Case Study 1: Cyclistic data analysis of user comparison

Sharon

Business Task Summary

In order to maximize the company's profit, increment of annual members is an effective means. A data analysis task has been assigned to figure out the differences between casual riders and annual members.

Data used

Dataset used in this project is Cyclistic's historical trip data from 1st July 2022 to 30th June 2023. included information below:

User information: user ID and the member types (casual riders or annual members)

Riding type: the bike used in the riding trip.

Time: the starting and ending time of the riding trip.

Location: the starting and ending location (station, latitude, longitude) of the riding trip.

Data Processing and Analysis

For large dataset, R Studio is more suitable for data processing, analysis and visualization. The data processing and analysis are carried out following steps below:

1. Import the data
   The downloaded data are saved in a folder named Project data, which is located in document folder in C drive. After checking the consistency of data, and confirming all columns have the same naming and type, I have merged the data. Also, I use glimpse() and str() function to check the data type.

```
path<-"~/Project data/divvy_trip_data_csv"
all_files<-list.files(path,pattern="*.csv",full.names=TRUE)
tripData <- data.frame()
for(file in all_files){
  data<-read.csv(file)
  tripData<-bind_rows(tripData,data)
}
```

```
glimpse(tripData)
## Rows: 5,779,444
## Columns: 13
## $ ride_id            <chr> "954144C2F67B1932", "292E027607D218B6", "5776585258…
## $ rideable_type      <chr> "classic_bike", "classic_bike", "classic_bike", "cl…
## $ started_at         <chr> "2022-07-05 08:12:47", "2022-07-26 12:53:38", "2022…
## $ ended_at           <chr> "2022-07-05 08:24:32", "2022-07-26 12:55:31", "2022…
## $ start_station_name <chr> "Ashland Ave & Blackhawk St", "Buckingham Fountain …
## $ start_station_id   <chr> "13224", "15541", "15541", "15541", "TA1307000117",…
## $ end_station_name   <chr> "Kingsbury St & Kinzie St", "Michigan Ave & 8th St"…
## $ end_station_id     <chr> "KA1503000043", "623", "623", "TA1307000164", "TA13…
## $ start_lat          <dbl> 41.90707, 41.86962, 41.86962, 41.86962, 41.89147, 4…
## $ start_lng          <dbl> -87.66725, -87.62398, -87.62398, -87.62398, -87.626…
## $ end_lat            <dbl> 41.88918, 41.87277, 41.87277, 41.79526, 41.93625, 4…
## $ end_lng            <dbl> -87.63851, -87.62398, -87.62398, -87.59647, -87.652…
## $ member_casual      <chr> "member", "casual", "casual", "casual", "member", "…
str(tripData)
## 'data.frame':    5779444 obs. of  13 variables:
##  $ ride_id           : chr  "954144C2F67B1932" "292E027607D218B6" "57765852588AD6E0" "B5B6BE4
4314590E6" ...
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "2022-07-05 08:12:47" "2022-07-26 12:53:38" "2022-07-03 13:58:49"
 "2022-07-31 17:44:21" ...
##  $ ended_at          : chr  "2022-07-05 08:24:32" "2022-07-26 12:55:31" "2022-07-03 14:06:32"
 "2022-07-31 18:42:50" ...
##  $ start_station_name: chr  "Ashland Ave & Blackhawk St" "Buckingham Fountain (Temp)" "Buckin
gham Fountain (Temp)" "Buckingham Fountain (Temp)" ...
##  $ start_station_id  : chr  "13224" "15541" "15541" "15541" ...
##  $ end_station_name  : chr  "Kingsbury St & Kinzie St" "Michigan Ave & 8th St" "Michigan Ave
& 8th St" "Woodlawn Ave & 55th St" ...
##  $ end_station_id    : chr  "KA1503000043" "623" "623" "TA1307000164" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.8 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr  "member" "casual" "casual" "casual" ...
```

2. Data processing

As we seen, the time information(started_at & ended_at) are in a wrong type. Thus I change the data type.

```
tripData$start_dt<-ymd_hms(tripData$started_at)
tripData$end_dt<-ymd_hms(tripData$ended_at)
```

