New yoga class in Toronto

Bardakov Viktor

August 3, 2019

1. Introduction

1.1 Background

From internet (like site https://www.thegoodbody.com/yoga-statistics/) we can find out that Yoga has grown massively in popularity over the past few years, with passionate yogis stretching around the world. The facts show that it has many health benefits, particularly for those suffering from back pain, and with the spend on yoga products growing annually the trend shows no sign of slowing down. Also yoga market has massive annual spends Worldwide each year. So it's interesting trend and we can participate in it.

1.2 Problem

Need to find a good place for creating new yoga class in Toronto. Based on different open data source we should find on which attributes of Neighbourhoods have more influence on yoga studios and where is the best place for it, depending of base surroundings. So our definition of done will be list of neighbourhoods of Toronto, which have lack of yoga studios, calculated by some model.

2. Data

2.1 Data sources

From open sources (see link above) we can know, that most of yoga visitors are 30-49 years old (43 percent).. Also we can get, that most of yoga practitioners (75%) also take other exercises. So we need to get base demographical data upon the different neighbourhoods of Toronto. This data I got from Neighbourhood Profiles 2016 of official site of Toronto (https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/locations-and-mapping/#8c732154-5012-9afe-d0cd-ba3ffc813d5a). It contained a lot of data about age, sex, marital status, education and employment.

Also we need to define surroundings of yoga classes, which can be attractable for yoga visitors or participated by them. For that purpose I got data from Foursquare.com, which contained information about most public places of Toronto and also allow to get nearest of them.

As city map (topojson and geojson) I used data from open sources (from GitHub - https://github.com/jasonicarter/toronto-geojson)

2.2 Data cleaning

Demographical data downloaded from official site and were combined into one table. There were a lot unnecessary information, so I left there only info about how many:

- male persons are splitted on ages
- female person are splitted on ages
- married persons
- not married persons
- person have no certificate\diploma\degree
- person have secondary certificate\diploma\degree
- person have postgraduate certificate\diploma\degree
- employed persons
- unemployed

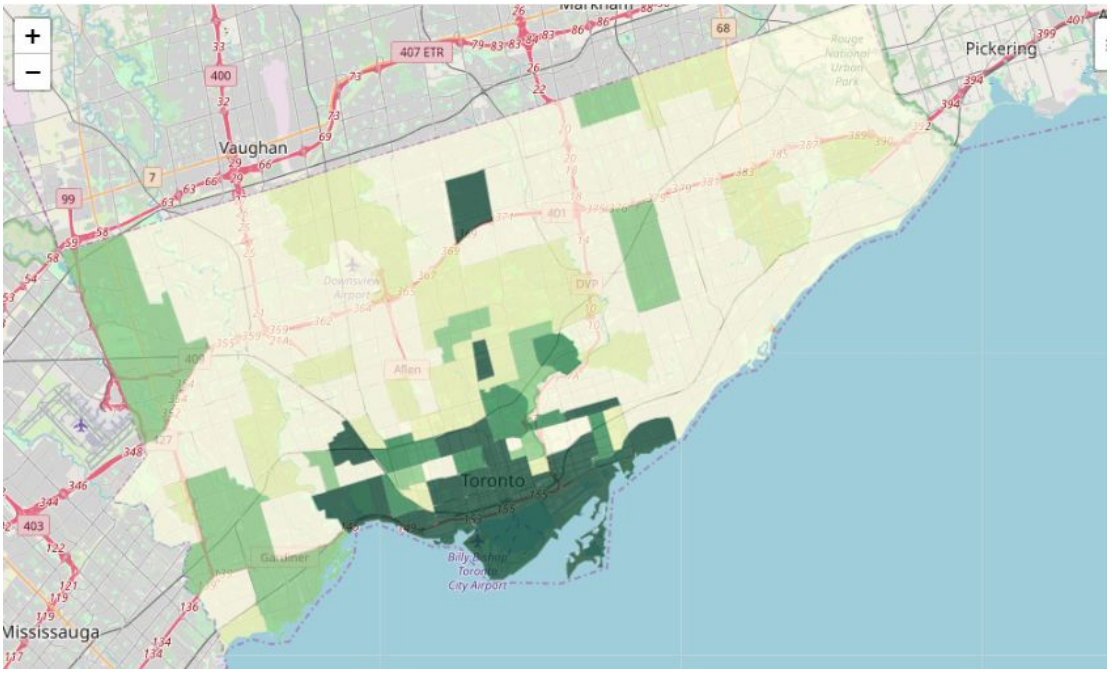
Empty data were filled by zeros, All data were grouped by different neighbourhoods and changed to relative values (certain value / total population of city)

