Cyclistic Data Analysis

Seiya Kudo

2023-05-31

Introduction

Hi my name is Seiya. I'm not only a company employee but also a university student. I'm seeking a job as a data analyst so this project is my portfolio.

Scenario

I am a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

About the company

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime. Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members. Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs. Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics.

About Data

This data is for clients who used the service between January 2022 and December 2022.

Download the previous 12 months of Cyclistic trip data here. (Note: The datasets have a different name because Cyclistic is a fictional company. For the purposes of this case study, the datasets are appropriate and

will enable you to answer the business questions. The data has been made available by Motivate International Inc. under this license.

And I am already using bigquery to consolidate monthly data by quarter.

- cyclistic_df : Merged divvy_trips_2022_Q1, divvy_trips_2022_Q2, divvy_trips_2022_Q3, divvy_trips_2022_Q4
- ride_id : Customer ID
- rideable_type : Bicycle types offered in the service (classic_bike, electric_bike, docked_bike)
- started_at : Start time of use
- ended_at : End time of use
- start_station_name : Station name when you started using the service
- end_station_name : Station name when you ended using the service
- start_station_id : Station id when you started using the service
- end_station_id : Station id when you ended using the service
- start_lat : Latitude at the start of service use
- start_lng : Longitude at the start of service use
- end_lat : Latitude at the end of service use
- end_lng : Longitude at the end of service use
- member_casual : Membership type (casual, member)
- month : Month name of date of use
- day : Day name of the week of use
- usage_time_m: Time of use in minutes

About analysis

- Compare monthly usage rates
- Compare daily utilization rates
- Compare average daily usage_time_m_m
- Compare the utilization of different types of bicycles
- Showing the top 10 stations used by casual members

※ In this case, we have not done all the analysis in order to make it easier to view as a portfolio.

Data clean and create new rows

1. clean_names(): This is a function from the janitor package that cleans up the column names of the dataframe by making them all lowercase and removing inappropriate characters.
2. mutate(started_at = lubridate::ymd_hms(started_at)): This uses the ymd_hms function from the lubridate package to convert the started_at column data into a POSIXct data type, which is a type of data that handles dates and times.
3. mutate(ended_at = lubridate::ymd_hms(ended_at)): Similarly, this converts the ended_at column data into a POSIXct data type.
4. mutate(month = month(started_at, label = TRUE)): This uses the month function from lubridate to extract the month from the started_at column and create a new month column. The label = TRUE option causes it to return the name of the month (e.g., January, February, etc.) instead of a number.
5. mutate(day = wday(started_at, label = TRUE)): This uses the wday function from lubridate to extract the day of the week from the started_at column and create a new day column. The label = TRUE option causes it to return the name of the day of the week (e.g., Sunday, Monday, etc.) instead of a number.
6. mutate(usage_time_m = as.numeric(difftime(ended_at, started_at, units = "mins"))): This uses the difftime function to calculate the difference between ended_at and started_at (i.e., the usage time),

and stores the result in minutes in a new usage_time_m column. The as.numeric function is used here to convert the result into a numeric type.

```
cyclistic_df <- cyclistic_df %>%
  clean_names() %>%
  mutate(started_at = lubridate::ymd_hms(started_at)) %>%
  mutate(ended_at = lubridate::ymd_hms(ended_at)) %>%
  mutate(month = month(started_at, label = TRUE)) %>%
  mutate(day = wday(started_at, label = TRUE)) %>%
  mutate(usage_time_m = as.numeric(difftime(ended_at, started_at, units = "mins")))
```

