An Analysis of Neighbourhood Crime in the City of Toronto

Naresh Vempala

1. Background, problem, and questions to be answered:

Toronto is considered to be a safe city in comparison to other big cities. In an article in the Economist (2015), Toronto was ranked as the safest major city in North America and the eighth safest major city in the world, as cited in Wikipedia.

Despite being a relatively safe city, Toronto has its fair share of crime. The city consists of 140 officially recognized neighbourhoods along with many other unofficial, smaller neighbourhoods. As is the case with any big city, some neighbourhoods are considered to be less safe than others. Several reasons are attributed to higher crime – lower income, higher unemployment leading to crime, lower literacy and access to education, among other reasons.

An analysis of crime and neighbourhood data within Toronto will provide us with a good understanding of how many of these assumptions are true and to what degree. It might additionally reveal hidden patterns, trends or relationships between some independent variables and our dependent variable (i.e., major crimes) that would not be obvious at the outset. This was my first motivation. So, the problem may be articulated as – finding major crime trends in Toronto's neighbourhoods, identifying potential independent variables related to major crime, and using these variables to build a prediction model.

My second motivation was to use some of the available open datasets provided by the City of Toronto. This project gave me an opportunity to explore these freely available datasets and find any intrinsic value using a proper data science pipeline of data wrangling, data analysis, data visualization, prediction, and data storytelling.

For this project, my focus was on crime in Toronto's 140 official neighbourhoods. I ignored the unofficial neighbourhoods and explored the following questions and topics.

1. A summarized visualization of all the major crimes in Toronto

2. A comparison of 3-5 most crime prone neighbourhoods against 3-5 least crime prone neighbourhoods

3. Finding the most prominent age group of people in each neighbourhood (does this in any way affect crime?)

4. Finding the neighbourhood with the most change in major crime (What could be the reasons for this change? Can the data give us an answer?)

5. Finally, building a predictive model of major crime in each neighbourhood using machine learning.

6. Furthermore, after addressing #5, I wanted to identify the most salient features/variables used by the model for predicting major crime, within the limitations of my dataset.

2. Potential clients:

There are two different types of clients that could be interested in the findings from this project.

- Canadian/US online and print media that cover socio-economic and urban issues: These clients are magazines that take an active interest in stories driven by socially relevant issues and are backed by data analytics, for creating awareness within the public while simultaneously enhancing the quality of their readership.

  E.g. Canadian magazines such as *The Walrus* and *THIS Magazine*.

- I also anticipate interest from the following clients, but this would depend on the results of my project – Government funded bodies and non-profits offering job placement services, subsidized education services for youth and adults.

3. Datasets used, data wrangling, and data exploration:

The City of Toronto has an Open Data portal, which consists of over 200 datasets. These datasets are organized into 15 different categories. I used three datasets from the Open Data portal consisting of safety, demographics, and economics data for two years – 2011 and 2008.

3.1 Reading in data

Each dataset was provided as a raw dataset in excel with two sheets – one for 2008 and one for 2011. I converted these sheets into separate csv files, and imported them as pandas dataframes. So each raw dataset resulted in two pandas dataframes as shown below. The safety datasets are referred to as "TorontoCrime2011.csv" and "TorontoCrime2008.csv". Likewise, the demographics datasets are referred to as "TorontoDemographics2011.csv" and "TorontoDemographics2008.csv". Initially, I did not foresee any use for the economics data as I felt that the safety and demographics datasets would be sufficient to address my problem. I will explain what motivated me to use the economics data later in this document.

In [3]:
```
c = pd.read_csv('TorontoCrime2011.csv')
d = pd.read_csv('TorontoDemographics2011.csv')

# Also read in the 2008 data files
c2 = pd.read_csv('TorontoCrime2008.csv')
d2 = pd.read_csv('TorontoDemographics2008.csv')
```

3.2 Initial data exploration

I checked the first five rows in the crime and demographics dataframes for both 2008 and 2011, and noticed some differences in the two years. The 2011 crime data had a column for total major crimes whereas the 2008 data did not have one. I wanted consistent columns in both years to enable comparisons.

I noticed that the demographics dataframe for 2011 had 39 columns while the one for 2008 had 85 columns. This was because in 2008 the city of Toronto collected data in each neighbourhood on language and ethnicity of different groups whereas in 2011 the city only collected language data. I realized this would pose another challenge for making comparisons between 2008 and 2011.

The first five rows of the 2011 crime dataframe are shown below.

Out[5]:

	Neighbourhood	Neighbourhood Id	Ambulance Calls	Ambulance Referrals	Arsons	Assaults	Break & Enters	Drug Arrests	Fire Vehicle Incidents	Firearms Incidents	Fires & Fire Alarms	Ha In
0	West Humber-Clairville	1	3613	10	4	272	193	88	674	3	135	26
1	Mount Olive-Silverstone-Jamestown	2	2229	5	0	269	88	145	52	11	70	16
2	Thistletown-Beaumond Heights	3	793	5	0	66	30	27	20	0	26	48
3	Rexdale-Kipling	4	664	5	0	49	28	17	14	0	26	64
4	Elms-Old Rexdale	5	836	3	2	49	22	8	45	1	23	52

3.3 Checking for missing values

I checked to see if there were any missing values in these dataframes. To do this, I dropped all the 'NA' columns in the dataframes and checked to see if their size was equal to the original dataframes. The sizes were equal. So, there were no missing values.

3.4 Renaming column titles

The column titles of the crime dataframes for 2011 and 2008 are shown below.

```
Out[9]: Index([u'Neighbourhood', u'Neighbourhood Id', u'Arsons', u'Assaults',
               u'Break & Enters', u'Drug Arrests', u'Fire Medical Calls',
               u'Fire Vehicle Incidents', u'Fires & Fire Alarms',
               u'Hazardous Incidents', u'Murders', u'Robberies', u'Sexual Assaults',
               u'Thefts', u'Total Major Crime Incidents', u'Vehicle Thefts'],
              dtype='object')
```

```
Out[10]: Index([u'Neighbourhood', u'Neighbourhood Id', u'Ambulance Calls',
                u'Ambulance Referrals', u'Arsons', u'Assaults', u'Break & Enters',
                u'Drug Arrests', u'Fire Vehicle Incidents', u'Firearms Incidents',
                u'Fires & Fire Alarms', u'Hazardous Incidents', u'Murders',
                u'Robberies', u'Sexual Assaults', u'TCHC Safety Incidents', u'Thefts',
                u'Vehicle Thefts'],
               dtype='object')
```

The column titles were very long with spaces in between. To make data access and data operations more manageable, I shortened all the column titles into smaller, single-word names by using a dictionary.

