Analysis Report for Case1: How Does a Bike-Share Navigate Speedy Success

Bunthit W

2022-07-20

Introduction

The case is about the business challenge of the Cyclistic, a fictional bike-share company assuming to run a business in Chicago. The marketing director requires the higher business success. For the market comprising two clusters of customers: casual riders and annual members, the goal for this analysis is to underlie an initiated marketing strategy to convert casual riders into annual member.

To achieve this, the process will be according to the 6 phases of data analysis: Ask, Prepare, Process, Analyze, Share, and Act (APPASA). This report will cover the first five.

By the end of this analysis, the following 3 issues should be responded:

- How do annual members and casual riders use Cyclistic bikes differently?
- Why would casual riders buy Cyclistic annual memberships?
- How can Cyclistic use digial media to influence casual riders to become members?

1.Information (Ask)

Cyclistics' business model is a bike-share program, more than 5,800 bicycles and 600 docking stations with a unique offerings, e.g. electric bikes and special bikes for disabiities. Users are usually riding for leisure and some (about 30%) use them to commute to work daily.

Stakeholders

- **Marketing director** Responsible for developing campaigns and initiatives to promote the bike-share program thru social media and other channels
- **Cyclistic marketing analytics team** who collect, analyze and report data to guide the marketing strategies.
- **Cyclistic executive team** who are notoriously detail-oriented and will be persons who decide whether to approve the recommendedd marketing program.

2. Data Preparation

- Data requisition and organization The data are collected and provided online in monthly basis as a Zip file of ".CSV". The volume is quite big, e.g. about 600,000 records (of 13 columns) a month. Checking on some archives and found that there were changes in formats when compared with the current ones. However, the format of the lastest 2 years or so, data are quite consistent though the completion could be an issue.

For the analysis, we will download 12 months (June 2021 - May 2022) from this which is provided by Movivate Internatioanal Inc, under this license

We have organized the data from 12 separated files into a sinle dataframe and perform some verification to ensure the readiness for analysis.

```
data_all <- list.files(path="/Users/Shared/working_dir", full.names = TRUE) %>%
  lapply(read_csv) %>%
  bind_rows
```

```
glimpse(data_all)
```

```
## Rows: 5,860,776
## Columns: 13
## $ ride_id            <chr> "99FEC93BA843FB20", "06048DCFC8520CAF", "9598066F68…
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", …
## $ started_at         <dttm> 2021-06-13 14:31:28, 2021-06-04 11:18:02, 2021-06-…
## $ ended_at           <dttm> 2021-06-13 14:34:11, 2021-06-04 11:24:19, 2021-06-…
## $ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "Michigan Ave &…
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "13042", NA, NA…
## $ start_lat          <dbl> 41.80, 41.79, 41.80, 41.78, 41.80, 41.78, 41.79, 41…
## $ start_lng          <dbl> -87.59, -87.59, -87.60, -87.58, -87.59, -87.58, -87…
## $ end_lat            <dbl> 41.80000, 41.80000, 41.79000, 41.80000, 41.79000, 4…
## $ end_lng            <dbl> -87.6000, -87.6000, -87.5900, -87.6000, -87.5900, -…
## $ member_casual      <chr> "member", "member", "member", "member", "member", "…
```

Refer to the glimpse data, noted that 1. For the 5.8 millions records, there are missing data on station name and id (for some months, only). We decide to ignore location data (including the _lat. and _lng. for geoloations). 2. Ride_id is a hash key and there are no link to customer data, so we decide to ignore it for this stage. 3. The format of data are quite all rights: chr/text, numeric/dbl, and the timestamp/dttm. However, for the analysis, we need to transform/extract them into more appropriate values as will be discussed in the next section.

So, in terms of data quality, according to the ROCCC (reliable, original, comprehensive, current or cited), it is quite satisfactory for analysis.

