```
library(readxl)
transaction <- read_excel("Desktop/Quantium/QVI_transaction_data.xlsx")

library(readr)
purchase <- read_csv("Desktop/Quantium/QVI_purchase_behaviour.csv")

library(tidyverse)
library(lubridate)
library(stringr)
###clean transaction data
glimpse(transaction)
summary(transaction)

#card number = store number + card number

#change data type
transaction$DATE <- as.Date(transaction$DATE, origin = "1899-12-30")
transaction$STORE_NBR <- as.factor(transaction$STORE_NBR)
transaction$LYLTY_CARD_NBR <- as.factor(transaction$LYLTY_CARD_NBR)
transaction$PROD_NBR <- as.factor(transaction$PROD_NBR)
transaction$PROD_QTY <- as.numeric(transaction$PROD_QTY)
transaction$TXN_ID <- as.numeric(transaction$TXN_ID)

summary(transaction$PROD_NAME)
unique(transaction$PROD_NAME)

Brand <- word(transaction$PROD_NAME,1)
unique(Brand)
transaction$BRAND <- Brand

#clear error entry of product
convert <- which(transaction$BRAND == 'Dorito')
transaction$BRAND[convert]<- str_replace(transaction$BRAND[convert], 'Dorito','Doritos')

convert2 <- which(transaction$BRAND == 'Snbts')
transaction$BRAND[convert2]<- str_replace(transaction$BRAND[convert2], 'Snbts','Sunbites')

convert3 <- which(transaction$BRAND == 'Smith')
transaction$BRAND[convert3]<- str_replace(transaction$BRAND[convert3], 'Smith','Smiths')

convert4 <- which(transaction$BRAND == 'RED')
transaction$BRAND[convert4]<- str_replace(transaction$BRAND[convert4], 'RED','RRD')

convert5 <- which(transaction$BRAND == 'Red')
```

```
transaction$BRAND[convert5]<- str_replace(transaction$BRAND[convert5], 'Red','RRD')

transaction$BRAND <- as.factor(transaction$BRAND)

summary(transaction)

#remove salsa
tolower(transaction$PROD_NAME)
salsa <- which(str_detect(tolower(transaction$PROD_NAME), 'salsa'))
transaction <- transaction[-salsa, ]

#remove n/a
summary(transaction)
which(is.na(transaction) == 'True')

#outliers for quality
transaction[which(transaction$PROD_QTY == 200),] #same person
transaction[which(transaction$LYLTY_CARD_NBR == 226000),] #maybe for business use
transaction <- transaction[-which(transaction$PROD_QTY == 200),]
summary(transaction)

#missing date
num_tra_date <- transaction %>% group_by(DATE) %>%
  dplyr::summarise(num_transactions = n())
head(num_tra_date)
tail(num_tra_date)
date <- seq.Date(from = as.Date("2018/07/01",format = "%Y/%m/%d"),
            to = as.Date("2019/06/30",format = "%Y/%m/%d"),
            by = 'day')
date <- tibble('date'=date)
full_date <- right_join(num_tra_date, date, by = c('DATE'='date'))

#### Setting plot themes to format graphs
theme_set(theme_bw())
theme_update(plot.title = element_text(hjust = 0.5))

#### Plot transactions over time
ggplot(full_date, aes(x = DATE, y = num_transactions)) +
  geom_line() +
  labs(x = "Day", y = "Number of transactions", title = "Transactions over time") +
  scale_x_date(breaks = "1 month") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))

#Zoom in for December
```

```
Dec <- filter(full_date, DATE >= '2018/12/15' & DATE <= '2019/01/01' )

ggplot(Dec, aes(x = DATE, y = num_transactions)) +
  geom_line() +
  labs(x = "Day", y = "Number of transactions", title = "Transactions over time") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))

#Pack size
transaction$PACK_SIZE <- parse_number(transaction$PROD_NAME)

summary(transaction$PACK_SIZE)  #make sense

# Plot a histogram showing the number of transactions by pack size
num_tra_packsize <- transaction %>% group_by(PACK_SIZE) %>%
  dplyr::summarise(num_transactions = n())
head(num_tra_packsize)
hist(transaction$PACK_SIZE,
    main = 'Number of transactions by pack size',
    xlab = 'Pack Size',
    ylab = 'Number of Transctions')



#####clean customer data
glimpse(purchase)

#change data type
purchase$LIFESTAGE <- as.factor(purchase$LIFESTAGE)
purchase$PREMIUM_CUSTOMER <- as.factor(purchase$PREMIUM_CUSTOMER)

#### Merge transaction data to customer data
data <- merge(transaction, purchase, all.x = TRUE)

#Check for missing customer details
which(is.na(data$LIFESTAGE))
which(is.na(data$PREMIUM_CUSTOMER))

#save as csv
library(data.table)
fwrite(data, paste0("Desktop/Quantium/","QVI_data.csv"))


###- Who spends the most on chips (total sales), describing customers by lifestage and
#how premium their general purchasing behaviour is
```

```
summary(data)
spend_most <- data[which(data$TOT_SALES == 29.5),]
summary(spend_most) #都买的 Smith 且 5 包


