Objectives

* To carry out exploratory data analysis on big-data.
* To conduct analysis on the dataset and identify the customer purchasing behaviors.
* To find out which customer type are major buyers.
* To generate useful insights from the data and provide recommendations.

METHODOLOGY

The dataset consisted of the transaction data and customer data. The transaction data consisted of 2,64,836 rows and 8 columns namely, Date, Store Number, Loyalty Card number, Transaction Id, Product Number, Product Name, Product Quantity and Total Sales. This dataset contains information of all transaction done by the customers in one year from 1st July 2018 to 30 June 2019. The table, Customer data consisted of 72,638 rows and three columns namely Loyalty Card Number, Life stage and Premium Customer. This dataset helped analyzing which customers were interested to spend more.

The data sets were analyzed individually as well as in combined form to derive useful insights. Graphical methods were applied to find the pattern in the data. Basic Natural Language processing was in the data preparation part and in formation of word cloud. Decision tree was used to observe the classification of products contributing to the sales. The entire analysis was done in R Studio.

DATA PREPARATION AND DESCRIPTIVE STATISTICS

# Importing data  
library(readxl)  
transaction\_data=read\_excel("C:/Users/Dell/Desktop/pro/transaction\_data.xlsx")

#Checking Structure of data  
head(transaction\_data)

## # A tibble: 6 x 8  
## DATE STORE\_NBR LYLTY\_CARD\_NBR TXN\_ID PROD\_NBR PROD\_NAME PROD\_QTY TOT\_SALES  
## <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <dbl> <dbl>  
## 1 43390 1 1000 1 5 Natural Chi~ 2 6   
## 2 43599 1 1307 348 66 CCs Nacho C~ 3 6.3  
## 3 43605 1 1343 383 61 Smiths Crin~ 2 2.9  
## 4 43329 2 2373 974 69 Smiths Chip~ 5 15   
## 5 43330 2 2426 1038 108 Kettle Tort~ 3 13.8  
## 6 43604 4 4074 2982 57 Old El Paso~ 1 5.1

dim(transaction\_data)

## [1] 264836 8

str(transaction\_data)

## 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 ...

**## Here it can be observed that the DATE column is in numeric format. Also, CSV and Excel integer dates begin on 30 Dec 1899. Therefore, converting the DATE column to right format.**

transaction\_data$DATE=as.Date(transaction\_data$DATE,origin="1899-12-30")

#Summarizing Data  
head(transaction\_data)

## # A tibble: 6 x 8  
## DATE STORE\_NBR LYLTY\_CARD\_NBR TXN\_ID PROD\_NBR PROD\_NAME PROD\_QTY  
## <date> <dbl> <dbl> <dbl> <dbl> <chr> <dbl>  
## 1 2018-10-17 1 1000 1 5 Natural Chip ~ 2  
## 2 2019-05-14 1 1307 348 66 CCs Nacho Cheese~ 3  
## 3 2019-05-20 1 1343 383 61 Smiths Crinkle C~ 2  
## 4 2018-08-17 2 2373 974 69 Smiths Chip Thin~ 5  
## 5 2018-08-18 2 2426 1038 108 Kettle Tortilla ~ 3  
## 6 2019-05-19 4 4074 2982 57 Old El Paso Sals~ 1  
## # ... with 1 more variable: TOT\_SALES <dbl>

summary(transaction\_data)

## 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

#Adding Month column  
library(dplyr)

library(lubridate)

transaction\_data=transaction\_data %>%  
 mutate(MONTH=month(DATE))  
head(transaction\_data)

## # A tibble: 6 x 9  
## DATE STORE\_NBR LYLTY\_CARD\_NBR TXN\_ID PROD\_NBR PROD\_NAME PROD\_QTY  
## <date> <dbl> <dbl> <dbl> <dbl> <chr> <dbl>

