Task_One

Burhanudin Badiuzaman

2024-08-20

Setting Rmarkdown

Solution template for Task 1

- This file is a solution template for the Task 1 of the Quantium Virtual Internship. It will walk you through the analysis, providing the scaffolding for your solution with gaps left for you to fill in yourself.
- Look for comments that say "over to you" for places where you need to add your own code!
- Often, there will be hints about what to do or what function to use in the text leading up to a code block - if you need a bit of extra help on how to use a function, the internet has many excellent resources on R coding, which you can find using your favourite search engine.

Load required libraries and datasets

- Note that you will need to install these libraries if you have never used these before and make sure it works in the right work directory.

```
getwd()
```

```
## [1] "/Users/burhanudin/Study_burhanudin_6/R/Code-R/Data_Analyst_in_R"
```

Point the filePath to where you have downloaded the datasets to and assign the data files to data.tables

over to you! fill in the path to your working directory. If you are on a Windows machine, you will need to use forward slashes (/) instead of backshashes ()

```
setwd("/Users/burhanudin/Study_burhanudin_6/R/Code-R/Data_Analyst_in_R")
```

```
#### Example code to install packages
#install.packages("data.table")

#### Load required libraries
library(data.table)
library(ggplot2)
library(ggmosaic)
```

```r
library(readr)
library("readxl")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
#### Point the filePath to where you have downloaded the datasets to and assign the data files to data.



transactionData <- read_excel("data/QVI_transaction_data.xlsx")
customerData <- fread("data/QVI_purchase_behaviour.csv")
```

Exploratory data analysis

The first step in any analysis is to first understand the data. Let's take a look at each of the datasets provided.

Examining transaction data

We can use `str()` to look at the format of each column and see a sample of the data. As we have read in the dataset as a `data.table` object, we can also run `transactionData` in the console to see a sample of the data or use`head(transactionData)` to look at the first 10 rows.

Let's check if columns we would expect to be numeric are in numeric form and date columns are in date format.

```r
#### Examine transaction data
# Examine the data using one or more of the methods described above.
str(transactionData)
```

```
## tibble [264,836 x 8] (S3: tbl_df/tbl/data.frame)
## $ DATE : num [1:264836] 43390 43599 43605 43329 43330 ...
## $ STORE_NBR : num [1:264836] 1 1 1 2 2 4 4 4 5 7 ...
## $ LYLTY_CARD_NBR: num [1:264836] 1000 1307 1343 2373 2426 ...
## $ TXN_ID : num [1:264836] 1 348 383 974 1038 ...
## $ PROD_NBR : num [1:264836] 5 66 61 69 108 57 16 24 42 52 ...
## $ PROD_NAME : chr [1:264836] "Natural Chip Compny SeaSalt175g" "CCs Nacho
Cheese 175g" "Smiths Crinkle Cut Chips Chicken 170g" "Smiths Chip Thinly
S/Cream&Onion 175g" ...
## $ PROD_QTY : num [1:264836] 2 3 2 5 3 1 1 1 1 2 ...
## $ TOT_SALES : num [1:264836] 6 6.3 2.9 15 13.8 5.1 5.7 3.6 3.9 7.2 ...
```

```r
###E see a sample of the data
head(transactionData,10)
```

```
## # A tibble: 10 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 Ch~ 2 6
## 2 43599 1 1307 348 66 CCs Nacho ~ 3 6.3
## 3 43605 1 1343 383 61 Smiths Cri~ 2 2.9
## 4 43329 2 2373 974 69 Smiths Chi~ 5 15
## 5 43330 2 2426 1038 108 Kettle Tor~ 3 13.8
## 6 43604 4 4074 2982 57 Old El Pas~ 1 5.1
## 7 43601 4 4149 3333 16 Smiths Cri~ 1 5.7
## 8 43601 4 4196 3539 24 Grain Wave~ 1 3.6
## 9 43332 5 5026 4525 42 Doritos Co~ 1 3.9
## 10 43330 7 7150 6900 52 Grain Wave~ 2 7.2
```

We can see that the DATE column is in an **integer** format. Let's change this to a date format.

```
#### Convert DATE column to a date format A quick search online tells us that CSV and Excel integer dat
transactionData$DATE <- as.Date(transactionData$DATE, origin = "1899-12-30")
```

We should check that we are looking at the right products by examining PROD_NAME.

```
#### Examine PROD_NAME
# Generate a summary of the PROD_NAME column.
summary(transactionData, PROD_NAME)
```

```
##       DATE              STORE_NBR     LYLTY_CARD_NBR        TXN_ID
##  Min.   :2018-07-01   Min.   :  1.0   Min.   :   1000   Min.   :      1
##  1st Qu.:2018-09-30   1st Qu.: 70.0   1st Qu.:  70021   1st Qu.:  67602
##  Median :2018-12-30   Median :130.0   Median : 130358   Median : 135138
##  Mean   :2018-12-30   Mean   :135.1   Mean   : 135550   Mean   : 135158
##  3rd Qu.:2019-03-31   3rd Qu.:203.0   3rd Qu.: 203094   3rd Qu.: 202701
##  Max.   :2019-06-30   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
```

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.

```
#### Examine the words in PROD_NAME to see if there are any incorrect entries such as products that are
productWords <- data.table(unlist(strsplit(unique(transactionData$PROD_NAME), "")))
setnames(productWords, 'words')
```

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()`.

```
# Over to you! Remove digits, and special characters, and then sort the distinct words by frequency of
library(stringi)
library(stringr)
#### Removing digits
productWords$words <- str_remove(productWords$words, "[[:digit:]]")
#### Removing special characters
productWords$words <- str_remove(productWords$words, "[[:punct:]]")
#### Let's look at the most common words by counting the number of times a word appears and
```

```
splitWords <- strsplit(productWords$words, " ")
splitWords.freq <- table(unlist(splitWords))
#### sorting them by this frequency in order of highest to lowest frequency
splitWords.freq <- as.data.frame(splitWords.freq)
splitWords.freq <- splitWords.freq[order(splitWords.freq$Freq, decreasing = T),]
```

There are salsa products in the dataset but we are only interested in the chips category, so let's remove these.

```
#### Remove salsa products
transactionData<- data.table(transactionData)
transactionData[, SALSA := grepl("salsa", tolower(PROD_NAME))]
transactionData <- transactionData[SALSA == FALSE, ][, SALSA := NULL]
```

Next, we can use `summary()` to check summary statistics such as mean, min and max values for each feature to see if there are any obvious outliers in the data and if there are any nulls in any of the columns (`NA's : number of nulls` will appear in the output if there are any nulls).

```
#### Summarise the data to check for nulls and possible outliers

# Over to you!
summary(transactionData)
```

```
##       DATE              STORE_NBR     LYLTY_CARD_NBR        TXN_ID
##  Min.   :2018-07-01   Min.   :  1.0   Min.   :   1000   Min.   :      1
##  1st Qu.:2018-09-30   1st Qu.: 70.0   1st Qu.:  70015   1st Qu.:  67569
##  Median :2018-12-30   Median :130.0   Median : 130367   Median : 135183
##  Mean   :2018-12-30   Mean   :135.1   Mean   : 135531   Mean   : 135131
##  3rd Qu.:2019-03-31   3rd Qu.:203.0   3rd Qu.: 203084   3rd Qu.: 202654
##  Max.   :2019-06-30   Max.   :272.0   Max.   :2373711   Max.   :2415841
##     PROD_NBR       PROD_NAME            PROD_QTY         TOT_SALES
##  Min.   :  1.00   Length:246742      Min.   :  1.000   Min.   :  1.700
##  1st Qu.: 26.00   Class :character   1st Qu.:  2.000   1st Qu.:  5.800
##  Median : 53.00   Mode  :character   Median :  2.000   Median :  7.400
##  Mean   : 56.35                      Mean   :  1.908   Mean   :  7.321
##  3rd Qu.: 87.00                      3rd Qu.:  2.000   3rd Qu.:  8.800
##  Max.   :114.00                      Max.   :200.000   Max.   :650.000
```

