Assignement

Shubham Khandelwal

November 4, 2019

0 - Loading the dataset

```
DiamondData <- read.csv("~/DiamondData_51309e6fb24ac297b10b919e60060c90.csv")
```

Importing the tidyverse package

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------ tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts --------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

1 - Cleanup of the dataset

We'll start by having a glimpse on the dataset and the type of its variables

```
## Observations: 50,000
## Variables: 10
## $ carat   <dbl> 1.50, 0.70, 1.22, 0.51, 2.02, 0.70, 0.46, 0.55, 0.51, ...
## $ cut     <fct> Fair, Ideal, Premium, Premium, Very Good, Very Good, I...
## $ color   <fct> G, E, G, E, J, E, G, F, G, H, J, E, D, H, E, H, G, J, ...
## $ clarity <fct> SI1, VS2, VS1, VS2, SI2, SI1, VVS1, SI2, VVS2, VS2, VS...
## $ depth   <dbl> 64.5, 61.4, 61.3, 62.5, 59.2, 63.2, 60.7, 60.6, 61.6, ...
## $ table   <dbl> 57, 57, 58, 60, 60, 61, 57, 58, 56, 55, 58, 56, 61, 56...
## $ price   <int> 10352, 2274, 8779, 1590, 11757, 2164, 1453, 1175, 1750...
## $ x       <dbl> 7.15, 5.72, 6.91, 5.08, 8.27, 5.61, 4.98, 5.28, 5.12, ...
## $ y       <dbl> 7.09, 5.78, 6.89, 5.10, 8.39, 5.49, 5.03, 5.31, 5.14, ...
## $ z       <dbl> 4.59, 3.53, 4.23, 3.18, 4.91, 3.51, 3.04, 3.21, 3.16, ...
```

The results show that our dataset contains 50,000 observations, described by 10 variables, with types ranging from factors to doubles to integers.

A quick exploration of the dataset shows that the `cut` variable got 6 different categories observations instead of 5 as follows.

```
## # A tibble: 6 x 2
##   cut       count
##   <fct>     <int>
## 1 Fair       1480
## 2 Good       4559
## 3 Ideal     19918
## 4 Premium   12826
## 5 Very Geod  2242
## 6 Very Good  8975
```

We will replace the `Very Geod` category with the `Very Good` category with the following code :

```
DiamondData$cut <- str_replace(as.character(DiamondData$cut), "Very Geod", "Very Good")
DiamondData %>%
  group_by(cut) %>%
  summarize(count = n())
```

```
## # A tibble: 5 x 2
##   cut       count
##   <chr>     <int>
## 1 Fair       1480
## 2 Good       4559
## 3 Ideal     19918
## 4 Premium   12826
## 5 Very Good 11217
```

A test for the existence of NA and missing values shows a number of them in the `depth`, `table`, `price`, `x`, `y` and `z` variables as follows :

```
## [1] "Missing values for Depth :"
```

```
## [1] 471
```

```
## [1] "Missing values for table :"
```

```
## [1] 390
```

```
## [1] "Missing values for Price :"
```

```
## [1] 253
```

```
## [1] "Missing values for x :"
```

```
## [1] 221
```

```
## [1] "Missing values for y :"
```

```
## [1] 333
```

```
## [1] "Missing values for z :"
```

```
## [1] 428
```

We observe that we only got missing values for quantitative variables. We inpute the missing values with the median of each variable as follows :

```
DiamondData <- DiamondData %>%
  mutate(z= replace(z,is.na(z),median(z, na.rm = TRUE)))

DiamondData <- DiamondData %>%
  mutate(y= replace(y,is.na(y),median(y, na.rm = TRUE)))

DiamondData <- DiamondData %>%
  mutate(x = replace(x, is.na(x), median(x, na.rm = TRUE)))

DiamondData <- DiamondData %>%
  mutate(price= replace(price,is.na(price),median(price, na.rm = TRUE)))

DiamondData <- DiamondData %>%
  mutate(table= replace(table,is.na(table),median(table, na.rm = TRUE)))

DiamondData <- DiamondData %>%
  mutate(depth= replace(depth,is.na(depth),median(depth, na.rm = TRUE)))
```

A following task would be to set the decimal precision for the `carat` variable to two decimals :

```
DiamondData$carat <- round(DiamondData$carat, digits=2)
```

A look through our and corrected dataset shows us that the cleanup process is complete :

```
glimpse(DiamondData)
```

```
## Observations: 50,000
## Variables: 10
## $ carat   <dbl> 1.50, 0.70, 1.22, 0.51, 2.02, 0.70, 0.46, 0.55, 0.51, ...
## $ cut     <chr> "Fair", "Ideal", "Premium", "Premium", "Very Good", "V...
## $ color   <fct> G, E, G, E, J, E, G, F, G, H, J, E, D, H, E, H, G, J, ...
## $ clarity <fct> SI1, VS2, VS1, VS2, SI2, SI1, VVS1, SI2, VVS2, VS2, VS...
## $ depth   <dbl> 64.5, 61.4, 61.3, 62.5, 59.2, 63.2, 60.7, 60.6, 61.6, ...
## $ table   <dbl> 57, 57, 58, 60, 60, 61, 57, 58, 56, 55, 58, 56, 61, 56...
## $ price   <int> 10352, 2274, 8779, 1590, 11757, 2164, 1453, 1175, 1750...
## $ x       <dbl> 7.15, 5.72, 6.91, 5.08, 8.27, 5.61, 4.98, 5.28, 5.12, ...
## $ y       <dbl> 7.09, 5.78, 6.89, 5.10, 8.39, 5.49, 5.03, 5.31, 5.14, ...
## $ z       <dbl> 4.59, 3.53, 4.23, 3.18, 4.91, 3.51, 3.04, 3.21, 3.16, ...
```

2 - Initial analysis :

Variable summary & Data visualisation :

