**CAPSTONE PROJECT #1**

****

|  |  |
| --- | --- |
| **Introduction** | **2** |
| **Data** | **2** |
| **Exploratory Data Analysis** | **3** |
| **Machine Learning** | **3 & 4 & 5** |
| **Conclusion** | **5** |

**Introduction:**

Starting in 2013, Citibike has been operational in New York City. The program currently covers the boroughs of Manhattan, Queens, and Brooklyn. It has also extended across the water to Jersey City.

In 2018 there were approximately 17million rides taken in total.

Given the large amount of available data from the rides, unexpected patterns and trends can be potentially found. The goal of this project is to use machine learning methods to explore the data and to see if there are any unique findings that can help improve the bike sharing program’s operation.

**Data:**

The rider data comes from Citibikenyc.com. Data for the entirety of 2018 was downloaded but due to computing limitations, a single month of data has been used. March data set (201803-citibike-tripdata.csv) is used as it is a largest full month of data that can be computed in a reasonable amount of time. As such, an assumption is made in that the observations we make in this exercise is consistent with the other 11 months.

Jersey City’s data came in a separate file which has been concatenated into the NYC file. March 2018 data is used for JC as well.

The files contain trip duration, start time, stop time, start station, end station, geographic coordinates, usertype, birth year, and gender.

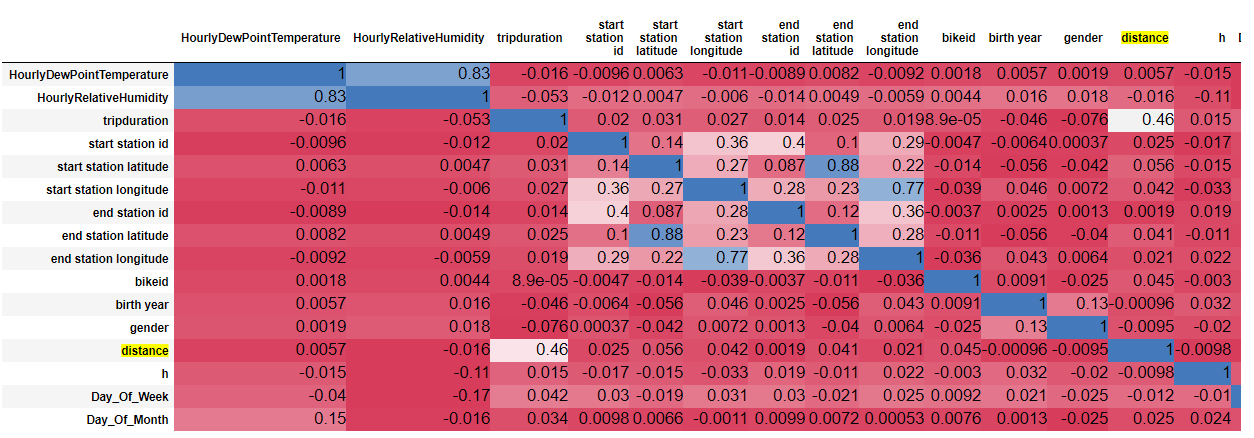
Weather data has also been concatenated into the file. Data on temperature and humidity are acquired from Weather.gov based on the Central Park Station.

Data Clean Up**:**

* Time data has minutes and seconds removed, leaving only hour and day of month data.
* The weather data is merged based on this ‘cleaned’ date column.
* NYC and Jersey City data sets are concatenated vertically. They contain the same type of ride data.
* Time of day and Weekday vs Weekend metrics are created based on the cleaned date column.
* Distance is calculated via haversine based on kilometer using the starting and ending station coordinates. The limitation here is that ‘city distance’ isn’t exactly the same as a direct line from point A to point B. However, ‘city distance’ needs paid service from Google so we are using haversine as an approximation.
* Blank weather data are filled using bfill as temperature shouldn’t change too much from one hour to another.
* Checked bike data for NaNs in the distance data but none was found.
* Outliers – There are a few trips which are extremely long. Any trips longer than 50,000 seconds (~13.9hours) are filtered out.

**Exploratory Data Analysis:**

A correlation matrix is ran on the entire data set. Results comes out to expectation with nothing particular unique that standing out. Duration & Distance, Start Station Location & End Station Location, Humidity & Temperature pairings have strong correlations.



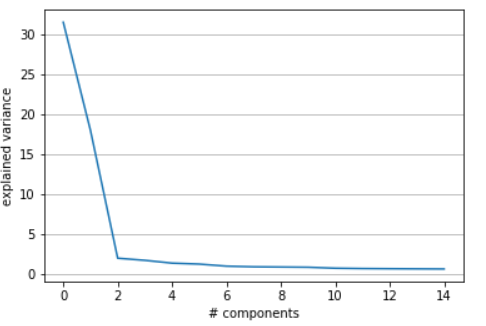
**Machine Learning**

As the data available are not labeled, unsupervised learning is used for this exercise.

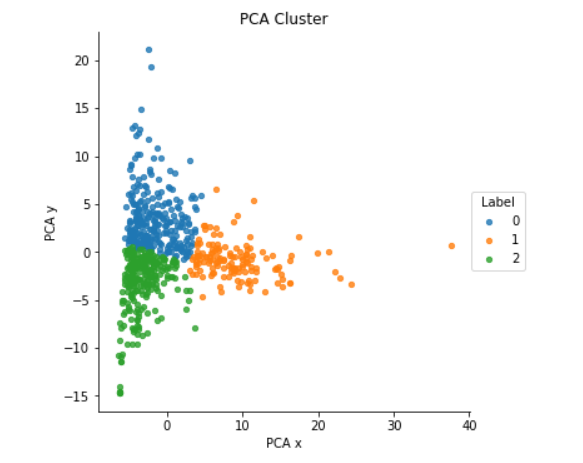
The learning is ran based on distance data split out by day of the month, weekday, weekend, morning, afternoon, and night.

