Assignment: Hypothesis testing on Diamonds data

```
> library(ggplot2)
> data(diamonds)
> summary(diamonds)
     carat               cut         color        clarity          depth      
 Min.   :0.2000   Fair      : 1610   D: 6775    SI1    :13065    Min.   :43.00  
 1st Qu.:0.4000   Good      : 4906   E: 9797    VS2    :12258    1st Qu.:61.00  
 Median :0.7000   Very Good:12082   F: 9542    SI2    : 9194    Median :61.80  
 Mean   :0.7979   Premium   :13791   G:11292    VS1    : 8171    Mean   :61.75  
 3rd Qu.:1.0400   Ideal     :21551   H: 8304    VVS2   : 5066    3rd Qu.:62.50  
 Max.   :5.0100                     I: 5422    VVS1   : 3655    Max.   :79.00  
                                    J: 2808    (Other): 2531                   

     table           price             x                y         
 Min.   :43.00    Min.   :  326    Min.   : 0.000   Min.   : 0.000  
 1st Qu.:56.00    1st Qu.:  950    1st Qu.: 4.710   1st Qu.: 4.720  
 Median :57.00    Median : 2401    Median : 5.700   Median : 5.710  
 Mean   :57.46    Mean   : 3933    Mean   : 5.731   Mean   : 5.735  
 3rd Qu.:59.00    3rd Qu.: 5324    3rd Qu.: 6.540   3rd Qu.: 6.540  
 Max.   :95.00    Max.   :18823    Max.   :10.740   Max.   :58.900  

       z         
 Min.   : 0.000  
 1st Qu.: 2.910  
 Median : 3.530  
 Mean   : 3.539  
 3rd Qu.: 4.040  
 Max.   :31.800  


> ####Q.1#Does the mean price differ between diamonds with clarity = VVS1 and clarity =
IF?####
> a1=diamonds[diamonds$clarity=="VVS1",]$price
> a2=diamonds[diamonds$clarity=="IF",]$price
> t.test(a1,a2,conf.level =0.95)

        Welch Two Sample t-test

data:  a1 and a2
t = -3.169, df = 3091.6, p-value = 0.001545
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -553.1600 -130.2889
sample estimates:
mean of x mean of y 
 2523.115  2864.839 
```

Here, P-value: 0.001545 < Level of significance: 0.05
Hence, we reject H0.
Therefore we can say that mean price differ between diamonds with clarity = VVS1 and clarity = IF.

```
> ####Q.2#Does the mean price differ between diamonds with cut = Fair and cut = Ideal?##
> b1=diamonds[diamonds$cut=="Fair",]$price
> b2=diamonds[diamonds$cut=="Ideal",]$price
> t.test(b1,b2,conf.level =0.95)

      Welch Two Sample t-test

data:  b1 and b2
t = 9.7484, df = 1894.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  719.9065 1082.5251
sample estimates:
mean of x mean of y
 4358.758  3457.542
```

Here, P-value: 2.2e-16 < Level of significance: 0.05
Hence, we reject H0.
Therefore we can say that mean price differ between diamonds with cut = Fair and cut = Ideal.

```
> ####Q.3#Does the mean price differ between diamonds with color = D and color = J?#####
> c1=diamonds[diamonds$color=="D",]$price
> c2=diamonds[diamonds$color=="J",]$price
> t.test(c1,c2,conf.level =0.95)

      Welch Two Sample t-test

data:  c1 and c2
t = -23.121, df = 4197.9, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2336.496 -1971.232
sample estimates:
mean of x mean of y
 3169.954  5323.818
```

Here, P-value: 2.2e-16 < Level of significance: 0.05
Hence, we reject H0.
Therefore we can say that mean price differ between diamonds with color = D and color = J.

```
> ####Q.4#Is there a relationship between the price of a diamond and the width of its
table?#####
> cor.test((diamonds$price),(diamonds$table))

      Pearson's product-moment correlation

data:  (diamonds$price) and (diamonds$table)
t = 29.768, df = 53938, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1188223 0.1354277
sample estimates:
      cor
0.1271339
```

