Introduction

This case study is my capstone project for the GOOGLE DATA ANALYTICS CERTIFICATE. I have chosen to work with the fictional company Cyclistic for my analysis.
Cyclistic is a bike sharing company located in Chicago, looking for recommendations following a thorough analysis of their DATA. Although Cyclistic is a fictitious company, the data I will use for this project is real data continuously collected from a bike sharing project based in Chicago.

Scenario

Cyclistic is a bike-share program that features more than 5,800 bicycles and 600 docking stations. Bikes can be undocked from one station and docked the bike back at any other station around the city.

Management has tasked the Data Analytics team with the goal of designing marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics, and finding any trends within the data.

♦ ASK

Business objective

Cyclistic is looking to create a marketing campaign to increase profitability by converting casual rider to annual members. As a junior Analyst, I was task with answering the following question:

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

♦ PREPARE

Data source

The Data is uploaded on a monthly basis on the company's website as CSV format. For this project, the last twelve months of data were used (**05/01/2021 to 04/30/2022**).

LICENSE

The company has granted a non-exclusive, royalty-free, limited, perpetual license to access, reproduce, analyze, copy, modify, distribute in your product or service and use the Data for any lawful purpose.

How is the data organized?

After downloading the CSV files, I uploaded them to Microsoft Excel in order to get to know the data I am dealing with. The data is organized into 13 columns with a unique rider ID. The number of rows varies Month from Month between 500K and 800K rows.

Tools

After a preliminary investigation, I decided to use:

- **EXCEL** for the sorting and organizing of the data.
- **SQL** for the data cleaning, exploration and analysis.
- **Tableau** to create visualizations of my finding.

DATA ORGANIZATION AND VERIFICATION WITH EXCEL

I started by making a copy and saving the original data as a backup. I then opened each individual files with excel to start familiarizing myself with the data.

Problems with the Data

After looking at the data, fixing the headers and adjusting the columns and the rows, I started sorting and filtering the columns to start my investigation. The following are a few pre-processing problems I found with the data that I took note of in order to investigate deeper in SQL:

- Duplicates: certain sheet contain duplicates of data that need to be cleaned.

- Format: Some station names contain “(temp)” or (*). Furthermore, “Ride_Id” should be 16 characters long and many rows are not formatted as such.

- Missing fields: many rows contain missing values which may be unusable for my analysis.

- name inconsistencies: some station names are not valid as they are test, repair or maintenance stations.

Before moving the data to SQL for processing, I decided to organize my data better to make it easier to clean and analyze.

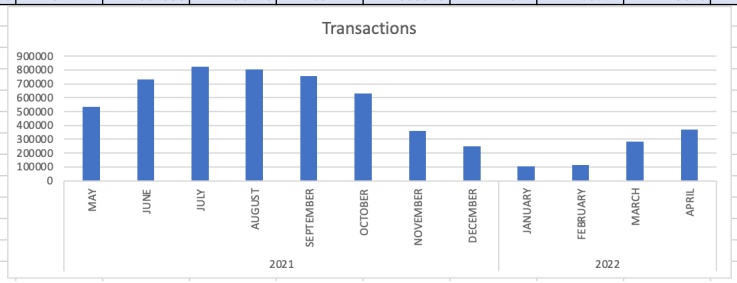
First, I decided to fill anywhere with missing data with “NULL” in order to better track it. I then created a column where I would calculate the ride length, by subtracting the arrival time of each ride from the starting time, to give me an idea of how long the trips may be on average. After creating this column, I

formatted the start and end date from “MM-DD-YYYY” to a more SQL appropriate format “YYYY-MM-DD”. I finally opted to separate the days, months and time into individual column in order to allow for more granulated analysis of the data.

Initial hypothesis

Before cleaning the data, I wanted to see what the data looked like throughout the year. I took the number of rows from each of the sheets and mapped it on a bar graph by months in Excel.

Year	2021								2022				
Month	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER	JANUARY	FEBRUARY	MARCH	APRIL	Total
Rides	531635	729596	822411	804353	756148	631227	359979	247541	103771	115610	284043	371250	5757564



From the result, it seems like the company offer a very seasonal service with the summer months being significantly higher than Winter.

