**Safeguarding the drunk New York City**

**Merrymel George**

**March 28, 2019**

1. **Introduction**
   1. **Background**

Being among the top 15 largest cities of the world and having around 9 million population, New York city is one of the major cities of United States of America. According to reports that came out back in February 2018, around 91% of the New Yorkers drink alcohol. But, the more shocking fact is 78% of the New Yorkers does barhopping, takes drugs and stayed up until sunrise or had a hangover even during weeknights. The percentage of crimes because of the usage of alcohol or drugs is not just significant but alarmingly dangerous. As such, NYPD must be very vigilant during their evenings throughout the night to next morning to avoid/reduce crimes that may arise in this regard.

* 1. **Problem**

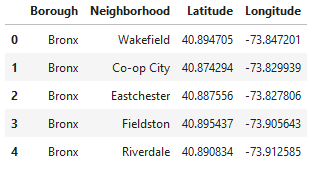
To safeguard New York, NYPD can do continuous patrolling as well as setting up temporary, mobile outposts where NYPD mobile police units/cars can be deployed at nights. There are about 600 pubs, bars and nightclubs spread across the different boroughs of New York. So, these police outposts must be placed in such a way that, they cover the entire pubs, bars and nightclubs. This project aims at finding the location of these outposts. The idea behind the project is that the entire pubs, bars and nightclubs would be bucketed into different groups, and for each group, a location, which is equidistant from all the pubs/bars/nightclubs of that group is chosen as the location for outpost for that location. Based on the opening of new bars, pubs, nightclubs or closing of an existing one, the location of these outposts must be adjusted.

* 1. **Interest**

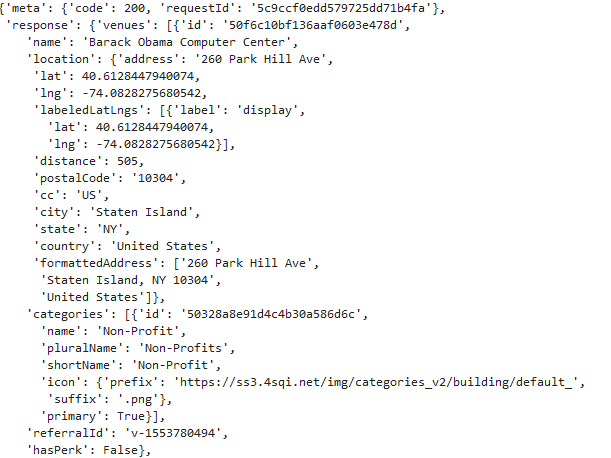
The project, obviously can help the NYPD to better place their mobile units to safeguard the city. This scope of the project can be extended on to various levels, like if combined with crime-location data, we will be able to find locations for outposts which will be more productive, or can be used to find the location to place docking stations for on-demand mobile app-based bicycle rentals, or can be used for mobile app-based cab services like Uber etc.

1. **Data Acquisition and Cleaning**
   1. **Data sources**

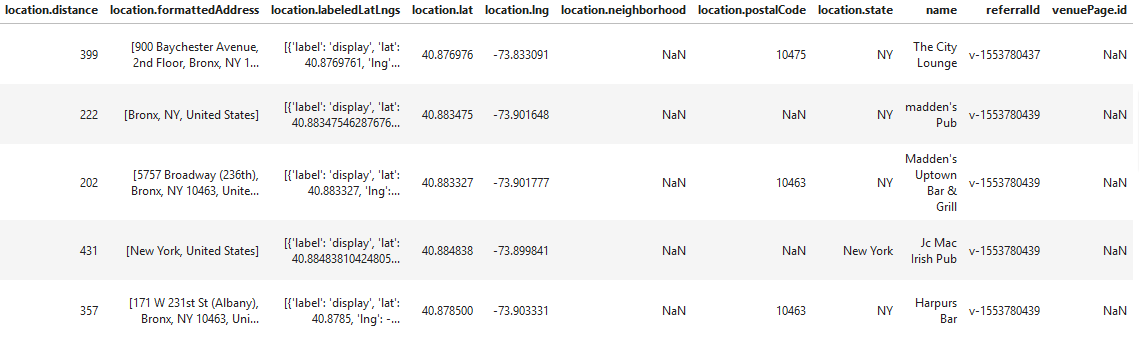
The data that is required for doing the project is acquired from two sources. First dataset, which gives the required data related New York, the different boroughs, neighborhoods, latitudes and longitudes are acquired from a data source which is freely available - <https://cocl.us/new_york_dataset> . Then, with the help of Foursquare API, the other dataset, which is data of the pubs, bars and nightclubs in New York are acquired. A set of sample records of the acquired data is shown below:



Then, with the help of Foursquare API, the other dataset, which is data of the pubs, bars and nightclubs in New York are acquired. Required Foursquare credentials are created and is used to pull data with details like name of the pub/bar/nightclub, location details etc for each neighborhood of the boroughs of New York. A sample of the data that is pulled using the Foursquare API is shown below.