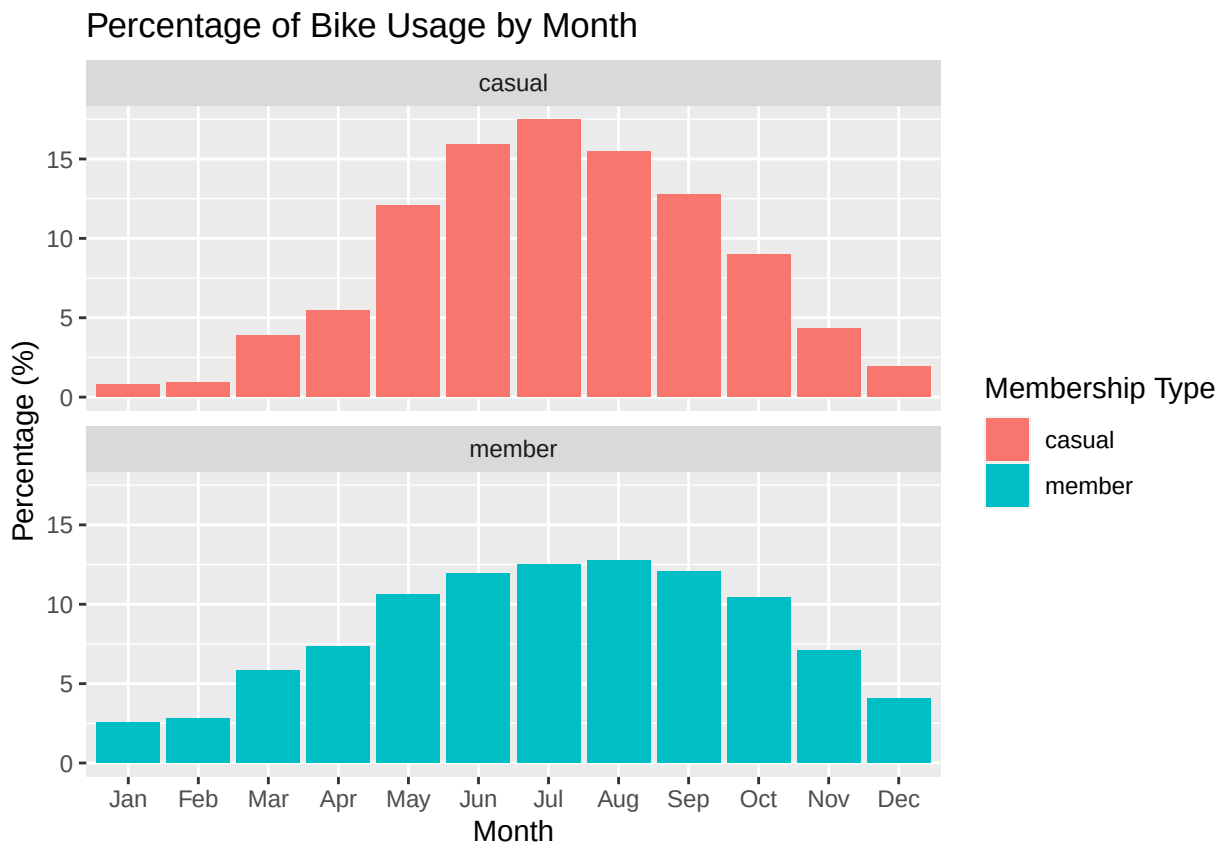
Data head

```
head(cyclistic_df)
```

```
## # A tibble: 6 x 16
##   ride_id          rideable_type started_at          ended_at
##   <chr>            <chr>         <dttm>              <dttm>
## 1 9A40ADC1C8AA2AC9 electric_bike 2022-02-10 18:33:36 2022-02-10 19:16:30
## 2 95491BF7DF1A1514 electric_bike 2022-02-19 08:29:16 2022-02-19 08:29:26
## 3 D50A4C7276D9194A electric_bike 2022-02-08 15:16:55 2022-02-08 15:36:13
## 4 FD3AEAA09333C42A electric_bike 2022-02-10 14:05:54 2022-02-10 14:23:13
## 5 61540135D7E8C96F electric_bike 2022-02-01 08:31:33 2022-02-01 08:34:07
## 6 46EF4712670222DD electric_bike 2022-02-21 13:48:14 2022-02-21 14:11:21
## # i 12 more variables: start_station_name <chr>, start_station_id <chr>,
## #   end_station_name <chr>, end_station_id <chr>, start_lat <dbl>,
## #   start_lng <dbl>, end_lat <dbl>, end_lng <dbl>, member_casual <chr>,
## #   month <ord>, day <ord>, usage_time_m <dbl>
```

