Cyclistic Bike Share Analysis

Jacob B Horton

2025-02-10

Purpose

The goal for this analysis is to determine differences in ridership behavior between casual and member riders and assess whether revenue generated from casual riders compensates for their lower trip volume.

Our primary business question:

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

Secondary questions:

- Why would casual riders buy Cyclistic annual memberships?
- How can Cyclistic use digital media to influence casual riders to become members?

Data and permissions

Cyclistic historical data ranges from 2013 to 2024 and we used as much as we could of it. for trip data we used the full range, for station data we used a subset (2020 to 2024).

Cyclistic ridership data was drawn from the data archives here: https://divvy-tripdata.s3.amazonaws.com/index.html

It was lent to us for analysis purposes under the license here: https://divvybikes.com/data-license-agreement

---

Process

Note that I conducted significant new research during this process and consulted various generative AI tools to help me identify processes, arrange syntax, and design code.

Step 1: Cleaning to align columns

Some 60 CSV files contained varying and evolving approaches to labeling and organization. To align these, I had to first pull all header names, along with samples of data in each column, from each file. This was done using Python. Once I had an understanding of the data organization, I had to create a high-level standardized organization, aligning headers and datatypes. This was done in Excel.

Example Python code

```python
output_rows = []

for column_name, details in column_data.items():
    row = {
        "column_name": column_name,
        "files_appeared_in": "; ".join(details["year_files"]),
        **details["samples"],  # Spread samples into the row dictionary
    }
    output_rows.append(row)

output_df = pd.DataFrame(output_rows)
```

Sample Spreadsheet data

Source.header	Standard.header	Source.datatype	Standard.datatype
Bike ID	id_bike	INT64	FLOAT64
Local End Time	end_time	STRING	TIMESTAMP
Local Start Time	start_time	STRING	TIMESTAMP

Step 2: Combining files into years and standardizing the columns

The data was organized into both years and months and I wanted to combine it into fewer CSV files, organized by year, mapping source columns to standardized columns. These files were 2 to 6 million rows long. I used a set of 8 Python transformation programs, each designed to align a single column to a specific datatype, while also uniformly managing problem values like blank cells or NaN values.

List of transformation programs run in Python

1. Timestamp transformation
2. Duration calculation in seconds
3. Basic string calculation
4. Float and Int transformations
5. Specialized Rider-type transformation
6. Specialized Station name transformations
7. Dealing with NULL data

Step 3: Aggregating granular data to higher-level data

The source data was organized by individual trips, which is why the datasets were so large. I needed to aggregate this data to higher level insights, such as trips per month.

I used Python and a few API tools to upload the 12 annual files to BigQuery. There, I used SQL statements to UNION the data into my desired, higher-level data groups.

Sample SQL statements

```
    -- 2024 Arrivals
    SELECT
        rider_type,
        TRIM(CAST(id_end_station AS STRING)) AS station_id,
        'arrival' AS trip_type,
        EXTRACT(YEAR FROM trip_end_time) AS year_num,
        EXTRACT(MONTH FROM trip_end_time) AS month_num,
        EXTRACT(DAYOFWEEK FROM trip_end_time) AS day_num,
        duration_seconds
    FROM `divvy-analysis-aggregat.2024_trips`
    WHERE duration_seconds BETWEEN 0 AND 12000
    AND rider_type != 'dependent'
)
SELECT
    rider_type,
    station_id,
    trip_type,
    year_num,
    month_num,
    day_num,
    COUNT(*) AS total_trips,
    AVG(duration_seconds) AS avg_trip_duration,
    MAX(duration_seconds) AS max_trip_duration,
    MIN(duration_seconds) AS min_trip_duration
FROM combined_trips
GROUP BY rider_type, station_id, trip_type, year_num, month_num, day_num
ORDER BY year_num DESC, month_num DESC, day_num
```

Step 4: Second round of cleaning

Once I had smaller, higher-level data files, I was able to identify problems in the data that I could not recognize at a lower level due to time, processing, and storage constraints. This process involved making certain key values uniform, reference, average, and sum columns. I did this cleaning in Excel.

