**Bike Share Case Study**

**Case Background:**  
  
In this case study, the business task is to analyze historical bike trip data of Cyclistic, a bike-share company in Chicago, to understand how casual riders and annual members use Cyclistic bikes differently. Cyclistic operates a successful bike-share business with 5,824 bicycles and 692 stations in Chicago. The company offers two membership types:

* **Casual Riders:** These are customers who purchase single-ride or full-day passes, indicating they use the service infrequently or on a pay-as-you-go basis.
* **Annual Members:** These are subscribers who pay for annual memberships, providing them with unlimited access to the bikes throughout the year.

The primary objective is to gain insights from the data that will help the marketing team develop a new strategy to convert casual riders into annual members. To do this, the analysis will likely involve examining various aspects of bike usage by both types of customers, including ride frequency, trip duration, and popular riding days. The findings can be used to tailor marketing strategies that encourage casual riders to become annual members and thereby increase the company's recurring revenue.

**About the Company:**

Cyclistic launched its bike-share program in 2016, which has since grown to include 5,824 GPS-tracked bikes across 692 stations in Chicago. Riders can pick up a bike from one station and return it to any other within the network, offering flexibility and convenience.

Cyclistic's marketing strategy has traditionally focused on raising awareness and attracting a broad customer base with flexible pricing options, including single-ride passes, full-day passes, and annual memberships. Casual riders purchase single-ride or full-day passes, while annual memberships offer more value and consistency.

Financial analysis shows that annual members are more profitable than casual riders. With future growth in mind, marketing lead Moreno believes the key is converting casual riders into annual members. Casual riders are already familiar with Cyclistic and use it for their mobility needs, making them a prime target for membership.

Moreno’s goal is to develop marketing strategies focused on this conversion. To support this, the marketing team is working to understand the differences between casual riders and members, explore what would motivate casual riders to switch, and assess how digital media can enhance these efforts. They’ll be analyzing historical bike trip data to uncover trends that can guide their approach.

**Google Data Analysis Process:**

The Google Data Analytics Professional Certificate uses a methodology for data analysis which involves 6 phrases: ASK, PREPARE, PROCESS, ANALYSE, SHARE AND ACT

**Phrase 1: Ask – Understanding the project and problem that needs solving**

**Three questions that will guide the marketing project:**

* 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?

**Business Task: Define the Problem:**

The challenge at hand is to create a data-driven marketing campaign that will effectively convert casual riders of Cyclistic, a bike-share company in Chicago, into annual members. The future success of Cyclistic heavily relies on maximizing the number of annual memberships. The director of marketing, Lily Moreno, envisions a comprehensive strategy that not only attracts new customers but also capitalizes on the existing awareness of the Cyclistic program among casual riders.

**Stakeholders:**

* **Lily Moreno:** Director of Marketing
* **Cyclistic executive team:** The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.
* **Cyclistic Marketing Analytics Team:** who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy.

**Deliverables for the project:**

1. A clear statement of the business task

2. A description of all data sources used

3. Documentation of any cleaning or manipulation of data

4. A summary of the analysis

5. Supporting visualizations and key findings

6. Top three recommendations based on the analysis

**Phrase 2: Prepare – Extract and prepare the data for analysis**   
  
**Data Acquisition and Verification:**

* Data Source: Obtain Cyclistic historical bike trip data from the official source: [Cyclistic Historical Data](https://divvy-tripdata.s3.amazonaws.com/index.html).
* Data License: Ensure data availability under the license agreement provided by Motivate International Inc.: [Data License Agreement](https://divvybikes.com/data-license-agreement).
* Public Accessibility: Verify that the data is publicly accessible.

**Data Integrity Verification:**

* Data Duration: Utilize a 12-month dataset spanning from August 2022 to July 2023 to identify trends.
* ROCCC Analysis:
  + Reliable: Confirm that the dataset is unbiased and reliable.
  + Original: Ensure that the dataset used in the analysis is original and publicly accessible.
  + Comprehensive: Check that the dataset is comprehensive and free from significant missing or incomplete information.
  + Current: Verify that the data is up to date and updated on a monthly basis.
  + Citation: Confirm that the dataset is appropriately cited.

