Case_study_Bike_share

Ji

2022-09-28

Case study: How Does a Bike-Share Navigate Speedy Success?

Introduction

This case study is set for Cyclistic, a bike-share company in Chicago. The study aims to discover the different usage of Cyclistic bikes of casual riders and annual members. By review the results, suggestions will be given on how to maximize the number of annual memberships.

Description of data sources used

Data was downloaded from this website: https://divvy-tripdata.s3.amazonaws.com/index.html (https://divvy-tripdata.s3.amazonaws.com/index.html). The data has been made available by Motivate International Inc.under this license: https://ride.divvybikes.com/data-license-agreement (https://ride.divvybikes.com/data-license-agreement). The data is composed by ride_id, rideable_type, start and end time, station name, station id, station location, and the member type. It provides riding counts and riding time/duration for both casual rider and annual members. The limitation is that it does not cover the cost of each ride. The time period of the data was from 2021/09 to 2022/08. Any time before or after this time period was not used for analysis.

Cleaning and data manipulation

A R file is presented in this repository for repeatable study. ### Load and combine data of different month into one data frame The data was saved as zip files by each month. It is better to combine them into one data frame for cleaning and analysis.

```
## check if the directory exists
if(!file.exists("./raw_data")){
  dir.create("./raw_data")
}

## load ggplot2 and patchwork for visualization
library(ggplot2)
#install.packages("patchwork")
#library(patchwork)
## down load the zip files
url<-"https://divvy-tripdata.s3.amazonaws.com"
file_name <- c("/202109-divvy-tripdata.zip",
               "/202110-divvy-tripdata.zip",
               "/202111-divvy-tripdata.zip",
               "/202112-divvy-tripdata.zip",
               "/202201-divvy-tripdata.zip",
               "/202202-divvy-tripdata.zip",
               "/202203-divvy-tripdata.zip",
               "/202204-divvy-tripdata.zip",
               "/202205-divvy-tripdata.zip",
               "/202206-divvy-tripdata.zip",
               "/202207-divvy-tripdata.zip",
               "/202208-divvy-tripdata.zip")
## unzip
for (i in file_name){
  download.file(paste(url,i,sep=""),paste("./raw_data/",i,sep=""))
  unzip(zipfile=paste("./raw_data",i,sep=""),exdir="./raw_data")
}

## read all csv files into one data frame: data_all
list_csv_files <- list.files(path = "./raw_data/",pattern="*.csv")
data_all <-data.frame()
for (i in 1:12){
  data_all<-rbind(data_all,read.csv(paste("./raw_data/",list_csv_files[i],sep="")))
}
```

The data_all is composed of 5883043 observations and 13 variables.

```
## view data_all
head(data_all) ## ride_id, type, station, latitude and longitude, member type
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 9DC7B962304CBFD8 electric_bike 2021-09-28 16:07:10 2021-09-28 16:09:54
## 2 F930E2C6872D6B32 electric_bike 2021-09-28 14:24:51 2021-09-28 14:40:05
## 3 6EF72137900BB910 electric_bike 2021-09-28 00:20:16 2021-09-28 00:23:57
## 4 78D1DE133B3DBF55 electric_bike 2021-09-28 14:51:17 2021-09-28 15:00:06
## 5 E03D4ACDCAEF6E00 electric_bike 2021-09-28 09:53:12 2021-09-28 10:03:44
## 6 346DE323A2677DC0 electric_bike 2021-09-28 01:53:18 2021-09-28 02:00:02
##   start_station_name start_station_id end_station_name end_station_id start_lat
## 1                                                                         41.89
## 2                                                                         41.94
## 3                                                                         41.81
## 4                                                                         41.80
## 5                                                                         41.88
## 6                                                                         41.87
##   start_lng end_lat end_lng member_casual
## 1    -87.68   41.89  -87.67        casual
## 2    -87.64   41.98  -87.67        casual
## 3    -87.72   41.80  -87.72        casual
## 4    -87.72   41.81  -87.72        casual
## 5    -87.74   41.88  -87.71        casual
## 6    -87.75   41.88  -87.74        casual
```

```
str(data_all) ## dates are chr, needs to convert to date when use
```

```
## 'data.frame':    5883043 obs. of  13 variables:
##  $ ride_id           : chr  "9DC7B962304CBFD8" "F930E2C6872D6B32" "6EF72137900BB910" "78D1DE
133B3DBF55" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike"
...
##  $ started_at        : chr  "2021-09-28 16:07:10" "2021-09-28 14:24:51" "2021-09-28 00:20:1
6" "2021-09-28 14:51:17" ...
##  $ ended_at          : chr  "2021-09-28 16:09:54" "2021-09-28 14:40:05" "2021-09-28 00:23:5
7" "2021-09-28 15:00:06" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.8 41.8 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 42 41.8 41.8 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
nrow(data_all) ##5883043
```

```
## [1] 5883043
```

Data cleaning and manupulation

Steps and results were shown below: 1. Remove duplication: there is 0 duplication of the data.

```
## any duplication of data?
nrow(data_all[duplicated(data_all$ride_id),]) # no duplication
```

```
## [1] 0
```

2. Make time stamp consistent: POSIXct time with year-month-day hour:min:second format is used for all the time stamp. I assumed the time in the original data is the local time in Chicago.

```
## is time stamp consistent? The time unit is not specified, I assumed it used
## the Chicago local time, which is EST.
data_all$started_at <- as.POSIXct(data_all$started_at, format="%Y-%m-%d %H:%M:%S")
data_all$ended_at <- as.POSIXct(data_all$ended_at, format="%Y-%m-%d %H:%M:%S")
class(data_all$started_at)
```

```
## [1] "POSIXct" "POSIXt"
```

```
class(data_all$ended_at)
```

```
## [1] "POSIXct" "POSIXt"
```

3. Extract year_month and day from the started_time, and make new columns of them.

```
## add the year-month as a column to the very right of the data, this will be used to group the
data
data_all$year_month<-format(data_all$started_at, "%y/%m")
## add the weekdays
data_all$weekday<-weekdays(data_all$started_at)
```

4. Solve for riding_time by subtracting started_time from ended_time, and make a new column of it.

```
## add the riding time
data_all$riding_time <-difftime(data_all$ended_at,data_all$started_at,units = "mins")
```

5. Remove riding_time <= 0: 606 rows of data were removed.

```
## remove riding_time less than 0
data_all_v1 <- data_all[!(data_all$riding_time <= 0),]
nrow(data_all_v1) # 5882437
```

```
## [1] 5882437
```

