Final Project of Applied Data Science Capstone by IBM

The Battle of Neighborhoods

Sercan Cakir

26.12.2019

Content

1. **Introduction/Business Problem**
   - Problem Background & Description
   - Data

2. **Methodology**

3. **Results/Discussion**

1. Introduction/Business Problem

- Problem Background & Description

- The population growth has always a demand for new housing, improving the facilities, enhancing the number of places of entertainment, restaurants, and especially shopping malls. Nowadays, Most of the small families opt for entertaining in shopping malls during weekend where they can find vast amount of opportunities for themselves as parents and for their children. Therefore, New York city has its potential for need-based built shopping malls. With this study, we will conduct a research on finding useful insights that would be necessary for determining the location for a new shopping mall.
- This study aims at helping a group of stakeholders find the right location to build a new shopping mall in New York city. Regardless of any region of the city (e.g. north, west, etc.), each neighborhood is of interest to the stakeholders. Therefore, all of them will be clearly examined and the one that essentially needs a shopping mall will be found out.

1. Introduction/Business Problem

- Data

- In this study, we will need three types of datasets:

  1) New York city (NYC) data including geographic coordinates of neighborhoods

  2) NYC population data: since the population growth is one of the key parameters, we will need borough based population.

  3) In addition to the above data, Foursquare API will be used to access venues in each neighborhood.

1. Introduction/Business Problem

– Data

1.)

- First type of data will be taken from our previous lab (Segmenting and Clustering Neighborhoods in New York City) which was taken originally from NYU Spatial Data Repository via the link: https://geo.nyu.edu/catalog/nyu_2451_34572
- In this data, we will access to the boroughs, neighborhoods and their geographical coordinates (latitudes and longitudes).

2.)

- Second type of data will be scraped from Wikipedia via the link: https://en.wikipedia.org/wiki/Demographics_of_New_York_City
- With this data, we will have access to the population of each borough (from 2017) and also have the opportunity to see the population growth from previous years which will help us make assumptions.
- First and second dataset will be merged on names of boroughs.

3.)

- With our client ID and secret, we will have opportunity to use Foursquare API where we will see the venues (restaurants, shopping malls, etc.) within each borough and neighborhood.

2. Methodology

This study as it is already mentioned aims at finding out the most suitable borough for building a new shopping mall in the city of New York.

To be able to make this aim realise, we need to determine some criteria that are the essentials for building a new shopping mall:

- Current population, historical population growth and future population of each borough
- Persons per square kilometers
- Gross domestic product per capita
- The borough size
- Existing venues in each borough such as shopping malls, restaurants, cafés, etc.
- The number of neighborhoods included by each borough