**Data Extraction and Organization:**

* Data Download: Download 12 CSV files, each covering a specific month from August 2022 to July 2023. Store these files on your computer for analysis.
* Data Variables: Identify that each CSV file contains the same 13 variables, including Distinct Ride ID, bike types, station information (names, IDs, longitude, and latitude), member/casual rider status, and ride start and end times.
* Data Organization: Organize the data by grouping it by month and year to facilitate analysis.
* Data Filtering and Sorting: Filter and sort the data to identify patterns and relationships related to ride frequency and ride duration, which are essential for understanding membership adoption.

**Insight Identification:**

* Consider explaining common factors among memberships, such as ride frequency and ride duration, to better understand how these behaviors relate to membership adoption.
* Plan to present insights that suggest correlations between usage patterns and annual membership purchases.

This organized process will ensure that the data is ready for in-depth analysis and deriving valuable insights for Cyclistic's business needs.

**Process -** Deliverable Documentation of any cleaning or manipulation of data

During the data preparation phase, I used MS Excel 2016 as my chosen tool for data cleaning and manipulation. The data was obtained from Cyclistic's internal records, which are maintained with a high degree of credibility and integrity. However, it's important to note that there may be inherent limitations and biases in the data, as it primarily represents user interactions within the Cyclistic bike-sharing system.

To ensure the integrity of the data, I thoroughly understood its source, generation process, and potential limitations. I maintained a rigorous approach by tracking any changes made during the data preparation process and keeping backups of the original data, ensuring that we had a clear audit trail.

The initial data inspection revealed a few issues that required attention. These issues included errors in data entry, and inconsistent data formatting. Addressing these issues was critical to obtaining accurate and reliable insights from the data.

In particular, I made a significant decision during the data cleaning and preprocessing phase, which directly relates to our focus questions. To understand how annual members and casual riders use Cyclistic bikes differently (Focus Question 1), I decided to exclude riders with less than 1 minute of recorded riding time from our analysis. This decision was driven by practical considerations related to the operational dynamics of bike-sharing systems. We wanted to focus on meaningful bike rides that represent typical user behaviors, as this would help us uncover why casual riders might choose to purchase Cyclistic annual memberships (Focus Question 2).

Overall, the data preparation process was essential in ensuring that the data was accurate, consistent, and aligned with the objectives of addressing our focus questions. It allowed us to obtain a clean and reliable dataset that we could use to analyze how annual members and casual riders differ in their usage patterns and explore the motivations behind casual riders purchasing Cyclistic annual memberships. Additionally, the cleaned data formed the basis for our subsequent analysis of how Cyclistic could use digital media to influence casual riders to become members (Focus Question 3).

**Excel:**

1. Added ride\_length formula: consisting of the time difference between Start\_at and ended\_at
2. Deleted all records under 1 minute (see explanation in appendix).
3. Check for Duplications: No Duplications
4. Remove unnecessary columns to reduce file size: start\_lat, start\_lng, end\_lat, end\_lng, started\_stationID, and ended\_StationID

(See Appendix for Pivot Tables and other advance features)

**Analyze - A summary of your analysis -** Deliverable summary of your analysis

**Excel -** I organize the data in several ways to analyze different trends. (see appendix B for a printout of the summary sheet:

**Total Riders**

Mean of rider length  
Max ride length  
Mode of day of week

Max ride length by member, casual riders, and by total riders  
Minium rider length by member, casual riders, and by total riders

Average ride length per bike type for member vs casual

By rider grouped by member, casual, and total rider per:

* + By Weekday
  + By Month
  + By Season

By ride grouped by bike type for Member, Casual and total riders for each:

* Classic bike
* Docked bike
* Electric bike

**Day of week**  
Average ride length per day of week for member, casual riders, and by total riders

Average ride length for member, casual riders per by bike type

**Column Formatting:**

started\_by and ended\_by were formatted as Datetime M:DD:YYYY, h:mm:ss

**Data trends:**   
Analyzing the data for each day of the week, we find a remarkable pattern: while members constitute 55% of the total riders, casual riders, who make up the remaining 45%, consistently have an average daily ride duration that is twice as long as that of members. This trend holds true for each day of the week, highlighting the significant contrast in ride durations between casual riders and members.