6. Remove rides with abnormal riding time: any riding time larger than 1 day were removed from the data. 5198 rows of data were removed.

```
## remove irregular riding time
data_all_v1$riding_time<-as.numeric(data_all_v1$riding_time)
summary(data_all_v1$riding_time) ## max 40705.02 weird data
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.02     6.05    10.72    19.76    19.33 40705.02
```

```
# remove riding_time larger than 1 day = 1440 min
data_all_v1<-data_all_v1[which(data_all_v1$riding_time<=1440),]
nrow(data_all_v1)  # 5877239
```

```
## [1] 5877239
```

```
summary(data_all_v1$riding_time)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0167    6.0500   10.7167   17.0034   19.3000 1439.3667
```

7. Remove unnecessary columns: station_id and location will not be used in this study. Related columns were removed and a new data frame was made: data_all_v1.

```
## remove unnecessary columns
data_all_v1 <- data_all_v1[,-which(names(data_all_v1) %in%
                           c("start_station_id","end_station_id","start_lat","start_lng","end_l
at","end_lng")),drop=FALSE]
```

8. Categorize the time to am_peak, pm_peak and other: the category of time was assigned to each row based on the started time. 6:00-8:00 is am peak, 16:00-18:00 is pm peak, the rest is other. A new column time_cat was made to show these values.

```
## add the time of rides use the start time
## assign am peak (6:00-8:00), pm peak (16:00-18:00), and other to time
## don't use for loop, it is very slow
data_all_v1$time <- as.numeric(format(data_all_v1$started_at, "%H"))
data_all_v1$time_cat<- NA

## use match
data_all_v1 <- data_all_v1[order(data_all_v1$time),]
am_match_result <- which(data_all_v1$time %in% c(6,7,8))
data_all_v1[am_match_result,"time_cat"]="am_peak"

pm_match_result <- which(data_all_v1$time %in% c(16,17,18))
data_all_v1[pm_match_result,"time_cat"]="pm_peak"

na_match_result <- which(is.na(data_all_v1$time_cat))
data_all_v1[na_match_result,"time_cat"]="other"
```

9. Make all string characters lower case. Replace any missing station name by a string “unknown station”.

```
## use all lower case for station name, replace missing value with "unknown station"
## is it NA or blank? it is "" (nothing)
data_all_v1$start_station_name <- tolower(data_all_v1$start_station_name)
data_all_v1$end_station_name <- tolower(data_all_v1$end_station_name)
missing <- which(data_all_v1$start_station_name %in% "")
data_all_v1$start_station_name[missing] = "unknown station"
missing <- which(data_all_v1$end_station_name %in% "")
data_all_v1$end_station_name[missing] = "unknown station"

## use all lower case for bike type
data_all_v1$rideable_type <- tolower(data_all_v1$rideable_type)
```

10. Check if the data is complete. There is no missing values. The data is complete. The data has no duplication, no missing values, no mixer of upper and lower cases, no abnormal data, and proper manipulation has been done for further analysis. The data is clean.

```
## check is the data frame is complete
complete <- complete.cases(data_all_v1)
length(which(complete=="TRUE")) #5877239
```

```
## [1] 5877239
```

```
nrow(data_all_v1) # no missing values, data is complete
```

```
## [1] 5877239
```

11. Save the cleaned data.

```
## save this data frame
write.csv(data_all_v1,"./data_all_v1.csv",row.names = FALSE)
```

Summary of analysis

Descriptive analysis

The data has 2463604 casual users and 3413635 members. The ratio of number of members/number of casual users is 1.38. It should provide unbiased information of the two groups.

```
table(data_all_v1$member_casual)
```

```
##
##  casual  member
## 2463604 3413635
```

High-level comparision between casual users and members

Casual user and members have very close min and max riding time. The difference is in the median and mean. The median and mean riding time of members is smaller than that of the casual users.

```
summary(data_all_v1[which(data_all_v1$member_casual=="member"),"riding_time"])
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0167    5.1833    8.9667   12.5803   15.5500 1435.4667
```

```
summary(data_all_v1[which(data_all_v1$member_casual=="casual"),"riding_time"])
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0167    7.9000   13.8833   23.1322   25.4333 1439.3667
```

```
## what is the mean riding time for casual/member?
tapply(data_all_v1$riding_time,data_all_v1$member_casual,mean) # casual 23,member 12
```

```
##   casual   member
## 23.13218 12.58025
```

