Cyclistic 12 Month Report

Grant Wills

2023-05-03

My goal in this case study is to identify and report on the trends of casual users of Chicago bike sharing company Cyclistic. A **casual** user of the service is one who pays for a single-day pass or is charged per ride, as opposed to Cyclistic **members** who buy annual memberships. Both types of users use Cyclistic differently. I aim to outline those differences clearly supported by collected data from **March 2022** to **February 2023**. Identifying the trends and habits of casual users could help Cyclistic advertise memberships to this group.

Some questions immediately arise: How do casual users use Cyclistic differently? Why are casual users not buying memberships? Is it even reasonable to expect this group to buy memberships or do their habits completely contradict purchasing one?

The Data

I'm using historical data made available by Motivate International Inc. https://divvy-tripdata.s3.amazonaws.com/index.html The data is stored in 12 .csv files

---

202203-divvy-tripdata.csv

202204-divvy-tripdata.csv

202205-divvy-tripdata.csv

202206-divvy-tripdata.csv

202207-divvy-tripdata.csv

202208-divvy-tripdata.csv

202209-divvy-publictripdata.csv

202210-divvy-tripdata.csv

202211-divvy-tripdata.csv

202212-divvy-tripdata.csv

202301-divvy-tripdata.csv

202302-divvy-tripdata.csv

Each file on average contains a few hundred thousand rows of data. Each row represents an individual bike ride taken during that month denoted by a ride_id column. Each ride has a start date/time and an end date/time, and a starting and ending station where users can pickup and return the bikes. Location data is also represented as longitude and latitude coordinates. We get information on what type of bike was used, and finally whether the trip was taken by a casual user or a member.

Data Cleaning and Manipulation

Since each file is already too large for spreadsheet applications to run efficiently - and I aim to combine them into one table - I will be using SQL for the majority of the cleaning process. All SQL code is running using MySQL. To see the script with all cleaning documented, please visit https://github.com/grantwills01/Cyclistic-Capstone.

- Load each .csv file into a new table in MySQL
- Combine 12 months of data into one table called all_tripdata; roughly 5 million entries
- Create column for ride length using the difference of columns ended_at and started_at
  - 101 entries have a negative ride length - meaning the started_at is later than the ended_at - so those entries have been removed
  - When I queried for largest values of ride_length, I found that many rides lasted days which did not seem right. Most of these rides were from casual users, so it is apparent that because casual riders aren't as familiar with the Cyclistic service as the annual members, they often are forgetting to end their trip using the app. I chose to cut off all rides exceeding 24 hours. This ended up being 5,411 entries consisting of mostly casual users and some members.
- Create column for day of the week of each ride using the started_at column of type DATE. This returns an integer 1-7 (1=Sunday, 7=Saturday).
- Create column for distance of each ride, rounded to 2 decimal places
  - Used a MySQL function called ST_Distance_Sphere() which takes coordinate points and returns distance in meters

