Divvy Bike Share Project

Google Data Analytic Professional Certificate

Jeffrey Teng

Identifying Business Task:

Cyclistic is a bike share program that features more than 5,800 bicycles and 600 docking stations. There are two types of customers, members and casual, according to the Cyclistic's finance analysts, annual members are much more profitable than the casual riders. In the effort of increased revenue, there are some questions that I need to answer in order to achieve the goals.

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. What is the most common starting and ending station?
4. How and when can Cyclistic use digital media to influence casual riders to become members?

Prepare Phase:

The data used for this analysis project was collected by Cyclistic. The data contains the riders' patterns from July 2019 to June 2021. The data was located on the Divvy server and was saved as a csv file. It is a public open data and is provided according to the Divvy Data License Agreement.

The data contained the following fields:
ride_id: id number per ride
rideable_type: the type of bike used
started_at: the date and time of the bike was checked out
ended_at: the date and time of the bike was checked in
start_station_name: the name of the station when the bike was checked out
end_station_name: the name of the station when the bike was checked in
start_station_id: unique id for the start station
end_station_id: unique id for the end station
start_lat: the latitude of the start station
start_lng: the longitude of the start station
end_lat: the latitude of the end station
end_lng: the longitude of the end station
member_casual: indicating whether the bike was checked out by the casual customer or the member

New field added to the data set:
day_of_week: the day of the week based started_at
season: the season based on started_at
total_ride_length_in_minutes: the total duration of ride lengths for all customers

average_ride_length_in_minutes: the average duration of ride lengths
number_of_trips: total number of trips

Given this is an internal data by the Cyclistic, it is safe to assume that this is unbiased, credible, and secured. Even though there are some errors in the data such as that ended_at was before the started_at, some missing informations in start_station_name and end_station_name, and the durations of the rides were more than 10,000 minutes.

Process Phase:

Tool used for the analysis:

Excel: Used this software for initial data cleaning and processing such as changing the date format to YYYY:MM:DD HH:MM:SS and filling the empty cells with 0 in order to import csv file into MySQL workbench

MySQL: Used this to import, clean, process, extract, and export data. Also used it to analyze the average ride length, total ride length, total number of trips, top 5 start station name, and top 5 end station name based on member or casual, day of week, season, and types of bikes.

Tableau: Used this to create multiple visualizations from the csv file exported from MySQL Workbench. Also created a dashboard of the project using this software.

Excel Cleaning and Manipulation:

For the “started_at” and “ended_at” columns, I converted them to datetime format of YYYY:MM:DD HH:MM:SS in order to use the load file function of MySQL Workbench.

For the “member_casual” column, the data before March 2020 was using Subscriber and Customer as data input so in order to keep the data consistent, I replace Subscriber as member and Customer as casual.

For the lat and lng column, since in this project, latitude and longitude will not be the focus and for the data before March 2020 lat and lng are not included so to keep it consistent, I delete the lng and lat from the rest of the excel files.

For the empty column, in order to load csv file into MySQL Workbench, null entry is not accepted so I replace empty entries with 0.

MySQL Cleaning and Manipulation:

Firstly, I created a database called project_bike_share then created table with the name and datatypes based on the excel sheets. Then I imported the csv file using the load data infile methods

```sql
-- Create database
create database project_bike_share;
use project_bike_share;

-- Create datatable
CREATE TABLE bike_share_data (
ride_id VARCHAR(255),
    rideable_type VARCHAR(255),
    started_at DATETIME,
    ended_at DATETIME,
    start_station_name VARCHAR(255),
    start_station_id VARCHAR(255),
    end_station_name VARCHAR(255),
    end_station_id VARCHAR(255),
    member_casual VARCHAR(255)
);

-- Import CSV into bike_share_data table
LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/Data.csv'
INTO TABLE main_data
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
```

Then with some data exploration, I found out that some of the “ended_at” value comes before the “started_at” value so in this case I created two more column called “started_at_new” and “ended_at_new” and used case statement to see if I need to swap the value of “started_at” and “ended_at” or keep the same base on the condition of if “started_at” is greater than “ended_at”. Then, drop the original “started_at” and “ended_at” column.

