# **Project Summary**

Contents

[**Project Summary** 1](#_Toc25797449)

[**Business Problem** 2](#_Toc25797450)

[**Dataset** 2](#_Toc25797451)

[Structure of Dataset 2](#_Toc25797452)

[Dummy Variables 2](#_Toc25797453)

[Missing Values 2](#_Toc25797454)

[Normalization 2](#_Toc25797455)

[**EDA** 2](#_Toc25797456)

[Package 2](#_Toc25797457)

[**Model Building** 2](#_Toc25797458)

[Splitting 2](#_Toc25797459)

[Vanilla Model (Model – 1) 3](#_Toc25797460)

[Model -2 3](#_Toc25797461)

[VIF/AV Plots 3](#_Toc25797462)

[Model - 3 3](#_Toc25797463)

[Model Deletion Diagnostics 3](#_Toc25797464)

[Model – 4 5](#_Toc25797465)

[Model - 5 5](#_Toc25797466)

[Step AIC 5](#_Toc25797467)

[Model - 6 6](#_Toc25797468)

[**Predictions** 6](#_Toc25797469)

[RMSE 6](#_Toc25797470)

## **Business Problem**

Dataset consists of **53,940 observations** and **27 variables** of various diamonds.

Our task is to predict the price of a diamond by using Multivariate Regression Technique.

## **Dataset**

### Structure of Dataset

The dataset has 3 categorical features (Cut Quality, Diamond color and Clarity)

### Dummy Variables

In order to perform Multiple Linear regression, we needed to convert these variables into dummy variables and then add those dummy variables in the dataset excluding the categorical features.

### Missing Values

We then check for **Missing values** but there is **none**.

### Normalization

Data is then normalized and then prepared for further analysis.

## **EDA**

EDA is done and is presented in a different report.

### Package

We have used dlookr package from R for the EDA part which provides us with the correlation and the information about normality between the features.

## **Model Building**

After the pre-processing and EDA, we proceed for Model building

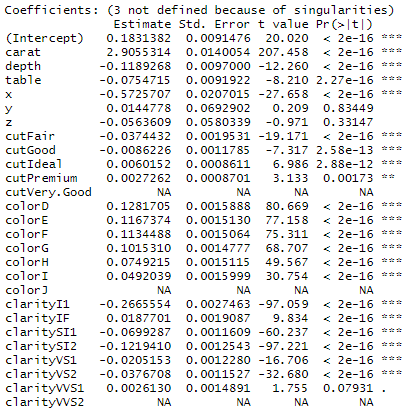
### Splitting

Initially the data is splitted into Train (80%) and Test (20%).

Model that we are going to use is **Multiple Linear Regression** as the target variable is continuous and there are multiple input variables.

### Vanilla Model (Model – 1)

We build the initial model and get R2 value as 0.9194 which indicates that the model is very good and most of the features are significant except y,z and clarity WS1



### Model -2

Removing those variables and building the 2nd model we get the same R2 value as 0.9194

All the features are significant

### VIF/AV Plots

Now we check for the VIF and find that there is a collinearity issue within features.



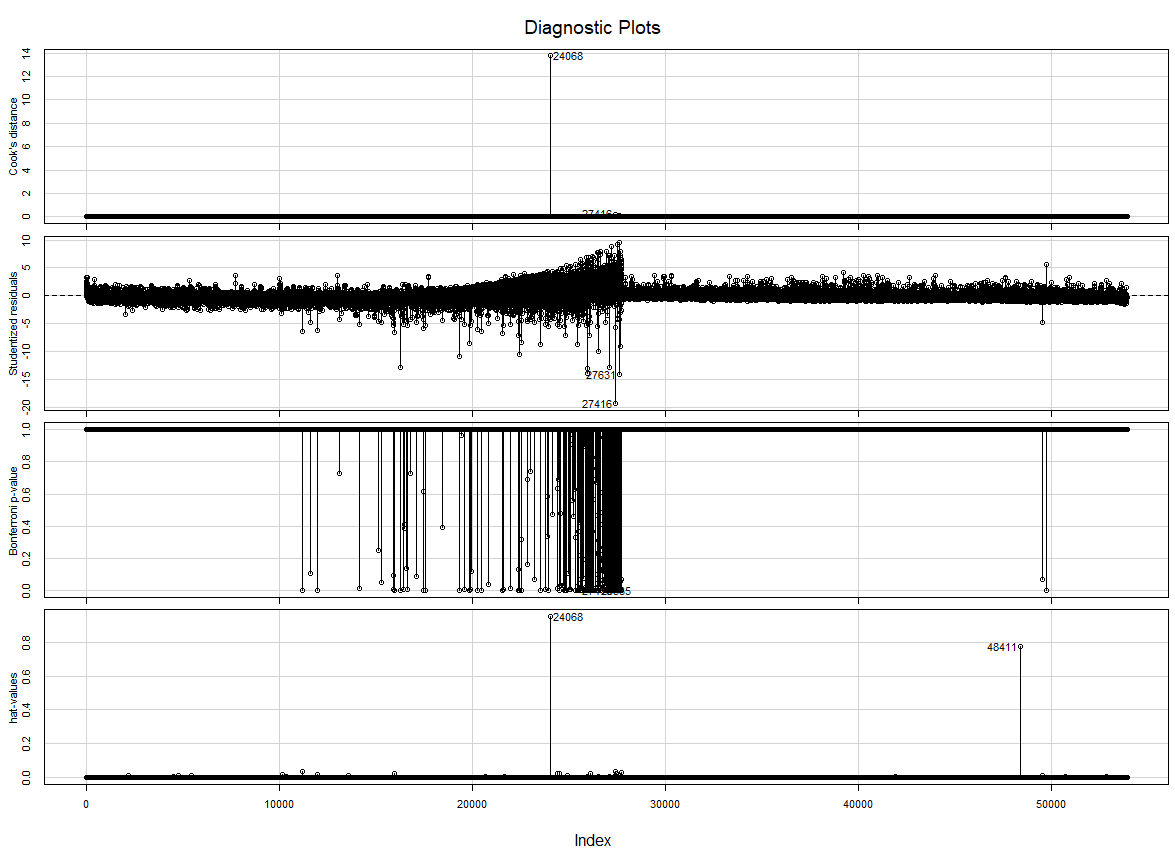
After doing analysis with VIF and AV Plots, we find that X has highest VIF value and can be deleted.

