Battle of Neighbourhoods – New York
Naween Kumar
Feb 2021

Report submitted towards fulfilment of IBM Data Science Course on Coursera

1. Introduction

1.1 Background

New York is divided into 55 Community Districts (CDs). These CDs are subjected to various government stimuli for community development and are governed through a community board.

1.2 Problem statement

The purpose of this project is to establish a 'liveability index' to rank these CDs. The project will also cluster these CDs based on socio-economic parameters.

2. Data sources and cleaning

The data is mainly acquired from four sources:

2.1 **Venue data from Foursquare**: Data on restaurants, night life spots, schools, colleges, stores, medical facilities. To increase the number of results in neighborhood search, the search was be done with specific category IDs detailed at https://developer.foursquare.com/docs/build-with-foursquare/categories/

2.2 **Crime data**: Historical crime data by precinct is made available by NYPD on the website https://www1.nyc.gov/site/nypd/stats/crime-statistics/historical.page

2.3 **Socio-economic data**: Data on income, racial composition, education, poverty, commute times etc. was obtained from the various editions of the American Community Survey (ACS). Data is downloaded from the census website at https://data.census.gov/cedsci/

2.4 **Map data**: Geojson map data for community districts was be obtained from https://www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page

2.5 Since the data is available for different geographical aggregation levels for each type of data, a mapping was needed among these various geographical levels. Hence, precinct-> neighbourhood-> PUMA-> Community District relationships were constructed. Additionally, naming conventions of same area differ in the various datasets so an equivalence mapping was also constructed. In cases where an exhaustive listing could not be obtained, the website https://boundaries.beta.nyc/ which lists different administrative demarcations in map form, was used

2.6 Category codes used for Foursquare API were as follows:

#	Category	Code
1	Food	4d4b7105d754a06374d81259
2	College and Education	4d4b7105d754a06372d81259
3	Shop and service	4d4b7105d754a06378d81259
4	School	4bf58dd8d48988d13b941735
5	Nightlife spot	4d4b7105d754a06376d81259
6	Medical Centre	4bf58dd8d48988d104941735

3. Methodology

The following steps were executed:

3.1 Get venue data in various categories using Foursquare API
3.2 Since venues are gathered as nearby locations from NY neighbourhoods, the scraping via Foursquare yields a high number of duplicates.
3.3 The geojson map obtained is cleaned to bring it in line with the community districts for which census data is available (the others are parks, airports or empty tracts of land)
3.4 After removing duplicates from venues data, each venue is placed in a Community District based on geojson map data for the districts
3.5 Venue and crime data is standardized based on population of the community district. Hence, the districts are compared on a 'per capita' prevalence of facilities
3.6 Data is cleaned and streamlined to have coherence with census data granularity. This means that if two community districts are presented in a combined manner in census data, the venue data and crime data also have to be accordingly combined to have a like-to-like comparison.
3.7 A regression is run to check if prevalence of crime in a Community District can be explained by the socio-economic characteristics (racial mix, income distribution, education profile)
3.8 The Community Districts are clustered using K-Means clustering, based on similarity on the abovementioned socio-economic characteristics

3.9 Each Community District is ranked based on the per capita prevalence of venues (restaurants, stores, medical facilities, night life spots, schools, colleges). An additional input of commute time is added as a proxy for availability of suitable employment opportunities.

3.10 A composite rank is formulated using an unweighted sum of the ranks obtained in the previous step. This composite rank is the liveability index and community districts are ranked as per this liveability index

4. Results and discussion

4.1 Regression between crime prevalence and socio-economic characteristics:

Prevalence of crime is explained poorly by the various socio-economic characteristics with a R^2 of only 34%. This indicates that factors other than these characteristics are at play.

Secondary research[1] suggests that for a 13-year (2000-2012) state-level US data, the largest determinants of crime are police spending, inequality, % of population that is foreign-born and education. Of these, the present project included all but police spending. There might be two reasons behind the low R^2:

a. The causation might be less pronounced at a community district or neighbourhood level

b. Police spending is the biggest determinant in the secondary research. The absence of this data might have lowered the R^2.

