How Does a Bike-Share Navigate Speedy Success?

For this capstone project, I will perform the real-world tasks of a junior data analyst for the marketing team at Cyclists, a bike-share company in Chicago.

To answer business questions, I performed six phases of data analysis: Ask, Prepare, Process, Analyze, Share and Act.

The tools I used in this analysis:

I did the data familiarization and cleaning using **Microsoft Excel**.

I performed the research and data processing directly using Google **BigQuery SQL**.

Data visualization was performed using **Tableau**.

In this analysis, I wanted to use as many basic tools as possible in practice, so the choice was exactly that.

Some basic information about the program of the company for which the analysis is being performed:

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 them to commute to work each day.

Stakeholders in this analysis:

Lily Moreno: The director of marketing. 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.

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

1. ASK

The purpose of this analysis is to develop marketing strategies in order to increase the number of annual members using this Cyclistic program. To do this, it is necessary to better understand how casual riders differ from annual members, and why random participants buy a membership in this Cyclistic program.

2. PREPARE

Data source: data on users of the Cyclistic program was taken for the last 12 months from 12.2021 to 01.2022. Each data file contains data for a month.

File Content: Each excel file contains 13 columns containing information related to ride id, rideable type, start stations and end stations names and information about stations etc.

Data Security: Rider's personal identifiable information is hidden through tokenization.

Data Limitations: As rider's 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

To familiarize myself with the data, I chose the Microsoft Excel tool. With the help of which I was able to understand how the data should look and find distorted data, gaps and errors in the data.

It is not possible to combine all the files and inspect the annual data at once because the file will contain too many rows and greatly increases the possible limit for use in Microsoft Excel.

During my familiarization and data cleaning, I was able to find data in several columns that do not correspond to a certain format that each column carries.

Examples:

In the ride id column, each unique id must consist of 16 unique characters, so all data that did not correspond to this format was cleared.

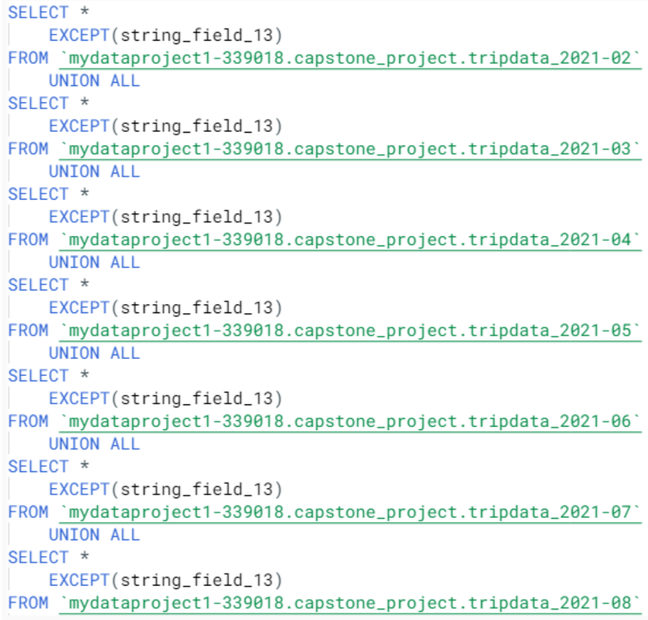
In the columns start station, start station id, end station, end station id, certain data were identified that related to test bike rides for the purpose of their technical verification, these names had the prefix WH-TEST and a certain unique id.

Then, using the available data, I calculated the length of each trip and also on what day of the week it was, then I decided to delete the duration of the trip less than one minute because they are not valuable for our analysis, columns with incorrect data were also deleted.

4. ANALYZE

To work with data and perform analysis, I selected the Google BigQuery SQL tool.

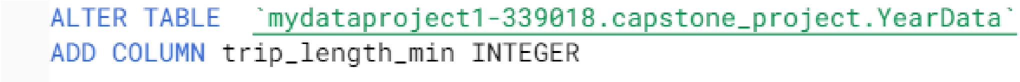
Using SQL, I combined each file with travel data for a month into one file for the whole year, for faster and more convenient analysis of all data for the past year at once:



```
SELECT *
    EXCEPT(string_field_13)
FROM `mydataproject1-339018.capstone_project.tripdata_2021-02`
    UNION ALL
SELECT *
    EXCEPT(string_field_13)
FROM `mydataproject1-339018.capstone_project.tripdata_2021-03`
    UNION ALL
SELECT *
    EXCEPT(string_field_13)
FROM `mydataproject1-339018.capstone_project.tripdata_2021-04`
    UNION ALL
SELECT *
    EXCEPT(string_field_13)
FROM `mydataproject1-339018.capstone_project.tripdata_2021-05`
    UNION ALL
SELECT *
    EXCEPT(string_field_13)
FROM `mydataproject1-339018.capstone_project.tripdata_2021-06`
    UNION ALL
SELECT *
    EXCEPT(string_field_13)
FROM `mydataproject1-339018.capstone_project.tripdata_2021-07`
    UNION ALL
SELECT *
    EXCEPT(string_field_13)
FROM `mydataproject1-339018.capstone_project.tripdata_2021-08`
```

