Cyclistic Bike-Sharing Data Analysis

Thanakorn Thanakraikiti

June 18, 2023

Contents

1 Intoduction

This section provides an overview of the project, including the purpose, objective, and stakeholders. It is divided into three main parts: (1) the business problem, (2) the business task, and (3) the stakeholders.

1.1 Business Problem

Cyclistic is a bike-share company in Chicago that launched a successful bike-share offering in 2016. Its marketing strategy primarily focused on building general awareness and appealing to broad consumer segments. Cyclistic offers a variety of pricing plans, including single-ride passes, full-day passes, and annual memberships. Cyclistic classifies its riders into two groups based on these plans: casual riders are those who purchase single-ride passes or full-day passes and annual members are those who purchase an annual membership.

Cyclistic's flexible pricing plans attract more customers, but financial analysts have found that annual members are more profitable. This suggests that a marketing campaign that targets existing customers is likely more effective at expanding revenue than a campaign that targets new customers.

Therefore, Cyclistic's marketing team is interested in understanding how casual riders and annual members use Cyclistic bikes differently. By understanding these differences, the marketing team can develop more targeted marketing strategies to convert casual riders into annual members.

1.2 Business Task

The marketing team will analyze Cyclistic's historical bike trip data from 2022 in order to identify how annual members and casual riders use Cyclistic bikes differently. The insights will be used to develop marketing strategies that appeal to casual riders and encourage them to subscribe to annual memberships.

1.3 Stakeholders

Key stakeholders of this project include:

- Marketing analytics team who collect, analyze, and report data to help guide Cyclistic's marketing strategy
- Marketing manager who supports the marketing campaign to promote Cyclistic's bike-share program
- Executive team who approve the recommended marketing program

2 Data Sources

This section provides a brief overview data sources for the project. It also assesses the credibility of data source and documents the process of data cleaning and data manipulation. It is divided into three main parts: (1) data description, (2) data credibility, and (3) data cleaning and manipulation.

2.1 Data Description

The marketing team uses Cyclistic's historical trip data from January 2022 to December 2022 as its primary source. The data can be downloaded from the company's website via this link. It was collected by the City of Chicago (a government agency) and published by Cyclistic. The data contains historical trip data for all of the company's customer (rider) data since 2013, including the following information: ride history, the location of stations for each ride, and other relevant information

The data is stored in twelve CSV files, one for each month of 2022 (`202201-divvy-tripdata.csv` through `202212-divvy-tripdata.csv`). Each file contains a set of fields (columns) and records (rows). Each record represents a single bike trip, which is uniquely identified by an identifier for the ride.

Every ride contains the following 13 fields:

- `ride_id` : A unique identifier for the ride
- `rideable_type` : The type of bike for the ride (classic bike, electric bike or docked bike)
- `started_at` : The date and time at which the ride started
- `ended_at` : The date and time at which the ride ended
- `start_station_id` : The ID of the station where the ride started
- `start_station_name` : The name of the station where the ride started
- `end_station_id` : The ID of the station where the ride ended
- `end_station_name` : The name of the station where the ride ended
- `start_lat` : The latitude of the station where the ride started
- `start_lng` : The longitude of the station where the ride started
- `end_lat` : The latitude of the station where the ride ended
- `end_lng` : The longitude of the station where the ride ended
- `member_casual` : The type of user who took the ride (member or casual)

2.2 Data Credibility

The data was evaluated for bias and credibility using the ROCCC method:

- **Reliable**: The data is considered reliable because it is open data provided by Cyclistic and the City of Chicago. It is a comprehensive and accurate dataset of Cyclistic's historical bike trips. The data includes information on the start and end location of each trip, the time of day the trip was taken, the type of user (casual or annual member), and the length of the trip. There are some missing values in some fields, but these can be used to explore how different customer types use Cyclistic bikes.
- **Original**: The data is considered to be original, based on a primary source of information from the company.
- **Comprehensive**: It is important to note that Cyclistic is subject to data privacy restrictions that prohibit the use of riders' personally identifiable information (PII). This means that the company cannot connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes, as well as identify the rider's gender and age. Despite these restrictions, the data is complete and contains all the information needed to understand how members and casual riders use Cyclistic bikes. This includes information such as the number of rides taken, the length of each ride, the start and end stations of each ride, the time of day each ride was taken.
- **Current**: The data is current and updated monthly. The selected datasets are the most recent data available.
- **Cited**: The data is properly cited as it contains the dataset name, author, and published year.

