

# Google Data Analysis Capstone Project

Presented by

**Phi Minh Quach**

September 9, 2022



## **SECTION A: INTRODUCTION**

### **1. Background**

Having recently finished 8 courses from the Google Professional Data Analytics certificate offered by Coursera, I am encouraged to hand on a capstone project and share all the knowledge I have learned throughout my journey by applying it to the project. In this project, I will go through 6 individual steps in the data analysis process, which have been primarily focused on during the learning process and showcase what I have obtained after all.

### **2. Scenario**

I will play as 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, your 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.

### **3. Information Summary**

Cyclistic is a program located in Chicago which successfully launched bike-sharing offering service, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

Customers who purchase single-ride or full-day passes are referred to as casual riders and 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.

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, the team believes that maximizing the number of annual members will be key to

future growth.

#### 4. Project Report

<b>Title</b>	The Road From Casual to Annual
<b>Industry focus</b>	Marketing
<b>Problem statement</b>	Although the pricing flexibility helps Cyclistic attract more customers, the board team believes that maximizing the number of annual members will be key to future growth
<b>Business use case</b> (what are you solving for?)	<ol style="list-style-type: none"><li>1. Investigating insight about the disparity between casual users and annual members</li><li>2. Understanding the reasons to subscribe for annual membership</li><li>3. Build recommendations in marketing campaign to influence casual customers to become annual members</li></ol>
<b>Goals / metrics</b>	Investigating, understanding the insight and identify a marketing campaign to raise the annual membership subscriptions in the mid-term period
<b>Deliverables</b>	A documentation of analytical process aligning with presentation
<b>Are datasets available</b>	Yes
<b>Data list</b>	Link embed in the case
<b>Websites to obtain the data needed</b>	Link of the case: <a href="#">Capstone Case</a>

### SECTION B: PROCESS

#### 1. Ask

**Problem:** members and casual customers of Cyclistic share nearly the same percentage in total trips and the stakeholders believed that succeeding in turning casual into members would be more beneficial to the program compared to attracting new customers.

**Stakeholder:** Marketing team leader Lily Moreno.

**Task:** what are the main disparities between members and casual customers?

**Other questions:**

- a. Why would casual riders buy Cyclistic annual memberships?
- b. What might be the reasons that make casual riders don't want to buy annual memberships?

c. How can Cyclistic use digital media to influence casual riders to become members?

## 2. Prepare

In this section, I will use available data called Cyclistic's historical trip data to identify and analyze the insight to solve the problem. The data has been made available to the public by Motivate International Inc. with all the links embedded in the Case Capstone (mentioned above). The data is formed in 12 months of trip data with 12 separate .zip files. I only need to download files, extract them and store on my local device (or upload them into other cloud platforms such as google drive, ... etc). The data has also been given as .csv file for each month ([Capstone Case](#)).

The quality of the data appears to meet the standards based on **ROCCC**:

**Reliability:** From prestigious company's database

**Original:** Primary data from Motivate International Inc. was collected directly from their users.

**Comprehensive:** Each table has 12 different attributes, which contribute to the comprehensive of the dataset about detail of every sale within months.
















































**Current:** The data was about 12 months in 2021 (latest data published)

**Cited:** data has been up to date and from reliable source

The dataset will be downloaded and stored in 3 different places:

Local storage: publisher laptop

























Cloud storage: google drive and Google Cloud Storage

	 <a href="#">202101-divvy-tripdata.csv</a>	17.5 MB	text/csv	29 Aug 2...	Standard	29 Aug 20...	Not public	–	Google-managed key		
	 <a href="#">202102-divvy-tripdata.csv</a>	8.9 MB	text/csv	29 Aug 2...	Standard	29 Aug 20...	Not public	–	Google-managed key		
	 <a href="#">202103-divvy-tripdata.csv</a>	41.5 MB	text/csv	29 Aug 2...	Standard	29 Aug 20...	Not public	–	Google-managed key		
	 <a href="#">202104-divvy-tripdata.csv</a>	61.1 MB	text/csv	29 Aug 2...	Standard	29 Aug 20...	Not public	–	Google-managed key		
	 <a href="#">202105-divvy-tripdata.csv</a>	95.3 MB	text/csv	29 Aug 2...	Standard	29 Aug 20...	Not public	–	Google-managed key		
	 <a href="#">202106-divvy-tripdata.csv</a>	130.1 MB	text/csv	18 Aug 2...	Standard	18 Aug 20...	Not public	–	Google-managed key		
	 <a href="#">202107-divvy-tripdata.csv</a>	146.9 MB	text/csv	18 Aug 2...	Standard	18 Aug 20...	Not public	–	Google-managed key		
	 <a href="#">202108-divvy-tripdata.csv</a>	144 MB	text/csv	18 Aug 2...	Standard	18 Aug 20...	Not public	–	Google-managed key		
	 <a href="#">202109-divvy-tripdata.csv</a>	134.6 MB	text/csv	18 Aug 2...	Standard	18 Aug 20...	Not public	–	Google-managed key		
	 <a href="#">202110-divvy-tripdata.csv</a>	110.7 MB	text/csv	18 Aug 2...	Standard	18 Aug 20...	Not public	–	Google-managed key		
	 <a href="#">202111-divvy-tripdata.csv</a>	62.3 MB	text/csv	18 Aug 2...	Standard	18 Aug 20...	Not public	–	Google-managed key		
	 <a href="#">202112-divvy-tripdata.csv</a>	42.6 MB	text/csv	18 Aug 2...	Standard	18 Aug 20...	Not public	–	Google-managed key		

### 3. Process

1<sup>st</sup> step: Open 1 month excel file to have a general idea about the data combining with looking into “schema” table in Bigquery to check for the data type in each column.

2<sup>nd</sup> step: Importing 12 month files into Google Cloud Storage and import them into SQL working console in Bigquery.

	month_1	
	month_10	
	month_11	
	month_12	
	month_2	
	month_3	
	month_4	
	month_5	
	month_6	
	month_7	
	month_8	
	month_9	

3<sup>rd</sup> step: Merging the data of 12 months into 1 “aggregate\_table” using UNION function (had 5595063 rows in total).

```

1  SELECT *
2  FROM `beaming-surfer-359923.trip_data_2021.month_1`
3  union all
4  SELECT *
5  FROM `beaming-surfer-359923.trip_data_2021.month_2`
6  union all
7  SELECT *
8  FROM `beaming-surfer-359923.trip_data_2021.month_3`
9  union all
10 SELECT *
11 FROM `beaming-surfer-359923.trip_data_2021.month_4`
12 union all
13 SELECT *
14 FROM `beaming-surfer-359923.trip_data_2021.month_5`
15 union all
16 SELECT *
17 FROM `beaming-surfer-359923.trip_data_2021.month_6`
18 union all
19 SELECT *
20 FROM `beaming-surfer-359923.trip_data_2021.month_7`
21 union all
22 SELECT *
23 FROM `beaming-surfer-359923.trip_data_2021.month_8`
24 union all
25 SELECT *

```

4<sup>th</sup> step: Check whether all the trip\_Id are unique or not and the result matched the total numbers of rows in a table → unique.

```

1  SELECT
2  count(distinct(ride_id))
3  FROM `beaming-surfer-359923.trip_data_2021.aggregate_2021`

```

Row	fo_
1	5595063

5<sup>th</sup> step: Calculating the date, time difference (day and minute), create new table with these 2 new columns, arranging the table regarding the order of date, minute difference respectively.

```

1  SELECT
2  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, member_casual,
3  date_diff(ended_at, started_at, day) as date_dif,
4  datetime_diff(ended_at, started_at, minute) as minute_dif
5  FROM `beaming-surfer-359923.trip_data_2021.aggregate_2021`
6  order by date_dif, minute_dif

```

6<sup>th</sup> step: Erasing all the columns with null value in any columns (the data is huge enough for us to erase since the null value may skew the result)

```

1  delete
2  FROM `beaming-surfer-359923.trip_data_2021.
   aggregate_before_cleaning_1`
3  where rideable_type is null OR
4  started_at is null OR
5  ended_at is null or
6  start_station_name is null or
7  start_station_id is null or
8  end_station_name is null or
9  end_station_id is null or
10 start_lat is null or
11 start_lng is null or
12 end_lat is null or
13 end_lng is null or
14 member_casual is null

```



This statement removed 1,006,761 rows from aggregate\_before\_cleaning\_1.

7<sup>th</sup> step: Erasing absurd data based on date, minute difference (can not be negative and can not be 0 in both 2 columns because it does not make sense)

```

1  delete
2  FROM `beaming-surfer-359923.trip_data_2021.
   aggregate_before_cleaning_1`
3  where
4  (date_dif < 0 or minute_dif < 0) or (date_dif = 0 and minute_dif
   = 0)

```



This statement removed 59,369 rows from aggregate\_before\_cleaning\_1.

