Quantium Virtual Internship - Retail Strategy and Analytics - Task 1

Eda AYDIN

Load required libraries

```
library(readxl)
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
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

Define File Path and Load Data

```
filePath <- ""  # Set working directory
# read excel file
transactions <- readxl::read_excel(paste0(filePath, "QVI_transaction_data.xlsx"))
customers <- fread(paste0(filePath, "QVI_purchase_behaviour.csv"))
```

Display the data

```r
head(transactions)
```

```
## # A tibble: 6 x 8
##    DATE STORE_NBR LYLTY_CARD_NBR TXN_ID PROD_NBR PROD_NAME    PROD_QTY TOT_SALES
##   <dbl>     <dbl>          <dbl>  <dbl>    <dbl> <chr>           <dbl>     <dbl>
## 1 43390         1           1000      1        5 Natural Chi~        2       6
## 2 43599         1           1307    348       66 CCs Nacho C~        3       6.3
## 3 43605         1           1343    383       61 Smiths Crin~        2       2.9
## 4 43329         2           2373    974       69 Smiths Chip~        5      15
## 5 43330         2           2426   1038      108 Kettle Tort~        3      13.8
## 6 43604         4           4074   2982       57 Old El Paso~        1       5.1
```

```r
head(customers)
```

```
##    LYLTY_CARD_NBR              LIFESTAGE PREMIUM_CUSTOMER
##             <int>                 <char>           <char>
## 1:           1000  YOUNG SINGLES/COUPLES          Premium
## 2:           1002  YOUNG SINGLES/COUPLES       Mainstream
## 3:           1003         YOUNG FAMILIES           Budget
## 4:           1004  OLDER SINGLES/COUPLES       Mainstream
## 5:           1005 MIDAGE SINGLES/COUPLES       Mainstream
## 6:           1007  YOUNG SINGLES/COUPLES           Budget
```

Summary Statistics

```r
summary(transactions)
```

```
##       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          TOT_SALES      
##  Min.   :  1.00   Length:264836      Min.   :  1.000   Min.   :  1.500  
##  1st Qu.: 28.00   Class :character   1st Qu.:  2.000   1st Qu.:  5.400  
##  Median : 56.00   Mode  :character   Median :  2.000   Median :  7.400  
##  Mean   : 56.58                      Mean   :  1.907   Mean   :  7.304  
##  3rd Qu.: 85.00                      3rd Qu.:  2.000   3rd Qu.:  9.200  
##  Max.   :114.00                      Max.   :200.000   Max.   :650.000  
```

```r
summary(customers)
```

```
##  LYLTY_CARD_NBR     LIFESTAGE         PREMIUM_CUSTOMER  
##  Min.   :   1000   Length:72637       Length:72637      
##  1st Qu.:  66202   Class :character   Class :character  
##  Median : 134040   Mode  :character   Mode  :character  
##  Mean   : 136186                                        
##  3rd Qu.: 203375                                        
##  Max.   :2373711                                        
```

```r
# number of rows
nrow(transactions)
```

```
## [1] 264836
```

```r
nrow(customers)
```

```
## [1] 72637
```

Variables Description

The transaction data contains the following variables:

- **DATE**: Date of purchase
- **STORE__NBR**: Store number
- **LYLTY__CARD__NBR**: Customer loyalty card number
- **TXN__ID**: Transaction ID
- **PROD__NBR**: Product number
- **PROD__NAME**: Product name
- **PROD__QTY**: Quantity of product purchased
- **TOT__SALES**: Total sales ($)

The customer data contains the following variables:

- **LYLTY__CARD__NBR**: Customer loyalty card number
- **LIFESTAGE**: Customer lifestage
- **PREMIUM__CUSTOMER**: Customer premium status

Data Cleaning

Missing Values

```r
colSums(is.na(transactions))
```

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

