Capstone Case Study- 1 How Does a Bike-Share Navigate Speedy Success?

Sarang Narayanrao Chandekar

2022-07-05

Scenario

As 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, the team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, the team will design a new marketing strategy to convert casual riders into annual members.But first, Cyclistic executives must approve our recommendations, so they must be backed up with compelling data insights and professional data visualizations.

GUIDING QUESTIONS:

Prepare

- Where is your data located?

  Ans: The Data is located on a cloud server click here. I'll be doing analysis on 12-Months Data sets(JUNE-21 to May-22).

- How is the data organized?

  Ans: The Data sets are provided by Motivate International Inc.on monthly basis in .csv file format.

- Are there issues with bias or credibility in this data? Does your data ROCCC?

  Ans: The data is Reliable, Original, Comprehensive, Current, and Cited as the data is collected directly from the customer's and published by the company so it is not biased and is consistent.

- How are you addressing licensing, privacy, security, and accessibility?

  Ans: The data was collected by Motivate International Inc. by the license click here

- How did you verify the data's integrity?

  Ans:The data is complete as it contains all the required components for the analysis.There are no irregularity in the datasets as the data is maintained according to month/years.hence, the data is consistent and creadible.

- How does it help you answer your question?

  Ans:To increase the number of members i.e from casual riders to membership holders I have created few extra features which will be usefull to put a comparison between the members and the casual riders.

- Are there any problems with the data?

  Ans: There were few missing values which were removed other than that the data was fine.

Process

- What tools are you choosing and why?

  Ans: I will be using R Programming Language as a tool for analysis because the data is vast and using R will be usefull to view the data easily.

- Have you ensured your data's integrity?

  Ans: Yes, The data set consists of all the required data or values neede for the analysis. The Data is not biased and also consistent.

- What steps have you taken to ensure that your data is clean?

  Ans: First as the datasets were individual according to months I had to combine them into one single data frame. After merging the datasets into one I checked for the na values and I found Inspite of being consistent, the data had some "na" values which were removed other than that no dulpicate values were present. So data was cleaned and ready for the analysis.

- How can you verify that your data is clean and ready to analyze?

  Ans: After removing the "na" values, I viewed the data frame and checked whether there are any na values present and found no inconsistency in the data so it was ready to use.

- Have you documented your cleaning process so you can review and share those results?

  Ans: The whole data preparing and cleaning process is present in the below code:

```
library(tidyverse)
```

Importing Packages

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(geosphere)
```

Importing Datasets & Merging them into One single dataframe:

- 12-Months Datasets(JUNE-21 to JUNE-22) are extracted and stored in the Folder- Trip_csv.
- The Datasets then imported and combined to create one single data frame named - Trip.

```
Trip = list.files(path='D:/Capstone-Case Study 1/Trip_csv', full.names = TRUE) %>%
  lapply(read_csv) %>%
  bind_rows
```

- After combining the datasets the data frame looks like

```
head(Trip)
```

```
## # A tibble: 6 x 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>
```

- After looking the Trip Dataframe we notice that it contains na values which we will remove to avoid inconsistency in our data