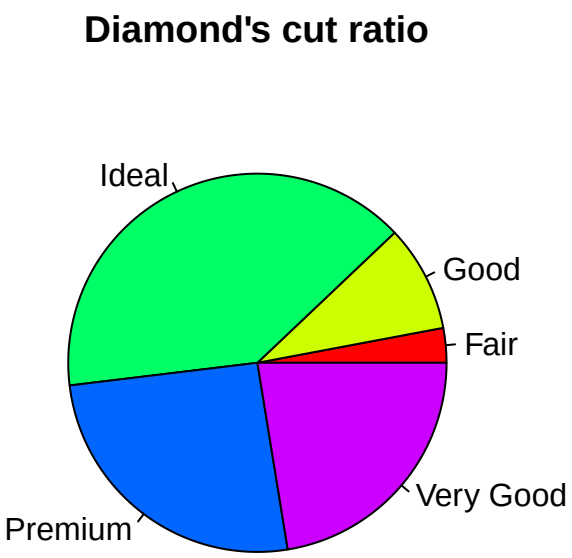
- A basic summary of the `carat` variable :

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.200   0.400   0.700   0.907   1.050  49.990
```

The `carat` values seem mostly centered around the `0 - 1` values in most of the observations, with a quarter of them going beyond `1` toward much higher values, which shows through the `maximum value` and the `Mean` that is so close to the `3rd Quarter`, and that the majority of the diamonds are on the low end in terms of `carat`.

- A basic summary of the `cut` variable :

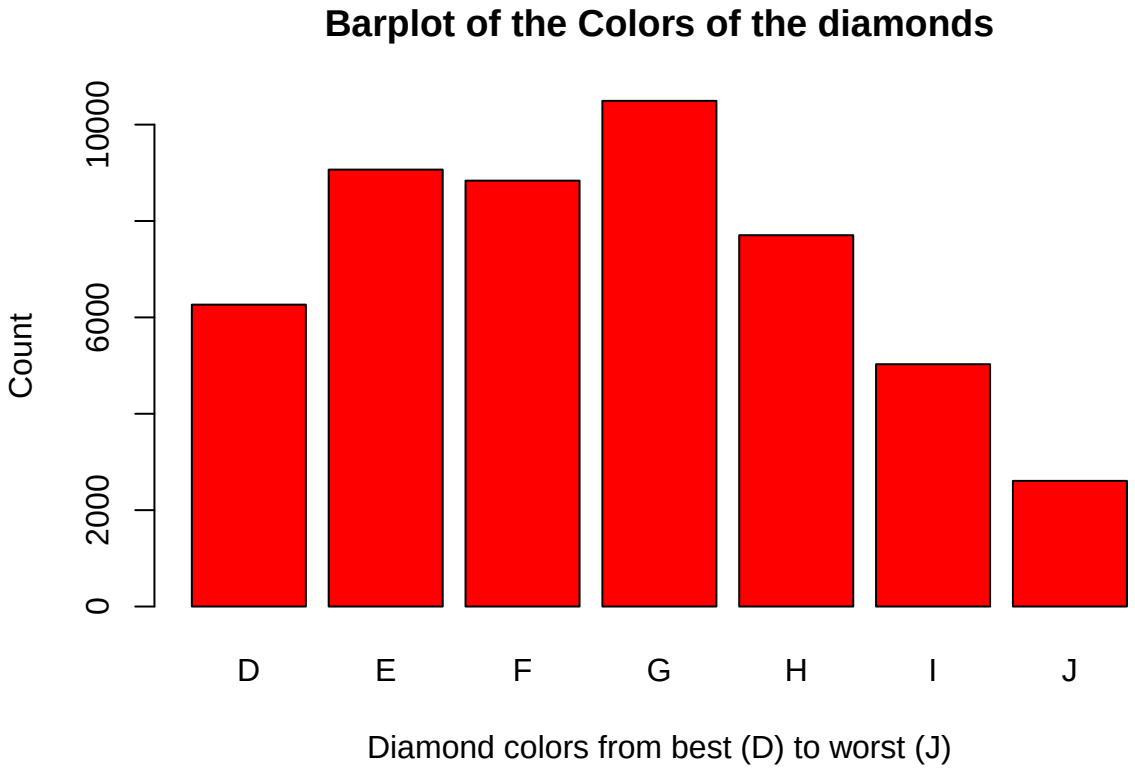
```
## # A tibble: 5 x 2
##   cut       count
##   <chr>     <int>
## 1 Fair       1480
## 2 Good       4559
## 3 Ideal     19918
## 4 Premium   12826
## 5 Very Good 11217
```



Most of the values of the `cut` variable are centered in the `Ideal` value, followed by the `Premium` value and `Very Good` respectively. This obviously shows that the dominating majority of the diamonds pieces in our dataset are of very superior cut.

- A basic summary of the `color` variable :

```
##     D     E     F     G     H     I     J
##  6264  9066  8837 10493  7705  5028  2607
```



The majority of the diamonds in the dataset appear to be of medium colors, with values ranging `E` to `H`, followed by a high count of diamonds with the best color `D`, and a very low count of diamonds with the worst color `J`. This shows that the dataset contains mostly diamonds with medium to high quality colors.

- A basic summary of the `clarity` variable :

```
##    I1    IF   SI1   SI2   VS1   VS2  VVS1  VVS2
##   690  1666 12120  8486  7563 11406  3377  4692
```

The `clarity` variable shows a distribution mostly focused on the medium quality, with the values `SI1` and `VS2` respectively, with the count decreasing as we increase or decrease clarity The majority of the diamonds in the dataset are of medium clarity.

- A basic summary of the `depth` variable :

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.00   61.10   61.80   61.75   62.50   79.00
```