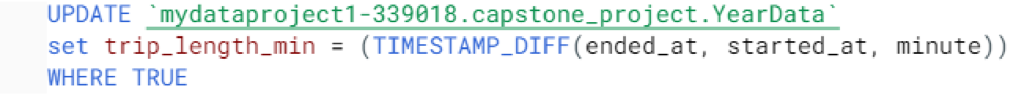
SQL query to connect all 12 files into one

After combining all the files into one, I created two more columns and calculated the length of each trip, as well as what day of the week it was:



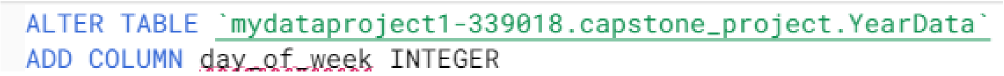
```
ALTER TABLE  `mydataproject1-339018.capstone_project.YearData`
ADD COLUMN trip_length_min INTEGER
```

Creating a new column "trip_length_min"



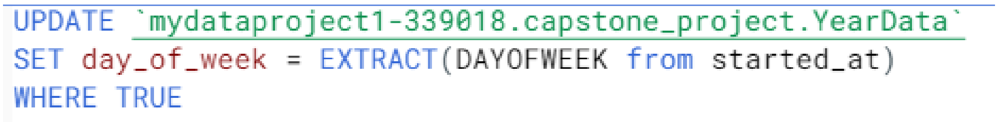
```
UPDATE `mydataproject1-339018.capstone_project.YearData`
set trip_length_min = (TIMESTAMP_DIFF(ended_at, started_at, minute))
WHERE TRUE
```

Calculating the length of each trip



```
ALTER TABLE `mydataproject1-339018.capstone_project.YearData`
ADD COLUMN day_of_week INTEGER
```

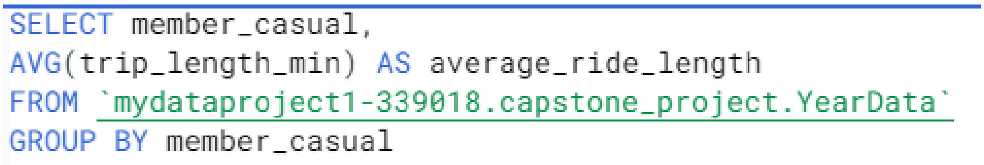
Creating a new column "day_of_week"



```
UPDATE `mydataproject1-339018.capstone_project.YearData`
SET day_of_week = EXTRACT(DAYOFWEEK from started_at)
WHERE TRUE
```

Calculation of the day of the week in which the trip was

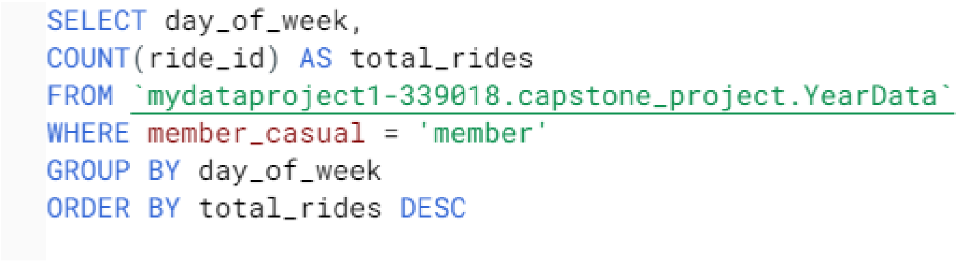
After calculating the length of each trip and I calculated the average ride time to compare the average time of a casual user of the program and the annual member:



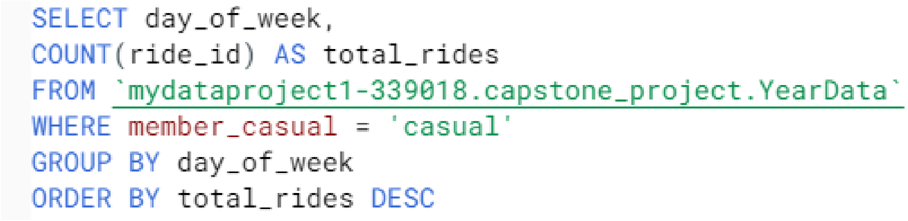
```
SELECT member_casual,
AVG(trip_length_min) AS average_ride_length
FROM `mydataproject1-339018.capstone_project.YearData`
GROUP BY member_casual
```

Analysis the average duration of a trip over the past year

Next, I found out which days of using the program are the most popular, for casual users and annual members separately for further comparison:



