R script code:

# Aadhithya Dinesh

# MIS 545 Section 02

# Lab03DineshA.R

# Import and prepare a large dataset of grocery store transactions, assign

# data types, view summary statistics, generate a histogram and a boxplot

# install.packages("tidyverse")

library(tidyverse)

# set the working directory

setwd("~/MIS/Classes/MIS545/Assignments/Lab03")

# read the csv file with column types specified

groceryTransactions1 <- read\_csv(file = "GroceryTransactions.csv",

col\_types = "iDffffifffffffin",

col\_names = TRUE)

# display the tibble

print(groceryTransactions1)

# display the first 20 rows

print(head(groceryTransactions1, n=20))

# display the structure of the tibble

str(groceryTransactions1)

# display the summary of the tibble

print(summary(groceryTransactions1))

# using dplyr summarize function to display mean of Revenue

print(summarize(.data = groceryTransactions1, mean(Revenue)))

# using dplyr summarize function to display median of UnitsSold

print(summarize(.data = groceryTransactions1, median(UnitsSold)))

# using dplyr summarize function to display standard deviation of Revenue

print(summarize(.data = groceryTransactions1, sd(Revenue)))

# using dplyr summarize function to display inter-quartile-range of Units sold

print(summarize(.data = groceryTransactions1, IQR(UnitsSold)))

# using dplyr summarize function to display min of revenue

print(summarize(.data = groceryTransactions1, min(Revenue)))

# using dplyr summarize function to display max of children

print(summarize(.data = groceryTransactions1, max(Children)))

# creating a new tibble with the below mentioned columns

groceryTransactions2 <- select(.data = groceryTransactions1,

PurchaseDate,

Homeowner,

Children,

AnnualIncome,

UnitsSold,

Revenue)

# display all features for transactions made by non-homeowners with at least

# 4 children

print(filter(.data = groceryTransactions2,

Homeowner == 'N' & Children >=4))

# display all of the records and features that were either made by customers

# in the $150K + annual income category OR had more than 6 units sold

print(filter(.data = groceryTransactions2,

AnnualIncome == '150k +' | UnitsSold > 6))

revnue <- select(.data = groceryTransactions1,

Revenue)

# display the average transaction revenue grouped by annual income level

print(select(.data = groceryTransactions1,

Revenue,

AnnualIncome)%>%

group\_by(AnnualIncome)%>%

summarize(averageRevnue = mean(Revenue))%>%

arrange(desc(averageRevnue)))

# calculate average price per unit as revenue/ units sold

groceryTransactions3 <- groceryTransactions2 %>%

mutate(AveragePricePerUnit = Revenue/UnitsSold)

print(groceryTransactions3)

histogramAveragePricePerUnit <- ggplot(data = groceryTransactions3,

aes(x=AveragePricePerUnit))

# creating the histogram for average price per unit

histogramAveragePricePerUnit + geom\_histogram(binwidth = 1,

color = "black",

fill = "orange",

alpha = 0.5

) +

ggtitle("Avergage Price Per Unit Histogram")

boxplotRevnue <- ggplot(data = groceryTransactions3,

aes(x=Revenue))

# creating the box plot for revenue

boxplotRevnue + geom\_boxplot(color = "#0C234B",

fill = "#AB0520")

Visualizations:

Chart, histogram

Description automatically generatedChart

Description automatically generated

Rapid Miner Process:

Graphical user interface, text, application

Description automatically generated

Results:

Graphical user interface, table

Description automatically generated