Cyclistic bike-share analysis case study

Google Data Analytics Professional Certificate Capstone

Introduction

This case study serves as the capstone requirement for the Google Data Analytics Professional Certificate. In this project, I assumed the role of a junior data analyst working for Cyclistic, a fictional bike-share company based in Chicago. Since its inception in 2016, Cyclistic 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 has two types of customers, customers who purchase single ride are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

In order to increase revenue, the company wants to come up with a new marketing strategy to convert casual riders into annual members.

The analysis will follow the 6 phases of the Data Analysis process: Ask, Prepare, Process, Analyze, and Act (APPAA).

Ask phase:

In an effort to grow the company, the Marketing department, led by Lily Moreno, wants to come up with a creative marketing strategy that leads to convert the casual riders to annual subscribers.

I have been assigned to analyze the way that the annual members and casual riders use our application differently.

Data-driven insight into the trends should help the Marketing department to plan the most effective marketing strategy using the digital media that makes the casual users of Cyclistic buy an annual membership.

At the end of the analysis, I will include some recommendations that try to achieve the goal of growing the company by increasing the annual subscribers. These recommendations will help the executive team to take the decision of marketing strategy approval.

<u>In the analysis, I'm trying to answer these questions:</u>

1- **What is the number of trips for both the casual riders and the members?**
2- **Find the average ride length for both the casual riders and the members?**
3- **Identify the most common start and end stations for both the casual riders and the members?**
4- **Which days of the week have the regular peak in using our services for both?**

Consider key stakeholders:

- Marketing department led by Lily Moreno.
- The Cyclistic marketing analytics team, which I'm one of this team.
- Cyclistic executive team, the decision-makers to approve or disapprove the recommendations and the marketing strategy.

Prepare phase:

Where is your data located?

The dataset used in this case study is actual public data (the Divvy datasets) that have been made available by Motivate International Inc, which operates the City of Chicago's Divvy bicycle-sharing service.

Cyclisitc is a fictional company using the Divvy datasets to explore the behaviors of both the members and casual riders.

How is the data organized?

The data were organized as separate files by month and year. The data was saved as .csv files within .zip folders. Analysis for this case study is made using one-year data from April 2020 to October 2021. During the analysis, I stored original copies of the data on a secured hard drive and worked with copies of the data on my pc.

The data included the following fields:

ride_id: a unique ID per ride

rideable_type: the type of bicycle used

started_at: the date and time that the bicycle was checked out

ended_at: the date and time that the bicycle was checked in

start_station_name: the name of the station at the start of the trip

start_station_id: a unique identifier for the start station

end_station_name: the name of the station at the end of the trip

end_station_id : a unique identifier for the end station

Start_lat: the latitude of the start station

start_lng: the longitude of the start station

end_lat: the latitude of the end station

end_lng: the longitude of the end station

member_casual: a field indicating whether the bicycle was taken about by a member or a casual

<u>Are there issues with bias or credibility in this data? Does your data ROCCC?</u>

The data is credible, it is first-party information, it is safe to assume that it is unbiased.

Data is Reliable, Original, Comprehensive, Current where the latest data is Oct 202, and Cited where the data is provided by the company on the site:

https://www.divvybikes.com/system-data

according to the license:

https://www.divvybikes.com/data-license-agreement

<u>How are you addressing licensing, privacy, security, and accessibility?</u>

Privacy is protected by using the *ride_id* as opposed to the riders' personal information, although it might have been useful to have a *ride_id* and total spent per ride so we could track the differences in spending between members and casuals.

<u>Are there any problems with the data?</u>

1- Choose the tools:

The combined size of all the datasets is close to 1.4 GB. Data cleaning in spreadsheets will be time-consuming and slower than using SQL or R. I will use R for data wrangling, analysis, and visualizations.

2- Check the data for errors:

To check the errors, we need to load the datasets.

First, load the libraries:

```
# Load Libraries
library(tidyverse)
library(ggplot2)
library(readr)
library(dplyr)
library(janitor)
library(data.table)
library(tidyr)
library(lubridate)
library(skimr)
```

Second, will load the datasets,which are in .CSV fromat, so, will use "*read.csv* " function to save each dataset in a variable has the name of its month.

```
# Load datasets
nov20<-read.csv("C:/BikeShare-datasets/202011-divvy-tripdata.csv")
dec20<-read.csv("C:/BikeShare-datasets/202012-divvy-tripdata.csv")
jan21<-read.csv("C:/BikeShare-datasets/202101-divvy-tripdata.csv")
feb21<-read.csv("C:/BikeShare-datasets/202102-divvy-tripdata.csv")
mar21<-read.csv("C:/BikeShare-datasets/202103-divvy-tripdata.csv")
apr21<-read.csv("C:/BikeShare-datasets/202104-divvy-tripdata.csv")
may21<-read.csv("C:/BikeShare-datasets/202105-divvy-tripdata.csv")
jun21<-read.csv("C:/BikeShare-datasets/202106-divvy-tripdata.csv")
jul21<-read.csv("C:/BikeShare-datasets/202107-divvy-tripdata.csv")
aug21<-read.csv("C:/BikeShare-datasets/202108-divvy-tripdata.csv")
sep21<-read.csv("C:/BikeShare-datasets/202109-divvy-tripdata.csv")
oct21<-read.csv("C:/BikeShare-datasets/202110-divvy-tripdata.csv")
```

Then will create a list has all variables.

```
ds<-
list(nov20,dec20,jan21,feb21,mar21,apr21,may21,jun21,jul21,aug21,sep21,oct21)
```

Now, will check the columns' names to ensure that all datasets have the same columns' names.

```
for(i in 1:length(ds)){
        print(colnames(ds[[1]]))
}
```

```
> for(i in 1:length(ds)){
+ print(colnames(ds[[1]]))
+ }
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
> 
```

In the next steps, I'm going to make a lot of changes on the data to prepare it for analysis. So, Instead of doing these changes in each data file, I combine all files in one file has all data for the whole year.

```
all_trips<-do.call('rbind',ds)
```

Now, *all_trips* is the file that has all data that we need for analysis.

