# Case of study 1

‘Cyclistic’

## Introduction

This case of study is the capstone project for the Google Data Analytics Professional Certificate. The scenario presented is that Cyclistic is a bike-share company in Chicago, that gave us the task of analyzing the differences between casual riders and annual members to develop a marketing strategy to convert casual riders into members. Cyclistic's marketing strategy has relied on building general awareness and appealing to broad consumer segments with flexible pricing plans. However, Cyclistic's director of marketing believes that future growth depends on maximizing the number of annual memberships, as annual members are more profitable than casual riders.

## Ask

The first step in every data analysis project is to know what we are looking for. What information and insights is the data going to provide us to solve a particular problem. We need to make ourselves questions oriented to our needs.. In this particular project, the marketing team needs help to answer a specific question that may be the key to generate more profit for the company.

**The question that I will answer is:**

- How do annual members and casual riders use Cyclistic bikes differently?

**In order to do that, I need to make a few more questions:**

- How many rides did members and casual riders do? When is it used the most?
- What day of the week are the bikes used the most (mode)?
- How is the distribution by users through the days of the week?
- What is the most common ride length? (Mode)
- How long is the average ride duration for annual members and casual riders?
- What type of bicycles do users use? Are there any differences of choice between members and casual riders?
- How much time have the bikes been used in total?

## Prepare

### Organization

I used data provided by google Data Analytics Professional Certificate which is hosted on AWS S3[1]. This data is in “.csv” format and it was imported into a postgresql database for further analysis. This file contains information from all bike rides made through the Divvy bike-sharing program of the City of

[1] S3 Bucket: https://divvy-ridedata.s3.amazonaws.com/index.html

Chicago during 2022. These rides are grouped by month in separate files. The data is organized in columns where every row is a different bike ride.

Headers description:

- **ride_id:** string representing a unique ride/ride.
- **rideable_type:** string representing the type of bike used for a particular ride.
- **started_at:** Timestamp with the start date and time of a ride.
- **ended_at:** Timestamp with the end date and time of a ride.
- **start_station_name:** string containing the start station name.
- **start_station_id:** string containing the start station id. They follow no clear pattern.
- **end_station_name:** string containing the start station name.
- **end_station_id:** string containing the start station id.
- **start_lat:** numeric data type representing the starting point latitude.
- **start_lng:** numeric data type representing the starting point longitude.
- **end_lat:** numeric data type representing the ending point latitude.
- **end_lng:** numeric data type representing the ending point longitude.
- **member_casual:** string representing one of two categories, casual or member.

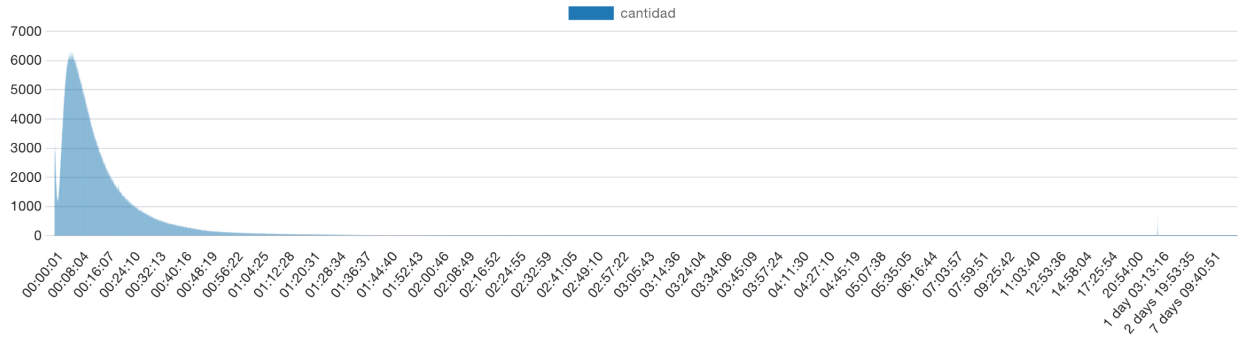
* Code in the appendix

## Data credibility

The dataset I worked with 'ROCCCs', which means it is reliable, original, comprehensive, current, and cited. The data is considered reliable and not biased since it contains the information about every single bike ride that took place in Chicago in 2022 through the Divvy bike sharing service system. We got the data from its original source and it contains the information needed to perform the analysis. It can also be considered current because the files contain information from 2022, which is 3 months before the starting date of this project.

I want to make a clarification regarding the reliability of the data. I realized that there are some outliers in the duration of the rides. For example there are records of rides with a duration of just a few seconds and some rides with duration longer than 1 day , with cases even reaching 28 days.

Graph 1



* Code in the appendix

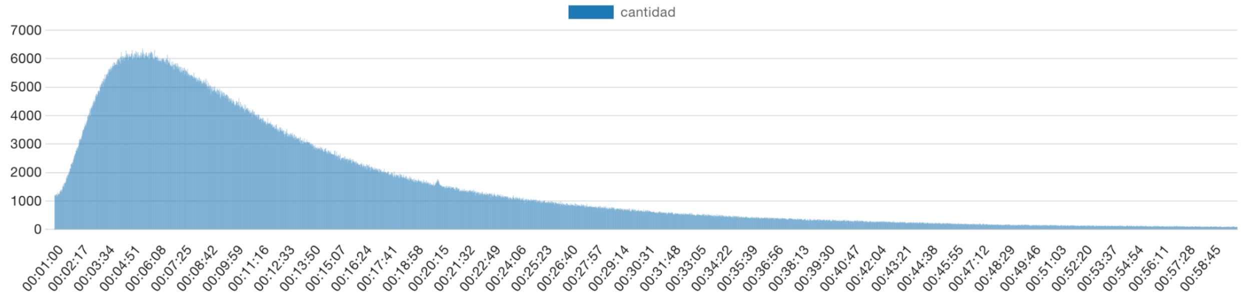
In order to leave outliers out of the analysis I choose only the rides whose duration was greater than 1 minute and 60 minutes. Those rides represent 94.87% of the total rides.

Table 1

| Rides between 1 and 60 min | Total rides | Percentage |
|---|---|---|
| 5.376.260 | 5.667.186 | 94.87% |
| * Code in the appendix | | |

This is the data distribution by ride duration after removing the outliers.

Graph 2



* Code in the appendix

## License

Regarding licensing, privacy, security, and accessibility I want to clarify some aspects. "Bikeshare hereby grants to me a non-exclusive, royalty-free, limited, perpetual license to access, reproduce, analyze, copy, modify, distribute in my product or service and use the data for any lawful purpose."[2] So I am entitled to use the data as long as I do not use the data in any unlawful manner or for any unlawful purpose, or attempt to correlate the data with names, addresses, or other information of customers or Members of Bikeshare. I downloaded the .csv files and stored them in my personal folders. I am not going to share them as attachments of the project, just going to document the findings.

## Data integrity

To verify the dataset's integrity I run some queries to check some values. First I checked that the ride_id field is not null, it's not duplicated and all of them have the same length and format. Also gathered information of the amount of total missing cells in the entire dataset.

[2] Divvy Data License Agreement: https://ride.divvybikes.com/data-license-agreement

```sql
-- Ride ID not null check
SELECT 'ride_id' AS variable, COUNT(*)
FROM rides
WHERE ride_id IS NULL;

-- No duplicates check
SELECT ride_id, COUNT(ride_id)
FROM rides
GROUP BY ride_id
HAVING COUNT(ride_id) > 1;

-- Ride ID length check
SELECT LENGTH(ride_id) AS ride_id_length, COUNT(*) AS count_of_length
FROM rides
GROUP BY ride_id_length
ORDER BY ride_id_length;

-- Missing cells check (created the VIEW table ‘table_columns_missing’ for that purpose. See appendix
for the code)
SELECT SUM(null_cells.count) AS missing_cells
FROM table_columns_missing AS null_cells;
```

Then I did a dataset overview where I checked how many missing cells were in the dataset. There is only a 4.7% of missing cells which is not significant for the analysis.

Dataset overview:
Table 2

| | |
|---|---:|
| **Number of variables** | 13 |
| **Number of observations** | 5.667.717 |
| **Total cells (variables x observations)** | 73.680.321 |
| **Missing cells** | 3.463.328 |
| **Missing cells (%)** | 4.7% |
| **Duplicate rows** | 0 |
| **Duplicate rows (%)** | 0% |

Also analyzed how the missing cells were distributed in the dataset. To do so I ran the following queries. First I generated a VIEW Table with the amount of null values for each variable and then I calculated the percentage with a CROSS JOIN.

I realized that there are 14.7% of missing Start Station values and 15.75% missing End Station values.

Table 3

| Variable | Null count | % of null |
|---|---:|---:|
| ride_id | 0 | 0% |
| rideable_type | 0 | 0% |
| started_at | 0 | 0% |
| ended_at | 0 | 0% |
| start_station_name | 833064 | 14.7% |
| start_station_id | 833064 | 14.7% |
| end_station_name | 892742 | 15.75% |
| end_station_id | 892742 | 15.75% |
| start_lat | 0 | 0% |
| start_lng | 0 | 0% |
| end_lat | 5858 | 0.10% |
| end_lng | 5858 | 0.10% |
| member_casual | 0 | 0% |
| * Code in the appendix | | |

So I wrote a query to understand how many records are affected by either having no information for the start station or the end station.

As seen in the table below, there are 22.91% of records affected by the lack of information either in the starting station field or the ending station field. But considering that more than 4 million records are complete I can say that the dataset is good to be used for the analysis.

