Exploratory Data Analysis on dataset : Diamonds

Group 17: Neethu Mariya, Nada Saiyed, Shadh Shanavas

1/31/2020

This is a document on the Exploratory Data Analysis of the default dataset Diamonds available with the package ggplot2 in RStudio to discover an intersting feature of the diamond price.

The Data

This data frame contains the prices and 9 attributes of Diamonds collected from around 54,000 pieces of this valuable gem. The features considered in the analysis are carat, cut, color, clarity, depth, table(width of the diamond top) and the fundamental dimensions of the diamonds (length, width and depth). The carat defines the weights of the diamonds in the range 0.2 to 5.01. The cut denotes the quality of the cut for the diamonds catagorized Fair, Good, Very Good, Premium & Ideal. Clarity is a measure of how clear the diamond is, I1 being the worst, through SI2, SI1, VS2,VS1, VVS2, VVS1, IF being the best. Color describes the colour of the diamond, from D(best) to J(worst). x,y,z gives the length, width and depth in mm. Depth calculates the percentage of depth of the diamond in relation to its average length and width. Table gives the width of top of diamond relative to widest point, ranges from 43 to 95.

```
## Data Frame Summary
## diamonds
## Dimensions: 53940 x 10
## Duplicates: 146
##
## ------------------------------------------------------------------------------------------------------------------------
## No   Variable            Stats / Values                Freqs (% of Valid)      Graph                  Valid    Missing
## ---- ------------------- ----------------------------- ----------------------- ---------------------- -------- ---------
## 1    carat               Mean (sd) : 0.8 (0.5)         273 distinct values     :                      53940    0
##      [numeric]           min < med < max:                                      : .                    (100%)   (0%)
##                          0.2 < 0.7 < 5                                         : :
##                          IQR (CV) : 0.6 (0.6)                                  : : .
##                                                                                : : : .
##
## 2    cut                 1. Fair                        1610 ( 3.0%)                                  53940    0
##      [ordered, factor]   2. Good                        4906 ( 9.1%)           I                      (100%)   (0%)
##                          3. Very Good                  12082 (22.4%)           IIII
##                          4. Premium                    13791 (25.6%)           IIIII
##                          5. Ideal                      21551 (40.0%)           IIIIIII
##
## 3    color               1. D                           6775 (12.6%)           II                     53940    0
##      [ordered, factor]   2. E                           9797 (18.2%)           III                    (100%)   (0%)
##                          3. F                           9542 (17.7%)           III
##                          4. G                          11292 (20.9%)           IIII
##                          5. H                           8304 (15.4%)           III
##                          6. I                           5422 (10.1%)           II
##                          7. J                           2808 ( 5.2%)           I
##
## 4    clarity             1. I1                           741 ( 1.4%)                                  53940    0
##      [ordered, factor]   2. SI2                         9194 (17.0%)           III                    (100%)   (0%)
##                          3. SI1                        13065 (24.2%)           IIII
##                          4. VS2                        12258 (22.7%)           IIII
##                          5. VS1                         8171 (15.2%)           III
##                          6. VVS2                        5066 ( 9.4%)           I
##                          7. VVS1                        3655 ( 6.8%)           I
##                          8. IF                          1790 ( 3.3%)
##
## 5    depth               Mean (sd) : 61.7 (1.4)        184 distinct values                :           53940    0
##      [numeric]           min < med < max:                                                 :           (100%)   (0%)
##                          43 < 61.8 < 79                                                   :
##                          IQR (CV) : 1.5 (0)                                             . :
##                                                                                         : :
##
## 6    table               Mean (sd) : 57.5 (2.2)        127 distinct values          :                53940    0
##      [numeric]           min < med < max:                                           :                (100%)   (0%)
##                          43 < 57 < 95                                               :
##                          IQR (CV) : 3 (0)                                           : :
##                                                                                     : :
##
## 7    price               Mean (sd) : 3932.8 (3989.4)   11602 distinct values     :                      53940    0
##      [integer]           min < med < max:                                        :                      (100%)   (0%)
##                          326 < 2401 < 18823                                      :
##                          IQR (CV) : 4374.2 (1)                                   : : .
##                                                                                  : : : : . . .
```

