Cyclistic Bike-Sharing Data Analysis

Thanakorn Thanakraikiti (Nun)

July 03, 2023

Contents

1 Intoduction

Cyclistic is a successful American bicycle-sharing program that was established in 2016. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system at any time. Cyclistic users are more likely to ride for leisure, but approximately 30% use them to commute to work each day.

Cyclistic's marketing strategy has primarily focused on building general awareness and appealing to broad consumer segments. The program offers a variety of pricing plans, including single-ride passes, full-day passes, and annual memberships. Cyclistic classifies its riders into the following groups based on these plans: *casual riders* (users who purchase single-ride passes or full-day passes) and *annual members* (users who purchase an annual membership).

Cyclistic's flexible pricing plans attract a larger customer base, but financial analysts have determined that annual members are more profitable. However, casual riders are already aware of the Cyclistic program and have chosen Cyclistic to meet their mobility needs. This suggests that a marketing campaign that targets existing customers is likely to be more effective at expanding the business than a campaign that targets only new customers.

Therefore, Cyclistic's marketing analytics team is interested in understanding how casual riders and annual members use Cyclistic bikes differently. By understanding these differences, the marketing analytics team can develop more targeted marketing strategies to convert casual riders into annual members.

1.1 Purpose

This report analyzes Cyclistic's historical bike trip data from 2022 in order to identify how annual members and casual riders use Cyclistic bikes differently. The report aims to extract insights and develop the most appropriate marketing strategies that appeal to casual riders and encourage them to subscribe to annual memberships. This will enable the business to improve productivity and respond to economic changes in a productive and efficient manner. The report outlines the behavioral differences in Cyclistic bike usage between annual members and casual riders. The report further demonstrates that the recommended solutions will help Cyclistic to expand its business growth.

1.2 Stakeholders

Key stakeholders include:

- Marketing analytics team who collect, analyze, and report data to help guide Cyclistic's marketing strategy
- Marketing manager who supports the marketing campaign to promote Cyclistic's bike-share program
- Executive team who approve the recommended marketing program

2 Data Sources

This report uses Cyclistic's historical trip data from January 2022 to December 2022 as its primary source. The data sets can be downloaded from the company's website, which can be find in the following details:

- Data Source: Trip History Data
- Data Website: System Data Website
- License: Data License Agreement

The data sets are collected by the City of Chicago (a government agency) and published by Cyclistic. They contain historical trip data for all of the company's customer (rider) data since 2013, including the following information: ride history, the location of stations for each ride, and other relevant information

2.1 Data Description

The data set is stored in twelve CSV files, one for each month of 2022 (*202201-divvy-tripdata.csv* through *202212-divvy-tripdata.csv*). Each file contains a set of fields (columns) and records (rows). Each record represents a single bike trip, which is uniquely identified by an identifier for the ride.

The data sets contain the following fields:

- `ride_id` : A unique identifier for the ride
- `rideable_type` : The type of bike for the ride (classic bike, electric bike or docked bike)
- `started_at` : The date and time at which the ride started
- `ended_at` : The date and time at which the ride ended
- `start_station_id` : The ID of the station where the ride started
- `start_station_name` : The name of the station where the ride started
- `end_station_id` : The ID of the station where the ride ended
- `end_station_name` : The name of the station where the ride ended
- `start_lat` : The latitude of the station where the ride started
- `start_lng` : The longitude of the station where the ride started
- `end_lat` : The latitude of the station where the ride ended
- `end_lng` : The longitude of the station where the ride ended
- `member_casual` : The type of user who took the ride (member or casual)

2.2 Data Credibility

The data is evaluated for bias and credibility using the ROCCC method:

- **Reliable**: The data is considered reliable because it is open data provided by Cyclistic and the City of Chicago. It is a comprehensive and accurate dataset of Cyclistic's historical bike trips. The data includes information on the start and end location of each trip, the time of day the trip was taken, the type of user (casual or annual member), and the length of the trip. There are some missing values in some fields, but these can be used to explore how different customer types use Cyclistic bikes.

