# Cyclistic Case Study

## 1 - The Task At Hand

In summary, the Cyclistic marketing director Lily Moreno wants to initiate a campaign aimed to convert *casual users* (those who simply use the company's 24-hour passes as needed) to *members* (those who have purchased an annual membership). In order to pursue this goal strategically, she has tasked the analysts with answering three questions, and as a recent hire to the analytics team, I have been specifically assigned to the first of these:

| | |
|---|---|
| 1 | How do *members* and *casual users* differ in their use of the company's services? |
| 2 | What would motivate a *casual user* to become a member? |
| 3 | How could digital media marketing best be utilised to achieve this conversion? |

I have been made aware that the executives in charge of approving the marketing team's proposed campaign are particularly detail-oriented, and will appreciate a very thorough analysis.

# 2 - The Raw Data

The data files I have been provided in order to answer my question have been made publicly available at https://divvy-tripdata.s3.amazonaws.com/index.html.[1] They consist of a series of .csv files containing data from every logged use of the company's products since launching in 2013.

For the purposes of this case study I initially decided to pretend it was currently early 2020, and use the data reported from the two years preceding.[2] I could have included earlier data, but it would likely be less relevant to the "current" trends in usership. The formatting of these tables has changed over time. In my chosen range of time, they contained the following 12 columns:

- trip_id: an ID assigned to each trip taken
- start_time: day and time trip started, in CST
- end_time: day and time trip ended, in CST
- bikeid: an ID assigned to each physical bike
- tripduration: the length of the trip taken, in seconds
- from_station_id, from_station_name: ID and name of the station where trip originated
- to_station_id, to_station_name: ID and name of the station where trip terminated
- usertype: either "Customer" for a *casual user*, or "Subscriber" for a *member*
- gender: either "Male" or "Female" (seriously?)
- birthyear: in YYYY format

For some reason, the files for 2018 Q1 and 2019 Q2 had different column names from the rest, so I simply edited the first line of each (in notepad) to make these consistent. I

[1] Licensed by the real-world bicycle rental service Divvy, who "Cyclistic" are based on for the purposes of this case study.
[2] Prior to 2020 the data was reported quarterly, or bi-quarterly, but this changed to monthly. Using the older data thus meant I had to upload fewer tables to my SQL server, saving some time and effort on the initial data entry. Also this avoids the likely obfuscating factors of covid-19 lockdowns periods.

then imported each of the 8 files into my SQL server, with the following column data types:

| | |
|---|---|
| **varchar** | trip_id |
| **nvarchar** | from_station_name, to_station_name, usertype, gender |
| **datetime2** | start_time, end_time |
| **smallint** | bikeid, from_station_id, to_station_id, birthyear |
| **float** | tripduration |

The choice to do my initial overview and cleaning of the data in an SQL server, rather than simply importing them directly into a spreadsheet application, was motivated by a couple of different factors. Firstly, since the files are pretty large (some over a million rows) I figured opening them with anything more sophisticated than a basic text editor would be a slow affair. I also wanted the chance to stretch my SQL muscles, and I'm glad I did. It had been a few months since I last ran any queries, and I was a little rusty at first.

# 3 - Verification & Preparation

Each table had a number of rows for which the tripduration field was inconsistent with the difference between start_time and end_time. In fact, almost half of the logged trips in the second half of 2019 had this issue, though most of these were only off by 1 second, likely due to rounding errors. The more significant discrepancies tended to be very close to 1 hour, and were mostly in cases where the bike was not returned for weeks or months[3], so again this is probably a rounding error. I chose to excise all rows for which this discrepancy was greater than 10 seconds, and dropped the tripduration column, in favour of simply computing this directly from the timestamps.

Continuing in the verification process, I used a series of SQL queries to confirm the following for each table:

- In each remaining row, the start_time field precedes the end_time
- The trip_id column was a unique identifier for each table
- The station_ids are in 1-1 correspondence with the station_names
- The gender column only contains nulls, and Male/Female entries
- Only the gender and birthyear columns contain nulls.

I then investigated the distribution of entries in the birthyear column. In each table, there were around 100-200 trips for which the reported birthyear was prior to 1920. Even if a centenarian was to use Cyclistic's services, they probably aren't the target demographic for the forthcoming marketing campaign. The more likely situation is that most of these are people using the service under falsified information, whether by accident, for privacy reasons, or otherwise. Since we still have a huge amount of data, I felt it was fair to remove these entries.