3. Process

Based on the acquired data, after we organized them into a single dataframe, we decided to transform:

- distinction between started_at and ended_at into absolute durations for each trip,
- add new attributes of day_of_week for insights of demand in each day of the week.
- extract the month-year for seeing trend of comparative riding patterns, esp. between members and casual.

```
data_all_tmp = mutate(data_all, trip_duration = round(as.double(difftime(ended_at, started_at, units="min
s")), digits=2)) %>%
  mutate(data_all, day_of_week = weekdays(started_at, abbreviate = TRUE)) %>%
  mutate(data_all, month_yr = format(as.Date(started_at), "%Y-%m"))

data_all_0 = data_all_tmp %>% select(rideable_type, member_casual, trip_duration, day_of_week, month_yr)

glimpse(data_all_0)
```

```
## Rows: 5,860,776
## Columns: 5
## $ rideable_type <chr> "electric_bike", "electric_bike", "electric_bike", "elec…
## $ member_casual <chr> "member", "member", "member", "member", "member", "membe…
## $ trip_duration <dbl> 2.72, 6.28, 5.98, 25.83, 4.13, 6.75, 6.18, 6.30, 8.77, 9…
## $ day_of_week   <chr> "Sun", "Fri", "Fri", "Thu", "Fri", "Thu", "Thu", "Thu", …
## $ month_yr      <chr> "2021-06", "2021-06", "2021-06", "2021-06", "2021-06", "…
```

After this stage, by the output of glimpse, we can gain quite a neat and compact data for analysis. We actually saved some space and acquired quite an appropriate format for data for further analysis, too. We, finally, check the completeness of data before moving to the analysis.

```
sapply(data_all_0, function(x) sum(is.na(x)))
```

```
## rideable_type member_casual trip_duration   day_of_week      month_yr
##             0             0             0             0             0
```

Next, we validate of trip_duration by using a simple min max calculation. The result is as follows:

```
stat0 <- data_all_0 %>%
  summarize(avg_ride = mean(trip_duration),
            min_ride = min(trip_duration),
            max_ride = max(trip_duration))
stat0
```

```
## # A tibble: 1 × 3
##   avg_ride min_ride max_ride
##      <dbl>    <dbl>    <dbl>
## 1     20.7    -58.0   55944.
```

We can observe the unusable of negative ride duration. We will delete these items for sure. We check further about those ride with unreasonably short duration, e.g. 1 - 3 mins or very long duration like over a day and found that

```
print(paste('trips shorter than or equal to 3  minutes = ',
            nrow(subset(data_all_0,trip_duration <= 3))))
```

```
## [1] "trips shorter than or equal to 3  minutes =  378400"
```

```
print(paste('trips lonager than 1 days = ',
            nrow(subset(data_all_0,trip_duration >= 24*60))))
```

```
## [1] "trips lonager than 1 days =  4406"
```

However, since we have no enough strong reason to get rid of these data, we decide to delete only those records with negative duration.

```
data_all_1 = data_all_0[!(data_all_0$trip_duration < 0),]
glimpse(data_all_1)
```

```
## Rows: 5,860,637
## Columns: 5
## $ rideable_type <chr> "electric_bike", "electric_bike", "electric_bike", "elec…
## $ member_casual <chr> "member", "member", "member", "member", "member", "membe…
## $ trip_duration <dbl> 2.72, 6.28, 5.98, 25.83, 4.13, 6.75, 6.18, 6.30, 8.77, 9…
## $ day_of_week   <chr> "Sun", "Fri", "Fri", "Thu", "Fri", "Thu", "Thu", "Thu", …
## $ month_yr      <chr> "2021-06", "2021-06", "2021-06", "2021-06", "2021-06", "…
```

After deleting 139 trips with duration less than zero, we had quite satisfactory level of data reliability, we move to the next phrase of analysis. Since, the data is quite big, we decide to use the R program with help of R-Studio for exercising this case. The process of data preparation seems to be smoothen and speedy, so far.

