Cyclistic Case Study

Introduction

For the capstone project, I have selected the Cyclistic bike share analysis case study to work on. For the case study, I will perform the real-world tasks of a junior data analyst for the marketing team at Cyclistic, a bike-share company in Chicago.

To answer key business questions, I followed the six steps of the data analysis process taught in the course which are: **Ask, Prepare, Process, Analyze, Share and Act**.

- Detailed documentation of code is available in GitHub.
- Initial analysis of datasets provided by Cyclistic using Microsoft Excel.
- Data cleaning, validation and exploration using Microsoft SQL.
- Data Visualization using Tableau Public.

Background

Cyclistic:
A bike share program that features around **5828 bicycles** and **692 docking stations** in 2016. Cyclistic differentiate itself from competition by also offering reclining bikes, hand tricycles, and cargo bikes. Most riders opt for traditional bikes and about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but 30% use them to commute to work each day.

1. Ask

Identify the business task:
Strategy to maximize the number of annual memberships by converting casual riders into annual riders.

Consider key stakeholders:
Lily Monero & the Executive team

Stakeholder perspective:
Monero believes company's future success depends on maximizing the number of annual memberships. She believes rather than creating a marketing campaign targeting all new customers, there is a very good chance to convert casual riders into members

Questions to Analyze:

- 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?

Monero has assigned the first question for the junior data analyst to analyze.

2. Prepare

Data Source:

Past 12 month of original bike share dataset
from **01/10/2020** to **30/02/2021** were extracted as 12 zipped .csv files. The data is made available and licensed by Motivate International Inc.

Data Organization & Description:

- File naming convention: Cyclistic_TripData_YYYYMM
- File Type: Converted from csv to xslx format to enable importation to Microsoft SQL.
- File Content: Each excel files contains 13 columns containing information related to ride id, ridership type, ride time and location and location etc. Number of rows varies between 49k to 531k from different excel files.

Data Security:

- Riders' personal identifiable information is hidden through tokenization.
- Original files are backed up in a separate folder.

Data Limitations:

As riders' personal identifiable information is hidden, thus will not be able to connect pass purchases to credit cards numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

3. Process

Tools I have selected for data verification and cleaning:

- Original files are backed up in a separate folder.
- Microsoft **Excel**
- Microsoft **SQL server**

Reasons:

- By scanning through data in Excel worksheet, the general outline and basic information can be found which enable me to get familiarize with the dataset. I can perform simple check on formatting, missing information, sorting and filtering from the spreadsheet as well.

- The 12 datasets combined will contain more than 4 million rows of data. Excel worksheet **limitation is 1,048,576 rows**.
- Thus, Microsoft SQL server is used to perform such task. Microsoft SQL server will be also used to extract and generate new table for desired information which will be used for data visualization via Tableau Public.

Initial assessment on the dataset in Microsoft Excel

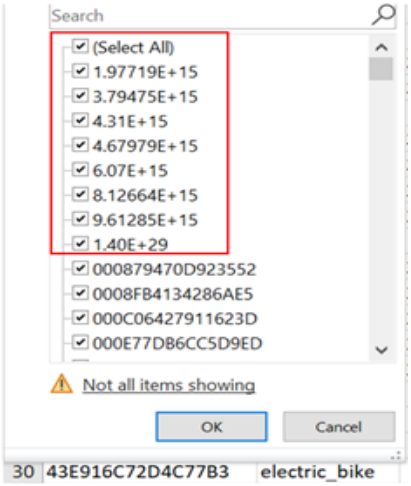
Prior to data cleaning, I used number of rows in each excel sheet to present the total number of rides per month and plot it out using a simple bar chart. It Shows that ridership peaked in Aug 2020 and dropped to the lowest point in Feb 2021 which might have correlation with seasonal change, as weather slowly turn cold from Aug and spring arrives at around March. Keep this in might and check again after data is cleaned.

