Staying Caffeinated: Customer Retention Strategies for a Coffee Store

Introduction

Coffee House is a multinational chain of coffeehouses and roastery reserves headquartered in Seattle, Washington. It is the world's largest coffeehouse chain. We have a large customer base and have consistently experienced high retention rates in the past.

Over the past year, Coffee House noticed a decline in the number of repeat customers and a corresponding decline in revenue. After reviewing customer feedback and conducting surveys, they have determined that the decline in retention rates is likely due to a combination of factors, including a lack of variety in our product offerings, poor customer service, and increased competition from other retailers. They have decided to conduct a data analysis project to better understand the factors contributing to the decline in retention rates and identify potential solutions to improve retention and increase revenue.

Project Objectives

An end-to-end analysis should be conducted to achieve the following objectives:

1. Cleaning and preparing the given dataset for analysis.
2. Explore data to identify trends and patterns.
3. Model data using statistics to identify factors contributing to retention rates.
4. Developing solutions to the given problem statement in the light of your data insights using visualizations.

Methodology

The entirety of this project was done using R. Different variables were compared against the response variable (continue_buying) to identify possible causes of the problem. Chi square test was adopted as the relevant statistical test due to the nature of the dataset, since most variables have been treated as categorical variables.

Data Pre-processing

Load the relevant packages which would be utilized during the course of this analysis.

```
suppressPackageStartupMessages(library(tidyverse))
suppressPackageStartupMessages(library(janitor))
library(ggplot2)
library(forcats)
```

Read the data from excel file into R workspace

```
coffee_house <-  readxl::read_xlsx("Data/coffee-house-satisfactory-survey.xlsx")
```

The structure of the dataset has to be inspected to have an idea of the data we're working with.

```
head(coffee_house)
```

```
## # A tibble: 6 x 21
##   Timestamp      1. Yo~1 2. Yo~2 3. Ar~3 4. Wh~4 5. Ho~5 6. Ho~6 7. Ho~7 8. Th~8
##   <chr>          <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
## 1 2019/10/01 12~ Female  From 2~ Student Less t~ Rarely  Dine in Betwee~ within~
## 2 2019/10/01 12~ Female  From 2~ Student Less t~ Rarely  Take a~ Below ~ 1km - ~
## 3 2019/10/01 12~ Male    From 2~ Employ~ Less t~ Monthly Dine in Betwee~ more t~
## 4 2019/10/01 12~ Female  From 2~ Student Less t~ Rarely  Take a~ Below ~ more t~
## 5 2019/10/01 12~ Male    From 2~ Student Less t~ Monthly Take a~ Betwee~ 1km - ~
## 6 2019/10/01 12~ Female  From 2~ Student Less t~ Rarely  Dine in Betwee~ more t~
## # ... with 12 more variables:
## #   `9. Do you have Coffee House membership card?` <chr>,
## #   `10. What do you most frequently purchase at Coffee House?` <chr>,
## #   `11. On average, how much would you spend at Coffee House per visit?` <chr>,
## #   `12. How would you rate the quality of Coffee House compared to other brands (Coffee Bean, Old T
## #   `13. How would you rate the price range at Coffee House?` <dbl>,
## #   `14. How important are sales and promotions in your purchase decision?` <dbl>, ...
```

```
glimpse(coffee_house)
```

```
## Rows: 122
## Columns: 21
## $ Timestamp
## $ `1. Your Gender`
## $ `2. Your Age`
```

```
## $ `3. Are you currently....?`
## $ `4. What is your annual income?`
## $ `5. How often do you visit Coffee House?`
## $ `6. How do you usually enjoy Coffee House?`
## $ `7. How much time do you normally  spend during your visit?`
## $ `8. The nearest Coffee House's outlet to you is...?`
## $ `9. Do you have Coffee House membership card?`
## $ `10. What do you most frequently purchase at Coffee House?`
## $ `11. On average, how much would you spend at Coffee House per visit?`
## $ `12. How would you rate the quality of Coffee House compared to other brands (Coffee Bean, Old Tow
## $ `13. How would you rate the price range at Coffee House?`
## $ `14. How important are sales and promotions in your purchase decision?`
## $ `15. How would you rate the ambiance at Coffee House? (lighting, music, etc...)`
## $ `16. You rate the WiFi quality at Coffee House as..`
## $ `17. How would you rate the service at Coffee House? (Promptness, friendliness, etc..)`
## $ `18. How likely you will choose Coffee House for doing business meetings or hangout with friends?`
## $ `19. How do you come to hear of promotions at Coffee House? Check all that apply.`
## $ `20. Will you continue buying Coffee House?`
```

A lot of these columns have very long names and it would be a better approach to shorten the lengths of the titles of each column to present for better viewing.

```
colnames(coffee_house)[1:5] <- c("time", "gender", "age_range",
                                 "employment_status", "annual_income")
colnames(coffee_house)[6:15] <- c("visit", "order_preference", "spent_time",
                                  "branch_distance", "membership", "frequent_purchase",
                                  "avg_spend", "coffee_house_rating",
                                  "price_range_rating", "promo_importance")
colnames(coffee_house)[16:21] <- c("ambiance_rating", "wifi_rating", "overall_service_rating",
                                   "future_business_likelihood", "hear_promo",
                                   "continue_buying")
```

