Data Analysis Report: Ridership Differences at Divvy

Prepared For: Divvy Executive Leadership

Prepared By: Alec Ciapara

Abstract

As a Junior Data Analyst at Divvy, I analyzed differences in bikeshare usage between casual riders and members to encourage conversions. Utilizing unbiased public data devoid of personally identifiable information, **I employed SQL in Google's BigQuery for data processing and examination. The analysis reveals that members, predominantly university students or professionals, undertake frequent, short trips for commuting purposes throughout the year. Casual riders, who engage in longer rides for leisure or exercise, primarily utilize bikeshare services during spring and summer. New strategies may be necessary to convert casual riders into members, focusing campaigns with reduced rates prior to peak usage periods.**

Scenario

Divvy is Chicago's bike share program operated by Lyft. The data is real and available here through a license. **We assume the role of a Junior Data Analyst for Divvy, which provides both electric and classic bikes through a bike share app.** Divvy maintains a fleet of over 6000 bicycles that can be retrieved or returned to docking stations or left in designated areas. Divvy caters to Members, who purchase annual membership plans for reduced fares, and Casual users, who do not possess a membership and utilize the service on an ad-hoc basis.

Ask

Lily Moreno, Director of Marketing, has tasked us with increasing annual memberships to ensure the company's success. We will present our findings to the executives, explaining how members and casual users utilize the bike share differently, why casual users would become members, and how digital media can influence this shift. I am responsible for the first question: **"How do annual members and casual riders use Divvy bikes differently?"** This will help the team understand why casual users might purchase a membership.

Prepare

To answer Lily's question, we will analyze Divvy's latest bike trip data. This data covers usage by casual users and members, helping us identify patterns between the two groups. Stored in Amazon AWS servers as monthly CSV files, the data is credible and bias-free, collected by Divvy under a public license. It excludes personally identifiable information (PII), so we must treat user groups collectively. The dataset includes three types of bikes: electric, classic, and docked (the latter referring to classic bikes).

Process

While the single files can be opened in Excel, the combined 5.5+ million rows of data would not be manageable. Therefore, SQL and Google's BigQuery were used to process, explore, and clean the data. The data processing involved uploading 12 CSV files, from February 2022 through January 2023, to a bucket in the Google Cloud service. **The combined data included 5,754,248 rows across 13 columns.**

Field name	Type
ride_id	STRING
rideable_type	STRING
started_at	TIMESTAMP
ended_at	TIMESTAMP
start_station_name	STRING
start_station_id	STRING
end_station_name	STRING
end_station_id	STRING
start_lat	FLOAT
start_lng	FLOAT
end_lat	FLOAT
end_lng	FLOAT
member_casual	STRING

Schema for Divvy data.

After processing the data, I began exploring it. **A copy of the SQL code I used to process and explore the data can be found on GitHub.**

```
C5.start_station_name & C6.start_station_id & C7.end_station_name & C8.end_station_name
-Check for leading/trailing/double spaces.
-Verify naming consistency.
*/

SELECT DISTINCT(start_station_name)
FROM ride_data
GROUP BY start_station_name
ORDER BY start_station_name;

SELECT DISTINCT(end_station_name)
FROM ride_data
GROUP BY end_station_name
ORDER BY end_station_name;

SELECT start_station_name
FROM ride_data
WHERE start_station_name LIKE '%   %';

SELECT COUNT(start_station_id)
FROM ride_data;

SELECT start_station_id
FROM ride_data
WHERE start_station_id LIKE '%  %';

SELECT *
FROM ride_data
WHERE start_station_name IS NULL OR end_station_name IS NULL AND
  rideable_type = 'classic_bike';
```

Sample SQL code used.

- Ride ID- This column contains a specific identifier connecting it to one specific instance of bike use, from start to finish. **I verified each Ride ID as distinct and unique, and confirmed each was a string of only 16 characters.**
- Rideable Type- This column names the type of bike used in the ride, classic or electric. I confirmed there were no empty rows, but docked bike is present. **Docked bike is an outdated term for classic bikes, thus we will change docked bike to classic bike in the cleaning process.**