```
sapply(Trip, function(x) sum(is.na(x)))
```

```
##            ride_id      rideable_type         started_at           ended_at
##                  0                  0                  0                  0
## start_station_name   start_station_id   end_station_name     end_station_id
##             823167             823164             878338             878338
##          start_lat          start_lng            end_lat            end_lng
##                  0                  0               5036               5036
##      member_casual
##                  0
```

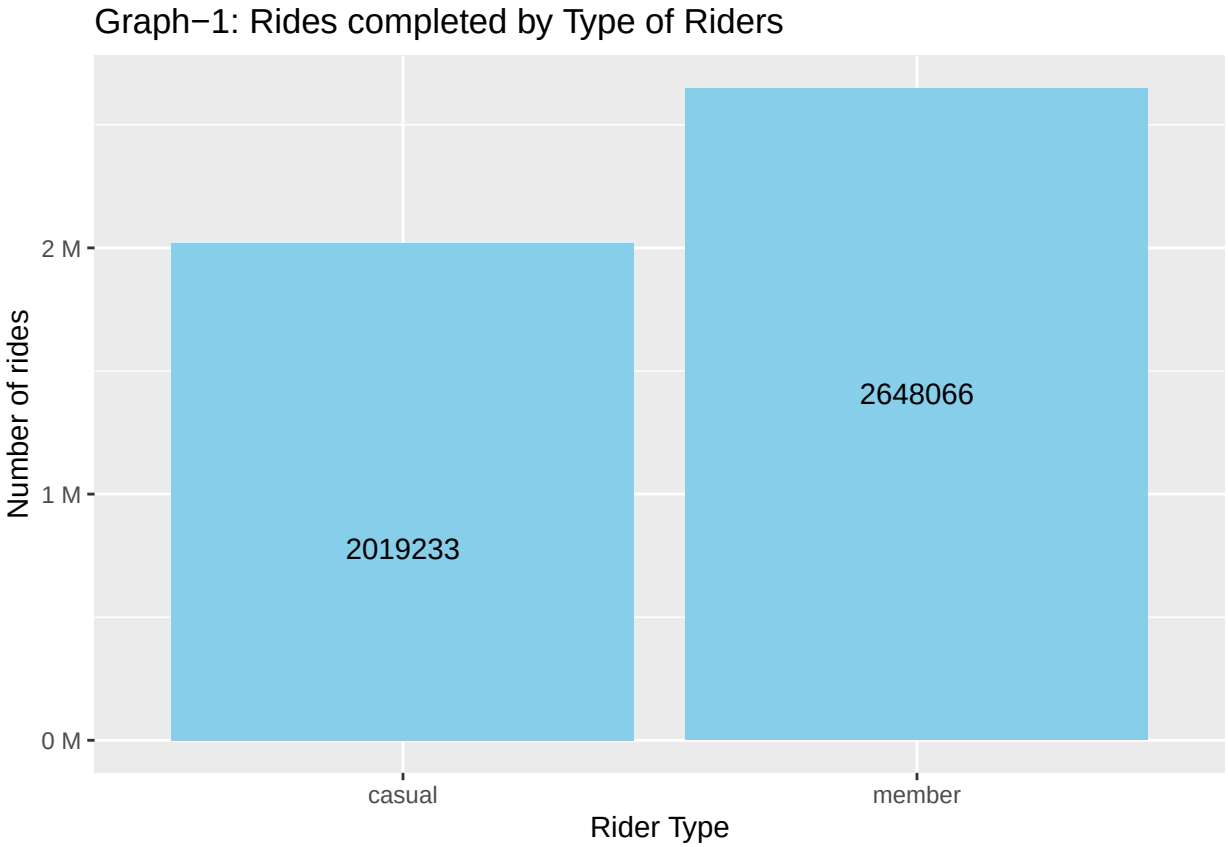
```
Trip1 = na.omit(Trip)
View(Trip1)
```

DATA ANALYSIS AND VISUALIZATION:

Number of Rides VS Type of Riders

- The below chart shows that the number of membership holders is approximately 26% more than the casual riders.

