Capstone Project: The Battle of Neighborhoods

Finding the Optimal Location for a Veterinary Clinic in Toronto

1. Introduction: Description of the business problem

In today's world it is of utmost importance to perform detailed analysis of the market and design a thorough business plan before launching a new enterprise. The times when skills and craftsmanship were sufficient for a success in business have long gone. Part of the market analysis is to collect and process the data relevant to the business one is planning on starting and use them to make a sound, fact-based business decision. One interesting example might be the idea to open a veterinary clinic in a large city. Finding an office space, making sure it is fit for the purpose and within the available budget is the groundwork done the old way. However, finding out how many veterinary clinics already exist in the city, how distributed they are and what affect it will have on the new business may be the crucial factor when it comes to deciding on the optimal location of the business. It would not be to difficult to walk around and a count exactly how many similar businesses are there in a small city. In a large city that spans over hundreds of square miles it would be an impossible mission. From the clients' perspective having more choices is good as it gives clients many options to choose from; from the business owners perspective though it is clearly a disadvantage as they need to take a chunk of the market while competing with many other similar businesses.

In summary, I would like to address the following hypothetical problem:
A person wants to open a veterinary clinic in Toronto. Before deciding on the optimal location, the person wants to find out:

- How many veterinary clinics exist in Toronto?
- Where they are located?
- How are they distributed, i.e. how many clinics there are in each city area?
- What is the population density?
- What is the correlation between the distribution of the existing veterinary clinics and the population density in all areas across the city?

In the end, the decision on the location of the new veterinary clinic will be based on the assumption that for a small start-up business it is best to have:

(i) small number of similar businesses in the same area, and
(ii) large population

In the real world this would not be the only nor the most important factor in the decision, but for the purpose of this analysis it will serve as a good scenario.

2. Data: What data will be used for solving the problem and what will be the source of the data?

The data used in the analysis will be derived from multiple sources.

First, we will need the list of FSAs (Forward Sortation Areas) assigned in Toronto. The list is available on the Wikipedia and must be scraped, processed and formatted as a dataframe to make it possible to merge it with other datasets.
List of postal codes of Canada: M - Wikipedia

A Canadian postal code is a six-character string that forms part of a postal address in Canada. Canada's postal codes are alphanumeric. They are in the format A1A 1A1, where A is a letter and 1 is a digit, with a space separating the third and fourth characters. A forward sortation area (FSA) is a geographical region in which all postal codes start with the same three characters. The dataset will consist of FSA codes, borough names and the list of neighbourhoods in each borough. This will be used as the basis for further data analysis, as outlined in the next paragraphs.

Second, we will need geospatial data to pair the FSAs with their geographical latitude and longitude. This will allow us to create a map and plot the FSAs as part of the data visualization component. This dataset can be obtained through python libraries where FSA information is passed as parameter. Several Python packages have been developed to make working with geospatial data in Python easier. The list of FSA codes from the first dataset will be used as input to retrieve latitude and longitude of each FSA.

Third, we will need the veterinary clinic information extracted from Foursquare. Foursquare City Guide is a local search engine and discovery mobile app that helps users discover new places from a community of peers. It provides personalized recommendations of places to go near a user's current location based on the user's previous visits, likes and check-in history. The data will be retrieved via API calls, processed, cleaned and formatted to make it possible to merge it with other datasets. The API calls have a defined format with several pieces of information are concatenated and sent to Foursquare:
https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={}&radius={}&limit={}&query={}
Parameters used:

- client_id, client_secret = tokens assigned to each registered user

- v = version, i.e. date of the search result dataset update
- radius = limit results to venues within this many meters of the specified location; defaults to a city-wide area; only valid for requests that use categoryId or query; the maximum supported radius is currently 100,000 meters
- limit = number of results to return
- query = a search term to be applied against venue names

In return, Foursquare generates a dataset in ".json" format, which will be extracted, flattened and formatted as the dataframe compatible with the other datasets as mentioned above. The dataset will have venue names, categories (in this case the key word used for data retrieval will be 'veterinary'), venue latitude/longitude and the FSA the venue is located in (the first three characters of the postal code).

