Introduction:

We will be exploring various characteristics and factors of the neighborhoods situated in Toronto and the relationship between them and the volume of crime committed in each associated neighborhood. The analysis will be performed through multiple linear regression, where we will build a model which uses a number of explanatory variables in order to predict the outcome of a response variable. The purpose of this analysis will be to create a model that is descriptive and interpretable in the context of the study and simultaneously is able to make decent predictions on the crime volume of a neighbourhood.

The homicide rates this year in Toronto were at an all-time high since 2005 and it still remains as having the third-highest rate of firearm-related homicides after Winnipeg and Edmonton. While this has been acknowledged and identified as a public health issue, there has been limited research on the trends and demographic patterns associated with the sub-groups of Toronto leading in the reported volume of general crime. A study conducted in 2011 by the FBI reported some factors which affect the volume of crime occurring in communities to be: population density, youth concentration, economic conditions, job availability, cultural factors, and poverty levels. There is a high likelihood some of these factors will surface in our own analysis as well and thus we will keep these mind during the variable selection process. For this analysis, crime is defined whether it is classified a major crime indicator of the following categories: assault, auto-theft, break and enter, robbery, homicide, and shootings.

Note: Census of population is collected every 5 years across Canada and the last reported data was from 2016, so our investigation will only entail data from the year 2016.

Methods:

The first step in the regression analysis is to formulate a data-frame which we will be directly working with in our statistical software. This will require us to clean the two original datasets obtained from the open data portals to produce an easier to work with dataset. Subsequently, we must devise which variables are to be chosen for inclusion in the model. We will also randomly split our data into a training and testing dataset (60-40) in order to validate our model later.

The first set of predictors will be chosen based on context and consistency from other reports and literature examined in the background section of the report. From this set of predictors, we must test for multicollinearity among them as a strong linear relationship between a pair of predictors can make our regression surface unstable. We will first fit a full model (from training dataset) and from there calculate the Variance Inflation Factor (VIF) for each predictor which measures how inflated variances will be due to the relationship of one predictor to the others. Predictors with VIF values exceeding 5 will be identified and we will attempt to reduce VIF values by removing various suspected collinear predictors. If multicollinearity is present in the model, we should arrive at a reduced model. To reduce the model further, manual backwards selection will be performed (required assumptions will be checked prior to this) based on adjusted R^2, corrected Akaike's Information Criterion (AIC), and Bayesian Information Criterion (BIC) (criterions based on maximized likelihood function for each model). We will remove predictor which increases the R^2 and decreases the AIC and BIC, but will not be done blindly and will have context incorporated. We will compare the final model to the one created by automated selection.

After a final model has been constructed and all required assumptions are satisfied, we must perform a validation test for our model using the validation dataset. This will be done by fitting the same model, but under the test dataset where we will inspect if we have minimal differences in the estimated regression

