Google Data Analytics Capstone Project: Cyclistic Bike Sharing Data Analysis

by Devroop Banerjee

Introduction

"You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations."

This project will follow the steps of the data analysis process**: ask, prepare, process, analyze, share, and act.**

About the Company

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

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.

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital

media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Ask

This is the first phase of the data analysis process which comprises of asking and stating the issues that require solving as well as identifying the stakeholders and understanding their expectations. The three questions that will guide the rest of this project as well as the future marketing program are:

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

Business Task

From the questions asked, it is clear that the business task is to analyze the difference in the usage between casual riders and annual members. Upon analysis, what can Cyclistic do differently in terms of its business and marketing strategy in order to increase the conversion rate from casual riders to annual members.

Key Stakeholders

Lily Moreno: The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.

Cyclistic Marketing Analytics Team: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy.

Cyclistic Executive Team: The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

Prepare

This phase consists of the collection of data, ensuring that it meets certain standards in terms of quality, integrity, licensing, bias.

Where is your data located?

Since Cyclistic is a fictional company, it does not have its own data. The datasets provided belong to a company called Motivate International Inc. under the following license. The datasets have been made public however in order to ensure data privacy, any information that identifies the riders have been redacted.

How is the data organized?

The data used covers the following period: June 2021 to May 2022, where each month consists of millions of entries but only thirteen columns.

Are there issues with bias or credibility in this data? Does your data ROCCC?

There are no issues with bias or credibility with this data. Since the data is **R**eliable, **O**riginal, **C**omprehensive, **C**urrent, and **C**ited, it does ROCCC.

How are you addressing licensing, privacy, security, and accessibility?

This link contains the data license agreement. The datasets are accessible to everyone since they have been made public however in order to ensure data privacy, any information that identifies the riders have been redacted.

How did you verify the data's integrity?

Upon searching Motivate International Inc, it turned out to be a valid and popular biking service company based in New York. Next the dataset was searched on a reputable site such as Kaggle. Since the same dataset was available, its integrity can be verified.

How does it help you answer your question?

Since the data has not been analyzed, it does not answer any questions yet. However, knowing that the data has been obtained from a credible source and that the integrity of the data can be verified means that any insights obtained post analysis will be truthful and not misleading.

Are there any problems with the data?

As with any dataset, this one contained null values that were deleted. While this gives us less data to work with, it also reduces bias in our analysis.

Process

This phase consists of cleaning and modifying the data, ensuring it is tailored for our use.

What tools are you choosing and why?

I have used Rstudio for processing the data since it has numerous libraries that are essential for this stage of the data analysis process. Rstudio is also capable of handling the large amount of data that I have been provided with.

Have you ensured your data's integrity?

The data's integrity has been ensured. The information provided has not been altered, but the datatypes have been changed or additional columns have been created based on the information provided.

What steps have you taken to ensure that your data is clean?

The following are the steps taken to clean the data:

- Combined all twelve individual months into a single dataframe called "Combined_trips".
- Created new columns to extract the day, month, year and day of the week for each ride entry.
- Created new column called "ride_length" which calculates the difference between "end_time" and "start_time" in order to find the duration of each ride entry.
- Changed datatype of "ride_length" from time to numeric.

- Deleted all rows containing null entries, as the incomplete data entries can not be used and might lead to anomalous points in our visualization, or skew the results of the analysis.
- Deleted entries where ride_length < 0, since this is invalid data or the numbers entered in start_time and end_time are incorrect.
- Created new dataframe excluding the longitude and latitude columns, since they are not used in the analysis. This reduces clutter in the data.
- Created a new dataframe where “started_at” and “ended_at” columns were combined to get an aggregate route for each trip.
- Organized the months chronologically from June 2021 to May 2022 and the days of the week from Sunday to Saturday.

How can you verify that your data is clean and ready to analyze?

Performing operations and calculations on the data does not give error results. The layout of the data has been simplified for easier navigation and verification. Noisy data has been eliminated.