For check a summary of our data

```
skim(all_trips)
```

```
-- Data Summary ------------------------
                           Values
Name                       all_trips
Number of rows             5378834
Number of columns          13

_______________________
Column type frequency:
  character                9
  numeric                  4

________________________
Group variables            None

-- Variable type: character ------------------------------------------------------------------
# A tibble: 9 x 8
  skim_variable      n_missing complete_rate   min   max  empty n_unique whitespace
* <chr>                  <int>         <dbl> <int> <int>  <int>    <int>      <int>
1 ride_id                    0             1    16    16      0  5378625          0
2 rideable_type              0             1    11    13      0        3          0
3 started_at                 0             1    19    19      0  4487412          0
4 ended_at                   0             1    19    19      0  4479067          0
5 start_station_name         0             1     0    53 600479      815          0
6 start_station_id         24434       0.995     0    36 576152     1304          0
7 end_station_name           0             1     0    53 646471      812          0
8 end_station_id         26826         0.995     0    36 619722     1299          0
9 member_casual              0             1     6     6      0        2          0

-- Variable type: numeric --------------------------------------------------------------------
# A tibble: 4 x 11
  skim_variable n_missing complete_rate  mean     sd    p0   p25   p50   p75  p100 hist
* <chr>             <int>         <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 start_lat             0             1  41.9 0.0455  41.6  41.9  41.9  41.9  42.1 □□□□□
2 start_lng             0             1 -87.6 0.0280 -87.8 -87.7 -87.6 -87.6 -87.5 □□□□□
3 end_lat            4831         0.999  41.9 0.0456  41.5  41.9  41.9  41.9  42.2 □□□□□
4 end_lng            4831         0.999 -87.6 0.0282 -88.1 -87.7 -87.6 -87.6 -87.4 □□□□□
```

From red square no.1, the format of the values the *start_at* and *end_at* columns is “chr” . I need to change the format to “datetime” , so, all of them will have the same format.

```
all_trips$started_at<-as_datetime(all_trips$started_at)
all_trips$ended_at<-as_datetime(all_trips$ended_at)
```

I It’s clear from red square no.2, there are multiple empty values in the tables.

To solve these empty values in data, will fill it with “N/A”

```
all_trips$start_station_name[all_trips$start_station_name=='']<-'N/A'
all_trips$end_station_name[all_trips$end_station_name=='']<-'N/A'
```

Because the stations’ names and ID’s have the same meaning, and at the same time the ID’s have a lot of missing values, I will use the names in the analysis and delete the ID’s.

About the coordinates, it wont help us in the analysis so I will delete them also.

```
all_trips<-all_trips%>%select(-c(start_station_id , end_station_id,
start_lat:end_lng))
```

Process phase:

During the analysis, I need to add some fields to help me during the analysis:

- ***ride_length***: the length of the ride calculated as *ended_at — started_at*.
- ***hour***: the hour of the trip for *started_at*
- ***day***: the day of the trip for *started_at*
- ***month***: the month and the year of the trip for *started_at*

Transform the data and Document the cleaning process:

1- create *ride_length* in minutes.

```
all_trips$ride_length<-
(as.double(difftime(all_trips$ended_at,all_trips$started_at)))/60
```

Some logged in to Cyclistic were for TEST, these values not useful and should delete it

```
all_trips<-all_trips[!((all_trips$start_station_name %like% "TEST")),]
nrow(subset(all_trips, start_station_name %like% "TEST"))
```

Check the *ride_length* summary

```
summary(all_trips$ride_length)
```

```
> summary(all_trips$ride_length)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-29049.97      6.97     12.38     20.49     22.43  55944.15
```

From summary, we find that there are periods in minus, and the maximum period is 55944 minutes, which equal about 30 days. These data will affect negatively in our analysis and the final recommendations, so will make some changes on it to be more appropriate for our purpose.

In the real life, no one will use Cyclistic for one or two minutes, so will delete any data has period less than 5 minutes. At the same time will assume that if the user used the bike-share for the whole day, he will be connected to Cyclistic for maximum 18 hours.

```
all_trips<-filter(all_trips,ride_length>5)
all_trips<-filter(all_trips,ride_length<1080)
```

1- create *hour* field

```
all_trips$hour<-format(all_trips$started_at,'%H')
all_trips$hour<-as.POSIXct(all_trips$hour,format="%H")
```

2- create *day* field

```
all_trips$day<-format(all_trips$started_at,'%A')
all_trips$day <- ordered(all_trips$day, levels=c("Monday", "Tuesday",
"Wednesday", "Thursday", "Friday", "Saturday", "Sunday"))
```

3- create *month* field

```
all_trips$month<-format(as.Date(all_trips$started_at),"%b-%y")
all_trips$month <- ordered(all_trips$month, levels=c('Nov-20','Dec-20','Jan-21','Feb-21','Mar-21','Apr-21','May-21','Jun-21','Jul-21','Aug-21','Sep-21','Oct-21'))
```

Analysis phase:

In the Process phase of the Data Analysis process, my tasks were to:

- Aggregate your data so it's useful and accessible.
- Organize and format your data.
- Perform calculations.
- Identify trends and relationships.
- My guiding question is **How might we influence casual riders to purchase annual subscriptions based on their riding habits?** So, a good place to start is to see when and how our Cyclistic's riders are using the service.

Note: The code for the analysis will be at the attached file.

- **Number of trips by user type**

A quick look at the down table and Pie-chart, gives us a general understanding of the current makeup of our customer base. From the pie chart, we find that the number of rides taken by members is about 51% (2265342 members' trips), which is slightly more than casual riders' number is about 49% (2330808 casual riders' trips) in the 12 months under review.

These numbers mean that there is a significant capability for converting casual riders into members, which is the primary goal of this analysis.

```
# A tibble: 2 x 2
  member_casual number_of_rides
  <chr>                   <int>
1 casual                2265342
2 member                2330808
```

Figure 1: Number of rides by user types.

