Capstone Project - The Battle of Neighborhoods
Report

- Sushma Danturi

1. Introduction

1.1 Background

NYC is the most populous city in United States, often described as culture, financial, technology, education, tourism capital of Westrern world. Its liguistically most diverse city in the world with millions of foreign-born population of any city in the world.

New york City is a place for migrating people or individuals who arrives to this place with dream of starting there lives away from there comforting families. For any individual or a family to move to a new place will be a hard decision, and moving to a metropolitan like New York City will be even more intimidating.

This projects vision is to make it even more easier for individuals to take decision of which borough to choose to live in the city.

1.2 Business Problem

The objective of this Capstone Project is to analyse and choose the safest borough in the New York City based on the total crimes. This will help the people to buy/rent a home who are newly arriving to NYC. Exploring the neighbourhood and select the best among the five boroughs Brooklyn, Queens, Manhattan, The Bronx and Staten Island it has.

1.3 Who will use it?

The target audience for this problem will be all the individual or the families moving to this new place to make a decision of which location is safe and will be suitable for there preferences.

2. Data Description

2.1 Sources

The data for this project has been acquired from three different sources.
First one is for the crime data of NYC I have used the below link.

https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

The data is downloaded and the uploaded on Jyupter with the name NYPD_Crime_Data.csv. The data frame information is attained and it looks like the below Fig(a).

```
 #   Column               Non-Null Count    Dtype
---  ------               --------------    -----
 0   Crime_No             482337 non-null   int64
 1   Crime_DT             482337 non-null   object
 2   Crime_Reported_DT    482337 non-null   object
 3   Classification_Code  482337 non-null   int64
 4   Offence_Desc         482317 non-null   object
 5   Internal_Code        481968 non-null   float64
 6   Level                482337 non-null   object
 7   Borough              481961 non-null   object
 8   Latitude             475612 non-null   float64
 9   Longitude            475612 non-null   float64
10   Lat_Lon              475612 non-null   object
11   No_of_crimes         482337 non-null   int64
dtypes: float64(3), int64(3), object(6)
memory usage: 44.2+ MB
```

Fig: (a)

Second data is scrapped from a the link
https://www.citypopulation.de/en/usa/newyorkcity/
that's contains the data related to boroughs and its population in every 10 years. The following is the information of the data. Fig (b)

```
Data columns (total 7 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Name                         6 non-null      object
 1   Status                       6 non-null      object
 2   PopulationCensus1990-04-01   6 non-null      int64
 3   PopulationCensus2000-04-01   6 non-null      int64
 4   PopulationCensus2010-04-01   6 non-null      int64
 5   PopulationEstimate2019-07-01 6 non-null      int64
 6   Unnamed: 6                   5 non-null      object
dtypes: int64(4), object(3)
memory usage: 464.0+ bytes
```

Fig: (b)

Fig (C) is the representaion of the table I got after scrapping the link mentioned above.

	Borough	Status	Population-1990	Population-2000	Population-2010	Population-2019	Unnamed
0	Bronx	Borough	1203789	1332244	1384580	1418207	→
1	Brooklyn (Kings County)	Borough	2300664	2465689	2504721	2559903	→
2	Manhattan (New York County)	Borough	1487536	1538096	1586381	1628706	→
3	Queens	Borough	1951598	2229394	2230619	2253858	→
4	Staten Island (Richmond County)	Borough	378977	443762	468730	476143	→

Fig (c)

I have utilised only the Population 2019 as the crimes from the first source was only from the 2019 year. Refer Fig(d)

|	Borough	Status	Population-2019
0	Bronx	Borough	1418207
1	Brooklyn (Kings County)	Borough	2559903
2	Manhattan (New York County)	Borough	1628706
3	Queens	Borough	2253858
4	Staten Island (Richmond County)	Borough	476143

Fig(d)

After merging the NYPD crime data with the NYC census the data was Fig(e).