- **Original**: The data is considered to be original, based on a primary source of information from the company.
- **Comprehensive**: It is important to note that Cyclistic is subject to data privacy restrictions that prohibit the use of riders' personally identifiable information (PII). This means that the company cannot connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes, as well as identify the rider's gender and age. Despite these restrictions, the data is complete and contains all the information needed to understand how members and casual riders use Cyclistic bikes. This includes information such as the number of rides taken, the length of each ride, the start and end stations of each ride, the time of day each ride was taken.
- **Current**: The data is current and updated monthly. The selected datasets are the most recent data available.
- **Cited**: The data is properly cited as it contains the dataset name, author, and published year.

Overall, the Cyclistic bike data is considered a credible source for analysis based on the Reliability, Originality, Comprehensiveness, Currency, and Citation (ROCCC) method. The data can be used to explore how different customer types use Cyclistic bikes. For example, the data can be used to determine which customer segments are most likely to use Cyclistic bikes for commuting, recreation, or transportation. The data can also be used to identify trends in bike usage, such as the time of day when bikes are most likely to be used or the types of bikes that are most popular.

While there are some restrictions on rider's personal information, there is still a great deal of information that can be learned from it. However, the data needs to be processed before analysis can be performed. This includes cleaning the data to remove errors and inconsistencies, and transforming the data into a format that is suitable for analysis. The data processing will be performed in the next section.

3 Data Cleaning and Manipulation

This section provides a documentation of data cleaning and manipulation. The ***R programming language*** is used in conjunction with the ***RStudio*** integrated development environment (IDE) to complete data cleaning's tasks, including loading packages, importing and combining the data, cleaning the data, transforming the data and removing outliers. R script is available on the GitHub repository.

3.1 Loading the Packges

Prior to cleaning the data, it is important to set up the R environment first. This requires loading the following packages into the environment:

- `tidyverse` for importing, wrangling and visualizing the data
- `skimr` for skimming the data
- `janitor` for exploring and cleaning the data

Note: *Make sure to install required packages using* `install.packages()` *before loading packages through* `library()` .

3.2 Importing and Combining the Data

Once packages are completely loaded, it is ready to import the data into the R environment. Each CSV file was imported into the environment. Subsequently, these data frames were combined into a single data frame, named ***trip_data***. Upon initial inspection, the data frame contains 5,667,717 rows and 13 columns. The table below provides a preview of the *trip_data* data frame.

Table 1: A preview of the Cyclistic's historical trip data

ride_id	98D355D9A9852BE9
rideable_type	classic_bike
started_at	2022-01-01 00:00:05
ended_at	2022-01-01 00:01:48
start_station_name	Michigan Ave & 8th St
start_station_id	623
end_station_name	Michigan Ave & 8th St
end_station_id	623
start_lat	41.87277
start_lng	-87.62398
end_lat	41.87277
end_lng	-87.62398
member_casual	casual

3.3 Cleaning the Data

The result of the `skim_without_charts()` function provides a concise summary of the *trip_data* dataset. The dataset contains 5,667,717 rows and 13 columns. The figure shows that the unique number of rows of `ride_id` is 5,667,717 rows, which is the same as the total number of rows. Therefore, it can be concluded that there are no duplicate values in the dataset.

Figure 1: A broad overview of dataset summary in the Cyclistic's historical trip data (*trip_data*) from `skim_without_charts()`

There are three data types in the dataset, including character, numeric and POSIXct. First, the **character** data type represents text or string values. Second, the **numeric** data type represents decimal values. Last, the **POSIXct** data type represents data and time values. Each data type include the following columns:

- **character:** `ride_id`, `rideable_type`, `start_station_name`, `start_station_id`, `end_station_name`, `end_station_id` and `member_casual`.
- **numeric:** `start_lat`, `start_lng`, `end_lat` and `end_lng`.
- **POSIXct:** `started_at` and `ended_at`.

