**Google Data Analytics Capstone: Complete a Case Study on Cyclistic Bike sharing**

**About the Company**

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.

Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers  
who purchase annual memberships are Cyclistic members.

**Objective:-**

The marketing director 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.

**Phase 1- ASK**

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?
4. Which age group and for what purpose members and casual riders use bike?

**Phase 2:- Prepare**

Cyclistic is a fictional company with the data set is public. You can download it [here](https://divvy-tripdata.s3.amazonaws.com/index.html). I used the 2019 data set from the lists. This data has been made available by motivate International Inc. under this [license](https://ride.divvybikes.com/data-license-agreement).

* I downloaded the data comprises of 4 data sets for four quarters in 2019.
* Extracted them and stored them in a folder with appropriate file naming
* Save another copy of the data sets in a sub folder to keep the original unchanged
* Open the data sets by using MS excel to observe how the data is organized
* I saw there are lots of null values in some fields
* Tried to filter and sort by different criteria
* N.B I used MS excel just to observe the data but I didn’t change the structure or anything as the data is too big it will create loss of data.

**Phase 3:- Process**

* In this phase I realized that I have to chose big query to make data cleaning and organizing of data because the data set is too big which is not able to be handled by excel or spreadsheet.
* I uploaded the four data sets on BIG query by creating a bucket in google cloud

Graphical user interface

Description automatically generated with medium confidence

* I checked the data errors
* Identify which of the data fields are necessary to answer the business task.
* By using query commands, I filtered, cleaned, and organize the data
* Document the cleaning and manipulating process I performed

The second data set which is the quarter 2 data was recorded with different field names than the other data sets. Renaming of those fields by using query command.

Text

Description automatically generated

The data field of gender and birth date is not complete, with lots of empty fields and I decided to not use because they will lead the analysis to bias.

Filter out the data fields data that doesn't have relation with the business task to minimize the the execution time and file size. and I decided to not use because they will lead the analysis to bias.

By using a UNION operator to combine the four data sets

Graphical user interface, text, application, email

Description automatically generated

 our data is ready for analysis

**Phase 4:- Analyze**

Perform different calculations like

* Average duration of ride

Graphical user interface, application

Description automatically generated

Text

Description automatically generated with medium confidence

* Maximum trip duration
* Number of casual riders and members per month, per days

**phase 5:- Share**

I used tableau for data visualization

Chart, line chart

Description automatically generated

* In the first quarter the minimum number of riders are recorded for both Customers and subscribers
* The number Riders become high in the weekend for customers
* The number Riders is in the weekend for Subscribers

Chart, line chart

Description automatically generated

* On working Days, the number of trips are high for Subscribers
* On Weekends, the number of trips are high for Customers
* The number of trips become increasing in spring for both user types

Chart, line chart

Description automatically generated

* The number of trips become the highest on Summer
* On working Days, the number of trips are high for Subscribers
* On Weekends, the number of trips are high for Customers

Chart, line chart

Description automatically generated

* The number of trips become decreasing when autumn comes.

***Trip duration VS number of trips***

Chart, pie chart

Description automatically generated

* The average trip duration for customers is almost 3 times higher than the duration of subscribers.
* Even if the customers made a smaller number of trips, they use bikes for longer duration

***Summary***

* The number of bike riders are highly related to season especially Casual riders are more season dependent.
* In summer, the number of bike riders become higher and on winter it is lower.
* On weekends the duration as well as the number of trips are peaked for casual Riders.
* The members use bike mostly on working days which seems like they are using bike for commuting work.
* The casual riders ride for long duration than the Members.

**Recommendation**

* Prepare a different kinds of plan by targeting casual riders like summertime plan, weekend plan, introducing welcoming free trips for new subscribers and others.
* Better to make further analysis by collecting other data attributes like riders Age, purpose of trips, sex, price available plans, reasons of choosing bikes, and others to address with a best and efficient advertisement campaign.