Applied Data Science Capstone

The Battle of Neighborhoods

- Week 5 –

Project: Where to built a new gym taking geographic data into account by using the Foursquare API

Submitted by Julian Eppelsheimer

04.03.2021

1. Introduction | Business Problem:

As a data scientist it is your job to advise stakeholder considering all kind of problems based of your gained insight of available data. This project deals with the theoretical scenario, where a client instructs us to search the best place to build a new gym at. The company is operating internationally, so they want to build their branch either in the USA or in Canada. Moreover, the client specifies the location rather clear:

The new branch should be either built in the Long Island City neighbourhood in New York (USA), in the Allapattah neighbourhood in Miami (USA) or in St. James Town in Toronto (Canada). Therefor it is our job to gather and select data, get an insight into it to be able to advise our client.

This final project shall recap the use of the python library 'BeautifulSoup' to scrap websites. In addition to that 'Geocoder' library and the Foursquare API combined will deliver necessary geographic data. As a classification method the cluster algorithm 'k-means' will be used and demonstrated.

2. Data

The main used data will be geographic data. Therefor the 'Pandas' and 'BeautifulSoup' libraries will be used to scrap websites for their tables. In order to get the zip codes from the desired neighbourhoods, the following links will be used:

- Neighbourhoods of Toronto:
- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Neighbourhoods of Miami:
- https://www.zipdatamaps.com/nh-miami-neighborhood-allapattah
- Neighbourhoods of New York City:
- https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm

Scraping the mentioned websites:

1) Toronto - St. James Town

The scraped table was stored in a Pandas DataFrame:

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
8	M9A	Etobicoke	Islington Avenue, Humber Valley Village
9	M1B	Scarborough	Malvern, Rouge
11	M3B	North York	Don Mills
12	M4B	East York	Parkview Hill, Woodbine Gardens
13	M5B	Downtown Toronto	Garden District, Ryerson
14	M6B	North York	Glencairn
17	M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grov...
18	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
20	M3C	North York	Don Mills

	Postal Code	Borough	Neighbourhood
22	M5C	Downtown Toronto	St. James Town

2) Miami - Allapattah

The scraped table was stored in a Pandas DataFrame:

Key	Value
Neighborhood Name	Allapattah
City	Miami
County	Miami-Dade
Zip Code	33142
Area Code	305 / 786
Time Zone	Eastern Standard Time
Current Time:	EST
Population	4904
Majority Race	White 73.9%
Unemployment Level	5.9%

3) New York - Long Island City

The scraped table was stored in a Pandas DataFrame:

	Postal Code	Neighbourhood	Population
0	11004	Glen Oaks	14.016
1	11005	Floral Park	1.806
2	11101	Long Island City	25.484
3	11102	Astoria	34.133
4	11103	Astoria	38.780
5	11104	Sunnyside	27.232
6	11105	Astoria	36.688
7	11106	Astoria	38.875

The gathered Zip Codes were stored in a separate Pandas Dataframe:

	City	Neighbourhood	Postal Code
0	Toronto	St. James Town	M5C
1	Miami	Allappatah	33142
2	New York	Long Island City	11101

Coordinates of neighbourhoods:

By using the 'Geocoder' library I wanted to get the coordinates for the selected neighbourhoods for the scrapped postal codes. But as discussed before in week 4, by using it with Google, its not very reliable. That’s the reason why I had to get the coordinates manually by using the developer section of google under the following link:

https://developers.google.com/maps/documentation/geocoding/overview

These coordinates (latitude and longitude) were appended to the existing Pandas Dataframe:

City	Neighbourhood	Postal Code	Latitude	Longitude
Toronto	St. James Town	M5C	43.670867	-79.373306
Miami	Allappatah	33142	25.812779	-80.237708
New York	Long Island City	11101	40.744309	-73.941860

Explore the neighbourhoods:

In order to explore the neighbourhoods I used the Foursquare API. This API granted me different features to work with. To keep it simple, I selected the important features and dropped the unnecessary ones. The features are listed as follows:

Used features:
location.name, location.categories, location.latitude, location.longitude

Dropped features:
referralId, reasons.count, reasons.items, id, location.address, location.crossStreet, location.labeledLatLngs, location.distance, location.postalCode, location.cc, location.city, location.state, location.country, location.formattedAddress, delivery.id, delivery.url, delivery.provider.name, delivery.provider.icon.prefix, delivery.provider.icon.sizes, delivery.provider.icon.name, photos.count, photos.groups, venuePage.id

The features for each neighbourhood have been stored in separate Pandas Dataframes:

Long Island City – first 10 rows of the used Dataframe:

	name	categories	lat	lng
0	Wendy's	Fast Food Restaurant	25.811238	-80.239864
1	KFC	Fried Chicken Joint	25.811704	-80.240005
2	U-Haul Moving & Storage at 36th St	Storage Facility	25.808879	-80.237750
3	Subway	Sandwich Place	25.813109	-80.240756
4	McDonald's	Fast Food Restaurant	25.809014	-80.232281
5	Shell	Gas Station	25.808809	-80.239825
6	Taco Bell	Fast Food Restaurant	25.811701	-80.239949
7	El Presidente Supermarket	Food & Drink Shop	25.809744	-80.231959
8	L' Boulevard Cafe	Nightclub	25.809720	-80.238601
9	Studio 60	Nightclub	25.809007	-80.234309

Allapattah – first 10 rows of the used Dataframe:

	name	categories	lat	lng
0	Wendy's	Fast Food Restaurant	25.811238	-80.239864
1	KFC	Fried Chicken Joint	25.811704	-80.240005
2	U-Haul Moving & Storage at 36th St	Storage Facility	25.808879	-80.237750
3	Subway	Sandwich Place	25.813109	-80.240756
4	McDonald's	Fast Food Restaurant	25.809014	-80.232281
5	Shell	Gas Station	25.808809	-80.239825
6	Taco Bell	Fast Food Restaurant	25.811701	-80.239949
7	El Presidente Supermarket	Food & Drink Shop	25.809744	-80.231959
8	L' Boulevard Cafe	Nightclub	25.809720	-80.238601
9	Studio 60	Nightclub	25.809007	-80.234309

St. James Town – first 10 rows of the used Dataframe:

	name	categories	lat	lng
0	Maison Selby	Bistro	43.671232	-79.376618
1	Mr. Jerk	Caribbean Restaurant	43.667328	-79.373389
2	Cranberries	Diner	43.667843	-79.369407
3	Rooster Coffee House	Coffee Shop	43.669654	-79.379871
4	Murgatroid	Restaurant	43.667381	-79.369311
5	The Keg Steakhouse + Bar	Steakhouse	43.666756	-79.378302
6	F'Amelia	Italian Restaurant	43.667536	-79.368613
7	Absolute Bakery & Café	Bakery	43.667469	-79.369277
8	Cabbagetown Brew	Café	43.666923	-79.369289
9	Merryberry Cafe + Bistro	Café	43.666630	-79.368792

Population Data:

Population data for Allapattah und St. James Town have already been scraped from the web tables mentioned before. The population data for Long Island City can be found under the following link:

https://www.point2homes.com/US/Neighborhood/NY/Queens/Long-Island-City-Demographics.html

3. Exploratory Data Analysis

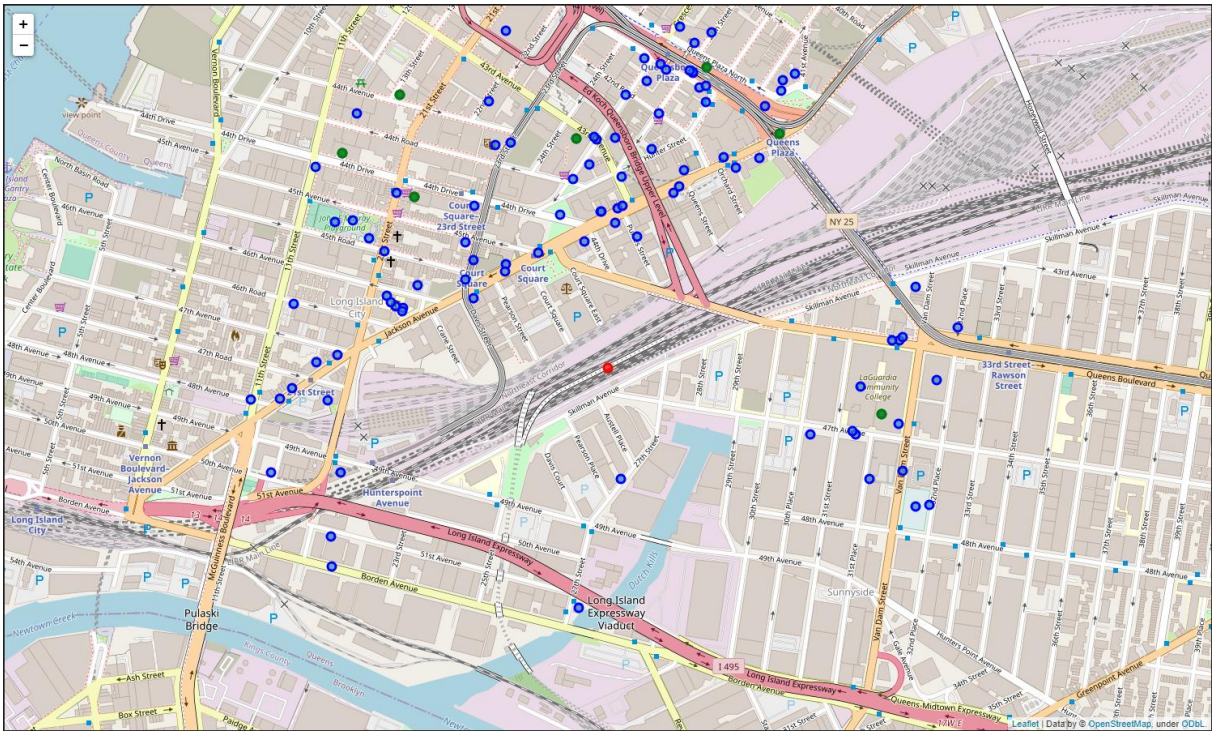
The needed data has been stored in a Pandas Dataframe:

Neighbourhood	City	Gyms	Venues	Population
Long Island City	New York	7	100	48188
Allapattah	Miami	0	20	4904
St. James Town	Toronto	2	80	25484

i) Visualizing the explored venues for every neighbourhood

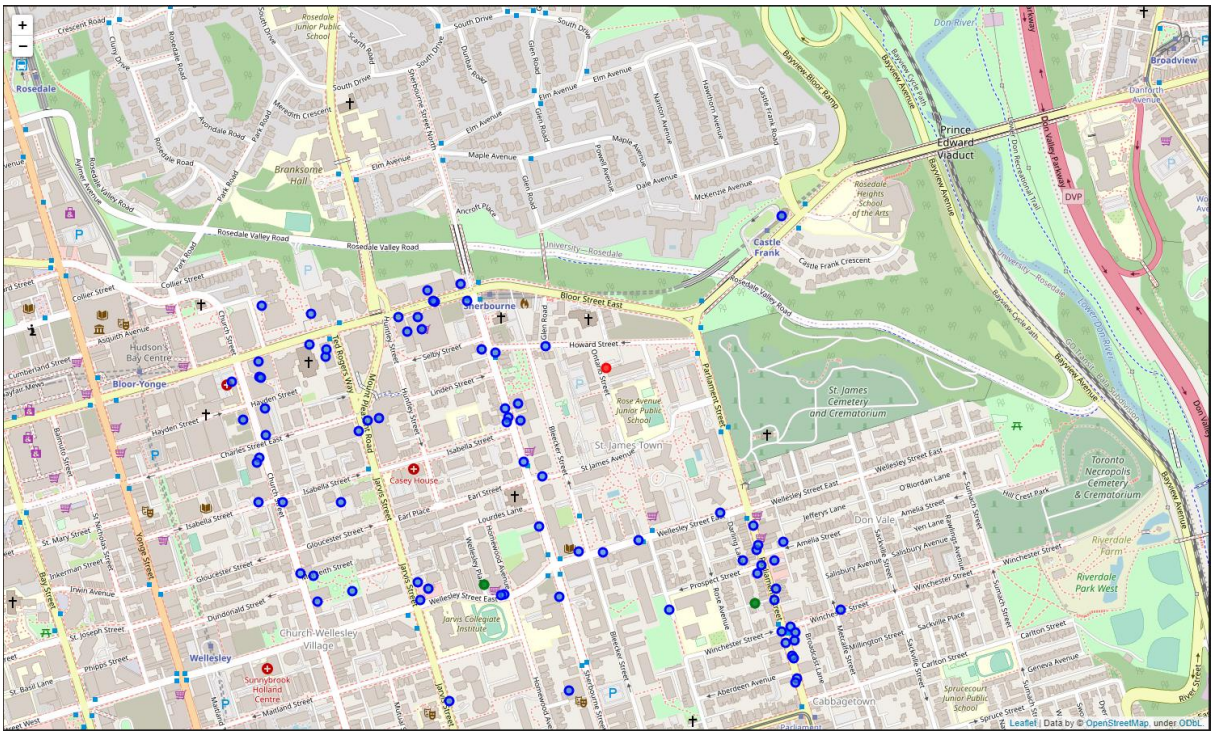
To get an overview of all the venue locations in every neighbourhood, the 'Folium' API was used. To visualize the data the 'folium' library was used. The resulting maps are displayed in the following graphics. The red marker represents the center of the neighbourhood, the green marker represents a gym venue and the blue marker display other explored venues.