The fourth and final dataset will be the population by FSA available on the Statistics Canada website: https://www12.statcan.gc.ca/ The website offers a multitude of open datasets available in various formats, most often as comma or tab separated values text files. This dataset will give us the information on how many people live in each city area, which will be used for the comparative analysis with the distribution of the veterinary clinics dataset as described above.

3. Methodology

Data scraping, also known as web scraping, is the process of importing information from a website into a spreadsheet or local file saved on your computer. It is one of the most efficient ways to get data from the web, and in some cases to channel that data to another website. This method will be used for obtaining the first dataset: the list of FSAs (Forward Sortation Areas) assigned in Toronto. The dataset is available on Wikipedia (https://en.wikipedia.org/) as an HTML table. This phase includes the following activities:

- opening the webpage
- looping through the tables on the webpage
- identifying the index ('order of appearance') of the table
- downloading the dataset
- loading the dataset to a list
- importing the list to a pandas dataframe

Data wrangling is the process of gathering, selecting, and transforming data to answer an analytical question. Before the dataset is ready for further processing, these additional data preparation steps will be necessary:

- assigning column names, as columns are originally unnamed

- data parsing: extract the first 3 characters (i.e. FSA) of the input string and saving the output as 'PostalCode' column
- data parsing: extract the substring of the input string beginning with the 4th character and ending before the first occurrence of the delimiter (in this case: left parenthesis) and saving the output as 'Borough' column
- data parsing: extract the substring of the input string beginning after the first occurrence of left parenthesis and ending before the first occurrence of right parenthesis, and saving the output as 'Neighbourhoods' column
- data parsing: replace forward slash with a comma in the 'Neighbourhoods' column
- data cleansing: remove the rows where FSA is not assigned
- data cleansing: drop the unneeded columns – the original input string as well as any temporary columns created during the data wrangling phase

The next step in the data wrangling phase is to download the latitude and longitude data and merge with the previously prepared dataframe. As part of this step, "pgeocode" is imported. It is a Python library for high performance off-line querying of GPS coordinates, region name and municipality name from postal codes. Utilizing a "for loop", FSRs are retrieved from the 'PostalCode' column of the dataframe and passed as input parameters to a "pgeocode" object using the "query_postal_code" method. The values returned by the call are inserted to the 'Latitude' and 'Longitude' columns in the dataframe.

The next phase in the process is to build the list of Foursquare queries using the pre-defined list of values and appending them to the dynamically built API calls. Foursquare API calls can only run with one set of input parameters at a time. This means that one API call can only retrieve the list of venues in vicinity of the location specified in the call. The solution to this problem is to loop through all FSAs ('PostalCode' column) and dynamically build one API call for each FSA. The API call statements are then stored in a new dataframe called 'df_foursquare_queries', which will be used for executing and retrieving the venue information from Foursquare. Other predefined parameters built into the Foursquare API call are:

- CLIENT_ID
- CLIENT_SECRET
- VERSION
- RADIUS
- LIMIT
- QUERY

To facilitate the next steps, it is necessary to install additional libraries that will be used for data processing and visualization: folium, json and matplotlib. Also, a function for exctracting the venue category from the Foursquare result set.

One of the key steps is store the dataset into a pandas dataframe and to flatten the .json format to a tabular view. The Foursquare result includes a multitude of potentially useful information, but for the purpose of this solution some of the columns are not needed and will be excluded from further analysis. The relevant columns will be extracted and stored in the 'df_neighbourhoods_venues' dataframe. This dataframe will have one row per venue. This may be useful for venue related analysis, but the focus of this solution the correlation between FSAs and geographical distribution of venues. Therefore, it will be necessary to group venues by FSA and count the venues in each FSA. The results will be stored in the 'df_venue_count' dataframe

The following methods will be used in theudata processing will include scraping, wrangling, analysis, merging/concatenation, dynamically building and execution of API queries, clustering, normalization and plotting a map with colour coded markers for easy identification of the areas with different venue counts. Markers will be superimposed on the Toronto area map with labels showing venue counts and location marker pop-ups showing FSA codes.

The next dataset is the population distribution by FSA. The dataset is available as a text file with comma separated values. The file will be downloaded and stored in a new dataframe called 'df_fsa_population'. Similar to the previously described situation, the dataset consists of several columns with some of columns not being relevant to the problem at hand – these columns will be dropped from the dataframe. The remaining columns will be renamed to more meaningful names to avoid any confusion. As the last step in this phase the 'df_venue_count' and the 'df_fsa_population' dataframes will be merged to form the 'df_fsa_population_venue_count' dataframe.

Normalization refers to rescaling real-valued numeric attributes into a 0 to 1 range. Data normalization is used in machine learning to make model training less sensitive to the scale of features. Two key columns in the 'df_fsa_population_venue_count' dataframe are 'Population' and 'VenueCount'. As their value ranges are very different it will be helpful to calculate the 'Population' / 'VenueCount' ratio and store the values in a separate column called 'PopulationPerVenue', and normalize the values to enable further data analysis steps.

Data analysis phase focus is on finding the optimal 'Population' / 'VenueCount' ratio, with the assumption that the highly desirable situation for a startup business is to have a huge potential client base (i.e. large population) and lack of competition (i.e. low number of similar businesses serving the same target area). The goal is to analyze and cluster the values, plot the map, identify the outliers (if any) analyze the feasibility of choosing the optimal location for the veterinary clinic with this approach.

4. Results

The starting point is the following table with FSAs assigned to Toronto:

M1A Not assigned	M2A Not assigned	**M3A** North York (Parkwoods)	**M4A** North York (Victoria Village)	**M5A** Downtown Toronto (Regent Park / Harbourfront)	**M6A** North York (Lawrence Manor / Lawrence Heights)	**M7A** Queen's Park (Ontario Provincial Government)	M8A Not assigned	**M9A** Etobicoke (Islington Avenue)
M1B Scarborough (Malvern / Rouge)	M2B Not assigned	**M3B** North York (Don Mills) North	**M4B** East York (Parkview Hill / Woodbine Gardens)	**M5B** Downtown Toronto (Garden District, Ryerson)	**M6B** North York (Glencairn)	M7B Not assigned	M8B Not assigned	**M9B** Etobicoke (West Deane Park / Princess Gardens / Martin Grove / Islington / Cloverdale)
M1C Scarborough (Rouge Hill / Port Union / Highland Creek)	M2C Not assigned	**M3C** North York (Don Mills) South (Flemingdon Park)	**M4C** East York (Woodbine Heights)	**M5C** Downtown Toronto (St. James Town)	**M6C** York (Humewood-Cedarvale)	M7C Not assigned	M8C Not assigned	**M9C** Etobicoke (Eringate / Bloordale Gardens / Old Burnhamthorpe / Markland Wood)
M1E Scarborough (Guildwood / Morningside / West Hill)	M2E Not assigned	M3E Not assigned	**M4E** East Toronto (The Beaches)	**M5E** Downtown Toronto (Berczy Park)	**M6E** York (Caledonia-Fairbanks)	M7E Not assigned	M8E Not assigned	M9E Not assigned
M1G Scarborough (Woburn)	M2G Not assigned	M3G Not assigned	**M4G** East York (Leaside)	**M5G** Downtown Toronto (Central Bay Street)	**M6G** Downtown Toronto (Christie)	M7G Not assigned	M8G Not assigned	M9G Not assigned
M1H Scarborough (Cedarbrae)	**M2H** North York (Hillcrest Village)	**M3H** North York (Bathurst Manor / Wilson Heights / Downsview North)	**M4H** East York (Thorncliffe Park)	**M5H** Downtown Toronto (Richmond / Adelaide / King)	**M6H** West Toronto (Dufferin / Dovercourt Village)	M7H Not assigned	M8H Not assigned	M9H Not assigned
M1J	**M2J**	**M3J**	**M4J**	**M5J**	**M6J**	M7J	M8J	M9J

Figure 1: FSAs assigned in Toronto

Once the table is downloaded, parsed and cleansed, it is transformed to this dataframe:

	PostalCode	Borough	Neighbourhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Figure 2: dataframe 'df_neighbourhoods'

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.8113	-79.1930
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.7878	-79.1564
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.7678	-79.1866
3	M1G	Scarborough	Woburn	43.7712	-79.2144
4	M1H	Scarborough	Cedarbrae	43.7686	-79.2389

Figure 3: dataframe 'df_neighbourhoods' with latitude and longitude

Dynamically building the Foursquare API calls generates this dataframe:

	PostalCode	Borough	Latitude	Longitude	FSquery
0	M1B	Scarborough	43.811300	-79.193000	https://api.foursquare.com/v2/venues/explore?&client_id=1UJCVJ3INGBNBAIP4HGIA0PY0ZWYG1IES5NKIX0XJW1UCEEN&client_secret=AD33OJ12KFJO
1	M1C	Scarborough	43.787800	-79.156400	https://api.foursquare.com/v2/venues/explore?&client_id=1UJCVJ3INGBNBAIP4HGIA0PY0ZWYG1IES5NKIX0XJW1UCEEN&client_secret=AD33OJ12KFJO
2	M1E	Scarborough	43.767800	-79.186600	https://api.foursquare.com/v2/venues/explore?&client_id=1UJCVJ3INGBNBAIP4HGIA0PY0ZWYG1IES5NKIX0XJW1UCEEN&client_secret=AD33OJ12KFJO
3	M1G	Scarborough	43.771200	-79.214400	https://api.foursquare.com/v2/venues/explore?&client_id=1UJCVJ3INGBNBAIP4HGIA0PY0ZWYG1IES5NKIX0XJW1UCEEN&client_secret=AD33OJ12KFJO
4	M1H	Scarborough	43.768600	-79.238900	https://api.foursquare.com/v2/venues/explore?&client_id=1UJCVJ3INGBNBAIP4HGIA0PY0ZWYG1IES5NKIX0XJW1UCEEN&client_secret=AD33OJ12KFJO
5	M1J	Scarborough	43.746400	-79.232300	https://api.foursquare.com/v2/venues/explore?&client_id=1UJCVJ3INGBNBAIP4HGIA0PY0ZWYG1IES5NKIX0XJW1UCEEN&client_secret=AD33OJ12KFJO
6	M1K	Scarborough	43.729800	-79.263900	https://api.foursquare.com/v2/venues/explore?&client_id=1UJCVJ3INGBNBAIP4HGIA0PY0ZWYG1IES5NKIX0XJW1UCEEN&client_secret=AD33OJ12KFJO

Figure 4: dataframe 'df_foursquare_queries' with dynamically built list of Foursquare queries

	PostalCode	Borough	Latitude	Longitude	VenueName	VenueCategory	VenueLat	VenueLong
0	M1B	Scarborough	43.8113	-79.193	West Hill Animal Clinic	Veterinarian	43.8113	-79.138002
1	M1B	Scarborough	43.8113	-79.193	Markham Road Animal Hospital	Veterinarian	43.8113	-79.229159
2	M1B	Scarborough	43.8113	-79.193	Whites Road Animal Hospital	Veterinarian	43.8113	-79.122314
3	M1B	Scarborough	43.8113	-79.193	Ashcott Veterinary Clinic	Veterinarian	43.8113	-79.290771
4	M1B	Scarborough	43.8113	-79.193	Markham Veterinary Clinic	Veterinarian	43.8113	-79.238510

Figure 5: dataframe 'df_neighbourhoods_venues' with venue information

Grouping venues by FSA and calculating the venue count for each FSA area produces the following dataframe:

	PostalCode	VenueCount	Latitude	Longitude
0	M4P	44	43.7135	-79.3887
1	M6C	44	43.6915	-79.4307
2	M5N	43	43.7113	-79.4195
3	M6G	43	43.6683	-79.4205
4	M5P	43	43.6966	-79.4120

Figure 6: dataframe 'df_venue_count'

This is the visual representation of the venues clustered by FSA and superimposed on the map of Toronto:

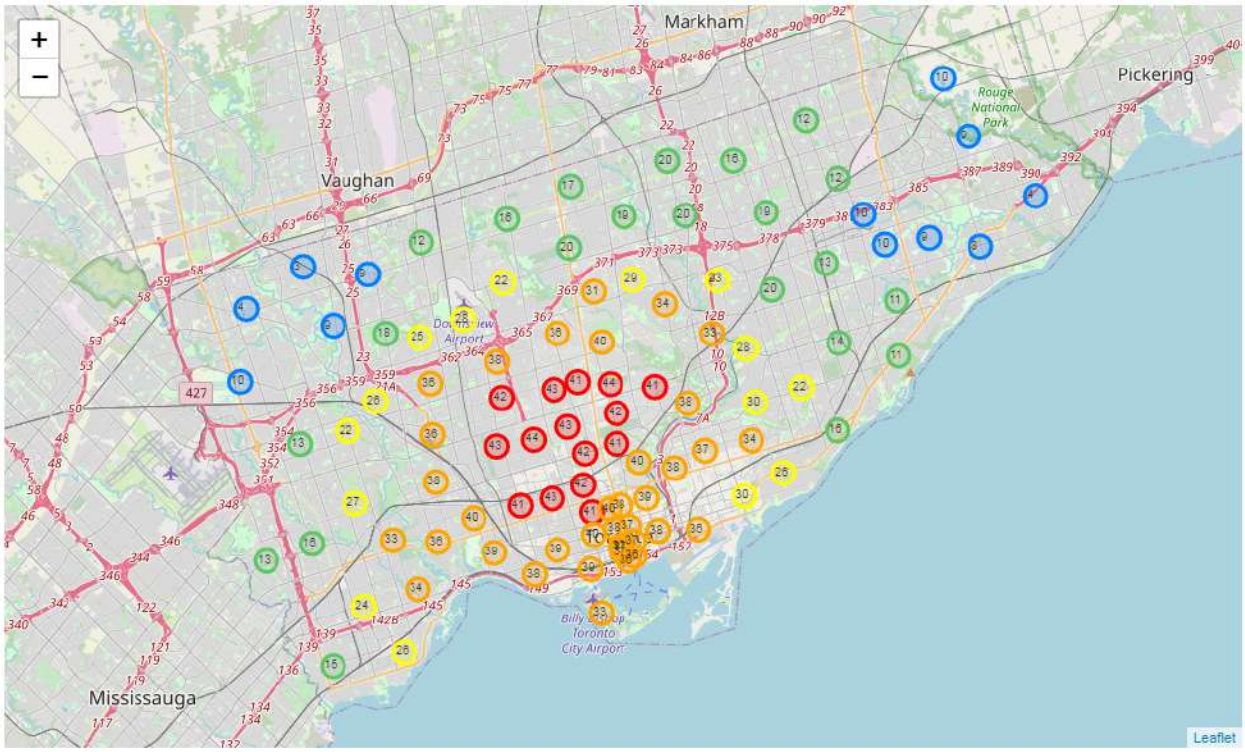

Figure 7: venue count by FSA on the map of Toronto

The map uses colour coding for easy identification of areas with different numbers of veterinary clinics across FSAs in Toronto:

red: 40+
orange: 31 – 40
yellow: 21 – 30
green: 11 – 20
blue: 0-10

The map indicates that the highest concentration of veterinary clinics is in the central areas of the city (red). The concentration gradually declines (yellow, green, blue) as we move towards the city boundaries. Another interesting fact indicated by the map is that the counts are somewhat lower in the Downtown core (orange). Further analysis will reveal that there is a significant lack of correlation between population and clinic count, as the Downtown is primarily a business area with low number of residential units.

The following is the dataframe where population and venue count by FSA are brought together for the first time in the process:

	PostalCode	VenueCount	Latitude	Longitude	Population
0	M4P	44	43.7135	-79.3887	20039.0
1	M6C	44	43.6915	-79.4307	24596.0
2	M5N	43	43.7113	-79.4195	16610.0
3	M6G	43	43.6683	-79.4205	32086.0
4	M5P	43	43.6966	-79.4120	19423.0

Figure 8: df_fsa_population_venue_count

The dataframe is further expanded to include population per venue ratio, and the population per venue ratio normalized using ‘min-max’ normalization method.

