Stat 416 Project: Statistics for Efficeint Marketing Campaign

Keerthi Sreenivas Konjety, Vishnu Elangovan

12/16/2021

Abstract: In this project we have used a marketing campaign dataset which consists of Segments of customers, the number of orders they placed and the number of different products they bought. Our study aims to find insights between these variables and inferences for better marketing. Our approach consists of 3 hypothesis. The first hypothesis involves testing whether Number of Web Orders(Q) placed by customers of different marital status(C) is different to focus advertising on a particular group and we found that mean of web orders placed by different marital groups is same[Parametric]. We also performed a second test to see which age group consumes more amount of wine and found out that customers whose age is above 45 consume more wine than whose age is below 45[Non-Parametric]. Second hypothesis tests to see if the number of children(C) a customer has and amount of sweet products(C) he/she purchases is dependent, and we found that these two variables are actually dependent on each other. Lastly, the third hypothesis is to check if amount of wines purchased and age are linearly related. Further we do a linear regression between to predict amount of wines purchased from meat purchases, multiple regression between amount of wines purchased and kidHome and educational status. We then perform boot strapping approach on the same multiple regression equation and also an additional model which could be predictors to wines purchased. As additional work, we have included Model fitting using AIC values(forward,backward and stepwise) and some exploratory data analysis.

We chose a Marketing campaign dataset from kaggle.
Source:https:https://www.kaggle.com/rodsaldanha/arketing-campaign
These are that are listed here are variables from the actual dataset.

```r
mkt = read.csv("marketing_campaign.csv")
head(mkt)
```

```
##     ID Year_Birth  Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524       1957 Graduation         Single  58138       0        0      4/9/12
## 2 2174       1954 Graduation         Single  46344       1        1      8/3/14
## 3 4141       1965 Graduation       Together  71613       0        0  21-08-2013
## 4 6182       1984 Graduation       Together  26646       1        0     10/2/14
## 5 5324       1981        PhD        Married  58293       1        0  19-01-2014
## 6 7446       1967     Master       Together  62513       0        1      9/9/13
##   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
```

```
#summary(mkt)
colnames(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"
```

This dataset consists of data about customers, we can see the column names of the data in the output of the above cell above. We will be working with the following variables:

1. Year_Birth: Birth Year of the customer (Time series)

2. Education: Education Level of Customer (Categorical)

3. Marital_Status: Marital Status of the customer (Categorical)

4. Income: Yearly household income of the customer (Quantitative)

5. Kidhome: Number of children in customer's household (Quantitative)

6. Teenhome: Number of teenagers in customer's household (Quantitative)

7. MntWines: Amount spent on wine in last 2 years (Quantitative)

8. MntFruits: Amount spent on fruits in last 2 years (Quantitative)

9. MntMeatProducts: Amount spent on meat in last 2 years (Quantitative)

10. MntFishProducts: Amount spent on fish in last 2 years (Quantitative)

11. MntSweetProducts: Amount spent on sweets in last 2 years (Quantitative)

12. MntGoldProds: Amount spent on gold in last 2 years (Quantitative)

13. NumDealsPurchases: Number of purchases made with a discount (Quantitative)

14. NumWebPurchases: Number of purchases made through the company's website (Quantitative)

15. NumStorePurchases: Number of purchases made directly in stores (Quantitative)

16. NumWebVisitsMonth: Number of visits to company's website in the last month(Quantitative)

```
#dropping columns that are not of our interest
mkt = subset(mkt, select = -c(ID, Dt_Customer, Recency, AcceptedCmp3,AcceptedCmp4,AcceptedCmp5,AcceptedC
#displaying the column names
colnames(mkt)
```

```
##  [1] "Year_Birth"          "Education"           "Marital_Status"
##  [4] "Income"              "Kidhome"             "Teenhome"
##  [7] "MntWines"            "MntFruits"           "MntMeatProducts"
## [10] "MntFishProducts"     "MntSweetProducts"    "MntGoldProds"
## [13] "NumDealsPurchases"   "NumWebPurchases"     "NumCatalogPurchases"
## [16] "NumStorePurchases"   "NumWebVisitsMonth"
```

To convert the time series data of Year_Birth, we will be converting it into age

```
mkt$Age = 2021-mkt$Year_Birth #creating a new varaible age
```

Part1(A):

Let's assume we are working with "Spencer's" retail store data set. Spencer's wants to see if they need to spend more money on advertising about their App towards a particular category of customers based on their Marital status. To get an insight about this, we will be performing a hypothesis test on "Marital_Status"(C) and "NumWebPurchases" (Q) check if all groups have placed same number of orders.

Listing the different categories in Marital_Status and number of customers in each category.

```
table(mkt$Marital_Status)
```

```
##
##   Absurd    Alone Divorced  Married   Single Together    Widow     YOLO
##        2        3      232      864      480      580       77        2
```

We have chosen to drop columns with Marital_Status "Absurd","Alone" and "YOLO". They just seem to ridiculous and noisy.

```
mkt = subset(mkt, mkt$Marital_Status != "Absurd")
mkt = subset(mkt, mkt$Marital_Status != "Alone")
mkt = subset(mkt, mkt$Marital_Status != "YOLO")
table(mkt$Marital_Status)
```

```
##
## Divorced  Married   Single Together    Widow
##      232      864      480      580       77
```

```
mean(mkt$NumWebPurchases)
```

```
## [1] 4.081505
```

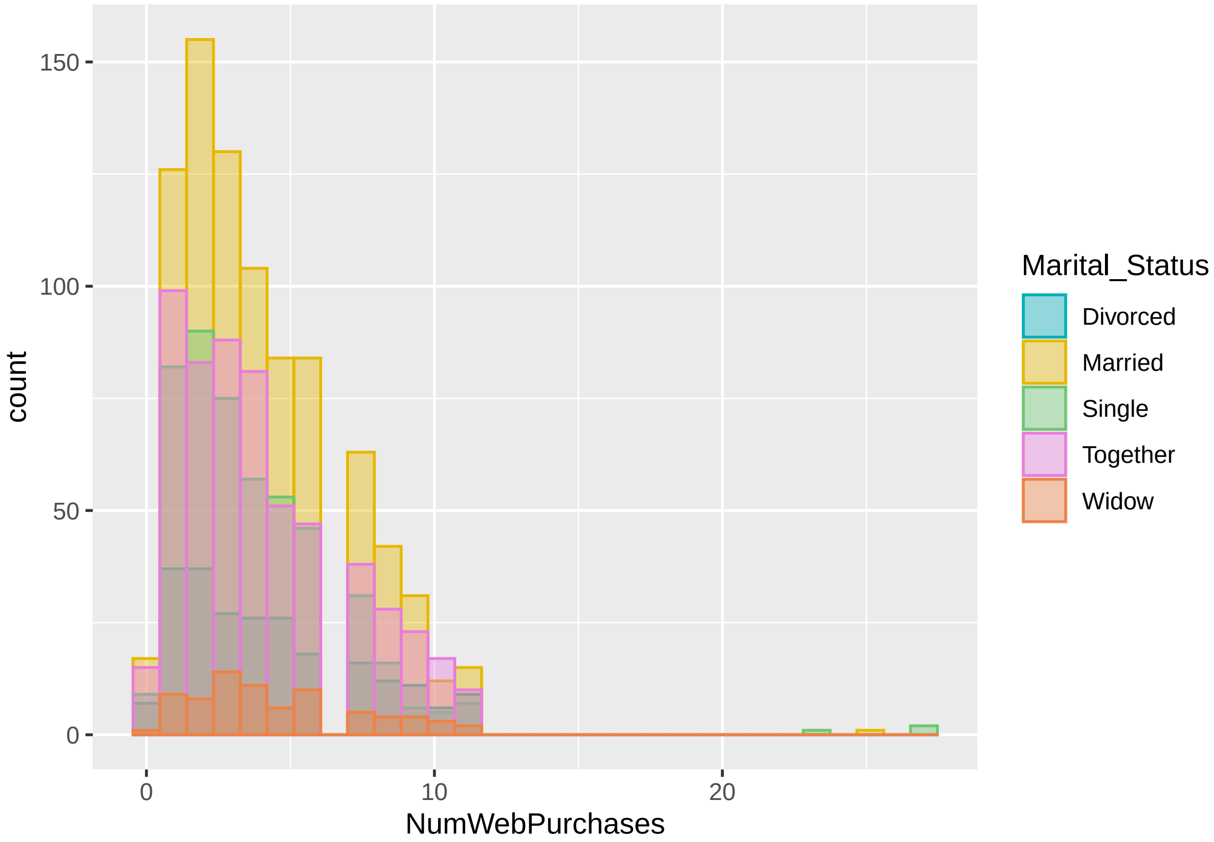
```
# library
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

