# REPORT

## Problem Statement

The fictional company, Cyclistic has 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. Finance analysts have concluded that annual members are much more profitable than casual riders. This aim of this report is to answer what is the distribution and behavior of annual members and casual riders in order to form marketing strategy. In general, four things about annual member and casual user should be discovered: active day or month, length of trip, preferred type of bike, distribution of used station on map.

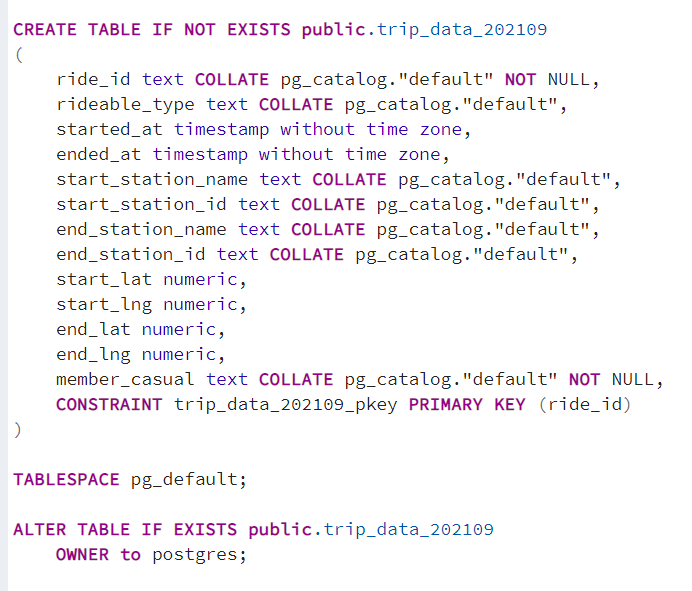
## Data sources

The data used is the record of Cyclistic trip data from September of 2021 till August of 2022 (12 months). It includes attributes: trip id, type of bicycle, time trip started and ended, id of the station, location of bicycle station, latitude and longitude of the station as well as type of customer.

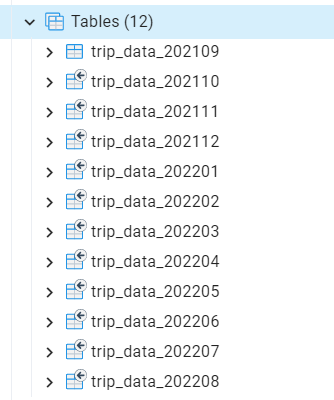
## Documentation of any cleaning or manipulation of data

3.1 Create table and import data

The RDBMS used is PostgreSQL. After database Cyclistic Project created. Table for September is 2021 is created first and imported. Below are settings of columns. Other than primary key ride\_id, member\_casual column is set as not null because of two reasons. First, the aim of the analysis is to differentiate member and casual user. Second, it is not deducible base on data exist.

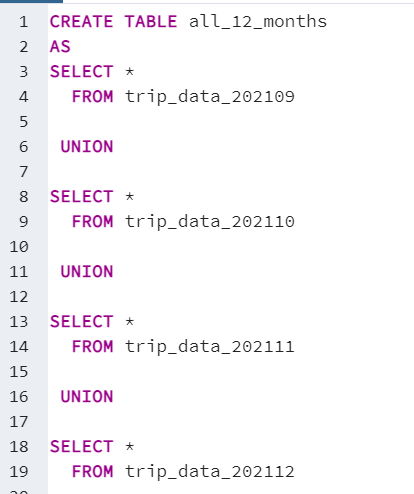


All 12 months data inherit the same settings. All the data is imported as well after table created.

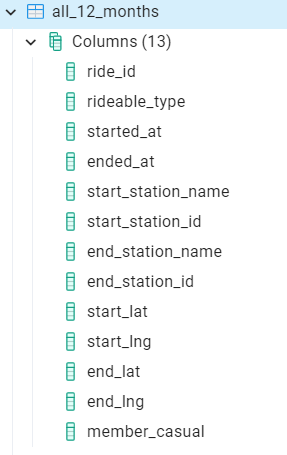


3.2 Combine tables to form new table

Union can be used to stack all of the records. Below is part of the query. Union will remove duplicated results so union is used otherwise select distinct is needed.

