Starbucks Data Analysis
Phase 1

Loading libraries & checking for missing values or null values:

Code:

```
library(readr)
library(tidyr)
library(stringr)
library(plyr)
library(dplyr)
library(gmodels)
library(corrplot)
library(RVAideMemoire)

#Setup Working Directory
setwd("/Users/manpreetkaurgurtatta/Library/CloudStorage/Box-Box/Sem 1/Data Stats/Assignments/Assignment 3- ANOVAs, Correlations, and Visualizations")

#Loading and Reading the Dataset
Starbucks <- read.csv("Starbucks_data-2.csv")
View(Starbucks)
head(Starbucks)

##Checking Dimensions of Dataset
dim(Starbucks)
#122 rows  23 columns

#Checking Variable Names
names(Starbucks)

#Checking Data structure
str(Starbucks)

#Renaming files to simplify data
names(Starbucks) <- c('Timestamp', 'Gender', 'Age', 'Status', 'Income', 'Visits', 'ServiceMode', 'Timespend', 'Nearestlocation', 'Membershipcard',
                      'Frequentpurchase', 'SpendPurchase', 'QualityRate', 'PriceRate', 'PromoRate', 'Ambiance', 'WiFiRate', 'ServiceRate',
                      'chooseRate', 'PromoMethod', 'loyalty', 'BeforePromoSatRate', 'AfterPromoSatRate')

Starbucks_Updated <- Starbucks
str(Starbucks_Updated)

#Checking for Missing Data
colnames(Starbucks)
sum(is.na(Starbucks))

#If Missing data, use listwise deletion.
Starbucks_Updated <- na.omit(Starbucks)
sum(is.na(Starbucks_Updated))
View(Starbucks_Updated)

#**********************************************************************************************************************
```

****

1. 3. Are you currently....?(question 3) affect How would you rate the ambiance at Starbucks? (lighting, music, etc...)(question 15)?

   Code:

```r
#Q1. 3.Are you currently....?(question 3) affect How would you rate the ambiance at Starbucks? (lighting, music, etc...)
#(question 15)?

Starbucks_Updated$Status <- revalue(Starbucks_Updated$Status, c("Employed"=0))
Starbucks_Updated$Status <- revalue(Starbucks_Updated$Status, c("Housewife"=1))
Starbucks_Updated$Status <- revalue(Starbucks_Updated$Status, c("Self-employed"=2))
Starbucks_Updated$Status <- revalue(Starbucks_Updated$Status, c("Student"=3))

Starbucks_Updated$Status = as.numeric(Starbucks_Updated$Status)

head(Starbucks_Updated$Status)
View(Starbucks_Updated)

#Test for Normality
library(RVAideMemoire)
shapiro.test(Starbucks_Updated$Ambiance)
shapiro.test(Starbucks_Updated$Status)

#Performing Kruskal-Wallis test since the given data is not normal
kruskal.test(Ambiance~Status, data = Starbucks_Updated)

#CONCLUSION:We did a Normality test for starbucks's Ambiance and Status where the p value was 1.027e-08 and 1.33e-13
#respectively, which is less than 0.05. meaning that the data is not normal. In furtherance, we performed Kruskal-Wallis
#test for Ambiance and Status where the p-value is 0.2062, which is greater than 0.05. This proves that we accept the null
#hypothesis which also means that there is no relation between ambiance and status.
```

