Quantium Virtual Internship - Retail Strategy and Analytics - Task 1

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.

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

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

filePath <- "/Users/al/career/Quantium/"
```

```
transactionData <- fread(paste0(filePath,"QVI_transaction_data.csv"))
customerData <- fread(paste0(filePath,"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.

```
#### Examine transaction data
# Over to you! Examine the data using one or more of the methods described above.
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 Crinkle Cut Chips Chicken 170g" "Smiths Chip Thinly S/Cream&Onion 175g"
...
## $ 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>
```

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 dates begin on 30
#### Dec 1899
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
# Over to you! Generate a summary of the PROD_NAME column.
head(transactionData$PROD_NAME)
```

```
## [1] "Natural Chip        Compny SeaSalt175g"
## [2] "CCs Nacho Cheese    175g"
## [3] "Smiths Crinkle Cut  Chips Chicken 170g"
## [4] "Smiths Chip Thinly  S/Cream&Onion 175g"
## [5] "Kettle Tortilla ChpsHny&Jlpno Chili 150g"
## [6] "Old El Paso Salsa   Dip Tomato Mild 300g"
```

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.

```r
#### Examine the words in PROD_NAME to see if there are any incorrect entries
#### such as products that are not chips
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()`.

```r
# Over to you! Remove digits, and special characters, and then sort the distinct
# words by frequency of occurrence.

#### Removing digits
productWords <- productWords[!grepl("[0-9]", productWords$words),]
#### Removing special characters
productWords <- productWords[!grepl("[[:punct:]]", productWords$words)]
#### Removing empty strings
productWords <- productWords[productWords$words != ""]

#### Let's look at the most common words by counting the number of times a word
#### appears and sorting them by this frequency in order of highest to lowest
#### frequency
freqdf <- data.frame(table(productWords$words))
freqdf[rev(order(freqdf$Freq)),]
```

```
##                 Var1 Freq
## 29             Chips   21
## 127           Smiths   16
## 49               Cut   14
## 42           Crinkle   14
## 71            Kettle   13
## 117             Salt   12
## 19            Cheese   12
## 93          Original   10
## 116            Salsa    9
## 53           Doritos    9
## 26              Chip    9
## 114              RRD    8
## 107         Pringles    8
## 39              Corn    8
## 168               WW    7
## 22           Chicken    7
## 130             Sour    6
## 120              Sea    6
## 161          Vinegar    5
## 151            Thins    5
## 150           Thinly    5
## 44            Crisps    5
## 24           Chilli    5
## 143          Supreme    4
## 113             Rock    4
## 109              Red    4
## 86           Natural    4
## 67          Infuzions    4
## 50              Deli    4
## 41          Cream    4
## 167     Woolworths    3
## 158       Twisties    3
## 156       Tostitos    3
## 155       Tortilla    3
## 154         Tomato    3
## 144          Sweet    3
## 133            Soy    3
## 122     Sensations    3
## 101           Popd    3
## 96            Paso    3
## 89             Old    3
## 80            Mild    3
## 73            Lime    3
## 54              El    3
## 51             Dip    3
## 36            Cobs    3
## 17             CCs    3
## 164          Waves    2
## 159       Tyrrells    2
## 149           Thai    2
## 147          Tangy    2
## 146            Swt    2
## 137             SR    2
## 126          Smith    2
## 119         Salted    2
## 111          Rings    2
## 104         Potato    2
## 85           Nacho    2
## 77          Medium    2
## 72         Lightly    2
## 64           Honey    2
## 62           Grain    2
## 57          French    2
## 30          Chives    2
## 27          ChipCo    2
## 21        Cheezels    2
## 20         Cheetos    2
## 15          Burger    2
## 9              BBQ    2
## 2              And    2
## 166         Whlgrn    1
## 165        Whlegrn    1
## 163         Vingar    1
## 162         Vinegr    1
## 160            Veg    1
## 157        Truffle    1
## 153            Tom    1
## 152          Tmato    1
## 148          Tasty    1
## 145     SweetChili    1
## 142       Sunbites    1
## 141          Strws    1
## 140          Sthrn    1
## 139           Steak    1
## 138         Stacked    1
## 136          Splash    1
## 135           Spicy    1
## 134              Sp    1
## 132        Southern    1
## 131       SourCream    1
## 129           Snbts    1
## 128          Smoked    1
## 125             Slt    1
## 124            Slow    1
## 123         Siracha    1
## 121 Seasonedchicken    1
## 118           saltd    1
## 115             Rst    1
## 112           Roast    1
## 110             Rib    1
## 108           Puffs    1
## 106           Prawn    1
## 105       PotatoMix    1
## 103             Pot    1
## 102            Pork    1
## 100            Plus    1
## 99            Pesto    1
## 98           Pepper    1
## 97               Pc    1
## 95          Papadums   1
## 94          Originl    1
## 92            Orgnl    1
## 91         OnionDip    1
## 90            Onion    1
## 88               Of    1
## 87              NCC    1
## 84                N    1
## 83          Mzzrlla    1
## 82          Mystery    1
## 81       Mozzarella    1
## 79         Mexicana    1
## 78          Mexican    1
## 76              Med    1
## 75            Mango    1
## 74              Mac    1
## 70              Jam    1
## 69         Jalapeno    1
## 68           Infzns    1
## 66              Hot    1
## 65             Hony    1
## 63          GrnWves    1
## 61          Gcamole    1
## 60           Garden    1
## 59            Fries    1
## 58     FriedChicken    1
## 56          Flavour    1
## 55              Fig    1
## 52          Dorito    1
## 48          Crnkle    1
## 47        Crnchers    1
## 46             Crn    1
## 45             Crm    1
## 43           Crips    1
## 40        Crackers    1
## 38          Compny    1
## 37         Coconut    1
## 35              Co    1
## 34          Chutny    1
## 33             Chs    1
## 32             Chp    1
## 31           Chnky    1
## 28        Chipotle    1
## 25      Chimuchurri   1
## 23           Chili    1
## 18            Ched    1
## 16        Camembert   1
## 14          Btroot    1
## 13             Box    1
## 12        Bolognese   1
## 11             Big    1
## 10           Belly    1
## 8            Basil    1
## 7         Barbeque    1
## 6         Barbecue    1
## 5            Balls    1
## 4              Bag    1
## 3            Bacon    1
## 1            Aioli    1
```