```r
colSums(is.na(customers))
```

```
##   LYLTY_CARD_NBR        LIFESTAGE PREMIUM_CUSTOMER
##                0                0                0
```

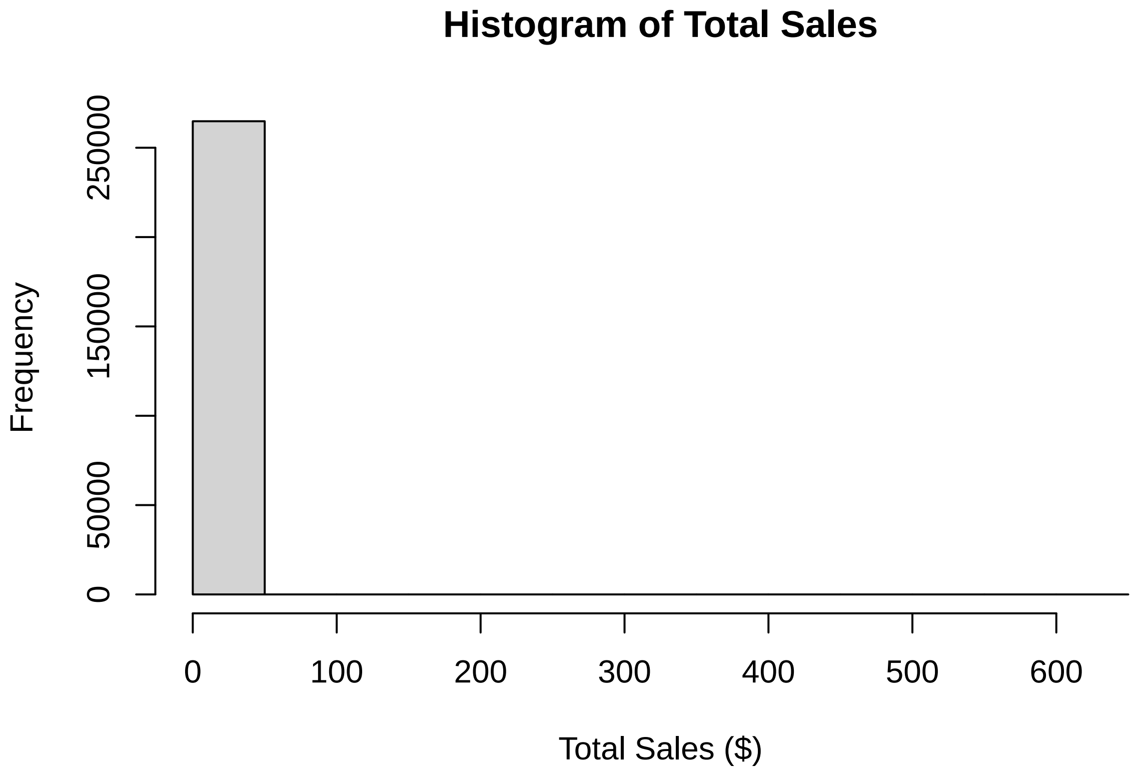
Fix the type

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

```
## # A tibble: 6 x 8
##   DATE       STORE_NBR LYLTY_CARD_NBR TXN_ID PROD_NBR PROD_NAME         PROD_QTY
##   <date>         <dbl>          <dbl>  <dbl>    <dbl> <chr>                <dbl>
## 1 2018-10-17         1           1000      1        5 Natural Chip    ~        2
## 2 2019-05-14         1           1307    348       66 CCs Nacho Cheese~        3
## 3 2019-05-20         1           1343    383       61 Smiths Crinkle C~        2
## 4 2018-08-17         2           2373    974       69 Smiths Chip Thin~        5
## 5 2018-08-18         2           2426   1038      108 Kettle Tortilla ~        3
## 6 2019-05-19         4           4074   2982       57 Old El Paso Sals~        1
## # i 1 more variable: TOT_SALES <dbl>
```

Outlier Detection