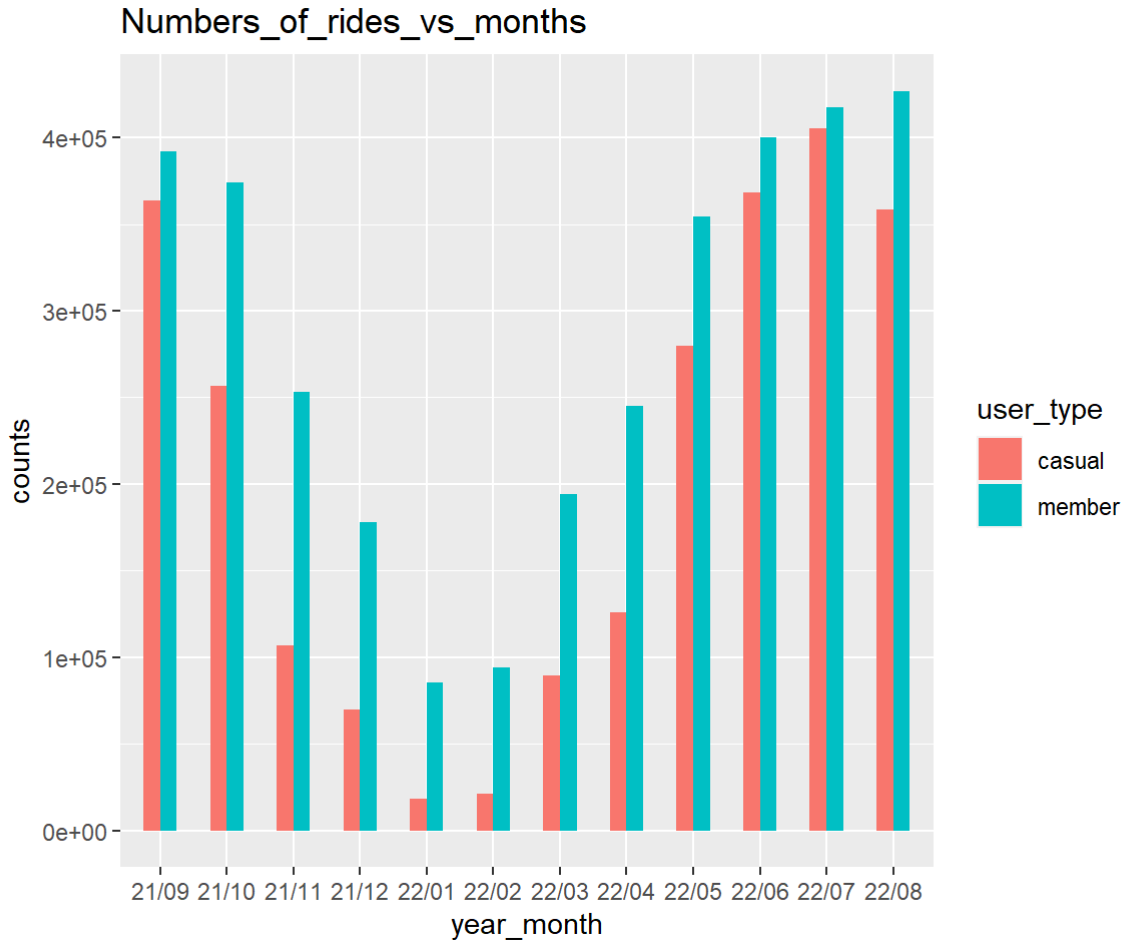
Visualizations and key findings

Key findings 1. Time of each ride. Members have consistent riding time for different months, days and time. Casual users always have longer riding time than members, and show a peak of riding time in June to September, and on weekends. 2. Numbers of rides. Members have more rides than casual users on weekdays. Casual users use more on weekends. 3. Peak hours. The peak hours for casual users are 16:00-18:00 (pm peak). 4. Peak locations. The top 10 stations that have more rides for casual users are: streeter dr & grand ave
dusable lake shore dr & monroe st
millennium park
michigan ave & oak st
dusable lake shore dr & north blvd shedd aquarium
theater on the lake
wells st & concord ln
clark st & armitage ave
clark st & lincoln ave. 5. Bike type. casual users likes electric bikes the most, followed by classic bikes. They don’t use a lot docked bikes.

Comparision between casual users and members for different months

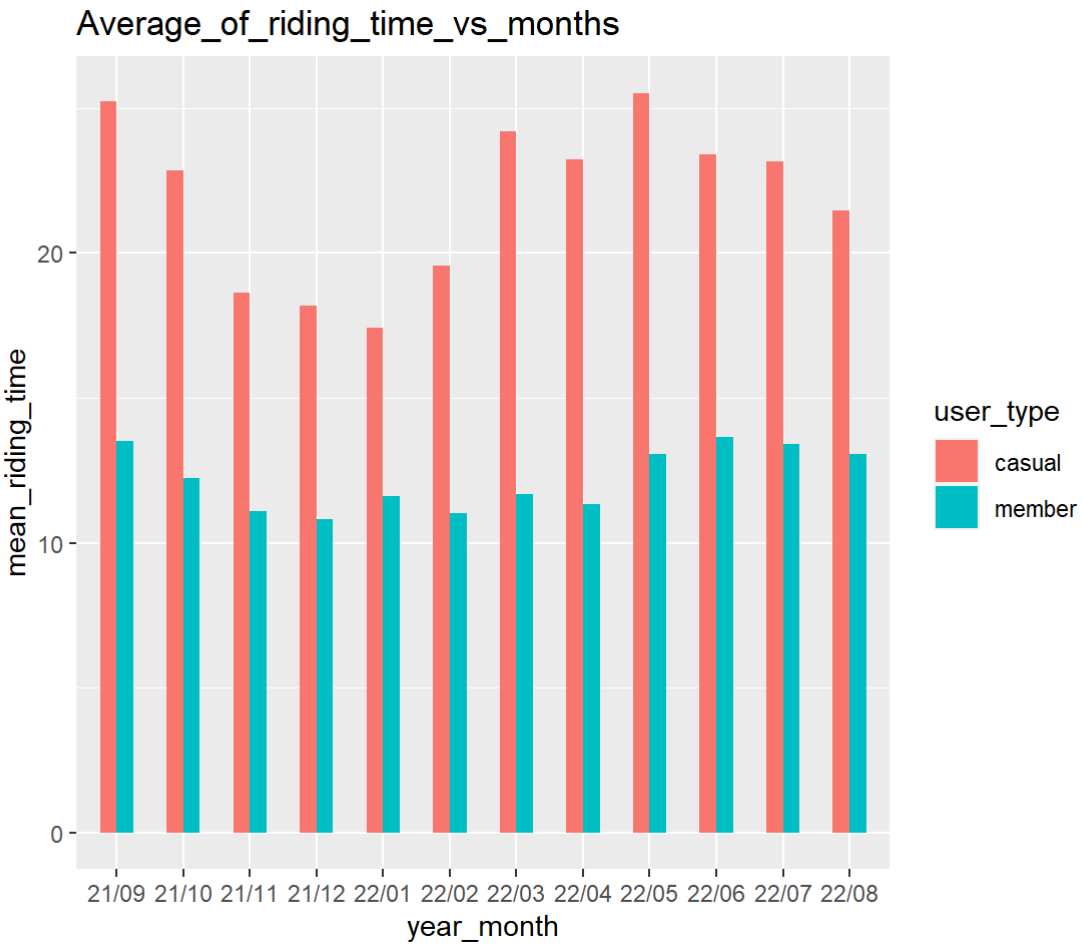
As shown in the Numbers_of_rides_vs_months figure, members always have more rides than casual users. Members and casual users both have **more numbers of rides in summer**, and low numbers of rides in winter. The months with top usages for casual users are 22/07, 22/06, 21/09 and 22/08.

```
## let's see usage for members and casual users for different months
year_month_count<-as.data.frame(table(data_all_v1$member_casual,data_all_v1$year_month))
colnames(year_month_count)<-c("user_type","year_month","counts")
write.csv(year_month_count,"./year_month_count.csv",row.names = FALSE)

## mean
year_month_mean<-as.data.frame(aggregate(data_all_v1$riding_time,
                                         list(data_all_v1$member_casual,data_all_v1$year_mont
h), FUN=mean))
colnames(year_month_mean)<-c("user_type","year_month","mean_riding_time")
write.csv(year_month_mean,"./year_month_mean.csv",row.names = FALSE)

## visualization
ggplot(data=year_month_count,aes(x = year_month, y = counts, fill = user_type))+geom_col(width=
0.5, position = position_dodge(width=0.5))+ theme(aspect.ratio = 1) + ggtitle("Numbers_of_rides
_vs_months")
```



