Google Data Analytics Capstone

Case study: Cyclistic bike-share company analysis

Marcelo Hirata

2022-07-24

1 Introduction

1.1 About the company Cyclistic bike-share

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.

Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments with 3 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**.

1.2 Scenario

You are 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.

1.3 Stakeholders

- **Lily Moreno**: The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program.
- **Cyclistic marketing analytics team**: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. You joined this team six months ago and have been busy learning about Cyclistic's mission and business goals — as well as how you, as a junior data analyst, can help Cyclistic achieve them.
- **Cyclistic executive team**: The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program. | |

1.3 Business task

Three questions will guide 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?

In this analysis, the first question will be addressed: How do annual **members** and **casual** riders use Cyclistic bikes differently?

2 Data preparation and process

2.1 Data sources

The data used in this analysis comprehends the last 12 months of Cyclistic's historical trip data (from 2021/06 to 2022/05), stored in monthly CSV files with structured wide datasets. *Data source* (https://divvy-tripdata.s3.amazonaws.com/index.html)

This is a public dataset that can be used to explore how different customer types are using Cyclistic bikes. *Data license* (https://ride.divvybikes.com/data-license-agreement)

Prepare the environment and Import the data to R

```
library(tidyverse) # data manipulation
library(janitor) # data cleaning
library(skimr) # summary statistics
library(lubridate) # work with date-times and time-spans
library(ggplot2) # visualize data
library(leaflet) # interactive maps
library(hydroTSM) # time series used in hydrology
library(geosphere) # geographic applications
library(scales) # graphical scales
library(stringr) # string operations
```

```
trip_2106 <- read_csv("202106-divvy-tripdata.csv")
trip_2107 <- read_csv("202107-divvy-tripdata.csv")
trip_2108 <- read_csv("202108-divvy-tripdata.csv")
trip_2109 <- read_csv("202109-divvy-tripdata.csv")
trip_2110 <- read_csv("202110-divvy-tripdata.csv")
trip_2111 <- read_csv("202111-divvy-tripdata.csv")
trip_2112 <- read_csv("202112-divvy-tripdata.csv")
trip_2201 <- read_csv("202201-divvy-tripdata.csv")
trip_2202 <- read_csv("202202-divvy-tripdata.csv")
trip_2203 <- read_csv("202203-divvy-tripdata.csv")
trip_2204 <- read_csv("202204-divvy-tripdata.csv")
trip_2205 <- read_csv("202205-divvy-tripdata.csv")
```

Check whether the set of data.frames are row-bindable and unite them

```
if(compare_df_cols_same(trip_2106, trip_2107, trip_2108, trip_2109, trip_2110, trip_2111, trip_2112, trip_2201, trip_2202, trip_2203, trip_2204, trip_2205))
  {
  all_trips <- rbind(trip_2106, trip_2107, trip_2108, trip_2109, trip_2110, trip_2111, trip_2112, trip_2201, trip_2202, trip_2203, trip_2204, trip_2205)
  } else {
    print("Check the variables")
  }
```

2.2 Data description and Cleansing

Getting familiar with the dataset

```
skim_without_charts(all_trips)
```

Data summary

Name	all_trips
Number of rows	5860776
Number of columns	13
_______________________	
Column type frequency:	
character	7
numeric	4
POSIXct	2
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1.00	16	16	0	5860776	0
rideable_type	0	1.00	11	13	0	3	0
start_station_name	823167	0.86	3	53	0	1105	0
start_station_id	823164	0.86	3	44	0	1063	0
end_station_name	878338	0.85	9	53	0	1112	0
end_station_id	878338	0.85	3	44	0	1068	0
member_casual	0	1.00	6	6	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
start_lat	0	1	41.90	0.05	41.64	41.88	41.90	41.93	45.64
start_lng	0	1	-87.65	0.03	-87.84	-87.66	-87.64	-87.63	-73.80
end_lat	5036	1	41.90	0.05	41.39	41.88	41.90	41.93	42.17
end_lng	5036	1	-87.65	0.03	-88.97	-87.66	-87.64	-87.63	-87.49

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
started_at	0	1	2021-06-01 00:00:38	2022-05-31 23:59:56	2021-09-23 17:33:23	4896834
ended_at	0	1	2021-06-01 00:06:22	2022-06-02 11:35:01	2021-09-23 17:49:29	4893478

```
head(all_trips)
```

```
## # A tibble: 6 × 13
##   ride_id rideable_type started_at          ended_at            start_station_n…
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 99FEC9… electric_bike 2021-06-13 14:31:28 2021-06-13 14:34:11 <NA>
## 2 06048D… electric_bike 2021-06-04 11:18:02 2021-06-04 11:24:19 <NA>
## 3 959806… electric_bike 2021-06-04 09:49:35 2021-06-04 09:55:34 <NA>
## 4 B03C0F… electric_bike 2021-06-03 19:56:05 2021-06-03 20:21:55 <NA>
## 5 B9EEA8… electric_bike 2021-06-04 14:05:51 2021-06-04 14:09:59 <NA>
## 6 62B943… electric_bike 2021-06-03 19:32:01 2021-06-03 19:38:46 <NA>
## # … with 8 more variables: 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>
```

The data set has a record for each trip made by the users in the period mentioned. In total, there are 5,860,776 observations (trip records) and 13 variables (characteristics).Because of its size, I'll be using the R programming language to conduct the analysis.

One of the dataset limitations is the absence of the user's data. It'd be very useful to know at least their quantity or some demographic data, but as it hasn't been provided, this analysis will focus only on the trip's characteristics.

- Variables description and evaluation:
  - **ride_id** – Each trip unique ID containing 16 characters. In this variable, there are no missing values
  - **rideable_type** – The types of bicycles. It has 3 unique and no missing values.
  - **start_at** and **ended_at** – The date and time in which the rides start and end. They are in date-time format and there are no missing values.
  - **start_station_name**, **start_station_id**, **end_station_name** and **end_station_id** – The names of the variables explain themselves. There are more than 800,000 missing values and more than 1000 unique values, which is not consistent with the number of available stations (692). These variables are not reliable and can't be used in the analysis.
  - **start_lat**, **start_lng**, **end_lat** and **end_lng** – The coordinates(latitude and longitude) of the rides start and end location. There are only 5036 missing values at the trip ending coordinates, so it won't affect the analysis.

