An Analysis of Neighbourhood Crime in the City of Toronto

Naresh Vempala

October 31, 2016

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)[1], Toronto was ranked as the safest major city in North America and the eighth safest major city in the world, as cited in Wikipedia[2].

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, 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 factors and major crime that would not be obvious at the outset. The problem may be articulated as – finding major crime trends in Toronto's neighbourhoods, identifying potential factors related to major crime, and using these factors to build a predictive model.

Another 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 analyses.

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

[1] http://safecities.cope.economist.com/wp-content/uploads/sites/5/2015/06/Safe_cities_index_2015_EIU_report-1.pdf
[2] https://en.wikipedia.org/wiki/Crime_in_Toronto

4. Identifying the neighbourhood with the most change in major crime and potential reasons for the change

5. Building a predictive model of major crime in each neighbourhood using machine learning

6. Finally, 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. The first type of clients would be Canadian or 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. For example, Canadian magazines such as *The Walrus* and *THIS Magazine* would fall under this category. I also anticipate interest from Government funded bodies and non-profits offering job placement services, and subsidized education services for youth and adults.

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

The City of Toronto has an Open Data[3] 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[4], demographics[5], and economics[6] 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 data

[3] http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=9e56e03bb8d1e310VgnVCM10000071d60f89RCRD

[4] http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=b3ff80ece073b410VgnVCM10000071d60f89RCRD

[5] http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=4482904ade9ea410VgnVCM10000071d60f89RCRD

[6] http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=aec380ece073b410VgnVCM10000071d60f89RCRD

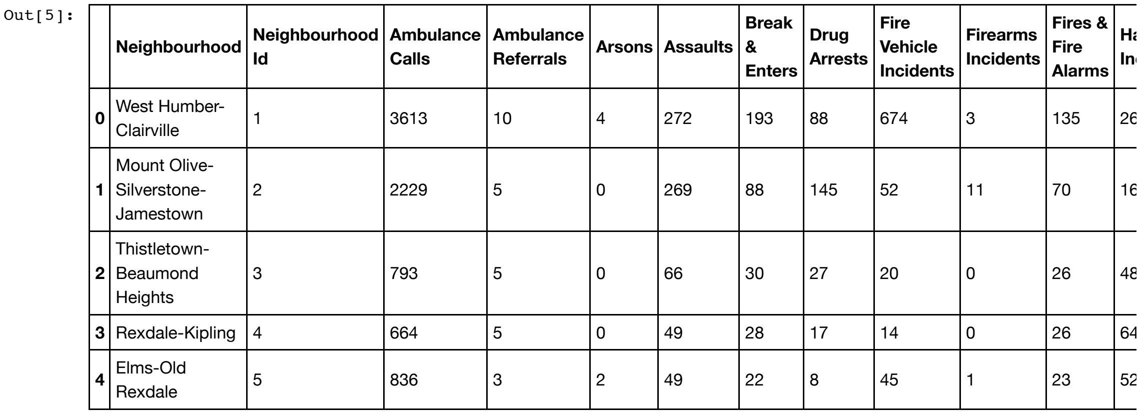
frames. Each raw dataset resulted in two Pandas data frames. Initially, I did not foresee any use for the economics data as I felt that the safety and demographics datasets would be sufficient to address my problem. I will explain what motivated me to use the economics data in Section 3.7.

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 data frame for 2011 had 39 columns while the data 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 data frames for 2011 and 2008 are shown below.

