ASSIGNMENT 3

By,

ARUL VINCY MARY

INTRODUCTION

- A dataset containing the prices and other attributes of almost 54,000 diamonds.

- There are 10 variables measuring various pieces of information about the diamonds.

Variable description:

Variables	description	categories
Carat	weight of the diamond (0.2--5.01)	-
Cut	cut quality of the diamond	Fair, Good, Very Good, Premium, Ideal (quality in increasing order)
Color	color of the diamond	D, E, F, G, H, I , J (D being the best and J the worst)
Clarity	how obvious inclusions are within the diamond	I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)
depth	depth % - the height of a diamond, measured from the culet to the table, divided by its average girdle diameter	-

variables	description	categories
Table	table% - the width of the diamond's table expressed as a percentage of its average diameter	-
Price	price of the diamond in US dollars	-
x	length of the diamond in mm (0--10.74)	-
y	width of the diamond in mm (0--58.9)	-
z	depth of the diamond in mm (0--31.8)	-

Problem statement:

- Develop an algorithm to predict the price of the diamonds with their attributes.

About the data:

The 4 Cs of Diamonds:-

- **carat (0.2--5.01):** The carat is the diamond’s physical weight measured in metric carats. One carat equals 1/5 gram and is subdivided into 100 points. Carat weight is the most objective grade of the 4Cs.
- **cut (Fair, Good, Very Good, Premium, Ideal):** In determining the quality of the cut, the diamond grader evaluates the cutter’s skill in the fashioning of the diamond. The more precise the diamond is cut, the more captivating the diamond is to the eye.
- **color, from J (worst) to D (best):** The colour of gem-quality diamonds occurs in many hues. In the range from colourless to light yellow or light brown. Colourless diamonds are the rarest. Other natural colours (blue, red, pink for example) are known as "fancy,” and their colour grading is different than from white colorless diamonds.
- **clarity (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)):** Diamonds can have internal characteristics known as inclusions or external characteristics known as blemishes. Diamonds without inclusions or blemishes are rare; however, most characteristics can only be seen with magnification.

Dimensions:

- x length in mm (0--10.74)
- y width in mm (0--58.9)
- z depth in mm (0--31.8)

depth (43--79) :

- The depth of the diamond is its height (in millimetres) measured from the culet (bottom tip) to the table (flat, top surface).

Table (43--95):

- A diamond's table refers to the flat facet of the diamond seen when the stone is face up. The main purpose of a diamond table is to refract entering light rays and allow reflected light rays from within the diamond to meet the observer’s eye. The ideal table cut diamond will give the diamond stunning fire and brilliance.

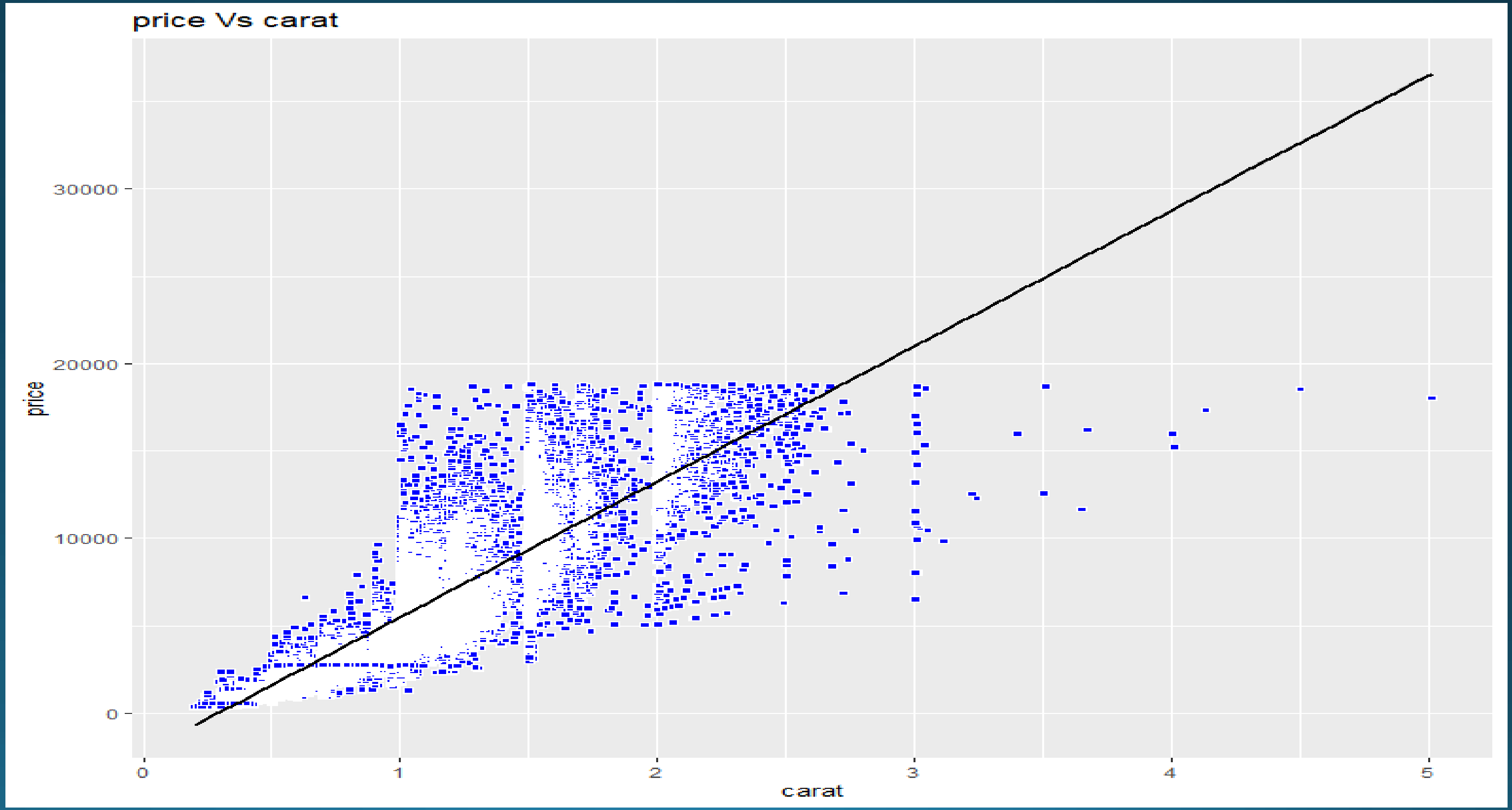
Data pre-processing & Exploratory data analysis:

- Read the data.
- See the structure, dim of the data.(53940,11)
- Take the copy of the dataset.
- Check the missing values and there is no missing values.
- In this dataset , the first column is an index (X) and thus we have to remove it. Now , we have 10 variables.
- To see the min, max of the dataset we have to use summary function.
- Min value of "x", "y", "z" are zero this indicates that there are faulty values in data that represents a dimensionless or 2-dimensional diamonds. So we need to filter out those as it clearly faulty data points. Now, we have 53920 rows and 10 columns.

Identify the variables & plotting:

Carat:

- From the scatterplot, we can see the regression.
- We can use this variable to build the model.
- It has some outliers too.



Depth:

- We can use depth for the price prediction model.
- But still it has few outliers.

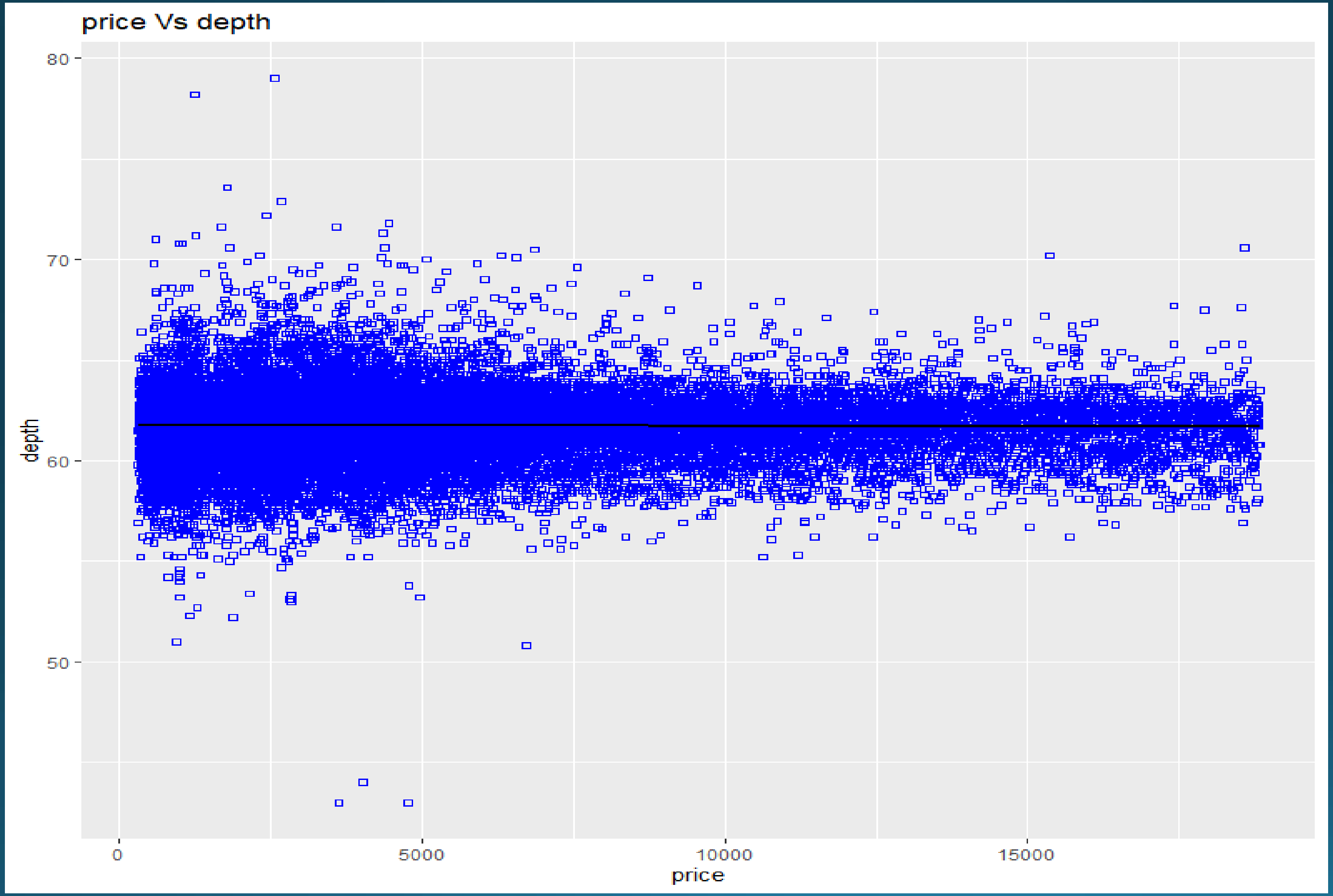
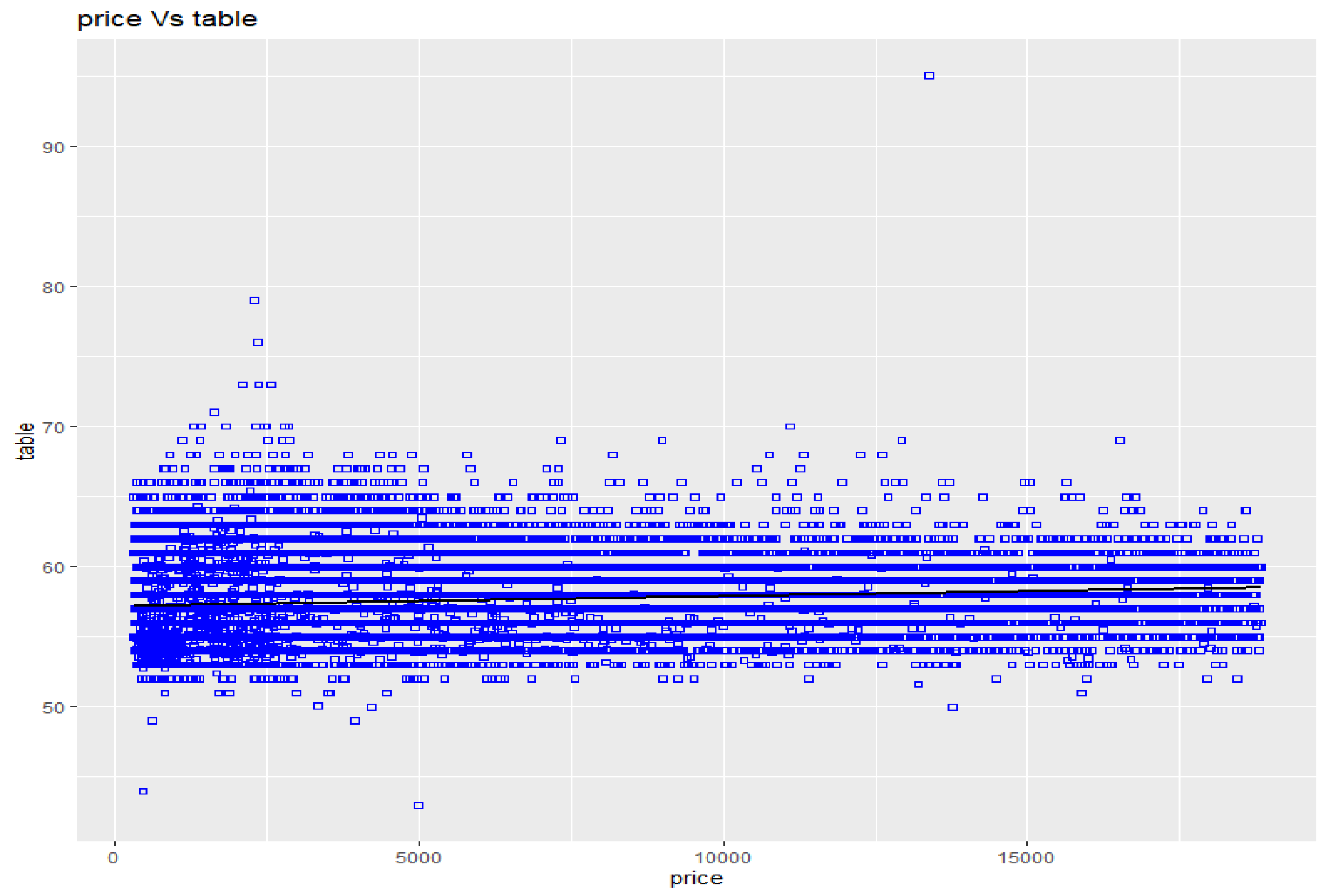