Sample from a cleaned, higher-level spreadsheet

---

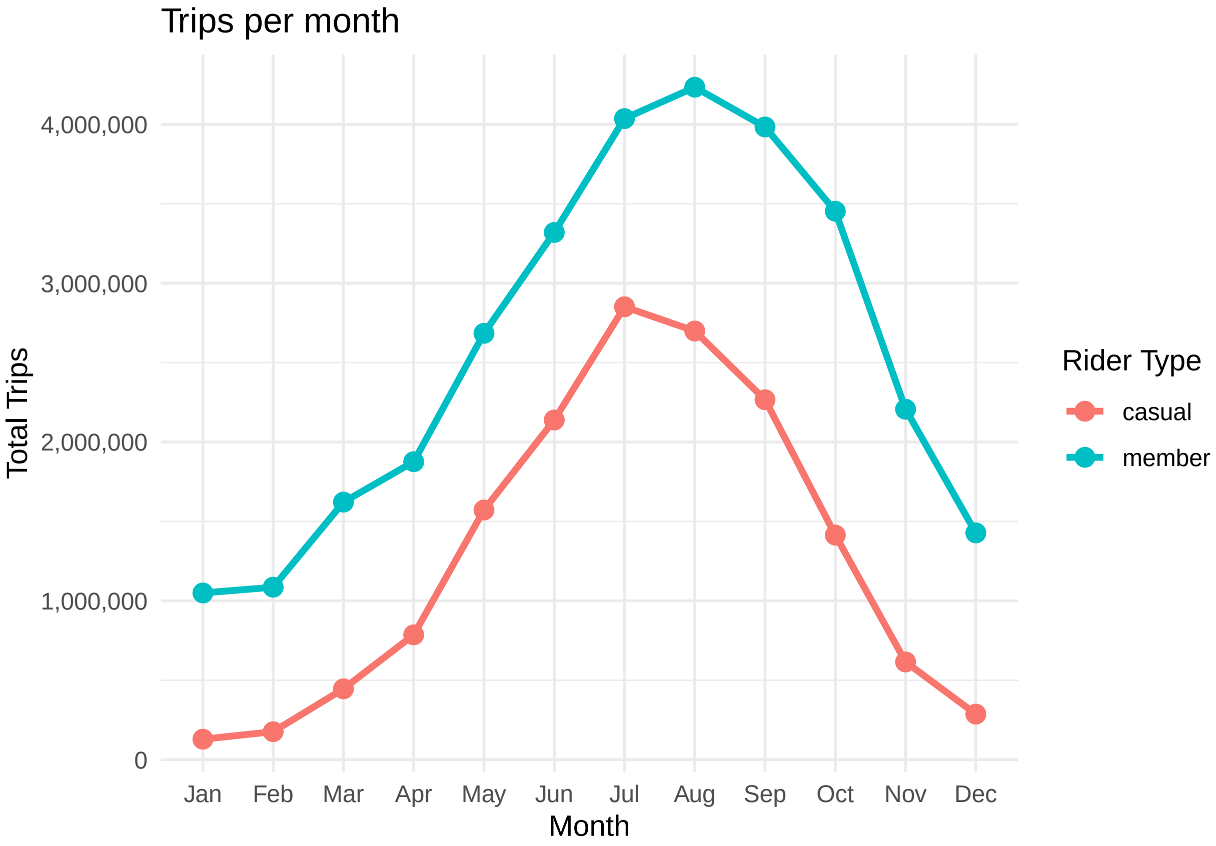
rider_type	year	month_num	day_num	hour_num	trips_total	duration_avg_secs
member	2024	8	1	11	3386	874.7
member	2017	5	7	14	2859	934.5
casual	2020	2	4	20	976	1345.2

