**Introduction**

Diamonds are a very rare and valuable jewel that is held in high regard by many throughout the years. Many see it as a symbol of love with the use of diamonds as engagement rings. Yet, there are many different uses for diamonds that people have found to gift to themselves or their loved ones. However, the diamond price range varies quite significantly from a couple of hundred dollars to tens of thousands. Diamonds are priced based on factors like carat which is the weight of a diamond, color, length and width of the diamond, cut, clarity, and total depth percentage. The dataset utilized cuts some poor-quality diamonds; for instance, color can go from D to M, while clarity can include I2 and I3 as well. This dataset only includes D through J and clarity only goes to I1.

**Data Visualization**

*Figure 1*

Chart, pie chart

Description automatically generated

According to figure 1, the clarity distribution is more leaned toward the middle of how clear the diamond is. While the extreme opaque (I1) is very tiny but also the extremely transparent (IF) is also very rare in the distribution. It leans a bit more toward the opaque side, as SI2, SI1, and VS2 are the larger values in the distribution and are on the opaque side. People are willing to sacrifice clarity for other more important, to them, variables. It is even possible for smaller diamonds with low clarity to still look like a perfect diamond. If people are getting an engagement ring, then they might even realize that the diamond has low clarity and purchase it. With all this knowledge, clarity does not seem like a huge factor in determining whether a diamond is chosen or not.

*Figure 2*

Chart, bar chart

Description automatically generated

In the color distribution, as seen in figure 2, D is the best diamond color and J is the worst diamond color. Although G is the most frequent, it is more skewed to the better color being more frequent. This is different than the clarity of the diamond. When rationally thinking about this, it makes sense that people care more about the color of the diamond than the clarity because people would notice the color first of the diamond and then the clarity as an afterthought. Most colors are generally appealing, so people can choose a lower color but still be satisfied with a lower price.

*Figure 3*

Chart, histogram

Description automatically generated

According to figure 3, people are most likely to purchase a diamond in the $1000 range. This is understandable because the everyday person who needs a diamond for engagement rings and other small jewelry pieces is not willing and not able to spend thousands of hard-earned dollars for something that they could get for $1000 and still be satisfied with their purchase. Of course, there are rich elites who can spend money on anything they want, so a diamond that costs more than 15,000 is a reasonable and acceptable price for them. What people want in a diamond is very subjective, if those who have money want to spend it on top-quality products, there is a market for them, however this is not true for most people. There is still a viable market for those who do not want to break the bank and spend exuberant amounts of money on something that may only be used for special occasions.

*Figure 4*

Chart, pie chart

Description automatically generated

The cut of the diamond is clearly an important factor in people’s decision on buying a diamond. This compared to all of the other variables mentioned before and the carat, has a much higher need for positive frequency than negative. The very ‘good’, ‘premium’, and ‘ideal’ categories take up almost the entire graph, while the ‘good’ and ‘fair’, do not even take up 25%. This indicates that people are willing to pay more for a high-quality cut above the other variables. People probably want their diamond to have a good quality cut more than color because that is the first thing that a person will notice about the diamond if everything else was average.

*Figure 5*

Chart, histogram

Description automatically generated

On the other hand, the diamond’s carat does not seem to be a big factor in people’s choice of a diamond. People cannot visually see the weight of the diamond and therefore cannot negatively or positively speculate on it. Beauty is more of a factor than weight. If the diamond is an engagement ring, people do not want it to be too heavy on the wearer’s finger so a low carat would make the most sense. Carat number also does not mean that the cut, clarity, or color is better. A low-carat diamond that all has favorable traits toward cut, clarity, and color is more significant than one with a high carat but has low traits in the other variables. Many people think that the cut means the shape of the diamond, however, the cut reflects on the diamond’s quality of cut.

**Linear Regression Analysis**

While doing linear regression, this report used every variable since none of them have an extreme number of categorical values or had NAs. Every variable besides the depth in mm, z, I1 clarity, color D, y, and cut fair is statistically significant at less than a .001 p-value. In the adjusted R-squared value, 91.98% of the variability in the diamond’s price can be explained by cut, clarity, carat, color, table, x, and y. A price of a diamond starts at $2,198. The color correlates with a negative price, but the worst color has the most devastating impact on the price of the diamond, while the best color has the least negative impact on the price of the diamond. Clarity has a high positive correlation with price. The average of IF clarity adds an extra 5345 dollars. The weight of the diamond, the carat, also has a significant impact on the price. The length by itself has a correlation of -1000 relating to price. However, when adding y as an interaction term, it becomes a positive 752. This is because no one wants a very wide diamond with no length or a long diamond with no width. People generally want an equal-looking diamond in terms of length and width.

**Clustering Analysis**

The dataset was quite large with 50,000 sample units so to do clustering, it was limited to 60% of the population and removed the categorical variables such as color, clarity, and cut. This was done in order to process and get results for clustering. Therefore, kmeans was chosen as the clustering method as it only deals with numerical variables and not categorical which would need to have an even larger dataset to be created. It was found that the optimal number of k-means was 4 which was chosen by the elbow point, as seen in figure 6.

*Figure 6*

A picture containing text, diagram, screenshot, line

Description automatically generated

*Figure 7*

Chart, scatter chart

Description automatically generated

Comparing price and carat, figure 7, before it was scaled, there were very clear distinctions between price and carat, which means that price had a larger range and thus more impact on the model, than carat range. Typically, the lower the carat, the lower the price.

*Figure 8*

*Chart, scatter chart

Description automatically generated*

As you can see in figure 8, after the price was scaled relating to carat, there is much more overlap in the clusters. Even though there is still a very strong correlation between price and carat. There is still a range between the weight of the diamond and the price since there are other factors that relate to the price and that are still statistically significant which could explain the carat being lower and the price still being at the extreme value of price. There are very few diamonds out there that have the extreme weight of five and all higher carat weight has a higher price correlation.

**Classification**

Using k-nearest neighbor and optimal number of clustering, 4, the model predicted that 595 diamonds were expensive when they were not and 529 were predicted as not expensive when they really were. It is obvious that buying a supposedly expensive diamond and then finding out later that it was not is more detrimental to the customer. A customer would be rightfully angry and disappointed that they paid a premium for something that actually was not premium. On the flip side, if someone purchased a diamond that they thought was not expensive but it turned out to be, they would be very satisfied with their purchase. Some errors that occurred were in the middle lower zones of the carat, depth, table, x, y, and z sections. The model found it difficult to estimate these diamonds since they were in the middle lower zones which could go either way.