Venue mapping of Long Island City:

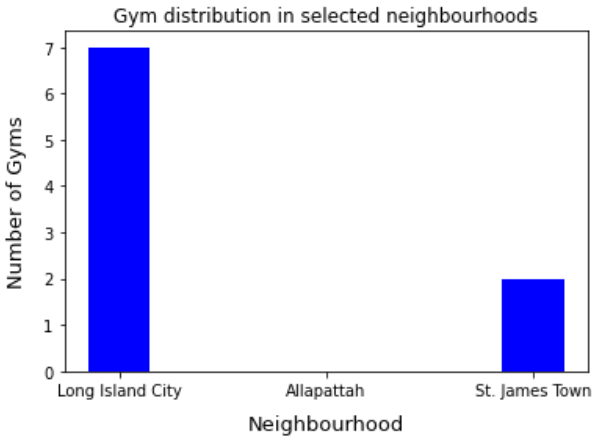


Venue mapping of Allapattah:

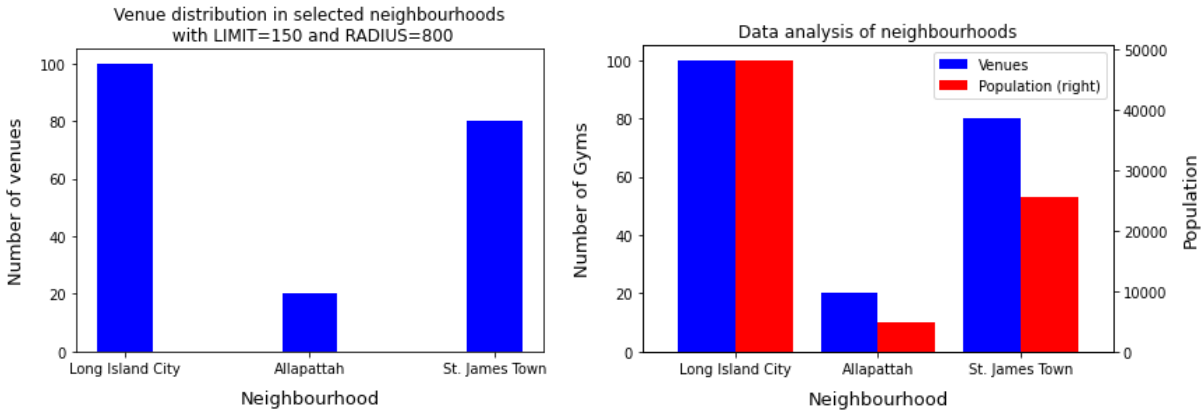
Venue mapping of St. James Town:



ii) Visualizing the number of gyms in every neighbourhood



In order to avoid unnecessary competition, Allapattah would be the best choice at first sight. However, we have to take the customer potential into account. This can be expressed by the overall amount of venues for each neighbourhood, as well as by the neighbourhood population:



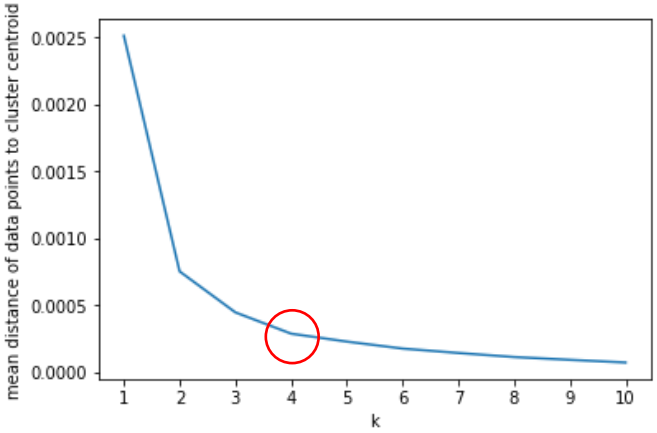
It could be shown, that in St. James Town the number of venues as well as the population is higher than in Allapattah, which offers a higher customer potential.

