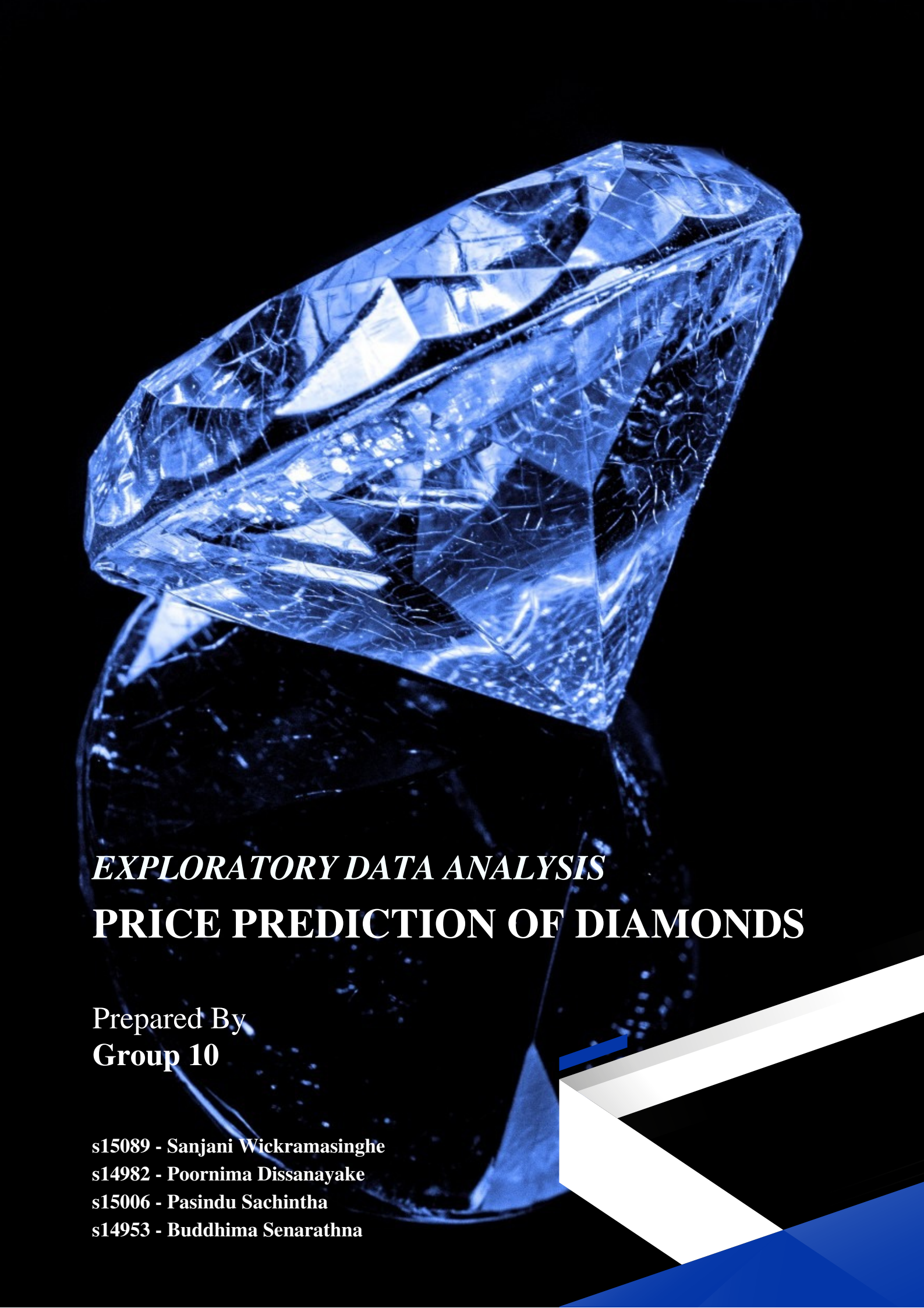

*EXPLORATORY DATA ANALYSIS*

# PRICE PREDICTION OF DIAMONDS

Prepared By

**Group 10**


**s15089 - Sanjani Wickramasinghe**
**s14982 - Poornima Dissanayake**
**s15006 - Pasindu Sachintha**
**s14953 - Buddhima Senarathna**

# *ABSTRACT*

The diamond trade has been under fire in recent times due to controversies surrounding pricing clarity, environmental effects, and working conditions. The lack of clarity in the field has caused widespread uncertainty about the actual value of diamonds, leading to instances where consumers overpay for a diamond. This is a result of the intricate network of suppliers, wholesalers, and retailers involved in the diamond business, making it hard for buyers to accurately compare prices and determine the worth of a diamond. Hence, the aim of our study is to build a model which predicts the price of a diamond based on its characteristics with higher accuracy.

