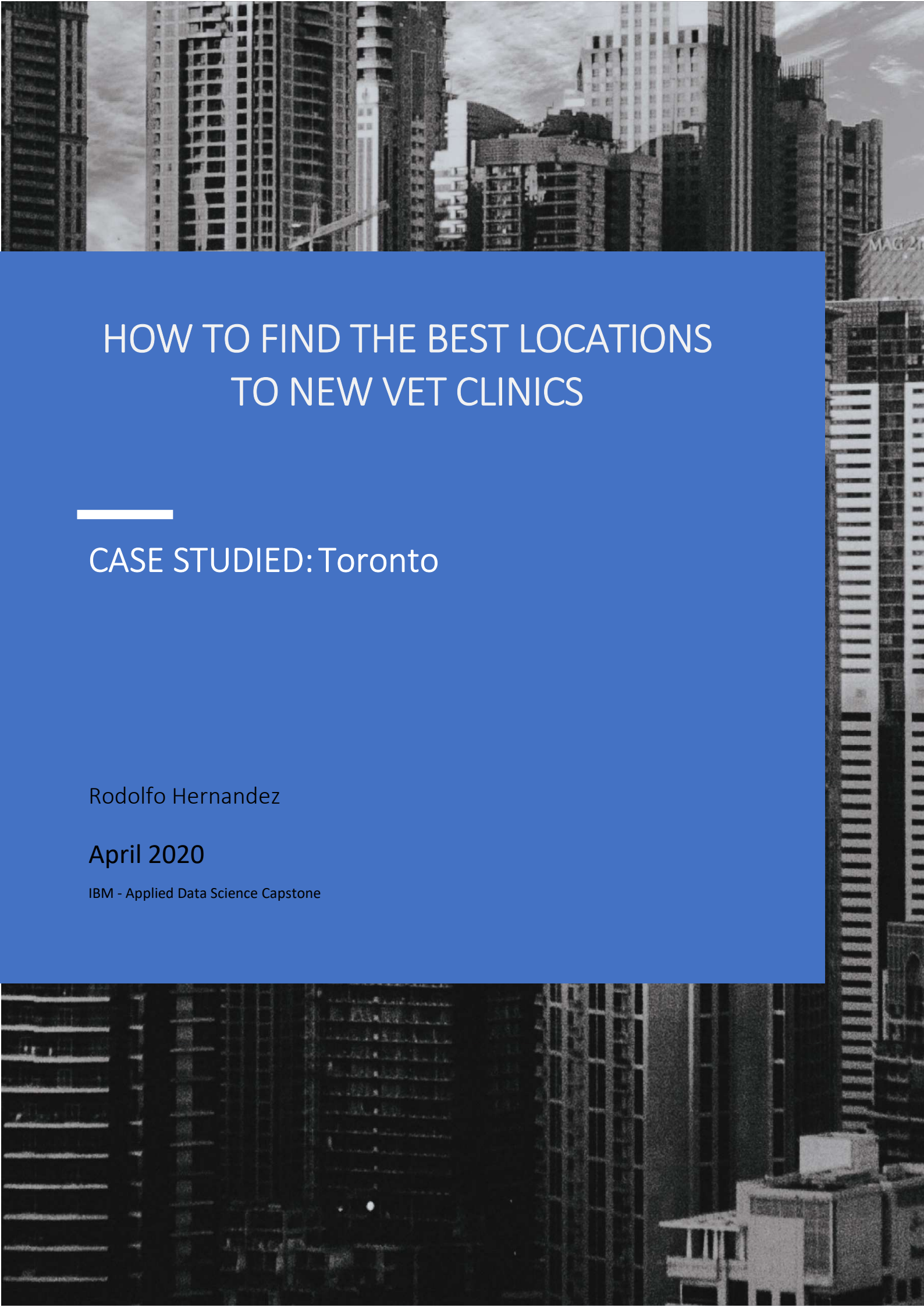

HOW TO FIND THE BEST LOCATIONS TO NEW VET CLINICS

CASE STUDIED: Toronto

Rodolfo Hernandez

April 2020

IBM - Applied Data Science Capstone

1. Introduction

According to recent statistics (2017[1]), over 57% of the Canadian homes had pets (mainly dogs and cats) and from 2016 to 2018 the population of dogs had increased while the cat population remains stable. In this context a Company specialized in pets care is searching for locations to new vet clinics and considers that Toronto could be a good option because of its population and the total number of pets (there are over 230,000 dogs in the city according to Toronto City's webpage).

Although the core business of the Company it's related to vet services, they are interested to sell basic complements for pets in the clinics (like food, toys, collars, leashes, etc.) as a secondary activity.

To search the best locations we can take into account the following factors:

1)The number of potential clients in each area. It's the key factor to ensure the viability of the business. We need accurate information about the number of pets in each neighborhood.

2) The distance to existing vet clinics and pet stores. It's important to identify where are situated existing businesses of the vet care sector in Toronto and surroundings. Although our core sector is vet clinics, the pet stores should be considered too.

3)The visibility of clinics is an important factor too, so we can analyze what kind of neighborhoods are the most suitable for new clinics (for example, locations closer to parks or dog runs could be more interesting than others). In this case, clustering analysis of the neighborhoods can be our starting point.

It's possible to consider additional features (socioeconomic factors, etc.) although for our specific analysis the former factors explained will be a good reference.

[1] Extracted from: Migiro, Geoffrey . "How Much Do Canadians Spend On Their Pets?" WorldAtlas, Jan. 7, 2020, worldatlas.com/articles/how-much-do-canadians-spend-on-their-pets.html.

2. Data Description

According to the basic ideas exposed above, we are going to use the following sources of data:

1)The number of potential clients. From the Toronto Open Data repository[2] we can download a dataset which contains the number of licensed dogs and cats in each Toronto's neighborhood[3].

https://www.toronto.ca/community-people/animals-pets/pets-in-the-city/dogs-in-the-city/

This information can be used combined with the population of the neighborhood (we are going to estimate the number of pets/number of inhabitants ratios to identify suitable areas to locate our clinics).

2)Distance to existing vet clinics and pet stores. Our first source of data is FOURSQUARE, we can query for venues related to the pet care sector (pet shops and vet clinics). Besides, from the Toronto City's webpage[4], we can access to list of vet clinics using scraping (also we can get the names and the coordinates). Finally, we can complete the clinics' information using Google services (for instance, we can get .kml files from Google Earth with location data). This additional source can be useful to find data in the Toronto surroundings (in the limits of the city we should consider venues located in the following areas: Mississauga, Brampton, Vaughan, Markham, Ajax and Pickering).

