Case_Study

Ismoil Ismoilov

2024-08-26

Case Study: How Does a Bike-Share Navigate Speedy Success?

In this article, I'll share my insights and offer well-informed recommendations on marketing strategies that will drive the company's success. To ensure a comprehensive and structured analysis, I'll be following Google's six-step data analysis process: ask, prepare, process, analyze, share, and act.

Background Cyclistic, a Chicago-based bike-share program, has grown significantly since its launch in 2016, now featuring over 5,800 bikes and nearly 700 docking stations. The company offers various bike options, including those accessible to people with disabilities. While the majority of riders use traditional bikes, a small percentage opt for assistive bikes. Cyclistic's pricing plans include single-ride passes, full-day passes, and annual memberships, with annual members being more profitable than casual riders.

The marketing team, led by Lily Moreno, aims to convert casual riders into annual members, as they believe this is key to the company's future growth. To achieve this, the team plans to analyze historical bike trip data to better understand the differences between casual riders and annual members. These insights will help in designing targeted marketing strategies, which will need executive approval before implementation.

The objective of this case study Address three key questions that will shape the future marketing program:

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

The director of marketing has assigned me the task of addressing the first question: In what ways do annual members and casual riders use Cyclistic bikes differently

For this assignment, I will produce a report that includes the following deliverables:

- A clear statement of the business task
- A description of all data sources used
- Documentation of any cleaning or manipulation of data
- A summary of your analysis
- Supporting visualizations and key findings

Ask Moreno has assigned me the first question to address: In what ways do annual members and casual riders use Cyclistic bikes differently?

Prepare To perform a comprehensive analysis, we need to acquire Cyclistic's historical trip data from August 2023 to July 2024. This data has been provided by Motivate International Inc. under the specified license.

Once the datasets are downloaded from the designated website, we will organize them in a dedicated folder named "Case_Study1" with each file in .csv format.

Process To clean, analyze, and aggregate the extensive monthly data stored in the folder, we will use RStudio.

To begin, we'll create a new script or query to use as a workspace for performing operations on the Cyclistic data. We'll also set the working directory to the path of the folder where the data is stored.

```
#setting the working directory
setwd("F:/Case_Study1/")
```

Next, we'll load all the necessary packages required for data manipulation.

```
library(tidyverse)
library(lubridate)
Sys.setlocale("LC_TIME", "English")
```

With all the packages loaded, we will aggregate the data in our Case_Study1 folder using the `map_df()` function.

```
#get all the files with .csv data type in the working directory which is our folder containing the data
aggregate_files <- list.files(pattern = "*.csv")
# read the .csv files in the aggregate_files and map it into a single data frame
aggregate_data <- map_df(aggregate_files, read_csv)
```

To examine the structure of the newly created aggregate_data data frame, we will use the `str()` function.

```
str(aggregate_data)
```

```
## spc_tbl_ [5,715,693 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5715693] "903C30C2D810A53B" "F2FB18A98E110A2B" "D0DEC7C94E4663DA" "E0DE
##  $ rideable_type     : chr [1:5715693] "electric_bike" "electric_bike" "electric_bike" "electric_bik
##  $ started_at        : POSIXct[1:5715693], format: "2023-08-19 15:41:53" ...
##  $ ended_at          : POSIXct[1:5715693], format: "2023-08-19 15:53:36" ...
##  $ start_station_name: chr [1:5715693] "LaSalle St & Illinois St" "Clark St & Randolph St" "Clark St
##  $ start_station_id  : chr [1:5715693] "13430" "TA1305000030" "TA1305000030" "KA1504000135" ...
##  $ end_station_name  : chr [1:5715693] "Clark St & Elm St" NA NA NA ...
##  $ end_station_id    : chr [1:5715693] "TA1307000039" NA NA NA ...
##  $ start_lat         : num [1:5715693] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5715693] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:5715693] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5715693] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:5715693] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
```

```
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

In the second line of the output, we can observe the total number of rows and columns in the aggregate_data table, listed as `spc_tbl_ [5,715,693 × 13]`. Below this, we'll find the columns along with a preview of the fields each column contains. The data types for each column are also indicated by different colors.

Since there are no errors regarding the data types or other issues, we can proceed with cleaning the dataset. We will create a new data frame with the cleaned data, which we will name cyclistic_clean_data.