By conducting a detailed descriptive analysis, this study aims to address the need for a comprehensive evaluation of the factors affecting diamond price. The study was conducted with the use of statistical data representation techniques and numerical summaries. Comparisons between predictor variables and response variable has been carried out to check significance and the relationship between variables. Afterwards a Kruskal-Wallis test has been conducted to check if there are statistically significant differences between two or more groups in each of the independent variables. Finally a Principal Component Analysis was implemented to identify outliers and to reduce the dimensionality of dataset and furthermore to check correlation between factors. A Cluster Analysis has been conducted to recognize clusters in the dataset. Based on the findings from the above analysis, some suggestions for a better advanced analysis has also been mentioned.

The findings of this overview will be used to build an accurate predictive model which will be continued in the advanced analysis, by identifying the important elements that influence the price of a diamond.

# *TABLE OF CONTENT*

## *LIST OF FIGURES*



## *LIST OF TABLES*



# *INTRODUCTION*

Diamonds are one of the world's most valuable stones. It is also one of the most expensive stones, hence its price is quite erratic. A diamond is a type of carbon that has undergone intense pressure and heat for an extended period. It is the toughest naturally occurring material and is valued for its impressive qualities such as strength, radiance, and sparkle. Diamonds have been associated with wealth, power, and affection for centuries and are still a sought-after option for jewelry and decorative items.

The worth of a diamond is evaluated based on its cut, color, clarity, and weight measured in carats, and these elements can greatly affect its value.
The price of a diamond is determined by several factors, known as the "**Four Cs**":

- **Carat** weight: The carat weight of a diamond is a measure of its size, with one carat equaling 0.2 grams. Larger diamonds are generally more valuable, although other factors such as cut, clarity, and color also have a significant impact on the price.
- **Cut**: The cut of a diamond refers to the angles and proportions of the stone, which determine its brilliance and fire. A well-cut diamond will reflect light better and appear more sparkling, making it more valuable.
- **Clarity**: The clarity of a diamond is a measure of its internal and external flaws, known as inclusions and blemishes. Flawless diamonds are extremely rare and therefore more valuable.
- **Color**: The color of a diamond is another factor that affects its value, with the most valuable diamonds being colorless or close to colorless. Diamonds with yellow or brown tints are less valuable.

In addition to these factors, the diamond's origin and the demand for certain diamond shapes, such as round or princess cut, can also have an impact on the price.
However, the purpose of this study is to figure out the how the ‘Four C’s’ and other factors impact the retail price of a diamond and create an accurate predictive model that predicts the price of a diamond.

## PROBLEM STATEMENT

The initial step in creating a reliable predictive model is to comprehend the relationship between the natural features and appearance of diamonds and their price behavior. With this understanding, a diamond retailing company can anticipate future demand and establish the necessary standards to meet that demand given a set of inputs. Thus, the objective of this analysis is to develop the optimal predictive model that predicts the price for a diamond given its important features.

## DESCRIPTION OF THE DATASET

The Diamond dataset acquired from the Kaggle website, which is also an inbuilt dataset in the ggplot2 library in R, contains 53940 records under 10 variables including the Price as the response variable (cnt). Table 1 lists down the variables along with a brief description about the variables.

| Variable Name | Variable Type | Description |
|---|---|---|
| **carat** | Quantitative | Weight of the diamond (ranges from 0.2 to 5.01 with a median of 0.7) |
| **cut** | Qualitative | Quality of the cut (Includes 5 types of ordinal values comprising Fair, Good, Very Good, Premium, Ideal) |
| **color** | Qualitative | Color of the diamond (ranges from D(best) to J(Worst) including 7 types of ordinal values) |
| **clarity** | Qualitative | A measurement of how clear the diamond is (includes 8 types of ordinal values as I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, and IF (best)) |
| **depth** | Quantitative | Total depth percentage of the diamond (ranges from 43 to 79 with a median of 61.80) |
| **table** | Quantitative | Width of the top of the diamond relative to widest point (ranges from 43 to 95 with a median of 57) |
| **price** | Quantitative | Price in US dollars (ranges from $326 to $18,800 USD with a median of $2,400 USD) |
| **x** | Quantitative | Length of the diamond in millimeters (ranges from 0 to 10.74 in mm with a median of 5.70mm) |
| **y** | Quantitative | Width of the diamond in millimeters (ranges from 0 to 58.9 in mm with a median of 5.71mm) |
| **z** | Quantitative | Depth of the diamond in millimeters (ranges from 0 to 31.8 in millimeters with a median of 3.53mm) |