An example of shortened column titles for the 2011 crime dataframe is shown below:

Out[11]:

	N	NId	Ars	Ass	BE	DA	FMC	FVI	FFA	HI	M	R	SA	T	TMCI	VT
135	West Hill	136	3	387	102	87	1145	78	338	142	1	71	52	3	749	46
136	Woburn	137	2	412	128	77	1469	219	504	223	3	107	29	7	808	45
137	Eglinton East	138	0	239	88	48	720	76	223	111	1	66	17	10	492	23
138	Scarborough Village	139	1	226	93	31	652	35	180	94	1	62	31	3	474	27
139	Guildwood	140	0	44	32	9	284	24	48	48	0	14	7	2	113	5

Here, *N* is the Neighbourhood, and *NId* is the neighbourhood ID. These two columns were named the same for both 2011 and 2008. A comparison with the original column titles tells us what the shortened column titles represent.

An example of shortened column titles for the 2008 crime dataframe is shown below:

Out[12]:

	N	NId	AC2	AR2	Ars2	Ass2	BE2	DA2	FVI2	FI2	FFA2	HI2	M2	R2	SA2	TCHCSI2	T2	VT2
135	West Hill	136	2323	10	3	357	90	179	74	7	111	181	0	59	28	721	9	70
136	Woburn	137	3607	29	2	325	129	72	254	9	118	196	1	78	17	373	15	152
137	Eglinton East	138	1500	8	1	171	83	105	92	5	67	104	2	45	10	285	3	92
138	Scarborough Village	139	1364	10	1	170	52	74	37	7	62	95	3	35	8	276	3	57
139	Guildwood	140	688	0	0	50	30	20	17	0	33	51	0	12	2	0	0	12

All columns in 2008 that had corresponding columns in 2011, were named with the suffix "2" added at the end. For example, *Break & Enters* was renamed as *BE* in 2011 and *BE2* in 2008. There were other columns in 2008 that did not have corresponding columns in

2011. These columns were renamed initially but ignored for the analysis. The column titled *TMCI* in 2011 represents all the major crimes committed. It is the sum of eight different crime categories – *Assaults, Break & Enters, Drug Arrests, Murders, Robberies, Sexual Assaults, Thefts*, and *Vehicle Thefts*. My focus for this project was on this summed crime category, which I refer to as major crime. As mentioned earlier, the 2008 crime data did not have this summed category. So, I computed this category for 2008 and added it as an additional column called *TMCI2*.

3.5 Normalizing crime data

I realized while generating some initial plots that population would be a major confounding variable. In other words, a neighbourhood might have more number of major crimes occurring merely because of having a higher population density. This would overpower other potential contributors to crime such as joblessness. To avoid this effect, I decided to normalize the data by looking at major crime per capita. The crime data was initially normalized by dividing all crimes for each neighbourhood with the neighbourhood's population. The population data was obtained from the demographics dataframes for 2008 and 2011. But these normalized values were small decimal values. To get a better feel for the data, I calculated major crimes per 1000 people.

Sample output of the first five rows of the normalized 2011 crime data is shown below.

```
In [18]: c_norm.head() #2011 crime data
```

Out[18]:

	N	Nld	Ars	Ass	BE	DA	FMC	FVI	FFA	HI	M	R	S
0	West Humber-Clairville	1	0.117302	11.436950	5.131965	1.818182	38.739003	14.721408	20.674487	6.158358	0.000000	2.404692	1.
1	Mount Olive-Silverstone-Jamestown	2	0.091491	9.637084	1.860323	2.744739	30.985056	1.799329	11.009454	5.367490	0.030497	2.378774	2.
2	Thistletown-Beaumond Heights	3	0.000000	8.382643	3.550296	1.577909	31.854043	4.733728	8.875740	3.353057	0.000000	1.676529	2.
3	Rexdale-Kipling	4	0.000000	5.627086	3.051979	1.430615	29.089175	3.242728	8.965188	5.245589	0.095374	1.525990	1.
4	Elms-Old Rexdale	5	0.104712	8.062827	2.617801	1.465969	33.612565	7.434555	11.204188	4.502618	0.000000	2.408377	0.

3.6 Further exploration

Next, I compared the means of all the crimes that fell under major crime for 2011 with 2008. I found that *Assaults, Drug Arrests,* and *Break & Enters* were the main major crime contributors for both these years. *Murders* and *Thefts* were the lowest two contributors of major crime. This finding perhaps justifies why Toronto is generally considered to be a safe city.

Mean normalized major crime data for 2011:

```
Out[23]: Ass    8.057472
         BE     4.113241
         DA     2.026643
         R      1.859598
         VT     1.518993
         SA     0.958315
         T      0.346375
         M      0.019289
         dtype: float64
```

Mean normalized major crime data for 2008:

```
Out[24]: Ass2    6.885194
         DA2     3.780567
         BE2     3.740126
         VT2     2.442250
         R2      1.622161
         SA2     0.568509
         T2      0.358737
         M2      0.025703
         dtype: float64
```

Then I explored topics 1 and 2 in more detail, in no particular order. As a reminder the two topics are listed below:

1. A summarized visualization of all the major crimes in Toronto

2. A comparison of 3-5 most crime prone neighbourhoods against 3-5 least crime prone neighbourhoods

All the major crime values in 2011 and 2008 (i.e., *TMCI* and *TMCI2*) were sorted to find the top five most crime prone neighbourhoods for both these years.

