Coursera Applied Data Science Capstone Project

The Battle of Neighborhood

Author: Anoop Kohale, IBM

# Introduction/Business Problem

Market analysis is an important part of any business start-up. The success or failure depends on the location where the business is opened. In big cities like New York, there is lot of competition to attract customers with your offerings. However, selection a perfect neighbourhood is often difficult and require lot of work.

Opening any business in any neighbourhood requires carefully analysing number of aspects of the business for it to become successful. There are number of factors influencing the decision, viz. will there be enough customers to buy my offerings, is there any competition around that can take away by revenue, etc. To success the businessman must carefully study these factors and come up with the strategy or plan of operating the business in the neighbourhood.

Fortunately, Advanced data analysis and machine learning will help taking this decision with the information available in abundance around the internet. Foursquare is such an information provider. Foursquare provides data about the interesting venues around any neighbourhood. We can utilize the machine learning algorithms and find out the clustering of specific business in the neighbourhood. This will empower us with the understanding of demographics and we can then take better decision that will result in making the business a success.

The main beneficiary of this project will be any entrepreneur who wishes to open a business in big city like New York. The project will try to find a suitable neighbourhood to open a business. For example, someone wants to open a bakery and is looking for a suitable neighbourhood, this project will give an insight on to the venues in a neighbourhood and then can decide whether opening a bakery there will be a suitable option. E.g. if there are other bakeries in the neighbourhood it may not be a suitable option.

# Data to be used

Any location you open a business has number of other similar businesses operating. One needs to analyse the data available at your hands to come up with a decision. During this project we will gather data from various data sources. Below is the list of the data sources used for this project.

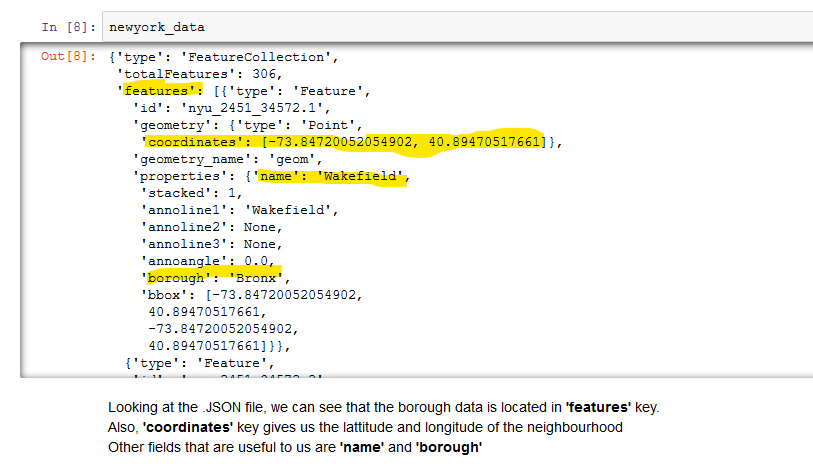
1. Foursquare Venues data
   1. Type: API Call to Foursquare. JSON data about the venue.
   2. Description: The data has various venues around a location within specific radius. Venues are categorized and reviewed by users of Foursquare
   3. Source: <https://www.foursquare.com>
2. Geocoder data
   1. Type: Latitude and Longitude data for given location
   2. Description: The latitude and longitude data of a given location can be extracted using GEOCODER library
   3. Source: Geocoder library
3. Neighbourhood data
   1. Type: Neighbourhoods around New York city
   2. Description: Neighbourhoods of New York city
   3. Format: GeoJSON data
   4. Source: <https://geo.nyu.edu/catalog/nyu-2451-34572>

The main features of the data will be neighbourhood and their latitude and longitude. Foursquare API will provide data about the venues near by the latitude and longitude of the neighbourhood. This data will include category of the venue, its popularity in terms of user ratings and other related data.

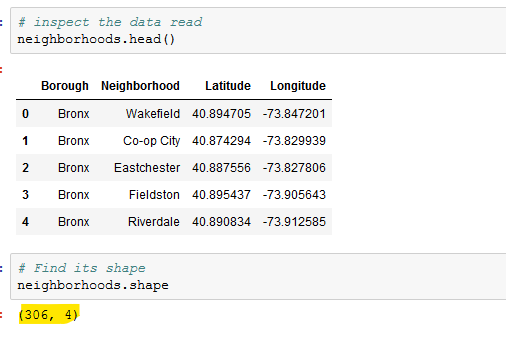
Using this information, we can cluster the venues using clustering techniques. The clusters then will be visualized to take decision.