```
# change fill and outline color manually
ggplot(mkt, aes(x = NumWebPurchases)) +
  geom_histogram(aes(color = Marital_Status, fill = Marital_Status),
                 position = "identity", bins = 30, alpha = 0.4) +
  scale_color_manual(values = c("#00AFBB", "#E7B800", "#6DC770","#e87ddd","#ed8345")) +
  scale_fill_manual(values = c("#00AFBB", "#E7B800","#6DC770","#e87ddd","#ed8345"))
```



First Hypothesis: CQ: Stating our hypothesis:

H0: Mean on number of Online orders placed by different groups(Divorced,Married,Single,Together and Widow) is same. H1: Atleast one of the means differs.

We shall be performing an ANOVA on the means of different Marital_Status 5 groups (Divorced,Married,Single,Together and Widow) to see if the mean of number of orders placed online differs significantly.

```
anova(aov(NumWebPurchases ~ Marital_Status, data= mkt))
```

```
## Analysis of Variance Table
##
## Response: NumWebPurchases
##                  Df  Sum Sq Mean Sq F value Pr(>F)
## Marital_Status    4    55.7 13.9197  1.8078 0.1246
## Residuals      2228 17155.5  7.6999
```

From the output of Anova, we see that the p-value is 0.1246 < alpha at 0.05 significance. Hence, we fail to reject the null Hypothesis and conclude that the means of number of online orders placed by different marital groups do not differ significantly.

Testing Assumptions for ANOVA:

Now, lets test the assumptions for conducting an ANOVA and check whether its really correct to test our hypothesis this way?

1.Independent populations and Independent Samples within observations: 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.We must have a equal variances across populations: 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 ~ Marital_Status, 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  1.4509 0.2147
##       2228
```

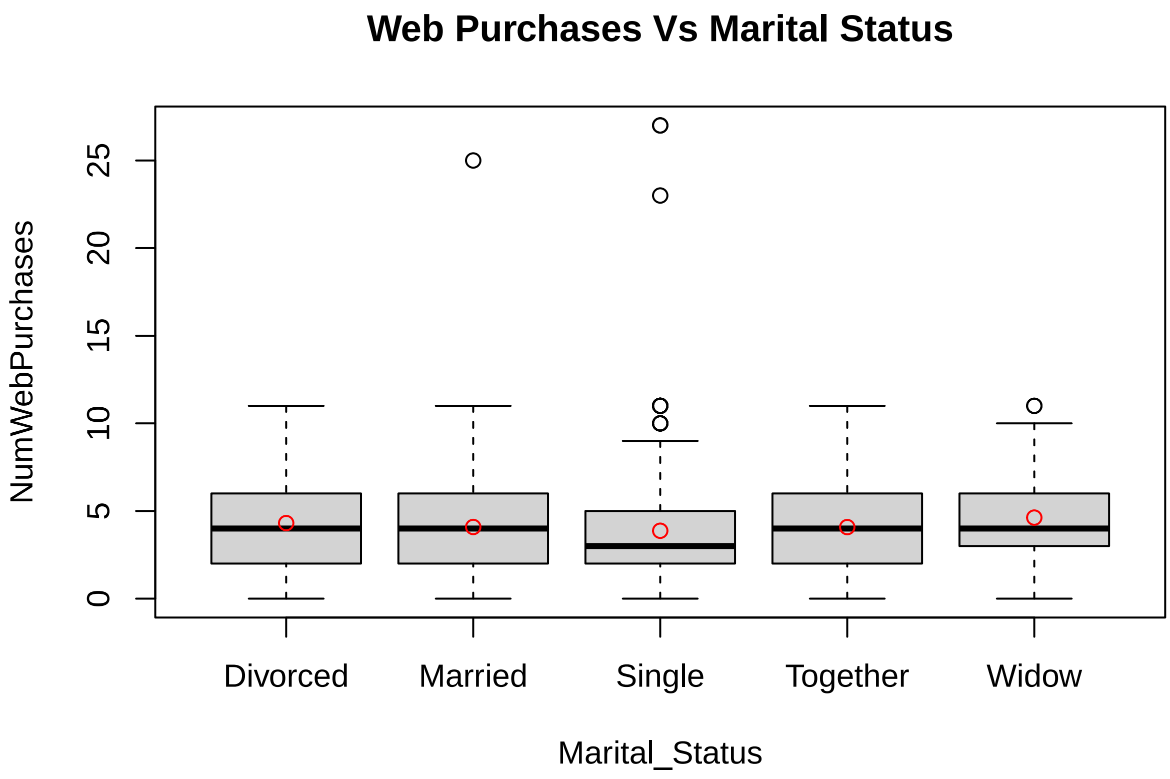
The p-value for our Levene's test is 0.2147 > alpha at 0.05 significance level, aiding us in concluding that variance across different marital groups is not different.

3.Finally, we check for Normality in each population:

```
#Box Plot
boxplot(NumWebPurchases ~ Marital_Status, data=mkt, main="Web Purchases Vs Marital Status")
(group.means <- tapply(mkt$NumWebPurchases, mkt$Marital_Status, mean))
```