3)Other data. We need additional information related to:

-Neighborhoods' basic data. We'll use several dataframes obtained previously (in the previous assignment) that can be useful again such as Toronto's neighborhoods basic data (name, postal code, coordinates), neighborhoods' clustering dataframe (it's an interesting aspect to analyze) and Toronto venues dataframe obtained from FOURSQUARE (which contains data related to venues in neighborhoods).

-Neighborhoods' population: from the "Statistics Canada" website we can download datasets (in .csv or .tab format) with population and other demographic facts organized by geographical areas[5]. Too, in the Toronto Open Data repository we can find additional demographic data of the city[6].

[2] Source: https://www.toronto.ca/city-government/data-research-maps/open-data/
[3] Source: https://open.toronto.ca/dataset/licensed-dogs-and-cats-reports/
[4] Source: https://www.toronto.ca/community-people/animals-pets/pet-licensing/bluepaw-partners/vets-hospitals/
[5] Source: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/comprehensive.cfm
[6] Source: https://open.toronto.ca/dataset/neighbourhood-profiles/

3. Methodology

3.1. Data acquisition & cleaning

Vet clinics and pet shops

First, we are going to search for venues which are included in our sector (pet care and shops). Searching for 'Venue Category' in FOURSQUARE we can identify just the category 'Pet Stores' as related terms to our seek with a total of ten results, without results for 'vet clinics' and similar (although some venues classified as 'Pet stores' are clinics too), so we have to use other kinds of sources to complete the data we need.

Although pet stores aren't our core activity (we are more interested in vet clinics sector) we are going to search for new pet stores in the surroundings of Toronto (Mississauga, Brampton, Vaughan, Markham, Ajax and Pickering) focusing on our Foursquare's seeking in these places (especially in the areas closer to the Toronto's limits) because Toronto's clients can access to these areas too.

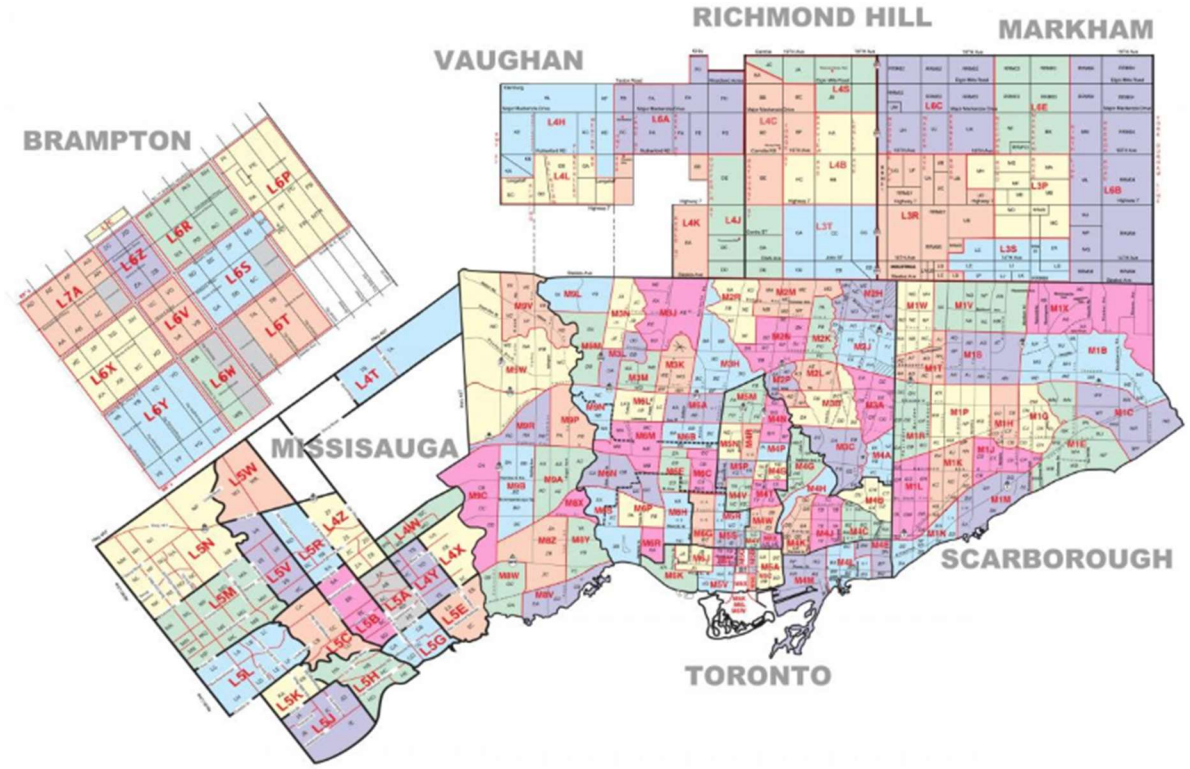

FIGURE 1: TORONTO AND SURROUNDINGS' MAP[7]

[7] Source: https://maps-mississauga.com/mississauga-postal-code-map

For this, we define a list with Toronto's neighborhoods situated in the limits (we use postal codes to identify them):

```
surroundings=['M8W','M9C','M9W','M9L','M3N','M3J','M2R','M2M','M2H','M1W','M1V','M1X','M1B','M1C','M1E']
```

If we repeat a 'Category Venue' query (searching for 'Pet stores' venues) in the selected neighborhoods with a radius limit of 1000 m, FOURSQUARE doesn't give back additional results.

Finally, our dataframe with pet stores information has the following structure (represented the head of the dataframe):

	Name	Latitude	Longitude
0	Big Al's Pet Supercentre	43.759279	-79.278325
1	PetSmart	43.769139	-79.412522
2	East York Animal Clinic	43.705921	-79.312196
3	PetSmart	43.712682	-79.362636
4	Pet Valu	43.666979	-79.314665

Vet clinics

According to our business approach vet clinics are our core business (pet stores are just a secondary activity) and the number of existing clinics obtained from FOURSQUARE it's too low. To get a suitable dataset of existing venues, we've used two kinds of additional sources:

1) For Toronto area is available a list of vet clinics from Toronto Open Data Service.

2) For the surroundings (Mississauga, Brampton, Vaughan, Markham, Ajax and Pickering) we can use other sources such as Google Earth service downloading a .kml file which contains coordinates and the name of the vet clinics.

For vet clinics we are going to consider the whole surrounding areas in our seek (no limited to a radius) because:

-The primary existing venues to analyze are vet clinics (our core business), so we have to use the most complete dataset.

-Vet services are expensive so it's likely that customers can search in a wider range of clinics (also in locations situated outside of Toronto).

