NEW YORK CITY NEIGHBOURHOODS

July14, 2021

Choosing best place for starting new restaurant in New york city

1. Introduction

1.1 Background

One of the greatest things about New York City is the diversity of the people; And there's no better way to experience New York City's melting pot than by digging into its food. New York City is a gigantic city with thousands and thousands of restaurants. In recent days people are more inclining towards variety of cuisines when they are out. No one mostly wish to eat same routine food when they visit restaurants. So huge variety of traditional restaurants have been opening in new york city. This is what inspired me to take up this project .

1.2 problem

Given data of neighbourhoods of NYC and scrape near bt venues of each neighbourhood from FOURSQUARE api, then apply (k-means)cluster analysis and find optimum place and type of restaurant to open in NYC

1.3 interests

Clients who are preparing to open a new restaurant or raw material retailers(veggies for Indian or meat for turkey ,etc) and for people who are exploring different traditions of food this will help them to organise each restaurant and also the best one in the region of their own choice

2. Data acquisition and cleaning

source of the data is https://geo.nyu.edu/catalog/nyu_2451_34572
NYC has many neighbourhoods(nothing but major landmarks). the above source provide many such major neighbourhoods in NYC
here comes https://foursquare.com/ we use this for getting all venues within provided radius of all neighborhoods
- we got 305 neighbours
- venues are taken from foursquare API
- This data may include offices, universities, crowded areas like theaters and parks in NYC
- data will have latitudes and longitudes of such workplaces and we'll find optimum location in various clustered region

Country/City: NYC
Goal: Open a restaurant for workers in weekdays
So, I will cross data from working days, and localisations.
I will use the following API:
Foursquare API: to find restaurant/venues

Data from url would be this

Out[16]:

	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
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446
10	Bronx	Baychester	40.866858	-73.835798

Data from foursquare api would be looked like this

	index	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue
0	3	Wakefield	40.894705	-73.847201	Burger King	40.895540	-73.856460	Fast Food R
1	4	Wakefield	40.894705	-73.847201	Ali's Roti Shop	40.894036	-73.856935	Caribbean R
2	9	Wakefield	40.894705	-73.847201	Golden Krust Caribbean Restaurant	40.903773	-73.850051	Caribbean R
3	10	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant	40.898083	-73.850259	Caribbean R
4	11	Wakefield	40.894705	-73.847201	Paula's Soul Cafe	40.893328	-73.855142	Southern / S R

3. Exploratory Data Analysis

As we will be dealing with data of latitudes and longitudes we won't be doing much of scaling and standardising

4. Methodology

Now, we have the neighborhoods data of nyc (305 neighborhoods). We also have the most popular venues in each neighborhood obtained using Foursquare API. A total of 2749 venues have been obtained in the whole city and 97 unique categories.

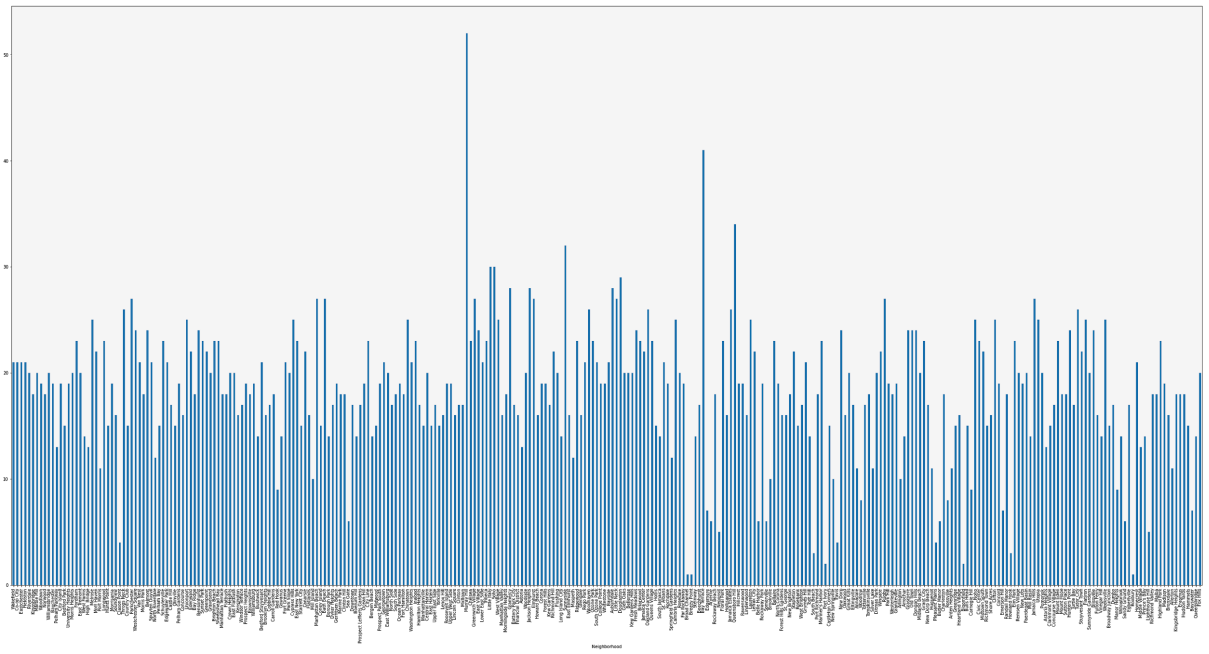
But as seen we have multiple neighborhoods with less than 10 venues returned. In order to create a good analysis let's consider only the neighborhoods with more than 10 venues. We can perform one hot encoding on the obtained data set and use it find the 10 most common venue category in each neighborhood.