```
hist(transactions$TOT_SALES, main="Histogram of Total Sales", xlab="Total Sales ($)")
```

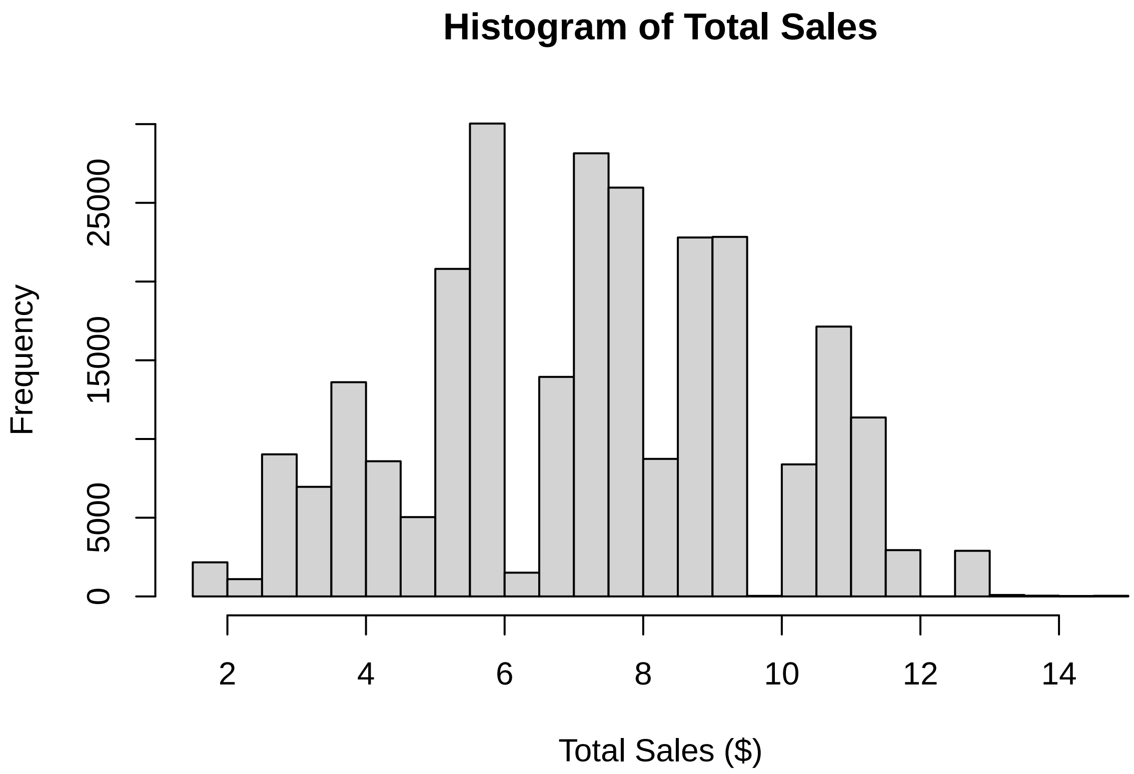


```
# remove outliers
q1 <- quantile(transactions$TOT_SALES, 0.25, na.rm = TRUE)  # First quartile
q3 <- quantile(transactions$TOT_SALES, 0.75, na.rm = TRUE)  # First quartile
IQR <- q3 - q1

lower_bound <- q1 - 1.5 * IQR
upper_bound <- q3 + 1.5 * IQR
```

```
transactions <- transactions[transactions$TOT_SALES <= upper_bound, ]
```

```
hist(transactions$TOT_SALES, main="Histogram of Total Sales", xlab="Total Sales ($)")
```



Merge data

```
merged_data <- merge(transactions, customers, by="LYLTY_CARD_NBR")
head(merged_data)
```

```
##   LYLTY_CARD_NBR       DATE STORE_NBR TXN_ID PROD_NBR
## 1           1000 2018-10-17         1      1        5
## 2           1002 2018-09-16         1      2       58
## 3           1003 2019-03-08         1      4      106
## 4           1003 2019-03-07         1      3       52
## 5           1004 2018-11-02         1      5       96
## 6           1005 2018-12-28         1      6       86
##                                 PROD_NAME PROD_QTY TOT_SALES
## 1 Natural Chip        Compny SeaSalt175g        2       6.0
## 2  Red Rock Deli Chikn&Garlic Aioli 150g        1       2.7
## 3 Natural ChipCo      Hony Soy Chckn175g        1       3.0
## 4  Grain Waves Sour    Cream&Chives 210G        1       3.6
## 5         WW Original Stacked Chips 160g        1       1.9
## 6                    Cheetos Puffs 165g        1       2.8
##                LIFESTAGE PREMIUM_CUSTOMER
## 1  YOUNG SINGLES/COUPLES          Premium
## 2  YOUNG SINGLES/COUPLES       Mainstream
## 3         YOUNG FAMILIES           Budget
## 4         YOUNG FAMILIES           Budget
## 5  OLDER SINGLES/COUPLES       Mainstream
## 6 MIDAGE SINGLES/COUPLES       Mainstream
```

Exploratory Data Analysis

