Course: Google Data Analytics Capstone: Complete a Case Study

Introduction

In this case study, I will perform many real-world tasks of a junior data analyst at a fictional company, Cyclistic. In order to answer the key business questions, I will follow the steps of the data analysis process:

ASK PREPARE PROCESS ANALYZE SHARE ACT

Quick links:

Data Source: divvy_tripdata

SQL Queries:
01. Data Combining
02. Data Exploration
03. Data Cleaning

Data Visualizations: Power BI

Background

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.

Until now, Cyclistic’s marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach 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.

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 (the director of marketing and my manager) 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.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst 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. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Scenario

I am assuming to be a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing 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, my team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve our recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Ask

Business Task

Devise marketing strategies to convert casual riders to members.

Analysis Questions

Three questions will guide the future marketing program:

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?

Moreno has assigned me the first question to answer: How do annual members and casual riders use Cyclistic bikes differently?

Prepare

Data Source

I will use Cyclistic’s historical trip data to analyze and identify trends from Jan 2023 to Dec 2023 which can be downloaded from divvy_tripdata. The data has been made available by Motivate International Inc. under this license.

This is public data that can be used to explore how different customer types are using Cyclistic bikes. But note that data-privacy issues prohibit from using riders’ personally identifiable information. This means that we won’t be able to connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

Data Organization

There are 12 files with naming convention of YYYYMM-divvy-tripdata and each file includes information for one month, such as the ride id, bike type, start time, end time, start station, end station, start location, end location, and whether the rider is a member or not. The corresponding column names are ride_id, rideable_type, started_at, ended_at, start_station_name, start_station_id, end_station_name, end_station_id, start_lat, start_lng, end_lat, end_lng and member_casual.

Process

PostGre SQL is used to combine the various datasets into one dataset and clean it.
Reason:
A worksheet can only have 1,048,576 rows in Microsoft Excel because of its inability to manage large amounts of data. Because the Cyclistic dataset has more than 5.6 million rows, it is essential to use a platform like BigQuery that supports huge volumes of data.

Combining the Data

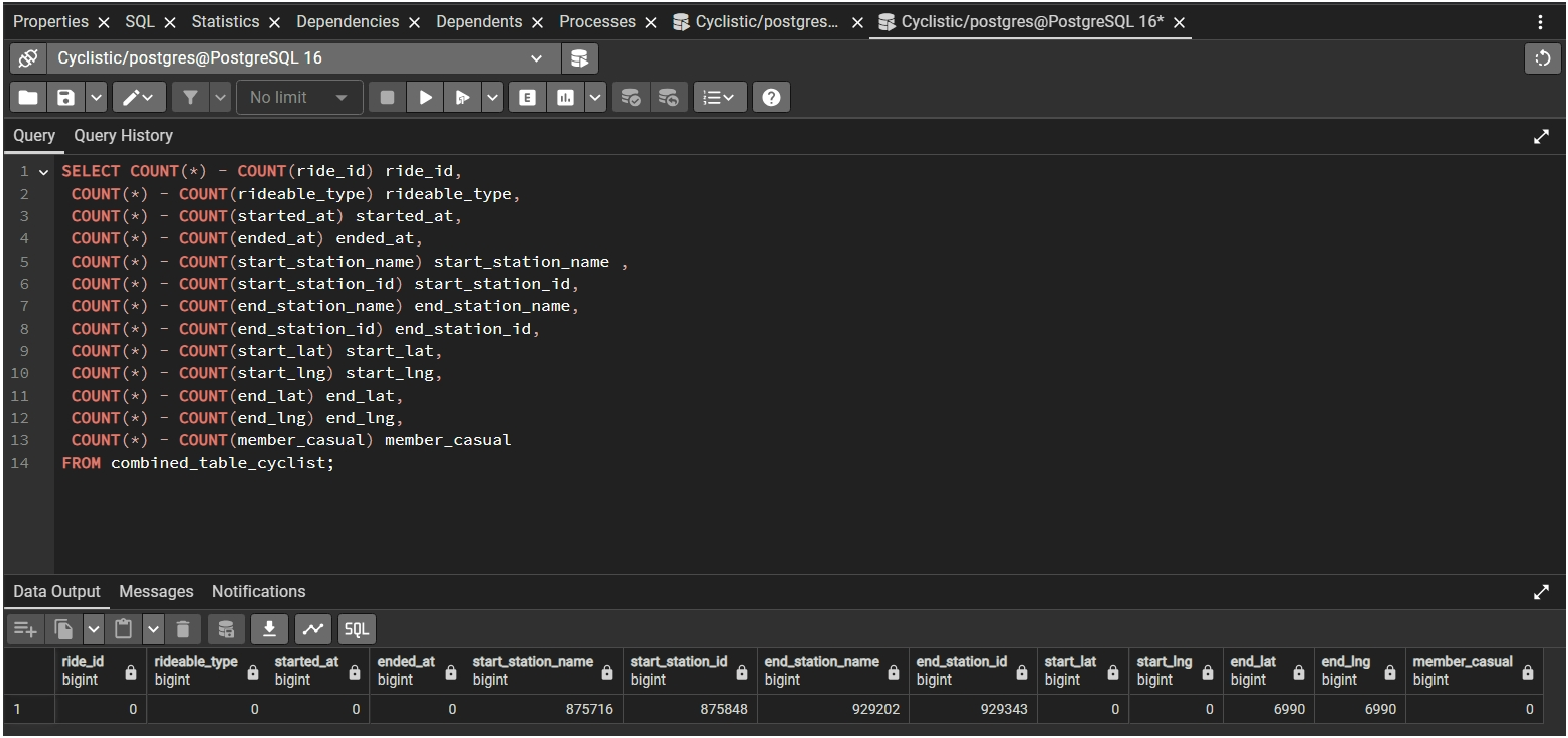
SQL Query: Data Combining
12 csv files ('202301-divvy-tripdata.csv' to '202312-divvy-tripdata.csv') are combined using the COPY command and a Table named "combined_data_cyclist" is created, containing 5,719,877 rows of data for the entire year.

Data Exploration

SQL Query: Data Exploration
Before cleaning the data, I am familiarizing myself with the data to find the inconsistencies.

Observations:

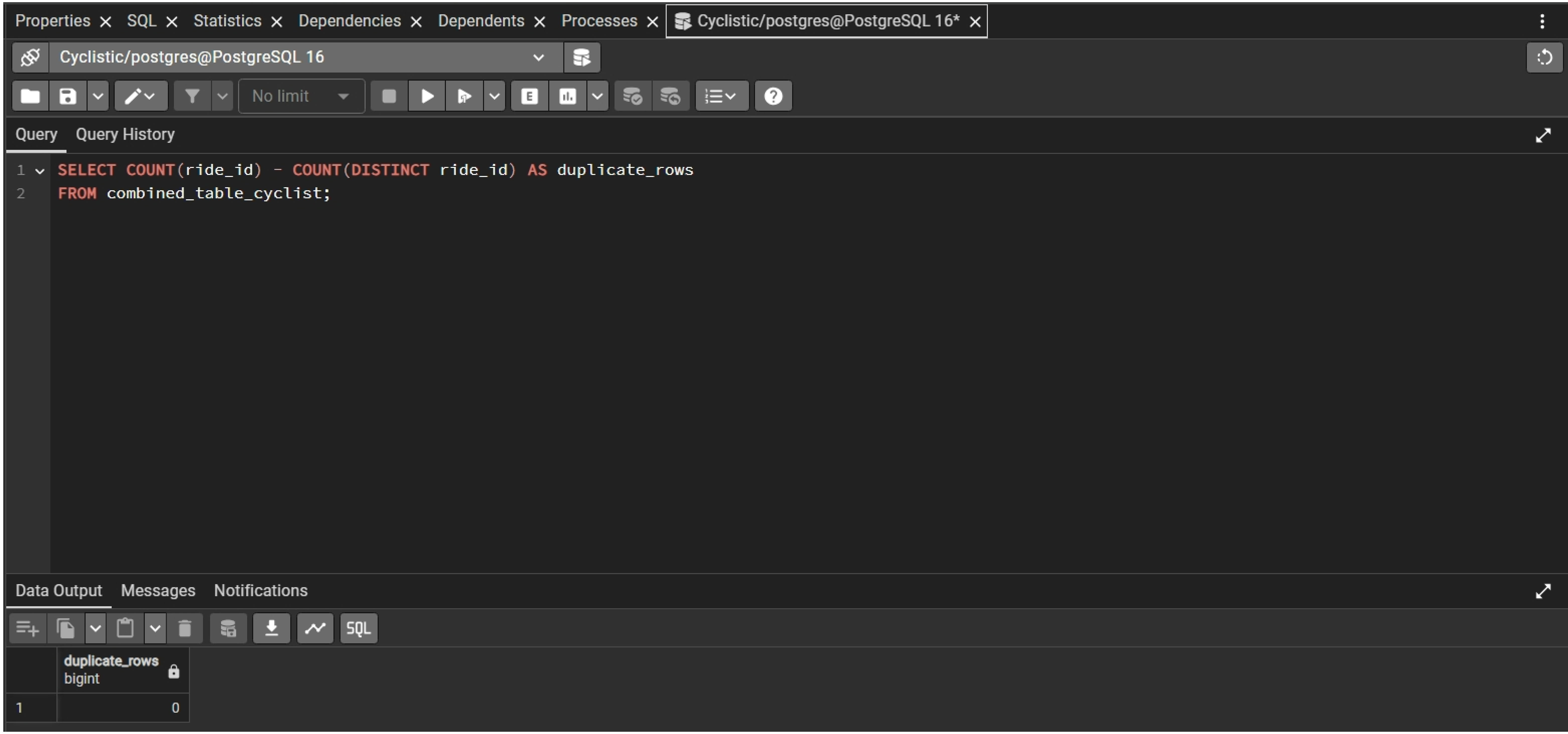
1. The following table shows number of **null values** in each column.



```sql
SELECT COUNT(*) - COUNT(ride_id) ride_id,
 COUNT(*) - COUNT(rideable_type) rideable_type,
 COUNT(*) - COUNT(started_at) started_at,
 COUNT(*) - COUNT(ended_at) ended_at,
 COUNT(*) - COUNT(start_station_name) start_station_name ,
 COUNT(*) - COUNT(start_station_id) start_station_id,
 COUNT(*) - COUNT(end_station_name) end_station_name,
 COUNT(*) - COUNT(end_station_id) end_station_id,
 COUNT(*) - COUNT(start_lat) start_lat,
 COUNT(*) - COUNT(start_lng) start_lng,
 COUNT(*) - COUNT(end_lat) end_lat,
 COUNT(*) - COUNT(end_lng) end_lng,
 COUNT(*) - COUNT(member_casual) member_casual
FROM combined_table_cyclist;
```

