Assignment 01

Jawad Adil - 3049429

3/29/2021

Getting the data from CSV file

```
#Import the dataset
DiamondData <- read.csv("C:/Users/jawad adil/Downloads/DiamondData.csv")

#Get a subset of data for testing
s <- sample(nrow(DiamondData), size=10000, replace = FALSE, prob = NULL)
s <- DiamondData[s, ]

#Getting original data again for final report
s<- DiamondData
```

Task 1 - omit/replace NA/incorrect values

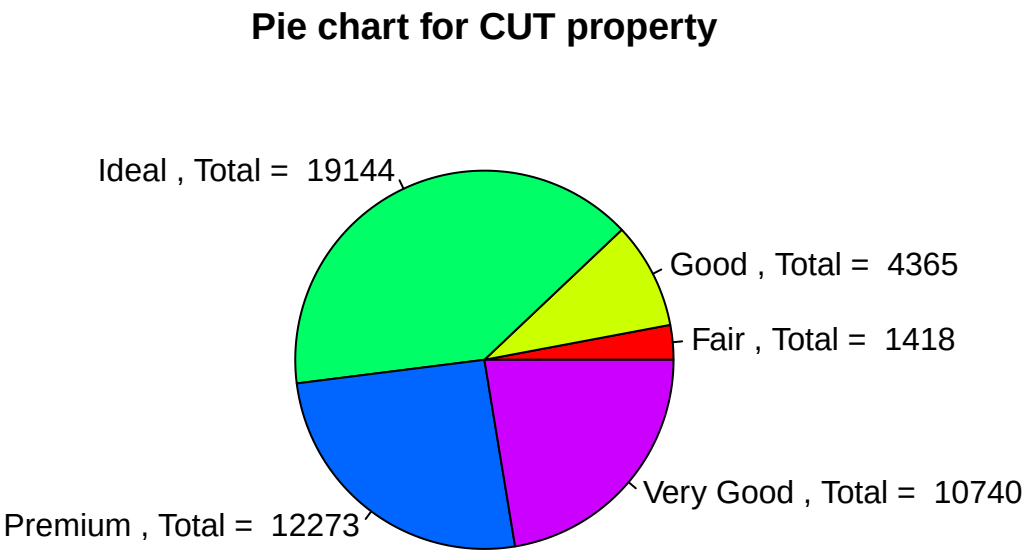
```
#omitting NA values
s<-na.omit(s)

#correcting spelling for cut attribute
s$cut[s$cut == "Very Geod"] <-"Very Good"


#limiting the carat to given range, if the carat is greater than given range,
#set it to maximum value
s$carat<-replace(s$carat,s$carat>5.01,5.01)
```

Task 2 - Get summary of variables and draw pie,bar char, histogram and scatter plot for appropriate variables

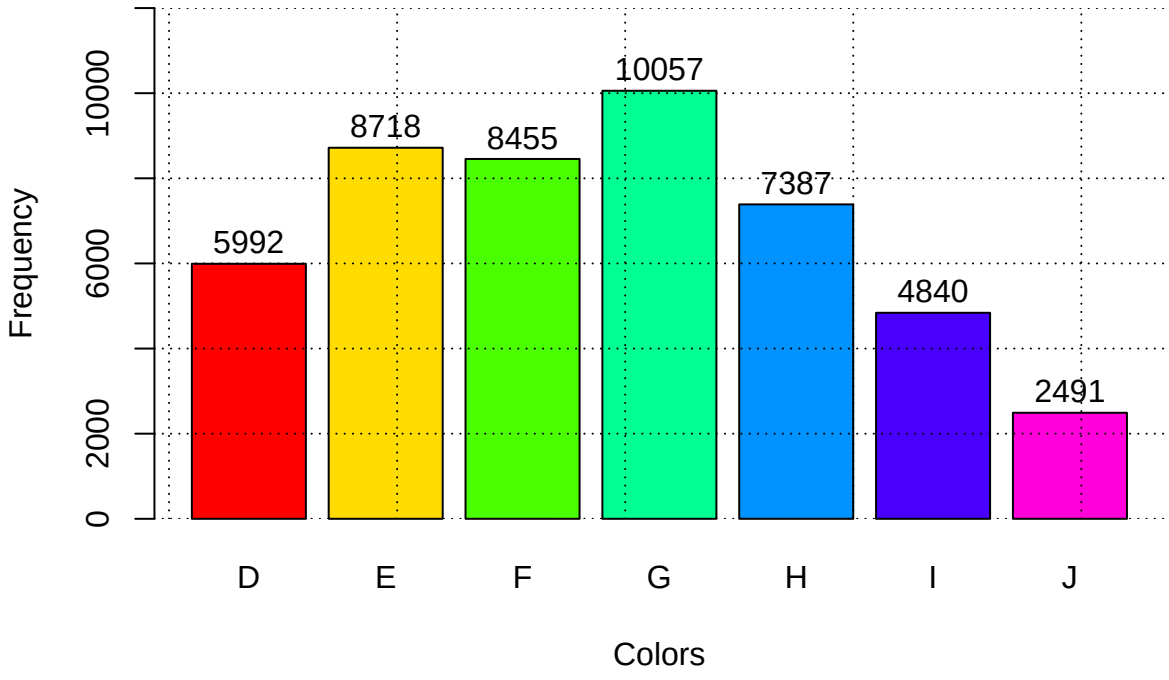
```
#Get the summary of variables
summary(s)
```

```
##      carat            cut                color              clarity
##  Min.   :0.2000   Length:47940       Length:47940       Length:47940
##  1st Qu.:0.4000   Class :character   Class :character   Class :character
##  Median :0.7000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.8112
##  3rd Qu.:1.0500
##  Max.   :5.0100
##      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.: 5345   3rd Qu.: 6.540
##  Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.230
##        y                z
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 4.720   1st Qu.: 2.910
##  Median : 5.710   Median : 3.520
##  Mean   : 5.734   Mean   : 3.539
##  3rd Qu.: 6.540   3rd Qu.: 4.040
##  Max.   :31.800   Max.   :31.800
```

