**Analyzing Neighborhoods and Recommending an area to start a restaurant in Brooklyn, New York**

**IBM Capstone Project**

**Prajwal Nagaraj | January 2021**

**Contents:**

Introduction 3

Data 4

Methodology 5

Data Visualisation 6

Data Wrangling 9

Clustering 12

Discussion & Results 13

Conclusion 16

**Introduction:**

With the increasing number of restaurants it is getting harder and harder to start and sustain a new restaurant in huge cities such as New York. In 2018 alone there were approximately 27 thousand restaurants in New york city out of which around 6500 restaurants were in Brooklyn alone. And these numbers are expected to grow with time. With this huge number of restaurants, competition will obviously arise and some restaurants become self sufficient while others fail.

How to reduce the chance of failure? Humans cannot control adverse weather effects or natural disasters or a Global Pandemic like the current situation which will inevitably result in the failure of not only restaurants any business for that matter. Although predicting a success or failure of a restaurant is difficult it is not impossible as we can control some of the conditions such as the location of the restaurant which can act as one of the main attributes to the success of a restaurant. Can we use the location of the neighborhood and the venues around it to predict success or failure of the restaurant? Yes, by examining the other venues around a neighborhood, we can predict if a new restaurant is going to succeed or not.

In this project I will be using KMeans clustering which is an unsupervised clustering algorithm to cluster neighborhoods which will be suitable to start a new restaurant by considering other venues in the neighborhood and ranking them from most common to least common.

**Data:**

The data used in this project is a dataset from New York city Department of Finance and consists properties that were from january 2020 to december 2020 and all the details along with it like the zip code, neighborhood name, tax class, sale price, block number, lot number, building class, land square feet, address and the sale date.

But we will only be using the Neighborhood, Address and the zip code to build our model as we will use zip codes to merge another dataframe which was retrieved from [GeoNames](https://www.geonames.org/) will give us the coordinates of every neighborhood. Once merged the dataframe will consist of Neighborhood, Address, Zip code, latitude and longitude.

Once the above dataframe is ready, we will convert all the multiple entries of the neighborhood into a single entry and take the centre of the neighborhood as its coordinates. This will be our final dataframe which will be used to retrieve venues data by Foursquare Api.

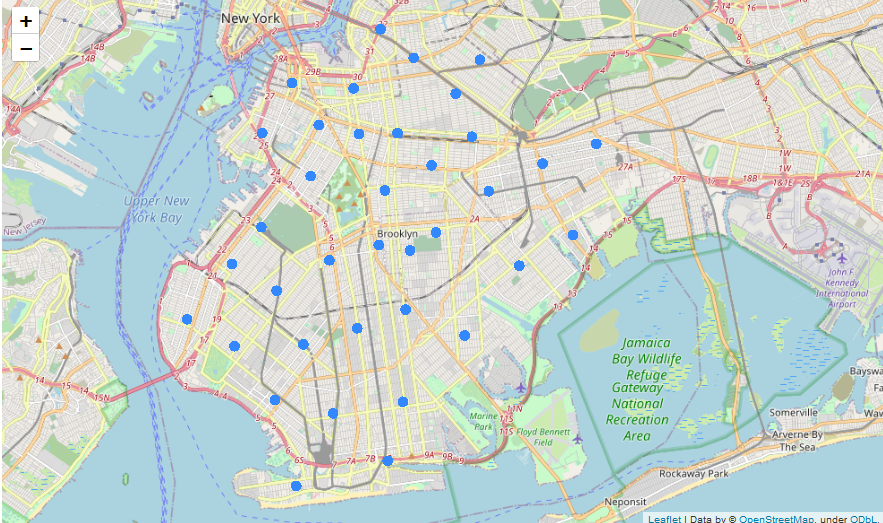
The Foursqaure api will retrieve a maximum of 100 venues within a 3km radius of every neighborhood. It will retrieve the venue’s name, venue’s category and its coordinates which will later be grouped by most common to least common venues.