*Table 1 : Variable Description*

## DATA PRE-PROCESSING

Initially the diamond dataset contained 53940 records. When examining the dataset, some unusual observations were noticed. It's a fact that measurements such as length(x), width(y) and depth(z) cannot take 0 values. Hence, the records containing 0 in any of these variables were removed. Next, we calculated a new total depth percentage for each record using the x, y and z values. Then, this calculated variable was compared with the original depth variable (with a small tolerance) to identify other remaining unusual observations. Furthermore, a new variable called vol was created using the formula x*y*z.

# *IMPORTANT RESULTS OF DESCRIPTIVE ANALYSIS*

Diamond pricing involves a complex mechanism influenced by multiple factors such as carat, cut, color, clarity etc. This section summarizes some important results and relationships identified through an in-depth descriptive analysis.

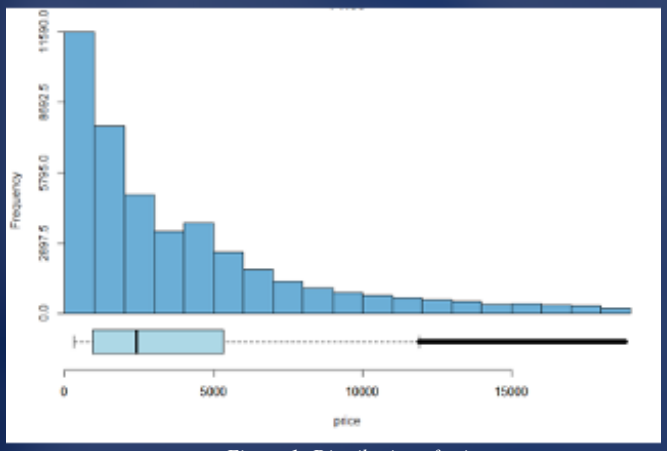


*Figure 1: Distribution of price*

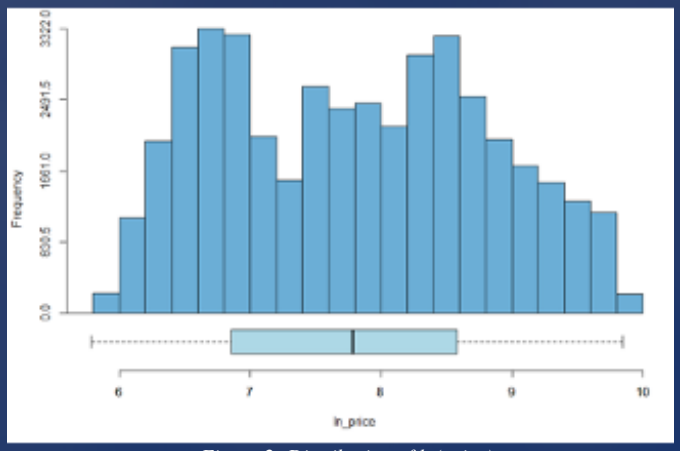


*Figure 2: Distribution of ln(price)*

The distribution of price is positively skewed, meaning that most of the values will be concentrated in the lower range, with a few extremely high values. This is because diamonds with larger carats and dimensions will be more expensive and less common. Also, when we plot the boxplot of this distribution, there seems to have many outliers. After applying the log transformation, the outliers are not present, and the skewness of the distribution is reduced.

