
# Cyclistic Bike-Sharing Data Analysis

Thanakorn Thanakraikiti (Nun)

August 12, 2023

# Contents

# 1 Intoduction

This section provides an overview of Cyclistic's company detail and its current problem.

## 1.1 Problem Statement

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'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.

## 1.2 Purpose

The purpose of the report is to analyze Cyclistic's historical bike trip data from January 2022 to December 2022 in order to identify how annual members and casual riders use Cyclistic bikes differently. It also aims to extract insights and develop the most appropriate marketing strategies that appeal to casual riders and encourage them to subscribe to annual memberships.

# 2 Dataset

This section provides a comprehensive overview of the dataset used in the analysis. In order to understand the insights, it is essential to understand the data that has contributed to shaping the findings. This section also aims to ensure that the analysis is based on reliable information, making our conclusions strong and trustworthy.

## 2.1 Dataset Overview

The primary dataset used in the project is Cyclistic's historical trip data from January 2022 to December 2022. The data is collected by the City of Chicago (a government agency) and published by Cyclistic. It contains the historical trip data since 2013, whic can be found and downloaded in the following details:

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

The data was downloaded and stored in ZIP files on the secure location. After extracting files, there are twelve CSV files, one for each month of 2022, including *202201-divvy-tripdata.csv* through *202212-divvy-tripdata.csv*.

## 2.2 Data Dictionary

Each dataset contains the following fields:

| Column | Data Type | Definition | Possible Value |
|---|---|---|---|
| ride_id | String | Unique ID number for all rides | - |
| rideable_type | String | Type of bikes | 3 |
| started_at | Date/Time | Which date and time the ride started | - |
| ended_at | Date/Time | Which date and time the ride ended | - |
| start_station_id | String | ID number of the stating station | - |
| start_station_name | String | Name of the the stating station | - |
| end_station_id | String | ID number of the ending station | - |
| end_station_name | String | Name of the the ending station | - |
| start_lat | Float | Latitude of the stating location | - |
| start_lng | Float | Longitude of the ending location | - |
| end_lat | Float | Latitude of the stating location | - |
| end_lng | Float | Longitude of the ending location | - |
| member_casual | String | Type of users | 2 |

## 2.3 Data Credibility

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

| Name | Rating | Reason |
|---|---|---|
| Reliable | High | Data includes useful information, such as location, date and user type |
| Original | High | Data comes from the first-party provider, provided by the company |
| Comprehensive | Medium | Data has missing values in some columns |
| Current | High | Data has been updated monthly since 2013 |
| Cited | High | Data contains data process method and license from the company |

Overall, the data is considered a credible source for analysis based on the Reliability, Originality, Comprehensiveness, Current, 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. This process uses the ***R programming language*** and ***RStudio*** integrated development environment (IDE) to complete the following tasks, including loading the 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 Packages

To perform the data cleaning process, it requires the following packages:

- `tidyverse` for importing, wrangling and visualizing the data
- `skimr` for skimming the data
- `janitor` for exploring and cleaning the data
- `scales` for converting from data values to perceptual properties

## 3.2 Importing and Combining the Data

After loading all necessary packages, it is important to import all 12-month data and then merge them into a single data, called ***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

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


The picture also reveals that the table has 13 columns of 3 data types: 7 **character**, 4 **numeric**, and 2 **POSIXct**. 6 columns have missing values, as follows:

- `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 contains many missing values and are not required for this analysis. Finally, the ***trip_data_v2*** data has only 5 columns, including `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 data, called `ride_length`, `day_of_week`, and `month`, in order to facilitate data analysis and visualization.

- The `ride_length` column will store the duration of the trip in minutes
- The `day_of_week` column will store the day of the week on which the trip was taken
- The `month` column will store the month in which the trip was taken

| | |
|---|---|
| ride_id | 98D355D9A9852BE9 |
| started_at | 2022-01-01 00:00:05 |
| ended_at | 2022-01-01 00:01:48 |
| rideable_type | classic_bike |
| member_casual | casual |
| ride_length | 1.716667 |
| day_of_week | Sat |
| month | Jan |

## 3.5 Removing Outliers

The last step is to remove unwanted data out of the table. The `skim_without_charts()` function is used to provides a concise summary of the ***trip_data_v2*** data again. It shows that there are minus number in the minimum values (min) in the `ride_length` column. There are also higher number with 1,439.933 mins in the `ride_length` column. This means that there are outliers in the data. Outliers can distort the results of an analysis, so that 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.


