```r
#Importing necessary libraries
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
#Loading the dataset
mkt <- read.csv('marketing_campaign.csv',sep = ';')
head(mkt)
```

```
##     ID Year_Birth  Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524       1957 Graduation         Single  58138       0        0  2012-09-04
## 2 2174       1954 Graduation         Single  46344       1        1  2014-03-08
## 3 4141       1965 Graduation       Together  71613       0        0  2013-08-21
## 4 6182       1984 Graduation       Together  26646       1        0  2014-02-10
## 5 5324       1981        PhD        Married  58293       1        0  2014-01-19
## 6 7446       1967     Master       Together  62513       0        1  2013-09-09
##   Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1      58      635        88             546             172               88
## 2      38       11         1               6               2                1
## 3      26      426        49             127             111               21
## 4      26       11         4              20              10                3
## 5      94      173        43             118              46               27
## 6      16      520        42              98               0               42
##   MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases
## 1           88                 3               8                  10
## 2            6                 2               1                   1
## 3           42                 1               8                   2
## 4            5                 2               2                   0
## 5           15                 5               5                   3
## 6           14                 2               6                   4
##   NumStorePurchases NumWebVisitsMonth AcceptedCmp3 AcceptedCmp4 AcceptedCmp5
## 1                 4                 7            0            0            0
## 2                 2                 5            0            0            0
## 3                10                 4            0            0            0
## 4                 4                 6            0            0            0
## 5                 6                 5            0            0            0
## 6                10                 6            0            0            0
##   AcceptedCmp1 AcceptedCmp2 Complain Z_CostContact Z_Revenue Response
## 1            0            0        0             3        11        1
## 2            0            0        0             3        11        0
## 3            0            0        0             3        11        0
## 4            0            0        0             3        11        0
```

```
## 5            0            0        0            3        11        0
## 6            0            0        0            3        11        0
```

```
#colnames
names(mkt)
```

```
##  [1] "ID"                 "Year_Birth"          "Education"
##  [4] "Marital_Status"     "Income"              "Kidhome"
##  [7] "Teenhome"           "Dt_Customer"         "Recency"
## [10] "MntWines"           "MntFruits"           "MntMeatProducts"
## [13] "MntFishProducts"    "MntSweetProducts"    "MntGoldProds"
## [16] "NumDealsPurchases"  "NumWebPurchases"     "NumCatalogPurchases"
## [19] "NumStorePurchases"  "NumWebVisitsMonth"   "AcceptedCmp3"
## [22] "AcceptedCmp4"       "AcceptedCmp5"        "AcceptedCmp1"
## [25] "AcceptedCmp2"       "Complain"            "Z_CostContact"
## [28] "Z_Revenue"          "Response"
```

```
#Basic descriptive statistics
summary(mkt)
```

```
##        ID          Year_Birth    Education         Marital_Status
##  Min.   :    0   Min.   :1893   Length:2240        Length:2240
##  1st Qu.: 2828   1st Qu.:1959   Class :character   Class :character
##  Median : 5458   Median :1970   Mode  :character   Mode  :character
##  Mean   : 5592   Mean   :1969
##  3rd Qu.: 8428   3rd Qu.:1977
##  Max.   :11191   Max.   :1996
##
##      Income          Kidhome          Teenhome      Dt_Customer
##  Min.   :  1730   Min.   :0.0000   Min.   :0.0000   Length:2240
##  1st Qu.: 35303   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median : 51382   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   : 52247   Mean   :0.4442   Mean   :0.5062
##  3rd Qu.: 68522   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :666666   Max.   :2.0000   Max.   :2.0000
##  NA's   :24
##     Recency         MntWines         MntFruits     MntMeatProducts
##  Min.   : 0.00   Min.   :   0.00   Min.   :  0.0   Min.   :   0.0
##  1st Qu.:24.00   1st Qu.:  23.75   1st Qu.:  1.0   1st Qu.:  16.0
##  Median :49.00   Median : 173.50   Median :  8.0   Median :  67.0
##  Mean   :49.11   Mean   : 303.94   Mean   : 26.3   Mean   : 166.9
##  3rd Qu.:74.00   3rd Qu.: 504.25   3rd Qu.: 33.0   3rd Qu.: 232.0
##  Max.   :99.00   Max.   :1493.00   Max.   :199.0   Max.   :1725.0
##
##  MntFishProducts  MntSweetProducts  MntGoldProds    NumDealsPurchases
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  3.00   1st Qu.:  1.00   1st Qu.:  9.00   1st Qu.: 1.000
##  Median : 12.00   Median :  8.00   Median : 24.00   Median : 2.000
##  Mean   : 37.53   Mean   : 27.06   Mean   : 44.02   Mean   : 2.325
##  3rd Qu.: 50.00   3rd Qu.: 33.00   3rd Qu.: 56.00   3rd Qu.: 3.000
##  Max.   :259.00   Max.   :263.00   Max.   :362.00   Max.   :15.000
##
##  NumWebPurchases  NumCatalogPurchases NumStorePurchases NumWebVisitsMonth
```

```
##  Min.   : 0.000   Min.   : 0.000      Min.   : 0.00     Min.   : 0.000
##  1st Qu.: 2.000   1st Qu.: 0.000      1st Qu.: 3.00     1st Qu.: 3.000
##  Median : 4.000   Median : 2.000      Median : 5.00     Median : 6.000
##  Mean   : 4.085   Mean   : 2.662      Mean   : 5.79     Mean   : 5.317
##  3rd Qu.: 6.000   3rd Qu.: 4.000      3rd Qu.: 8.00     3rd Qu.: 7.000
##  Max.   :27.000   Max.   :28.000      Max.   :13.00     Max.   :20.000
##
##   AcceptedCmp3      AcceptedCmp4      AcceptedCmp5      AcceptedCmp1
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.07277   Mean   :0.07455   Mean   :0.07277   Mean   :0.06429
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##   AcceptedCmp2        Complain        Z_CostContact   Z_Revenue
##  Min.   :0.00000   Min.   :0.000000   Min.   :3     Min.   :11
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:3     1st Qu.:11
##  Median :0.00000   Median :0.000000   Median :3     Median :11
##  Mean   :0.01339   Mean   :0.009375   Mean   :3     Mean   :11
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:3     3rd Qu.:11
##  Max.   :1.00000   Max.   :1.000000   Max.   :3     Max.   :11
##
##     Response
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1491
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