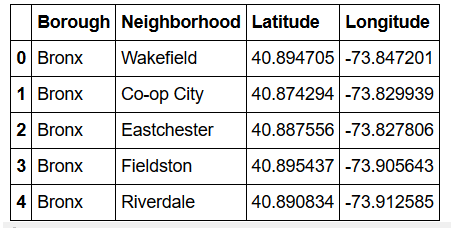


This data needs to be cleansed, wrangled and munged to the required data like pub/bar/nightclub name, location, latitude, longitude etc can be extracted. A sample set of extracted records are shown below.



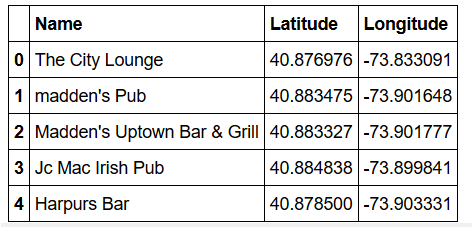
**2.3 Data cleansing, munging and wrangling**

The downloaded New York data is loaded in data frame with proper column names as shown below.



The data related to the pubs, bars and nightclubs that are there in New York, which is pulled using the Foursquare API is loaded as a data frame by normalizing to extract the tags and values corresponding to the different columns. Then, all the records which has NULL values for latitude, longitude and name columns are dropped, followed by dropping the duplicates for the three column combination of name, latitude and longitude as well. The data that was pulled using Foursquare API has certain records which doesn’t have latitude and longitude values. These records are bad records and hence have to be deleted. Similarly, the data also contains records which duplicate. Any 2 records with same combination of name, latitude and longitude must be considered as duplicates, and hence, only one of those records should be retained. After doing this cleansing, another dataframe is created from this, with only the columns that are required for analysis.

The Foursquare API provided 647 records. After the data cleansing, we got 564 records, meaning there are 564 pubs/bars/nightclubs that are there in New York.

