Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

The Best Neighborhood for Dog Walking in Toronto

Table of contents

Introduction: Business Problem

A client wants to start a dog walking business in Toronto and has asked for advice about where it should be located. In this project we will try to find the optimal location for a dog walker to center their business. Dog walkers spend a lot of time picking up and dropping off dogs. Ideally we want to target an area with a large number of pets as well as a large number of places to walk them. That way we can minimize the amount of time the dog walker spends driving between dog walking clients, and the amount of time driving the dogs to their walking location.

To do this analysis, we will use the following information:

- The name and population of each neighborhood in Toronto.
- The number of pets in each neighborhood.
- The number of appropriate venues for walking dogs in each neighborhood (trails, parks and dog runs).

Using the above information, we will

- Look at the proportion of pets/population in each neighborhood.
- Cluster neighborhoods to find areas with a relatively high number of pet licenses and high number of dog related venues.

Data

The following data sources will be needed to extract/generate the required information:

- **Population by Postal Code from Census 2016:** https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/comprehensive.cfm
  - This information will be used to calculate the proportion of pet ownership in each neighborhood of Toronto.
  - We will use the 3 columns: FSA, Province and Population, 2016.
  - The FSA is the "forward sortation area" or first three digits of the postal code.
- **Toronto Open Data: Licensed Dogs and Cats Reports for 2013 through 2017:** https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/community-services/#a666d03a-bafe-943a-e256-3c2d14b07b10
  - This data has four columns: FSA / # Cat Licenses Issued / # Dog Licenses Issued / Total Licenses.
- **FourSquare Api:**

        - We will query to get the number of places to walk a dog per neighborhood in Toronto.
        - Specifically, this query will be filtered to look for only the following category Ids:
            - Dog Run 4bf58dd8d48988d1e5941735
            - Park 4bf58dd8d48988d163941735
            - Trail 4bf58dd8d48988d159941735
        - A sample query looks like this: https://api.foursquare.com/v2/venues/explore?client_id=&client_secret=E&v=X&ll=43.642960,-79.371613&radius=400&limit=100&categoryId=4bf58dd8d48988d1e5941735,4bf58dd8d48988d163941735,4bf58dd8d48988d159941735
- **geopy:**
    - We will use this to get coordinates for each postal code, for use when calling the Foursquare API.
- **Wikipedia:** https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
    - We will use this to get the neighborhood name for each FSA.

Gathering and Cleaning Data

Toronto Neighborhood Data

First we found the latitude & longitude coordinates for each neighborhood in Toronto. We did this by first scraping the neighborhood FSA, or postal code, data from Wikipedia into a pandas dataframe. Then we loaded the coordinates for each postal code from geopy into another dataframe. Finally, we merged the two dataframes into a single dataframe which contained the coordinates, neighborhood name and postal code for each neighborhood in Toronto. This is the first 5 rows of the resulting dataframe:

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

Next we loaded the population data from the 2016 census and added that to our above dataframe, so we had the population for each neighborhood. Here are the first 5 rows after bringing in the population data:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Pop
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353	66108.0
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	35626.0
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711	46943.0
3	M1G	Scarborough	Woburn	43.770992	-79.216917	29690.0
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	24383.0

Next we loaded and cleaned the pet licensing data from 2017 and added that to our dataframe as well.

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Pop	DOG
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353	66108.0	627
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	35626.0	775
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711	46943.0	963
3	M1G	Scarborough	Woburn	43.770992	-79.216917	29690.0	385
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	24383.0	309

There were a few outlier neighborhoods with very low population (0-15 people), so we dropped those rows.

Finally we called Foursquare for each neighborhood to get the list of parks, trails and dog runs in each neighborhood, and merged that into our dataframe.