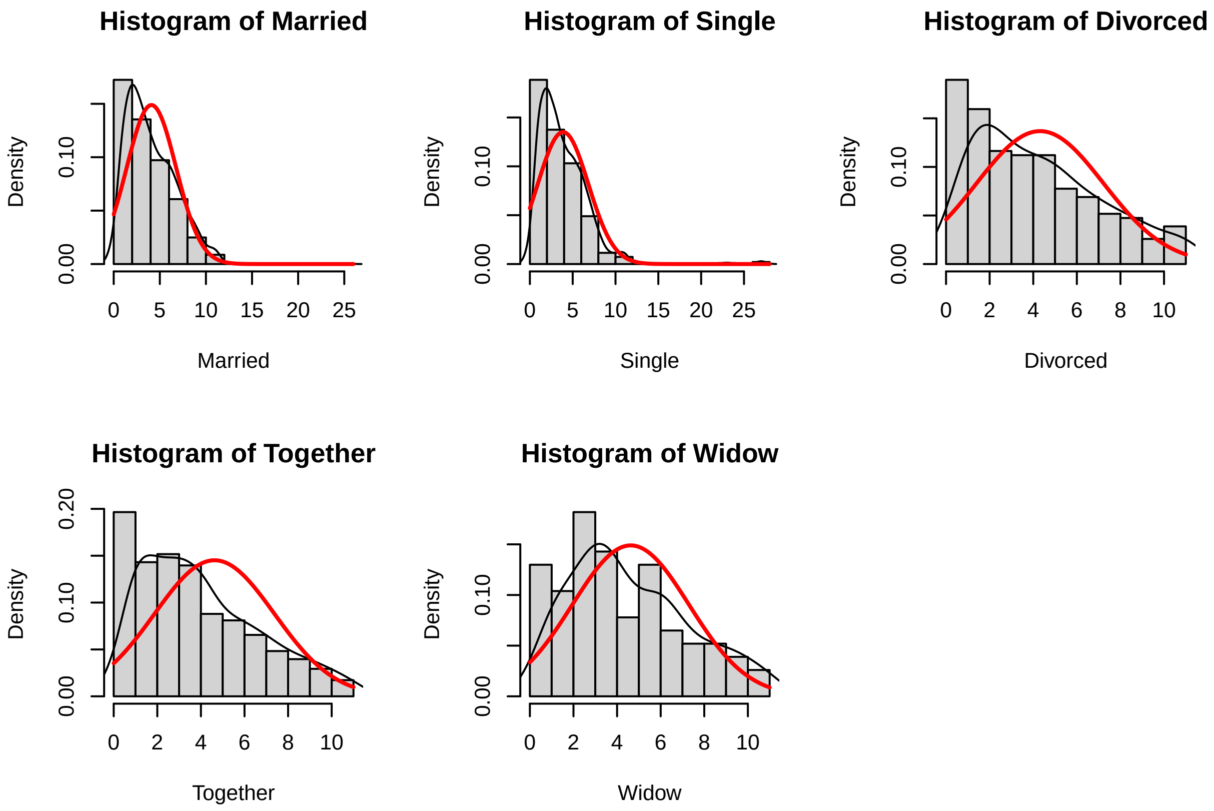
```
## Divorced  Married   Single Together    Widow
## 4.310345 4.087963 3.872917 4.081034 4.623377
```

```
points(1:5, group.means, col = "red")
```



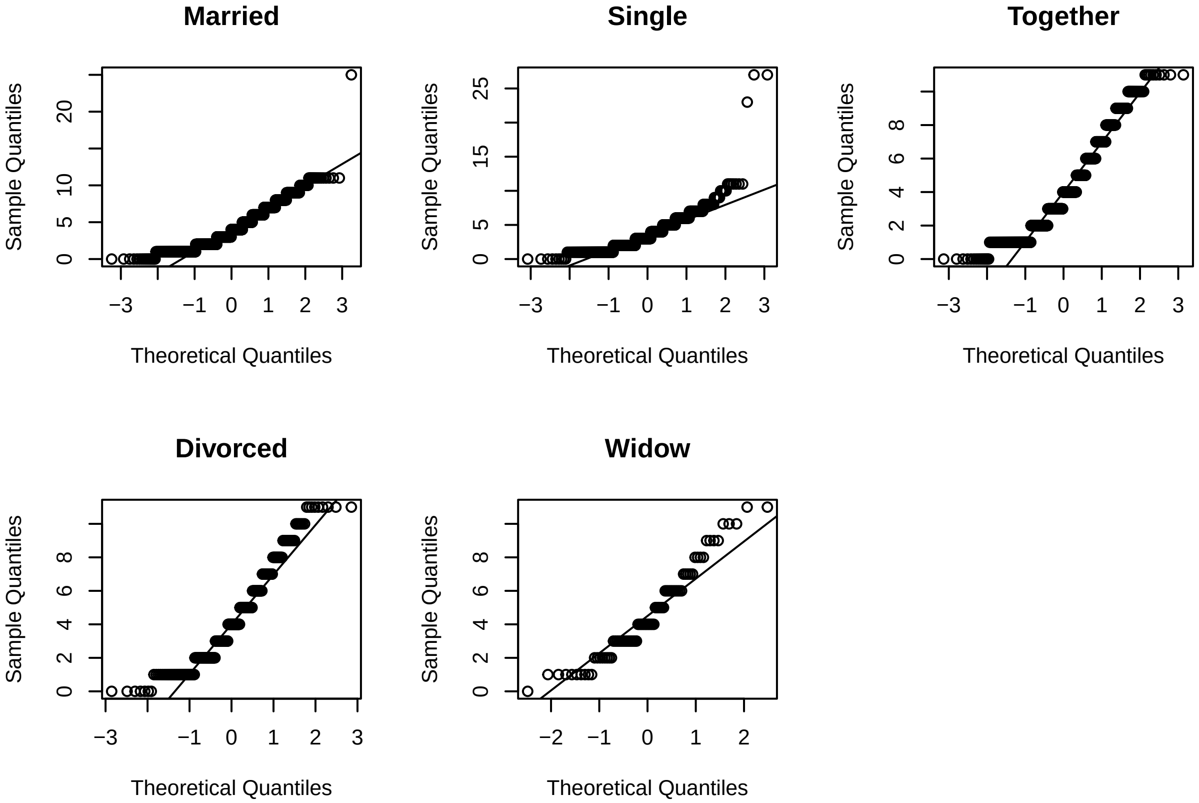
```
# Histograms
par(mfrow = c(2,3))
hist(mkt$NumWebPurchases[mkt$Marital_Status == "Married"], freq = F, main = "Histogram of Married", xla
lines(density(mkt$NumWebPurchases[mkt$Marital_Status == "Married"]))
curve(dnorm(x, mean=mean(mkt$NumWebPurchases[mkt$Marital_Status == "Married"]), sd=sd(mkt$NumWebPurchas
hist(mkt$NumWebPurchases[mkt$Marital_Status == "Single"], freq = F, main = "Histogram of Single", xlab =
lines(density(mkt$NumWebPurchases[mkt$Marital_Status == "Single"]))
curve(dnorm(x, mean=mean(mkt$NumWebPurchases[mkt$Marital_Status == "Single"]), sd=sd(mkt$NumWebPurchase
hist(mkt$NumWebPurchases[mkt$Marital_Status == "Divorced"], freq = F, main = "Histogram of Divorced", xl
lines(density(mkt$NumWebPurchases[mkt$Marital_Status == "Divorced"]))
curve(dnorm(x, mean=mean(mkt$NumWebPurchases[mkt$Marital_Status == "Divorced"]), sd=sd(mkt$NumWebPurcha
hist(mkt$NumWebPurchases[mkt$Marital_Status == "Together"], freq = F, main = "Histogram of Together", xl
lines(density(mkt$NumWebPurchases[mkt$Marital_Status == "Together"]))
curve(dnorm(x, mean=mean(mkt$NumWebPurchases[mkt$Marital_Status == "Widow"]), sd=sd(mkt$NumWebPurchases
hist(mkt$NumWebPurchases[mkt$Marital_Status == "Widow"], freq = F, main = "Histogram of Widow", xlab =
lines(density(mkt$NumWebPurchases[mkt$Marital_Status == "Widow"]))
curve(dnorm(x, mean=mean(mkt$NumWebPurchases[mkt$Marital_Status == "Widow"]), sd=sd(mkt$NumWebPurchases

#Normal QQ plots
par(mfrow = c(2,3))
```



```
qqnorm(mkt$NumWebPurchases[mkt$Marital_Status == "Married"], main = "Married")
qqline(mkt$NumWebPurchases[mkt$Marital_Status == "Married"])
qqnorm(mkt$NumWebPurchases[mkt$Marital_Status == "Single"], main = "Single")
qqline(mkt$NumWebPurchases[mkt$Marital_Status == "Single"])
qqnorm(mkt$NumWebPurchases[mkt$Marital_Status == "Together"], main = "Together")
qqline(mkt$NumWebPurchases[mkt$Marital_Status == "Together"])
qqnorm(mkt$NumWebPurchases[mkt$Marital_Status == "Divorced"], main = "Divorced")
qqline(mkt$NumWebPurchases[mkt$Marital_Status == "Divorced"])
qqnorm(mkt$NumWebPurchases[mkt$Marital_Status == "Widow"], main = "Widow")
qqline(mkt$NumWebPurchases[mkt$Marital_Status == "Widow"])

#Skew and kurtosis
par(mfrow=c(2,3))
```



```
library(pastecs)
```

```
##
## Attaching package: 'pastecs'

## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
stat.desc(x=mkt$NumWebPurchases[mkt$Marital_Status == "Married"], norm=TRUE)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
## 8.640000e+02 1.700000e+01 0.000000e+00 0.000000e+00 2.500000e+01 2.500000e+01
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 3.532000e+03 4.000000e+00 4.087963e+00 9.112324e-02 1.788491e-01 7.174177e+00
##      std.dev     coef.var     skewness     skew.2SE     kurtosis     kurt.2SE
## 2.678465e+00 6.552079e-01 1.115715e+00 6.705894e+00 3.590402e+00 1.080225e+01
##   normtest.W   normtest.p
## 9.136152e-01 9.233207e-22
```

```
stat.desc(x=mkt$NumWebPurchases[mkt$Marital_Status == "Single"], norm=TRUE)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
## 4.800000e+02 9.000000e+00 0.000000e+00 0.000000e+00 2.700000e+01 2.700000e+01
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 1.859000e+03 3.000000e+00 3.872917e+00 1.347252e-01 2.647254e-01 8.712417e+00
##      std.dev     coef.var     skewness     skew.2SE     kurtosis     kurt.2SE
## 2.951680e+00 7.621337e-01 2.893194e+00 1.297907e+01 1.765725e+01 3.968706e+01
##   normtest.W   normtest.p
## 7.882003e-01 1.689244e-24
```

```r
stat.desc(x=mkt$NumWebPurchases[mkt$Marital_Status == "Divorced"], norm=TRUE)
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##  2.320000e+02  7.000000e+00  0.000000e+00  0.000000e+00  1.100000e+01
##         range           sum        median          mean       SE.mean
##  1.100000e+01  1.000000e+03  4.000000e+00  4.310345e+00  1.914179e-01
##  CI.mean.0.95           var       std.dev      coef.var      skewness
##  3.771482e-01  8.500672e+00  2.915591e+00  6.764171e-01  6.126669e-01
##      skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
##  1.917099e+00 -5.639270e-01 -8.859742e-01  9.292852e-01  4.153854e-09
```

```r
stat.desc(x=mkt$NumWebPurchases[mkt$Marital_Status == "Together"], norm=TRUE)
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##  5.800000e+02  1.500000e+01  0.000000e+00  0.000000e+00  1.100000e+01
##         range           sum        median          mean       SE.mean
##  1.100000e+01  2.367000e+03  4.000000e+00  4.081034e+00  1.125393e-01
##  CI.mean.0.95           var       std.dev      coef.var      skewness
##  2.210350e-01  7.345754e+00  2.710305e+00  6.641221e-01  6.757154e-01
##      skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
##  3.330360e+00 -3.785149e-01 -9.343700e-01  9.291906e-01  6.420599e-16
```

```r
stat.desc(x=mkt$NumWebPurchases[mkt$Marital_Status == "Widow"], norm=TRUE)
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##  77.000000000   1.000000000   0.000000000   0.000000000  11.000000000
##         range           sum        median          mean       SE.mean
##  11.000000000 356.000000000   4.000000000   4.623376623   0.313208301
##  CI.mean.0.95           var       std.dev      coef.var      skewness
##   0.623808394   7.553656869   2.748391688   0.594455506   0.521564091
##      skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
##   0.952078658  -0.613967343  -0.566955682   0.945343525   0.002383707
```

```r
#Shapiro-Wilk's  test for each group:
tapply( mkt$NumWebPurchases, mkt$Marital_Status, shapiro.test)
```

```
## $Divorced
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.92929, p-value = 4.154e-09
##
##
## $Married
```

```
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.91362, p-value < 2.2e-16
##
##
## $Single
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.7882, p-value < 2.2e-16
##
##
## $Together
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.92919, p-value = 6.421e-16
##
##
## $Widow
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.94534, p-value = 0.002384
```