Then clustering can be performed on the dataset. Here K - Nearest Neighbor clustering technique have been used. To find the optimal number of clusters silhouette score metric technique is used. The clusters obtained can be analyzed to find the major type of venue categories in each cluster. This data can be used to suggest business people, suitable locations based on the category.

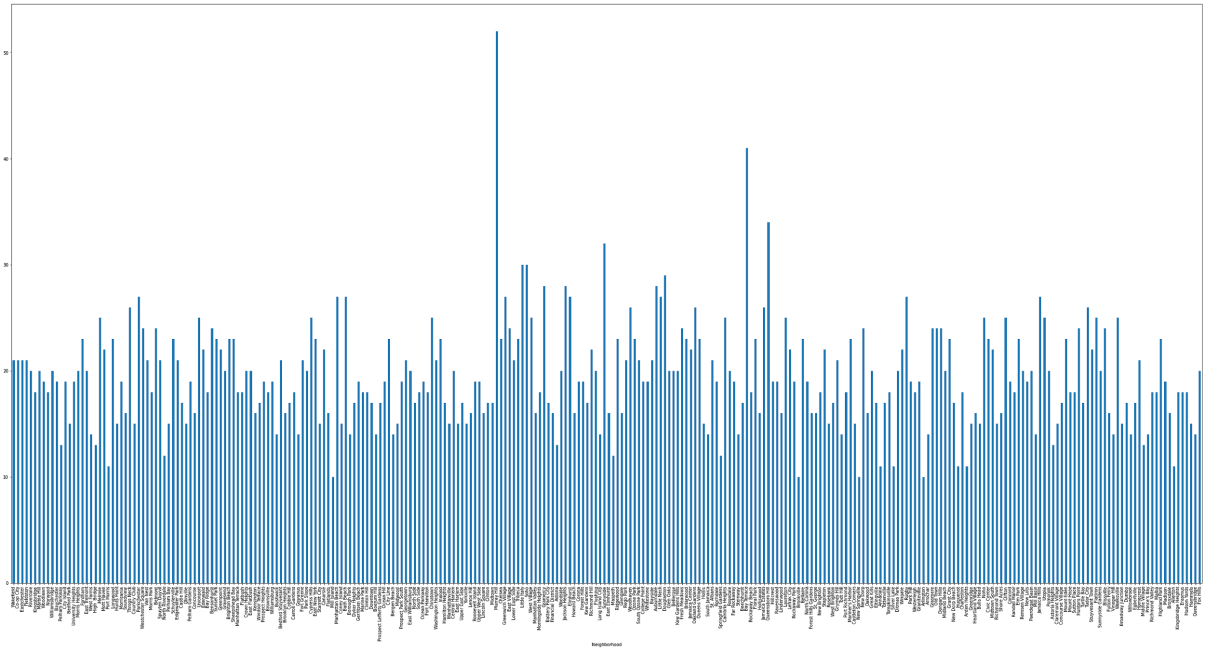
5. ANALYSIS

Looking into the dataset we found that there were many neighborhoods with less than 10 venues which can be remove before performing the analysis to obtain better results. The following plot shows only the neighborhoods from which 10 or more than 10 venues were obtained. The resultant dataset consists of 37 neighborhoods as shown in Fig 5.1.

