Google Data Analytics Course Capstone: Cyclistic

You Jia Chuang

2022-09-08

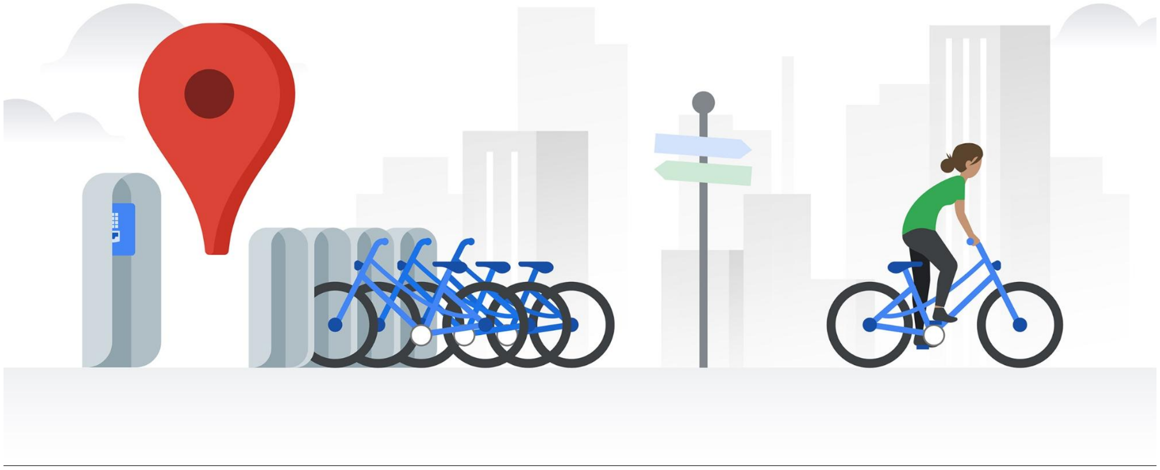


Figure 1: Source: Google The Keyword

Introduction

For the capstone project, I had selected the Cyclistic bike share analysis case to apply my knowledge about data analytics into a real-world task. In this scenario, I acted as a junior data analysis working in the marketing analyst team at Cyclistic, a bike-share company located in Chicago.

In order to successfully answer the key business questions, I followed the 6 major steps of the data analysis process alone the project, which are **Ask**, **Prepare**, **Process**, **Analyze**, **Share**, and **Act**.

Background

The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, our team wants to understand how casual riders (single-ride passes and full-day passes) and annual members use Cyclistic bikes differently. To be noted that customers who classified as casual riders are already aware of the Cyclistic program and have chosen it for their mobility needs. From these insights, our team will design a new marketing strategy to convert casual riders into annual members. However, in order to convince Cyclistic executives to approve our recommendations, the team must provide compelling data insights and professional data visualizations to support our point of view.

- *Cyclistic*: A bike-share program that feature more than 5800 bicycles and 600 docking station across Chicago in 2016. Cyclistic offers reclining bikes, hand tricycles, and cargo bikes which sets itself apart from other companies. It allows people with disabilities and riders who can't use a stander two-wheeled bike can also share the convenience of bike-share system. In addition, there are also three different type of passes that people can purchase, single-ride passes, full-day passes, and annual memberships. The majority of riders opt for traditional bikes: about 8% of riders use the accessorial options. The Cyclistic users tend to ride for leisure, but about 30% use them to commute to work each day.
- *Lily Moreno*: The director of marketing and my manager. Moreno is responsible for the development of campaigns and initiative to promote the bike-share program.
- *Cyclistic marketing analytics team*: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy.

1. Ask

Identify the business task:

Design strategies to increase and maximize the number of annual memberships by focusing on analyzing users behaviours.

Major stakeholders:

Lily Moreno and Cyclistic marketing analytics team.

Stakeholders perspective:

As the leader of the team, Moreno believes that focusing on converting casual riders into annual members would be more feasible for future growth than trying to develop all-new customers.

Questions to analyze

- How do annual members and casual riders use Cyclistic bikes differently?
- Why would casual riders but Cyclistic annual membership?
- How can yclistic use digital media to influence casual riders to become member?