2. Methodology

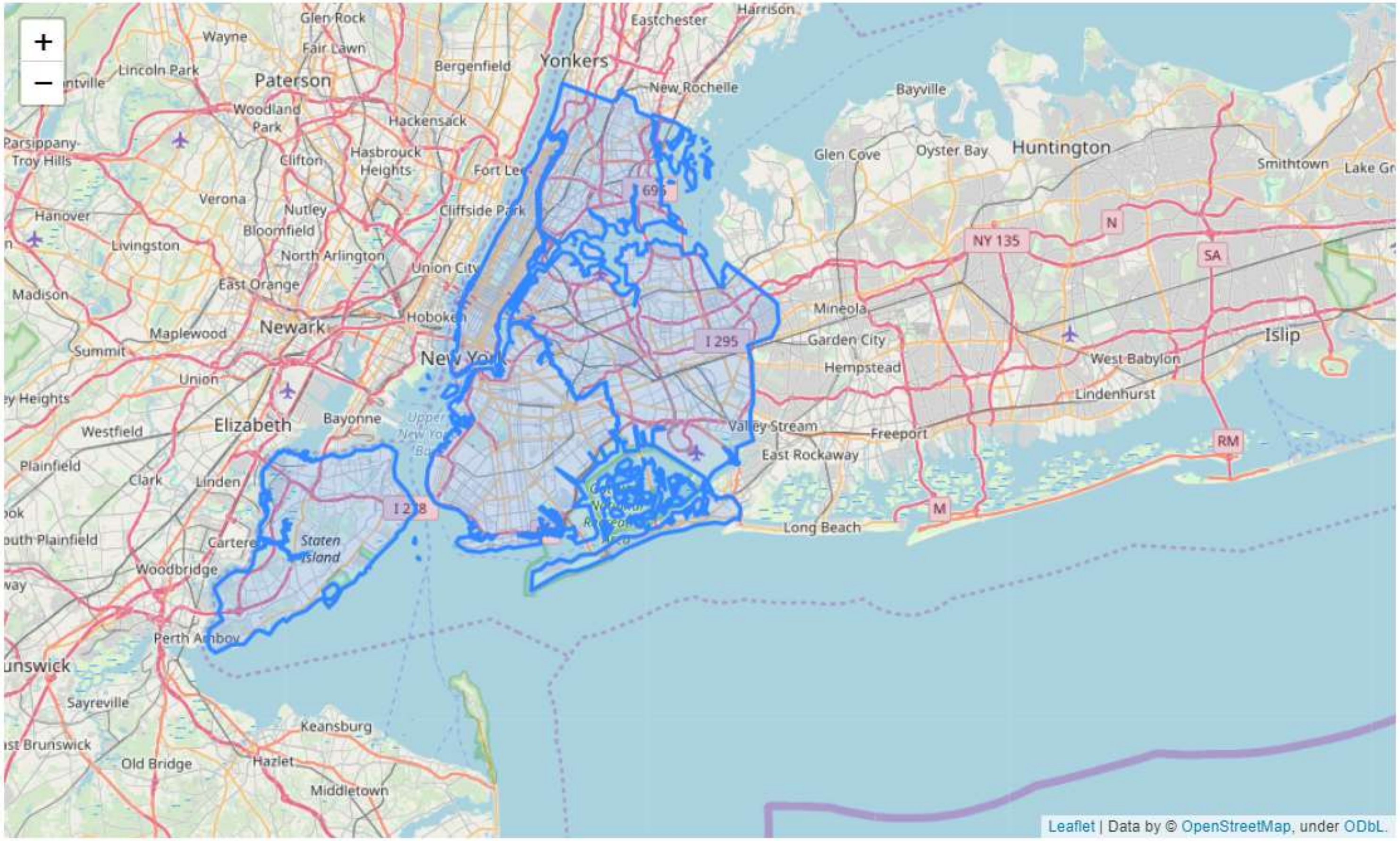
All criteria mentioned above should have an importance score from 0 to 10 that then influences the final score of each borough. The higher the score the better area to build up a shopping mall. With this methodology, we will be helping stakeholders to make the best decision.

The methods that we will follow up during this study can be seen as follows:

1. We will access the borough data including the geographical coordinates of each polygon's borders, and the neighborhoods also including their coordinates
2. The population data and the growth over 150 years will be scraped from Wikipedia and the data will be cleansed based on any need.
3. Based on the population growth data, a regression model will be developed for each borough in order to predict potential populations of each borough for the years 2020 and 2030 for the future plan.
4. Foursquare API will be used for accessing the venues in each borough.
5. Use all the above mentioned methods to implement a decision making methodology for calculating a final score for each borough.

Downloading NYC boroughs with their geographical coordinates

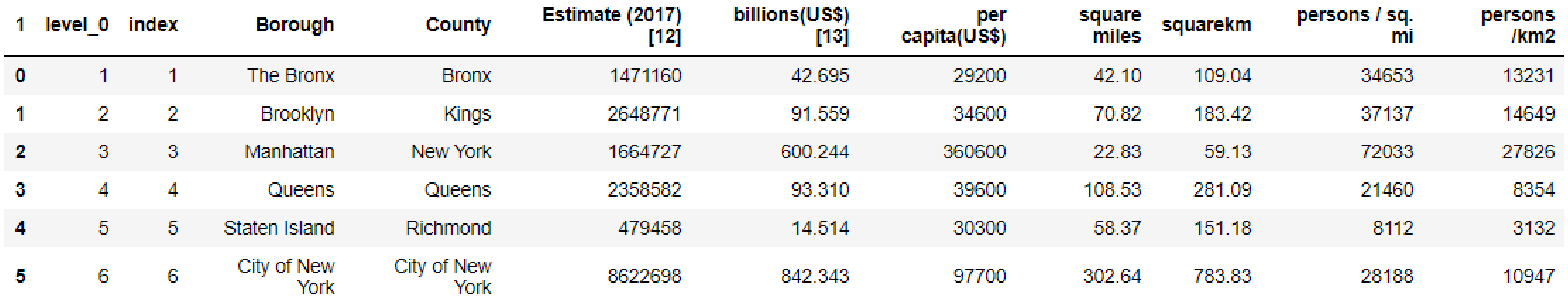
- The data is downloaded as geojson from https://geo.nyu.edu/download/file/nyu-2451-34154-geojson.json and visualised as follows:



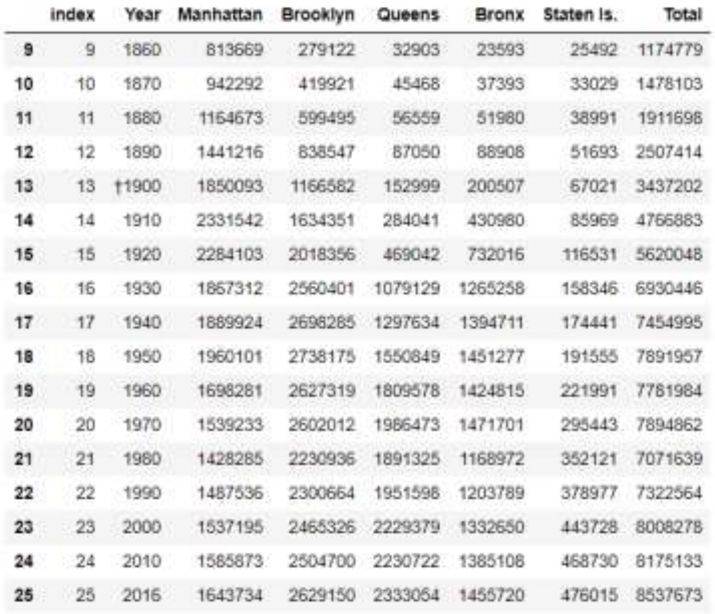
Leaflet | Data by © OpenStreetMap, under ODbL.

Scraping Wikipedia NYC census page

- To access the population of each borough the census table is scraped.
- To analyze the population growth in each borough, the historical census of NYC is also scraped

