**Cyclistic Bike Share Analysis Case Study**

This analysis is a capstone project case study in the Google Data Analytic Certificate course through Coursera. In this project a public dataset will be used and the data analysis process (**Ask, Prepare, Process, Analyse, Share, Act**) will be implemented to develop the product.

1. **(ASK) Introduction / Background**

Cyclistic, a bike-share company in Chicago, is operating a bike sharing program that features more than 5,800 bicycles and 600 docking stations. The bikes can be unlocked from one station and returned to any other station in the system anytime.

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. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day. Customers who purchase annual memberships are Cyclistic members. The director of marketing in Cyclistic believes the company’s future success depends on maximizing the number of annual memberships.

This report discusses marketing analyst team’s finding during the course of qualitative analysis of Cyclistic historical bike trip dataset.

The aim of the investigation is to:

* Illustrate the difference between the use of Cyclistic bikes among annual members and casual riders.
* Depict the reason that casual riders buy Cyclistic annual memberships
* Complement data findings to inspire the strategy for Cyclistic to use digital media to target casual riders and convert them into annual members

The key stakeholders in this data analysis are:

* **Lily Moreno:** The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.
* **Cyclistic marketing analytics team:** A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy.
* **Cyclistic executive team:** The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

1. **(PREPARE) Data Collection**

**Identify data source**

Cyclistic’s historical trip data will be used to for data analysis, Click here to access the public dataset –> [link](https://divvy-tripdata.s3.amazonaws.com/index.html)

A screenshot of a phone

Description automatically generated

The data is located on AWS file server where the zip files can be downloaded and named correctly. The data dated from April 2023 to March 2024 was downloaded and stored locally on desktop for the data cleaning and further analytical processes. The data is current that the usefulness of data is secured. The data is indexed by month (2023-2024) .

The public data has been made available by Motivate International Inc. under license (<https://ride.divvybikes.com/data-license-agreement>). The downloaded original dataset from Cyclistic is reliable - accurate and comprehensive records with complete and unbiased information and it proven fit for use.

Note: Data-privacy issues prohibit data-user from using riders’ personally identifiable information. All rider personal information is hidden or kept private to Cyclistic only. Data users won’t be able connect the pass purchases to credit card numbers to o determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

**Filename used from the server**

202304-divvy-tripdata.zip, 202305-divvy-tripdata.zip, 202306-divvy-tripdata.zip,

202307-divvy-tripdata.zip, 202308-divvy-tripdata.zip, 202309-divvy-tripdata.zip,

202310-divvy-tripdata.zip, 202311-divvy-tripdata.zip, 202312-divvy-tripdata.zip,

202401-divvy-tripdata.zip, 202402-divvy-tripdata.zip, 202403-divvy-tripdata.zip

**Determine the credibility and integrity of the data - ROCCC Analysis**

**Reliability:** This data is reliable. There is a huge sample size (> 5 million of trip record per year) has been used

**Originality:** This is the original dataset as collected and made available by Motivate International Inc.

**Comprehensiveness:** This data is comprehensive. There is information about each ride trip: member status of bike-user, ride date and time, starting location and ending location of trip, duration of bike ride. The data is unbiased and randomised where all other data is not covered: gender, age, race etc.

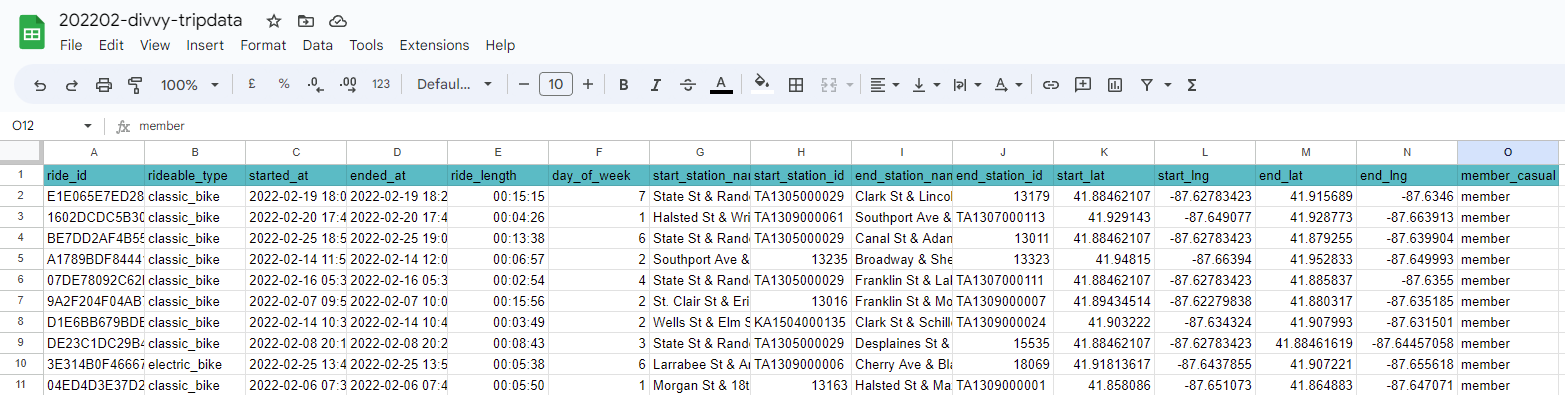
**Current:** This data was collected in 2023, The data is even currently updated monthly. The dataset covered a whole year record from March 2023- April 2024, where this report is drafted in 2024.