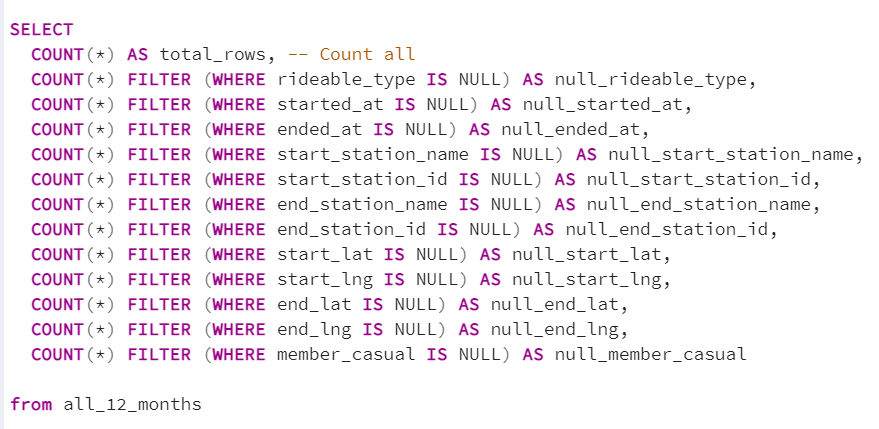


12 months trip table created.

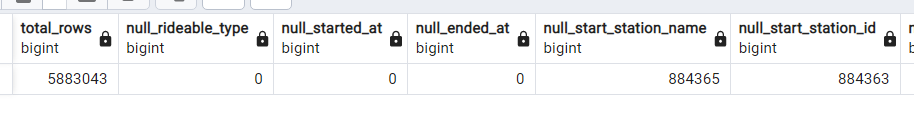


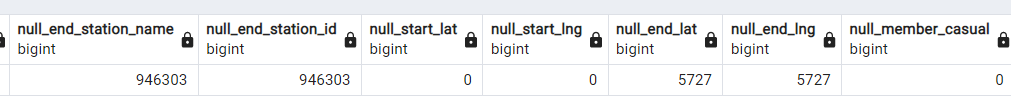
3.3 Check for null value

Count for null in each column



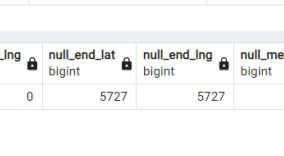
Returned results.



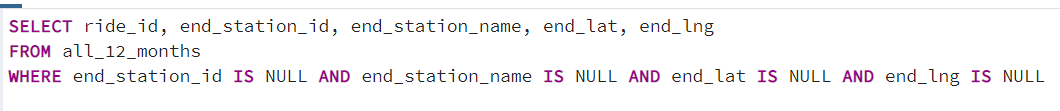


3.3.1 Null data handling

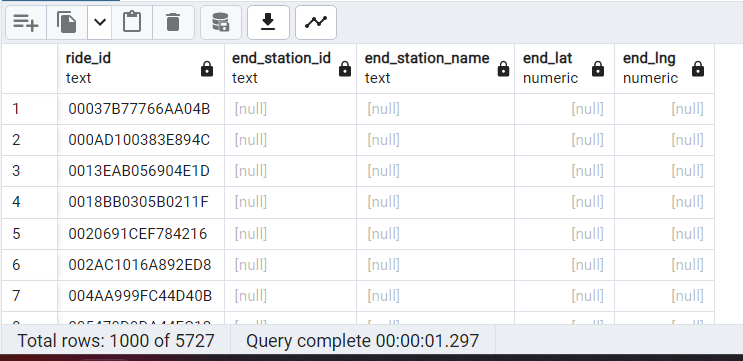
Null value seems exist in data related with station’s information. Name of station and id of station should be unique to each other. The coordinate and name of station are also related. To answer the question about region, coordinate or name of the place must be fulfilled for Tableau to interpret geographic data. Based on the previous result, there are 5727 rows do not have coordinate.