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.

```
#### Filter the dataset to find the outlier
# Over to you! Use a filter to examine the transactions in question.
prodQyt_200 <- transactionData[PROD_QTY == 200,]
print(prodQyt_200)
```

```
##          DATE STORE_NBR LYLTY_CARD_NBR TXN_ID PROD_NBR
##        <Date>     <num>          <num>  <num>    <num>
## 1: 2018-08-19       226         226000 226201        4
## 2: 2019-05-20       226         226000 226210        4
##                          PROD_NAME PROD_QTY TOT_SALES
##                             <char>    <num>     <num>
## 1: Dorito Corn Chp     Supreme 380g      200       650
## 2: Dorito Corn Chp     Supreme 380g      200       650
```

There are two transactions where 200 packets of chips are bought in one transaction and both of these transactions were by the same customer.

```
#### Let's see if the customer has had other transactions

# Over to you! Use a filter to see what other transactions that customer made.
sameCustomer <- transactionData[LYLTY_CARD_NBR == 226000,]
print(sameCustomer)
```

```
##          DATE STORE_NBR LYLTY_CARD_NBR TXN_ID PROD_NBR
##        <Date>     <num>          <num>  <num>    <num>
## 1: 2018-08-19       226         226000 226201        4
## 2: 2019-05-20       226         226000 226210        4
##                          PROD_NAME PROD_QTY TOT_SALES
##                             <char>    <num>     <num>
## 1: Dorito Corn Chp     Supreme 380g      200       650
## 2: Dorito Corn Chp     Supreme 380g      200       650
```

It looks like this customer has only had the two transactions over the year and is not an ordinary retail customer. The customer might be buying chips for commercial purposes instead. We'll remove this loyalty card number from further analysis.

```
#### Filter out the customer based on the loyalty card number
# Over to you!
transactionData <- transactionData[LYLTY_CARD_NBR != 226000]
#### Re-examine transaction data
# Over to you!
summary(transactionData)
```

```
##       DATE              STORE_NBR     LYLTY_CARD_NBR        TXN_ID
##  Min.   :2018-07-01   Min.   :  1.0   Min.   :   1000   Min.   :      1
##  1st Qu.:2018-09-30   1st Qu.: 70.0   1st Qu.:  70015   1st Qu.:  67569
##  Median :2018-12-30   Median :130.0   Median : 130367   Median : 135182
##  Mean   :2018-12-30   Mean   :135.1   Mean   : 135530   Mean   : 135130
##  3rd Qu.:2019-03-31   3rd Qu.:203.0   3rd Qu.: 203083   3rd Qu.: 202652
##  Max.   :2019-06-30   Max.   :272.0   Max.   :2373711   Max.   :2415841
##     PROD_NBR       PROD_NAME            PROD_QTY       TOT_SALES
##  Min.   :  1.00   Length:246740      Min.   :1.000   Min.   : 1.700
##  1st Qu.: 26.00   Class :character   1st Qu.:2.000   1st Qu.: 5.800
##  Median : 53.00   Mode  :character   Median :2.000   Median : 7.400
##  Mean   : 56.35                      Mean   :1.906   Mean   : 7.316
##  3rd Qu.: 87.00                      3rd Qu.:2.000   3rd Qu.: 8.800
##  Max.   :114.00                      Max.   :5.000   Max.   :29.500
```

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.

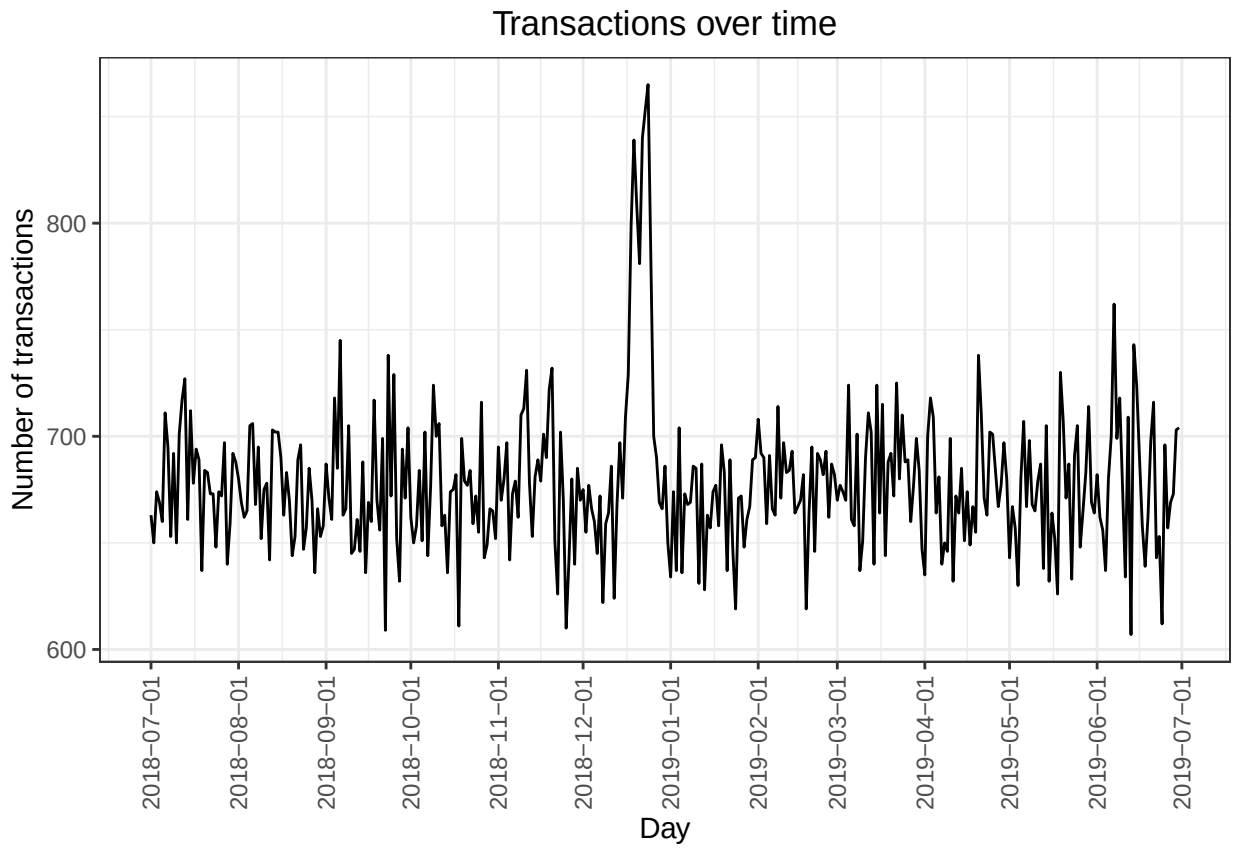
```
#### Count the number of transactions by date
# Over to you! Create a summary of transaction count by date.
transactionDataCount <- count(transactionData, transactionData$DATE)
nrow(transactionDataCount)
```

```
## [1] 364
```

