



HUMBOLDT UNIVERSITY BERLIN

SUMMER TERM 2018

Economics of Crime

A CROSS-SECTIONAL ANALYSIS OF TORONTO NEIGHBOURHOODS

Gabriel Blumenstock, Felix Degenhardt and Haseeb Warsi

STATISTICAL PROGRAMMING LANGUAGES

supervised by

Petra BURDEJOVA

August 6, 2018

Abstract






Your abstract.

All scripts can be found on:

Contents

List of Figures	ii
List of Tables	iii
1 Introduction	1
1.1 Motivation	1
1.2 Literature review	1
2 Initial Explanatory Analysis	2
2.1 Data Structure	2
3 Feature Engineering	2
3.1 Exploratory Data Analysis	4
3.2 Cluster Analysis	5
4 Regression and Results	8
4.1 Performing an automated regression analysis	8
4.2 Regression Results	11
4.3 Validity of the Regression Results	13
5 Further Regression Approaches	14
5.1 log-Transformed Data	15
5.2 Spatial Regression	15
5.3 Generalized Linear Model with Poisson distribution	16
5.4 Validity	16
6 Conclusion	17
Bibliography	20
A Appendix	21

List of Figures

1	Average Income by Neighbourhood adjusted.	6
2	k-means cluster analysis crime.	7
3	Heatmap of clusters of crime	7
4	Major Crime Indicators Toronto.  Readme.md	21
5	Histogram of total crime Toronto.  Readme.md	22
6	Heatmap of Assaults by Neighbourhood.  Readme.md	22
7	Heatmap of Unemployment Rate by Neighbourhood.	23
8	Heatmap of Auto Theft by Neighbourhood.	23
9	Heatmap of Average Income by Neighbourhood.	24
10	Auto Theft by Neighbourhood adjusted.	25
11	Elbow curve for k selection.  Readme.md	25
12	Distribution of original and transformed crime types.  Readme.md	26
13	Ceres-plots of the break.and.enter-regression with transformed independent variable	27
14	Ceresplot of total.crimes regression with log-transformed variable.	28
15	Ceresplot of total.crimes Poisson-regression.	28

List of Tables

1	Sample of Census Data	3
2	Structure of Merged Data	4
3	Results of original data	12
4	Results of Basic-Power-Transformation	12
5	OLS-assumptions for original dependent variables	13
6	OLS-assumptions for transformed dependent variables	13
7	OLS-assumptions of log-transformed data	16
8	Results of log-transformed data	21
9	Results of Spatial Regression	24
10	Results of Poisson Regression	26
11	Comparison of Different Models for assault	29
12	Comparison of Different Models for auto.theft	29
13	Comparison of Different Models for break.and.enter	30
14	Comparison of Different Models for robbery	30
15	Comparison of Different Models for theft.over	31
16	Comparison of Different Models for drug.arrests	31
17	Comparison of Different Models for total.crime	32

1 Introduction

1.1 Motivation

Already in 1996, the United Nations' "World Resources Institute" published a study that proposed the ongoing urbanization as one of the main challenges of the 21st Century. Predicting that more than 50% of the population of both developing and developed countries will live in urban areas until 2025, today's worldwide population already reached that point according to World Bank data ([WorldBank \(2018\)](#)). Reconsidering that, the [United Nations \(2018\)](#) forecast for 2050, where 68% of the World's population are prognosticated to live in urban areas, strengthens the effects of urbanization in general.

While the population growth in general brings several problems, the corresponding effects of the growth of cities are just as complex. While firms and property owners might see themselves having a better infrastructure or higher profits, urban population runs the risk of increasing air pollution, higher rents or less spaces for daycare programs. All those developments are somewhat visible in every bigger city in industrialized countries. Besides that, a lot of literature suggests that crime rates are higher in cities than in rural areas (e.g. [Malik \(2016\)](#) or [Glaeser und Henderson \(2017\)](#)). Putting the effects of a rising urban population and a higher crime rate in urban areas together, the motivation for this paper was to find appropriate methods to find how a city's characteristics and - to be more specific - its distinct neighbourhoods can explain differences in inner-city crime rates. That is, if there are differences in one city and which characteristics are responsible for those differences and how they can be computed. The remainder of this paper is therefore as follows: After a short literature review of the broad field of economics of crime, we will give an overview of the data selected and the features that are found in the data. Second, we seek to obtain results by OLS with original and transformed data and, in the following sections, also by variations of linear models and generalized linear models. A main goal of this paper is correspondingly to find an appropriate model for the type of data we analyzed. The last section concludes.

1.2 Literature review

The field of studies our topic falls in - the economics of crime - describe a relatively young discipline of economics that first appeared in the 1970s as an approach to explain crime by its incentives ([Becker \(1968\)](#)). This is on one hand the very beginning of economics of crime and, on the other hand, one of three major parts of this field of economics. While this behavioral approach deals with the investigation of rational choices as well as its limitations ([Wright et al. \(2005\)](#)), the two other parts of economics of crime deal with the investigation of external influences on the extent of crime, namely policy - and, more specific, police - measurements as well as socio-economic reasons. The former can be seen as both the effectiveness of policy measurements and its effects on areas where these measurements have not taken place. As an example, one could think about the effectiveness of a more modern police station in one city and an increase in criminal activity in its neighboring city. Both effects go back to the work of [Guerette \(2009\)](#). In contrast to those fields of economics of crime, the analysis socio-economic causes of crime deals with e.g. the effect of unemployment, demographic differences or cultural backgrounds on crime types or the sheer number of crime appearances (e.g. [Lochner und Moretti \(2004\)](#), [Hall \(2007\)](#)). Additionally - as in most of economic disciplines - different kinds of analysis can also be found in the field of economics of crime. Therefore, there can be found several approaches of the analysis of crime rates over time (e.g. [Corman und Mocan \(2000\)](#) for time series analysis or e.g. [Cornwell und Trumbull \(1994\)](#) for

panel data) as well as pure cross-sectional approaches (e.g. [Kelly \(2000\)](#)).

2 Initial Explanatory Analysis

2.1 Data Structure

In an effort to be more transparent, cities are now taking part in an OpenData initiative, resulting in publicly available crime and census data. All data used in this report was collected from the City of Toronto's Open Data portal. The two main datasets used were the 2016 census data from the city of Toronto and crime data set released by Toronto police. Prior to performing any analysis, the dataset had to be gathered, filtered and formatted in a manner that could be used in R. The crime dataset required minimal preprocessing before it was ready to be used, however, the census data had to be filtered and formatted significantly. This entailed creating a function to gather variables of interest from the census dataset and join them into a single data frame. Over 50 variables were collected from the census data, relating to topics such as income, demographics, population, education and housing characteristics, etc.

```
1 #####Create function to get data from main dataset
2 getData <- function(x, characteristic, new_col_name = characteristic) {
3   a <- subset(x, Characteristic == characteristic) #subset dataframe by
4     characteristic
5   a <- as.data.frame(colSums(a[, -which(names(a) %in% c("Characteristic", "Topic"))
6     ]) #Remove characteristic column, leaving only vector of values
7   colnames(a) <- new_col_name #rename colname to characteristic
8   return(a)
9 }
```

Listing 1: Function to get the data

3 Feature Engineering

In addition to the variables that were collected directly from census data, additional variables were also created. The crime statistics provided by the in the crime dataset simply state the number of occurrences of a crime in a neighbourhood, however this number does not take into account the population of the neighbourhood. To make crime statistics more comparable across neighbourhoods with different populations each crime statistic was converted into a crime rate per 10 000, similar to the commonly used crime rate per 100 000.

```
1 #create a function to turn crime variable into crime per hundred thousand rate
2 crime.per.tenthshnd <- function(x) {
3   x / agg.2016[, "population.2016"] * 10000
4 }
```

Listing 2: Crime rate function

In an effort to gain insights about the economic wealth of a neighbourhood we used the average income of each neighbourhood, obtained from the Toronto census data. The average, though, provides an incomplete picture, as it can be heavily skewed by a few high-earning individuals. It ignores the distribution of incomes and income inequality in each neighbourhood, which is considered to be positively correlated with violent crime rates around the world ([Fajnzylber et al. \(2002\)](#)). To get a more complete picture of the economic wealth of each neighbourhood it became pertinent

to look at the median income in each neighbourhood as well. The median income data was not readily available from the census data, but the distribution of incomes in each neighbourhood was. Below is a sample of the census data that showed the distribution of incomes in each neighbourhood:

Table 1: Sample of Census Data

	Characteristic	Agincourt.North	Agincourt.South.Malvern.West	Alderwood
970	Under \$10,000 (including loss)	5170.00	4535.00	1365.00
971	\$10,000 to \$19,999	6325.00	4505.00	1505.00
972	\$20,000 to \$29,999	3520.00	2715.00	1360.00
973	\$30,000 to \$39,999	2465.00	2020.00	1095.00
974	\$40,000 to \$49,999	1895.00	1560.00	950.00
975	\$50,000 to \$59,999	1265.00	1125.00	825.00
976	\$60,000 to \$69,999	865.00	825.00	690.00
977	\$70,000 to \$79,999	655.00	570.00	530.00
978	\$80,000 to \$89,999	435.00	435.00	395.00
979	\$90,000 to \$99,999	365.00	315.00	370.00
981	\$100,000 to \$149,999	530.00	525.00	620.00
982	\$150,000 and over	135.00	165.00	225.00

Using a method of calculating the median from a set of grouped intervals from Statistics Canada (<https://www.statcan.gc.ca/edu/power-pouvoir/ch11/median-mediane/5214872-eng.htm>) and the census information, a function was created to find the median income of each neighbourhood, resulting in the median income of each neighbourhood.

```

1 ###Change Characteristic Vector to specific form
2 census.tmp$Characteristic <- c("0-9999", "10000-19999", "20000-29999", "30000-39999",
3   "40000-49999", "50000-59999",
4   "60000-69999", "70000-79999", "80000-89999", "90000-99999",
5   "100000-149999", "150000-1000000") #create income
6   intervals increasing by 10000, last interval has max of
7   1000000 as assumption
8
9 ###Create Function to Calculate median income using groups
10 Grouped_Median <- function(frequencies, intervals, sep = NULL, trim = NULL) {
11   # If "sep" is specified, the function will try to create the
12   # required "intervals" matrix. "trim" removes any unwanted
13   # characters before attempting to convert the ranges to numeric.
14   if (!is.null(sep)) {
15     if (is.null(trim)) pattern <- ""
16     else if (trim == "cut") pattern <- "\\[[\\]\\(\\)"
17     else pattern <- trim
18     intervals <- sapply(strsplit(gsub(pattern, "", intervals), sep), as.numeric)
19   }
20
21   Midpoints <- rowMeans(intervals) #midpoint of interval
22   cf <- cumsum(frequencies)
23   Midrow <- findInterval(max(cf)/2, cf) + 1
24   L <- intervals[1, Midrow] # lower class boundary of median class
25   h <- diff(intervals[, Midrow]) # size of median class
26   f <- frequencies[Midrow] # frequency of median class
27   cf2 <- cf[Midrow - 1] # cumulative frequency class before median class
28   n_2 <- max(cf)/2 # total observations divided by 2
29
30   unname(L + (n_2 - cf2)/f * h)
31 }

```



```

28
29 median.income <- cbind.data.frame(neigh.codes, #apply grouped median function to
    income intervals from census data
30                                     as.data.frame(sapply(census.tmp[, -which(names(
    census.tmp) %in% c("Topic", "Characteristic")
    )], function(x) {Grouped_Median(x, intervals =
    census.tmp$Characteristic, sep = "-")}))
31 colnames(median.income) <- c(colnames(neigh.codes), "median.income") #rename
    columns
32 agg.2016 <- join(agg.2016, median.income[, -which(names(median.income) %in% c("
    Neighbourhood"))], by = "Hood_ID") #join median income to agg.2016

```

Listing 3: Function for grouped median

Table 2: Structure of Merged Data

	Neighbourhood	Hood_ID	avg.income	median.income
1	Agincourt.North	129	30414.00	20901.90
2	Agincourt.South.Malvern.West	128	31825.00	22237.35
3	Alderwood	20	47709.00	36711.66
4	Annex	95	112766.00	43035.87
5	Banbury.Don.Mills	42	67757.00	42493.22
6	Bathurst.Manor	34	45936.00	29612.30
7	Bay.Street.Corridor	76	56526.00	27511.69
8	Bayview.Village	52	52035.00	32818.51

Looking at the above table, it becomes readily apparent that certain neighbourhoods have a much higher average income than median income, indicating the presence of a few high earners and income inequality. With all variables collected into a single data frame, analysis could begin.

