Diamonds

Group2 (Akshita, Khalid, Nilesh)

02/02/2020

1. Understanding the Data

This **diamonds** dataset describes the prices and other variables such as the clarity of diamonds, their depth, size, color etc. **The purpose of the report is to explore the dataset, analyse the trends and record suggestions to improve the data**. The dataset has **53940 rows and 10 columns**, with no missing values (checked using *sum(is.na(diamonds))* function).

Structure of Diamond Dataset:

Firstly, we will describe the general structure of diamonds Dataset describing the variables (their datatypes and levels)

Structure

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    53940 obs. of  10 variables:
##  $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

Using *str(diamonds)* function, we get the datatype and structural details.
Description The *description* of variables are as shown below:
carat: weight of the diamond (0.2–5.01)
cut: quality of the cut with 5 factors: Fair, Good, Very Good, Premium, Ideal
color: diamond colour (D,E,F,G,H,I,J), from D (best) to J (worst)
clarity: measure of clarity of diamond (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
depth: total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)
table: width of top of diamond relative to widest point (43–95)
price: price in US dollars ($326–$18,823)
x: length in mm (0–10.74)
y: width in mm (0–58.9)
z: depth in mm (0–31.8)

2. Analyzing the data

Summary of the diamonds dataset:

Here, we will focus on numeric variables **carat, depth, table, price, x, y and z - giving a sense of their central tendency, dispersion, range, levels, etc.** There are 7 numeric variables (carat, depth, table, price, x, y and z) for which the measure of central tendency, dispersion and range are calculated. Following is the summary of the Factor variables, followed by the numeric variables:

```
##      Carat               Cut        Color
##  Min.   :0.2000   Fair     : 1610   D: 6775
##  1st Qu.:0.4000   Good     : 4906   E: 9797
##  Median :0.7000   Very Good:12082   F: 9542
##  Mean   :0.7979   Premium  :13791   G:11292
##  3rd Qu.:1.0400   Ideal    :21551   H: 8304
##  Max.   :5.0100                     I: 5422
##                                     J: 2808
```

Measures of central tendency

```
##    Measure | carat |  depth |  table |   price |     x |     y |     z
## 1:     Min |   0.2 |     43 |     43 |     326 |     0 |     0 |     0
## 2: 1st Qu. |   0.4 |     61 |     56 |     950 |  4.71 |  4.72 |  2.91
## 3:       - 
## 4:    Mean | 0.798 | 61.749 | 57.457 |  3932.8 | 5.731 | 5.735 | 3.539
## 5:  Median |   0.7 |   61.8 |     57 |    2401 |   5.7 |  5.71 |  3.53
## 6:    Mode |   0.3 |     62 |     56 |     605 |  4.37 |  4.34 |   2.7
## 7:       - 
## 8: 3rd Qu. |  1.04 |   62.5 |     59 | 5324.25 |  6.54 |  6.54 |  4.04
## 9:     Max |  5.01 |     79 |     95 |   18823 | 10.74 |  58.9 |  31.8
```

Measures of Dispersion

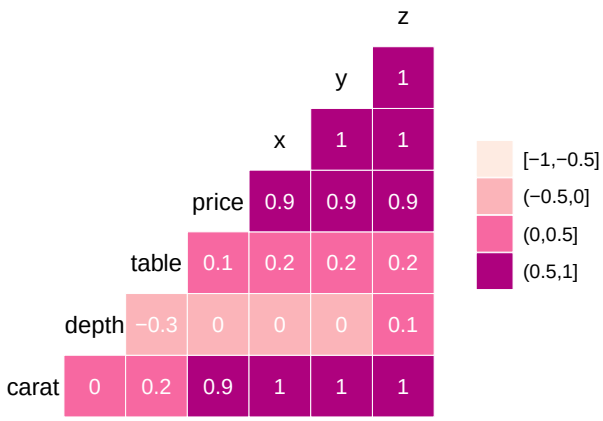
```
##    Measure |    carat | depth | table |       price |       x |      y |      z
## 1:     Var |     0.22 |  2.05 |  4.99 | 15915629.42 |    1.26 |    1.3 |    0.5
## 2:   S.Dev |     0.47 |  1.43 |  2.23 |     3989.44 |    1.12 |   1.14 |   0.71
## 3:     IQR |     0.64 |   1.5 |     3 |     4374.25 |    1.83 |   1.82 |   1.13
## 4:   Range | 0.2 5.01 | 43 79 | 43 95 |   326 18823 | 0 10.74 | 0 58.9 | 0 31.8
```

