A
REPORT
On
Battle of Neighborhoods
Of New York City

by
JAYANTH KUMAR

Introduction/Business Problem

Every city has different parts like Residential, Office area, Entertainment hubs etc. Grouping these into clusters we can see which regions dominate particular part. This will help new business enterprises to establish their business at right place. This will also be useful to people who are moving into the city. We will cluster neighborhoods of New York City based on venues data from Foursquare location data. The choice of New York city because it has lots of location data compared to cities in India where I live because it is used more there. But this model can be used for other cities by including their respective city Data.

Data

We will use the New York City Neighborhood data used in our lab. This is a JSON file which contains list of Neighborhoods of New York with their Borough name and approximate Latitude and Longitude.

Then we get venues in the neighborhood from above latitude and longitude up to 1000 mt radius and Limit of 100 venues(that is max given by foursquare). We get Venue Name , latitude and longitude and venue sub category

Then we get categories and subcategories that go into them. Categories are - Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport

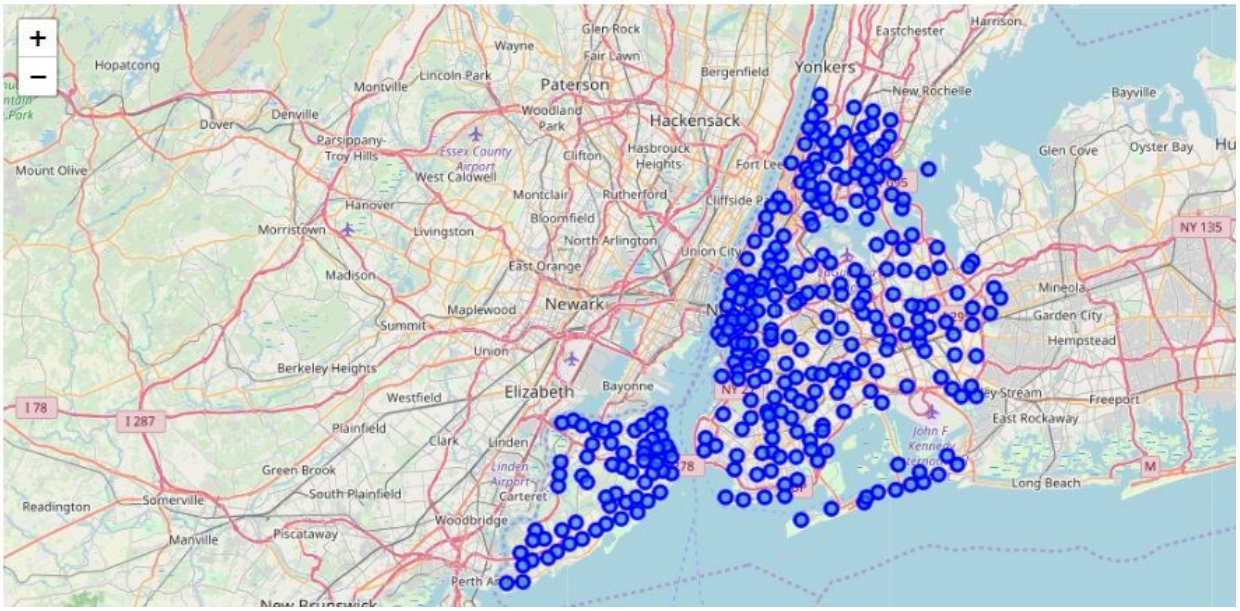
Methodology

Getting New York City Neighborhood Data

We get data of NYC Neighborhoods from here. It is a JSON file. We extract the Neighborhood, Borough, Longitude and Latitude information from the file. Then store it in pandas dataframe **ny_neighborhood.**

	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

View The Neighborhood points on the Map



Get Categories and Sub Categories From Foursquare

We will get the list of categories and Subcategories using the url

'https://api.foursquare.com/v2/venues/categories?&client_id={}&client_secret={}&v={}'

The response is parsed and data is extracted and stored in **sub_cat** dictionary such that dictionary keys are Categories and they contain list of subcategories.

There are some subcategories values are missing. So I add subcategories which appear more than 10 times into the dictionary and remaining are omitted.

