market\_analysis.R

admin

2023-03-02

#import the data sets  
library(readr)  
  
Subscription <- read\_csv("Subscription.csv")

## Rows: 298 Columns: 7  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): gender, ownHome, subscribe, Segment  
## dbl (3): age, income, kids  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

amusement <- read\_csv("amusement.csv")

## Rows: 500 Columns: 8  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): weekend  
## dbl (7): num.child, distance, rides, games, wait, clean, overall  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# PART1 - Working with Subscription data set  
# QUESTION 1  
  
# i. number of rows and columns  
dim(Subscription)

## [1] 298 7

# ii. data type of each variable  
  
str(Subscription)

## spc\_tbl\_ [298 × 7] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ age : num [1:298] 23.3 25.9 24.7 61.3 24.2 ...  
## $ gender : chr [1:298] "Male" "Female" "Female" "Female" ...  
## $ income : num [1:298] 11985 12406 12545 15347 15493 ...  
## $ kids : num [1:298] 1 0 1 0 1 0 0 3 2 0 ...  
## $ ownHome : chr [1:298] "ownNo" "ownNo" "ownNo" "ownYes" ...  
## $ subscribe: chr [1:298] "subNo" "subNo" "subYes" "subNo" ...  
## $ Segment : chr [1:298] "Urban hip" "Urban hip" "Urban hip" "Travelers" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. age = col\_double(),  
## .. gender = col\_character(),  
## .. income = col\_double(),  
## .. kids = col\_double(),  
## .. ownHome = col\_character(),  
## .. subscribe = col\_character(),  
## .. Segment = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

# iii. first and last 3 rows  
head(Subscription, n = 3)

## # A tibble: 3 × 7  
## age gender income kids ownHome subscribe Segment   
## <dbl> <chr> <dbl> <dbl> <chr> <chr> <chr>   
## 1 23.3 Male 11985. 1 ownNo subNo Urban hip  
## 2 25.9 Female 12406. 0 ownNo subNo Urban hip  
## 3 24.7 Female 12545. 1 ownNo subYes Urban hip

tail(Subscription, n = 3)

## # A tibble: 3 × 7  
## age gender income kids ownHome subscribe Segment   
## <dbl> <chr> <dbl> <dbl> <chr> <chr> <chr>   
## 1 54.9 Female 106430. 0 ownYes subNo Travelers  
## 2 64.6 Male 113457. 0 ownNo subNo Travelers  
## 3 58.9 Male 114278. 0 ownYes subNo Travelers

# iv. Check for any missing values  
any(is.na(Subscription))

## [1] FALSE

# v. converting character type variables to factor type  
Subscription$gender <- as.factor(Subscription$gender)  
Subscription$ownHome <- as.factor(Subscription$ownHome)  
Subscription$subscribe <- as.factor(Subscription$subscribe)  
Subscription$Segment <- as.factor(Subscription$Segment)  
  
class(Subscription$gender)

## [1] "factor"

# vi. Summary statistics on the data  
summary(Subscription)

## age gender income kids ownHome   
## Min. :19.26 Female:157 Min. : 11985 Min. :0.000 ownNo :159   
## 1st Qu.:32.97 Male :141 1st Qu.: 39966 1st Qu.:0.000 ownYes:139   
## Median :39.43 Median : 52200 Median :1.000   
## Mean :41.12 Mean : 51298 Mean :1.279   
## 3rd Qu.:47.85 3rd Qu.: 61437 3rd Qu.:2.000   
## Max. :80.49 Max. :114278 Max. :7.000   
## subscribe Segment   
## subNo :258 Moving up : 70   
## subYes: 40 Suburb mix:100   
## Travelers : 78   
## Urban hip : 50   
##   
##

# QUESTION TWO  
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

subscribed <- Subscription %>%  
 filter(subscribe == "subYes")  
  
head(subscribed)