Top five major crime prone areas in 2011:

Out[25]:

	N	**TMCI**
30	Yorkdale-Glen Park	37.521280
78	University	37.997433
77	Kensington-Chinatown	46.972973
75	Bay Street Corridor	52.571724
72	Moss Park	56.468424

Top five major crime prone areas in 2008:

Out[26]:

	N	TMCI2
77	Kensington-Chinatown	52.486834
65	Danforth	53.577982
75	Bay Street Corridor	63.054830
78	University	71.126164
72	Moss Park	72.157623

Four of the top five neighbourhoods matched in 2008 and 2011. This is because Yorkdale-Glen Park, Neighbourhood ID: 31 (indexed as 30) was ranked eighth in 2008 and had a slight increase in crime in 2011, therefore showing up in the 2011 top five. Also, Danforth, Neighbourhood ID: 66 (indexed as 65), which is generally considered to be a high crime neighbourhood, had a drop in crime in 2011.

A closer look at Danforth showed that while Danforth had an overall reduction in major crime categories, there was a considerable reduction in *Drug Arrests* by approximately 75%, which may have accounted for Danforth not showing up in the top five neighbourhoods for 2011.

Next, I found what the least crime prone neighbourhoods for these two years were.

Bottom five major crime prone areas in 2011:

Out[27]:

	N	TMCI
48	Bayview Woods-Steeles	7.092723
51	Bayview Village	7.468175
45	Pleasant View	7.866212
132	Centennial Scarborough	8.167939
116	L'Amoreaux	8.170990

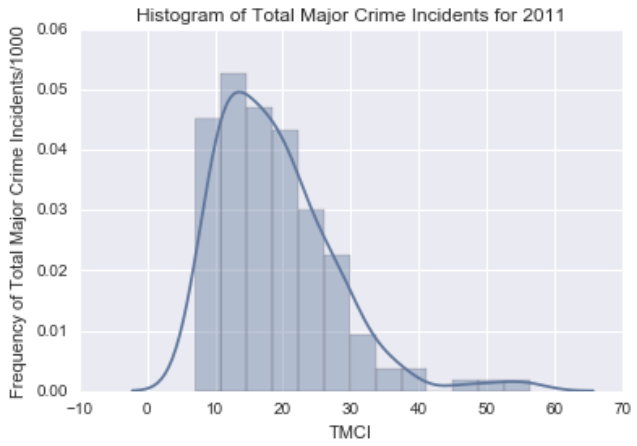
Bottom five major crime prone areas in 2008:

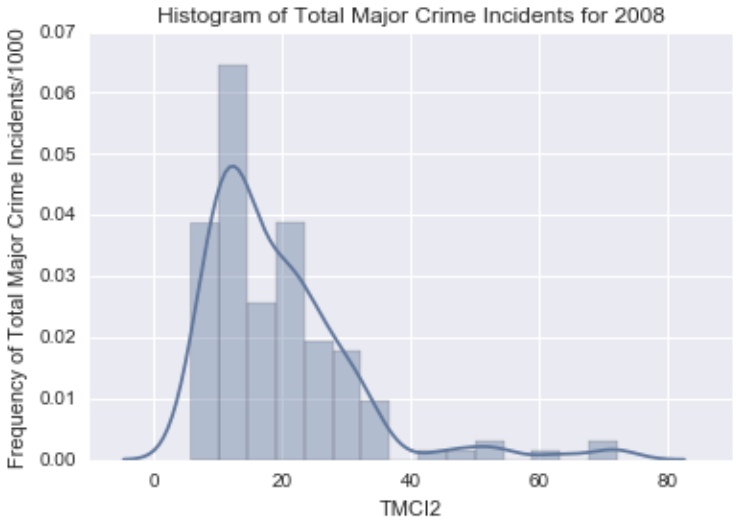
Out[28]:

	N	TMCI2
48	Bayview Woods-Steeles	5.716435
45	Pleasant View	6.560337
51	Bayview Village	6.961614
132	Centennial Scarborough	7.068273
10	Eringate-Centennial-West Deane	7.229566

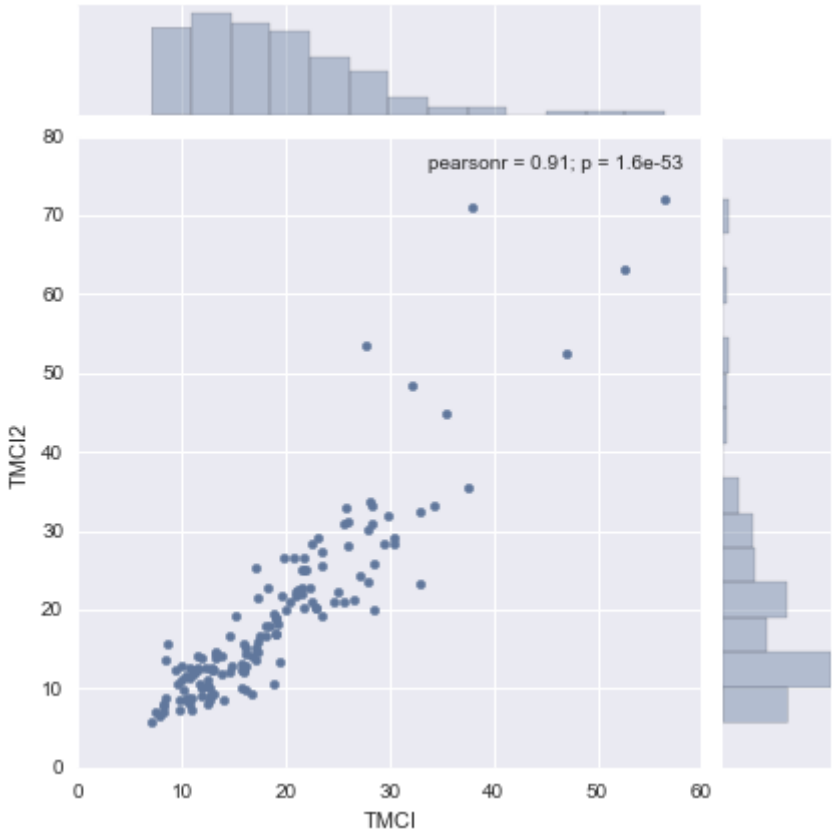
Again, I noticed that four of the five neighbourhoods match in 2008 and 2011, suggesting that major crime had more or less been stable in Toronto, in these three years.

