**Springboard—Intermediate Data Science with Python**

**Proposal for Capstone Project – Bike Share Advertising in New York City**

**Andrew Harris**

**July 2019**

**Statement of Problem:**

Citi Bike is a public bicycle sharing system operating in New York City. Citi Bike have over 800 bicycle sharing stations located throughout the city which have space for small billboard advertising. Citi Bike wishes to profile their users so that they can determine the value of their advertising space and identify the products and services which should be marketed at a given station.

Citi Bike has collected data for every trip taken from July 2013 onward. The data contains information such as the start time, end time, start station and end station. The data also contains rider specific information such as the gender, birth year and whether or the not the rider is a subscriber (Annual Member) or customer (24-hour pass or 3-day pass user).

This trip data can be segmented based on rider type (customer vs. subscriber), gender, age, start station and end station so that the station user demographics can be determined for individual stations and used to support marketing activities.

**Description of Client:**

The client for this problem is Citi Bike itself. Citi Bike have over 800 active bicycle sharing stations located throughout New York City. These stations currently have payment machines which have small billboards on the sides perpendicular to the payment machine user interface. One of the two billboards is typically a map of the surrounding area while the other billboard is an advertisement for Citi Bike itself or for services offered by Citi Bank.

The segmentation described in the “Statement of Problem” section will allow the client to determine the user demographic for each station. Citi Bike could use this information in two ways:

1. Banking services offered by Citi Bank which are geared toward specific demographics can be advertised at stations which are most frequently used by that demographic.
2. If Citi Bike decided to lease the advertising space to a third-party organization, they would have data supporting the user demographic of each station. User demographic data would allow Citi Bike to determine the value of the advertising space and identify the organizations to which they should market the advertising space.

**Data Set:**

The source of data for these models is the Citi Bike website at <https://www.citibikenyc.com/system-data>. The data set itself is called “tripdata” and is reported on a monthly basis with a .csv file for each month. Each line in the data is an individual trip with 15 attributes. The 15 attributes are as follows:

* start time, start station id, start station name, start station latitude and start station longitude.
* End time, end station id, end station name, end station latitude and end station longitude.
* Trip duration, bike id, user type, birth year and gender.

Citi Bike has processed the data to remove any trips that were taken by staff as part of service and inspection activities or were between “test” stations. Citi Bike has also processed the data to remove any trips below 60 seconds in length as such trips are potentially false starts or users attempting to re-dock a bike to verify that it is secure. Thus, the trips present in the data set are actual paid trips taken by users.

The past six months of data (January 2019 to June 2019) contains over nine million trips taken between over 800 bicycle stations.

**Approach:**

The approach used to execute this capstone project will be as follows:

1. Preparation of Project Data Sets: this work will involve cleaning and manipulating the data to develop a working data set to be used for Exploratory Data Analysis and Inferential Statistics. From the working data set, a data set that has been formatted specifically for use with K-Means clustering will be developed.
2. Perform Exploratory Data Analysis: this work will involve preparing summary statistics and visualizations using the working data set in order to understand the data and look for patterns.
3. Perform Inferential Statistics: this work will involve applying inferential statistics to further explore that data and uncover relationships between variables.
4. Prepare a Baseline Clustering Model: this work will involve using the K-Means clustering algorithm with an appropriately defined value of K and distance function to perform clustering on the data set. The results of the clustering will be analyzed in the context of the business problem.
5. Perform Additional Cluster Modeling: this work will involve implementing other clustering algorithms such a t-SNE, adding or engineering features, or applying different distance functions to see if better clustering can be achieved beyond the baseline model. The results of the additional clustering will be analyzed in the context of the business problem.

**Deliverables:**

The deliverables for this capstone project are as follows:

1. Capstone Project Milestone Report – This report will be prepared and submitted after the Baseline Clustering Model has been prepared under Step 4 of the Approach Section.
2. Capstone Project Final Report – This report will be submitted after the Additional Cluster Modeling has been prepared under Step 5 of the Approach Section. This report will be submitted as a draft in order to obtain mentor feedback to incorporate into the final submission.
3. A presentation slide deck documenting the Capstone Project.
4. All Python code used for data cleaning and manipulating, for exploratory data analysis, for statistical inference and for machine learning model development and evaluation.