The picture also shows that there are unique values in the `rideable_type` column. After checking, there are `classis_bike`, `electric_bike`, and `docked_bike`. Because there are only `classis_bike` and `electric_bike`, `docked_bike` needed to be remove from the data. The data is completely processed and is ready for the data analysis. The data contains 5,667,717 rows and 13 columns.

# 4 Data Analysis

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.

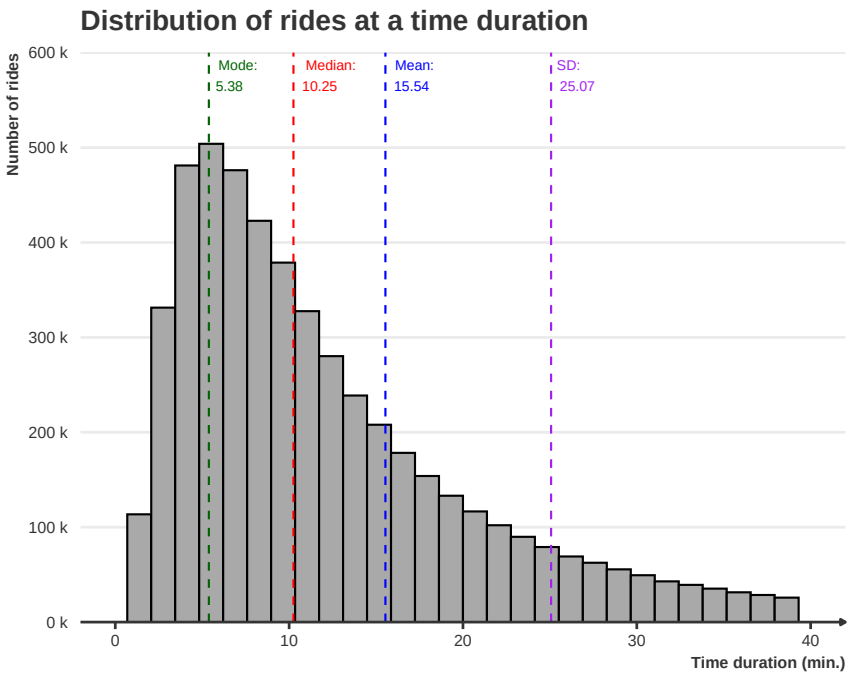
## 4.1 Overall Cyclistic Rides in 2022

As shown in the below table, it presents a statistical summary of time duration for Cyclistic's users, including casual riders and annual members.

Table 5: Descriptive Statistics by Time Duration

| Mean | Min | Mode | Median | Max | SD | Size | Popular_Day |
|---|---|---|---|---|---|---|---|
| 15.53885 | 1 | 5.383333 | 10.25 | 1439.933 | 25.07091 | 5367347 | Sat |

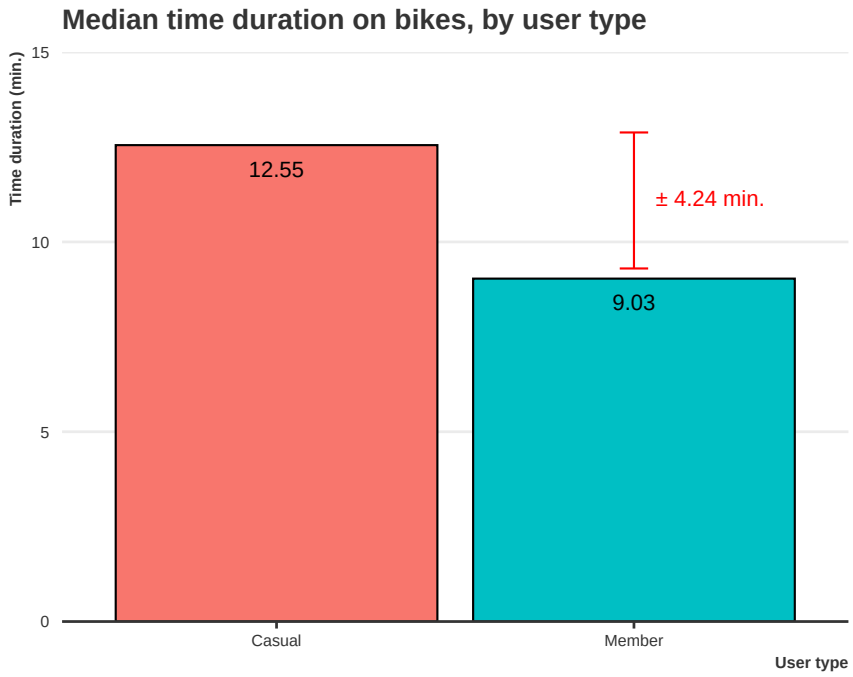
The time duration of Cyclistic's rides can also be visualized by a histogram. The histogram shows that the distribution of ride times is positively skewed, with a high cluster of lower values and a spread-out tail on the right. This means that there are a few rides that are much longer than the majority of rides.