Console:

```
> #Q1. 3.Are you currently....?(question 3) affect How would you rate the ambiance at Starbucks? (lighting, music, etc...)
> #(question 15)?
>
> Starbucks_Updated$Status <- revalue(Starbucks_Updated$Status, c("Employed"=0))
> Starbucks_Updated$Status <- revalue(Starbucks_Updated$Status, c("Housewife"=1))
> Starbucks_Updated$Status <- revalue(Starbucks_Updated$Status, c("Self-employed"=2))
> Starbucks_Updated$Status <- revalue(Starbucks_Updated$Status, c("Student"=3))
>
> Starbucks_Updated$Status = as.numeric(Starbucks_Updated$Status)
>
> head(Starbucks_Updated$Status)
[1] 3 3 0 3 3 3
> View(Starbucks_Updated)
>
> #Test for Normality
> library(RVAideMemoire)
> shapiro.test(Starbucks_Updated$Ambiance)

        Shapiro-Wilk normality test

data:  Starbucks_Updated$Ambiance
W = 0.86417, p-value = 1.027e-08

> shapiro.test(Starbucks_Updated$Status)

        Shapiro-Wilk normality test

data:  Starbucks_Updated$Status
W = 0.70868, p-value = 1.33e-13

>
> #Performing Kruskal-Wallis test since the given data is not normal
> kruskal.test(Ambiance~Status, data = Starbucks_Updated)

        Kruskal-Wallis rank sum test

data:  Ambiance by Status
Kruskal-Wallis chi-squared = 4.5688, df = 3, p-value = 0.2062

>
> #CONCLUSION:We did a Normality test for starbucks's Ambiance and Status where the p value was 1.027e-08 and 1.33e-13
> #respectively, which is less than 0.05. meaning that the data is not normal. In furtherance, we performed Kruskal-Wallis
> #test for Ambiance and Status where the p-value is 0.2062, which is greater than 0.05. This proves that we accept the null
> #hypothesis which also means that there is no relation between ambiance and status.
```

Conclusion: We did a Normality test for starbucks's Ambiance and Status where the p value was 1.027e-08 and 1.33e-13 respectively, which is less than 0.05. meaning that the data is not normal. In furtherance, we performed Kruskal-Wallis test for Ambiance and Status where the p-value is 0.2062, which is greater than 0.05. This proves that we accept the null hypothesis which also means that there is no relation between ambiance and status.

****

2. What is the relationship or association between How would you rate the service at Starbucks? (Promptness, friendliness, etc..) (question 17) and price range (question 13)?

Code:

```
#Q2. What is the relationship or association between How would you rate the service at Starbucks? (Promptness, friendliness, etc..)
#(question 17) and price range (question 13)?

#Solution

#Test for Normality
library(RVAideMemoire)
shapiro.test(Starbucks_Updated$ServiceRate)
shapiro.test(as.numeric(Starbucks_Updated$PriceRate))

#Data is not Normal.

#Checking Spearman Correlations:
cor.test(Starbucks_Updated$ServiceRate,Starbucks_Updated$PriceRate, method="spearman" , exact = FALSE)

#CONCLUSION: For this question, we did a normality test on starbucks' Service Rate and Price rate, where the p-value is 8.837e-09
#and 1.131e-06 respectively proving that the given data is not normal. Hence, we then checked Spearman Correlations between starbucks'
#Service Rate and Price rate, where the p-value is 0.004018, which is less than 0.05, meaning for us to reject the null hypothesis and
#accept the alternate hypothesis. Overall, there is a relationship/co-relation between starbucks' Service Rate and Price rate.

#**********************************************************************************************************************
```

Console:

```
> #Q2. What is the relationship or association between How would you rate the service at Starbucks? (Promptness, friendliness, etc..)
> #(question 17) and price range (question 13)?
>
> #Solution
>
> #Test for Normality
> library(RVAideMemoire)
> shapiro.test(Starbucks_Updated$ServiceRate)

        Shapiro-Wilk normality test

data:  Starbucks_Updated$ServiceRate
W = 0.86258, p-value = 8.837e-09

> shapiro.test(as.numeric(Starbucks_Updated$PriceRate))

        Shapiro-Wilk normality test

data:  as.numeric(Starbucks_Updated$PriceRate)
W = 0.90838, p-value = 1.131e-06

>
> #Data is not Normal.
>
> #Checking Spearman Correlations:
> cor.test(Starbucks_Updated$ServiceRate,Starbucks_Updated$PriceRate, method="spearman" , exact = FALSE)

        Spearman's rank correlation rho

data:  Starbucks_Updated$ServiceRate and Starbucks_Updated$PriceRate
S = 170968, p-value = 0.004018
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2697928

>
> #CONCLUSION: For this question, we did a normality test on starbucks' Service Rate and Price rate, where the p-value is 8.837e-09
> #and 1.131e-06 respectively proving that the given data is not normal. Hence, we then checked Spearman Correlations between starbucks'
> #Service Rate and Price rate, where the p-value is 0.004018, which is less than 0.05, meaning for us to reject the null hypothesis and
> #accept the alternate hypothesis. Overall, there is a relationship/co-relation between starbucks' Service Rate and Price rate.
>
> #**********************************************************************************************************************
>
```

CONCLUSION: For this question, we did a normality test on starbucks' Service Rate and Price rate, where the p-value is 8.837e-09 and 1.131e-06 respectively proving that the given data is not normal. Hence, we then checked Spearman Correlations between starbucks' Service Rate and Price rate, where the p-value is 0.004018, which is less than 0.05, meaning for us to reject the null hypothesis and accept the alternate hypothesis. Overall, there is a relationship/co-relation between starbucks' Service Rate and Price rate.