2008 Columns	**2011 Columns**	**2008 Shortened Name**	**2011 Shortened Name**
Neighbourhood	Neighbourhood	N	N
Neighbourhood Id	Neighbourhood Id	NId	NId
Ambulance Calls	-	AC2	-
Ambulance Referrals	-	AR2	-
Arsons	Arsons	Ars2	Ars
Assaults	Assaults	Ass2	Ass
Break & Enters	Break & Enters	BE2	BE
Drug Arrests	Drug Arrests	DA2	DA
-	Fire Medical Calls	-	FMC
Fire Vehicle Incidents	Fire Vehicle Incidents	FVI2	FVI
Firearms Incidents	-	FI2	-
Fires & Fire Alarms	Fires & Fire Alarms	FFA2	FFA
Hazardous Incidents	Hazardous Incidents	HI2	HI
Murders	Murders	M2	M
Robberies	Robberies	R2	R
Sexual Assaults	Sexual Assaults	SA2	SA
TCHC Safety Incidents	-	TCHCSI2	-
Thefts	Thefts	T2	T
-	Total Major Crime Incidents	-	TMCI
Vehicle Thefts	Vehicle Thefts	VT2	VT

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.

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*.

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

	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

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

	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

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 data frames 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 capita (1000 people).

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 and 2008:

Crime	**2011 Per Capita**	**2008 Per Capita**
Assaults	8.057	6.885
Break & Enters	4.113	3.74
Drug Arrests	2.027	3.781
Robberies	1.86	1.622
Vehicle Thefts	1.519	2.442
Sexual Assaults	0.958	0.569
Thefts	0.346	0.359
Murders	0.019	0.026

3.7 Comparison of most and 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:

Neighbourhood	Total Crimes
Yorkdale-Glen Park	37.52
University	38.0
Kensington-Chinatown	46.97
Bay Street Corridor	52.57
Moss Park	56.47

Top five major crime prone areas in 2008:

Neighbourhood	Total Crimes
Kensington-Chinatown	52.49
Danforth	53.58
Bay Street Corridor	63.05
University	71.13
Moss Park	72.16

Four of the top five neighbourhoods matched in 2008 and 2011. This is because Yorkdale-Glen Park was ranked eighth in 2008 and had a slight increase in crime in 2011, therefore showing up in the 2011 top five. Also, Danforth, 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 compared the least crime prone neighbourhoods for these two years.

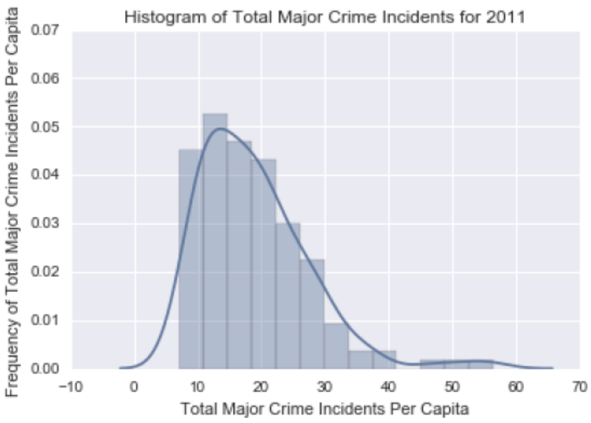
Bottom five major crime prone areas in 2011:

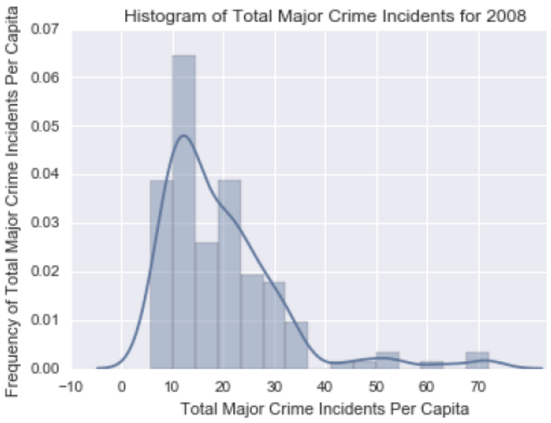
Neighbourhood	Total Crimes
Bayview Woods-Steeles	7.092
Bayview Village	7.468
Pleasant View	7.866
Centennial Scarborough	8.168
L’Amoreaux	8.171

Bottom five major crime prone areas in 2008:

Neighbourhood	Total Crimes
Bayview Woods-Steeles	5.716
Pleasant View	6.56
Bayview Village	6.962
Centennial Scarborough	7.068
Eringate-Centennial-West Deane	7.23

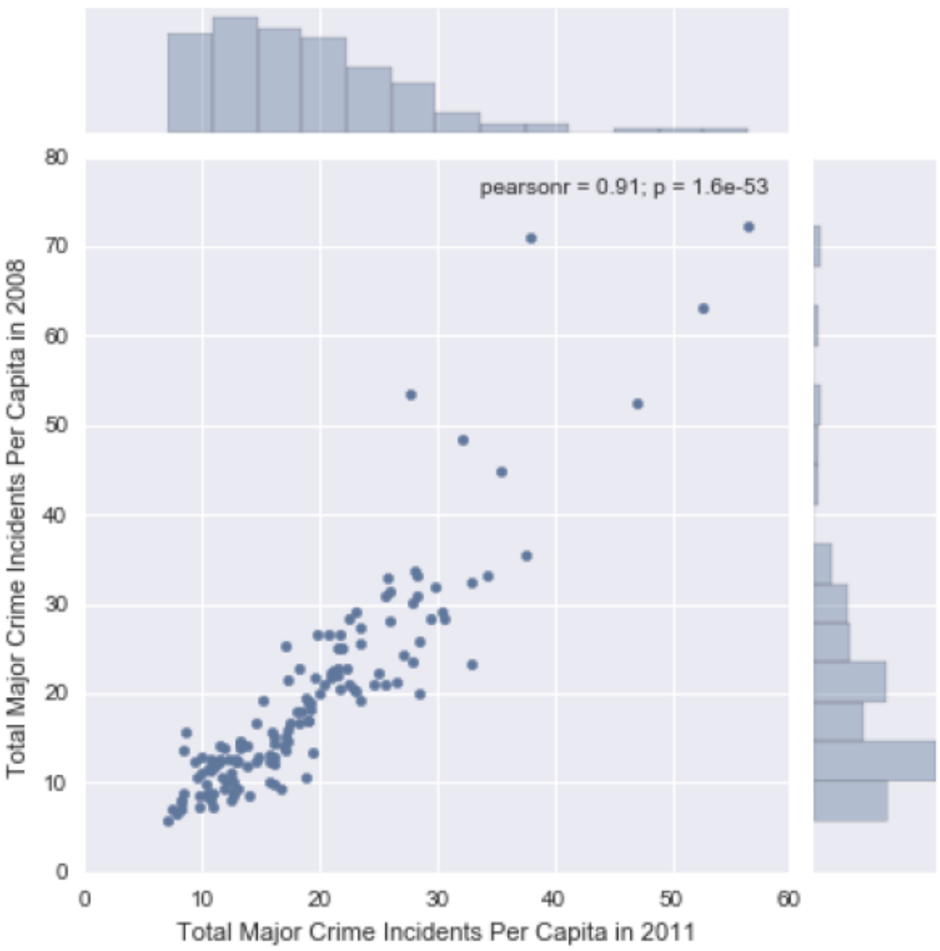
Again, I noticed that four of the five neighbourhoods match in 2008 and 2011, suggesting that major crime had more or less been stable in Toronto, in these two years. Next, I plotted a few visualizations of the major crime data. I compared major crimes in 2011 and 2008 as univariate distributions using Seaborn by plotting a histogram and a kernel density estimate. A kernel density estimation allows us to estimate the probability density function of a random variable from a finite set of data. So, it allows us to look at the major crime data as a continuous probability distribution rather than a histogram.





I realized that an easier way to compare both years would be to show the major crimes in a scatterplot. This also showed how correlated both were. 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 points, which did not fall along the general trend. These two points refer to major crime per capita in two neighbourhoods. These are neighbourhoods where the major crime in 2008 is between 45 and 55, and the major crime in 2011 is between 25 and 35. I found that these two neighbourhoods were Danforth and Waterfront Communities. 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.



3.8 Neighbourhoods with maximum change in crime

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.

