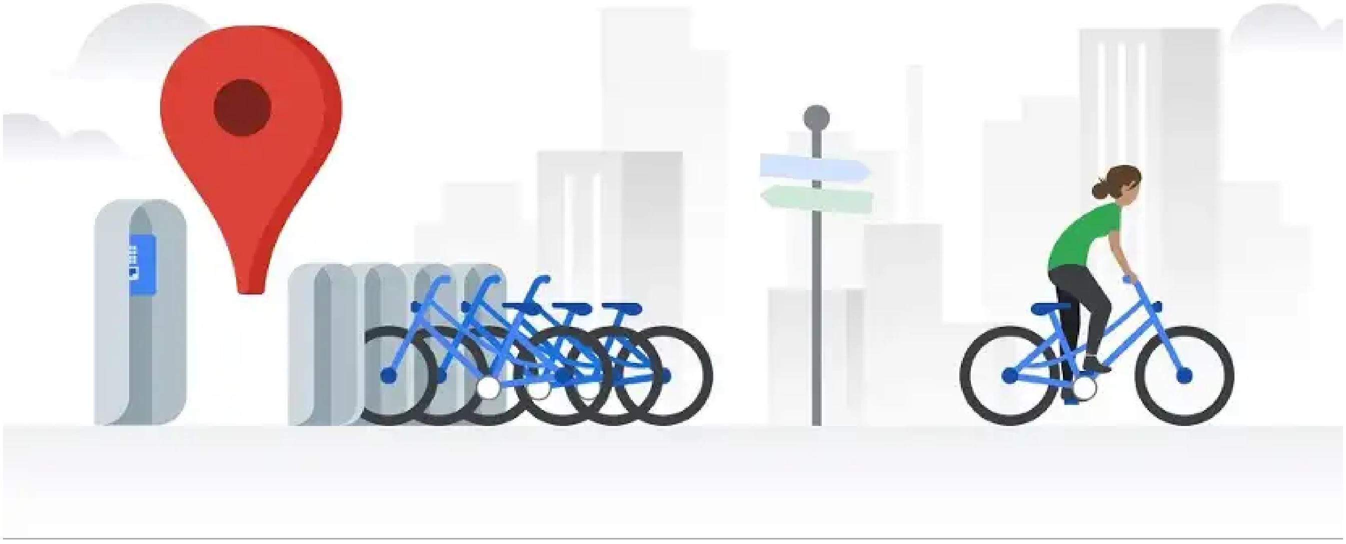
Ahmed Anees Zaveri

Dec 19 · 10 min read · Listen

Save

Cyclistic — Google Data Analytics Capstone Project



Introduction

Cyclistic is a bike-share program launched in Chicago in 2016 that features 5,824 bicycles and 692 docking stations. The bikes can be unlocked from one station and returned to any other station in the system anytime. 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.

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer s h that helped make these things possible was the flexibility of its 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. Customers who purchase annual memberships are Cyclistic members.

Scenario

I am a junior data analyst working in the marketing analyst team at Cyclistic. The director of marketing, Lily Moreno, believes the company's future success depends on maximizing the number of annual memberships. Therefore, my team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, the team will design a new marketing strategy to convert casual riders into annual members. In order to answer these key business questions, the six steps of the data analyst process will be followed: **Ask**, **Prepare**, **Process**, **Analyze**, **Share** and **Act**.

Step 1: Ask

Ask phase requires the analyst to ask the key questions which would set the course of the whole data analysis process. The three questions which the director of Marketing, Lily Moreno, needs answered are as follows:

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

Business Idea:

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs. For this reason she has tasked me with the first of the three questions listed above. Therefore, my analysis would revolve around finding how the annual members and casual riders use the Cyclistic bikes differently.

Key Stakeholders:

- **Lily Moreno:** The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program.
- **Cyclistic marketing analytics team:** A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy.
- **Cyclistic executive team:** The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

Step 2: Prepare

Data Source:
I have been asked to make use of the Cyclistic's historical trip data to analyze and identify trends. For the purposes of this case study, the datasets are appropriate and will be sufficient to answer the business questions. The data has been made available by Motivate International Inc. and can be found *here*.

Quality of Data:
The data is provided by Motivate International Inc. under the license found *here*. It is public data that we can use to explore how different customer types are using Cyclistic bikes. The data is **R**eliable and **C**ited as it comes from a licensed source. It is **O**riginal as it is acquired from a primary source. It is **C**omprehensive because it covers all the aspects of the data which would be required to perform the analysis. Lastly, the data is also **C**urrent as data from the last 12 months have been used for the analysis. Concisely put, it adheres to the ROCCC standards and hence can be classified as good data.

Datasets used:
Trip data from October 2021 to September 2022 has been used for the analysis. The data is in the form of a separate csv file for each month.

Step 3: Process

Choice of Tools: Microsoft Excel, Microsoft SQL Server Management Studio and Tableau Public

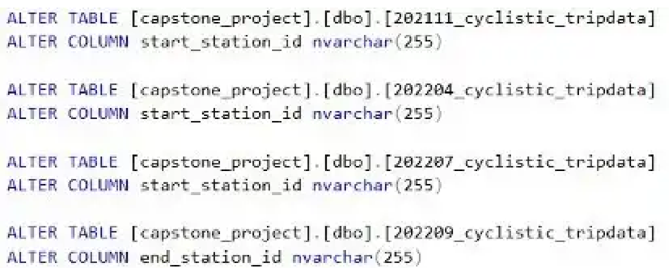
- **Microsoft Excel:** A spreadsheet serves as the best tool for basic Data Exploration and Cleaning. It also allows quick sorting and filtering options which helps in familiarizing with the dataset and make some basic sense of the data.
- **Microsoft SQL Server:** Once the 12 datasets are combined, the total amount of data will surpass the maximum limit allowable for an Excel Worksheet. Hence, SQL will be a much more suitable fit to perform Data Transformation on the huge dataset and also for extracting desired data in the form of concise tables.
- **Tableau Public:** The concise tables are copied to a spreadsheet from where they are imported into Tableau Public. This is where the tabulated data is visualized in the form of meaningful illustrations, charts and maps.