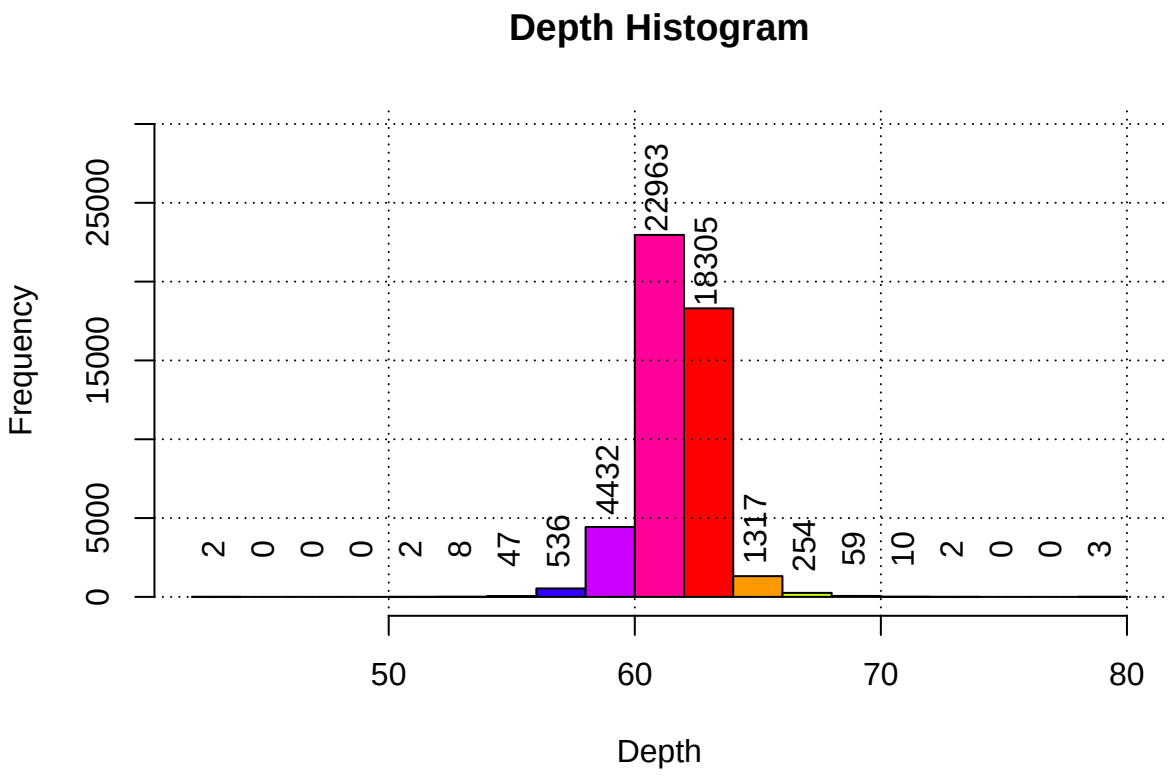
```
#pie chart for cut variable
pie(table(factor(s$cut)),main = "Pie chart for CUT property",labels = paste(levels(factor(s$cut)),
    paste(", Total = ",table(s$cut))),col=rainbow(nrow(table(s$cut))))
```



```r
#bar plot for the color variable
barplot <- barplot(table(factor(s$color)),col=rainbow(nrow(table(s$color))),
                   xlab = "Colors",ylab = "Frequency",ylim=c(0,12000))
#printing counts on the bars of bar chart
text(barplot,table(s$color)+500,paste(table(s$color)) ,cex=1)
grid(col = "black")
```

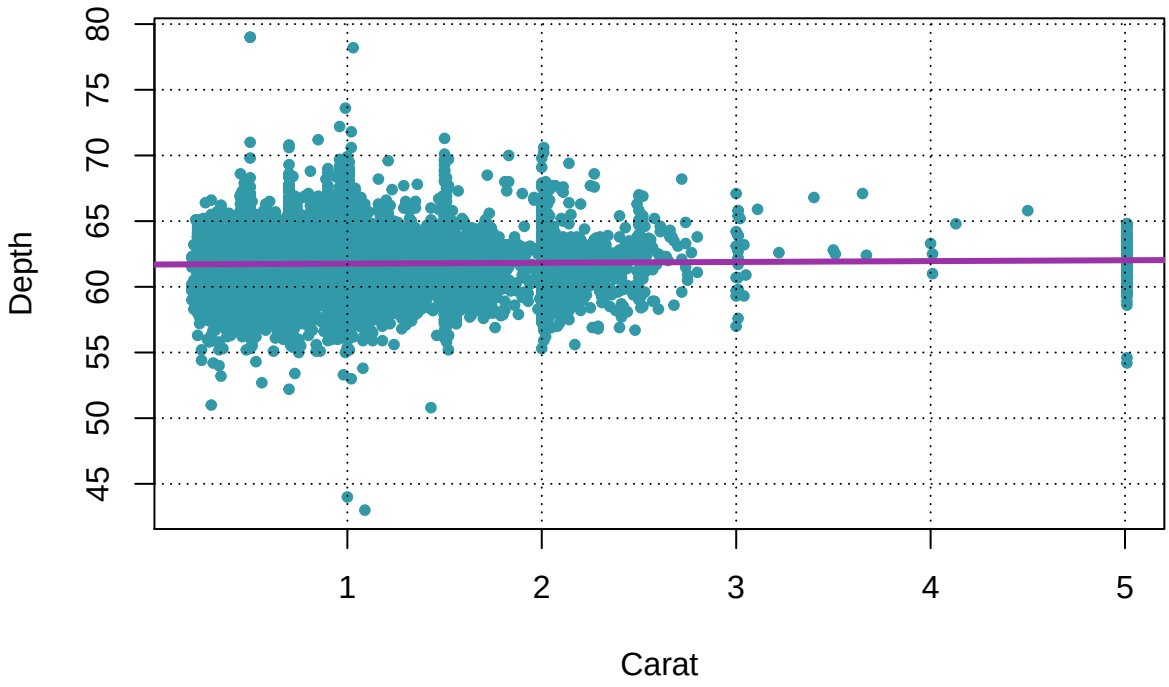


```r
#histogram for depth variable
hist1 <- hist(s$depth,main = "Depth Histogram", ylab = "Frequency",xlab = "Depth",
              ,ylim=c(0,30000),col=rainbow(10))
#printing counts on the histogram
text(hist1$mids+0.8,hist1$counts+3000,labels = hist1$counts,adj=c(0.5, -0.5),srt=90)
grid(col = "black")
```