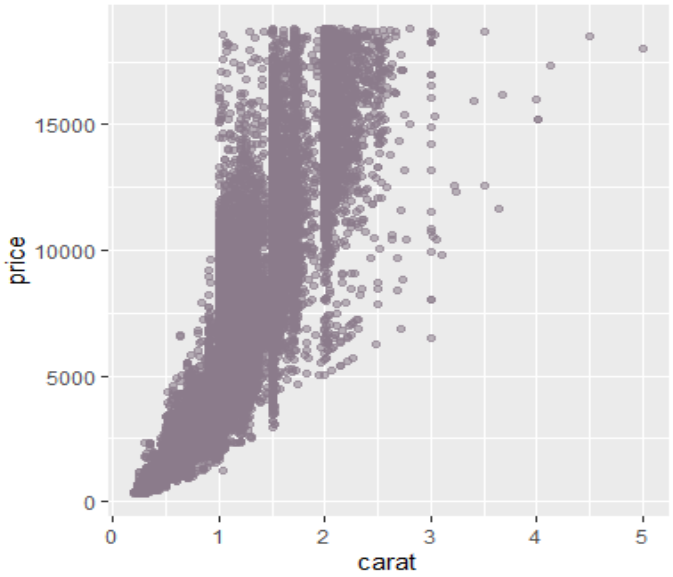
```
##
## 8    x                 Mean (sd) : 5.7 (1.1)         554 distinct values            :               53940    0
##      [numeric]         min < med < max:                                             : .             (100%)   (0%)
##                        0 < 5.7 < 10.7                                               : : :
##                        IQR (CV) : 1.8 (0.2)                                         : : :
##                                                                                   . : : : :
##
## 9    y                 Mean (sd) : 5.7 (1.1)         552 distinct values     :                      53940    0
##      [numeric]         min < med < max:                                        : :                    (100%)   (0%)
##                        0 < 5.7 < 58.9                                          : :
##                        IQR (CV) : 1.8 (0.2)                                    : :
##                                                                                : :
##
## 10   z                 Mean (sd) : 3.5 (0.7)         375 distinct values       :                    53940    0
##      [numeric]         min < med < max:                                          :                    (100%)   (0%)
##                        0 < 3.5 < 31.8                                         : :
##                        IQR (CV) : 1.1 (0.2)                                   : :
##                                                                               : :
## ------------------------------------------------------------------------------------------------------------------------
```

Analysis

Don’t Get Ripped Off When Buying Diamonds!

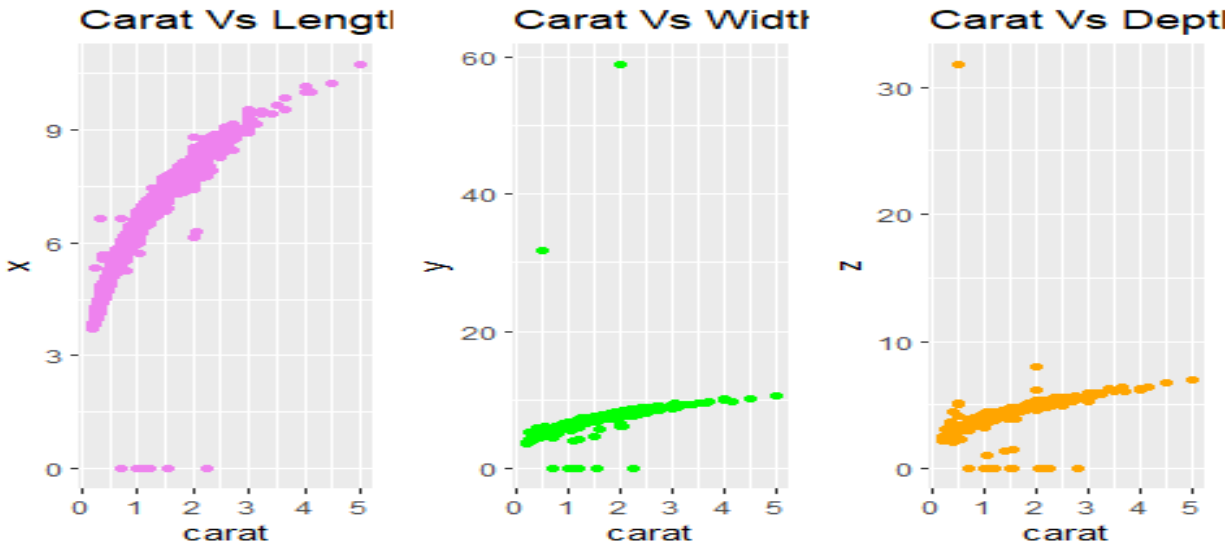
Carat is definitely the most determining ’C’s when it comes to the price of a diamond! But should that be the major determining factor for you as a shopper? Well, lets get into the details before getting ripped off while gifting this precious stone to your loved ones.

