Cyclistic: A Case Study

Adzuira Musule Palacios
Google Data Analytics Capstone Project
April, 2025

Executive summary

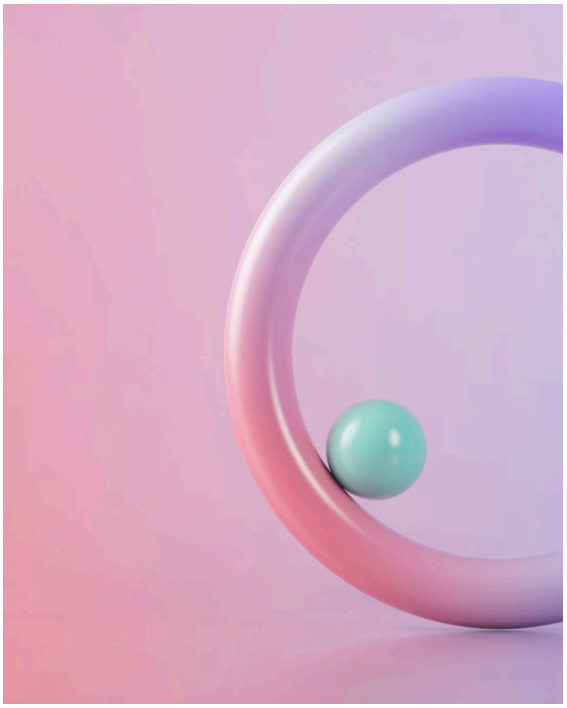
01. Overview

From casual riders to members

Cyclistic, a bike-share company in Chicago with more than 5,800 bicycles and 600 docking stations, is looking to analyze data from the past year to look for opportunities to maximize the number of annual memberships. The director of marketing from Cyclistic would like to design a strategy based on the findings of this study. This analysis is made using R.

For this project, data from 2021 to 2022 was used to create visualizations and gather insights, which were then used to inform recommendations on potential marketing strategies to meet business goals.

Casual usership spikes in warmer months and most often on weekends. Frequent users who use the service a lot during these times could be a good market segment to promote memberships to. A surprising number of casual users who take very long rides stood out in the dataset, this group could also benefit from a marketing campaign aimed at informing them of the benefits of a membership.



02. Key findings

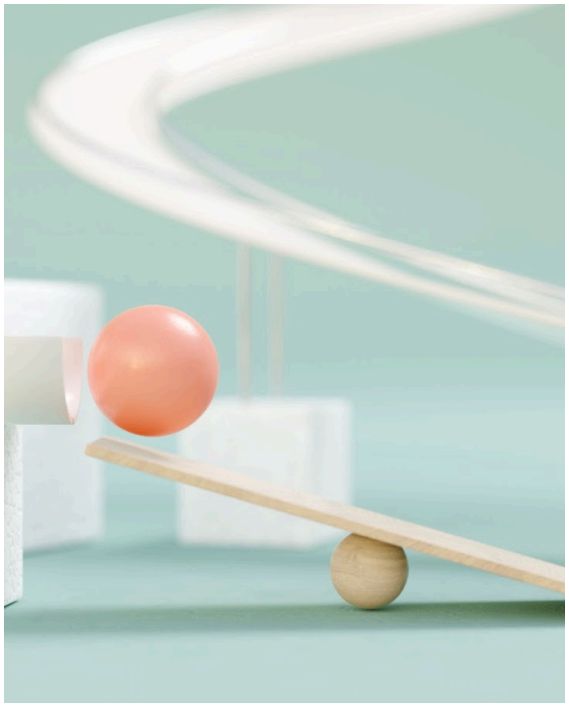
a. Casual riders could be more willing to purchase a membership on the weekends.
b. Many casual riders use the Cyclistic services for many hours.
c. When casual riders use the service, they use it significantly more than members.
d. The popular months in which Cyclistic has more business from both casual and members are spring and summer.

03. Proposed solutions

a. Prioritize starting marketing campaigns during the weekends and in the months corresponding to spring and summer, as this is when casual usership peaks.

b. Provide special offers to casual users that spend more money on their rides than the cost of membership.

c. Provide short-term membership for the busiest months.