Exclude the variables that won't be used in this analysis.

```
all_trips <- select(all_trips, 'ride_id', "rideable_type", "started_at", "ended_at", "m
ember_casual",
                    "start_lat", "start_lng", "end_lat", "end_lng")
```

Create a variable with the trip duration in seconds, and verificate if there are negative values.

```
all_trips <- mutate(all_trips, "trip_dur" = difftime(ended_at, started_at, units = "sec
s"))
```

```
summary(all_trips$trip_dur < 0)
```

```
##    Mode   FALSE    TRUE
## logical 5860637     139
```

- There are only 139 trips with inconsistent start and end times, their exclusion won't affect the analysis.

Exclude observations with negative trip duration values.

```
all_trips <- subset(all_trips, trip_dur > 0)
```

Verify if the data is clean.

```
summary(all_trips$trip_dur < 0)
```

```
##    Mode   FALSE
## logical 5860130
```

2.3 Wrangle

Transform the data to make it more accessible, simplify the code and make it more readable.

Rename columns for more meaningful names

```
all_trips <- all_trips %>%
rename(trip_id = ride_id, bike_type = rideable_type)
```

Create variables for year season, months, day_of_week, time of day, and trip distance

```
hr <- hour(all_trips$started_at)
all_trips <- mutate(all_trips, "year_season" = time2season(all_trips$started_at, out.fm
t = "seasons", type="default"),
                    "month" = month(started_at, label = TRUE, abbr = TRUE),
                    "day_of_week" = weekdays(started_at, abbreviate = TRUE),
                    "day_time" = case_when(hr > 6 & hr < 12 ~ "morning",
                                           hr >= 12 & hr < 16 ~ "afternoon",
                                           hr >= 16 & hr < 20 ~ "evening",
                                           TRUE ~ "night")
                    )
```

```
all_trips$year_season <- str_replace(all_trips$year_season, "autumm", "autumn")
```

```
all_trips <- all_trips %>% rowwise %>%
  mutate("trip_dist" = distm(x = c(start_lng, start_lat), y = c(end_lng, end_lat), fun
= distGeo))
```

```
summary(all_trips$trip_dist)
```

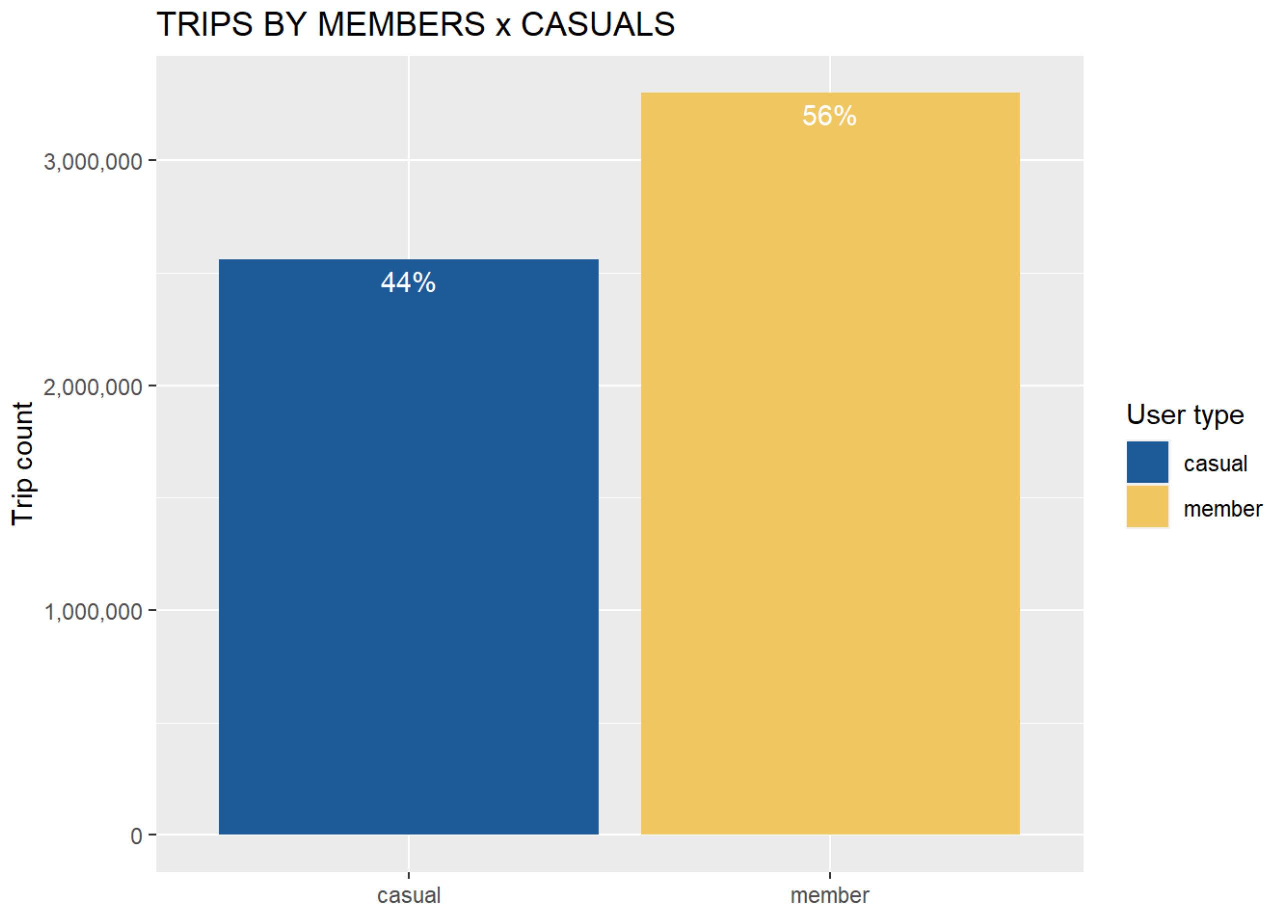
```
##        V1
##  Min.   :      0.0
##  1st Qu.:    891.4
##  Median :   1611.2
##  Mean   :   2168.8
##  3rd Qu.:   2846.6
##  Max.   :1192245.6
##  NA's   :5036
```

3 Analysis

Let’s organize and visualize the data to categorize, discover connections, find patterns, and identify themes.