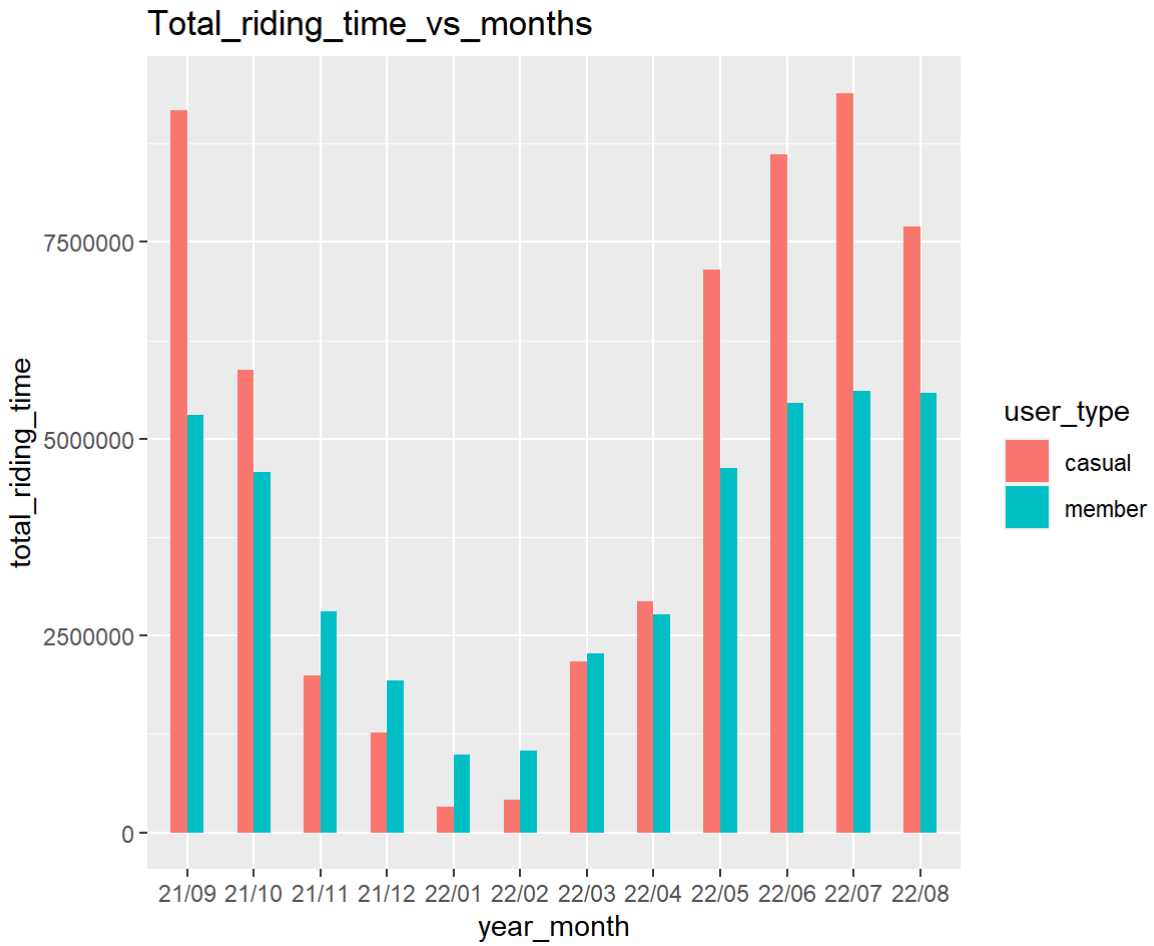
As shown in the Average_of_riding_time_vs_months figure, members have consistent **average riding time** all over the past year. In contrast, casual users didn't ride longer than 20 minutes from 21/11-22/02.

```
ggplot(data=year_month_mean,aes(x = year_month, y = mean_riding_time, fill = user_type))+geom_c
ol(width=0.5, position = position_dodge(width=0.5))+ theme(aspect.ratio = 1)+ggtitle("Average_o
f_riding_time_vs_months")
```



As shown in the total_riding_time_vs_months figure, the total riding time of members is similar or higher than casual users from 21/11-22/04. In contrast the total riding time of casual users are higher in 21/09, and 22/05-22/08.

```
total_by_month<-as.data.frame(aggregate(data_all_v1$riding_time,list(data_all_v1$member_casual, data_all_v1$year_month), FUN=sum))
colnames(total_by_month)<-c("user_type","year_month","total_riding_time")
ggplot(data=total_by_month,aes(x = year_month, y = total_riding_time, fill = user_type))+geom_col(width=0.5, position = position_dodge(width=0.5))+ theme(aspect.ratio = 1)+ggtitle("Total_riding_time_vs_months")
```