**Notable Trend: Casual Riders Take Longer Trips**

Our analysis shows an interesting trend: although annual members make up 55% of total riders, casual riders, who account for 45%, tend to take longer trips. On average, casual riders' trips are twice as long as those of members, regardless of the day of the week. This suggests that casual riders often opt for longer, more leisurely rides, while members use the service for shorter, more frequent trips.

**Relationship:**Casual Riders Are a Key Opportunity for Conversion Casual riders, though using the service less often, tend to enjoy longer rides. This suggests they might be more inclined to convert to annual memberships if they can see the value in terms of convenience and cost savings for their extended trips.

**How are the three guiding questions answered?**

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

Casual riders take longer but less frequent trips, while members typically use the bikes for

shorter, more frequent commutes or errands.

* 1. **Why would casual riders be interested in annual memberships?**

Casual riders might benefit from memberships by enjoying unlimited rides without worrying about additional costs, especially since their rides tend to be longer.

* 1. **How can Cyclistic use digital media to encourage casual riders to become members?**

By using the data to personalize ads that highlight the perks of membership—such as cost savings on longer rides and more flexibility—Cyclistic can attract casual riders who already appreciate the service but haven’t committed to a membership.

**RStudio   
installing and loading necessary packages:**

install.packages("tidyverse")

install.packages("ggplot2")

install.packages(“dplyr”)

install.packages(“janitor”)

install.packages(“lubridate”)

library(tidyverse)

library(ggplot2)

library(dplyr)

library(janitor)

library(lubridate)

**# Import 12 monthly files**

df1 <- read.csv("202208-divvy-tripdata.csv")

df2 <- read.csv("202209-divvy-tripdata.csv")

df3 <- read.csv("202210-divvy-tripdata.csv")

df4 <- read.csv("202211-divvy-tripdata.csv")

df5 <- read.csv("202212-divvy-tripdata.csv")

df6 <- read.csv("202301-divvy-tripdata.csv")

df7 <- read.csv("202302-divvy-tripdata.csv")

df8 <- read.csv("202303-divvy-tripdata.csv")

df9 <- read.csv("202304-divvy-tripdata.csv")

df10 <- read.csv("202305-divvy-tripdata.csv")

df11 <- read.csv("202306-divvy-tripdata.csv")

df12 <- read.csv("202307-divvy-tripdata.csv")

**# Merge all 12 files into one called "bike\_rides"**

Bike\_rides <- rbind(df1, df2, df3, df4, df5, df6, df7, df8, df9, df10, df11, df12)

**Data Information in RStudio**

nrow(bike\_rides) #563615

ncol(bike\_rides) #9

head(bike\_rides) #see the first 6 rows of the data frame

tail(bike\_rides) #see the last 6 rows of the data frame

str(bike\_rides) #see list of columns and data types(d

summary(bike\_rides) #statistical summary of data

colnames(bike\_rides) #list of column names

# Column names

[1] "ride\_id"   
[2] "rideable\_type"

[3] "started\_at"   
[4] "ended\_at"

[5] "start\_station\_name"   
[6] "end\_station\_name"

[7] "member\_casual"

[8] “ride\_length”

[9] “weekday”

Dim(Bike\_rides) Results: 563615 rows, 9 columns

Export into one file for Bigquery and tableau

**SQL – In Big Query**

**# Union of table with total record count**

SELECT COUNT(\*) AS TotalRecords

FROM (

    SELECT A.\*, B.\*

    FROM `da-bike-share-final-project.All\_bikeshare\_Data.2022\_tripdata` AS A

    LEFT JOIN `da-bike-share-final-project.All\_bikeshare\_Data.2023\_tripdata` AS B

    ON A.ride\_id = B.ride\_id

    WHERE COALESCE(A.ride\_id, '') <> ''