As a junior data analyst, the first question was assigned to me by the team leader.

2. Prepare

Data source:

Extract past 12 months data about Cyclistic's historical trip data which are from July 2021 to June 2022 from the website. The data has been made available by Motivate International Inc. under this license However, data-privacy issue prohibit you from using riders' personally identifiable information, which means i won't be able to connect pass purchase to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased mutiple sigle passes.

Data preparations before analysis:

Firsts of all, after the data from July 2021 to June 2022 were downloaded, Excel was used to open up each file. Some necessary modifications were applied into the data for further analysis. The data originally contained 13 columns, including various information, except the **duration of each trip** and **the day of the week**. The data type of the date in the data was changed and subtraction was applied between the dates in order to calculate the **duration** in a new column. **The day of the week** was showed in a new column named as **weekday** by counted the started date.

Moreover, all the data were exported and the process was continued on another platform, Rstudio.

In Rstudio, the packages that might need were installed first for the convenience of analysis.

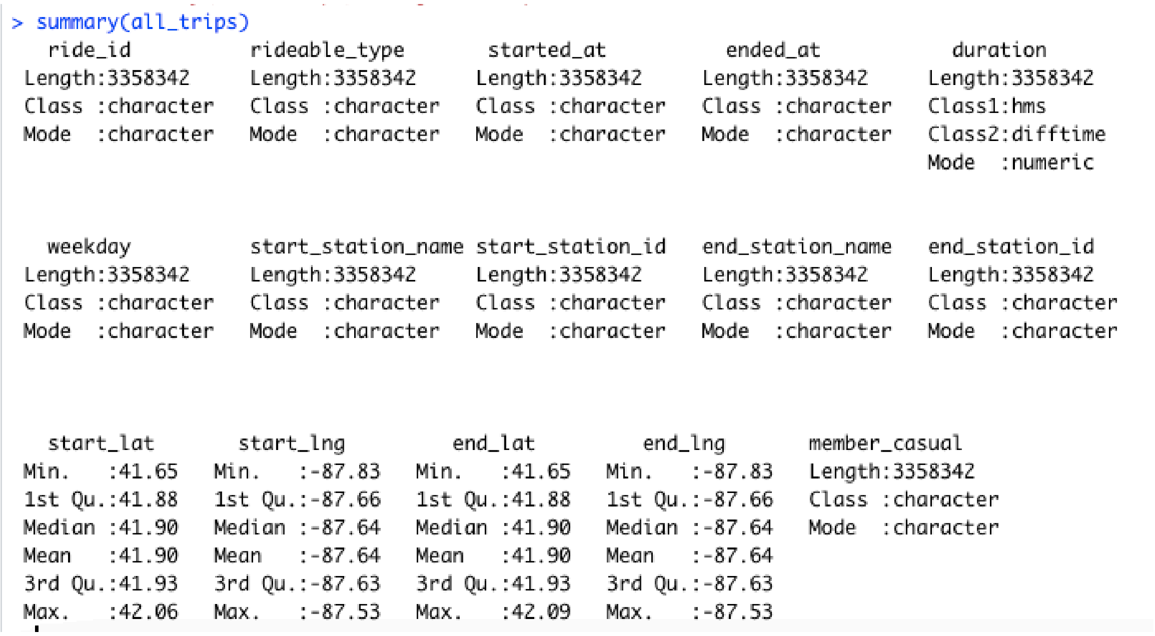
All the 12 datasets were imported in Rstudio and were combined into one large data frame, named as **all_trips**.

3. Process

Data cleaning

When **all_trips** was opened and read, there were 16 columns and almost 6 millions rows in the dataset. In order to make sure the results from the analysis were accurate, cleaning data was an essential step.

