Google Data Analytics Capstone: Cyclistic Case Study

Course: Google Data Analytics Capstone: Complete a Case Study

Introduction

I've been learning about data analysis with the Google Data Analytics Certificate courses on Coursera. This project is the final result of my efforts. I'll be using Excel, SQL, Big Query, and Tableau to explore and compare ride-share patterns between annual members and casual riders. I am Bushra Mihdawi, and this project is a testament to my newfound skills in data analysis.

Links to code and visuals will be posted at the bottom of the page, but I will include them here for convenience as well.

- SQL Code Link: https://github.com/Bushra-YM/Capstone-Google-Data-Analytics-Project
- Tableau Visual:

Background

Cyclistic operates a bike-sharing program with over 5,800 bikes and 600 docking stations, providing inclusive options like reclining bikes and hand tricycles. Established in 2016, the program has expanded to 692 tracked stations across Chicago, allowing users to unlock and return bikes at any station.

Historically, Cyclistic focused on broad consumer awareness, offering various pricing plans such as single-ride passes, full-day passes, and annual memberships. Casual riders, opting for single-ride or full-day passes, coexist with Cyclistic members, who subscribe annually.

My Role

As a junior data analyst at Cyclistic, my team is tasked with developing marketing strategies targeting the conversion of casual riders into annual members.

Overall Goal

Devise effective marketing strategies to encourage casual riders to become Cyclistic annual members.

Business Questions

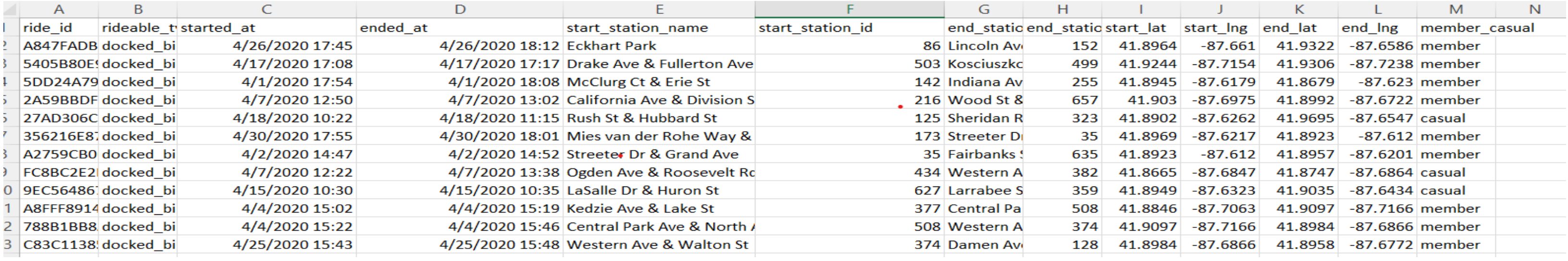
"How do the behaviors of annual members and casual riders differ in their use of Cyclistic bikes?" The following sections outline the step-by-step approach employed in addressing this project.

Ask: Business Task

I'll examine data on how current Cyclistic bike users use the bikes, looking for patterns like usage frequency and other trends. This analysis aims to understand the general habits of two customer groups: annual members and casual riders.

Prepare: Data Source Description

The information used to address the business question was sourced directly from Cyclistic's internal metrics, capturing data from its stations and bikes. The dataset spans a 5-month duration, from January 2023 to December 2023, organized into individual files for each month. Each table includes 13 columns, featuring ride IDs, bike types, ride start and end times, station identifiers (including name and location), and the rider's membership type.



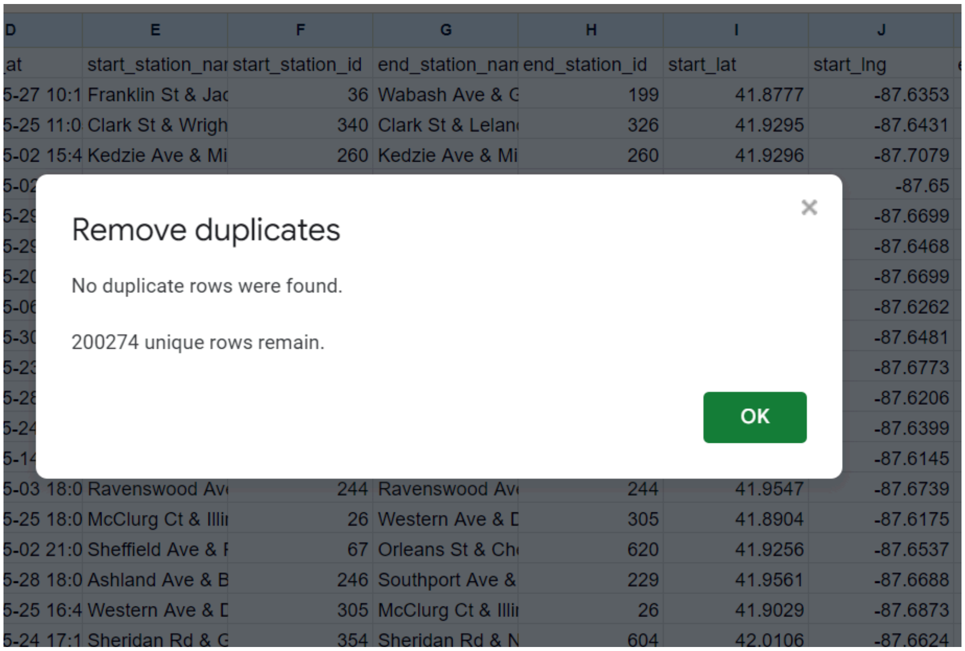
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ride_id	rideable_t	started_at	ended_at	start_station_name	start_station_id	end_static	end_static	start_lat	start_lng	end_lat	end_lng	member_casual	
2	A847FADB	docked_bi	4/26/2020 17:45	4/26/2020 18:12	Eckhart Park	86	Lincoln Av	152	41.8964	-87.661	41.9322	-87.6586	member	
3	5405B80E	docked_bi	4/17/2020 17:08	4/17/2020 17:17	Drake Ave & Fullerton Ave	503	Kosciuszko	499	41.9244	-87.7154	41.9306	-87.7238	member	
4	5DD24A79	docked_bi	4/1/2020 17:54	4/1/2020 18:08	McClurg Ct & Erie St	142	Indiana Av	255	41.8945	-87.6179	41.8679	-87.623	member	
5	2A59BBDF	docked_bi	4/7/2020 12:50	4/7/2020 13:02	California Ave & Division S	216	Wood St &	657	41.903	-87.6975	41.8992	-87.6722	member	
6	27AD306C	docked_bi	4/18/2020 10:22	4/18/2020 11:15	Rush St & Hubbard St	125	Sheridan R	323	41.8902	-87.6262	41.9695	-87.6547	casual	
7	356216E8	docked_bi	4/30/2020 17:55	4/30/2020 18:01	Mies van der Rohe Way &	173	Streeter D	35	41.8969	-87.6217	41.8923	-87.612	member	
8	A2759CB0	docked_bi	4/2/2020 14:47	4/2/2020 14:52	Streeter Dr & Grand Ave	35	Fairbanks	635	41.8923	-87.612	41.8957	-87.6201	member	
9	FC8BC2E2	docked_bi	4/7/2020 12:22	4/7/2020 13:38	Ogden Ave & Roosevelt Rc	434	Western A	382	41.8665	-87.6847	41.8747	-87.6864	casual	
10	9EC56486	docked_bi	4/15/2020 10:30	4/15/2020 10:35	LaSalle Dr & Huron St	627	Larrabee S	359	41.8949	-87.6323	41.9035	-87.6434	casual	
11	A8FFF8914	docked_bi	4/4/2020 15:02	4/4/2020 15:19	Kedzie Ave & Lake St	377	Central Pa	508	41.8846	-87.7063	41.9097	-87.7166	member	
12	788B1BB8	docked_bi	4/4/2020 15:22	4/4/2020 15:46	Central Park Ave & North	508	Western A	374	41.9097	-87.7166	41.8984	-87.6866	member	
13	C83C1138	docked_bi	4/25/2020 15:43	4/25/2020 15:48	Western Ave & Walton St	374	Damen Av	128	41.8984	-87.6866	41.8958	-87.6772	member	