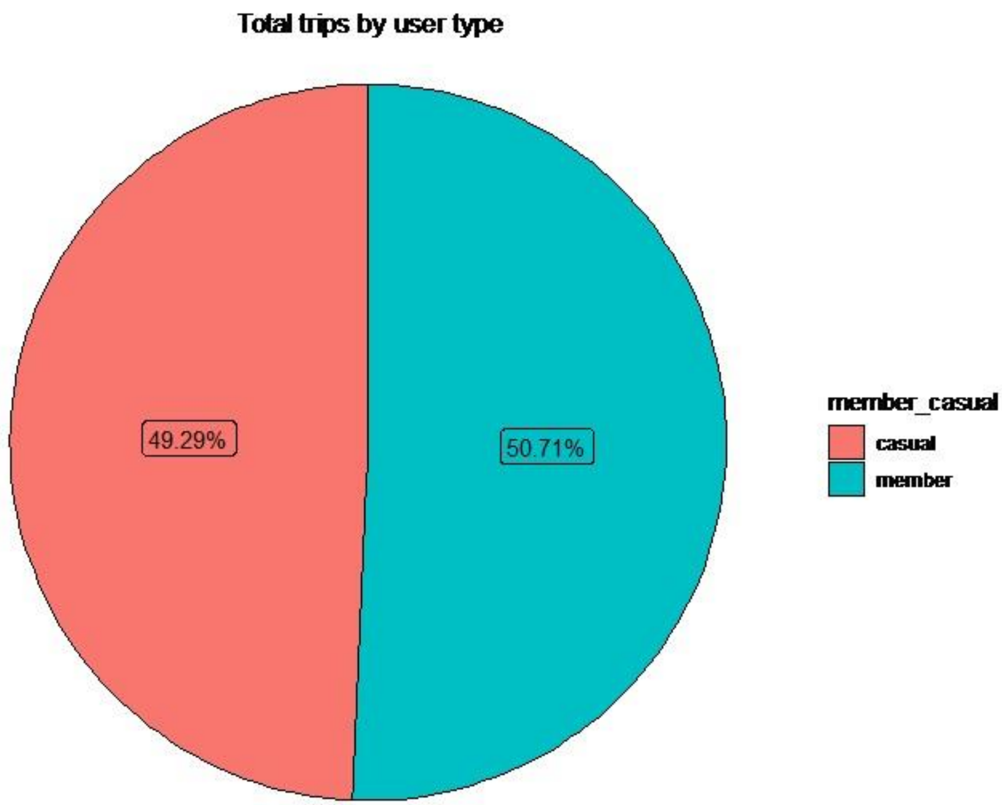

Figure 2: Total trips by user types

- **Ride Duration by Minutes**

From the table.5&6, the mean ride length of the members (16.2 mins) is always lower than the mean ride length for all trips (22.5 mins). On the other hand, the casual riders' mean ride length (29 mins) is always more than the mean ride length for all trips. That back us to the hypothesis that the members use Cyclistic to reach specific places, while casual riders use the bikes for leisure and joyrides.