There's only 364 rows, meaning only 364 dates which indicates a missing date. Let's create a sequence of dates from 1 Jul 2018 to 30 Jun 2019 and use this to create a chart of number of transactions over time to find the missing date.

```
#### Create a sequence of dates and join this the count of transactions by date
# Over to you - create a column of dates that includes every day from 1 Jul 2018 to 30 Jun 2019, and jo
```

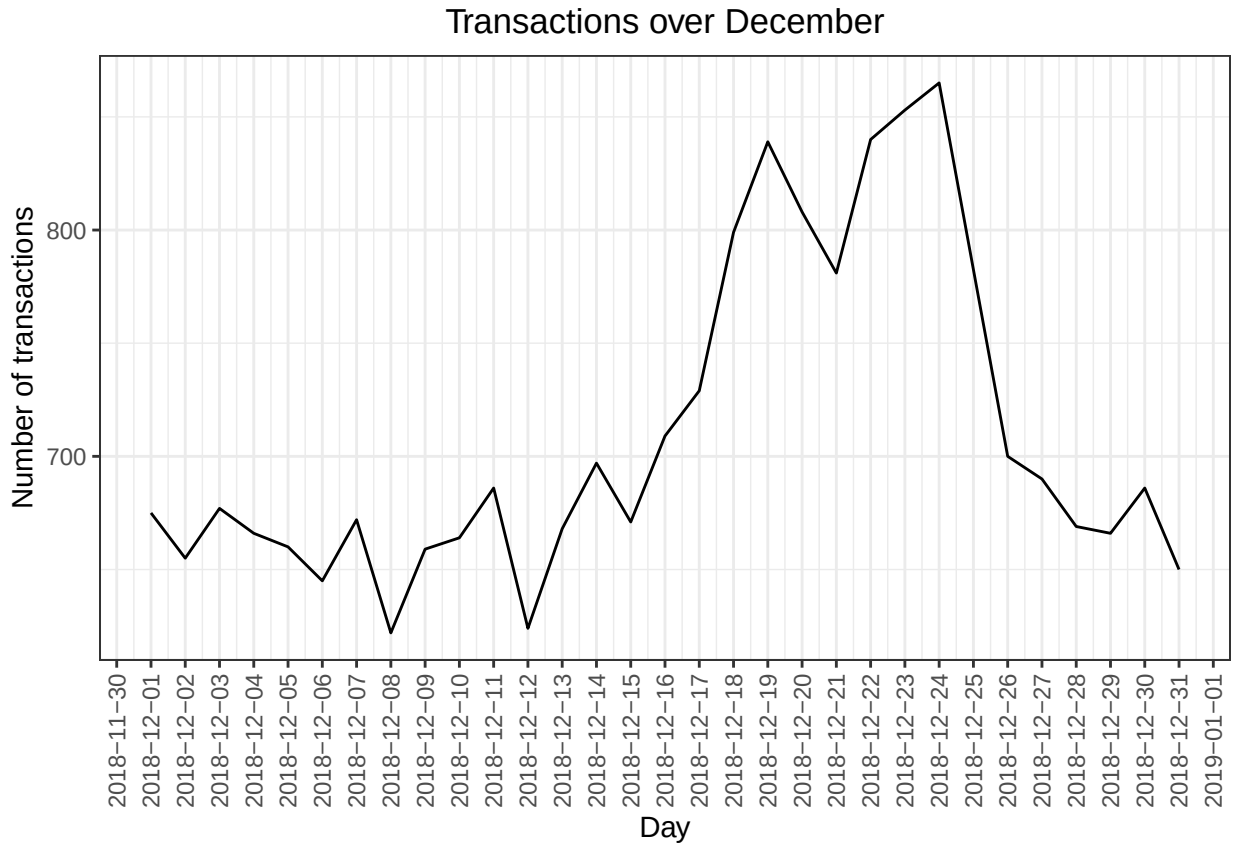
```
transactions_by_day <- transactionDataCount %>%
  filter(
    transactionDataCount$`transactionData$DATE` >= "2018-07-01" &
      transactionDataCount$`transactionData$DATE` <= "2019-06-30"
  )

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

#### Plot transactions over time
N <- transactions_by_day$n
DATE <- transactions_by_day$`transactionData$DATE`
ggplot(transactions_by_day, aes(x = DATE, y = N)) + geom_line() + labs(x = "Day", y = "Number of transa
```



We can see that there is an increase in purchases in December and a break in late December. Let's zoom in on this.

```
#### Filter to December and look at individual days
# Over to you - recreate the chart above zoomed in to the relevant dates.
dec_date <- transactions_by_day %>%
  filter(
    transactions_by_day$`transactionData$DATE` >= "2018-12-01" &
      transactions_by_day$`transactionData$DATE` <= "2018-12-31"
  )
N_DEC <- dec_date$n
DATE_DEC <-dec_date$`transactionData$DATE`
```

```
ggplot(dec_date, aes(x = DATE_DEC, y = N_DEC))+
  geom_line()+
  labs(x = "Day", y = "Number of transactions", title = "Transactions over December")+
  scale_x_date(breaks = "1 day") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
```



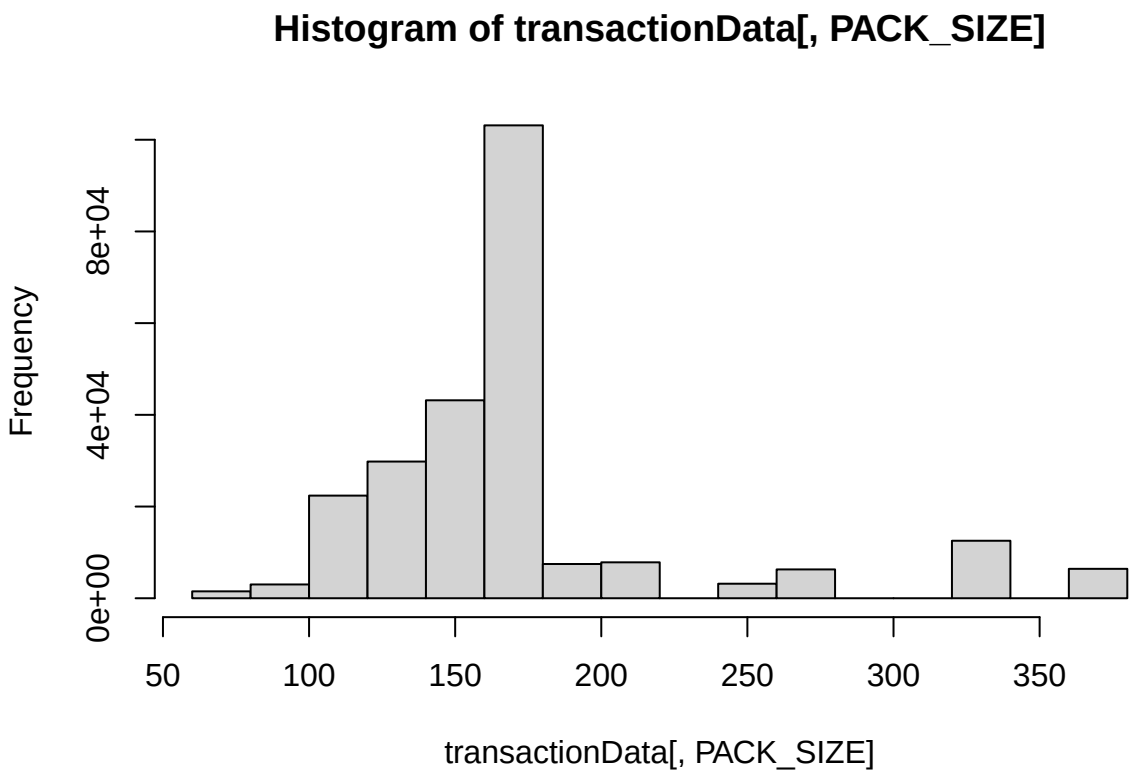
We can see that the increase in sales occurs in the lead-up to Christmas and that there are zero sales on Christmas day itself. This is due to shops being closed on Christmas day.