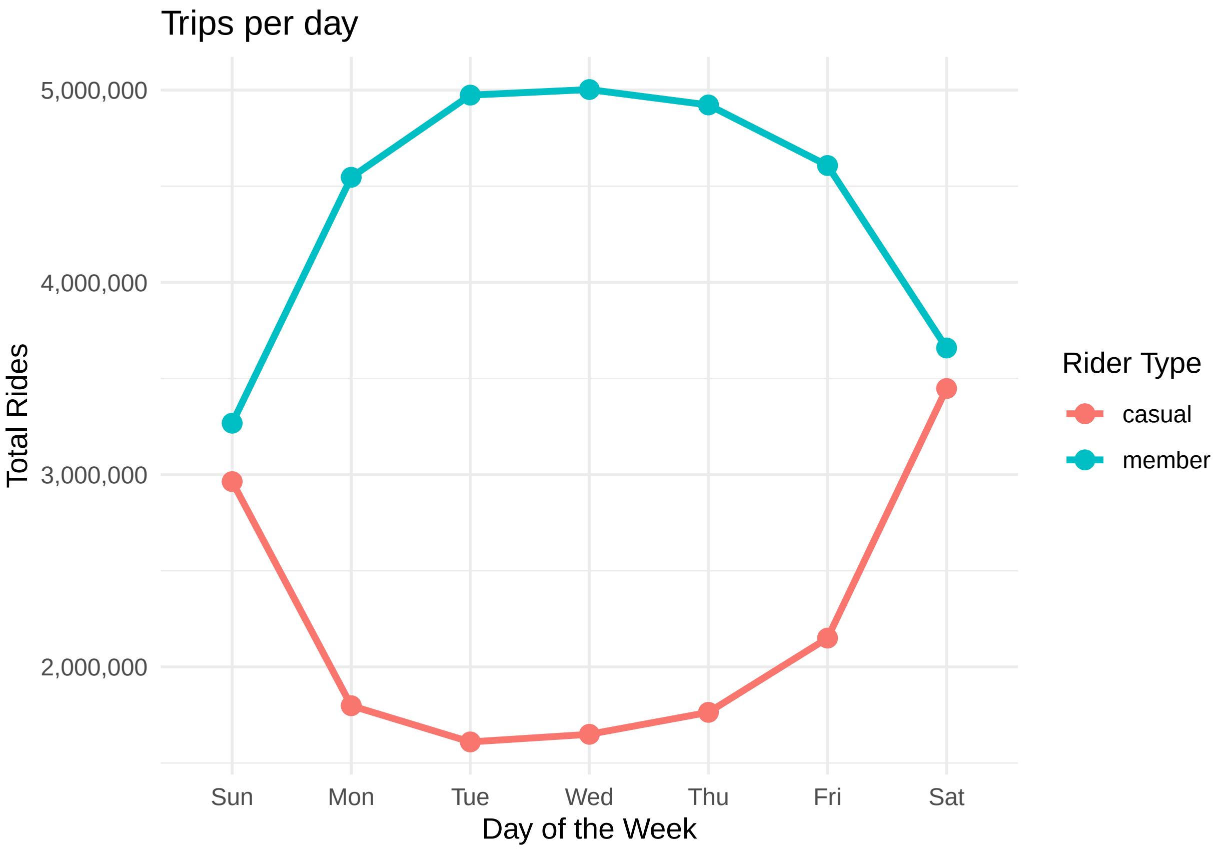
Step 5: Visualization

Once the aggregated datesets were cleaned, the could be uploaded intl RStudio for analysis and visualization. This document is the end product of visualization created using the RMarkdown library in RStudio.

Behavior of casual riders and members

How many trips do rider groups take?