Table 4

| Variables | Null count | % of null |
|---|---|---|
| NO start OR end station | 1.298.357 | 22.91% |
| * Code in the appendix | | |

Then I decided to run 3 checks to validate the integrity of the data contained in the date fields.

1. Verify that all records have date and time.
2. Verify if all rides in the database are from the year 2022.
3. Verify if the start date and time are before the end date and time in every case.

1. I previously checked that there are no null values in the date fields, which means that every record in the original .csv dataset has a date. Date records with only year, month, and day, and the time set to 00:00:00 may indicate that the time was not uploaded correctly since that is the default value that the database assigns if a value is missing.

In order to determine if that is the case, I ran the following query.

```
    SELECT ride_id, started_at, ended_at
        FROM rides
        WHERE (EXTRACT(HOUR FROM started_at) = 0
              AND EXTRACT(MINUTE FROM started_at) = 0
              AND EXTRACT(SECOND FROM started_at) = 0)
                  OR (EXTRACT(HOUR FROM ended_at) = 0
              AND EXTRACT(MINUTE FROM ended_at) = 0
              AND EXTRACT(SECOND FROM ended_at) = 0);
```

Only 56 records met the criteria, so I was able to check every case manually to verify that the data makes sense, which means that all rides are valid for the analisis.

2. At first I ran a query to retrieve every ride that took place between 01/01/2022 as the starting date and 31/12/2023 as the ending date which brought me back a result of 99.99%. So I did some research in the dataset and realized that some rides started in 2022 but ended in 2023. So I updated the query and 100% of the rides match the criteria and verify that every record is good for analysis.

```
SELECT (COUNT(*) * 100.0 / (SELECT COUNT(*) FROM rides)) AS percentage
FROM rides
WHERE started_at IS NOT NULL AND ended_at IS NOT NULL
  AND DATE(started_at) BETWEEN '2022-01-01'::DATE AND '2022-12-31'::DATE
  AND DATE(ended_at) BETWEEN '2022-01-01'::DATE AND '2023-01-01'::DATE;
```

3. For the dates also to be valid the starting date can not be after the ending date. In 99.99% of the cases it is so. There were 531 records where the end date is earlier than the starting date. Because the amount of rides is not significant we will leave those records outside from the analysis.

```
SELECT (COUNT(*) * 100.0 / (SELECT COUNT(*) FROM rides)) AS percentage
FROM rides
WHERE started_at IS NOT NULL AND ended_at IS NOT NULL
  AND started_at < ended_at;
```

Cases were the dates are not consistent:

```
SELECT COUNT(*)
FROM rides
WHERE started_at IS NOT NULL AND ended_at IS NOT NULL
  AND started_at >= ended_at;
```

### Conclusion

Without diving further in the analysis it seems that the dataset doesn’t have major issues and would be enough to answer the questions I have.

## Process

To process the data I chose SQL and used PGAdmin as the query tool. Due to the huge amount of observation I am dealing with I thought it would be a better tool than excel. After I ensured myself about data’s integrity I started cleaning the data set to start later on with the proper analysis and answering the questions I have.

I deleted from the dataset the 531 records where the starting date was after the ending date. For that operation I used the following query:

```
DELETE FROM rides
WHERE started_at IS NOT NULL
  AND ended_at IS NOT NULL
  AND started_at >= ended_at;
```

I also deleted the outliers. Those are the rides whose duration were less than 1 minute or greater than 60 minutes. 290.926 records were deleted.

```
DELETE FROM rides
WHERE (EXTRACT(epoch FROM ended_at - started_at) / 60.0) < 1.0
   OR (EXTRACT(epoch FROM ended_at - started_at) / 60.0) > 60.0;
```

For this data set there are no more steps required to make it ready for analysis. In the preparation phase I made sure that the data is good to be worked on.

Other actions I could have to do if needed were:

- Removing or merging duplicate records
- Replacing null/missing values with a default value
- Removing irrelevant records
- Trimming whitespace from text fields
- Removing leading or trailing characters from text fields
- Standardizing date formats
- Converting data types
- Normalizing text fields
- Splitting text into separate columns

As part of the process phase I also added 2 columns that will be very useful for the analysis phase.

The first column I added is the duration of the ride in minutes for every single ride, which I get by subtracting the starting time to the ending time.

```
ALTER TABLE rides
ADD COLUMN duration_ride INTEGER GENERATED ALWAYS AS (EXTRACT(EPOCH FROM (ended_at - started_at))/60)
STORED;
```

The second column added is the season in which the bike was hired for every particular ride. This will help later to understand the behavior of the users in relation to the time of the year.

```
ALTER TABLE rides ADD COLUMN season text;

UPDATE rides SET season =
    CASE
        WHEN EXTRACT(MONTH FROM started_at) IN (12, 1, 2) THEN 'Winter'
          WHEN EXTRACT(MONTH FROM started_at) IN (3, 4, 5) THEN 'Spring'
          WHEN EXTRACT(MONTH FROM started_at) IN (6, 7, 8) THEN 'Summer'
          WHEN EXTRACT(MONTH FROM started_at) IN (9, 10, 11) THEN 'Fall'
    END;
```

# Analyze

In this phase I started to run queries and perform calculations to understand the data better and get the insights. That way I can identify trends and relationships between the data. The key tasks in this phase is to start organizing, formatting and aggregating the data to answer the questions I asked myself and that are relevant for the stakeholders.

Once I ran the queries and got the information I used Google Spreadsheets to format the tables and visualize the data. All queries for the tables and graphs are in the appendix of this document.

As I mentioned at the beginning of this document, the main question I want to answer is: *How do annual members and casual riders use Cyclistic bikes differently?*

So to come to a conclusion is needed to get the answer of more specific questions to understand the behavior of the two different user types. Those questions are the following:

1. **How many rides did members and casual riders do? When is it used the most?**

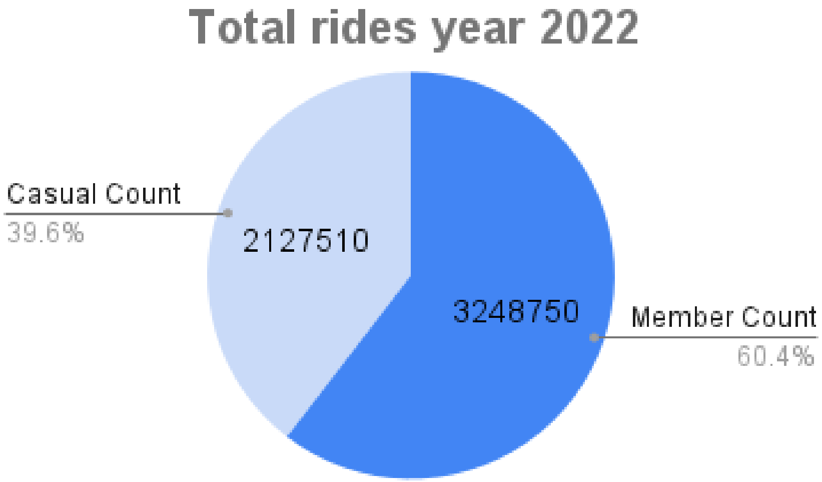
<u>Annually</u>

The first thing we can notice looking at the data is that through the year members use the service more in comparison to casual users. In 2022 members were responsible for 60% of all the rides done, as shown in table 5 and graph 3.

Table 5

| Year | Member Count | Member Percentage | Casual Count | Casual Percentage | Total |
|---|---|---|---|---|---|
| **2022** | 3248750 | 60.43 | 2127510 | 39.57 | 5376260 |

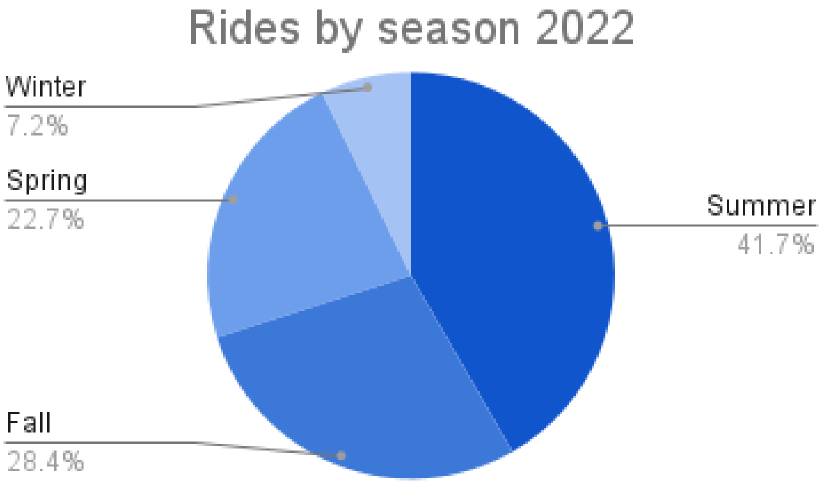
Graph 3



Seasonal

Carrying on with the analysis it was worth it to understand better the trends by making the analysis more specific and see if the seasons play a role in the distribution of the rides. As expected, in hotter seasons the use is higher. There is a correlation between the use and the seasons, summer being the season with more rides (41.7%), followed by fall (28.4%) and spring (22.7%), and winter being the quieter time of the year (7.2%).

