Assignment 1: Case 3

Jyotsna Dalal, Pallav Walia, Shashsank Singhal

04-02-2020

Introduction to Data

The dataset Diamonds has 53940 observtaion with 11 variables. Out of 11 variables 3 are categorical variables i.e. cut, color & clarity and rest of the varaibles like price, carat etc. are continuous variables. The data set looks like:

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1 0.23  Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2 0.21  Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3 0.23  Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4 0.290 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5 0.31  Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6 0.24  Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

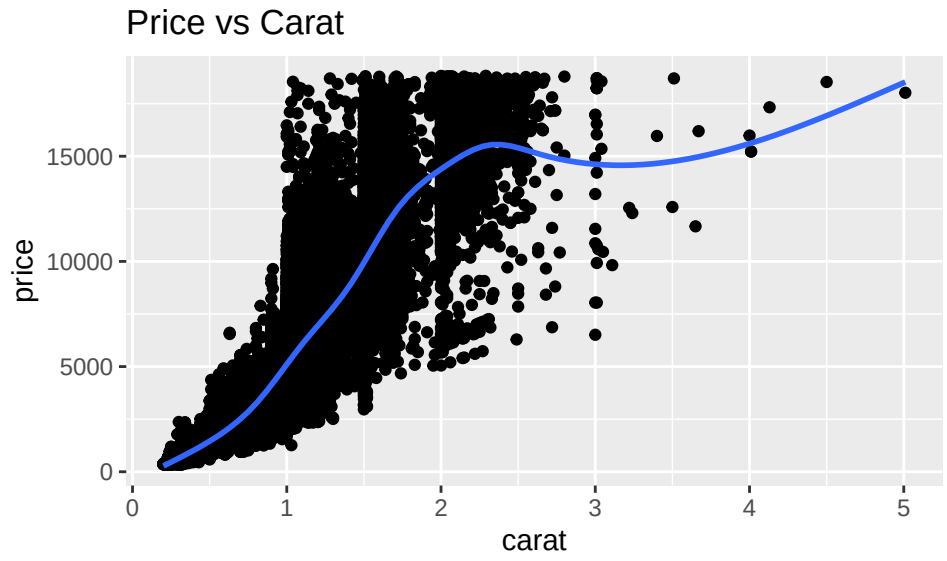
The table displays the detailed information about the prices of diamonds and its properties. The price ranges from $326 to $18823 with a mean of $3933 and approx. 60% of the diamonds are of "premium" or "ideal" cut. Half of the diamonds are less than 0.7 carat with highest value at 5.01 and the distribution of data is right skewed.

Therefore, initially we can focus our analysis on observing a relationship between price and its properties.

Analysis

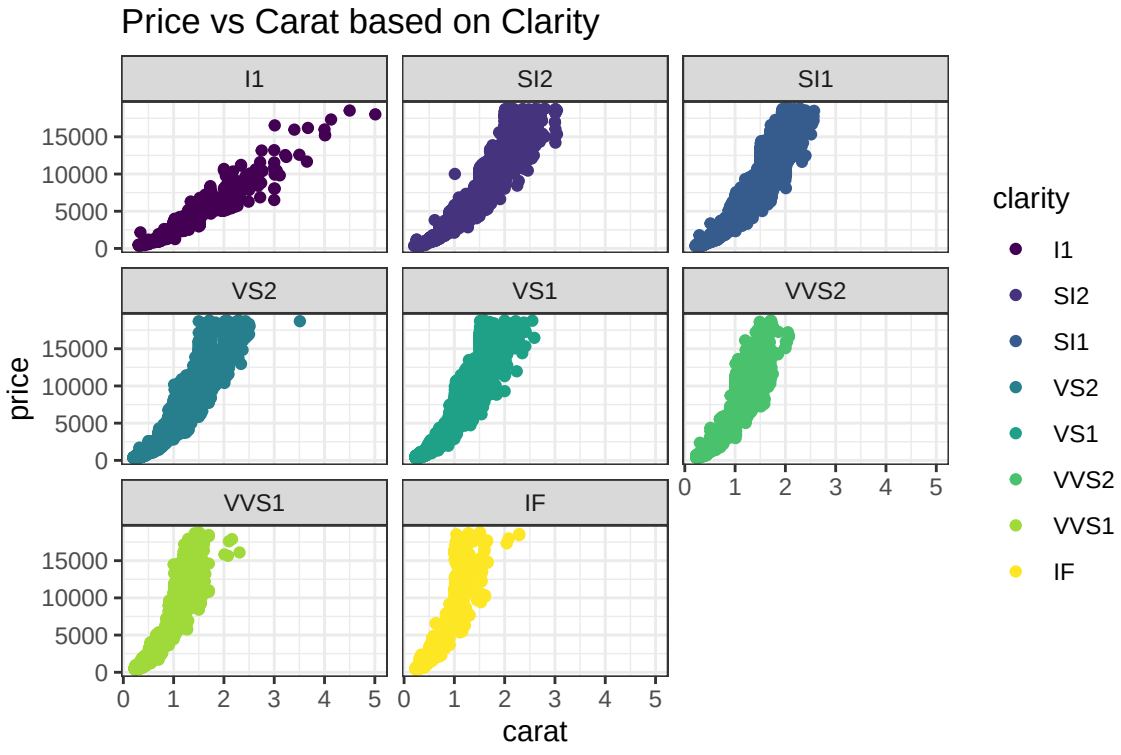
The dataset has many variables, we will focus on specific variables to get an understanding how these variables are related to Price and with each other. We will take price, carat, clarity, color, depth & cut into consideration. Let's have a closer look on each variable-