Number of trips by user types - “members” and “casuals”

```
plot_col <- c("#1d5a98", "#f0c660")
legend_title <- "User type"
ggplot(all_trips, mapping = aes(x = member_casual , fill = member_casual), labels = lab
el_percent())+
  geom_bar()+
  geom_text(aes(label = scales::percent((..count..)/sum(..count..))), stat = "count", v
just = 1.5, colour = "white")+
  scale_fill_manual(legend_title, values = plot_col)+
  labs(title = "TRIPS BY MEMBERS x CASUALS",
       x = NULL,
       y = "Trip count")+
  scale_y_continuous(labels = comma)
```

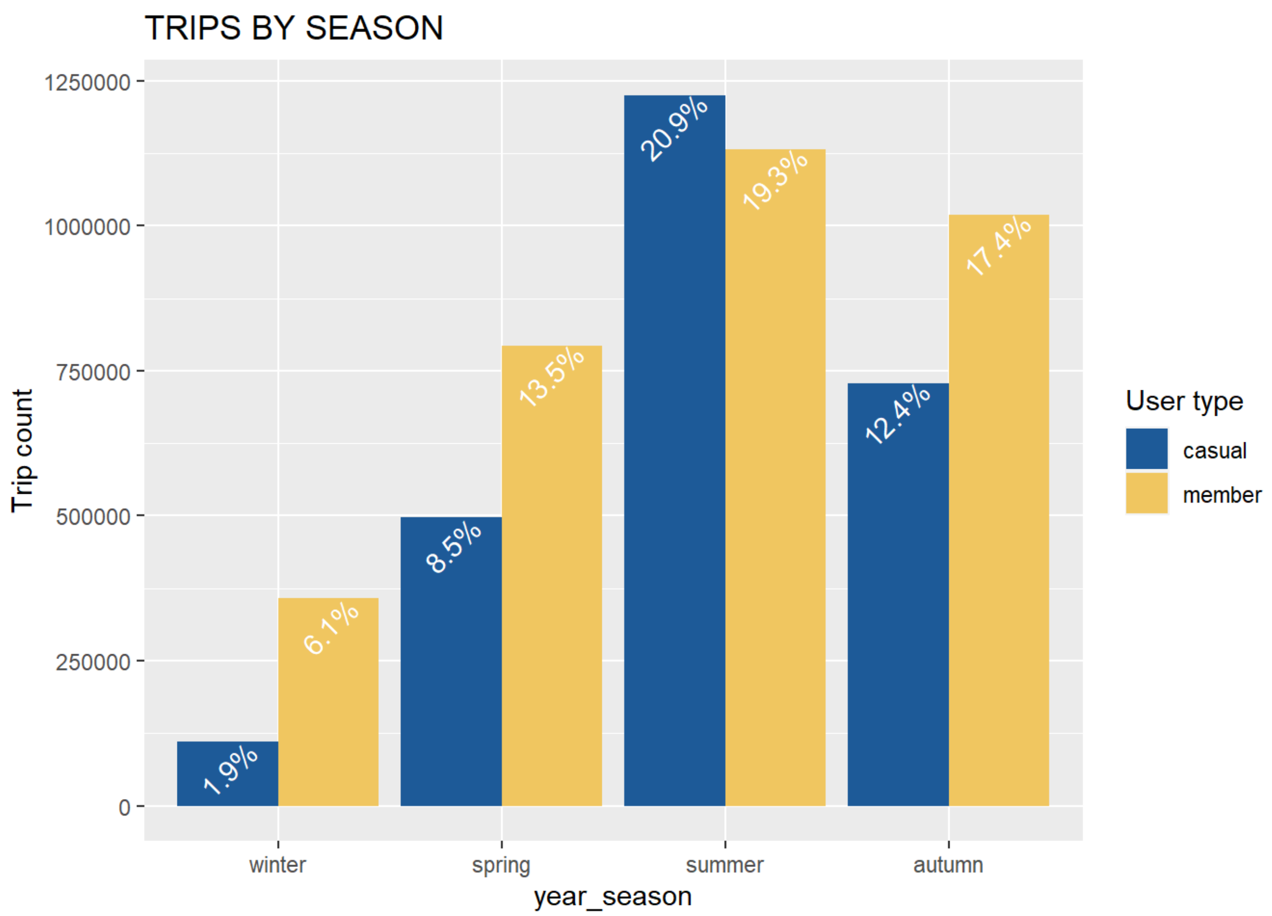


- In one year, there are more trips made by members.

