Coffee Shop Customer Propensity Analysis

1. Approach overview
2. Data cleaning and preparation
3. Approach 1: Identification of customer loyalty - exploratory approach
4. Approach 2: Sales Distribution Pattern
5. Approach 3: Customer profiling and high priority profile
6. Approach 4: Purchase pattern of priority profile
7. Approach 5: Pricing strategy
8. Conclusions & recommendations

Approach overview

Central Perk wants to smoothen their demand across the day. They hold an assumption that they have a loyal customer base that drives a majority of their sales. We believe that understanding the behavioral patterns of customers will provide us with vital information about whether Central Perk has a loyal customer base. As a next step, such patterns could be leveraged to extend this behavior to a wider customer base to achieve desired modifications in the sales trend. On the other hand, if Central Perk does not possess a loyal customer base, we would want to understand the reasons behind the churn. This can help us in identifying areas of improvement towards better customer satisfaction.

The overall sales pattern across a day for Central Perk aligns with typical life cycle of working professionals. To summarize:

1. During the weekdays, a sales peak is observed at around 8am which then falls through the afternoon and again slightly rises around the lunch duration followed by a drop by the end of the day
2. During the weekends, sales go on rising till around 10am indicating a late start of the day which is followed by consistent drop across the rest of the day

Since, we expect sales to be higher in the first half of the day for a coffee-shop business, attempting to shift this early morning demand across the day would violate the typical ‘need’ pattern of the customers and could result in undesirable effects.

On the other hand, our approach focuses on **understanding the behavior of loyal customers in low-sales periods and proposing a strategy to extend this behavior to more customers** by providing incentives to them. We expect a **lift in sales because of this strategy aligning it with high sales period, thus flattening the demand as the end outcome.**

Hence, the high-level approach is as follows:

1. Uncovering customer loyalty
2. Understanding behavioral patterns of loyal customers in areas of focus (Low sales) OR understanding reasons for the churn
3. Extending relevant behavior to a wider customer base by providing incentives and thereby lifting the sales
4. Conclusively, aligning the sales across the day and overall time period

---

Data cleaning and preparation

The data is procured for 3 years separately i.e., for 2016, 2017 and 2018. However, the data for 2016 and 2018 are not complete. The range of the data is from 15th July 2016 to 24th August 2018.

```
data_16 <- read.csv("Central Perk Item Sales Summary 2016.csv",
stringsAsFactors = F, na.strings = c('', NA, ' '))
data_17 <- read.csv("Central Perk Item Sales Summary 2017.csv",
stringsAsFactors = F, na.strings = c('', NA, ' '))
data_18 <- read.csv("Central Perk Item Sales Summary 2018.csv",
stringsAsFactors = F, na.strings = c('', NA, ' '))

data_total <- rbind(data_16, data_17, data_18)
```

The data is cleaned in the following four steps –

- Data is checked for NA values; the table below shows the number of NA values in each column of data

```
missing <- lapply(data_total, is.na)
NAs <- sapply(missing, sum)
NAs
```

Field	Count of NA values
Notes	221471
Customer ID	78077
Price Point Name	86
Date, Time, category, Item, Quantity, Gross Sales, Discounts, Net Sales, Tax, Event Type	1

Among the fields with a lot of NA values, customer ID is the only column that is valuable for us in the analysis. So, we ignore the fields Notes and Price Point Name. However, we need customer ID majorly for the customer analysis. Hence, we remove NA values from customer ID specifically for the customer analysis and from the rest of the columns having 1 NA value each for all the analyses.

```
data_total <- data_total[complete.cases(data_total[,-c(11,13)]),]
```

- Date and Time fields are concatenated and are converted into POSIXct object for the datetime reference

```
data_total$Date <- as.Date(data_total$Date, "%m/%d/%y")
data_total$Date_time <- paste(data_total$Date, data_total$Time)
data_total$Date_time <- as.POSIXct(data_total$Date_time, "%m/%d/%y %H:%M:%S")
```

- Sales fields such as Gross Sales, Net Sales, Tax and Discounts are cleaned to convert to numeric values (code is similar for the other fields)

```
data_total$Gross.Sales <- gsub("\\$", "", data_total$Gross.Sales)
data_total$Gross.Sales <- gsub("\\(", "-", data_total$Gross.Sales)
data_total$Gross.Sales <- gsub("\\)", "", data_total$Gross.Sales)
data_total$Gross.Sales <- as.numeric(data_total$Gross.Sales)
```

- Value in the Item field is cleaned

```
data_total[data_total$Item == "ðŸ\u008d‹LemonadeðŸ\u008d‹","Item"] <- Lemonade"
```

Approach 1: Identification of customer loyalty - exploratory approach

Description and rationale for the analysis:

As a first step, we want to validate whether loyalty among Central Perk customers exists as claimed by the business. Identification of loyal customers would help us in providing them with incentives for better customer experience while helping us in devising a strategy to make more customers follow that pattern. In case of absence of loyal customers, we would like the store management to be aware of this issue and dig more into the reasons for customer dissatisfaction or any other causes leading to customer churn. The crudest way to observe loyalty would be to understand if a small chunk of customers is driving significant amount of sales - Pareto's principle. However, we would like to perform thorough checks to support our findings.

Key Goal: Identify whether a loyal customer base exists for Central Perk

The first step to identify loyalty among customers is to determine the share of sales generated through repeating customers. Based on Pareto's 80-20 rule, we expect majority share of the sales to be driven by a small proportion of the customers who pay repeated visits to our store. Keeping only the customer data that has a customer id associated to it and then determining various customer metrics such as lifetime purchase and tenure (day of last transaction - day of first transaction) of the customer

```
loyalty <- cp %>% filter(!is.na(Customer.ID)) %>% group_by(Customer.ID) %>%
  summarise(first_purchase = min(Date_time),
            last_purchase = max(Date_time),
            distinct_trans = n_distinct(Date_time),
            lifetime_net = sum(Net.Sales),
            lifetime_purs = sum(Qty)) %>%
  mutate(tenure = as.Date(last_purchase,  '%Y-%m-%d')  -
as.Date(first_purchase, '%Y-%m-%d'),
         avg_trans = lifetime_net / distinct_trans)

## Warning in as.POSIXlt.POSIXct(x, tz = tz): unknown timezone '%Y-%m-%d'

## Warning in as.POSIXlt.POSIXct(x, tz = tz): unknown timezone '%Y-%m-%d'
```

Initially we will check the share of sales by customers for those customers who have:

1. A customer ID
2. Visited more than once

as compared to the overall sales across 3 years:

```
print(paste('The % share of registered customers that have visited more than
once is: ',
      round(nrow(loyalty[which(loyalty$tenure > 0),]) / nrow(loyalty) * 100,
2)))
```

```
## [1] "The % share of registered customers that have visited more than once
is:  22.28"
```

```
print(paste('The % share of net sales by just 22% of registered customers is:
',
            round(sum(loyalty[which(loyalty$tenure > 0), 'lifetime_net']) /
sum(cp$Net.Sales) * 100, 2)))
```

```
## [1] "The % share of net sales by just 22% of registered customers is:
43.28"
```

About 43% of net sales over 3 years are being driven by just 22% of registered customers who have visited more than once

```
loyalty <- data.frame(loyalty)
print(paste('The % share of total transactions driven by 22% of registered
customers is: ',
            round(nrow(cp[which(cp$Customer.ID %in%
loyalty[which(loyalty$tenure > 0), 'Customer.ID']),]) / nrow(cp) * 100, 2)))
```

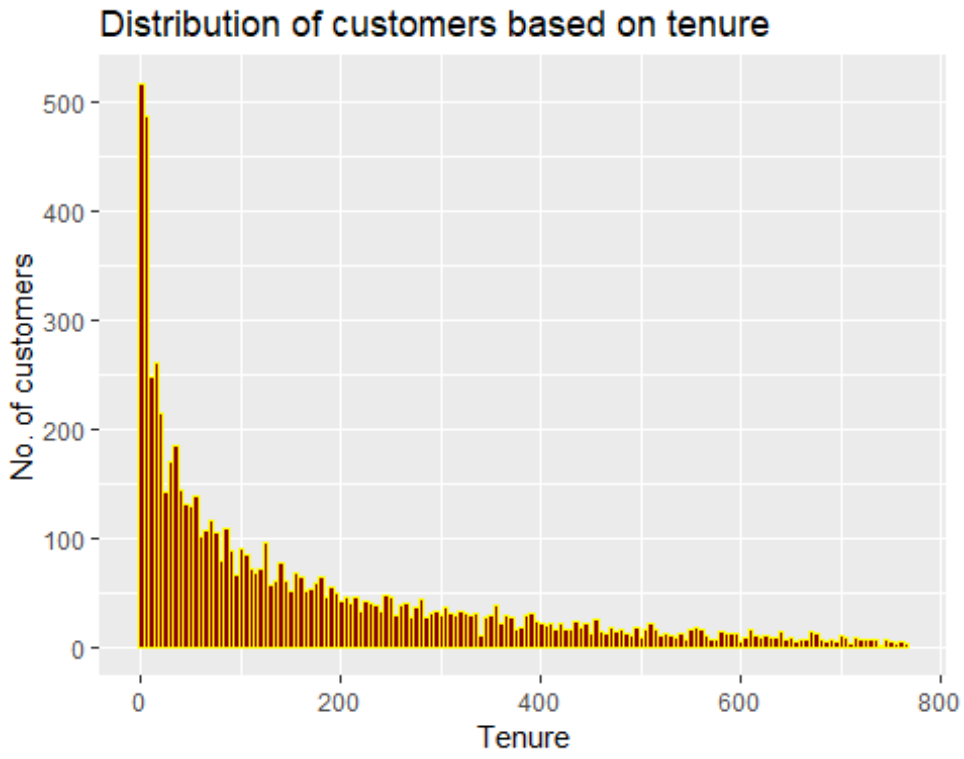
```
## [1] "The % share of total transactions driven by 22% of registered
customers is:  42.75"
```

Hence, we can conclude that a large number of transactions and sales are being generated by a small number of customers by which we can safely assume presence of loyal customers.

Looking at the distribution of tenure of the customers

```
loyalty <- filter(loyalty, tenure > 0)
ggplot(loyalty, aes(x = tenure)) +
  geom_histogram(binwidth = 5, color = 'yellow', fill = 'darkred') +
  ggtitle('Distribution of customers based on tenure') + xlab('Tenure') +
ylab('No. of customers')
```

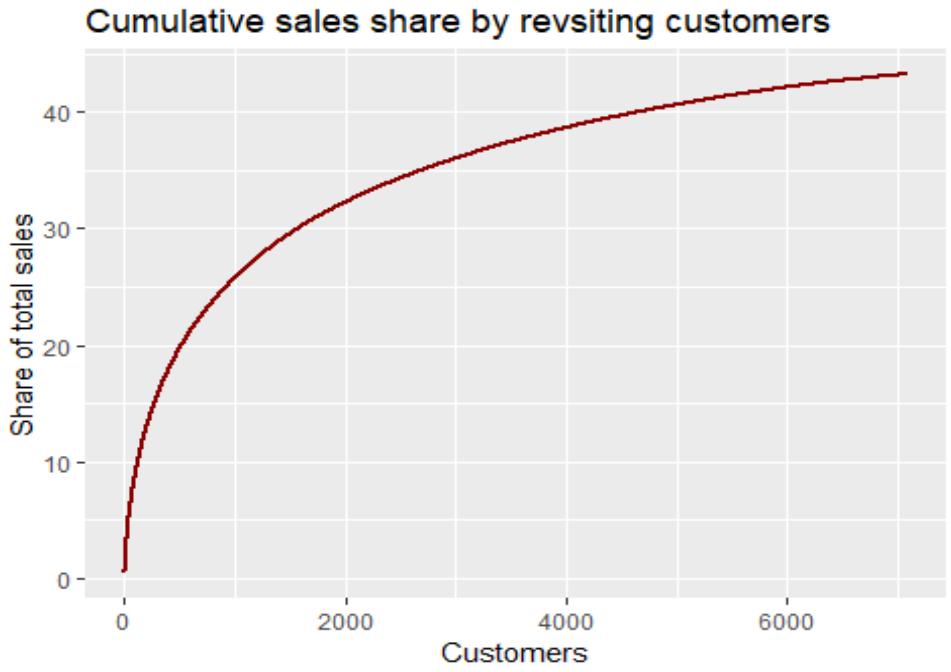
```
## Don't know how to automatically pick scale for object of type difftime.
Defaulting to continuous.
```



We can observe that a significant portion of the customers have tenure greater than about a month (30 days).