Visualize bike usage rate by day of the week for each membership type

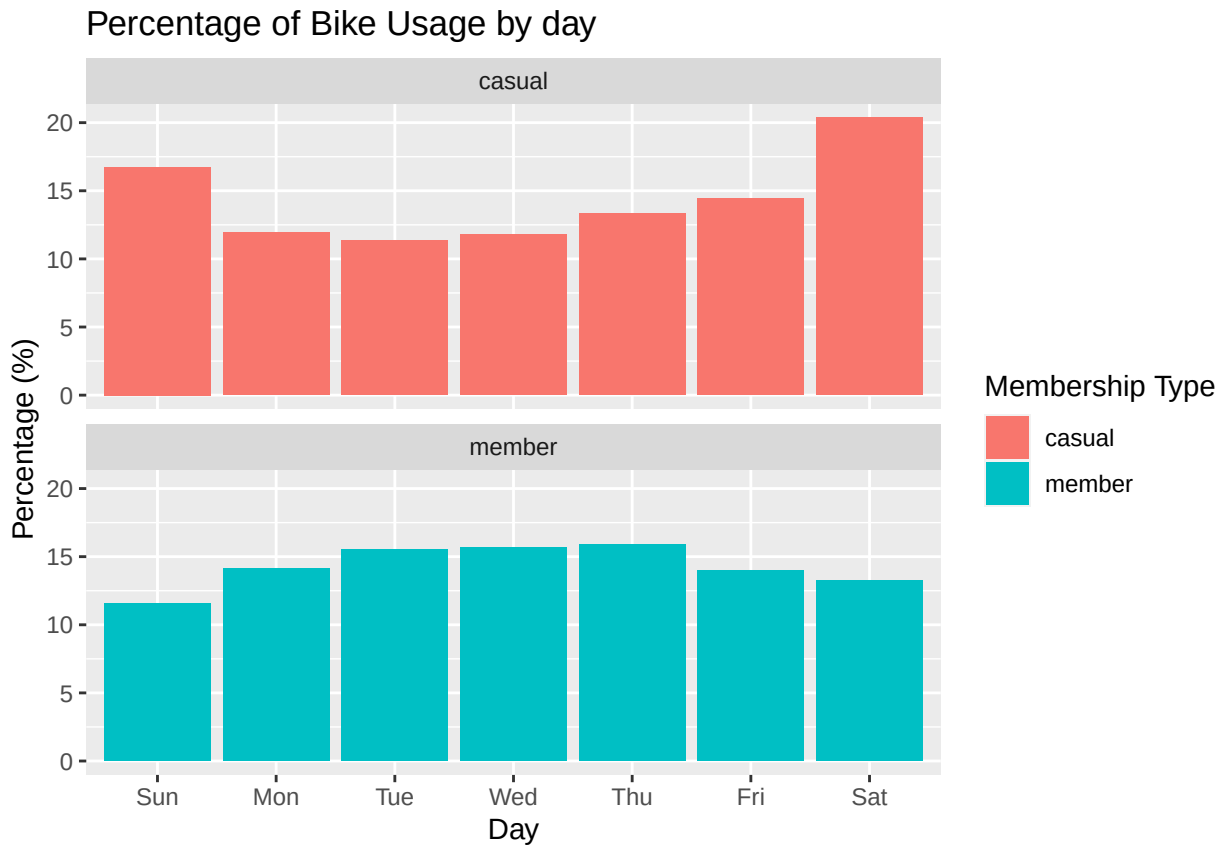
```
cyclistic_df %>%
  select(member_casual, month) %>%
  group_by(member_casual, month) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(member_casual) %>%
  mutate(percentage = count / sum(count) * 100) %>%
  ggplot(aes(x = month, y = percentage, fill = member_casual)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~member_casual, ncol = 1) +
  labs(title = "Percentage of Bike Usage by Month",
       x = "Month",
       y = "Percentage (%)",
       fill = "Membership Type")
```



