Exploratory analysis of the Diamonds dataset

1. **What information is in the dataset? What do the measures of cut, clarity, and colour mean? Which are the best and the worst?**

```
Console   Terminal ×   Background Jobs ×
R 4.2.2 · ~/
> str(diamonds)
tibble [53,940 × 10] (S3: tbl_df/tbl/data.frame)
 $ carat  : num [1:53940] 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 [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
 $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
 $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
> head(diamonds)
# A tibble: 6 × 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  0.29 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
>
```

Diamonds data set contains information of 53,940 diamonds. For each diamond dataset provides 10 different variables namely, carat, cut, colour, clarity, depth, table, price, x, y, z. Among which cut, clarity and colour are the categorical variables having data type ‘Ord.factor’.

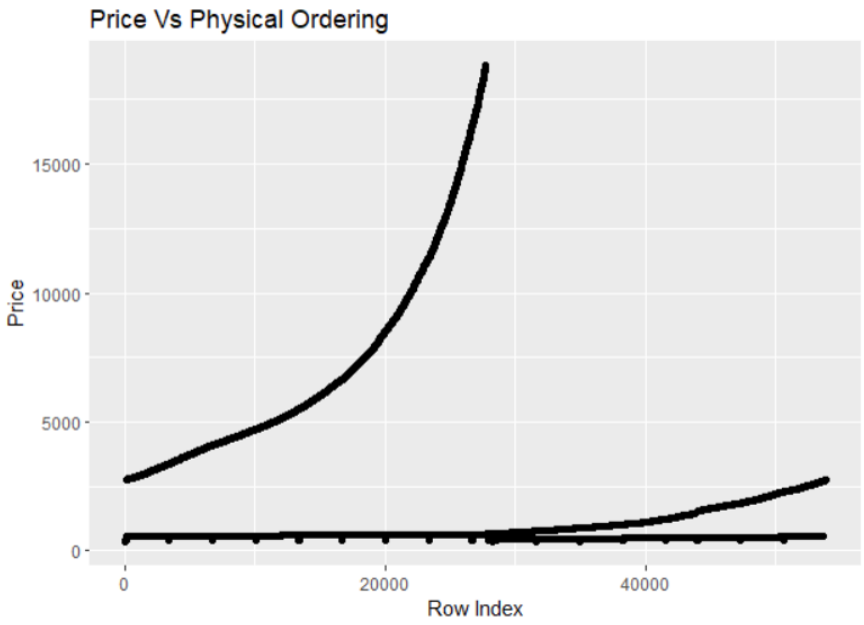
1. Cut:
   A diamond’s cut refers to how well-proportioned the dimensions of a diamond are, and how these surfaces, or facets, are positioned to create sparkle and brilliance [1].
   Dataset has 5 levels of variable Cut ‘Fair<Good<Very Good<Premium<Ideal’.
   The worst type of cut is Fair and the best type is Ideal.
2. Clarity:
   Diamond clarity is the assessment of small imperfections on the surface and within the stone [1].
   Dataset has 8 levels of variable Clarity ‘I1 < SI2 < SI1 < VS2 < VS1 < VVS2 < VVS1 < IF’.
   I1 is the worst type and IF is the best.
   I1, SI2 and SI1 has inclusions that can be seen by naked eye. As we go further (VVS2, VVS1) it becomes difficult even for the trained eyes to see the difference.
3. Colour:
   Colourless diamonds are very rare to find. In fact, diamonds are found in almost any naturally occurring colour, including grey, white, yellow, green, brown, and pink [1].

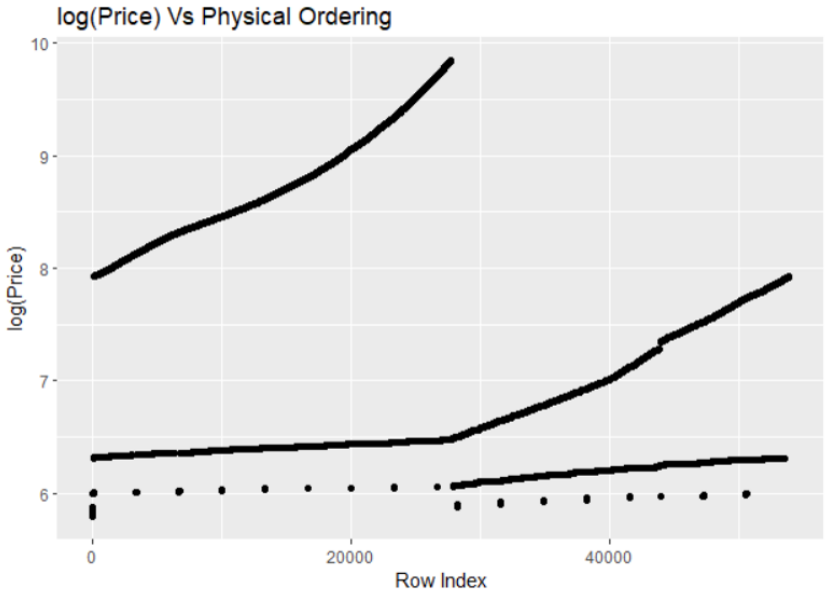
Dataset has 7 levels of variable Colour 'D < E < F < G < H < I < J'.
D being the highest graded and J being the lowest.
J level is nearly colourless and the light yellow tone cannot be easily seen by naked eye unless compared side by side with diamond of level I.
Colour difference between D, E and F is so minute that only gemmologist can detect.

To my opinion, these different levels of colour and clarity which appear exactly the same to the naked eye doesn't affect the value of diamond significantly. But there might be a slight difference in value as these diamonds have difference in their making cost.

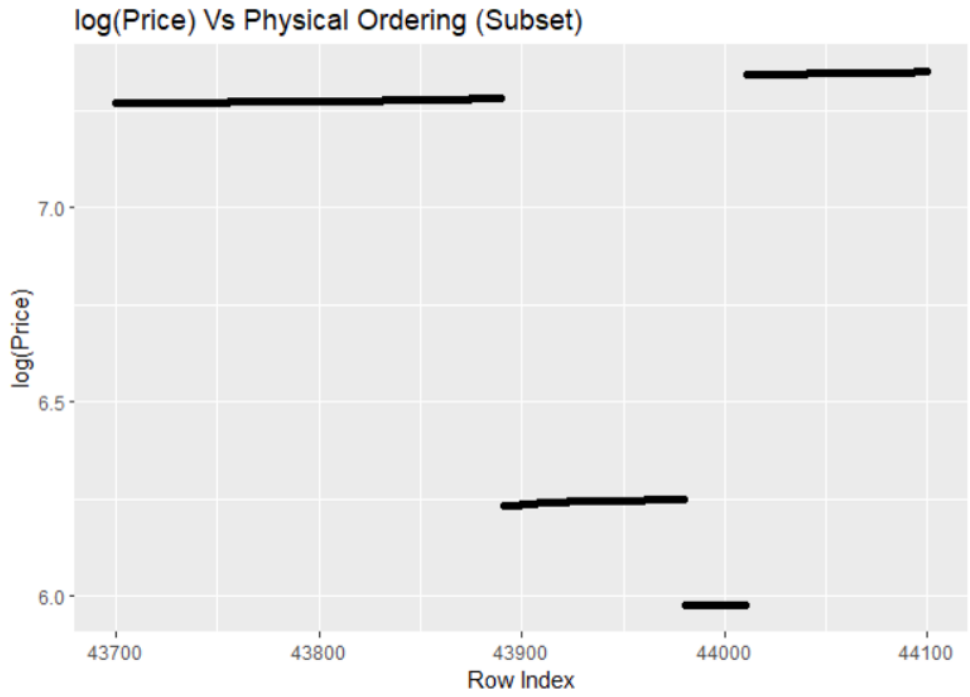
2. **Does the dataset look as if it is correctly constructed? What basic checks could we do?**

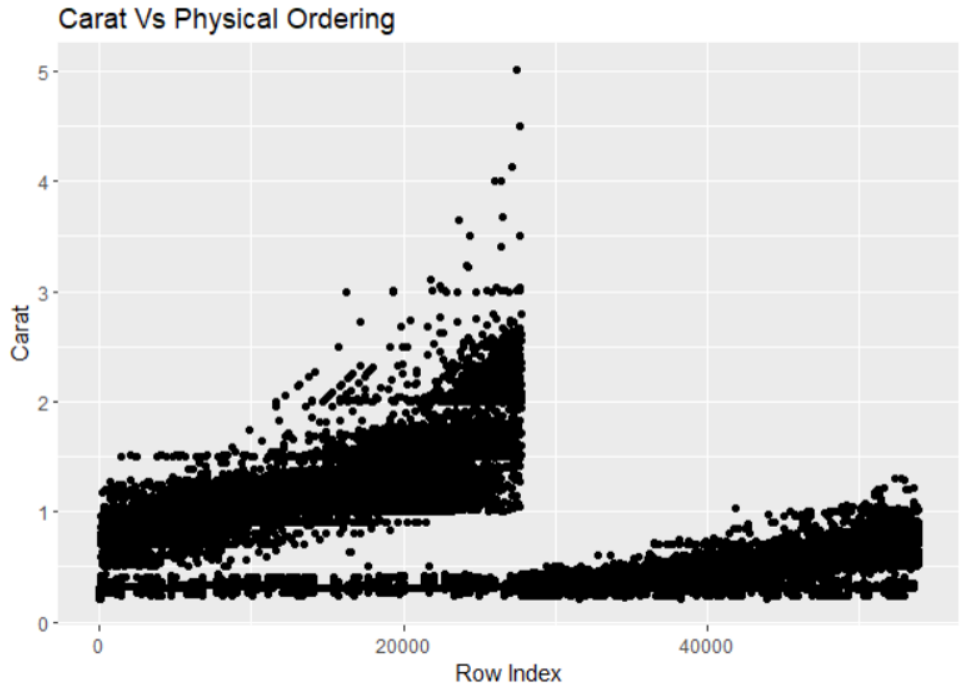
2.1 Plot price and weight against the physical ordering of the data.