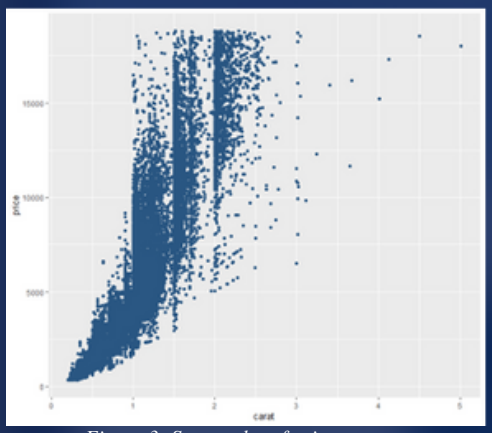


*Figure 3: Scatterplot of price vs carat*

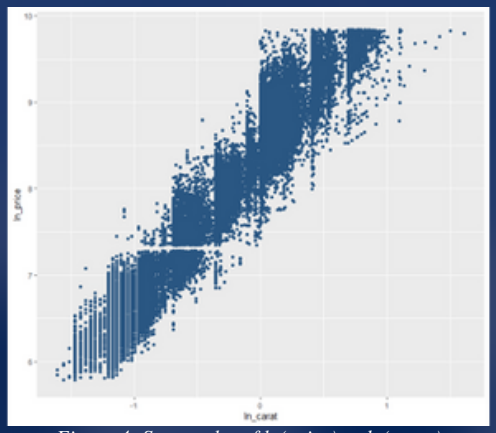


*Figure 4: Scatterplot of ln(price) vs ln(carat)*

As we can see in figure 3, the price of a diamond generally increases with carat weight.
It is because larger diamonds are more rare and therefore more valuable. Carat weight is the unit of measurement used to determine the size of a diamond.

The larger a diamond is, the rarer it is and the more valuable it becomes. This is because larger diamonds are more difficult to mine, cut, and polish, and they also offer a larger surface area for light to reflect and create sparkle.
Additionally, larger diamonds are often seen as more prestigious and desirable, which also drives up their value. Because of these factors, diamonds with a larger carat weight are often more expensive per carat compared to smaller diamonds.
To remove the exponential trend, we will apply a one-to-one transformation where the ln(price) is plotted against ln(carat) and it has a strong positive linear relationship.

It's important to note that other factors, such as cut, clarity, and color, can also impact the price of a diamond, even if the carat weight is the same. A larger diamond with a lower cut, clarity, or color grade may still be less valuable than a smaller diamond with a higher grade in these categories.
The price of a diamond increases with respect to cut, clarity, and color because these are all factors that determine the quality and beauty of a diamond.

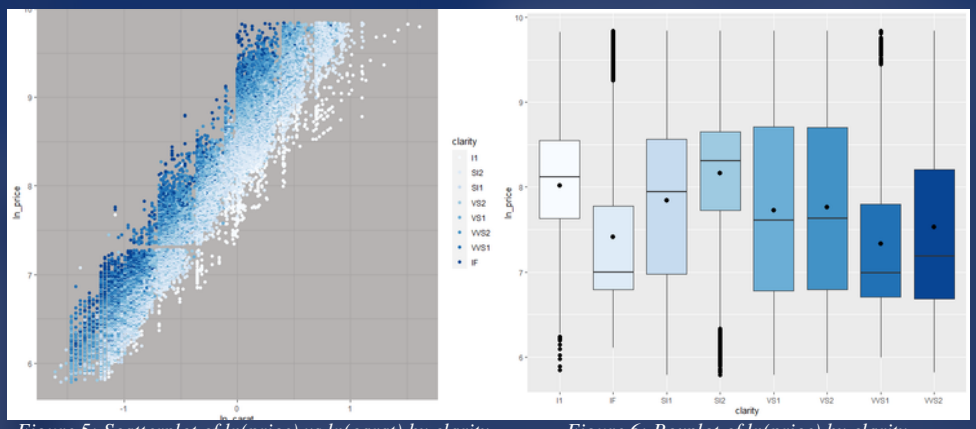


*Figure 5: Scatterplot of ln(price) vs ln(carat) by clarity*




*Figure 6: Boxplot of ln(price) by clarity*

Natural diamonds are the result of carbon exposed to tremendous heat and pressure deep in the earth. This process can result in a variety of internal characteristics called inclusions and external characteristics called blemishes. Clarity refers to the presence of these inclusions and blemishes within a diamond.

Inclusions are internal imperfections, while blemishes are external imperfections. The fewer inclusions and blemishes a diamond has, the higher its clarity and the more valuable it is. This is why a diamond with high clarity such as IF is often more expensive than a diamond with low clarity such as I1 which has the same carat weight. This can be clearly seen by figure 5, where the darkness of the blue color increases which indicates higher clarity diamonds are more expensive.

