Chinmay Vijaya Kumar_2957148_MSCBD_BDA_Assignment 1

Chinmay Vijaya Kumar_2957148

03/11/2019

Summary of the Diamond Dataset Before Cleaning

```
##   carat       cut color clarity depth table price    x    y    z
## 1  0.74 Very Good     D     VS2  59.8    58  3476 5.90 5.94 3.54
## 2  0.72     Ideal     H     VS1  61.6    59  2642 5.75 5.78 3.55
## 3  0.36     Ideal     D    VVS2  61.9    53   957 4.57 4.60 2.84
## 4  0.31   Premium     I    VVS1  61.0    58   732 4.39 4.33 2.66
## 5  1.00   Premium     H     SI2  59.1    62  3640 6.50 6.47 3.83
## 6  0.50   Premium     F     SI2  61.4    61  1172 5.14 5.09 3.14
```

```
## 'data.frame':    50000 obs. of  10 variables:
##  $ carat  : num  0.74 0.72 0.36 0.31 1 0.5 1.07 0.53 1.5 1.01 ...
##  $ cut    : Factor w/ 5 levels "Fair","Good",..: 5 3 3 4 4 4 1 4 2 3 ...
##  $ color  : Factor w/ 7 levels "D","E","F","G",..: 1 5 1 6 5 3 3 2 3 3 ...
##  $ clarity: Factor w/ 8 levels "I1","IF","SI1",..: 6 5 8 7 4 4 3 5 6 4 ...
##  $ depth  : num  59.8 61.6 61.9 61 59.1 61.4 60.6 58.5 63.6 62.9 ...
##  $ table  : num  58 59 53 58 62 61 66 61 55 57 ...
##  $ price  : int  3476 2642 957 732 3640 1172 4554 1950 13853 4858 ...
##  $ x      : num  5.9 5.75 4.57 4.39 6.5 5.14 6.65 5.39 7.27 6.35 ...
##  $ y      : num  5.94 5.78 4.6 4.33 6.47 5.09 6.46 5.28 7.22 6.41 ...
##  $ z      : num  3.54 3.55 2.84 2.66 3.83 3.14 3.97 3.12 4.61 4.01 ...
```

```
##      carat               cut        color        clarity          depth      
##  Min.   :0.2000   Fair     : 1500   D: 6269   SI1    :12110   Min.   :43.00  
##  1st Qu.:0.4000   Good     : 4539   E: 9097   VS2    :11365   1st Qu.:61.00  
##  Median :0.7000   Ideal    :20011   F: 8854   SI2    : 8501   Median :61.80  
##  Mean   :0.7974   Premium  :12770   G:10463   VS1    : 7582   Mean   :61.75  
##  3rd Qu.:1.0400   Very Good:11180   H: 7666   VVS2   : 4692   3rd Qu.:62.50  
##  Max.   :5.0100                     I: 5029   VVS1   : 3404   Max.   :79.00  
##                                     J: 2622   (Other): 2346                  
##      table           price             x                y        
##  Min.   :43.00   Min.   :  326   Min.   : 0.00   Min.   : 0.000  
##  1st Qu.:56.00   1st Qu.:  949   1st Qu.: 4.71   1st Qu.: 4.720  
##  Median :57.00   Median : 2401   Median : 5.70   Median : 5.710  
##  Mean   :57.45   Mean   : 3925   Mean   : 5.73   Mean   : 5.732  
##  3rd Qu.:59.00   3rd Qu.: 5312   3rd Qu.: 6.54   3rd Qu.: 6.540  
##  Max.   :95.00   Max.   :18823   Max.   :10.74   Max.   :31.800  
##                                                                  
##        z         
##  Min.   : 0.000  
```

```
##  1st Qu.: 2.910
##  Median : 3.520
##  Mean   : 3.538
##  3rd Qu.: 4.030
##  Max.   :31.800
##
```

```
## [1] 50000
```

The above data shows summary of the raw Diamond dataset (Structure, Summary and Row Count)

Task 1

Cleaning the Diamonds Data Set (Removing NA's and Outliers)

```
##      carat               cut        color        clarity          depth
##  Min.   :0.2000   Fair      : 1500   D: 6269   SI1    :12110   Min.   :43.00
##  1st Qu.:0.4000   Good      : 4539   E: 9097   VS2    :11365   1st Qu.:61.00
##  Median :0.7000   Very Good:11180   F: 8854   SI2    : 8501   Median :61.80
##  Mean   :0.7974   Premium   :32781   G:10463   VS1    : 7582   Mean   :61.75
##  3rd Qu.:1.0400                     H: 7666   VVS2   : 4692   3rd Qu.:62.50
##  Max.   :5.0100                     I: 5029   VVS1   : 3404   Max.   :79.00
##                                     J: 2622   (Other): 2346
##      table           price             x                y
##  Min.   :43.00   Min.   :  326   Min.   : 3.730   Min.   : 3.680
##  1st Qu.:56.00   1st Qu.:  949   1st Qu.: 4.710   1st Qu.: 4.720
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
##  Mean   :57.45   Mean   : 3925   Mean   : 5.731   Mean   : 5.733
##  3rd Qu.:59.00   3rd Qu.: 5312   3rd Qu.: 6.540   3rd Qu.: 6.540
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :31.800
##
##        z
##  Min.   : 1.070
##  1st Qu.: 2.910
##  Median : 3.520
##  Mean   : 3.539
##  3rd Qu.: 4.030
##  Max.   :31.800
##
```

```
## [1] 50000
```