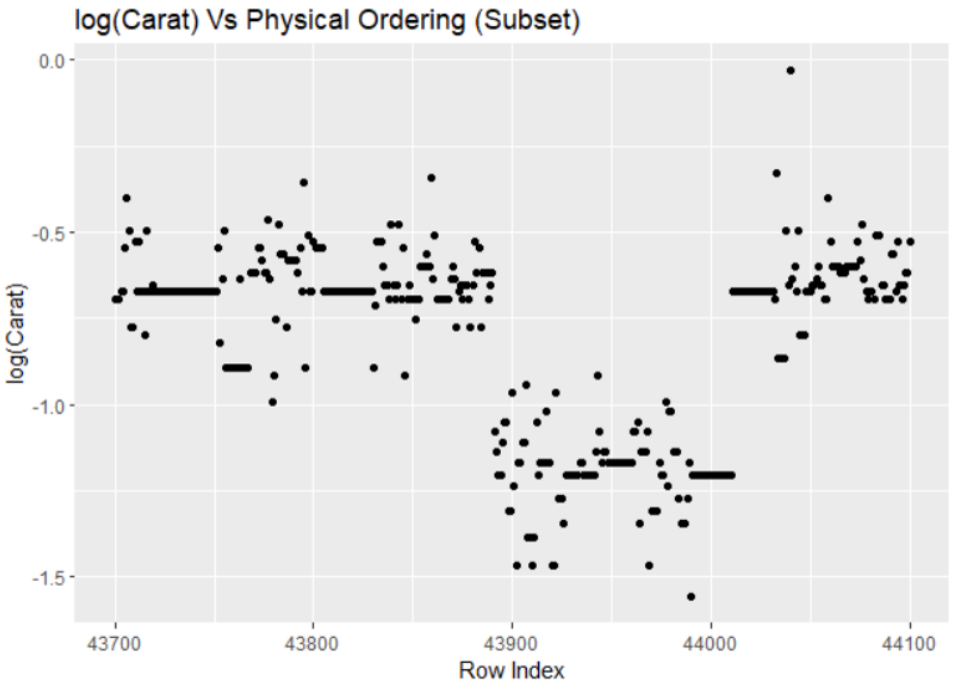




There seem to be two halves of the dataset, and it looks as if there has been a problem during data collection of the second half of the data. There is a missing range of prices.

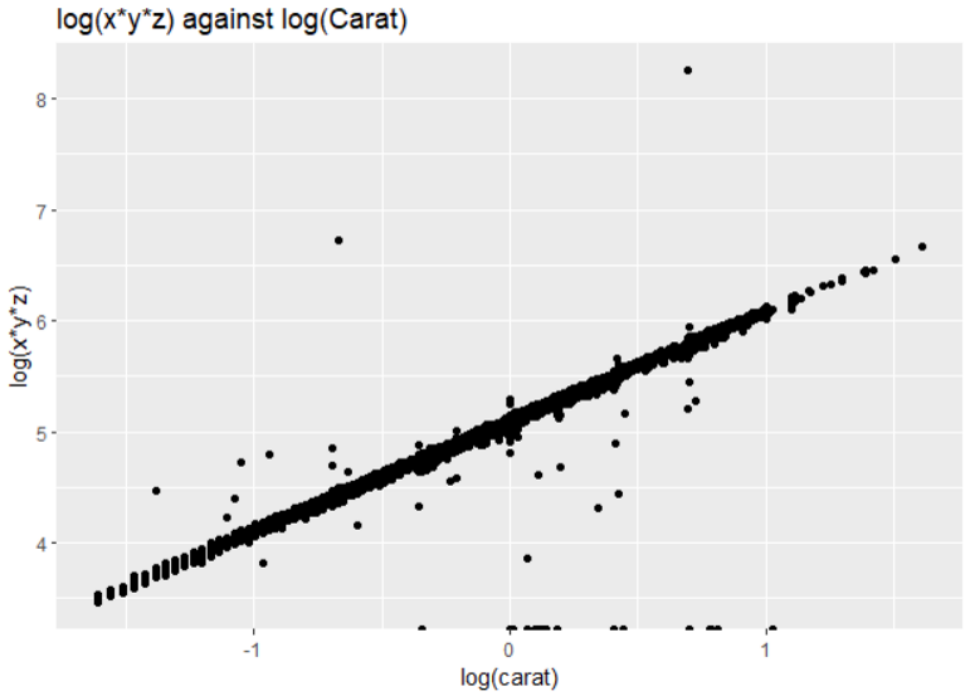






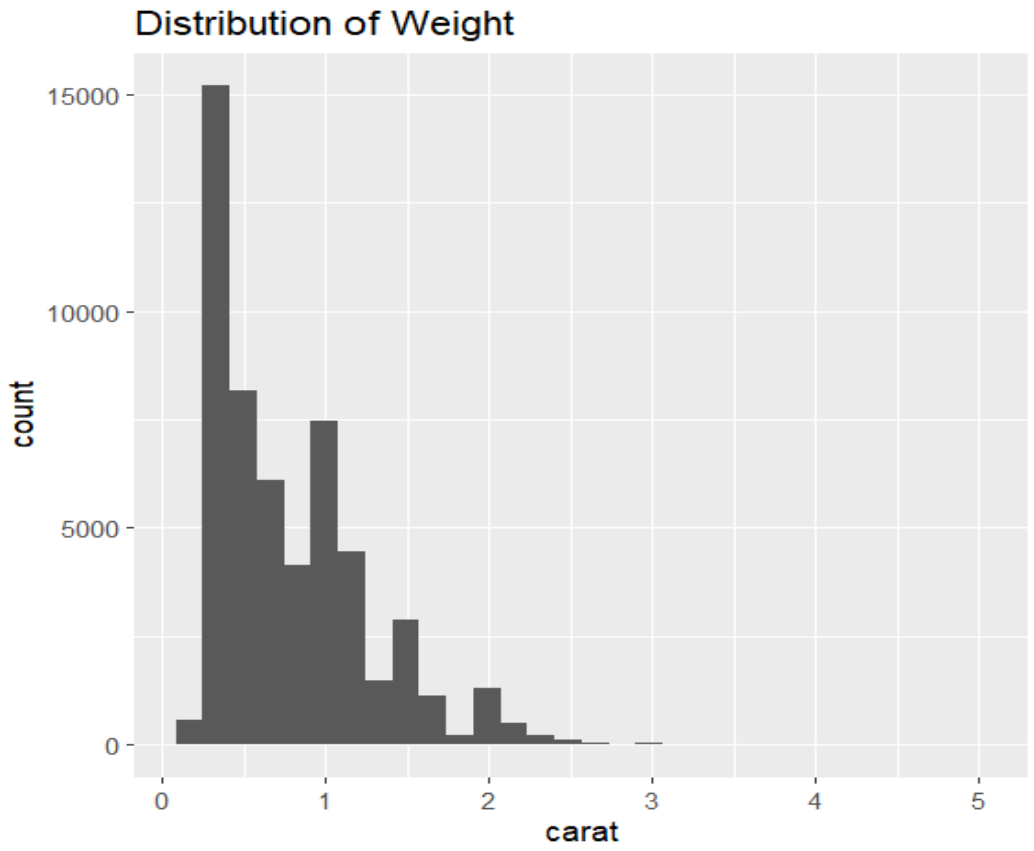
When plotted log(carat) Vs physical ordering of subset, we can observe that they form clusters of weights in carats.

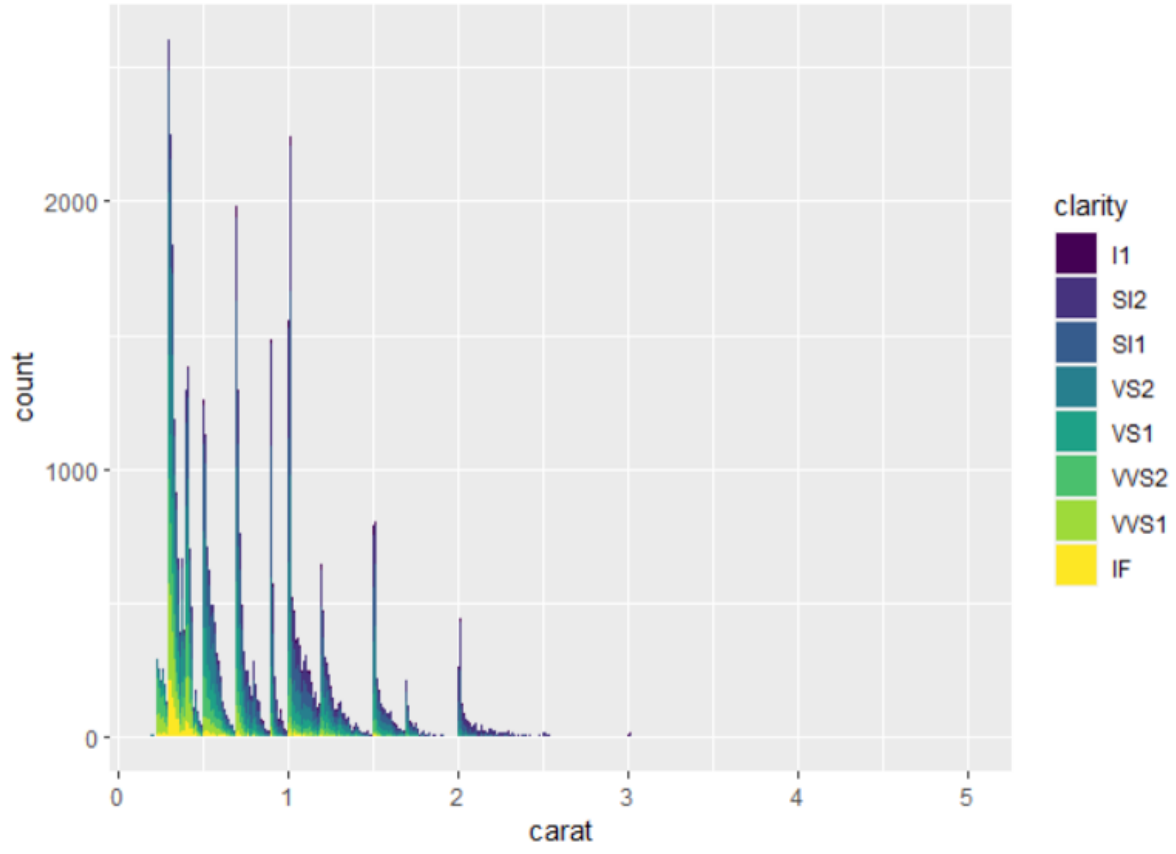
2.2 Checking consistency of x, y, and z measurements with weight (carat)