Summarizing the above measures grouped by class

3. Exploring the Data

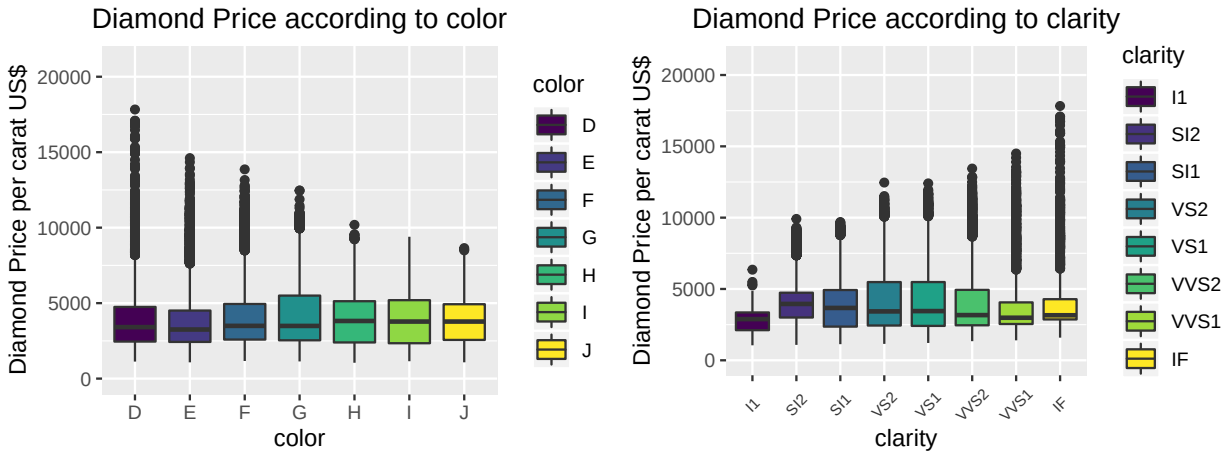
Let's begin by finding the correlation for all the variables

```
## Warning in ggcorr(diamonds[, 1:10], nbreaks = 4, label = TRUE, palette =
## "RdPu", : data in column(s) 'cut', 'color', 'clarity' are not numeric and were
## ignored
```



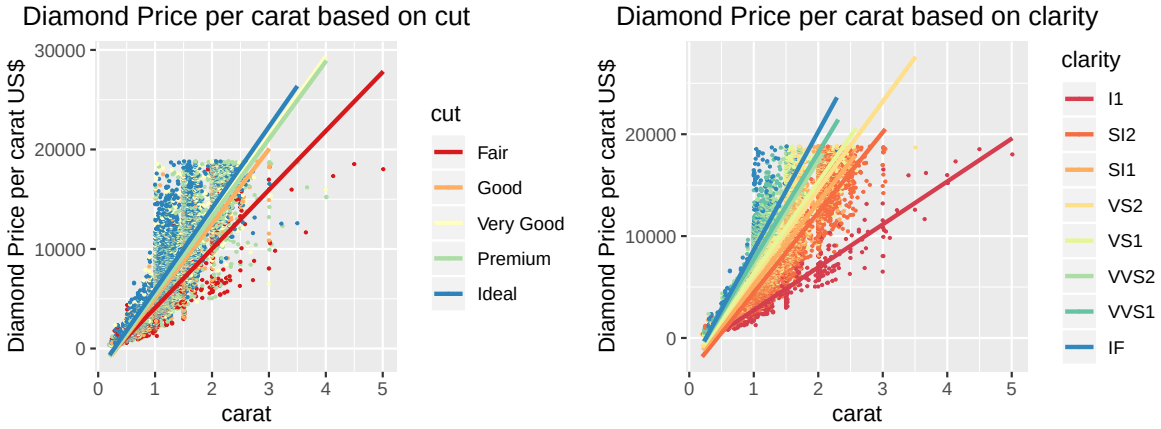
We see that *price* & *carat* are highly correlated. Also, x,y,z (i.e the length, width and depth) are correlated with price, which makes sense in real life as bigger the diamond, higher it will sell for! Since we know price & carat are highly correlated, we'll analyse the effect of other variables on price~carat

