How to select your best neighborhood in Toronto

Denis Deis

March 2020

1. Introduction

1.1. Background

Toronto is the most populous city in Canada, with a population of 3 million. The city is the anchor of the Golden Horseshoe, it is an international center of business, finance, arts, culture, and it is recognized as one of the most multicultural and cosmopolitan cities in the world. (based on https://en.wikipedia.org/wiki/Toronto).

1.2. Problem

In this paper, we will analyze how to select your best neighborhood, when you are moving to Toronto, or when you are moving within Toronto. Considering that all families and their needs are different, we will develop a short questionnaire with weights to understand what is essential for a particular family to make the best suggestion.

1.3. Interest

All people who are considering moving to Toronto can use this tool to find the best neighborhood for their purposes.

2. Data acquisition and cleaning

2.1. Data sources

We will get data about Toronto neighborhoods from the Toronto Open Data Portal https://open.toronto.ca/

Neighborhoods data sample:

	_id	AREA_ID	AREA_ATTR_ID	PARENT_AREA_ID	AREA_SHORT_CODE	AREA_LONG_CODE	AREA_NAME	AREA_DESC	X	Y	LONGITUDE	LATIT
0	3781	25886861	25926662	49885	094	094	Wychwood (94)	Wychwood (94)	None	None	-79.425515	43.67
1	3782	25886820	25926663	49885	100	100	Yonge-Eglinton (100)	Yonge-Eglinton (100)	None	None	-79.403590	43.70
2	3783	25886834	25926664	49885	097	097	Yonge-St.Clair (97)	Yonge-St.Clair (97)	None	None	-79.397871	43.68
3	3784	25886593	25926665	49885	027	027	York University Heights (27)	York University Heights (27)	None	None	-79.488883	43.76
4	3785	25886688	25926666	49885	031	031	Yorkdale-Glen Park (31)	Yorkdale-Glen Park (31)	None	None	-79.457108	43.71

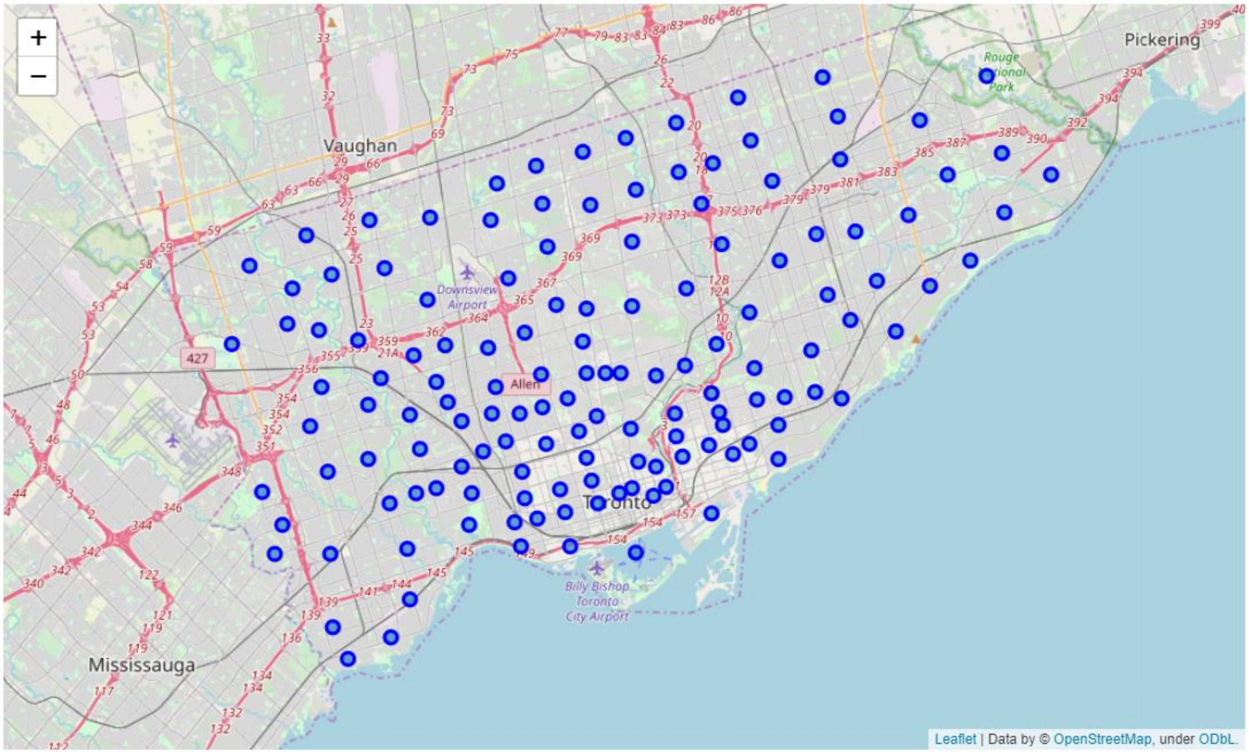
And then, we will enrich the data with venues using http://foursquare.com/ API.