3.1 Exploratory Data Analysis

A first step to analyzing crime in Toronto was a look at the prevalence of different crime types in Toronto. Using the *ggplot* and *dplyr* packages in *r* to aggregate the data by crime type and plot it, one can see the most common type of crime in the city is assault, which overshadows every other type of crime (Figure 15 in Appendix A).

```

1 ###Group crimes by MCI
2 mci.group <- group_by(crime.dt, MCI)
3 crime.by.mci <- dplyr::summarise(mci.group, n=n()) #count of events by MCI
4 crime.by.mci <- crime.by.mci[order(crime.by.mci$n, decreasing = TRUE),] #order
    crime by type from most to least
5
6 plot(ggplot(aes(x = reorder(MCI, n), y = n), data = crime.by.mci) +
7     geom_bar(stat = 'identity', width = 0.5) +
8     geom_text(aes(label = n), stat = 'identity', data = crime.by.mci, hjust = -0.1,
9         size = 3.5) +
10    coord_flip() +
11    xlab('Major Crime Indicators') +
12    ylab('Number of Occurrences') +
13    ggtitle('Major Crime Indicators Toronto 2016') +
14    theme_bw() +
15    theme(plot.title = element_text(size = 16),
16        axis.title = element_text(size = 12, face = "bold")))

```

Listing 4: ggplot function code

To see the distribution of our crime statistics a function was created to plot a histogram (Figure 16 in Appendix A). Since the shape of a histogram can be heavily influenced by the bin widths used to divide a continuous variable, an additional argument allowing the user to specify bin widths was added to the function in order to add flexibility.

Using this function on the different crime variables it is apparent that the distribution of crime in Toronto is left-tailed, with many neighbourhoods having a low level of crime. However, there are a few neighbourhoods that are on the far-right of the distribution, indicating a significantly high number of crime occurrences. To gain a better understanding of the differences between neighbourhoods it was pertinent to see the data plotted on a heat map of the city. With a quick visual inspection we can see how each neighbourhood differs. The process of creating heat maps involved first reading the necessary shapefile (obtained from the City of Toronto Open Data Catalogue) and joining it to the retrieved census data.

```

1 #Heatmap of toronto by population
2 # Read the neighborhood shapefile data and plot
3 geo.data <- data.frame(agg.2016)
4 geo.data$Hood_ID <- str_pad(geo.data$Hood_ID, width = 3, side = 'left', pad = '0')
5
6 # the path to shape file
7
8 toronto <- readOGR(dsn = "Shapefiles/Neighbourhoods_Toronto" , "NEIGHBORHOODS_WGS84"
9 )
10
11 # fortify and merge: muni.df is used in ggplot
12 toronto@data$id <- rownames(toronto@data)
13 toronto.geo <- fortify(toronto)
14 toronto.geo <- join(toronto.geo, toronto@data, by="id")
15 names(toronto.geo)[names(toronto.geo) == 'AREA_S_CD'] <- 'Hood_ID'
16
17 toronto.geo <- join(geo.data, toronto.geo, by = "Hood_ID")

```

Listing 5: Crime rate function

With the data in a usable format, a function was then created to plot each variable according to each neighbourhood.

Several problems occurred when plotting variables using this manner, particularly the presence of outliers. The presence of an outlier in the data resulted in a scale that made it difficult to visualize the difference between the majority of neighbourhoods. In the case of Toronto, this is evident in the occurrence of auto thefts by neighbourhood. One neighbourhood had a significantly higher number of auto thefts than the rest, which made it impossible to see any differences between other neighbourhoods.

The solution to this problem was to create a function for plotting a heat map that had an upper limit. Any number above this limit would appear as one colour and the scale of the heat map would be preserved, resulting in a more readable and insightful heat map.

3.2 Cluster Analysis

With 140 neighbourhoods in the city, it may be helpful to group neighbourhoods that exhibit similar characteristics together. One popular and often used method of clustering is the k means clustering algorithm. The k means algorithm is used to assign each observation to a cluster, based on similar characteristics in order to simplify the number of groups to look at. The algorithm first

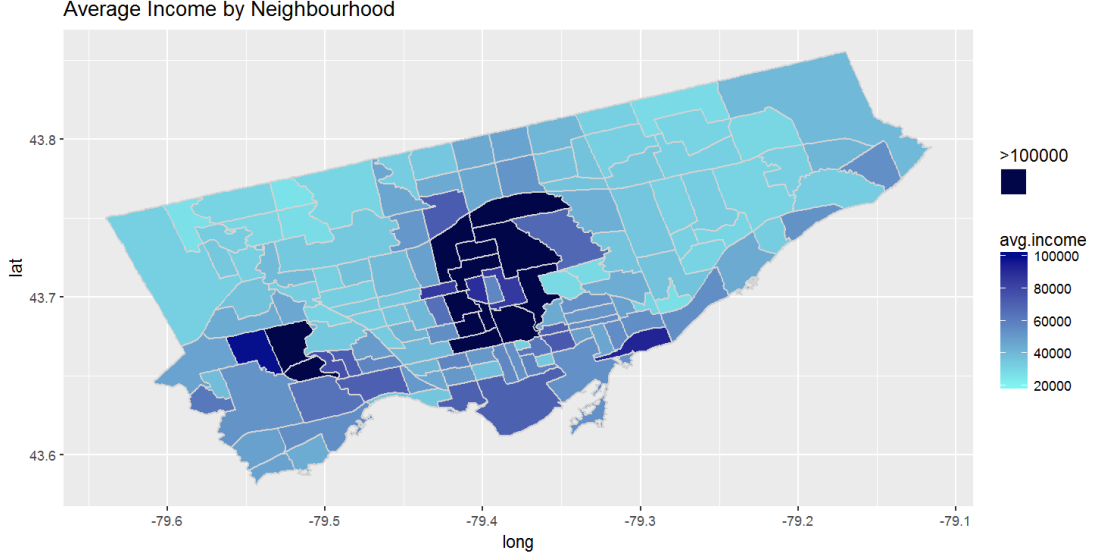


Figure 1: Average Income by Neighbourhood adjusted.

places a specified number of centroids and along the specified features. Each centroid defines a cluster. Each data point is then assigned to its nearest centroid, based on the squared Euclidean distance. The algorithm then moves each centroid towards the mean of all data points assigned to that centroid. The new clusters are calculated and the process is repeated until some stopping criteria is met. The goal being to minimize the sum of the distance between points within each cluster.

$$J(V) = \sum_{i=1}^C \sum_{j=1}^{C_i} (\|x_i - v_j\|)^2 \quad (1)$$

where

$V = \{v_1, v_2, \dots, v_c\}$ are the set of centers,

$\|x_i - v_j\|$ is the Euclidian distance between x_i and v_j ,

c_i is the number of data points in the i^{th} cluster and

c is the number of cluster centers.

The first thing to be decided was the number of clusters to initially specify. Too few clusters would ignore nuances between groups and too many clusters would not simplify anything. A common method of determining the optimal number of clusters involves testing a varying number of clusters and looking at the sum of squares within clusters. The optimal number is the point at which the sum of squares decreases most sharply and any further increases in the number of clusters results in only a small decrease in the sum of squares within clusters. The easiest method of visualizing this is plotting the total sum of squares within clusters with the number of clusters and finding the “elbow” of the curve.

Above we can see the optimal number of clusters to be 8, based on using the different crime statistics of each neighbourhood. Plotting a cluster plot shows us the neighbourhoods that belong to each cluster. However, it is not possible to see exactly which neighbourhood belongs to each cluster. To see this, a new variable with each neighbourhood’s cluster number was added to the aggregated dataset and plotted using a new heat map function.

From the figure above, it is apparent that the physical location of each neighbourhood does not determine which cluster the neighbourhood belongs to. Although we see several neighbourhoods

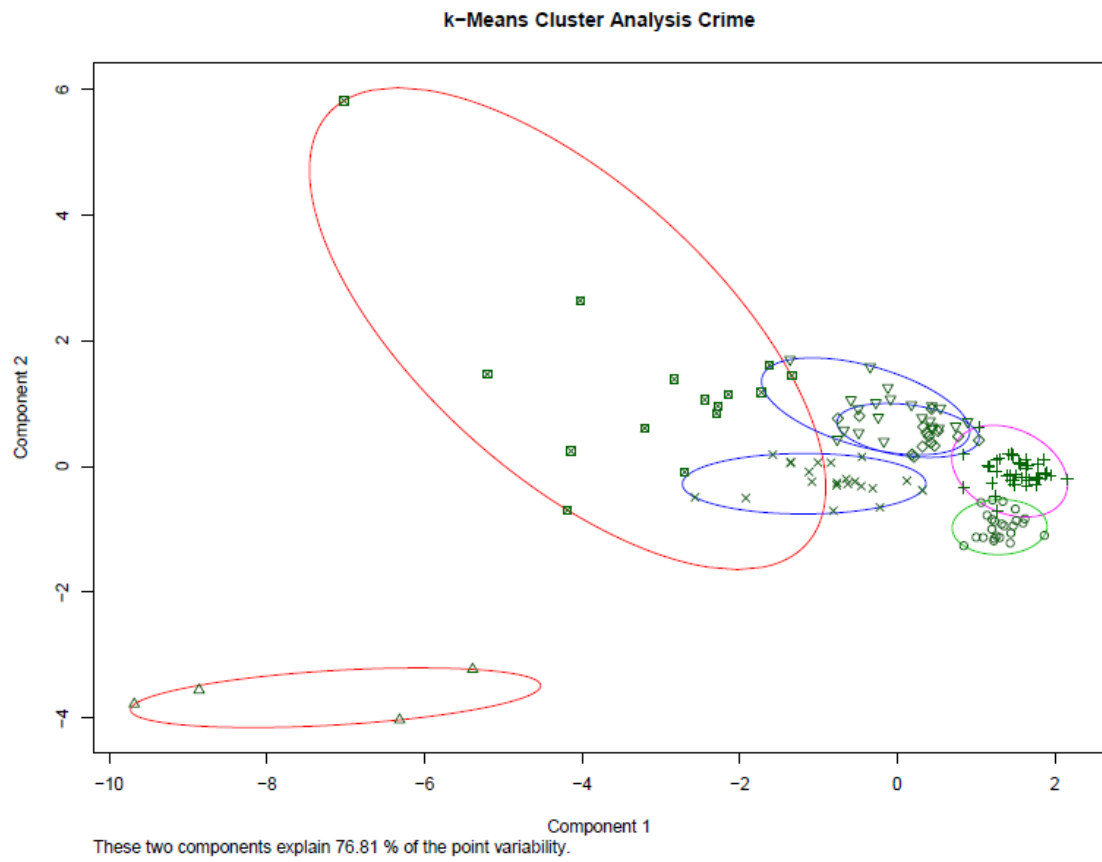


Figure 2: k-means cluster analysis crime.