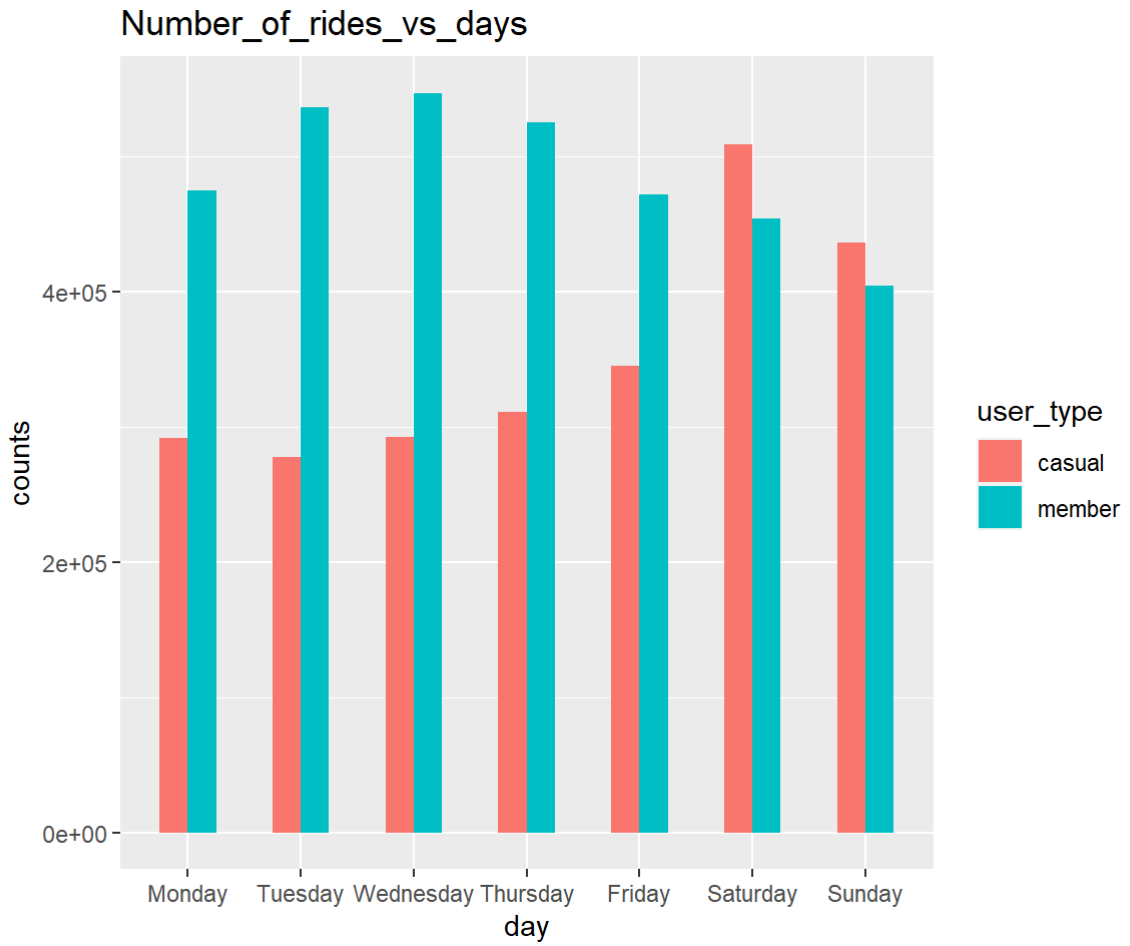


Conclusion of monthly analysis: casual users are active from June to September.

Comparision between casual users and members for different days

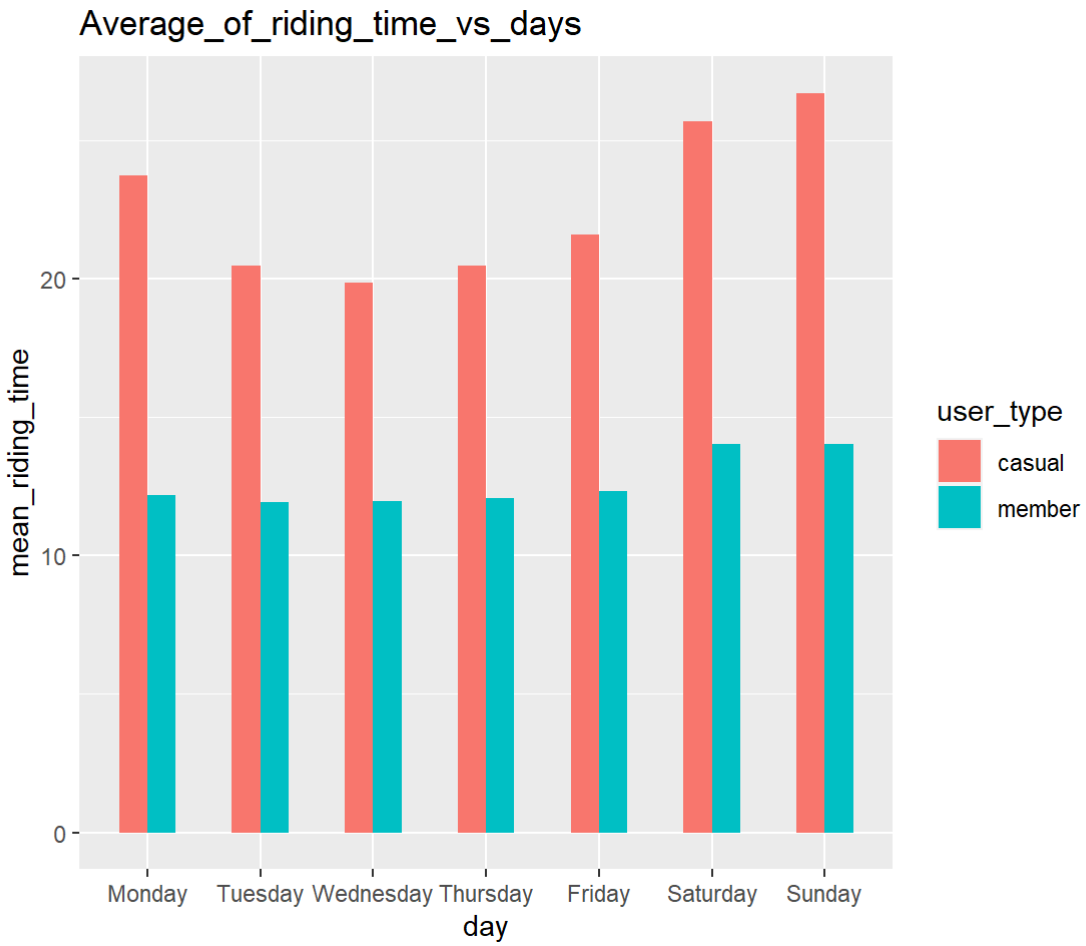
As shown in the Number_of_rides_vs_days figure, members have more rides on weekdays. In contrast, casual users have more rides on weekends.

```
day_count<-as.data.frame(table(data_all_v1$member_casual,data_all_v1$weekday))
colnames(day_count)<-c("user_type","day","counts")
day_count$day <- ordered(day_count$day, levels=c("Monday", "Tuesday", "Wednesday", "Thursday",
"Friday", "Saturday", "Sunday"))
day_count<-day_count[order(day_count$day),]
write.csv(day_count,"./day_count.csv",row.names = FALSE)
## mean
day_mean<-as.data.frame(aggregate(data_all_v1$riding_time,
                                  list(data_all_v1$member_casual,data_all_v1$weekday), F
UN=mean))
colnames(day_mean)<-c("user_type","day","mean_riding_time")
day_mean$day <- ordered(day_mean$day, levels=c("Monday", "Tuesday", "Wednesday", "Thursday", "F
riday", "Saturday", "Sunday"))
day_mean<-day_mean[order(day_mean$day),]
write.csv(day_mean,"./day_mean.csv",row.names = FALSE)
## visualization
ggplot(data=day_count,aes(x = day, y = counts, fill = user_type))+geom_col(width=0.5, position
 = position_dodge(width=0.5))+ theme(aspect.ratio = 1)+ggtitle("Number_of_rides_vs_days")
```



As shown in the Average_of_riding_time_vs_days figure, members has consistent riding time all over the week. Casual users have longer riding time on weekends. The average time of casual users is longer than that of members all over the week.

```
ggplot(data=day_mean,aes(x = day, y = mean_riding_time, fill = user_type))+geom_col(width=0.5,
 position = position_dodge(width=0.5))+theme(aspect.ratio = 1)+ggtitle("Average_of_riding_time_
vs_days")
```