    UNION ALL

    SELECT A.\*, B.\*

    FROM `da-bike-share-final-project.All\_bikeshare\_Data.2022\_tripdata` AS A

    RIGHT JOIN `da-bike-share-final-project.All\_bikeshare\_Data.2023\_tripdata` AS B

    ON A.ride\_id = B.ride\_id

    WHERE COALESCE(B.ride\_id, '') <> ''

);

**#Results: 2,817,480   
In Big Query: All Records 1 minute and over 5,634,868**

**# Join tables to determine Max, Min, and Avg ride lengths over all**

WITH All\_RideShares AS (

  SELECT \*

  FROM `da-bike-share-final-project.All\_bikeshare\_Data.2022\_tripdata` AS A

  UNION ALL

  SELECT \*

  FROM `da-bike-share-final-project.All\_bikeshare\_Data.2023\_tripdata` AS B

)

SELECT

  TIME(TIMESTAMP\_SECONDS(MAX(TIMESTAMP\_DIFF(ended\_at, started\_at, SECOND)))) AS Max\_Length,

  TIME(TIMESTAMP\_SECONDS(MIN(TIMESTAMP\_DIFF(ended\_at, started\_at, SECOND)))) AS Min\_Length

SELECT  
 TIME(TIMESTAMP\_SECONDS(CAST(AVG(TIMESTAMP\_DIFF(ended\_at, started\_at, SECOND)) AS INT64))) AS AVG\_Length   
FROM All\_RideShares;

**#Results:** Max\_Length: 17:42:00 Min\_Length: 10:00 Avg\_Length: 31:41 (formatted as h:mm:ss)

**# Join Tables then determine day of the week from started\_at field.** **Grouped total\_rides in Desc order to determine the most popular weekday for riding.**

WITH All\_RideShares AS (

  SELECT \*,

         EXTRACT(DAYOFWEEK FROM started\_at) AS day\_of\_week

  FROM `da-bike-share-final-project.All\_bikeshare\_Data.2022\_tripdata` AS A

  UNION ALL

  SELECT \*,

         EXTRACT(DAYOFWEEK FROM started\_at) AS day\_of\_week

  FROM `da-bike-share-final-project.All\_bikeshare\_Data.2023\_tripdata` AS B

)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SELECT    CASE      WHEN day\_of\_week = 1 THEN 'Sunday'      WHEN day\_of\_week = 2 THEN 'Monday'      WHEN day\_of\_week = 3 THEN 'Tuesday'      WHEN day\_of\_week = 4 THEN 'Wednesday'      WHEN day\_of\_week = 5 THEN 'Thursday'      WHEN day\_of\_week = 6 THEN 'Friday'      WHEN day\_of\_week = 7 THEN 'Saturday'      ELSE 'Unknown' -- Handle any unexpected values    END AS day\_of\_week\_text,    COUNT(\*) AS total\_rides  FROM All\_RideShares  GROUP BY day\_of\_week  ORDER BY total\_rides desc; | | Row | day\_of\_week\_text | total\_rides | | --- | --- | --- | | 1 | Unknown | 680517 | | 2 | **Saturday** | **507840** | | 3 | **Friday** | **417991** | | 4 | Thursday | 398892 | | 5 | Sunday | 395455 | | 6 | Wednesday | 376573 | | 7 | Tuesday | 369664 | | 8 | Monday | 351065 | |

**# Results: The populator Bike riding day is Saturday follow by Friday**

#

WITH All\_RideShares AS (

  SELECT \*,

         EXTRACT(DAYOFWEEK FROM started\_at) AS day\_of\_week

  FROM `da-bike-share-final-project.All\_bikeshare\_Data.2022\_tripdata` AS A

  UNION ALL

  SELECT \*,

         EXTRACT(DAYOFWEEK FROM started\_at) AS day\_of\_week

  FROM `da-bike-share-final-project.All\_bikeshare\_Data.2023\_tripdata` AS B

)