Overall, the Cyclistic bike data is considered a credible source for analysis based on the Reliability, Originality, Comprehensiveness, Currency, and Citation (ROCCC) method. The data can be used to explore how different customer types use Cyclistic bikes. For example, the data can be used to determine which customer segments are most likely to use Cyclistic bikes for commuting, recreation, or transportation. The data can also be used to identify trends in bike usage, such as the time of day when bikes are most likely to be used or the types of bikes that are most popular.

While there are some restrictions on rider's personal information, there is still a great deal of information that can be learned from it. However, the data needs to be processed before analysis can be performed. This includes cleaning the data to remove errors and inconsistencies, and transforming the data into a format that is suitable for analysis. The data processing will be performed in the next section.

2.3 Data Cleaning and Manipulation

This part provides the document of data cleaning and manipulation. It includes the following tasks:

1. Removing duplicate records
2. Fixing station name
3. Identifying and correcting missing values
4. Adding new calculated columns
5. Removing outlier values

The R programming language will be used in conjunction with the RStudio integrated development environment (IDE) to complete these tasks. Packages are also required to load to R environment for the project, including `tidyverse` for importing and wrangling the data, `janitor` for cleaning data, `ggplot2` for visualizing the data, `ggmap` for visualizing the data in a map.

Prior to cleaning the data, it is necessary to set up the data in a structured manner. Each CSV file was imported into a separate data frame, and then these data frames were combined into a single data frame called *trip_data_v2*. The table below shows the preview of the *trip_data_v2* data frame that was created.

ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual
98D355D9A9852BE9	classic_bike	2022-01-01 00:00:05	2022-01-01 00:01:48	Michigan Ave & 8th St	623	Michigan Ave & 8th St	623	41.87277	-87.62398	41.87277	-87.62398	casual
04706CA7F5BD25EE	electric_bike	2022-01-01 00:01:00	2022-01-01 00:04:39	Broadway & Waveland Ave	13325	Broadway & Barry Ave	13137	41.94907	-87.64863	41.93758	-87.64410	casual
42178E850B92597A	electric_bike	2022-01-01 00:01:16	2022-01-01 00:32:14	Clark St & Ida B Wells Dr	TA1305000009	Clark St & Ida B Wells Dr	TA1305000009	41.87592	-87.63119	41.87593	-87.63058	casual
6B93C46E8F5B114C	classic_bike	2022-01-01 00:02:14	2022-01-01 00:31:07	Michigan Ave & 8th St	623	Michigan Ave & 8th St	623	41.87277	-87.62398	41.87277	-87.62398	casual
466943353EAC8022	classic_bike	2022-01-01 00:02:35	2022-01-01 00:31:04	Michigan Ave & 8th St	623	Michigan Ave & 8th St	623	41.87277	-87.62398	41.87277	-87.62398	casual
AC1F67BDCDDD5988	electric_bike	2022-01-01 00:03:36	2022-01-01 00:04:02	Adler Planetarium	13431	Adler Planetarium	13431	41.86616	-87.60729	41.86610	-87.60727	member

Upon initial inspection, the data frame contains 5,667,717 rows and 13 columns. It is ready to use to perform the data cleaning tasks.

2.3.1 Removing Duplicate Records

The first step in data cleaning and manipulation is to detect and eliminate duplicate data. Duplicate data refers to rows or cells that contain the same data. It can occur when data is entered incorrectly or when data is merged from multiple sources. Duplicate data can skew the results of analysis, so it is important to eliminate it before proceeding.

