Google Data Analytics

Capstone Project Outline

June 13th, 2022

# Stages

1. **Ask**
   1. What are you exploring?
   2. What problem are you trying to resolve?
   3. What metrics will you use to achieve your objective?
   4. Who is your audience?
   5. How will this data help stakeholders make decisions?
2. **Prepare**
   1. Where is your data located?
   2. How is the data organized?
   3. Are there issues with bias or credibility in this data? Does the data ROCCC?
   4. How are you addressing licensing, privacy, security and accessibility?
   5. How did you verify the data’s integrity?
   6. How does it help you answer your question?
   7. Are there any problems with the data?
3. **Process**
   1. What tools are you choosing and why?
   2. Have you ensured your data’s integrity?
   3. What steps have you taken to ensure the data is clean?
   4. How can you verify that your data is clean and ready to analyze?
   5. Have you documented your cleaning process?
4. **Analyze**
   1. How should you organize your data to perform analysis on it?
   2. Has your data been properly formatted?
   3. What surprises did you discover in the data?
   4. What trends or relationships have you found in the data?
   5. How do these insights answer your question or solve the problem?
5. **Share**
   1. What story does your data tell?
   2. How do your findings relate to your original question?
   3. Who is your audience? What is the best way to communicate with them?
   4. Can data visualization help you share your findings?
   5. Is your presentation accessible to your audience?
6. **Act**
   1. What is your conclusion based on your analysis?
   2. How can you apply your insights?
   3. Are there any next steps you or your stakeholders can take based on your findings?
   4. Is there additional data you could use to expand on your findings?
   5. How can you feature your case study in your portfolio?

# Ask

## Who:

**Cyclistic:** A Bike Share program that features over 5800 bikes and 600 docking stations. Cyclistic sets itself apart by offering a wide range of different bikes, such as reclining, hand-trikes and cargo bikes, which makes Cyclistic more inclusive to the disabled or those who cannot ride a traditional two-wheeled bike. The majority of users opt for traditional bikes while about 8% use the assistive options.

**Lily Moreno:** Director of marketing and my manager. Responsible for the development and deployment of various marketing campaigns for all channels such as email, social media and more.

**Cyclistic marketing analysis team:** Responsible for the collection, organization, analysis and reporting of Cyclistic data, helping to guide the marketing strategies of the company. This is the team I joined roughly 6 months ago.

**Cyclistic Executive Team:** The final decision makers responsible for the final approval of new marketing strategies and programs. Notoriously detail oriented.

## What:

1. How do annual membership riders and casual riders use Cyclistic bikes differently?
2. Why would casual riders want to buy Cyclistic memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

**The question I have been tasked with answering:**

How do annual membership riders and casual riders use Cyclistic bikes differently?

## Why:

Cyclistic has determined that membership riders are more profitable than casual riders. Moreno believes that there exists a very good chance to convert casual riders into annual members, noting that casual riders are already aware of the Cyclistic program and have already chosen Cyclistic for their mobility needs. Moreno’s current goal is to develop new marketing materials and strategies targeted at these casual riders to convert them to annual membership riders.

# Prepare

## What Data?

Cyclistic historical rides over the last 12 months, not including any personally identifiable information.

## How is the data organized?

The current state of the data included a ride ID, start and end datetimes, start and end station names and station ids, the type of bike used, whether the rider was a member or casual rider, and longitude and latitude coordinates of the start and end locations.

## How will the data be stored?

The data is currently stored in CSV format. I plan to load this data into a PostgreSQL database for ease of access. The original data, and any formatted, cleaned, or processed CSVs will be stored in the GitHub repository along with all other case materials and code.

# Process

## Tools Used

Because the data is stored in multiple CSV files, I will be using Python notebooks along with Pandas and NumPy for the cleaning and processing tasks. Once complete, the resulting dataset will be imported into a new PostgreSQL database for further analysis via SQL queries and BI reports in Tableau\*. I will also use Python Notebooks to create a more technical version of the BI reports in PDF format, using Matplotlib and Seaborn to generate data visualizations.

*\* Tableau Public does not allow the connection of a database, therefore the data will be loaded into Tableau using the CSV files, however I will still be importing the data to PostgreSQL for the purpose of creating SQL queries as if I were able to work directly with a database in Tableau or other platforms.*

## Integrity Check

Upon initial visual inspection, the dataset appears to be complete, with minimal to no errors. The start and end station names and IDs are missing from many rows.

## Cleaning and Processing Documentation

1. Load all csv files into a single dataframe (to be later exported as a single csv instead of many csv’s)
2. Convert all datetime columns into Python datetime objects
3. Create a new timedelta column describing the number of days, hours, minutes and seconds a ride lasted. Column name: ride\_duration
4. Extract the day of the week each ride took place from the start datetime. These will be labeled as integers. 0 = Monday, 6 = Sunday. Column name: day\_of\_week