****

3. Show and explain a visualization of correlations of questions 12, 13, 14, 15, and 16. Hint: create one visualization showing all the correlations.

Code:

```
#Q3.Show and explain a visualization of correlations of questions 12, 13, 14, 15, and 16. Hint:  create one visualization showing
#all the correlations.

names(Starbucks_Updated)

DataViz <- Starbucks_Updated[, c(13,14,15,16,17)]

str(DataViz)

#Checking Spearman Correlations

library(corrplot)

corrplot(cor(DataViz, method = "spearman"))

corrplot(cor(DataViz, method = "spearman"), method="number")


#**********************************************************************************************************************
```

Console and Data visualization:

```
> names(Starbucks_Updated)
 [1] "Timestamp"          "Gender"             "Age"                "Status"             "Income"             "Visits"
            "ServiceMode"
 [8] "Timespend"          "Nearestlocation"    "Membershipcard"     "Frequentpurchase"   "SpendPurchase"      "Quality
Rate"          "PriceRate"
[15] "PromoRate"          "Ambiance"           "WiFiRate"           "ServiceRate"        "chooseRate"         "PromoMe
thod"          "loyalty"
[22] "BeforePromoSatRate" "AfterPromoSatRate"
>
> DataViz <- Starbucks_Updated[, c(13,14,15,16,17)]
>
> str(DataViz)
'data.frame':   112 obs. of  5 variables:
 $ QualityRate: int  4 4 4 2 3 4 5 4 5 4 ...
 $ PriceRate  : int  3 3 3 1 3 3 5 2 4 3 ...
 $ PromoRate  : int  5 4 4 4 4 5 5 3 4 3 ...
 $ Ambiance   : int  5 4 4 3 2 5 5 3 4 4 ...
 $ WiFiRate   : int  4 4 4 3 2 4 3 3 4 3 ...
>
> #Checking Spearman Correlations
>
> library(corrplot)
>
> corrplot(cor(DataViz, method = "spearman"))
>
> corrplot(cor(DataViz, method = "spearman"), method="number")
>
```

Explanation: For this question, we created a data visualization using Spearman Correlations to find relationship between QualityRate, Pricerate, Promorate, ambiance and wifirate and discovered the following the co-relations:

- The co-relation between QualityRate and PriceRate is 0.43.
- The co-relation between QualityRate and PromoRate is 0.16.
- The co-relation between QualityRate and Ambiance is 0.56.
- The co-relation between QualityRate and WiFiRate is 0.27.
- The co-relation between PriceRate and PromoRate is 0.07.
- The co-relation between PriceRate and Ambiance is 0.32.
- The co-relation between PriceRate and WiFiRate is 0.22.
- The co-relation between PromoRate and Ambiance is 0.39.
- The co-relation between PromoRate and WiFirate is 0.47.

****

4. Create a visualization and answer the questions below, which will provide an interesting story or insight within this data.
   a. Who is your audience?

b. What is the application insight?
c. What does this application insight mean for the audience? Why is it important for the audience to know?

Code:

```
#Q4.Create a visualization and answer the questions below, which will provide an interesting story or insight within this data.
#a.  Who is your audience?
#b.  What is the application insight?
#c.  What does this application insight mean for the audience?  Why is it important for the audience to know?

#Note- Earlier we revalued category of people to numeric values where:
#Employed = 0
#Housewife = 1
#Self-employed = 2
#Student = 3

# Stacked Bar Plot for Timespent by different category of people at Starbucks
counts <- table(Starbucks_Updated$Timespend, Starbucks_Updated$Status)
barplot(counts, main="Timespent by different set of people at Starbucks",
        xlab="Category of People", col=c("yellow","purple", "red", "green","pink"),
        legend = rownames(counts) , beside=TRUE)


#*****************************************************************************************************************************
```