Furthermore, I make the data calculation to conclude the information dataset given, which included duration, weekday and distance for each trip. Firstly, I can use the geosphere package to calculate the distance between the starting point and ending point by the latitude and longitude given.

```
library(geosphere)
## The legacy packages maptools, rgdal, and rgeos, underpinning the sp package,
## which was just loaded, will retire in October 2023.
## Please refer to R-spatial evolution reports for details, especially
## https://r-spatial.org/r/2023/05/15/evolution4.html.
## It may be desirable to make the sf package available;
## package maintainers should consider adding sf to Suggests:.
## The sp package is now running under evolution status 2
##        (status 2 uses the sf package in place of rgdal)
start_point<-cbind(lng=tripData$start_lng,lat=tripData$start_lat)
end_point<-cbind(lng=tripData$end_lng,lat=tripData$end_lat)
tripData$distance<-distHaversine(start_point,end_point,r=6378137)
```

Time duration and day of a week is calculated. After checking the data, I found some bugs which time duration is smaller than 0, which may caused by tracking error of APP. The whole dataset included 500+ error samples which is only occupied less than 0.01% of the data, filtered it out will not effect the data analysis result.

```
tripData$weekday<-weekdays(tripData$start_dt,abbreviate=TRUE)
tripData$duration<-as.numeric(tripData$end_dt - tripData$start_dt)
```

As the dataset is too large, I select the columns that are useful in the analysis process.

```
#Dispose the unnormal data
tripDataFilter<-filter(tripData,duration>0 & distance>0)
#Filter the columns
tripDataNeed<-select(tripDataFilter,c("ride_id","rideable_type","weekday","duration","distance",
"member_casual"))
```

Duplicates, Outliers and missing values also should be dropped before analysis. As I mentioned, there are no duplicates in the ride_id, which means that frequency cannot be analyzed. I use a box-plot to filter out the outliers, and filled them with N/A. Finally I deleted all N/A values to dropped all outliers and missing values.

```
#check duplicate
duplicated(tripDataNeed$ride_id)
#Replace outlier by NA
for (x in c('duration','distance')){
 value = tripDataNeed[,x][tripDataNeed[,x] %in% boxplot.stats(tripDataNeed[,x])$out]
 tripDataNeed[,x][tripDataNeed[,x] %in% value] = NA
}

#Check NA
as.data.frame(colSums(is.na(tripDataNeed)))
##                colSums(is.na(tripDataNeed))
## ride_id                                   0
## rideable_type                             0
## weekday                                   0
## month                                     0
## duration                             383935
## distance                             323520
## member_casual                             0
#Drop NA
tripDataNeed<-na.omit(tripDataNeed)
```

