Diamonds Dataset

Akshay Kapoor and Rishi Shah

01/02/2020

Introduction

The dataset is about diamonds and the attributes that affect the price of a diamond. This dataset features *prices* and various attributes such as *carat*, *cut*, *color*, *clarity* and some physical measurement features like *depth*, *table* and *x,y,z* dimensional data for 53,400 diamonds.

We can observe that data has 53,940 samples and have data corresponding to 10 features for each of these observations.

Let's look at each feature of data individually :

1. Carat: It stored as a numerical ratio variable and represents the weight of the diamond (one carat = 200 milligrams). It varies from 0.2 to 5.01 with a mean of 0.7979 for this dataset.
2. Cut: It is an ordered categorical variable which represents the quality of cut for the diamond. It is categorized as "Fair", "Good", "Very Good", "Premium", "Ideal" (in increasing order).
3. Color: It is an ordered categorical variable which represents the color of the diamond. It is categorized as "Fair", "Good", "Very Good", "Premium", "Ideal" (in increasing order).
4. Clarity: It is an ordered categorical variable which exhibits the measurement of how clear the diamond is. It is categorized as "I1" (worst), "SI2", "SI1", "VS2", "VS1", "VVS2", "VVS1", "IF"(best).
5. x: It is a numerical variable, representing length of diamond in mm, it ranges from 0 to 10.74mm in this dataset.
6. y: It is a numerical variable, representing breadth of diamond in mm, it ranges from 0 to 58.9 mm in this dataset.
7. z: It is a numerical variable, representing depth of diamond in mm, it ranges from 0 to 31.8 mm in this dataset.
8. Depth: It is quantitative measure of total depth percentage and hence is a numerical-ratio variable, It ranges from 43 to 79 and is calculated using the x,y and z dimensions.
9. Table: It is a numerical variable, which represents width of top of diamond relative to widest point and it ranges from 43 to 95.
10. Price: It is a numerical variable, which represents the price of diamond in US dollars, it ranges from $326 to $18,823.

Summarising Data Features

Before heading off to in-depth analysis of dataset, let's have a quick glance over the statistics of each variable present in the diamonds dataset. We would be using a summary command for understanding the variation of each feature.

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

Observations from feature summary

Assessing the above information, we can see we have quantitative predictor variables as carat,x,y,z,table,depth and qualitative predictor variables as cut, clarity and color. All these features affect the dependent price variable for an individual sample. Also, we can form a basic notion about the data, that most of diamonds present in the data are of "Ideal" cut (21551 in number), "SI1" clarity (13065 in number) and of color "G" (11292 in number).

Analysis on the data

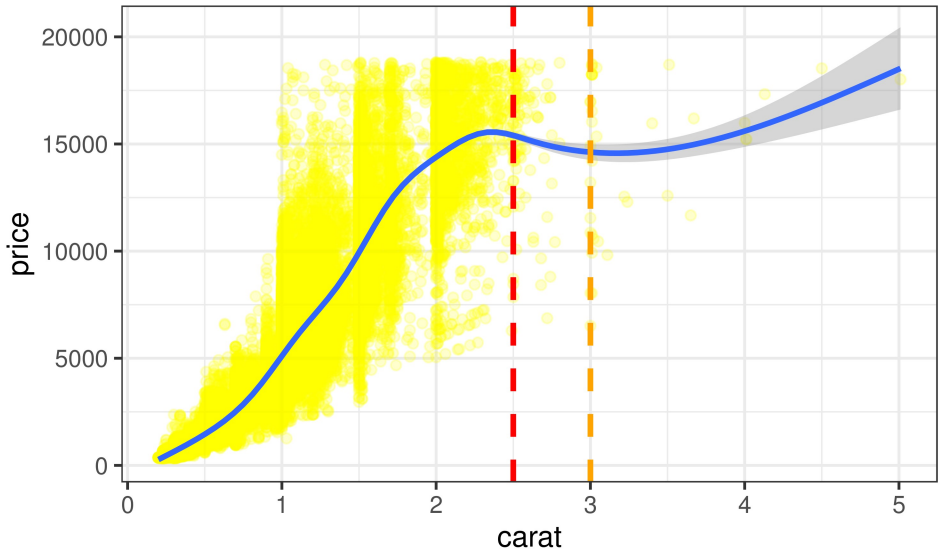
To begin with the basic analysis, let's calculate correlation among features using ggpair() from GGally library.

Correlation between different variables

From the correlation calculations, We have the following observations: 1. There is a strong correlation among Price, carat and dimensional parameters(all correlation values are greater than 0.85). 2. Price exhibits very limited or no relationship with parameter table and depth. 3. Carat is highly correlated to x,y,z, which is as per our expectations i.e. volume and weight would be proportional to each other.