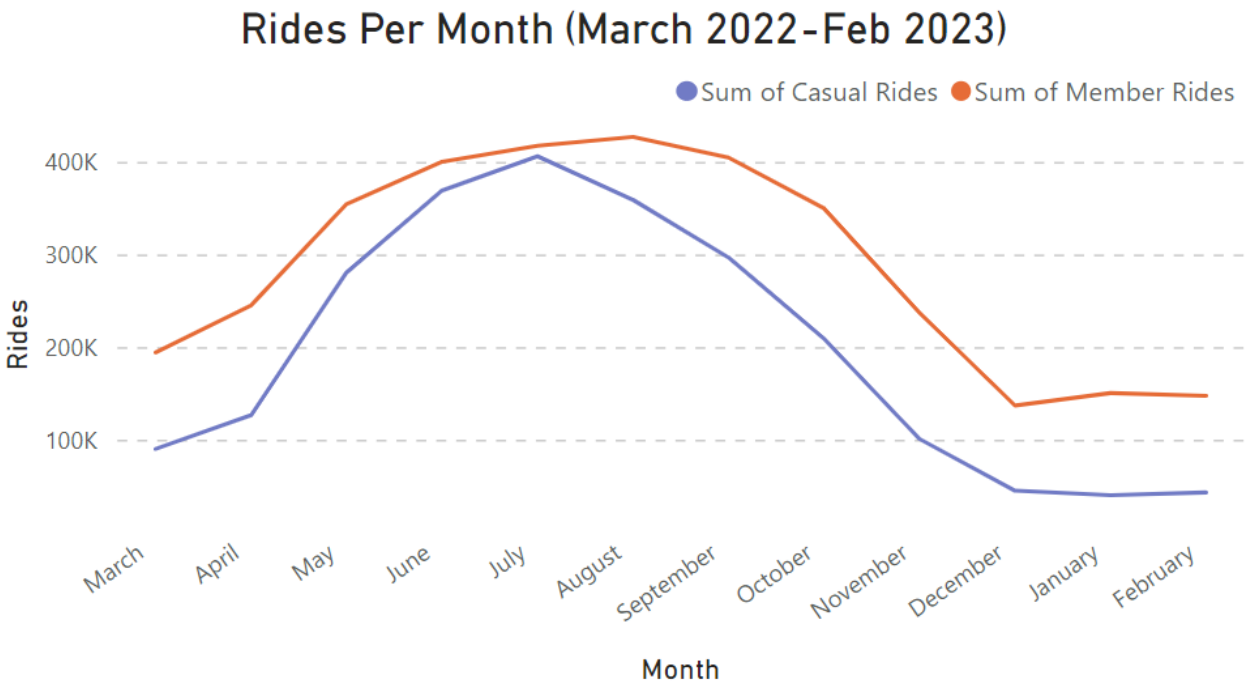
Analysis

At this point I will be querying the data to gain some initial insights into the trends of casual Cyclistic users. To see the SQL script I'm using to perform this analysis, please visit https://github.com/grantwills01/Cyclistic-Capstone.

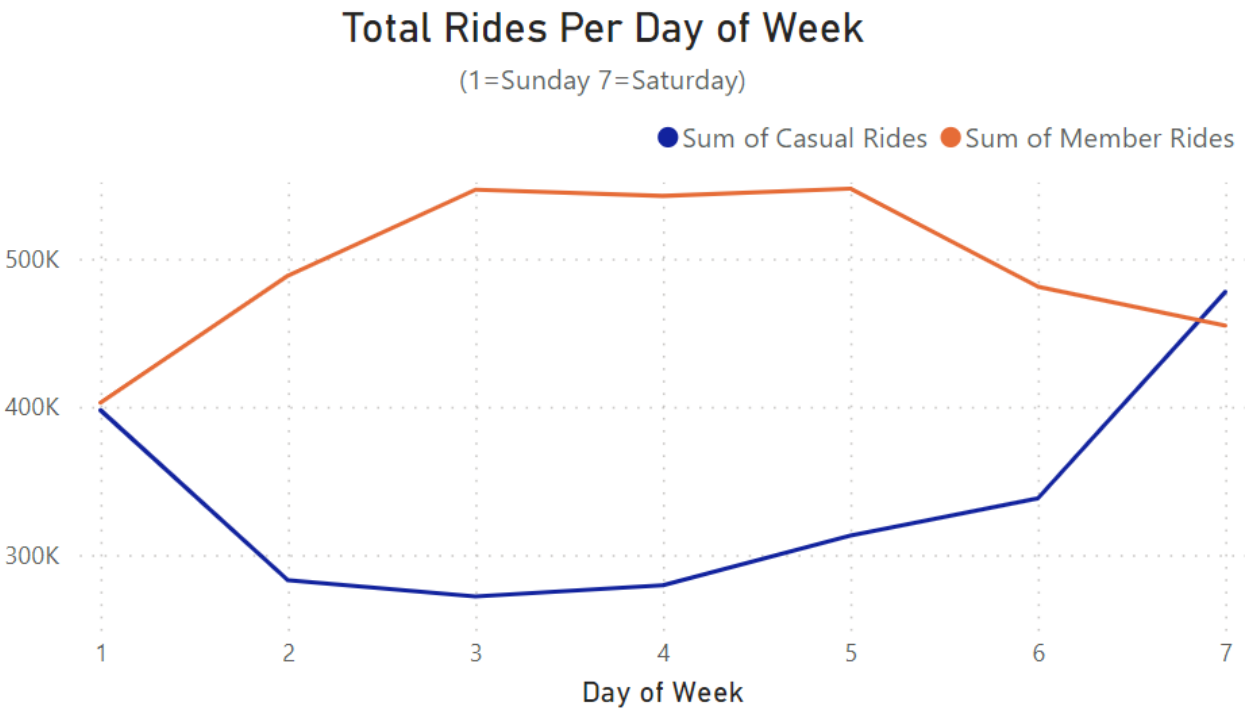
- Found that members use the service more than casuals throughout the year, but that casuals use Cyclistic almost just as much during the Summer. There is a large dip in both users as it gets closer to the holidays, but members still use it during those months while casuals are almost all gone. For example, in the month of January there were only 40,000 rides taken by casual users while members took over 150,000 trips. Immediately I have a theory that members are using Cyclistic for transportation and commuting to work, whereas casual users are using it more for recreational purposes.
- Found that on average there is an inverse relationship between members and casuals throughout the week. For members there is a dip in usage during the weekends and a steady rise during workdays. The opposite is true for casual users. This supports the running theory that casual users ride bikes for fun while members ride for transportation.
- Both average distance and average time of each ride for casual users are about double of the members. Member rides on average last for 12 minutes. Casual rides go for 25 minutes. Members ride for 2,111 meters on average. Casuals ride for 5,098 meters. This could mean that casual users are often tourists wanting to explore the city on bikes. It could also be people just wanting exercise.
- There is a pretty big jump in casual user traffic at station 13022. This kind of jump does not exist in members. They are more consistent throughout Chicago. Casual users seem to prefer a particular station. This could be a spot with a lot of tourists or just a lot of foot traffic.