- Started At & Ended At- These columns represent timestamps of a bike ride being initiated and then completed. As they are two timestamps, I verified them together. The time is created as Chicago local time, but appears in the data as UTC, this will be corrected. **Using a TimeStamp_Diff function we can subtract the starting time from the ending time to verify the ride duration. There are 5,390 rides over a day, and 229,452 rides less than a minute or null. These are likely user or system errors and will be purged during the cleaning process.**
- Start Station Name/End Station Name/Start Station ID/End Station ID- As these are all strings meant to identify where a ride started and where it ended, I will verify them together. As is common with strings, there are trailing and leading spaces which must be removed for naming consistency. **More problematically, over 800,000 Classic bike rides, which must end in a dock, have no ending station logged, and so we will correct this in cleaning using latitude and longitude data available to us later in the schema. Start and End station IDs have no value to us and are part of Divvy's interior statistics, as such we will remove these columns during the cleaning process.**
- Start Lat/Start Lng/End Lat/End Lng- These columns all contain floats which give us coordinate data on where a ride started and ended. **I verified 5,899 rows contain null values for at least one of these values, these will have to be purged from our final data so that we can map the coordinates. An end station at Green St. & Madison shows 8 rows of null coordinate data, but as we have the data in other instances of this station, we can correct this. Finally, we will use this coordinate data to correct the missing station names as stated in the bullet point above.**
- Member/Casual- This column simply lists the membership status of the account associated with the bike ride. **I confirmed there were only 2 possible values present and no null values. No further cleaning necessary.**

Having completed my exploration, I used SQL to clean the data. **A copy of the SQL code used to clean the data can be found here. Cleaning the data was performed by creating 2 temporary tables, then joining them in a third table to be queried for the analysis.**

```
47    cleaned_data AS (
48      SELECT *
49      FROM
50        (SELECT ride_id,
51          CASE
52            WHEN rideable_type = 'docked_bike' THEN 'classic_bike'
53            ELSE rideable_type
54            END AS bike_type,
55            started_at, ended_at, TRIM(INITCAP(start_station_name)) AS start_station_name_c,
56            TRIM(INITCAP(end_station_name)) AS end_station_name_c, start_lat, start_lng,
57          CASE
58            WHEN end_lat = 0.0 THEN 41.881827376571351
59            ELSE end_lat
60            END AS end_lat_c,
61          CASE
62            WHEN end_lng = 0.0 THEN -87.648831903934479
63            ELSE end_lng
64            END AS end_lng_c,
65            member_casual
66        FROM total_data
67        )
68      WHERE start_lat IS NOT NULL AND
69        start_lng IS NOT NULL AND
70        end_lat_c IS NOT NULL AND
71        end_lng_c IS NOT NULL
72    ),
```

Query cleaning existing data.

1. The first temporary table (line 47) corrected the ending latitude and longitude for Green Street and Madison AVE., replace instances of docked_bike with classic_bike, and removes unwanted columns start_station_id and end_station_id. I also cleaned text strings for station names by trimming unnecessary spaces and correcting capitalizations and removed rides missing any coordinate information.

2. The second temporary table creates columns for analysis (line 78 on GitHub). I extract the day of the week, day of month, month, year, and ride duration from data present in the data set, making sure to use Chicago as time zone to correct the instances of UTC. I also dropped rides with less than a minute and more than 24-hour ride duration.

3. The third table (line 118) uses an INNER JOIN function to create a singular table from the other two. This table, ride_data, will be used for our analysis.