# Methodology

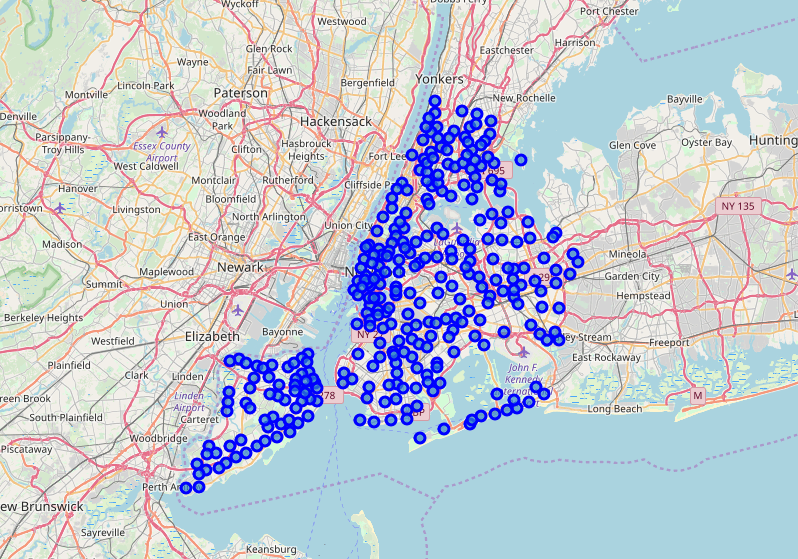
The data about the Ney York neighbourhoods is made publicly available by NYU. The data is in GeoJSON format. The sample data is shown below.



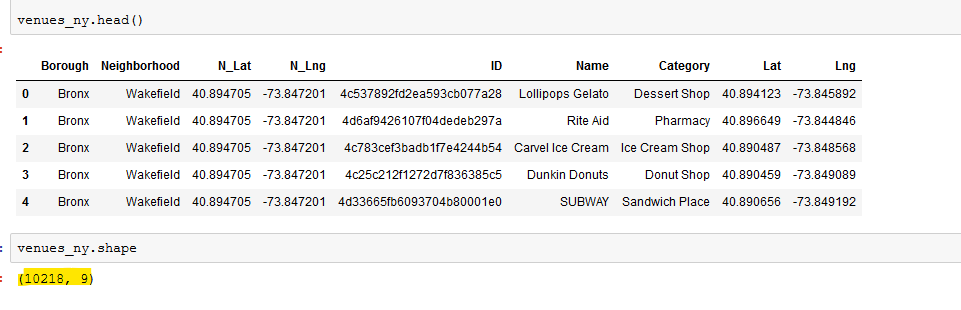
After data clean up relevant fields are populated in a dataframe. There are total 306 neighbourhoods in New York.



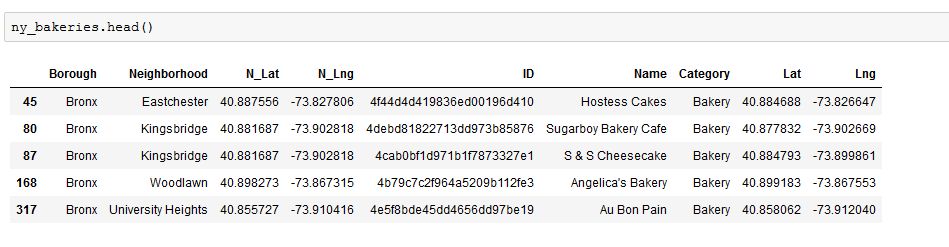
The neighbourhoods are plotted on the Folium map. Below is the visualization of the neighbourhoods in New York.



