Survey Methodology – Data Science Department

Team\_6, Phase 6

**Data Analysis code**

A picture containing text

Description automatically generated

#install.packages("effects")

library(dplyr)

library(corrplot)

library(readr)

library(ggstatsplot)

library(ggplot2)

library(ggcorrplot)

library(Hmisc)

library(vcd)

library(data.table)

require(foreign)

require(MASS)

require(reshape2)

data <- read.csv("C:\\Users\\Menna Yasser\\Documents\\R\\All data (1) .csv ")

#Dropping columns we don't want

data=data[,-c(1,7,29,30,32,27,28)]

#Descriptive statistics:

describe(data)

#chi-square tests

chisq.test(table(data$Last\_Watched, data$Reason\_of\_Watch))

#plot chi-square tests

ggbarstats(

  data = data,

  x = Last\_Watched,

  y = Reason\_of\_Watch

) +

  labs(caption = NULL)

# Correlation Matrix Between categorical value Variables

c=table(data$Last\_Watched,data$Reason\_of\_Watch)

cor(c)

corrplot(cor(c),method = "pie",type = ("upper"))

#chi-square tests

chisq.test(table(data$best.day.c., data$best.time.c.))

#plot chi-square tests

ggbarstats(

  data = data,

  x = best.day.c.,

  y = best.time.c.

) +

  labs(caption = NULL)

# Correlation Matrix Between categorical value Variables

c=table(data$best.day.c.,data$best.time.c.)

cor(c)

corrplot(cor(c),method = "pie",type = ("upper"))

#chi-square tests

chisq.test(table(data$Search\_Before\_Watch, data$Ads\_effect))

#plot chi-square tests

ggbarstats(

  data = data,

  x = Search\_Before\_Watch,

  y = Ads\_effect

) +

  labs(caption = NULL)

# Correlation Matrix Between categorical value Variables

c=table(data$Search\_Before\_Watch,data$Ads\_effect)

cor(c)

corrplot(cor(c),method = "pie",type = ("upper"))

#chi-square tests

chisq.test(table(data$Search\_Before\_Watch, data$Reason\_of\_Watch))

#plot chi-square tests

ggbarstats(

  data = data,

  x = Search\_Before\_Watch,

  y = Reason\_of\_Watch

) +

  labs(caption = NULL)

# Correlation Matrix Between categorical value Variables

c=table(data$Search\_Before\_Watch, data$Reason\_of\_Watch)

cor(c)

corrplot(cor(c),method = "pie",type = ("upper"))

#chi-square tests

chisq.test(table(data$best.time.c., data$usuall.way.to.watch.movies.at.home))

#plot chi-square tests

ggbarstats(

  data = data,

  x = best.time.c.,

  y = usuall.way.to.watch.movies.at.home

) +

  labs(caption = NULL)

# Correlation Matrix Between categorical value Variables

c=table(data$best.time.c.,data$usuall.way.to.watch.movies.at.home)

cor(c)

corrplot(cor(c),method = "pie",type = ("upper"))

#--------------------------------------------#

# Correlation Matrix Between Numeric Variables

correlation=cor(data[sapply(data,is.numeric)])

correlation

#plot Correlation Matrix Between Numeric Variables

ggcorrplot(correlation, title = "Correlation Matrix",hc.order = T,lab=TRUE

,type = "lower",lab\_size =1 ) + theme(plot.title=element\_text(hjust = 1,size = 20),

 axis.text.y = element\_text(size = 10),axis.text.x = element\_text(size = 8),legend.text = element\_text(size = 5))

#--------------------------------------------#

# Correlation Matrix Between all Variables

# for(j in 1:34){

#   for(i in 1:33){

#     m=cor(table(data[,j], data[,i]))

#     print( paste0(colnames(data)[j] ," and ", colnames(data)[i]))

#     print(m)

#

#   }

# }

# #--------------------------------------------#

# # chi-square tests all Variables

# for(j in 1:33){

#  for(i in 1:ncol(data)){

#     m=chisq.test(table(data[,j], data[,i]))

#

#     print(m)

#     print(paste0("chi-square tests between ",colnames(data)[j] ," and ", colnames(data)[i]))}}

#

#--------------------------------------------#

# Create the relationship model.

#Ordering the dependent variable

input <- data[,c("Genre","Reviews","Quality\_.Story","Comedy\_Rate" ,"Sci\_fi\_Rate")]

input$Genre = factor(input$Genre, levels = c( "1","2","3","4","5"), ordered = TRUE)

input$Reviews = factor(input$Reviews, levels = c( "1","2","3","4","5"), ordered = TRUE)

input$Quality\_.Story = factor(input$Quality\_.Story, levels = c("1","2","3","4","5"), ordered = TRUE)

input$Comedy\_Rate = factor(input$Comedy\_Rate, levels = c("1","2","3","4","5"), ordered = TRUE)

input$Sci\_fi\_Rate = factor(input$Sci\_fi\_Rate, levels = c("1","2","3","4","5"), ordered = TRUE)

summary(input)

table(input$Genre,input$Reviews)

table(input$Genre,input$Quality\_.Story)

table(input$Genre,input$Comedy\_Rate)

table(input$Genre,input$Sci\_fi\_Rate)

