Multiple Linear Regression Analysis: Calgary Community Crime Rate

Data 603: Statistical Modeling with Data

University of Calgary

Group: Jianling Xie, Alan Cheun, Zane Wu

April 9, 2024

Contents

List of Figures

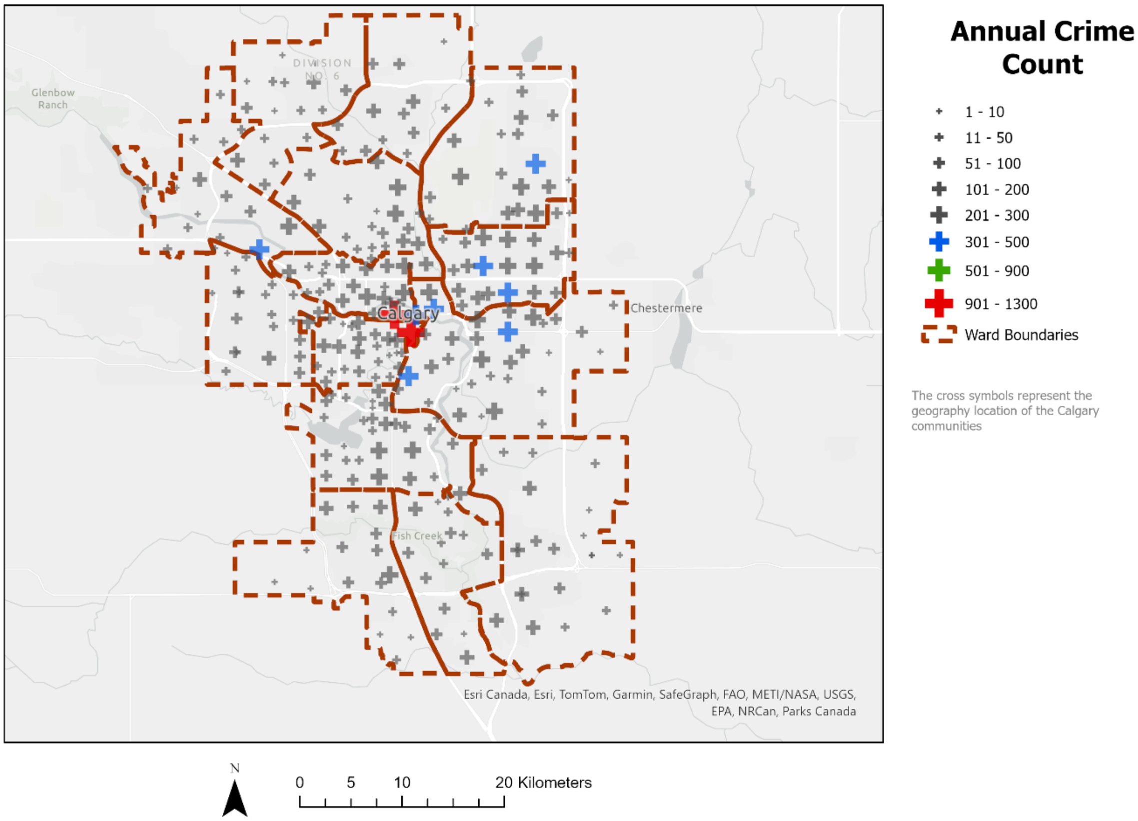
List of Tables

Introduction

Crime has been an important indicator for various aspects of community well-being and safety. It can reflect underlying social issues, economic disparities, and the effectiveness of law enforcement and community programs. Additionally, crime statistics are often used to assess trends over time, allocate resources for crime prevention, and inform policy decisions. Understanding crime patterns can help identify areas that may need additional support or intervention, leading to targeted efforts to improve community safety.

According to the police-reported crime statistics in Canada, as measured by the Crime Severity Index (CSI), crime increased by 5% from 2021 to 2022, reaching 5,668 incidents per 100,000 population. The Violent CSI also rose in 2022, reaching its highest point since 2007 (1).The CSI index for Canada in 2022 was 78.1, with Calgary slightly below the national average at 75.2. However, two Alberta cities ranked high in the CSI, with Edmonton at 100.4 and Lethbridge at 119 (2). Moreover, in a survey conducted in 2023 on perspectives of safety in Calgary, 41% of Calgarians thought crime has increased, and only 33% felt safe riding a C-Train alone after dark, down from 47% in 2022 (3). Studies have shown that, social and economic disadvantage, such as unemployed or employed in low-paying, unskilled jobs were found to be strongly associated with crime (4). There is also an ecological theory postulates that crime will always display an uneven geographical distribution and that this variation is the result of the interrelationship between humans (or groups of humans) and their surroundings (5). Data from the Calgary Open Data Portal (6) also supported this theory, as it shows that the distribution of crime varied greatly across the Calgary comminutes (Figure 1).

Figure 1 2023 Calgary Community Annual Crime Count



Therefore, our project aims to construct a linear regression model to quantify the associations between these spatial and economic factors and the crime rate in the community, and to use this model to predict the crime rates for the communities that will have the future green LRT stations in 2030.

Methodology

Data Source

We obtained the necessary datasets between year 2013 and 2023 for the Calgary communities from City of Calgary's Open Data Portal (6-10).These datasets included the crime statistics, census data, unemployment rates, number of building permits, median property assessment, geospatial center point of the communities, geospatial center point of Transit light-rail transit (LRT) stations, geospatial center point of police service locations. Each dataset was downloaded as CSV files. As the raw crime statistics dataset is large containing 67,262 records, recording nine categories of crimes (Assault-Non-domestic, Break & Enter – Commercial, Break & Enter – Dwelling, Break & Enter – Other Premises, Commercial Robbery, Street Robbery, Theft FROM Vehicle, Theft OF Vehicle, Violence Other-Non-domestic). We then loaded the dataset to MySQL Workbenth and used the Structured query language (SQL) to join the multiple datasets by the unique community identities. The data we needed for our regression analysis included the dependent variable community annual crime rate per 100,000 population as a function of 10 independent variables, including: Year of the crime rate assessment, Calgary Sectors, shortest distance to a LRT station, shortest distance to a police station, percentage of male of the community, percentage of people aged 75 and older of a community, total number of building permit of the community, Calgary (CMA) average hourly wage rate, median property assessment of the community, and percentage of Canada unemployment rate.

In the raw dataset of crime statistics, the crime counts were recorded as per month per community per a type of crime. As our depended variable is the annual crime rate per 100,000 population of the community, we need to do some data wrangling. In MySQL Workbenth, we aggregated the crime count by month by category of crime to create the annual crime count per each community. This process utilize the SQL "SELECT", "WHERE", "SUM COUNT" syntaxes and clauses.

Finally, we created a master dataframe from the previous developed dataframe including all the dependent and independent variables as columns. The master dataframe included crime rate for 209 communities, and for each community the crime rates were measured for 6 years from 2018 to 2023, so there were 1256 rows in our final data frame, with each row representing a community. We then read the dataframe into R using function "read.csv", for analysis.

The datasets from the City of Calgary's Open Data Portal are licensed by Open Government License (15), the information provider, City of Calgary grants us a worldwide, royalty-free, perpetual, non-exclusive license to use the Information, including for commercial purposes, subject to the terms the license (11).

Variable Explanation and Data Assumptions

The data for the Calgary crime statistics data is provided monthly by the Calgary Police Service. This data is considered cumulative as late-reported incidents are often received well after an offence has occurred. Crime count is based on the most serious violation (MSV) per incident. Violence: These figures include all violent crime offences as defined by the Centre for Canadian Justice Statistics Universal Crime Reporting (UCR) rules. Domestic violence is excluded. Break and Enter: Residential B&E includes both House and 'Other' structure break and enters due to the predominantly residential nature of this type of break in (e.g. detached garages, sheds). B&Es incidents include attempts.

We collected the following spatial and economic variables from the open data portal as previous studies have shown that geographic factors such as residential mobility, and economic factors such as neighborhoods characterized by poverty, are attributable to crime(12, 13).

Sectors are the city's geographical division. There are eight sectors in Calgary: centre, east, north, northeast, northwest, south, southeast, west.

Shortest distance between community and LRT was calculated as the shortest distance between the center geometric point of a community and the center geometric point of an LRT station.

Shortest distance between community and a police station was calculated as the shortest distance between the center geometric point of a community and the center geometric point of a police station.

We also collected the following economic variables available from the open data portal:

Total building permit is the total number of building permits issued by the City of Calgary's Planning & Development department for the community in each year.

Community property assessment contains annual median assessment of the community in Canadian dollars.

Calgary average hour wage rate and the Canada unemployment rate are the sets of economic indicators monitored by Corporate Economics.

We also collected the community's demographic variables for our regression analysis, as previous literatures suggested there may be effects of neighborhood structural characteristics on crime (12, 13).

Percentage of male and the percentage of people ages 75 years and older are the census data for the communities.

Year was included as an independent variable to account for potential trend change over time.

