Identifying the Best Neighbourhood

Saurav Banerjee

14th June 2020

1. Introduction and Problem Statement:

1.1. Problem Background:

Moving to a new city is always a challenging task for any individual or family. Multiple factors need to be considered based on the specific needs of the individual(s) to make daily life more convenient. While such decisions can be made relative easily via manual observation for a small city, it becomes much harder for a large metropolitan city.

Studies have shown that in today's modern world the majority of people are migrating from small rural cities and towns to large urban metros in search of better job or education opportunities or a more modern lifestyle. This trend brings about a need for a more analytical approach to identifying the best neighbourhoods to move to based on the needs of the individual(s).

1.2. Problem Description:

A metropolis is a large city or conurbation which is a significant economic, political, and cultural center for a country or region, and an important hub for regional or international connections, commerce, and communications. Such rapid growth and expansion attracts a large number of talented individuals from all across the globe.

Let us consider a basic question: Why are they moving there?

The individual(s) could be moving to start a business. This could be in the form of a small business. This leads to the addition of a unique venue to that particular neighbourhood, providing the neighbourhood with an amenity that would fulfil the needs of other individuals. Thus for any individual moving to that city the best neighbourhood to move to would first and foremost be decided based on the basic amenities/services that the neighbourhood can provide for the needs of the individual.

Unfortunately due to the large number of the same types of amenities scattered around such cities. Finding the right neighbourhood with all the required amenities using manual methods is highly time consuming. For example the basic amenities needed by an individual such as myself would be grocery stores, gym, indian restaurants, gas station, park, theatre etc. But these amenities are typically available scattered around the city in large numbers. analysing each neighbourhood would be very tedious.

Additionally there could be amenities that would produce undesirable effects in a neighbourhood. For example bars, clubs, hotels, etc produce a large amount of noise and sometimes neighbourhoods with a large concentration of such amenities could also result in high crime rates. Preferring peace and quiet during the night time when I am most likely to be at home, I would not like to move to such neighbourhoods.

The power of data science will thus be used to make this decision for us.

1.3. Target Audience:

I am the target audience for this analysis since I plan to move to Toronto by the end of 2020.

While this project is primarily motivated by my need to identify the right neighbourhood to move to Toronto, the same approach can be used by any individual who wants to move to any major city to reduce the list of neighbourhoods to consider.

2. Data:

The different data sources and how they are used:

1. Toronto Postal Code Data

First the total list of postal Codes in the Ontario region is extracted using the Beautiful Soup package from the Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This list was then filtered for only the postal codes in Toronto and a list of neighbourhoods is extracted.

2. Latitude and Longitude Data from the Google API

The Google API is used in conjecture with the Postal codes to pull their corresponding Latitude and Longitude information.

3. Four Square API data (and basic amenities list)

The FourSquare API is used with the Latitude and Longitude information as input to pull a list of the venues within each Postal Code. This data is used to perform the K-Means clustering to cluster the neighbourhoods.

This data is also further used to identify the uniques Venues within each cluster thus categorizing them

4. Noise Pollution Data

Additionally Noise Pollution data was used to further make a decision on the ideal Neighbourhood.

Noise pollution levels:
https://www.toronto.ca/wp-content/uploads/2017/11/8f4d-tph-Environmental-Noise-Study-2017.pdf

2.1. Data Clean up

The data was pulled from the Wikipedia page using beautiful soup.

1. All rows with empty Boroughs were eliminated from the list.
2. Each element in the dataframe had a new line character i.e. '\n' because of the formatting of the website table. this was removed.
3. My goal is to do analysis for only toronto hence all boroughs outside of toronto were removed. This was done by eliminating any rows without the word Toronto in the Borough name.
4. We now have a clean list of all Postal Codes, Boroughs and their corresponding Neighbourhoods.

	Postal Code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

2.2. Data Preparation and Exploratory Analysis

To the prepared table of postal codes the geo-location was added to each corresponding Postal code using the google arcgis method.

Using the Foursquare API I pulled the list of venues in the first Postal code i.e. "M5A" to test the functionality of the API and also to view the JSON response in order to parse the necessary information into a table.

Once verified I pulled a list of the closed 100 venues for each postal code using the geolocations as the centre and a radius of 1500 m.

I picked this distance because this is the most comfortable distance which is a either a short walk or a short car drive to any venue within this radius.