Idx	Male: 0 to 04 years	Male: 05 to 09 years	Male: 10 to 14 years	Male: 15 to 19 years	Male: 20 to 24 years	Male: 25 to 29 years	Male: 30 to 34 years	Male: 35 to 39 years	Male: 40 to 44 years	Male: 45 to 49 years	...	Female: 100 years and over	Married	
Neighbourhood														
Agincourt North	0.009443	0.010022	0.010162	0.011315	0.010419	0.008911	0.007668	0.007229	0.008783	0.009795	...	0.015385	0.012020	(
Agincourt South-Malvern West	0.008227	0.007787	0.007083	0.010506	0.010265	0.009174	0.007530	0.006644	0.007049	0.008365	...	0.000000	0.009485	(
Alderwood	0.005151	0.003893	0.003464	0.003839	0.003644	0.003117	0.003765	0.004837	0.004854	0.004843	...	0.007692	0.004855	(
Annex	0.006367	0.005263	0.005004	0.006263	0.012472	0.018261	0.014785	0.011215	0.009649	0.009355	...	0.038462	0.010391	(
Banbury-Don Mills	0.008155	0.009517	0.010393	0.009631	0.007186	0.005663	0.006750	0.007813	0.009418	0.011116	...	0.030769	0.011207	(

Also I got Toronto coordinates and get from foursquare.com all yoga classes across the city and it's coordinates. This was not enough, cause I had to split this yogas across different neighbourhoods. I got coordinates of each neighbourhoods Toronto from open source geojson file and create polygon object. After that I checked each yoga studio from my list and count it in certain neighbourhood, if its coordinated had been included into neighbourhood poligon. So that's how I got data about **count of yoga classes** across different neighbourhoods.

id	coordinates	geometry.type	Neighbourghood	type	Poligon	Centre	Venues	VenueCount
097	[[[-79.39119482699992, 43.68108112399995], [-7...	Polygon	Yonge-St.Clair (97)	Feature	Path(array([[-79.39119483, 43.68108112],\n ...	(-79.39853901899993, 43.68808772449996)	{'meta': {'code': 200, 'requestId': '5d45c4147...	0.006061
027	[[[-79.50528791599992, 43.759873493999955], [-...	Polygon	York University Heights (27)	Feature	Path(array([[-79.50528792, 43.75987349],\n ...	(-79.4922534984999, 43.76490100749996)	{'meta': {'code': 200, 'requestId': '5d45c4152...	0.006061
038	[[[-79.43998431099992, 43.761557654999955], [-...	Polygon	Lansing-Westgate (38)	Feature	Path(array([[-79.43998431, 43.76155765],\n ...	(-79.42441129199992, 43.75211632699995)	{'meta': {'code': 200, 'requestId': '5d45c415c...	0.000000
031	[[[-79.43968732599991, 43.70560981799996], [-7...	Polygon	Yorkdale-Glen Park (31)	Feature	Path(array([[-79.43968733, 43.70560982],\n ...	(-79.45555520499992, 43.71487854199995)	{'meta': {'code': 200, 'requestId': '5d45c4154...	0.000000
016	[[[-79.49262119699992, 43.6474363499999], [-79...	Polygon	Stonegate-Queensway (16)	Feature	Path(array([[-79.4926212 , 43.64743635],\n ...	(-79.49884145449992, 43.63527967899996)	{'meta': {'code': 200, 'requestId': '5d45c416a...	0.000000

Map with sum of yoga studios by different neighbourhoods



Then I got all nearest places from foursquare.com for each yoga studio and grouped them by each neighbourhood. That's how I got all **nearest objects to yoga classes** which could be used as surroundings.