If the rows has station name and station id, the latitude and longitude can be referred by other record with same station name or station id. In that case, window function along with aggregate function can be used to form new table with less null values. Check for the rows with null value in coordinates.



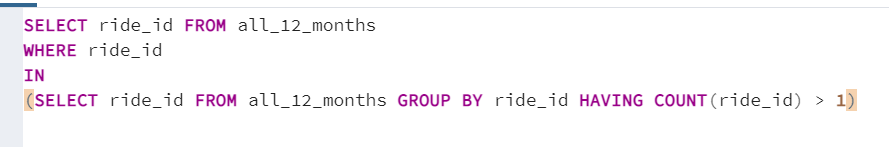
Apparently, 5727 rows with null value in coordinate also have null value in station id and station name , there is no way to deduce the coordinate base on the data.



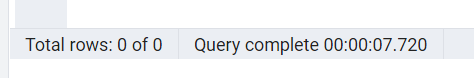
The population size is 5884043. 5884043 - 5257 = 5877316. The sample size is 5877316. Accuracy is important for coordinates, it should has very little margin of error. By fit the sample size, population, 99.999% confidence level in calculation, margin of error is 0.00%, therefore it will be fine to eliminate these 5257 rows when showing distribution on map.

3.4 Check for duplicate value

Duplicate rows already eliminated when union the rows. The column ride\_id is primary key, it should not duplicated.

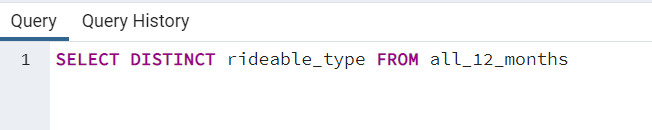


Zero result returned. Which mean there is no duplicate ride id.

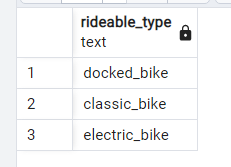


3.5 Check for Structural Error

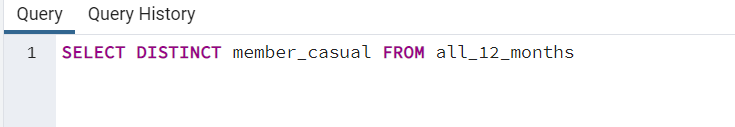
Check for type of bicycle.



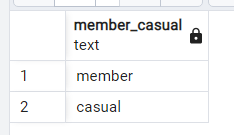
Bike type has no structural error.



Check for type of user.



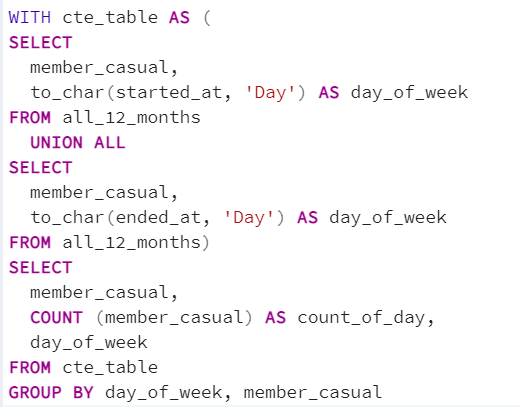
User type has no structural error.



