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

```
In [8]: newyork_data

Out[8]: {'type': 'FeatureCollection',
         'totalFeatures': 306,
         'features': [{'type': 'Feature',
           'id': 'nyu_2451_34572.1',
           'geometry': {'type': 'Point',
            'coordinates': [-73.84720052054902, 40.89470517661]},
           'geometry_name': 'geom',
           'properties': {'name': 'Wakefield',
            'stacked': 1,
            'annoline1': 'Wakefield',
            'annoline2': None,
            'annoline3': None,
            'annoangle': 0.0,
            'borough': 'Bronx',
            'bbox': [-73.84720052054902,
             40.89470517661,
             -73.84720052054902,
             40.89470517661]}},
          {'type': 'Feature',
```

Looking at the .JSON file, we can see that the borough data is located in **'features'** key.
Also, **'coordinates'** key gives us the lattitude and longitude of the neighbourhood
Other fields that are useful to us are **'name'** and **'borough'**

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

```
: # inspect the data read
  neighborhoods.head()
```

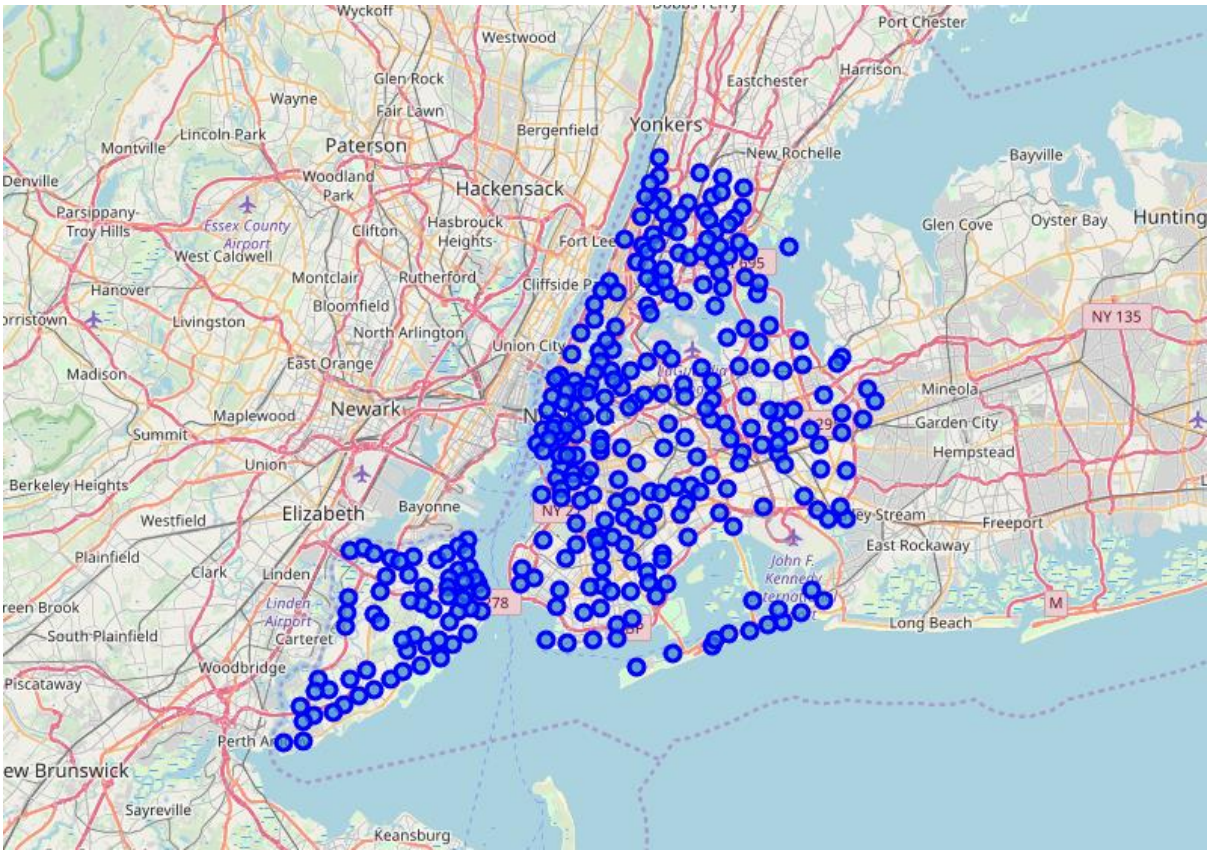
:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

```
: # Find its shape
  neighborhoods.shape
```

: (306, 4)

