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

**Introduction:**This case study is part of the journey to earn a [Google Data Analytics Professional Certificate](https://www.coursera.org/professional-certificates/google-data-analytics).In this case study, the steps of the data analysis process: Ask, Prepare,Process, Analyze, Share and Act are performed to show how casual riders and annual members use Cyclistic bikes differently.

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

Tools: Spreadsheet (Google sheet), SQL (BigQuery), Tableau

Visualization: Go to my[Tableau Profile](https://public.tableau.com/app/profile/srinidhi.b8847)

**Characters and teams**

● **Cyclistic:** 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 the bikes to commute to work each day.

● **LilyMoreno**:The director of marketing and 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.

**1.Ask**

#### **Scenario:**

The marketing team at Cyclistic, a bike-share company located in Chicago, has the task of devising marketing strategies making casual bike riders to annual members However, in order to achieve this goal, the analyst team must gain a clearer understanding of the differences between annual members and casual riders and can drive the business decision by analyzing the historical data of the bike riders to drive insights.

#### **Stakeholders:**

* Director of marketing
* Cyclistic executive team
* Cyclistic marketing analytics team

**Deliverable:**

* Getting knowledge about the unique usage patterns of annual members and casual riders with regards to Cyclistic bikes ride.
* Presenting insightful data driven decision visualization to validate the differences.
* Using the insights to take actionable recommendations to convert casual riders into annual members.

**2.Prepare**

**Data Sources**

For the year 2023, a total of 12 datasets have been made available. Every ride that Cyclistic customers have logged is detailed in each dataset. The personal information of the riders has been removed from this publicly available data.

* The data is organized by month and year and is stored as a zip file.
* There are no issues with bias or credibility in this data.
* Data is ROCCC i-e; as follows
* **Reliable**:The data is reliable and credible.
* **Origina**l: The data has been validated by the source.
* **Comprehensive**:The data has all the elements required for the data analysis process.
* **Current**:The data is current as it has been constantly updated to the current month.
* **Cited**:The data has been cited by the source.

**Documentation, Cleaning and Preparation of data for analysis**

In order to conduct an adequate analysis,Data integrity can be achieved by checking data redundancy , accuracy, verification,data validation using spreadsheets and would need to employ either SQL or R, as the total size of the dataset is too large to be handled by a simple spreadsheet. I chose SQL(BigQuery) because I could handle data manipulation, analysis on the same platform.

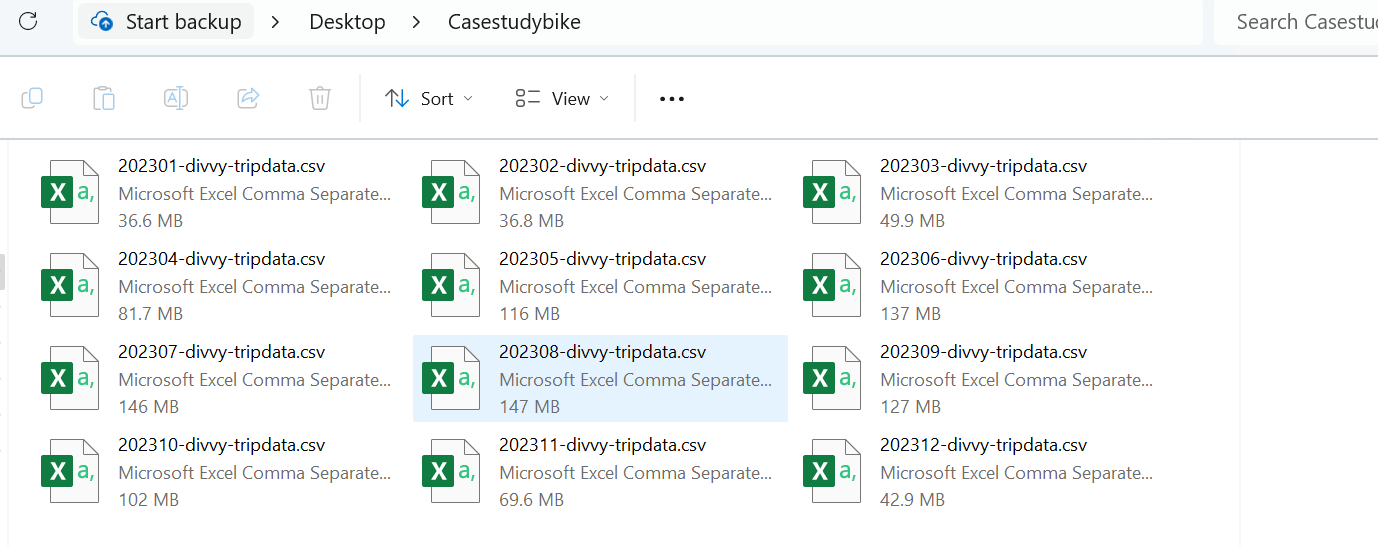
Addressing licensing, privacy, security, and accessibility involves implementing comprehensive strategies in each of the areas to ensure compliance with regulations, protect user data, and provide inclusive access.

**Key Tasks**

* Download data and store it appropriately.
* Identify how it’s organized.
* Sort and filter the data.
* Determine the credibility of the data.

**Deliverable**

* A description of data sources used is 2023 (January 2023-December2023) bike share data i-e: 12 months dataset of bike-share.
* 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: January 2023 - December 2023.[Dataset](https://divvy-tripdata.s3.amazonaws.com/index.html)



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

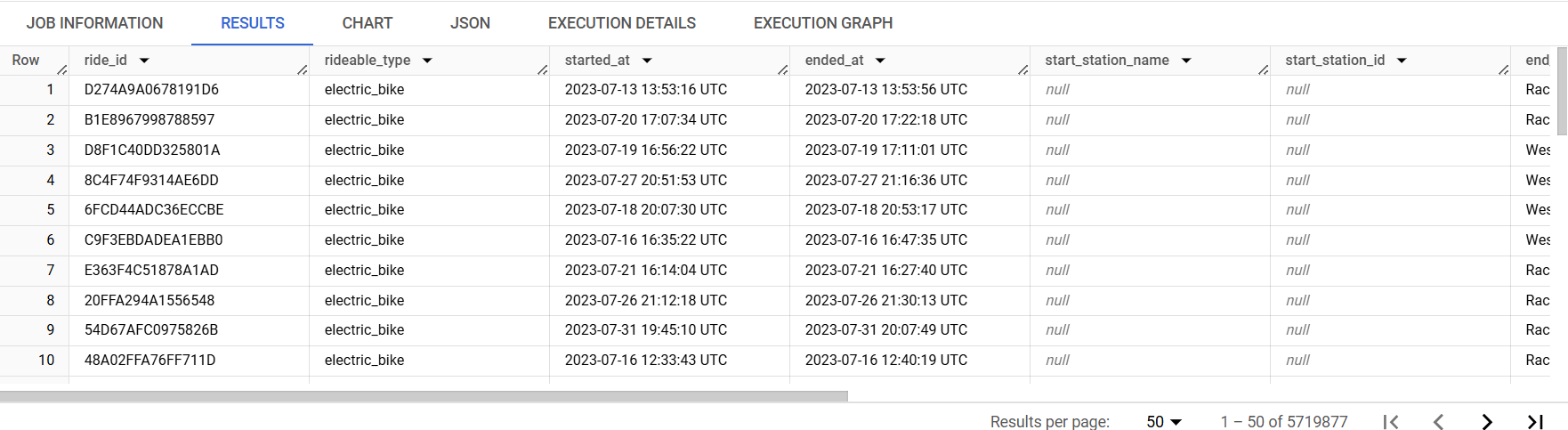
I tried performing some basic calculations to understand the range of the data in Google Sheet as the dataset is too large and found difficulty so chose the SQL in BigQuery for further calculations and analysis.By Google Sheet got to know few knowledge about the dataset as follows:

