Photo by Joe Caione on Unsplash

Exploring Toronto neighbourhoods – where to open the next Doghotel

CAPSTONE PROJECT

Karin Leisink | IBM Coursera Data Science Introduction | September 2019

0 Preface

This report describes the research I performed for the Data Science Capstone project to complete the Data Science introductory course offered by IBM and Coursera. A problem regarding neighbourhoods is addressed. Solving this problem requires location data. In this report I will outline the problem under investigation and mention the stakeholders and the people who may possibly take an interest in the outcome of the research or in the way the research is performed. The datasets used during the analysis phase will be stated and briefly explained. I will describe the methodology and present the results. My observations and recommendations will also be discussed. A final conclusion will complete this report.

1 Introduction

1.1 Business Problem and Background Discussion

My friend Chelsea, who lives in Toronto, found out that I enrolled in a Data Science course and asked me to help her realise a childhood dream of hers, namely establish a Dog hotel that offers boarding and sitting services to dogs. It means dog owners can bring their dog for a day or a couple of days on occasions where they are not able to take care of the dog themselves. It enables them to have the hospital treatment, do the home renovation or go on holiday, knowing that their dog is in good hands.

Now, Chelsea is wondering whether there is a neighbourhood in Toronto where she could open the very first dog hotel. And not only that, she hopes that the surrounding neighbourhoods are also deprived of any doghotels. Chelsea has given me two extra requirements:

- there should be 'enough' potential clients for her in that neighbourhood, that means dogs who are chipped (registered) and vaccinated.
- the owners of the clients ought to be able to afford the service. The cost of living in Toronto is very high [1]. It is estimated you need to earn at least $50,000 a year to meet your expenses. Keeping a dog is not cheap ($2,000 a year provided you have a healthy dog) and the high quality dog hotel service which Chelsea is going to offer, will cost around $40 for a stay of one day and $60 for an overnight stay. Chelsea expects that dog owners, who receive a yearly income of $70,000 or more, are willing to spend that amount of money for their dog's stay in the hotel.

The following stakeholders and interested parties can be identified:

1. My friend Chelsea who wants to establish a dog hotel

2. People who would like to establish a dog hotel (preferably not in Toronto just yet, let Chelsea set up a business first)
3. Future Data Scientist for whom this research might be an interesting case study.
4. My colleagues who are software engineers and take an interest in the domain of Data Science and in the Data Science Introductory course that I am taking. This research gives them an idea of what the course is about.

1.2 Datasets and usage of data

Nr	Dataset	Source
1	List of Toronto postal codes with corresponding Neighbourhood and Borough names	Scraped from a wikipage[2]
2	List of Geospatial Codes of Toronto neighbourhoods.	Supplied by Coursera[3]
3	List of registered cats and dogs in Toronto (2019)	Toronto Open Data[4]
4	List of neighbourhood profiles in Toronto from the 2016 census	Toronto Open Data[5]
5	List consisting of postal codes and Neighbourhood names to link the neighbourhoods mentioned in the neighbourhood profiles list to the scraped Toronto postal codes	Self-constructed with the help of wordpostalcode.com [6]
6	List of dog boarding and dog sitting services in Toronto.	Self-constructed with the help of the google search engine[7].

Some remarks have to be made about the acquisition of the datasets and the processing of the data:

- The postal codes were scraped from a webpage with the help of the BeautifulSoup4[9] package. Postal codes without an assigned borough and neighbourhood were discarded. Postal codes without an assigned neighbourhood received the borough name as neighbourhood name.
- The cats were omitted from Toronto open data dataset with the registered cats and dogs.
- The list with the 'Neighbourhood data from Toronto open data' contains lots of data about the inhabitants of the various neighbourhoods. I am only interested in the annual income data of households (income data of individuals were also given, but dogs are kept in a household and households can have double incomes, hence the household income data were used). In the dataset the income data are divided into income categories like the following:

  Under $5,000

  $5,000 to $9,999

  $10,000 to $14,999

  $15,000 to $19,999

  .....