Data Verification in Microsoft Excel

- Check individual columns for the assigned data type.
- Check for invalid / incorrect / unusable data, remove them if necessary.
- Create a column named “ride length” which calculates the time period length of each ride.
- Create another column named “day_of_week” to calculate the day of the week

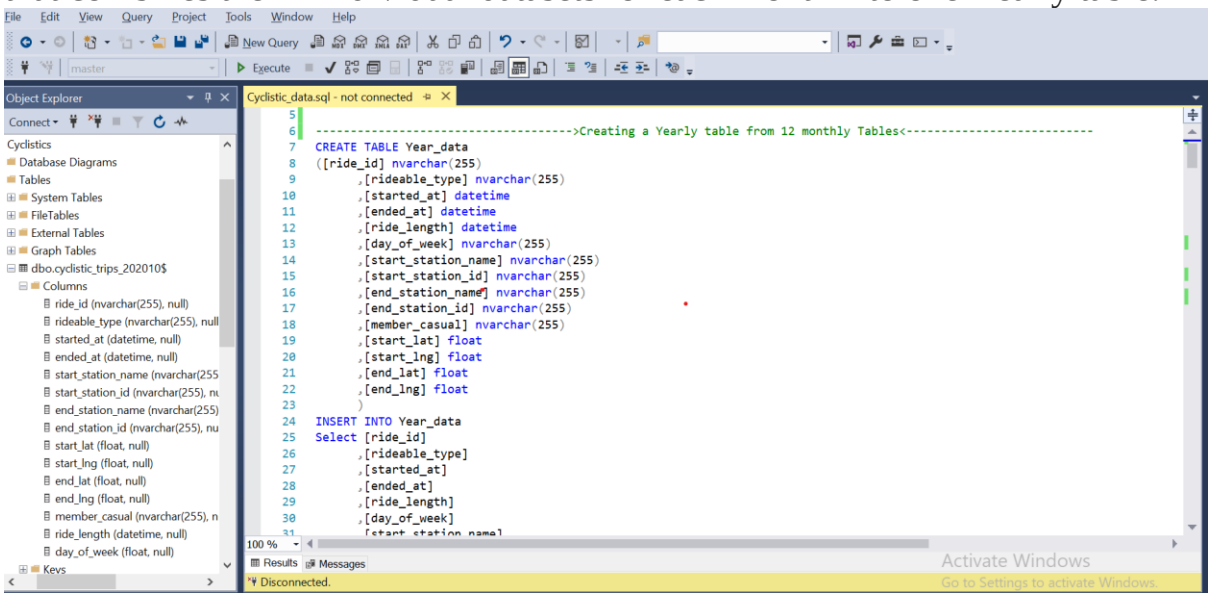
- Some columns contain empty cells which I have used the replace function in Excel to replace empty cell with “NULL” string.

- By using filter function, abnormalities in the ride_id column is spotted. Standard ride_id contains 16 characters. Take note to remove rows that does not contains 16 characters for ride_id.