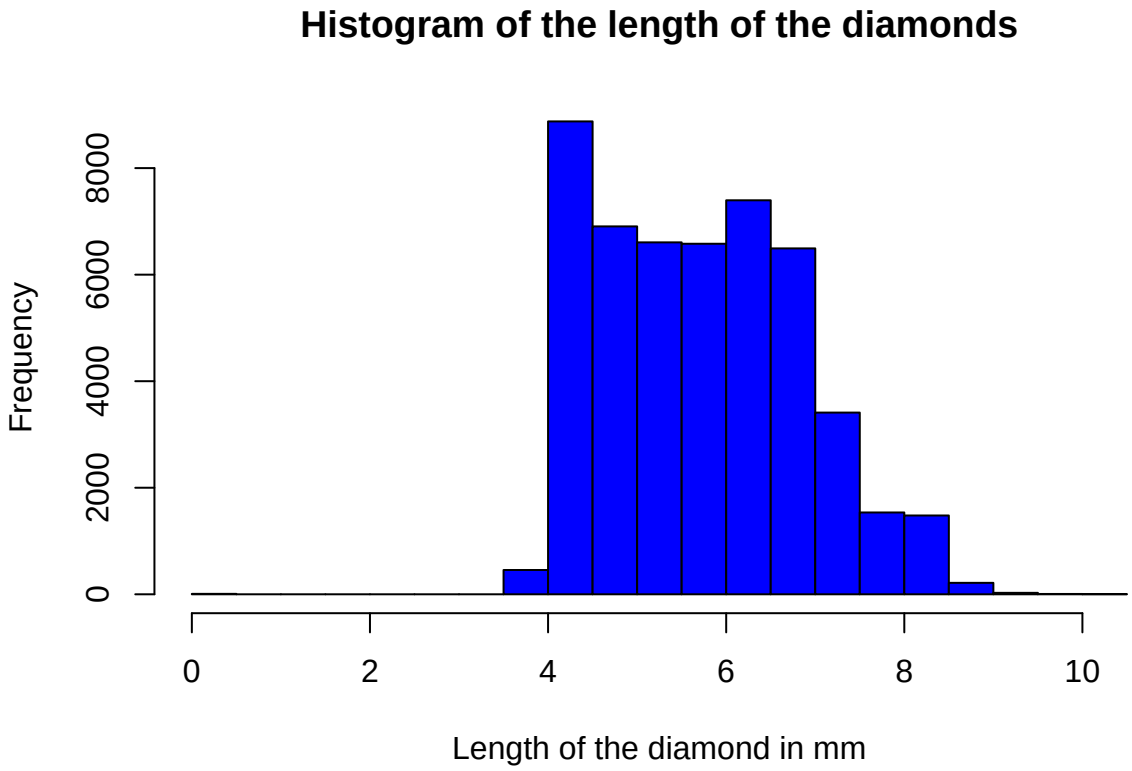
- A basic summary of the `table` variable :

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.00   56.00   57.00   57.46   59.00   95.00
```

The `Depth` and `table` variables show a usual normal distribution.

- A basic summary of the `x` variable :

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.710   5.700   5.732   6.540  10.230
```



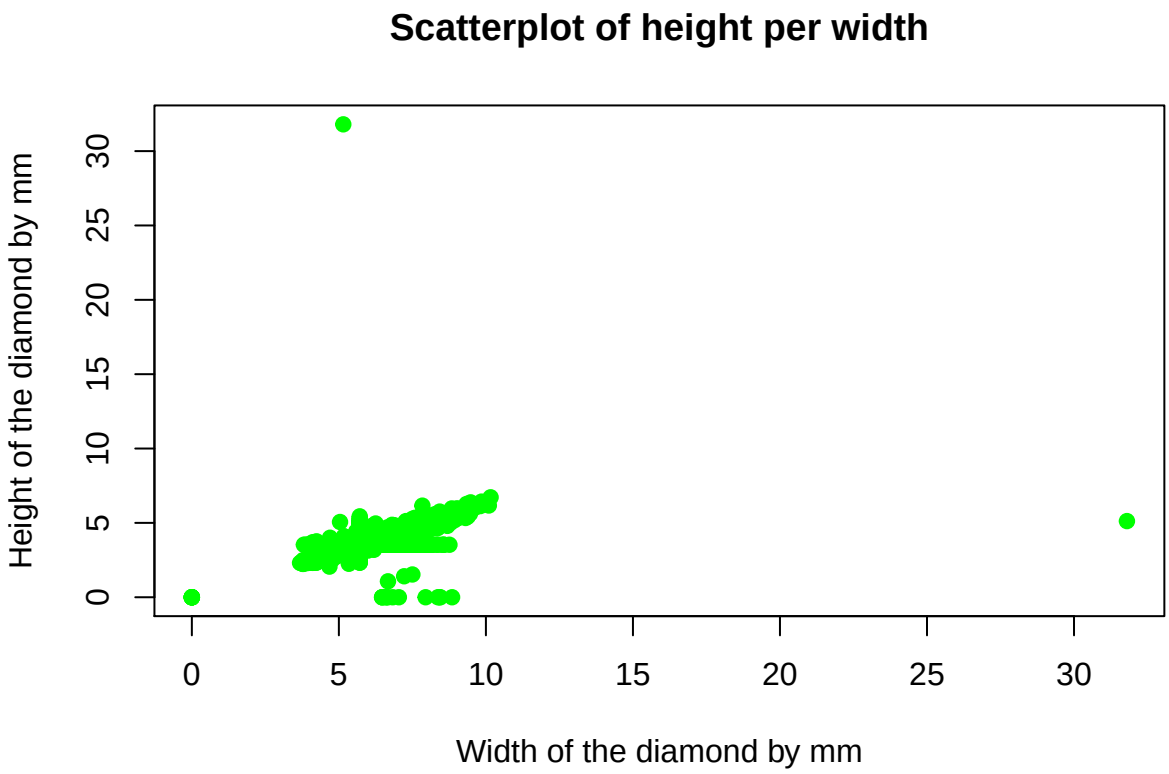
The count of small length diamonds appear to be higher and slowly decreases as the length increase.