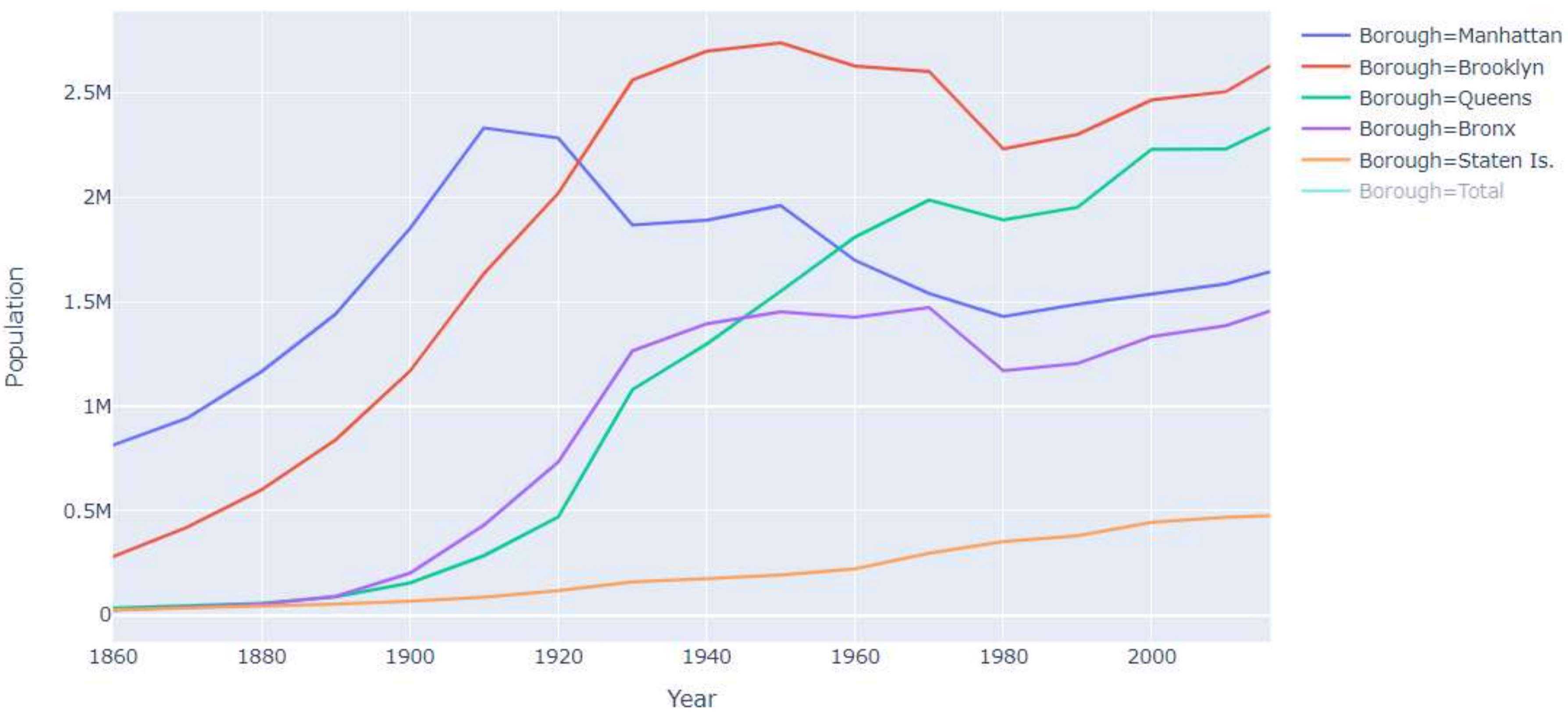


1	level_0	index	Borough	County	Estimate (2017) [12]	billions(US$) [13]	per capita(US$)	square miles	squarekm	persons / sq. mi	persons /km2
0	1	1	The Bronx	Bronx	1471160	42.695	29200	42.10	109.04	34653	13231
1	2	2	Brooklyn	Kings	2648771	91.559	34600	70.82	183.42	37137	14649
2	3	3	Manhattan	New York	1664727	600.244	360600	22.83	59.13	72033	27826
3	4	4	Queens	Queens	2358582	93.310	39600	108.53	281.09	21460	8354
4	5	5	Staten Island	Richmond	479458	14.514	30300	58.37	151.18	8112	3132
5	6	6	City of New York	City of New York	8622698	842.343	97700	302.64	783.83	28188	10947



	index	Year	Manhattan	Brooklyn	Queens	Bronx	Staten Is.	Total
9	9	1860	813669	279122	32903	23593	25492	1174779
10	10	1870	942292	419921	45468	37393	33029	1478103
11	11	1880	1164673	599495	56559	51980	38991	1911698
12	12	1890	1441216	838547	87050	88908	51693	2507414
13	13	†1900	1850093	1166582	152999	200507	67021	3437202
14	14	1910	2331542	1634351	284041	430980	85969	4766883
15	15	1920	2284103	2018356	469042	732016	116531	5620048
16	16	1930	1867312	2560401	1079129	1265258	158346	6930446
17	17	1940	1889924	2698285	1297634	1394711	174441	7454995
18	18	1950	1960101	2738175	1550849	1451277	191555	7891957
19	19	1960	1698281	2627319	1809578	1424815	221991	7781984
20	20	1970	1539233	2602012	1986473	1471701	295443	7894862
21	21	1980	1428285	2230936	1891325	1168972	352121	7071639
22	22	1990	1487536	2300664	1951598	1203789	378977	7322564
23	23	2000	1537195	2465326	2229379	1332650	443728	8008278
24	24	2010	1585873	2504700	2230722	1385108	468730	8175133
25	25	2016	1643734	2629150	2333054	1455720	476015	8537673