[GO TO TABLE](#)

8<sup>th</sup> step: Checking for duplicate values in all the columns → still have the same number of rows

```
1 SELECT
2 distinct *
3 FROM `beaming-surfer-359923.trip_data_2021.aggregate_cleaned`
```

9<sup>th</sup> step: Checking the integrity of categorized columns

```
1 SELECT
2 distinct(member_casual)
3 FROM `beaming-surfer-359923.trip_data_2021.
  aggregate_before_cleaning_1`
```

```
1 SELECT
2 distinct(rideable_type)
3 FROM `beaming-surfer-359923.trip_data_2021.
  aggregate_before_cleaning_1`
```

10<sup>th</sup> step: Regarding the started day, turned it into day in a week for analyzing purposes.



```

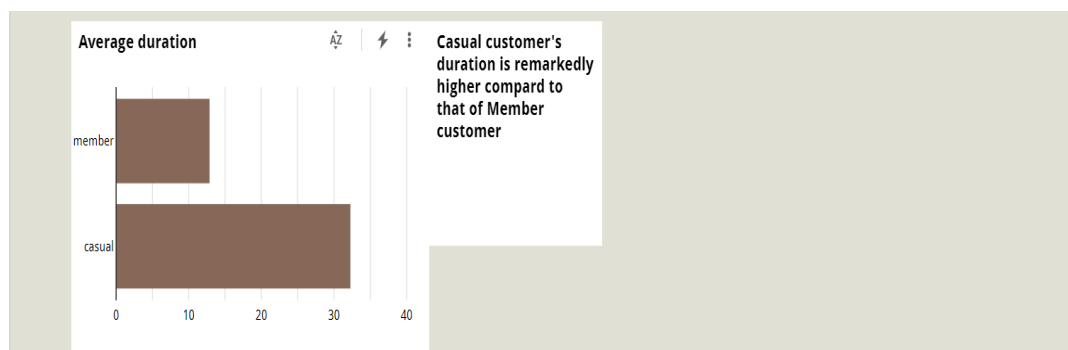
1 With day_of_week_temp as
2 (
3 SELECT
4 ride_id,
5 CASE
6 when EXTRACT(DAYOFWEEK FROM started_at) = 1 then "Sunday"
7 when EXTRACT(DAYOFWEEK FROM started_at) = 2 then "Monday"
8 when EXTRACT(DAYOFWEEK FROM started_at) = 3 then "Tuesday"
9 when EXTRACT(DAYOFWEEK FROM started_at) = 4 then "Wednesday"
10 when EXTRACT(DAYOFWEEK FROM started_at) = 5 then "Thursday"
11 when EXTRACT(DAYOFWEEK FROM started_at) = 6 then "Friday"
12 else "Saturday"
13 end day_of_week
14 FROM `beaming-surfer-359923.trip_data_2021.
    aggregate_before_cleaning_1`
15 )
16
17 select
18 main_table.*, day_of_week_temp.day_of_week
19 from `beaming-surfer-359923.trip_data_2021.
    aggregate_before_cleaning_1` as main_table
20 left join day_of_week_temp
21 on main_table.ride_id = day_of_week_temp.ride_id
22 order by date_dif, minute_dif

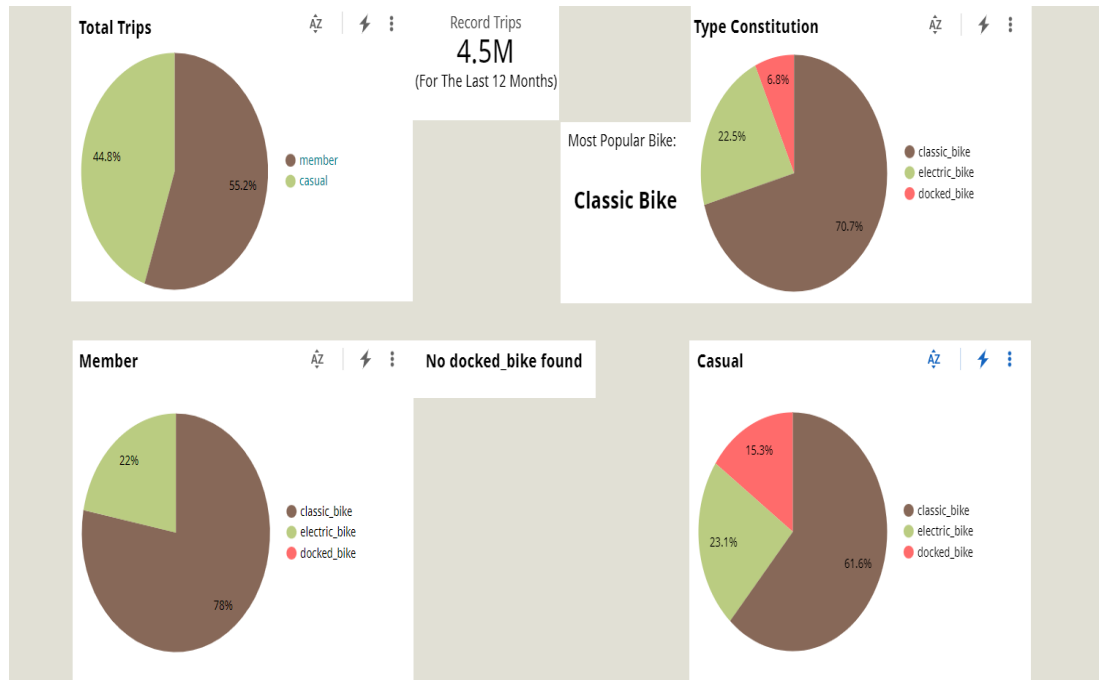
```

