Capstone Project - The Battle of Neighborhoods

Helping Newbie in Toronto

Designed and Created by

Naved Akhtar

04-02-2020

1- Introduction:

1.1 Background:

Toronto is one of the best cities to live in. It has great atmosphere and good surrounding with nice people around. Good things comes with price. It has good malls with good business centers and companies. This makes it quite expensive to live. Apart from good environment Toronto has some crimes also, although its quite less than other countries but still a new person wants to check for good area for living. When we say good area obviously it should have less crime, nice environment, less rent along with good to do the business.

1.2 Business Problem:

When everything is available on internet and its very easy to search then why we need to code. The issue is a new person will not know the areas near Toronto. As Toronto is too big to analyze each and every area for crime, rent and business. That's why I come up with a program which will help a new person to give his preference like Max rent, Max crime rate and also which business he wants to open. By this its will be very easy for him to choose among different areas.

1.3 Interest:

As lot of people migrate to Canada due its kind attitude toward migrants, this project will be will be very helpful for any person who have no idea from where to start. Best thing is it will show all the areas over map and graphs by which it become very easy to choose best area.

2. Data Mining:

2.1 Data Acquisition:

To achieve this I have to get crime data and rental data from kaggle. Kaggle is one of the best source of data for data mining.

CrimeData: https://github.com/naveakht/Coursera_Capstone/blob/master/Homicide.csv

Number Of Records: 1016

This file contains all the necessary information regarding crime till 2018. Below are the important columns:

Row in Datacolumn	Desciption
Index_,	Index Number
Occurrence_year,	Year of occurrence of Crime
Homicide_Type,	Crime type
Occurrence_Date	Actual date of crime
Neighbourhood,	Address of crime
Lat,	Latitude of crime address
Long	Longitude of crime address

As Neighbourhood is address and its difficult to get the exact address, I have to Lat and Long to get the actual address and post code.

RentData:
https://raw.githubusercontent.com/naveakht/Coursera_Capstone/master/Toronto_apartment_rentals_2018.csv

This data is used for finding the rental in Toronto area. Below are column names.

Row in Datacolumn	Desciption
Bedroom	Number of bedroom
Bathroom	Number of bathrooms
Address	Address
Lat	Latitude of property
Long	Longitude of property
Price	Price in Canada Dollar

From Address it's very difficult to get the exact address, So I have to Lat and Long to get the actual address and post code.

2.2 Data Cleaning

CrimeData: This data is cleaned keeping in mind that we need to get the result on the basis of 3 word postcode ex: M5J. There are lot of error for which I used excel sheet to correct it.

Records after cleaning: 1016

RentData: This data is cleaned keeping in mind that we need to get the result on the basis of 3 word postcode ex: M5J. There are lot of error for which I used excel sheet to correct it.

Also selecting only the records which has Toronto in its address.

Records after cleaning: 788

Also in both of the datasets I have used Latitude and longitude instead of address. As its best to get the exact address.

2.3 Data Processing and Assumption

Our aim is to process both Crime data and Rental data to group by using postcode.

MIN_BEDROOM=1 #Minimum number of bedrooms in need

MIN_BATHROOM=1 #Minimum Number of Bathroom

MAX_RENT = 1700 #Max Rent

MAX_CRIME= 10 #Max crime

BUSINESS_TYPE='Dog' #Type of business person wants to start

Data processing over Crimedata:

Used geolocator.reverse to get the postcode using Latitude and longitude. Use substring to save only first 3 characters of postcode for easy processing. As it's a big program I

created a new dataframe and save the output including Postcode in github (https://github.com/naveakht/Coursera_Capstone/blob/master/Toronto_CrimedataWithPostcode.csv). I use this file for further processing.

Assumption over Crimedata:

It has been assumed that if there is no record then there is no crime.

After processing I got the data like:

```
In [127]: ShortCodeCrime_df.head()
```

Out[127]:

Shortpostcode	NumberOfCrime
L3R	2
M1B	21
M1C	5
M1E	10
M1G	5

Data processing over Rent data:

Used geolocator.reverse to get the postcode using Latitude and longitude. Use substring to save only first 3 characters of postcode for easy processing. As it's a big program I created a new dataframe and save the output including Postcode in github (https://github.com/naveakht/Coursera_Capstone/blob/master/RentTorontoWithPostcode.csv). I use this file for further processing.

Also I have calculated the mean price if there are more than 1 rental data in a ShortPostcode.

