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  1.50      Fair     G     SI1  64.5    57 10352 7.15 7.09 4.59
## 2  0.70     Ideal     E     VS2  61.4    57  2274 5.72 5.78 3.53
## 3  1.22   Premium     G     VS1  61.3    58  8779 6.91 6.89 4.23
## 4  0.51   Premium     E     VS2  62.5    60  1590 5.08 5.10 3.18
## 5  2.02 Very Good     J     SI2  59.2    60 11757 8.27 8.39 4.91
## 6  0.70 Very Good     E     SI1  63.2    61  2164 5.61 5.49 3.51
```

```
## 'data.frame':    50000 obs. of  10 variables:
##  $ carat  : num  1.5 0.7 1.22 0.51 2.02 0.7 0.46 0.55 0.51 0.5 ...
##  $ cut    : Factor w/ 6 levels "Fair","Good",..: 1 3 4 4 6 6 3 4 3 3 ...
##  $ color  : Factor w/ 7 levels "D","E","F","G",..: 4 2 4 2 7 2 4 3 4 5 ...
##  $ clarity: Factor w/ 8 levels "I1","IF","SI1",..: 3 6 5 6 4 3 7 4 8 6 ...
##  $ depth  : num  64.5 61.4 61.3 62.5 59.2 63.2 60.7 60.6 61.6 62.3 ...
##  $ table  : num  57 57 58 60 60 61 57 58 56 55 ...
##  $ price  : int  10352 2274 8779 1590 11757 2164 1453 1175 1750 1921 ...
##  $ x      : num  7.15 5.72 6.91 5.08 8.27 5.61 4.98 5.28 5.12 5.09 ...
##  $ y      : num  7.09 5.78 6.89 5.1 8.39 5.49 5.03 5.31 5.14 5.12 ...
##  $ z      : num  4.59 3.53 4.23 3.18 4.91 3.51 3.04 3.21 3.16 3.18 ...
```

```
##      carat               cut        color        clarity     
##  Min.   : 0.200   Fair     : 1480   D: 6264   SI1    :12120  
##  1st Qu.: 0.400   Good     : 4559   E: 9066   VS2    :11406  
##  Median : 0.700   Ideal    :19918   F: 8837   SI2    : 8486  
##  Mean   : 0.907   Premium  :12826   G:10493   VS1    : 7563  
##  3rd Qu.: 1.050   Very Geod: 2242   H: 7705   VVS2   : 4692  
##  Max.   :49.990   Very Good: 8975   I: 5028   VVS1   : 3377  
##                                     J: 2607   (Other): 2356  
##      depth           table           price             x         
##  Min.   :43.00   Min.   :43.00   Min.   :  326   Min.   : 0.000  
##  1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  949   1st Qu.: 4.710  
##  Median :61.80   Median :57.00   Median : 2401   Median : 5.700  
##  Mean   :61.75   Mean   :57.46   Mean   : 3939   Mean   : 5.732  
##  3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5339   3rd Qu.: 6.540  
##  Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.230  
##  NA's   :471     NA's   :390     NA's   :253     NA's   :221     
##        y               z         
##  Min.   : 0.000   Min.   : 0.000  
##  1st Qu.: 4.720   1st Qu.: 2.910  
##  Median : 5.710   Median : 3.530  
##  Mean   : 5.734   Mean   : 3.539  
##  3rd Qu.: 6.540   3rd Qu.: 4.040
```

```
##  Max.   :31.800   Max.   :31.800
##  NA's   :333      NA's   :428
```

```
## [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
##  Min.   :0.2000   Fair      : 1433   D: 6034   SI1    :11697
##  1st Qu.:0.4000   Good      : 4396   E: 8783   VS2    :11043
##  Median :0.7000   Very Good:10825   F: 8516   SI2    : 8200
##  Mean   :0.8073   Premium   :31659   G:10145   VS1    : 7299
##  3rd Qu.:1.0500                     H: 7452   VVS2   : 4530
##  Max.   :4.9990                     I: 4874   VVS1   : 3274
##                                     J: 2509   (Other): 2270
##      depth           table           price             x
##  Min.   :43.00   Min.   :43.00   Min.   :  326   Min.   : 3.730
##  1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  949   1st Qu.: 4.710
##  Median :61.80   Median :57.00   Median : 2403   Median : 5.700
##  Mean   :61.75   Mean   :57.46   Mean   : 3942   Mean   : 5.733
##  3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5351   3rd Qu.: 6.540
##  Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.230
##
##        y               z
##  Min.   : 3.680   Min.   : 1.070
##  1st Qu.: 4.720   1st Qu.: 2.910
##  Median : 5.710   Median : 3.530
##  Mean   : 5.736   Mean   : 3.541
##  3rd Qu.: 6.540   3rd Qu.: 4.040
##  Max.   :31.800   Max.   :31.800
##
```

```
## [1] 48313
```

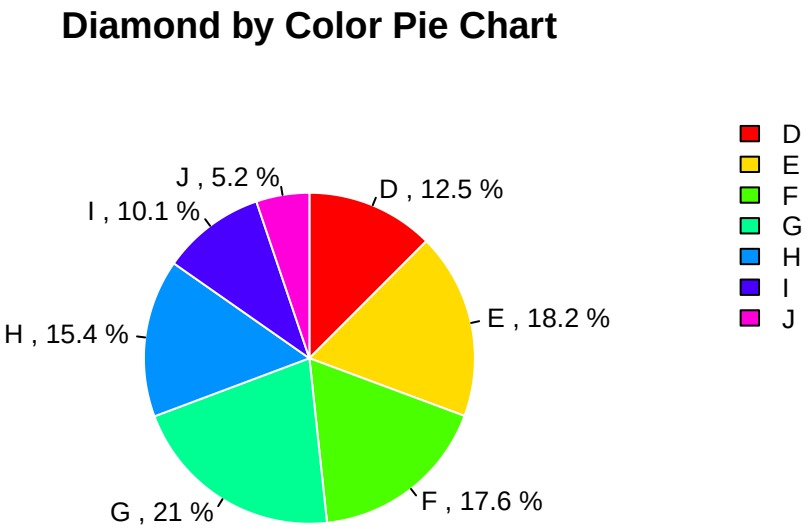
The above data shows summary of the cleaned Diamond dataset.