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.

```
venues_ny.head()
```

	Borough	Neighborhood	N_Lat	N_Lng	ID	Name	Category	Lat	Lng
0	Bronx	Wakefield	40.894705	-73.847201	4c537892fd2ea593cb077a28	Lollipops Gelato	Dessert Shop	40.894123	-73.845892
1	Bronx	Wakefield	40.894705	-73.847201	4d6af9426107f04dedeb297a	Rite Aid	Pharmacy	40.896649	-73.844846
2	Bronx	Wakefield	40.894705	-73.847201	4c783cef3badb1f7e4244b54	Carvel Ice Cream	Ice Cream Shop	40.890487	-73.848568
3	Bronx	Wakefield	40.894705	-73.847201	4c25c212f1272d7f836385c5	Dunkin Donuts	Donut Shop	40.890459	-73.849089
4	Bronx	Wakefield	40.894705	-73.847201	4d33665fb6093704b80001e0	SUBWAY	Sandwich Place	40.890656	-73.849192

```
venues_ny.shape
```

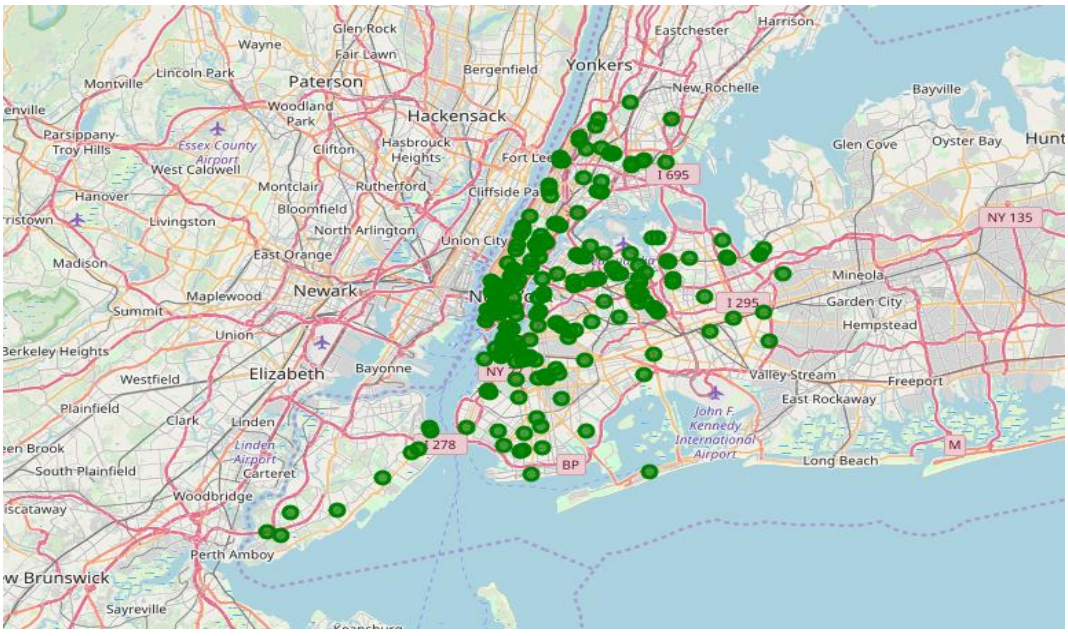
```
(10218, 9)
```

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

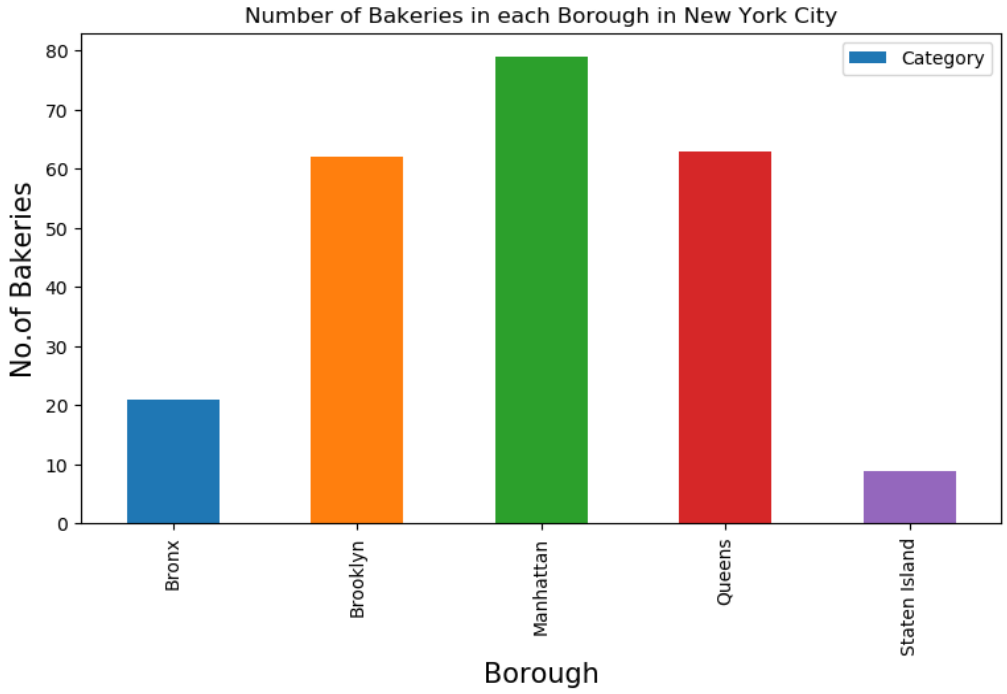
```
ny_bakeries.head()
```

	Borough	Neighborhood	N_Lat	N_Lng	ID	Name	Category	Lat	Lng
45	Bronx	Eastchester	40.887556	-73.827806	4f44d4d419836ed00196d410	Hostess Cakes	Bakery	40.884688	-73.826647
80	Bronx	Kingsbridge	40.881687	-73.902818	4debd81822713dd973b85876	Sugarboy Bakery Cafe	Bakery	40.877832	-73.902669
87	Bronx	Kingsbridge	40.881687	-73.902818	4cab0bf1d971b1f7873327e1	S & S Cheesecake	Bakery	40.884793	-73.899861
168	Bronx	Woodlawn	40.898273	-73.867315	4b79c7c2f964a5209b112fe3	Angelica's Bakery	Bakery	40.899183	-73.867553