$60,000 to $69,999

$70,000 to $79,999

……

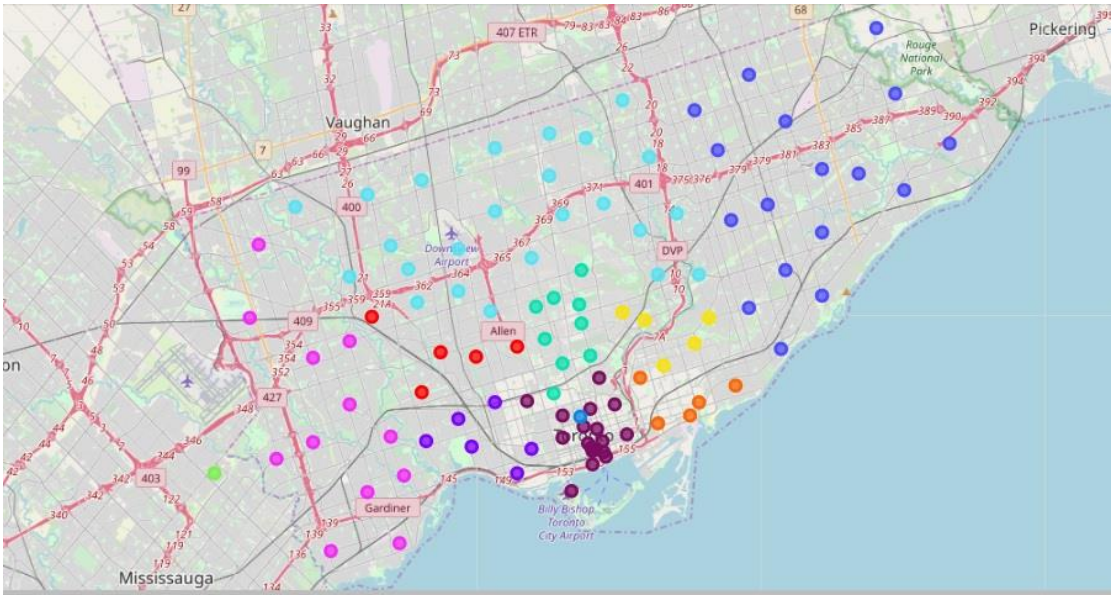
$200,000 and over

For each category and each neighbourhood the number of households are given. At first I kept all the household income columns. But in order to calculate the percentage of households with an income of $70,000 or more, I first calculated the total number of households per neighbourhood. Then I added up all the households in the categories of $70,000 to $79,999 and the categories above. With these two totals the percentage of households with an annual income of at least $70,000 could easily be determined:

*(1 / total amount of households) * total amount of households earning at least $70,000*

- The neighbourhoods in the dataset with the Toronto neighbourhood data differ from the 'scraped' neighbourhoods and sadly, no postal code was given. In order to join the two datasets I produced a postal code with every neighbourhood name in the dataset with the Toronto neighbourhood data. The neighbourhood names with the looked up postal codes were stored in a dataset.
- Initially I wanted to use Foursquare to retrieve the dog sitting and dog boarding venues [8]. However, dog sitting and dog boarding are not a specific venue category with Foursquare. The venue category 'pet service' comes closest. I retrieved the pet service venues for every Toronto neighbourhood, but the venue name only could point out whether it was a dog sitting or dog boarding facility. With names like 'My Pet Food 'N More', 'Paws & Affection', 'Velvet Paws', it is just not clear what their business is. If I want confirmation I should look up these venues on the internet. After filtering the dog carers I had 9 venues. In a city of 2,5 million inhabitants and over 50,000 registered dogs, there have to be more venues. So I decided to look up all the dog sitting and dog boarding facilities in Toronto with the help of a search engine [7]. I decided to ignore the venues that I retrieved with Foursquare and use the results of my thorough search on the internet to construct my own dataset with dog boarding and dog sitting venues.

With the scraped neighbourhood data, the geospatial codes and the Folium library, the area of activity can be shown in a map with the neighbourhoods depicted as circles. Neighbourhoods belonging to the same borough have identical colours.



Now that all the datasets are collected, the data can be analysed.

2. Methodology

First I will show the base dataset on which all the analysis is performed. Then I will produce a few scatterplots to explore the relationship between the numeric columns in the base dataset. Next I will segment the data using the K-means clustering technique. The obtained clusters will be visualized on a map. Potential dog hotel locations will be considered by looking at the map. Remarkable situations will also be discussed

