**Note: This is a synopsis, full writeup available [here](Coming Soon)**

# 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

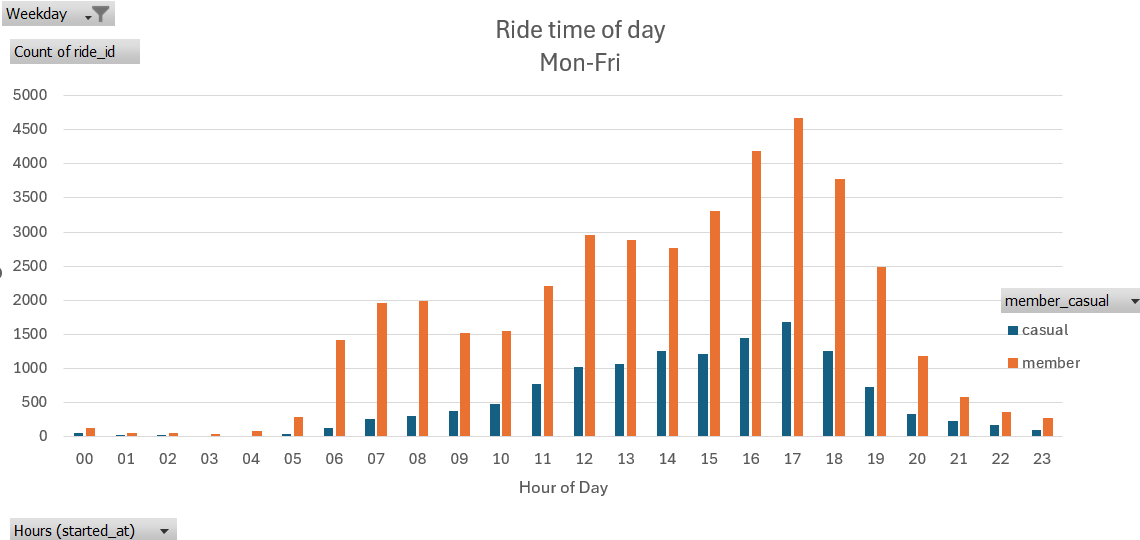
All entries matched the variable type. 102,004 entries were duplicates

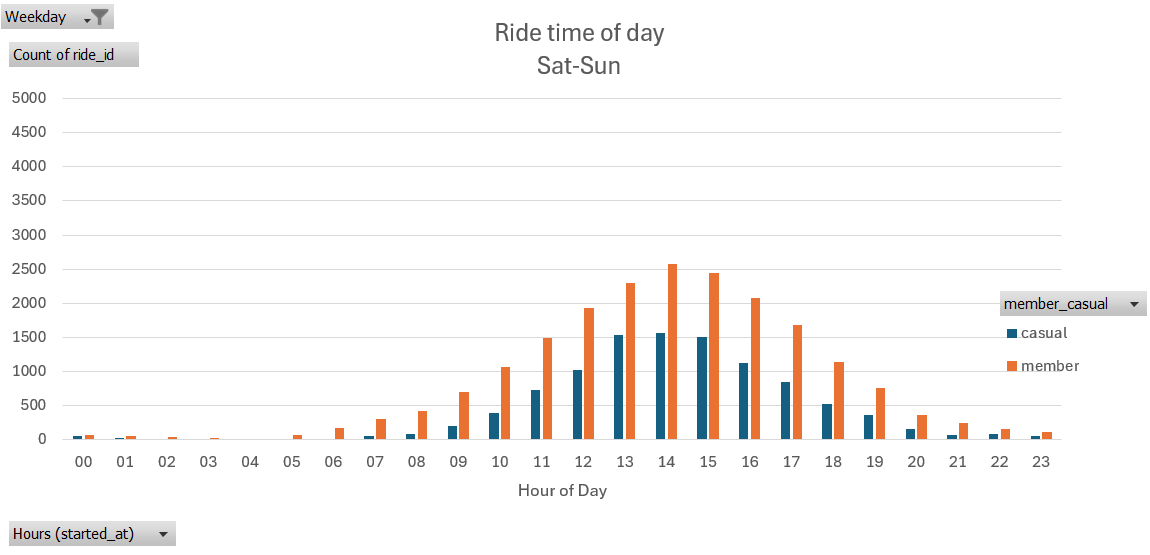
|  |  |
| --- | --- |
| Variable | Pre-cleaning notes (in some cases nulls were blankspace) |
| ride\_id | All 16 characters distinct alphanumeric.  No Nulls |
| rideable\_type | There are three types of rideable\_type: electric\_bike, classic\_bike, and docked\_bike.  Nulls present |
| started\_at and ended\_at | Some durations infeasible e.g. negative, < 1 minute (with different start/stop locations), over 1 day.  Nulls present |
| 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. |
| 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 |
| Member\_casual | All columns contained ‘member’ or ‘casual’ |

A screenshot of a computer

Description automatically generated

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:

A map of a city

Description automatically generated

End points casual:

A map of a city

Description automatically generated

Member:

A map of a city with a red spot

Description automatically generated

Casual:

A map with a red spot

Description automatically generated