**Cited:** Yes

The data integrity and credibility is clearly insufficient to provide relevant insights to Cyclistic. Therefore, the following data processing and analysis will be done accordingly to provide hints and directions to the key questions in this case study.

1. **(PROCESS) Data Structure and Data Cleaning**

The format of each data file is CSV. All 12 .csv files have the same set of column names (attributes):



* **ride\_id,** character, unique primary key for each ride trip
* **rideable\_type,** character, the type of bike used in bike trip
* **started\_at,** time-date data, the starting date and time for each bike trip
* **ended\_at,** time-date data, the ending date and time for each bike trip
* **start\_station\_name,** character, starting station name for the bike trip
* **start\_station\_id,** character, ID used for identifying starting station
* **end\_station\_name,** character, ending station name for the bike trip
* **end\_station\_id,** character, ID used for identifying ending station
* **start\_lat,** geographic data, latitude of starting station of the bike trip
* **start\_lng,** geographic data, longitude of starting station of the bike trip
* **end\_lat,** geographic data, latitude of ending station of the bike trip
* **end\_lng,** geographic data, longitude of ending station of the bike trip
* **member\_casual,** character, the member status of bike user, fell into only casual (user) and member

**Data importing**

The data processing and part of data analysis will be done in R. The dataset files has been imported into R and renamed

A blue and white text on a white background

Description automatically generated

**Wrangle data and combine into a single file**

The column names of all datasets are examined, compared. When the names in the columns in datasets was not matching perfectly, the column names needs to be renamed before joining datasets into a single file.

The column names of all data sets are checked by R command.

Using colname () to find out the column name if all datasets are coherent.

The column in the dataset file is then examined.

Using compare\_df\_cols to compare the data types held by each column, if they are all coherent in all files

Since the column name for all dataset is uniform, the datasets are then combined into a single dataset file, bike\_rides



using rbind() to combine all 12 dataset files into a single dataset file

The head (first 6 rows) of the combined dataset is then examined using head() to show first 6 rows of records in the combined dataset.

A number of numbers on a white background

Description automatically generated

The dimension of dataset is examined



using dim() to show the number of rows (5750177) and number of columns (15) in the dataset.

The str function in R is used to display the internal structure of an R object. When you call str(bike\_rides), it provides a concise summary of the bike\_rides data frame, including information about the data types and a preview of the data in each column.

A screenshot of a computer program

Description automatically generated

1. **Clean up and add data to prepare for analysis**

The `bike\_rides` dataset was inspected using functions to view column names, row count, dimensions, first few rows, data structure, and statistical summary, ensuring data integrity and understanding.

A close-up of a text

Description automatically generated

The columns with missing values was removed from the `bike\_rides` dataset, specifically `start\_station\_name`, `start\_station\_id`, `end\_station\_name`, and `end\_station\_id`, to ensure data quality and relevance.



Now we added columns to the bike\_rides dataset to extract and list the date, month, day, year, and day of the week for each ride. This enables us to aggregate ride data at different temporal levels, enhancing our ability to analyze trends over time.

A computer code with text

Description automatically generated with medium confidence

The "ride\_length" column was added to the bike\_rides dataset to calculate the duration of each ride in seconds. This calculation subtracts the started\_at timestamp from the ended\_at timestamp using the difftime function.

Calculate Ride Length: The difftime function computes the time difference between ended\_at and started\_at, storing the result in the new column ride\_length.

bike\_rides$ride\_length <- difftime(bike\_rides$ended\_at, bike\_rides$started\_at)

Inspect Structure: The str function displays the structure of bike\_rides to verify the addition of ride\_length and its data type.

str(bike\_rides)

To enable calculations on ride durations, the "ride\_length" column in the bike\_rides dataset was converted from a factor to numeric format. This transformation allows for quantitative analysis and statistical computations on ride durations.

Check Data Type: Initially, we confirmed whether the "ride\_length" column was stored as a factor using is.factor.

is.factor(bike\_rides$ride\_length)

Convert to Numeric: To perform calculations, the column was converted from factor to numeric using as.numeric(as.character(...)). This ensures the values are treated as numeric data.

bike\_rides$ride\_length <- as.numeric(as.character(bike\_rides$ride\_length))

Verify Numeric Type: We used is.numeric to confirm that "ride\_length" was successfully converted to numeric format.

is.numeric(bike\_rides$ride\_length)

Inspect Structure: Finally, the str function was used to review the structure of bike\_rides, ensuring the "ride\_length" column is now numeric.

str(bike\_rides)