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[, 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.
transactionData[transactionData$PROD_QTY == 200]
```

```
##          DATE STORE_NBR LYLTY_CARD_NBR TXN_ID PROD_NBR
##        <Date>     <int>          <int>  <int>    <int>
## 1: 2018-08-19       226         226000 226201        4
## 2: 2019-05-20       226         226000 226210        4
##                           PROD_NAME PROD_QTY TOT_SALES
##                              <char>    <int>     <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.
transactionData[transactionData$LYLTY_CARD_NBR == 226000]
```

```
##          DATE STORE_NBR LYLTY_CARD_NBR TXN_ID PROD_NBR
##        <Date>     <int>          <int>  <int>    <int>
## 1: 2018-08-19       226         226000 226201        4
## 2: 2019-05-20       226         226000 226210        4
##                           PROD_NAME PROD_QTY TOT_SALES
##                              <char>    <int>     <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[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.

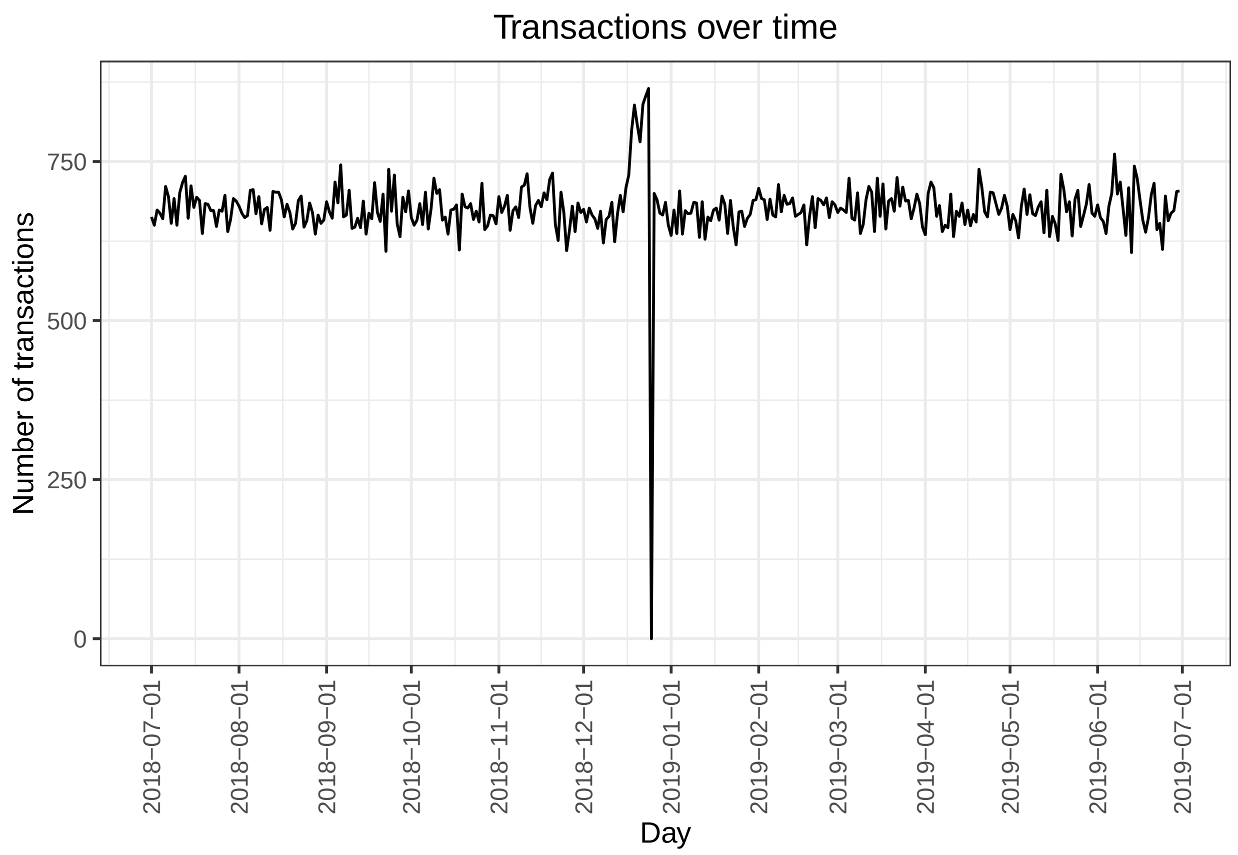
```r
#### Count the number of transactions by date

# Over to you! Create a summary of transaction count by date.
vec_transactions_by_day <- tapply(transactionData, transactionData$DATE, nrow)
length(vec_transactions_by_day)
```

