Introduction

A business wants to open up a Pet Store in the city of Toronto but the management is not sure where they should do it. In this capstone, we will try to solve this business problem for the management using the data.

Data

1) A list of postal codes of neighborhoods in Toronto is taken from the Wikipedia page from the URL given below. To scrape this Wikipedia page we will use the "Wikipedia" package.

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2) The link below will provide us with the data on the number of pets (Cats and Dogs) in every FSA (Postal Code). This will help in determining if there is any correlation (Week or Strong) between the population in a neighborhood and the number of licensed pets within the neighborhood.

https://open.toronto.ca/dataset/licensed-dogs-and-cats-reports/

3) The GeoSpatial data for the Toronto Neighborhoods is taken from the link below. This data gives us Longitude and Latitude for every FSA (Postal Code) in Toronto.

https://cocl.us/Geospatial_data'

4) Statistical Canada – This data will give us the total population in every FSA (Postal Code) in Toronto.

https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/CompFile.cfm?Lang=Eng&T=1201&OFT=FULLCSV

5) FourSquare API – From FourSquare we will acquire the details of every PetStore in the neighborhood of Toronto. We will need this data to analyze the number of Pet Stores a neighborhood has. This will be useful to determine if any neighborhoods do not have Pet Stores.

Category ID that we have used is '4bf58dd8d48988d100951735,5032897c91d4c4b30a586d69'

Note: Throughout the report, we will use the term Pet store for the venues that are Pet Stores as well as for the venues that provide Pet Services.

Methodology

We will combine all the data to create a dataset to see whether a Pet Store or services exist in a neighborhood or not. Data related to the number of pets Toronto open data portal was available only for 97 Postal codes, as a result, our entire analysis was done on the same Postal Codes. Data related to other Postal Codes were discarded.
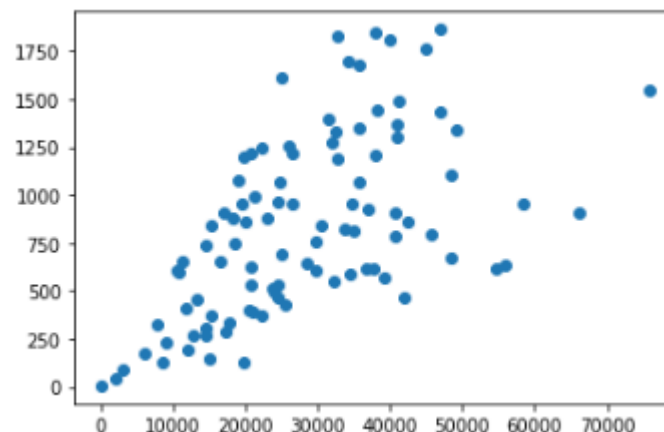
First, we wanted to check if there's a correlation between the population and the number of pets within the neighborhoods that are contained in the Postal Codes. For this, we used the Pearson Correlation coefficient and that came out to be 0.57. The correlation is not strong but is significant and is positive. This indicates that the more the

population within a Postal Code (FSA), the higher the number of pets it should have. Also, as the population increases, the number of pets within a Postal Code (FSA) should increase.