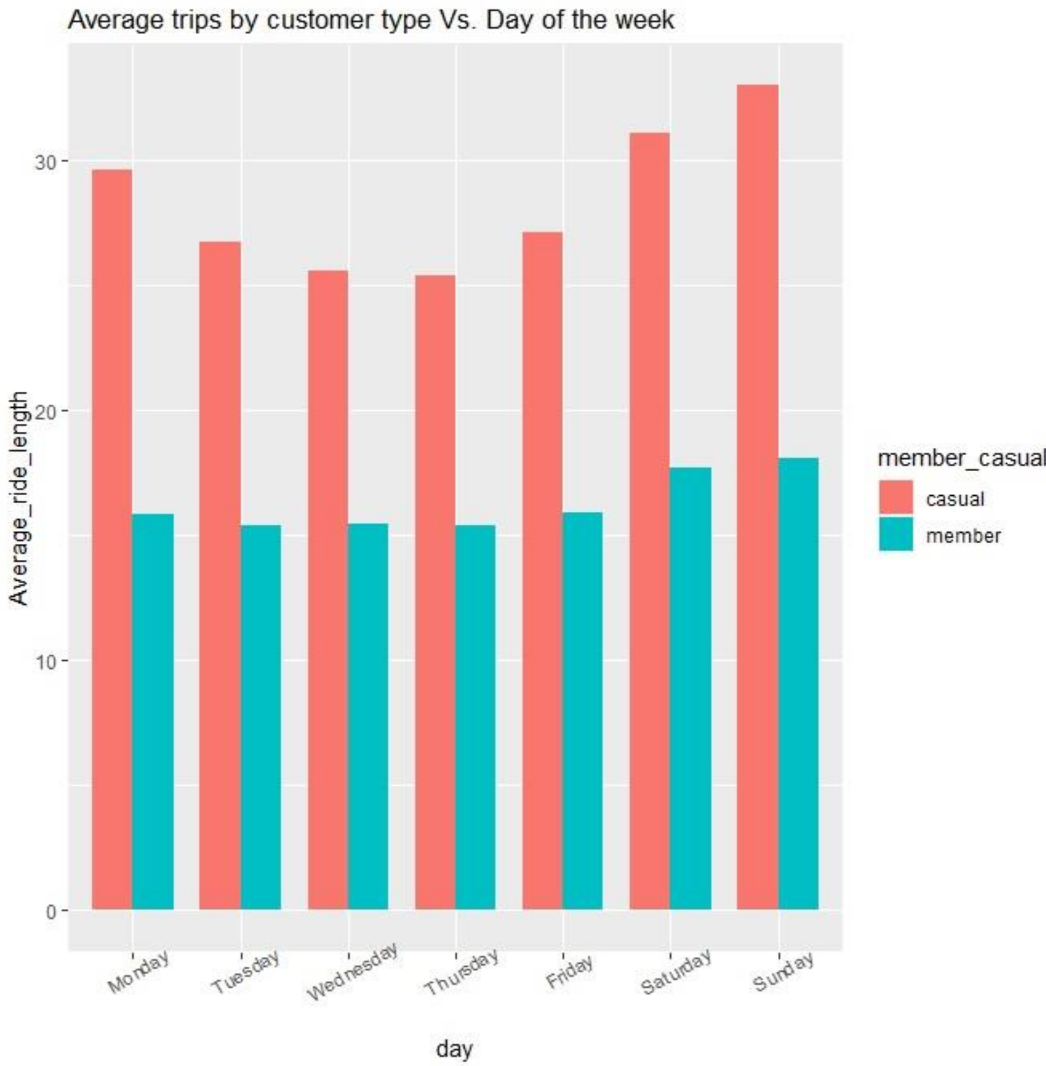

Figure 3: Average trips by customer type Vs. Day of the week

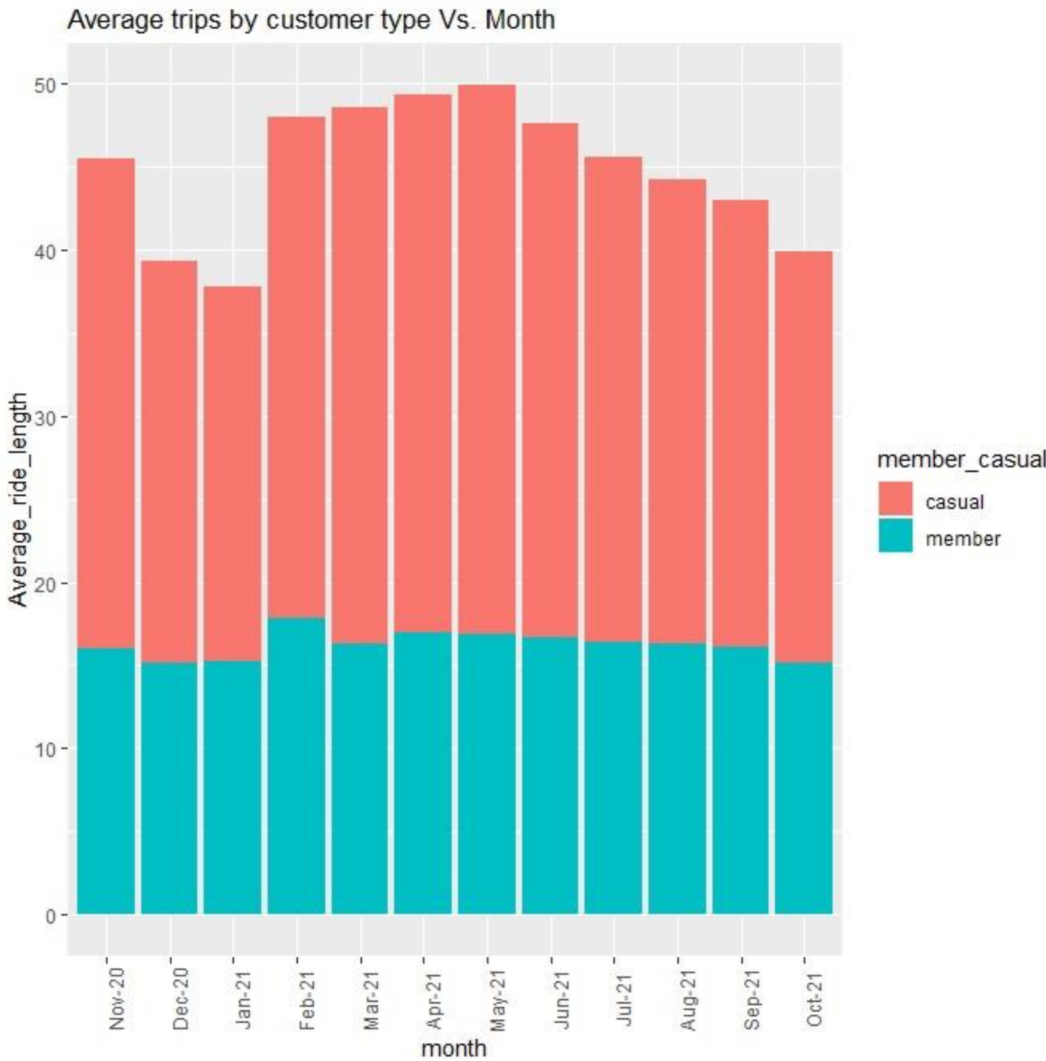

Figure 4: Average trips by customer type Vs. month

```
+ summarise(Average_ride_length=mean(ride_length))
  Average_ride_length
1            22.56082
>
> #Average ride length per user type
> all_trips%>%group_by(member_casual)%>%
+ summarise(Average_ride_length=mean(ride_length))
# A tibble: 2 x 2
  member_casual Average_ride_length
  <chr>                       <dbl>
1 casual                       29.1
2 member                       16.2
>
> #Total ride length per user type
> all_trips%>%group_by(member_casual)%>%
+ summarise(Total_ride_length=sum(ride_length))
# A tibble: 2 x 2
  member_casual Total_ride_length
  <chr>                     <dbl>
1 casual                65837359.
2 member                37855552.
```

Figure 5: Average and total ride duration by user type

- **Peak by day**

From *table.1* and *graph.1*, Over the year, Saturday was the busiest day of the week for casual riders and Tuesday for the members, while weekdays had the fewest rides by casual riders and Sundays had the fewest rides by the members.

There is a fairly consistent riding pattern during the week from the members who use Cyclistic. This nearly consistent pattern leads to presume most of the members rely on Cyclistics' bikes for commutes are job goers, thus have more consistent riding patterns day to day.

On the other hand, casual riders primarily use Cyclistics' bikes in the weekends, particularly on Saturdays. This fits the hypothesis that casual riders are mostly using the service for leisure as opposed to commuting, although there are still a good number of trips have been booked during the weekdays.

