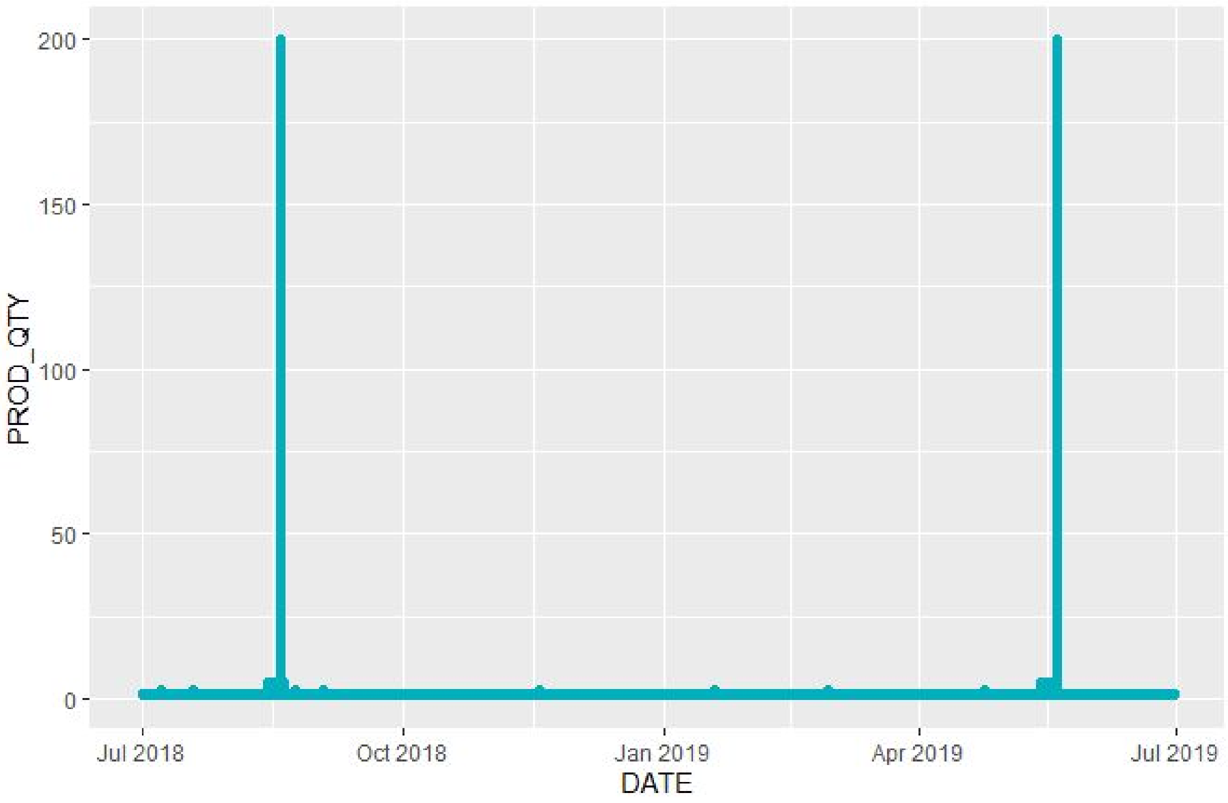
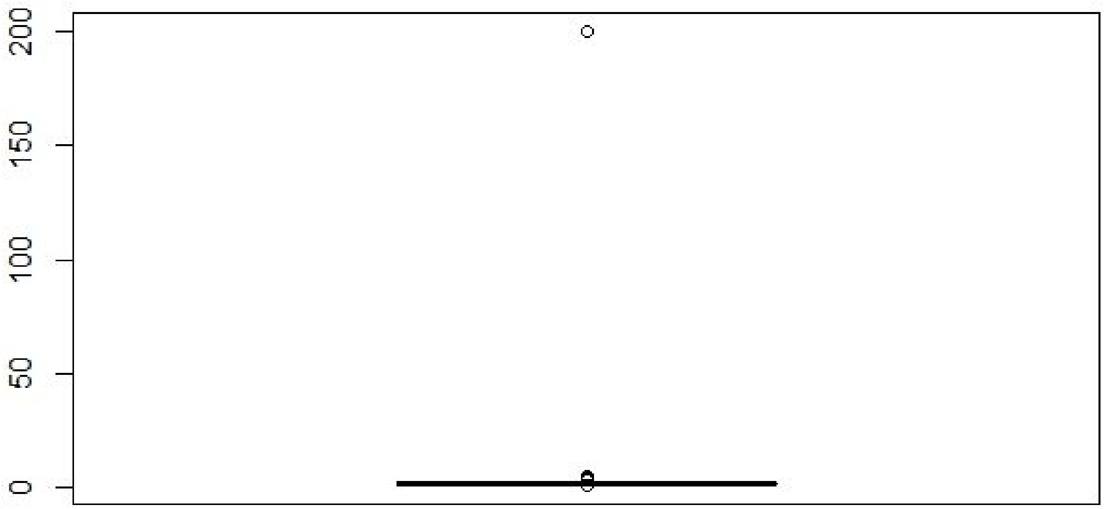
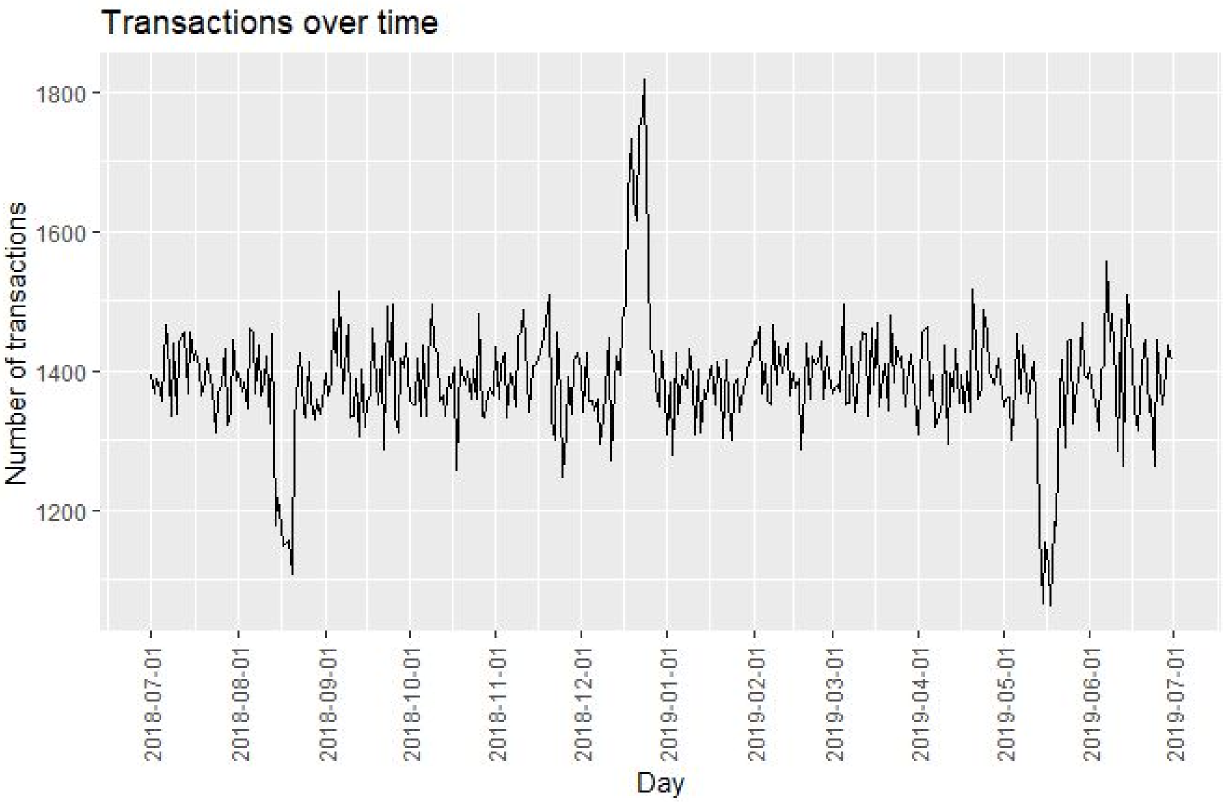
```
## Load required libraries and datasets
install.packages("data.tables")
install.packages("ggplot2")
install.packages("ggmosaic")
install.packages("readr")
install.packages("stringr")
install.packages("arulesViz")
library(data.table)
library(ggplot2)
library(ggmosaic)
library(readr)
library(dplyr)
library(stringr)
library(arulesViz)
library(knitr)
#### Point the filePath to where you have downloaded the datasets to and
#### assign the data files to data.tables
transactiondata <- QVI_transaction_data_1_
purchasedata <- QVI_purchase_behaviour_1_
`transactiondata`
## Exploratory data analysis
  #examining transaction data
head(transactiondata)
class(transactiondata)
sapply(transactiondata, class)
class(transactiondata$DATE)

#date column is in integer format changing that to date format
transactiondata$DATE <- as.Date(transactiondata$DATE, origin = "1899-12-30")
summary(transactiondata$PROD_NAME)
head(transactiondata$PROD_NAME)

##Looks like we are definitely looking at potato chips but how can we check that
##these are all chips? We can do some basic text analysis by summarising the
##individual words in the product name.
productWords <- data.frame(unlist(strsplit(unique(transacdata[, transacdata$PROD_NAME]), "
")))
setnames(productWords, 'words')
options(max.print = 100000000)


##As we are only interested in words that will tell  us if the product is chips or
##not, let's remove all words with digits  and special characters such as '&' from our
##set of product words. We can do this using `grepl()`.
```

```
grepl("&", transacdata)
x <- gsub("&", " ", transactiondata$PROD_NAME)
transacdata$PROD_NAME <- gsub('[[:digit:]]+', '', transacdata$PROD_NAME)
gsub("Hot & Spicy", "", transacdata$PROD_NAME)
gsub("Light&", "", transacdata$PROD_NAME)
is.na(transacdata)
gsub("[[:punct:]]", "", transacdata$PROD_NAME)
gsub("&", "", transacdata$PROD_NAME)
transacdata$PROD_NAME <- gsub("/", "", transactiondata$PROD_NAME)
transacdata$PROD_NAME

##There are salsa products in the dataset but we are only interested in the chips
##category, so let's remove these.