Venues data sample:

	code	neigborhood_latitude	neigborhood_longitude	venue	venue_latitude	venue_longitude	venue_category_id
0	1	43.71618	-79.596356	Comfort Hotel	43.716058	-79.594135	4bf58dd8d48988d1fa931735
1	1	43.71618	-79.596356	Burger King	43.719967	-79.601043	4bf58dd8d48988d16e941735
2	1	43.71618	-79.596356	Zaytoun	43.714980	-79.593675	503288ae91d4c4b30a586d67
3	1	43.71618	-79.596356	Tim Hortons	43.714657	-79.593716	4bf58dd8d48988d1e0931735
4	1	43.71618	-79.596356	Domino's Pizza	43.719329	-79.594570	4bf58dd8d48988d1ca941735
...	...	...	...	...	...	...	...
114	1	43.71618	-79.596356	Woodbine Ribfest	43.718170	-79.597435	56aa371be4b08b9a8d57350b
115	1	43.71618	-79.596356	Baby Donut	43.720369	-79.599806	4bf58dd8d48988d148941735
116	1	43.71618	-79.596356	Urban Behavior	43.720200	-79.600691	4bf58dd8d48988d103951735
117	1	43.71618	-79.596356	Canes Family Health Team	43.716920	-79.592241	4bf58dd8d48988d177941735
118	1	43.71618	-79.596356	Fusia	43.720073	-79.600666	4bf58dd8d48988d145941735

2.2. Data cleaning

We have 140 different neighborhoods from https://open.toronto.ca/, and a lot of data about them, including geometry data. It is not essential for our purpose so that we will keep only Codes, Neighborhood Names, and their respective coordinates.



We have information about 14197 venues in these 140 neighborhoods. To aggregate this data, we need to categorize it. The most common way to do it is to apply the highest level of venue category. There are ten different categories at the highest level of venue categories in Foursquare data.

2.3. Feature selection

We can use our categories as features, but we want to remove non-important. Events are temporary, so we should not consider them. And we are not interested in the number of residences. We could be interested in details, but a number tells almost nothing about the neighborhood.

Finally, we will use the following features:

1. Arts & Entertainment
2. College & Education
3. Food
4. Nightlife
5. Outdoors & Recreation
6. Professional
7. Shops

8.   Travel

We calculate the number of venues of each category per neighborhood and store it in the data frame for analysis.

Our dataset should have a following look:

```
In [26]: neighborhoods_data.head()
```

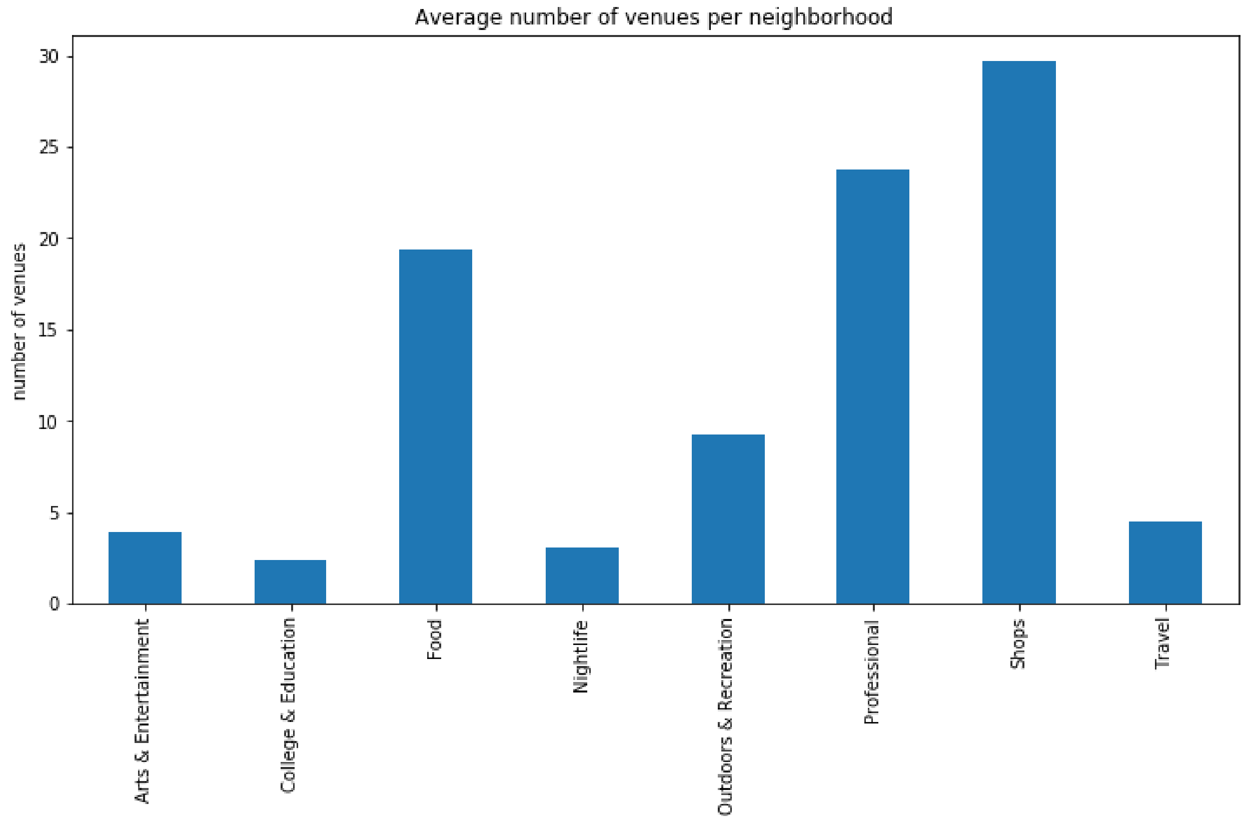
Out[26]:

code	neighborhood	longitude	latitude	Arts & Entertainment	College & Education	Food	Nightlife	Outdoors & Recreation	Professional	Shops	Travel
1	West Humber-Clairville	-79.596356	43.716180	4	1	38	0	3	15	52	5
2	Mount Olive-Silverstone-Jamestown	-79.587259	43.746868	2	2	26	1	1	20	32	3
3	Thistletown-Beaumond Heights	-79.563491	43.737988	2	3	32	2	2	17	32	2
4	Rexdale-Kipling	-79.566228	43.723725	6	1	22	2	5	36	33	6
5	Elms-Old Rexdale	-79.548983	43.721519	1	0	32	2	7	17	38	8

3. Exploratory Data Analysis

3.1. Average figures

First of all, let's take a look at the average number of venues between neighborhoods:



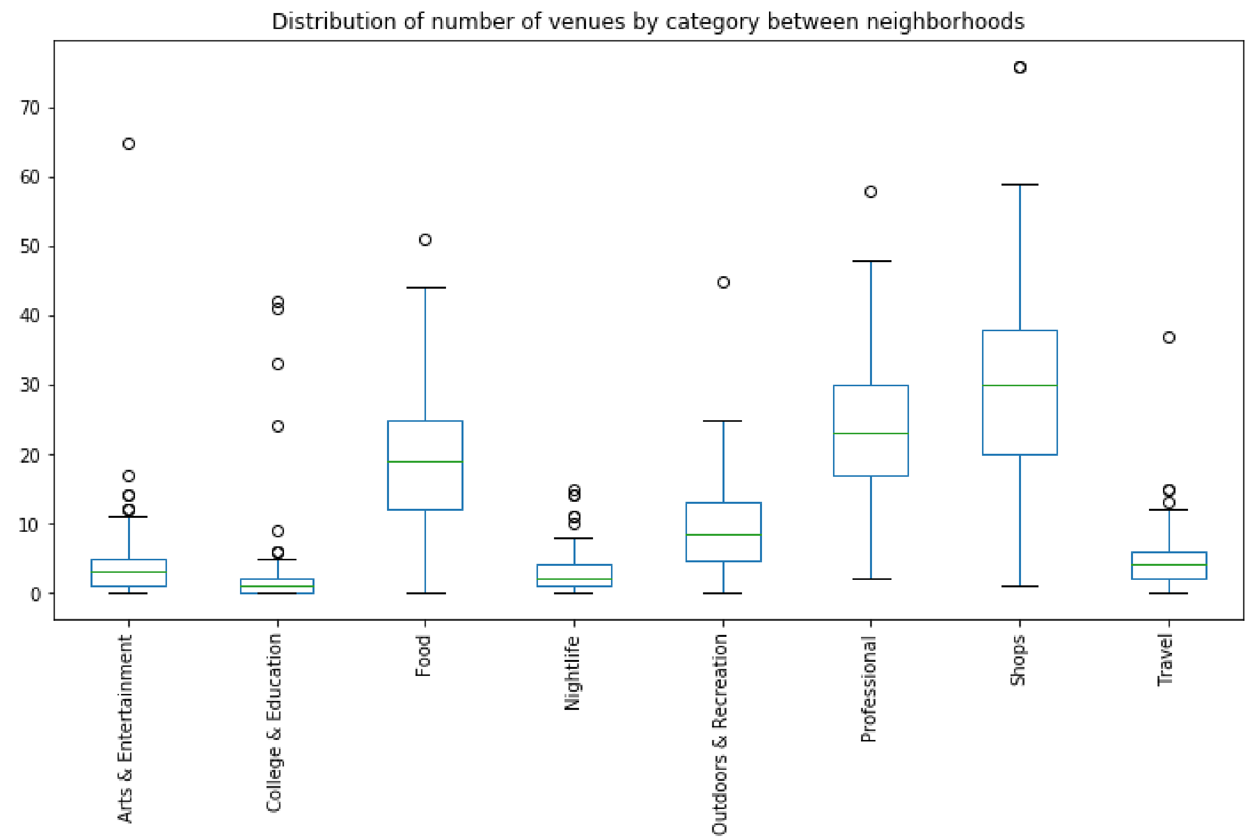
Average figures are the following:

```
Arts & Entertainment       3.914286
College & Education        2.378571
Food                      19.364286
Nightlife                  3.028571
Outdoors & Recreation      9.214286
Professional              23.800000
Shops                     29.642857
Travel                     4.471429
```

We can see that the most popular categories are "Shops", "Professional", and "Food". This is reasonable, as there are a lot of small businesses within these categories.