2.1 The base dataset

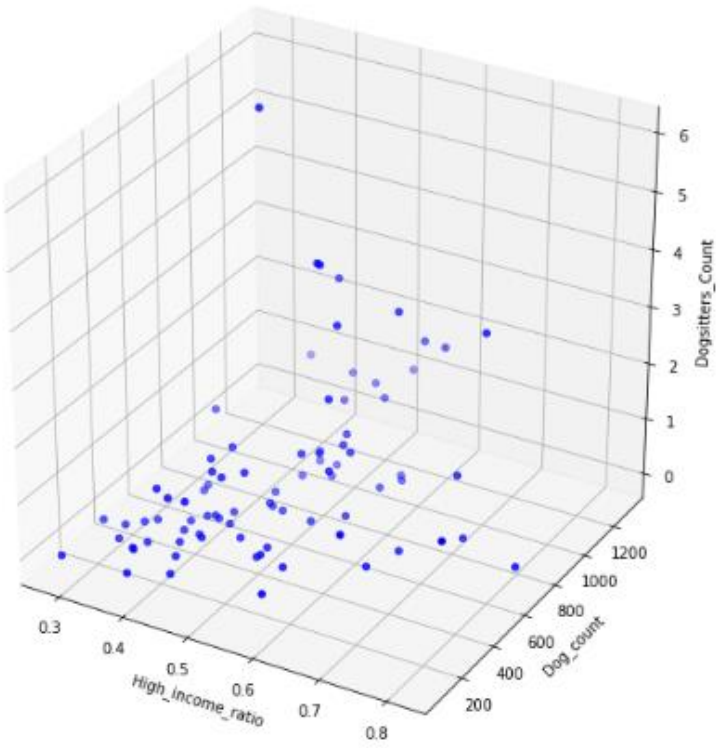
So, let's take a peek at the base dataset. It consists of 82 neighbourhoods.

	FSA	70000_plus_ratio	DOG	Latitude	Longitude	Dogsitters_Count
0	M1B	0.552097	627	43.806686	-79.194353	1.0
1	M1C	0.722772	775	43.784535	-79.160497	0.0
2	M1E	0.453927	963	43.763573	-79.188711	0.0
3	M1G	0.387412	385	43.770992	-79.216917	0.0
4	M1J	0.342124	398	43.744734	-79.239476	0.0

The column '70000_plus_ratio' contains the percentage of household with an annual income that amounts to 70,000 dollars or more. I will also use the name 'high_income_ratio' to refer to this column The column 'DOG' contains the number of registered dogs in the neighbourhood. As you can see, not every neighbourhood has a dog sitting service.

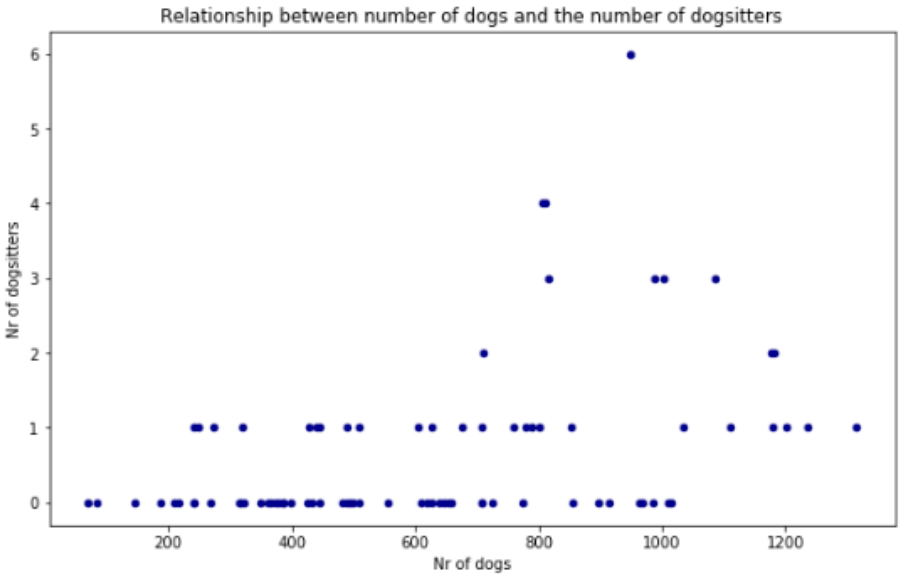
2.2 Explore relationships between columns by using scatterplots

The 3 numerical columns, 70000_plus_ratio, DOG and Dogsitters_Count are visualised in a 3D plot.