Examining a Boxplot:
From examining the box plot we see that all group means and group variances have similar values, and the mean is close to median indicating that all groups.

Examining Histograms and QQ plots:
From examining the histograms, we see that Married and Single groups are right skewed, and Widow, Divorced and Together groups are fairly normal looking.The Normal QQ Plots show some evidence of Non-normality.

From the descriptive statistics of each group we see that,
Married: Skew.2SE = 6.705894e+00, kurt.2SE = 1.080225e+01
Single : Skew.2SE = 1.297907e+01, kurt.2SE = 3.968706e+01 Divorced: skew.2SE = 1.917099e+00, kurt.2SE = -8.859742e-01
Together: skew.2SE = 3.330360e+00, kurt.2SE = -9.343700e-01
Widow: skew.2SE = 0.952078658, kurt.2SE= -0.566955682

Except for the group "Widow" we do not observe any skew.2SE and kurt.2SE values falling in (-1,1) range. All other groups seem to violate normality.

From Shapiro Wilk Normality test on all groups, We observe that,
Divorced group has p-value = 4.154e-09 Married group has p-value = 2.2e-16
Single group has p-value = 2.2e-16
Together group has p-value = 6.421e-16

Widow group has p-value = 0.002384

All p-values are less than alpha at 0.05 significance, hence we reject our null hypothesis and conclude that normality assumption is violated. But the sample sizes are large enough for CLT to be applied.
Hence, We consider ANOVA to be an appropriate test in this scenario.

For the sake of this experiment, since normality is violated by most of the groups, we can try using a non-parametric test Kruskal wallis test (Normality is violated and Unbalanced design), to see if the conclusion given by ANOVA still holds.

```
kruskal.test( NumWebPurchases ~ Marital_Status, mkt )
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  NumWebPurchases by Marital_Status
## Kruskal-Wallis chi-squared = 8.6508, df = 4, p-value = 0.07044
```

Using Kruskal Wallis Test we get a test statistic of 8.6508 and p-value = 0.07044> alpha at 0.05 significance, from which we can conclude that the Mean of Number of orders purchased by Different Marital status groups is same and the conclusion we obtained from ANOVA still holds.

Now, Spencer's can decide that they can allot almost the same budget to advertise their products towards different Marital Status groups.