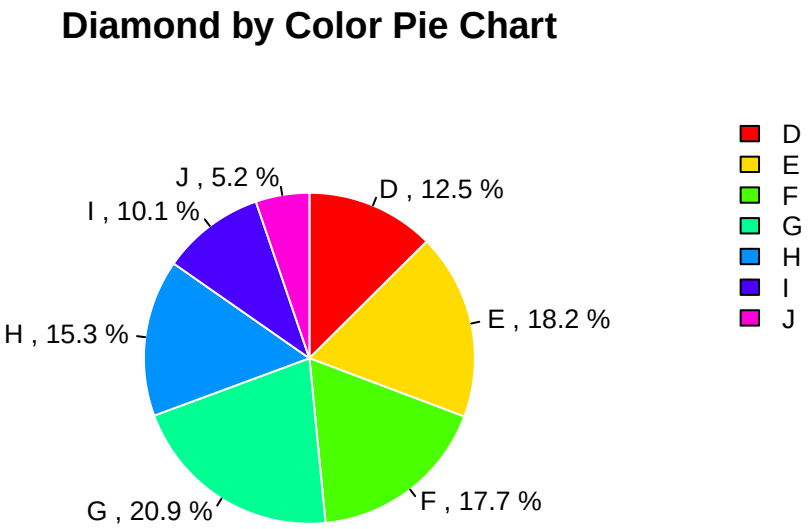
The above data shows summary of the cleaned Diamond dataset.

Every data set has to be cleaned before using it i.e. removing incomplete rows, columns or any insignificant columns. The raw Diamonds dataset had 50000 observations, after removal of NA's and Outliers the row count reduced to 48313 observations. 1687 observations which were incomplete were removed.