3. Data Analysis

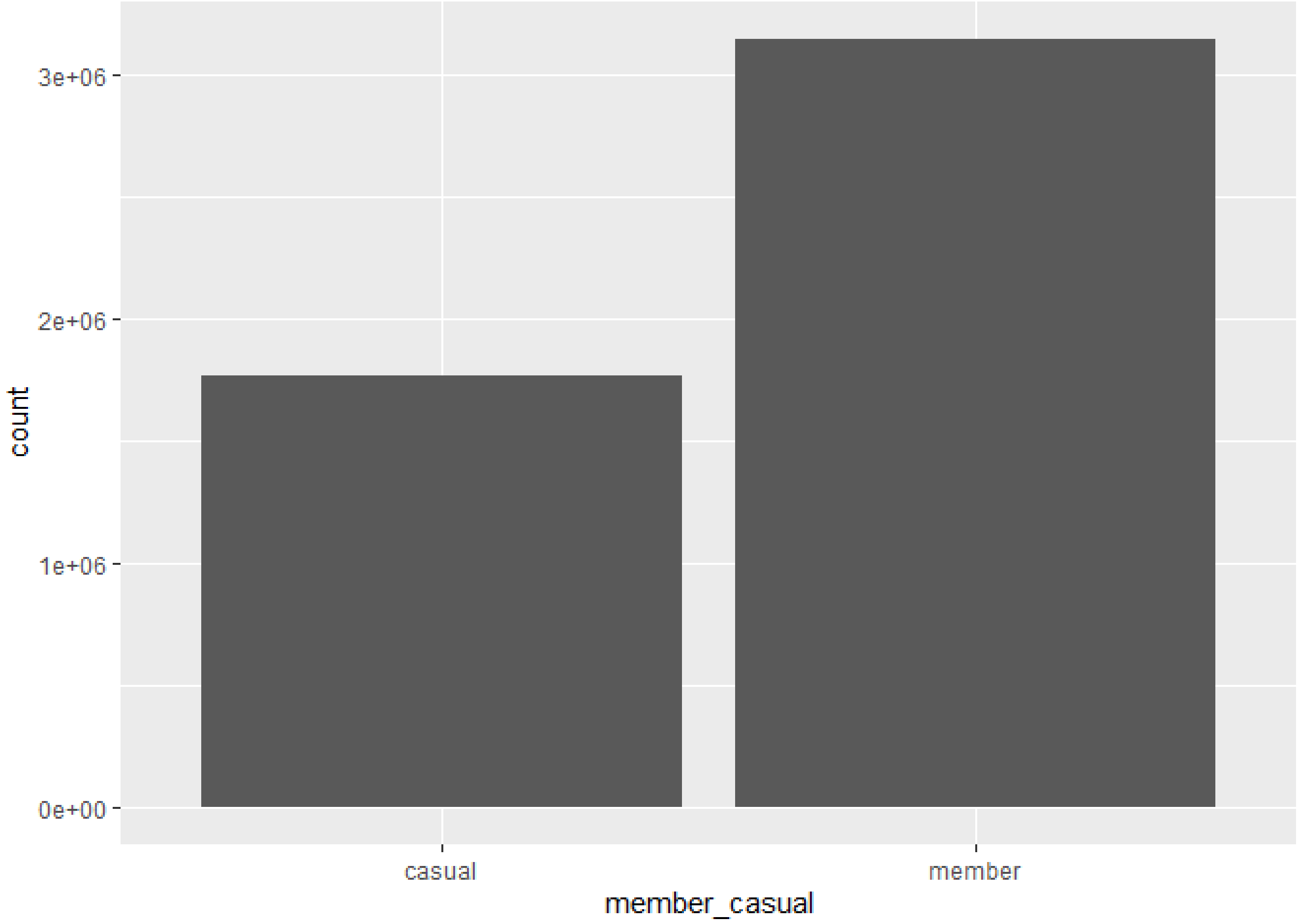
By grouping the riders’ type, I can check the differences.

```
tripDataNeed%>%
 group_by(member_casual)%>%
 summarize(n=n(),mean_time=mean(duration),min_time=min(duration),
 max_time=max(duration),mean_dis=mean(distance),
 min_dis=min(distance),max_dis=max(distance))
## # A tibble: 2 × 8
##   member_casual       n mean_time min_time max_time mean_dis min_dis max_dis
##   <chr>           <int>     <dbl>    <dbl>    <dbl>    <dbl>   <dbl>   <dbl>
## 1 casual        1763458      726.        1     2050    1913.  0.0186   5632.
## 2 member        3139011      576.        1     2050    1821.  0.0202   5632.
```