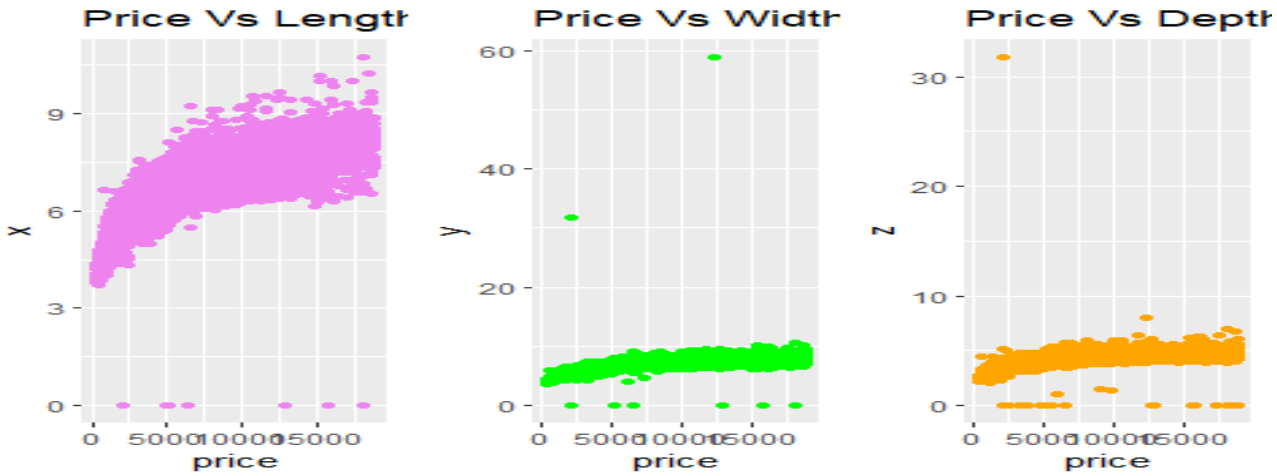
Let’s analyze how the price varies with carats.



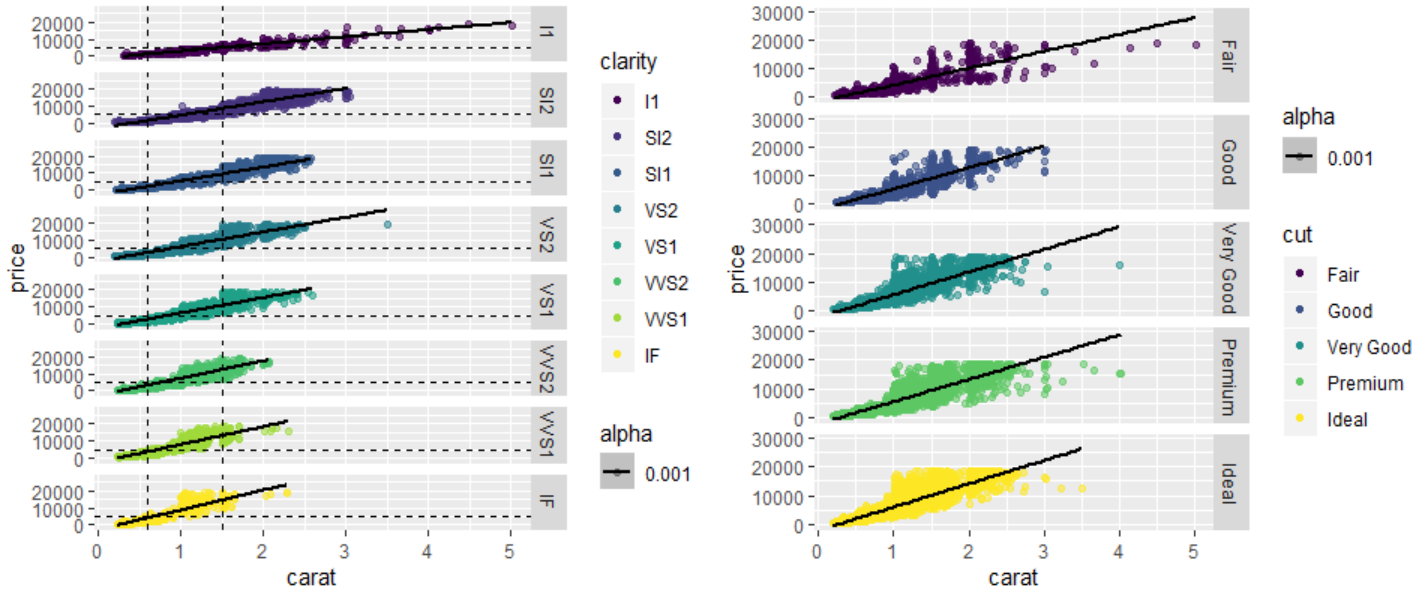
We see that the price varies exponentially with carat.