```
ggplot(Trip1, aes(x=member_casual))+geom_bar(fill = "skyblue")+
  labs(title = "Graph-1: Rides completed by Type of Riders",x = "Rider Type",y = "Number of rides")+
  scale_y_continuous(labels = label_number(suffix = " M", scale = 1e-6))+
  geom_text(stat='count', aes(label=..count..), color="black",vjust=15)
```



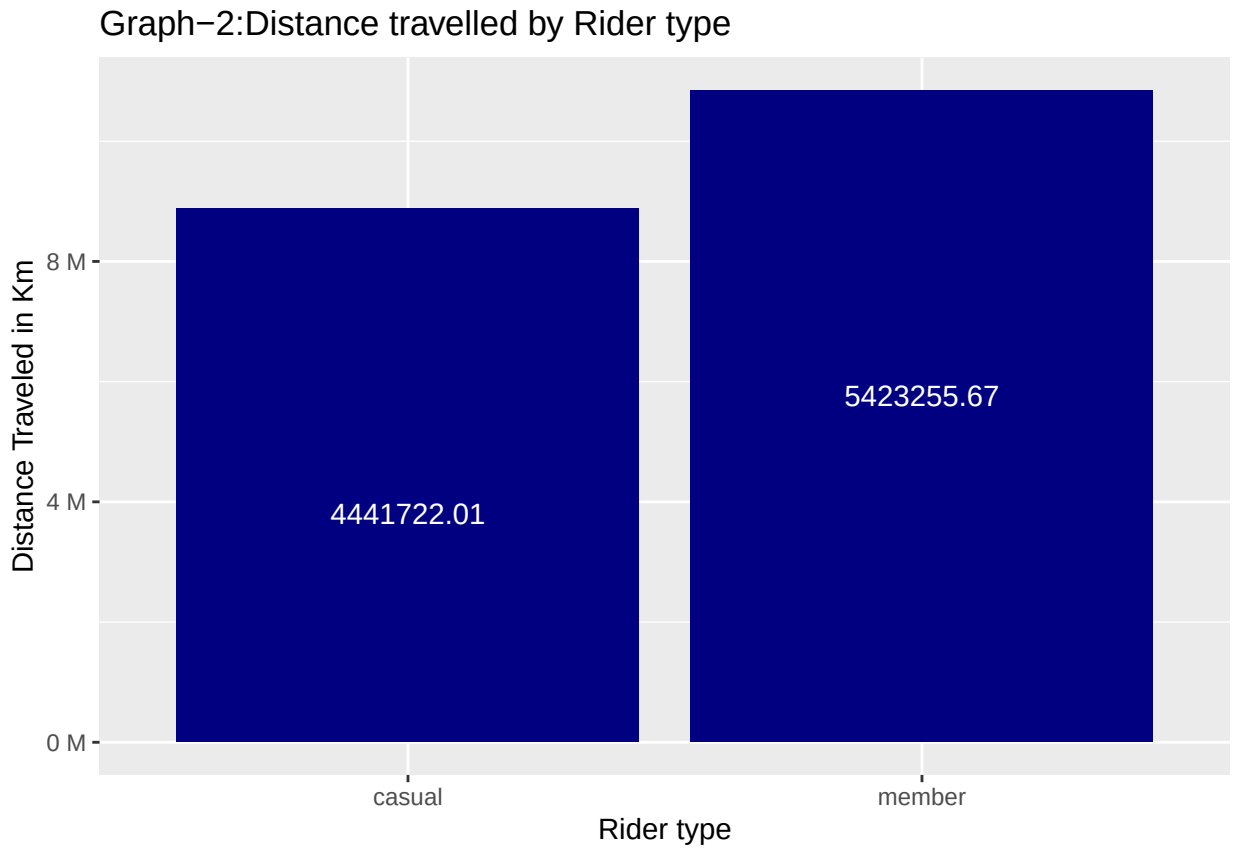
Total Distance Traveled by Rider Type

- Creating new Feature or Column Named = distance: using Longitude and Latitude given inside the data frame
- After creating new feature a graph is plotted according to the kilometers of distance traveled by the riders.
- After examining the graph, it is concluded that around 19.0% of members with annual membership traveled more than casual riders.
- When calculated average distance traveled by the riders it is noted that there is not much of difference between the member and the casual riders.

