Choosing a location for a pet care center in Toronto

Capstone Project - IBM Data Science Professional Certificate

Valentina Tushkova
February 2019

Table of contents

Introduction **2**

Data **2**

Data Sources 2

Data Cleaning 3

Methodology **5**

Exploratory Data Analysis 5

Classification modeling 6

Results **9**

Discussion **10**

Conclusion **10**

Introduction

For many of us, our pets are family members. We try to provide them with the best possible care, including food, toys, walks, grooming, vet check-ups and much more. And if you live in a big city, you prefer to have all your pet supplies and facilities nearby.

Our hypothetical client, A Pet Company, is planning to open their first pet care center in Toronto, and they need help in choosing the right neighborhood for that. As a data scientist, I will conduct analysis using methods learned throughout the course and offer my insights and recommendations to the business owner.

As a starting point, the client's idea of the best location is this:

1. Many pets living in the neighborhood (i.e. prospective customers living within a walking distance).
2. Less or no competitor venues in the neighborhood.
3. Some park or dog run nearby to walk the pets.

Data

Data Sources

To complete this project, we will need these data sources:

1) **Toronto neighborhoods geodata**

   The following Wikipedia page was scraped using BeautifulSoup: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, in order to obtain the table of postal codes and to transform the data into a pandas dataframe.

The file with latitudes and longitudes of postal areas was kindly provided by the course instructor: https://cocl.us/Geospatial_data

2) **Data about pets living in Toronto, by neighborhood**

   In Canada, all pet owners must register and license their pets. That's good for us because this data is also made available by the government on the web:

   https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/community-services/#a666d03a-bafe-943a-e256-3c2d14b07b10

   The file contains information about the number of registered cats and dogs, and total (cats+dogs), grouped by Forward Sortation Area (FSA).

   In terms of this project, the FSA=Postal Code. These are essentially the same values but named differently by data sources.

3) **Data about existing pet venues**

   To get a list of existing pet venues in Toronto and to analyze their density by neighborhood, we will use the **Foursquare API.**

4) **Coordinates of Toronto** were obtained by geopy package (Nominatim).

5) **Geojson file of Toronto FSAs**

   Unfortunately, geojson file is not readily available from the Canadian Open Data Catalogue or any other public source. Statistics Canada does have Census Boundary Files for Canadian FSAs but in a shapefile format that has to be converted to GeoJSON format.

   https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2016-eng.cfm

   I have learned how to do this from this blogpost by A Gordon: https://medium.com/dataexplorations/generating-geojson-file-for-toronto-fsas-9b478a059f04

Data Cleaning

For **Toronto neighborhoods geodata**, the following data cleaning steps were taken:

- Scrape data from the web and transform the obtained list into a pandas dataframe.
- Remove cells with a borough that is Not assigned.
- If more than one neighborhood exists in one postal code area, combine them into one row with the neighborhoods separated with a comma.
- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

For **geospatial data** for Toronto, no cleaning was required, I simply appended the lat and lon values to the dataframe:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Morningside, Guildwood, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

For **pets data**, minimal work was required:

- Drop columns for Cats and Dogs, and keep only Total counts.
- Append the numbers of pets to the main dataframe.

Methodology

Exploratory Data Analysis

By simply sorting the dataframe by “Pets” column, the largest populations of pets are found in neighborhoods of East York, Etobicoke, and West Toronto. At the same time, downtown and central neighborhoods have almost none.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Pets
0	M4C	East York	Woodbine Heights	43.695344	-79.318389	1864
1	M8V	Etobicoke	Mimico South, New Toronto, Humber Bay Shores	43.605647	-79.501321	1848
2	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572	1828
3	M6P	West Toronto	High Park, The Junction South	43.661608	-79.464763	1809
4	M6H	West Toronto	Dovercourt Village, Dufferin	43.669005	-79.442259	1761
5	M6S	West Toronto	Swansea, Runnymede	43.651571	-79.484450	1693
6	M4J	East York	East Toronto	43.685347	-79.338106	1681
7	M4E	East Toronto	The Beaches	43.676357	-79.293031	1611
8	M2N	North York	Willowdale South	43.770120	-79.408493	1547
9	M6N	York	The Junction North, Runnymede	43.673185	-79.487262	1487

After querying the Foursquare API and wrangling the data, the following dataframe was created. It contains 4 main features that will be used for analysis and modeling:

- Pets: number of pets registered in the area
- NumberOfVenues: count of pet shops, pet services and vets within 700 m radius
- NearbyParks: count of dog runs and parks within 700 m
- VenuesPerPet: pet to venue ratio

