Note: This is a short synopsis, some details and steps are omitted.

Introduction

This capstone project was completed as part of the Google Data Analytics Certificate program. It focuses on analysing user behaviour for Cyclistic, a fictional bike-share company in Chicago (based on the real company *Lyft*). The objective of the project was to leverage historical bike trip data to uncover insights on how casual riders and annual members differ in their usage patterns. Based on this analysis, the goal was to recommend data-driven marketing strategies aimed at **converting casual riders into annual members**.

Company outline

Fleet Size: Over 5,800 bicycles with more than 600 docking stations across Chicago.

USP: In addition to traditional bikes, offers reclining bikes, hand tricycles, and cargo bikes to accommodate people with additional assistance requirements.

Pricing Plans: Single-ride 'ticker' price, full-day passes, and annual memberships. (Lyft 'Divvy' pricing outline [here) .

User Base: Traditional bike users make up 92% of rides. 30% use bikes for daily commuting.

Existing Strategy: Building general awareness and appealing to broad consumer segments.

Future Strategy: Focused on increasing the number of annual members, who are more profitable than casual riders.

Initial considerations

I took care in approaching and fully defining the task to align the outcomes with the best interest of the stakeholders (the marketing team). I postulated potential actionable insights between the two groups that could be marketed to casual members from existing competitor membership marketing campaigns. Whilst these postulates guided the analysis, these did **not** influence any results - every decision was made in the interest of integrity and minimising bias – and all statistically insignificant results were also reported.

Data

- Publicly available on the Lyft website [here)
- Organized by year, quarter, and month, covering the last decade.
- Original, current, and cited (ROCCC).

I personally reviewed the licensing agreement to ensure this is fair usage. Once pulled all data was stored locally.

Variable Summary

The dataset contained detailed ride data for each bike trip, including the following variables:

Variable	Type	Description	Example Values / Range
ride_id	Identifier	Unique identifier for each ride.	Alphanumeric strings (e.g., "A12B3C4D567E").
rideable_type	Categorical	Type of bike used for the trip.	"classic_bike", "electric_bike", "docked_bike".
started_at	Datetime	Timestamp when the ride started.	e.g., "2023-11-01 12:45:30"
ended_at	Datetime	Timestamp when the ride ended.	e.g., "2023-11-01 13:15:45"
start_station_name	Categorical	Name of the station where the ride started.	e.g., "Millennium Park", "State St & Pearson St".
start_station_id	Categorical	Unique identifier for the start station.	e.g., 113
end_station_name	Categorical	Name of the station where the ride ended.	e.g., "Millennium Park", "Broadway & Sheridan Rd".
end_station_id	Categorical	Unique identifier for the end station.	e.g., 567
start_lat	Numerical (float)	Latitude of the start station or location (in decimal degrees).	e.g., 41.8811015 (Chicago coordinates)
start_lng	Numerical (float)	Longitude of the start station or location (in decimal degrees).	e.g., -87.62408183333334 (Chicago coordinates)
end_lat	Numerical (float)	Latitude of the end station or location (in decimal degrees).	e.g., 41.949422717
end_lng	Numerical (float)	Longitude of the end station or location (in decimal degrees).	e.g., -87.646384716
member_casual	Categorical	Membership type of the rider: whether the user is a casual rider or a member.	"member", "casual".

Survey data and customer ID numbers (for instance proxied by hashed billing information or home addresses) were not available on request.

Processing

With the integrity and relevance of the data verified, I pulled the twelve most recent months of data and joined the dataset locally in MySQL. Twelve months of data was used as this allowed me to control for seasonal, holiday and tourist trends during the analysis. Data back to the company's inauguration in 2016 was available, however the level of general awareness of the bike-share service is likely to have steadily increased year on year. As we are recommending future marketing strategies to be implemented in the existing level of consumer awareness, data prior to the previous twelve months is deemed to offer less insight in this context.