	Borough	Felony	Misdemeanor	Violation	Status	Population-2019	Total
1	BROOKLYN	46631	70504	21247	Borough	2559903	138382
2	MANHATTAN	38903	66785	15862	Borough	1628706	121550
0	BRONX	30356	57102	17367	Borough	1418207	104825
3	QUEENS	31369	49857	15975	Borough	2253858	97201
4	STATEN ISLAND	5059	10727	4217	Borough	476143	20003

Fig(e)

Third source to find the list of the Nighbourhood of Staten Island as it is the most safest place according to our data. Fig(f)
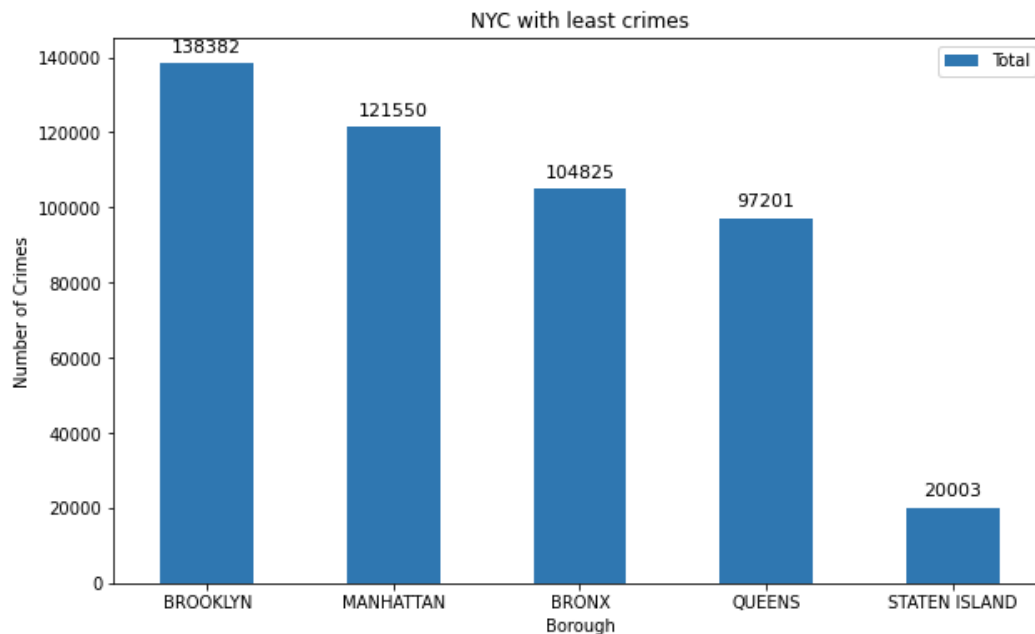
Fig(f)

The link to get the neighbourhood dataset is https://geo.nyu.edu/catalog/nyu_2451_34572

And I downloaded 'newyork_data.json'.
The Coordinates of the neighbourhood are acquired by Geopy library to get the latitude and longitude values of NYC.

The new dataset acquired about neighbourhood is explored and segmented using Foursquare API.

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 Statistical Summary

The Describe function is python is used to get statistics. Here I used it to obtaine the NYC crime data, which in return gives the statistics like count, mean, standard deviation, minimun, maximum, Quartiles like 25%, second 50%, third 75% . Refer Fig(g).

```
NYC_Crime_Table.describe()
```

	Felony	Misdemeanor	Violation	Population-2019	Total
count	5.000000	5.000000	5.000000	5.000000e+00	5.000000
mean	30463.600000	50995.000000	14933.600000	1.667363e+06	96392.200000
std	15643.156037	23927.115883	6375.194334	8.098120e+05	45560.768087
min	5059.000000	10727.000000	4217.000000	4.761430e+05	20003.000000
25%	30356.000000	49857.000000	15862.000000	1.418207e+06	97201.000000
50%	31369.000000	57102.000000	15975.000000	1.628706e+06	104825.000000
75%	38903.000000	66785.000000	17367.000000	2.253858e+06	121550.000000
max	46631.000000	70504.000000	21247.000000	2.559903e+06	138382.000000