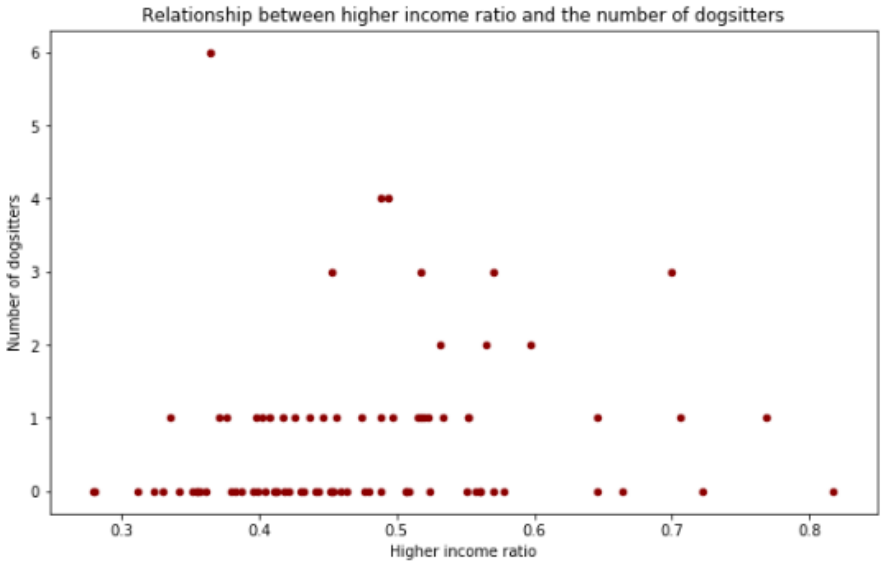


The 3D scatter plot is quite hard to read. There is an outlier visible on the dogsitters axis. It is not clear though where this outlier is located on the High_income_ratio axis and on the dog_count_axis. There are some more neighbourhoods with more than one dog sitting facility, but again for those neighbourhoods I can't say what the related income or dog count is. Let's take a look at the 2D scatterplot of the following relations:

1. the amount of dogs and the amount of dog sitters
2. the number households with an annual income of more than $70,000 and the amount of dog sitters



In a few neighbourhoods, the venues that offer dog sitting services increases when the number of dogs is over 700. But it must be said there are neighbourhoods with the same amount of registered dogs that don't have a dog sitting venue at all. What is significant is that in neighbourhoods where the amount of registered dogs is a little over 1000 there is at least one dog minding venue. There is an outlier for a neighbourhood with 950 dogs and 6 dog minding facilities. I presume this is a neighbourhood in a central area where a lot of people work. These facilities might take advantage of the fact that dog owners like to drop off their dog on their way to work. So Chelsea's wish to establish a dog hotel near the homes of the dogs may not be such a good idea. Maybe it is better do establish a service near the working places of the dog owners. We may find out later whether the neighbourhood with the 6 facilities is indeed in a central, working are, when segmentation into clusters is done and the clusters are plotted on a map.

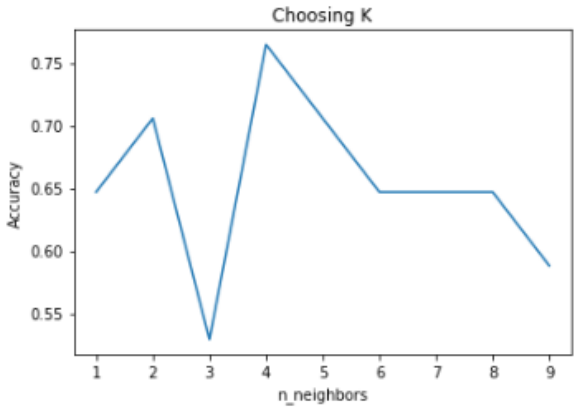


The outlier that we remarked earlier appears to be in a neighbourhood where only 37% of households earn a yearly income of at least 70,000 dollars. But as we should ignore outliers, let's consider the rest of the plot. Half of the dog sitting venues are in neighbourhoods where more than half of the annual household incomes are below 70,000 dollars. Apparently it is not so important for these venues that they are in a neighbourhood with a lot of customers that can afford to take their dog to a dog sitting service. Maybe when this is indeed a neighbourhood at a crossroads, having neighbourhoods around you with a lot of dogs and enough well earning owners will supply enough clientele. But as I said, we will see later. When the majority of incomes is over 70,000 in a neighbourhood (greater than 50%), there is a chance of more than 1 dog sitting venue.

2.3 Segmentation with k-means clustering

2.3.1 Optimal value for k

K-means clustering is used to find out the optimal amount of clusters in the data. The accuracy classification score is used to calculate how accurate the classification is. Accuracy is most optimal with a value of 0,7647058823529411 when k has a value of 4. Below a plot is given with the accuracy score for different values of k.



2.3.2 Segmentation of the data into 4 clusters

K-means is run again to segment the data into 4 clusters. The cluster labels are added to the Toronto base dataset. It is now time to have a look at the characteristics of each cluster and to come up with a name for the cluster.