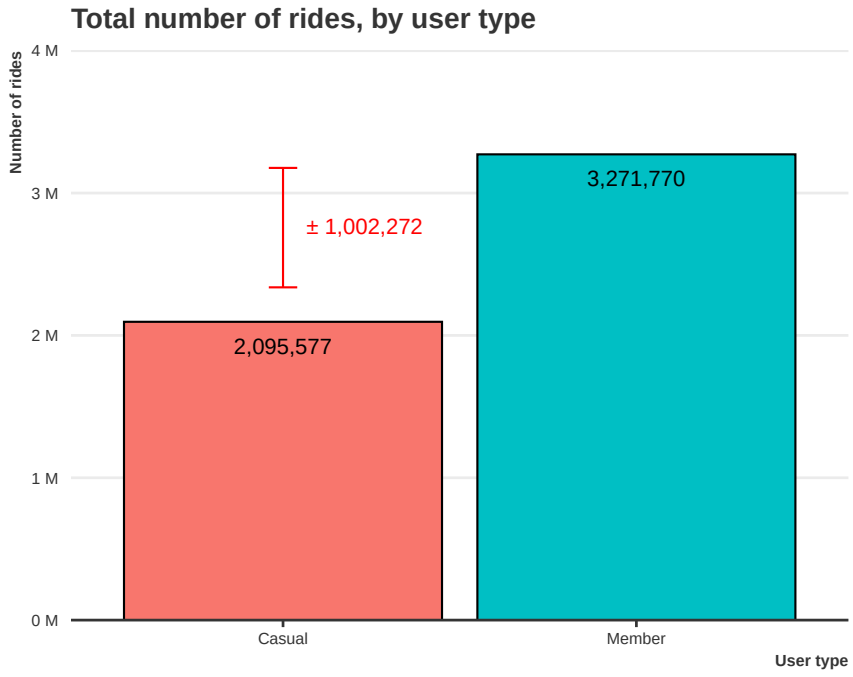



Because of the positive skew, the mean ride time is not a very accurate measure of the typical ride time. The median ride time is a more accurate measure, as it is not as affected by the outliers. So, the median ride time will be chosen to conduct further descriptive analysis.

## 4.2 Casual Riders vs Annual Members

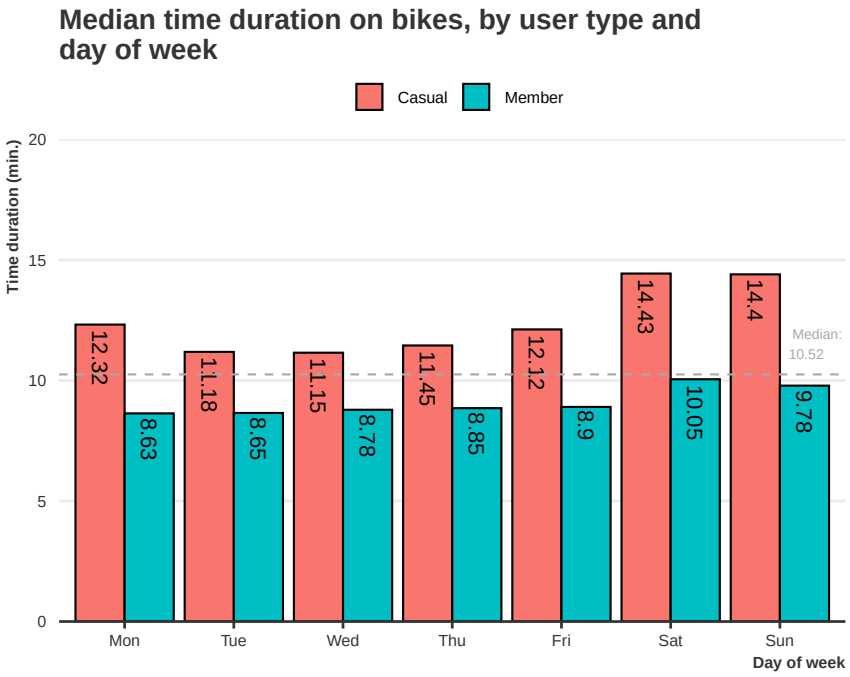
Casual riders appear to ride for longer periods of time than annual members. The median time duration for casual riders is 13.27 minutes, which is higher compared to the median time duration of 9.03 minutes for annual members. However, the data also shows that casual riders took a total of 2,269,498 rides, while annual members took 3,271,770 rides. From these findings, it can be concluded that annual members tend to be more regular users of the bike-sharing service. On the other hand, casual riders are more likely to take longer trips when they do use the service, even though they use it less frequently.