3.1 Difference between “WHEN” casual and members riders use Cyclist

Number of trips by members and casuals in each season of the year

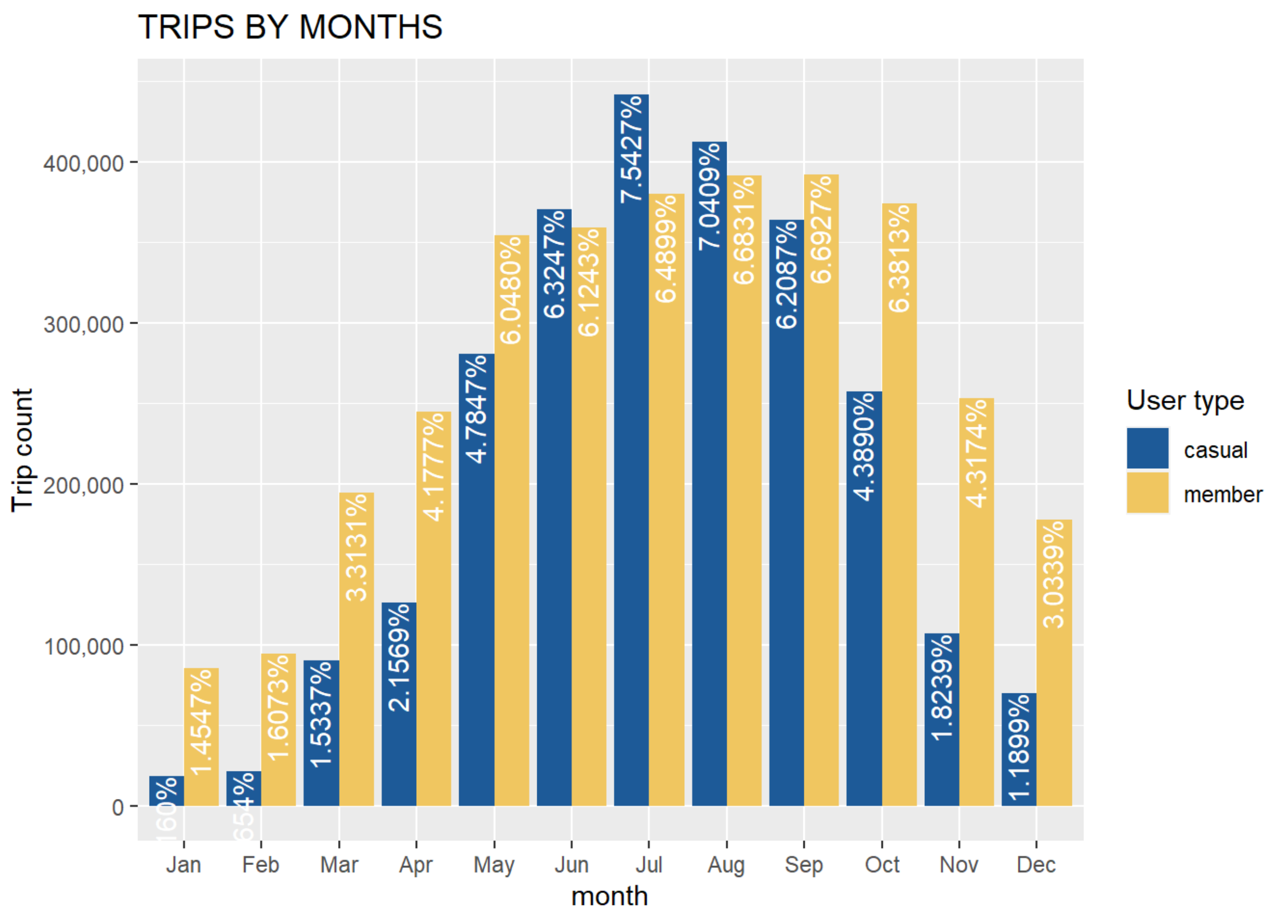
```
ggplot(data = all_trips, mapping = aes(x = year_season, fill = member_casual))+
    geom_bar(position = "dodge")+
    geom_text(aes(label = scales::percent(..count../sum(..count..))), stat = "count", v
just = 1.4, hjust = 0.8, colour = "white", position = position_dodge(width = 0.9), angl
e = 45)+
    labs(title = "TRIPS BY SEASON",
         X = NULL,
         y = "Trip count")+
  scale_fill_manual(legend_title, values = plot_col)+
  scale_x_discrete(limits = c("winter", "spring", "summer", "autumn"))
```



- The trip rates are expected to variate, but we can observe that the casual trips variate more than the members. This can indicate that casual riders don’t necessarily need to use Cyclist bikes. They may have alternative transport or are riding just for leisure.

Number of trips by members and casuals in each month

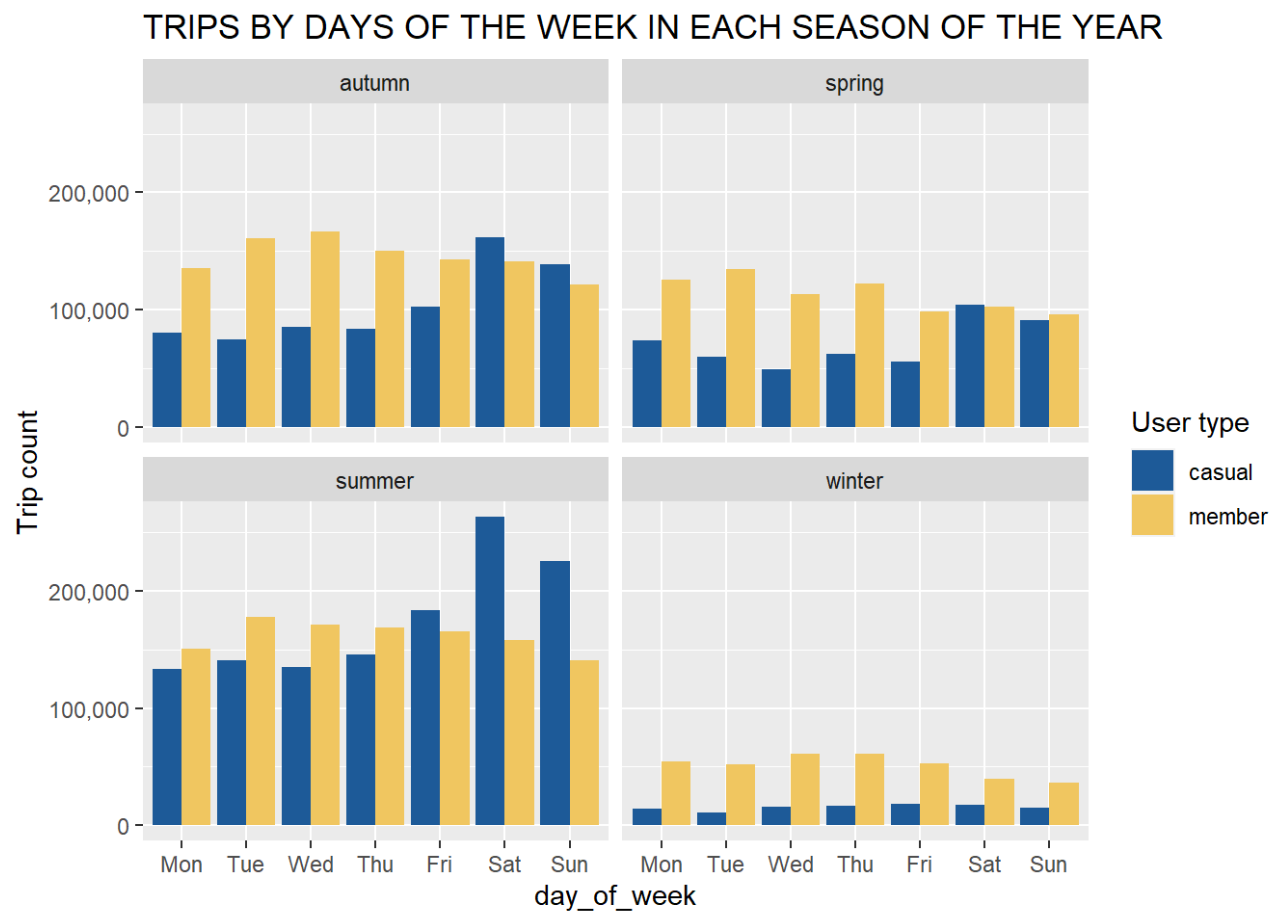
```
ggplot(data = all_trips, mapping = aes(x = month, fill = member_casual))+
    geom_bar(position = "dodge")+
    geom_text(aes(label = scales::percent(..count../sum(..count..))), stat = "count", v
just = 0.4, hjust = 1, colour = "white", position = position_dodge(width = 0.9), angle
= 90)+
    labs(title = "TRIPS BY MONTHS",
         X = NULL,
         y = "Trip count")+
  scale_fill_manual(legend_title, values = plot_col)+
  scale_y_continuous(labels = comma)
```



- Here we can see that, in the individual months, there isn't a big difference from the season trend with the casual trips topping by +- 1% in July and August.