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

```r
# length(unique(transactionData$DATE))
```

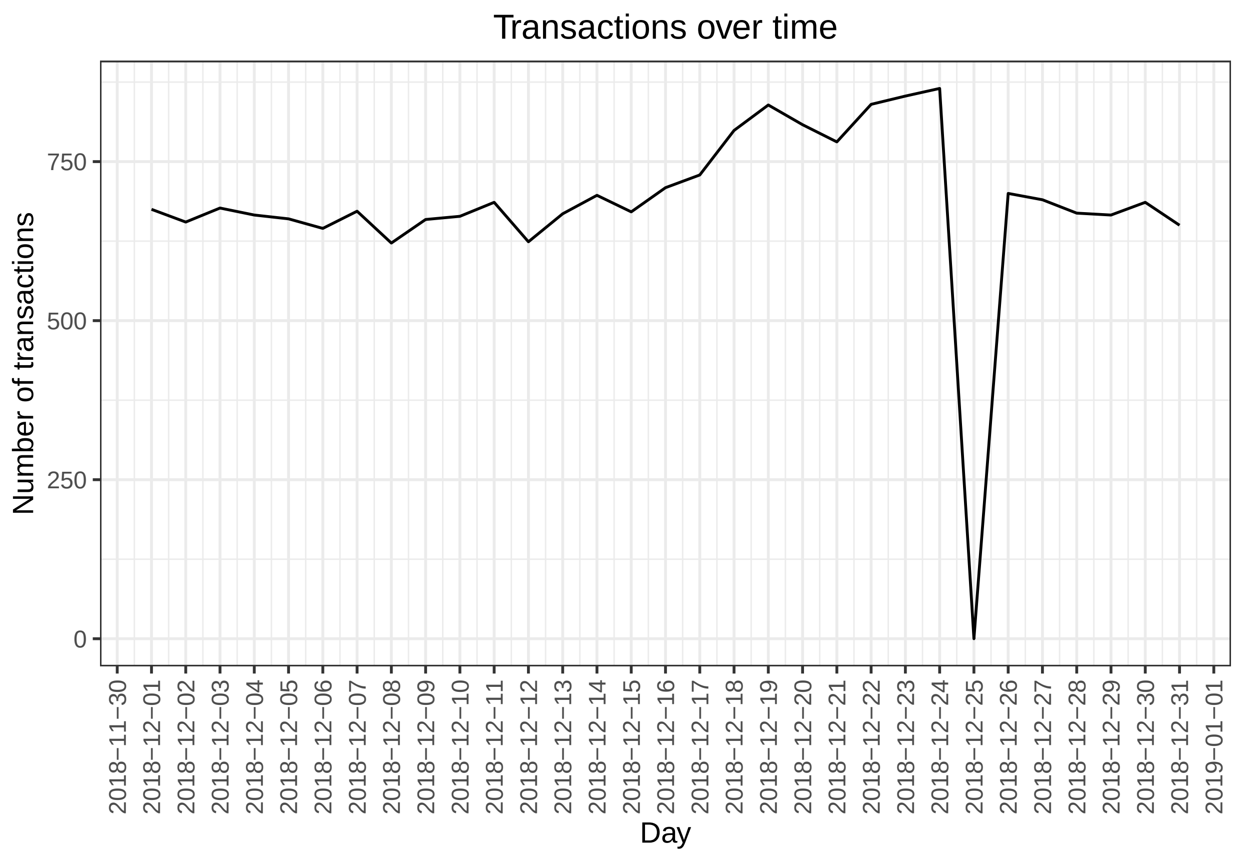
There's only 364 rows, meaning only 364 dates which indicates a missing date. Let's create a sequence of dates from 1 Jul 2018 to 30 Jun 2019 and use this to create a chart of number of transactions over time to find the missing date.

```r
#### 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 join it onto the data to fill in the missing day.
transactions_by_day <- data.frame(DATE = as.Date(names(vec_transactions_by_day)),
                                  N = as.integer(vec_transactions_by_day))

all_dates <- data.frame(DATE = seq(from = as.Date('2018-07-01'),
                                   to = as.Date('2019-06-30'),
                                   by = "day"))

transactions_by_day <- merge(all_dates, transactions_by_day, by = "DATE", all.x = TRUE)
transactions_by_day$N[is.na(transactions_by_day$N)] <- 0

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

#### Plot transactions over time
ggplot(transactions_by_day, aes(x = DATE, y = N)) +
 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 there is an increase in purchases in December and a break in late December. Let's zoom in on this.