Now that we are satisfied that the data no longer has outliers, we can move on to creating other features such as brand of chips or pack size from PROD_NAME. We will start with pack size.

```
#### Pack size We can work this out by taking the digits that are in PROD_NAME

transactionData[, PACK_SIZE := parse_number(PROD_NAME)]
#### Always check your output Let's check if the pack sizes look sensible transactionData[, .N, PACK_SI
```

The largest size is 380g and the smallest size is 70g - seems sensible!

```
#### Let's plot a histogram of PACK_SIZE since we know that it is a categorical variable and not a cont
# Over to you! Plot a histogram showing the number of transactions by pack size.
hist(transactionData[, PACK_SIZE])
```



Pack sizes created look reasonable. Now to create brands, we can use the first word in PROD_NAME to work out the brand name...

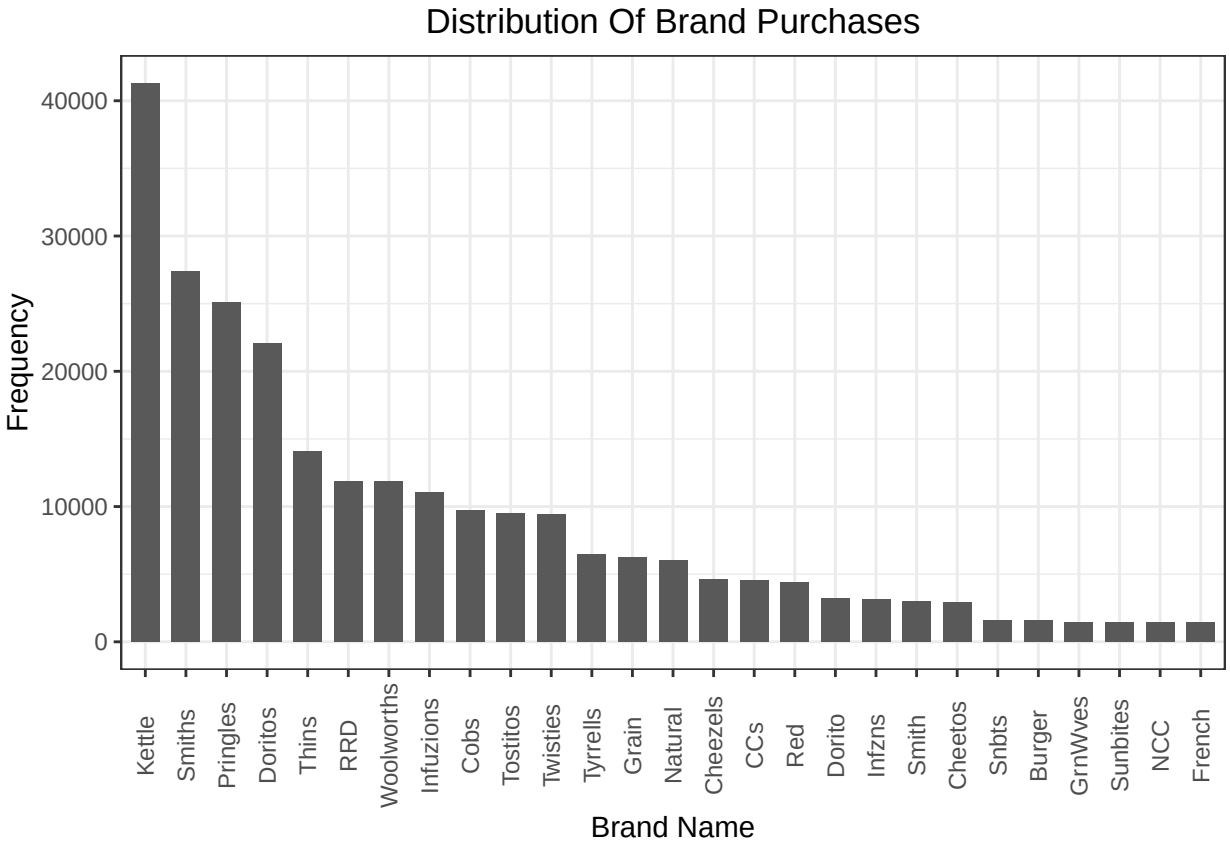
```
#### Brands
# Over to you! Create a column which contains the brand of the product, by extracting it from the produ
transactionData$BRAND_NAME <- gsub("([A-Za-z]+).*", "\\1", transactionData$PROD_NAME)
#### Checking brands
# Over to you! Check the results look reasonable.
brands <- data.table(unlist(unique(transactionData$BRAND_NAME)))
print(brands)
```

```
##            V1
##         <char>
##  1:    Natural
##  2:        CCs
##  3:     Smiths
##  4:     Kettle
##  5:      Grain
##  6:    Doritos
##  7:   Twisties
##  8:         WW
##  9:      Thins
## 10:     Burger
## 11:        NCC
## 12:   Cheezels
## 13:     Infzns
## 14:        Red
## 15:   Pringles
## 16:     Dorito
## 17:  Infuzions
```

```
## 18:      Smith
## 19:    GrnWves
## 20:   Tyrrells
## 21:       Cobs
## 22:     French
## 23:        RRD
## 24:   Tostitos
## 25:    Cheetos
## 26: Woolworths
## 27:      Snbts
## 28:   Sunbites
##             V1
```

Some of the brand names look like they are of the same brands - such as RED and RRD, which are both Red Rock Deli chips. Let's combine these together.

```
#### Clean brand names transactionData[BRAND == "RED", BRAND := "RRD"]
# Over to you! Add any additional brand adjustments you think may be required.
transactionData[BRAND_NAME == "RED", `:=`(BRAND_NAME, "RRD")]
transactionData[BRAND_NAME == "GRAIN", `:=`(BRAND_NAME, "GrnWves")]
transactionData[BRAND_NAME == "INFZNS", `:=`(BRAND_NAME, "Infuzions")]
transactionData[BRAND_NAME == "WW", `:=`(BRAND_NAME, "Woolworths")]
transactionData[BRAND_NAME == "SNBTS", `:=`(BRAND_NAME, "Sunbites")]
#### Check again
# Over to you! Check the results look reasonable.
brands <- data.frame(sort(table(transactionData$BRAND_NAME), decreasing = T))
setnames(brands, c("Brand", "freq"))
ggplot(brands, aes(x = Brand, y = freq))+
  geom_bar(stat = "identity", width = 0.7)+
  labs(x = "Brand Name", y = "Frequency", title = "Distribution Of Brand Purchases")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
```