I have recalculated the price for 2 BedRoom and 1 Bathroom using other sizes. For ex: If 1 room + 1 Bathroom price is 1000 then 2 Bedroom and 1 Bathroom will be (1000/2)*3. In this way I get the approx. price even if there is no data for 2 Bedroom and 1 bathroom

Assumptions on Rent data:

Assuming price of apartment has a linear dependency over size of apartment in same postcode.

After processing data I got below output.

```
In [23]: MeanRentalPrice_df.head()
```

Out[23]:

Shortpostcode	PriceFor2Room
M1T	1250.000000
M2J	1927.000000
M2N	1500.000000
M2R	1530.833333
M3A	800.000000

Data Processing Over Business Search:

Using the above Crime and Rental data, I have comeup with the best possible combination by considering MAX_RENT and MAX_CRIME. By using this I got the below dataframe and then I used Foursquare API to find near by business which deal in Dog.

```
In [96]: df_Crime_Rental.head()
```

Out[96]:

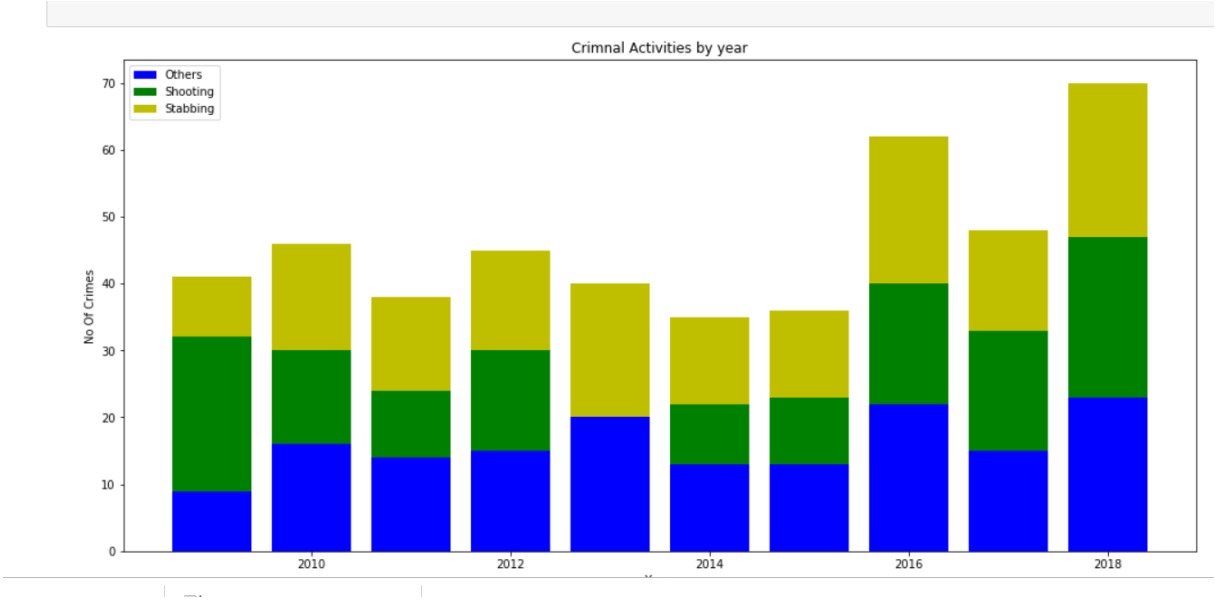
	Shortpostcode	NumberOfCrime	NoOfDogShop	Pr
0	M3A	2.0	0.000000	
1	M6G	8.0	0.758621	
2	M5G	2.0	1.000000	
3	M1T	6.0	1.000000	
4	M4E	4.0	0.517241	

Assumption over Business:

It has been assumed that the best place for the business is where we see less competition. So finding a place where there is less shops regarding the business has been taken as better option.

