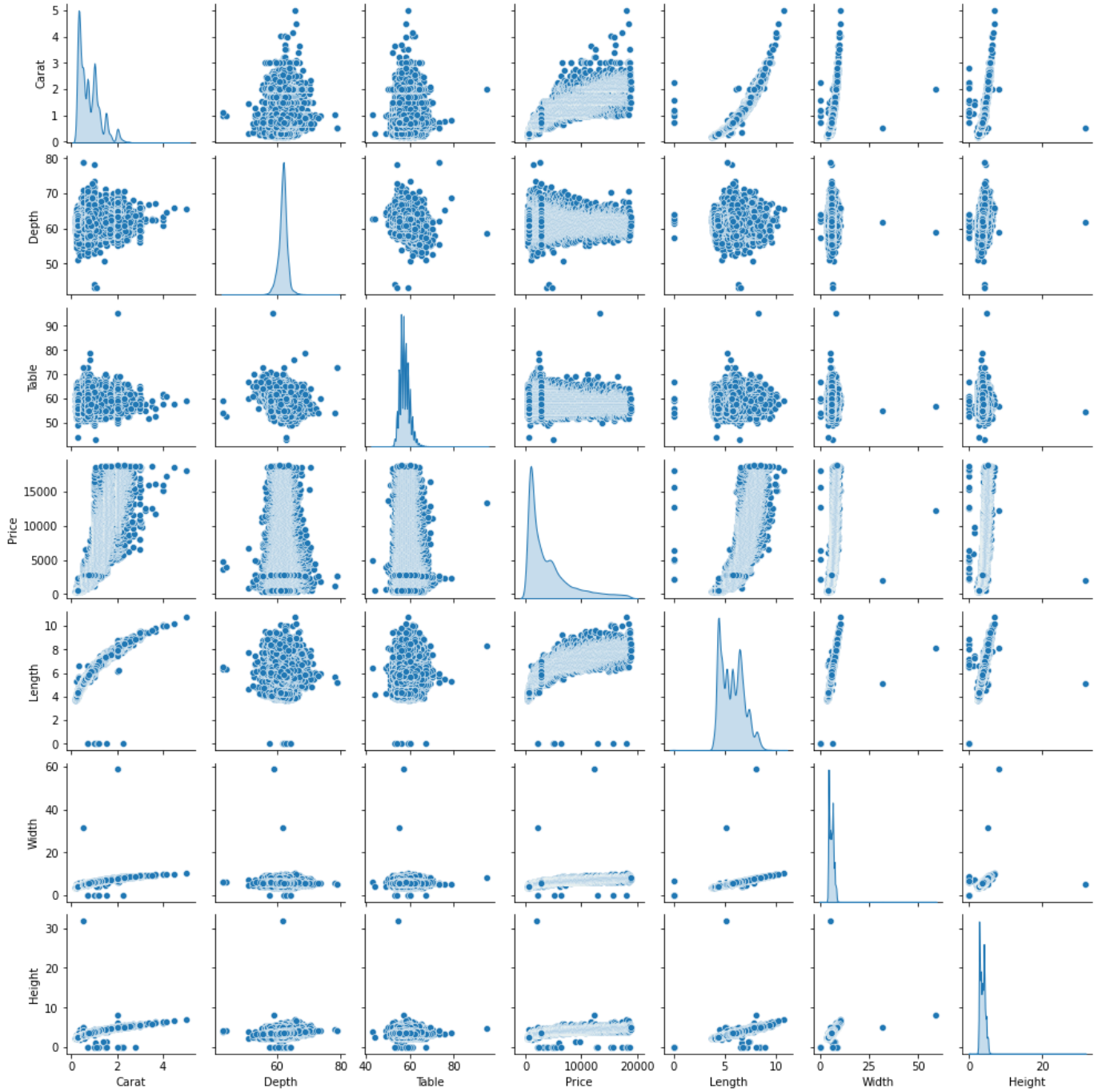
RESULTS

EDA

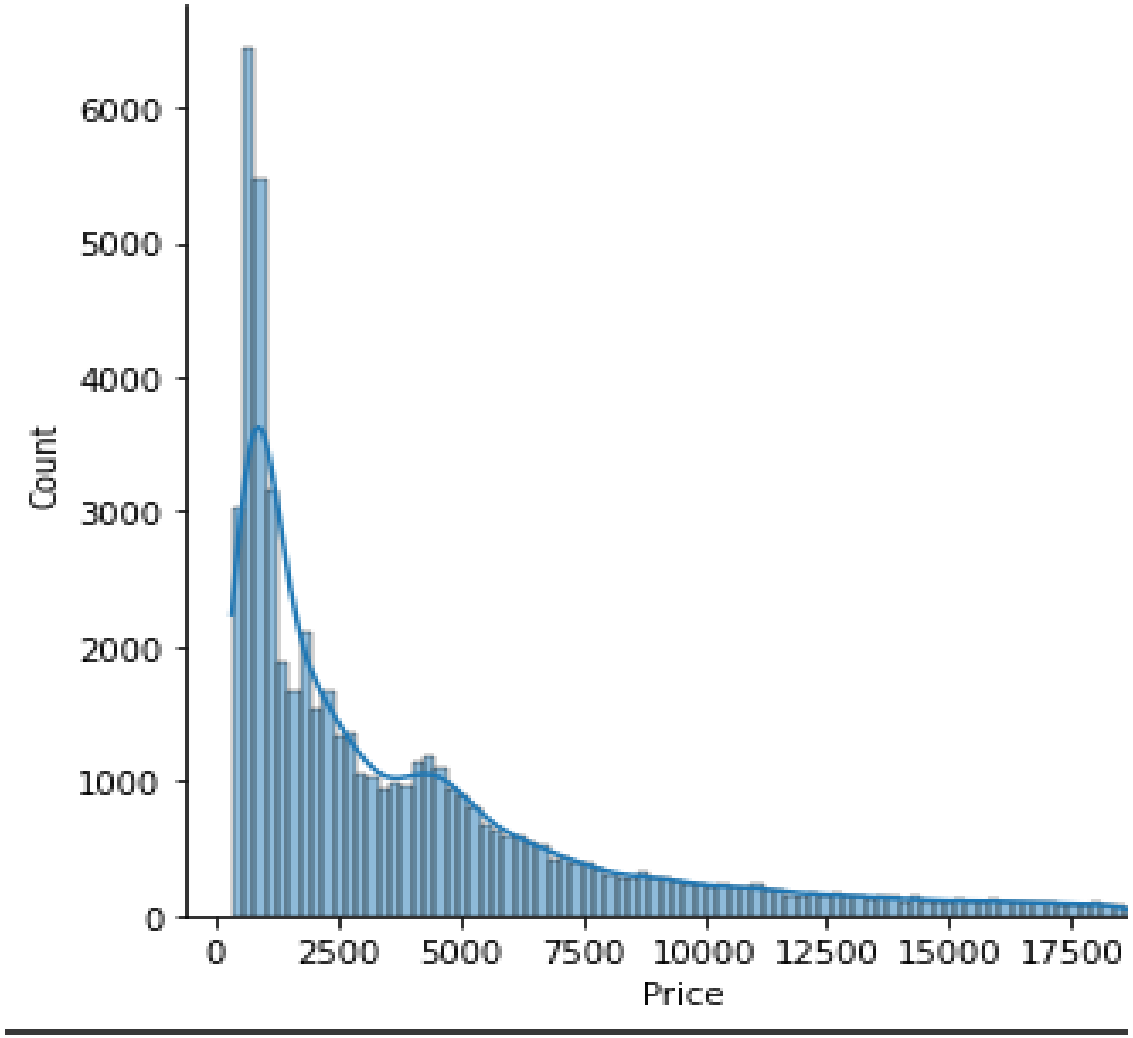
1. The Dataset did not contain any missing values.
2. Even though the boxplot shows many outliers we have taken the data as it is for the model building, as the diamond features may vary with different types.
3. We have converted categorical columns into numerical using LabelEncoder



