IBM Capstone Project

BUSINESS ANALYSIS
VETERINARY CLINIC IN TORONTO, CANADA

By Thibault Dody

credit: https://www.uplacevet.com/

Introduction

Business owners can leverage the power of data analysis and machine learning technics to get valuable insights before opening a new place. By inspecting the distribution of their direct competitors and patterns in the population, one can optimize the location of a new business. In this example, we focus on a new veterinary clinic in Toronto. Using publicly available data, we want to find a good candidate neighborhood to start open a new veterinarian clinic.

Our analysis will be divided into the following phases:

1. Gather data
2. Get useful insights into the current status of the vet market
3. Leverage machine learning technic to identify areas lacking veterinarians

Data

Postal Codes

The first dataset is based on a Wikipedia article regarding the different neighborhoods in Toronto, Canada. This article contains a table providing the neighborhoods of the city of Toronto along with their postal codes and boroughs names. We will eventually use this dataset to produce plots of the city of Toronto.

The following features are included in this set:

1. Postcode
2. Borough
3. Neighborhood

Note that this set contains several records with unassigned values and will, therefore, need to be cleaned up.

Data location: Here

Census Data

The second dataset is provided by the government of Canada. It consists of a subset of the results of the 2016 census for the Forward sortation areas. This dataset will provide valuable input regarding the distribution of the population across the city of Toronto.

The set contains the following features:

1. Geographic code
2. Geographic name
3. Province or territory
4. Incompletely enumerated Indian reserves and Indian settlements, 2016
5. Population, 2016
6. Total private dwellings, 2016
7. Private dwellings occupied by usual residents, 2016

Data location: Here

Animal Registry

The final datasets consist of the pet registry (dogs and cats) in the city of Toronto for the years 2013 and 2017. These two datasets will be useful to observe trends in the number of registered animals in each neighborhood.

The datasets contain the following features:

1. FSA (postal code)
2. Cat: number of registered cats
3. Dog: number of registered dogs
4. Total: number of registered dogs and cats.

Data location: Here and Here

Foursquare API

Finally, the Foursquare API will be used to obtain indications of the current competitors (i.e. veterinary clinics). The API can be queried to obtain all existing veterinary clinics located within a certain distance of a neighborhood.

Methodology

Data Cleaning

The downloaded datasets cannot be used as is, there are missing records and the labels do not match between sets.

Postal Codes

The following transformations are performed on the postal code dataset:

1. Delete row where the Borough is defined as "**Not assigned**"
2. Concatenate neighborhoods with the same PostalCode
3. Replace unassigned Neighborhood by the Borough name

We end up with a data frame containing 103 records.

Census Data

The census dataset contains information regarding the entire territory of Canada. With over 1600 FSA contained in the set, the first transformation was to filter the census data set to only cover the neighborhoods of Toronto. The census data set was filtered using the postal codes contained in the Postal Code dataset.

In addition, the census dataset contains 6 neighborhoods with less than 15 inhabitants. These were filtered as they are outliers in term of the number of inhabitants as the mean neighborhood population is greater than ten thousand.

Animal Registry

The animal registry was only filtered to contain only the postal codes contained in the Postal Code dataset.

Dataset compatibility

Finally, all three datasets were filtered using the intersection of all the postal codes contained in the three datasets. The final sets contained information related to 96 neighborhoods.

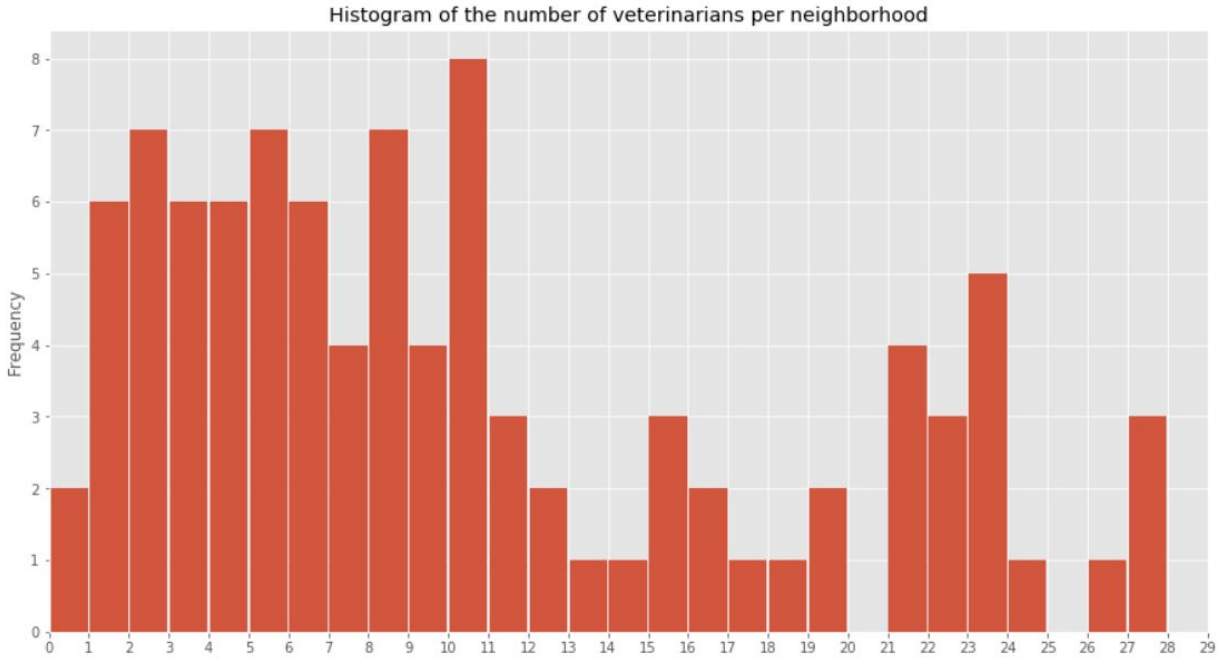
New Set - Existing Veterinarians

The ArcGIS geocoder is used to retrieve the latitude and longitude of the neighborhoods of Toronto. These coordinates are then used to query the Foursquare venues using the following properties:

1. Neighborhood latitude and longitude
2. Radius of 2000m
3. Limit to 50 results (never reached)
4. Venue category as *4d954af4a243a5684765b473* corresponding to **veterinarian** per the Foursquare documentation.

Note that the radius of 200 m will lead to results outside the neighborhood. This is not an issue as users do not limit their search by neighborhood and 2000 m seems to be a reasonable distance to travel when looking for a veterinarian. Once the API search is

terminated, a total of venues **129** unique venues were identified. The plot below shows the distribution of neighborhood as a function of the number of nearby veterinarians.



Feature Engineering

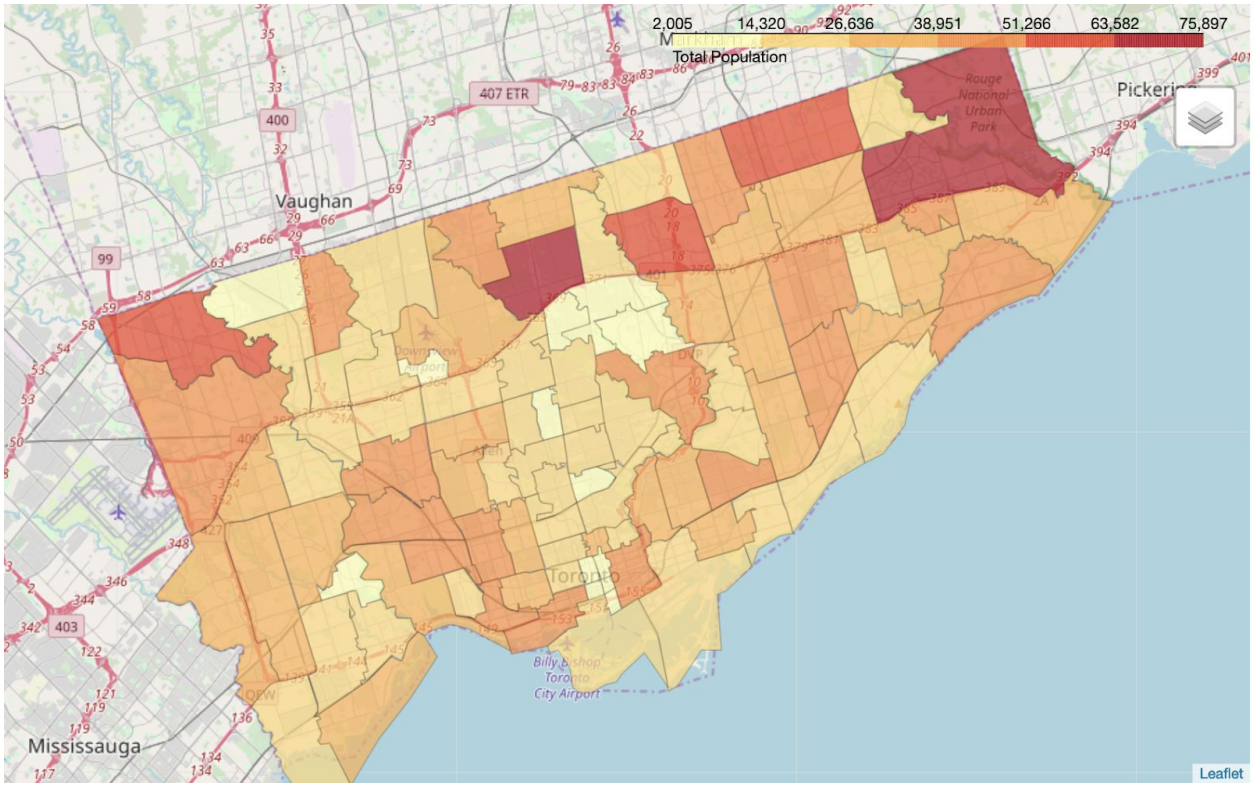
The following features were considered meaningful for this problem and were therefore added to the data set.

1. Number of veterinarians per 1000 inhabitants
2. Number of registered pets per 1000 inhabitants
3. Number of registered cats per 1000 inhabitants
4. Number of registered dogs per 1000 inhabitants
5. Number of vets per 1000 registered pets
6. Number of vets per 1000 registered inhabitants
7. Number of registered pets per 1000 dwellings
8. Number of vets per 1000 dwellings
9. Proportion of cats
10. Change in the cat registration number
11. Change in the dog registration number
12. Change in the total pet registration number

Data Exploration

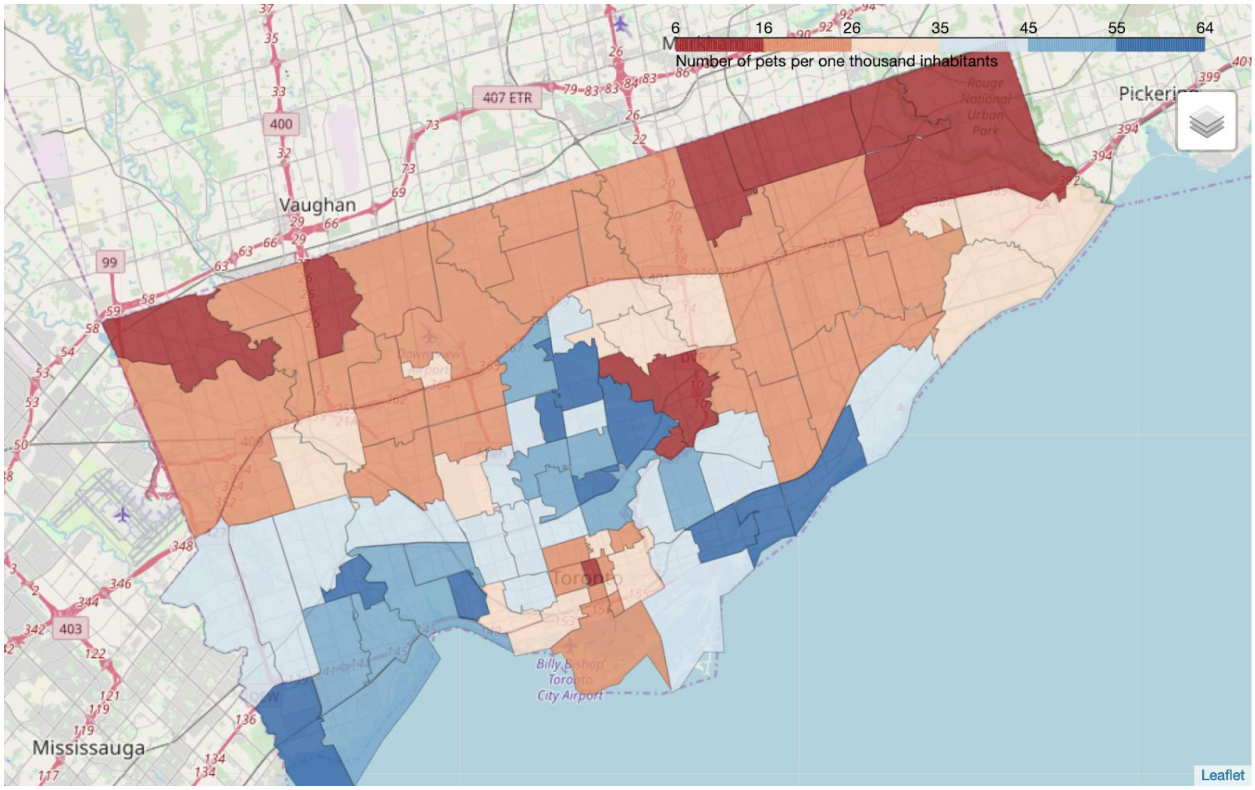
Population Distribution

As depicted by the map below, the population of Toronto is unevenly distributed between its neighborhoods. If we were to use the number of veterinarian per neighborhood instead of the ratio of the number of veterinarians and the population of the neighborhood, we would not properly capture the need for new veterinarians in densely populated neighborhoods.



Animal Distribution

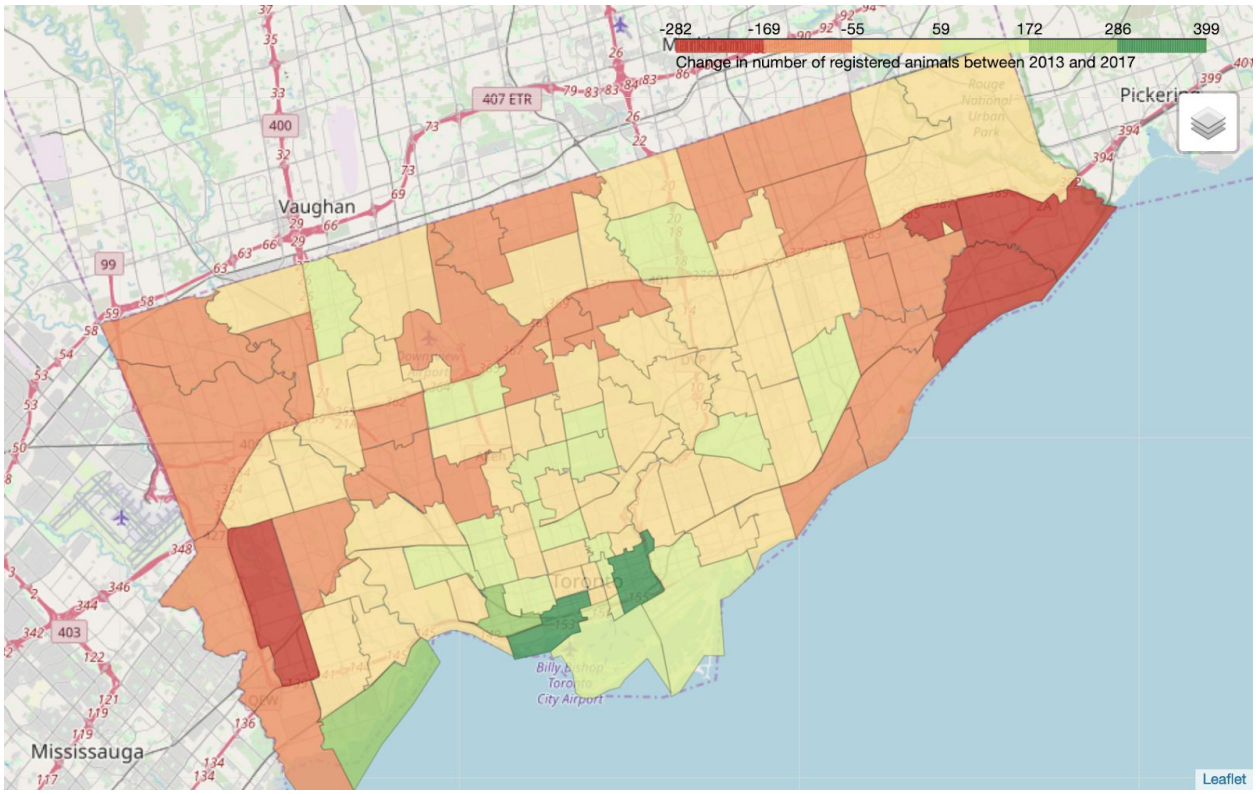
As shown by the map below, the neighborhoods in the center of the city (south-west to north-east direction) have the largest ratio of pets per inhabitants while the northern neighborhoods have the smallest ratios.



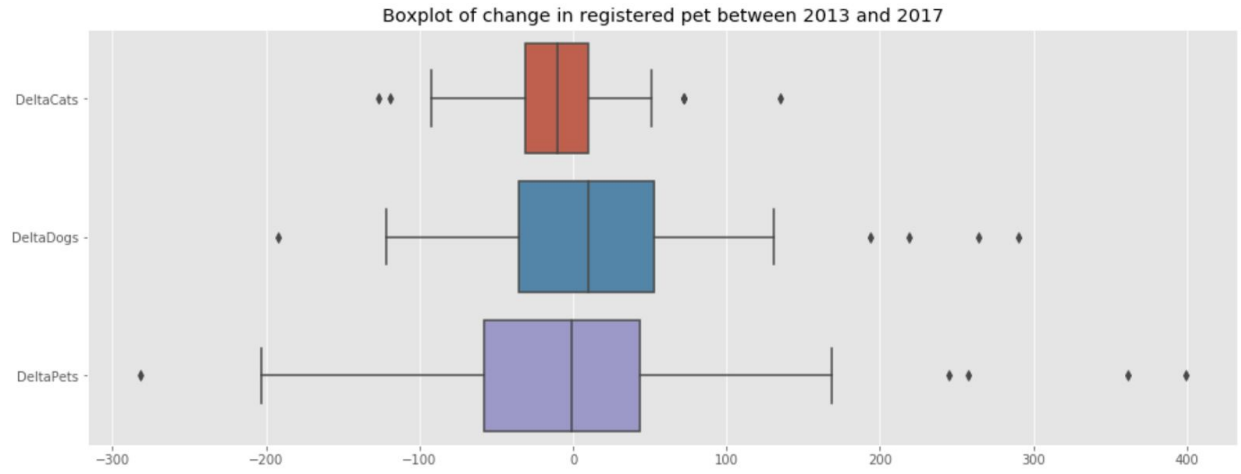
The results from this visualization will need to be combined with the veterinarian distribution in order to target the best market location for our new clinic.

Change is Animal Registration

Although the pet to people ratio is a fundamental index in our study, we also need to look at trends in the data. Indeed when considering a new location for a business, one needs to consider the evolution of its customer population. In our case, we want to know if people are registering more pets or less. This approach will help us distinguish promising locations. For instance, the small neighborhoods located on the south shore of the city were identified as having the smallest pet to inhabitant ratio but they appear to also be the fastest growing neighborhoods for pet registration.



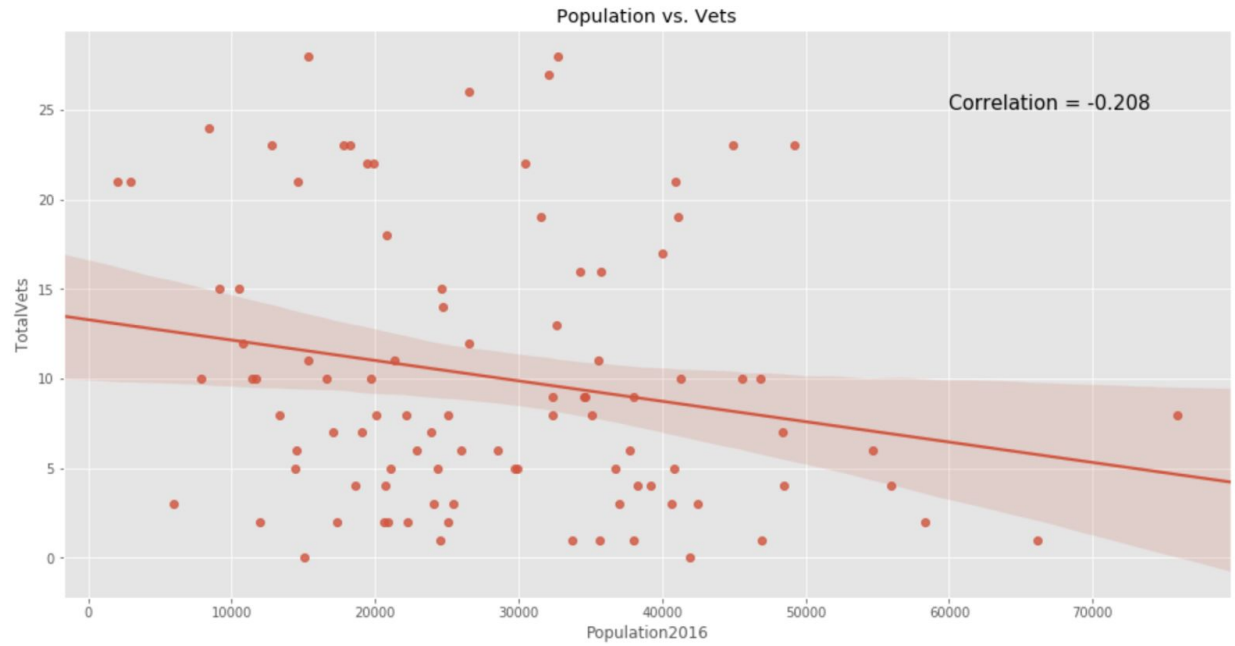
Finally, the boxplot below contains valuable information regarding the evolution of cats and dogs between 2013 and 2017. On average, the number of registered pets remains unchanged or decreases slightly but it appears that people tend to have fewer cats and more dogs.



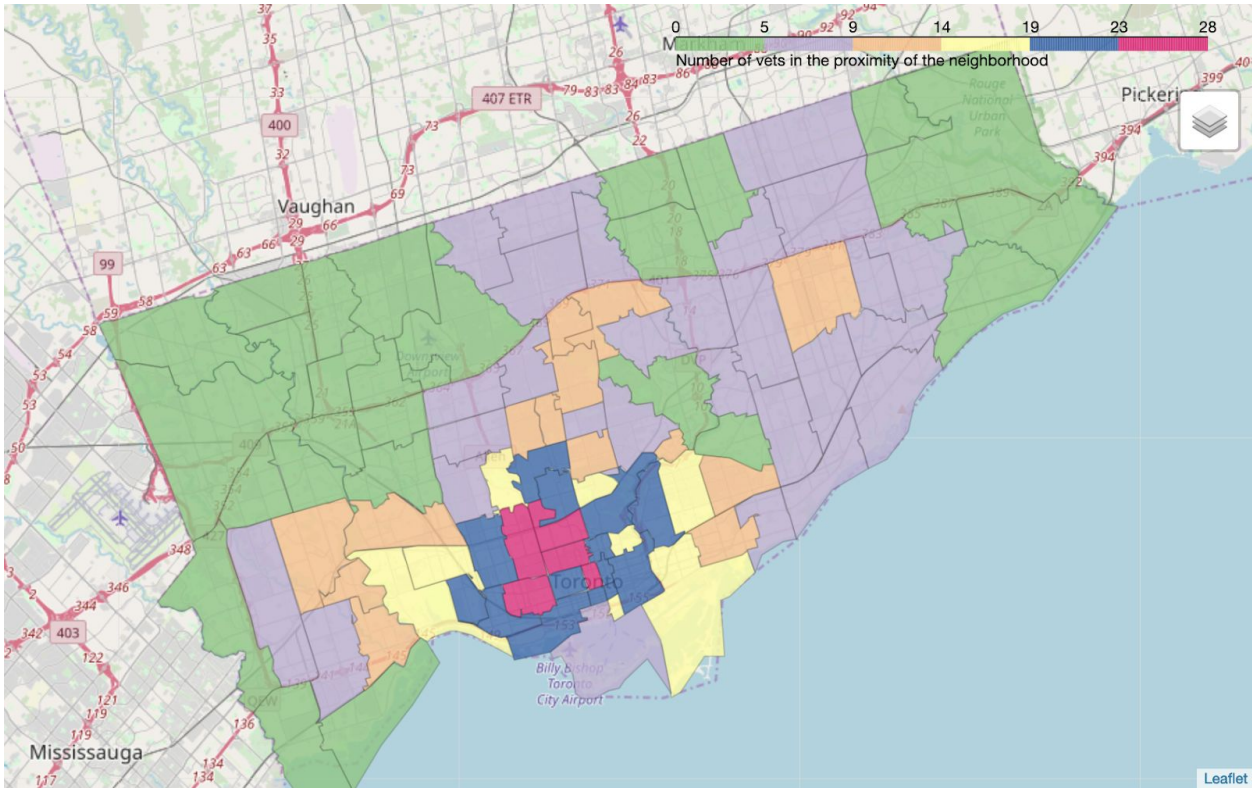
Vet Distribution

The last piece of our puzzle consists of understanding the distribution of vets in Toronto. The goal is to answer whether the vets are located near densely populated areas are where most pets are.

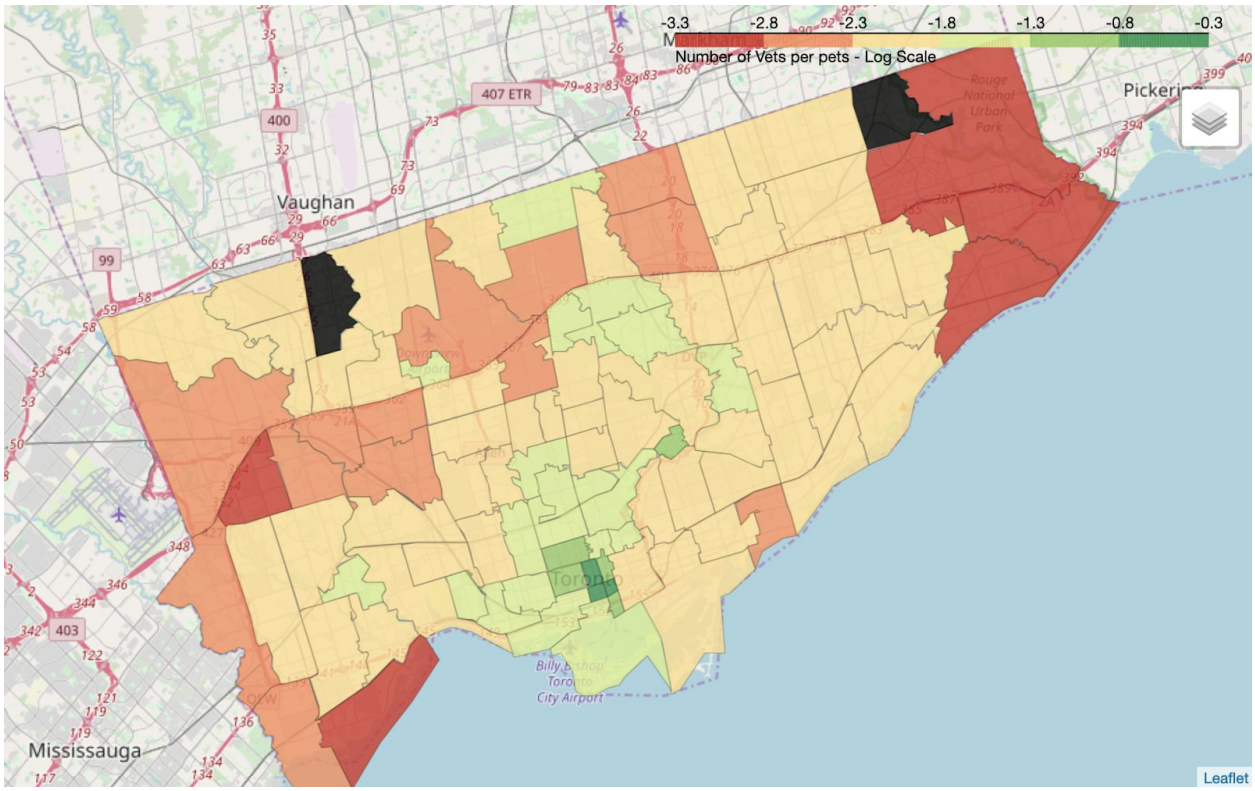
The plot below shows the results of a linear regression fitted on the number of vets per neighborhood as a function of the neighborhood population. As we can see, the negative correlation shows us that the vets are not concentrated in densely populated neighborhoods.



The map below shows the number of vets located within 2km of a neighborhood center. The neighborhoods located in the center of Toronto are all located near a large number of vets. As we move away from the city center, the number of nearby vets drops significantly. The neighborhoods located along the perimeter of the city have between 0 to 5 vets within a radius of 2 km.



When combining the number of registered pets per neighborhood with the number of nearby vets per neighborhood, we can see the east and west sides have the smallest ratio of vets per pets. This is a good indicator because it means that pet owners living in these neighborhoods have to travel further away to find a vet. Note that the index used in the map below is based on a $\log_{10}$ scale for clarity.



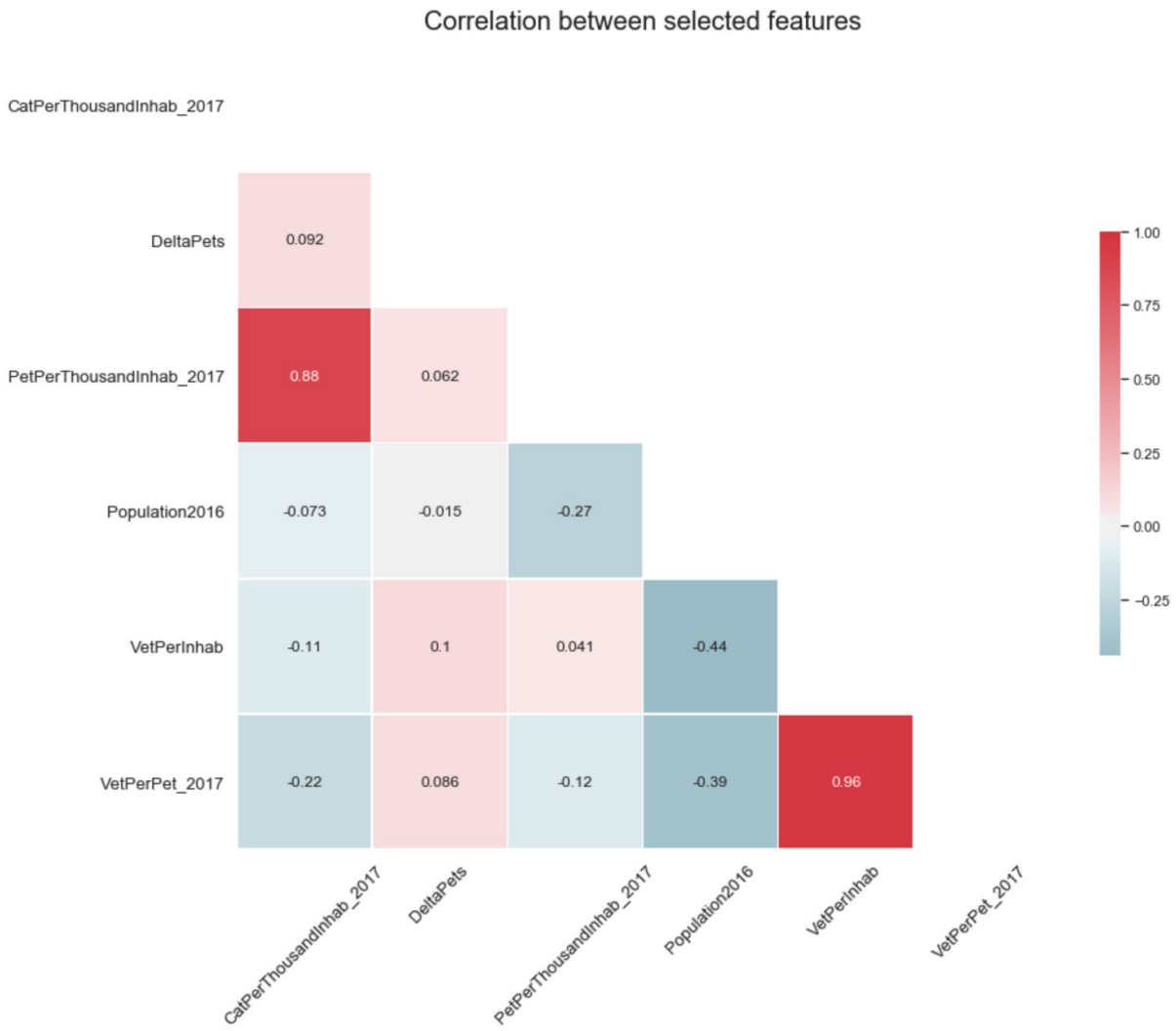
From the preliminary inspection of the data and multiple plots, we can predict that the best location for the new veterinary clinic will be one of the neighborhoods highlighted in yellow, orange, or red above.

Results

In this section, we leverage the power of clustering in order to group the neighborhoods into meaningful clusters. Before we do, the dataset features need to be filtered so that only the most meaningful ones are used. The following are kept:

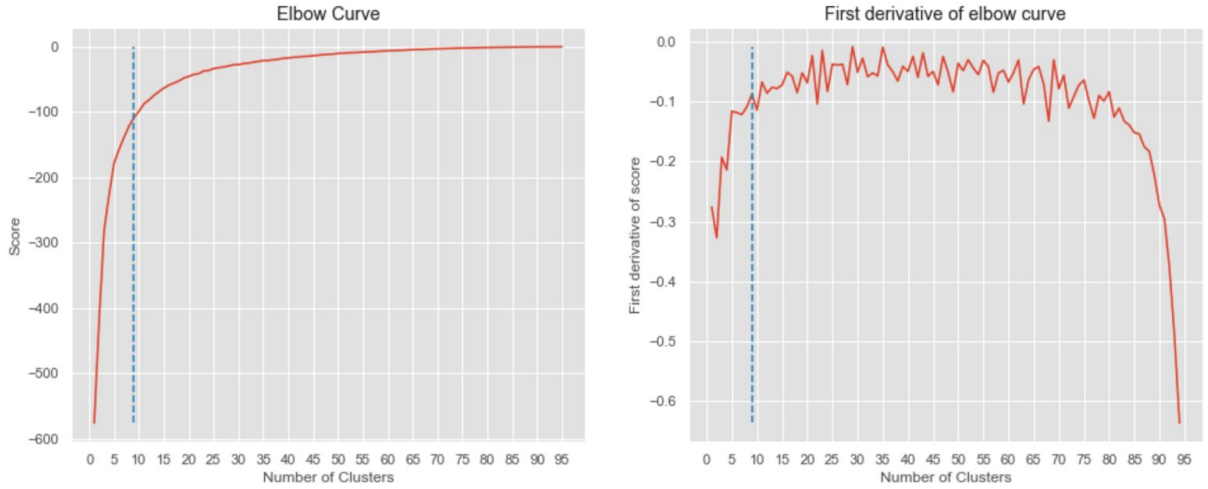
1. CatPerThousandInhab_2017
2. DeltaPets
3. PetPerThousandInhab_2017
4. Population2016
5. VetPerInhab
6. VetPerPet_2017'

We choose to use K-mean clustering to group the neighborhoods into meaningful clusters. After performing a search for the best value of the hyperparameter K, we trained our final model with 9 clusters.

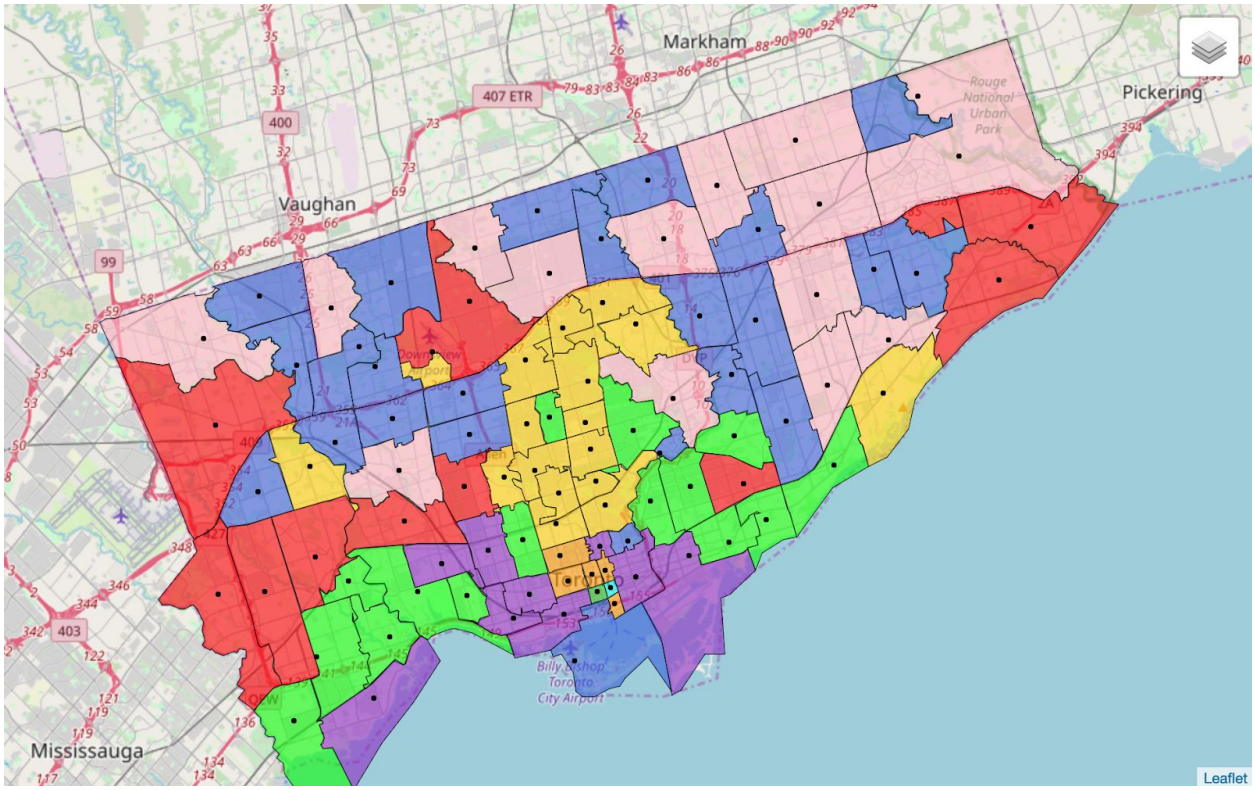


The above figure presents the correlation between the selected variables. As one can expect, there are a few strongly correlated features (Vets per inhabitants and vets per pets...)

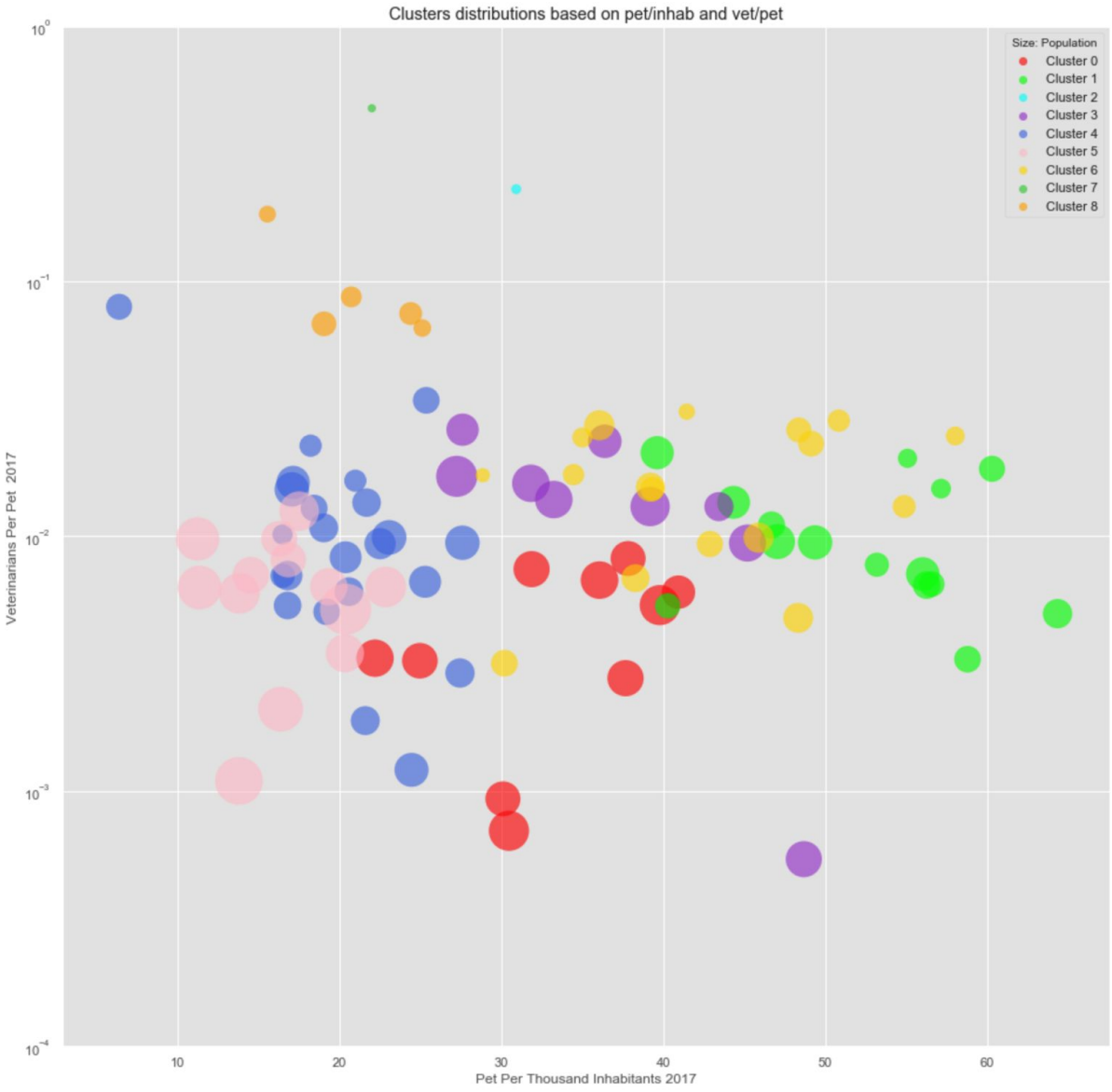
The figure below shows the results of model tuning. We select the value of K which results in the best fit of the data without overfitting.



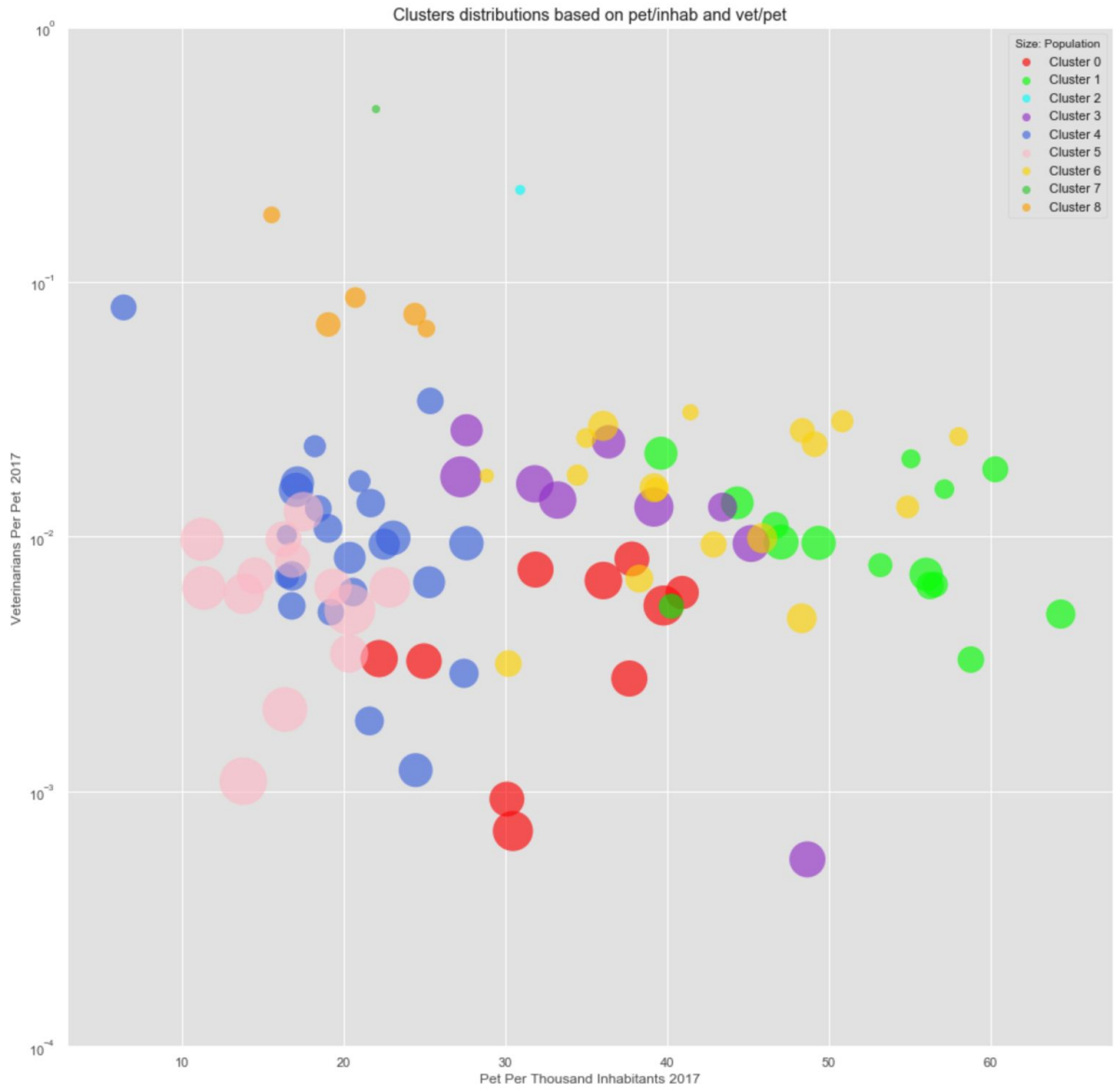
When plotted by clusters, the distribution of neighborhoods gives a more meaningful representation of the segmentation of the city. Even though the latitudes and longitudes were not included in the model, neighborhoods of the same cluster are located near each other.



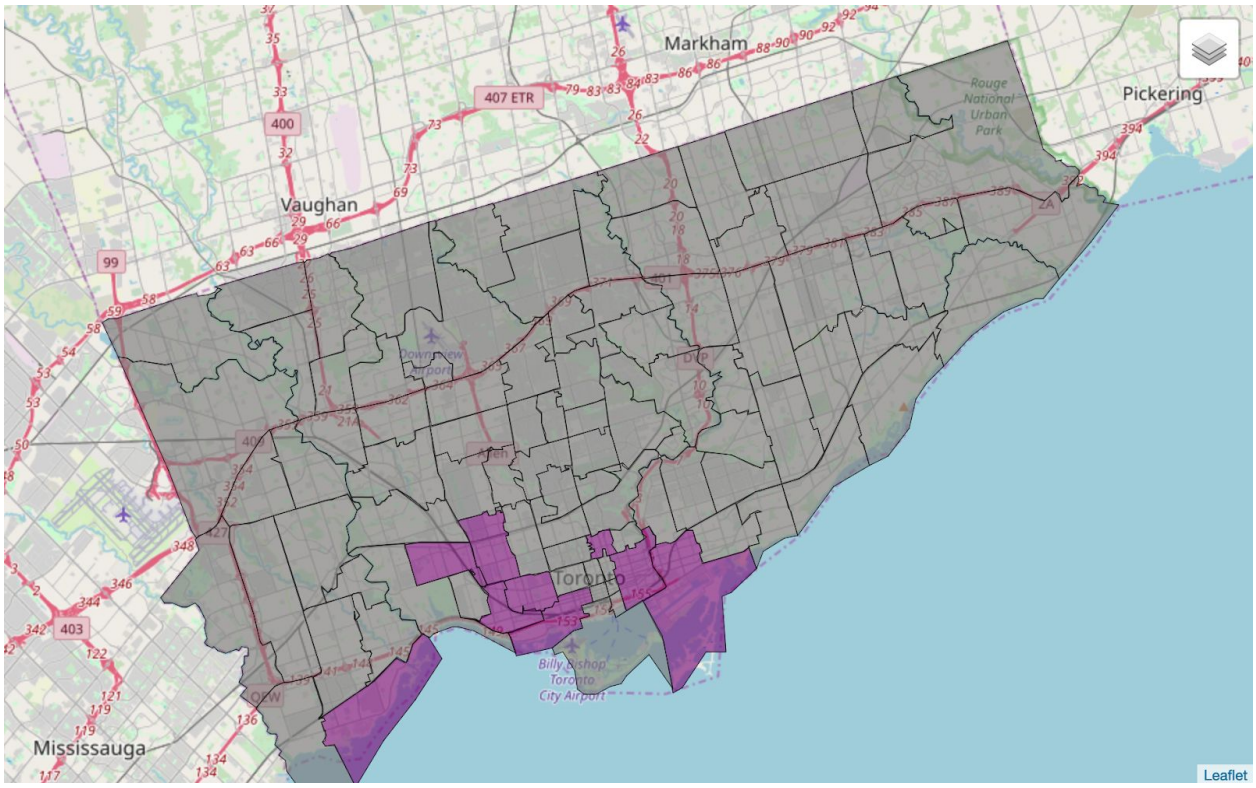
The next step of the analysis consists of producing plots using the selected features and group the data points per cluster. The figure depicts the neighborhood distribution based on the “pet per thousand inhabitants” and “vet per pet” features. From this visualization, it seems that the red (cluster 0), orange (cluster 8), green (cluster 1), and purple (cluster 3) clusters are all located in the area where the “pet per thousand inhabitants” is high while the “vet per pet ratio” is low.



In order to distinguish these four potential candidates, the “vet per pet” and “change in registered pets” are examined.



From the above plot, only one of our four candidates presents good characteristics. Indeed, with a low “vet per pet” ratio and a positive change in the number of registered pets, Cluster 3 appears to be an ideal candidate.



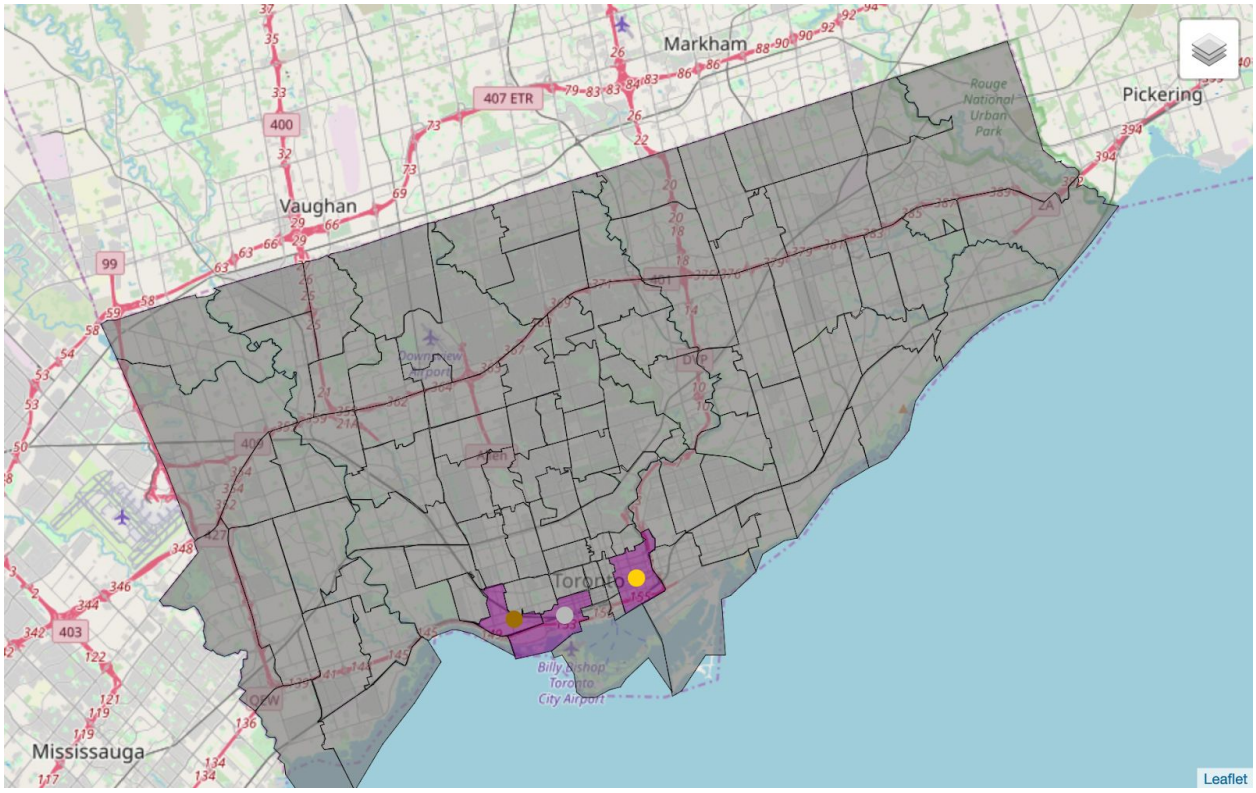
In order to select the best three neighborhood for the new veterinarian clinic, we first sort and filter the best five neighborhoods of cluster 3 based on the pets to vets ratio. This leads to the following subset:

PostalCode	Vet per Pet
M4Y	**0.026159**
M6J	**0.023549**
M5V	**0.017164**
M6K	0.016117
M5A	0.013919
M6H	0.013061
M4M	0.013060
M6P	0.009397
M8V	0.000541

We now look at our last feature, the change in registered pets between 2013 and 2017. From the plot below, it appears that all the neighborhoods of Cluster 3 have seen an increase in registered pets.

PostalCode	Change in Registered Pets
M5A	**399**
M5V	**362**
M6K	**258**
M6J	169
M4Y	168

From the above, the three best candidates are the neighborhoods corresponding to the postal code M5A, M5V, and M6K. They are plotted on the map below: M5A (gold), M5V (silver), and M6K (bronze).



Discussion

From our analysis, we can conclude that based on the available data, the location for a new veterinarian can be narrowed down to a cluster of 8 neighborhoods (Cluster 4) and more specifically to 3 neighborhoods with the neighborhood corresponding to M5A being the most promising candidate.

However, it is important to note the following aspects of the analysis and assumptions that were considered for this study:

1. The study relies heavily on the fact that the number of registered pets is a good representation of the actual number of pets per neighborhood.
2. The number of pets per neighborhood is simplified by being taken as the sum of registered pets and dogs. It does not include smaller animals which can represent a significant number of individuals.
3. The requests made on Foursquare does not account for any ranking or reviews of the existing veterinarian places.

In conclusion, this study constitutes the initial step of an involving and complex problem to solve than a stand-alone analysis.

Conclusion

In conclusion, data analysis technics and machine learning processes can be leveraged on publicly available datasets. In our case, with three simple datasets and a simple clustering algorithm, we were able to make simple yet effective observations of trends in the data and make a data-informed decision regarding the potential best location for a new business.