Cyclistic Trip Data

James Hung

2022-07-17

Cyclistic Trip

Business Task

Analyze the Cyclistic trip data from the previous 12 months to identify the differentiation of Cyclistic bikes usage habit between Casual Riders and Cyclistic Members. This allows the business is understand how Casual Riders and Cyclistic Members are differ, what causes the different, and how to close this usage gap between Casual Riders and Cyclistic Members. Therefore, it helps the business's marketing department to design an effective market strategy to promote Cyclistic Membership and convert Casual Riders into annual members.

Data Sources Description

The data source is the previous 12 months Cyclistic trip data. It is first-party data as it is collected internally based on the actual Cyclistic trip transactions data. The data source is open-source per the Data License Agreement, click here Link. It is stored on AWS cloud platform in zip files separated by month (i.e., monthly files). The data is in structured data format, in the form of tabular data, in table rows and columns.

The data is considered reliable as it is first-party data generated and gathered directly and internally by the Cyclistic. Additionally, it is considered original as it is directly representing each bike trip transaction taken by Cyclistic users and recorded by Cyclistic. Furthermore, it is considered comprehensive as it included data, trip start and end datetime information which can be used to determine the usage habit of the riders. Moreover, it is relatively current as the latest data offered is in the current year (i.e., 2022). Lastly, it is cited as the data hasn't been cited and vetted by Cyclistic internally. Therefore, the data source does ROCCC.

Per Cyclistic, each trip transaction is anonymized. All personal identification information (PII), i.e., Rider ID, has been encrypted. Therefore, privacy and security are addressed. In terms of data integrity, after initial review via sorting and filtering the data, there are some data points are missing and have errors, for example: Trip start/end stations are missing, and trip end datetime is before trip start datetime.

The data included columns: Trip start day and time, Trip end day and time, Trip start station, Trip end station, Rider type (Member, Casual), which can provide insights on usage habit of each trip for each rider type.

Data Process - Data Import

Previous 12 months Cyclistic trip data is downloaded from https://divvy-tripdata.s3.amazonaws.com/index.html. The month ride data are imported month by month using R language in R studio into individual month data frames: colnames(bike_trip_202106) ~ colnames(bike_trip_202205)

After imported all the monthly files, we need to inspect the the structure of the data frame, i.e., number of columns, to ensure data integrity: colnames(bike_trip_202106) ~ colnames(bike_trip_202205). We need to make sure no missing or additional column and all columns are aligned.

With all the all the monthly data imported into their individual data frames, it is to be merged into a single data frame with 12 months of data. This will be the main data frame to be used for data cleaning, data manipulation, and data analysis: bike_trip_12mo <- rbind(bike_trip_202106 ~ bike_trip_202205)

After all the monthly files merged, we can remove all the individual month data frame to save system memory resources: remove(bike_trip_202106 ~ bike_trip_202205)

We can check consolidated data frame structure to ensure all data are aligned:

- colnames(bike_trip_12mo) # List of cols
- nrow(bike_trip_12mo) # Num of rows
- dim(bike_trip_12mo) # dimensions of the df
- head(bike_trip_12mo) # Check 1st 6 rows
- str(bike_trip_12mo) # See list of columns and data types (numeric, character, etc)
- summary(bike_trip_12mo) # Statistical summary of data. Mainly for numeric

Data Process - Data Cleaning and Manipulation

First, we need to make sure the values in column **member_casual** and **rideable_type** are consistent:

```
unique(bike_trip_12mo$member_casual)
```

```
## [1] "member" "casual"
```

```
unique(bike_trip_12mo$rideable_type)
```

```
## [1] "electric_bike" "classic_bike"  "docked_bike"
```

Second, we need to calculate the length/duration of each bike ride and record them in a new column **ride_length**:

```
bike_trip_12mo$ride_length <- difftime(bike_trip_12mo$ended_at, bike_trip_12mo$started_at)
```