transactiondata$PROD_NAME <-  filter(!grep("Salsa", transactiondata$PROD_NAME))
boxplot(transactiondata$PROD_QTY)
transactiondata <- subset(transactiondata, transactiondata$PROD_QTY <200)

#### Summarise the data to check for nulls and possible outliers
boxplot(transactiondata$PROD_QTY)
ggplot(data = transactiondata, aes(x = DATE, y = PROD_QTY))+
  geom_line(color = "#00AFBB", size = 2)
```





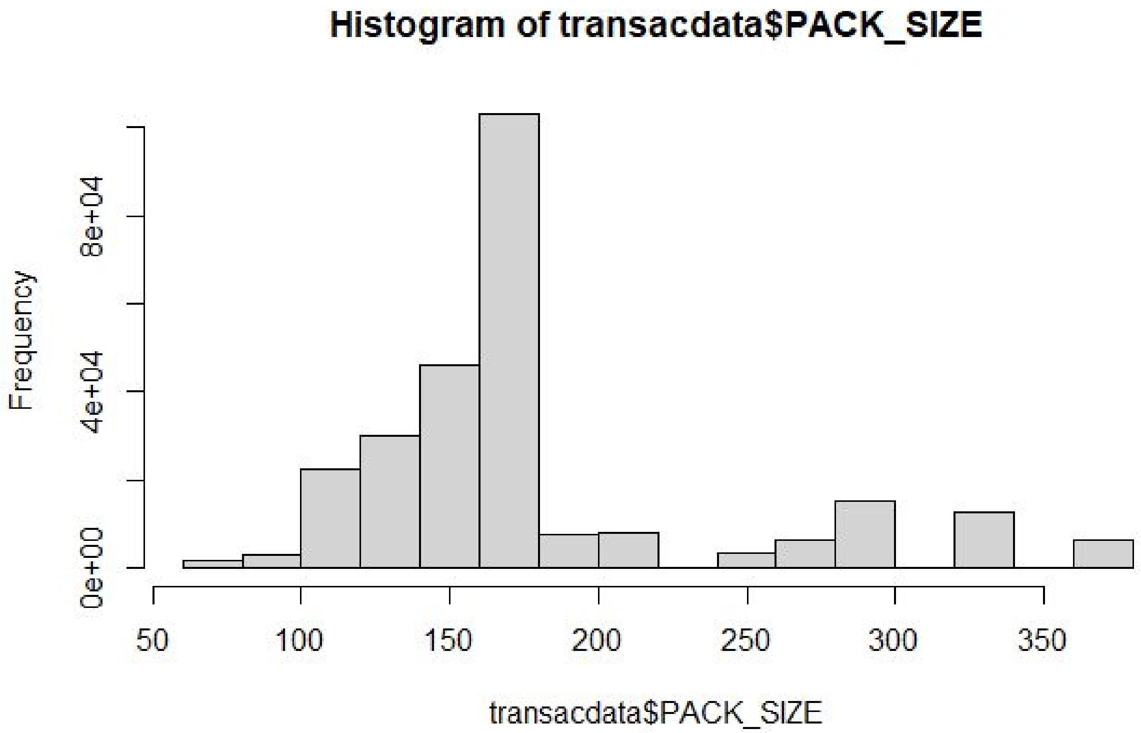
```
##There are no nulls in the columns but product quantity appears to have an outlier
##which we should investigate further. Let's investigate further the case where 200
##packets of chips are bought in one transaction.
subset(transacdata, PROD_QTY < 200)

##That's better. Now, let's look at the number of transaction lines over time to see
##if there are any obvious data issues such as missing data.
x <- seq(as.Date("2018-7-1"), as.Date("2019-6-30"), by = "day")
transacdata$DATE
transaction_by_day <- aggregate(transacdata$PROD_QTY, by=list(transacdata$DATE), sum)

#### Setting plot themes to format graphs
theme_set(theme_bw())
theme_update(plot.title = element_text(hjust = 0.5))
#### Plot transactions over time
ggplot(transaction_by_day, aes(x = Group.1, y = x)) +
  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))
```



##We can see that the increase in sales occurs in the lead-up to Christmas and that
##there are zero sales on Christmas day itself. This is due to shops being closed on
##Christmas day.

```
#### Pack size
#### We can work this out by taking the digits that are in PROD_NAME
table1[, PACK_SIZE := parse_number(PROD_NAME)]
packsize <- table1[, .N, PACK_SIZE][order(PACK_SIZE)]
```

```
#### Let's plot a histogram of PACK_SIZE since we know that it is a categorical
##variable and not a continuous variable even though it is numeric.