♦ PROCESS

I opted for SQL over EXCEL to process the data because of the size of the documents. Compiling all the data together before cleaning will amount to 5,757,564 rows of Data, and Microsoft

EXCEL only allows for 1,048,576 rows which would be problematic for the case study.

I first removed the duplicates in EXCEL using the “remove duplicates” function and uploaded each individual EXCEL sheets (saved as CSV) to SQL.

DATA CLEANING WITH SQL.

I used the “UNION” function to put all the data together into a single “YEAR_TABLE” that I could later alter as I go.

```
Select *
    INTO [cycle_DB].[dbo].[Year_Table]
FROM
(
Select ride_id,rideable_type, start_station_name,end_station_name,
member_casual,day_of_week,month
FROM [cycle_DB].[dbo].['202105']
UNION
Select ride_id, rideable_type, start_station_name,end_station_name,
member_casual,day_of_week,month
FROM [cycle_DB].[dbo].['202106']
UNION
Select ride_id,rideable_type, start_station_name,end_station_name,
member_casual,day_of_week,month
FROM [cycle_DB].[dbo].['202107']
UNION
Select ride_id,rideable_type, start_station_name,end_station_name,
member_casual,day_of_week,month
FROM [cycle_DB].[dbo].['202108']
UNION
```

The next step was to deal with the missing data that appeared as “NULL” in my table. After looking into the possibilities of filling the data, I decided that the best course of action was to

remove any rows with missing information as they would be of no use to the analysis. 790207 rows where removed.

```
DELETE
FROM [cycle_DB].[dbo].[Year_Table]
WHERE start_station_name LIKE '%NULL%'
AND end_station_name LIKE '%NULL%'
```

Then, I used a CASE statement to change the day and month columns from numbers to characters to make the table easier to read.
For this, I started with the months, and altered the table to replace the month number with the month name:

```
UPDATE [cycle_DB].[dbo].[Year_Table]
SET month_year = CASE
    WHEN month = 1 THEN 'January'
    WHEN month = 2 THEN 'February'
    WHEN month = 3 THEN 'March'
    WHEN month = 4 THEN 'April'
    WHEN month = 5 THEN 'May'
    WHEN month = 6 THEN 'June'
    WHEN month = 7 THEN 'July'
    WHEN month = 8 THEN 'August'
    WHEN month = 9 THEN 'September'
    WHEN month = 10 THEN 'October'
    WHEN month = 11 THEN 'November'
    ELSE 'December'
    END
```

Then I did the same with the days of the week using a similar query.

```
UPDATE [cycle_DB].[dbo].[Year_Table]
SET week_day = CASE
    WHEN day_of_week = 1 THEN 'Sunday'
    WHEN day_of_week = 2 THEN 'Monday'
    WHEN day_of_week = 3 THEN 'Tuesday'
    WHEN day_of_week = 4 THEN 'Wednesday'
    WHEN day_of_week = 5 THEN 'Thursday'
    WHEN day_of_week = 6 THEN 'Friday'
    ELSE 'Saturday'
    END
```

To make sure the ride_id (Primary Key) stayed consistent, filtered and removed any ride_id that had less than 16 characters

```
DELETE
FROM [cycle_DB].[dbo].[Year_Table]
WHERE LEN(ride_id) <> 16
```

Next, I removed the “LBS-WH-TEST” and the “DIVVY CASSETTE REPAIR MOBILE STATION”. According to the company, these are trips that are taken by staff as they service and inspect the system and do not count as rides in our analysis.

```sql
DELETE
FROM [cycle_DB].[dbo].[Year_Table]
WHERE start_station_name LIKE '%TEST%'
AND start_station_name LIKE '%REPAIR%'
```

A similar query was used for the “end_station_name” column.

Using both the TRIM and REPLACE functions, I cleaned both the start and end station names with “*” and “(temp)” and I proceeded to TRIM the station names to make sure no extra spaces were left. Doing this allowed me to have clean and unique station names, which we will be able to match with their respective geographical coordinates later.

```sql
UPDATE [cycle_DB].[dbo].[Year_Table]
SET clean_start_station = TRIM(
    REPLACE( REPLACE(start_station_name, '*',''), '(TEMP)','')
```

```sql
UPDATE [cycle_DB].[dbo].[Year_Table]
SET clean_end_station = TRIM (
    REPLACE(REPLACE(end_station_name, '*',''), '(TEMP)','')
```

