**Capstone Project**

**A Case Study on Bi-Cycle Ride Share**

**Prepared as part of Google Data Analytics Certificate course**

**Prepared By:** **Marium**

**Overview:**

Cyclist is a bike sharing platform where people rent bike for a particular time period. They offer annual membership and casual renting. As there are not a lot of data about how they operate, I’m following ‘**Bike Share, Toronto**’ business model to base my assumption.

**Offerings at a glance:**

1. Annual Membership  
   a) CAD 100- 30 mins unlimited ride time   
   b) CAD 115- 45 mins unlimited ride time
2. Casual Membership  
   a) CAD 3.25- 30 mins  
   b) CAD 4.00- 45 mins  
   c) CAD 7.00- 24 hours 30 mins long unlimited membership  
   d) CAD 15.00- 72 hours 30 mins long unlimited membership

The organization believes that there is a very good chance to convert casual riders into members. So, my goal is to analyse their cycling behavior, pattern, how casual differs from members, share those findings with the team and help them to make better decision to convert casual riders into members.

**Business task statement:**

My task is to find out how casual riders differ from members. I will investigate data and analyse their cycling behavior. It will help us to find their cycling pattern. That will help the business to understand casual riders better, offer even better membership value. Eventually it will help to grow business financially, improve society, lead the world to greener future and lessen carbon footprint.

**Prepare:**

**Data Source:**

Data source needs to be **reliable, original, comprehensive, current** and **cited** (R.O.C.C.C).

I have downloaded the data from the link provided by Coursera. All data owns by google and the organizations.

Out off the shelve, the data were nicely organized and structured. As I’ve downloaded the data, data collection was out of my project scope. So, my assumption here is that the data is produced internally, and it is the original data. Here, I have both qualitative (Name, description or any data that cannot be measures or counted, sentence that is not numbers) and quantitative data. Some quantitative data are discrete, and some are continuous data.

**Data Ethics & Privacy:**

I tried to get some idea how data has been collected, shared and used. Again, as I do not have full access on those processes, I cannot make any comments here.

As for sensitive personal information, there is no data which can link to any individual to identify them personally. It looks like the data I have got is originated from organization’s own application.

For user consent on why data is collected, how it will be stored and how long it will be stored or on transactional transparency, I can only assume that they have used and followed proper processes as I don’t have any visibility here as well.

**Cleanup:**

To cleanup data, I looked for unwanted data, extra spaces, formatting, duplicates, and blank cells.

**Cleanup Process & Assumption:**

I have added few columns to get better understanding of data and one of the columns is ride length. There are some ride lengths which are too short and too long.

As I cannot talk to anybody from management about those data, I have no choice but to assume how long or short ride length I should keep. Those could be someone took a bicycle for maintenance.

So, I have deleted all the data with ride length –

1. More than one (01) day and
2. Less than three (03) minutes

**Blank Cells:** There were a lot of blank cells for Start Station Name, Start Station ID, End Station Name and End Station ID. I have deleted rows where there were either both Start Station Name and ID missing or both End Statin ID and Name were missing.

Before deleting those rows with blank cells, I tried to fill up them from alternative sources. In my case I tried to fill up the both start stop name and ID from latitude and longitude. But latitude/longitude data value after decimal point are not same everywhere (example, ‘Clark St & Touhy Ave’ latitude is 42.0127863333333 in one place and 42.012701 somewhere else). There are different value latitudes and longitudes associated with a single location. Because of the inconsistency, I decided not to use latitude/longitude as alternative source.

**Unwanted Data:** As latitude/longitude was not usable for alternative sources, I decided not to include them in my analysis.

To check data integrity, I’ve checked for duplicate user ID to make sure I don’t have same value more than once.

**Format Change:** I have changed the format of the Ride Length as there were a lot of value which was over a day and time format cannot show value over 24 hours. So I used custom ([h]:mm:ss) format to show average ride length. For all other places I used float format which shows value in days as sum of ride length is huge in hour format.

**Extra Space & Spelling or Typos:** I couldn’t find any extra space, spelling or typos in any of the files.

**Process:**

To process data, I used both Excel and SQL.