transacdata <- table1
hist(x = transacdata$PACK_SIZE)
```

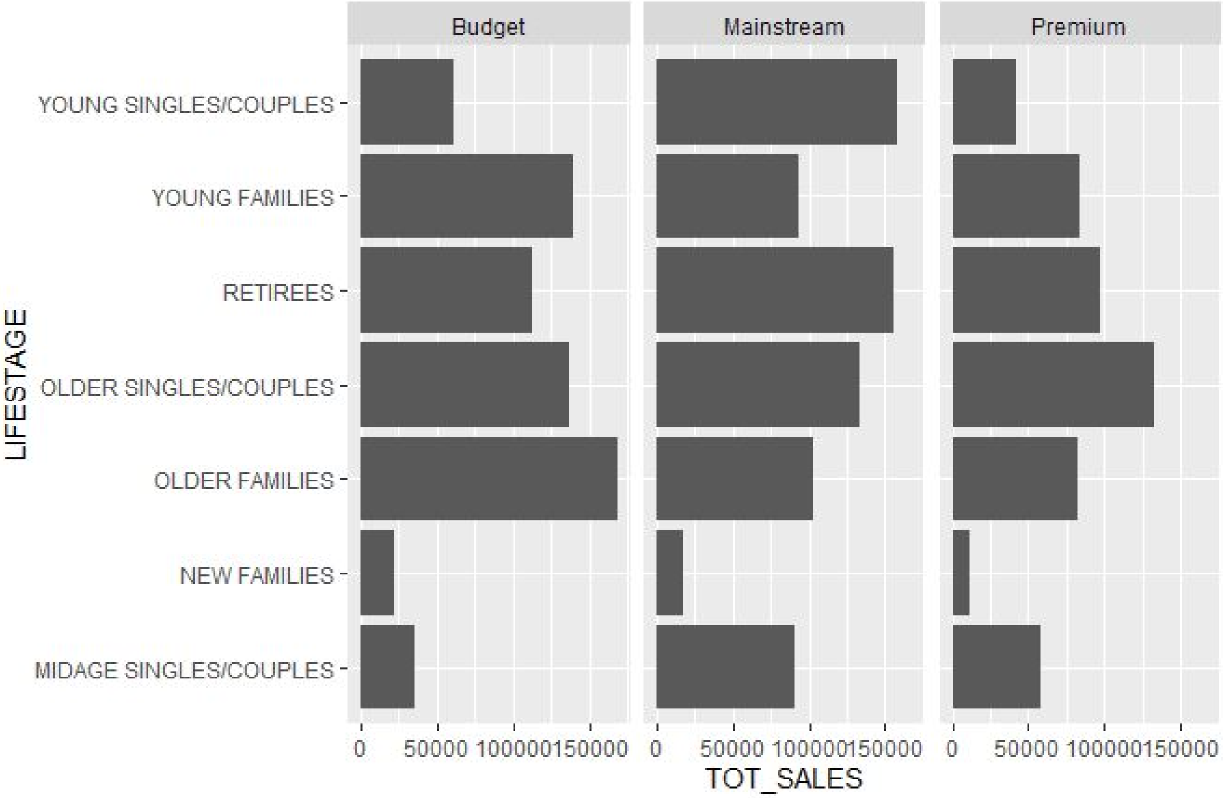


```
#### Brands
transacdata$BRAND_NAME <- word(transacdata$PROD_NAME, 1)
#### Checking brands
unique(transacdata$BRAND_NAME)
transacdata$BRAND_NAME <-  gsub("RED", "Red", transacdata$BRAND_NAME)
transacdata$BRAND_NAME
unique(transacdata$BRAND_NAME)
#### Clean brand names
transacdata$BRAND_NAME <-  gsub("Snbts", "Sunbites", transacdata$BRAND_NAME)
transacdata$BRAND_NAME <-  gsub("Burger", "BurgerRings", transacdata$BRAND_NAME)


summary(purchasedata)
unique(purchasedata)
is.na(purchasedata)