Fig(g)

3.1.2 After merging the crime data with the NYC census data the table obtained looked like the Fig (h).

|	Borough	Felony	Misdemeanor	Violation	Status	Population-2019	Total
1	BROOKLYN	46631	70504	21247	Borough	2559903	138382
2	MANHATTAN	38903	66785	15862	Borough	1628706	121550
0	BRONX	30356	57102	17367	Borough	1418207	104825
3	QUEENS	31369	49857	15975	Borough	2253858	97201
4	STATEN ISLAND	5059	10727	4217	Borough	476143	20003

Fig(h)

3.1.3 Comparing the five boroughs of NYC it is observed that the Fig(i) shows the borough brooklyn and Manhattan are rated with highest crime.

```
NYC_Crime_Table.sort_values(['Total'], ascending = False, axis = 0, inplace = True )
NYC_Crime_Table
```

|	Borough	Felony	Misdemeanor	Violation	Status	Population-2019	Total
1	BROOKLYN	46631	70504	21247	Borough	2559903	138382
2	MANHATTAN	38903	66785	15862	Borough	1628706	121550
0	BRONX	30356	57102	17367	Borough	1418207	104825
3	QUEENS	31369	49857	15975	Borough	2253858	97201
4	STATEN ISLAND	5059	10727	4217	Borough	476143	20003

Fig(i)

3.1.4 Comparing the five boroughs of NYC for the lowest crime rated borough being the safest is so evidently would be Staten Island.
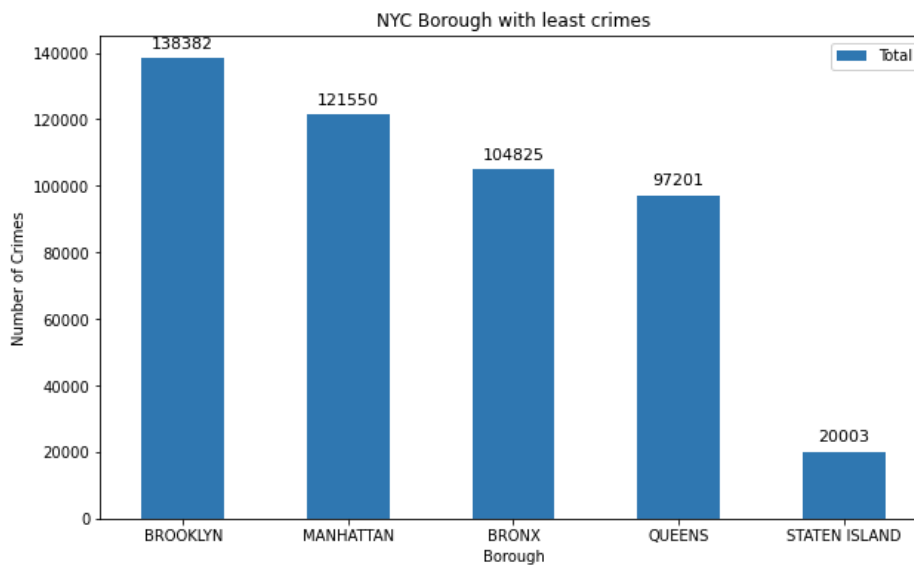
Fig(j)