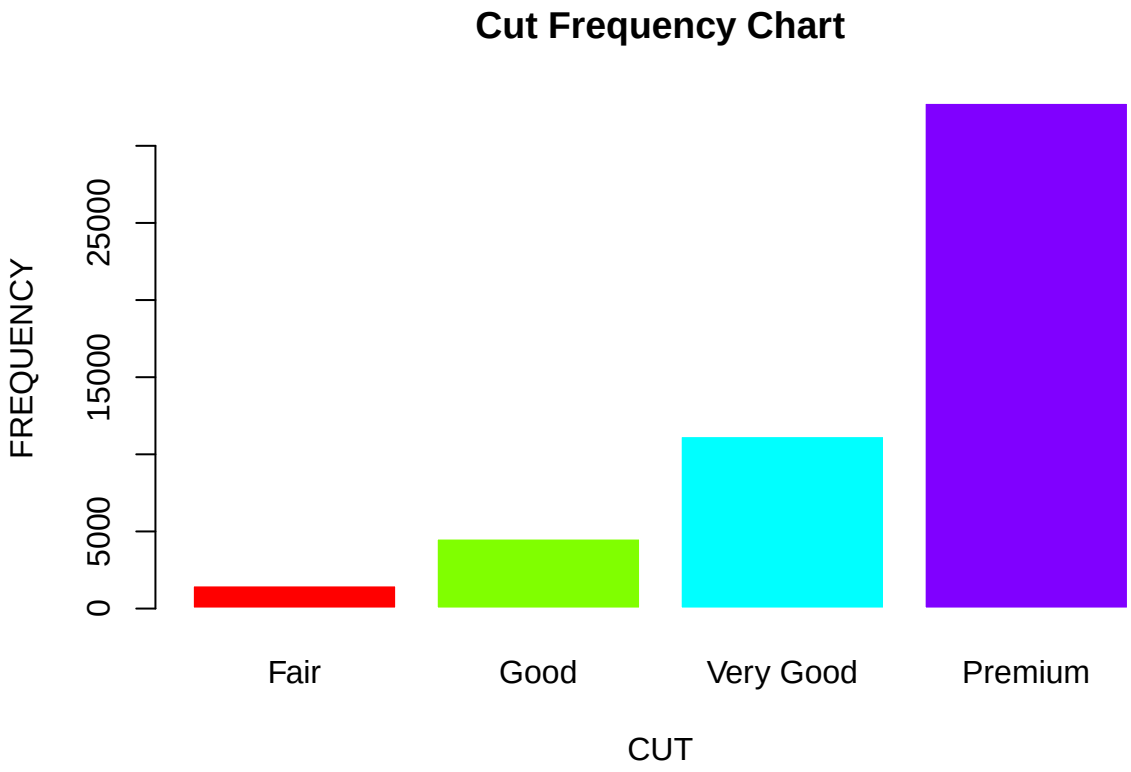
Task 2

4 Plots: Pie Chart, Bar Chart, Histogram, Scatter Plot

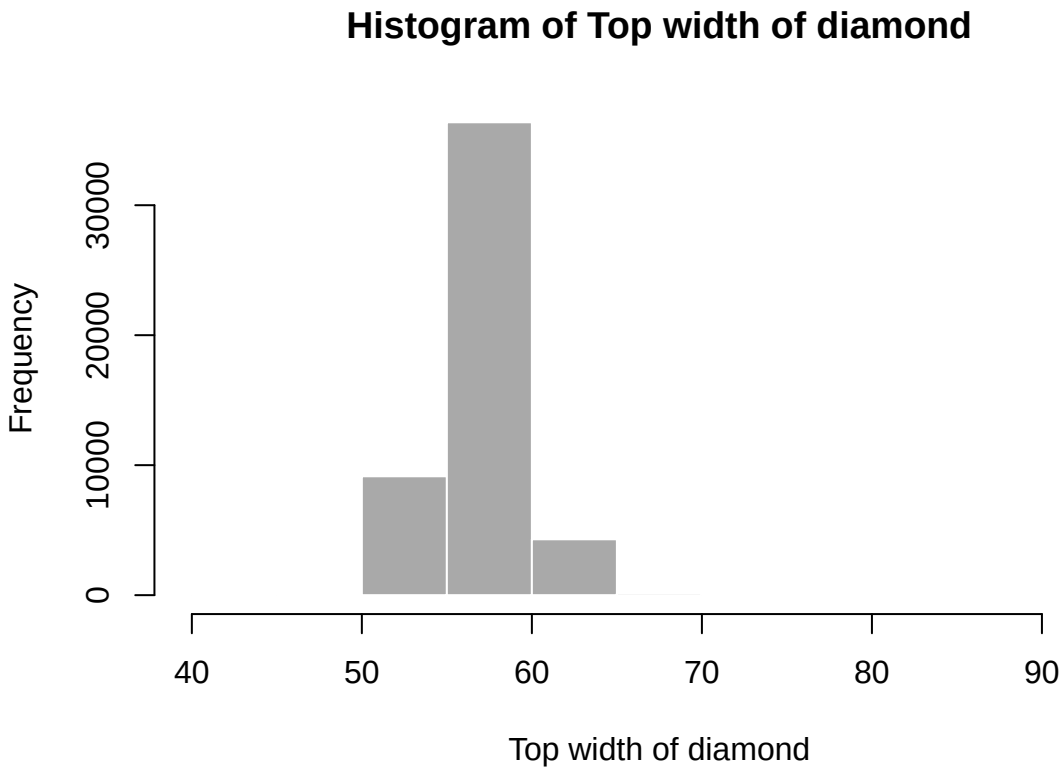
Pie Chart



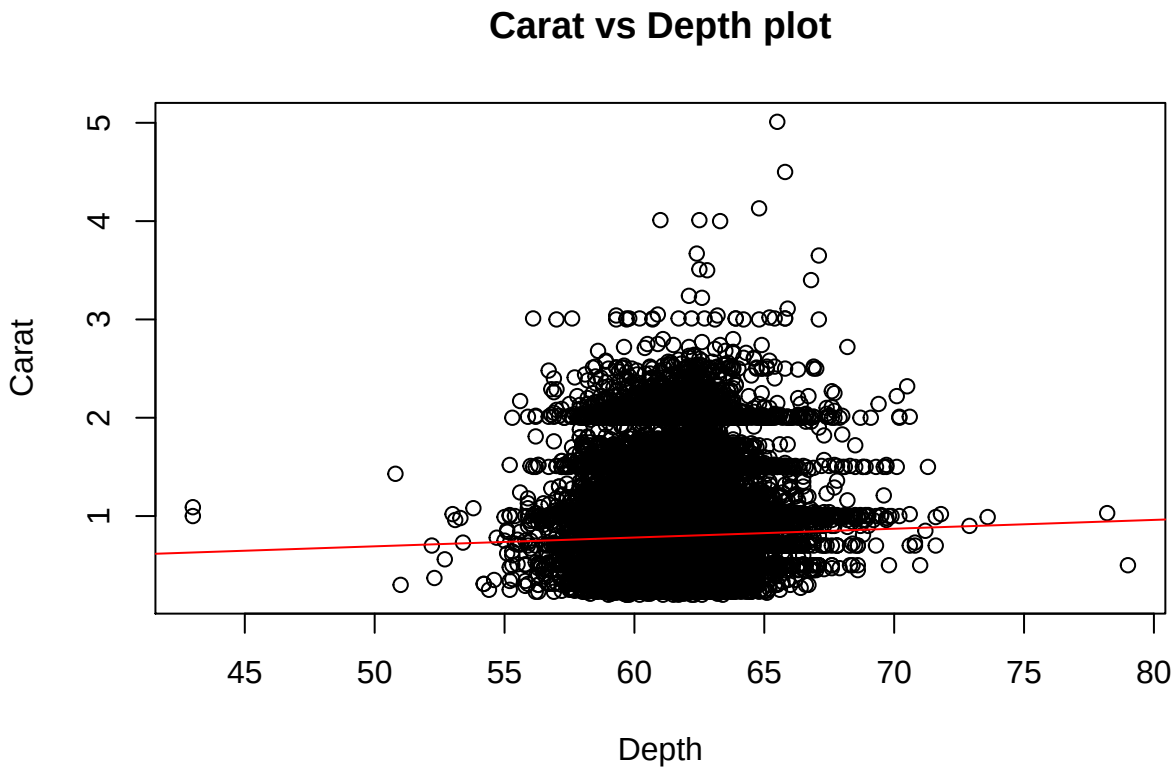
Bar Chart


Cut Frequency Chart
FREQUENCY
0
5000
15000
25000
Fair
Good
Very Good
Premium
CUT

Histogram


Histogram of Top width of diamond
Frequency
0
10000
20000
30000
40
50
60
70
80
90
Top width of diamond