A screenshot of a computer code

Description automatically generated

To enhance data quality, a new version (bike\_rides\_v2) of the dataset was created by removing "bad" data entries where ride durations were negative or bikes were checked for quality or maintenance. This process ensures that only valid ride records are retained for accurate analysis and insights.

bike\_rides\_v2 <- bike\_rides[!(bike\_rides$ride\_length<0),]

str(bike\_rides\_v2)

dim(bike\_rides\_v2)

A screenshot of a computer program

Description automatically generated

1. **(ANALYSE) Data analysis of the trip data**

Statistical summary of the variables in dataset is created

* **Compare count of rides among member and casual users**

A close-up of a number

Description automatically generated

R script to summarise number of ride: Member vs Causal users

A screenshot of a computer

Description automatically generated

Summary of Number of ride: Member vs Causal users



R script ggplot () to generate plot for User count of memnber and casual users

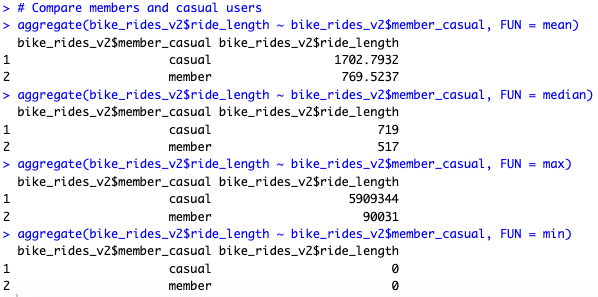
A graph with a red and blue rectangular bar

Description automatically generated

Plot to show number of rides between casual and member users by ggplot

**Finding:** There are more rides from member users (approx. 64%) than the causal users (approx. 36%)

* Comparison between member and casual users



According to the above statistics, the average ride length (in secs) for casual users is significantly greater than for member users. The Average ride length for casual user is 1702.8 secs , whereas for member users it is 769.6 secs

* Compare average ride and No of time by each day for members vs casual users

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

According to the above statistics, Monday - Friday, Casual customers are less likely to ride bikes in comparison to Member customers. However, on Saturday and Sunday, Casual customers are more privy to ride bikes. The below visualization provide these insights.

A graph of blue and red bars

Description automatically generated

The average duration during the week is higher for Casual customers in comparison to Member customers.

A graph of different colored bars

Description automatically generated

R script to summarise average trip duration (avg\_ride\_length) between casual and member users

A screenshot of a computer program

Description automatically generated

A graph of a number and a bar

Description automatically generated with medium confidence

Plot to show average trip duration between casual and member users

**Finding:** Longer trip duration by casual users than the member users

* Compare bike type used by casual user and member users

A computer code with blue text

Description automatically generated

R script to summarise bike type usage between casual and member users

A close-up of numbers

Description automatically generated

R script to plot bike type used by percentage between casual and member users



A screenshot of a graph

Description automatically generated

Plot to show bike type used by percentage between casual and member users

**Findings:**

1. Docked bikes are the least popular type of bike to be used
2. Casual user: Electrical bike is more popular to be used than classical bike
3. Member user: Classical bike is nearly equally popular as electric bikes, member users never use docked bikes

* **Ride count & Average Ride Length (Weekdays, casual users vs member users)**

A graph of blue and red bars

Description automatically generated

A graph of different colored bars

Description automatically generated

**Findings:**

1. Casual users: more trips in the weekend than in weekdays (Monday to Friday), longer trip duration in weekend than in weekdays
2. Member users: More trips and shorter trip duration in the weekday than in weekend, and the trip duration is relatively stable throughout days of a week.

* **Ride count & Average Ride Length (Yearly, casual users vs member users)**

A graph with red and blue lines

Description automatically generated

A graph with red line and blue line

Description automatically generated

**Findings:**

1. Summer time (May to Sep) is both promising in terms of ride count and trip duration. High usage remains throughout the third quarter with a gradual decline until year end. Both ride count and trip duration are more than that of the rest months of the year.
2. Casual users has more trip duration but member user have more rides during summer time.
3. **(SHARE / ACT) Recommendation from the findings**

* **Reason that casual riders buy Cyclistic annual memberships**

It is found that member users are steroeotyped stable users to use bike at commuting hours (morning and mainly evening). The convenience to use bike on roads with heavy traffic in the commercial areas hints the subscription.

The casual users are so-called users who use bike for casual reasons: sightseeing or leisure. They have peak season of summertime, peak usage on weekend and off-rush hour.

We can conclude that member users and casual users are quite different group of people and hence it is unlikely to convince casual users to pick up member user using pattern.

Besides membership should base on casual users' preference to secure the subscription rate.

* **Strategy for Cyclistic to use digital media to target casual riders and convert them into annual members**

To convert casual riders into annual members, the following marketing strategies can be implemented:

1. Offer occasional membership subscription discount to casual riders on summer and weekends
2. Put banners or special discount advertisements at tourist attractions that would target casual users which might influence them to become members
3. Push Cyclistic's advertisement online during the weekends to target casual users
4. Encouraging casual users to have membership by lowering the renting price of bikes (member's deal) for weekend.