Initial Data Cleaning using Microsoft Excel
A new column called “ride_length” is created in each of the 12 worksheets which calculates the length of each ride by subtracting the column “started_at” from the column “ended_at” and then converting the format of this new column to “Time” in order to get the values in the “HH:MM:SS” format.

Next, a new column called “day_of_week” was created. Calculation of this new column was done using the “WEEKDAY” built-in Excel Function.

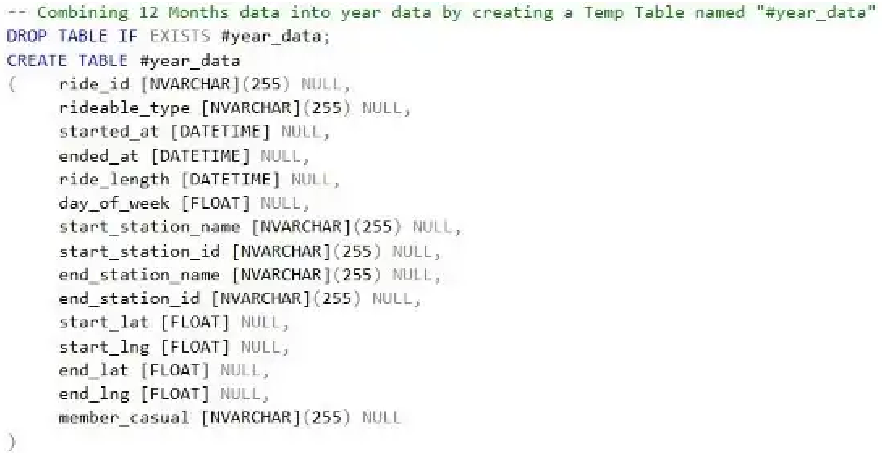
Step 4: Analyze

Data Cleaning and Transformation using Microsoft SQL Server
The first transformation was done to change the data types of a few columns which were incorrectly decided by Microsoft SQL Server while it was being imported.



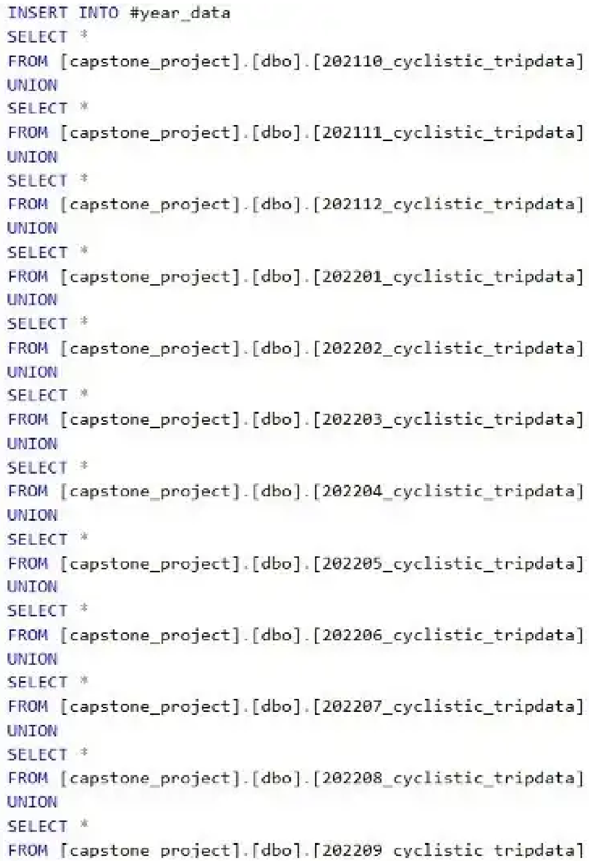
```
ALTER TABLE [capstone_project].[dbo].[202111_cyclistic_tripdata]
ALTER COLUMN start_station_id nvarchar(255)

ALTER TABLE [capstone_project].[dbo].[202204_cyclistic_tripdata]
ALTER COLUMN start_station_id nvarchar(255)

ALTER TABLE [capstone_project].[dbo].[202207_cyclistic_tripdata]
ALTER COLUMN start_station_id nvarchar(255)

ALTER TABLE [capstone_project].[dbo].[202209_cyclistic_tripdata]
ALTER COLUMN end_station_id nvarchar(255)
```

The next step was to combine the 12 months data into a single temp table.



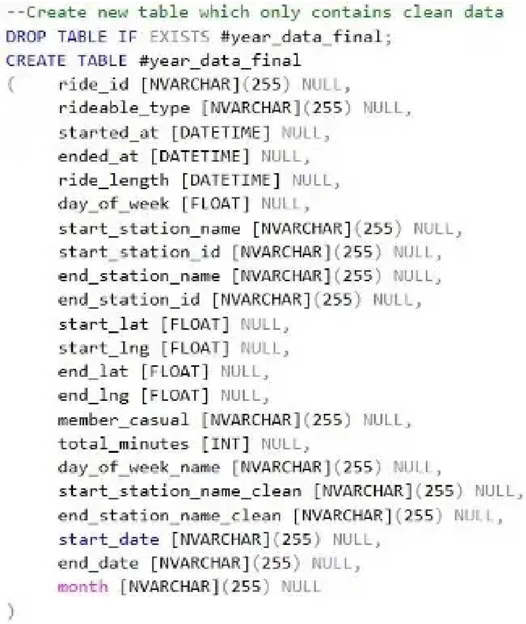
```
-- Combining 12 Months data into year data by creating a Temp Table named "#year_data"
DROP TABLE IF EXISTS #year_data;
CREATE TABLE #year_data
(     ride_id [NVARCHAR](255) NULL,
      rideable_type [NVARCHAR](255) NULL,
      started_at [DATETIME] NULL,
      ended_at [DATETIME] NULL,
      ride_length [DATETIME] NULL,
      day_of_week [FLOAT] NULL,
      start_station_name [NVARCHAR](255) NULL,
      start_station_id [NVARCHAR](255) NULL,
      end_station_name [NVARCHAR](255) NULL,
      end_station_id [NVARCHAR](255) NULL,
      start_lat [FLOAT] NULL,
      start_lng [FLOAT] NULL,
      end_lat [FLOAT] NULL,
      end_lng [FLOAT] NULL,
      member_casual [NVARCHAR](255) NULL
)
```