* There are 13 columns in the files.
* There are three types of bikes: Classic Bike, Docked\_Bike and Electric Bike
* There are started\_at and ended\_at time stamps which can be used to find the duration of each trip.
* 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.
* There are no duplicate values in any of the files.
* 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.



Below is the query result table:



The combined data set has **13** columns and **5719877** rows. I saved the data set as a new BigQuery table “Combined\_12months\_dataset “

**3.Process**

The 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.

**Guiding questions**

● **What tools are you choosing and why?**

Tool choosing is SQL in BigQuery as the dataset is too large; it's easy for the calculation and analysis purpose.

● **Have you ensured your data’s integrity?**

Yes, data is reliable by checking data validation,redundancy,accuracy.

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

* Review data documentation and metadata.
* Removing duplicate records and the blank/empty cell rows based on the relevance.
* Convert data types to appropriate format.
* Validating data integrity.
* Standardize date formats, text casing, and numerical precision.
* Ensure documentation is clear and accessible for future reference.

**Key tasks**

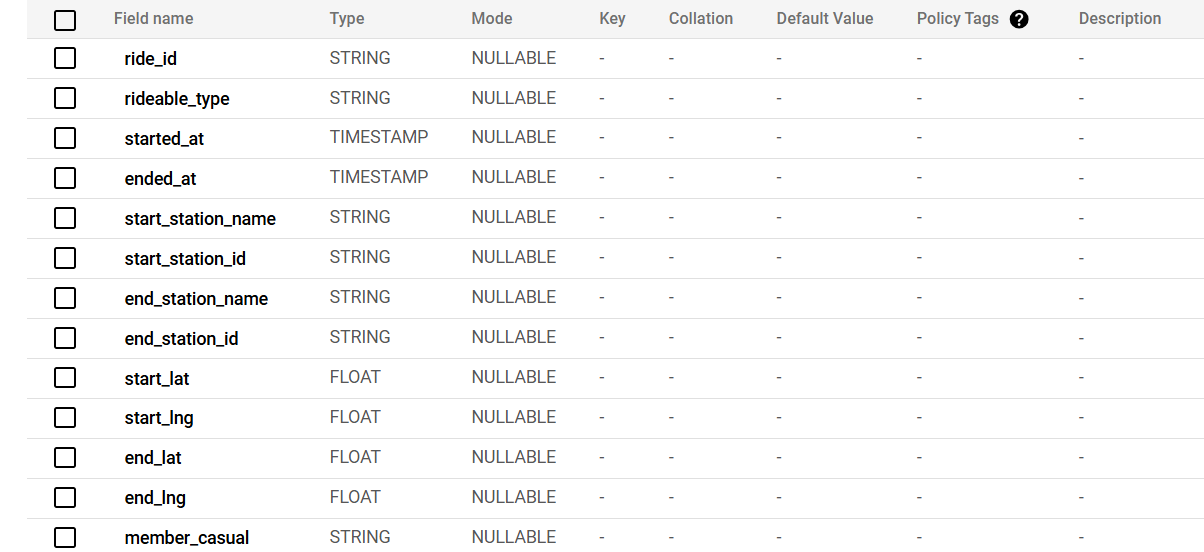
● Check the data for errors.

● Choose your tools.

● Transform the data so you can work with it effectively.

● Document the cleaning process.

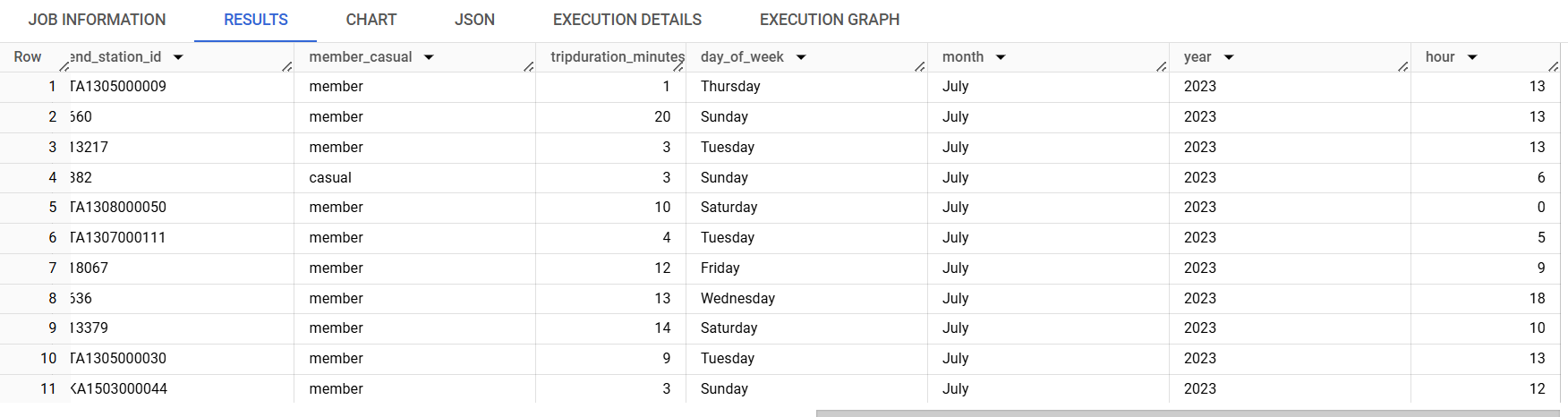
Steps followed in the **PROCESS:**

**Step 1:**Combine the table of 12 months as **“Combined\_12months\_dataset”** to make changes and updates for analysis. Below is the schema for the table.

**Step 2:**

Created new columns as follows tripduration\_mintues, day\_of\_ week, month, year and hour by running the query below.