Visualising NYC historical census data from 1860 to 2016

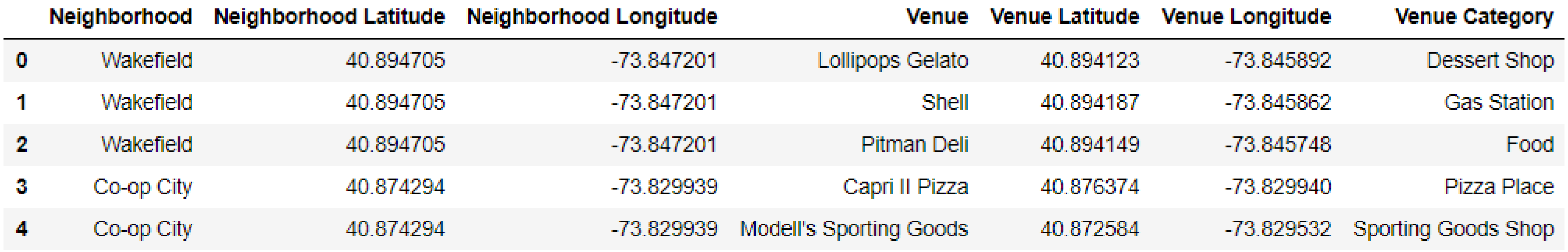


Linear Regression

- Since we got the NYC historical census data, it was possible to create a regression model in order to predict future population growth of each borough.
- A regression model was created and predictions were made for the years of 2020 and 2030.
- Making predictions was especially important for the future plan of the shopping mall.

Foursquare API

- For accessing the number of venues in each borough, Foursquare API was used within a radius of 250 meters.
- Below table is an example to the extracted data via Foursquare API



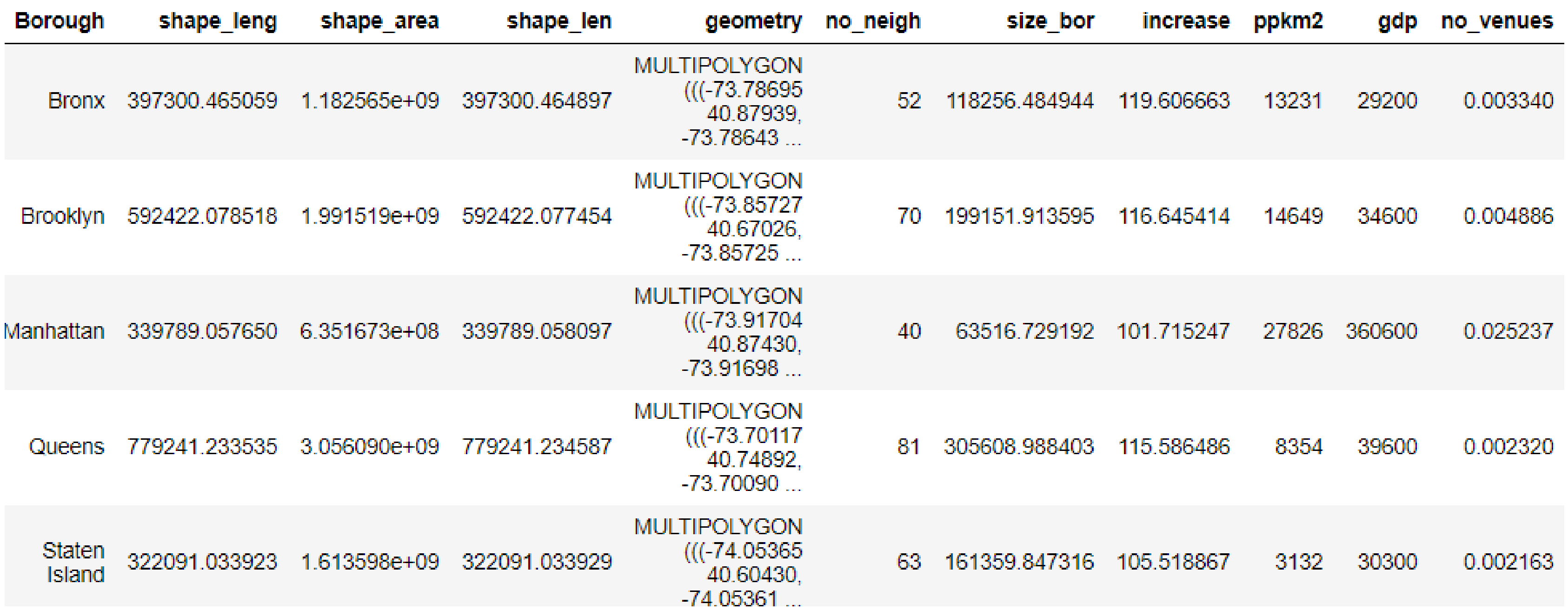
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station
2	Wakefield	40.894705	-73.847201	Pitman Deli	40.894149	-73.845748	Food
3	Co-op City	40.874294	-73.829939	Capri II Pizza	40.876374	-73.829940	Pizza Place
4	Co-op City	40.874294	-73.829939	Modell's Sporting Goods	40.872584	-73.829532	Sporting Goods Shop