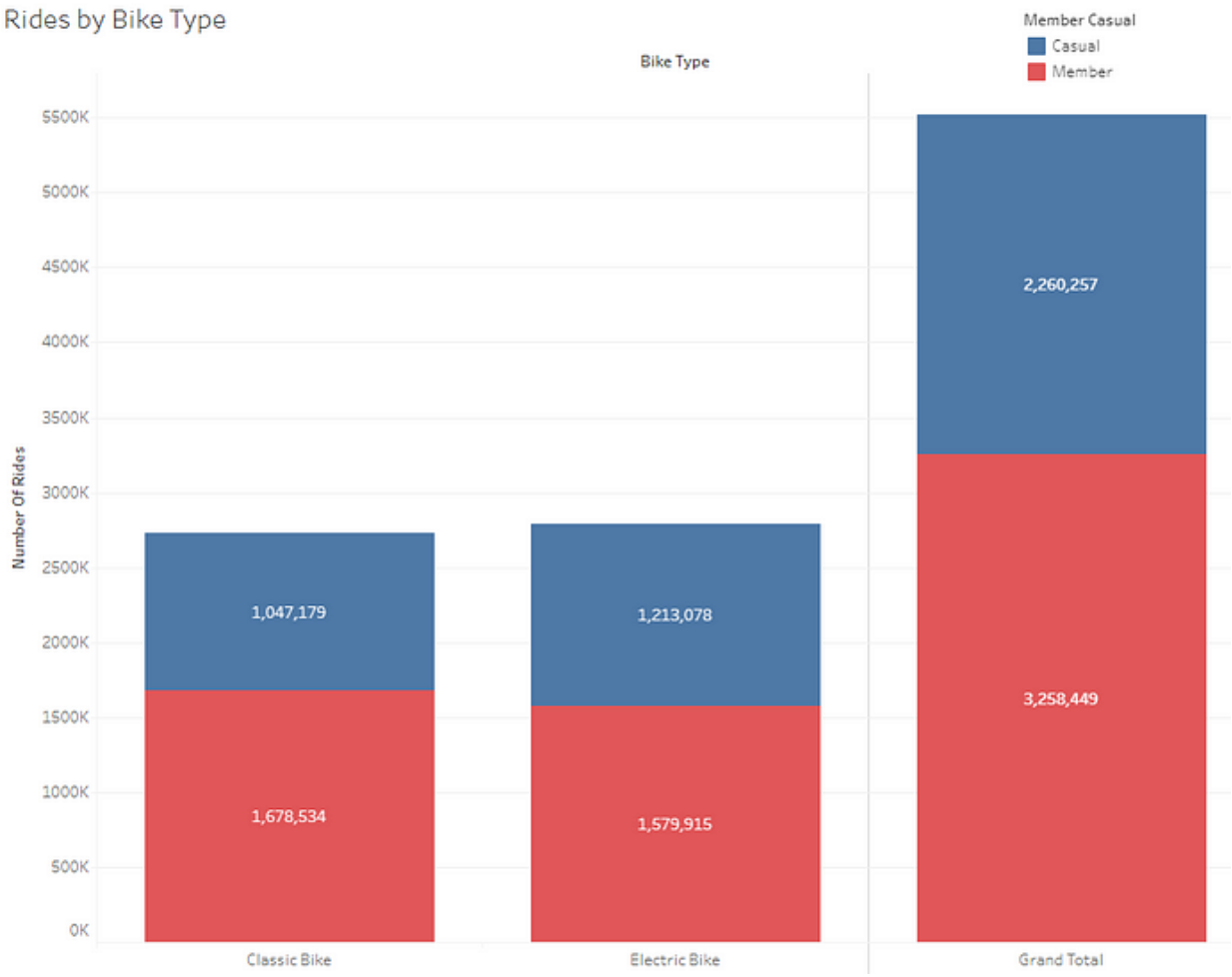
Analyze

The analysis was conducted by querying the ride_data temporary table to create CSVs for export. **Find the queries here, on line 126.** The goal was to compare patterns between casual riders and members. Queries included ride number, duration by hour, weekday, month, and popular start/end locations for both groups. Results are below.

Visualize

"How do annual members and casual riders use Divvy bikes differently?"