4.Analyze the data

4.1 Findings summary for executive

In the scope of 12 months (shifted full year of June 2021 - May 2022), we have found some interesting points as follows:

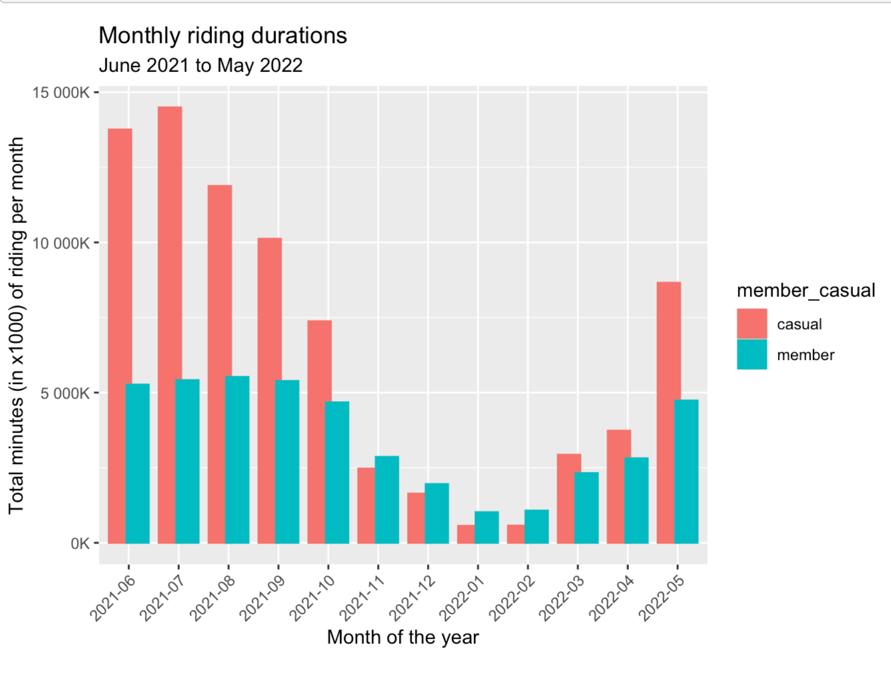
- Considering monthly data, there are patterns of riding where demand is quite low in the year-end or -early (winter) and go huge in the mid-year (summer time). This is true for both of member and casual users. – This can confirm that annual membership will help to smoothen the income.
- If considering the full year, we found that there are interesting patterns of riding among casual and members. Members have larger in number of trips and total distance, but lesser in terms of total hours spent.– this can help in campaign on social
- If looking a week scope of data,we found that most casual riders prefer spending time in weekends, while the members spent riding time quite flatly through the week. The leisure program can be a candidate for promotion to new types of annual members.

4.2 Monthly pattern of riding

This can give a full picture for the demand pattern of riding month-by-month and also the higher level,e.g. peak and trough of usage.

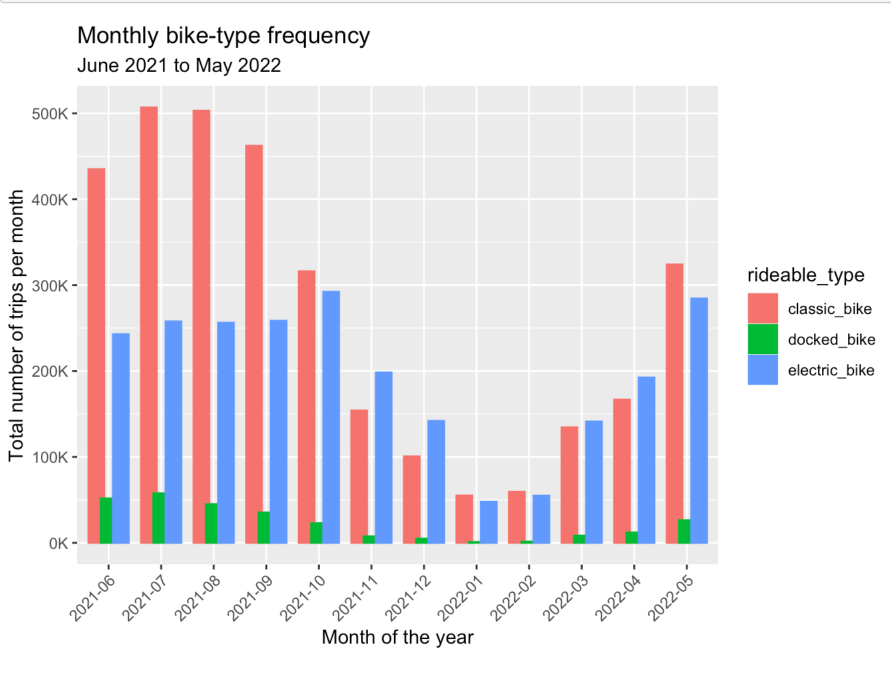
```
stat1 = data_all_1 %>%
  group_by(month_yr, member_casual) %>%
  summarize(ride_monthly =   sum(trip_duration/1000))

ggplot(data=stat1, aes(x=factor(month_yr), y=ride_monthly, fill = member_casual, colour=member_casual))+
  geom_bar(stat="identity", position = position_dodge(width=0.7))+
  theme(axis.text.x = element_text(angle = 45, hjust=1))+
  labs(
    title = "Monthly riding durations",
    subtitle = "June 2021 to May 2022",
    x = "Month of the year",
    y = "Total minutes (in x1000) of riding per month") +
    scale_y_continuous(labels = label_number(suffix = "K"))
```