Shows no significant difference in monthly utilization by membership.

Visualize bike usage rate by day of the week for each membership type

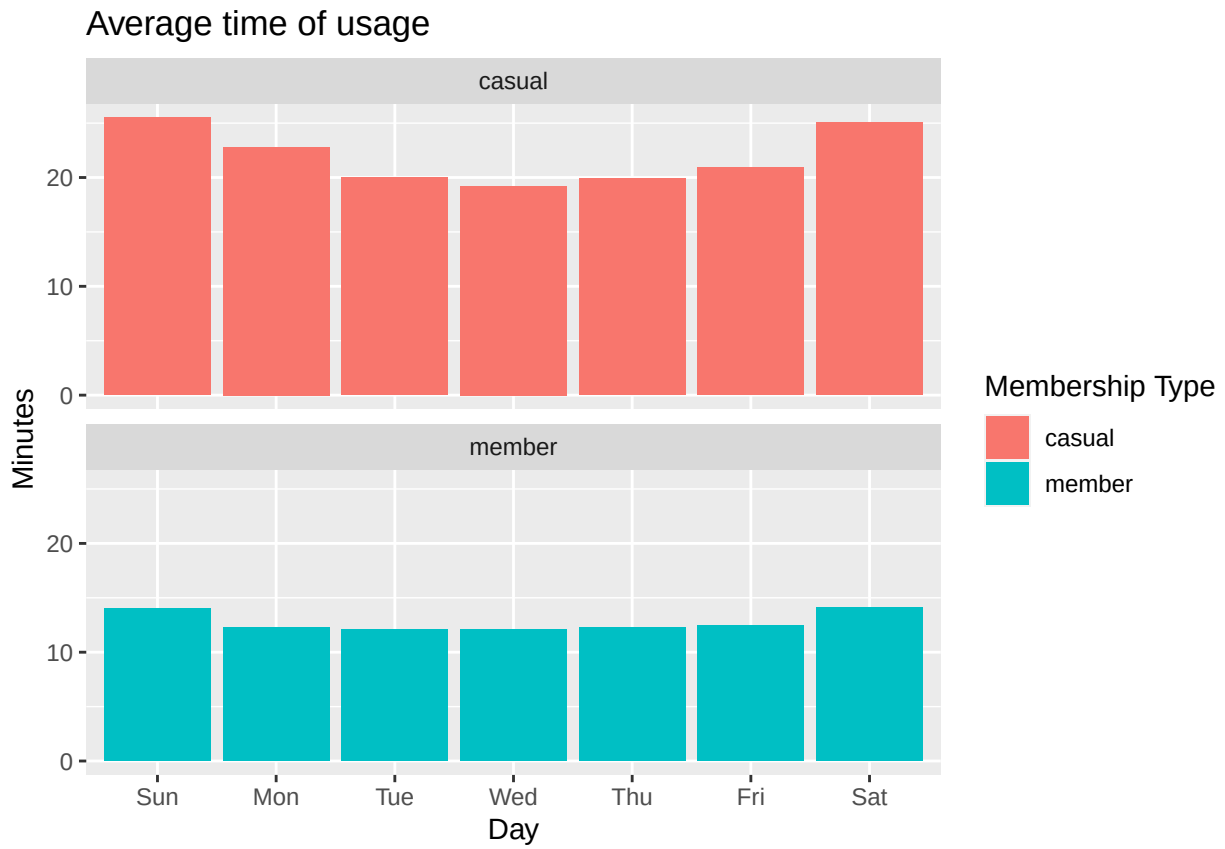
```
cyclistic_df %>%
  select(member_casual, day) %>%
  group_by(member_casual, day) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(member_casual) %>%
  mutate(percentage = count / sum(count) * 100) %>%
  ggplot(aes(x = day, y = percentage, fill = member_casual)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~member_casual, ncol = 1) +
  labs(title = "Percentage of Bike Usage by day",
       x = "Day",
       y = "Percentage (%)",
       fill = "Membership Type")
```