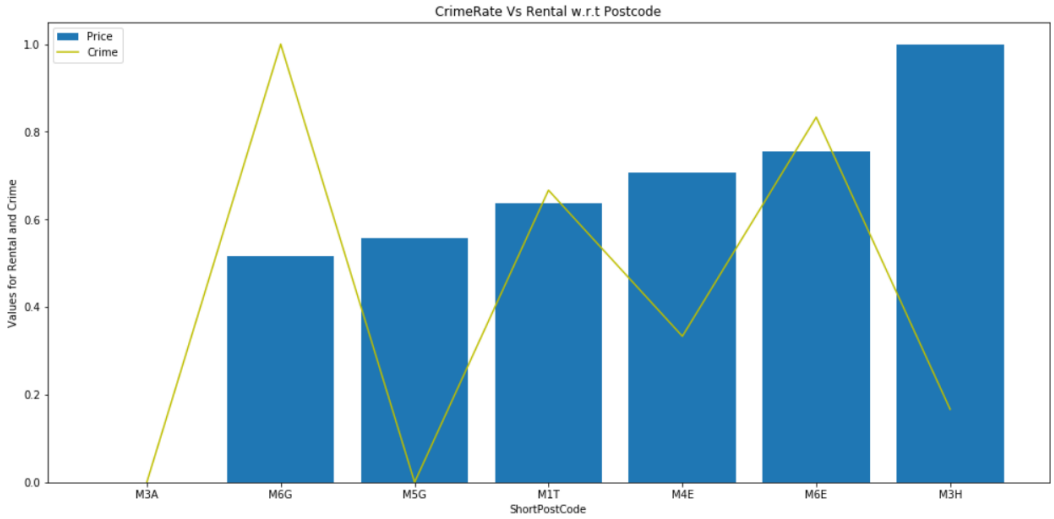
Methodology:

By using crime data over last 10 years it has been found that crime is bit Increased in all the Type (Shooting, Stabbing and Others).



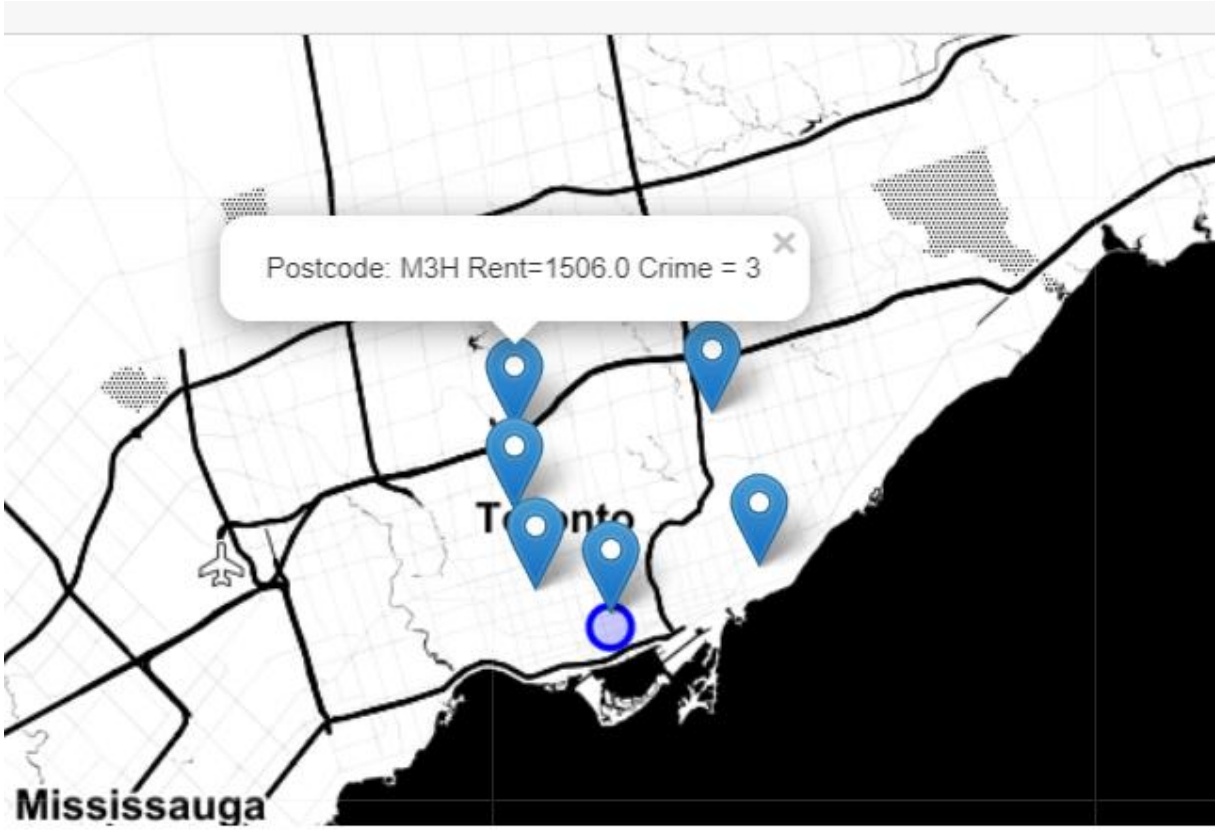
So it's not a gradual increase. It means that there will be some specific events of crimes in few years. And it's increasing gradually. Hopefully it might decrease in coming years.