Examining customer data

Now that we are happy with the transaction dataset, let's have a look at the customer dataset.

```
#### Examining customer data
# Over to you! Do some basic summaries of the dataset, including distributions of any key columns.
str(customerData)
```

```
## Classes 'data.table' and 'data.frame': 72637 obs. of 3 variables:
## $ LYLTY_CARD_NBR : int 1000 1002 1003 1004 1005 1007 1009 1010 1011 1012 ...
## $ LIFESTAGE : chr "YOUNG SINGLES/COUPLES" "YOUNG SINGLES/COUPLES" "YOUNG
FAMILIES" "OLDER SINGLES/COUPLES" ...
## $ PREMIUM_CUSTOMER: chr "Premium" "Mainstream" "Budget" "Mainstream" ...
## - attr(*, ".internal.selfref")=<externalptr>
```

Let's have a closer look at the LIFESTAGE and PREMIUM_CUSTOMER columns.

```
### Examining the values of lifestage and premium_customer
lifestageCat <- data.frame(sort(table(customerData$LIFESTAGE),decreasing = TRUE ))
setnames(lifestageCat,c("lifestage","N"))
lifestageCat
```

```
##                lifestage     N
## 1              RETIREES 14805
## 2  OLDER SINGLES/COUPLES 14609
## 3  YOUNG SINGLES/COUPLES 14441
## 4         OLDER FAMILIES  9780
## 5         YOUNG FAMILIES  9178
## 6 MIDAGE SINGLES/COUPLES  7275
```

```
## 7           NEW FAMILIES  2549
```

```
### PERMIUM_CUSTOMER
permiumCust <- data.frame(sort(table(customerData$PREMIUM_CUSTOMER),decreasing = TRUE ))
setnames(permiumCust,c("lifestage","N"))
permiumCust
```

```
##    lifestage     N
## 1 Mainstream 29245
## 2     Budget 24470
## 3    Premium 18922
```

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

As the number of rows in `data` is the same as that of `transactionData`, we can bevsure that no duplicates were created. This is because we created `data` by setting `all.x = TRUE` (in other words, a left join) which means take all the rows in `transactionData` and find rows with matching values in shared columns and then joining the details in these rows to the `x` or the first mentioned table.

Let's also check if some customers were not matched on by checking for nulls.

```
# Over to you! See if any transactions did not have a matched customer.
sum(is.na(data))
```

```
## [1] 0
```

Great, there are no nulls! So all our customers in the transaction data has been accounted for in the customer dataset. Note that if you are continuing with Task 2, you may want to retain this dataset which you can write out as a csv

```
write.csv(data, file=file.path('output/','QVI_data.csv'))
```

Data exploration is now complete!

Data analysis on customer segments

Now that the data is ready for analysis, we can define some metrics of interest to the client: - Who spends the most on chips (total sales), describing customers by lifestage and how premium their general purchasing behaviour is - How many customers are in each segment - How many chips are bought per customer by segment - What's the average chip price by customer segment

We could also ask our data team for more information. Examples are: - The customer's total spend over the period and total spend for each transaction to understand what proportion of their grocery spend is on chips - Proportion of customers in each customer segment overall to compare against the mix of customers who purchase chips Let's start with 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.

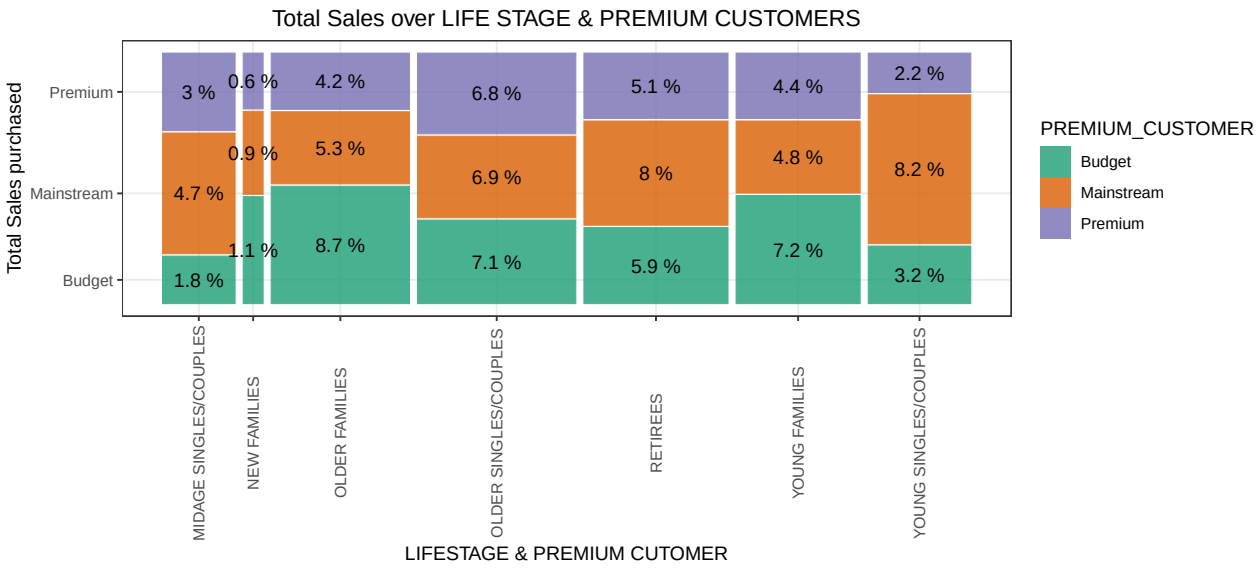
```
#### Total sales by LIFESTAGE and PREMIUM_CUSTOMER
# Over to you! Calculate the summary of sales by those dimensions and create a plot.
sales <- data[, .(SALES = sum(TOT_SALES)), .(LIFESTAGE, PREMIUM_CUSTOMER)]

p <- ggplot(sales)+
  geom_mosaic(aes(weight = SALES, x = product(PREMIUM_CUSTOMER, LIFESTAGE), fill = PREMIUM_CUSTOMER))+
  labs( x= "LIFESTAGE & PREMIUM CUTOMER", y = "Total Sales purchased", title = "Total Sales over LIFE S
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))+
  scale_fill_brewer(palette = "Dark2")

p + geom_text(data = ggplot_build(p)$data[[1]], aes(x = (xmin + xmax)/2 , y =
```

```
(ymin + ymax)/2, label = as.character(paste(round(.wt/sum(.wt),3)*100,
'%'))))
```

```
## Warning: The `scale_name` argument of `continuous_scale()` is deprecated as of ggplot2
## 3.5.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The `trans` argument of `continuous_scale()` is deprecated as of ggplot2 3.5.0.
## i Please use the `transform` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

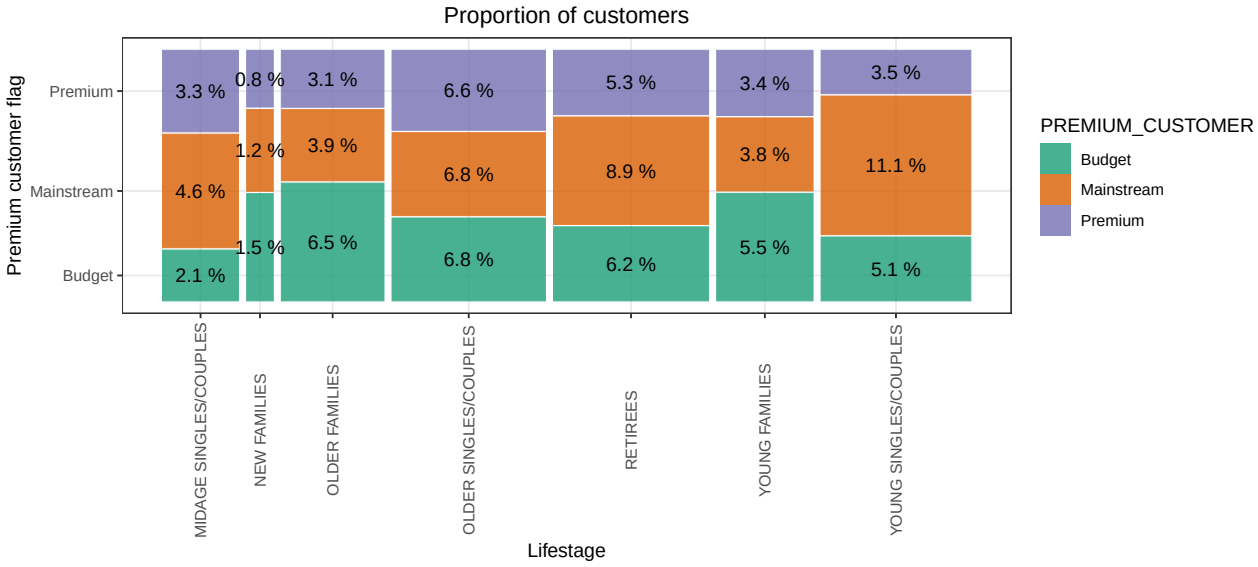
```
## Warning: `unite_()` was deprecated in tidyr 1.2.0.
## i Please use `unite()` instead.
## i The deprecated feature was likely used in the ggmosaic package.
##   Please report the issue at <https://github.com/haleyjeppson/ggmosaic>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Sales are coming mainly from Budget - older families, Mainstream - young singles/couples, and Mainstream - retirees Let's see if the higher sales are due to there being more customers who buy chips.