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.

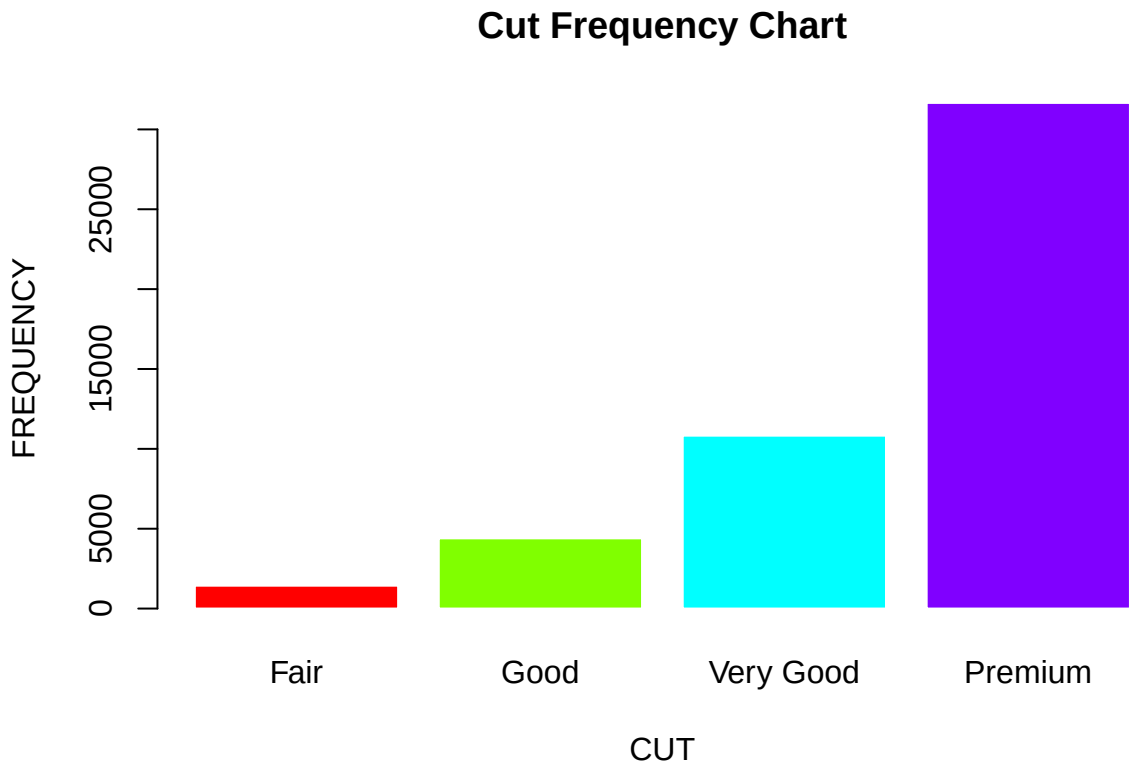
Task 2

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

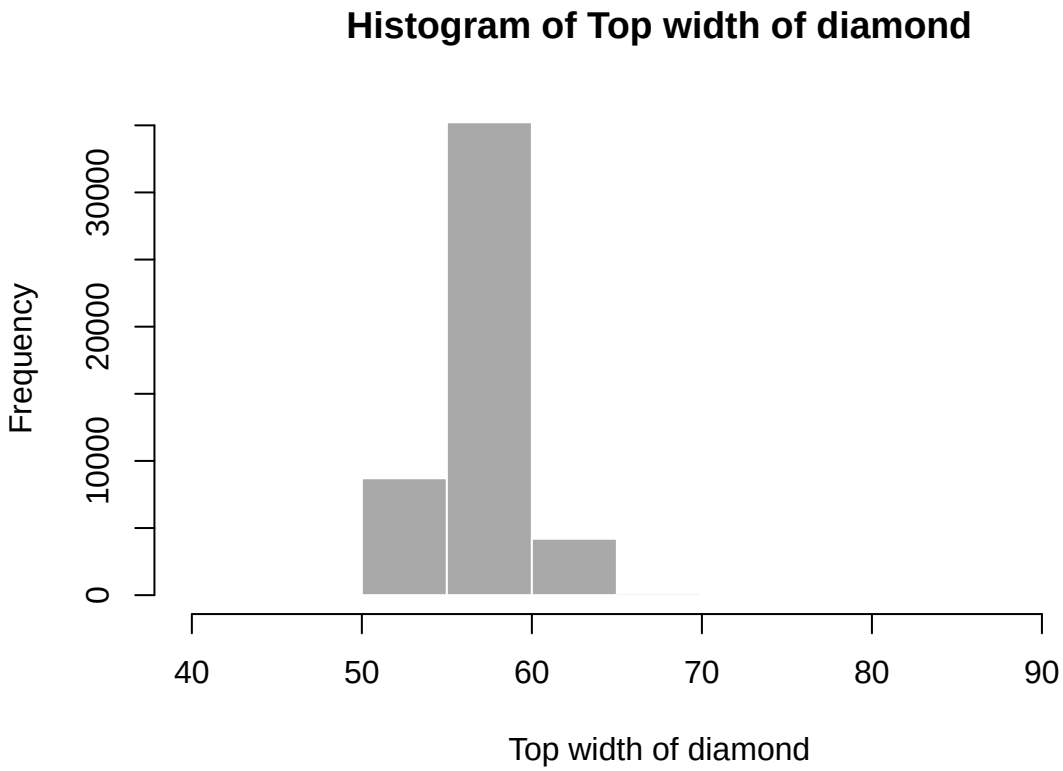
Pie Chart



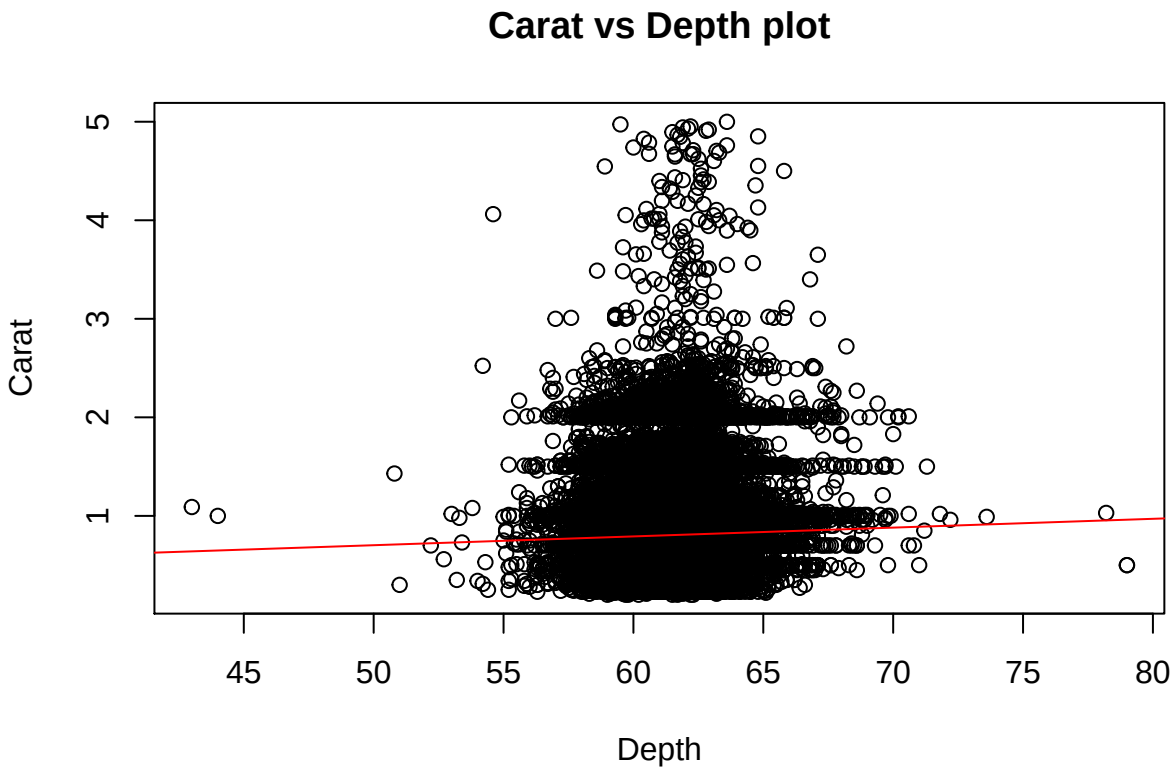
Bar Chart


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

