Assignment\_1

Stephen Gombos

2025-09-07

# 1.) Download a dataset from the web. You may use any source, but specify the source in your code. Also ensure that the data has a mix of quantitative and qualitative (categorical) variables.\*\*

##Done outside of RStudio

## 1.1) Set up RStudio file

##install.packages("dplyr") Commented out, only need to install once  
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

# 2.) Import the dataset into R

BathSoap <- read.csv("/Users/stephengombos/Documents/KSU MBA PROGRAM/Fall 2025 FUNDAMENTALS OF MACHINE LEARNING (BA-64060-002)/CSV Files/BathSoap.CSV")

## 2.1) Visualize the Dataset

head(BathSoap)

## Member.id SEC FEH MT SEX AGE EDU HS CHILD CS Affluence.Index No..of.Brands  
## 1 1010010 4 3 10 1 4 4 2 4 1 2 3  
## 2 1010020 3 2 10 2 2 4 4 2 1 19 5  
## 3 1014020 2 3 10 2 4 5 6 4 1 23 5  
## 4 1014030 4 0 0 0 4 0 0 5 0 0 2  
## 5 1014190 4 1 10 2 3 4 4 3 1 10 3  
## 6 1017020 4 3 10 2 3 4 5 2 1 13 3  
## Brand.Runs Total.Volume No..of..Trans Value Trans...Brand.Runs Vol.Tran  
## 1 17 8025 24 818.0 1.41 334.38  
## 2 25 13975 40 1681.5 1.60 349.38  
## 3 37 23100 63 1950.0 1.70 366.67  
## 4 4 1500 4 114.0 1.00 375.00  
## 5 6 8300 13 591.0 2.17 638.46  
## 6 26 18175 41 1705.5 1.58 443.29  
## Avg..Price Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..  
## 1 10.19 100% 0% 0%  
## 2 12.03 89% 10% 2%  
## 3 8.44 94% 2% 4%  
## 4 7.60 100% 0% 0%  
## 5 7.12 61% 14% 24%  
## 6 9.38 100% 0% 0%  
## Br..Cd..57..144 Br..Cd..55 Br..Cd..272 Br..Cd..286 Br..Cd..24 Br..Cd..481  
## 1 38% 13% 0% 0% 0% 0%  
## 2 2% 8% 0% 0% 0% 6%  
## 3 3% 55% 0% 3% 0% 0%  
## 4 40% 60% 0% 0% 0% 0%  
## 5 5% 14% 0% 0% 0% 0%  
## 6 8% 7% 0% 0% 0% 0%  
## Br..Cd..352 Br..Cd..5 Others.999 Pr.Cat.1 Pr.Cat.2 Pr.Cat.3 Pr.Cat.4  
## 1 0% 0% 49.2% 23% 56% 13% 7%  
## 2 0% 14% 69.9% 29% 55% 9% 6%  
## 3 0% 2% 37.9% 12% 32% 56% 0%  
## 4 0% 0% 0.0% 0% 40% 60% 0%  
## 5 0% 0% 80.7% 0% 5% 14% 81%  
## 6 0% 0% 85.7% 22% 45% 7% 27%  
## PropCat.5 PropCat.6 PropCat.7 PropCat.8 PropCat.9 PropCat.10 PropCat.11  
## 1 50% 0% 0% 0% 0% 0% 0%  
## 2 46% 35% 3% 2% 1% 0% 6%  
## 3 24% 12% 3% 1% 1% 0% 0%  
## 4 40% 0% 0% 0% 0% 0% 0%  
## 5 81% 0% 0% 5% 0% 0% 0%  
## 6 49% 10% 0% 1% 7% 0% 0%  
## PropCat.12 PropCat.13 PropCat.14 PropCat.15  
## 1 3% 0% 13% 34%  
## 2 0% 0% 8% 0%  
## 3 2% 0% 56% 0%  
## 4 0% 0% 60% 0%  
## 5 0% 0% 14% 0%  
## 6 0% 0% 7% 27%

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

summary(BathSoap)