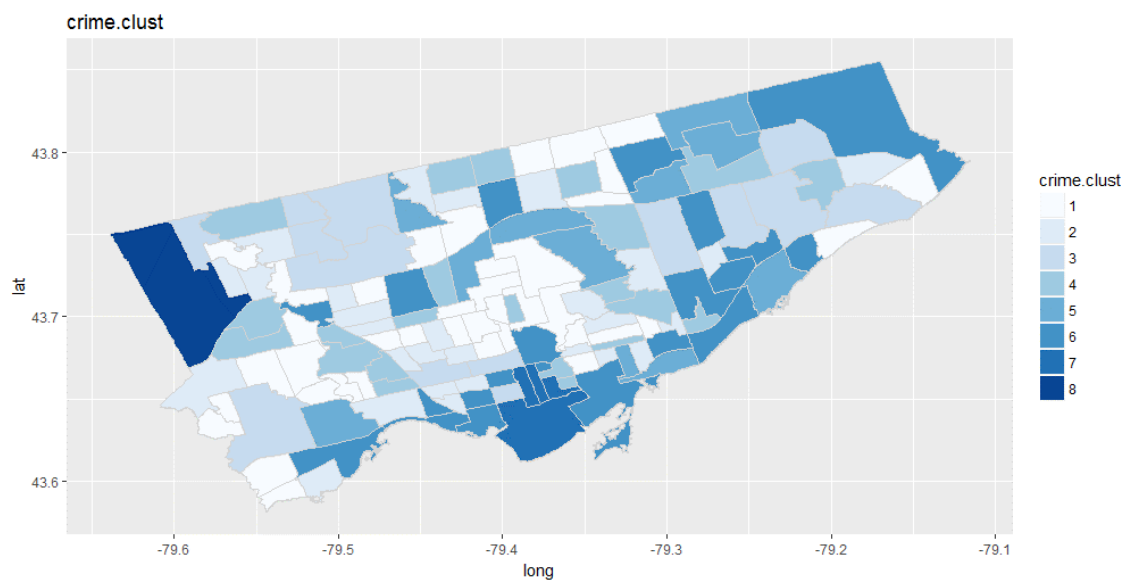


Figure 3: Heatmap of clusters of crime

that are near each other belonging to the same cluster, there are several neighbourhoods that are on opposite ends of the city but still belong to the same cluster.

```

1 #Define function to label each neighbourhood with cluster label
2 clust_func <- function(x, no.of.clust, title) {
3   kc <- kmeans(x, no.of.clust) #use optimal number of clusters
4   z1 <- data.frame(x, kc$cluster) #create data frame with cluster number for each
      neighbourhood
5   print(clusplot(z1, kc$cluster, color=TRUE, shade=F, labels=0, lines=0, main=paste
      ('k-Means Cluster Analysis', title, sep = " ", collapse = NULL))) #cluster
      plot
6   y <- as.factor(z1$kc.cluster)
7   return(y)
8 }
9 agg.2016$ethnic.clust <- clust_func(agg.kmeans.ethnic, 4, "Ethnic Characteristics")

```

Listing 6: Cluster Function

```

1 #Define function to generate heat maps for cluster variables (input dataframe and
      desired cluster)
2 heat_map_clust <- function(data, x) {
3   plot(ggplot(data= data, aes(x=long, y=lat, group=group)) +
4     geom_polygon(aes_string(fill= x)) + # draw polygons and add fill with
      variable
5     geom_path(color="light grey") + # draw boundaries of neighbourhoods
6     coord_equal() +
7     geom_tile() + #plot fill as geom tiles
8     scale_fill_brewer(palette="Blues") + #choose colour palette
9     labs(title= x)) #set title # render the map
10 }

```

Listing 7: Heatmap Cluster Function

4 Regression and Results

4.1 Performing an automated regression analysis

To examine and to quantify the effects that different neighbourhood characteristics have on different crime types, we perform linear regressions. More precisely, we run one regression for each of the seven crime types.

While the crime types are always set to be the dependent variable, we have various neighbourhood characteristics available as possible explanatory variables. Certainly, it is not reasonable to include all of them at once: As some of the variables like *youth* and *male.youth* are highly correlated, we would face the problem of imperfect multicollinearity, leading to high standard errors and unstable estimators. Therefore, it is necessary to select those variables that appear to be the most influential on crime and on the same time describe a different neighbourhood characteristic and to exclude all other variables prior to running the regression. After having tried out different combinations of variables and examining their regression results regarding the R squared, standard errors of the estimators, and the correlation among the regressors, the following variables appeared the most reasonable for us to be included: The number of young males between 15 and 24 years living in the respective neighbourhood, the number of people who have less than a high school degree, the number of people who belong to the low-income class by earning less than 40.000 CAD,

and the number of immigrants living in the neighbourhood. To better compare the effects among different crime types, we used the same combination of independent variables for each regression. As described in later subchapters, however, we could not completely eliminate the correlations between the regressors as most variables like income and age variables are already correlated despite referring to different neighbourhood characteristics.

After having defined the dependent and independent variables, we are able to perform the regression analysis as done in the (Q) Automated Regression Quantlet. As implied by its name, we automated the seven regressions by implementing a for-loop.

Prior to running the for-loop, however, some preparations had to be made: Besides converting the data table *agg.2016* from the (Q) Merging Quantlet into a data frame called *r*, creating empty lists and data frames for the regression results and their diagnosis tests to be stored in, and defining the vector *crimetypes* which contains the names of the different crime types to loop over, we defined the function *FindBestExponent*:

```
1 FindBestExponent <- function (p, x) {
2   y <- bcPower(x, p)
3   shapiro.test(y)$statistic
4 }
```

Listing 8: Automated Regression Quantlet – rows 5-15 (without comments)

Based on a numeric vector *x* and a value *p*, the function performs a basic power transformation based on *p* as the exponent and the values of *x* as bases, and saves the transformed data inside *y*. The transformed data is then tested for normality by using the Shapiro-Wilk test. The null hypothesis in this case states that the assessed variable is normally distributed. Therefore, a high p-value implies that the assumption of normal distribution cannot be rejected. We later use this function to find the best basic power transformation to shift the dependent variable towards the normal distribution. OLS-estimations do not necessarily require this condition to yield unbiased estimates, and we will assess in the next subchapters whether the transformation has led to a qualitative improvement of the regression results or not.

After these preparations, the for-loop is then performed, leading to seven regression results and their diagnosis tests, stored in the previously created lists and data frames:

```
1
2 for (i in crimetypes){
3
4   r$tmp <- r[,i]
5
6   firstmodel <- lm(tmp~male.youth + less.than.high.school + low.income
7                     + immigrants, data=r)
8   regressionresults.first[[i]]<-summary(firstmodel)
9   regressionstargazer.first[[i]] <- (firstmodel)
10
11   ols.ass.first[i, "means"] <- mean(firstmodel$residuals)
12
13
14   ols.ass.first[i, "bptests"] <- bptest(firstmodel)$p.value
15
16   ols.ass.first[i, "swtests"] <- shapiro.test(residuals(firstmodel))$p.value
```

```

17
18 ols.ass.first[i, "vif1"] <- vif(firstmodel)[1]
19 ols.ass.first[i, "vif2"] <- vif(firstmodel)[2]
20 ols.ass.first[i, "vif3"] <- vif(firstmodel)[3]
21 ols.ass.first[i, "vif4"] <- vif(firstmodel)[4]
22 #corrplot::corrplot(cor(r_assault[c()],))
23
24 ols.ass.first[i, "cortest1"] <- cor.test(r$male.youth,
25                                         firstmodel$residuals)$p.value
26 ols.ass.first[i, "cortest2"] <- cor.test(r$less.than.high.school,
27                                         firstmodel$residuals)$p.value
28 ols.ass.first[i, "cortest3"] <- cor.test(r$low.income,
29                                         firstmodel$residuals)$p.value
30 ols.ass.first[i, "cortest4"] <- cor.test(r$immigrants,
31                                         firstmodel$residuals)$p.value
32
33 crPlots(firstmodel)
34 ceresplots.first[[i]] <- recordPlot()
35
36 rm(firstmodel)
37 }

```

Listing 9: Automated Regression Quantlet – rows 38-87 (without comments)

To simplify the loop regarding column referencing, the respective dependent variable is first temporarily stored in an extra column in the *r* data frame, *r\$tmp*.

To transform the dependent variables towards a normal distribution, outliers of the dependent variables are removed first. It is reasonable to do this step before the actual transformation, as the optimal transformation parameter would be heavily influenced by those extreme values. After having examined the distribution of all seven crime types, it appeared to be the most reasonable to define those neighbourhoods as outliers, which differed from the mean of all neighbourhoods by more than 2.5 interquartile ranges. Between four and seven neighbourhoods were therefore eliminated from each regression. The data frame *rtmp* is then created, which excludes the outlying neighbourhoods.

As we then want to perform a basic power transformation, we need all observed numbers of crimes to be different from zero. Therefore, we artificially convert possibly occurring zeros into ones, which should not affect the validity of our results. We then calculate the exponent which maximizes the p-value of the Shapiro-Wilk test for our basic power transformation. This is done by implementing the previously defined *FindBestExponent*-function inside the *optimize*-command.

Having obtained the optimal exponent, we then transform the dependent variable and store it in an extra column *tmp.bp* in the *rtmp*-data frame. As we can see in the following graphs, the transformed dependent variables follow much more a normal distribution than the original ones and the p-value of the respective Shapiro-Wilk test is increased in most of the cases. Two exceptions are Theft Over and Total Crime, where the transformation failed to shift the distribution towards a normal distribution. To better compare the distributions with the normal distribution, a normal density line is added to each graph (See Figure 15, Appendix A).

Using the transformed dependent variable, we then perform the seven regressions again and store them in the *regressionresults*-list. In the second part of the for-loop, we then examine whether the OLS-assumptions are met for all regressions: Regarding the error term, we test whether it has

a population mean of zero, whether it is homoscedastic using the Breusch-Pagan test and whether it is normally distributed using again the Shapiro-Wilk test. The last condition is an optional one as it is not required for the validity of the OLS-assumption, however, normally distributed error terms allow us to perform statistical hypothesis testing of the estimated parameters and generate reliable confidence intervals. All these results are stored in the *ols.ass*-data frame. As the observation order is random, we do not check for serial correlation of the error term.

Regarding the independent variables we need to assess whether there is no multicollinearity among them. This is done by calculating the variance inflation factor. As a rough guideline, a variance inflation factor that is higher than 10 indicates serious multicollinearity. Furthermore, we test whether there is a correlation of the independent variables with the error terms using the correlation test and extracting its p-value. The null hypothesis in this case states that there is no correlation between the variables. Again, the results are stored in the *ols.ass* data frame.

Finally, we check the assumption that the relation between each independent variable and the dependent variable is linear by creating so-called Ceres-plots and saving them inside the *ceresplots* list. In these plots, the terms from each independent variable are added to the residuals and then plotted against the independent variable itself. One can then access via the graphs whether the assumption of linearity is adequate or not.

To be able to assess whether we could achieve a qualitative improvement when transforming the dependent variable, we finally run the for-loop again, this time without excluding outliers and transforming the dependent variable, and saving the results in *regressionresults.first*, *ols.ass.first*, and *ceresplots.first*. This could have been easily implemented inside the first loop, however, we decided to perform two loops so that each loop is more clearly arranged.

4.2 Regression Results

After having performed the automated regressions, we now examine the results that are stored in the data frames *regressionresults* and *regressionresults.first*.

The following two tables contain the estimated parameters, their level of significance, and standard errors. Furthermore, the number of observations, the (Adjusted) R Squared, the Residual Standard Error, and the F-Statistic are reported for each regression. The first one refers to the seven regressions using transformed data while the second table documents the results when we do not transform the dependent variable.

Overall, the results do not seem to change significantly between the two tables: While the (Adjusted) R Squared stays roughly the same within each of the seven regressions, we obtain the same amount of significant estimations in each table. Furthermore, with an exception of few cases, the direction of each parameter also stays the same. It therefore does not seem to make a difference whether we use the first or the second table to draw our conclusions about which neighbourhood characteristics affect crime rates in Toronto in which way. As we will see in the next subchapter, some diagnostic tests performed better in the case of transformed data. Therefore, we will now use the first table for our interpretation.

Before interpreting the results, however, we should consider what type of results we are expect-

Table 3: Results of original data

	<i>Dependent variable:</i>						
	assault (1)	auto.theft (2)	break.and.enter (3)	robbery (4)	theft.over (5)	drug.arrests (6)	total.crime (7)
male.youth	0.001 (0.001)	0.0004 (0.0004)	0.001** (0.0004)	0.001*** (0.0004)	0.0002 (0.0004)	0.0001 (0.001)	0.001** (0.001)
less.than.high.school	0.0004* (0.0002)	0.0004** (0.0002)	-0.001*** (0.0002)	0.001*** (0.0002)	-0.0002 (0.0002)	0.001*** (0.0002)	0.0003 (0.0002)
low.income	0.001*** (0.0001)	0.0001 (0.0001)	0.001*** (0.0001)	0.0002** (0.0001)	0.0004*** (0.0001)	0.001*** (0.0001)	0.001*** (0.0001)
immigrants	-0.001*** (0.0001)	-0.00005 (0.0001)	-0.0004*** (0.0001)	-0.0003*** (0.0001)	-0.0002*** (0.0001)	-0.0004*** (0.0001)	-0.001*** (0.0001)
Constant	4.625*** (0.315)	2.381*** (0.234)	3.889*** (0.246)	1.700*** (0.237)	0.396** (0.200)	1.938*** (0.276)	7.197*** (0.310)
Observations	136	135	136	136	133	135	135
R ²	0.618	0.361	0.453	0.507	0.330	0.382	0.647
Adjusted R ²	0.606	0.341	0.436	0.492	0.309	0.363	0.636
Residual Std. Error	1.545	1.197	1.210	1.218	1.051	1.391	1.509
F Statistic	52.896***	18.370***	27.096***	33.627***	15.760***	20.126***	59.565***

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 4: Results of Basic-Power-Transformation

	<i>Dependent variable:</i>						
	assault (1)	auto.theft (2)	break.and.enter (3)	robbery (4)	theft.over (5)	drug.arrests (6)	total.crime (7)
male.youth	0.097*** (0.020)	0.023*** (0.008)	0.024*** (0.007)	0.023*** (0.005)	0.009*** (0.002)	0.026*** (0.007)	0.202*** (0.037)
less.than.high.school	-0.005 (0.008)	0.008** (0.003)	-0.012*** (0.003)	0.005*** (0.002)	-0.003*** (0.001)	-0.001 (0.003)	-0.008 (0.014)
low.income	0.032*** (0.005)	-0.001 (0.002)	0.011*** (0.002)	0.003** (0.001)	0.002*** (0.0005)	0.008*** (0.002)	0.054*** (0.009)
immigrants	-0.028*** (0.003)	-0.0003 (0.001)	-0.008*** (0.001)	-0.004*** (0.001)	-0.002*** (0.0003)	-0.008*** (0.001)	-0.051*** (0.006)
Constant	-46.266*** (11.174)	0.092 (4.633)	-0.020 (3.927)	-4.532* (2.722)	-3.713*** (1.110)	-7.662* (4.010)	-62.123*** (20.347)
Observations	140	140	140	140	140	140	140
R ²	0.661	0.246	0.587	0.488	0.522	0.397	0.668
Adjusted R ²	0.651	0.223	0.574	0.472	0.508	0.379	0.658
Residual Std. Error (df = 135)	58.116	24.097	20.426	14.155	5.773	20.856	105.825
F Statistic (df = 4; 135)	65.698***	10.998***	47.871***	32.119***	36.897***	22.242***	68.002***

Note:

*p<0.1; **p<0.05; ***p<0.01

ing: As all of our four independent variables, the number of young males, the number of people with less than a high school degree, the number of people belonging to the low-income class, and the number of immigrants, are considered as driving forces of crime in general, we should only expect positive estimated parameters.