The Toronto web page contains some valuable data about the vet clinics in Toronto. However, the data isn't available as a file, so we've extracted it using web scraping techniques. We can observe the structure of the webpage in the following image:

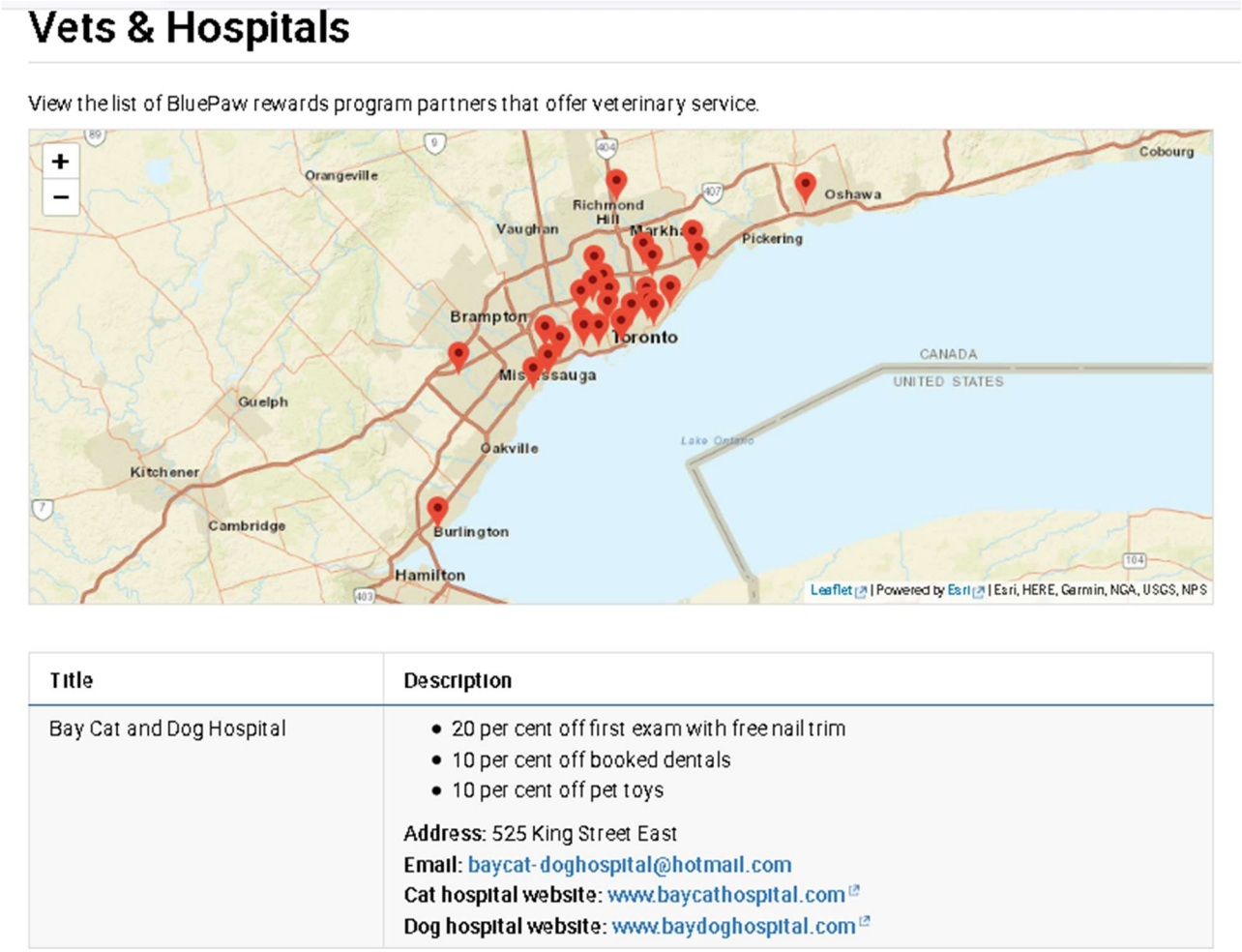

FIGURE 2: SCREENSHOT FROM TORONTO'S VETERINARY SERVICE REPOSITORY.

From the webpage we obtain the following data:

-Name of the vet clinic.

-Coordinates (latitude, longitude).

This web stores information about clinics situated in Toronto mainly. Now, we are going to extend our searching limits to closer locations (see the general map in page nº4) such as:

-Pickering.

-Vaughan.

-Mississauga.

-Brampton.

-Markham.

-Ajax.

For each location we've executed a seek using Google Earth and downloading a .kml file with the names and coordinates of related venues. Using the Geopandas library we can the data we need (clinics' names and coordinates). From all sources of data, we've obtained a total of 88 venues although some values are duplicated (it's possible because we've combined several sources of data). So, we've eliminated duplicated values obtaining a total of 66 clinics. In the following map we can see their geographical distribution:

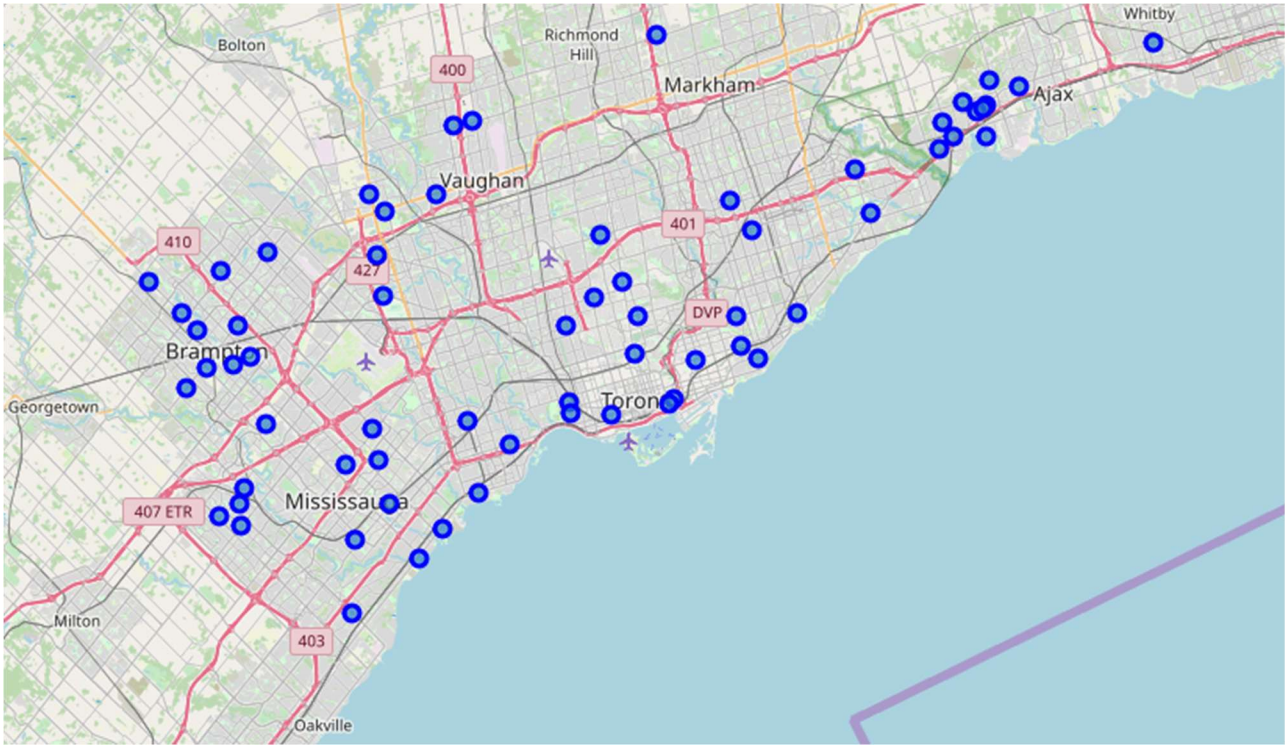

FIGURE 3: LOCATION OF VET CLINICS OBTAINED FROM ALL DATA SOURCES COMBINED

The structure of the vet clinics' dataframe (called df_clinics) is the following (represented the head of the dataframe):

	Name	Longitude	Latitude
0	Bay Cat and Dog Hospital	-79.35861190000003	43.6552217
1	Bay Cities Animal Hospital (VCA Canada Hospital)	-79.78532059999998	43.33892729999999
2	Birchmount Animal Hospital (VCA Canada Hospital)	-79.29011489999999	43.76167840000001
3	Bridletowne Warden Animal Hospital	-79.30896389999998	43.7810853
4	Cachet Village Animal Hospital (VCA Hospital)	-79.3738907	43.8860006

FIGURE 4: HEAD OF CLINICS' DATAFRAME

3.2. Modelling

We'll consider the following steps to set up our model:

3.2.1. Clustering analysis

One of the aspects we can analyze is the influence of the neighborhoods' types in the election of best locations for our clinics. In this sense, the approach based on performing a clustering analysis using basic features of Toronto's neighborhoods (mainly using venues queries from FOURSQUARE) it's a great starting point.

The basic procedure consists of:

a) First, using FOURSQUARE queries for all neighborhoods, we can identify what types of venues are most frequent in each of them (we need to select just the most frequent):

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adelaide, King, Richmond	Coffee Shop	Restaurant	Café	Bar	Bakery	Seafood Restaurant	Cosmetics Shop	Lounge	Steakhouse	Clothing Store
1	Agincourt	Lounge	Skating Rink	Latin American Restaurant	Breakfast Spot	Clothing Store	Dumpling Restaurant	Distribution Center	Dog Run	Doner Restaurant	Donut Shop
2	Agincourt North, L'Amoreaux East, Milliken, St...	Park	Playground	Women's Store	Donut Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant
3	Albion Gardens, Beaumond Heights, Humbergate, ...	Grocery Store	Liquor Store	Sandwich Place	Fast Food Restaurant	Beer Store	Fried Chicken Joint	Pharmacy	Pizza Place	Comic Shop	Concert Hall
4	Alderwood, Long Branch	Pizza Place	Gym	Sandwich Place	Skating Rink	Pharmacy	Coffee Shop	Pub	Dim Sum Restaurant	Diner	Discount Store

FIGURE 5: HEAD OF TORONTO'S VENUES DATAFRAME

b) Secondly, we perform a cluster analysis by a K-means approach setting a number of clusters (K) equals 3. We can represent graphically on Toronto's map using Folium library:

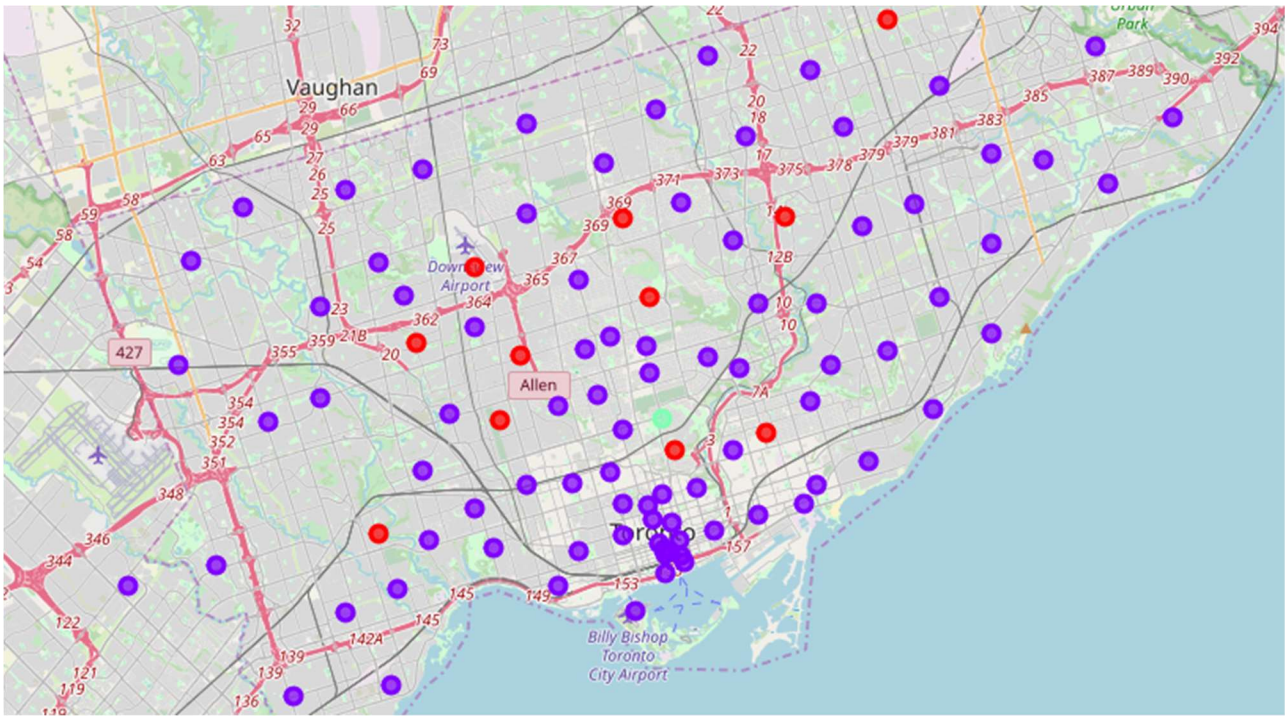

FIGURE 6: LOCATION OF TORONTO'S CLUSTERS ACCORDING TO MOST FREQUENT VENUES

In the biggest cluster (represented by blue dots, with a total of 86 neighborhoods) the most frequent venues are related to coffee shops, restaurants, etc.):