Casual members have a higher usage rate on weekends, while annual members have a higher usage rate on weekdays, suggesting that casual members use the service for leisure on weekends and annual members use the service for commuting on weekdays.

Visualize average travel time by day

```
cyclistic_df %>%
  select(member_casual, day, usage_time_m) %>%
  filter(usage_time_m > 1, usage_time_m <= 1440) %>%
  group_by(member_casual, day) %>%
  summarise(avg_travel_time = mean(usage_time_m), .groups = "drop") %>%
  ggplot(aes(x = day, y = avg_travel_time, fill = member_casual)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~member_casual, ncol = 1) +
  labs(title = "Average time of usage",
       x = "Day",
       y = "Minutes",
       fill = "Membership Type")
```

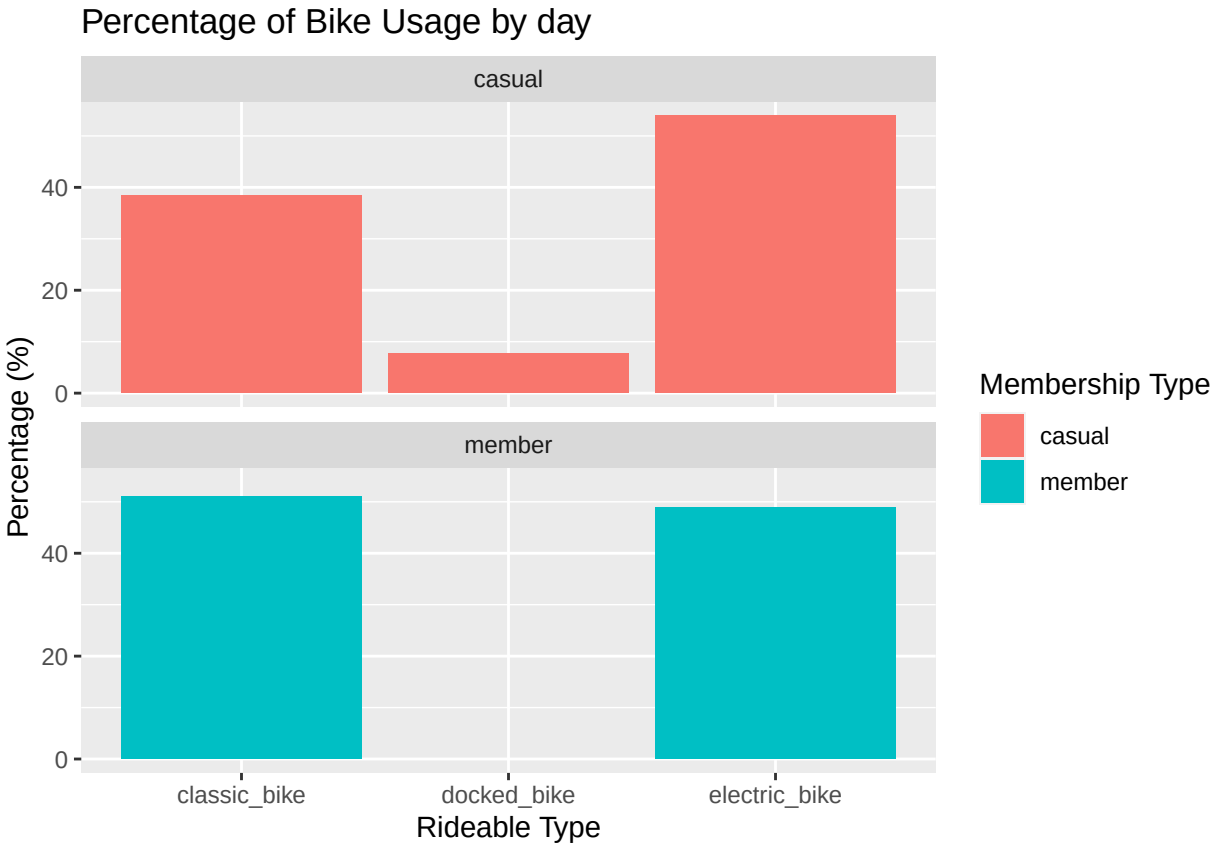


The average time spent by casual members is longer than that of annual members, around 20 minutes on any day of the week; it is assumed that annual members use this service for commuting, but it is not known what casual members use it for, so additional research is needed.

※ To avoid bias, we limit the filtering of data to use of more than 1 minute and less than 24 hours.

Visualize bike usage rate by day of the week for each membership type

```
cyclistic_df %>%
  select(member_casual, rideable_type) %>%
  group_by(member_casual, rideable_type) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(member_casual) %>%
  mutate(percentage = count / sum(count) * 100) %>%
  ggplot(aes(x = rideable_type, y = percentage, fill = member_casual)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~member_casual, ncol = 1) +
  labs(title = "Percentage of Bike Usage by day",
       x = "Rideable Type",
       y = "Percentage (%)",
       fill = "Membership Type")
```