Official Findings

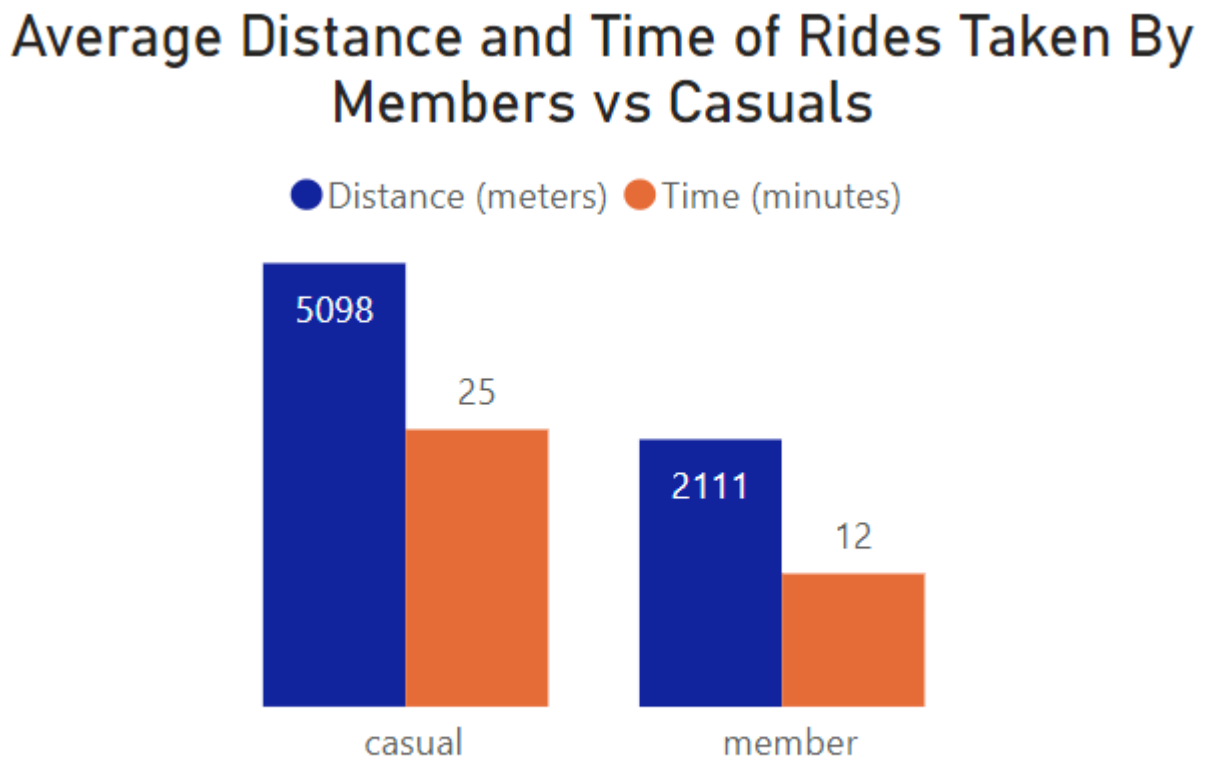
All visuals were done in Power BI.