	Postcode	Borough	Neighborhood	Latitude	Longitude	Pop	DOG	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M5H	Downtown Toronto	Adelaide,King,Richmond	43.650571	-79.384568	2005.0	38	NaN	NaN	NaN	NaN	NaN	NaN
1	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	2951.0	71	43.651494	-79.375418	St. James Park	43.650425	-79.372311	Park
2	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	2951.0	71	43.651494	-79.375418	Berczy Park	43.648048	-79.375172	Park
3	M3K	North York	CFB Toronto,Downsview East	43.737473	-79.464763	5997.0	122	43.737473	-79.464763	Ancaster Park	43.734706	-79.464777	Park
4	M2P	North York	York Mills West	43.752758	-79.400049	7843.0	259	43.752758	-79.400049	Tournament Park	43.751257	-79.399717	Park

Methodology

In this project we are locating areas of Toronto that have high dog population and large numbers of places to walk them.

In the first step we collected the required data: **dog walking venues, and pet and human populations in each Toronto neighborhood**.

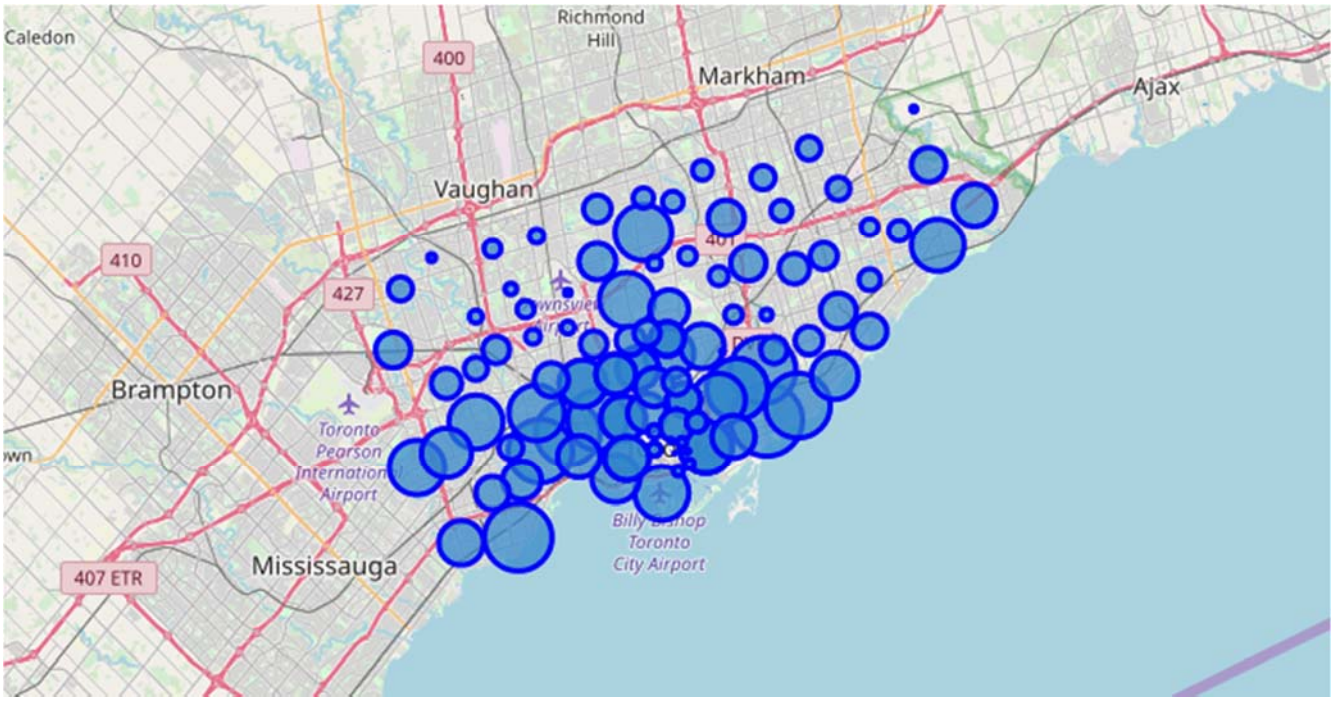
The second step in our analysis will be the calculation and exploration of '**dog population density**' and '**dog walking venue density**' across different areas of Toronto.