Price vs Carat



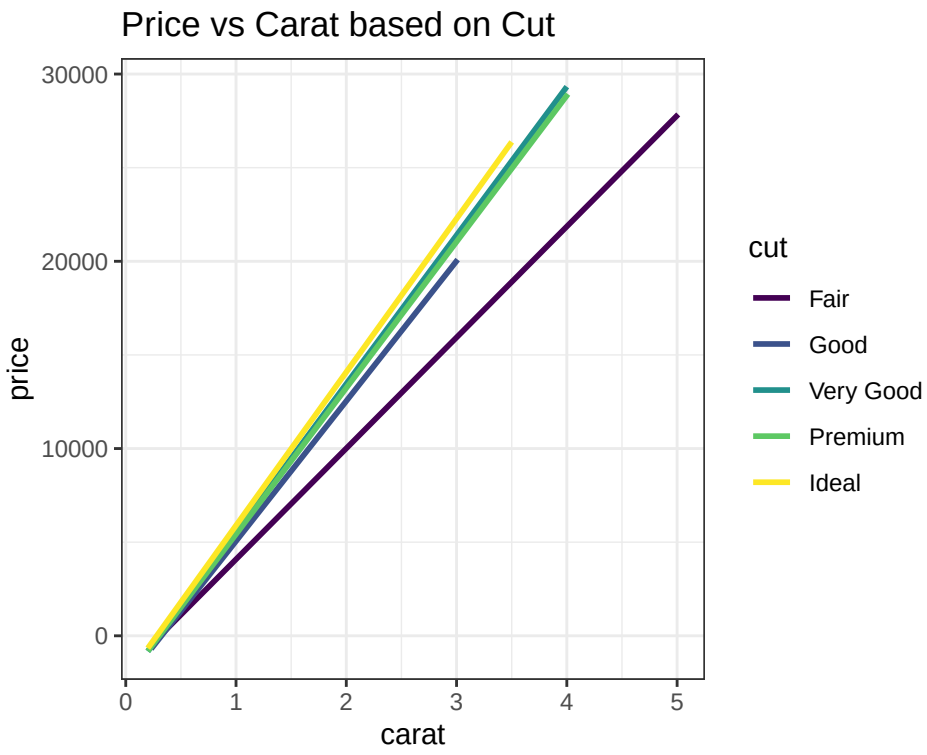
The scatter plot between price and carat shows a postive correation with some outliers. Therefore we further need to investigate the effect of other variables on price.