The following is a complete list of variables used in our modelling process. All variables were reported annually at a community level (unit shown in parentheses):

1. crime_rate –community crime rate (per 100,000 of the community population) *Dependent variable
2. Year - Year of assessment (Year) *Independent Variable
3. Sectors - City's geographical divisions (factor with 8 categories) *Independent Variable
4. SHORTEST_DISTANCE_TO_LRT_METERS - shortest distance (meters) between the center geometry points of the community and an LRT station*Independent Variable
5. SHORTEST_DISTANCE_TO_POLICE_METERS - shortest distance (meters) between the center geometry points of the community and a police station *Independent Variable
6. male_percentage – percentage of male population in the community (%) *Independent Variable
7. age_75_plus_percentage - percentage of population aged 75 years and older in the community (%) *Independent Variable
8. TotalPermits - total number of building permits issued by the City of Calgary's Planning & Development department *Independent Variable
9. calgary_cma_average_hourly_wage_rate - Calgary average hourly wage (dollar) *Independent Variable
10. property_assessment_median - median community property assessment (dollar) *Independent Variable
11. canada_unemployment_rate - Canada unemployment rate (%) *Independent Variable

One dataset will be used to make prediction regarding projected community crime rate for the communities that will have the Green LRT stations in year 2023. We hypothesized that the geospatial location of LRT stations may be associated with community crime rates. Thus, we obtained the geospatial center points of the future Green LRT stations and calculated to future shortest distance between the Green LRT stations and communities that will have the Green LRT stations. For the communities that have their shortest distance to LRT changed after the implementation of the future Green LRT, we will then use the predict() function in R to predict their community crime rate

Modeling Plan

We will apply the methods we have learned in Data 603 for building our regression model, step by step. We will first examine the distribution of the dependent variable of crime rate, to see if any transformation is needed. The justification for this step is that as if the dependent variable is highly skewed, the normality assumption of linear regression would likely be violated. We then run a linear regression model using all independent variables and test variables for multicollinearity. The justification for this step is that, with extremely large VIF values across predictive variables, running a stepwise regression may eliminate important predictors due to their multicolinearity. Once we have removed the variables with high multicollinearity, we will use step-wise regression to select a model of main effects. We will then perform a partial F-test to compare our full model and the reduced model.

Once we have decided our main effects, we will use the individual t-test to check for significant interactions and higher-order terms. We will then run another F-test to evaluate if the interactions and the higher-order terms are significant. Any significant interactions and higher-order terms will be added to the main effects to produce our final model. We will then test our model for the following six assumption for multiple linear regression model:

1. Linearity Assumption – Review residual plots
2. Independent Assumption – Review residual against year (time), and residual against sectors (spatial variable)
3. Normality Assumption – Review residual normal qq plot and use Shapiro-Wilk normality test
4. Equal Variance Assumption (heteroscedasticity) – Review residual plot and use Breusch-Pagan test
5. Multicollinearity – Using variance inflation factors (VIF)
6. Outliers – check Cook's distance and leverage

If our model does not meet any of these assumptions, we will review our modeling approach to see if any improvement could be made. Once our model satisfies all of these assumptions, we will then use the model to predict future crime rate for the communities that will have the future Green LRT stations.

Statistical analysis is conducted using R.4.3.2.

Workload Distribution

Workload have been equally distributed to each of the team members:

Data collection: Alan Cheun, Zane Wu, Jianling Xie

Data cleaning and wrangling: Alan Cheun, Zane Wu, Jianling Xie

Statistical analysis: Jianling Xie, Alan Cheun, Zane Wu

Critical review and result interpretation: Alan Cheun, Zane Wu, Jianling Xie,

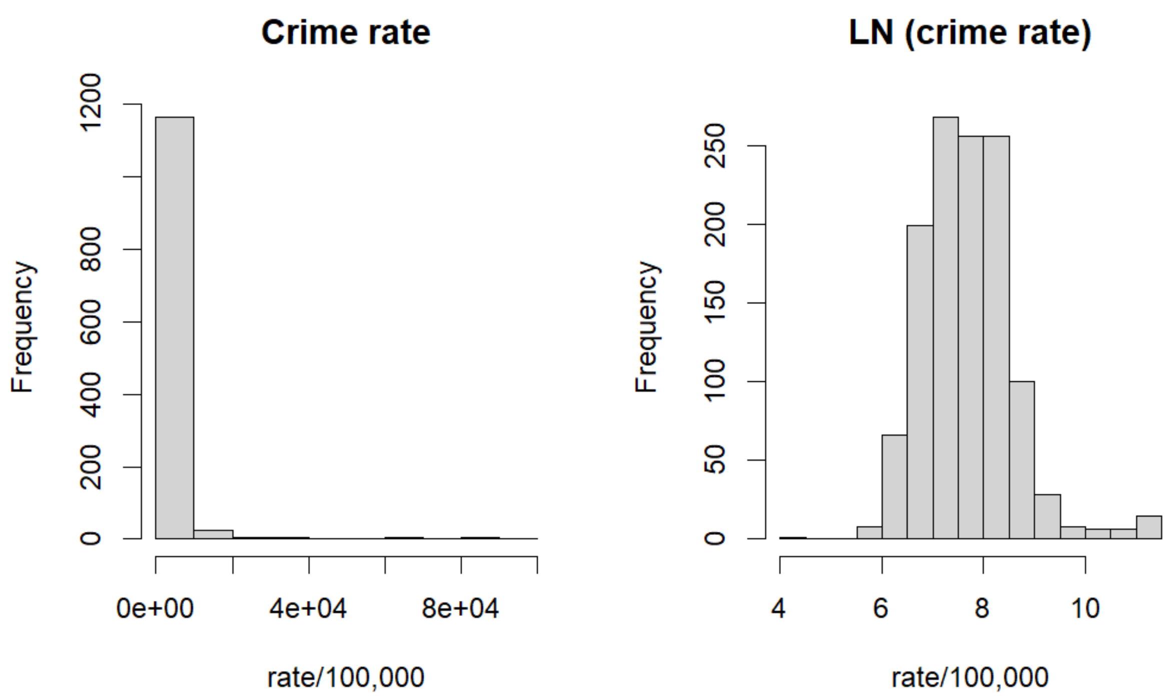
Final Report writing: Jianling Xie, Alan Cheun, Zane Wu

Results

Variable Selection Procedures

We first evaluated the distribution of the dependent variable crime_rate using a histogram and found that the distribution was highly right skewed. Considering that this may violate the normality assumption for the linear regression model, we performed a natural log transformation on crime_rate. After log transformation, the distribution approximate to normal distribution (Figure 2). Will use this variable ln_crime_rate as our dependent variable for our model building.

Figure 2 Histogram of the dependent variable



We then checked for missing data in the dataset, and found that there were 35 rows (3%) of the total of 1256 rows had missing data in either *age_75_plus_percentag* or *TotalPermits*, as the proportion of missingness was small, we filtered out the rows with missing data and performed a complete case analysis for our project.

We then built a first-order model that comprised all candidate independent variables, as shown below. This will be helpful as we continue to select different variables through various selection procedures.

First order Model

$$\widehat{Y_{\log(crime\_rate)}} = \hat{\beta}_0 + \hat{\beta}_1 X_{Year} + \hat{\beta}_2 X_{Sectors}$$

$$+\hat{\beta}_3 X_{SHOREST\_DISTANCE\_TO\_LRT\_METERS}$$

$$+\hat{\beta}_4 X_{SHOREST\_DISTANCE\_TO\_POLICE\_METERS}$$

$$+\hat{\beta}_5 X_{male\_percentage}$$

$$+\hat{\beta}_6 X_{age\_75\_plus\_percentage}$$

$$+\hat{\beta}_7 X_{TotalPermits}$$

$$+\hat{\beta}_8 X_{calgary\_cma\_average\_hourly\_wage\_rate}$$

$$+\hat{\beta}_9 X_{property\_assessmnt\_median}$$

$$+\hat{\beta}_{10} X_{canada\_unemployment\_rate}$$

To begin the variable selection procedure, we first ran the above full model and check for multicollinearity using VIF, and dropped off any variable with high VIF values, rather than start with running the stepwise regression. The justification for this step lies in that we were not sure if there were any potential multicollinearity of our variables. Running a stepwise regression with variables with high VIF values may eliminate important predictors due to their multicollinearity. So, we decided to run several VIF tests and remove predictors with high VIF.

At the first iteration of checking VIF, we found that Year (VIF=38.73) and calgary_cma_average_hourly_wage_rate (VIF=37.40) had high VIF values. We decided to keep year in the model as it helps us explain the potential time trend in crime rates. At the second iteration, the VIF values for all predictors were smaller than 5.0 (Table 2), suggesting there may be moderate collinearity, but it is not severe enough to warrant corrective measures.

Table 1 VIF values as variables are removed sequentially

Independent Variables	VIF iteration 1	VIF iteration 2
Year	38.993	1.0380
SectorsEAST	1.505	1.5052
SectorsNORTH	2.7094	2.7094
SectorsNORTHEAST	1.3904	1.3903
SectorsNORTHWEST	1.7401	1.7401
SectorsSOUTH	1.5008	1.5008
SectorsSOUTHEAST	2.1638	2.1638
SectorsWEST	1.3881	1.3881
SHORTEST_DISTANCE_TO_LRT_METERS	4.1931	4.1930
SHORTEST_DISTANCE_TO_POLICE_METERS	2.3935	2.3934
male_percentage	1.7300	1.7299
age_75_plus_percentage	1.3276	1.3275
TotalPermits	1.2984	1.2974
calgary_cma_average_hourly_wage_rate	37.6036	N/A
property_assessment_median	1.2543	1.2541
canada_unemployment_rate	2.4450	1.0418

Hypothesis Statement for Individual T-tests:

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$
$$i = Year, Sectors, SHORTEST\_DISTANCE\_TO\_LRT\_METERS,$$
$$SHORTEST\_DISTANCE\_TO\_POLICE\_METERS,$$
$$male\_percentage, \quad age\_75\_plus\_percentage, TotalPermits,$$
$$property\_assessment\_median, canada\_unemployment\_rate$$

$$\alpha = 0.05$$

Main Effects Individual T-tests:

$$Year: t = -7.203, p = 1.04^{-12}$$
$$SectorsEast: t = -1.924, p = 0.0546$$
$$SectorsNORTH: t = -6.599, p = 6.19^{-11}$$
$$SectorsNORTHEAST: t = -8.541, p < 2^{-16}$$
$$SectorsNORTHWEST: t = -9.233, p < 2^{-16}$$
$$SectorsSOUTH: t = -15.807, p < 2^{-16}$$
$$SectorsSOUTHEast: t = -3.398, p = 0.00070$$