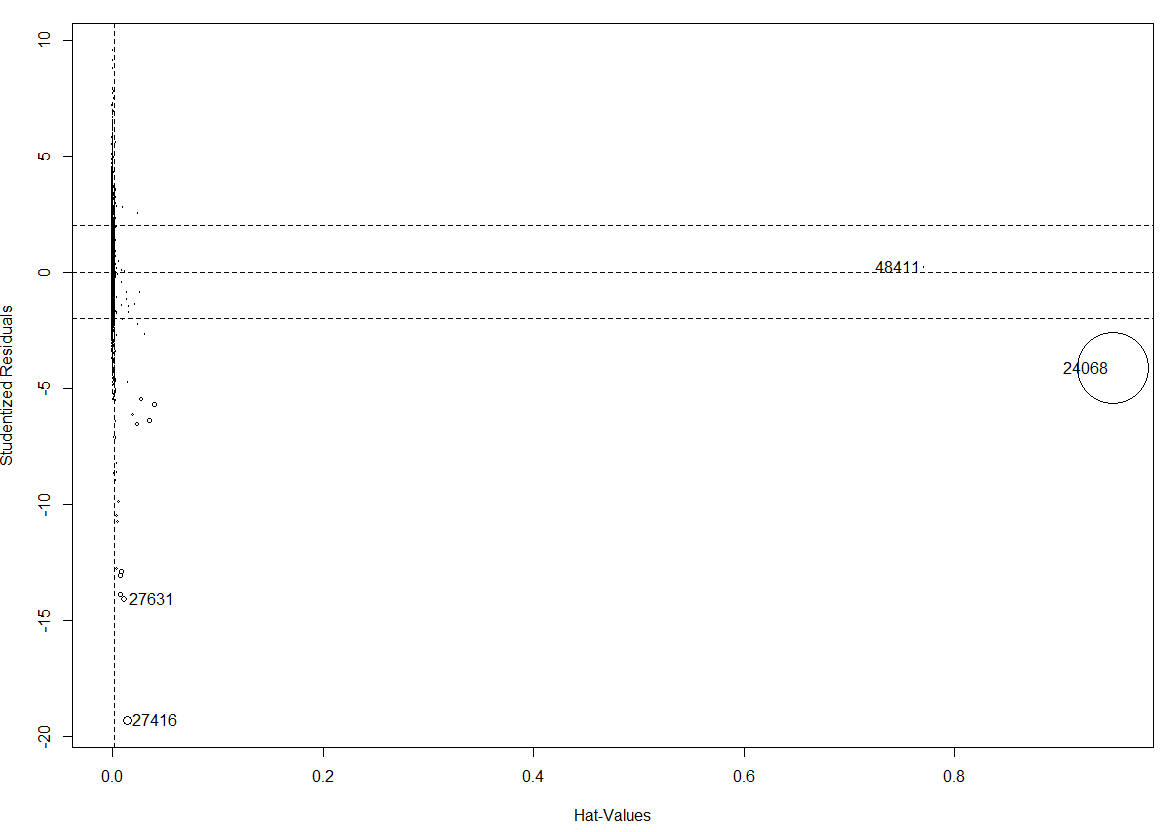
### Model - 3

Building the model after this provided R2 value as 0.9158

### Model Deletion Diagnostics

Since the R2 value has decreased, we decide to retain the features and try dropping influential observations from the model



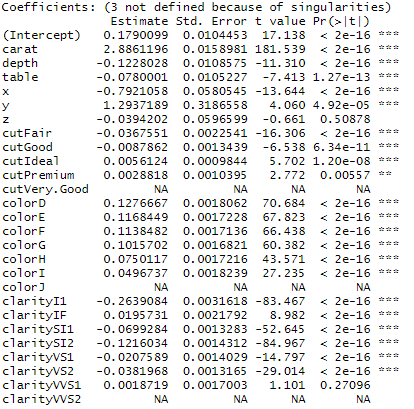


We have identified 3 influential observations - 24068, 27631 and 27416 based on Cook Distance and Hat Values.

Since the observation is from Test data which is obtained after splitting the original dataset, hence we remove a range of observations from 15000 to 25000 and thus we have removed these observations.

### Model – 4

We build a new model and now obtained R2 as 0.9194



We can see that the influential observations are deleted but there are some insignificant features also so removing those features and making the final model

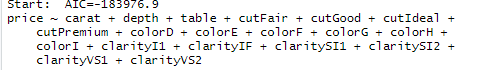
### Model - 5

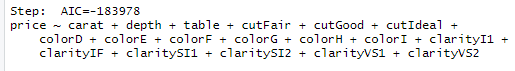
The final model has a R2 value of 0.9158

### Step AIC

Now we run a stepAIC function to check the Akaike Information Criterion of a particular model.

The lower the AIC value, the better is the model.





We find that the final model without cutPremium feature is the better model as it is an insignificant feature.

### Model - 6

Building the improved model

The R2 value is 0.9158

## **Predictions**

We perform the predictions on the test data and then calculate the RMSE of the model.

### RMSE

RMSE value is 0.0623

**--- End of Report ---**