EDA on Diamonds Dataset

Dhaval Thakur, Rushi Bhuva, Tejas Pandit

29/01/2020

Data set in study: Diamonds

Introduction

A dataset containing the prices and other attributes of 53,940 diamonds. This is a dataset which is preloaded.

The dataset contains information on prices of diamonds, as well as various attributes of diamonds, some of which are known to influence their price that are **carat, cut, color,** and **clarity** , as well as some physical measurements (depth, table, price, x, y, and z).

Observation : After running the structure command on the dataset we get see that there are the 10 variables present in the dataset namely

carat (of num type) ,cut (ordinal factor.. with 5 levels),color (ordinal factor with 7 Levels) ,clarity (ordinal factor with 8 levels) ,depth (of num type) ,table (of num type) ,price (of int type),x (of num type),y (of num type),z (of num type)

Data Quality

On further data checking on diamonds dataset, we found that the data has 20 observations that has x or y or z dimensions as 0. These observations are noisy data because there should not be any diamond with any one dimension as 0 value.

Summary of the Data

Before going in-depth in the dataset of diamonds its always good to have a quick glance over the stats of each variable present in the diamonds dataset. We use summary command for the same.

Observation From the above summary we got, we can see that the measurable variables present in diamonds dataset such as carat,depth,price etc have their central tendencies, ranges. whereas categorical variables such as clarity, color, we can see their frequencies listed for each level. From the summary we get a basic idea that the ideal cut diamonds are mostly present in the dataset with 21551 to be specific in number, and G color is present the maximum in the dataset with 11292 frequency. The variable carat has mean of 0.7979 and median 0.70 .On the other end for Price variable means is 3933 and median 2401 dollars respectively. Range for **carat** variable is 0.2, 5.01 and range for **price** is 326, 18823.

Analysis on the data

We did the re-ordering of factor levels and clarity so that the order is shown consistently from worst to best. (proof present in R file) Now to start the basic analysis lets have a distribution of the sample diamond data so that we get to obtain a better understanding of price range we have in our sample data.

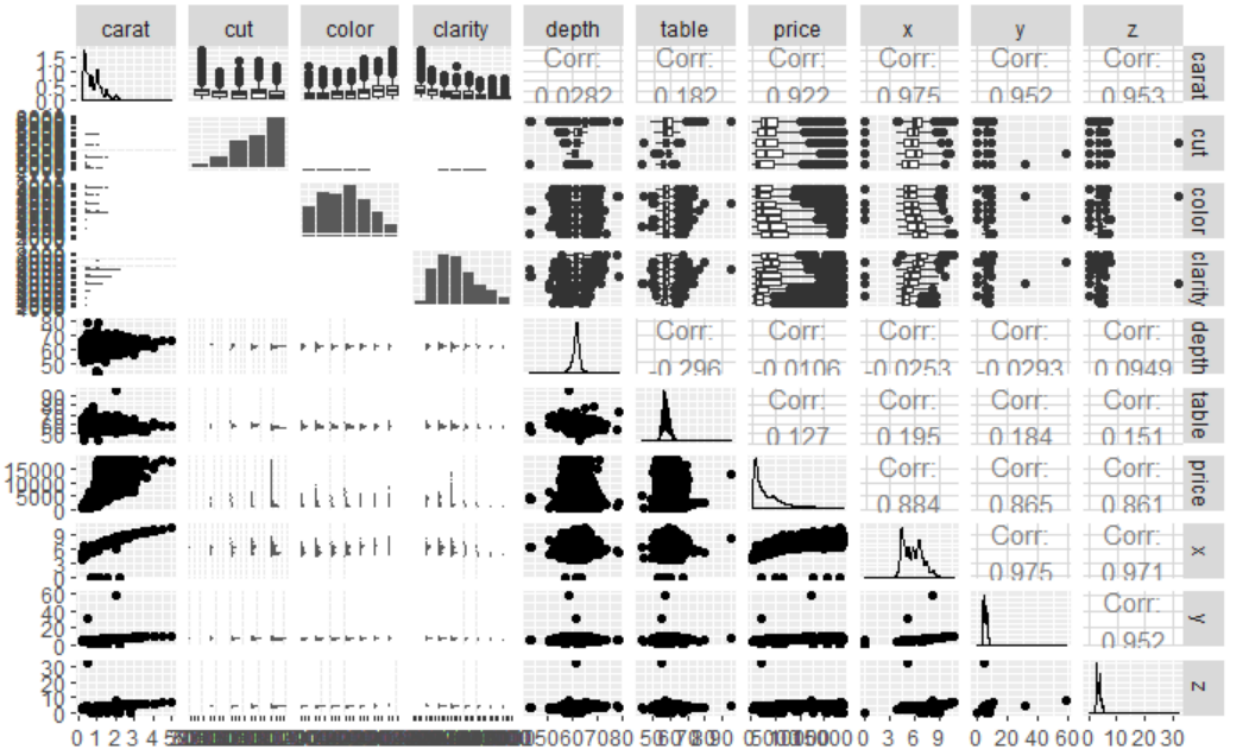
But first lets setup a Hypothesis.