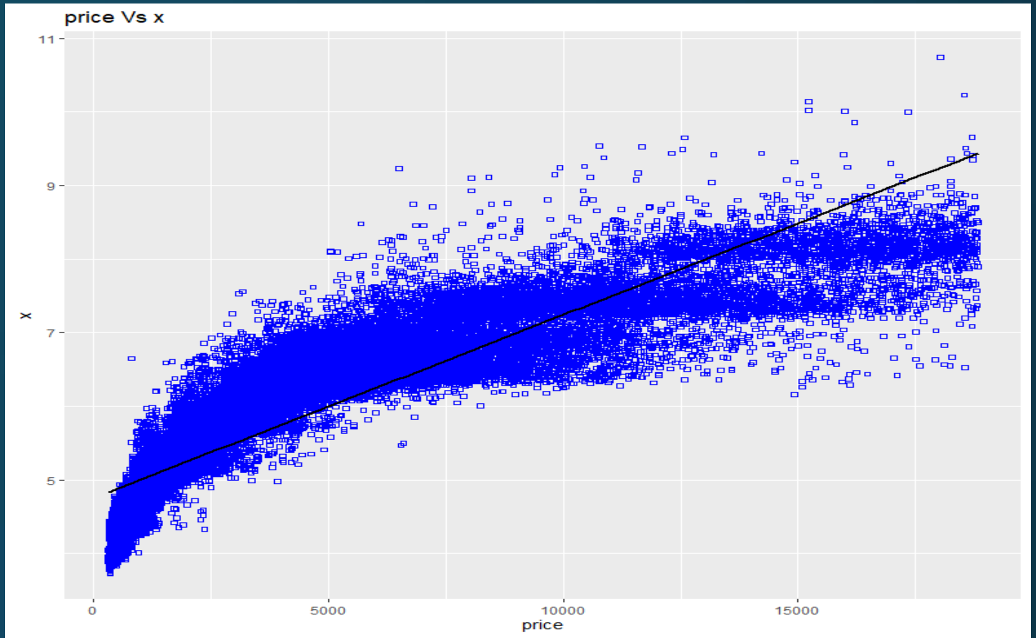
Table:

