Exploratory Data Analysis - Diamonds

Shabnam Hajian

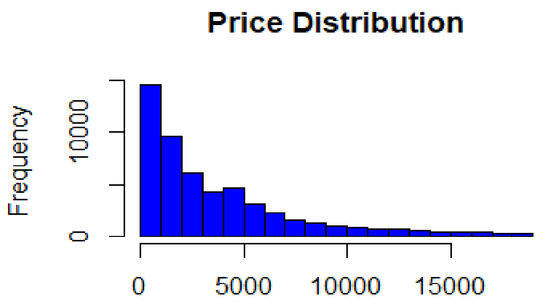
2020-04-09

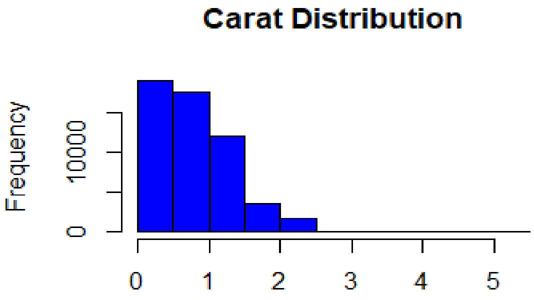
Data

This dataset involves **53,940** observations of different diamonds grouped in **10** different variables: **1)price**, the value of each diamond in US dollars varying from $326 to $18,823, **2)carat**, the weight of each diamond in carat (each carat is .2 gram) varying from 0.2 to 5.01 carat, **3)Clarity**, measuring how clear a diamond is (fewer inclusions, more clearness) classifying in 8 different categories from worst=I1 to best=IF , **4)color**, measuring how free from any color a diamond is (colorless to near colorless) classifying from D=best to J=worst, **5)cut**, the manner a diamond has been shaped and polished (from its beginning as a rough stone to its final gem) classifying in 5 different categories from fair to ideal, **6,7,8)x**=length, **y**=width, and **z**=depth that used to show diamonds' dimensions, **9,10) depth** and **table** which talked about the shape of diamonds.

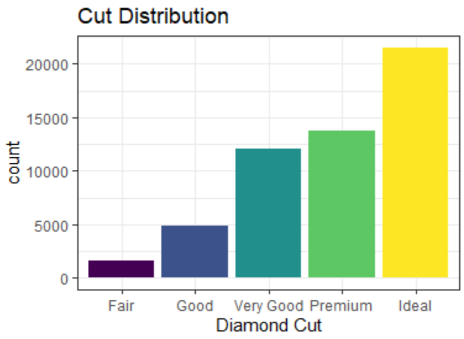
We will use exploratory data analysis to understand how diamonds **price** varies according to carat, cut, color, and clarity, informally known as the "four Cs", that commonly used as the basic descriptors of diamonds. This factor was developed by **Gemological Institute of America** in 1953 as internationally recognized standard parameters to evaluate diamonds characteristics. The other 5 variables are approximately subsets of these 4Cs; since **depth** and **table** are two of the most important factors in determining the quality of a diamond's *cut*, and **x**, **y**, and **z** are highly correlated with *carat*. This correlation is respectively 0.975, 0951, 0.953 which means the bigger the diamond became, the heavier it would be. Also, *Carat* shows a higher correlation compared to dimension variables with price, as the main outcome variable (the price correlation with carat is 0.92 while its correlations with x, y and z are 0.88, 0.86 and 0.86). Furthermore, *carat* is commonly used to determine a diamond's price in the real market. As a result, we choose **cut** over *depth* and *table*, and choose **carat** over *X*, *Y* and *Z* as the size variable. We will continue with analyzing the relations of *price* with the 4 Cs, especially the carat as the most important "C" in diamond's value determination.

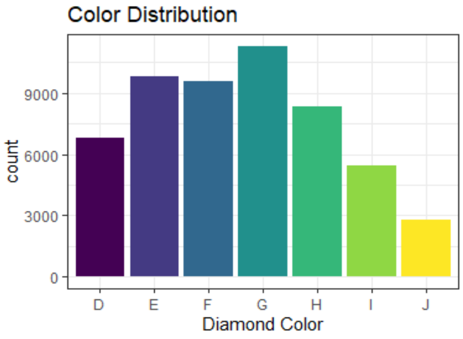
price distribution is skewed to right with the mean and standard deviation of $3,933 and $3,989.44 respectively. The **carat**'s distribution is also skewed to right. The main reason for the skewness of this variable is that diamonds are not sold per stone but per carat and larger diamonds are rare in the market. One important note in this dataset is that 6,253 of our observations (about %11.6) are out of three standard deviation of carat's mean but since we do not have enough reasoning and we believe these points are due to rare stones , we would not consider them as outlier and we will continue with all available observations. Carat distribution has Mean=0.8 and Standard Deviation=0.45.