Neighbourhood	Percentage Change in Crime from 2008 to 2012
Highland Creek	57.34
Bedford Park-Nortown	63.68
Princess-Rosethorn	65.28
Morningside	76.68
Newtonbrook East	79.72

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

Neighbourhood	**Percentage Change in Crime from 2008 to 2012**
Danforth	-48.25
University	-46.58
Casa Loma	-44.71
Forest Hill South	-38.32
Waterfront Communities-The Island	-33.7

Positive values of *Percentage Change in Crime from 2008 to 2012* 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 *Percentage Change in Crime* value which shows a 48.5% decrease from 2008 to 2011. We also noticed the decrease in major crime for Waterfront Communities in the scatterplot. In support of this finding, the *Percentage Change in Crime* data reveals a 33.7% decrease in major crime for Waterfront Communities from 2008 to 2011.

3.9 Identifying prominent age group and economics factors

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 data frame, except 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 data frame, 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 data frames, selected only the most relevant columns: number of businesses in each neighbourhood, number of people employed in each neighbourhood, and number of social assistance recipients. Then I 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 crime. These variables were normalized to values per 100 people.

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

Neighbourhood	**Neighbourhood Id**	**Businesses per 100 people**	**Number employed per 100 people**	**Social assistance recipients per 100 people**
West Hill	136	1.6	16.95	18.73
Woburn	137	2.01	30.35	12.65
Eglinton East	138	1.3	10.3	18.27
Scarborough Village	139	1.37	11.14	24.88
Guildwood	140	0.68	9.41	3.15

4. Prediction using machine learning

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 are 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 fitting, 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 data frames 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 correlations among variables 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.

Full Name	**2011 Column**	**2008 Column**
Percentage of males	PM	PM2
Percetage of females	PF	PF2
Children per 100 people	C	C2
Youth per 100 people	Y	Y2
Adults per 100 people	A	A2
Seniors per 100 people	S	S2
Businesses per 100 people	Bus	Bus2
Employed per 100 people	Emp	Emp2
Social Assistance recipients per 100 people	SAR	SAR2
Total Major Crimes per capita (1000 people)	TMCI	TMCI2

Cross-correlation table for 2011 is provided below.