	ride_id bigint	rideable_type bigint	started_at bigint	ended_at bigint	start_station_name bigint	start_station_id bigint	end_station_name bigint	end_station_id bigint	start_lat bigint	start_lng bigint	end_lat bigint	end_lng bigint	member_casual bigint
1	0	0	0	0	875716	875848	929202	929343	0	0	6990	6990	0

2. As ride_id has no null values, let's use it to check for duplicates.

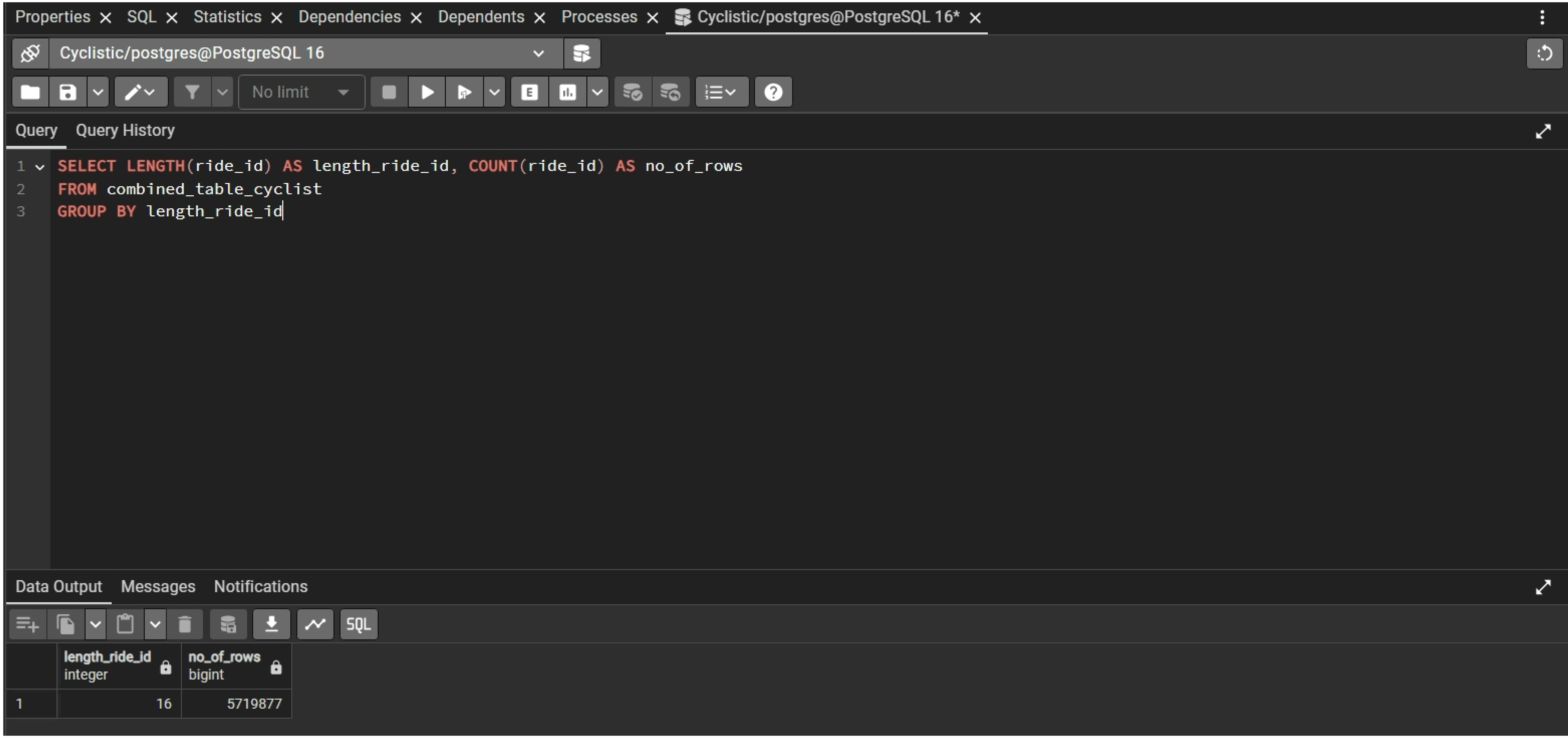


```
SELECT COUNT(ride_id) - COUNT(DISTINCT ride_id) AS duplicate_rows
FROM combined_table_cyclist;
```

	duplicate_rows bigint
1	0

There are no **duplicate** rows in the data.

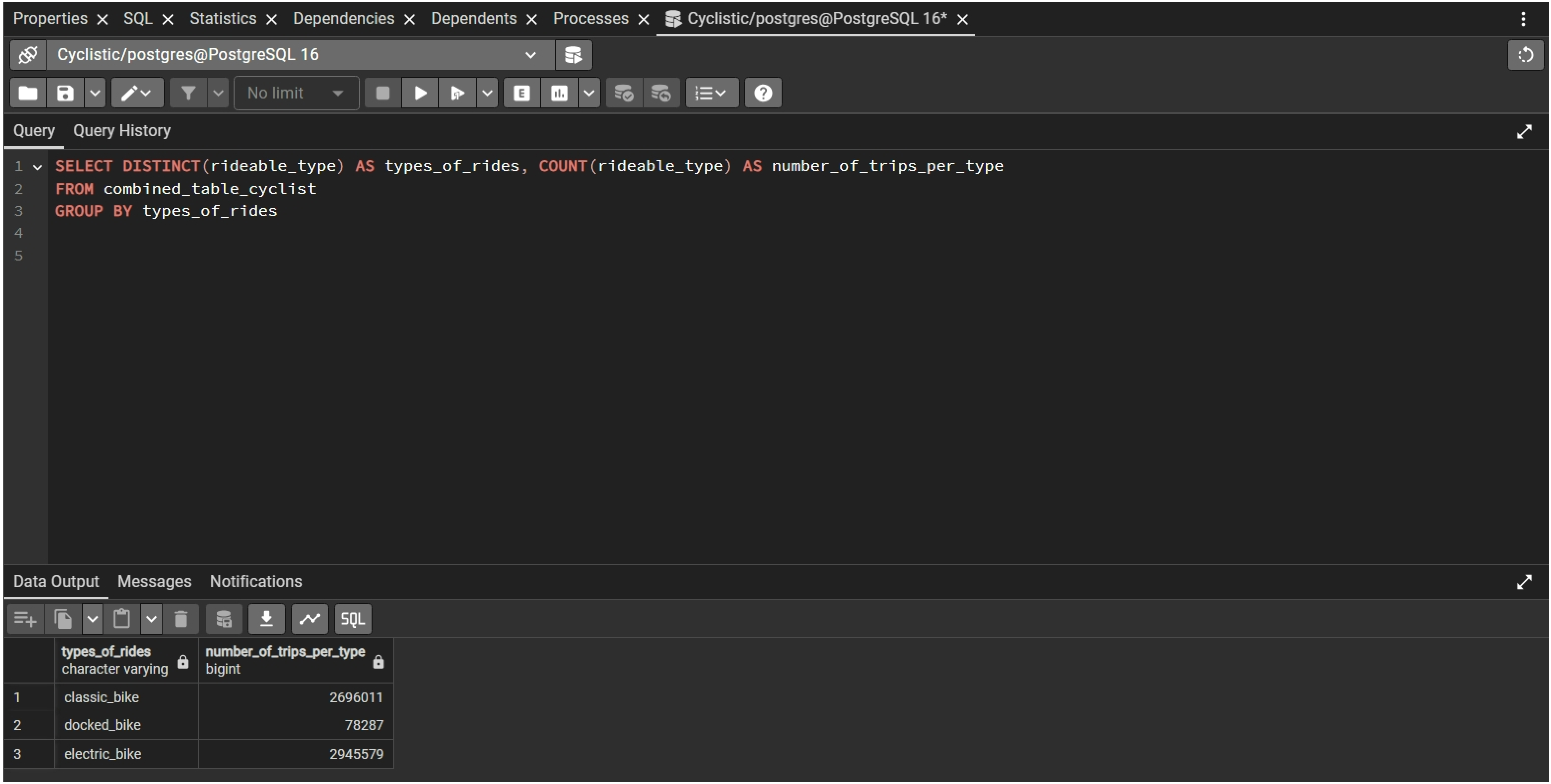
3. All **ride_id** values have length of 16 so no need to clean it.



```
SELECT LENGTH(ride_id) AS length_ride_id, COUNT(ride_id) AS no_of_rows
FROM combined_table_cyclist
GROUP BY length_ride_id
```

