```
#---- install packages ("data.table", "ggplot2", "readr")
#install.packages("ggplot2")
#install.packages("ggmosaic")
#install.packages("readr")
#install.packages("readxl")

#---- Load required libraries

library(data.table)
library(ggplot2)
library(ggmosaic)
library(readr)
library(readxl)

#---- Load the data

filePath <- "D:\\Projects\\Proj#3_Quantium_Data_Analysis_Retail_ChipsCategory\\"
transactionData <- read_excel(paste0(filePath, "QVI_transaction_data.xlsx"))
customerData <- fread(paste0(filePath, "QVI_purchase_behaviour.csv"))

#---- Know the transaction data

str(transactionData)
head(transactionData)
summary(transactionData)

#---- Convert the numeric date format to actual date format

transactionData$DATE <- as.Date(transactionData$DATE, origin = "1899-12-30")

#---- View all product names and remove unwanted items, eg. Salsa

  # View product names
unique(transactionData$PROD_NAME)
setDT(transactionData)
  # Remove salsa
transactionData[, SALSA := grepl("salsa", tolower(PROD_NAME))]
transactionData <- transactionData[SALSA == FALSE][, SALSA := NULL]

#---- Identify and remove outliers (such as, many quantity purchases

transactionData[PROD_QTY > 100]
  # Remove customer who bought 200 units
transactionData <- transactionData[LYLTY_CARD_NBR != 226000]

#---- Plot transactions over time (01 July, 2018 to 01 June, 2019)

transactions_by_day <- transactionData[, .N, by = DATE]

  # To create the plot graph
ggplot(transactions_by_day, aes(x = DATE, y = N)) +
geom_line() +
labs(title = "Transactions Over Time") +
scale_x_date(breaks = "1 month") +
theme(axis.text.x = element_text(angle = 90))

#---- Converts PROD_NAME into structured variables to help with grouping and segmentation
  # Extract pack size
transactionData[, PACK_SIZE := parse_number(PROD_NAME)]

  # Extract brand name
transactionData[, BRAND := tstrsplit(PROD_NAME, ' ')[[1]]]

  # Normalize brand names
transactionData[BRAND == "RED", BRAND := "RRD"]

#---- Analyze Customer Dataset (purchase behaviour.csv)
summary(customerData)
  # Merge with transactions
data <- merge(transactionData, customerData, all.x = TRUE)
```

```
#---- Analysze Segments
  #Sales by segment
sales_summary <-data[, .(TOTAL_SALES =sum(TOT_SALES)), by = .(LIFESTAGE, PREMIUM_CUSTOMER)]
  # Units per customer
units_summary <- data[, .(UNITS = sum(PROD_QTY)/uniqueN(LYLTY_CARD_NBR)), by = .(LIFESTAGE, PREMIUM_
  # Price per unit
price_summary <- data[, .(AVG_PRICE = mean(TOT_SALES/PROD_QTY)), by = .(LIFESTAGE, PREMIUM_CUSTOMER)

#----- Example: compare price per unit between segments
premium_group <- data[LIFESTAGE == "YOUNG SINGLES/COUPLES" & PREMIUM_CUSTOMER == "Premium"]
mainstream_group <- data[LIFESTAGE == "YOUNG SINGLES/COUPLES" & PREMIUM_CUSTOMER == "Mainstream"]

t.test(premium_group$TOT_SALES / premium_group$PROD_QTY,
       mainstream_group$TOT_SALES / mainstream_group$PROD_QTY)

#---- Find favorite brands for Mainstream Young Singles
segment_data <- data[LIFESTAGE == "YOUNG SINGLES/COUPLES" & PREMIUM_CUSTOMER == "Mainstream"]
segment_data[, .N, by = BRAND][order(-N)]

#---- Export Cleaned Dataset

fwrite(data, paste0(filePath, "Task_1_QVI_data.csv"))
```