Lastly, find out the duration of the ride length and delete all the rows that duration last more than 5 hours.

```sql
-- Data Cleaning
alter table main_data add column started_at_new datetime;
alter table main_data add column ended_at_new datetime;
UPDATE main_data
SET
    started_at_new = CASE
        WHEN started_at > ended_at THEN ended_at
        ELSE started_at
    END;
UPDATE main_data
SET
    ended_at_new = CASE
        WHEN started_at > ended_at THEN started_at
        ELSE ended_at
    END;

alter table main_data drop column started_at, drop column ended_at;
```

Analysis Phase:

Total number of trips and the percent from total trips between member and casual customer

```sql
-- Total number of trips based on casual or member
SELECT
    member_casual AS member_or_casual,
    COUNT(ride_id) AS number_of_trips,
    ROUND((COUNT(ride_id) / (SELECT
                    COUNT(ride_id)
                FROM
                    main_data) * 100),
            2) AS percentage_of_total_trips
FROM
    main_data
GROUP BY member_casual;
```

Member or Casual	Number of Trips	Percentage from Total
Member	4,580,009	63.20
Casual	2,666,818	36.80

Total number of trips, total number of ride length in minutes, average number of ride length in minutes group by day of week, member or casual customer, and different type of bikes.

```sql
#Calculate the average ride_length for users by day_of_week
SELECT
    DAYNAME(started_at_new) AS day_of_week,
    member_casual AS member_or_casual,
    rideable_type AS type_of_ride,
    COUNT(ride_id) AS number_of_trips,
    ROUND(SUM(TIMESTAMPDIFF(MINUTE,
                started_at_new,
                ended_at_new)),
            2) AS total_ride_length_in_minutes,
    ROUND(AVG(TIMESTAMPDIFF(MINUTE,
                started_at_new,
                ended_at_new)),
            2) AS average_ride_length_in_minutes
FROM
    (SELECT
        ride_id,
            rideable_type,
            DAYNAME(started_at_new),
            ended_at_new,
            started_at_new,
            member_casual
    FROM
        main_data) a
GROUP BY day_of_week , member_casual , rideable_type
ORDER BY DAYOFWEEK(started_at_new) , member_casual , rideable_type
LIMIT 1 , 100;
```

Total number of trips, total number of ride length in minutes, average number of ride length in minutes group by season, member or casual customer, and different type of bikes.

```sql
#Different season data exploration
SELECT
    season,
    member_casual AS member_or_casual,
    rideable_type AS type_of_ride,
    COUNT(ride_id) AS number_of_rides,
    ROUND(SUM(TIMESTAMPDIFF(MINUTE,
                started_at_new,
                ended_at_new)),
            2) AS total_ride_length_in_minutes,
    ROUND(AVG(TIMESTAMPDIFF(MINUTE,
                started_at_new,
                ended_at_new)),
            2) AS average_ride_length_in_minutes
FROM
    (SELECT
        ride_id,
            started_at_new,
            ended_at_new,
            member_casual,
            rideable_type,
            CASE
                WHEN MONTH(started_at_new) IN (3 , 4, 5) THEN 'Spring'
                WHEN MONTH(started_at_new) IN (6 , 7, 8) THEN 'Summer'
                WHEN MONTH(started_at_new) IN (9 , 10, 11) THEN 'Autumn'
                WHEN MONTH(started_at_new) IN (12 , 1, 2) THEN 'Winter'
            END AS season
    FROM
        main_data) t1
GROUP BY season , member_casual , rideable_type
ORDER BY CASE
    WHEN season = 'Spring' THEN 1
    WHEN season = 'Summer' THEN 2
    WHEN season = 'Autumn' THEN 3
    WHEN season = 'Winter' THEN 4
END , member_casual ASC , rideable_type ASC
LIMIT 1 , 100;
```

Most Common Start and End Station for member and casual:

```sql
SELECT
    member_casual AS member_or_casual,
    start_station_name,
    COUNT(start_station_name)
FROM
    main_data
WHERE
    start_station_name != '0'
GROUP BY member_or_casual , start_station_name
ORDER BY member_casual , COUNT(start_station_name) DESC;
SELECT
    member_casual AS member_or_casual,
    end_station_name,
    COUNT(end_station_name)
FROM
    main_data
WHERE
    end_station_name != '0'
GROUP BY member_or_casual , end_station_name
ORDER BY member_casual , COUNT(end_station_name) DESC;
```

Total and Average Ride Lengths:

Member or Casual	Total length of ride in minutes	Average length of ride in minutes
Member	84,188,117	31.57
Casual	61,530,642	13.43

Busiest Weekdays:

Over the two-year period the busiest day of the week for member is Wednesday with 371,140 number of trips and for casual customer is Saturday with 325401 number of trips.