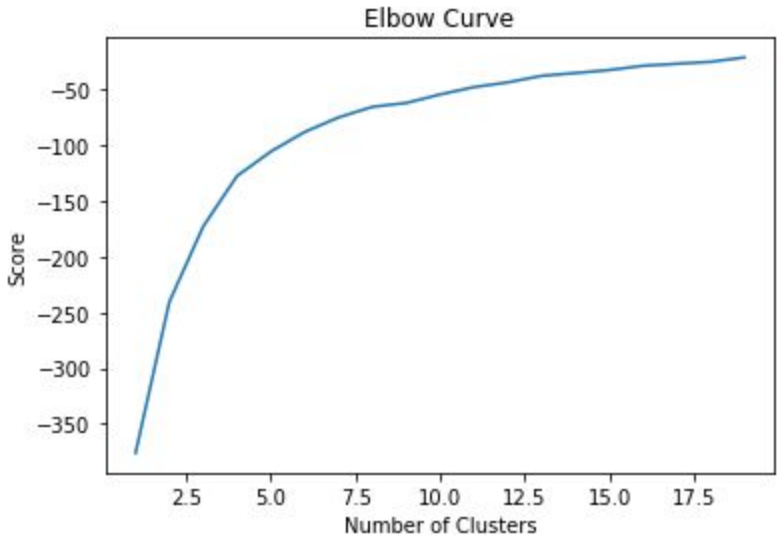
	PostalCode	Borough	Neighborhood	Latitude	Longitude	Pets	NumberOfVenues	NearbyParks	VenuesPerPet
0	M4C	East York	Woodbine Heights	43.695344	-79.318389	1864	1.0	0.0	0.000536
1	M8V	Etobicoke	Mimico South, New Toronto, Humber Bay Shores	43.605647	-79.501321	1848	2.0	0.0	0.001082
2	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572	1828	4.0	0.0	0.002188
3	M6P	West Toronto	High Park, The Junction South	43.661608	-79.464763	1809	3.0	1.0	0.001658
4	M6H	West Toronto	Dovercourt Village, Dufferin	43.669005	-79.442259	1761	2.0	4.0	0.001136

Classification modeling

First, the data had to be normalized with StandardScaler. The resulting dataframe is below. Now our data is ready for modeling.

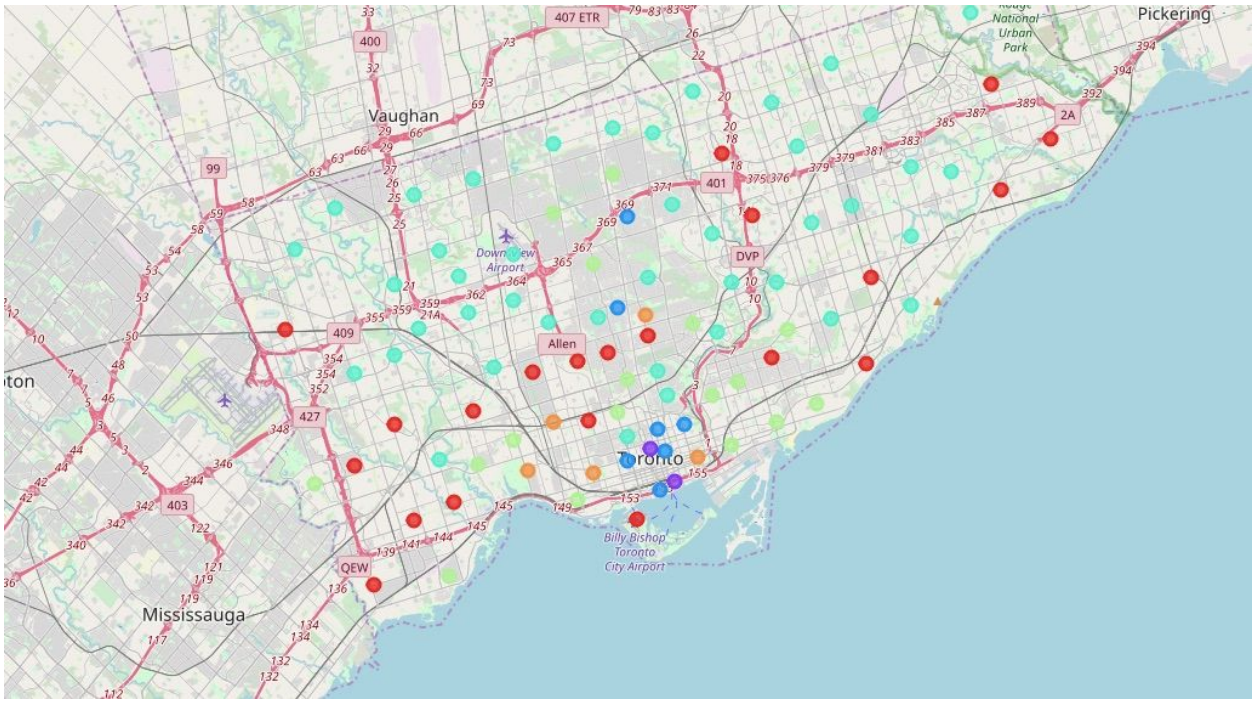
	Pets	NumberOfVenues	NearbyParks	VenuesPerPet
0	2.240520	-0.250314	-0.524265	-0.432929
1	2.204825	0.385620	-0.524265	-0.278035
2	2.160206	1.657488	-0.524265	0.035838
3	2.117818	1.021554	0.524265	-0.114526
4	2.010732	0.385620	3.669855	-0.262860

For clustering, I used k-means algorithm because it is relatively simple and fast, and can be run multiple times to optimize results. To determine the optimal value of *k,* visual elbow method was used.