We calculated a few handy metrics and added them as new columns to our dataframe:

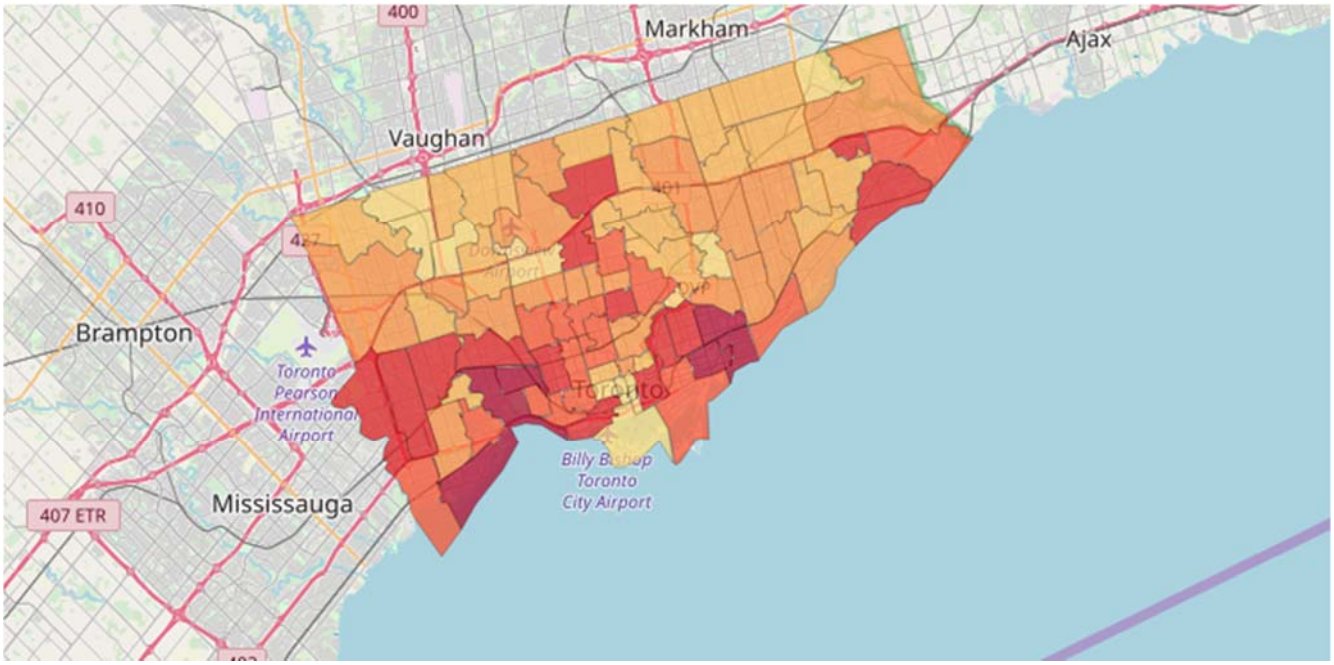
- PropTotal: proportion of dog liceneses granted in Toronto in 2017 that were granted to this FSA
- PerCapitaTotal: number of dogs registered in FSA per person

We found that the highest pet ownership rates are in East Toronto (M4L) and Etobicoke (M8V), and the lowest pet ownership rates are in Downtown toronto (M5H) and East York (M4H).