```r
# Analyze the young singles/couples
young_singles_couples <- merged_data %>% filter(LIFESTAGE %in% c("YOUNG SINGLES/COUPLES"))
head(young_singles_couples)
```

```
##   LYLTY_CARD_NBR       DATE STORE_NBR TXN_ID PROD_NBR
## 1           1000 2018-10-17         1      1        5
## 2           1002 2018-09-16         1      2       58
## 3           1007 2018-12-05         1      8       10
## 4           1007 2018-12-04         1      7       49
## 5           1010 2018-09-09         1     10       51
## 6           1010 2018-12-14         1     11       59
##                                  PROD_NAME PROD_QTY TOT_SALES
## 1   Natural Chip        Compny SeaSalt175g        2       6.0
## 2    Red Rock Deli Chikn&Garlic Aioli 150g        1       2.7
## 3      RRD SR Slow Rst     Pork Belly 150g        1       2.7
## 4 Infuzions SourCream&Herbs Veg Strws 110g        1       3.8
## 5                 Doritos Mexicana    170g        2       8.8
## 6  Old El Paso Salsa   Dip Tomato Med 300g        1       5.1
##               LIFESTAGE PREMIUM_CUSTOMER
## 1 YOUNG SINGLES/COUPLES          Premium
## 2 YOUNG SINGLES/COUPLES       Mainstream
## 3 YOUNG SINGLES/COUPLES           Budget
## 4 YOUNG SINGLES/COUPLES           Budget
## 5 YOUNG SINGLES/COUPLES       Mainstream
## 6 YOUNG SINGLES/COUPLES       Mainstream
```