[3] These bikes do get lost or stolen occasionally, and I'm sure there are more than a few in the Chicago river

Finally I ran a query to count the number of rows containing null entries in the gender and birthyear columns. Each table contained at most a few thousand of these for which the usertype was ‘Subscriber’, since *annual members* were ostensibly required to supply this information on signing up. However, the number of the rows corresponding to *casual users* which contained these nulls was a significant proportion of their total. For this reason, I decided it was best to keep all of these, and to instead take into account possible nulls in my later analysis.

At this point the row counts for each table were as follows:

| **Table** | **# Rows** | **# Rows with nulls** |
|---|---|---|
| Trips_2018_Q1 | 386974 | 22483 |
| Trips_2018_Q2 | 1059401 | 196661 |
| Trips_2018_Q3 | 1513271 | 294979 |
| Trips_2018_Q4 | 642452 | 48301 |
| Trips_2019_Q1 | 364945 | 19707 |
| Trips_2019_Q2 | 1108028 | 185555 |
| Trips_2019_Q3 | 1640391 | 287320 |
| Trips_2019_Q4 | 703818 | 66557 |

Based on this I made a couple observations, which I will investigate further with more in-depth analysis. Firstly, usership in general is lower in the first and fourth quarters of both years, as one might expect during the colder months in Chicago. Moreover, judging by the proportion of rows containing null entries, it seems the *casual* usership is especially affected by the time of year, which again would make sense, as I would imagine *annual members* are more likely to continue using the bikes, despite the city's notorious winter weather.

At this stage I thought the data was in good shape to begin the analysis proper, which I decided I would do using R. I exported each of the 8 tables as .csv files, then loaded these into dataframes in a new RStudio project. As a rudimentary check, I confirmed the row count for each table was preserved in this process, before combining them into a single dataframe "trip_data" (of 7,419,280 rows). Using tidyverse's mutate function, I made the following changes to this new frame:

- Added column "duration", storing the difference (in minutes, rounded to 4 significant figures) of the start_time and end_time fields.
- Added column "weekday", storing the day of the week for the start_time field, in abbreviated (3 character) format e.g. Mon, Tue.
- Changed data types of the "trip_id", "bikeid", "from_station_id" and "to_station_id" columns to "character", and the "duration" column to "numeric".
- Edited the entries in the "usertype" column, replacing "Subscriber" with "member" and "Customer" with "casual", in keeping with the terminology of this report.

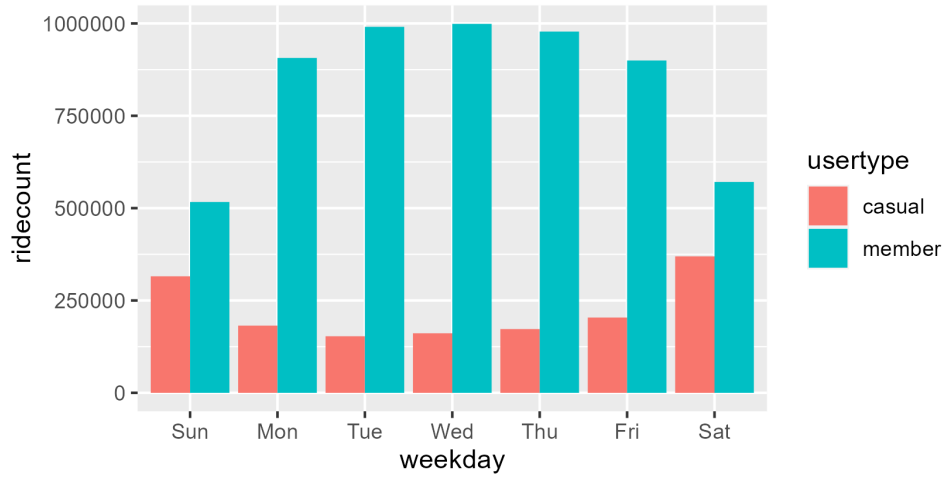
Inspecting the data closer, I found something I hadn't picked up on before. One of the possible "station_name" entries was "HUBBARD ST BIKE CHECKING (LBS-WH-TEST)", clearly a testing or repair facility for the bikes. Many of the trips for which this was the reported destination lasted days or weeks, likely during inspection or repairs, and I decided to remove these rows as well.