3.1.5 Fig (k) shows the different crimes within the Staten Island borough based on various categories.

NYC Boroughs with the least Crimes
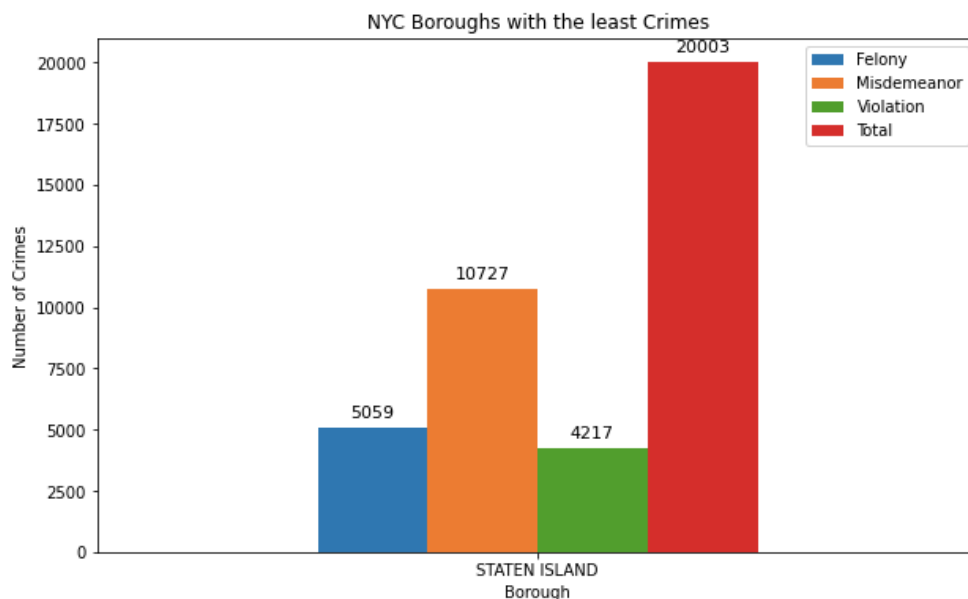
Fig(k)

3.2 Data Segmenting and Clustering

The data is downloaded from newyork_data.json and it is loaded. The loaded data is tranformed into pandas dataframe.

Once examining the resulting dataframe I made sure it shows 5 boroughs and 306 neighborhoods. Fig(l)

```
[10]:  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

```
[11]:  print('The dataframe has {} boroughs and {} neighborhoods.'.format(
           len(neighborhoods['Borough'].unique()),
           neighborhoods.shape[0]
       )
    )

The dataframe has 5 boroughs and 306 neighborhoods.
```

Fig (l)

3.2.1 Exploring the Neighborhood of Staten Island

The geographyical coordinates of Staten Island is found using the geolocator. And the below visualization Fig(m) shows the nieghborhood of the borough.
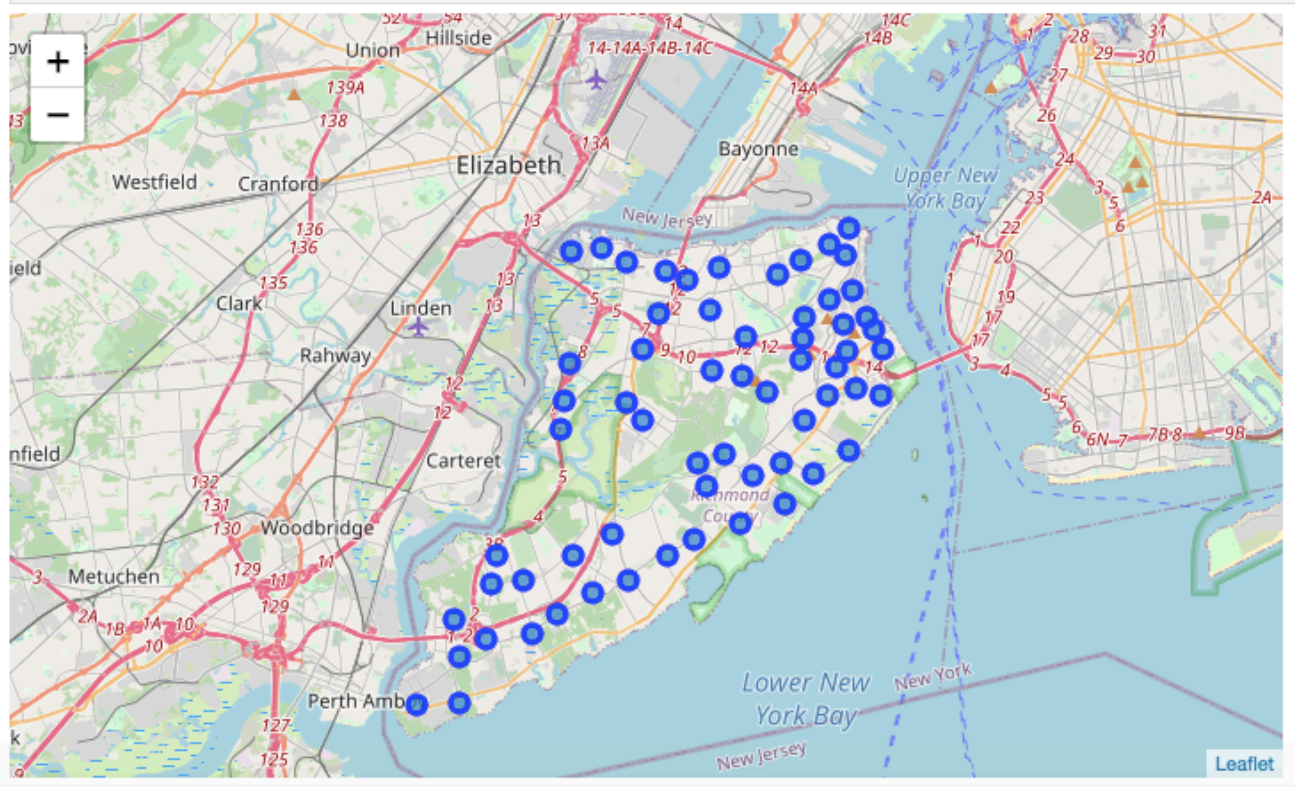
Fig(m)

3.2.2 Utilizing the Foursquare API for exploring
Here the radius is choosen to be 500 and the limit of venues to be returned is to be 100.

3.2.3 Analyzing each Neighborhood of Staten Island

One hot encoding is done on the venues data. The venues data is then grouped by Neighborhood and the mean of the venues are calculated, finally the common venues are calculates for each of the neighborhood.

	Neighborhood	Accessories Store	African Restaurant	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	...	Tourist Information Center	Toy / Game Store	Trail
0	Annadale	0.0	0.0	0.100000	0.000000	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0
1	Arden Heights	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0
2	Arlington	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0
3	Arrochar	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.043478	...	0.0	0.0	0.0
4	Bay Terrace	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
58	Travis	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0
59	West Brighton	0.0	0.0	0.026316	0.000000	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0
60	Westerleigh	0.0	0.0	0.000000	0.333333	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0
61	Willowbrook	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0
62	Woodrow	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0

63 rows × 181 columns

Fig(n)

3.2.4 Clustering of the Neighborhood

For finding the similar neighborhoods in the safest borough the clustering similar neighbors using K – means clustering the supervised machine learning algorithm to cluster the data based on the size of 5.

The reason to use the K-means clustering is to cluster neighborhoods with silmilar venues together so that people can choose the interesting areas based on there interests.

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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 V
0	Staten Island	St. George	40.644982	-74.079353	2	Clothing Store	Sporting Goods Shop	Italian Restaurant	Park	Bar	Pizza Place	Outlet Mall	Bas Sta
1	Staten Island	New Brighton	40.640615	-74.087017	0	Deli / Bodega	Bus Stop	Park	Discount Store	Playground	Flower Shop	Bowling Alley	Fi Resta
2	Staten Island	Stapleton	40.626928	-74.077902	2	Pizza Place	Discount Store	Sandwich Place	Bank	Restaurant	Spanish Restaurant	Fast Food Restaurant	Skate
3	Staten Island	Rosebank	40.615305	-74.069805	2	Pharmacy	Italian Restaurant	Grocery Store	Breakfast Spot	Beach	Pizza Place	Deli / Bodega	Cosm
4	Staten Island	West Brighton	40.631879	-74.107182	2	Coffee Shop	Pharmacy	Bank	Italian Restaurant	Music Store	Breakfast Spot	Bar	Bus

Fig(o)

4. Results

The vision of this project is to help indivuduals/families who want to relocate to the safest borough in New York City, anyone can explore the neighborhoods to which they want to relocate based on the most common venues in it.

