Quantium Virtual Internship - Retail Strategy and Analytics - Task 1

Load required libraries and datasets

```
filePath <- "" # Set your working directory
transactionData <- fread(paste0(filePath,"QVI_transaction_data.csv"))
customerData <- fread(paste0(filePath,"QVI_purchase_behaviour.csv"))
```

Exploratory Data Analysis

```
str(transactionData)
```

```
## Classes 'data.table' and 'data.frame':   264836 obs. of  8 variables:
##  $ DATE          : int  43390 43599 43605 43329 43330 43604 43601 43601 43332 43330 ...
##  $ STORE_NBR     : int  1 1 1 2 2 4 4 4 5 7 ...
##  $ LYLTY_CARD_NBR: int  1000 1307 1343 2373 2426 4074 4149 4196 5026 7150 ...
##  $ TXN_ID        : int  1 348 383 974 1038 2982 3333 3539 4525 6900 ...
##  $ PROD_NBR      : int  5 66 61 69 108 57 16 24 42 52 ...
##  $ PROD_NAME     : chr  "Natural Chip        Compny SeaSalt175g" "CCs Nacho Cheese    175g" "Smiths C
##  $ PROD_QTY      : int  2 3 2 5 3 1 1 1 1 2 ...
##  $ TOT_SALES     : num  6 6.3 2.9 15 13.8 5.1 5.7 3.6 3.9 7.2 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
summary(transactionData)
```

```
##       DATE         STORE_NBR     LYLTY_CARD_NBR        TXN_ID
##  Min.   :43282   Min.   :  1.0   Min.   :   1000   Min.   :      1
##  1st Qu.:43373   1st Qu.: 70.0   1st Qu.:  70021   1st Qu.:  67602
##  Median :43464   Median :130.0   Median : 130358   Median : 135138
##  Mean   :43464   Mean   :135.1   Mean   : 135550   Mean   : 135158
##  3rd Qu.:43555   3rd Qu.:203.0   3rd Qu.: 203094   3rd Qu.: 202701
##  Max.   :43646   Max.   :272.0   Max.   :2373711   Max.   :2415841
##     PROD_NBR       PROD_NAME            PROD_QTY
##  Min.   :  1.00   Length:264836      Min.   :  1.000
##  1st Qu.: 28.00   Class :character   1st Qu.:  2.000
##  Median : 56.00   Mode  :character   Median :  2.000
##  Mean   : 56.58                      Mean   :  1.907
##  3rd Qu.: 85.00                      3rd Qu.:  2.000
##  Max.   :114.00                      Max.   :200.000
##    TOT_SALES
##  Min.   :  1.500
```

```
##  1st Qu.:  5.400
##  Median :  7.400
##  Mean   :  7.304
##  3rd Qu.:  9.200
##  Max.   :650.000
```

```
colSums(is.na(transactionData))
```

```
##           DATE      STORE_NBR LYLTY_CARD_NBR         TXN_ID
##              0              0              0              0
##       PROD_NBR      PROD_NAME       PROD_QTY      TOT_SALES
##              0              0              0              0
```

Data Cleaning

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

```
transactionData[, SALSA := grepl("salsa", tolower(PROD_NAME))]
transactionData <- transactionData[SALSA == FALSE, ][, SALSA := NULL]
```

```
remove_outliers <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR_value <- Q3 - Q1
  return(x[x >= (Q1 - 1.5 * IQR_value) & x <= (Q3 + 1.5 * IQR_value)])
}
transactionData <- transactionData %>% filter(PROD_QTY %in% remove_outliers(PROD_QTY))
```

Feature Engineering

```
transactionData <- transactionData %>%
  mutate(
    Brand = word(PROD_NAME, 1),
    Pack_Size = as.numeric(str_extract(PROD_NAME, "\\d+"))
  )
```