For SQL,

1. Opened the downloaded CSV file in Excel and saved them as .xlsx file
2. Then I used Export Import Utility tool from SQL Server Management Studio (SSMS) to import data from excel file to MS SQL database
3. Once import is done, I have run multiple queries to check data integrity, Null value, exceptionally long/short ride length
4. I have also run multiple queries to get more insight on data to get better understanding on data (all queries are in GitHub: <https://github.com/SonaliMarium/googleproject>)

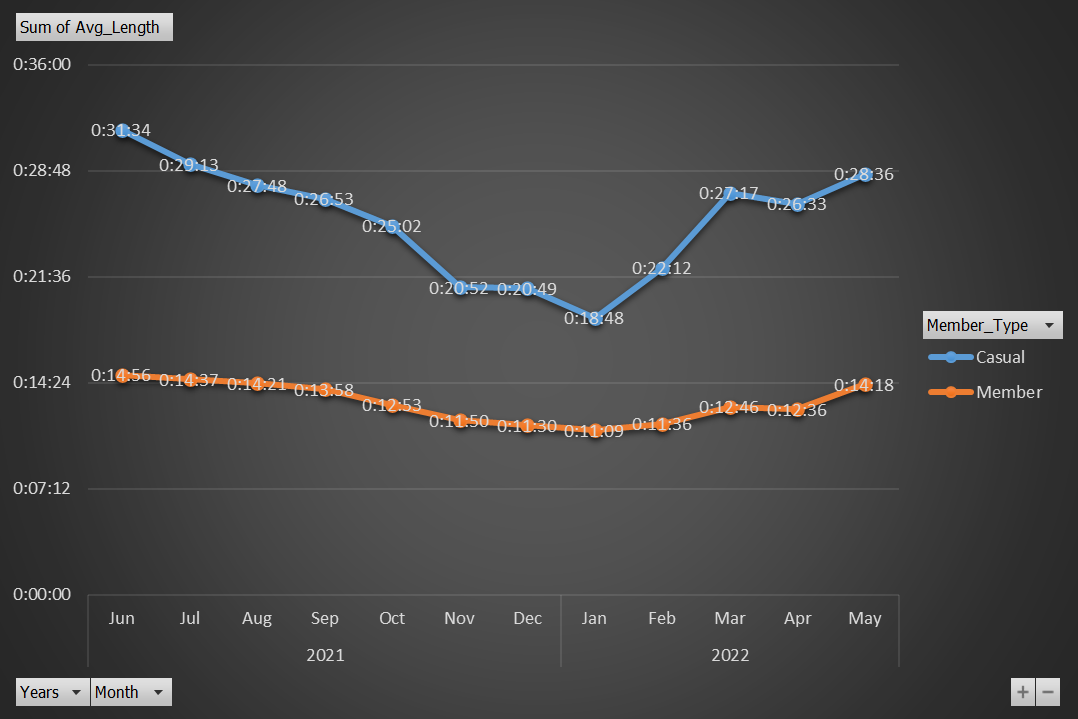
For Excel,

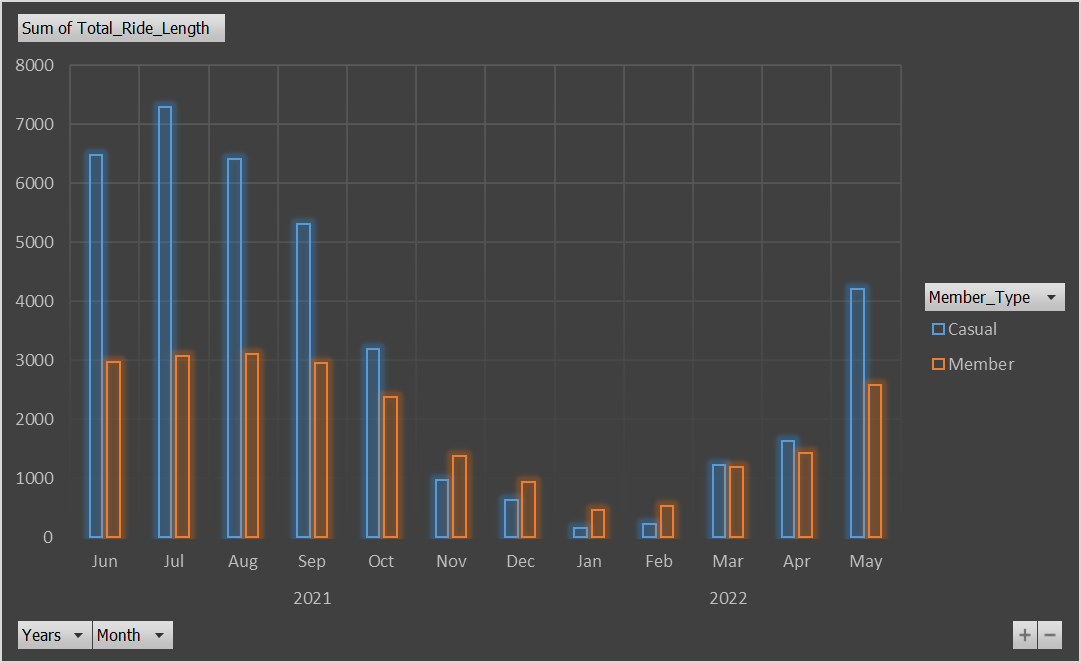
1. Opened the downloaded CSV file in Excel and saved them as .xlsx file
2. Used ‘TRIM’ to get rid of extra space
3. Used ‘Go To dialogue box’ by pressing F5 and used special to select all Blank Cell to delete them
4. Used ‘Proper’ for Case change
5. Calculated min, max and mode of the day of week
6. Calculated ride length (=end\_at - started\_at)
7. Calculated avg. ride length