Looking at the cumulative sales share based on share of customers visiting more than once:

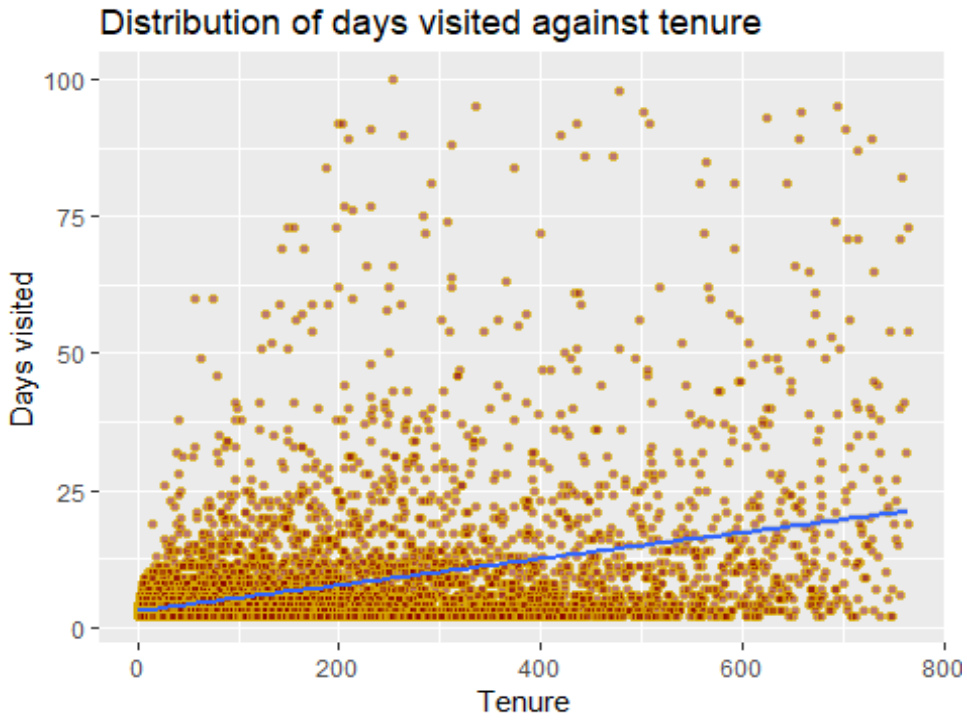
```
loyalty <- loyalty %>% arrange(-lifetime_net) %>%
  mutate(cumshare = 100 * cumsum(lifetime_net) / sum(cp$Net.Sales))

ggplot(loyalty, aes(x = seq(1, nrow(loyalty)), y = cumshare)) +
  geom_line(size = 1, color = 'darkred')+
  ggtitle('Cumulative sales share by revisiting customers') +
xlab('Customers') + ylab('Share of total sales')
```



Hence based on the saturation of the curve we can deduce that Central Perk has a loyal customer base of about 4K customers.

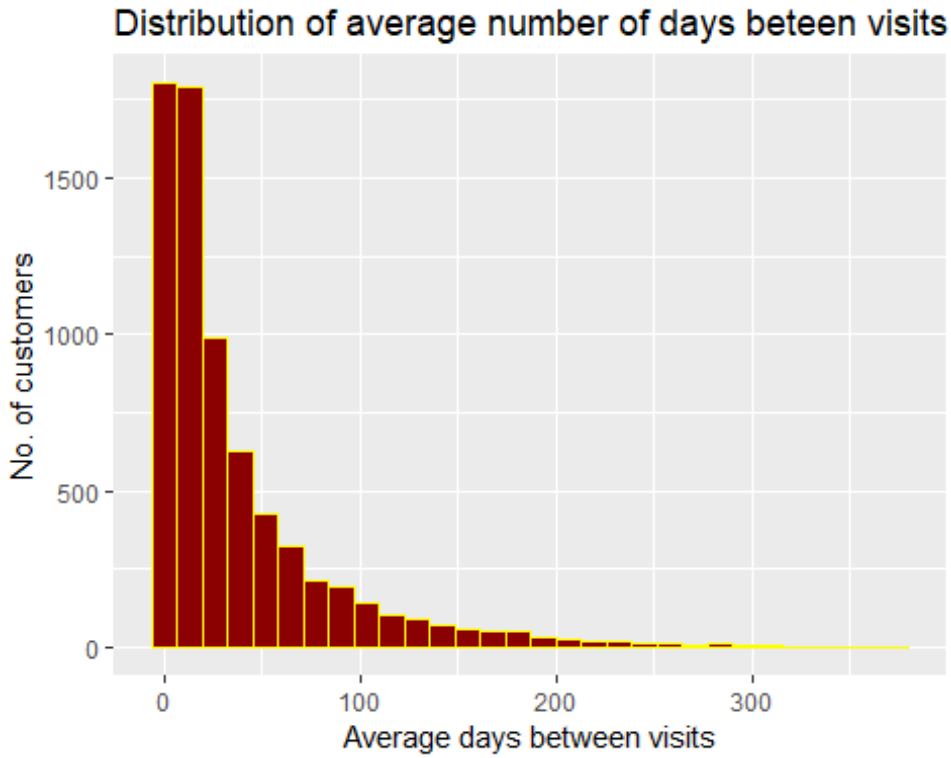
Finally, we would like to see trends in tenure vs transactions.

```
## Don't know how to automatically pick scale for object of type difftime.
Defaulting to continuous.
```



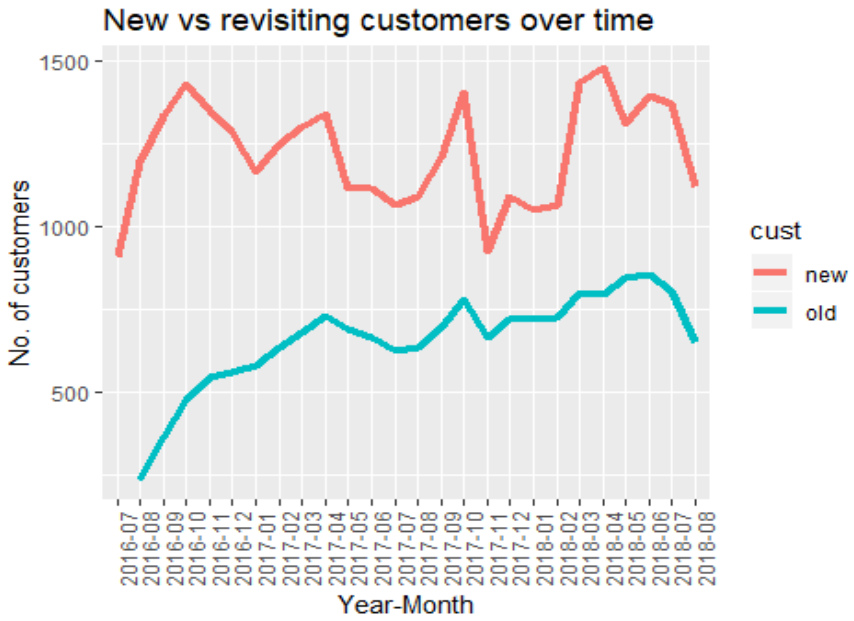
The general pattern of visits against tenure shows increasing pattern which indicates that loyal customers tend to visit repeatedly over time

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From here, it can be concluded that a high number of loyal customers have up to 50 days of gap between subsequent visits. **Hence, we have decent chunk of loyal customers**

Customers can churn over period, but we want to ensure that our customer base consistently has certain portion of loyal customer base retained hence we would like to check overall population of the loyal customers over time, In order to get this metric, we will check number of repeating customers month on month



Here we can ignore the dip in last month since the data is incomplete, but we can observe that overall repeating customers stay constant and even show a slight increasing trend even though the number of new customers fluctuates, hence we can say that Central Perk has a constant moving customer base with loyalty over specific time period keeping the overall number of repeating/loyal customers constant

Hence, we can safely assume that loyalty among customers exists. Further insights about the loyal customers can be found out through clustering

Inference and conclusions:

1. Central Perk has a significant loyal customer base that contributes to about 43% of the overall sales
2. The loyal customer base is moving – keeps changing over the years with customers spanning from 2 months to 8 months of tenure, however overall number of loyal customers stays relatively flat
3. This means that even though some of the older customers are churning out, new customers are being converted to loyal customers in similar rate
4. The customers with high tenure are also having a greater number of visits i.e. it the case of a customer visiting only twice in 2.5 years is less likely

Next approach based on the above conclusions:

1. Since we have established the presence of loyal customers, we would like to understand the attributes that define these customers
2. Being cognizant about characteristics and purchase behavior of the loyal customers will help us in providing recommendations to lift the sales in relevant timeframes

---

Approach 2: Sales Distribution Pattern:

Description and rationale for the analysis:

For the next step, we would like to understand the sales patterns of all the customers at an overall level. This exploratory analysis will help us in uncovering information such as popular products, popular days of visit for various customers and times of the day during which we need to focus on the sales. Since we have established that we have a loyal customer base, this step aims to identify the sales information that will be useful to identify areas of improvement.

Key Goal: Determine the sales patterns and areas of improvement towards sales lift (smoothening)

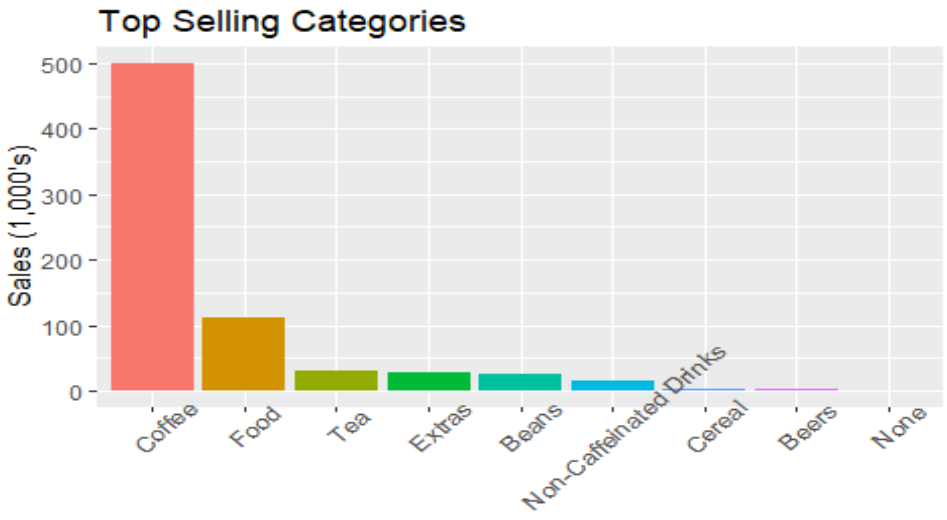
1. Sales by Categories

We first started by identifying the sales of each category in order to analyze their sales patterns. Being a coffee shop, we believe that the top category will be coffee.

Execution and Output:

```
##  # ##
##  <fct> <dbl>
##  1 Coffee 500
##  2 Food 113
##  3 Tea 31.3
```

```
A tibble: 3 x 2
Category gross_sales($thousands)
```



After plotting the categories, we identified the top three categories were: Coffee, Food, and Tea which represent 90% of the total sales ($644K / 717K). Coffee alone contributed 70% of the total sales($500K of the 717K) and this confirms our initial assertions.

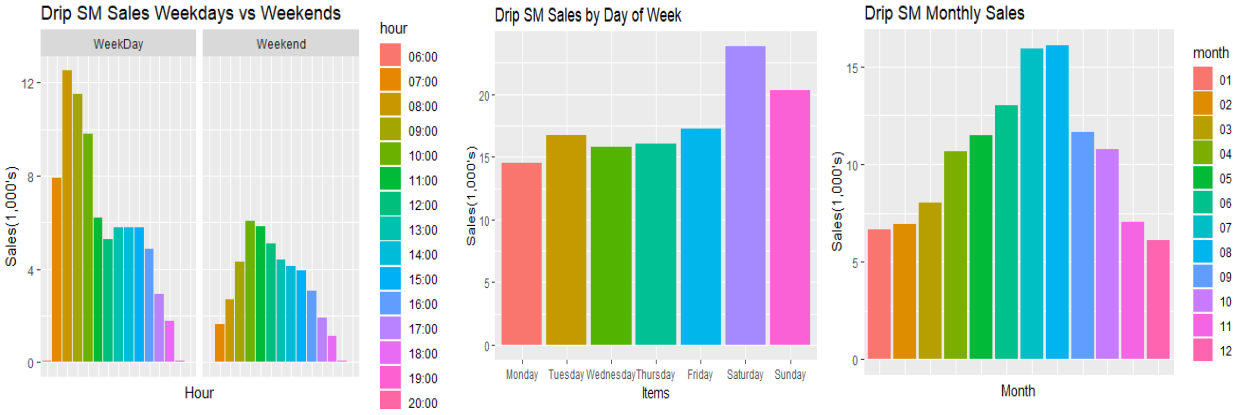
2. Top Selling Items

a. Coffee

After we reviewed the sales of each category, we proceeded to focus our analysis on the top three categories, starting with Coffee.