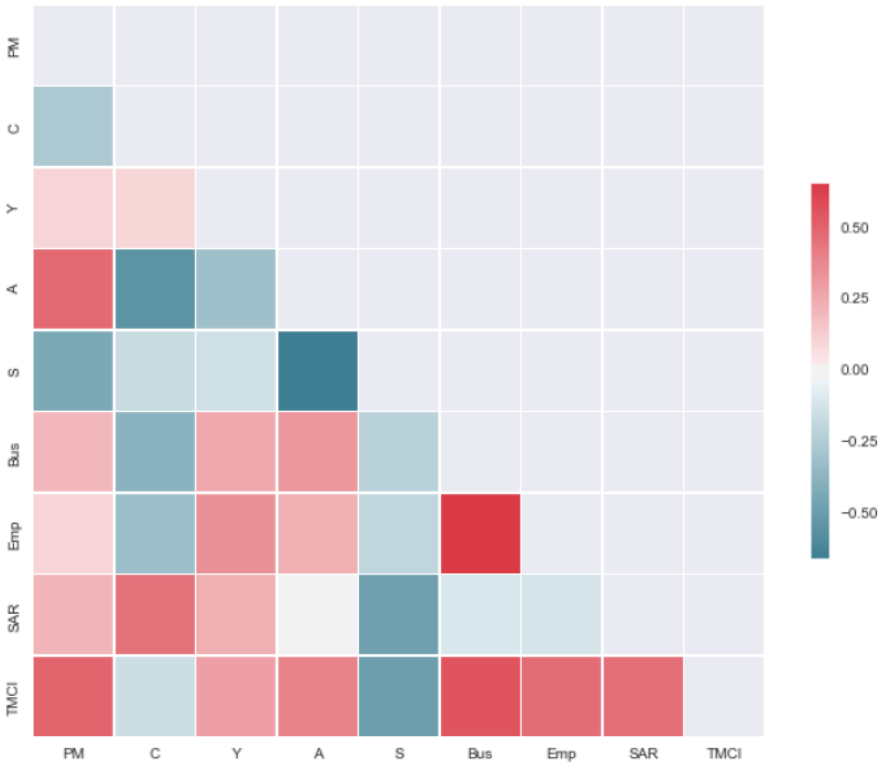
	PM	**PF**	**C**	**Y**	**A**	**S**	**Bus**	**Emp**	**SAR**	**TMCI**
PM	1.0	-1.0	-0.262	0.105	0.485	-0.434	0.206	0.107	0.218	0.505
PF	-1.0	1.0	0.259	-0.107	-0.484	0.436	-0.207	-0.108	-0.22	-0.506
C	-0.262	0.259	1.0	0.102	-0.546	-0.167	-0.38	-0.322	0.457	-0.151
Y	0.105	-0.107	0.102	1.0	-0.311	-0.139	0.263	0.346	0.237	0.295
A	0.485	-0.484	-0.546	-0.311	1.0	-0.658	0.328	0.24	0.011	0.396
S	-0.434	0.436	-0.167	-0.139	-0.658	1.0	-0.218	-0.193	-0.479	-0.497
Bus	0.206	-0.207	-0.38	0.263	0.328	-0.218	1.0	0.861	-0.103	0.563
Emp	0.107	-0.108	-0.322	0.346	0.24	-0.193	0.861	1.0	-0.121	0.475
SAR	0.218	-0.22	0.457	0.237	0.011	-0.479	-0.103	-0.121	1.0	0.463
TMCI	0.505	-0.506	-0.151	0.295	0.396	-0.497	0.563	0.475	0.463	1.0

Cross-correlation table for 2008 is provided below.

	PM2	PF2	C2	Y2	A2	S2	Bus2	Emp2	SAR2	TMCI2
PM2	1.0	0.629	-0.088	0.007	0.197	-0.163	0.212	0.072	0.134	0.412
PF2	0.629	1.0	0.096	-0.1	-0.177	0.173	0.043	-0.029	-0.038	-0.015
C2	-0.088	0.096	1.0	0.176	-0.546	-0.264	-0.403	-0.324	0.496	-0.212
Y2	0.007	-0.1	0.176	1.0	-0.341	-0.147	0.107	0.171	0.172	0.146
A2	0.197	-0.177	-0.546	-0.341	1.0	-0.597	0.39	0.286	-0.019	0.432
S2	-0.163	0.173	-0.264	-0.147	-0.597	1.0	-0.17	-0.136	-0.467	-0.397
Bus2	0.212	0.043	-0.403	0.107	0.39	-0.17	1.0	0.886	-0.097	0.635
Emp2	0.072	-0.029	-0.324	0.171	0.286	-0.136	0.886	1.0	-0.118	0.499
SAR2	0.134	-0.038	0.496	0.172	-0.019	-0.467	-0.097	-0.118	1.0	0.399
TMCI2	0.412	-0.015	-0.212	0.146	0.432	-0.397	0.635	0.499	0.399	1.0