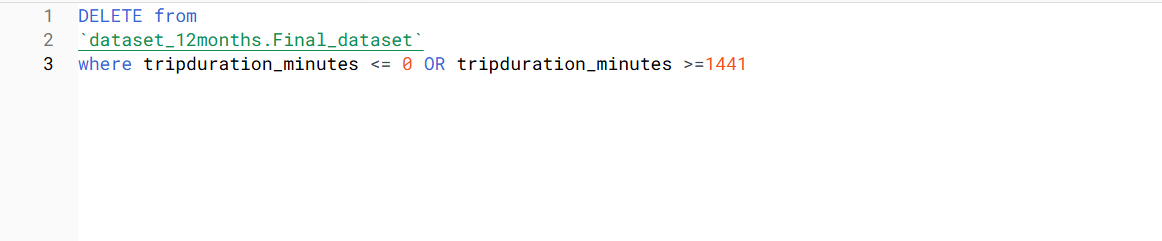


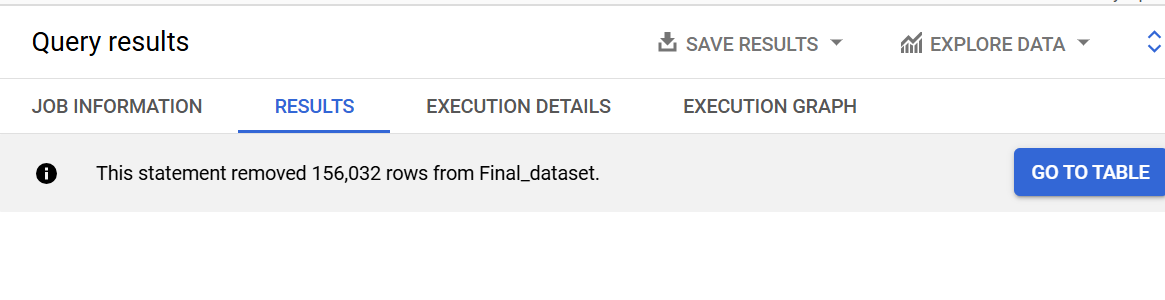
**Step:3**

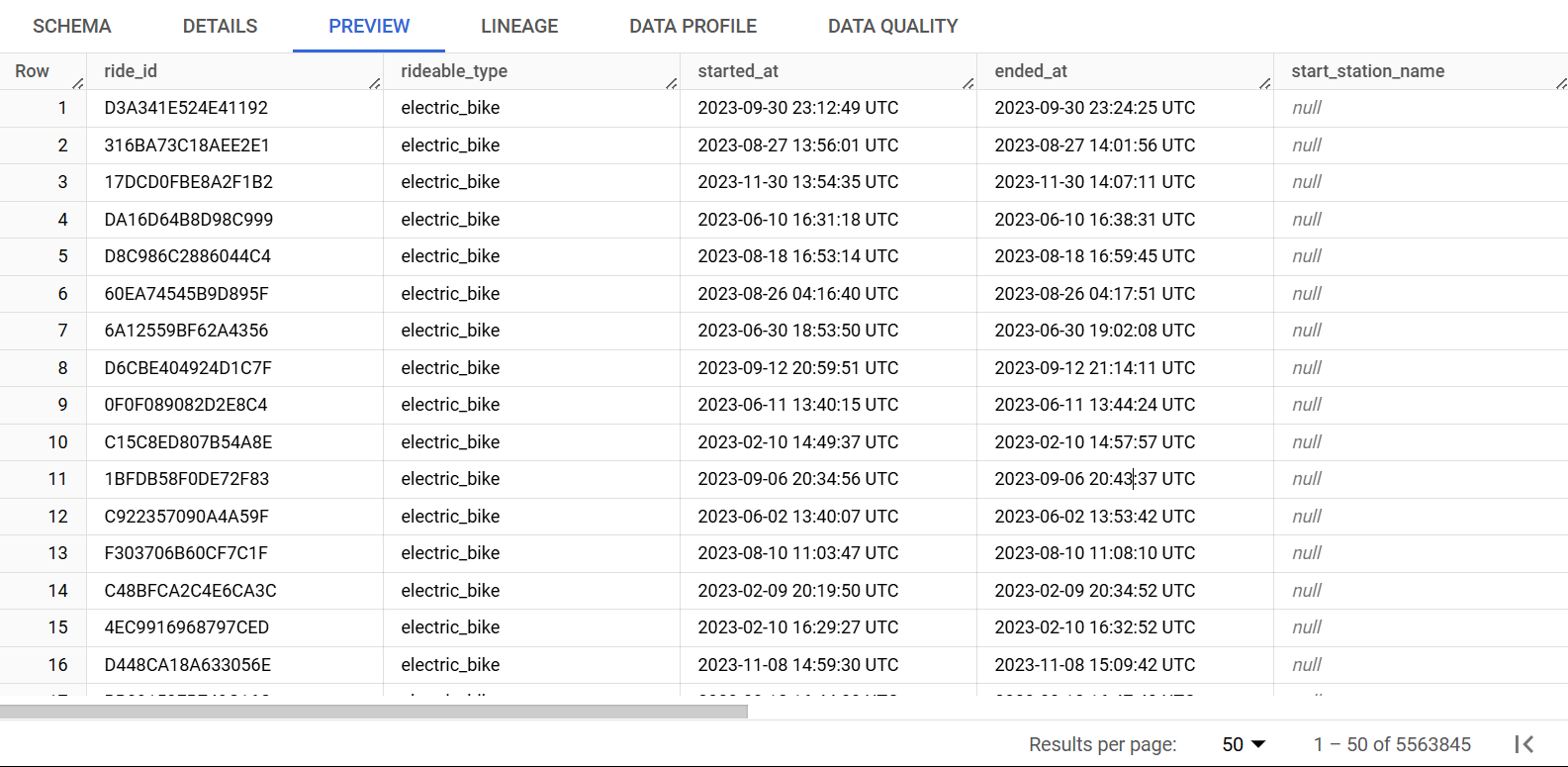
Save the new table with the additional columns as “Final\_dataset”.

**Step:4**

Noticed that some cells in column **tripduration\_minutes = 0** and some longer than 24 hours (1440 minutes) which might not provide a correct average trip duration for all rides.so,ran a query to delete rows with tripduration\_minutes greater than a day or equal to zero. The query deleted 156,032 from the table “Final\_dataset”. Now, the total number of rows in the table are 5,563,845







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 did not delete the rows as deleting these rows would significantly reduce the data set.

**4.Analyze**

**Explanation:** In this phase, analyze existing relationships across columns to identify key trend insights by extracting small data sets from the final table. These tables helped to understand and to make data driven decisions about the difference between the riding of casual riders vs member riders.