```
cyclistic_clean_data <- aggregate_data %>%
  select("ride_id", "rideable_type", "started_at", "ended_at",
         "start_station_name", "end_station_name", "member_casual") %>%
  na.omit() %>%
  mutate(trip_length = as.numeric(difftime(ended_at, started_at, units = "mins")),
         weekday = format(as.Date(started_at), "%A")) %>%
  select("ride_id", "rideable_type","started_at", "ended_at", "start_station_name", "end_station_name",
         "weekday", "trip_length", "member_casual") %>%
  filter(trip_length >=1, trip_length <= (24*60))
```

Let's break down the code chunk step by step to better understand the operations needed for cleaning our data:

Stage 1: Defining a variable

```
cyclistic_clean_data <- aggregate_data %>%
```

The line of code above suggests that the output from aggregate_data is used as input for the following operations, which will be applied sequentially. The final result of these operations will be assigned to the cyclistic_clean_data object, creating a new cleaned dataset.

Stage 2: Choosing the relevant columns for analysis

```
select("ride_id", "rideable_type", "started_at", "ended_at",
         "start_station_name", "end_station_name", "member_casual") %>%
```

We applied the select() function to filter specific columns from the aggregate_data output, which serves as input for this function.

Stage 3: Eliminating all missing values

```
na.omit() %>%
```

The na.omit() function will eliminate all rows with missing values (NA) from the selected columns of the aggregate_data. This step helps clean the data and remove elements that could impact the analysis.

Stage 4: Adding new columns required for analysis

We are reminded of our business task: designing marketing strategies to convert casual riders into annual members, and specifically, comparing how casual riders and members use Cyclistic bikes.

To gain valuable insights, we need to extract crucial information from the current data and create new columns. For instance, we can add columns for trip length and weekday. These columns will help us analyze when casual riders and members use the bikes and the duration of their rides. To achieve this, we can use and incorporate the following code into our existing script:

```
mutate(trip_length = as.numeric(difftime(ended_at, started_at, units = "mins")),
         weekday = format(as.Date(started_at), "%A")) %>%
```

We use the mutate() function to add two new columns to the aggregate_data data frame: trip_length and weekday. * The trip_length column is calculated as the difference between the ended_at and started_at columns, expressed in minutes, and is formatted as numeric. * The weekday column is derived by formatting the started_at column to display the day of the week using the format() function.

Stage 5: Select the `trip_length` and `weekday` columns

```
select("ride_id", "rideable_type","started_at", "ended_at", "start_station_name", "end_station_name","w
```

Stage 6: Filter the `trip_length` column

We will filter the `trip_length` column to include only values greater than 1 minute and less than 24 hours, using the `filter()` function, and apply this to the cyclistic_clean_data data frame.

```
filter(trip_length >=1, trip_length <= (24*60))
```

The cleaned data frame is saved as cyclistic_clean_data, containing the selected and added columns, with missing values removed and filtered.

We still need to verify if further cleaning is required. To start, we'll check for NA values in each column by running this function:

```
colSums(is.na(cyclistic_clean_data))
```

```
##            ride_id      rideable_type         started_at
##                  0                  0                  0
##           ended_at start_station_name   end_station_name
##                  0                  0                  0
##            weekday        trip_length      member_casual
##                  0                  0                  0
```

Since there are 0 NA values in each column, we will next check for any duplicate ride_id values and remove them if found.

```
cyclistic_clean_data <- cyclistic_clean_data[!duplicated(cyclistic_clean_data$ride_id),]
any(duplicated(cyclistic_clean_data$ride_id))
```

```
## [1] FALSE
```

The returned value is FALSE, indicating that there are no duplicates in the data frame.

Now, let's recheck the minimum and maximum values of the trip_length column.