1. Drip SM

We first began by taking a deeper look into the bestselling food item, Drip SM's sales along different time frames.

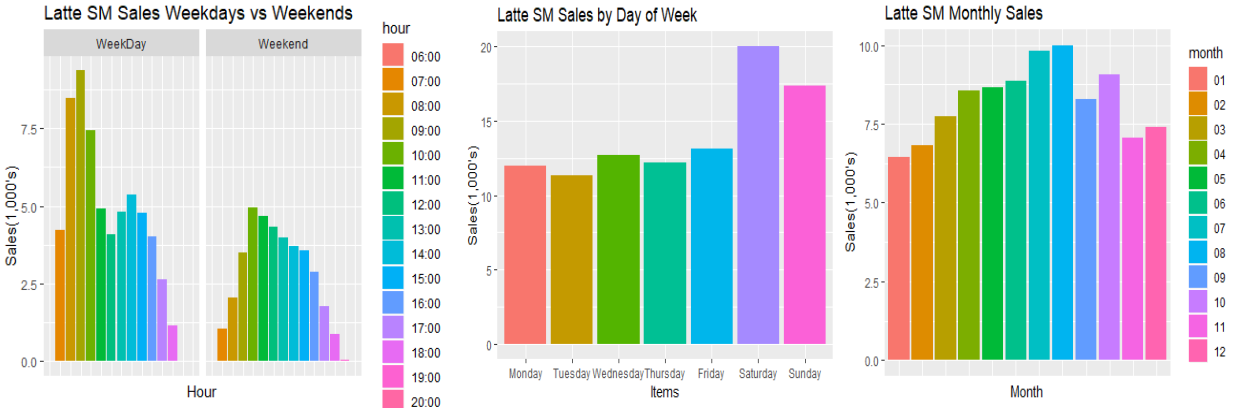


- Drip SM has the most sales at 8:00 am on weekdays i.e. breakfast time, on the other hand, the peak hour is on 10:00 am on the weekends.
- Weekends i.e. Saturday and Sunday appear to have the most sales for Drip SM. The sales on weekdays appear to be relatively similar except for Friday that has a slight bump in sales. They also seem to buy more items on weekends which indicates that they might be spending more time at the coffee shop.
- Drip SM has the most sales in July and August and the least sales in December and January. This implies that this item has more appeal in the warm summer month and the least appeal in the cold wintry months.

In summary, we can assume that customers tend purchase Drip SM on weekdays as they head to work and have more purchases in the summer months. It might indicate that they are people who predominantly walk to work during the summer months and thus more likely to get into the store.

2. Latte SM

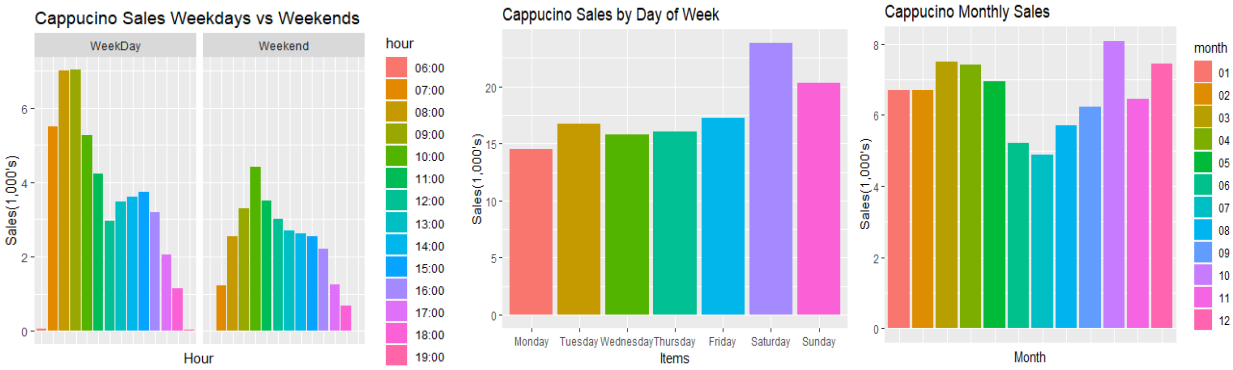
The second top selling item in the coffee category is the Latte SM. We proceeded to assess its sales pattern next based on the different time scales as well.



- Latte SM appears to have the highest sales at 9am during the weekdays and at 10am during the weekends.
- Latte SM appear to have the highest sales on Saturday and Sunday and relatively similar sales for the rest of the week.
- Latte SM have the most sales in July & August while having the least sales in November and January. This implies that the Latte SM have a greater appeal during the warm summer months while having the least appeal during the colder wintry months.

3. Cappuccino

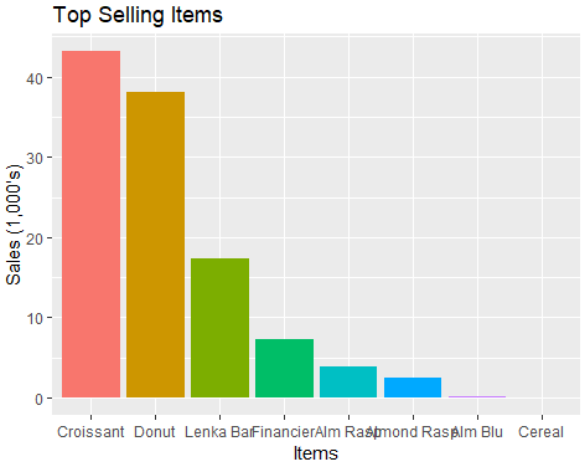
Lastly, we take a look at the third top selling item in coffee category i.e. Cappuccino. We proceeded to assess its sales pattern based along different time scales.



- Cappuccino appears to have the highest sales at 8am and 9am during the weekdays and 10am during the weekends.
- Cappuccino appear to have the highest sales on Saturday and Sunday and relatively similar sales for the rest of the week.
- Cappuccino has the highest sales in October, December, March & April and the lowest sales during the months of June, July, and August. It appears that Cappuccino is mostly bought during the colder months of the year while being preferred the least during the warmer summer months.

b. Food

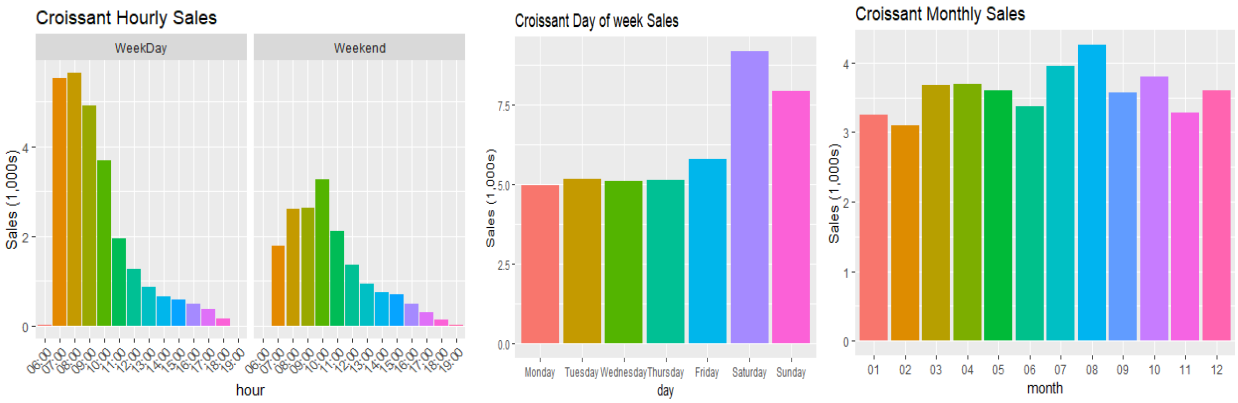
As the second most popular category in the shop, we want to break down sales amount of each product within it.



From the above figure, we can see that the top 3 bestselling items are Croissant, Donut and Lenka bar, which contribute to 87.5% sales of Food category.

1. Croissant

Let us look deeper into the bestselling food item, croissant's sales across different time scales.

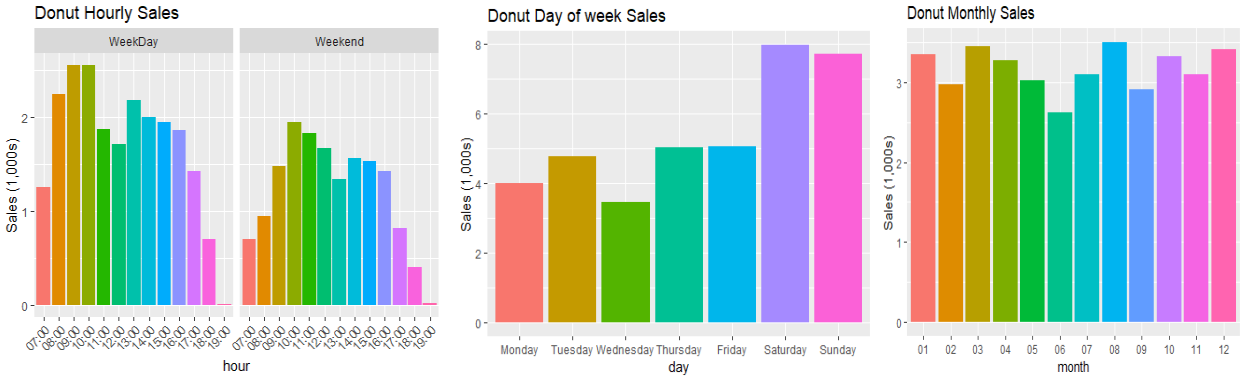


- Croissants were sold most during 8:00 am to 9:00 am on weekdays, which is exactly the breakfast time, on the other side, peak hour is on 10:00 am on weekend. when approaching to lunch time (from 11:00 am), sales of croissant drop substantially.
- Within the entire week, croissants were sold most on the weekend compared to weekdays. Saturday has the highest sales, which is almost 50% more sales than Monday.
- In general, summer(July, August) has comparatively higher sales of croissant than other seasons, and in winter(January, February), less customers purchased croissant.

In summary, we can infer that most customers purchased croissant for breakfast and most of them visit Central Perk on the weekends.

2. Donut

Next, we want to know the second largest sales (10% less than croissant) item, Donut's sales distribution on different time scale for comparison.

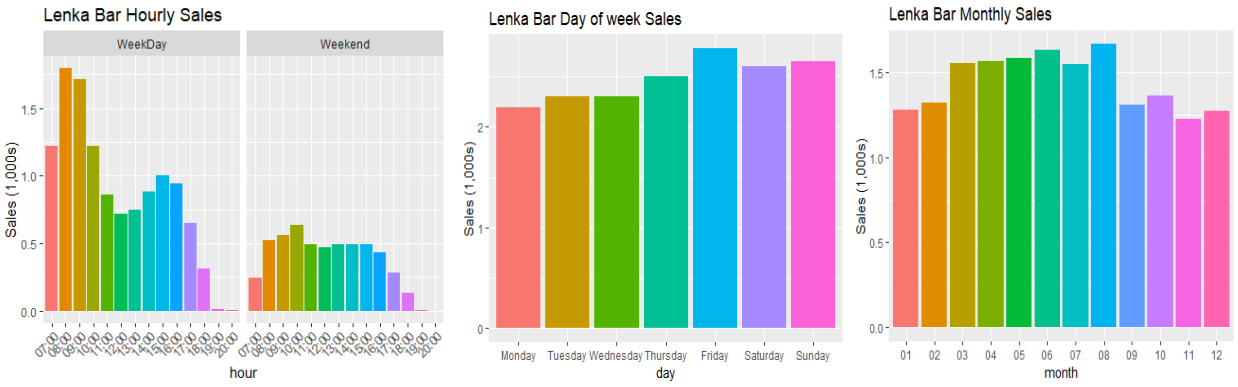


- Similar to croissant, we observe that the highest sales hour of donut weekdays is one hour earlier than on weekend. In contrast, donuts' sales amounts do not significantly drop like croissant. In addition, the demand of donuts maintains at a certain level before dinner time
- Same as croissant, the bestselling day of donuts is on the weekends and the amount sold are centered on Saturday and Sunday
- Unlike croissant's obvious seasonality, donut's sales are distributed evenly at around $3,000 per month.

In brief, we can conclude that the Central Perk's donuts maintain its popularity across the months.

3. Lenka bar

Lastly, although Lenka bar's sales amount lag far behind (less than half of the sales amount croissant and donuts), we think it is probably being driven by a distinct customer segment; therefore, it's worthwhile to take a look.