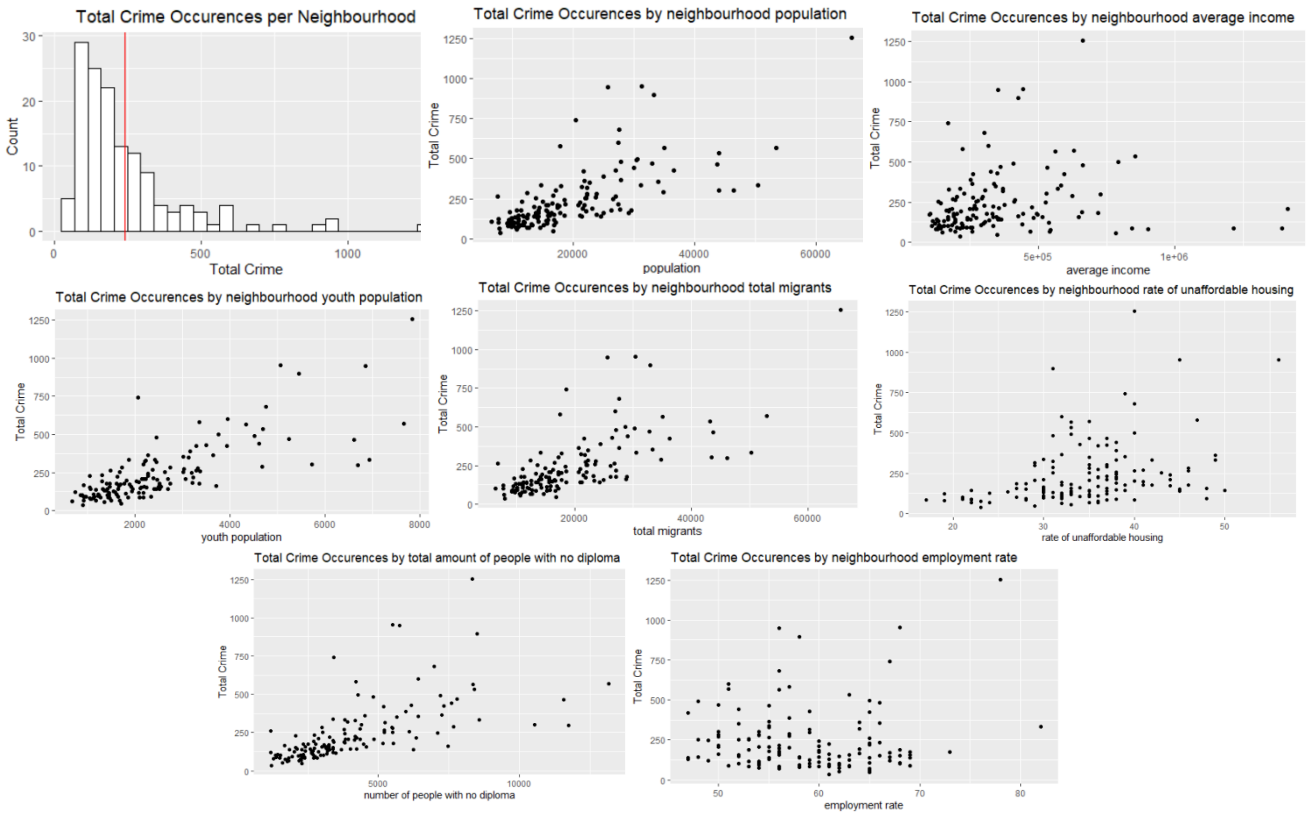
coefficients (inspect summary of model), same predictors appearing significant (build model with automated selection), no new model violations, and a similar adjusted R^2.

Regarding model diagnostics, these will be checked first initially and then periodically throughout the analysis after each significant change to the model. The first two conditions our model should satisfy are (1) our conditional mean response is a single function of a linear combination of the predictors, and (2) the conditional mean of each predictor is a linear function with another predictor. Condition 1 will be checked with the use of a plot of the response against the fitted values, where we would want to see the points randomly scattered around the identity function. To check Condition 2, we will plot all pairs of predictors and inspect if relationships appear linear.

With these conditions satisfied, we are able to meaningfully interpret our residual plots in order to check the remaining assumptions. We must ensure that the true relationship between the mean response and the predictor are linear (mean zero errors – errors should be centered at 0 in plots). The errors must also be uncorrelated and must have common variance (homoscedasticity), where the residual plots should not have clustering, discernible patterns, and fanning (should have equal spread). Errors should also be normally distributed, and this is checked by plotting a QQ plot where we would have points mostly on the line with minimal lifting on tails. Depending on which assumption is not met, we will correct it through the use of a power transformation of either the predictors or the response or both. Assumptions must be verified in order to have an unbiased model with minimum variance. Normality must be satisfied in order to use backwards selection and make meaningful inferences with our model.