3.2. Distributions

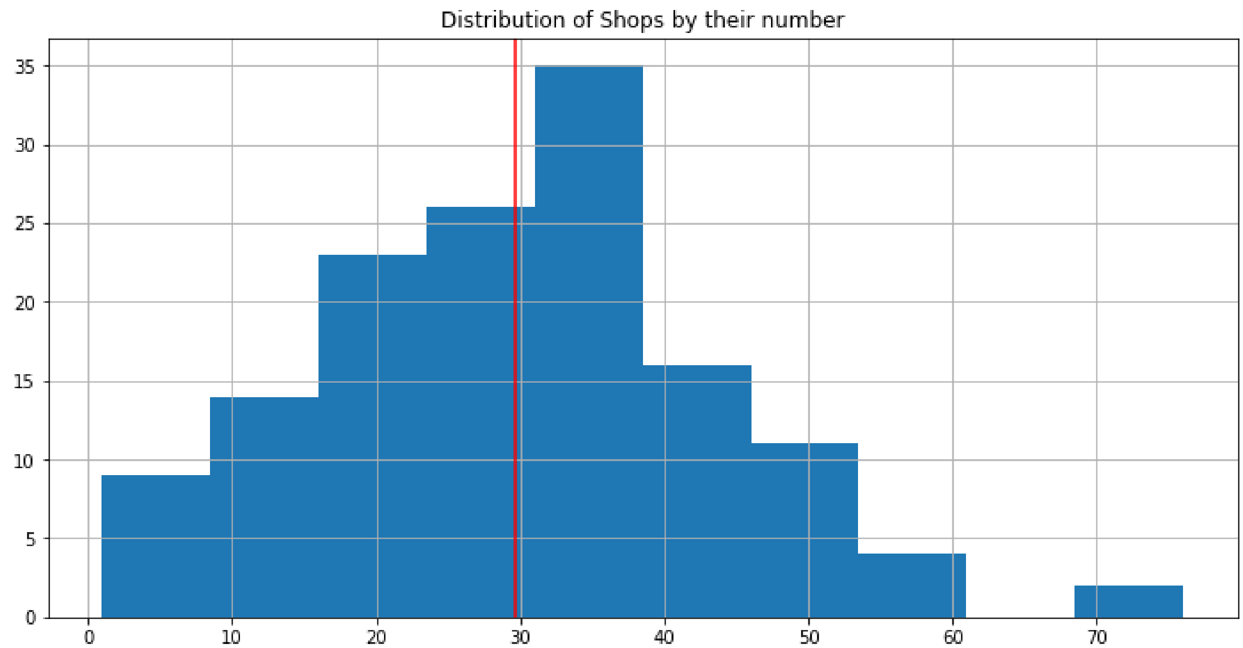
Let's take a look at the distribution of these categories with the help of a box plot diagram.