Pre-Cleaning

All entries matched the variable type. 102,004 rows were duplicates

Variable	Pre-cleaning notes (in some cases nulls were blankspace)
ride_id	All 16 characters distinct alphanumeric. No Nulls - **Action:** None
rideable_type	There are three types of rideable_type: electric_bike, classic_bike, and docked_bike. Nulls present - **Action**: Investigate the null entries and assess if imputing or categorizing them as "unknown" is feasible
started_at and ended_at	Some durations infeasible e.g. negative, < 1 minute (with different start/stop locations), over 1 day. Nulls present - **Action**: Remove entries where duration is negative or where trip duration exceeds 1 day (unless further investigation justifies keeping these). - Analyse whether these anomalies are tied to specific stations, times of day, or bike types to identify potential systemic errors (e.g., system outages, misreported times).
start_station_name, end_station_name, and start_station_id	• There are 1,552 unique start stations, with "null" being the most frequent. 1,579 end stations, with "null" also as the most frequent. • Reviewed and searched for names containing "priv" "test" "demo" and "temp" or no letters, none returned. However, inspecting end_station NOT IN start_station returned "Base - 2132 W Hubbard" which is the Divvy warehouse. • All station names and station IDs matched. • Nulls Present, however while classic/docked bikes must start and end at docking stations, electric bikes can lock up near docking stations, meaning trips may not always start or end at a station. - **Action** For trips with "null" stations, investigate if the rideable_type is an electric bike (which could explain lack of docking). Consider labeling

	electric bike trips with missing station data as **"on-bike lock"** if there's enough consistency.
Start_lat start_lng end_lat end_lng	All within the Chicago range (41.6-42.0° N, 87.5-87.9° W) Nulls Present - **Action**: Investigate null lat/long values. If tied to specific ride types or stations, impute or investigate. For classic/docked bikes, missing lat/longs are concern, but for electric bikes station-based docking isn't required. - Analyse if nulls are systematically associated with certain station names or specific areas to identify patterns in missing data. Could signal operational or data collection issues.
Member_casual	All columns contained 'member' or 'casual', but lots of values have some parsed ride_id's concatenated e.g. "member001FH3JI1". **- Action:** Issue with 'new line' in CSV file import in MySQL, further investigation confirms this. Fortunately cross referencing with raw ride_ids shows that a simple string parsing will fix the issue.

Orthogonality of null, duplicate and erroneous values

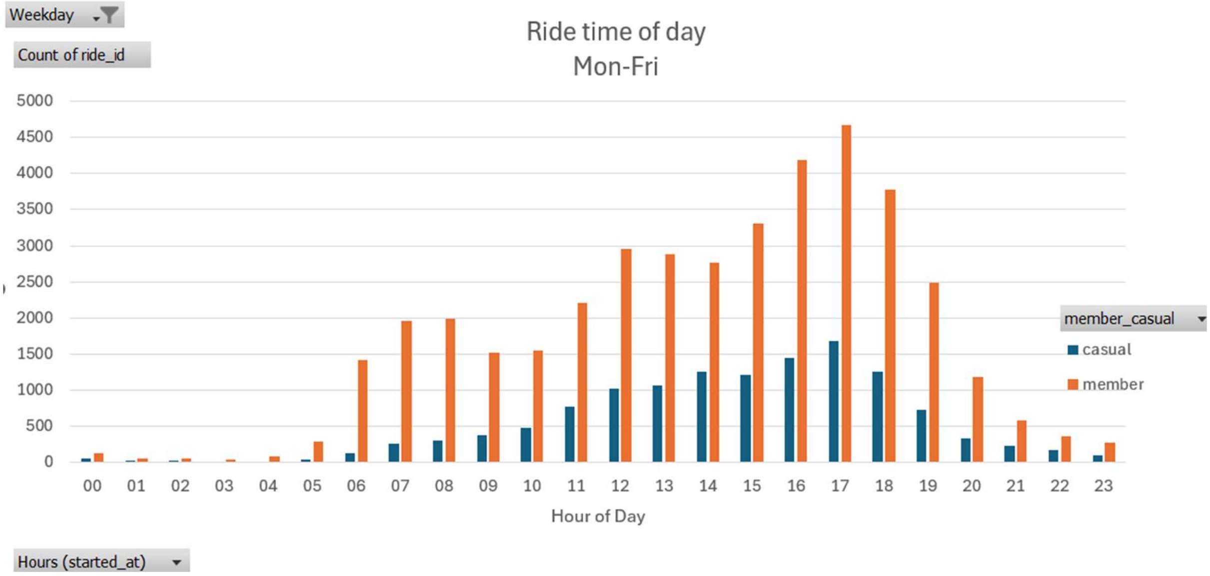
I investigated how the erroneous values were distributed/clustered across the variables in the dataset. This better informs the data cleaning decisions – e.g. deleting, interpolating, reassigning values – in preventing bias to arise in our analysis. Given the nature of the data, insightful conclusions are heavily dependent on the integrity of the assumptions. If the proportion of errors is higher in certain instances than we'd expect - e.g. for electric bikes, during certain times of day or in certain areas - this could skew any inferences about user behaviour or trip patterns.