	length_ride_id integer	no_of_rows bigint
1	16	5719877

4. There are 3 unique types of bikes(**rideable_type**) in our data.



```sql
SELECT DISTINCT(rideable_type) AS types_of_rides, COUNT(rideable_type) AS number_of_trips_per_type
FROM combined_table_cyclist
GROUP BY types_of_rides
```

	types_of_rides character varying	number_of_trips_per_type bigint
1	classic_bike	2696011
2	docked_bike	78287
3	electric_bike	2945579

5. The **started_at** and **ended_at** shows start and end time of the trip in YYYY-MM-DD hh:mm:ss UTC format. New column ride_length can be created to find the total trip duration. There are 0 trips which has duration longer than a day and 151070 trips having less than a minute duration or having end time earlier than start time so we need to remove them. Other columns day_of_week and month can also be helpful in analysis of trips at different times in a year.

6. Total of 875716 rows have both **start_station_name** and **start_station_id** missing which needs to be removed.

7. Total of 929202 rows have both **end_station_name** and **end_station_id** missing which needs to be removed.

8. Total of 6990 rows have both **end_lat** and **end_lng** missing which needs to be removed.

9. Columns that need to be removed are start_station_id and end_station_id as they do not add value to analysis of our current problem. Longitude and latitude location can be used to visualise a map.

Data Cleaning

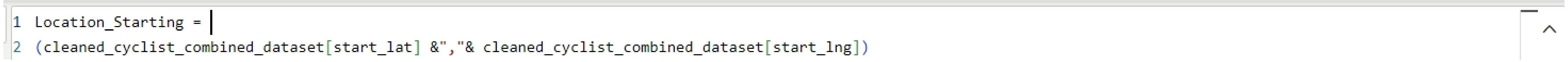
SQL Query: Data Cleaning

1. All the rows having missing values are deleted.

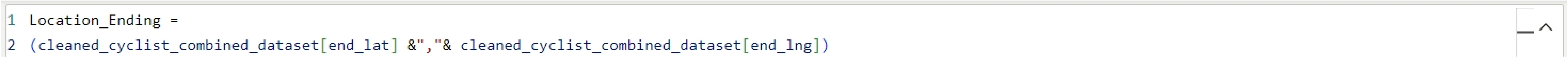
2. 3 more columns ride_length for duration of the trip, day_of_week and month are added.
3. Trips with duration less than a minute are excluded.
4. Total 1,475,449 rows are removed in this step.

Analyze and Share

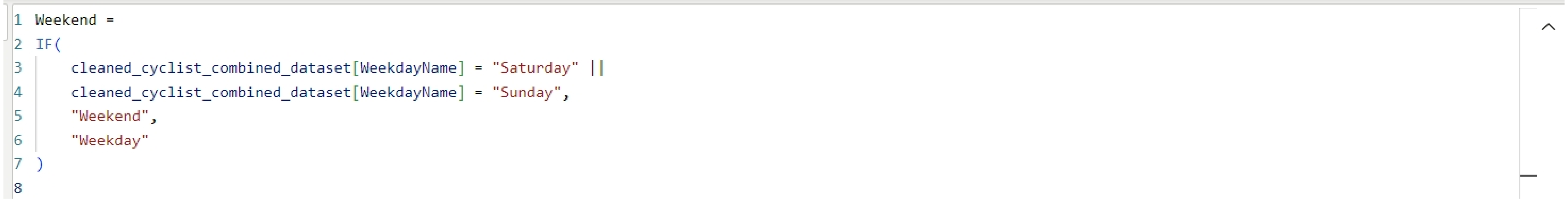
Data Visualization: Power BI
The data is stored appropriately and is now prepared for analysis. I used Power BI Desktop to analyse the 'cleaned_cyclist_combined_dataset.csv' file, I developed four calculated columns named Location_Starting, Location_Ending, Weekend, and Seasonal_Category. These columns were derived using the latitude and longitude data for the start and end points of the journeys, the day of the week data, and the month data, respectively. DAX Code for the respective columns are -



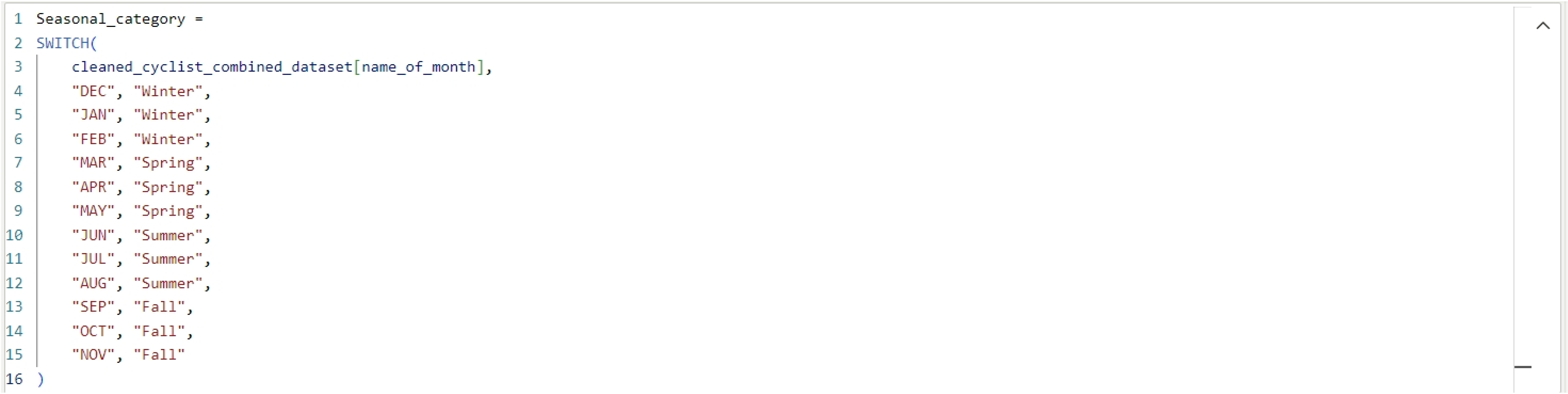
```
Location_Starting =
(cleaned_cyclist_combined_dataset[start_lat] &","& cleaned_cyclist_combined_dataset[start_lng])
```



```
Location_Ending =
(cleaned_cyclist_combined_dataset[end_lat] &","& cleaned_cyclist_combined_dataset[end_lng])
```



