Google-Data-Analytics-Capstone-Case-Study-1

Manuel Madrid Gonzalez

2023-08-04

Introduction

Welcome to the Cyclistic bike-share analysis case study! In this case study, I will perform many real-world tasks of a junior data analyst. I will work for a fictional company, Cyclistic, and meet different characters and team members. In order to answer the key business questions, I will follow the steps of the data analysis process: ask, prepare, process, analyze, share, and act.

Quick links:

Data source: divvy-tripdata

Upload files: Upload files

Data preparation and exploration: Data preparation and exploration

Data cleaning: Data cleaning

Data analysis: Data analysis

Data visualization: Tableau 1 Tableau 2

Background

Cyclistic

A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members. Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual

members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Scenario

I am assuming to be a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, my team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, my team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve our recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Ask

Business task

Devise marketing strategies to convert casual riders to members.

Analysis Questions

Three questions will guide the future marketing program:

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

Moreno has assigned me the first question to answer: How do annual members and casual riders use Cyclistic bikes differently?

Prepare

Data source

I will analyze and identify trends with Cyclistic's historical trip data from January 2022 to December 2022 which can be downloaded at divvy-tripdata. Data provided by Motivate International Inc. under this license.

This is general information that can be used to understand how different types of customers use Cyclistic bikes. However, note that the Privacy Policy prohibits the use of personally identifiable information.

Data Organization

There are 12 files with the naming convention YYYYMM-divvy-tripdata each containing one month of information. Each file includes 13 columns with the following names ride_id, rideable_type, started_at, ended_at, start_station_name, start_station_id, end_station_name, end_station_id, start_lat, start_lng, end_lat, end_lng and member_casual.

Process

R is used to combine multiple data sets into one and clean it.

Reason: A worksheet in Microsoft Excel can only have 1048576 rows because it cannot handle large amounts of data. And the Cyclistic dataset contains more than 5.6 million rows.

Data preparation and exploration

R code: Upload files

R code: Data preparation and exploration

The 12 cvs files were uploaded as tables. Exploring the data allowed me to realize that the columns started_at and ended_at should be in date format but are in character format and also that table X202201 has a date format d/m/Y H:ma and the other tables have a date format Y-m-d H:m:s. So first the tables from 202202 to 202212 were combined in a new table called c_t and then the format of the columns started_at and ended_at was changed to date type for both table 202201 and table c_t, later both tables were combined in the table c_t.

I will use the ride_id column to search for duplicates, but as shown in the image the number of unique values of ride_id is 5667717 the same number of rows as the data frame so there are no duplicates or missing data to delete.

```
── Variable type: character ──────────────────────────────────────────────────────────────────────────────────────
  skim_variable      n_missing complete_rate min max empty n_unique whitespace
1 ride_id                    0         1       8  16     0  5667717          0
2 rideable_type              0         1      11  13     0        3          0
3 start_station_name    833064         0.853   7  64     0     1674          0
4 start_station_id      833064         0.853   2  44     0     1330          0
5 end_station_name      892742         0.842   9  64     0     1692          0
6 end_station_id        892742         0.842   2  44     0     1335          0
7 member_casual              0         1       6   6     0        2          0

── Variable type: numeric ────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate  mean     sd    p0   p25   p50   p75  p100
1 start_lat             0         1      41.9 0.0463  41.6  41.9  41.9  41.9  45.6
2 start_lng             0         1     -87.6 0.0300 -87.8 -87.7 -87.6 -87.6 -73.8
3 end_lat            5858         0.999  41.9 0.0681   0    41.9  41.9  41.9  42.4
4 end_lng            5858         0.999 -87.6 0.108  -88.1 -87.7 -87.6 -87.6   0

── Variable type: POSIXct ────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min                 max                 median              n_unique
1 started_at            0             1 2022-01-01 06:00:00 2022-12-31 23:59:26 2022-07-22 15:03:59  4676869
2 ended_at              0             1 2022-01-01 06:01:00 2023-01-02 04:56:45 2022-07-22 15:24:44  4689842
```

Now each value in the ride_id column should have a length of 16, so to find out if there is data that does not meet this parameter, a new column called ride_id_length will be created.