Third, in some cases the ride **end_at** is less than **started_at** which would generate a ***negative ride length***. Therefore we need to filter out all data that has ride_length greater than 0 into a new cleaned data frame and remove the raw data frame:

```
bike_trip_12mo_cleaned <- bike_trip_12mo %>%
  filter(ride_length >= 0)

# Remove raw data frame
remove(bike_trip_12mo)
```

Forth, we need format ride_length from *difftime* datatype to *numeric* datatype to ensure all futur calculation based on this column is proper:

```
bike_trip_12mo_cleaned$ride_length <- as.numeric(as.character(bike_trip_12mo_cleaned$ride_length))
```

Fifth, we need to calculate day-of-week based on the **started_at** column:

```
# Add day_of_week column
bike_trip_12mo_cleaned$day_of_week <- wday(bike_trip_12mo_cleaned$started_at)
```

The newly added **day_of_week** column will have numbers from 1 to 7 which representing Sunday to Saturday respectively.

Lastly, we calculate **trip_distance** variable which derived from start_lng, start_lat, end_lng, and end_lat (in that order):

```
bike_trip_12mo_cleaned <- bike_trip_12mo_cleaned %>%
  rowwise %>%
  mutate(trip_distance = as.vector(distm(x = c(start_lng, start_lat), y = c(end_lng, end_lat), fun = di
```

With trip_distance data, it can provide a different angle on the data usage habit between memeber riders and causal riders.

Data Analyze

First, we conduct general **Descriptive Analysis** as follow:

```
summary(bike_trip_12mo_cleaned$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     382     680    1241    1236 3356649
```

```
table(bike_trip_12mo_cleaned$member_casual)
```

```
##
##  casual  member
## 2559796 3300841
```

```
mean(bike_trip_12mo_cleaned$ride_length)        # In second (1,241.409 sec)
```

```
## [1] 1241.409
```

```
max(bike_trip_12mo_cleaned$ride_length)         # In second (3,356,649 sec)
```

```
## [1] 3356649
```

```
find_mode(bike_trip_12mo_cleaned$day_of_week)   # Saturday appears the most
```

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

Second, we aggregate the data to calculate the average **ride_length** by user types:

```
avg_ride_length_usertype <- aggregate(bike_trip_12mo_cleaned$ride_length,
          by = list(bike_trip_12mo_cleaned$member_casual), FUN = mean)

colnames(avg_ride_length_usertype) <- c("member_casual", "avg_rider_length")
```

Third, we aggregate the data to calculate the average **ride_length** by user types and day-of-week:

```
avg_ride_length_usertype_dayofweek <- aggregate(bike_trip_12mo_cleaned$ride_length,
          by = list(bike_trip_12mo_cleaned$member_casual, bike_trip_12mo_cleaned$day_of_week), FUN = mea

colnames(avg_ride_length_usertype_dayofweek) <- c("member_casual", "day_of_week", "avg_ride_length")

avg_ride_length_usertype_dayofweek <- avg_ride_length_usertype_dayofweek %>%
  mutate(day_of_week = wday(day_of_week, label=TRUE, abbr=FALSE))

avg_ride_length_usertype_biketype_dayofweek <- aggregate(bike_trip_12mo_cleaned$ride_length,
                                                by = list(bike_trip_12mo_cleaned$member_casual,
                                                          bike_trip_12mo_cleaned$rideable_type,
                                                          bike_trip_12mo_cleaned$day_of_week),
                                                FUN = mean)

colnames(avg_ride_length_usertype_biketype_dayofweek) <- c("member_casual", "rideable_type", "day_of_wee

avg_ride_length_usertype_biketype_dayofweek <- avg_ride_length_usertype_biketype_dayofweek %>%
  mutate(day_of_week = wday(day_of_week, label=TRUE, abbr=FALSE))
```