```
Trip1 = Trip1 %>%
  mutate(Trip1, distance = distHaversine(cbind(start_lng, start_lat), cbind(end_lng, end_lat))/1000)

Trip2 = Trip1 %>%
  group_by(member_casual) %>%
  summarise(distance=sum(distance))

ggplot(Trip2, aes(x=member_casual, y=distance)) +
  geom_bar(stat = "identity", fill= "navyblue") +
  labs(title = "Graph-2:Distance travelled by Rider type",x = "Rider type",y = "Distance Traveled in Km
  scale_y_continuous(labels = label_number(suffix = " M", scale = 2e-6)) +
  geom_text(aes(label=round(stat(y),2)), color="white",vjust=15)
```




```
Trip2.1 =Trip1%>%
  group_by(member_casual) %>%
  summarise(average_distance_km=mean(distance))
print(Trip2.1)
```

```
## # A tibble: 2 x 2
##   member_casual average_distance_km
##   <chr>                       <dbl>
## 1 casual                       2.20
## 2 member                       2.05
```

Duration of Riders Rode the Bike

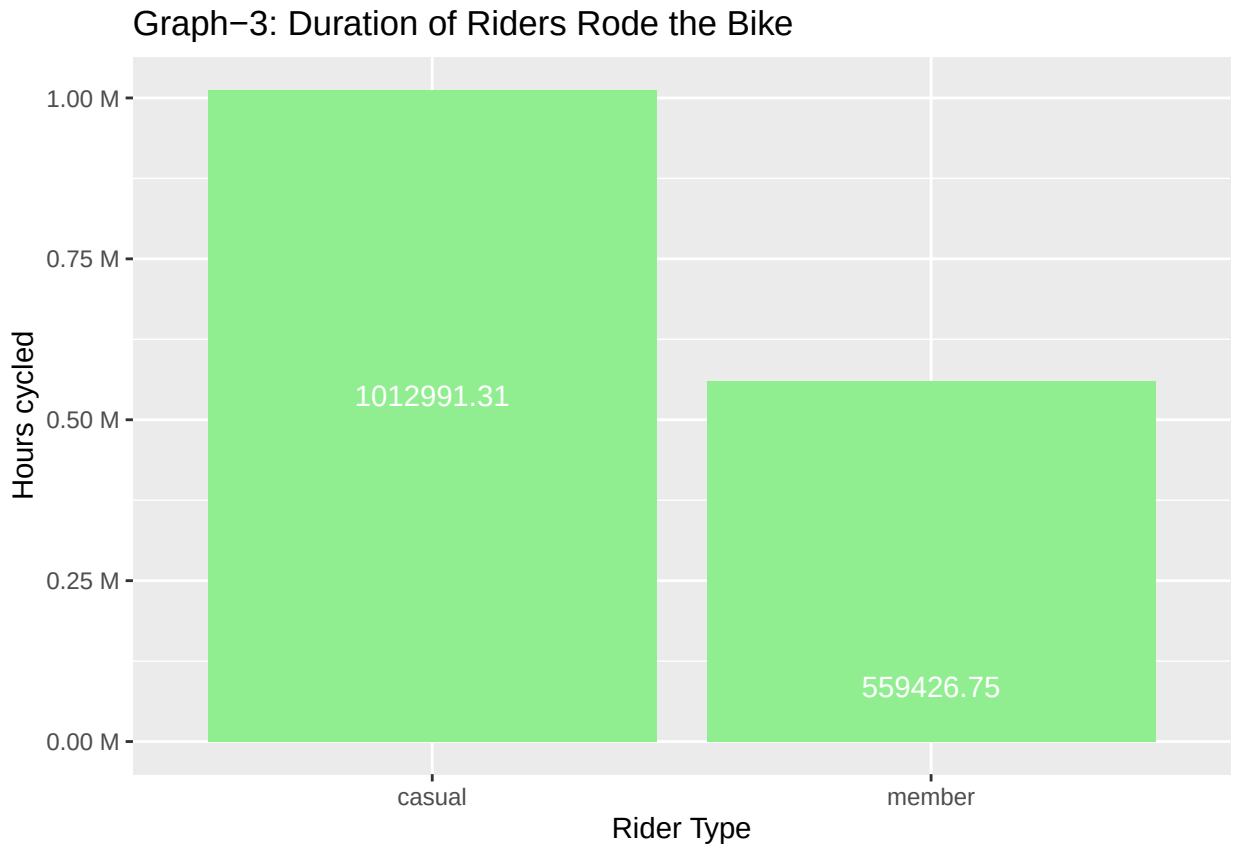
- New column named time_difference is created with unit in hours to know how long the rides have been taken
- 57% more casual riders rode bikes for longer duration than the membership holders.
- The average time difference for casual riders is 0.29hr more than the members.