Firstly, the columns names were inspected, one unknown column was found and removed by used the *subset* function. After that, the data structure and ingredients were checked. Secondly, the **member_casual** column was cleaned by filter out rows that contained word other than **member** or **casual**. The **duration** column was also checked by removed durations under 5 minutes to exclude those purchases might make by mistakes. Any cell that showed **NA** was removed in the whole dataset during the cleaning process by used the *na.omit* function. The summary of **all_trips** after cleaning was showed below:



```
> summary(all_trips)
   ride_id           rideable_type        started_at          ended_at            duration
 Length:3358342     Length:3358342     Length:3358342     Length:3358342     Length:3358342
 Class :character   Class :character   Class :character   Class :character   Class1:hms
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Class2:difftime
                                                                             Mode  :numeric


   weekday          start_station_name start_station_id   end_station_name   end_station_id
 Length:3358342     Length:3358342     Length:3358342     Length:3358342     Length:3358342
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character



   start_lat       start_lng          end_lat         end_lng        member_casual
 Min.   :41.65   Min.   :-87.83   Min.   :41.65   Min.   :-87.83   Length:3358342
 1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66   Class :character
 Median :41.90   Median :-87.64   Median :41.90   Median :-87.64   Mode  :character
 Mean   :41.90   Mean   :-87.64   Mean   :41.90   Mean   :-87.64
 3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63
 Max.   :42.06   Max.   :-87.53   Max.   :42.09   Max.   :-87.53
```

There were 15 columns and a little bit over 3 millions rows remained after cleaning the data. The columns information were:

- *ride_id*: The unique id for each ride

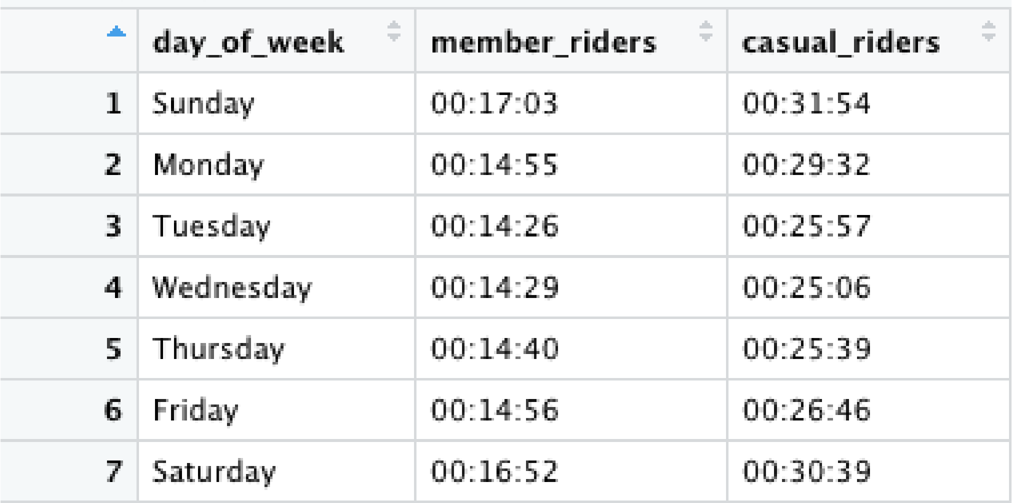
- *rideable_type*: The type of bike ridden, classic_bike, docked_bike, or electric_bike.
- *started_at*: The started date and time of each trip.
- *ended_at*: The ended date and time of each trip.
- *duration*: The length of each trip.
- *weekday*: The day of the week of the trip started.
- *start_station_name*: Start station name.
- *start_station_id*: Start station id.
- *end_station_name*: End station name.
- *end_station_id*: End station id.
- *start_lat*: The latitude of the start station.
- *start_lng*: The longitude of the start station.
- *end_lat*: The latitude of the end station.
- *end_lng*: The longitude of the end station.
- *member_casual*: The type of membership: member or casual.

Overall, the data was ready for further analysis.

Prepare for analysis

Data transformation, reorganization, and visualization were needed before the data could be analyzed.