H0 : Diamonds pricing is not effected by other variables such as cut,carat,color

H1 : Diamonds pricing is effected by other variables such as cut,carat,color

Corelation between different variables

We need to plot different corelation graphs in order to understand which variables would be beneficial to study upon and do analysis. Using the GGAlly package we plotted a comprehenvice plot, whose image is below using GGpairs function as it gave us a quick glance over the various co relations that may be present in between the variables.



Observations For layman, let us break down the conclusions from the above graph, and for any statistian, he/she can see evidently that there is high correlation between variables such as price to carat, cut etc.

From the above correlation plot, we observe that:

1. There is a relationship in between log Price and carat.
2. Price exhibits very limited or no relationship with table and depths.
3. Carat is highly correlated to x,y,z. The weight (carat) of a diamond would definitely be affected by its length (x), width(y) and width (z).

Cut distribution

First we saw the summary of the cut distribution which came out to be: Fair -1610 ,Good - 4906 ,Very Good - 12082,Premium-13791 ,Ideal- 21551

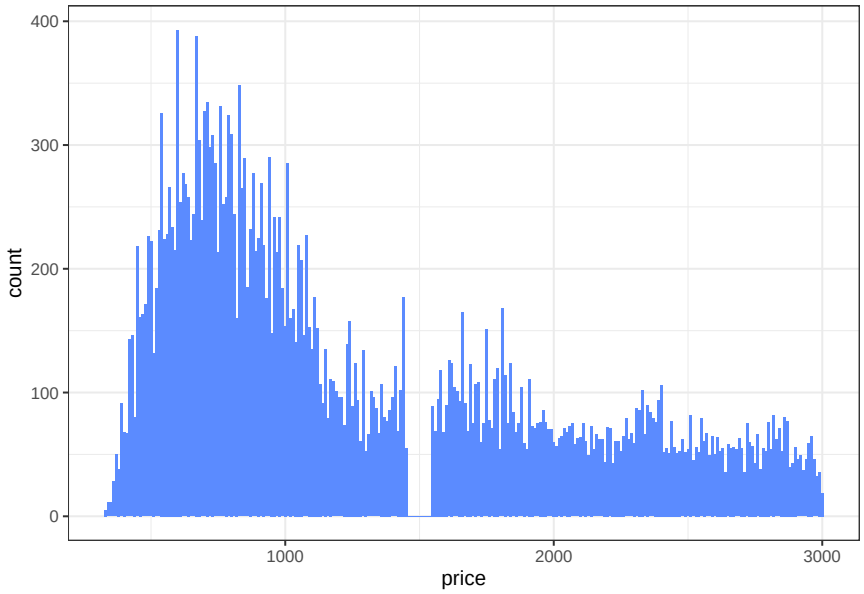
Now we got to know that there are 5 types of cuts , now we plotted each cut wrt price variable(count vs cost graph graph). and found out that these all distributions are rightly skewed.