```
Trip1 = Trip1 %>%
  mutate(Trip1, time_difference = difftime(ended_at, started_at, units = "hours"))

Trip3 = Trip1 %>%
  group_by(member_casual) %>%
  summarise(duration=sum(time_difference))
```

```
ggplot(Trip3, aes(x=member_casual, y=duration)) +
  geom_bar(stat = "identity", fill = "lightgreen") +
  labs(title = "Graph-3: Duration of Riders Rode the Bike",x = "Rider Type",y = "Hours cycled") +
  scale_y_continuous(labels = label_number(suffix = " M", scale = 1e-6)) +
  geom_text(aes(label=round(duration,2)), vjust=15, color="white")
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```




```
Trip3.1 = Trip1 %>%
  group_by(member_casual) %>%
  summarise(average_time=mean(time_difference))
print(Trip3.1)
```

```
## # A tibble: 2 x 2
##   member_casual average_time
##   <chr>         <drtn>
## 1 casual        0.5016713 hours
## 2 member        0.2112586 hours
```

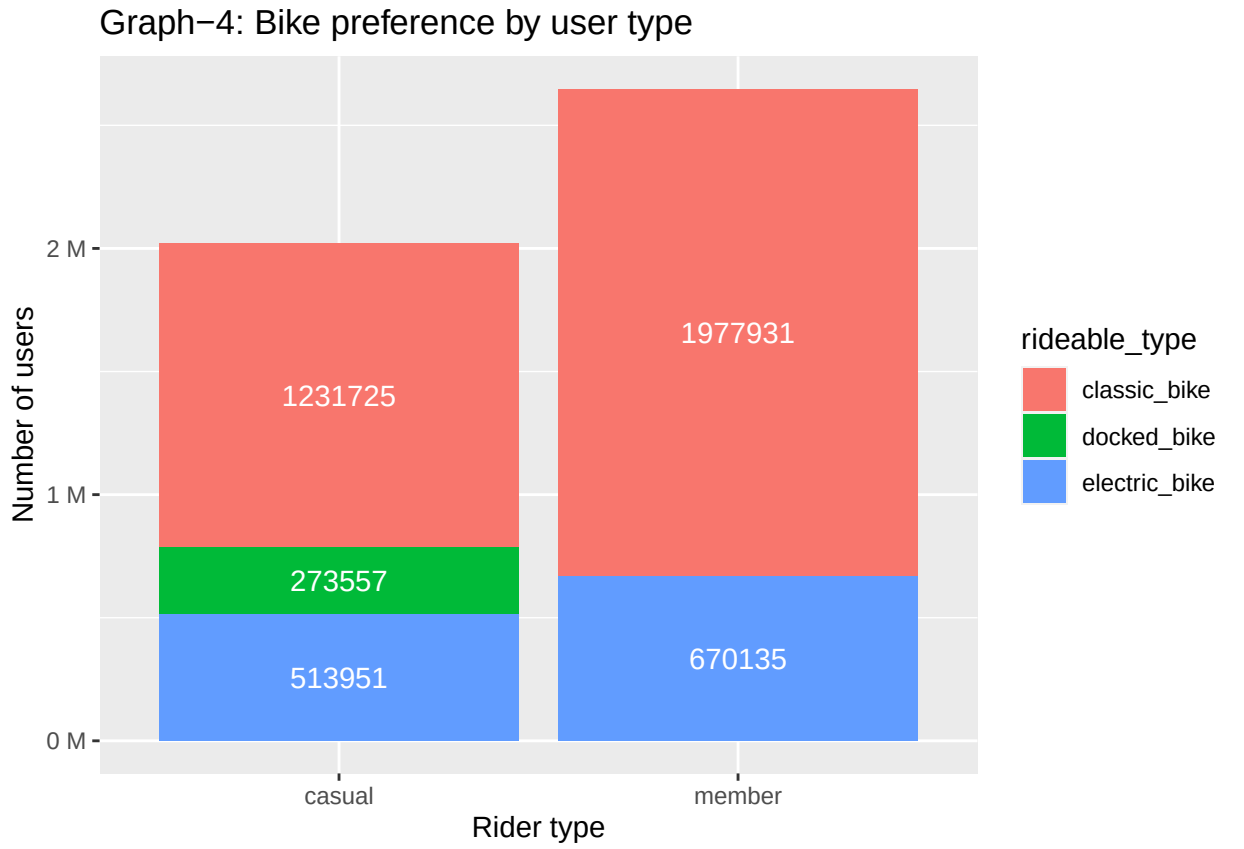
Bike Preference by The Rider Type