## 1 2018-10-17 1 1000 1 5 Natural Chip ~ 2  
## 2 2019-05-14 1 1307 348 66 CCs Nacho Cheese~ 3  
## 3 2019-05-20 1 1343 383 61 Smiths Crinkle C~ 2  
## 4 2018-08-17 2 2373 974 69 Smiths Chip Thin~ 5  
## 5 2018-08-18 2 2426 1038 108 Kettle Tortilla ~ 3  
## 6 2019-05-19 4 4074 2982 57 Old El Paso Sals~ 1  
## # ... with 2 more variables: TOT\_SALES <dbl>, MONTH <dbl>

#Adding column Pack\_size to data  
library(dplyr)  
library(purrr)  
library(stringr)  
transaction\_data=transaction\_data %>%  
 mutate(PACK\_SIZE = readr::parse\_number(as.character(PROD\_NAME)))  
head(transaction\_data)

## # A tibble: 6 x 10  
## DATE STORE\_NBR LYLTY\_CARD\_NBR TXN\_ID PROD\_NBR PROD\_NAME PROD\_QTY  
## <date> <dbl> <dbl> <dbl> <dbl> <chr> <dbl>  
## 1 2018-10-17 1 1000 1 5 Natural Chip ~ 2  
## 2 2019-05-14 1 1307 348 66 CCs Nacho Cheese~ 3  
## 3 2019-05-20 1 1343 383 61 Smiths Crinkle C~ 2  
## 4 2018-08-17 2 2373 974 69 Smiths Chip Thin~ 5  
## 5 2018-08-18 2 2426 1038 108 Kettle Tortilla ~ 3  
## 6 2019-05-19 4 4074 2982 57 Old El Paso Sals~ 1  
## # ... with 3 more variables: TOT\_SALES <dbl>, MONTH <dbl>, PACK\_SIZE <dbl>

**##Data preparation part is done now. We have added extra columns like Month, Pack size to the data which will be useful for further analysis.**

STATISTICAL ANALYSIS

summary(transaction\_data)

## 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

## MONTH PACK\_SIZE   
## Min. : 1.000 Min. : 70.0   
## 1st Qu.: 4.000 1st Qu.:150.0   
## Median : 7.000 Median :170.0   
## Mean : 6.536 Mean :182.4   
## 3rd Qu.:10.000 3rd Qu.:175.0   
## Max. :12.000 Max. :380.0

**##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 products are bought in one transaction.**

#Finding and excluding outliers  
which(transaction\_data$PROD\_QTY==200)

## [1] 69763 69764

transaction\_data[c(69763,69764),]

# A tibble: 2 x 10

DATE STORE\_NBR LYLTY\_CARD\_NBR TXN\_ID PROD\_NBR PROD\_NAME PROD\_QTY TOT\_SALES

<date> <dbl> <dbl> <dbl> <dbl> <chr> <dbl> <dbl>

1 2018-08-19 226 226000 226201 4 Dorito Corn~ 200 650

2 2019-05-20 226 226000 226210 4 Dorito Corn~ 200 650

**## There are two transactions where 200 products are bought in one transaction and both of these transactions where done by the same customer.**

transaction\_data=transaction\_data[-(69763:69764),]  
dim(transaction\_data)

## [1] 264834 10

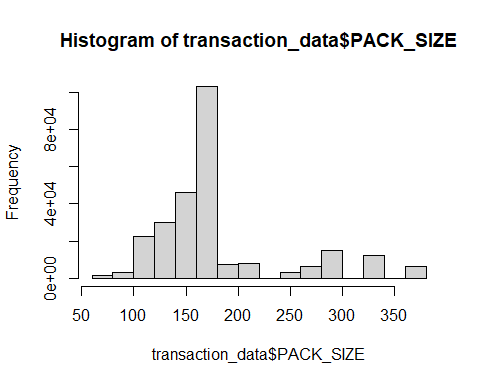
**##It looks like this customer has only had the two transactions over the year and is not an ordinary retail customer. The customer might be buying chips for commercial purposes instead. We’ll remove this loyalty card number from data frame.**

str(transaction\_data)