# 4 - Analysis

In order to address our task, we can compare the *casual users* and *members* across the following categories of statistics:

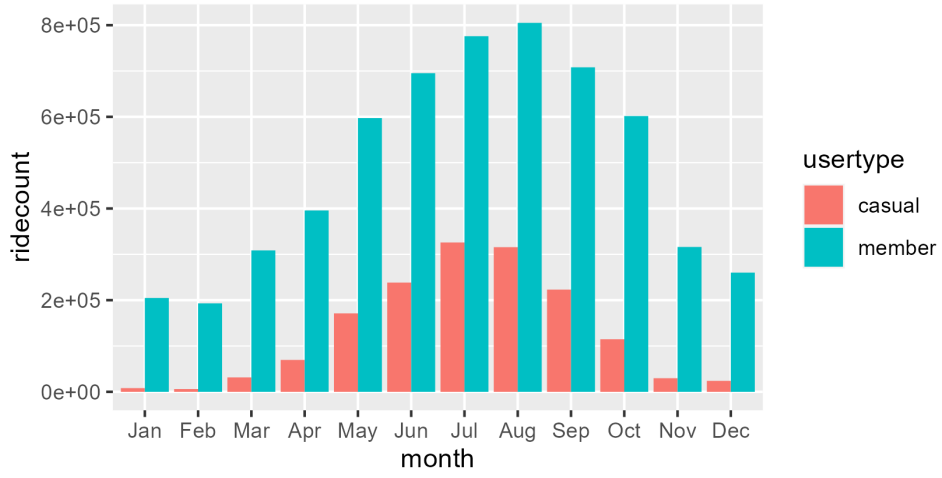
- Age demographics
- Trip length
- Day of week
- Time of day
- Time of year

To begin, I produced a pivot table computing the mean trip duration and total number of rides for each usertype and weekday. Immediately this revealed that, regardless of the day, *casual users* tend to use the bikes for around an hour at a time, whereas *members* only ride for about a quarter of that. From this table, I produced the following bar graph, which shows the total number of rides in each category:




This demonstrates another striking comparison between our two types of rider: whilst *casual users* tend to use the service more frequently on the weekend, *annual members* tend to use it more during the work week. It's also clear from this that the majority of rides taken are by *members*. I then ran similar code to produce a pivot table computing how ridership varied over the calendar months for the two rider types.

The resulting bar graph (below) reflects my earlier observation that a significant proportion of *annual members* will continue using the service during the winter, when *casual* usership falls away almost completely.




I repeated this twice more, to see how ridership varied by user age, as well as the time of day (using start_time), again grouped by usertype. While constructing the first of these, I noticed that there were no data points for casual users under the age of 15, despite a significant number of annual members in this age range. It turns out that the company doesn't allow anybody under the age of 16 to use the bikes, so I assume this was once again due to members falsifying their date of birth on sign-up. I decided to discount that analysis entirely, as it seems pretty unreliable. Finally, comparing usership by time of day of ride, didn't appear to spell out any new differences between the two rider categories, so I abandoned that too.

# 5 - Results & Recommendations

Outputting the two pivot tables I constructed in Rstudio (ridership by weekday/month) to csv files, I then took this data and constructed a dashboard in Tableau Public, giving a polished (and interactive) version of the above bar graphs. Ultimately, the three major differences I observed in the data were as follows:

1. Whilst *casual users* tend to ride for just under an hour on average, *annual members* only ride for around 10-15 minutes at a time.
2. Though both user types tend to use the service more during the summer, there are a large number of *annual members* who continue riding year-round.
3. *Annual members* tend to use the service more during the work-week, whereas *casual users* use it more on the weekend.

It would appear from this that *annual members* are more likely than *casual users* to use the bikes for commuting to/from work, rather than for leisure.

As for my recommendations, I think the following could be effective strategies:

- Emphasising the health benefits, both physical and mental, of continuing to ride in the winter. Something like “beat the winter blues”, playing off the blue-coloured bikes?
- Pushing the idea of “approaching your commute like the weekend”. People heading to work dressed for the beach etc.
- Position the bikes as an alternative to commuting via rideshare services or public transit, with a substantially smaller carbon footprint.[4]

[4] It would be useful to be able to cross-reference our data against use of these other transport services.