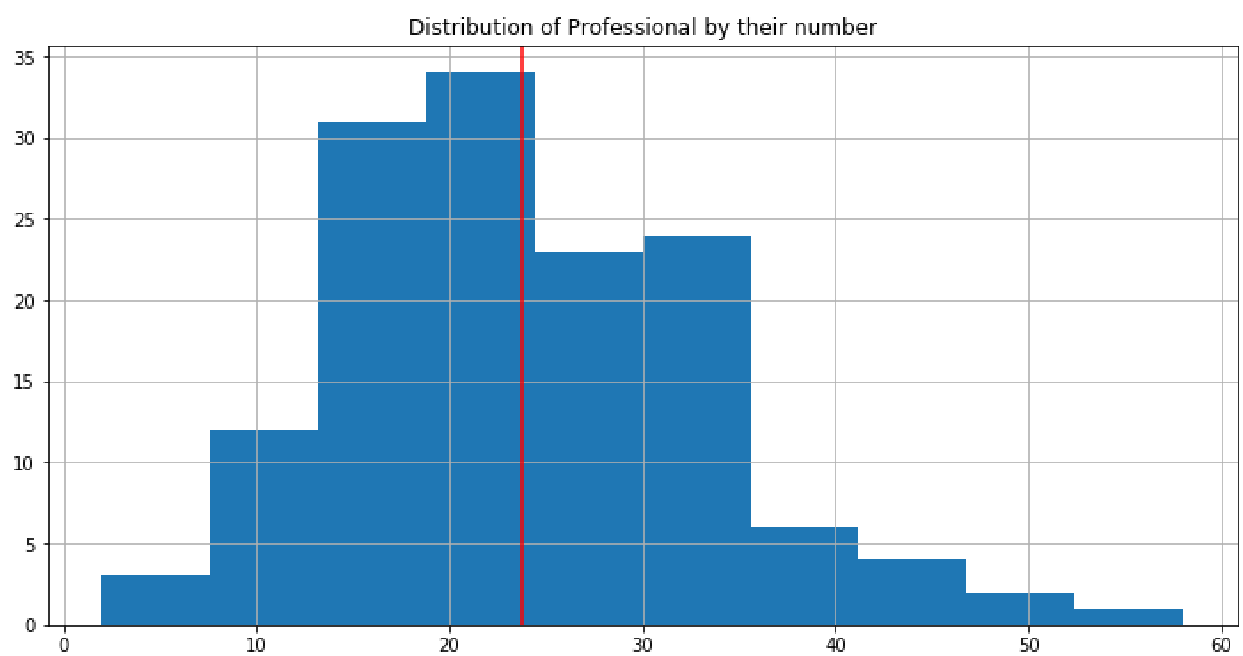


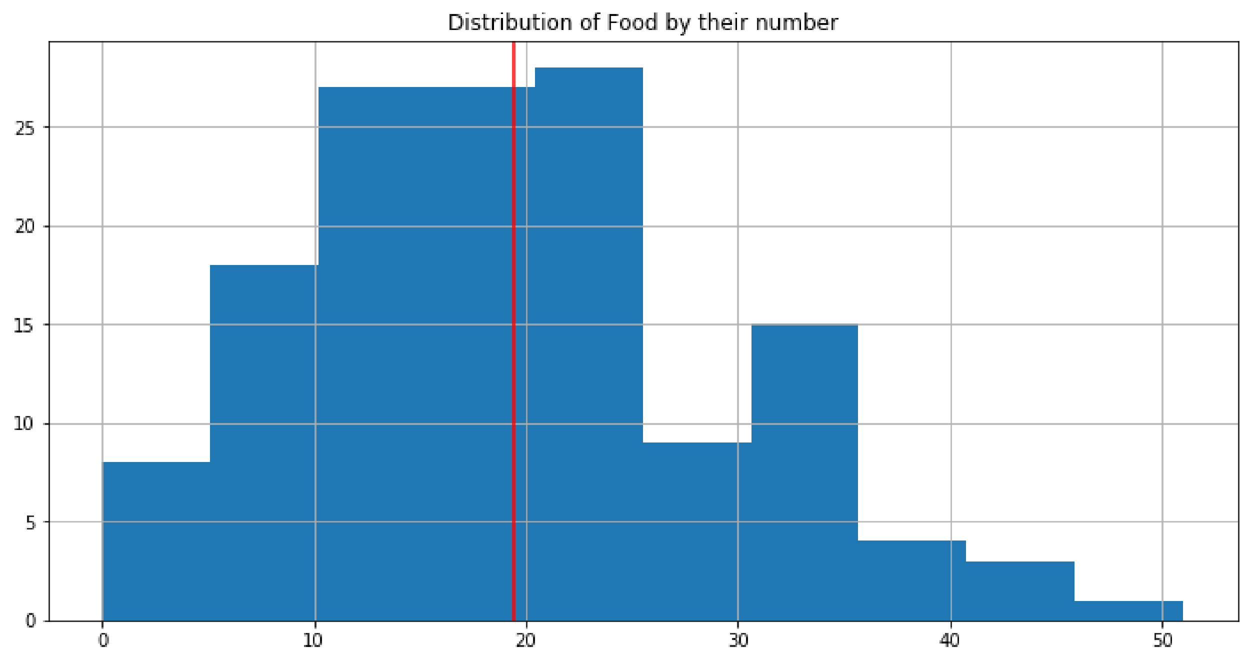
We can see that the variance of features is different. Let's look at particular values of the standard deviation of these features.

```
Arts & Entertainment       6.060232
College & Education        5.968102
Food                       9.781211
Nightlife                  2.851083
Outdoors & Recreation      6.322036
Professional               9.488501
Shops                     13.692445
Travel                     4.004391
```

"Shops", "Professional", and "Food" are the categories with the highest variance. Let's take a look at their distribution. A red vertical line is an average value.