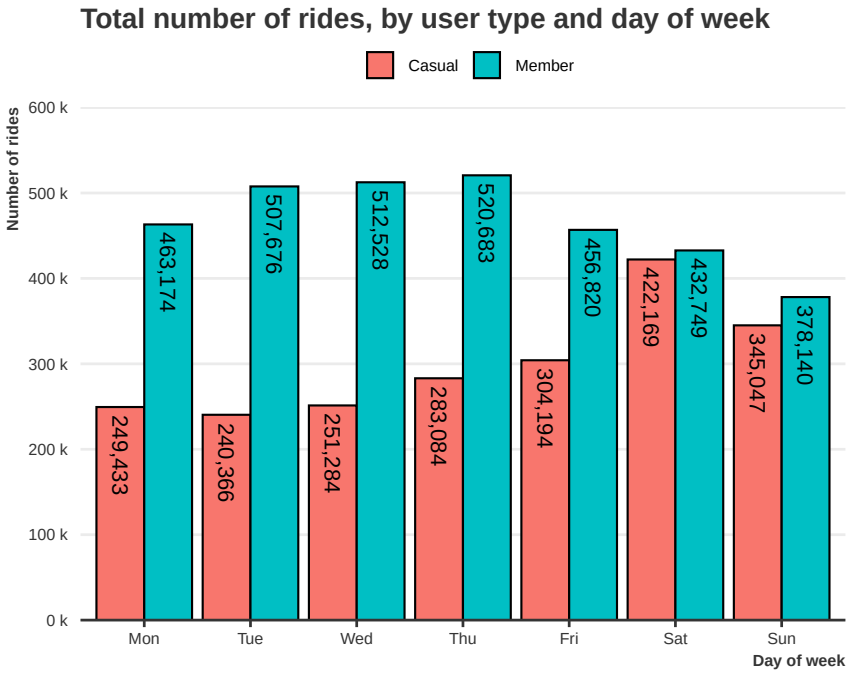


## 4.3 Day of Week

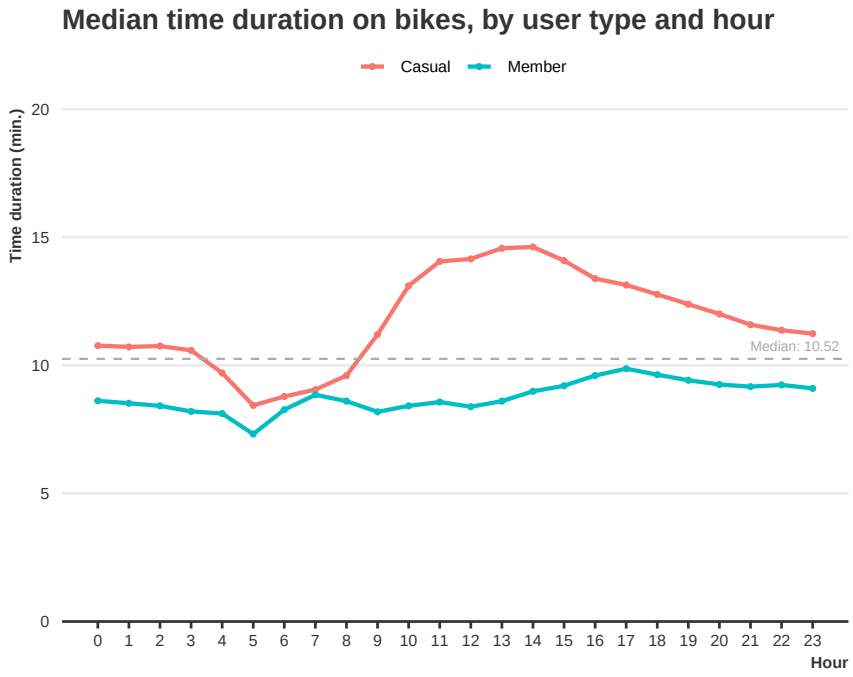
Casual riders tend to take longer bike rides than annual members, especially on the weekend. On average, casual riders spend about 15 minutes longer on bikes than annual members on the weekend, but annual members take time cycling at a consistent level throughout the day. In contrast, annual members use Cyclistic bikes more frequently on the weekday and their usage gradually decreases as the weekend approaches. Casual riders, on the other hand, tend to take more bike rides on weekends. This suggests that the longer ride times of casual riders are highly correlated with their increased usage during the weekends.