Second, after acquiring the data from FourSquare, it was observed that 22 Postal Codes don't have a Pet Store but they do have a considerable population and the number of Pets. Note that each Postal Codes might contain more than one neighborhood. These 22 FSAs are displayed in the map below:
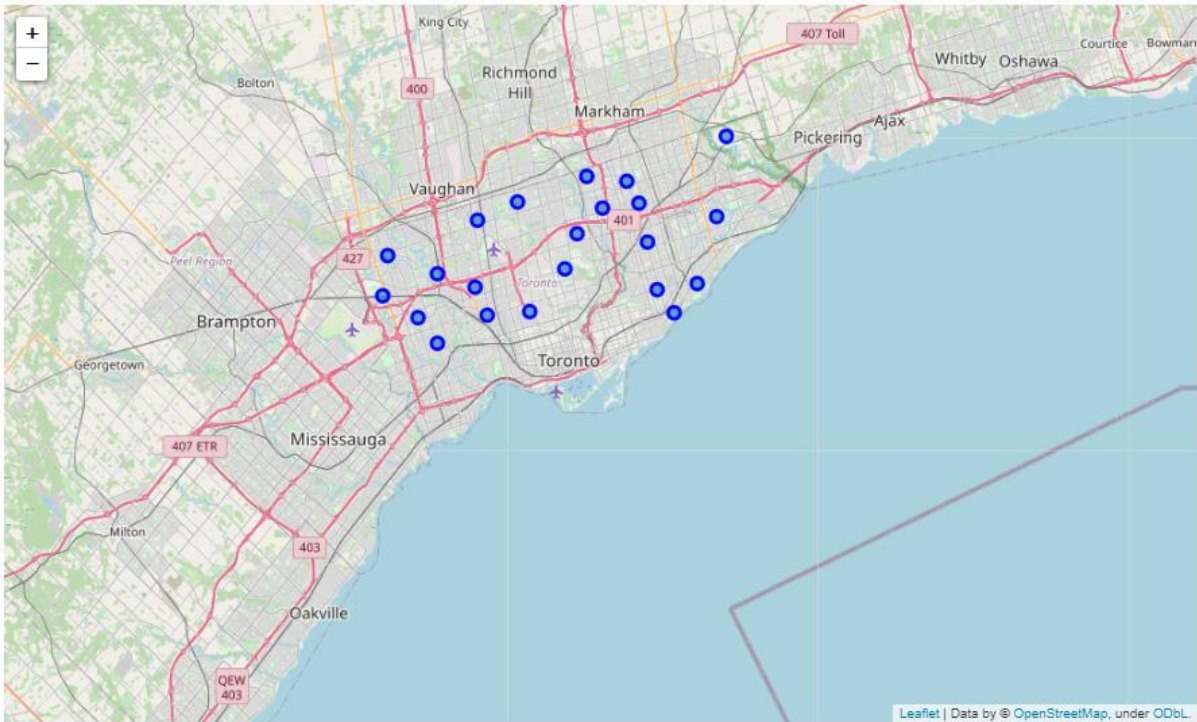
	Postal Code	Population, 2016	Total
0	M2J	58293.0	954
1	M9V	55959.0	636
2	M1W	48471.0	670
3	M6M	42434.0	863
4	M2R	40792.0	789

The Postal Code M2J among these 22 FSAs has maximum population and pets but does not have a single Pet Store. This seems to be an ideal place to open up a store.

But, now the question arises, can the business open up a large Pet Store to serve multiple Postal Codes?

To answer this we analyzed the data set further using an unsupervised Machine Learning algorithm called K-Means to divide these 22 FSAs into clusters. All figures below show pets per population:

	Pets Per Capita
All the FSAs	0.0296
FSAs that has Pet Stores	0.0318
FSAs that does not have Pet Stores	0.0231

|	Total Population	(% of Total Population)		Total Pets	(% of Total Pets)
Total Population	2.73211e+06		Total Pets	80954	
Population in FSAs that has Pet Stores	2.04228e+06	74.751	Total Pets in FSAs that has Pet Stores	64997	80.289
Population in FSAs that does not have Pet Stores	689825	25.249	Total Pets in FSAs that does not have Pet Stores	15957	19.711

The statistical analysis on the FSAs that have Pet Stores and the ones that does not reveals that the Pets per capita are considerably less in the ones that don't have one. Moreover, the FSAs that do not have Pet Stores makeup nearly 25% of the overall population but have only 19.7% of the Pets in the city.

Could this be the reason why other businesses have not to open the shops over there yet?

To understand this, we analyzed and compared the Pets per capita of individual FSAs that do not have Pet Stores to the average pets per capita of FSAs that have Pet Stores using a bar chart.

