EBIS 4043 Big Data Analysis and Applications

Individual Assignment

The purpose of this assignment is to make sure that **you are picking up the R based analytics skills (Please do not use other tools to generate the answers!)** that have been introduced in this class and check your ability to translate research questions into analytical approaches.

As you write your answers, please make sure to follow the instructions below:

* + Use the datasets that were uploaded on iSpace.
  + Make sure to have the entire process from data loading to analysis and interpretation in the submission.
  + All your answers including your identity, codes, interpretation should be in one file (HTML or PDF). Any sort of multiple files will be graded as zero mark.
  + Explain your approach to the analysis; no explanation = no mark
  + You can discuss the coding with your friends. However, any visible overlap in your interpretation will be considered plagiarism.
  + There can be more than one correct answer to every question. Use any technique that you learned from the classroom.

1. Greg Metcalf works for a national credit card company, and he is performing a customer value analysis on a subset of credit card customers. In order to perform the RFM analysis on the customers, Greg has compiled the accompanying data file ***Credit\_Cards*** that contains the dates of the last transaction (LastTransactionDate), total number of transactions in the past two years (Frequency), and total spending during the past two years (Spending). [50 marks]
   1. Greg wants to calculate the number of days between January 1, 2022, and the last transaction date. Create a new variable “DaysSinceLast” that contains the number of days since the last transaction. What is the average number of days since the last purchase for all the customers?
   2. Create the RFM scores for each customer. How many customers have an RFM score of 555? What is their average spending?

*The recency, frequency, monetary value (RFM) model assigns a firm's customer base a particular trait, which can be used to improve marketing analysis. For each attribute (recency, frequency, and monetary value), customers are given a score from 1 (lowest) to 5 (best) based on their observed purchasing behavior.*

* 1. Create a new variable called “LogSpending” that contains the natural logarithms for the total spending during the past two years. Bin the logarithm values into five equal-interval groups. Label the groups usingnumbers 1 (lowest values) to 5 (highest values). How many observations are in group 2?
  2. Create a new variable called “AverageOrderSize” that contains the average spending per order. This is calculated by dividing total spending (Spending) by total number of transactions (Frequency) in the past two years. Bin the values of AverageOrderSize into five equal-interval groups. Label the groups using numbers 1 (lowest values) to 5 (highest values). How many observations are in group 2?
  3. Compare the groups in parts c and d. Are the groupings the same or different? Interpret your comparison.

# 加载必要的包

library(dplyr)

# 读取数据文件 Credit\_Cards

credit\_cards <- read.csv("Credit\_Cards.csv") # 假设数据文件名为 Credit\_Cards.csv

# a. 计算DaysSinceLast

credit\_cards$LastTransactionDate <- as.Date(credit\_cards$LastTransactionDate, format = "%Y-%m-%d")

credit\_cards$DaysSinceLast <- as.numeric(difftime("2022-01-01", credit\_cards$LastTransactionDate, units = "days"))

# 计算平均天数

average\_days\_since\_last <- mean(credit\_cards$DaysSinceLast, na.rm = TRUE)

cat("平均天数自最后一次购买:", average\_days\_since\_last, "\n")

# b. 创建RFM得分

credit\_cards$RFM\_Score <- paste0(

ifelse(credit\_cards$DaysSinceLast <= 365, 5, ifelse(credit\_cards$DaysSinceLast <= 730, 4, 3)),

ifelse(credit\_cards$Frequency >= 100, 5, ifelse(credit\_cards$Frequency >= 50, 4, 3)),

ifelse(credit\_cards$Spending >= 10000, 5, ifelse(credit\_cards$Spending >= 5000, 4, 3))

)

# 统计RFM得分为555的客户数量

rfm\_555\_count <- sum(credit\_cards$RFM\_Score == "555")

cat("RFM得分为555的客户数量:", rfm\_555\_count, "\n")

# 计算RFM得分为555的客户的平均消费

rfm\_555\_avg\_spending <- mean(credit\_cards$Spending[credit\_cards$RFM\_Score == "555"], na.rm = TRUE)

cat("RFM得分为555的客户的平均消费:", rfm\_555\_avg\_spending, "\n")

# c. 创建LogSpending变量并分组

credit\_cards$LogSpending <- log(credit\_cards$Spending)

credit\_cards$LogSpendingGroup <- cut(credit\_cards$LogSpending, breaks = 5, labels = FALSE)

# 统计分组2的观测数量

group\_2\_count <- sum(credit\_cards$LogSpendingGroup == 2)

cat("LogSpending分组2的观测数量:", group\_2\_count, "\n")