```
#Check for missing values
colSums(is.na(mkt))
```

```
##                  ID          Year_Birth           Education      Marital_Status
##                   0                   0                   0                   0
##              Income             Kidhome            Teenhome         Dt_Customer
##                  24                   0                   0                   0
##             Recency            MntWines           MntFruits     MntMeatProducts
##                   0                   0                   0                   0
##     MntFishProducts    MntSweetProducts        MntGoldProds   NumDealsPurchases
##                   0                   0                   0                   0
##     NumWebPurchases NumCatalogPurchases   NumStorePurchases   NumWebVisitsMonth
##                   0                   0                   0                   0
##        AcceptedCmp3        AcceptedCmp4        AcceptedCmp5        AcceptedCmp1
##                   0                   0                   0                   0
##        AcceptedCmp2            Complain       Z_CostContact           Z_Revenue
##                   0                   0                   0                   0
##            Response
##                   0
```

The income column has 24 missing values. Let's fill it with it's median value.

```
mkt$Income[is.na(mkt$Income)] <- median(mkt$Income, na.rm=TRUE)
```

As Z_CostContact and Z_Revenue has same unique value, lets drop those columns first.

```
mkt <- subset(mkt,select=-c(Z_CostContact,Z_Revenue,ID))
```

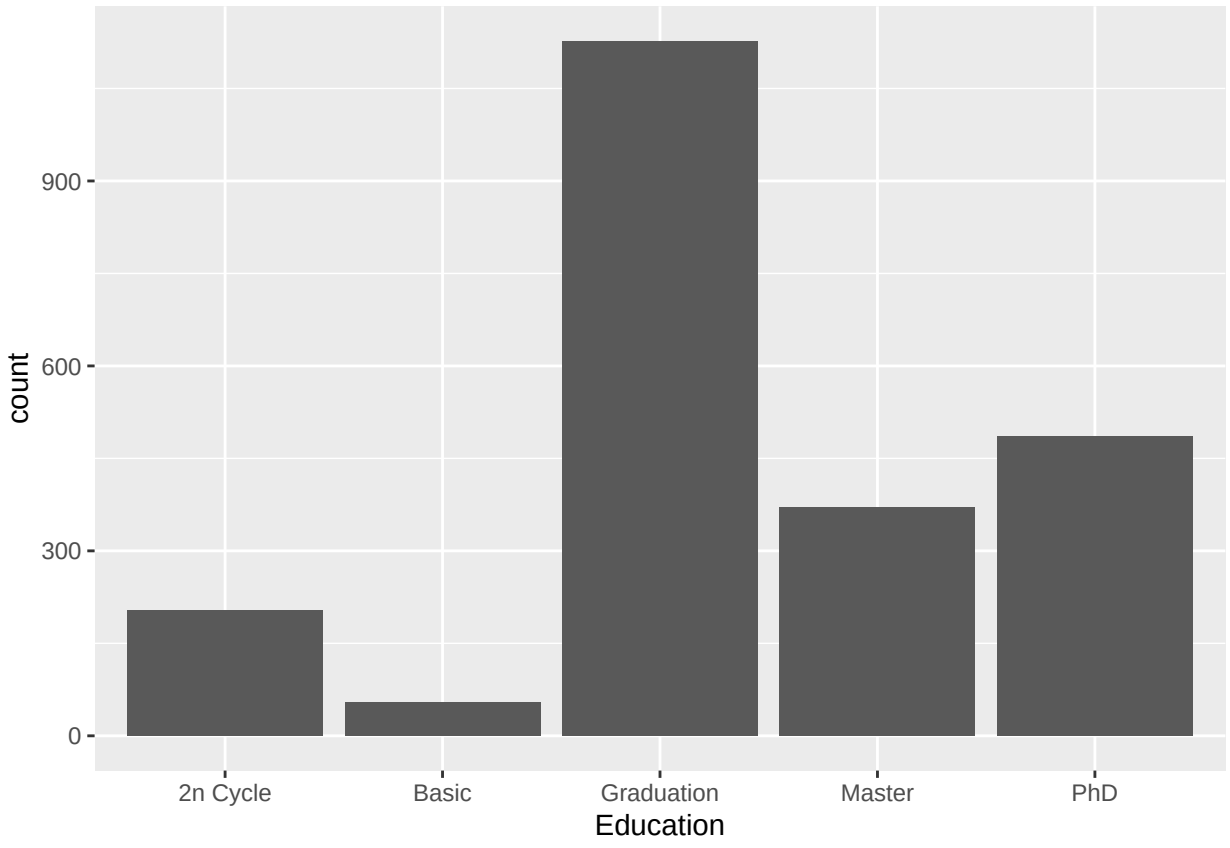
Part1: Let's assume we are working with "Reliance" retail store data set.Reliance wants to see if they want to spend more money on advertising about their app towards a particular category of customers based on Education. To get an insight about this let's perform a hypothesis test on Education and NumWebPurchases variables.

Let's see what different categories we have within the Education column.

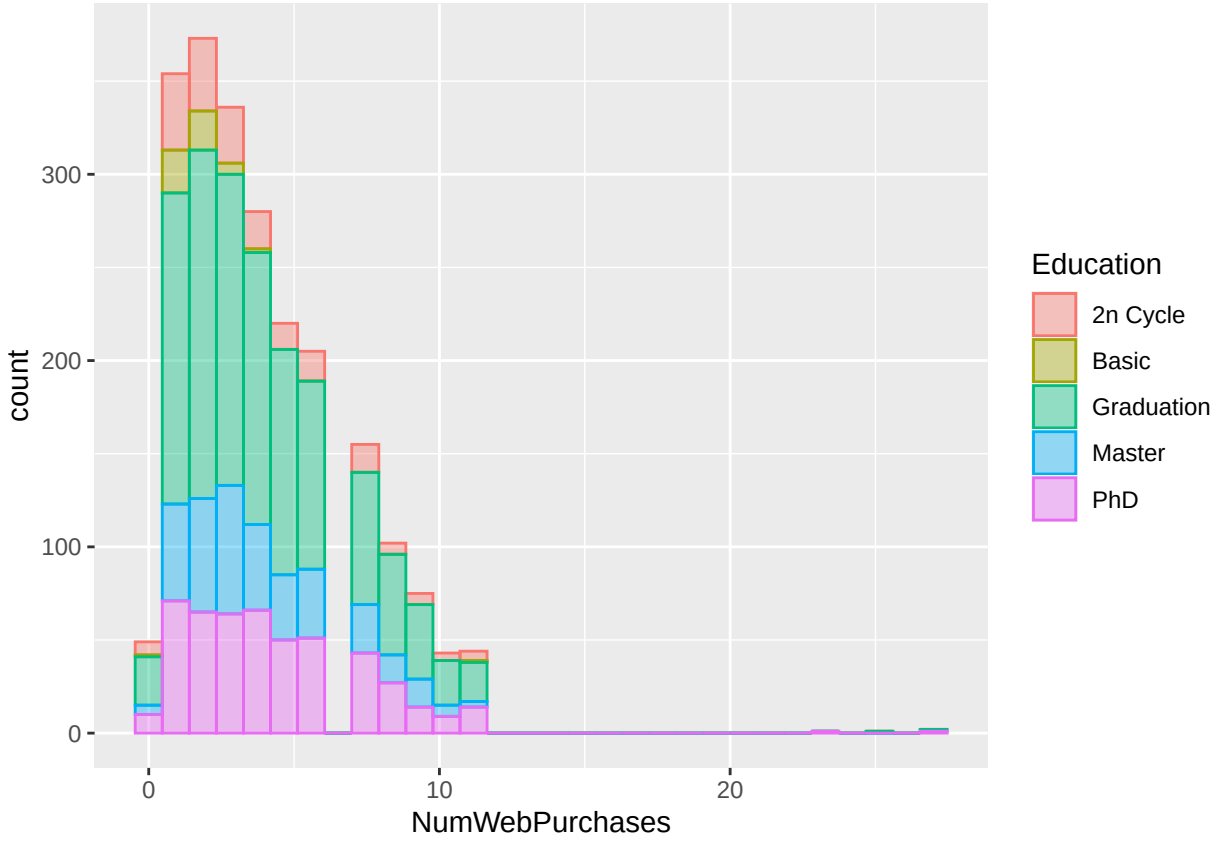
```
table(mkt$Education)
```

