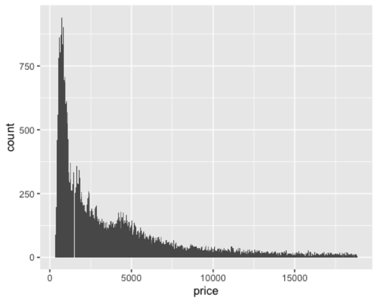
Study on the dataset ‘Diamonds’ through data visualization

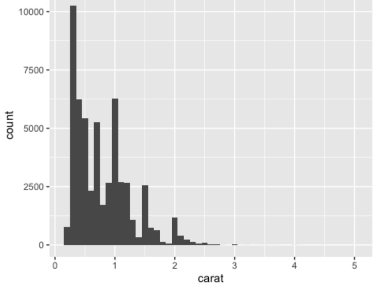
1. Data Description

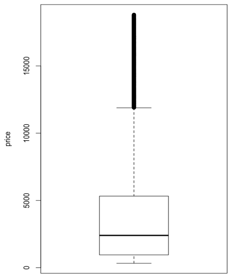
In the dataset **Diamonds**, there are 53,940 observations and 10 variables, including seven continuous variables and 3 categorical variables. The seven continuous variables are *price* (price in US dollars), *carat* (weight of the diamond), *x* (length in mm), *y* (width in mm), *z* (depth in mm), *depth* (total depth percentage), *table* (width of top of diamond relative to widest point). The three categorical variables are *cut* (quality of the cut), *color* (diamond color), and *clarity* (a measurement of how clear the diamond is).

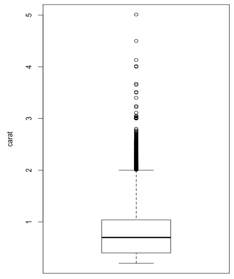
Currently, the variables, *cut* and *clarity*, are arranged by the alphabetic order in the dataset. We rearrange them based on their quality. For example, we rearrange the data order for *cut* from “fair”, “good”, “very good”, “premium”, to “ideal.” In this way, we are able to conduct data analysis logically.

The summary of each variable can be found in the Appendix A. Here, we plot the histograms and boxplots of two continuous variables, *price* and *carat*.









The above figures show that *price* and *carat* are both highly right-skewed, with many outliers falling beyond the third quartile. This is not surprising, because in real world situations, diamond prices can vary dramatically based on their quality and scarcity. These valuable diamonds reflect the market prices and should not be excluded from the dataset as outliers. Therefore, it would be more helpful to use median as the measurement to analyze the two variables rather than using their average values, confidence intervals.

For the readers’ reference, Appendix B illustrating central tendency, dispersion, and range of the seven continuous variables are attached. The average price is $3,933 and the price median is $2,401; The mean of carat is 0.80 and the median is 0.70. The standard deviation of price is $3,989 and the range is $18,497; The standard deviation of carat is 0.47 and the range is 4.81. However, due to a large number of meaningful outliers, we would use median rather mean or standard deviation as our measurement to avoid biases in our analysis.

2. Data Analysis

From this dataset, two major areas of interest are developed - the pricing mechanism of diamonds and the purchasing behavior of diamonds. We will first make some assumptions to study the price of a diamond, followed by which more analysis will be done to find some patterns in consumer’s purchasing behavior.

We conducted market research on diamonds to better understand this dataset and equipped ourselves with the basic knowledge of diamonds. From the research, we learned that the most crucial factors which determine the price of a diamond are cut, carat, color, and clarity. (Appendix C: reference website) Therefore, our analysis on diamonds will be based on these factors.

Analysis Part I: The Pricing Mechanism

To begin with, three assumptions are worth noticing:

I: The price of a diamond will increase as the carat (weight) increases.

II: Given the same carat, different levels of cut, color, and clarity have different diamond prices.

III: The price per carat of a diamond has different degrees of sensitivity towards cut, color, and clarity.

Based on these assumptions, we create the following figures using ggplot2 package from R. The y-axis is price and x-axis is carat. The factors - cut, color, and clarity - are plotted independently in three separate figures. Note that the slopes of lines represent the price per carat.