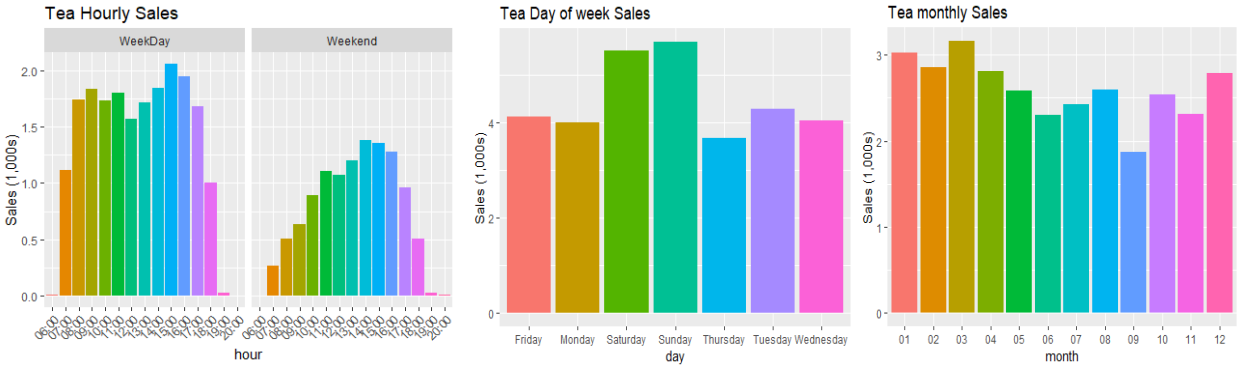
- Again, similar to croissant and donut, highest selling time lags one hour on weekend. Although its demand drops substantially after breakfast, interestingly there is a second peak occurring in the afternoon.
- Unlike croissant and donut, buying pattern across week is not obvious.
- Here we noticed that Lenka bars' sales amounts are about 20% higher in Spring and Summer (March to August). Highest selling month is in August.

As we know, Lenka bar is kind of healthy light food which is convenient to provide certain amounts of calories during meals or even for people who have no time eating breakfast.

For food category, the top selling items show clearly that highest sales hour of weekend would lag 1 to 2 hours compared to weekday. It is reasonable that people tend to wake up late when they have day off.

C. Tea

The proportion of tea in Central Perk's total sales amount is quite low (less than 5%) but considering for those who do not drink coffee, the demand of tea is indispensable.



- In contrast to coffee, tea's highest selling hour is in the afternoon, from 2:00 pm to 3:00 pm
- In general, sales amounts of tea is about 15% higher on weekends than on weekdays.

- We observed a trend that tea's sales gradually increased from winter and achieved its peak in Spring. In summer, tea's sales amount is relatively low.

So far, we did not see different items in tea category, and from the data showing the opposite trend that the demand of tea usually goes up in the afternoon. Perhaps we could have more tea items on the menu for customer who visit in the afternoon. In addition, based on the monthly sales, we infer that probably Central Perk only provide hot tea. If they could add iced tea on the menu, it might attract customers in summer.

Interpretation:

Upon analyzing the seasonal demand patterns, we noticed that the demand varies for almost all the products across hours of the day, week, and months of the year. The demand patterns vary between the items and by the sizes. This shows that there is an uneven distribution of demand and gives us a sense of when it has to be risen to make it smooth.

Inference and Conclusion:

At an overall level for various products and categories, the demand could be increased in the afternoon on weekdays.

Next approach based on the above conclusions:

Now that we know that there is an opportunity to increase the demand to make it smoother, we need to further analyze the customers and their buying patterns in a detailed manner. This will help us tailor the marketing strategies according to the purchase preferences of the customers.

Approach 3: Customer profiling and high priority profile

Description and rationale for the analysis:

We expect that the behavior of loyal customers would be significantly different than the regular customers in terms of their purchase patterns. We not only expect to see a segment of loyal customers but also distinct segments of loyal customers who have different behavioral patterns than other segments. For e.g., a customer who regularly purchases coffee in the morning during weekday and a customer coming on every weekend for a brunch represent loyal customers but with significantly different behavioral patterns. Hence, we will perform a cluster analysis to understand various behavioral patterns of different segments of loyal customers. This will help us in incentivizing them accordingly for improved customer experience and promoting this behavior among other customer segments.

Key Goal: Identify clusters of customers to observe distinction in behavioral patterns.

Execution and Output:

The first step of clustering is to identify the level at which the clusters are to be made and using which attributes. As we are finding the segment of customers, the level is Customer ID.

As a next step, we need to understand which attributes of these customers can be used to segment them. The only data that we have is their purchase data. Hence, it is imperative to extract the features from this data which will reflect the various areas of their purchase patterns.

To segment the customers, there must be an identification for each of them to associate their transactions. This makes the transactions without Customer ID in the data non-usable for the clustering analysis. Hence, the data is filtered for the transactions with Customer ID present.

```
cp <- data_total
cp$Date_time <- as.POSIXct(cp$Date_time, "%Y-%m-%d %H:%M:%S",
tz=Sys.timezone())

cust_attr <- cp[!is.na(cp$Customer.ID),]
```

The feature extraction for the clustering analysis is done as below:

- Extracting hour of the day, day of the week, month and time of the day (Morning, Afternoon & Evening) from the date-time field to understand the seasonality across various levels

```
cust_attr <- cust_attr %>% mutate(month = format(Date_time, '%b'),
                                  day = format(Date_time, '%a'),
                                  hour = format(Date_time, '%H'),
                                  time_of_day = if_else(hour<12, 'Morning',
                                                        if_else(hour<16,
'Afternoon', 'Evening')))
```

- Extracting features related to the customers' purchase behavior such as
    - Average basket size and Unique items – to understand whether the customers buy a variety of products or just loyal to a single product
    - Number of visits, Number of weeks and Number of unique weeks – to determine the loyal customers
    - Min date and Max date – to estimate the total tenure the customer been visiting
    - Total sales , Total quantity, Sales per week and Quantity per week – to identify the customers that bring more value
    - Number of visits per week – to identify the frequent visitors
    - Most purchased category and item – to identify the popular products for various segments of customers
    - Usual time of day, Usual day of week and Usual month- to determine the peak hours, days and months

- Total categories and items – to estimate the potential customers for various products

```
man_mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

cust_attr_1 <- cust_attr%>%group_by(Customer.ID, Date_time)%>%
  summarise(basket_size = sum(Qty),
            sum_sales = sum(Gross.Sales))%>%
  ungroup()%>%
  group_by(Customer.ID)%>%
  summarise(avg_basket = mean(basket_size),
            num_visits = length(Date_time),
            min_date = min(Date_time),
            max_date = max(Date_time),
            total_sales = sum(sum_sales),
            total_qty = sum(basket_size))%>%
  mutate(num_weeks =  if_else(num_visits ==
1,1,as.numeric(difftime(max_date,min_date,units='weeks'))),
         num_weeks = if_else(num_weeks < 1, 1, num_weeks),
         sales_per_week = total_sales/num_weeks,
         qty_per_week = total_qty/num_weeks,
         num_visits_per_week = num_visits/num_weeks)
cust_attr_1 <- cust_attr_1[, -c(4,5,7)]
```

```
cust_attr_2 <- cust_attr %>% group_by(Customer.ID) %>%

summarise(most_pur_cat = man_mode(Category),
most_pur_item = man_mode(Item),
unique_items = n_distinct(Item),
usual_time = man_mode(time_of_day),
usual_dow = man_mode(day),
usual_month = man_mode(month),
total_qty = sum(Qty),
total_cat = n_distinct(Category),
total_item = n_distinct(Item))
```

```
cust_attr_3 <- unique(cust_attr[,c('Customer.ID','Date_time')])%>%
  mutate(year = format(as.Date(Date_time),"%Y"),
         week_num = strftime(as.POSIXct(Date_time),format="%W"),
         year_week = paste0(year,week_num))%>%
  group_by(Customer.ID)%>%
  summarize(num_unique_weeks = n_distinct(year_week))
```

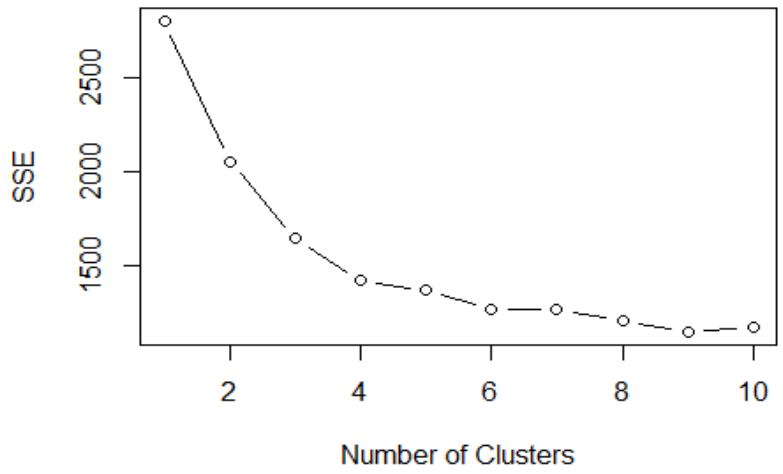
```
merge_1 <- merge(cust_attr_1, cust_attr_2, by='Customer.ID')
cust_attr_final <- merge(merge_1, cust_attr_3, by='Customer.ID')
cust_attr_clust <- cust_attr_final[,-1]
```

Now that we have the attributes of customers, we need to normalize the numeric values and discretize the categorical values to bring them all on the same scale. This makes sure that the attributes are given equal importance when the clusters are being formed using them.

```
normalize <- function(x){return ((x - min(x))/(max(x) - min(x)))}

cust_attr_clust <- cust_attr_clust%>%

mutate(avg_basket = normalize(avg_basket),
num_visits = normalize(num_visits),
total_sales = normalize(total_sales),
num_weeks = normalize(num_weeks),
sales_per_week = normalize(sales_per_week),
qty_per_week = normalize(qty_per_week),
num_visits_per_week = normalize(num_visits_per_week),
unique_items = normalize(unique_items),
total_qty = normalize(total_qty),
total_cat = normalize(total_cat),
total_item = normalize(total_item),
num_unique_weeks = normalize(num_unique_weeks),
most_pur_cat = as.factor(most_pur_cat),
most_pur_item = as.factor(most_pur_item),
usual_time = as.factor(usual_time),
usual_dow = as.factor(usual_dow),
usual_month = as.factor(usual_month))
```

As we have both categorical and numerical attributes, it is recommended to use k-Prototypes clustering technique. The clusters are formed with the values of k ranging from 1 to 10 and the corresponding SSE is calculated to plot the elbow curve.

```
set.seed(13723)
SSE_curve <- c()
for (k in 1:10){
  kpro <- kproto(as.data.frame(cust_attr_clust), k, nstarts = 10000)
  sse <- sum(kpro$withinss)
  SSE_curve[k] <- sse
}
```



From the above graph, we can see that there is a significant decrease in the SSE of clusters till the value of k is 4. Hence, it would be appropriate to create 4 clusters from the data.

```
set.seed(13723)
kpro <- kproto(as.data.frame(cust_attr_clust), 4, nstarts = 10000)

cust_attr_final$cluster <- kpro$cluster
```

Inference and conclusions:

This means that there exist segments of customers who are innately similar in their purchase behavior, i.e., in terms of the attributes that we used for clustering.

Next approach based on the above conclusions:

Now that we have seen that there are natural clusters in the customer base, the next step would be to analyze each of the clusters to arrive at actionable insights.

Customer Profiling

Description and rationale for the analysis:

Analyzing the customer segments to help us understand how they behave; what attributes are common among the customers belonging to a cluster and what attributes separate them from the customers of other segments. This will help us incentivize the customers according to their interests which increases the engagement with the customers.