No values greater than 16 were found, but 78 values less than 16 were found.

The start_station_name, start_station_id, end_station_name, end_station_id, end_lat, and end_lng columns had missing data (816804, 816804, 874815, 874815, 5772 and 5772 respectively).

A new column ride_length will be created, which will contain the length of the ride in minutes. In this new column, values equal to or greater than 1440 minutes and values equal to or less than one minute will be searched. Finally, the week_day and month columns were created that show the day and month in which the trip was made.

Data cleaning

R code: Data cleaning

- All rows with missing values were deleted.
- 4 columns were added (ride_id_length, ride_length, day_week and month)
- Trips of less than one minute and more than one day were excluded.
- Values in the ride_is column with a length less than 16 were excluded.
- A total of 1376578 rows were deleted in this step.

Analyze and share

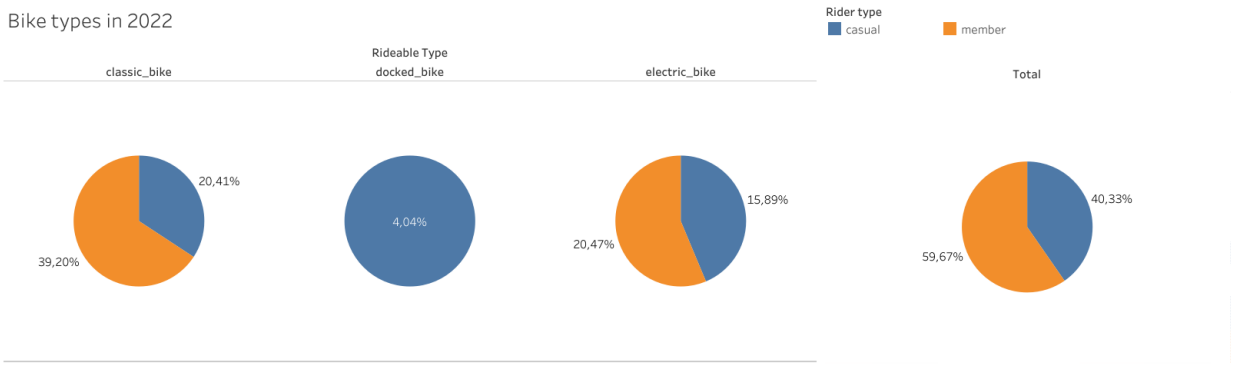
R Query: Data analysis

Data Visualization: Tableau 1 Tableau 2

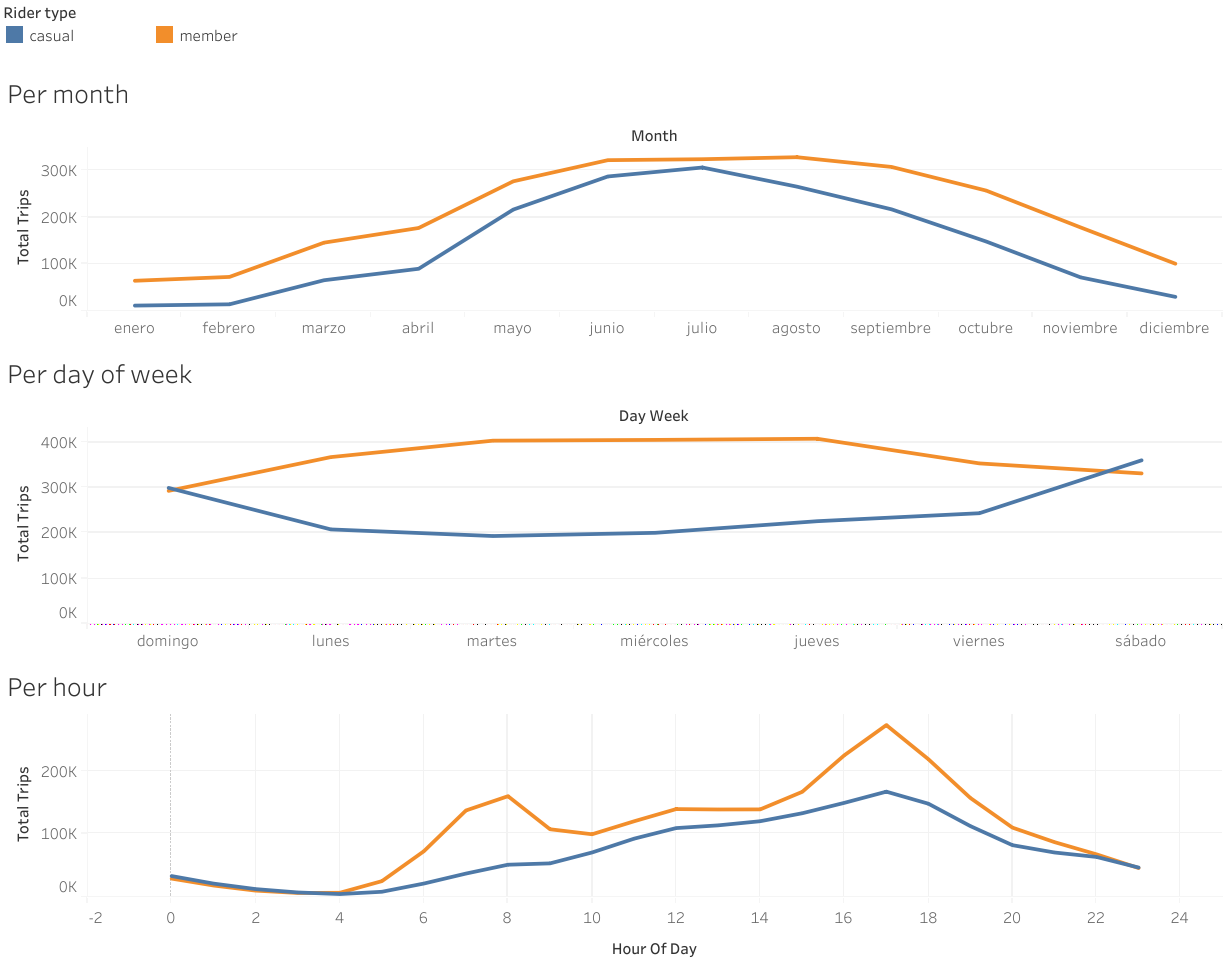
The data is now properly saved and ready for analysis. I perform various queries to create tables that were analyzed and visualized in Tableau.

The analysis question is: How do annual members and casual riders use Cyclistic bikes differently?

For the first question a comparison by type of bicycle used by annual and casual riders.



Each bike type graph shows a percentage of the total.The most used bicycles are classic bicycles followed by electric bicycles. Docked bikes are only used by casual riders and are the least used bikes.The members make 59.7% of the total while remaining 40.3% constitutes casual riders.