5. Discussion

In the clusters formed after the data is explored a individual can look for a neighborhood with good public transportation, food places we can see that Clusters 2 has bus stops, restaurants as the most common venues.

Cluster- 2

	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
26	Graniteville	Boat or Ferry	Grocery Store	Filipino Restaurant	Gas Station	Furniture / Home Store	French Restaurant	Food Truck	Food & Drink Shop	Food	Flower Shop
53	Howland Hook	Boat or Ferry	German Restaurant	Gas Station	Furniture / Home Store	French Restaurant	Food Truck	Food & Drink Shop	Food	Flower Shop	Fish & Chips Shop

If a person is looking for a neighborhood with stores and restaurants in a close proximity then the neighborhoods in the cluster 3 is suitable for them. For a family looking for a neighborhood the Cluster 3 is more suitable as it shows parks, beach, grocery stores and Gyms.

Cluster- 3

	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 M Comm Ve
0	St. George	Clothing Store	Sporting Goods Shop	Italian Restaurant	Park	Bar	Pizza Place	Outlet Mall	Baseball Stadium	Bus Stop	Bus Stat
2	Stapleton	Pizza Place	Discount Store	Sandwich Place	Bank	Restaurant	Spanish Restaurant	Fast Food Restaurant	Skate Park	New American Restaurant	Optical S
3	Rosebank	Pharmacy	Italian Restaurant	Grocery Store	Breakfast Spot	Beach	Pizza Place	Deli / Bodega	Cosmetics Shop	Ice Cream Shop	Eas Europ Restau
4	West Brighton	Coffee Shop	Pharmacy	Bank	Italian Restaurant	Music Store	Breakfast Spot	Bar	Bus Stop	Café	Sandw Pl
7	South Beach	Pier	Beach	Athletics & Sports	Deli / Bodega	Diner	Discount Store	Furniture / Home Store	French Restaurant	Food Truck	Food & D SI
8	Port Richmond	Rental Car Location	Martial Arts Dojo	Donut Shop	Pizza Place	Dim Sum Restaurant	Fast Food Restaurant	French Restaurant	Food Truck	Food & Drink Shop	F
11	Castleton Corners	Bank	Pizza Place	Japanese Restaurant	Sandwich Place	Go Kart Track	Grocery Store	Mini Golf	Tattoo Parlor	Bagel Shop	
12	New Springville	Chinese Restaurant	Coffee Shop	Mobile Phone Shop	Bagel Shop	Accessories Store	Donut Shop	Sandwich Place	Restaurant	Pizza Place	Pharm

6. Conclusion

This project helps an individual to get a better exposure to the neighborhoods in terms of the crimes occuring in the borough and the most common venues in that neighborhood.

A project like this will be helpful to many people, it is always helpful to make use of technology and to understnad about the location online instead of being present there in the location itself or even before moving to the new place.

We have just taken safety as a primary concern of everyone and has shortlisted to the safest borough in New york city and then finding the most common places in the neighbourhoods and presenting the different clusters to choose from according to once preferences.

FYI:- *Due to some technical issues I have made two PDF copies of code as*
Battle_of_NeighborhoodCODE1
Battle_of_NeighborhoodCODE2

Please check both of them.