#### Merge transaction data to customer data
data <- merge(transacdata, purchasedata, all.x = TRUE)
is.na.data.frame(data)
##saving csv for task 2
fwrite(data, paste0("c:/Users/PAARTH/Desktop/rprojs/insidesherpa","QVI_data.csv"))
write.csv(data,"c:/Users/PAARTH/Desktop/rprojs/insidesherpa", row.names = FALSE)
```

## Data analysis on customer segments

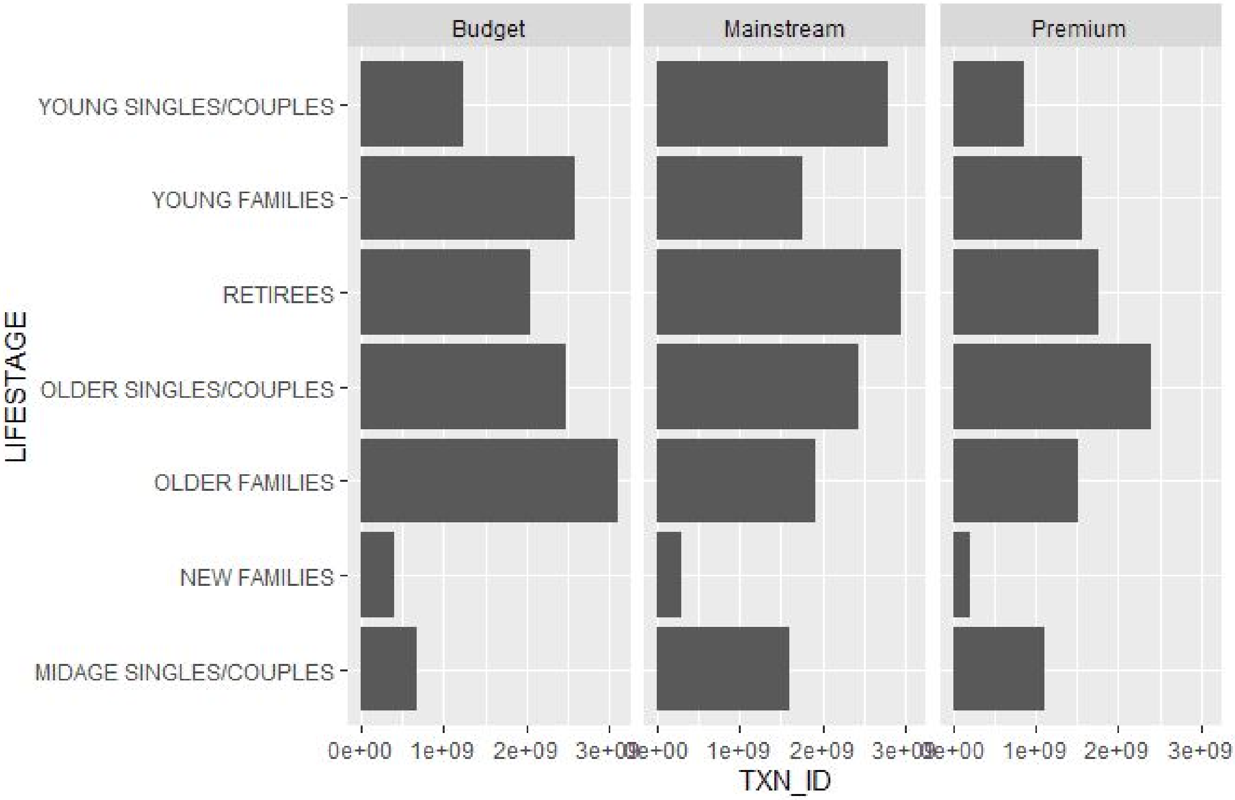
#### Total sales by LIFESTAGE and PREMIUM_CUSTOMER

```
ggplot(data = data, aes(x=TOT_SALES, y = LIFESTAGE))+
  geom_col()+
  facet_grid(~PREMIUM_CUSTOMER)
##Sales are coming mainly from Budget - older families, Mainstream - young