**Key Tasks**

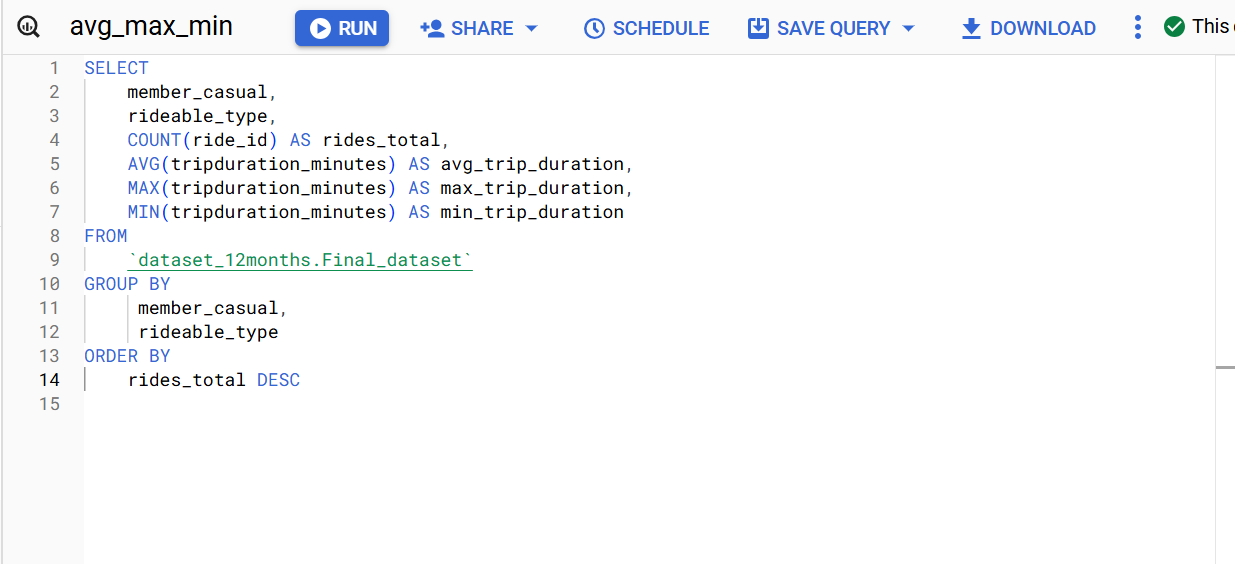
● Aggregate the data so it’s useful and accessible.

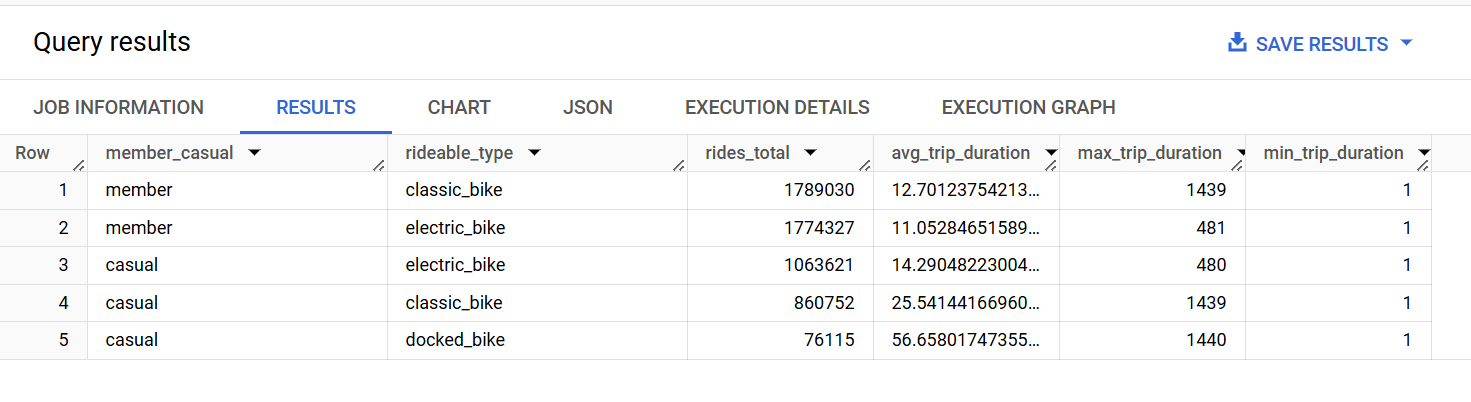
● Organize and format the data.

● Perform calculations.