Forth, we do a count on rider by user types and day-of-week:

```
count_rides_usertype_dayofweek <- bike_trip_12mo_cleaned %>%
  count(member_casual, day_of_week, name = "number_of_rides")

count_rides_usertype_dayofweek <- count_rides_usertype_dayofweek %>%
  mutate(day_of_week = wday(day_of_week, label=TRUE, abbr=FALSE))

count_rides_usertype_biketype_dayofweek <- bike_trip_12mo_cleaned %>%
  count(member_casual, rideable_type, day_of_week, name = "number_of_rides")

count_rides_usertype_biketype_dayofweek <- count_rides_usertype_biketype_dayofweek %>%
  mutate(day_of_week = wday(day_of_week, label=TRUE, abbr=FALSE))
```

Fifth, we are to calculate the average, minimum, maximum trip distance by user types and day-of-week:

```
# Avg, min, and max trip distance for each type of users by day_of_week
bike_trip_distance_stat <- bike_trip_12mo_cleaned %>%
  mutate(day_of_week = wday(started_at, label = TRUE, abbr = FALSE)) %>%
  group_by(member_casual, day_of_week) %>%
  summarise(average_trip_distance = mean(trip_distance, na.rm = TRUE),
            max_trip_distance = max(trip_distance, na.rm = TRUE),
            min_trip_distance = min(trip_distance, na.rm = TRUE)) %>%
  arrange(member_casual, day_of_week)
```

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

Please note that we exclude the **NA** values in the data to ensure the statistic accuracy

Lastly, we export varies summary information to CSV files for further analysis on findings and visualization purpose:

```
write_csv(avg_ride_length_usertype, "avg_ride_length_usertype.csv")
write_csv(avg_ride_length_usertype_dayofweek, "avg_ride_length_usertype_dayofweek.csv")
write_csv(count_rides_usertype_dayofweek, "count_rides_usertype_dayofweek.csv")
write_csv(bike_trip_distance_stat, "bike_trip_distance_stat.csv")
```

Analysis Findings

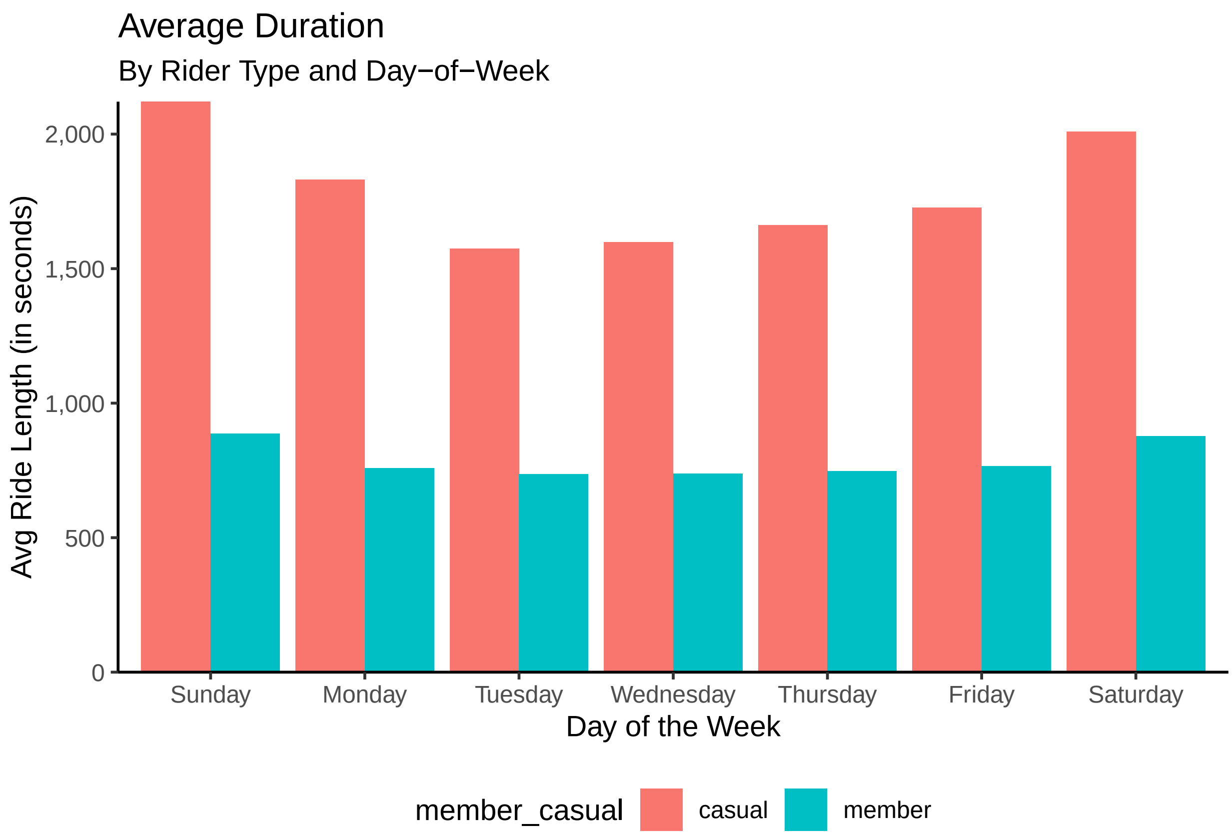
Based on the descriptive analysis on the past 12 months ridership data:

- The day of the week with highest numbers of rides is: Saturday
- On average, the ride duration is: 20.69015 minutes (1,241.409 sec)

As the following graph shows, for the average ride duration (in seconds), Cyclistic Members are relatively constant. However, for Casual Riders, the ridership dipped slightly during the weekdays period. That said, the graph depicts that Casual Riders, in general, spent longer time riding than Cyclistic Members. We can derive a **hypothesis that Cyclistic Members are using the Cyclistic bikes for their daily commute to school or work, which usually in close proximity. On the other hand, for Casual Riders with longer ride duration, this might indicates that they tended to travel to further area for leisure purpose**.

```
avg_ride_length_usertype_dayofweek %>%
  ggplot(aes(x = day_of_week,
             y = avg_ride_length,
             fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_y_continuous(labels = comma,
```

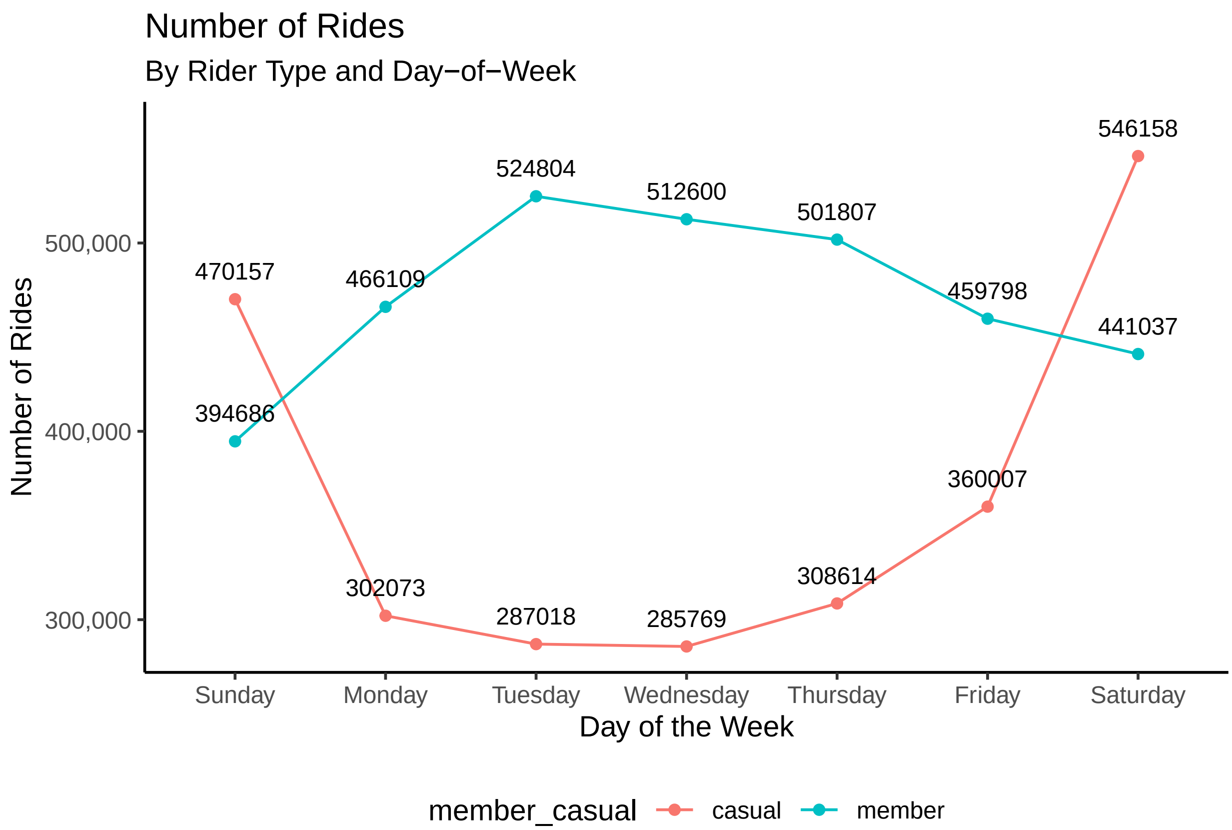
```
                      expand = c(0, 0)) +
  theme_classic() +
  theme(legend.position = "bottom") +
  labs(title = "Average Duration",
       subtitle = "By Rider Type and Day-of-Week",
       x = "Day of the Week",
       y = "Avg Ride Length (in seconds)")
```