$SectorsWEST: t = -12.563, p < 2^{-16}$
$SHORTEST\_DISTANCE\_TO\_LRT\_METERS: t = -2.781, p = 0.00551$
$SHORTEST\_DISTANCE\_TO\_POLICE\_METERS: t = -6.389, p = 2.39^{-10}$
$male\_percentage: t = 20.870, p < 2^{-16}$
$age\_75\_plus\_percentage: t = 19.792, p < 2^{-16}$
$TotalPermits: t = 6.850, p = 1.17^{-11}$
$property\_assessment\_median: t = -3.041, p = 0.00241$
$canada\_unemployment\_rate: t = -2.776, p = 0.00558$

Individual T-test were also used in our variable selection to determine the best predictors based on a significance level of $\alpha = 0.05$. From the results of these tests, we would reject the null hypothesis in favor of the alternative. This suggested that year, sectors, shortest distance to an LRT station, shortest distance to a police station, percentage of male population in of the community, percentage of community population aged 75 years and older, total number of building permits issued for the community, median assessment of property, and Canada's unemployment rate are significant predictors for the community crime rate on their own. Therefore, all these variables will be added to our model for further comparison between interaction and higher order terms. Our main effect model is shown below:

$$
\begin{aligned}
Y_{\log\,(\widehat{crime\_rate})} = \hat{\beta}_0 + \hat{\beta}_1 X_{Year} + \hat{\beta}_2 X_{Sectors} \\
+\hat{\beta}_3 X_{SHOREST\_DISTANCE\_TO\_LRT\_METERS} \\
+\hat{\beta}_4 X_{SHOREST\_DISTANCE\_TO\_POLICE\_METERS} \\
+\hat{\beta}_5 X_{male\_percentage} \\
+\hat{\beta}_6 X_{age\_75\_plus\_percentage} \\
+\hat{\beta}_7 X_{TotalPermits} \\
+\hat{\beta}_8 X_{property\_assessmnt\_median} \\
+\hat{\beta}_9 X_{canada\_unemployment\_rate}
\end{aligned}
$$

We looked into the presence of possible two-way interaction effects between our predictive variables, we found there were interactions between the following variables: year and percentage of male population, year and total number of building permits issued, sectors and shortest distance to LRT stations, sector and percentage of male population, sector and percentage of population aged 75 years and older , sectors and total number of building permits issued, sectors and the community's median property assessment value, shortest distance between community and LRT station and shortest distance between community and police station, shortest distance between community and LRT station and percentage of male population, shortest distance between community and LRT station and percentage of population aged 75 years and older, shortest distance between community and LRT station and total number of building permits issued, shortest distance between community and police station and total number of building permits issued, percentage of male population and percentage of population aged 75 years and older, and percentage of male population and the community's median property assessment value. After removing the non-significant of individual interaction terms form the individual t-test (median community property assessment value and Cananda unemployment rate, p=0.127) and re-running, a summary of individual t-test, we are left with the results below:

Hypothesis Statement for Individual T-tests (Interaction Terms):

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$
$$i = all\ possible\ 2\ way\ interactions\ betwen\ these\ varaibles:$$
$$Year, Sectors, SHORTEST\_DISTANCE\_TO\_LRT\_METERS,$$
$$SHORTEST\_DISTANCE\_TO\_POLICE\_METERS,$$
$$male\_percentage,\ \ age\_75\_plus\_percentage, TotalPermits,$$
$$property\_assessment\_median, canada\_unemployment\_rate$$

$$\alpha = 0.05$$

Interaction Term T-tests:

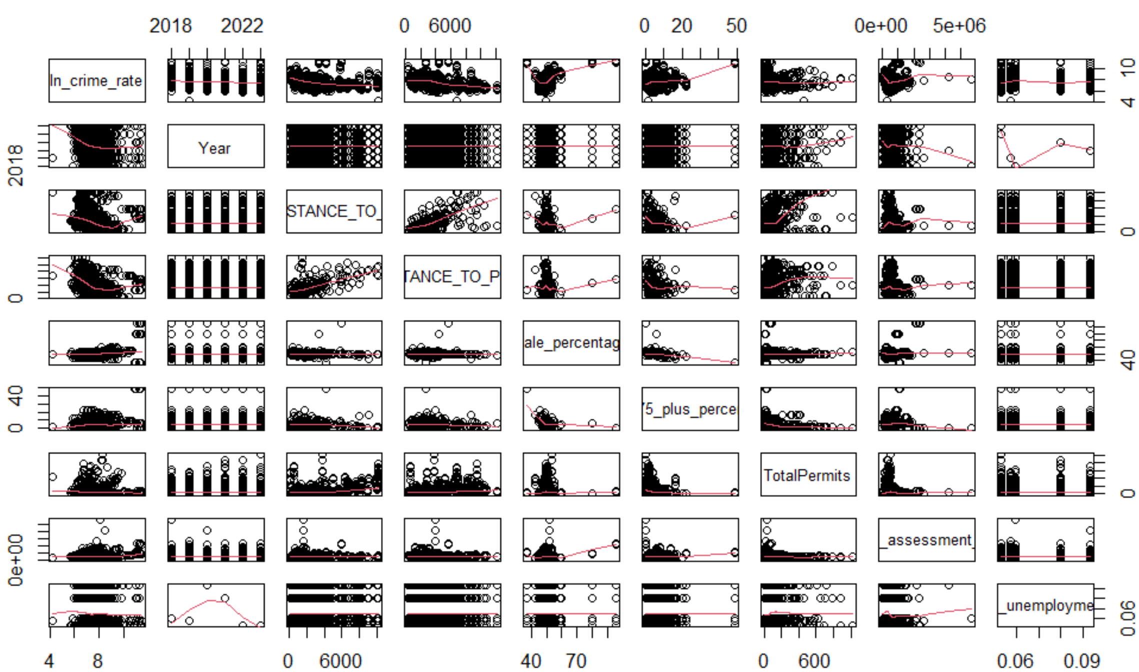
$Year * male\_percentage: t = 2.4747, p = 0.0135$
$Year * TotalPrmits: t = 5.992, p = 2.78^{-9}$
$SectorsEAST * SHORTEST\_DISTANCE\_TO\_LRT\_METERS = -3.078, p = 0.00213$
$SectorsNORTH * SHORTEST\_DISTANCE\_TO\_LRT\_METERS = 2.371, p = 0.0018$
$SectorsSOUTH * SHORTEST\_DISTANCE\_TO\_LRT\_METERS = -2.856, p = 0.0044$
$SectorsSOUTHEAST * SHORTEST\_DISTANCE\_TO\_LRT\_METERS = 2.186, p = 0.029$
$SectorsWEST * SHORTEST\_DISTANCE\_TO\_LRT\_METERS = 4.849, p = 1.41^{-6}$
$SectorsEAST * SHORTEST\_DISTANCE\_TO\_POLIC\_METERS = 2.761, p = 0.0059$
$SectorsSOUTH * SHORTEST\_DISTANCE\_TO\_POLIC\_METERS = 2.040, p = 0.0417$
$SectorsWEST * SHORTEST\_DISTANCE\_TO\_POLIC\_METERS = -3.256, p = 0.0012$
$SectorsNORTH * male\_percentage = -4.110, p = 4.24^{-5}$
$SectorsNORTHEAST * male\_percentage = -2.507, p = 0.0123$
$SectorsNORTHWEST * male\_percentage = 2.202, p = 0.0278$
$SectorsSOUTH * male\_percentage = 2.206, p = 0.028$
$SectorsSOUTH * age\_75\_plus\_percentage = 3.03, p = 0.025$
$SectorsSOUTHEAST * age\_75\_plus\_percentage = 5.884, p = 5.25^{-9}$
$SectorsNORTH * TotalPermits = -3.199, p = 0.0014$
$SectorsNORTHEAST * TotalPermits = -5.707, p = 1.47^{-8}$
$SectorsNORTHWEST * TotalPermits = 4.258, p = 2.23^{-5}$
$SectorsNORTHEAST * TotalPermits = -4.356, p = 1.44^{-5}$
$SectorsNORTHEAST * property\_assessment\_median = -2.617\ p = 0.0090$
$SectorsNORTHWEST * property\_assessment\_median = 6.046\ p = 2.01^{-9}$
$SectorsSOUTH * property\_assessment\_median = 7.922\ p = 5.47^{-15}$
$SectorsSOUTHEAST * property\_assessment\_median = 2.484\ p = 0.0131$

$SHORTEST\_DISTANCE\_TO\_LRT\_METERS * SHORTEST\_DISTANCE\_TO\_POLICE\_METERS = 1.976\, p = 0.048$
$SHORTEST\_DISTANCE\_TO\_LRT\_METERS * male\_percentage = 4.741, p = 2.40^{-6}$
$SHORTEST\_DISTANCE\_TO\_LRT\_METERS * age\_75\_plus\_percentage = 3.77, p = 0.00017$
$SHORTEST\_DISTANCE\_TO\_LRT\_METERS * TotalPermits = 2.879, p = 0.0041$
$SHORTEST\_DISTANCE\_TO\_POLICE\_METERS * property\_assessment\_median = -2.859, p = 0.0043$
$male\_percentage * age\_75\_plus\_percentage = -4.767, p = 2.10^{-6}$
$male\_percentage * property\_assessment\_median = -4.211, p = 2.74^{-5}$

As these interaction terms are significant predictors of community crime rate, they will be added to the regression model. This also makes sense, as research in criminology reveals that certain social characteristics are linked with a greater likelihood of involvement in criminal activity (14).

