**Introduction**

This is analysis of [Google Data Analytics Capstone: Complete a Case Study](https://www.coursera.org/learn/google-data-analytics-capstone) course. In this case study, I will perform many real-world tasks of a junior data analyst for marketing Team at Cyclistic, a bike-share company in Chicago.

**Background**

In 2016, Cyclistic launched a successful bike-share offering. Since then, 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.

Until now, Cyclistic’s marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

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, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

**Scenario**

You are 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, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

**Question to Analyze**

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

**Prepare**

**Data Source**

Dataset from 01-01-2021 to 31-12-2021 available from [Cyclistic Bike Sharing Data](https://divvy-tripdata.s3.amazonaws.com/index.html). The data has been made available by Motivate International Inc. under this [license](https://ride.divvybikes.com/data-license-agreement). Each csv files contain 13 columns containing information related to ride id, ridership type, ride time and location and location etc. File naming: YYYYMM-divvy-tripdata.csv and stored by monthly data file (total 12 files). Next, these datasets will be organized by PostgreSQL.

**Process**

First, we need to create a database for our dataset. We can use right-clicking of databases icon in menu bar and create database. In this case, the name of database is ‘Cyclist’. Next, we can activate Query Tools and create the tables for our datasets by querying it. The name of headers is based on the name of csv's file header and the name of this table is 'divvy\_tripdata\_2021'. Finally, we can import our csv files by right-clicking 'divvy\_tripdata\_2021' table under schema of ‘Cyclist’ database.

After that, we can check the structure of our previous table using PSQL tool by right clicking 'Cyclist' Database. Then we can input the command '\d divvy\_tripdata\_2021' without quotation mark.

Table

Description automatically generated with medium confidence

We also can check the preview of first n-values in our tables in Query Tool. In this case, I want to check first 10-values in 'divvy\_tripdata\_2021'.



Then we can also check the total number of All rows from this table.

Graphical user interface, application

Description automatically generated

**Data Cleaning**

Next step is cleaning the data. First, I want to investigate any possibility of duplicate data. In this case, I want to investigate the possibility of EXACT duplicate data in each column using this query.

Text

Description automatically generated

The result shows there is no duplicate data which means, this table has the unique values in each row.

A picture containing text

Description automatically generated

We may also check the duplicate data based on any certain columns to investigate it deeper. In this next case, I want to check any possibility of duplicate values even the ride\_id is unique. This is likely to happen because the data is inputted more than once by accident.

Graphical user interface, text, application, Word

Description automatically generated

The last results also show its consistency if there is no duplicate data in this table.

Next, we must check the integrity of dataset. It means the data must contains right and logic value. To know it, I use duration of trips for each ride\_id. In this part, I use two parameters using difference of time in the column 'diff\_time' and difference times in second in the column "diff\_in\_second". For notes, PostgreSQL don’t have the function of ‘datediff’ like other SQL application to extract time difference from timestamps data. In this case, I use ‘date\_part’ function from PostgreSQL.

Text, application

Description automatically generated

Table

Description automatically generated with medium confidence

Everything seems to be correct, but we must ensure it deeper, if there are not duration of trips in in negative values. I use these two queries to double check deeper the integrity of dataset.

Text

Description automatically generated

Text

Description automatically generated

Voila, we find it. There are 147 rows with negative values. These rows are wrong data, and we need to remove these rows from dataset.

After removing the wrong data, we need to deal with Nulls values. A NULL value indicate that a data value does not exist in the database. In other words, it is just a placeholder to denote values that are missing or that we do not know. NULL can be confusing and cumbersome at first.

From the previous data preview, we can see if Nulls seems to be appeared some columns. First, I want to investigate the number of Null per column.



From this result, we know most of missing values mostly happen in stations columns, both in 'Start\_station\_name' and 'end\_station\_name'. The result shows, there are maximum 739.149 rows with at least 1 missing value in their columns. It means this table has 13,21% Null values (from 5.594.916 rows).

Next, I want to investigate where does the Null mostly occur in dataset. The easiest parameter is using 'rideable\_type' column because it contains 3 variables for comparison.

Table

Description automatically generated

The query results show that the majority of NULLs appear in 'electric\_bike' data. There are 729.816 nulls or 98,07% of the total Nulls that appear ONLY in 'electric\_bike' data. This number is also equivalent to 35.92% of total 'electric\_bike' data. So, instead of deleting all data with Null, which means it will delete a lot of data and cause a possible inaccurate analysis. I will treat Null data differently.

--- CASE 1: If total dif\_in\_second is Not 0 AND start\_lat or start\_lng is Not Null AND end\_lat or end\_lng is NOT Null THEN there is a transaction on this ride\_id even though the column of 'start\_station\_name' or 'end\_station\_name' contains Null values.

--- CASE 2: If total dif\_in\_second is 0 AND start\_lat or start\_lng is NULL OR Not Null AND end\_lat or end\_lng is Null THEN there is no transaction on this ride\_id. WE MUST REMOVE THIS ROWS.