Scatter Plot

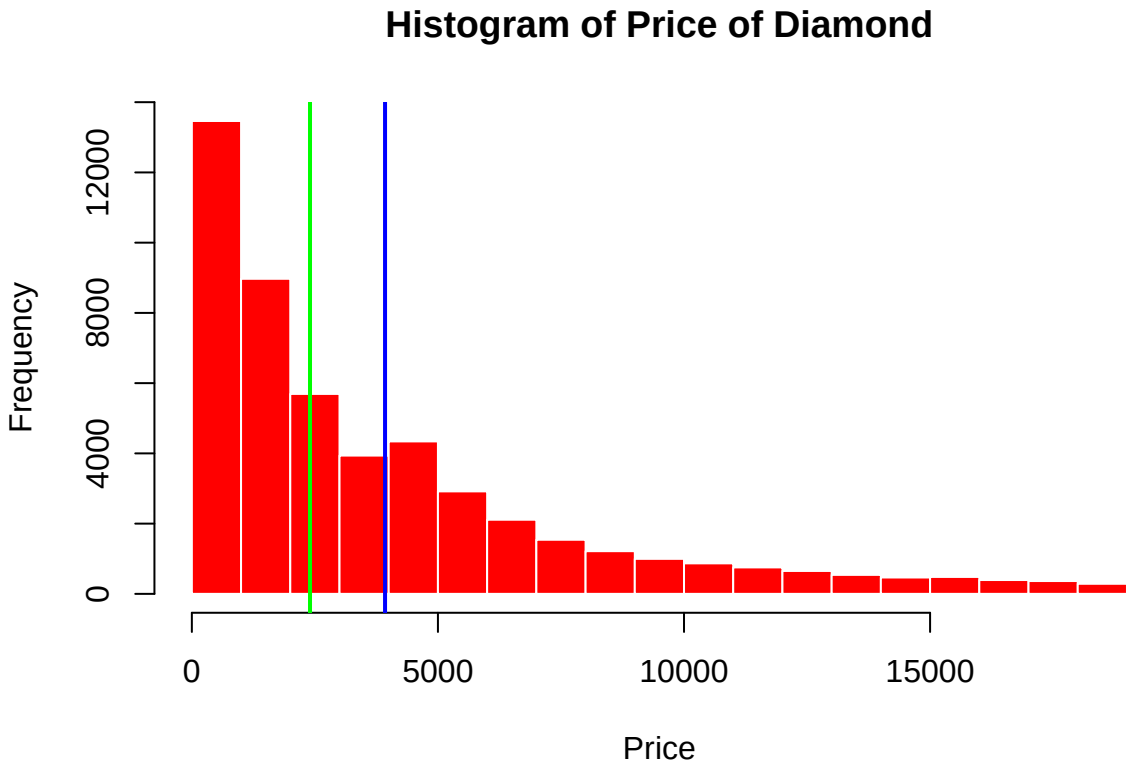


The distribution for the all the above graphs is positively skewed. This means lesser plots are at the larger numeric value. The mean typically gets pulled toward the tail, and is greater than the median.

Task 3

Analysis on the Price Variable

Price Histogram




```
## [1] 15840563
```

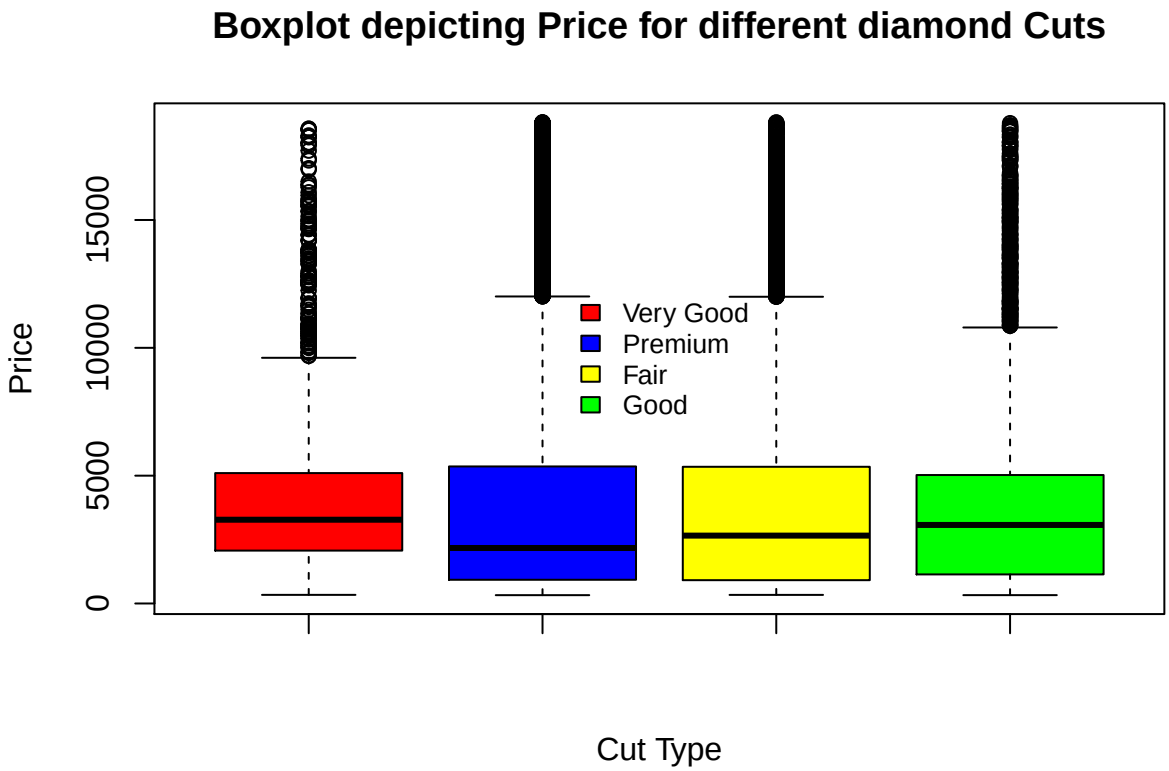
The Histogram for the price depicts that the frequency reduces as the price increases. The mean line is shown by the blue line. The median is shown by the yellow line. Variance is a huge value equaling 15979107 which cannot be displayed on the graph.