Once, the temp table was created, it needs to be populated with the data from 12 months



```
INSERT INTO #year_data
SELECT *
FROM [capstone_project].[dbo].[202110_cyclistic_tripdata]
UNION
SELECT *
FROM [capstone_project].[dbo].[202111_cyclistic_tripdata]
UNION
SELECT *
FROM [capstone_project].[dbo].[202112_cyclistic_tripdata]
UNION
SELECT *
FROM [capstone_project].[dbo].[202201_cyclistic_tripdata]
UNION
SELECT *
FROM [capstone_project].[dbo].[202202_cyclistic_tripdata]
UNION
SELECT *
FROM [capstone_project].[dbo].[202203_cyclistic_tripdata]
UNION
SELECT *
FROM [capstone_project].[dbo].[202204_cyclistic_tripdata]
UNION
SELECT *
FROM [capstone_project].[dbo].[202205_cyclistic_tripdata]
UNION
SELECT *
FROM [capstone_project].[dbo].[202206_cyclistic_tripdata]
UNION
SELECT *
FROM [capstone_project].[dbo].[202207_cyclistic_tripdata]
UNION
SELECT *
FROM [capstone_project].[dbo].[202208_cyclistic_tripdata]
UNION
SELECT *
FROM [capstone_project].[dbo].[202209_cyclistic_tripdata]
```

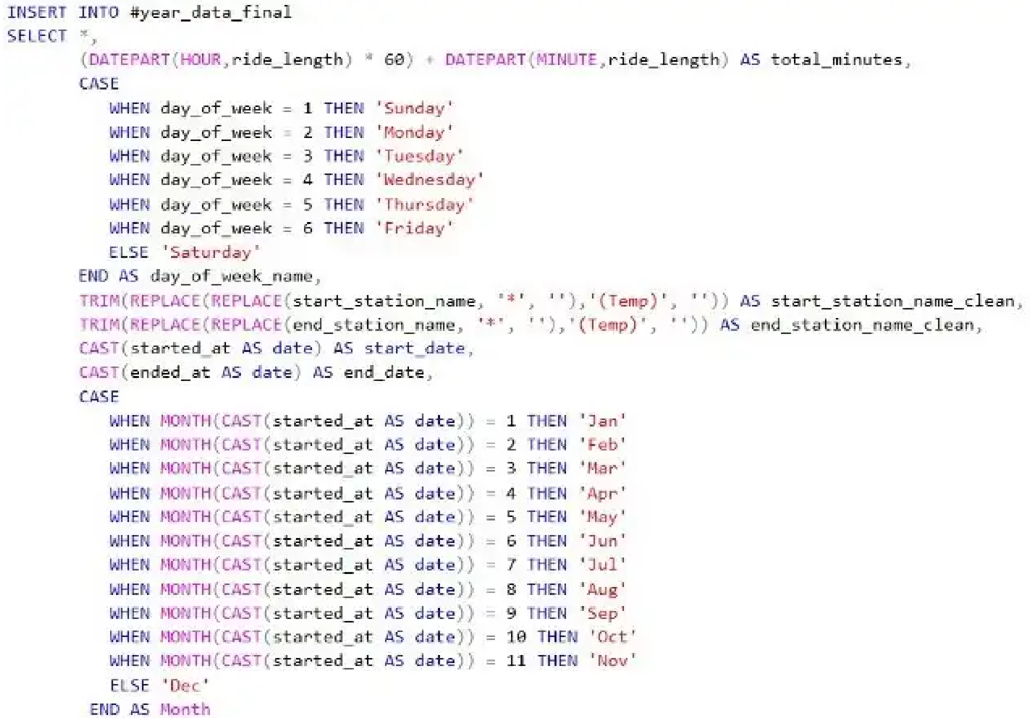
The data we have at hand is still pretty much in raw form and a lot of transformations need to be performed before actual analysis can begin. For this purpose, another temp table is created which will house the transfomed data.



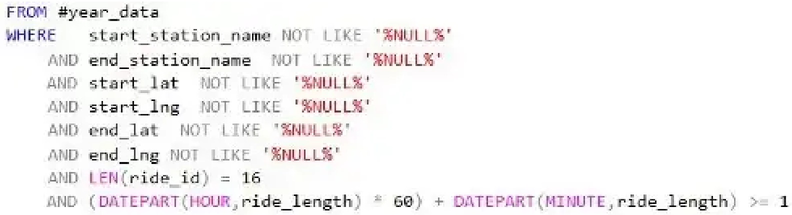
```
--Create new table which only contains clean data
DROP TABLE IF EXISTS #year_data_final;
CREATE TABLE #year_data_final
(    ride_id [NVARCHAR](255) NULL,
     rideable_type [NVARCHAR](255) NULL,
     started_at [DATETIME] NULL,
     ended_at [DATETIME] NULL,
     ride_length [DATETIME] NULL,
     day_of_week [FLOAT] NULL,
     start_station_name [NVARCHAR](255) NULL,
     start_station_id [NVARCHAR](255) NULL,
     end_station_name [NVARCHAR](255) NULL,
     end_station_id [NVARCHAR](255) NULL,
     start_lat [FLOAT] NULL,
     start_lng [FLOAT] NULL,
     end_lat [FLOAT] NULL,
     end_lng [FLOAT] NULL,
     member_casual [NVARCHAR](255) NULL,
     total_minutes [INT] NULL,
     day_of_week_name [NVARCHAR](255) NULL,
     start_station_name_clean [NVARCHAR](255) NULL,
     end_station_name_clean [NVARCHAR](255) NULL,
     start_date [NVARCHAR](255) NULL,
     end_date [NVARCHAR](255) NULL,
     month [NVARCHAR](255) NULL
)
```