```r
#### Filter to December and look at individual days

# Over to you - recreate the chart above zoomed in to the relevant dates.
dec_transactions_by_day <- transactions_by_day[month(transactions_by_day$DATE) == 12,]

#### Plot transactions over time
ggplot(dec_transactions_by_day, aes(x = DATE, y = N)) +
 geom_line() +
 labs(x = "Day", y = "Number of transactions", title = "Transactions over time") +
 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.

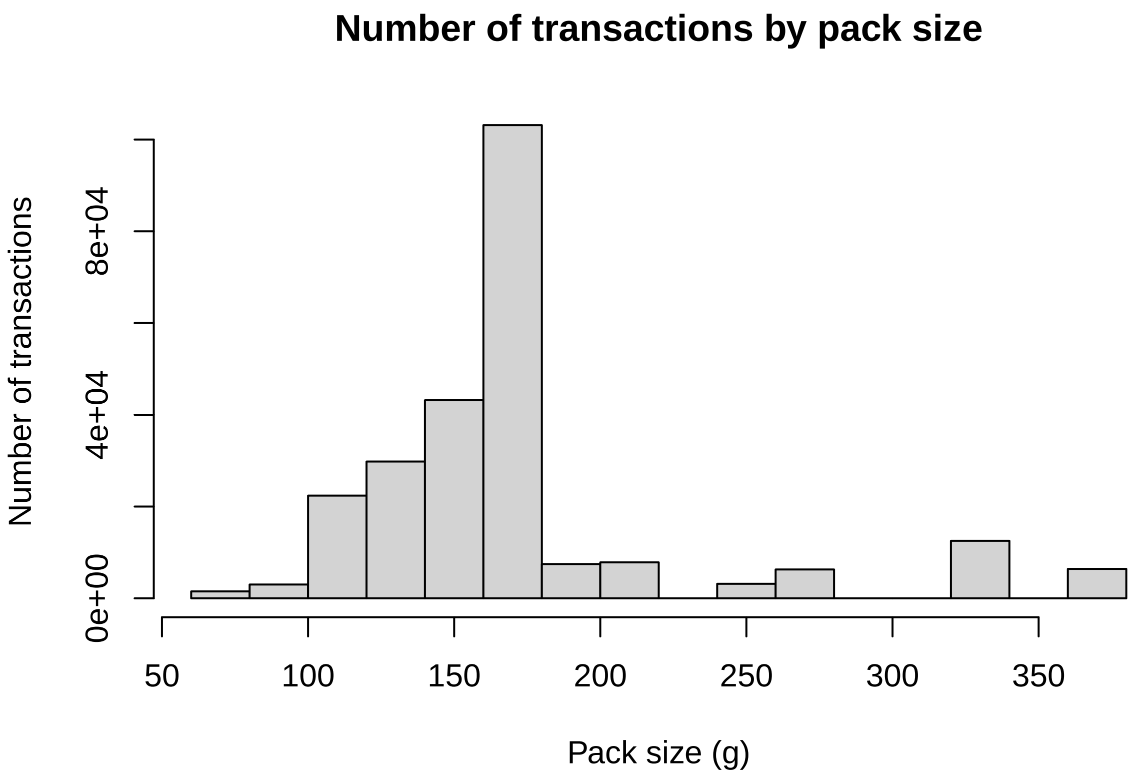
Now that we are satisfied that the data no longer has outliers, we can move on to creating other features such as brand of chips or pack size from PROD_NAME. We will start with pack size.

```
#### 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_SIZE][order(PACK_SIZE)]
```

```
##     PACK_SIZE     N
##         <num> <int>
##  1:        70  1507
##  2:        90  3008
##  3:       110 22387
##  4:       125  1454
##  5:       134 25102
##  6:       135  3257
##  7:       150 40203
##  8:       160  2970
##  9:       165 15297
## 10:       170 19983
## 11:       175 66390
## 12:       180  1468
## 13:       190  2995
## 14:       200  4473
## 15:       210  6272
## 16:       220  1564
## 17:       250  3169
## 18:       270  6285
## 19:       330 12540
## 20:       380  6416
##     PACK_SIZE     N
```

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

```r
#### 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.

# Over to you! Plot a histogram showing the number of transactions by pack size.
hist(transactionData[, PACK_SIZE], main="Number of transactions by pack size",
     ylab="Number of transactions", xlab="Pack size (g)")
```



Pack sizes created look reasonable.