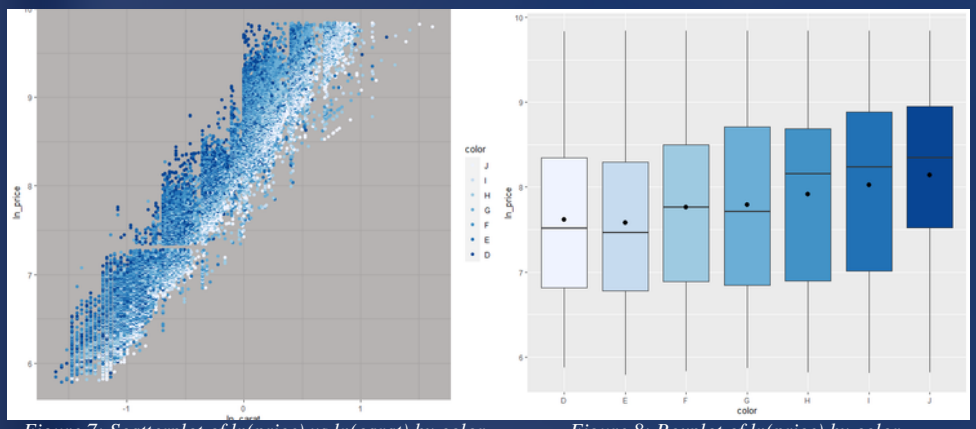


*Figure 7: Scatterplot of ln(price) vs ln(carat) by color*




*Figure 8: Boxplot of ln(price) by color*

Color refers to the presence of any yellow or brown tint in a diamond. The purer and whiter a diamond is, the more valuable it is. This is why diamonds that are nearly colorless or have a very faint yellow tint are often the most valuable and expensive. Here, J represents the worst color which is more yellowish, and D represents the best color which is nearly colorless. Figure 8 clearly shows how the distribution and the means increases with respect to color.

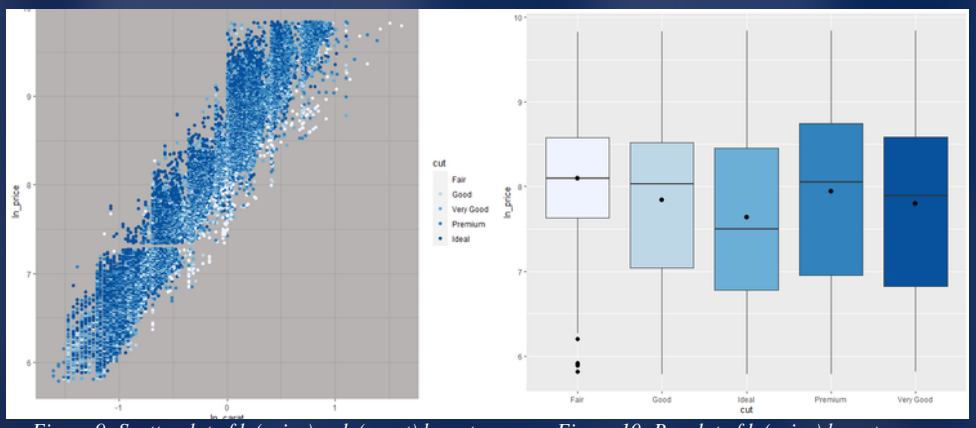


*Figure 9: Scatterplot of ln(price) vs ln(carat) by cut*




*Figure 10: Boxplot of ln(price) by cut*

Diamonds are renowned for their ability to transmit light and sparkle so intensely. We often think of a diamond's cut as shape, but what diamond cut actually does mean is how well a diamond's facets interact with light which implies precision and symmetry. A well-cut diamond will reflect light more effectively and will appear brighter and more sparkling than a poorly cut diamond. This is why a well-cut diamond is often more valuable and commands a higher price than a poorly cut diamond.

Overall, cut, clarity, and color are important factors in determining the value and beauty of a diamond, and these factors will generally have a significant impact on the price of a diamond.
Furthermore, Kruskal Wallis test was conducted to determine if there are statistically significant differences between two or more groups in each of the independent variables clarity, cut and color on the ln_price variable and it provided significant results for all the variables.

| Variable | Test statistic | P-value |
|---|---|---|
| Clarity | 2115.5 | < 2.2e-16 |
| Cut | 755.24 | < 2.2e-16 |
| Color | 1053.2 | < 2.2e-16 |