This histogram is showing normality. Count of each value is shown on the top of bars.

```r
#scatter plot for carat vs depth variables
plot(s$carat,s$depth,xlab = "Carat",ylab = "Depth",pch=20,col="#3299a8")
abline(lm(s$depth~s$carat),col="#9c32a8",lwd="3")
grid(col = "black")
```




```r
#finding correlation
cor(s$carat,s$depth)
```

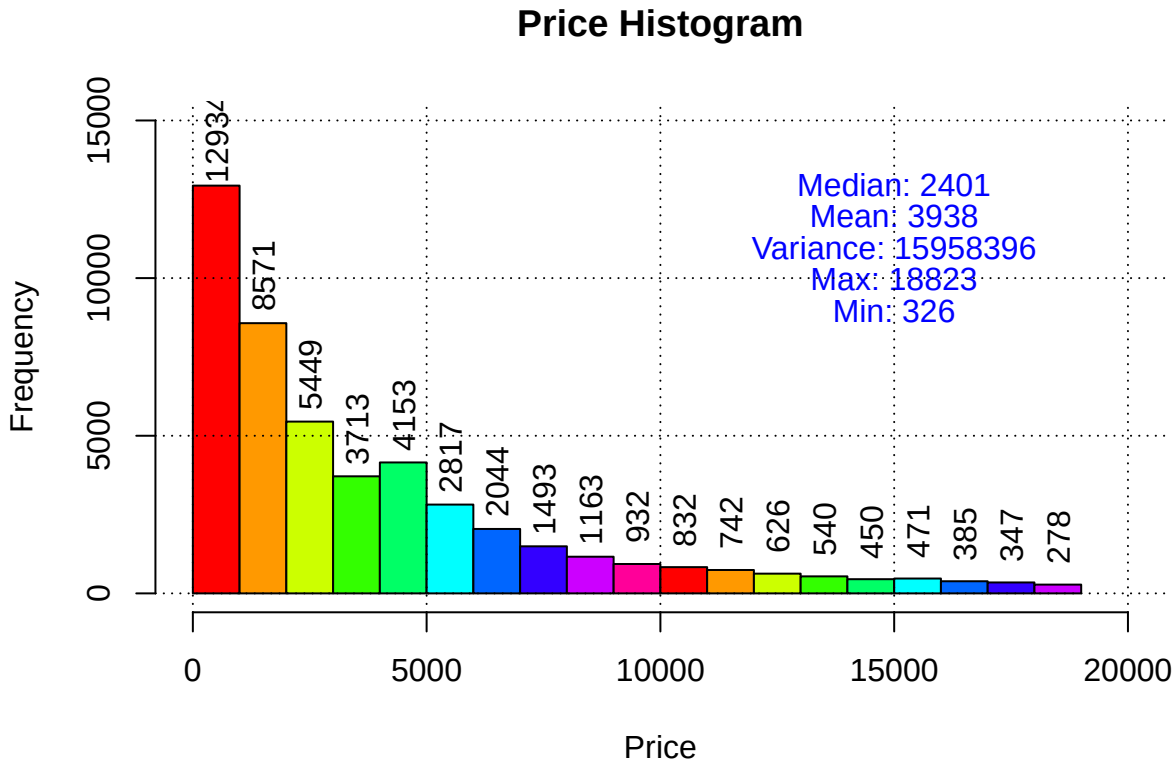
```
## [1] 0.02391426
```

This Graph is showing normality and the relationship among them is very weak. Correlation function also results the same.

Task 3:

3A: histogram of price variable

```
#histogram of price variable
histPrice <- hist(s$price,main = "Price Histogram", ylab = "Frequency",xlab = "Price",
     col=rainbow(10),xlim=c(0,20000),ylim=c(0,15000))
#printing counts on the bars of histogram
text(histPrice$mids+500,histPrice$counts+1500,labels = histPrice$counts,adj=c(0.5, -0.5),srt=90)
#printing the summary on the histogram using text function
text(15000,12000,paste("Median:",summary(s$price)[3]),pos=3,col="Blue")
text(15000,11000,paste("Mean:",as.integer(summary(s$price)[4])),pos=3,col="Blue")
text(15000,10000,paste("Variance:",as.integer(var(s$price))),pos=3,col="Blue")
text(15000,9000,paste("Max:",summary(s$price)[6]),pos=3,col="Blue")
text(15000,8000,paste("Min:",summary(s$price)[1]),pos=3,col="Blue")
grid(col = "black")
```