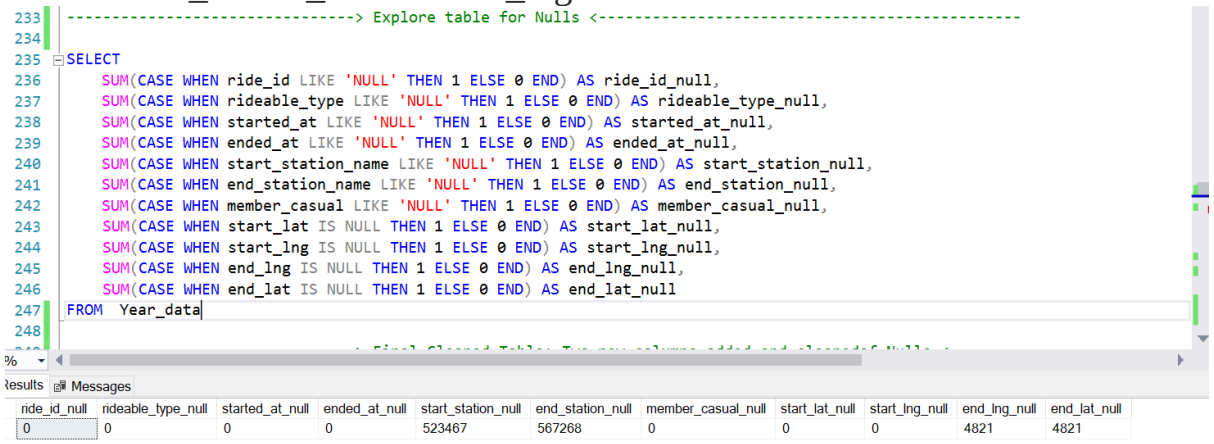


Data Cleaning and Data Manipulation using Microsoft SQL

I started with the **CREATE TABLE** clause, the purpose is to create a new table that combines the 12 individual datasets for each month into one Yearly table.



By scanning through the combined dataset, “NULL’ strings show up from column start_station_name to end_lng. These rows of data are considered to be



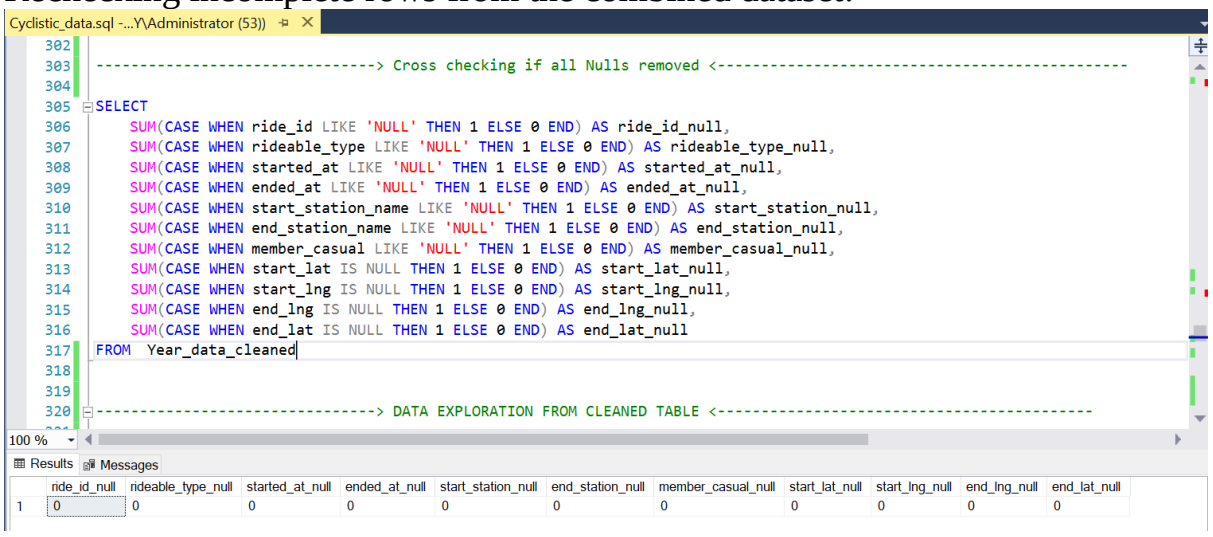
```
----------------------------------> Explore table for Nulls <-----------------------------------------------

SELECT
    SUM(CASE WHEN ride_id LIKE 'NULL' THEN 1 ELSE 0 END) AS ride_id_null,
    SUM(CASE WHEN rideable_type LIKE 'NULL' THEN 1 ELSE 0 END) AS rideable_type_null,
    SUM(CASE WHEN started_at LIKE 'NULL' THEN 1 ELSE 0 END) AS started_at_null,
    SUM(CASE WHEN ended_at LIKE 'NULL' THEN 1 ELSE 0 END) AS ended_at_null,
    SUM(CASE WHEN start_station_name LIKE 'NULL' THEN 1 ELSE 0 END) AS start_station_null,
    SUM(CASE WHEN end_station_name LIKE 'NULL' THEN 1 ELSE 0 END) AS end_station_null,
    SUM(CASE WHEN member_casual LIKE 'NULL' THEN 1 ELSE 0 END) AS member_casual_null,
    SUM(CASE WHEN start_lat IS NULL THEN 1 ELSE 0 END) AS start_lat_null,
    SUM(CASE WHEN start_lng IS NULL THEN 1 ELSE 0 END) AS start_lng_null,
    SUM(CASE WHEN end_lng IS NULL THEN 1 ELSE 0 END) AS end_lng_null,
    SUM(CASE WHEN end_lat IS NULL THEN 1 ELSE 0 END) AS end_lat_null
FROM  Year_data
```