Graph 4



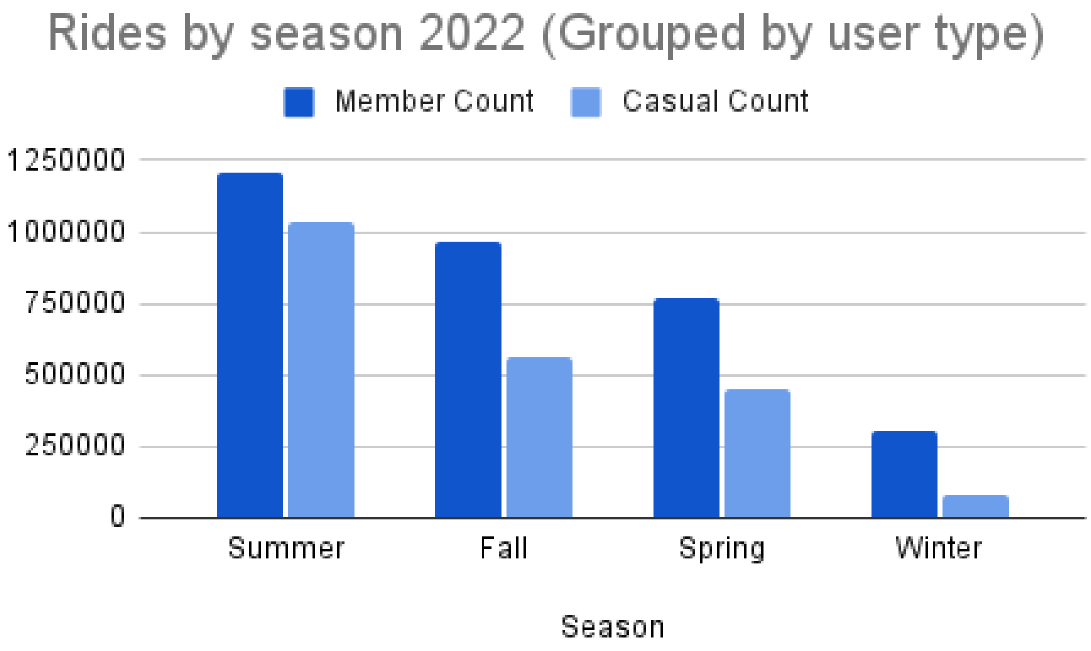
The interesting behavior we can see is that the gap in the usage between member and casual riders increases the colder it gets. For the months of summer members represent 53.85% of the total number of rides of the season whereas in winter it increases representing almost 80% of them. Most likely members see the service as an alternative way of transportation and for casual riders is a recreational activity.

Table 6

| Season | Total rides | Member Count | Member Percentage | Casual Count | Casual Percentage |
|---|---|---|---|---|---|
| **Summer** | 2241447 | 1207104 | 53.85 | 1034343 | 46.15 |
| **Fall** | 1526129 | 961909 | 63.03 | 564220 | 36.97 |
| **Spring** | 1221161 | 772415 | 63.25 | 448746 | 36.75 |
| **Winter** | 387523 | 307322 | 79.30 | 80201 | 20.70 |

In the graph below we can see clearly how the gap in uses increases depending on the seasons of the year. For casual riders we see how it decreases drastically for other seasons compared to summer.

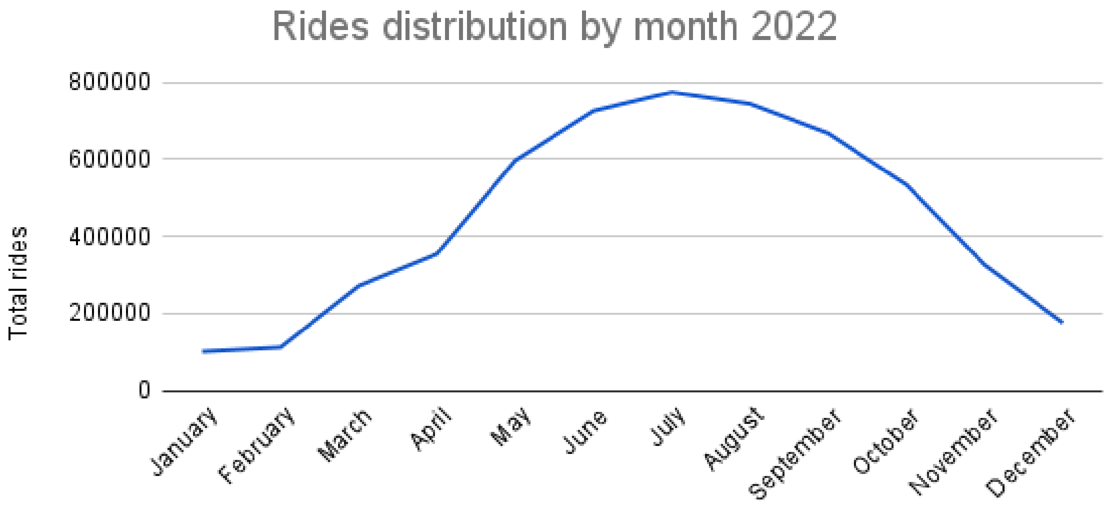
Graph 5



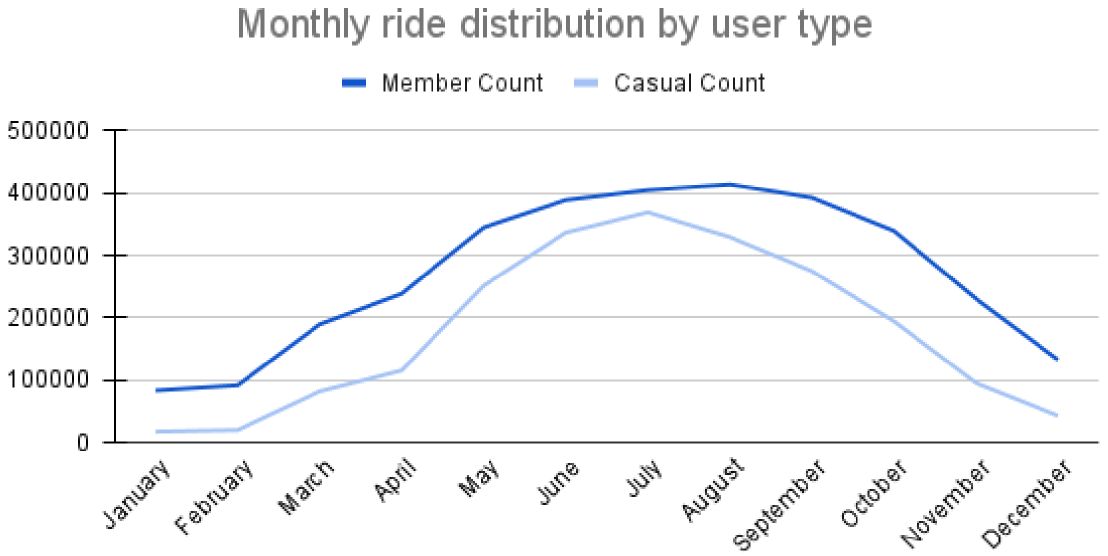
<u>Monthly</u>

I decided to dig deeper and see the behavior for months. Maybe in the hottest month of the year casual rides are higher than the member ones. But it is not like that. The pattern is the same through the whole year, with more use of the bike service in the summer season peaking in July and member use being always above casual use as can be seen in graphs 6 and 7.

Graph 6



Graph 7



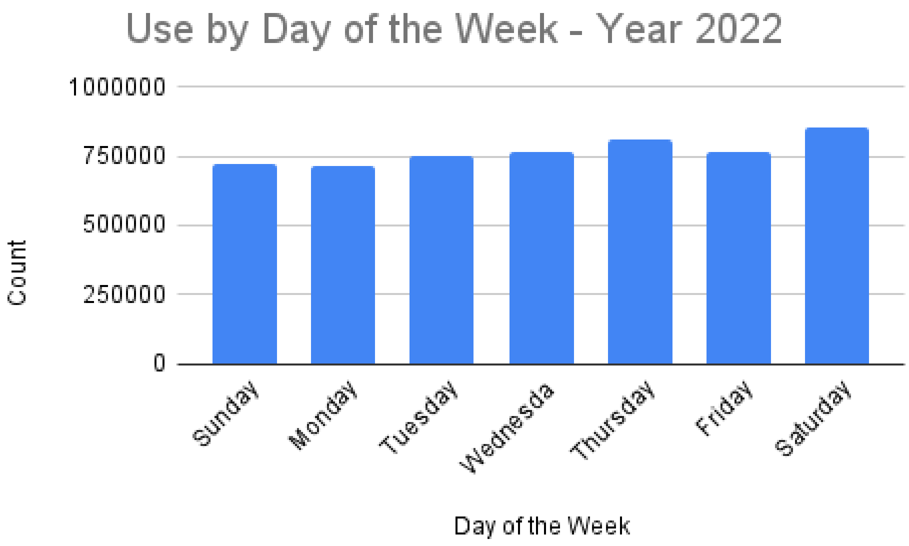
The difference we can identify between members and casual riders is that the curve of use for members is smoother. The use grows gradually peaking in August and decreases gradually again. For casual riders the peak builds steeper in July, being the use very low in the winter months.

2. **What day of the week are the bikes used the most (mode)?**

Annually

Regarding the use over the year in the different days of the week there is a general tendency on Saturdays being the day with more rides followed by Thursday and Friday.