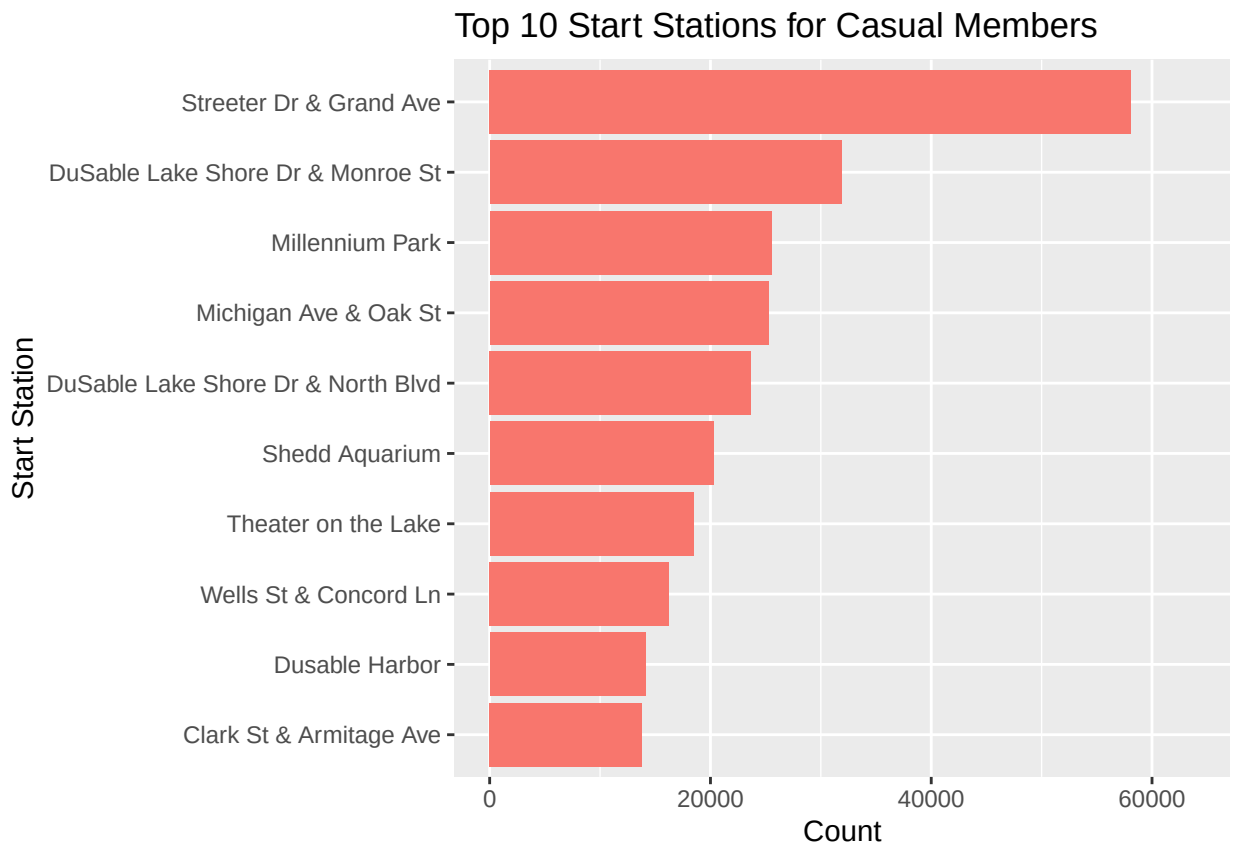


Casual members utilize docked bikes approximately 8% of the time, whereas annual members do not use docked bikes at all (0%). Given the assumption that docked bike users are likely family members, offering family-oriented plans and advertisements would be highly effective.

Visualize casual member top 10 start station

```
# Save the intermediate results to a variable
casual_top_stations <- cyclistic_df %>%
  select(member_casual, start_station_name) %>%
  drop_na() %>%
  filter(member_casual == "casual") %>%
  group_by(start_station_name) %>%
  summarise(count = n(), .groups = "drop") %>%
  top_n(10, wt = count)

# Use the variable for plotting
casual_top_stations %>%
  ggplot(aes(x = reorder(start_station_name, count), y = count)) +
  geom_bar(stat = 'identity', fill = "#F8766D") +
  coord_flip() +
  ylim(0, max(casual_top_stations$count) * 1.1) +  # Add 10% to the maximum count
  labs(title = "Top 10 Start Stations for Casual Members",
       x = "Start Station",
       y = "Count",
       fill = "Membership Type")
```