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
count	86	86.0	86	86	86	86	86	86	86	86	86	86
unique	10	NaN	45	58	50	55	48	49	48	49	51	45
top	Downtown Toronto	NaN	Coffee Shop	Coffee Shop	Coffee Shop	Coffee Shop	Diner	Distribution Center	Distribution Center	Distribution Center	Doner Restaurant	Doner Restaurant
freq	18	NaN	18	7	7	5	8	6	9	9	9	8

FIGURE 7: STATISTICS FROM TORONTO'S CLUSTER Nº1 DATAFRAME

However, in the following cluster classified (labeled as nº2, in the map are represented by red dots with a total of 11 neighborhoods) if we analyzed the top venues, we detect parks, playgrounds and dogs runs between the most common venues:

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
count	11	11.0	11	11	11	11	11	11	11	11	11	11
unique	7	NaN	1	10	9	7	6	4	4	4	4	3
top	North York	NaN	Park	Playground	Women's Store	Women's Store	Dim Sum Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop
freq	5	NaN	11	2	2	3	3	5	5	5	5	7

FIGURE 8: STATISTICS FROM TORONTO'S CLUSTER Nº2 DATAFRAME

According to our 'visibility criteria' (places used more frequently by pets' owners), these neighborhoods are very interesting for us, so we'll study carefully them later with the rest of the factors.

3.2.2. Descriptive statistics

From our cluster classification defined above, we are going to set up some metrics to evaluate:

-The number of potential clients in each area. First, using data about the number of pets in each neighborhood and population we can observe that the number of dogs in Toronto is higher (more than twice) than the number of cats:

	Postal Code	Neighborhood	Population, 2016	CAT	DOG	Total
count	97	97	97.000000	97.000000	97.000000	97.00000
unique	97	97	NaN	NaN	NaN	NaN
top	M6R	Woodbine Gardens, Parkview Hill	NaN	NaN	NaN	NaN
freq	1	1	NaN	NaN	NaN	NaN
mean	NaN	NaN	28166.072165	252.525773	582.051546	834.57732
std	NaN	NaN	14230.972172	155.503958	319.776757	465.58773
min	NaN	NaN	15.000000	6.000000	0.000000	6.00000
25%	NaN	NaN	18241.000000	138.000000	322.000000	495.00000
50%	NaN	NaN	25473.000000	237.000000	556.000000	798.00000
75%	NaN	NaN	37769.000000	353.000000	807.000000	1197.00000
max	NaN	NaN	75897.000000	676.000000	1315.000000	1864.00000

FIGURE 9: BASIC STATISTICS PETS POPULATION DATAFRAME

Too, we can estimate the number of pets per inhabitant. This is a simple but effective metric to identify interesting areas for vet care businesses:

Per each neighborhood:

$$pets_{ratio} = \frac{Total\ number\ of\ pets * 1000}{Total\ number\ of\ inhabitants}$$

$$dogs_{ratio} = \frac{Total\ number\ of\ dogs * 1000}{Total\ number\ of\ inhabitants}$$

$$cats_{ratio} = \frac{Total\ number\ of\ cats * 1000}{Total\ number\ of\ inhabitants}$$

If we calculate ratios in all neighborhoods and sort by the highests pets_ratios the results are (represented the head of the dataframe):

Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Population, 2016	CAT	DOG	Total	dogs_ratio	cats_ratio	pets_ratio
M4E	East Toronto	The Beaches	43.676357	-79.293031	1	25044.0	430	1181	1611	47.157004	17.169781	64.326785
M6R	West Toronto	Parkdale, Roncesvalles	43.648960	-79.456325	1	19857.0	417	780	1197	39.280858	21.000151	60.281009
M8W	Etobicoke	Alderwood, Long Branch	43.602414	-79.543484	1	20674.0	400	815	1215	39.421496	19.347973	58.769469
M4T	Central Toronto	Moore Park, Summerhill East	43.689574	-79.383160	2	10463.0	114	493	607	47.118417	10.895537	58.013954
M4R	Central Toronto	North Toronto West	43.715383	-79.405678	1	11394.0	170	481	651	42.215201	14.920133	57.135334

FIGURE 10: HEAD OF RATIOS' DATAFRAME SORTED BY HIGHEST PETS_RATIO

Using Folium we can check where are situated the neighborhoods with the highest ratios (previously, we've downloaded a geojson file with neighborhoods' limits[8]). We have to take into account that in some neighborhoods there isn't data available so in the maps they are represented using a dark color so we can identify them clearly:

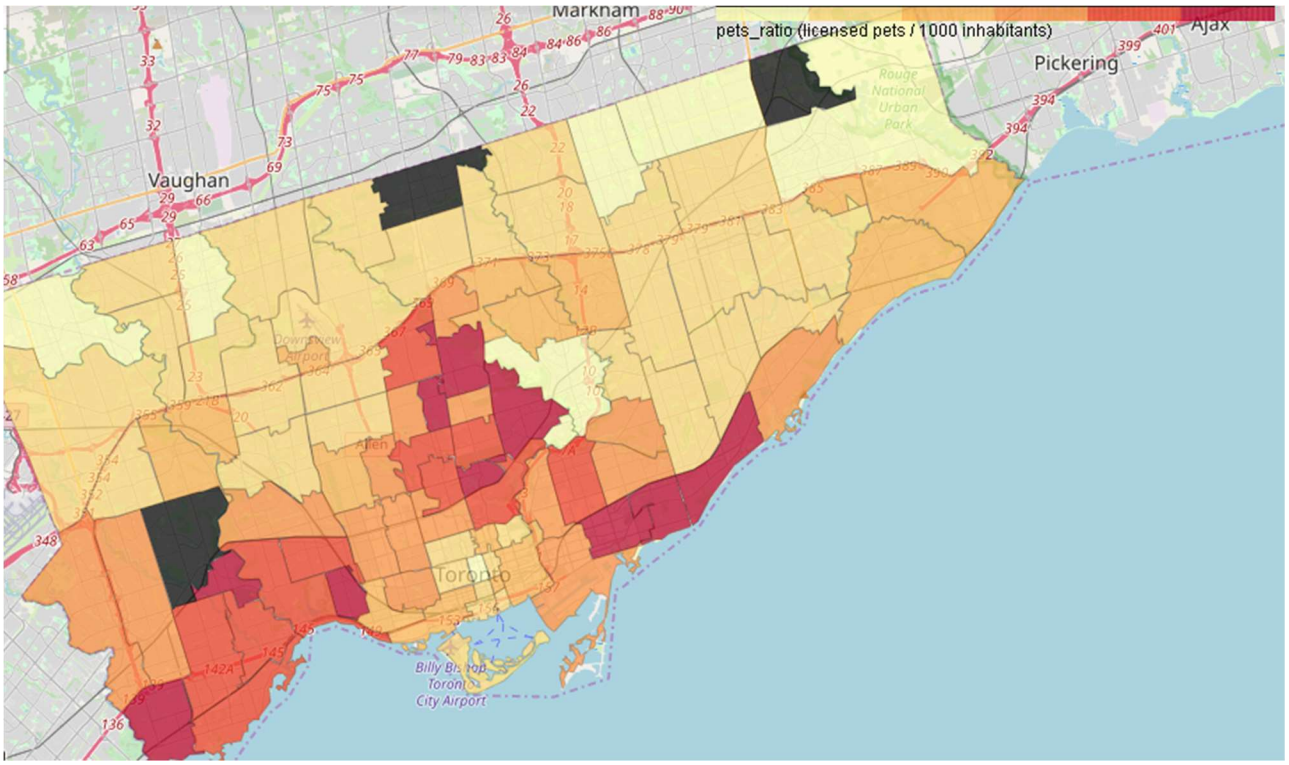

FIGURE 11: PETS_RATIO CHOROPLETH MAP.

[8] Source: https://www12.statcan.gc.ca/census-recensement/alternative_alternatif.cfm?l=eng&dispext=zip&teng=lfsa000b16a_e.zip&k=%20%20%20%2044221&loc=http://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/files-fichiers/2016/lfsa000b16a_e.zip

Besides, it's more interesting to analyze the relationship between clusters and ratios using graphs such as box-plots. For example, if we represent cluster labels vs. pets-ratio:

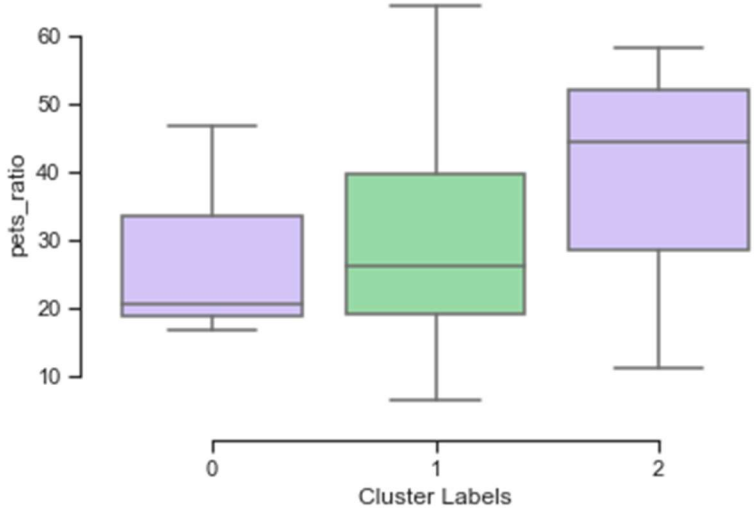

FIGURE 12: BOX-PLOT (CLUSTER LABELS VS. PETS_RATIO)

This graph is quite interesting: in neighborhoods included in cluster nº2 (residential areas with parks, dog runs, etc. in the top venues positions), the pets_general ratio is higher than other clusters.

If we repeat it using just the dogs_ratio instead of pets_ratio:

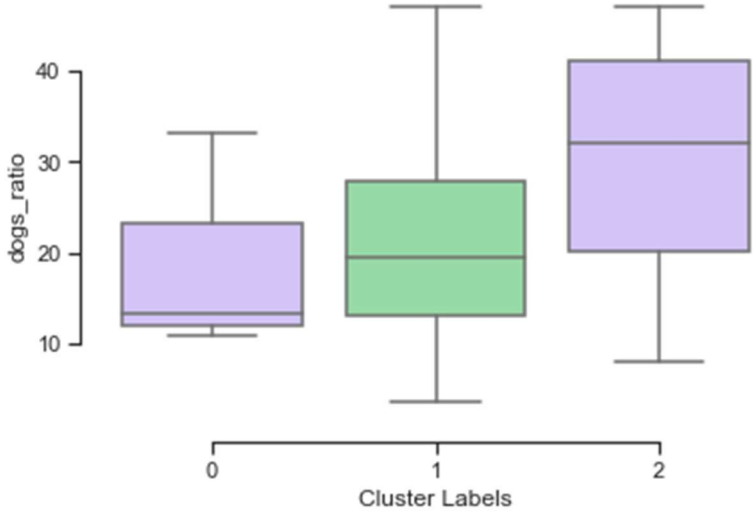

FIGURE 13: BOX-PLOT (CLUSTER LABELS VS. DOGS_RATIO)

We can observe that the gap between clusters if we choose the dogs_ratio is bigger than general_pets_ratio. This could be interesting for our analysis because the dog's vet spending is higher than cats (according to statistics[9]).

So, the neighborhoods included in the cluster nº2 should be carefully analyzed to our business.

- The distance to existing vet clinics and pet stores from potential locations. Now, we are going to include another variable to complete our analysis: our competitors (the existing vet clinics and pet stores). For this, we are going to estimate the distance between existing vet clinics and pet stores and each neighborhood using the Geopy library. For this, we'll use the coordinates data available for neighborhoods/clusters and existing pet care venues.

The process consists of the following steps:

-First, we'll calculate the distance between the clusters' position and existing vet clinics. We are going to consider just the geographical distance (it's a simpler approach than the 'real' distance taking account transport networks but it can be a good reference too).

-For each cluster we obtain the accumulated distance respect to all clinics (in kilometers) according to the following scheme:

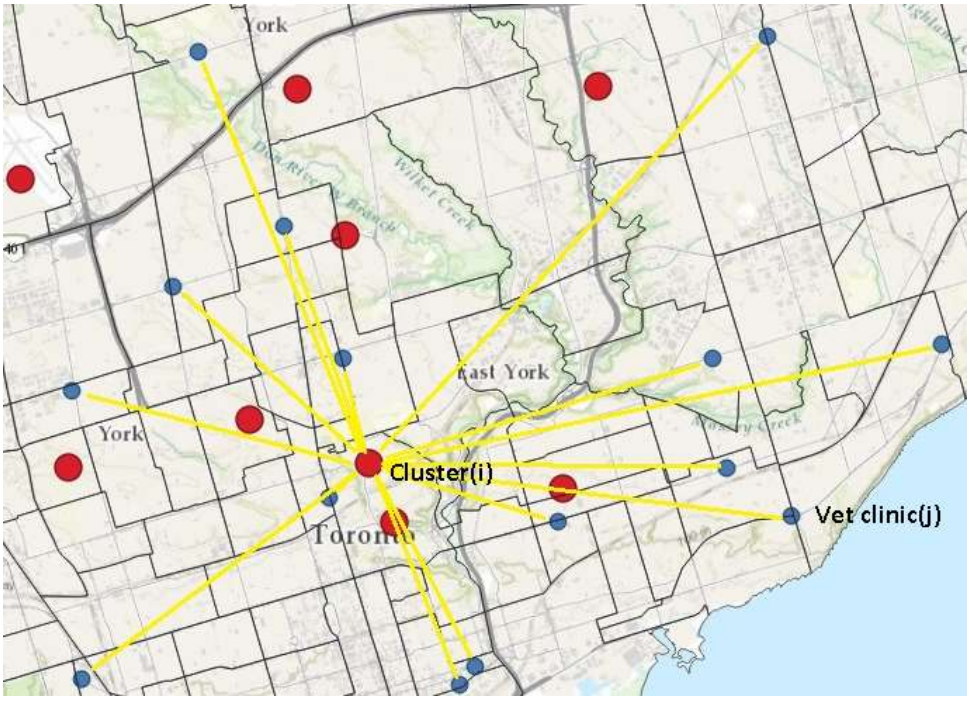

FIGURE 14: SCHEME OF CALCULATION OF ACCUMULATED DISTANCES BETWEEN CLUSTERS AND VET CLINICS/STORES

[9] Source: https://www.ratesupermarket.ca/blog/the-cost-of-owning-a-pet-in-canada/

$$distance_{cluster(i)} = \sum_{j=0}^{n_{vets\_clinics}} distance\ (cluster(i)\ to\ vet_{clinic(j)})$$

This parameter can be used as an indicator of how a specific neighborhood is covered by the existing vet clinics. To get an overall insight we can graph the relationship between pets_ratio (per neighborhood/cluster) and accumulated distance:

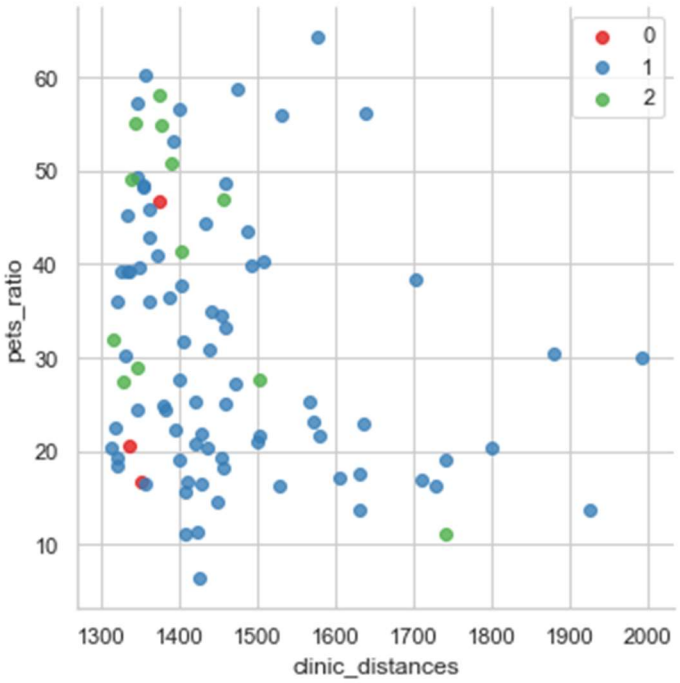

FIGURE 15: CLINICS-CLUSTERS_DISTANCES- VS. PETS_RATIO

In the upper scatter plot we can see that the highest pets_ratios are the most covered by existing vet clinics, although we can observe some remarkable exceptions (see the upper and lower right quadrant of the graph) which can be interesting placements to new clinics.

On the other hand, cluster nº2 locations (which we've identified previously as potential interesting placements because of their characteristics) are well covered in most cases.

We can repeat the same procedure using the pet stores dataframe. If we represent results graphically again:

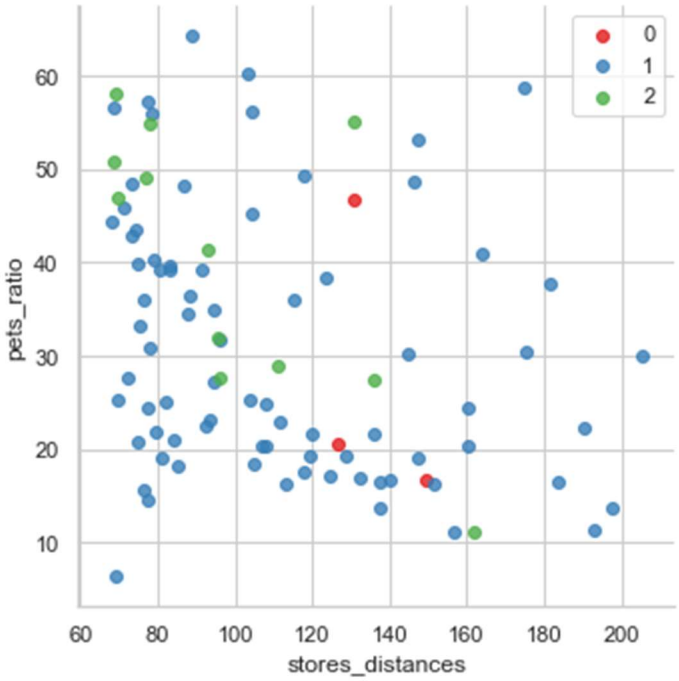

FIGURE 16: STORES-CLUSTERS_DISTANCES- VS. PETS_RATIO

Analyzing pet_stores distances respect to clusters we can reach similar conclusions as vet_clinic_distances: highest pets_ratios clusters are the best covered by existing pet stores.

On the other hand, cluster nº2 locations (which we've identified previously as potential interesting placements because of their characteristics) are well covered too.

3.3. A scoring approach to localize the best placements

We've seen some interesting insights about our problem, but now we need a procedure to localize the best placements to our new vet clinics. To get this target, a scoring system to classify all neighborhoods based on the data used in our analysis can be useful. A possible approach is:

$$Overall_{scoring} = Pets_{scoring} * Clinicdistances_{scoring} * Storesdistances_{scoring}$$

Where:

1) Clinicdistances & Storesdistances scorings are the accumulated distance to existing vet clinics & pet stores (max. distance equals to max. scoring).
2) Pets scoring is an estimation of the number of potential clients (using the total number of licensed pets in each neighborhood). In this case, the highest number of pets gets the highest scoring.

In both cases, we'll standarize the parameters (we'll use a max-min approach), so all factors will work at the same scale (between 0 and 1).