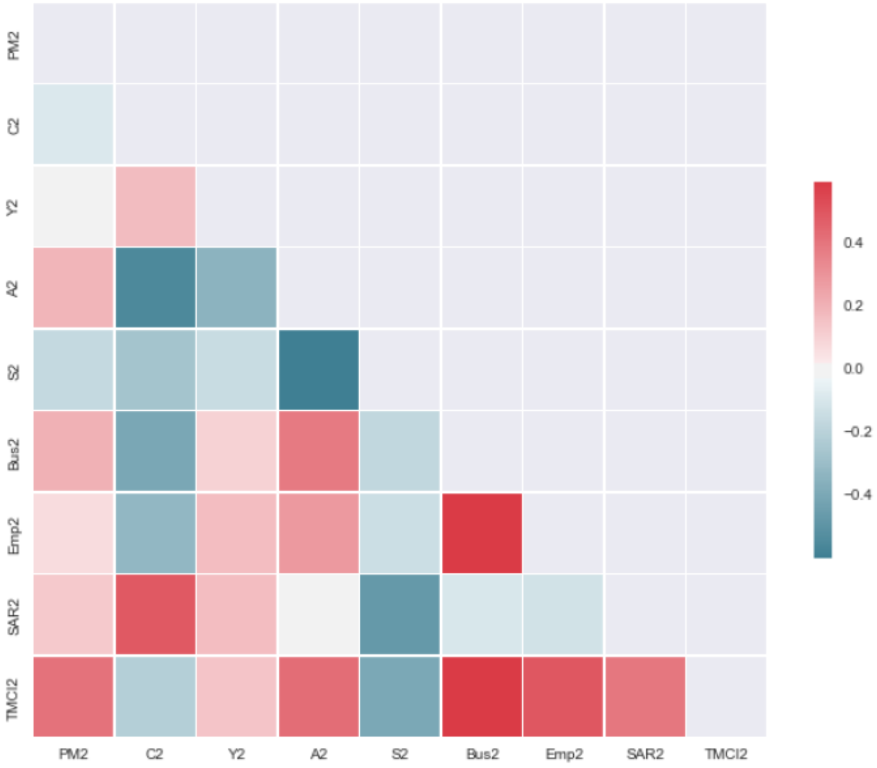
I plotted these correlations to make them more understandable.

Correlation plot for 2011:



For the 2011 data, there was a moderate to high positive correlation between the number of males per capita (i.e., per 100 people) and major crime (*TMCI*), number of adults per capita and major crime, number of people employed per capita and major crime, and number of social assistance recipients per capita and major crime. There was a negative correlation between the number of seniors per capita and major crime. There was also a strong positive correlation between the number of people employed per capita and the number of businesses per capita in that neighbourhood, suggesting that these features might be containing redundant information.

Correlation plot for 2008:



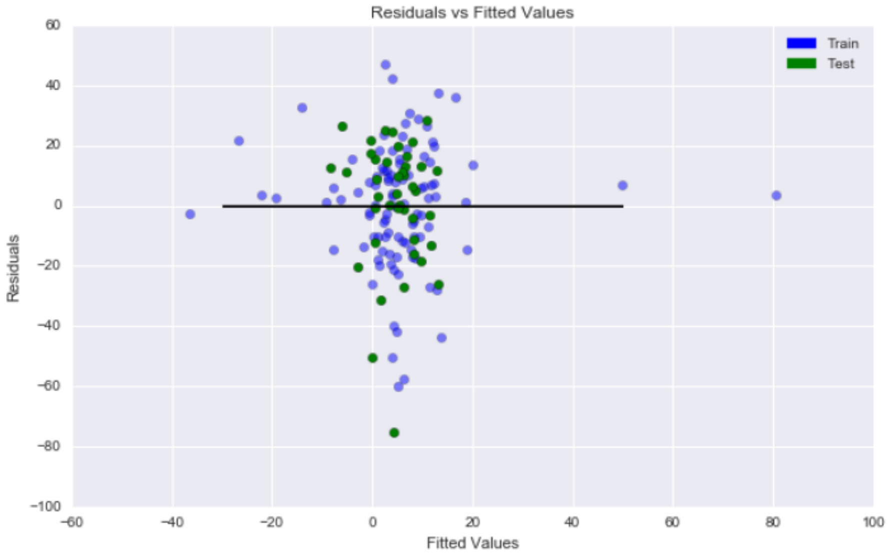
Major crime is primarily associated with males rather than females. This led to removing the percentage of females as a feature. I noticed a strong correlation between the number of people employed per capita and the number of businesses per capita in that neighbourhood for 2008 as well. Ideally, I would 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.

