README

An Analysis of the Bay Wheels Bike Share System Data

Author: Jiayi Zhang

Scope: Data wrangling and visual analysis was performed on a dataset of over 5 million, containing various data collected between 2017 and 2020 by the Bay Wheel bike share platform. The four steps of this wrangling effort will be briefly summarized below.

Step 1: Data Gathering

A total of 28 zip files containing datasets spanning across different months and years were programmatically downloaded, unzipped, opened and concatenated to form the exploratory dataset.

Step 2: Data Analysis

The focus of this section was to analyze our combined dataset for any potential tidiness and quality issues that should be addressed prior to any analysis is completed. The major points of concern are summarized as follows:

Tidiness Issues:

None

Quality Issues:

1. *bike_share_for_all_trip* should be *True* or *False*
2. quite a few columns contain missing values
3. drop duplicate rows
4. dropping unnecessary columns
5. *start_time* and *end_time* should be *datetime* format
6. *start_station_id*, *end_station_id* and *bike_id* should be str

Step 3: Data Cleaning

Most issues from step 2 were addressed with the exception of filling in the missing data values for *start_station_name* and *end_station_name*, approximately 12% of this data was missing. My attempt at finding these missing values was by creating a dictionary of lat/lon coordinates (which contained no missing values) and its unique station name, then using this dictionary to map the missing ids. This however did not work for all values as intended.

My suspicions as to why the method did not work for all values was due to the fact that some of the coordinates were just not mapped to any stations to begin with, which would of course not be part of the dataset. One of the biggest challenges was also picking an appropriate rounding point

for each lat/lon value, as some locations were identical in name, but had differing lat/long values up to the 3rd or 4th decimal points.
One other strategy that would potentially work is to perform reverse geocoding on each coordinate to get its location using the *geopy* library, however, with over 5million+ data points, the standard built in API would not be feasible.

Therefore, although unfortunate, I think maybe the best solution would be to drop all of the missing values with reference to *start_station_id* (this column contains the most missing values), as we will be using the station names as part of our analysis.

Step 4: Data Visualization

This section was intended for exploratory and explanatory data visualization and analysis of our cleaned dataset. Some of the worth mentioning findings from a series of univariate, bivariate, and univariate analysis are summarized below:

1. Over 6000 out of 16000 sharebikes have only around 100 trips, while the average number of trips per bike is 434. This tells us that there are a lot of new bikes in the rotation.
2. The most popular (frequent) starting location for a trip is Market St. at 10th St., while the most popular end location is the San Francisco Caltrain (Townsend and 4th.)
3. The trip between the San Francisco Ferry Building and the Embarcadero at Sansome St. is by far the most popular trip in terms of frequency
4. 8 am and 5 pm are the busiest times, which represents morning and afternoon rush hours. For day of week, there are a lot more users on weekdays than weekends, which may suggest that people use it mostly for work related commutes
5. There are 3.69 times more subscribers than there are regular customers, however, when it comes to bikeshare usage, it seems that the casual customers use the bikes for longer time and further trips than subscribers. This may mean that a large portion of subscribers aren't taking advantage of their subscriptions.
6. There is a positive relationship between the distance and duration of a trip.
7. Out of the 10 most popular trips, the average maximum commute length is just over 2000 meters
8. Users on average spend more time on sharebikes during weekends than weekdays, even though there is more usage during weekdays (see point 4). Tuesdays have the highest distanced travelled after Sunday, which seems kind of unintuitive.

Useful resources:

- Programmatically downloading zip files (https://stackoverflow.com/a/14260592)
- Programmatically concatenating multiple files (https://stackoverflow.com/a/21232849)
- Distance calculation based on coordinate points (https://stackoverflow.com/a/44446971)