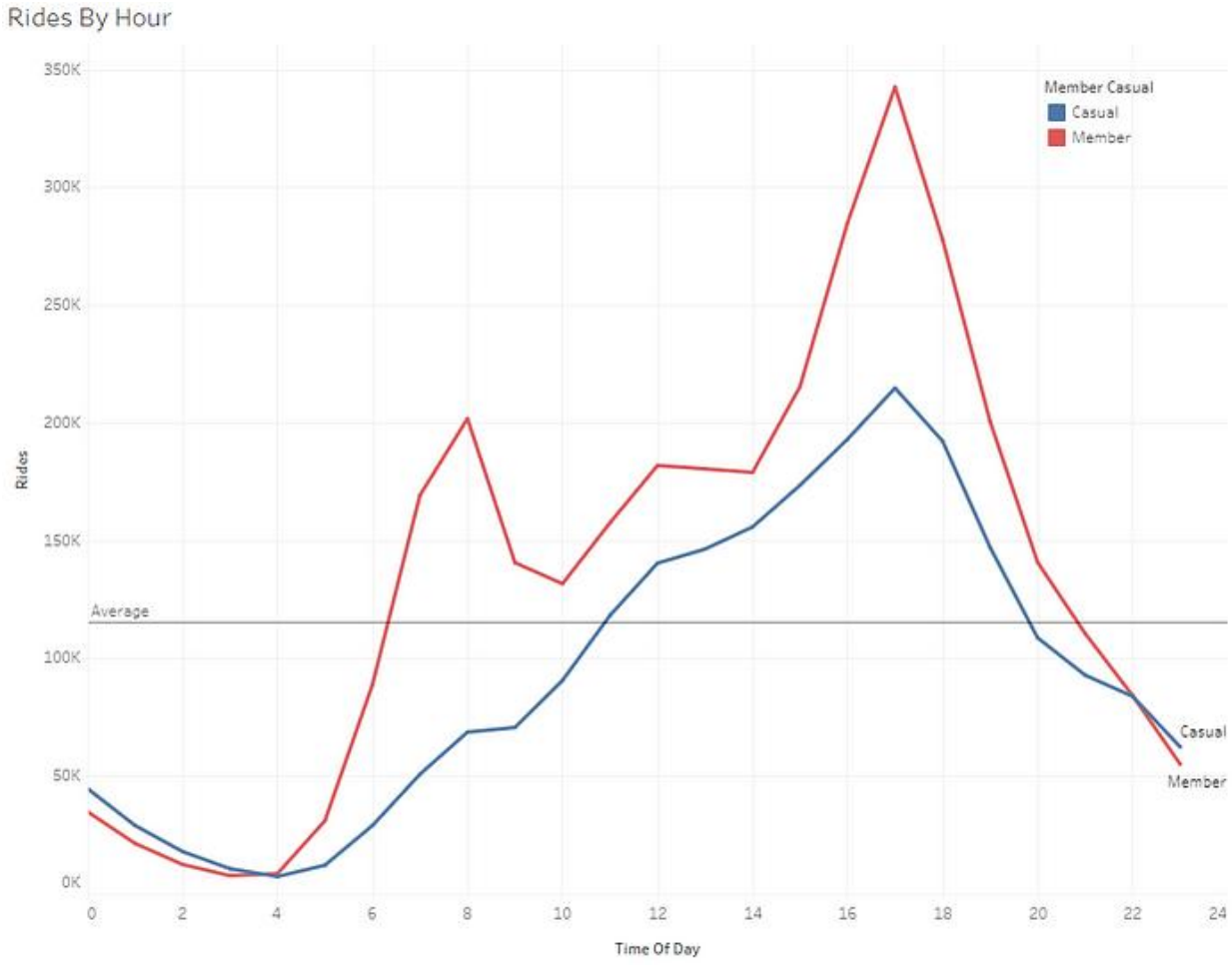
The visual representation of the analysis can be found here, on Tableau Public.