Introduction

01. Background

Cyclistic: A road to member increase

Cyclistic, a fictional bike-sharing company, is currently located in Chicago and features more than 5,000 bicycles as well as 600 docking stations. The company has extensive options for its users, offering reclining bikes, hand tricycles, and cargo bikes, which makes their products more inclusive for people with disabilities. The company reported that the majority of riders opt for traditional bikes, with approximately 8% of users opting for the assistive options. Furthermore, Cyclistic reported that about 30% of its users use the bikes to commute to work every day and the rest of its users are most likely to ride for leisure.

02. Objectives

a. Understand how casual riders and members use bikes differently.
b. Perform statistical analysis to gain insights from the data.
c. Based on the insights, design a new marketing strategy to convert casual riders into annual members using data visualization.

03. Problem

As part of its future marketing strategy, Cyclistic is looking to convert more of its casual riders to members. This would allow the company to have more customer loyalty and, therefore, gain more profit.

04. Stakeholders

- **Lily Moreno**: This report will be first presented to the director of marketing and manager Lily Moreno who is responsible for the development of campaigns and initiatives to promote the bike-sharing program.
- **Cysclictic Marketing Analytics Team:** This is a team of data analysts that help guide Cyclistic marketing strategy. The team is responsible for collecting, analyzing, and reporting data.
- **Cyclistic Executive Team:** This team is responsible for approving the recommended marketing program.

Recommendations

1. Prioritize starting marketing campaigns during the weekends and in the months corresponding to spring and summer, as this is when casual usership peaks.
2. Provide special offers to casual users that spend more money on their rides than the cost of membership.
3. Provide short-term membership for the busiest months.

01. Impact of recommendation

SOLUTION	EFFECTIVENESS	IMPACT	NOTES
Market during peak casual usership times	Highly effective	Casual users are more likely to see marketing materials and engage with them during periods of heavy usage.	
Provide special offers to customers who spend a lot on casual rides	Moderately effective	Users might not be aware that their casual use is more expensive than a membership. This could help convert some users.	
Provide special short-term duration memberships for peak season	Moderately effective	Casual users prefer warmer months, and a short-term membership for the season might entice them.	A follow-up strategy could be to offer a discount on a full year membership to the users who already paid a seasonal membership.

Analysis

01. Research methods

Sourcing the data

The data is provided by the link included in Google Analytics Cyclistic case study (as mentioned in the case study instructions, the data has been made available by Motivate International Inc. under this license). The first 12 months of data were downloaded from the source, read, and combined using R to prepare the dataset for analysis.

02. Approaches used

Preparing the data for analysis in R

The data from the past 12 months was combined in a single file, being April of 2020 the earliest data and March 2021 the latest data.

For seasonal analysis, the 12 months were divided into four seasons: spring, summer, fall, and winter, and a new column called “Season” was added.

First cleaning of the data

The dataset contained missing values in some of the columns, for which the analysis team does not have another data source to fill in the missing information. Therefore, assuming that for this analysis, incomplete data is not relevant for the analysis, those rows were removed for the scope of this project.

```
# Cleaning the data frame to remove empty information (if any) assuming that
for this analysis, incomplete data is not relevant for the analysis.

cyclistic_data <- janitor::remove_empty(cyclistic_data, which = "cols")
cyclistic_data <- janitor::remove_empty(cyclistic_data, which = "rows")
```

New columns

A new column called “ride_length” was created to know the duration of each trip. To create this new column, the "started_at" column and the "ended_at" column were set as a date format. After, the "started_at" column was subtracted from the "ended_at" column. After the subtraction, the team noted that some values in the ride_length column were negative. Since the analysis team does not have the tools to further investigate these negative values, these records were removed from the analysis.

To categorize the dataset according to the day of the week, a “day_of_week” column was created. Additionally, a “day_of_week_number” was created. This last column is useful for the creation of a viz later in the analysis.

```
  # Add new column called "ride_length" by subtracting the "started_at"
  column from the "ended_at" column.

  # Set the "started_at" column and the "ended_at" column as dates.

  cyclistic_data$started_at <- as_datetime(cyclistic_data$started_at)
  cyclistic_data$ended_at <- as_datetime(cyclistic_data$ended_at)

  # Subtract start from end date
```

```
cyclistic_data$ride_length <- cyclistic_data$ended_at -
cyclistic_data$started_at

# Removing negative values from the column "ride_length"

cyclistic_data <- cyclistic_data [cyclistic_data$ride_length >= 0, ]

# Create new column called "day_of_week" from the started_at column.

cyclistic_data$day_of_week <- weekdays(cyclistic_data$started_at)

# Create a new column called day_of_week_number to create a histogram.

cyclistic_data$day_of_week_number <- wday(cyclistic_data$started_at)
```

