**Case study: How does a bike-share navigate speedy success?**

****

**Introduction**

Welcome to my capstone project, part of the Google Data Analytics Professional Certificate. To complete this project and to answer the business questions, I will be using the knowledge and skills I have gained through the seven courses that I have completed as part of this certificate. To make the process effective and reliable I will be following the steps of the data analysis process: **Ask, Prepare, Process, Analyze, Share, and Act.** To complete this project, I will use Microsoft Excel and BigQuery for the data preparation, cleaning, and to organize the data for analysis. I will use Tableau for data visualization to make a final analysis and provide recommendations.

**Scenario**

I am a junior data analyst working on 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, the team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, my 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.

**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 at any time. 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, Director of Marketing 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 solid opportunity 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. To do that, however, the team needs to understand better 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.

**Steps followed for the Data Analysis process:**

1. **ASK:** This process involves understanding the context, considering the key stakeholders, and identifying the business task by asking the right questions.

The following three questions will guide the future marketing program:

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

b) Why would casual riders buy Cyclistic annual memberships?

c) How can Cyclistic use digital media to influence casual riders to become members?

As per the project roadmap, I have been assigned the first question to answer by my manager.

**The key stakeholders:**

* **Lily Moreno:** The director of marketing and my 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.

**Identifying the Business Task:**

Analyze how annual members and casual riders use Cyclistic bikes differently by using historical data of the last 12 months.

**Context:** The comparison between the data as mentioned in the task above will play a key role in devising a marketing strategy to convert the casual riders to annual members.

1. **PREPARE:** This process involves accessing the database, asking yourself a few questions to determine the credibility of the data, identifying how the data is organized, and performing steps to sort/filter the data.

Below are the guiding questions to determine the credibility of the data:

1. **Where is the data located?**

It is a public data set available on AWS [here](https://divvy-tripdata.s3.amazonaws.com/index.html). I will be using the historical data for the 12 months (March 2023 - February 2024).

1. **How is the data organized?**

The data is organized by month and year and is stored as a zip file.

1. **Are there issues with bias or credibility in this data?**

No, there are no issues with bias or credibility in this data.

1. **Does your data ROCCC?**

ROCCC stands for Reliable, Original, Comprehensive, Current, Cited

Reliable: Yes, the data is reliable and credible

Original: Yes, the data has been validated by the source

Comprehensive: Yes, the data has all the elements required for the data analysis process

Current: Yes, the data is current as it has been constantly updated to the current month.

Cited: Yes the data has been cited by the source

1. **How are you addressing licensing, privacy, security, and accessibility?**

The data has been made available by Motivate International Inc. under this [license](https://divvybikes.com/data-license-agreement).

Data-privacy issues prohibit you from using riders’ personally identifiable information.

Hence, I won’t be able to connect pass purchases to credit card numbers to determine

if casual riders live in the Cyclistic service area or if they have purchased multiple

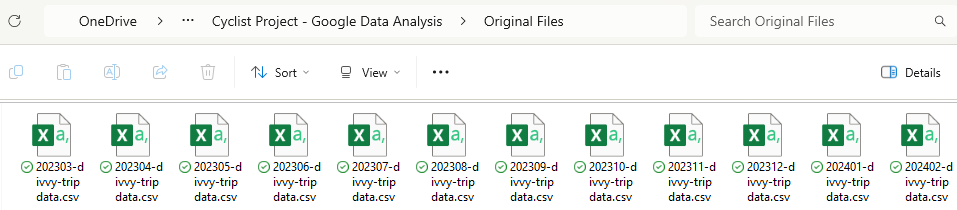
single passes.

After answering the above questions, the next steps are to Download and Prepare the

Data for analysis. Preparing data will involve observing the data and determining if the

data needs to be combined/cleaned or organized.

The data I have downloaded is for 12 months: March 2023 - February 2024.



The files are saved in both .csv and .xlxs format. To observe the data I used the .xlxs files.

I performed some basic calculations to understand the range of the data.

Below are a few of my observations about the data:

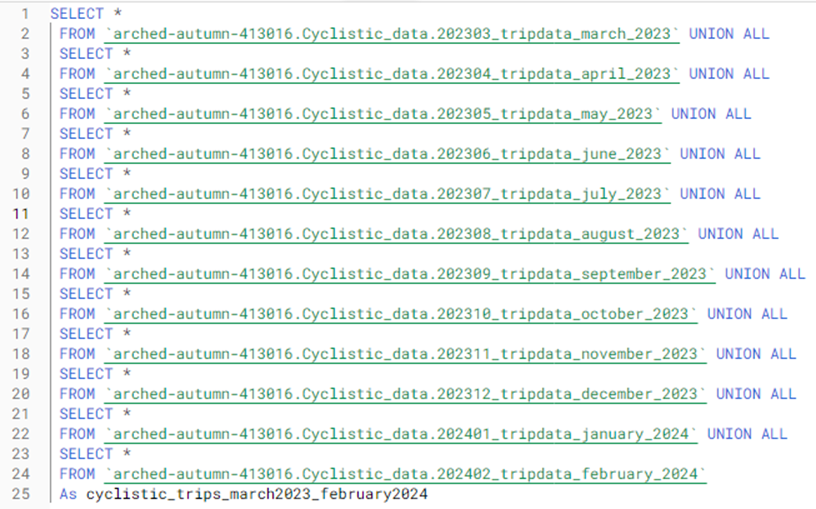
1. There are 13 columns in the files.
2. There are three types of bikes: Classic Bike, Docked\_Bike and Electric Bike
3. There are started\_at and ended\_at time stamps which can used to find the duration of each trip.
4. There are a few blank cells in the start\_station\_name and end\_station\_name. While working in SQL, I will determine if the data with blank cells needs to be removed from the data set.
5. There are no duplicate values in any of the files.
6. The combined file size of all 12 months will be too big to be cleaned and organized in Excel but these observations have helped to understand the first few queries that need to be run using SQL to clean and organize the data.

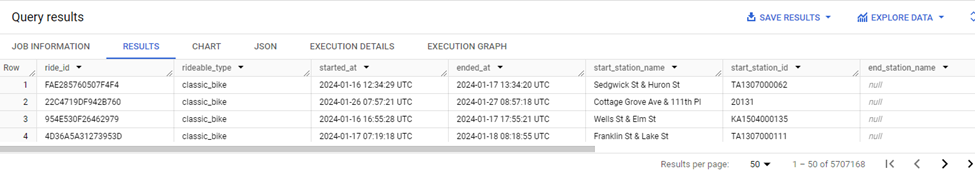
As I will be processing and analyzing the data using SQL in BigQuery, I uploaded all the

files in BigQuery and combined them into one table using the **UNION ALL** query as below.

**The combined data set has 13 columns and 5707168 rows.** I saved the data set as a

new BigQuery table **“data\_12months\_2023march\_2024february “**

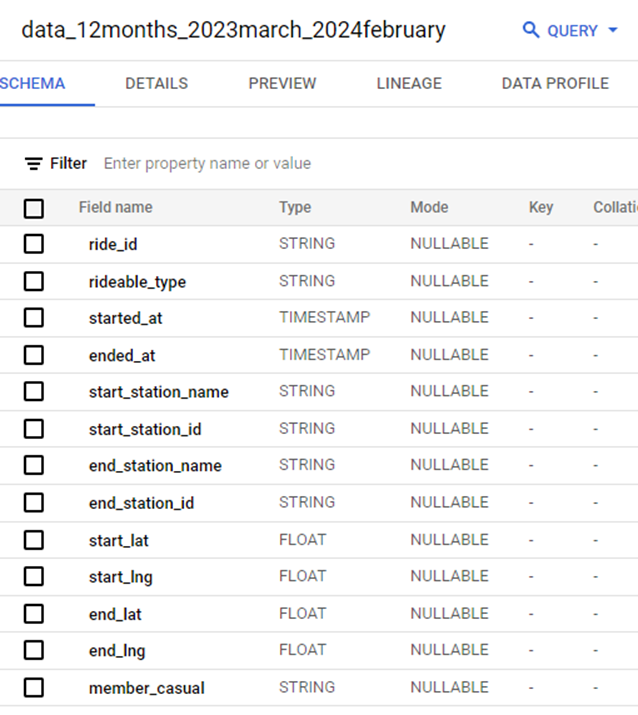




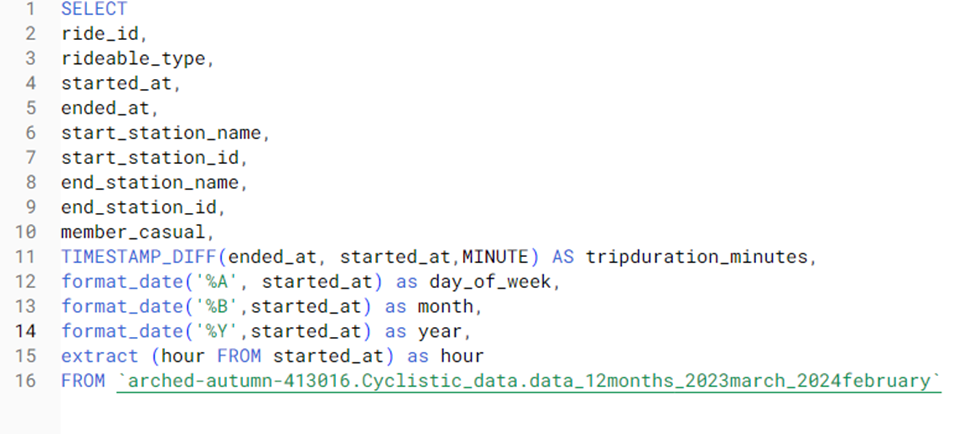
1. **PROCESS:** This process involves connecting the business task to the data analysis process, cleaning the data, documenting the data cleaning process, and saving the final data set that will be used for the analysis.