Next, it is time to perform the following transformations and populate this new temp table with the wrangled data simultaneously:

- Clearing empty cells from the dataset using NOT and LIKE operators along with % sign to identify NULLS
- Remove ride_ids where length of ride_id > 16
- Creating new column named total_minutes by using DATEDIFF command
- Using CASE statement to change the day_of_week from numerical to string data
- Remove rows where total_minutes < 1
- start_station_name and end_station_name cleaned by replacing '*' and 'temp' with blank spaces



```
INSERT INTO #year_data_final
SELECT *,
       (DATEPART(HOUR,ride_length) * 60) + DATEPART(MINUTE,ride_length) AS total_minutes,
       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 AS day_of_week_name,
       TRIM(REPLACE(REPLACE(start_station_name, '*', ''),'(Temp)', '')) AS start_station_name_clean,
       TRIM(REPLACE(REPLACE(end_station_name, '*', ''),'(Temp)', '')) AS end_station_name_clean,
       CAST(started_at AS date) AS start_date,
       CAST(ended_at AS date) AS end_date,
       CASE
          WHEN MONTH(CAST(started_at AS date)) = 1 THEN 'Jan'
          WHEN MONTH(CAST(started_at AS date)) = 2 THEN 'Feb'
          WHEN MONTH(CAST(started_at AS date)) = 3 THEN 'Mar'
          WHEN MONTH(CAST(started_at AS date)) = 4 THEN 'Apr'
          WHEN MONTH(CAST(started_at AS date)) = 5 THEN 'May'
          WHEN MONTH(CAST(started_at AS date)) = 6 THEN 'Jun'
          WHEN MONTH(CAST(started_at AS date)) = 7 THEN 'Jul'
          WHEN MONTH(CAST(started_at AS date)) = 8 THEN 'Aug'
          WHEN MONTH(CAST(started_at AS date)) = 9 THEN 'Sep'
          WHEN MONTH(CAST(started_at AS date)) = 10 THEN 'Oct'
          WHEN MONTH(CAST(started_at AS date)) = 11 THEN 'Nov'
          ELSE 'Dec'
        END AS Month
```

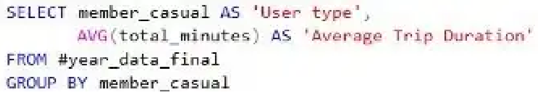


```
FROM #year_data
WHERE   start_station_name NOT LIKE '%NULL%'
    AND end_station_name  NOT LIKE '%NULL%'
    AND start_lat  NOT LIKE '%NULL%'
    AND start_lng  NOT LIKE '%NULL%'
    AND end_lat  NOT LIKE '%NULL%'
    AND end_lng NOT LIKE '%NULL%'
    AND LEN(ride_id) = 16
    AND (DATEPART(HOUR,ride_length) * 60) + DATEPART(MINUTE,ride_length) >= 1
```

Now that the new temp table is populated with the clean data, we can get rid of the previous temp table by using the **DROP TABLE** command in order to free up some space in the memory.

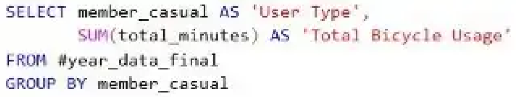
Data Exploration, Extraction and Analysis

1. Which user type on average uses the bicycle more?



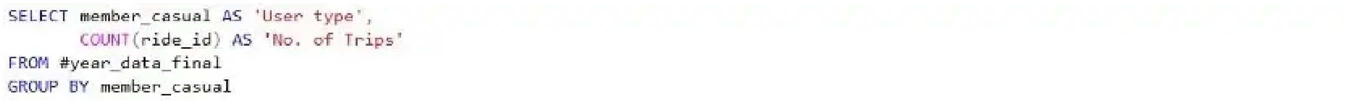
```
SELECT member_casual AS 'User type',
       AVG(total_minutes) AS 'Average Trip Duration'
FROM #year_data_final
GROUP BY member_casual
```

2. Which user type uses the bicycle more in total (by minutes)?



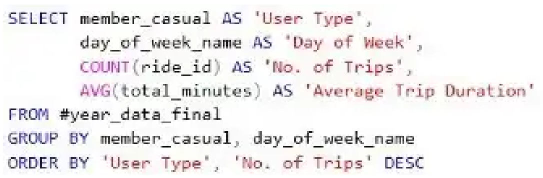
```
SELECT member_casual AS 'User Type',
       SUM(total_minutes) AS 'Total Bicycle Usage'
FROM #year_data_final
GROUP BY member_casual
```

3. Which user type makes the most number of trips?



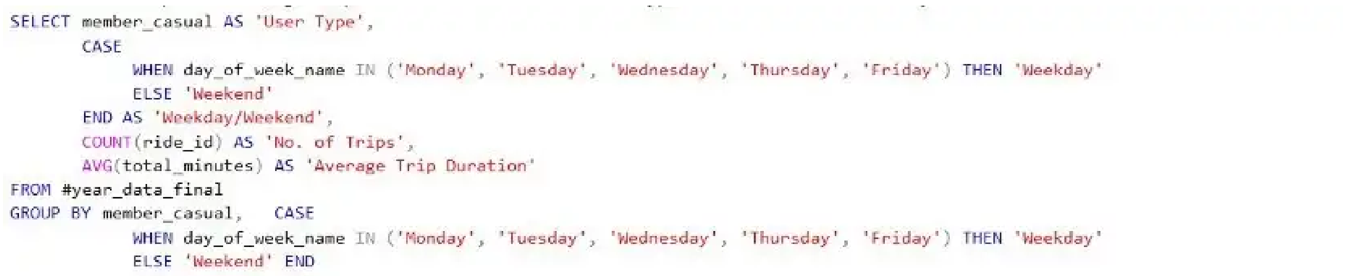
```
SELECT member_casual AS 'User type',
       COUNT(ride_id) AS 'No. of Trips'
FROM #year_data_final
GROUP BY member_casual
```