The data is generally unbiased, except for the absence of start and end location data for numerous rides. After a thorough inquiry, it was determined that information for these specific stations was unavailable. I downloaded each monthly CSV file and converted them into .xlsx files for utilization in Excel.

Process: Cleaning the Data

After acquiring the data, I examined each file separately in Excel to understand its structure. Given the extensive size of the data, I opted to clean each month's information independently in Excel before attempting any data merging. The following outlines the cleaning process:

- Examined all rows in each file for duplicates using Excel's built-in feature, and no duplicates were identified.



Remove duplicates

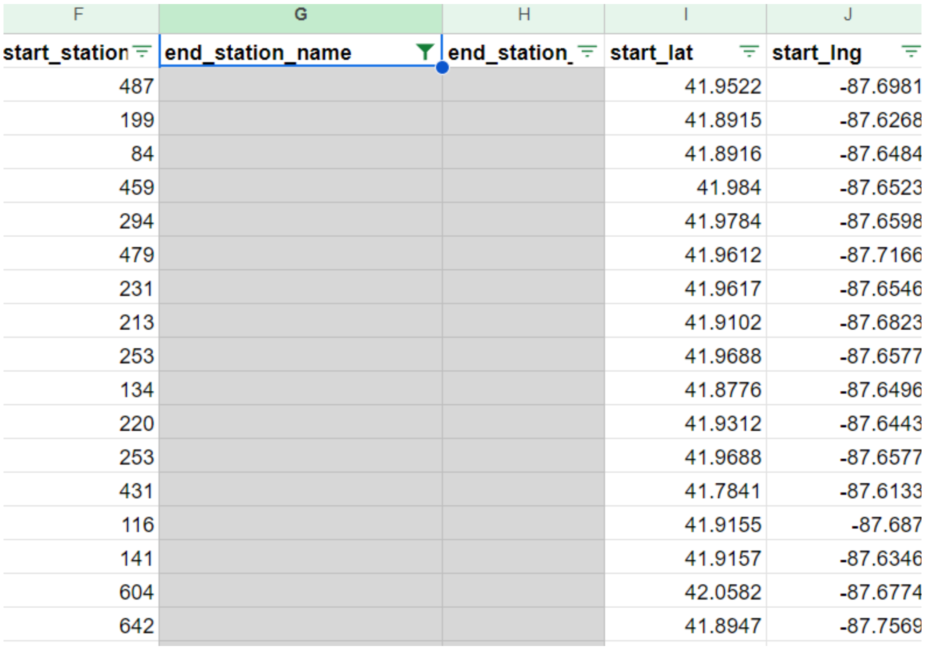
No duplicate rows were found.

200274 unique rows remain.

D	E	F	G	H	I	J
at	start_station_nar	start_station_id	end_station_nam	end_station_id	start_lat	start_lng
5-27 10:1	Franklin St & Jac	36	Wabash Ave & G	199	41.8777	-87.6353
5-25 11:0	Clark St & Wrigh	340	Clark St & Lelan	326	41.9295	-87.6431
5-02 15:4	Kedzie Ave & Mi	260	Kedzie Ave & Mi	260	41.9296	-87.7079
5-02						-87.65
5-29						-87.6699
5-29						-87.6468
5-20						-87.6699
5-06						-87.6262
5-30						-87.6481
5-23						-87.6773
5-28						-87.6206
5-24						-87.6399
5-14						-87.6145
5-03 18:0	Ravenswood Av	244	Ravenswood Av	244	41.9547	-87.6739
5-25 18:0	McClurg Ct & Illi	26	Western Ave & D	305	41.8904	-87.6175
5-02 21:0	Sheffield Ave & F	67	Orleans St & Ch	620	41.9256	-87.6537
5-28 18:0	Ashland Ave & B	246	Southport Ave &	229	41.9561	-87.6688
5-25 16:4	Western Ave & D	305	McClurg Ct & Illi	26	41.9029	-87.6873
5-24 17:1	Sheridan Rd & G	354	Sheridan Rd & N	604	42.0106	-87.6624

- Example of duplicate check results for May data

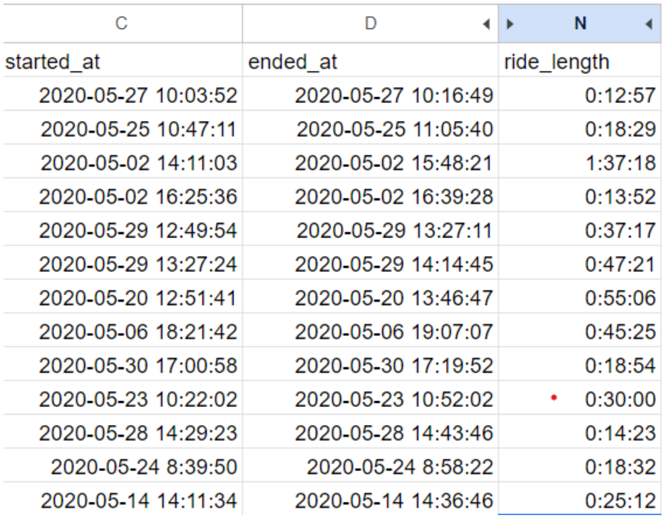
- I employed conditional formatting to highlight blank cells in gray, facilitating the identification of missing data. Subsequently, rows with incomplete information for both start and end station locations in several trips were excluded from the dataset.