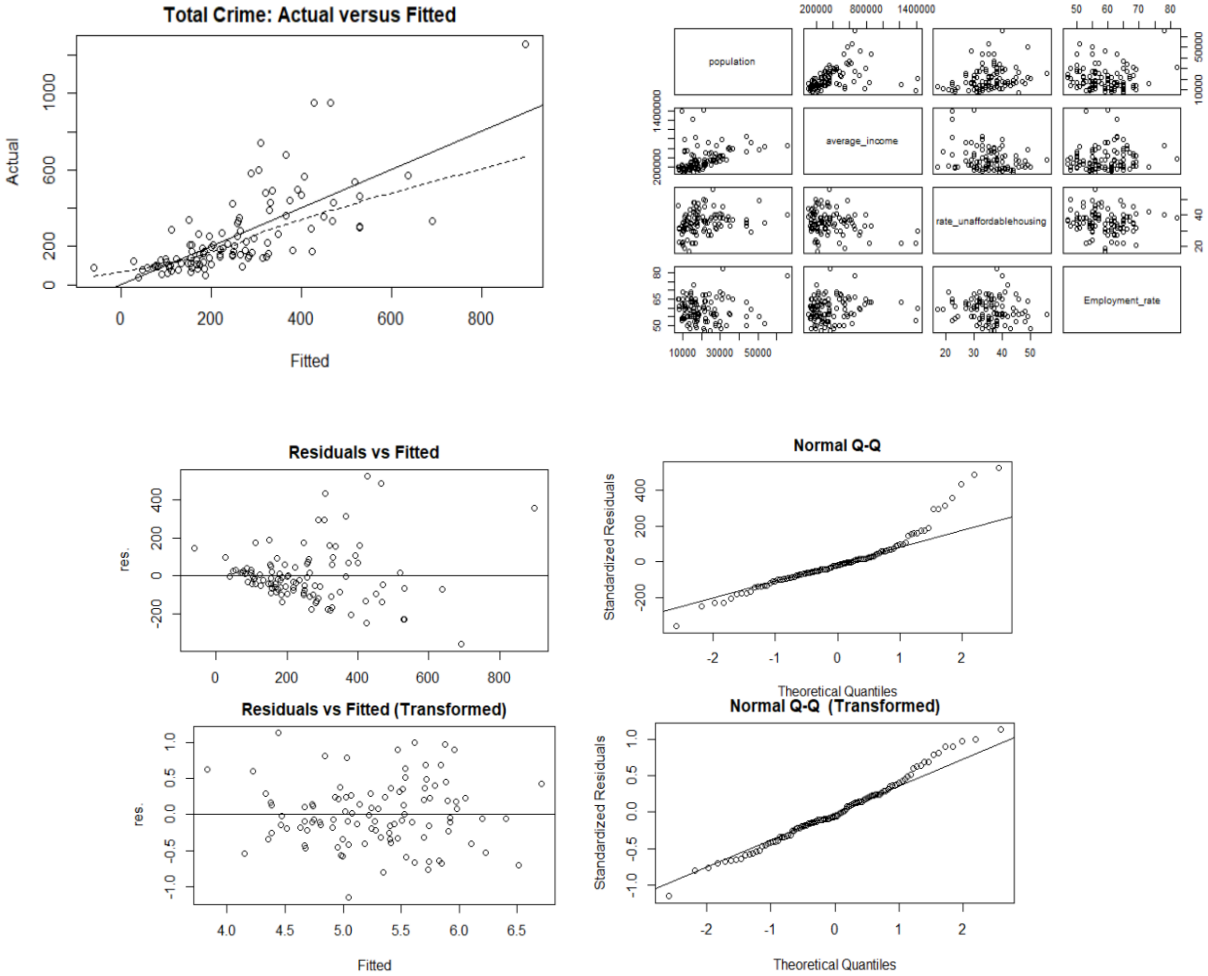
Results:

Based on background information and insight from various literature concerning the topic of study, the following predictor variables for each neighbourhood were chosen to be included in the initial model: population, average income, youth population, total migrants, rate of unaffordable housing, number of people with no diploma of any kind, employment rate. Each predictor seems to visually demonstrate a linear relationship with the response. The response seems to be right skewed which may be an issue in future steps of the analysis (potential transformation). Note: there was no missing data to account for.



Computed VIF values for the set {population, total migrants, total youth, number of people with no diploma} demonstrate severe multicollinearity. The model which produces all VIF values under 5 for all predictors removes the following initial predictors: youth population, total migrants, and number of people with no diploma. The context of these removed variables is in accordance with the VIF values as youth population, total migrants, and number of people with no diploma are subsets of the population variable.