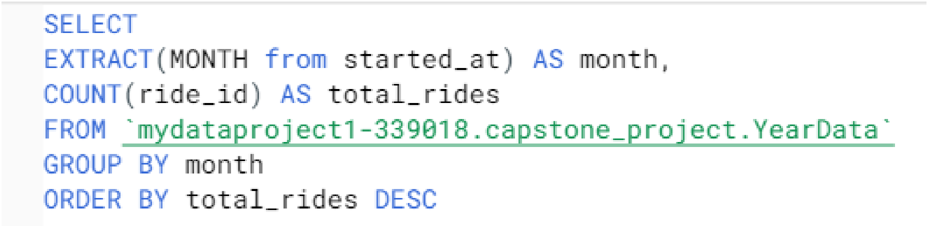
```
SELECT day_of_week,
COUNT(ride_id) AS total_rides
FROM `mydataproject1-339018.capstone_project.YearData`
WHERE member_casual = 'member'
GROUP BY day_of_week
ORDER BY total_rides DESC
```



```
SELECT day_of_week,
COUNT(ride_id) AS total_rides
FROM `mydataproject1-339018.capstone_project.YearData`
WHERE member_casual = 'casual'
GROUP BY day_of_week
ORDER BY total_rides DESC
```

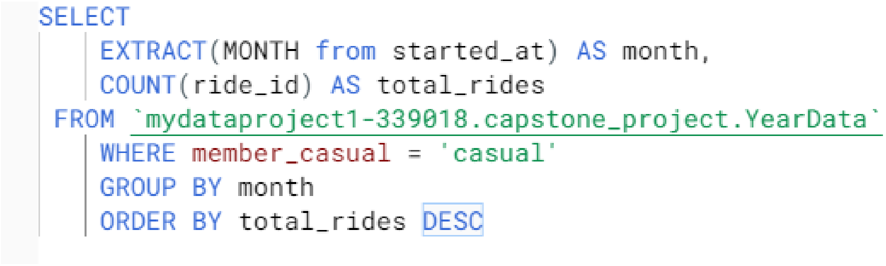
Analysis of the peak days of the week of using the program for ordinary users and annual members

Next, I found out which months are the most popular among users of the Cyclistic program:

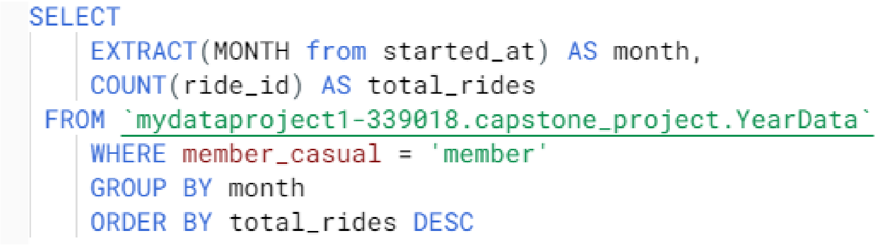


```
SELECT
EXTRACT(MONTH from started_at) AS month,
COUNT(ride_id) AS total_rides
FROM `mydataproject1-339018.capstone_project.YearData`
GROUP BY month
ORDER BY total_rides DESC
```

Analysis of the most peak months of using the program



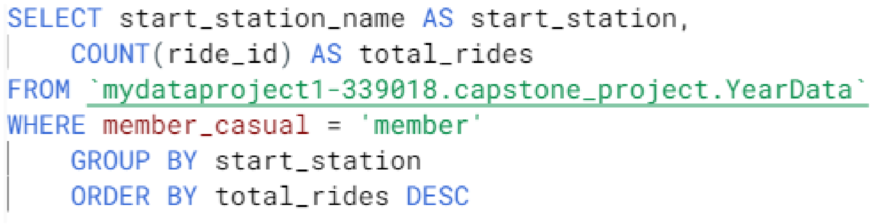
```
SELECT
    EXTRACT(MONTH from started_at) AS month,
    COUNT(ride_id) AS total_rides
 FROM `mydataproject1-339018.capstone_project.YearData`
    WHERE member_casual = 'casual'
    GROUP BY month
    ORDER BY total_rides DESC
```



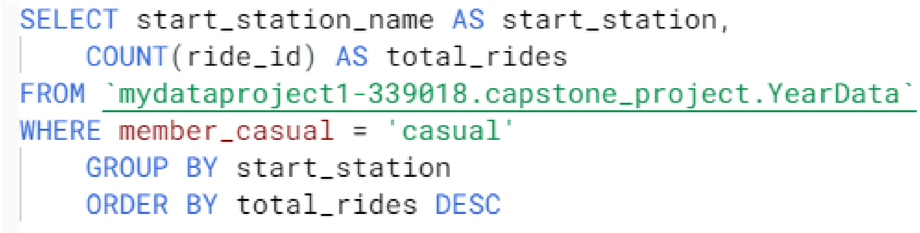
```
SELECT
    EXTRACT(MONTH from started_at) AS month,
    COUNT(ride_id) AS total_rides
 FROM `mydataproject1-339018.capstone_project.YearData`
    WHERE member_casual = 'member'
    GROUP BY month
    ORDER BY total_rides DESC
```

Analysis of the most peak months of using the program for ordinary users and annual member

Next, I analyzed the most popular stations on which casual riders and annual members, for further comparison and the concept of preferences of each of the parties:

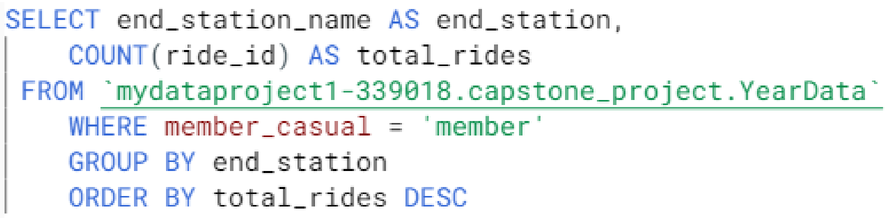


```
SELECT start_station_name AS start_station,
    COUNT(ride_id) AS total_rides
FROM `mydataproject1-339018.capstone_project.YearData`
WHERE member_casual = 'member'
    GROUP BY start_station
    ORDER BY total_rides DESC
```

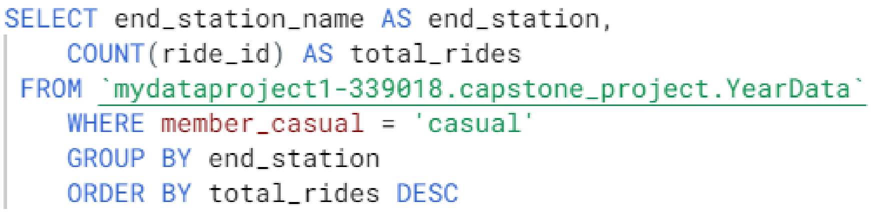


```
SELECT start_station_name AS start_station,
    COUNT(ride_id) AS total_rides
FROM `mydataproject1-339018.capstone_project.YearData`
WHERE member_casual = 'casual'
    GROUP BY start_station
    ORDER BY total_rides DESC
```

Analysis of the most popular departure stations among regular users and annual members



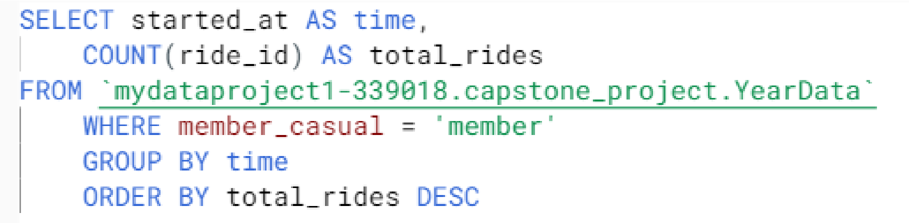
```
SELECT end_station_name AS end_station,
    COUNT(ride_id) AS total_rides
 FROM `mydataproject1-339018.capstone_project.YearData`
    WHERE member_casual = 'member'
    GROUP BY end_station
    ORDER BY total_rides DESC
```



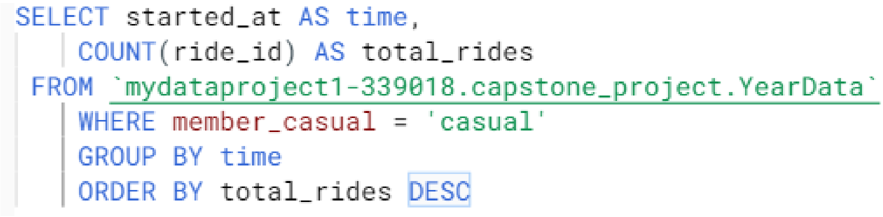
```
SELECT end_station_name AS end_station,
    COUNT(ride_id) AS total_rides
 FROM `mydataproject1-339018.capstone_project.YearData`
    WHERE member_casual = 'casual'
    GROUP BY end_station
    ORDER BY total_rides DESC
```

Analysis of the most popular arrival stations for regular users and annual members

Next, I analyzed the most popular time during the day of cycling, to compare the indicators of casual riders and annual members:



```
SELECT started_at AS time,
    COUNT(ride_id) AS total_rides
FROM `mydataproject1-339018.capstone_project.YearData`
    WHERE member_casual = 'member'
    GROUP BY time
    ORDER BY total_rides DESC
```



```
SELECT started_at AS time,
    COUNT(ride_id) AS total_rides
 FROM `mydataproject1-339018.capstone_project.YearData`
    WHERE member_casual = 'casual'
    GROUP BY time
    ORDER BY total_rides DESC
```