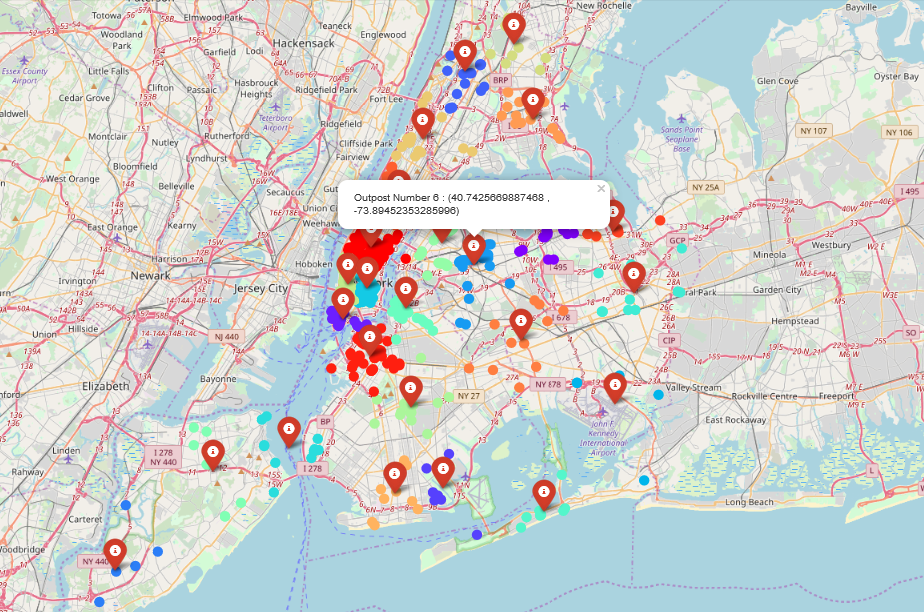


1. **Data analysis methodology**

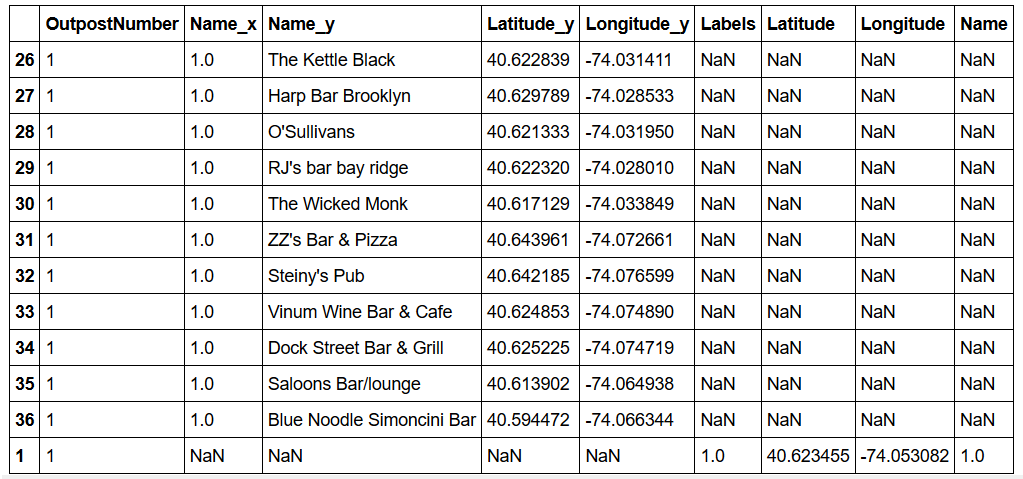
Using the latitude and longitude values of pub/bar/nightclub data collected, we create the data for clustering by using the standard scaler fit and transform. Then, using the decided number of planned police outpost number (this is any number chosen based on the resource availability of NYPD, the number is taken as 25 in the project), we run KMeans algorithm and created the corresponding best clustering for each latitude, longitude pair. Each of the latitude, longitude pair is then tagged with the cluster label of the cluster to which it was attached. The idea behind the project is to find locations for placing the Police outposts in a such way that these outposts are equidistant from all the pubs/bars that are tagged to it. Hence, clustering of the latitude, longitude values of the pubs/bars was done using KMeans. And, to find the outposts, the mean value of the latitude and longitude values of each of cluster was calculated. First the map is created for all the pubs/bars/night clubs latitudes and longitudes and marked into as different clusters. Then, on top of the same map, the centroids of each clusters are marked. These centroids are the location where the Outpost for each cluster should be placed. Based on the closing and opening of bars/pubs, the clustering has to be rerun to get the appropriate outpost location. Similarly, based on the available resource/infrastructure capability, the number of outposts can be increased or decreased, and correspondingly, the clustering and centroid allocation has to redone.

1. **Result**

The map clearly shows the location where the Outposts have to located so that all the pubs/bars in New York can be more or less patrolled on an efficient method. The map with the clusters(as different colored dots ) and the outpost locations (info-sign baloon) created is shown below.



We can find the pubs/bar/nightclubs which each Police Outpost has to monitor from the map and as well as from the data frame as shown below.



The records with the NaN values for Name\_x, Name\_y, Latitude\_y and Latitude\_y corresponds to the Outpost location, while all the other records corresponds to bars/pubs/nightclubs which are under that Outpost.

1. **Discussion**

Here, the clustering has been done on the basis of Euclidean distance and hence the cluster sizes are not equal. This means that, there are certain Police outposts which has more bars/pubs under it for patrolling. Hence, based on the number of points in each cluster, NYPD has to allocate sufficient resources to the Outposts. We can also do the same with equal sized clusters. The equidistant approach was used here because, choosing the centroid will help deployment and arrival of the backup police resources from the outpost in case of emergency situation, to any point in the cluster with the same time consumption.

1. **Conclusion**

Using the clustering technique and plotting the points on the map, it will be easy for NYPD to easily locate the Outpost location and set it up there. Thereafter, by allocating sufficient Police cars to each outpost, NYPD can be continuously patrol the entire New York area. Since, the Outposts are placed at a location which equidistant from each point in the cluster, in case of some emergency, the police resources can reach to the danger location with efficiency. By applying this, NYPD will be able to safeguard New York better.