Have you documented your cleaning process so you can review and share those results?

Yes, the cleaning process has been documented here.

Analyze

This phase consists of analyzing the cleaned data to find useful patterns, relationships, and trends so that we can start making data driven decisions in order to get a step closer to the business task.

How should you organize your data to perform analysis on it?

The null values have been eliminated and the data has been organized according to month and days of the week.

Has your data been properly formatted?

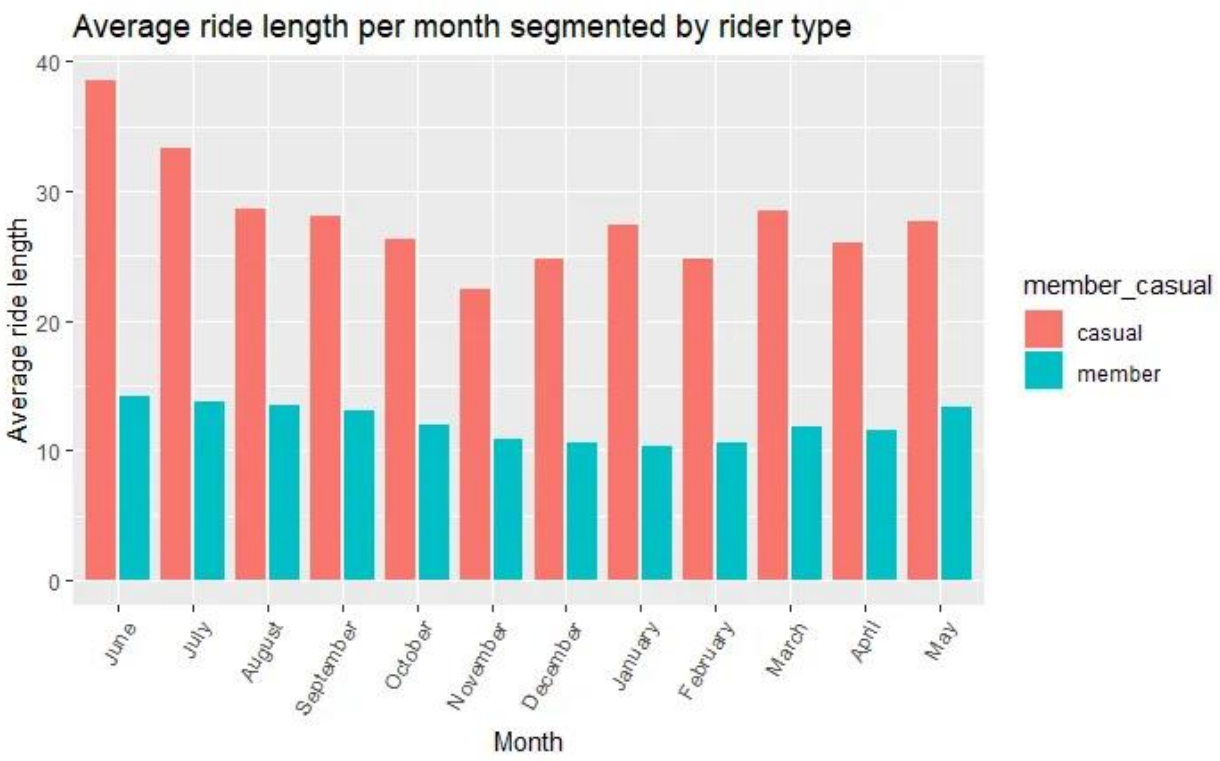
The data has been properly formatted for this analysis. The datatypes of each column and the dimensions of the dataframes have been checked throughout the analysis to maintain consistency and reduce chances of error.

What surprises did you discover in the data?