Does the increase in carat have a visual impact from a shopper’s perspective?



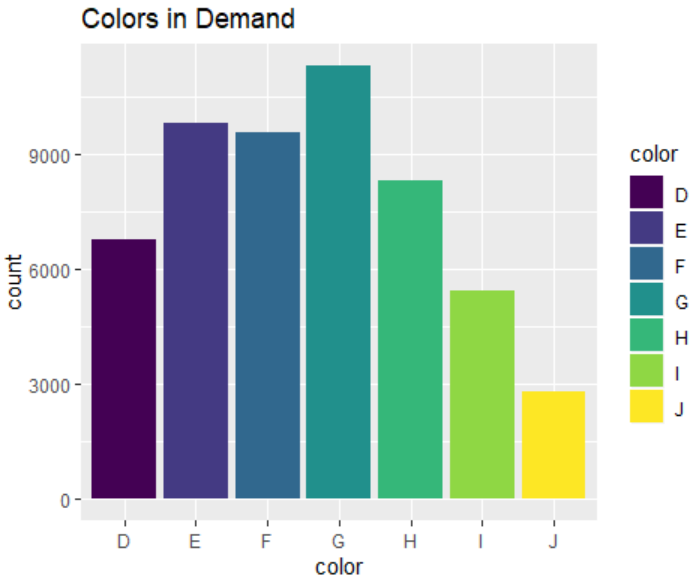


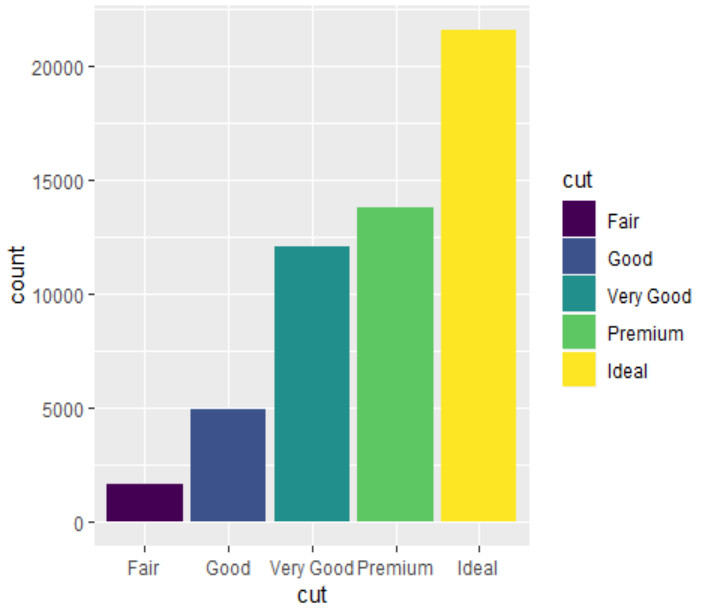
Yeah! We get bigger stones for higher price. But,wait! Is that worth the money? There are other C's that should be considered. The Cut and Clarity of the stone have their own pride!