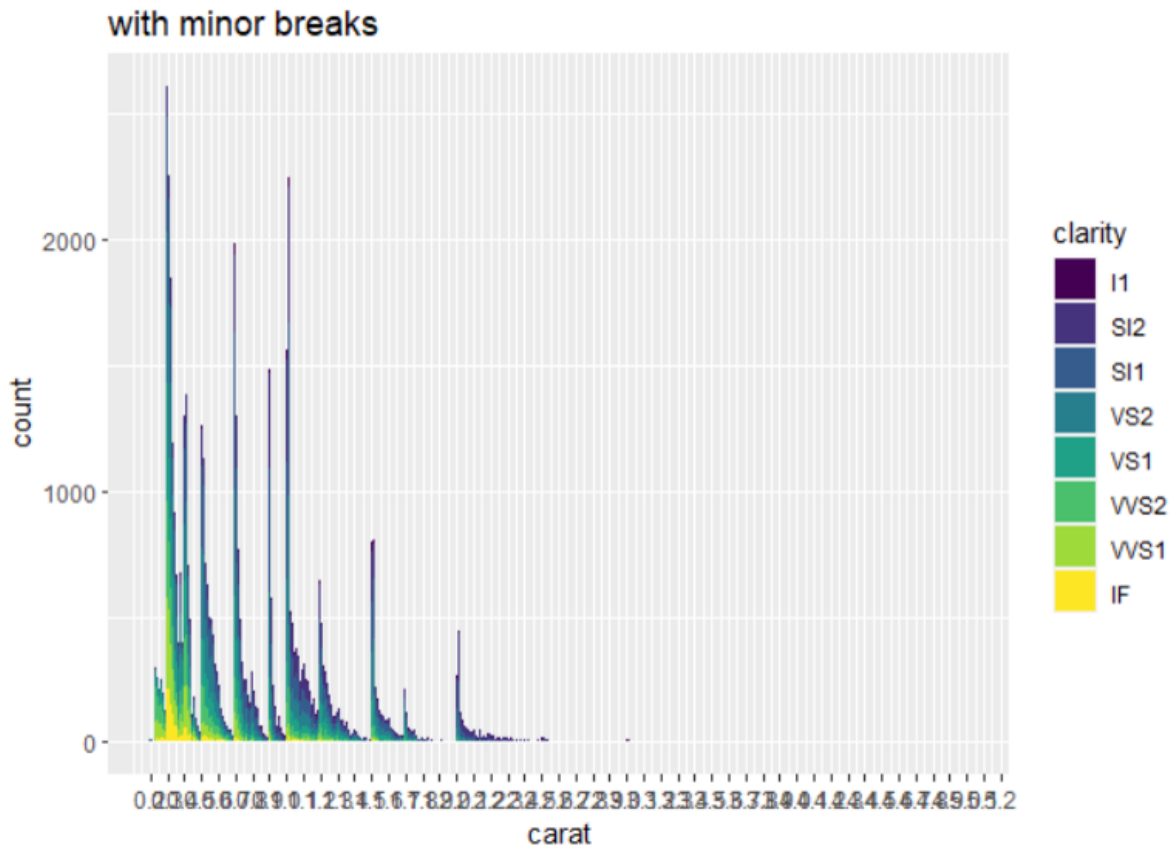


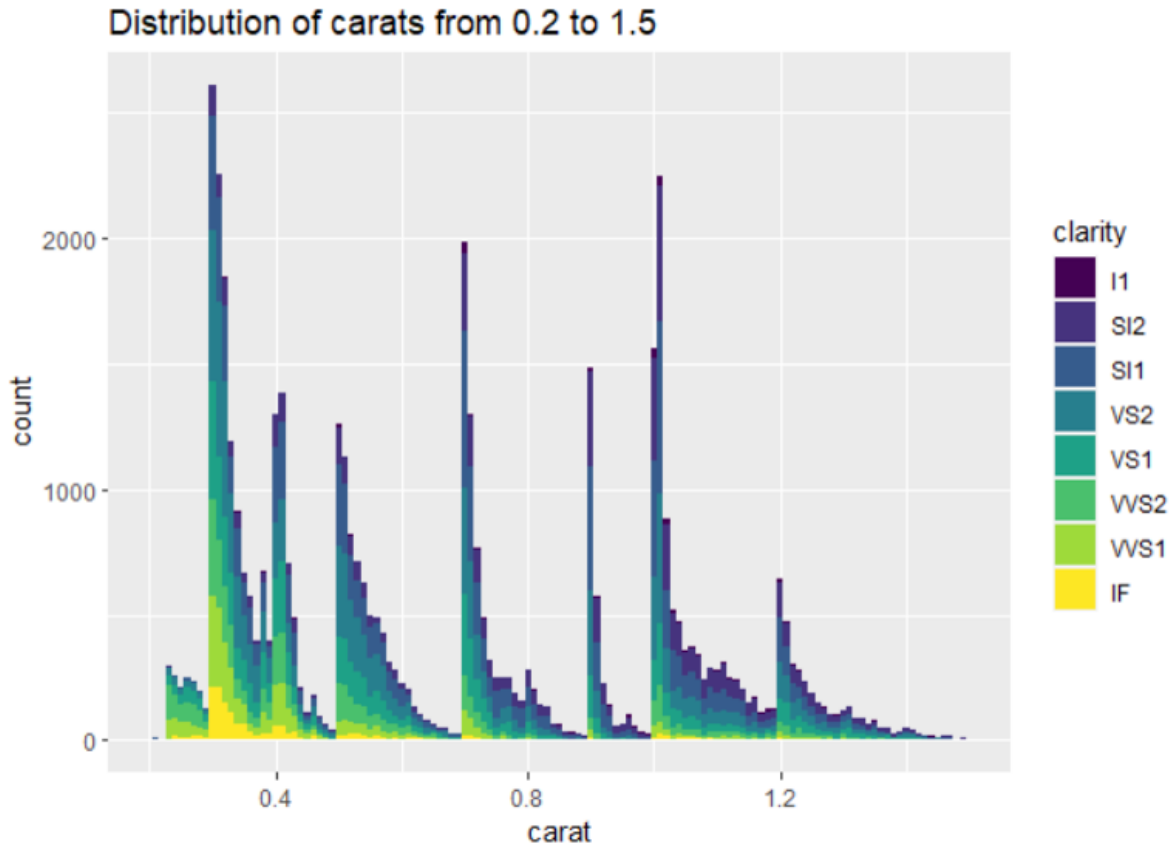
Some diamonds are on the 0 scale of log(x*y*z) on Y-axis. This shows that the error is in x, y, z measurements. This is not consistent with the rest of the diamonds that shows weight (carat) is proportional to x*y*z.

2.3 Studying the distributions of the variables



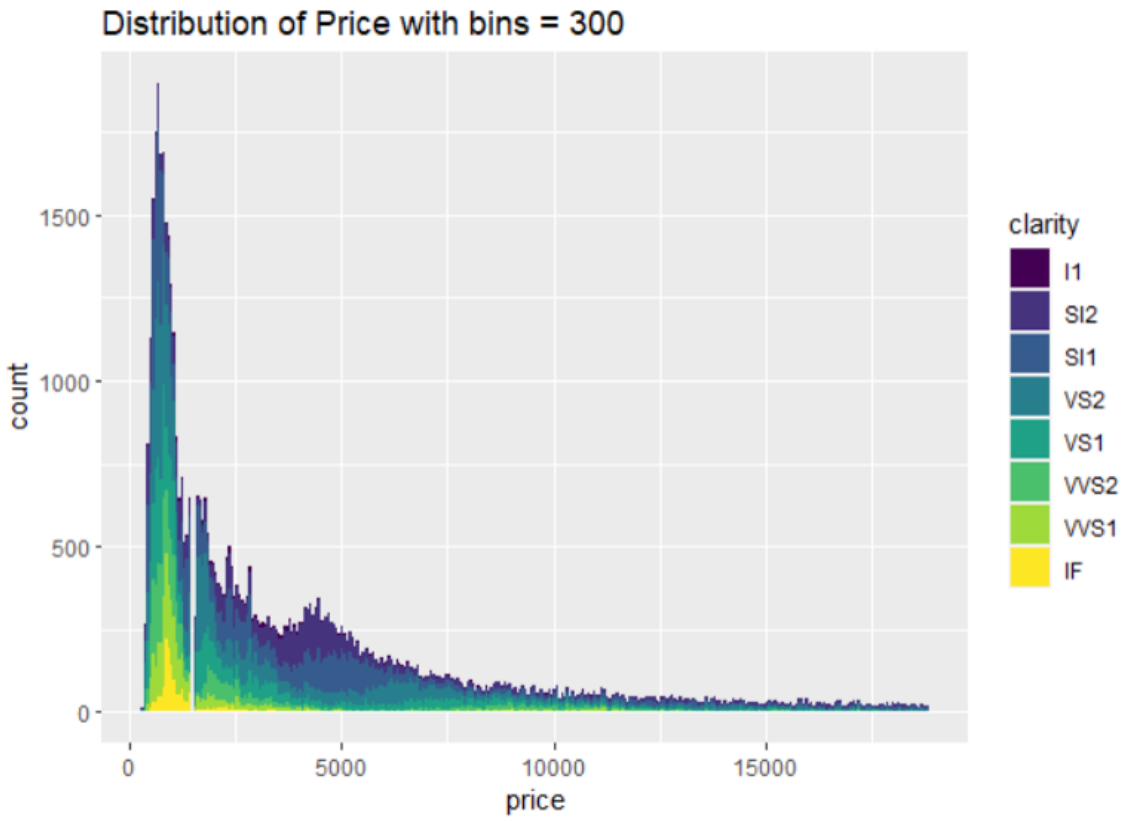


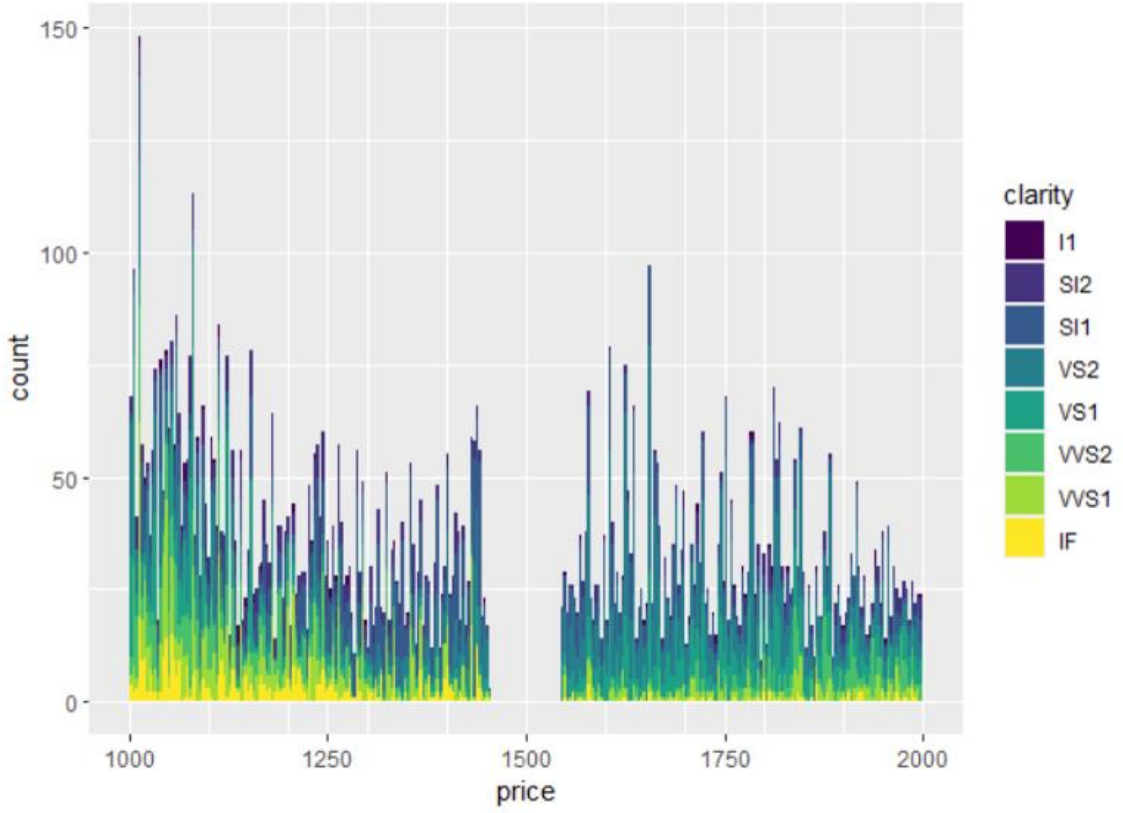




The picks are at the positions they are because we observe that certain standard is maintained while cutting the diamonds, they are kept above certain carats.