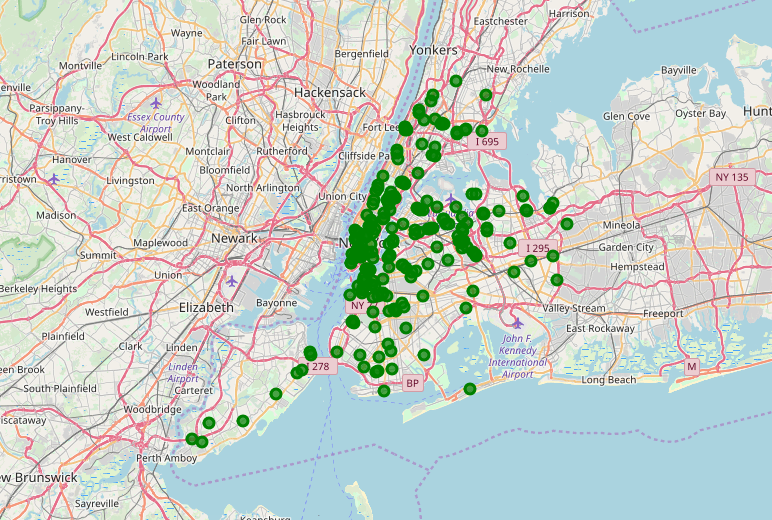
Foursquare API is used to gather the venues around each neighbourhood. The result is then compiled in a dataframe as shown below. We can see that total od 10218 venues are selected by Foursquare API.

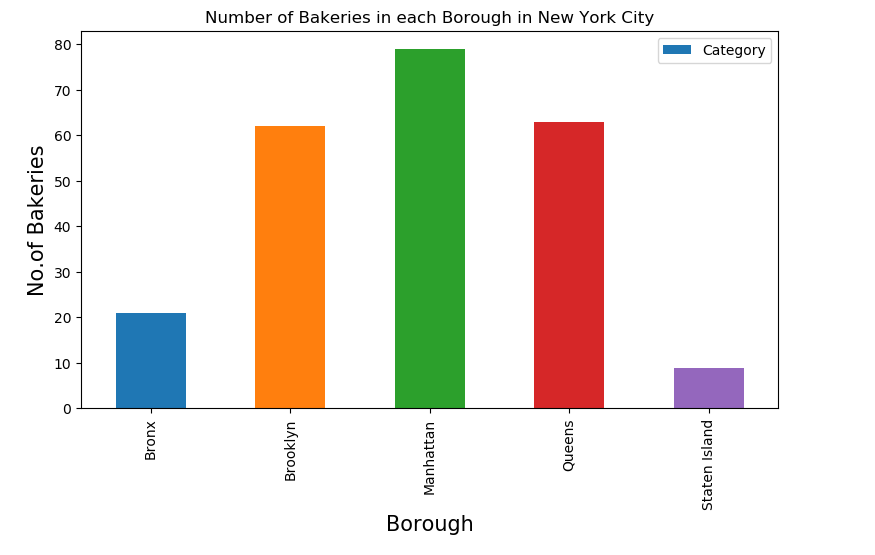


To perform exploratory analysis, we selected venues with **Category = ‘Bakery’**

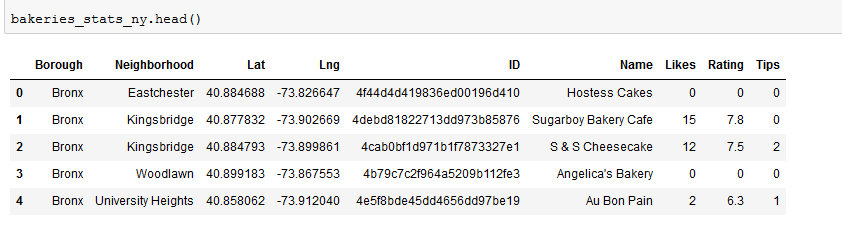


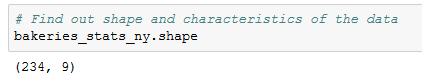
Bakeries in the neighbourhoods of New York are plotted on the map using Folium library.