```
min(cyclistic_clean_data$trip_length)
```

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

```
max(cyclistic_clean_data$trip_length)
```

```
## [1] 1439.867
```

It appears that we have correctly filtered the trip_length column, as it now shows values within 24 hours. Finally, we should check the structure of our new data frame to ensure that all data types match the expected column types.

```
str(cyclistic_clean_data)
```

```
## tibble [4,179,021 x 9] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:4179021] "903C30C2D810A53B" "6400344C80D626CA" "B56F0D2EC8B33085" "EE2
##  $ rideable_type     : chr [1:4179021] "electric_bike" "electric_bike" "electric_bike" "electric_bik
##  $ started_at        : POSIXct[1:4179021], format: "2023-08-19 15:41:53" ...
##  $ ended_at          : POSIXct[1:4179021], format: "2023-08-19 15:53:36" ...
##  $ start_station_name: chr [1:4179021] "LaSalle St & Illinois St" "Clark St & Randolph St" "Sheffiel
##  $ end_station_name  : chr [1:4179021] "Clark St & Elm St" "Dearborn Pkwy & Delaware Pl" "Sawyer Ave
##  $ weekday           : chr [1:4179021] "Saturday" "Friday" "Saturday" "Wednesday" ...
##  $ trip_length       : num [1:4179021] 11.72 7.62 22.28 8.28 5.17 ...