- Classic bike is the most preferred bike type among both of the rider type.

```
Trip4 = Trip1 %>%
  group_by(member_casual, rideable_type) %>%
  summarise(total_number_of_Riders = n())
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

```
ggplot(Trip4, aes(x=member_casual, y= total_number_of_Riders, fill=rideable_type)) +
  geom_bar(stat="identity") +
  labs(title = "Graph-4: Bike preference by user type",
       x = "Rider type",y = "Number of users") +
  geom_text(aes(label=total_number_of_Riders), position = position_stack(vjust = .5), color="white") +
  scale_y_continuous(labels = label_number(suffix = " M", scale = 1e-6))
```



Number of Rides Completed Monthly by Rider Type

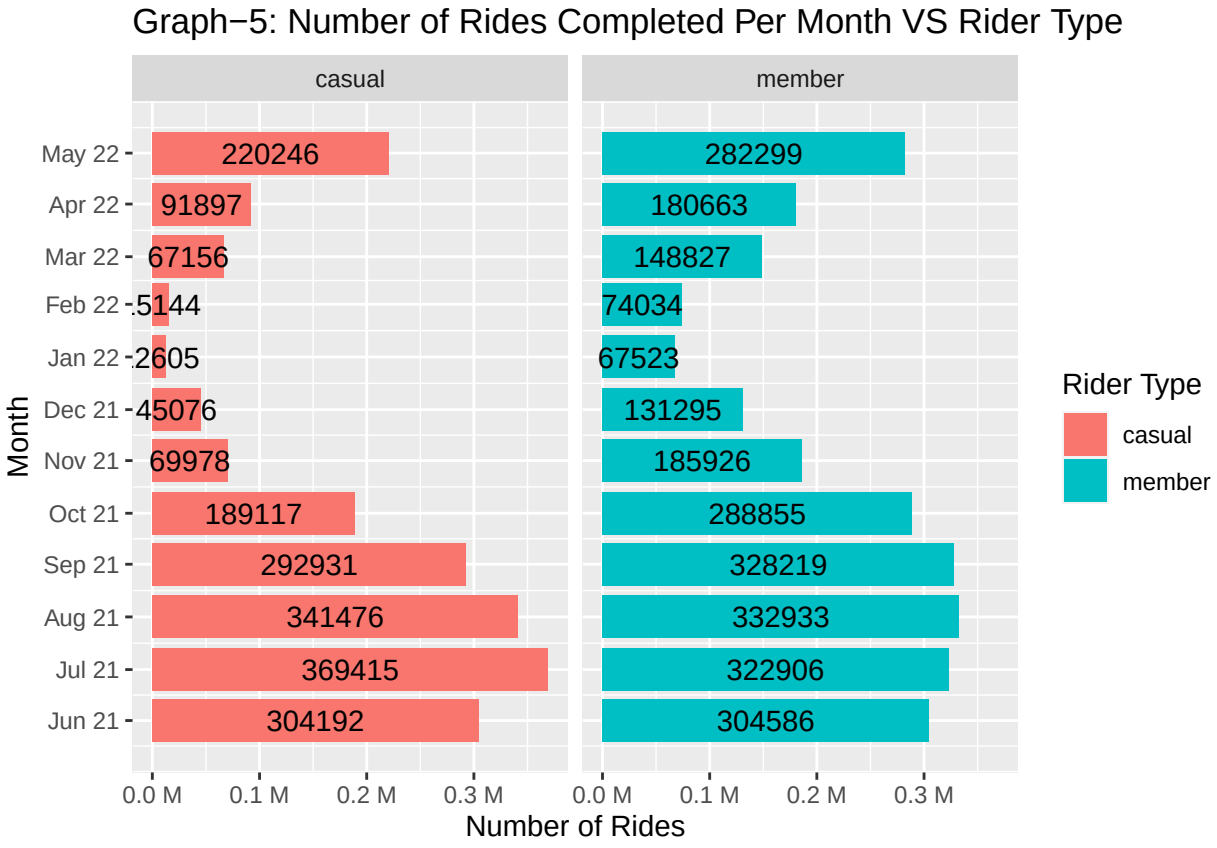
- There is increase in the use of bikes from month of June to September due to summer season as people go to visit places as its vacation.
- A sudden decrease is seen from month of October to February, may be due to the winter season which might be saying that people don't get more out in winter.