ii) Exploring the best location in St. James Town for the new gym

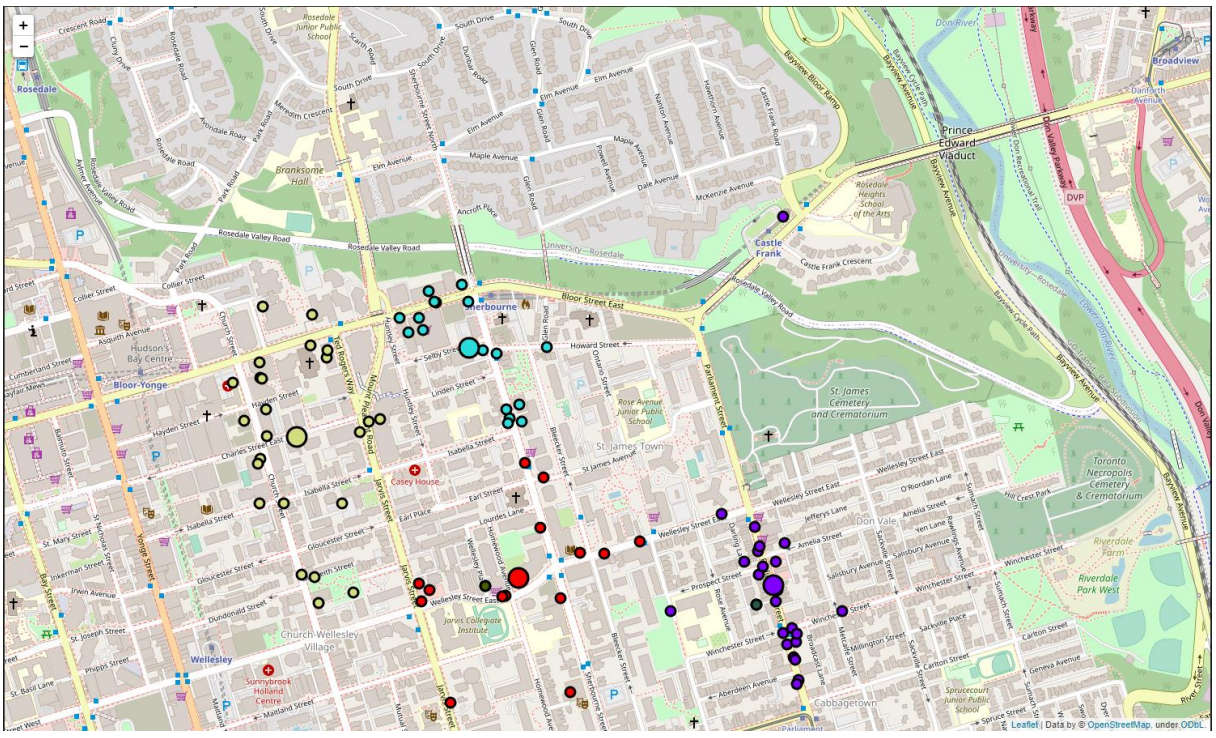
In order to determine the best location for the new gym, one can cluster the venues in St. James Town to get places of high interest by using k-means. Afterwards one can check if there are any gyms in the highlighted areas.

To get the best number of clusters for the 'k-means' algorithm, the model was fitted for a series of k values ranging from 1 to 10. The best k value will be determined by the mean distance of data points to their corresponding cluster centroid. The problem is that with increasing the number of clusters, the distance of centroids to data points will always

reduce. This means increasing K will always decrease the error. So, the value of the metric as a function of K is plotted and the elbow point is determined where the rate of decrease sharply shifts. It is the right K for clustering.



The elbow point can be found at k = 4, so the model was trained with four segment clusters. The resulting folium map looks like this:



Different colors are marking the venue clusters. Green color points display a gym venue. The big markers show the center of the corresponding clusters. The Gym should be build in St. James Town: Either in the area which is displayed by the yellow cluster segment or by the blue cluster segment, since there are no gyms yet.

4. Conclusion

By taking into account the number of gyms in Long Island City, Allapattah and St. James Town it turns out, that the new branch should definitely not be build in Long Island City, since there are already seven gyms. In Allapattah and St. James Town there are zero and two gyms, respectively. In order to avoid unnecessary competition, Allapattah would be the best choice at first sight. However, we must take the customer potential into account. This can be expressed by the overall number of venues for each neighbourhood, as well as by the neighbourhood population.

It could be shown, that in St. James Town the number of venues as well as the population is higher than in Allapattah, which offers a higher customer potential.

With the clustering algorithm 'k-means' areas of high interest have been clustered. The resulting folium plot shows the area with high business potential and no existing gyms. Therefor I could determine the two best areas for the new gym branch in St. James Town. These are marked in the following graphic with red circles.