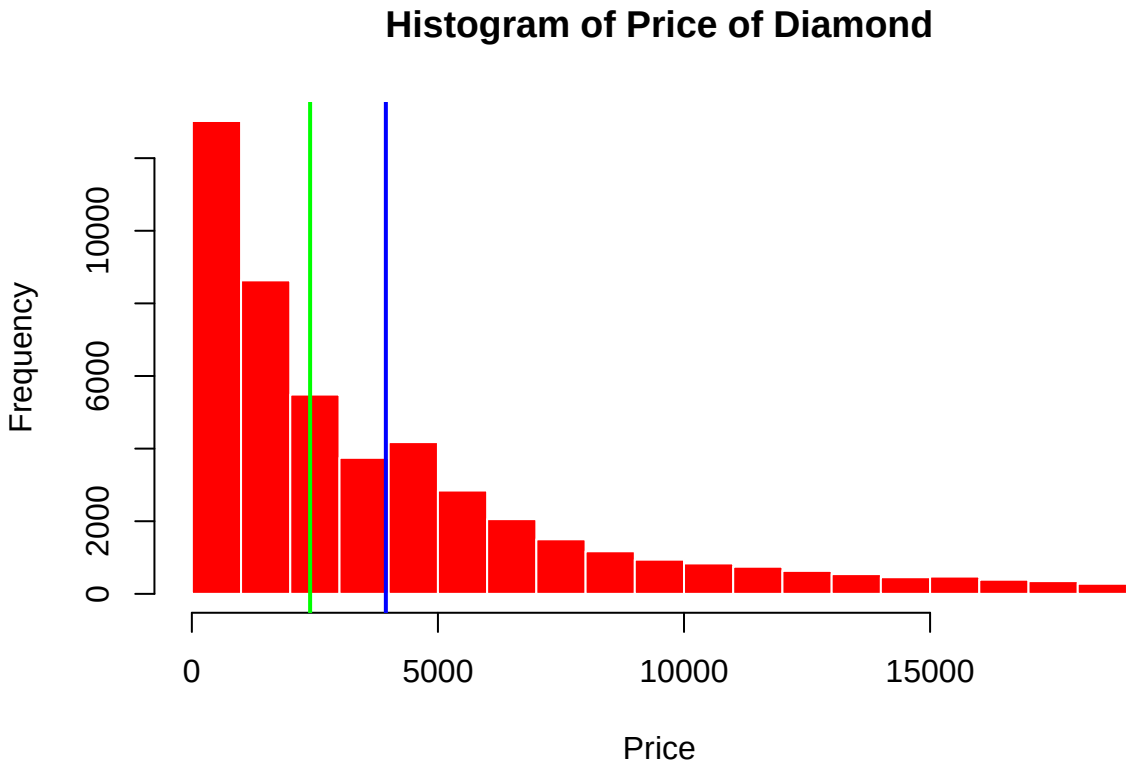


The distribution for the all the above graphs is positively skewed. This means lesser plots are at the larger numeric value. The mean typically gets pulled toward the tail, and is greater than the median.

Task 3

Analysis on the Price Variable

Price Histogram