## tibble [264,834 x 10] (S3: tbl\_df/tbl/data.frame)  
## $ DATE : Date[1:264834], format: "2018-10-17" "2019-05-14" ...  
## $ STORE\_NBR : num [1:264834] 1 1 1 2 2 4 4 4 5 7 ...  
## $ LYLTY\_CARD\_NBR: num [1:264834] 1000 1307 1343 2373 2426 ...  
## $ TXN\_ID : num [1:264834] 1 348 383 974 1038 ...  
## $ PROD\_NBR : num [1:264834] 5 66 61 69 108 57 16 24 42 52 ...  
## $ PROD\_NAME : chr [1:264834] "Natural Chip Compny SeaSalt175g" "CCs Nacho Cheese 175g" "Smiths Crinkle Cut Chips Chicken 170g" "Smiths Chip Thinly S/Cream&Onion 175g" ...  
## $ PROD\_QTY : num [1:264834] 2 3 2 5 3 1 1 1 1 2 ...  
## $ TOT\_SALES : num [1:264834] 6 6.3 2.9 15 13.8 5.1 5.7 3.6 3.9 7.2 ...  
## $ MONTH : num [1:264834] 10 5 5 8 8 5 5 5 8 8 ...  
## $ PACK\_SIZE : num [1:264834] 175 175 170 175 150 300 330 210 150 210 ...

**## Now we find that our data is in better format and good for further analysis.**

#Finding which preferred pack size  
hist(transaction\_data$PACK\_SIZE)



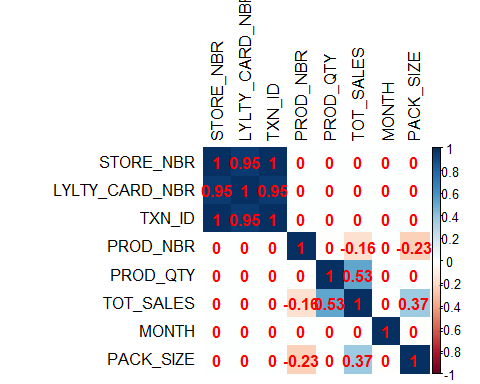
y=table(transaction\_data$PACK\_SIZE)  
names(y)[which(y==max(y))]

## [1] "175"

**## Pack sizes created look reasonable. From Histogram we can conclude the pack size of 175g is most preferred by customers.**

#To see Correlation between variables  
library(corrplot)

correlation=cor(transaction\_data[,-c(1,6)])  
corrplot(correlation, method="color", addCoef.col= "red",tl.col="black")



**##From above correlation plot it can be observed that store number and transaction id are completely correlated. Also, Loyalty Card number and Store number are highly correlated. While other correlations may be neglected as they are not much significant.**

#Exploring Consumer data  
library(readr)  
library(dplyr)  
Customer\_data=read\_csv("C:/Users/Dell/Desktop/pro/purchase\_data.csv")

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## LYLTY\_CARD\_NBR = col\_double(),  
## LIFESTAGE = col\_character(),  
## PREMIUM\_CUSTOMER = col\_character()  
## )

head(Customer\_data)

## # A tibble: 6 x 3  
## LYLTY\_CARD\_NBR LIFESTAGE PREMIUM\_CUSTOMER  
## <dbl> <chr> <chr>   
## 1 1000 YOUNG SINGLES/COUPLES Premium   
## 2 1002 YOUNG SINGLES/COUPLES Mainstream   
## 3 1003 YOUNG FAMILIES Budget   
## 4 1004 OLDER SINGLES/COUPLES Mainstream   
## 5 1005 MIDAGE SINGLES/COUPLES Mainstream   
## 6 1007 YOUNG SINGLES/COUPLES Budget

dim(Customer\_data)

## [1] 72637 3

table(Customer\_data$LIFESTAGE)

##   
## MIDAGE SINGLES/COUPLES NEW FAMILIES OLDER FAMILIES   
## 7275 2549 9780   
## OLDER SINGLES/COUPLES RETIREES YOUNG FAMILIES   
## 14609 14805 9178   
## YOUNG SINGLES/COUPLES   
## 14441

table(Customer\_data$PREMIUM\_CUSTOMER)