Group Diamond by Price range (Low, Medium, High)

```
## # A tibble: 3 x 8
##   range  count   mean median      var   min   max    sd
##   <fct>  <int>  <dbl>  <dbl>    <dbl> <int> <int> <dbl>
## 1 High    2535 15650.  15531 2753436. 13001 18823 1659.
## 2 Medium  7016  9103.   8820 3430178.  6500 12998 1852.
## 3 Low    40449  2293.   1723 2881565.   326  6499 1698.
```

A new column range has been added into the table and factorized. Display the summary which includes mean, median, standard deviation, variance, minimum and maximum.

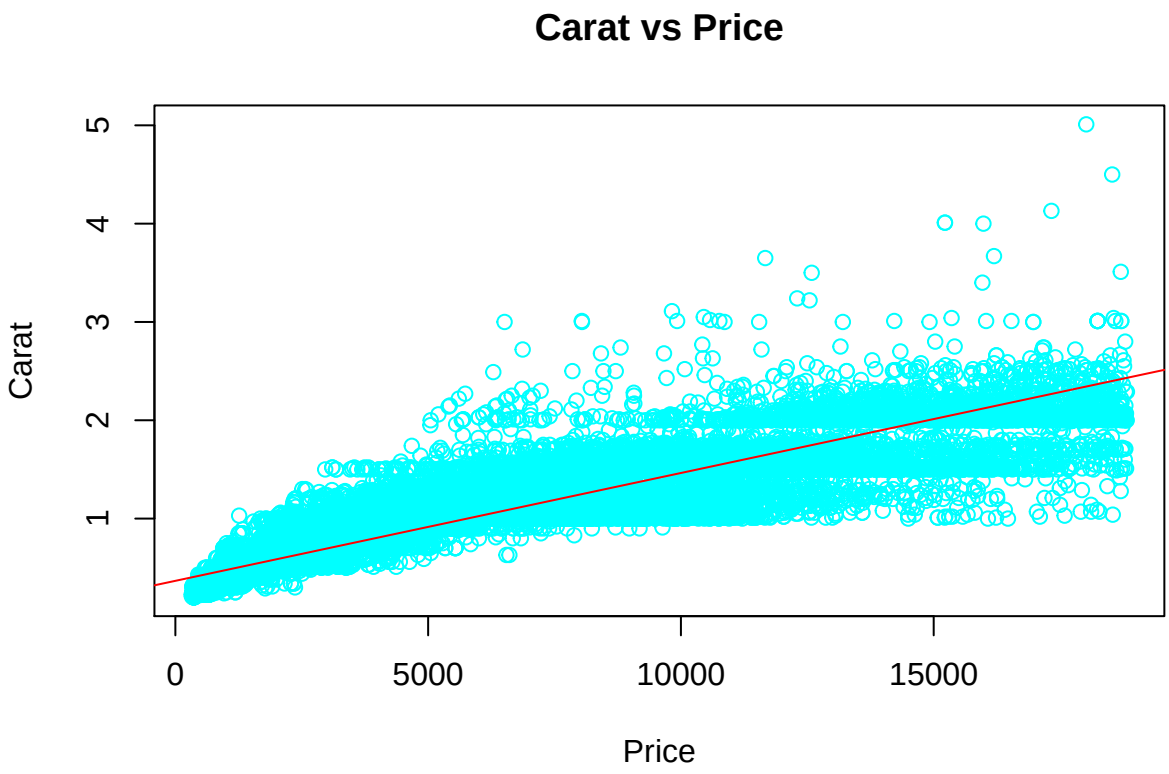
Explore prices for different cut types using Boxplot



Boxplot displays the different Prices for the 4 Cut types Fair, Premium, Very Good and Good.

Correlation between all attributes with Price

```
##          carat     table         x         y         z       depth
## [1,] 0.9213562 0.1278579 0.8871083 0.8836283 0.8667867 -0.01213396
```



The highest correlation for depth is with carat, x and y, highest being x. Plot for carat vs price is as shown by the graph above with the regression line.