Number of trips by members and casuals on each day of the week in different seasons

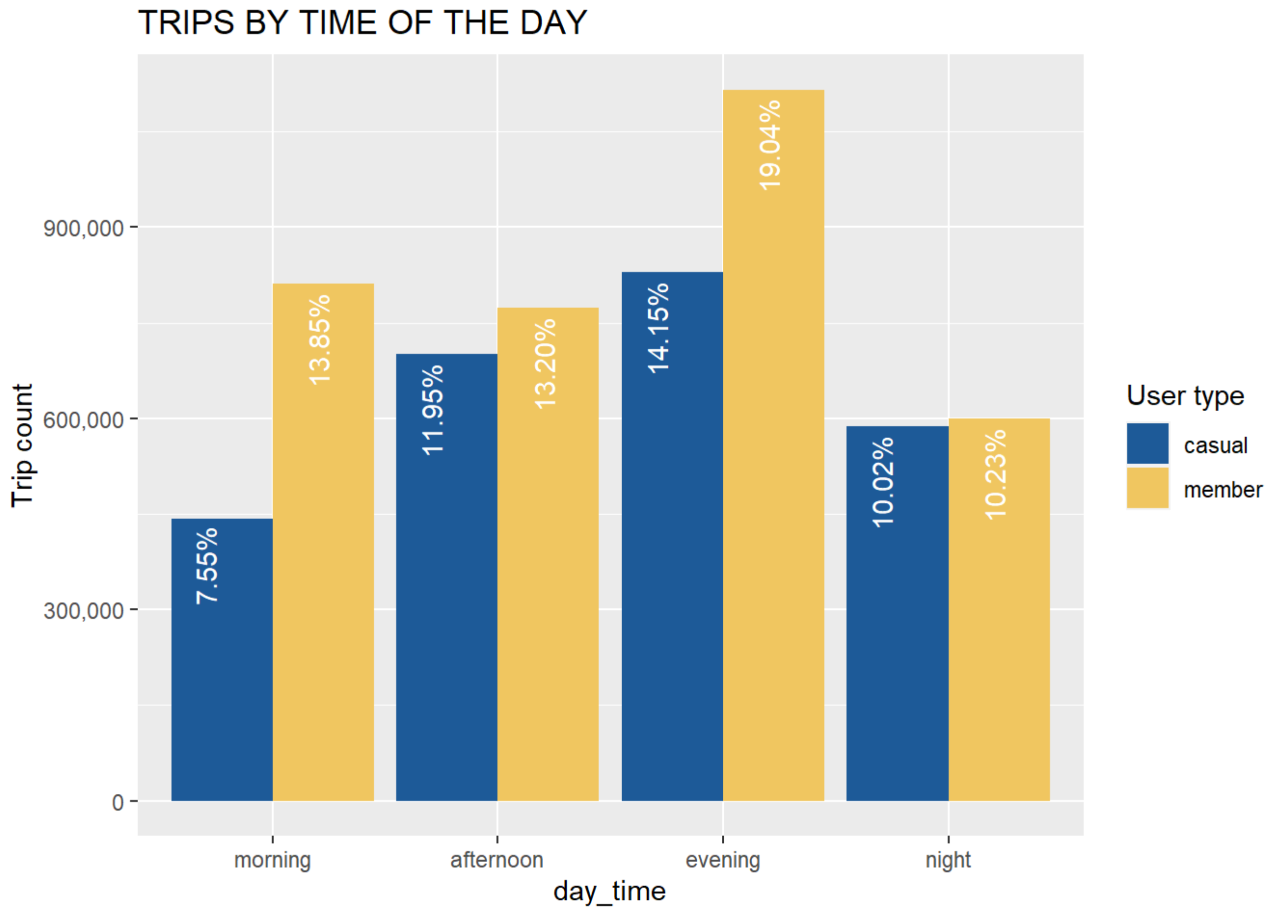
```
ggplot(data = all_trips, mapping = aes(x = day_of_week, fill = member_casual))+
  geom_bar(position = "dodge")+
  labs(title = "TRIPS BY DAYS OF THE WEEK IN EACH SEASON OF THE YEAR",
       X = NULL,
       y = "Trip count",)+
  scale_fill_manual(legend_title, values = plot_col)+
  scale_x_discrete(limits = c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"))+
  scale_y_continuous(labels = comma)+
  facet_wrap(~year_season)
```



- Looking at the data grouped by the days of the week, we can see that the casual clients' trips surpass the members at the weekends not only in the summer but also in the autumn and spring Saturdays. This confirms that casual clients use Cyclistic in their leisure time and members use it for everyday routine.