##  $ member_casual     : chr [1:4179021] "member" "member" "member" "member" ...
##  - attr(*, "na.action")= 'omit' Named int [1:1474331] 2 3 4 5 6 7 8 9 10 11 ...
##   ..- attr(*, "names")= chr [1:1474331] "2" "3" "4" "5" ...
```

With the raw data having 5,715,693 rows, our cleaned dataset now contains 4,179,021 rows and 9 columns. The data types of each column are as needed, so we're ready to start analyzing our cleaned data.

Analyze & Share Now that we have cleaned our data, we can revisit our business question and work towards answering it:

How do annual members and casual riders use Cyclistic bikes differently?

Using the data we've prepared, we can extract crucial insights to address this question. Here's how we can approach it based on the information available:

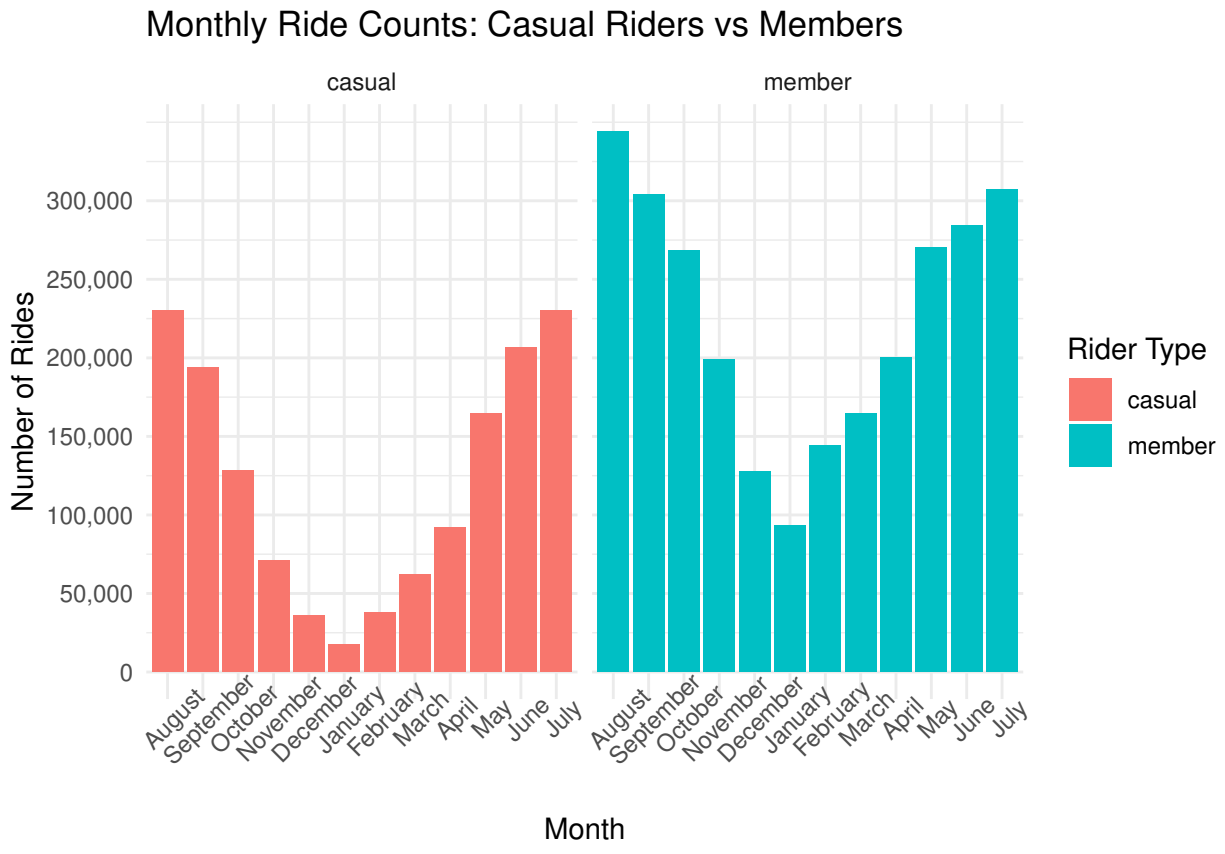
A. Evaluation of the monthly ride count We can examine and compare the monthly ride counts for casual riders and members. To achieve this, we can create a data frame by extracting the relevant columns from the `cyclistic_clean_data` — specifically the `started_at` column (converting the datetime to its corresponding month) and the `member_casual` column. We'll also add a `row_count` column to tally the bike rides each month.