Regarding model diagnostics, Condition 1 and 2 were checked and we can observe both conditions are satisfied (pairwise relationships seems linear, points scattered somewhat evenly across identity line). Successively, we plotted residuals by fitted values and the QQ-plot, and found we have violations for normality and linearity (severe tails on QQ-plot, pattern in residual plot). To correct violation, we performed a logarithmic transformation to both predictors and response as this was the recommended power yielding the highest likelihood from the BoxCox summary. This indeed corrected our violations (refer to figure).



With model assumptions verified, we performed manual backwards selection: refer to following table.

Step 1:

Model	Adjusted R^2	BIC	AIC
Full model	0.6101	149.8493	133.9256
Missing employment rate	0.614	145.1965	131.9266
Missing unaffordable housing rate	0.5992	149.1377	135.8679
Missing average income	0.5901	151.4883	138.2185
Missing Population	0.2808	210.5275	197.2577

Model without the employment rate predictor yields best criterions. We will remove this predictor.

Step 2:

Model	Adjusted R^2	BIC	AIC
Missing employment rate	0.614	145.1965	131.9266
Missing unaffordable housing rate and employment rate	0.6031	144.4839	133.868
Missing average income and employment rate	0.5935	146.9978	136.382
Missing Population and employment rate	0.2828	206.61	195.9941

Removing other predictors does not yield a better a model. Furthermore, automated stepwise selection yielded and identical model which is reassuring. Assumptions were re-checked and observed as satisfied.

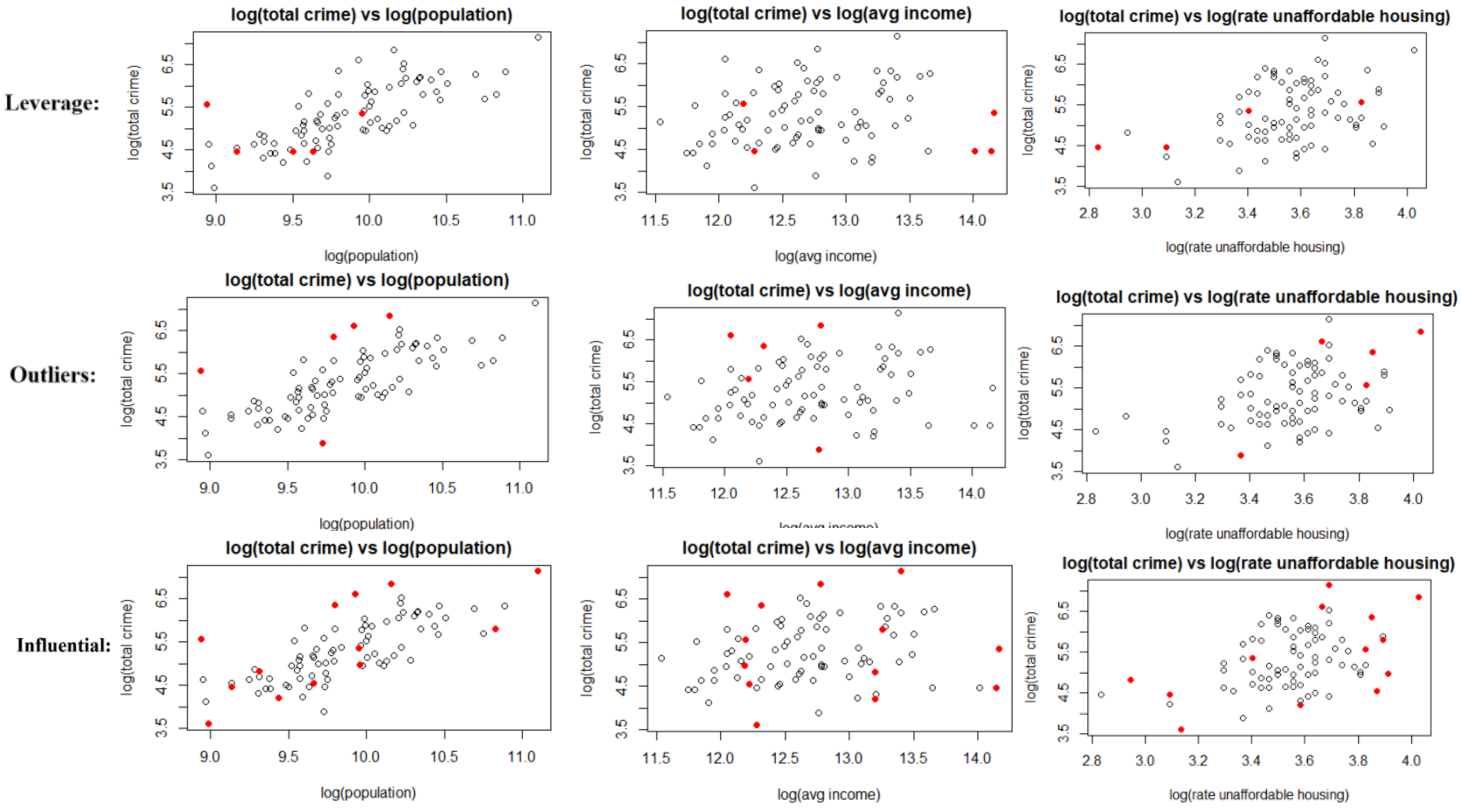
log(Volume of crime) ~ log(population) + log(average income) + log(rate of unaffordable housing)