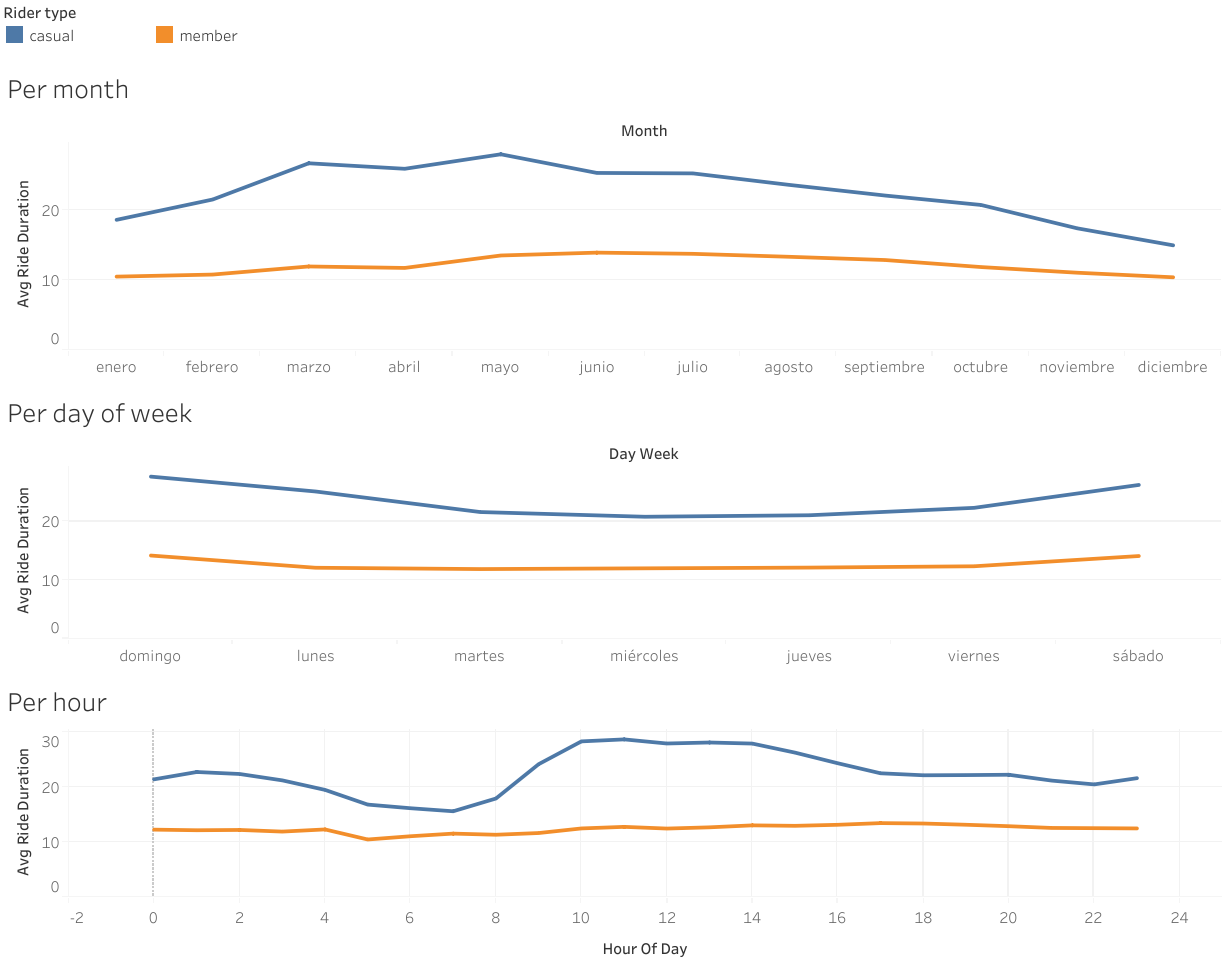


When comparing the number of trips per month, it was found that most of the trips occur in spring and summer and during the winter the number of trips decreases. Throughout the year the behavior of casual users and members is very similar and in the month of July, the difference in the number of trips decreases.