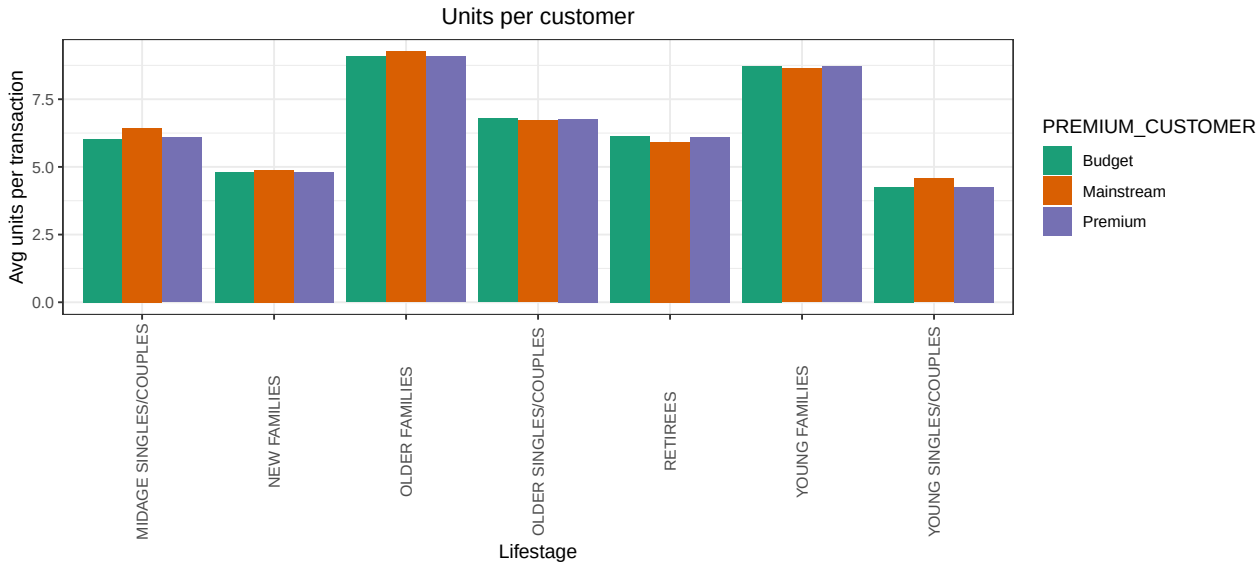
```
#### Number of customers by LIFESTAGE and PREMIUM_CUSTOMER
# Over to you! Calculate the summary of number of customers by those dimensions and create a plot.
customers <- data[, .(CUSTOMERS = uniqueN(LYLTY_CARD_NBR)), .(LIFESTAGE, PREMIUM_CUSTOMER)][order(-CUST

p <- ggplot(customers)+
  geom_mosaic(aes(weight = CUSTOMERS, x = product(PREMIUM_CUSTOMER, LIFESTAGE), fill = PREMIUM_CUSTOMER
  labs( x= "Lifestage", y = "Premium customer flag", title = "Proportion of customers")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))+
  scale_fill_brewer(palette = "Dark2")

p + geom_text(data = ggplot_build(p)$data[[1]], aes(x = (xmin + xmax)/2 , y =
```

```
(ymin + ymax)/2, label = as.character(paste(round(.wt/sum(.wt),3)*100,
'%'))))
```