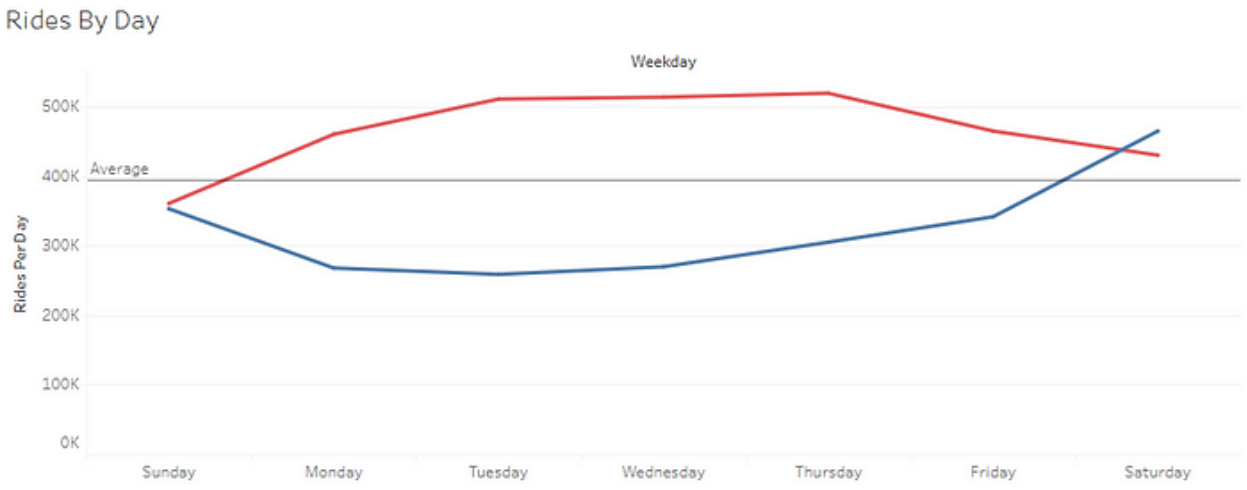
Total Rides by Bike Type

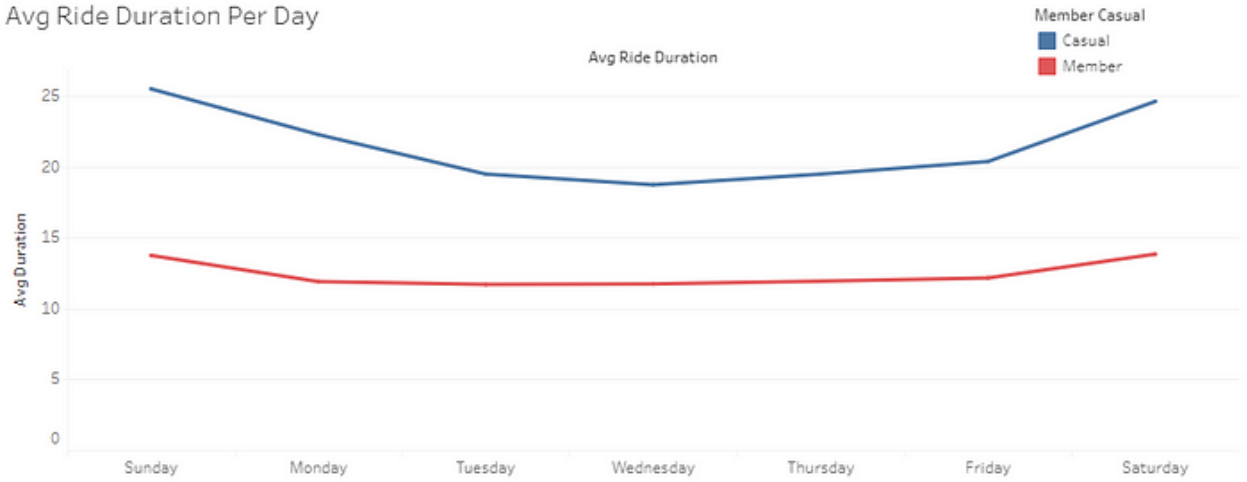
From February 2022 to January 2023, Members accounted for 59% of all rides. Both Members and Casual riders show no strong preference between Classic (manual) and Electric bikes. Members used Classic bikes in 51.51% of rides, while Casual riders chose

Electric bikes in 53.67% of rides. **Despite more rides by members, without personally identifiable information (PII), we cannot conclude that ride share users are more likely to be members.** It might be that a smaller number of members utilize the rideshare service more frequently throughout the year.