	YogaName	Yoga Latitude	Yoga Longitude	id	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Moksha Yoga Uptown	43.688799	-79.394484	097	The Bagel House	43.687374	-79.393696	Bagel Shop
1	Moksha Yoga Uptown	43.688799	-79.394484	097	Cava Restaurant	43.689809	-79.394932	Tapas Restaurant
2	Moksha Yoga Uptown	43.688799	-79.394484	097	9bars	43.688660	-79.391940	Café
3	Moksha Yoga Uptown	43.688799	-79.394484	097	DAVIDsTEA	43.688376	-79.394158	Tea Room
4	Moksha Yoga Uptown	43.688799	-79.394484	097	Capocaccia Café	43.685915	-79.393305	Italian Restaurant

This data were summarized and used for next researching.

2.3 Feature selection

All data were grouped by neighbourhoods and get mean values. After that all parameters were checked on correlation (using numpy).

Found, that across nearest places most correlated were

```
Latin American Restaurant    0.628757
Hotel Bar                    0.594660
Arepa Restaurant             0.581358
Street Art                   0.546436
Udon Restaurant              0.540545
Speakeasy                    0.524478
Aquarium                     0.509186
Other Nightlife              0.509186
Lake                         0.509186
Baseball Stadium             0.509186
```

From this list selected 'Aquarium','Basketball Stadium','Lake','Baseball Stadium', as which were most nearest to exercises and other sport activities.

From demographic parameters most correlated were

```
Male: 30 to 34 years      0.647936
Female: 25 to 29 years    0.630664
Female: 30 to 34 years    0.625783
Male: 25 to 29 years      0.611233
Male: 35 to 39 years      0.606367
Postsecondary diploma     0.562332
Employed                  0.526583
Female: 35 to 39 years    0.494748
Male: 40 to 44 years      0.488849
```

From that list selected 'Male: 30 to 34 years','Female: 25 to 29 years','Female: 30 to 34 years','Male: 25 to 29 years','Male: 35 to 39 years' as most correlated.

3. Methodology

My purpose was to find out, which neighbourhoods has lack of yoga studios. So first of all I had to to calculate model, which will count it. Because I had a lot of parameters, I use multi linear regression model.

```
In [56]: X = ComposedData[['Aquarium','Basketball Stadium','Lake','Baseball Stadium','Male: 30 to 34 years','Female: 25 to 29 years','Female: 30 to 34 years','Male: 25 to 29 years','Male: 35 to 39 years']]
         Y = ComposedData['VenueCount']

In [57]: regr = linear_model.LinearRegression()
         regr.fit(X, Y)

Out[57]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
                  normalize=False)

In [58]: print('Intercept: \n', regr.intercept_)
         print('Coefficients: \n', regr.coef_)

         Intercept:
          0.0016202586839977147
         Coefficients:
          [ 1.09156096  0.27289024  0.27289024  1.09156096  5.06191523  3.67004876
          -3.66992319 -4.43284991  0.13961876]
```

After that I could calculate predictable value of Yoga classes (relative to all city), and compare it with real data.

```
ComposedData['ModelValues']=regr.predict(X)
ComposedData['res']=ComposedData['ModelValues']-ComposedData['VenueCount']
```

Let's select 2 Neighbourhoods, where difference between model value and real quantity of yoga centres is the biggest

```
ComposedData.sort_values(by=(['res']), ascending=False).head(2)
```

]:

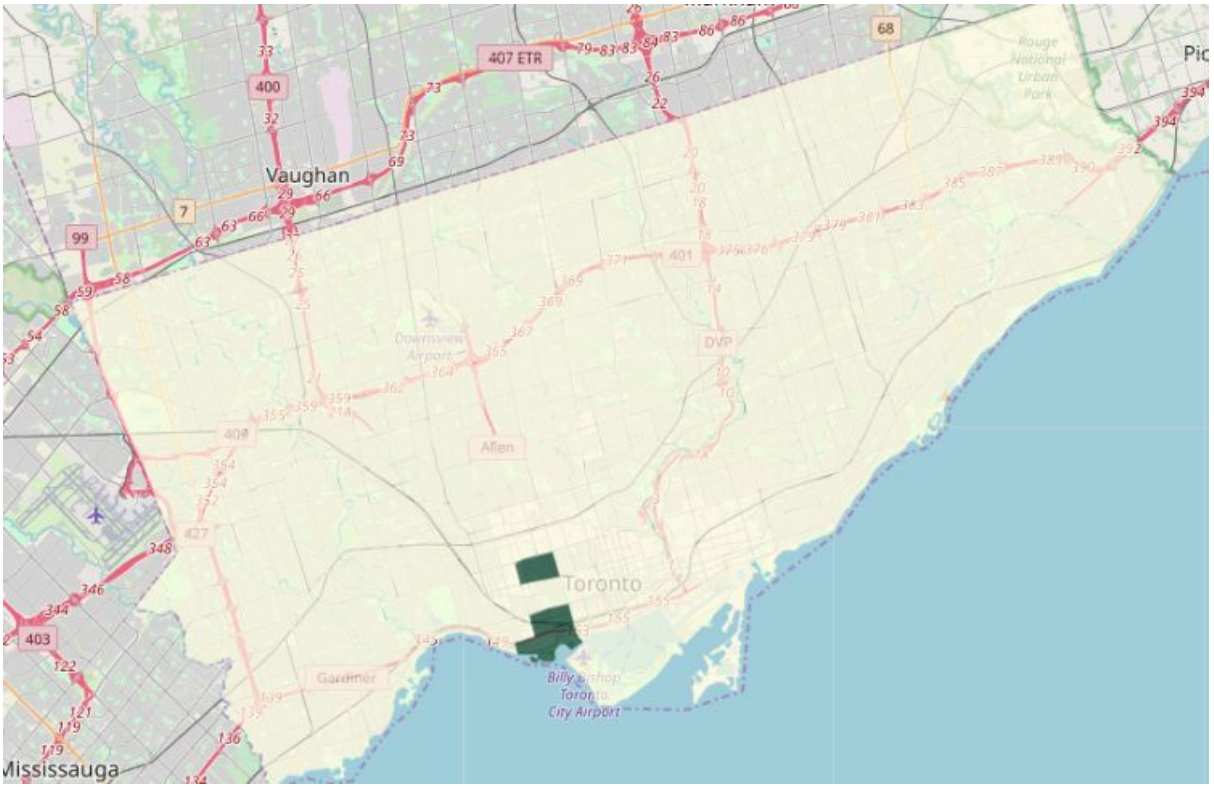
	Aquarium	Basketball Stadium	Lake	Baseball Stadium	Male: 30 to 34 years	Female: 25 to 29 years	Female: 30 to 34 years	Male: 25 to 29 years	Male: 35 to 39 years	VenueCount	ModelValues	res
082	0.0	0.0	0.0	0.0	0.031223	0.030326	0.029220	0.028664	0.023334	0.0	0.026323	0.026323
080	0.0	0.0	0.0	0.0	0.008357	0.011089	0.008169	0.009130	0.006272	0.0	0.015126	0.015126

```
They are 82 and 80 neighbourhoods
```

4. Results

As we can see above, we got 2 neighbourhoods of Toronto, which have lack of yoga classes. It's

Niagara and Palmerston-Little Italy



As we can see Niagara and Palmerston-Little Italy don't have yoga studios at all.

According model Niagara have enough population and nearest surroundings to support 4 yoga studios, and Palmerston-Little Italy for 2 new.

5. Discussion

Of course, process of selecting new place for yoga studio can't be such simple. There are tons of additional parameters, such as:

- nearest transport hubs (ex. metro)
- nearest office buildings (which are absent in foursquare)
- rent price

But this research compare basic parameters to find out which of them are better and gain direction to find out main answer on questions, which this work based on.

6. Conclusion

In this study, I analyzed the relationship between demographic and nearest neighbours of certain yoga classes to amount of classes itself (depending of each neighbourhood of Toronto). Worked a lot with geo data and data manipulation. Of course, for big real projects this amount of work is not enough, cause in real - there are much more parameters and dependencies, but even this amount can show the way for appropriate decision process of opening new yoga class.