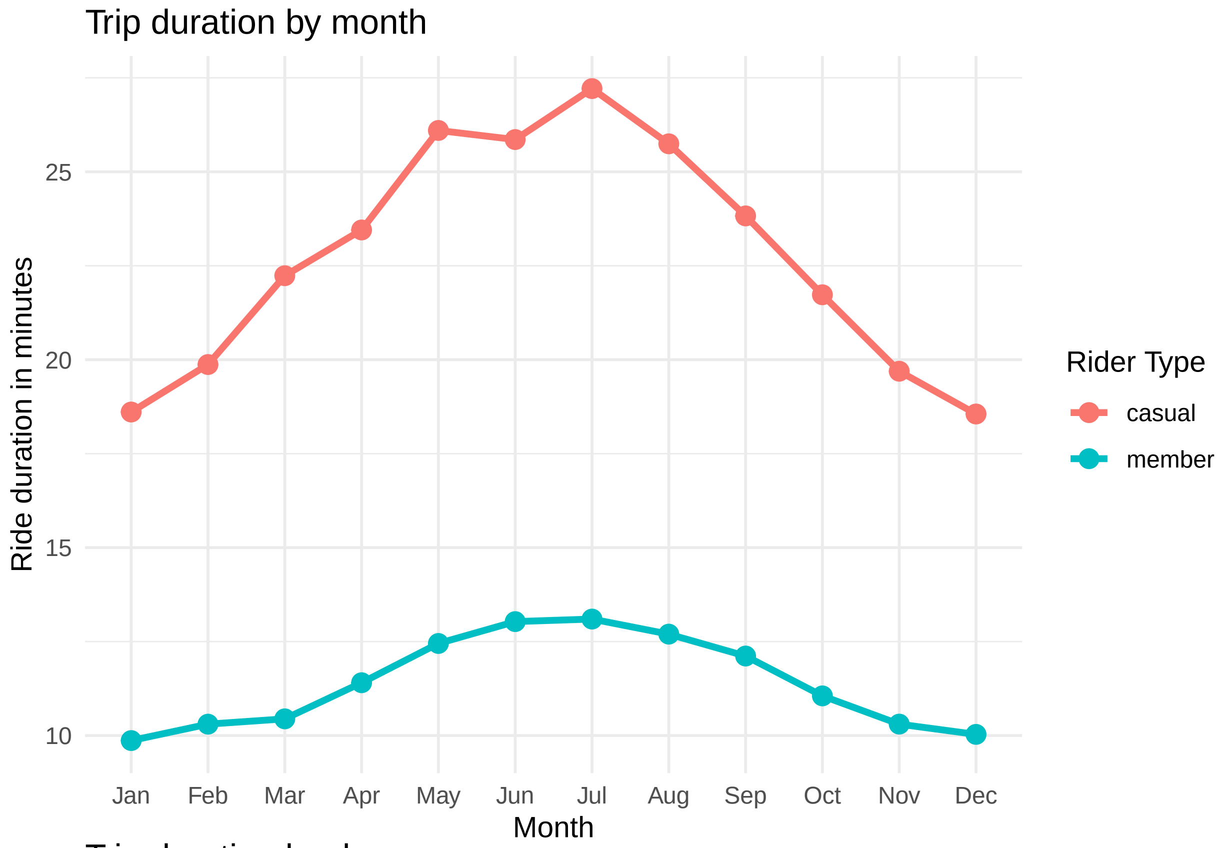


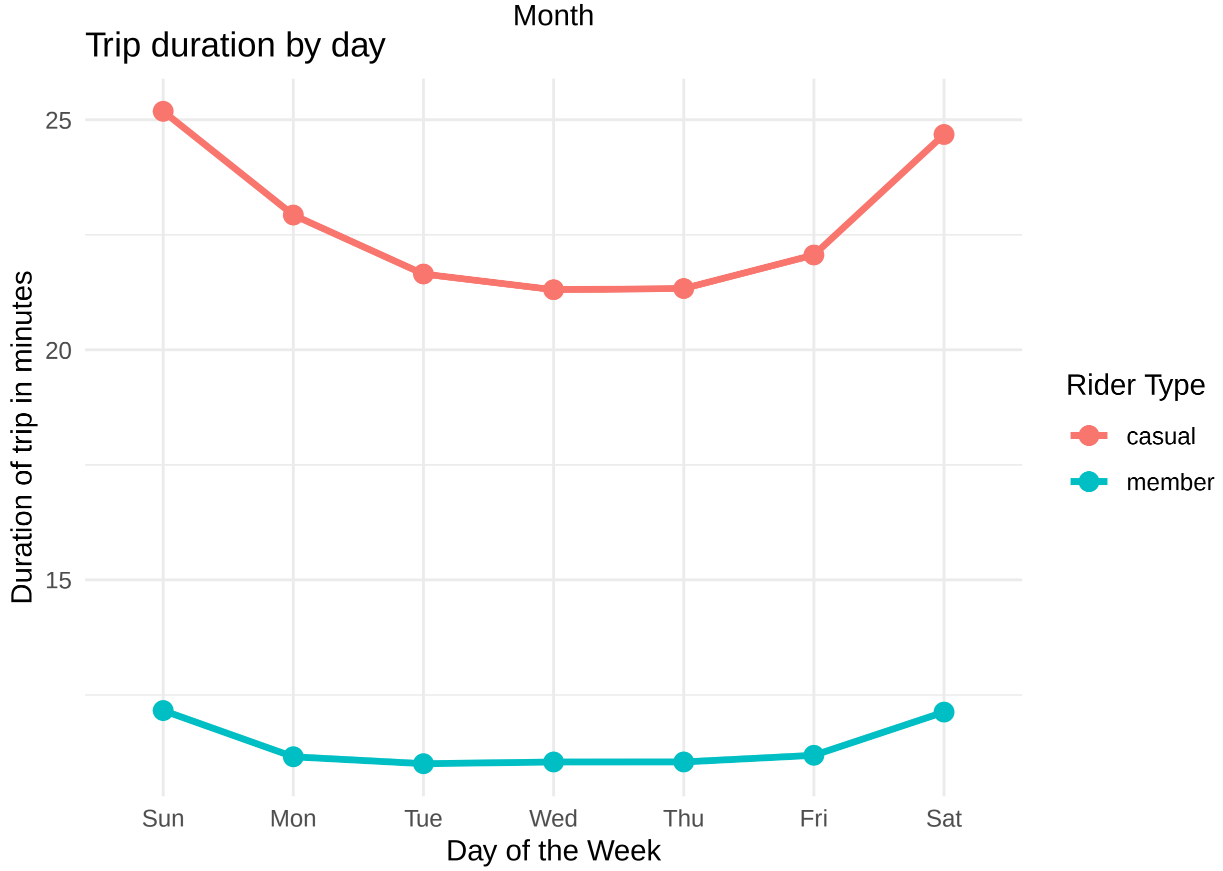


Notes on trip volume

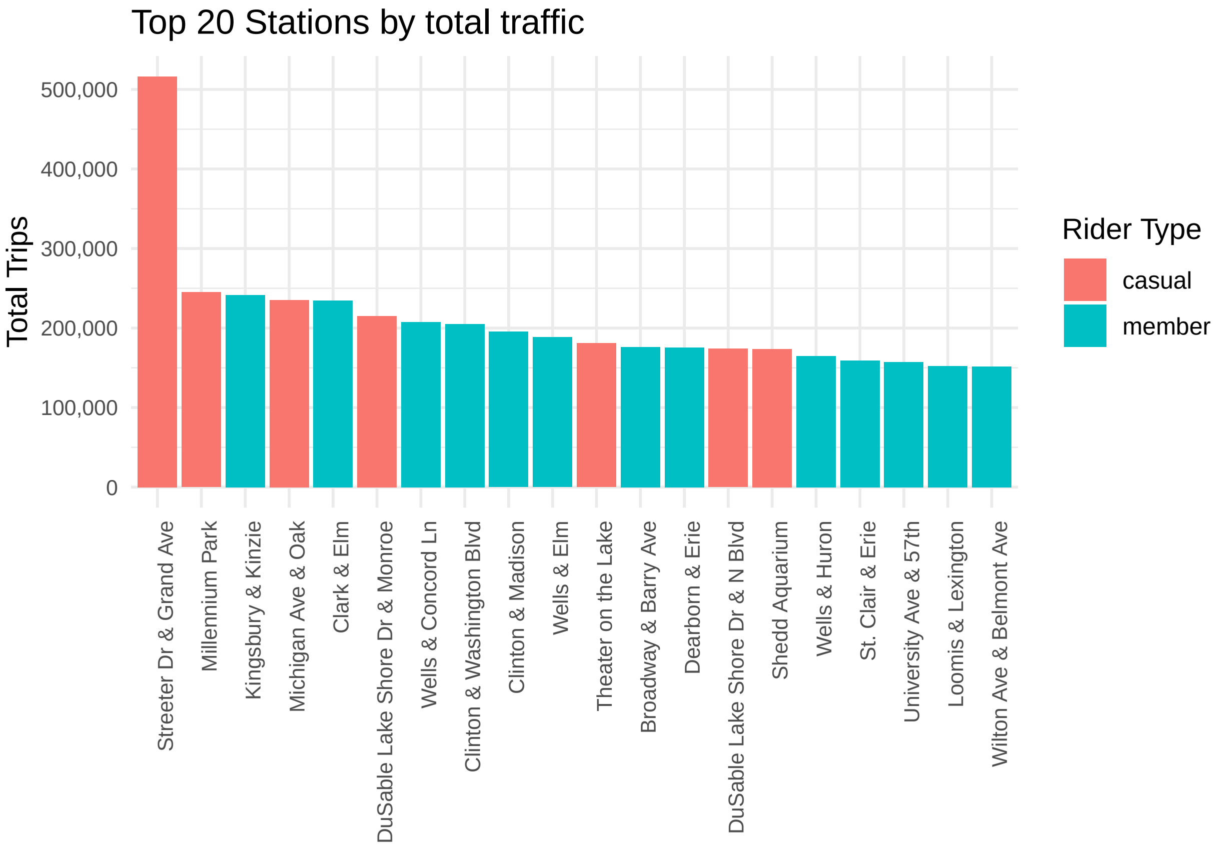
Members take more total trips than casual members, all week long and all year long. This suggests a simple explanation: Members ride for practical, habitual reasons, back and forth, such as commuting to and from work.

