Bike\_Functions.py

Purpose: Stores functions – mostly for writing to text files in different directories.

Data\_Load.py

Purpose: Initial download of data files from Capital Bikeshare’s website, which I downloaded directly to my computer and created a directory that could be walked through easily.

Inputs: Dataframes in the form of CSV files downloaded directly from Capital Bikeshare’s site (<https://s3.amazonaws.com/capitalbikeshare-data/index.html>)

Outputs:

* full\_data\_dict info - ORIGINAL.txt – contains information regarding all the raw, unwrangled data
* full\_data\_dict info - EDIT.txt – contains information regarding all the data, but now with changed feature names based on inconsistencies found in full\_data\_dict info - ORIGINAL.txt
* Create Zipcodes.txt – contains information regarding the addition of two new features (starting and ending zipcodes)

Select\_Subset.py

Purpose: Select only the subset of raw data I found to be useful and needed: 2015 Q3 – 2016 Q3.

Inputs: Dataframes in the form of CSV files downloaded directly from Capital Bikeshare’s site (<https://s3.amazonaws.com/capitalbikeshare-data/index.html>)

Outputs:

* subset\_data\_dict info.txt - contains information regarding all the relevant data
* full\_subset\_data.csv – new CSV file with all the subset data in one dataframe

Station\_Data.py

Purpose: Bring in code found at <https://kwkelly.com/blog/analyzing-capital-bikeshare-data-with-python-and-pandas/> and use it to map station numbers to corresponding latitude/longitude field.

Inputs: full\_subset\_data.csv

Outputs: full\_subset\_data - V2.csv

Distance\_Feature.py (Beware: This script can take 15-30 minutes to run)

Purpose: Create a new feature based on distance between each starting and ending point latitude/longitude.

Inputs: full\_subset\_data - V2.csv

Outputs: full\_subset\_data - V3.csv

Clean\_Stations\_Numbers.py

Purpose: Explore whether station numbers and station names always match, and, if not, show stats related to the stations that do not match. Output a more clean data file.

Inputs: full\_subset\_data - V3.csv

Outputs: full\_subset\_data - V4.csv

**THIS IS WHERE I SEE THE ISSUE**

Full\_Empty\_2.py

Purpose: Up until this point, I only have network type data in the form of Trip 1 took this person from Point A to point B and these were the locations and so on. My project proposal, however, is to be able to predict the number of bikes/open docks at any station. Since I do not have any starting numbers, I need to create them in a new target feature. The data I have so far (full\_subset\_data - V4.csv) shows all comings and goings between stations, and I am pairing it with another dataset from CaBi Tracker (<http://cabitracker.com/station_outage.php?id=149>) which, over the same time period, shows all times when a station is full and when that stops and ends, and when a station is empty and when that stops and ends. Putting these two data sets together, I should be able to back out station capacity, number of bikes at each station, and number of empty spaces at each station at each point in the original data set. To do this, I split up the two data sources into 6 different data frames to effectively discretize time from a network (Point A at Time X to Point B at Time Y in one row) to standalone, sortable points in time (Point A at Time X in one row, and Point B at Time Y in another row). Doing this, however, has shown me that the original data set has been scrubbed too much by the original company, Capital Bikeshare, and so, even combining these two datasets, I am not able to back out station information accurately at all. To test this theory, I started with station 31116, which you will find is the focus of this script. This station has a capacity of 19, but my resulting dataframe shows me much different information. **QUESTION: Is the data as unusable as I think it is, or is my code doing something I did not want it to do?**

Inputs: full\_subset\_data - V4.csv, full empty\_31116\_new.csv (CaBi Tracker Data)

Outputs: Still in progress, but this would be a new data file with my target features.