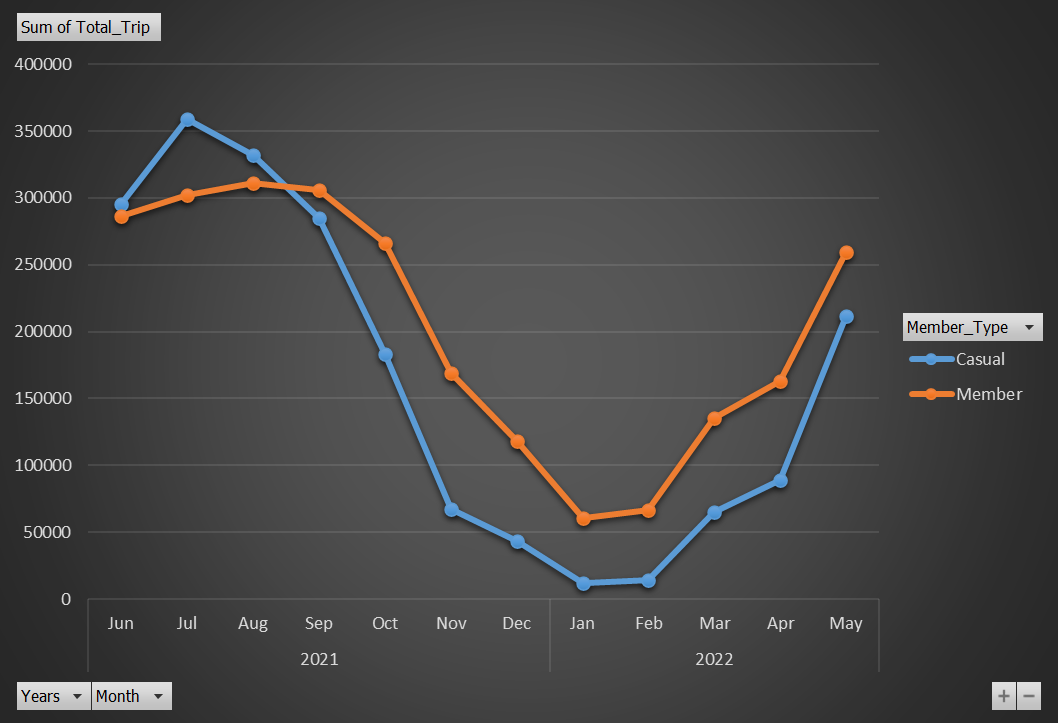
**Analyze:**

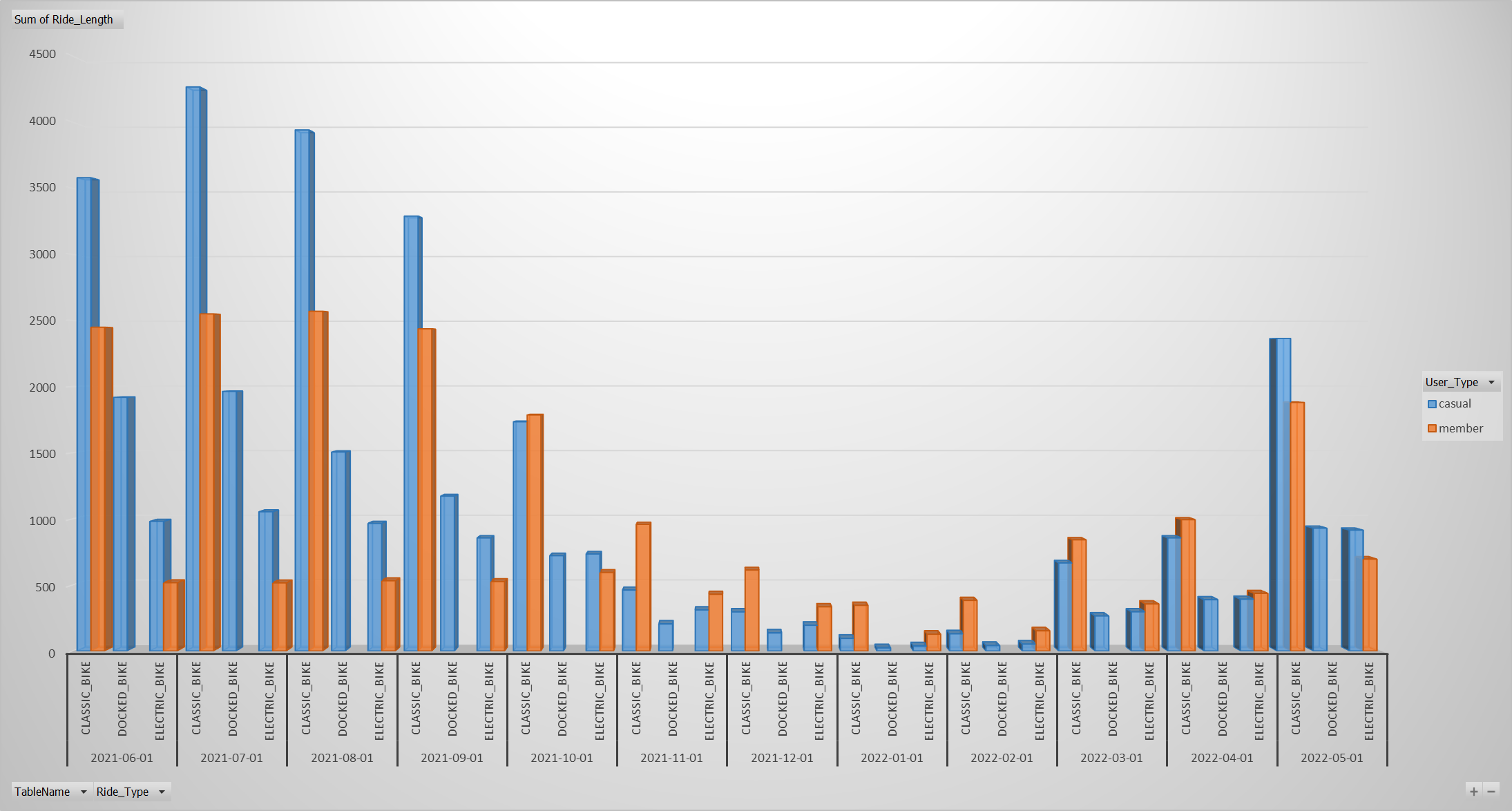
I have run multiple queries to see trends. Some of my observations are

