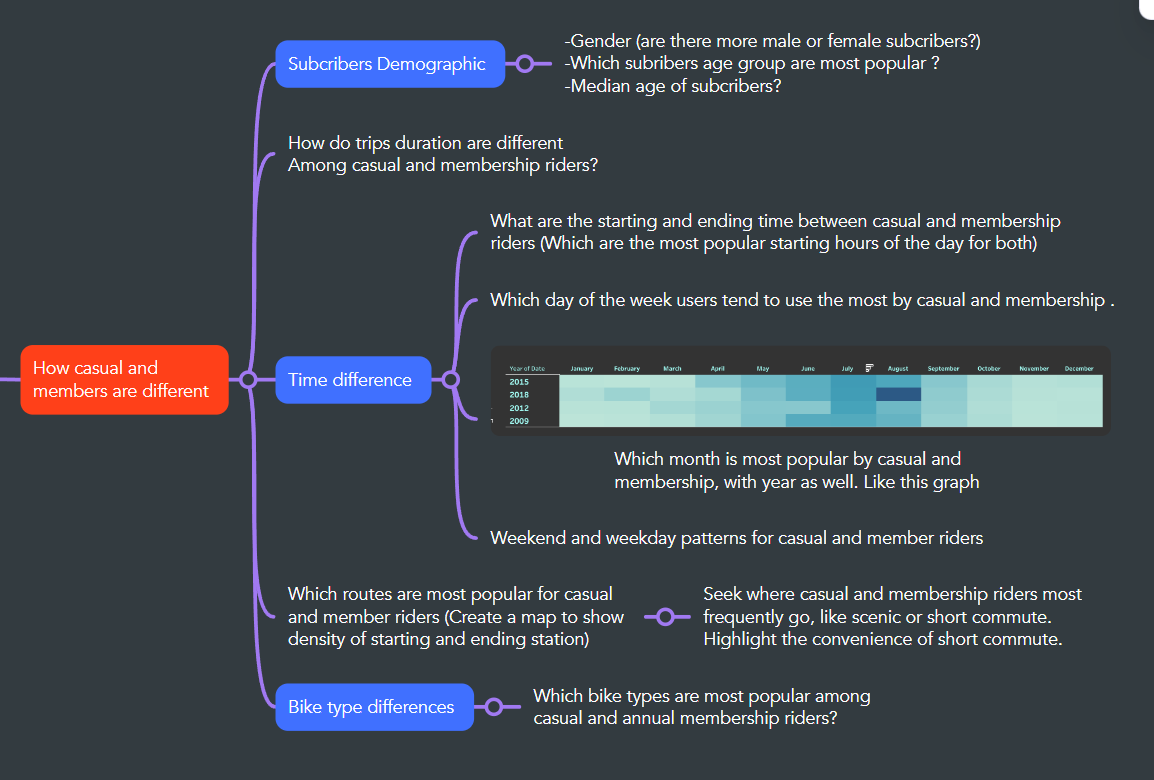
**Data Analysis**

## Objectives

My goal is to identify how do annual members and casual riders use Cyclistic bikes differently? These are my initial insights that may be helpful. All the queries results are stored in Excel\_Tableau.Xlsx for Tableau Dashboard.



## Subscribers Demographic

First, I will Identify what is the dominant gender for Subscribers. Note that for all the data results after querying, I put it to Excel sheet ‘Excel\_Tableau.csv’ to visualize in Tableau.

/\*DATA ANALYSIS \*/

--What is the dominant gender for the subscribers. Data are in Trips\_2013\_2019

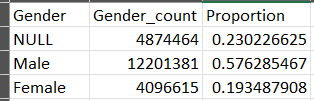
SELECT Gender,

COUNT (Gender) AS Gender\_count

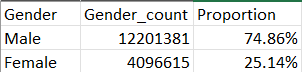
FROM Trips\_2013\_2019

GROUP BY Gender

We would have



But since our focus is to find whether male or female proportion, we can drop the nulls values and get the male and female proportion.



Next, I want to explore the age group and median age of the subscribers.

/\* What is the most popular age group? I create an extra age group column

< 18

18-35

35-45

45-60

60 +

\*/

--Create a temp table holding Trips\_2013\_2019

SELECT Birth\_Year INTO #Age\_group

FROM Trips\_2013\_2019

WHERE Birth\_Year is not null

--Add the age group column

ALTER TABLE #Age\_group

ADD Age\_Group nvarchar(50),

Age smallint

--Get the Age from the BirthYear Column

UPDATE #Age\_group

SET Age = 2024 - Birth\_Year

--Create an age group columns based on the Age column

UPDATE #Age\_group

SET Age\_Group = CASE

WHEN Age < 18 THEN '<18'

WHEN Age BETWEEN 18 AND 34 THEN '18-34'

WHEN Age BETWEEN 35 AND 44 THEN '35-44'

WHEN Age BETWEEN 45 AND 59 THEN '45-59'

ELSE '60+'

END;

Since there are some users errors when typing the answer, I consider filter out any users > 100 years old.