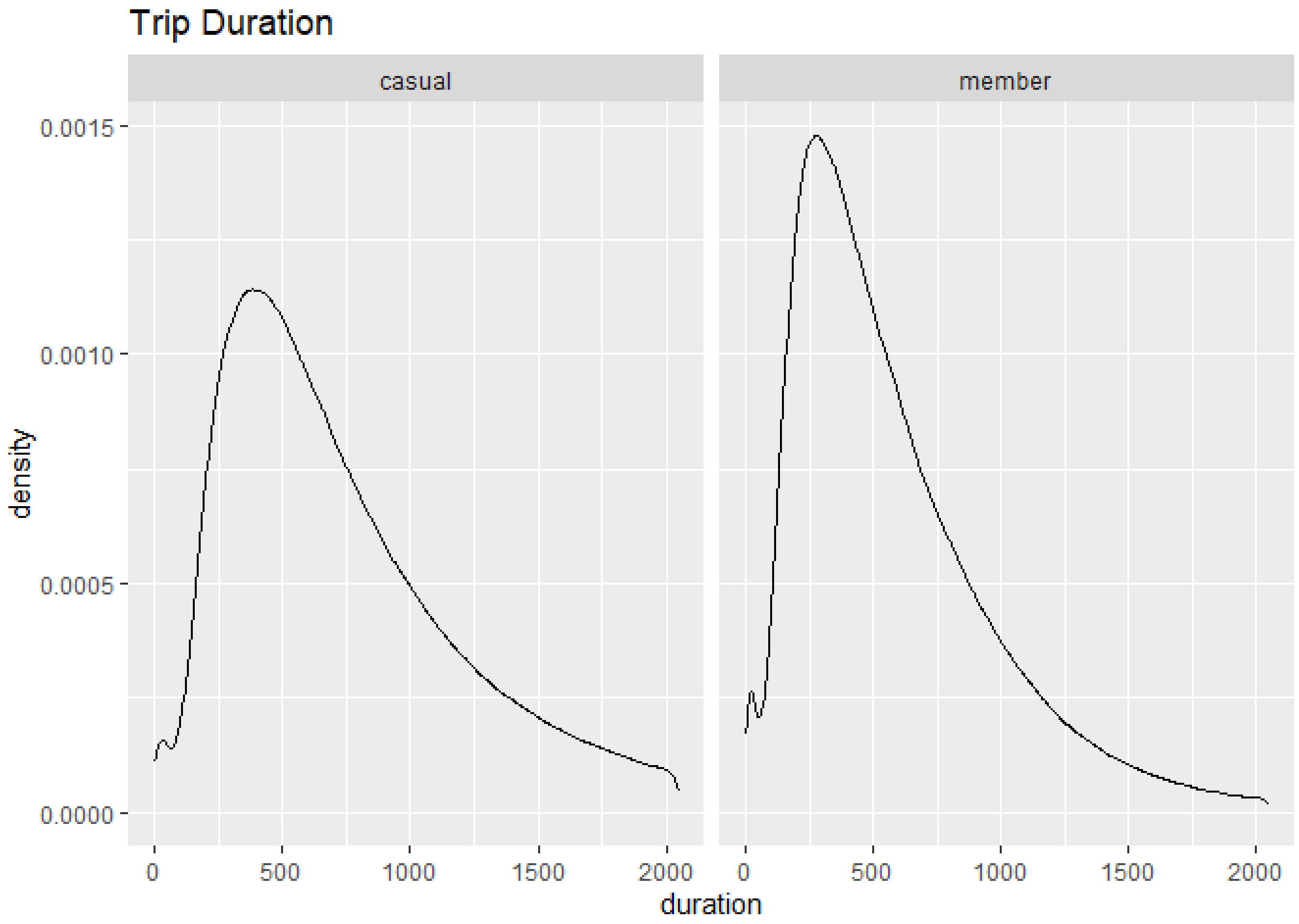
As the result above, members are 1.7 times more than casual riders, which is a good signal for company. The trip duration of casual riders are slightly more than the members（726sec vs 576sec）, and the trip distance of casual riders are slightly more than the members (1913m vs 1821m). It can be two reasons: **1) The casual riders are less professional in riding, which take more times to finish their trips, 2) The casual riders do take a longer single trip than the members.**

For easier understanding, data visualization is carried out.

```
#Visualization for numeric value
ggplot(tripDataNeed)+geom_bar(aes(member_casual))
```