```
##
##   2n Cycle      Basic Graduation     Master        PhD
##        203         54       1127        370        486
```

```
ggplot(mkt,aes(x=Education))+geom_bar()
```



```
ggplot(mkt,aes(x=NumWebPurchases))+geom_histogram(aes(color = Education,fill=Education),bins=30,alpha=0
```



Stating our hypothesis:

H0: Mean of number of online orders placed by different groups of Education is same.
H1: Atleast one of the mean differs.

We will perform an ANOVA on this to see if the mean number of online orders differ significantly.

```
#First Hypothesis Test
anova(aov(NumWebPurchases ~ Education,data=mkt))
```

```
## Analysis of Variance Table
##
## Response: NumWebPurchases
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Education    4   344.8  86.200  11.371 3.888e-09 ***
## Residuals 2235 16943.1   7.581
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As, we can see the p value is less than 0.05 so we reject the null hypothesis.

Let's test the assumptions of ANOVA:

1.Independence: We need our populations to be independent from one another, our data set has records of different customers and all the populations of different groups and samples and independent.

2.Homogeneity of Variance: The variances of the groups should be equal, meaning that the spread or dispersion of the data in each group should be similar.We can check this using Levene's test.

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
leveneTest(NumWebPurchases~Education, data = mkt)
```

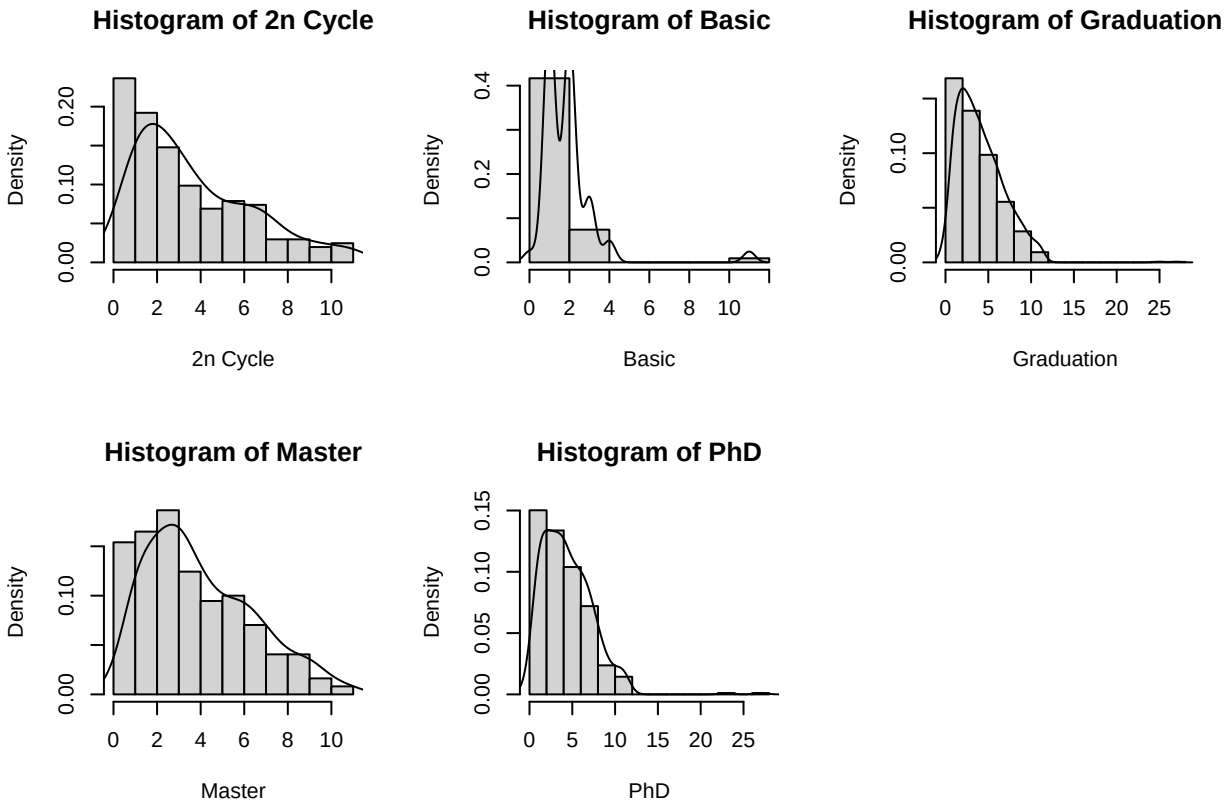
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    4  8.2344 1.368e-06 ***
##       2235
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value for Levene's test is less than 0.05 which means that the assumption of equal variances is violated.

3.Normality: We check for normality in each population.

```
par(mfrow = c(2,3))
hist(mkt$NumWebPurchases[mkt$Education == '2n Cycle'],freq=F, main = "Histogram of 2n Cycle", xlab = "2n
lines(density(mkt$NumWebPurchases[mkt$Education == "2n Cycle"]))
hist(mkt$NumWebPurchases[mkt$Education == 'Basic'],freq=F, main = "Histogram of Basic", xlab = "Basic")
lines(density(mkt$NumWebPurchases[mkt$Education == "Basic"]))
hist(mkt$NumWebPurchases[mkt$Education == 'Graduation'],freq=F, main = "Histogram of Graduation", xlab =
lines(density(mkt$NumWebPurchases[mkt$Education == "Graduation"]))