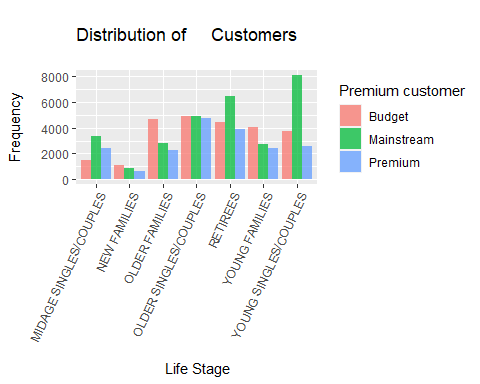
##   
## Budget Mainstream Premium   
## 24470 29245 18922

table(Customer\_data$LIFESTAGE,Customer\_data$PREMIUM\_CUSTOMER)

##   
## Budget Mainstream Premium  
## MIDAGE SINGLES/COUPLES 1504 3340 2431  
## NEW FAMILIES 1112 849 588  
## OLDER FAMILIES 4675 2831 2274  
## OLDER SINGLES/COUPLES 4929 4930 4750  
## RETIREES 4454 6479 3872  
## YOUNG FAMILIES 4017 2728 2433  
## YOUNG SINGLES/COUPLES 3779 8088 2574

***## It is clear that maximum customers are of mainstream with life stage young singles/Couples. We can visualize the above table graphically as following.***

info=data.frame(table(Customer\_data$LIFESTAGE,Customer\_data$PREMIUM\_CUSTOMER))  
colnames(info)=c("Life stage","Premium customer","Freq")  
  
library(ggplot2)  
ggplot(data =info, aes(x=`Life stage`, y=Freq, fill=`Premium customer`)) +  
 geom\_bar(stat = "identity", position = position\_dodge(), alpha = 0.75)+  
 labs(x="\n Life Stage", y="Frequency\n", title="\nDistribution of Customers\n")+  
 theme(axis.text.x = element\_text(angle =65, hjust = 1))



**## It can be concluded that the most of the customers are of Mainstream type. From those Young singles/Couples and Retirees being maximum. Also new Families being least customers.**

#Combining transaction and customer data set.  
trans\_data=transaction\_data %>%  
 group\_by(LYLTY\_CARD\_NBR) %>%  
 summarize(Prod\_qty=sum(PROD\_QTY),  
 Tot\_sales=sum(TOT\_SALES),  
 Store\_Nbr=mean(STORE\_NBR),  
 Pack\_size=sum(PACK\_SIZE));  
  
head(trans\_data)

## # A tibble: 6 x 5  
## LYLTY\_CARD\_NBR Prod\_qty Tot\_sales Store\_Nbr Pack\_size  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1000 2 6 1 175  
## 2 1002 1 2.7 1 150  
## 3 1003 2 6.6 1 385  
## 4 1004 1 1.9 1 160  
## 5 1005 1 2.8 1 165  
## 6 1007 2 6.5 1 260

dim(trans\_data)

## [1] 72636 5

comb\_data=data.frame(inner\_join(trans\_data,Customer\_data,by="LYLTY\_CARD\_NBR"))  
head(comb\_data)

## LYLTY\_CARD\_NBR Prod\_qty Tot\_sales Store\_Nbr Pack\_size LIFESTAGE  
## 1 1000 2 6.0 1 175 YOUNG SINGLES/COUPLES  
## 2 1002 1 2.7 1 150 YOUNG SINGLES/COUPLES  
## 3 1003 2 6.6 1 385 YOUNG FAMILIES  
## 4 1004 1 1.9 1 160 OLDER SINGLES/COUPLES  
## 5 1005 1 2.8 1 165 MIDAGE SINGLES/COUPLES  
## 6 1007 2 6.5 1 260 YOUNG SINGLES/COUPLES