**Step:1** I combined the table of 12 months as “**data\_12months\_2023march\_2024february**”

to make changes and updates for analysis. Below is the schema for the table.

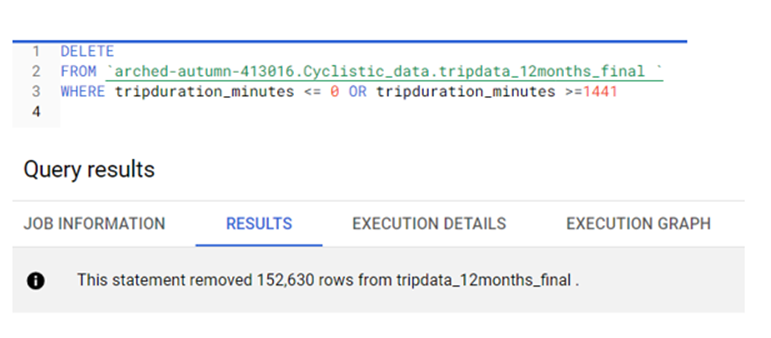


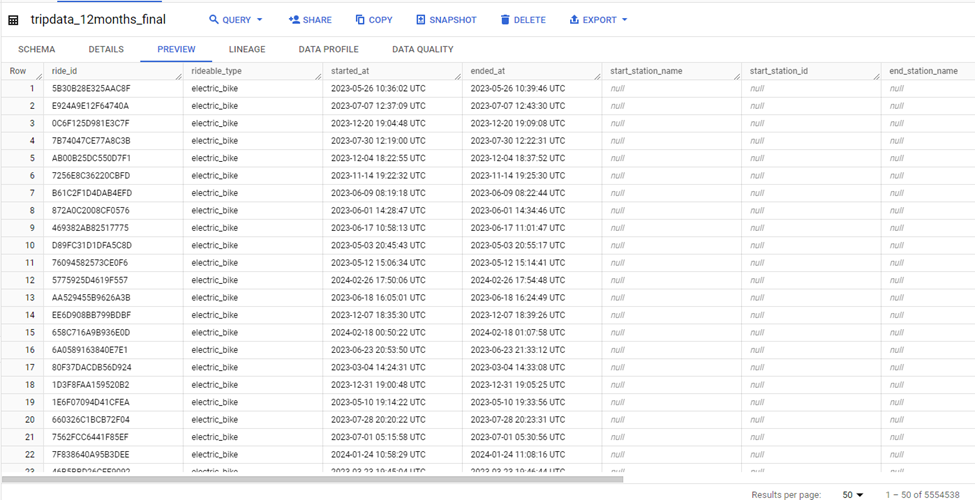
**Step:2** I created new columns **tripduration\_mintues**, **day\_of\_ week**, **month**, **year** and **hour** by running the query below.



**Step:3** I saved the new table with the additional columns as “tripdata\_12months\_final”.

**Step:4** I observed that some cells in column tripduration\_minutes are equal to “0” and some longer than 24 hours (1440 minutes) which might not provide a correct average trip duration for all rides. I ran a query to delete rows with tripduration\_minutes greater than a day or equal to zero. The query deleted 152,630 from the table “tripdata\_12months\_final”. Now, the total number of rows in the table are 5,554,538.

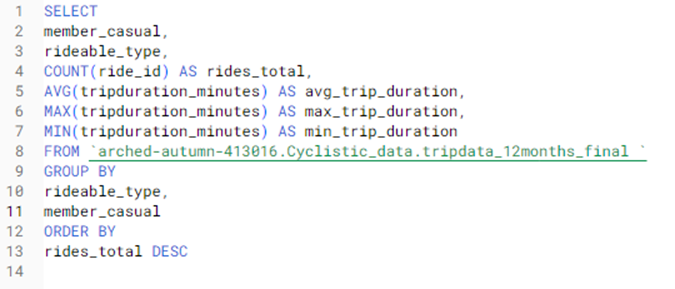


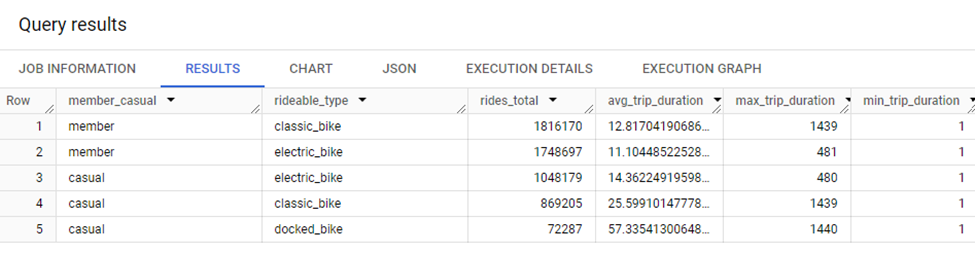


Some cells have null as start\_station\_name and end\_station\_name but the number of rows with **null** in the columns is quite significant so I did not delete the rows as deleting these rows would significantly reduce the data set.