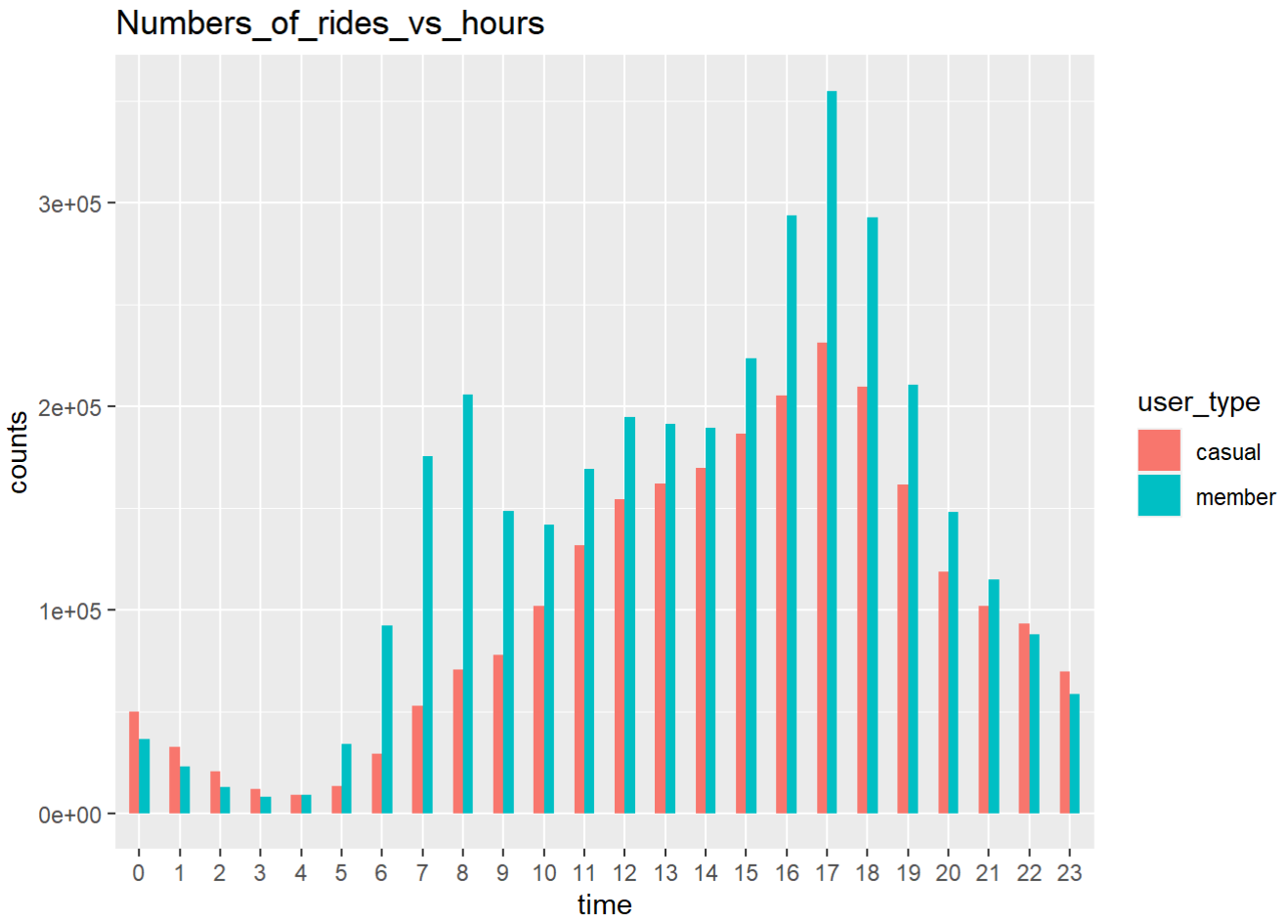


Conclusion of daily analysis: casual users have more numbers of rides and longer riding time on weekends.

Comparision between casual users and members for different hours

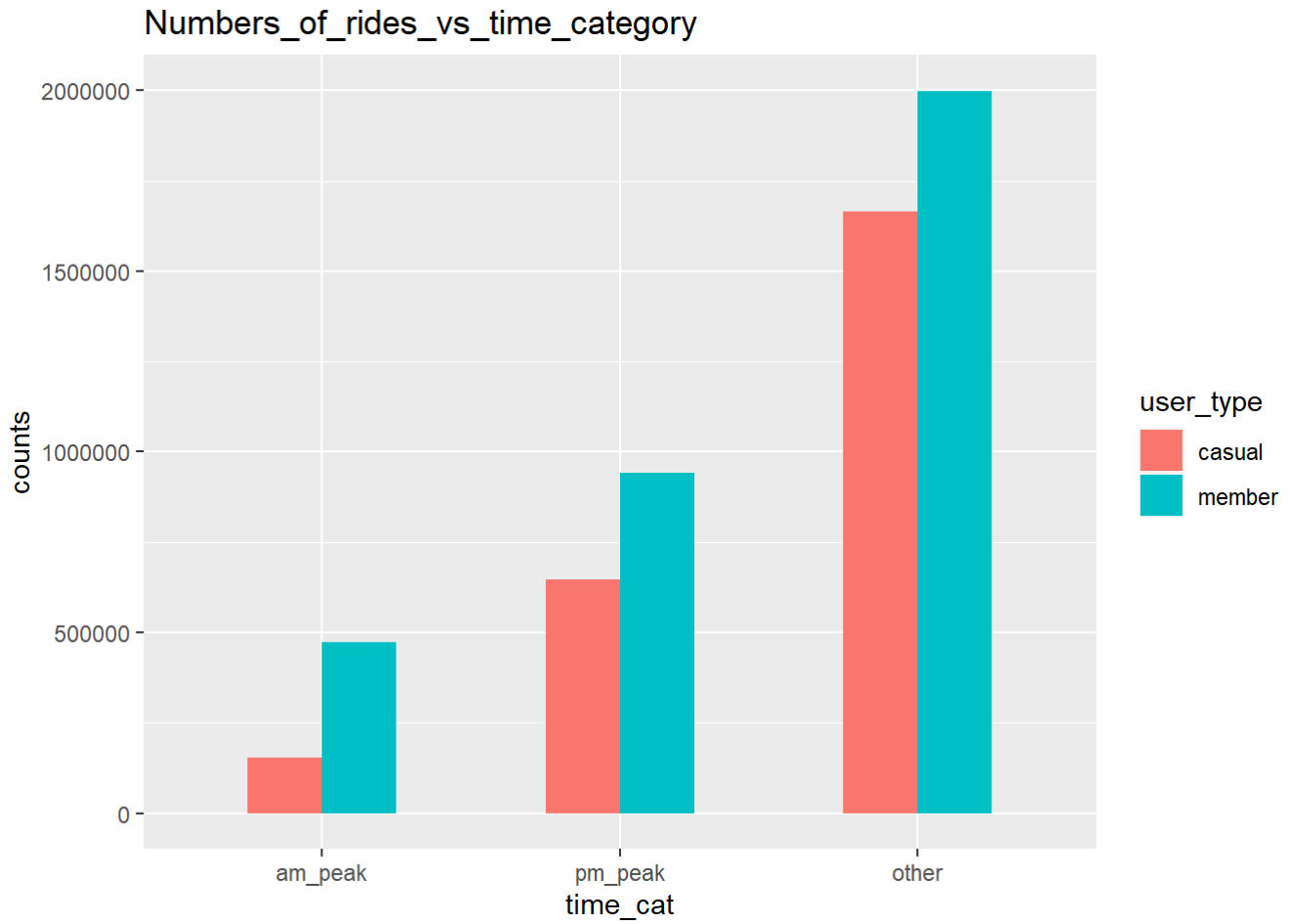
As shown in the Numbers_of_rides_vs_hours figure, members have morning peak and afternoon peak, while casual users only have the afternoon peak.

```
time_count<-as.data.frame(table(data_all_v1$member_casual,data_all_v1$time))
colnames(time_count)<-c("user_type","time","counts")
write.csv(time_count,"./time_count.csv",row.names = FALSE)
ggplot(data=time_count,aes(x = time, y = counts, fill = user_type))+geom_col(width=0.5, positio
n = position_dodge(width=0.5))+ggtitle("Numbers_of_rides_vs_hours")
```