Graph 8



<u>Seasonal</u>

But if we analyze the amount of rides per day of the week considering the different seasons we can appreciate that the mode changes. In winter the bikes are used the most during the week with quite an even distribution use, being Thursday on the top. On the other hand in the other seasons Saturdays are the main choice.

Graph 9

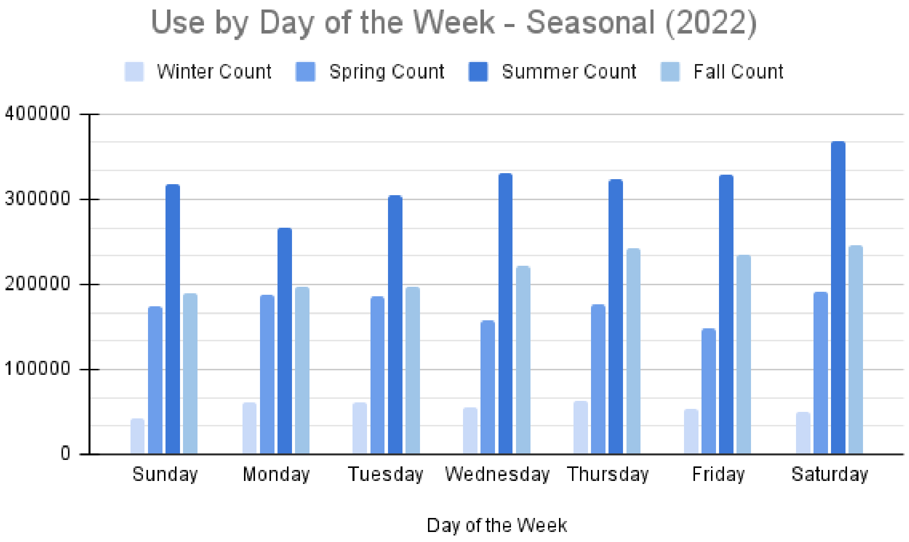


Table 7

| Day of the Week | Winter Count | Day of the Week | Spring Count | Day of the Week | Summer Count | Day of the Week | Fall Count |
|---|---|---|---|---|---|---|---|
| Thursday | 63252 | Saturday | 191241 | Saturday | 369163 | Saturday | 245139 |
| Tuesday | 61404 | Monday | 188265 | Wednesday | 331720 | Thursday | 241867 |
| Monday | 60360 | Tuesday | 186408 | Friday | 327954 | Friday | 234176 |
| Wednesday | 56297 | Thursday | 176585 | Thursday | 323546 | Wednesday | 221451 |
| Friday | 53389 | Sunday | 173952 | Sunday | 317284 | Monday | 197665 |
| Saturday | 49804 | Wednesday | 156453 | Tuesday | 304867 | Tuesday | 196815 |
| Sunday | 43017 | Friday | 148257 | Monday | 266913 | Sunday | 189016 |

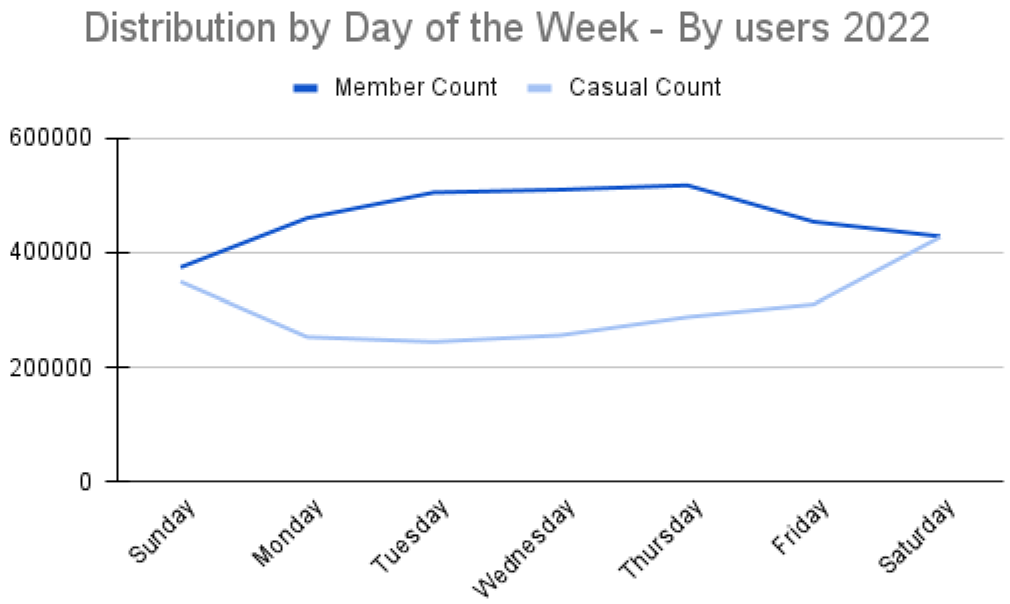
### 3. How is the distribution by users through the days of the week?

Because I want to understand how member and casual users' behavior differs from each other, it makes sense to analyze que ride distribution by day of the week considering the user type.

Annually

The first thing that we can appreciate quickly looking at graph 10 is that members use the service mainly during the week and casual users over the weekend. For casual users we can see how it peaks on saturday and sunday, where for member drops.

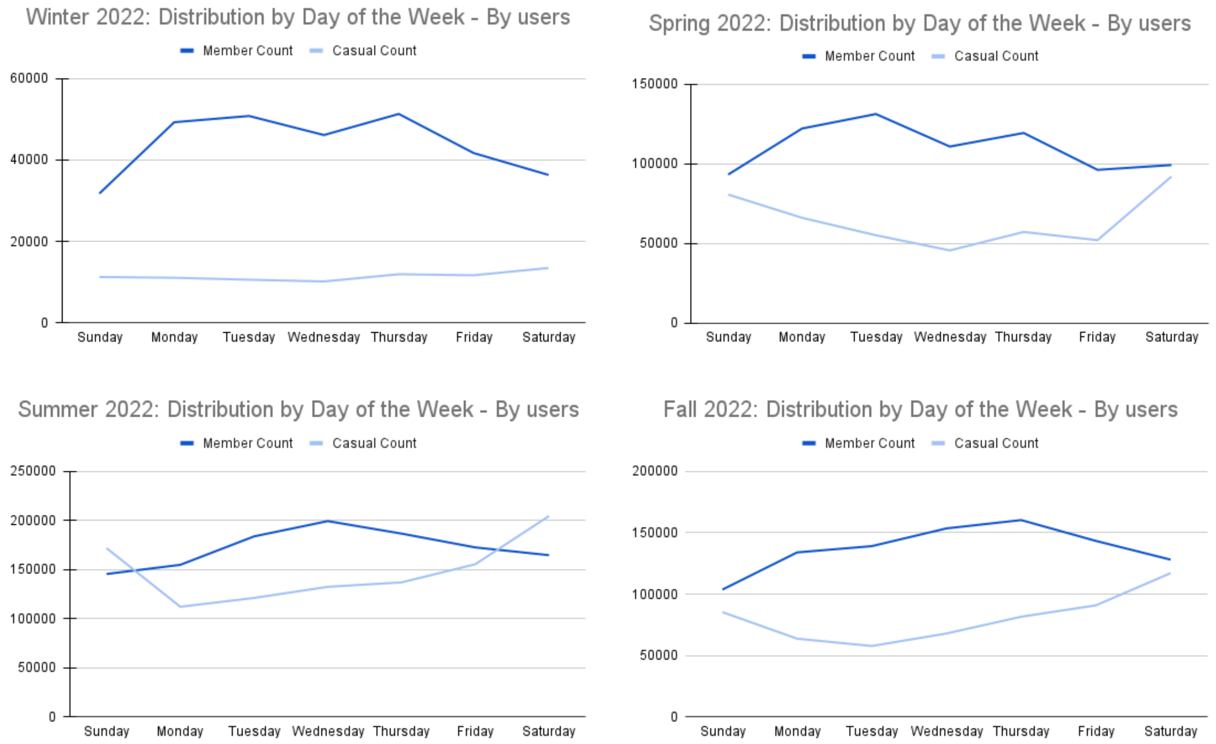
Graph 10



Seasonal

If we look at the amount of rides by day of the week separated by seasons we get interesting insights of the behavior of the users.

Graphs 11-14



In winter the amount of casual rides is pretty low and steady through the week in comparison with members. We can clearly see that members use the bikes mainly during the week which gives us the hint that it is mainly used for transportation purposes most likely to commute to work.

In spring we can start seeing the tendency for casual riders to prefer weekends. It is clearly a recreational use of the service. For members most of the use is still during the week.