```r
summary(young_singles_couples)
```

```
##  LYLTY_CARD_NBR         DATE              STORE_NBR         TXN_ID
##  Min.   :   1000   Min.   :2018-07-01   Min.   :  1.0   Min.   :     1
##  1st Qu.:  65345   1st Qu.:2018-09-30   1st Qu.: 65.0   1st Qu.: 63089
##  Median : 133221   Median :2018-12-29   Median :133.0   Median :137478
##  Mean   : 135616   Mean   :2018-12-30   Mean   :135.1   Mean   :135184
##  3rd Qu.: 205375   3rd Qu.:2019-03-30   3rd Qu.:205.0   3rd Qu.:204443
##  Max.   :2373711   Max.   :2019-06-30   Max.   :272.0   Max.   :270205
##     PROD_NBR       PROD_NAME            PROD_QTY       TOT_SALES
##  Min.   :  1.00   Length:36321       Min.   :1.000   Min.   : 1.50
##  1st Qu.: 28.00   Class :character   1st Qu.:2.000   1st Qu.: 5.40
##  Median : 55.00   Mode  :character   Median :2.000   Median : 7.40
##  Mean   : 56.19                      Mean   :1.828   Mean   : 7.14
##  3rd Qu.: 84.00                      3rd Qu.:2.000   3rd Qu.: 8.80
##  Max.   :114.00                      Max.   :5.000   Max.   :14.80  
##   LIFESTAGE         PREMIUM_CUSTOMER  
##  Length:36321       Length:36321      
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
##                                       
##                                       
##                                       
```

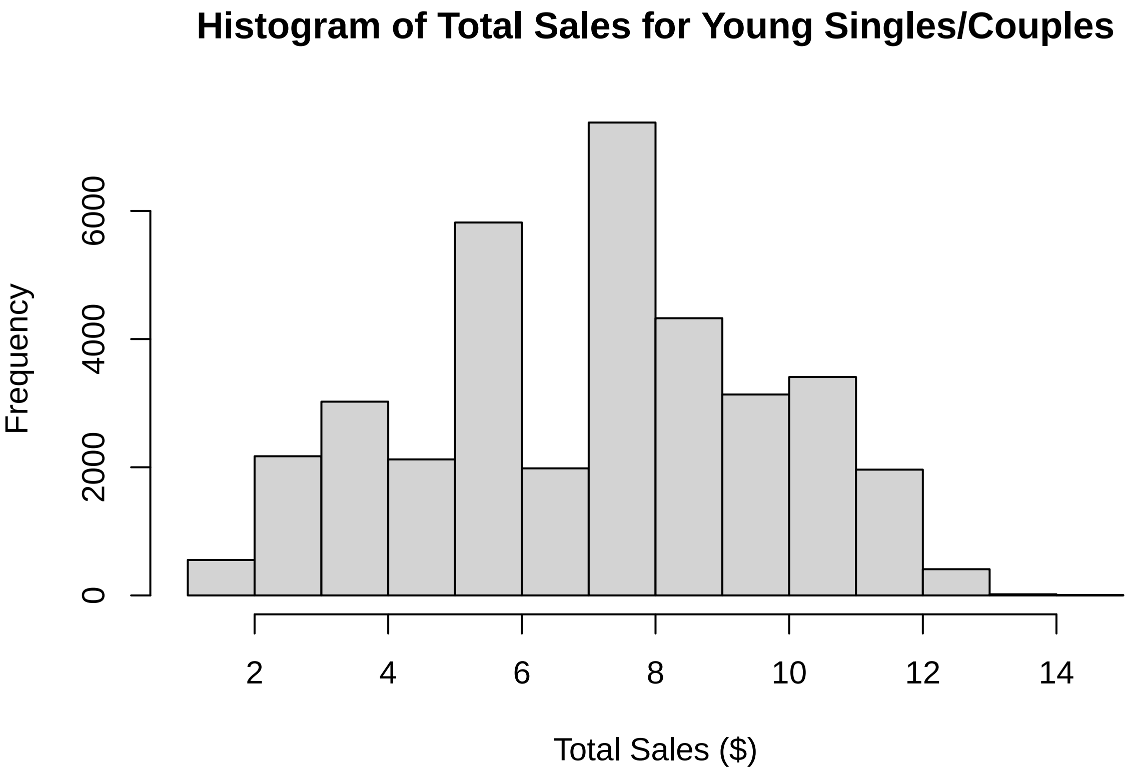
```r
str(young_singles_couples)
```

```
## ’data.frame’:    36321 obs. of  10 variables:
##  $ LYLTY_CARD_NBR  : num  1000 1002 1007 1007 1010 ...
##  $ DATE            : Date, format: "2018-10-17" "2018-09-16" ...
##  $ STORE_NBR       : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ TXN_ID          : num  1 2 8 7 10 11 22 23 24 26 ...
##  $ PROD_NBR        : num  5 58 10 49 51 59 3 97 38 19 ...
##  $ PROD_NAME       : chr  "Natural Chip       Compny SeaSalt175g" "Red Rock Deli Chikn&Garlic Aioli
##  $ PROD_QTY        : num  2 1 1 1 2 1 1 1 1 1 ...
##  $ TOT_SALES       : num  6 2.7 2.7 3.8 8.8 5.1 4.6 3 2.4 2.6 ...
##  $ LIFESTAGE       : chr  "YOUNG SINGLES/COUPLES" "YOUNG SINGLES/COUPLES" "YOUNG SINGLES/COUPLES" "Y
##  $ PREMIUM_CUSTOMER: chr  "Premium" "Mainstream" "Budget" "Budget" ...
```

