Annex I

BIKE SHARE CASE STUDY

Google Data Analytics Capstone Project

By Javier Litke

July 2022

Introduction

In this case study, I represent a junior data analyst working in the marketing analyst team at a fictional bike-sharing company in Chicago named Cyclistic. The company has two types of customers:

- Customers who purchase single-ride or full-day passes are referred to as Casual Riders
- Customers who purchase Annual Memberships are Cyclistic Members.

The company's finance analysts have concluded that Annual Members are more profitable tan Casual Riders, and the director of marketing believes that there is merit in trying to convert Casual Riders to Annual Members, since they are already aware of Cyclistic's program and have already used the service.

To design a marketing campaign aimed at converting Casual Riders into Annual Members, the director of marketing wants to first obtain a better understanding on how Casual Riders and Annual Members use Cyclistic bikes differently.

The following report will be done according to Google's six data analysis phases:

1. Ask
2. Prepare
3. Process
4. Analyse
5. Share
6. Act

Annex I

Content

Google Data Analytics Capstone Project 1

Introduction 1

Phase 1: Ask 3

Business task 3

Stakeholders 3

How can the stakeholders' questions be resolved? 3

Phase 2: Prepare 3

About the data 3

Google parameters for good data 3

Data Structure 3

Phase 3: Process 4

SQL 4

PowerBI 4

Phase 4: Analyse 4

SQL 5

PowerBI 5

Phase 5: Share 10

Phase 6: Act 10

Insights 10

Limitations 10

Moving Forward 10

Annex I

Phase 1: Ask

At this stage I will define the main parameters of the analysis.

Business task

- Find how Annual Members use Cyclistic differently from Casual Riders.

Stakeholders

- The executive team, who will decide whether to approve the recommended marketing program.
- The director of marketing, who is responsible for the development of campaigns and initiatives to promote the bike-share program.

How can the stakeholders' questions be resolved?

- By comparing behaviour of Casual Riders and Members and finding common ground.
- By analysing the behaviour of Casual Riders to find patterns that would benefit from an Annual Membership.

Phase 2: Prepare

Here I will determine the source and characteristics of the data being used.

About the data

- Stored in an AWS server as .csv files. It was downloaded to my hard drive where I built a database in PostgreSQL and imported it to PowerBI.
- Licencing - Licensed and made available by Motivate International Inc. Here are links to the datasets and the data license agreement.
- Privacy – The data is already made available void of any personal identifying information
- Accessibility – Raw data will not be made available to the public.

Google parameters for good data

- Reliable – Information seems to be accurate and unbiased.
- Original – The data is generated by the original company.
- Comprehensive – For privacy reasons, we are missing data that could identify single users, so we will miss out on analysing the behaviour of individuals. The user data that is there has some data missing (gender, age of birth). The location-based information (from and to bike stations) are expressed as street intersections instead of map coordinates, so they will not be useful unless converted. This will not enter the scope of this analysis.
- Current – For the sake of this exercise, I used data from 2019, but currently data is available from January 2013 until May 2022.
- Cited – The data is 1st party generated

Data Structure

- It consists of 4 .csv files with a total of 3.818.004 rows.
- A total of 12 columns. Of those, 3 give information on the user and the rest, on the trip.

  trip_id, start_time, end_time, bike_id, trip_duration_secs, from_station_id, from_station_name, to_station_id, to_station_name, user_type, user_gender, user_birth_year

Annex I

Phase 3: Process

Here I will manipulate and clean the data to enable proper analysis.

The tools I will be using are:

PostgreSQL – with which I built a database with only one table ('trip_data') where I imported the previously downloaded .csv files. I did minor cleaning tasks and obtained an overview of the data.

MS PowerBI – with which I did the main cleaning and analysing tasks. I decided to use PowerBI directly because the data did not need intensive cleaning.

SQL

The complete list of queries can be found in Annex I.

1. Modified the column names to be more descriptive when creating the table.
2. Reformated the trip_duration_secs column to allow it to be cast as numerical
3. Found that there is missing data:

TABLE 1: MISSING DATA IN TRIPS FOR 2019

'user_gender'	559.206 values	14,6%
'user_birth_year'	538.751 values	14,1%
Overlap between the previous two values	538.749 values	96,3%

The numbers are similar, so when checking the overlap between the two it is evident that the same people who don't complete their gender data are not completing their birth year.