From the data, members used the bike share service more during the weekdays and the causal customers used it more during the weekend. From this, one of the hypothesis is that the member used the service primarily for commute to work and casual customer used it when they are out on break.

Busiest Season:

Over the two-year period the busiest season for both member and casual customers are both summer with 1,874,339 of total number of trips for member and 1,409,428 of total number of trips for casual customer. As for average ride length in minutes, member is 14.61 minutes and casual customer is 34.15 minutes.

Rideable Type Preference:

Before 2020-07-13, there were only one type of bike option with is docked bike. After 2020-07-13, there were two more types of bike which are classical bike and electric bike. Overall docked bikes were the most popular bike that the customer used, second is electric bikes, and third is classical bikes.

Busiest Station:

<u>Top 5 Start Station Name for Member:</u>

1. Clinton St & Madison St
2. Clark St & Elm St
3. Canal St & Adams St
4. Kingsbury St & Kinzie St
5. Clinton St & Washington Blvd

<u>Top 5 Start Station Name for Casual:</u>

1. Streeter Dr & Grand Ave
2. Lake Shore Dr & Monroe St
3. Millennium Park
4. Michigan Ave & Oak St

5. Lake Shore Dr & North Blvd

Top 5 End Station Name for Member:

1. Clark St & Elm St
2. Clinton St & Madison St
3. Kingsbury St & Kinzie St
4. Canal St & Adams St
5. Clinton St & Washington Blvd

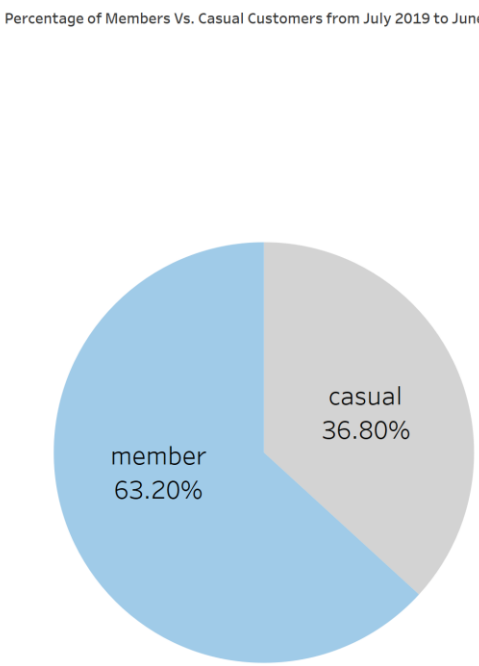
Top 5 End Station Name for Casual:

6. Streeter Dr & Grand Ave
7. Lake Shore Dr & Monroe St
8. Millennium Park
9. Lake Shore Dr & North Blvd
10. Michigan Ave & Oak St

Shared Phase:

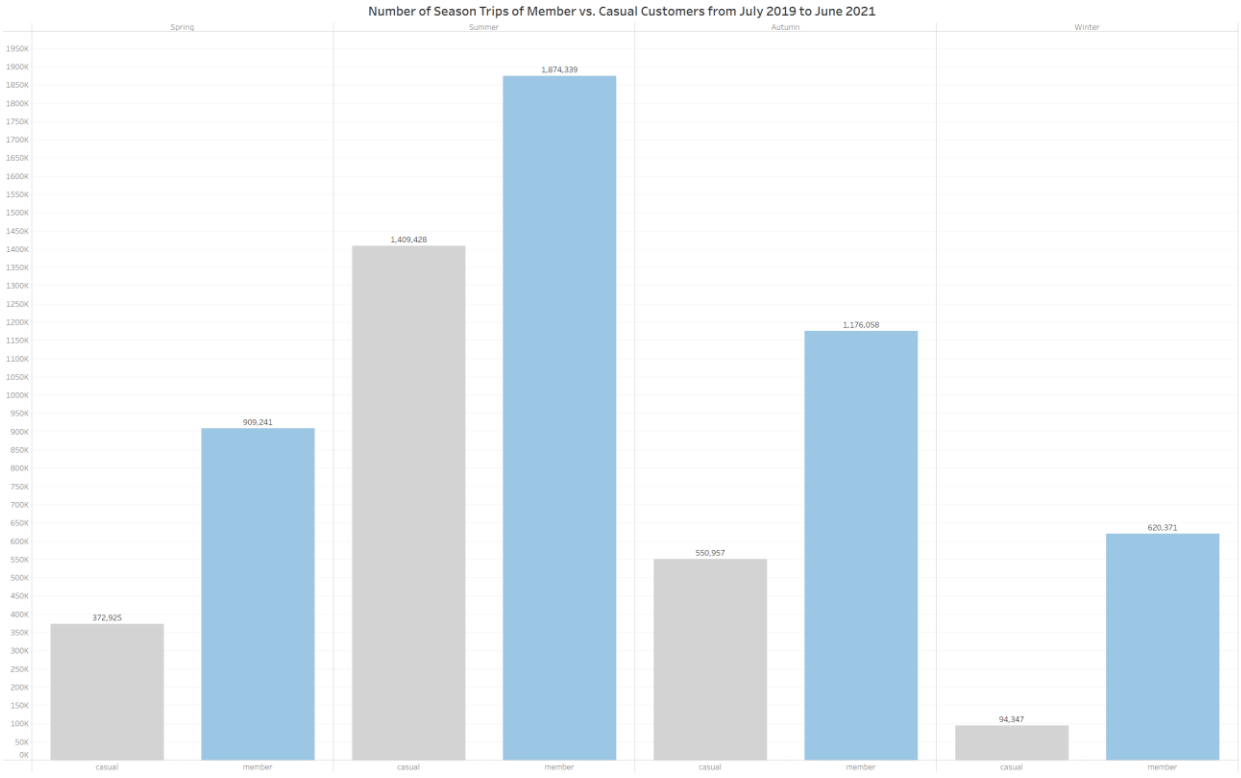
Total Number of Trips:

The primary goal of the analysis is to find a way to make casual rider to buy membership so it is important to know the current customer bases. As you can see in the chart below, the percentage that form the customer bases are member with 63.20 percent and casual customers with 36.80 percent. This indicates that there are still rooms for converting casual customers into members.



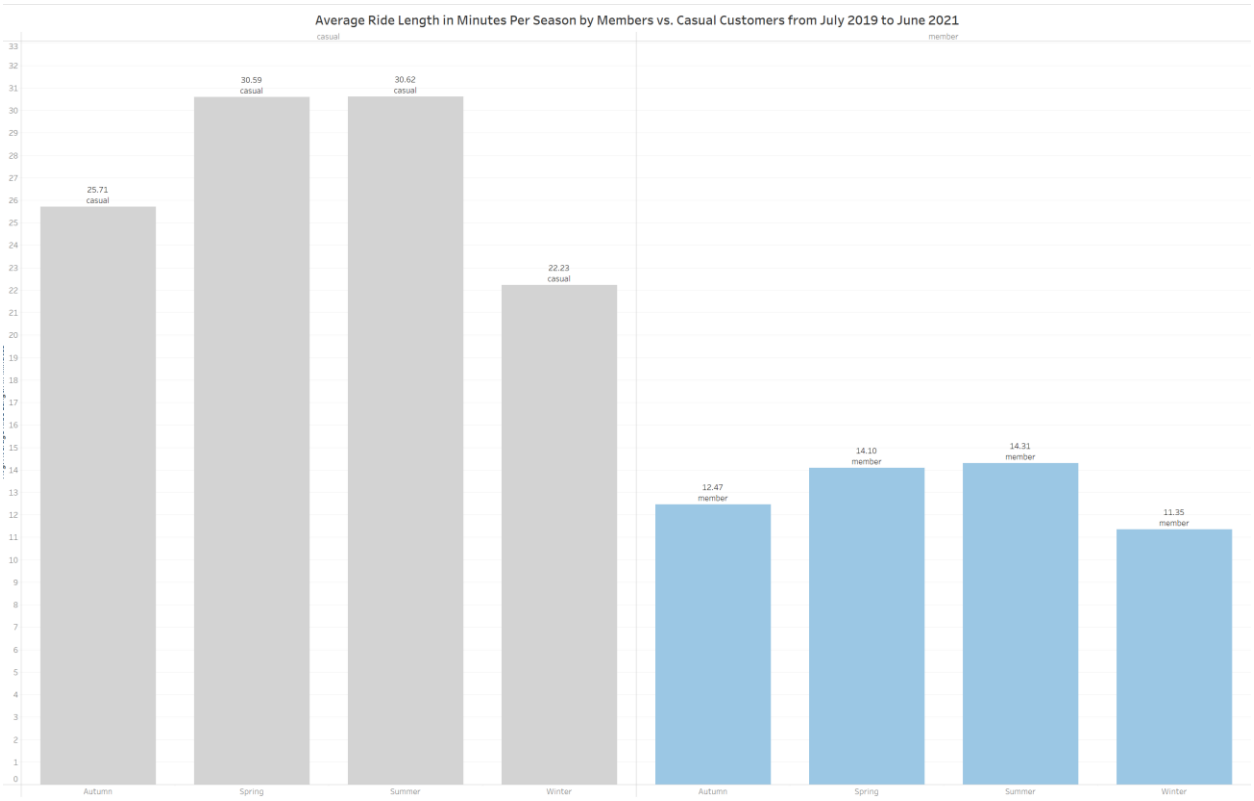
Season Trips:

In terms of season trips, from the chart below you can see that Summer is the seasons that most of the trips occurred and Winter is the season with the least number of trips occurred. From this chart, the hypothesis are that the warmer the weather the more people out enjoying the weather, so if the company wants to do target advertising, summer time will be the best choice.



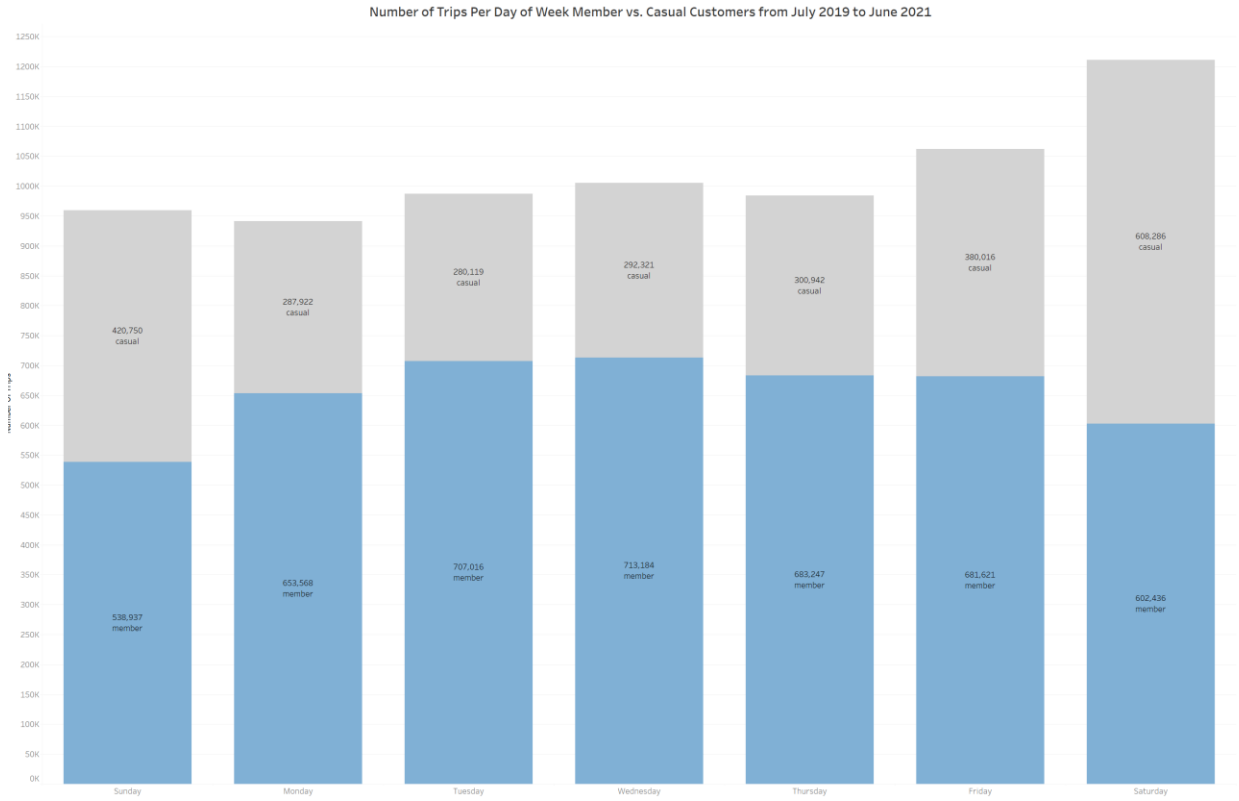
Total Ride Length and Average Ride Length:

The highest length of the average ride for member is 14.31 minutes in Summer and the highest for casual customer is 30.62 minutes in Summer. From the chart below we can see that the casual customers have a much higher length of average ride. One of the reason might be that the member are using the service primarily to commute to work and casual customers goes to more places and only check out when they are done visiting the places..