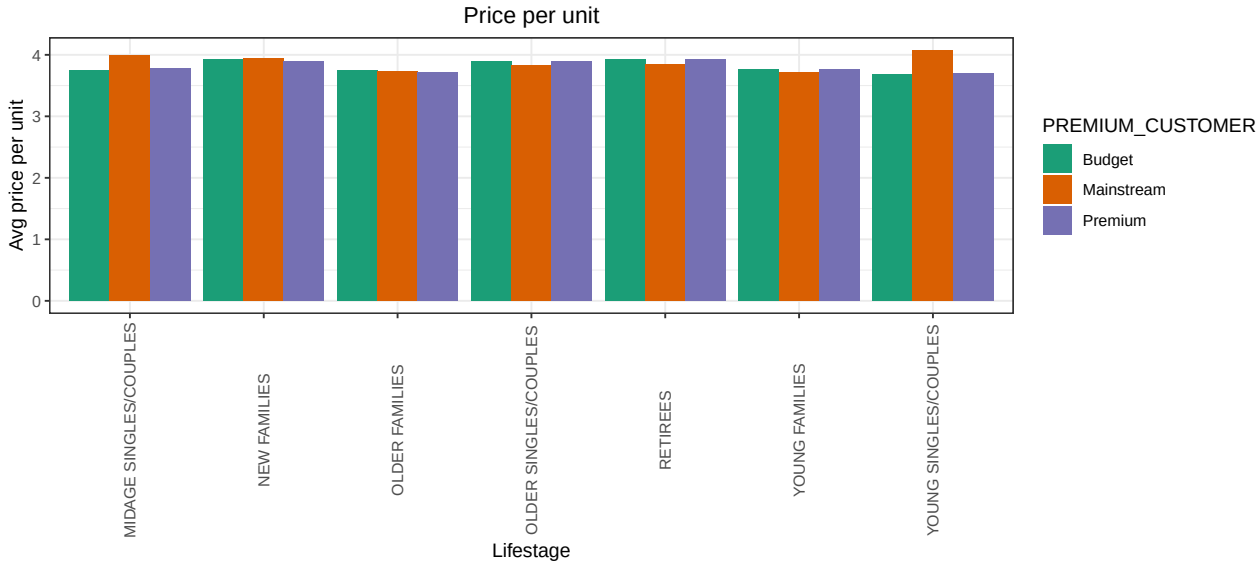
There are more Mainstream - young singles/couples and Mainstream - retirees who buy chips. This contributes to there being more sales to these customer segments but this is not a major driver for the Budget - Older families segment. Higher sales may also be driven by more units of chips being bought per customer. Let's have a look at this next.

```
#### Average number of units per customer by LIFESTAGE and PREMIUM_CUSTOMER
# Over to you! Calculate and plot the average number of units per customer by those two dimensions.
avgNUnit <- data[, .(AVG = sum(PROD_QTY)/uniqueN(LYLTY_CARD_NBR)), .(LIFESTAGE, PREMIUM_CUSTOMER)][orde

ggplot(avgNUnit, aes(weight = AVG, x = LIFESTAGE, fill = PREMIUM_CUSTOMER))+
  geom_bar(position = position_dodge())+
  labs( x= "Lifestage", y = "Avg units per transaction", title = "Units per customer")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))+
  scale_fill_brewer(palette = "Dark2")
```



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.

```
#### Average price per unit by LIFESTAGE and PREMIUM_CUSTOMER
# Over to you! Calculate and plot the average price per unit sold (average sale price) by those two cus
pricePerUnit <- data[, .(PRICE = sum(TOT_SALES)/sum(PROD_QTY)), .(LIFESTAGE, PREMIUM_CUSTOMER)][order(-

ggplot(pricePerUnit, aes(weight = PRICE, x = LIFESTAGE, fill = PREMIUM_CUSTOMER))+
  geom_bar(position = position_dodge())+
  labs( x= "Lifestage", y = "Avg price per unit", title = "Price per unit")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))+
  scale_fill_brewer(palette = "Dark2")
```



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. As the difference in average price per unit isn't

large, we can check if this difference is statistically different.

```
#### Perform an independent t-test between mainstream vs premium and budget midage and young singles an
# Over to you! Perform a t-test to see if the difference is significant.
PriceUnit <- data$TOT_SALES / data$PROD_QTY
t.test(
  c(
    data$LIFESTAGE %in% c("YOUNG SINGLES/COUPLES","MIDAGE SINGLES/COUPLES") & data$PREMIUM_CUSTOMER ==
  ),
    c(
    data$LIFESTAGE %in% c("YOUNG SINGLES/COUPLES","MIDAGE SINGLES/COUPLES") & data$PREMIUM_CUSTOMER !=
  )
)
```

```
##
## Welch Two Sample t-test
##
## data: c(data$LIFESTAGE %in% c("YOUNG SINGLES/COUPLES", "MIDAGE
SINGLES/COUPLES") & data$PREMIUM_CUSTOMER == "Mainstream", PriceUnit) and
c(data$LIFESTAGE %in% c("YOUNG SINGLES/COUPLES", "MIDAGE SINGLES/COUPLES") &
data$PREMIUM_CUSTOMER != "Mainstream", PriceUnit)
## t = 1.9479, df = 986950, p-value = 0.05143
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -4.921249e-05 1.589991e-02
## sample estimates:
## mean of x mean of y
## 1.978295 1.970370
```

The t-test results in a p-value of XXXXXXX, i.e. the unit price for mainstream, young and mid-age singles and couples [ARE / ARE NOT] significantly higher than that of budget or premium, young and midage singles and couples.

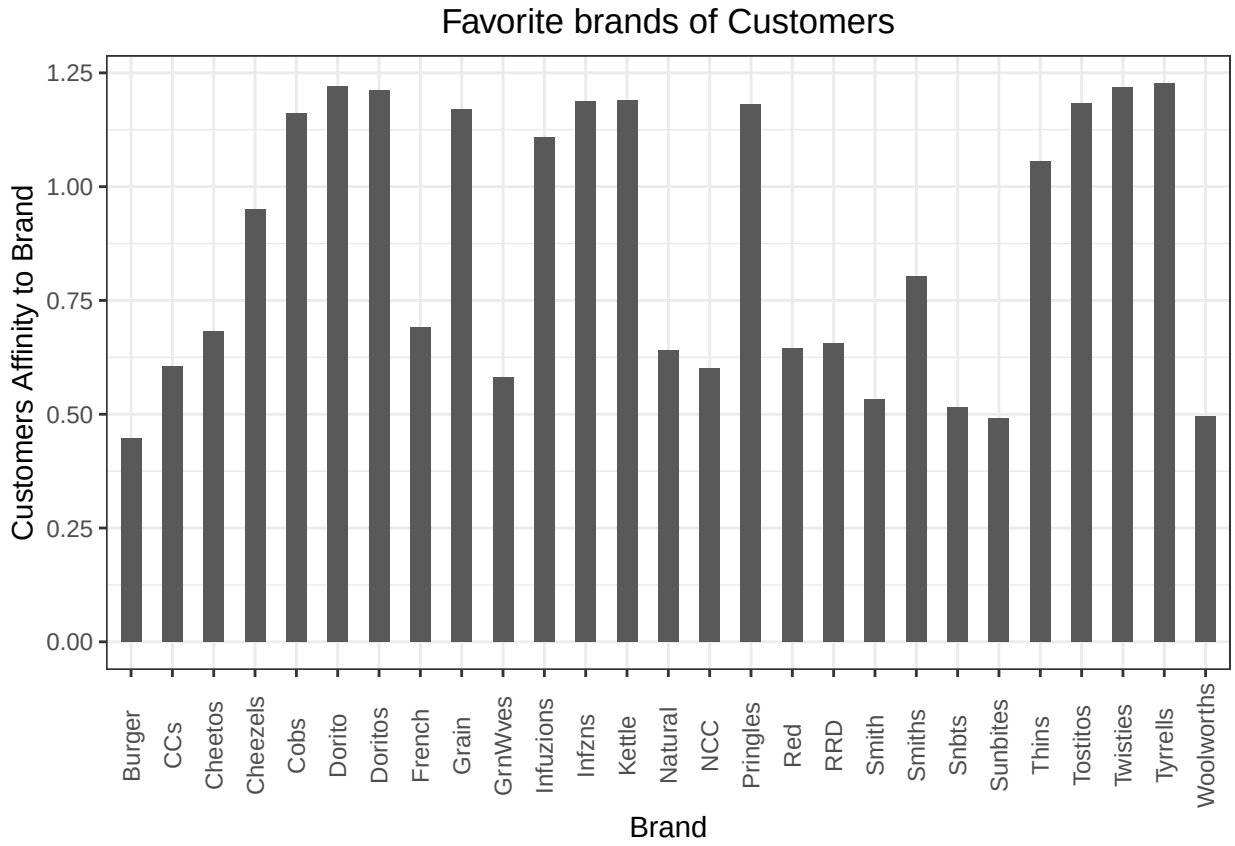
Deep dive into specific customer segments for insights

