Increase Members for Ride Sharing App

Thahseen AG

10/31/2021

In this case study we will analyze data from Divvy Bike Sharing App.

We will use R to analyse and visualize the relationship between casual and annual member riders. As the director of marketing believes that maximizing the number of annual member will be key to future growth, which has been backed by the financial analyst which states annual members are much more profitable that cacual members.

The Question trying to be answered:

- How do annual members and casual riders use Cyclists bikes differently?

Packages Required:

- **Tidyverse** : for data import and wrangling
- **lubridate** : for data functions
- **ggplot** : for data visualization
- **dplyr** : for data manipulation

Importing CSV files for last 12 months

We will be analyzing on the last 12 months data, September 2020-September 2021.

```
## Rows: 756147 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
```

```
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Wrangle And Combine Data Into a Single File

As we want to combine the data for the last 12 months and then analyse, first we check for any discrepancy among the data set, from different column names, difference in number of columns or difference in data types which will prevent the merging process.

- Compare column names for last 12 month reports
  - All have same 13 column names
- Check the structure of all the reports

```
colnames(Sep_2021)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
str(Sep_2020)
```

```
## spec_tbl_df [532,958 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:532958] "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC778E" "57F6
##  $ rideable_type     : chr [1:532958] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:532958], format: "2020-09-17 14:27:11" "2020-09-17 15:07:31" ...
##  $ ended_at          : POSIXct[1:532958], format: "2020-09-17 14:44:24" "2020-09-17 15:07:45" ...
##  $ start_station_name: chr [1:532958] "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway" "W Oakda
##  $ start_station_id  : num [1:532958] 52 NA NA 246 24 94 291 NA NA NA ...
##  $ end_station_name  : chr [1:532958] "Green St & Randolph St" "W Oakdale Ave & N Broadway" "W Oakda
##  $ end_station_id    : num [1:532958] 112 NA NA 249 24 NA 256 NA NA NA ...
##  $ start_lat         : num [1:532958] 41.9 41.9 41.9 42 41.9 ...
##  $ start_lng         : num [1:532958] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:532958] 41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num [1:532958] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:532958] "casual" "casual" "casual" "casual" ...
##  - 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_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(Sep_2021)
```

```
## spec_tbl_df [756,147 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:756147] "9DC7B962304CBFD8" "F930E2C6872D6B32" "6EF72137900BB910" "78D1
##  $ rideable_type     : chr [1:756147] "electric_bike" "electric_bike" "electric_bike" "electric_bike
```

```
##  $ started_at        : POSIXct[1:756147], format: "2021-09-28 16:07:10" "2021-09-28 14:24:51" ...
##  $ ended_at          : POSIXct[1:756147], format: "2021-09-28 16:09:54" "2021-09-28 14:40:05" ...
##  $ start_station_name: chr [1:756147] NA NA NA NA ...
##  $ start_station_id  : chr [1:756147] NA NA NA NA ...
##  $ end_station_name  : chr [1:756147] NA NA NA NA ...
##  $ end_station_id    : chr [1:756147] NA NA NA NA ...
##  $ start_lat         : num [1:756147] 41.9 41.9 41.8 41.8 41.9 ...
##  $ start_lng         : num [1:756147] -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:756147] 41.9 42 41.8 41.8 41.9 ...
##  $ end_lng           : num [1:756147] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:756147] "casual" "casual" "casual" "casual" ...
##  - 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>
```

- The Data Sets Sep 2020 to Nov 2020 has start station id and end station id as double. Converting start_station_id and end_station_id to character type for merging all the data set to one.

```
## spec_tbl_df [131,573 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:131573] "70B6A9A437D4C30D" "158A465D4E74C54A" "5262016E0F1F2F9A" "BE11
##  $ rideable_type     : chr [1:131573] "classic_bike" "electric_bike" "electric_bike" "electric_bike"
##  $ started_at        : POSIXct[1:131573], format: "2020-12-27 12:44:29" "2020-12-18 17:37:15" ...
##  $ ended_at          : POSIXct[1:131573], format: "2020-12-27 12:55:06" "2020-12-18 17:44:19" ...
##  $ start_station_name: chr [1:131573] "Aberdeen St & Jackson Blvd" NA NA NA ...
##  $ start_station_id  : chr [1:131573] "13157" NA NA NA ...
##  $ end_station_name  : chr [1:131573] "Desplaines St & Kinzie St" NA NA NA ...
##  $ end_station_id    : chr [1:131573] "TA1306000003" NA NA NA ...
##  $ start_lat         : num [1:131573] 41.9 41.9 41.9 41.9 41.8 ...
##  $ start_lng         : num [1:131573] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:131573] 41.9 41.9 41.9 41.9 41.8 ...
##  $ end_lng           : num [1:131573] -87.6 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:131573] "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>
```

- Combing all 12 months individual report to 1 report