The data also depicted ridership trend between Casual Riders and Cyclistic Members for the past 12 months by day-of-week:

```
count_rides_usertype_dayofweek %>%
  ggplot(aes(x = day_of_week,
             y = number_of_rides,
             group = member_casual,
             label = number_of_rides)) +
  geom_line(aes(color = member_casual)) +
  geom_point(aes(color = member_casual)) +
  geom_text(nudge_y = 15000,
            size = 3) +
  scale_y_continuous(labels = comma) +
  theme_classic() +
  theme(legend.position = "bottom") +
  labs(title = "Number of Rides",
       subtitle = "By Rider Type and Day-of-Week",
       x = "Day of the Week",
```

```
       y = "Number of Rides")
```

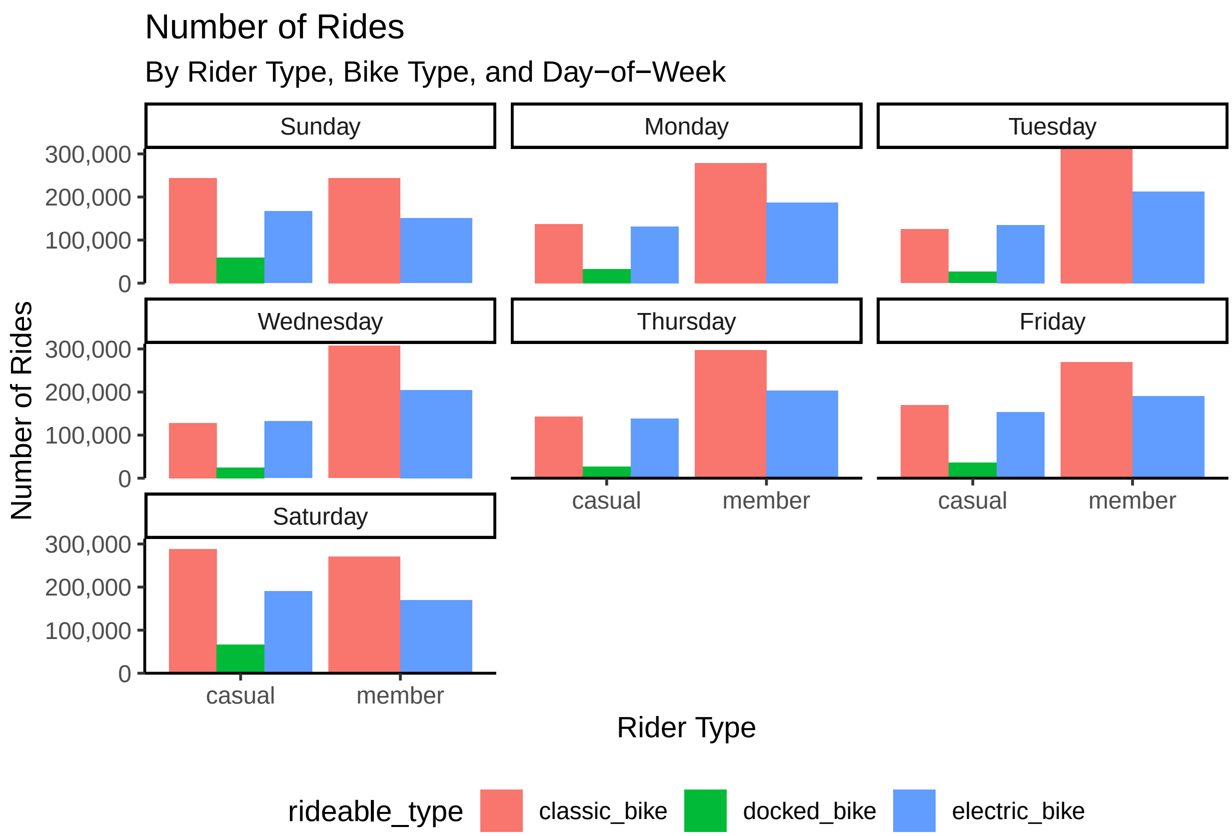


The line graph shows that during weekdays, Cyclistic Members have higher ridership than Casual Riders group. However, the ridership of Casual Riders increased significantly during each weekends throughout the 12 months period. This implies and reinforces our hypothesis indicated above, i.e., Cyclistic Members are using the Cyclistic bikes for their daily commute in close proximity and Casual Riders are using the bikes to travel to further for leisure.

Additionally, if we include rideable type (i.e., bike type) data into the analysis, we would see that the number of rides pertaining to Casual Riders is relatively low compare to Cyclistic Members during weekdays. On the other hand, the ridership number increased to the level on par of the ridership number of Cyclistic Members during weekends. In facts, the ridership number of Cyclistic Members is constant throughout the week. This implies that to convert Casual Rider to Cyclistic Member, the company will need change the Casual Ridership habit or behavior.

```
count_rides_usertype_biketype_dayofweek %>%