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 train and test sets using a 70:30 split, respectively. I decided to use linear regression because it provides a basic starting point for regression, under the assumption that the relationship between independent and dependent variables is linear. 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 moderate value indicating that the model was not flexible enough to fit the data. We also have to consider that we are neither using the actual features nor 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.

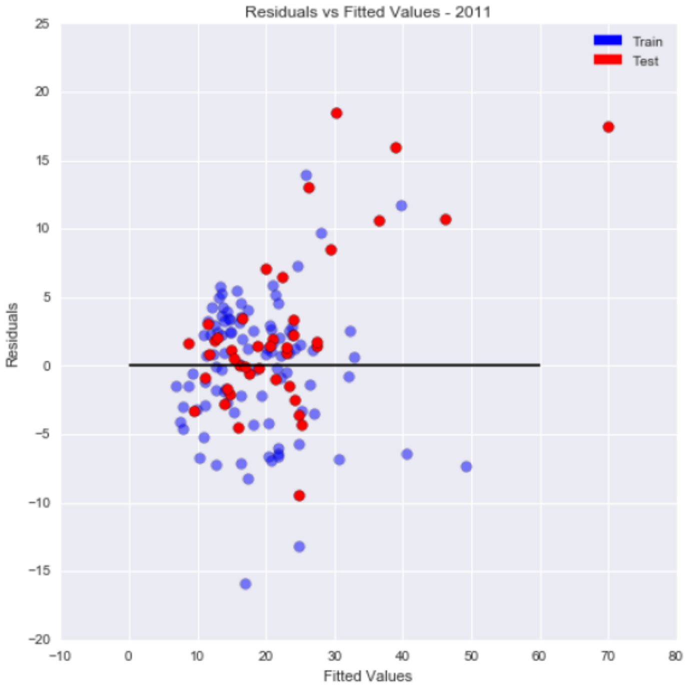


The random forest model had an R^2 value of 85.1% on the training data, which was a big improvement over the linear regression model. However, it performed poorly on the test data

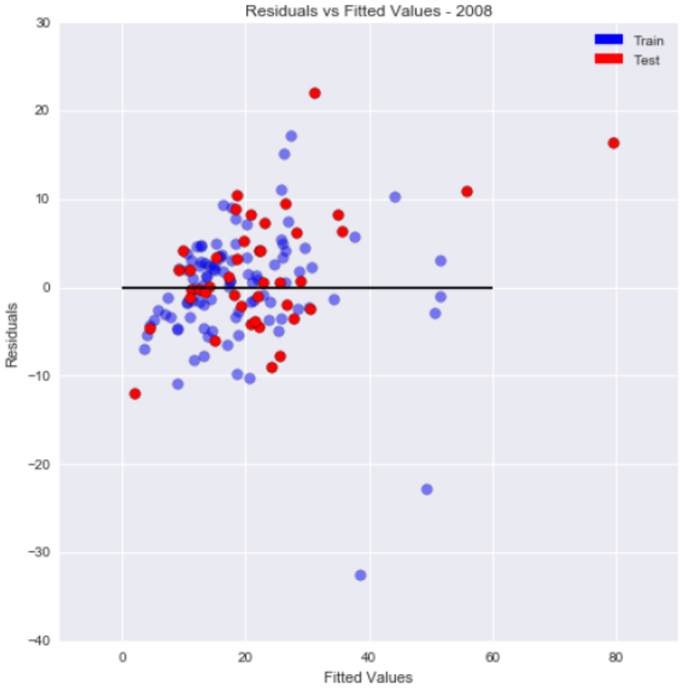
(R^2 value of 9.2%), suggesting that it does not generalize. This could also be noticed in the difference in mean squared error values for the training data (MSE = 83) and for the test data (MSE = 404.9). As an additional step, I did a k-fold cross-validation with k = 5. The average mean squared error 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. It is also likely that the model is wrong for the data. So, my conclusion was that with this model would not be able to successfully predict the percentage of change in major crime in a neighbourhood from 2008 to 2011.