## # A tibble: 6 × 7  
## age gender income kids ownHome subscribe Segment   
## <dbl> <fct> <dbl> <dbl> <fct> <fct> <fct>   
## 1 24.7 Female 12545. 1 ownNo subYes Urban hip  
## 2 22.8 Female 15911. 0 ownNo subYes Urban hip  
## 3 23.2 Male 17510. 1 ownNo subYes Urban hip  
## 4 21.2 Male 18419. 1 ownYes subYes Urban hip  
## 5 24.7 Male 18458. 1 ownNo subYes Urban hip  
## 6 25.2 Female 20126. 2 ownNo subYes Urban hip

summary(subscribed)

## age gender income kids ownHome   
## Min. :21.24 Female:21 Min. :12545 Min. :0.00 ownNo :22   
## 1st Qu.:29.69 Male :19 1st Qu.:35155 1st Qu.:0.00 ownYes:18   
## Median :36.57 Median :49107 Median :1.00   
## Mean :39.65 Mean :45934 Mean :1.25   
## 3rd Qu.:47.02 3rd Qu.:56145 3rd Qu.:2.00   
## Max. :80.49 Max. :82077 Max. :6.00   
## subscribe Segment   
## subNo : 0 Moving up :14   
## subYes:40 Suburb mix: 6   
## Travelers :10   
## Urban hip :10   
##   
##

not\_subscribed <- Subscription %>%  
 filter(subscribe == "subNo")  
  
head(not\_subscribed)

## # A tibble: 6 × 7  
## age gender income kids ownHome subscribe Segment   
## <dbl> <fct> <dbl> <dbl> <fct> <fct> <fct>   
## 1 23.3 Male 11985. 1 ownNo subNo Urban hip  
## 2 25.9 Female 12406. 0 ownNo subNo Urban hip  
## 3 61.3 Female 15347. 0 ownYes subNo Travelers  
## 4 24.2 Male 15493. 1 ownNo subNo Urban hip  
## 5 22.3 Male 16341. 0 ownNo subNo Urban hip  
## 6 21.4 Male 16646. 3 ownNo subNo Urban hip

summary(not\_subscribed)

## age gender income kids ownHome   
## Min. :19.26 Female:136 Min. : 11985 Min. :0.000 ownNo :137   
## 1st Qu.:33.16 Male :122 1st Qu.: 40637 1st Qu.:0.000 ownYes:121   
## Median :39.78 Median : 52883 Median :1.000   
## Mean :41.35 Mean : 52130 Mean :1.283   
## 3rd Qu.:47.85 3rd Qu.: 62830 3rd Qu.:2.000   
## Max. :78.20 Max. :114278 Max. :7.000   
## subscribe Segment   
## subNo :258 Moving up :56   
## subYes: 0 Suburb mix:94   
## Travelers :68   
## Urban hip :40   
##   
##

# QUESTION 3  
# create cross-tables and proportion-based cross tables between  
  
# i. subscription status and home ownership  
home\_owner\_table <- table(Subscription$subscribe, Subscription$ownHome)  
home\_owner\_table

##   
## ownNo ownYes  
## subNo 137 121  
## subYes 22 18

home\_owner\_prop\_table <- prop.table(home\_owner\_table, margin = 1)  
  
home\_owner\_prop\_table

##   
## ownNo ownYes  
## subNo 0.5310078 0.4689922  
## subYes 0.5500000 0.4500000

home\_owner\_prop\_table\_2 <- prop.table(home\_owner\_table, margin = 2)  
home\_owner\_prop\_table\_2

##   
## ownNo ownYes  
## subNo 0.8616352 0.8705036  
## subYes 0.1383648 0.1294964

# ii. subscription status and gender  
gender\_cross\_table <- table(Subscription$subscribe, Subscription$gender)  
gender\_cross\_table

##   
## Female Male  
## subNo 136 122  
## subYes 21 19

gender\_prop\_table <- prop.table(gender\_cross\_table, margin = 1)  
gender\_prop\_table

##   
## Female Male  
## subNo 0.5271318 0.4728682  
## subYes 0.5250000 0.4750000

gender\_prop\_table\_2 <- prop.table(gender\_cross\_table, margin = 2)  
gender\_prop\_table\_2