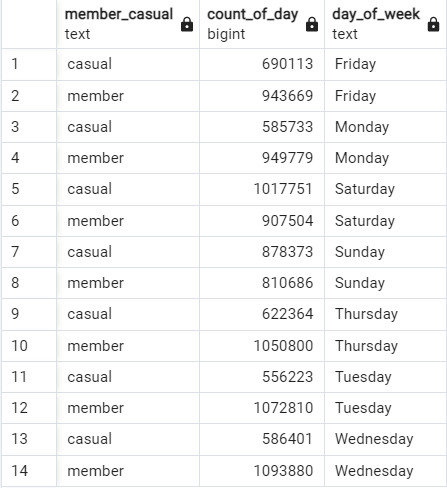
## A summary of your analysis

* 1. Analysis for active day or month

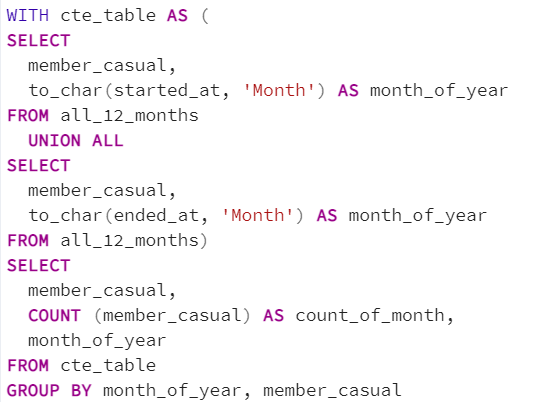
Days of users utilize extracted from started\_at and ended\_at column. Days to use the service counted.



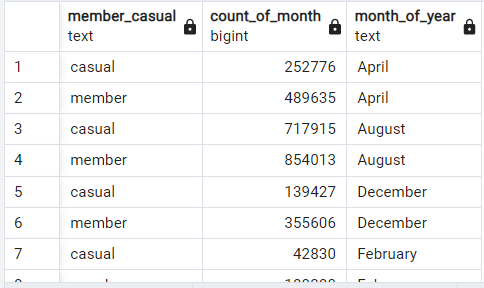
Part of the results.



Months of users utilize extracted from started\_at and ended\_at column. Months to use the service counted.

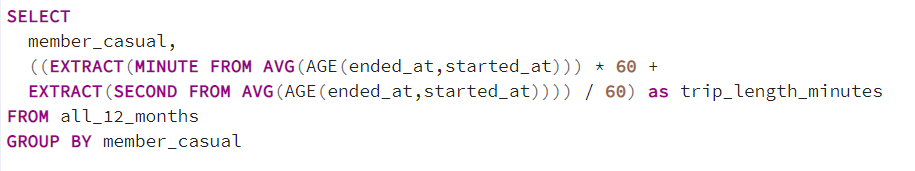


Part of the results.

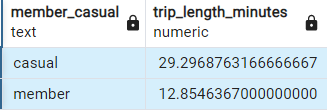


4.2 Analysis for length of trip

Average length of trip for member and casual user calculated.