Validation:

An EDA was performed on the validation data and variable distributions seemed similar enough to the training data. Fitting a model with same predictors and transformation as the final model on the validation data yields minimal differences in estimated regression coefficients and adjusted R^2 values (refer to summary). Furthermore, automated stepwise selection constructed a model with the same predictors implying both datasets have the same significant predictors.

Model	Intercept estimate	log(population) estimate	log(average income) estimate	log(rate of unaffordable housing) est.	Adjusted R^2
Train	-5.2104	1.2248	-0.2557	0.4761	0.614
Test	-4.2801	1.0207	-0.1647	0.4698	0.6088

Leverage points, Outliers, and Influential: Points



Discussion:

The final model can be interpreted as: if all other predictors are held constant, if the logarithmic population increases by one person, the average logarithmic volume of crime for that neighbourhood will increase by a factor of 1.2248. The model also has decent predictive power with an adjusted R^2 of 0.614. We can conclude that we have satisfied our initial goal as our model is also both descriptive and interpretable where we have three significant predictors (which have shown up in previous studies) with a clear linear relationship with the response under a natural logarithmic transformation.

From the results of the analysis, we can conclude that the population, average income, and rate of unaffordable housing of each neighborhood are most influential when it comes to predicting the average volume of crime for the corresponding neighborhood which is backed up contextually by various literature cited. Add more context

Regarding limitations, even after transformations, there is still evidence of non-normal occurrences as the QQ-plot of final model has deviation in the tails. This can be possibly related to the amount of problematic observations we had in the model where both the training and testing datasets had approximately 20% of the points to be influential, which can heavily skew the model. This contextually makes sense as Toronto has several both very wealthy and impoverished neighbourhoods. In the future, it would be beneficial to collect data from years other than 2016 to improve the variability of the model.

References:

FBI. (2012, November 5). Variables affecting crime. FBI. Retrieved December 20, 2021, from https://ucr.fbi.gov/hate-crime/2011/resources/variables-affecting-crime

Toronto crime rates in 2021. Vilkhov Law. (2021, September 14). Retrieved December 20, 2021, from https://vilkhovlaw.ca/toronto-crime-rates/

Reuters. (2021, November 26). *Murder rate in Canada hits 15-year high in 2020, StatsCan Reports*. nationalpost. Retrieved December 20, 2021, from https://nationalpost.com/news/canada/murder-rate-in-canada-hits-15-year-high-in-2020-statscan-reports

Portals/Websites data was obtained from:

Neighbourhood Profiles - City of Toronto Open Data Portal

Neighbourhood Crime Rates 2020 | Neighbourhood Crime Rates 2020 | Toronto Police Service Public Safety Data Portal