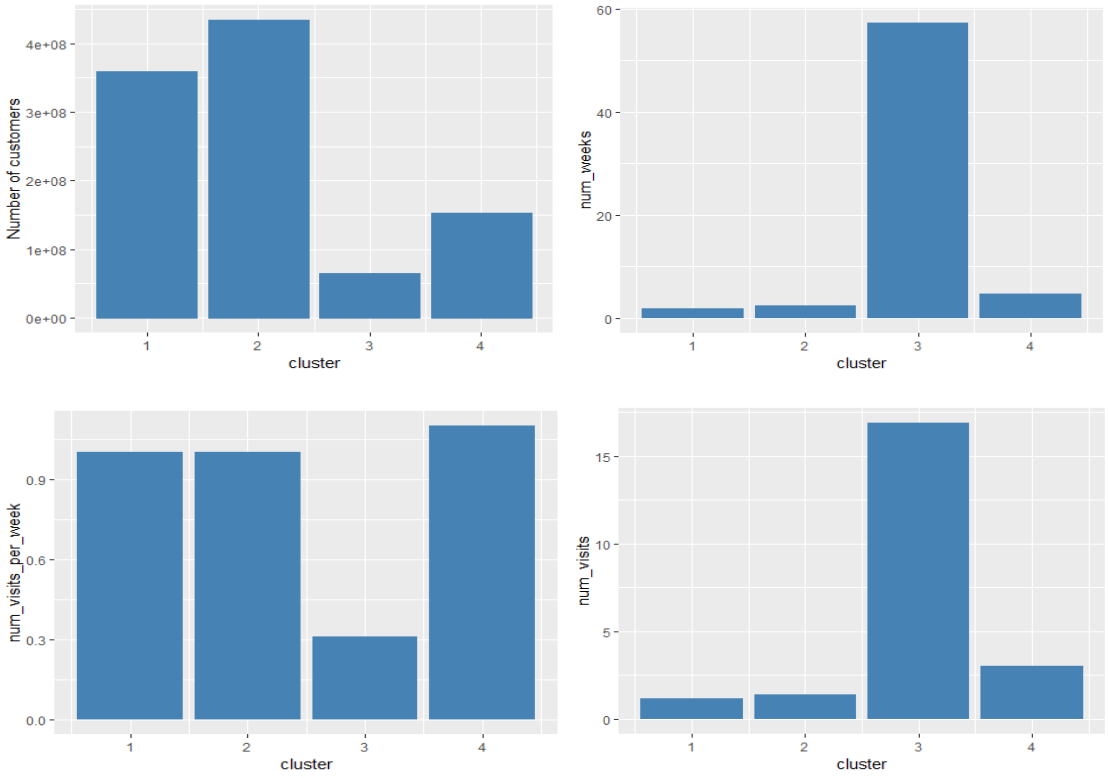
Key Goal: Analyze customers' attributes and get insights about the behavior of customers.

Execution and Output:

Now that we have clusters formed, we need to understand what kind of customers each cluster have. This will let us understand how to treat or target each of these clusters

appropriately. As a first step, we looked at the number of customers in each cluster and their engagement on a weekly basis.

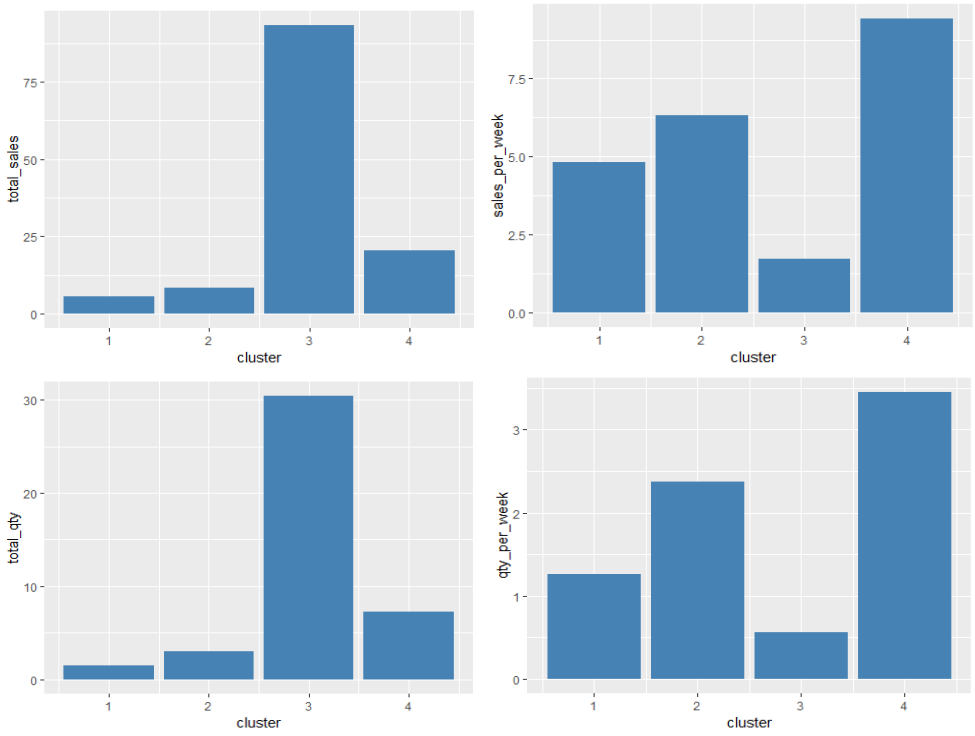
```
library(ggplot2)
ggplot(cust_attr_final, aes(x = cluster, y = n_distinct(Customer.ID)))+
  geom_bar(stat = 'identity', color = 'steelblue')+labs(y = 'Number of
customers')

ggplot(cust_attr_final, aes(x = cluster, y = num_weeks))+
  stat_summary(fun.y = 'mean', geom= 'bar', fill = 'steelblue')

ggplot(cust_attr_final, aes(x = cluster, y = num_visits))+
  stat_summary(fun.y = 'mean', geom = 'bar', fill = 'steelblue')

ggplot(cust_attr_final, aes(x = cluster, y = num_visits_per_week))+
  stat_summary(fun.y = 'mean', geom= 'bar', fill = 'steelblue')
```



We can see that cluster 3 has the least number of customers but the most loyal ones. Their average number of visits and the average number of weeks in the system is so high when compared with other customers. However, their average number of visits per week is lesser. This means that cluster 3 comprises of loyal customers who have been visiting for a long period of time but at a low frequency.

We should understand how much customers belonging to each of these clusters contribute to the overall revenue and sales quantity.

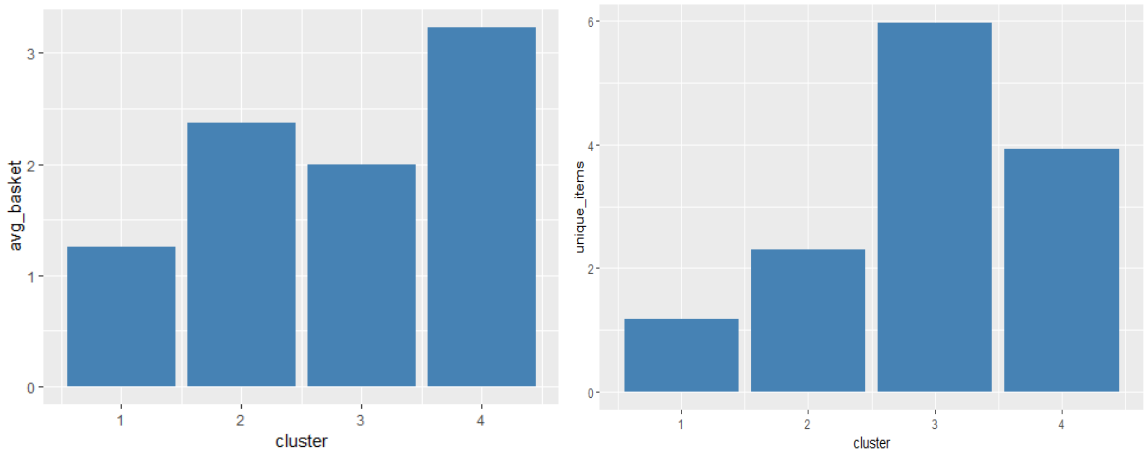
```
ggplot(cust_attr_final, aes(x = cluster, y = total_sales))+
  stat_summary(fun.y = 'mean', geom= 'bar', fill = 'steelblue')

ggplot(cust_attr_final, aes(x = cluster, y = sales_per_week))+
  stat_summary(fun.y = 'mean', geom= 'bar', fill = 'steelblue')

ggplot(cust_attr_final, aes(x = cluster, y = total_qty))+
  stat_summary(fun.y = 'mean', geom= 'bar', fill = 'steelblue')

ggplot(cust_attr_final, aes(x = cluster, y = qty_per_week))+
  stat_summary(fun.y = 'mean', geom= 'bar', fill = 'steelblue')
```



One interesting thing is that even though clusters other than 3 are new customers, cluster 4 seems to be more potential than that of the other clusters. This leaves clusters 3 and 4 to be a priority and need to be targeted according to their interests.

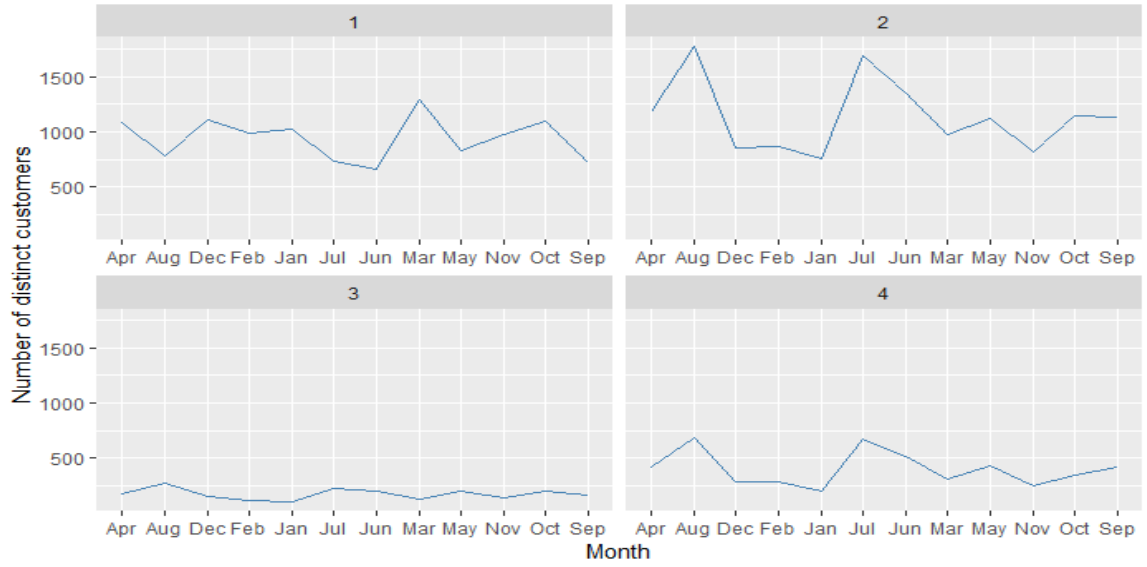
As there is a difference in the behavior of new customers in cluster 4 when compared to those of clusters 1 and 2, we need to further analyze to understand them better. Hence, we looked at the attributes that can uncover the potential of the customers i.e., average basket size and number of unique items bought by customers in this cluster.

```
ggplot(cust_attr_final, aes(x = cluster, y = avg_basket))+
  stat_summary(fun.y = 'mean', geom= 'bar', fill = 'steelblue')
```



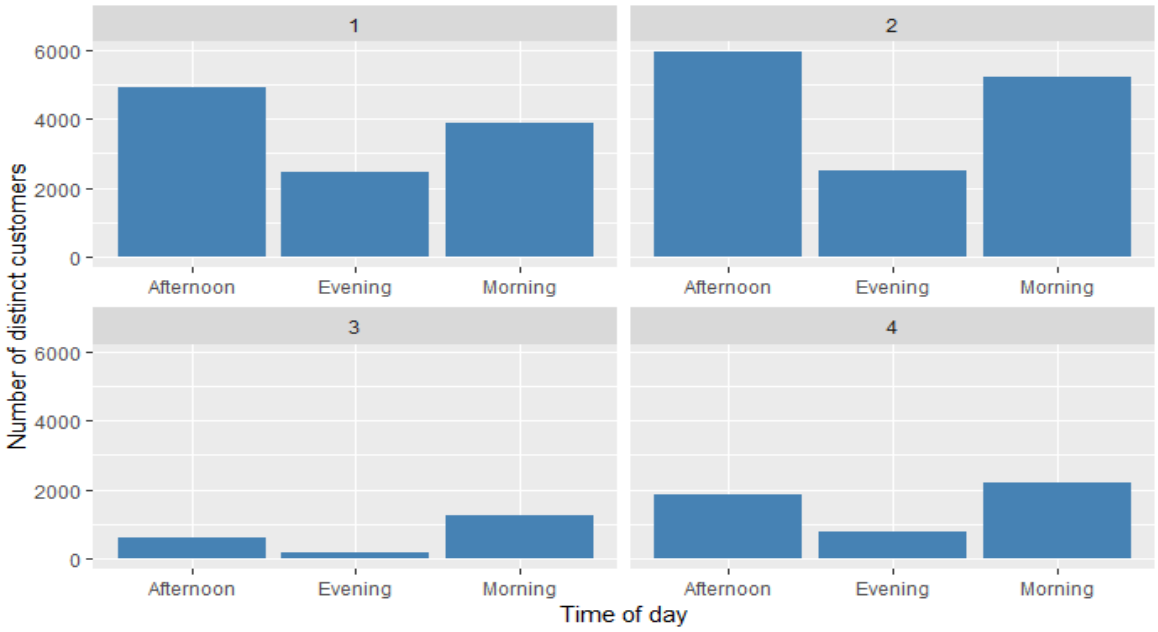
We establish that the cluster 3 has the most loyal customers and cluster 4 has the potential new customers who can be retained to turn into loyal customers if targeted properly. We expect the behavior of cluster 4 to be like cluster 3. Hence, we looked at the pattern of number of customers visiting the store on a given day and across months.

```
month_trend <- cust_attr_final%>%group_by(cluster,
usual_month)%>%summarize(cust_cnt = n_distinct(Customer.ID))
ggplot(month_trend, aes(x = usual_month, y = cust_cnt, group = cluster))+
  geom_line( color = 'steelblue')+
  ylab("Number of distinct customers")+
  xlab("Month")+
  facet_wrap(~cluster, scales = "free_x")
```




```
day_trend <- cust_attr_final%>%group_by(cluster,
usual_time)%>%summarize(cust_cnt = n_distinct(Customer.ID))
ggplot(day_trend, aes(x = usual_time, y = cust_cnt, group = cluster))+
  geom_bar(stat = 'identity', fill = 'steelblue')+
```

```
  ylab("Number of distinct customers")+
  xlab("Time of day")+
  facet_wrap(~cluster, scales = "free_x")
```



We could see that cluster 3 and 4 are behaving similarly when compared to cluster 1 and 2.