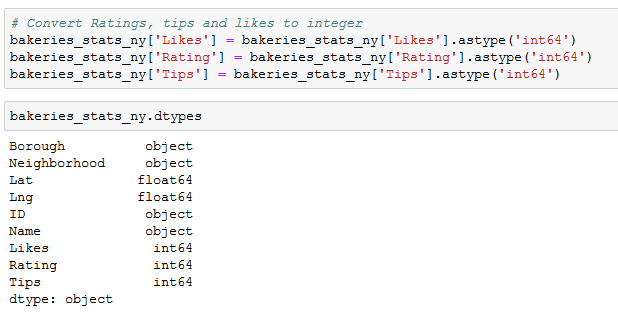


Same information is plotted into a Bar Plot

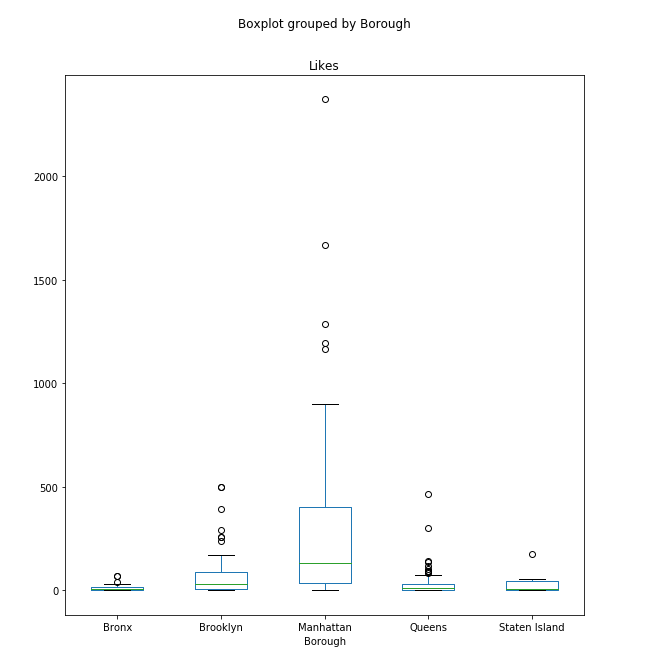
After that, we get the details of each Bakery using premium call to Foursquare API. The result is then captured in a dataframe as shown below.







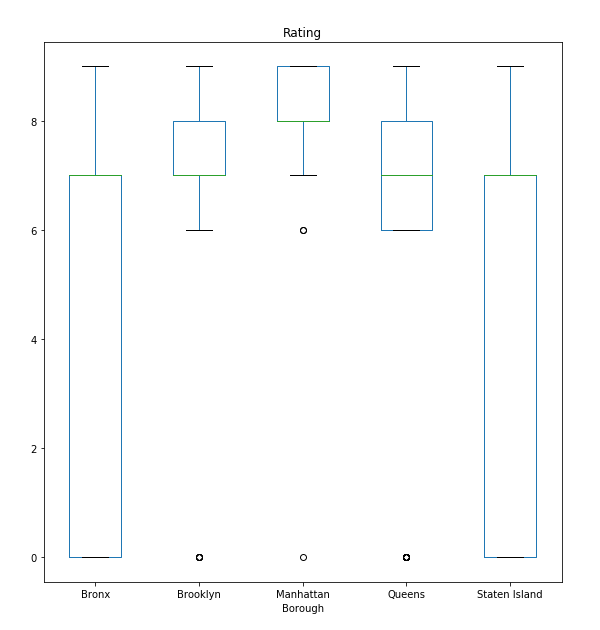
We can observe that there are some bakeries whose data is not available/returned by Foursquare API. We give them all rating, like and tips as 0. Later, to perform unbiased analysis, we remove those entries.



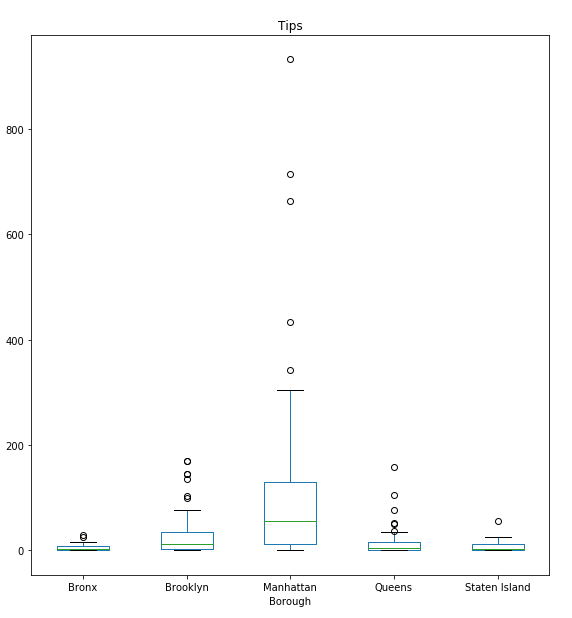
The Bakeries that received likes by the users of Foursquare is plotted in the form of Boxplot. This boxplot shows that Manhattan borough has the highest number of likes to the bakeries.