Figure 1

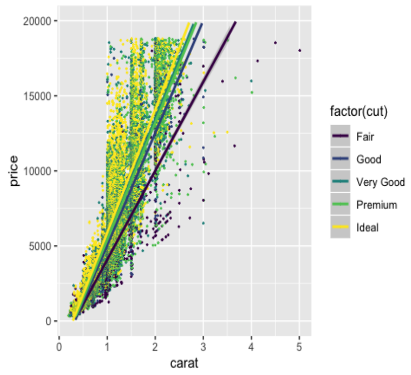


Figure 2

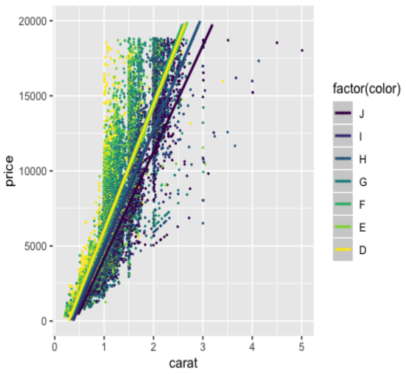
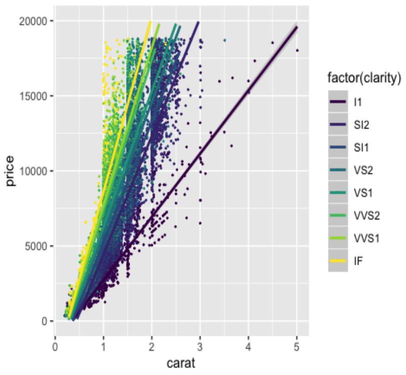


Figure 3

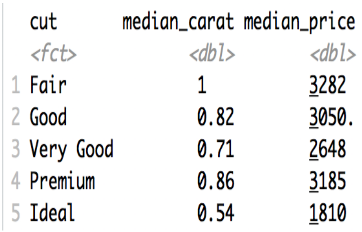


Looking at the three plots, the positive slopes of all the lines show that Assumption I holds, as the positive sloping means that the heavier the carat (weight), the more expensive the price of a diamond.

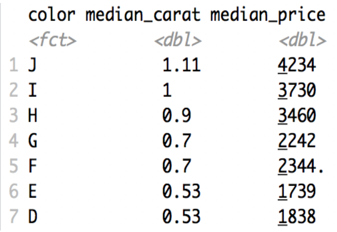
Looking at the three plots independently, we see that Assumption II also holds, as given the same carat, the points on the top are diamonds with more premium quality, better color, and higher clarity, while points in the bottom are diamonds with low-quality parameters.

From Figure 2 and 3, we can see that the points of different colors are well-scattered based on different factor levels, and in particular, the slopes of lines are more differentiated in Figure 3 compared with that in Figure 2. That is to say, as the clarity improves from I1(poorest) to IF (best), the price per carat increases more significantly compared with the price per carat changed when colors shift from J (poorest) to D (best). When looking at Figure 1, the differentiation happens only between cut level “Fair” and other categories. More specifically, as the cut level is not “Fair” and changes from the lower to the higher level, the price per carat does not change a lot. Therefore, Assumption III holds as well.

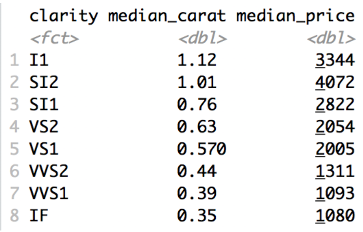
After proving the three assumptions, we now provide some descriptive data analysis on the continuous variable in the dataset. As previously mentioned, we use median as the measurement for analysis due to a number of meaningful outliers that cannot be removed in the dataset. The results are shown in the following figures.



```
  cut       median_carat median_price
  <fct>            <dbl>        <dbl>
1 Fair              1            3282
2 Good              0.82         3050.
3 Very Good         0.71         2648
4 Premium           0.86         3185
5 Ideal             0.54         1810
```



```
  color median_carat median_price
  <fct>        <dbl>        <dbl>
1 J             1.11         4234
2 I             1            3730
3 H             0.9          3460
4 G             0.7          2242
5 F             0.7          2344.
6 E             0.53         1739
7 D             0.53         1838
```



```
  clarity median_carat median_price
  <fct>          <dbl>        <dbl>
1 I1             1.12          3344
2 SI2            1.01          4072
3 SI1            0.76          2822
4 VS2            0.63          2054
5 VS1            0.570         2005
6 VVS2           0.44          1311
7 VVS1           0.39          1093
8 IF             0.35          1080
```

It may be surprising to notice that the median price of diamonds at the “Fair” level is almost as twice as that at the “Ideal” level. The question is how a poor-quality diamond is more expensive than a premium-quality diamond. After looking into the data more closely, we find that the median carat of diamonds at the “Fair” level is almost as twice as that at the “Ideal” level. So, carat (weight) seem to be an important factor that determine the price of a diamond. Other factors, such as diamond cut, color, and clarity, also seem to adjust the diamond price.