This is the case for three of our four independent variables: Except of two parameters, *male.youth*, *less.than.high.school*, and *low.income* all have positive estimates throughout the different crime types. Surprisingly, the variable *immigrants* has statistically significant negative estimates for all regressions. Apparently, immigrants do not increase crime rates as one might expect in the case of the city of Toronto.

Furthermore, some crime types seem to be better described by our regressions than others: While *assault*, *break.and.enter*, *robbery*, *theft.over*, *drug.arrests* and *total.crime* provide statistically significant parameters a 5% significance level in almost all cases, the parameters for *auto.theft* are not statistically significant in most cases. These differences are also reflected by the (Adjusted) R Squared of each regression.

4.3 Validity of the Regression Results

Surely, the results of the regressions are only valid if the OLS-assumptions are met. Therefore, we have implemented assumption checks inside the for loop, as described in the former subchapters. The following tables show their results:

Table 5: OLS-assumptions for original dependent variables

	means	bptests	swtests	vif1	vif2	vif3	vif4	cortest1	cortest2	cortest3	cortest4
assault	0.00	0.62	0.22	10.10	2.63	26.09	17.64	1.00	1.00	1.00	1.00
auto.theft	-0.00	0.35	0.92	8.91	2.32	24.75	14.56	1.00	1.00	1.00	1.00
break.and.enter	-0.00	0.89	0.83	8.48	2.65	28.38	17.90	1.00	1.00	1.00	1.00
robbery	-0.00	0.10	0.69	9.11	2.42	27.72	16.13	1.00	1.00	1.00	1.00
theft.over	-0.00	0.62	0.03	11.62	3.02	28.86	18.88	1.00	1.00	1.00	1.00
drug.arrests	0.00	0.49	0.59	10.67	2.67	27.04	18.85	1.00	1.00	1.00	1.00
total.crime	-0.00	0.22	0.55	10.25	2.59	25.73	17.30	1.00	1.00	1.00	1.00

Table 6: OLS-assumptions for transformed dependent variables

	means	bptests	swtests	vif1	vif2	vif3	vif4	cortest1	cortest2	cortest3	cortest4
assault	0.00	0.00	0.00	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
auto.theft	-0.00	0.01	0.00	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
break.and.enter	0.00	0.00	0.00	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
robbery	-0.00	0.01	0.00	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
theft.over	-0.00	0.00	0.00	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
drug.arrests	0.00	0.01	0.00	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
total.crime	0.00	0.00	0.00	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00

As an intercept was included in all cases, the population mean of the error term is, as expected, equal to zero for all regressions. Regarding the Breusch-Pagan test, the H0 of the error term being homoscedastic cannot be rejected at a level of significance of 5% in any case when using transformed independent variables. Without transformation, however, the H0 is rejected for all regressions. As the Breusch-Pagan-test is sensitive to the assumption of normal distribution and to test the optional OLS-assumption of normally distributed errors, we implemented the Shapiro-Wilk-test. While again the H0 of the error term being normally distributed is rejected in every case for the non-transformed data, it cannot be rejected in the case of transformed data at a level

of significance of 5%, except of one case.

Assessing the variance inflation factors which check for imperfect multicollinearity among the independent variables, most of them have a value higher than 10. As mentioned in the former subchapters, this number is a rough guideline to assess the severance of multicollinearity. The assumption of no multicollinearity therefore appears to not be met in this case.

Regarding the correlation test of each independent variable with the error term, the H_0 states that the correlation is equal to zero. As we only obtain p-values equal to 1, we cannot reject the H_0 for any regression, indicating that the OLS-assumptions are met in these cases.

When looking at the Ceres-plots, we can conclude that the assumption of a linear dependence between the regressors and the dependent variable is valid in all cases. While it is not reasonable to include all Ceres-plots in the paper, we present one plot from the *break.and.enter*-regression with transformed independent variable as an example:

While the variables *less.than.high.school*, *low.income*, and *immigrants* clearly show a linear dependence, there seems to be a slight deviation for *male.youth* in some areas. However, the assumption of a linear dependence still seems to be reasonable.

All in all, the OLS-assumptions seem to be met in the case of using transformed data. When looking at the Breusch-Pagan test and the Shapiro-Wilk test, we see a qualitative improvement when using transformed data instead of the original data. Although it does not infer with the necessary OLS-assumptions, the existence of imperfect multicollinearity might decrease the quality of regression results as it might lead to unstable parameter estimates and paradoxical results as the overall p-value might be low while all individual p-values being high. However, when we have a look at Table 1 and Table 2 from the previous chapter, these problems do not seem to occur for our regressions. We can therefore conclude that the regressions when using transformed data fulfill the standard OLS-assumptions.

Although this is very crucial for the validity of regression results in general, to transform data in that way might not necessarily be the best way to get valid results. To keep the interpretation of parameters and regression results more simple and to keep in mind the specific structure of the data, further regression methods are applied in the next section.

5 Further Regression Approaches

The methods applied in this section differ in the way they deal with characteristics of the data. To overcome the problem of potentially non-linearity, the first method is a linear regression on the log-transformed independent variable. In the second application, the focus is on spatial dependence. Since our subject of analysis are neighbourhoods in Toronto, there exists the possibility that the neighbourhoods are somewhat correlated just by their sheer location. Last, we construct a Poisson regression. In this case, the assumed distribution of the error terms is a Poisson distribution instead of a normal distribution. This variation of the regression is applied to deal with the specific structure of the independent variables, which is count data of crimes.

5.1 log-Transformed Data

As indicated above, the first variation of a linear regression model is the log-transformation of the independent variable. As for the example of *assault* as the independent variable, the model to be estimated is

```
1 logmodel <- lm(log(assault) ~ male.youth + less.than.high.school + low
  .income + immigrants, data=r)
```

where the regression is held very simple by means of OLS. As in the case for the untransformed data, the interpretation of the coefficient is very straightforward. Table X shows that the coefficients are - as an example to simplify the comparison - slightly different but similar to the OLS-coefficients with Basic-Power transformation (divided by 100, because of the interpretation of a log-lin-model). However, the coefficients differ in a huge way and even in signs for some variables.

5.2 Spatial Regression

For the spatial regression, which is - as part of the broad field of spatial econometrics - the consideration of spatial interdependence and asymmetries ([Anselin \(2002\)](#)). The most important examples of applications for this paper are spatial regressions for measuring spill-over effects (e.g. [Durlauf \(1994\)](#) or [Glaeser et al. \(1996\)](#)) for different neighbourhoods. The idea of the spatial regression is to conclude a weighting matrix in the regression function that contains the spatial dependences of the independent variable. In this fashion, this dependences are no longer in the error term. The choice of the weighting matrix itself is therefore crucial and can have several forms. One opportunity is to calculate the distance to the spatial center (so in our case the city center). However, we chose to construct a weighting matrix that indicates if a neighbourhood is neighbouring another one, where the row sum is equal to one. Listing X shows how the weighting matrix W is defined in the above sense and how the regression function is stated.

```
1 ####define the weighting matrix
2 shp <- readOGR("Shapefiles/Neighbourhoods_Toronto", "NEIGHBORHOODS_WGS84")
3 neigh <- poly2nb(shp, queen = TRUE)
4 W <- nb2listw(neigh, style="W", zero.policy=TRUE)
5
6 ####spatial regression function
7 spamodel <- lagsarlm(assault~male.youth+less.than.high.school
8                   +low.income+immigrants, data=r, W)
```

As can be seen in Table X, the regression coefficients are similar to those of the standard OLS model, although there are small differences. As can be found in the results in (QUANTLET), ρ , the coefficient of the weighting matrix, is not significantly different from zero on every reasonable level of significance since its p-value is around 0.5. This states that the effect of the weighting matrix and therefore the spatial regression is insignificant. Applying Moran's I-Test to measure spatial correlation on the data and the weighting matrix([Getis und Ord \(1992\)](#)), being not able to reject the null hypothesis of random data on every reasonable level of significance strengthens this statement. As a further confirmation, the coefficients of the spatial regression outputs shown in Tables XX - XX show the small differences between the standard OLS-approach and the Spatial regression.

5.3 Generalized Linear Model with Poisson distribution

A Poisson distribution instead of a normal distribution as an assumption for the error terms (and thus the dependent variables) seems reasonable if one takes a look at the plots of the histograms of the untransformed variables in Figure X. Since count data - and thus data that underlies a Poisson distribution - have their own characteristics and meanings, there are suggestions to not transform them to meet the OLS assumptions but to instead apply a Generalized Linear Model (in the following: *GLM*) (e.g. O'hara und Kotze (2010)). The various difficulties and additional challenges that are accompanied by applying GLM can of course not be discussed in this paper. However, one should keep in mind that important difficulties concerning very basic GLM regressions should be considered (as it will be in the validity part). Still, we set a model in the following form:

```
1 pomodel <- glm(assault ~ male.youth+less.than.high.school+low.income+immigrants,
2               family = "poisson", data = r)
```

Where the *glm* function estimates a model as shown in equation (1) and (2) (Osgood (2000)).

$$\log(\lambda_i) = \sum_{k=0}^K \beta_k x_{ik} \quad (2)$$

$$P(Y_t = y_t) = \frac{e^{-\lambda_t} \lambda_t^{y_t}}{y_t!} \quad (3)$$

Where (1) is a regression equation where the logarithm of the mean is equal to the sum of the product of the dependent variables and the regression coefficients. Equation (2) matches a Poisson-distribution to the probability of y_i (Osgood (2000)). The regression output can be found in Table X. As can be seen directly, the coefficients differ from those of the previous regressions. Also, the interpretation is somewhat different. While in the previous models we were able to interpret the coefficient in a simple way, one has to take the transformation of the independent variable as in (1) into account.

Whether or not the here proposed models outperform the standard OLS-models in the section above is strongly dependent on the degree to which the assumptions of the corresponding model are met.

5.4 Validity

For the log-model, the validation is very similar to that for the basic-power transformed and the original data. Table X shows the major results, where one can see that the OLS-assumptions are somewhat fulfilled as in the first place with $XXX = XXX$. A plot of the residuals can be found in the appendix and indicates an appropriate fit as for the Basic-Power-transformed variables.

Table 7: OLS-assumptions of log-transformed data

	means	bptests	swtests	vif1	vif2	vif3	vif4	cortest1	cortest2	cortest3	cortest4
assault	-0.00	0.55	0.21	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
auto.theft	0.00	0.16	0.71	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
break.and.enter	0.00	0.70	0.43	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
robbery	0.00	0.07	0.10	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
theft.over	-0.00	0.28	0.36	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
drug.arrests	-0.00	0.41	0.13	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00
total.crime	0.00	0.79	0.93	9.13	2.38	26.22	15.59	1.00	1.00	1.00	1.00

For the remaining two models, it is not possible to determine results as in the previous cases. That is, e.g., because the Breusch-Pagan-Test is not applicable for models of the type *lagsarm*. Instead, we decided to check for model-specific assumptions and if they are met. A first step for the Spatial Regression has taken place in the corresponding subsection, where the ρ -Parameter was found to be insignificant. As also indicated above, the necessary conditions of spatially correlated variables is not fulfilled by means of Moran's I-Test. Although there are no indicators in favor of a Spatial Regression model, an additional step to confirm that statement would be to calculate different types of weighting matrices (as e.g. in Wakefield (2006) a matrix that adjusts for differences of neighbourhoods to a centroid) to check if there is no spatial correlation.