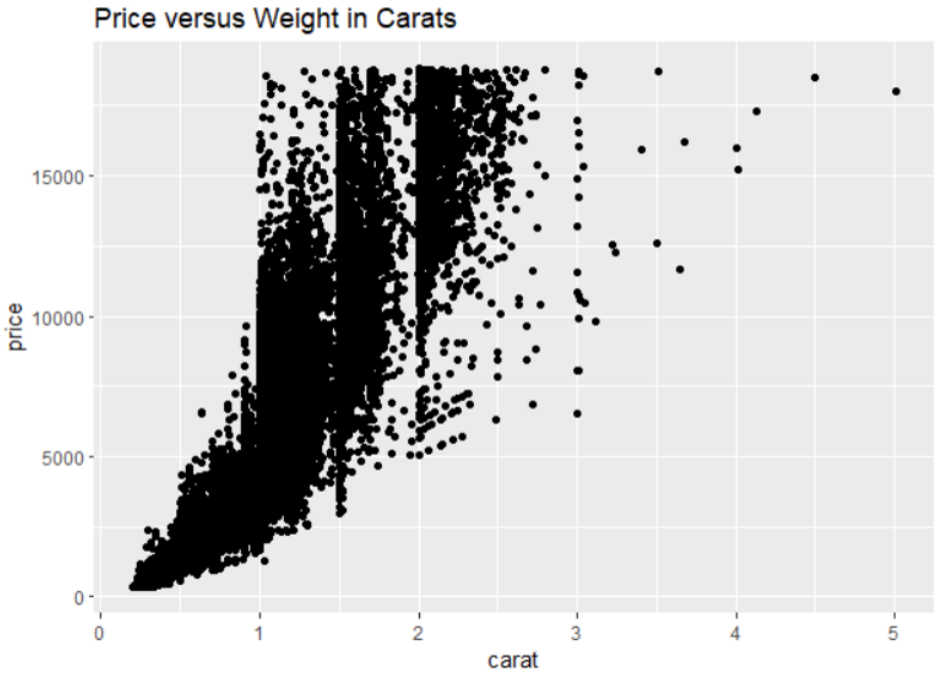
2.4 Checking the distribution of prices





The gap is between 1450 and 1550. This gap corresponds to the same gap we observed in Price against physical ordering plot. It seems like no diamonds at all are priced between 1450 and 1550.

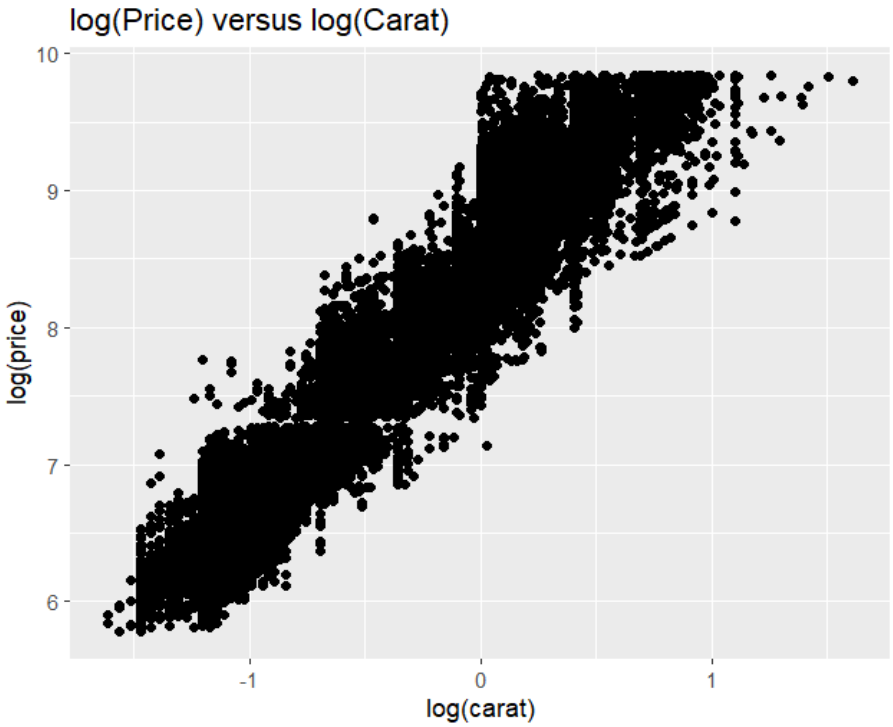
3. **How is weight (in carats) related to price?**



3.1 According to Messing's theory, price is proportional to exp (C x weight1/3). Is this plausible?

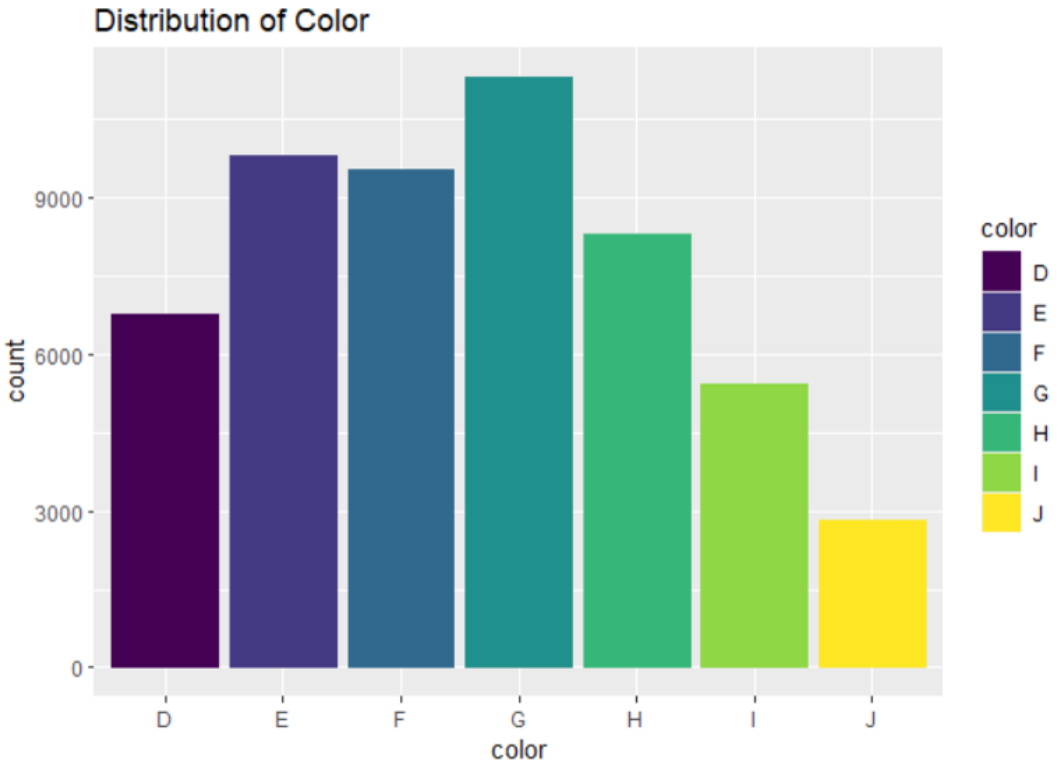
We can observe a slight curvature. Points doesn't lie along a straight line. On a log scale, even a slight curvature can be quite a large deviation.

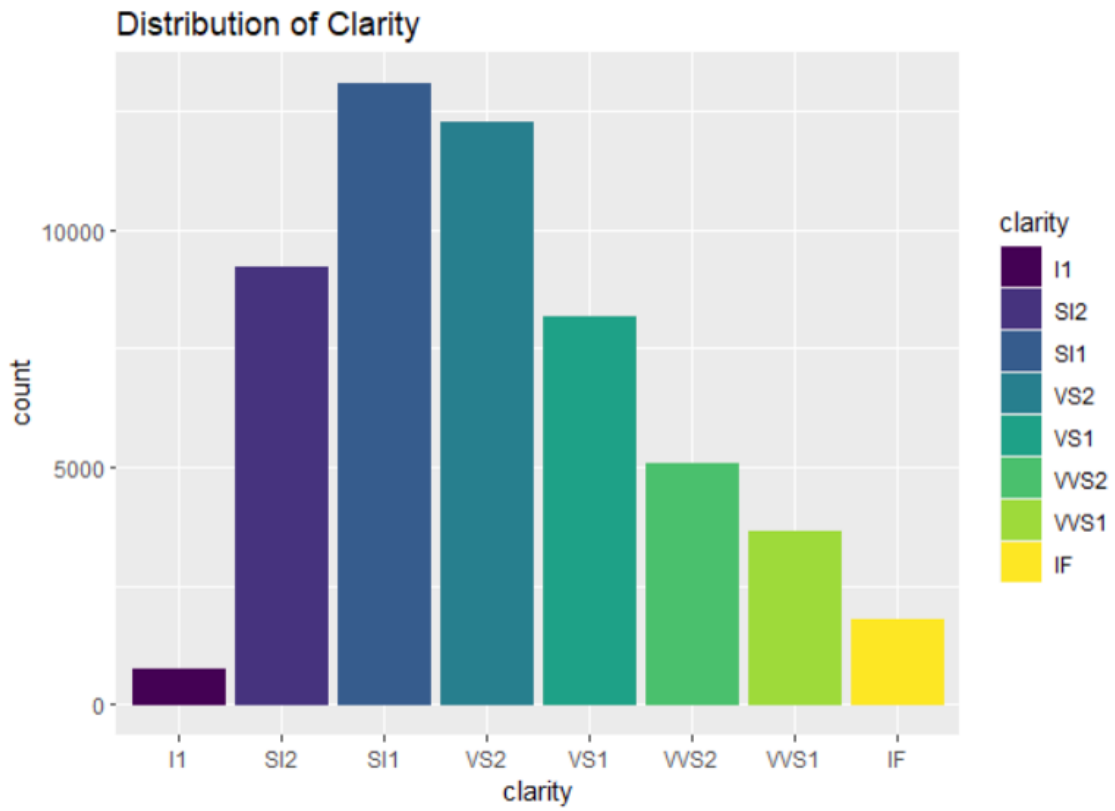
3.2 Might a power law fit the data better?