Part1b:

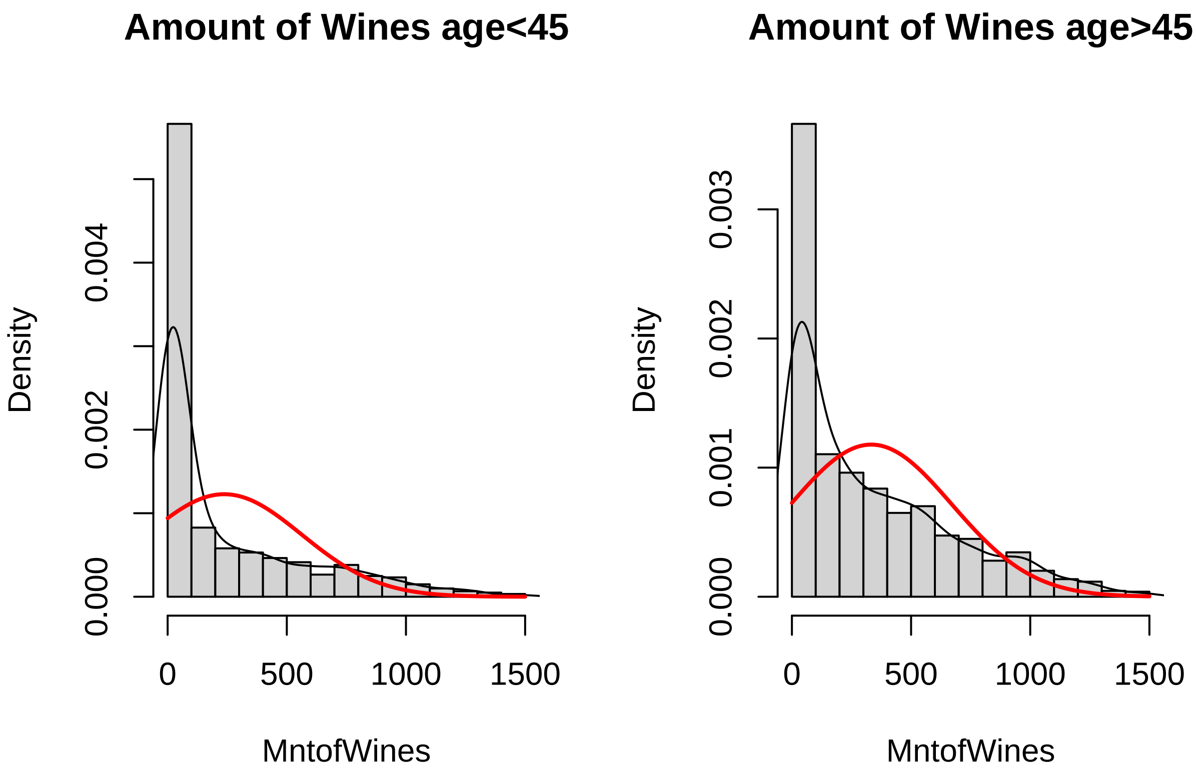
Spencer's has a special section for their exotic wines. Mr Nate, the product manager of Spencer's wines wants to understand his customers better, to segment his customers and market. So he has reached out to the analytics team with few questions from the sales data we have.

An article from a well-known research study claimed that, people above the age of 45 consume more wine more than the rest.
Nate wants to check whether this clam is valid or not.
As an analyst my hypothesis is here are as follows:
H0: Mean(MntWines) for age above45 = Mean(MntWines) for age below 45
H1: Mean(MntWines) for age above 45>Mean(MntWines) for age below 45
I am splitting my Age variable into two groups , one below 45 and one above 45.

```
#subsetting my dataset
lessthan45 <- subset(mkt,mkt$Age<45)
Greaterthan45 <- subset(mkt,mkt$Age>45)
```

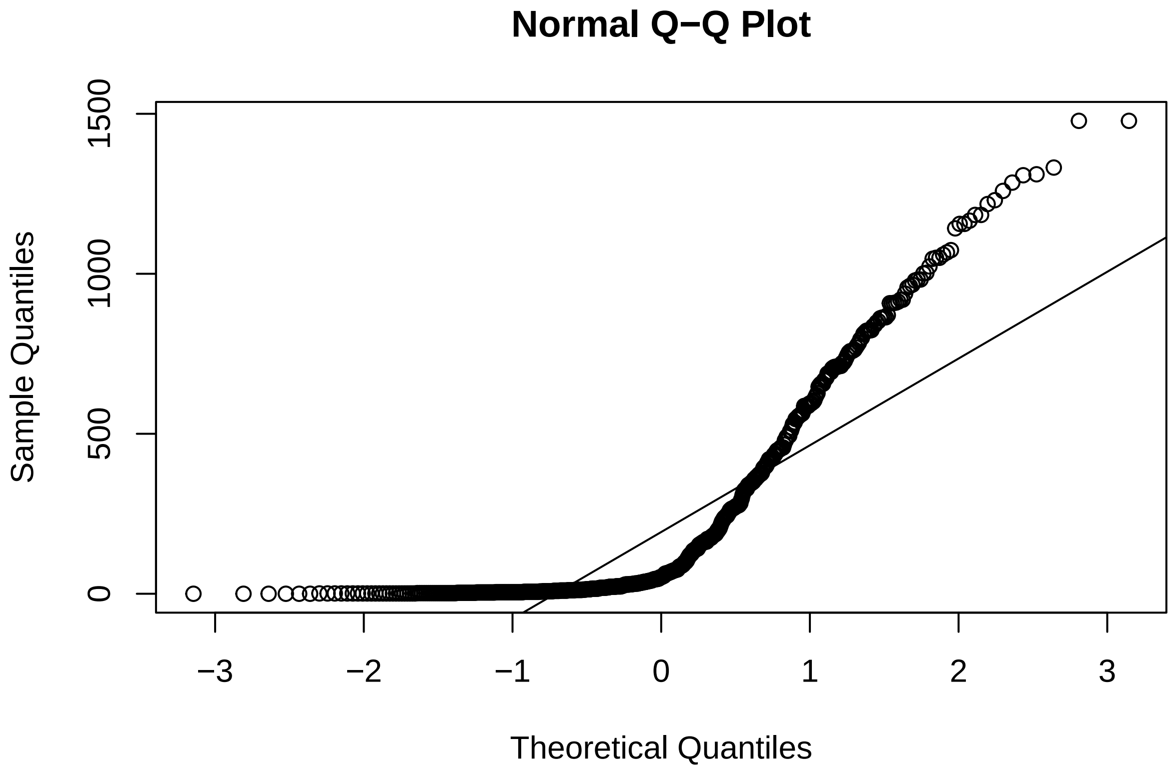
I plan to use the two-sample T-test to compare the means of the two groups. Before doing the test , I'm interested in checking my assumptions for the test.

```
par(mfrow=c(1,2))
#Histogram + normal curve
hist(lessthan45$MntWines,freq=F,main="Amount of Wines age<45",
        xlab="MntofWines")
lines(density(lessthan45$MntWines))
curve(dnorm(x, mean=mean(lessthan45$MntWines), sd=sd(lessthan45$MntWines)), col="red", lwd=2, add=T)
hist(Greaterthan45$MntWines,freq=F,main="Amount of Wines age>45",
        xlab="MntofWines")
lines(density(Greaterthan45$MntWines))
curve(dnorm(x, mean=mean(Greaterthan45$MntWines), sd=sd(Greaterthan45$MntWines)), col="red", lwd=2, add
```