Analysis Part II: The purchasing behavior of diamonds

In this part, the boxplots show how price and carat are plotted against the factors – cut, color, and clarity.

Looking at Figure 5 and Figure 6, we can see that as the quality of color and clarity grows, people tend to choose to decrease the carat to maintain a reasonable price which they can afford. In other words, there is a tradeoff between the weight and the quality (color and clarity) of a diamond. However, we can notice that in the category with higher quality parameters, say color D and clarity IF, although the medians of carat and price go down, there is still a number of outliers compared with lower-quality categories, say color J and I1. This shows that a small portion of rich people still purchase a very large diamond with high-quality parameters.

Looking at Figure 4 which shows the cut level, the most interesting thing is that neither the weight nor the price of a diamond goes down as the cut level increases. For instance, when cut level changes from ‘Very Good’ to a higher ‘Premium’, the mean carat increases rather than decrease. This provides another piece of evidence that there is not a significant change in the price per carat when the cut level changes compared with the level in color and clarity change.

Figure 4 (a) – Price vs. Cut

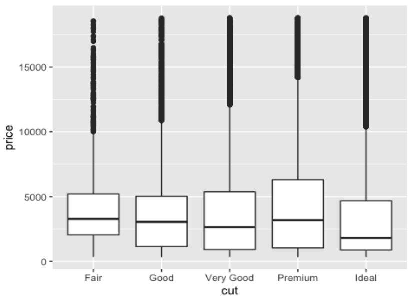


Figure 4 (b) - Carat vs. Cut

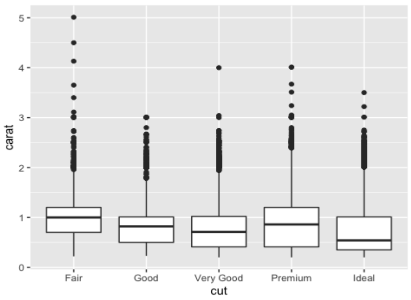


Figure 5 (a) – Price vs. Color

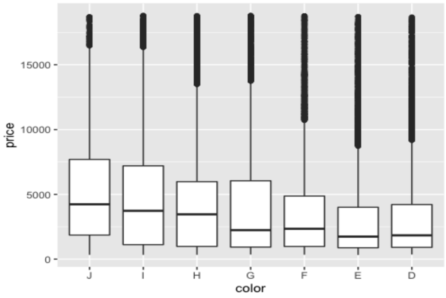


Figure 5 (b) – Carat vs. Color

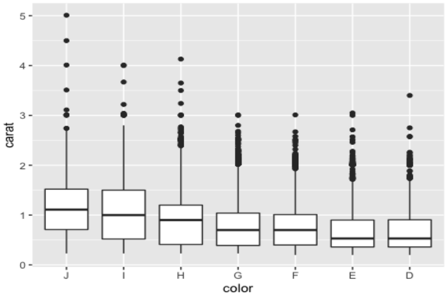


Figure 6 (a) – Price vs. Clarity

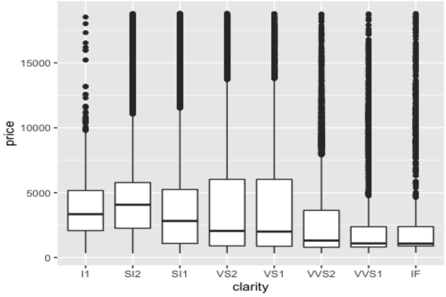
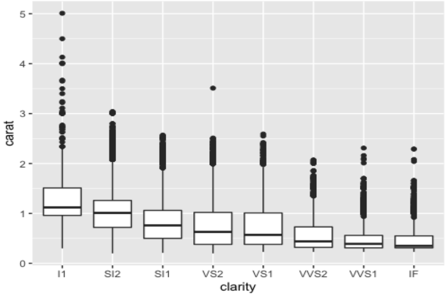


Figure 6 (b) – Carat vs. Clarity



3. Conclusion

I: The price of a diamond will increase as the carat (weight) increases. The weight of a diamond (carat) seems to be a very important factor when determining the price of a diamond.