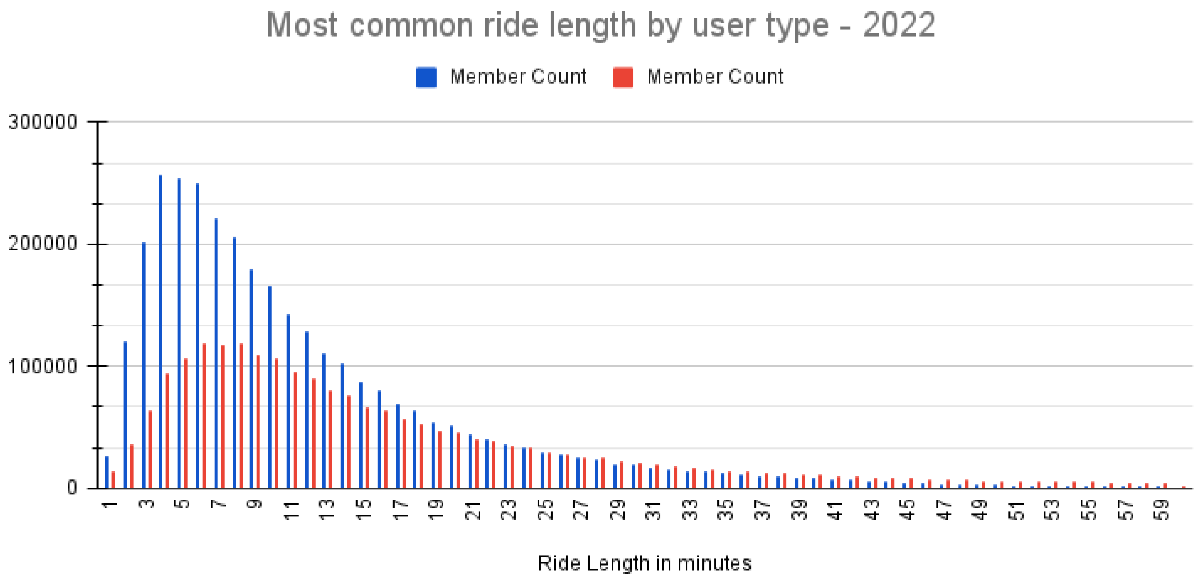
In summer the casual use on the weekends intensifies and it is the only time where there are more casual rides than member rides. Members' tendency is still the same.

In fall the amount of rides for both user types drops in general but maintaining the tendency seen in spring.

**4. What is the most common ride length? (Mode)**

<u>Annually</u>

Graph 15



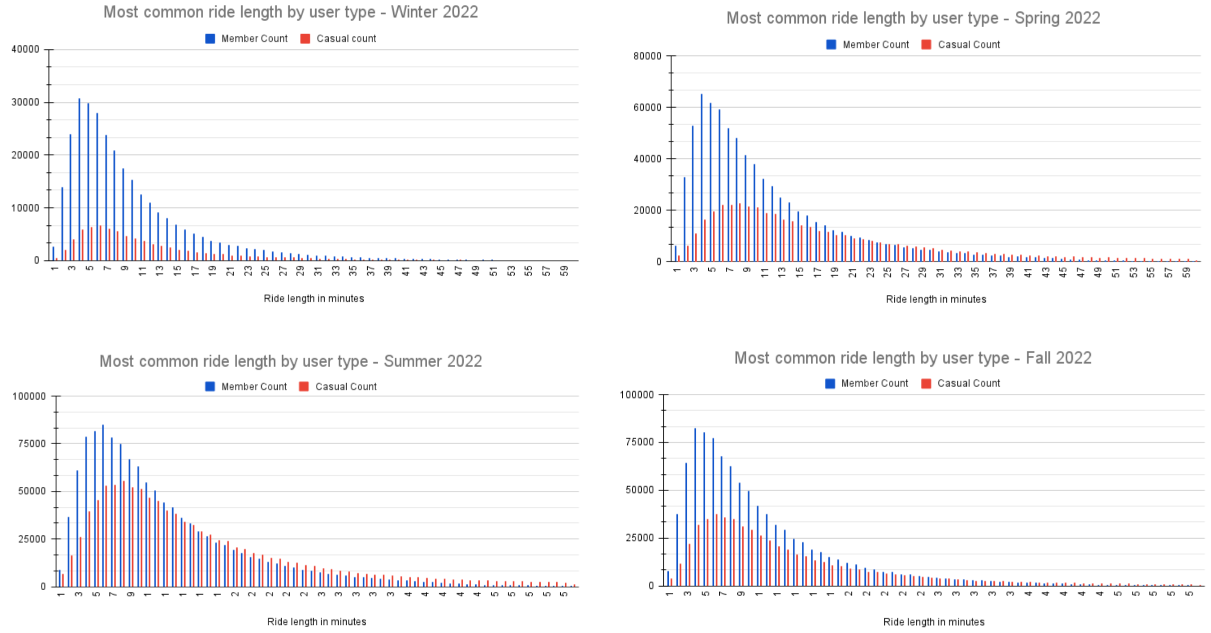
Considering the full year the most common ride length for members is 4 minutes whereas for casual riders is double the time with 8 minutes. We see as well that for rides lasting 25 minutes or more the amount of rides for casual riders is greater than members.

<u>Seasonal</u>

Analyzing the most common rides by season we can see that casual riders mode value is always higher. Even in winter where casual riders barely use the service compared to members.

For members, most of the ride duration is shorter than 10 minutes. We see in the graphs how it peaks for shorter rides and we see a great decrease in the amount of rides. For casual riders that peak is not as pronounced and the decrease in the amount of rides for longer rides is gradual.

Graphs 16-19



Focusing on each season individually we realized that in winter members most common ride length is 4 minutes and for casual riders is 6. For every single ride length member use the service more.

In spring, members' most common ride is 4 minutes but this time casual riders double the mode being for them 8 minutes. For rides of a duration of 25 minutes or longer there are more casual riders using the service.

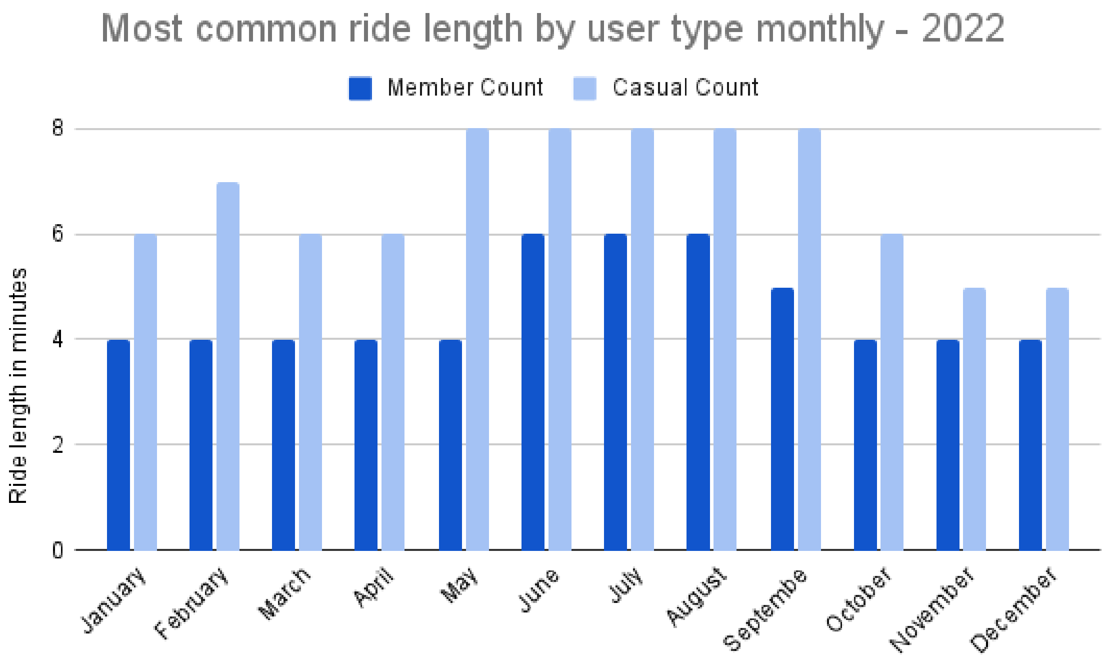
Summer is the main season where members and casual riders use the service the most. Is the time of the year where casual riders get close to members in terms of amount of rides. For members the most common ride length is 6 minutes and for casual riders is 8. And for rides lasting 17 minutes or more there are more casual riders using the service. Here it becomes clear that casual riders usage is recreational whereas member use keeps being mainly for transportation purposes throughout the year.

In fall the casual riders use decreases almost half compared to summer and for members it stays more consistent. The most common ride length for members is 4 minutes and for casual riders is 6.  For rides of a duration of 37 minutes or longer there are more casual riders using the service.

Monthly

We can see that monthly the most common ride is always longer for casual riders at an average of 2 minutes.