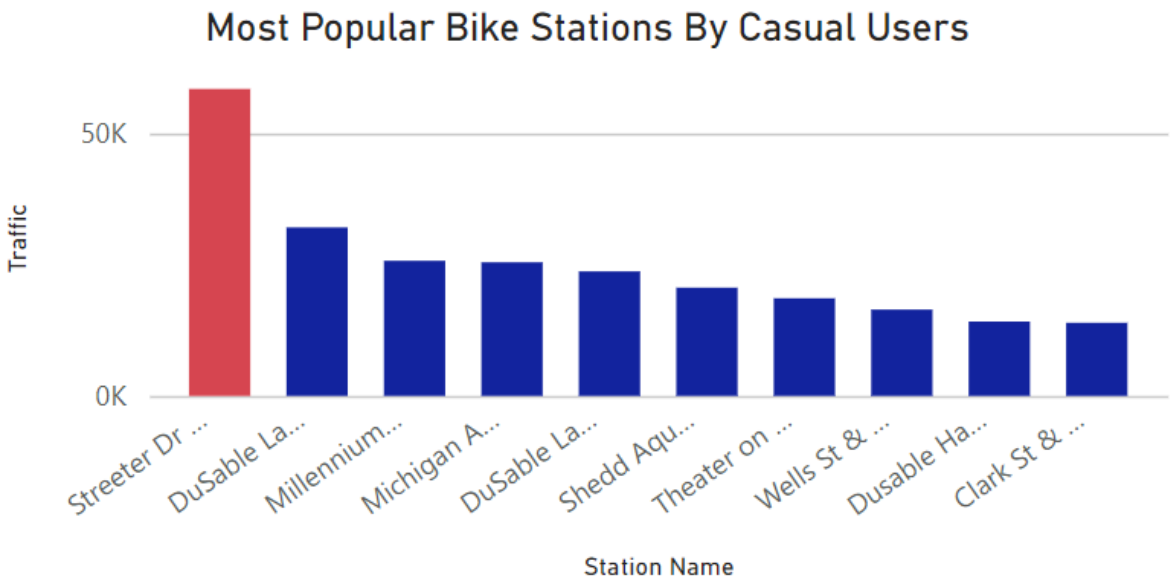
As you can see above, the number of rides for both casual users and members greatly increase in the Summer months and then dip in the holidays. Even in the downtime, Cyclistic annual members are still using the service far more than casuals who have come to a near halt. This is likely due to many members seemingly using the service for transportation means. Casual users seem to want to ride bikes in a warmer climate. Tourism could even be a contributing factor to this huge spike in the Summer. Let's take a closer look.



This is what the average week looks like for a Cyclistic user. We see an inverse relationship between the habits of annual members and casual users. Members use the service consistently throughout the work week and drop off during the weekend. Casuals spike in the weekend (even passing member numbers) and then dip considerably during the work week. A large subset of members clearly are using Cyclistic for their work commute.



The graph above shows that casual users ride Cyclistic bikes more than double the distance and for double the time compared to members. So what is stopping casual users from buying annual memberships? As mentioned before it could be that a large subset of casual users are tourists and are only around during the Summer months. For this group of people it is unreasonable to expect annual membership purchases if they are only in Chicago for a few months. However a large percentage of casual users could be Chicago residents who have never thought of using Cyclistic as a means of transportation. Perhaps if there were advertisements that convey the use of Cyclistic in that way, we might see an uptick in memberships. Now let's take a look at where in the city casual users are using Cyclistic the most.



The figure above lists the 10 stations with the highest traffic by casual users. There is a clear distinction as to which station is the most popular among this group. That would be station 13022 on Streeter Dr. and Grand Ave. This is likely a tourist location as it is right next to Navy Pier and the Children's Museum.

Top 3 Recommendations Based on Analysis

1. Advertise memberships heavily to casual users right before and during the Summer months. This is when they are riding bikes the most and are more likely to purchase a membership.
2. Convey in advertisements year-round the usage of Cyclistic for transportation and commuting to work. It is possible that casual users have never considered this use-case.
3. Advertise outside of Chicago. Tourism appears to be a clear factor in the casual usage of Cyclistic. Creating media that portrays Chicago as the city to bike in year-round could convince tourists to buy annual memberships.