Analysis

Casual riders vs Members: Ride length

To get a better sense of the data and to establish a first big picture, basic statistics, such as the mean, the max, and the mode of the "ride_length" column, were calculated. For the mean and max, the "ride_length" data will be used, and for the mode, the "day_of_week" column that was previously created will be used.

```
# Calculate mean, max, and the mode of ride_length to get a sense of the data
statistically.

mean(as.numeric(cyclistic_data$ride_length))
max(as.numeric(cyclistic_data$ride_length))
mode_calc <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

mode_calc(cyclistic_data$day_of_week)
```

These calculations were:

- Mean: 1677.049 seconds
- Max: 3523202 seconds
- Mode: Saturday

Based on max calculation, the team considered removing outliers. However, the team learned that thousands of rides are longer than a day, therefore, it was decided to keep these data as these could be real use cases of the service. A sum function in the “ride_lenght” column revealed that 2882 rides were longer than a day (86400 seconds).

Based on the large number of rides longer than a day, the team calculated the number of casual riders from this group to see if an approach could be arranged to identify a possible group that could be interested in a membership.

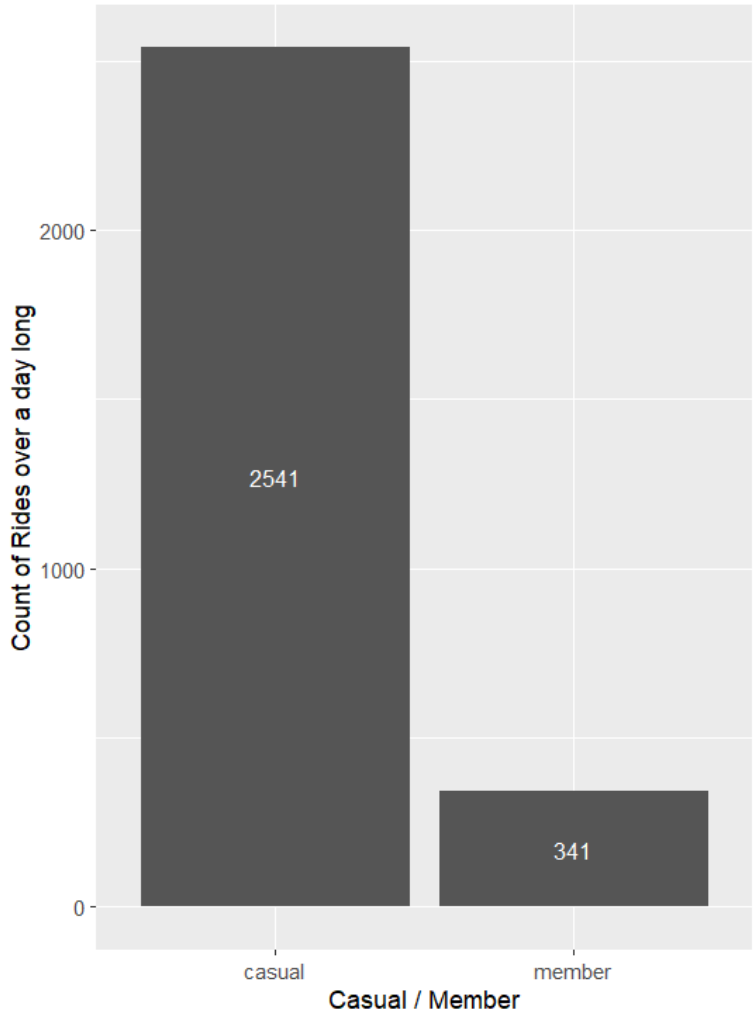
```
# Count of user with length of more than a day (86400 seconds)

# Casual Users
sum((cyclistic_data$ride_length > 86400 & cyclistic_data$member_casual ==
"casual"))

# Members
sum((cyclistic_data$ride_length > 86400 & cyclistic_data$member_casual ==
"member"))
rides_longer_than_a_day <- cyclistic_data[cyclistic_data$ride_length >
86400, ]
ggplot(data = rides_longer_than_a_day, aes(x=member_casual)) +
  labs(x = "Casual / Member",
       y = "Count of Rides over a day long") +
  geom_bar(stat = "count") +
  stat_count(geom = "text", colour = "white", size = 3.5,
             aes(label = ..count..),position=position_stack(vjust=0.5))
```