4. Found that 'user_birth_year' contains data that is invalid: birth years all the way back to 1759. This will be cleaned later in PowerBI.
5. Checked for duplicates using 'trip_id'. There were none.

PowerBI

1. After importing the data into PowerBi, I started by making sure that all the data types where correct
2. Changed 'user_type' values from 'Subscriber' and 'Customer' to 'Members' and 'Casual Riders' respectively, to coincide with the original nomenclature stated at the beginning of the report.
3. Truncated the 'user_birth_year' to cap the age at 100 years
4. Added columns to help with calculations and visualizations
    a. trip_duration_[min] to calculate the trip length in minutes
    b. day_number and day_name to facilitate time dependant visualizations
    c. user_age to calculate age based on year of birth

This is not an extensive list but represents the most important additions.

Phase 4: Analyse

At this stage I set out to find trends and correlations in the data

Annex I

SQL

I obtained some base numbers and calculations in SQL as a base for further analysis

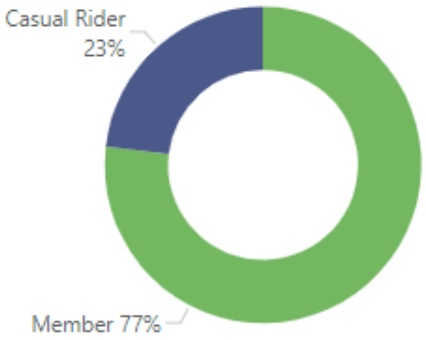
TABLE 2: INITIAL ANALYSIS DONE WITH SQL

Number of bikes		6017
Number of stations (from/to)		616/617
User Count	"Customer" "Subscriber"	880.637 2.937.367
Most used stations		"Streeter Dr & Grand Ave" "Canal St & Adams St" "Clinton St & Madison St" "Lake Shore Dr & Monroe St" "Clinton St & Washington Blvd" "Lake Shore Dr & North Blvd"
Average trip length		24 minutes
Average trip length for shorter trips (<4hs)[1]		18 minutes
Average trip length for longer trips (>4hs)		1,44 days

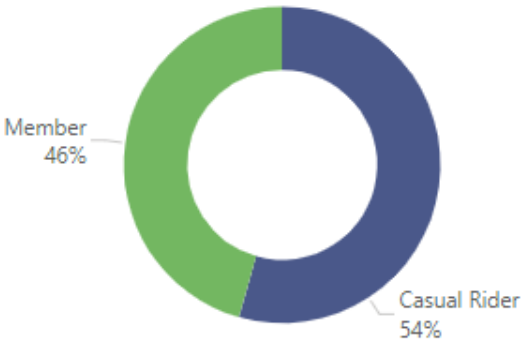
PowerBI

Comparison between Members and Casual Riders

GRAPH 1: DISTRIBUTION OF TRIPS TAKEN BY EACH USER _TYPE OVER THE TOTAL AMOUNT OF TRIPS IN 2019



GRAPH 2: DISTRIBUTION OF TRIP LENGTH BY EACH USER _TYPE OVER THE TOTAL AMOUNT OF TRIPS IN 2019



It seems that although the Annual Members take a lot more trips, the Casual Riders spend the most amount of time with the rentals.[2]

[1] It must be noted that the 'trip_duration_sec' accounts for the duration of the rental, not the real time spent riding the bikes.

[2] As users cannot be individually identified for privacy reason, it is not possible to determine weather a certain group of users make more trips or take longer rentals than others