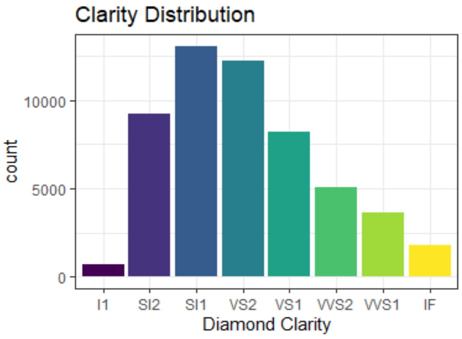




The below graphs also illustrate that most of the observed diamonds (about %88) have an ideal to very good **cut** and (about %85) have a D-H **color** label. While, lower levels of **clarity** (SI2, SI1, VS2) have a higher frequency which is sensible, as only about 20% of all mined diamonds have a high enough rate of clarity to be considered as an appropriate gemstone, and the rest 80% have industrial use.

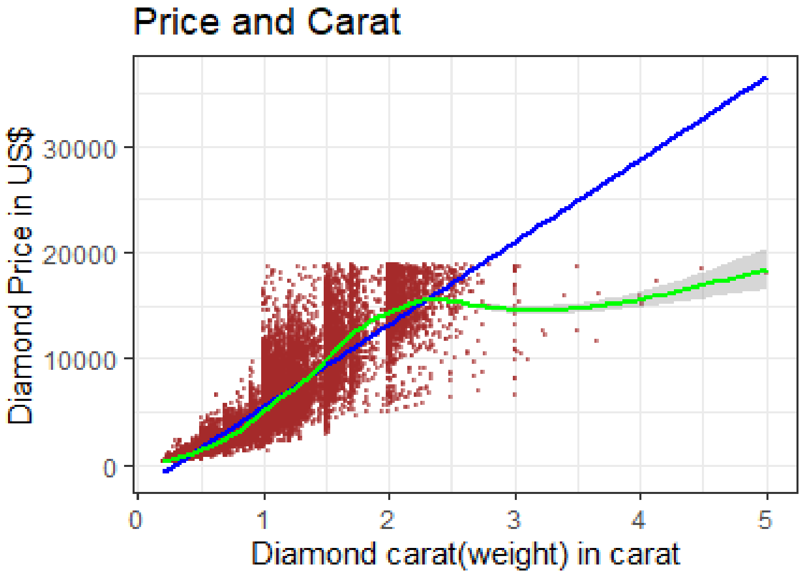






Analysis

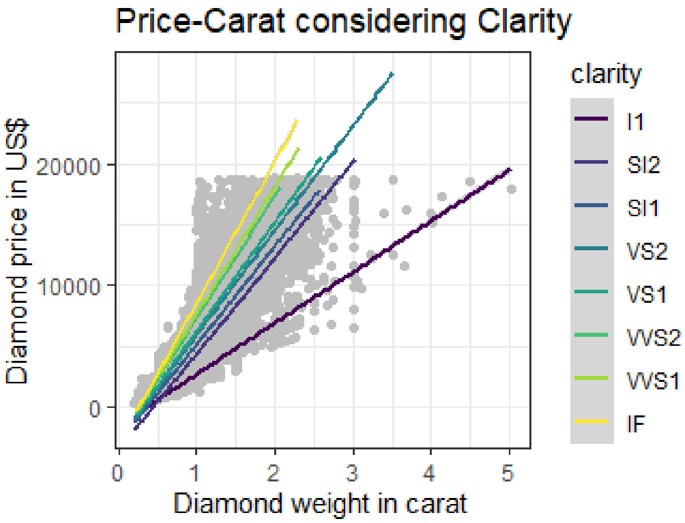
A) Price Relationship with carat:



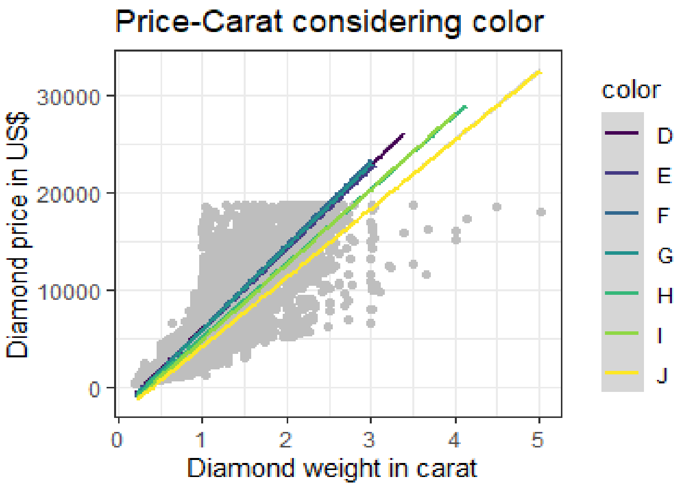
The graph shows that **price** and **carat** are positively correlated (cor=0.92), however the price does not behave linearly with carat in heavy diamonds. In the graph, the green line illustrates the real relationship between price and carat, while the blue line assumed a linear association. We can observe some heavy diamonds (carat around 5) that are just as expensive as lightweight diamonds (carat around 1), while we expect them to be $40,000 based on linear relationship assumption. Also, we can see that for each specific weight (carat), there are a wide range of prices. As an instance, the price of diamonds with 1.5 carat varies from $3,500 to $18,800. The graph shows that there is no magic number that can be multiplied to a diamonds' weight and find its price since other variables such as cut, clarity, and color might have an influential effect on diamond's price.