Finally, I created my “Final_Table” with the information necessary to start my analysis.

```
SELECT *
INTO [cycle_DB].[dbo].[Final_table]
    FROM (
    SELECT ride_id, rideable_type, member_casual, week_day,
    month_year,clean_start_station,clean_end_station, CAST(crid.total_time as time) as ride_lengt
    FROM [cycle_DB].[dbo].[Year_Table]
```

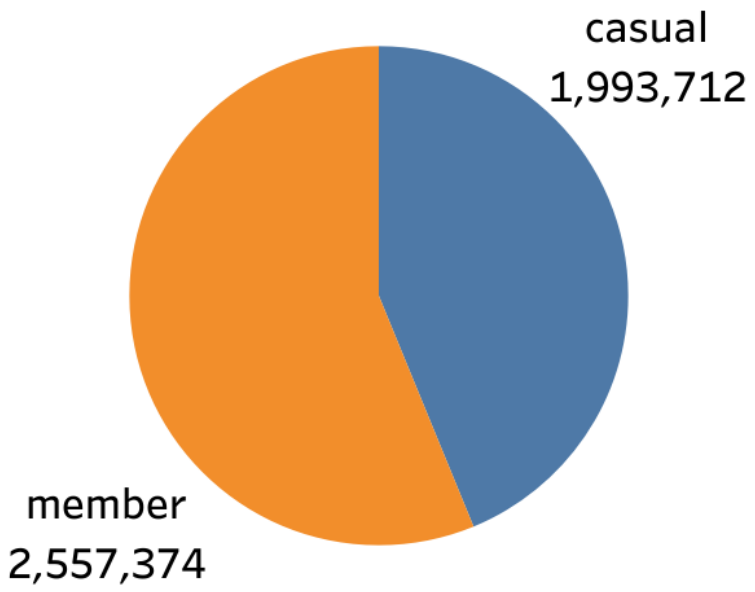
I then, filtered and removed any trips that were below 60 seconds in length as potentially false starts or users trying to re-dock a bike to ensure it was secure. For this reason, they these rides will be of no used to this analysis.

```
DELETE
FROM [cycle_DB].[dbo].[Final_table]
WHERE ride_length < '00:01:00'
```

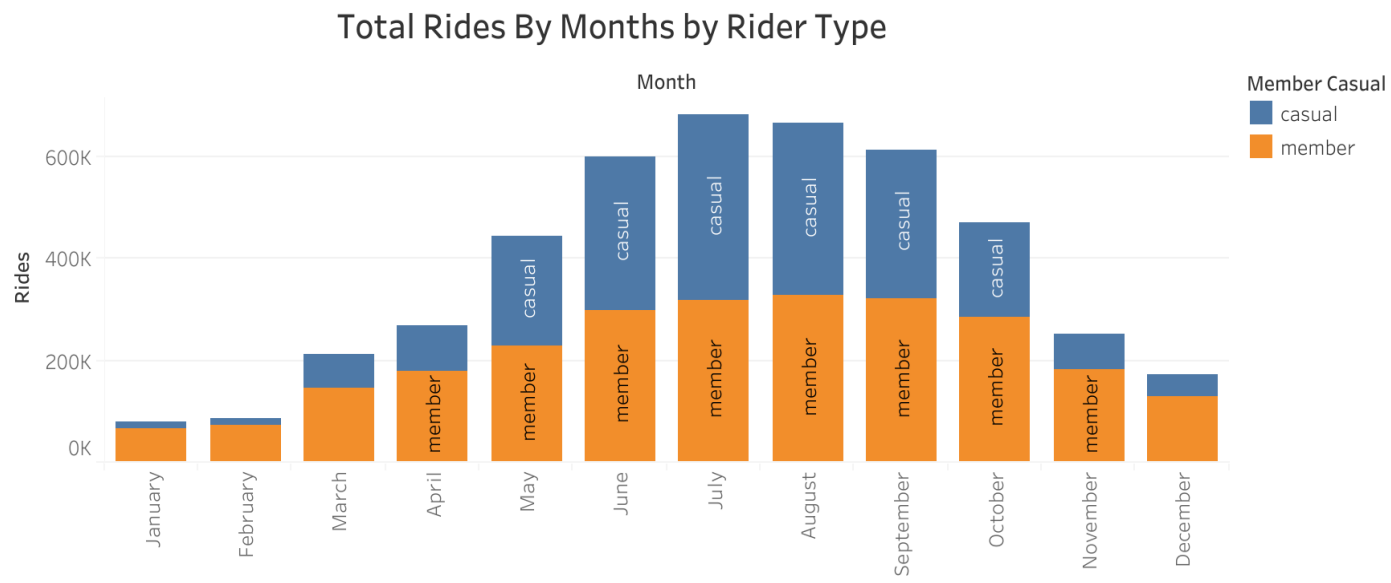
♦ ANALYZE

Now that the data has been cleaned and we have our final table. We can start exploring, analyzing and gaining insights from the data.