Get Venues nearby to neighborhood location

Using Latitude and Longitude from **ny_neighborhood** dataframe we get list of venues near to that point. We use constraints radius=1000 mts and Limit = 200. But we get only 100 venues which is maximum no we get from Foursquare using the URL. https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}

This data is parsed to get Venue's Name, Latitude, Longitude and Subcategory. This data is stored in **ny_venues_df**.

PreProcess Data for Modelling

- Add Venue Category column to **ny_venues_df** using **sub_cat** dict
- Drop Columns whose **VenueCategory** column has empty String
- Now use get_dummies function of pandas to transform **VenueCategory** column into Respective columns of Categories whose value is either 1 or 0 and which returns new dataframe **ny_merged**
- Use Groupby function on Neighborhood and sum function for the values.

	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Allerton	4	0	34	2	2	0	0	18	2
1	Annadale	1	0	8	0	2	0	0	1	1
2	Arden Heights	0	0	10	0	5	0	0	8	2
3	Arlington	0	0	10	1	1	0	0	5	4
4	Arrochar	2	0	10	0	4	0	0	3	4
5	Arverne	0	0	12	1	3	0	0	4	5
6	Astoria	2	0	66	13	0	0	0	12	0
7	Astoria Heights	1	0	24	3	4	0	0	17	11
8	Auburndale	2	0	57	4	1	0	0	28	2
9	Bath Beach	0	0	56	1	1	1	0	26	3

Use KMeans to Cluster the Neighborhoods

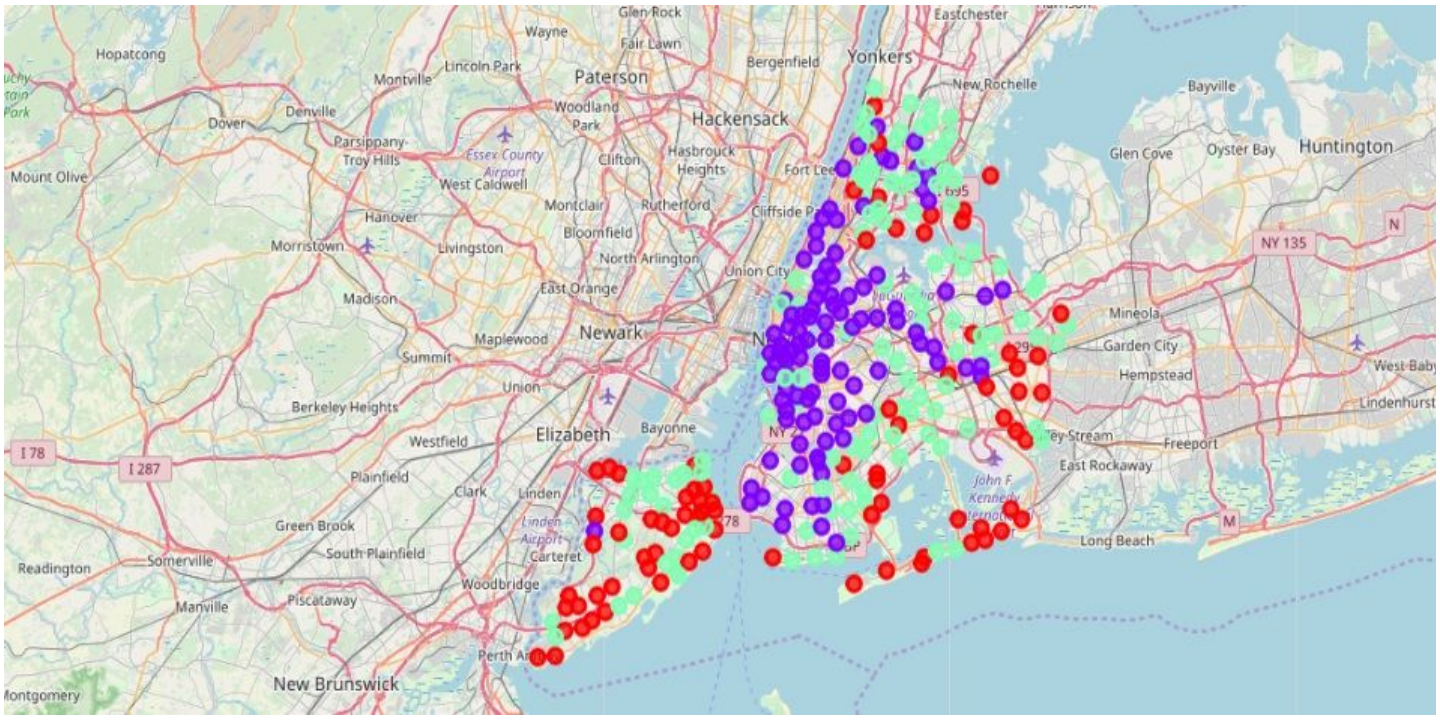
Use no of clusters as 3 in KMeans function of scikit-learn and fit it to the above PreProcessed Data to get Model.