Relationship between different memberships and average duration on different day during a week: The data was rearranged and reorganized to form a new table in order to make the analysis process more efficient and easier to understand. One set of code was ran for analysis, but **the day of a week** was not in order and the **membership types** were alternated. Firstly, **the day of a week** were organised in order by applied the *level* function. Then, same code was ran again and **the day of a week** were in order. Although the table was formed, the **membership types** did not arranged properly and its readability was low. 3 different lists were created by extracted rows individually from the table to form a new version of the table, the *data_frame* function was used. The table was showed below.



	day_of_week	member_riders	casual_riders
1	Sunday	00:17:03	00:31:54
2	Monday	00:14:55	00:29:32
3	Tuesday	00:14:26	00:25:57
4	Wednesday	00:14:29	00:25:06
5	Thursday	00:14:40	00:25:39
6	Friday	00:14:56	00:26:46
7	Saturday	00:16:52	00:30:39

Moving on, a set of code was ran to group and sort the usertype and weekday to calculate average duration in **all_trips** and created a data visualization by used the *ggplot* function. The data visualization would be showed at **Analyze** process.

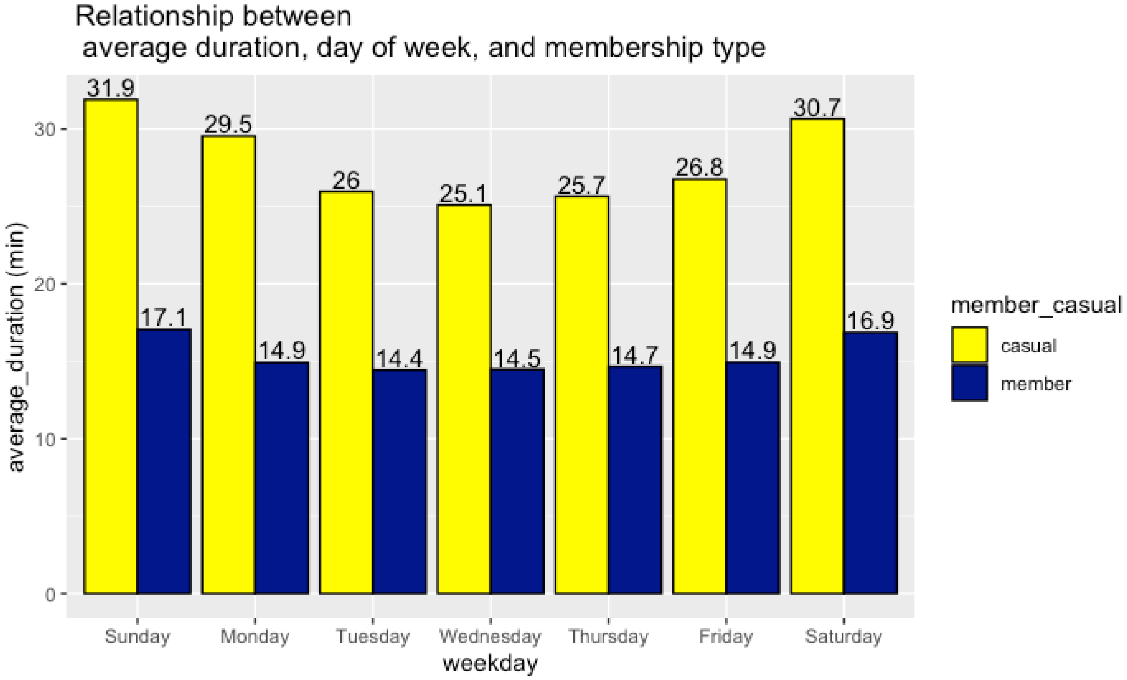
Relationship between different memberships and the number of ride on different day during a week: The relationship of different types of membership and their usage during each day in a week was notable. A set of code was ran to group and sort usertype and weekday to calculate the numbers of rides in **all_trips** and created a data visualization by used the *ggplot* function. The data visualization would be showed at **Analyze** process.

Relationship between popular station* and member types** Lastly, the data was sorted by the name of started and ended stations and different type of memberships to organize the top 5 popular station names for each category. Various function were used, and the data were plotted into 4 separate graphs. For a better comparison, the 4 graphs were rearranged into 1 diagram by used the *ggarrange* function.

To find out more about the coding, please refer to My Github Page.