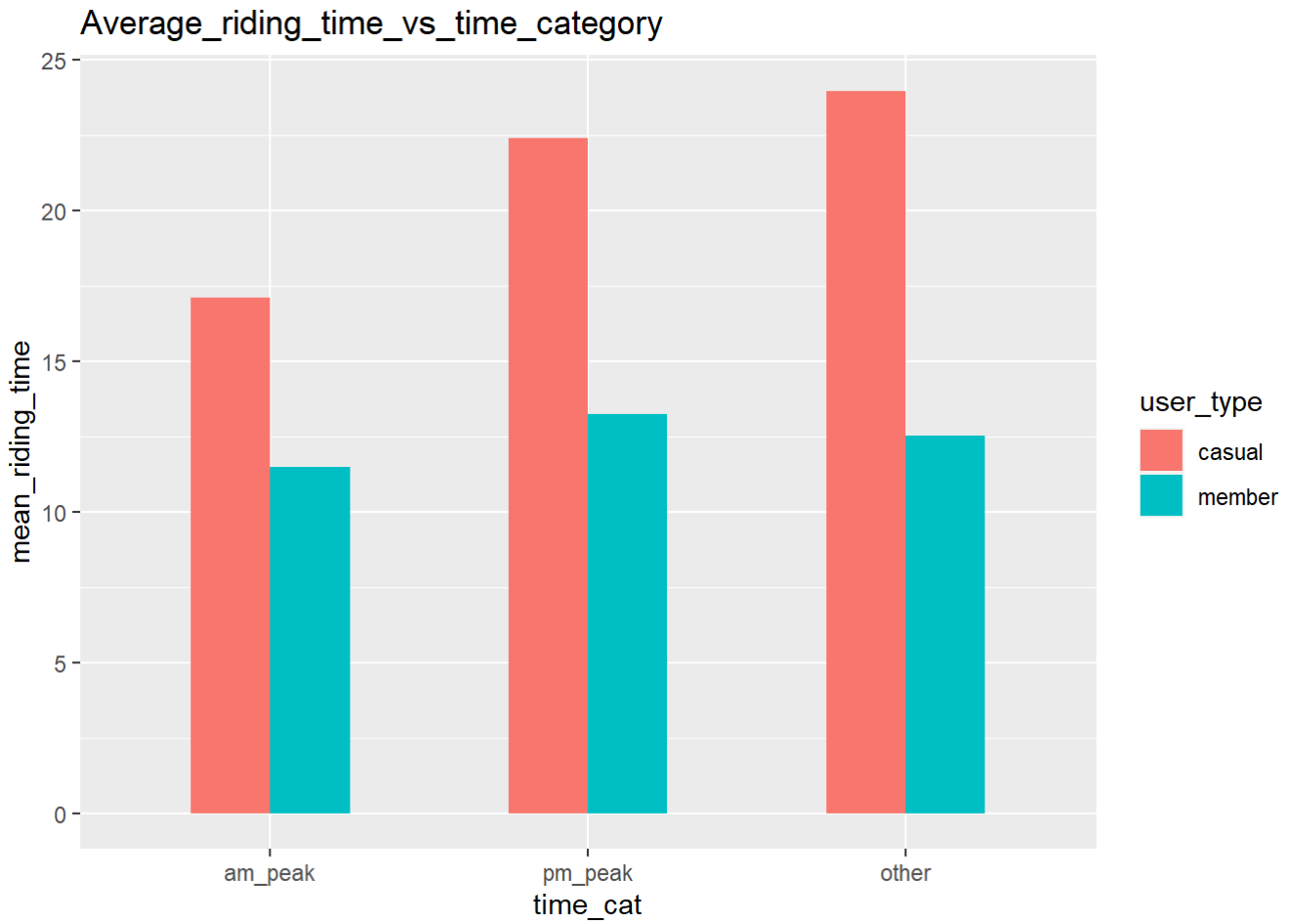


The Numbers_of_rides_vs_time_category figure shows that casual users have much less usage during the am peak. The Average_riding_time_vs_time_category figure shows that members have consistent short riding time all over the day. Casual users have longer riding time than members. And casual users take longer time of riding during pm peak.

```
## any pattern in peak hours and non-peak hours?
time_cat_count<-as.data.frame(table(data_all_v1$member_casual,data_all_v1$time_cat))
colnames(time_cat_count)<-c("user_type","time_cat","counts")
time_cat_count$time_cat <- ordered(time_cat_count$time_cat, levels=c("am_peak", "pm_peak", "oth
er"))
time_cat_count<-time_cat_count[order(time_cat_count$time_cat),]
write.csv(time_cat_count,"./time_cat_count.csv",row.names = FALSE)
## mean
time_cat_mean<-as.data.frame(aggregate(data_all_v1$riding_time,
                                       list(data_all_v1$member_casual,data_all_v1$time_cat), F
UN=mean))
colnames(time_cat_mean)<-c("user_type","time_cat","mean_riding_time")
time_cat_mean$time_cat <- ordered(time_cat_mean$time_cat, levels=c("am_peak", "pm_peak", "othe
r"))
time_cat_mean<-time_cat_mean[order(time_cat_mean$time_cat),]
write.csv(time_cat_mean,"./time_cat_mean.csv",row.names = FALSE)
## visualization
ggplot(data=time_cat_count,aes(x = time_cat, y = counts, fill = user_type))+geom_col(width=0.5,
position = position_dodge(width=0.5))+ggtitle("Numbers_of_rides_vs_time_category")
```




```
ggplot(data=time_cat_mean,aes(x = time_cat, y = mean_riding_time, fill = user_type))+geom_col(width=0.5, position = position_dodge(width=0.5))+ggtitle("Average_riding_time_vs_time_category")
```