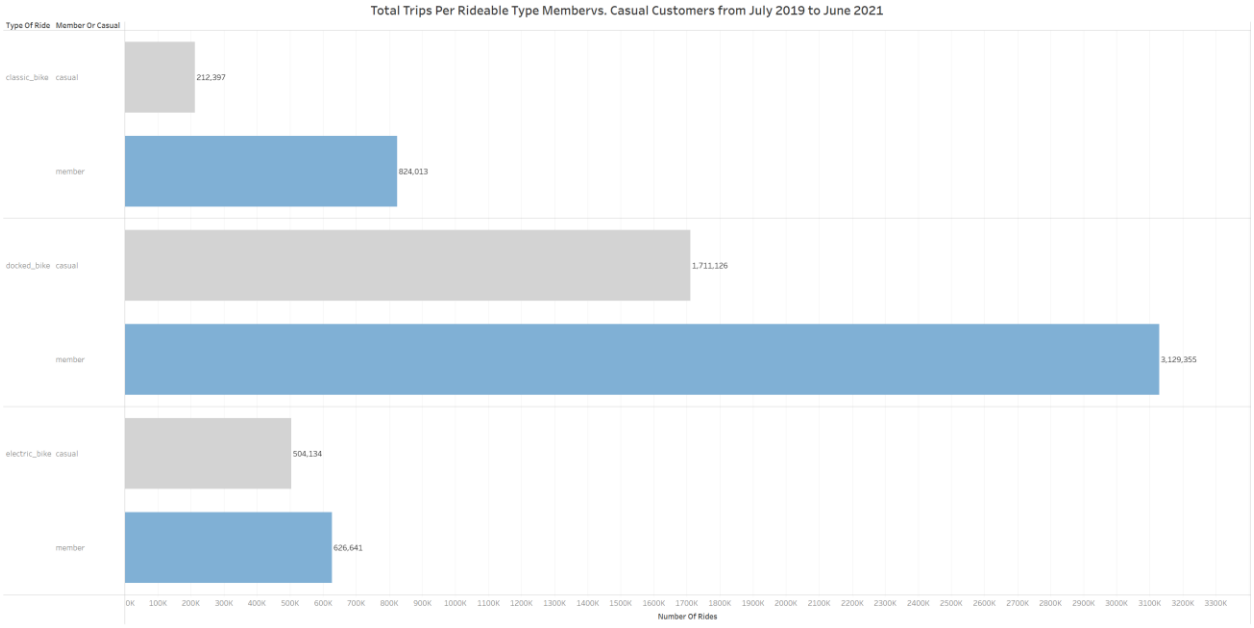
Busiest Weekday:

The chart below indicates that the Saturday is the most popular day of casual customer and Wednesday is the most popular day for member. From this chart we can see that for member the number of trips are over 60,000 rides for each day. On the other hand, for casual customers, only Saturday is the day that has over 60,000 rides. From this we can assume that the casual customers' use of the service are not primarily to commute to work but for enjoying the outdoor weather and members use this service primarily to commute to work.



Rideable type preference:

Docked bike is the most popular option of all for both member and casual customers. As for the second choice, members seems to like the classical bike better while the casual customer picks electric bike over classical bike.



Act Phase:

Based on the analysis above, these are my recommendations:

1. In order to attract more casual riders to subscribe to membership, advertising during the weekend will be better than weekdays due to the high number of trips of casual customers.
2. Warmer season such as Summer will be the best time to advertise, not only there are more casual customer out enjoying the weather, Summer is also where the students and teachers are out of school, which could be another target group of customers.
3. As for the poster advertisement or like station advertisement or board advertisement should be implemented at the TOP 5 casual customers' start and end station.
4. For the type of bike, company should focus on advertising and improving the usage of docked and electric bike and avoid classic bike in promotion due to lack of people using them

Some recommendation on collecting the data:

1. Assign unique id to both members and casual customers, this will allow the company to track the riding patterns in order to get a more accurate data. Also this can help to identify the casual riders' patter which can allow a more focus advertising options.
2. Adding how much each ride cost to the data, this will allow the company to see what kind of promotion they can do in order to attract more casual riders to convert to member.
3. Adding how much the bike is moving and stop. This will allow the company to get a more accurate ride duration time. Since right now there are like ride duration of 10000 minutes, it might be the bike was never parked until somebody found them. With adding how much time the bike is moving or stop can help company get a more accurate data and get a better analysis on total and average ride duration.