# d. 创建AverageOrderSize变量并分组

credit\_cards$AverageOrderSize <- credit\_cards$Spending / credit\_cards$Frequency

credit\_cards$AverageOrderSizeGroup <- cut(credit\_cards$AverageOrderSize, breaks = 5, labels = FALSE)

# 统计分组2的观测数量

group\_2\_count\_d <- sum(credit\_cards$AverageOrderSizeGroup == 2)

cat("AverageOrderSize分组2的观测数量:", group\_2\_count\_d, "\n")

# e. 比较LogSpending和AverageOrderSize的分组

if (identical(credit\_cards$LogSpendingGroup, credit\_cards$AverageOrderSizeGroup)) {

cat("LogSpending和AverageOrderSize的分组相同\n")

} else {

cat("LogSpending和AverageOrderSize的分组不同\n")

}

1. The accompanying data file ***Gas\_Prices*** shows the average price of gas (Price in $ per gallon) for the 50 states. Suppose that as a data analyst who works for government, you came to find out the outlies. [30 marks]
   1. Construct a boxplot for the Price variable. Does the boxplot suggest that outliers exist?

为Price变量构建一个箱线图，观察箱线图是否暗示存在异常值。箱线图可以直观地显示数据的分布和潜在的异常值。

# 读取数据文件 Gas\_Prices

gas\_prices <- read.csv("Gas\_Prices.csv") # 假设数据文件名为 Gas\_Prices.csv

# 构建箱线图

boxplot(gas\_prices$Price, main = "Gas Prices Boxplot", ylab = "Price ($ per gallon)")

* 1. Use z-scores to determine if there are any outliers for the Price variable. Are your results consistent with part a? Explain why or why not.

b. 使用z-分数来确定Price变量是否存在异常值。通常，z-分数大于3或小于-3的值可以被视为异常值。您可以使用如下代码计算z-分数：

# 计算z-分数

z\_scores <- scale(gas\_prices$Price)

# 寻找z-分数大于3或小于-3的观测

outliers <- which(z\_scores > 3 | z\_scores < -3)

cat("通过z-分数法找到的异常值：", length(outliers), "\n")

# 打印异常值的观测

cat("异常值的观测索引：", outliers, "\n")

* 1. What is the mean of the average price of gas without the outliers

c. 计算剔除异常值后的平均价格：

# 剔除异常值后的平均价格

mean\_without\_outliers <- mean(gas\_prices$Price[-outliers])

cat("剔除异常值后的平均价格：", mean\_without\_outliers, "\n")

1. A company that sells unisex t-shirts is interested in finding out the color and size of its best-selling t-shirt. The accompanying data file ***TShirts*** contains the size, color, and quantity of t-shirts that were ordered during the transactions. [20 marks]
   1. Construct a contingency table that shows the total quantity sold for each color and size combination. How many size M red t-shirts were sold? How many size XL purple t-shirts were sold?

a. 构建显示每种颜色和尺寸组合的总销售数量的列联表（contingency table）。然后，从表中获取特定尺寸和颜色的销售数量。

# 读取数据文件 TShirts

tshirts <- read.csv("TShirts.csv") # 假设数据文件名为 TShirts.csv

# 构建列联表

contingency\_table <- table(tshirts$Size, tshirts$Color)

# 获取特定尺寸和颜色的销售数量

size\_M\_red\_sold <- contingency\_table["M", "Red"]

size\_XL\_purple\_sold <- contingency\_table["XL", "Purple"]

cat("Size M红色T恤销售数量:", size\_M\_red\_sold, "\n")

cat("Size XL紫色T恤销售数量:", size\_XL\_purple\_sold, "\n")

* 1. Construct a heat map that displays colors or color intensity based on the total quantity sold. Which two color and size combinations are the most?

创建一个热图，其中颜色的强度基于总销售数量。您可以使用ggplot2包创建这个热图

# 加载必要的包

library(ggplot2)

# 创建数据框，其中包含颜色、尺寸和总销售数量

heatmap\_data <- data.frame(

Color = rep(levels(tshirts$Color), each = length(levels(tshirts$Size))),

Size = rep(levels(tshirts$Size), times = length(levels(tshirts$Color))),

TotalSold = as.vector(contingency\_table)

)

# 创建热图

heat\_map <- ggplot(heatmap\_data, aes(x = Size, y = Color, fill = TotalSold)) +

geom\_tile() +

scale\_fill\_gradient(low = "white", high = "blue") + # 颜色强度从白色到蓝色

labs(title = "T-Shirt Sales Heat Map", x = "Size", y = "Color", fill = "Total Quantity Sold")

# 显示热图

print(heat\_map)