--Identify the Age Group for the subscribers

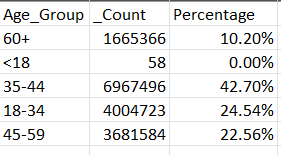
SELECT Age\_Group, COUNT(Age\_Group) AS \_Count

FROM #Age\_group

WHERE AGE < 100

GROUP BY Age\_Group

The results



Now I compute the average age

--Compute the median Age for subcribers

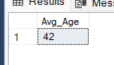
SELECT AVG(Age) as Avg\_Age

FROM #Age\_group

WHERE Age <= 100

--Drop temp table after use

DROP TABLE #Age\_group



Our Average subscriber age is 42.

## Trips duration

/\* Find the Trips duration for each customer types \*/

--Run the #Duration\_2020\_2023 temp table

SELECT Start\_Time, Stop\_Time,

CASE

WHEN User\_Type = 'casual' THEN 'Customer'

ELSE 'Subscriber'

END AS User\_Type2

INTO #Duration\_2020\_2023

FROM Trips\_2020\_2023;

--Compute the average trips duration with Trips\_2013\_2019 union #Duration\_2020\_2023

SELECT

User\_Type2,

AVG(CAST(Trip\_Duration AS FLOAT)) AS Avg\_Trip\_Duration

FROM

(

SELECT DATEDIFF(SECOND, Start\_Time, Stop\_Time) AS Trip\_Duration, User\_Type2

FROM #Duration\_2020\_2023

UNION ALL

SELECT Trip\_Duration, User\_Type

FROM [dbo].[Trips\_2013\_2019]

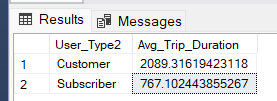
) AS Combined

GROUP BY User\_Type2

--Drop temp table after use

DROP TABLE #Duration\_2020\_2023

The results



## Most popular Bike types

In this part, I aim to identify which type of bike is most popular among casual and member users.

--What are the popular Bike Types by User\_Type?

SELECT

User\_Type,

Ride\_Type,

COUNT(Ride\_Type) as Total

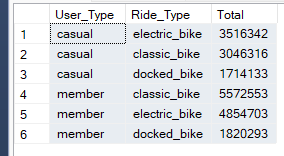
FROM [dbo].[Trips\_2020\_2023]

WHERE User\_Type is not null

GROUP BY User\_Type, Ride\_Type

ORDER BY 1, 3 DESC

The results



## Time period difference

Firstly, I aim to identify the Starting hours for the user’s type.

-----------------------------------------------------------------------------------------------

--What are the popular starting time and stop time for user type

--Rerun the temp table

SELECT Start\_Time,

CASE

WHEN User\_Type = 'casual' THEN 'Customer'

ELSE 'Subscriber'

END AS User\_Type2

INTO #TempTrips\_2020\_2023

FROM Trips\_2020\_2023;

--I layered the subqueries and aggregate the hour count by users type

SELECT

User\_Type,

Starting\_Hr,

COUNT (Starting\_Hr) as Count\_hour

FROM (

SELECT

DATEPART(HOUR,Start\_Time) AS Starting\_Hr,

User\_Type

FROM

(SELECT Start\_Time, User\_Type

FROM [dbo].[Trips\_2013\_2019]

UNION ALL

SELECT\*

FROM #TempTrips\_2020\_2023)

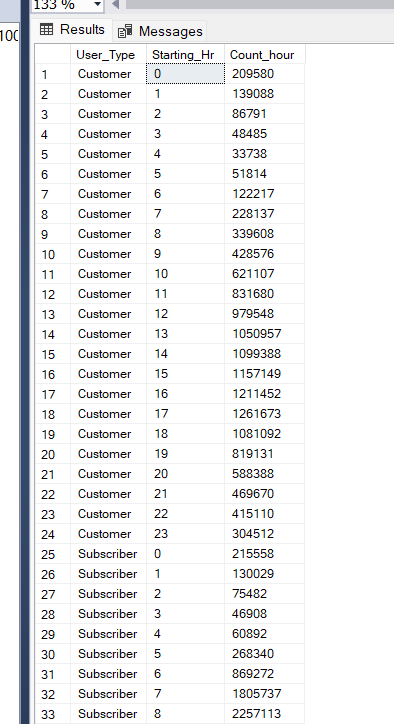
AS Combined2

) AS Combined3

GROUP BY User\_Type, Starting\_Hr

ORDER BY 1, 2,3 DESC

The results: I would have a list of hour and their counts with user types that I can later use for visualization.



Next, I want to explore the patterns for day of the week and whether there are differences between weekend and weekday.