The validity of the Poisson Regression is the one that is the hardest to check out of the presented ones. Since - as in the case of Spatial Regression - the Breusch-Pagan-test is not applicable just like other OLS-specification tests, we would have to apply other validity parameters. Since a detailed discussion would go beyond the scope of this paper, we would like to identify the main problems associated with a Poisson Regression. One of the biggest issues is that the mean of a Poisson distribution is equal to its variance. In the case that the variance is bigger than the mean, so-called overdispersion is the problem (Berk und MacDonald (2008)). Although there exist tests for overdispersion (e.g. as proposed by Cameron und Trivedi (1990)), we do not apply such tests for the reason stated above. As can be found by the regression results in the code, an indicator that overdispersion is a problem in our model is, since the ratio of the residual deviance and the degrees of freedom is larger than one (Cameron und Trivedi (1990)). For example, the ratio for the *assault* variable is $\frac{3175}{139}$, which is way larger than one. It is therefore not to decide finally if and how good the model fit of the Poisson regression is although the distributions look a lot like a Poisson distribution.

We have now accessed the performance and the validity of different models. The next and last section will conclude the main results and shows limitations of our methods.

6 Conclusion

This paper tried to seek for neighbourhoods' characteristics that have an effect on the crime rate. For the example of Toronto neighbourhoods, we merged two open datasets that included neighbourhood characteristics as well as crime occurrences. By visualizing the merged data in several ways, we were able to get an important impression of the different neighbourhoods and their corresponding crime statistics. We then applied Ordinary Least squares on the Basic-Power-transformed and the log-transformed dependent variables for all crime types that were available in the datasets. To deal with the data, that was characterized as both count data and spatial data, we included a spatial regression and a generalized linear model with a Poisson-specification. The results of those models differ in both their ability to catch the dependences and the violation of their underlying assumptions. Although this is the case, we could still get an impression on how specific neighbourhood characteristics influence the number of crimes - which we could measure for six different crime types. Unfortunately, we were not able to confirm our findings by an instrumental variable approach and suitable control variables since that would have gone beyond the scope of this paper. Furthermore, the differences in both the coefficient values in most of the regressions do not allow us to make final statements on the impacts of specific characteristics. However, our paper can serve as an overview of how one can analyze crime occurrences in a specific location and what the major steps are. Further research would thus have to focus on the right specification of

the models as has been proposed in other papers.

References

- Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural economics* 27(3), 247–267.
- Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime*, pp. 13–68. Springer.
- Berk, R. und J. M. MacDonald (2008). Overdispersion and poisson regression. *Journal of Quantitative Criminology* 24(3), 269–284.
- Cameron, A. C. und P. K. Trivedi (1990). Regression-based tests for overdispersion in the poisson model. *Journal of econometrics* 46(3), 347–364.
- Corman, H. und H. N. Mocan (2000). A time-series analysis of crime, deterrence, and drug abuse in new york city. *American Economic Review* 90(3), 584–604.
- Cornwell, C. und W. N. Trumbull (1994). Estimating the economic model of crime with panel data. *The Review of economics and Statistics*, 360–366.
- Durlauf, S. N. (1994). Spillovers, stratification, and inequality. *European Economic Review* 38(3-4), 836–845.
- Fajnzylber, P., D. Lederman, und N. Loayza (2002). What causes violent crime? *European economic review* 46(7), 1323–1357.
- Getis, A. und J. K. Ord (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis* 24(3), 189–206.
- Glaeser, E. und J. V. Henderson (2017). Urban economics for the developing world: An introduction. *Journal of Urban Economics* 98, 1–5.
- Glaeser, E. L., B. Sacerdote, und J. A. Scheinkman (1996). Crime and social interactions. *The Quarterly Journal of Economics* 111(2), 507–548.
- Guerette, R. T. (2009). *Analyzing crime displacement and diffusion*. US Department of Justice, Office of Community Oriented Policing Services Washington, DC.
- Hall, A. (2007). Socio-economic theories of crime. Technical report, Working Paper, Capella University.
- Kelly, M. (2000). Inequality and crime. *Review of economics and Statistics* 82(4), 530–539.
- Lochner, L. und E. Moretti (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American economic review* 94(1), 155–189.
- Malik, A. A. (2016). Urbanization and crime: A relational analysis. *J. HUMAN. & Soc. Sci.* 21, 68–69.
- Osgood, D. W. (2000). Poisson-based regression analysis of aggregate crime rates. *Journal of quantitative criminology* 16(1), 21–43.
- O’hara, R. B. und D. J. Kotze (2010). Do not log-transform count data. *Methods in Ecology and Evolution* 1(2), 118–122.

- United Nations (2018). 68% of the world population projected to live in urban areas by 2050, says un. <https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html>. Accessed: 2018/08/06.
- Wakefield, J. (2006). Disease mapping and spatial regression with count data. *Biostatistics* 8(2), 158–183.
- World Bank (2018). Urban population % of total. <https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS>. Accessed: 2018/08/06.
- Wright, R., F. Brookman, und T. Bennett (2005). The foreground dynamics of street robbery in britain. *British Journal of Criminology* 46(1), 1–15.

A Appendix

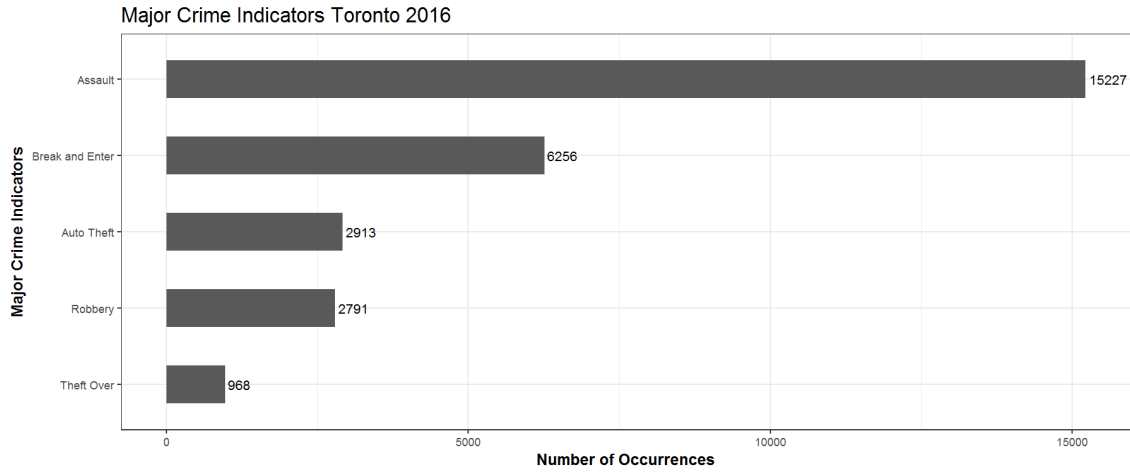


Figure 4: Major Crime Indicators Toronto. [Readme.md](#)

Table 8: Results of log-transformed data

	<i>Dependent variable:</i>						
	assault	auto.theft	break.and.enter	robbery	theft.over	drug.arrests	total.crime
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
male.youth	0.0002 (0.0002)	0.0003 (0.0002)	0.0003* (0.0002)	0.001** (0.0002)	0.0004 (0.0003)	0.0004 (0.0003)	0.0004*** (0.0001)
less.than.high.school	0.0001 (0.0001)	0.0002** (0.0001)	-0.0002*** (0.0001)	0.0003*** (0.0001)	-0.0002** (0.0001)	0.0002** (0.0001)	0.00005 (0.0001)
low.income	0.0003*** (0.00004)	0.0001 (0.0001)	0.0002*** (0.00004)	0.0002*** (0.0001)	0.0003*** (0.0001)	0.0003*** (0.0001)	0.0002*** (0.00003)
immigrants	-0.0002*** (0.00003)	-0.00003 (0.00004)	-0.0001*** (0.00003)	-0.0002*** (0.00004)	-0.0002*** (0.00004)	-0.0003*** (0.00004)	-0.0002*** (0.00002)
Constant	3.104*** (0.099)	1.773*** (0.129)	2.698*** (0.091)	1.450*** (0.134)	0.250* (0.144)	1.399*** (0.154)	4.072*** (0.077)
Observations	140	140	140	140	140	140	140
R ²	0.606	0.377	0.494	0.459	0.411	0.384	0.656
Adjusted R ²	0.594	0.358	0.479	0.443	0.394	0.366	0.646
Residual Std. Error	0.516	0.670	0.474	0.697	0.748	0.801	0.401
F Statistic	51.806***	20.396***	32.976***	28.650***	23.560***	21.056***	64.462***

Note:

*p<0.1; **p<0.05; ***p<0.01

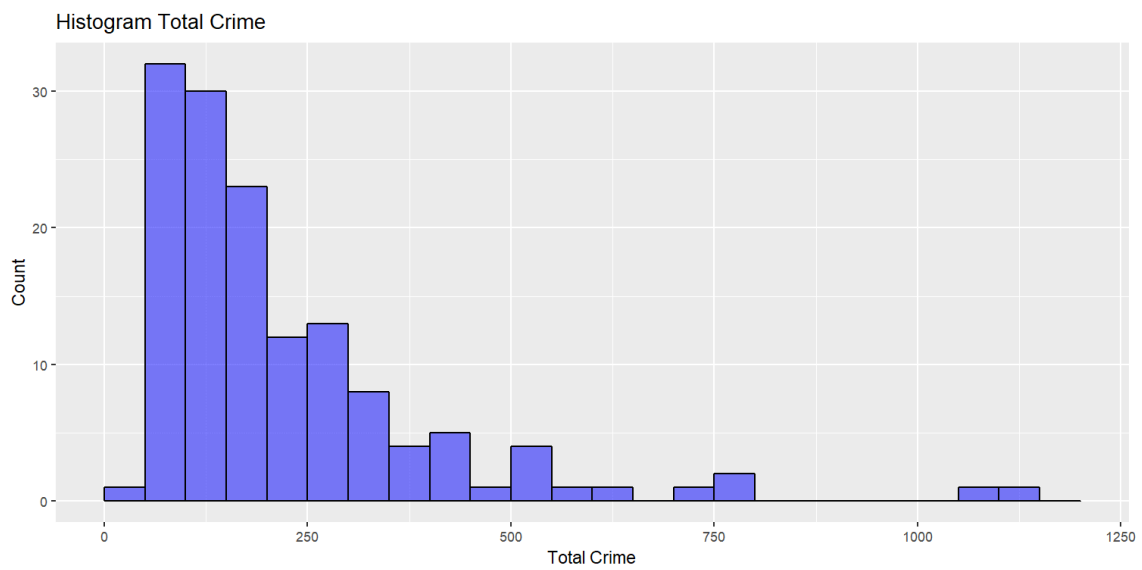


Figure 5: Histogram of total crime Toronto. [Readme.md](#)

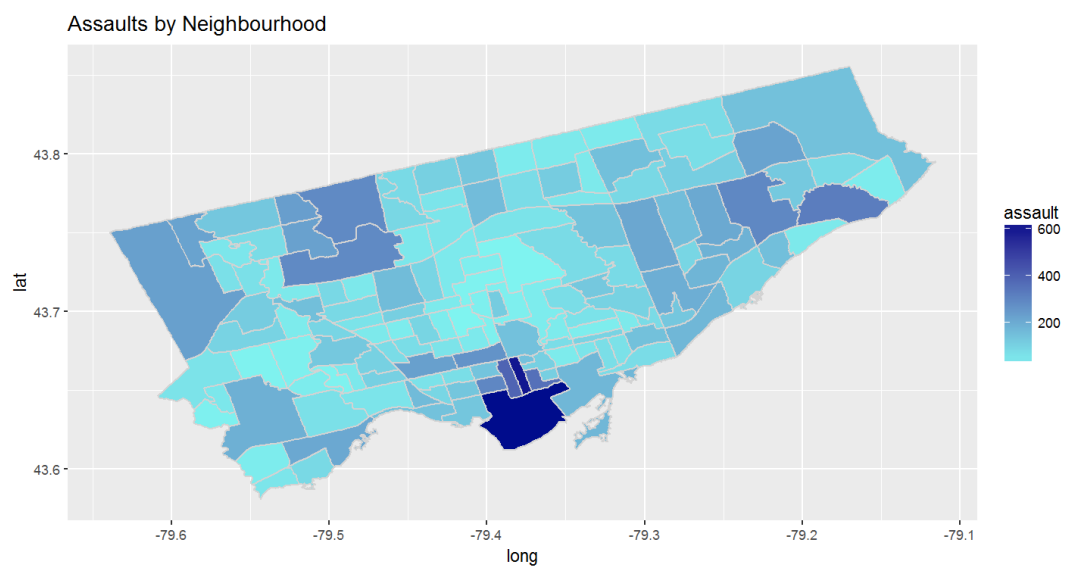


Figure 6: Heatmap of Assaults by Neighbourhood. [Readme.md](#)