The elbow is approximately at 6 clusters. So I moved forward with k=6 and trained the model.

Visualization of clusters on the map, created with Folium library.



Examining clusters and drawing insights

First, let's take a look at mean values of each cluster.

	Cluster Labels	Latitude	Longitude	Pets	NumberOfVenues	NearbyParks	VenuesPerPet
0	0	43.697682	-79.395005	1174.904762	0.714286	0.238095	0.000664
1	1	43.651362	-79.380345	180.000000	3.500000	2.000000	0.020184
2	2	43.679022	-79.388186	459.000000	3.714286	1.428571	0.008312
3	3	43.739974	-79.395566	550.404762	0.309524	0.142857	0.000607
4	4	43.682428	-79.402669	1357.588235	3.294118	0.294118	0.002612
5	5	43.666580	-79.413833	1274.200000	2.800000	3.400000	0.002284

Clusters 1, 2 and 3 don't look like good candidates, because they have very small average amount of pets per neighborhood. On top of that, clusters 1 and 2 already have one of the highest numbers of existing venues (i.e. our client's potential competitors). Therefore I am going to disregard them.

Cluster 0 has a good mean value of pets, plus relatively small numbers of existing venues. It has potential.

Clusters 4 and 5 have the biggest numbers of registered pets, but a high number of venues, too. Cluster 4 has the highest competition ("Venues per pet"), but not enough places to walk the dogs - only 0.29 per neighborhood on average.

Cluster 5 looks like a good candidate since it has lots of parks and a high number of registered pets, plus medium competition.

Now that we have narrowed down to clusters 0 and 5, we'll examine each of them separately.

Cluster 0

	Borough	Neighborhood	Pets	NumberOf Venues	NearbyParks	VenuesPerPet
22	Etobicoke	Alderwood, Long Branch	1215	0.0	2.0	0.000000
31	York	Humewood-Cedarvale	964	0.0	2.0	0.000000
21	Central Toronto	Davisville	1216	1.0	1.0	0.000822

Alderwood, Long Branch (Etobicoke) and Humewood-Cedarvale (in York) are good candidates for the business's location because they have lots of resident pets and zero competition nearby. Plus, each has 2 parks nearby to walk the dogs.

Cluster 5

	Borough	Neighborhood	Pets	NumberOf Venues	NearbyParks	VenuesPerPet
4	West Toronto	Dovercourt Village, Dufferin	1761	2.0	4.0	0.001136
25	West Toronto	Little Portugal, Trinity	1189	2.0	3.0	0.001682
43	Central Toronto	Davisville North	859	2.0	4.0	0.002328

13	Downtown Toronto	Harbourfront, Regent Park	1365	4.0	3.0	0.002930
24	West Toronto	Parkdale, Roncesvalles	1197	4.0	3.0	0.003342

Here, Dovercourt Village, Dufferin, Little Portugal, Trinity and Davisville North are the best candidates: they have less competition in the area and convenient location with a few parks nearby.

Results

Our analysis shows that not all the neighborhoods have large numbers of registered pets, especially in Central and Downtown Toronto there are very few pets. So it wouldn't make sense to open a business there. Furthermore, not all neighborhoods have a park or a dog run within a walking distance, and such neighborhoods are not suitable for our business model either.

After running a clustering algorithm on the data, I narrowed down the search to 2 potentially interesting clusters, and after filtering them by the number of competitors already existing in the area, I came up with a final list.

The following neighborhoods are recommended for our client's new business location:

1. Alderwood (Etobicoke)
2. Long Branch (Etobicoke)
3. Humewood-Cedarvale (York)
4. Dovercourt Village (West Toronto)
5. Dufferin (West Toronto)
6. Little Portugal (West Toronto)
7. Trinity (West Toronto)
8. Davisville North (Central Toronto)

Discussion

There are a few points to consider regarding the accuracy of the above results.

1. Accuracy and freshness of the pet data have room for improvement. The data set I used was dated 2017 and contained the number of pets registered that year. There is no telling how many cats and dogs live in Toronto unregistered by their owners, so potentially our client base can be much larger than estimated and even have a different distribution over neighborhoods.
2. For better model accuracy and more insights, other features can be included in future studies:
    - Population by neighborhood/pet per capita
    - Walk score of neighborhood
    - Foot traffic
    - Real estate prices, etc.

Conclusion

In this study, I analyzed the neighborhoods of Toronto and the facilities that it has for pets. I identified the number of pets per neighborhood, number of venues per neighborhood, the number of dog runs per neighborhood among the most important features. I built a classification model using k-means algorithm to cluster the neighborhoods.

By analyzing the clusters, I came to a conclusion about the list of the best prospective locations for the business owner to open a pet care center in Toronto.