The above chart shows the pattern of riding (in terms of total duration) which give insight into the fact that usually casual users spent more time of riding, except during the winter time. If we looks further in terms of rideable types that each class of users use mostly in each month, then we get data in the following chart.

```
stat3_1 <- data_all_1 %>%
  group_by(month_yr, rideable_type) %>%
  summarise(no_of_use = n())
ggplot(data=stat3_1, aes(x=factor(month_yr), y=no_of_use, fill = rideable_type, colour=rideable_type))+
  geom_bar(stat="identity", position = position_dodge(width=0.7))+
  theme(axis.text.x = element_text(angle = 45, hjust=1))+
  labs(
    title = "Monthly bike-type frequency",
    subtitle = "June 2021 to May 2022",
    x = "Month of the year",
    y = "Total number of trips per month") +
    scale_y_continuous(labels = label_number(suffix = "K", scale = 1e-3))
```



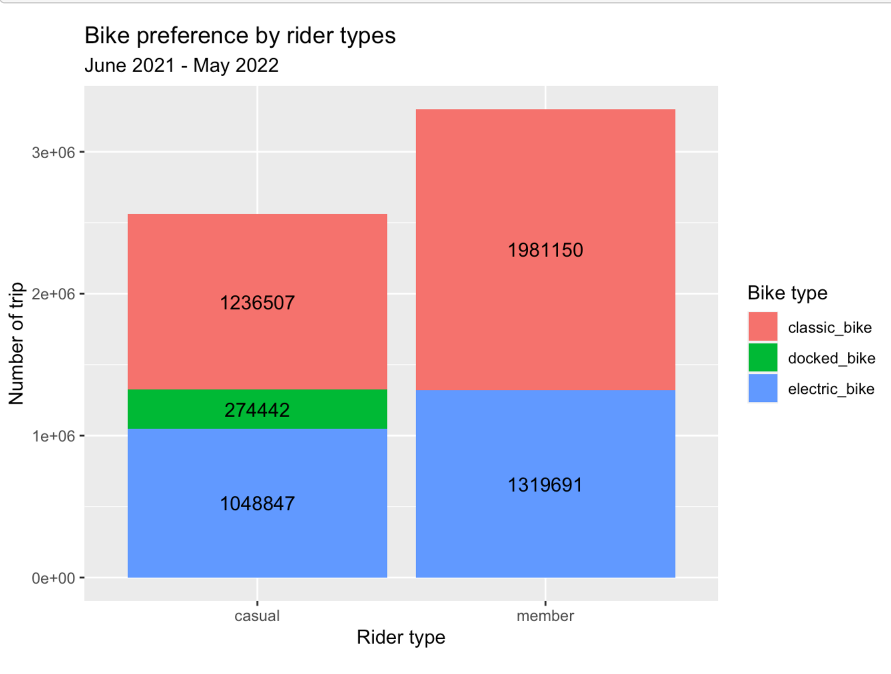
We can observe the trend of electric bike to be more and more popular when compared wo the other types. Well, the next question is that is this true for both of member and casual or not.