The column names have been shortened now using this approach, and the dataset can be inspected to ensure that the changes have been effected.

```
## Rows: 122
## Columns: 21
## $ time                      <chr> "2019/10/01 12:38:43 PM GMT+8", "2019/10/01~
## $ gender                    <chr> "Female", "Female", "Male", "Female", "Male~
## $ age_range                 <chr> "From 20 to 29", "From 20 to 29", "From 20 ~
## $ employment_status         <chr> "Student", "Student", "Employed", "Student"~
## $ annual_income             <chr> "Less than RM25,000", "Less than RM25,000",~
## $ visit                     <chr> "Rarely", "Rarely", "Monthly", "Rarely", "M~
## $ order_preference          <chr> "Dine in", "Take away", "Dine in", "Take aw~
## $ spent_time                <chr> "Between 30 minutes to 1 hour", "Below 30 m~
## $ branch_distance           <chr> "within 1km", "1km - 3km", "more than 3km",~
## $ membership                <chr> "Yes", "Yes", "Yes", "No", "No", "No", "Yes~
## $ frequent_purchase         <chr> "Coffee", "Cold drinks;Pastries", "Coffee",~
## $ avg_spend                 <chr> "Less than RM20", "Less than RM20", "Less t~
## $ coffee_house_rating       <dbl> 4, 4, 4, 2, 3, 4, 5, 4, 5, 4, 4, 3, 4, 4, 5~
## $ price_range_rating        <dbl> 3, 3, 3, 1, 3, 3, 5, 2, 4, 3, 1, 2, 3, 3, 2~
## $ promo_importance          <dbl> 5, 4, 4, 4, 4, 5, 5, 3, 4, 3, 4, 4, 2, 4, 5~
## $ ambiance_rating           <dbl> 5, 4, 4, 3, 2, 5, 5, 3, 4, 4, 5, 4, 4, 4, 5~
## $ wifi_rating               <dbl> 4, 4, 4, 3, 2, 4, 3, 3, 4, 3, 3, 3, 4, 4, 5~
## $ overall_service_rating    <dbl> 4, 5, 4, 3, 3, 5, 5, 3, 4, 3, 3, 4, 3, 4, 5~
```

```
## $ future_business_likelihood <dbl> 3, 2, 3, 3, 3, 4, 5, 3, 4, 4, 4, 4, 4, 3, 2~
## $ hear_promo                 <chr> "Starbucks Website/Apps;Social Media;Emails~
## $ continue_buying            <chr> "Yes", "Yes", "Yes", "No", "Yes", "Yes", "Y~
```

Noticeably, most variables are either in ***Character*** format or ***Double***, which will not be particularly useful for my analysis, so it becomes imperative to convert the class of these variables to ***Factor***

```
factor_cols <-coffee_house[, 2:19]
coffee_house_1 <- lapply(factor_cols, as.factor)
coffee_house_1$frequent_purchase = as.character(coffee_house_1$frequent_purchase)
coffee_house_1 <- data.frame(coffee_house_1)
coffee_house[, 2:19] <- coffee_house_1
coffee_house$continue_buying = as.factor(coffee_house$continue_buying)
```

The factors of some variables need to be reordered to make the data appear cleaner on plots. The `Forcats` function is used for varying levels of modifications to factors.

```
# change the levels of age_range factor
unique(coffee_house$age_range)
```

```
## [1] From 20 to 29 From 30 to 39 40 and above  Below 20
## Levels: 40 and above Below 20 From 20 to 29 From 30 to 39
```

```
coffee_house$age_range <- fct_recode(
  coffee_house$age_range, "<20" = "Below 20",
  "20-29" = "From 20 to 29", "30 -39" = "From 30 to 39", ">40" = "40 and above"
)
levels(coffee_house$age_range)
```

```
## [1] ">40"    "<20"    "20-29"  "30 -39"
```

```
#change the levels of annual_income factor
unique(coffee_house$annual_income)
```

```
## [1] Less than RM25,000    RM50,000 - RM100,000  RM25,000 - RM50,000
## [4] RM100,000 - RM150,000 More than RM150,000
## 5 Levels: Less than RM25,000 More than RM150,000 ... RM50,000 - RM100,000
```

```
coffee_house$annual_income <- fct_recode(
  coffee_house$annual_income, "<25,000" = "Less than RM25,000", ">150,000" = "More than RM150,000", "10
    "RM100,000 - RM150,000", "25,000-50,000" = "RM25,000 - RM50,000", "50,000 - 100,000" = "RM50,000 - 
)
levels(coffee_house$annual_income)
```

```
## [1] "<25,000"           ">150,000"          "100,000 - 150,000"
## [4] "25,000-50,000"     "50,000 - 100,000"
```

```
#change the levels of branch_distance
unique(coffee_house$branch_distance)
```

```
## [1] within 1km    1km - 3km     more than 3km
## Levels: 1km - 3km more than 3km within 1km
```

```
coffee_house$branch_distance <- fct_recode(coffee_house$branch_distance,
  "< 1km" = "within 1km", "> 3km" = "more than 3km"
)
levels(coffee_house$branch_distance)
```

```
## [1] "1km - 3km" "> 3km"     "< 1km"
```

```r
# change the levels of avg_spend
unique(coffee_house$avg_spend)
```

```
## [1] Less than RM20     Around RM20 - RM40 More than RM40     Zero
## Levels: Around RM20 - RM40 Less than RM20 More than RM40 Zero
```

```r
coffee_house$avg_spend <- fct_recode(coffee_house$avg_spend,
   "<20" = "Less than RM20", "20 - 40" = "Around RM20 - RM40",  ">40" = "More than RM40",  "0" = "Zero"
)
levels(coffee_house$avg_spend)
```

```
## [1] "20 - 40" "<20"     ">40"     "0"
```

```r
#change levels of spent_time
unique(coffee_house$spent_time)
```

```
## [1] Between 30 minutes to 1 hour Below 30 minutes
## [3] More than 3 hours            Between 1 hour to 2 hours
## [5] Between 2 hours to 3 hours
## 5 Levels: Below 30 minutes ... More than 3 hours
```

```r
coffee_house$spent_time <- fct_recode(coffee_house$spent_time,
"30 mins - 1 hour" = "Between 30 minutes to 1 hour", "< 30mins" = "Below 30 minutes", "> 3hours" = "Mor
"1 - 2 hours" = "Between 1 hour to 2 hours", " 2 - 3 hours" = "Between 2 hours to 3 hours"
)
levels(coffee_house$spent_time)
```

```
## [1] "< 30mins"         "1 - 2 hours"      " 2 - 3 hours"     "30 mins - 1 hour"
## [5] "> 3hours"
```