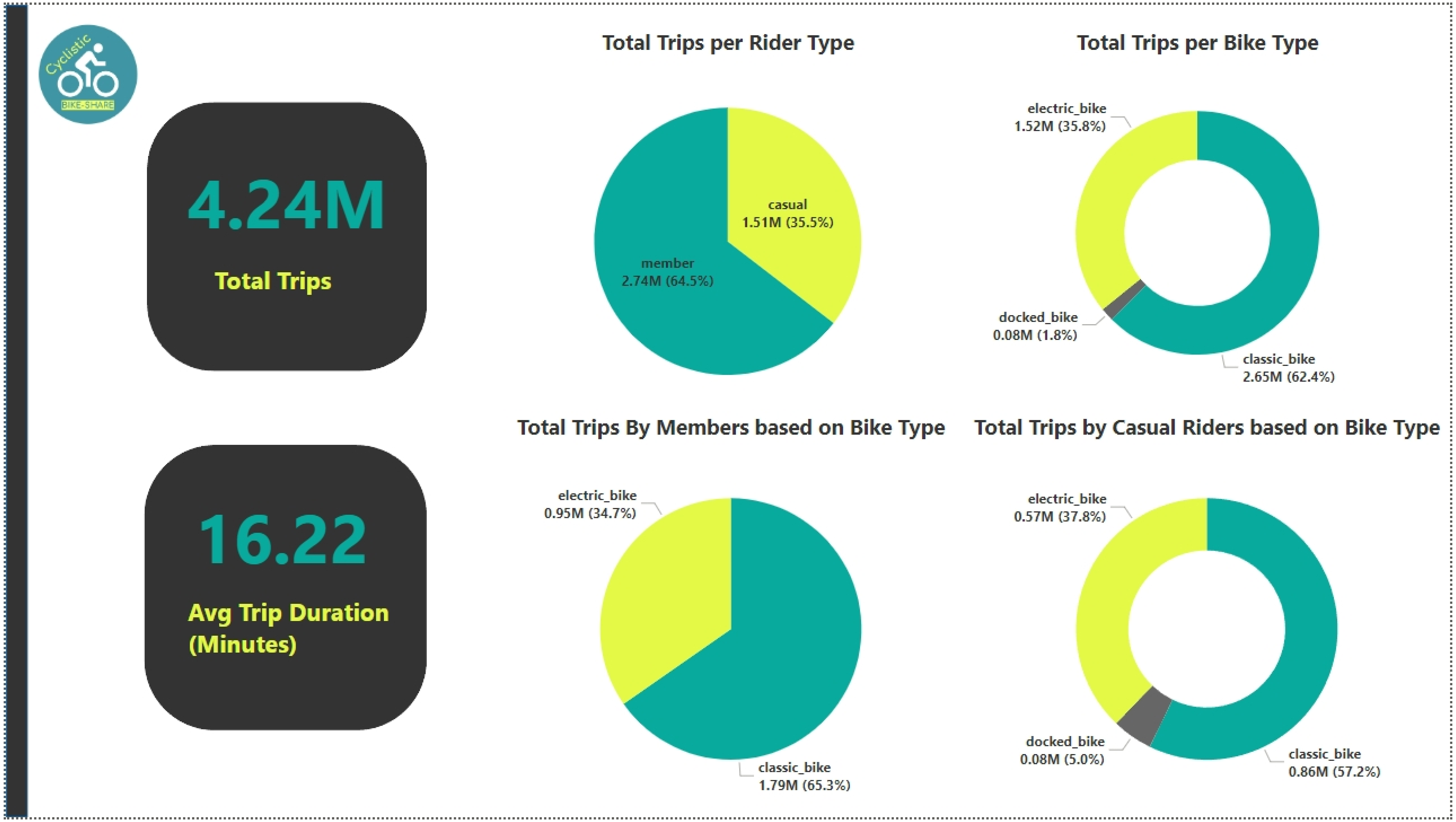
```
Weekend =
IF(
    cleaned_cyclist_combined_dataset[WeekdayName] = "Saturday" ||
    cleaned_cyclist_combined_dataset[WeekdayName] = "Sunday",
    "Weekend",
    "Weekday"
)

```



```
Seasonal_category =
SWITCH(
    cleaned_cyclist_combined_dataset[name_of_month],
    "DEC", "Winter",
    "JAN", "Winter",
    "FEB", "Winter",
    "MAR", "Spring",
    "APR", "Spring",
    "MAY", "Spring",
    "JUN", "Summer",
    "JUL", "Summer",
    "AUG", "Summer",
    "SEP", "Fall",
    "OCT", "Fall",
    "NOV", "Fall"
)
```

The analysis question is:

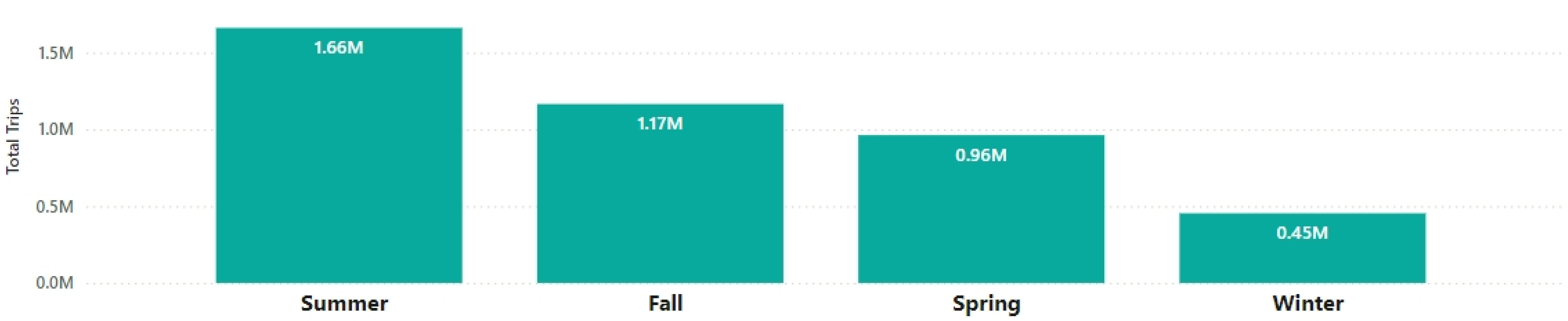
How do annual members and casual riders use Cyclistic bikes differently?



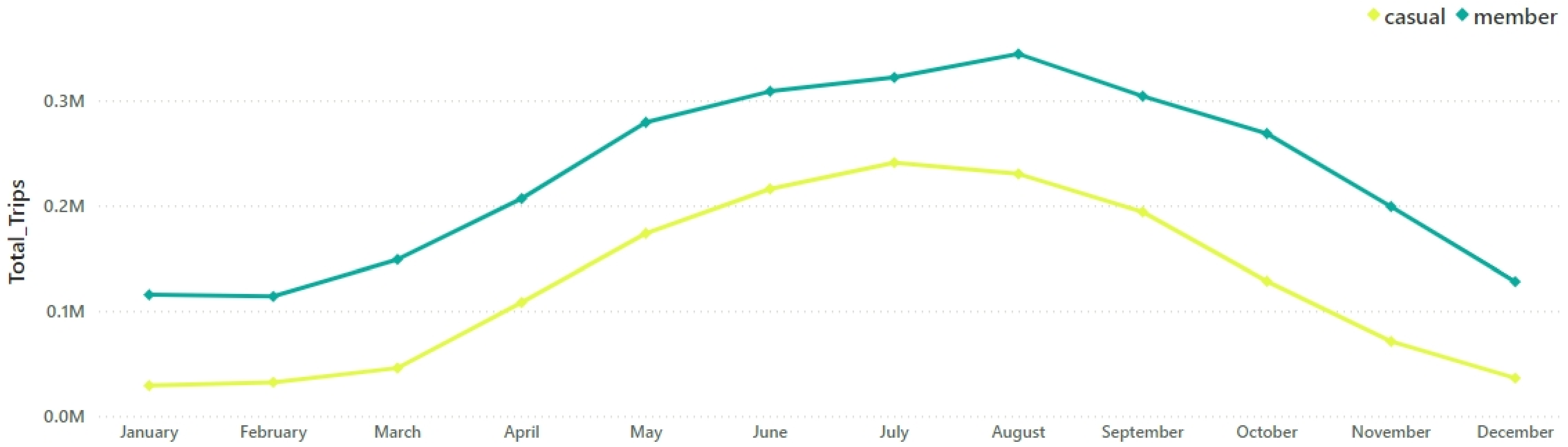
The members make 64.5% of the total while remaining 35.5% constitutes casual riders. Most used bike is classic bike followed by the electric bike. Docked bikes are used the least and that too only by casual riders.