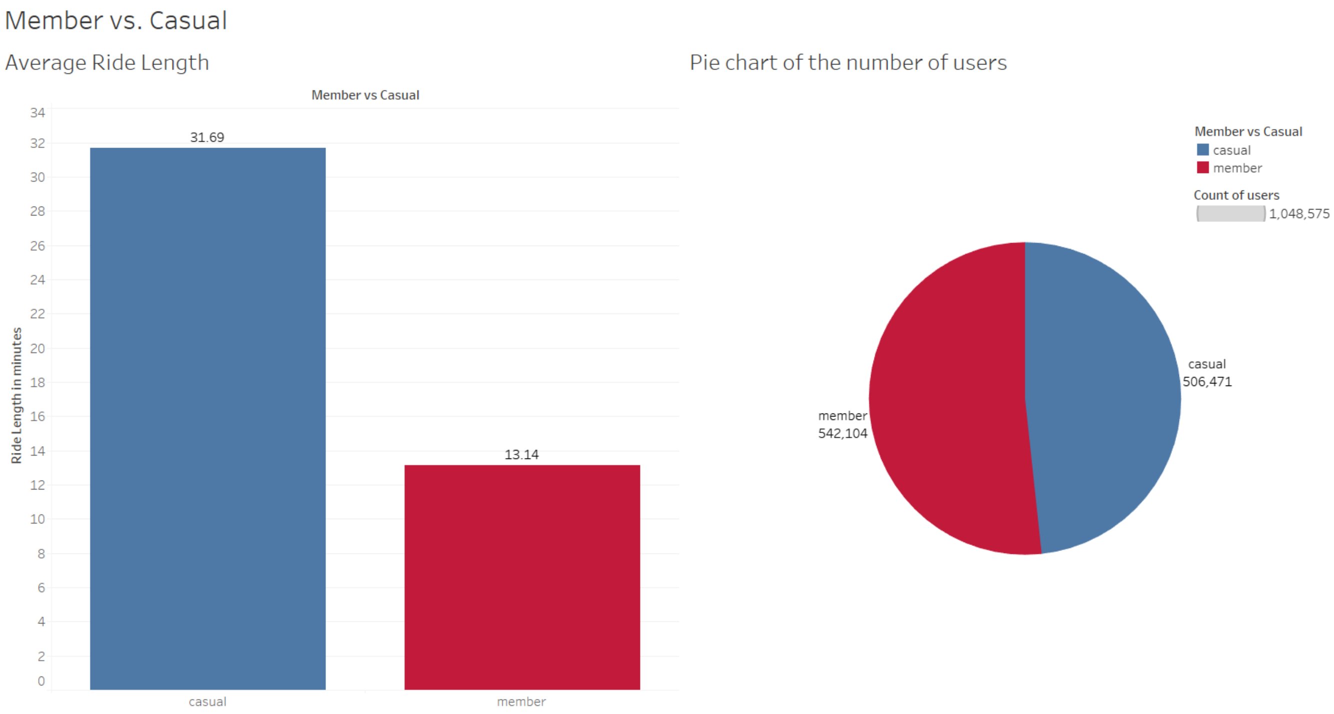
Analysis of peak hours of the day for regular users and annual members

An important insight was found during the analysis that members make the largest number of trips on weekdays, unlike casual riders of the service, who, in turn, have the largest number of trips on weekends. Based on this, we can draw an important conclusion that annual members mostly use the service on a daily basis, unlike ordinary users, who in turn use the service for weekend rides. This theory is confirmed by comparing the most popular stations among members and casual riders and comparing the time of day. This analysis confirms our theory about the difference in the use of the service between annual members and casual users of the Cyclistic program.

5. SHARE

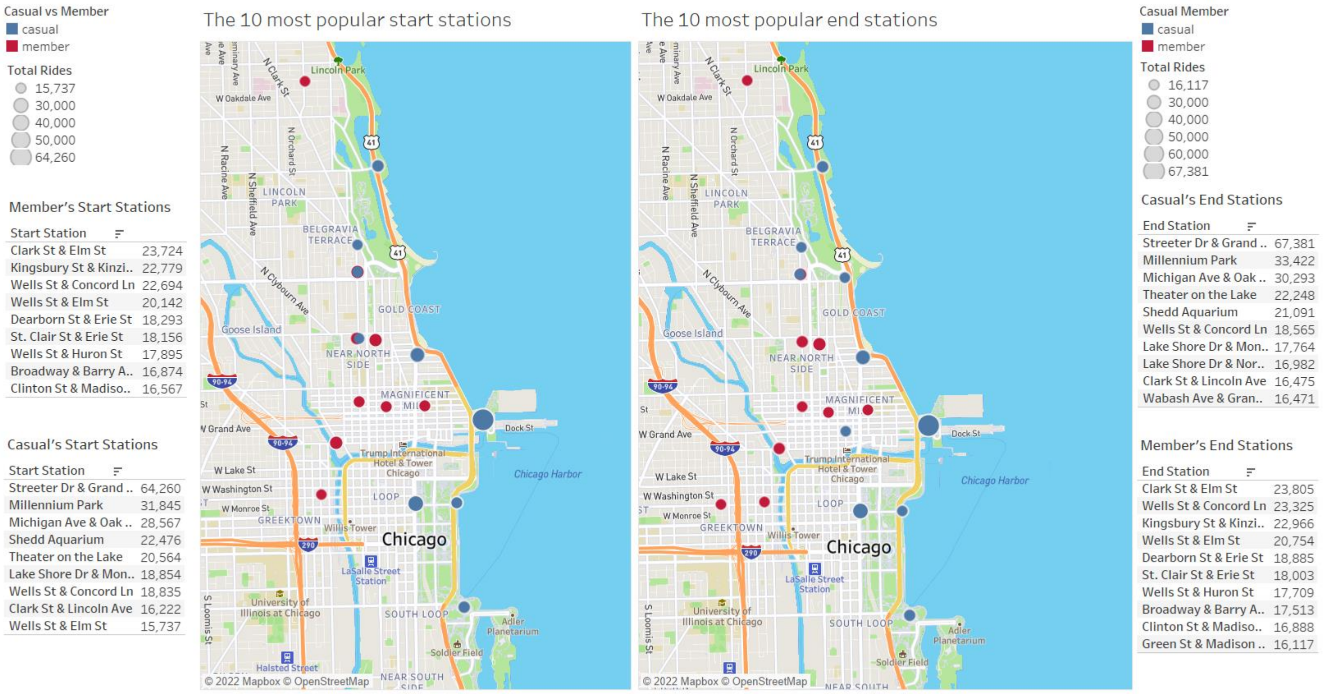
This section will consist of visualizing the data that I analyzed using the SQL tool.