The time column has to be converted to the right format.

```r
# convert the time column to date format
coffee_house$time = as.Date(coffee_house$time)
```

Finally, the data has to be inspected for missing rows and duplicate values.

```r
# check for rows with missing values
coffee_house[!complete.cases(coffee_house), ]
```

```
## # A tibble: 0 x 21
## # ... with 21 variables: time <date>, gender <fct>, age_range <fct>,
## #   employment_status <fct>, annual_income <fct>, visit <fct>,
## #   order_preference <fct>, spent_time <fct>, branch_distance <fct>,
## #   membership <fct>, frequent_purchase <chr>, avg_spend <fct>,
## #   coffee_house_rating <fct>, price_range_rating <fct>,
## #   promo_importance <fct>, ambiance_rating <fct>, wifi_rating <fct>,
## #   overall_service_rating <fct>, future_business_likelihood <fct>, ...
```

```r
# no missing data

#check for duplicate data
sum(duplicated(coffee_house))
```

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

```r
# no duplicate data, analysis can now begin
```

After all these have been modifications done, the dataset has to be inspected for correctness or any lagging issues yet to be corrected

```
head(coffee_house, 15)
```

```
## # A tibble: 15 x 21
##    time       gender age_range employmen~1 annua~2 visit order~3 spent~4 branc~5
##    <date>     <fct>  <fct>     <fct>       <fct>   <fct> <fct>   <fct>   <fct>
##  1 2019-10-01 Female 20-29     Student     <25,000 Rare~ Dine in 30 min~ < 1km
##  2 2019-10-01 Female 20-29     Student     <25,000 Rare~ Take a~ < 30mi~ 1km - ~
##  3 2019-10-01 Male   20-29     Employed    <25,000 Mont~ Dine in 30 min~ > 3km
##  4 2019-10-01 Female 20-29     Student     <25,000 Rare~ Take a~ < 30mi~ > 3km
##  5 2019-10-01 Male   20-29     Student     <25,000 Mont~ Take a~ 30 min~ 1km - ~
##  6 2019-10-01 Female 20-29     Student     <25,000 Rare~ Dine in 30 min~ > 3km
##  7 2019-10-01 Female 20-29     Student     <25,000 Rare~ Dine in < 30mi~ < 1km
##  8 2019-10-01 Male   20-29     Employed    50,000~ Rare~ Dine in 30 min~ > 3km
##  9 2019-10-01 Female 20-29     Student     <25,000 Rare~ Drive-~ < 30mi~ > 3km
## 10 2019-10-01 Male   20-29     Employed    <25,000 Mont~ Take a~ < 30mi~ > 3km
## 11 2019-10-01 Female 20-29     Student     <25,000 Rare~ Dine in < 30mi~ > 3km
## 12 2019-10-01 Female 20-29     Student     <25,000 Rare~ Dine in 30 min~ > 3km
## 13 2019-10-01 Female 20-29     Student     <25,000 Week~ Take a~ < 30mi~ 1km - ~
## 14 2019-10-01 Female 20-29     Student     <25,000 Rare~ Take a~ < 30mi~ 1km - ~
## 15 2019-10-01 Female 20-29     Student     <25,000 Rare~ Take a~ < 30mi~ < 1km
## # ... with 12 more variables: membership <fct>, frequent_purchase <chr>,
## #   avg_spend <fct>, coffee_house_rating <fct>, price_range_rating <fct>,
## #   promo_importance <fct>, ambiance_rating <fct>, wifi_rating <fct>,
## #   overall_service_rating <fct>, future_business_likelihood <fct>,
## #   hear_promo <chr>, continue_buying <fct>, and abbreviated variable names
## #   1: employment_status, 2: annual_income, 3: order_preference, 4: spent_time,
## #   5: branch_distance
```