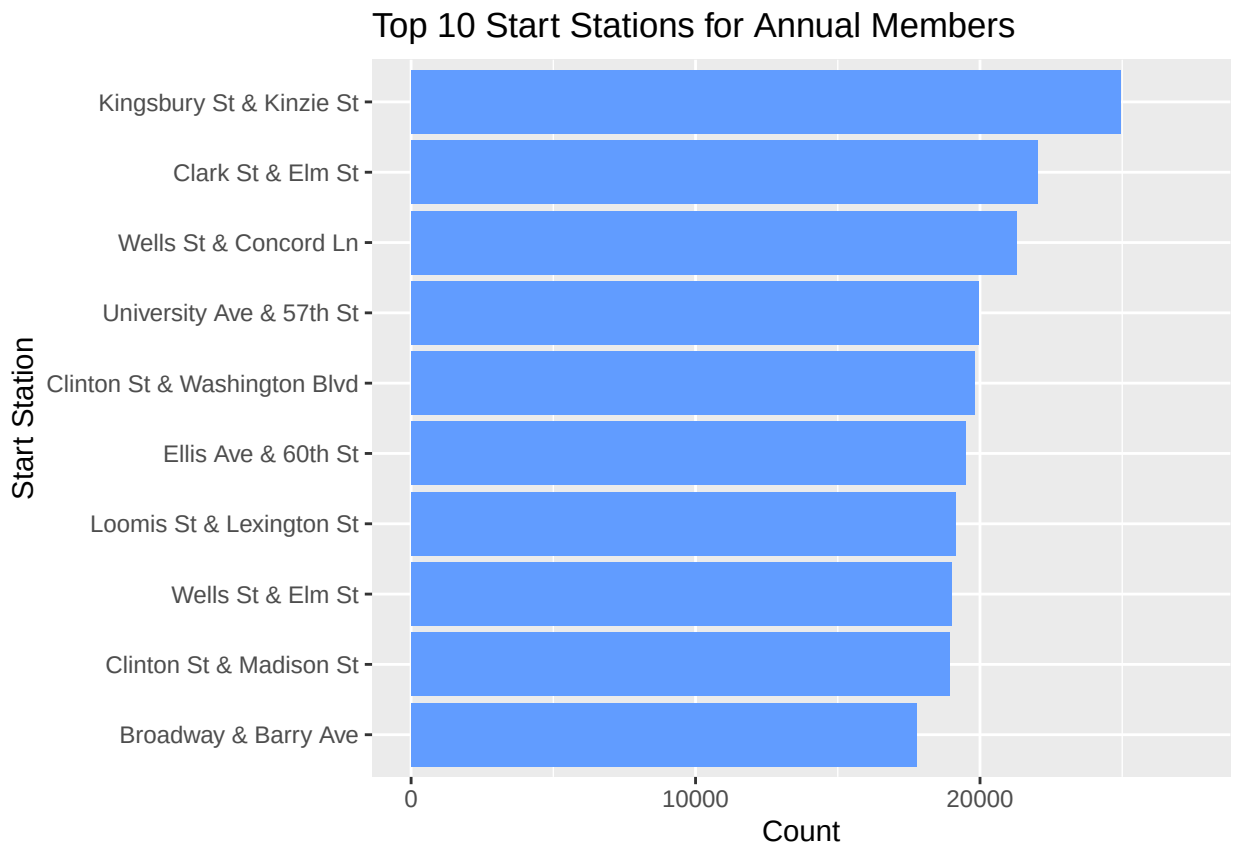


The Streeter Dr & Grand Ave station, which has the highest use, is twice as far away from the second place, so it would be helpful to check the area around this station to see why it is used, and to increase the use of other stations.

Visualize annual member top 10 start station

```
member_top_stations <- cyclistic_df %>%
  select(member_casual, start_station_name) %>%
  drop_na() %>%
  filter(member_casual == "member") %>%
  group_by(start_station_name) %>%
  summarise(count = n(), .groups = "drop") %>%
  top_n(10, wt = count)

# Use the variable for plotting
member_top_stations %>%
  ggplot(aes(x = reorder(start_station_name, count), y = count)) +
  geom_bar(stat = 'identity', fill = "#619CFF") +
  coord_flip() +
  ylim(0, max(member_top_stations$count) * 1.1) +  # Add 10% to the maximum count
  labs(title = "Top 10 Start Stations for Annual Members",
       x = "Start Station",
       y = "Count",
       fill = "Membership Type")
```



Although there is not as much difference in the top 10 as in Casual members, a more detailed discussion can be obtained by examining these stations.

Conclution

Data analysis did not reveal any significant differences in monthly usage rates by membership category. However, there are significant differences in usage patterns and duration of use. Casual members are more likely to use the service on weekends, with the average usage time often exceeding 20 minutes. On the other hand, annual fee members prefer weekdays and seem to use it to commute to work or school.

In addition, about 8% of casual members use docked bikes, suggesting that they are intended for family use. However, this is not the case at all for annual members, so it would be good to have a family-friendly plan to convert casual members to annual members.

The most heavily used station is also Streeter Dr & Grand Ave, far ahead of the second place and below. A thorough study of these stations would reveal a more nuanced discussion and strategies to improve usage.