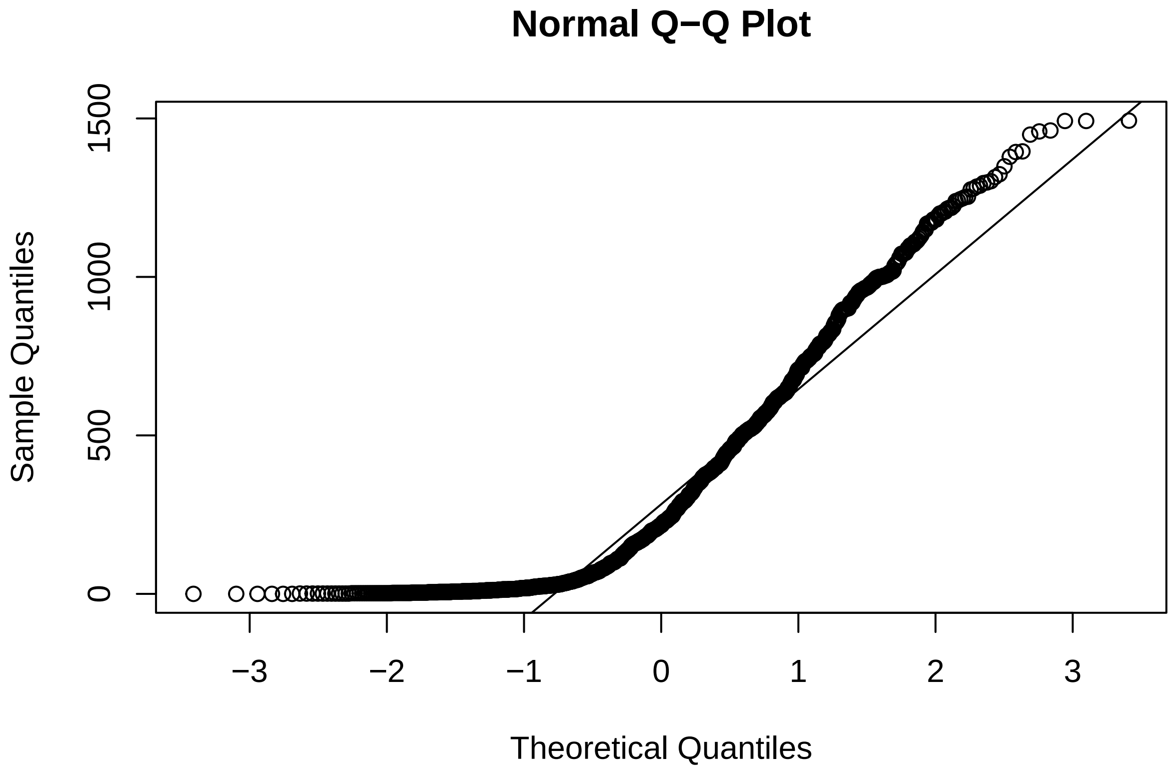


From the histogram 1 in case of Age<45 , I can say that , it is not following a normal distribution , since the histogram is right-skewed.
Also from the histogram 2 in the case of Age>45 , I can say that , it is not following a normal distribution , since the histogram is right-skewed.

```
## Normal QQ plots
qqnorm(lessthan45$MntWines)
qqline(lessthan45$MntWines)
```



```
qqnorm(Greaterthan45$MntWines)
qqline(Greaterthan45$MntWines)
```



The QQ plot in first and second graph tells me that the distribution over the points in the tail deviates from our fitted line and indicating that is not normal.

```r
library(pastecs)
stat.desc(x=lessthan45$MntWines, norm=TRUE)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range 
## 6.040000e+02 6.000000e+00 0.000000e+00 0.000000e+00 1.478000e+03 1.478000e+03 
##          sum       median         mean      SE.mean CI.mean.0.95          var 
## 1.433710e+05 5.250000e+01 2.373692e+02 1.322196e+01 2.596668e+01 1.055914e+05 
##      std.dev     coef.var     skewness     skew.2SE     kurtosis     kurt.2SE 
## 3.249483e+02 1.368957e+00 1.504641e+00 7.566941e+00 1.434096e+00 3.611979e+00 
##   normtest.W   normtest.p 
## 7.501346e-01 3.350187e-29
```

```r
stat.desc(x=Greaterthan45$MntWines, norm=TRUE)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range 
## 1.540000e+03 6.000000e+00 0.000000e+00 0.000000e+00 1.493000e+03 1.493000e+03 
##          sum       median         mean      SE.mean CI.mean.0.95          var 
## 5.123780e+05 2.170000e+02 3.327130e+02 8.627551e+00 1.692300e+01 1.146293e+05 
##      std.dev     coef.var     skewness     skew.2SE     kurtosis     kurt.2SE 
## 3.385696e+02 1.017602e+00 1.061942e+00 8.514871e+00 3.395116e-01 1.362016e+00 
##   normtest.W   normtest.p 
## 8.658626e-01 2.230324e-34
```

From descriptive statistics,
The Kurt.2SE in the case of "lessthan45" is 3.618
The Kurt.2SE in the case of "Greaterthan45" is 1.394
The kurtosis indicates that the data does not lie in the range of -1,1.
The skew.2SE in this case of lessthan45 is 7.572
The skew.2SE in the case of Greaterthan45 is 8.538
The skew.2SE indicates that the data does not lie in the range of -1,1

```r
shapiro.test(x=lessthan45$MntWines)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  lessthan45$MntWines
## W = 0.75013, p-value < 2.2e-16
```

```r
shapiro.test(x=Greaterthan45$MntWines)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  Greaterthan45$MntWines
## W = 0.86586, p-value < 2.2e-16
```

```r
#shapiro.test(x=twins$diff)
#The histogram displays a unimodal, symmetric, bell-shaped curve that matches closely with the normal d
```

The Shapiro-Wilk test has a W = 0.7501 and P-value is less than 2.2e-16 , since P-value is lesser than alpha , we reject the null hypothesis.
The Shapiro-Wilk test has a W = 0.8659 and P-value is less than 2.2e-16 , since P-value is lesser than alpha , we reject the null hypothesis.
We can confirm that our distributions do not follow a normal distribution based on Shapiro Wilk's test. Thus I would do a non-parametric version of the T-test which is Wilcoxon Rank Sum Test.

```
alpha=0.01
wilcox.test(lessthan45$MntWines,Greaterthan45$MntWines,alternative="less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  lessthan45$MntWines and Greaterthan45$MntWines
## W = 350106, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

My test statistic is 350106 and P-value is less than 2.2e-16 , thus I would reject my null hypothesis , and It would be reasonable to conclude that, people more than the age of 45 consume more wine. Hence, Nate can confidently spend on advertising more towards people whose age is greater than 45.

Part2 (CC) The marketing team of Spencer's 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 27.1.

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

```
## [1] 27.10837
```

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

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

```
mkt$sweet.cat =  cut(mkt$MntSweetProducts, c(0, 27.1, 263 ), c("Less","More")) #Cut function
table(mkt$sweet.cat)
```

```
##
## Less More
## 1173  642
```

```
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 222 670 257  24
##   More 370 240  27   5
```

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

```
##
##                0        1        2        3
##   Less 382.5983 588.1157 183.5438 18.74215
##   More 209.4017 321.8843 100.4562 10.25785
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 310.09, df = 3, p-value < 2.2e-16
```

The Chi-squared test gives us a test statistic X-squared = 310.09 and a p-value< 2.2e-16< alpha at 0.05 significance level, leading us to reject null hypothesis and we can conclude that these is a relationship between number of children each customer has and Amount of sweet he buys at Spencer's.

Let us try using a linear regression model to fit Amount of Sweet products and number of children (categorical)

```
mkt$Children = as.numeric(mkt$Children) #converting it back into numeric
lm1 = lm(mkt$Children ~ mkt$sweet.cat)
summary(lm1)
```

```
##
## Call:
## lm(formula = mkt$Children ~ mkt$sweet.cat)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07076 -0.48131 -0.07076  0.51869  2.51869
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.07076    0.01956  105.88   <2e-16 ***
## mkt$sweet.catMore -0.58945    0.03288  -17.93   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6698 on 1813 degrees of freedom
##   (418 observations deleted due to missingness)
## Multiple R-squared:  0.1506, Adjusted R-squared:  0.1501
## F-statistic: 321.3 on 1 and 1813 DF,  p-value: < 2.2e-16
```

Linear regression on Sweet.cat (Less and More) and Number of children gives us a significant relationship with test statistic 321.3 and p-value<2.2e-16 < alpha at 0.05 significance level.

The individual t-test on the categorical predictor variable(sweet.cat) also gives us a test statistic -17.93 and p-value<2e-16< alpha at 0.05 significance level implying it is a significant predictor of number of children.

Part 3a: (QQ) Let’s say We are trying to place things that sell closer to each other at Spencers’ to increase the chance of customer buying it. This approach is commonly called market basket analysis, where we try to find two products that can be sold together. Let’s predict wine purchases with number of meat purchases and try to find a linear relationship between them.
H0: beta1 = 0 (No relationship) H1: beta1 != 0 (Some relationship exists)