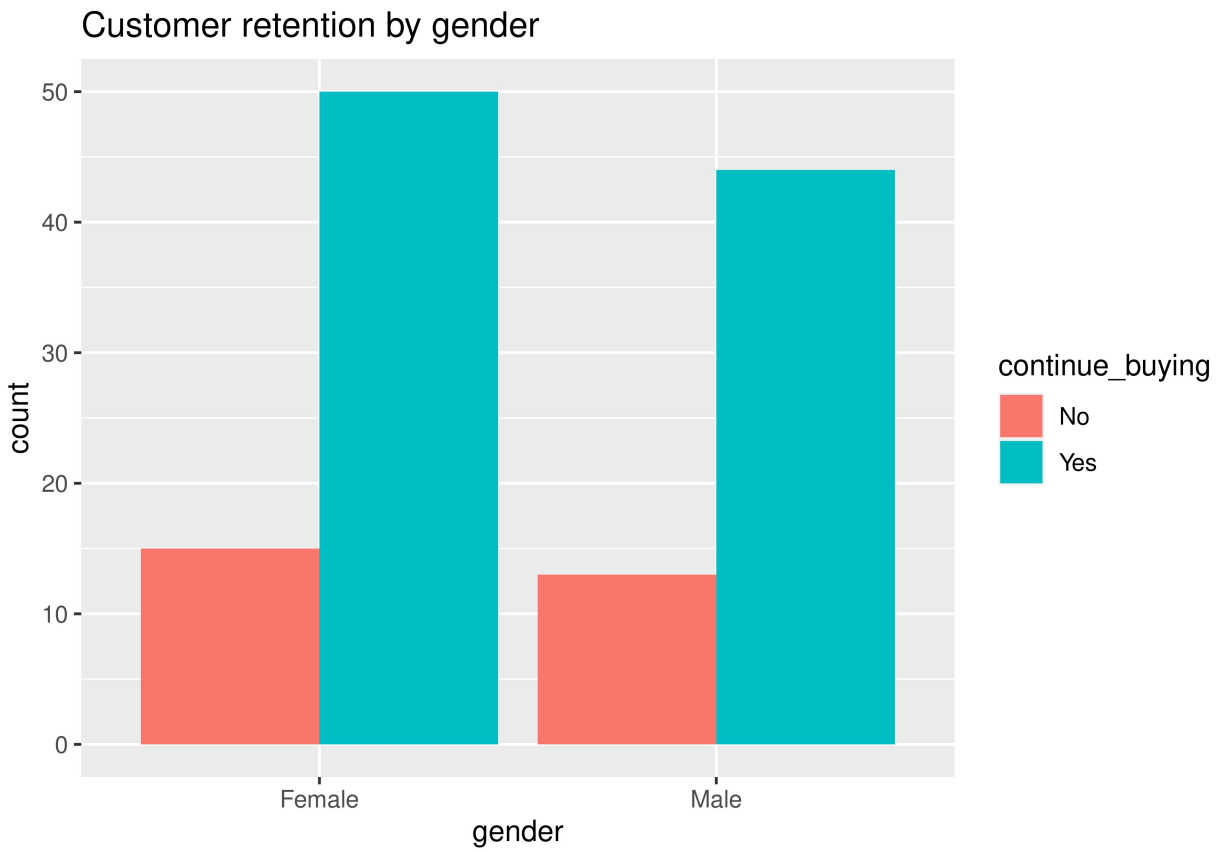
Looks like the dataset has shaped up nicely and it is **finally** ready for inspection and analysis!

Exploratory Data Analysis

In this section, the objective is clear- to find out what variables are directly related to customer retention i.e the customer_buying column.
The first analysis to be done is to test if the gender of customers is a factor to be considered when analysisng retention.

```
gender_eff <- coffee_house %>%
  select(gender, continue_buying) %>%
  group_by(gender) %>%
  table()
```

```
# plot bar graph of gender based on customer retention
ggplot(coffee_house, aes(fill =continue_buying, x = gender)) + geom_bar(position = "dodge") +
  labs(title = "Customer retention by gender")
```



From the plot, we can infer that; 1. there are more female customers 2. a slightly higher number of female customers do not want to purchase anything from the company again However, do these figures have any statistical significance? A chi square test can be used to test this hypothesis.