```
month_count = cyclistic_clean_data %>%
  group_by(months = month.name[month(started_at)], member_casual) %>%
  summarize(row_count = n()) %>%
  arrange(match(months,month.name))
```

The data frame we created is named `month_count`. The data frame is organized by the `months` and `member_casual` columns, showing the total ride counts for each month under the `row_count` column for both casual riders and members.

To make the data more accessible and easier to interpret, we will create a visual representation.

```
month_order <- c("August", "September", "October", "November", "December",
                 "January", "February", "March", "April", "May", "June", "July")

month_count$months <- factor(month_count$months, levels = month_order)

ggplot(data = month_count, aes(x = months, y = row_count, fill = member_casual)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_y_continuous(labels = scales::comma,
                     breaks = seq(0, max(month_count$row_count), by = 50000)) +
  labs(title = "Monthly Ride Counts: Casual Riders vs Members",
       x = "Month",
       y = "Number of Rides",
       fill = "Rider Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45)) +
  facet_wrap(~ member_casual)
```



The data shows a clear seasonal trend where casual riders have the highest usage in the summer months, peaking in July, and dropping significantly during the winter. On the other hand, member usage is more consistent throughout the year, with a notable peak also in July. This suggests that casual riders tend to

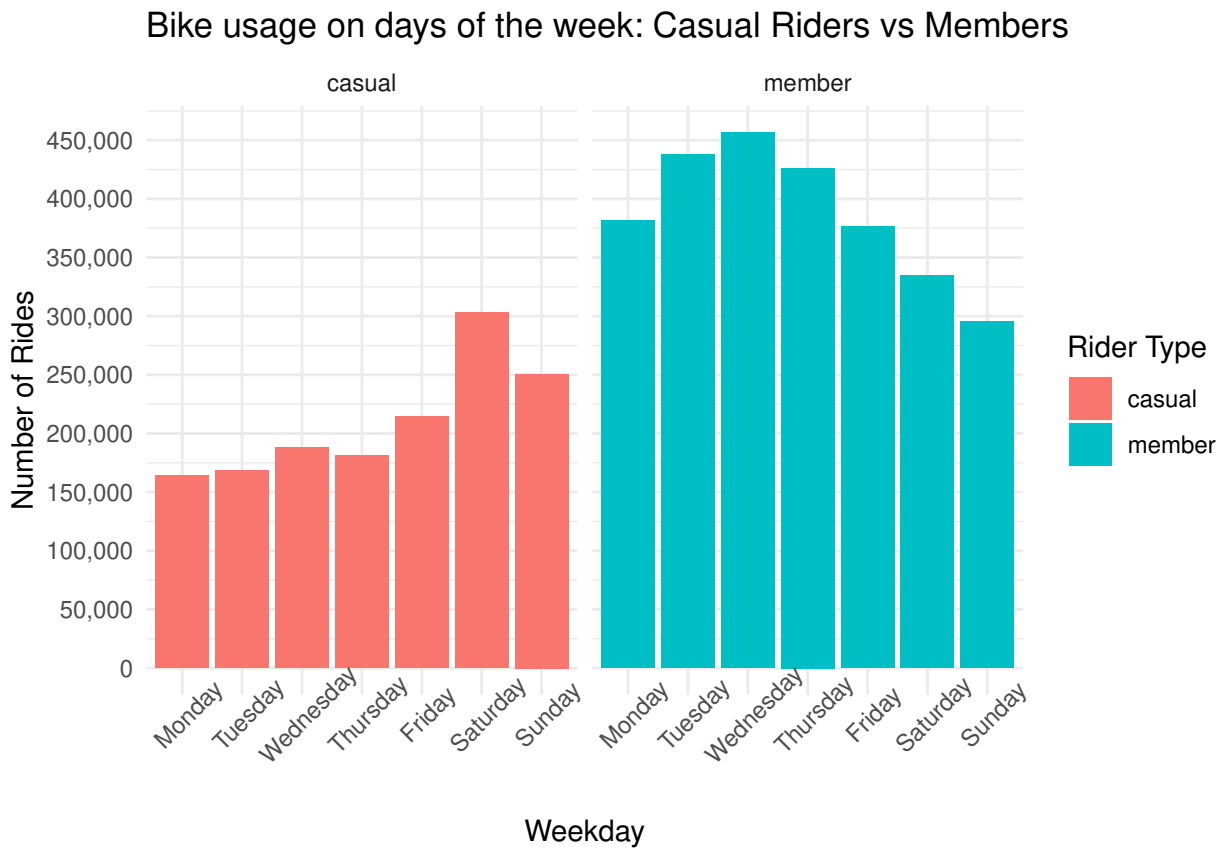
use the service more during warmer months, while members have a steadier usage pattern regardless of the season.

B. Evaluation of the weekly ride count We can analyze the differences in bike usage between casual riders and members across different days of the week using our data. We'll work with the `cyclistic_clean_data`, extract and group the `weekday` and `member_casual` columns, then create a new column to count the rides for each day of the week. The results will be stored in a data frame named `weekday_count`.

```
weekday_count = cyclistic_clean_data %>%
  group_by(weekday = weekday, member_casual = member_casual) %>%
  summarize(row_count = n())
```

Let's visualize this data to identify the trends in bike usage for casual riders and members across different days of the week. We will create a graph using the `weekday_count` data frame to compare the patterns for each group.

```
weekday_order <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday",
                   "Saturday", "Sunday")

weekday_count$weekday <- factor(weekday_count$weekday, levels = weekday_order)