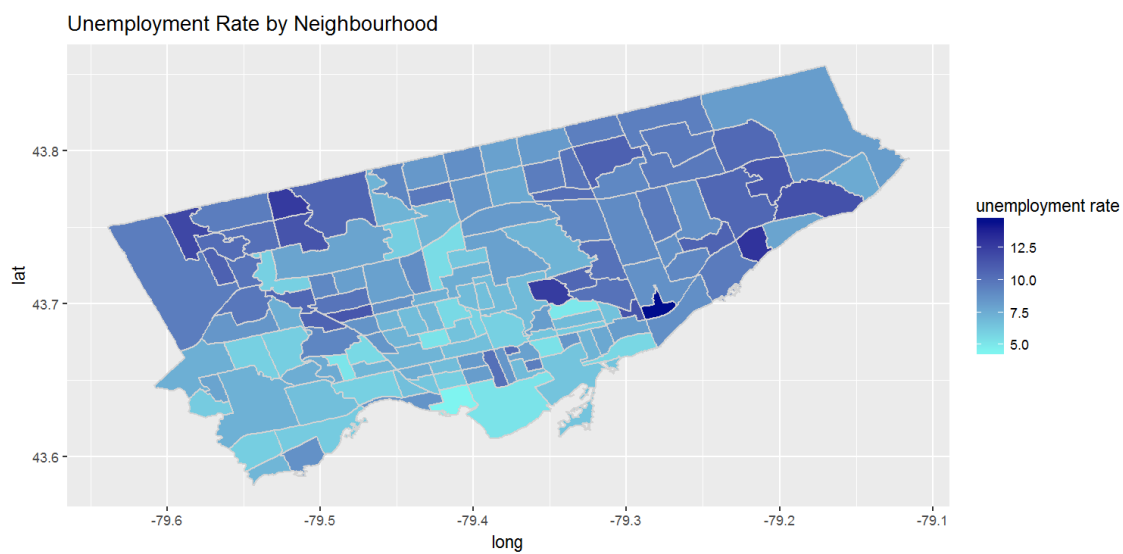


Figure 7: Heatmap of Unemployment Rate by Neighbourhood.

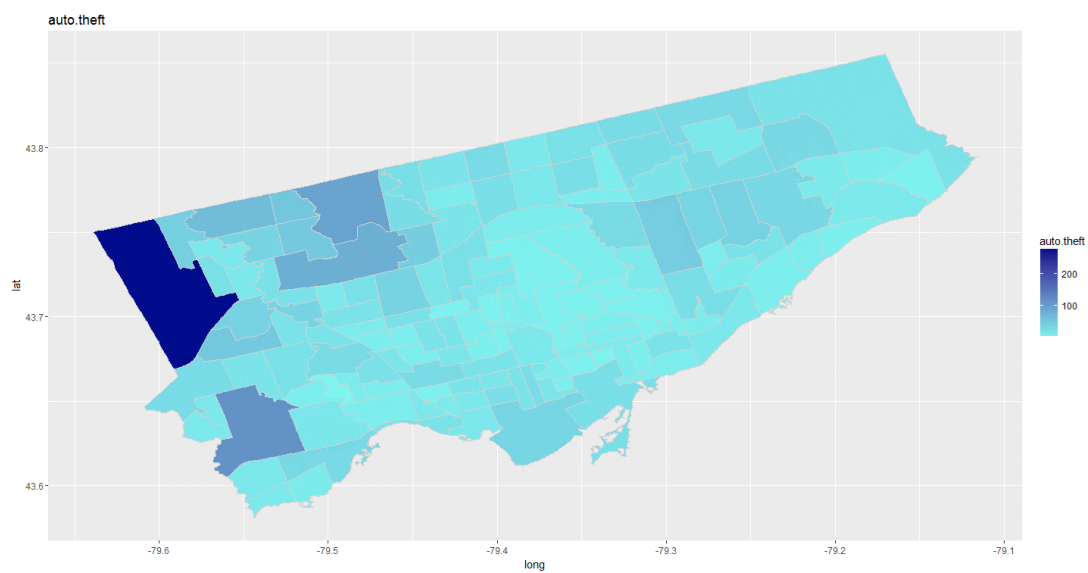


Figure 8: Heatmap of Auto Theft by Neighbourhood.

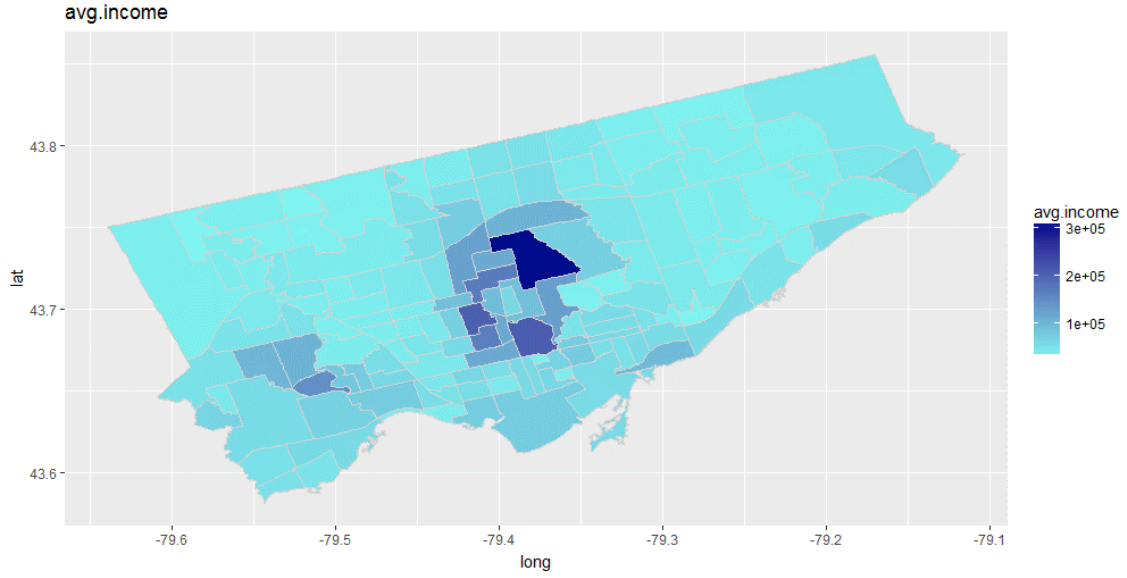


Figure 9: Heatmap of Average Income by Neighbourhood.

Table 9: Results of Spatial Regression

	<i>Dependent variable:</i>						
	assault	auto.theft	break.and.enter	robbery	theft.over	drug.arrests	total.crime
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
male.youth	0.097*** (0.020)	0.023*** (0.008)	0.024*** (0.007)	0.022*** (0.005)	0.009*** (0.002)	0.024*** (0.007)	0.198*** (0.037)
less.than.high.school	-0.005 (0.008)	0.007** (0.003)	-0.012*** (0.003)	0.005*** (0.002)	-0.003*** (0.001)	-0.002 (0.003)	-0.009 (0.014)
low.income	0.032*** (0.005)	-0.002 (0.003)	0.011*** (0.002)	0.003** (0.001)	0.002*** (0.0005)	0.008*** (0.002)	0.055*** (0.009)
immigrants	-0.028*** (0.003)	-0.0002 (0.002)	-0.008*** (0.001)	-0.004*** (0.001)	-0.002*** (0.0003)	-0.008*** (0.001)	-0.051*** (0.006)
Constant	-46.594*** (10.286)	3.919 (5.765)	1.093 (8.417)	-8.873*** (3.378)	-4.344*** (1.368)	-13.295*** (4.705)	-79.405*** (30.081)
Observations	140	140	140	140	140	140	140
Log Likelihood	-764.848	-640.618	-618.432	-565.134	-440.674	-619.247	-848.466
σ^2	3,256.851	549.182	402.124	186.060	31.676	402.282	10,743.300
Akaike Inf. Crit.	1,543.695	1,295.236	1,250.865	1,144.268	895.348	1,252.493	1,710.933
Wald Test		1.944	0.028	4.163**	0.801	4.425**	0.587
LR Test	0.001	1.957	0.047	3.964**	0.791	4.170**	0.579

Note:

*p<0.1; **p<0.05; ***p<0.01

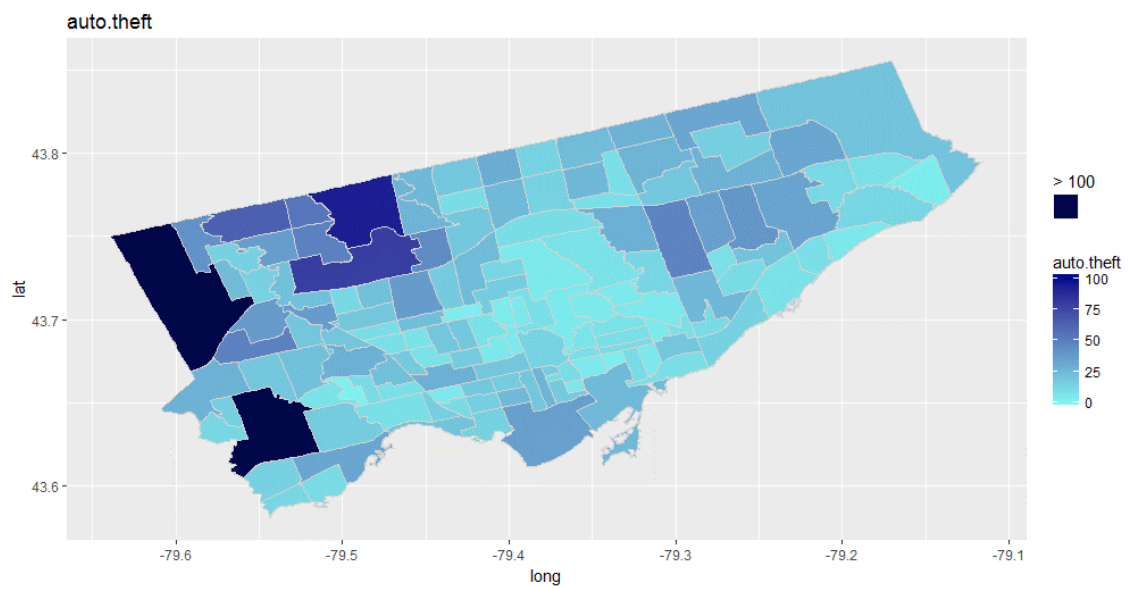


Figure 10: Auto Theft by Neighbourhood adjusted.

