# Capstone Project - The Battle of Neighborhoods

## Helping Newbie in Toronto

Designed and Created by

Naved Akhtar

04-02-2020

# Introduction:

## 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.

## 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.

## 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.

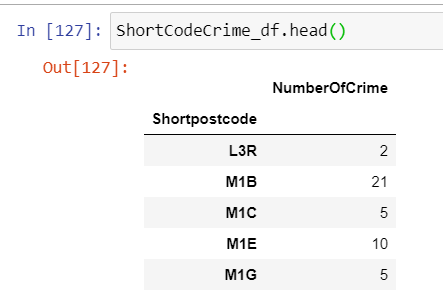
### Data processing over Crimedata:

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:



### 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.