317	Bronx	University Heights	40.855727	-73.910416	4e5f8bde45dd4656dd97be19	Au Bon Pain	Bakery	40.858062	-73.912040

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.

```
bakeries_stats_ny.head()
```

	Borough	Neighborhood	Lat	Lng	ID	Name	Likes	Rating	Tips
0	Bronx	Eastchester	40.884688	-73.826647	4f44d4d419836ed00196d410	Hostess Cakes	0	0	0
1	Bronx	Kingsbridge	40.877832	-73.902669	4debd81822713dd973b85876	Sugarboy Bakery Cafe	15	7.8	0
2	Bronx	Kingsbridge	40.884793	-73.899861	4cab0bf1d971b1f7873327e1	S & S Cheesecake	12	7.5	2
3	Bronx	Woodlawn	40.899183	-73.867553	4b79c7c2f964a5209b112fe3	Angelica's Bakery	0	0	0
4	Bronx	University Heights	40.858062	-73.912040	4e5f8bde45dd4656dd97be19	Au Bon Pain	2	6.3	1

```
# Find out shape and characteristics of the data
bakeries_stats_ny.shape
```

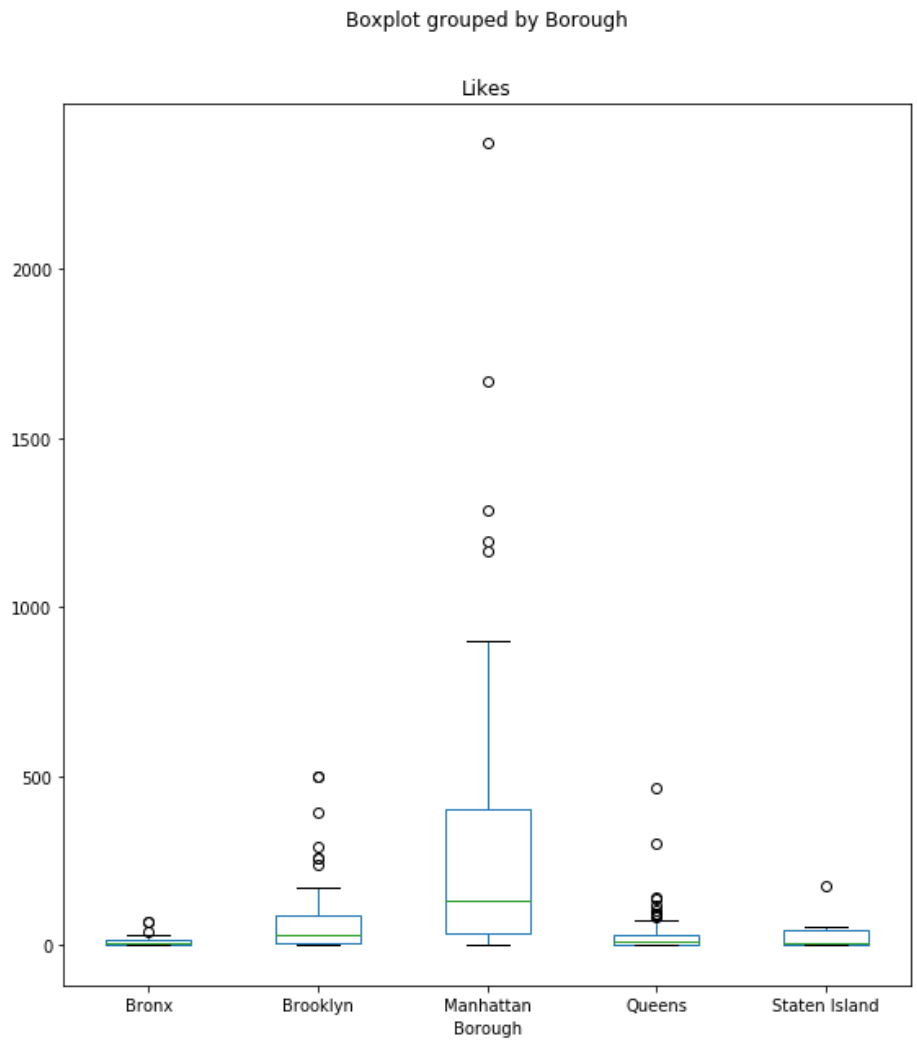
```
(234, 9)
```

```
# Convert Ratings, tips and likes to integer
bakeries_stats_ny['Likes'] = bakeries_stats_ny['Likes'].astype('int64')
bakeries_stats_ny['Rating'] = bakeries_stats_ny['Rating'].astype('int64')
bakeries_stats_ny['Tips'] = bakeries_stats_ny['Tips'].astype('int64')
```

```
bakeries_stats_ny.dtypes
```

```
Borough          object
Neighborhood     object
Lat             float64
Lng             float64
ID               object
Name             object
Likes             int64
Rating            int64
Tips              int64
dtype: object
```