Now to create brands, we can use the first word in PROD_NAME to work out the brand name…

```r
#### Brands
# Over to you! Create a column which contains the brand of the product, by
# extracting it from the product name.
transactionData[, BRAND := sub(" .*", "", PROD_NAME)]
#### Checking brands
# Over to you! Check the results look reasonable.
transactionData[, .N, BRAND][order(BRAND)]
```

```
##          BRAND     N
##         <char> <int>
##  1:     Burger  1564
##  2:        CCs  4551
##  3:    Cheetos  2927
##  4:   Cheezels  4603
##  5:       Cobs  9693
##  6:     Dorito  3183
##  7:    Doritos 22041
##  8:     French  1418
##  9:      Grain  6272
## 10:    GrnWves  1468
## 11:  Infuzions 11057
## 12:     Infzns  3144
## 13:     Kettle 41288
## 14:        NCC  1419
## 15:    Natural  6050
## 16:   Pringles 25102
## 17:        RRD 11894
## 18:        Red  4427
## 19:      Smith  2963
## 20:     Smiths 27390
## 21:      Snbts  1576
## 22:   Sunbites  1432
## 23:      Thins 14075
## 24:   Tostitos  9471
## 25:   Twisties  9454
## 26:   Tyrrells  6442
## 27:         WW 10320
## 28: Woolworths  1516
##          BRAND     N
```

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 == "RED", BRAND := "RRD"]
transactionData[BRAND == "Snbts", BRAND := "Sunbites"]
transactionData[BRAND == "WW", BRAND := "Woolworths"]
transactionData[BRAND == "Smith", BRAND := "Smiths"]
transactionData[BRAND == "Natural", BRAND := "NCC"]
transactionData[BRAND == "Infzns", BRAND := "Infuzions"]
transactionData[BRAND == "Grain", BRAND := "GrnWves"]
transactionData[BRAND == "Dorito", BRAND := "Doritos"]
#### Check again
# Over to you! Check the results look reasonable.
transactionData[, .N, BRAND][order(BRAND)]
```

```
##          BRAND     N
##         <char> <int>
##  1:     Burger  1564
##  2:        CCs  4551
##  3:    Cheetos  2927
##  4:   Cheezels  4603
##  5:       Cobs  9693
##  6:    Doritos 25224
##  7:     French  1418
##  8:    GrnWves  7740
##  9:  Infuzions 14201
## 10:     Kettle 41288
## 11:        NCC  7469
## 12:   Pringles 25102
## 13:        RRD 11894
## 14:        Red  4427
## 15:     Smiths 30353
## 16:   Sunbites  3008
## 17:      Thins 14075
## 18:   Tostitos  9471
## 19:   Twisties  9454
## 20:   Tyrrells  6442
## 21: Woolworths 11836
##          BRAND     N
```

Examining customer data

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

```r
#### 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>
```

```r
summary(customerData)
```

```
##  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
#### 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 be sure 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.

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

```
## [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

```r
fwrite(data, paste0(filePath,"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:

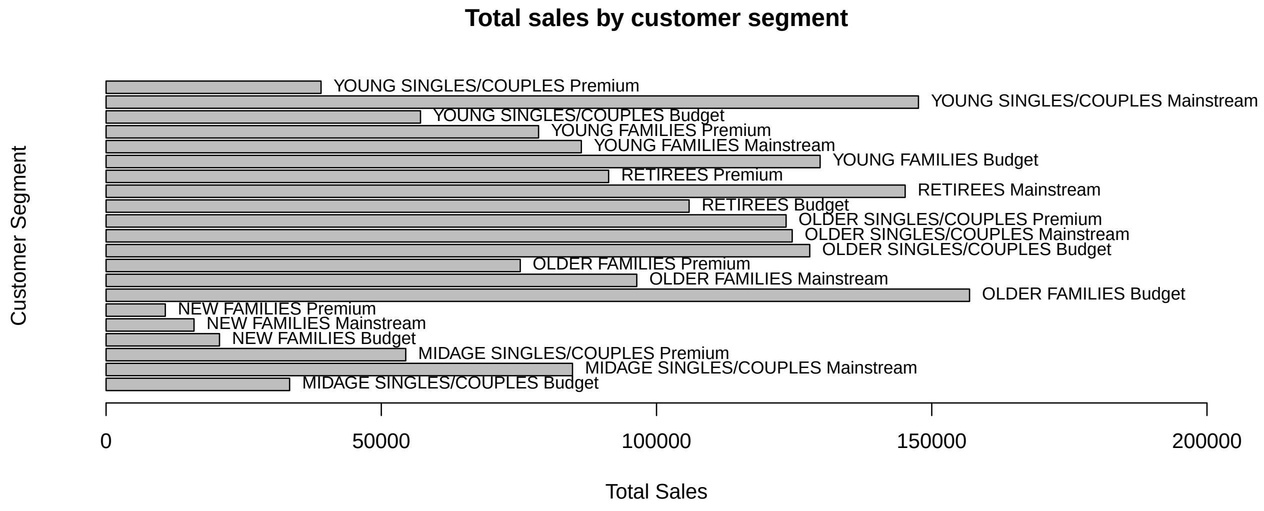
- The customer's total spend over the period and total spend for each transaction to understand what proportion of their grocery spend is on chips
- Proportion of customers in each customer segment overall to compare against the mix of customers who purchase chips

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.

```r
#### Total sales by LIFESTAGE and PREMIUM_CUSTOMER

# Over to you! Calculate the summary of sales by those dimensions and create a
# plot.
sales_summary <- data %>%
  group_by(LIFESTAGE, PREMIUM_CUSTOMER) %>%
  summarise(total_sales = sum(TOT_SALES))
```