##singles/couples, and Mainstream - retirees
```

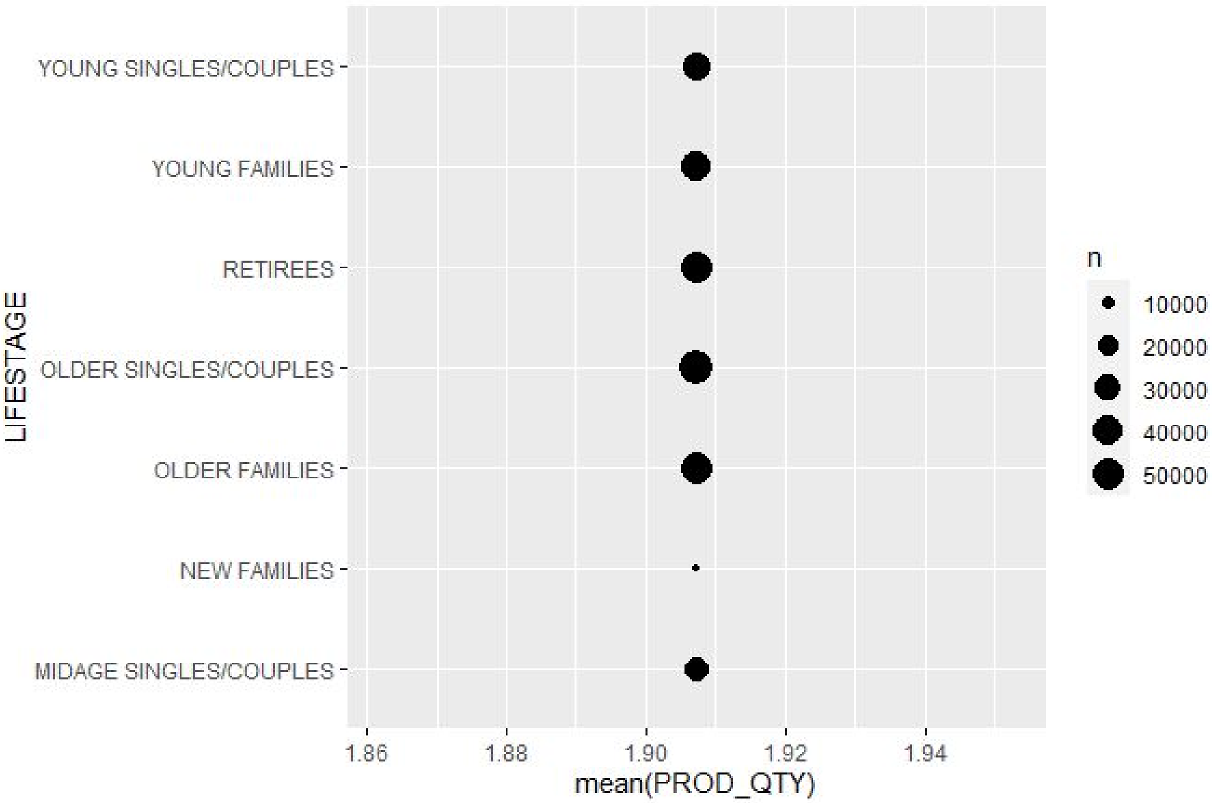


#### Number of customers by LIFESTAGE and PREMIUM_CUSTOMER

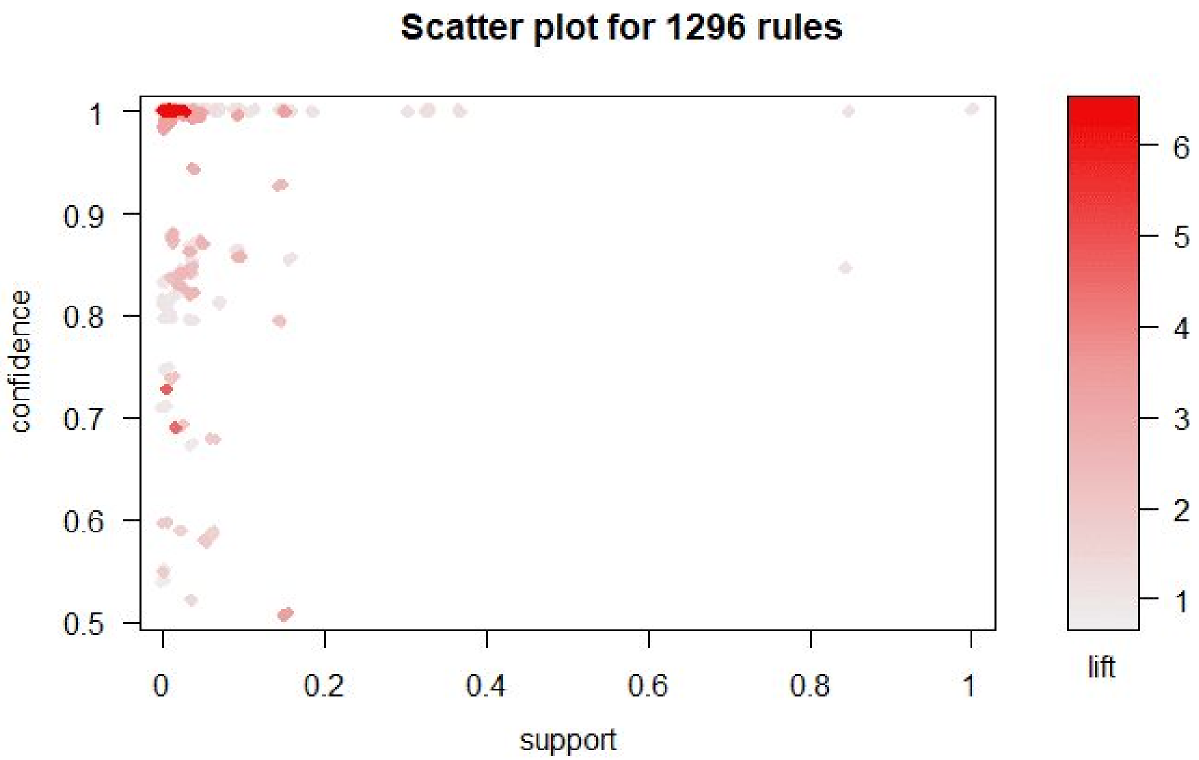
```
ggplot(data = data, aes(x=TXN_ID, y = LIFESTAGE))+
  geom_col()+
  facet_grid(~PREMIUM_CUSTOMER)
##There are more Mainstream - young singles/couples and Mainstream - retirees who buy
##chips. This contributes to there being more sales to these customer segments but
```



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