4. What is the number of trips and the average trip duration for each data type on each day of the week?



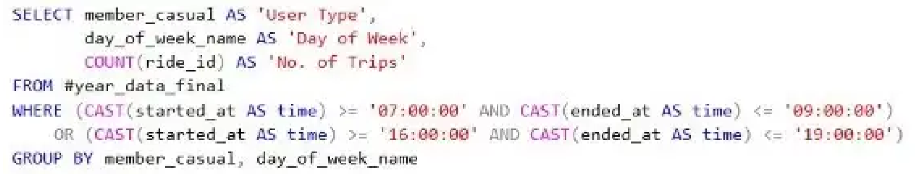
```
SELECT member_casual AS 'User Type',
       day_of_week_name AS 'Day of Week',
       COUNT(ride_id) AS 'No. of Trips',
       AVG(total_minutes) AS 'Average Trip Duration'
FROM #year_data_final
GROUP BY member_casual, day_of_week_name
ORDER BY 'User Type', 'No. of Trips' DESC
```

5. What is the number of trips and the average trip duration for each data type on weekdays in comparison to weekends?



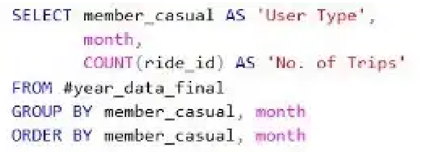
```
SELECT member_casual AS 'User Type',
       CASE
            WHEN day_of_week_name IN ('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday') THEN 'Weekday'
            ELSE 'Weekend'
       END AS 'Weekday/Weekend',
       COUNT(ride_id) AS 'No. of Trips',
       AVG(total_minutes) AS 'Average Trip Duration'
FROM #year_data_final
GROUP BY member_casual,   CASE
            WHEN day_of_week_name IN ('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday') THEN 'Weekday'
            ELSE 'Weekend' END
```

6. How does the usage of each user type differ during peak hours (07:00 — 09:00 and 16:00 — 19:00?



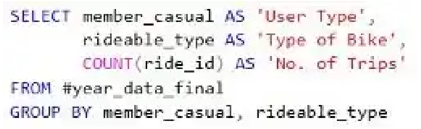
```
SELECT member_casual AS 'User Type',
       day_of_week_name AS 'Day of Week',
       COUNT(ride_id) AS 'No. of Trips'
FROM #year_data_final
WHERE (CAST(started_at AS time) >= '07:00:00' AND CAST(ended_at AS time) <= '09:00:00')
    OR (CAST(started_at AS time) >= '16:00:00' AND CAST(ended_at AS time) <= '19:00:00')
GROUP BY member_casual, day_of_week_name
```

7. How does the usage of each customer type vary during each month?



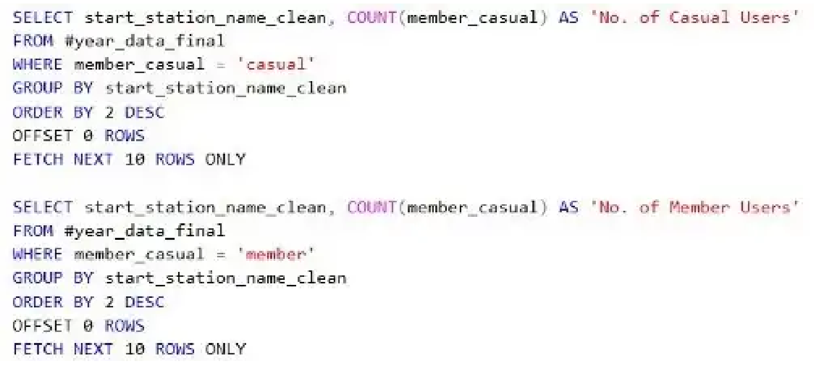
```
SELECT member_casual AS 'User Type',
       month,
       COUNT(ride_id) AS 'No. of Trips'
FROM #year_data_final
GROUP BY member_casual, month
ORDER BY member_casual, month
```

8. How differently does each type of user use the different types of bikes available?



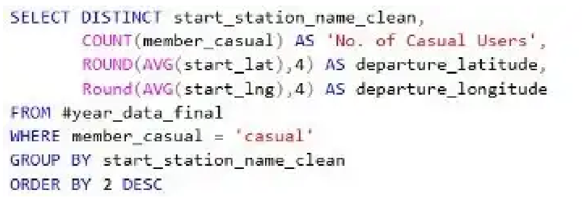
```
SELECT member_casual AS 'User Type',
       rideable_type AS 'Type of Bike',
       COUNT(ride_id) AS 'No. of Trips'
FROM #year_data_final
GROUP BY member_casual, rideable_type
```

9. What are the Top 10 most common “Departing” Stations for casual users and members?



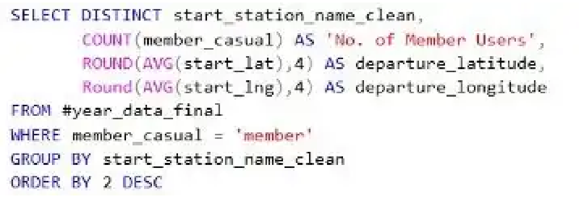
```
SELECT start_station_name_clean, COUNT(member_casual) AS 'No. of Casual Users'
FROM #year_data_final
WHERE member_casual = 'casual'
GROUP BY start_station_name_clean
ORDER BY 2 DESC
OFFSET 0 ROWS
FETCH NEXT 10 ROWS ONLY

SELECT start_station_name_clean, COUNT(member_casual) AS 'No. of Member Users'
FROM #year_data_final
WHERE member_casual = 'member'
GROUP BY start_station_name_clean
ORDER BY 2 DESC
OFFSET 0 ROWS
FETCH NEXT 10 ROWS ONLY
```