Price distribution

As our main focus is the price variable, thus, exploring the distribution of the price throughout the dataset would be a good start.

Observation: We plotted the price distribution graph normally (code present in the R file) but noticed something unusual. First There were no diamonds with a price of (between 1450-1550 dollars).Thus a gap is there in the distribution. and the other normal observation is that there is a bulge in the distribution around $750.

To further claim this observation we filtered the diamonds dataset to see all the diamonds less then 3000 dollars (just to get a clear picture on this un-usuality)



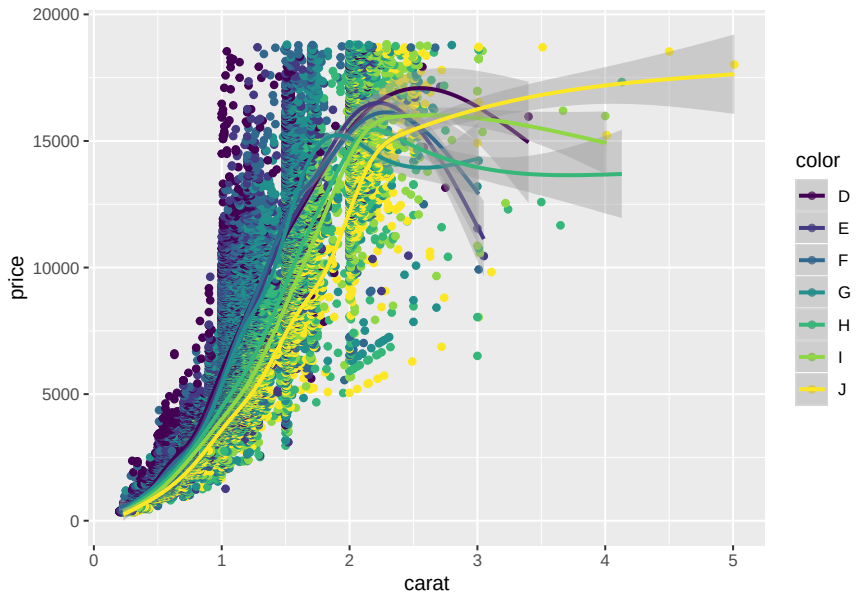
Relation between Price & Cut

Now lets visualise a graph to see what is / or if there is any relation between Price and the cut of the diamond. (plot code present in R file)

Observation : From the plot ,we observed that the highest price diamonds are premium cut, lowest price diamonds are ideal cut.

Price vs carat(with color)

Now after seeing the relationship between Cut and Price, we have another variable in our dataset ie carat which we think may effect the pricing of the dimonds present in the dataset. So just to make sure if this is the case or not we plotted the graph



Observation : It can be depicted from the graph that, price will increase as carat increases. Color-J means the diamond is *colourless* which has highest price for greater carat. Whereas Color D means it has brownish yellow color which has lower price for greater carat(>3.3) compared to J colored diamond. So, color should be considered as a factor of price. Apart from this, we plotted graphs for cut and clarity which has shown similar behaviour. That means carat, cut, clarity and color are the factors for the feature *price*.

Conclusion

We see that there are more smaller diamonds than bigger ones. Also in y and z dimensions, there are outliers. They could be errors or real diamonds that are exceptionally large. Also we find something interesting that somehow we don't have diamonds that are priced around $1500.

From all the distribution graphs we see that the distribution of all the variables are all right skewed and lastly they are multimodal or "spiky". And atlast we came to conclusion that Price is definitely related with carat, cut, clarity and color as evident from the graphs above and from various intermediate graphs present in R file.

Sources / References

For this R exploratory data analysis, we used:-

1. https://ggplot2.tidyverse.org/reference/
2. https://stackoverflow.com (for common errors)
3. https://www.bluenile.com/ca/education/diamonds (We didnt had knowledge about various aspects or measurements of the diamonds, so to get a sound knowledge about diamond's physical properties in real life)