Here, P-value: 2.2e-16 < Level of significance: 0.05
Hence, we reject H0.
Therefore we can say that there is some relationship between the price of a diamond and the width of its table.

```
> ####Q.5#Is there a relationship between the price of a diamond and its carat
weight?#####
> cor.test((diamonds$price),(diamonds$carat))

      Pearson's product-moment correlation

data:  (diamonds$price) and (diamonds$carat)
t = 551.41, df = 53938, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9203098 0.9228530
sample estimates:
      cor
0.9215913
```

Here, P-value: 2.2e-16 < Level of significance: 0.05
Hence, we reject H0.
Therefore we can say that there is relationship between the price of a diamond and its carat weight.

```
> ####Q.6#Is cut and color are associated with each other?####
> #e=ftable(diamonds$cut,diamonds$color)
> chisq.test((diamonds$cut),(diamonds$color))

        Pearson's Chi-squared test

data:  (diamonds$cut) and (diamonds$color)
X-squared = 310.32, df = 24, p-value < 2.2e-16
```

Here, P-value: 2.2e-16 < Level of significance: 0.05
Hence, we reject H0.
Therefore we can say that cut and color are associated with each other.

```
> ####Q.7#Is clarity and cut are associated with each other?#####
> chisq.test((diamonds$clarity),(diamonds$cut))

        Pearson's Chi-squared test

data:  (diamonds$clarity) and (diamonds$cut)
X-squared = 4391.4, df = 28, p-value < 2.2e-16
```

Here, P-value: 2.2e-16 < Level of significance: 0.05
Hence, we reject H0.
Therefore we can say that clarity and cut are associated with each other.

```
> ####Q.8#Is Clarity and color are associated with each other?#####
> chisq.test((diamonds$clarity),(diamonds$color))

        Pearson's Chi-squared test

data:  (diamonds$clarity) and (diamonds$color)
X-squared = 2047.1, df = 42, p-value < 2.2e-16
```

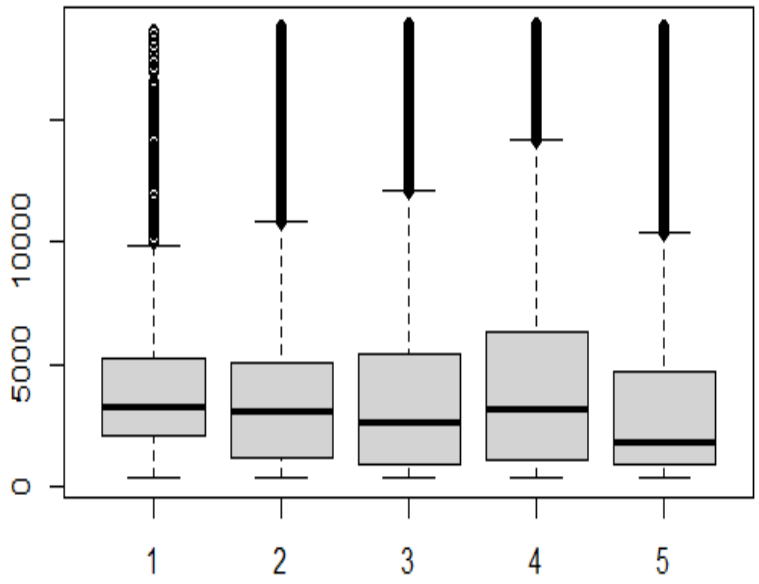
Here, P-value: 2.2e-16 < Level of significance: 0.05
Hence, we reject H0.
Therefore we can say that Clarity and color are associated with each other.

```
> ####Q.9#Typical diamonds of which cut have the highest depth? On average, does depth
increase or decrease as cut grade increase or decrease?
> f=aov(diamonds$depth~diamonds$cut)
> summary(f)
                Df Sum Sq Mean Sq F value Pr(>F)
diamonds$cut     4  13656    3414    1897 <2e-16 ***
Residuals    53935  97048       2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(f)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = diamonds$depth ~ diamonds$cut)

$`diamonds$cut`
                         diff        lwr         upr p adj
Good-Fair          -1.6757985 -1.7808932 -1.57070385     0
Very Good-Fair     -2.2234019 -2.3204792 -2.12632456     0
Premium-Fair       -2.7770044 -2.8733719 -2.68063694     0
Ideal-Fair         -2.3322761 -2.4268124 -2.23773972     0
Very Good-Good     -0.5476034 -0.6095482 -0.48565862     0
Premium-Good       -1.1012059 -1.1620322 -1.04037965     0
Ideal-Good         -0.6564776 -0.7143590 -0.59859608     0
Premium-Very Good -0.5536025 -0.5991981 -0.50800691     0
Ideal-Very Good    -0.1088742 -0.1504601 -0.06728823     0
Ideal-Premium       0.4447283  0.4048276  0.48462906     0
```