10. Group departing station name with distinct Latitude and Longitude for casual users.



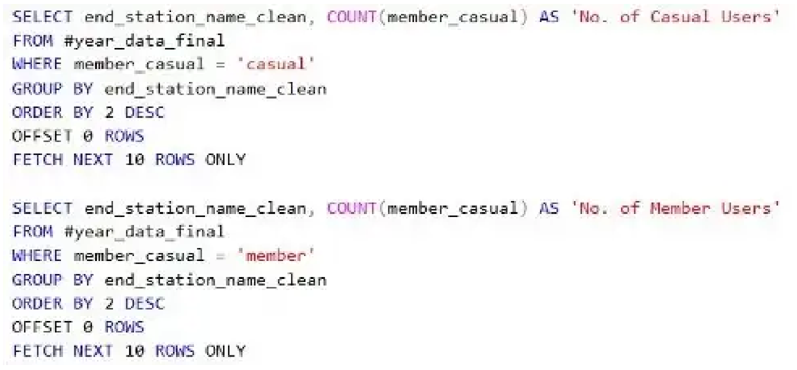
```
SELECT DISTINCT start_station_name_clean,
       COUNT(member_casual) AS 'No. of Casual Users',
       ROUND(AVG(start_lat),4) AS departure_latitude,
       Round(AVG(start_lng),4) AS departure_longitude
FROM #year_data_final
WHERE member_casual = 'casual'
GROUP BY start_station_name_clean
ORDER BY 2 DESC
```

11. Group departing station name with distinct Latitude and Longitude for members.



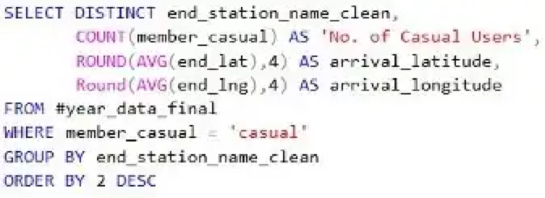
```
SELECT DISTINCT start_station_name_clean,
       COUNT(member_casual) AS 'No. of Member Users',
       ROUND(AVG(start_lat),4) AS departure_latitude,
       Round(AVG(start_lng),4) AS departure_longitude
FROM #year_data_final
WHERE member_casual = 'member'
GROUP BY start_station_name_clean
ORDER BY 2 DESC
```

12. What are the Top 10 most common “Arrival” Stations for casual users and members?



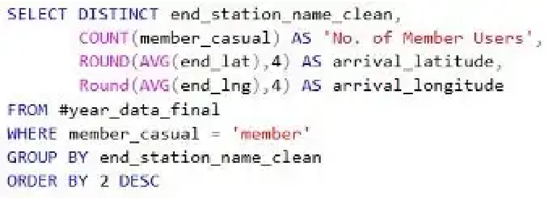
```
SELECT end_station_name_clean, COUNT(member_casual) AS 'No. of Casual Users'
FROM #year_data_final
WHERE member_casual = 'casual'
GROUP BY end_station_name_clean
ORDER BY 2 DESC
OFFSET 0 ROWS
FETCH NEXT 10 ROWS ONLY

SELECT end_station_name_clean, COUNT(member_casual) AS 'No. of Member Users'
FROM #year_data_final
WHERE member_casual = 'member'
GROUP BY end_station_name_clean
ORDER BY 2 DESC
OFFSET 0 ROWS
FETCH NEXT 10 ROWS ONLY
```

13. Group arrival station name with distinct Latitude and Longitude for casual users.



```
SELECT DISTINCT end_station_name_clean,
       COUNT(member_casual) AS 'No. of Casual Users',
       ROUND(AVG(end_lat),4) AS arrival_latitude,
       Round(AVG(end_lng),4) AS arrival_longitude
FROM #year_data_final
WHERE member_casual = 'casual'
GROUP BY end_station_name_clean
ORDER BY 2 DESC
```

14. Group arrival station name with distinct Latitude and Longitude for members.

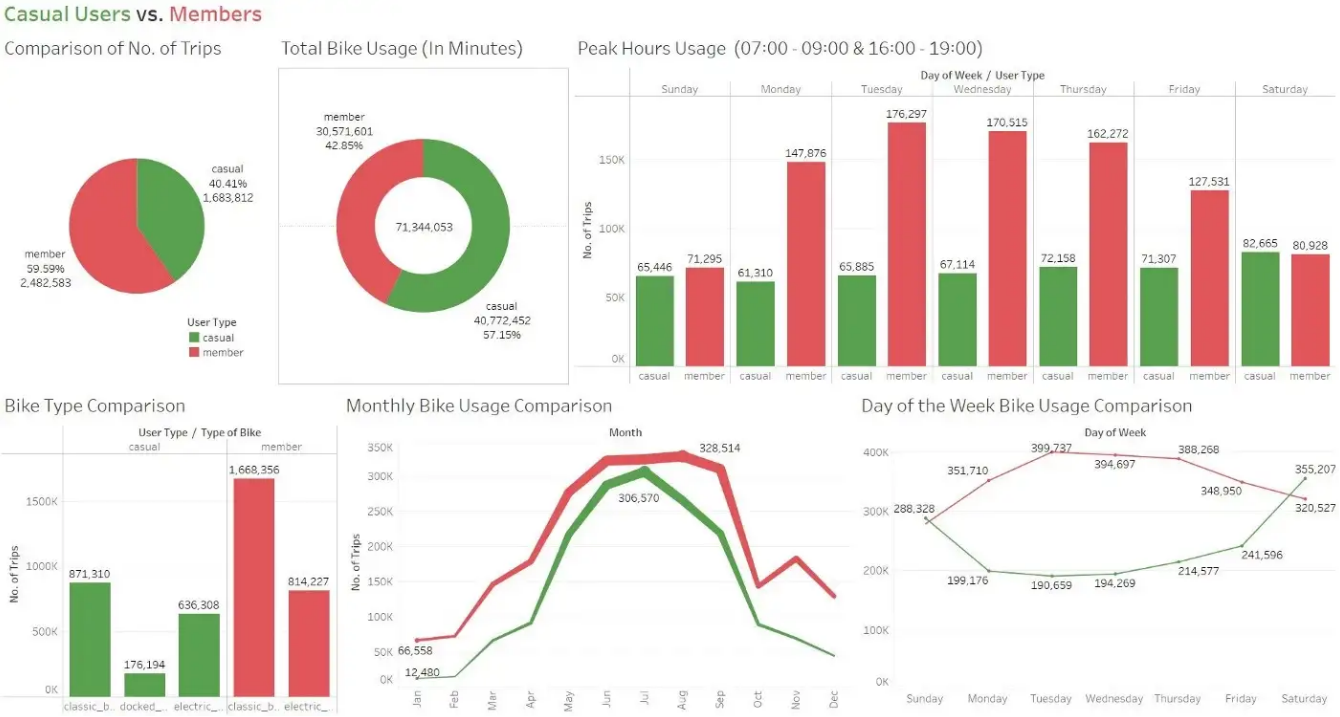


```
SELECT DISTINCT end_station_name_clean,
       COUNT(member_casual) AS 'No. of Member Users',
       ROUND(AVG(end_lat),4) AS arrival_latitude,
       Round(AVG(end_lng),4) AS arrival_longitude
FROM #year_data_final
WHERE member_casual = 'member'
GROUP BY end_station_name_clean
ORDER BY 2 DESC
```