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

```r
customer_seg <-  paste(sales_summary$LIFESTAGE, sales_summary$PREMIUM_CUSTOMER)

bar_pos <- barplot(sales_summary$total_sales, horiz = TRUE, xlim = c(0,200000),
                   xlab = 'Total Sales', ylab='Customer Segment', las = 1,
                   main = 'Total sales by customer segment')
text(sales_summary$total_sales, bar_pos, xpd = TRUE, cex = 0.8, pos = 4,
     labels = customer_seg)
```



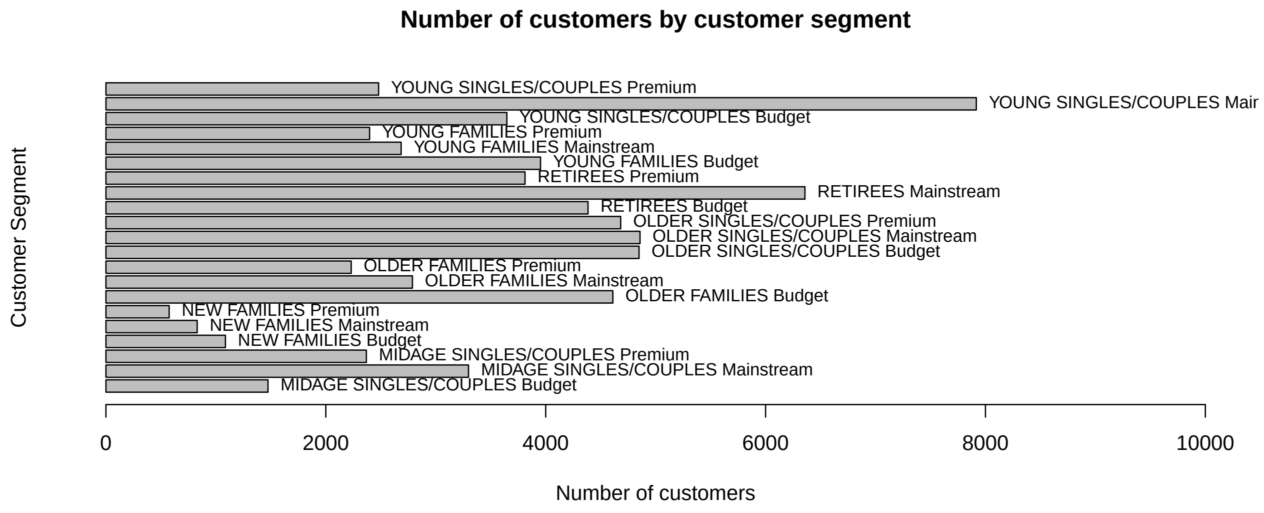
Sales are coming mainly from Budget - older families, Mainstream - young singles/couples, and Mainstream - retirees

Let's see if the higher sales are due to there being more customers who buy chips.

```r
#### 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.
total_cust_summary <- data %>%
  group_by(LIFESTAGE, PREMIUM_CUSTOMER) %>%
  summarise(total_cust = n_distinct(LYLTY_CARD_NBR))
```

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

```r
bar_pos <- barplot(total_cust_summary$total_cust, horiz = TRUE, xlim = c(0,10000),
                   xlab = 'Number of customers', ylab='Customer Segment', las = 1,
                   main = 'Number of customers by customer segment')
text(total_cust_summary$total_cust, bar_pos, xpd = TRUE, cex = 0.8, pos = 4,
     labels = customer_seg)
```



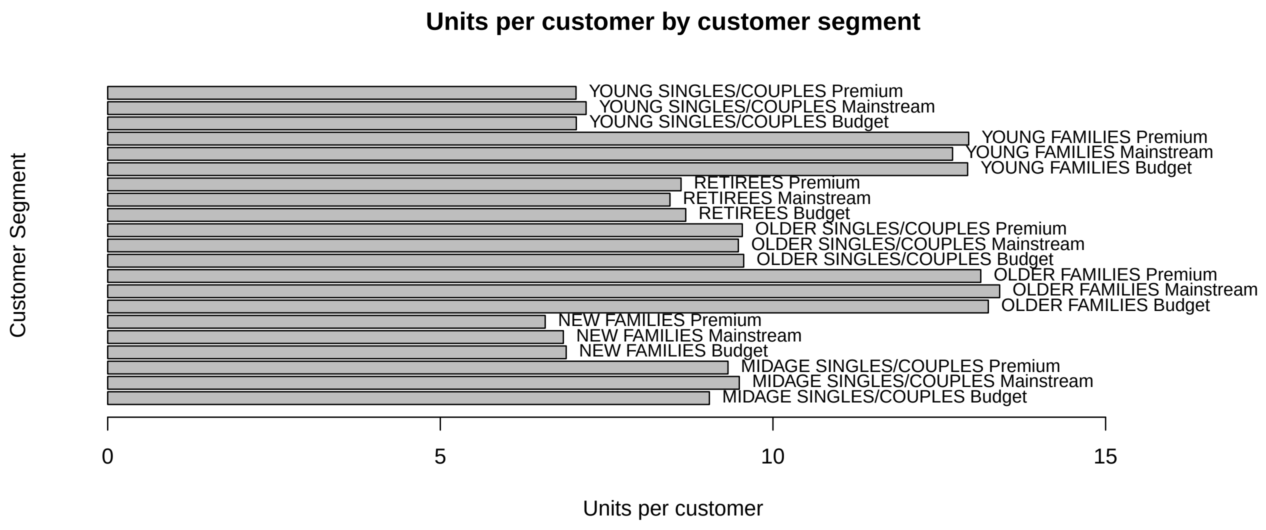
There are more Mainstream - young singles/couples and Mainstream - retirees who buy chips. This contributes to there being more sales to these customer segments but this is not a major driver for the Budget - Older families segment.