Next, I plotted a few visualizations of the major crime data. I compared major crimes in 2011 and 2008 as univariate distributions using Seaborn by plotting a histogram and a kernel density estimate.



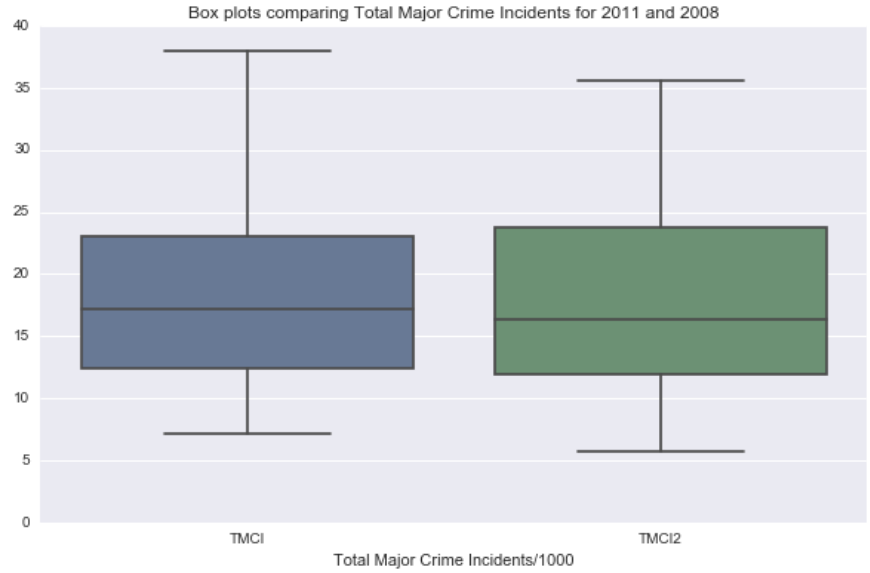


I realized that an easier way to compare both years would be to show the major crimes as bivariate distributions in a scatter plot. This also showed how correlated both distributions were. Higher correlation indicates lesser differences in major crime rates for 2008 and 2011. In the plot below, 2011 major crime data is on the x-axis and 2008 major crime data is on the y-axis.



As expected, there was a strong correlation of 0.91 in major crime for both years. In addition to outliers in the plot, I also noticed two interesting points which did not fall along the general trend. These two points refer to major crime (per 1000 people) in two neighbourhoods where 45 < TMCI2 (i.e., major crime in 2008) < 55, and 25 < TMCI (i.e., major crime in 2011) < 35. I found that these two neighbourhoods were Danforth (Neighbourhood ID: 66) and Waterfront Communities (Neighbourhood ID: 77). We discussed how Danforth, despite being a crime prone neighbourhood, had a reduction in major crimes in 2011. Likewise, Waterfront Communities seemed to have a drop in crime as well.

Next I plotted the distribution of major crime for both years as box plots. I also visualized these distributions as violin plots. The violin plots did not show any additional information beyond what the box plots did, since the data did not have a bimodal distribution.



Then I explored topic 4, which was finding the neighbourhoods with the maximum change in major crime. To do this, I calculated the percentage of increase or decrease in major crime from 2008 to 2011 by (a) computing the difference in major crime between the two years for each neighbourhood, (b) dividing this difference with the major crime in 2008 for that neighbourhood, and (c) converting this value into a percentage.

The five neighbourhoods with the maximum increase in major crime from 2008 to 2011 were the following.

	N	TCDiff
133	Highland Creek	57.341559
38	Bedford Park-Nortown	63.680435
9	Princess-Rosethorn	65.275058
134	Morningside	76.682444
49	Newtonbrook East	79.724978

The five neighbourhoods with the maximum decrease in major crime from 2008 to 2011 were the following.

	N	TCDiff
65	Danforth	-48.253243
78	University	-46.577419
95	Casa Loma	-44.713227
100	Forest Hill South	-38.321479
76	Waterfront Communities-The Island	-33.703986

Positive values of *TCDiff* indicate an increase in major crime and negative values indicate a decrease in major crime, from 2008 to 2011. I wanted to know how these neighbourhoods with the most increase and decrease in major crime compared against some of the top and bottom major crime neighbourhoods for 2011.

I already discussed how Danforth, generally a high major crime neighbourhood, had a decrease in crime in 2011. This becomes obvious when we look at the *TCDiff* value which shows a 48.5% decrease from 2008 to 2011. We also noticed the decrease in major crime for Waterfront Communities in the bivariate scatter plot. In support of this finding, the *TCDiff* data reveals a 33.7% decrease in major crime for Waterfront Communities from 2008 to 2011.

Next, I explored topic 3, which was about finding the most prominent age group of people in each neighbourhood. I looked at the demographics data for 2011 and 2008 and decided to focus on four different age groups – (1) Children (0-14 years), (2) Youth (15-24 years), (3) Adults (25-54 years), (4) Seniors (55 and over). Most of these categories were already available as columns in the 2011 demographics dataframe, excepting Adults. So, I computed the number of Adults in each neighbourhood by subtracting the sum of the remaining three age groups from the total population of each neighbourhood.

In the 2008 demographics dataframe, none of these categories were available as pre-existing age groups. Instead, population was divided into columns that were grouped by

5-year age categories (e.g., 0-4 years, 5-9 years etc.). Therefore, to make the 2008 data consistent with the 2011 data, I calculated the number of people in each of the four age groups – Children, Youth, Adults, Seniors. Summary statistics for the 2008 and 2011 demographics data showed that the Adults group is the most prominent population group across all neighbourhoods.

I also wanted to look at the median/mean household income in each neighbourhood for 2008 and 2011. Unfortunately, income data was available only with the 2008 safety data from the Open Data portal, and was unavailable for 2011. This is what motivated me to take a closer look at the economics data, something I had hinted at in Section 3.1. Surprisingly, the economics datasets did not contain any income data. However, they had other potentially important variables such as number of people employed, and number of people on social assistance.