4.3 Behaviral patterns of riding in overall

Here, we will analyse the whole data of all 12 months to have holistic view of the market.

The next bar chart show the comparison in terms of number of trips for each type of bike for the whole year.

```
stat3 <- data_all_1 %>%
  group_by(member_casual, rideable_type) %>%
  summarise(no_of_use = n())
ggplot(stat3, aes(x=member_casual, y= no_of_use, fill=rideable_type)) +
  geom_bar(stat="identity") +
  labs(
    title = "Bike preference by rider types",
    subtitle = "June 2021 - May 2022",
    fill = "Bike type",
    x = "Rider type",
    y = "Number of trip") +
  geom_text(aes(label=no_of_use), position = position_stack(vjust = .5), color="black")
```



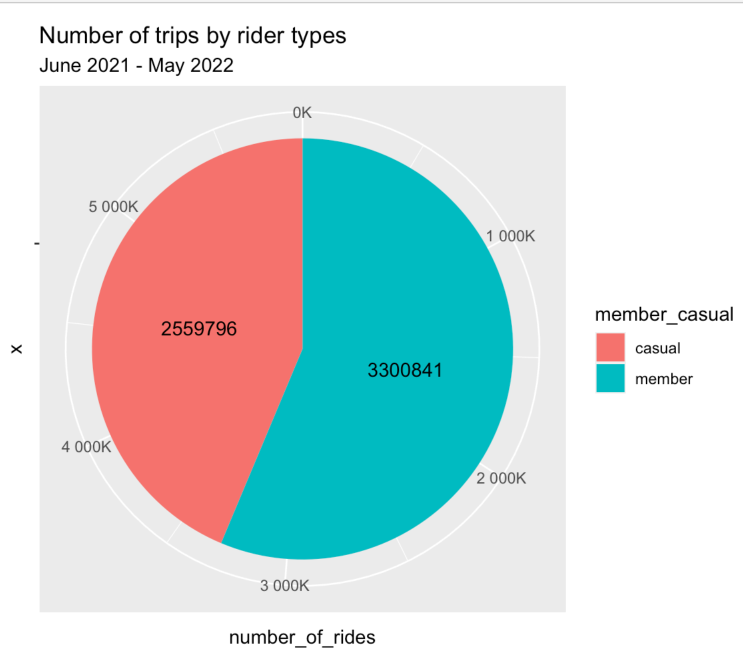
This give information that the preference of electric bike is quite a important in either class of user. Considering the given information about the trend of electric bike population for long hours usage, marketing team may be beneficial on this piece of data for a new marketing plot. Note that member users have no requirement of docked bikes. We move to the analysis of overall picture for gain further insight.

```
stat3_2 <- data_all_1 %>%
  group_by(member_casual, rideable_type) %>%
  summarize(avg_ride = mean(trip_duration),
            min_ride = min(trip_duration),
            max_ride = max(trip_duration))
```

```
## # A tibble: 5 × 5
## # Groups:   member_casual [2]
##   member_casual rideable_type avg_ride min_ride max_ride
##   <chr>         <chr>            <dbl>    <dbl>    <dbl>
## 1 casual        classic_bike      28.7        0    1560.
## 2 casual        docked_bike       83.1        0   55944.
## 3 casual        electric_bike     19.0        0     487.
## 4 member        classic_bike      13.7        0    1560.
## 5 member        electric_bike     12.0        0     481.
```

From the table, we can see that casual riders spent longer time per trip than the member riders in every type of bike. Note that the information indicate significant time per use for the docked bike. Last but not least, we can have a number of rides per rider types which indicate that the Cyclistic has a larger number of member than the casual (56.32% : 43.67%) which is conformed to the direction of marketing.

```
stat3_3 = data_all_1 %>%
  group_by(member_casual) %>%
  summarize(number_of_rides = n())
ggplot(stat3_3, aes(x="", y=number_of_rides, fill=member_casual))+
  geom_col()+
  coord_polar(theta="y")+
  scale_y_continuous(labels = label_number(suffix = "K", scale = 1e-3)) +
  geom_text(aes(label = number_of_rides),
          position = position_stack(vjust = 0.5)) +
  labs(
    title = "Number of trips by rider types",
    subtitle = "June 2021 - May 2022")
```