Table 2: Test Summary

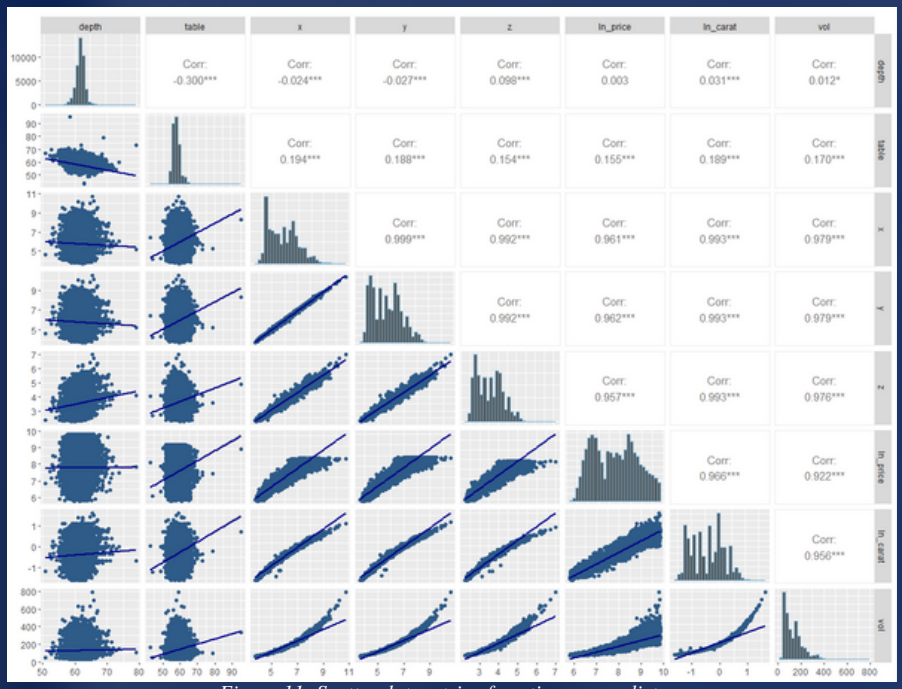


*Figure 11: Scatterplot matrix of continuous predictors*

By referring to the scatterplot matrix, we can clearly see that ln(carat), x, y, z and vol (x*y*z) have a strong positive correlation with ln(price) which implies that bigger diamonds tend to be more expensive. But depth and table have a very weak correlation with ln(price). Furthermore, it indicates that there are strong correlations between predictor variables which indicates multicollinearity.

# *FURTHER ANALYSIS*

## PRINCIPAL COMPONENT ANALYSIS

By PCA we can reduce the dimensionality of our data and use less no of variables(called principal components) in the analysis. Also, this method is useful in dealing with multicollinearity as we can see the variables are highly corelated with each other. Since our categorical variables are ordinal, we can treat them as numeric values.

### Optimal no.of components?

It can be seen that 4 components are enough in order to explain 88.9% of the data.
Other than proportion of variance we can also use the scree plot. The ideal pattern is a steep curve, followed by a bend, and then a straight line. Use the components in the steep curve before the first point that starts the line trend. Therefore using 6 components will be ideal by analyzing the scree plot.
Another method is using the Kaiser criterion, you use only the principal components with eigenvalues that are greater than 1. In this case the first three PCA's would be enough.
Considering all the cases 4 PCA's
will be sufficient.

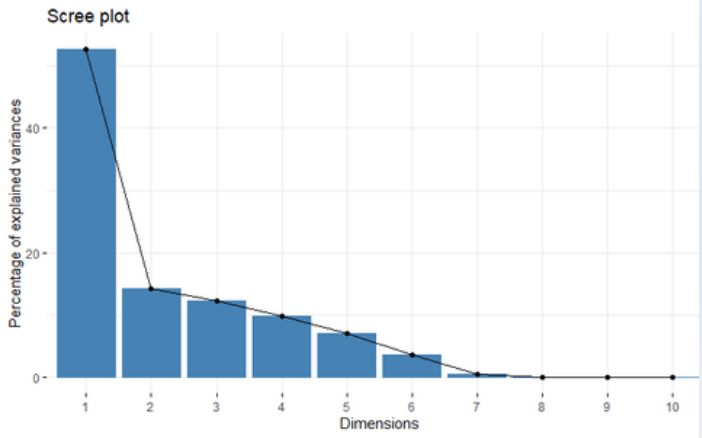


*Figure 12: Scree Plot*

| PC | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Eigen value | 5.26484 | 1.41742 | 1.23017 | 0.98704 | 0.35381 | 0.6995 | 0.0452 | 0.0010 | 0.0005 |