SELECT

  CASE

    WHEN day\_of\_week = 1 THEN 'Sunday'

    WHEN day\_of\_week = 2 THEN 'Monday'

    WHEN day\_of\_week = 3 THEN 'Tuesday'

    WHEN day\_of\_week = 4 THEN 'Wednesday'

    WHEN day\_of\_week = 5 THEN 'Thursday'

    WHEN day\_of\_week = 6 THEN 'Friday'

    WHEN day\_of\_week = 7 THEN 'Saturday'

    ELSE 'Unknown' -- Handle any unexpected values

  END AS day\_of\_week\_text,

  member\_casual,

  TIME(TIMESTAMP\_SECONDS(CAST(AVG(TIMESTAMP\_DIFF(ended\_at, started\_at, SECOND)) AS INT64))) AS AVG\_time

FROM All\_RideShares

GROUP BY day\_of\_week\_text, day\_of\_week, member\_casual

ORDER BY day\_of\_week;

**#Results:** Average Bike ride per day of the week - Showing Casual Riders – ride 2x as long compare to Members

**Share - Supporting visualizations and key findings**

**Act - Your top three recommendations based on your analysis**

**Added Note:**

**User Bike Preferences Analysis**

The data shows that electric bikes are popular with both members and casual users, likely because of their convenience and speed. Classic bikes tend to be more popular among members, possibly due to cost savings. On the other hand, docked bikes are used only by casual users, which might mean that members prefer the flexibility of undocked bikes. These insights offer valuable guidance for improving bike placement, adjusting pricing, and customizing marketing to better cater to the preferences of each user group.

**Appendix:**

Detail explanation for deleting riders with less than 1 minutes of riding time as follows:

**Data Cleaning and Preprocessing:**

As part of the data cleaning and preprocessing phase, we made the decision to exclude riders with less than 1 minutes of recorded riding time from our analysis. This decision was driven by practical considerations related to the operational dynamics of bike-sharing systems and the need to accurately represent meaningful bike rides in our analysis.

**Rationale for Exclusion:**

The choice to exclude riders with less than 1 minutes of recorded riding time is based on several key considerations:

**Check-in and Check-out Processes:**

In a bike-sharing system, riders must perform essential actions such as locating a bike, checking it out, and returning it to a docking station and checking it back in. These processes inherently introduce a minimum time threshold for a bike ride, which is often not accurately reflected in the dataset when ride durations are less than 1 minutes.

**Realistic Bike Usage:**

From an operational standpoint, a bike ride that lasts less than 1 minutes is frequently not a realistic or meaningful bike-sharing experience. Such short durations may not adequately capture genuine bike rides but could represent transient interactions with the bikes, such as brief stops or users testing the bike's functionality.

**Data Quality Assurance:**

The inclusion of extremely short rides can introduce noise and anomalies into our analysis. These brief rides may be the result of system glitches, erroneous data entries, or users who briefly interact with the bikes without engaging in substantial rides. By excluding them, we aim to enhance the quality and reliability of our dataset.

**Practical Insights:**

For our primary objective of understanding how annual members and casual riders utilize Cyclistic bikes differently, it is imperative to focus on rides that mirror authentic bike-sharing experiences. Including extremely short rides may distort our findings and limit our ability to derive actionable insights for our marketing strategy.

**Data Integrity and Validity:**

It is important to underscore that the decision to exclude riders with less than 1 minutes of riding time was taken to bolster the integrity and validity of our analysis. By removing entries of this nature, we aim to ensure that our conclusions are firmly grounded in realistic bike-sharing experiences, ultimately enabling us to provide Cyclistic executives with robust and actionable recommendations for the marketing program.

**Appendix B**  
Excel Summary sheet:

****