Calculation of 6 indicators

- Number of neighborhoods per borough
- Size of each borough
- Expected population growth from 2017 to 2030
- Persons per square kilometers in 2017
- Gross domestic product (GDP) per capita in 2017
- Number of existing venues per hectare in each borough such as shopping malls, restaurants, cafés, etc.

  – Since each indicator has roughly different unit, they are normalised using minimum-maximum normalisation method.

Table of indicators



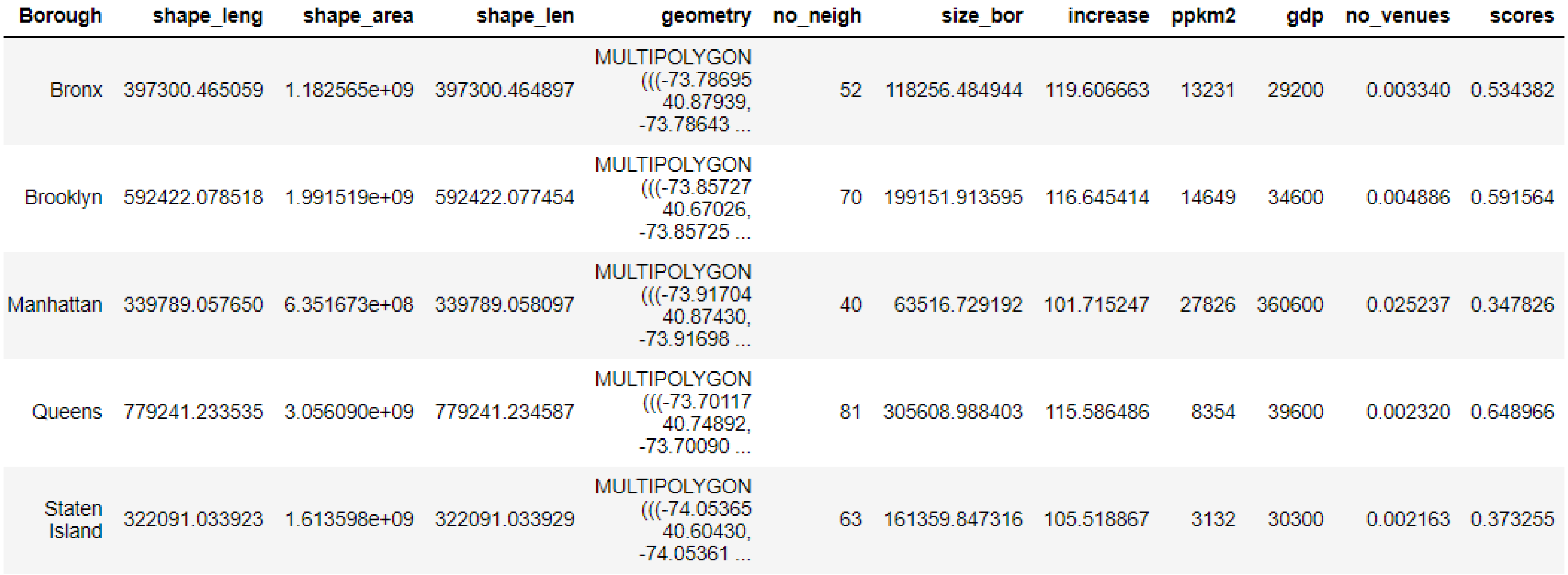
Borough	shape_leng	shape_area	shape_len	geometry	no_neigh	size_bor	increase	ppkm2	gdp	no_venues
Bronx	397300.465059	1.182565e+09	397300.464897	MULTIPOLYGON (((-73.78695 40.87939, -73.78643 ...	52	118256.484944	119.606663	13231	29200	0.003340
Brooklyn	592422.078518	1.991519e+09	592422.077454	MULTIPOLYGON (((-73.85727 40.67026, -73.85725 ...	70	199151.913595	116.645414	14649	34600	0.004886
Manhattan	339789.057650	6.351673e+08	339789.058097	MULTIPOLYGON (((-73.91704 40.87430, -73.91698 ...	40	63516.729192	101.715247	27826	360600	0.025237
Queens	779241.233535	3.056090e+09	779241.234587	MULTIPOLYGON (((-73.70117 40.74892, -73.70090 ...	81	305608.988403	115.586486	8354	39600	0.002320
Staten Island	322091.033923	1.613598e+09	322091.033929	MULTIPOLYGON (((-74.05365 40.60430, -74.05361 ...	63	161359.847316	105.518867	3132	30300	0.002163