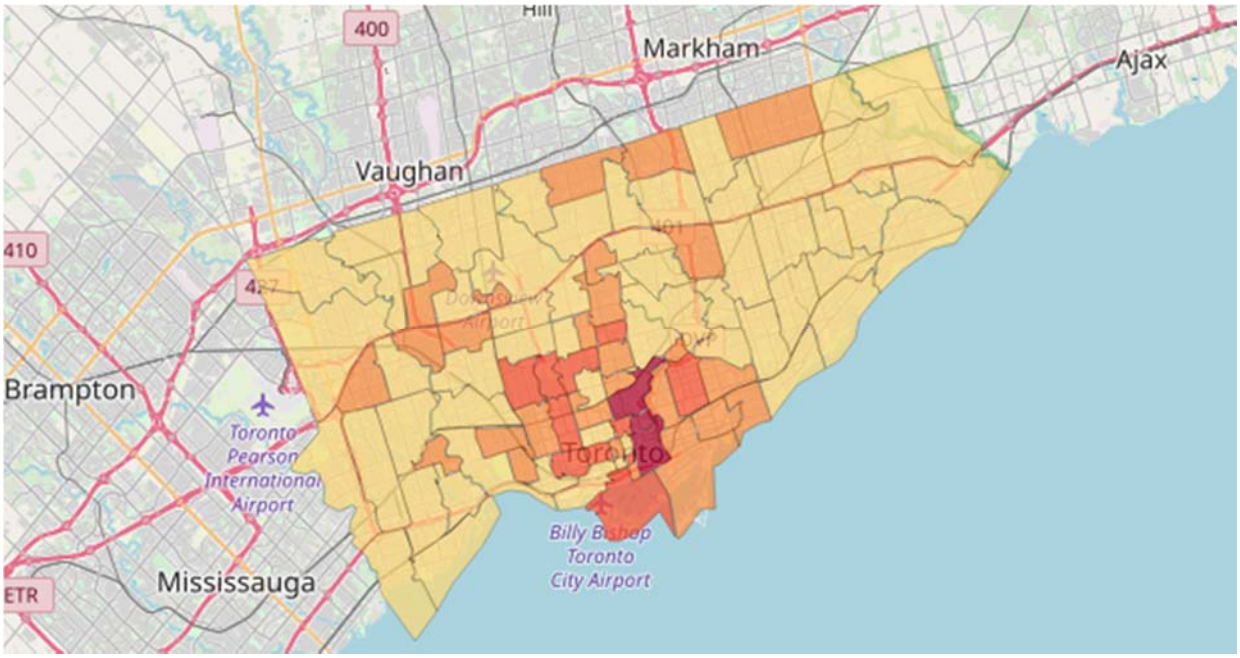
We then made a map with the dog ownership rates. Note that the size of the marker indicates the proportion of registered dogs in that area.



Then we created a choropleth map of the dogs in Toronto. The more red areas have more dogs.



Next we plotted a map of dog walking venues. The darker red areas have more parks, trails and dog runs.