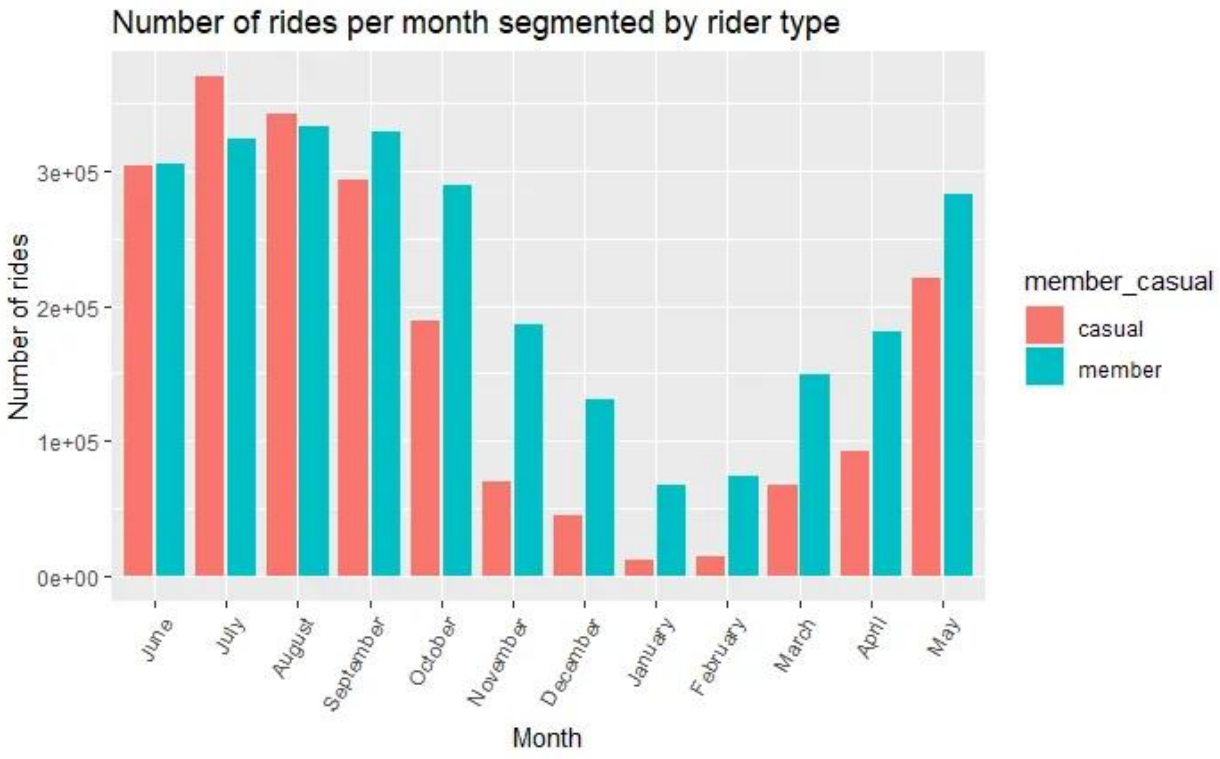
It was surprising to observe that casual riders rode for nearly twice as long as annual members and that members did not use the docked bike type at all.

What trends or relationships did you find in the data?