Case 1 Sample:

Graphical user interface, application, table

Description automatically generated

Case 2 Sample:



Total Count of Case 2

Table

Description automatically generated

The result shows, there are 506 deleted rows with Case 2 Criteria

After dealing with Nulls, next we need to deal with Outlier data. Outlier is a data point in the dataset that differs significantly from the other data or observations. Many statistic procedures are affected by the presence of outliers. So, in this case, removing the outlier may be an option. We can use standard deviation to remove outliers and get a "trimmed average". In PostgreSQL, the function is called stddev\_samp().

Text

Description automatically generatedText

Description automatically generated

The result:

Table

Description automatically generated

The result shows, there are 8636 outlier data with maximum Values 932 hour or 38 days for using bike and most of them coming from 'Casual Members'. In this case, I will use one way to deal with outliers to know the result of future analysis by removing all of outliers.

**ANALYSIS AND SHARE**

It’s time to analyze our cleaned data using R. R is the best option when we analyze and visualize data in one platform. First, I loaded necessary libraries and read the csv file. I use ‘fread’ function from ‘data.table’ library, instead of ‘read.csv’ function. ‘Fread’ function has faster and more convenient capability to import a large file. The name of data frame for this dataset is ‘trip\_data’.



Here is the summary of trip\_data:

A close-up of a document

Description automatically generated with medium confidence

Before starting the analysis, I want to create three new variables in the data frame. These new variables will allow us to make our analysis easier.

These variables are

trip\_data: distance of trip in meter. I use distHaversine function from geosphere library to extract the actual distance between two points of lon-lat data.

Text

Description automatically generated

duration\_in\_min: duration of trip in minutes as numeric data. I use difftime function from lubridate library to extract the time in minutes.

Text

Description automatically generated

day\_of\_week: the day of date time.

Text

Description automatically generated with low confidence

Now let’s start our analysis. First, for the basic one. Let’s start to know the comparison of total count between casual and member.

Text

Description automatically generated

The result shows, the member users have larger percentage than casual from the total users. For the notes, it’s not the actual member that represent the actual data. The data itself has the LIMITATION that all the transaction did not include the user id. So, the assumption is all the transaction is unique for each user.

Next, plot the bar chat to create first analysis to compare the casual and member in their distribution of rideable type.

Chart, bar chart

Description automatically generated

This plot shows that us member users mostly being dominant by using classic bike and little bit higher for using electric bike than casual users. In other hand, docked bike only being used by casual users.

Let’s move the analysis by the day of week.

Chart, bar chart

Description automatically generated

The plot shows us that member users use the service higher than casual users on weekdays. For the member users the usage services on weekdays are relatively same and even lower on weekends. In other hand, casual users use the service mostly at weekend, especially on Saturday. The usages on weekdays are relatively lower than on weekend for the users. We can assume if the reason for members using the services is for supporting daily activities like working and for casual users is for entertainment during the weekends.

Now let’s observe the duration in minutes between casual and member users using histogram. The result show, that the casual users take longer duration than the member users. But the maximum of duration is longer for member users.

Chart, histogram

Description automatically generated

Still using the histogram, it hard to observe the detail of difference between casual and member users for their distance of trip. But we can shortly conclude, if the member users have the highest count of distance traveled than casual users. In other hand, casual users seem to have the highest distance of traveled. Let’s breakdown it further.

Chart, histogram

Description automatically generated

First, I make the filtered variable among the member and casual users. Then I breakdown the summary of each user for their day of week, duration, and distance. Here is the clear summary of each user:

A picture containing table

Description automatically generated

The result shows some points:

* The highest peak for member users in day of week is Wednesday and for casual users is Saturday.
* The mean of distance is quietly similar although it is higher for casual users than member users. The maximum distance for casual users is far higher than member users.
* The casual users also have the longer mean of duration for using the services than member users. Both of users almost have the same maximum duration for services.

Next, I want to know the most popular starting stations between member and casual users.

Text

Description automatically generated

The result looks interesting. The count of starting stations for member users is almost same for each station. But the difference for the Top Popular starting station than others for casual users is far higher. Streeter Dr & Grand Ave is the most popular starting station for casual members.

**CONCLUSION**

From the previous analysis, it’s still hard to give the best suggestion based on current analysis because the data have the critical limitation, especially from the uniqueness of each transaction from each user. However, we still have some useful conclusions that might be applied for future decisions. Let’s summarize what we got:

* Classic bike is the most favorites choice for both casual and member users, but the largest user for this service is member users. Docked bike is the only used by casual users. Both of users quietly love electric bike service.
* The usage of service for member users is mostly during weekdays and it peak is Wednesday. For the casual users is during weekend and its peak is Saturday.
* The mean duration and distance of trip for casual users is longer than member users. A great opportunity if we can convert these users.
* The count of starting station name almost similar for member users but for casual users, there is a favorite starting station to use these services. That is Streeter Dr & Grand Ave station.

According to the previous conclusions, my suggestions to answer the second question which is converting casual to member users are:

* Grabbing momentum during the weekend in the most popular starting station. Placing the promotion like event/programs and other advertisement are the best movements like in Streeter Dr & Grand Ave station.
* Dive the best pricing for using docked bike to targeting casual users who join membership since we know docked bike is the popular service on casual users.
* We can make a gamification program like exclusive ‘quest’ for member that reach certain distance and duration trip with special rewards.
* During the process, we found if there are huge numbers of missing and outlier data from casual members that using electric bike service. It should be concern for the future recording.