3B: Group Diamonds by some price ranges and summerize them separately

```
#divide the price variable into 3 groups

#low are those values which are less than 1st Quartile
low <- subset(s,price<=summary(s$price)[2])

#med are those values which are less than 3rd Quartile
med <- subset(s,price<=summary(s$price)[5] & price>summary(s$price)[2])

#all other greater values are high values
high <- subset(s,price > summary(s$price)[5])

#printing the count of low values
nrow(low)
```

```
## [1] 12002
```

```
#summary of low variable values
summary(low)
```

```
##      carat            cut                color              clarity
##  Min.   :0.2000   Length:12002       Length:12002       Length:12002
##  1st Qu.:0.3000   Class :character   Class :character   Class :character
##  Median :0.3200   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.3438
##  3rd Qu.:0.3600
##  Max.   :5.0100
##      depth           table           price             x
##  Min.   :51.00   Min.   :44.00   Min.   :326.0   Min.   :3.730
##  1st Qu.:61.20   1st Qu.:55.00   1st Qu.:573.0   1st Qu.:4.310
##  Median :61.80   Median :57.00   Median :694.0   Median :4.400
##  Mean   :61.76   Mean   :56.97   Mean   :688.3   Mean   :4.433
##  3rd Qu.:62.40   3rd Qu.:58.00   3rd Qu.:810.0   3rd Qu.:4.570
##  Max.   :71.00   Max.   :68.00   Max.   :949.0   Max.   :6.650
##        y               z
##  Min.   :3.680   Min.   :2.240
##  1st Qu.:4.320   1st Qu.:2.670
##  Median :4.410   Median :2.720
##  Mean   :4.447   Mean   :2.742
##  3rd Qu.:4.580   3rd Qu.:2.820
##  Max.   :5.640   Max.   :3.580
```

```r
#printing the count of med values
nrow(med)
```

```
## [1] 23957
```

```r
#summary of med variable values
summary(med)
```

```
##      carat             cut                color              clarity
##  Min.   :0.2400   Length:23957       Length:23957       Length:23957
##  1st Qu.:0.5200   Class :character   Class :character   Class :character
##  Median :0.7000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.7317
##  3rd Qu.:0.9100
##  Max.   :5.0100
##      depth           table           price             x
##  Min.   :43.00   Min.   :43.00   Min.   : 950   Min.   :0.000
##  1st Qu.:61.00   1st Qu.:56.00   1st Qu.:1574   1st Qu.:5.170
##  Median :61.80   Median :57.00   Median :2401   Median :5.700
##  Mean   :61.76   Mean   :57.55   Mean   :2681   Mean   :5.668
##  3rd Qu.:62.60   3rd Qu.:59.00   3rd Qu.:3812   3rd Qu.:6.200
##  Max.   :79.00   Max.   :79.00   Max.   :5345   Max.   :8.110
##        y               z
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 5.170   1st Qu.: 3.190
##  Median : 5.710   Median : 3.530
##  Mean   : 5.669   Mean   : 3.501
##  3rd Qu.: 6.200   3rd Qu.: 3.850
##  Max.   :31.800   Max.   :31.800
```

```r
#printing the count of high values
nrow(high)
```

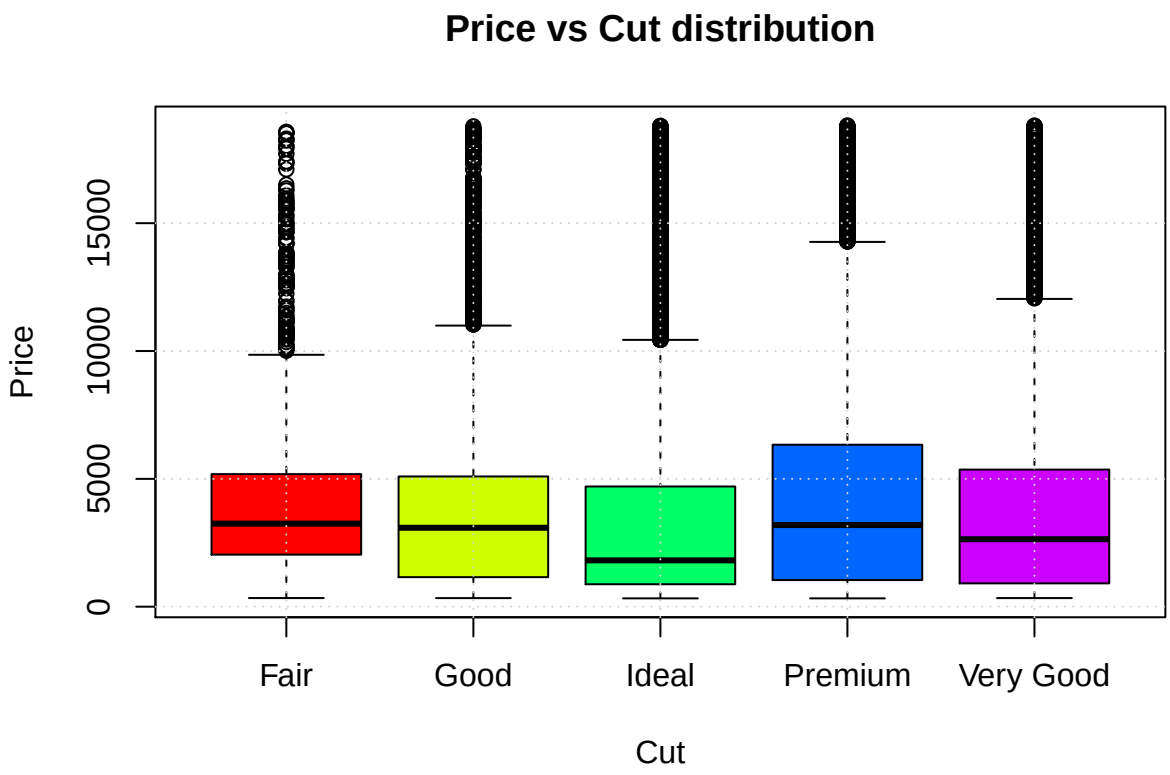
```
## [1] 11981
```

```r
#summary of high variable values
summary(high)
```

```
##      carat           cut                color              clarity
##  Min.   :0.630   Length:11981       Length:11981       Length:11981
##  1st Qu.:1.090   Class :character   Class :character   Class :character
##  Median :1.330   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1.438
##  3rd Qu.:1.590
##  Max.   :5.010
##      depth           table           price             x
##  Min.   :50.80   Min.   :50.00   Min.   : 5346   Min.   : 0.000
##  1st Qu.:61.00   1st Qu.:56.00   1st Qu.: 6608   1st Qu.: 6.630
##  Median :61.90   Median :58.00   Median : 8688   Median : 7.060
##  Mean   :61.72   Mean   :57.79   Mean   : 9709   Mean   : 7.159
##  3rd Qu.:62.60   3rd Qu.:59.00   3rd Qu.:12179   3rd Qu.: 7.530
##  Max.   :70.60   Max.   :95.00   Max.   :18823   Max.   :10.230
##        y                z
##  Min.   : 0.000   Min.   :0.000
##  1st Qu.: 6.640   1st Qu.:4.080
##  Median : 7.050   Median :4.350
##  Mean   : 7.153   Mean   :4.413
##  3rd Qu.: 7.520   3rd Qu.:4.640
##  Max.   :10.160   Max.   :6.720
```