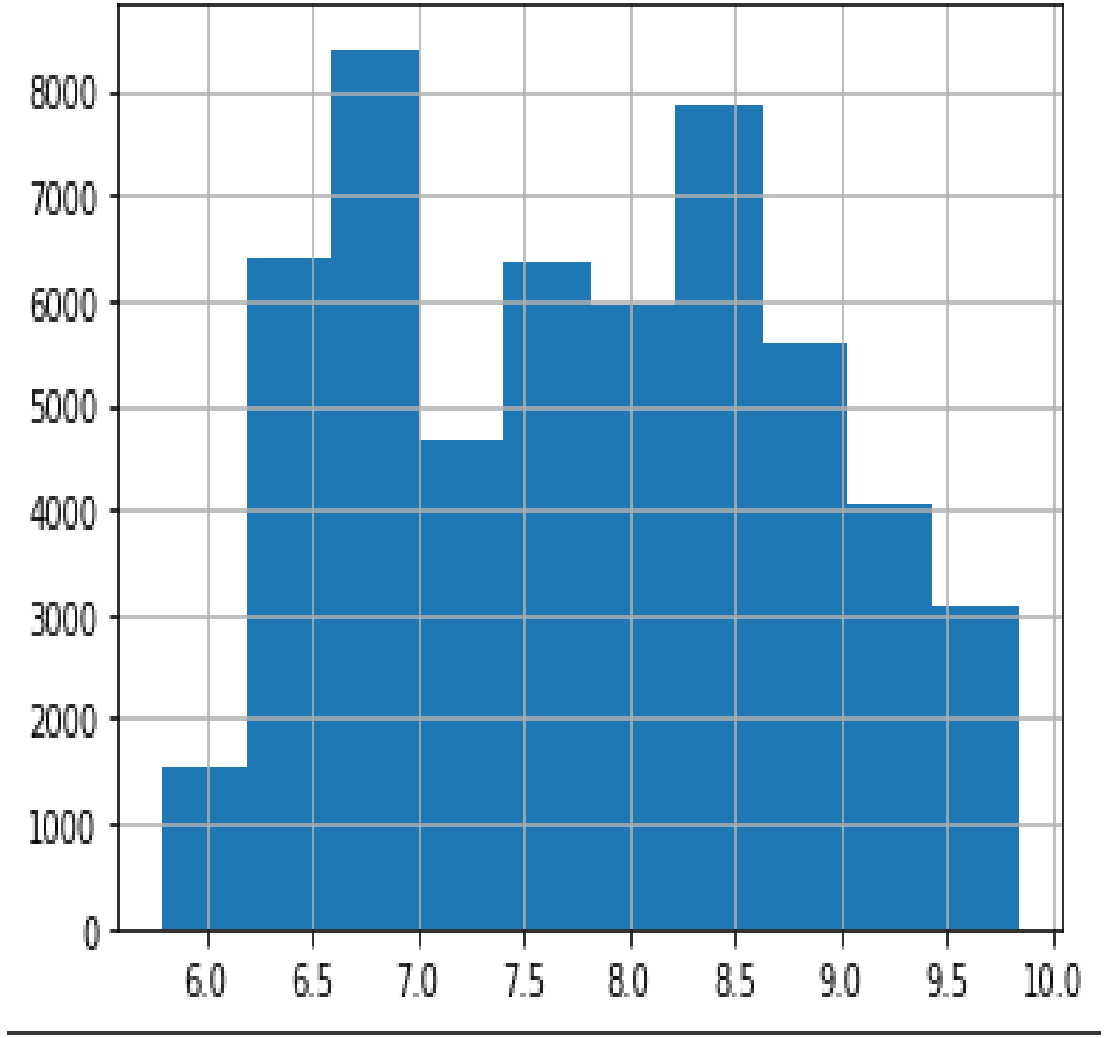
4. From the below pair plot we can say that the data is linear