Following a procedure similar to what I did with the safety and demographics data, I read the economics data for 2008 and 2011 into separate pandas dataframes, selected only the most relevant columns (i.e., number of businesses in each neighbourhood, number of people employed in each neighbourhood, and number of social assistance recipients), and shortened the column titles to simpler ones. All three variables seemed important because they are connected to income and employment, which are traditionally considered as important motivators for committing crime.

A sample of five rows showing the selected economics variables for 2011:

	N	NId	Bus	Emp	SAR
135	West Hill	136	1.596987	16.949153	18.734463
136	Woburn	137	2.011246	30.346767	12.646673
137	Eglinton East	138	1.296540	10.297854	18.269820
138	Scarborough Village	139	1.372254	11.140536	24.875113
139	Guildwood	140	0.682629	9.414162	3.148242

Here, *Bus* is the number of businesses per 100 people, *Emp* is the number employed per 100 people, and *SAR* is the number of social assistance recipients per 100 people.

4. Prediction using machine learning with model evaluation

Ideally, I would have liked to have all the major crime data over a series of years (approximately 20 years or so) along with the corresponding features for each year. This would have allowed me to perform crime prediction for a projected year. Since, I did not have this data available, I decided to perform two types of prediction exercises, both of which fall under regression problems.

First, I decided to look at the percentage change in major crime from 2008 to 2011 as my dependent variable, and the corresponding percentage changes in my features as my independent variables. Can we successfully predict the percentage of increase or decrease in crime from 2008 to 2011 using a machine learning model?

Second, I decided to predict the major crime in a neighbourhood using the features as independent variables. So, in the second problem, I am not looking at changes from 2008 to 2011. Instead, I am building two separate machine learning models – one for 2008 and one for 2011.

4.1 Feature selection

Prior to performing model selection, I wanted to take a look at the limited set of features I had and filter out any unnecessary features, keeping the two regression problems in mind.

I merged the age group data (the four age groups) and percentage of male and female data with the three selected features from the economics data, and major crime data, for each year separately. This resulted in obtaining two dataframes with 10 columns – 9 features/independent variables, and 1 dependent variable. Then I computed cross-correlations of all 10. This allowed me to look for any collinearities (variables correlated with each other) and remove variables that were confounding each other. Two strongly correlated variables can decrease the robustness of a machine learning model as one variable can potentially inhibit the effect of another.

Cross-correlation table for 2011 is provided below.

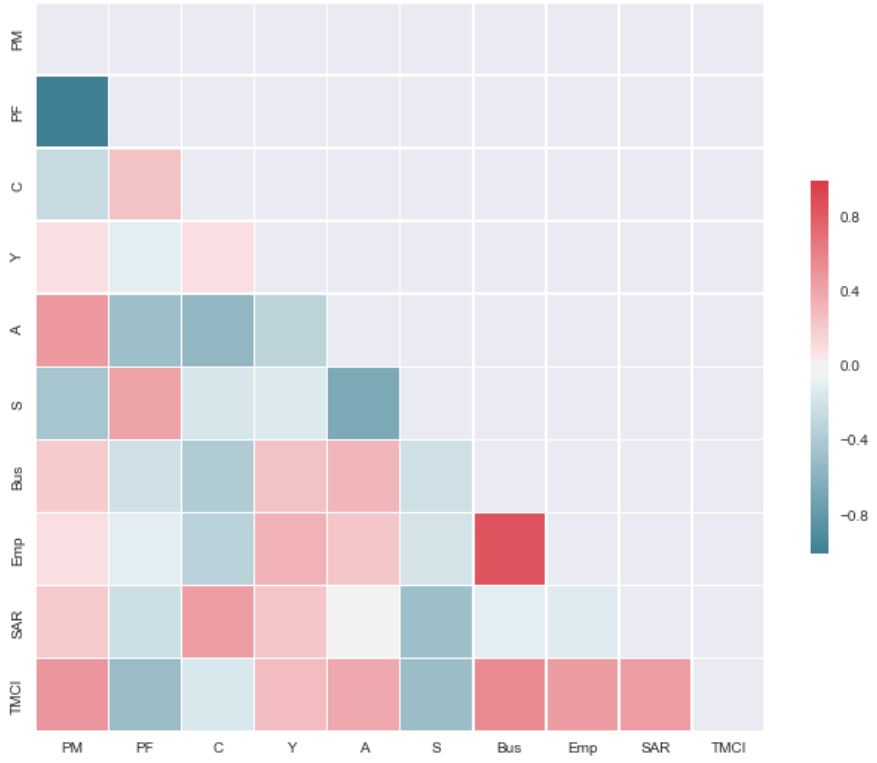
	PM	PF	C	Y	A	S	Bus	Emp	SAR	TMCI
PM	1.000000	-0.999708	-0.262159	0.105218	0.485240	-0.434347	0.206219	0.107014	0.218235	0.504714
PF	-0.999708	1.000000	0.259095	-0.106597	-0.483859	0.435686	-0.207220	-0.107741	-0.219820	-0.505878
C	-0.262159	0.259095	1.000000	0.101991	-0.545515	-0.166599	-0.380420	-0.321718	0.456916	-0.150601
Y	0.105218	-0.106597	0.101991	1.000000	-0.311823	-0.139468	0.263345	0.345554	0.236682	0.294891
A	0.485240	-0.483859	-0.545515	-0.311823	1.000000	-0.657665	0.327817	0.240374	0.011142	0.396014
S	-0.434347	0.435686	-0.166599	-0.139468	-0.657665	1.000000	-0.217517	-0.192954	-0.479418	-0.496818
Bus	0.206219	-0.207220	-0.380420	0.263345	0.327817	-0.217517	1.000000	0.860655	-0.102634	0.563295
Emp	0.107014	-0.107741	-0.321718	0.345554	0.240374	-0.192954	0.860655	1.000000	-0.120620	0.475120
SAR	0.218235	-0.219820	0.456916	0.236682	0.011142	-0.479418	-0.102634	-0.120620	1.000000	0.463434
TMCI	0.504714	-0.505878	-0.150601	0.294891	0.396014	-0.496818	0.563295	0.475120	0.463434	1.000000

Cross-correlation table for 2008 is provided below.