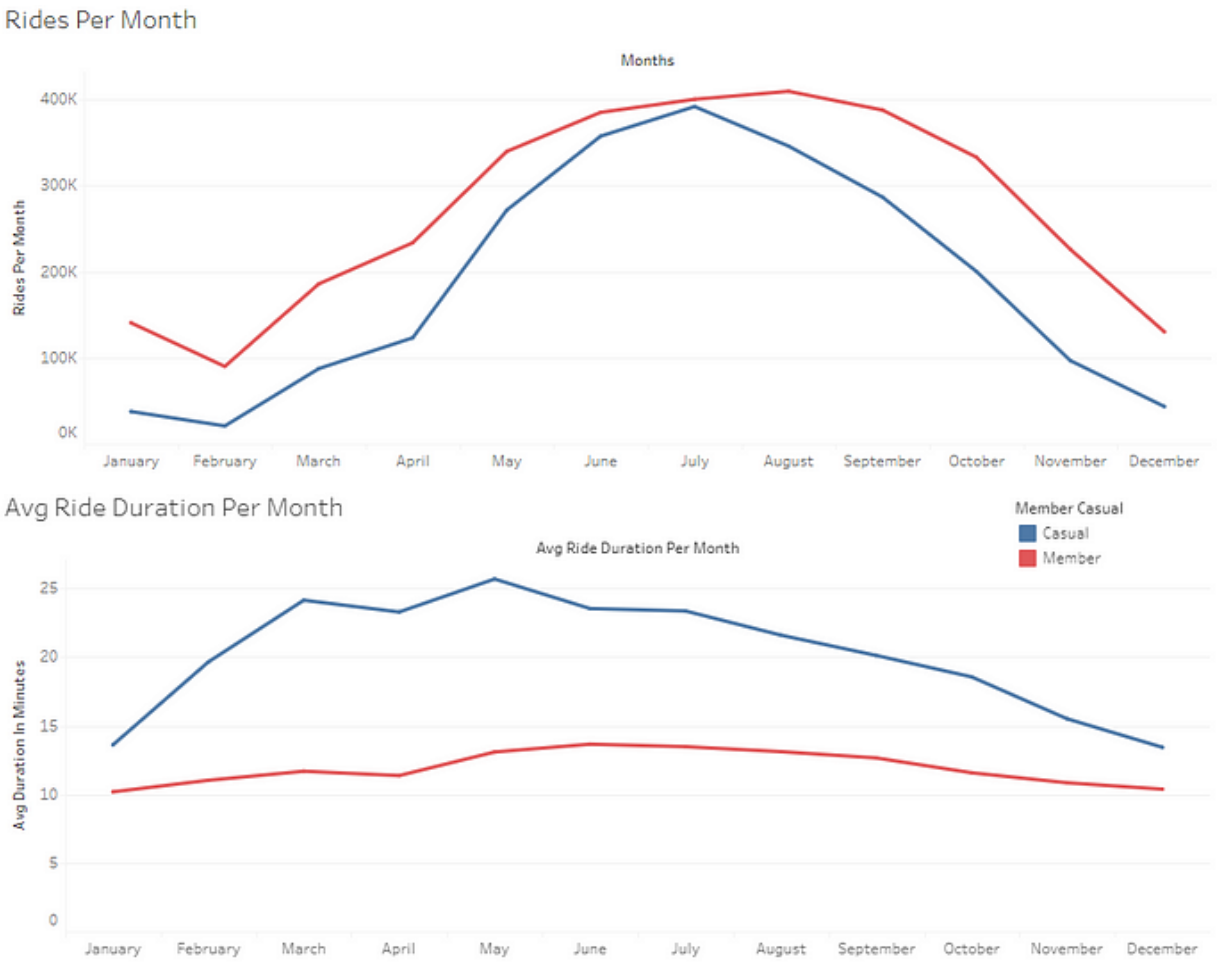


Next, we examine usage patterns throughout the day. The chart above indicates that the number of rides by members is generally higher than that of casual riders for most of the day. **The data suggests that members take advantage of their membership throughout the day. Noticeably, usage by members increases significantly during commute times, from 6–9 A.M. and 3–8 P.M., as well as during common lunch hours, 11 A.M. to 1 P.M..** Casual riders primarily use the service from about noon to 7 P.M.. **This information indicates that members may be frequent users who utilize bikeshare for commuting, while casual users might use the service less often, potentially as an infrequent alternative mode of transportation.**



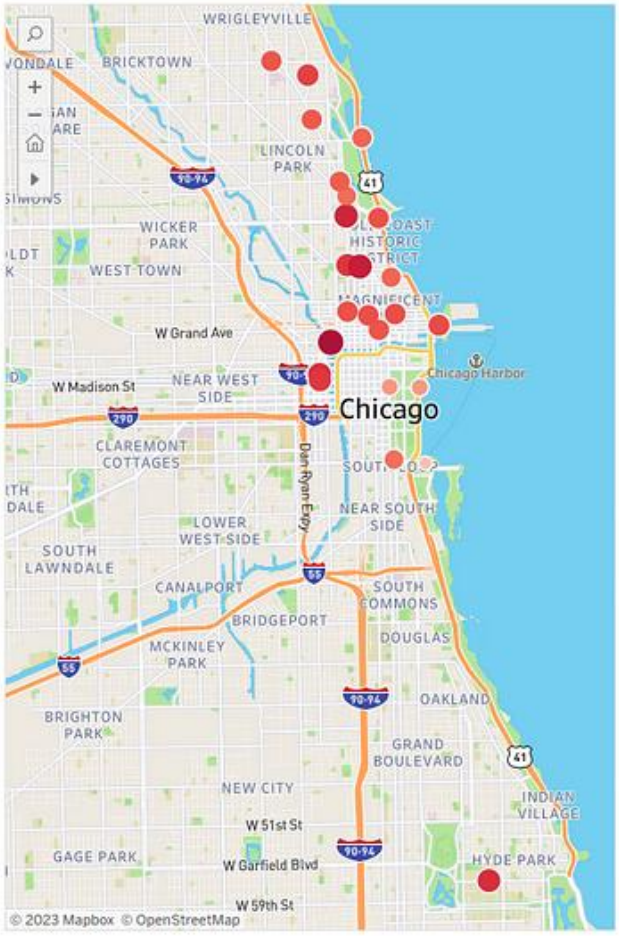


Next, we examine usage patterns by weekday. As indicated above, members exhibit a higher than average usage rate throughout the week, particularly on weekdays. In contrast, casual riders exceed the average usage rate primarily on Saturdays, with Sunday being their second highest day. When analyzing the average duration of rides per day, we observe that although casual users take fewer total rides per day overall, they tend to have longer ride durations regardless of the day. **Members predominantly utilize the service during weekdays, while casual usage peaks during weekends. This suggests that casual rides may be utilized for leisure, exercise, or extended journeys.**