- A basic summary of the `y` variable :

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.720   5.710   5.734   6.530  31.800
```

- A basic summary of the `z` variable :

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.910   3.530   3.539   4.030  31.800
```

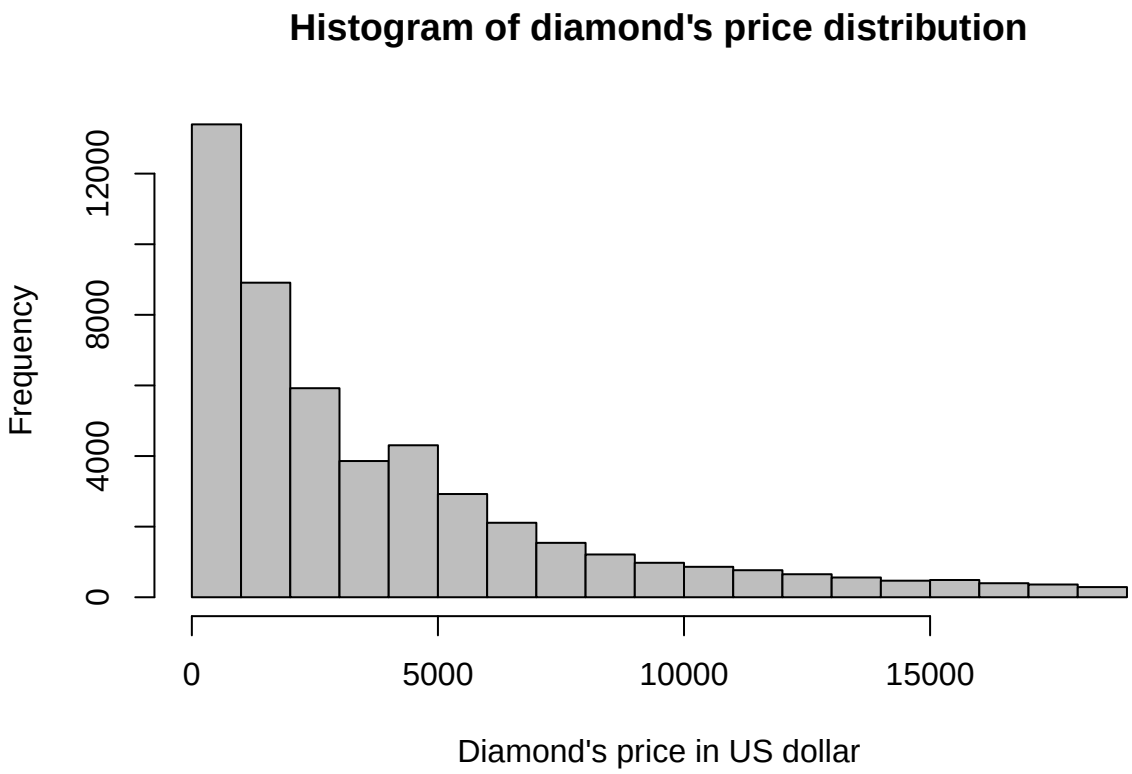


The variables `y` and `z` appear to be linearly correlated, as they are about the size and volume of the diamond, with some outliers.

3 - Price Variable Analysis :

a - Price variable description :

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326     954    2401    3931    5315   18823
```



We notice that most of the diamond prices are on the low end, justifying the low observed `Mean` and `3rd Quarter` compared to the maximum value, with very few having high prices.

b - Price groupings :

If we divide the price into three groups, the `low` group would go from `0` to `900` being almost the first Quarter. The `medium` group would be from `900` to `4000` being close to the mean. The `high` group would be from `4000` to `20000`

- A summary of diamonds with low prices :

```
## [1] "Summary of carat "

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2000  0.3000  0.3200  0.4329  0.3500 49.7300

## [1] "Summary of color "

##    D    E    F    G    H    I    J
## 1534 2302 1879 2444 1794 1047  377

## [1] "Summary of clarity "

##   I1   IF  SI1  SI2  VS1  VS2 VVS1 VVS2
##   44  435 2545  901 1966 2832 1138 1516
```

```
## [1] "Summary of depth "

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   54.20   61.20   61.80   61.76   62.40   71.00

## [1] "Summary of table "

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44.00   55.00   57.00   56.99   58.00   67.00

## [1] "Summary of x "

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.730   4.300   4.390   4.424   4.550   6.650

## [1] "Summary of y "

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.680   4.310   4.400   4.442   4.560   5.710

## [1] "Summary of z "

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.240   2.670   2.720   2.741   2.810   3.530
```

- A summary of diamonds with medium prices :

```
## [1] "Summary of carat "

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2400  0.5000  0.5700  0.7248  0.7200 49.9900

## [1] "Summary of color "

##    D    E    F    G    H    I    J
## 3052 4456 4108 4207 2381 1594  852

## [1] "Summary of clarity "

##   I1   IF  SI1  SI2  VS1  VS2 VVS1 VVS2
##  417  951 4591 3247 3131 4505 1731 2077

## [1] "Summary of depth "

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.00   61.00   61.80   61.73   62.50   79.00

## [1] "Summary of table "
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   49.00   56.00   57.00   57.38   59.00   79.00
```

```
## [1] "Summary of x "
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   5.050   5.350   5.408   5.790   8.540
```

```
## [1] "Summary of y "
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.05    5.36    5.41    5.80   31.80
```

```
## [1] "Summary of z "
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    3.11    3.30    3.34    3.58   31.80
```

- A summary of diamonds with high prices :

```
## [1] "Summary of carat "
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.54    1.02    1.20    1.42    1.51   49.52
```