First, I started by verifying the total number of rides we are dealing with. As each ride id is unique to each ride, we will consider each row to represent 1 ride. We started with 5,757,564 and after cleaning and processing the data, we are left with 4,551,086 rides.

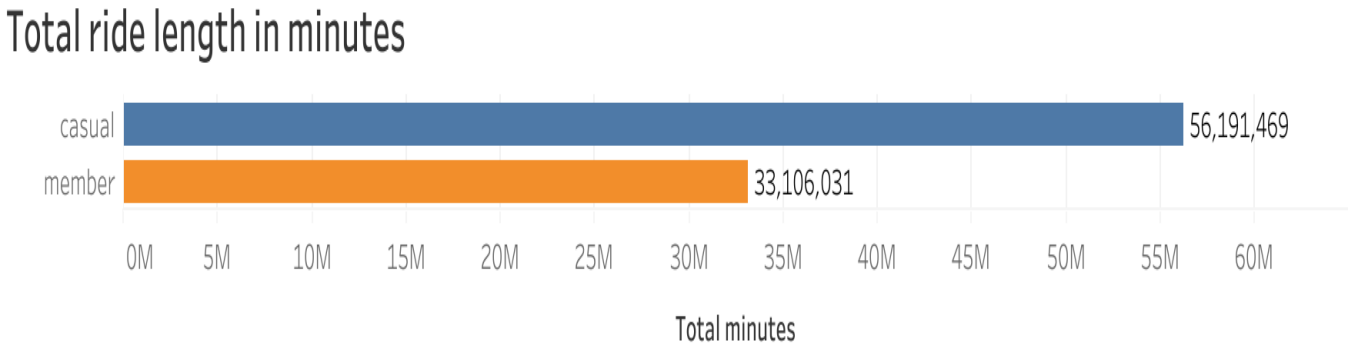


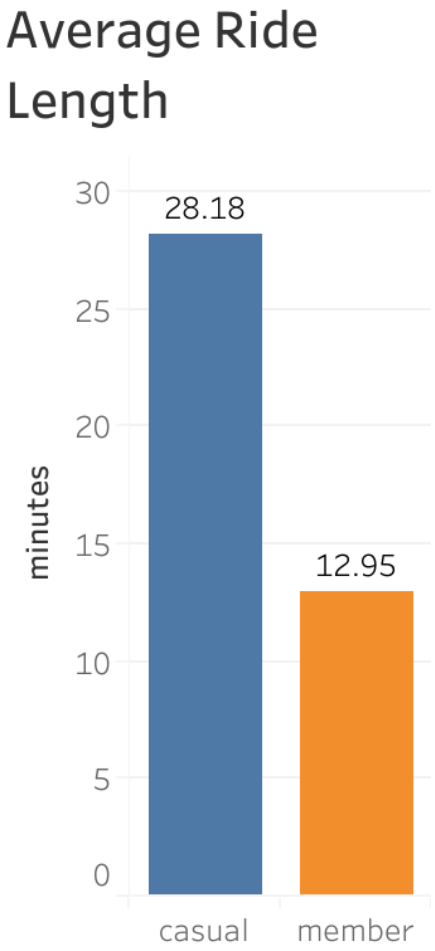
I than graphed the number of rides by months in order to verify my first hypothesis.



Looking at this bar graph, we can confirm that this is a seasonal service with a significantly higher demand between the months of May and October. This graph also shows that, while during the peak months, the number usage is similar between casual riders and annual member, however, annual members keep using the service longer than casual riders.