The results for these calculations were:

- Casual riders: 2541
- Member riders: 341



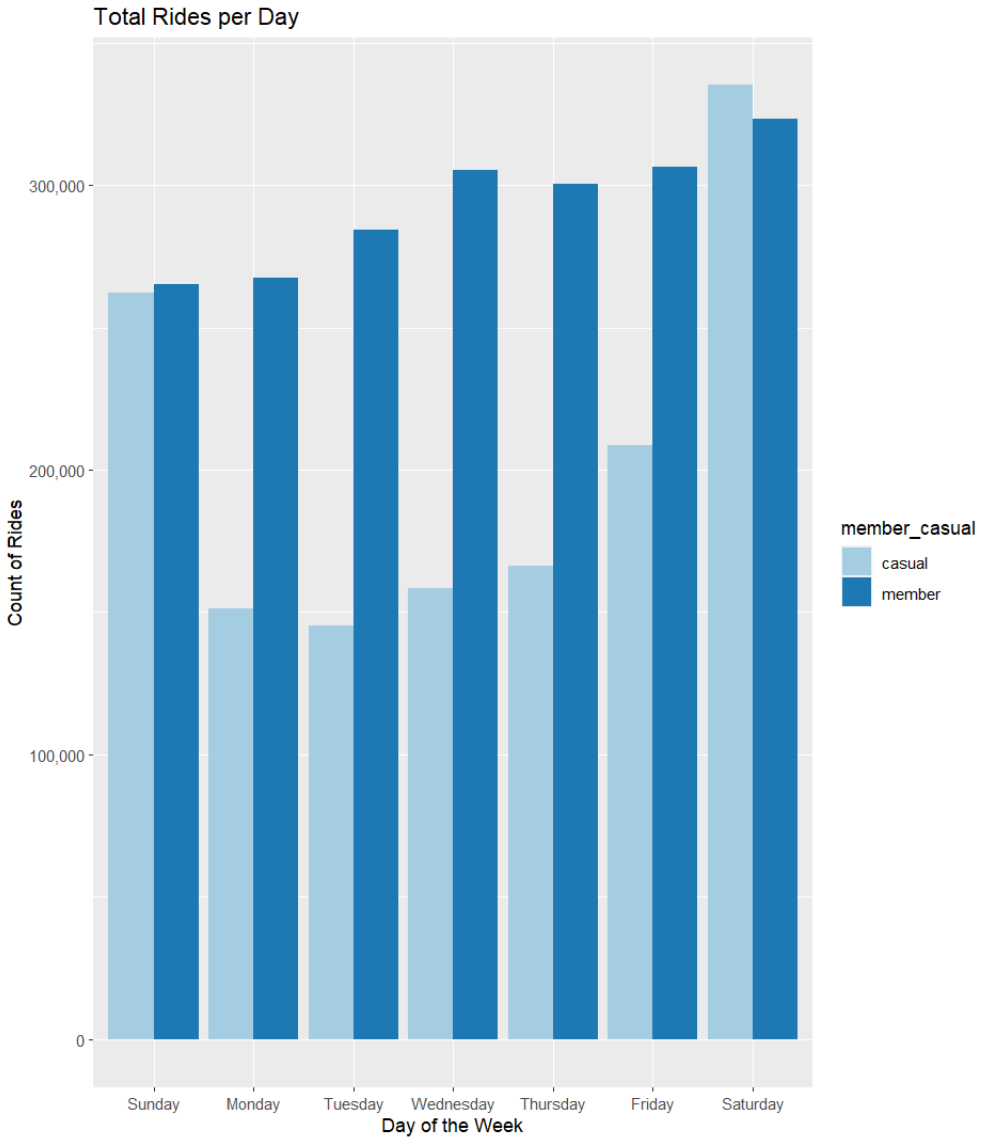
Knowing that most rides (88.16%) longer than a day corresponded to the casual riders group, the marketing team could identify these users and offer a membership plan.