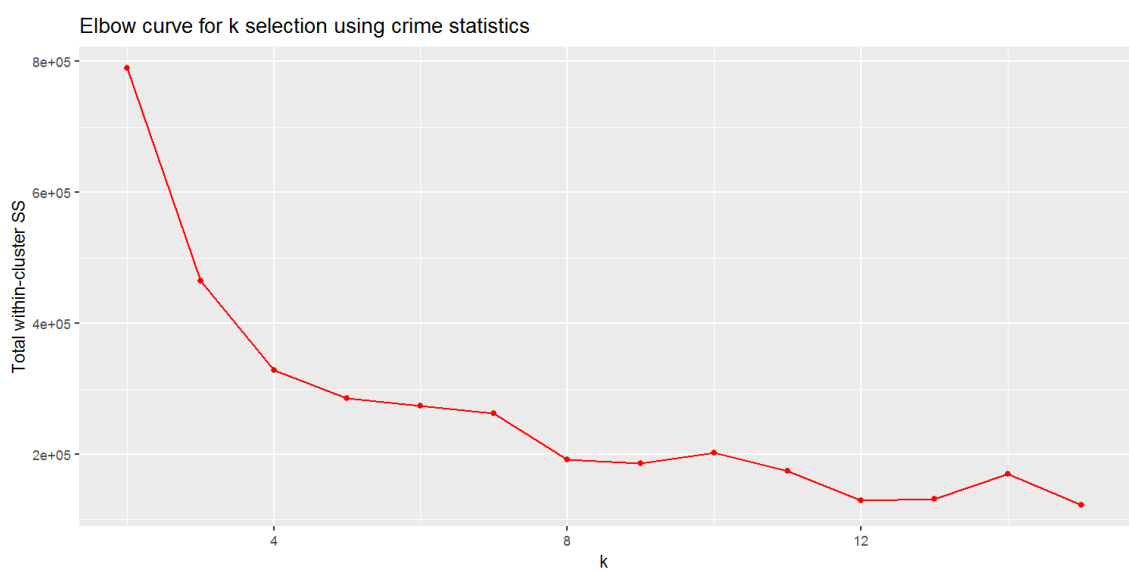


Figure 11: Elbow curve for k selection. [Readme.md](#)

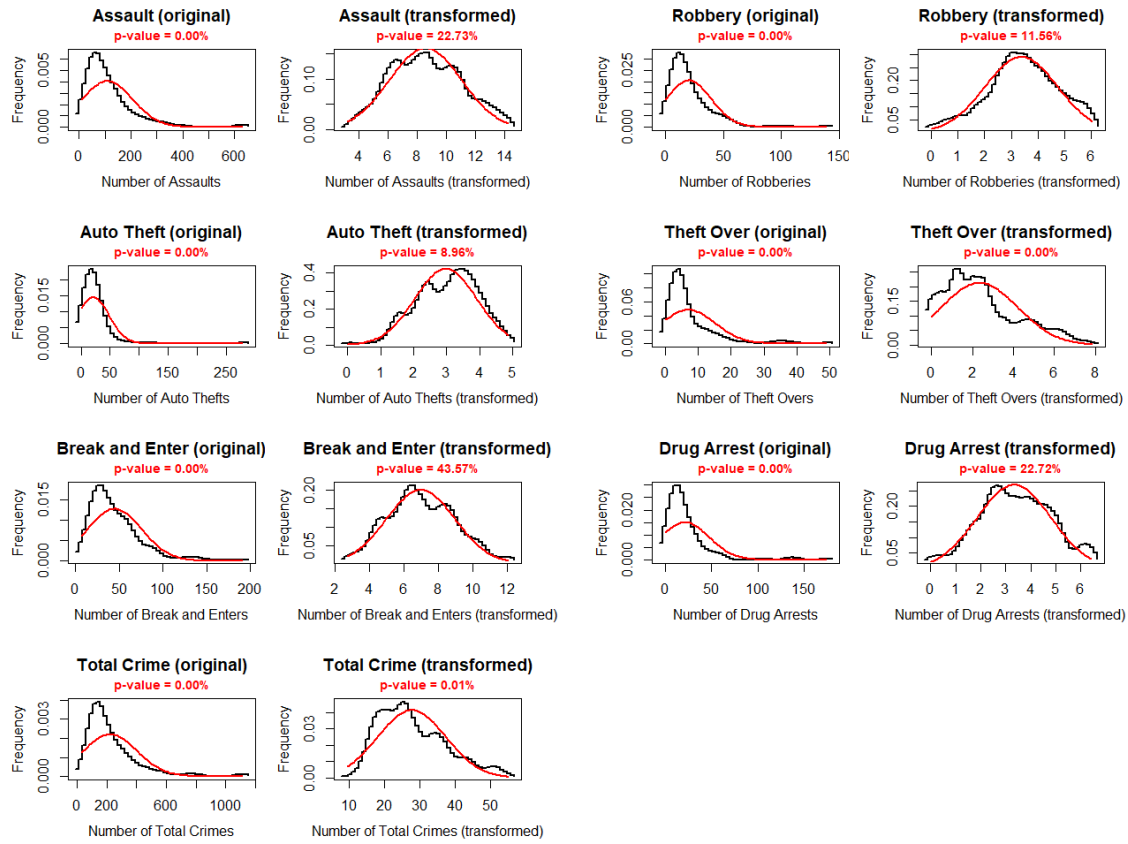


Figure 12: Distribution of original and transformed crime types. [Readme.md](#)

Table 10: Results of Poisson Regression

	<i>Dependent variable:</i>						
	assault	auto.theft	break.and.enter	robbery	theft.over	drug.arrests	total.crime
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
male.youth	0.0004*** (0.00003)	0.001*** (0.0001)	0.0002*** (0.00004)	0.001*** (0.0001)	0.001*** (0.0001)	0.001*** (0.0001)	0.0005*** (0.00002)
less.than.high.school	-0.00001 (0.00001)	0.0003*** (0.00002)	-0.0002*** (0.00002)	0.0002*** (0.00003)	-0.0002*** (0.00004)	0.00004 (0.00003)	-0.00000 (0.00001)
low.income	0.0002*** (0.00001)	-0.00003** (0.00002)	0.0002*** (0.00001)	0.0001*** (0.00002)	0.0002*** (0.00003)	0.0003*** (0.00001)	0.0002*** (0.00000)
immigrants	-0.0002*** (0.00000)	-0.00002 (0.00001)	-0.0001*** (0.00001)	-0.0002*** (0.00001)	-0.0002*** (0.00002)	-0.0003*** (0.00001)	-0.0002*** (0.00000)
Constant	3.502*** (0.019)	2.190*** (0.042)	2.949*** (0.029)	1.953*** (0.043)	0.732*** (0.074)	1.907*** (0.043)	4.330*** (0.013)
Observations	140	140	140	140	140	140	140
Log Likelihood	-2,022.340	-1,077.800	-950.521	-797.781	-453.514	-1,061.389	-3,004.484
Akaike Inf. Crit.	4,054.681	2,165.599	1,911.042	1,605.561	917.027	2,132.778	6,018.967

Note:

*p<0.1; **p<0.05; ***p<0.01

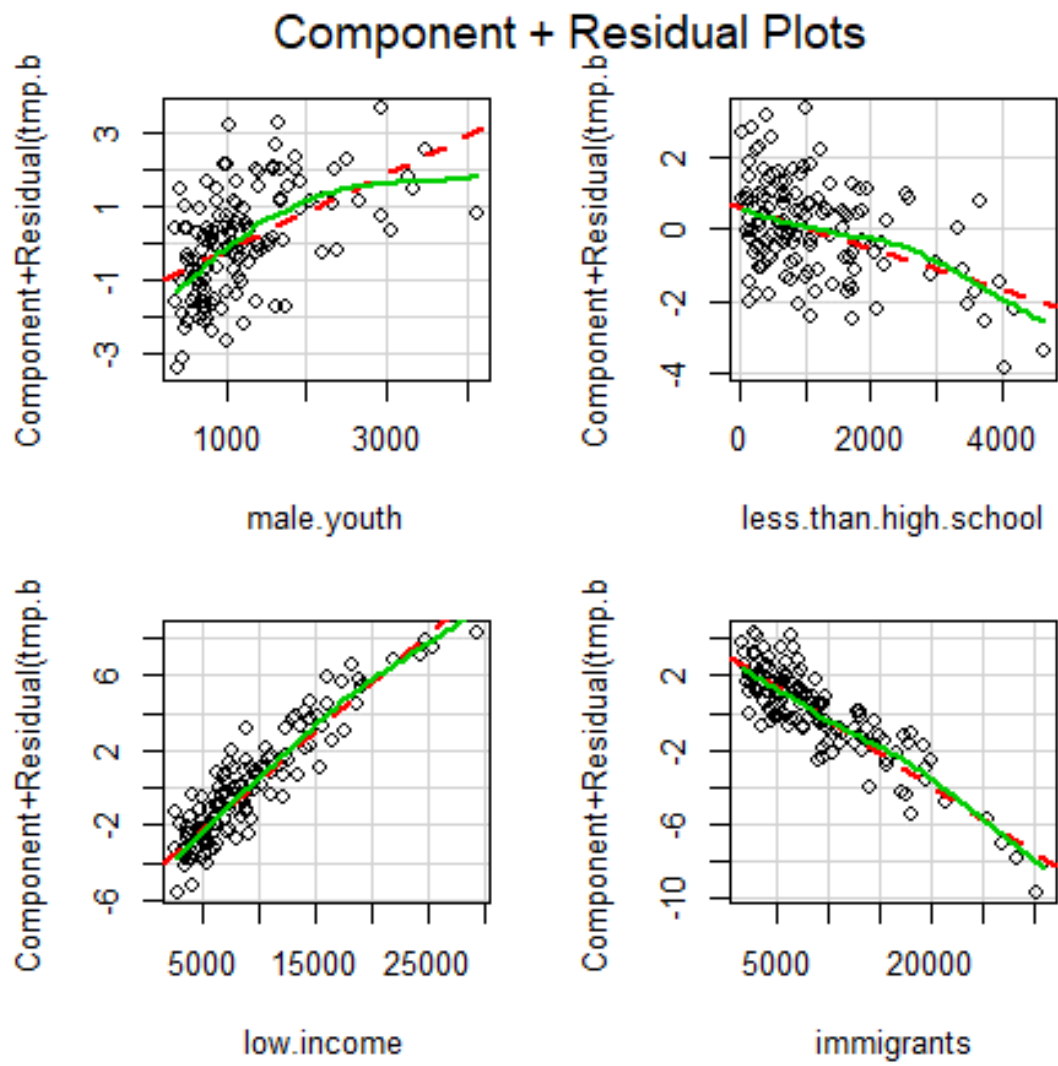


Figure 13: Ceres-plots of the break.and.enter-regression with transformed independent variable

component + residual plots total.crime (log transformed data)

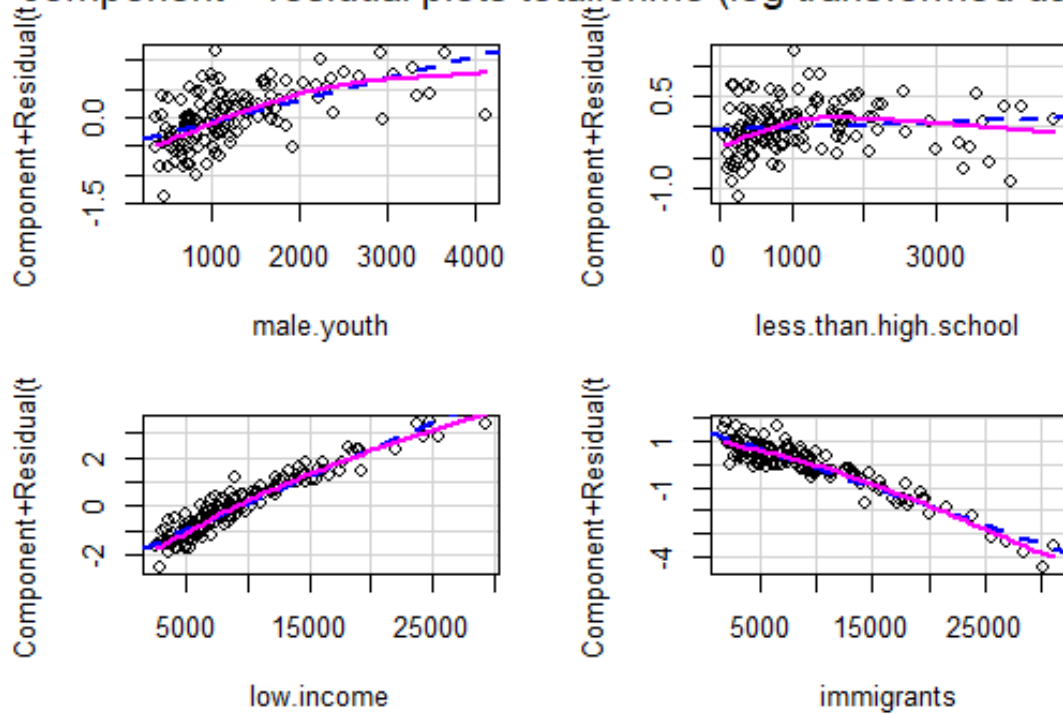


Figure 14: Ceresplot of total.crimes regression with log-transformed variable.

component + residual plots total.crime (poiss reg)

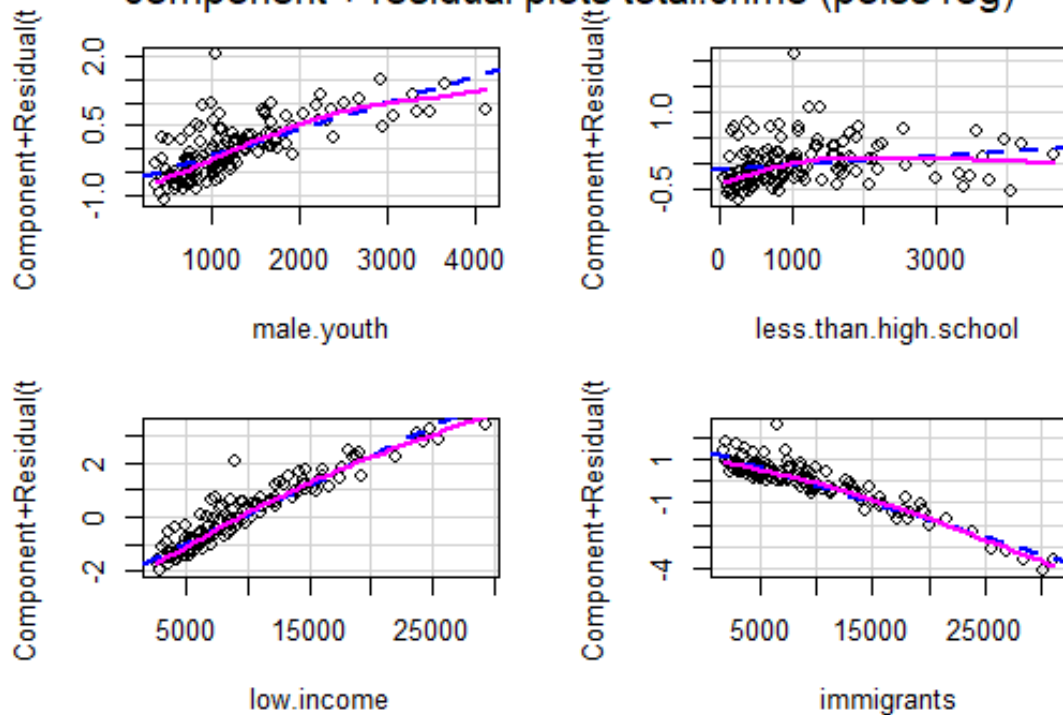


Figure 15: Ceresplot of total.crimes Poisson-regression.