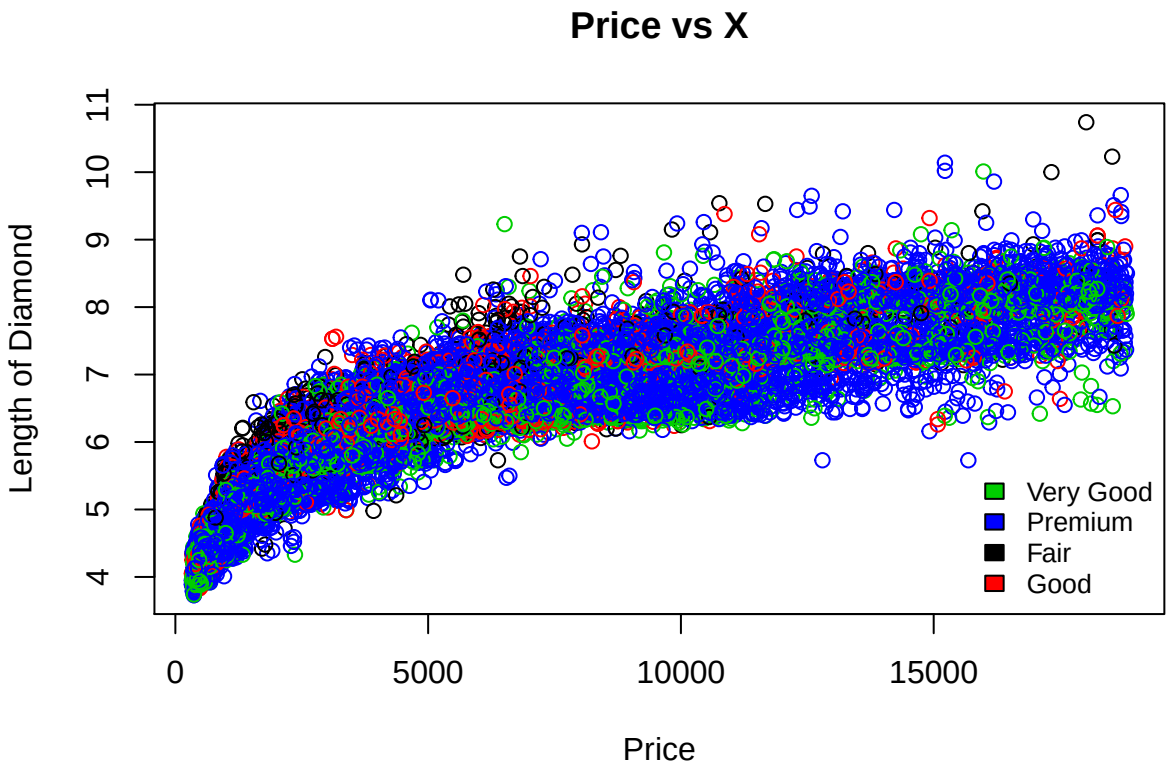
Task 4

Frequencies of Diamonds for various cuts and clarity

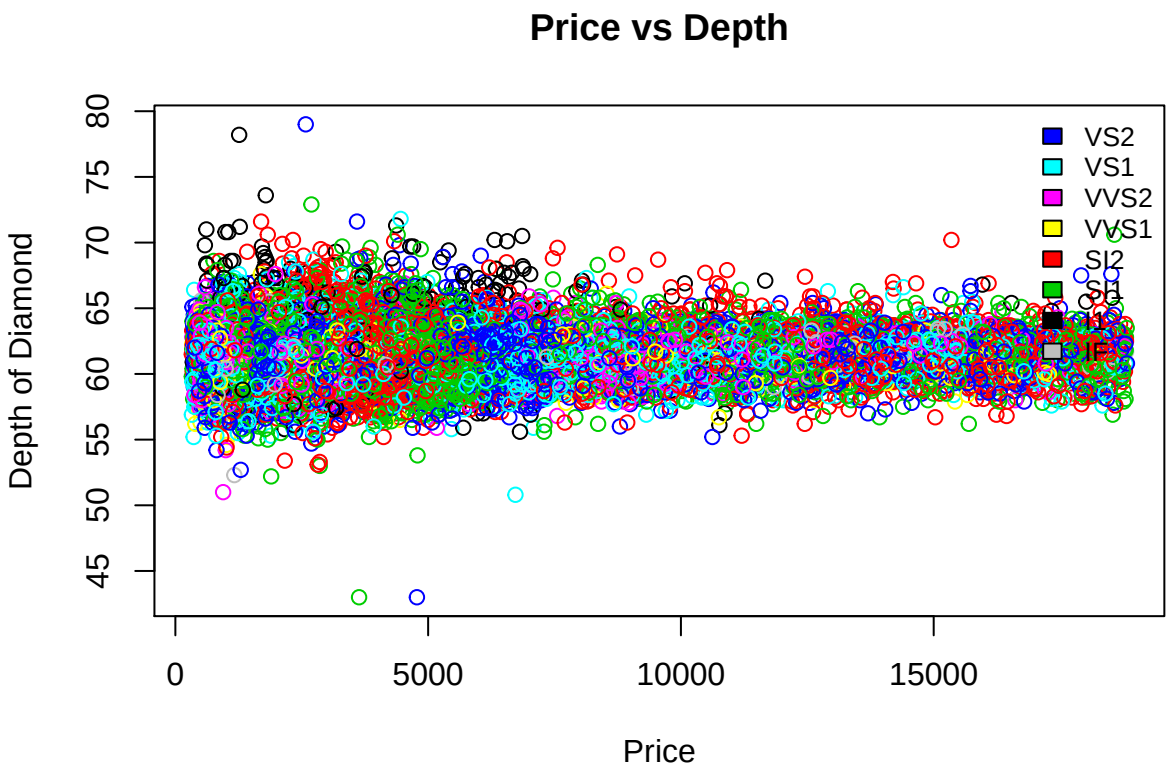
```
##
##      Fair      Good Very Good   Premium
##      1500      4539     11180     32781

##
##    I1   SI2   SI1   VS2   VS1  VVS2  VVS1    IF
##   679  8501 12110 11365  7582  4692  3404  1667
```

