1-analysis

Analysis of the diamonds data set

The code below shows the computation of the correlation between price and some of the variables used in the diamonds data set.

```
library(ggplot2)


# Variables for easier access
carat   = diamonds$carat
cut     = diamonds$cut
color   = diamonds$color
clarity = diamonds$clarity
depth   = diamonds$depth
price   = diamonds$price
volume  = diamonds$x * diamonds$y * diamonds$z
table   = diamonds$table

maxCorrelation = function()
{
  # Compute correlations between price and other numerical variables
  priceCaratCor   = cor(price, carat)
  priceDepthCor   = cor(price, depth)
  priceVolumeCor  = cor(price, volume)
  priceTableCor   = cor(price, table)

  # Put these values into a list
  correlationsList = c("Price-Carat with corr coefficient" = priceCaratCor,
                       "Price-Depth with corr coefficient " = priceDepthCor,
                       "Price-Volume with corr coefficient" = priceVolumeCor,
                       "Price-Table with corr coefficient" = priceTableCor)

  # Sort the list descending to get out the values with highest correlation first,
  # these will consequently drive the analysis
  sortedCorrelations = sort(correlationsList, decreasing = TRUE)

  return(sortedCorrelations)
}

main = function ()
{
  print("Correlations for: ")
  print( maxCorrelation())
}

main()
```

```
## [1] "Correlations for: "
##  Price-Carat with corr coefficient Price-Volume with corr coefficient
##                          0.9215913                          0.9023845
```

```
##  Price-Table with corr coefficient Price-Depth with corr coefficient
##                          0.1271339                         -0.0106474
```

Obviously, the price of a diamond mostly correlates with its carat as well as its volume. While depth and table almost have no correlation and therefore are not inspected in further detail during analysis.

The following plot shows the density distribution of the carat variable in the data set.