Next, 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 to back-predict total crimes for 2008 and one to back-predict total crimes 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. The residual plot for the 2011 data is provided below.

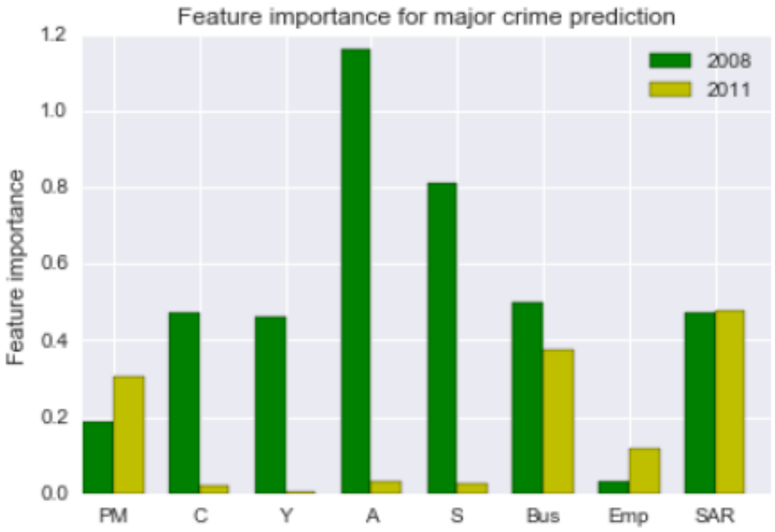


The residual plot for the 2008 data is provided below.



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 standardized coefficients of all the predictors for both models using statsmodels.



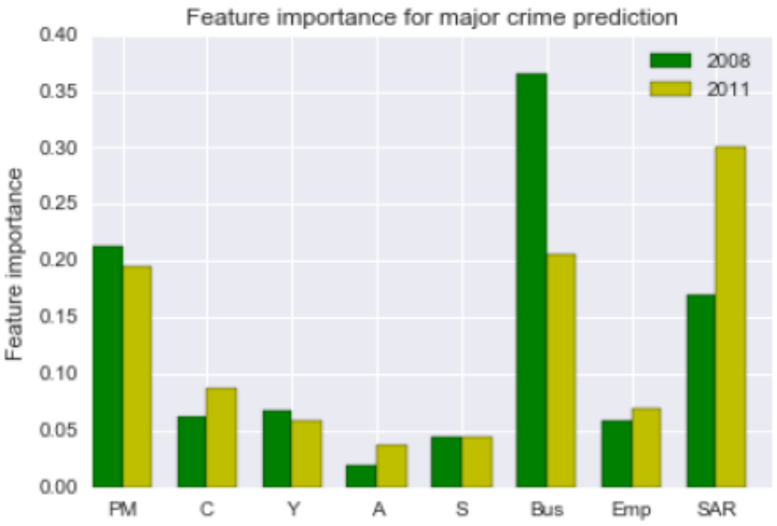
For the 2008 linear regression model, the normalized coefficient values suggest that almost all the features except the number of people employed, are important predictors for major crime in a neighbourhood. These features are people in each of the four age groups, percentage of males, number of businesses, 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.

Full Name	**X axis labels**
Percentage of males	PM
Children per 100 people	C
Youth per 100 people	Y
Adults per 100 people	A
Seniors per 100 people	S
Businesses per 100 people	Bus
Employed per 100 people	Emp
Social Assistance recipients per 100 people	SAR

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 random forest 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.

Since random forest models allow us to find out the most important features for that model, I plotted these.



For both 2008 and 2011, again the percentage of males, number of businesses, and number of social assistance recipients show up as the most important predictors of major crime, similar to the 2011 linear regression model.

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 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 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[7]. 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.

[7] https://en.wikipedia.org/wiki/Regent_Park_Revitalization_Plan

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[8]. 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 having more businesses in each neighbourhood is correlated with a reduction in 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. 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.

[8] http://www.mcss.gov.on.ca/en/mcss/programs/social/