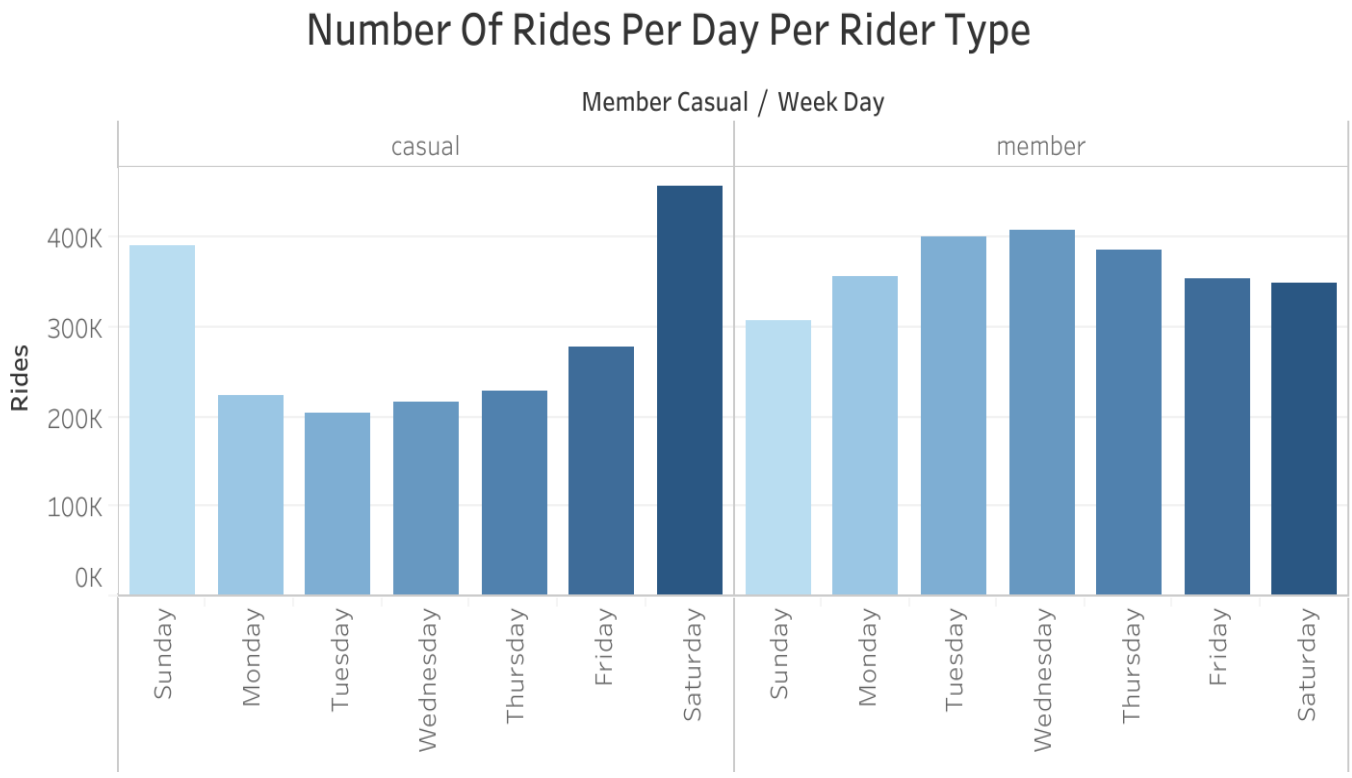
Next, I proceeded to aggregate and verify the total time spent riding for each rider type. I also averaged the amount of time of each type of riders.





The graphs above shows that casual riders tend to ride longer on average than Annual members.

Next, I wanted to look at how the two types of riders compared throughout the week.



This graph shows that annual members have more rides from Monday to Friday, while casual riders tend to use the bicycles significantly more during weekends.

GEO-MAPPING

Annual members and casual riders also differ with the locations where they tend to start and end their journey.

The dataset we were using for our analysis contained a few errors in the latitude and longitude of each docking station. Some stations with the same name had slightly different

coordinates. To circumvent this problem and obtain the information necessary, a new query had to be created from SQL.

Using the WITH statement in SQL, I averaged the latitude and longitudes for all the docking stations with the same name and grouped them by name. As a result, each unique station was paired with a unique latitude and longitude.
I did this with both the START and END stations and proceeded to put the results together using a JOIN statement.