To check for duplicate records in a data set, the following command can be used to count the number of rows and unique rides in the data set:

```
trip_data_v2 %>%
  summarize(total_ride = n(),
            unique_ride = length(unique(trip_data_v2$ride_id)))
```

total_ride	unique_ride
5667717	5667717

The result shows that there are 5,667,717 rides in total and 5,667,717 unique rides in the data set. This indicates that there are no duplicate values in the data set.

2.3.2 Fixing Station Names

The second step in data cleaning and manipulation is to standardize the data. This involves ensuring that all of the data is in a consistent format, with the same data types and values. This can be a time-consuming process, but it is essential to ensure the accuracy and reliability of the data.

Initially, irrelevant columns were removed as they were not required for this analysis. These columns included `start_station_id` and `end_station_id`.

```
trip_data_v2 <- trip_data_v2 %>%
  select(-c(start_station_id, end_station_id))
```

```
# Examine the starting and ending station names
start_station <- count(trip_data_v2, start_station_name, name = "total_station")
end_station <- count(trip_data_v2, end_station_name, name = "total_station")
```

An inspection of the start and end station columns were examined and revealed some inconsistent formattings. The following test stations were found:

- `Pawel Bialowas - Test- PBSC charging station`
- `Hastings WH 2`
- `DIVVY CASSETTE REPAIR MOBILE STATION`
- `Base - 2132 W Hubbard Warehouse`
- `Base - 2132 W Hubbard`
- `NewHastings`
- `WestChi`
- `WEST CHI-WATSON`.

This can be done by executing the command below:

```r
# Remove all rows that have eight test stations in both the starting and ending locations
trip_data_v2 <- trip_data_v2[
  !(trip_data_v2$start_station_name %in% c("Pawel Bialowas - Test- PBSC charging station",
                                            "Hastings WH 2",
                                            "DIVVY CASSETTE REPAIR MOBILE STATION",
                                            "Base - 2132 W Hubbard Warehouse",
                                            "Base - 2132 W Hubbard",
                                            "NewHastings",
                                            "WestChi",
                                            "WEST CHI-WATSON") |
   trip_data_v2$end_station_name %in% c("Pawel Bialowas - Test- PBSC charging station",
                                         "Hastings WH 2",
                                         "DIVVY CASSETTE REPAIR MOBILE STATION",
                                         "Base - 2132 W Hubbard Warehouse",
                                         "Base - 2132 W Hubbard",
                                         "NewHastings",
                                         "WestChi",
                                        "WEST CHI-WATSON")), ]
```

The following extra texts were found in station names: `*`, `- Charging`, `(Temp)`, and `amp;`.

```
# Remove all unwanted characters from the starting and ending station names
trip_data_v2 <- trip_data_v2 %>%
  mutate(start_station_name = str_replace_all(start_station_name, fixed("*"),           "")) %>%
  mutate(start_station_name = str_replace_all(start_station_name, fixed(" - Charging"), "")) %>%
  mutate(start_station_name = str_replace_all(start_station_name, fixed(" (Temp)"),     "")) %>%
  mutate(start_station_name = str_replace_all(start_station_name, fixed("amp;"),        "")) %>%
  mutate(end_station_name   = str_replace_all(end_station_name,   fixed("*"),           "")) %>%
  mutate(end_station_name   = str_replace_all(end_station_name,   fixed(" - Charging"), "")) %>%
  mutate(end_station_name   = str_replace_all(end_station_name,   fixed(" (Temp)"),     "")) %>%
  mutate(end_station_name   = str_replace_all(end_station_name,   fixed("amp;"),        ""))
```

2.3.3 Identifying and Correcting Missing Values

2.3.4 Adding new calculated columns

2.3.5 Removing outlier values

3 Analysis

3.1 Summary

3.2 Supporting Visualizations

3.3 Key Findings

-
-
-
-
-

4 Conclusion

5 Recommendations

-
-
-