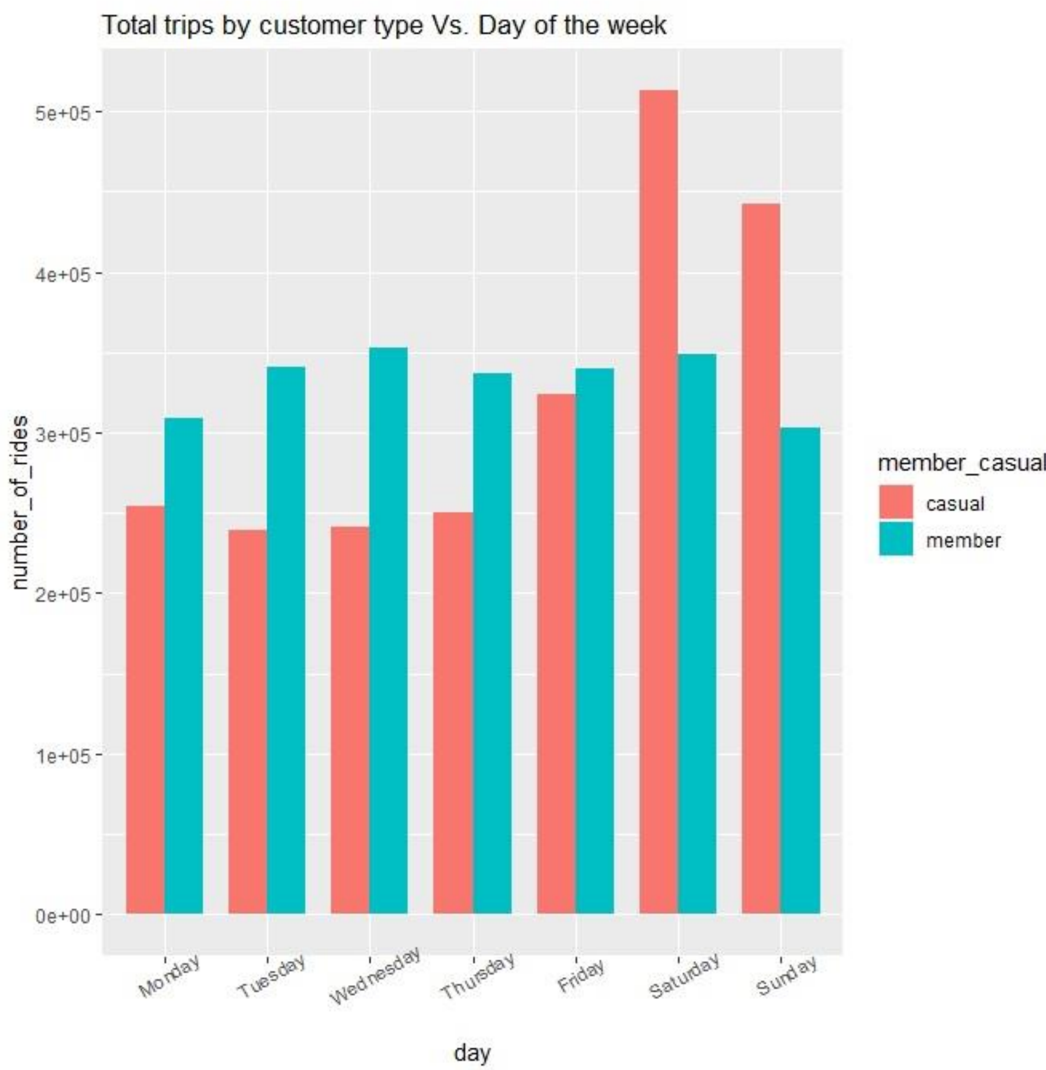

Figure 6: Total trips by customer type Vs. Day of the week

- **Peak by month**

From Table.2 and Bar-graph.3, it is obvious that the peak of demand is in the summer for both customer types.

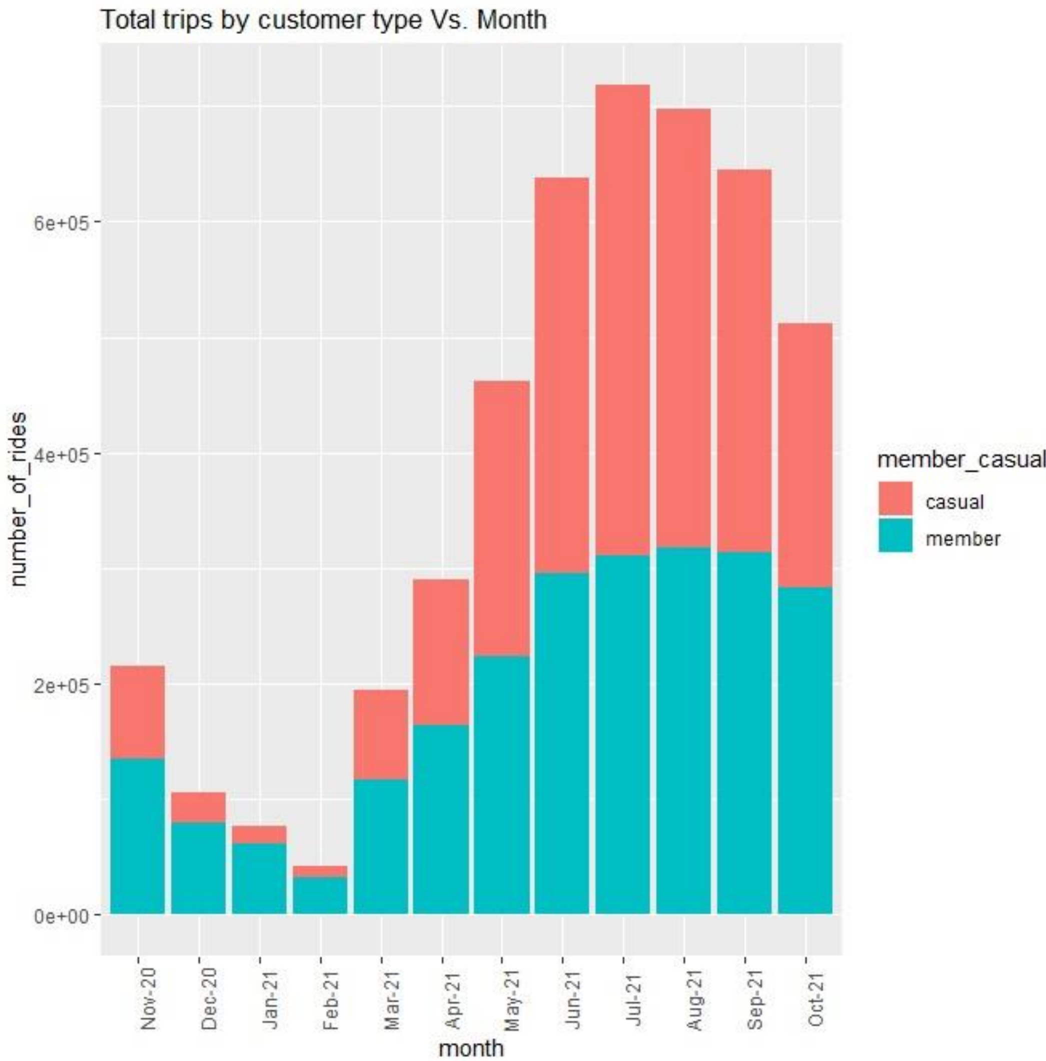

Figure 7: Total trips by customer type Vs. month

Casual riders and members both have similar year-round usage patterns, which increase in the summer months(June to October), particularly in July, and start to decrease, in the winter, from November to reach the lowest values in February.