##   
## Female Male  
## subNo 0.8662420 0.8652482  
## subYes 0.1337580 0.1347518

# QUESTION 4  
levels(Subscription$Segment)

## [1] "Moving up" "Suburb mix" "Travelers" "Urban hip"

# a) what the average income, average number of kids, and average age is for each segment  
  
# segment - Moving up  
Moving\_up\_stats <- Subscription %>%  
 filter(Segment == "Moving up") %>%  
 summarize(average\_income = mean(income), average\_no\_kids = mean(kids),   
 average\_age = mean(age))  
  
Moving\_up\_stats

## # A tibble: 1 × 3  
## average\_income average\_no\_kids average\_age  
## <dbl> <dbl> <dbl>  
## 1 53091. 1.91 36.3

# Suburb mix segement  
Sburb\_mix\_stats <- Subscription %>%  
 filter(Segment == "Suburb mix") %>%  
 summarize(average\_income = mean(income), average\_no\_kids = mean(kids),   
 average\_age = mean(age))  
  
Sburb\_mix\_stats

## # A tibble: 1 × 3  
## average\_income average\_no\_kids average\_age  
## <dbl> <dbl> <dbl>  
## 1 55034. 1.92 39.9

# Travelers Segment  
Travelers\_stats <- Subscription %>%  
 filter(Segment == "Travelers") %>%  
 summarize(average\_income = mean(income), average\_no\_kids = mean(kids),   
 average\_age = mean(age))  
  
Travelers\_stats

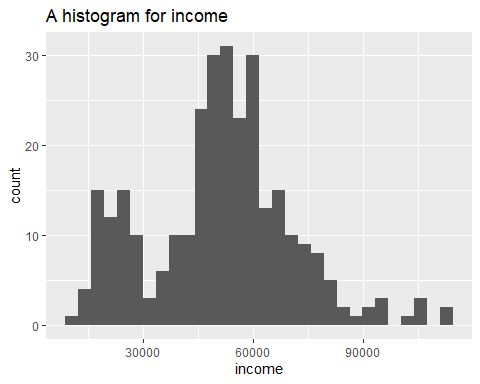
## # A tibble: 1 × 3  
## average\_income average\_no\_kids average\_age  
## <dbl> <dbl> <dbl>  
## 1 63885. 0 58.0

# Urban hip segment  
Urban\_hip\_stats <- Subscription %>%  
 filter(Segment == "Urban hip") %>%  
 summarize(average\_income = mean(income), average\_no\_kids = mean(kids),   
 average\_age = mean(age))  
  
Urban\_hip\_stats

## # A tibble: 1 × 3  
## average\_income average\_no\_kids average\_age  
## <dbl> <dbl> <dbl>  
## 1 21682. 1.1 23.9

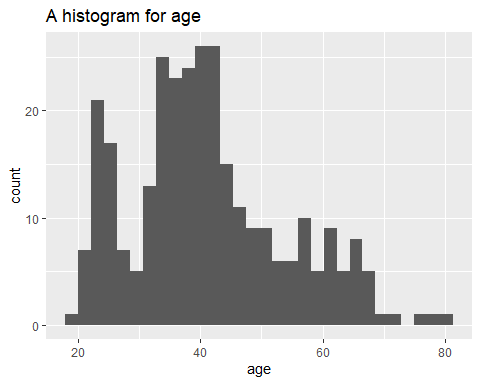
# QUESTION FIVE  
  
# i) a histogram for income and discuss its distribution  
library(ggplot2)  
  
ggplot(Subscription,aes(income)) + geom\_histogram() + labs(  
 title = "A histogram for income "  
)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

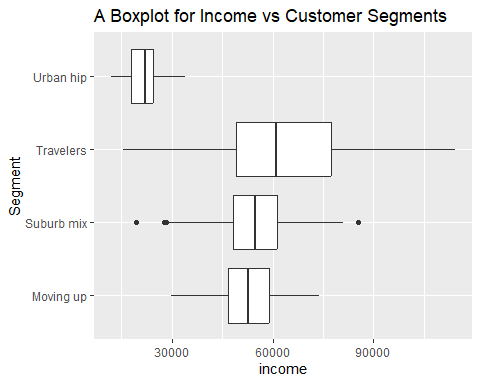


# ii) a histogram for age  
ggplot(Subscription,aes(age)) + geom\_histogram() + labs(  
 title = "A histogram for age "  
)