NOTE: Since we are taking only the center of the postal code as the point, we need to do a deeper analysis of the prospective chosen cluster with more granular zip codes for each of the residential areas within each postal code for a more in depth analysis. But this is not within the scope of this project.

An analysis was performed to see how many venues were pulled for each neighbourhood.

It can be seen that most postal codes have the max value of 100 venues extracted. The lowest count is for M5J with only 27 venues recorded, which is still good enough for the modelling.

	Postal Cdoe Latitude	Postal Code Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Postal Code						
M4E	100	100	100	100	100	100
M4K	84	84	84	84	84	84
M4L	100	100	100	100	100	100
M4M	85	85	85	85	85	85
M4N	63	63	63	63	63	63
M4P	99	99	99	99	99	99
M4R	100	100	100	100	100	100
M4S	100	100	100	100	100	100
M4T	90	90	90	90	90	90
M4V	100	100	100	100	100	100
M4W	100	100	100	100	100	100
M4X	100	100	100	100	100	100
M4Y	100	100	100	100	100	100
M5A	100	100	100	100	100	100
M5B	100	100	100	100	100	100
M5C	100	100	100	100	100	100
M5E	100	100	100	100	100	100
M5G	100	100	100	100	100	100
M5H	100	100	100	100	100	100
M5J	27	27	27	27	27	27
M5K	100	100	100	100	100	100
M5L	100	100	100	100	100	100
M5N	67	67	67	67	67	67
M5P	100	100	100	100	100	100
M5R	100	100	100	100	100	100
M5S	100	100	100	100	100	100
M5T	100	100	100	100	100	100
M5V	100	100	100	100	100	100
M5W	100	100	100	100	100	100
M5X	100	100	100	100	100	100
M6G	100	100	100	100	100	100
M6H	100	100	100	100	100	100
M6J	100	100	100	100	100	100
M6K	100	100	100	100	100	100

At this point I also pulled a list of all the unique venue categories that were pulled in the list. Using this I was able to select a list of 15 venues that I consider to be of highest importance to be in the 1500m vicinity. This is a very subjective list and can be set by the use as per his preference.

	Venue Category
0	Park
1	Theater
2	Gym / Fitness Center
3	Hotel
4	Breakfast Spot
5	Grocery Store
6	Supermarket
7	Shopping Mall
8	Gym
9	Indian Restaurant
10	Bank
11	Gas Station
12	Street Art
13	Rental Car Location
14	Track
15	Video Game Store
16	Gaming Cafe

One hot encoding was used to first convert the categorical venue data into a quantitative binary data. This table was then grouped by mean to normalize the data to get what is effectively the weighted occurrence of each venue category present in each postal code.

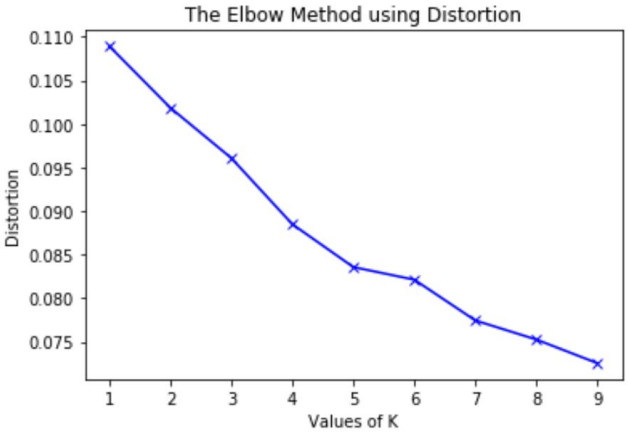
The data is now ready for Modelling.

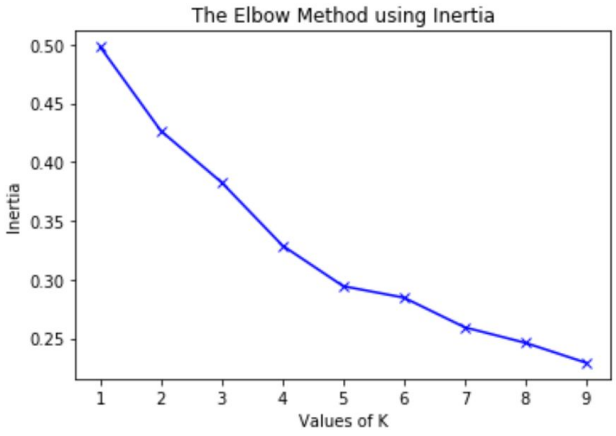
3. Modelling

K means modelling is used to fit the data and cluster the different postal codes based on the venues they contain.