```
Last_12_months <- bind_rows(Sep_2021,Aug_2021,Jul_2021,Jun_2021,May_2021,Apr_2021,Mar_2021,Feb_2021,Jan
```

Cleaning Data to Prepare for Analysis

First we will analyze just on how the casual and annual members use the ride differently irrespective of the geographical location. Hence we will remove information regarding station id, name and latitude and longitude details.

- Removing columns not required for analysis

```
Last_Year<-Last_12_months %>% select(-c(start_station_name,start_station_id,end_station_name,end_statio
colnames(Last_Year)
```

```
## [1] "ride_id"       "rideable_type" "started_at"    "ended_at"
## [5] "start_lat"     "start_lng"     "member_casual"
```

```
dim(Last_Year)
```

```
## [1] 5136261       7
```

```
head(Last_Year)
```

```
## # A tibble: 6 x 7
##   ride_id        rideable_type started_at          ended_at            start_lat
##   <chr>          <chr>         <dttm>              <dttm>                  <dbl>
## 1 9DC7B962304CB~ electric_bike 2021-09-28 16:07:10 2021-09-28 16:09:54      41.9
## 2 F930E2C6872D6~ electric_bike 2021-09-28 14:24:51 2021-09-28 14:40:05      41.9
## 3 6EF72137900BB~ electric_bike 2021-09-28 00:20:16 2021-09-28 00:23:57      41.8
## 4 78D1DE133B3DB~ electric_bike 2021-09-28 14:51:17 2021-09-28 15:00:06      41.8
## 5 E03D4ACDCAEF6~ electric_bike 2021-09-28 09:53:12 2021-09-28 10:03:44      41.9
## 6 346DE323A2677~ electric_bike 2021-09-28 01:53:18 2021-09-28 02:00:02      41.9
## # ... with 2 more variables: start_lng <dbl>, member_casual <chr>
```

```
summary(Last_Year)
```

```
##    ride_id          rideable_type        started_at
##  Length:5136261     Length:5136261     Min.   :2020-10-01 00:00:06
##  Class :character   Class :character   1st Qu.:2021-04-11 18:50:57
##  Mode  :character   Mode  :character   Median :2021-06-21 18:01:31
##                                        Mean   :2021-05-25 22:30:57
##                                        3rd Qu.:2021-08-11 21:13:51
##                                        Max.   :2021-09-30 23:59:48
##     ended_at                      start_lat       start_lng
##  Min.   :2020-10-01 00:05:09   Min.   :41.64   Min.   :-87.84
##  1st Qu.:2021-04-11 19:15:05   1st Qu.:41.88   1st Qu.:-87.66
##  Median :2021-06-21 18:20:59   Median :41.90   Median :-87.64
##  Mean   :2021-05-25 22:51:34   Mean   :41.90   Mean   :-87.65
##  3rd Qu.:2021-08-11 21:33:57   3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :2021-10-01 22:55:35   Max.   :42.08   Max.   :-87.52
##  member_casual
##  Length:5136261
##  Class :character
##  Mode  :character
##
##
##
```

*We will calculate how long each trip has take for this we need to calculate the difference between the start and the end time.

- As the starting and ending time is given as datetime object we will add new column "trip_duration" by calculation time difference between the trip end and start time in seconds

```
Last_Year$trip_duration<-difftime(Last_Year$ended_at,Last_Year$started_at)
```

- We want to see how different the casual and annual members differ during each weekdays.
- To get the weekdays when each trip was taken we will use the method weekdays()
- Calculating the Day of the week for each trips

```
Last_Year$day_of_week<-weekdays(Last_Year$started_at)
```

- Checking if there is any trip duration calculated as negative or zero and removing those from the data set.