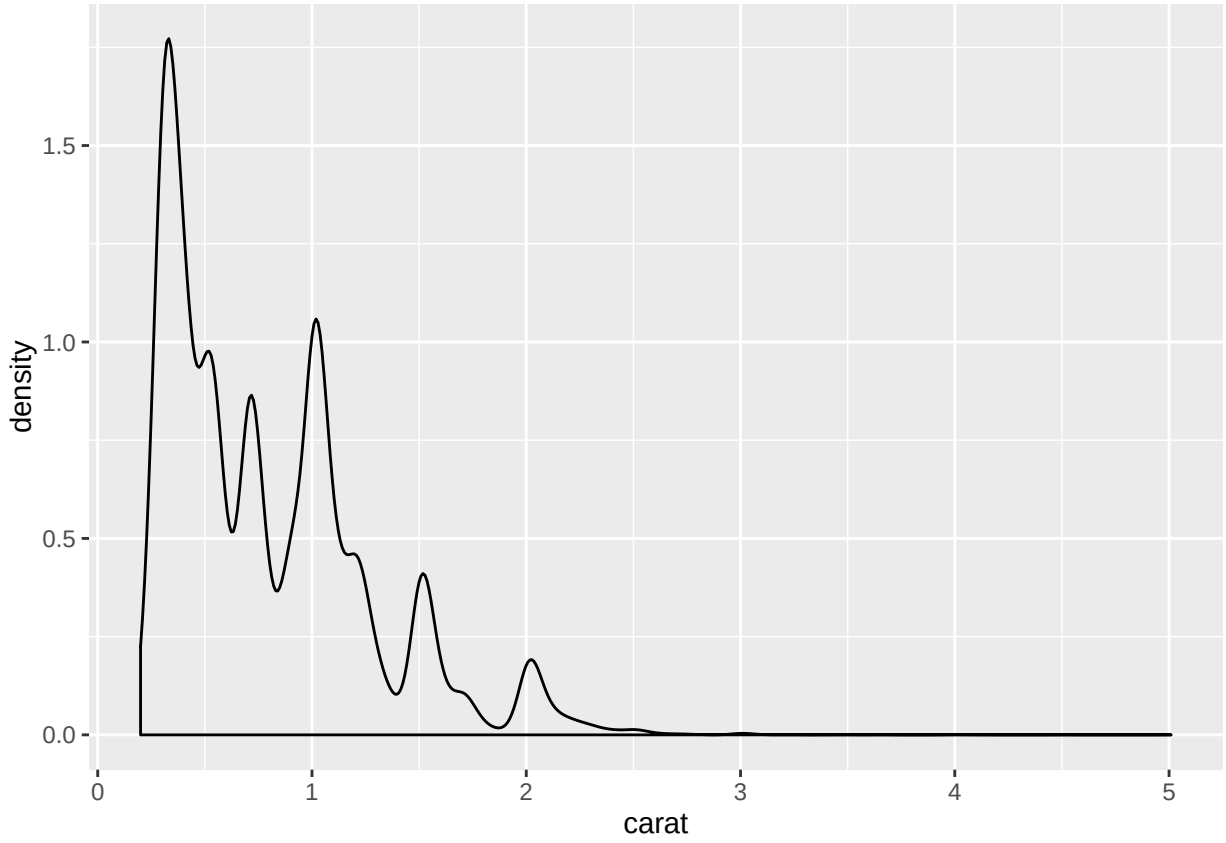

Figure 1: Desity plot for the distribution of carat values

Since the price and carat have a high correlation and the plot above reveals an atleast approximately exponentially decaying distribution, the first assumption is that the price might increase also exponentially with the carat. This assumption seems to be correct, which is expressed through the next plot below.

Figure 2 demonstrates that the relationship of price and carat is linear in logarithmic scale, thus it is exponential in normal scale. However, for high values of log(carat) namely values close to 1 clearly the variance starts to rise, which is due to the fact that not arbitrarily prices are paid for diamonds irrespective of their carat value. So, indeed, it can be further asssumed that the price-carat curve from Figure 2 will have rather an "S-shape" in normal scale as opposed to a strictly exponential curve as expected from Figure 2.

```
## Warning: Removed 20 rows containing non-finite values (stat_smooth).
```

Interestingly, Figure 3 shows that price has also a exponential relationship to volume, which is not that surprisng considering the high correlation between carat and volume shown below, but more that there are quiet a lot of outliers. The reason for that might be that volume and carat are not perfectly correlated, which was our initial assumption. The code below shows both mentioned facts.

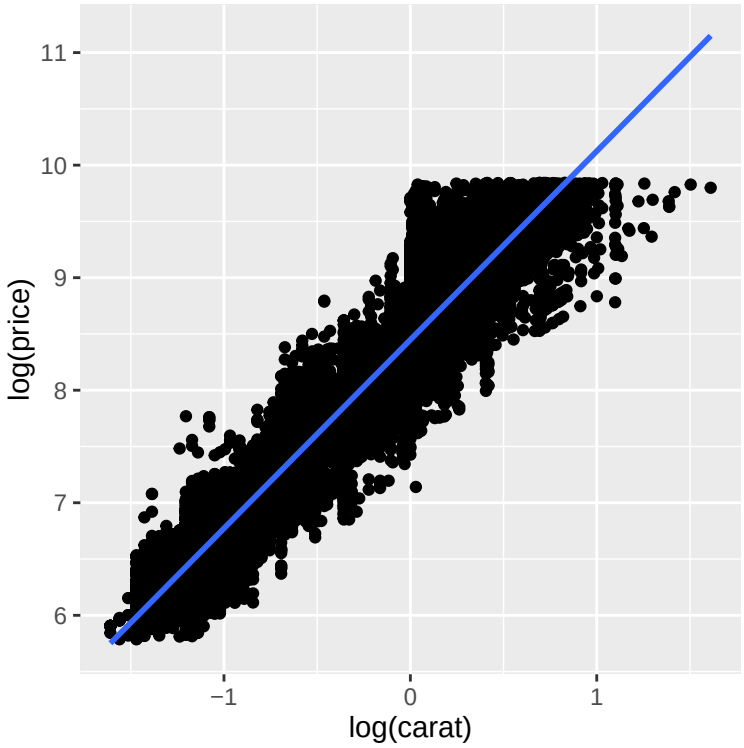

Figure 2: Scatterplot of price and carat values in logarithmic scale.