*Table 3:Eigen values of the principal components*

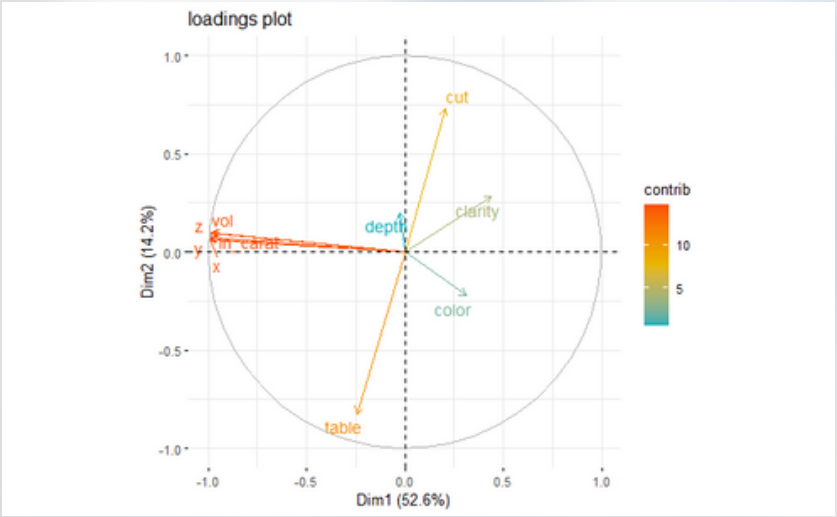


*Figure 13: Biplot of Loadings*

The loading plot shows the relationship between the PCs and the original variables. PC1 has a larger negative association with the variables **x, y, z, volume, ln_carat.** And also, the plot implies that those variables have a high correlation among them.
Also the variables cut and table has significant association with the PC2.
Furthermore the variables cut and table are almost orthogonal to each variable x,y,z and ln_carat.

# CLUSTER ANALYSIS

There are different kinds of clustering methods to create a feature, here Kmeans is used. First of all, scaling should be applied since if a columns have much higher values than the others, it may dominate the results. And then, to determine the best number of clusters here we used Nbclust package.

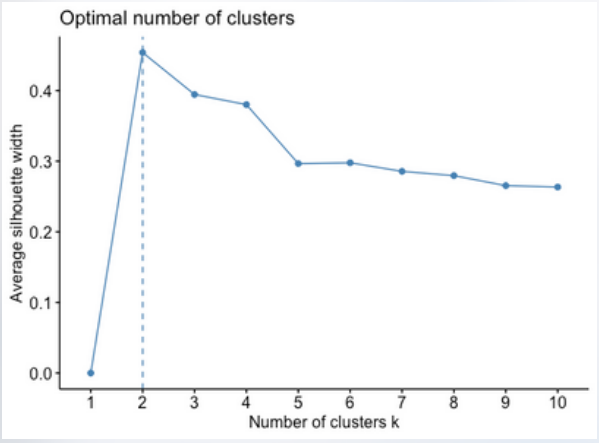


*Figure 14: Line graph of optimal number of clusters*

As can be seen, the **data has been grouped into two separate clusters**. According to the Kmeans clustering result, there are 2 groups or types of diamonds in the train dataset.

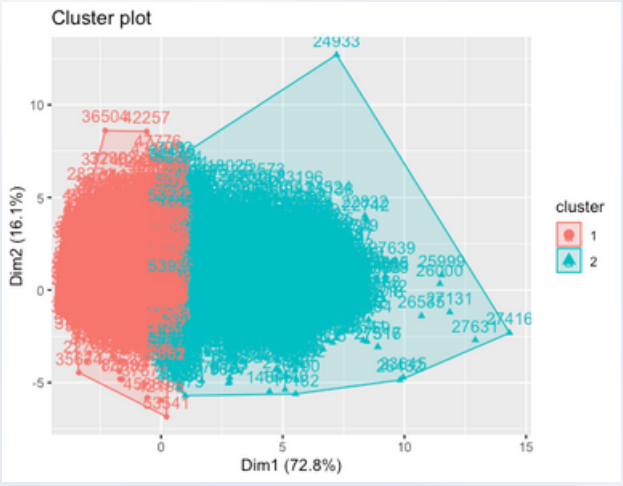


*Figure 15: Kmeans cluster plot*

**What are the characteristics of these clusters ?**

As table 4 shown below, Cluster 1 has a higher number of all columns except cut and clarity. The reverse happens for cluster 2.