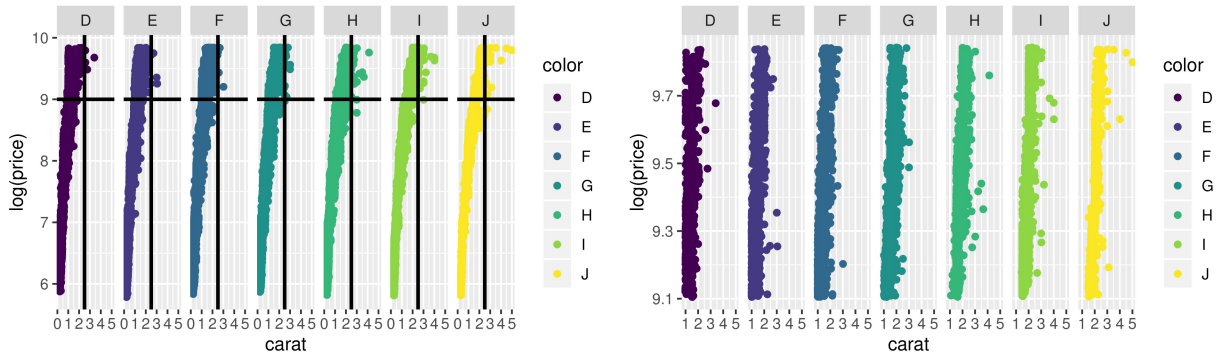
Because of high correlation value, we plotted carat vs. price scatter plot.

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



On plotting carat vs price, we have the following observations : 1. As carat i.e. the weight of diamond increases, the price of diamond increases. 2. Also, we have very few samples for higher carat diamonds, only 0.265% diamonds have a carat value greater than 2.5 and this proportion further reduces to 0.074% for carat value >3.

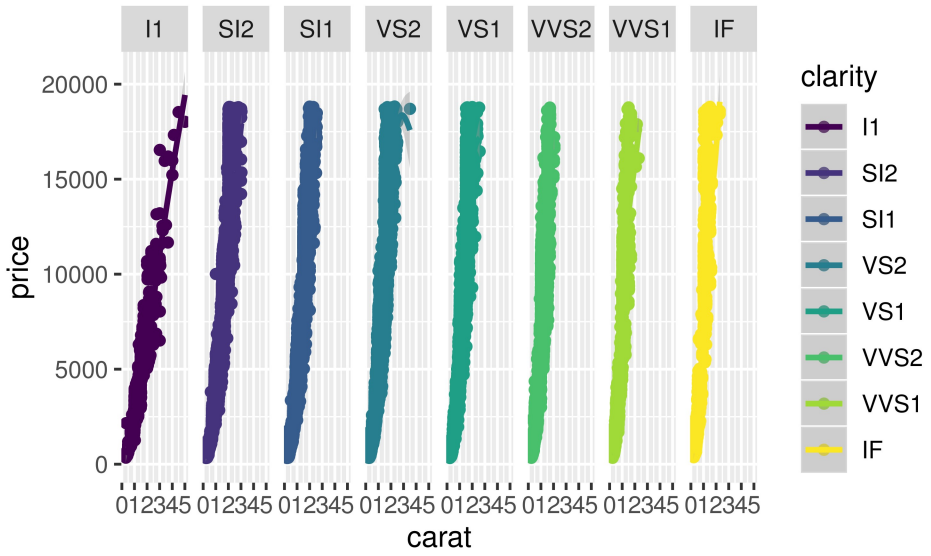
To dive deep into analysis, let's plot carat against the price for various colours.in order to visualise the relationship among price vs. carat for different diamond colors.



By looking at the above plot, we can visualize that as carat increases, the price of diamond increases and this holds true for every color of the diamonds. Furthermore,as the color of diamond increases (ie D being most transparent and J being most colorful) there is a slight curve in the graph, which gives an idea that high carat diamonds have dense distribution for more colorful diamonds compared to transparent diamonds. It looks like for higher price diamonds,the color quality is compensated by improving the carat value. To analyse further we would plot graph higher price diamonds (>9000 USD) vs carat for different diamond colors(Above right).

Inspired by the above information, we are motivated to see the effect of clarity on price vs. carat plots.

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



By looking at the above plot we deduce following observations:

Similar to the color quality, as clarity of diamond reduces (ie I1 worst clear and J Most clear),curve in price vs carat graph is observed in the graph, which points to an idea that less clarity diamonds tend to have high carat value to improve the price value.

These observations are supported by the following statistical calculations displayed below.

```
##      upper_price mean_price lower_price upper_carat mean_carat lower_carat
## ci_D    3249.895   3169.954    3090.013   0.6663585  0.6577948   0.6492312
## ci_E    3142.981   3076.752    3010.524   0.6651658  0.6578667   0.6505676
## ci_F    3800.840   3724.886    3648.933   0.7445169  0.7365385   0.7285600
## ci_G    4073.864   3999.136    3924.408   0.7793331  0.7711902   0.7630474
## ci_H    4577.360   4486.669    4395.979   0.9230116  0.9117991   0.9005867
## ci_I    5217.601   5091.875    4966.148   1.0423469  1.0269273   1.0115077
## ci_J    5488.044   5323.818    5159.592   1.1841832  1.1621368   1.1400903
```

Conclusion

Form the above graph and tables, We have following conclusions:

1. Our sample has few high carat values(>2.5) compared to lower.
2. These high carat values usually correspond to the diamonds having low color quality or poorer clarity.

Thus, we can conclude that the higher carats tend to compensate for the low color quality and poorer finish in high price diamonds.

Team Contribution

This report is been presented by Joint efforts of Akshay Kapoor and Rishi Shah. Akshay has contributed in covariance calculations and plotting the right graphs whereas Rishi has put together the report for knitting and carried out analysis for the various graphs plotted in the report.