```
Last_Year %>% count(trip_duration<=0)
```

```
## # A tibble: 2 x 2
##   `trip_duration <= 0`       n
##   <lgl>                  <int>
## 1 FALSE                5132499
## 2 TRUE                    3762
```

```
Last_Year_v2<-Last_Year[!(Last_Year$trip_duration<=0),]
```

Descriptive Analysis

We will analyse each type users and analyse the average duration of the trip,median max and minimum duration of trips.

- Comparing the Casual and Members of the Cyclist App

```
aggregate(Last_Year_v2$trip_duration~Last_Year_v2$member_casual,FUN=mean)
```

```
##   Last_Year_v2$member_casual Last_Year_v2$trip_duration
## 1                     casual             1970.5946 secs
## 2                     member              850.1194 secs
```

```
aggregate(Last_Year_v2$trip_duration~Last_Year_v2$member_casual,FUN=median)
```

```
##   Last_Year_v2$member_casual Last_Year_v2$trip_duration
## 1                     casual                  1000 secs
## 2                     member                   606 secs
```

```
aggregate(Last_Year_v2$trip_duration~Last_Year_v2$member_casual,FUN=max)
```

```
##   Last_Year_v2$member_casual Last_Year_v2$trip_duration
## 1                     casual               3356649 secs
## 2                     member                573467 secs
```