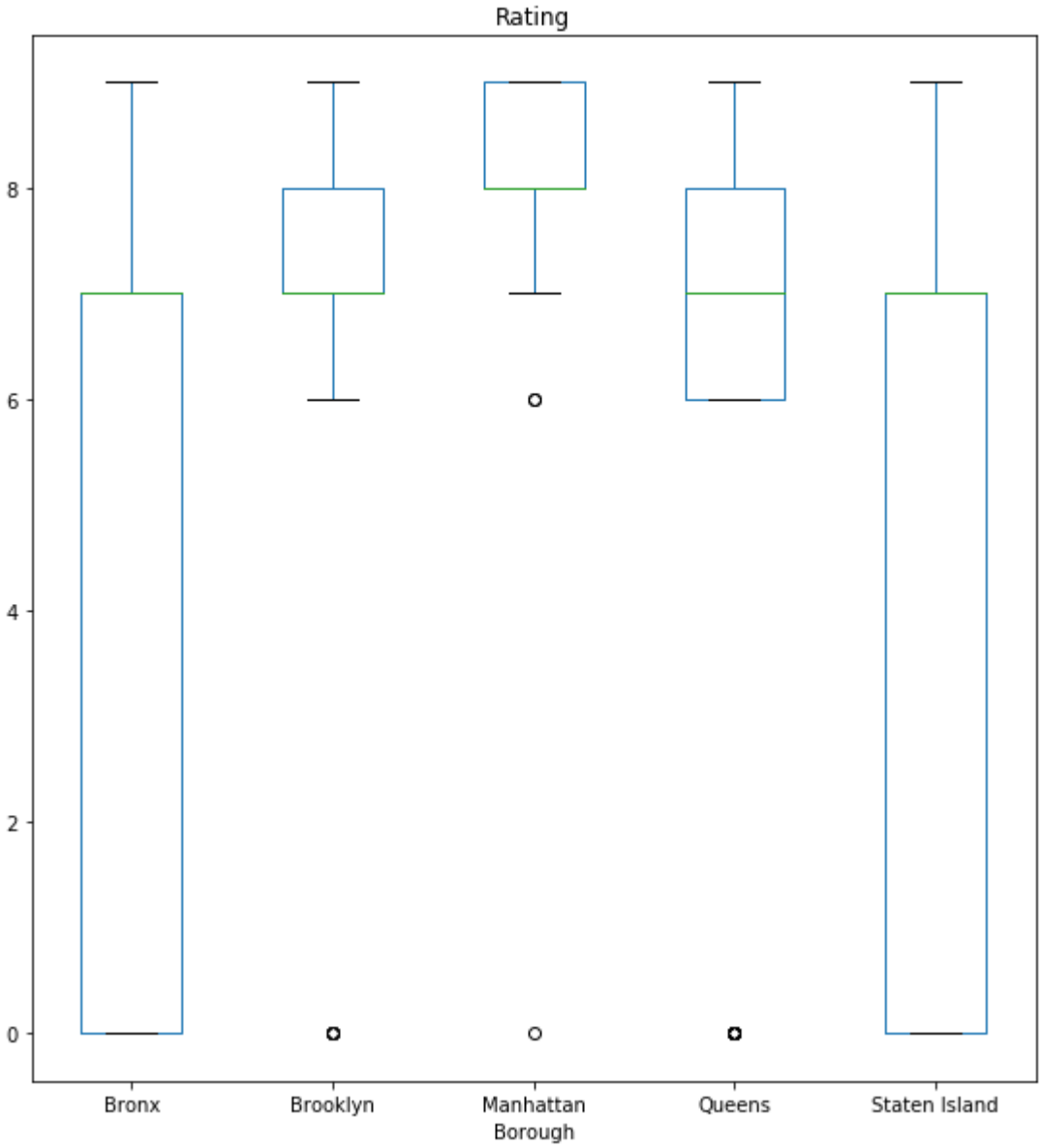
We can observe that there are some bakeries whose data is not available/returned by Foursquare API. We give them all rating, like and tips as 0. Later, to perform unbiased analysis, we remove those entries.