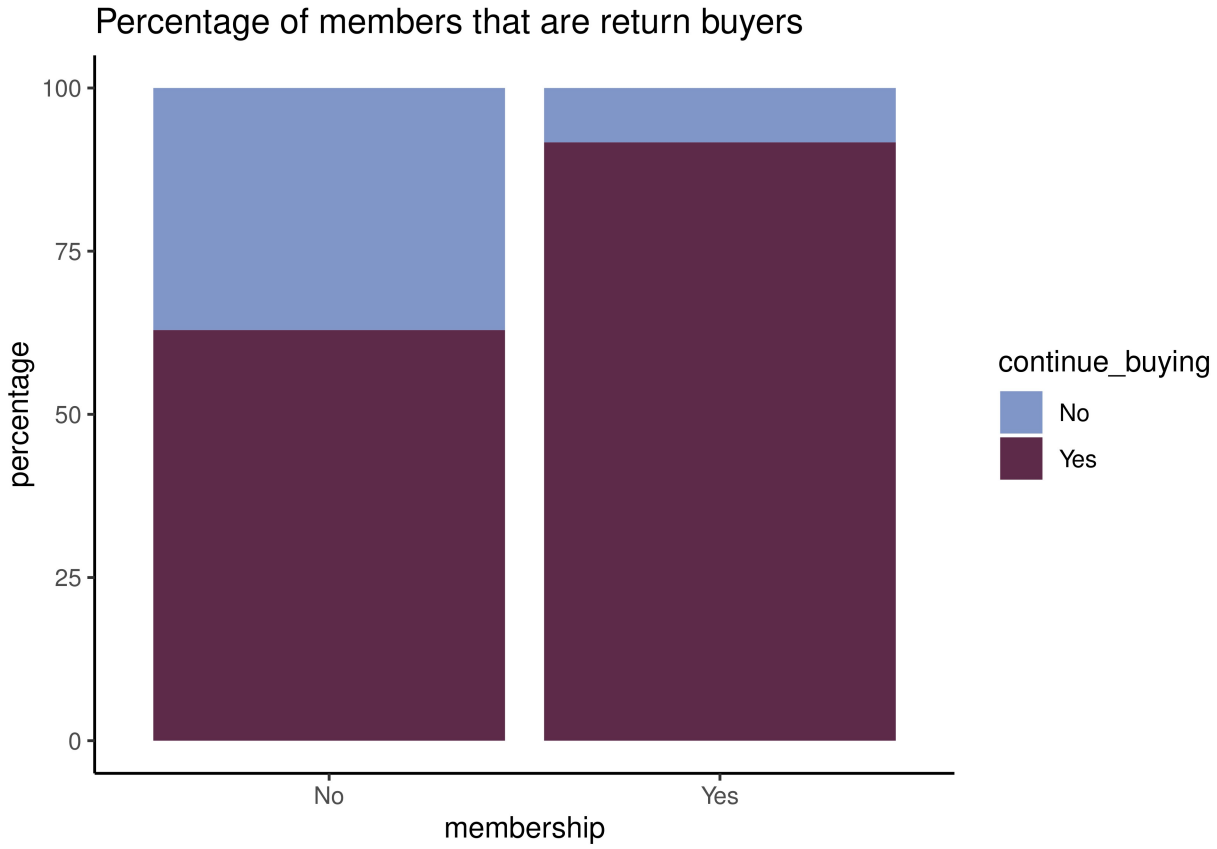
```
cont_table <-with(coffee_house, table(gender, continue_buying))
alpha = 0.05 # significance level of 5% is arbitrarily chosen
chisq.test(cont_table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cont_table
## X-squared = 1.8109e-30, df = 1, p-value = 1
```

A p-value of 1 means we have to fail to reject the hypothesis, thereby establishing that there is no relationship whatsoever between gender and customer retention. The coffee store has a membership scheme which this dataset doesn't indicate whether there are perks attached to that or not. But, you'd like to assume that surely there are some incentives for customers to become members or purchase membership tags. So, this could possibly have an effect on whether they choose to return or not. In this case, we're particularly interested in the factors behind customers not returning so it is important to consider the effect of membership on this.
* One pertinent question is, what percentage of customers are return buyers?

```
mem_return<- coffee_house %>%
  select(membership, continue_buying) %>%
  group_by(membership) %>%
  count(continue_buying) %>%
  mutate(percentage = n/sum(n)*100)
```

```r
# visualize this data for better inferences
 ggplot(mem_return, aes(x=membership, y = percentage, fill = continue_buying)) +  geom_col(position = "
    theme_classic() + labs(title = "Percentage of members that are return buyers") +
   scale_fill_manual(values = c("#8197c9", "#5e2a4b"))
```



Visualizing this data makes it immediately obvious that, compared to members, a larger percentage of non-members opt not to return to Coffee house. A chi square test is needed to verify that these variables are indeed correlated.

```r
cont_table_2 <- with(coffee_house, table(membership, continue_buying))
  alpha = 0.05
  chisq.test(cont_table_2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cont_table_2
## X-squared = 12.686, df = 1, p-value = 0.0003685
```

The p-value here is indeed less than `alpha` and therefore the null hypothesis has to be rejected.

> NOTE: In a chi squared test, the null hypothesis, H0, states that no relationship exists between the categorical variable.

This confirms the association between membership and customer retention, an important point to look out for. Usually, when you want to check out any new store, you have a look around for the reviews from other people who have visited that store. For this dataset, reviews of users have been captured in the ***"Overall_service_rating"*** variable.
This service rating is an important metric for stores and it is therefore imperative that any associations

between this variable and customer retention.
To test for this, I have initially converted the variable from a factor to a numeric variable, and then calculated for the average rating.

```
 coffee_house$overall_service_rating = as.numeric(coffee_house$overall_service_rating)
  class(coffee_house$overall_service_rating)
```

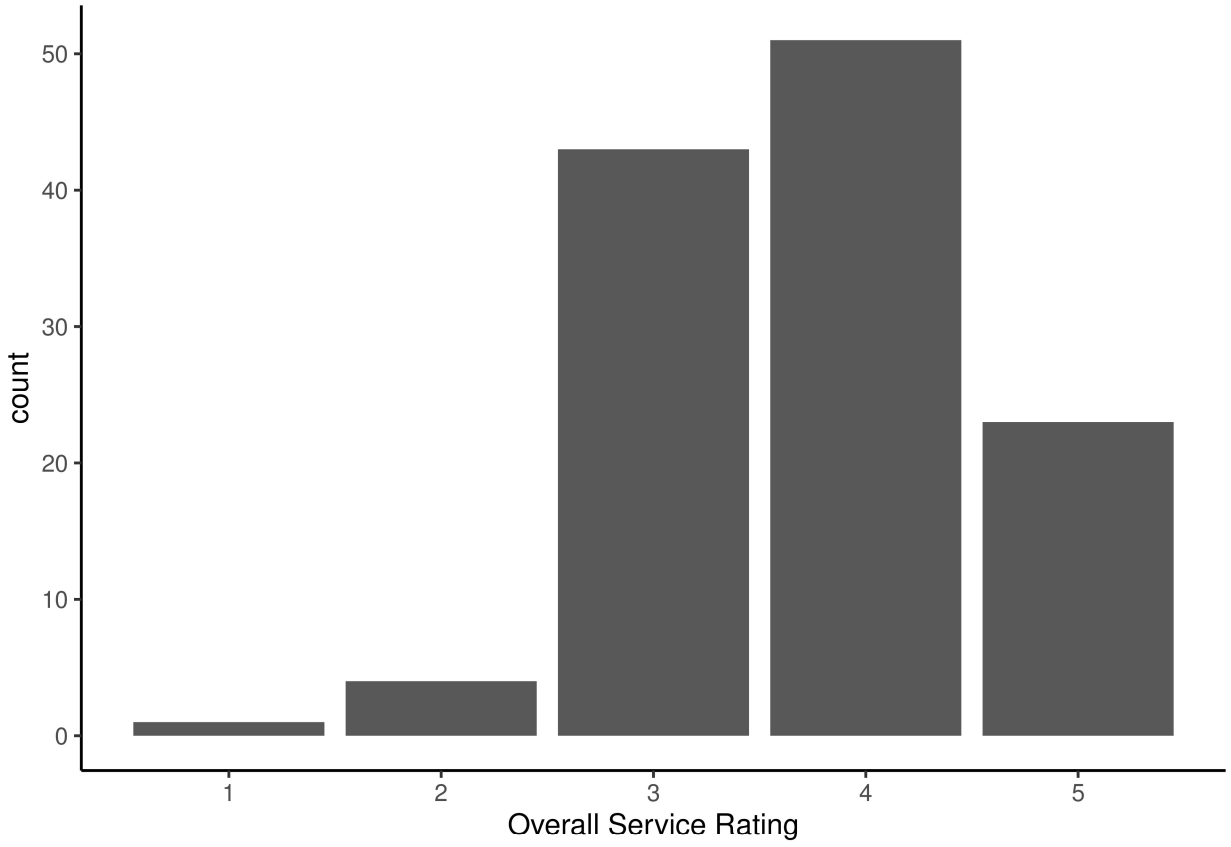
```
## [1] "numeric"
```

```
# calculate for average service rating
 avg_service <- coffee_house %>%
    summarise(avg_service = mean(overall_service_rating)) %>%
    pull(avg_service)
print(avg_service)
```

```
## [1] 3.745902
```

This relationship is now reprsented using a bar plot

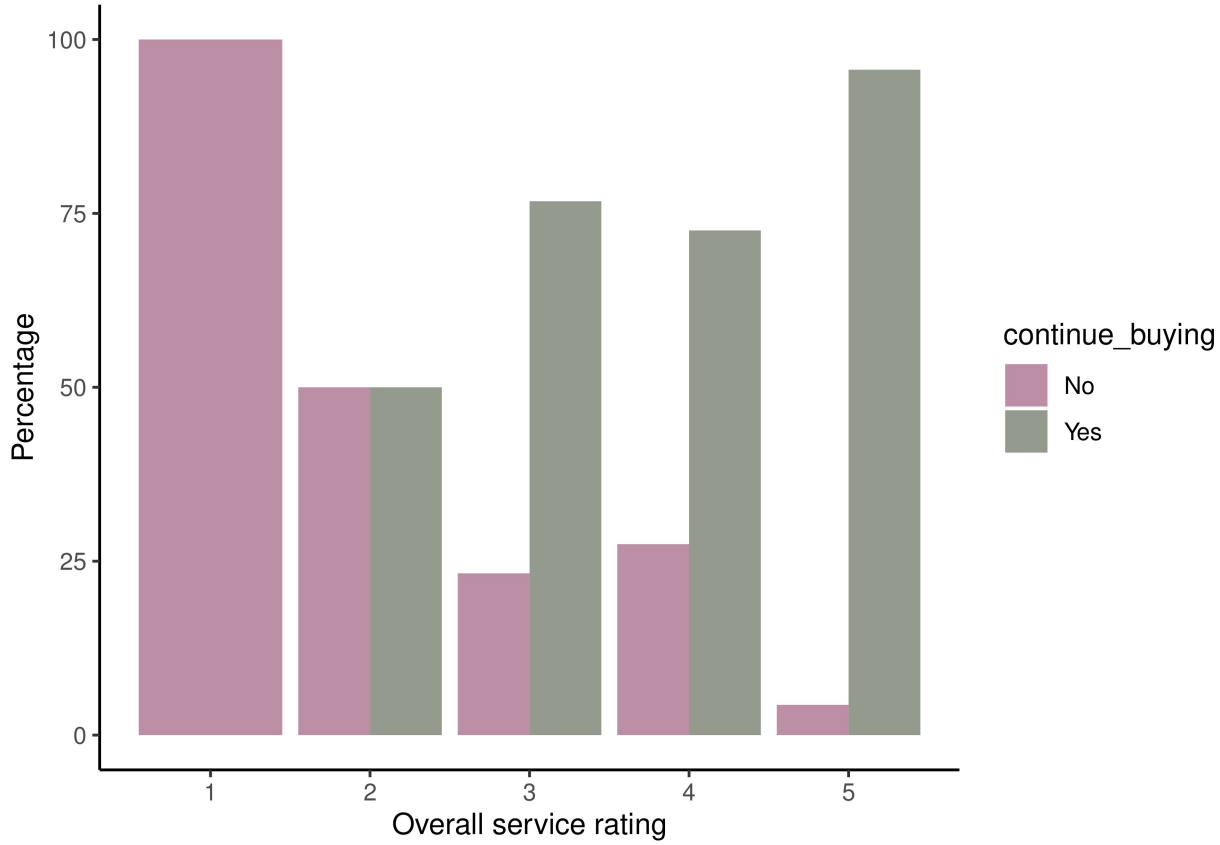
```
    ggplot(coffee_house, aes(x = overall_service_rating))  + geom_bar() + theme_classic() +
  labs(x = "Overall Service Rating")
```



The bar plot corroborates what we already know from the average service rating value of 3.74 gotten in the previous calculation. Majority of the ratings fall between 3 and 4,and the task for the company is clear, figure out how to completely eliminate those lower ratings.
From each of these ratings, let's find out what percentage of customers that give each of these would like to come back

```
rating_ret <- coffee_house %>%
      group_by(overall_service_rating) %>%
```

```
    count(continue_buying) %>%
    mutate(pct = (n/sum(n))*100)
  ggplot(rating_ret, aes(overall_service_rating, pct, fill = continue_buying)) + geom_col(position =
    theme_classic()
```




```
# they generally seem satisfied but is this statistically relevant to customer retention rate?
   cont_table_3 <- with(coffee_house, table(overall_service_rating, continue_buying))
   alpha = 0.05
  suppressWarnings(chisq.test(cont_table_3))
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont_table_3
## X-squared = 10.1, df = 4, p-value = 0.03878
```

```
  # our p value is less than alpha, which means there's an association
```

We have indeed established that overall service rating plays a major part in customer retention. Logically, a myriad of factors are considered by a customer before giving the overall service rating, so it is important for the company to know what some of these factors could be and try to improve upon them in order to increase customer retention and conseqeuntly, profits.
I've picked out two factors to analyze:

- wifi rating
- price range rating

```
# convert from factor to numeric
coffee_house$wifi_rating = as.numeric(coffee_house$wifi_rating)
   avg_wifi_rating <- coffee_house %>%
     summarise(avg_wifi_rating = mean(wifi_rating)) %>%
     pull(avg_wifi_rating)
   # is this mean less than the group mean
   avg_wifi_rating < avg_service
```

```
## [1] TRUE
```

```
# convert from factor to numeric
coffee_house$price_range_rating = as.numeric(coffee_house$price_range_rating)
  avg_price_rating <- coffee_house %>%
    summarise(avg_price_rating = mean(price_range_rating))
  # is it also less than the group mean?
  avg_price_rating<avg_service
```

```
##      avg_price_rating
## [1,]             TRUE
```

Analysis shouldn't be done in isolation, so it is important to corroborate this with the fisher Exact test. This is used instead of the chi sqaure test because it is more appropriate for smaller sample sizes.

```
cont_table_4 <- with(coffee_house, table(overall_service_rating, wifi_rating))
cont_table_5 <- with(coffee_house, table(overall_service_rating, price_range_rating))

suppressWarnings(fisher.test(cont_table_4, simulate.p.value = T))  # to decrease computational time of
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  cont_table_4
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

```
suppressWarnings(fisher.test(cont_table_5, simulate.p.value = T))
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  cont_table_5
## p-value = 0.0009995
## alternative hypothesis: two.sided
```

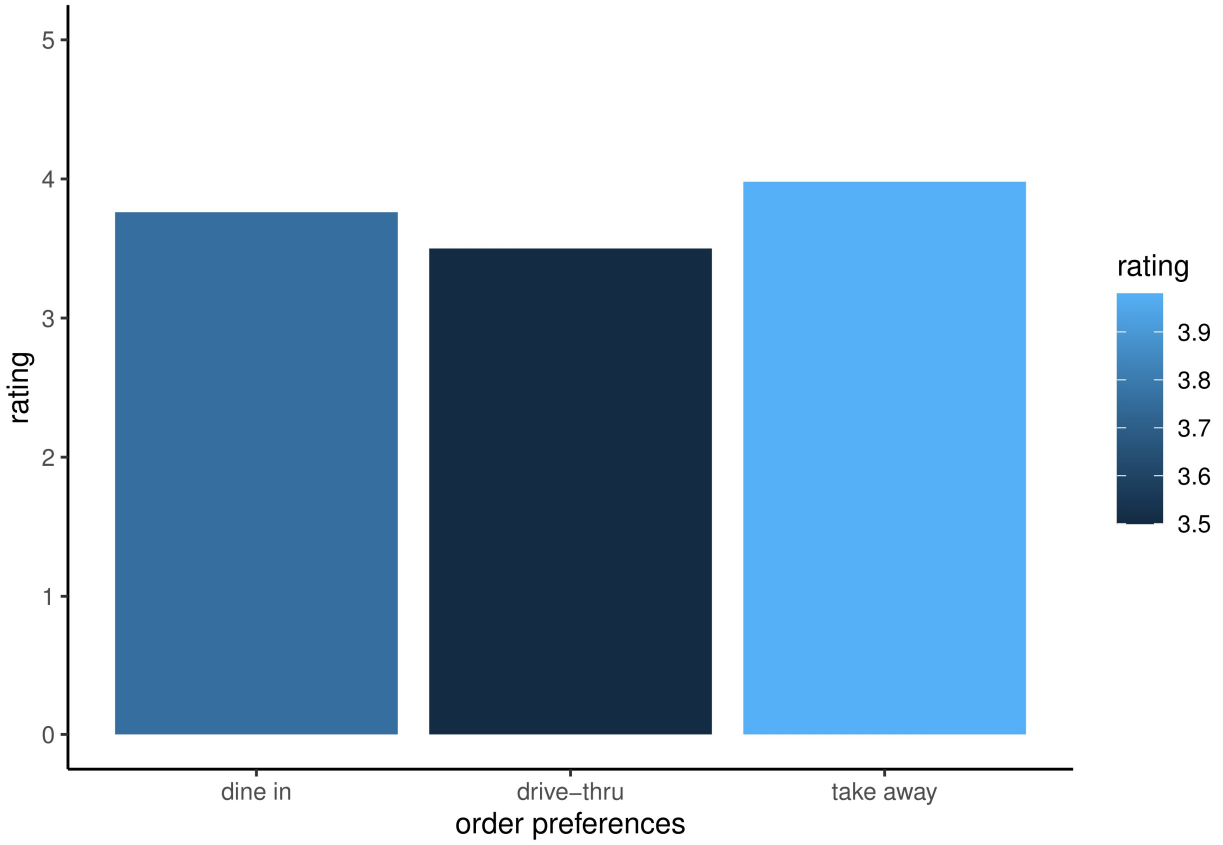
The p-value once again is less than the alpha value of 0.05, which means we have to reject the null hypothesis. Both of these variables are important factors when it comes to overall service rating, so seeking to improve them would ultimately have an effect on customer retention. Finally, I want to explore what observations can be drawn from the order preferences column. This will be done by getting the average rating by each preference type.

```
 # all values are to be converted to lower case because there's a "Never" and another "never" value
 coffee_house_2 <- coffee_house %>% # create a new df
   mutate(order_pref = tolower(order_preference))
coffee_house_2$overall_service_rating = as.numeric(coffee_house_2$overall_service_rating)
```

```r
# to get the average ratings
  pref_rating <-  coffee_house_2 %>%
   group_by(order_pref) %>%
   filter(order_pref != "na") %>%
   summarise(rating = mean(overall_service_rating)) %>%
    arrange(-rating)
  pref_rating
```

```
## # A tibble: 6 x 2
##   order_pref         rating
##   <chr>               <dbl>
## 1 take away            3.98
## 2 dine in              3.76
## 3 drive-thru           3.5
## 4 i dont like coffee   3
## 5 never buy            3
## 6 never                2.5
```

```r
  # visualize this relationship
  pref_rating_1 <-  pref_rating %>%
    filter(order_pref %in% c("take away", "dine in", "drive-thru"))
  pref_rating_1 %>%
    ggplot(aes(x = order_pref, y = rating, fill = rating)) + geom_col() + ylim(0, 5) + labs(x = "order
```



From the visual, customers using the drive-in method gave the least ratings on average.

Key Insights

1. Membership programs have a significant impact on customer retention and revenue for the coffee store.
2. Providing good quality and reliable wifi is important for customers, and improving this helps to increase customer satisfaction and attract new customers.
3. Price is an important factor for customers, and overpriced products led to lower ratings and decreased customer satisfaction.
4. Customer service is critical, especially for the drive-through medium, and improvements have to be made to increase customer satisfaction and loyalty

Recommendations

From my analysis of the coffee store dataset and the insights drawn, I will make the following recommendations:

1. Encourage more customers to become members by offering incentives such as discounts, free coffee, or a membership reward system like loyalty points.
2. Improve the quality and speed of the wifi to increase customer satisfaction and attract more customers who need a place to work or study.
3. Re-evaluate the pricing of certain products to ensure that they are competitive and in line with customers' expectations.
4. Provide additional training and support to staff working in the drive-through to improve their communication skills and speed of service.