ggplot(data = weekday_count, aes(x = weekday, y = row_count, fill = member_casual)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_y_continuous(labels = scales::comma,
                     breaks = seq(0, max(weekday_count$row_count), by = 50000)) +
  labs(title = "Bike usage on days of the week: Casual Riders vs Members",
       x = "Weekday",
       y = "Number of Rides",
       fill = "Rider Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45)) +
  facet_wrap(~ member_casual)
```



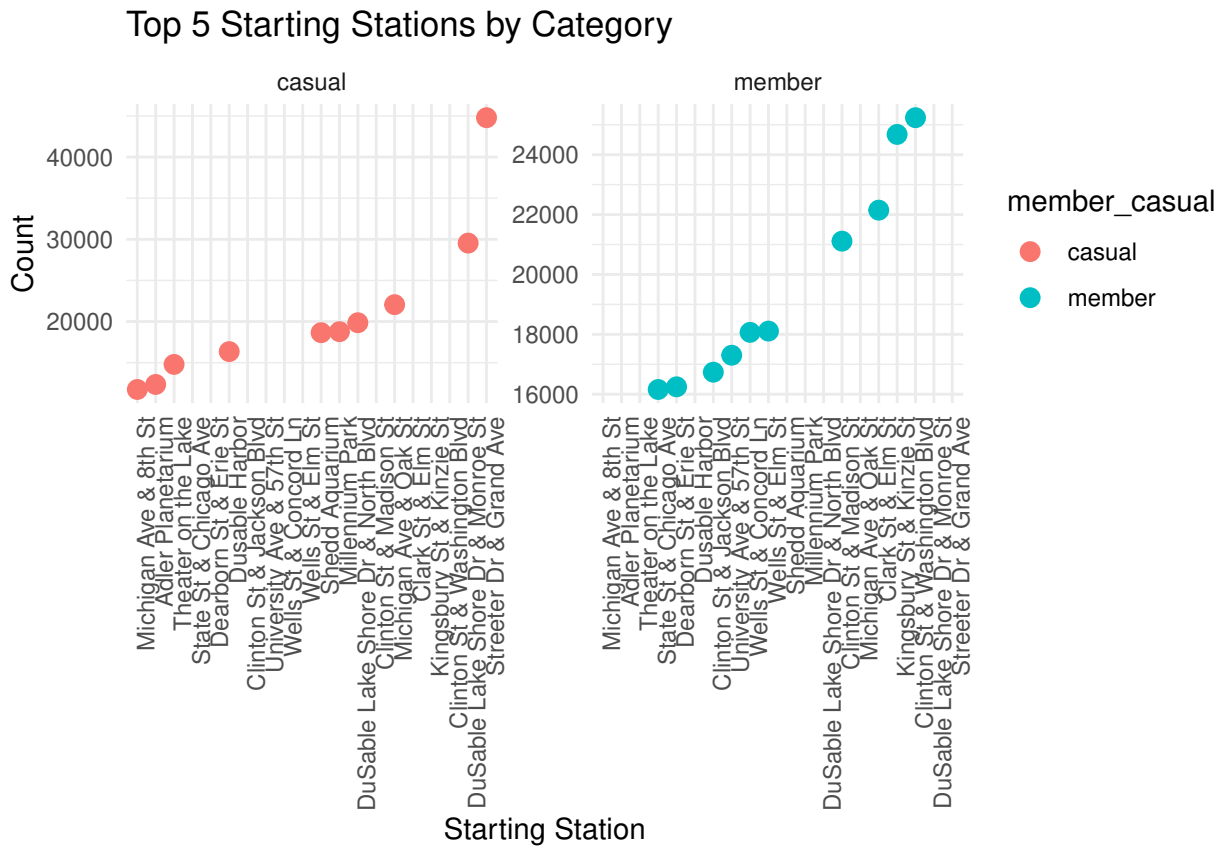
The graph reveals a clear distinction in bike usage between casual riders and members. Casual riders tend to use bikes more frequently on weekends, with the highest usage on Saturday and Sunday. In contrast, members show a consistent pattern of bike usage throughout the weekdays, peaking on Tuesday and Wednesday, with