```
## [1] "Summary of color "
```

```
##    D    E    F    G    H    I    J
## 1678 2265 2847 3830 3529 2386 1377
```

```
## [1] "Summary of clarity "
```

```
##   I1   IF  SI1  SI2  VS1  VS2 VVS1 VVS2
##  229  280 4973 4336 2460 4028  507 1099
```

```
## [1] "Summary of depth "
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.00   61.00   61.90   61.76   62.60   71.80
```

```
## [1] "Summary of table "
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.00   56.00   58.00   57.86   59.00   95.00
```

```
## [1] "Summary of x "
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.460   6.770   6.939   7.350  10.230
```

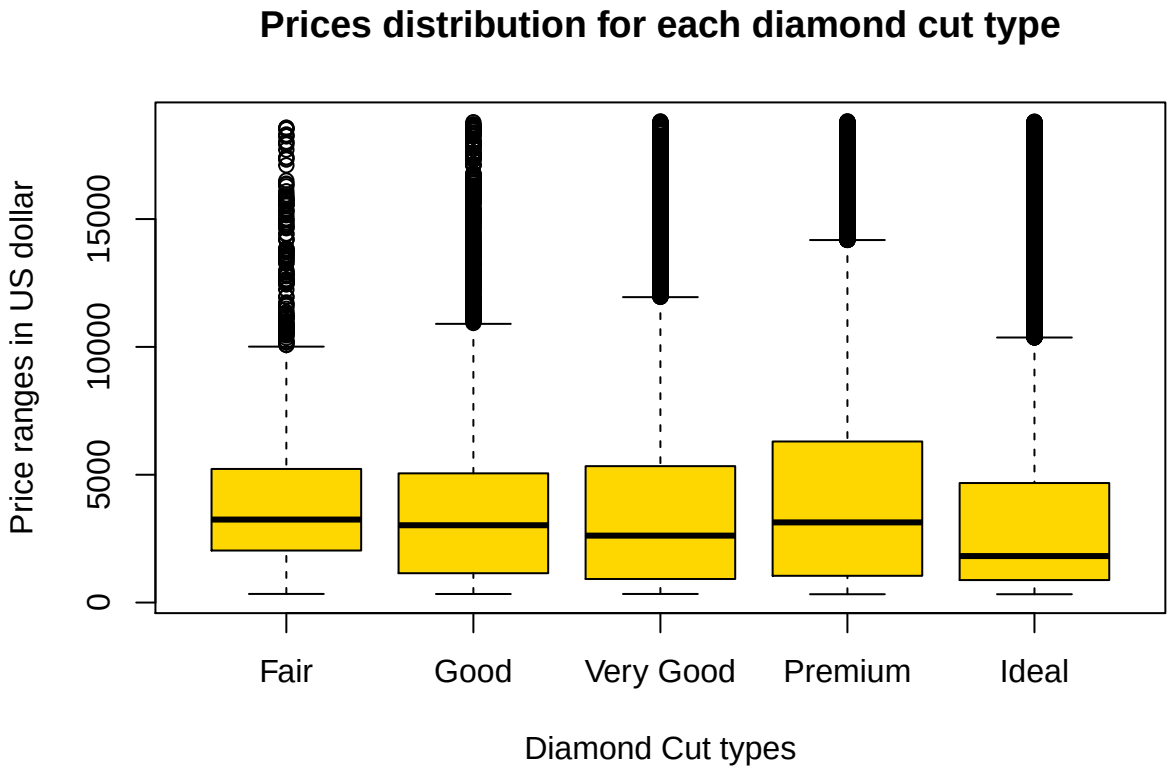
```
## [1] "Summary of y "

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.460   6.760   6.931   7.350  10.160