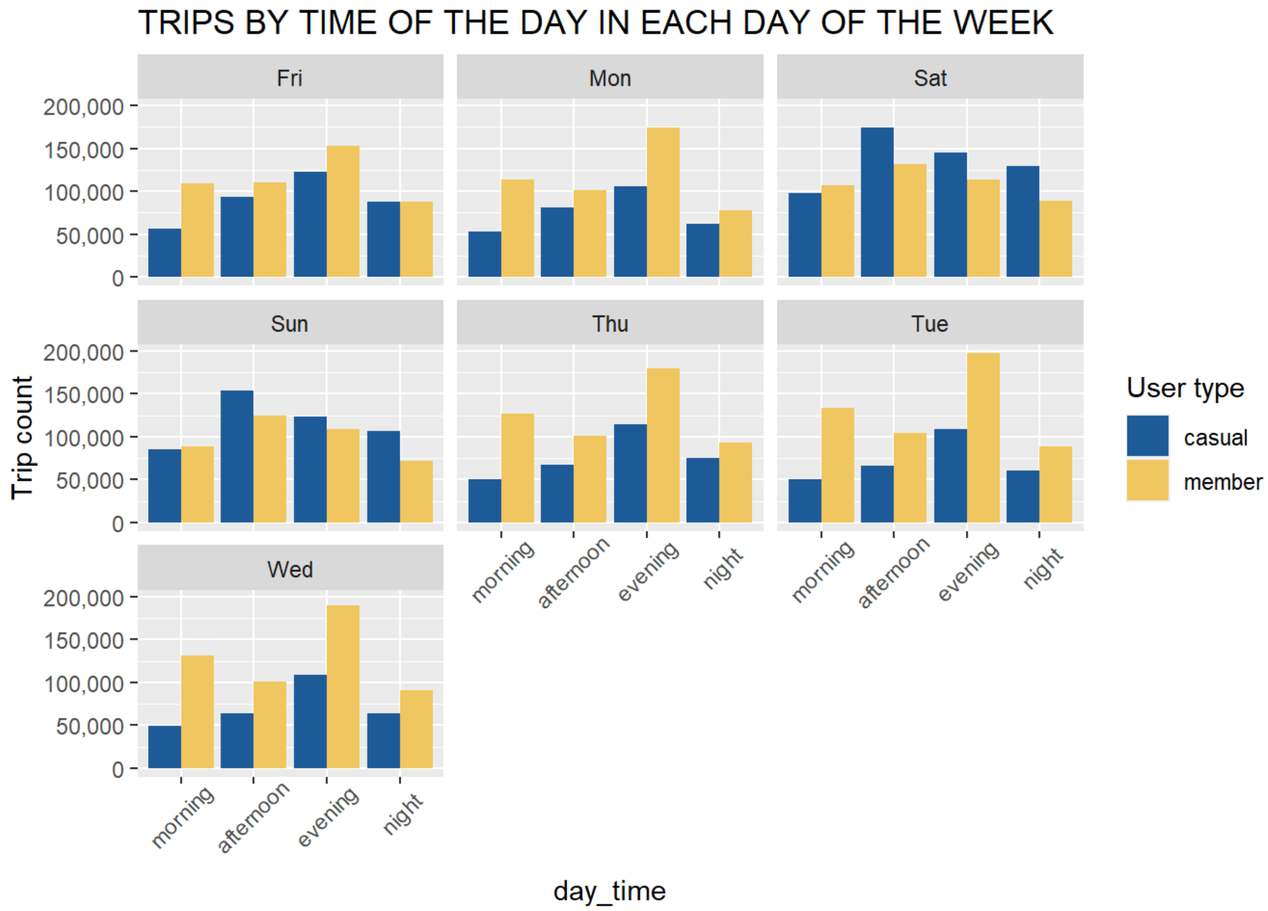
Number of trips by members and casuals by the time of the day

```
ggplot(data = all_trips, mapping = aes(x = day_time, fill = member_casual))+
  geom_bar(position = "dodge")+
  geom_text(aes(label = scales::percent(..count../sum(..count..))), stat = "count", vju
st = 0, hjust = 1.1, colour = "white",
            angle = 90 , position = position_dodge(width = 1))+
  labs(title = "TRIPS BY TIME OF THE DAY",
       X = NULL,
       y = "Trip count")+
  scale_fill_manual(legend_title, values = plot_col)+
  scale_x_discrete(limits = c("morning","afternoon","evening", "night"))+
  scale_y_continuous(labels = comma)
```



- While the trips made by casuals variate 3% during each time of the day, the trips made by members don't variate much between the morning and afternoon, but increase greatly in the evening and decrease at night matching the commute rush hours. |

```
ggplot(data = all_trips, mapping = aes(x = day_time, fill = member_casual))+
  geom_bar(position = "dodge")+
  labs(title = "TRIPS BY TIME OF THE DAY IN EACH DAY OF THE WEEK",
       X = NULL,
       y = "Trip count")+
  scale_fill_manual(legend_title, values = plot_col)+
  scale_x_discrete(limits = c("morning","afternoon","evening", "night"))+
  scale_y_continuous(labels = comma)+
  facet_wrap(~day_of_week)+
  theme(axis.text.x = element_text(angle = 45, vjust = 0.8, hjust=0.6))
```