2 Scatter plots, color the diamonds price by clarity and cuts.



Color code for Diamonds price by Cut



Color code for Diamonds price by Clarity

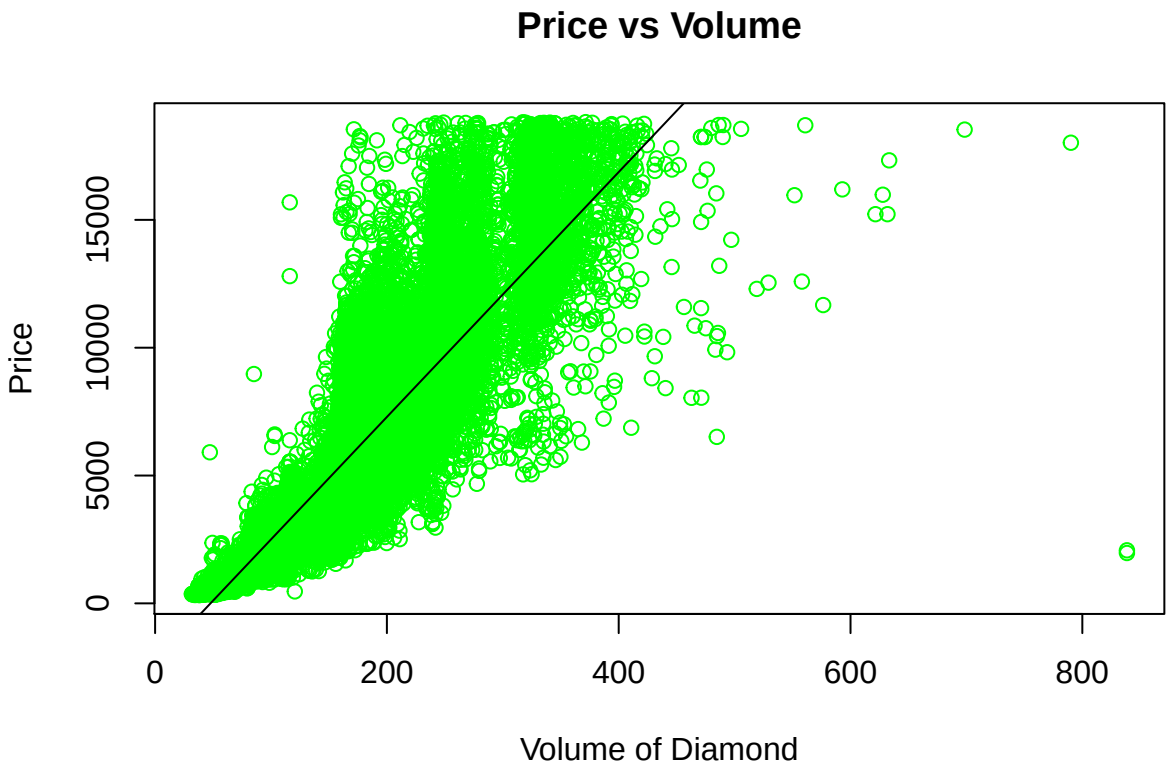
Task 5

Compute Volume variable from x, y, z

```
##      carat               cut        color        clarity          depth
##  Min.   :0.2000   Fair      : 1500   D: 6269   SI1     :12110   Min.   :43.00
##  1st Qu.:0.4000   Good      : 4539   E: 9097   VS2     :11365   1st Qu.:61.00
##  Median :0.7000   Very Good:11180   F: 8854   SI2     : 8501   Median :61.80
##  Mean   :0.7974   Premium   :32781   G:10463   VS1     : 7582   Mean   :61.75
##  3rd Qu.:1.0400                     H: 7666   VVS2    : 4692   3rd Qu.:62.50
##  Max.   :5.0100                     I: 5029   VVS1    : 3404   Max.   :79.00
##                                     J: 2622   (Other): 2346
##      table           price             x                y
##  Min.   :43.00   Min.   :  326   Min.   : 3.730   Min.   : 3.680
##  1st Qu.:56.00   1st Qu.:  949   1st Qu.: 4.710   1st Qu.: 4.720
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
##  Mean   :57.45   Mean   : 3925   Mean   : 5.731   Mean   : 5.733
##  3rd Qu.:59.00   3rd Qu.: 5312   3rd Qu.: 6.540   3rd Qu.: 6.540
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :31.800
##
##        z             range           volume
##  Min.   : 1.070   High  : 2535   Min.   : 31.71
##  1st Qu.: 2.910   Medium: 7016   1st Qu.: 65.16
```