Assigning importance scores

- Importance score for each criteria was assigned out 10
- **The most important criteria will be the number of already existing venues per hectare in each borough**
- Number of neighborhoods per borough: **5** (the higher the better in terms of diversity)
- Size of each borough: **6** (the larger the better)
- Expected population growth from 2017 to 2030: **9** (the higher the better)
- Person per square kilometers in 2017: **8** (the higher the better)
- Gross domestic product per capita in 2017: **8** (the higher the better)
- Number of existing venues per hectare in each borough: **10** (the lower the better)

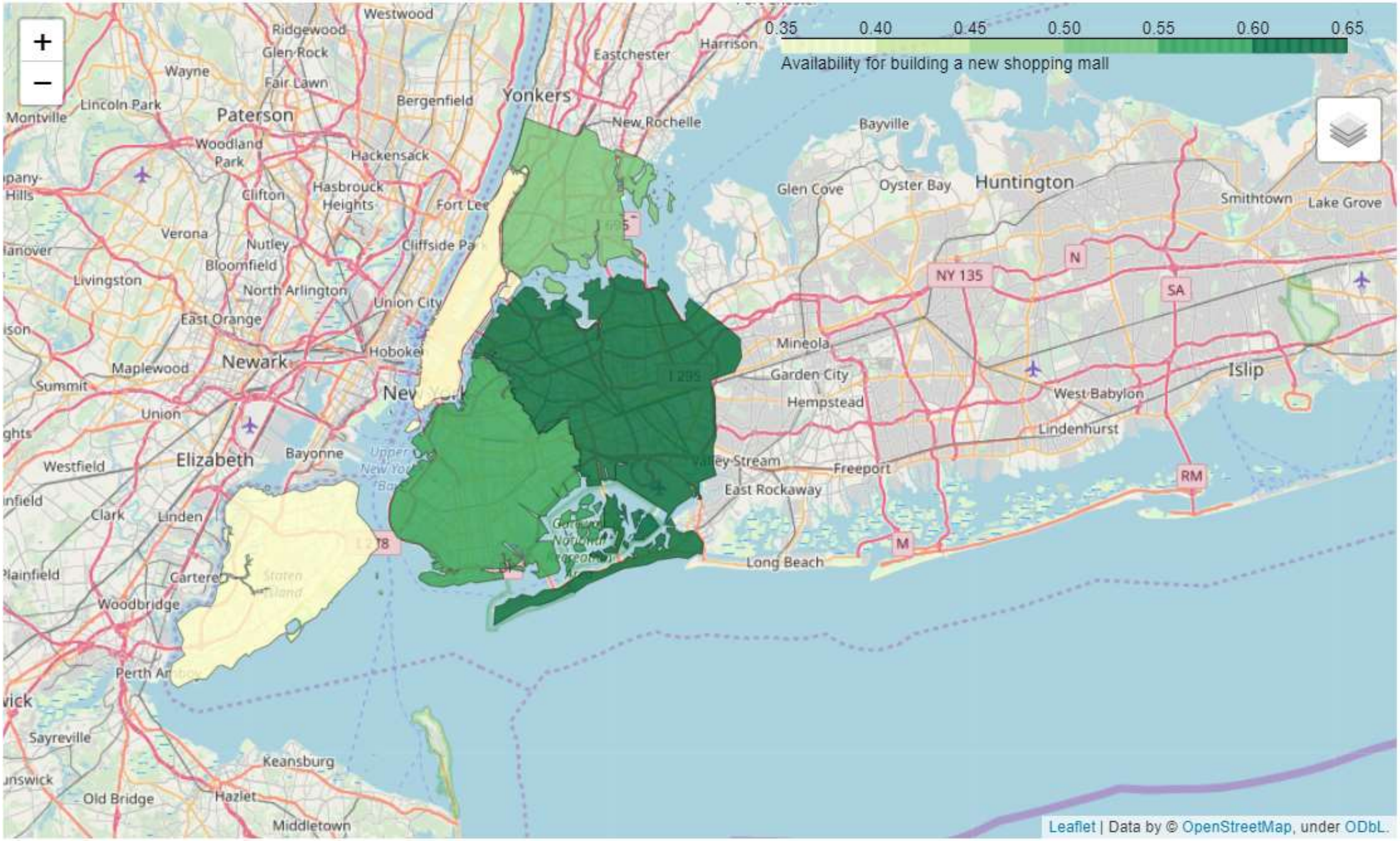
Calculating scores

- score = sum(weight * norm_value)
- Scores for each borough calculated summing the multiplications of normalised value of an indicator and its corresponding weight value.
- Borough Queens got the highest score



Borough	shape_leng	shape_area	shape_len	geometry	no_neigh	size_bor	increase	ppkm2	gdp	no_venues	scores
Bronx	397300.465059	1.182565e+09	397300.464897	MULTIPOLYGON (((-73.78695 40.87939, -73.78643 ...	52	118256.484944	119.606663	13231	29200	0.003340	0.534382
Brooklyn	592422.078518	1.991519e+09	592422.077454	MULTIPOLYGON (((-73.85727 40.67026, -73.85725 ...	70	199151.913595	116.645414	14649	34600	0.004886	0.591564
Manhattan	339789.057650	6.351673e+08	339789.058097	MULTIPOLYGON (((-73.91704 40.87430, -73.91698 ...	40	63516.729192	101.715247	27826	360600	0.025237	0.347826