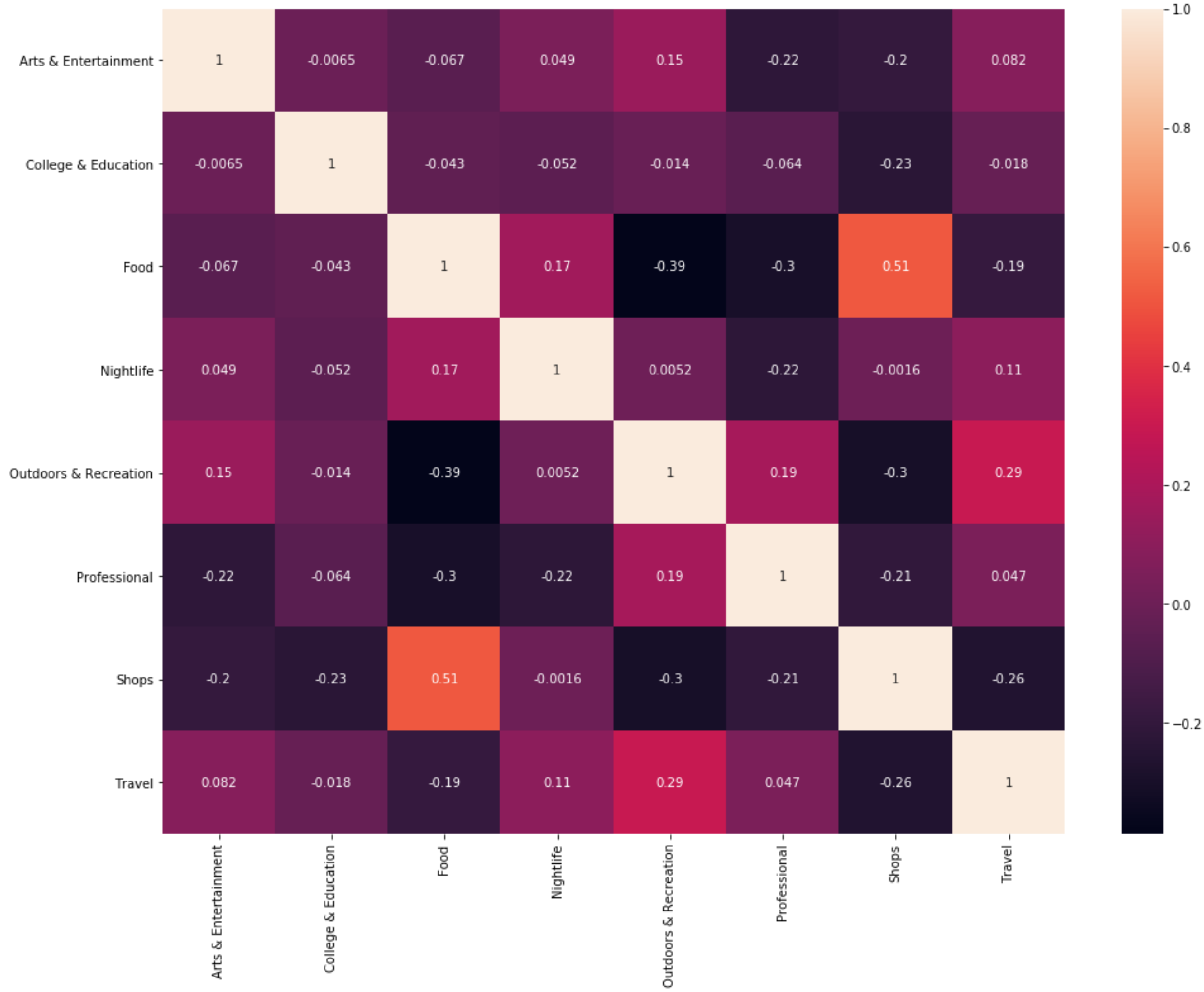




All three distributions look reasonable: peak in the middle and skewed to the right. We can consider it as a normal one.

3.3. Correlations

Let's take a look at correlations of the number of venues by neighborhood.



We can see a small positive correlation between Food and Shops categories, which is pretty reasonable. There are no strong correlations between these features, so we can use all of them for our analysis.

4. Modeling

4.1. Target values

The objective of this task is to help a family who is moving to Toronto to select an optimal neighborhood for their needs. Considering that all families have different needs, we will set up parameters for some "standard" family. However, these parameters can be adjusted in any way for other families. To select parameters, we need to understand how the family values different features. To do this, we will ask to set up parameters on a scale 0..100, where 0 means: "I don't want to have such venues in the neighborhood", and 100 says, "I want the available maximum of these venues in the neighborhood."

We've selected the following values:

```
arts_and_entertainment = 80 # the family likes arts and entertainment
college_and_education = 50 # they consider nearby colleges as not important
food = 80 # restaurants and coffee_shops are important
nightlife = 0 # they dislike nightlife and want to avoid it
outdoors_and_recreation = 90 # outdoors and recreations are very important
professional = 20 # they prefer residential areas, these will limit business districts but will keep some local businesses
shops = 60 # they want to have some variety of shops, but not too much
travel = 0 # travel amenities could bring disturbance, and they want to avoid them
```

We will use these values as percentiles of the particular category.

4.2. Data preprocessing

As we have different variance and numbers of venues, first of all, we need to standardize this data. We will use StandardScaler from sklearn.preprocessing to do it.

Normalized data should look in the following way:

```
In [38]: neighborhoods_norm.head()
Out[38]:
```

code	neighborhood	longitude	latitude	Arts & Entertainment	College & Education	Food	Nightlife	Outdoors & Recreation	Professional	Shops	Travel
1	West Humber-Clairville	-79.596356	43.716180	0.014195	-0.231819	1.912097	-1.066067	-0.986486	-0.930768	1.638672	0.132472
2	Mount Olive-Silverstone-Jamestown	-79.587259	43.746868	-0.317011	-0.063660	0.680850	-0.714064	-1.303976	-0.401923	0.172767	-0.368773
3	Thistletown-Beaumond Heights	-79.563491	43.737988	-0.317011	0.104499	1.296474	-0.362061	-1.145231	-0.719230	0.172767	-0.619396
4	Rexdale-Kipling	-79.566228	43.723725	0.345400	-0.231819	0.270435	-0.362061	-0.668996	1.290384	0.246063	0.383094
5	Elms-Old Rexdale	-79.548983	43.721519	-0.482614	-0.399978	1.296474	-0.362061	-0.351506	-0.719230	0.612539	0.884340