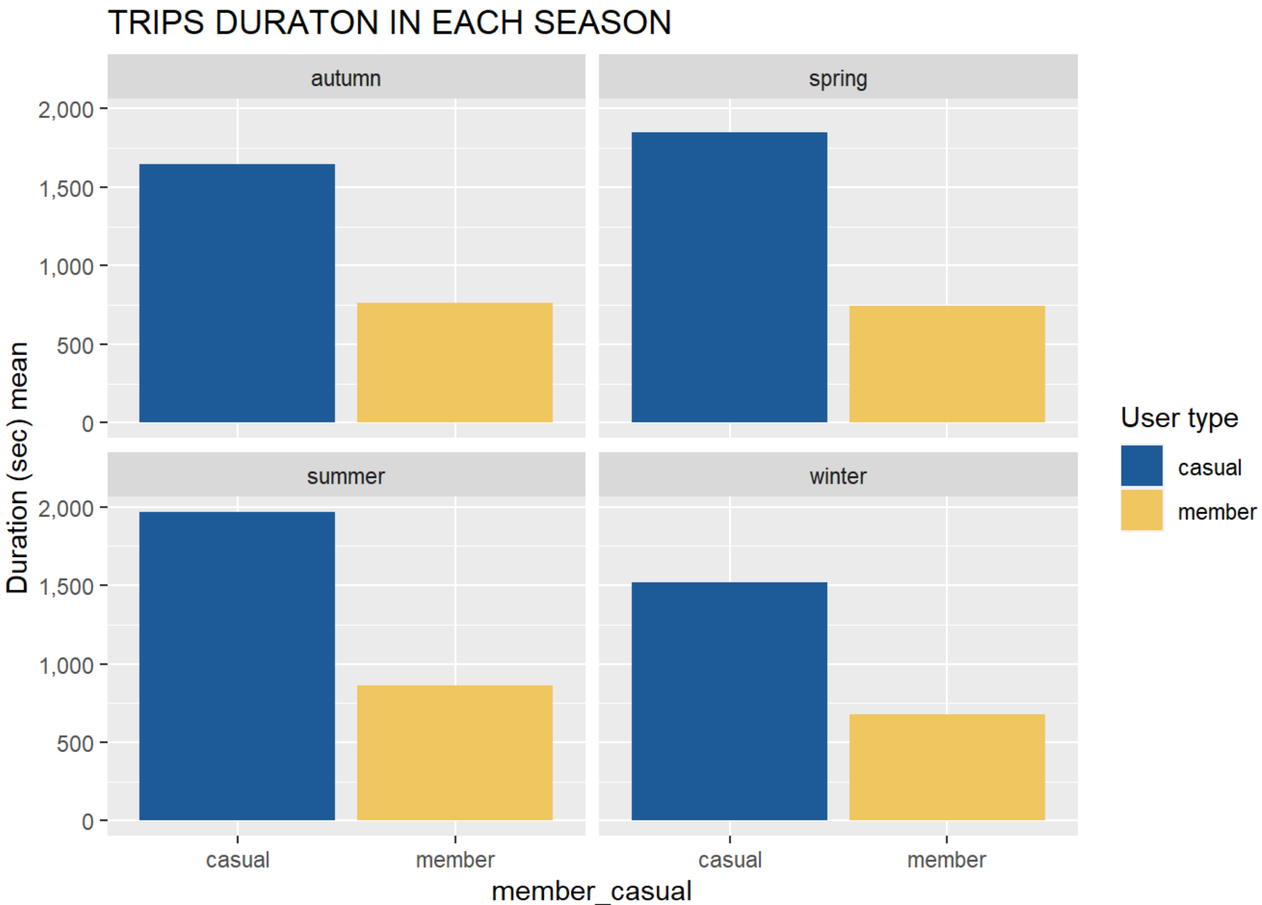


- Looking at each day of the week, we can establish an important difference. At weekends, besides the trips made by casuals increase, the majority of trips start in the afternoon, uncovering another characteristic of casual users.

3.1 Difference between “HOW” casuals and members riders use Cyclist

Duration of trips by members and casuals

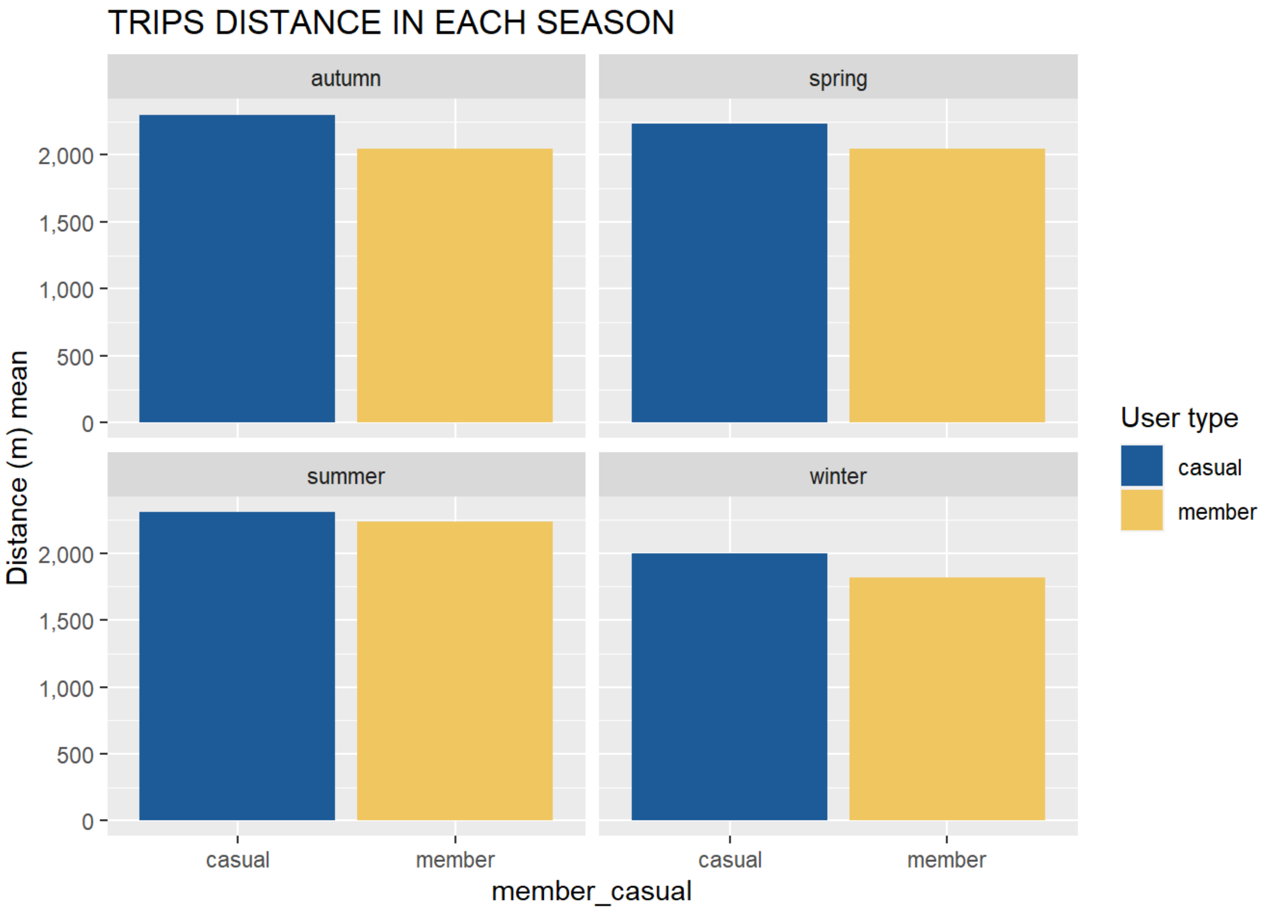
```
all_trips %>%
  group_by(year_season, member_casual) %>% summarise(duration_mean = mean(trip_dur), .g
roups = 'drop') %>%
ggplot(mapping = aes(x = member_casual, y = duration_mean, fill = member_casual))+
  geom_col()+
    labs(title = "TRIPS DURATON IN EACH SEASON",
        X = NULL,
        y = "Duration (sec) mean")+
  scale_fill_manual(legend_title, values = plot_col)+
  scale_y_continuous(labels = comma)+
  facet_wrap(~year_season)
```