Casual riders vs Members: Rides per day of week

The team also wanted to identify if more days of the week were more popular for the casual riders compared to the members. For this, a plot showing the number of rides per day of the week was created.

```
# Plot rides per day of the week
week_order <- c("Sunday", "Monday", "Tuesday", "Wednesday",
                      "Thursday", "Friday", "Saturday")
ggplot(data = cyclistic_data) +
  labs(title = "Total Rides per Day",
       x = "Day of the Week",
       y = "Count of Rides") +
  scale_fill_brewer(palette = "Paired") +
  geom_bar(mapping = aes(x = factor(day_of_week, week_order), fill =
```

```
member_casual), position = "dodge") +
  scale_y_continuous(labels = label_comma ())
```



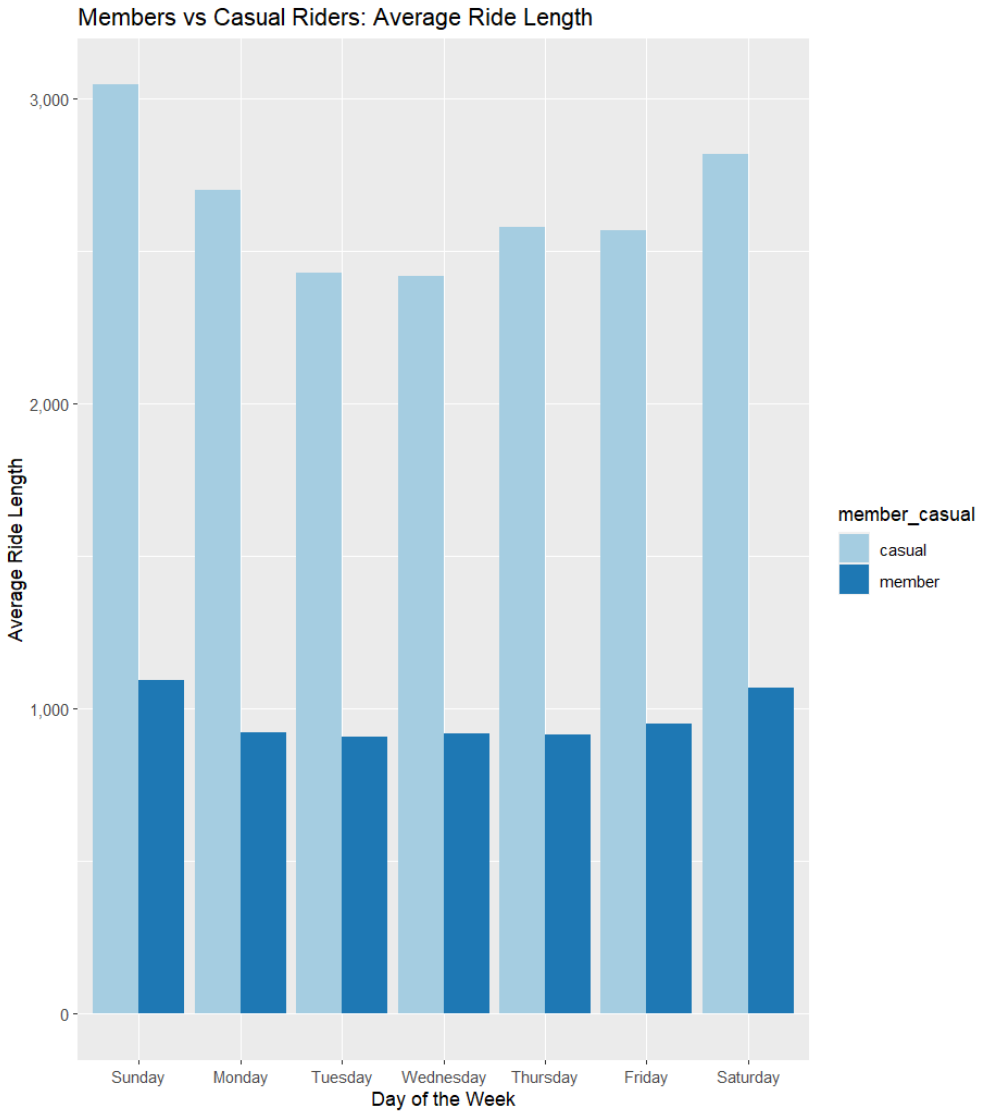
Based on this graph there are a few insights that were identified. It is evident that members use the Cyclistic services more in general. However, on Saturdays more casual riders use the service. This observation could be due to the fact that members could probably be using the services for commuting and casual riders use the services for leisure most of the time. The marketing team could take this information to send offers to its casual riders during the booking time on the weekends, when casual riders use the services more and therefore, could be more willing to purchase a membership.