```r
sum(young_singles_couples$TOT_SALES, na.rm = TRUE)
```

```
## [1] 259340
```

```r
hist(young_singles_couples$TOT_SALES, main="Histogram of Total Sales for Young Singles/Couples", xlab="
```



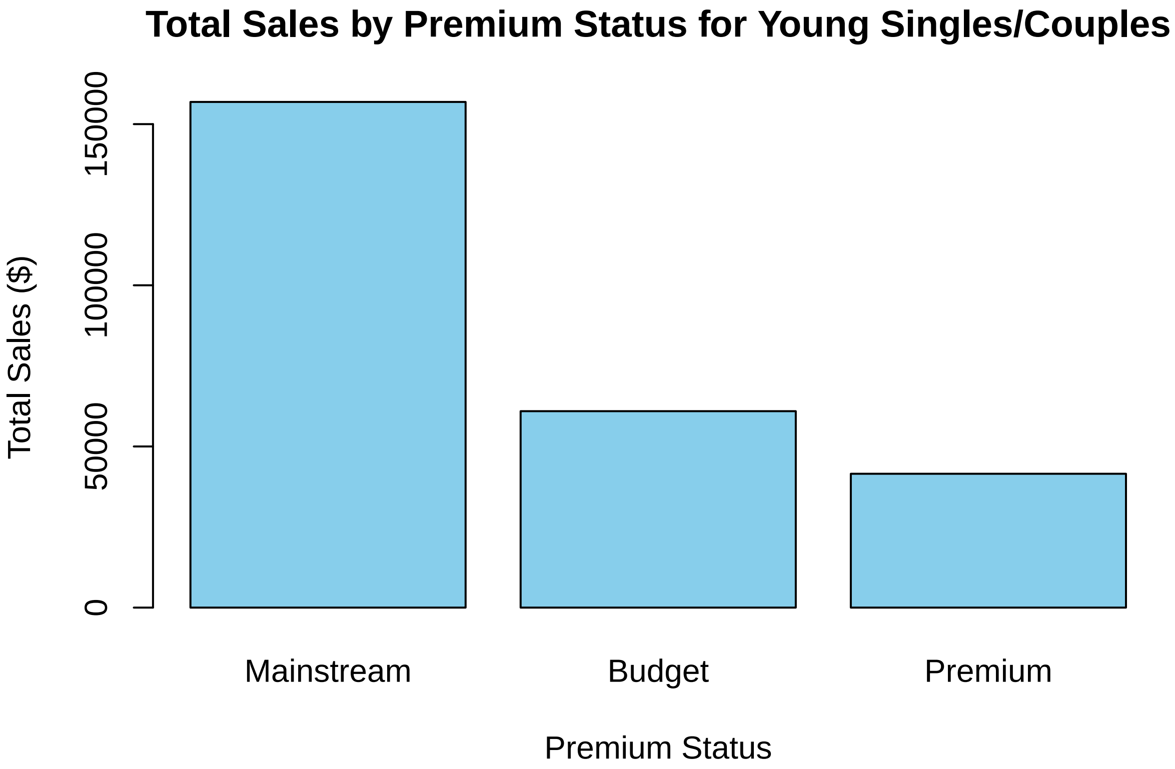
```
# Summarize total sales by product name
product_sales <- young_singles_couples %>%
  group_by(PROD_NAME) %>%
  summarise(Total_Sales = sum(TOT_SALES, na.rm = TRUE)) %>%
  arrange(desc(Total_Sales))  # Optional: Sort by total sales in descending order
head(product_sales)
```

```
## # A tibble: 6 x 2
##   PROD_NAME                                Total_Sales
##   <chr>                                          <dbl>
## 1 Dorito Corn Chp     Supreme 380g               5655
## 2 Smiths Crnkle Chip  Orgnl Big Bag 380g         5192
## 3 Kettle Mozzarella   Basil & Pesto 175g         5119.
## 4 Smiths Crinkle Chips Salt & Vinegar 330g       4930.
## 5 Doritos Cheese      Supreme 330g               4839.
## 6 Kettle Sweet Chilli And Sour Cream 175g        4709.
```

```
# Summarize total sales by premium status
premium_status_sales <- young_singles_couples %>%
        group_by(PREMIUM_CUSTOMER) %>%
        summarize(Total_sales = sum(TOT_SALES, na.rm = TRUE)) %>%
        arrange(desc(Total_sales))  # Optional: Sort by total sales in descending order
premium_status_sales
```