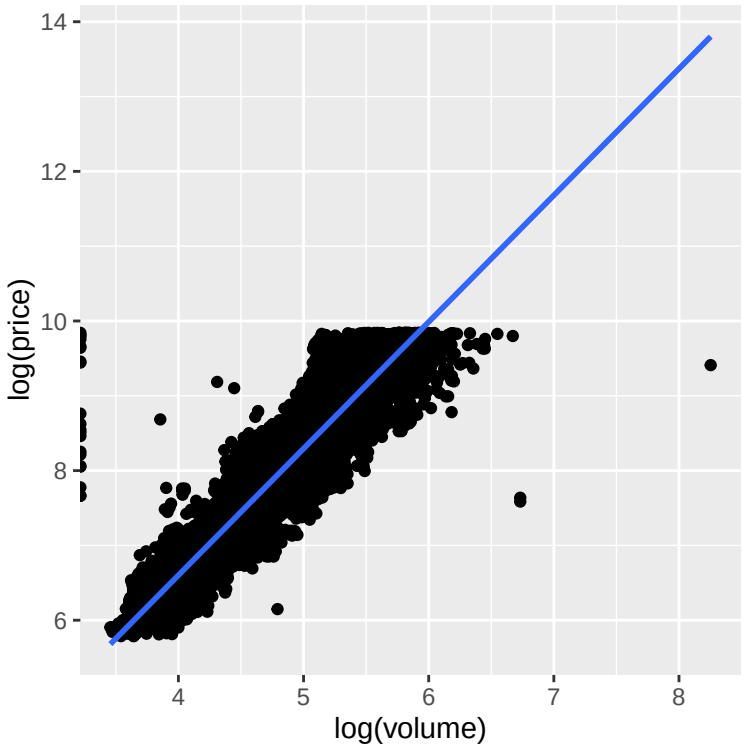

Figure 3: Scatterplot of price and volume values in logarithmic scale.

```
# Strong correlation between carat and volume obviously, but it is not perfect, i.e. not 1.
print(cor(carat,volume))
```

```
## [1] 0.9763084
```

The next variables to consider for the analysis are the clarity and cut variables. Intuitively, we would assume both to have a hugh impact on the price. Figure 4 below shows that this is partely true for the clarity variable.