To check for higher order terms, we used the function pairs() to see how the response looks with respect to each independent variable. We look at all pairwise combinations of continuous variables. It looked like there might be some concavity in the crime rate vs. year, crime rate vs. shortest distance to LRT station, and crime rate vs. shortest distance to police station (Figure 3).

Figure 3 Scatter plots of pairwise combinations of continuous variables



Hypothesis Statement for Individual T-tests (Interaction Terms):

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

$$i = Year^2, SHORTEST\_DISTANCE\_TO\_LRT\_METERS^2,$$
$$SHORTEST\_DISTANCE\_TO\_POLICE\_METERS^2,$$
$$\alpha = 0.05$$

Higher Order Term Individual T-tests:

$$Year^2 : t = -2.114, p = 0.034$$
$$SHORTEST\_DISTANCE\_TO\_LRT\_METERS^2 : t = 4.196, p = 2.92^{-5}$$
$$SHORTEST\_DISTANCE\_TO\_LRT\_METERS^3 : t = -3.448, p = 0.000584$$
$$SHORTEST\_DISTANCE\_TO\_POLICE\_METERS^2 : t = -1.703, p = 0.0883$$

In the individual T-tests, the higher order terms $Year^2$, $SHORTEST\_DISTANCE\_TO\_LRT\_METERS^2$, and $SHORTEST\_DISTANCE\_TO\_LRT\_METERS^3$ are found to be significant.

To ensure that the higher terms are significant in the presence of interaction variables, we ran an ANOVA test to compare the model with main effect, interaction and higher order terms to the model with main effect and interaction.

Hypothesis Statement for ANOVA Test:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0 : Higerher\ order\ terms\ are\ not\ significant$$
$$H_a : at\ least\ one\ \beta_p \neq 0 : At\ leaset\ one\ higher\ order\ term\ is\ significant$$

Table 2 ANOVA Table

Source of Variation	DF	Sum of Squares	Mean Square	F-statistics	P value
Regression	2	3.7778	1.8889	9.8681	0.00005632
Residual	1152	220.51	0.1914149		
Total	1154	224.2878			

From the results of the ANOVA (F=9.8681, p=0.00005632), we have evidence to reject the null hypothesis. This indicates that the higher order terms do significantly predict the community crime rate. As a result, they will be added to our model.

Since the higher terms for year and shortest distance to LRT station were significant predictors of community crime rate, they will be added to our model.

To ensure that the higher terms and interaction are significant predictors, we ran an ANOVA test to compare the model with main effect, interaction and higher order terms to the model with main effect only.

Hypothesis Statement for ANOVA Test:

$$H_0: Model\ with\ main\ effect\ only\ is\ the\ same\ as\ model\ with\ interact\ and\ higher\ order\ terms$$
$$H_a: Model\ with\ interact\ and\ higher\ order\ terms\ is\ significantly\ better\ than\ model\ with\ main\ effect\ only$$
$$\alpha = 0.05$$

Table 3 ANOVA Table

Source of Variation	DF	Sum of Squares	Mean Square	F-statistics	P value
Regression	53	177.31	3.345472	17.477	< 2.2e-16
Residual	1152	220.51	0.1914149		
Total	1205	397.82			

From the results of the ANOVA (F=17.477, p< 2.2e-16), we have evidence to reject the null hypothesis. This indicates that the higher order term and interaction do significantly predict the community crime rate. As a result, they will be added to our model.

The model with interaction and higher order terms is show below:

$$\begin{aligned} \widehat{Y_{\log(crime\_rate)}} = \hat{\beta}_0 &+ \hat{\beta}_1 X_{Year} + \hat{\beta}_2 X_{Sectors_i} \\ &+\hat{\beta}_3 X_{SHOREST\_DISTANCE\_TO\_LRT\_METERS} \\ &+\hat{\beta}_4 X_{SHOREST\_DISTANCE\_TO\_POLICE\_METERS} \\ &+\hat{\beta}_5 X_{male\_percentage} \\ &+\hat{\beta}_6 X_{age\_75\_plus\_percentage} \\ &+\hat{\beta}_7 X_{TotalPermits} \\ &+\hat{\beta}_8 X_{property\_assessmnt\_median} \\ &+\hat{\beta}_9 X_{canada\_unemployment\_rate} \\ &+\hat{\beta}_{10} X_{Year*male\_percentage} \\ &+\hat{\beta}_{11} X_{Year*TotalPrmits} \end{aligned}$$

$$
\begin{aligned}
&+\hat{\beta}_{12} X_{Sectors_i * SHORTEST\_DISTANCE\_TO\_LRT\_METERS} \\
&+\hat{\beta}_{13} X_{Sectors_i * SHORTEST\_DISTANCE\_TO\_POLICE\_METERS} \\
&+\hat{\beta}_{14} X_{Sectors_i * male\_percentage} \\
&+\hat{\beta}_{15} X_{Sectors_i * age\_75\_plus\_percentage} \\
&+\hat{\beta}_{16} X_{Sectors_i * TotalPemits} \\
&+\hat{\beta}_{17} X_{Sectors_i * property\_assessment\_median} \\
&+\hat{\beta}_{18} X_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS * male\_percentage} \\
&+\hat{\beta}_{19} X_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS * age\_75\_plus\_percentage} \\
&+\hat{\beta}_{20} X_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS * property\_assessment\_median} \\
&+\hat{\beta}_{21} X_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS * TotalPermits} \\
&+\hat{\beta}_{22} X_{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS * property\_assessment\_median} \\
&+\hat{\beta}_{23} X_{male\_percentage * age\_75\_plus\_percentage} \\
&+\hat{\beta}_{24} X_{male\_percentage * property\_assessment\_median} \\
&+\hat{\beta}_{25} X_{Year}^2 \\
&+\hat{\beta}_{26} X_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^2 \\
&+\hat{\beta}_{27} X_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^3
\end{aligned}
$$

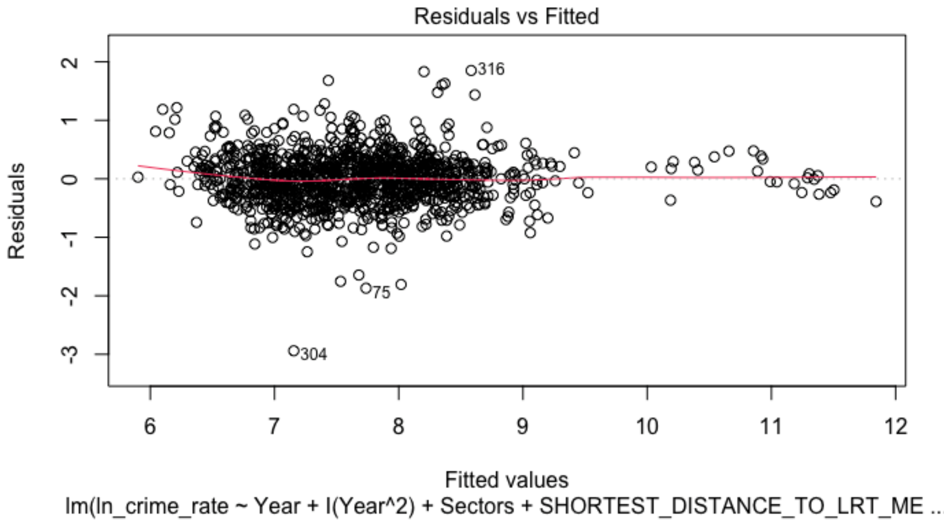
Multiple Regression Assumptions

This section will demonstrate how we test our model to meet various assumptions associated with multiple linear regression. These assumptions must be tested, to ensure that our model are to an extent valid and trustworthy.

Linearity Assumption

The core premise of multiple linear regression is the existence of a linear relationship between the dependent variable and the independent variables. We used residual plots as shown below (Figure 4). This linearity can be visually inspected if there are any discernible patters that are non-linear. From the plot, we see that there is no pattern showing in the trend in our data, suggesting that it passed the linearity assumption.