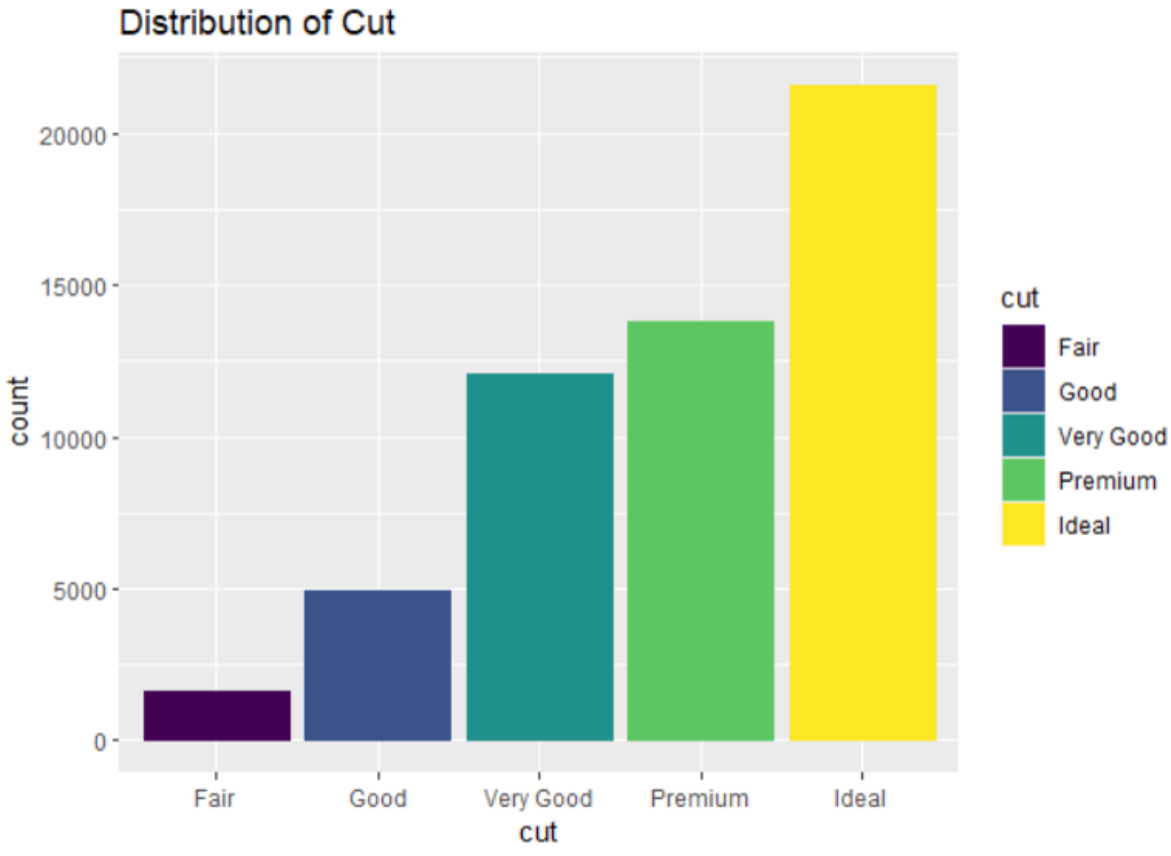


The log-log plot looks more like a straight line than the previous plot. There is a wide scatter of points, and other factors besides weight affect the price.

4. What are the distributions of clarity, colour, and cut?

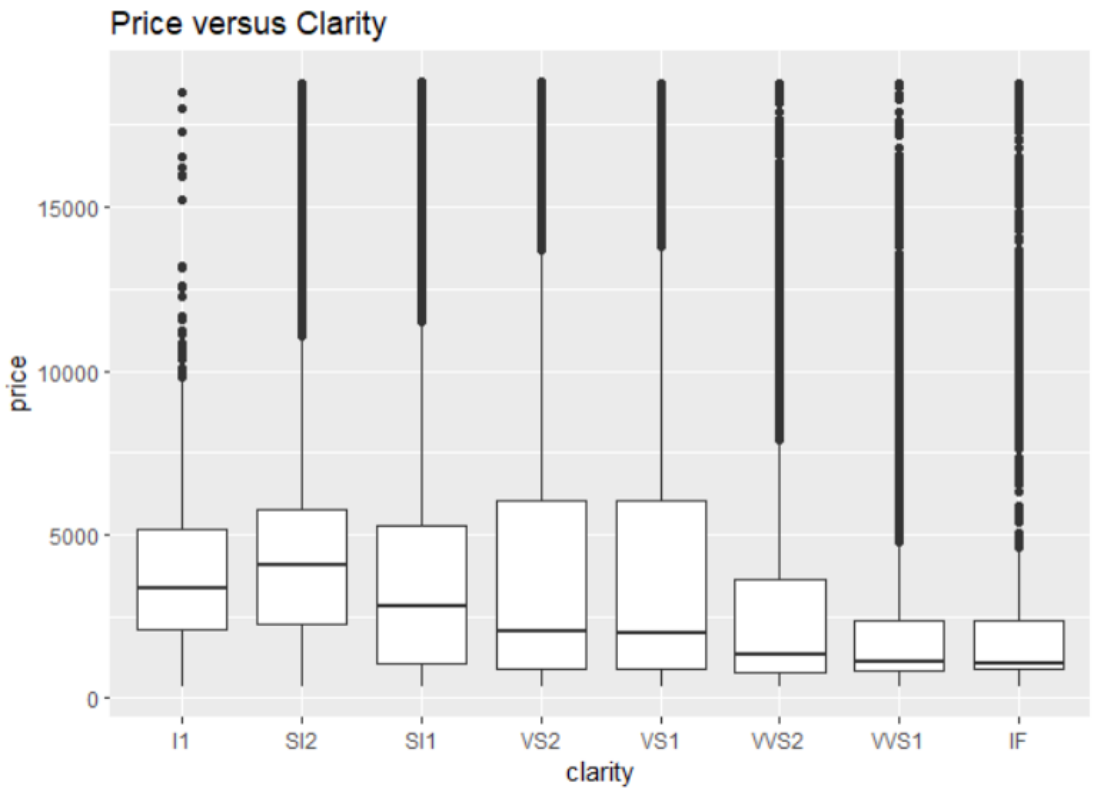



Distribution of Clarity
count
10000
5000
0
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
clarity
clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF


Distribution of Cut
count
20000
15000
10000
5000
0
Fair
Good
Very Good
Premium
Ideal
cut
cut
Fair
Good
Very Good
Premium
Ideal

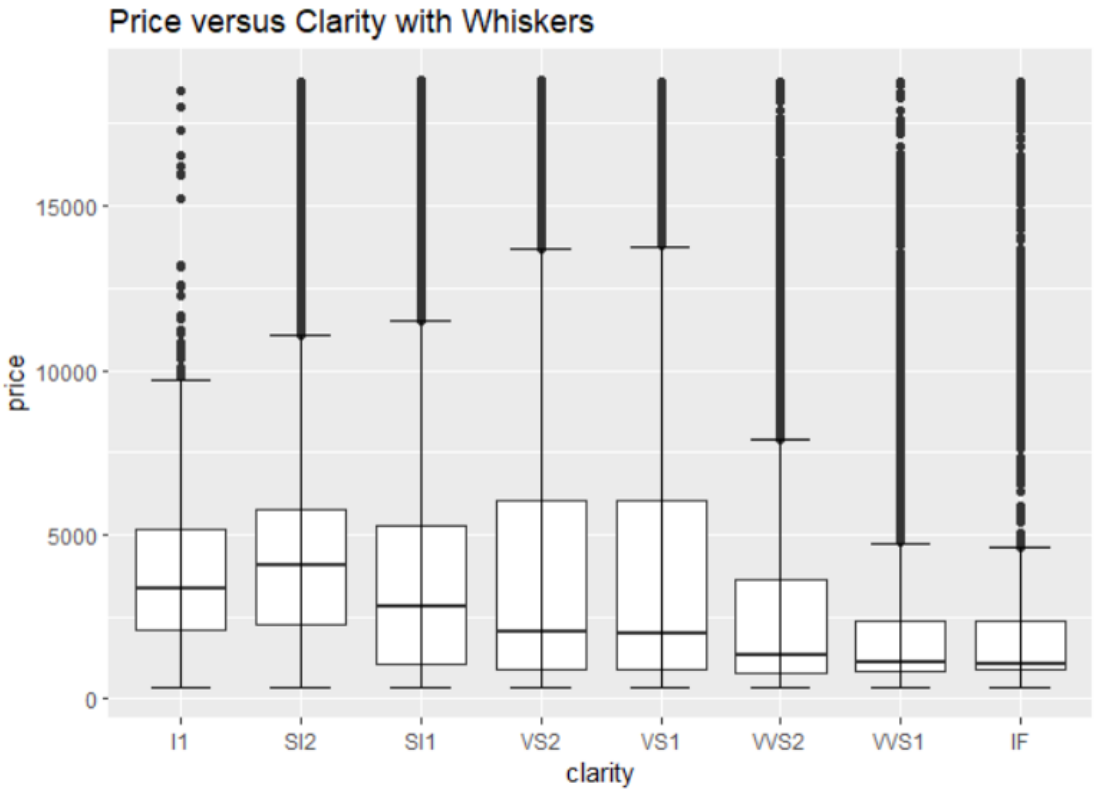
5. How does clarity, colour, and cut affect price?