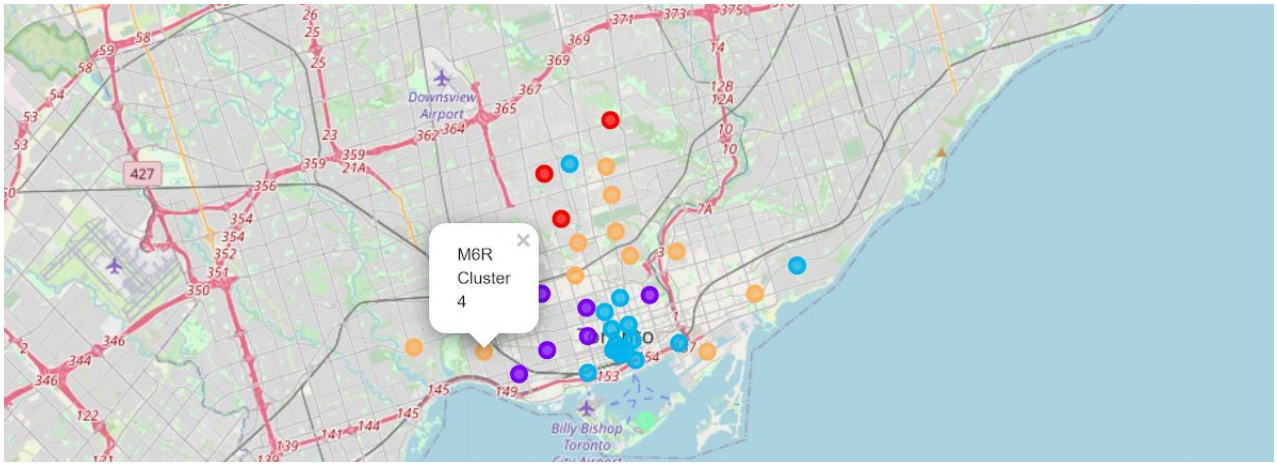
An initial test was made using different values for K i.e. different number of clusters.

The Elbow method was then used to determine the optimal number of clusters.





5 was found to be the optimal number and the postal codes were thus divided into 5 clusters.



3.1. Further analysis to Label the clusters

The different clusters are now further analysed to identify a general pattern of venues present and thus label them for easy interpretation.

3.1.1. Top 10 venues

The top 10 occuring venues in each cluster are listed.

Unfortunately, due to the high concentration of Restaurants and Parks in each postal code in Toronto, the unique features of each cluster is not evident through this analysis. As we can see in the graphs below nearly all the clusters have parks and restaurants as the most occurring venues which is normal in a big city.

Thus an alternative analysis needs to be conducted.

3.1.2. Unique Venues

Since restaurants are the most commonly occurring venues with a lot of listings for the different types of restaurants, any venue with a restaurant in its category name was removed from the listed venues for each cluster.

Further analysis of the unique venues in each cluster (i.e. these venues are not present in any of the other clusters) is also conducted and the resulting analysis has led to the successful labelling of the Clusters and selection of a postal code for further research.

3.1.3. Key Venues analysis

A count of the Key venues in each cluster is considered to reveal the best cluster that contains a high concentration of the venues that are needed by the user.

4. Results and Discussion

4.1. Cluster Labelling

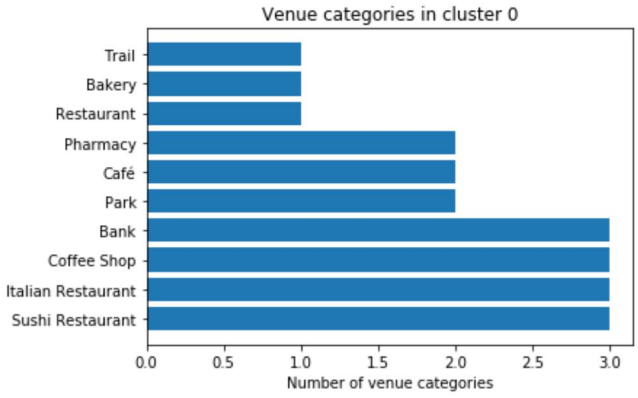
Cluster 0

This Cluster is named as the University cluster due to the presence of Colleges, College facilities, Schools and other venues. For example this Cluster has a high concentration of Fast Food and other restaurants that are typically found in college campuses.

