**Introduction/Business Problem**

The goal of this project is to provide meaningful information to someone who wants to plan a bike ride using the existing Citi Bike network in New York City. It can be someone who already lives there and either needs to go on an occasional bike trip or even wants to plan regular trips. But the target audience can also include somebody who is moving in and wants to rely on the bike share program, and tourists who wants to visit the city in an active way. In any case, information regarding the most common trips, their durations and the point of interests at the destination are useful for planning.

The assumption in this work is that a start station where to pick up the bicycle is chosen, and the user is looking for information regarding the final destinations. Specifically, the final goal is to highlight the type of venues and point of interest located near the end station of the trip. To do this, different steps are taken. First, an existing bike station is selected as starting point, and its data in terms of past trips to other existing stations are cleaned and analyzed. Then, the Foursquare API explore feature is exploited to find the most popular venues around the destination stations. Also, clustering techniques are used to find patterns, and from the clustering results and the location data, conclusion are drawn.

**Data**

The first part of this work is based on the monthly data relative to a selected bike station of the existing Citi Bike network. A list of the existing station in JSON format is available:

<https://feeds.citibikenyc.com/stations/stations.json>

This JSON file is used to visualize the stations maps and find the information relative to chosen station, such as ID and geographical coordinates.

The monthly data can be extracted from the spreadsheets available on this website:

<https://www.citibikenyc.com/system-data>

This same website has an interactive map with real time data, which gives an idea of the location and status of bike stations and docks. One spreadsheet of a certain month and year is chosen. It contains information about the trips to and from all stations, such as start time, duration, start and end station coordinates and so on. Not all the data in this spreadsheet are used. For example, the information relative to the end stations, such as name and coordinates, are kept for grouping and visualization purposes. They are also used as inputs to Foursquare. Other information such as trip duration and type of customer are kept for a possible analysis of any correlations.

The second part of this projects relies on the Foursquare API to find the venues around selected destinations. The obtained data are used in a clustering algorithm to highlights the common characteristics of the final destinations.

**Methodology**

The first step consists of importing the necessary libraries, reading the JSON file of the bike stations and visualizing them with Folium. Next the CSV spreadsheet related to the trips data of the chosen month is imported into a Pandas data frame and a start station is also chosen. The name of the chosen station is used to select only the rows relative to it and deal with a smaller data frame. The data are cleaned and analyzed. For example, any rows with empty values are deleted, after checking that their number is small. The categorical values indicating the type of customer are converted using the “get\_dummies” method. Columns such gender are dropped, as well as the trip’s end time, since the start time and duration already contain the information. The trip duration is expressed in minutes instead of seconds.

The clean data frame is grouped by end station features, i.e. name and geographical coordinates. These stations are also visualized on a map through Folium.

The Foursquare API is then used to explore the areas around the end stations and find the most popular point of interest around them. Only the most common stations are considered, i.e. those stations that are the end points of at least 5 trips in the month. In general, all the stations reached in the month can be considered.

KMeans is used to find patterns in the trips’ most common end stations.

**Observations**

For the chosen month and starting point, there’s no significant correlation between destination, duration, type of customer and birth year. The birth year of the customer has a low standard deviation. The trip duration has many outliers. In general, a different choice of start station and monthly data can lead to different results and show correlations.

According to KMeans results, most of the destinations are grouped in one cluster, which is characterized by having restaurants, coffee shops and bars. Some stations are quite close to each other, so probably the same or similar venues are found by Foursquare, which might explain why they’re grouped together. Some are instead further. In general, this shows that in various locations and final destinations it’s common to find food venues. The other clusters contain one or two stations. The two stations in one cluster have a park has the most common venue. The other clusters with only one station each are characterized by a variety of point of interest. One of them has multiple markets.

A couple of tests have been run by using a slightly different number of clusters and/or a different number of destinations. The results are similar, i.e. most of the arrival stations are grouped in one cluster, with the other clusters having one or two stations at the most.

**Conclusions**

This project shows that interpretation of past bike trips data and of geographical information can be used in bike trips planning. The process can be generalized and repeated with different starting points and using data from different months. In general, the results will be different than the ones reported here, but the process is the same. Once a starting point is chosen, the data analyses and the algorithm uses can give an idea of the most popular places that surround the other stations of the bike share city network.