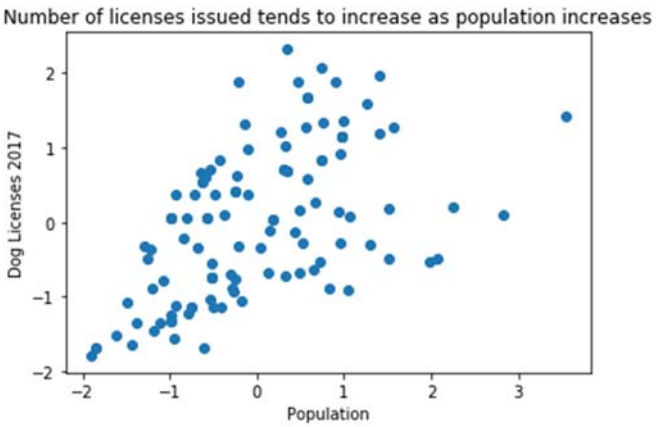
Neighborhood Clustering¶

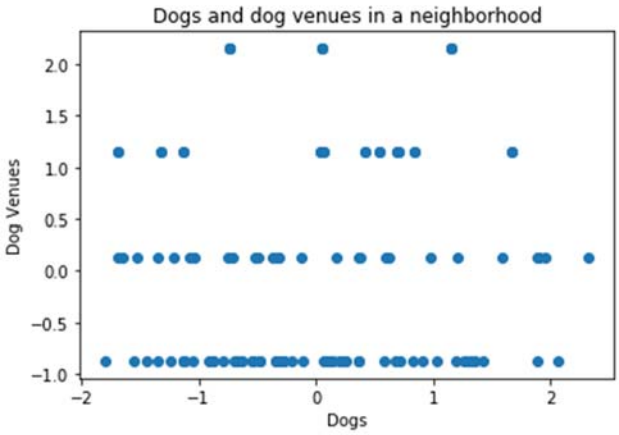
Next we used k-means clustering to cluster our dog population and venues. Since the population and counts are on different scales, they should be scaled before doing any analysis. We used StandardScaler from SciKitlearn.

We clustered based on:

- Population
- Total Dogs in 2017
- Number of existing venues

Exploring the data, we made a few scatter plots:





In [46]:

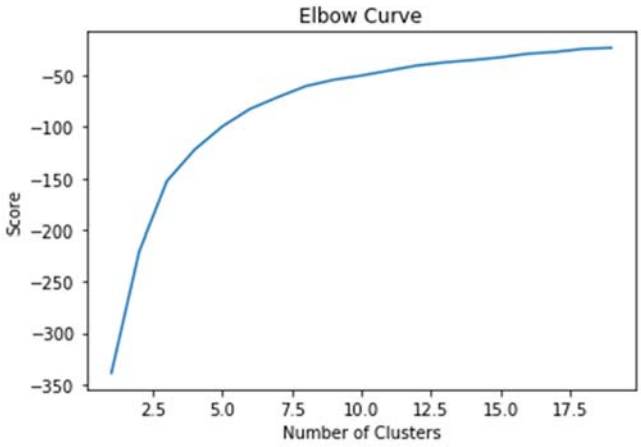
In the first plot we can see that as the human population increases, the number of dog licenses does as well. However, when plotting the number of dogs against the number of dog walking venues, the relationship is not as clear. There are neighborhoods with a larger number of dogs and a small number of venues to walk them. We will avoid these neighborhoods for our business.

We then ran the .corr function on our training set to see the correlation between each data set.