Merge Data for Analysis

- Add cluster labels which we get from the model to the data frame **ny_merged**
- Add Latitude and Longitude Data to the **ny_merged** dataframe from **ny_neighborhood**

Results

Use ny_merged data to plot a map with latitude, longitude of neighborhoods and using different colors for different colors.



1. Red is cluster with label 0(cluster1)
2. Green is cluster with label 2(cluster3)
3. Violet is cluster with label 1(cluster2)

Lets see the tables of above clusters

- cluster1

```
print(cluster1.shape)
cluster1.head(10)
```

```
(79, 11)
```

]:

	index	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	1	Annadale	1	0	8	0	2	0	0	1	1
1	2	Arden Heights	0	0	10	0	5	0	0	8	2
2	3	Arlington	0	0	10	1	1	0	0	5	4
3	4	Arrochar	2	0	10	0	4	0	0	3	4
4	5	Arverne	0	0	12	1	3	0	0	4	5
5	15	Bayswater	0	0	3	0	2	0	0	2	0
6	20	Belle Harbor	0	0	8	1	14	1	0	4	0
7	24	Bergen Beach	0	0	8	0	2	0	0	1	0
8	26	Bloomfield	6	0	4	0	1	0	0	7	2
9	29	Breezy Point	0	0	1	0	3	0	0	0	0

- cluster3

```
print(cluster3.shape)
cluster3.head(10)
```

```
(120, 11)
```

:

	index	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	0	Allerton	4	0	34	2	2	0	0	18	2
1	7	Astoria Heights	1	0	24	3	4	0	0	17	11
2	10	Battery Park City	8	0	35	2	19	2	0	22	4
3	12	Bay Terrace	6	0	34	1	6	0	1	35	4
4	13	Baychester	4	0	40	0	4	0	0	45	2
5	16	Bedford Park	8	0	35	4	12	0	0	13	4
6	18	Beechhurst	2	0	26	2	2	0	0	14	2
7	19	Bellaire	1	0	28	0	6	0	0	12	6
8	21	Bellerose	1	0	20	2	0	0	0	17	4
9	25	Blissville	5	0	37	2	2	0	0	14	7

- cluster2

```
print(cluster2.shape)
cluster2.head(10)
```

```
(103, 11)
```

]:

	index	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	6	Astoria	2	0	66	13	0	0	0	12	0
1	8	Auburndale	2	0	57	4	1	0	0	28	2
2	9	Bath Beach	0	0	56	1	1	1	0	26	3
3	11	Bay Ridge	1	0	64	8	1	0	0	21	0
4	14	Bayside	2	0	62	9	2	0	0	21	1
5	17	Bedford Stuyvesant	3	0	58	11	7	0	0	18	0
6	22	Belmont	5	0	59	1	9	0	0	22	0
7	23	Bensonhurst	0	0	61	0	1	0	0	32	0
8	30	Briarwood	1	0	54	2	2	0	0	13	4
9	34	Bronxdale	3	0	61	2	4	0	0	25	1

Discussion

There are 2 observations that can be made from above results

1. We see that cluster color changes as we move outwards from Manhattan Borough from violet to green to Red. We also see No of places mainly Categories Food and Shop & Service decrease from cluster2 to cluster3 to cluster1, which gives us that outer parts(red, cluster1) are mostly Residential while the inside part(Violet, cluster2) are Work areas. While areas with Green(cluster3) color are mix of both.
2. We also observe that more no of Arts & Entertainment, Outdoor & Recreation, Travel & Transport are high in cluster3(Green) than in other 2 clusters

Conclusion

We used Data analysis on New York City to get neighborhood types by using venues list from Foursquare API. The results show fine boundary between different clusters and also no of venues from the tables show decreasing order as we go away from violet to green to red. But there was a limitation as we get max 100 values for a query which could have give better insights