F	G	H	I	J
start_station	end_station_name	end_station_	start_lat	start_lng
487			41.9522	-87.6981
199			41.8915	-87.6268
84			41.8916	-87.6484
459			41.984	-87.6523
294			41.9784	-87.6598
479			41.9612	-87.7166
231			41.9617	-87.6546
213			41.9102	-87.6823
253			41.9688	-87.6577
134			41.8776	-87.6496
220			41.9312	-87.6443
253			41.9688	-87.6577
431			41.7841	-87.6133
116			41.9155	-87.687
141			41.9157	-87.6346
604			42.0582	-87.6774
642			41.8947	-87.7569

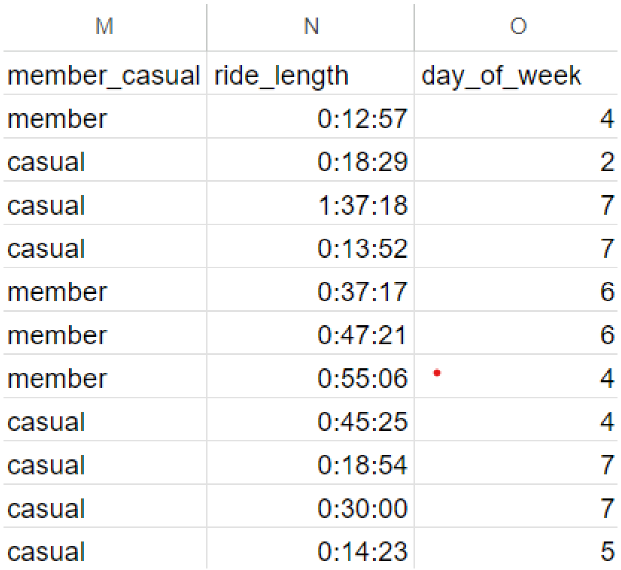
Empty fields were identified for station names.

- Established a new column called "ride_length" to determine the duration of each bike ride in minutes. This involved calculating the time difference between the end and start times.



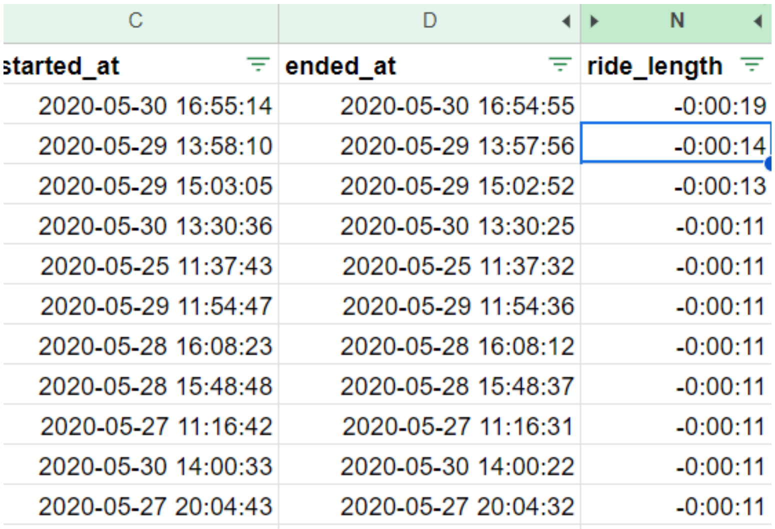
C	D	N
started_at	ended_at	ride_length
2020-05-27 10:03:52	2020-05-27 10:16:49	0:12:57
2020-05-25 10:47:11	2020-05-25 11:05:40	0:18:29
2020-05-02 14:11:03	2020-05-02 15:48:21	1:37:18
2020-05-02 16:25:36	2020-05-02 16:39:28	0:13:52
2020-05-29 12:49:54	2020-05-29 13:27:11	0:37:17
2020-05-29 13:27:24	2020-05-29 14:14:45	0:47:21
2020-05-20 12:51:41	2020-05-20 13:46:47	0:55:06
2020-05-06 18:21:42	2020-05-06 19:07:07	0:45:25
2020-05-30 17:00:58	2020-05-30 17:19:52	0:18:54
2020-05-23 10:22:02	2020-05-23 10:52:02	0:30:00
2020-05-28 14:29:23	2020-05-28 14:43:46	0:14:23
2020-05-24 8:39:50	2020-05-24 8:58:22	0:18:32
2020-05-14 14:11:34	2020-05-14 14:36:46	0:25:12

- Established a new column called "day_of_week" to extract the day name from each start time and saved it as a numerical representation ranging from 1 to 7.



M	N	O
member_casual	ride_length	day_of_week
member	0:12:57	4
casual	0:18:29	2
casual	1:37:18	7
casual	0:13:52	7
member	0:37:17	6
member	0:47:21	6
member	0:55:06	4
casual	0:45:25	4
casual	0:18:54	7
casual	0:30:00	7
casual	0:14:23	5

- I reviewed the maximum and minimum values in all rows to identify outliers. During this examination, I found negative times, and it became evident that Daylight Saving Time (DST) had affected the recorded ride lengths. To address this issue, I introduced a new formula specifically for these rides to rectify the ride length data.



C	D	N
started_at	ended_at	ride_length
2020-05-30 16:55:14	2020-05-30 16:54:55	-0:00:19
2020-05-29 13:58:10	2020-05-29 13:57:56	-0:00:14
2020-05-29 15:03:05	2020-05-29 15:02:52	-0:00:13
2020-05-30 13:30:36	2020-05-30 13:30:25	-0:00:11
2020-05-25 11:37:43	2020-05-25 11:37:32	-0:00:11
2020-05-29 11:54:47	2020-05-29 11:54:36	-0:00:11
2020-05-28 16:08:23	2020-05-28 16:08:12	-0:00:11
2020-05-28 15:48:48	2020-05-28 15:48:37	-0:00:11
2020-05-27 11:16:42	2020-05-27 11:16:31	-0:00:11
2020-05-30 14:00:33	2020-05-30 14:00:22	-0:00:11
2020-05-27 20:04:43	2020-05-27 20:04:32	-0:00:11

Negative ride times caused by DST time change

- Identified and eliminated rows with zero-minute ride times. Subsequently, removed any remaining negative ride times from the dataset.
- Conducted a count of distinct ride IDs to ensure uniqueness, and verified that the calculated count aligned with the number of rows in each dataset.