**Methodology:**

In this project exploratory data analysis is done on Brooklyn, New york to determine which neighborhood is most suitable to start a new restaurant.First the Brooklyn dataset is cleaned for errors and combined with another dataset containing Co-ordinates for every neighborhood according to the zip codes and visualised.Then Foursquare Api is used to fetch the nearby venues in every neighborhood and the top most common venues is identified in decreasing order of occurence. The Neighborhood data and the venues data are combined and unsupervised clustering is done by KMeans clustering to get the Neighborhoods with similar venues. After clustering each neighborhood is examined and analyzed and a cluster is chosen as the best cluster to start a new restaurant based only on the other venues in that cluster.

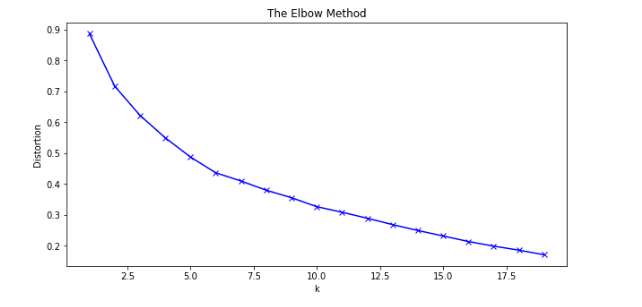
**Data Visualisation:**

Visualising the data for better understanding of the dataframe and to remove any outliers.

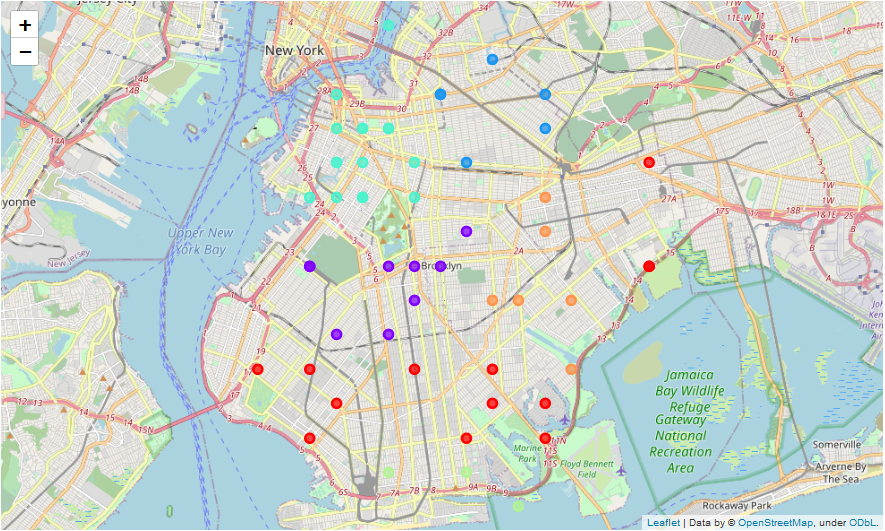
****

All the neighborhoods in the brooklyn dataset visualised on a map using Folium. When clustering the Neighborhoods we need to first determine the number of clusters, for this we use elbow method to determine the number of clusters.

In this method,the distortion is plotted against the K value the inflection point of the graph gives us the optimal value of K as shown in the figure below.

****

After clustering is done, we again generate a map of brooklyn using Folium with the cluster labels as shown.

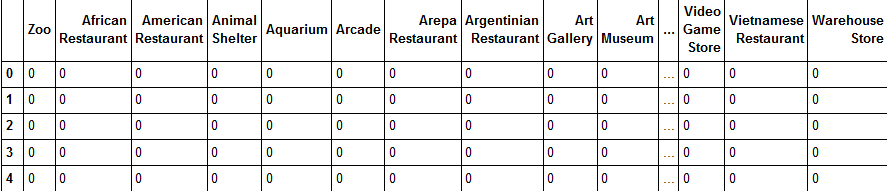
****

**Data Wrangling:**

In the Dataset that we’re dealing with, we first check for any missing values and if there were any missing values we dropped that entire row because we cannot find mean for neighborhoods, address because they are strings.

Zip codes and Coordinates on the other hand are numerical values and hence finding mean is possible, but once we find mean replacing the empty values with the mean does not make sense because it would change to location to some ambiguous coordinates and hence will lead to errors in the future.This dataframe’s columns are then converted from object type to either integer, floating point or string type as the Folium library cannot understand object type data

Once the missing data is dealt with,we use the Foursquare Api to retrieve all the venues in that neighborhood and create another dataframe. We then transform the whole data set by One-Hot Encoding which converts categorical variable into dummy variable by assigning either 1’s or 0’s.