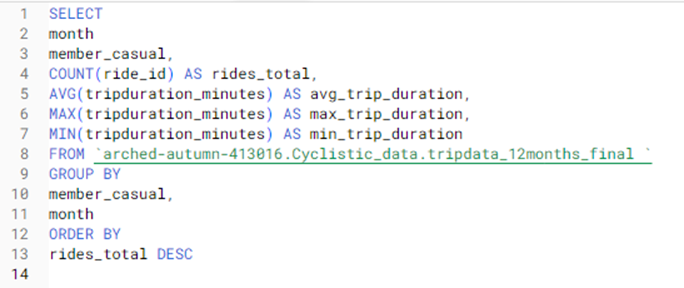
1. **ANALYZE:** In this phase, I analyzed existing relationships across columns to identify key trends by extracting small data sets from the master table. These tables helped me understand the difference between the riding habits of casual riders vs member riders.

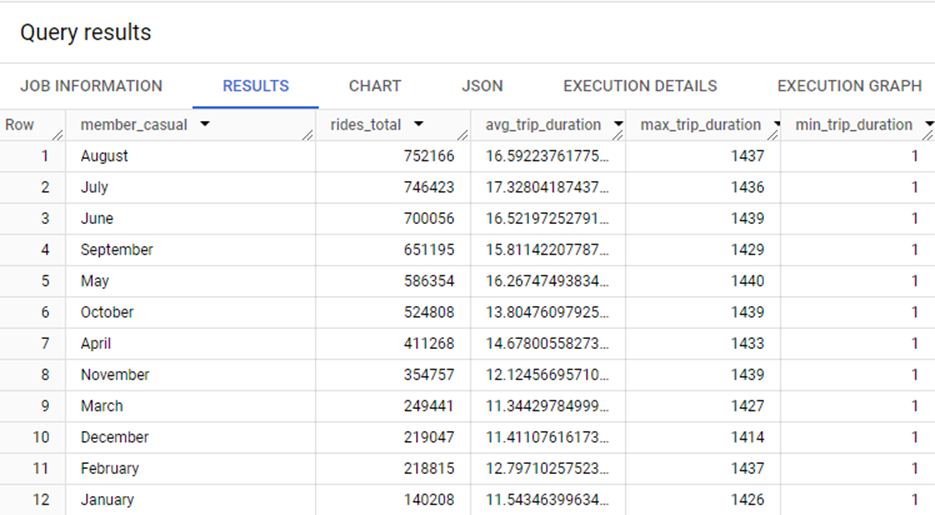
**Step:1** I ran a query to find the ride total, average, maximum and minimum trip duration for different ride types by casual and member riders.



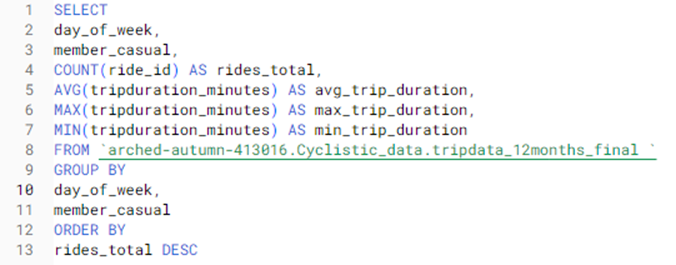


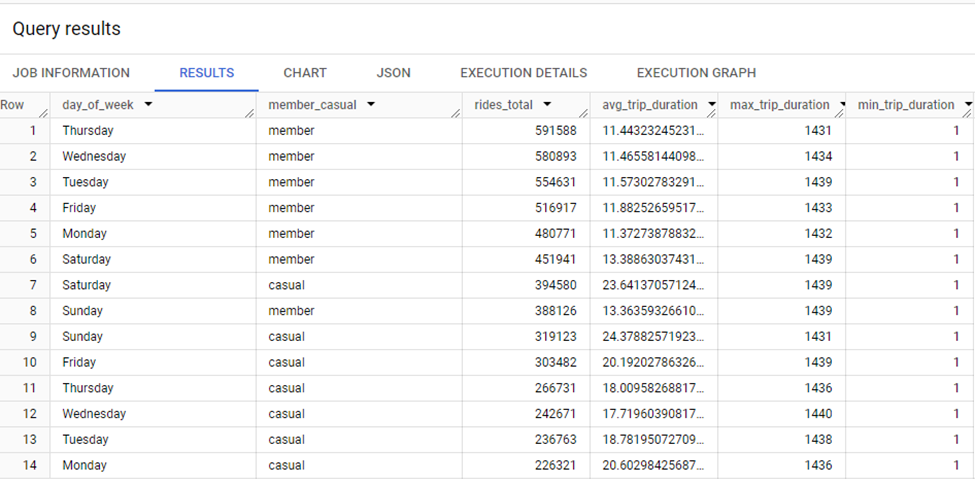
**Step:2** I ran a query to find the ride total, average, maximum and minimum trip duration for each month by casual and member riders.



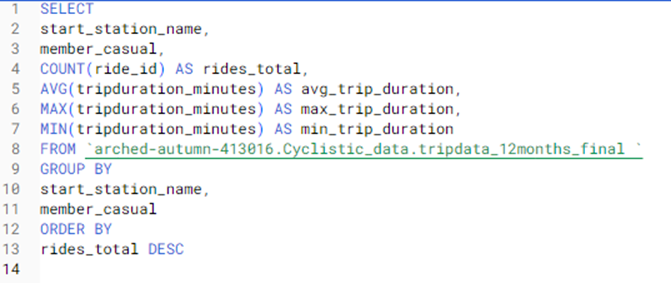


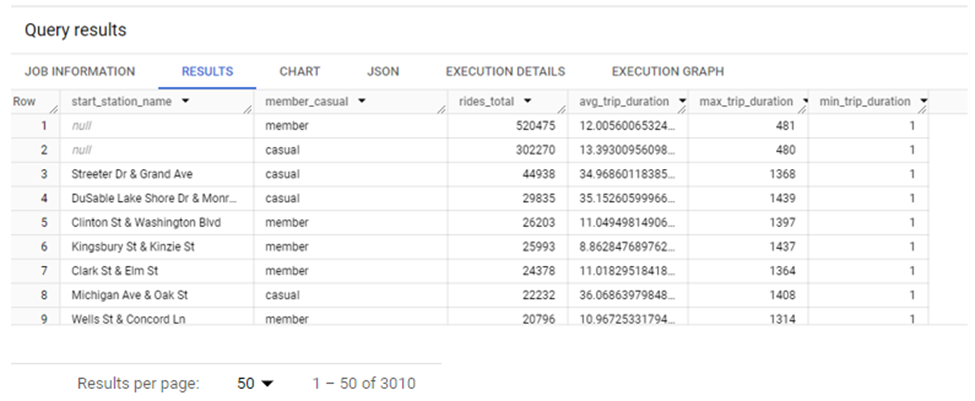
**Step:3** I ran a query to find the ride total, average, maximum and minimum trip duration for each day of the week by casual and member riders.





**Step:4** I ran a query to find the ride total, average, maximum and minimum trip duration at each start station by casual and member riders.



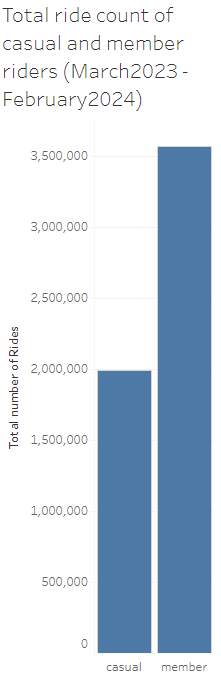


1. **VISUALIZE AND SHARE:** In this phase of data analysis, I visualized key relationships and trends using Tableau. These visualisations will help the tables observed using SQL queries above, easy to understand and interpret to derive analysis.

**Compared number of rides between Casual vs Member riders:** The visualization below

highlights that frequency of bike usage among member riders is more compared to the

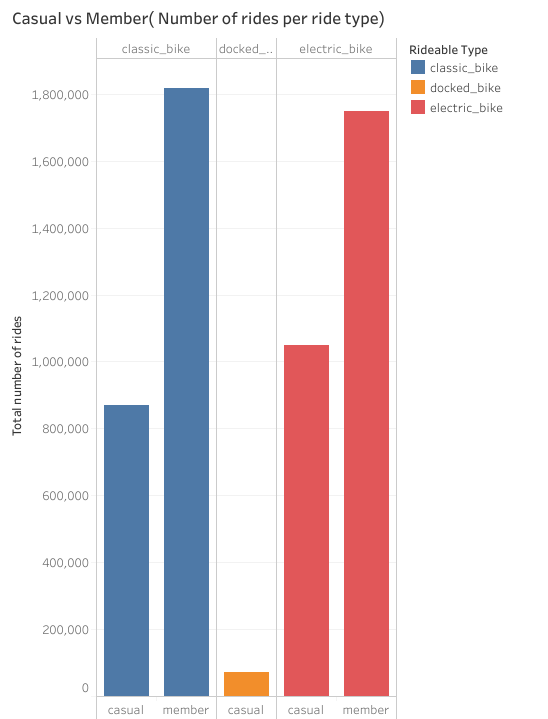
casual riders.



**Compared number of rides per ride type between casual and member riders:**

The visualization below shows that classic bikes are preferred by member riders, while casual

riders prefer using electric bikes.