Interpretation:

Upon analyzing the clusters, we found that cluster 3 has most of the loyal customers who constitute to only 6.45% of the total customers. However, they contribute to a whopping 73% to the sales amount. Also, the percentage of contribution to the overall quantity is huge. The next potential cluster is cluster 4 which comprises of new customers. But they behave more like the loyal customers in terms of their time of visit, their purchase quantity and diversified products.

Inference and conclusions:

This leads us to establish the most significant customer profiles, which when targeted with the right marketing strategies will bring more revenue. Customers belonging to cluster 3 are to be targeted for increasing the frequency of their visits and customers belonging to cluster 4 are to be targeted to increase their loyalty.

Next approach based on the above conclusions:

Furthermore, we need to understand the product portfolio of these customers to be able to provide targeted recommendations. As we have seen already, many of the customers in general buy coffee most of the times. Hence, it is better to look at the most purchased products by customers at the cluster level. This will help us understand the purchase preferences of different segments of the customers.

Approach 4: Purchase pattern of priority profile

Description and rationale for the analysis:

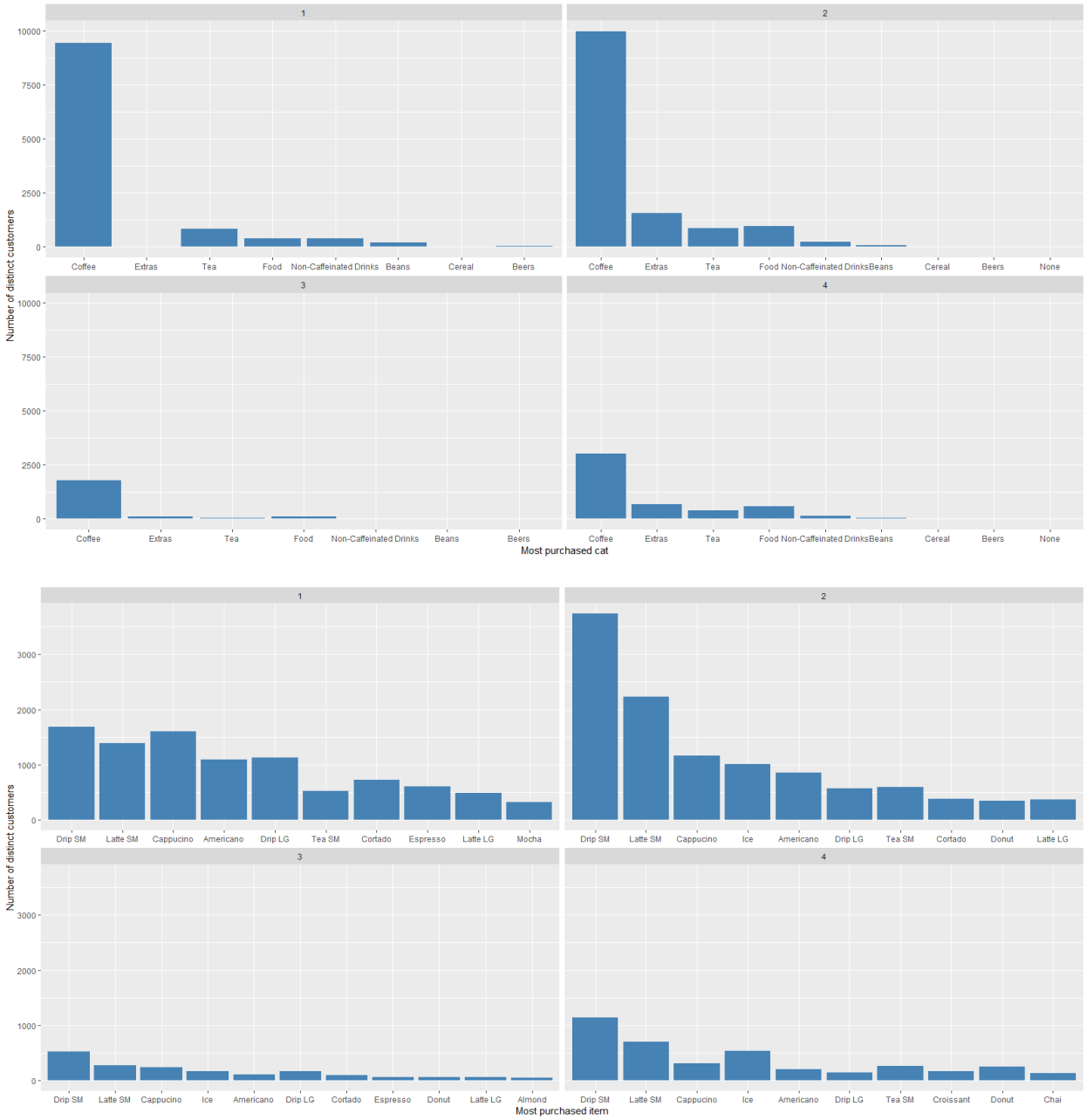
Understanding the market basket of customer profile that contributes to significant sales during the desired time will help us in coming up with bundling strategies that could result in cross-selling and up-selling for rest of the customers. For example, a loyal customer might be a result of him/her purchasing two products that go well together. Hence, we have implemented association rule mining to get an understanding of product combinations preferred by our loyalists who make most of the purchases during the time period when we target to improve sales. This will help us in designing recommendations for new customers towards pulling more crowd and improving customer experience. This will also result in improved sales during low-sales period.

Key Goal: Identify popular product combinations among loyal customers to devise upselling and cross-selling strategies

Execution and Output:

We looked at the categories and items that are mostly bought by the customers in each cluster to understand their preferences. This gives us a sense of the demand of categories and items at a granular level.

```
cat<- cust_attr_final%>%group_by(cluster, most_pur_cat)%>%summarize(cust_cnt
= n_distinct(Customer.ID))
ggplot(cat, aes(x = reorder(most_pur_cat, -cust_cnt), y = cust_cnt, group =
cluster))+
  geom_bar(stat = 'identity', fill = 'steelblue')+
  ylab("Number of distinct customers")+
  xlab("most purchased category")+
  facet_wrap(~cluster, scales = "free_x")
```



From the above figures, we illustrate that the clusters 2 and 4 behave similarly when compared to other clusters. Customers belonging to these clusters purchase products that belong to other categories as well whereas customers from cluster 1 and 3 purchase coffee most of the times. This gives us an opportunity to up sell and cross sell food items along with coffee, which is the main purchased item in the entire menu.

Association Rules

We performed association rule mining on individual clusters to identify the products that are most often purchased together. This gives us a sense co-occurrence that exists in the selling of products.

The results of association rule mining in cluster 2 resulted in:

```
library(tidyr)
library(dplyr)

library(stringr)
library(arules)

data_total <- read.csv("cleaned_data.csv")
cust_attr_final<- read.csv("customer_clusters_4.csv")
data_ar <- merge(data_total, cust_attr_final[,c("Customer.ID","cluster")], on
= "Customer.ID", how = "left")
data_ar <- data_ar[data_ar$cluster==2,]

cust_level <- data_ar%>%filter(!is.na(Customer.ID))%>%
  select(Customer.ID,Date_time,Category, Item, Price.Point.Name)%>%
  group_by(Customer.ID, Date_time)

cust_cat_trans <- unique(cust_level[,c('Customer.ID','Date_time','Item')])

if(sessionInfo()['basePkgs']=="dplyr" | sessionInfo()['otherPkgs']=="dplyr"){
  detach(package:dplyr, unload=TRUE)
}

library(plyr)

cust_cat_trans <- ddply(cust_cat_trans,c("Customer.ID","Date_time"),
                        function(df1)paste(df1$Item,
                                           collapse = ","))

cust_cat_trans$Customer.ID <- NULL
cust_cat_trans$Date_time <- NULL
colnames(cust_cat_trans) <- c("itemList")

write.csv(cust_cat_trans,"ItemList.csv", quote = FALSE, row.names = TRUE)

txn = read.transactions(file="ItemList.csv", rm.duplicates= TRUE,
format="basket",sep=",",cols=1);
txn@itemInfo$labels <- gsub("\"","",txn@itemInfo$labels)

rules <- apriori(txn,parameter = list(sup = 0.01, conf = 0.1))

rules <- sort(rules, by = "lift", decreasing = T)
inspect(rules)
```

```
##                 lhs            rhs    support confidence     lift count
## [1]    {Almond,Ice} =>  {Latte SM} 0.01781713  0.8146341 3.742007   334
## [2]           {Oat} =>  {Latte SM} 0.01061560  0.6297468 2.892731   199
## [3]      {Latte LG} =>    {Almond} 0.01056225  0.2554839 2.591613   198
## [4]           {Soy} =>  {Latte SM} 0.01845727  0.5562701 2.555217   346
## [5]      {Latte SM} =>    {Almond} 0.05393151  0.2477334 2.512993  1011
```

```
## [6]  {Ice,Latte SM} =>    {Almond} 0.01781713  0.1847345 1.873936   334
## [7]       {Drip LG} => {Croissant} 0.01269604  0.1746148 1.865145   238
## [8]    {Croissant} =>   {Drip LG} 0.01269604  0.1356125 1.865145   238
## [9]       {Almond} => {Cappucino} 0.02117785  0.2148268 1.751694   397
## [10]    {Croissant} => {Cappucino} 0.01819055  0.1943020 1.584335   341
## [11]        {Donut} =>   {Drip LG} 0.01050891  0.1116147 1.535092   197
## [22]        {Donut} => {Cappucino} 0.01637683  0.1739377 1.418284   307
```

Similarly, for cluster 3 and 4 the association rules formed are as follows:

Cluster 3 –

```
##             lhs            rhs    support confidence     lift count
## [1]    {Almond} =>  {Latte SM} 0.03234455  0.5213755 3.280013  1122
## [2]    {Almond} => {Cappucino} 0.01144455  0.1844796 1.404920   397
## [3] {Croissant} => {Cappucino} 0.01153103  0.1367989 1.041804   400
```

Cluster 4 –

```
##                  lhs            rhs    support confidence     lift count
## [1]     {Almond,Ice} =>  {Latte SM} 0.01472004  0.7725632 3.632446   214
## [2]         {Almond} =>  {Latte SM} 0.05323979  0.5604634 2.635193   774
## [3]            {Soy} =>  {Latte SM} 0.01891594  0.5018248 2.359486   275
## [4]  {Ice,Lenka Bar} =>   {Drip SM} 0.02166735  0.6589958 2.072800   315
## [5]           {Soy} => {Cappucino} 0.01114321  0.2956204 2.014876   162
## [6]     {Donut,Ice} =>   {Drip SM} 0.03783189  0.6030702 1.896892   550
## [7] {Croissant,Ice} =>   {Drip SM} 0.02730775  0.5846834 1.839058   397
## [8]  {Ice,Latte SM} =>    {Almond} 0.01472004  0.1617536 1.702805   214
## [9]     {Lenka Bar} => {Cappucino} 0.01802174  0.2116317 1.442429   262
## [10]     {Cappucino} => {Lenka Bar} 0.01802174  0.1228317 1.442429   262
## [11] {Croissant,Ice} =>  {Latte SM} 0.01416976  0.3033873 1.426470   206
## [12]        {Almond} => {Cappucino} 0.01864080  0.1962346 1.337487   271
## [13]     {Cappucino} =>    {Almond} 0.01864080  0.1270511 1.337487   271
## [14]     {Lenka Bar} =>    {Almond} 0.01045536  0.1227787 1.292510   152
## [15]        {Almond} => {Lenka Bar} 0.01045536  0.1100652 1.292510   152
## [16]  {Ice,Latte SM} =>     {Donut} 0.01616453  0.1776266 1.284105   235
## [17]   {Drip SM,Ice} =>     {Donut} 0.03783189  0.1718750 1.242525   550
## [18]     {Donut,Ice} =>  {Latte SM} 0.01616453  0.2576754 1.211541   235
## [19]  {Ice,Latte SM} => {Croissant} 0.01416976  0.1557067 1.205359   206
## [20]     {Croissant} =>  {Latte SM} 0.03232907  0.2502662 1.176705   470
```

Interpretation:

From the above rules, we can infer that there are food products that are sold together with coffee. For cluster 3, which is our loyal customer base, we can cross sell Almond and Croissant along with Coffee(Latte and Cappuccino) at a reduced price by bundling them together. Similarly, for cluster 4, which consists of potential new customers who tend to behave more like loyal customers Almond, Soy, Croissant and Donut can be bundled with Latte, Cappuccino and Drip SM and Lenka Bar with Cappuccino. For cluster 2 who are new customers with less visits, Almond, Oat, Croissant and Donut can be bundled with Drip LG and Latte SM.