Figure 4 Residual vs. fitted value plot to check for linearity

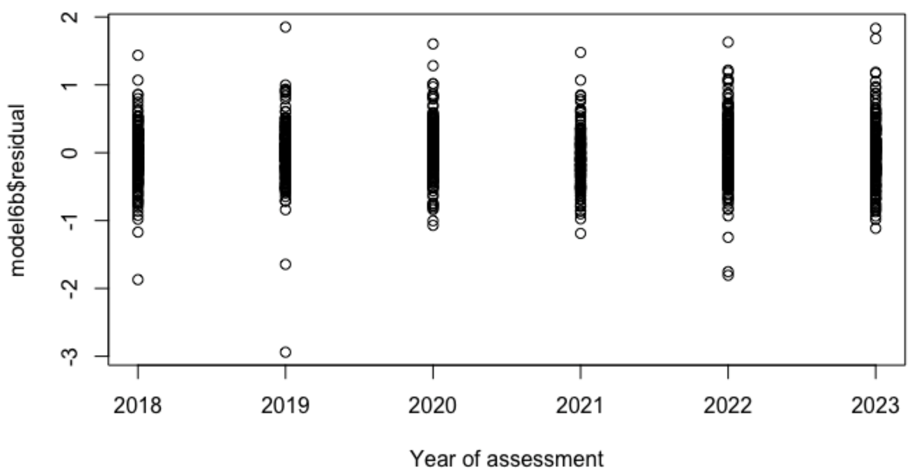


Independence Assumption

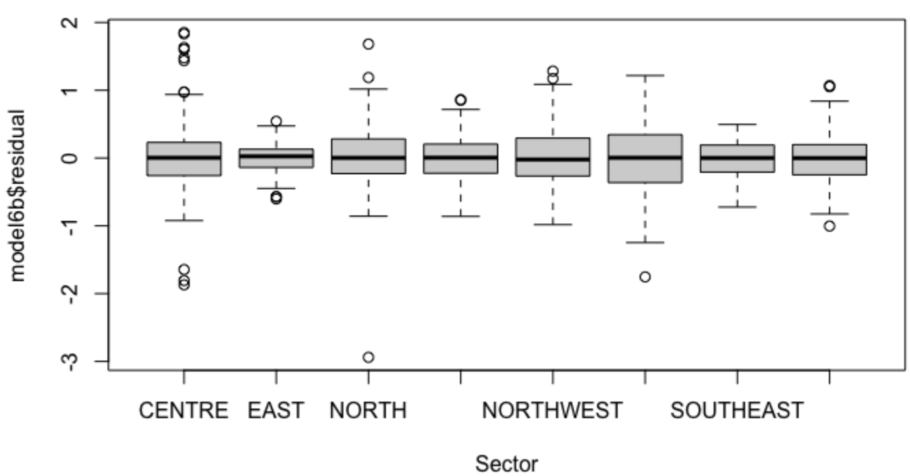
A linear regression model assumes that each observation is independent of the others. The independence assumption would be violated if the observations are clustered, such as if there are repeated measures from the same individual, as in longitudinal data. In our data, communities are clustered in sectors, and crime rates were assessed for communitas from year 2018 to 2023. Because of these, we need to check if the assumption of independent errors is satisfied, using the scatter plot of residual against year, and boxplot of residual by the spatial variable of sectors (Figure 5). We the plots below, we can see that there was no prominent pattern in the plot and the errors were uncorrelated, suggesting that it passed the independence assumption.

Figure 5 Plots to check for independent assumption

a. scatter plot of residual vs. Year of assessment



b. *box plot of residual vs. sectors*



Normality Assumption

The analysis assumes that the residuals are normally distributed. This assumption can be assessed by examining histograms or Q-Q plots of the residuals, or through statistical tests such as the Shapiro-Wilk normality test. the Shapiro-Wilks test reveal W = 0.96852, p-value = 1.332e-15. Thus, at α=0.05, we have evidence to reject the null hypothesis. However, by visual inspection of the histogram of residual, the distribution follows a fairly normal trend with some data points occurring near the tail ends. Additionally, a normal probability plot of residuals is provided. Again, we see that most of the data points approximate the normal line, however, there are a few points deviate the straight at tails indicating the presence of possible outliers.

The normality tests are supplementary to the graphical assessment of normality, for large sample sizes, significant results would be derived even in the case of a small deviation from normality (15, 16), although this small deviation will not affect the results of a parametric test. Thus, based on the visual assessment of normality, we consider that our model meets the normality assumption.

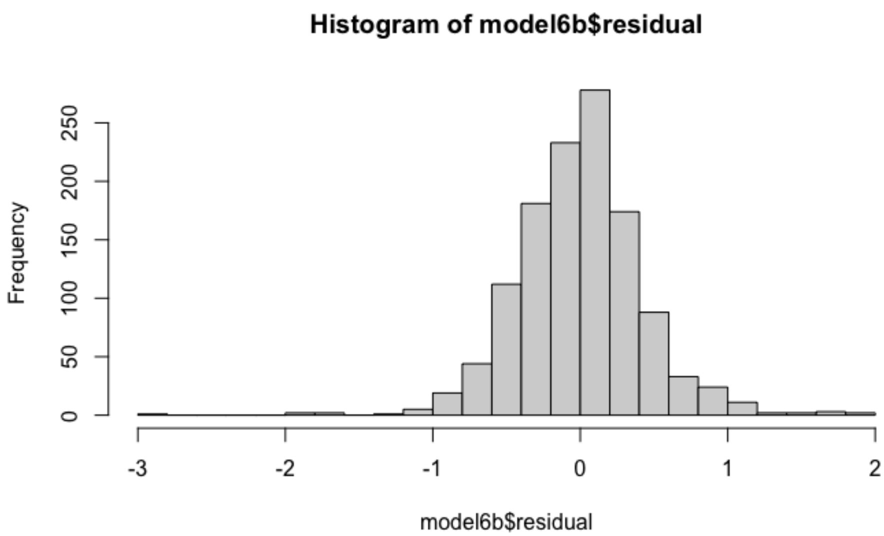
Hypothesis Statement for Shapiro-Wilks Test:

$$\text{Null Hypothesis H}(0)\text{: The sample data is normally distributed}$$
$$\text{Alt. Hypothesis H}(A)\text{: The sample data is not normally distributed}$$

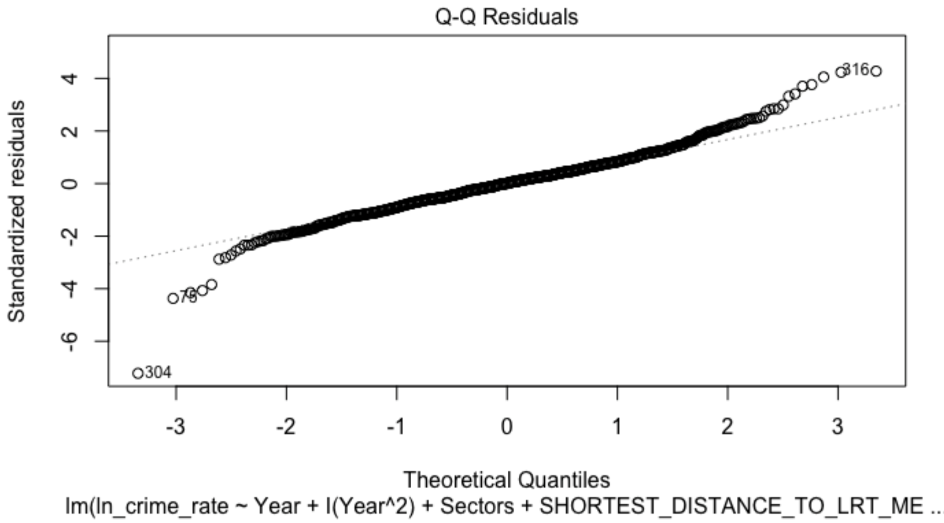
$$\alpha = 0.05$$

Figure 6 Plots to check for normality distribution

a. Histogram of residual



b. Q-Q plots of the residuals



Equal Variance Assumption

Hypothesis Statement for Breusch-Pagan Test:

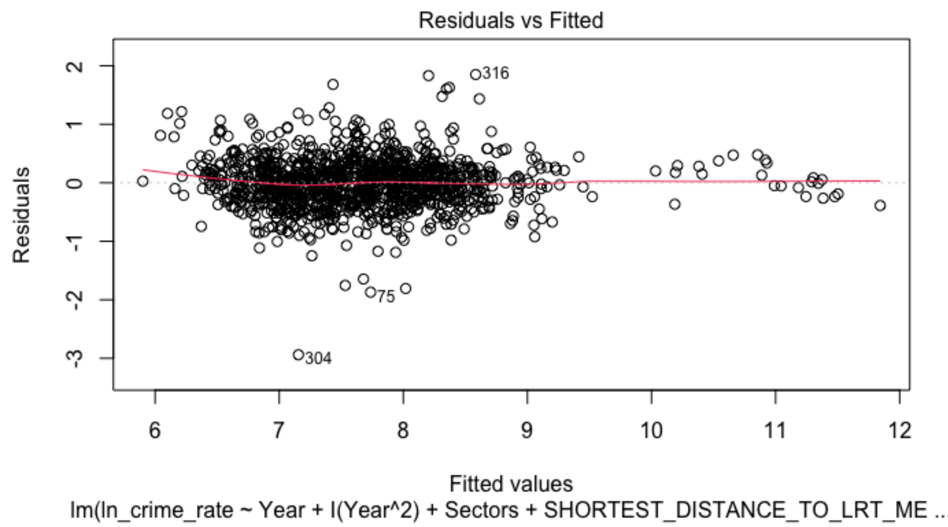
$$\text{Null Hypothesis H}(0)\text{: Heteroscedascity is not present}$$
$$\text{Alt. Hypothesis H}(A)\text{: Heteroscedascity is present}$$
$$\alpha = 0.05$$

Next, we tested to see if our data is homoscedastic through a plot of fits to residuals as well as the Breusch-Pagan test. Looking at the plot of fits to residuals, we see that the residuals are spread equally along the ranges of the fitted values. On the scale-location plot between fitted values and standardized residuals, we see a horizontal line with equally (randomly) spread points. These plots suggest that our model likely meet the equal variance assumption. However, results of the Breusch-Pagan test (BP = 184.6, p-value = 1.729e-12). we would reject the null hypothesis in favor of the alternative, suggesting that our model fails to be homoscedastic.

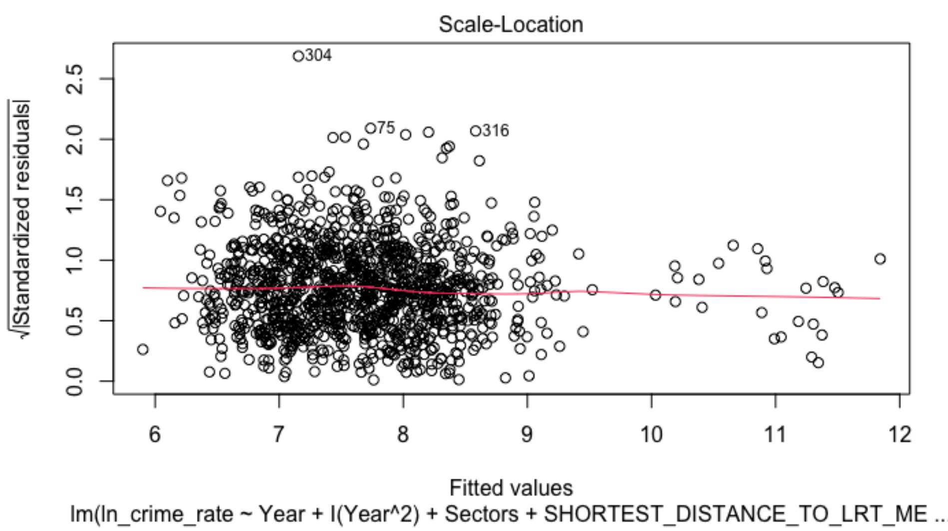
We consider that for large sample sizes, significant result of Breusch-Pagan test would be derived even in the case of a small deviation from equal variance, and from the residual vs fitted plot, we can see that there are less samples with large, fitted values, which may also contribute to a significant Breusch-Pagan test. Thus, we consider this small deviation of the equal variance will not affect the validity of our model.