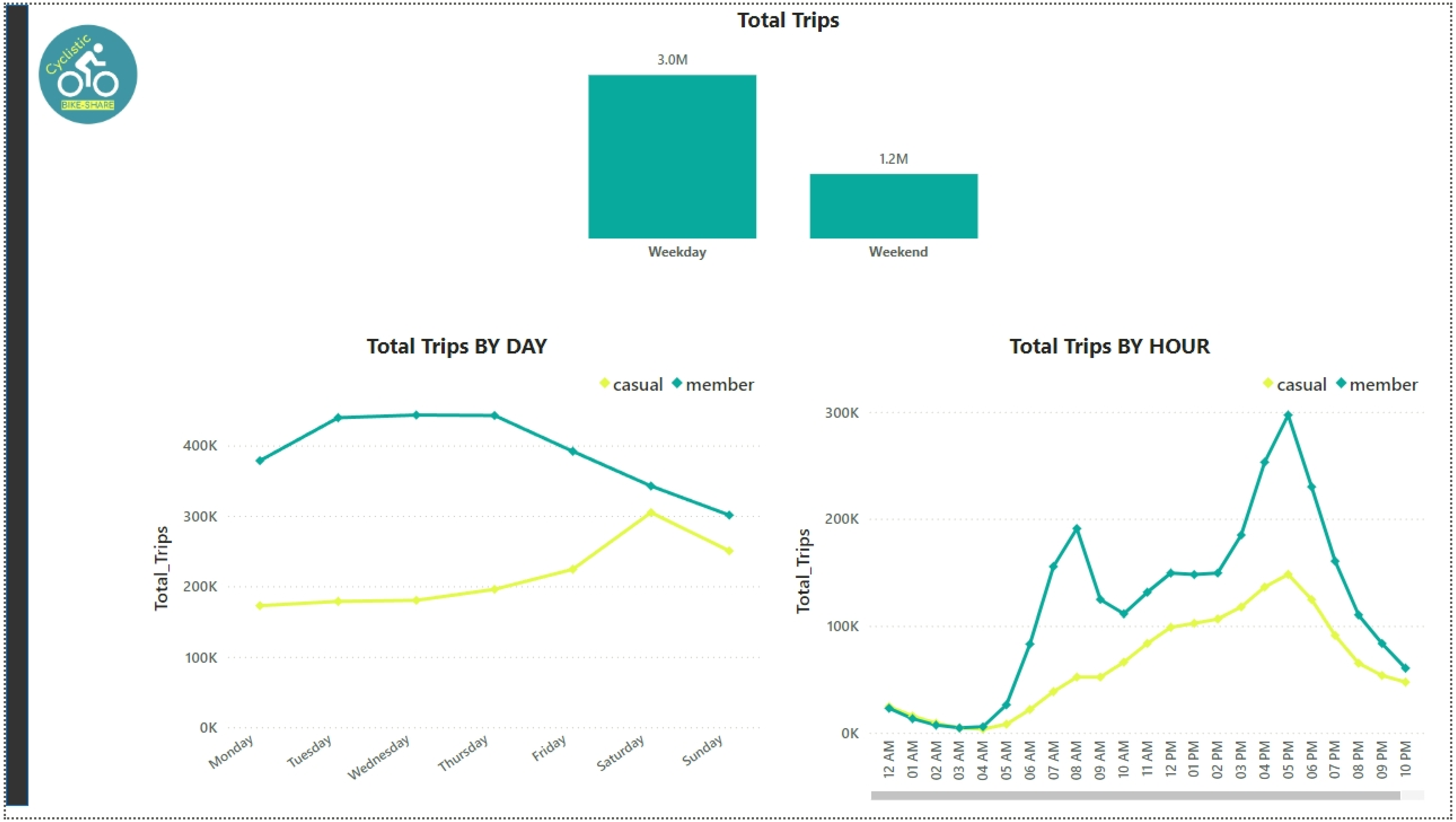
Next the number of trips distributed by the months, by seasons, days of the week and hours of the day are examined.

Total Trips BY SEASONS



Total Trips BY MONTH





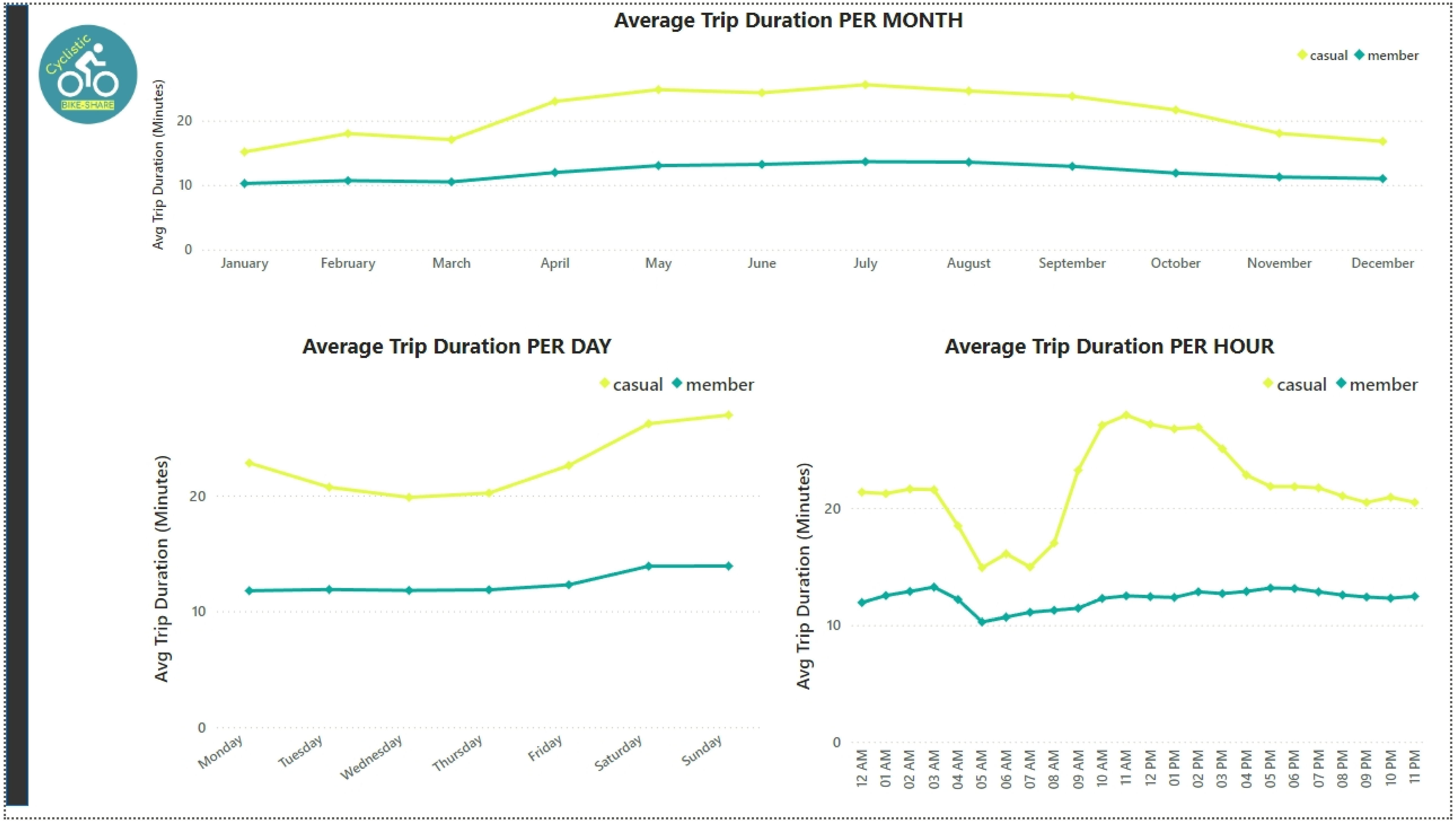
Months: When it comes to monthly trips, both casual and members exhibit comparable behavior, with more trips in the summer and fall and fewer in the spring and least in winter. The gap between casuals and members is closest in the month of july in summmer.

Days of Week: When the days of the week are compared, it is discovered that casual riders make more journeys on the weekends, particularly on Saturday, as compared to weekdays while members show a decline over the weekend in contrast to weekdays. Still the total trips over Weekdays are more than double in number to Weekends.

Hours of the Day: The members shows 2 peaks throughout the day in terms of number of trips. One is early in the morning at around 7 am to 8 am and other is in the evening at around 4 pm to 6 pm while number of trips for casual riders increase consistently over the day till evening (5-6 pm) and then decrease afterwards.

We can infer from the previous observations that member may be using bikes for commuting to and from the work in the week days while casual riders are using bikes throughout the day, more frequently over the weekends for leisure purposes. Both are most active in summer and fall.

Ride duration of the trips are compared to find the differences in the behavior of casual and member riders.