How long do they ride on each trip?





What can we learn from station data?



Station data underlines emerging casual rider hypotheses

Trip data show us that Casual riders take more rides on weekends and during warm months. Duration data shows us that casual riders also take much longer rides.

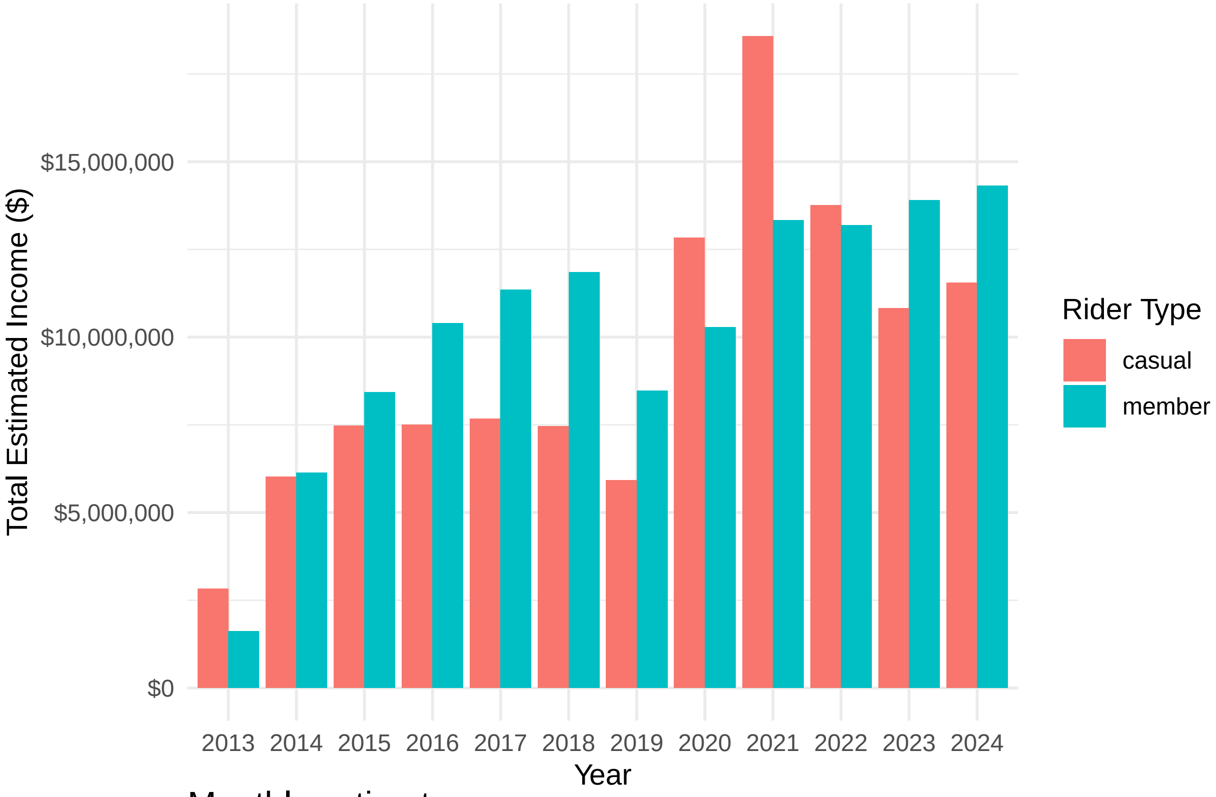
When we look at station traffic, we see that Streeter and Grand station and Millennium Park are the highest traffic stations. These stations are located at the heart of a waterside entertainment and public park district.

All of this suggests that casual riders use the service most often for entertainment and leisure activities rather than a daily commute.