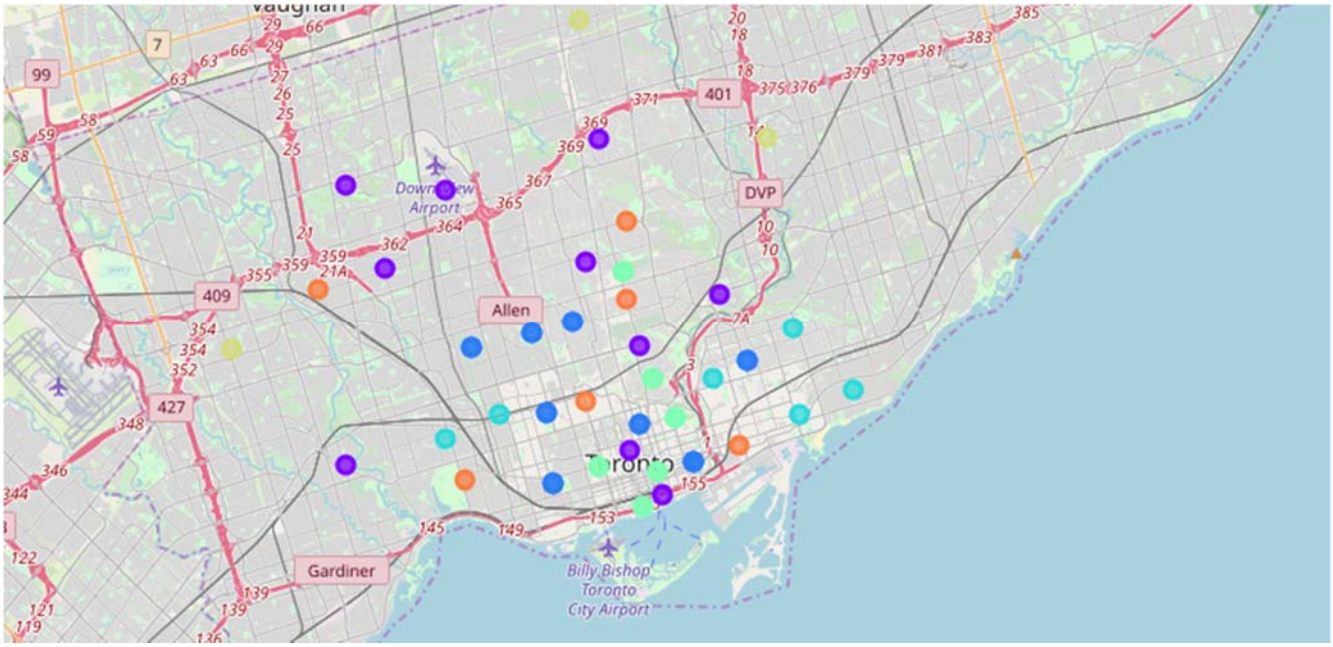
	Pop	DOG	NumVenues
Pop	1.000000	0.533844	-0.206125
DOG	0.533844	1.000000	0.073937
NumVenues	-0.206125	0.073937	1.000000

We can see in the above chart that our interpretation of the scatter plots was correct- the population to dog correlation is positive and moderate (.53) and the correlation between dogs and number of places to walk them is very weak (.07). Also notice there is a weak negative correlation between the population and the number of parks/trails/dog runs in a neighborhood (-.2).

Now we needed to determine how many clusters we should divide the neighbourhoods into, so we used the visual Elbow method. We clustered using 1 to 20 clusters and get the scores (the sum of distances of samples to their closest cluster center) in a list of scores to plot.



It looks like the elbow is around 7 clusters, so we used that value and re-ran KMeans to get our Clusters. Then we plotted the clusters on a map.