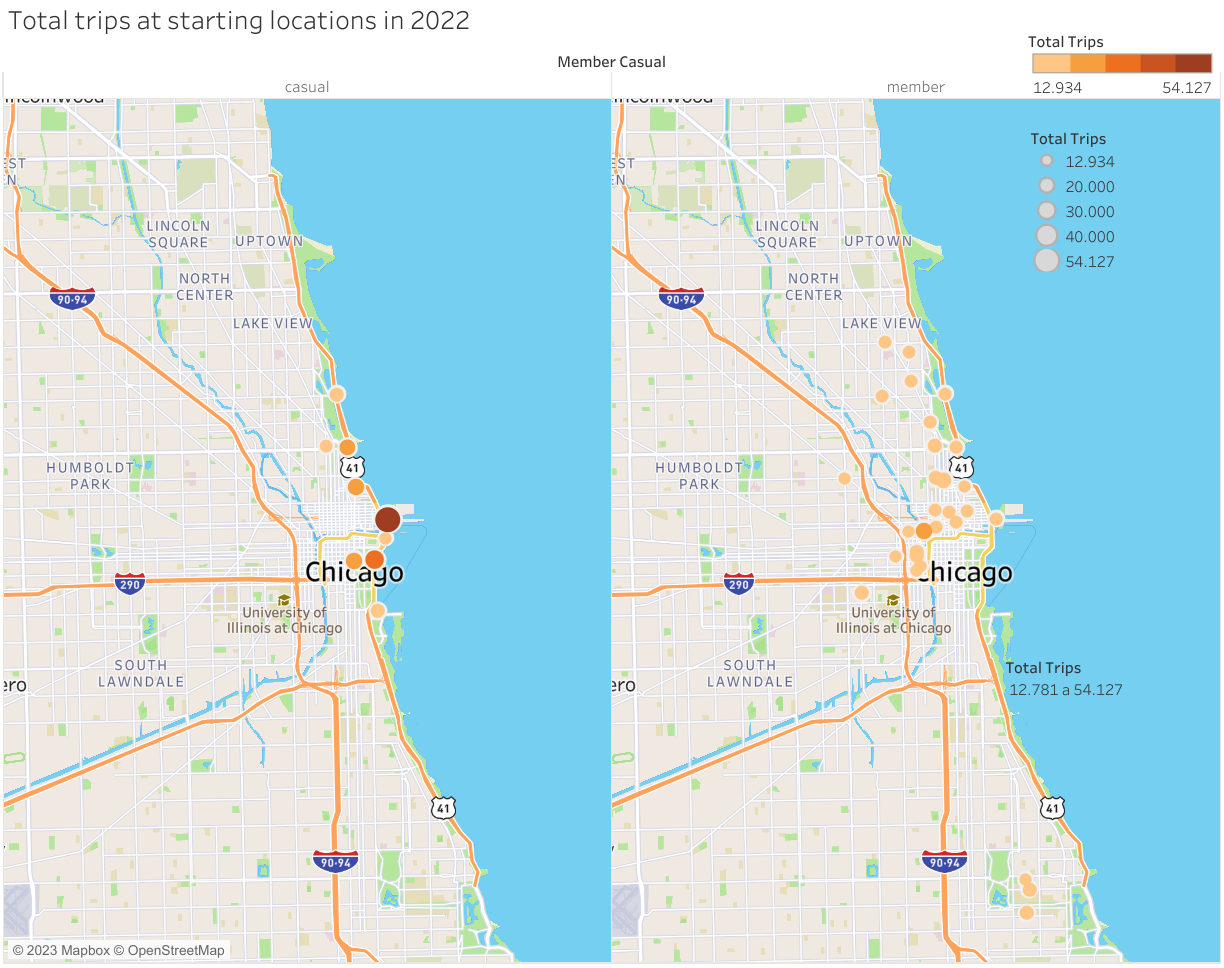
When observing the behavior of the users during the days of the week, it was found that the number of trips that members make decreases throughout the week, on the contrary, the number of trips that casual riders make increases during the weekends.

Throughout the day the members show two peaks in the number of trips, the first in the morning between 6 am and 8 am and the second peak is in the afternoon between 4 pm and 8 pm, on the other hand, the number of trips of casual riders increases steadily during the day until the evening when they begin to decline.