None of the Postal Codes are near the location of my office and hence this cluster is removed from consideration.

- Top 10 Venues Analysis

Inconclusive.



- Unique Venues Analysis

Examples of Unique Venues: 'College Quad', 'College Gym', 'Food Court', 'Skating Rink', 'Tennis Court', 'Field', 'Soccer Field', 'High School'

- Key Venues Analysis

Not Conducted

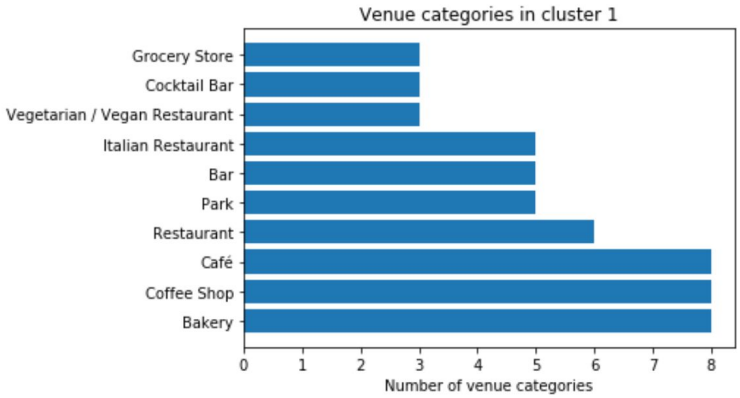
Cluster 1

This cluster is labelled as the Nightlife cluster as it has the highest concentration of Bars, Clubs and other assorted Nightlife related activities.

This was revealed by both the Top 10 venues and the Unique Values analysis.

As areas popular for Nightlife typically have high noise pollution levels in the night and high crime rates this cluster was removed from consideration.

- Top 10 Venues Analysis



Presence of a high number of Bars revealed.

- Unique Venues Analysis

Examples of Unique Venues: 'Cocktail Bar','Wine Bar', 'Indie Movie Theater', 'Comedy Club','Music Store', 'Nightclub','Gay Bar', 'Beer Store', 'Whisky Bar','Dive Bar', 'Beach Bar', 'Supermarket', 'Donut Shop','Flower Shop', 'Hotel Bar', 'Jazz Club','Street Art', 'Smoke Shop',
'Rock Club', 'Performing Arts Venue'.

- Key Venues Analysis

Not Performed

Cluster 2

This is the Business district Cluster and can contain one of the potential Postal Codes for my new house/apartment since the map reveals that it is close to my new office.

The Unique venues analysis prompts the labelling of this cluster.

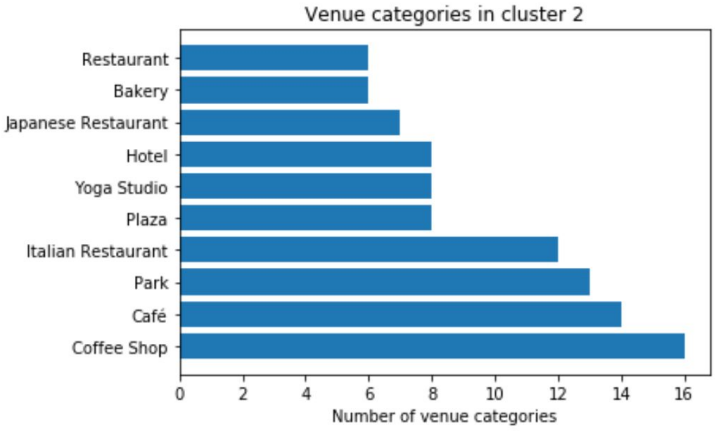
A Key Venue analysis of this cluster reveals that there are a large number of Hotels in this cluster. This means that there is a high traffic of tourists and travellers in this district.

Noise pollution figures also show that there is a high noise pollution level here at all times of the day.

Hence this cluster is removed from consideration.

- Top 10 Venues Analysis

Inconclusive



- Unique Venues Analysis

'Tech Startup', 'Farmers Market', 'Café', 'Circus', 'Coffee Shop', 'Dessert Shop', 'Distribution Center','Coworking Space','Concert Hall', 'Monument / Landmark', 'Train Station', 'Opera House', 'Indie Movie Theater', 'IT Services', 'Roof Deck'.