Annex I

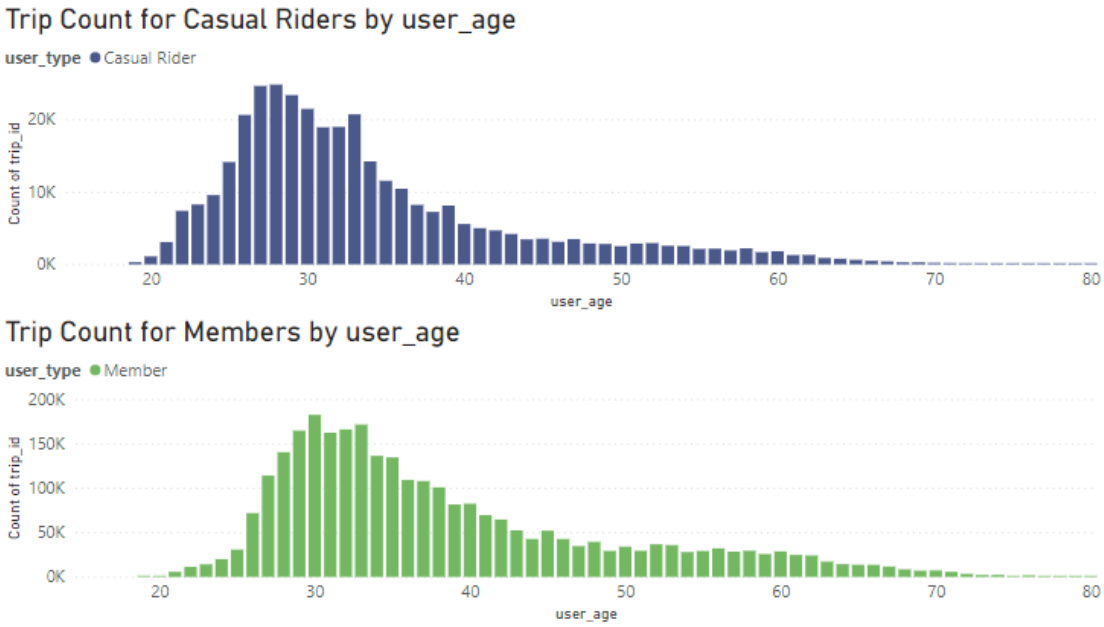
In the following table, it is made even more evident that Casual Riders consistently rent the bike for longer periods of time. The average trip for Casual Riders is larger both for shorter (<24hs) and longer (≥24hs) rentals

TABLE 3: COMPARISON OF TRIP DURATIONS BETWEEN MEMBERS AND CASUAL RIDERS

user_type	Number of Trips	Max of trip_duration_[min]	Average of trip_duration_[min]	Median of trip_duration_[min]	Mode of trip_duration_[min]	Average of trips_under_24hs_length_[min]	Average of trips_over_24hs_length_[hs]
Casual Rider	0.9M	177K	57	26	12	39	192
Member	2.9M	151K	14	10	5	13	137
Total	**3.8M**	**177K**	**24**	**12**	**5**	**19**	**177**

The age of the users are similar for both groups, and concentrated around de 20 – 40 year age group. The centre of the distribution for Casual Riders is marginally to the left of that of Members (28 vs 31 years old). [3]

GRAPH 3: TRIP COUNTS BY USER_AGE



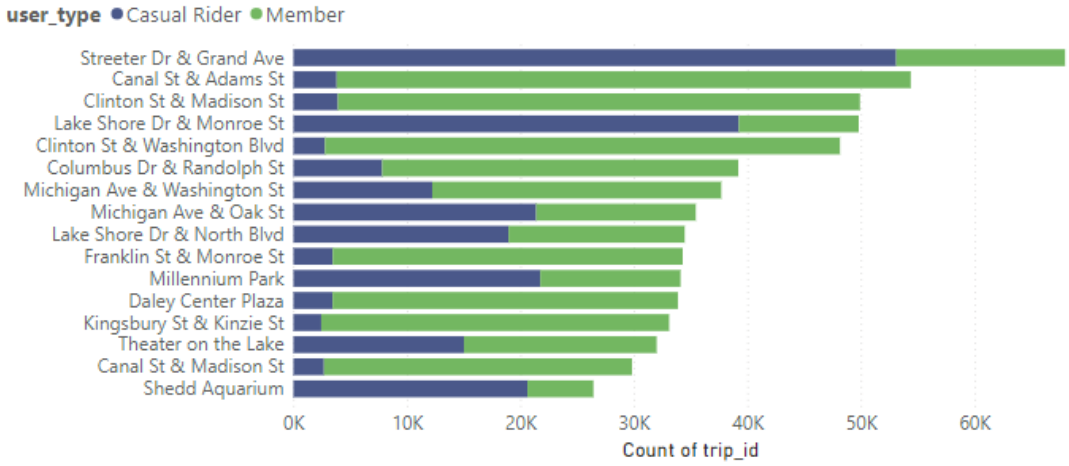
[3] Obtained from visual analysis. Statistical analysis may be useful

Annex I

Stations

When analysing the stations used, there is a clear preference by both groups for different stations, both for departing and arriving stations (the graph for arriving stations is not included for it is very similar to Graph 4). More could be analysed if there where spatial coordinates.[4]