- **<u>Hourly usage</u>**

From Table.2 and Line-graph.2, it seems that the members' usage peaks are in two times, one in the morning around 7:00-8:00 am and the other one is in the evening around 5:00 pm, which supports the hypothesis of using Cyclistic among the members, such as office-goers, for commutes.

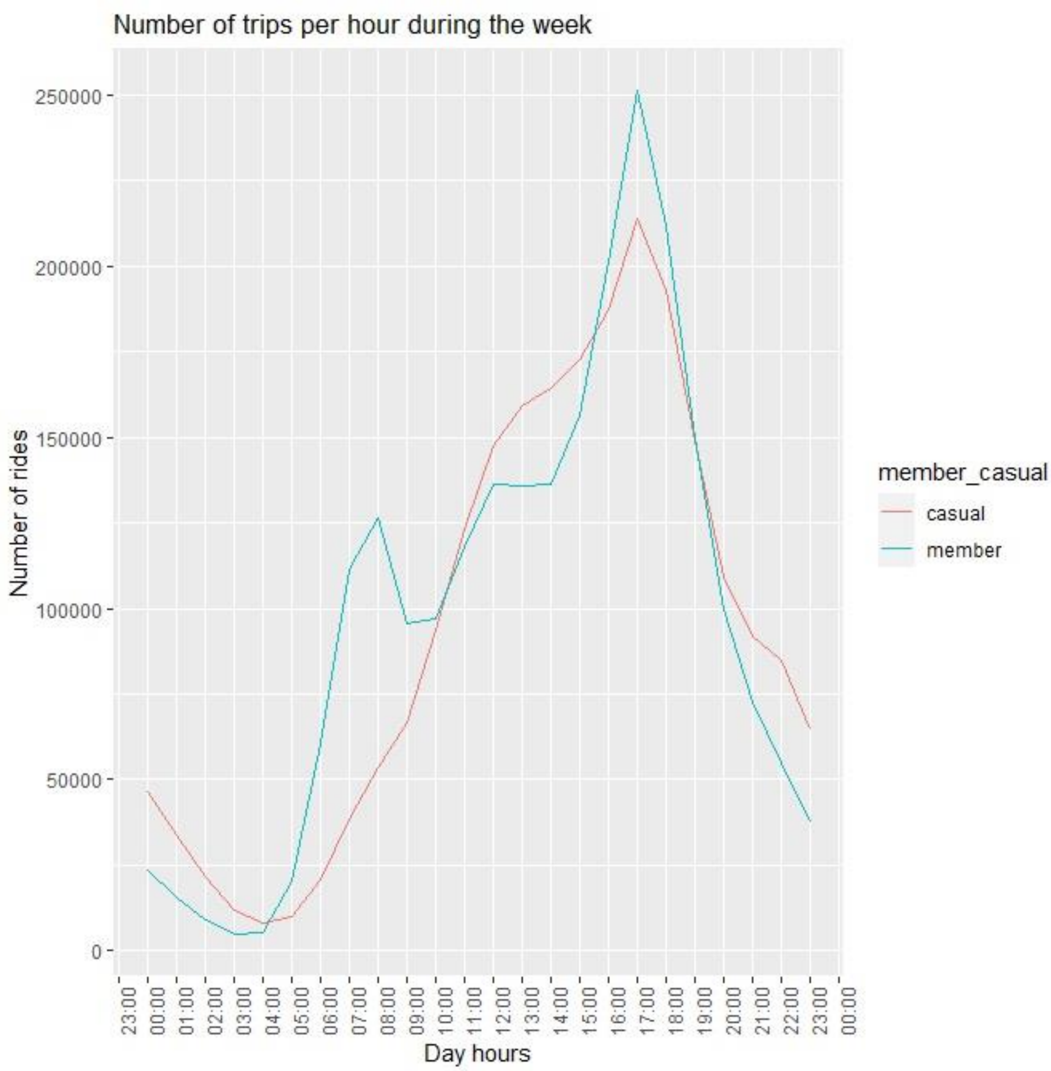

Figure 8: Number of trips per hour during the week

In the meanwhile, the number of casual riders starts to increase gradually from the early morning to reach the usage spike around 5:00 pm, which lets us think that the casual riders may use the bikes for joyrides.

- **Bike Types**

the analysis of the rideable types indicated that classic bikes were the most popular all around, followed by electric bikes and docked bikes in last place. This trend was consistent across members and casual riders.

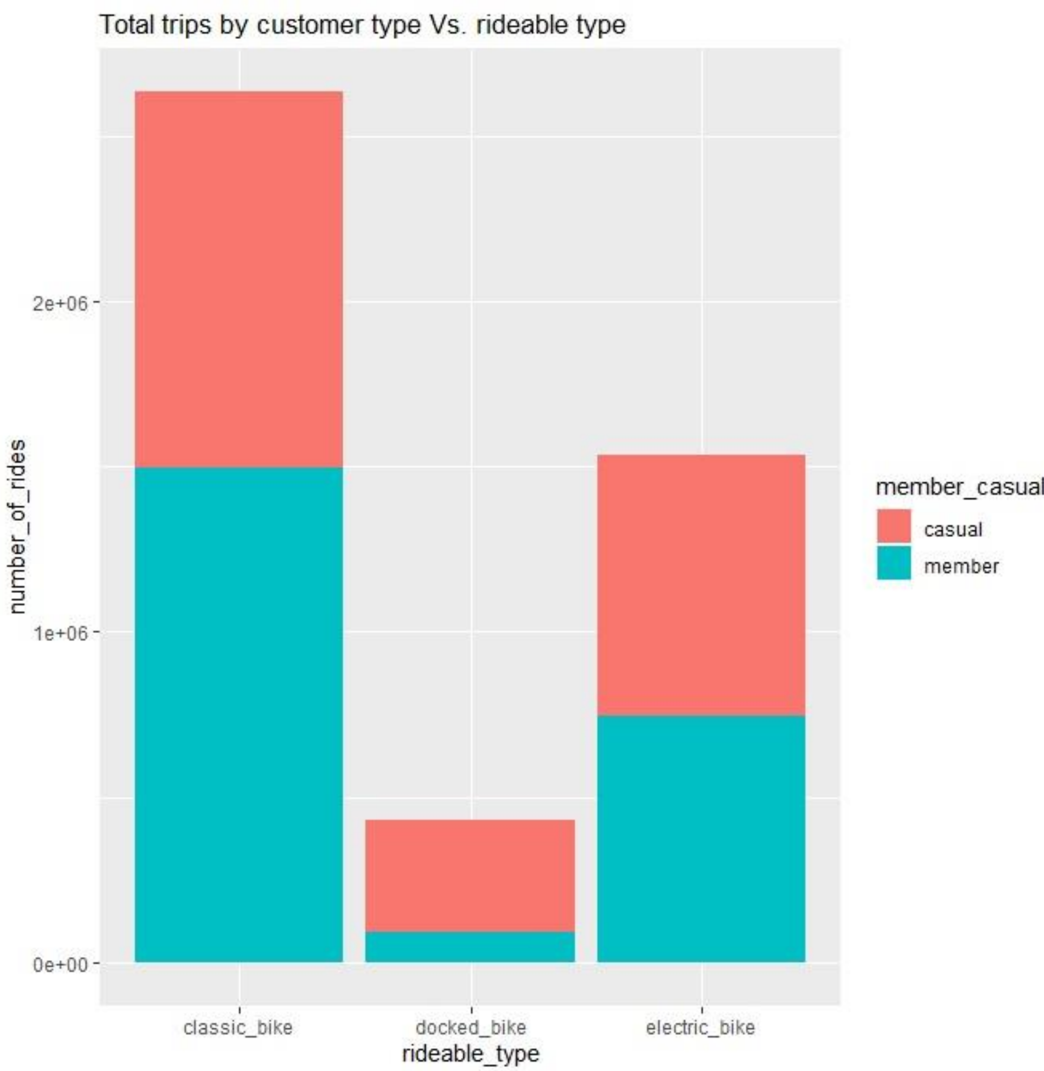

Figure 9: Total trips by customer type Vs. rideable type

- **<u>Stations' locations</u>**

If we look at the heat map.6&7 that show the most popular starting stations in Cyclistic's network during the week, we can note that 'Streeter Dr & Grand Ave', 'Michigan Ave & Oak St' and 'Millennium Park' are significantly more popular among casual riders than they are among members, where most of these stations and rides are concentrated around the city center. If we focused on the casual riders' heat map, we could see that on weekends casual riders have a density around the city center. Member riders on the other hand, are much more spread, especially on weekdays. this adds more weight of the assumption that they use the service mostly work commutes.