## [1] "Summary of z "

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.000   4.170   4.279   4.550   6.720
```

c - Price exploration for different cut type :



The price distribution for the different types of cut does not appear to have many differences, with all price distributions being mostly located on the 'low' range of diamond prices, with closely similar medians and with only a very slightly pricier `premium` type. The price does not appear to depend on the cut type.

d - Attributes correlated with price :

Based on the summaries at Part 3-b, the `carat` isn't well correlated with price, as it has almost the same distribution in all three price ranges. THe same could be said about the `depth` and `table`, as they remain with almost the same distribution of values, mean and median values over the three price ranges.

Color is also not well correlated, as for all price ranges, its distribution of frequency remains unchanging.`clarity`'s distribution changes a bit over the three price ranges, but the change aren't found to be significant, with just variations over the medium clarity types.

Based on the boxplot in Part 3-c, the `cut` variable is also not very well correlated, as it remains with the same range distribution of prices over all the cut types.

The most correlated variables appear to be `x`, `y` and `z`, that appear to grow linearly with the price range (with `y` and `z` having an outlier in the medium price range.)

4 - Diamond prices by clarity and cuts :

- Frequency of diamonds by clarity level :

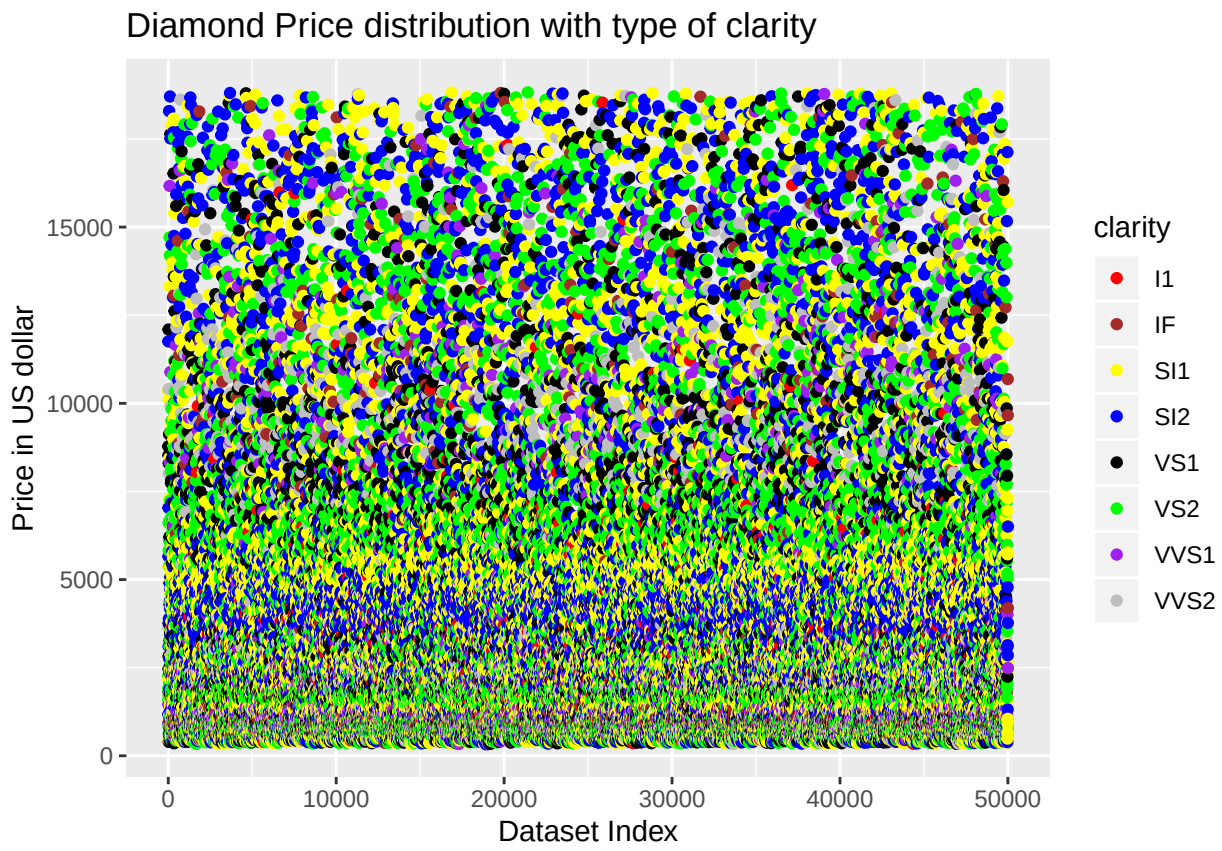
```
## # A tibble: 8 x 2
##   clarity count
##   <fct>   <int>
## 1 I1        690
## 2 IF       1666
## 3 SI1     12120
## 4 SI2      8486
## 5 VS1      7563
## 6 VS2     11406
## 7 VVS1     3377
## 8 VVS2     4692
```

- Frequency of diamonds by cut type :

```
## # A tibble: 5 x 2
##   cut       count
##   <chr>     <int>
## 1 Fair       1480
## 2 Good       4559
## 3 Ideal     19918
## 4 Premium   12826
## 5 Very Good 11217
```

```
p <- ggplot(data = DiamondData, aes(1:50000, price, color = cut)) +
  geom_point() +
  scale_color_manual(values = c("Fair" = "red", "Good" = "blue", "Very Good" = "yellow", "Premium" = "g
p <- p + labs(x = "Dataset Index",y = "Price in US dollar",title = "Diamond Price distribution with typ
p
```


```
m <- ggplot(data = DiamondData, aes(1:50000, price, color = clarity)) +
  geom_point() +
  scale_color_manual(values = c("I1" = "red", "SI2" = "blue", "SI1" = "yellow", "VS2" = "green","VS1" =
m <- m + labs(x = "Dataset Index",y = "Price in US dollar",title = "Diamond Price distribution with typ
m
```



5 - Analysis of the carat, depth, table and the dimensions (x,y,z) :

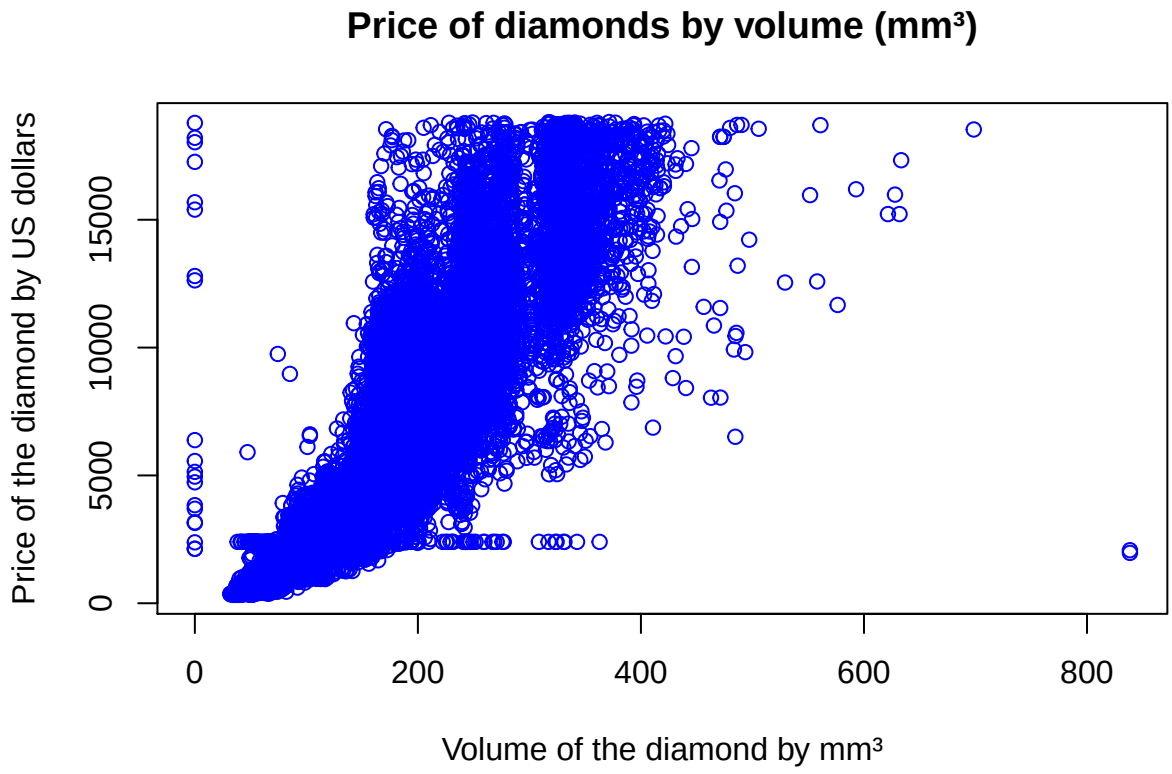
a - Volume variable :

Computing of the new Volume variable and insertion into the dataset is as follows :