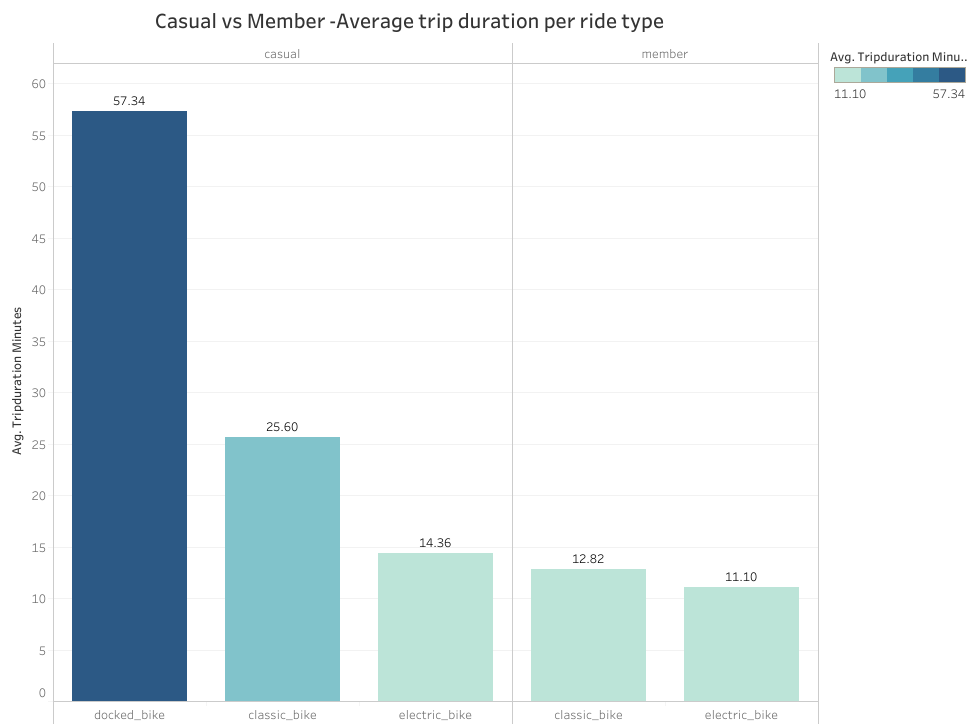


**Compared average trip duration per ride type between casual and member riders:**

The visualization indicates that casual riders tend to have longer average trip durations when using

classic bikes, whereas member riders typically have shorter average trip durations when using

electric bikes.



**Compared number of rides per month between casual and member riders:**

The visualization below highlights that August 2023 recorded the highest number of bike rides

among member riders with July 2023 closely trailing behind. In contrast, July 2023 took the lead for

casual riders closely followed by August 2023. January 2024 had the least number of bike rides by

both casual and member riders.



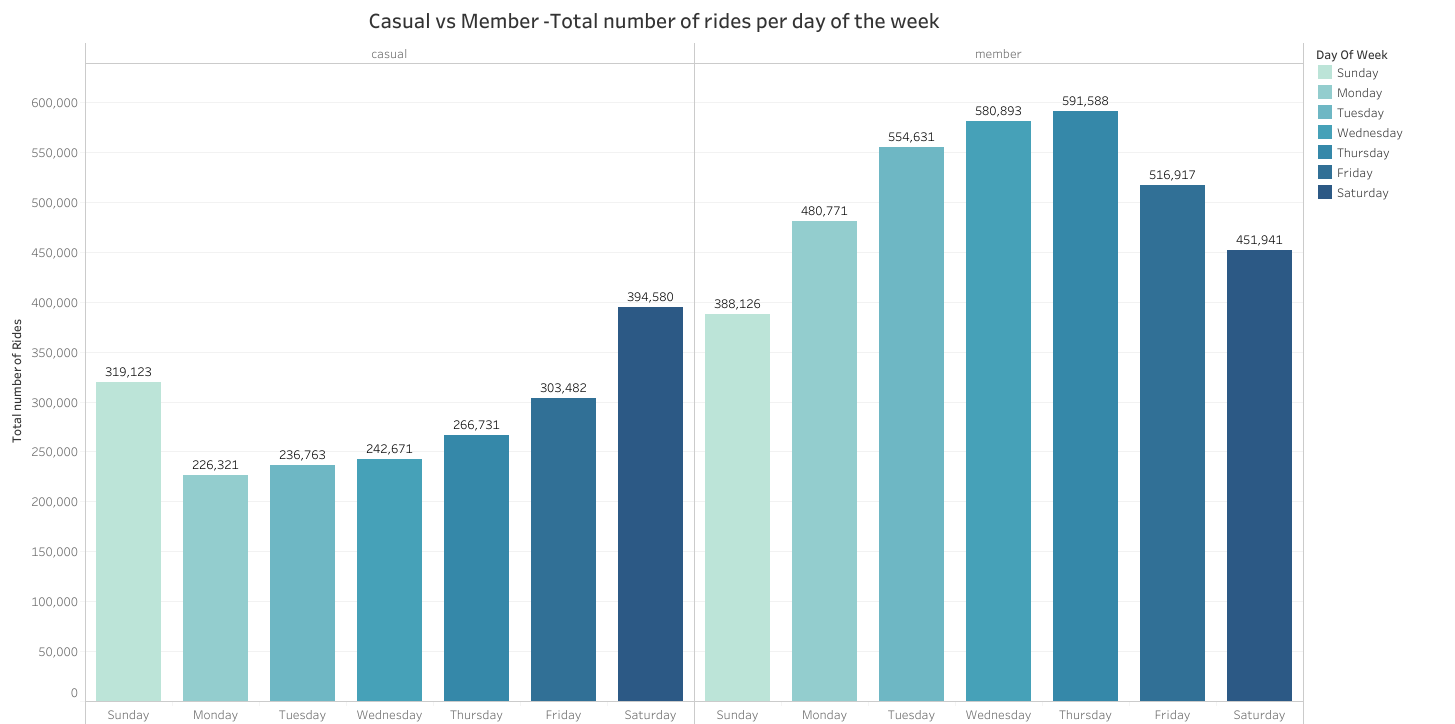
**Compared number of rides per day of the week between casual and member riders:**