Fig 5.1,Filtered Neighborhood Dataset Next, before removing less than neighbours



After removing less than 10 venues neighbours



we will perform one hot encoding on the filtered data to obtain the venue categories in each neighborhood. Then group the data by neighborhood and take the mean value of the frequency of occurrence of each category. A sample output is shown in Fig 4.2.

```
nyc_onehot.head()
```

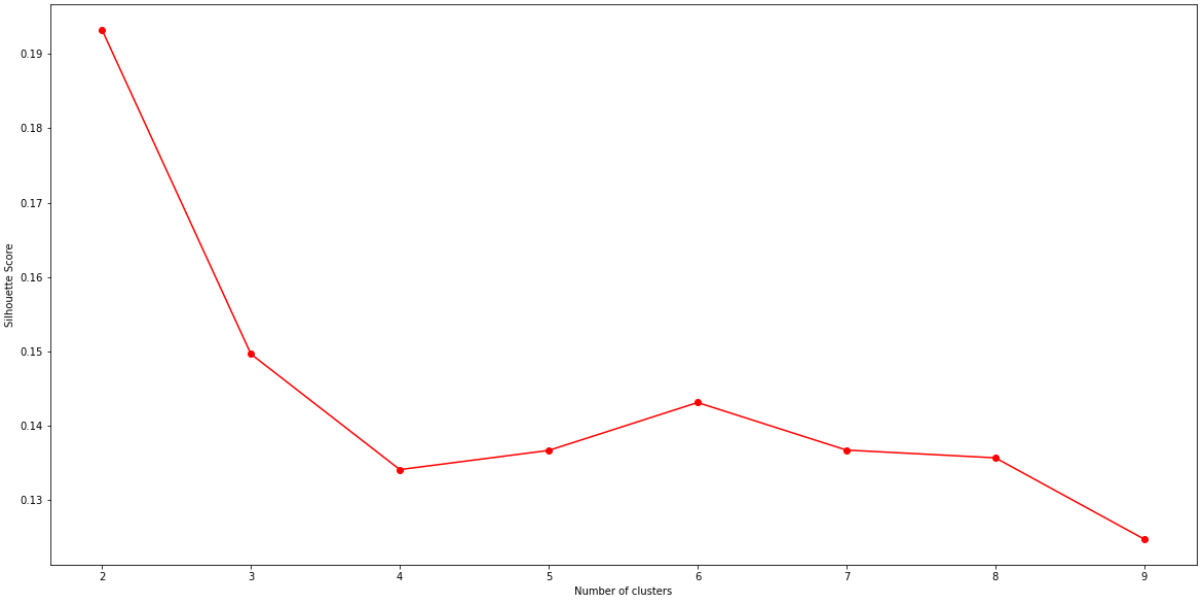
```
(5369, 97)
```

Out[60]:

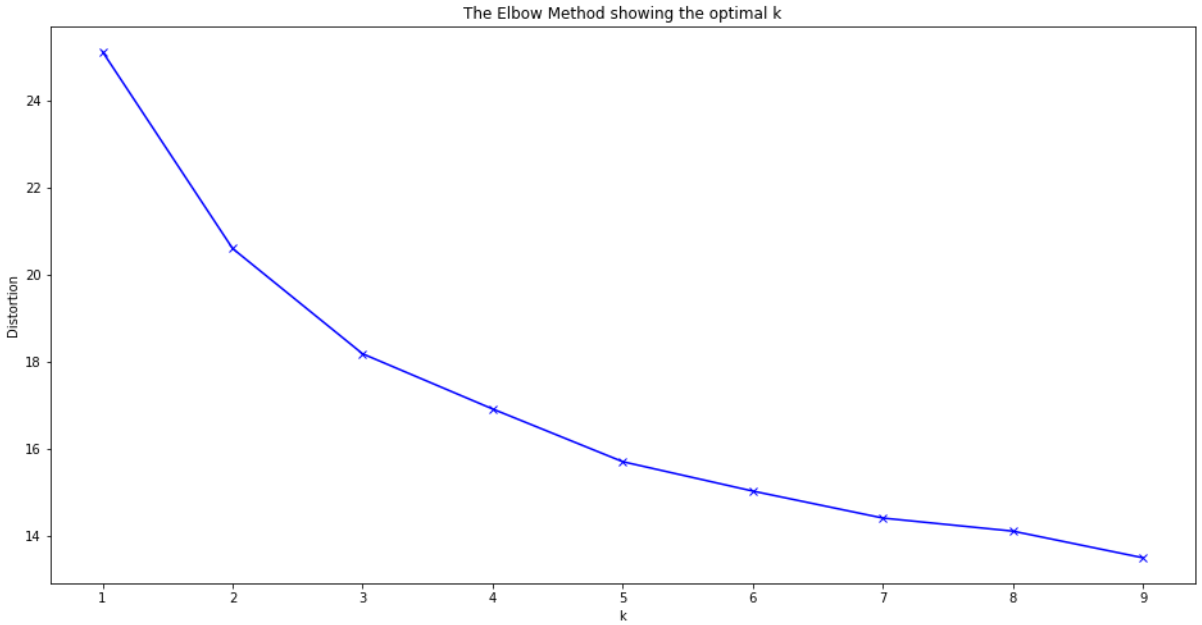
	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Armenian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	Bangladeshi Restaurant	Brazilian Restaurant	Bu Resta
0	Wakefield	0	0	0	0	0	0	0	0	0	0	0	
1	Wakefield	0	0	0	0	0	0	0	0	0	0	0	
2	Wakefield	0	0	0	0	0	0	0	0	0	0	0	
3	Wakefield	0	0	0	0	0	0	0	0	0	0	0	
4	Wakefield	0	0	0	0	0	0	0	0	0	0	0	