	PostalCode	VenueCount	Latitude	Longitude	Population	PopulationPerVenue	PopulationPerVenueNorm
0	M4P	44	43.7135	-79.3887	20039	455.431818	0.032555
1	M6C	44	43.6915	-79.4307	24596	559.000000	0.039958
2	M5N	43	43.7113	-79.4195	16610	386.279070	0.027612
3	M6G	43	43.6683	-79.4205	32086	746.186047	0.053338
4	M5P	43	43.6966	-79.4120	19423	451.697674	0.032288

Figure 9: dataframe 'df_fsa_population_venue_count' with population/venue ratio

The values in the ‘PopulationPerVenue’ column are visualized in the following boxplot:

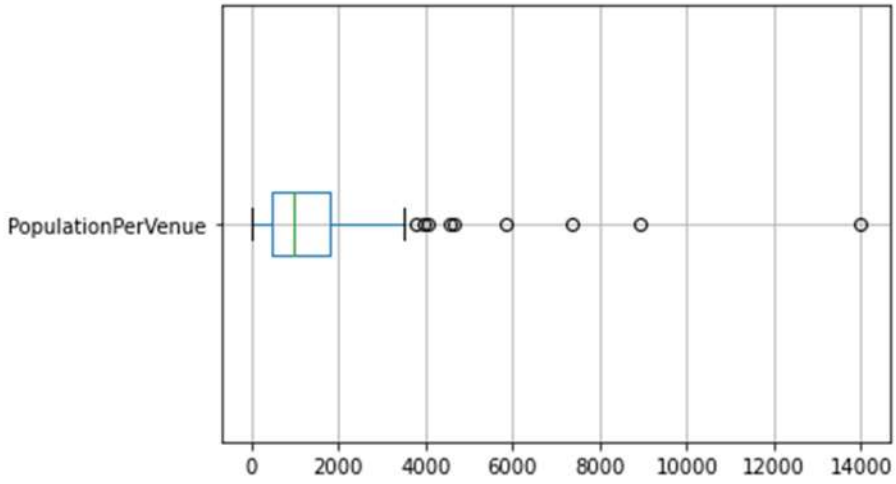

Figure 10: Population / Venue boxplot

The boxplot strongly indicates tight distribution of values around the median value with only a few outliers. This is a strong case that the outliers could be the focus of the business strategy. They could point the city areas with higher-than-average disproportion between population and veterinary clinics.

The following is an attempt to make use of the normalized population / venue count ratio:

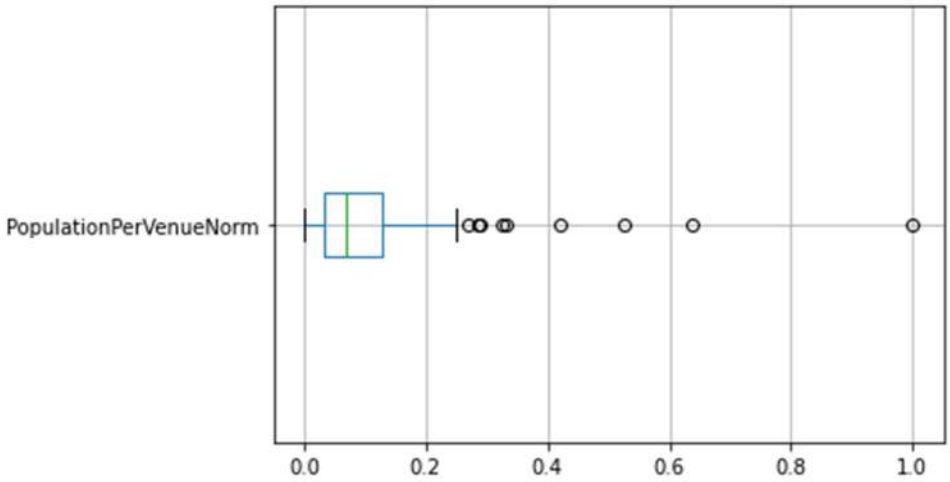

Figure 11: Population / Venue boxplot (normalized values)

There are no differences between the two boxplots. The shape is the same as is the distribution of outliers. This is expected as both sets of values (population and venue count) are normalized. As a consequence, the ratio between normalized values does not differ from the original ratio between the original (i.e. non-normalized) values.

We will now implement another machine learning technique: linear regression. Let us create a scatter plot using two columns: Population and PopulationPerVenueNorm.

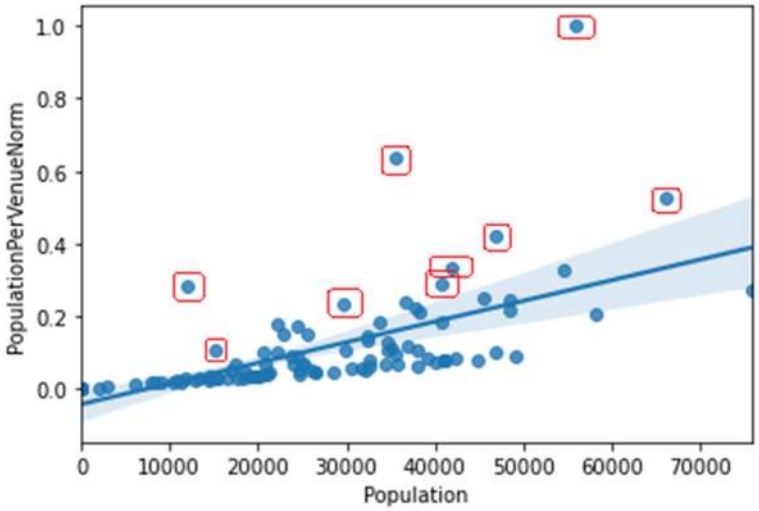

Figure 12: Linear Regression Plot with Outliers Outside of the Shaded Area

The model used in this example is linear regression with default value of 95% for confidence interval. Some of the points are below the regression line but the deviation is relatively small. Much more interesting are the outliers above the regression line; some of the outliers are high above the expected value. These points will be shown in the following illustrations.

Finally, once the final version of the 'df_fsa_population_venue_count' dataset is plotted on the map, the areas where the disproportion between population and venue count is the highest are clearly visible:

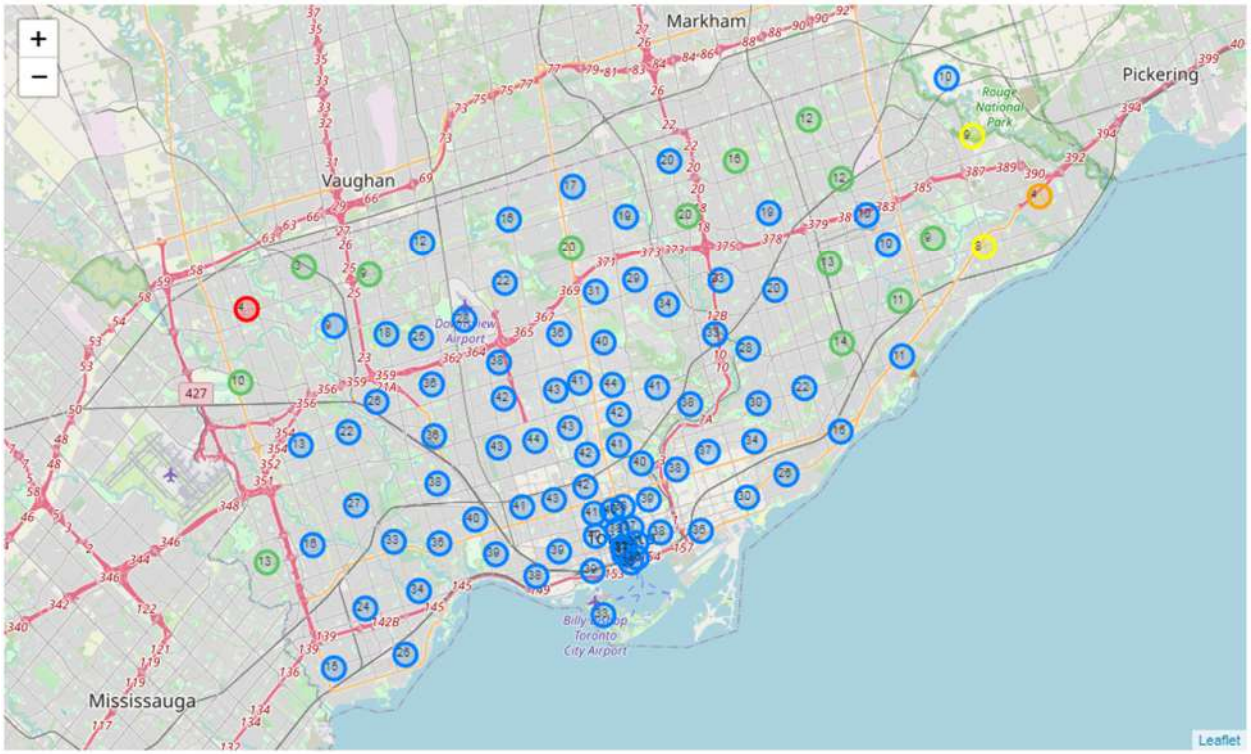

Figure 13: dataset 'df_fsa_population_venue_count' with colour coded population / venue count ratio

The map uses colour coding for easy identification of areas with different values of the population / venue count ration umbers of veterinary clinics across FSAs in Toronto:

red: 40+
orange: 31 – 40
yellow: 21 – 30
green: 11 – 20
blue: 0-10

5. Discussion

The results have revealed that the existing distribution of veterinary clinics across Toronto is more or less even, with only a few areas where the ratio suggests that the large local population may benefit from more veterinary clinics in the area. section where you discuss any observations you noted and any recommendations you can make based on the results. Ultimatelly, the dataframe sorted by the 'PopulationPerVenue' column in descending order shows the optimal candidate:

- FSA 'M9V' Etobicoke located in the north-western part of the city has the population of 55,959 and features only 4 veterinary clinics: 13,989.75 ratio.

The other two close candidates are:

- FSA 'M1C' with 4 clinics serving the population of 35,626 (8906.5 ratio), and

- 'M1B' with 9 clinics serving the population of 66,108 (7,345.33 ratio).

Both FSAs are located in Scarborough, on the far east of the city.

	PostalCode	VenueCount	Latitude	Longitude	Population	PopulationPerVenue	PopulationPerVenueNorm
100	M9V	4	43.7432	-79.5876	55959	13989.750000	1.000000
99	M1C	4	43.7878	-79.1564	35626	8906.500000	0.636645
96	M1B	9	43.8113	-79.1930	66108	7345.333333	0.525051
98	M1E	8	43.7678	-79.1866	46943	5867.875000	0.419441
97	M3N	9	43.7568	-79.5210	41958	4662.000000	0.333244
87	M1V	12	43.8177	-79.2819	54680	4556.666667	0.325715
91	M9W	10	43.7144	-79.5909	40684	4068.400000	0.290813
101	M9L	3	43.7598	-79.5565	11950	3983.333333	0.284732
68	M2N	20	43.7673	-79.4111	75897	3794.850000	0.271259
82	M1P	13	43.7612	-79.2707	45571	3505.461538	0.250574
81	M1K	14	43.7298	-79.2639	48434	3459.571429	0.247293

Figure 14: Optimal Location Candidates for a Veterinary Clinic in Toronto

6. Conclusion

The time and resources available for the work on this assignment are not unlimited and so unfortunately the analysis and investigation of this business problem have to be concluded with the above findings. It would be great though if other parameters could be included such as income, education age and probably the most interesting factor: time – how all these values have changed over the years and if any meaningful conclusion could be arrived at if data were available to support such an analysis. Nevertheless, it has been a very interesting work that has revealed the results that would be impossible to guess without the use of data science tools and methodology.