The figure also shows that some columns have missing values. The following columns include:

- `start_station_name` (833,064 rows)
- `start_station_id` (833,064 rows)
- `end_station_name` (892,742 rows)
- `end_station_id` (892,742 rows)
- `end_lat` (5,858 rows)
- `end_lng` (5,858 rows)

There are multiple ways to handle missing data, but in this case, it was decided to remove those columns as they are not required for this analysis. The following R code was used to select only the useful columns and then store the new dataset as ***trip_data_v2***:

```
trip_data_v2 <- trip_data %>%
  select(ride_id, started_at, ended_at, rideable_type, member_casual)
```

Finally, the *trip_data_v2* dataset has only 5 columns, which are `ride_id`, `started_at` and `ended_at`, `rideable_type` and `member_casual`.

3.4 Transforming the Data

The next step is to add three new columns to the dataset, namely `month`, `day_of_week`, and `ride_length`, in order to facilitate data analysis and visualization. The `month` column will store the month in which the trip was taken, the `day_of_week` column will store the day of the week on which the trip was taken, and the `ride_length` column will store the duration of the trip in minutes. The following R codes can be used to add these three new columns to the dataset:

```
trip_data_v2$ride_length <- as.double(
  difftime(trip_data_v2$ended_at,
           trip_data_v2$started_at,
           units = "mins"))
trip_data_v2$day_of_week <- weekdays(trip_data_v2$started_at)
trip_data_v2$month <- format(trip_data_v2$started_at, "%b")
```

3.5 Removing Outliers

Once three new columns have been added to the dataset, it can be used for data analysis and visualization. The `skim_without_charts()` function is used to provides a concise summary of the *trip_data_v2* dataset again. The figure shows that there are minus number in minimum value (min) in the `ride_length` column.

Figure 2: A broad overview of dataset summary in the Cyclistic's historical trip data (*trip_data_v2*) from `skim_without_charts()`

This means that there are outliers in the datset. Outliers can be caused by errors in data entry, measurement, or other factors. They can distort the results of an analysis, so it is important to remove them before proceeding. In this case, all trip data that were less than 60 seconds were filtered out. These trips are likely to be false starts or users attempting to re-dock a bike to ensure its security. Trips that exceeded a

period of 24 consecutive hours were also removed. This is because these trips are likely to be the result of the rider not returning the bike. This was accomplished by running the following R code:

```
error_rows <- trip_data_v2$ride_length < 1
timeout_rows <- trip_data_v2$ride_length > 1440

trip_data_v2 <- trip_data_v2[!(error_rows | timeout_rows), ]
```

This completes the data cleaning and manipulation process, and the data is now ready for analysis.