Cluster label 0

	FSA	70000_plus_ratio	DOG	Dogsitters_Count	Neighbourhood
48	M5J	0.560368	186	0.0	Harbourfront East, Toronto Islands, Union Station
52	M5S	0.426254	249	1.0	Harbord, University of Toronto
53	M5T	0.323024	242	0.0	Chinatown, Grange Park, Kensington Market
55	M6A	0.412477	241	0.0	Lawrence Heights, Lawrence Manor
62	M6L	0.376389	272	1.0	Downsview, North Park, Upwood Park
75	M9L	0.430039	145	0.0	Humber Summit
76	M9M	0.401980	241	1.0	Emery, Humberlea

The above cluster consists of 16 neighbourhoods where dogs are relatively few, approximately 40% of households have an income of at least 70,000 dollars. Few dog sitting business can be found here.

Cluster label 1

	FSA	70000_plus_ratio	DOG	Dogsitters_Count	Neighbourhood
59	M6G	0.487944	812	4.0	Christie
61	M6J	0.493375	807	4.0	Little Portugal, Trinity
63	M6M	0.356486	620	0.0	Del Ray, Keelesdale, Mount Dennis, Silverthorn
66	M6R	0.334792	780	1.0	Parkdale, Roncesvalles
69	M8W	0.517753	815	3.0	Alderwood, Long Branch
71	M8Y	0.517052	708	1.0	Humber Bay, King's Mill Park, Kingsway Park So...
81	M9W	0.476506	639	0.0	Northwest

The above cluster consists of 26 neighbourhoods with a medium dog count. On average 50% of annual households earn at least 70,000 dollars. Most neighbourhoods have zero or one dog sitting facility. There are however 3 neighbourhoods with 3 or 4 dog sitting business.

Cluster label 2

	FSA	70000_plus_ratio	DOG	Dogsitters_Count	Neighbourhood
3	M1G	0.387412	385	0.0	Woburn
4	M1J	0.342124	398	0.0	Scarborough Village
6	M1L	0.397935	510	1.0	Clairlea, Golden Mile, Oakridge
9	M1P	0.411975	500	0.0	Dorset Park, Scarborough Town Centre, Wexford ...
10	M1S	0.436039	428	1.0	Agincourt
11	M1T	0.411034	387	0.0	Clarks Corners, Sullivan, Tam O'Shanter
12	M1V	0.463256	433	0.0	Agincourt North, L'Amoreaux East, Milliken, St...

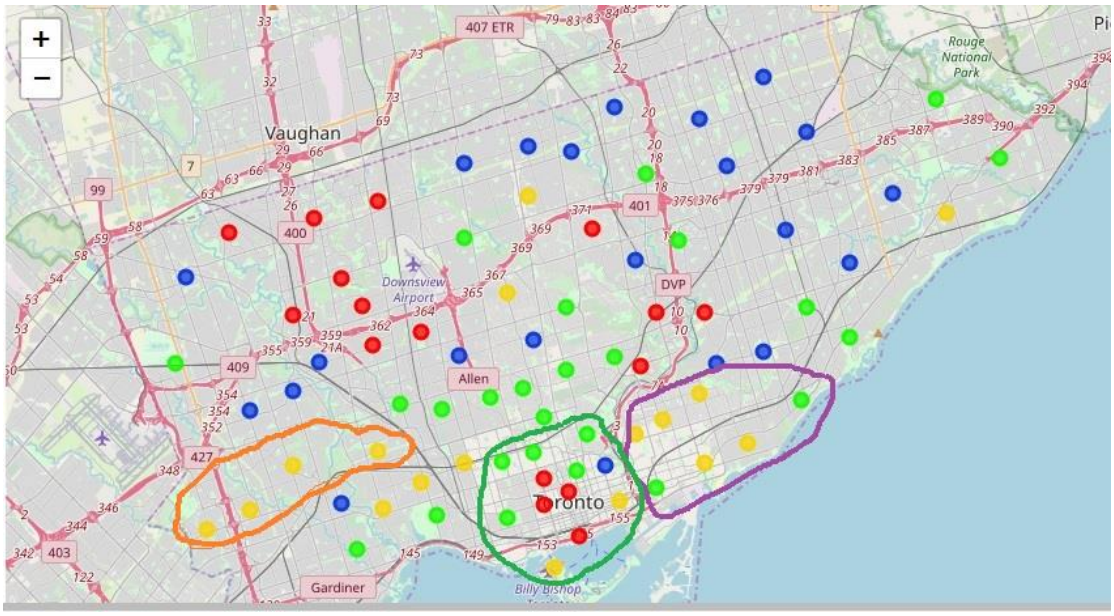
In the above cluster the neighbourhoods have a low to medium dog count. Annual household incomes are in 45% of the cases 70,000 dollars or more. Practically no dog sitters can be found in these neighbourhoods. Of the 22 neighbourhoods, only 5 have a dog sitting facility.