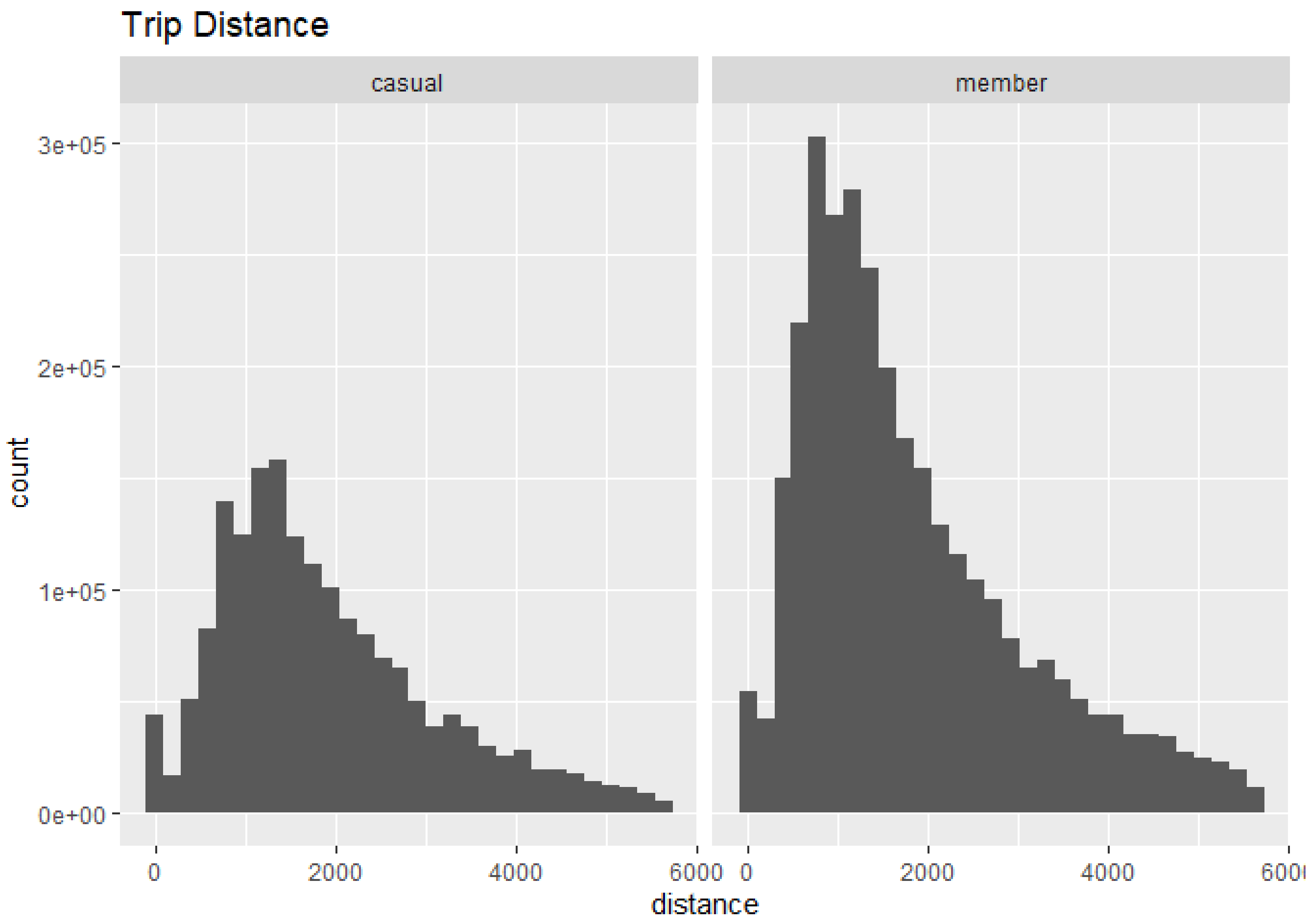


```
ggplot(tripDataNeed)+geom_density(aes(duration))+facet_wrap(facet=~member_casual)+labs(title='Tr
ip Duration')
```



```
ggplot(tripDataNeed)+geom_histogram(aes(distance))+facet_wrap(facet=~member_casual)+labs(title='
Trip Distance')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

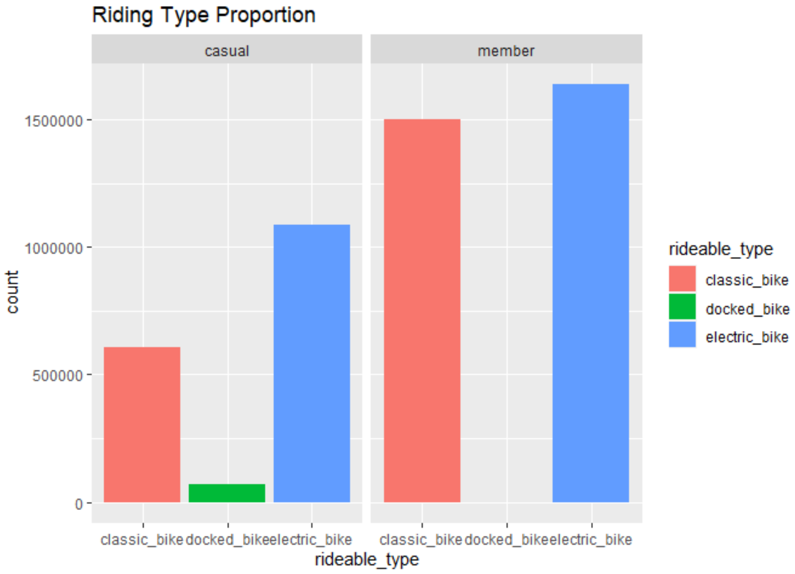


The duration and distance of both casual riders and members are showed in a skewed distribution, where showed an information that **users mainly use the bikes in a short trip.**

For categorical data, data visualization is also an effective way to figure out the differences.

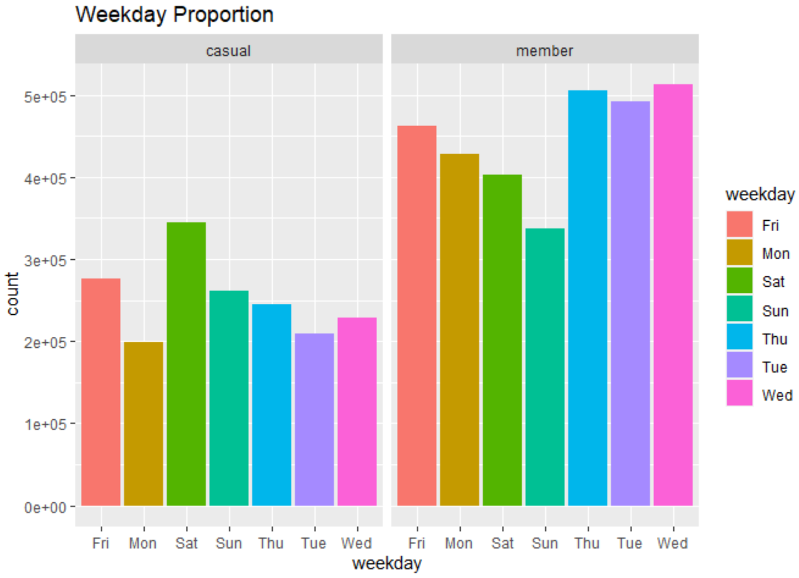
```
#Visualization for categorical value