```
Trip5 = Trip1 %>%
  mutate(Trip1, start_month_year = floor_date(as_date(started_at), "month")) %>%
```

```
  group_by(start_month_year, member_casual) %>%
  summarise(number_of_rides = n())
```

```
## `summarise()` has grouped output by 'start_month_year'. You can override using
## the `.groups` argument.
```

```
ggplot(Trip5, aes(x=start_month_year, y=number_of_rides, fill=member_casual))+
  geom_bar(stat="identity") +
  labs(title = "Graph-5: Number of Rides Completed Per Month VS Rider Type",x = "Month",y = "Number of
  geom_text(aes(label=number_of_rides), position = position_stack(vjust = .5), color="black", angle = 3
  scale_y_continuous(labels = label_number(suffix = " M", scale = 1e-6)) +
  scale_x_date(date_labels = "%b %y", date_breaks = "1 month") +facet_wrap(~member_casual) + coord_flip
```



Number of Rides Completed: Per Day VS Rider Type

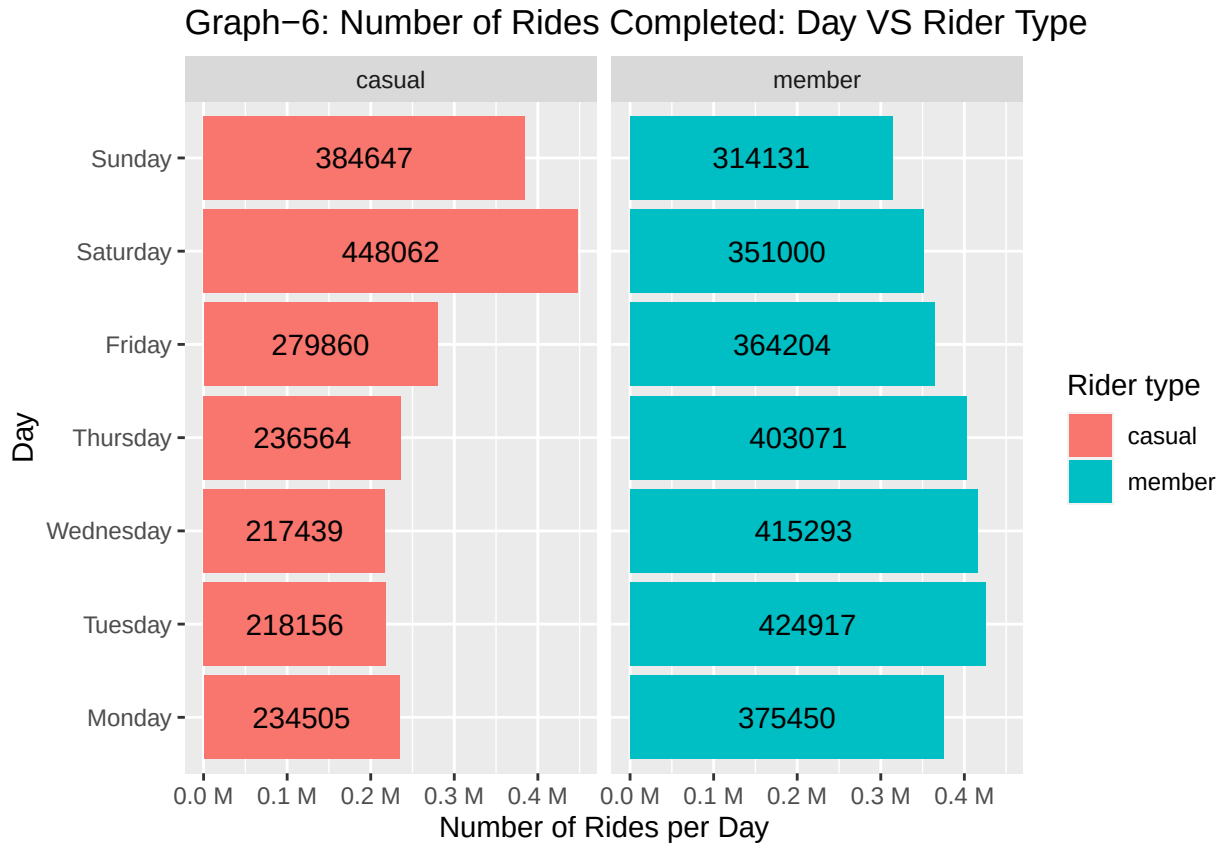
- The plot shows that during week days the number of riders with membership is higher than casual members, this may mean that the membership holders might ride for non-leisure purpose mostly work.
- And, during the weekends the numbers of riders who are not members is higher than membership holders, this suggests that the casual riders used the bikes for leisure purposes.