6. Model :

6.1. We will use K-means model for fitting , so ti find optimum k we did Silhouette Score for different Number of Clusters.



6.2. And elbow method as well

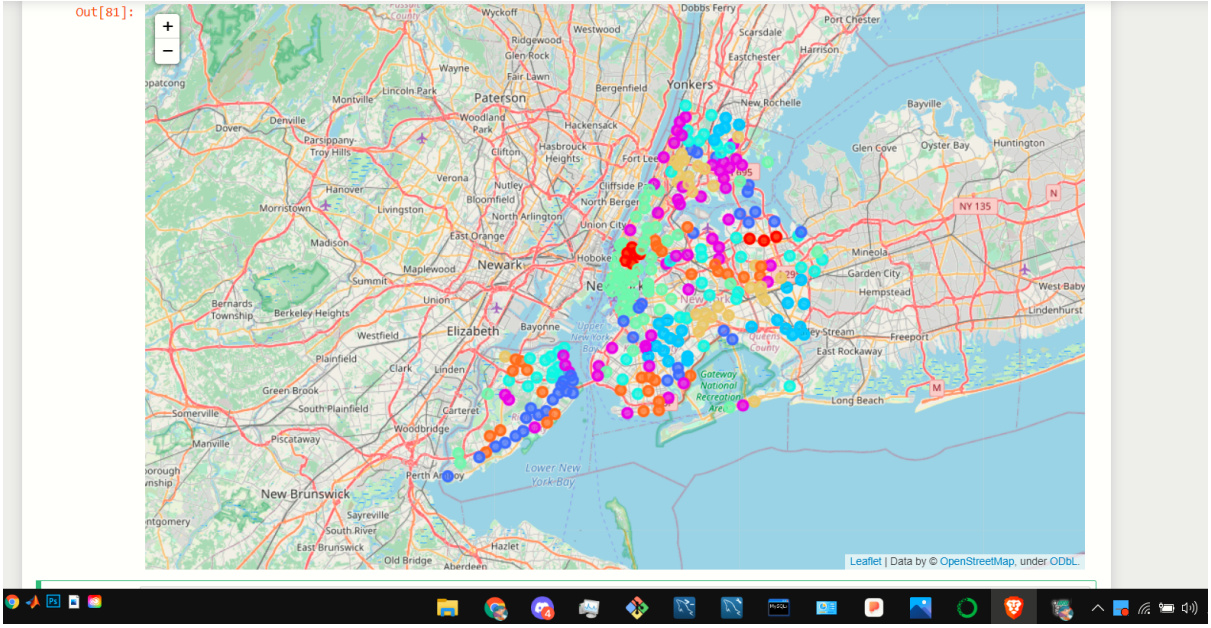


Both suggest k ==6 or 8, but our data is huge so I used 8 clusters

7. Results

The clusters after clustering are examined and plotted for better understanding

Visualisation of clusters of restraunt types in nyc



I will discuss just two clusters which has good number of restaurants and another one with less

7.1. cluster0

Korean Restaurant 9
Name: 1st Most Common Venue, dtype: int64

American Restaurant 4
Japanese Restaurant 2
Greek Restaurant 1
Mexican Restaurant 1
Italian Restaurant 1
Name: 2nd Most Common Venue, dtype: int64