It seems we can label cluster 1 diamonds as big and expensive while cluster 2 diamonds can be labeled as smaller and less expensive.

| cluster | carat | cut | color | clarity | depth | table | price | x | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.3280435 | 3.739130 | 4.347826 | 3.478261 | 62.11522 | 57.58478 | 7987.13 | 6.953913 | 6.963478 | 4.322826 |
| 2 | 0.4485246 | 4.295082 | 3.131148 | 4.590164 | 61.71311 | 57.04918 | 1269.23 | 4.861311 | 4.870656 | 3.002131 |

*Table 4: Variable means by clusters*

Also when we plot the our response variable - price vs above two clusters we can get the below graph, it also indicates cluster 1 diamonds are expensive while cluster 2 diamonds are less expensive.

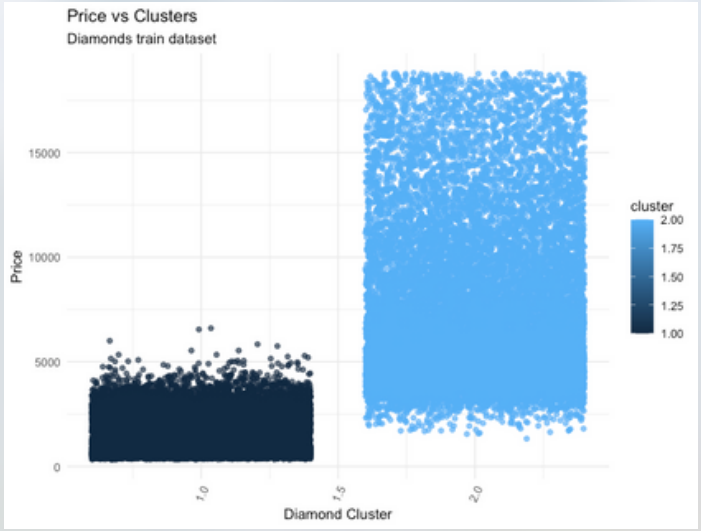


*Figure 16: Visualizing clusters against price*

# *SUGGESTIONS FOR QUALITY ADVANCED ANALYSIS*

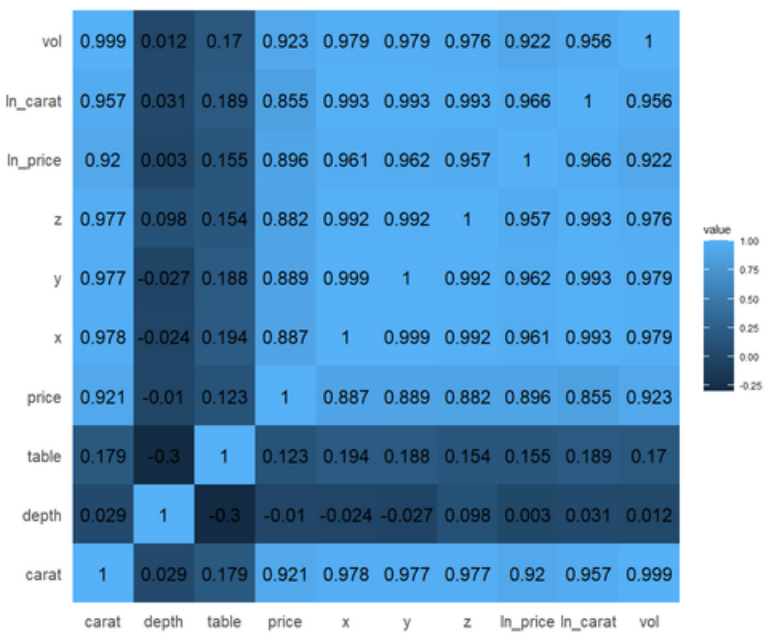


*Figure 17: Correlation matrix*

- Figure 11 indicates that most of the continuous predictors are highly correlated with ln(price). Also, there are some predictors which are highly correlated with each other which indicates multicollinearity. Hence, we can try fitting a multiple linear regression by dropping some of these highly correlated predictors. Moreover, we can also try fitting a model using the principal components we obtained.

- Existence of multicollinearity can be corrected through the utilization of several different regularization and variable reduction techniques such as Ridge Regression, LASSO regression, and Elastic Nets.

- Random Forest can be used which is a strong ensemble learning method that may be used to solve a wide range of prediction problems, including classification and regression. Because the method is based on an ensemble of decision trees, it offers all the benefits of decision trees, such as high accuracy, ease of use, and the absence of the need to scale data.

- Furthermore, it is possible that there are clusters in our dataset. In this case, we can try fitting separate models for each cluster and check the accuracy of the models.

# *APPENDIX*

R code: https://github.com/st3082group10/price_prediction_on_diamonds