#Note:

Please be aware that additional cleaning tasks were carried out later in the process to address unexpected issues. Refer to the SQL code in the next section to observe any cleaning actions taken on the combined data.

Analyze and Share: Analysis of Data Through Big Query and Tableau

We use BigQuery to bring together and clean up different datasets. This is because Microsoft Excel can only handle up to 1,048,576 rows, and since the Cyclistic dataset has over 5.6 million rows, we need a platform like BigQuery that can handle such large amounts of data.

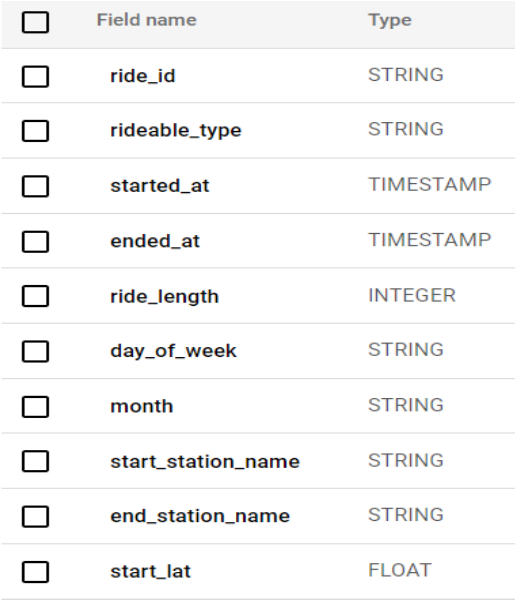
Combining the Data

SQL Query:

Tables for 12 CSV files have been added to the '2023_tripdata' dataset. Furthermore, a table named "combined_data" has been established, encompassing a total of 5,719,877 rows of data representing the entire year.

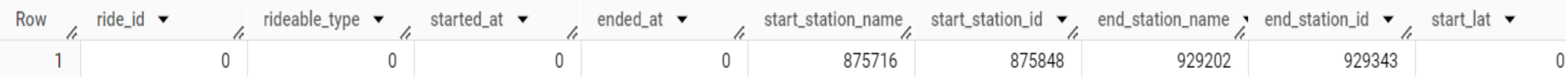
Findings:

1- The following table displays the names of all columns along with their corresponding data types.



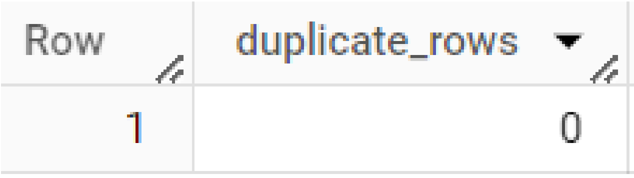
□	Field name	Type
□	ride_id	STRING
□	rideable_type	STRING
□	started_at	TIMESTAMP
□	ended_at	TIMESTAMP
□	ride_length	INTEGER
□	day_of_week	STRING
□	month	STRING
□	start_station_name	STRING
□	end_station_name	STRING
□	start_lat	FLOAT

2- The table below illustrates the count of null values in each column.



Row	ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat
1	0	0	0	0	875716	875848	929202	929343	0

3- Since ride_id has no null values, we can utilize it to identify and check for duplicates.



Row	duplicate_rows
1	0

4- The columns started_at and ended_at display the trip's start and end times in the YYYY-MM-DD hh:mm:ss UTC format. To determine the overall trip duration, a new column named ride_length can be generated.

Data Cleaning

SQL Query:

1- Rows containing missing values have been removed.

2- Trips with a duration of less than a minute or longer than a day have been excluded.

3- During this step, 1,476,445 rows in total have been removed.

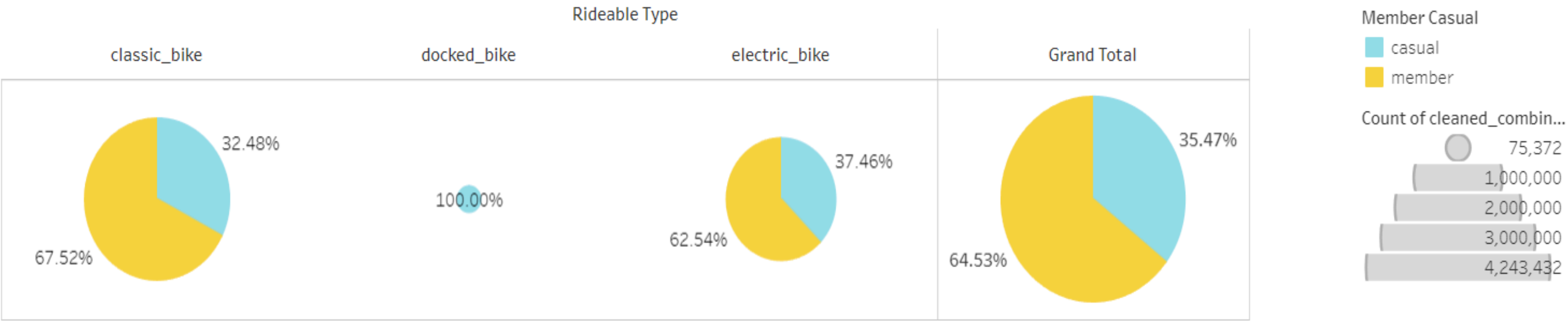
Analyze and Share

SQL Query:

Data Visualization:

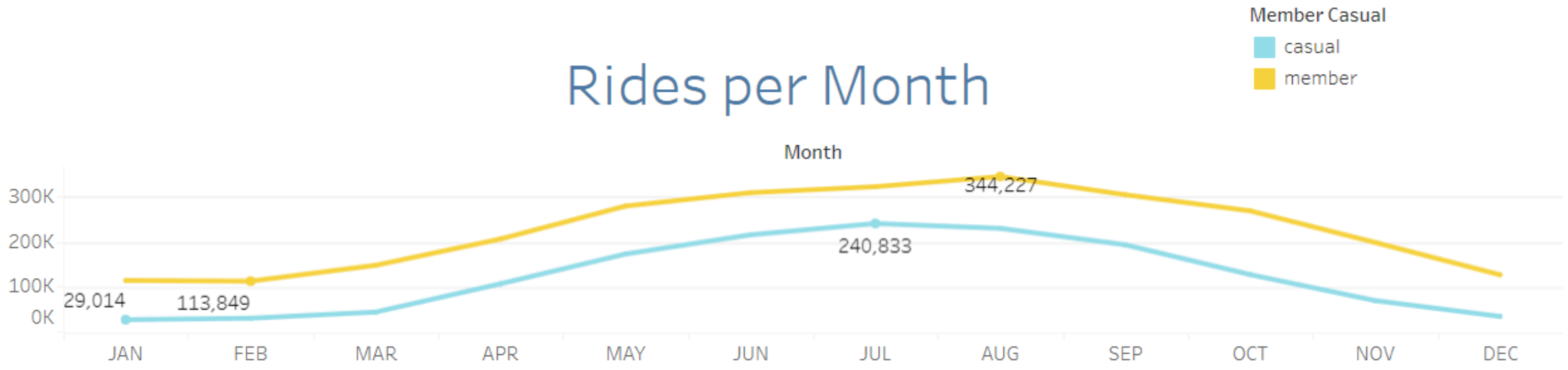
The data has been organized and is now ready for analysis. I performed queries on various pertinent tables and visualized the results using Tableau.

The initial analysis involves comparing member and casual riders, focusing on the types of bikes they use.

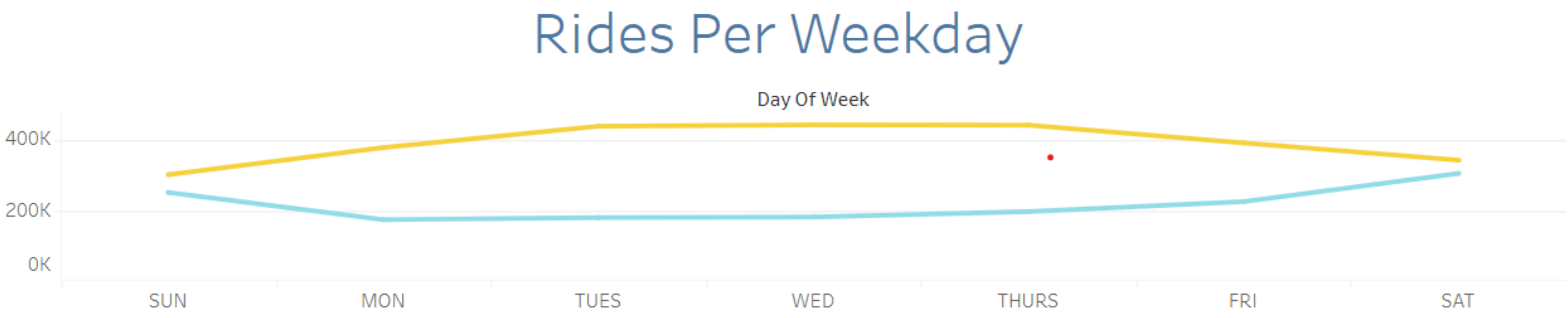


More than half (64.53%) of the riders are members, and the rest (35.47%) are casual riders. The percentage charts for each bike type show how much they are used compared to the total. Classic bikes are the most popular, followed by electric bikes. Docked bikes are used the least, mostly by casual riders.

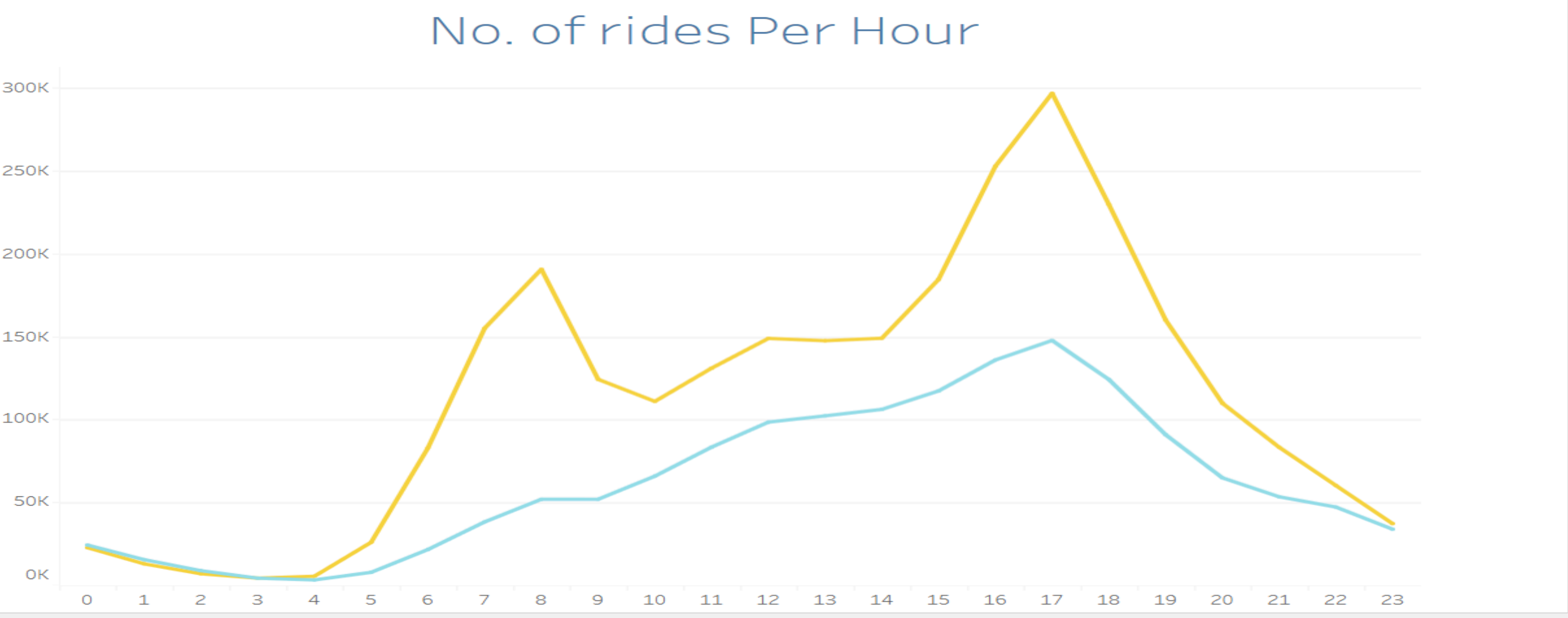
Next, we analyze the distribution of trips based on months and days of the week.