	PM2	PF2	C2	Y2	A2	S2	Bus2	Emp2	SAR2	TMCI2
PM2	1.000000	0.629499	-0.088499	0.007301	0.196905	-0.162688	0.212224	0.072305	0.134476	0.412011
PF2	0.629499	1.000000	0.096254	-0.099926	-0.176978	0.173348	0.043809	-0.029058	-0.038105	-0.015050
C2	-0.088499	0.096254	1.000000	0.175526	-0.545894	-0.264382	-0.402654	-0.324367	0.496280	-0.211898
Y2	0.007301	-0.099926	0.175526	1.000000	-0.340884	-0.146744	0.107260	0.171481	0.171874	0.146219
A2	0.196905	-0.176978	-0.545894	-0.340884	1.000000	-0.597055	0.390237	0.285942	-0.019473	0.432165
S2	-0.162688	0.173348	-0.264382	-0.146744	-0.597055	1.000000	-0.169916	-0.135863	-0.466623	-0.397451
Bus2	0.212224	0.043809	-0.402654	0.107260	0.390237	-0.169916	1.000000	0.885773	-0.096853	0.635169
Emp2	0.072305	-0.029058	-0.324367	0.171481	0.285942	-0.135863	0.885773	1.000000	-0.117931	0.499110
SAR2	0.134476	-0.038105	0.496280	0.171874	-0.019473	-0.466623	-0.096853	-0.117931	1.000000	0.399231
TMCI2	0.412011	-0.015050	-0.211898	0.146219	0.432165	-0.397451	0.635169	0.499110	0.399231	1.000000

I plotted these correlations to make them more understandable.

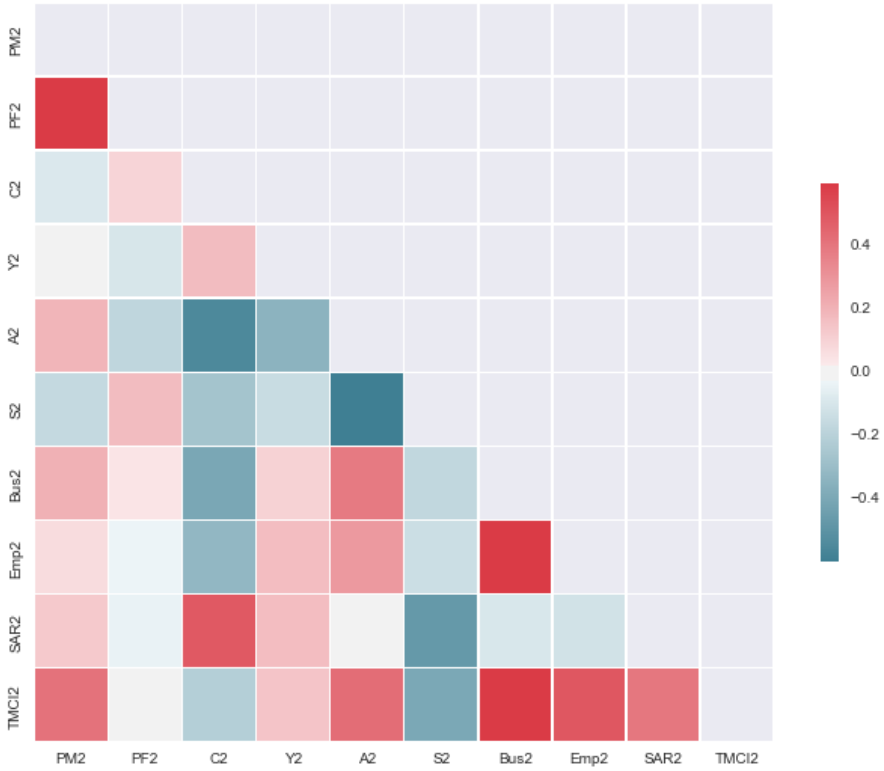
Correlation plot for 2011:



For the 2011 data, there was a moderate to high positive correlation between the number of males and major crime (*TMCI*), number of adults and major crime, number of people employed and major crime, and number of social assistance recipients and major crime.

There was a negative correlation between the number of seniors and major crime. There was also a strong positive correlation between the number of people employed and the number of businesses in that neighbourhood, suggesting that these features might be containing redundant information. I also noticed a strong negative correlation between the percentage of female population and the percentage of male population.

Correlation plot for 2008:



For the 2008 data, surprisingly, there was a moderate positive correlation between the percentage of females and the percentage of females, unlike in 2011, indicating that there was no consistent pattern. In any case, major crime is primarily associated with males rather than females. This led to removing the percentage of females as a feature. I noticed a strong correlation between the number of people employed and the number of businesses in that neighbourhood for 2008 as well. Ideally, I would have removed one of these features. However, I decided against it based on the following reasoning. First, it is important to acknowledge that there is likely to be some correlation between these two variables. The number of businesses in a neighbourhood are likely to be a good indicator of the overall economic health of that neighbourhood. However, it is also possible that the businesses in a specific neighbourhood are not necessarily employing the people from that neighbourhood. Several businesses exist in downtown areas of cities, despite which neighbourhoods closer to downtown could still be high crime prone areas. Additionally, the nature of businesses can vary. A neighbourhood might consist entirely of small businesses such as sole proprietorships and small partnerships, which might not be

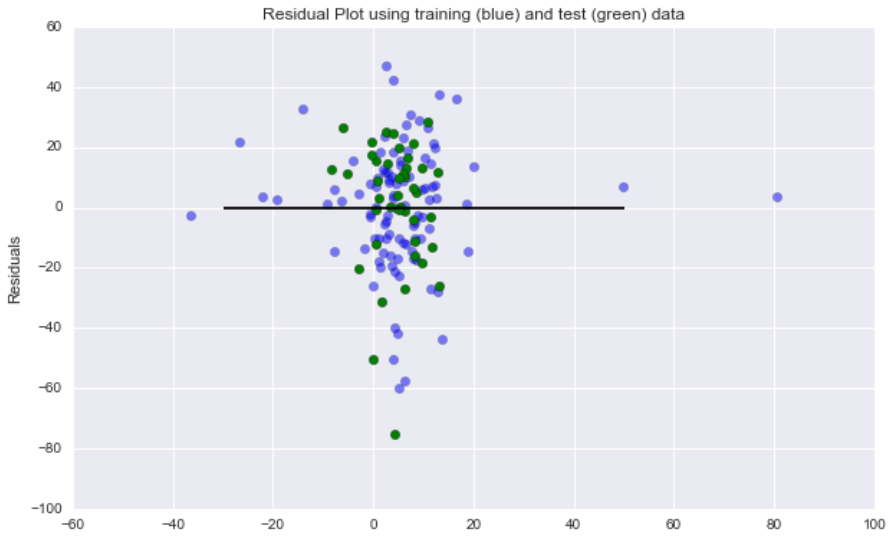
having a sizeable number of employees. Given all these reasons, I included both these variables within the feature set. So, in summary, I used eight features.