1) **Effect of color on price per carat**



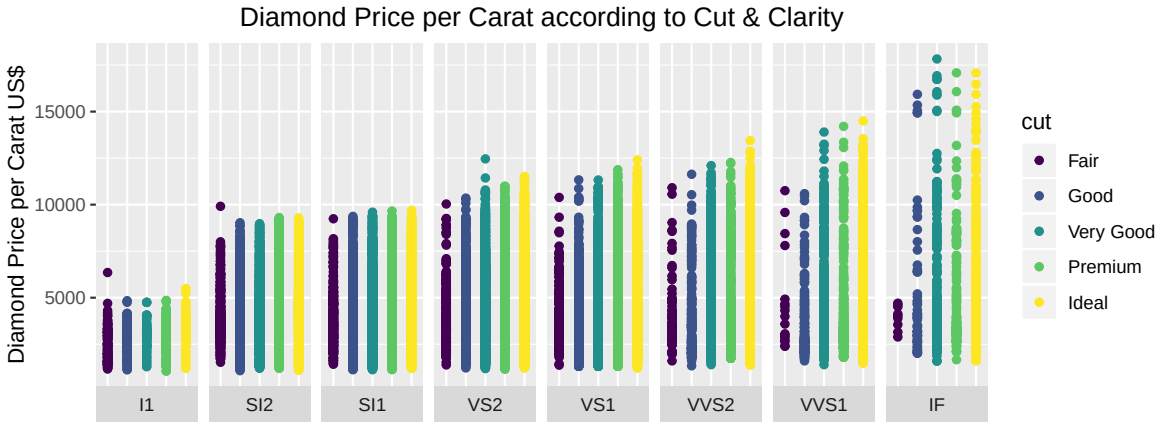
From the above graph, it is observed that the color variation doesn't have any effect on Price per carat of the Diamond. So, analysis is done on cut and clarity factors.

2) **Effect of cut on price per carot** vs **effect of clarity on price per carot**



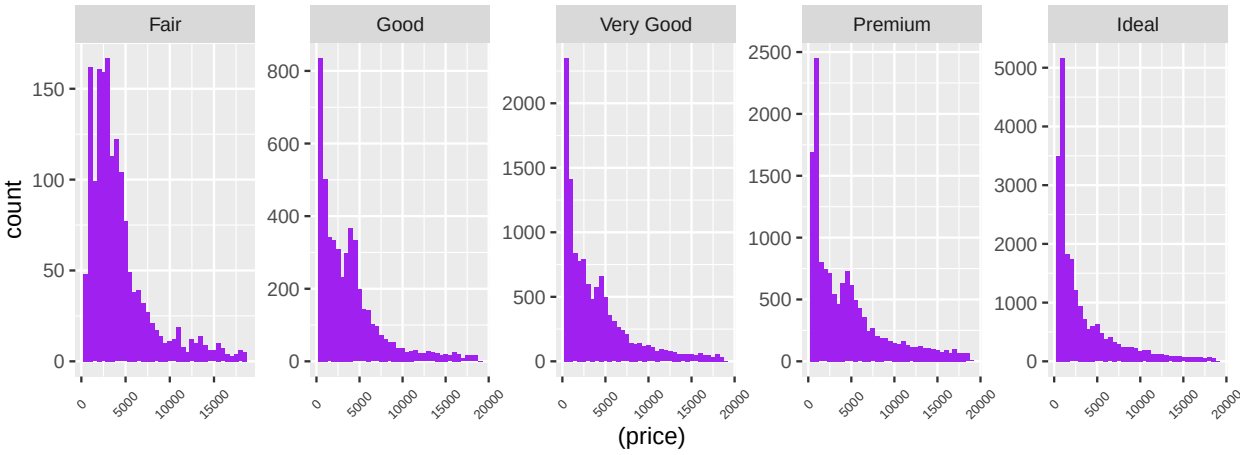
It is observed that the Blue Line (*Ideal Cut* & *IF Clarity*, respectively) is more sensitive w.r.t carat and Price per carat. This is can be further clarified and explained using the following graph:

3) **Effect of cut and clarity together on price per carot**



From the above graph, it is observed that as the cut moves towards *Ideal Cut* and clarity towards *IF (Highest Clarity)*, the price shoots more than *15000* US$, but this is not the case in every situation, probably due to other factors such as customer demand, market economy, environmental impacts, etc.

4) **Trend of different cuts**



It is evident that as the price increases, the number of diamonds i.e., count, on y-axis decreases exponentially. Which is understood in real-world scenario i.e., few people can only afford to buy expensive diamonds.

4. Conclusion

Conclude

1. Contrary to general notion, color has minimal effect on the price of diamonds.
2. With clarity and cut, it is observed that the price increases steeply with high clarity and Ideal cut.
3. There are more outliers in this dataset, so the dataset should be collected in a way that it should convey an accurate information for future prediction.

Future Scope

1. For ease of understanding to to get better insight about the diamonds dataset, the variable name should be properly defined (x, y, z).
2. Time-Series data could be introduced with other variables to make more precise predictions & inferences.