```
## [1] 15979107
```

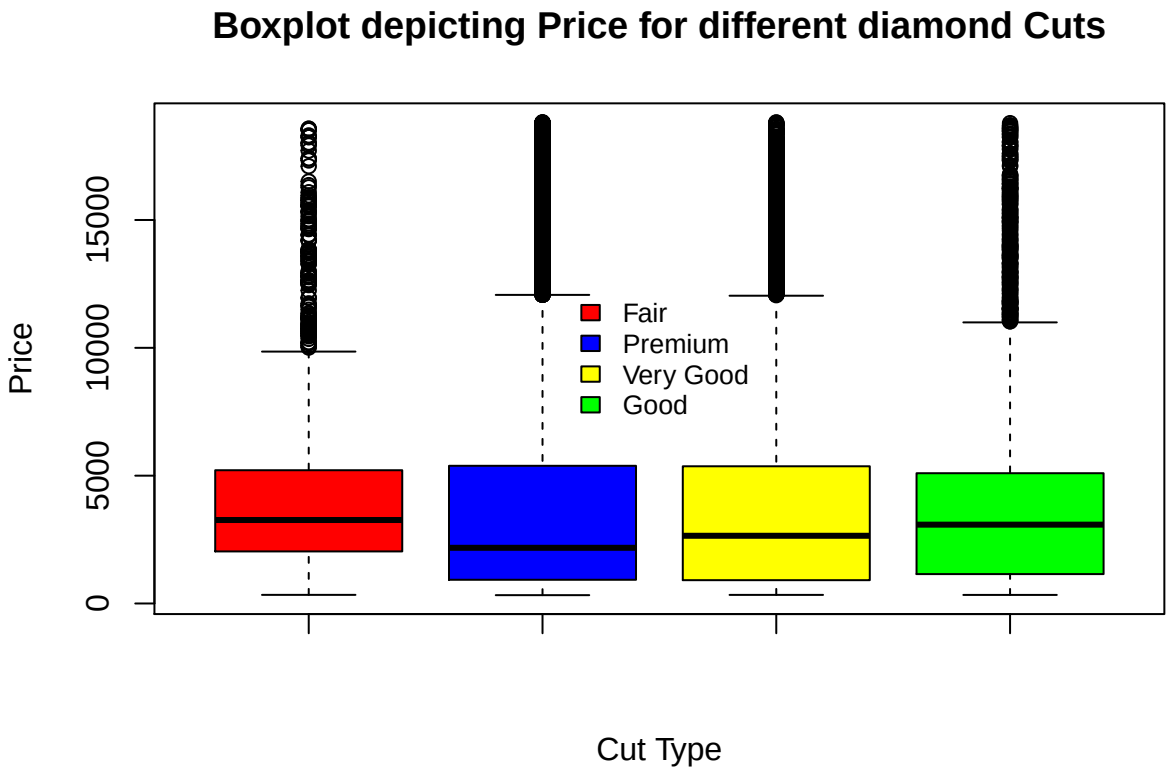
The Histogram for the price depicts that the frequency reduces as the price increases. The mean line is shown by the blue line. The median is shown by the yellow line. Variance is a huge value equaling 15979107 which cannot be displayed on the graph.