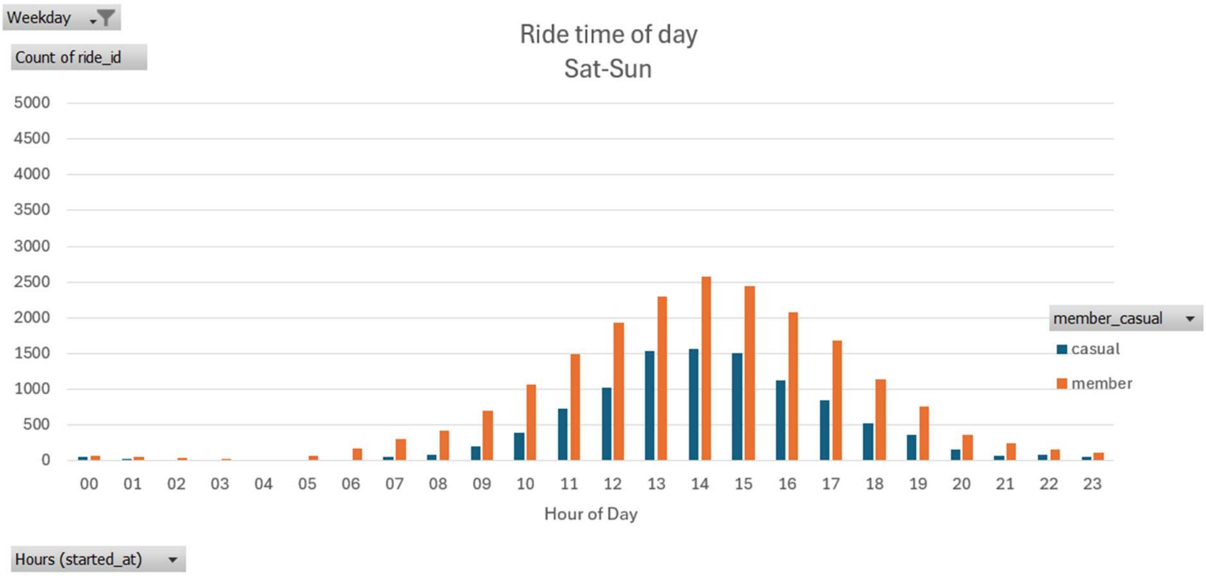
Duplicate proportion by start station, relative to all rides

start_station_name	duplicate_count	duplicate_percentage	all_rides	all_rides_percentage
Streeter Dr & Grand Ave	3974	0.98	27794	1.41
University Ave & 57th St	3888	0.96	12225	0.62
Ellis Ave & 60th St	3758	0.93	10905	0.55
Kingsbury St & Kinzie St	3092	0.77	15879	0.81
DuSable Lake Shore Dr & Monroe St	2978	0.74	18591	0.95
Clark St & Elm St	2950	0.73	14700	0.75
Clinton St & Washington Blvd	2958	0.73	14508	0.74
Clinton St & Madison St	2638	0.65	13109	0.67
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Two stations had higher than expected duplicate numbers – Uni Ave & 57th and Ellis Ave & 60th – but duplicate errors appear randomly distributed across all others. This stations will be analysed closely when analysing other nulls.

