**Google Data Analytics Capstone**

**Cyclistic Bike Share Analysis** (Power BI, SQL, Python)

**Overview**:

A bike-sharing company named Cyclistic Bikes has 3 pricing plans: single ride passes, a single day pass, and an annual membership pass. This company believes that annual memberships are more profitable than casual riders (single ride & single day passes). Their overall goal is to push sales of their annual membership pass by converting casual riders into members.

**Ask**:

1. How do annual members and casual riders use Cyclistic bikes differently?

**The problem(s) I am trying to solve:**

* In what ways are annual members using bikes differently than casual riders, and how can I help convince casual riders to get an annual membership pass?

These insights can drive business decisions by showing what annual members do differently than casual riders and how we can market more specifically to these casual riders.

**Business task**:

Analyze data to see how annual members data compares to casual riders.

**Stakeholders**:

The key stakeholders of Cyclistic:

* Lily Moreno, the director of marketing.
* Cyclistic Marketing analytics team
* Cyclistic Executive team

**Deliverable**:

To give recommendation(s) to convert Casual members to Annual members.

### ****Prepare:****

**Data Location and Organization:**

I've downloaded all of the Cyclistic CSV data sets from January 2023 to June 2023 and organized them in separate files on my desktop by year. This analysis is to track progress over the half yearly progress, so I have not included the data before this period in my analysis. (<https://divvy-tripdata.s3.amazonaws.com/index.html>)

**Credibility**:

Given that this is first party data collected by Cyclistic bikes themselves, I have no reasons to think that the data is lacking in credibility or has inherited any bias. The data is current, original, and comprehensive.

**Licensing** for use of data:

<https://ride.divvybikes.com/data-license-agreement>

**Integrity**:

All of the CSV files have matching column fields.

**Answering my Question:**

I think the data can be aggregated and trends can dictate where the problem lies.

**Problems with the Data**:

The main problems with the data are numerous missing values in the Station Name and Station ID columns. I'm having a lot of trouble matching these values as there are over 1.7 million observances where numerous observances are missing from both the station name and ID in conjunction.

### ****Process****:

I initially chose Excel to analyze and clean up the data, but the data has in excess of 1.7 million observances, so I decided to switch to MySQL. I tried loading table with Data Import wizard which was really time consuming then I decided to load the tables with MySQL-Python-Connector where it took only minutes to load records. After the cleaning Process I used PowerBI to create a Dashboard to make informed Decisions and to give recommendation on the chances of converting Casual rider to membership rider.

**Loading data:**

**Using Python-MySQL connector**

import mysql.connector  
import csv  
  
mydb = mysql.connector.connect(  
 host='localhost',  
 user='root',  
 password='PASSWORD',  
 port='3306',  
 database='bicycle\_insights'  
)  
  
cursor = mydb.cursor()  
  
file = open('D://Data Analyst Projects/BICYCLE\_PROJECT/cleaned\_data/202306-divvy-tripdata.csv')  
reader = csv.reader(file,delimiter=',')  
next(reader)  
  
for row in reader:  
 cursor.execute('INSERT INTO june (ride\_id,rideable\_type,started\_at,ended\_at,ride\_duration\_min,ride\_duration,day\_of\_week,start\_station\_name,start\_station\_id,end\_station\_name,end\_station\_id,start\_lat,start\_lng,end\_lat,end\_lng,member\_casual) VALUES (%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s)',row)  
  
mydb.commit()  
  
cursor.close()  
  
mydb.commit()  
  
print("rows inserted")

Note:

I have created separate table for each month and loaded the data with corresponding csv files for easy cleaning purpose.

**Cleaning the Data:**

* I have cleaned the data by removing columns with null values that aren't essential to my analysis (station names and station id's).

# delete the null values

delete from bicycle\_insights.january

where

start\_station\_name =''

or

start\_station\_id =''

or

end\_station\_name =''

or

end\_station\_id =''

or

END\_LAT =''

or

end\_lng =''

;

* I have cleaned the records where the ride time is less than one minute since there can’t be a ride less than one minute.

#check for rides within one minute within same station

select count(\*)

FROM bicycle\_insights.january

where trim(RIDE\_DURATION) < time '00:01:00'

and start\_station\_name = end\_station\_name

and start\_lat = end\_lat

and start\_lng = end\_lng;

* I have the loaded clean data to the final table which is connected to the Power BI.

**Transformation in PowerBI:**

* I have created new columns for Day, Month, Average Time to have better analysis of data and visualize to find the trends.
* I have counted the rides with Count (ride\_id).

### ****Analyze****:

In the previous "Process" step, I organized and aggregated my data together in Power BI. This stage is when I took care of any formatting issues as well.

**Trends**:

As I began to look at my data, I started to notice a trend of members riding bike are in more instances than casuals, but casuals on average seemed to be riding for longer period of time.

Rides by "Day of Week" seemed to show a lot of casual riders over the weekend, but a huge drop during the weekdays. Member riders remained fairly consistent throughout the week, although lower than casual riders on the weekend.

Finally, across the months of the year, members also tended to be more consistent with the amount of rides. Casuals really had a dramatic increase in May and June which may be due to festivals that are in those months. This data also led to the conclusion of members being consistent than casual riders across all months.

**Analysis Summary:**

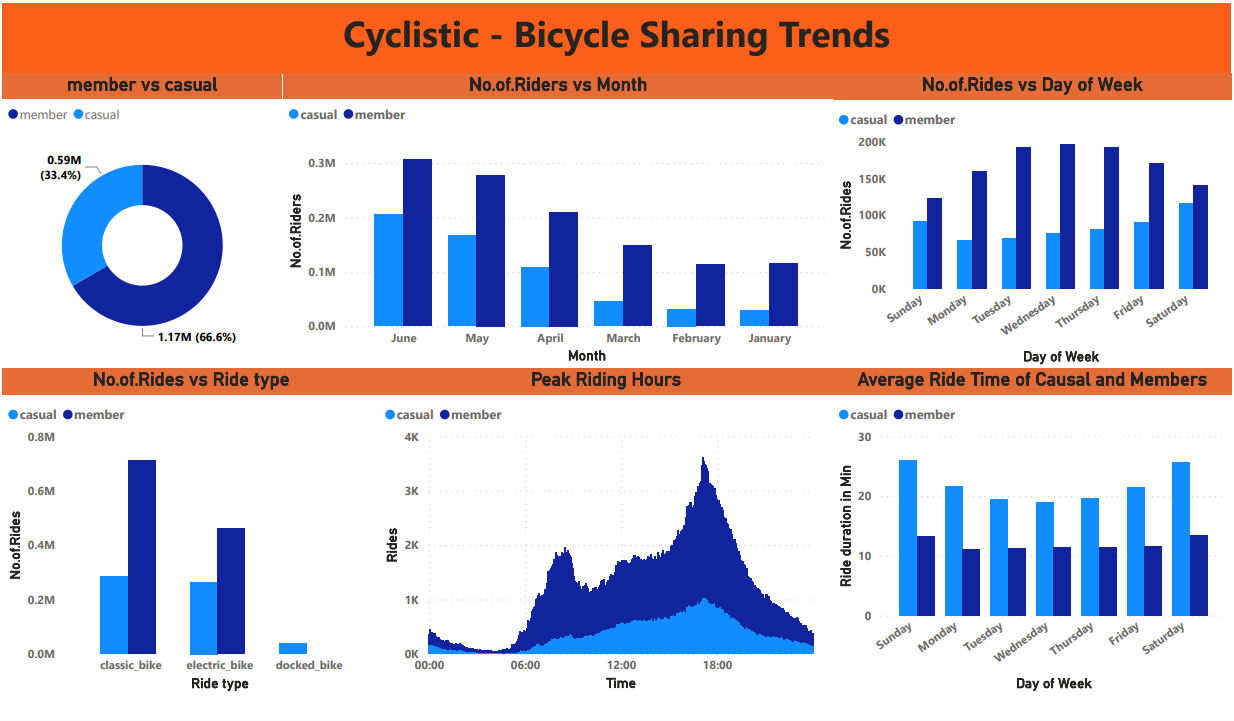
From what I can gather, it seems members are riding more consistently across all time metrics, but riding for shorter periods of time than casual riders.

Casual riders seem to have a lot of rides in the May and June months and on the weekends, but other than that they don't show a lot of consistency. It seems the casual riders prefer to ride for longer periods of time on a single ride, as opposed to members.

### ****Share****:

All of my visualizations are based on metrics of casual vs. member riders so I can see the data that will directly answer my business question.

I have created the dashboard with PowerBI, Here is Snapshot of that.



### ****Act****:

**Conclusion:**

I think Cyclistic should create a marketing campaign to show how much money can be saved using a bike as one's regular means of commuting to a workplace. In addition, we could incorporate the health benefits associated with riding a bike vs driving a car.

Also, Cyclistic could compare the benefits of using a bike to rising fuel costs, the hassle of finding parking, dealing with morning and evening traffic, and the reduction in carbon emissions.

**Deliverable:**

* Creating a Campaign to compare the savings of using bikes vs using cars for commuting.
* Provide membership discounts for casual riders who travel more on weekends without affecting profit margin.
* Can Collaborate with Environmental governance companies to promote e-bikes usage rather than other vehicles.
* Make an app to compete rider with other riders in terms of greener earth points and health dashboards and personal well being.