3C: Exploring different cut types using boxplot

```
#Price vs Cut box plot
boxplot(s$price~s$cut,main="Price vs Cut distribution",xlab = "Cut",ylab = "Price",
        col=rainbow(nrow(table(s$cut))))
grid()
```



3D: Finding relation of attributes with price

```
#separate numeric data from actual dataset so we can directly use cor() function
numeric_data <- data.frame(s$price,s$carat,s$depth,s$table,s$x,s$y,s$z)

#finding the correlation between variables.
cor(numeric_data)
```

```
##             s.price    s.carat     s.depth    s.table        s.x        s.y
## s.price  1.00000000 0.82644437 -0.01337194  0.1277515  0.8845208  0.8810516
## s.carat  0.82644437 1.00000000  0.02391426  0.1633404  0.8729700  0.8672019
## s.depth -0.01337194 0.02391426  1.00000000 -0.2950651 -0.0278515 -0.0308265
## s.table  0.12775151 0.16334035 -0.29506506  1.0000000  0.1952386  0.1874497
## s.x      0.88452082 0.87296998 -0.02785150  0.1952386  1.0000000  0.9923832
## s.y      0.88105158 0.86720194 -0.03082650  0.1874497  0.9923832  1.0000000
## s.z      0.85918803 0.85069073  0.09243840  0.1506000  0.9684744  0.9634559
##               s.z
## s.price 0.8591880
## s.carat 0.8506907
## s.depth 0.0924384
## s.table 0.1506000
## s.x     0.9684744
## s.y     0.9634559
## s.z     1.0000000
```

Price is correlated most with the X variable, Y is the 2nd high correlated variable with price and Z is the 3rd most correlated variable. Simply, the order of correlation is:

Most Strong:

X variable

Mid Strong:

Y variable

Least Strong:

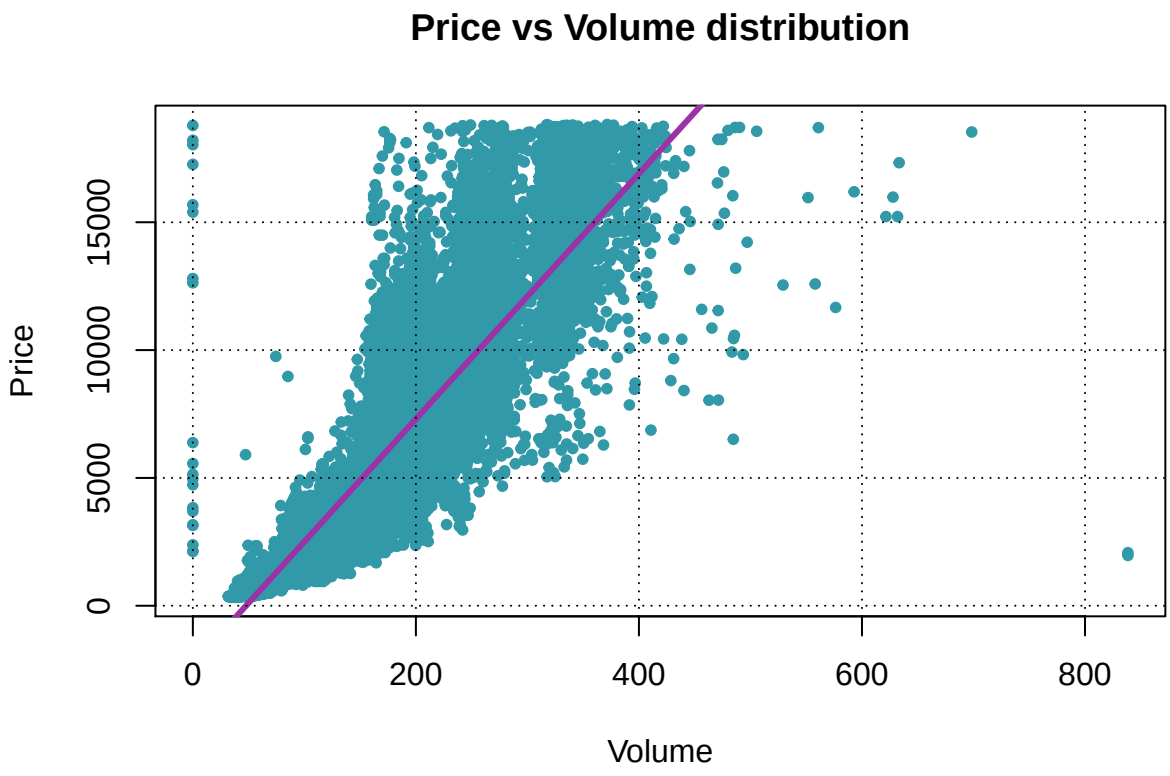
Z variable

All other variables have less strong relationship with price than these variables.