ggplot(tripDataNeed)+geom_bar(aes(rideable_type,fill=rideable_type))+facet_wrap(facet=~member_ca
sual)+labs(title='Riding Type Proportion')
```



From the comparison of rideable type of the trip, the electric bike is more popular than others, especially in casual riders group. As we seen from the picture above, **the choice of classic bike and electric bike showed more significant difference in the casual riders group.** Also, only casual riders use the docked bike.

```
ggplot(tripDataNeed)+geom_bar(aes(weekday,fill=weekday))+facet_wrap(facet=~member_casual)+labs(t
itle='Weekday Proportion')
```



The result of weekday showed a valuable information. Casual riders mostly start a trip on Friday, Saturday and Sunday while the members mostly start a trip on Tuesday, Wednesday and Thursday. **The differences showed that casual riders mainly use Cyclistic in weekend for leisure. The members use Cyclistic in weekday for daily usage purpose.**

Conclusion

1. Different in purpose

The main difference between casual riders and members is the purpose of using Cyclistic. As the result showed from the weekday comparison drawn a conclusion that when people have a daily need in using Cyclistic, they have more probability to become a member.

From the given information, only 30% users use Cyclistic to commute to work each day, thus it is a huge potential market space. I recommend to put into more marketing resources in the daily work scene such as the elevator in an office building and Subway station.

2. Different in rideable type

The riding type also showed a difference. Firstly, the docked bike is not popular in casual riders and not useful in members. I recommend to dispose the docked bike line to transfer the resources to other lines. Secondly, the majority of riders opt for electric bike, it may be a trend nowadays. I recommend to put more attention in the electric bike, but also without ignoring the classic bike.

3. Different in duration and distance

The slight differences of the duration and distance pointed out the casual riders will take a longer trip than members, where consistent to the conclusion above. Casual riders mainly take a leisure trip while members mainly use for the daily purpose. I recommend to provide some packages for the longer trip to attract casual riders become a member.