#Dividing data into training and test set

#Random sampling

samplesize = 0.8\*nrow(input)

set.seed(100)

index = sample(seq\_len(nrow(input)), size = samplesize)

#Creating training and test set

datatrain = input[index,]

datatest = input[-index,]

#Build ordinal logistic regression model

model <- polr(Genre~Quality\_.Story+ Reviews+Comedy\_Rate+Sci\_fi\_Rate, data = datatrain, Hess=T)

summary(model)

#Compute confusion table and misclassification error

predictGenre = predict(model,datatest)

predictGenre

table(datatest$Genre, predictGenre)

table(datatest$Genre,datatest$Genre)

#Plotting the effects

library("effects")

Effect(focal.predictors = "Quality\_.Story",model)

plot(Effect(focal.predictors = "Quality\_.Story",model))

plot(Effect(focal.predictors = "Reviews",model))

plot(Effect(focal.predictors = c("Quality\_.Story", "Sci\_fi\_Rate"),model))

#--------------------------------------------#

#visualization

#gender and age

Table=table(data$Gender,data$Age)

Table

barplot(height=Table,beside=T,

        col=c("#8C0000","#FFBE0F"),

        xlab="Age",

        ylab="Num of males and females")

legend("topright",

       c("Female","Male"),

       fill = c("#8C0000","#FFBE0F")

)

#watching places

watchingcounts=table(data$Watching\_Place)

watchingcounts

library(lessR)

PieChart(watchingcounts, values = "%",

         main = "Number of watchers at home vs cinema",

         hole = 0,

         fill =c("#8C0000","#FFBE0F") )

#Reason\_of\_Watch

reasonOfWatch=table(data$Reason\_of\_Watch)

par(mar=c(9, 8, 3, 5))

barplot(reasonOfWatch,las=2,main='Reason of watch counts',

        ylab="Count",

        col="#BD2000",

        border="#8C0000"

)

#search before

search=table(data$Search\_Before\_Watch)

search

library(lessR)

PieChart(search, values = "%",

         main = "how often people search before watching and thier count",

         hole = 0,

         fill =c("#FFBE0F","#FA1E0E","#8C0000","#BD2000"))

#search before and rates

par(mfrow=c(3,3))

#Action

ActionSearch=table(data$Search\_Before\_Watch,data$Action\_Rate)

barplot(ActionSearch,main='search and action rates',

        xlab="Rate",

        ylab="Count",

        beside=T,

        border="blue",

        las=1,

        col=c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

legend("topleft", cex=0.45,

       c("Maybe","Never","sometimes","usually"),

       fill =c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

#Adventure

AdventureSearch=table(data$Search\_Before\_Watch,data$Adventure\_Rate)

barplot(AdventureSearch,las=2,main='search and Adventure',

        ylab="Count",

        xlab="Rate",

        beside=T,

        border="blue",

        las=1,

        col=c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

legend("topleft", cex=0.45,

       c("Maybe","Never","sometimes","usually"),

       fill = c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

#comedy

ComedySearch=table(data$Search\_Before\_Watch,data$Comedy\_Rate)

barplot(AdventureSearch,las=2,main='search and comedy',

        ylab="Count",

        xlab="Rate",

        beside=T,

        border="blue",

        las=1,

        col=c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

#

legend("topleft", cex=0.45,

       c("Maybe","Never","sometimes","usually"),

       fill = c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

#Drama

DramaSearch=table(data$Search\_Before\_Watch,data$Drama\_Rate)

barplot(DramaSearch,las=2,main='search and drama',

        ylab="Count",

        xlab="Rate",

        beside=T,

        border="blue",

        las=1,

        col=c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

legend("topleft", cex=0.45,

       c("Maybe","Never","sometimes","usually"),

       fill = c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

# Romance

RomanceSearch=table(data$Search\_Before\_Watch,data$ Romance\_Rate)

barplot( RomanceSearch,las=2,main='search and romance',

         ylab="Count",

         xlab="Rate",

         beside=T,

         border="blue",

         las=1,

         col=c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

legend("topleft", cex=0.4,

       c("Maybe","Never","sometimes","usually"),

       fill = c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

#Horror

HorrorSearch=table(data$Search\_Before\_Watch,data$ Horror\_Rate)

