Cyclistic bike-share analysis

About the company

Cyclistic is a bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

Business task: *Design marketing strategies aimed at converting casual riders into annual members.*

About data and sources

The dataset contains the record of Cyclistic trip data from July 2013 to May 2022. The datasets are saved in a folder with 12 subfolders.
Ten of the 12 subfolders contain ride data for the years 2013 to 2022 and are named after the year in which each dataset was collated.

There are CSV and Excel files folders that contain datasets of Cysclistic bikes stations in CSV and xlsx formats respectively.
The data has been made available by Motivate International Inc. under this license. The data is reliable, original, comprehensive, current and cited.

The structures of the datasets are slightly different in terms of the number of columns and the unique identifiers for each ride. All datasets were made consistent with the structure of the 2020 dataset in terms of the number of columns, data types, and column arrangement.

Data Cleaning Process

The datasets were cleaned using Postgresql

- The columns of the datasets of each year were first rearranged and transformed into appropriate data types, using the 2020 dataset as a template so they can be merged into a single table.
- Each transformed dataset is saved as a new version of the original dataset

- All new versions of the transformed datasets were merged into a single table using the UNION operator which combines the rows of multiple tables and ensures no duplicate rows are appended.
- The member_casual column contained 5 categories: member, subscriber, customer, dependent and casual. I reduced the categories into member and casual by replacing all “subscribers” with “members” and all “customers” with “casual”, and lastly, drop rows with the dependent user types
- I added new columns to the data set;
    - year_month_day column from the started_at column (timestamp data type) using the CAST() clause to populate the field with only the date part of the started_at column
    - day_of_week column from the started_at column using the EXTRACT() clause to obtain values for every day of the week, with values ranging from 0 (Sunday) - 6 (Saturday)
    - ride_length column by extracting the difference between the start and end time of each ride in seconds

  These were done for effective aggregation of the data as the timestamp columns were too granular for analysis

- Lastly i deleted rows having:
    - start_station name as “HQ QR”: these are rides from maintenance activities
    - ride_length less that 0 sec
    - member_casual values = “dependent”

Analyzing the data

I started by using a simple query where I grouped the number of rides by date (year_month_day column) and user type (member_casual column) to explore how the number of member and casual rides differs by the day. The result was a long list as you would predict but it shows there’s no linearity between the number of rides causal and member rides. I saved this result in a spreadsheet for further visualization and aggregation using tableau.

Moving forward, I analyzed the average, minimum, maximum and median rides lengths of each rider type. There were outliers in the results as shown by the ridiculous values below:

member_casual	avg_ride_length	min_ride_length	max_ride_length	med_ride_length
casual	2259.71	0	14340041	1259
member	795.59	0	13561217	588

I computed the sample standard deviation of the ride lengths from which we specified upper and lower boundaries of the ride lengths, and kept only the 95% data closest to the average with this formula:

*(Average - Standard Deviation * 2) < DATA WE KEEP < (Average + Standard Deviation * 2)*
or
Lower boundary < DATA WE KEEP < upper boundary

Below are the results obtained

member_casual	avg_ride_length	min_ride_length	max_ride_length	med_ride_length
casual	1796.33	0	37225	1254
member	748.19	0	37212	588

As we can see, these results are more reasonable as casual riders are only able to purchase single-ride and full-day passes. We had maximum ride lengths longer than a day's ride from the previous table..