Results | Messages

ride_id_null	rideable_type_null	started_at_null	ended_at_null	start_station_null	end_station_null	member_casual_null	start_lat_null	start_lng_null	end_lng_null	end_lat_null
0	0	0	0	523467	567268	0	0	0	4821	4821

Command **WHERE**, **NOT LIKE AND** was used to remove nulls. Rechecking incomplete rows from the combined dataset.



Cyclistic_data.sql -...Y\Administrator (53))

```
----------------------------------> Cross checking if all Nulls removed <-----------------------------------------------

SELECT
    SUM(CASE WHEN ride_id LIKE 'NULL' THEN 1 ELSE 0 END) AS ride_id_null,
    SUM(CASE WHEN rideable_type LIKE 'NULL' THEN 1 ELSE 0 END) AS rideable_type_null,
    SUM(CASE WHEN started_at LIKE 'NULL' THEN 1 ELSE 0 END) AS started_at_null,
    SUM(CASE WHEN ended_at LIKE 'NULL' THEN 1 ELSE 0 END) AS ended_at_null,
    SUM(CASE WHEN start_station_name LIKE 'NULL' THEN 1 ELSE 0 END) AS start_station_null,
    SUM(CASE WHEN end_station_name LIKE 'NULL' THEN 1 ELSE 0 END) AS end_station_null,
    SUM(CASE WHEN member_casual LIKE 'NULL' THEN 1 ELSE 0 END) AS member_casual_null,
    SUM(CASE WHEN start_lat IS NULL THEN 1 ELSE 0 END) AS start_lat_null,
    SUM(CASE WHEN start_lng IS NULL THEN 1 ELSE 0 END) AS start_lng_null,
    SUM(CASE WHEN end_lng IS NULL THEN 1 ELSE 0 END) AS end_lng_null,
    SUM(CASE WHEN end_lat IS NULL THEN 1 ELSE 0 END) AS end_lat_null
FROM  Year_data_cleaned


----------------------------------> DATA EXPLORATION FROM CLEANED TABLE <-----------------------------------------
```

100 % | Results | Messages

	ride_id_null	rideable_type_null	started_at_null	ended_at_null	start_station_null	end_station_null	member_casual_null	start_lat_null	start_lng_null	end_lng_null	end_lat_null
1	0	0	0	0	0	0	0	0	0	0	0

Next, I proceed to aggregate the data. To find out the ride length of each ride, DATEDIFF clause was used and defined in **MINUTE**. **CASE** clause is used to convert given condition into different strings from Monday to Sunday.

```
Cyclistic_data.sql -...Y\Administrator (53))
268 INSERT INTO Year_data_cleaned
269 Select [ride_id]
270       ,[rideable_type]
271       ,[started_at]
272       ,[ended_at]
273       ,DATEDIFF(minute, started_at, ended_at) AS duration_mins
274       , CASE
275       WHEN day_of_week = 1 THEN 'Sunday'
276       WHEN day_of_week = 2 THEN 'Monday'
277       WHEN day_of_week = 3 THEN 'Tuesday'
278       WHEN day_of_week = 4 THEN 'Wednesday'
279       WHEN day_of_week = 5 THEN 'Thursday'
280       WHEN day_of_week = 6 THEN 'Friday'
281         ELSE
282             'Saturday'
283         END
284             AS Day_Week
285       ,[start_station_name]
286       ,[start_station_id]
287       ,[end_station_name]
288       ,[end_station_id]
289       ,[member_casual]
290       ,[start_lat]
291       ,[start_lng]
292       ,[end_lat]
293       ,[end_lng]
294 From Year_data
```

```
294 From Year_data
295 Where start_station_name NOT LIKE 'NULL'
296 AND end_station_name NOT LIKE 'NULL'
297 AND LEN(ride_id) = 16
298
```