barplot(HorrorSearch,las=2,main='search and horror',

        ylab="Count",

        xlab="Rate",

        beside=T,

        border="blue",

        las=1,

        col=c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

legend("topright", cex=0.45,

       c("Maybe","Never","sometimes","usually"),

       fill = c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

#Sci-fi

Sci\_fiSearch=table(data$Search\_Before\_Watch,data$Sci\_fi\_Rate)

barplot(Sci\_fiSearch,las=2,main='search sci fi',

        ylab="Count",

        xlab="Rate",

        beside=T,

        border="blue",

        las=1,

        col=c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

legend("topleft", cex=0.45,

       c("Maybe","Never","sometimes","usually"),

       fill = c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

#Musical

MusicalSearch=table(data$Search\_Before\_Watch,data$ Musical\_Rate)

barplot(MusicalSearch,las=2,main='search and Musical',

        xlab="Rate",

        ylab="Count",

        beside=T,

        border="blue",

        las=1,

        col=c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

legend("topright", cex=0.45,

       c("Maybe","Never","sometimes","usually"),

       fill = c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

#Family

FamilySearch=table(data$Search\_Before\_Watch,data$ Family\_Rate)

barplot(FamilySearch,las=2,main='search and Family',

        xlab="Rate",

        ylab="Count",

        beside=T,

        border="blue",

        las=1,

        col=c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

#

legend("topright", cex=0.5,

       c("Maybe","Never","sometimes","usually"),

       fill = c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

#Factors

#Factors and Action Rate

typeofEachFactorinAction<-group\_by(data,Factors)

typeofEachFactorinAction<-summarise(typeofEachFactorinAction,sumofactionrate = sum(Action\_Rate))

print(typeofEachFactorinAction)

p1=pie(x=typeofEachFactorinAction$sumofactionrate,

       labels=typeofEachFactorinAction$Factors,

       main="Factors and Action Rate",

       radius = 1.05,

       col =c("#FFBE0F","#FA1E0E","#8C0000","#BD2000"))

# Factors and Musical Rate

typeofEachFactorinMusical<-group\_by(data,Factors)

typeofEachFactorinMusical<-summarise(typeofEachFactorinMusical,sumofMusicalrate = sum(Musical\_Rate))

print(typeofEachFactorinMusical)

p2=pie(x=typeofEachFactorinMusical$sumofMusicalrate,

       labels=typeofEachFactorinMusical$Factors,

       main="Factors and Musical Rate",

       radius =1.05,

       col =c("#FFBE0F","#FA1E0E","#8C0000","#BD2000"))

#factors that you consider when deciding to see a movie.

par(mar=c(9, 8, 3, 5))

boxplot(data[,18:23],las=1,horizontal = T,col = c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F"),main="factors effect and its rank ")

#Seats and Snacks of the Cinema effect on  enjoying a movie?

data <- read.csv("C:\\Users\\Menna Yasser\\Documents\\R\\All data (1) .csv ")

data=data[,-c(1,7,29,30,32,27)]

seats=table(data$`Seats.and.Snacks.of.the.Cinema.effect.on..enjoying.a.movie.`)

seats

library(lessR)

PieChart(seats, values = "%",

         main = "Seats and Snacks of the Cinema effect on  enjoying a movie?",

         hole = 0,

         fill =c("#FFBE0F","#8C0000"))

#Ads\_effect and how often a a movie Disappoints you

table(data$Usually\_Disappoint,data$Ads\_effect)

par(mar=c(9, 8, 3, 5))

barplot( table(data$Usually\_Disappoint,data$Ads\_effect),main='Ads\_effect and how often a a movie Disappoints you',

         beside=T,

         border="blue",

         las=2,

         col = c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F")

)

legend("topright",

       c("Almost Always","Never","sometimes","usually"),

       fill =c("#8C0000", "#BD2000","#FA1E0E","#FFBE0F"),

       cex = 0.8

)

#freq going to cinema  and sum of Drama Rate

freqGoToCinemaAndDrama<-group\_by(data,Frequent\_Go\_Cinema)

freqGoToCinemaAndDrama<-summarise(freqGoToCinemaAndDrama,sumofDramarate = sum(Drama\_Rate))

print(freqGoToCinemaAndDrama)

pie(x=freqGoToCinemaAndDrama$sumofDramarate,

    labels=freqGoToCinemaAndDrama$Frequent\_Go\_Cinema,

    main="Factors and Drama Rate",

    col =c("#FFBE0F","#FA1E0E","#8C0000","#BD2000"))