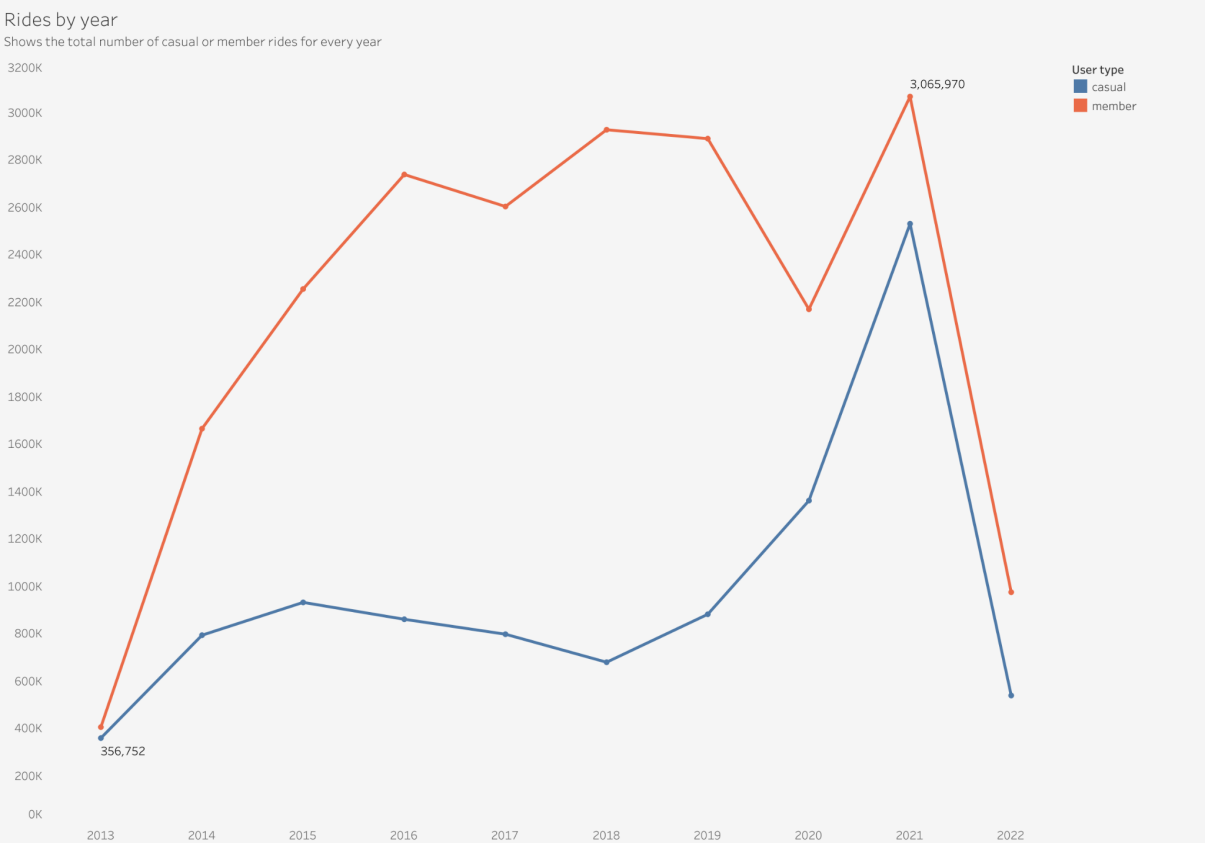
The above process was utilized in aggregating ride lengths for each day to obtain more accurate results. The result was saved in a spreadsheet.

Finally, we queried the top 20 bike stations by number of visits (how many times they appeared in the full table using the end_station_name column. I used the end_station_name column because when I queried the table for unique values for start and end-station columns, there were more values in the end_station column and each station serves as starting point of a ride or its point.