Here we can say that depth increase or decrease as cut grade increase or decrease.

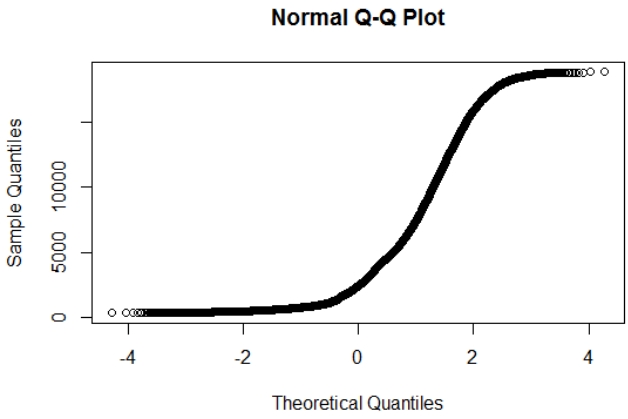
####Q.10#Compare the distribution of price for the different cuts. Does anything seem unusual? Describe.

```
> b1=diamonds[diamonds$cut=="Fair",]$price
> b2=diamonds[diamonds$cut=="Good",]$price
> b3=diamonds[diamonds$cut=="Very Good",]$price
> b4=diamonds[diamonds$cut=="Premium",]$price
> b5=diamonds[diamonds$cut=="Ideal",]$price
> boxplot(b1,b2,b3,b4,b5)
```



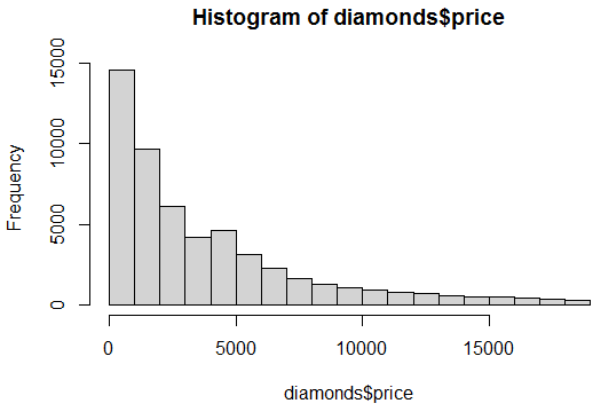
From the Boxplot we can see that the prices for cuts are gradually increasing. According to that we expect that the price of “Ideal cut” diamonds should have maximum price among all. But we can see that the prices for Ideal cut diamonds are suddenly getting low. This is unusual in this dataset.

```
> ####Q.11#Check whether the price of diamond is normalydistributed or not?
> a=qqnorm(diamonds$price)
> qqplot(a)
> hist(diamonds$price)
```



If data is from normal distribution then we should see the points forming a line that's roughly straight.

Here, no such straight line is observed. Therefore, we can say that the data is not normal.



Here, we can see that data is positively skewed.