--Explore weekday and weekend patterns

--Which is the most popular day for the user type

SELECT

User\_Type,

Date\_name,

COUNT (Date\_name) as Count\_day

FROM

(SELECT

User\_Type,

DATENAME(dw, Start\_Time) as Date\_name

FROM

(SELECT Start\_Time, User\_Type

FROM [dbo].[Trips\_2013\_2019]

UNION ALL

SELECT\*

FROM #TempTrips\_2020\_2023)

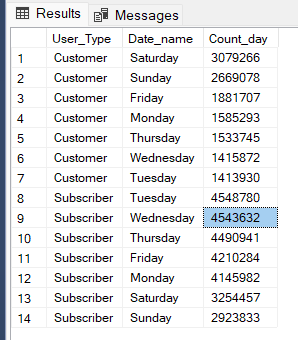
AS Combined2

) AS Combined3

GROUP BY User\_Type, Date\_name

ORDER BY 1 ,3 DESC

The results



Finally, I want to identify busiest month of the year.

--Identify the busiest month of each year

SELECT

User\_Type,

Year\_,

Month\_,

COUNT(Month\_) as Month\_Count

FROM

(SELECT

User\_Type,

DATEPART(yy, Start\_Time) AS Year\_,

DATEPART(m, Start\_Time) AS Month\_

FROM

(SELECT Start\_Time, User\_Type

FROM [dbo].[Trips\_2013\_2019]

UNION ALL

SELECT\*

FROM #TempTrips\_2020\_2023)

AS Combined2

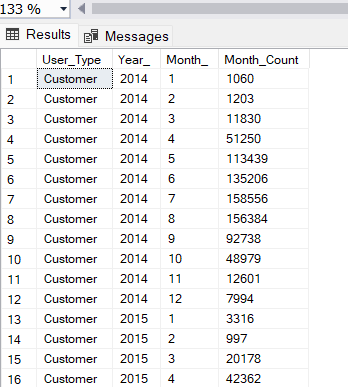
) AS Combined3

WHERE Year\_ NOT IN (2013, 2024) -- As 2013 and 2024 don’t have adequate months

GROUP BY User\_Type, Year\_, Month\_

ORDER BY 1, 2, 3 ASC

The results: I would have a list of year and month count for each user types



## Most popular stations

In the final part, I will explore which starting and ending stations are busiest for casual and annual member customers. Due to incomplete data, I will use from 2020-2023 for the most updated Stations.

-- Find the most popular starting point for Casual members

SELECT

From\_Station\_Name,

Start\_cnt,

latitude,

longitude

FROM

(

SELECT

From\_Station\_Name,

COUNT(From\_Station\_Name) as Start\_cnt

FROM [dbo].[Trips\_2020\_2023]

WHERE User\_Type = 'casual'

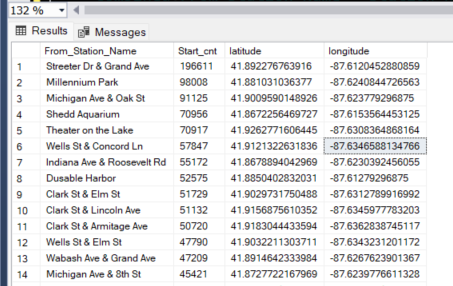
GROUP BY From\_Station\_Name

) AS Combined INNER JOIN [dbo].[Station\_name] ON

Combined.From\_Station\_Name = Station\_name.name

ORDER BY Start\_cnt DESC

The results: we would have a table with starting station name, count of rides, latitude and longitude for casual members. This can be used for visualization later.



Now repeat to find the ending station for causal members and repeat for Subscriber members

-- Find the most popular ending point for Casual members

SELECT

To\_Station\_Name,

To\_cnt,

latitude,

longitude

FROM

(

SELECT

To\_Station\_Name,

COUNT(To\_Station\_Name) as To\_cnt

FROM [dbo].[Trips\_2020\_2023]

WHERE User\_Type = 'casual'

GROUP BY To\_Station\_Name

) AS Combined INNER JOIN [dbo].[Station\_name] ON

Combined.To\_Station\_Name = Station\_name.name

ORDER BY To\_cnt DESC

-- Find the most popular starting point for Subscriber members

SELECT

From\_Station\_Name, Start\_cnt, latitude, longitude

FROM

(

SELECT

From\_Station\_Name,

COUNT(From\_Station\_Name) as Start\_cnt

FROM [dbo].[Trips\_2020\_2023]

WHERE User\_Type = 'member'

GROUP BY From\_Station\_Name

) AS Combined INNER JOIN [dbo].[Station\_name] ON

Combined.From\_Station\_Name = Station\_name.name

ORDER BY Start\_cnt DESC

-- Find the most popular ending point for Subscriber members

SELECT

To\_Station\_Name, To\_cnt, latitude, longitude

FROM

(

SELECT

To\_Station\_Name,

COUNT(To\_Station\_Name) as To\_cnt

FROM [dbo].[Trips\_2020\_2023]

WHERE User\_Type = 'member'

GROUP BY To\_Station\_Name

) AS Combined INNER JOIN [dbo].[Station\_name] ON

Combined.To\_Station\_Name = Station\_name.name

ORDER BY To\_cnt DESC