4.2 Model selection and evaluation

My first task was to predict the percentage of increase or decrease in crime from 2008 to 2011 using a machine learning model. I computed the percentage changes in all eight features from 2008 to 2011, and the percentage change in major crime, for each neighbourhood. I used a linear regression model imported from scikit-learn. I partitioned the data into training and test sets using a 70:30 split for training and testing, respectively. I decided to use linear regression because it provides a basic starting point for regression. It always leaves us with the option of choosing a more flexible model later, if needed.

I evaluated the model by first checking the R^2 value of the model on the training set, which was 27.3% - a relatively low number indicating that the model was not flexible enough to fit the data. We also have to consider that neither are we using the actual features, nor are we predicting major crime rates. Instead, we are using percentage changes of these attributes between 2008 and 2011 as features for predicting the percentage change in major crime. So, some of the initial relationships between the features and major crimes, as indicated by their correlations, may have been compromised. As an additional evaluation test, I computed the mean squared error on the training data (MSE = 404.9) and on the test data (MSE = 439.0). Both the training and test mean squared errors are extremely high.

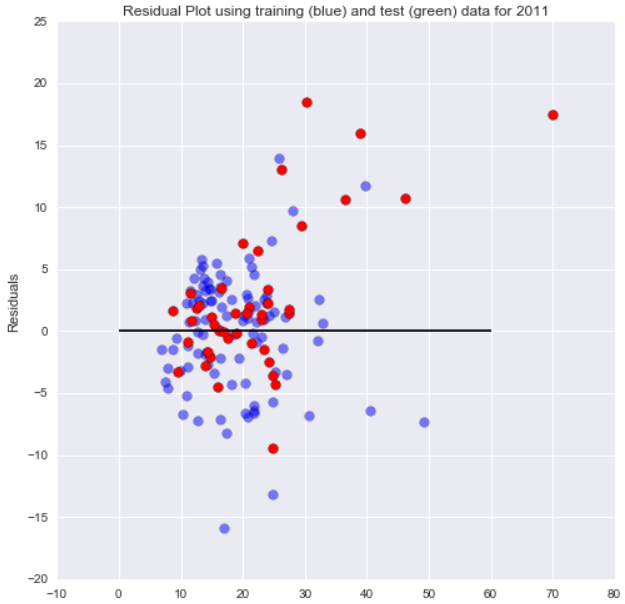
I plotted the residuals for the linear regression model. Blue indicates residuals for the training data and green indicates residuals for the test data. In both cases, the variance is not constant. The residual plot does not show homoscedasticity. So, I decided to use a more flexible model – random forests.



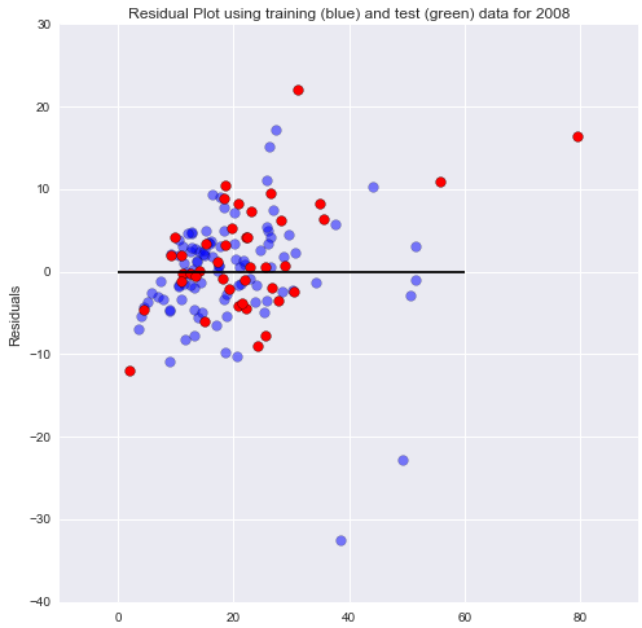
The random forest model had an R^2 value of 85.1% on the training data, which was a big improvement over the linear regression model. However, it performed poorly on the test data (R^2 value of 9.2%), suggesting that it could not generalize. This could also be noticed in the difference in mean squared error values for the training data (MSE = 83) and for the test data (MSE = 404.9). As an additional step, I did a k-fold cross-validation with k = 5. The average mean squared error across all five folds was still high (MSE = 571.6).

These results indicated that the features did not have enough information that allowed the possibility of deriving a model that could make reasonable predictions. So, my conclusion was that with this dataset and features that I had selected I would not be able to successfully predict the percentage of change in major crime in a neighbourhood from 2008 to 2011.

So, I moved on to my second regression problem where I had decided to predict major crime in each neighbourhood using the eight original features as independent variables. For the second problem, I wanted to have two separate machine learning models – one for 2008 and one for 2011. Following a similar approach to what I did previously, I started with linear regression as my choice of machine learning model. I used the same randomized partitioning as before – 70% of the data was used for training and 30% for testing. Both the models were able to explain the variance in the training data reasonably well, as reflected by their R^2 values (71% for 2011 and 72% for 2008). For the 2011 model, mean squared errors were 22.3 on the training data and 40.7 on the test data, whereas for the 2008 model, mean squared errors were 40.9 on the training data and 46.6 on the test data. So, although the performance for the 2011 model seemed to be superior, the 2008 model seemed more robust with respect to generalizability, given that the difference between the MSE for training and test data was quite low. Residual plot for the 2011 data is provided below.