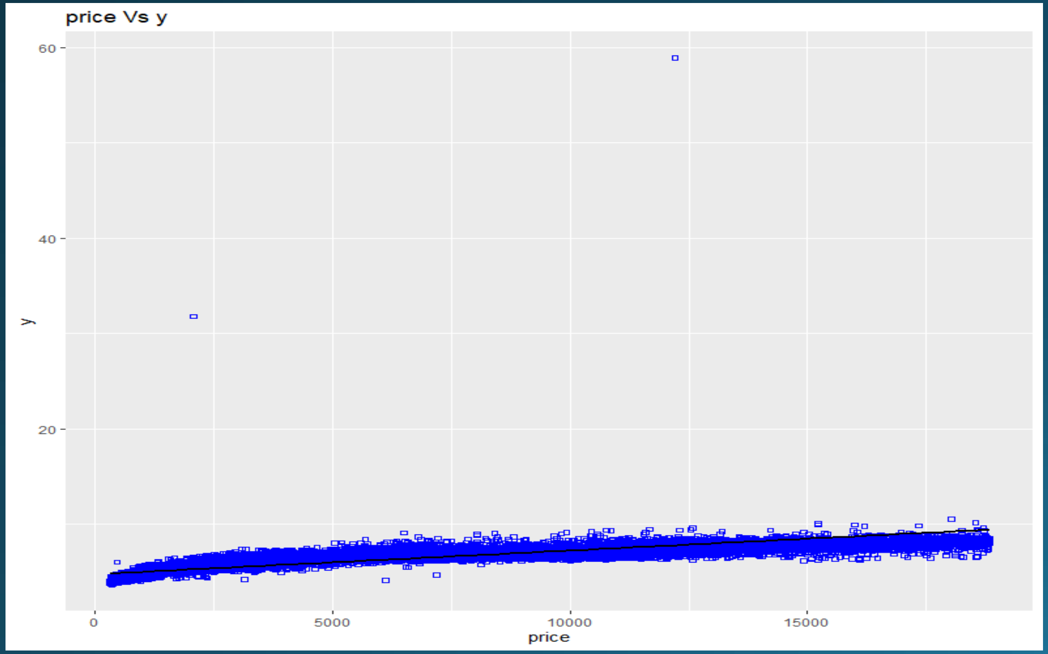
- The variable table showing the linearity with high outiers.
- We have to drop it and use for model building.

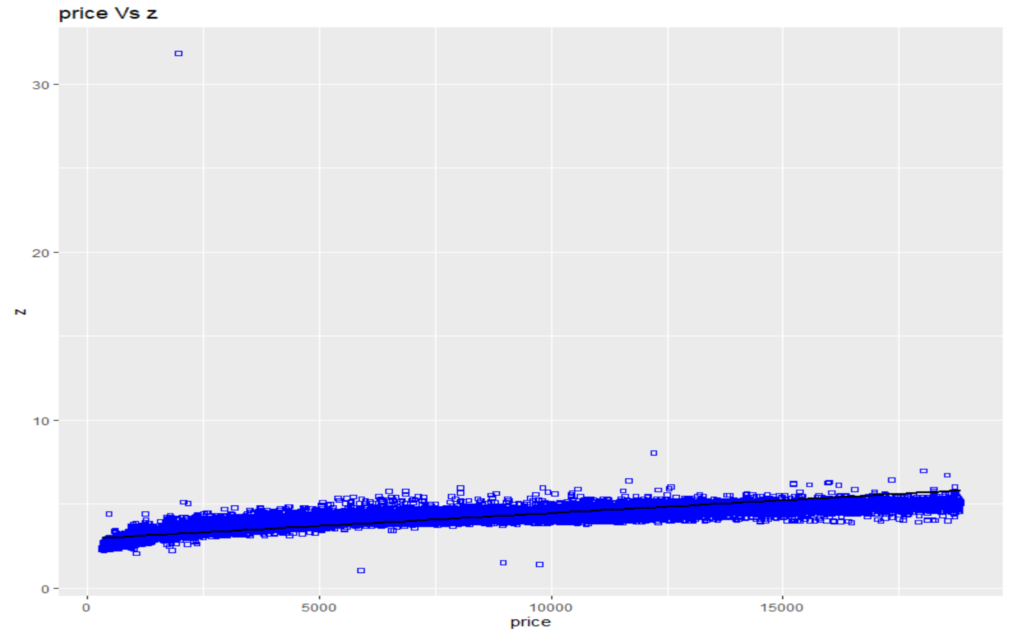


x,y,z:

- "y" and "z" have some dimensional outlies in our dataset that needs to be eliminated. After drop the Outliers by setting the working range we have 53907 rows and 10 columns.