## PREMIUM\_CUSTOMER  
## 1 Premium  
## 2 Mainstream  
## 3 Budget  
## 4 Mainstream  
## 5 Mainstream  
## 6 Budget

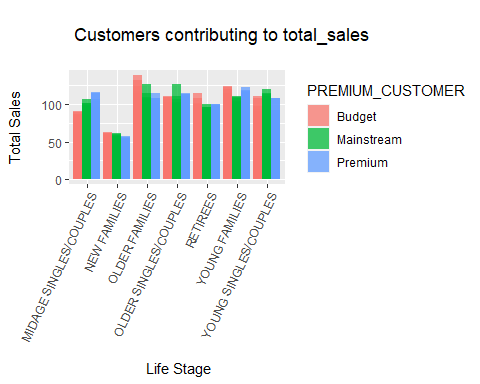
dim(comb\_data)

## [1] 72636 7

summary(comb\_data)

## LYLTY\_CARD\_NBR Prod\_qty Tot\_sales Store\_Nbr   
## Min. : 1000 Min. : 1.000 Min. : 1.50 Min. : 1.0   
## 1st Qu.: 66202 1st Qu.: 2.000 1st Qu.: 9.10 1st Qu.: 66.0   
## Median : 134040 Median : 6.000 Median : 21.70 Median :133.0   
## Mean : 136185 Mean : 6.949 Mean : 26.61 Mean :135.1   
## 3rd Qu.: 203374 3rd Qu.:10.000 3rd Qu.: 40.00 3rd Qu.:203.0   
## Max. :2373711 Max. :36.000 Max. :138.60 Max. :272.0   
## Pack\_size LIFESTAGE PREMIUM\_CUSTOMER   
## Min. : 70.0 Length:72636 Length:72636   
## 1st Qu.: 280.0 Class :character Class :character   
## Median : 520.0 Mode :character Mode :character   
## Mean : 665.1   
## 3rd Qu.: 965.0   
## Max. :3395.0

library(ggplot2)  
ggplot(data =comb\_data, aes(x=`LIFESTAGE`, y=`Tot\_sales`, fill=`PREMIUM\_CUSTOMER`)) +  
 geom\_bar(stat = "identity", position = position\_dodge(), alpha = 0.75)+  
 labs(x="\n Life Stage", y="Total Sales\n", title="\n Customers contributing to total\_sales\n")+theme(axis.text.x = element\_text(angle =65, hjust = 1))



**##It can be concluded from above graph that Budget, Mainstream and Premium customers within each life stage is almost similar.** **Older families contributing most towards total sales followed by young families and Retiress. Again, new families being least.**

# Cleaning and Analyzing Product name column  
name=data.frame(table((transaction\_data$PROD\_NAME)))  
library(wordcloud2)

library(RColorBrewer)  
library(tm)

getwd()

## [1] "C:/Users/Dell/Desktop/pro"

library(readr)  
write\_tsv(name, file = "name.txt")  
text=readLines(file.choose())  
docs=Corpus(VectorSource(text))  
library(dplyr)

docs=docs %>%  
 tm\_map(removeNumbers) %>%  
 tm\_map(removePunctuation) %>%

tm\_map(stripWhitespace)

docs=tm\_map(docs, content\_transformer(tolower))

docs=tm\_map(docs, removeWords, stopwords("english"))

dtm=TermDocumentMatrix(docs)   
matrix=as.matrix(dtm)   
words=sort(rowSums(matrix),decreasing=TRUE)   
df=data.frame(word = names(words),freq=words)  
set.seed(678)  
wordcloud2(data=df,size=1.2, color='random-dark')



**##From above word cloud we can say that** **the product “Chips” is most purchased product and especially of “Kettle” brand followed by “RRD” brand.**

#Checking total sales  
total\_sales=sum(transaction\_data$TOT\_SALES);total\_sales

## [1] 1933115

***##The total sales from last year is 1933115 units.***

#Checking Sales by Store number  
d\_table=data.frame(sort(table(transaction\_data$STORE\_NBR)))

f=d\_table$Freq

d\_table %>% mutate(cum\_per=(cumsum(f)/sum(f)\*100))