4. Results

Applying our scoring scheme, we can obtain a sorted dataframe with the best locations for new vet clinics in Toronto (sorted by overall_scoring):

Postal Code	Neighborhood	Cluster Labels	pets_scoring	clinic_distances_scoring	stores_distances_scoring	overall_scoring
M1C	Highland Creek, Rouge Hill, Port Union	1	0.564835	1.000000	1.000000	0.564835
M1E	Guildwood, Morningside, West Hill	1	0.761538	0.833647	0.781212	0.495956
M1B	Rouge, Malvern	1	0.476923	0.904417	0.943867	0.407125
M1G	Woburn	1	0.308242	0.719094	0.672445	0.149051
M1V	Agincourt North, L'Amoreaux East, Milliken, St...	2	0.313736	0.630588	0.682561	0.135037
M8V	Humber Bay Shores, Mimico South, New Toronto	1	0.991209	0.216960	0.569518	0.122477
M8W	Alderwood, Long Branch	1	0.643407	0.237976	0.776452	0.118887
M1S	Agincourt	1	0.313736	0.612370	0.609683	0.117134
M1M	Cliffcrest, Cliffside, Scarborough Village West	1	0.457692	0.574855	0.404906	0.106533
M1K	East Birchmount Park, Ionview, Kennedy Park	1	0.584066	0.476412	0.317487	0.088343

FIGURE 17: TOP TEN OF HIGHEST OVERALL SCORES (NEIGHBORHOODS)

M1C, M1E and M1B are the locations with the best scoring about our procedure. Finally, we can use Folium library to locate geographically the best places to new clinics:

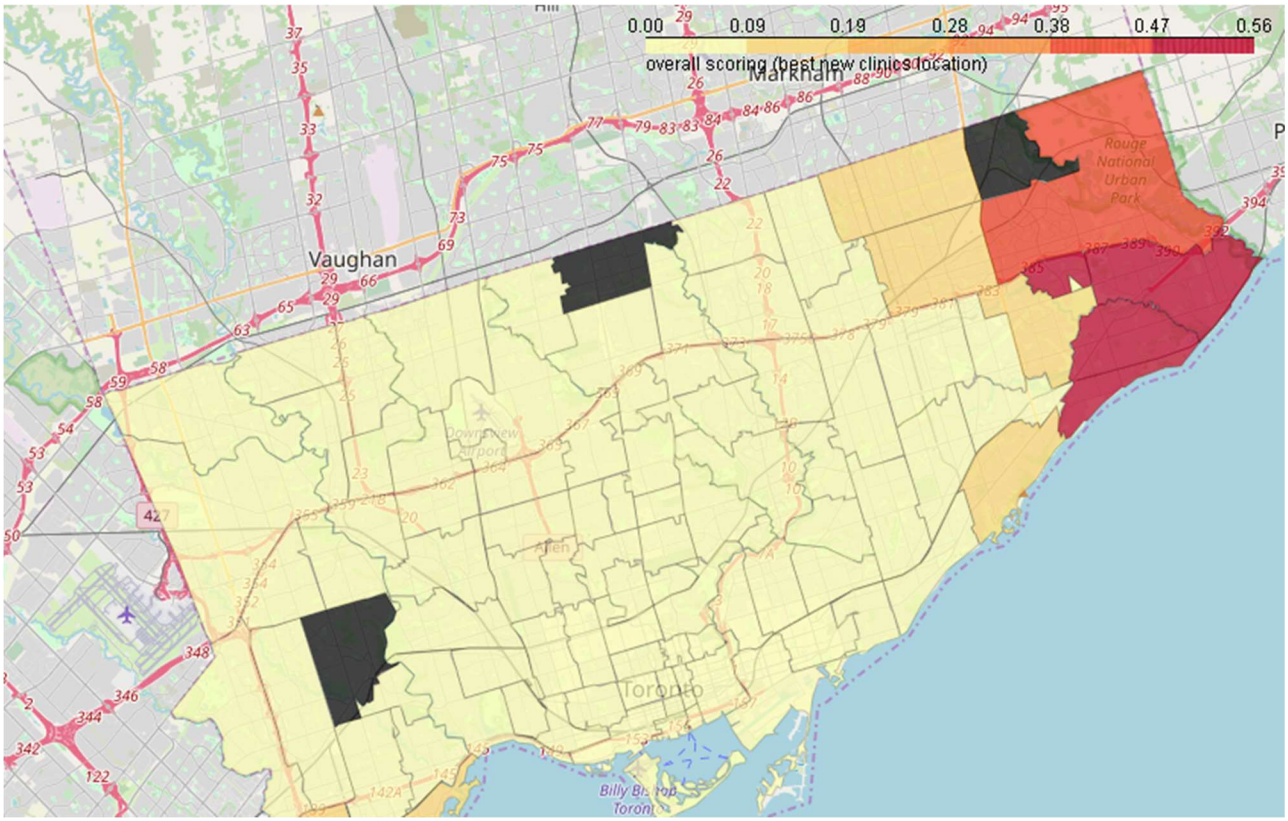

FIGURE 18: CHOROPLETH MAP (OVERALL_SCORINGS)

The map shows us an interesting aspect: the best locations are concentrated in a specific area (the east of Toronto) so new clinics can cover a wider area with interesting features (high number of potential clients in a closer radius of influence).

5. Discussion

Our proposal can be useful to set up an analytical procedure to search new locations in the vet care sector and not only for the analysis of best locations for new clinics: using the same data we can search for existing clinics that could be interesting to acquire (if our stakeholders are interested in the existing business instead of opening new clinics). Too, this approach can be used to analyze business opportunities in other cities, and even comparing them with Toronto's possibilities.

A great advantage of the proposed scoring procedure is its flexibility: we can modify it adding weighting factors if we consider that one or some of the parameters are more significant than the rest.

Respect to the clustering analysis, although finally our best locations aren't not especially related to a specific cluster (although we identified cluster nº2 as potentially suitable locations previously), we should check it because can add value to our proposal (for example, identifying places with special conditions for our business). In this case, neighborhoods with parks, dog runs, etc. looked like suitable places for our clinics. However, according to our analysis, logically our competence realized it before. But not always will have to be so.

6. Conclusions

In this study, we've developed a procedure to search for the best locations for new vet clinics in Toronto, using information from open data sources and specialized API such as FOURSQUARE. The starting point of the method is a clustering analysis which it's useful to identify what types of neighborhoods can we find. After, through a combination of different ratios (such as the number of pets per neighborhood and accumulated distances from existing vet care venues to the different areas) we've set up a scoring system to classify the neighborhoods according to their potential to locate new clinics. In this specific case, we've proposed three specific locations situated in the east of Toronto according to our estimation method.

We consider our approach it's quite flexible and can be adapted to other areas of study (new cities, also new types of businesses). Besides, we can modify it easily to change our target (for instance, analyze existing vet clinics to invest instead of opening new clinics).