```
volume <- DiamondData$x  * DiamondData$y * DiamondData$z
DiamondData$volume <- volume
DiamondData$volume <- round(DiamondData$volume, digits=2)
glimpse(DiamondData)
```

```
## Observations: 50,000
## Variables: 11
## $ carat   <dbl> 1.50, 0.70, 1.22, 0.51, 2.02, 0.70, 0.46, 0.55, 0.51, ...
## $ cut     <chr> "Fair", "Ideal", "Premium", "Premium", "Very Good", "V...
## $ color   <fct> G, E, G, E, J, E, G, F, G, H, J, E, D, H, E, H, G, J, ...
## $ clarity <fct> SI1, VS2, VS1, VS2, SI2, SI1, VVS1, SI2, VVS2, VS2, VS...
## $ depth   <dbl> 64.5, 61.4, 61.3, 62.5, 59.2, 63.2, 60.7, 60.6, 61.6, ...
## $ table   <dbl> 57, 57, 58, 60, 60, 61, 57, 58, 56, 55, 58, 56, 61, 56...
## $ price   <int> 10352, 2274, 8779, 1590, 11757, 2164, 1453, 1175, 1750...
## $ x       <dbl> 7.15, 5.72, 6.91, 5.08, 8.27, 5.61, 4.98, 5.28, 5.12, ...
## $ y       <dbl> 7.09, 5.78, 6.89, 5.10, 8.39, 5.49, 5.03, 5.31, 5.14, ...
## $ z       <dbl> 4.59, 3.53, 4.23, 3.18, 4.91, 3.51, 3.04, 3.21, 3.16, ...
## $ volume  <dbl> 232.68, 116.71, 201.39, 82.39, 340.68, 108.10, 76.15, ...
```

```r
plot(x=DiamondData$volume , y=DiamondData$price,col="blue",main="Price of diamonds by volume (mm³)",xlab
```



Diamond's `price` appears to be correlated exponentially to `volume` of the diamond.

b - Correlation of the carat and volume variables :

```r
print("Pearson correlation test result :")
```

```
## [1] "Pearson correlation test result :"
```

```r
cor(DiamondData$carat,DiamondData$volume,method="pearson")
```

```
## [1] 0.2267053
```

```r
print("Spearman correlation test result :")
```

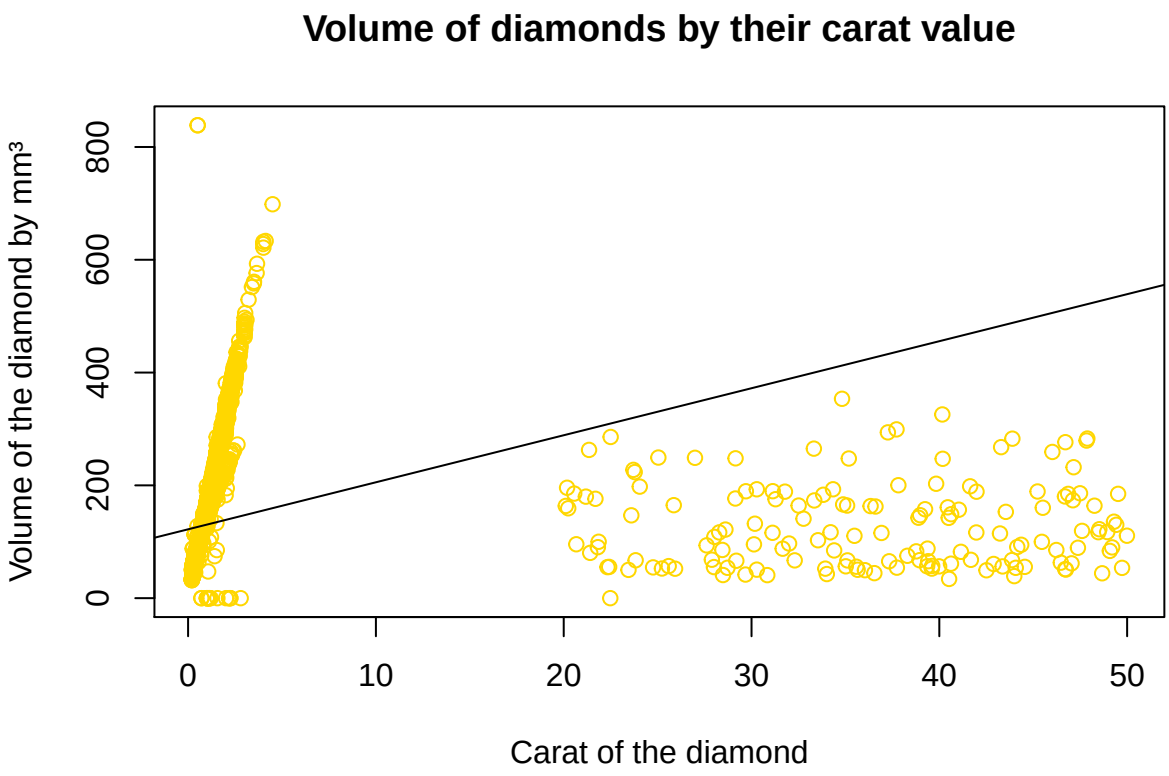
```
## [1] "Spearman correlation test result :"
```