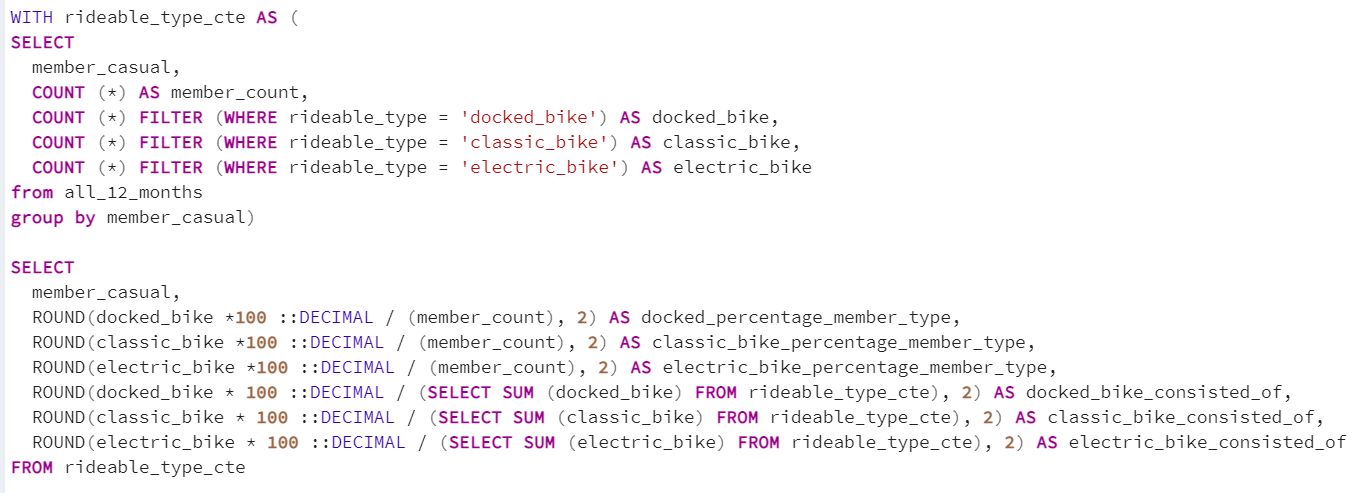


Results.

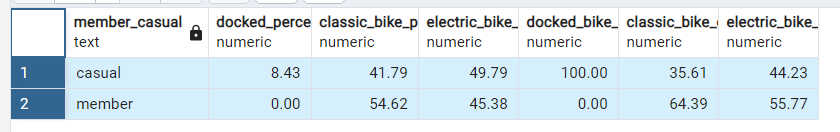


4.2 Analysis for bike type

Two aspects here, percentage of type of bike for member and casual user, percentage of member type for each type of bike calculated.

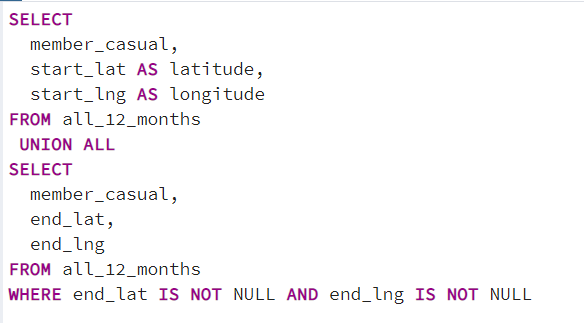


Results.

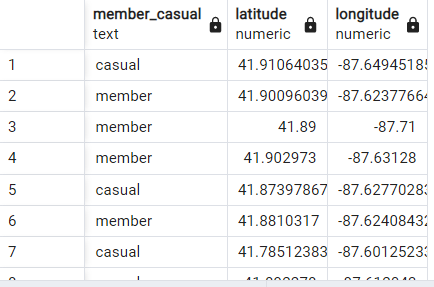


4.4 Analysis for station location

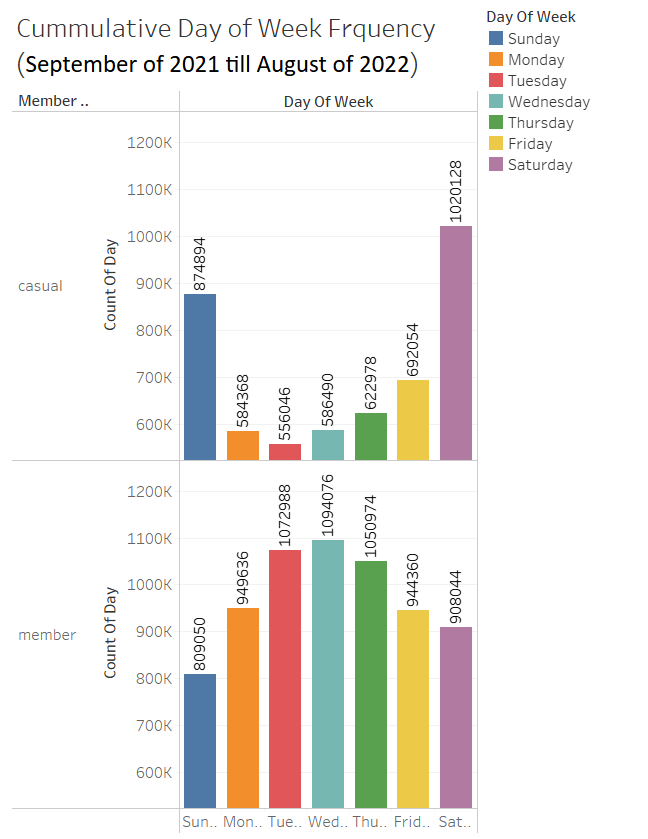
Forming table with geographical data with respect of member type. The aim is to find out any region has significant different amount of member type, start and end stations are stacked together.