4.2 Clustering of Community Districts

A k-means clustering algorithm was run on the data set of Community Districts with the following data fields:

a. Per capita incidence of total crime in the category of seven major felonies of murder, rape, robbery, assault, grand larceny, larceny of motor vehicle and burglary

b. Per capita prevalence of restaurants and night life spots

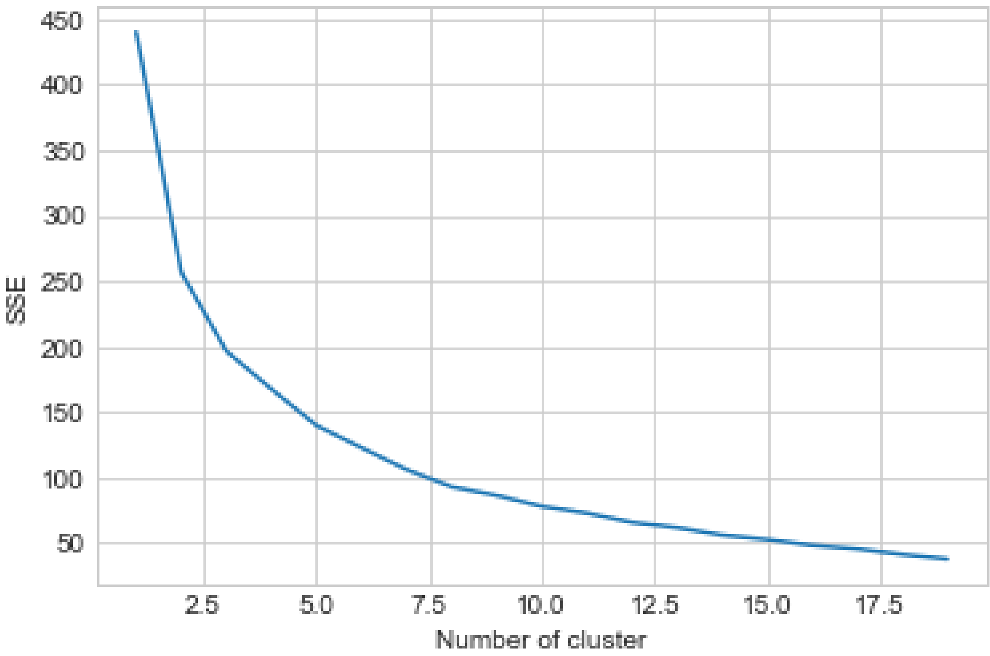
c. Per capita prevalence of schools and colleges

d. Per capita prevalence of medical facilities

e. Per capita prevalence of shops and stores

To find the most optimum k, the clustering was run for a range of values for k and the following elbow plot was generated:

[1] The Economic Determinants of Crime (Giovanni Cerulli, Maria Ventura, and Christopher F Baum) | Retrieved from http://fmwww.bc.edu/EC-P/wp948.pdf