We have found quite a few interesting insights that we can dive deeper into. We might want to target customer segments that contribute the most to sales to retain them or further increase sales. Let's look at Mainstream - young singles/couples. For instance, let's find out if they tend to buy a particular brand of chips.

```
#### 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. Y
segmentOne <- data[
  LIFESTAGE == "YOUNG SINGLES/COUPLES" & PREMIUM_CUSTOMER == "Mainstream",
]
otherSegment <- data[
  LIFESTAGE != "YOUNG SINGLES/COUPLES" & PREMIUM_CUSTOMER != "Mainstream"
]
qSegmentOne <- sum(segmentOne$PROD_QTY)
qOtherSement <- sum(otherSegment$PROD_QTY)

qSegmentOneByBrand <- segmentOne[, .(tragetSegment = sum(PROD_QTY)/ qSegmentOne), by = BRAND_NAME]
qOtherSementByBrand <- otherSegment[, .(noTragetSement = sum(PROD_QTY)/ qOtherSement), by = BRAND_NAME]

brandPortion <- merge(qSegmentOneByBrand, qOtherSementByBrand)[, affinityByBrand :=tragetSegment /noTrag

brandPortion[order(-affinityByBrand)]
```

```
##     BRAND_NAME tragetSegment noTragetSement affinityByBrand
##          <char>         <num>          <num>           <num>
##  1:   Tyrrells   0.031552795    0.025714871       1.2270252
##  2:     Dorito   0.015707384    0.012859333       1.2214774
##  3:   Twisties   0.046183575    0.037932945       1.2175057
##  4:    Doritos   0.107053140    0.088311629       1.2122202
##  5:     Kettle   0.197984817    0.166560423       1.1886666
##  6:     Infzns   0.014934438    0.012570956       1.1880113
##  7:   Tostitos   0.045410628    0.038350332       1.1841000
##  8:   Pringles   0.119420290    0.101110251       1.1810898
##  9:      Grain   0.029123533    0.024891479       1.1700202
## 10:       Cobs   0.044637681    0.038448988       1.1609586
## 11:  Infuzions   0.049744651    0.044838812       1.1094105
## 12:      Thins   0.060372671    0.057159336       1.0562172
## 13:   Cheezels   0.017971014    0.018903864       0.9506530
## 14:     Smiths   0.089772257    0.111852290       0.8025965
## 15:     French   0.003947550    0.005706827       0.6917242
## 16:    Cheetos   0.008033126    0.011758947       0.6831501
## 17:        RRD   0.032022084    0.048895061       0.6549145
## 18:        Red   0.011787440    0.018289166       0.6445039
## 19:    Natural   0.015955832    0.024899068       0.6408204
## 20:        CCs   0.011180124    0.018444738       0.6061417
## 21:        NCC   0.003643892    0.006059709       0.6013312
## 22:    GrnWves   0.003588682    0.006177337       0.5809432
## 23:      Smith   0.006597654    0.012366057       0.5335293
## 24:      Snbts   0.003478261    0.006754090       0.5149858
## 25: Woolworths   0.024099379    0.048747078       0.4943759
## 26:   Sunbites   0.002870945    0.005858604       0.4900392
## 27:     Burger   0.002926156    0.006537808       0.4475745
##     BRAND_NAME tragetSegment noTragetSement affinityByBrand
```

```
ggplot(brandPortion, aes(x = BRAND_NAME, y = affinityByBrand))+
  geom_bar(stat = "identity", width = 0.5)+
  labs(x = "Brand", y = "Customers Affinity to Brand", title = "Favorite brands of Customers") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
```



We can see that : INSIGHTS

- Let's also find out if our target segment tends to buy larger packs of chips.

```
#### Preferred pack size compared to the rest of the population
# Over to you! Do the same for pack size.
qSegmentOneByPack <- segmentOne[, .(tragetSegment = sum(PROD_QTY)/ qSegmentOne), by = PACK_SIZE]
qOtherSementByPack <- otherSegment[, .(noTragetSement = sum(PROD_QTY)/ qOtherSement), by = PACK_SIZE]

brandPortionPack <- merge(qSegmentOneByPack, qOtherSementByPack)[, affinityByPack :=tragetSegment /noTra

brandPortionPack[order(-affinityByPack)]
```

```
##     PACK_SIZE tragetSegment noTragetSement affinityByPack
##         <num>         <num>          <num>          <num>
##  1:       270   0.031828847    0.025069818      1.2696083
##  2:       380   0.032160110    0.025711077      1.2508271
##  3:       330   0.061283644    0.050974410      1.2022433
##  4:       110   0.106280193    0.089575175      1.1864916
##  5:       134   0.119420290    0.101110251      1.1810898
##  6:       210   0.029123533    0.024891479      1.1700202
##  7:       135   0.014768806    0.012931427      1.1420863
##  8:       250   0.014354727    0.012863127      1.1159594
##  9:       170   0.080772947    0.080347115      1.0052999
## 10:       150   0.157598344    0.163069544      0.9664487
## 11:       175   0.254989648    0.271457518      0.9393354
## 12:       165   0.055652174    0.061587439      0.9036286
## 13:       190   0.007481021    0.012130802      0.6166964
```

```
## 14:       180   0.003588682    0.006177337      0.5809432
## 15:       160   0.006404417    0.012221868      0.5240129
## 16:       125   0.003008972    0.005976232      0.5034898
## 17:        90   0.006349206    0.012612695      0.5033981
## 18:       200   0.008971705    0.018471299      0.4857105
## 19:        70   0.003036577    0.006283581      0.4832558
## 20:       220   0.002926156    0.006537808      0.4475745
##     PACK_SIZE tragetSegment noTragetSement affinityByPack
```

INSIGHTS

- It looks like Mainstream young singles/couples are 27% more likely to purchase a 270g pack of chips compared to the rest of the population but let's dive into what brands sell this pack size.

```
data[PACK_SIZE == 270, unique(PROD_NAME)]
```

```
## [1] "Twisties Cheese     270g" "Twisties Chicken270g"
```

Conclusion

Let's recap what we've found!

- Sales have principally been because of Budget - older families,thought young singles/couples, and thought - retirees shoppers.
- we have a tendency to found that the high pay on chips for mainstream young singles/couples and retirees is due to there being a lot of of them than alternative buyers. thought, mid-age, and young singles and couples are more seemingly to pay more per packet of chips. this is often indicative of impulse shopping for behavior.
- We've also found that Mainstream young singles and couples are 23% more likely to get Tyrrells chips compared to the remainder of the population.
- The class Manager might want to extend the category'sperformance by off-locating some Tyrrells and smaller packs of chips in discretionary area close to segments wherever young singles and couples frequent a lot of typically to increase visibility and impulse behavior.
- Quantium will facilitate the class Manager with recommendations of where these segments are and morehelp them with measure the impact of the modified placement.