Task 4:

4A: compute the volume & plot price vs volume graph

```
#separating x, y, z variables from dataset for better understanding
x <- s[,8]
y <- s[,9]
z <- s[,10]

#calculating volume variable
volume <- x*y*z

#binding old dataset and volume and storing them in new variable so old one
#remain unchanged
newData=cbind(s,volume)

#getting price from new dataset
price=newData[,7]

#ploting graph between price & volume
plot(volume,price,xlab = "Volume",ylab = "Price", main = "Price vs Volume distribution",
     pch=20,col="#3299a8")
abline(lm(price~volume),col="#9c32a8",lwd="3")
grid(col="black")
```



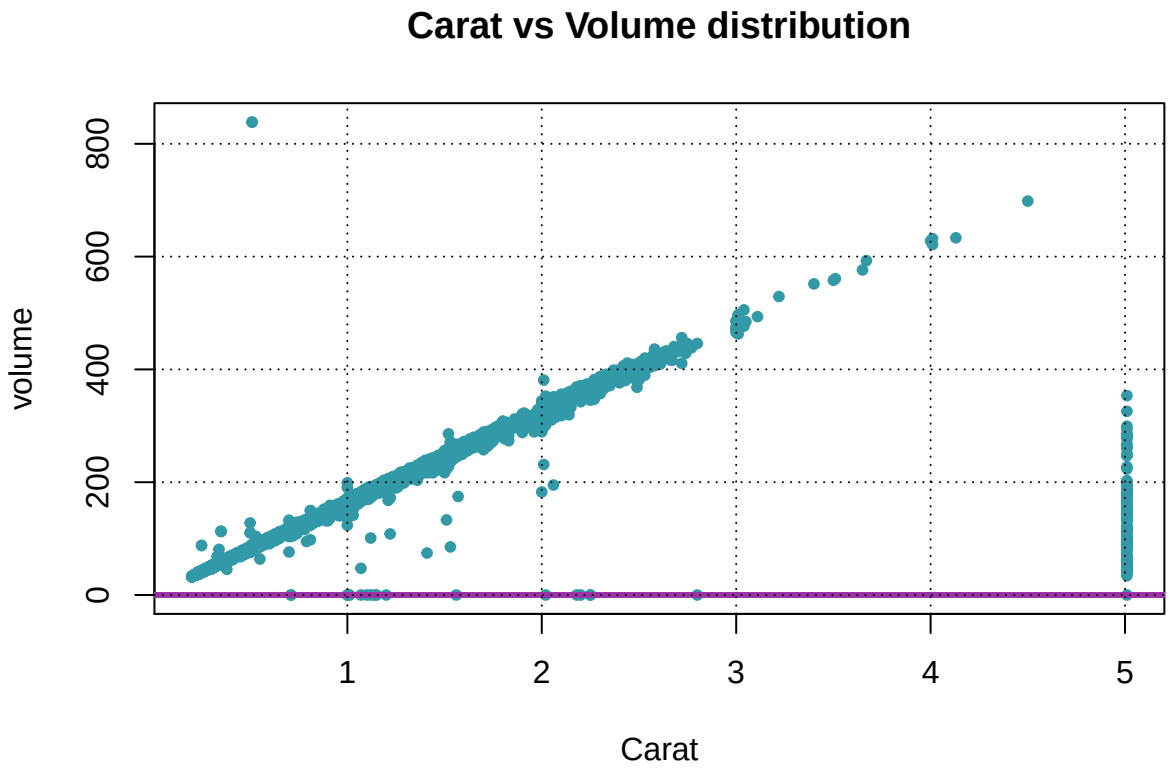
```
#finding correlation between price and volume
cor(price,volume)
```

```
## [1] 0.920117
```

The graph shows direct and very strong relationship between price and volume, strength is also proven by correlation function.

4B: Finding correlation between carat and volume & plotting them with regression line

```
#plotting the graph between carat and volume
plot(newData$carat,volume,xlab = "Carat",main = "Carat vs Volume distribution",
     pch=20,col="#3299a8")

#drawing a regression line on graph
abline(lm(newData$carat ~ volume),col="#9c32a8",lwd="3")
grid(col="black")
```




```
#finding correlation between volume and carat
cor(newData$carat, volume)
```

```
## [1] 0.8903231
```

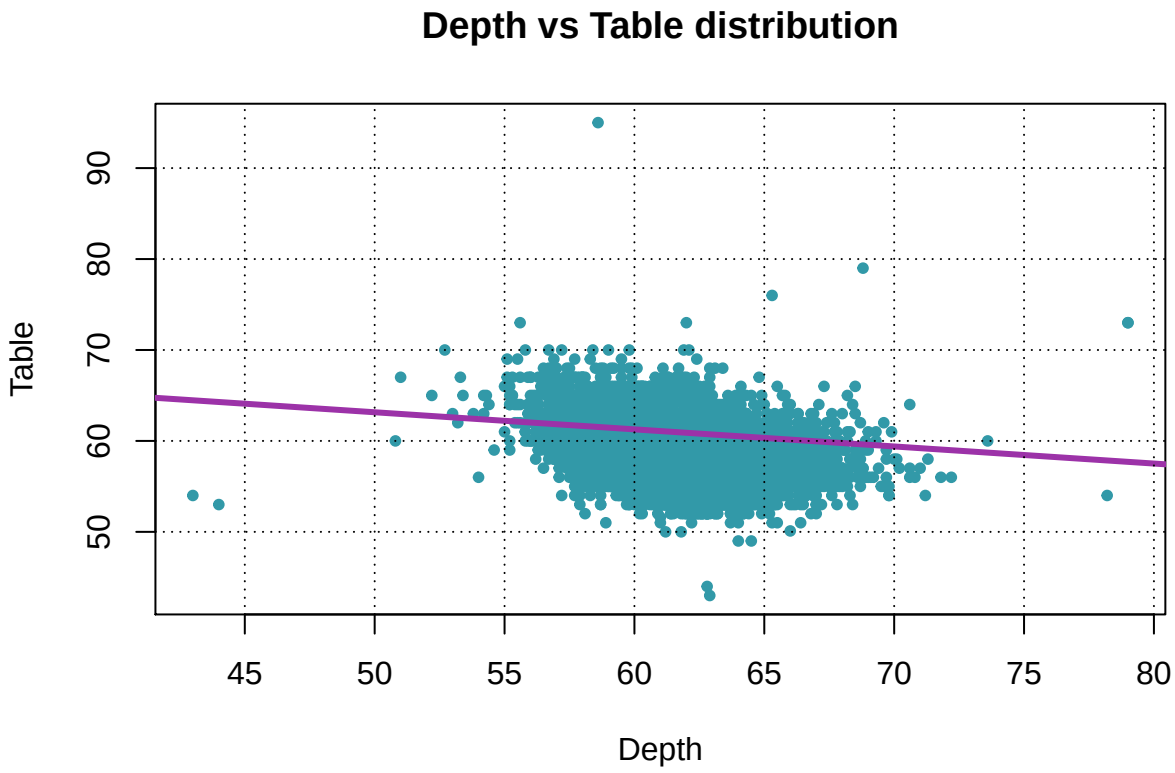
The graph shoes the strong correlation between volume and carat which is also proved by cor() function.