- The trips made by casuals take more time than the members' trips all year round

Distance of trips by members and casuals

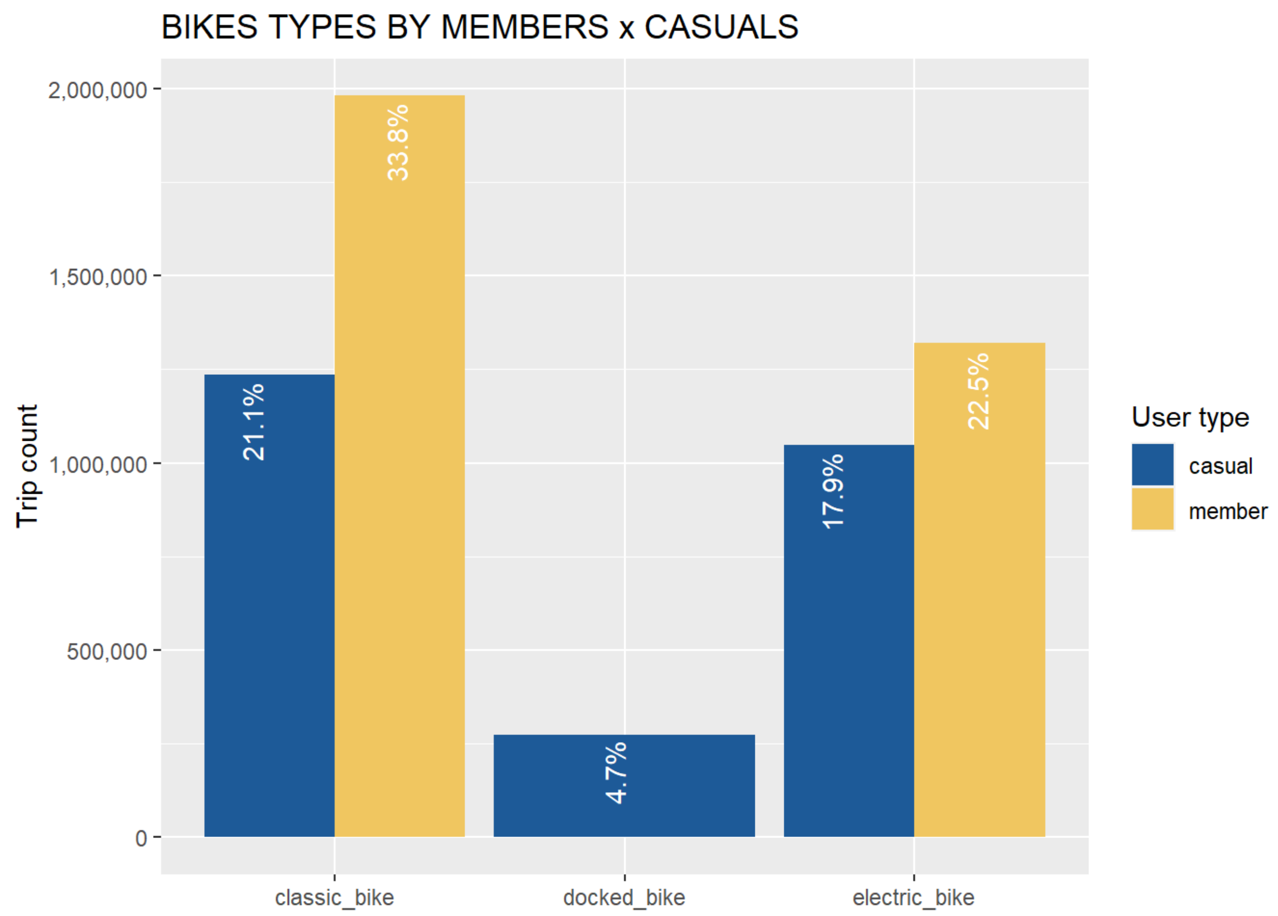
```
all_trips %>%
  group_by(year_season, member_casual) %>% summarise(distance_mean = mean(trip_dist[!i
s.nan(trip_dist)]), .groups = 'drop') %>%
ggplot(mapping = aes(x = member_casual, y = distance_mean, fill = member_casual))+
  geom_col()+
  labs(title = "TRIPS DISTANCE IN EACH SEASON",
        X = NULL,
        y = "Distance (m) mean")+
  scale_fill_manual(legend_title, values = plot_col)+
  scale_y_continuous(labels = comma)+
  facet_wrap(~year_season)
```



- When comparing the trips by time duration and distance, we realize that the trips made by casuals are longer and take more time, but the distance increase rate is not so big as the duration increase rate. Therefore, casual rides are a little longer, but much more time. Another confirmation is that the casuals use Cyclistc for leisure.

Bikes types by members and casuals

```
ggplot(all_trips, mapping = aes(x = bike_type, fill = member_casual), labels = label_pe
rcent())+
  geom_bar(position = "dodge")+
  geom_text(aes(label = scales::percent(..count../sum(..count..))), stat = "count", vju
st = 0, hjust = 1.1, colour = "white",
            angle = 90 , position = position_dodge(width = 1))+
  scale_fill_manual(legend_title, values = plot_col)+
  labs(title = "BIKES TYPES BY MEMBERS x CASUALS",
       x = NULL,
       y = "Trip count")+
  scale_y_continuous(labels = comma)
```



- There is no relevant trend for this analysis.

4 Last considerations

4.1 Differences

- Most of the Casuals users use Cyclistic for leisure and the members for everyday routine, most likely to commute. That is the reason for the following differences:
    - Members' trips number don't variate so much as the number of casuals in each season of the year.
    - Casuals ride more at weekends and members on weekdays.
    - Casual trips mostly start in the afternoon and members' trips in the evening.
    - Casual trips are longer in distance and time.

4.2 Recommendations

- Due to the casual's seasonality, it may be interesting for Cyclistic to create segmented memberships just for warmer days, weekends, or afternoons.

- As the casuals use Cyclistic mainly in their free time, the marketing campaigns targeting them should include themes like "Chicago sightseeing", "family bicycle riding", and "bicycle riding dates".