Cluster label 3

	FSA	70000_plus_ratio	DOG	Dogsitters_Count	Neighbourhood
2	M1E	0.453927	963	0.0	Guildwood, Morningside, West Hill
19	M2N	0.474468	1034	1.0	Willowdale South
31	M4C	0.446370	1201	1.0	Woodbine Heights
32	M4E	0.565171	1181	2.0	The Beaches
35	M4J	0.519594	1110	1.0	East Toronto
36	M4K	0.479452	967	0.0	The Danforth West, Riverdale
37	M4L	0.515428	1315	1.0	The Beaches West, India Bazaar

The above cluster consists of neighbourhoods where the dog count is quite high. Approximately 50% of households receive an annual income of at least 70,000 dollars. Two thirds of the neighbourhoods (12 out of 18) have dog business. The neighbourhood with the exceptional high amount of 6 dog facilities is in this cluster. There are 3 neighbourhoods with 3 businesses and 2 neighbourhoods with 2 businesses.

2.4 Visualizing the clusters on a map



Let us summarize the clusters and link the marker colours in the map to the clusters.

- Cluster 0 (red): low dog count, just 40% of households earning $70,000 or more, very few dog sitting businesses.
- Cluster 1(green): medium dog count, 50% of households earning $70,000 or more, - a few neighbourhoods with 3 to 4 dog sitting businesses
- Cluster 2(blue): low to medium dog count,45% of households earning $70,000 or more, very few dog sitting businesses.

- Cluster 2(yellow): high dog count,50% of households earning 70,000 or more, 2 in 3 neighbourhoods have dog sitting business.

The neighbourhoods in the yellow and green clusters are the most attractive neighbourhoods to set up a dog sitting business for Chelsea, provided there are no other businesses established yet.

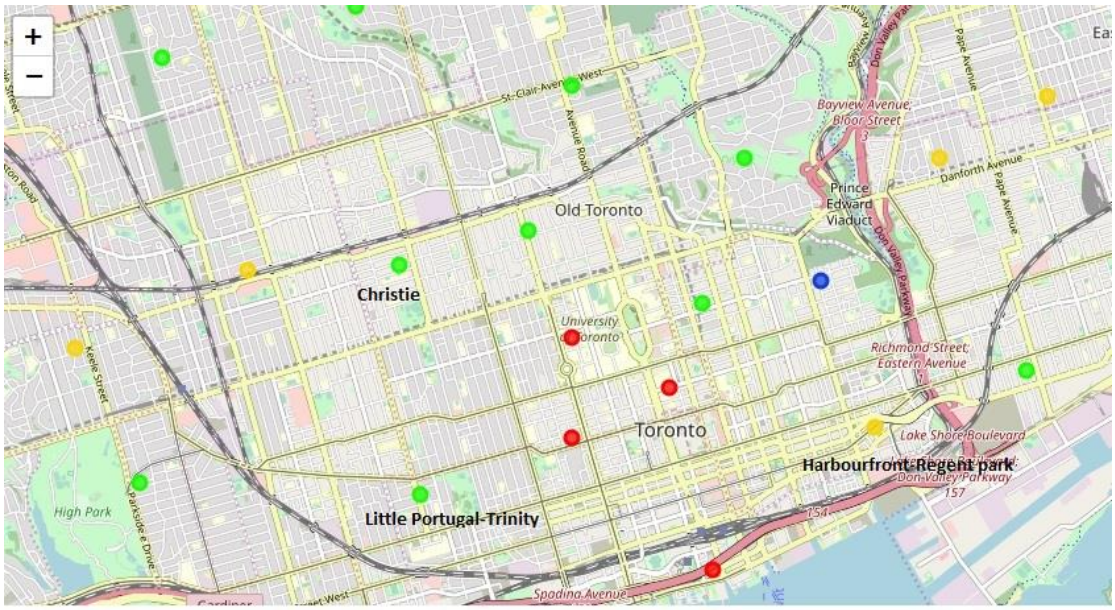
2.5 Investigating interesting areas on the map

I encircled 3 areas on the map that I would like to zoom into and investigate further:

- Green | the centre of Toronto the neighbourhood with the 6 dog sitting venues is located here and there are 2 other neighbourhoods with each 4 business.
- Orange | West of Toronto (area around Islington)| A potential area for Chelsea to establish her dog hotel
- Purple | East York | A Potential area for Chelsea to establish her dog hotel

2.5.1 Centre of Toronto

Below is a detailed map of the centre of Toronto.