Table 11: Comparison of Different Models for assault

	assault				
	Basic Power Transformation	OLS	Log-OLS	Spatial Regression	Poisson Regression
	(1)	(2)	(3)	(4)	(5)
male.youth	0.001 (0.001)	0.097*** (0.020)	0.0002 (0.0002)	0.097*** (0.020)	0.0004*** (0.00003)
less.than.high.school	0.0004* (0.0002)	-0.005 (0.008)	0.0001 (0.0001)	-0.005 (0.008)	-0.00001 (0.00001)
low.income	0.001*** (0.0001)	0.032*** (0.005)	0.0003*** (0.00004)	0.032*** (0.005)	0.0002*** (0.00001)
immigrants	-0.001*** (0.0001)	-0.028*** (0.003)	-0.0002*** (0.00003)	-0.028*** (0.003)	-0.0002*** (0.00000)
Constant	4.625*** (0.315)	-46.266*** (11.174)	3.104*** (0.099)	-46.594*** (10.286)	3.502*** (0.019)
Observations	136	140	140	140	140
R ²	0.618	0.661	0.606		
Adjusted R ²	0.606	0.651	0.594		
Log Likelihood				-764.848	-2,022.340
σ^2				3,256.851	
Akaike Inf. Crit.				1,543.695	4,054.681
Residual Std. Error	1.545 (df = 131)	58.116 (df = 135)	0.516 (df = 135)		
F Statistic	52.896*** (df = 4; 131)	65.698*** (df = 4; 135)	51.806*** (df = 4; 135)		
LR Test				0.001 (df = 1)	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 12: Comparison of Different Models for auto.theft

	auto.theft				
	Basic Power Transformation	OLS	Log-OLS	Spatial Regression	Poisson Regression
	(1)	(2)	(3)	(4)	(5)
male.youth	0.0004 (0.0004)	0.023*** (0.008)	0.0003 (0.0002)	0.023*** (0.008)	0.001*** (0.0001)
less.than.high.school	0.0004** (0.0002)	0.008** (0.003)	0.0002** (0.0001)	0.007** (0.003)	0.0003*** (0.00002)
low.income	0.0001 (0.0001)	-0.001 (0.002)	0.0001 (0.0001)	-0.002 (0.003)	-0.00003** (0.00002)
immigrants	-0.00005 (0.0001)	-0.0003 (0.001)	-0.00003 (0.00004)	-0.0002 (0.002)	-0.00002 (0.00001)
Constant	2.381*** (0.234)	0.092 (4.633)	1.773*** (0.129)	3.919 (5.765)	2.190*** (0.042)
Observations	135	140	140	140	140
R ²	0.361	0.246	0.377		
Adjusted R ²	0.341	0.223	0.358		
Log Likelihood				-640.618	-1,077.800
σ^2				549.182	
Akaike Inf. Crit.				1,295.236	2,165.599
Residual Std. Error	1.197 (df = 130)	24.097 (df = 135)	0.670 (df = 135)		
F Statistic	18.370*** (df = 4; 130)	10.998*** (df = 4; 135)	20.396*** (df = 4; 135)		
Wald Test				1.944 (df = 1)	
LR Test				1.957 (df = 1)	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 13: Comparison of Different Models for break.and.enter

	break.and.enter				
	Basic Power Transformation	OLS	Log-OLS	Spatial Regression	Poisson Regression
	(1)	(2)	(3)	(4)	(5)
male.youth	0.001** (0.0004)	0.024*** (0.007)	0.0003* (0.0002)	0.024*** (0.007)	0.0002*** (0.00004)
less.than.high.school	-0.001*** (0.0002)	-0.012*** (0.003)	-0.0002*** (0.0001)	-0.012*** (0.003)	-0.0002*** (0.00002)
low.income	0.001*** (0.0001)	0.011*** (0.002)	0.0002*** (0.00004)	0.011*** (0.002)	0.0002*** (0.00001)
immigrants	-0.0004*** (0.0001)	-0.008*** (0.001)	-0.0001*** (0.00003)	-0.008*** (0.001)	-0.0001*** (0.00001)
Constant	3.889*** (0.246)	-0.020 (3.927)	2.698*** (0.091)	1.093 (8.417)	2.949*** (0.029)
Observations	136	140	140	140	140
R ²	0.453	0.587	0.494		
Adjusted R ²	0.436	0.574	0.479		
Log Likelihood				-618.432	-950.521
σ^2				402.124	
Akaike Inf. Crit.				1,250.865	1,911.042
Residual Std. Error	1.210 (df = 131)	20.426 (df = 135)	0.474 (df = 135)		
F Statistic	27.096*** (df = 4; 131)	47.871*** (df = 4; 135)	32.976*** (df = 4; 135)		
Wald Test				0.028 (df = 1)	
LR Test				0.047 (df = 1)	

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 14: Comparison of Different Models for robbery

	robbery				
	Basic Power Transformation	OLS	Log-OLS	Spatial Regression	Poisson Regression
	(1)	(2)	(3)	(4)	(5)
male.youth	0.001*** (0.0004)	0.023*** (0.005)	0.001** (0.0002)	0.022*** (0.005)	0.001*** (0.0001)
less.than.high.school	0.001*** (0.0002)	0.005*** (0.002)	0.0003*** (0.0001)	0.005*** (0.002)	0.0002*** (0.00003)
low.income	0.0002** (0.0001)	0.003** (0.001)	0.0002*** (0.0001)	0.003** (0.001)	0.0001*** (0.00002)
immigrants	-0.0003*** (0.0001)	-0.004*** (0.001)	-0.0002*** (0.00004)	-0.004*** (0.001)	-0.0002*** (0.00001)
Constant	1.700*** (0.237)	-4.532* (2.722)	1.450*** (0.134)	-8.873*** (3.378)	1.953*** (0.043)
Observations	136	140	140	140	140
R ²	0.507	0.488	0.459		
Adjusted R ²	0.492	0.472	0.443		
Log Likelihood				-565.134	-797.781
σ^2				186.060	
Akaike Inf. Crit.				1,144.268	1,605.561
Residual Std. Error	1.218 (df = 131)	14.155 (df = 135)	0.697 (df = 135)		
F Statistic	33.627*** (df = 4; 131)	32.119*** (df = 4; 135)	28.650*** (df = 4; 135)		
Wald Test				4.163** (df = 1)	
LR Test				3.964** (df = 1)	

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 15: Comparison of Different Models for theft.over

	theft.over				
	Basic Power Transformation	OLS	Log-OLS	Spatial Regression	Poisson Regression
	(1)	(2)	(3)	(4)	(5)
male.youth	0.0002 (0.0004)	0.009*** (0.002)	0.0004 (0.0003)	0.009*** (0.002)	0.001*** (0.0001)
less.than.high.school	-0.0002 (0.0002)	-0.003*** (0.001)	-0.0002** (0.0001)	-0.003*** (0.001)	-0.0002*** (0.00004)
low.income	0.0004*** (0.0001)	0.002*** (0.0005)	0.0003*** (0.0001)	0.002*** (0.0005)	0.0002*** (0.00003)
immigrants	-0.0002*** (0.0001)	-0.002*** (0.0003)	-0.0002*** (0.00004)	-0.002*** (0.0003)	-0.0002*** (0.00002)
Constant	0.396** (0.200)	-3.713*** (1.110)	0.250* (0.144)	-4.344*** (1.368)	0.732*** (0.074)
Observations	133	140	140	140	140
R ²	0.330	0.522	0.411		
Adjusted R ²	0.309	0.508	0.394		
Log Likelihood				-440.674	-453.514
σ^2				31.676	
Akaike Inf. Crit.				895.348	917.027
Residual Std. Error	1.051 (df = 128)	5.773 (df = 135)	0.748 (df = 135)		
F Statistic	15.760*** (df = 4; 128)	36.897*** (df = 4; 135)	23.560*** (df = 4; 135)		
Wald Test				0.801 (df = 1)	
LR Test				0.791 (df = 1)	

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 16: Comparison of Different Models for drug.arrests

	drug.arrests				
	Basic Power Transformation	OLS	Log-OLS	Spatial Regression	Poisson Regression
	(1)	(2)	(3)	(4)	(5)
male.youth	0.0001 (0.001)	0.026*** (0.007)	0.0004 (0.0003)	0.024*** (0.007)	0.001*** (0.0001)
less.than.high.school	0.001*** (0.0002)	-0.001 (0.003)	0.0002** (0.0001)	-0.002 (0.003)	0.00004 (0.00003)
low.income	0.001*** (0.0001)	0.008*** (0.002)	0.0003*** (0.0001)	0.008*** (0.002)	0.0003*** (0.00001)
immigrants	-0.0004*** (0.0001)	-0.008*** (0.001)	-0.0003*** (0.00004)	-0.008*** (0.001)	-0.0003*** (0.00001)
Constant	1.938*** (0.276)	-7.662* (4.010)	1.399*** (0.154)	-13.295*** (4.705)	1.907*** (0.043)
Observations	135	140	140	140	140
R ²	0.382	0.397	0.384		
Adjusted R ²	0.363	0.379	0.366		
Log Likelihood				-619.247	-1,061.389
σ^2				402.282	
Akaike Inf. Crit.				1,252.493	2,132.778
Residual Std. Error	1.391 (df = 130)	20.856 (df = 135)	0.801 (df = 135)		
F Statistic	20.126*** (df = 4; 130)	22.242*** (df = 4; 135)	21.056*** (df = 4; 135)		
Wald Test				4.425** (df = 1)	
LR Test				4.170** (df = 1)	

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 17: Comparison of Different Models for total.crime

	total.crime				
	Basic Power Transformation	OLS	Log-OLS	Spatial Regression	Poisson Regression
	(1)	(2)	(3)	(4)	(5)
male.youth	0.001** (0.001)	0.202*** (0.037)	0.0004*** (0.0001)	0.198*** (0.037)	0.0005*** (0.00002)
less.than.high.school	0.0003 (0.0002)	-0.008 (0.014)	0.00005 (0.0001)	-0.009 (0.014)	-0.00000 (0.00001)
low.income	0.001*** (0.0001)	0.054*** (0.009)	0.0002*** (0.00003)	0.055*** (0.009)	0.0002*** (0.00000)
immigrants	-0.001*** (0.0001)	-0.051*** (0.006)	-0.0002*** (0.00002)	-0.051*** (0.006)	-0.0002*** (0.00000)
Constant	7.197*** (0.310)	-62.123*** (20.347)	4.072*** (0.077)	-79.405*** (30.081)	4.330*** (0.013)
Observations	135	140	140	140	140
R ²	0.647	0.668	0.656		
Adjusted R ²	0.636	0.658	0.646		
Log Likelihood				-848.466	-3,004.484
σ^2				10,743.300	
Akaike Inf. Crit.				1,710.933	6,018.967
Residual Std. Error	1.509 (df = 130)	105.825 (df = 135)	0.401 (df = 135)		
F Statistic	59.565*** (df = 4; 130)	68.002*** (df = 4; 135)	64.462*** (df = 4; 135)		
Wald Test				0.587 (df = 1)	
LR Test				0.579 (df = 1)	

Note:

*p<0.1; **p<0.05; ***p<0.01