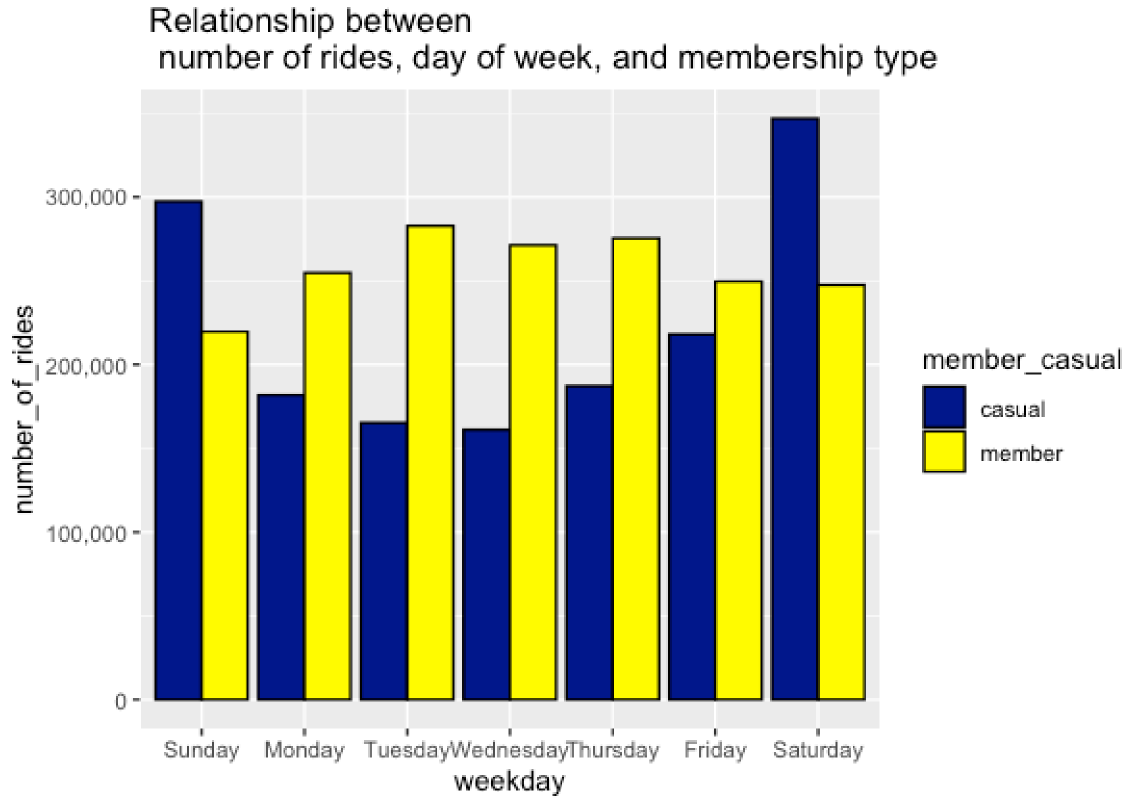
4 & 5. Analyze and Share insights

Relationship between different memberships and average duration on different day during a week: The relationship of average duration and membership types was studied and the imagine was showed below.



