Sam Gutentag

Bay Area Bike Share Data Wrangling

The data provided by the Bay Area Bike Share Program [1] is very well organized and concise. Broken into six month chunks, the data covers August 29th, 2013 through August 31st, 2016 and is split into four files for each chunk. The files are 'status_data', 'station_data', 'trip_data', and 'weather_data'. The data in each file is barely clean to start with if a bit verbose. to clean each file I will use a few different techniques depending on what portion of the proposal I am attempting to complete.

The 'status_data' files are the largest by a large margin, each file consisting of roughly 17 millions records of the status of each stations' dock availability. This data is a time series data set of each station reporting the number of bikes that are currently parked and how many open docks are available for bikes to be parked there on one minute intervals. For my uses, I will look to aggregate this data down to half hour or possibly one hour records of the min, max, and average bikes available and docks available at each station. Once heavy usage times, such as major commute hours, are identified, I will trim the data set to only include those smaller time frames on a minute by minute basis for a more granular view of the data.

The 'station_data' files are records of the individual stations. Each record includes the station id, station name (typically the cross street or closes address), the latitude and longitude coordinates, a dock count, the nearest landmark (often the nearest city name), and the installation date. This data is quite clean to start, the only notable key detail is that stations may appear multiple times in the data, as a few stations have been relocated, warranting a new record with an updated installation date, and a few stations have been expanded by the addition of more docks, also indicated by a record with a later installation date.

The 'trip_data' files are records of individual trips taken as part of the program. Each record contains a trip id, the duration of the trip in seconds, the start date and time in PST, start station, start terminal, end station, end terminal, the bike number, subscription type, and zip code. In earlier dates, the subscription type was not recorded. for most of my analysis these null values are simply ignored, but when I start to do analysis by subscription type, I will drop the null values. Zip codes are only available for subscription plan members, so again suffers from the same issue that some records do not specify. in later dates, the program expanded to allow users to manually enter their home zip codes for any user type. however it is specifically noted that this data may not be reliable. For this data I will look to drop null values, and drop zip codes that are not known US zip codes to clean out errors.

The 'weather_data' files are also quite clean to begin with, but they are very verbose, so my initial cleaning will be to pare down the data to just include date, average temperatures, humidity levels, wind speed, events, cloud cover, and zip codes. This seems like a lot, but from the initial list of 24 different columns, this is a trim. The event records are full of null values, so these will only be used as an auxiliary reference point if noticeable changes in ridership are spotted.

The most notable outliers are found in the trip_data, specifically in the ride duration. There is a single trip that lasted 287840 minutes, which is just under 200 days! The data set of just over 980,000 trips consists of just 4 trips that are over 10 days long, so these are being removed from the data set. The most extreme outlier is a single trip that is just over 199 day long! and is most likely a reporting error, or else a very sad, over billed rider.

For the DarkSky dataset, there was a bit more cleaning to be done. The web API allows for all kinds of calls to historical weather data but it is returned as a large JSON payload that I first convert to a dictionary and then work through the dictionary to flatten it into a data frame of hourly weather data. It also includes a large number of entries for each record so I can first pare down to just a few select columns.

The data which will be used in this project is publicly available through The Bay Area Bike Share Open Data site [1], supplemental weather data may be gathered from DarkSky [2].

References

[1] http://www.bayareabikeshare.com/open-data
[2] https://darksky.net/