Inference and conclusions:

Based on the insights found so far, the following recommendations are to be implemented for these clusters to maximize revenue by retaining the new customers and increasing the frequency of visits of old customers.

Cluster 3 (Loyal Customers)

- Bundle Food products such as Almond and Croissant along with Coffee(Latte and Cappuccino) to sell them together at a reduced price
- Offer to be valid in the afternoon and evening on weekdays to smoothen the demand

Cluster 4 (Potential New Customers)

- Bundle Almond, Soy, Croissant and Donut with Latte, Cappuccino and Drip SM and Lenka Bar with Cappuccino for free upgradation of coffee size when the corresponding food items are bought together
- Offer to be valid in the afternoon and evening on weekdays to smoothen the demand

Cluster 2 (One-time New Customers)

- Bundle Almond, Oat, Croissant and Donut with Drip LG and Latte SM for free upgradation of coffee size when corresponding food items are bought together
- Offer to be valid in the evening on weekdays to smoothen the demand

Next approach based on the above conclusions:

Upon identifying the customer profiles and the products in each segment that are sold together most often, we want to analyze for these product categories and items how the price of a product would affect the tenure of the customer and the demand of a product.

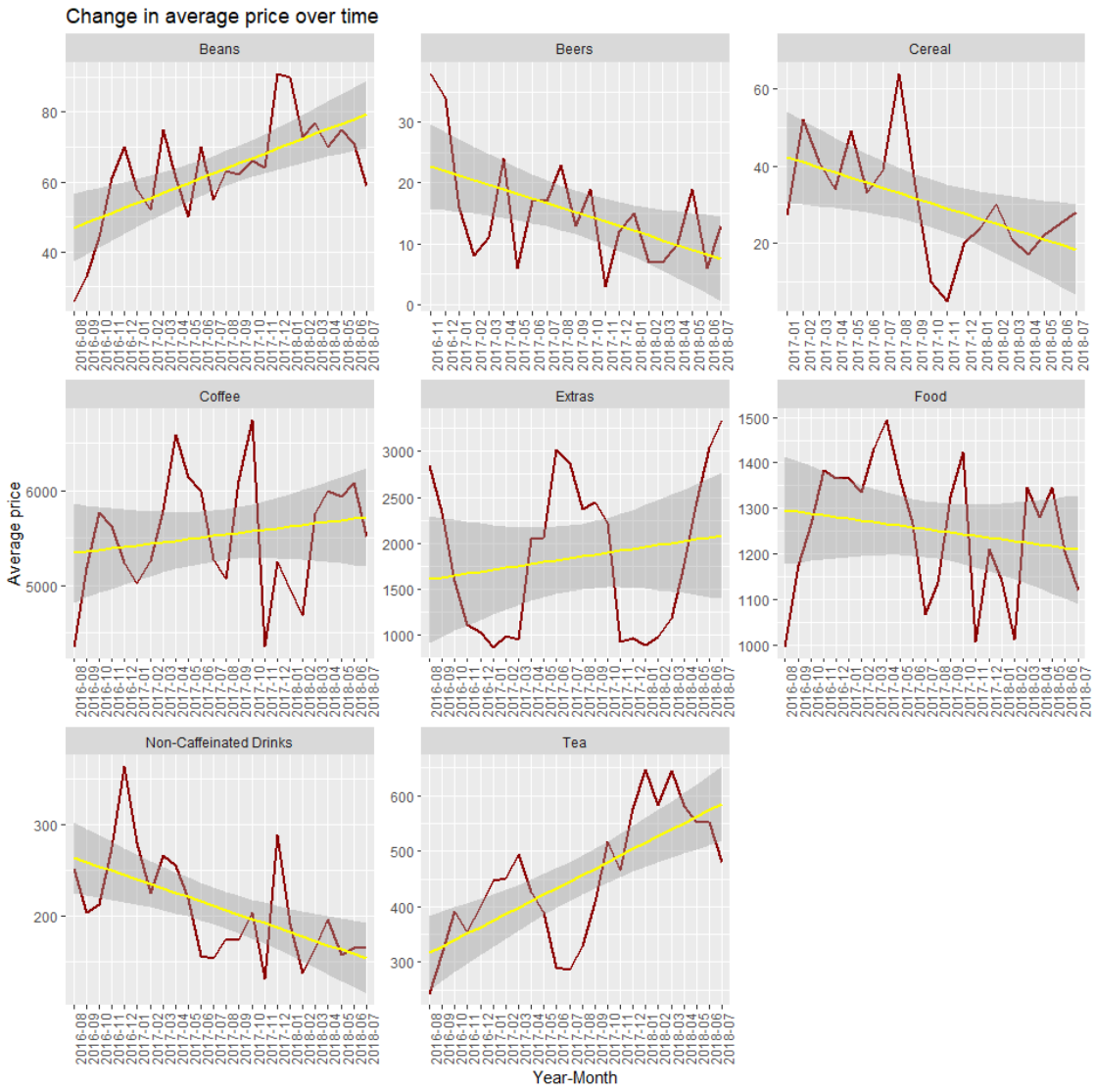
Approach 5: Pricing strategy

Description and rationale for the analysis:

Having identified the purchase behavior we need to extend, getting an estimate regarding the pricing threshold which would boost our sales while keeping the profit margin significant would further improve sales in low sales periods. Hence as the last step, understanding the effects of price fluctuations and other prices on sales would lead us to directional recommendations regarding readjusting product prices.

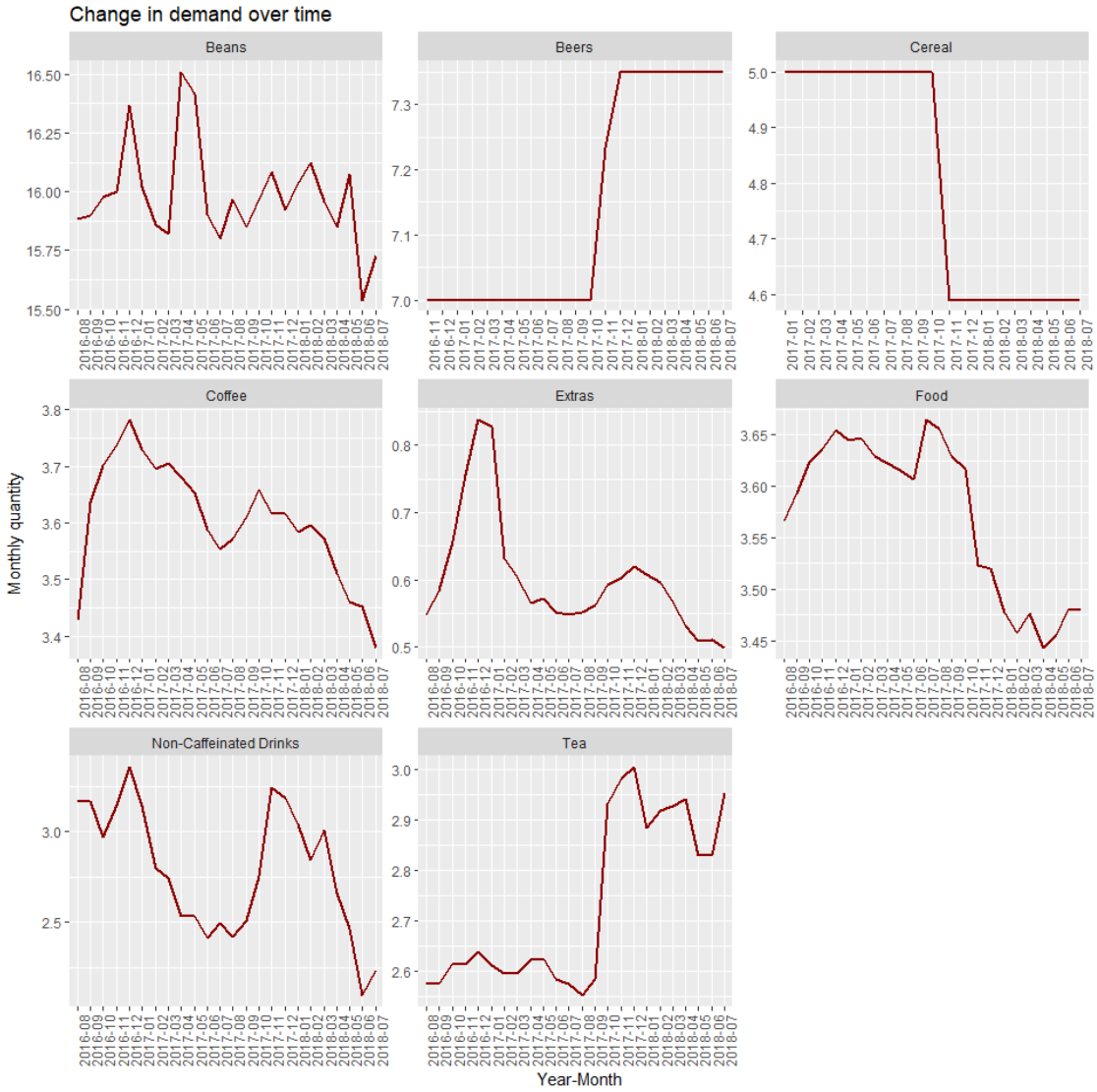
Key Goal: Provide recommendations for price adjustment that will lead to improved sales

Initially we would want to look at how average price per category is changing against demand over time



The demand of Coffee except a few seasonal drops has stayed pretty much constant which is the main product under consideration. The demand of Tea and beans has been increasing over time. Overall, the demand of cereal, beers and non-caffeinated drinks has been dropping

We need to establish if the price of the products have any role to play in these demand changes over time



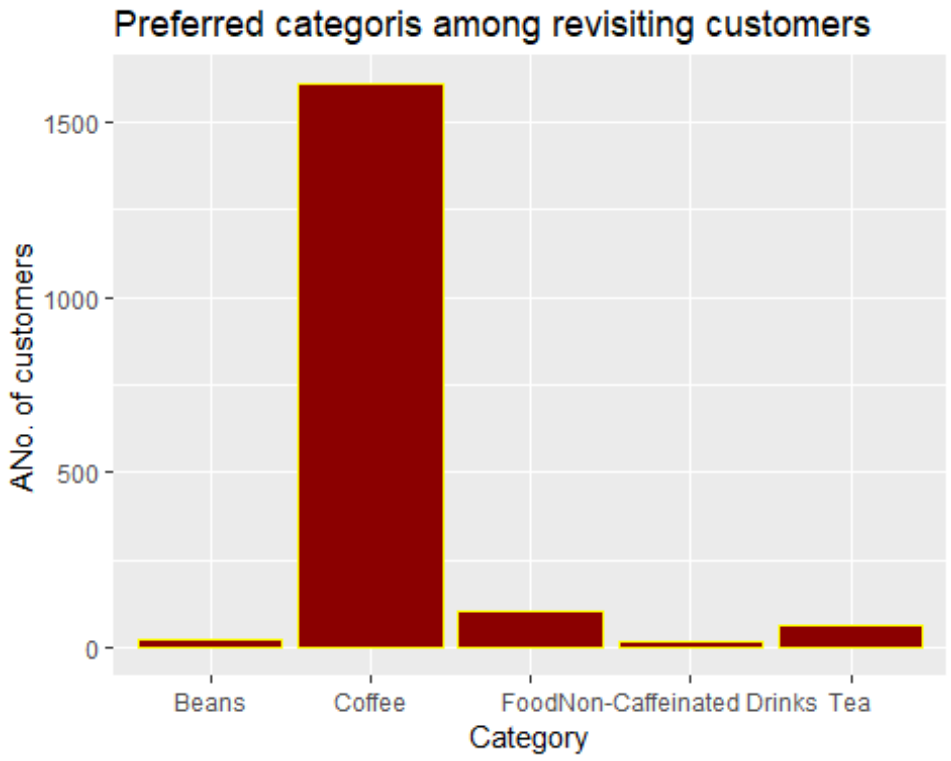
We can observe following patterns in the price changes:

3. Despite increase in tea prices, the demand has been going up
4. The price of coffee has dropped but demand has not shown significant increase
5. Price of food has dropped and so has the demand
6. Price of caffeinated drinks shows an overall drop but demand is also dropping
7. Price of beans as a food item has been fluctuating

Provided the demand rise in tea despite price change, constant demand of key product coffee profitability from beans and to lift demand of cereal and non-caffeinated drink for profitability we need to understand if price change affects the customer tenure negatively. This will help us in retaining the loyal customers and helping more customers turn loyal as they drive maximum revenue.