Figure 7 Plots to check for normality distribution

a. Residual vs. fitted value



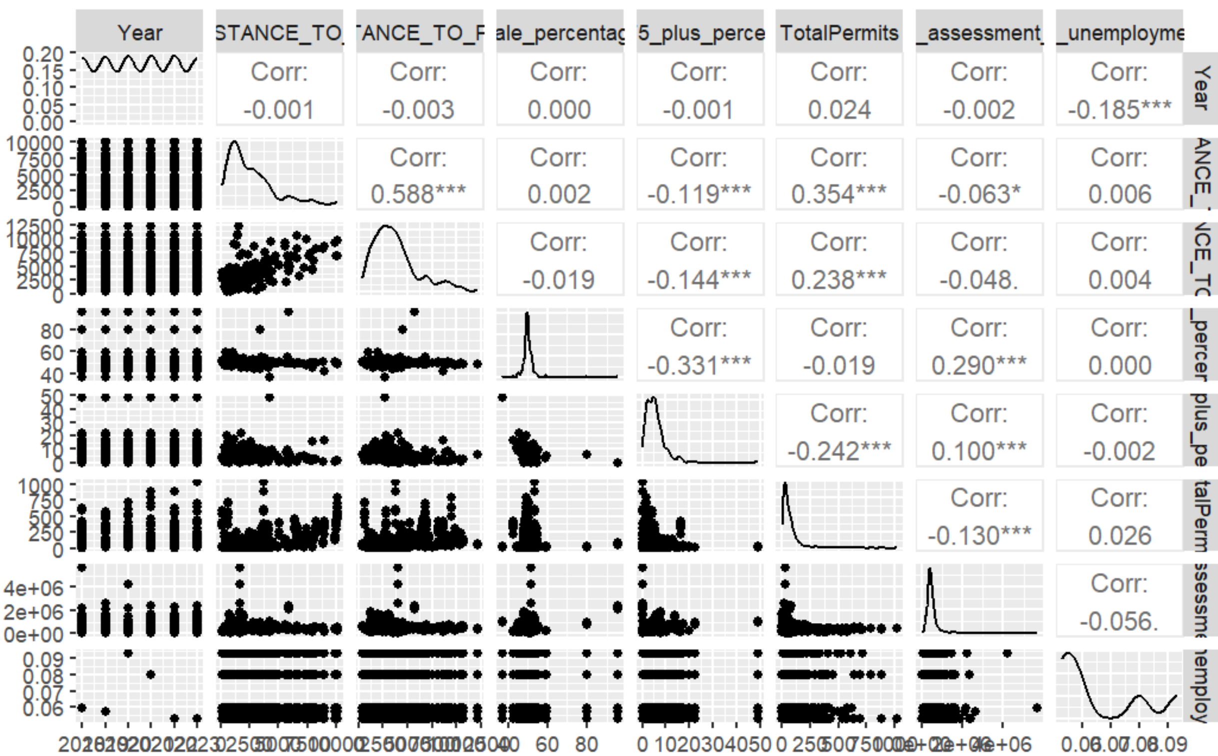
b. Scale location plot



Multicollinearity Assumption

To test for multicollinearity in our model, we have examined the VIF and the VIF values for all predictors were smaller than 5.0 (Table 2), suggesting there may be moderate collinearity, but it is not severe enough to warrant corrective measures. We also ran a ggpair function to ensure that there were no extremely high correlation (r>0.8) between the predictors in our model (Figure 8).

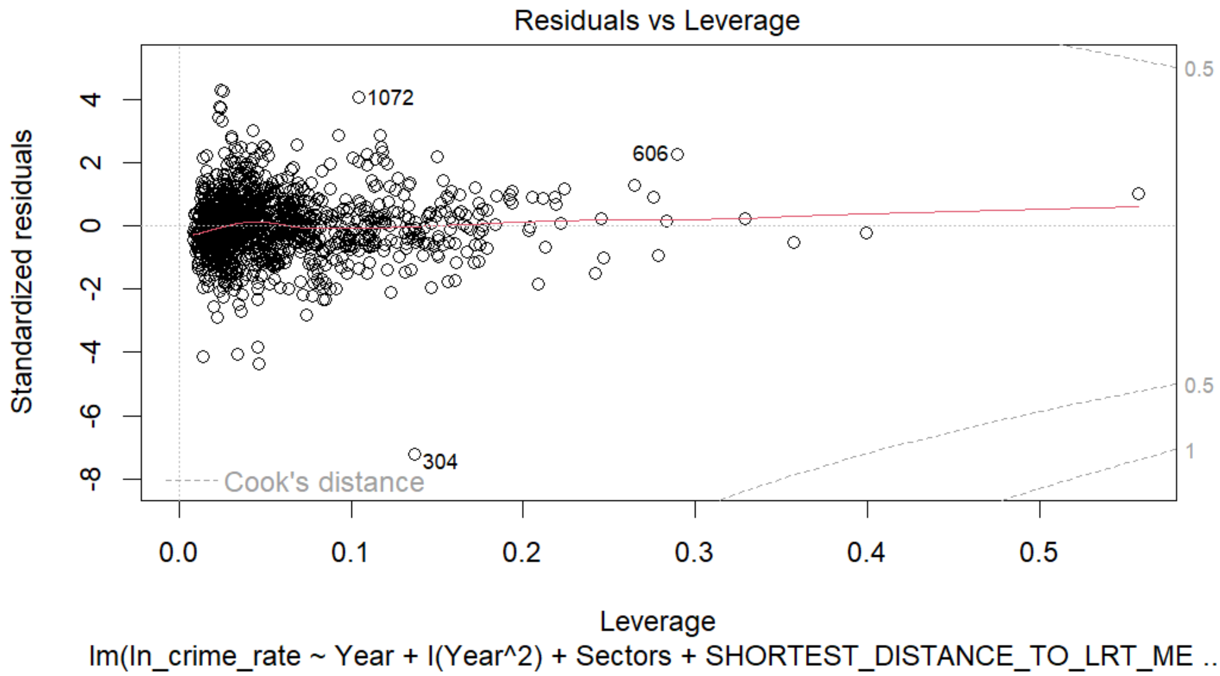
Figure 8 GGPAIR Plot to check for multicollinearity



Influential Points and Outliers

In linear regression we assume that the relationship between the independent and dependent variables is linear. Outliers may indicate a non-linear relationship or the presence of influential points that violate this assumption leading to skewed predictions. To check for this, we plot the values against the Cook's distance. From the residual vs. leverage plot (Figure 9) we can see that there are no data points beyond Cook's distance. This suggests that there are no influential points that have an effect on our regression results.

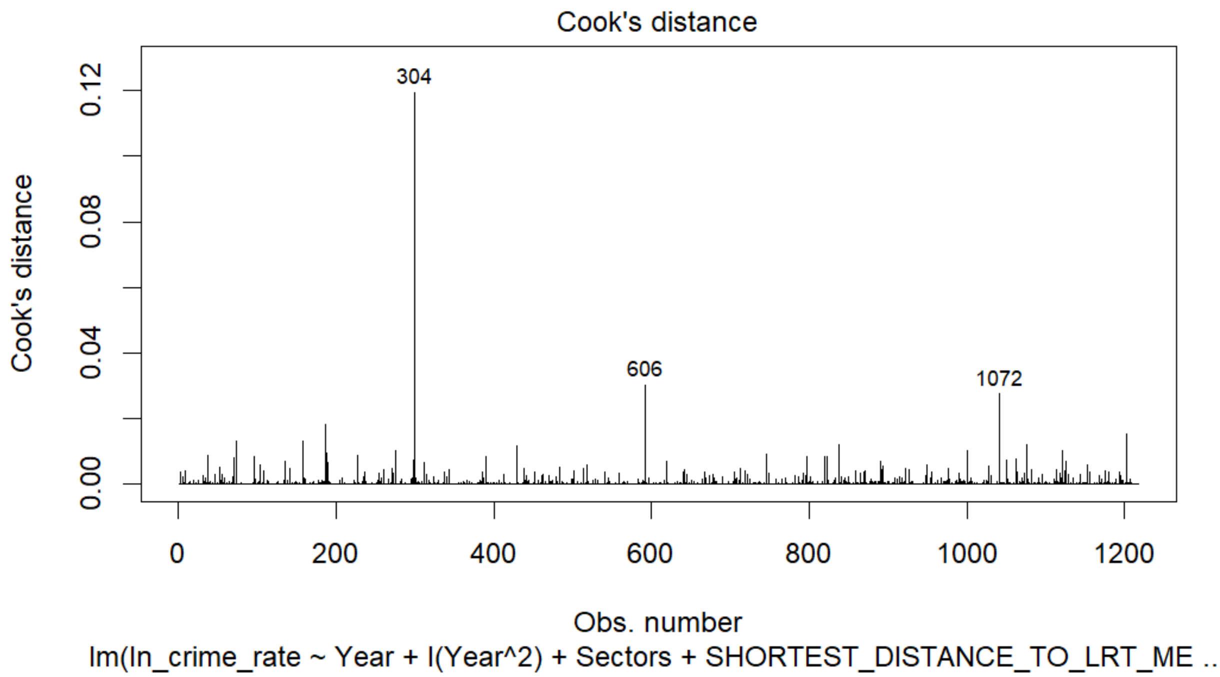
Figure 9 Plot to check for influential points