I first cleaned the departing stations for casual riders

```
WITH casual_dep AS (

  SELECT distinct(clean_start_station), ROUND(AVG(cast(start_lat as float)),4) as dep_lat,
  ROUND(AVG(cast(start_lng as float)),4) as dep_lng, count(ride_id) as casual
  FROM [cycle_DB].[dbo].[Year_Table]
  WHERE member_casual = 'casual'
  GROUP BY clean_start_station
  ),
```

A similar query was used for the departing stations for annual members.

```
member_dep AS(

  SELECT distinct(clean_start_station), ROUND(AVG(cast(start_lat as float)),4) as dep_lat,
 ROUND(AVG(cast(start_lng as float)),4) as dep_lng, count(ride_id) as member
 FROM [cycle_DB].[dbo].[Year_Table]
 WHERE member_casual = 'member'
 GROUP BY clean_start_station
 ),
```

And then I joined both member and casual departing stations in order to export to Tableau for visualization.

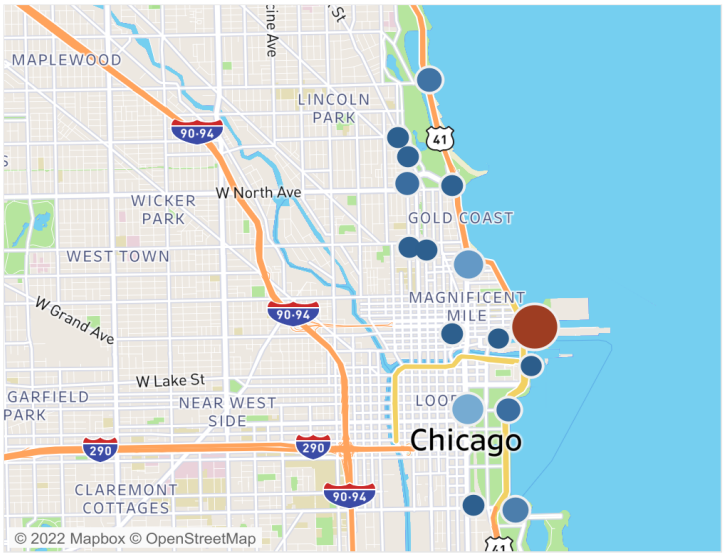
```
Depart_station_viz AS (

  Select cd.clean_start_station, cd.dep_lat, cd.dep_lng,cd.casual, md.member
  From [cycle_DB].[dbo].[casual_dep] cd
  JOIN [cycle_DB].[dbo].[member_dep] md
  ON md.clean_start_station = cd.clean_start_station
  ORDER BY md.member DESC )
```

The result of the above query allowed me to map the departing for both annual members and casual riders respectively. A similar procedure was used to map the arriving stations for both types of riders.

The results for casual riders can be seen here:

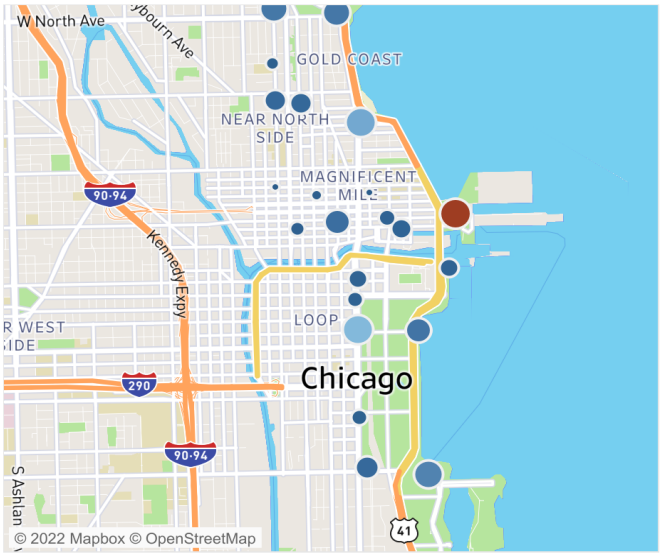
Departing Stations



Top 5 Start Station casual

Streete..	Millenn..	Michiga..	Shedd ..	Theater..
65,238	30,815	27,566	21,922	19,661