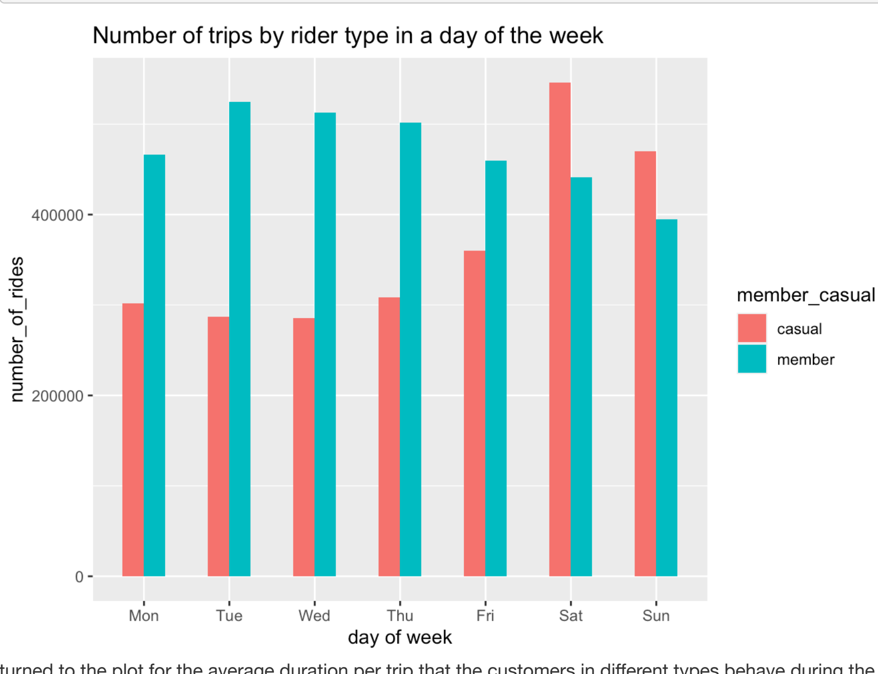
4.4 Behaviral patterns of riding in days of week

The last analysis will be dedicated to the data segmented into days of week which could give some insights about behaviors of users during weekdays and weekends.

We start by having a look on number of trips per each rider class for each day of the week. The chart below showing that the casual riders are peak at the weekend and surpass the number of rides by member users who normally greater during weekdays. However, number of member users riding in weekend is not much different than those in the weekdays. This information gives notice on the potential for setting annual programs for those casual who enjoy riding weekend.

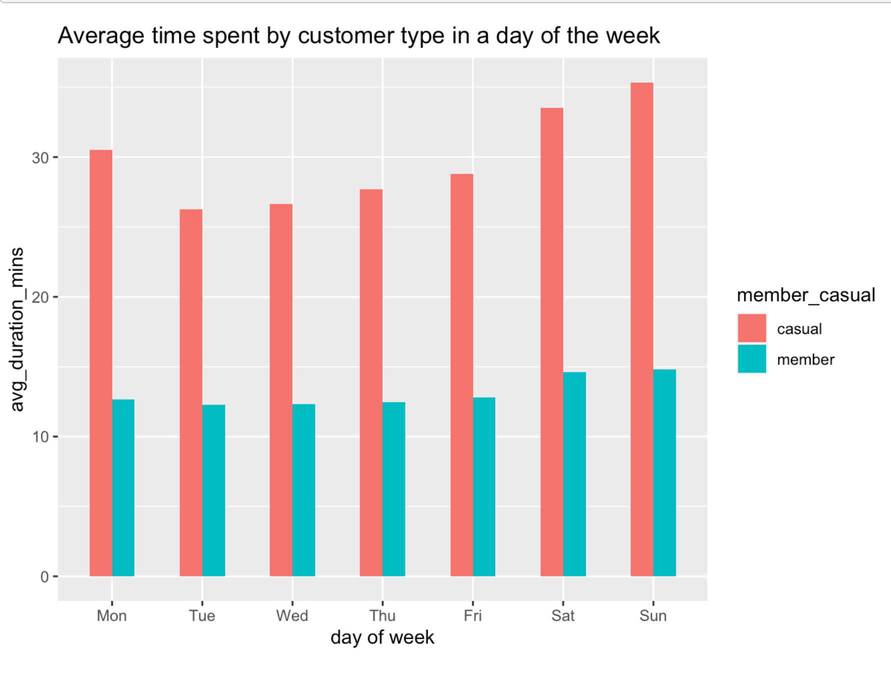
```
stat4 = data_all_1 %>%
  group_by(member_casual, day_of_week) %>%
  summarize(number_of_rides = n(), avg_duration_mins = mean(trip_duration))

stat4 %>% ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  labs(title ="Number of trips by rider type in a day of the week") +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  scale_x_discrete(name ="day of week",
                   limits=c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"))
```