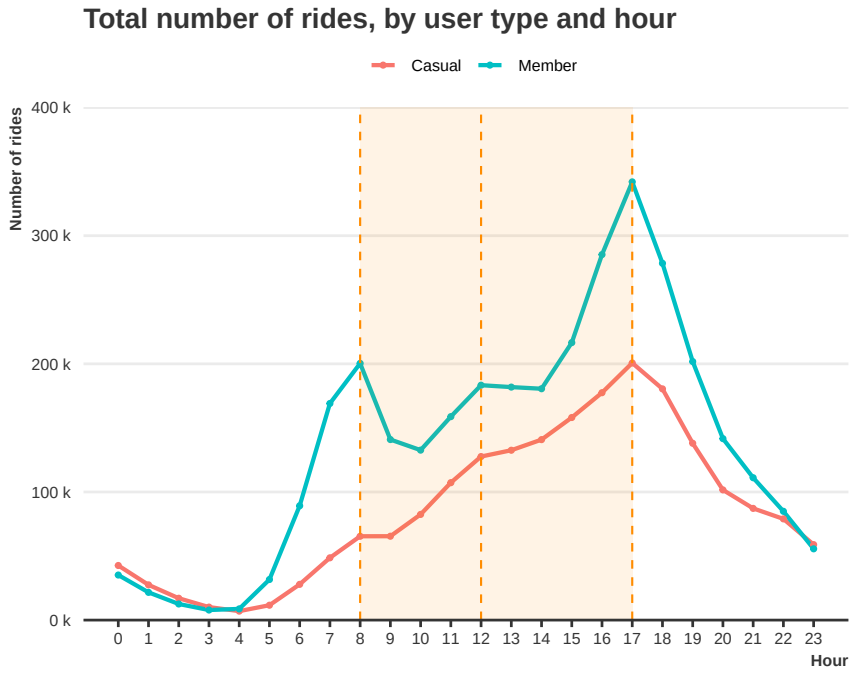


## 4.4 Hour

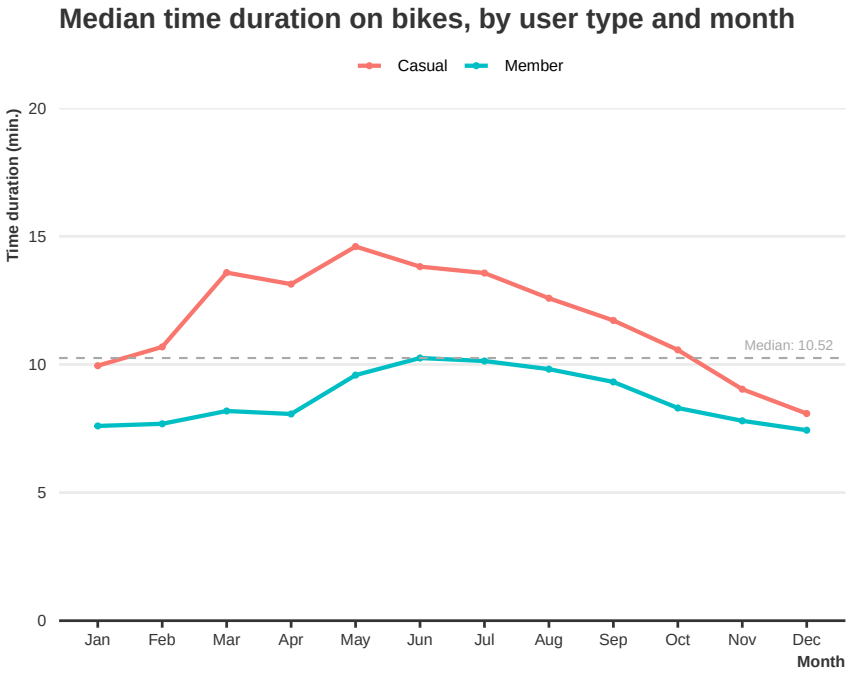
Casual riders are most likely to take longer bike rides after the midday hours, from 1 PM to 3 PM. After that, their usage decreases slightly. Annual members, on the other hand, take time on bikes at a consistent level throughout the day, with no significant spikes or dips in usage. By the number of rides, annual members use Cyclistic bikes at a consistent rate throughout the day, with three peaks in usage: at 8 AM, at 12 PM, and at 5 PM, whereas casual riders start using Cyclistic bikes at 5 AM and gradually increase their usage until they peak at 5 PM.