In this panel, I have displayed a comparison of the total number of regular users and annual membership holders who have used the program over the past year, as well as a comparison of their average one-time trip time.



Visualization of average travel time and comparison of the number of regular users and annual membership holders over the past year

Next, I compared in which parts of the city the arrival and departure stations are most popular with ordinary users and annual members.



Casual vs Member
🟦 casual
🟥 member

Total Rides
● 15,737
● 30,000
● 40,000
● 50,000
● 64,260

Member's Start Stations

Start Station	
Clark St & Elm St	23,724
Kingsbury St & Kinzi..	22,779
Wells St & Concord Ln	22,694
Wells St & Elm St	20,142
Dearborn St & Erie St	18,293
St. Clair St & Erie St	18,156
Wells St & Huron St	17,895
Broadway & Barry A..	16,874
Clinton St & Madiso..	16,567

Casual's Start Stations

Start Station	
Streeter Dr & Grand ..	64,260
Millennium Park	31,845
Michigan Ave & Oak ..	28,567
Shedd Aquarium	22,476
Theater on the Lake	20,564
Lake Shore Dr & Mon..	18,854
Wells St & Concord Ln	18,835
Clark St & Lincoln Ave	16,222
Wells St & Elm St	15,737

The 10 most popular start stations

The 10 most popular end stations

Casual Member
🟦 casual
🟥 member

Total Rides
● 16,117
● 30,000
● 40,000
● 50,000
● 60,000
● 67,381

Casual's End Stations

End Station	
Streeter Dr & Grand ..	67,381
Millennium Park	33,422
Michigan Ave & Oak ..	30,293
Theater on the Lake	22,248
Shedd Aquarium	21,091
Wells St & Concord Ln	18,565
Lake Shore Dr & Mon..	17,764
Lake Shore Dr & Nor..	16,982
Clark St & Lincoln Ave	16,475
Wabash Ave & Gran..	16,471

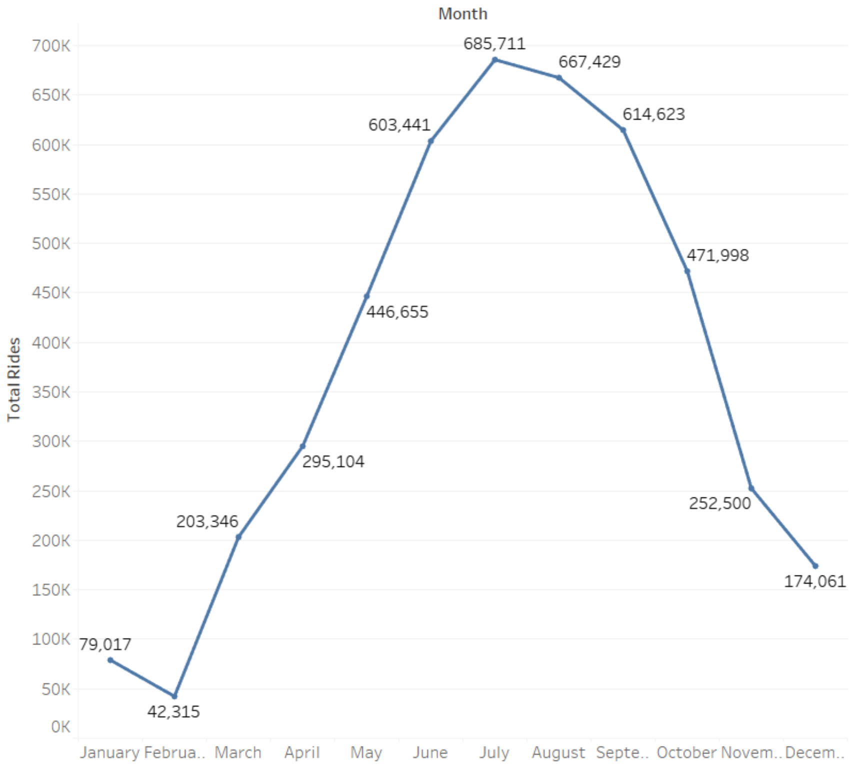
Member's End Stations

End Station	
Clark St & Elm St	23,805
Wells St & Concord Ln	23,325
Kingsbury St & Kinzi..	22,966
Wells St & Elm St	20,754
Dearborn St & Erie St	18,885
St. Clair St & Erie St	18,003
Wells St & Huron St	17,709
Broadway & Barry A..	17,513
Clinton St & Madiso..	16,888
Green St & Madison ..	16,117

Visualization of the most popular departure and arrival stations between regular users and holders of annual memberships for the last year