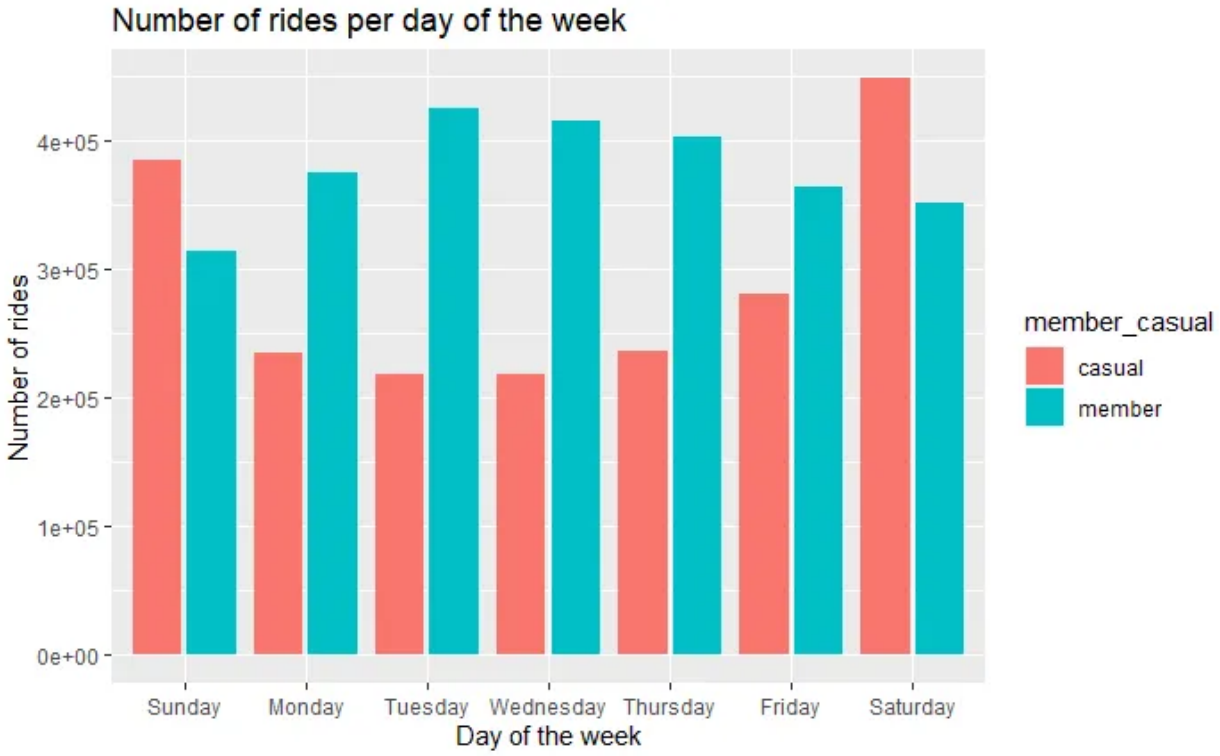
- Casual riders had a much higher maximum average ride length than that of members. This makes sense because despite having six hundred thousand more records for members as compared to casual riders, the latter rode for twice as long as members.



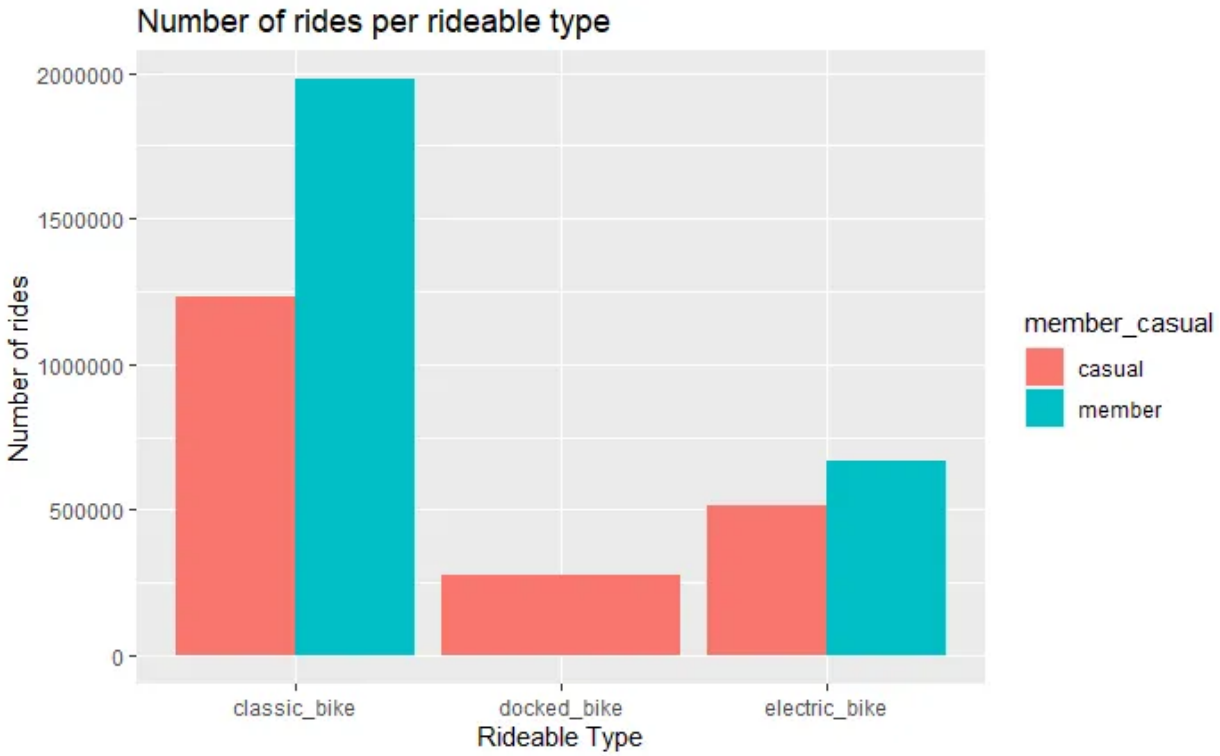
- Casual riders experienced a peak in July while members experienced their peak in August. However, both casual and member riders used Cyclistic services the least in January.