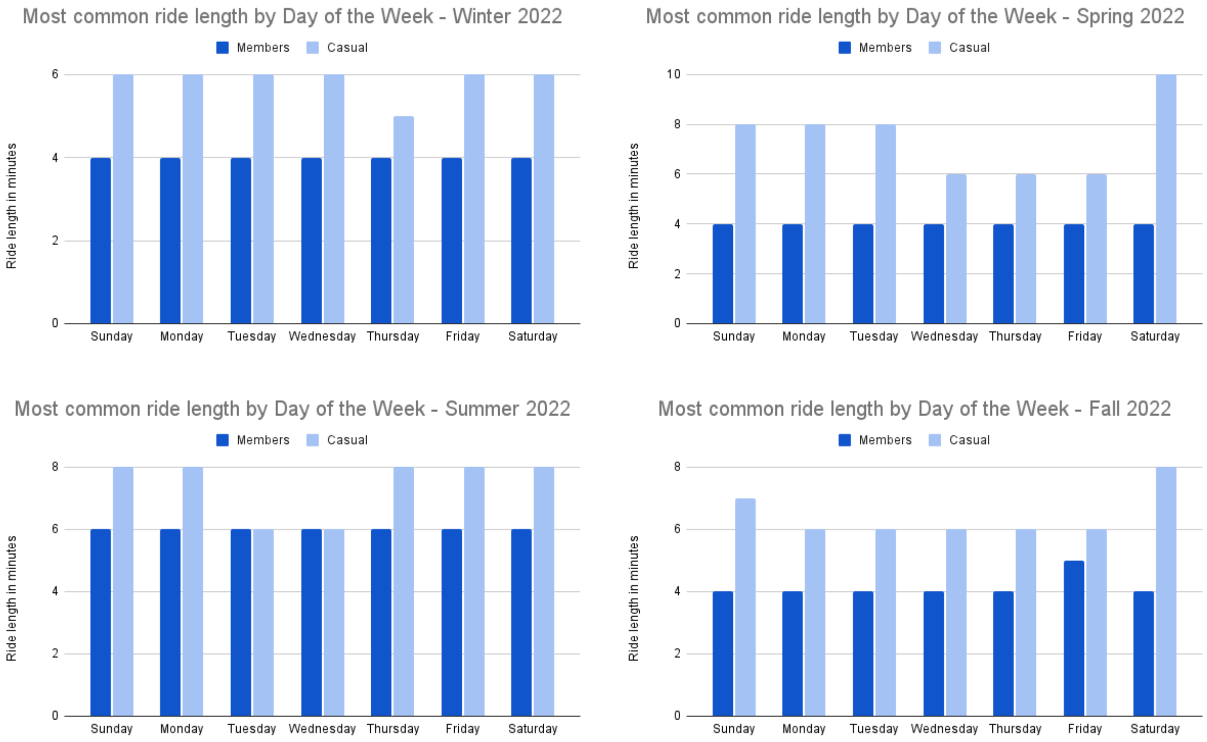
Graph 20



Day of the week

If we focus the analysis by day of the week by season the tendency is the same as monthly, casual users most common daily ride length is always longer except for summer. In summer, although casual riders' most common ride is longer for most days, members' ride length mode raises to 6 for every day of the week matching casual riders choice for Tuesdays and Wednesdays.

Graphs 21-24



## 5. How long is the average ride duration for annual members and casual riders?

The mode in the previous section shows us the most common ride length by the different user types, but what is the average ride length depending on the user type? The main take away from this analysis is that casual riders' average ride is always longer than members no matter the timeframe we choose to perform the analysis (annually, monthly, seasonal or by day of the week).

<u>Annually</u>

If we calculate the average ride duration throughout the year, the rides of casual users are on average 4.3 minutes longer.

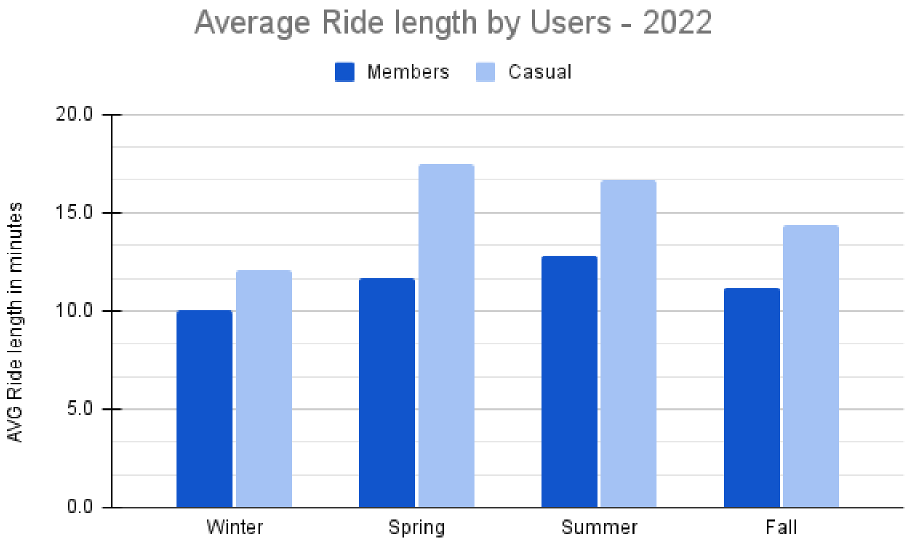
Table 8

| User Type | AVG Ride length |
|---|---|
| **Casual** | 16.1 |
| **Members** | 11.8 |

Seasonal

If we take under consideration the seasons, the biggest difference we find is in spring where casual rides are on average 5.8 minutes longer, followed by summer with 3.9, fall with 3.2 and winter with 2.1 minutes.

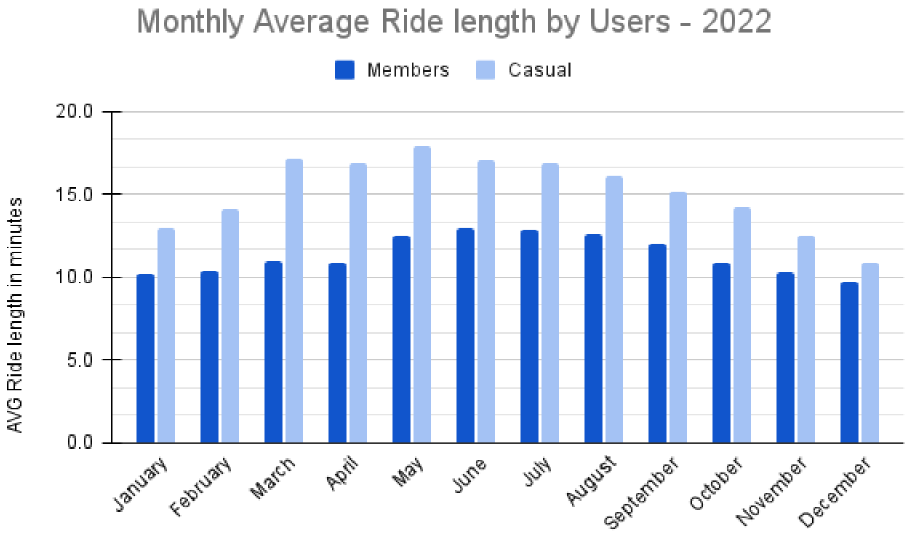
Graph 25



Monthly

The greatest difference is spotted in March with 6.2 minutes longer rides on average for casual users, decreasing progressively as the month passes by. But in terms of the average longest ride for both user types we find it in May for casual users with 17.9 minutes and in June for members with an average of 13 minutes per ride.

Graph 26

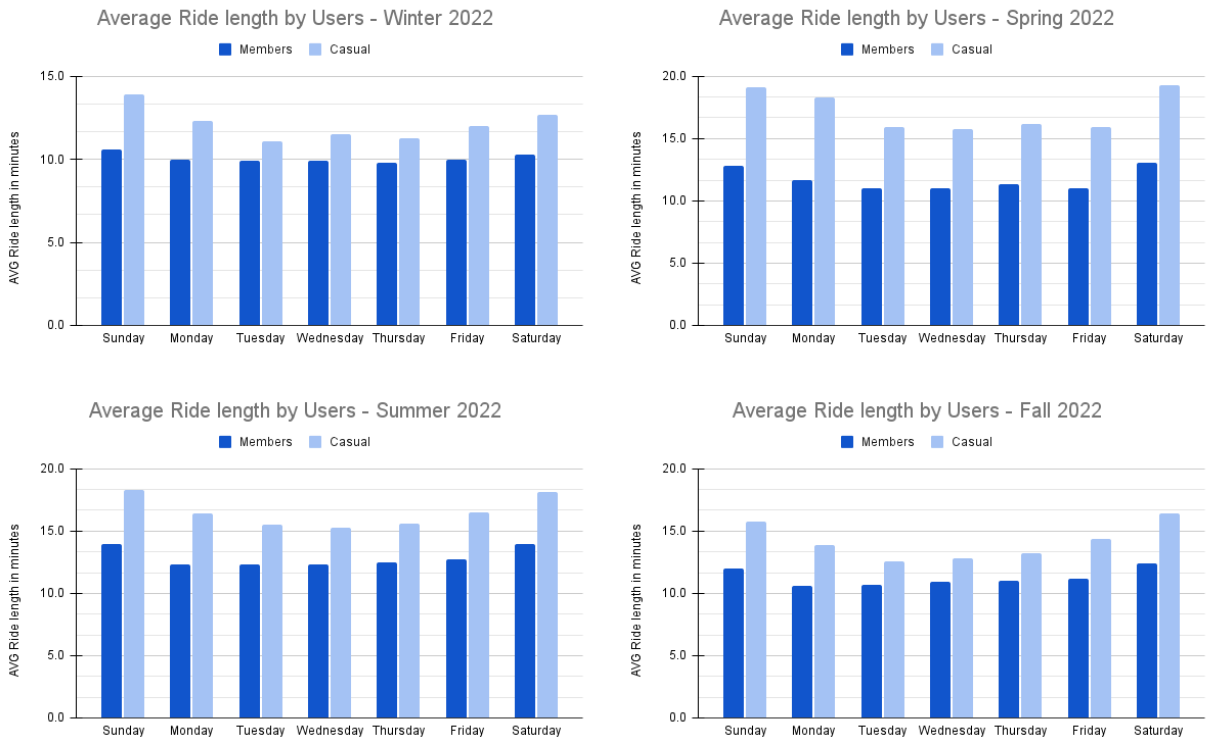


<u>Day of the week</u>

Considering the days of the week, the average longest ride for casual users, we find it in spring on Saturdays and Sundays with 19.3 and 19.1 minutes respectively.