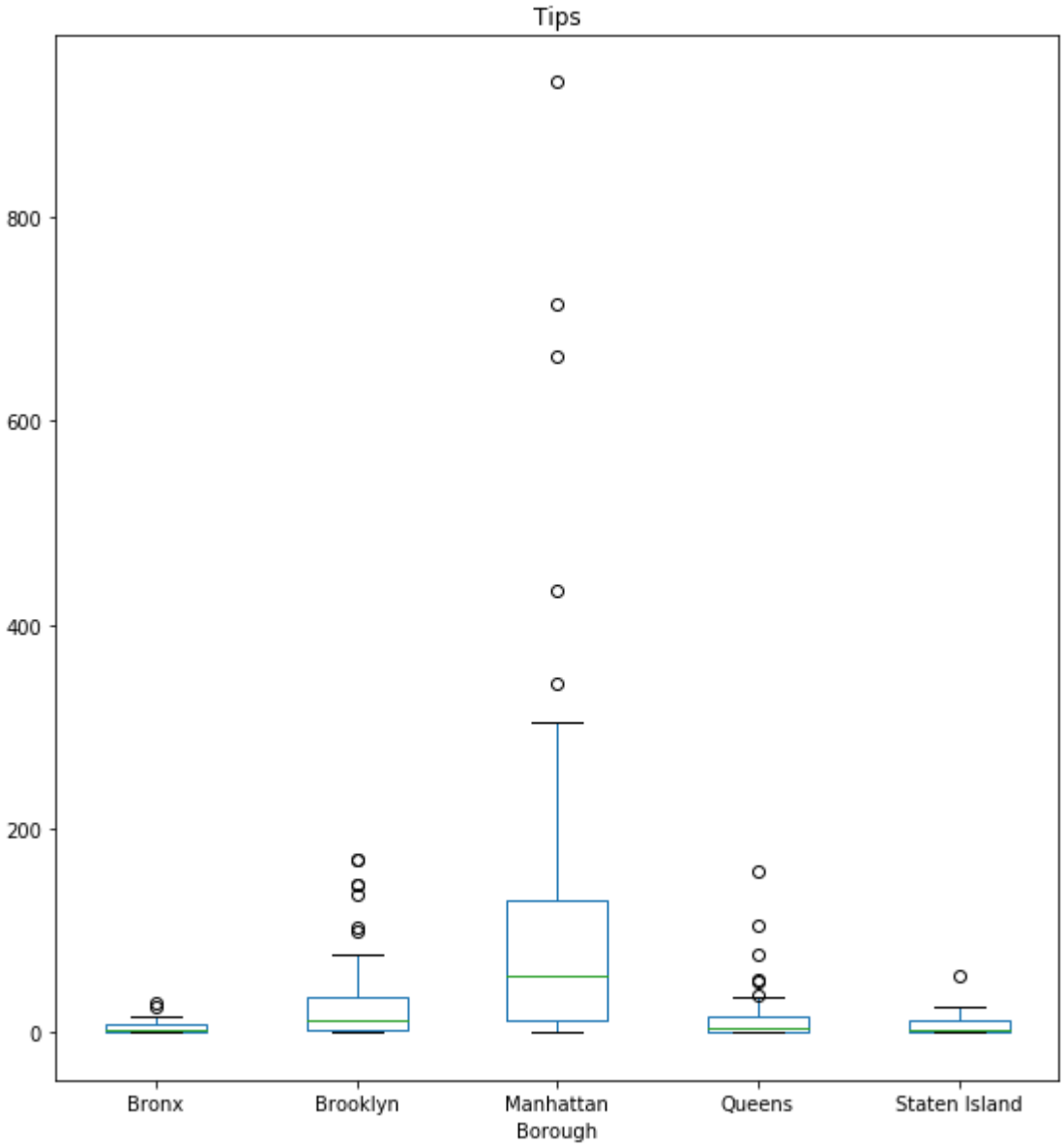
The Bakeries that received likes by the users of Foursquare is plotted in the form of Boxplot. This boxplot shows that Manhattan borough has the highest number of likes to the bakeries.