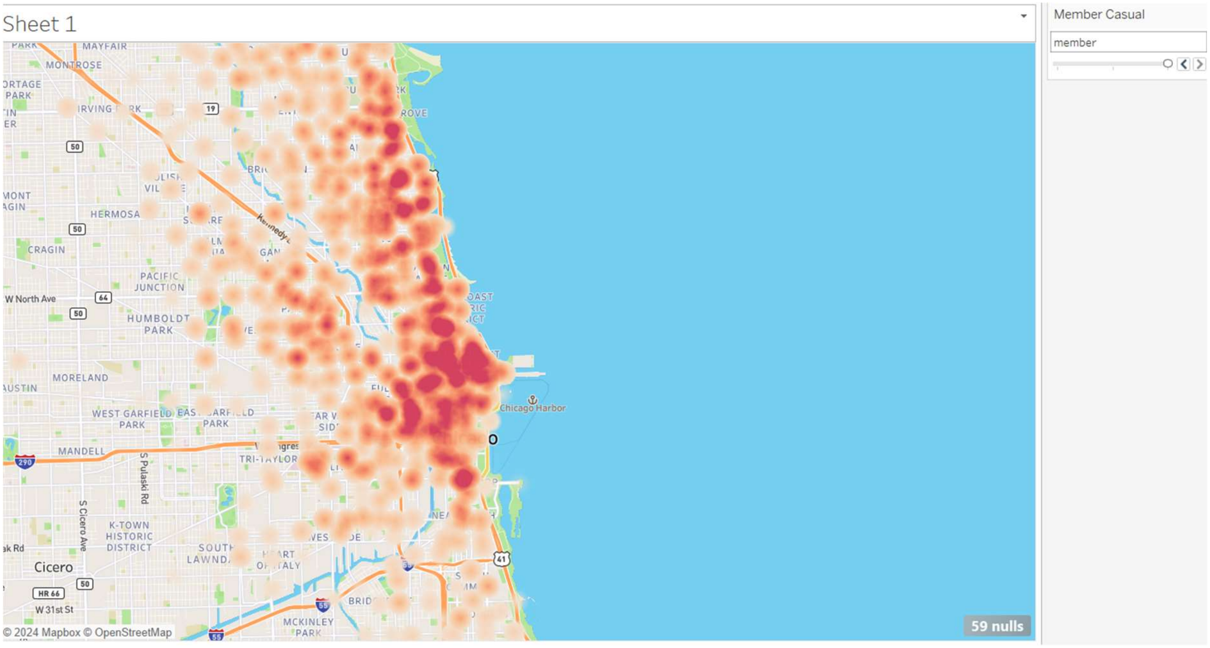
End points:



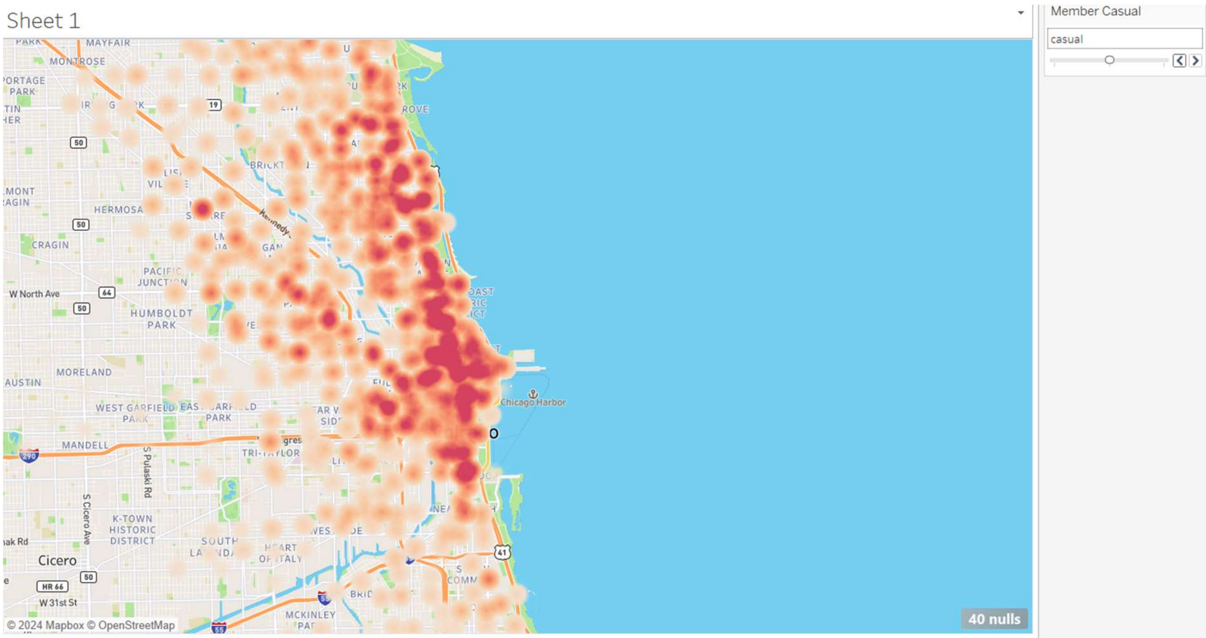


Assume normal distribution of sat-sun is a non-commuting baseline, can see the peaks in casual vs member at commute times relative to this normal curve

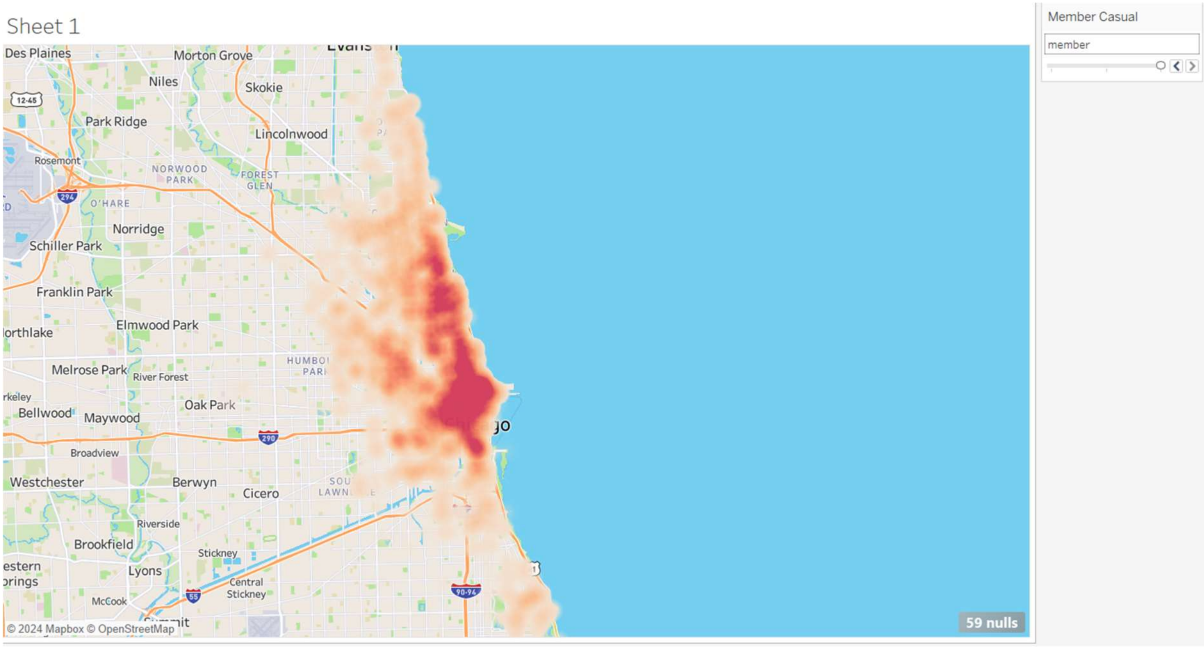
End points Member:



End points casual:



Member:



Casual: