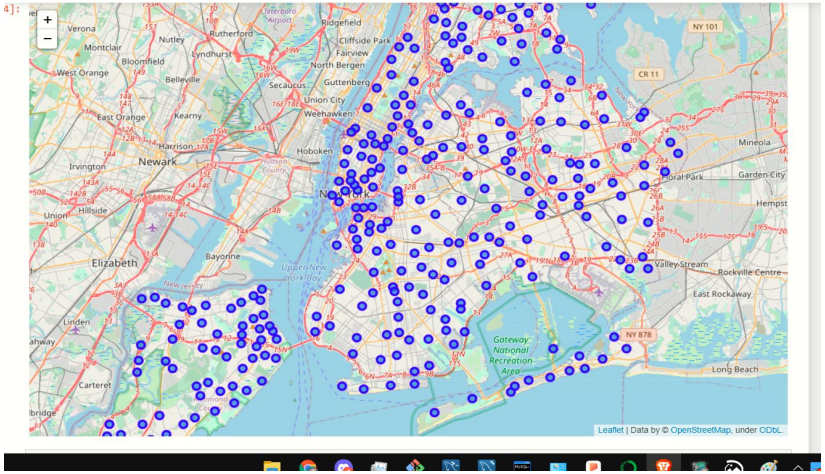
IBM -APPLIED DATA SCIENCE CAPSTONE PROJECT

-SAI CHARAN

Introduction

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

Neighbourhood locations map



1. Data acquisition and cleaning

source of the data is https://geo.nyu.edu/catalog/nyu_2451_34572
here comes https://foursquare.com/ we use this for getting all venues within provided radius of all neighborhoods
- we got 305 neighbours and 2749 unique venues
- venues are taken from foursquare API

```
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

Neighbourhood data

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

Venues data

Data processing

1, now the structure of data is , for every neighborhood several venues are there

2, take only food venues, from that only restaurant venues

3, do one hot encoding and group data by neighbourhoods

4, calculate average of all restaurants for respective neighbourhoods

5, then calculate top average restaurants that gives top most restaurants for each neighbourhood

One hot encoding results

```
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	

```
neighborhoods_venues_sorted.head()
```

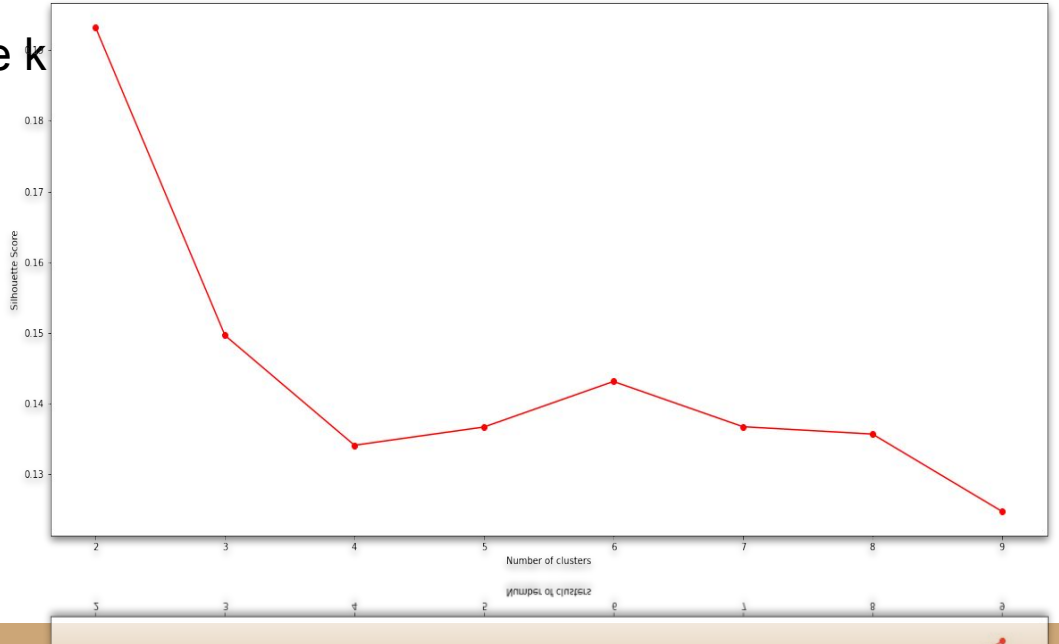
Out[67]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allerton	Chinese Restaurant	Mexican Restaurant	Fast Food Restaurant	Eastern European Restaurant	Afghan Restaurant	Polish Restaurant	Seafood Restaurant	Scandinavian Restaurant	Salvadoran Restaurant	Russian Restaurant
1	Annadale	Italian Restaurant	American Restaurant	Sushi Restaurant	Restaurant	Middle Eastern Restaurant	Chinese Restaurant	Afghan Restaurant	Polish Restaurant	Seafood Restaurant	Scandinavian Restaurant
2	Arden Heights	American Restaurant	Italian Restaurant	Sushi Restaurant	Restaurant	Chinese Restaurant	Mexican Restaurant	Afghan Restaurant	Polish Restaurant	Seafood Restaurant	Scandinavian Restaurant
3	Arlington	American Restaurant	Fast Food Restaurant	Peruvian Restaurant	Chinese Restaurant	Spanish Restaurant	Asian Restaurant	Caribbean Restaurant	Polish Restaurant	Shabu-Shabu Restaurant	Seafood Restaurant
4	Arrochar	Italian Restaurant	Restaurant	Chinese Restaurant	Polish Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Japanese Restaurant	Afghan Restaurant	Seafood Restaurant	Scandinavian Restaurant

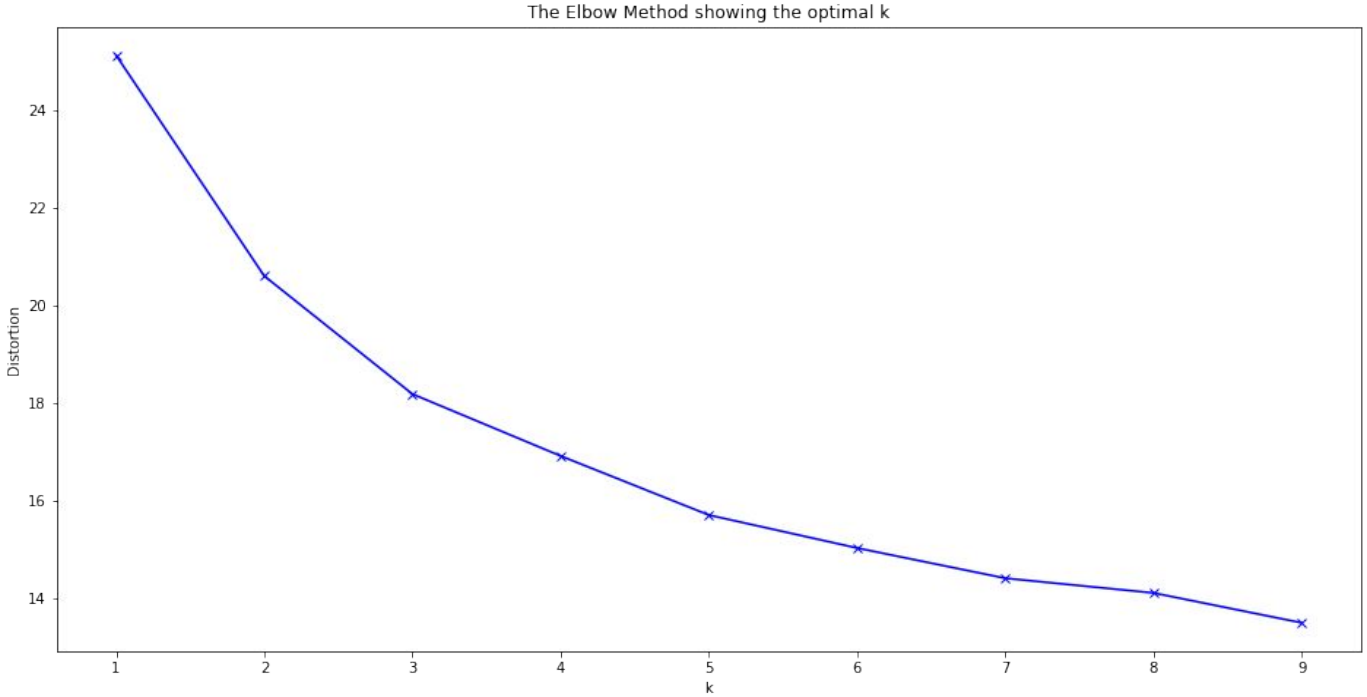
Model

- we will use k-means clustering algorithm
- k value ranging from 2 to 10 is taken to calculate Silhouette Score for different Number of Clusters.
- Highest score and respective k