The horizontal line depicts the Pets per Capita in the FSAs that have Pet Stores.
The figure shows the Pets per capita of the FSAs that do not have Pet Stores.
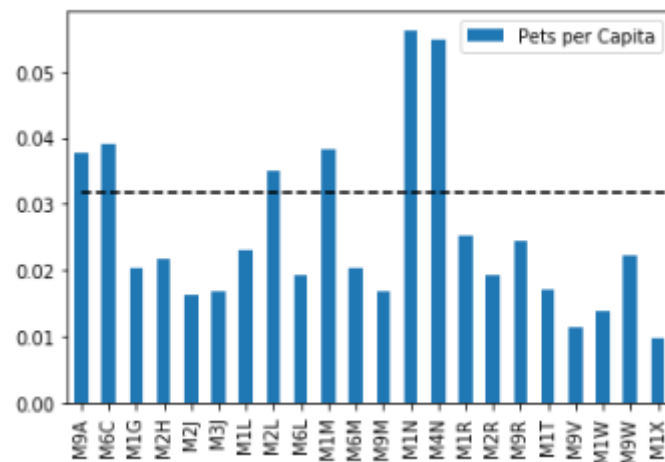

From the chart above we can observe that some Postal Codes like M1N, M4N, M1M, M2L, M6C, M9A have more Pets per Capita than the horizontal line drawn. Our initial analysis of shortlisted M2J should be the place where a Pet Shop should be setup. But the Pets per Capita at Postal Code M2J is way below the horizontal line.

Lots of Postal codes are nearby, what if we open up a store to cater to the needs of neighborhoods in multiple Postal Codes?

To find out even if the above is possible, we will have to divide the FSAs that do not have Pet Stores into clusters. To do this, we chose the unsupervised machine learning algorithm like KMeans.

First and foremost, we normalized the data, specifically, Latitude and Longitude of the 22 FSAs using MinMaxScalar from the Sklearn library. After this, to determine the number of clusters we used the elbow method to determine the Optimal K. To do that, we plotted a graph between Sum of Squared Errors and Number of Clusters (K).
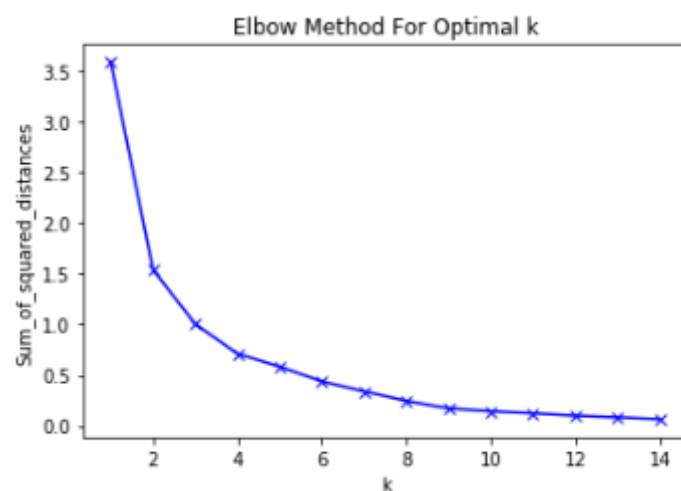
From the graph, we determined the optimal number of clusters to be 8 as after this point the inertia decreases negligibly for the increase in the number of clusters.