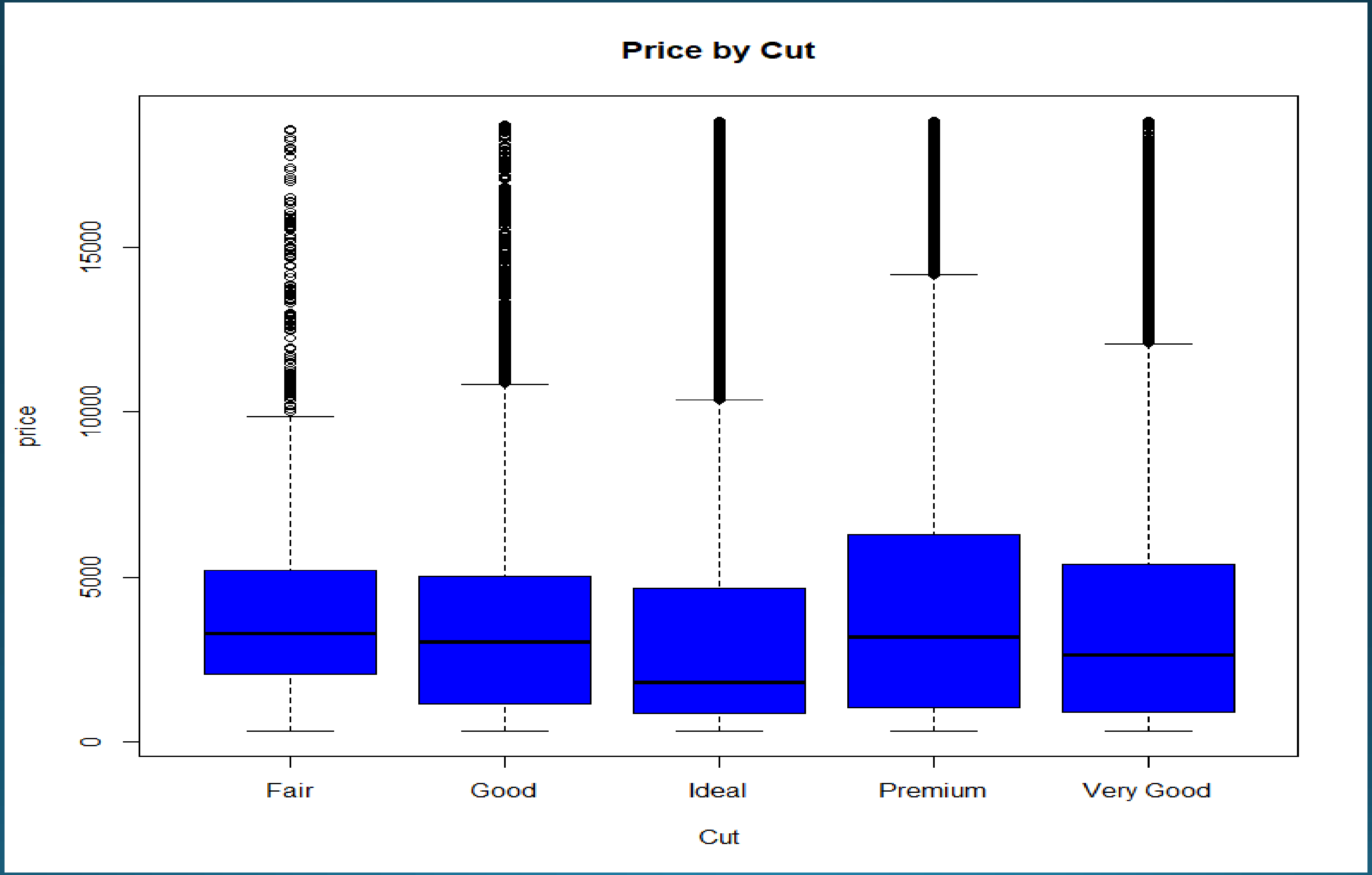






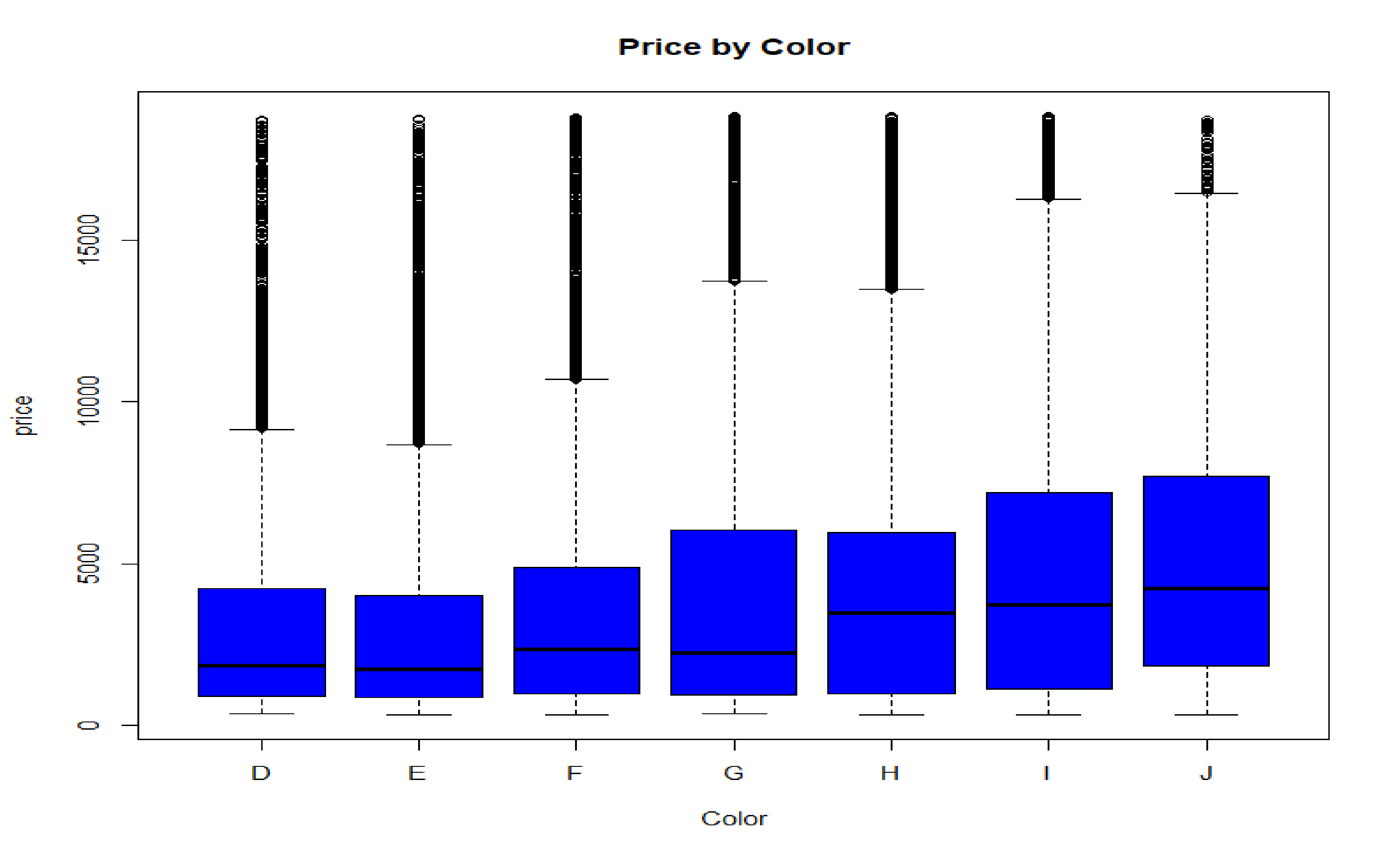
Cut:

- From the plot, we can see that the lower the quality of cut, the higher the number of outliers except for the Ideal cut type. Also, each category type has the same maximum and minimum price.