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.

```
# set number of clusters
kclusters = 5

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(bakeries_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_
```

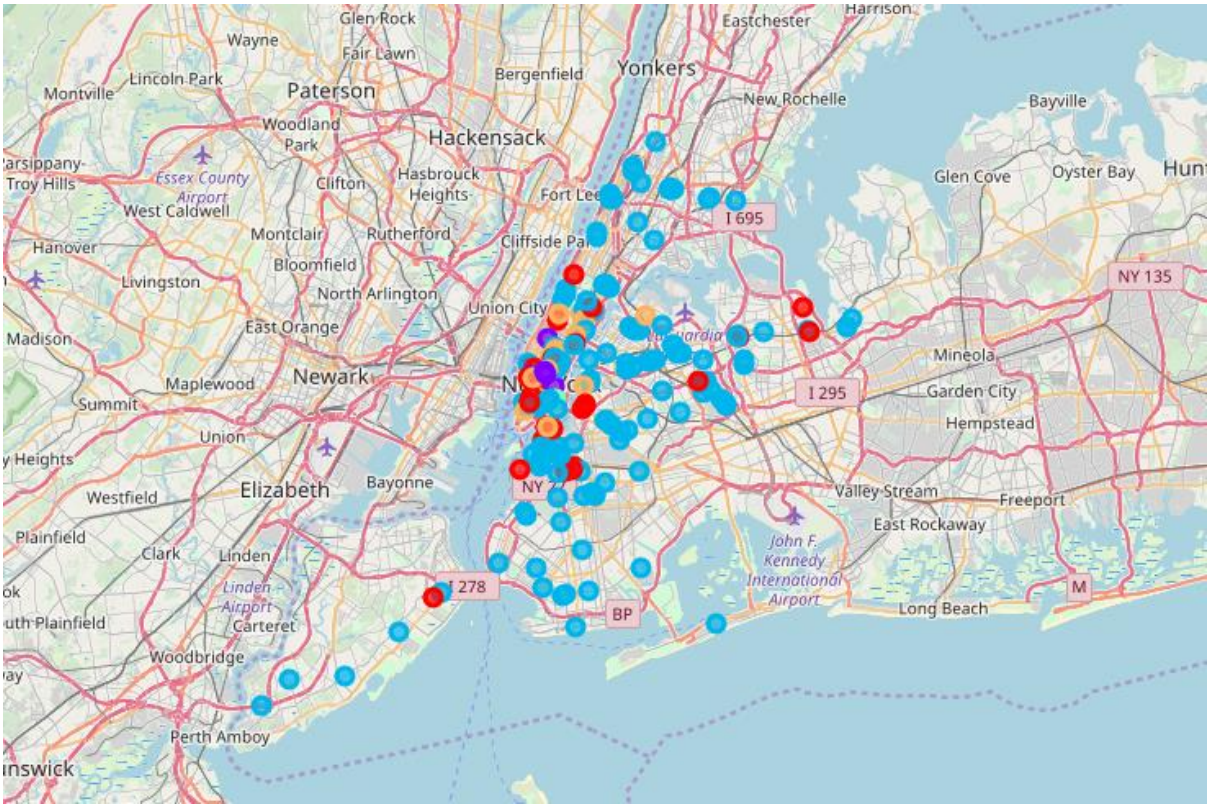
```
array([2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
```