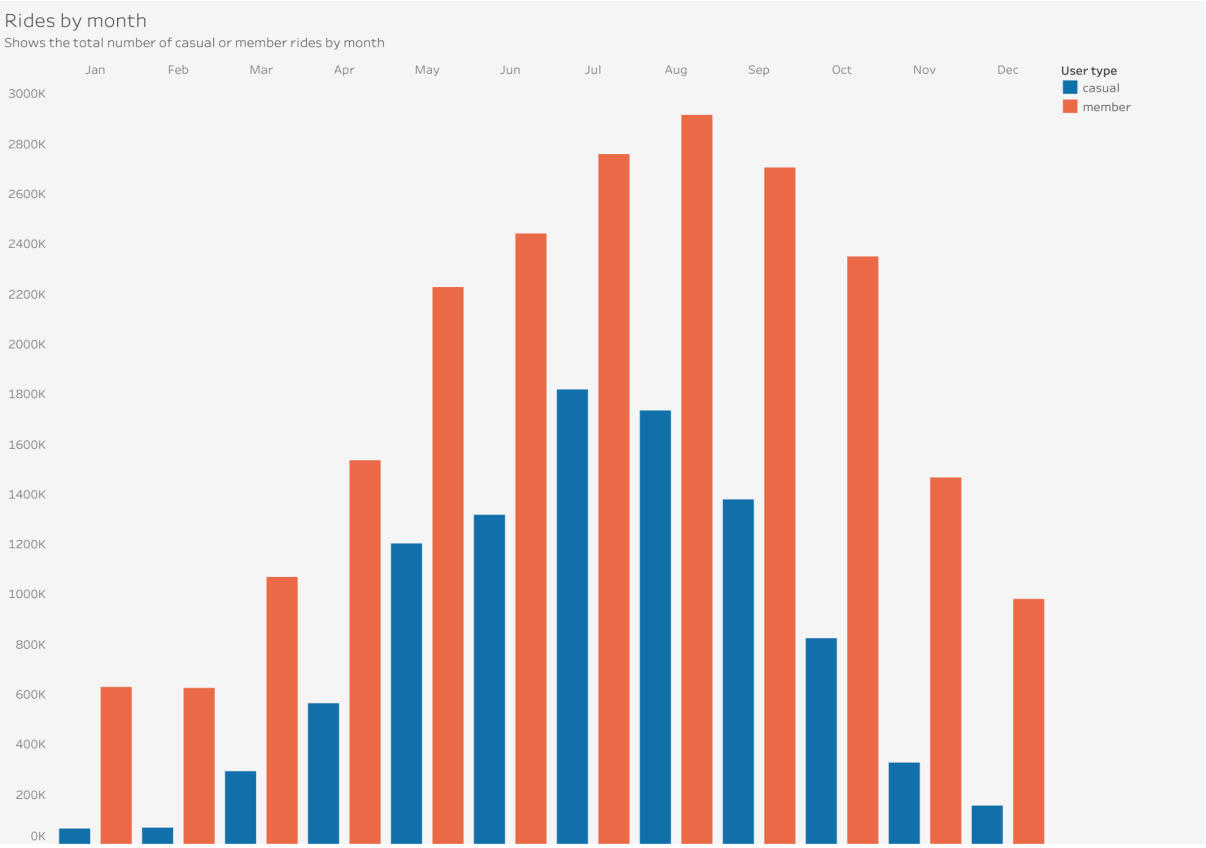
Visualizations and Insights

The uptrend in member and casual bikes from 2013 to 2021 signifies increasing demand for Cyclistic bikes. However, the steep decline in the number of rides from 2021 to 2022 is due to our having only 5 months of data for 2022.

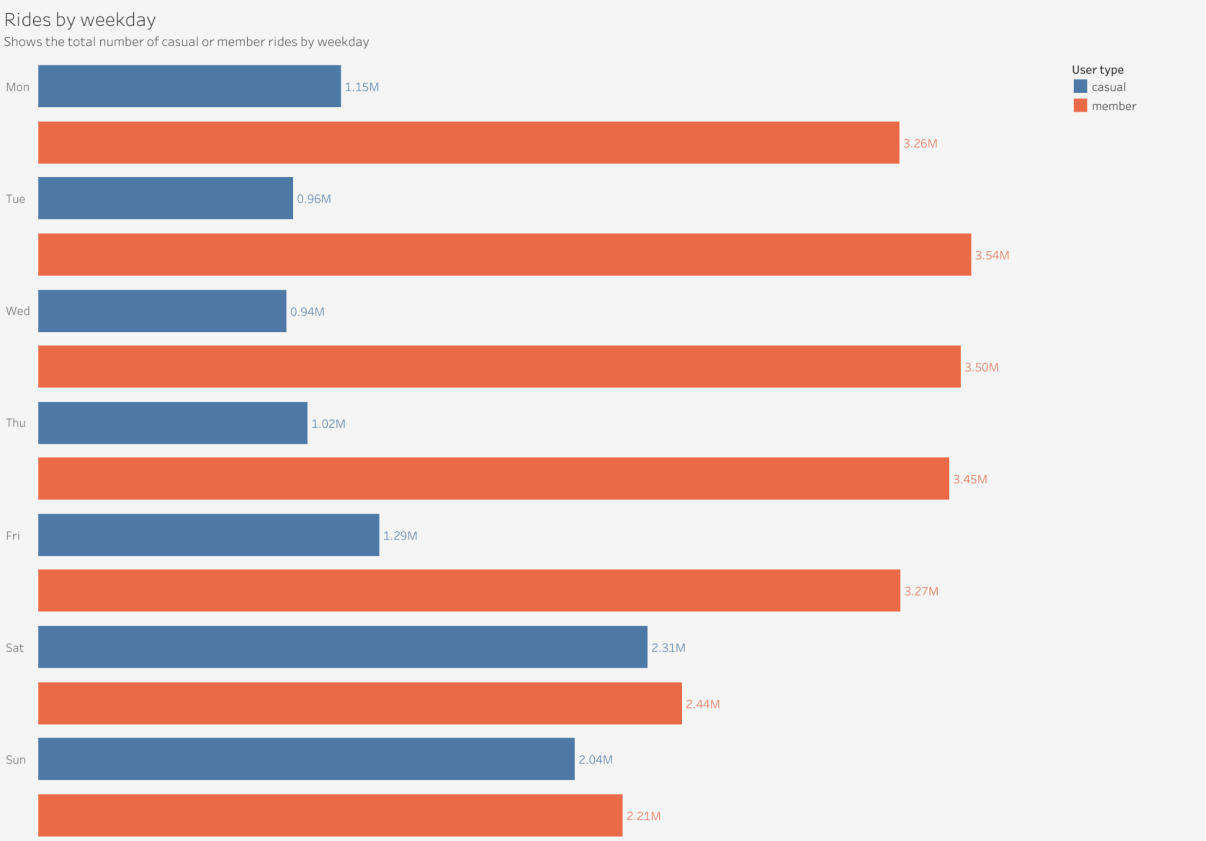
Each year, there's an average and maximum ride length of 87.13k and 1.32M minutes for casual riders, and also average and maximum ride lengths of 37.5k and 1.28M for member rides.