hist(mkt$NumWebPurchases[mkt$Education == 'Master'],freq=F, main = "Histogram of Master", xlab = "Master
lines(density(mkt$NumWebPurchases[mkt$Education == "Master"]))
hist(mkt$NumWebPurchases[mkt$Education == 'PhD'],freq=F, main = "Histogram of PhD", xlab = "PhD")
lines(density(mkt$NumWebPurchases[mkt$Education == "PhD"]))
```



From examining the histograms we can see that all of the groups are right skewed
Let's examine the normality using Shapiro's Wilk test.

```
tapply( mkt$NumWebPurchases, mkt$Education, shapiro.test)
```

```
## $`2n Cycle`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.90196, p-value = 2.694e-10
##
##
## $Basic
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.56774, p-value = 2.305e-11
##
##
## $Graduation
##
##  Shapiro-Wilk normality test
##
```

```
## data:  X[[i]]
## W = 0.89927, p-value < 2.2e-16
##
##
## $Master
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.93407, p-value = 9.941e-12
##
##
## $PhD
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.88357, p-value < 2.2e-16
```

From Shapiro Wilk's normality test, we can observe that for all groups the p values is less than 0.05, which means that the assumption of normality is violated.

Since 2 of the assumptions are violated let's try using a non-parametric test called Kruskal Wallis test.

```
kruskal.test(NumWebPurchases~Education, data=mkt)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  NumWebPurchases by Education
## Kruskal-Wallis chi-squared = 59.462, df = 4, p-value = 3.763e-12
```

As, the p value of Kruskal-Wallis test is less than 0.05, we reject the null hypothesis and conclude that the mean of number of online orders differ significantly within the groups of Education.

Part1b: Reliance has a special section for Gold Products. Mr Ryan, the product manager of Reliance Gold Products wants to understand his customers better to segment the customers based on the data. So, he has reached out to the analyst with the sales data we have.

Based on the research, it is known that the people who have high income spend more on Gold Products. Let's test this hypothesis:

H0: There is no significant difference between amount spent on gold products between high income and low income customers.
H1: The mean amount spent on gold products is higher for high income groups than low income groups.

Let's split the Income column into two categories based on the median.