Rides per Month



Rides Per Weekday



No. of rides Per Hour

Monthly: In terms of monthly trips, both casual riders and members demonstrate similar trends, with increased trips in spring and summer and reduced activity in winter. The distinction between casual riders and members is minimal, particularly during July in the summer season.

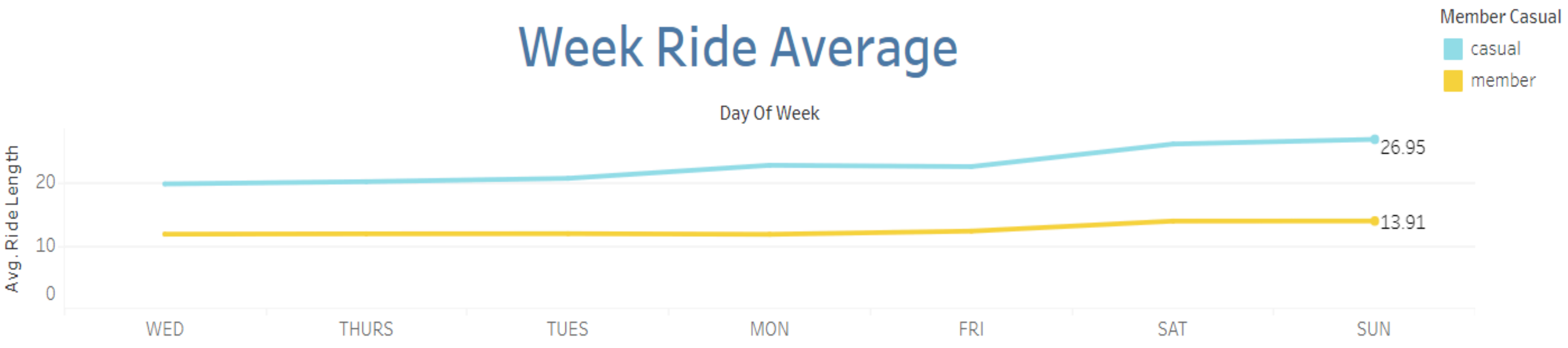
Weekly: When examining days of the week, it's observed that casual riders increase their trips on weekends, while members, although experiencing a decline over the weekend, still make the most journeys compared to other days of the week.

Hourly: The data indicates two notable peaks in the number of trips for members: one during the early morning hours (around 6 am to 8 am) and another in the evening (around 4 pm to 8 pm). In contrast, casual riders display a gradual increase in trip numbers throughout the day, peaking in the evening and tapering off later.

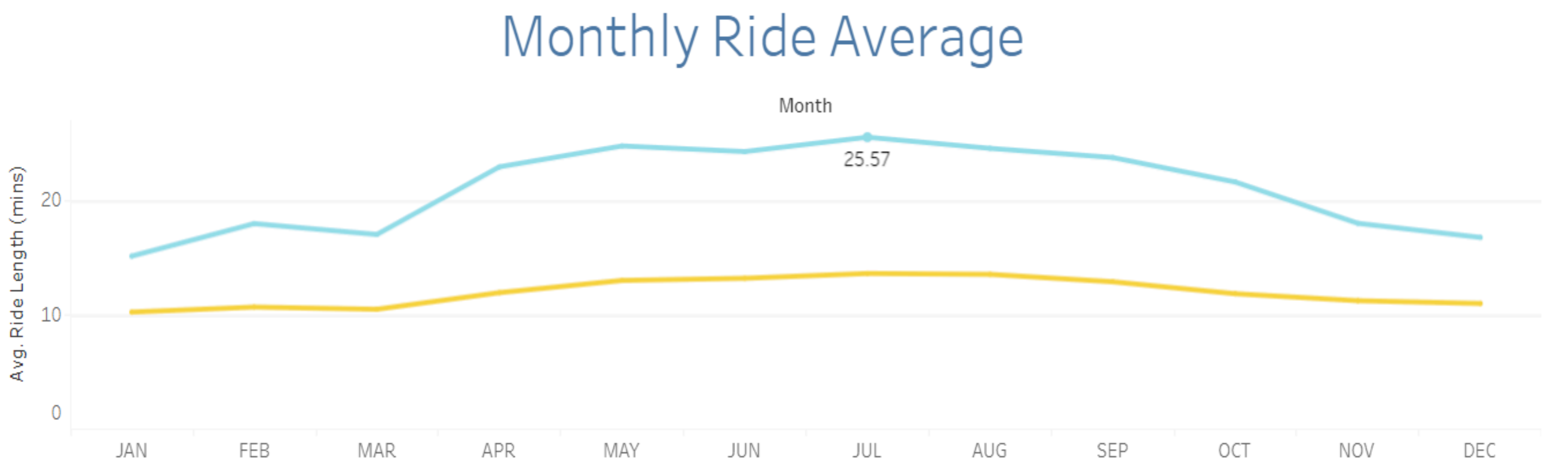
From these observations, we can infer that members likely use bikes for commuting to and from work on weekdays, while casual riders prefer bikes consistently throughout the day, with a higher usage frequency on weekends for leisure. Both groups exhibit heightened activity during the summer and spring seasons.

To delve deeper into the distinctions in behavior between casual and member riders, an analysis of the durations of their respective trips can provide valuable insights.