```
## # A tibble: 3 x 2
##   PREMIUM_CUSTOMER Total_sales
##   <chr>                  <dbl>
## 1 Mainstream           156882
## 2 Budget                60938.
## 3 Premium               41520.
```

```
# Barplot for total sales by premium status
barplot(premium_status_sales$Total_sales, names.arg = premium_status_sales$PREMIUM_CUSTOMER, main = "To
```



```
# Sample Mainstream Customers
mainstream_customers <- young_singles_couples[young_singles_couples$PREMIUM_CUSTOMER == "Mainstream", ]

# Summarize total sales by product name for mainstream customers
product_sales <- mainstream_customers %>%
  group_by(PROD_NAME) %>%
  summarise(Total_Sales = sum(TOT_SALES, na.rm = TRUE)) %>%
  arrange(desc(Total_Sales))  # Optional: Sort by total sales in descending order
head(product_sales)
```

```
## # A tibble: 6 x 2
##   PROD_NAME                                Total_Sales
##   <chr>                                          <dbl>
## 1 Dorito Corn Chp     Supreme 380g               3660.
## 2 Smiths Crnkle Chip  Orgnl Big Bag 380g         3481
## 3 Kettle Mozzarella   Basil & Pesto 175g         3359.
## 4 Smiths Crinkle Chips Salt & Vinegar 330g       3317.
## 5 Doritos Cheese      Supreme 330g               3169.
## 6 Cheezels Cheese 330g                           3089.
```

```
# Sample Mainstream Customers
budget_customers <- young_singles_couples[young_singles_couples$PREMIUM_CUSTOMER == "Budget", ]

# Summarize total sales by product name for mainstream customers
product_sales <- budget_customers %>%
  group_by(PROD_NAME) %>%
  summarise(Total_Sales = sum(TOT_SALES, na.rm = TRUE)) %>%
  arrange(desc(Total_Sales))  # Optional: Sort by total sales in descending order
head(product_sales)
```

```
## # A tibble: 6 x 2
##   PROD_NAME                                Total_Sales
##   <chr>                                          <dbl>
## 1 Dorito Corn Chp     Supreme 380g               1164.
## 2 Doritos Cheese      Supreme 330g               1077.
## 3 Kettle Sea Salt     And Vinegar 175g           1069.
## 4 Kettle Mozzarella   Basil & Pesto 175g         1004.
## 5 Smiths Crinkle Chips Salt & Vinegar 330g        969
## 6 Old El Paso Salsa   Dip Tomato Med 300g         949.
```

```
# Sample Mainstream Customers
premium_customers <- young_singles_couples[young_singles_couples$PREMIUM_CUSTOMER == "Premium", ]