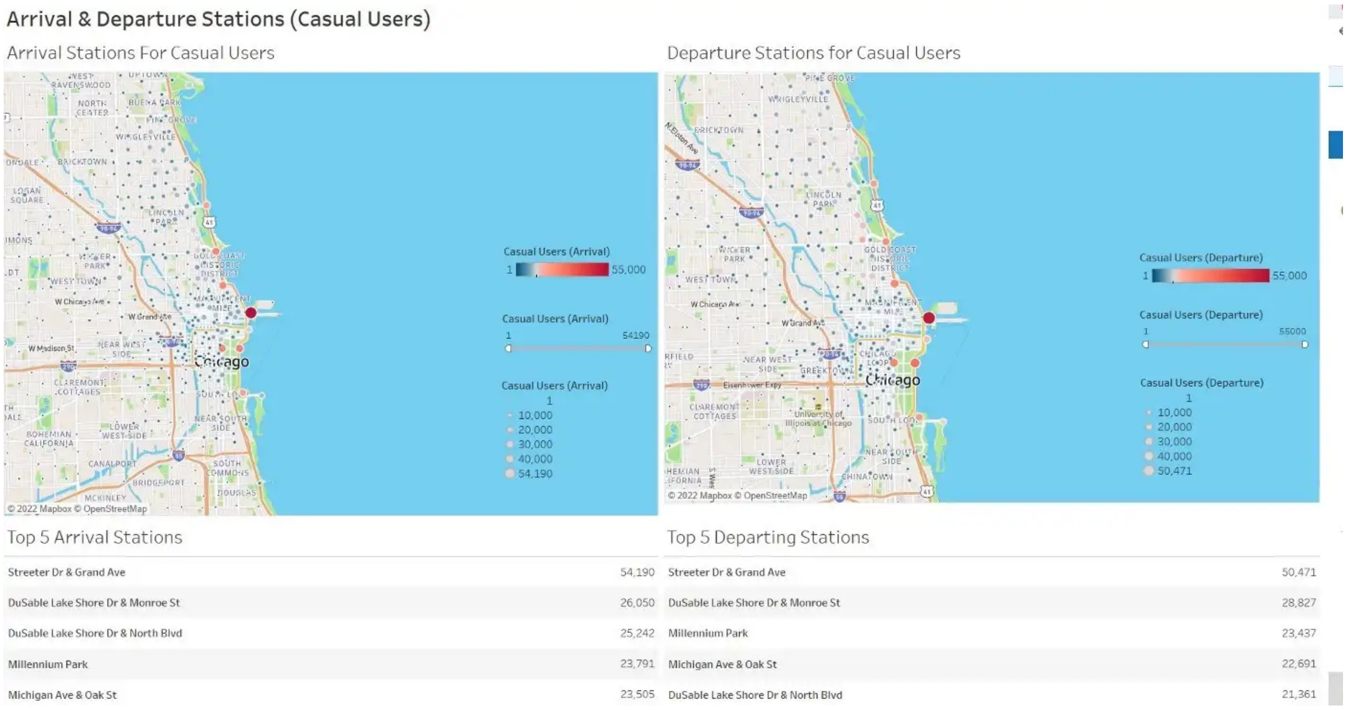
The tabulated data acquired as a result of these SQL queries was copied to a spreadsheet workbook which was then imported into Tableau Public for the purpose of visualization.

Step 5: Share

Multiple Visualizations were created using Tableau Public from the data that was extracted using SQL. With the help of these visualizations, the following dashboard was created which compares how the annual members and casual riders use the Cyclistic bikes differently.



The following maps show how the departing and arrival stations differed for Casual Users and Members.