```
median(mkt$Income)
```

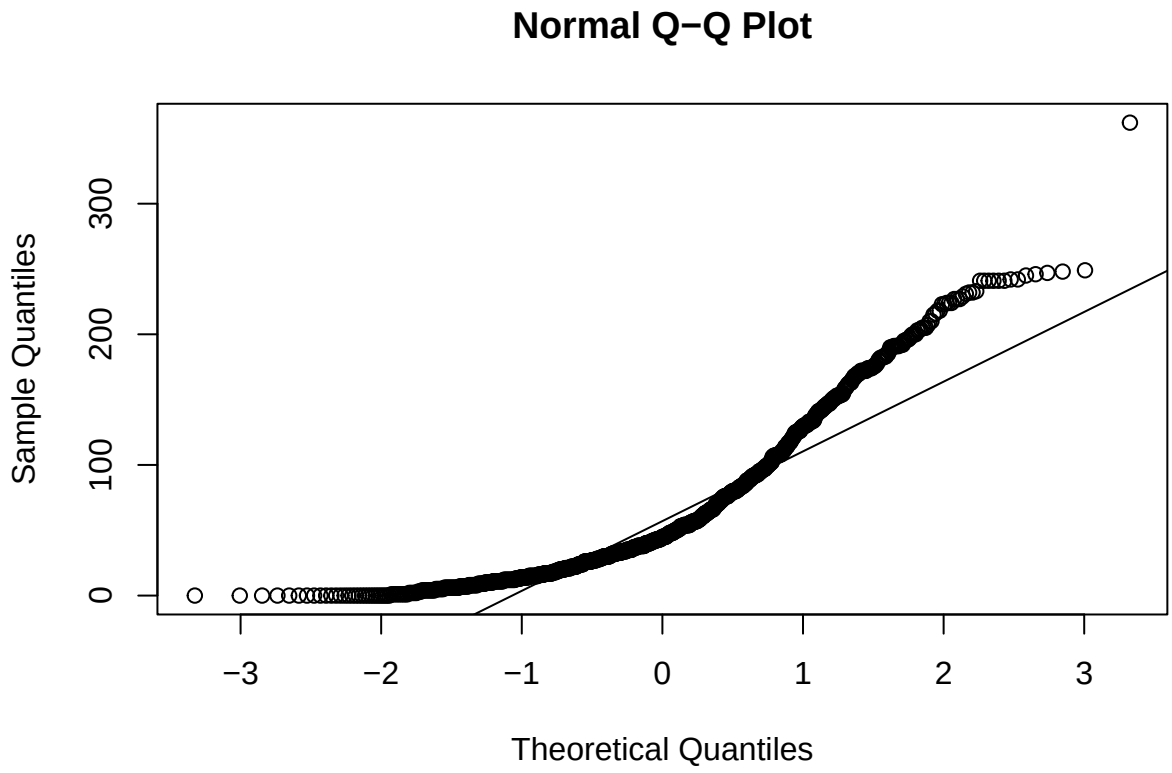
```
## [1] 51381.5
```