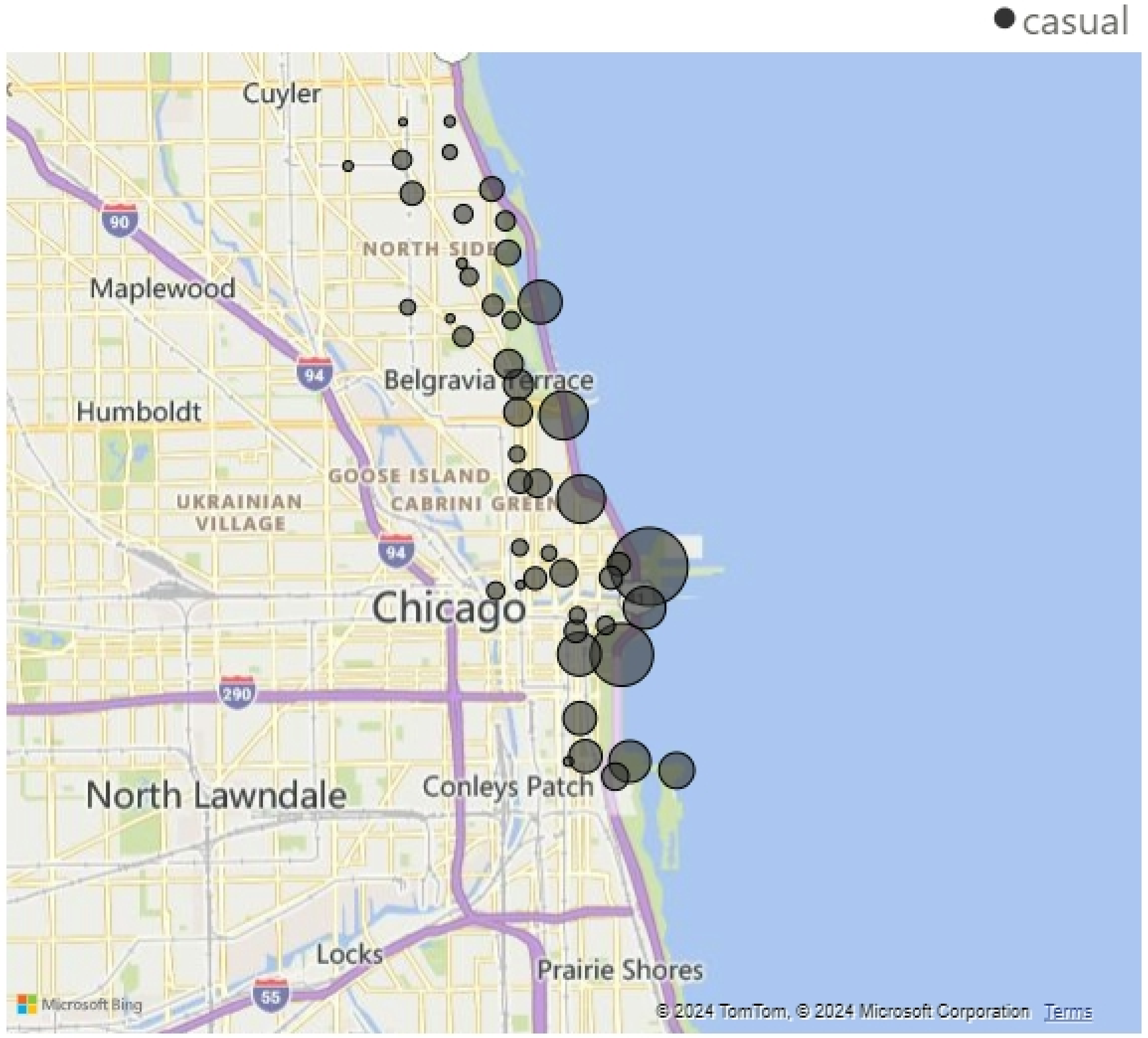


Take note that casual riders tend to cycle longer than members do on average. The length of the average journey for members doesn't change much throughout the month, day or hour. However, there are variations in how long casual riders cycle. In the summer and fall, on weekends, and from 10 am to 2 pm during the day, they travel greater distances. Between five and eight in the morning, they have brief trips.

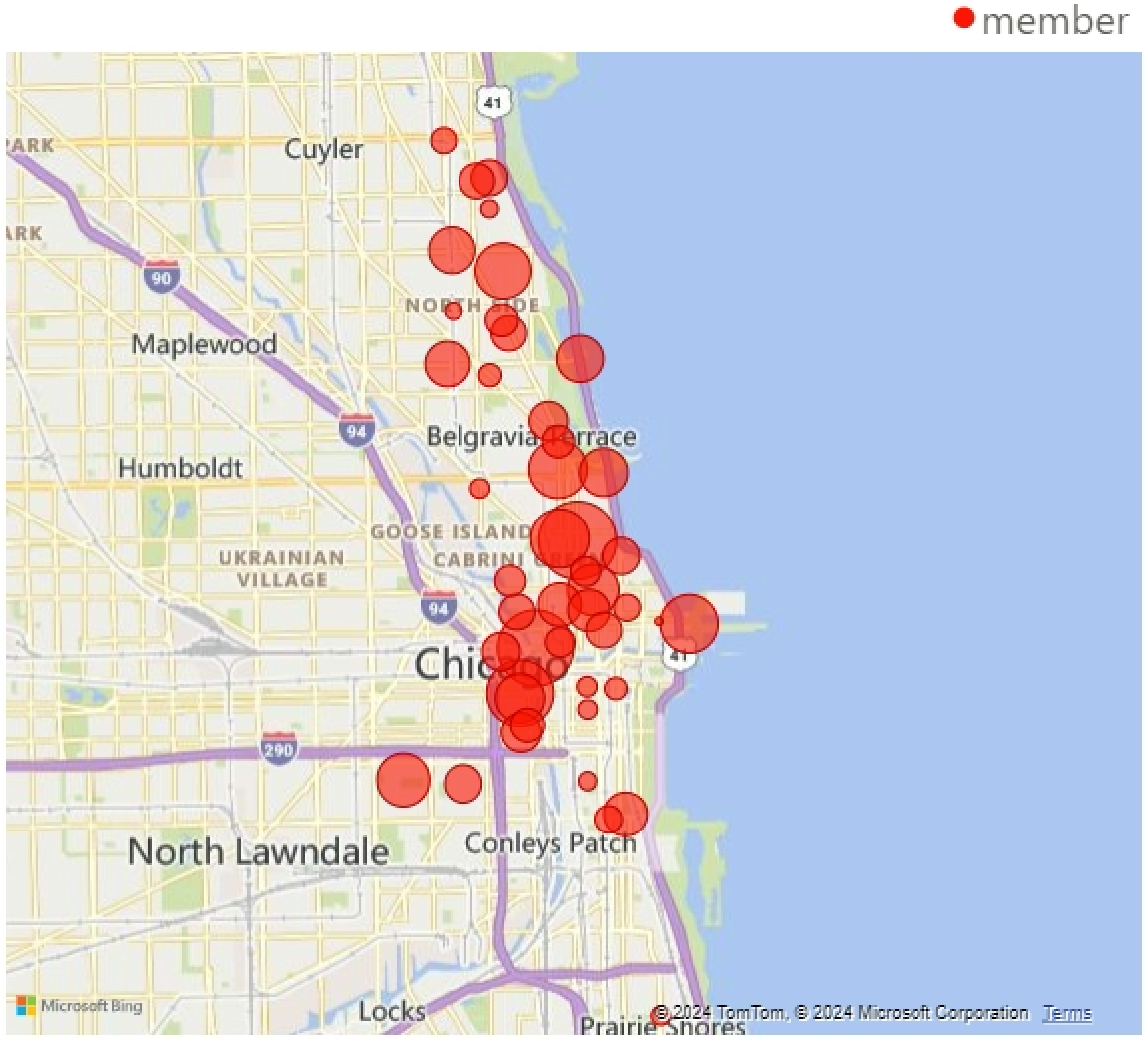
These findings lead to the conclusion that casual commuters travel longer (approximately 2x more) but less frequently than members. They make longer journeys on weekends and during the day outside of commuting hours and in summer and fall season, so they might be doing so for recreation purposes.

To further understand the differences in casual and member riders, locations of starting and ending stations can be analysed, Stations with the most trips (Top 15) are considered using filters to draw out the following conclusions.