For another dimension of data, we turned to the plot for the average duration per trip that the customers in different types behave during the day of the week. Interestingly, we found that most member users spent a bit more than 10 minutes per ride on average no matter what day of the week. The casual riders spent more than 30 minutes on average and spend a bit longer time per ride on the weekends.

```
stat4 %>% ggplot(aes(x = day_of_week, y = avg_duration_mins, fill = member_casual)) +
  labs(title ="Average time spent by customer type in a day of the week") +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  scale_x_discrete(name ="day of week",
                   limits=c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"))
```



5. Share and Recommendation

Based on the shared information we have perform through the process of analysis: ask, prepare, process, and analysis, we have got insights which were shared as executive summary in 4. and I repeat below:

Insight (How do annual members and casual riders use Cyclistic bikes differently?)

- Considering monthly data, there are patterns of riding where demand (total hours riding) is quite low in the year-end or -early (winter) and go huge in the mid-year (summer time). This is true for both of member and casual users. – This can confirm that annual membership will help to smoothen the income.
- If considering the full year, we found that there are interesting patterns of riding among casual and members. Members have larger in number of trips and total distance, but lesser in terms of total hours spent.– this can help in campaign on social
- If looking a week scope of data,we found that most casual riders prefer spending time in weekends, while the members spent riding time quite flatly through the week. The leisure program can be a candidate for promotion to new types of annual members.

We put additional findings here which may give some hints for recommendation in the discussion, next.

- For the behavior in day of the week, we found that casual riders spent much longer time per trip than those member. The proportion might be about 3:1.
- Electric bike are more and more popular choices for overall and also true for both groups of customers. Classic bikes are still the most popular in overall.

Recommendation

- Since the casual riders enjoy his time on weekend, the marketing campaign could be relevant to the weekend, e.g. organize marketing events and give special discount on the next year for member who rides up to a certain hours/distances/number of trips during weekend.
- On social, the activitity with electric bike that reinforces the pleasure of riding, e.g. long hours riding without tired, providing community to share the pleasure of riding and tips for enjoying riding, appointments for riding together.
- Dynamic price program by reducing price in winter and increasing when demand is high during summmer to convince those casual to save costs by entering into our membership with special discount package.
- Create a new membership type that suit for the behavior of casual, e.g. weekend riders with a lower costs than normal member packed with a bulk of discount coupons for riding in weekdays.

Further direction

Since we had utilized only a few attributes for analysis, Cyclistic could make further steps in analysis and marketing program in so many possible ways, e.g.:

- Analyze the behavior of users in utilizing stations, e.g. the top 10 most congested, so that Cyclistic can make a plan for improving services and customer satisfaction
- Using geolocation to study the track and trace the route so that Cyclistic can plan for new stations on high traffic routes, and invent some activities/campaign on those low traffic areas with business potency to lead sales to new customers.
- Cyclistic had ever collected some demographic data that might be so huge value for Social media activities and marking campaign, e.g. analyse and get influencers on social, clusters the users for marketing and services, e.g. matching customers and stations for designing and organizing the facilities, special events for big group of members with similar styles.