If we compare this box plot to the earlier bar chart it is obvious that large number of bakeries in Manhattan received large number of likes by the users. Also, interesting to note here is that there are some bakeries in Manhattan that are very popular receiving more than 2000 likes.

In terms of bakeries in other locations, Brooklyn and Queens perform in a similar way receiving similar pattern of likes.



When we looked at the ratings data, we see similar picture. Manhattan bakeries received on an average rating of 8. Bakeries in Brooklyn and Queens are in 2nd place.



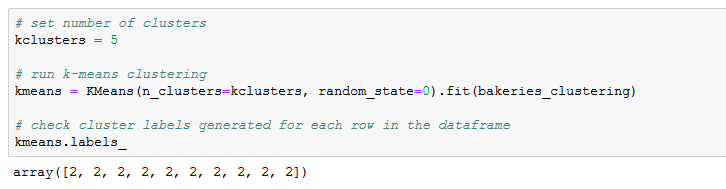
Similar trend is observed for the tips users posted for Bakeries.

# Clustering

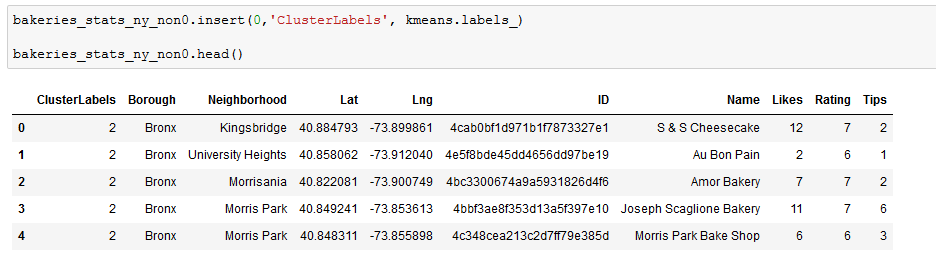
Now the bakeries are clustered using labels “Likes”, ”Rating”, and “Tips”. These attributes define how popular the venue is. More of these means the venue is more popular among the customers. Also, these labels provide insight into the way consumers in the neighbourhood use the certain venue.

We use K-Means clustering algorithm from ScikitLearn library. The K-Means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as prototype of cluster.

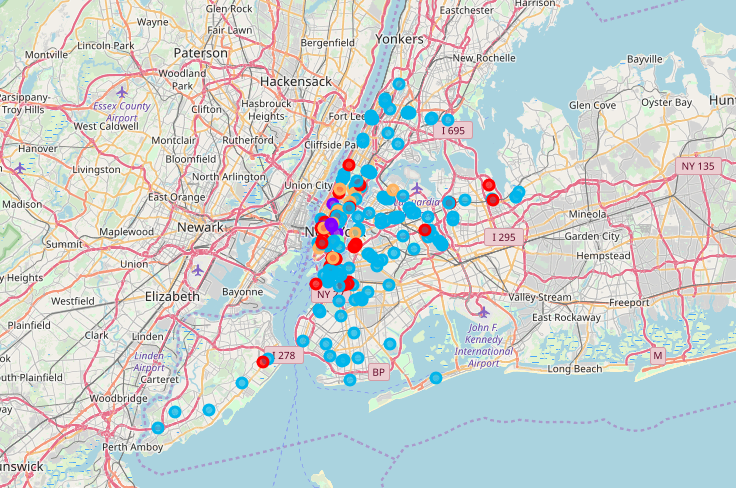
It has been successfully used in market segmentation, computer vision, and astronomy among many other domains. It often is used as a pre-processing step for other algorithms, for example to find a starting configuration.