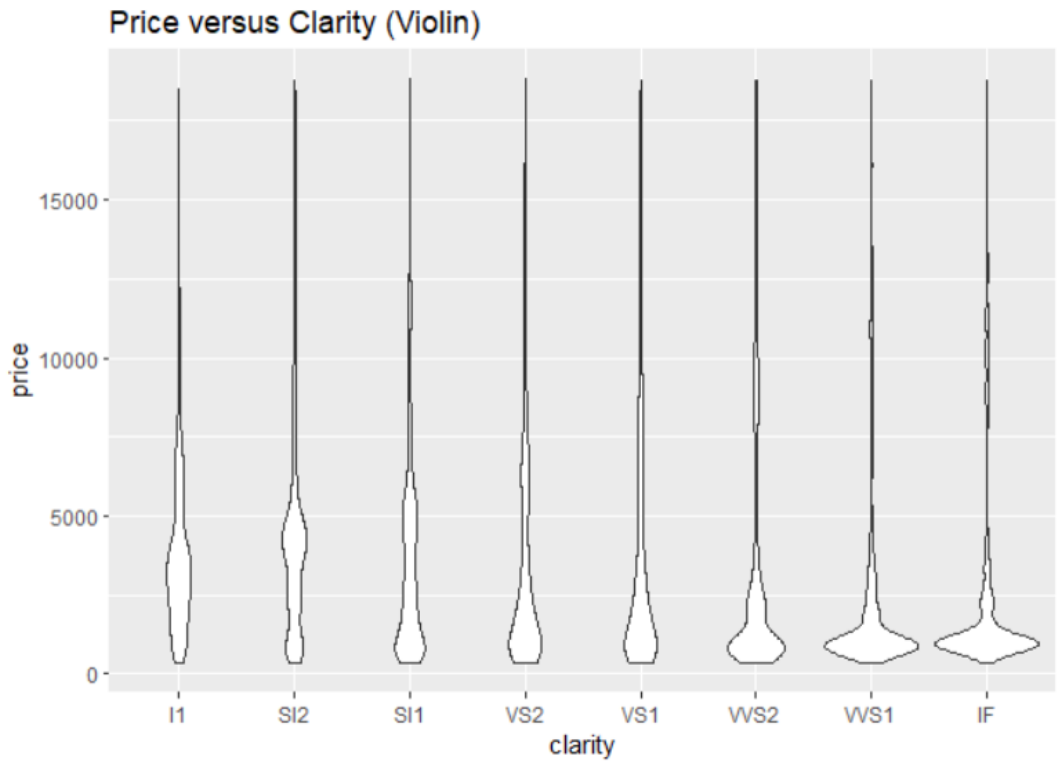
5.1 Boxplots of price versus clarity: First attempt



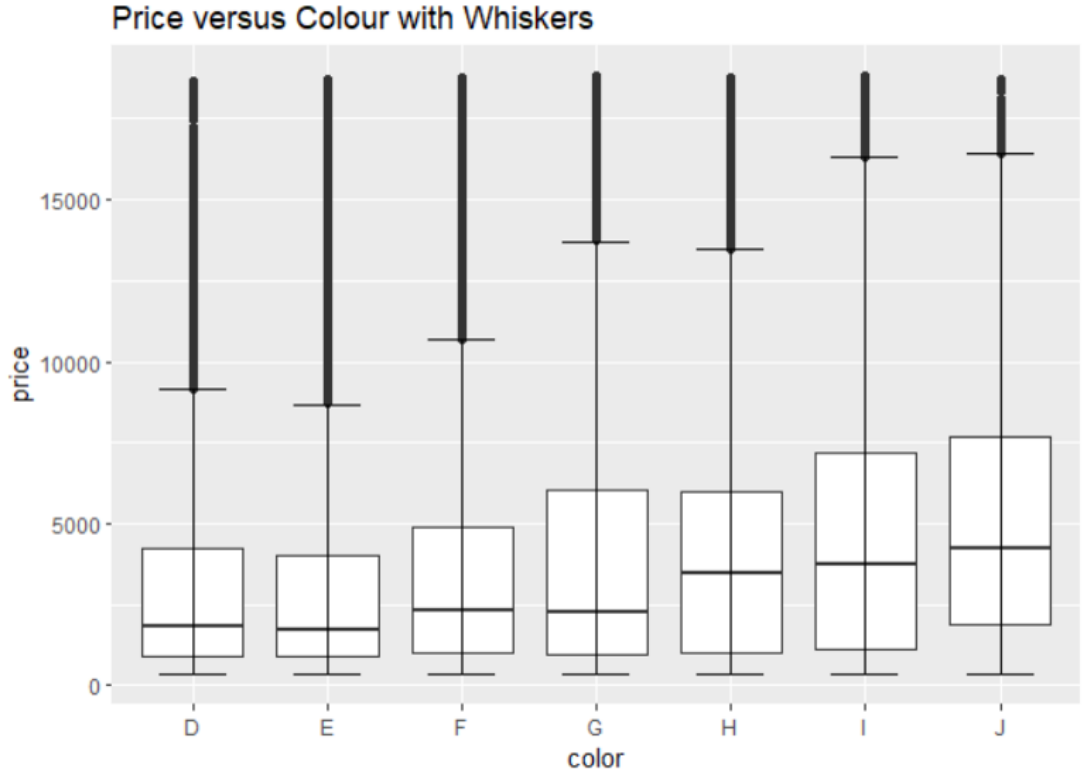
Best level of clarity is IF and worst is I1. Median prices in the boxplot shows that even though IF being the best level of clarity, price decreases as we move from level SI2 to IF.

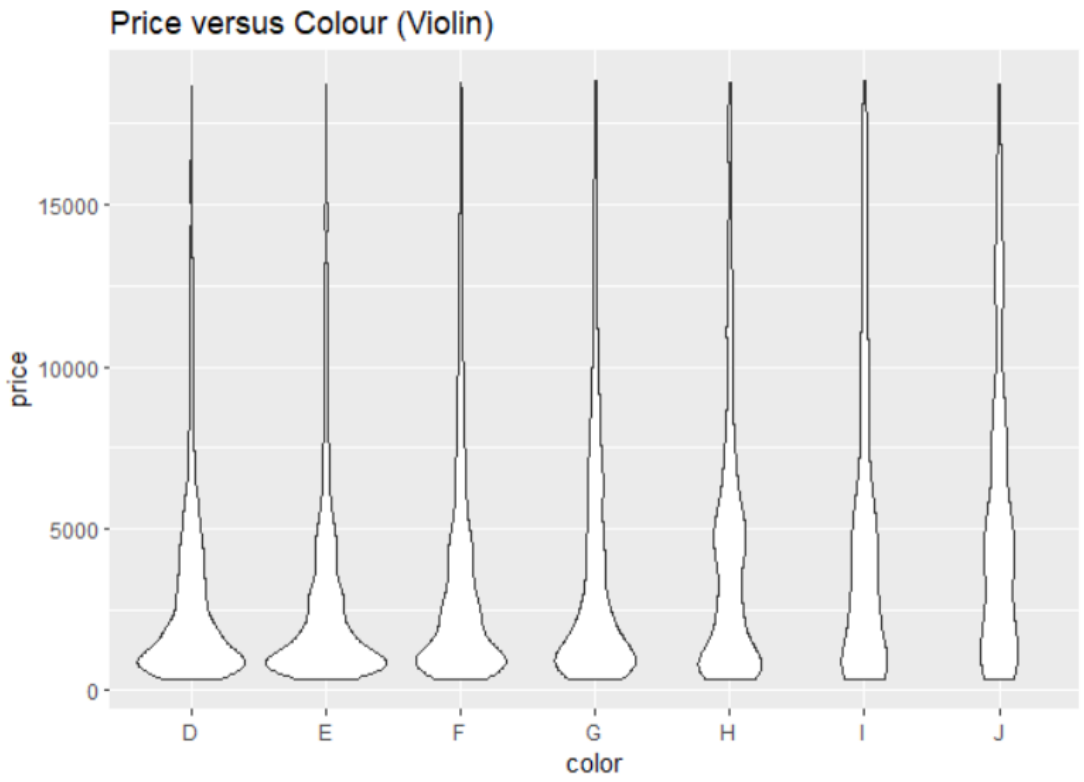
VS2 and VS1 have almost equal prices.