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

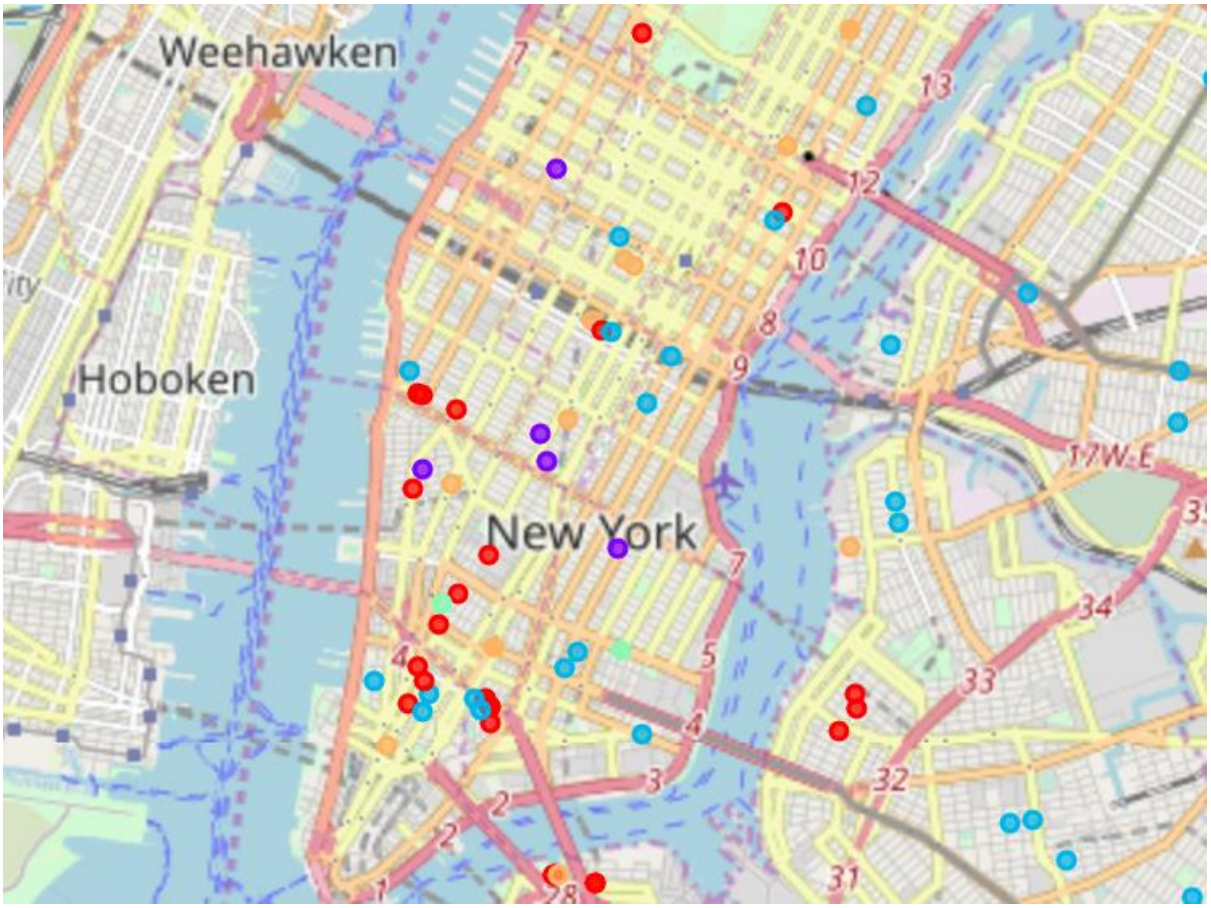
```
bakeries_stats_ny_non0.insert(0,'ClusterLabels', kmeans.labels_)

bakeries_stats_ny_non0.head()
```

	ClusterLabels	Borough	Neighborhood	Lat	Lng	ID	Name	Likes	Rating	Tips
0	2	Bronx	Kingsbridge	40.884793	-73.899861	4cab0bf1d971b1f7873327e1	S & S Cheesecake	12	7	2
1	2	Bronx	University Heights	40.858062	-73.912040	4e5f8bde45dd4656dd97be19	Au Bon Pain	2	6	1
2	2	Bronx	Morrisania	40.822081	-73.900749	4bc3300674a9a5931826d4f6	Amor Bakery	7	7	2
3	2	Bronx	Morris Park	40.849241	-73.853613	4bbf3ae8f353d13a5f397e10	Joseph Scaglione Bakery	11	7	6
4	2	Bronx	Morris Park	40.848311	-73.855898	4c348cea213c2d7ff79e385d	Morris Park Bake Shop	6	6	3

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

```
bakeries_stats_ny_non0.loc[bakeries_stats_ny_non0['ClusterLabels'] == 0, bakeries_stats_ny_non0.columns[[1,2] + list(range(5,
```

	Borough	Neighborhood	ID	Name	Likes	Rating	Tips
24	Brooklyn	Prospect Heights	548eeada498e4c3cf0acd057	Little Cupcake Bakeshop	143	8	27
26	Brooklyn	Prospect Heights	44f9a2c4f964a52066381fe3	Joyce Bakeshop	135	8	76
30	Brooklyn	Brooklyn Heights	4f86cb13e4b05dd564791c7d	Le Pain Quotidien	143	7	39
38	Brooklyn	Red Hook	44dc6750f964a520a6361fe3	Baked	291	9	135
39	Brooklyn	Park Slope	4b7d51f3f964a52013b82fe3	Cousin John's Cafe and Bakery	141	8	55
56	Brooklyn	North Side	57f2b5a5498eb5093e5fb8bb	Martha's Country Bakery	236	8	44
57	Brooklyn	North Side	5241e37c498e5ed64d425a21	Caprices by Sophie	136	9	55
58	Brooklyn	North Side	4197f180f964a520151e1fe3	Fabiane's Cafe & Pastry Shop	168	7	103
60	Manhattan	Chinatown	482c4142f964a520d14f1fe3	Fay Da Bakery	132	8	56
61	Manhattan	Chinatown	49d55b3bf964a5208d5c1fe3	Mei Li Wah	241	8	141
62	Manhattan	Chinatown	3fd66200f964a520bce61ee3	La Bella Ferrara	161	9	91
86	Manhattan	Lincoln Square	56ba9f99498ef6f5b55a33dc	Breads Bakery - Lincoln Center	159	9	38
93	Manhattan	Chelsea	4a284b61f964a52007951fe3	Fat Witch Bakery	144	8	67
94	Manhattan	Chelsea	4a33b48ff964a520379b1fe3	Amy's Bread	198	8	107