Var1 Freq cum\_per

1 76 1 3.775922e-04

2 92 1 7.551843e-04

3 11 2 1.510369e-03

4 31 2 2.265553e-03

5 206 2 3.020737e-03

6 211 2 3.775922e-03

7 252 2 4.531106e-03

8 85 3 5.663883e-03

9 193 3 6.796659e-03

10 117 46 2.416590e-02

11 198 47 4.191273e-02

12 258 47 5.965956e-02

13 146 56 8.080472e-02

14 177 56 1.019499e-01

15 44 58 1.238502e-01

16 244 59 1.461282e-01

17 140 61 1.691613e-01

18 263 61 1.921944e-01

19 42 64 2.163603e-01

20 99 64 2.405262e-01

21 14 66 2.654473e-01

22 267 66 2.903684e-01

23 158 67 3.156671e-01

24 161 67 3.409657e-01

25 218 67 3.662644e-01

26 224 67 3.915631e-01

27 139 68 4.172393e-01

28 159 68 4.429156e-01

29 204 68 4.685919e-01

30 127 72 4.957785e-01

31 135 73 5.233427e-01

………………

………………

………………

239 72 1628 7.864150e+01

240 270 1630 7.925697e+01

241 157 1631 7.987283e+01

242 69 1633 8.048943e+01

243 223 1636 8.110718e+01

244 26 1640 8.172643e+01

245 94 1641 8.234606e+01

246 130 1641 8.296568e+01

247 259 1642 8.358569e+01

248 81 1649 8.420834e+01

249 112 1651 8.483174e+01

250 201 1654 8.545628e+01

251 203 1661 8.608346e+01

252 71 1676 8.671631e+01

253 4 1678 8.734991e+01

254 168 1682 8.798502e+01

255 199 1687 8.862202e+01

256 179 1702 8.926468e+01

257 133 1703 8.990772e+01

258 152 1707 9.055227e+01

259 100 1711 9.119833e+01

260 128 1714 9.184552e+01

261 63 1742 9.250329e+01

262 58 1743 9.316143e+01

263 156 1747 9.382108e+01

264 213 1747 9.448074e+01

265 230 1751 9.514190e+01

266 40 1764 9.580797e+01

267 43 1771 9.647669e+01

268 237 1785 9.715069e+01

269 165 1819 9.783753e+01

270 93 1832 9.852928e+01

271 88 1873 9.923651e+01

272 226 2022 1.000000e+02

**##****The stores with store number 226, 88, 93, 165 and 237 contributes the most. It can be observed that some stores contribute very less to the total Sales. The first 30 stores contribute less than 0.5%. Therefore, some action may be taken to improve the performance of those stores.**

#Sorting Total Sales month-wise  
transaction\_data %>%  
 group\_by(MONTH) %>%  
 summarize(tot\_sales=sum(TOT\_SALES))

## # A tibble: 12 x 2  
## MONTH tot\_sales  
## \* <dbl> <dbl>  
## 1 1 162642.  
## 2 2 150665   
## 3 3 166265.  
## 4 4 159845.  
## 5 5 156718.  
## 6 6 160539.  
## 7 7 165275.  
## 8 8 158081.  
## 9 9 160522   
## 10 10 164416.  
## 11 11 160234.  
## 12 12 167913.

***##It can be seen that maximum sales were in the month of December and also the sales don’t vary much month-wise.***

data=transaction\_data %>%  
 group\_by(DATE) %>%  
 summarize(tot\_sales=sum(TOT\_SALES)) %>%  
 mutate(month=month(DATE))

head(data)

## # A tibble: 6 x 3  
## DATE tot\_sales month  
## <date> <dbl> <dbl>  
## 1 2018-07-01 5372. 7  
## 2 2018-07-02 5315. 7  
## 3 2018-07-03 5322. 7  
## 4 2018-07-04 5310. 7  
## 5 2018-07-05 5081. 7  
## 6 2018-07-06 5605. 7

dim(data)

## [1] 364 3

**##There are only 364 rows, meaning only 364 dates which indicates a missing date. So, we create a sequence of dates from 1 Jul 2018 to 30 Jun 2019 and use this to find the missing date.**

date=data$DATE

date\_range=seq(min(date), max(date), by = 1)