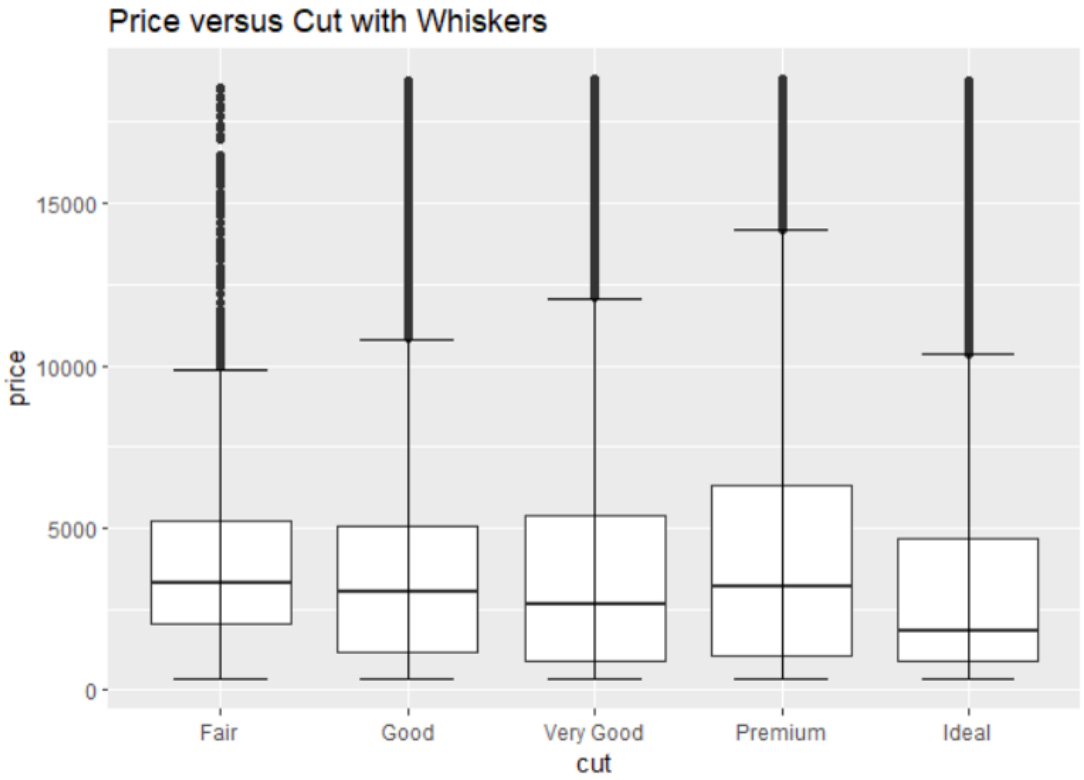


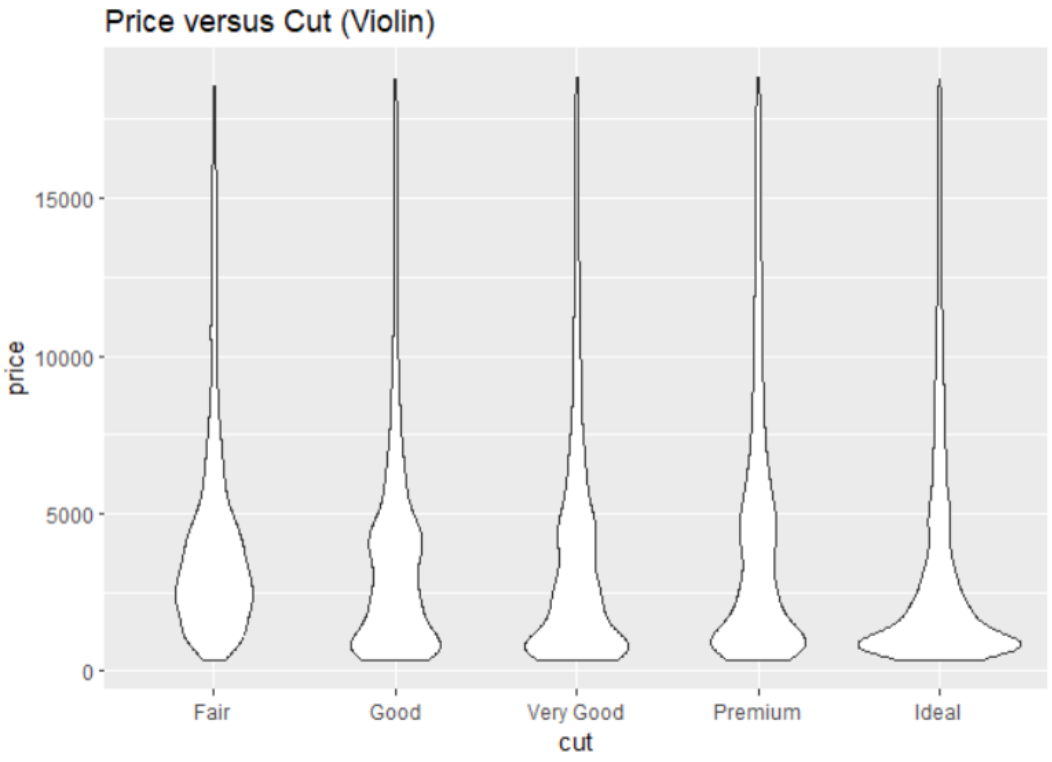
As clarity is increasing, price distribution is becoming more compact. And we see price is decreasing.





We can observe similar trend in colour as well. As quality of colour increases price seems to be decreasing. This was not the expected behaviour.

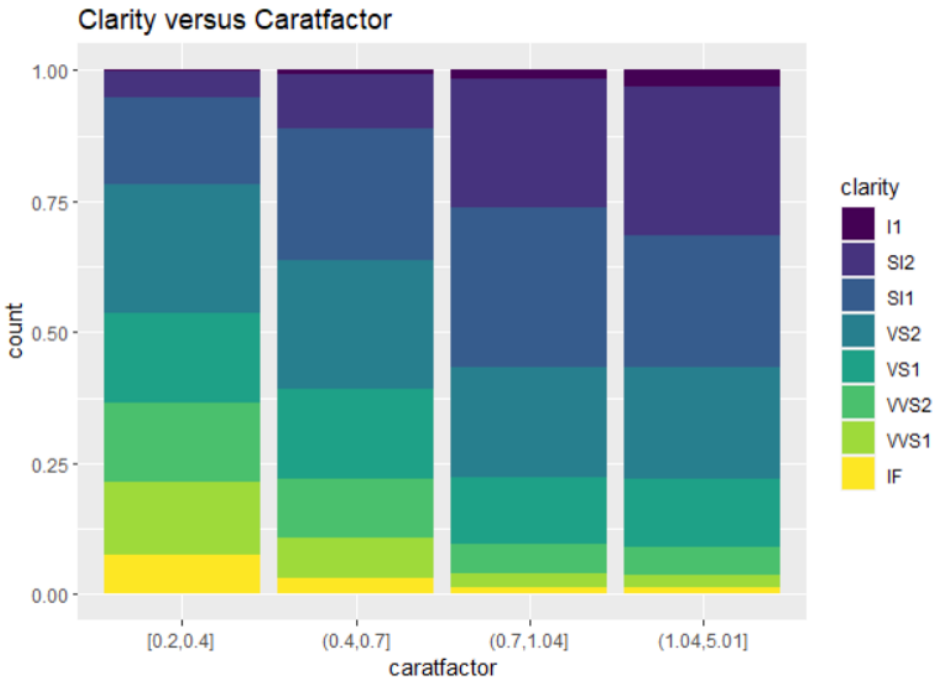




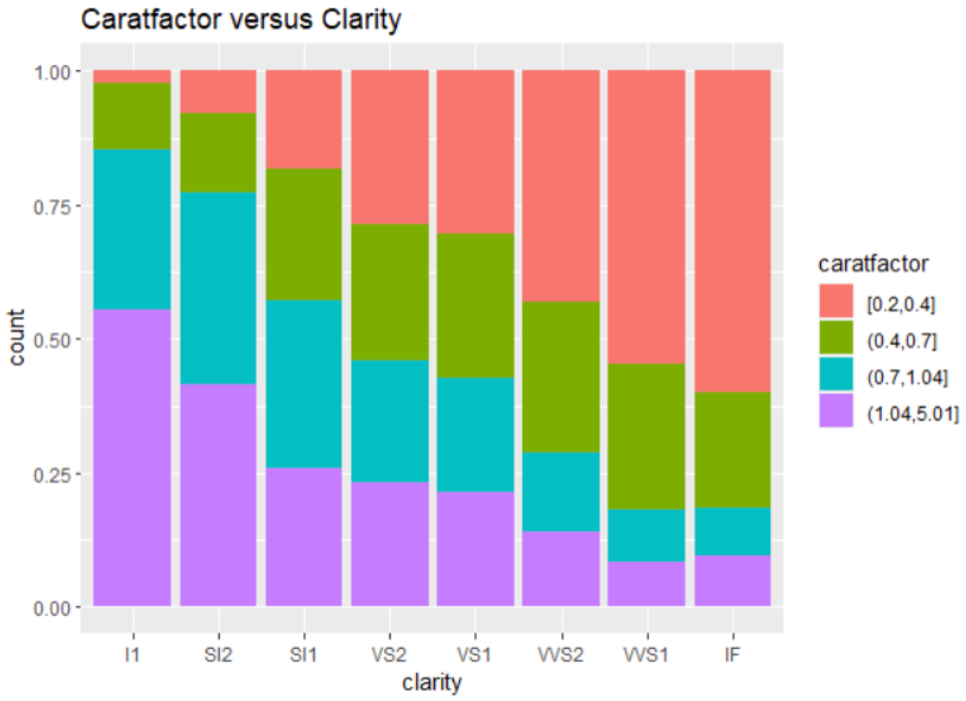
Trend in Price Vs Cut is little bit mixed. First, we see that price in increasing till premium level and then it is decreasing

The “best” levels of clarity and colour appear to have the lowest median prices might be because of the size and weight of the diamond. As we saw in previous plots, size and weight has significant effect on the price of diamonds.

5.2 Does the distribution of clarity, colour, and cut change with carat-weight?



The relative proportions of diamonds of different levels of clarity are not the same at different weights. There is a greater number of diamonds of best clarity in the lowest weight bar. I1, which is the worst clarity level has the highest proportion of diamonds in the largest weight bar. In short, the proportion of best clarity diamond decreases with weight and worst clarity diamonds increases with weight.

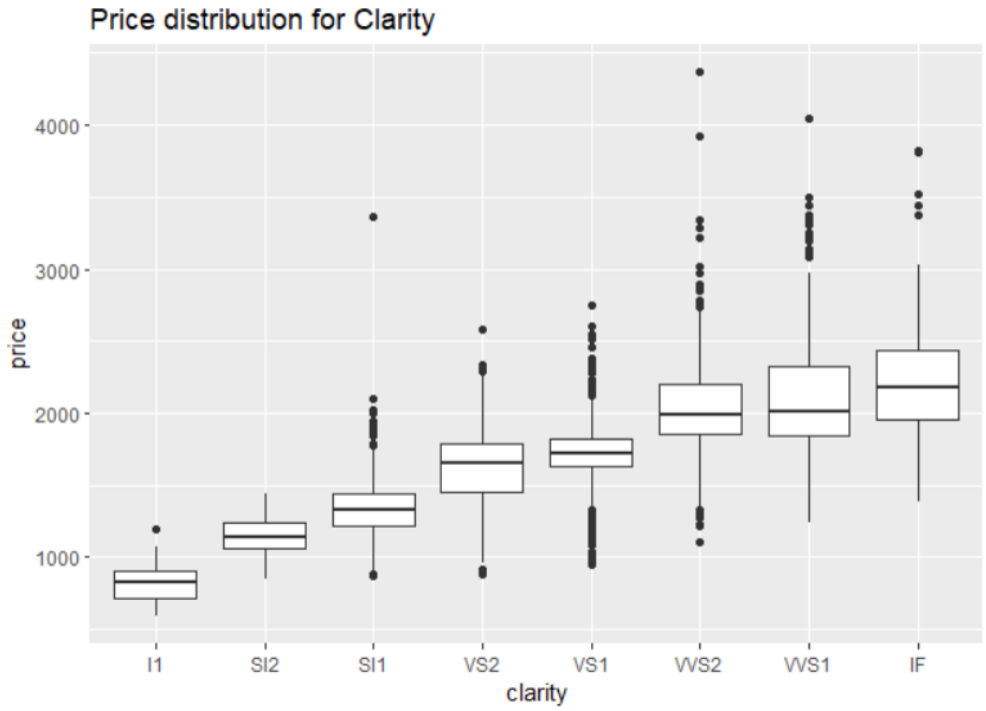


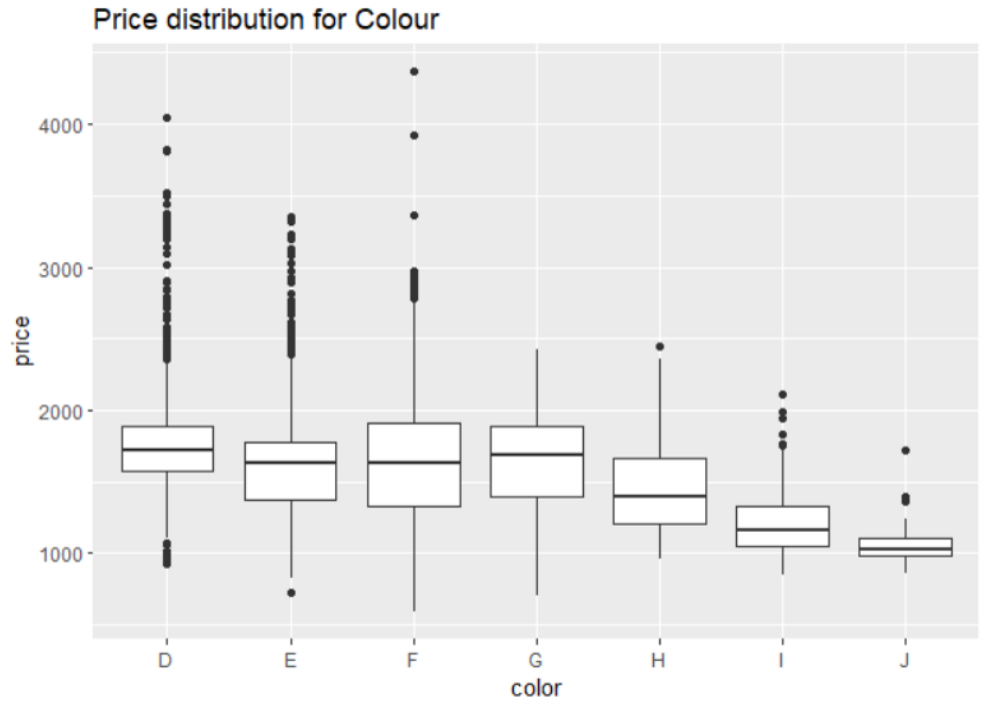
5.3 The effect of clarity, colour, and cut on price: Second attempt