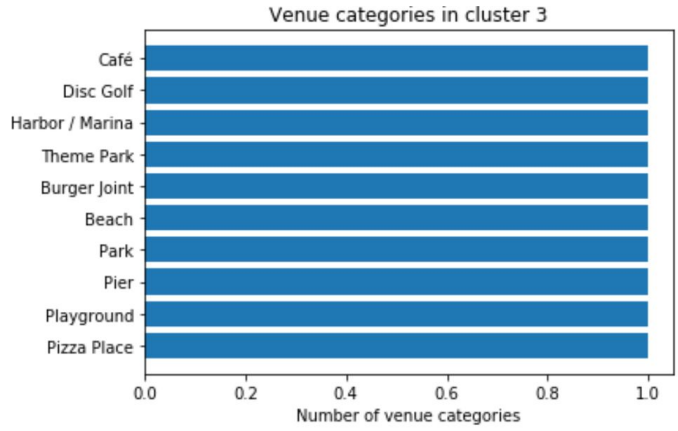
- Key Venues Analysis

	Postal Code
Venue Category	
Park	61
Hotel	33
Gym	23
Theater	22
Grocery Store	17
Gym / Fitness Center	16
Supermarket	15
Shopping Mall	12
Breakfast Spot	11
Street Art	5
Gas Station	3
Track	3
Bank	2
Indian Restaurant	2
Gaming Cafe	1

Cluster 3

There is only one Postal Code in this Cluster and a quick study of the Venues clearly shows this as the Harbour Cluster.

Classifying Venues: 'Park', 'Theme Park', 'Scenic Lookout', 'Beach', 'Pizza Place', 'Harbor / Marina', 'Boat or Ferry', 'Pier'



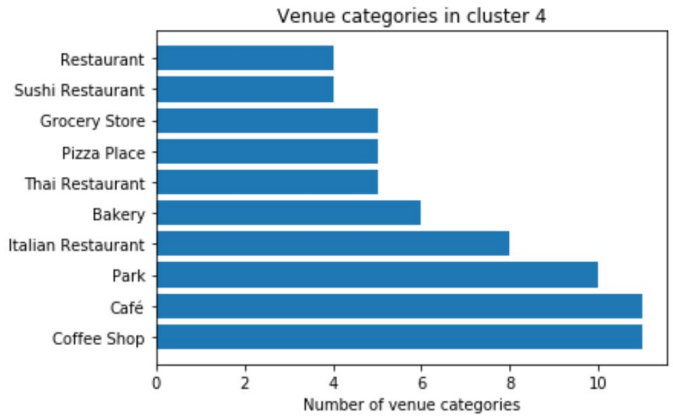
Cluster 4 - The Best Neighbourhood Cluster

After analysing the 5 clusters I have determined that this is the best cluster for me to live in due to the following reasons:

1. Based on the count/spread of the number of key venues.
2. Based on the location of my new office I have selected "M6R" as the best neighbourhood to live in as it is one of the closest Neighbourhoods to the considered location.
3. It is also close to at least one neighbourhood from 3 other clusters, hence easy for me to avail the benefits of the relevant venues based on my needs.

- Top 10 Venues Analysis

Inconclusive



- Unique Venues Analysis

Inconclusive. This indicates that it has a mix of venues from all the clusters.