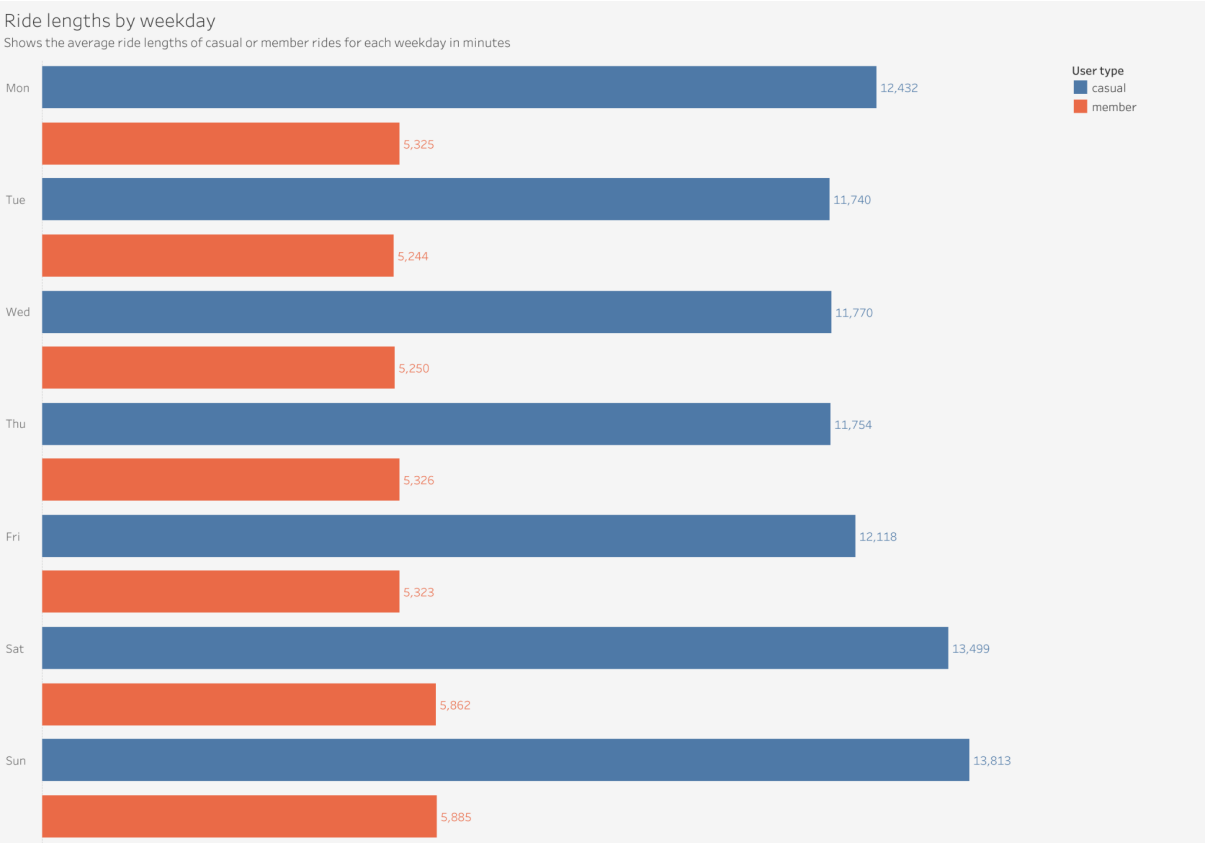
There's a significant increase in casual rides from late Spring (May) to early Autumn (September) each year.



Cyclistic members use bikes more during weekdays and less on the weekends while casual riders use cyclistic's bikes more on the weekends and less on weekdays.



Casual riders use bikes for twice as much time as members everyday.



click the link for interactive dashboard

Recommendations

- Set up a marketing campaign that runs from late spring to early autumn with casual riders as the target audience.
- Offer discounts on annual memberships.
- Create a survey plan and survey to collect data on casual riders' ride purposes, and bike-pass satisfaction levels. This will provide Cyclistic valuable insights into how casual riders and members use rides for different puposes, what weekend offers could be of great interest to casual riders and most effective for a marketting campaign.
- Considering there are larger numbers of member rides than casual rides on weekdays but longer durations of casual rides, this could mean that Cyclistic membership is purchased more by people who have jobs and is not simply affordable.
- Develop a more flexible pricing plan to make membership more affordable.

Conclusion

There are different ways in which casual users and members use Cyclistic bikes as I have established through my analysis. There could be other high-level insights to draw from this data through other different or more advanced means than the work I have done on this project.

There is information that was not provided which could improve the efficiency of our analysis; the prices of different ride passes and Cyclistic membership for example.

I shall explore other projects on this data for my personal development and find out different ways this work could be improved.