After those 10 steps of checking and processing the data, I assumed that the data was clean and ready for deeper analysis.

## 4. Analyze

### a) Overall Summary





From the bar chart above, we can see that casual customers' average duration (nearly 32 minutes) is remarkably higher compared to that of members at only 12.5 minutes.

The four pie charts in the second picture show us some primary information about the customers. The first pie illustrates the constitution of 2 customer groups which are casual and members with dominance belonging to members at over 55%. Secondly, the pie on the top right shows that the classic bike is the most preferred option for all the customers since it remarkably constitutes up to over 70% of the choices with the following from electric bike at only 22%. The two bottom pies surprisingly reveal that only casual customers chose to use docked bikes while there is no docked bike constitution found in the member's chart.

## b) Customers Insights



The top left columns chart shows us that the number of total trips experienced an amazing escalation from May (the start of summer vacation followed by warm weather after a long cold season in Chicago) to October last year with its peak in July and August at around 800k trips from both types of customers. Opposingly, the number of total trips went through a severe cutback from November till April and this may be because of the end of summer vacation followed by the cold and snowing season, which is dangerous for such a vehicle as bicycle and any kind of bike in general.

Coming the top right corner of the picture, we can see that there is a difference between 2 groups of customers: while the member's columns appear to share the stable average duration of each trip throughout the week, that of casual seems to have a slight fluctuation between days especially we can see a slight increase during the weekend compared to the weekday. This may be because of the different usages that each group has: the member groups mainly use it for commuting routines such as work (so the columns are stable and low since bicycle proved to be an optimal option for short distance transportation, which leads to shorter trip duration) while the casual group may mostly use it for other purposes other than commuting to work, which makes its columns way higher compared to that of members

The bottom charts show the number of total trips during the weekday of 2 groups

and once again we can see there is a huge disparity between them. With member group, we can easily see the stability over 7 days a week, which is possible because of the long-term usage that this group's customer has determined before subscribing for membership. When it comes to casual group, it shows us the poor number of trips during 5 days in a week and a critical boost of total trips during 2 days of the weekend and this could be explained by the fact that most casual customers use this service for recreational purposes during the weekend when they have more free time or maybe as an alternative option.