****

All the entries which have zoo or african restaurant as the venue will be assigned 1’s and those which dont have these venues will be assigned 0’s.

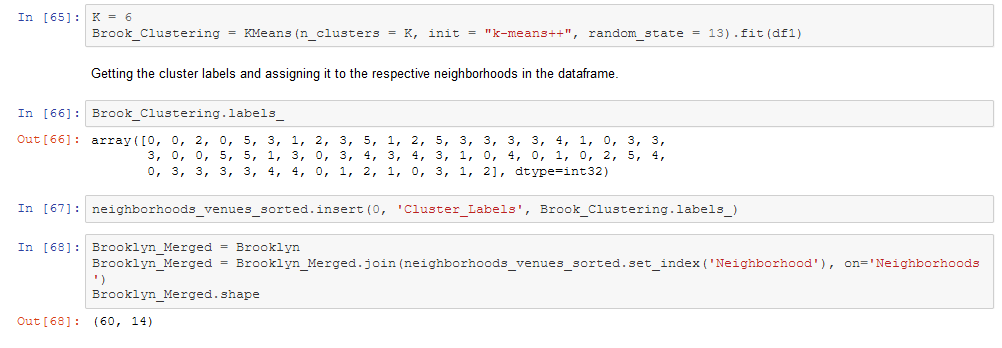
****

Once the Dataframe is one hot encoded we will use a function to group the venues by their decreasing order of occurrence and add back the neighborhood column to the dataframe.

**Machine Learning Models:**

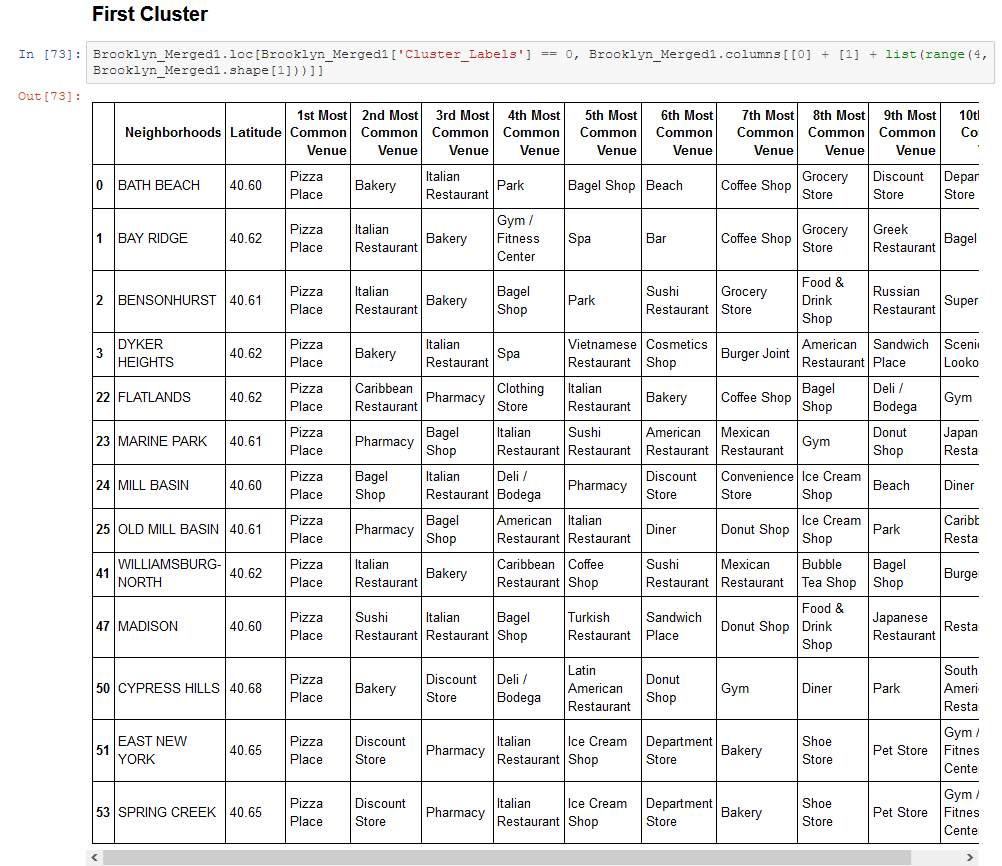
Kmeans Clustering:

K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

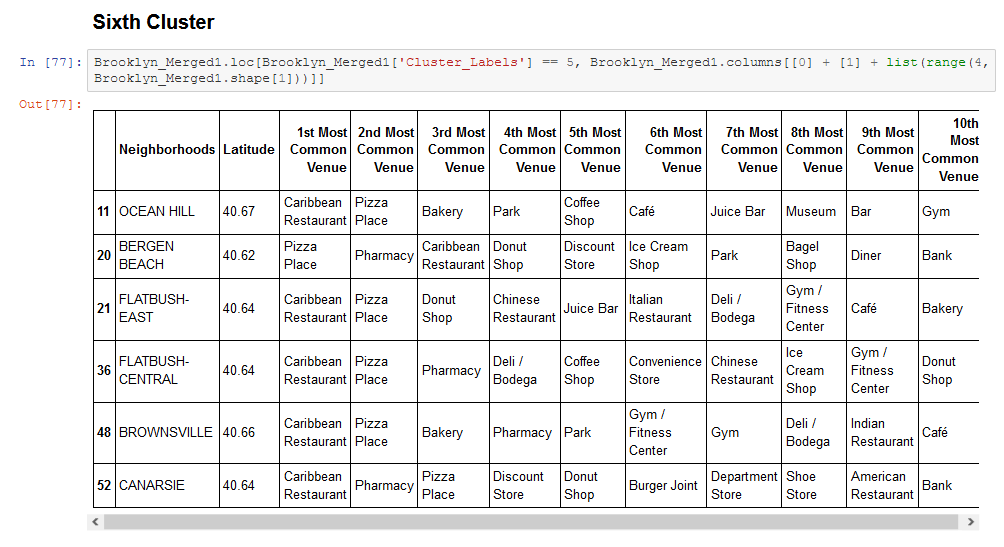


After choosing the best K value, we cluster the neighborhood and get the cluster labels and add the labels to the dataframe so that it is easier to visualise it using Folium.

**Discussion & Results:**

****

By examining all the clusters obtained we can see that a few clusters are very well suited to open a new restaurant, whereas others are suited to house other venues such as a bar, cafe, gym etc. Neighborhoods in the first and the sixth cluster are ideal to open a new restaurant as the first most common venues in both the clusters are restaurants. The neighborhoods in the third and fourth cluster on the other hand are more accepting of a new bar or a cafe as those two venues are the most common in that cluster. Neighborhoods in the second and the fifth cluster have a wide range of venues around them but the second cluster is leaning more towards cafe's and restaurants but venues on the fifth cluster on the other hand is very broad and has everything from beaches to bars to grocery stores. Thus only the first and sixth cluster are suitable to open a restaurant.



Both the first and the sixth cluster have a lot of restaurants with different cuisines such as carribean restaurant, pizza place, American restaurant and Chinese restaurant but in the first cluster both the 1st most common venues and the 2nd most common venues are either pizza place's or italian restaurants with an exception of one carribean restaurant. On the other hand, in the sixth cluster the 1st most common venues are all carribean restaurants with an exception of one pizza place. Because both the clusters are contrasting with respect to the cuisines offered in the restaurants, it comes down to the choice of cuisine offered by the new restaurant, if the cuisine is italian then the first cluster is most suitable to open an italian restaurant or a pizza place. This is because even though the 6th cluster has majority of pizza places as the second most common venues, the number of venues or in the first cluster is double the number of venues in the sixth cluster and the venues in the first cluster have italian restaurant or pizza place as the top most common venue. If the cuisine offered by the new restaurant is carribean, then the 6th cluster is the most suitable to open a carribean restaurant.

**Conclusion:**

All the clusters were examined and the neighborhoods in first and sixth cluster are best suited to start a restaurant depending on the new restaurant’s cuisine as the venues in these clusters are mostly made up of either italian or carribean restaurant.Valuable insights were gained into the venues and popular places in Brooklyn, New york and a better understanding of the diverse places and their influence on a new restaurant.