Casual riders vs Members: Average ride length

Additionally, the team was also interested in comparing the average ride length between casual and members using a bar chart.

```
# Plot comparing members with casual users using average ride length

cyclistic_data %>%
  group_by(day_of_week, member_casual) %>%
  summarise(avg_ride_length = mean(ride_length)) %>%
  ggplot() +
  labs(title = "Members vs Casual Riders: Average Ride Length",
       x = "Day of the Week",
       y = "Average Ride Length") +
  scale_fill_brewer(palette = "Paired") +
  geom_bar(mapping = aes(x = factor(day_of_week, week_order), y =
avg_ride_length, fill = member_casual), stat = "identity", position =
"dodge") +
  scale_y_continuous(labels = label_comma ())
```



In this graph, it can be seen that even though members use the service almost every day, when casual riders use the service, they use it more significantly more time on average compared to the member riders. This could also be an opportunity for the marketing team to send out offers to the casual users that use the bikes more time on average. Therefore, the casual users could probably save money considering they are already probably spending more than what the cost of the membership is. Cyclistic could benefit from this since it could create loyalty in its casual riders.

Casual riders vs Members: Average ride length by Season

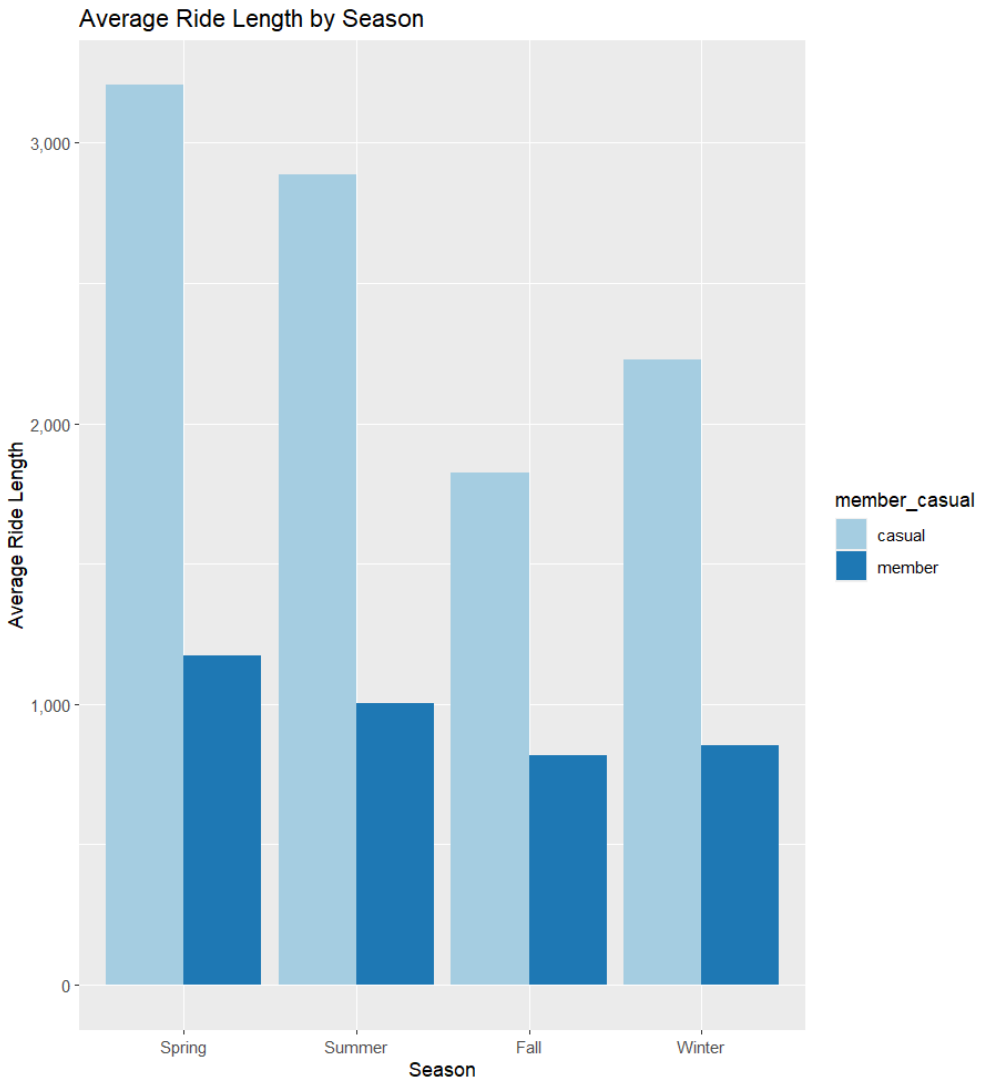
Finally, the team was also interested in comparing the average ride length by season between casual and member riders. As mentioned previously, for seasonal analysis, the 12 months were divided into four seasons: spring, summer, fall, and winter.,Since