For members the average longest ride is also on Saturdays and Sundays with 14 minutes each day, but in summer instead of spring.

Graphs 27-30



In winter, members' average ride is quite steady through the whole week rounding about the 10 minute average ride while for casual riders we see slightly longer rides on the weekends.

In spring, members' average ride is 2 minutes longer during weekends and for casual riders it peaks on saturdays, sundays and mondays with  3 more minutes on average then rest of the week.

In summer, members' average ride is 1.5 minutes longer during weekends, while for casual riders we can appreciate a curve with symmetrical distribution with peaks on Saturdays and Sundays of 3 minutes longer rides compared to Wednesday where the shorter average rides are found.

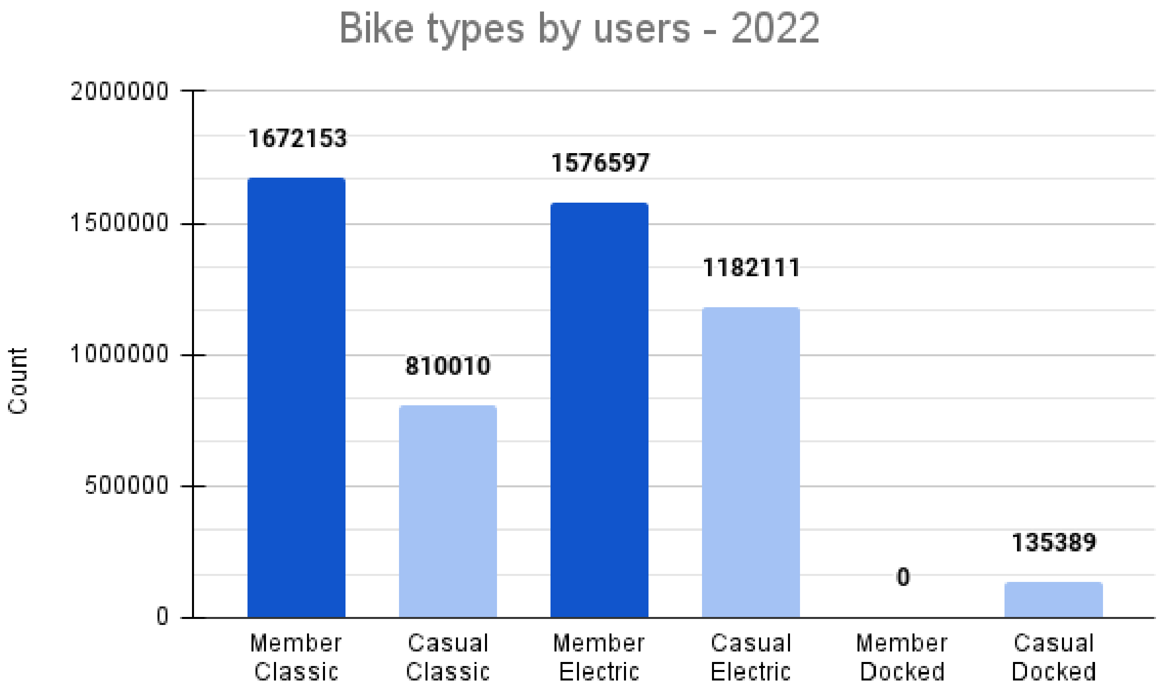
In fall, members' and casual riders' tendency is similar to summer with the difference that the shortest average ride for casual riders shifts to Tuesdays.

6. **What type of bicycles do users use? Are there any differences of choice between members and casual riders?**

Annually

Doing an annual overview we can realize that members have a slight preference for classic bikes whereas casual riders prefer by far electric bikes. On what refers to docked bikes (tricycles), there is not much engagement in general. Members don't use docked bikes at all and only a few casual riders choose them throughout the year.

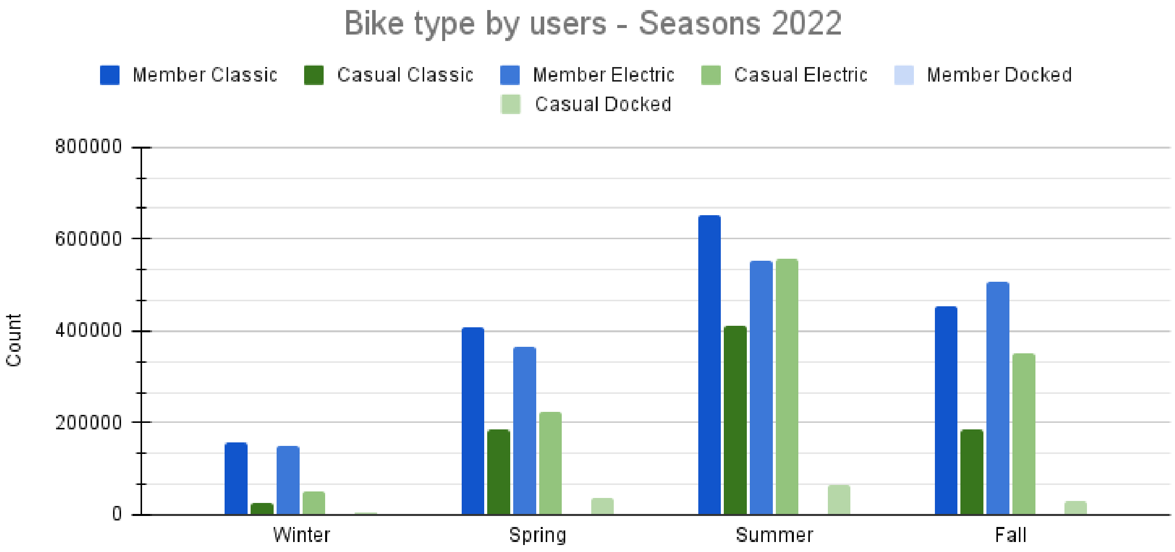
Graph 31



<u>Seasonal</u>

Seasonally we see the same tendency as when we analyze the whole year with the exception that in fall the members choice changes being electric bikes the preferred ones.

Graph 32

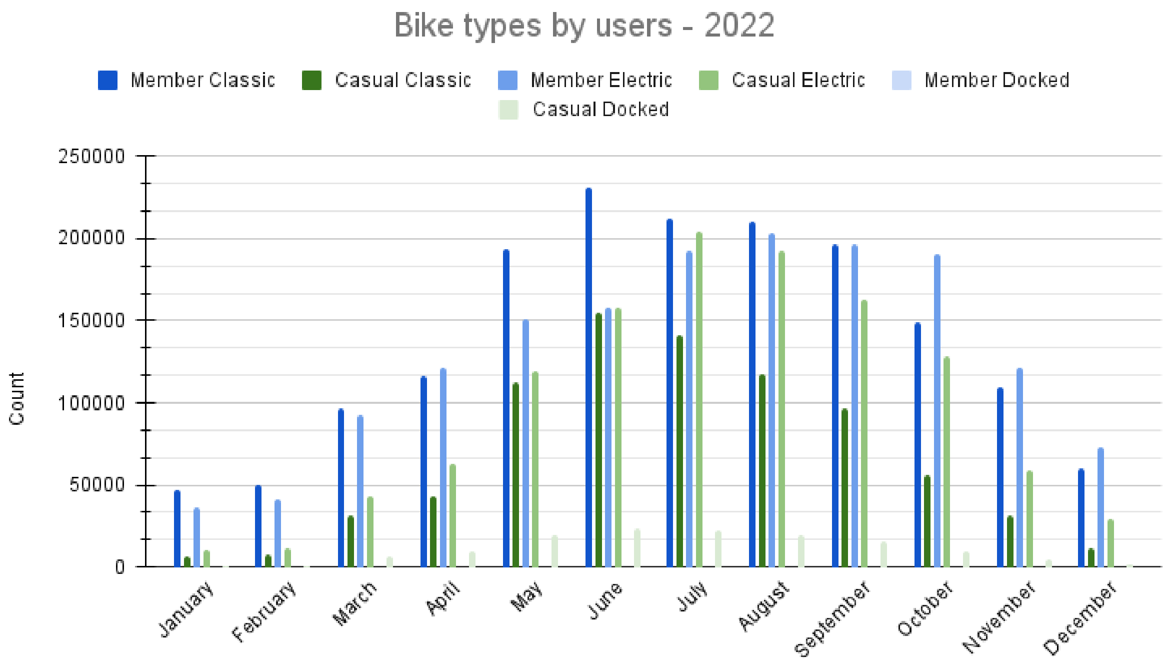


<u>Monthly</u>

When observing the behavior monthly we realize that there are 5 months of the year where members choose electric bikes as much as classic bikes or more. Those months are April, September, October, November and December.

For casual riders, electric bikes choice is a constant throughout the year.