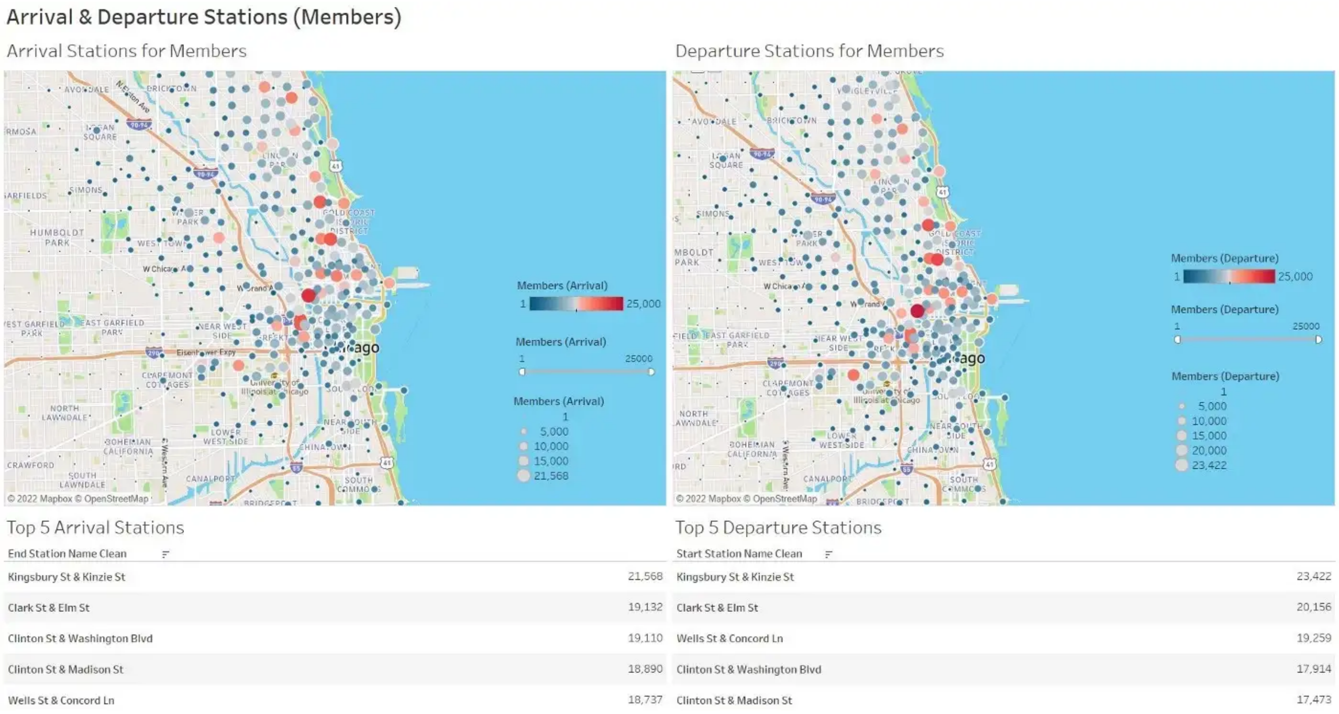
Arrival & Departure Stations (Casual Users)

Top 5 Arrival Stations

Streeter Dr & Grand Ave	54,190
DuSable Lake Shore Dr & Monroe St	26,050
DuSable Lake Shore Dr & North Blvd	25,242
Millennium Park	23,791
Michigan Ave & Oak St	23,505

Top 5 Departing Stations

Streeter Dr & Grand Ave	50,471
DuSable Lake Shore Dr & Monroe St	28,827
Millennium Park	23,437
Michigan Ave & Oak St	22,691
DuSable Lake Shore Dr & North Blvd	21,361



Arrival & Departure Stations (Members)

Top 5 Arrival Stations

End Station Name Clean	
Kingsbury St & Kinzie St	21,568
Clark St & Elm St	19,132
Clinton St & Washington Blvd	19,110
Clinton St & Madison St	18,890
Wells St & Concord Ln	18,737

Top 5 Departure Stations

Start Station Name Clean	
Kingsbury St & Kinzie St	23,422
Clark St & Elm St	20,156
Wells St & Concord Ln	19,259
Clinton St & Washington Blvd	17,914
Clinton St & Madison St	17,473

Step 6: Act

Following are some key findings from the visualizations made using Tableau:

- Members tend to ride the bikes for shorter durations in comparison to casual riders. This is evident from the different pictures portrayed by the pie chart and the doughnut chart. The pie chart shows that members make around 60% of all trips. However, as per the doughnut chart, they only constitute approximately 43% of total bikes usage which conclusively means shorter trips for members. One reason for this could be the short ride transit from train stations to their offices or homes and vice versa, assuming that members category comprises mainly of adults who use the bikes daily for travelling to and from work.

- Members tend to ride the classic bikes twice as often as the electric ones. However, such drastic difference of usage between the different types of bikes can not be observed with casual users. Moreover, members don't prefer to use the docked bikes as all the trips using docked bikes have been made by casual riders.
- For both user types, there is a dramatic increase in the usage of bikes from April and a similar decrease in usage in September. This shows that regardless of the user type, riders prefer a warmer weather.
- A weekday vs. weekend comparison and the peak hours usage comparison shows us clearly why my aforementioned assumption regarding members primarily belonging to the working adults category holds true. Members and casual members barely differ in their usage over the weekend. However, during weekdays the members' usage is roughly twice as much as the casual riders' usage. Furthermore, during the peak hours, i.e. hours during which people normally commute to and from work, members show a clear spike in usage during weekdays in comparison to the casual riders, for whom the usage remains fairly constant regardless of which day of the week it is.
- Casual riders are likely to be comprised mostly of tourists or families who visit the beaches for engaging in leisure activities. This explains why the most visited stations by casual riders are all concentrated around the coastal area. On the other hand, for members, the trips are more spread out as most users use the bike to commute to work and offices are scattered all across the city.

In light of the aforementioned insights, the next step is to put forward a few recommendations which would help accomplish the overall goal of converting the casual riders to members.

- Keeping in mind the fact that the usage of bikes sharply increases in the month of April, the marketing campaign should be launched around that time in order to attract the most number of users.
- Prices during the off-peak hours and weekends can be reduced for members only. Since the existing members mainly belong to the working adults category,

their usage wouldn't change thereby not impacting the revenue generated from their usage. However, many casual riders might switch to buying memberships in order to make use of reduced prices.

- A different pricing method can be employed for Members only. For instance, rides which last longer than 10 minutes to be charged at half the normal price. It wouldn't in any way negatively impact the existing members or the revenue generated. However, as we already know casual riders tend to make longer trips, they would be attracted to subscribe to the membership in order to take advantage of the additional convenience granted to members only.

Cyclistic Sql Tableau Google Data Analytics