GRAPH 4: TRIP COUNT BASED ON DEPARTING STATION BY USER_TYPE



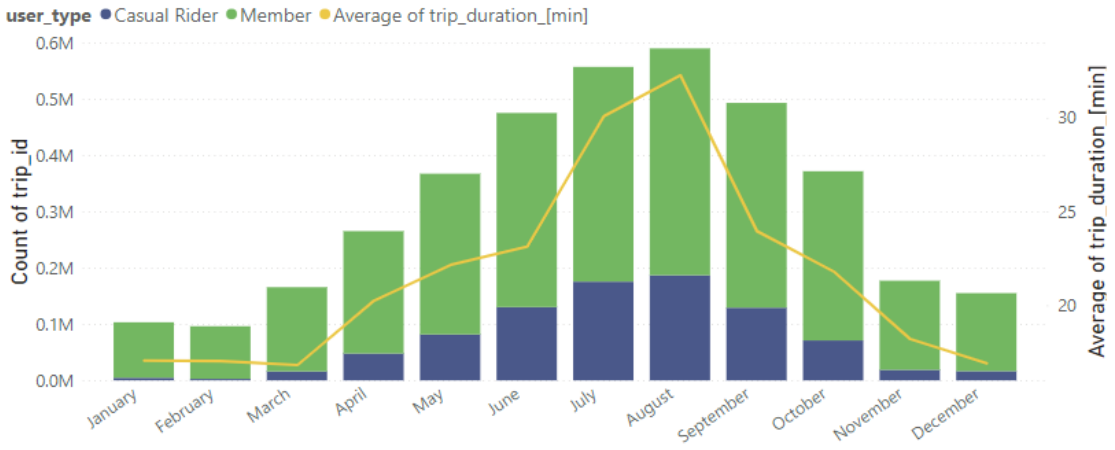
Time series analysis

YEARLY

The following graph shows how the trip count of both Casual Riders and Members increases sharply in the summer months, and so does the average trip duration. So, at the peak months, people take more and longer trips.

In winter, we see a severe drop in usage. Members take a lot fewer trips and shorter trips, while Casual Riders seem to almost not use Cyclistic at all.

GRAPH 5: TRIP COUNT AND TRIP DURATION BY MONTH AND USER TYPE



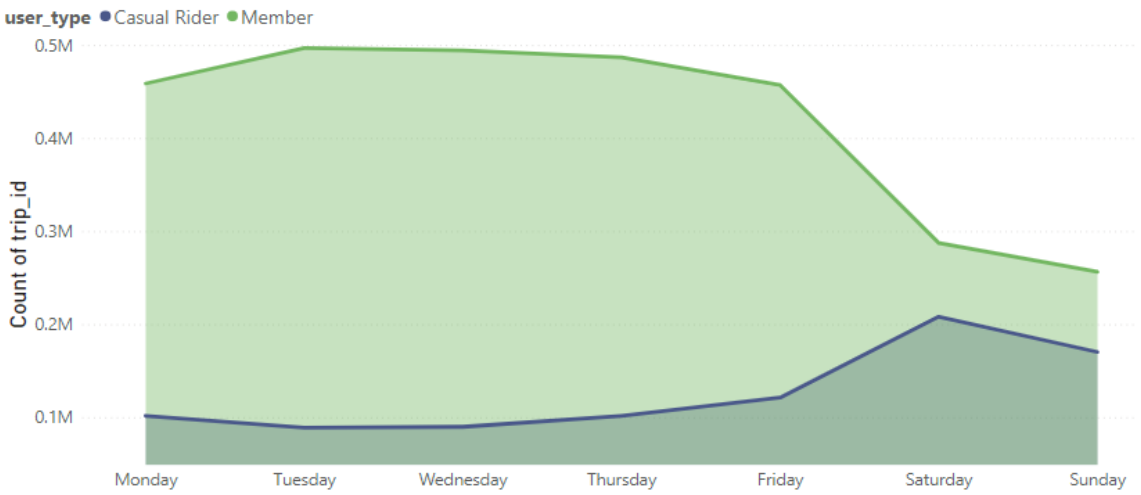
WEEKLY

When looking at the weekly behaviour of the users, Members and Casual Riders seem to have opposite behaviour. Member trip count shows a dip most pronounced on Weekends, while these

[4] In data from following years, there appears to be better spatial information on which to perform a deeper analysis on.

days are the ones that see the most activity from Casual Riders. Weekends appear to be the peak time for Casual Riders while Members see a dip in usage on these days.