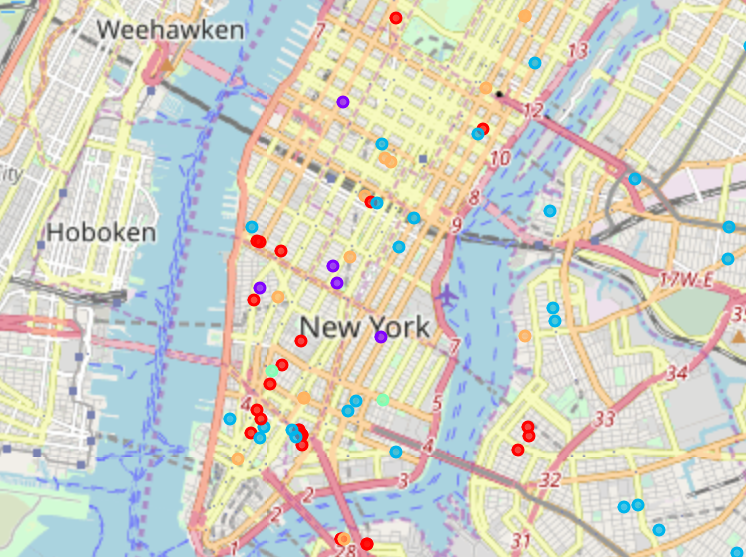
The results of the K-Means clustering results into cluster labels for each row. We then assign the labels to the dataframe.



The K-Means labels are also visualized using Folium Map.