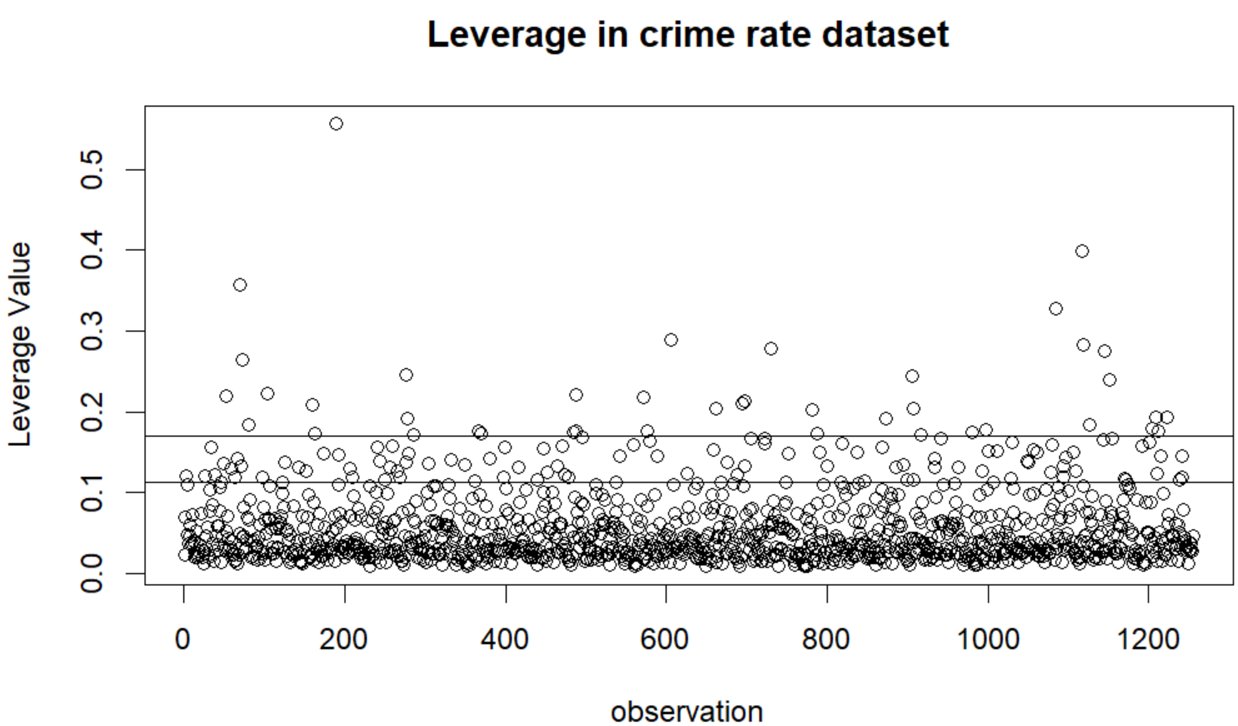
The plot for Cook's distance by observations (Figure 10 a) shows that there are three observations (304, 606, 1072) that have the highest Cook's distance, however, their Cook's distance values are all less than 0.5. so they are not influential. Then we use the leverage plot to identify outlier beyond 2p/n, and 3p/n. We re-run the model with removing both of these the outliers, the R^2_{adj} value decrease from 75.9% to 69% after removing the outlier beyond 2p/n, and 3p/n, thus these outliers deemed influential.

Figure 10 Plot to check for outliers

a.



b.



Our final and best fitted model includes main effect, interaction and higher order terms. The final model can be written as:

$$
\begin{aligned}
Y_{\log\,(\widehat{crime\_rate})} = \hat{\beta}_0 + \hat{\beta}_1 X_{Year} + \hat{\beta}_2 X_{Sectors} \\
&+\hat{\beta}_3 X_{SHOREST\_DISTANCE\_TO\_LRT\_METERS} \\
&+\hat{\beta}_4 X_{SHOREST\_DISTANCE\_TO\_POLICE\_METERS} \\
&+\hat{\beta}_5 X_{male\_percentage} \\
&+\hat{\beta}_6 X_{age\_75\_plus\_percentage} \\
&+\hat{\beta}_7 X_{TotalPermits} \\
&+\hat{\beta}_8 X_{property\_assessmnt\_median} \\
&+\hat{\beta}_9 X_{canada\_unemployment\_rate} \\
&+\hat{\beta}_{10} X_{Year*male\_percentage} \\
&+\hat{\beta}_{11} X_{Year*TotalPrmits} \\
&+\hat{\beta}_{12} X_{Sectors_i*SHORTEST\_DISTANCE\_TO\_LRT\_METERS} \\
&+\hat{\beta}_{13} X_{Sectors_i*SHORTEST\_DISTANCE\_TO\_POLICE\_METERS} \\
&+\hat{\beta}_{14} X_{Sectors(i)*male\_percentage} \\
&+\hat{\beta}_{15} X_{Sector(i)*age\_75\_plus\_percentage} \\
&+\hat{\beta}_{16} X_{Sectros(i)*TotalPemits} \\
&+\hat{\beta}_{17} X_{Sectors*property\_assessment\_median} \\
&+\hat{\beta}_{18} X_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS*male\_percentage} \\
&+\hat{\beta}_{19} X_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS*age\_75\_plus\_percentage} \\
&+\hat{\beta}_{20} X_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS*property\_assessment\_median} \\
&+\hat{\beta}_{21} X_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS*TotalPermits} \\
&+\hat{\beta}_{22} X_{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS*property\_assessment\_median} \\
&+\hat{\beta}_{23} X_{male\_percentage*age\_75\_plus\_percentage} \\
&+\hat{\beta}_{24} X_{male\_percentage*property\_assessment\_median} \\
&+\hat{\beta}_{25} X^2_{Year} \\
&+\hat{\beta}_{26} X^2_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS} \\
&+\hat{\beta}_{27} X^3_{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}
\end{aligned}
$$

R^2_{adj} and Residual standard error of the Final Best fitted Model

R^2_{adj}= 0.759, this means that 75.9% variation of the dependent variable log(crime rate) can be explained by the final model.

Residual standard error = 0.4375, this value means that the standard deviation of the unexplained variance by the model in estimation of dependent variable log(crime rate) is 0.4375.

Interpreting Coefficients

The variable “sectors” is a categorical variable and includes 8 values. In our model, the value CENTER” is the reference group. According to our final model, we obtained the equations below.

When sector is ESAT, the equation is

$$
\begin{aligned}
\hat{Y}_{\log(\text{crime\_rate})} = & -78870 + 78.4 \cdot X_{\text{Year}} - 0.01948 \cdot X_{\text{Year}^2} - 1.324 \cdot X_{\text{SectorsEAST}} \\
& - 0.00243 \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 7.361 \times 10^{-8} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^2} \\
& - 4.518 \times 10^{-12} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^3} + 1.04 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& - 8.586 \cdot X_{\text{male\_percentage}} + 0.2464 \cdot X_{\text{age\_75\_plus\_percentage}} - 0.9201 \cdot X_{\text{TotalPermits}} + 4.281 \times 10^{-6} \cdot X_{\text{property\_assessment\_median}} \\
& - 5.171 \cdot X_{\text{canada\_unemployment\_rate}} + 0.004294 \cdot X_{\text{Year}} \cdot X_{\text{male\_percentage}} + 0.0004562 \cdot X_{\text{Year}} \cdot X_{\text{TotalPermits}} \\
& - 0.0004616 \cdot X_{\text{SectorsEAST}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 0.0004969 \cdot X_{\text{SectorsEAST}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& + 0.001377 \cdot X_{\text{SectorsEAST}} \cdot X_{\text{male\_percentage}} + 0.0806 \cdot X_{\text{SectorsEAST}} \cdot X_{\text{age\_75\_plus\_percentage}} + 0.0008713 \cdot X_{\text{SectorsEAST}} \cdot X_{\text{TotalPermits}} \\
& + 3.807 \times 10^{-7} \cdot X_{\text{SectorsEAST}} \cdot X_{\text{property\_assessment\_median}} + 3.873 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{male\_percentage}} \\
& + 1.727 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{age\_75\_plus\_percentage}} + 4.484 \times 10^{-7} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{TotalPermits}} \\
& - 1.599 \times 10^{-10} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \cdot X_{\text{property\_assessment\_median}} - 0.005405 \cdot X_{\text{male\_percentage}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& - 8.529 \times 10^{-8} \cdot X_{\text{male\_percentage}} \cdot X_{\text{property\_assessment\_median}}
\end{aligned}
$$