### calculating total sales by LIFESTAGE and PREMIUM_CUSTOMER and
# plotting the split by these segments to describe which customer segment contribute
# most to chip sales.

total_sales_customers <- data.frame(aggregate(data$TOT_SALES,
                            by=list(data$LIFESTAGE,data$PREMIUM_CUSTOMER), FUN=sum))
colnames(total_sales_customers) <- c('life_stage','member','total_sale')
total_sales_customers <-
total_sales_customers[order(total_sales_customers$total_sale,decreasing = TRUE),]

ggplot(data = total_sales_customers[order(total_sales_customers$total_sale,decreasing =
TRUE),],
      mapping = aes(x = life_stage,
                    y = total_sale)) +
   geom_bar(stat = 'identity', aes(fill = member)) +
   theme(axis.text.x = element_text(angle=90, vjust=0.6)) +
   labs(title = 'Total Sales by Lifestage and Premium',
        x = 'Life Stage',
        y = 'Total Sales ($)',
        fill = 'Member',
        size = 5) +
   theme(panel.grid.major = element_blank(),
         panel.grid.minor = element_blank())

ggplot(data = total_sales_customers[order(total_sales_customers$total_sale,decreasing =
TRUE),],
      mapping = aes(x = member,
                    y = total_sale)) +
   geom_bar(stat = 'identity', aes(fill = life_stage)) +
   theme(axis.text.x = element_text(angle=90, vjust=0.6)) +
   labs(title = 'Total Sales by Member',
        x = 'Life Stage',
        y = 'Total Sales ($)',
        fill = 'Life Stage',
        size = 5) +
   theme(panel.grid.major = element_blank(),
         panel.grid.minor = element_blank())
```

```
ggplot(data = total_sales_customers[order(total_sales_customers$total_sale,decreasing =
TRUE),],
      mapping = aes(x = life_stage,
                    y = total_sale,
                    shape = member,
                    color = member)) +
   geom_point() +
   theme(axis.text.x = element_text(angle=90, vjust=0.6)) +
   labs(title = 'Total Sales by Lifestage',
        x = 'Life Stage',
        y = 'Total Sales ($)',
        size = 5) +
   theme(panel.grid.major = element_blank(),
         panel.grid.minor = element_blank())
#Sales are coming mainly from Budget - older families, Mainstream - young
# singles/couples, and Mainstream - retirees
# if the higher sales are due to there being more customers who buy chips
#### Number of customers by LIFESTAGE and PREMIUM_CUSTOMER
summary(data$LIFESTAGE) #number of transaction insteads of customer

num_customers <- data %>% group_by(LIFESTAGE, PREMIUM_CUSTOMER) %>%
   dplyr::summarise(
      num_transctions = n(),
      num_customers = length(unique(LYLTY_CARD_NBR)))

num_customers[order(num_customers$num_customers,decreasing = TRUE),]

ggplot(data = num_customers, mapping = aes(x = LIFESTAGE,
                                y = num_customers,
                                shape = PREMIUM_CUSTOMER,
                                color = PREMIUM_CUSTOMER)) +
   geom_point()+
   theme(axis.text.x = element_text(angle=90, vjust=0.6)) +
   labs(title = 'Customer Numbers by Lifestage and Premium',
        x = 'Life Stage',
        y = 'Customer Number',
        fill = 'Member',
        size = 5) +
   theme(panel.grid.major = element_blank(),
         panel.grid.minor = element_blank())


#There are more Mainstream - young singles/couples and Mainstream - retirees who buy
# chips. This contributes to there being more sales to these customer segments but
```

```
#this is not a major driver for the Budget - Older families segment

#Higher sales may also be driven by more units of chips being bought per customer.
#### Average number of units per customer by LIFESTAGE and PREMIUM_CUSTOMER
avg_qty <- data %>% group_by(LIFESTAGE, PREMIUM_CUSTOMER) %>%
  dplyr::summarise(
    num_transactions = n(),
    avg_qty = sum(PROD_QTY)/length(unique(LYLTY_CARD_NBR)))
avg_qty[order(avg_qty$avg_qty, decreasing = TRUE),]


ggplot(data = avg_qty, mapping = aes(x = LIFESTAGE,
                          y = avg_qty,
                          shape = PREMIUM_CUSTOMER,
                          color = PREMIUM_CUSTOMER)) +
  geom_point()+
  theme(axis.text.x = element_text(angle=90, vjust=0.6)) +
  labs(title = 'Average Quantity by Lifestage and Premium',
       x = 'Life Stage',
       y = 'Customer Number',
       fill = 'Member',
       size = 5) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
#Older families and young families in general buy more chips per customer

#Let's also investigate the average price per unit chips bought for each customer