GRAPH 6: TRIP COUNT PER DAY BY USER TYPE

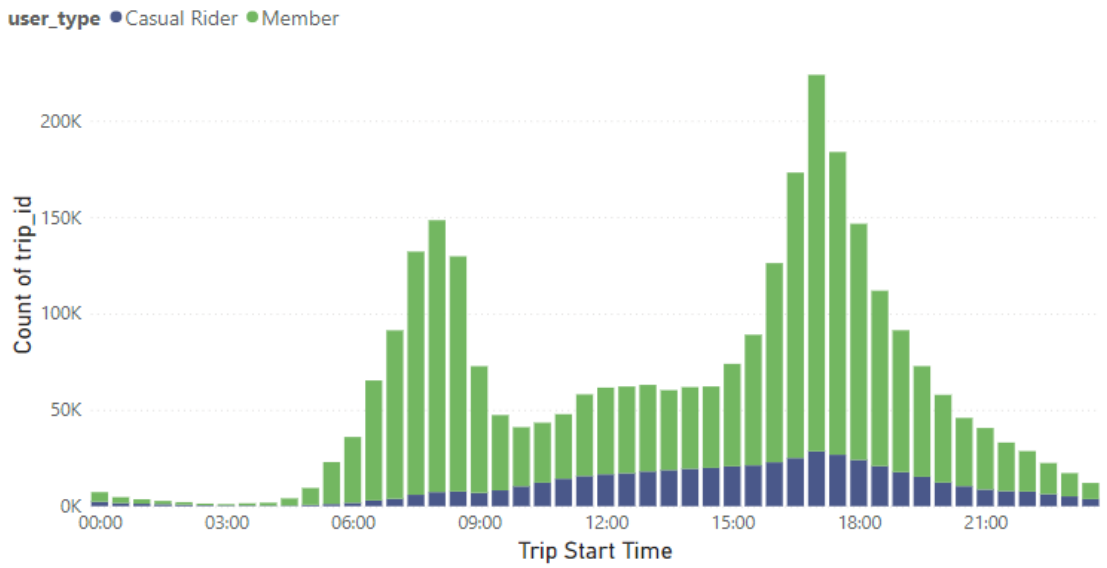


HOURLY

WEEKDAYS

Finally, by looking at an hourly breakdown of the trip counts, we can see that Members appear to be using the bikes for daily commutes on typical work hours (9am to 5pm). This reinforces the behaviour seen in Graph 5. Casual Riders do not show this tendency but show a maximum of usage on the afternoons (5pm).