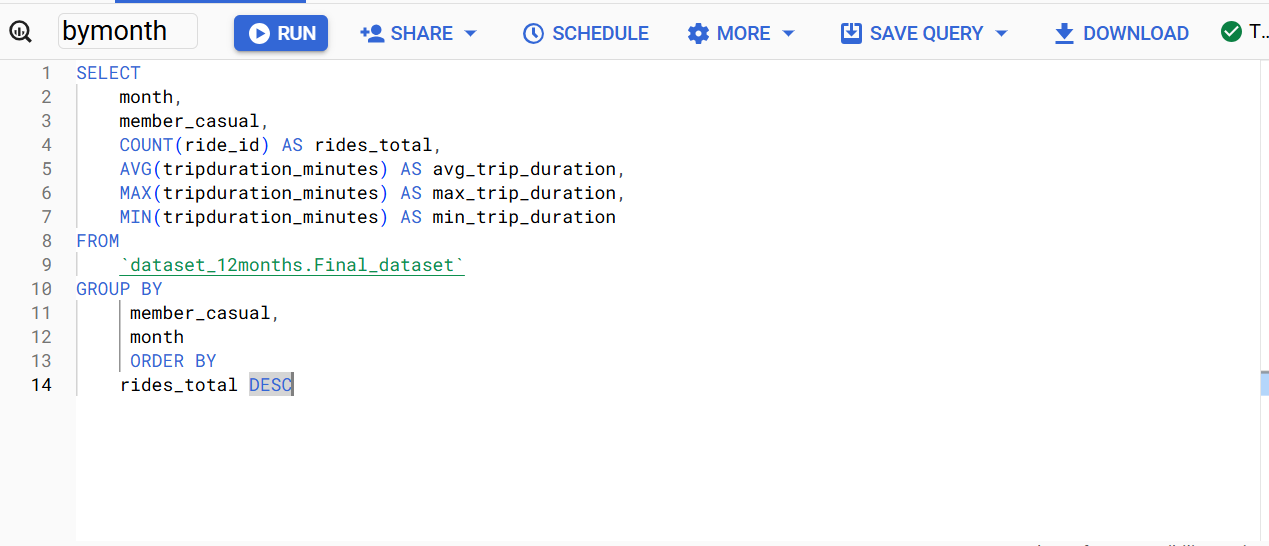
● Identify trends and relationships

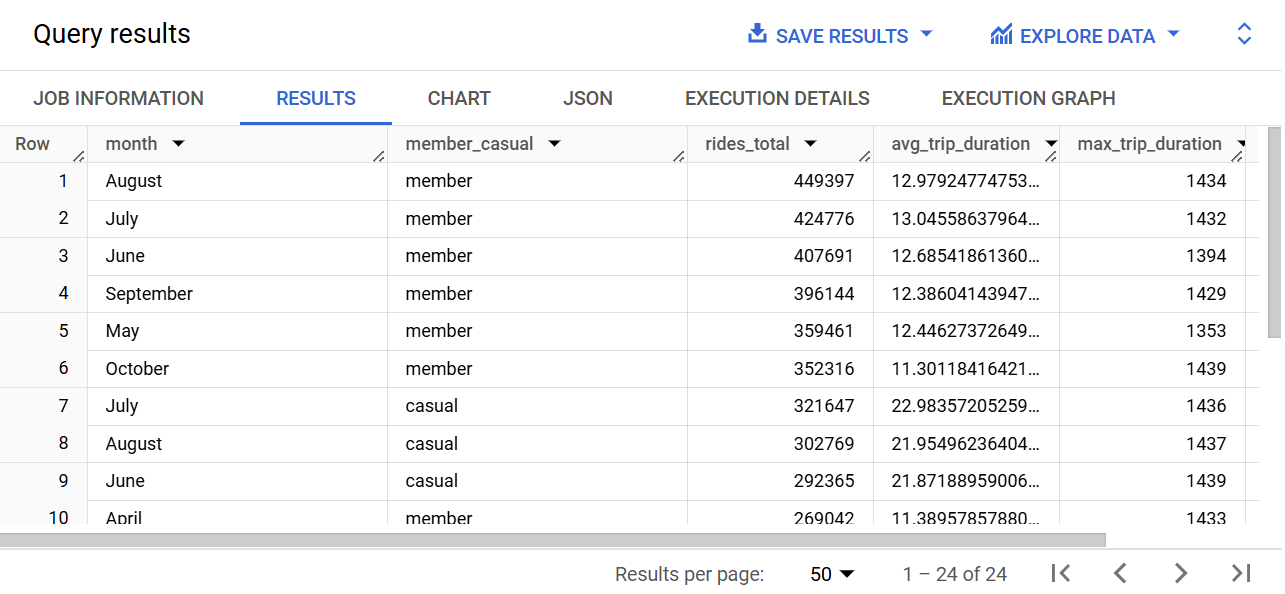
**Step:1** Executed a query to find the ride total, average, maximum and minimum trip duration for different ride types by casual and member riders.



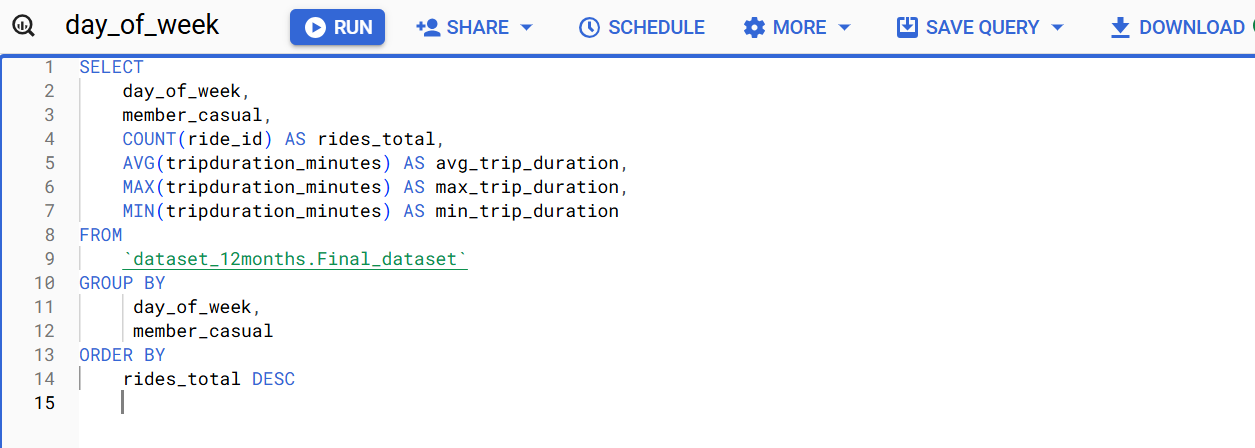


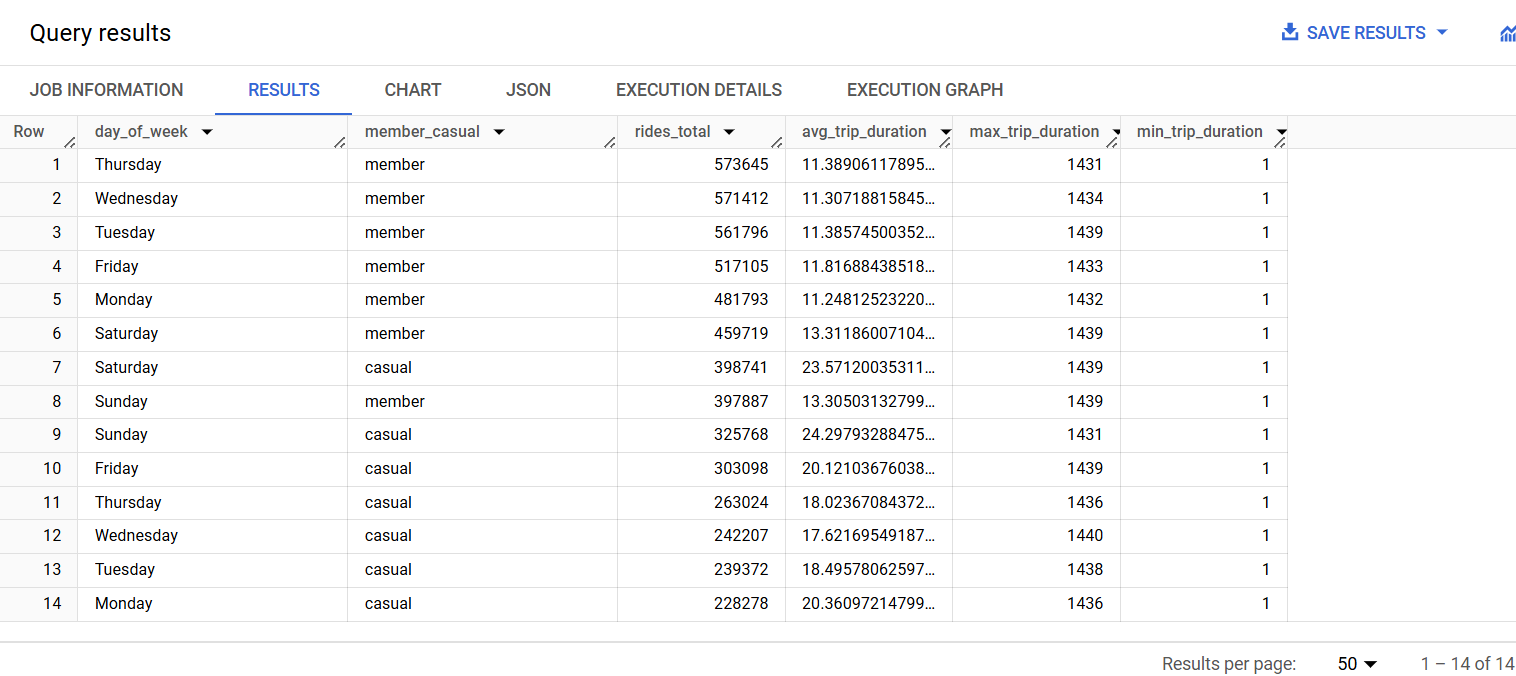
Step:2 Executed a query to find the ride total, average, maximum and minimum trip duration for each month by casual and member riders.