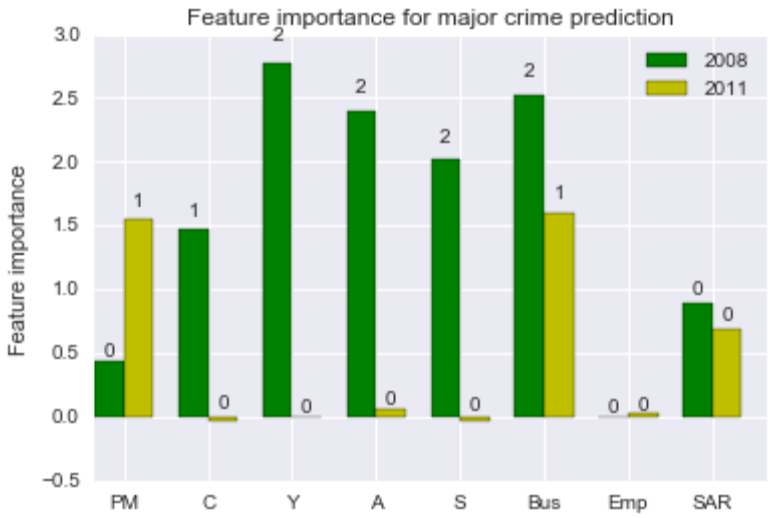


Blue indicates residuals for the training data and red indicates residuals for the test data. Residual plot for the 2008 data is provided below.



For the most part, the models look homoscedastic, further confirming that linear regression is a reasonable choice for a regression model. They show a small amount of heteroscedasticity and ceiling effects.

Since the models looked reasonable after evaluation, I examined the beta coefficients of all the predictors for both models using the "coef_" attribute for linear regression in scikit-learn.



For the 2008 linear regression model, the beta coefficient values suggest that almost all the features except the number of people employed, are important predictors for major crime in a neighbourhood. These features are people in each of the four age groups, percentage of males, number of businesses, and number of social assistance recipients in each neighbourhood. For the 2011 linear regression model, however, only the percentage of males, number of businesses, and number of social assistance recipients show up as important predictors of major crime. So, the models for both years are not consistent with each other.

As a next step, I used a random forest regression model to see if there would be an improvement in prediction performance, with the same 70:30 partitioning of the data into training and test data.

Both the models were able to explain the variance in the training data very well, as reflected by their R^2 values (90.8% for 2011 and 92.4% for 2008). For the 2011 model, mean squared errors was 153.3 on the test data, whereas for the 2008 model, mean squared error was 156.4 on the test data. These results indicate that the regression models, despite having better R^2 values, performed poorer than linear regression. In other words, they were too flexible and as a result overfitted to the training data. From a generalizability and robustness perspective, linear regression might be the better option as a regression model for predicting major crime in each neighbourhood.

5. Key findings, limitations, and future extensions

My most important finding from the 2011 linear regression model and the random forest models was that among the limited set of independent variables that I were available to me, the percentage of males, the number of businesses, and the number of social assistance recipients within a neighbourhood are the most important predictors of major crime in that neighbourhood.

This finding cannot be generalized because the data was limited to only two years – 2008 and 2011. Besides, there were several other features that should have ideally been included as independent variables (i.e., features in a machine learning model) which I could not include as they were unavailable for both years. Perhaps the most important missing feature was income data. I would also have liked to include a feature that either quantified urbanization or captured gentrification in these neighbourhoods. Toronto has been undergoing a lot of change in the form of major construction projects in low, and mid income neighbourhoods. Some of these are considered to be a part of revitalization projects. For example, the city of Toronto began an initiative in 2005 known as the Regent Park Revitalization Plan. The plan involved transforming an area from a social housing neighbourhood into a thriving mixed income neighbourhood by implementing construction in three phases that included a mix of rental and condominium buildings, townhouses, commercial space with community facilities, active parks and open spaces. Currently, phases 2 and 3 are underway. Variables that capture urbanization can definitely show how the demographics of the city are being reshaped and how these changes are affecting crime in that neighbourhood. To obtain some of these features I

would have to explore beyond the Open Data portal made available by the City of Toronto. This would be a possible extension to the current work.

Toronto is also going through a condominium construction boom, which is increasingly escalating rental and housing prices, as well as affecting affordability of living for low-income residents. How much of this change could be affecting crime? Housing and rental prices could serve as important independent variables that affect crime.

One final aspect I would investigate is the effectiveness of social assistance programs. My findings show that the greater the number of people on social assistance, the more the crime in that area. But does this mean that social assistance is causing more crime? Clearly no. However, it is a reflection of the income needs of people in that neighbourhood. The expectation is that with greater social assistance, the income of each person and therefore the overall economic health of that neighbourhood will change. But is this really happening? One way to investigate this is to look at neighbourhoods that received more social assistance, and see if the crime rates in that neighbourhood reduced in a few years from the point of receiving higher social assistance.

From a prediction standpoint, obtaining more data for at least 15-20 years would have allowed the models to capture predictable trends. Despite these limitations, the data provided us with interesting findings and a set of action items for future extension of this work.

6. Recommendations

These are the three recommendations for my clients.

Recommendation 1 – The findings show that we need more businesses in each neighbourhood to reduce the amount of major crime. Having more businesses is a sign of urbanization and impacts the overall economic health of a neighbourhood.

Recommendation 2 – My second recommendation is a derived finding from 1. Given that the number of businesses in a neighbourhood is important, we need to facilitate the growth of businesses. Therefore, more people within each neighbourhood would need to be provided with the necessary training and self-financing opportunities for starting their own entrepreneurial ventures. The city should provide various economic incentives for entrepreneurs and ensure that people in every neighbourhood are made aware of these opportunities.

Recommendation 3 – The findings show a correlation between the number of people receiving social assistance and the amount of major crime in a neighbourhood. This probably reflects the immediate housing and income needs of people in that neighbourhood. So, the city should consider ways in which social assistance programs can assist people not just in the short-term by way of income and housing help, but also in the long-term by offering accessible education and employability programs to these

people. The city can also do an assessment of these long-term programs for their effectiveness in reducing crime over a 2-5 year period.