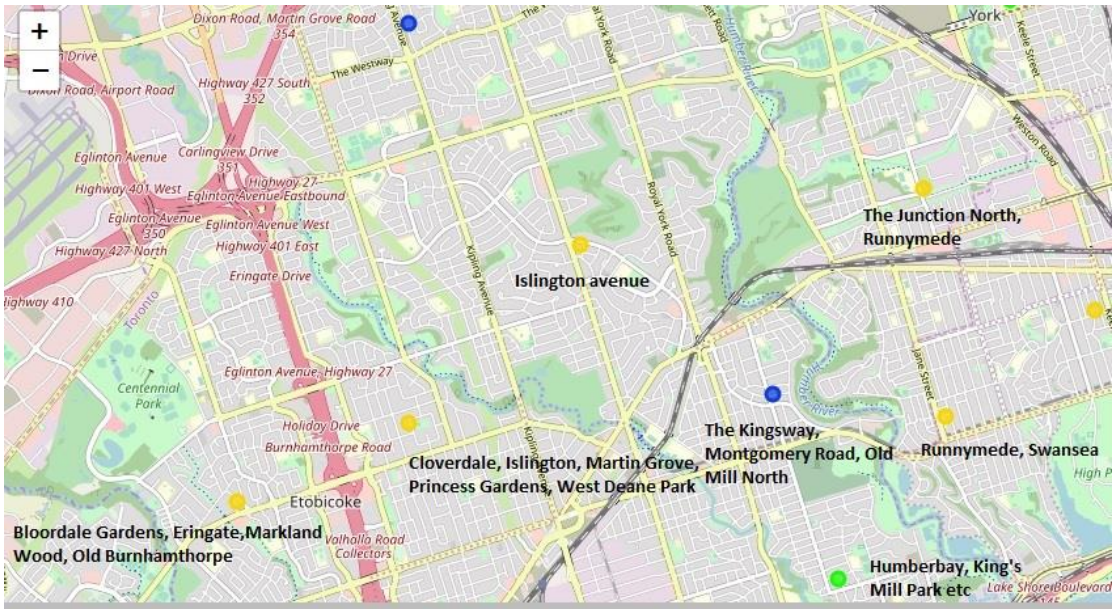
Harbourfront-Regent park is the 'outlier' neighbourhood with the 6 dog sitting businesses that keeps me puzzled and preoccupied. Moreover there are 2 neighbourhoods, 'Christie' and 'Little Portugal-Trinity' with each 4 dog sitting businesses close to each other. What could explain the presence of so many businesses in this area? In the Harbourfront-Regent Park there is a high dog count, but this dog count alone can not provide enough customers for the 6 businesses. It must be said that the centre of Toronto provides a lot of jobs to the inhabitants of Toronto. It may be convenient for dog owners to drop off their dog on their way to work. We also notice on the map that the neighbourhood 'Harbourfront Regent park' is right next to an important central avenue or main road. This suggests that access by car is an

important condition for the inflow of enough clientele. Dog owners don't want to make a big detour dropping off their dog in the morning or picking up their dog in the evening. Here we have 2 extra indicators for finding the best location for Chelsea's dog hotel:

- a high amount of working places
- easy access or short distances from the central avenues or main roads.

2.5.2 West of Toronto (area around Islington)

In area in West Toronto, around the Islington neighbourhood, caught my eye, because there is a chain of 4 yellow coloured neighbourhoods. Have a look at the detailed map.



Yellow coloured markers mean neighbourhoods with a high dog count and half the households living there should be able to afford their dog staying in a dog sitting facility. All 4 neighbourhoods have no dog sitting facilities, so Chelsea may enjoy a substantial number of customers if she sets up business here. On the map we notice that two neighbourhoods are next to a main road. If easy access to main roads is important to attract customers living outside the area than I would advise Chelsea to set up business in 'Cloverdale Islington', because it is next to the main road and in between two neighbourhoods with lots of potential customers.

2.5.3 East York

I want to present at least two options to Chelsea for establishing her dog hotel. So I looked for another area with high potential and found it in East York, the area east of the Toronto centre. Here is a detailed map.

Contrary to west Toronto there are already some dog sitting facilities in the yellow marked neighbourhoods. Nearly all the labelled (named) neighbourhoods have one dog sitting facility. Only 'The Beaches' has two facilities and 'The Danforth West, Riverdale' has no businesses yet. However, I have a great deal of faith in Chelsea's talents for running a dog sitting business. A little competition is good for business and the quality of dog care. In this area she can choose a neighbourhood close to the main road if she wants to attract customers from outside the area. She may also profit from dog owners who are on their way to work in Central Toronto and living in the eastern part of Toronto.