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    2498 15633. 15516  2748908. 13001 18823 1658.
## 2 Medium  6800  9105.  8808. 3463320.  6500 12998 1861.
## 3 Low    39015  2294.  1723  2893716.   326  6499 1701.
```

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.

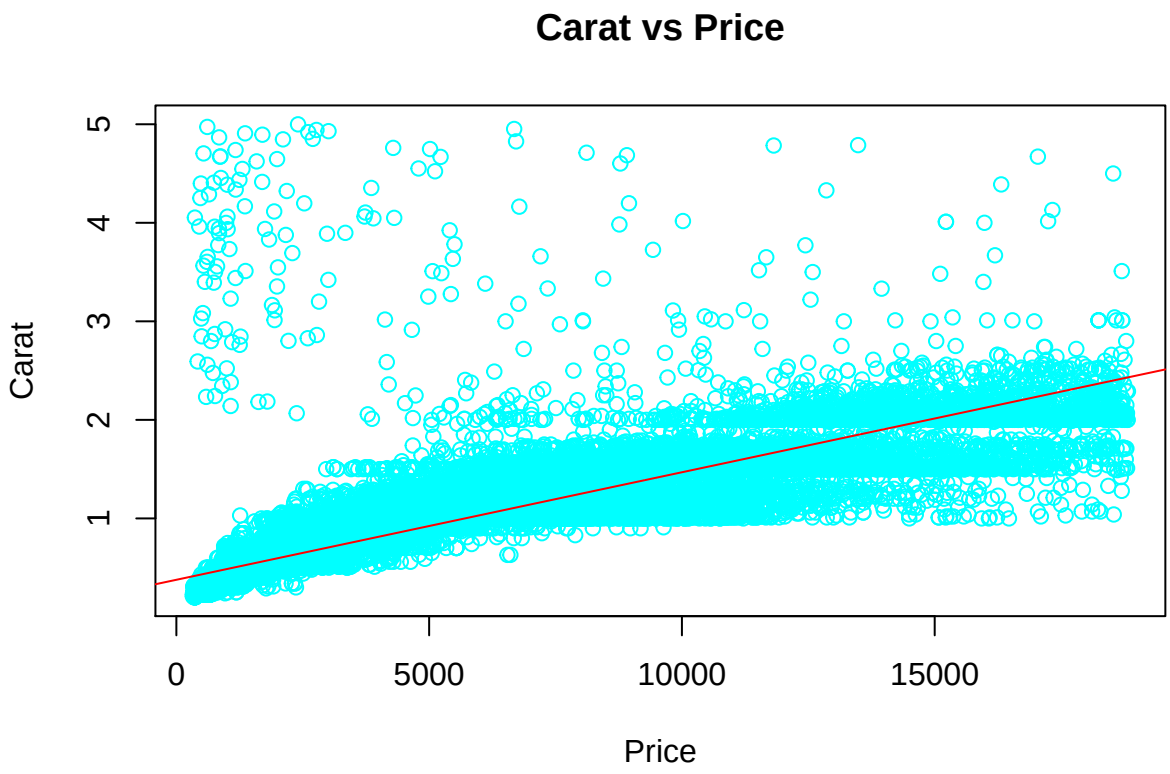
Explore prices for different cut types using Boxplot



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,] 0.8705155 0.1271754 0.8873361 0.8836565 0.866359 -0.01273964
```



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.