a slight decrease over the weekend. This suggests that casual riders are more likely to use bikes for leisure on weekends, while members use bikes more regularly for commuting or daily activities during the week.

C. Top 5 stations for starting and ending rides We can identify the top five starting and ending stations for both casual riders and members, which will help us uncover the popular locations for each group. To begin, we'll create a new data frame that extracts the `start_station_name`, `end_station_name`, and `member_casual` columns. This will allow us to analyze and compare the stations most frequently used by casual riders and members.

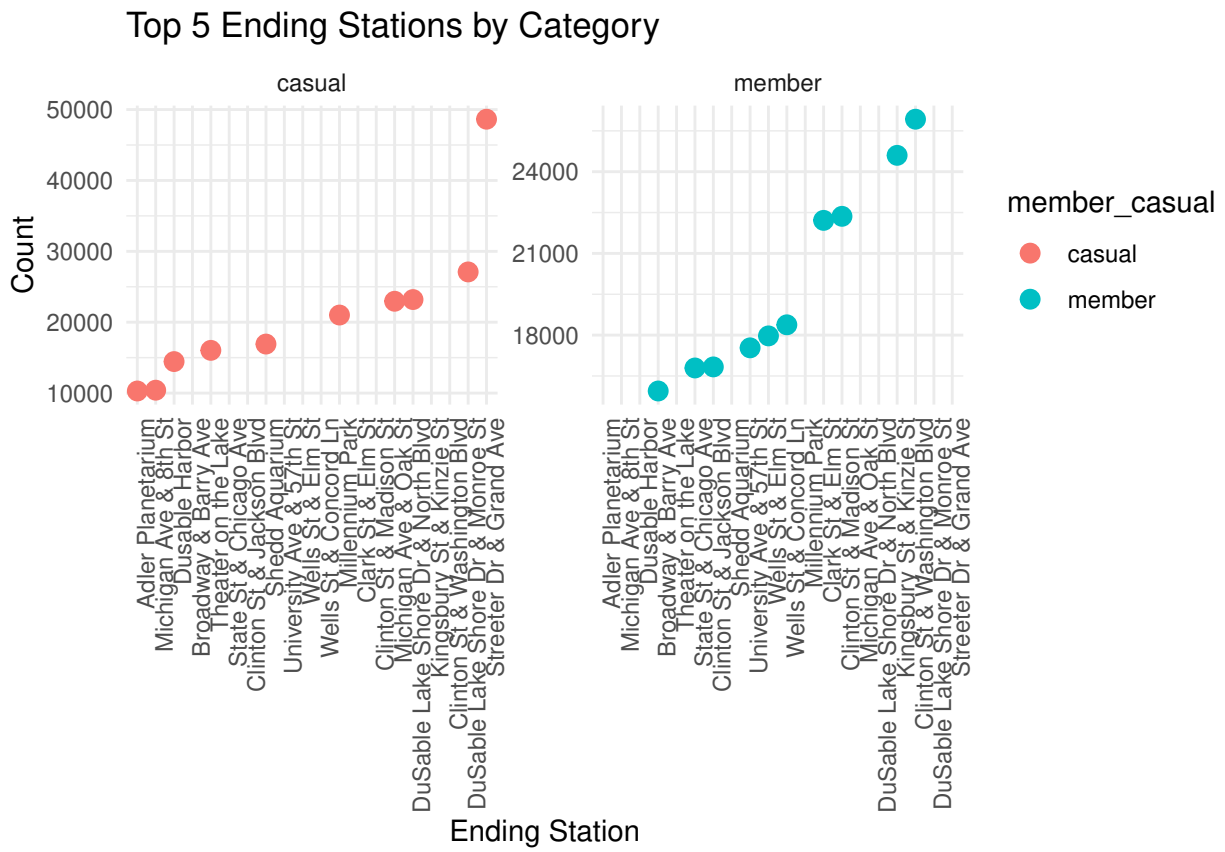
```
top_start_station <- cyclistic_clean_data %>%
  group_by(start_station_name, member_casual) %>%
  summarize(row_count = n()) %>%
  arrange(desc(row_count))