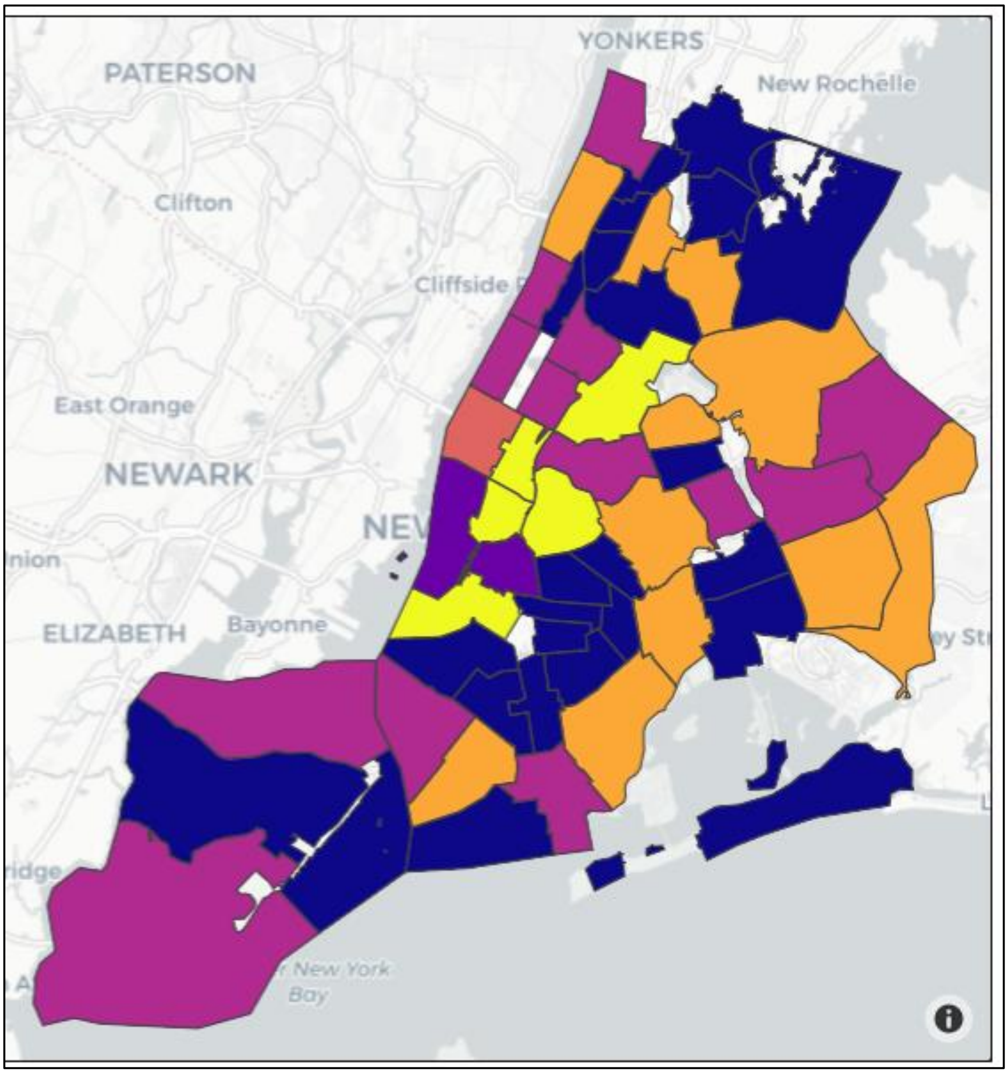


According to the above, an optimum value of k=6 was chosen and the clusters were labelled as follows:

Cluster	List of Community Districts in cluster
1	BRON01_02, BRON04, BRON05, BRON07, BRON10, BRON11, BRON12, BROO03, BROO4, BROO7, BROO08, BROO09, BROO12, BROO13, BROO14, BROO16, BROO17 MANH10, QUEE04, QUEE09, QUEE10, QUEE14, STAT02
2	BROO02, MANH01_02
3	BRON08, BROO10, BROO15, MANH07, MANH08, MANH09, MANH11, QUEE02, QUEE06, QUEE08, QUEE11, STAT01, STAT03
4	MANH04_05
5	BRON03_O6, BRON09, BROO05, BROO11, BROO18, MANH12, QUEE03, QUEE05, QUEE07, QUEE12, QUEE13
6	BROO01, BROO06, MANH03, MANH06, QUEE01

To check that the clusters are different from each other, the intra-cluster mean of various values was calculated:

Difference in means of various per capita parameters across clusters							
Clusters	**Medical**	**Restaurants**	**Stores**	**Colleges**	**Schools**	**Nightlife**	**Crime**
1	3.64	12.10	13.40	1.75	3.38	1.81	114.18
2	11.47	38.96	58.06	14.10	9.30	16.03	168.01
3	7.72	19.35	23.50	5.86	5.34	3.12	81.78
4	20.14	48.24	72.51	18.41	15.00	26.48	366.31
5	3.87	12.49	14.04	1.97	3.13	1.65	104.95
6	8.76	29.46	36.56	3.88	6.27	13.19	149.70



Depiction of clusters on map

4.3 Ranking of Community Districts on liveability index

To calculate the liveability index, each Community District is ranked on each parameter. The unweighted sum of ranks is taken as the liveability index. The results are as follows (same ranked districts are given 0.5 values):

Final ranking of various community districts on the constructed Liveability Index					
CD	**Final Rank**	**CD**	**Final Rank**	**CD**	**Final Rank**
BROO16	1	BRON01_02	23	STAT03	46
BROO05	2	BROO14	24	BROO01	47
BRON09	3	MANH10	25	MANH06	48
BRON05	4	BROO08	26	BROO06	49
BRON07	5	BROO03	27	MANH08	50
QUEE12	6	MANH12	28	STAT01	51
BRON04	7	QUEE05	29.5	BROO02	52
BROO18	8	QUEE13	29.5	MANH01_02	53.5
BROO17	9.5	BROO07	31	MANH04_05	53.5
QUEE10	9.5	QUEE08	32		
BROO13	11	MANH11	33		
BROO11	12	BRON10	34		
BRON03_06	13	BROO15	35		
QUEE14	14	QUEE07	36		
QUEE09	15	MANH09	37		
QUEE04	16	QUEE06	38		
QUEE03	17	BRON08	39.5		
BRON12	18	QUEE02	39.5		
BROO09	19	BROO10	41		
BRON11	20	QUEE01	42		
BROO12	21	MANH07	43		
BROO04	22	QUEE11	44.5		
		MANH03	44.5		

5. Conclusion and discussion

- The analysis achieved two objectives
  - Clustering New York Community Districts based on socio-economic characteristics
  - Ranking of community districts based on livability characteristics
- For a person newly moving into New York, the ranking on livability can be used. However, for a person moving within the city, he/she would like to consider the next location based on similarity on socio-economic characteristics.
- The analysis can be further improved by adding weights to the different characteristics based on personal preferences.