Merging Data

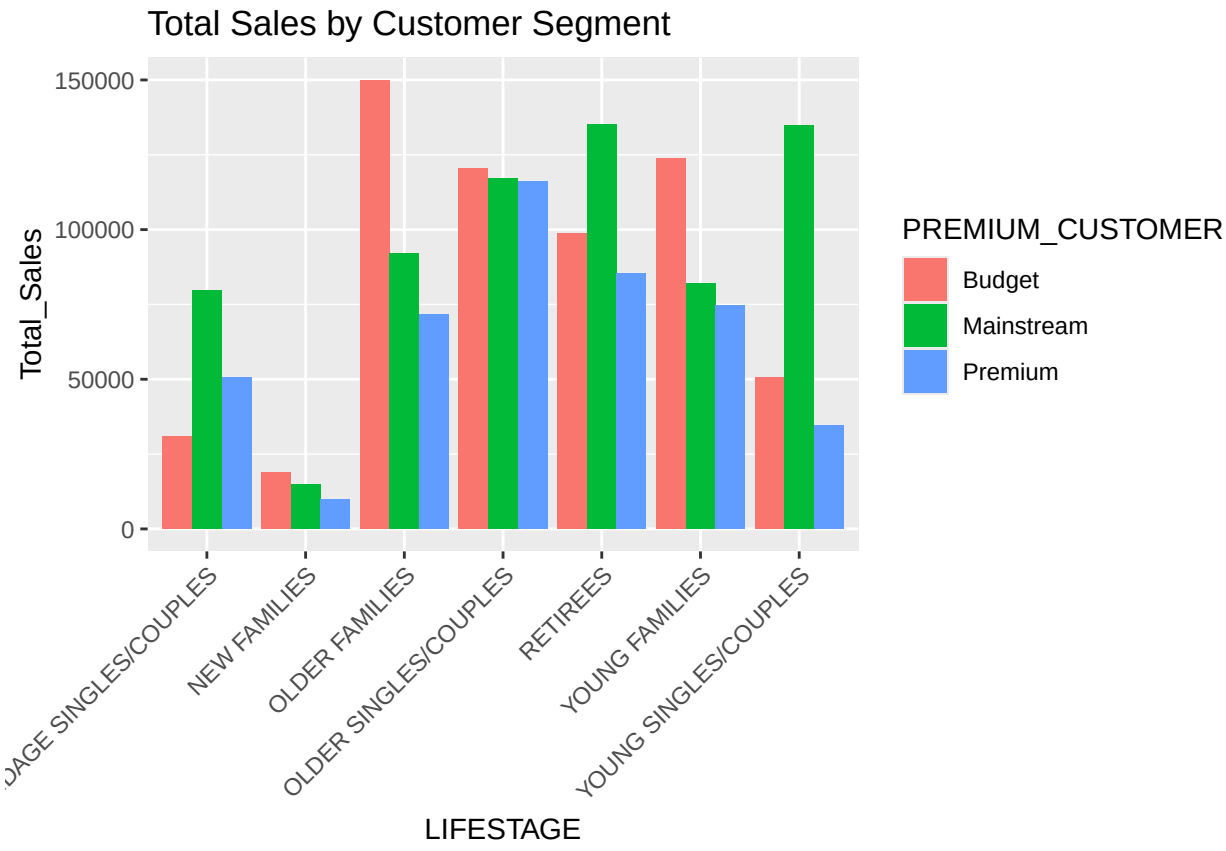
```
data <- merge(transactionData, customerData, all.x = TRUE)
```

Customer Segmentation Analysis

```
sales_summary <- data %>%
  group_by(LIFESTAGE, PREMIUM_CUSTOMER) %>%
  summarise(Total_Sales = sum(TOT_SALES, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'LIFESTAGE'. You can override using
## the `.groups` argument.
```

```
ggplot(sales_summary, aes(x=LIFESTAGE, y=Total_Sales, fill=PREMIUM_CUSTOMER)) +
  geom_bar(stat='identity', position='dodge') +
  labs(title='Total Sales by Customer Segment') +
  theme(axis.text.x = element_text(angle=45, hjust=1))
```



Save Final Cleaned Data

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