```
Trip6 =Trip1 %>%
  mutate(Trip1, day = weekdays(started_at)) %>%
  group_by(day, member_casual) %>%
  summarise(Number_of_Rides = n())
```

```
## `summarise()` has grouped output by 'day'. You can override using the `.groups`
## argument.
```

```
week = c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday')

ggplot(Trip6, aes(x=factor(day, level = week), y=Number_of_Rides, fill=member_casual))+
  geom_bar(stat="identity") +
  labs(title = "Graph-6: Number of Rides Completed: Day VS Rider Type",x = "Day",y = "Number of Rides p
  geom_text(aes(label=Number_of_Rides), position = position_stack(vjust = .5), color="black", angle = 3
  scale_y_continuous(labels = label_number(suffix = " M", scale = 1e-6))+facet_wrap(~member_casual)+ co
```




```
T = Trip1 %>%
  group_by(member_casual, start_station_name) %>%
  summarise(number_rides=n(),.groups ='drop' ) %>%
  arrange(desc(number_rides))
View(T)
T2 = filter(T,member_casual == 'member') %>% slice(1:5)
head(T2)
```

Finding out Top5 Stations

```
## # A tibble: 5 x 3
```

```
##   member_casual start_station_name       number_rides
##   <chr>         <chr>                           <int>
## 1 member        Kingsbury St & Kinzie St        25066
## 2 member        Clark St & Elm St               23893
## 3 member        Wells St & Concord Ln           23148
## 4 member        Wells St & Elm St               20239
## 5 member        Clinton St & Madison St         18573
```

```
T3 = filter(T,member_casual == 'casual') %>% slice(1:5)
View(T3)
```

Top Three recommendations based on the Analysis:

Following are some recommendations which can be applied to increase the membership holders:

- Introduce special offers for membership holders
- Put restrictions on the use of bikes for casual riders, such as limit the duration casual riders can ride bikes.
- As casual riders use bikes more on weekends, increase the rates on those days.