K-Means is chosen for this exercise.

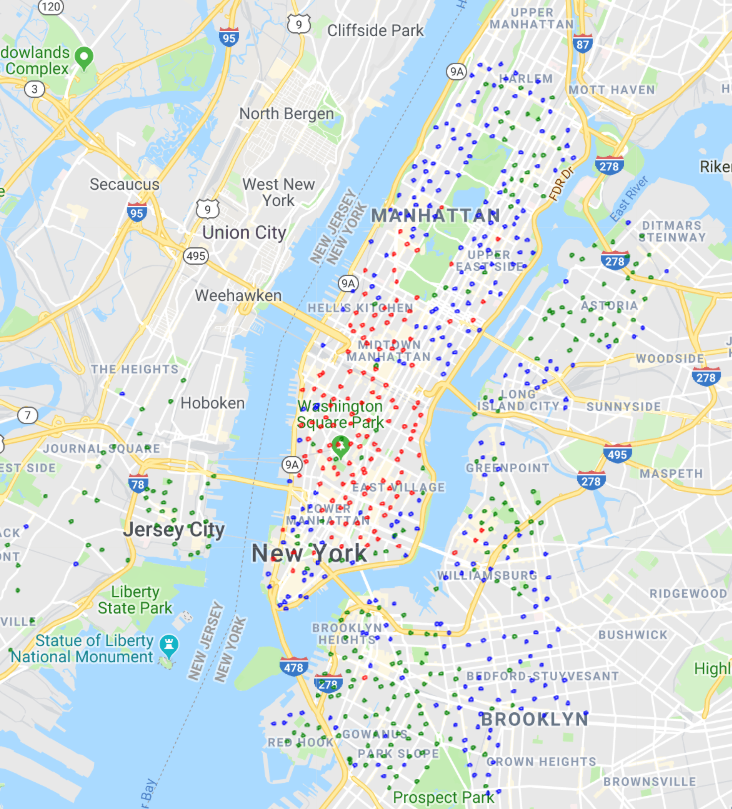
PCA with K=2 is used based on Elbow method.



Cluster of 3 is chosen as it splits the data most cleanly and the clusters do not overlap much. Cluster of 2, 4, and 5 have been tested as well.



Ran on distance of rides for starting station. Also ran on ending station and the results are very similar.



The three clusters are split into Red, Green, and Blue points.

Red Cluster –This is largest cluster accounting for 55% of all rides within the system. Geographically it is located south of Central Park and north of Financial District.

Blue Cluster –This cluster accounts for 28% of all rides. Geographically it is spread out in the outer part of Brooklyn, Long Island City, Upper West Side, Upper East Side, and South of Houston Street.

Green Cluster –This cluster accounts for the remaining 17% riders. Geographically it is spread in central Brooklyn, Jersey City, Astoria, and East Harlem.

Blue Cluster has the longest trips whilst Green has the shortest trips across all periods of a week. This significant differential is quite interesting as both clusters overlap quite a bit in Queens and Brooklyn.

* A likely reason has to do with the location of work for the residences.
* Blue Cluster seems to be a group of people living further away from central business hubs and are thus have longer commuters. This cluster peaked at 2.1KM on Weekday mornings which is significantly higher than all other time periods & clusters. The Upper West, Upper East, and stations near Bridges are most likely used by commuters to Manhattan business districts, and the area on the eastern edge of Brooklyn system is most likely used for people who work in central Brooklyn.
* Green Cluster on the other hand seems to be aggregating riders who are commuting in the local areas in the Brooklyn, Astoria, Jersey City, and East Harlem areas. It also has a peak on Weekday mornings.
* It interesting to note that areas that are similar to each other in terms of demographics and are geographically close by such as Astoria & Long Island City, and Greenpoint & Williamsburg, are clearly split into separate clusters. The primary difference between them is that one is right next to a bridge to Manhattan and the other is not.
* There seems to be a threshold for commuting distance where people switch from taking the train to riding bikes instead.

Red Cluster is unique compared to the Blue and Green Cluster. It does not overlap as much with the other two clusters as it sits almost exclusively in central Manhattan.

* The distance traveled in all periods of the week are relatively consistent. There is no significant increase in distance during Weekday mornings like the other two clusters.
* This illustrates a point the Manhattan riders most likely only commute within Manhattan itself and do not go to other boroughs. The exception to this once again lies near areas close by to bridges.
* The additional population density and cost of owning vehicles in Manhattan adds to the incentive for people to choose bike as a primary source of transportation for a broader range of purpose besides commuting to work.

**Conclusion:**

Based on the observations, we can conclude that rider behaviors is heavily dependent on the location of the starting station.

The information can be applied to help determine placement of future docking stations as it provides insight into how the station will be utilized depending on geographic location.

For these new stations, this analysis can help forecast the rate of bike replacement, utilization rate, loss rate, etc. To get the full forecast, additional data will be needed such as loss data and dock idle data, but that is beyond the scope of this exercise.

An area to expand on for this analysis is to include all 12 months of the year into the data set rather than one month. This will require additional computing power.

Also, utilizing Google Map paid service to determine city distance rather than geometric distance will provide better accuracy on actual distance traveled by the riders.