GRAPH 7: TRIP COUNT PER HOUR BY USER TYPE - WEEKDAYS

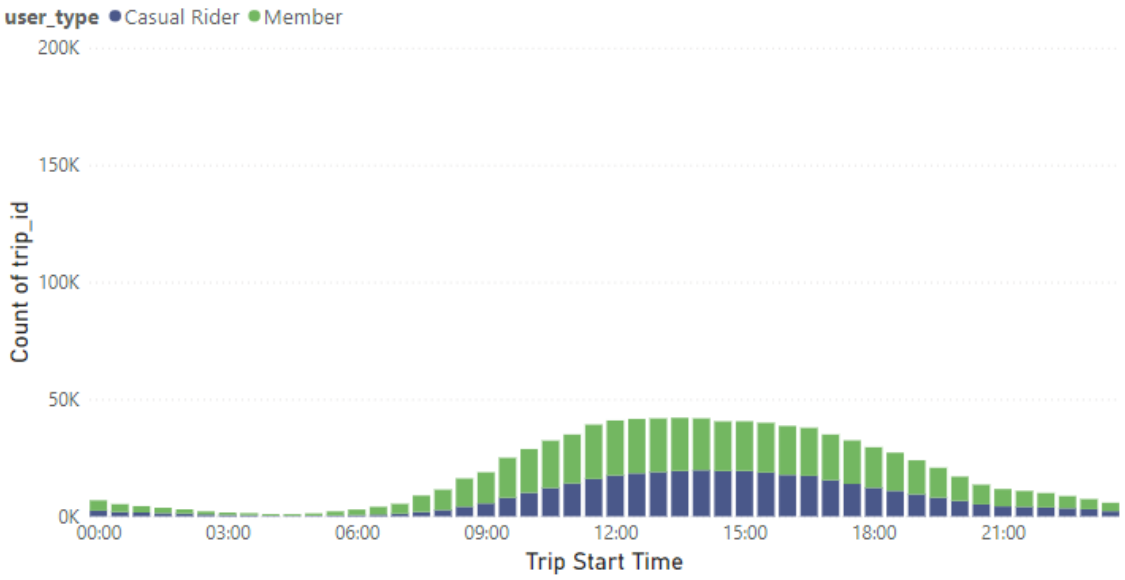


WEEKENDS

On weekends, behaviour of both Members and Casual Riders coincide, showing both a bell curve behaviour with a maximum sometime between noon and 3pm. The number of trips for Members however is much smaller on weekends, while Casual Riders maintain similar numbers

Annex I

GRAPH 8: TRIP COUNT PER HOUR BY USER TYPE - WEEKENDS



Phase 5: Share

To communicate my findings to the stakeholders (the executive team and the marketing director) I prepared a concise report on PowerBI which can be found on Annex II

Phase 6: Act

The business task at hand is to find how Annual Members use Cyclistic differently from Casual Riders. On that account, these are the insights that can be taken from the analysed data.

Insights

Main differences between Annual Members and Casual Riders

TABLE 4: MAIN INSIGHTS DERIVED FROM PREVIOUS ANALYSIS

User Age	Casual Riders are marginally younger than Members
Trip Count	There are much less Casual Riders than Members using Cyclistic
Stations used	Each group favours different stations to both start and end trips
Trip Length	Casual Riders are consistently taking longer trips, whether they are less or more than a day long.
Breakdown by Month	There is a growth in app usage in the summer for both groups. The evident reason would be better weather and temperatures for riding.
Breakdown by Week	Casual Users make more trips on weekends, while Members tend to use the app more on weekdays.
Breakdown by Hour	Casual Riders tend to make more trips in the afternoon, while on average the Annual Members use it more for commuting to and from work.

Limitations

The limitations I ran onto while performing this analysis were the following:

- The users are not individually identifiable, so there is no way of analysing behaviour of recurring clients. That would be especially beneficial for analysing behaviour of Casual Riders. If the same users make repeated trips, it would be a good selling point for Annual Memberships.
- Location data for stations was given in the form of crossroads instead of coordinates.

Moving Forward

After understanding what the differences are between Casual Riders and Annual Members, the following steps have already been established by the marketing director, and they are to answer the following questions:

1. Why would Casual Riders buy Cyclistic Annual Memberships?
2. How can Cyclistic use digital media to influence Casual Riders to become Members?