3. Results

By applying the k-means algorithm on the data and thus segment the data into clusters, we got an insight into which neighbourhoods are not such good places for starting a dog hotel and which neighbourhoods could be successful places to start a dog hotel. There is a strong relationship between the amount of registered dogs living in the neighbourhood and the presence of a dog hotel. The relationship between the number of households earning at least 70,000 dollars a year and the presence of a dog hotel is not so strong. The percentage of households earning at least 70,000 dollars a year living in neighbourhoods in the two clusters that are not so favourable for establishing a dog hotel (red and blue) is 40 to 45%. The percentage of households earning at least 70,000 dollars a year living in the neighbourhoods in the other two clusters that are more favourable for starting a dog hotel is only just a little higher, 50%. By investigating an area in Toronto with neighbourhoods that enjoy the presence of a lot of dog sitting facilities (4 to 6), we discovered 2 additional indicators that could proof to be an advantage for attracting customers from outside the area:

1. the presence of a substantial amount of working places in the area. Customers can drop off and pick up their dog on the way to or from their work.

2. The proximity of a main road. The dog hotel is easy accessible for customers from outside the area if it is located near a main road.

In my recommendations for the best location in Toronto to establish a dog hotel I tried to take these newly discovered indicators into account But actually these indicators are worth another, more detailed investigation.

For now I can advise Chelsea the following 2 areas for starting a dog hotel:

1. The area around Islington in West Toronto. The neighbourhood 'Cloverdale Islington' is in the middle of a chain of neighbourhoods with a lot of dogs and practically no dog sitting facilities. Furthermore, a dog hotel in this neighbourhood is easy accessible from the main road.
2. The area of East York. The neighbourhoods 'The Danforth West, Riverdale', 'East Toronto' and 'Woodbine Heights' have a high dog count and are close to the main road. A dog hotel in these neighbourhoods will profit from people coming from the eastern part of Toronto and working in Central Toronto. The few dog sitting business which are already present in these neighbourhoods may turn out to be challenging competitors. My personal opinion is that this competition is good for business and may be very beneficial for the quality of care to dogs.

4. Observations and Recommendations

In this section I will describe the other issues I have noticed and that I haven't mentioned in the Results section.

I observed that the relation between earning a good annual salary and the presence of a dog sitting facility in the neighbourhood is not very strong. Maybe the 70,000 annual income divide is not so well chosen and should be lowered. It could be that dog owners love their pets so much, they are willing to spend a significant amount of money on their care and wellbeing even if they are not earning such a good salary. Possibly they may spend less money on other matters just to be able to keep a dog. The spending habits of the potential customers should be investigated.

We may have discovered 2 extra indicators, but we don't know whether potential customers find it convenient that the dog hotel or dog sitting facility is in the neighbourhood where they live or in the neighbourhood where they work. This is something that needs to be looked into.

Finally I would like to remark that Foursquare is not doing a very good job in registering the pet services in Toronto. There is a large difference between the pet services concerning dog sitting and dog boarding that Foursquare produced and the dog sitting and dog boarding services that I found on the internet. I also think that the Foursquare pet services category should be divided in more fine grained categories like 'pet service to cats', 'pet service to dogs', 'pet service to rodents' etcetera.

5. Conclusion

I have been able to recommend two locations for starting up a dog hotel business based on the initial two requirements Chelsea has given me (enough dogs living in the area and the owners of the dogs should be able to afford the dog hotel service). During analysis two additional indicators were discovered:

- proximity to a main road so that easy access to the dog hotel is provided.
- Proximity of a commuter car traffic route so that dog owners can easily drop off and pick up their dog on their road to and from work.

I would advise Chelsea to get some additional research into the best location for starting a dog hotel in Toronto and focus on these two newly discovered indicators.

I wish Chelsea every success in setting up her dog hotel business. If I can be of any help by doing some additional investigation using Data Science techniques, she can call upon me. I enjoyed very much doing this data science research. I find animal welfare an important issue and I think pet owners should take good care of their pets. I am very pleased to be able to contribute to these issues even if it involves just a small step like finding the best location for a dog hotel. I am sure that Chelsea is going to take excellent care of the dogs that are entrusted upon her in her future dog hotel.

References

[1] Cost of living in Toronto

[2] List of postal codes of Canada: M

[3] Toronto Geospatial codes

[4] List of registered dogs and cats published on Toronto open data (updated 23 July 2019)

[5] List of Toronto neighbourhood profiles from the 2016 census published on Toronto open data

[6] Postal Codes Search by Address, Country, City

[7] Google search engine

[8] Foursquare

[9] Beautifulsoup4