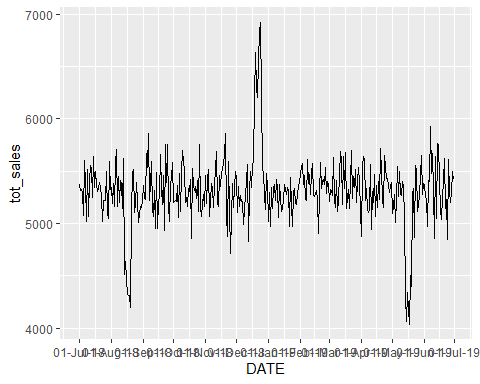
date\_range[!date\_range %in% date]

## "2018-12-25"

**##****It can be observed that there were no transactions on 25th December 2018. It may be that the stores would be closed due to Christmas and Christmas may be the reason that the sales were highest in the month of December.**

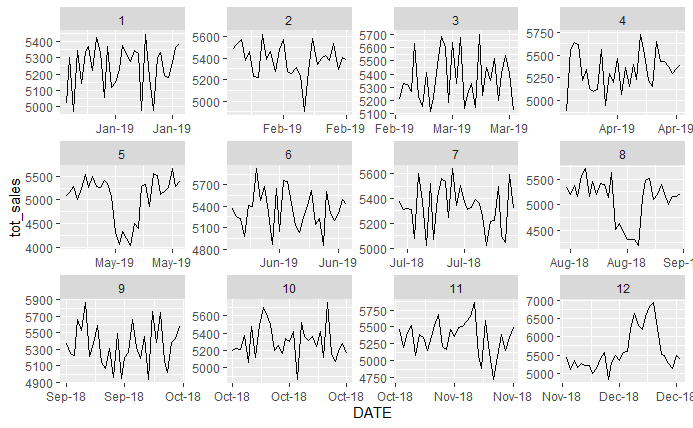
**##Visualizing the total sales**

library(readr)   
library(lubridate)  
library(scales)  
library(ggplot2)  
ggplot(data, aes(x =DATE, y =tot\_sales))+geom\_line()+  
 scale\_x\_date(labels=date\_format("%d-%b-%y"),breaks = ("1 month"))



***##TOTAL SALES MONTH WISE***

ggplot(data, aes(x =DATE, y =tot\_sales))+geom\_line()+  
 scale\_x\_date(labels=date\_format("%b-%y"),breaks=("15 days"))+  
 facet\_wrap(~month,scales = "free")



## Decision tree

#In order to obtain a decision tree, we added a column “level” which would facilitate the binary classification of data along the parameter “total sales”. Total sales greater than 7 are classified as yes meaning it contributed highly in the sales of product and total sales less than 7 implying less contribution comparatively. 7.304 being the mean of the total sales column we chose the reference value to be 7.

level=ifelse(transaction\_data$TOT\_SALES<=7,"No","Yes")  
dec\_data=cbind(transaction\_data[,-c(1,3,4,5,6,8)],level)  
head(dec\_data)

## STORE\_NBR PROD\_QTY MONTH PACK\_SIZE level  
## 1 1 2 10 175 No  
## 2 1 3 5 175 No  
## 3 1 2 5 170 No  
## 4 2 5 8 175 Yes  
## 5 2 3 8 150 Yes  
## 6 4 1 5 300 No

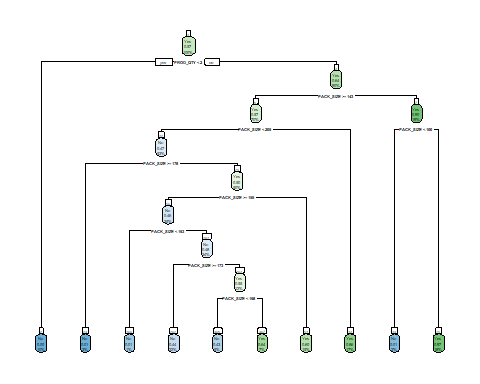
***# We divide 80% data as train data and 20% data as test data.***

library(caTools)  
set.seed(987)  
sample=sample.split(dec\_data, SplitRatio=0.8)  
train=subset(dec\_data, sample==TRUE)  
test=subset(dec\_data,sample==FALSE)  
library(rpart)  
tree=rpart(level~.,data=train,method = "class")  
tree