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.
total_cust_units <- data %>%
  group_by(LYLTY_CARD_NBR) %>%
  summarise(total_units = sum(PROD_QTY))

unit_per_cust_summary <- data %>%
  left_join(total_cust_units, by = "LYLTY_CARD_NBR") %>%
  group_by(LIFESTAGE, PREMIUM_CUSTOMER) %>%
  summarise(unit_per_cust = mean(total_units))
```

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

```
bar_pos <- barplot(unit_per_cust_summary$unit_per_cust, horiz = TRUE, xlim = c(0,17),
                   xlab = 'Units per customer', ylab='Customer Segment', las = 1,
                   main = 'Units per customer by customer segment')
text(unit_per_cust_summary$unit_per_cust, bar_pos, xpd = TRUE, cex = 0.8,
     pos = 4, labels = customer_seg)
```



Older families and young families in general buy more chips per customer

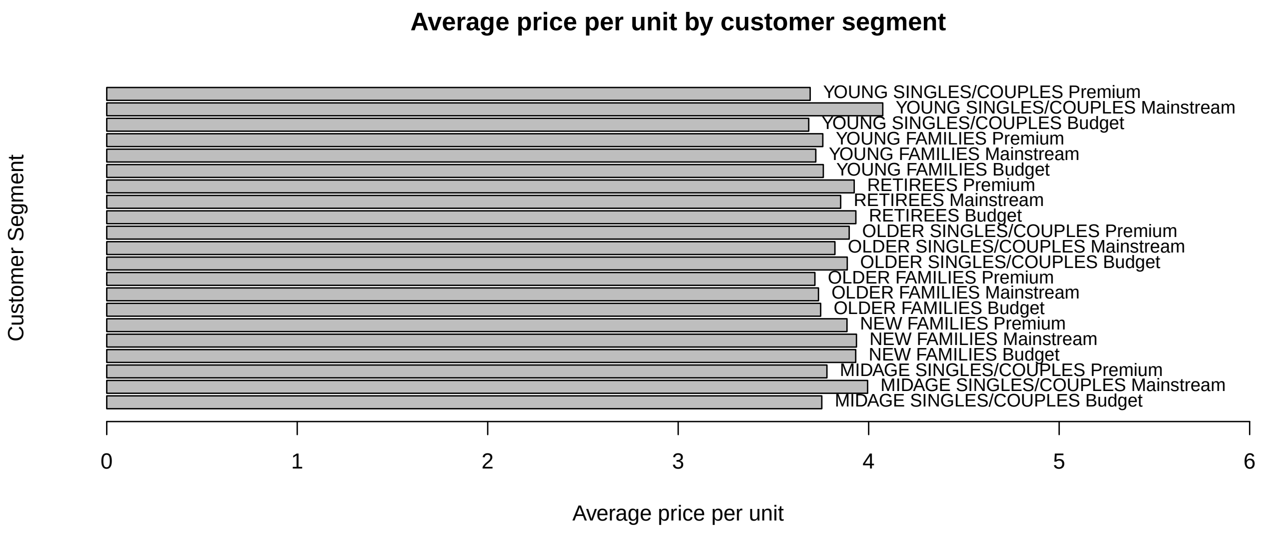
Let's also investigate the average price per unit chips bought for each customer segment as this is also a driver of total sales.

```
#### 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 customer dimensions.
```

```
price_per_unit_summary <- data %>%
  group_by(LIFESTAGE, PREMIUM_CUSTOMER) %>%
  summarise(price_per_unit = sum(TOT_SALES)/sum(PROD_QTY))
```

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

```
bar_pos <- barplot(price_per_unit_summary$price_per_unit, horiz = TRUE, xlim = c(0,6),
                   xlab = 'Average price per unit', ylab='Customer Segment', las = 1,
                   main = 'Average price per unit by customer segment')
text(price_per_unit_summary$price_per_unit, bar_pos, xpd = TRUE, cex = 0.8,
     pos = 4, labels = customer_seg)
```



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 and couples

# Over to you! Perform a t-test to see if the difference is significant.
data[, price_per_unit := TOT_SALES/PROD_QTY]

lifestages <- c("YOUNG SINGLES/COUPLES", "MIDAGE SINGLES/COUPLES")
mainstream_customers <- data[LIFESTAGE %in% lifestages &
                               PREMIUM_CUSTOMER == "Mainstream",
                             price_per_unit]
non_mainstream_customers <- data[LIFESTAGE %in% lifestages &
                                   PREMIUM_CUSTOMER != "Mainstream",
                                 price_per_unit]
t.test(mainstream_customers, non_mainstream_customers, alternative = "greater")
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  mainstream_customers and non_mainstream_customers
## t = 37.624, df = 54791, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.3187234       Inf
## sample estimates:
## mean of x mean of y
##  4.039786  3.706491
```