Price vs Carat based on Clarity



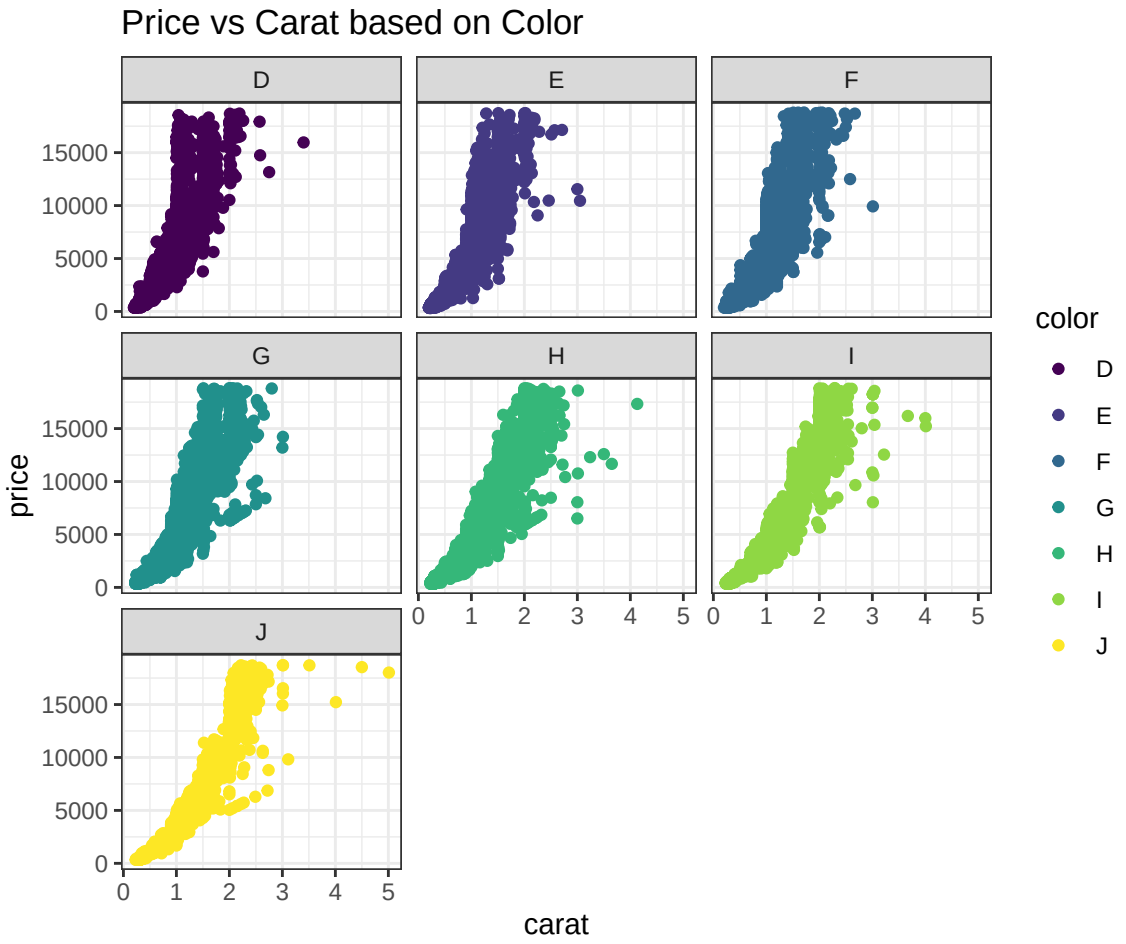
Interestingly the price of higher clarity grade diamond (IF) increases exponentially as compared with low quality diamond (I1) which has a linear relationship. As we move from lower clarity grade to higher clarity grade, the slople of the graph becomes steeper.

Price vs Carat based on Cut



Based on the graph, it looks like only fair cut category of diamonds are impacting the price as the slope is flatter than the other cut types. We can not say much from this information on the impact of cut type on price of diamonds.

Price vs Carat based on Color



The graph depicts that the price of higher color grade diamond (D) increases exponentially as compared with lower color grade (J). As we move from higher color grade to lower color grade, the slople of the graph becomes flatter.

Also when we look for cooreation between the continuous variables, only price and carat appear to be coorelated as evident from the following results-

```
##             price      carat       depth      table
## price  1.0000000 0.92159130 -0.01064740  0.1271339
## carat  0.9215913 1.00000000  0.02822431  0.1816175
## depth -0.0106474 0.02822431  1.00000000 -0.2957785
## table  0.1271339 0.18161755 -0.29577852  1.0000000
```

Conclusion

From the above analysis we can infer that there are multiple factors which act as price driver for diamonds. Carat seems to be the one of the most influencial factors among all.However, it is difficult to determine how much the price of a diamond can move with a unit change in any of the independent variables or predictors. To determine the relationship between price and other variables we need to dig deeper into the data and try to establish a linear regression model which will help us in estimating the relation between variables.