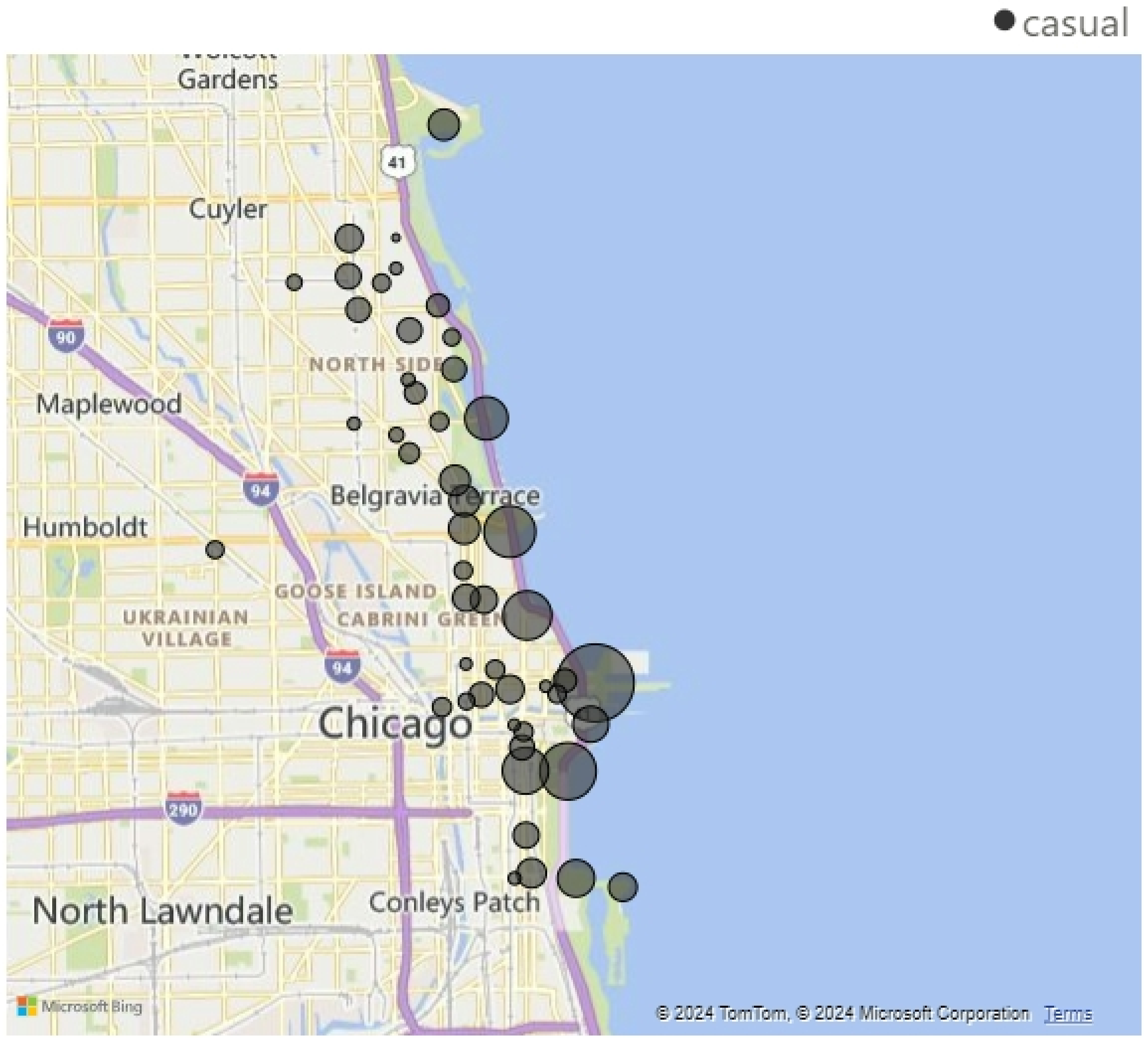
Top 50 Starting locations with Most Trips



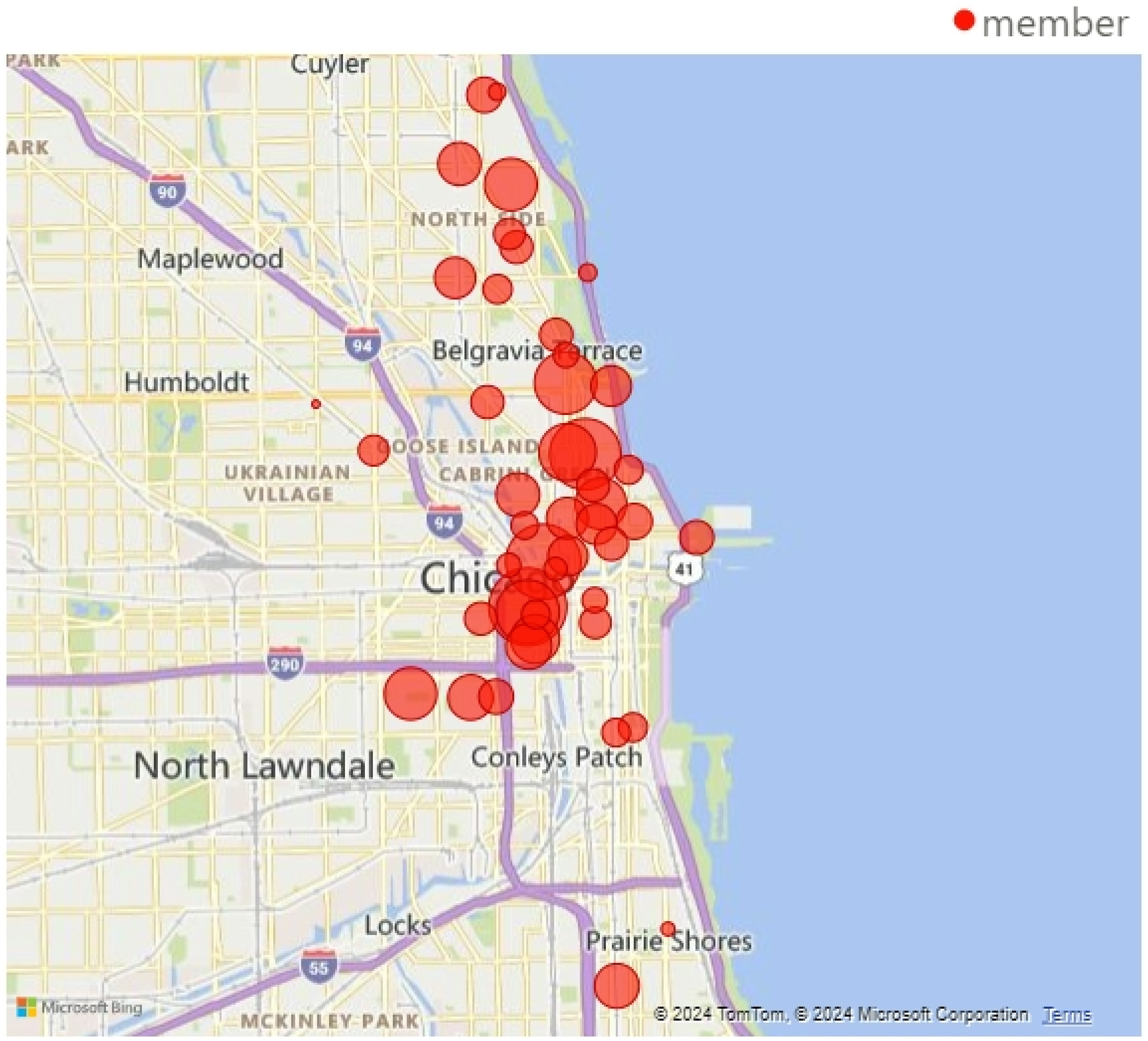
Top 50 Starting locations with Most Trips