```
aggregate(Last_Year_v2$trip_duration~Last_Year_v2$member_casual,FUN=min)
```

```
##   Last_Year_v2$member_casual Last_Year_v2$trip_duration
## 1                     casual                     1 secs
## 2                     member                     1 secs
```

*Analyze the average trip duration for each type of users on each day of the week

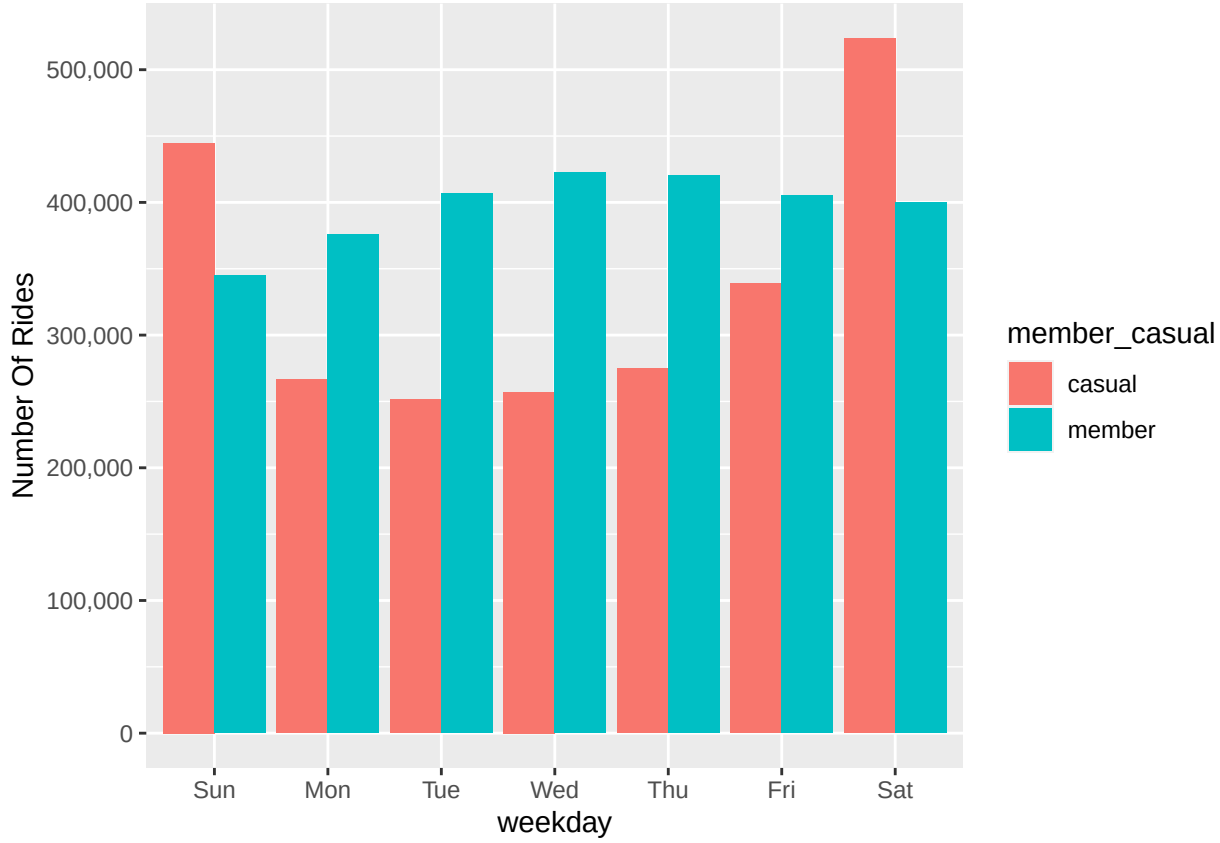
```
Last_Year_v2 %>% mutate(weekday = wday(started_at,label=TRUE)) %>%
    group_by(member_casual, weekday) %>%
      summarise(number_of_rides = n(),average_duration = mean(trip_duration)) %>%
        arrange(member_casual,weekday)
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int> <drtn>
##  1 casual        Sun              444899 2284.4402 secs
##  2 casual        Mon              266394 1953.2472 secs
##  3 casual        Tue              251449 1759.7402 secs
##  4 casual        Wed              257161 1718.2030 secs
##  5 casual        Thu              274615 1702.9358 secs
##  6 casual        Fri              339111 1883.5209 secs
##  7 casual        Sat              523626 2134.7323 secs
##  8 member        Sun              345033  962.7708 secs
```