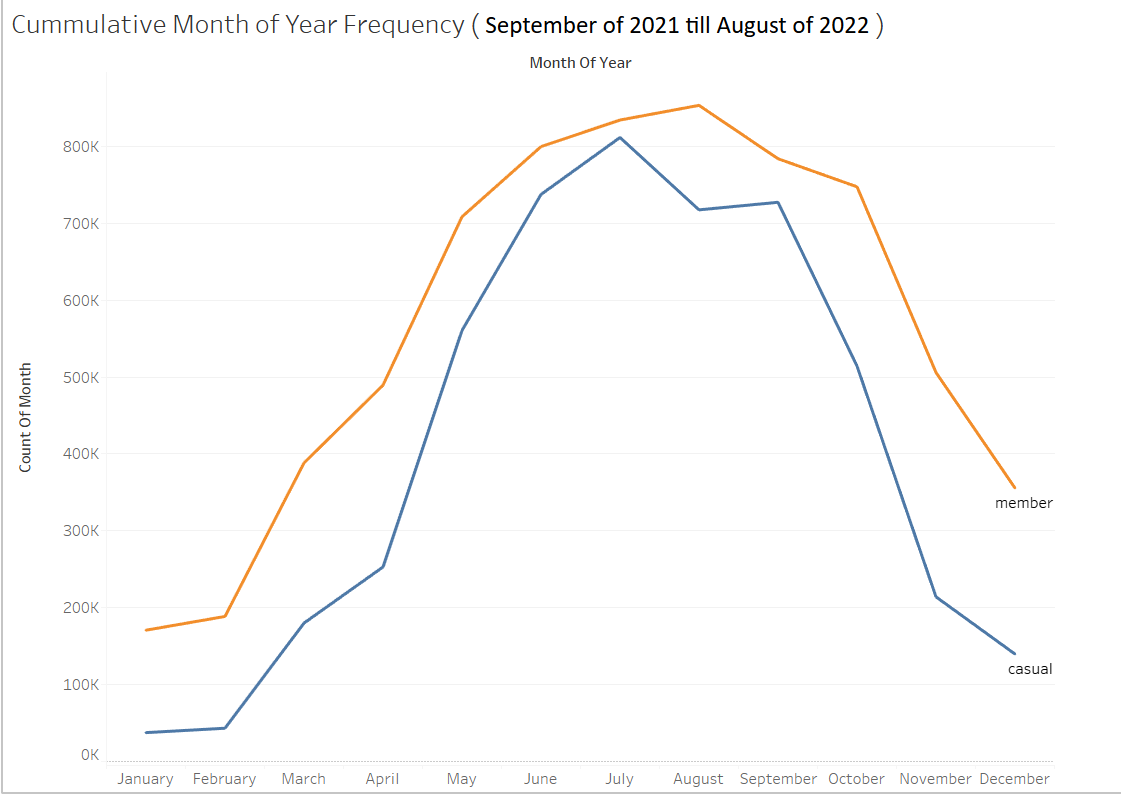
Part of the results.