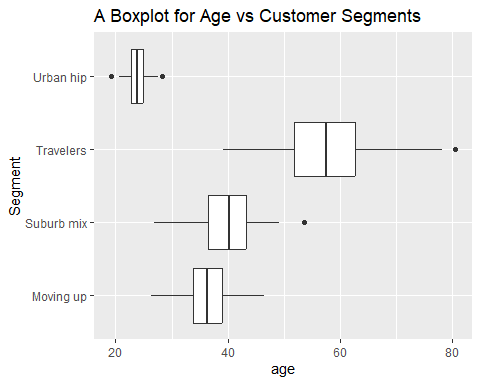
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# iii) a boxplot for income vs. customer segments   
ggplot(Subscription, aes(x = income, y = Segment )) + geom\_boxplot() +  
 labs(title = "A Boxplot for Income vs Customer Segments")



# iv) a boxplot for age vs. customer segments  
ggplot(Subscription, aes(x = age, y = Segment )) + geom\_boxplot() +  
 labs(title = "A Boxplot for Age vs Customer Segments")



# PART 2 - Relationship between continuous variables  
# Using the amusements.csv data set  
  
# QUESTION 7  
  
# i) the number of rows (respondents) and number of columns (variables)  
dim(amusement)

## [1] 500 8

# ii) the data types of each variable  
str(amusement)

## spc\_tbl\_ [500 × 8] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ weekend : chr [1:500] "yes" "yes" "no" "yes" ...  
## $ num.child: num [1:500] 0 2 1 0 4 5 1 0 0 3 ...  
## $ distance : num [1:500] 114.6 27 63.3 25.9 54.7 ...  
## $ rides : num [1:500] 87 87 85 88 84 81 77 82 90 88 ...  
## $ games : num [1:500] 73 78 80 72 87 79 73 70 88 86 ...  
## $ wait : num [1:500] 60 76 70 66 74 48 58 70 79 55 ...  
## $ clean : num [1:500] 89 87 88 89 87 79 85 83 95 88 ...  
## $ overall : num [1:500] 47 65 61 37 68 27 40 30 58 36 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. weekend = col\_character(),  
## .. num.child = col\_double(),  
## .. distance = col\_double(),  
## .. rides = col\_double(),  
## .. games = col\_double(),  
## .. wait = col\_double(),  
## .. clean = col\_double(),  
## .. overall = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

# iii) whether there is any missing value in the data set  
any(is.na(amusement))

## [1] FALSE

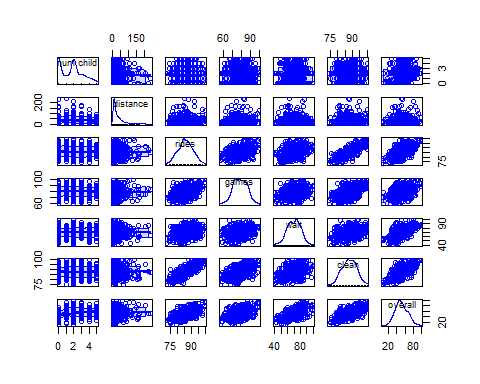
# iv) run summary statistics of the data set.  
summary(amusement)

## weekend num.child distance rides   
## Length:500 Min. :0.000 Min. : 0.5267 Min. : 72.00   
## Class :character 1st Qu.:0.000 1st Qu.: 10.3181 1st Qu.: 82.00   
## Mode :character Median :2.000 Median : 19.0191 Median : 86.00   
## Mean :1.738 Mean : 31.0475 Mean : 85.85   
## 3rd Qu.:3.000 3rd Qu.: 39.5821 3rd Qu.: 90.00   
## Max. :5.000 Max. :239.1921 Max. :100.00   
## games wait clean overall   
## Min. : 57.00 Min. : 40.0 Min. : 74.0 Min. : 6.00   
## 1st Qu.: 73.00 1st Qu.: 62.0 1st Qu.: 84.0 1st Qu.: 40.00   
## Median : 78.00 Median : 70.0 Median : 88.0 Median : 50.00   
## Mean : 78.67 Mean : 69.9 Mean : 87.9 Mean : 51.26   
## 3rd Qu.: 85.00 3rd Qu.: 77.0 3rd Qu.: 91.0 3rd Qu.: 62.00   
## Max. :100.00 Max. :100.0 Max. :100.0 Max. :100.00

