Assignment 1

Tongxiang Lu

2022-09-11

# Import the dataset into R

1. Dataset source:

download from <https://www.superdatascience.com/pages/rcourse>,

the file is under section 6 advanced visualization—movie ratings

1. Steps:

# 1). Navigate using the Files Tab (bottom right pane) to where my data file is. It is in my desktop.

# 2). Click on the \* P2-Movie-Ratings.csv \* file.

# 3). Click on \*Import Dataset\*.

# 4). Pay attention to the dataset name (lower left corner). It will be something like \* P2-Movie-Ratings.csv \*. This will be how I will use the data. Also, note the R code that is being used to read the file.

# 5). Click import.

# Print out descriptive statistics for a selection of quantitative and categorical variables

Steps:

# use below console to find my file path.

getwd()

setwd(“C:/Users/Tongxiang Lu/Desktop”)

getwd()

mydata <- read.csv(“P2-Movie-Ratings.csv”)

# Use summary to calculate all of my variable measures.

summary(mydata)

# Calculate the mean and standard deviation of a quantitative variable.

mean(mydata

Budget)

# To see more descriptive statistics output, use stat.desc() from the package {pastecs}.

install.packages(“pastecs”)

library(pastecs)

stat.desc(mydata)

# Transform at least one variable

# Use Logarithm to see my variable transformation output, the benefit of which are listed as below:

1. It reduces the scatterness of the data.
2. It linearizes the non-linear data.
3. It mixes the data smooth.

log\_Budget <- log10(mydata$Budget)

log\_Budget

log\_Budget <- log10(mydata$Audience.Ratings)

log\_Budget

# Plot at least one quantitative variable, and one scatterplot

Steps:

# use hist(mydata$quantitative variable) to see my histogram output

hist(mydata$Budget)

# use plot( my data quantitative variable x quantitative variable y) to see my scatterplot output.

plot(mydataAudience.Ratings..)