## n= 211867   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 211867 90226 Yes (0.425861507 0.574138493)   
## 2) PROD\_QTY< 1.5 22027 0 No (1.000000000 0.000000000) \*  
## 3) PROD\_QTY>=1.5 189840 68199 Yes (0.359244627 0.640755373)   
## 6) PACK\_SIZE>=142.5 148967 64029 Yes (0.429820027 0.570179973)   
## 12) PACK\_SIZE< 205 111999 53149 No (0.525451120 0.474548880)   
## 24) PACK\_SIZE>=177.5 6201 32 No (0.994839542 0.005160458) \*  
## 25) PACK\_SIZE< 177.5 105798 52681 Yes (0.497939470 0.502060530)   
## 50) PACK\_SIZE>=155 74794 34625 No (0.537061796 0.462938204)   
## 100) PACK\_SIZE< 162.5 2057 13 No (0.993680117 0.006319883) \*  
## 101) PACK\_SIZE>=162.5 72737 34612 No (0.524148645 0.475851355)   
## 202) PACK\_SIZE>=172.5 47485 20759 No (0.562830367 0.437169633) \*  
## 203) PACK\_SIZE< 172.5 25252 11399 Yes (0.451409789 0.548590211)   
## 406) PACK\_SIZE< 167.5 10894 4645 No (0.573618506 0.426381494) \*  
## 407) PACK\_SIZE>=167.5 14358 5150 Yes (0.358685054 0.641314946) \*  
## 51) PACK\_SIZE< 155 31004 12512 Yes (0.403560831 0.596439169) \*  
## 13) PACK\_SIZE>=205 36968 5179 Yes (0.140094135 0.859905865) \*  
## 7) PACK\_SIZE< 142.5 40873 4170 Yes (0.102023341 0.897976659)   
## 14) PACK\_SIZE< 100 3189 19 No (0.994042019 0.005957981) \*  
## 15) PACK\_SIZE>=100 37684 1000 Yes (0.026536461 0.973463539) \*

**##It comes out that the product quantity and pack size being the major split node.**

# Visualize the decision tree with rpart.plot  
library(rpart.plot)  
rpart.plot(tree,nn=TRUE)



Prediction1=predict(tree,newdata=test,type = 'class')

# Confusion matrix  
library(caret)  
library(e1071)  
confusionMatrix(as.factor(Prediction1),as.factor(test$level))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 16684 6402  
## Yes 5877 24004  
##   
## Accuracy : 0.7682   
## 95% CI : (0.7646, 0.7718)  
## No Information Rate : 0.5741   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5274   
##   
## Mcnemar's Test P-Value : 2.259e-06   
##   
## Sensitivity : 0.7395   
## Specificity : 0.7894   
## Pos Pred Value : 0.7227   
## Neg Pred Value : 0.8033   
## Prevalence : 0.4259   
## Detection Rate : 0.3150   
## Detection Prevalence : 0.4359   
## Balanced Accuracy : 0.7645   
##   
## 'Positive' Class : No   
##

**##We find a decision tree with 76.82% accuracy, 73.95% Sensitivity and 78.9% specificity.**

CONCLUSIONS:

Some of the major insights drawn from analysis were:

1) Older families were contributing most towards total sales followed by young families and Retirees. New families contributing least, its due to the fact that new-families themselves being least customers in number.

2) The product “Chips” is most purchased product and especially of “Kettle” brand followed by “RRD” brand.

3)Pack size of 175g is most preferred by customers.

4)Maximum sales were in the month of December and also the sales don’t vary much month-wise.

5) It can be observed that there were no transactions on 25th December 2018. It may be that the stores would be closed due to Christmas and Christmas may be the reason that the sales were highest in the month of December. It the only day when there were no transactions.

6)The stores with store number 226, 88, 93, 165 and 237 contributes the most***.***

7) While creating a Decision tree it comes out that the product quantity and pack size being the major split node.

8) We find a decision tree with 76.82% accuracy, 73.95% Sensitivity and 78.9% specificity.