4 Analysis and Key Findings

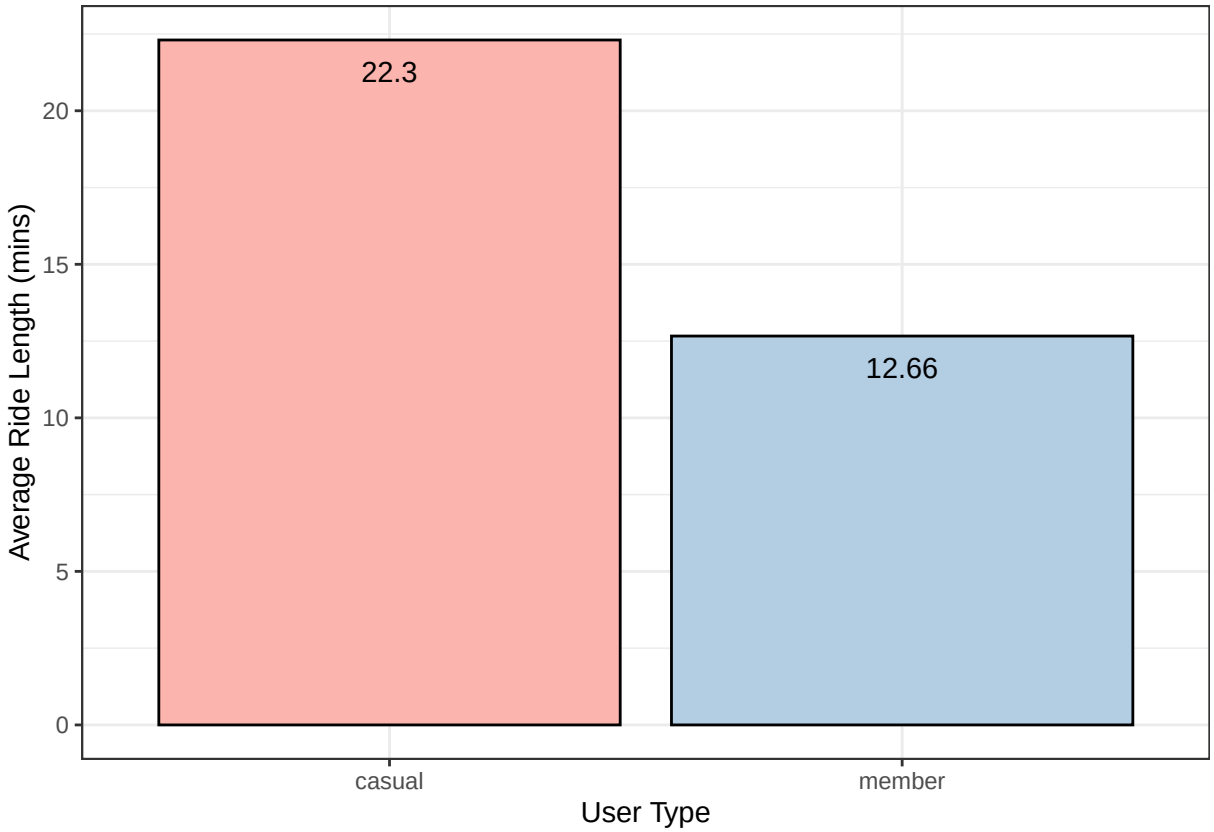
This section presents an analysis of Cyclistic's historical trip data from January 2022 to December 2022, with the objective of identifying the differences in the use of Cyclistic bikes between annual members and casual riders. The R script used for this analysis is available on the GitHub repository.