So map Crime and Rent on a same chart will be a good so that we can see the relation between rent and crime. To do this I normalized both Crime and Rent data and come up below graph.



These postcodes are only filtered postcode using MaxRent and MaxCrime.

Lets maps this postcode over map to see its distance from Toronto Main city.



Blue circle is Main City Toronto. So all the postcode are not too far from the city.

Next is to find the Dog business near these localities. To find this I have used Foursquare API. To get the Latitude and longitude of these 2 postcodes I have again used my dataset and get the cordicates on the basis of postcode. It comes like below dataframe.

```
featureDf
```

Out[37]:

	postcode	Lat	Long
M3A	M3A	-79.325851	43.743374
M6G	M6G	-79.428719	43.669308
M5G	M5G	-79.384895	43.658512
M1T	M1T	-79.315231	43.784492
M4E	M4E	-79.298027	43.678886
M6E	M6E	-79.441109	43.702255
M3H	M3H	-79.440086	43.736111

For using Foursquare api, below parameters has been used:

Creteria: Venues/search

VERSION = '20190605'

Query= BUSINESS_TYPE (Dog)

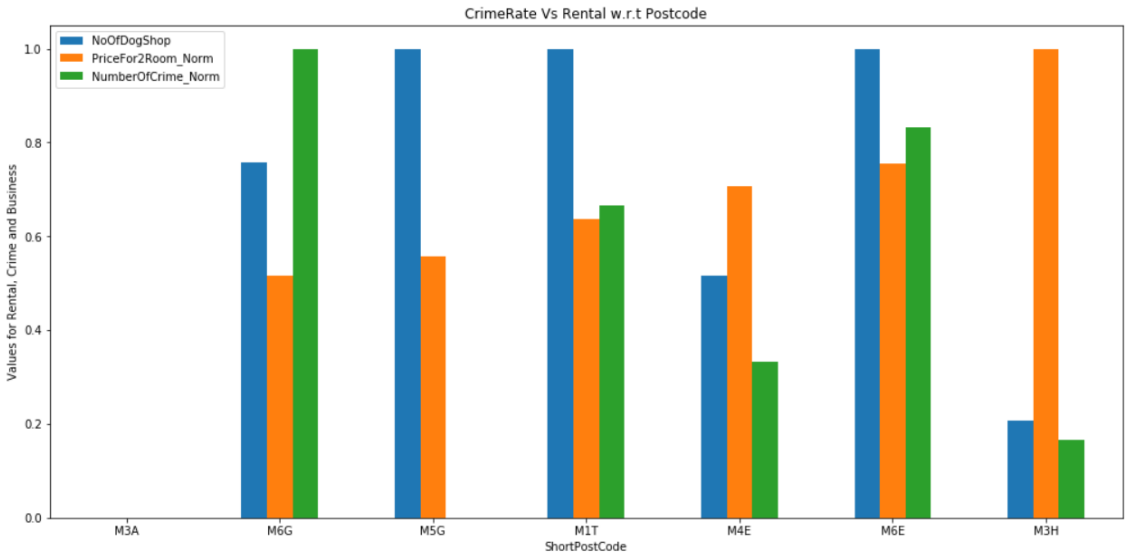
limit=50

This API will give the details of shops which deals with Dog. Using Json format I counted the number business in each area. As to show it on the graph I have to again normalize the Business column in dataframe using MinMaxScalar. I got the below dataframe:

Out[139]:

	Shortpostcode	NoOfDogShop	NumberOfCrime_Norm	PriceFor2Room_Norm
0	M3A	0.000000	0.000000	0.000000
1	M6G	0.758621	1.000000	0.516129
2	M5G	1.000000	0.000000	0.557042
3	M1T	1.000000	0.666667	0.637293
4	M4E	0.517241	0.333333	0.708104
5	M6E	1.000000	0.833333	0.755311
6	M3H	0.206897	0.166667	1.000000

Once I got all the data in place, I put them into graph to see the best combination.



Result and Discussion:

By looking the above graph it can be easily seen that **M3H** postcode is best according to selected criteria. It has rent just below 1500$ which is less than maximum limit of 1700$ also crime rate is much below to 3 which is also less than 10 (Max crime limit). Apart from this the Dog business is very less compared to other localities. Also looking at map it can be easily checked that this area is bit far from Toronto main city.

2nd Postcode will be **M4E**, which has bit more crime rate but less rent and bit more competition in business. But this area is close to Toronto main city.

Conclusion:

I would like to conclude the report by saying that this is report will provide genuine filter criteria on the basis of different factors. As this report is used keeping in mind that crime data and rent data of many localities are not present which is quite hard to manipulate the exact result.