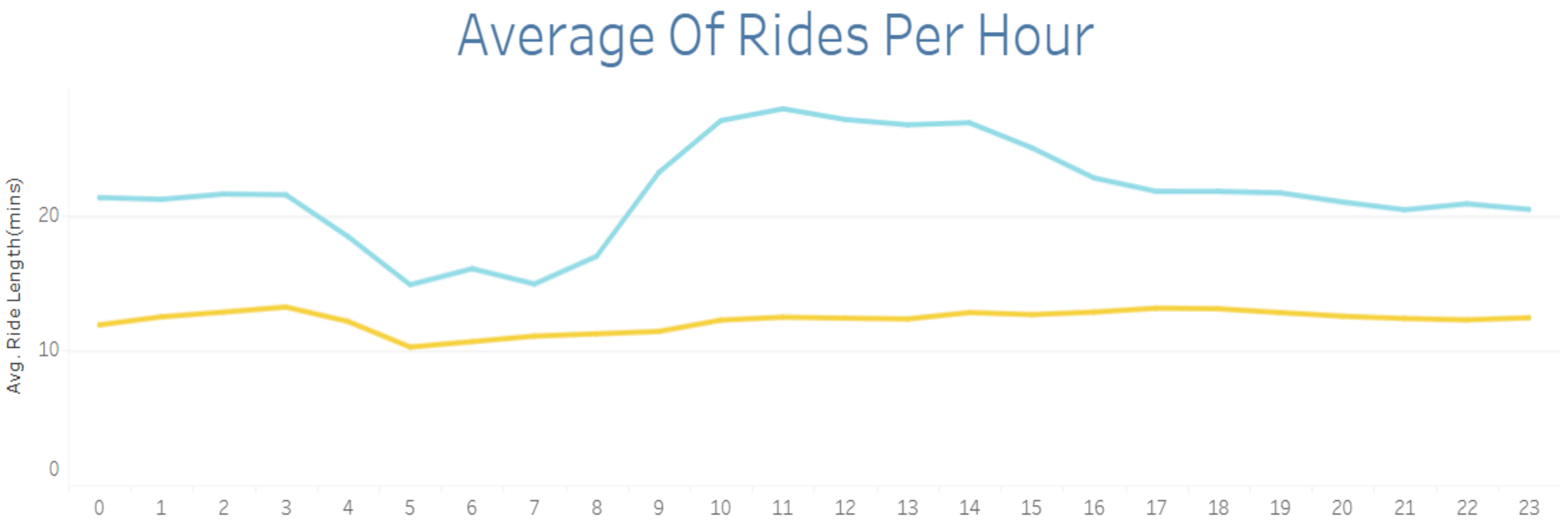
Average Ride Durations in 2023 Trips



Week Ride Average



Monthly Ride Average



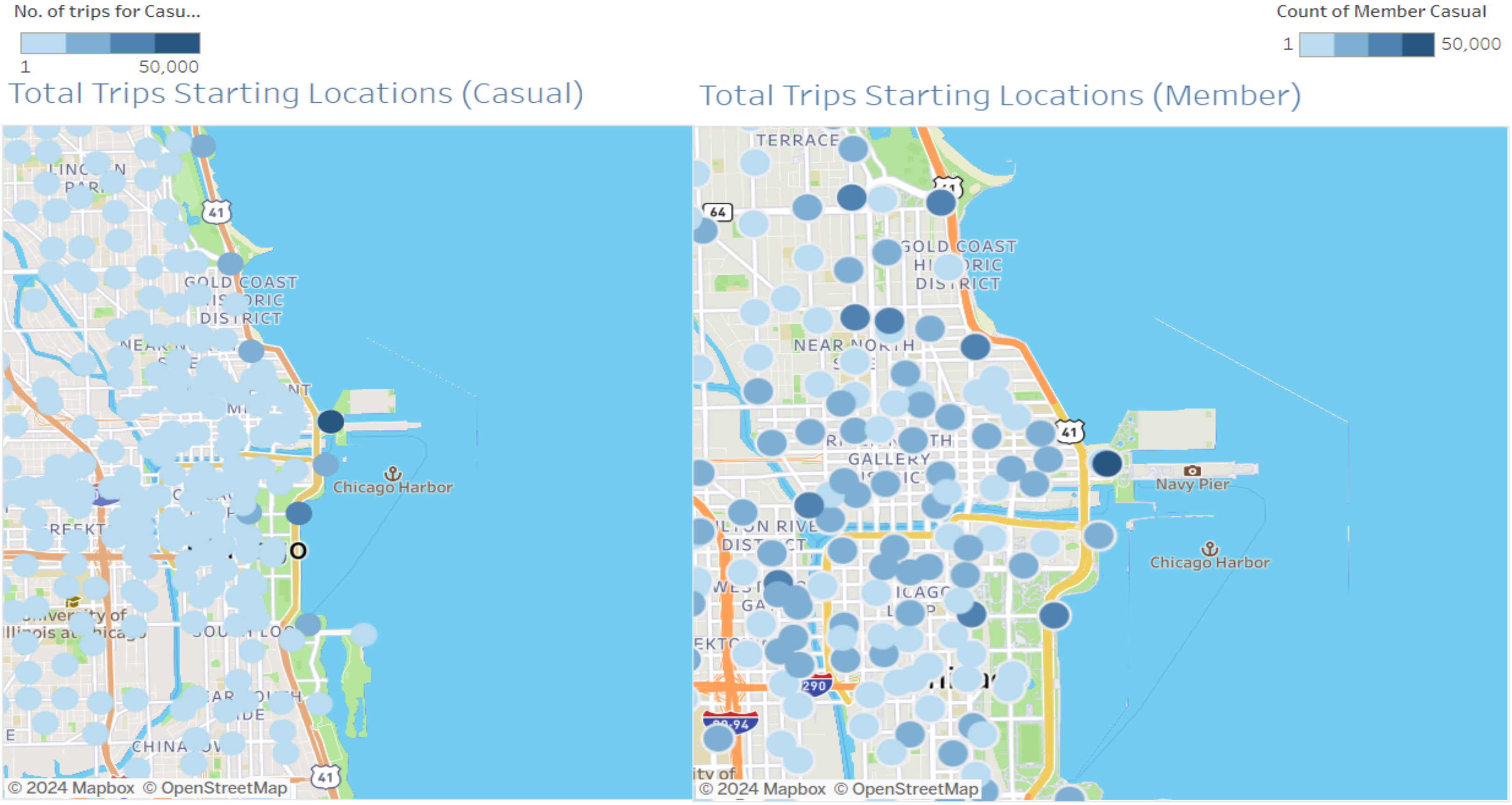
Average Of Rides Per Hour

It's noteworthy that casual riders, on average, have longer cycling durations compared to members. The average journey length for members remains consistent throughout the year, week, and day. In contrast, casual riders exhibit variations in their cycling durations. They cover greater distances in the spring and summer, on weekends, and between 10 am to 2 pm during the day. Additionally, their trips tend to be brief between five and eight in the morning.

These findings lead to the conclusion that casual commuters cover longer distances (approximately twice as much) but with less frequency compared to members. Their longer journeys on weekends, during the day outside of commuting hours, and in the spring and summer seasons suggest a recreational purpose behind their cycling activities.

To gain a deeper understanding of the distinctions between casual and member riders, an analysis of the starting and ending station locations can provide valuable insights. By applying filters to identify stations with the highest trip frequencies, we can draw conclusions based on the following observations.