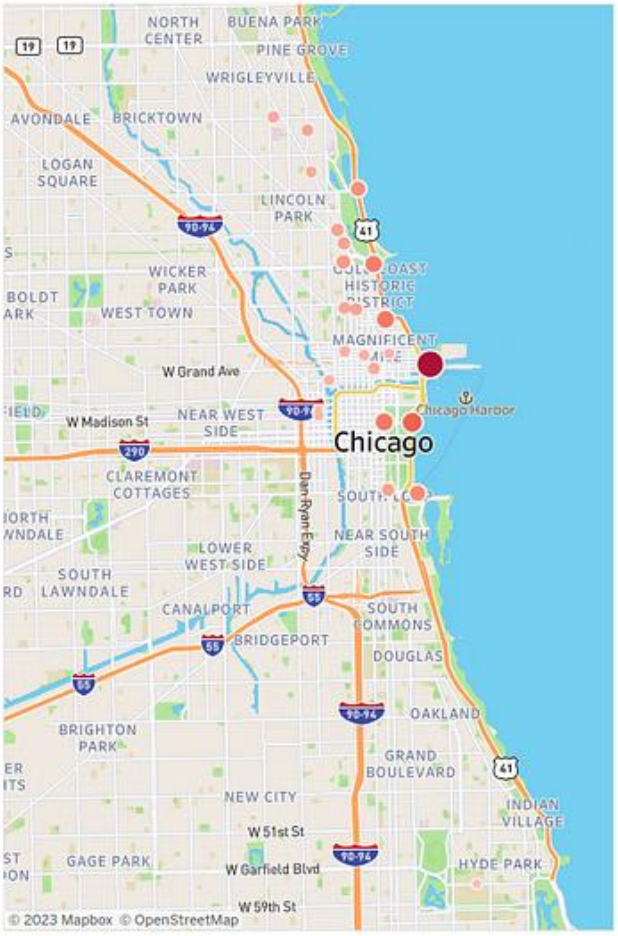


The chart above shows usage patterns per month. The total number of rides for both groups decreases during winter months, which is typical for Chicago's cold winters. Members show a gradual increase in rides throughout the remainder of the year, while casual riders have a sharper rise in usage during late spring and a steeper drop mid-autumn. Average ride duration per month indicates that casual riders experience an increase in ride duration particularly during spring and summer, whereas member ride durations remain consistent throughout the year. **This strengthens the hypothesis members may use the bikeshare as a primary mode of transportation, while casual users may use it for leisure or exercise.**