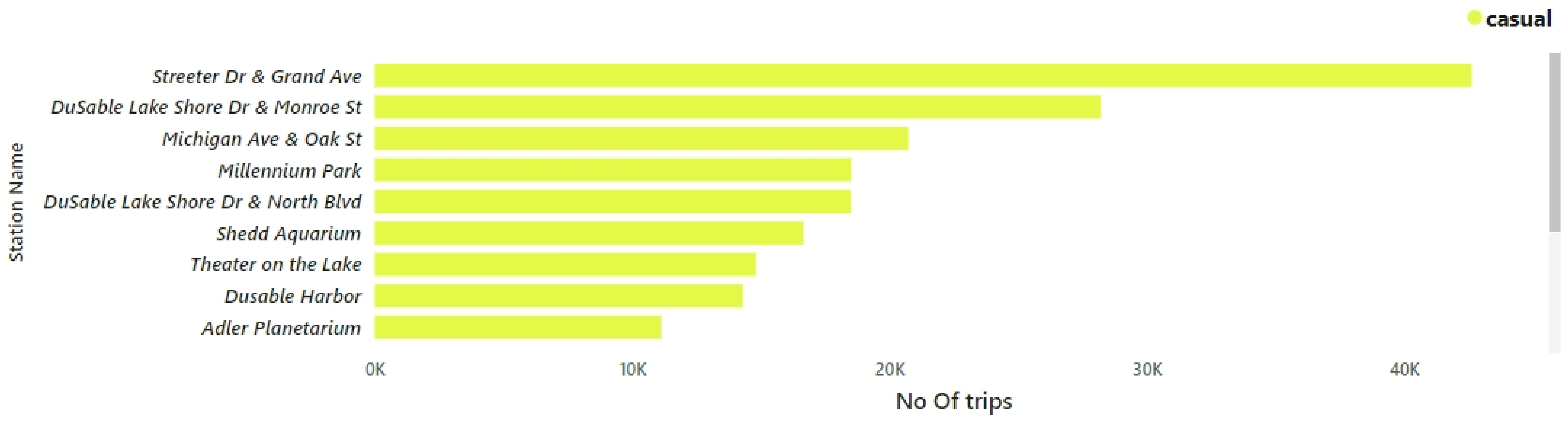
Top 50 Ending locations with Most Trips



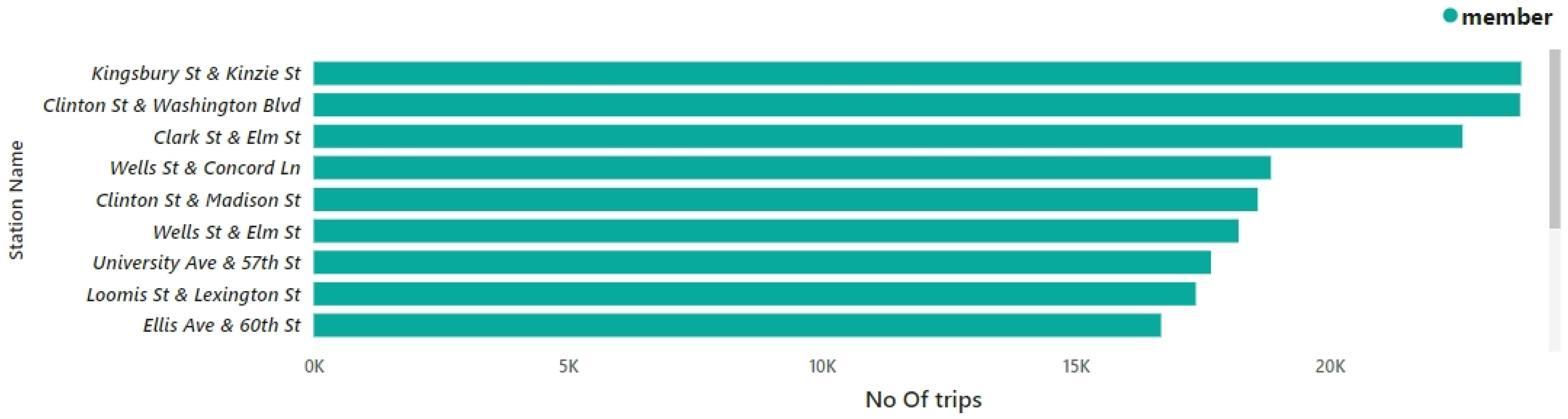
Top 50 Ending locations with Most Trips

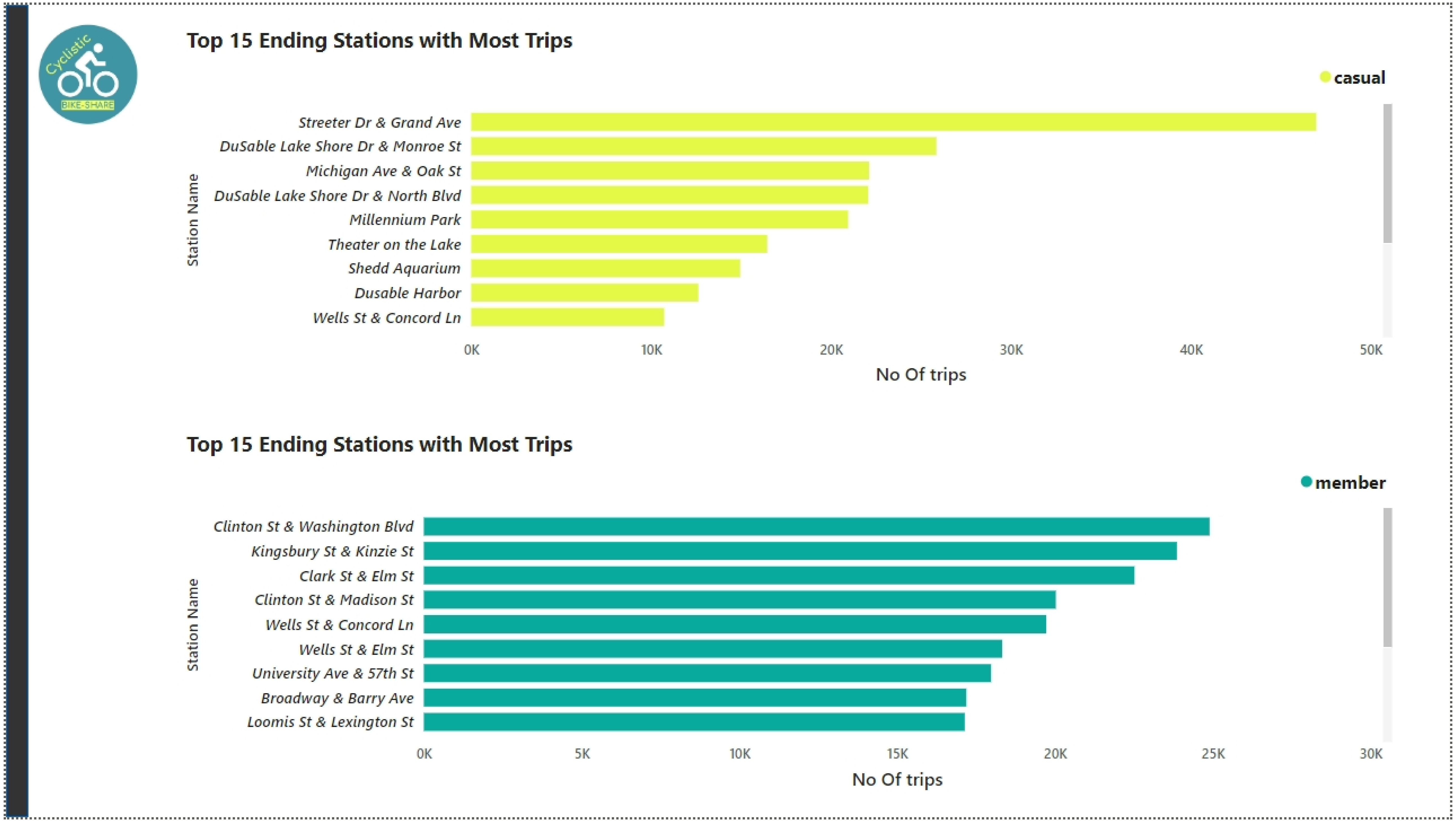


Top 15 Starting Stations with Most Trips



Top 15 Starting Stations with Most Trips





Casual riders have frequently started their trips from the stations in vicinity of commercial districts, museums, parks, beach, harbor points, aquarium (near the bay area) while members have begun their journeys from stations close to universities, residential areas, restaurants, hospitals, grocery stores, theatre, schools, banks, factories, train stations, parks and plazas. So, members tend to expand their rides to Suburban areas possibly for commuting purposes.

Similar trend can be observed in ending station locations. Casual riders end their journey near parks, museums and other recreational sites whereas members end their trips close to universities, residential and commmercial areas. So this proves that casual riders use bikes for leisure activities while members extensively rely on them for daily commute.

Summary:

Casual	**Member**
Prefer using bikes throughout the day, more frequently over the weekends in summer and spring for leisure activities.	Prefer riding bikes on week days during commute hours (8 am / 5pm) in summer and spring.
Travel 2 times longer but less frequently than members.	Travel more frequently but shorter rides (approximately half of casual riders' trip duration).
Start and end their journeys near parks, museums, along the coast and other recreational sites.	Start and end their trips close to universities, residential and commercial areas.

Act

After identifying the differences between casual and member riders, marketing strategies to target casual riders can be developed to persuade them to become members.

Recommendations:

1. Marketing campaigns might be conducted in spring and summer at tourist/recreational locations popular among casual riders.
2. Casual riders are most active on weekends and during the summer and spring, thus they may be offered seasonal or weekend-only memberships.
3. Casual riders use their bikes for longer durations than members. Offering discounts for longer rides may incentivize casual riders and entice members to ride for longer periods of time.
4. Consider alternatives to conversion, such as new service and price offerings Weekend Pass • Yearly subscription providing an unlimited pass for every weekend • Introduces a middle pricing tier that could be the basis for pricing optimization
5. Explore ways to convey the benefits of more frequent biking (Social Media Nudge Model), to influence consumer behavior subtly by leveraging psychological insights.

   Example -

   By using Cyclistic...

   You saved $105.50 otherwise used on gas

   You burned 10,000 calories

   You increased your life expectancy by 7 years

   You saved the planet from the equivalent of 3 gas tanks of CO2
6. Surge Pricing Model - Implement surge pricing during peak demand times to incentivize casual riders to opt for more cost-effective annual memberships, thereby maximizing profitability and promoting growth for Cyclistic.
7. A word of caution can be that the evidence is inconclusive, There may be more than meets the eye, Moving forward as is with a conversion marketing strategy is risky.