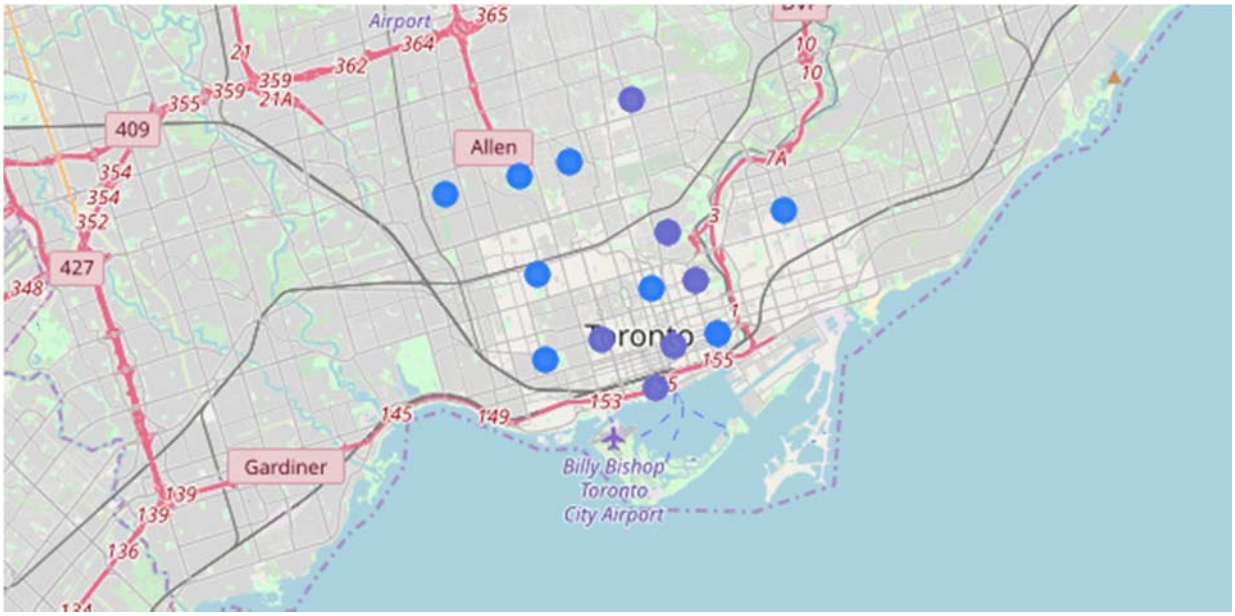


Looking at the mean values for each cluster, we get this chart:

	Cluster Labels	DOG	NumVenues	Pop
4	4	368.142857	2.428571	15479.642857
2	2	834.882353	2.176471	32312.588235
0	0	433.000000	1.000000	54680.000000
1	1	270.272727	1.000000	13287.727273
3	3	1154.666667	1.000000	36853.000000
5	5	525.333333	1.000000	33559.333333
6	6	729.666667	1.000000	22992.000000

We can see that clusters 4 and 2 have the largest number of dog walking venues, but cluster 2 has twice as many dogs on average as cluster 4.

Let's see our clusters of interest on a map.



Results

Dog Ownership in Toronto 2017:

- Highest pet ownership rates are in: East Toronto (M4L) and Etobicoke (M8V)
- Lowest pet ownership rates are in: Downtown toronto (M5H) and East York (M4H)

Dog Walking Venues

- The largest number of parks, trails and dog runs in any neighborhood is 3. The following neighborhoods have 3 venues for walking a dog:
  - Rosedale (M4W)
  - Harbourfront,Regent Park (M5A)
  - Cabbagetown,St. James Town (M4X)

Candidates for dog walking location

We identified 2 clusters that appear like good candidates for a dog walking business.

- these areas have
    - High pet ownership
    - High number of places to walk a dog
- Within these clusters, 2 FSAs jump out as particularly ideal for our purposes
    - M5A (Harbourfront,Regent Park) with 949 dogs and 3 walking venues
    - M4J (East Toronto) with 1110 dogs and 2 walking venues

Discussion

Caveats

Looking at the map, East Toronto (M4J) is relatively isolated whereas Harbourfront,Regent Park (M5A) has other neighborhoods in clusters 2 and 4 nearby. Next time we could add latitude and longitude values into the clustering to look for neighborhoods that are close to each other.

Not all parks allow dogs. I was not able to find data on which Toronto parks allow dogs. This information would be helpful for getting a more accurate estimate of dog walking venues.

I'm using newly issued licenses to estimate dog ownership in a neighborhood, but this may not be a perfect measure since 1) not every owner registers their dog and 2) dogs in certain neighborhoods may be registered at lower rates due to income or some other factor.

Recommendations

My recommendation would be to target Harbourfront,Regent Park (M5A) when looking for a location to open a dog walking business.

This area has

* A large number of registered dogs

* A high number of venues nearby for walking dogs

* Proximity to other neighborhoods with these qualities

Conclusion

In this project, I wanted to identify areas in Toronto that might be good candidates for centering a dog walking business. Data was collected from a number of sources including open data portals, API calls and website scraping. This data provided a picture about the various Toronto neighbourhoods, including population, number of new dog licenses issued and number of venues for walking a dog. Based on this data, I used the KMeans algorithm to cluster and identify areas that had high population, a high number of dog licenses and a high number of locations to walk a dog.

This analysis suggested 2 clusters of locations, and within these I would recommend focusing on M5A, the Harbourfront,Regent Park area of Toronto.