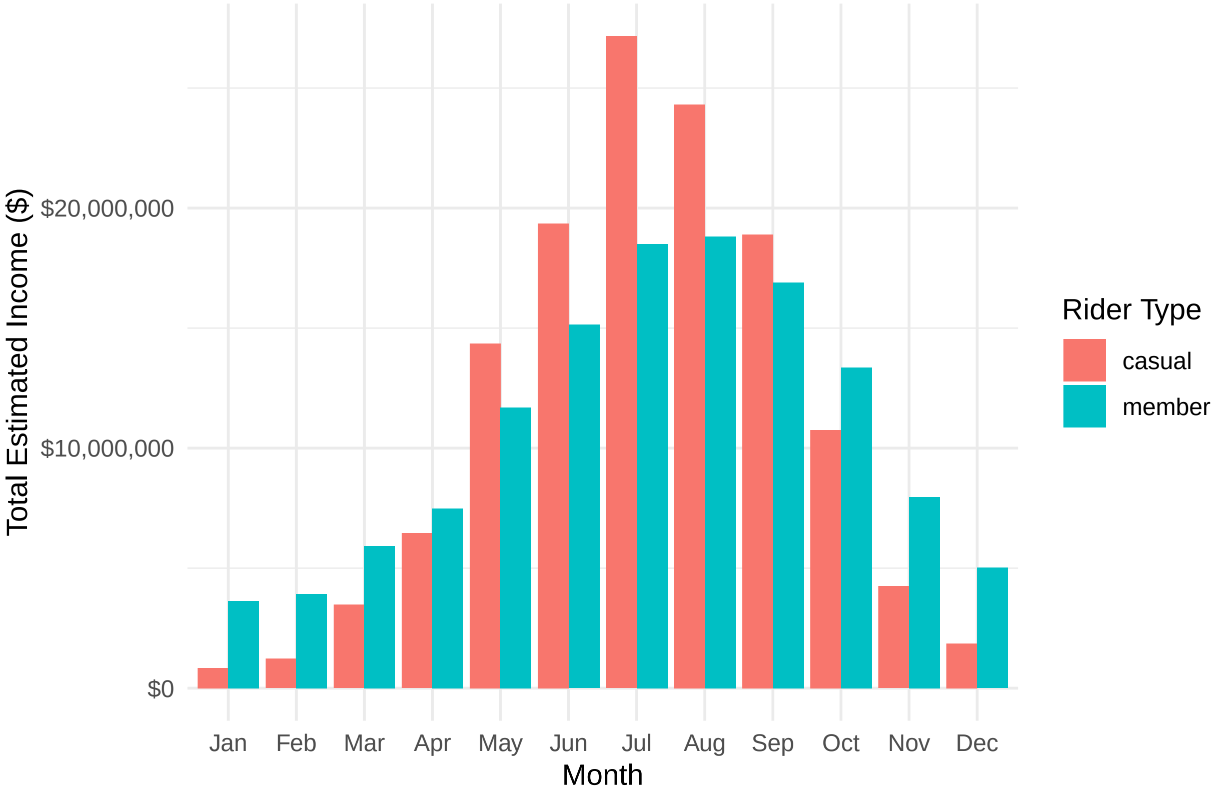
What is the bottom line?