- Casual riders rode the most on Saturdays but members rode the most on Tuesdays.



- The most commonly used cycle type was classic bike in both groups but members did not use docked bikes at all.



How will these insights help answer your business questions?

These insights will help the stakeholders decide how to alter their marketing/promotional strategies, eg: special membership offers targeting casual riders on specific days of the week or times of the year. It could even influence them to discontinue certain types of cycles, such as docked bikes, since members do not use them.

Share

All the findings are relayed to the relevant stakeholders.

This phase was completed using RStudio and Tableau. Click here to view the tableau visualization.

Cyclistic Bike Sharing Dashboard 1

Min/Max/Avg/Median Ride Length

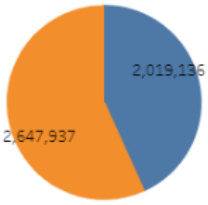
Member Casu..	Count of Member Casual	Min. Ride Length	Avg. Ride Length	Max. Ride Length	Median Ride Length
casual	2,019,136	0	30	55,944	16
member	2,647,937	0	13	1,496	9

Rider type
- casual
- member

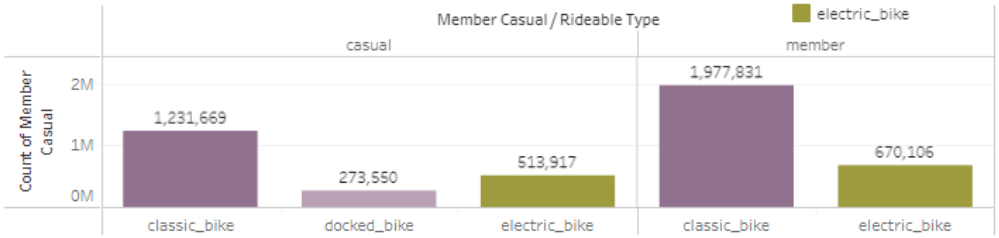
Total number of riders
4,667,073

Rideable Type
- classic_bike
- docked_bike
- electric_bike

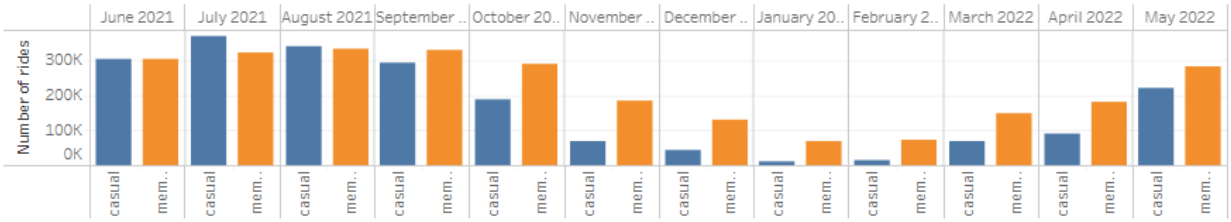
Number of riders by type



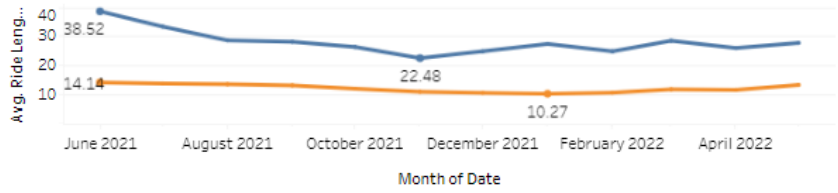
Distribution of rideable types by riders



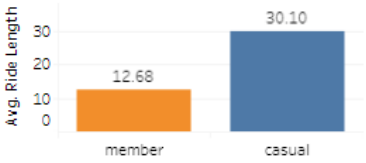
Number of rides per month by rider type



Average ride length per month by rider type



Average ride length by rider type



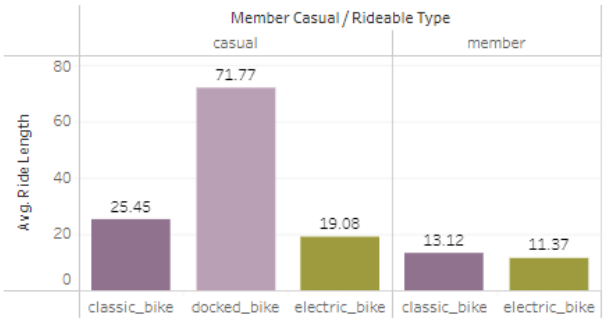
Cyclistic Bike Sharing Dashboard 2

Min/Max/Avg/Median Ride Length

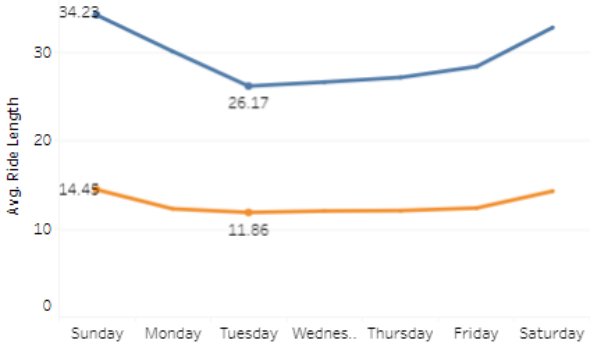
Member Casu..	Count of Member Casual	Min. Ride Length	Avg. Ride Length	Max. Ride Length	Median Ride Length
casual	2,019,136	0	30	55,944	16
member	2,647,937	0	13	1,496	9

Member Casual: casual, member
Rideable Type: classic_bike, docked_bike, electric_bike

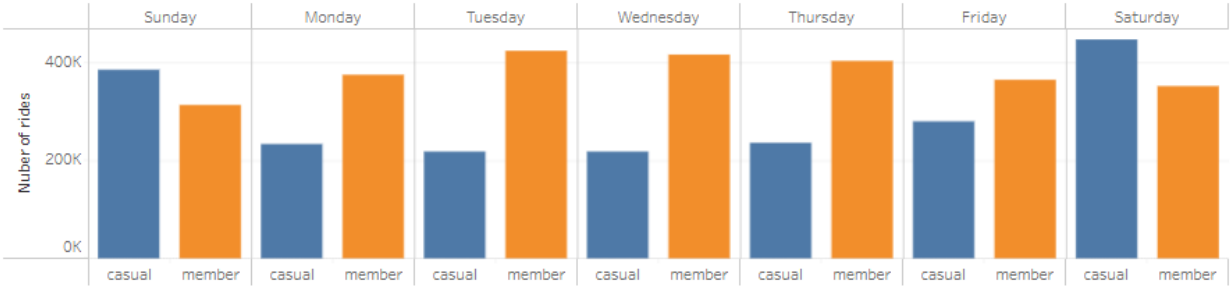
Average ride length per riderable type by rider type