#### Average price per unit by LIFESTAGE and PREMIUM_CUSTOMER
ggplot(data = data, aes(x= mean(PROD_QTY), y = LIFESTAGE))+
  geom_count()
  facet_grid(~PREMIUM_CUSTOMER)
```

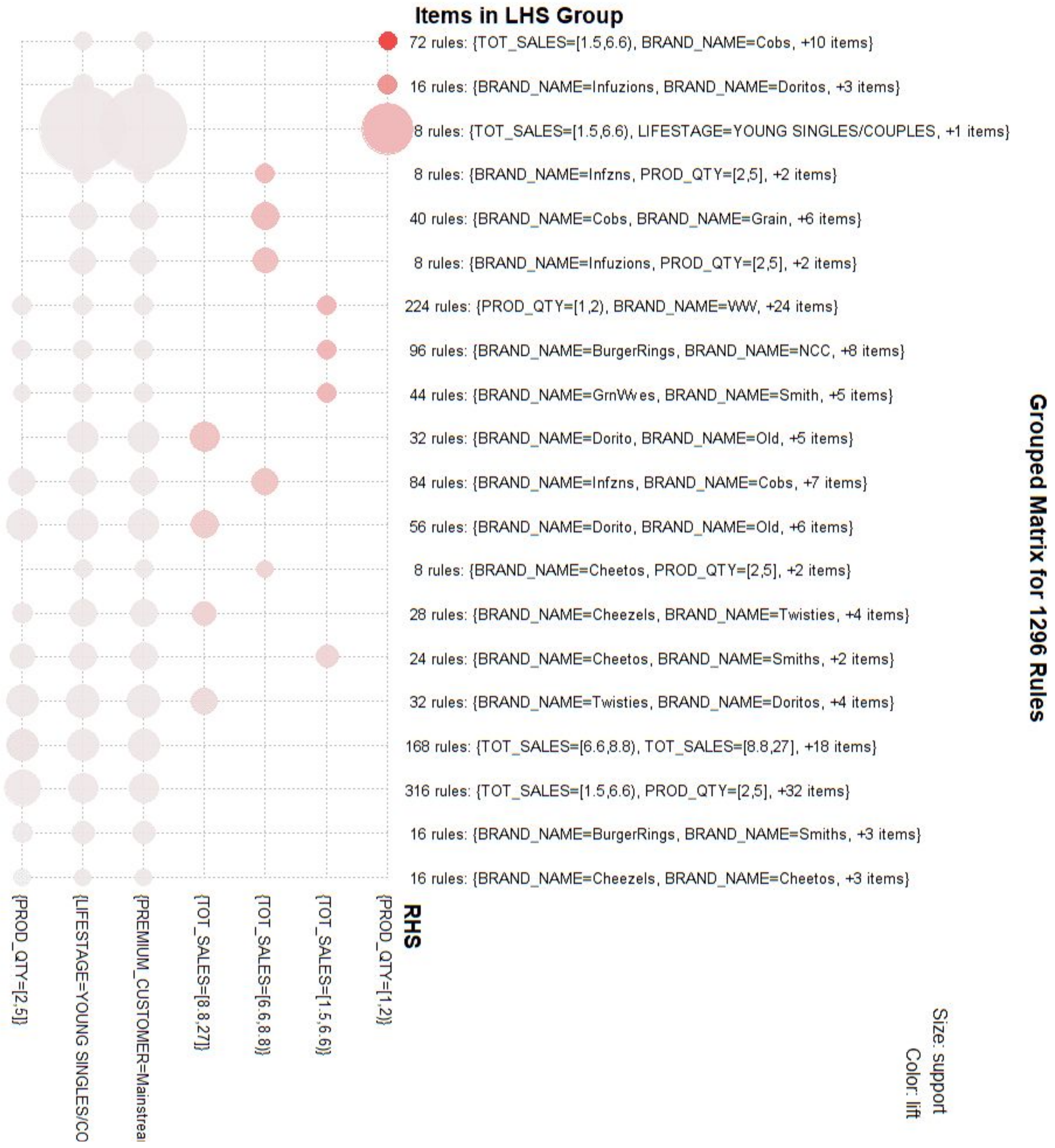


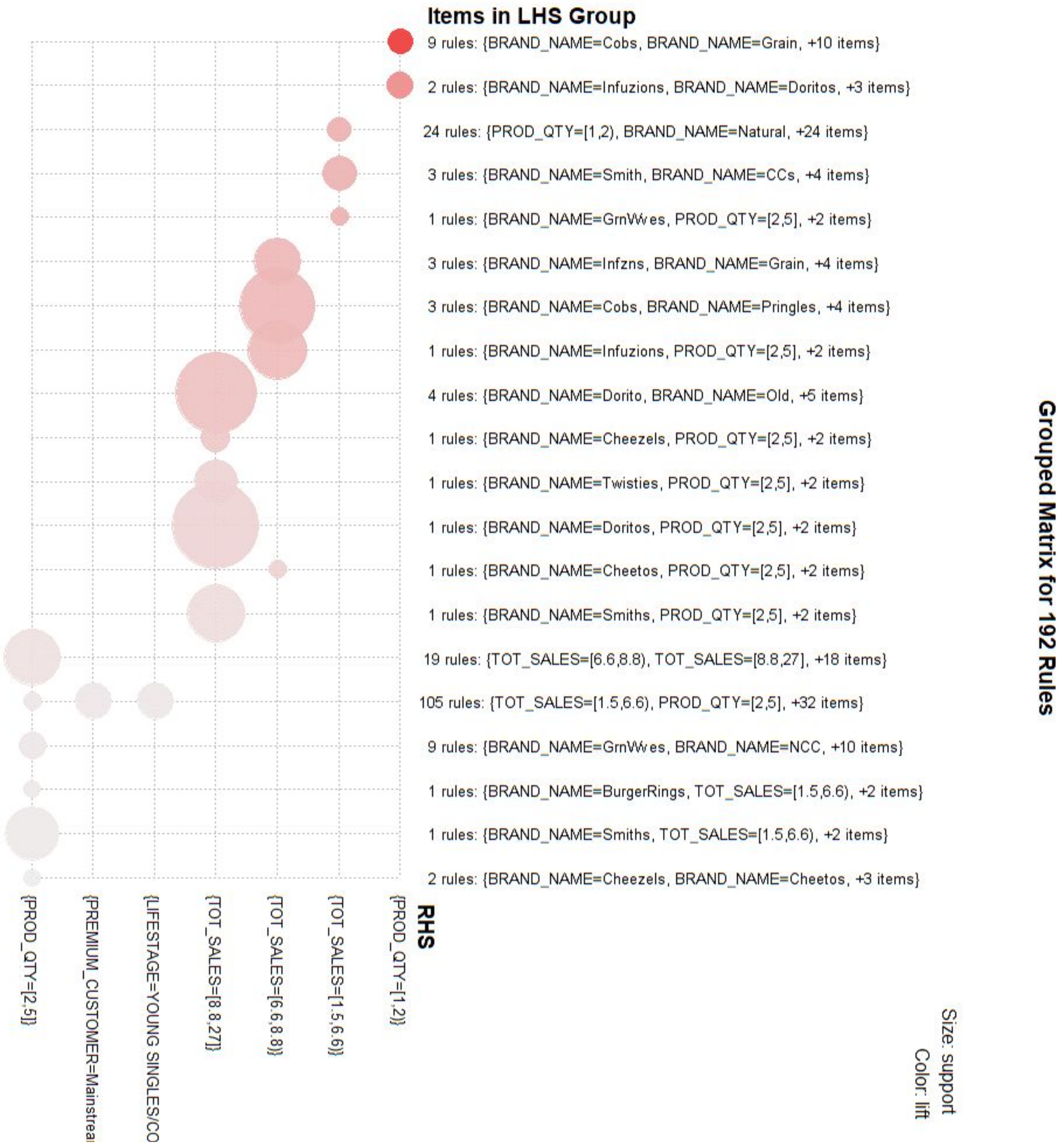
```
#### Deep dive into Mainstream, young singles/couples
g <- data.frame()
q <- data.frame()
q <- subset(g, g$PREMIUM_CUSTOMER == "Mainstream")

##Apriori Analysis
rule1 <- apriori(q, parameter = list(support=0.002, confidence = 0.5))
summary(rules)
inspect(head(rule1,20))
inspect(head(sort(rule1, by="lift"), 20))
plot(rule1)
plot(rule1, method = "grouped")
```



```