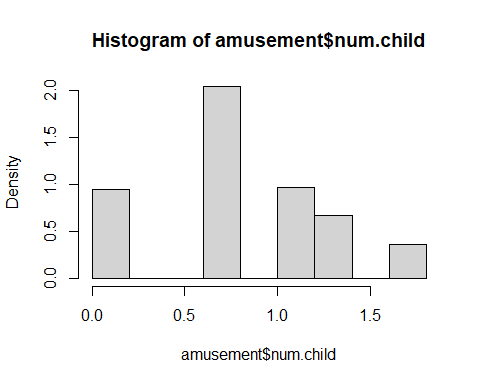
# QUESTION 8  
library(car)

## Loading required package: carData  
##   
## Attaching package: 'car'  
##   
## The following object is masked from 'package:dplyr':  
##   
## recode

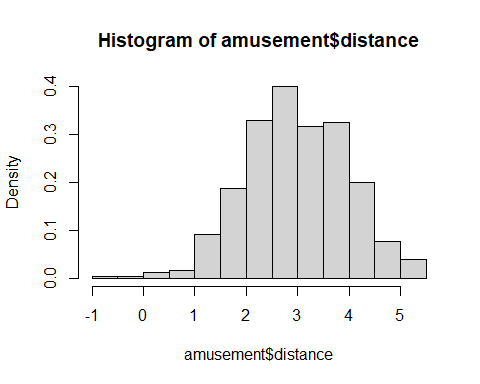
scatterplotMatrix(formula=~num.child+distance+rides+games+wait+clean+overall,data=amusement)



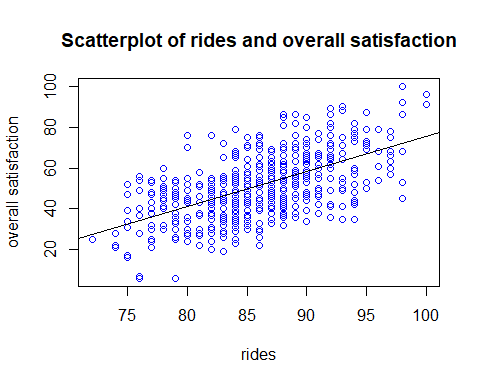
# transform variables which are not normaly distributed  
amusement$num.child <- log(amusement$num.child)  
amusement$distance <- log(amusement$distance)  
  
hist(amusement$num.child, probability = TRUE)



hist(amusement$distance, probability = TRUE)



# QUESTION 9  
# Plot a scatter plot for overall satisfaction and rides satisfaction  
  
plot(amusement$rides, amusement$overall, col = "blue", main = 'Scatterplot of rides and overall satisfaction'  
 , xlab = 'rides', ylab = 'overall satisfaction')  
  
abline(lm(amusement$overall~amusement$rides ))



# QUESTION 10 - Run a correlation matrix with corrplot.mixed()   
library(corrplot)

## corrplot 0.92 loaded

cor\_matrix <- cor(amusement[, 2:8],use = "complete.obs")  
cor\_matrix

## num.child distance rides games wait clean  
## num.child 1 NaN NaN NaN NaN NaN  
## distance NaN 1.000000000 -0.01102768 0.001868728 0.01746093 0.02212374  
## rides NaN -0.011027676 1.00000000 0.455185111 0.31419951 0.78956505  
## games NaN 0.001868728 0.45518511 1.000000000 0.29910498 0.51697987  
## wait NaN 0.017460929 0.31419951 0.299104980 1.00000000 0.36788467  
## clean NaN 0.022123740 0.78956505 0.516979874 0.36788467 1.00000000  
## overall NaN 0.076327893 0.58598628 0.437467872 0.57262166 0.63939818  
## overall  
## num.child NaN  
## distance 0.07632789  
## rides 0.58598628  
## games 0.43746787  
## wait 0.57262166  
## clean 0.63939818  
## overall 1.00000000

corrplot(cor\_matrix, method = "ellipse")