SQL queries

```
-- Create table
create table trip_data (
    trip_id primary key not null,
    start_time timestamp without time zone,
    end_time date,
    bike_id integer,
    trip_duration_secs character varying(50),
    from_station_id integer,
    from_station_name character varying(100),
    to_station_id integer,
    to_station_name character varying(100),
    user_type character varying(50),
    user_gender character varying(50),
    user_birth_year integer  NULL
);

-- trip_duration_secs could not be cast as integer, so it must be
cleaned first

-- Overview of trip_duration_secs
select trip_duration_secs
from trip_data
order by trip_duration_secs desc
limit 500

-- The problem seems to be the number formating (eg. 1,000.0
instead of 1.000,0)
update trip_data
set trip_duration_secs = replace(trip_duration_secs, '.0', '')

-- We will also need to remove the separating comma ','for the
thousands and millions
update trip_data
set trip_duration_secs = replace(trip_duration_secs, ',', '')

-- Cast trip_duration_secs as integer
alter table trip_data
alter column trip_duration_secs type integer
USING trip_duration_secs::integer

-- Overview of the data
select * from trip_data
limit 10

--It appears that some columns have null values. Specifically in
user_gender and user_birth_year
select count(*) as gender_nulls
from trip_data
where user_gender is null
select count(*) as birth_year_nulls
from trip_data
where user_birth_year is null

-- They are similar numbers, so what is the overlap?
select count(*) as null_overlap
from trip_data
```

```
where user_gender is null and user_birth_year is null

-- A few calculations on this matter
select
      gender_null_percent
      ,birth_year_null_percent
      ,100*birth_year_nulls/gender_nulls as overlap
from
      (select
            (select count(*)
            from trip_data
            where user_gender is null)*1.0 as gender_nulls
            ,(select count(*)
            from trip_data
            where user_birth_year is null)*1.0 as birth_year_nulls
            ,100*(select count(*) as gender_nulls
            from trip_data
            where user_gender is null)*1.0/count(*) as
gender_null_percent
            ,100*(select count(*) as birth_year_nulls
            from trip_data
            where user_birth_year is null)*1.0/count(*) as
birth_year_null_percent
      from trip_data) x

-- Analyzing user_birth_year
select min(user_birth_year), max(user_birth_year)
from trip_data

-- Check for duplicates using trip_id
select
      count(*) as all_rows
      ,count(distinct trip_id) as disctinct_rows
from trip_data

-- Count number of stations
select count(distinct from_station_id) as from_station_count,
count(distinct to_station_id) as to_station_count
from trip_data

-- Count number of bikes
select count(distinct bike_id)
from trip_data

-- See what the most used departing stations are
select distinct from_station_name, count(from_Station_name) as
use_count
from trip_data
group by from_station_name
order by use_count desc
limit 5

-- See what the most used arriving stations are
select distinct to_station_name, count(to_Station_name) as cnt
from trip_data
group by to_station_name
order by cnt desc
limit 10

-- Calculate the average trip length [min]
```

```
select avg(trip_duration_secs)/60 as avg_trip_length
from trip_data

-- Calculate the average trip length [min] for trips not longer
than 4hs
select avg(trip_duration_secs)/60 as avg_trip_length
from trip_data
where trip_duration_secs < 4*60*60

-- Calculate the average trip length [days] for trips longer than
4hs
select avg(trip_duration_secs/60/60/24) as avg_trip_length
from trip_data
where trip_duration_secs > 4*60*60
```