The visualization below indicates distinct day preferences between casual and member riders.

For casual riders, Saturday ranks highest in popularity, whereas Monday experiences the

lowest activity. On the other hand, member riders favor Thursdays, with Sundays seeing the

least number of rides.

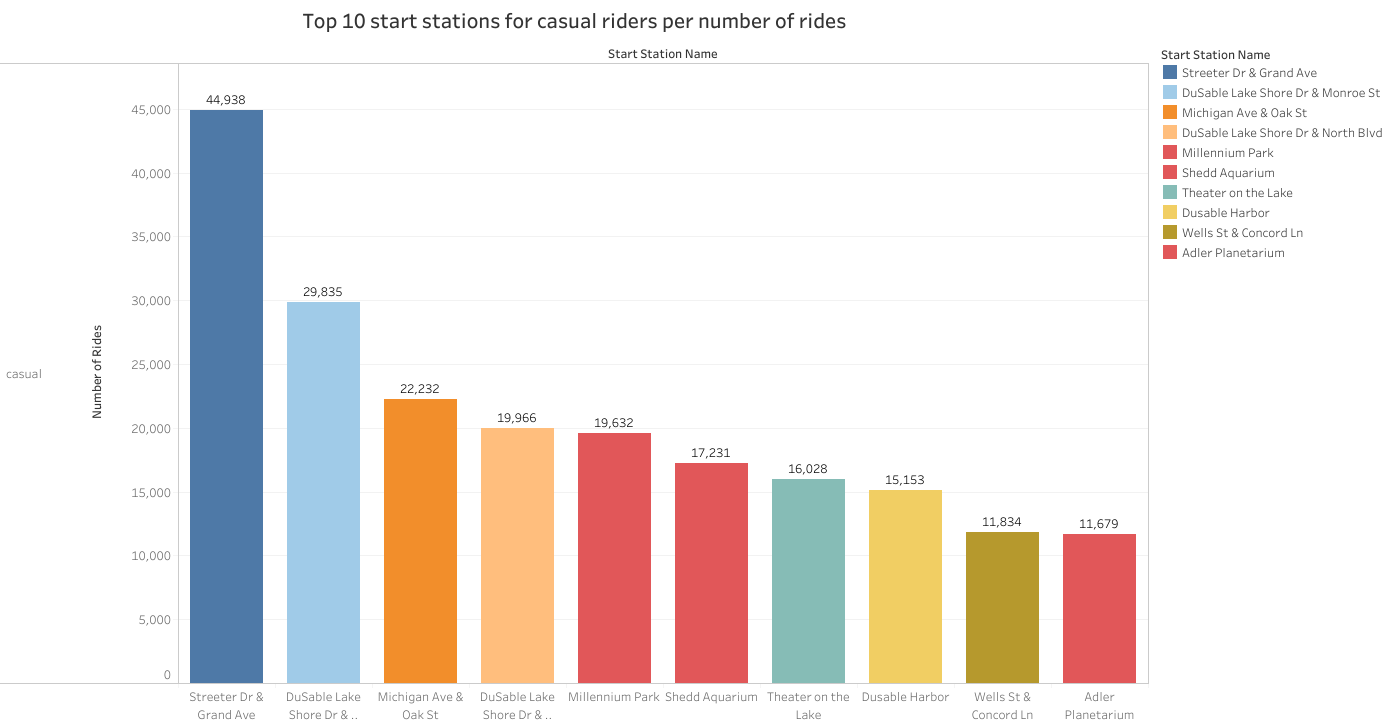


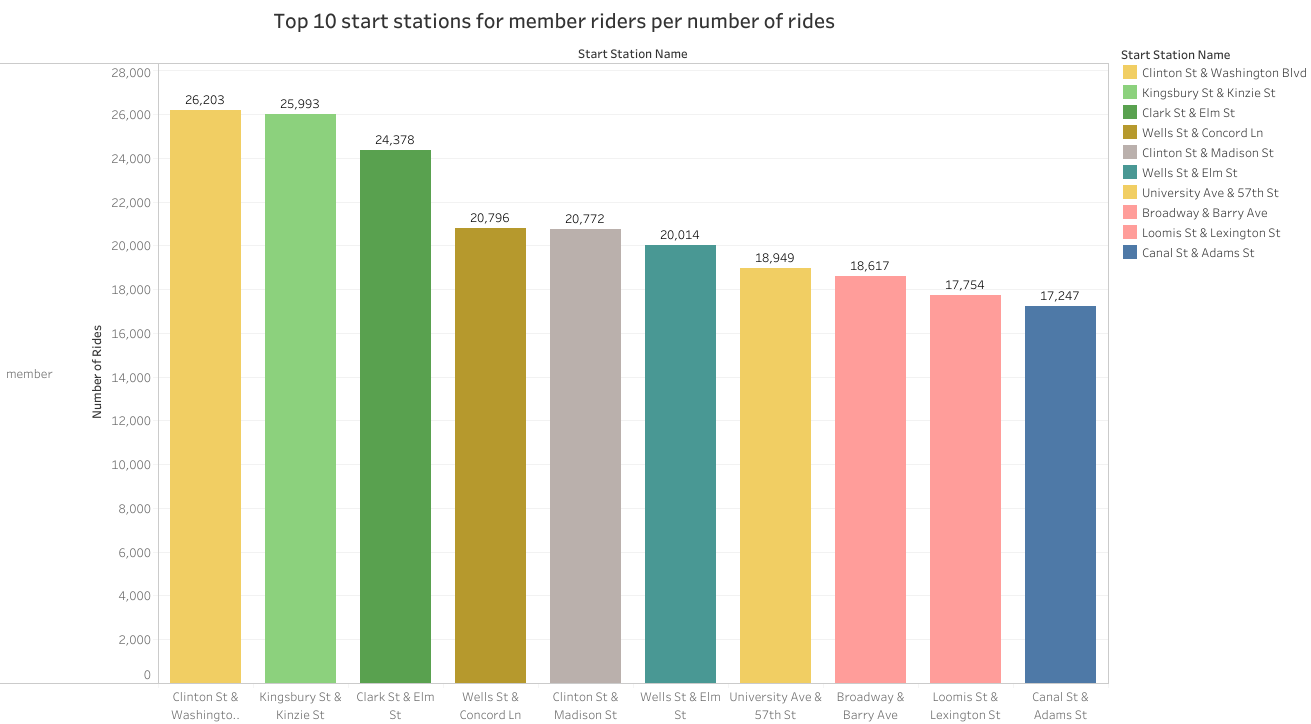
**Identified top 10 start stations for casual and member riders to study trends:**

The graphs show the top ten start stations for both casual and member riders. Understanding why

certain stations are popular among members can help improve those attributes among the popular

stations for casual riders too.





1. **ACT:** In this phase, I will be using the analysis done in the previous two phases to provide conclusions and recommendations to the stakeholders. Based on the findings, the company can make key decisions to convert casual riders to members.

**Conclusions:**

* Members use the bikes more than casual riders
* Members prefer using classic bikes to electric bikes whereas electric bikes are more popular among casual riders
* Only casual members use docked bikes
* Members have shorter ride durations than casual riders
* The months of July and August 2023 experienced the highest number of rides among casual and member riders.
* January 2024 being the winter month experiences the least number of rides
* Saturday is the most popular day among casual riders whereas Thursday is the most popular among member riders.
* The top three stations for casual riders are Streeter Dr & Grand Ave, DuSable Lake Shore Dr& Monroe St, Michigan Ave & Oak St
* The top three stations for member riders are Clinton St & Washington Blvd, Kingsbury St & Kinzie St, Clark St & Elm St.

**Recommendations:**

* **Promote Electric and Docked bikes with Membership incentives:** Start a focused marketing push for electric and docked bikes since casual riders favor them. Provide special discounts and perks for these bikes when people sign up for membership. This way, casual riders will feel more tempted to become members.
* **Summer Promotion Campaign:** Since July and August see the highest number of rides among both members and casual riders, membership should be promoted during these months with a focus on summer-specific promotion plans offering lower fees or bonus rides to encourage frequent rides.
* **Day-Specific Promotions:** Considering casual riders' increased bike usage over the weekend, it's strategic to run a targeted promotion campaign specifically on Fridays, Saturdays, and Sundays. This approach will incentivize casual riders to convert to members during their peak riding days.
* **Station-Specific Marketing Campaign:** A campaign to incentivize membership sign-ups at the top three stations, preferred by casual riders can be implemented. If the progress is significant after the campaign, it can be extended to 7-10 more stations.
* **Referral Programs:** A referral program can be implemented where existing members can earn rewards or credits for referring friends and family to join as members. This will help existing members to earn incentives and also encourage new sign-ups.

**Additional Observation:**

During the analysis, it was noticed that a significant number of rows did not have a start station and end station. This missing information is a key data point that is helpful in identifying popular stations. It needs to be examined why this information is not being recorded and essential steps need to be taken to fix the issue.