4C: Exploring relationship of table and depth variables:

```
#finding correlation between table and depth variables
cor(newData$depth,newData$table)
```

```
## [1] -0.2950651
```

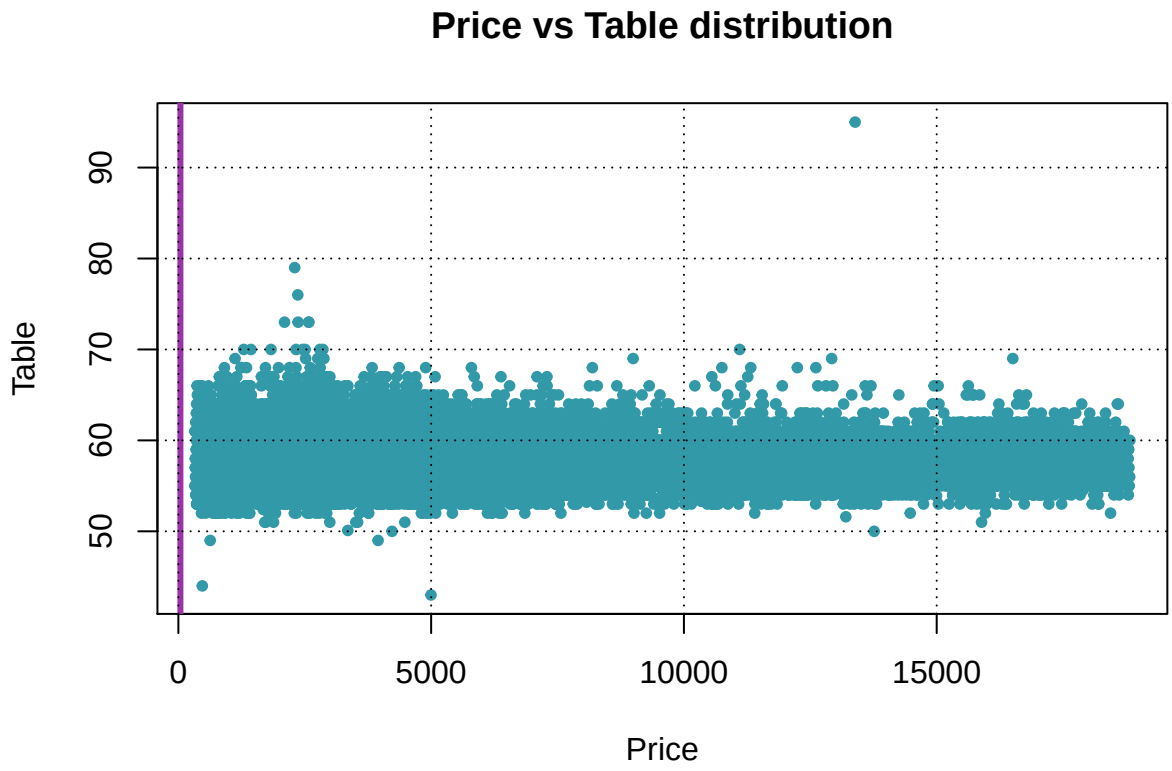
```
#plotting the relationship graph
plot(newData$depth,newData$table,xlab = "Depth",ylab = "Table",
     main = "Depth vs Table distribution",pch=20,col="#3299a8")

#drawing regression line
abline(lm(newData$depth~newData$table),col="#9c32a8",lwd="3")
grid(col = "black")
```



The graph shows negative relationship which is also proven by the cor() function.

4D: Exploring relationship of table and rest of the variables:

```
#ploting relationship between price and table
plot(newData$price,newData$table,xlab = "Price",ylab = "Table",
     main = "Price vs Table distribution",pch=20,col="#3299a8")

#drawing regression line
abline(lm(newData$price~newData$table),col="#9c32a8",lwd="3")
grid(col = "black")
```