# Summarize total sales by product name for mainstream customers
product_sales <- premium_customers %>%
  group_by(PROD_NAME) %>%
  summarise(Total_Sales = sum(TOT_SALES, na.rm = TRUE)) %>%
  arrange(desc(Total_Sales))  # Optional: Sort by total sales in descending order
head(product_sales)
```

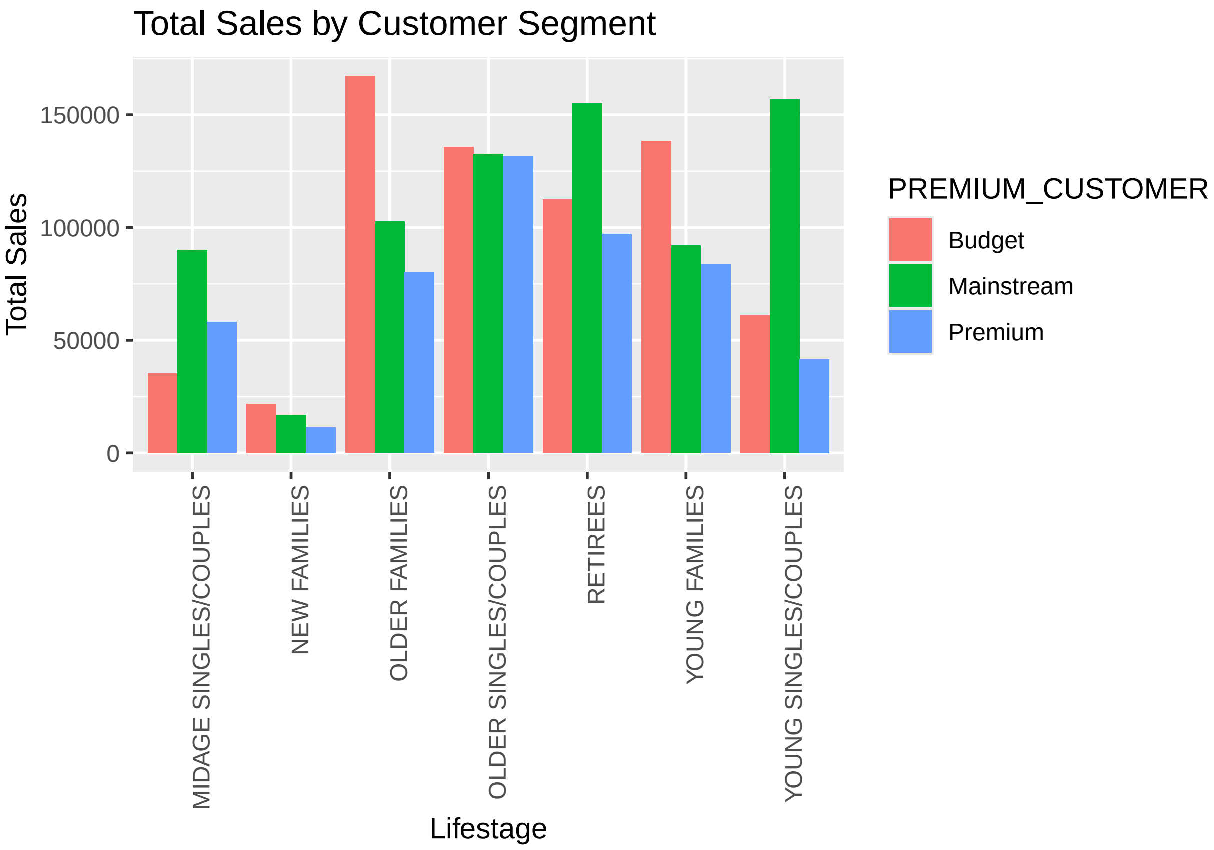
```
## # A tibble: 6 x 2
##   PROD_NAME                               Total_Sales
##   <chr>                                         <dbl>
## 1 Dorito Corn Chp     Supreme 380g               832
## 2 Smiths Crnkle Chip  Orgnl Big Bag 380g         808.
## 3 Kettle Mozzarella   Basil & Pesto 175g         756
## 4 Kettle Sweet Chilli And Sour Cream 175g        724.
## 5 Tostitos Splash Of  Lime 175g                  713.
## 6 Smiths Crinkle      Original 330g              707.
```

```
library(ggplot2)
library(data.table)

#### Aggregate Sales Data
sales_summary <- merged_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.
```

```
#### Plot Total Sales by Customer Segment
ggplot(sales_summary, aes(x = LIFESTAGE, y = Total_Sales, fill = PREMIUM_CUSTOMER)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  geom_col(position = "dodge") +
  labs(title = "Total Sales by Customer Segment", x = "Lifestage", y = "Total Sales")
```



```r
t_test <- t.test(
  merged_data$TOT_SALES[merged_data$PREMIUM_CUSTOMER == "Mainstream" & merged_data$LIFESTAGE == "YOUNG S
  merged_data$TOT_SALES[merged_data$PREMIUM_CUSTOMER == "Premium" & merged_data$LIFESTAGE == "YOUNG SIN
)

print(t_test)
```

```
##
##  Welch Two Sample t-test
##
## data:  merged_data$TOT_SALES[merged_data$PREMIUM_CUSTOMER == "Mainstream" & merged_data$LIFESTAGE ==
## t = 24.334, df = 9685.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8447455 0.9927645
## sample estimates:
## mean of x mean of y
##  7.536606  6.617851
```