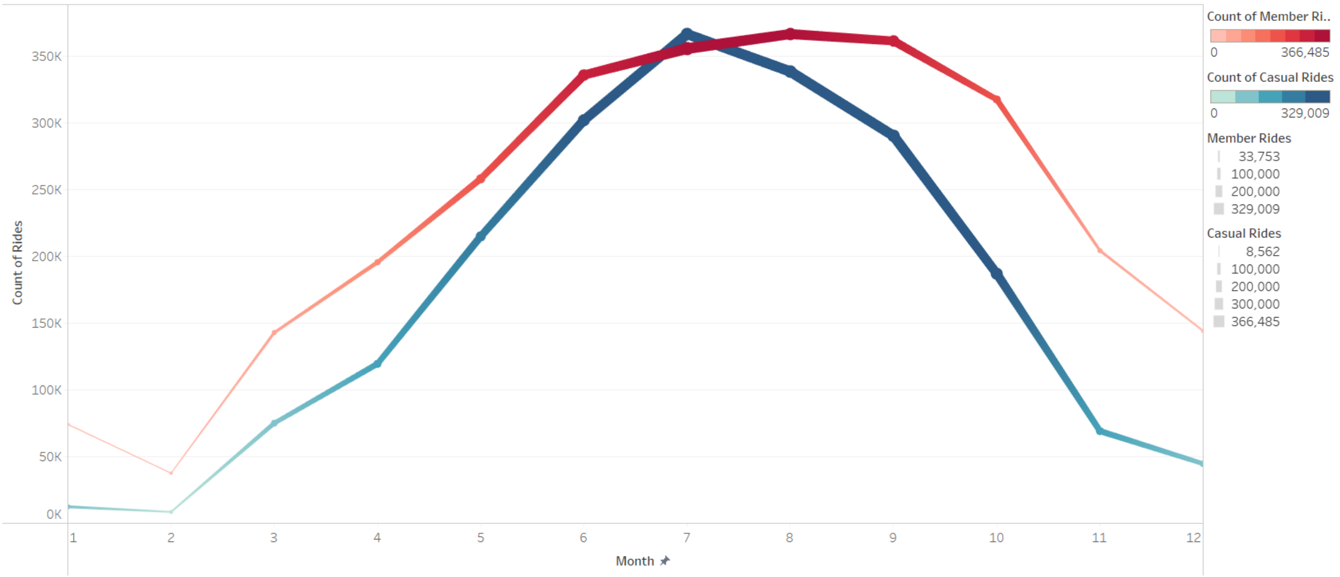
Next, I was interested to find out at what time of the year, and in which months specifically, users most often use the program. And then I visualized for comparison the peak months of ordinary users and annual members.

Total Trips by Month



Visualization of the popularity of months of using the program over the past year

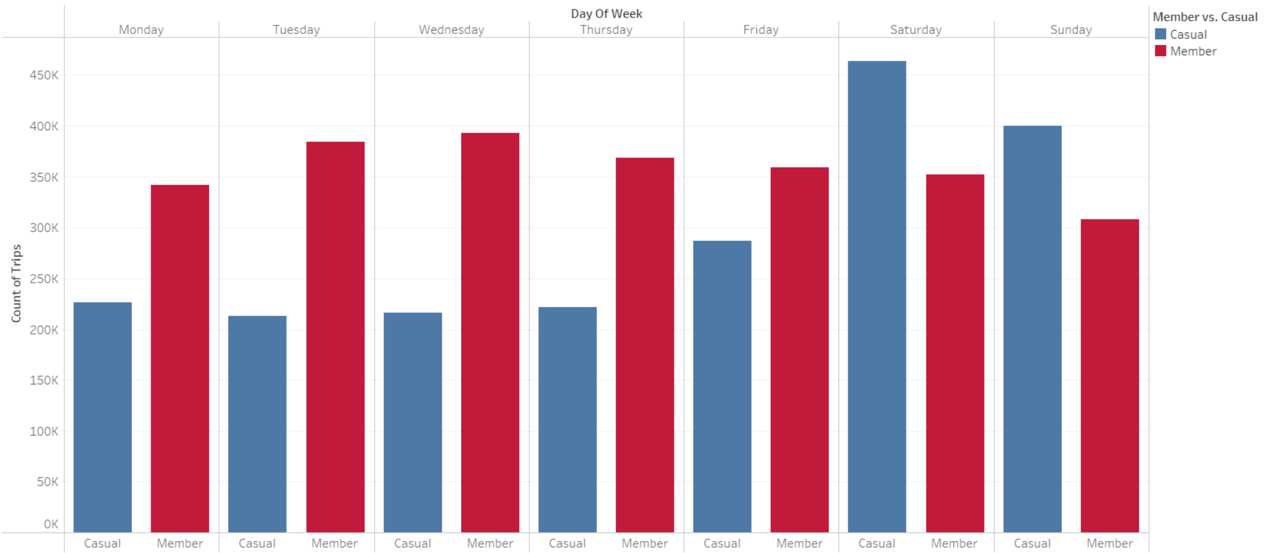
Number of Trips by Month



Visualization of the comparison of the popularity of the months of using the program over the past year between regular users and annual members

Then I went further and analyzed, and then visualized a comparison of the most popular days of the week of using the program between an ordinary user and an annual member.