At which station did casual users use bikes the most? The top 10 stations where casual users use bikes the most frequently are: streeter dr & grand ave
dusable lake shore dr & monroe st
millennium park
michigan ave & oak st
dusable lake shore dr & north blvd shedd aquarium
theater on the lake
wells st & concord ln
clark st & armitage ave
clark st & lincoln ave

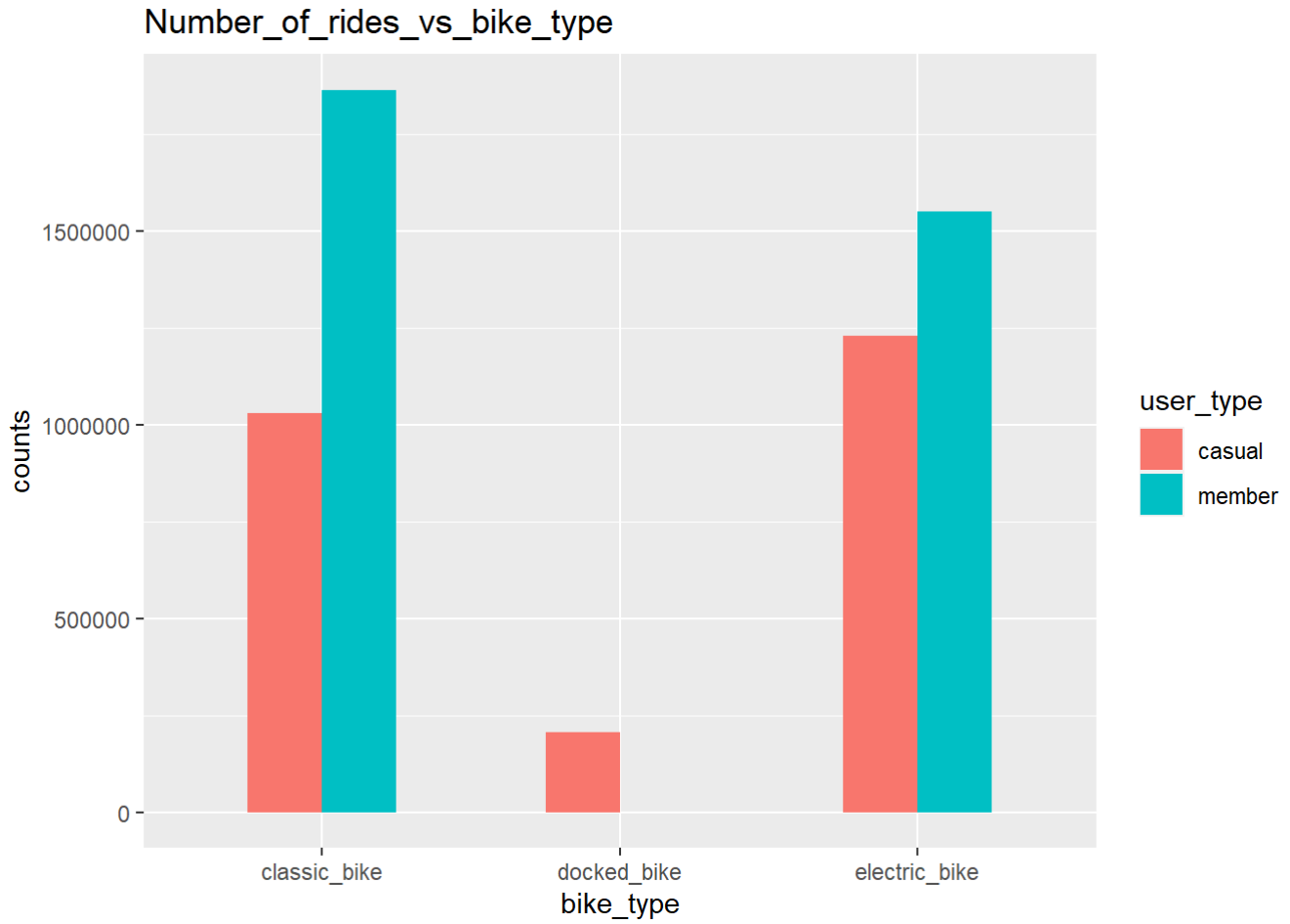
```
start_station<-as.data.frame(table(data_all_v1$member_casual,data_all_v1$start_station_name))
colnames(start_station)<-c("user_type", "start_station","counts")
start_station_casual<-start_station[which(start_station$user_type=="casual"),]
start_station_casual<-start_station_casual[order(start_station_casual$counts,decreasing=TRUE),]
start_station_casual[2:11,"start_station"] # show the most visited 10 stations
```

```
##  [1] streeter dr & grand ave           dusable lake shore dr & monroe st
##  [3] millennium park                   michigan ave & oak st
##  [5] dusable lake shore dr & north blvd shedd aquarium
##  [7] theater on the lake               wells st & concord ln
##  [9] clark st & armitage ave           clark st & lincoln ave
## 1439 Levels: 10101 s stony island ave ... zapata academy
```

How about the type of bike?

As shown in “Number_of_rides_vs_bike_type”, casual users prefer electric bikes while members like classic bikes.

```
bike_type<-as.data.frame(table(data_all_v1$member_casual,data_all_v1$rideable_type))
colnames(bike_type)<-c("user_type", "bike_type","counts")
write.csv(bike_type,"./bike_type.csv",row.names = FALSE)

ggplot(bike_type,aes(x = bike_type, y = counts, fill = user_type))+geom_col(width=0.5, position
= position_dodge(width=0.5))+ggtitle("Number_of_rides_vs_bike_type")
```



Recommendations

1. Given that casual users have more rides from June to September, advertisement of members' benefits should be launched in these months, and promotions should be placed after September to help casual users remain members.
2. Given that casual users have more rides on weekends, advertisement of members' benefits should be launched on weekends, and promotions should be placed in weekdays to help casual users remain members.
3. Given that casual users always prefer long rides than short rides, special promotions for rides longer than 20 minutes may be effective.

Additionally, as we obtained the top 10 most used start station and the preferred bike type, the advertisement and promotion should target such locations on the usage of electric and classic bikes.

Limitation: The reason why casual users prefer long rides and like to ride during weekends is unknown. This may be solved by collecting more personal information from the users.