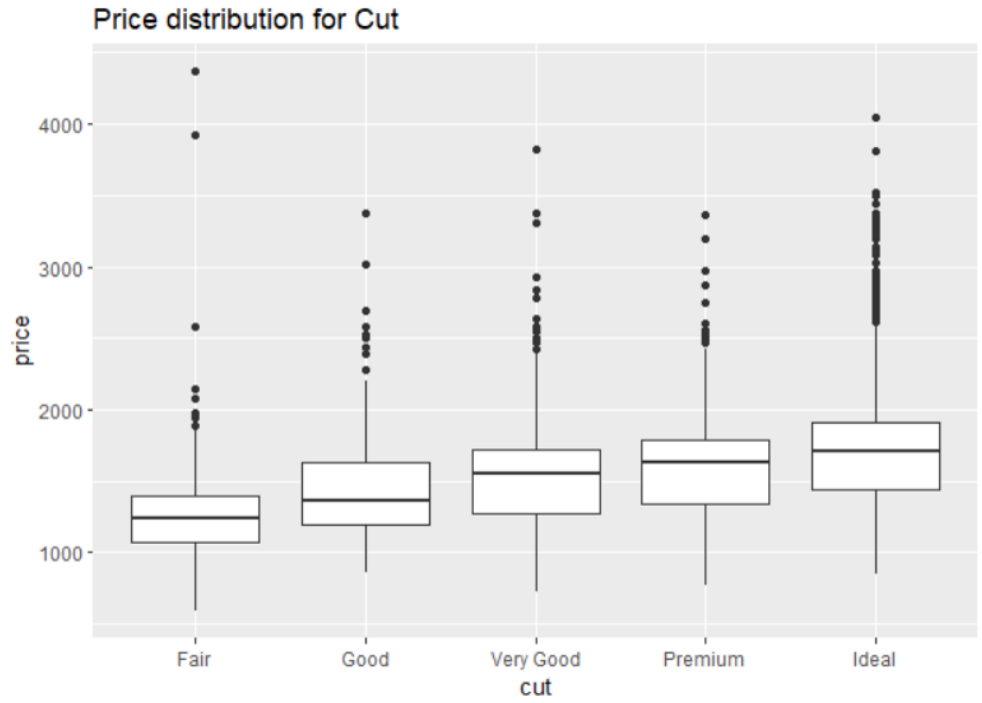
```
Console   Terminal ×   Background Jobs ×
R 4.2.2 · ~/
> str(halfcaratdiamonds)
tibble [4,536 × 11] (S3: tbl_df/tbl/data.frame)
 $ carat      : num [1:4536] 0.54 0.54 0.52 0.53 0.52 0.51 0.51 0.53 0.51 0.53 ...
 $ cut        : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 5 5 3 5 5 5 5 5 5 ...
 $ color      : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 3 1 3 3 1 3 3 2 ...
 $ clarity    : Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 6 6 7 6 8 7 7 8 7 7 ...
 $ depth      : num [1:4536] 61.6 61.5 61.3 57.5 62.2 62 61.7 61.9 62 61.9 ...
 $ table      : num [1:4536] 56 57 55 64 55 57 56 54 57 55 ...
 $ price      : int [1:4536] 2776 2776 2778 2782 2783 2787 2797 2802 2812 2821 ...
 $ x          : num [1:4536] 5.25 5.24 5.19 5.34 5.14 5.11 5.12 5.22 5.15 5.2 ...
 $ y          : num [1:4536] 5.27 5.26 5.22 5.37 5.18 5.15 5.16 5.25 5.11 5.21 ...
 $ z          : num [1:4536] 3.24 3.23 3.19 3.08 3.21 3.18 3.17 3.24 3.18 3.22 ...
 $ caratfactor: Factor w/ 4 levels "[0.2,0.4]","(0.4,0.7]",..: 2 2 2 2 2 2 2 2 2 2 ...
>
```

```
Console   Terminal ×   Background Jobs ×
R 4.2.2 · ~/
> str(onecaratdiamonds)
tibble [9,260 × 11] (S3: tbl_df/tbl/data.frame)
 $ carat      : num [1:9260] 1.17 1.01 1.01 1.01 1.05 1.05 1 1.01 1.04 1 ...
 $ cut        : Ord.factor w/ 5 levels "Fair"<"Good"<..: 3 4 1 4 3 1 4 1 4 4 ...
 $ color      : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 7 3 2 5 7 7 6 2 4 7 ...
 $ clarity    : Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 1 1 1 2 2 2 2 2 1 2 ...
 $ depth      : num [1:9260] 60.2 61.8 64.5 62.7 63.2 65.8 58.2 67.4 62.2 62.3 ...
 $ table      : num [1:9260] 61 60 58 59 56 59 60 60 58 58 ...
 $ price      : int [1:9260] 2774 2781 2788 2788 2789 2789 2795 2797 2801 2801 ...
 $ x          : num [1:9260] 6.83 6.39 6.29 6.31 6.49 6.41 6.61 6.19 6.46 6.45 ...
 $ y          : num [1:9260] 6.9 6.36 6.21 6.22 6.45 6.27 6.55 6.05 6.41 6.34 ...
 $ z          : num [1:9260] 4.13 3.94 4.03 3.93 4.09 4.18 3.83 4.13 4 3.98 ...
 $ caratfactor: Factor w/ 4 levels "[0.2,0.4]","(0.4,0.7]",..: 4 3 3 3 4 4 3 3 3 3 ...
>
```

There are 4,436 diamonds in ‘halfcaratdiamonds’ and 9,260 in ‘onecaratdiamonds’.







For these same-weight subsets we can now observe a trend that is as quality of Clarity, Colour and Cut increases, price also increases.

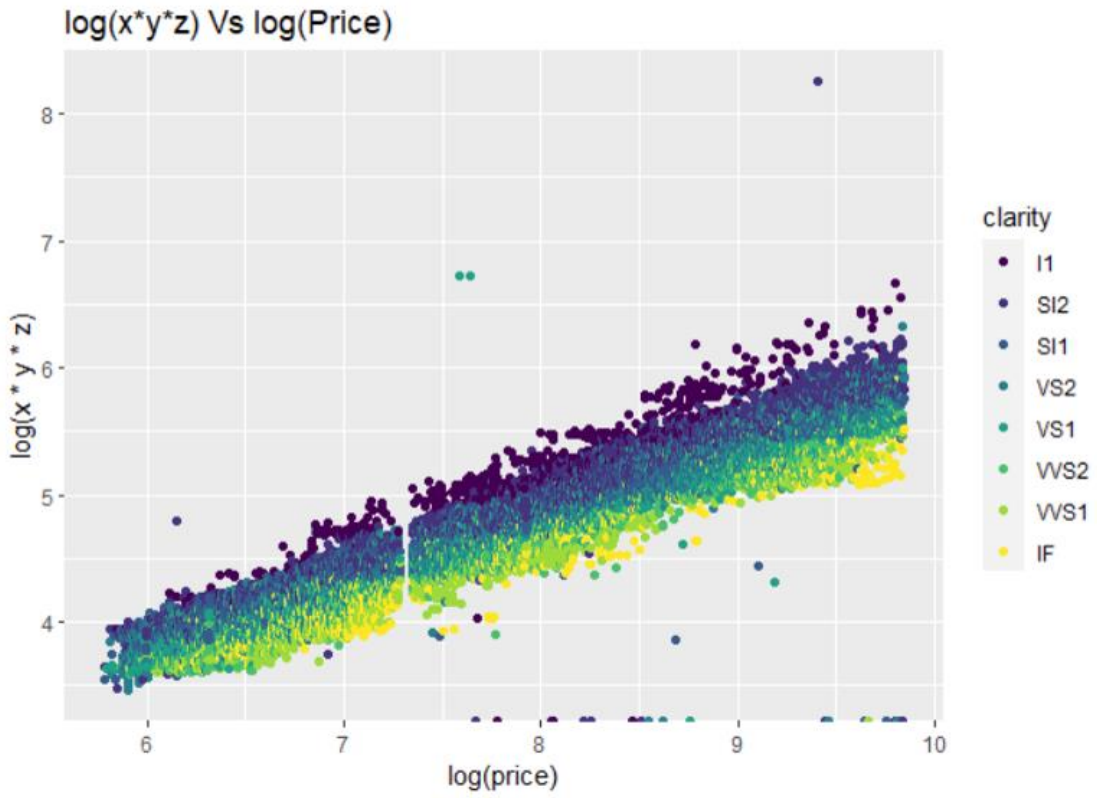
Clarity seems to have most powerful effect on price.

6. How is x, y, z related to price and clarity?

Log(x*y*z) versus price shows a curved relation. Inner side of the curve is dominated by IF level clarity that is the highest level of clarity and eventually decreases towards the outer edge of the curve.

If we plot log-log plot of approximate volume versus price we can observe linear relation.

```
p <- ggplot(diamonds,aes(x=log(price), y=log(x*y*z), color= clarity)) + geom_point()
p + ggtitle("log(x*y*z) Vs log(Price)")
```



The price gap is observed same as we saw previously in question 2.4.

Reference

1. https://www.bluenile.com/ca/education/diamonds?track=SideNav