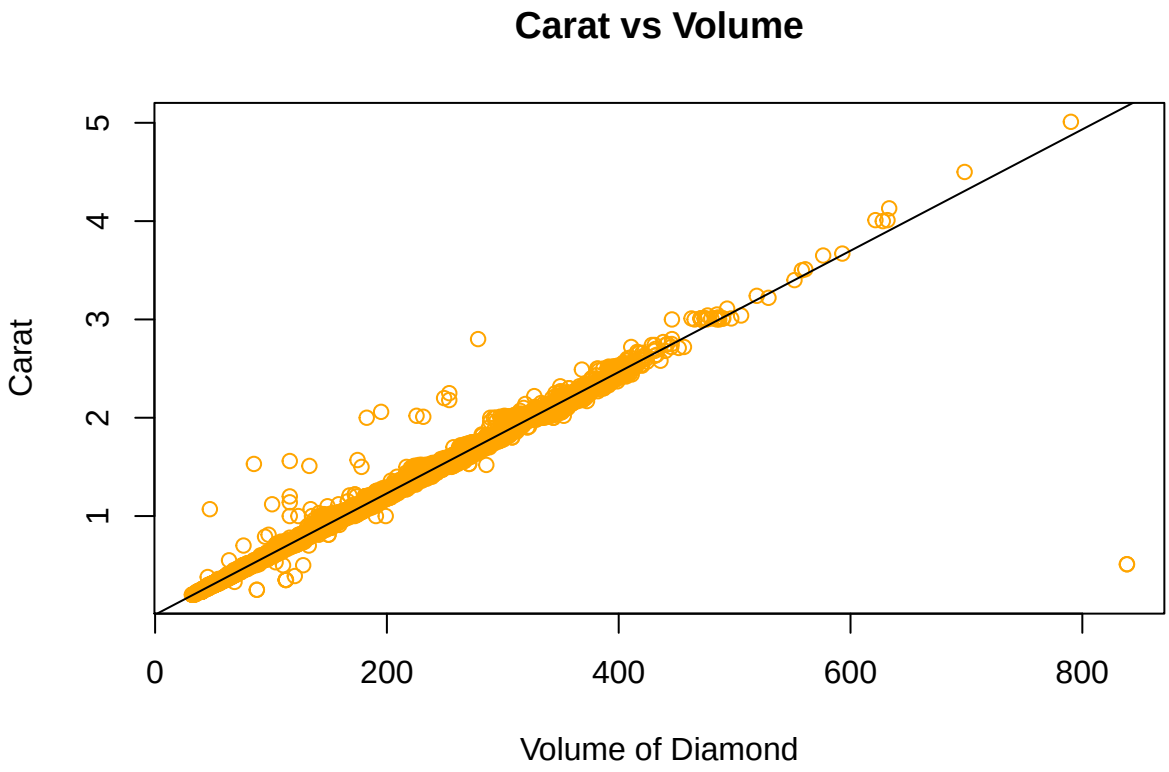
```
##  Median : 3.520   Low   :40449   Median :114.81
##  Mean   : 3.539                   Mean   :129.75
##  3rd Qu.: 4.030                   3rd Qu.:170.76
##  Max.   :31.800                   Max.   :838.50
##
```

Price vs Volume



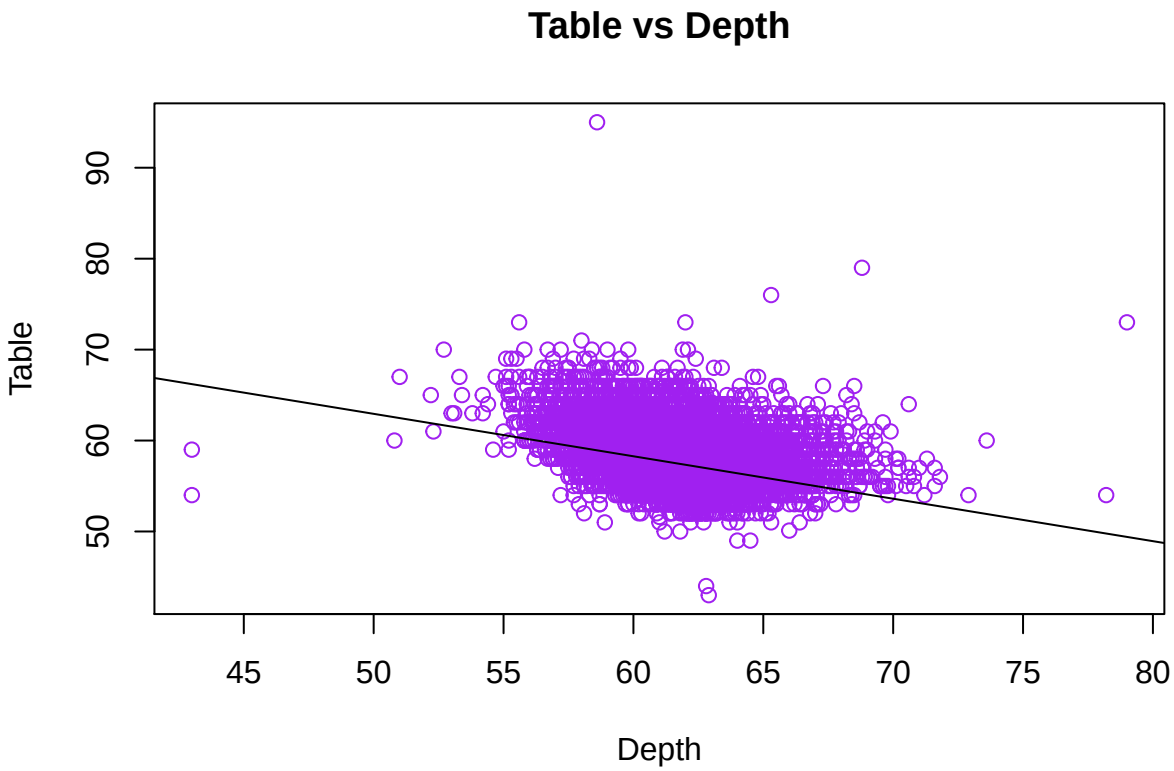
There is an exponential increase in Price with the increase in the volume of the diamond. The regression line shows a good correlation between price and volume.

Carat vs Volume



There is a linear progression in Carat with the increase in the volume of the diamond. The regression line shows a good correlation between carat and volume.

Relationship between Table and Depth



There is a negative correlation between Table and Depth close to zero. They are not correlated according to the graph.

Correlation between table and all other numeric columns

```
##          carat     price         x         y         z     depth    volume
## [1,] 0.1830079 0.1278579 0.1975854 0.1898482 0.1536813 -0.297866 0.1728202
```

Table is least correlated with each of the numeric values in the dataset. It has a negative and least correlation with depth. It has a positive correlation and least correlated with all the other variables which is close to zero.