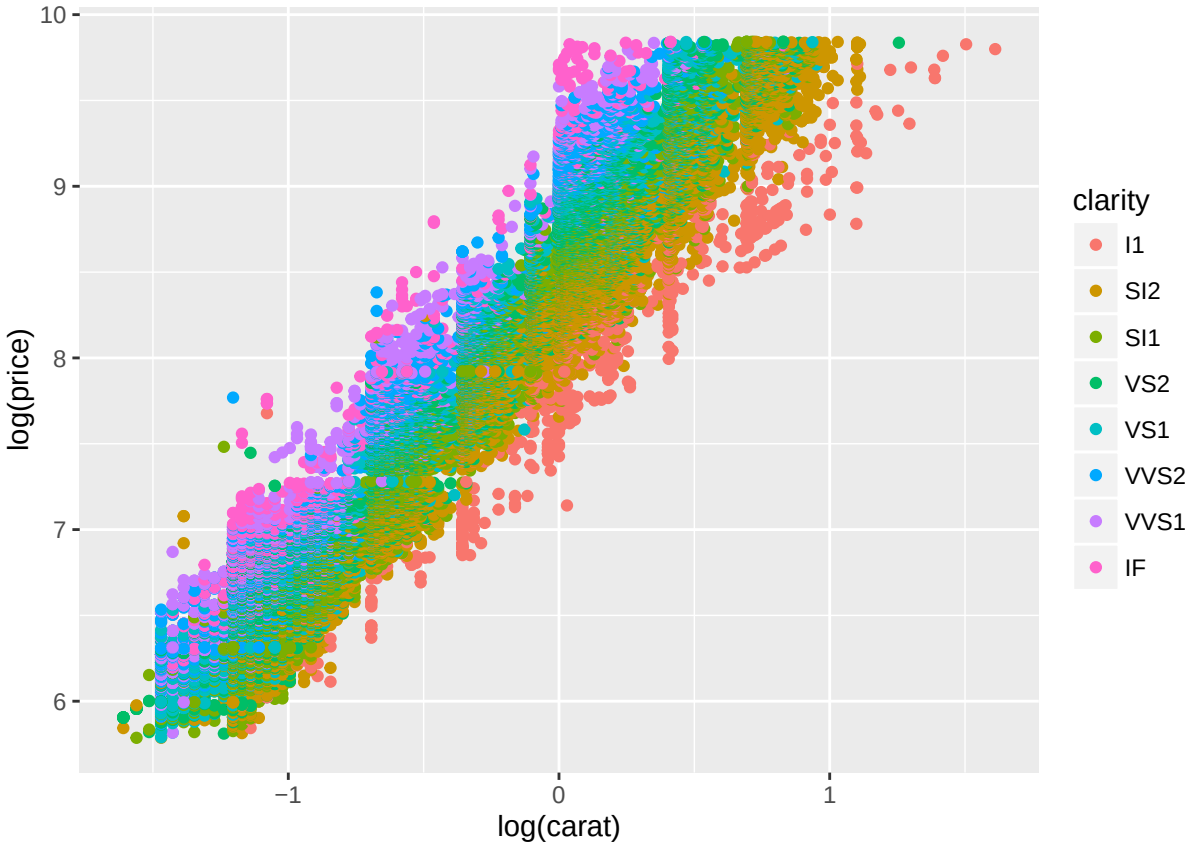

Figure 4: Scatterplot of price and carat values in logarithmic scale coloured differently for varying clarity values.

This plot really nicely illustrates that foreach of the clarity categories such as IF, VVS1, I1, ... one can imagine a line for each set of points with one of these categories. Apparently, these lines would have approximately the same slope while being shifted along the log(price) axis. Which means they can be transformed into each other by adding a scalar. Recall that it is logarithmic scale, so adding a scalar means actually multiplying by a scalar in normal space. So, the clarity makes up a difference of a factor.

Clearly looking at Figure 5, the cut does not have a great impact on the price, besides maybe for the cut "Fair", where the variance is quiet large but this maybe due to less data available with the respective annotation

The boxplot in Figure 6 gives an idea about the importance of colour, since the medians and interquartile ranges do not differ that much, the importance for the price seems to be rather small. However, for colour type D there are many outliers and with colour E,F,... the count of outliers gets smaller, so this might indicate that colour is in particular cases a very important criterion for the price.