Arriving Stations



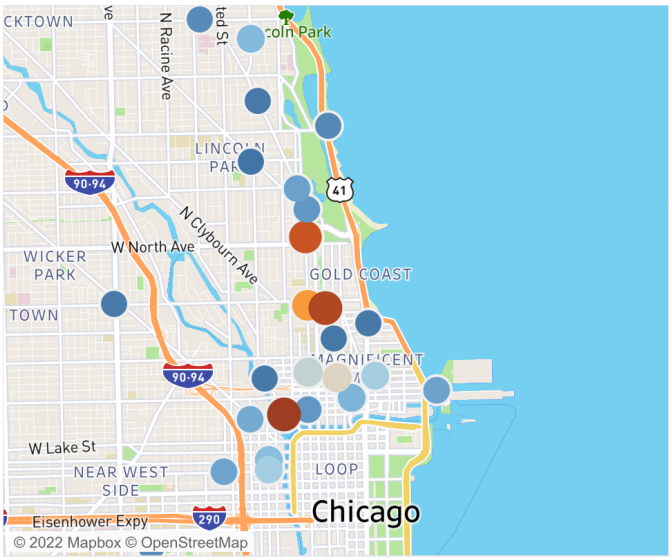
Top 5 End Station casual

Streete..	Millenn..	Michiga..	Theater..	Shedd ..
68,471	32,309	29,141	21,136	20,724

The map above allows us to see where the concentration of casual riders is more important. It also gives us some insight on the top 5 most popular stations are.

The results for annual members can be seen here:

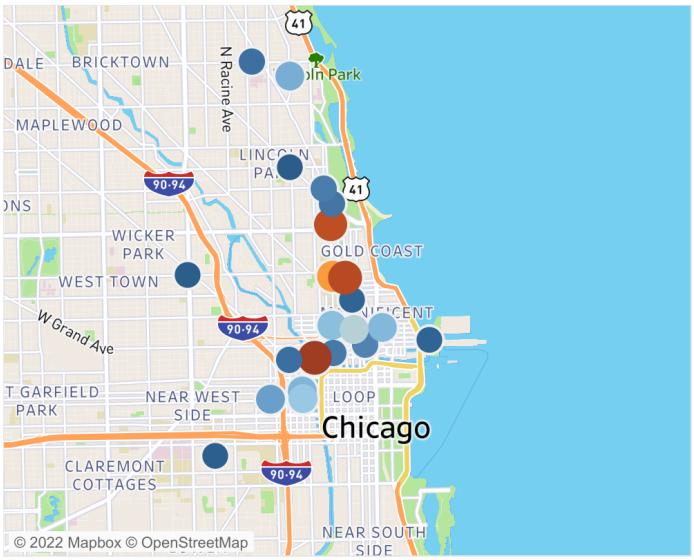
Departing Stations



Top 5 Start Station

Kingsbu..	Clark St..	Wells S..	Wells S..	Dearbo..
24,280	23,543	22,657	19,939	18,176

Arriving Stations



Top 5 End Station

Kingsbu..	Clark St..	Wells S..	Wells S..	Dearbo..
24,232	23,396	23,191	20,480	18,666

This map allows us to see how the annual members differ from the casual rider in their station preferences. We can see that they are more concentrated toward the city center, unlike the casual riders who tend to go more towards the coastline.

♦ SHARE

Details of the SQL queries used for this case study can be found on <GITHUB>, and the Interactive visualizations can be found on <Tableau Public>

♦ **ACT** (recommandations)

Insights

- The Cyclistic Bike Sharing offers a seasonal service where the highest demands is during the summer and the lowest during winter.

- For the period we are working with, annual members have more rides in total than casual riders, while casual riders tend to ride longer on average than annual members.

- Ridership for annual member is higher from Monday to Friday, while the ridership for Casual rider avec highest during weekends.

- On average, annual members start and end their journey at Kingsbury St. & Kinzie St. station and stay mostly concentrated on the city center.

- On the other hand, Casual riders mostly start and end their journey at Streeter Dr & Grand Av. Station and stay mostly concentrated on the coastline.

Recommendations

- A marketing campaign should be focus on the most popular casual rider stations.

- As it is a seasonal service, the campaign should start in March and finish at the beginning of November. To reach out the most riders, The marketing campaign should be more aggressive during the weekend as it is the time of the week when most casual riders are using the service.

- Cyclist bike share should implement a new member onboarding discount and focus on yearly weekend-only membership to encourage casual riders to subscribe.

Thank you for reading my case study.