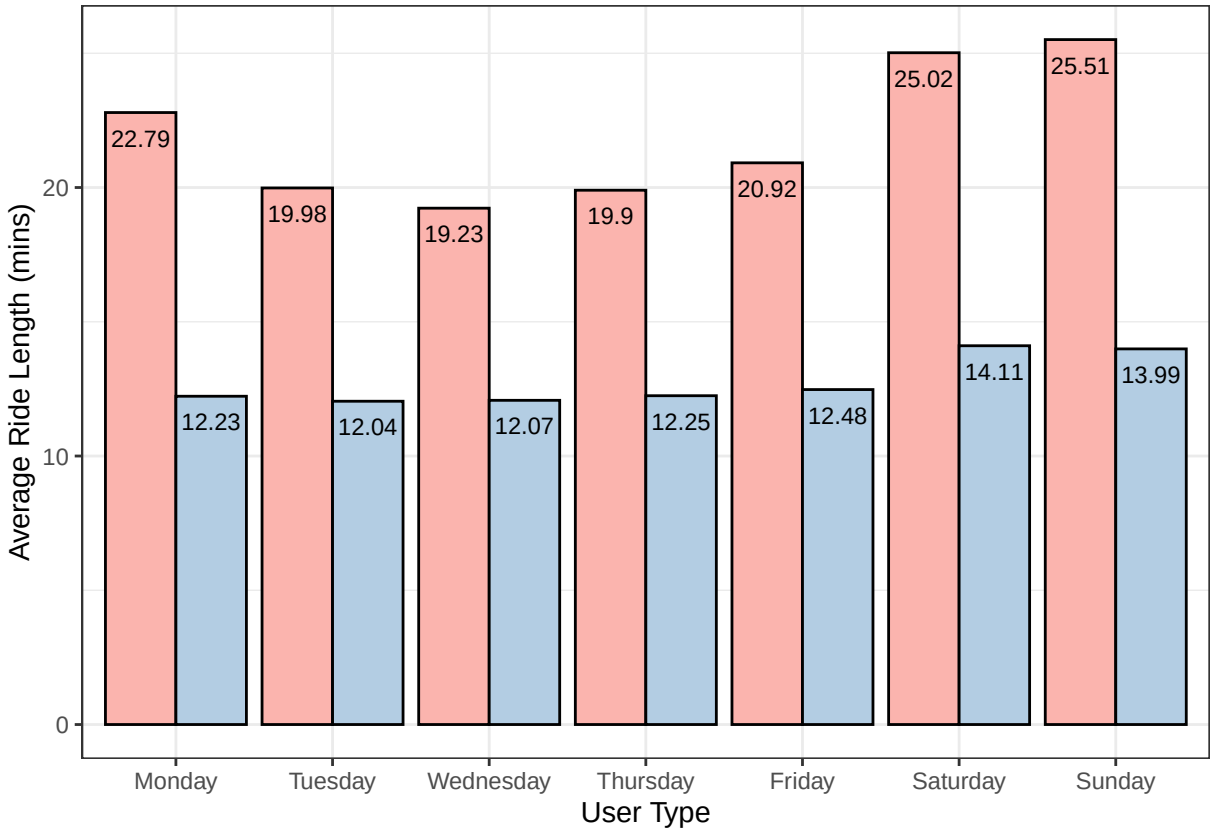
4.1 Overall Summary

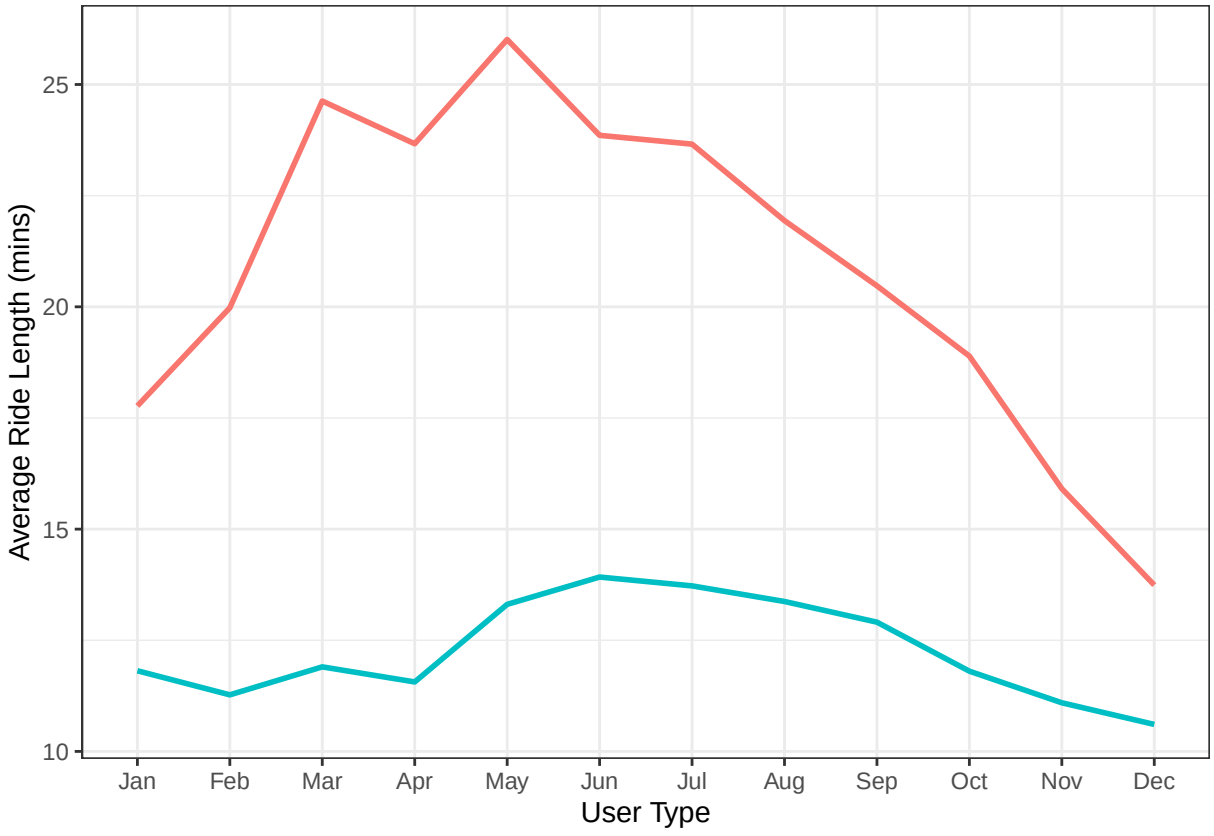
Table 2: A statistical summary of the duration of rides in the Cyclistic's historical trip data

