Case Study – Cyclistic

Objective

This case study consists in analyzing Cyclistic historical data, to find out patterns and insights that can help improve the annual membership numbers. The final report will be used by the director of marketing (Lily Moreno) and the Cyclist executive team to drive business decisions.

Data sources

The dataset that will be used includes information about bicycle trips from the first and last quarters of 2019 (Q2 and Q3 could not be used due to size restrictions). They were imported into a table in a database (Postgres), containing the following information (inside parenthesis are the column names):

- ID of the trip, a unique identifier (trip_id)
- Start and end time of the trip (MM/DD/YYYY HH:MM format) (start_time, end_time)
- ID of the bicycle used (bike_id)
- Trip duration (in seconds) (trip_duration)
- IDs and names of the start and end stations (from_station_id, from_station_name, to_station_id and to_station_name)
- User type ("Subscriber", for annual users, and "Customer", for casual) (user_type)
- Gender and birth (gender and birth_year)
  Two more columns were added after the data cleaning process (ride_length and day_of_week) and are described further in this document.

Data integrity

The following steps were executed to ensure data integrity:

- Checking of end time x start time: 13 records had an end time that were less than start
  time. In all of them, the trip duration remains consistent if we add an extra hour to the end time. For example, one record has a start time of 1:43, end time of 1:09 and duration of 1594 seconds (or 26.5 minutes). If we add one hour to the end time (2:09), the difference between 2:09 and 1:43 is 26 minutes, which is consistent with the trip duration. This was fixed using SQL queries.
- Null values: the columns gender and birth_year were found to contain null values, occurring in about 8% of records. Since this information does not affect trip duration or station destination, the records were kept, but this might affect the precision of a demographic analysis.
- Duplicate records: since trip_id was defined as primary key and no import errors happened, we can be sure that no duplicate transactions exist in the database.

Data cleaning and manipulation

There were errors when trying to import the first CSV file into Postgres. After some investigation, it was concluded that the trip duration information was causing this error, because of the presence of a comma as a thousand separator. The formatting of the column in the CSV file was changed in Excel to hide it without any data loss (1,800.00 became 1800, for example). After that, importing proceed without any issues.

Two columns were added to the table to facilitate future analysis:

- ride_length: the trip duration in the HH:MM:SS format
- day_of_week: the day of the week that the ride took place, in written form (Sunday, Monday, etc.)

Data analysis

By using SQL queries, we find out that:

- Casual members use bikes, on average, for longer periods of time (37 minutes) and most commonly on Sundays. Annual members use bikes for shorter runs (11.5 minutes) and most commonly on Tuesdays.
- Performing more queries, we find out that the number of casual members on weekends is almost twice more than on weekdays. However, the average ride length stays almost constant - weekend rides are less than 10% longer.
- The trend is reversed with annual members: there are almost twice as more users on weekdays than on weekends. But the ride length also stays constant, only about one minute longer on weekends

These data relationships are further explored in an Excel spreadsheet, that contains tables and data visualizations.

Recommendations

- Creation of plans that benefit weekend users and users that ride for longer. This covers most casual users.
- Marketing campaign to attract more female users. Both casual and subscribers are composed majority by men; there is a huge market opportunity here and can greatly expand Cyclistic userbase
- Customer retention. Although attracting casual members is an important part of business, annual members still are the majority of userbase. Keeping these customers satisfied is key for a successful business.