Now, as we have the optimal number of clusters that we should create, we will run the K-Means algorithm.
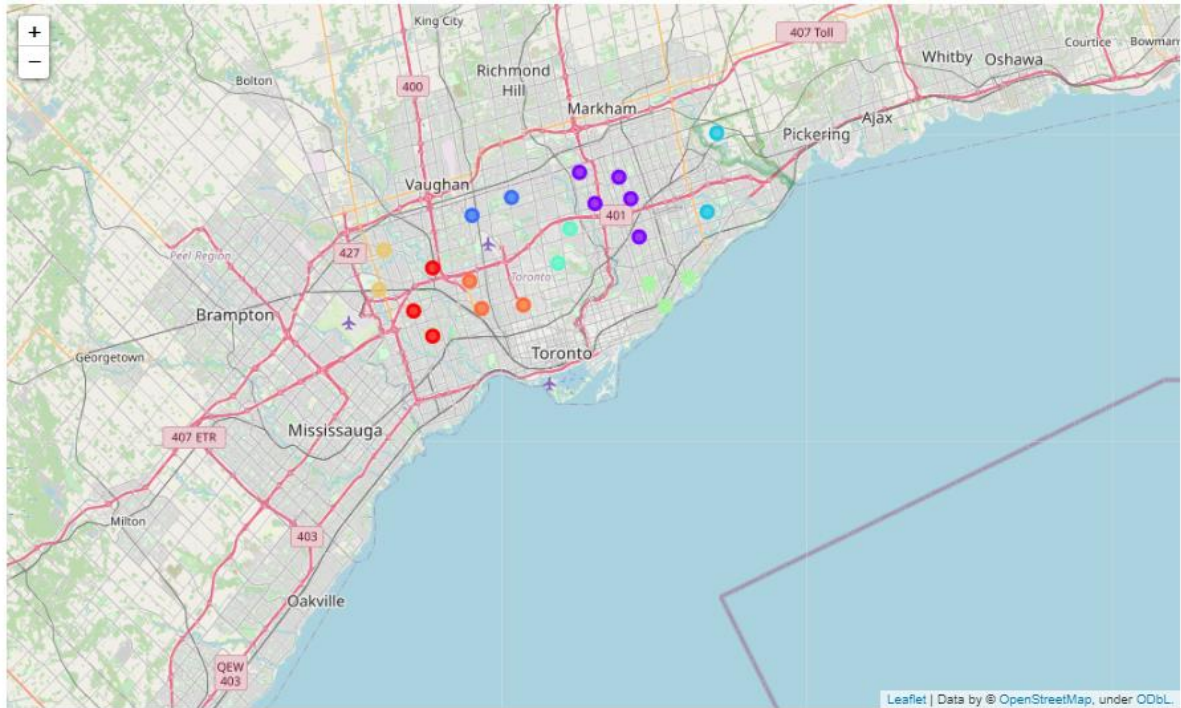
As can be seen from above the FSAs in clusters are close to each other. So, a large Pet Store can be opened up to cater to the needs of multiple Postal Codes.

Now, we need to determine which cluster should the business prioritize to open a Pet Store. For this, we analyze the data to see the maximum population for every cluster and the total number of pets in that cluster.

Cluster Labels	Total	Population, 2016	Postal Code
1	3498	195707.0	M2H,M2J,M1R,M1T,M1W
5	2931	80130.0	M1L,M1M,M1N
0	2545	91600.0	M9A,M9M,M9R
7	2223	87646.0	M6C,M6L,M6M
6	1539	96643.0	M9V,M9W
4	1251	27047.0	M2L,M4N
2	1217	66265.0	M3J,M2R
3	753	44787.0	M1G,M1X