```
high_income <- subset(mkt, mkt$Income>=51381.5)
low_income <- subset(mkt,mkt$Income<51381.5)
```

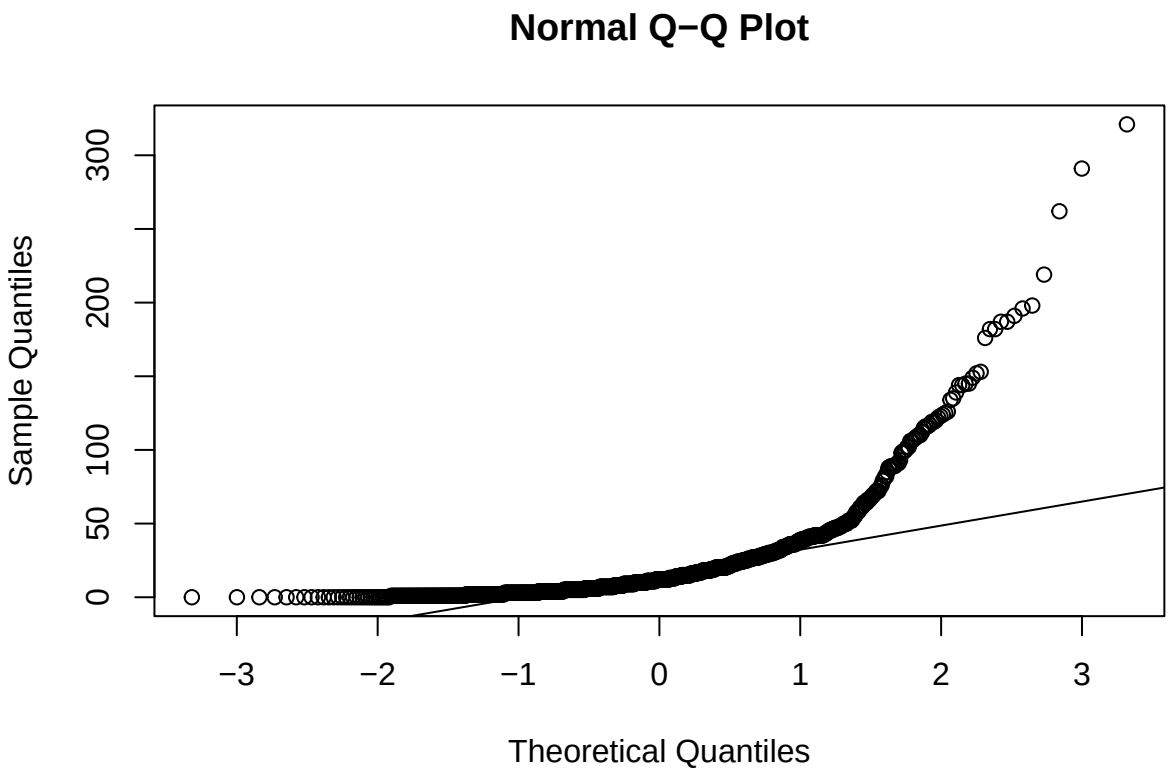
I want to use a two sample t test to check the difference between mean of 2 groups. Let's check the assumptions of t test.

Checking the assumption of normality:

```
qqnorm(high_income$MntGoldProds)
qqline(high_income$MntGoldProds)
```




```
qqnorm(low_income$MntGoldProds)
qqline(low_income$MntGoldProds)
```



Both the qqplots deviate from the line at the end, so we can say that they don't follow a normal distribution. Checking the normality with a Shapiro-Wilk normality test:

```
shapiro.test(x=high_income$MntGoldProds)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  high_income$MntGoldProds
## W = 0.86822, p-value < 2.2e-16
```

```
shapiro.test(x=low_income$MntGoldProds)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  low_income$MntGoldProds
## W = 0.60988, p-value < 2.2e-16
```

The p value is less than 0.05 in both cases, so we can conclude that the groups don't follow a normal distribution.

As the normality assumption is violated we can't use a t test for this. Instead, we use a non parametric t test called Wilcoxon rank sum test.

```r
wilcox.test(high_income$MntGoldProds, low_income$MntGoldProds, alternative='greater')
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  high_income$MntGoldProds and low_income$MntGoldProds
## W = 979288, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

As, the p value is less than 0.05, we reject the null hypothesis and conclude that higher income group spend more on Gold Products than the lower income group.

Second hypothesis test:

Part2: The marketing team of Reliance also wants to know if the families with more number of Children have bought more sweet products in the past year. Let's make a new variable called Children to sum up both kids and teens. If we categorize the Amount of sweet products into two categories by mean by calling Amount of Sweet below average to be "less" and above average to be "more". Let us create a variable called 'sweet.cat'. We see the mean amount of sweet products is 27.06.

```r
mean(mkt$MntSweetProducts)
```

```
## [1] 27.06295
```

```r
max(mkt$MntSweetProducts)
```

```
## [1] 263
```

```r
mkt$sweet.cat <-  cut(mkt$MntSweetProducts, c(0, 27.06, 263 ), c("Less","More")) #Cut function
table(mkt$sweet.cat)
```

```
##
## Less More
## 1178  643
```

```r
mkt$Children <- mkt$Kidhome + mkt$Teenhome
mkt$Children <- as.factor(mkt$Children)
```

Hypothesis:

H0: The amount of sweet products a customer buys and the number of children he/she has are independent.
H1: The amount of sweet products a customer buys and the number of children he/she has is dependent.

