##### HOMEWORK 1 #####

getwd()

library(readr)

IMDB.df <- read.csv( "IMDB Movie Dataset.csv" , header = TRUE , stringsAsFactors = TRUE )

IMDB.df

##### Summary statistics #####

# Dataset Description

#display all variable names

names(IMDB.df)

#describe and classify data type

summary(IMDB.df)

# Summary Statistics Table

#filter out variables that are numeric

numeric <- Filter(is.numeric,IMDB.df)

#summary statistics of numeric variables

summary(numeric)

# Missing Data from Variables

sort(colSums(is.na(numeric)), decreasing=TRUE)

# Variables Containing Outliers

boxplot(IMDB.df$duration, # numeric.var ~ categorical.var

xlab = "Duration",

main = "Distribution of Movie Duration")

# Correlation Table with Heatmap

#filter data frame with numeric variables only

IMDBnumeric.df <- Filter( is.numeric, IMDB.df )

head(IMDBnumeric.df)

#find correlation (correlations only for non-missing data)

round( cor( IMDBnumeric.df , use = "complete.obs" ) , 2 )

#install package

install.packages("gplots", dependencies = TRUE)

library(gplots)

#define color for heatmap

colfunc <- colorRampPalette(c("green", "white", "red"))

#create heatmap

heatmap.2(cor( IMDBnumeric.df , use = "complete.obs" ),

Rowv = FALSE, Colv = FALSE,

dendrogram = "none",

col = colfunc(15),

cellnote = round( cor( IMDBnumeric.df , use = "complete.obs" ) , 2 ),

notecol = "black",

key = FALSE, trace = 'none',

margins = c(12,12))

##### Building Chart #####

# Histogram 1

hist( IMDB.df$imdb\_score,

xlim = c( 1 , 10 ),

breaks = 9,

xlab = "IMDB Score",

ylab = "Number of film",

main = "Histogram of IMDB Score",

col = "sky blue"

)

# Histogram 2

hist( IMDB.df$duration,

xlim = c( 0 , 334 ),

xlab = "Duration",

ylab = "Number of film",

main = "Histogram of duration",

col = "yellow"

)

# Bar chart 1

#create data to be displayed

data.for.plot <- aggregate( IMDB.df$movie\_title,

by = list( IMDB.df$ï..color ),

FUN = length

)

data.for.plot

names(data.for.plot) <- c("Color","CountOfFilms")

#create bar chart

barplot( height = data.for.plot$CountOfFilms,

names.arg = data.for.plot$Color,

xlab = "Color",

ylab = "Count of Films",

main = "Films by colors",

col = c( "grey" , "black" , "pink" )

)

# Bar chart 2

#create data to be displayed

data.for.plot <- aggregate( IMDB.df$movie\_title,

by = list( IMDB.df$content\_rating ),

FUN = length

)

data.for.plot

names(data.for.plot) <- c("ContentRating","CountOfFilms")

#create bar chart

barplot( height = data.for.plot$CountOfFilms,

names.arg = data.for.plot$ContentRating,

xlab = "Content Rating",

ylab = "Number of Films",

main = "Films by Content Rating",

col = "light green"

)

# Box plot 1

boxplot(IMDB.df$facenumber\_in\_poster ~ IMDB.df$language,

ylim = c(0,10),

xlab = "Language",

ylab = "# of Faces in the Movie Poster",

main = "Relationship between the Movie Genre and Facebook Likes")

# Box plot 2

boxplot(IMDB.df$imdb\_score ~ IMDB.df$content\_rating,

ylim = c(0,10),

xlab = "Content Rating", ylab = "IMDB Score",

main = "Relationship between the Content Rating and IMDB Score")

# Scatterplot 1

plot( IMDB.df$actor\_1\_facebook\_likes[-c(1224,1903,2270,2940,4046,4410,4593,4705)]

~ IMDB.df$cast\_total\_facebook\_likes[-c(1224,1903,2270,2940,4046,4410,4593,4705)],

xlab = "Cast Total Facebook Likes",

ylab = "Actor 1 Facebook Likes",

main = "Scatterplot of Cast Total VS. Actor 1 Facebook Likes"

)

# Scatterplot 2

plot( IMDB.df$duration[-1711]

~ IMDB.df$imdb\_score[-1711],

xlab = "IMDB Scores",

ylab = "Duration",

main = "Scatterplot of IMDB Scores VS. Duration"

)

##### Handling missing data #####

# Method 1

summary(IMDB.df$duration)

IMDB.df <- IMDB.df[!is.na(IMDB.df$duration), ]

dim(IMDB.df)

# Method 2

IMDB.df$budget[is.na(IMDB.df$budget)] <- median(IMDB.df$budget, na.rm = TRUE)

summary(IMDB.df$budget)

##### Binary Dummy Variable #####

IMDB.df <- read.csv( "IMDB Movie Dataset.csv" , header = TRUE , stringsAsFactors = TRUE )

#install Packages

install.packages( "fastDummies" , dependency = TRUE )

library( fastDummies )

#create Dummy Variables

dummy\_cols( IMDB.df,

select\_columns = "content\_rating",

remove\_first\_dummy = TRUE,

remove\_selected\_columns = TRUE

)

IMDB.df <- dummy\_cols( IMDB.df,

select\_columns = "content\_rating",

remove\_first\_dummy = TRUE,

remove\_selected\_columns = TRUE

)

head(IMDB.df)