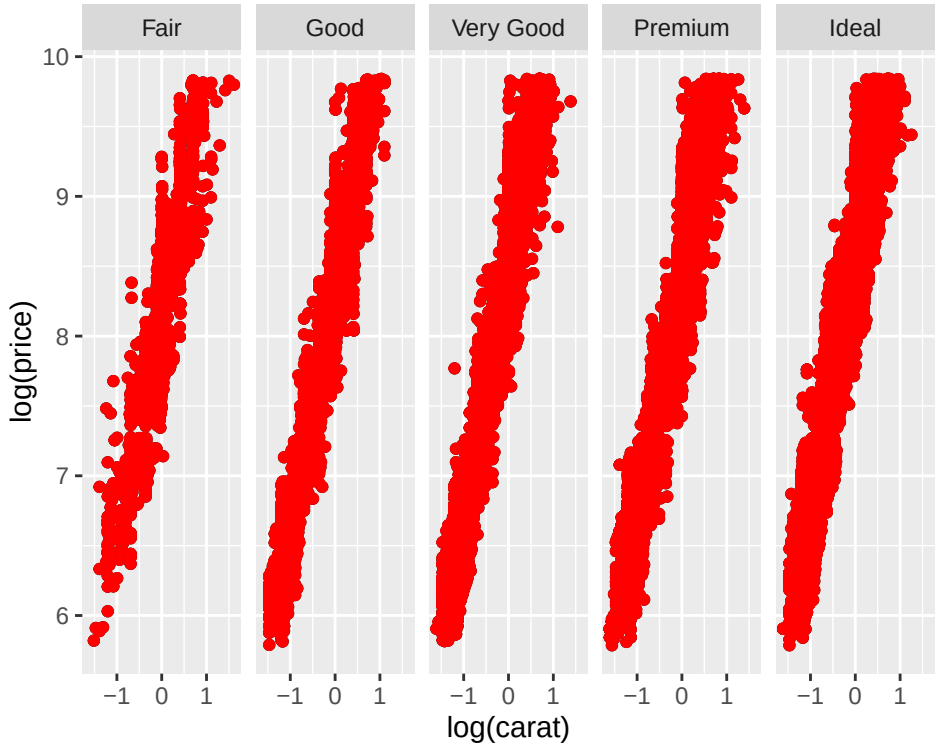

Figure 5: Scatterplot of price and carat values in logarithmic scale for varying cut values.

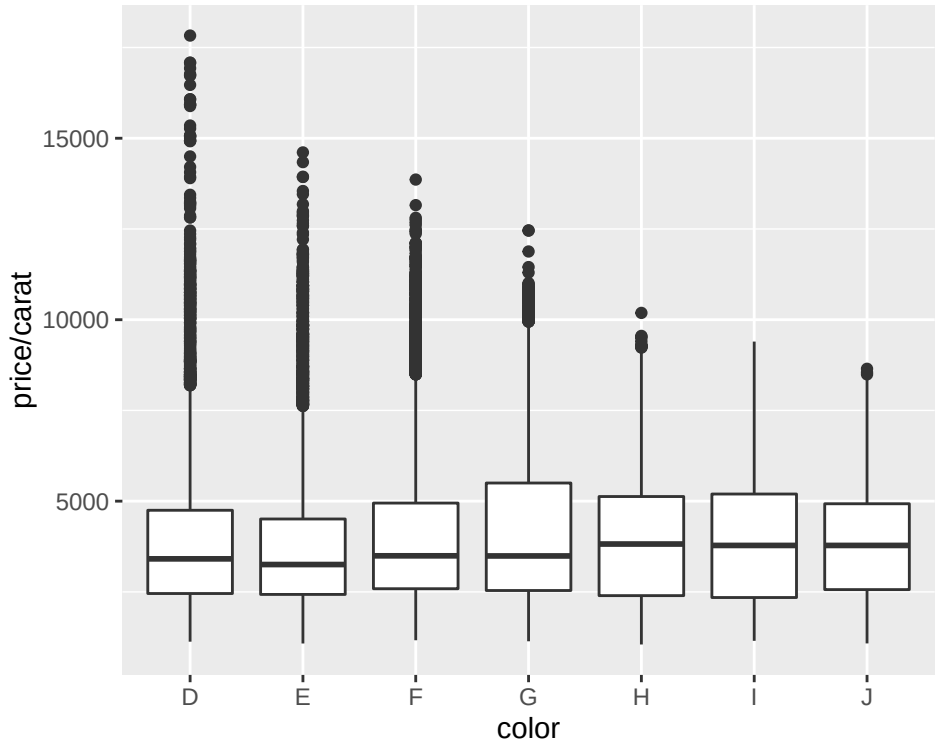

Figure 6: Boxplot of price per carat and color values.