II: Given the same carat, different levels of cut, color, and clarity have different diamond prices.

III: The price per carat of a diamond has different degrees of sensitivity towards cut, color, and clarity.

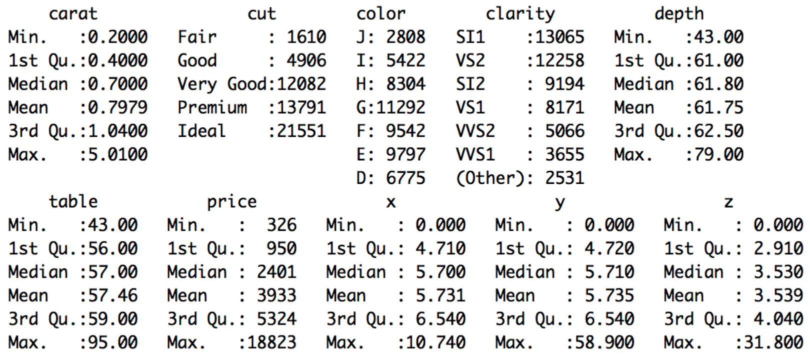
IV: People tend to make a tradeoff between a diamond's weight and its quality (cut, color, and clarity), except a small portion of rich people may choose for the diamond's weight and quality at the same time.

4. Limitation and Future Study

For this report, we mainly used descriptive analysis such as graphing the dataset to generate the conclusions. To get statistically significant results, more advanced statistical analysis should be conducted. For instance, i) Kruskal-Wallis Test can be used to test whether there is a difference in price between different levels of colors, clarity and cut, ii) a regression model can be established to test whether factors like carat (weight), depth and table make a significant impact on prices. Furthermore, as we have found that the changes in diamond cut, color, and clarity impact the price per carat differently, further research can be done to investigate whether they have different levels of impact on the overall price of a diamond. Last but not least, a case study on several most common types of diamonds in the market can be conducted to study the more detailed pricing mechanism.

Appendix

Appendix A – Summary of variables in dataset **Diamonds**



```
     carat              cut        color      clarity        depth      
 Min.   :0.2000   Fair     : 1610   J: 2808   SI1    :13065   Min.   :43.00  
 1st Qu.:0.4000   Good     : 4906   I: 5422   VS2    :12258   1st Qu.:61.00  
 Median :0.7000   Very Good:12082   H: 8304   SI2    : 9194   Median :61.80  
 Mean   :0.7979   Premium  :13791   G:11292   VS1    : 8171   Mean   :61.75  
 3rd Qu.:1.0400   Ideal    :21551   F: 9542   VVS2   : 5066   3rd Qu.:62.50  
 Max.   :5.0100                     E: 9797   VVS1   : 3655   Max.   :79.00  
                                    D: 6775   (Other): 2531                  
     table           price            x                 y                z         
 Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910  
 Median :57.00   Median : 2401   Median : 5.700   Median : 5.710   Median : 3.530  
 Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735   Mean   : 3.539  
 3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040  
 Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900   Max.   :31.800  
```

Appendix B – Central Tendency, Dispersion, and Range of Question 3 **Diamonds**

	price	**carat**	**depth**	**table**	**x**	**y**	**z**
Mean	3933	0.80	61.75	57.46	5.73	5.74	3.54
Median	2401	0.70	61.80	57.00	5.70	5.71	3.53
Min. Value	326	0.20	43.00	43.00	0.00	0.00	0.00
Max. Value	18823	5.01	79.00	95.00	10.74	58.90	31.80
Range	18497	4.81	36.00	52.00	10.74	58.90	31.80
Standard Deviation	3989	0.47	1.43	2.23	1.12	1.14	0.71

Appendix C – Reference website of dataset **Diamonds** :

https://www.withclarity.com/education/

Appendix D –Please find the R code for this report in the file named “Diamonds.R” under the “eric-jiang-1997/stats-and-machine-learning/diamonds_dataset_visualization” folder.