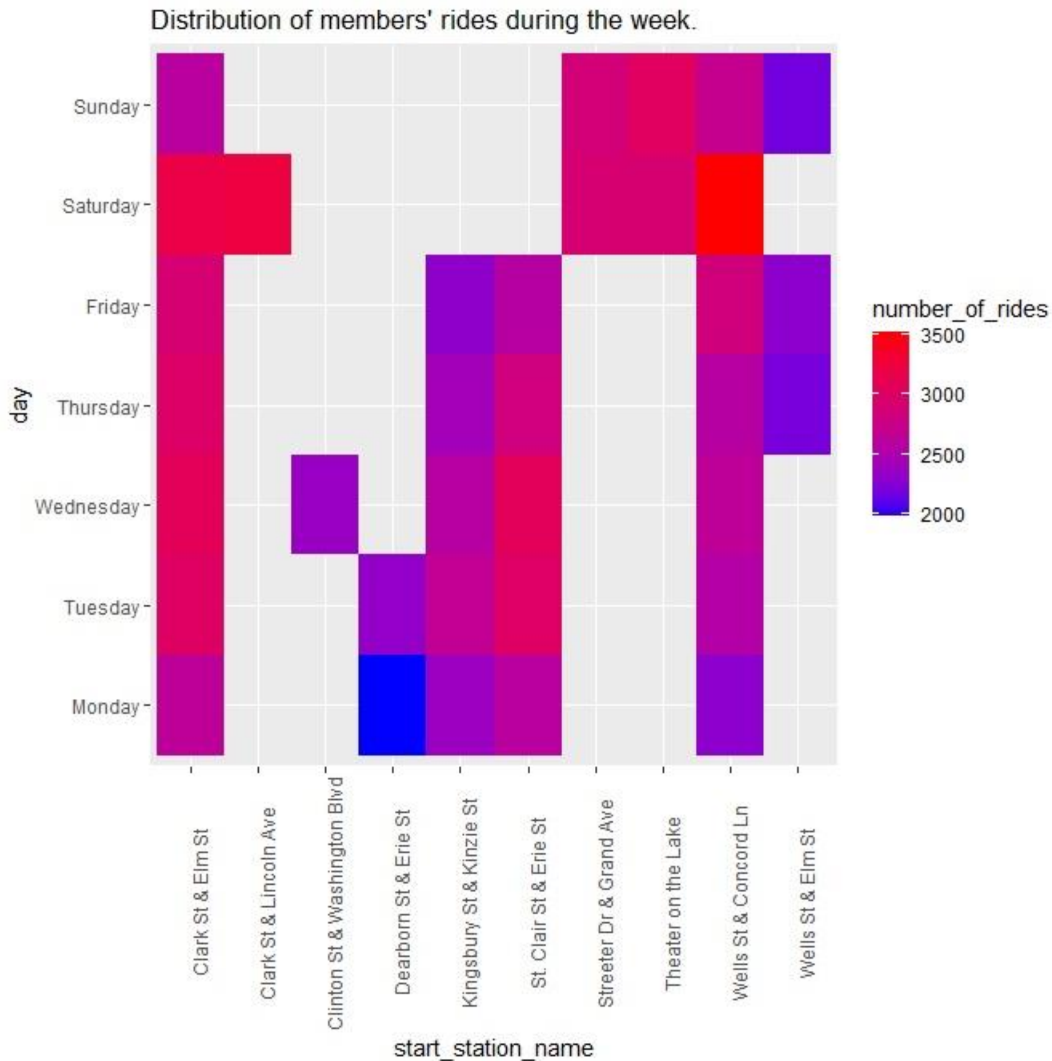

Figure 10: Distribution of members' rides during the week

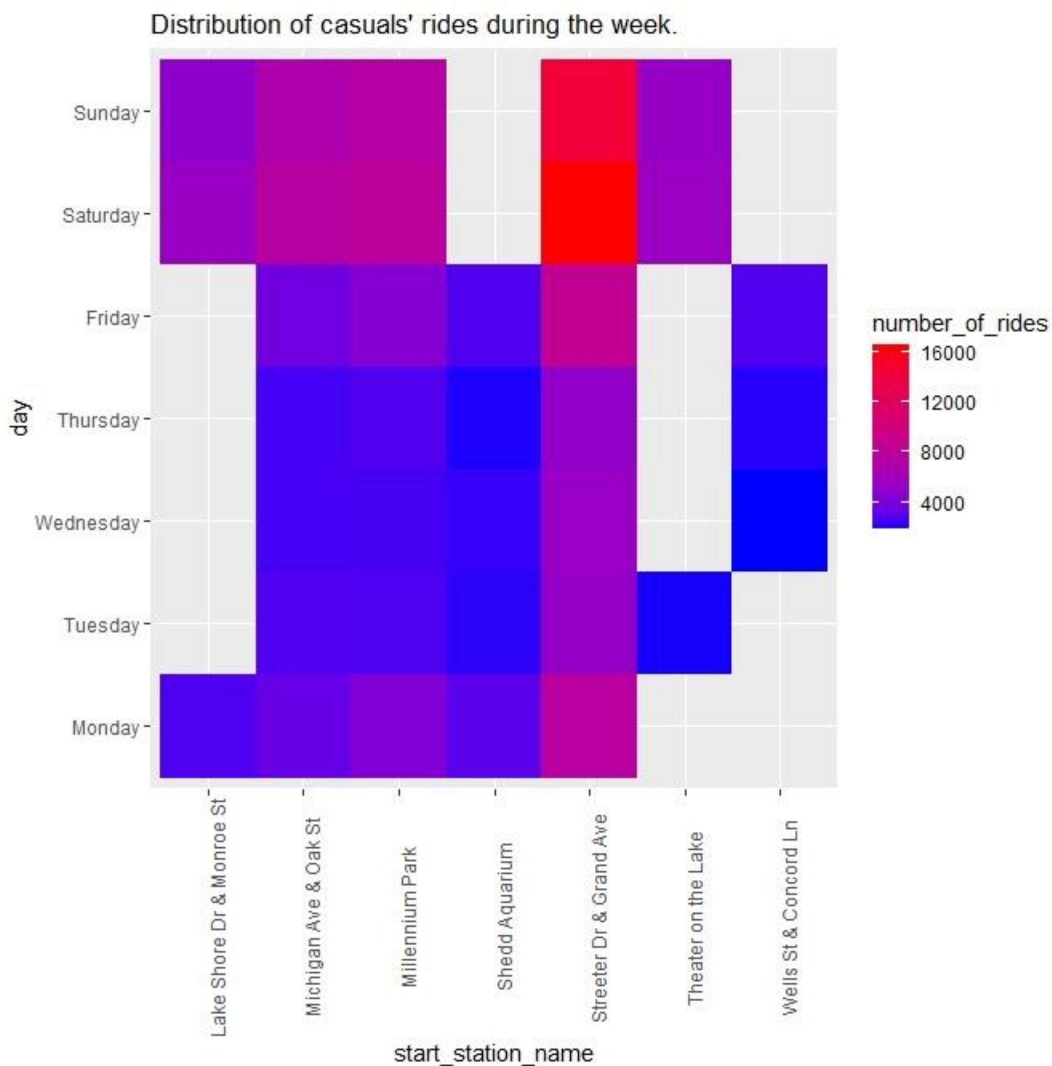

Figure 11: Distribution of casual riders' rides during the week

Act phase:

Summary:

- Both user types share about half the number of rides.
- The average ride duration by casual riders is nearly twice that of member riders.
- Casual customers use Cyclistic services mostly during weekends, while members use them consistently over the entire week.
- In general, casual riders ride longer during summer than the members by almost double, while both use the bike-share relatively similar number of trips in winter.

Recommendations:

Depending on the numbers from the analysis, summer is the peak season of usage. This means that the casual riders would have kind of hesitate to subscribe for an annual subscription when the most usage for the bikes is in summer, so, in winter the subscription will be useless. This lets them think that the pay by ride is cheaper for them than pay for an annual service that the actual benefits are just for about half of the year.

The numbers and graphs showed us also that the casual riders use Cyclistic on the weekends most of the time, for long distances, and around tourist places. That means they don't use it as Cyclistic members for the commute.

An annual subscription system may call it (**Summer+**) or similar names, seeks to satisfy casual and new clients, based on the idea of exploiting the high demand in summer and its decline in winter. This subscription is a package of offers, which satisfy the potential subscriber and at the same time guarantee his annual subscription:

1- Decreasing the annual subscription fees by 30-40%, with free rides during the bike rides season, which is from May to October.
2- During the rest of the year, which is from November to April, the member with *Summer+* will have to pay per ride, but with a discount of 10-15% than the casual riders. This will provide a feeling for the members that they don't pay fees for a not fully used service and at the same time give the member advantage over the casual riders.
3- Give the member an option to select two road lines for free the whole year, whether summer or winter. This will induce the 'work goers' casual riders to switch to members and use Cyclistic as a commute and will encourage them to use bike-share during the weekdays.

I advise the marketing department to create a marketing campaign that starts in the peak usage time for both member and casual riders, which is from spring and goes through September.

The campaign should consider the times of day when riders are most active and focus advertisements from around 7:00 am to 6:00 pm on weekdays, and 11:00 am to 6:00 pm on weekends.

Concentrate advertisements in the parks, companies and factories areas, schools, and universities that a potential member could be interested to use Cyclistic as a commute, especially in the downtown where people suffering the traffic jams.

Encourage the Marketing Department to find institutes and organizations that have common interests to support the marketing campaign, such as Chicago Traffic Department, the Chicago tourism department, the Chicago health department, Universities and schools in Chicago, and fitness clubs. That will attract a wider range of people that are interesting in fitness and health, suffering from traffic jams, and like bike trips.

**Note: all the recommended discount percentages are estimated. To expand the scope of the analysis, additional data should provide:*

1- Usage cost details for members and casual riders - Based on this data, we could study the cost structure for both user types and provide membership plans without affecting the profit margin.
2- More data about the users, such as age, gender, occupation, and address that could help in the analysis and the marketing strategy.

References:

1- Stack Overflow
2- Kaggle community
3- Medium
4- Stack Overflow
5- RDocumentation
6- RStudio community