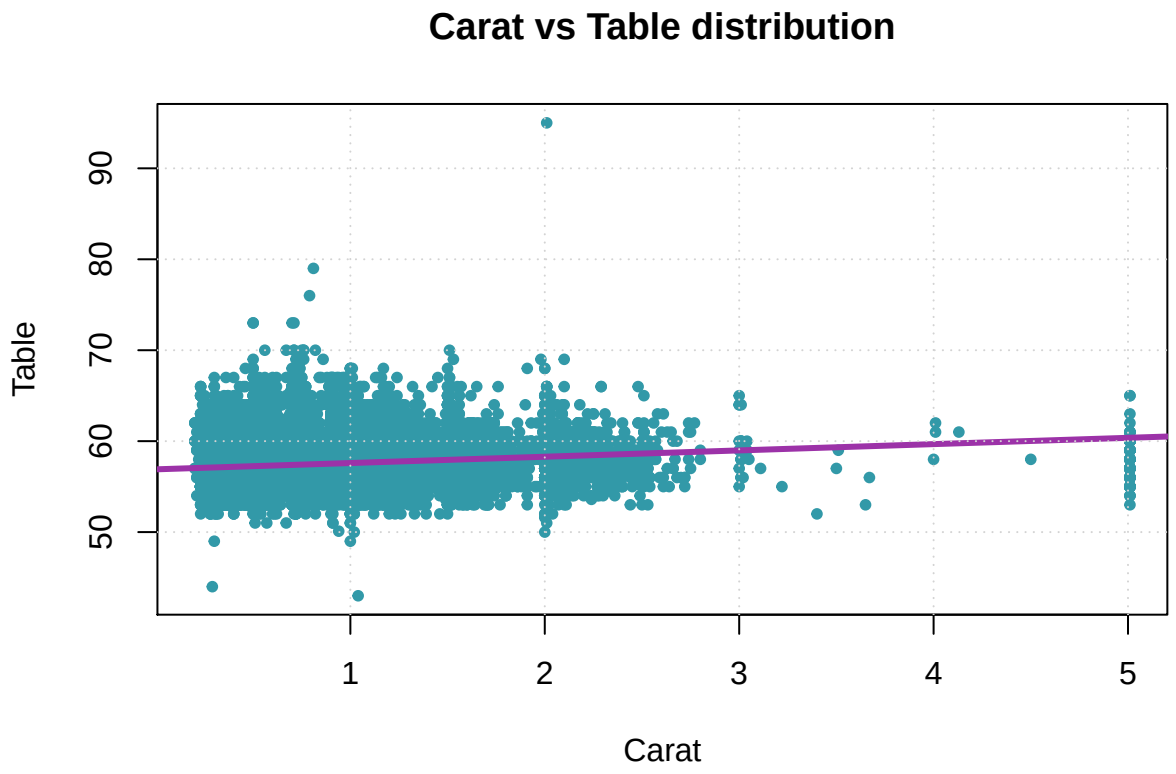



```
#finding correlation between price and table
cor(newData$table,newData$price)
```

```
## [1] 0.1277515
```

The graph between price and table shows weak but positive relationship which is also proven by the cor() function.

```
#plotting carat vs table relationship
plot(newData$carat,newData$table,xlab = "Carat",ylab = "Table"
     ,main = "Carat vs Table distribution",pch=20,col="#3299a8")

#plotting regression line
abline(lm(newData$table~newData$carat),col="#9c32a8",lwd="3")
grid()
```

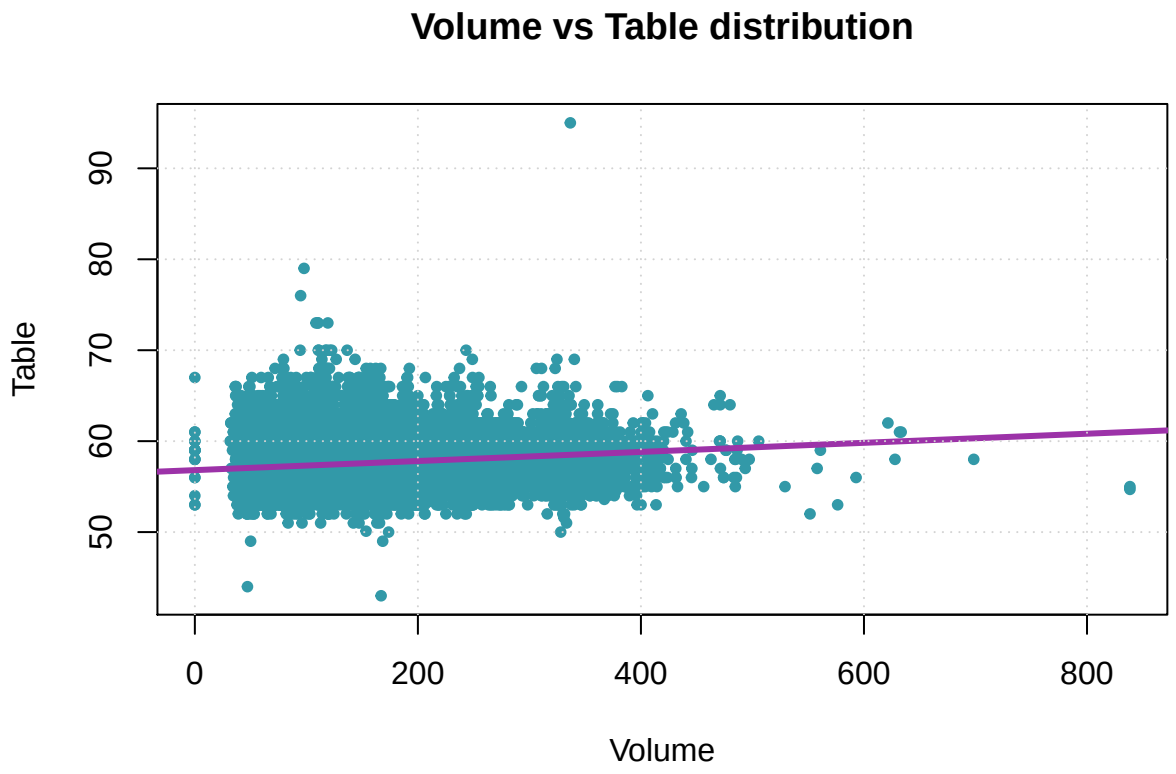



```
#finding correlation between table and carat
cor(newData$table,newData$carat)
```

```
## [1] 0.1633404
```

The Graph show weak but positive relationship between carat and table which is also proven by correlation function. Also, this graph is showing normality.

```
#plotting carat vs table relationship
plot(newData$volume,newData$table,xlab = "Volume",ylab = "Table"
     ,main = "Volume vs Table distribution",pch=20,col="#3299a8")

#plotting regression line
abline(lm(newData$table~newData$volume),col="#9c32a8",lwd="3")
grid()
```




```
#finding correlation between table and carat
cor(newData$table,newData$volume)
```

```
## [1] 0.1711106
```

The Graph between volume & table also shows a weak but positive relation which is also proven by correlation function. This graph is also showing normality as the line is going slighlty upward as we move right from left.