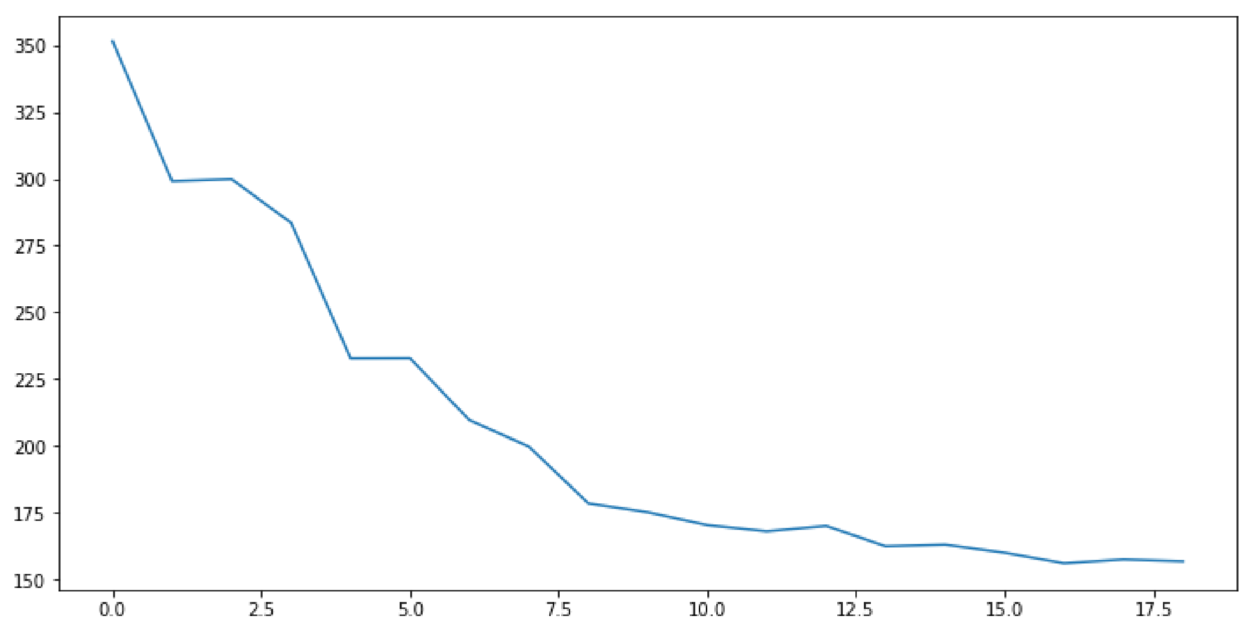
It is crucial to standardize our target values with the same category scalers. Target values will have the following values after the standardization.

```
'Arts & Entertainment': 0.1797971950648584,

'College & Education': -0.23181933066201452,

'Food': 0.701371111775053,

'Nightlife': -1.0660671961357695,
```

```
'Outdoors & Recreation': 1.2359420611099186,

'Professional': -0.740384028647716,

'Shops': 0.17276729197101026,

'Travel': -1.1206407611166134
```

4.3. Clustering

It is reasonable to cluster neighborhoods into similar groups. To do it, we will use the k-means clustering. This will allow us to have a set of possible options that could be reasonable for a particular family. Let's try to find an optimal K to do it.



It looks like we have an "elbow" point approximately near 8, so let's use K=8 to make our clusters. Let's take a look at them.

	Arts & Entertainment	College & Education	Food	Nightlife	Outdoors & Recreation	Professional	Shops	Travel
0	-0.119561	-0.231819	1.221494	-0.348522	-0.748369	-0.759911	1.246824	-0.397691
1	-0.267330	-0.167919	-0.152294	-0.369101	-0.126089	0.784807	0.036438	-0.223412
2	0.149130	-0.063660	-0.679602	-0.114355	1.247701	0.314957	-0.571043	0.754387
3	-0.125931	-0.167143	-1.110772	-0.199597	-0.534674	-1.044674	-1.248032	-0.233823
4	10.115958	-0.399978	-0.755605	-1.066067	0.759708	-2.305767	-1.732908	-0.118151
5	1.339016	-0.399978	0.065227	1.749959	2.029666	-0.613461	-1.659613	8.152393
6	0.014195	5.485590	-0.011726	-0.362061	-0.192762	-0.322596	-1.238166	-0.180806
7	0.142997	-0.157082	0.521244	1.886850	-0.166304	-0.566453	0.319358	-0.173845

We need to value our clusters, to do it, let's calculate a distance between a cluster center and our target values. We will need a helper function to calculate a distance.