  is considered



Elbow method also we can use for verification of above method

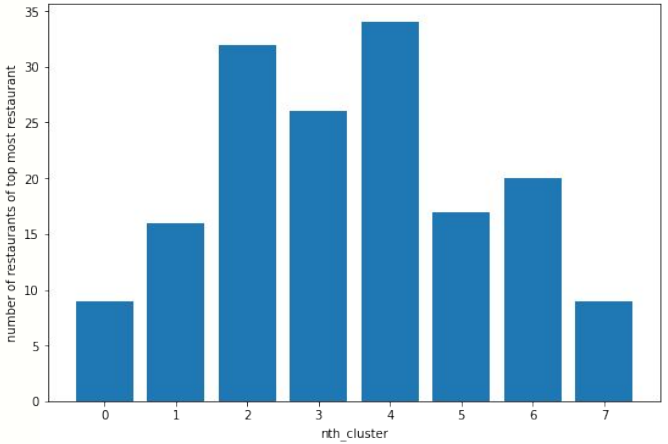


From above and this method k=4,6,8 looks optimum
Let’s take k=8

Results

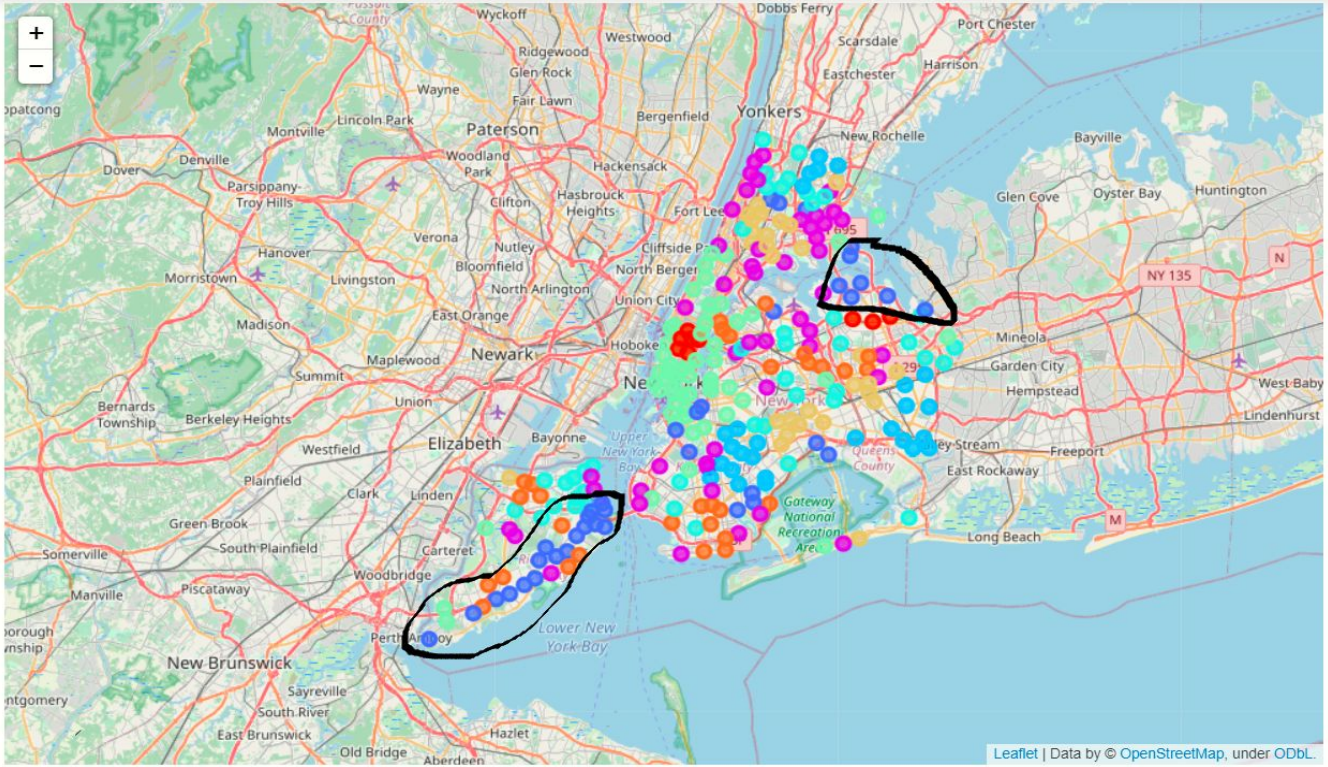
We have applied k-means algorithm for our data set and finally obtained these results example:
If client is willing to open a chinese restaurant then he is advised to choose cluster 4 region, still there are option where chinese is opted as second most but this is best case scenario

	cluster	top_most_restaurant	number_of_top_restaurants
0	0	Korean Restaurant	9
1	1	Fast Food Restaurant	16
2	2	Italian Restaurant	32
3	3	Caribbean Restaurant	26
4	4	Chinese Restaurant	34
5	5	Italian Restaurant	17
6	6	Fast Food Restaurant	20
7	7	Sushi Restaurant	9



Cluster marked is optimum, if client interested in opening italian restaurant

Out[81]:



Key points to discuss

- Let’s bring the key analysis of our algorithm
- so cluster2 -Italy and in cluster4- chinese traditional food have tremendous data, this could be because of the prosperity of the region or available of schools,offices where week-day work is more. we can see that after italian there are very less number of other restaurants
- let's also discuss why there are only few restaurants in cluster0 regions, either the foursquare api must have missed data or the service, place might not be comfortable or may be taste wasn't good
- so can we clearly say we can open a Italian restaurant or chinese restaurant in corresponding clusters

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 confirm 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

THANKYOU