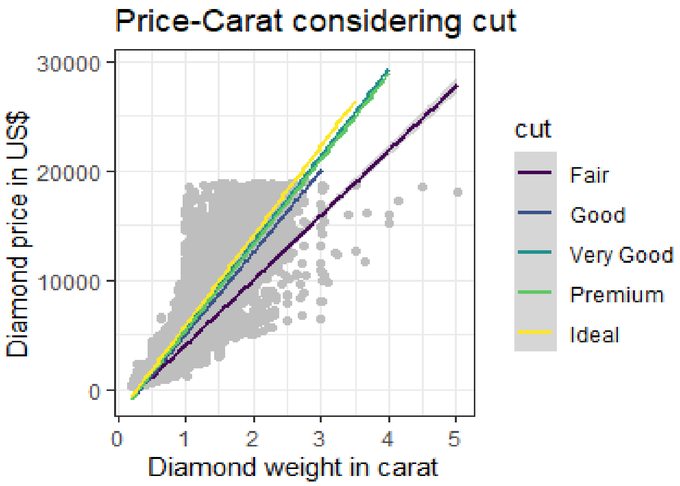
B) Price relationship with carat considering the other three "C"s:

The below graph shows, as the **clarity** improves and moves from I1=Worst to IF=Best (from dark lines to brighter lines), the diamond price becomes more sensitive to the weight. This means that in better clarity levels the relationship between carat and price is steeper and the slop of relation lines increases by clarity level improvement. As a result, a change in the weight of a clear diamond has a stronger effect on its price compared with a similar change of not clear diamonds. For instance, if a diamond weight for the best clarity diamond(IF) increase from 1 carat to 2 carat, the price of that diamond increase from $10,000 to $20,000; however, this weight change for the worst clarity diamond(I1) can increase the price by only $3,000 (nearly one- third). The other important note from this chart is about the carat ranges. We could see that diamonds with better clarity have a smaller range of changes in carat(weight) plus they are lighter, while worse clarity 's carat (like I1) varies from 0.2 to 5.01 and are available in bigger carats.



We can also see in following graphs that the improvement of **color**, means moving from J=Worst to D=Best (from brighter lines to darker lines), or improvement of **cut,** means moving from fair to ideal (from darker lines to brighter lines), shows similar effects on the relationship between carat(weight) and price as the clarity does. This means that the line that shows the relationship between price and carat(weight) for less favorite diamonds are flatter in all these three categories of clarity, color and cut. Moreover, better colors and better cut diamonds are more available in lighter weights. We can also see that there is not a substantial difference between D-F lines in the color graph. This could be explained by the fact that diamonds graded D–F are considered "colorless" and G–J are considered "near-colorless" as a more general group. Thus, the little differences among colorless diamonds do not affect the diamonds' value in real market. We can see that this overlapping of lines happened among Ideal to very good levels in cut graph too; since people cannot distinguish among these differences by unaided eyes.





Conclusion

Most gem diamonds are traded on the wholesale market based on a single value(price). However, this value is affected by diamonds characteristic knowing as carat, clarity, color, and cut (four Cs). Thus, although the price is positively related to all these four characteristics, knowing only some of these variables is not enough to determine diamond value. For example, a 5-carat diamond might be as expensive as a 2.5-carat diamond just because its other characteristics are not as favorable as the lighter diamond. As a result, consumers who individually purchase diamonds are often advised to use all four Cs (carat, clarity, color, and cut) to pick the diamond that is "right" for them. For example, knowing that a diamond is rated as 1.5-carats (300 mg), VS2 clarity, F color, and the excellent cut is enough to reasonably establish an expected price range.

We also found that people become more sensitive about the proportion of diamond prices to weigh when they find a high-quality diamond. While the value of a poor-quality diamonds cannot increase considerably as its weight increase.

In this report, although we found out that each of these 4 Cs has an important effect on price, we could not compare which characteristic has a higher effect. If we used statistical models in addition to visual diagrams, we could get a better conception from all these relations at the same time. We also did not distinguish between heavier diamonds and lighter diamonds which have different uses and demands in the real market. However, different markets could affect these relations. For future follow-ups, we suggest to group diamonds based on their market, and then develop and compare prediction models for price in each group based on these four important characteristics.