	Arts & Entertainment	College & Education	Food	Nightlife	Outdoors & Recreation	Professional	Shops	Travel	distance
1	-0.267330	-0.167919	-0.152294	-0.369101	-0.126089	0.784807	0.036438	-0.223412	2.534453
0	-0.119561	-0.231819	1.221494	-0.348522	-0.748369	-0.759911	1.246824	-0.397691	2.547379
2	0.149130	-0.063660	-0.679602	-0.114355	1.247701	0.314957	-0.571043	0.754387	2.832830
3	-0.125931	-0.167143	-1.110772	-0.199597	-0.534674	-1.044674	-1.248032	-0.233823	3.188272
7	0.142997	-0.157082	0.521244	1.886850	-0.166304	-0.566453	0.319358	-0.173845	3.416659
6	0.014195	5.485590	-0.011726	-0.362061	-0.192762	-0.322596	-1.238166	-0.180806	6.229774
5	1.339016	-0.399978	0.065227	1.749959	2.029666	-0.613461	-1.659613	8.152393	9.984970
4	10.115958	-0.399978	-0.755605	-1.066067	0.759708	-2.305767	-1.732908	-0.118151	10.401546

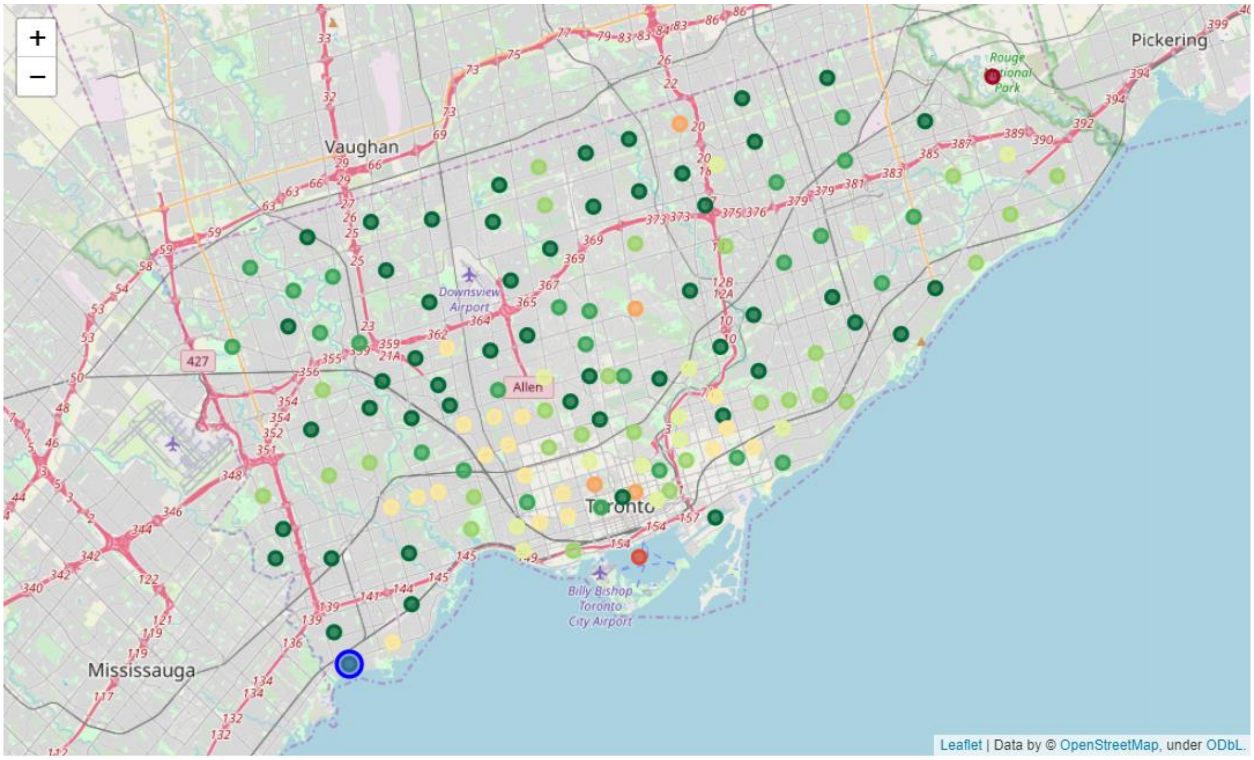
We can see that clusters 1 and 0 are both very close to the target values. So, let's find the best neighborhood directly by calculating the distance between the neighborhood and target values

	neighborhood	longitude	latitude	Arts & Entertainment	College & Education	Food	Nightlife	Outdoors & Recreation	Professional	Shops	Travel	cluster	distance
code													
19	Long Branch	-79.533345	43.592362	3	3	31	2	12	17	40	2	0	1.478682

5. Results

5.1. Recommendations

We can plot our recommendations on the map, differentiating neighborhoods by colors, where green=better and red=worse. Also, we will mark the best neighborhood with a blue circle as the best option.



5.2. Conclusions and Future Directions

Good news! We have a lot of matching neighborhoods for this family all around the city. It means that they have a lot of choices, and can consider other factors, like proximity of job or some other points of interest for the family.

We can see that the best neighborhood is at the border of the city. We can consider improving our model by adding coordinates of a point of interest, e.g., job, college, or even downtown coordinates to consider it's proximity as one of the features.