Data: 09/08/2021 – Ferrara

Google Data Analytics Certificate – Capstone

Case Study – How Does a Bike-Share Navigate Speedy Success?

# Introduction

This is the last part of the Google Data Analytics Certificate that I’ve made in Coursera platform. In this case study, I will perform many real-world tasks of a junior data analyst. I will work for a fictional company, *Cyclistic*, and meet different characters and team members. In order to answer the key business questions, I will follow the steps of the data analysis process: **ask, prepare, process, analyze, share,** and **act**.

# Scenario

The ***Cyclistic*** is a bike-share company in Chicago. The director of marketing believes the company’s future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand **how casual riders and annual members use Cyclistic bikes differently**. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, *Cyclistic* executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

*Cyclist* is a bike-share program that features more than 5,800 bicycles and 600 docking stations. *Cyclistic* sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can’t use a standard two-wheeled bike. Most riders opt for traditional bikes; about 8% of riders use the assistive options. *Cyclistic* users are more likely to ride for leisure, but about 30% use them to commute to work each day.

**Lily Moreno** is 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.

The Cyclistic **Marketing Analytics team** is a team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide *Cyclistic* marketing strategy.

The Cyclistic **Executive team** is the notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

## 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.

Until 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**.

*Cyclistic’s* finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, the director of marketing believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, she 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 marketing director has set a clear goal: 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.

# Ask

Three questions will guide the future marketing program:

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?

My goal is to answer the first question: How do annual members and casual riders use Cyclistic bikes differently?

## Business Task

Analyze the trip data from the last 12 months in order to identify the differences between the way annual members and casual riders use the Cyclistic service with the objective of maximize the number of annual members.

# Prepare

I will use Cyclistic’s historical trip data to analyze and identify trends. The datasets have a different name because Cyclistic is a fictional company. For the purposes of this case study, the datasets are appropriate and will enable me to answer the business questions. The data has been made available by ***Motivate International Inc.*** under this license.

Specifically, the spreadsheets that will be used for this case-study comes from July/2020 to June/2021

* **202007-divvy-tripdata.csv ~ 202106-divvy-tripdata.csv**

The data is located in the platform AWS, a cloud computing service of Amazon. It can be found here: <https://divvy-tripdata.s3.amazonaws.com/index.html>

This is a Structured Data, that is the data is organized in rows and columns in Spreadsheets.

This is a First-Party data, that is the data is collected by an individual or group (Motivate International In.) using their own resources. This confirms that the data is clean, accurate, relevant, and timely.

To ensure the data’s integrity, first I downloaded the previous 12 months of Cyclistic trip data and opened all files to confirm that the data were equivalent – that is, all the files have the same columns and type of data. After that, I merge all the files in only one TXT file.

I loaded the file called **“202007-202106-divvy-tripdata.txt”** in ***R Studio*** with the name **tripdata\_raw**. The final file had **4.460.151** observations (rows) and 13 variables (columns).

I created a smaller dataset called **tripdata\_1** as a random sample of **tripdata\_raw**. I used these parameters to have an idea of the size of my sample:

|  |  |
| --- | --- |
| **Random Sampling** | |
| Population Size | 4.460.151 |
| Confidence Level | 99% |
| Margin of Error | 2% |
| Sample Size | 4.143 |

*Sample size calculated from raosoft.com/samplesize.html*

Even though the calculated sample size was around 4.200 observations, I decided to use exactly **5%** of the tripdata\_raw, that resulted in **223.008** observations for the dataset tripdata\_1.

*>glimpse(tripdata\_1)*

Rows: 223,008

Columns: 13

$ ride\_id <chr> "D409FEE4E8F2E637", "0C1DB7F657A54CA9", "82CE55232588113C", ~

$ rideable\_type <chr> "classic\_bike", "docked\_bike", "classic\_bike", "docked\_bike", ~

$ started\_at <chr> "2021-05-27 19:40:19", "2020-07-26 09:49:39", "2021-03-11 14:48:24", ~

$ ended\_at <chr> "2021-05-27 22:29:09", "2020-07-26 10:04:50", "2021-03-11 15:11:35", ~

$ start\_station\_name <chr> "Delano Ct & Roosevelt Rd", "Noble St & Milwaukee Ave", "Wabash Ave &~

$ start\_station\_id <chr> "KA1706005007", "29", "TA1307000117", "27", "13150", "180", "35", ~

$ end\_station\_name <chr> "Indiana Ave & Roosevelt Rd", "Clark St & Elm St", "State St & Van "~

$ end\_station\_id <chr> "SL-005", "176", "TA1305000035", "177", "13138", "156", "90", ~

$ start\_lat <dbl> 41.86749, 41.90068, 41.89147, 41.91021, 41.85781, 41.90687, 41.89228,~

$ start\_lng <dbl> -87.63219, -87.66260, -87.62676, -87.64350, -87.62455, -87.62622, ~

$ end\_lat <dbl> 41.86789, 41.90297, 41.87718, 41.92628, 41.87725, 41.93650, 41.88103 ~

$ end\_lng <dbl> -87.62304, -87.63128, -87.62784, -87.63083, -87.63937, -87.64754, ~

$ member\_casual <chr> "casual", "member", "casual", "casual", "member", "casual", “member”,~

# Process

After creating the sample dataset, I started the cleaning process.

To do so, I have taken the following steps:

* Convert the data types of the variable’s **member\_casual** and **rideable\_type** from *char* to *factor*
* Convert the data types of the variable’s **start\_at** and **ended\_at** from *char* to *date-time*