Average ride length per week by rider type



Number of rides per week by rider type



Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently?

Yes, this question was answered as a result of my findings. I looked into questions pertaining to the nature of user, types of bikes used, duration of bike ridden, popular routes as well as trends in bike usage over time.

What story does your data tell?

The data tells me that there are more members than casual users however the latter have a much higher average ride length per month than members. Members generally use classic bikes followed by electric bikes, but never docked bikes. Casual riders follow a similar trend but a small portion of them use docked bikes. Average number of rides per month gives a very interesting insight. Both types of users show a similar trend where the highest number of rides are during the months of July and August (especially for casual riders) before dropping during the winter months from November to February. This could mean that the main reason why casual riders are not members is because they only use the bikes during the summer, maybe for cruising the city. Upon analyzing the number of rides for each hour of the day and each day of the week, it can be inferred that the casual riders use the bikes mainly for

leisure since their number of rides are highest on the weekends. Whereas peak riding hours for members are 8am (to go to work) and 5pm (to return from work); their number of rides are higher on the week days than the weekends.

How do your findings relate to your original question?

My findings tell me the differences in how casual riders and members use their bikes differently, which was the original question.

Who is your audience? What is the best way to communicate with them?

My audience is the stakeholders. The best way to communicate with them would be to hold a meeting where these visualizations are on a presentation and I am explaining my findings to them.

Can data visualization help you share your findings?

Yes, data visualization can help share my findings as the trends in the graphs help demonstrate the points that I would make. The changes that would need to be implemented are based on these trends as well.

Is your presentation accessible to your audience?

Yes, the link to my tableau visualization in presented here. The R markdown file is available here.

Act

What is your final conclusion based on your analysis?

There are certain trends that have been identified from the data provided and these results give us insight into the changes that can be implemented in order to convert more casual riders to members. The recommended changes are discussed in the next part of this section.

How could your team and business apply your insights?

There are a few recommendations I would make in order to convert casual users to members:

- An introduction of monthly/seasonal plans for casual riders would be prudent as the data indicates that a lot of casual riders are affected by season/weather, ie:- more rides in summer, between June to September, as compared to fewer rides in the winter, between November to February. Cyclistic could offer a free month's worth of trial to casual riders during the winter months so that the perks of a membership is realized and memberships are taken up during the summer for the sake of ease and convenience.
- Since the most popular start and end stations have been identified, these stations can be used for targeted physical ads and campaigns. Strategically placed billboards and posters introducing casual riders to new membership plans and offers could prove to be influential.
- Using ads that highlight financial incentives for members showing how a membership is cheaper in the long run as opposed to renting bikes casually for a few months may persuade casual riders. Coupons/bonuses could encourage more frequent rides. A weekly scoreboard showing the number of rides and average ride length could be used to decide winners at the end of the month who would get a seasonal pass or annual membership.

What next steps would you or your stakeholders take based on your findings?

Based on my findings, the next steps should be the introduction of seasonal passes, strategic placements of billboards and posters, spotlighting financial incentives of membership over casual usage. If the casual riders can be encouraged to use bikes more often then they will realize the value of having an annual membership.

Is there additional data you could use to expand on your findings?

Additional data such a longer or different time period would be insightful since 2022 was the tail end of the pandemic and quarantine as well as remote work became a big part of life which eliminated the need for bikes.