To ensure that all the ride_id only contains 16 characters, **LEN** command is used. There are some ride lengths which is less than 1 minutes. Those will be treated as error rides and filtered out.

Finally, the final table is formed and this completes the part on data cleaning. The next part will focus on data exploration and data visualization.

4 & 5. Analyze & Share Insights

Analyze (Step 4) and Share (Step 5) are combined together in this section.

Tools I have selected for Data Exploration and Data Visualization:

- **Microsoft SQL server**
- **Tableau Public**

To analyze the cleaned data table, my first step is to find out ***how many member / casual cyclists are departing or arriving at different bike stations***. I used **COUNT** and **GROUP BY** command to perform this operation.

To find out number of casual riders departing from different bike stations

Cyclistic_data.sql -...Y\Administrator (53))

```
----------------------------------> DATA EXPLORATION FROM CLEANED TABLE <----------------------------------------



---------------------------------> no.of rides by member type and type of bike<----------------
Select member_casual, count(member_casual) as number_of_rides, rideable_type
From Year_data_cleaned
Group by member_casual, rideable_type

---------------------------------> av. ride duration by member type<-------------------------
Select member_casual, AVG(duration_mins) as average_time_mins
From Year_data_cleaned
Group by member_casual

---------------------->no.of ridestrips weekly by member_type<--------------------------------
Select member_casual, day_week, count(day_week) as number_of_rides
From Year_data_cleaned
Group by day_week, member_casual
Order by member_casual
---------------------> no.of ridestrips hourly by member_type<--------------------
Select DATEPART(HOUR, started_at) as Hour, count(started_at) as number_of_rides , member_casual
From Year_data_cleaned
Group by DATEPART(HOUR, started_at), member_casual
Order by Hour

---------------------> no.of ridestrips monthly by member_type<--------------------
Select DATENAME(MONTH, started_at) as Month, count(started_at) as number_of_rides , member_casual
From Year_data_cleaned
```

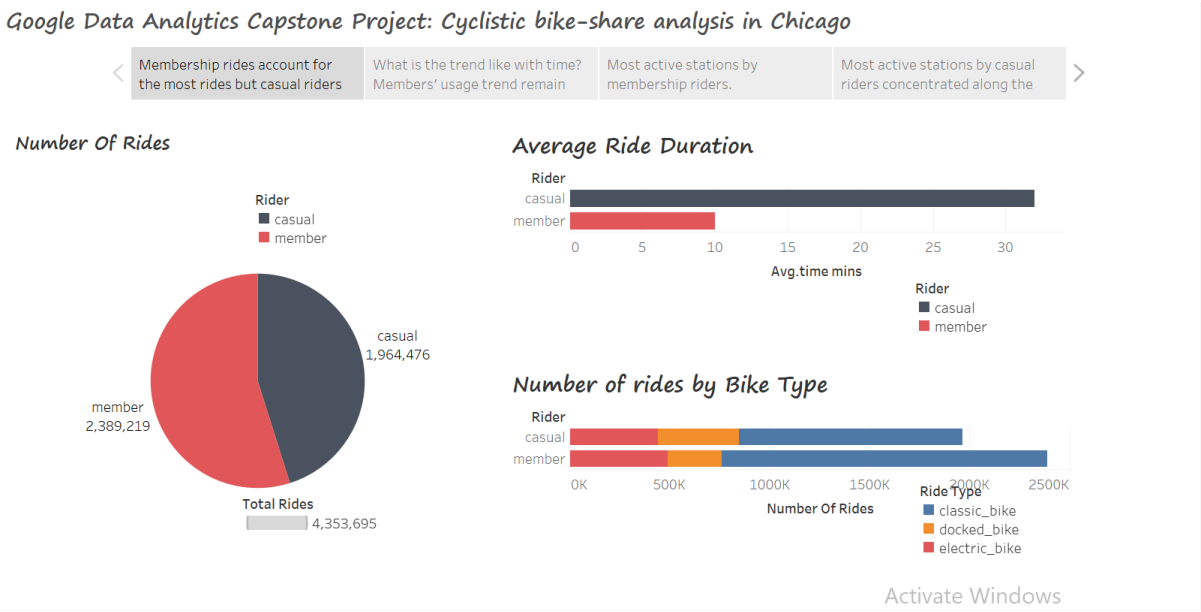
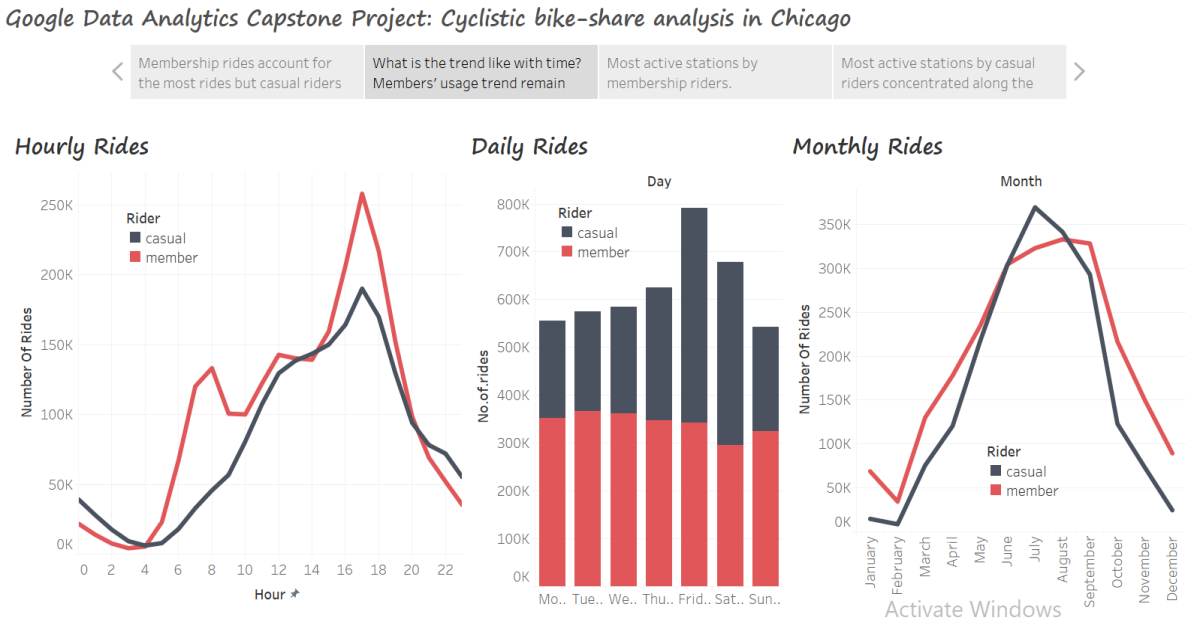


Tableau Visualization for Average Ride Time and Overall Rider Count based on Ridership Type

Average ride time for casual rider is significantly higher compared to member riders. More than half of all riders are member riders.

Next, I wanted to find out ***how the ridership for casual or member varies throughout the year accordingly***. Thus, for the SQL query, I used **COUNT** command to count the number of casual riders, and group them by each day.



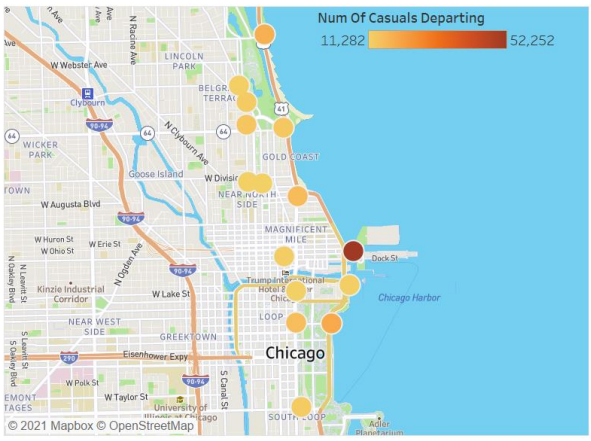
From the Tableau Visualization, it indicated for both casual and member ridership peaked around ***July*** and hit the lowest at ***February*** before rebounding up sharply and it also shows that the trend for member riders is relatively consistent throughout the week, with two peaks at 8am and 5pm with also a slight drop on weekends. In contrast, for casual riders, weekdays bike trips are significantly lower compared for members, and peaking on Friday and Saturday.

```
Cyclistic_data.sql -...Y\Administrator (53))
349
350  ---------------------->location where members depart from<--------------------
351  Select distinct TOP 15 start_station_name, COUNT(member_casual) as num_of_members_departing,start_lat,start_lng
352  From Year_data_cleaned
353  Where member_casual = 'member'
354  Group by start_station_name, start_lat, start_lng
355  Order by num_of_members_departing DESC
356
357  ---------------------->location where casuals depart from<--------------------
358
359  Select distinct TOP 15 start_station_name, COUNT(member_casual) as num_of_casuals_departing,start_lat,start_lng
360  From Year_data_cleaned
361  Where member_casual = 'casual'
362  Group by start_station_name, start_lat, start_lng
363  Order by num_of_casuals_departing DESC
364
365  ---------------------->location where members arrive to<--------------------
366  Select distinct TOP 15 end_station_name, COUNT(member_casual) as num_of_members_arriving,end_lat,end_lng
367  From Year_data_cleaned
368  Where member_casual = 'member'
369  Group by end_station_name, end_lat, end_lng
370  Order by num_of_members_arriving DESC
371
372  ---------------------->location where casuals arrive to<--------------------
373  Select distinct TOP 15 end_station_name, COUNT(member_casual) as num_of_casuals_arriving,end_lat,end_lng
374  From Year_data_cleaned
375  Where member_casual = 'casual'
```

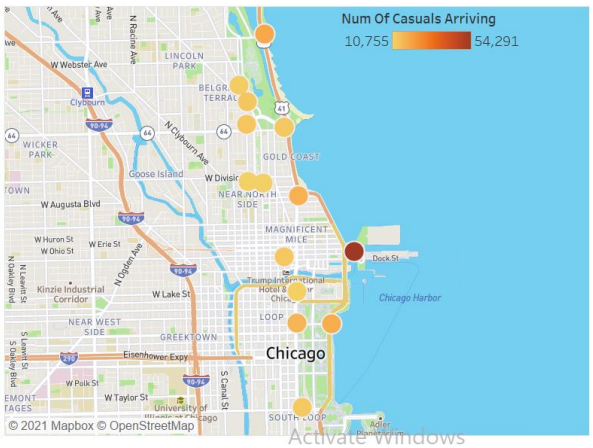
Google Data Analytics Capstone Project: Cyclistic bike-share analysis in Chicago

Membership rides account for the most rides but casual riders | What is the trend like with time? Members' usage trend remain | Most active stations by membership riders. | Most active stations by casual riders concentrated along the

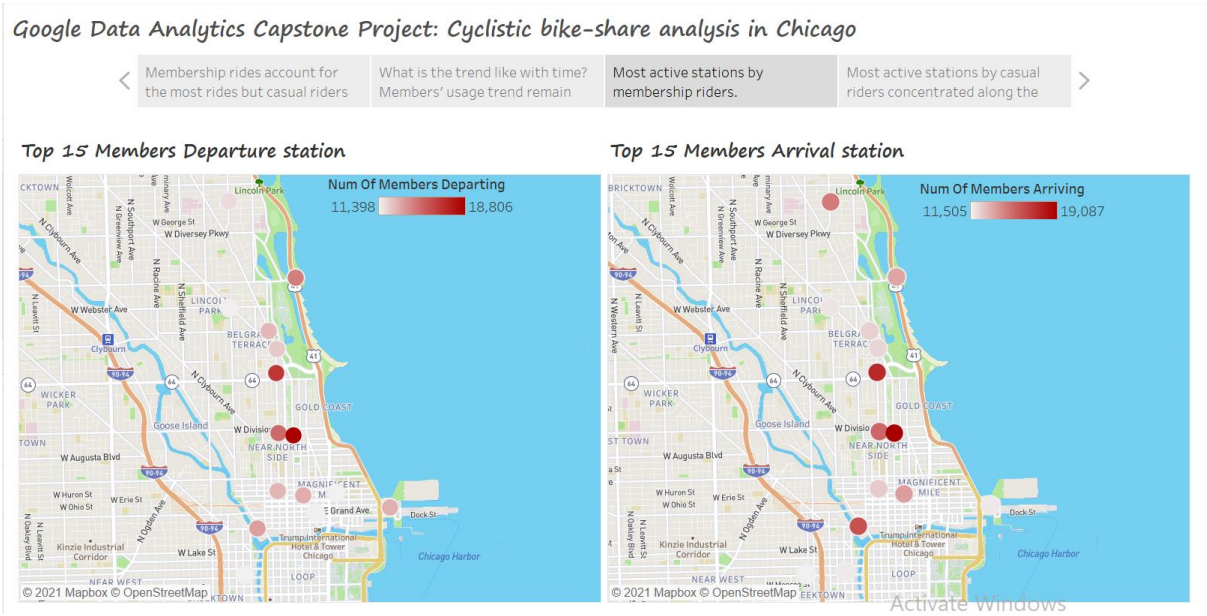
Top 15 Casual Departure station



Top 15 Casual Arrival station



For the visualization above, it presents the **frequency of visit for casual riders** at various bike stations. The bike stations with the highest rider visit frequency can be easily spotted. The top 15 location are stated and coincidently they are all located at close proximity to the coast of the lake.



For the visualization above, it presents the **frequency of visit for member riders** at various bike stations. Compared to the Geodata for casual riders, the visits for respective bike stations are more spread out instead of concentrated at stations near the coast.

6. Act

Summary of the insights gained from Tableau Visualization

- Casual riders are more concentrated around the coast area, whereas member riders are more spread out around the office area. Casual riders also peaked during weekends, there is a high chance they are tourists or families who are visiting the coastline for leisure activities such as sightseeing during the weekend. Longer average ride time for casual rider provide further suggestion on the previous mentioned point.
- Ridership starts to pick up from February and start to decrease in August. It might have correlation to the seasonal changes. As the weather start to get warmer in February (start of Spring), more riders start to cycle, and inversely when the weather to start to turn cold in August (end of Autumn)
- Length of ride for members are relatively shorter compared to casual riders. This might be due to short ride transit from train stations to their offices / home for member rider type.
- More than half of the riders are members, indicating that the company have already sustained some level of loyalty among their bike users. Thus, the company has chance to convert more casual riders to members.

In additional to sharing the insights gathered to *Lily Monero and the executive stakeholder*. I would like to propose a few recommendations based on data evidence:

- Based on the trips made, the **marketing campaign** should be **launched between February to August** as the number of trips made by cyclists **starts to build up**.
- As **casual rider usage** often **peaks on the weekend**, the marketing campaign can include **weekend only membership membership subscription** at lower price to attract casual riders to convert to members
- Modification to membership subscription, such as **ride length-based charges which charges lesser as ride length increases**. This provides more incentive for the member rides to cycle longer distances. With such modification, it could also encourage casual riders to convert to members to enjoy the ride length discounts.