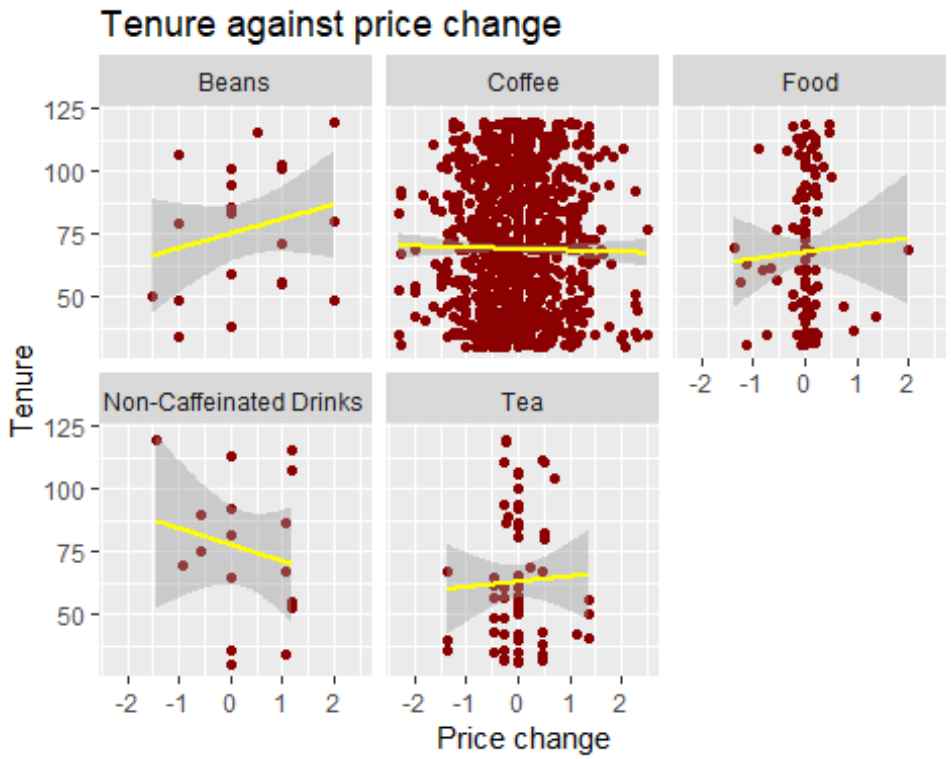
Hence determining relevant customer metrics. We will consider only the customers that have tenure between 1 month to 4-month period to remove extreme loyalists and short tenure people which might skew the results

```
price_analysis <- cp %>%
  filter(!is.na(Customer.ID) & Category != 'None') %>%
  group_by(Customer.ID) %>%
  summarise(first_purchase = min(Date_time),
            last_purchase = max(Date_time),
            distinct_trans = n_distinct(Date_time),
            lifetime_net = sum(Net.Sales),
            lifetime_purs = sum(Qty),
            lifetime_disc = sum(Discounts),
            most_pur_cat = man_mode(Category),
            most_pur_item = man_mode(Item),
            refunds = sum(Event.Type == 'Refund'),
            unique_items = n_distinct(Item)) %>%
  mutate(tenure = as.Date(last_purchase,  '%Y-%m-%d')  -
as.Date(first_purchase, '%Y-%m-%d'),
         avg_trans = lifetime_net / distinct_trans) %>%
  filter(tenure >= 30 & tenure <= 120 & !most_pur_cat %in% c('Beers',
'Cereal', 'Extras'))

## Warning in as.POSIXlt.POSIXct(x, tz = tz): unknown timezone '%Y-%m-%d'

## Warning in as.POSIXlt.POSIXct(x, tz = tz): unknown timezone '%Y-%m-%d'

# We observe that coffee is highest preferred product among these customers
ggplot(price_analysis, aes(x = most_pur_cat)) +
  geom_histogram(stat = 'count', fill = 'darkred', color = 'yellow') +
  ggtitle('Preferred categoris among revisiting customers') +
xlab('Category') + ylab('ANo. of customers')

## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```r
# Getting product prices in different periods
prod_prices <- cp %>% group_by(Category, Date_time) %>% summarise(old_price =
mean(price),
                                                                  new_price =
mean(price))

# Generating price for product when customer first visited and last visited
the store
price_analysis <- merge(price_analysis, prod_prices[,c('Category',
'Date_time', 'old_price')],
                        by.x = c('first_purchase', 'most_pur_cat'), by.y =
c('Date_time', 'Category'), all.x = TRUE)
price_analysis <- merge(price_analysis, prod_prices[,c('Category',
'Date_time', 'new_price')],
                        by.x = c('last_purchase', 'most_pur_cat'), by.y =
c('Date_time', 'Category'), all.x = TRUE)


# Replacing NAs with average price category
avg_lt_price <- cp %>% group_by(Category) %>% summarise(avg_lt = mean(price))
price_analysis <- merge(price_analysis, avg_lt_price, by.x = 'most_pur_cat',
by.y = 'Category')
price_analysis$old_price <- ifelse(is.na(price_analysis$old_price),
price_analysis$avg_lt, price_analysis$old_price)
price_analysis$new_price <- ifelse(is.na(price_analysis$new_price),
price_analysis$avg_lt, price_analysis$new_price)
```

```r
# Calculating price change
price_analysis$price_change <- price_analysis$new_price -
price_analysis$old_price
price_analysis$tenure <- as.numeric(price_analysis$tenure)

ggplot(price_analysis, aes(x = price_change, y = tenure)) +
  geom_point(color = 'darkred') +
  geom_smooth(formula = y ~ x, method = 'lm', color = 'yellow') +
  facet_wrap(~ most_pur_cat)+
  ggtitle('Tenure against price change') + xlab('Price change') +
ylab('Tenure')
```




```r
# Based on the graph, we can conclude that Beers and Cereal could be
disregarded
```

A **mixed model** was implemented to determine the impact of price change based on category. At the same time, we also wanted to observe the effect of lifetime unique items purchased on the tenure of the customer as we expect that the customers trying out more items will develop a better understanding of the menu and tend to visit more due to variety of choices.

```r
price_rec <- lmer(tenure ~ unique_items + (1 + price_change | most_pur_cat),
                  data = price_analysis, control = lmerControl(optimizer
="Nelder_Mead"))
summary(price_rec)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: tenure ~ unique_items + (1 + price_change | most_pur_cat)
##    Data: price_analysis
## Control: lmerControl(optimizer = "Nelder_Mead")
##
## REML criterion at convergence: 17008.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0382 -0.8770 -0.1042  0.7914  2.1049
##
## Random effects:
##  Groups       Name         Variance  Std.Dev. Corr
##  most_pur_cat (Intercept)    0.54179  0.7361
##               price_change   0.01707  0.1307  -1.00
##  Residual                   667.39432 25.8340
## Number of obs: 1821, groups:  most_pur_cat, 5
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   61.2723     1.4827  41.324
## unique_items   2.1223     0.3419   6.207
##
## Correlation of Fixed Effects:
##             (Intr)
## unique_itms -0.815
```

```
coef(price_rec)
```

```
price_analysis$preds <- predict(price_rec, data=price_analysis)
```

```
## Warning in predict.merMod(price_rec, data = price_analysis): unused
## arguments ignored
```

```
price_analysis$ape <- abs(price_analysis$preds - price_analysis$tenure) /
price_analysis$tenure
mean(price_analysis$ape)
```

```
## [1] 0.3866583
```

```
## $most_pur_cat
##                         price_change (Intercept) unique_items
## Beans                    -0.02863816    61.43367     2.122272
## Coffee                   -0.02845079    61.43261     2.122272
## Food                      0.03103127    61.09753     2.122272
## Non-Caffeinated Drinks   -0.02029575    61.38667     2.122272
## Tea                       0.04301460    61.03002     2.122272
##
## attr(,"class")
## [1] "coef.mer"
```

Inference and conclusions:

1. Based on these results we can conclude that despite price changes, demand of tea has been overall rising and small change in price would not affect the tenure of the customers adversely
2. Similarly pertaining to drop in food demand over time, we will be able to generate more revenue if with slight price hike in food items as it will not affect customer loyalty adversely
3. Hence, we recommend introducing small price hike on food and tea products which are being used in bundling to increase revenue
4. However, even though coffee is primary category and beans provides us the most revenue, further price increases in these categories could lead to undesirable effects on customers and might result in churn
5. Most importantly, we observe that for each new item that a customer tries at Central Perk, the tenure increases by two days
6. Hence, we recommend providing discount coupons to similar products in typical purchase category of the customers – This will provide them incentive to try out more products and improve their experience with our store

---

Conclusions & recommendations

1. **Product Bundling:**

Customers belonging to cluster 3 are to be targeted for increasing the frequency of their visits and the rest of the customers especially in cluster 4 are to be targeted to increase their retention and loyalty. We've come up with the following bundling strategies after performing market basket analysis:

- Cluster 3 (Loyal Customers): bundle food products such as Almond and Croissant along with Coffee (Latte and Cappuccino) at discounted price.
- Cluster 4 (New Customers with high potentials): bundle Almond, Soy, Croissant and Donut with Latte, Cappuccino and Drip SM and Lenka Bar with Cappuccino for free upgradation of coffee size.
- Cluster 2 (One-time New Customers): bundle Almond, Oat, Croissant and Donut with Drip LG and Latte SM for free upgradation of coffee size or discount.

2. **Price Adjustment:**

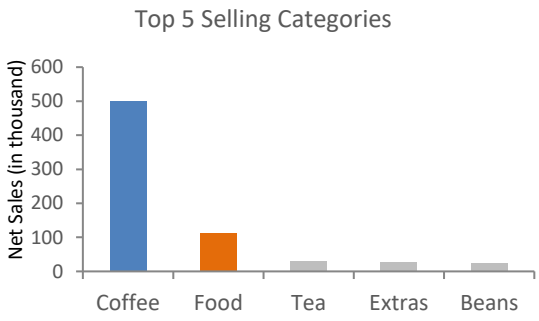
Based on the regression analysis, we can conclude that **price of tea and food category items** can be hiked by a small margin to improve revenue without negatively affecting the customer base since:

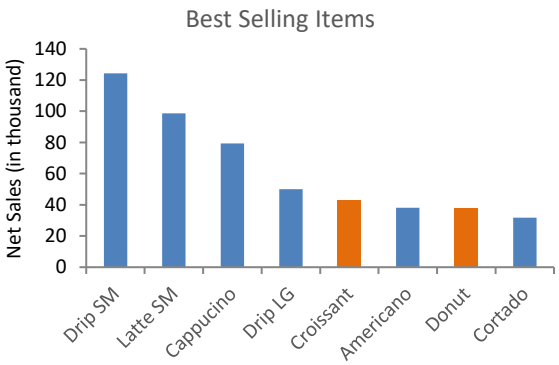
- Demand of tea has increased over time despite price changes
- Price increment in food and tea categories does not negatively affect the customer tenure

- Price of beans, non-caffeinated drinks and coffee categories should not be increased as that will negatively affect the customer tenure
- Increasing price on food category will open avenues to gain more revenue through food items in the afternoon slot to increase revenue during this time interval as food items show higher sales in this duration

3. **Demand Smoothing & Sales Boosting:**

- All the product bundling offers need to be enforced in the non-peak hours (e.g.: the evening of weekdays).
- Introduce pre-order system to keep track of the incoming orders and make preparation in advance to release the burden in peak hour and improve customer in-store experience.
- Provide special offers on food items to loyal customers coming in morning that can be redeemed in afternoon hours.
- Design special menu of the Afternoon Set to increase the demand in non-peak hour. Launch more food especially dessert items to match with tea and coffee as among the top 8 most selling items, only two items (donut, croissant) are food while food category is the second largest revenue generator.





- Give the “early bird” discount on peak hour popular categories to divert the high-volume demand. For example, Central Perk can provide 10% discount on the first 30 customers for lunchbox.
- Central Perk needs to improve the training of cashiers and build up performance measurement standards as we observed that of all the returning transactions made throughout recent three years, over 50% happened in 2018 caused by “accidental charge”, suggesting the incompetence of cashier performance since 2018.
- Regression analysis show s that there’s a positive relationship between the tenure of customer and number of unique items the customer has purchased. We recommend providing product suggestions in the frequently purchased category to revisiting customers which would potentially result in more visits and longer tenure.