Bike Share Client Analysis

Legend

Casual Rider Member

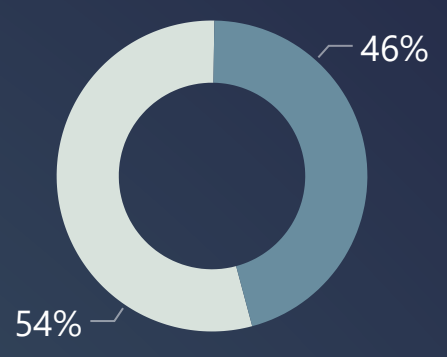
Filters

Weekdays Weekend

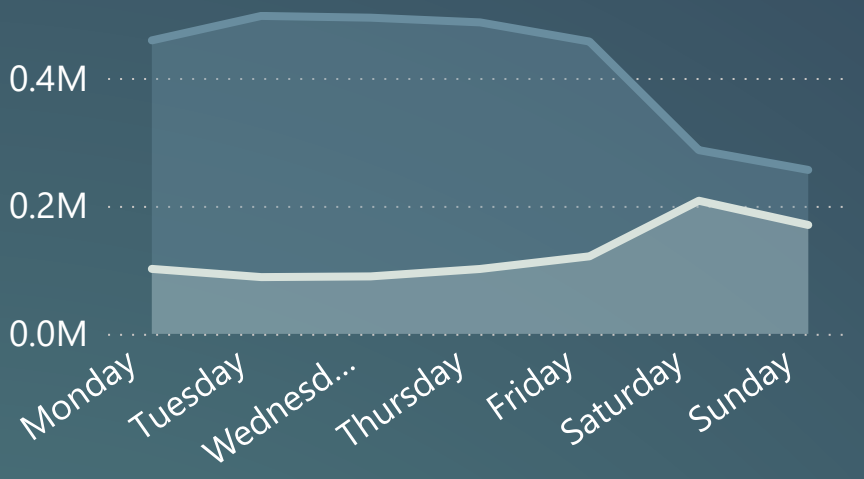
Female Male

Trip Count by User Type

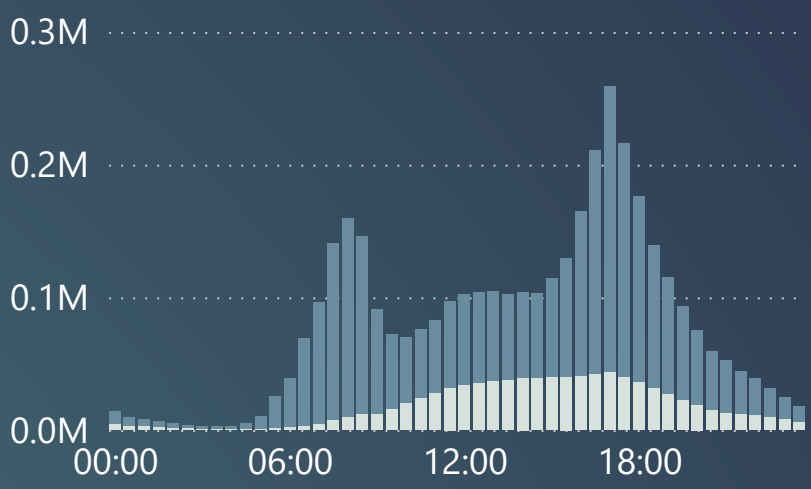
Trip Duration by User Type



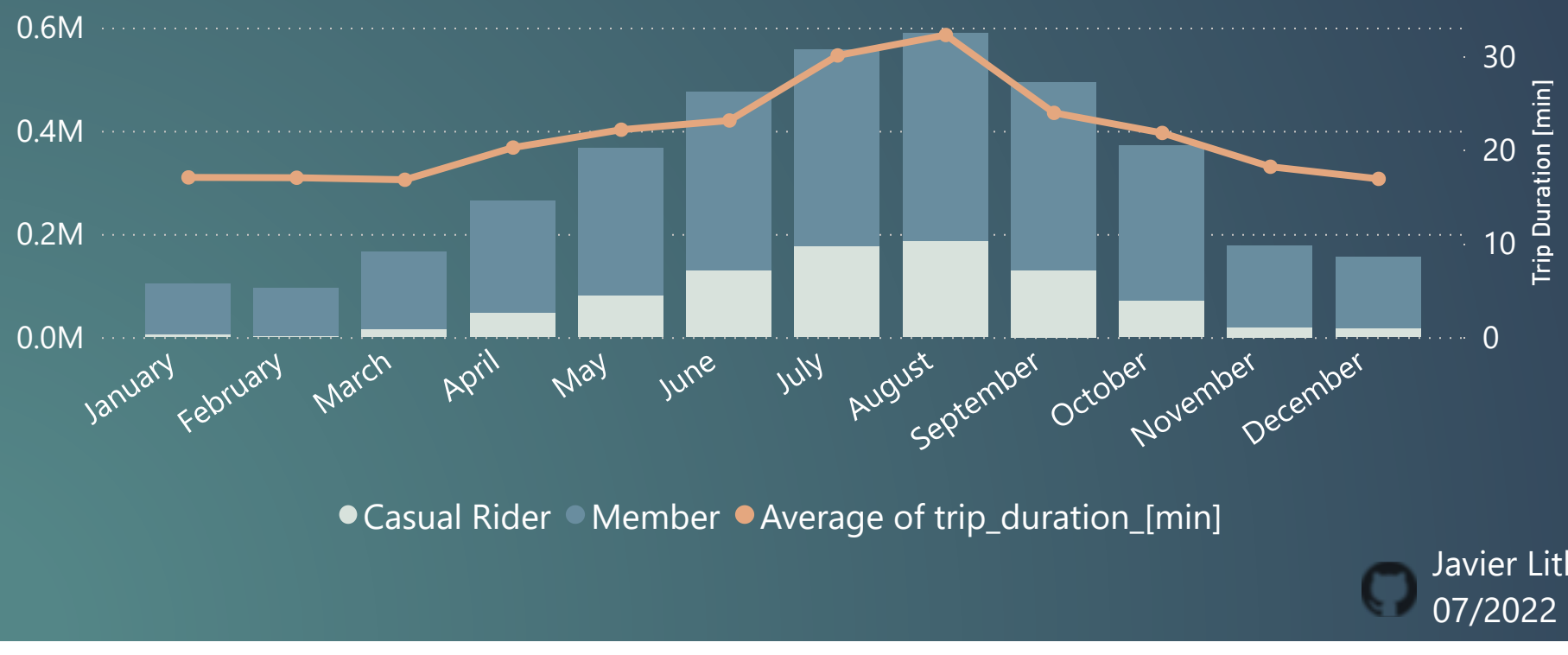
Trip Count by Day and User Type



Trip Count by Start Time and User Type



Trip Count and Average Trip Duration by Month and User Type



Javier Litke
07/2022