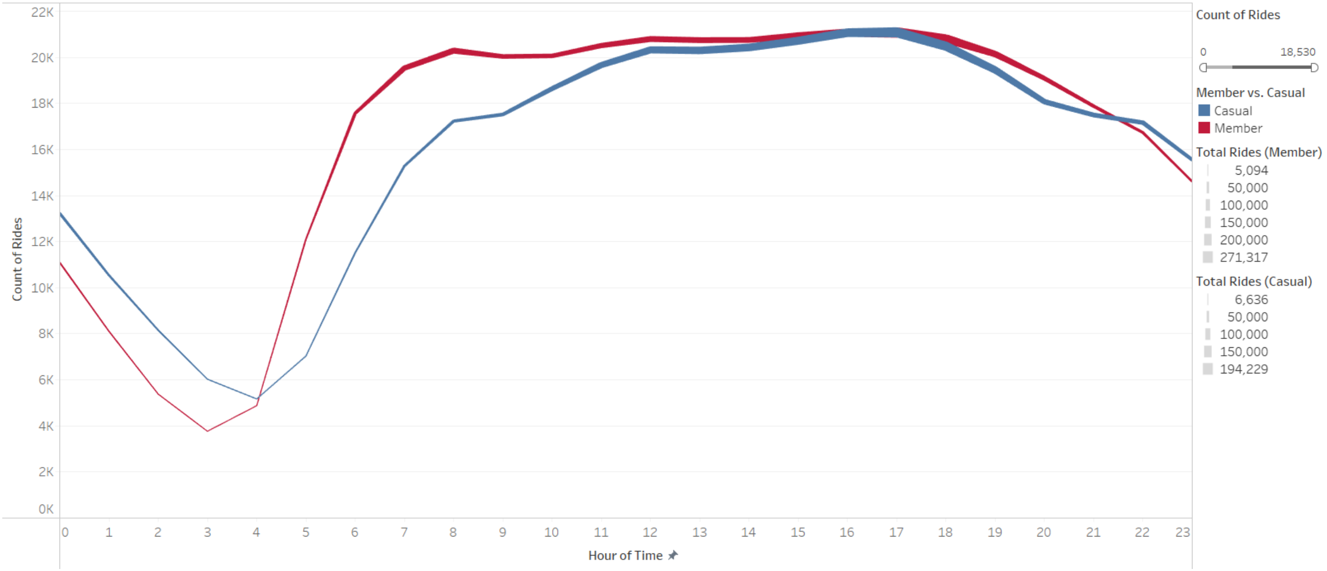
Number of Trips by Day of the Week



Visualization of the popularity of the days of the week of using the program over the past year between regular users and annual members

And at the end, I visualized at what time of the day ordinary users and annual members use the program most often and then visualized comparisons of these indicators between them.

Number of Trips by Hours of the Day



Visualization of the peak time of day of program usage over the past year between regular users and annual members

6. ACT

Based on the analysis and visualization of the relevant data, this section will describe the main recommendations aimed at a marketing strategy to successfully convert the largest number of random users into annual members of the Cyclistic program.

Recommendations:

- A promotional campaign to promote the program for ordinary users is recommended to be held in July (June - July - August - September, depending on the length of the promo campaign that the company wants to conduct).

Based on the analysis and visualization of the analyzed data, it was revealed that the most popular time of the year for using the Cyclistic program by ordinary users is the summer season of the year, as well as the beginning of autumn. Namely months: **June, July, August, September**.

- It is also recommended to conduct a promotional campaign on the most popular stations among ordinary users in order to reach the largest number of ordinary users with the Cyclistic program. (the number of stations depends on the budget allocated for the promo campaign)

Based on the analysis and visualization of the analyzed data, it can be concluded that ordinary users of the program most often (twice as often as all other stations) use the **Streeter Dr & Grand Ave station** because this station is located near Lake Michigan, and there are a large number of tourists, as well as ordinary vacationers on weekends. This is also confirmed by the analysis of the comparison of the peak days of the week of use by the program for ordinary users.

- It is recommended to use promotions related to the more profitable purchase of annual membership so that ordinary users are more inclined to use membership on a regular basis. This recommendation requires an additional questionnaire or feedback after the completion of the bike ride in order to understand what attracts or repels an ordinary user from acquiring an annual membership.

Based on the analysis of all available data on Cyclistic program users and the comparison of indicators between an ordinary rider and annual membership holders, it can be concluded that ordinary riders most often use the service for weekend bike rides. This statement is confirmed by the following indicators:

1. The most popular stations. For ordinary riders - **near, or in parks, or near a lake.** For annual members - **near the subway, shops, shopping centers or ordinary buildings.**
2. Peak days of the week of using the program. Annual members have **Tuesdays, Wednesdays (weekdays).** For regular riders - **Saturday, Sunday (weekends).**
3. Peak time of day. Annual members begin to actively use the service from **8 a.m.**, while regular riders from **12 p.m.**

Thank you for taking the time to me and my project!

The analysis was performed by Maksym Artemenko