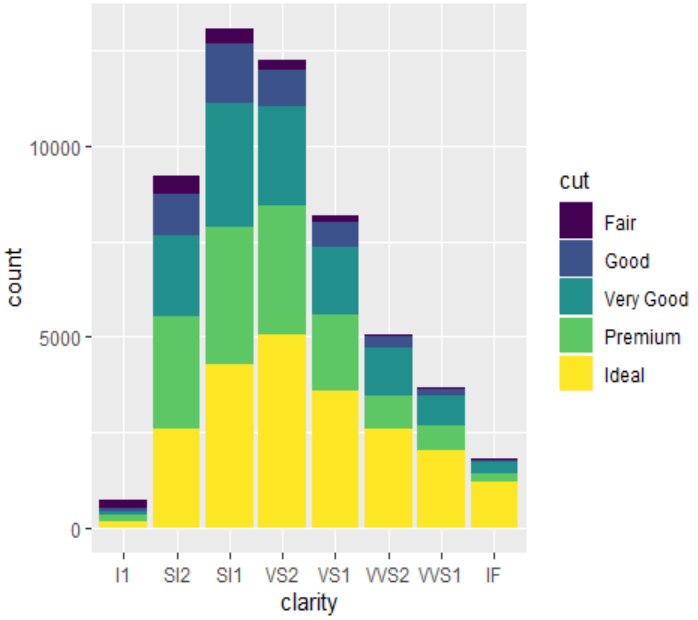


What does the market demand?

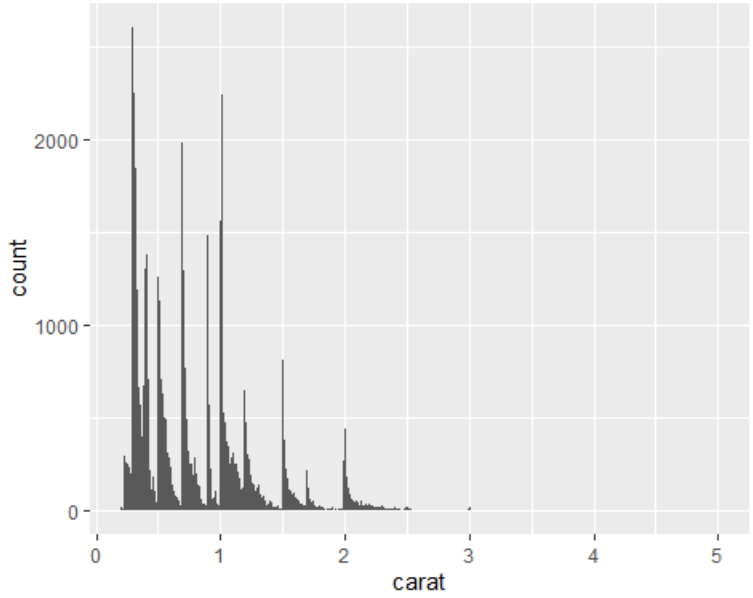
It would definitely answer many of our questions by knowing what the general trend is. To a first-time shopper this should help to decide on which 'C' to consider important and how to trade-off among them.







This must be fascinating!



Why are there more diamonds at whole carats and common fractions of carats? Wondering!? This interesting plot related back to the 'Magic Sizes'. How often have you seen stones that weigh 0.48 or 0.98 carats? By design, such stones are rare oddities because they fall below the "magic sizes" that can fetch more money. Instead, what you would observe from the graph is that the bulk of the stones are cut to nice figures like 0.50 or 1.00 carats.

Conclusion

Here, we see that the price of diamond varies exponentially with carat. ie: the cost of a diamond with carat 2.00 is significantly higher than the price of **two** carat 1.00 diamonds. We also infer that for a fixed budget, the clarity and cut is compromised for bigger carats. So by blindly chasing after a psychological 1.00 carat mark, you often settle for lower clarity/ cut stones which are more likely to have inclusions visible to the naked eye. It is quite clear from the graph that the market demand is dominated by the ideal cut diamonds(40%) and the ones having clarity SI2 through VS1. We now understand the difference and meaning of these factors in a diamond and with this, we can now economically trade-off among cut, clarity and carat to own a brilliantly sparkling diamond fitting to your budget. So, now that you know what to look for, don't wait, go get these gem stones for yourself or your loved ones without getting ripped.