Battle of neighborhoods - New York

Introduction

- The project aims to compare the 55 community districts in New York on their liveability.
- Liveability is determined by the prevalence of facilities in the area:
  - Restaurants and night life (‘social life parameters’)
  - Proximity of schools, colleges and stores (‘basic needs’)
  - Proximity of medical facilities and low incidence of crime (‘hygiene)
  - Commute times as a proxy for availability of employment opportunities
- The project will also cluster the community districts on socio-economic parameters:
  - Education profile (less than high school | high school | some college | grad or higher)
  - Racial profile (% of various races in the population)
  - Income distribution (% of population in various income brackets)

Data sources

Data was sourced from 4 major sources:

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. **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 .

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/

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

Methodology (1 of 2)

The following steps were followed:

1. Get venue data in various categories using Foursquare API

2. Since venues are gathered as nearby locations from NY neighbourhoods, the scraping via Foursquare yields a high number of duplicates.

3. After removing duplicates, venues are placed in Community District based on geojson map data for the districts

4. Restaurants and crime data is standardized based on population of the community district. Hence, the districts are compared on a ‘per capita’ prevalence of facilities

5. Data is cleaned and streamlined to have coherence with census data granularity

Methodology (2 of 2)

6. 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)

7. The Community Districts are clustered using K-Means clustering, based on similarity on the abovementioned socio-economic characteristics

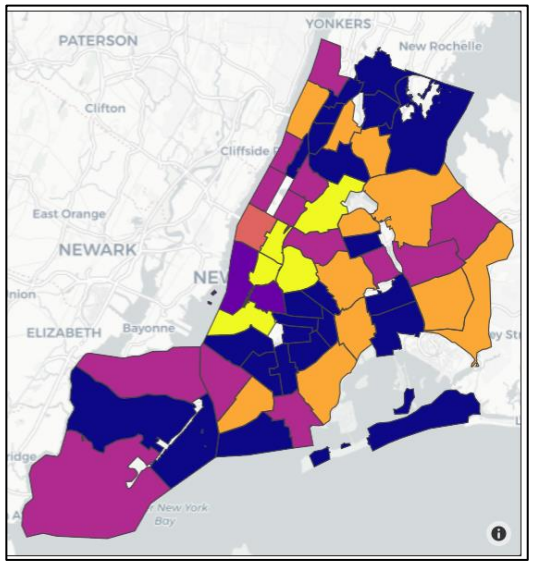
8. Each Community District is ranked based on the per capita prevalence of venues (restaurants, stores, medical facilities, night life spots, schools, colleges). An additional input of commute time is added as a proxy for availability of suitable employment opportunities.

9. 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

Results (1 of 2)

1. Prevalence of crime is explained poorly by the various socio-economic characteristics with a R2 of only 34%. This indicates that factors other that these characteristics are at play.

2. The 55 Community Districts can be clustered into 6 clusters based on similarity of socio-economic characteristics



Cluster	
1	BRON01_02, BRON04, BRON05, BRON07, BRON10, BRON11, BRON12, BRO003, BRO04, BRO07, BRO008, BRO009, BRO012, BRO013, BRO014, BRO016, BRO017 MANH10, QUEE04, QUEE09, QUEE10, QUEE14, STAT02
2	BRO002, MANH01_02
3	BRON08, BRO010, BRO015, MANH07, MANH08, MANH09, MANH11, QUEE02, QUEE06, QUEE08, QUEE11, STAT01, STAT03
4	MANH04_05
5	BRON03_06, BRON09, BRO005, BRO011, BRO018, MANH12, QUEE03, QUEE05, QUEE07, QUEE12, QUEE13
6	BRO001, BRO006, MANH03, MANH06, QUEE01

Results (2 of 3)

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

Results (3 of 3)

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
BROOO5	2	BROO14	24	BROOO1	47
BRON09	3	MANH10	25	MANH06	48
BRON05	4	BROOO8	26	BROOO6	49
BRON07	5	BROOO3	27	MANH08	50
QUEE12	6	MANH12	28	STAT01	51
BRON04	7	QUEE05	29.5	BROOO2	52
BROO18	8	QUEE13	29.5	MANH01_02	53.5
BROO17	9.5	BROOO7	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		
BROOO9	19	BROO10	41		
BRON11	20	QUEE01	42		
BROO12	21	MANH07	43		
BROOO4	22	QUEE11	44.5		
		MANH03	44.5		

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.