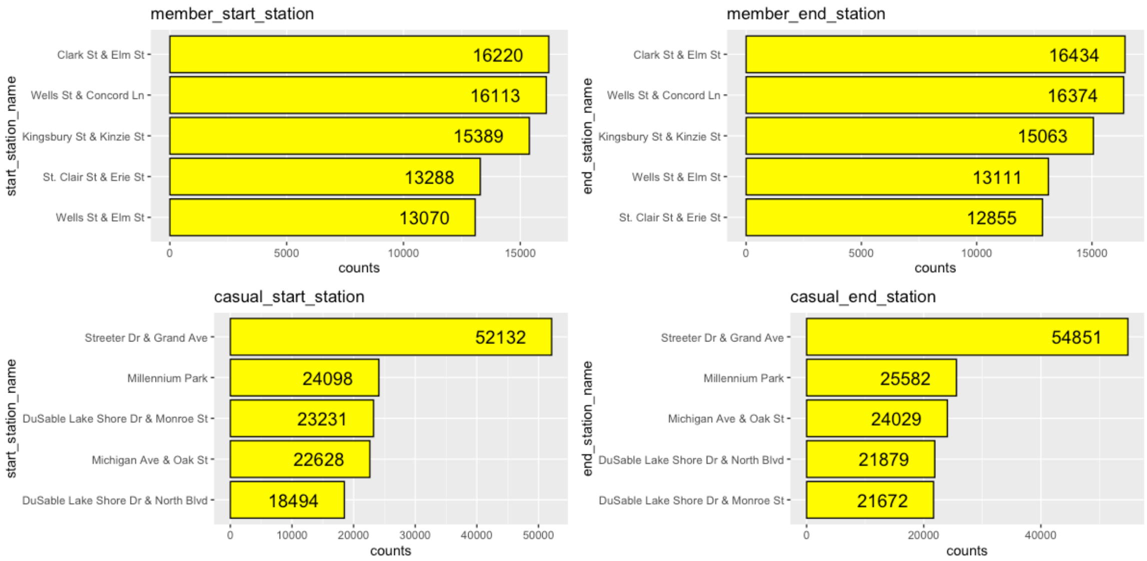
Clearly, **casual riders tended to use the bike longer** compared to member riders everyday during a week. Moreover, the average duration for member riders were falling between 14.5 and 17.1 minutes, while for the casual rider, the range were from 25.1 to 31.9 minutes. **The peaks for the average duration happened on Saturday and Sunday for casual riders**, while member riders showed more steady trend on the average duration but also showed slightly longer time on the weekend.

Relationship between different memberships and the number of ride on different day during a week: Firstly, the relationship between annual member and causal member and their usage on each day of a week was discussed. As the diagram shown below.



It was very obvious that the usage of casual riders compared to member riders was much more fluctuate during a week. In addition, **the number of rides for casual riders were mostly concentrated on weekend, such as Saturday and Sunday**; while annual members tended to use the bike during the weekday which showed slightly higher numbers than weekend. Another key feature was that there were **only Saturday and Sunday that the usage of bikes for casual riders were higher than annual member**. What's more, **the 2 biggest numbers for the usage were both happened to casual riders during the weekend** .

Relationship between popular station and member types The graph below showed the top 5 stations for member riders and casual riders by started station and ended station.



According to the graph, the top 5 stations for member users were the same for started and ended station, and the **casual users showed the same top 5 stations for started and ended** as well. However, the top 5 stations for started and ended weren't the same for member users and casual users. The most popular station for member users was the station at Clark St & Elm St, while **the most popular station for casual users was the station at Streeter Dr & Grand Ave**. Another point to pay attention was that the gap for the number of rides between the top 5 stations for member users was very small, on the other hand, **the gap between the top 1 and 2 stations for the number of rides for the casual users was significantly huge**. In addition, **the usages of the most popular station were more than double of the usages of the second popular station for casual riders**.

6. Act

Potential solutions and outline

1. According to the analysis about *the relationship between average duration and membership types*, **casual riders showed relatively longer usage time** than member riders. In order to convince casual rider to purchase the membership, **working on how to provide more benefits to longer ride users while remaining the profits** is recommended. After that, **showing the results of how the riders could be benefit when the ride is longer by joining the annual membership to casual users** might increase the their interest.

2. Based on the *relationship between number of rides and membership type during the week* analysis, **casual riders showed much more usage during the weekend** compared to the number of rides during the weekday. Therefore, **focusing on how to provide as many conveniences as possible or more access to the bikes during the weekend in the membership to make a clear difference between the annual member and casual user** to convince those weekend casual users to purchase the membership, such as **certain numbers of bikes are only for annual member on the weekend at a station** and **cheaper fare then buying a single ride or day pass** etc.

3. In a reference to the * relationship between popular stations and member types* analysis, **the casual riders tended to use one specific station more then other station**. As a result, **the company should put more effort on certain stations or areas on how to convert casual riders to pay for annual membership rather than other options**.

4. Regrading to advertising, **the company should target on how to show casual riders the benefits of annual member when it comes to a longer ride** and **focusing on those popular stations at the weekend to disseminate the information**.