the data was divided by month, this study considers January, February, and March as Winter, April, May and June as Spring, July, August and September as Summer, and October, November and December as Fall.

```
# Plot comparing average ride length by season

season_order <- c("Spring", "Summer", "Fall", "Winter")

cyclistic_data %>%
  group_by(season, member_casual) %>%
  summarise(avg_ride_length = mean(ride_length)) %>%
  ggplot() +
  labs(title = "Average Ride Length by Season",
       x= "Season",
       y= "Average Ride Length") +
  scale_fill_brewer(palette = "Paired") +
  geom_bar(mapping = aes(x = factor(season, season_order), y =
avg_ride_length, fill = member_casual), stat = "identity", position =
"dodge") +
  scale_y_continuous(labels = label_comma())
```



This bar chart shows that the season in which more people is riding is spring and sommer. It could be assumed that this is due to the higher weather temperatures. So the marketing team could focus its marketing efforts to convert its casual riders to members in the popular months.

03. Relevant facts and information

a. Many casual riders use the Cyclistic services for many hours, making them potential members.

b. Casual riders could be more willing to purchase a membership on the weekends since it's when they use the Cyclistic services more, making it a good chance to send offers during the booking time.

c. Even though members use the service almost every day, when casual riders use the service, they use it significantly longer time. The marketing team could send out offers to the casual users that use the bikes more time on average.
d. The popular months in which Cyclistic has more business from both casual and members are spring and summer. The marketing team could focus on releasing its marketing strategies on those months with higher weather temperatures.

Conclusion

01. Summary of findings

The analysis showed that the casual riders could be more willing to purchase a membership on the weekends of spring and summer, therefore, the marketing team could prioritize its strategies to those months when casual usership peaks. Also, the team also observed that when casual users use the bikes, they use them for more time than members on average. Additionally, the analysis showed that some casual users use the service for more than 24 hours consecutively. This could suggest that some users hold on to the bikes to reserve one for the next day or that the user left the service running.

As a potential strategy, the marketing team could group the casual users based on further data such as member ID to offer a membership to those users that are already paying more than the cost of membership.

Another strategy could involve providing short-term membership for the busiest months.

02. Limitations of this study and follow-up analysis

One of the limitations that the analysis team encountered was that the data did not provide user IDs. Therefore, the thousands of trips with longer duration could not be grouped by user ID to know if these rides were performed by a recurring group of users or accidental cases were the user was unable to end the trip. Therefore, the analysis team had to consider this data in the study since there was no further way to rule out these possible outliers. The data also did not provide information about distance travelled; this could have provided additional insights useful for alternative marketing strategies. A follow-up analysis could request this data to reach a more thorough conclusion.

Another limitation is that some of the incomplete data does not have station IDs or station names; therefore, the team could not fill in the information that was missing,` and some of the data had to be excluded from the analysis. A follow-up investigation could be performed to know why this information is not being recorded. This effort could also result in a more thorough conclusion.

References

Google Career Certificates. Coursera. Case study: How does a bike-share navigate speedy success?

How to put labels over geom_bar in R with GGPLOT2. Stack Overflow. https://stackoverflow.com/questions/6455088/how-to-put-labels-over-geom-bar-in-r-with-ggplot2

Count observations greater than a particular value. Stack Overflow. https://stackoverflow.com/questions/22690255/count-observations-greater-than-a-particular-value/39273973

R - how to remove outliers from dataset by two different groups. Stack Overflow. https://stackoverflow.com/questions/66411056/r-how-to-remove-outliers-from-dataset-by-two-different-groups

John. (2020, January 19). *How to remove outliers in R | R-bloggers*. R-bloggers. https://www.r-bloggers.com/2020/01/how-to-remove-outliers-in-r/