We can use a Chi-square test of independence to see if these two variables are dependent.

Testing Assumptions to perform a chi-square test of Independence are: All assumptions are met.
-Each of the expect values should be greater than or equal to 5

-Both variables have atleast two categories.
-The samples must be independent.

```
tab <- table(mkt$sweet.cat,mkt$Children)
tab # a customer can have 0,1,2,3 children and can purchase less More.
```

```
##
##          0   1   2   3
##   Less 222 673 259  24
##   More 371 240  27   5
```

```
chisq.test(tab)$exp
```

```
##
##               0        1        2        3
##   Less 383.6101 590.6172 185.0126 18.76002
##   More 209.3899 322.3828 100.9874 10.23998
```

```
chisq.test(tab, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 313.3, df = 3, p-value < 2.2e-16
```

The Chi-Squared test has given a p value less than 0.05, which means we reject the null hypothesis and conclude that there is a relationship between number of children each customer has and Amount of sweet people buys at Reliance.

THIRD HYPOTHESIS TEST

Part3: A manager at Reliance claims that they should consider the number of kids a customer has and his education status before they market their wines to people. To test if this claim is true , we Consider predictors KidHome and Education as two predictors for MntWines.Let us try to fit a multiple linear regression equation.

H0: The number of kids the customer has and their Education status have no significant impact on the amount they spend on wines.
H1: The number of kids the customer has and their Education status have a significant impact on the amount they spend on wines.

```
lm <- lm(mkt$MntWines~mkt$Kidhome+mkt$Income)
summary(lm)
```

```
##
## Call:
```

```
## lm(formula = mkt$MntWines ~ mkt$Kidhome + mkt$Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3877.5  -143.3   -44.9    87.7  1110.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.537e+01  1.634e+01   4.612 4.21e-06 ***
## mkt$Kidhome -1.915e+02  1.122e+01 -17.067  < 2e-16 ***
## mkt$Income   6.004e-03  2.413e-04  24.882  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 258.7 on 2237 degrees of freedom
## Multiple R-squared:  0.4097, Adjusted R-squared:  0.4092
## F-statistic: 776.3 on 2 and 2237 DF,  p-value: < 2.2e-16
```

By performing the multiple linear regression, we got the p value to be less than 0.05, which means we reject the null hypothesis and conclude that the number of kids and the level of Education the customer has will impact the amount of wines they purchase.

Let us use certain predictors to predict the amount of wines sold.

```
lm1 <- lm( MntWines ~ Income + NumWebPurchases + NumCatalogPurchases + NumStorePurchases + NumWebVisitsl
summary(lm1)
```

```
##
## Call:
## lm(formula = MntWines ~ Income + NumWebPurchases + NumCatalogPurchases +
##     NumStorePurchases + NumWebVisitsMonth + Education + Kidhome +
##     MntMeatProducts, data = mkt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1470.45   -93.40   -16.90    55.11  1029.73
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -3.315e+02  2.615e+01 -12.679  < 2e-16 ***
## Income               2.286e-03  2.490e-04   9.181  < 2e-16 ***
## NumWebPurchases      1.977e+01  1.987e+00   9.949  < 2e-16 ***
## NumCatalogPurchases  2.885e+01  2.380e+00  12.122  < 2e-16 ***
## NumStorePurchases    2.941e+01  1.846e+00  15.934  < 2e-16 ***
## NumWebVisitsMonth    2.193e+01  2.499e+00   8.777  < 2e-16 ***
## EducationBasic       4.464e+01  3.190e+01   1.400   0.1618
## EducationGraduation  3.921e+01  1.569e+01   2.499   0.0125 *
## EducationMaster      9.824e+01  1.796e+01   5.469 5.04e-08 ***
## EducationPhD         1.320e+02  1.725e+01   7.654 2.89e-14 ***
## Kidhome             -6.423e+01  1.022e+01  -6.285 3.92e-10 ***
## MntMeatProducts      2.063e-01  2.974e-02   6.935 5.29e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 205.3 on 2228 degrees of freedom
## Multiple R-squared:  0.6296, Adjusted R-squared:  0.6278
## F-statistic: 344.4 on 11 and 2228 DF,  p-value: < 2.2e-16
```

The p value is less than 0.05, indicating that all the variables have a significant impact on the amount of wines purchased.