Before logging



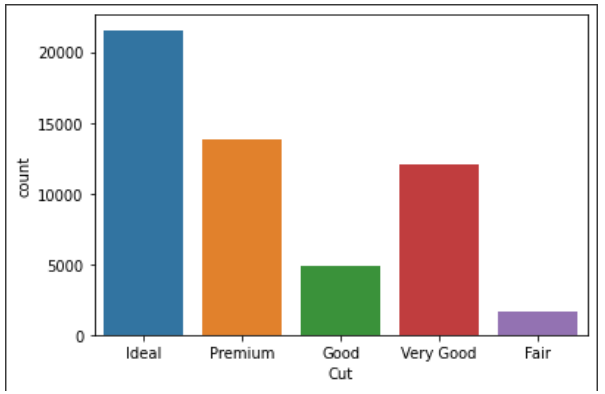
After logging

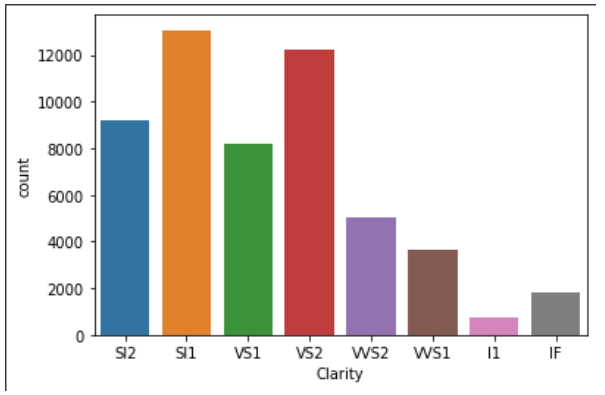


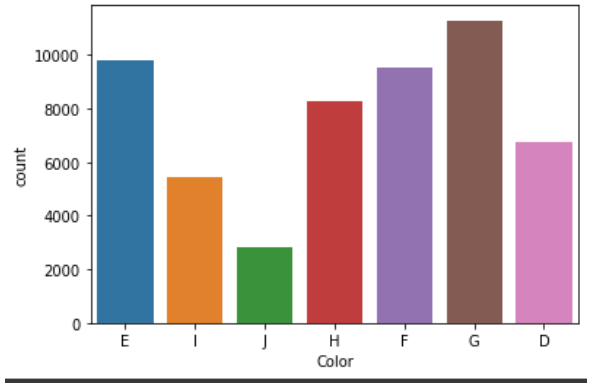
5 . We transferred target column price from skewed data to normal data

6. Counts of Categorical columns

1. In the cut feature of Diamond Ideal cut has highest value count with 21551
2. In the color feature of Diamond G type has highest count with 11292
3. The clarity feature of diamond has SI1 type with highest count of 13065