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

```
bakeries_stats_ny_non0.loc[bakeries_stats_ny_non0['ClusterLabels'] == 1, bakeries_stats_ny_non0.columns[[1,2] + list(range(5,
```

	Borough	Neighborhood	ID	Name	Likes	Rating	Tips
88	Manhattan	Clinton	4fab5090e4b0eeffb5c77119	Schmackary's	902	9	276
100	Manhattan	East Village	3fd66200f964a5208be41ee3	Veniero's Pasticceria & Caffe	897	9	301
113	Manhattan	West Village	3fd66200f964a5203be71ee3	Magnolia Bakery	1167	8	433
186	Manhattan	Flatiron	4079dc00f964a52070f21ee3	The City Bakery	1288	8	663
187	Manhattan	Flatiron	503fb6d4ebca66a84f029bd8	Breads Bakery	1195	9	305

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

```
bakeries_stats_ny_non0.loc[bakeries_stats_ny_non0['ClusterLabels'] == 2, bakeries_stats_ny_non0.columns[[1,2] + list(range(5,
```

	Borough	Neighborhood	ID	Name	Likes	Rating	Tips
0	Bronx	Kingsbridge	4cab0bf1d971b1f7873327e1	S & S Cheesecake	12	7	2
1	Bronx	University Heights	4e5f8bde45dd4656dd97be19	Au Bon Pain	2	6	1
2	Bronx	Morrisania	4bc3300674a9a5931826d4f6	Amor Bakery	7	7	2
3	Bronx	Morris Park	4bbf3ae8f353d13a5f397e10	Joseph Scaglione Bakery	11	7	6
4	Bronx	Morris Park	4c348cea213c2d7ff79e385d	Morris Park Bake Shop	6	6	3
5	Bronx	Belmont	4bb6496e46d4a59398fdc5c0	Madonia Bakery	69	8	29
6	Bronx	Belmont	4af1a473f964a520c5e121e3	Egidio Pastry Shop	39	8	15
7	Bronx	Belmont	4b366de6f964a5202d3525e3	Addeo & Sons Bakery	23	8	11

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

```
bakeries_stats_ny_non0.loc[bakeries_stats_ny_non0['ClusterLabels'] == 3, bakeries_stats_ny_non0.columns[[1,2] + list(range(5,
```

	Borough	Neighborhood	ID	Name	Likes	Rating	Tips
98	Manhattan	Greenwich Village	4eb13d68e5e8c0f5bd2fd983	Dominique Ansel Bakery	2372	9	714
101	Manhattan	Lower East Side	40a55d80f964a52020f31ee3	Clinton St. Baking Co. & Restaurant	1668	9	932

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_stats_ny_non0.loc[bakeries_stats_ny_non0['ClusterLabels'] == 4, bakeries_stats_ny_non0.columns[[1,2] + list(range(5,
```

	Borough	Neighborhood	ID	Name	Likes	Rating	Tips
17	Brooklyn	Greenpoint	4c4bbf66c668e21ec3a57afa	Ovenly	394	9	100
75	Manhattan	Upper East Side	4ae37940f964a520869521e3	Lady M Cake Boutique	393	9	125
78	Manhattan	Upper East Side	4fc95a1ee4b0f9aeca457705	Maison Kayser	476	9	153
79	Manhattan	Yorkville	4a06f82df964a52010731fe3	Two Little Red Hens	767	9	342
81	Manhattan	Lenox Hill	4fc95a1ee4b0f9aeca457705	Maison Kayser	476	9	153
87	Manhattan	Lincoln Square	4a271f0bf964a5205b911fe3	Magnolia Bakery	667	9	273
90	Manhattan	Midtown	527e819a11d281648dd93142	Maison Kayser	456	9	84
91	Manhattan	Midtown	51deeb27498e0bcc94b6538c	Lady M Cake Boutique	416	8	103

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!