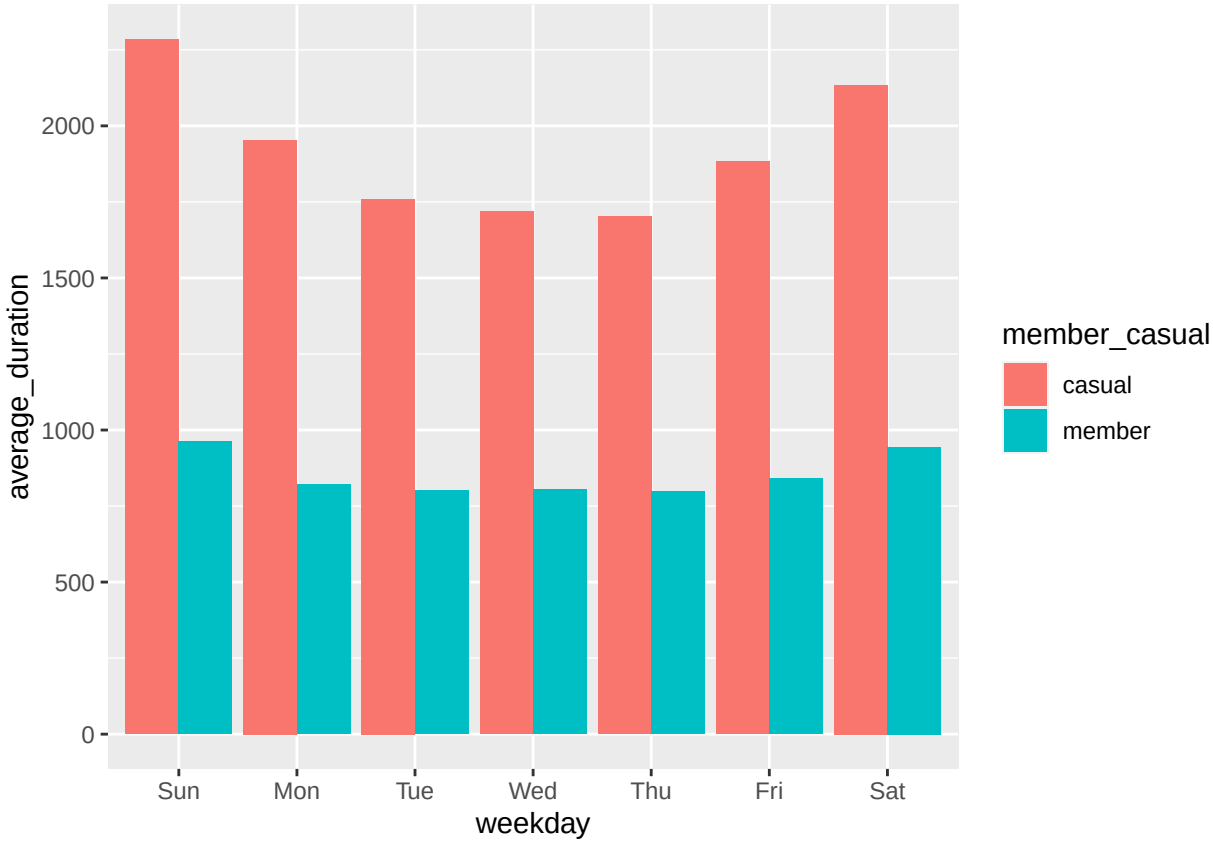
```
##  9 member        Mon              375661  820.6713 secs
## 10 member        Tue              406632  800.0397 secs
## 11 member        Wed              422539  804.3196 secs
## 12 member        Thu              420288  797.8204 secs
## 13 member        Fri              405231  841.1131 secs
## 14 member        Sat              399860  944.0037 secs
```

Visualizing Data

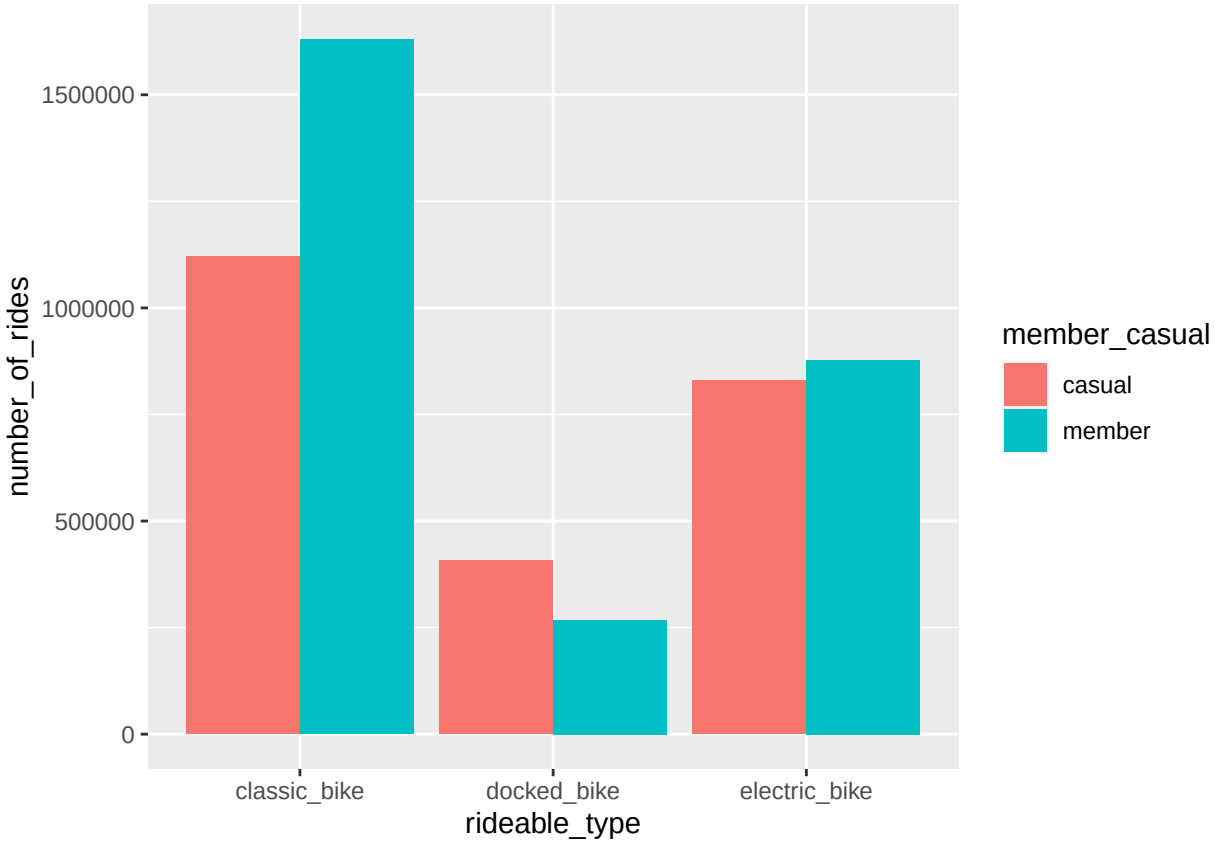
Visualizing by the number of rides for each type of users for each day of the week



Visualizing by the trip duration for each type of users for each day of the week



Visualizing the Ride Type used by Casual and Member users



From the first part of analysis we can see that the number of casual riders are usually more during weekends and they prefer longer duration trips.

Recommendations

- Increasing membership incentives for Long duration trips and separate campaign and offers on weekends for member users to increase annual memberships.