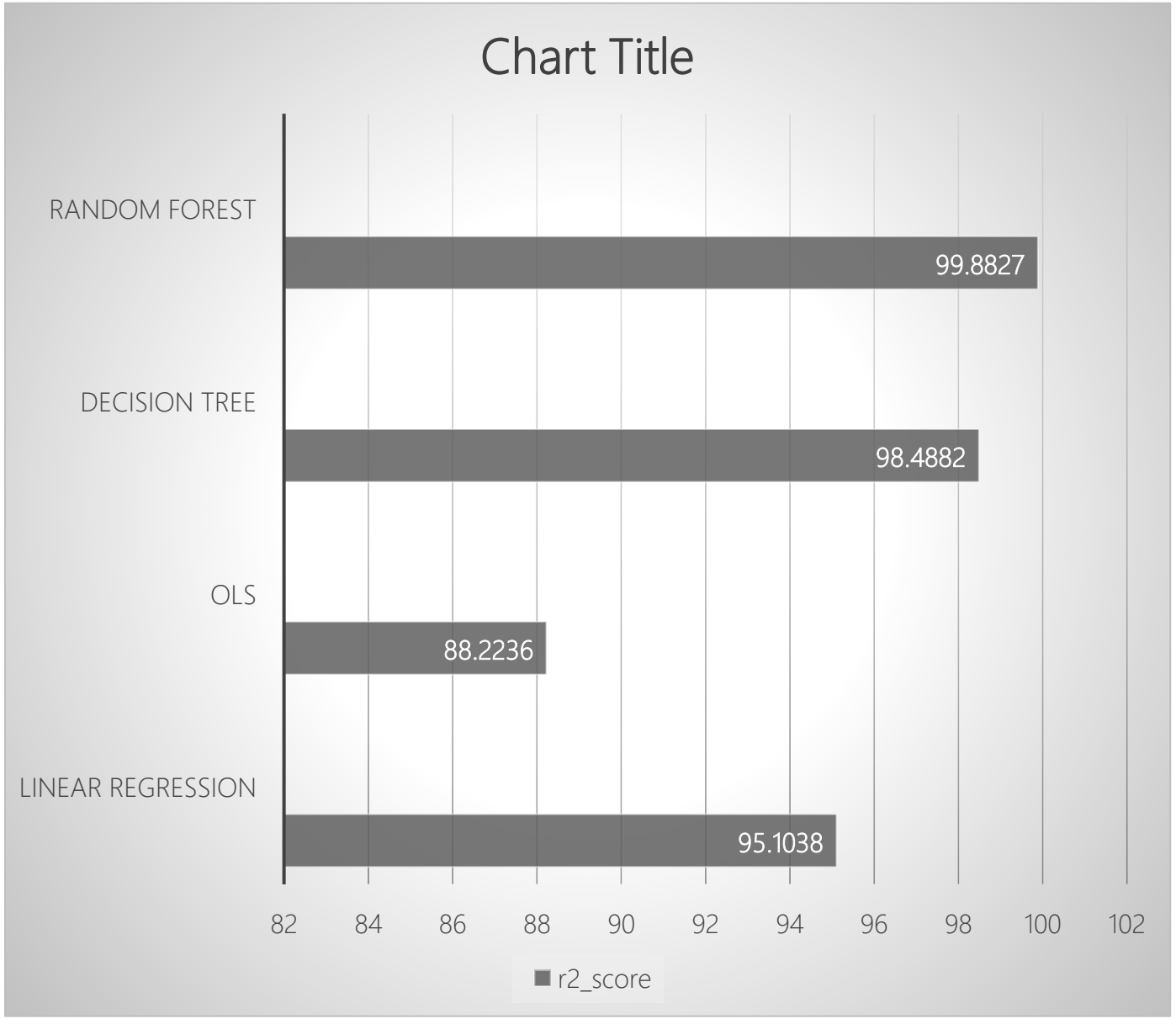






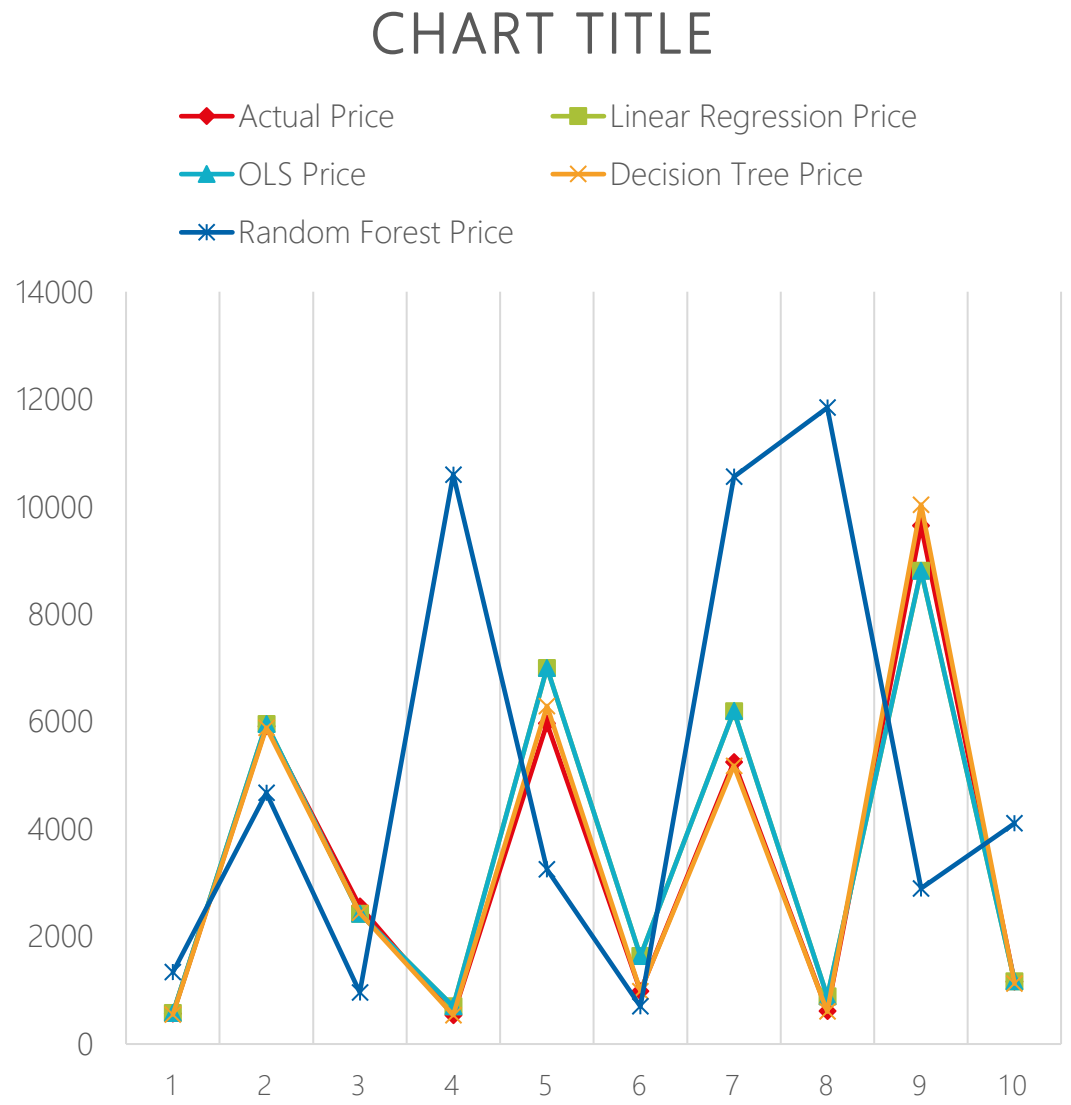
Models

- We have built 4 regression models i.e., Linear Regression, OLS, Decision Tree, Random Forest



Predictions

- We can observe the predictions of 1st 10 data points in the linear graph shown below



CONCLUSION

- 53,940 data points from 10 columns have been processed to build the regression models.
- Ideal cut of diamonds are used or sold more in the market
- G colored diamonds are more frequent
- Diamond with SI1 clarity have appeared more in the dataset
- With 99.88% r2_score we can conclude that Random Forest Regressor give the best model.
- We have taken 10 prices from each model and compared it to actual prices. Even though random forest is displaying varied prices in the 1st 10 points, it has very low error rate when processed through 50 odd thousand data points.


THANK YOU