Graph 33

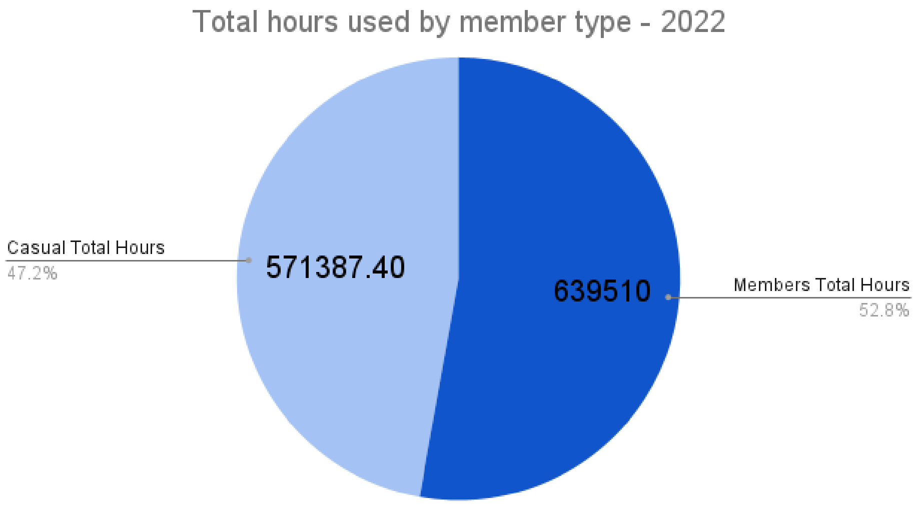


7. **How much time have the bikes been used in total?**

<u>Annually</u>

In terms of hours of use, members represent 52.8% of the time, being casual users 47.2%. That is a significant amount of time for casual riders.

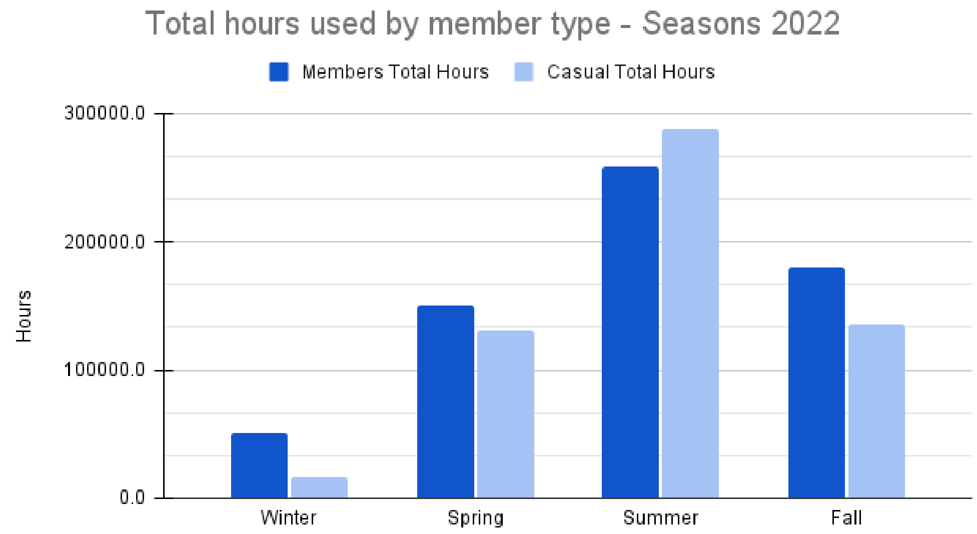
Graph 34



Seasonal

By seasons, we can see that in summer casual users use the service more if we add the duration of every trip.

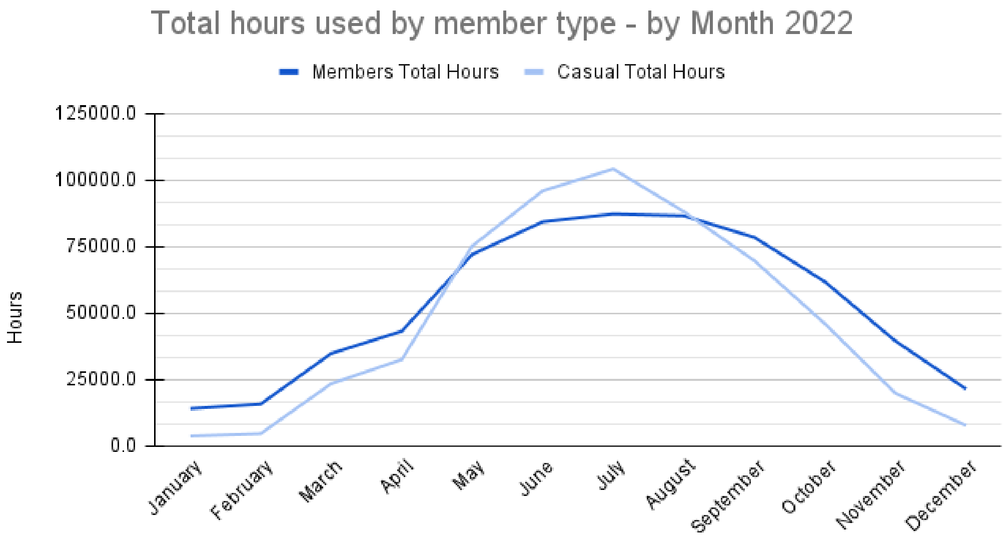
Graph 35



Monthly

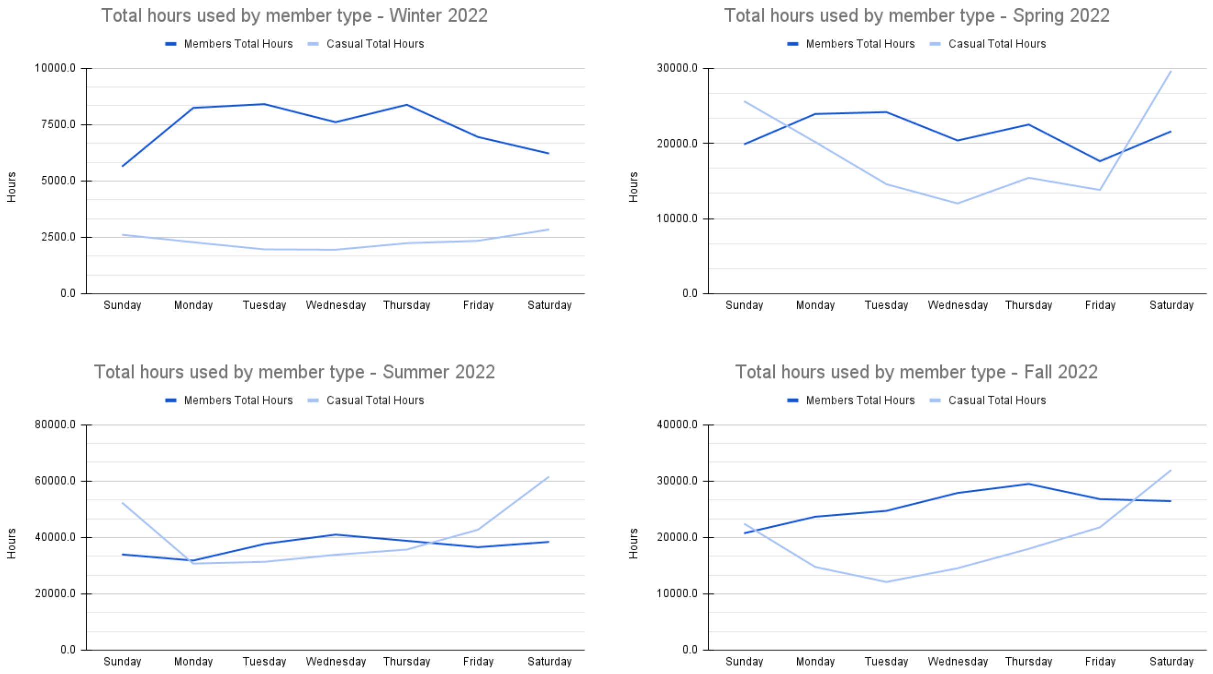
I decided to analyze the time of use by month to have a better picture of the behavior of both user types. In May, June, July and August casual riders make more use of the service. July being the month with the biggest gap where casual riders used the bikes for 16913.5 more hours, which represents 16.2% more time.

Graph 36



Day of the week

Graphs 37-40



Winter is the season where casual riders barely use the bike sharing service. Compared to members, casual riders use the bikes ⅓ than members on weekdays and ½ of the time on weekends.

For the other three seasons we can see that on weekends, casual users use the service for more time.

In summer time, casual riders use the service for longer on Fridays as well.

## Conclusion

- Members use the service more in terms of amount of rides, but considering the amount of time of use casual riders and members use the service evenly.
- Summer is the peak season where casual riders' use time is greater than members’.
- Members use the service more consistently during the year and specially during weekdays. Whereas casual riders prefer to use the service in hot weather, specially in summer and weekends.
- Throughout the year the average ride length is always longer for casual riders.
- Members tend to prefer classic bikes whereas casual riders prefer electric ones.

# Share

TABLEAU DASHBOARD

# Act

Based on the results of the analysis there are a few recommendations I can give to help the marketing team reach the goal of converting casual riders into members.

It would be interesting that the marketing team runs a survey in January-February among all casual riders trying to gather information about what kind of service, features and benefits would they consider getting a membership.

After the survey and having in mind that in winter the service is barely used by casual riders, the best time of the year to promote and launch a campaign to gain members is in March-April. Right before the usage starts to increase in Spring. Probably would be good to offer seasonal passes as an alternative to annual memberships. Apart from Summer, in Spring and Fall there are still many casual riders that chose the service. This way, casual riders can save money in winter when they don't use the service. Another option is to allow members to put the annual membership on hold for up to three months which can be very convenient and a good feature that may lead to greater conversion and retention rates.

Apart from annual memberships with 'On Hold' option and seasonal passes, it can also be an option to consider different types of membership with different benefits at different prices. For example: 'x' amount of minutes pass (full week or only for weekends), monthly passes or even passes only for a specific type of bike.