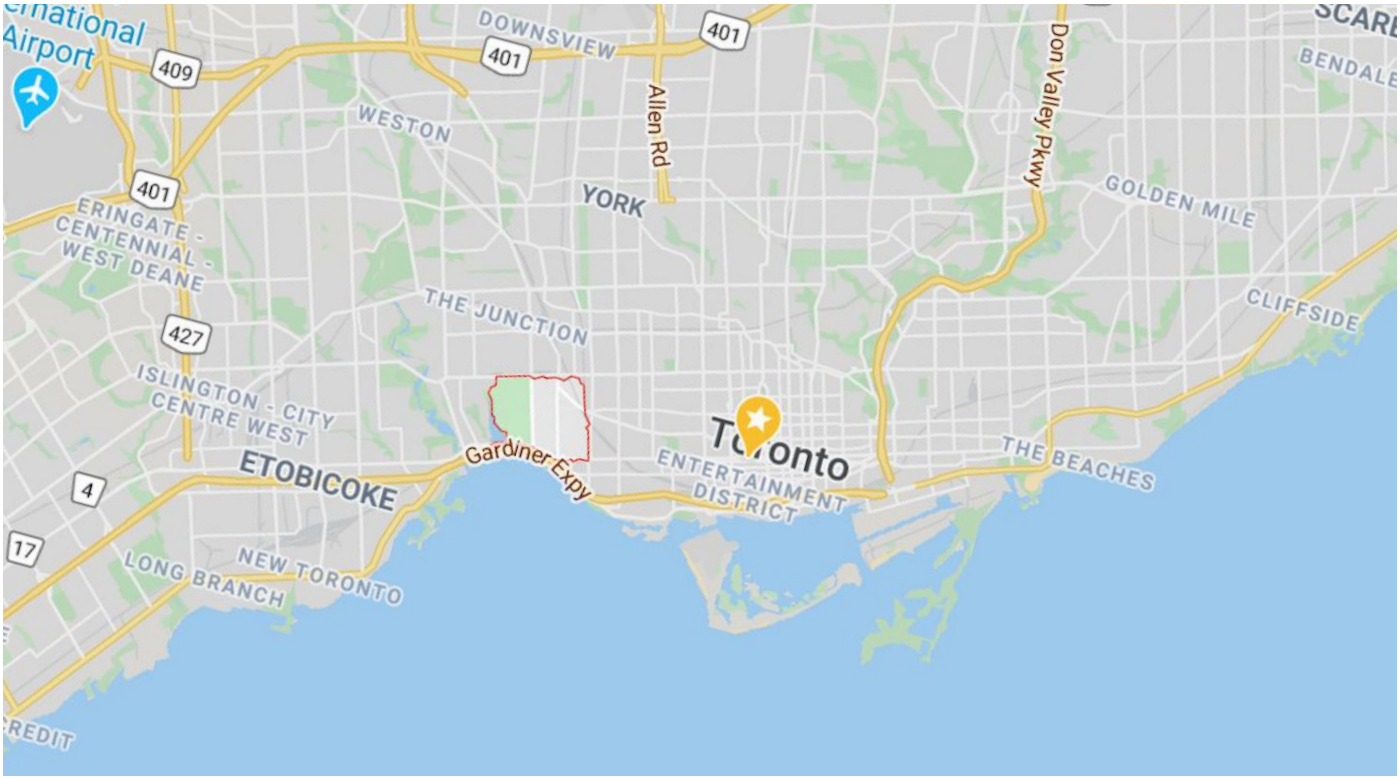
- Key Venues Analysis

It has a good concentration of both the primary key venues such as grocery stores, gym, park etc. and of secondary venues such as hotels and banks and theatres that are needed in low concentrations.

	Postal Code
Venue Category	
Park	52
Grocery Store	23
Indian Restaurant	21
Gym	14
Breakfast Spot	12
Bank	11
Hotel	8
Supermarket	6
Gas Station	4
Gym / Fitness Center	3
Track	2
Shopping Mall	1
Theater	1