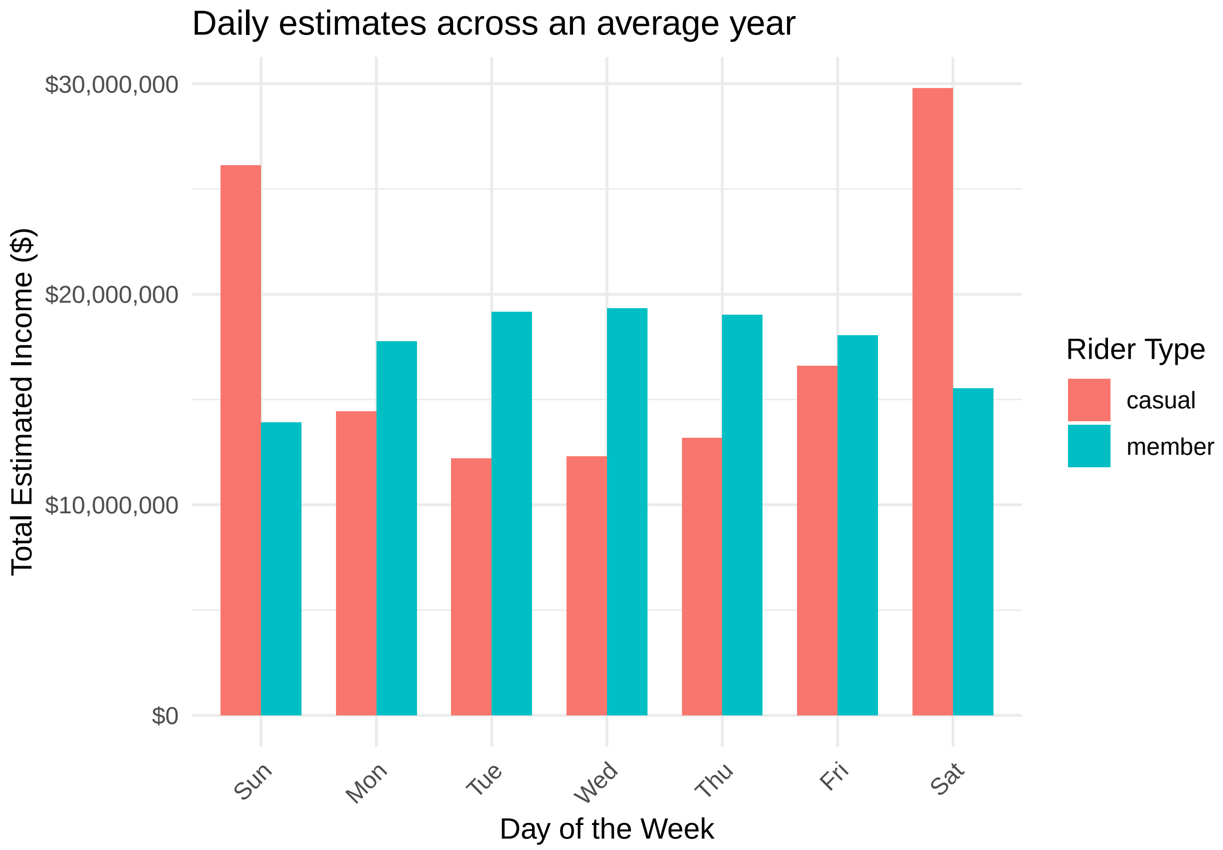
If Casual riders take fewer rides, but for a longer time, does the revenue balance against members, who ride more regularly, but for less time?

Annual income estimates



Monthly estimates across an average year





Answering questions

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

- Members ride more frequently but for less time. This suggests they use the service as part of their routine, to commute to and from their homes and work, or for running errands.
- Casual riders primarily take trips on the weekends, during warmer months, and for longer periods of time. Likely, the service is part of their relaxation and free time.

Why would casual riders buy Cyclistic annual memberships?

1. Do we want to commit to a conversion strategy?

For 5 of our most lucrative months, May to September, Casual riders are the leading source of revenue. Do we really want to alter the leading revenue group during the high revenue season so that we have a few more riders in the low revenue season? Is it worth the marketing dollars?

2. We should test the viability of conversion.

Casual riders may discover through targeted advertising and incentives that the service will fit into their routine. On the other hand, they may not want to think about their routine while riding for pleasure. These riders may find the service best suits them as a leisure activity rather than a commuter activity. Some may live to far to commute by bike, others may be tourists or seasonal residents. And keep in mind the value of the Casual rider to our historical revenue trends. We need to know more about those willing to bridge from Casual to Member.

How can Cyclistic use digital media to influence casual riders to become member

1. Test conversion effectiveness.

- Identify Casual riders and prepare a markeint campaign.
- Reach out to the Casual rider as the peak Casual months come to a close (August and September) as well as on Sundays as the work week begins. Emphasize benefits including the wide distribution of stations throughout the city, the health and convenience benefits of regular cycling, and the cost savings. Offer incentives for conversion.
- Track the conversion effectiveness and tweak messaging to improve conversion. Track the historical and post-conversion behavior of the converted users. Run another analysis against the general trends of Casual riders and Members. Are the Converted riders positively affecting revenue during peak months? During down months?

2. Consider encouraging the Casual rider pool

- Invest in advertising campaigns targeted at Casual and Non-riders who use competing transportation methods. Lean into the existing trends.
- Run campaigns during March and April with the aim of ramping-up for the high income months. d
- Run end-of-week campaigns and incentives to encourage even more Casual ridership on weekends.
- Emphasize city attractions, hidden gems, emerging hot spots, and the joys of longer trips with friends and family.
- Differentiate messaging between the city visitor and the city resident.

Questions and comments are welcome!