# segment as this is also a driver of total sales.
avg_price <- data %>% group_by(LIFESTAGE, PREMIUM_CUSTOMER) %>%
  dplyr::summarise(
    num_transactions = n(),
    total_sale = sum(TOT_SALES),
    avg_sale = sum(TOT_SALES)/length(unique(LYLTY_CARD_NBR)),
    avg_sale_transaction = sum(TOT_SALES)/sum(PROD_QTY))

avg_price[order(avg_price$avg_sale_transaction, decreasing = TRUE),]

##Mainstream midage and young singles and couples are more willing to pay more per
#packet of chips compared to their budget and premium counterparts. This may be due
#to premium shoppers being more likely to buy healthy snacks and when they buy
#chips, this is mainly for entertainment purposes rather than their own consumption.
#This is also supported by there being fewer premium midage and young singles and
#couples buying chips compared to their mainstream counterparts.
```

```
#The difference of avg_sale_tr isn't large, then we can check for significant difference
#t-test

#### Perform an independent t-test between mainstream vs premium and budget midage and
#### young singles and couples
data <- as.data.table(data)
pricePerUnit = data[, price := TOT_SALES/PROD_QTY]
t.test(data[LIFESTAGE %in% c("YOUNG SINGLES/COUPLES", "MIDAGE SINGLES/COUPLES")
         & PREMIUM_CUSTOMER == "Mainstream", price],
     data[LIFESTAGE %in% c("YOUNG SINGLES/COUPLES", "MIDAGE SINGLES/COUPLES")
         & PREMIUM_CUSTOMER != "Mainstream", price],
     alternative = "greater")
#p<0.05, significant difference, mainstream are significantly higher than budget or premium

## Deep dive into specific customer segments for insights
#### Deep dive into Mainstream, young singles/couples
# Over to you! Work out of there are brands that these two customer segments prefer
# more than others. You could use a technique called affinity analysis or a-priori
# analysis (or any other method if you prefer)
library(arules)
library(arulesViz)

main_young <- data[data$LIFESTAGE == 'YOUNG SINGLES/COUPLES' &
                   data$PREMIUM_CUSTOMER == 'Mainstream', ]
main_young_brand <- main_young$BRAND

write.csv(main_young_brand, "try.csv")
tr <- read.transactions("try.csv", format = 'basket')

itemFrequencyPlot(tr, topN = 10, type = 'absolute',
            main = 'Absolute Item Frequency Plot of Mainstream, young singles/couples',
            xlab = 'Brands',
            ylab = 'Item Absolute Frequency')

other <- data[!(LIFESTAGE == "YOUNG SINGLES/COUPLES" & PREMIUM_CUSTOMER
=="Mainstream"),]
#### Brand affinity compared to the rest of the population
quantity_main_young <- main_young[, sum(PROD_QTY)]
quantity_other <- other[, sum(PROD_QTY)]

quantity_main_young_by_brand <-
  main_young[, .(targetSegment = sum(PROD_QTY)/quantity_main_young), by = BRAND]
quantity_other_by_brand <-
  other[, .(other = sum(PROD_QTY)/quantity_other), by = BRAND]
```

```
brand_proportions <- merge(quantity_main_young_by_brand, quantity_other_by_brand)[,
affinityToBrand := targetSegment/other]
brand_proportions[order(-affinityToBrand)]

ggplot(brand_proportions,
aes(brand_proportions$BRAND,brand_proportions$affinityToBrand)) +
  geom_bar(stat = "identity",fill = "green") +
  labs(x = "Brand",
       y = "Customers Affinity to Brand",
       title = "Favorite brands of Customers") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
## Mainstream young singles/couples are 23% more likely to purchase Tyrrells chips compared
to the rest of the population
## Mainstream young singles/couples are 56% less likely to purchase Burger Rings compared to
the rest of the population

#### Preferred pack size compared to the rest of the population

quantity_main_young_by_pack<-
  main_young[, .(targetSegment = sum(PROD_QTY)/quantity_main_young), by = PACK_SIZE]
quantity_other_by_pack <-
  other[, .(other = sum(PROD_QTY)/quantity_other), by = PACK_SIZE]

pack_proportions <- merge(quantity_main_young_by_pack, quantity_other_by_pack)[,
affinityToPackSize := targetSegment/other]
pack_proportions[order(-affinityToPackSize)]

ggplot(pack_proportions,
aes(as.factor(pack_proportions$PACK_SIZE),pack_proportions$affinityToPackSize)) +
  geom_bar(stat = "identity",width =0.5, fill = "green") +
  labs(x = "Pack Size",
       y = "Customers Affinity to Pack Size",
       title = "Favorite pack sizes of Customers") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))+
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```