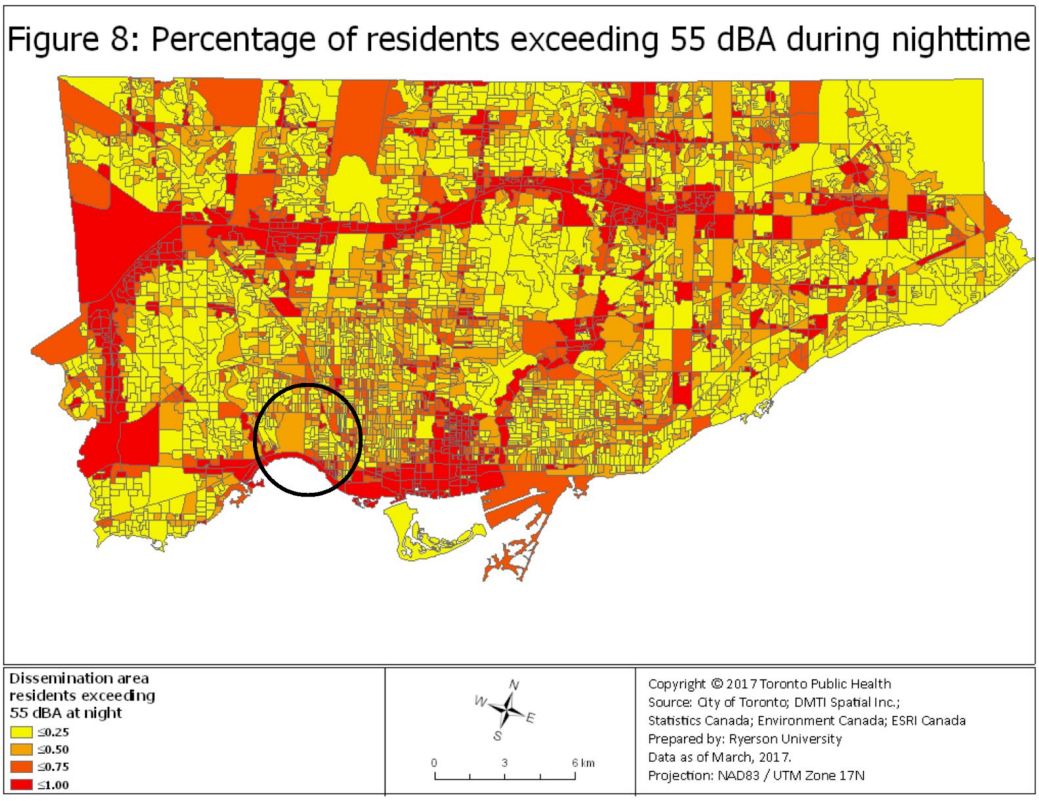
Noise Pollution level Analysis of Postal Code “M6R”

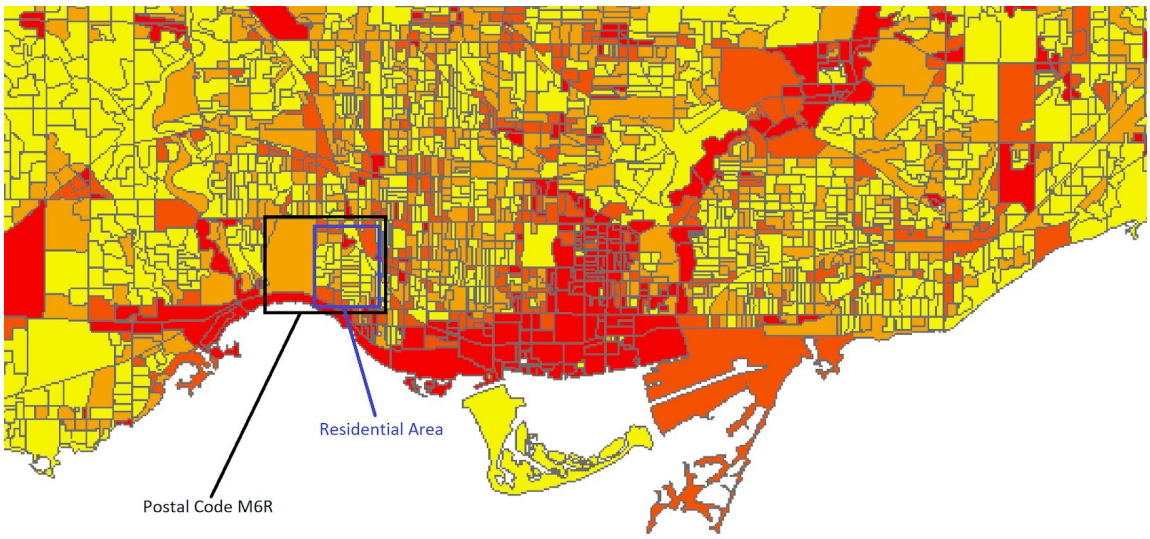
Geo Boundary of Postal Code “M6R” from Google Maps.



Correlating the same information on the Noise Pollution level maps reveals that the percentage of residencies that have noise pollution levels greater than 55 dBA is less than 25% which is the ideal scenario.

Figure 8: Percentage of residents exceeding 55 dBA during nighttime





Future Goals

To this data additional parameters can be used to make a more conclusive decision with further analysis.

Depending on availability of data, in the future variables I would consider are:

1. Renting cost
2. Cost of Houses
3. Location of Public Transportation
4. Air Pollution Levels
5. Climate variables
6. Insect/pest concentrations
7. Crime Rate