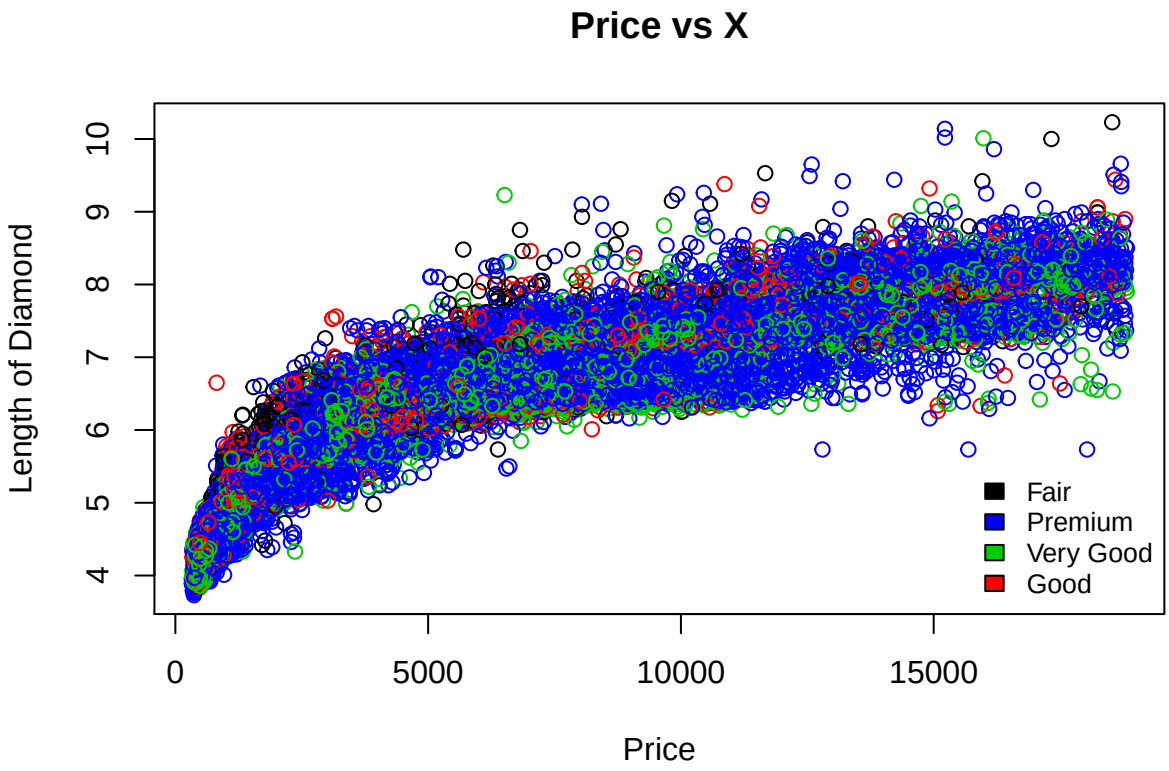
Task 4

Frequencies of Diamonds for various cuts and clarity

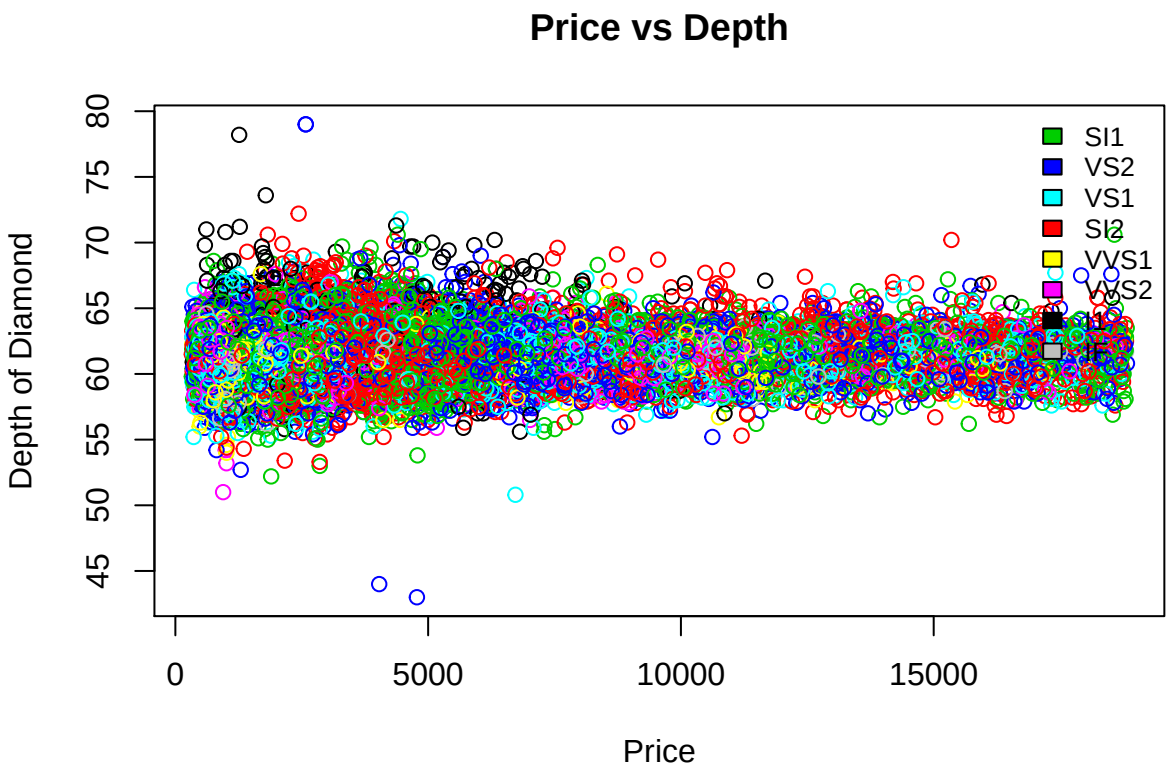
```
##
##      Fair      Good Very Good   Premium
##      1433      4396     10825     31659

##
##    I1   SI2   SI1   VS2   VS1  VVS2  VVS1    IF
##   663  8200 11697 11043  7299  4530  3274  1607
```

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

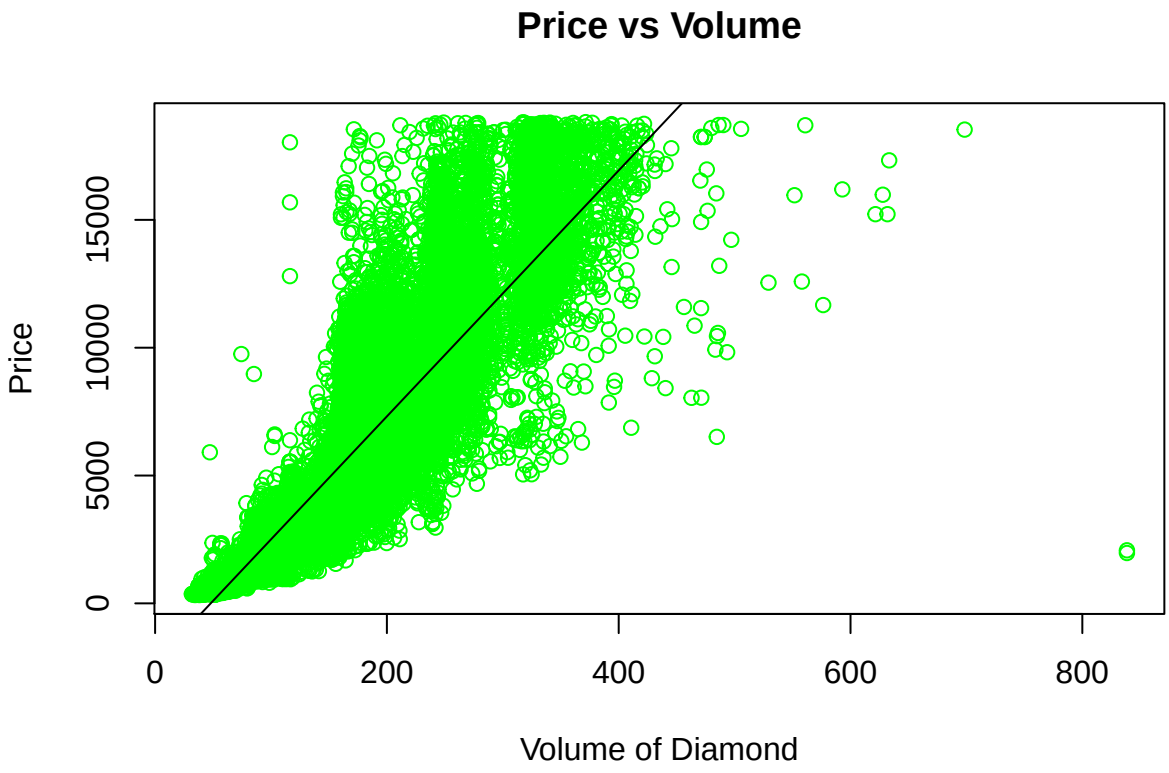
Task 5

Compute Volume variable from x, y, z

```
##      carat               cut        color        clarity
##  Min.   :0.2000   Fair      : 1433   D: 6034   SI1     :11697
##  1st Qu.:0.4000   Good      : 4396   E: 8783   VS2     :11043
##  Median :0.7000   Very Good:10825   F: 8516   SI2     : 8200
##  Mean   :0.8073   Premium   :31659   G:10145   VS1     : 7299
##  3rd Qu.:1.0500                     H: 7452   VVS2    : 4530
##  Max.   :4.9990                     I: 4874   VVS1    : 3274
##                                     J: 2509   (Other): 2270
##      depth           table           price             x
##  Min.   :43.00   Min.   :43.00   Min.   :  326   Min.   : 3.730
##  1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  949   1st Qu.: 4.710
##  Median :61.80   Median :57.00   Median : 2403   Median : 5.700
##  Mean   :61.75   Mean   :57.46   Mean   : 3942   Mean   : 5.733
##  3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5351   3rd Qu.: 6.540
##  Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.230
##
##        y                z             range           volume
##  Min.   : 3.680   Min.   : 1.070   High  : 2498   Min.   : 31.71
##  1st Qu.: 4.720   1st Qu.: 2.910   Medium: 6800   1st Qu.: 65.13
```

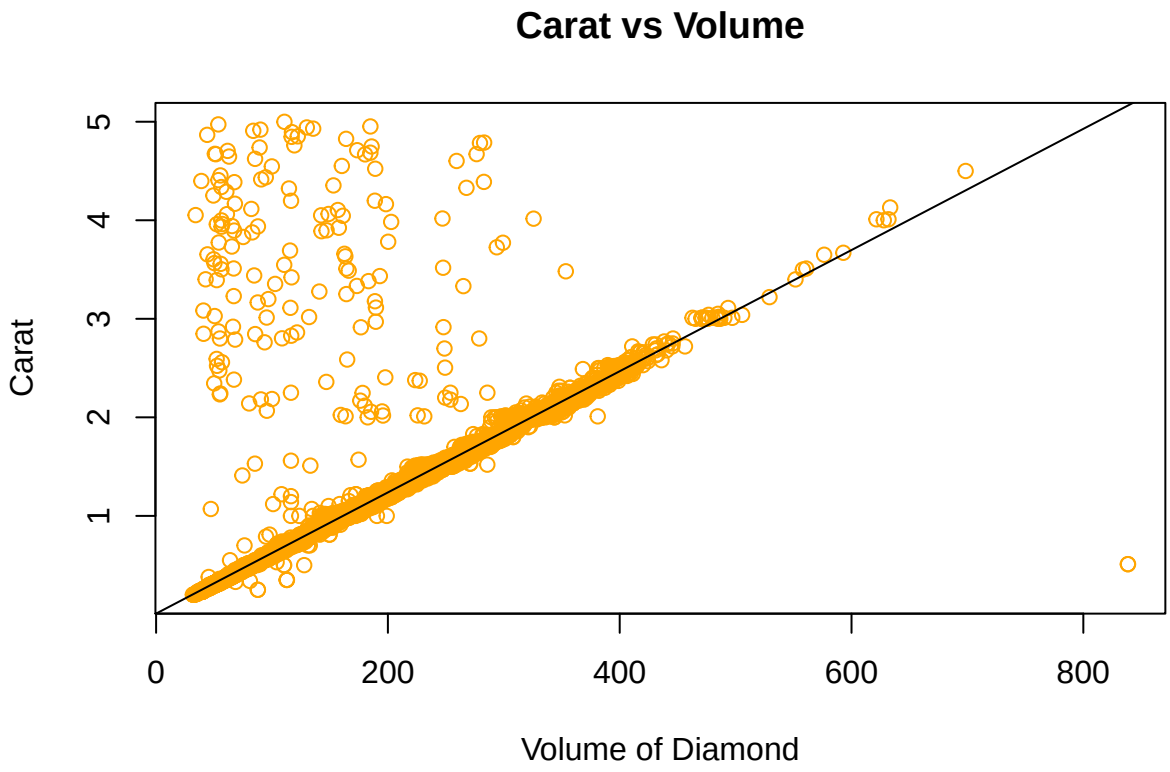
```
##  Median : 5.710   Median : 3.530   Low   :39015   Median :114.89
##  Mean   : 5.736   Mean   : 3.541                  Mean   :129.94
##  3rd Qu.: 6.540   3rd Qu.: 4.040                  3rd Qu.:170.92
##  Max.   :31.800   Max.   :31.800                  Max.   :838.50
##
```

Price vs Volume



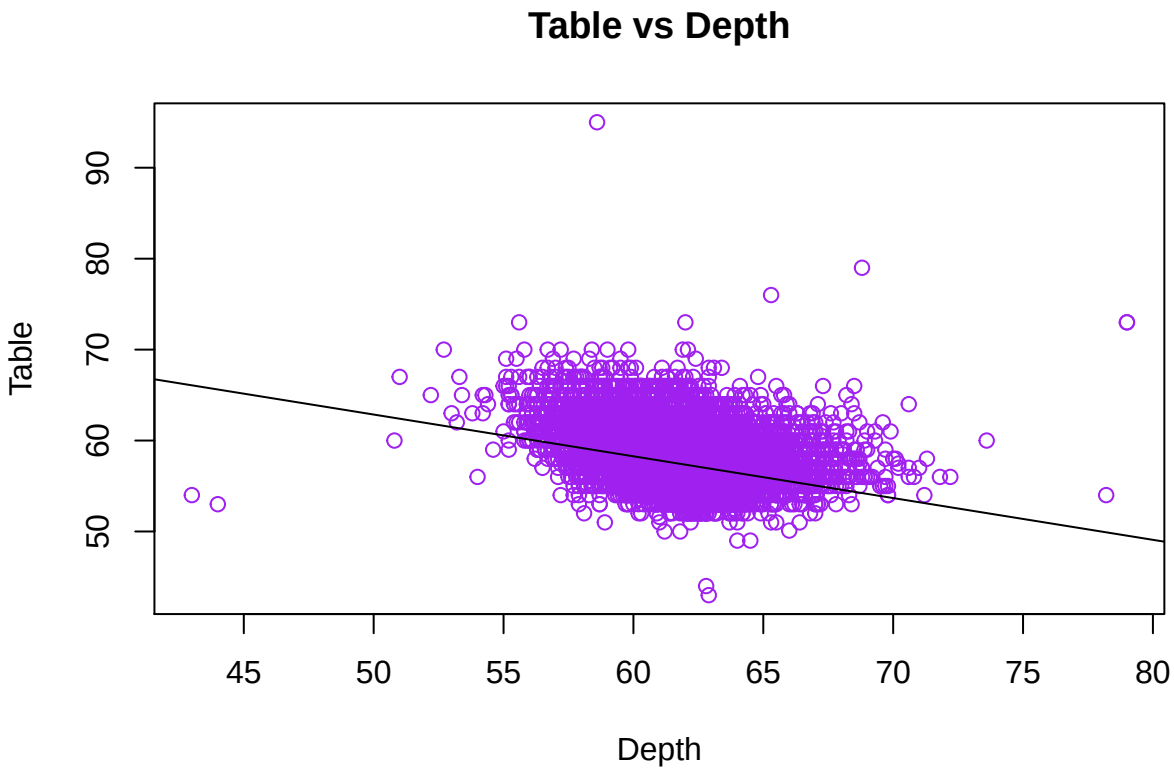
There is an exponential increase in Price with the increase in the volume of the diamond. The regression line shows a good correlation between price and volume.

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
## [1,] 0.1702997 0.1271754 0.1952626 0.1875182 0.1516615 -0.2938471
##         volume
## [1,] 0.1708346
```

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.