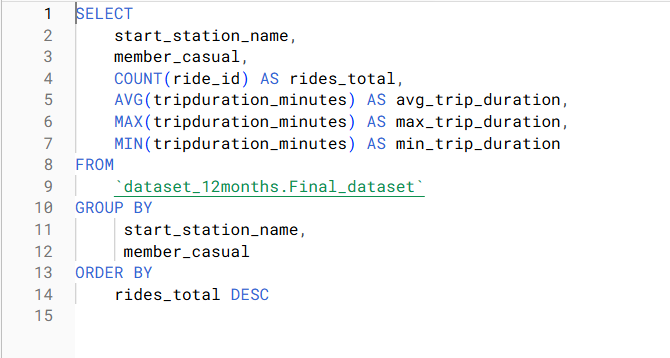


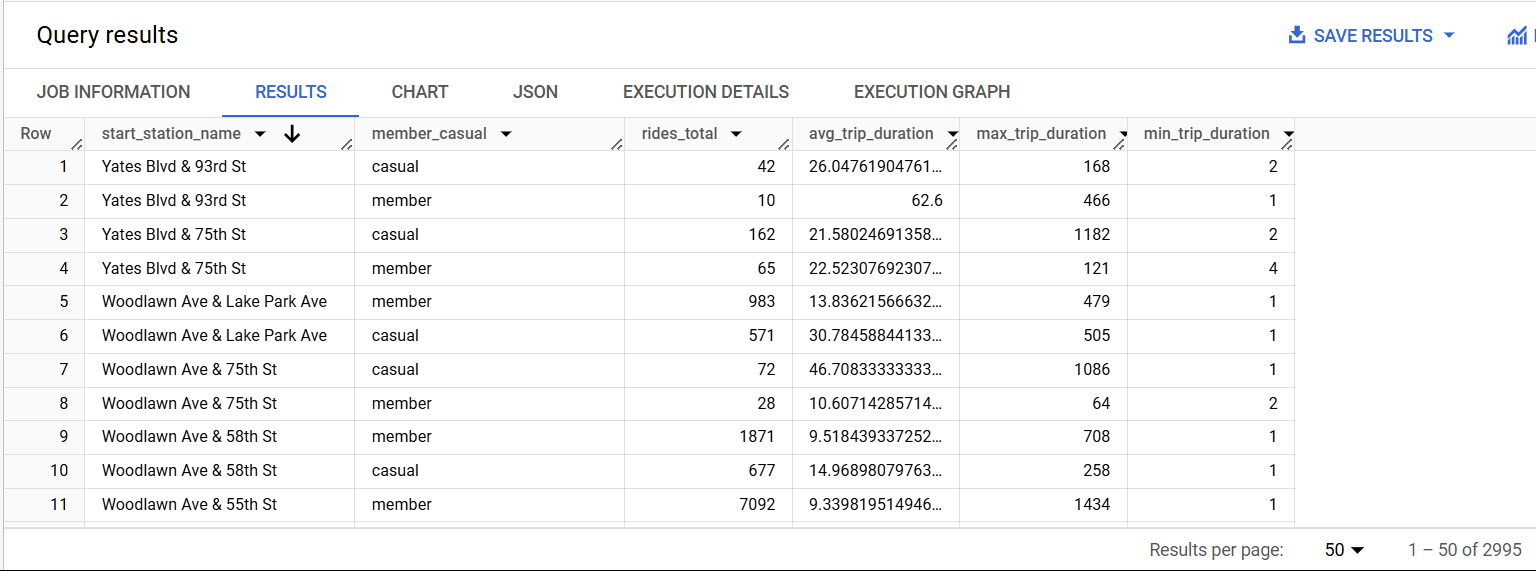
**Step:3** Executed a query to find the ride total, average, maximum and minimum trip duration for each day of the week by casual and member riders.



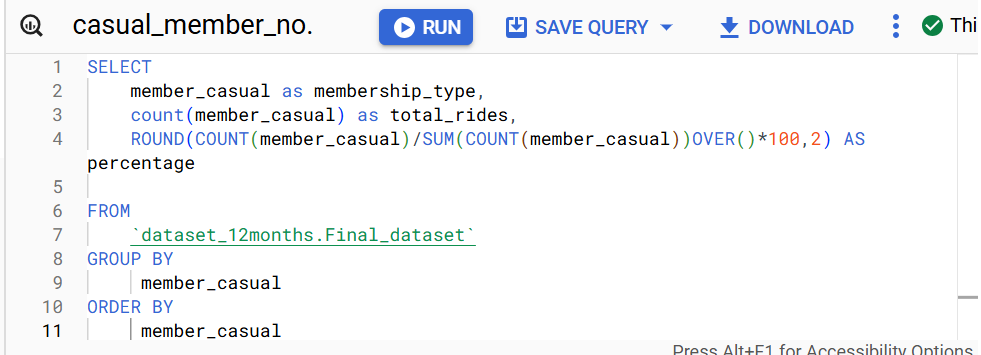


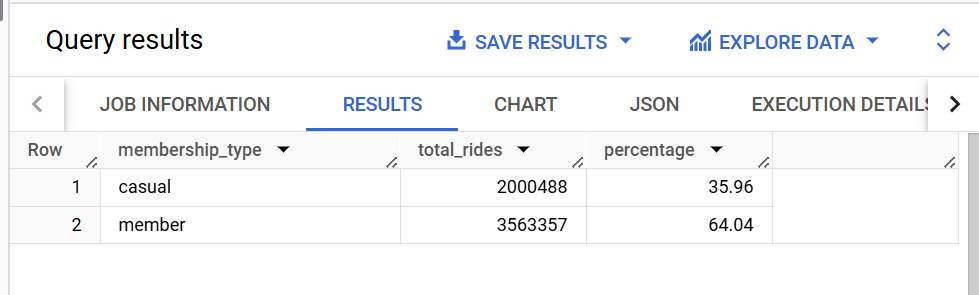
**Step:4** Executed a query to find the ride total, average, maximum and minimum trip duration at each start station by casual and member riders.



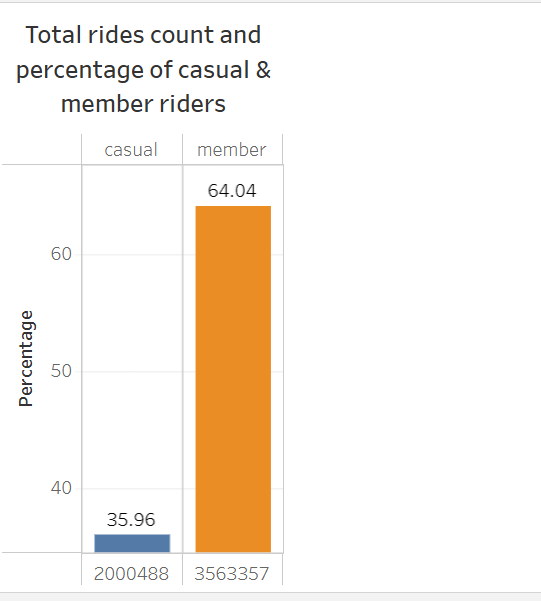


**Step 5:** Execute a query to find the total number of rides and its percentage taken by each rider type, sorted in alphabetical order and the percentage rounded to two decimal places.

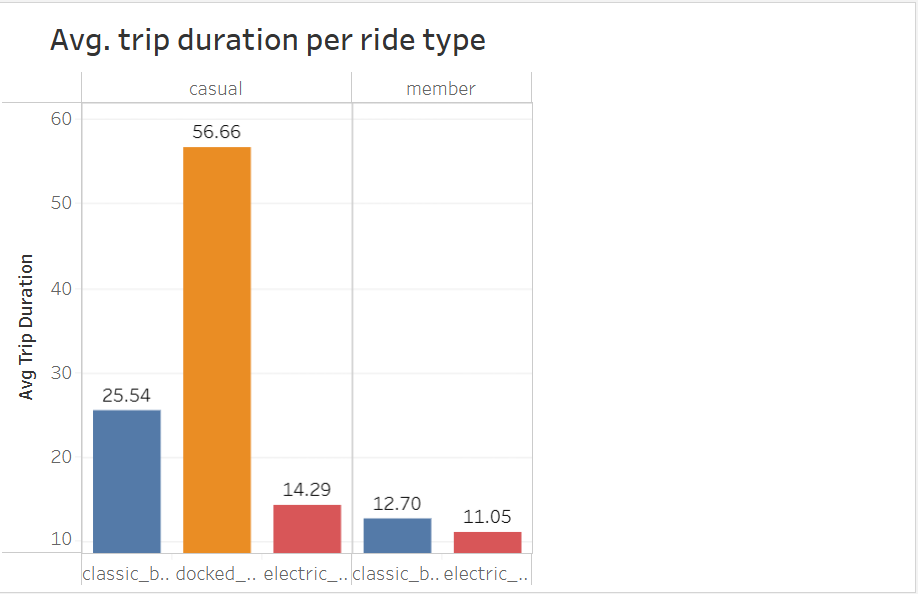


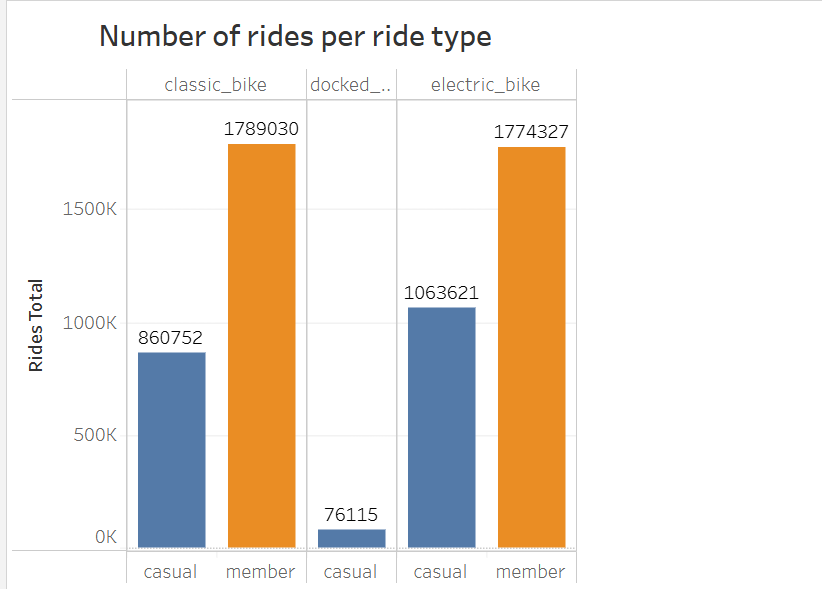


**5.Share:**In this phase of data analysis, visualize key relationships and trends using Tableau. These visualizations will help the tables observed using SQL queries above, easy to understand and interpret to data driven analysis.

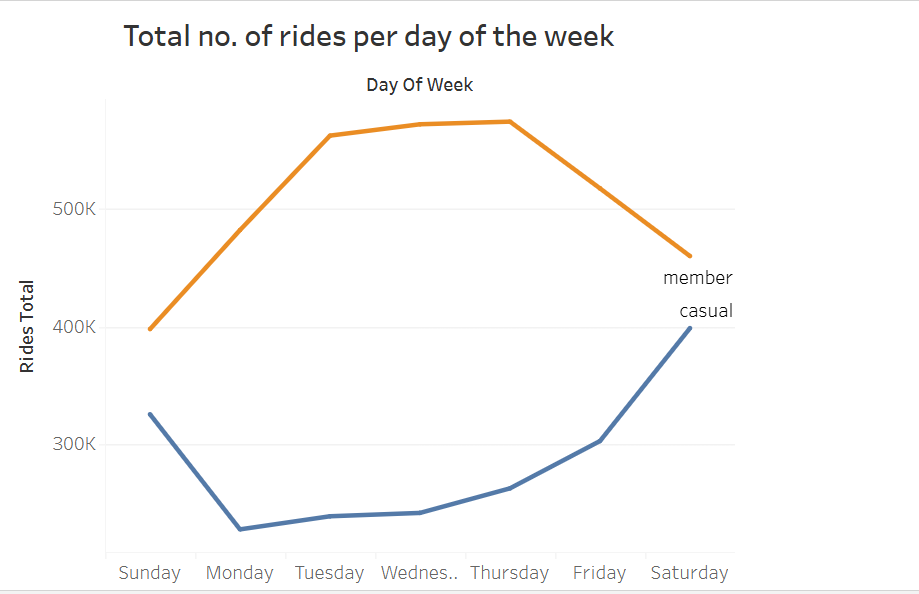
**Compared number of rides between Casual vs Member riders**: The visualization below highlights that the frequency of bike usage among member riders is more compared to casual riders.

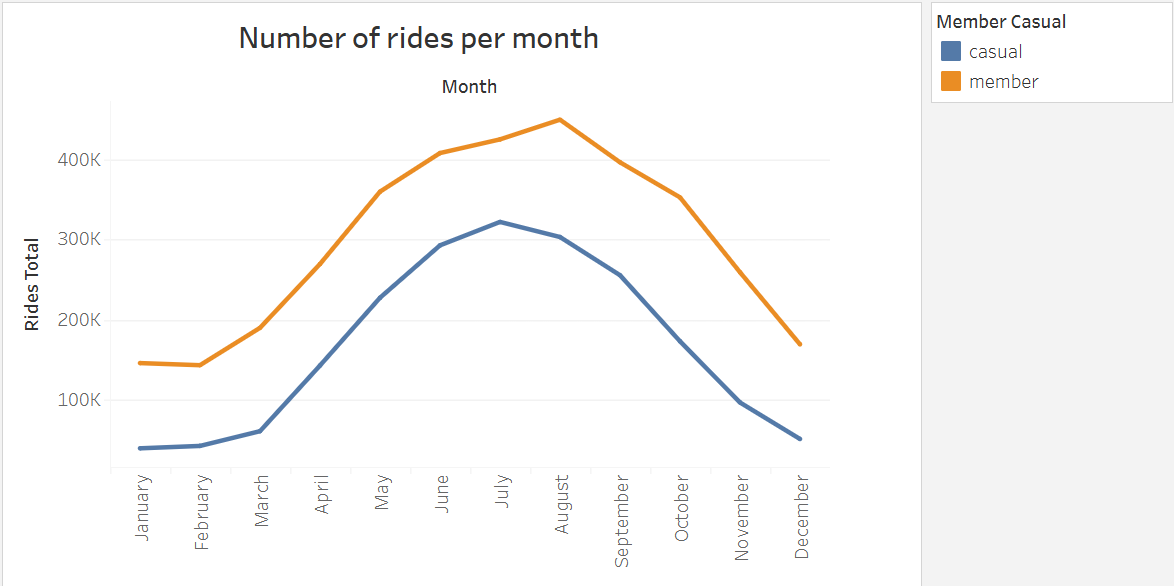
**Compared average trip duration per ride type between casual and member riders:** The visualization indicates that casual riders tend to have longer average trip durations when using classic bikes, whereas member riders typically have shorter average trip durations when using electric bikes.



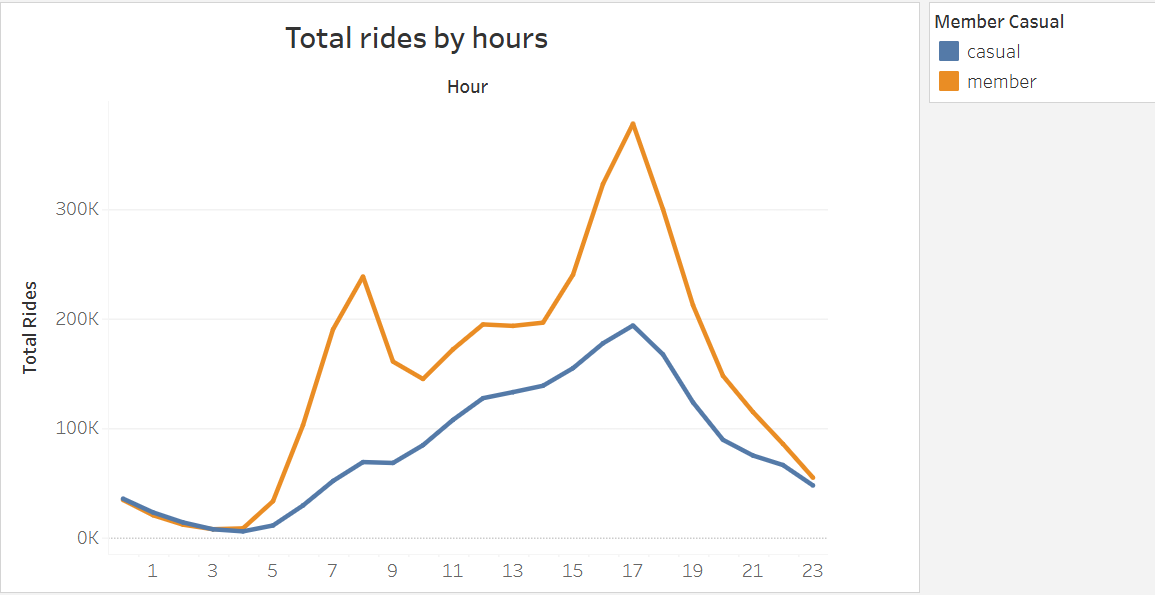
**Compared number of rides per ride type between casual and member riders:** The visualization below shows that classic bikes are preferred by member riders, while casual riders prefer using electric bikes. 

**Compared 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.



**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 2023 had the least number of bike rides by both casual and member riders.

**Compared number of rides per hour between casual and member riders:** Member riders rent bikes most frequently during rush hours (around 8 AM and 5 PM), while casual riders tend to rent bikes primarily around 5 PM.



**Explanation:**

* Able to answer the question of how annual members and casual riders use Cyclistic bikes differently by representing in visual form.

* The story provides a holistic view of how people move around the city, the factors influencing their choices, informs future improvements and strategic decisions to enhance the system’s efficiency and user satisfaction.
* By providing actionable insights that can be used to improve operations, enhance user satisfaction, and drive strategic planning. By continuously analyzing and relating data back to the core questions, businesses can make informed decisions that align with the goals and respond effectively to user needs and market dynamics.

Audience are team leaders ,Investors,stakeholders, analysts,team members,marketing and sales team.The best way to communicate is by visualization and by utilizing effective tools.

* Data visualization analysis helps to share findings and presentations accessible to the audience.

**Key Tasks**

● Determine the best way to share the findings.

● Create effective data visualizations.

● Present your findings

● Ensure your work is accessible.

**Deliverable**

* Supporting visualizations and key finding
* Creating the visualization by tableau.
* Ensuring the clear meaning through the proper use of common elements such as headlines, subtitles, and labels.
* Refining the data visualization by applying deep attention to detail.

#### 

**7.Act**

## **Conclusion:**

* Casual riders made 36% of total trips and member riders makeup 64% during 2023
* Annual members take more trips while casual riders take longer trips.
* Casual customers use bikeshare services more during weekends, while members use them consistently over the entire week.
* Both types, but especially casual riders bike significantly less during the winter months
* 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 2023 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.

## **Recommendations**

* Offer seasonal passes for riders who don’t want to bike during the winter.
* Introduce a member only rewards program based on trip duration to incentivize casual riders to sign up as members and be eligible for the rewards.
* Offer discounted pricing during non-busy hours so that casual riders might choose to use bikes more often and level out demand over the day.
* Understanding why certain stations are popular among members can help improve those attributes among the popular stations for casual riders too.

### A**dditional data that could expand scope of analysis**

* **Occupation of member riders** - this data could be used to target non-members who come under similar occupation
* **Age and gender profile** - Again, this data could be used to study the category of riders who can be targeted for attracting new members.
* **Students** - This data could be used to target students
* **Pricing details for members and casual riders** - Based on this data, we might optimize cost structure for casual riders or provide discounts without affecting the profit margin.
* Address/ neighborhood details of members to investigate if there are any location specific parameters that encourage membership.