The t-test results in a p-value of 2.2e-16, i.e. the unit price for mainstream, young and mid-age singles and couples ARE 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 if 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)
target <- data[LIFESTAGE == "YOUNG SINGLES/COUPLES" & PREMIUM_CUSTOMER == "Mainstream",]
others <- data[!(LIFESTAGE == "YOUNG SINGLES/COUPLES" & PREMIUM_CUSTOMER
 ↪  =="Mainstream"),]

target_brand_quantity <- target[, .(
  target_segment = sum(PROD_QTY) / target[, sum(PROD_QTY)]
), by = BRAND]

other_brand_quantity <- others[, .(
  other_segment = sum(PROD_QTY) / others[, sum(PROD_QTY)]
), by = BRAND]

brand_proportions <- merge(target_brand_quantity, other_brand_quantity)[,
  affinity := target_segment / other_segment
]

brand_proportions <- brand_proportions[order(-affinity)]
brand_proportions
```

```
##          BRAND target_segment other_segment  affinity
##         <char>          <num>         <num>     <num>
##  1:   Tyrrells    0.031552795   0.025692464 1.2280953
##  2:   Twisties    0.046183575   0.037876520 1.2193194
##  3:    Doritos    0.122760524   0.101074684 1.2145526
##  4:     Kettle    0.197984817   0.165553442 1.1958967
##  5:   Tostitos    0.045410628   0.037977861 1.1957131
##  6:   Pringles    0.119420290   0.100634769 1.1866703
##  7:       Cobs    0.044637681   0.039048861 1.1431238
##  8:  Infuzions    0.064679089   0.057064679 1.1334347
##  9:      Thins    0.060372671   0.056986370 1.0594230
## 10:    GrnWves    0.032712215   0.031187957 1.0488733
## 11:   Cheezels    0.017971014   0.018646902 0.9637534
## 12:     Smiths    0.096369910   0.124583692 0.7735355
## 13:     French    0.003947550   0.005758060 0.6855694
## 14:    Cheetos    0.008033126   0.012066591 0.6657329
## 15:        RRD    0.032022084   0.049150801 0.6515069
## 16:        Red    0.011787440   0.018342876 0.6426168
## 17:        NCC    0.019599724   0.030853989 0.6352412
## 18:        CCs    0.011180124   0.018895650 0.5916771
## 19:   Sunbites    0.006349206   0.012580210 0.5046980
## 20: Woolworths    0.024099379   0.049427188 0.4875733
## 21:     Burger    0.002926156   0.006596434 0.4435967
##          BRAND target_segment other_segment  affinity
```

We can see that :

Mainstream - young singles/couples prefer Tyrells, Twisites and Doritos compared to the rest of the population, and are much less likely to buy Burger rings, Woolworths or Sunbites.

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.
target_pack_size <- target[, .(
  target_segment = sum(PROD_QTY) / target[, sum(PROD_QTY)]
), by = PACK_SIZE]

other_pack_size <- others[, .(
  other_segment = sum(PROD_QTY) / others[, sum(PROD_QTY)]
), by = PACK_SIZE]

size_proportions <- merge(target_pack_size, other_pack_size)[,
  affinity := target_segment / other_segment
]

size_proportions <- size_proportions[order(-PACK_SIZE)]
size_proportions
```

```
##     PACK_SIZE target_segment other_segment  affinity
##         <num>          <num>         <num>     <num>
##  1:       380    0.032160110   0.025584213 1.2570295
##  2:       330    0.061283644   0.050161917 1.2217166
##  3:       270    0.031828847   0.025095929 1.2682873
##  4:       250    0.014354727   0.012780590 1.1231662
##  5:       220    0.002926156   0.006596434 0.4435967
##  6:       210    0.029123533   0.025121265 1.1593180
##  7:       200    0.008971705   0.018656115 0.4808989
##  8:       190    0.007481021   0.012442016 0.6012708
##  9:       180    0.003588682   0.006066692 0.5915385
## 10:       175    0.254989648   0.270006956 0.9443818
## 11:       170    0.080772947   0.080985964 0.9973697
## 12:       165    0.055652174   0.062267662 0.8937572
## 13:       160    0.006404417   0.012372920 0.5176157
## 14:       150    0.157598344   0.163420656 0.9643722
## 15:        135    0.014768806   0.013075403 1.1295106
## 16:        134    0.119420290   0.100634769 1.1866703
## 17:        125    0.003008972   0.006036750 0.4984423
## 18:        110    0.106280193   0.089791190 1.1836372
## 19:         90    0.006349206   0.012580210 0.5046980
## 20:         70    0.003036577   0.006322350 0.4802924
##     PACK_SIZE target_segment other_segment  affinity
```

Mainstream - young singles/couple have an affinity for larger pack size, preferring the top 4 pack sizes over the rest of the population, and buying the smallest two pack sizes much less than the rest of the population.