## Member.id SEC FEH MT   
## Min. :1010010 Min. :1.00 Min. :0.000 Min. : 0.000   
## 1st Qu.:1065295 1st Qu.:1.75 1st Qu.:1.000 1st Qu.: 4.000   
## Median :1106235 Median :2.50 Median :3.000 Median :10.000   
## Mean :1104188 Mean :2.50 Mean :2.048 Mean : 8.178   
## 3rd Qu.:1148292 3rd Qu.:3.25 3rd Qu.:3.000 3rd Qu.:10.000   
## Max. :1167670 Max. :4.00 Max. :3.000 Max. :19.000   
## SEX AGE EDU HS   
## Min. :0.000 Min. :1.000 Min. :0.000 Min. : 0.000   
## 1st Qu.:2.000 1st Qu.:3.000 1st Qu.:3.000 1st Qu.: 3.000   
## Median :2.000 Median :3.000 Median :4.500 Median : 4.000   
## Mean :1.738 Mean :3.213 Mean :4.043 Mean : 4.192   
## 3rd Qu.:2.000 3rd Qu.:4.000 3rd Qu.:5.000 3rd Qu.: 5.000   
## Max. :2.000 Max. :4.000 Max. :9.000 Max. :15.000   
## CHILD CS Affluence.Index No..of.Brands   
## Min. :1.000 Min. :0.0000 Min. : 0.00 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:1.0000 1st Qu.:10.00 1st Qu.:2.000   
## Median :4.000 Median :1.0000 Median :15.00 Median :3.000   
## Mean :3.233 Mean :0.9317 Mean :17.02 Mean :3.637   
## 3rd Qu.:4.000 3rd Qu.:1.0000 3rd Qu.:24.00 3rd Qu.:5.000   
## Max. :5.000 Max. :2.0000 Max. :53.00 Max. :9.000   
## Brand.Runs Total.Volume No..of..Trans Value   
## Min. : 1.00 Min. : 150 Min. : 1.00 Min. : 20.0   
## 1st Qu.: 8.00 1st Qu.: 6825 1st Qu.: 22.00 1st Qu.: 789.6   
## Median :15.00 Median :10360 Median : 28.00 Median :1216.0   
## Mean :15.75 Mean :11915 Mean : 31.15 Mean :1337.4   
## 3rd Qu.:21.00 3rd Qu.:15344 3rd Qu.: 40.00 3rd Qu.:1675.8   
## Max. :74.00 Max. :50895 Max. :138.00 Max. :6371.9   
## Trans...Brand.Runs Vol.Tran Avg..Price Pur.Vol.No.Promo....  
## Min. : 1.000 Min. : 94.43 Min. : 5.62 Length:600   
## 1st Qu.: 1.420 1st Qu.: 250.51 1st Qu.: 9.76 Class :character   
## Median : 1.845 Median : 361.52 Median :11.25 Mode :character   
## Mean : 2.618 Mean : 415.05 Mean :11.83   
## 3rd Qu.: 2.690 3rd Qu.: 490.89 3rd Qu.:13.42   
## Max. :23.000 Max. :2525.00 Max. :33.33   
## Pur.Vol.Promo.6.. Pur.Vol.Other.Promo.. Br..Cd..57..144 Br..Cd..55   
## Length:600 Length:600 Length:600 Length:600   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Br..Cd..272 Br..Cd..286 Br..Cd..24 Br..Cd..481   
## Length:600 Length:600 Length:600 Length:600   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Br..Cd..352 Br..Cd..5 Others.999 Pr.Cat.1   
## Length:600 Length:600 Length:600 Length:600   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Pr.Cat.2 Pr.Cat.3 Pr.Cat.4 PropCat.5   
## Length:600 Length:600 Length:600 Length:600   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## PropCat.6 PropCat.7 PropCat.8 PropCat.9   
## Length:600 Length:600 Length:600 Length:600   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## PropCat.10 PropCat.11 PropCat.12 PropCat.13   
## Length:600 Length:600 Length:600 Length:600   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## PropCat.14 PropCat.15   
## Length:600 Length:600   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##

# 4) Transform at least one variable. It doesn’t matter what the transformation is.

First lets look at the column names

colnames(BathSoap)

## [1] "Member.id" "SEC" "FEH"   
## [4] "MT" "SEX" "AGE"   
## [7] "EDU" "HS" "CHILD"   
## [10] "CS" "Affluence.Index" "No..of.Brands"   
## [13] "Brand.Runs" "Total.Volume" "No..of..Trans"   
## [16] "Value" "Trans...Brand.Runs" "Vol.Tran"   
## [19] "Avg..Price" "Pur.Vol.No.Promo...." "Pur.Vol.Promo.6.."   
## [22] "Pur.Vol.Other.Promo.." "Br..Cd..57..144" "Br..Cd..55"   
## [25] "Br..Cd..272" "Br..Cd..286" "Br..Cd..24"   
## [28] "Br..Cd..481" "Br..Cd..352" "Br..Cd..5"   
## [31] "Others.999" "Pr.Cat.1" "Pr.Cat.2"   
## [34] "Pr.Cat.3" "Pr.Cat.4" "PropCat.5"   
## [37] "PropCat.6" "PropCat.7" "PropCat.8"   
## [40] "PropCat.9" "PropCat.10" "PropCat.11"   
## [43] "PropCat.12" "PropCat.13" "PropCat.14"   
## [46] "PropCat.15"

Next lets confirm the dataclass for “Pur.Vol.No.Promo….”

class(BathSoap$Pur.Vol.No.Promo....)

## [1] "character"

The data in this column is a percentage. For the transfirmation we will remove the % sign and transform it to a number and then to a decimal

BathSoap$Pur.Vol.No.Promo.... <- gsub("%", "", BathSoap$Pur.Vol.No.Promo....)  
BathSoap$Pur.Vol.No.Promo.... <- as.numeric(BathSoap$Pur.Vol.No.Promo....)  
BathSoap$Pur.Vol.No.Promo.... <- BathSoap$Pur.Vol.No.Promo.... / 100

confirm the class has changed

class(BathSoap$Pur.Vol.No.Promo....)

## [1] "numeric"

verify data

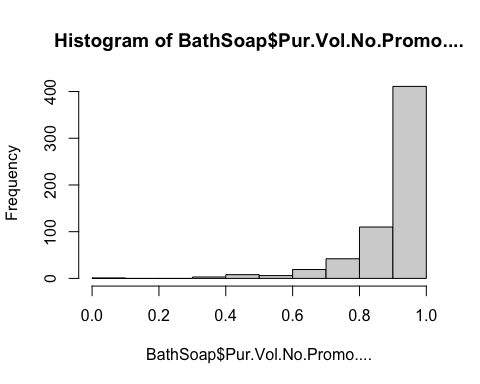
head(BathSoap$Pur.Vol.No.Promo....)

## [1] 1.00 0.89 0.94 1.00 0.61 1.00

# 5) Plot at least one quantitative variable, and one scatterplot

## 5.1) Quantitative variable plot of the newly transformed BathSoap$Pur.Vol.No.Promo…. using a histogram

hist(BathSoap$Pur.Vol.No.Promo....)



## 5.2) Scatterplot of the Total Volume vs Value

plot(BathSoap$Total.Volume, BathSoap$Value,  
 main = "Total Volume vs. Value",  
 xlab = "Total Volume",  
 ylab = "Value",  
 col = "blue")