rule2 <- apriori(q, parameter = list(support=0.002, confidence = 0.5, minlen = 5))
inspect(head(rule2,5))
```

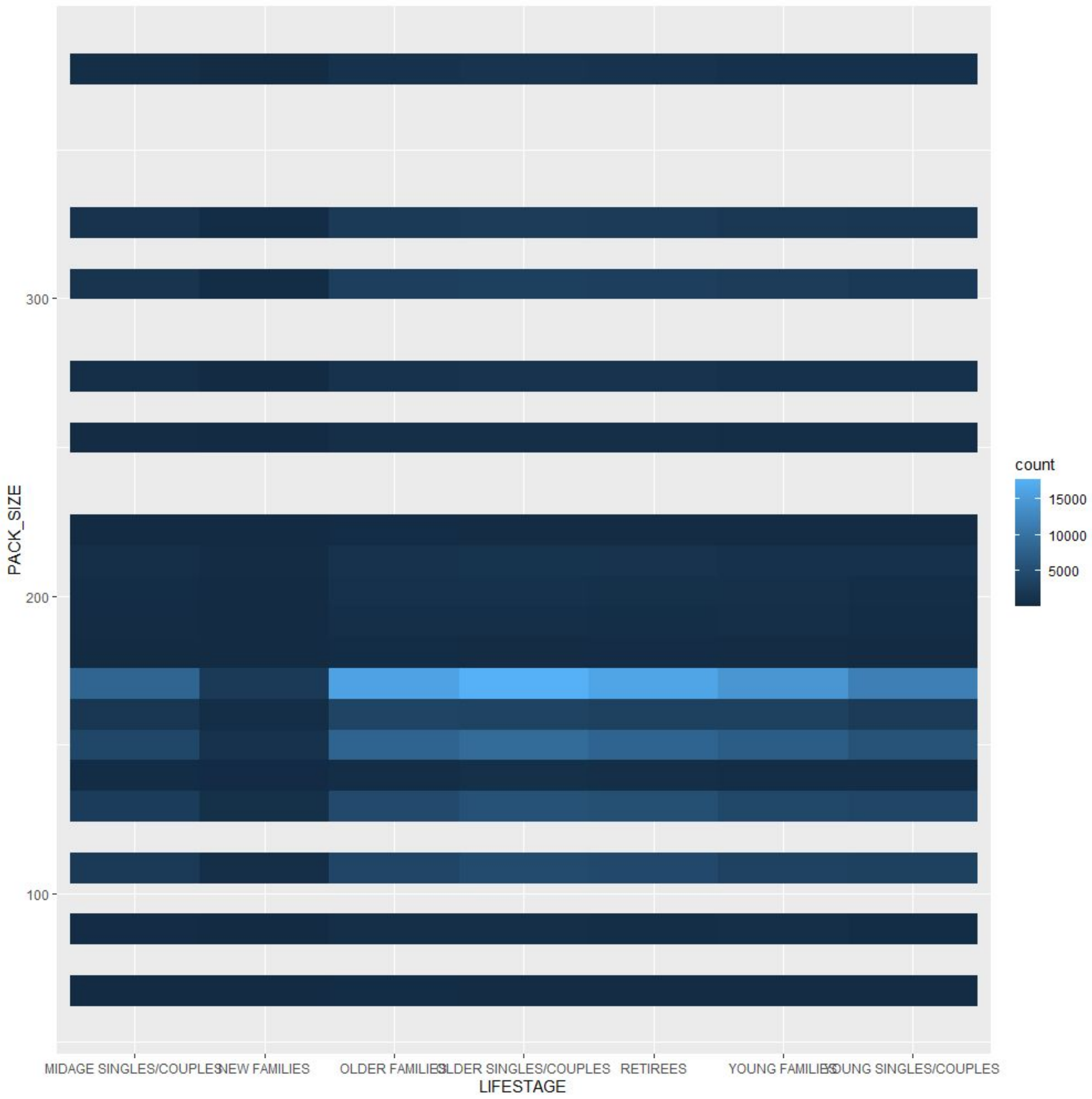
```
plot(rule2, method = "grouped")
```





We can see that Doritos have the most number of quantity purchased by customers
The red dots show the sale of packets of each brand in Main stream young/singles
##Let's also find out if our target segment tends to buy large packs of chips.

```
ggplot(data , aes(x = LIFESTAGE, y= PACK_SIZE))+
  geom_bin2d()
```



The Bind2k plot shows that the target segment tends to buy a larger pack size as compared to the rest of the customers.