Console:

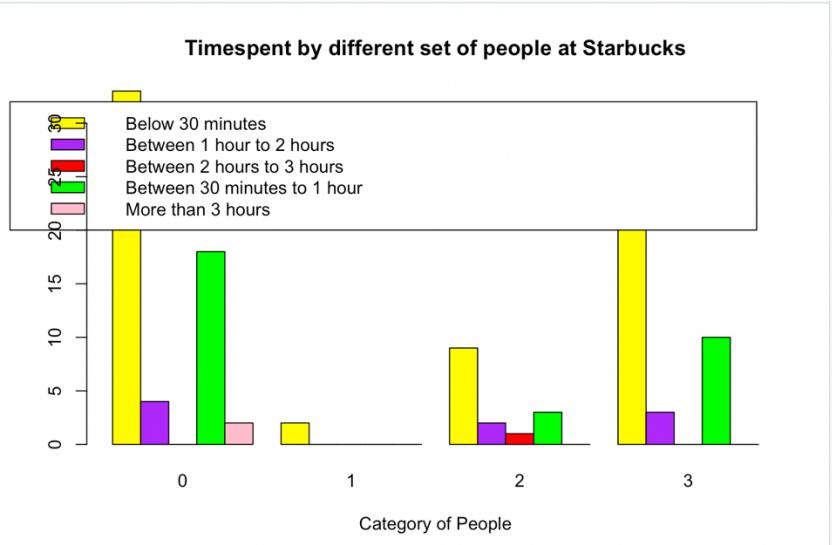
```
> names(Starbucks_Updated)
 [1] "Timestamp"          "Gender"             "Age"                "Status"             "Income"             "Visits"             "ServiceMode"

 [8] "Timespend"          "Nearestlocation"    "Membershipcard"     "Frequentpurchase"   "SpendPurchase"      "QualityRate"        "PriceRate"

[15] "PromoRate"          "Ambiance"           "WiFiRate"           "ServiceRate"        "chooseRate"         "PromoMethod"        "loyalty"

[22] "BeforePromoSatRate" "AfterPromoSatRate"
>
> DataViz <- Starbucks_Updated[, c(13,14,15,16,17)]
>
> str(DataViz)
'data.frame':   112 obs. of  5 variables:
 $ QualityRate: int  4 4 4 2 3 4 5 4 5 4 ...
 $ PriceRate  : int  3 3 3 1 3 3 5 2 4 3 ...
 $ PromoRate  : int  5 4 4 4 4 5 5 3 4 3 ...
 $ Ambiance   : int  5 4 4 3 2 5 5 3 4 4 ...
 $ WiFiRate   : int  4 4 4 3 2 4 3 3 4 3 ...
>
> #Checking Spearman Correlations
>
> library(corrplot)
>
> corrplot(cor(DataViz, method = "spearman"))
>
> corrplot(cor(DataViz, method = "spearman"), method="number")
> #Q4.Create a visualization and answer the questions below, which will provide an interesting story or insight within this data.
> #a.  Who is your audience?
> #b.  What is the application insight?
> #c.  What does this application insight mean for the audience?  Why is it important for the audience to know?
>
> #Note- Earlier we revalued category of people to numeric values where:
> #Employed = 0
> #Housewife = 1
> #Self-employed = 2
> #Student = 3
>
> # Stacked Bar Plot for Timespent by different category of people at Starbucks
> counts <- table(Starbucks_Updated$Timespend, Starbucks_Updated$Status)
> barplot(counts, main="Timespent by different set of people at Starbucks",
+         xlab="Category of People", col=c("yellow","purple", "red", "green","pink"),
+         legend = rownames(counts) , beside=TRUE)
>
>
> #*****************************************************************************************************************************
>
```

Data visualization:



a) Who is your audience?

The audience for this presentation is Starbucks' management team.

b) What is the application insight?

The visualization clearly indicates that only 2% of employed people spend more than 3 hours at starbucks. Also, it shows that among all the categories, housewives rarely visit the café showing 2%. We can also see that Most people spend less than 30 minutes at starbucks.

c) What does this application insight mean for the audience? Why is it important for the audience to know?

This visualization can be useful for the starbucks management to analyze different categories of customers. Using this visualization they can identify their target audience and come up with offers and discounts specific to that category.

****