Popular Member Starts

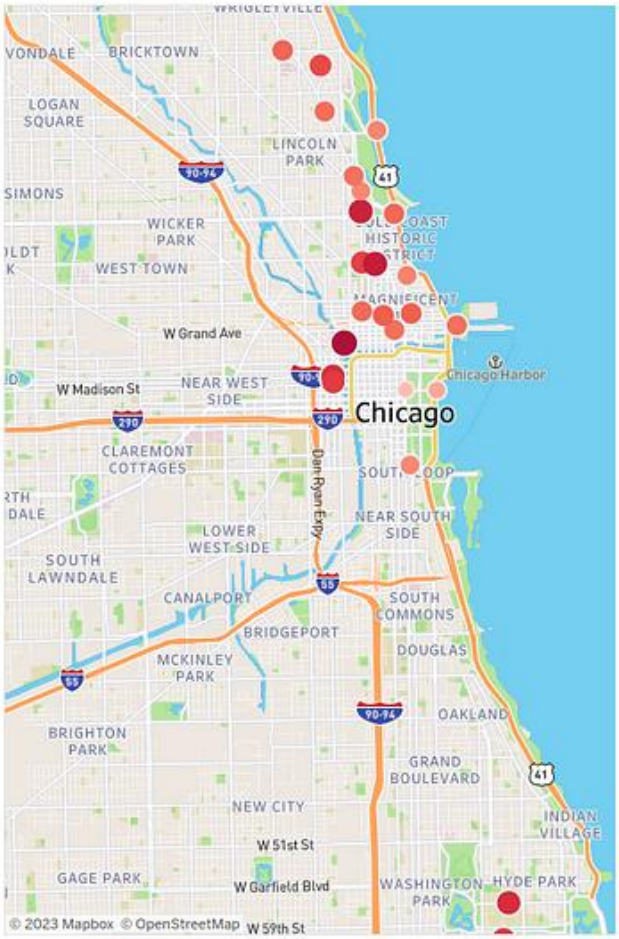


Popular Casual Starts

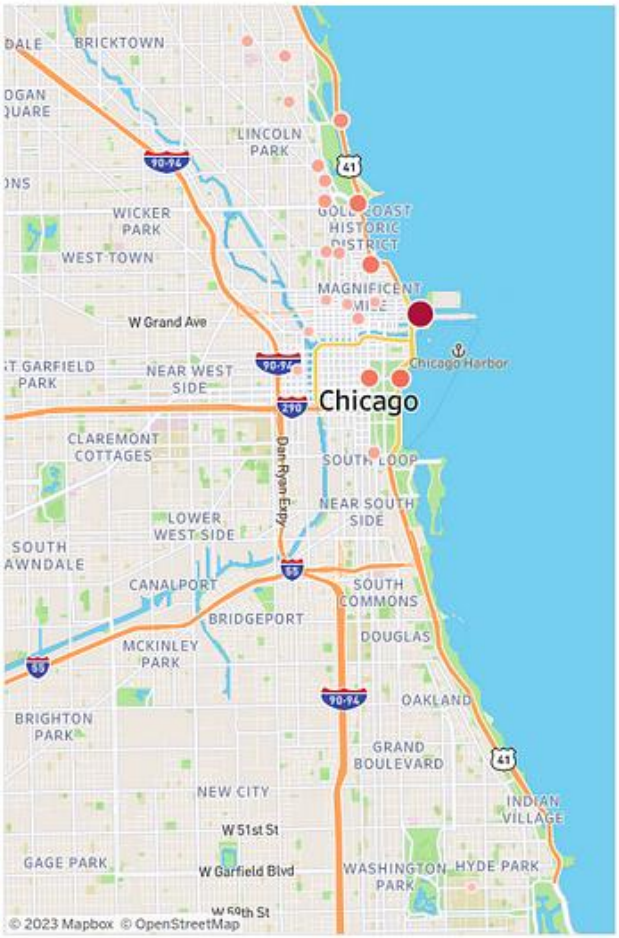


To test for casual users taking longer routes, I mapped the starting docked locations for members on the left, and casual users on the right. The top 25 starts are shown above. Members frequently start and end their rides at docks, whereas casual users may leave bikes locked to other city property instead. Below, I also mapped out the popular end docks. Casual riders use docks with less preference for specific locations, resulting in a more even spread amongst docks used even outside of the top 25. Members predominantly use docks in the northeast, University of Chicago, and University of Illinois campuses.

Popular Member Ends



Popular Casual Ends



Act

“How do annual members and casual riders use Divvy bikes differently?”

We can form our analysis into insights:

Members consist of university students and working professionals who rely on Divvy as their primary mode of transportation, particularly for commuting to and from work or school. Their journeys are typically direct and brief yet occur more frequently, with usage persisting throughout the year except during the winter months.

Casual riders can be any resident or tourist, riding for leisure or exercise. They generally take longer rides which start evenly distributed around the city but tend to end near the east side, around water features and attractions. They ride most frequently in spring and summer, particularly on weekends and later in the day when it is warmer.

Campaigns targeting casual riders should commence in the late winter to early spring months, as this period is when casual riders anticipate utilizing the service throughout the remainder of spring and summer. Implementing discounted plans for new members during this time may encourage trial adoption, leading to sustained membership. Additionally, a reduced rate plan could be established for weekends to cater to the higher usage by casual members or for extended ride durations. Moreover, the app could feature routes designed for individuals seeking exercise opportunities or highlighting city views and attractions.

Divvy bikes can be locked up on city property outside of a Divvy dock for an additional fee. Members, who use the bikeshare more frequently, often dock their bikes, while casual riders do not. This fee, combined with the frequency of use by members, can add up to a significant cost, so it is typically avoided. It has been proposed that the fee could be raised for non-members while remaining the same for members, in order to encourage individuals to join. **However, this may result in casual users ceasing to use the rideshare service due to increased costs and inconvenience.**