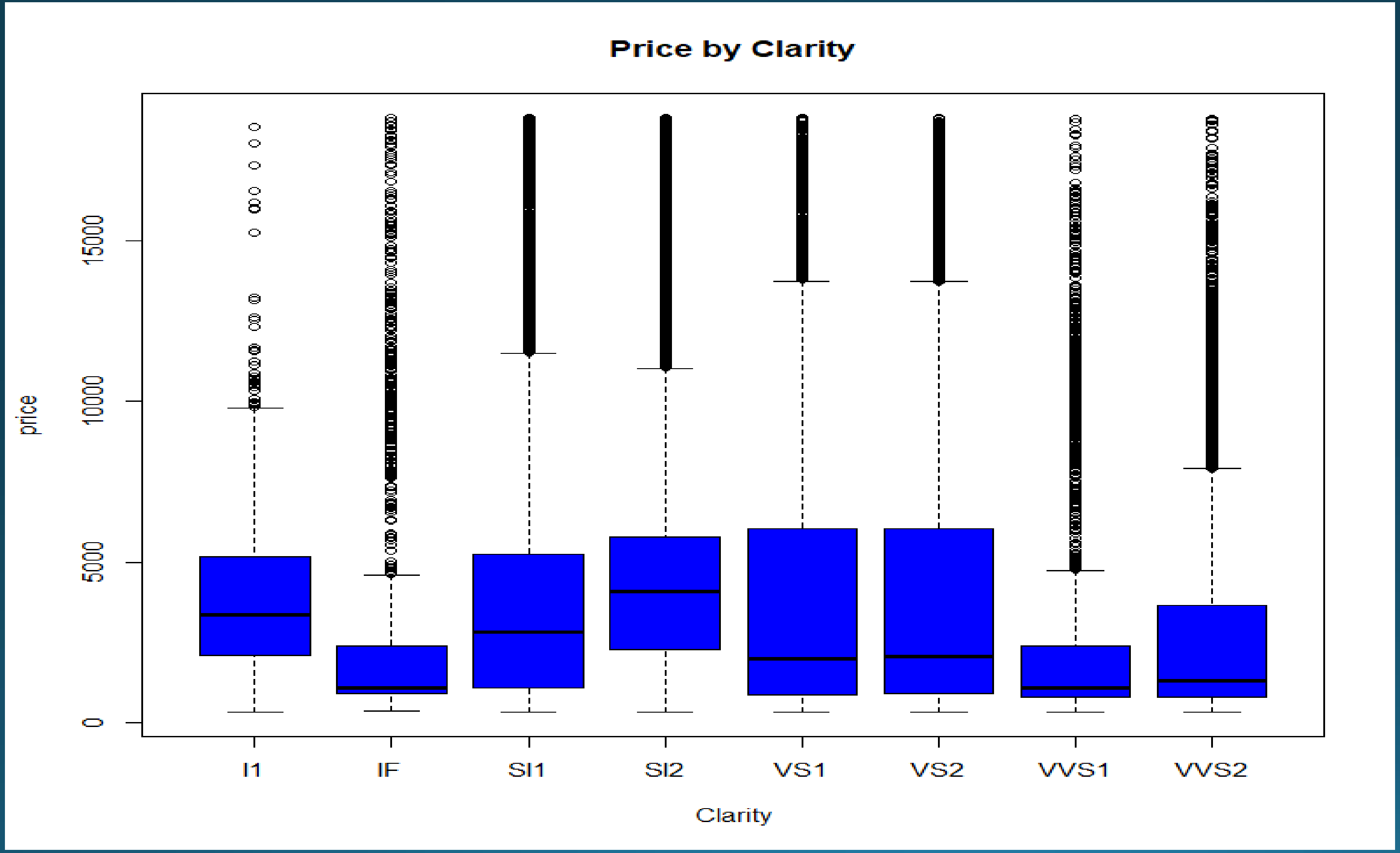
Color:

- From the plot, we can see that G, H, I and J type color has less number of outliers compared to D and E.It suggests that the better the quality of color the higher the outliers except for G type color. Also, each category type has the same maximum and minimum price.