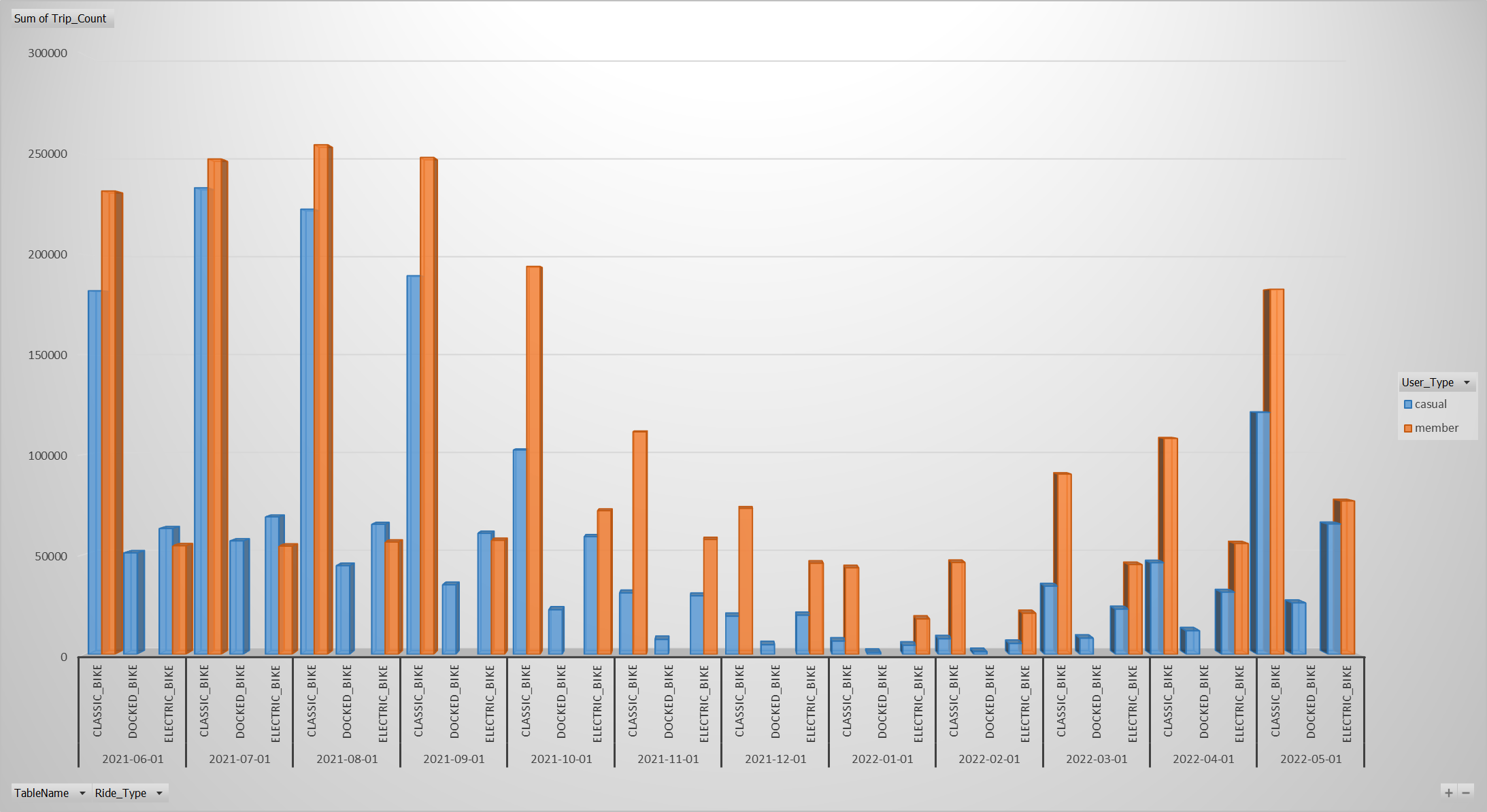
1. Classic bike is the most popular bike among both types of riders. It is true in terms of ride length and ride count as well
2. Only casual members used ‘Docked Bike’
3. On an average, casual riders rode more length
4. Casual riders rode most all over the year except November to Feb, means around winter season. Around that time their riding went down drastically
5. Member riders also used bike less around wintertime but not as much
6. Members used bike more in terms of ride count means most of their ride lengths are short
7. Casual members used bike more on weekend mostly on Saturday
8. The most used station for casual cyclist is Streeter Dr. & Grand Ave.











**Recommendation:**

Based on the above analysis, some of my recommendation are

1. Casual riders ride mostly on weekend. So, we can offer some deal only for weekend.
2. Casual riders use ‘Docked Bike’ exclusively. So, we could make some offer on docked bike.
3. Casual riders are mostly seasonal bikers. Some season-based offering could be beneficiary
4. Also, they tend to ride more length. Length based offering could be helpful as well