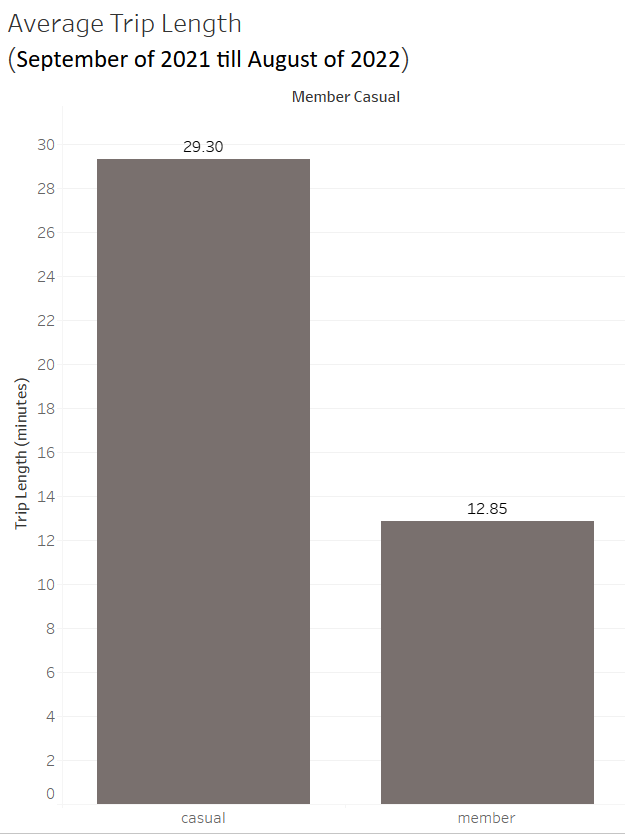
## Supporting visualizations and key findings



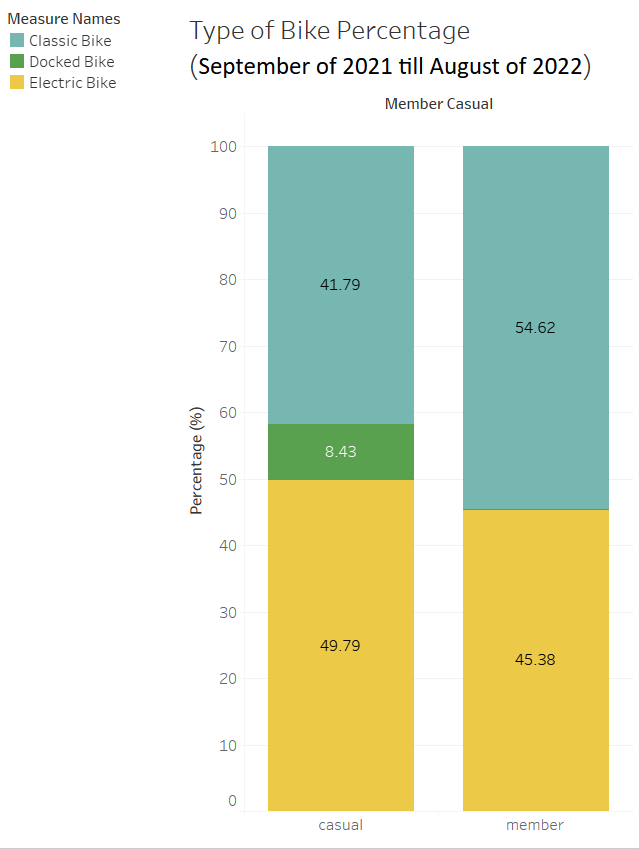
Casual user use the service frequently at weekends while member user usually use the service during weekdays.



Both casual user and member user have similar trend in month of year frequency but casual user has its mode in July while member user has its mode in August.



Average trip length of casual user is 29.30 minutes while member user is 12.85 minutes. Casual user travel longer than member user.



All docked bike is used by casual user. Member user prefer classic bike over electric bike compare with casual user.

### Top three recommendations based on your analysis

Recommendations

- Introduce weekends only membership since casual user use the service on weekends frequently.

- Charge the trip of casual user by length of trip as casual user travel longer.

- Install more electric bike so that casual user get to use the type of bike they use more.