mean	median	min	max	sd	count
16.61257	10.51667	1	1439.933	29.60487	5541268

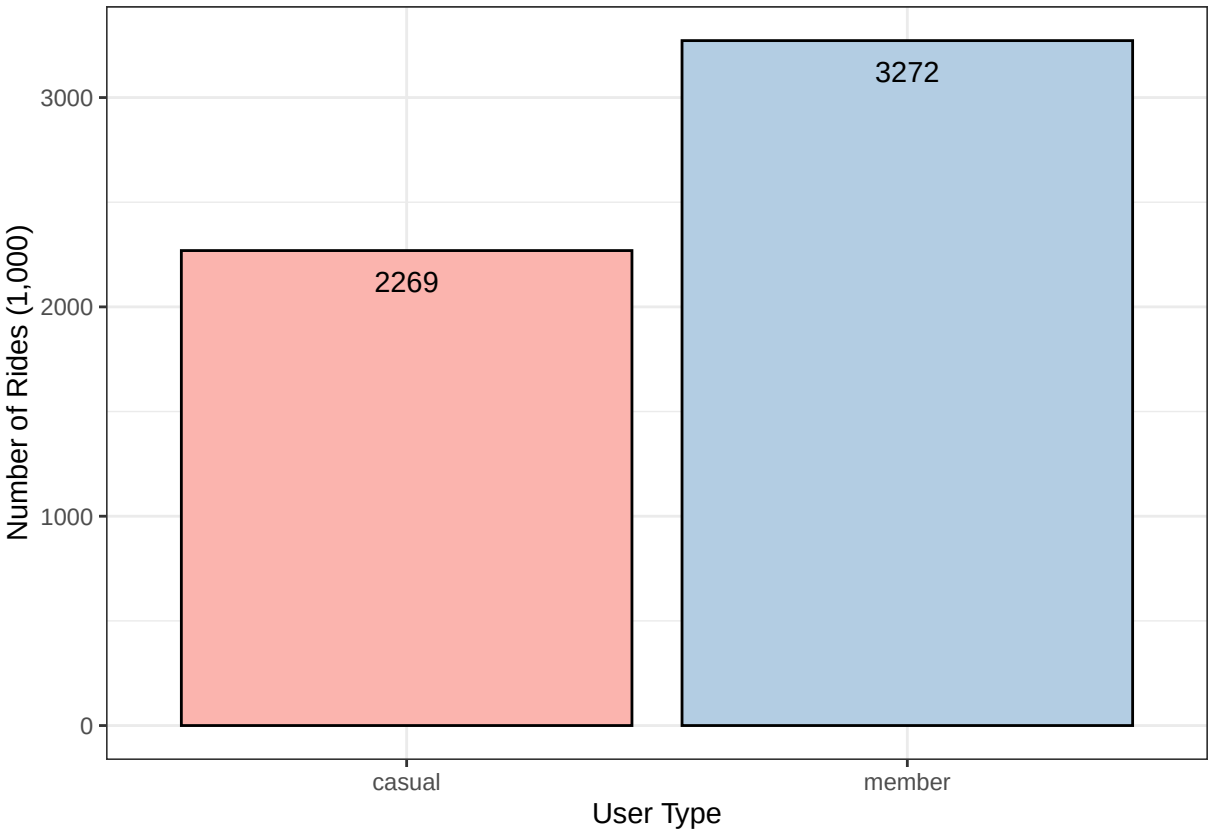
4.2 The Duration of Rides

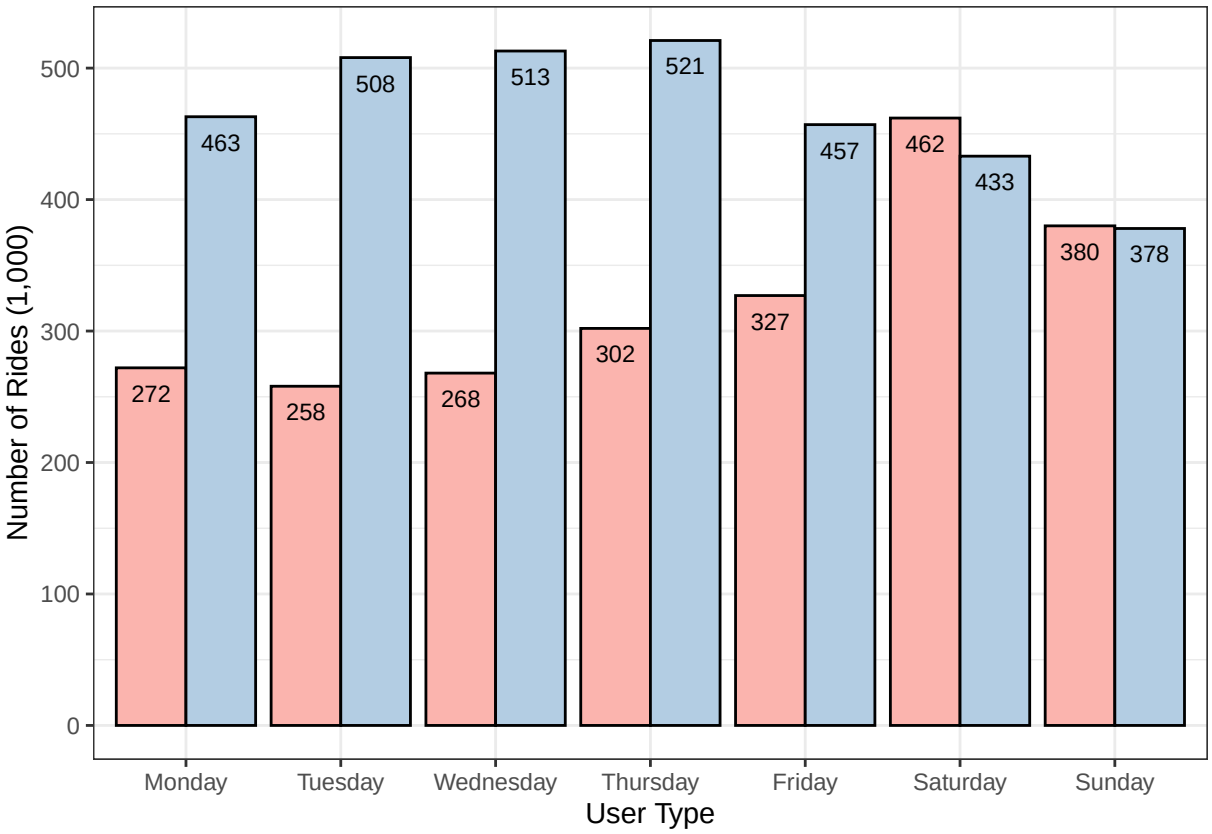


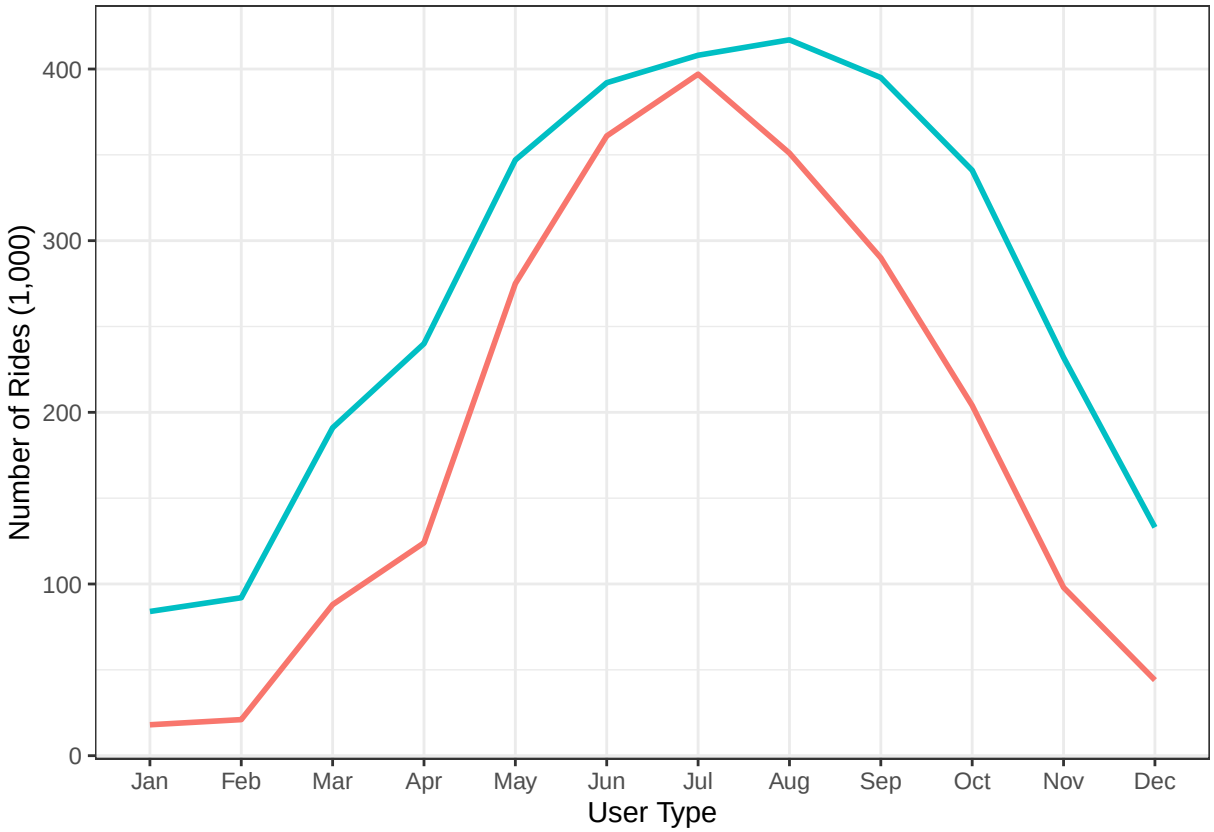

Average Ride Length (mins)
0
10
20
22.79
12.23
19.98
12.04
19.23
12.07
19.9
12.25
20.92
12.48
25.02
14.11
25.51
13.99
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
User Type


Average Ride Length (mins)
10
15
20
25
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
User Type

4.3 The Number of Rides






400
300
200
100
0
Number of Rides (1,000)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
User Type

5 Conclusion and Recommendations

To be written