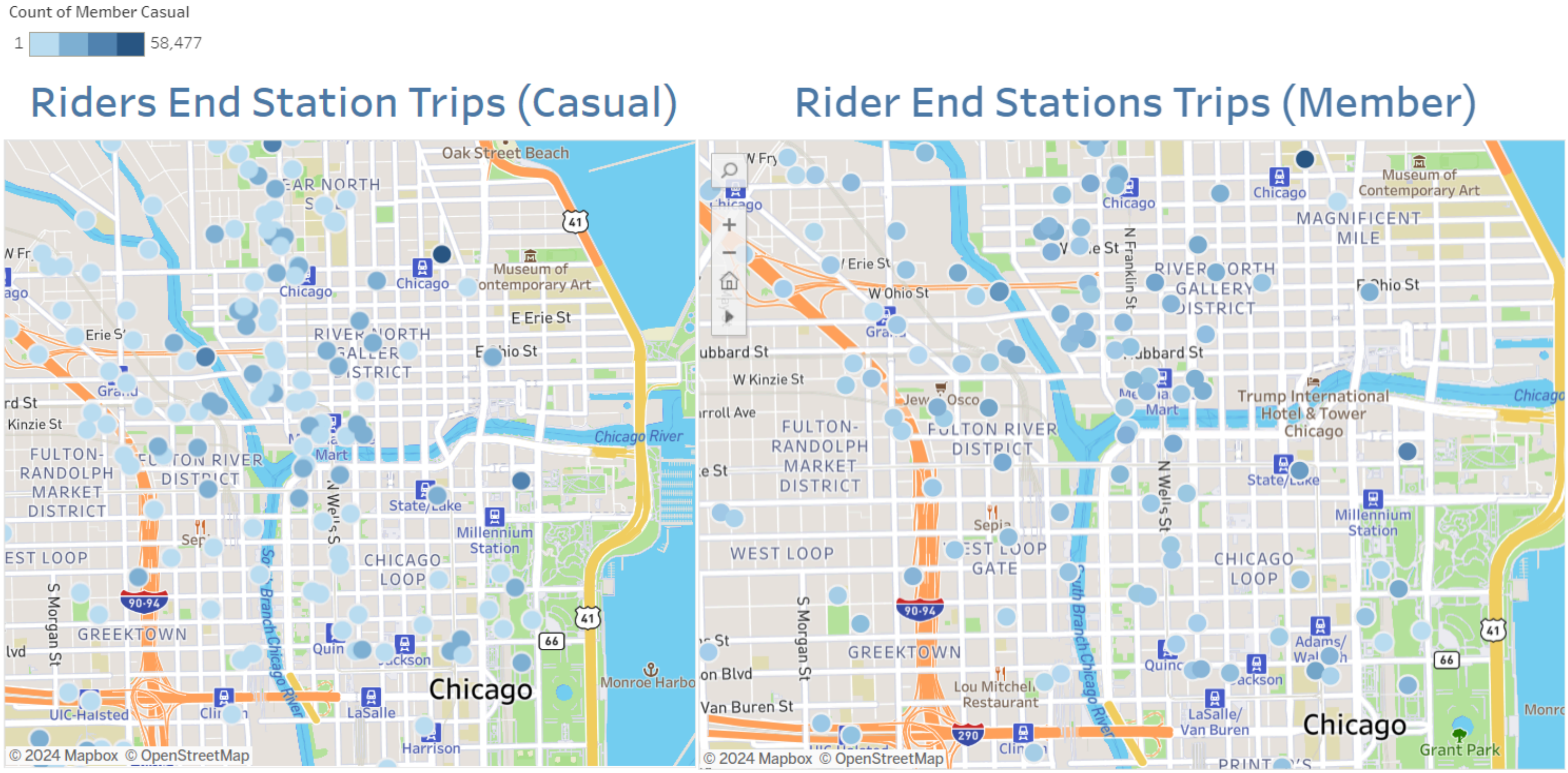
Total Trips Start Locations in 2023



Total Trips Starting Locations (Casual)

Total Trips Starting Locations (Member)

Total Trips End Locations in 2023



Riders End Station Trips (Casual)

Rider End Stations Trips (Member)

Casual riders frequently choose to start their trips from stations situated near museums, parks, beaches, harbor points, and aquariums. Notably, the Chicago Harbor stands out as a prominently attended location for commencing trips, drawing considerable attention from casual riders as a preferred starting point for their journeys. In contrast, members tend to initiate their journeys from stations near universities, residential areas, restaurants, hospitals, grocery stores, theaters, schools, banks, factories, train stations, parks, and plazas. This distinction in starting station locations suggests diverse preferences and purposes for bike usage between casual and member riders.

When it comes to where riders finish their trips, a similar pattern emerges. Casual riders often end their journeys near parks, museums, and other fun places, showing a preference for leisure. On the other hand, members finish their trips near universities, residential areas, and business areas, suggesting a trend of using bikes for everyday commuting. This difference in where they end their trips supports the idea that casual riders mainly use bikes for fun, while members rely on them a lot for their daily commutes.

Summary

Casual	Casual & Member	Member
1- Casual riders engage in longer, less frequent rides, indicating a preference for leisure and recreational activities.	Both members and casual riders demonstrate a higher demand for bikes during the middle of the week. However, casual riders exhibit more activity throughout the midday hours, while members predominantly utilize bikes during typical work commute times.	1- Members engage in more frequent rides, but the duration of their trips is shorter, approximately half of the trip duration observed in casual riders.
2- They tend to start and end their journeys near recreational sites such as parks, museums, and along the coast.		2- The starting and ending points of members' trips are often close to universities, residential areas, and commercial zones, emphasizing their inclination toward using bikes for daily commuting purposes.
3- Casual riders prefer using bikes consistently throughout the day, with a notable increase in activity over weekends in the summer and spring.		3- Members prefer bike rides on weekdays, particularly during commute hours between 8 am and 5 pm, and this trend is more prominent in the summer and spring seasons.

Act

Once we understand how casual and member riders differ, we can create marketing plans to attract casual riders and encourage them to become members.

1- Acknowledge that casual riders tend to use their bikes for longer durations than members and consider introducing discounts for extended rides to incentivize both casual riders and members to opt for longer journeys.

2- Recognize the peak activity of casual riders on weekends and during the summer and spring, leading to the potential offering of seasonal or weekend-only memberships.

3- Explore partnership opportunities, such as bundled membership deals with other transportation modes like elevated trains or buses, to attract commuters.

4- Conduct targeted marketing campaigns during spring and summer at tourist and recreational locations popular among casual riders.

5- Reevaluate the marketing approach for students in the area, considering a more student-friendly membership aligned with the academic calendar to boost memberships among college students who rely on bikes for commuting to class.

6- Tailor marketing campaigns for classic bike riders, focusing on promoting biking opportunities lasting around 20 minutes in nearby areas like parks or bike paths.