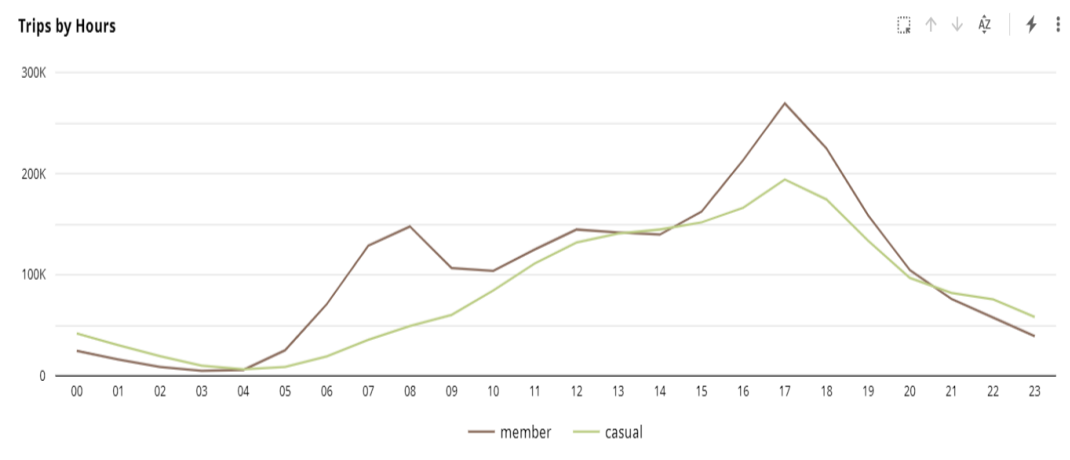
### c) Popular Stations



The charts above show us the popular stations of 2 groups divided into 3, one for casual, one for member and the one left for the combine of 2 groups. Through the charts, it is easily observed that each group has different popular stations (Streeter Dr & Grand Ave for casual and Clark St & Elm St for member). These differences once again show the contrast between the usage of 2 groups, one mostly for working commuting and one mostly for tourist or recreational purposes since Streeter Dr & Grand Ave station is in a park near a coastal town in Chicago while Clark St & Elm St itself is park of Chicago downtown where many companies and building locate. However, both type of customers appears to share mostly the same stations based on

what the combined chart shows.

#### d) Rush Hours During the Day



The line chart illustrates the number of total trips during the day for 2 groups exhibited by 2 different lines. From the chart, we can see that the 2 lines appear to share the same trend during the day with only some small differences. The number of trips started to slow down after 8 P.M till 6 A.M because of the low demand during that period. After 7 A.M casual group slowly increases and goes through the rush hour from 5 to 7 P.M with its peak at 200k trips in just nearly an hour. Back to the member group, it maintains a stable number of trips during the day with 2 peaks: the smaller one at 7 A.M with 150k trips and the bigger one at 5:30 P.M with nearly 300k trips. This could be explainable since 5 to 7 P.M are the rush hours for people to get back from work, which pushes the number of total trips to be higher.

## 5. Share

### a) Key findings

Member group constitutes 55% of the total rides with the dominance belongs to classic bike option → there are still 45% of market available for the project to turn them into annual member.

Only casual members use docked bike options → having the new market and chance to turn casual customers who use docked bike to become annual members.

The average duration of member is intensely lower to that of casual customers → casual customers prefer longer trips in both time and distance.

The busiest period during the year would be from June to September → chance to showcase the benefit of being an annual member to casual customers when they have a

tendency to use the service more often.

Annual group has the stable rides during the week, which is understandable based on their usage, which is mostly for routinely short distance commuting. Oppositely, casual group has a poor number of rides during 5 days in a week with the remarkable jump during the weekend → we can partly figure out that this group may use service mostly for recreational purposes or some specific situations.

Member and casual group have the same trend of bike during the day: slow after 8 P.M till 6 A.M and have 2-rush-hour period from 5 P.M to 7 P.M everyday → Customers tend to use the service more for going to work and returning home.

## **b) Visualization**

I chose to use Data Studio as a tool for both analyzing and sharing my insight in the process since it is easier for me to directly link the work that I have done in Bigquery to data studio because they are both from the Google Ecosystem. Moreover, since the data studio has quite the same interface and mechanism compared to other Power BI tools, there was no such difference in working with the data studio compared to other applications like Tableau and Power BI.

Link: [Project Visualization](#)

## **6. Act**

Regarding the insights have been driven by the analysis above, there are some possible practical recommendations that can improve the goal of the program:

Categorizing the group of casual customers who share the same traits with members such as: short distance/time trips, same popular stations, time mostly using the service,... and target them as potential customers to ask for membership subscription.

Designing some customized membership package for specific groups of customers (weekend membership would be one of the potential packages with high demand).

Increasing the renting price for casual customers during the peak hours or busy day regarding the high demand of service and maybe the lack of bikes to serve.

Offering fringe benefits exclusively for members regarding the distance they use the bike for possibly compel more casual customers.