**Cyclistic Bike- Share Case Study**

**(Google Data Analytics Capstone Project)**

By

Christopher Achubie

**Introduction**

Welcome to my Cyclistic Bike-Share Analysis Case Study!

As part of the Google Data Analytics Professional Certificate, I performed various tasks, as a Junior Data Analyst for Cyclist, a fictional bike-sharing company in Chicago.

Prior to 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.

According to the cyclistic’s finance analysts, annual members are much more profitable than casual riders. Lily Moreno, the director of marketing believes the company’s future success and key to future growth depends on maximizing the number of annual memberships. 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.

The procedure and approach I used in this case study is the five (5phases) of data analysis which are;

ASK, PREPARE, PROCESS, ANALYZE and SHARE.

**Phase 1: Ask**

**Defining the Problem**

Annual members are key to future growth, but an enormous number of riders are not annual members.

**Business Task**

* The marketing analyst team needs to better understand how annual members and casual riders differ.
* Why casual riders would buy a membership.
* How digital media could affect their marketing tactics
* Analyzing the Cyclistic historical bike data to identify trends.

**Key Stakeholders**

* Cyclistic Executive Team
* Director of Marketing, Lily Moreno
* Cyclistic Marketing Analytics Team

**Phase 2: Prepare**

**Preparing Data for Exploration**

* **Source of Data:** The Cyclistic historical trip data used for this case study is obtained from

[Index of bucket “divvy-tripdata”](https://divvy-tripdata.s3.amazonaws.com/index.html) is a public data made available by Motivate International

Inc.under this license, [Data License Agreement | Divvy Bikes](https://ride.divvybikes.com/data-license-agreement). And since Cyclistic is a fictional

company, the datasets have different names.

* **Data Organization:** The Cyclistic trip data for the previous 12 months (August 2021- July 2022) was downloaded and unzipped as 12 .csv files. The data is structured in 13 variables (columns)
* **Data Credibility:** The data appears to be intact when scanned through, although there are some missing values. We can also check for the credibility of the data using a process called **ROCCC.**

A credible data **ROCCCs** if it is **R**eliable, **O**riginal, **C**omprehensive, **C**urrent and **C**ited.

* **R**eliable**:** I think the data is reliable, although it’s not unclouded if it was vetted.
* **O**riginal**:** The data is validated with the original source, so it is original.
* **C**omprehensive: I don’t think the data is entirely comprehensive because some information

are missing.

* **C**urrent: Data for the previous 12months is used for this study.
* **C**ited: The source of the data is cited

**Phase 3: Process**

**Data Cleaning**

In this phase we will need to clean up data to get rid of any possible errors, inaccuracies, or

inconsistencies, removing repeated entries and checking as much as possible for bias in the

data.

The tools used for the cleaning and manipulating of this data are **Microsoft Excel** and

**R Programming.**

I used Excel for initial cleaning and processing of the csv. files, checked for duplicates,

mismatched data types, extra spaces using the “**TRIM** function”

and renamed the columns meaningfully (columns renamed below), and then imported

the cleaned datasets to R.

* ride\_id = trip\_id
* rideable\_type = bike\_id
* started\_at = start\_time
* ended\_at = end\_time
* start\_station\_name = from\_station\_name
* start\_station\_id = from\_station\_id
* end\_station\_name = to\_station\_name
* end\_station\_id = to\_station\_id
* member\_casual = user\_type

**Data Integrity**

Data Integrity is the accuracy, completeness, consistency, and trustworthiness of data

throughout its lifecycle. This is an important phase when processing data from dirty

to clean. Taking into cognizance that data integrity can be compromised, I ensured carefulness

when replicating, transferring, and manipulating the data.

**Data Aggregation**

This is the process of gathering data from multiple sources, in order to combine it into a single

summarized collection. With this data we can identify trends, make comparisons, and gain insights.

The twelve (12) datasets were imported into R.