Queens	779241.233535	3.056090e+09	779241.234587	MULTIPOLYGON (((-73.70117 40.74892, -73.70090 ...	81	305608.988403	115.586486	8354	39600	0.002320	0.648966
Staten Island	322091.033923	1.613598e+09	322091.033929	MULTIPOLYGON (((-74.05365 40.60430, -74.05361 ...	63	161359.847316	105.518867	3132	30300	0.002163	0.373255

Visualisation of scores of each borough on a map



Discussion

- By looking at the final map, it is clear to see that the most suitable borough for building a new shopping mall is Queens which has the highest score and the greenest one on the map.
- Borough Queens has the highest score. When the criteria values of this borough is examined, it is worth mentioning that Queens has the biggest area which is playing a very big role on the final score even though the importance score is the second least one compared to others.
- The most important criteria is the number of existing venues per hectare in each borough which was given by me. By looking at the value that Queens has for this is one of the lowest compared to other boroughs. Since this criteria has a negative impact on the final score, having a low value means that the score will be high.
- The least suitable borough for building a shopping mall is Manhattan which is then followed by Staten Island.
- Even though the GDP per capita is the highest one by far, the area size of Manhattan is the smallest one compared to others. In addition to the GDP, the population growth is not shiny compared to other boroughs. These reasons has made Manhattan in the last place in terms of availability.