```r
lm1 = lm(mkt$MntWines~  mkt$MntMeatProducts)
summary(lm1)
```

```
##
## Call:
## lm(formula = mkt$MntWines ~ mkt$MntMeatProducts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1612.1  -162.4  -114.8   115.3  1244.1
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          163.6085     7.3312   22.32   <2e-16 ***
## mkt$MntMeatProducts    0.8403     0.0261   32.20   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 278.5 on 2231 degrees of freedom
## Multiple R-squared:  0.3173, Adjusted R-squared:  0.317
## F-statistic:  1037 on 1 and 2231 DF,  p-value: < 2.2e-16
```

Adjusted R-squared: 0.317

F-statistic: 1.04e+03 and p-value: <0.0000000000000002, concluding that our relationship is significant.

1.When no meat is purchased, we expect the number of wine purchases to be 163.6085
2.Withe one unit purchase of meat we expect the wine purchase to increase by 0.8403.

Using this insight, we can place the wine Aisle next to the Meat Aisle at the Super market.

(QC) A manager at spencer's claims that they should consider the number of kids a customer has and his education status before they market their wines to them. To test if this clain is true , we Consider predictors KidHome(Quantitative) and Education(Categorical) as two predictors for MntWines.Let us try to fit a mutiple linear regression equation.

H0: Beta1 = beta2 = 0
H1: Atleast one of the slopes is not zero.

```
lm1 = lm(mkt$MntWines ~ mkt$Kidhome + mkt$Education )
summary(lm1)
```

```
##
## Call:
## lm(formula = mkt$MntWines ~ mkt$Kidhome + mkt$Education)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -525.47 -172.16  -72.16  131.84 1318.89
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               343.12      20.74  16.545  < 2e-16 ***
## mkt$Kidhome              -303.31      11.24 -26.978  < 2e-16 ***
## mkt$EducationBasic       -144.90      43.73  -3.313 0.000936 ***
## mkt$EducationGraduation    76.04      21.77   3.494 0.000486 ***
## mkt$EducationMaster       127.31      24.95   5.102 3.64e-07 ***
## mkt$EducationPhD          184.36      23.89   7.718 1.77e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 285.4 on 2227 degrees of freedom
## Multiple R-squared:  0.2843, Adjusted R-squared:  0.2826
## F-statistic: 176.9 on 5 and 2227 DF,  p-value: < 2.2e-16
```

Adjusted R-squared: 0.283 and F-statistic: 177 with p-value: <0.0000000000000002 concludes that the relationship between these variables is significant.

Looking at individual t-test statistics and p-values, we see that all variables are significant for the relationship.

1.The base value of amount of wines purchased is 343.1
2.With increase in 1 kid at home the wines purchased drops by 303.3, keeping other variables constant
3.When customer has Basic education, we expect the wines purchased to reduce by 144.9, keeping other

variables constant.
4.When customer has Graduate level education, we expect the wines purchased to increase by 76, Keeping other variables constant.
5.When the customer has Master level education, we expect the wines purchased to increase by 127.3, Keeping other variables constant.
6.When the customer has a PHD level education, we expect the wines purchased to increase by 184.4, Keeping other variables constant.

Part 3b:Boot Strapping Sometimes, it can be difficult to derive a standard error or a confidence interval formula when we don't meet the assumptions. To be sure of the results obtained, we can do bootstrapping which his a more robust approach.

Boot strapping treats our sample data as a population and takes one mini sample from the sample data at a time and computes its test statistic. The distribution/ standard error of the each of the sample test statistic in each is taken as a more robust measure of the Test - statistic of our original sample. Boot strapping is a non-parametric approach and does not require assumptions of a certain test to be met. It can also be used to mitigate doubt when we are not confident about our estimates obtained using the entire sample at once.

Steps in boot strapping: -> We resample the observations with replacement X times (X is a very large number).
-> For a single sample, calculate the estimates and test statistic
-> Repeat this process for each of the X samples (calculate the parameter estimates and the test Statitics)
-> The distribution of the parameter estimates and test-statistic is obtained from X samples is the bootstrapped sampling distribution. The standard deviation of the test-statistic and parameter estimates give us the standard error in each of them.

Sources for code and reading:
1.https://aedmoodle.ufpa.br/pluginfile.php/401852/mod_resource/content/5/Material_PDF/1.Discovering%20Statistics%2
2.http://www.utstat.toronto.edu/~brunner/oldclass/appliedf12/lectures/2101f12BootstrapR.pdf 3.https://towardsdatascienc
regression-in-r-98bfe4ff5007

```
#Using the same equation from QC multiple regression

bstar = NULL # Rows of bstar will be bootstrap vectors of regression coefficients.
n = length(mkt$MntWines); B = 1000
for(draw in 1:B)
  {
#Randomly sample from the rows with replacement
  Dstar = mkt[sample(1:n,size=n,replace=T),]
  model = lm(mkt$MntWines ~ mkt$Kidhome + mkt$Education )
  bstar = rbind(bstar, coef(model))
} # Next draw
bstar[1:3,]
```

```
##      (Intercept) mkt$Kidhome mkt$EducationBasic mkt$EducationGraduation
## [1,]     343.116   -303.3149          -144.8992                76.04175
## [2,]     343.116   -303.3149          -144.8992                76.04175
## [3,]     343.116   -303.3149          -144.8992                76.04175
##      mkt$EducationMaster mkt$EducationPhD
## [1,]            127.3122         184.3562
## [2,]            127.3122         184.3562
## [3,]            127.3122         184.3562
```

We find that the parameter estimates haven't changed much after bootstrapping approach to the equation. We believe the possible reason for this could be the large sample size that is already giving us robust results.

```
#Let us use certain predictors to predict the amount of winessold.
lm1 = lm( MntWines ~ Income + NumWebPurchases + NumCatalogPurchases + NumStorePurchases + NumWebVisitsMo
summary(lm1)
```

```
##
## Call:
## lm(formula = MntWines ~ Income + NumWebPurchases + NumCatalogPurchases +
##     NumStorePurchases + NumWebVisitsMonth + Education + Kidhome +
##     MntMeatProducts, data = mkt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1460.01   -94.21   -16.63    56.17  1031.20
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -3.254e+02  2.656e+01 -12.250  < 2e-16 ***
## Income               2.268e-03  2.515e-04   9.019  < 2e-16 ***
## NumWebPurchases      2.082e+01  2.080e+00  10.011  < 2e-16 ***
## NumCatalogPurchases  2.843e+01  2.438e+00  11.663  < 2e-16 ***
## NumStorePurchases    2.868e+01  1.881e+00  15.245  < 2e-16 ***
## NumWebVisitsMonth    2.117e+01  2.553e+00   8.291  < 2e-16 ***
## EducationBasic       4.421e+01  3.208e+01   1.378   0.1683
## EducationGraduation  3.852e+01  1.587e+01   2.427   0.0153 *
## EducationMaster      9.852e+01  1.819e+01   5.418 6.70e-08 ***
## EducationPhD         1.332e+02  1.745e+01   7.637 3.30e-14 ***
## Kidhome             -6.399e+01  1.039e+01  -6.162 8.53e-10 ***
## MntMeatProducts      2.087e-01  3.077e-02   6.781 1.53e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 206.2 on 2197 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.6291, Adjusted R-squared:  0.6273
## F-statistic: 338.8 on 11 and 2197 DF,  p-value: < 2.2e-16
```