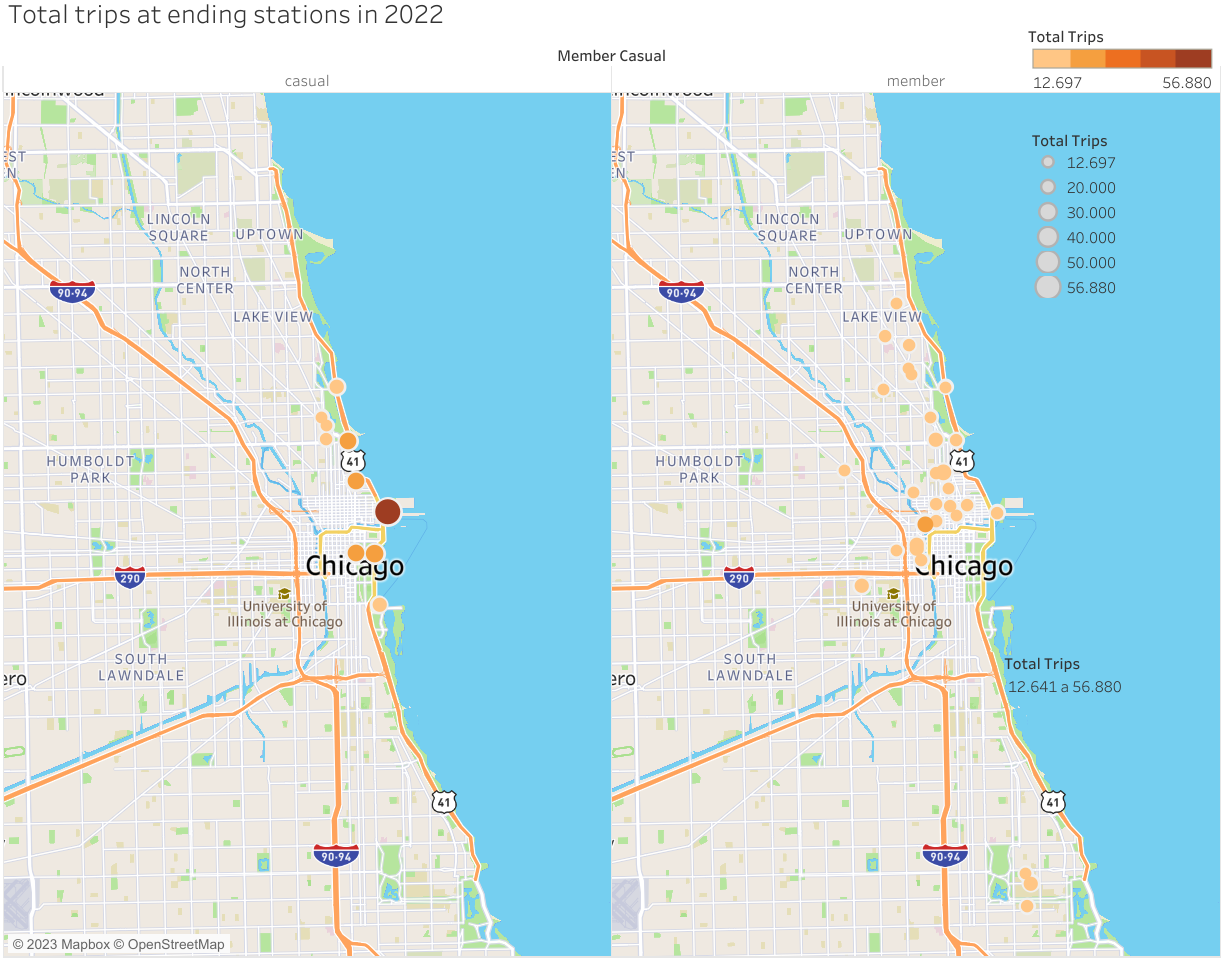
It can be said from observations that members may be using the bikes to get to and from work, while casual riders are using the bikes throughout the day and more frequently on weekends for leisure purposes. Both casual cyclists and members are most active during the spring and summer.



The average length of trips made by members during the year, day of the week, and time of day do not vary. While casual riders make longer trips during the spring and summer, on weekends, and between 10 am and 2 pm. It can be concluded that casual riders make trips twice as long but less frequently than members.



Members start their trips at stations near residential areas, hospitals, schools, banks, universities, train stations, parks, squares, and grocery stores. On the contrary, casual riders start their trips at stations close to parks, beach, aquariums, museums and harbor points.



Members and casual riders have similar behavior at ending station locations. Members and casual riders have similar behavior at ending-station locations. Members end their rides in locations close to commercial areas, residential areas, and universities while the casual riders end their rides near parks, museums, and other recreational sites. Coinciding with the above, casual riders use bikes for leisure, and members use them mainly for daily activities.

Summary:

Casual	Member
Use the bikes throughout the day, mostly on weekends during the spring and summer for leisure activities. They make trips twice as long but less frequent. Start and end their trips near parks, museums, and other recreational sites.	They use the bikes during the weekdays and during commute hours in spring and summer. They make shorter but more frequent trips. Start and end their trips near universities, residential areas, and commercial areas.

Act

After identifying the differences between casual and member riders, marketing strategies to target casual riders can be developed to persuade them to become members. Recommendations:

- Marketing campaigns might be conducted in spring and summer at tourist/recreational locations popular among casual riders.
- Casual riders are most active on weekends and during the summer and spring, thus they may be offered seasonal or weekend-only memberships.
- Casual riders use their bikes for longer durations than members. Offering discounts for longer rides may incentivize casual riders and entice members to ride for longer periods of time.