Sushi Restaurant 2
Italian Restaurant 2
Ramen Restaurant 2
Cajun / Creole Restaurant 1
French Restaurant 1
Australian Restaurant 1

In cluster labeled 0 korean style food is adored most where australian style is not good choice

7.2. Cluster2 :I removed few restaurants below as we need only high and low

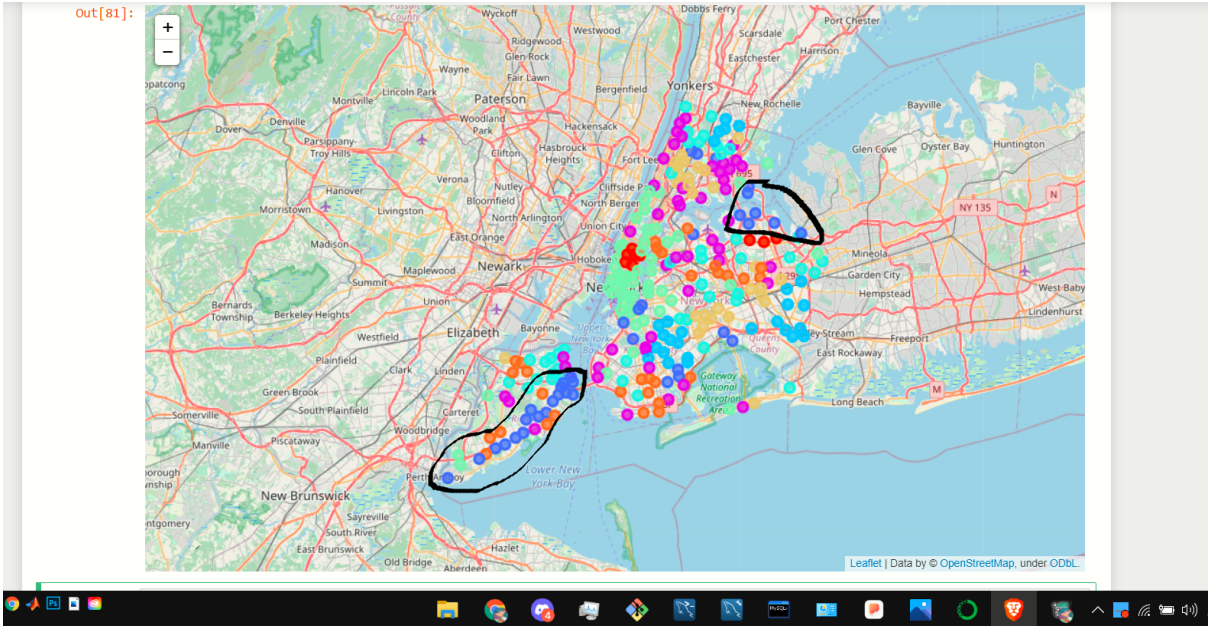
Italian Restaurant 34
Name: 1st Most Common Venue, dtype: int64

Mexican Restaurant 7
Fast Food Restaurant 6
Chinese Restaurant 6
Restaurant 4

Name: 2nd Most Common Venue, dtype: int64

Restaurant 5
Middle Eastern Restaurant 1
Puerto Rican Restaurant 1
Halal Restaurant 1
Israeli Restaurant 1
South American Restaurant 1
Afghan Restaurant 1
Name: 6th Most Common Venue, dtype: int64

So in cluster 2, region of blue dots, Italian restaurant works well but afghan, halal, such restaurants are not working good



CONCLUSION

Purpose of this project was to analyze the neighborhoods of nyc and create a clustering model to suggest personals places to start a new business based on the category. The neighborhoods data was obtained from an online source and the Foursquare API was used to find the major venues in each neighborhood. But we found that many neighborhoods had less than 10 venues returned. In order to build a good Data Science model, we filtered out these locations. The remaining locations were used to create a clustering model. The best number of clusters i.e. 8 was obtained using the silhouette score and I did elbow method to conform it and it worked well . Each cluster was examined to find the most venue categories present, that defines the characteristics for that particular cluster.

This project can conclude by providing me a bundle of knowledge based on real life project. The applications we can hereafter apply in this project will be helpful in any city,place.

The `stakeholder` takeaway of this project will be considering the number of most famous restaurants type in each cluster and proceed to open restaurant choosing profitable cluster