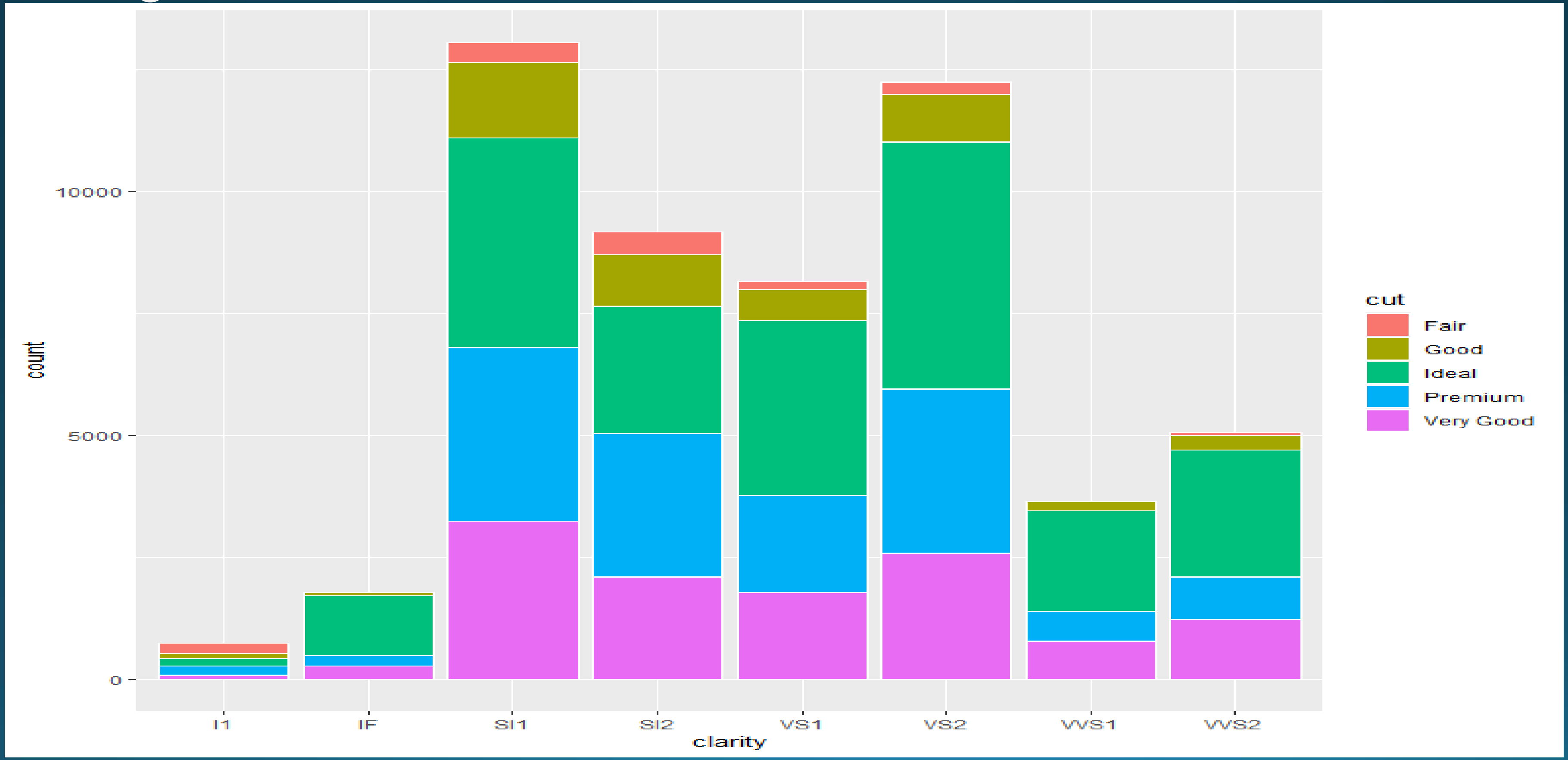
Clarity:

- From the plot, we can see that IF, VVS1 and VVS2 have a high number of outliers compared to other categories of color. Moreover VS1,VS2 are having less number of outliers compared to others .Also, each category type has the same maximum and minimum price.



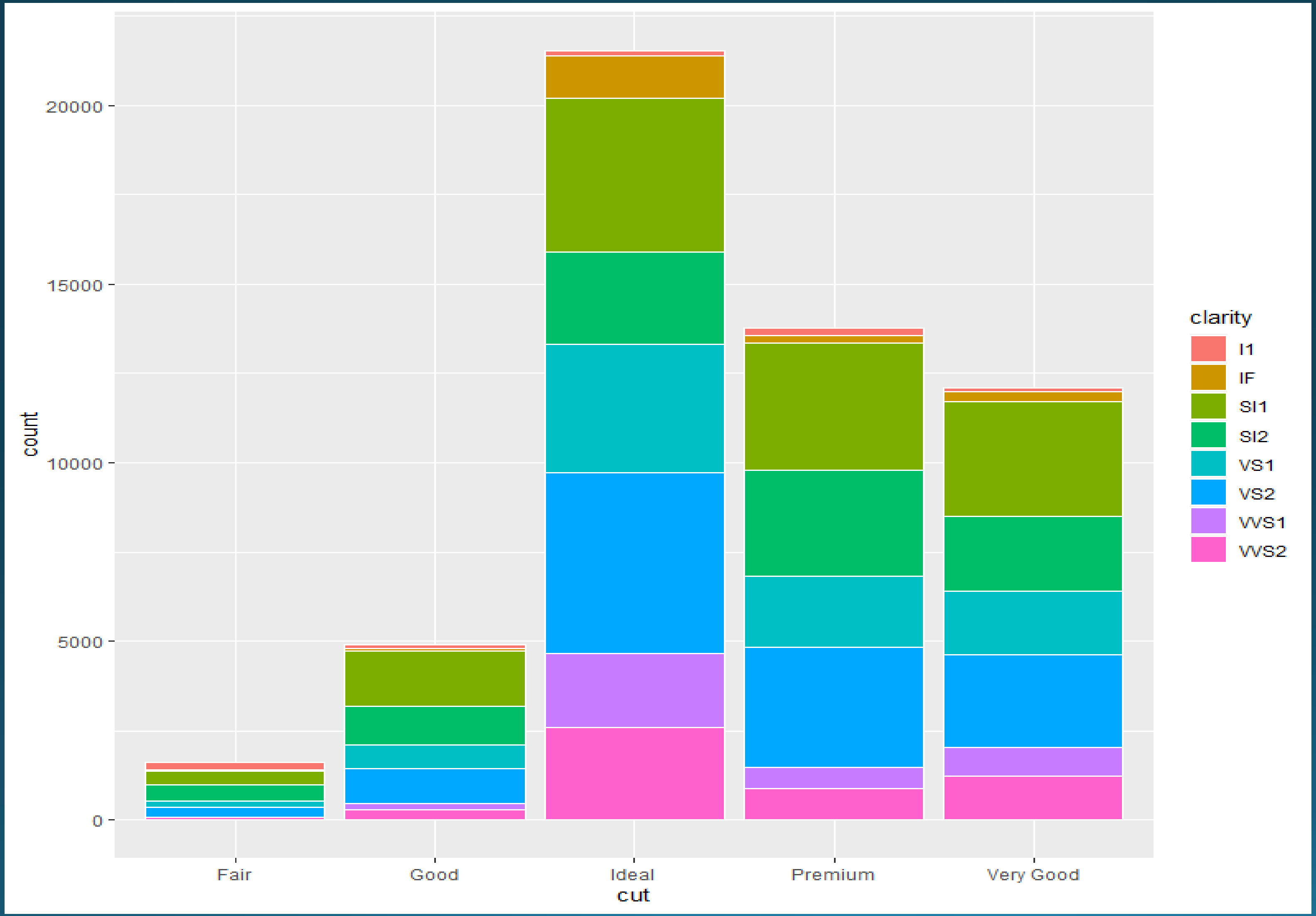
Clarity vs cut:

- We can see that from the plot that most of the people prefer to buy diamond of SI1 clarity followed by VS2, SI2, and VS1.In that, the cut they prefer is Ideal, Premium, and very good's diamond cut category. Moreover, we can infer that people are not taking the highest clarity diamonds, such as IF or VVS1 and others . and are ready to sacrifice on clarity but are more focusing on the cut of the diamonds.



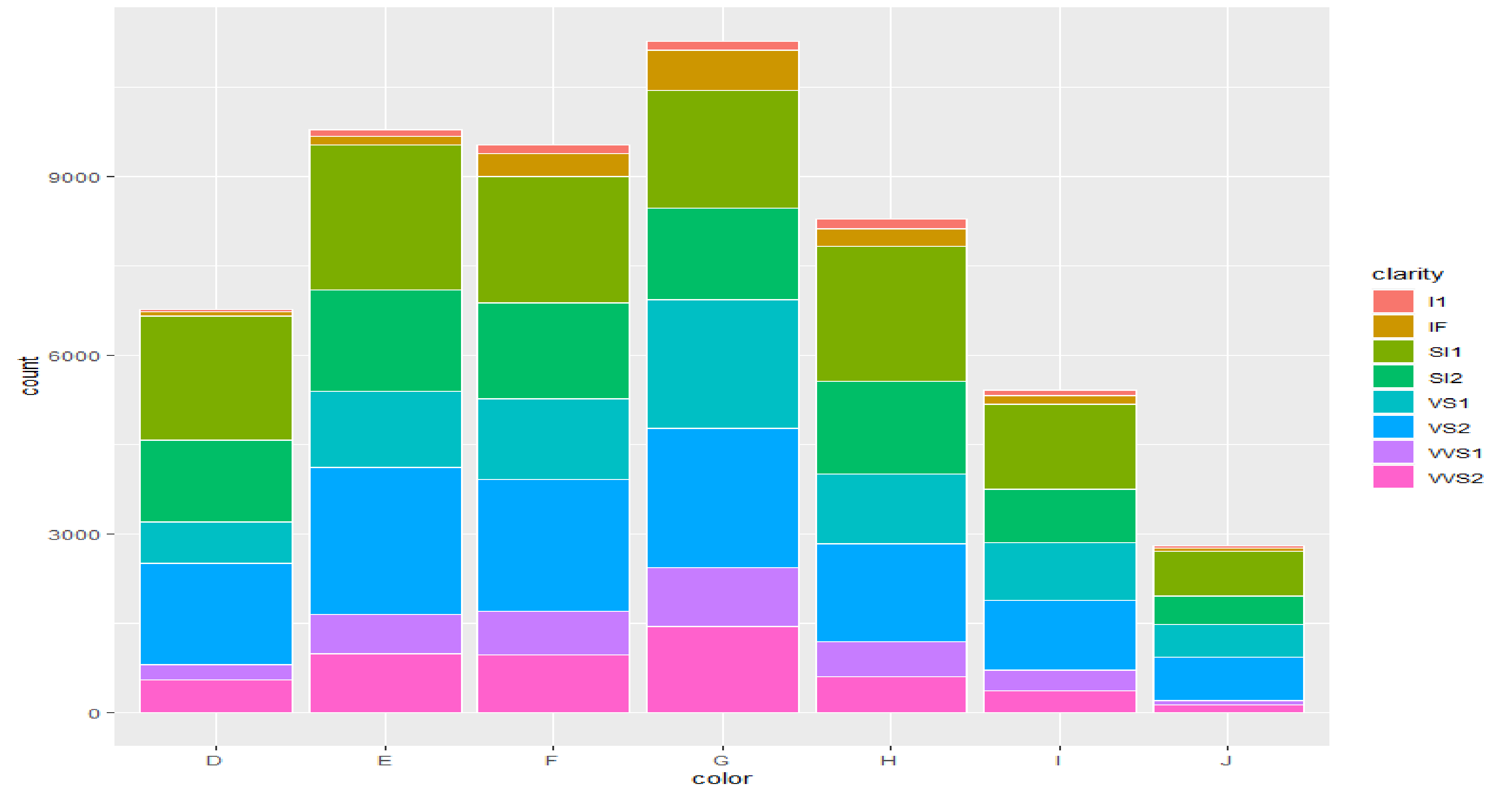
Cut Vs clarity:

- We can see that people prefer Ideal cut over any other cut diamonds followed by Premium and Very Good. It suggests that people are focusing on cut than clarity.



Color Vs clarity:

- We can see that from the plot that most of the people prefer G color followed by E, F, and H.In that the clarity they mostly prefer SI1 or SI2 category.
- Therefore from all the plots, we can conclude that carat has high importance followed by cut, color, and clarity in predicting the price of a diamond.



Conclusion:

- From the above analysis, we could say that carat,length, width, depth are an essential factor in deciding the price of a diamond. However, other features also play an essential role such as cut, clarity, and color. However, some of the features have a considerable number of outliers. Therefore, We have to use regression-based algorithms to determine the price of a diamond based on some of the potential features such as Linear, and regression algorithms to create our model.

THANK YOU!