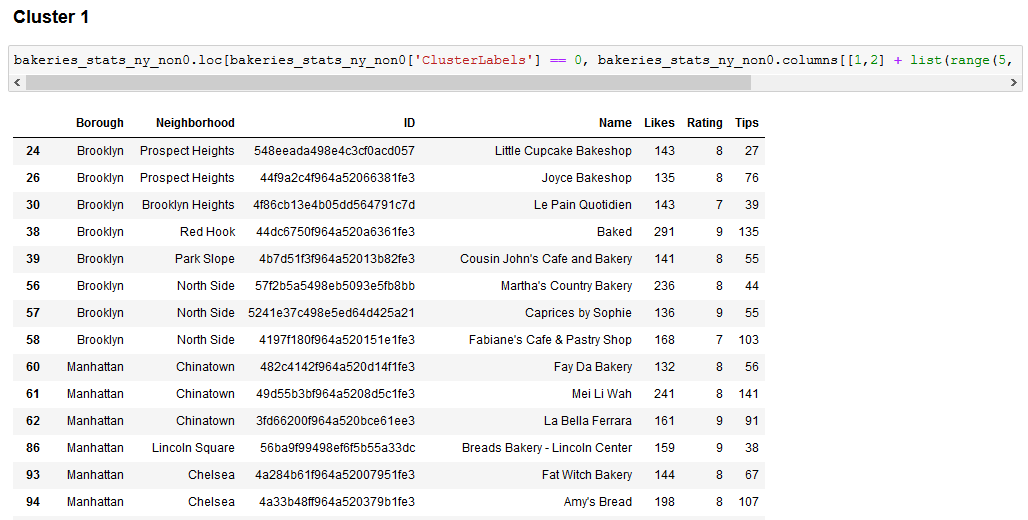


Zoomed in map of New York which displays all the five clusters

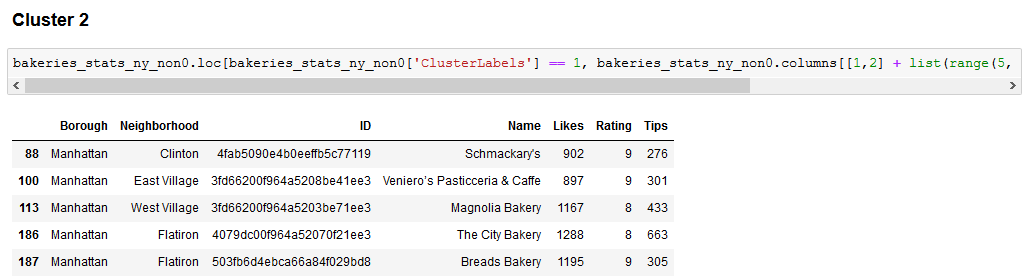


# Review Results

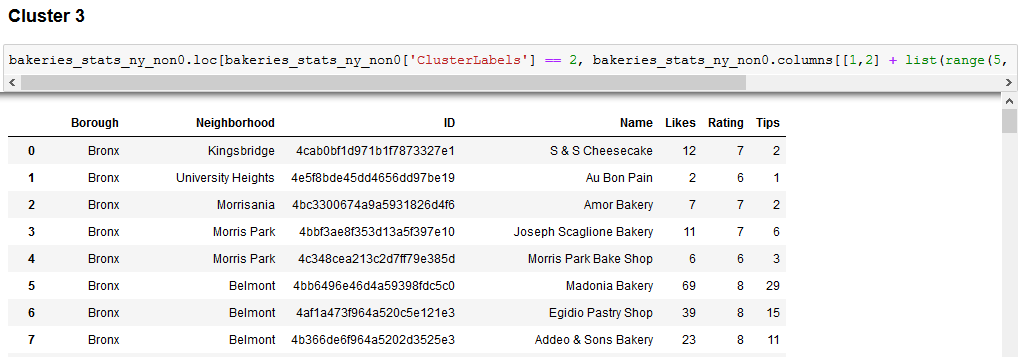
Lets review the results of the clustering exercise and examine each of the clusters. A closer look at each of the clusters is shown below



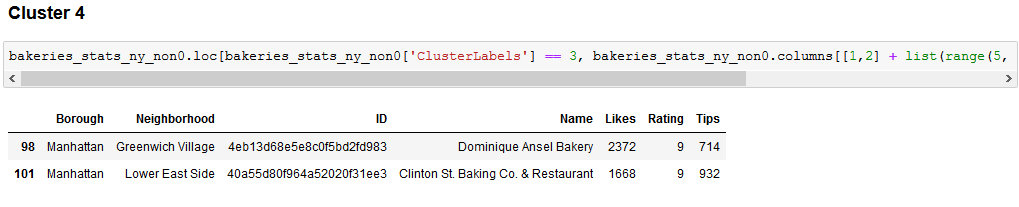
Cluster 1 are the bakeries with moderate likes, tips and ratings. Ratings between 7 to 9, likes between 100 to 300 and tips less than 150.



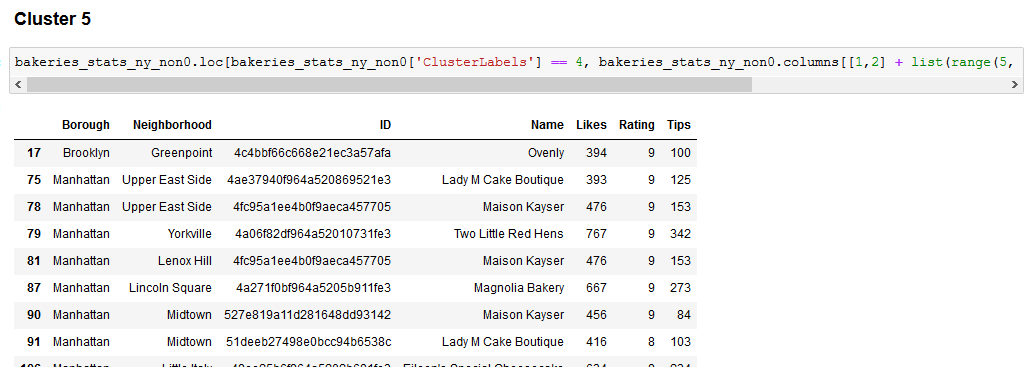
Cluster 2 are the bakeries with high likes, tips and ratings. Ratings between 8 to 9, likes between 900 to 1300 and tips less than between 270 to 650.



Cluster 3 have bakeries with lowest likes, ratings and tips. These bakeries may be in unpopular neighbourhood or they are not so popular among the customers.



Cluster 4 are the highest liked and rated Bakeries. These are the most popular award winning bakeries and restaurants in the New York city.



Cluster 5 bakeries are the moderately popular venues in New York.

# Discussion

As with K-Means clustering, it provides starting point of any decision making process. With our analysis, we can look for starting bakery business in a neighbourhood where there is less number of bakeries and also look at the competition offered around the neighbourhoods. With this approach there is a scope of improvement, some of them include:

* We can add the data about demographics like population density.
* Look at nearby Business to Business opportunities to widen the reach of the business since breads and other bakery products are main ingredients
* Cross verify the location data provided by Foursquare by other location providers like Google Places API.

# Conclusion

While this project does not provide definitive results to the stakeholders or is scientifically accurate, it provides a good starting point to the problem of finding out the business location to open a bakery outlet. The existing bakeries that are popular should be avoided and moderately popular neighbourhoods where there are a smaller number of bakeries can be selected for the new venture.

This exercise also showcases the great potential of data science in the field of location scouting without visiting the exact venues. The data visualization tools such as Folium map library, various plots offer great insights in the data and help make decision.

Thank you for reading the report!