```
bstar = NULL # Rows of bstar will be bootstrap vectors of regression coefficients.
n = length(mkt$MntWines); B = 1000
for(draw in 1:B)
  {
#Randomly sample from the rows with replacement
  Dstar = mkt[sample(1:n,size=n,replace=T),]
  model = lm(MntWines ~ Income + NumWebPurchases + NumCatalogPurchases +
    NumStorePurchases + NumWebVisitsMonth + Education + Kidhome +
    MntMeatProducts , data=mkt)
  bstar = rbind(bstar, coef(model))
} # Next draw
bstar[1:3,]
```

```
##      (Intercept)      Income NumWebPurchases NumCatalogPurchases
## [1,]   -325.3659 0.002268237        20.82196            28.4348
## [2,]   -325.3659 0.002268237        20.82196            28.4348
## [3,]   -325.3659 0.002268237        20.82196            28.4348
##      NumStorePurchases NumWebVisitsMonth EducationBasic EducationGraduation
## [1,]          28.67668          21.16812       44.21493            38.51909
## [2,]          28.67668          21.16812       44.21493            38.51909
## [3,]          28.67668          21.16812       44.21493            38.51909
##      EducationMaster EducationPhD   Kidhome MntMeatProducts
## [1,]        98.52172     133.2345 -63.99202       0.2086529
## [2,]        98.52172     133.2345 -63.99202       0.2086529
## [3,]        98.52172     133.2345 -63.99202       0.2086529
```

Adjusted R-squared(Coefficent of determination) : 0.629
F-statistic: 251 and p-value: <0.0000000000000002, indicating the overall equation is significant.

Interpreting Parameter estimates after bootstrapping:
1.The base amount of wines purchased when all the predictor variables are zero is -325.
2.One Unit change in income changes(increase) the expected amount of wines purchased by 0.0022, keeping other predictor variables constant.
3.One Unit change in NumWebPurchases changes(increase) the expected amount of wines purchased by 20.822, keeping other predictor variables constant.
4.One Unit change in NumCatalogPurchases changes(increase) the expected amount of wines purchased by 28.4348, keeping other predictor variables constant.
5.One Unit change in NumStorePurchases changes(increase) the expected amount of wines purchased by 28.6767, keeping other predictor variables constant.
6.One Unit change in NumWebVisitsMonth changes the expected amount of wines purchased by 21.1681, keeping other predictor variables constant.
7.Having EducationBasic changes(decrease) the expected amount of wines purchased by 44.2149, keeping other predictor variables constant.
8.Having EducationMaster changes(increase) the expected amount of wines purchased by 98.6029, keeping other predictor variables constant.
9.Having EducationPhd changes(increase) the expected amount of wines purchased by 143.601, keeping other predictor variables constant.
10.Having a KidHome changes(decrease) the expected amount of wines purchased by 193.335 , keeping other predictor variables constant.
11.Having Marital_StatusMarried changes(decrease) the expected amount of wines purchased by -9.95365 , keeping other predictor variables constant.
12.Having Marital_StatusSingle changes(decrease) the expected amount of wines purchased by -12.0183, keeping other predictor variables constant.
13.Having Marital_StatusTogether changes(decrease) the expected amount of wines purchased by -9.60396 , keeping other predictor variables constant.
14.Having Marital_StatusWidow changes(decrease) the expected amount of wines purchased by -22.0109, keeping other predictor variables constant.

From the inference of the above linear regression equation, We find our key insights to be that customer's with PHD's purchase more wines. Customers with a Kid home would purchase less wines compared to the ones without a kid.

Summary:

From this analysis we try to find insights to approach marketing in a better manner. From the first hypothesis(CQ), we concluded using ANOVA that mean of Number of weborders of different marital groups is the same and we need not pay attention towards one particular group for advertising online.For the next hypothesis(CQ), by performing wilcoxon we found out that age>45 customers could be potential market for wines and advertising to them would be effective. Second hypothesis(CC) tests to see if the number of children(C) a customer has and amount of sweet products(C) he/she purchases is dependent. From Chi-square test of independence we found that they do have a relationship, helping us to guage the expectations of customers with kids. In the third hypothesis, we find a significant linear relationship between Meat products and wines purchased, helping us to validate if we should place these two products nearby in the supermarket. In the same part, we do a multiple regression between amount of wines purchased and kidHome and educational status and find a significant relations ship between these variables too. Performing bootstrapping has not changed much in the estimates, probably because of the large sample size.

Additional Work interpretation:

##As an analyst,I want to know what factors play a role in purchasing wines based on Customer panel data.So the natural thing to do is run a regression model across all my variables and interpret the results to answer this question.

Full Model: My F-statistic is 54.5 and P-value is significantly low to reject my null hypothesis.
I am doing a Multiple regression with all my predictor Variables and I found that Income, KidHome,Teenhome, NumWebPurchases, NumCatalogPurchases,NumWebVisitsMonth,Education_Master, Education_PhD, Marital_Status_Widow are significant in predicting AmountofWine Purchased.

To choose my best model with the right parameters, I explore various Model Selection methods to see which are the more significant parameters which can help me in predicting better outcome variable.

#1.Forward selection: We start with no predictors in the model, iteratively add the most contributive predictors, and stops when the improvement is no longer statistically significant.
When I did Forward Selection, At AIC=6570.72 ,the value of AIC doesn't get reduced.

The F-statistic is 115 and P-value is really low , less than 0.05 , hence I reject my null hypothesis.

Thus when we prefer this Forward Selection model , with right parameters being Income ,NumWebPurchases,NumCatalogPurchases,NumWebVisitsMonth,Education_PhD, Kidhome,Teenhome,NumStorePurchase,Education_Ma
and Marital_Status_Married.

##Backward Elimination:
Backward selection (or backward elimination), which starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant

At AIC=6570.36,the value of AIC doesn't get reduced.

The F-statistic is 106 and P-value is really low , less than 0.05 , hence I reject my null hypothesis.

Thus we prefer this Backward Elimination model , with right significant parameters being Income ,NumWebPurchases,NumCatalogPurchases,
NumWebVisitsMonth,Education_PhD, Kidhome,Teenhome,NumStorePurchases,Education_Master,Marital_Status_Widow
and Marital_Status_Married.

We see Education_Basic,Education_Graduation are some new parameters being added in this model. Also we see parameter like Marital_Status_Married being eliminated from the Forward Selection Model.

##Stepwise selection: Stepwise selection which is a combination of forward and backward selections. You start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection).

At AIC=6570.72,the value of AIC doesn't get reduced.

The F-statistic is 115 and P-value is really low , less than 0.05 , hence I reject my null hypothesis.

Thus we prefer this Stepwise Selection Model , with right parameters being Income ,NumWebPurchases,NumCatalogPurchases,NumWebVisitsMonth,Education_PhD,Kidhome,Teenhome,NumStorePurchases,Education_Master,Marital_Status_Widow and Marital_Status_Married.

We see that Stepwise Selection Model gives us the same set of significant predictors like we see in Backward Elimination.

Conclusion:
Thus as an analyst I would recommend my manager these are my most significant factors which play a role in Predicting Wine Purchases.Thus we prefer this Stepwise Selection Model , with right parameters being Income,NumWebPurchases,NumCatalogPurchases,NumWebVisitsMonth,Education_PhD,Kidhome,Teenhome,NumStorePurch
and Marital_Status_Married.