```r
cor(DiamondData$carat,DiamondData$volume,method="spearman")
```

```
## [1] 0.9894253
```

The weak and close to 0 coefficient of Pearson shows that there is very weak linear correlation between the carat and volume variables. But the coefficient of Spearman shows the existence of a strong monotonic correlation between the two variables.

The next plot shows it more clearly :



c- Exploration of relationships between table & depth

```
print("Pearson correlation test result :")
```

```
## [1] "Pearson correlation test result :"
```

```
cor(DiamondData$table,DiamondData$depth,method="pearson")
```

```
## [1] -0.2935653
```

```
print("Spearman correlation test result :")
```

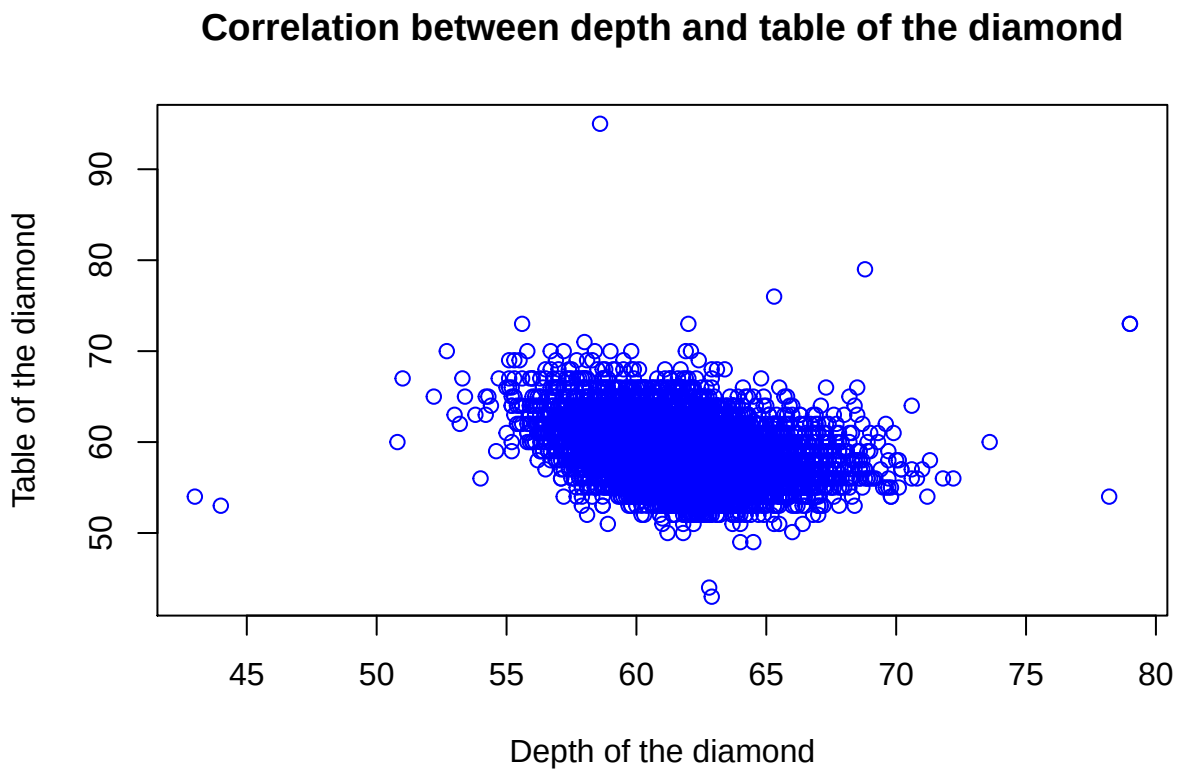
```
## [1] "Spearman correlation test result :"
```

```
cor(DiamondData$table,DiamondData$depth,method="spearman")
```

```
## [1] -0.2437824
```

The coefficients of correlation show with how close to 0 they are that there is almost no correlation, whether linear or monotonic between the two variables.

```
plot(y=DiamondData$table,x=DiamondData$depth,col="blue",xlab="Depth of the diamond",ylab="Table of the d
```



The plot shows no apparent linear or non linear correlation.

d - Exploration of relationship of table with other vaiables :

- Relationship of table with volume :

```
## [1] "Pearson correlation test result :"
```

```
## [1] 0.1701368
```

```
## [1] "Spearman correlation test result :"
```

```
## [1] 0.1820559
```

- Relationship of table with x :

```
## [1] "Pearson correlation test result :"
```

```
## [1] 0.1936679
```

```
## [1] "Spearman correlation test result :"
```

```
## [1] 0.2012749
```

- Relationship of table with y :

```
## [1] "Pearson correlation test result :"
```

```
## [1] 0.1863486
```

```
## [1] "Spearman correlation test result :"
```

```
## [1] 0.1953082
```

- Relationship of table with z :

```
## [1] "Pearson correlation test result :"
```

```
## [1] 0.149407
```

```
## [1] "Spearman correlation test result :"
```

```
## [1] 0.1592651
```

- Relationship of table with price :

```
## [1] "Pearson correlation test result :"
```

```
## [1] 0.1269137
```

```
## [1] "Spearman correlation test result :"
```

```
## [1] 0.1711841
```

- Relationship of table with carat :

```
## [1] "Pearson correlation test result :"
```

```
## [1] 0.03971152
```

```
## [1] "Spearman correlation test result :"
```

```
## [1] 0.1941334
```

It appears that the `table` variable doesn't correlate with the other variables linearly or monotonically, with all the coeficcients of correlation close to zero.