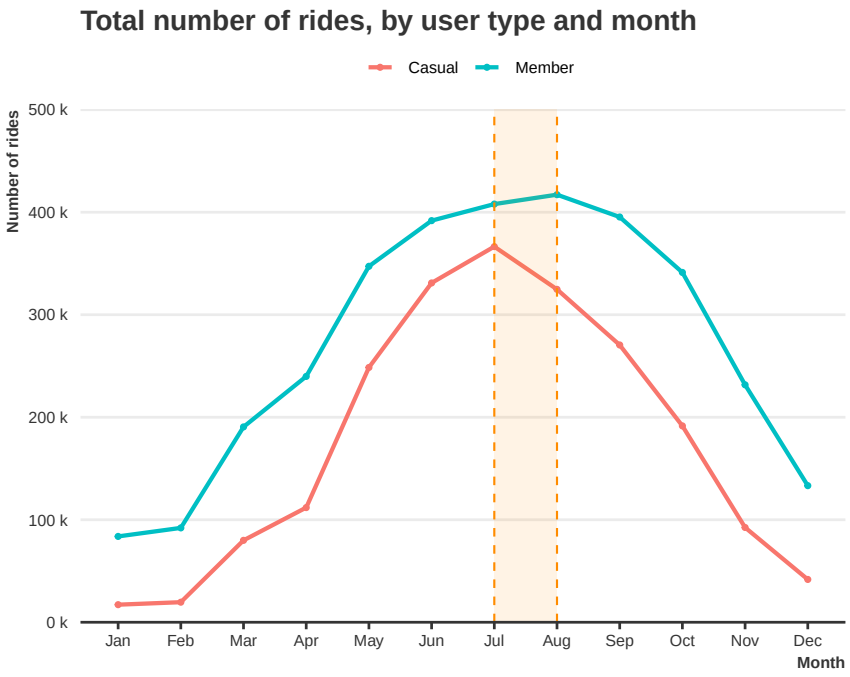


## 4.5 Month

The findings show that the median ride time for both casual and member riders is relatively consistent throughout the year, with the exception of three months: March, April and May. However, casual riders tend to spend more time on bikes than annual members. They also shows that the number of Cyclistic rides increases from month to month, with a peak in July for casual riders and August for annual members. After that, the number of rides suddenly declines. This can be concluded that



# 5 Summary

This section provides concluding thought and recommendations of this

## 5.1 Conclusion

Based on the data collected, it can be concluded that Quis ad elit anim dolor nisi tempor cillum. Occaecat do ex reprehenderit velit ipsum cupidatat officia proident cupidatat do ullamco incididunt. Labore qui labore anim et proident deserunt aliqua. Enim irure aliquip nostrud tempor sit anim ex. Incididunt eu et adipisicing aute minim exercitation aliquip ex dolore deserunt fugiat ullamco quis irure magna. Ut non labore consequat ullamco occaecat exercitation officia cillum culpa fugiat pariatur.

Based on data analysis, :

- 
- 
- 
- 
- 

## 5.2 Recommendations

The report makes the following recommendations:

1. 
2. 
3. 

## 5.3 Limitations

Quis ad elit anim dolor nisi tempor cillum. Occaecat do ex reprehenderit velit ipsum cupidatat officia proident cupidatat do ullamco incididunt. Labore qui labore anim et proident deserunt aliqua. Enim irure aliquip nostrud tempor sit anim ex. Incididunt eu et adipisicing aute minim exercitation aliquip ex dolore deserunt fugiat ullamco quis irure magna. Ut non labore consequat ullamco occaecat exercitation officia cillum culpa fugiat pariatur.