From the table above, it can be seen that Cluster 1 containing Postal Codes M2H, M2J, M1R, M1T, and M1W has the maximum population and number of pets among all the clusters followed by Clusters 5, 0, and 7.
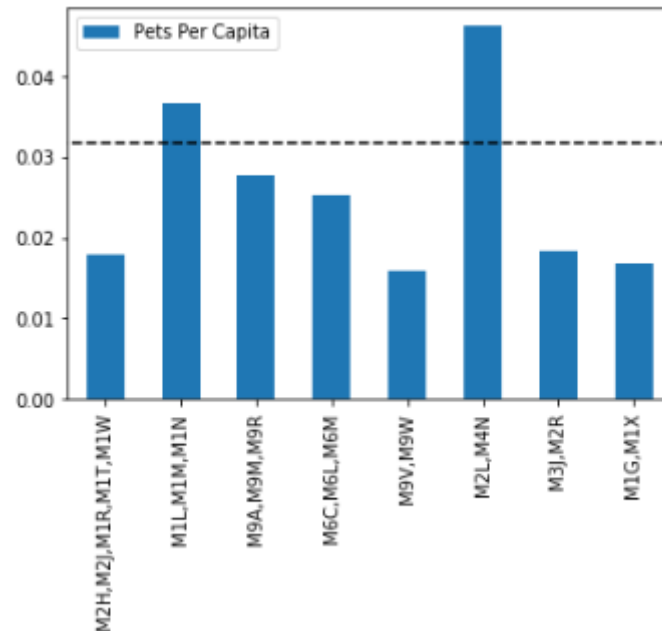
However, if we compare the Pets per capita of different clusters with the average Pets per capita of FSAs that has Pet Stores as in the figure above, we find that the Clusters 4 (M2L, M4N) & 5 (M1L, M1M, M1N) has more Pets per capita than the average followed by clusters 0 (M9A, M9M, M9R) and 7 (M2H, M2J, M1R, M1T, M1W). Cluster 1 has a maximum population and pets, but the Pets per capita in that cluster is significantly less. Clusters 2, 3, 6 neither has a significant number of pets nor do they have higher pets per capita, so, they are not worth further discussion.

Results

1) A positive correlation exists between the total population in an FSA and the total licensed pets (Cats and Dogs) it has.

2) There are exactly 22 FSAs where there are no Pet Stores but these neighborhoods do have a considerable number of pets. Out of these 22, the FSA M2J has the highest population and the maximum number of pets. But if we compare the Pets Per capita of these 22 FSAs with average Pets per Capita of all the FSAs that have Pet Stores, then M1N, M4N, M1M, M2L, M6C, M9A seems to be a better choice than M2J to open a Pet store.

3) After dividing the 22 FSAs into eight clusters using the K-Means algorithm, initially, it seemed that the cluster 1 containing the Postal Codes M2H, M2J, M1R, M1T, and M1W contains the maximum population and pets among all other clusters. If a business can open up a store at such a location between these Postal Codes, then it will experience significantly higher footfall than if it targets a particular FSA. Finding this particular location, however, is beyond the scope of this project.

   However, after comparing the Pets per Capita of different clusters with the Pets per Capita of FSAs that already has Pet stores we observe that Clusters 4 (M2L, M4N), 5 (M1L, M1M, M1N), 0 (M9A, M9M, M9R), 7 (M2H, M2J, M1R, M1T, M1W) might be better choices to open up the store in than in cluster 1 (M2H, M2J, M1R, M1T, and M1W).

Discussion

In my opinion, the business should open its stores in M1N, M4N, M1M, M2L, M6C, M9A, and not only in M2J as Pets per capita matters but the total amount of pets matter too. The business might have to open more stores or larger stores in FSAs with higher population and multiple neighborhoods. As a result, businesses will have to do calculate cost/benefit analysis on its plan to expand business in these areas as the investment required might be higher in terms of capital costs and marketing costs.

Similarly, in the case of clusters, clusters 4,5,0,7 looks better than cluster 1 but again, for cluster 1 the business might have to open multiple stores and that may or might not be the case with other clusters.

Analyzing these factors is beyond the scope of these reports but should be taken into consideration by the business before investing.

Conclusion

We can conclude from our analysis that there are multiple FSAs where the Pet Shops does not exist and any business that wants to set up Pet Stores should start its store there. Also, many of these FSAs are close to each other, so, a single Pet Store can be opened to cater to all these FSAs at a location between them at a Marketplace. But the firm should perform cost/benefit analysis to make decisions on how many stores it wants to open to cover the population and the capacity of these stores. FSAs or clusters that have higher pets per capita can be considered for such analysis while others can be ignored as they are not attractive in anyway. However, it is beyond the scope of this report to comment on any of these factors or analyse them.