  ggplot(aes(x = member_casual,
             y = number_of_rides,
             fill = rideable_type)) +
  geom_col(position = "dodge") +
  scale_y_continuous(labels = comma,
                     expand = c(0, 0)) +
  facet_wrap(vars(day_of_week), ncol = 3) +
  theme_classic() +
  theme(legend.position = "bottom") +
  labs(title = "Number of Rides",
       subtitle = "By Rider Type, Bike Type, and Day-of-Week",
```

```
      x = "Rider Type",
      y = "Number of Rides")
```



Recommending Actions

In conclusion, based on the past 12 months ridership data, there is a significant usage gap between Casual Rider and Cyclistic Member during the weekdays. This might be contributed by different habit/behavior is bike usage, i.e., based on the data Cyclistic Member ridership is constant throughout the week which might indicates that they are using the bikes for their daily commute. As for Casual Rider, we only see a usage number flipped during weekends time, thus this indicates they are using the bike for leisure purpose.

Therefore, to close the usage gap between the ridership habit between the two groups, the following are the top three recommendation:

1. Encourage Casual Riders to use bikes for their daily commutes by providing membership discount for the first 3 months. This allows the Casual Riders get a taste of using bikes for daily commutes, thus to direct their usage habit.
2. 3 months discount membership for Casual Riders with electric bike and classic bike can be used with no extra charge/fee. With electric bike, this allows the Casual Riders to change their commute habit without breaking a sweat, thus provide more upside of possible continue usage.
3. Promote bike riding as health improving exercise and eco-friendly to combat global warming.