top_end_station <- cyclistic_clean_data %>%
  group_by(end_station_name, member_casual) %>%
  summarize(row_count = n()) %>%
  arrange(desc(row_count))

top_start_station_filtered <- top_start_station %>%
  group_by(member_casual) %>%
  slice_head(n = 10)
```

```r
top_end_station_filtered <- top_end_station %>%
  group_by(member_casual) %>%
  slice_head(n = 10)
```

```r
ggplot(top_start_station_filtered, aes(x = reorder(start_station_name, row_count), y = row_count, color
  geom_point(size = 3) +
  facet_wrap(~ member_casual, scales = "free_y") +
  labs(title = "Top 5 Starting Stations by Category", x = "Starting Station", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```r
ggplot(top_end_station_filtered, aes(x = reorder(end_station_name, row_count), y = row_count, color = m
  geom_point(size = 3) +
  facet_wrap(~ member_casual, scales = "free_y") +
  labs(title = "Top 5 Ending Stations by Category", x = "Ending Station", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



The charts illustrate the top 5 starting and ending stations for casual riders and annual members, categorized separately. For both starting and ending stations, it's evident that certain locations are more popular among members compared to casual riders. Notably, stations like DuSable Lake Shore Dr & Monroe St and Streeter Dr & Grand Ave are prominent among casual riders, while members frequent locations like Kingsbury St & Kinzie St and Clinton St & Washington Blvd.

D. Trip duration To compare the trip lengths of casual riders and annual members, we can analyze the average duration of rides for each group. This will help us identify any significant differences in ride lengths between the two types of riders.

```
ride_length <- cyclistic_clean_data %>%
  group_by(member_casual) %>%
  summarize(mean_trip_length = mean(trip_length))

ride_length
```

```
## # A tibble: 2 x 2
##   member_casual mean_trip_length
##   <chr>                    <dbl>
## 1 casual                    24.3
## 2 member                    12.7
```

The table displays the average trip length for two types of riders: casual and member. Casual riders have a mean trip length of approximately 24.29 minutes, while members have a shorter mean trip length of around 12.67 minutes. This suggests that casual riders tend to take longer trips compared to members.

Casual riders and members exhibit distinct usage patterns throughout the year and across different days of the week. Casual riders demonstrate the highest usage during the summer months, particularly in July, with a significant decline in winter. They are more inclined to use bikes on weekends, especially Saturdays and Sundays, indicating a preference for leisure activities. In contrast, members maintain a consistent usage pattern throughout the year, with a slight peak in July, and they primarily use bikes on weekdays, peaking on Tuesdays and Wednesdays, likely for commuting or daily activities. Additionally, casual riders prefer specific popular stations like DuSable Lake Shore Dr & Monroe St, whereas members frequent different stations, such as Kingsbury St & Kinzie St. Casual riders also tend to have longer trip durations, averaging around 24.29 minutes, compared to the 12.67-minute average for members, suggesting that casual riders use the service more for extended trips.

Act Based on the analysis of casual and member rider behaviors, here are marketing strategies to convert casual riders into annual members:

- Seasonal Membership Discounts: Since casual riders peak during the summer, offering discounted or promotional membership rates during this period could encourage them to commit to an annual membership. For example, offering a "Summer Special" that provides a discounted annual rate if they sign up during the peak months could entice casual riders to transition to members.
- Weekend Membership Perks: Casual riders tend to use bikes more on weekends, so introducing a membership tier that offers weekend-specific benefits could appeal to them. For instance, an "Active Weekender" membership could include perks like additional ride minutes, free bike rentals on weekends, or access to exclusive events, encouraging casual riders to see the value in becoming members.
- Leisure and Experience-Based Campaigns: Highlighting the benefits of membership through leisure-focused campaigns can attract casual riders who primarily use the service for recreation. Marketing strategies could include showcasing the convenience of having a membership for spontaneous weekend adventures, or featuring testimonials from members who enjoy leisurely rides with added flexibility and cost savings.
- Targeted Station Promotions: Focus on promoting membership at popular casual rider stations like DuSable Lake Shore Dr & Monroe St. Special promotions at these stations could include pop-up sign-up events, instant rewards for becoming a member, or offering a trial membership with a seamless upgrade path to an annual membership.
- Ride Time Incentives: Given that casual riders take longer trips, introduce a membership benefit that rewards longer rides, such as extending the maximum ride time without additional fees for members. This could appeal to casual riders who frequently find themselves exceeding typical ride durations, offering them a more economical and stress-free option.

Thank you for accompanying me on this analytical journey into the dynamics of bike-sharing usage. I hope you found the insights not only informative but also enjoyable to explore. This case study has been a great opportunity to delve into data and uncover actionable strategies for enhancing rider engagement.