When sector is NORTH, the equation is

$$
\begin{aligned}
\hat{Y}_{\log(\text{crime\_rate})} = & -78870 + 78.4 \cdot X_{\text{Year}} - 0.01948 \cdot X_{\text{Year}^2} + 14.07 \cdot X_{\text{SectorsNORTH}} \\
& - 0.00243 \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 7.361 \times 10^{-8} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^2} \\
& - 4.518 \times 10^{-12} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^3} + 1.04 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& - 8.586 \cdot X_{\text{male\_percentage}} + 0.2464 \cdot X_{\text{age\_75\_plus\_percentage}} - 0.9201 \cdot X_{\text{TotalPermits}} \\
& + 4.281 \times 10^{-6} \cdot X_{\text{property\_assessment\_median}} - 5.171 \cdot X_{\text{canada\_unemployment\_rate}} \\
& + 0.004294 \cdot X_{\text{Year}} \cdot X_{\text{male\_percentage}} + 0.0004562 \cdot X_{\text{Year}} \cdot X_{\text{TotalPermits}} \\
& + 2.601 \times 10^{-5} \cdot X_{\text{SectorsNORTH}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \\
& - 1.568 \times 10^{-5} \cdot X_{\text{SectorsNORTH}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& - 0.2949 \cdot X_{\text{SectorsNORTH}} \cdot X_{\text{male\_percentage}} \\
& - 0.02659 \cdot X_{\text{SectorsNORTH}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& - 0.003733 \cdot X_{\text{SectorsNORTH}} \cdot X_{\text{TotalPermits}} \\
& + 7.487 \times 10^{-7} \cdot X_{\text{SectorsNORTH}} \cdot X_{\text{property\_assessment\_median}} \\
& + 3.873 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{male\_percentage}} \\
& + 1.727 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& + 4.484 \times 10^{-7} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{TotalPermits}} \\
& - 1.599 \times 10^{-10} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \cdot X_{\text{property\_assessment\_median}} \\
& - 0.005405 \cdot X_{\text{male\_percentage}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& - 8.529 \times 10^{-8} \cdot X_{\text{male\_percentage}} \cdot X_{\text{property\_assessment\_median}}
\end{aligned}
$$

When sector is NORTHEAST, the equation is

$$
\begin{aligned}
\hat{Y}_{\log(\text{crime\_rate})} = & -78870 + 78.4 \cdot X_{\text{Year}} - 0.01948 \cdot X_{\text{Year}^2} + 6.831 \cdot X_{\text{SectorsNORTHEAST}} \\
& - 0.00243 \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 7.361 \times 10^{-8} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^2} \\
& - 4.518 \times 10^{-12} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^3} + 1.04 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& - 8.586 \cdot X_{\text{male\_percentage}} + 0.2464 \cdot X_{\text{age\_75\_plus\_percentage}} - 0.9201 \cdot X_{\text{TotalPermits}} \\
& + 4.281 \times 10^{-6} \cdot X_{\text{property\_assessment\_median}} - 5.171 \cdot X_{\text{canada\_unemployment\_rate}} \\
& + 0.004294 \cdot X_{\text{Year}} \cdot X_{\text{male\_percentage}} + 0.0004562 \cdot X_{\text{Year}} \cdot X_{\text{TotalPermits}} \\
& - 2.455 \times 10^{-5} \cdot X_{\text{SectorsNORTHEAST}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \\
& + 3.835 \times 10^{-5} \cdot X_{\text{SectorsNORTHEAST}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& - 0.1353 \cdot X_{\text{SectorsNORTHEAST}} \cdot X_{\text{male\_percentage}} + 0.101 \cdot X_{\text{SectorsNORTHEAST}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& - 0.00314 \cdot X_{\text{SectorsNORTHEAST}} \cdot X_{\text{TotalPermits}} - 1.763 \times 10^{-6} \cdot X_{\text{SectorsNORTHEAST}} \cdot X_{\text{property\_assessment\_median}} \\
& + 3.873 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{male\_percentage}} \\
& + 1.727 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& + 4.484 \times 10^{-7} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{TotalPermits}} \\
& - 1.599 \times 10^{-10} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \cdot X_{\text{property\_assessment\_median}} \\
& - 0.005405 \cdot X_{\text{male\_percentage}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& - 8.529 \times 10^{-8} \cdot X_{\text{male\_percentage}} \cdot X_{\text{property\_assessment\_median}}
\end{aligned}
$$

When sector is NORTHWEST, the equation is

$$
\begin{aligned}
\hat{Y}_{\log(\text{crime\_rate})} = & -78870 + 78.4 \cdot X_{\text{Year}} - 0.01948 \cdot X_{\text{Year}^2} - 7.486 \cdot X_{\text{SectorsNORTHWEST}} \\
& - 0.00243 \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 7.361 \times 10^{-8} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^2} \\
& - 4.518 \times 10^{-12} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^3} + 1.04 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& - 8.586 \cdot X_{\text{male\_percentage}} + 0.2464 \cdot X_{\text{age\_75\_plus\_percentage}} - 0.9201 \cdot X_{\text{TotalPermits}} \\
& + 4.281 \times 10^{-6} \cdot X_{\text{property\_assessment\_median}} - 5.171 \cdot X_{\text{canada\_unemployment\_rate}} \\
& + 0.004294 \cdot X_{\text{Year}} \cdot X_{\text{male\_percentage}} + 0.0004562 \cdot X_{\text{Year}} \cdot X_{\text{TotalPermits}} \\
& + 0.0001077 \cdot X_{\text{SectorsNORTHWEST}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \\
& - 3.832 \times 10^{-5} \cdot X_{\text{SectorsNORTHWEST}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& + 0.1243 \cdot X_{\text{SectorsNORTHWEST}} \cdot X_{\text{male\_percentage}} \\
& + 0.006047 \cdot X_{\text{SectorsNORTHWEST}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& + 0.004583 \cdot X_{\text{SectorsNORTHWEST}} \cdot X_{\text{TotalPermits}} \\
& + 7.823 \times 10^{-7} \cdot X_{\text{SectorsNORTHWEST}} \cdot X_{\text{property\_assessment\_median}} \\
& + 3.873 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{male\_percentage}} \\
& + 1.727 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& + 4.484 \times 10^{-7} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{TotalPermits}} \\
& - 1.599 \times 10^{-10} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \cdot X_{\text{property\_assessment\_median}} \\
& - 0.005405 \cdot X_{\text{male\_percentage}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& - 8.529 \times 10^{-8} \cdot X_{\text{male\_percentage}} \cdot X_{\text{property\_assessment\_median}}
\end{aligned}
$$

When sector is SOUTH, the equation is

$$
\begin{aligned}
\hat{Y}_{\log(\text{crime\_rate})} = & -78870 + 78.4 \cdot X_{\text{Year}} - 0.01948 \cdot X_{\text{Year}^2} - 4.812 \cdot X_{\text{SectorsSOUTH}} \\
& - 0.00243 \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 7.361 \times 10^{-8} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^2} \\
& - 4.518 \times 10^{-12} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^3} + 1.04 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& - 8.586 \cdot X_{\text{male\_percentage}} + 0.2464 \cdot X_{\text{age\_75\_plus\_percentage}} - 0.9201 \cdot X_{\text{TotalPermits}} \\
& + 4.281 \times 10^{-6} \cdot X_{\text{property\_assessment\_median}} - 5.171 \cdot X_{\text{canada\_unemployment\_rate}} \\
& + 0.004294 \cdot X_{\text{Year}} \cdot X_{\text{male\_percentage}} + 0.0004562 \cdot X_{\text{Year}} \cdot X_{\text{TotalPermits}} \\
& - 0.0001764 \cdot X_{\text{SectorsSOUTH}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \\
& + 0.0001026 \cdot X_{\text{SectorsSOUTH}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& + 0.06792 \cdot X_{\text{SectorsSOUTH}} \cdot X_{\text{male\_percentage}} + 0.02133 \cdot X_{\text{SectorsSOUTH}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& - 0.0007758 \cdot X_{\text{SectorsSOUTH}} \cdot X_{\text{TotalPermits}} + 1.557 \times 10^{-6} \cdot X_{\text{SectorsSOUTH}} \cdot X_{\text{property\_assessment\_median}} \\
& + 3.873 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{male\_percentage}} \\
& + 1.727 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& + 4.484 \times 10^{-7} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{TotalPermits}} \\
& - 1.599 \times 10^{-10} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \cdot X_{\text{property\_assessment\_median}} \\
& - 0.005405 \cdot X_{\text{male\_percentage}} \cdot X_{\text{age\_75\_plus\_percentage}} - 8.529 \times 10^{-8} \cdot X_{\text{male\_percentage}} \cdot X_{\text{property\_assessment\_median}}
\end{aligned}
$$

When sector is SOUTHEAST, the equation is

$$
\begin{aligned}
\hat{Y}_{\log(\text{crime\_rate})} = & -78870 + 78.4 \cdot X_{\text{Year}} - 0.01948 \cdot X_{\text{Year}^2} - 4.992 \cdot X_{\text{SectorsSOUTHEAST}} \\
& - 0.00243 \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 7.361 \times 10^{-8} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^2} \\
& - 4.518 \times 10^{-12} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^3} + 1.04 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& - 8.586 \cdot X_{\text{male\_percentage}} + 0.2464 \cdot X_{\text{age\_75\_plus\_percentage}} - 0.9201 \cdot X_{\text{TotalPermits}} \\
& + 4.281 \times 10^{-6} \cdot X_{\text{property\_assessment\_median}} - 5.171 \cdot X_{\text{canada\_unemployment\_rate}} \\
& + 0.004294 \cdot X_{\text{Year}} \cdot X_{\text{male\_percentage}} + 0.0004562 \cdot X_{\text{Year}} \cdot X_{\text{TotalPermits}} \\
& + 0.0001219 \cdot X_{\text{SectorsSOUTHEAST}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \\
& + 0.0001316 \cdot X_{\text{SectorsSOUTHEAST}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& + 0.04239 \cdot X_{\text{SectorsSOUTHEAST}} \cdot X_{\text{male\_percentage}} + 0.1931 \cdot X_{\text{SectorsSOUTHEAST}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& - 0.005772 \cdot X_{\text{SectorsSOUTHEAST}} \cdot X_{\text{TotalPermits}} + 1.327 \times 10^{-6} \cdot X_{\text{SectorsSOUTHEAST}} \cdot X_{\text{property\_assessment\_median}} \\
& + 3.873 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{male\_percentage}} \\
& + 1.727 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& + 4.484 \times 10^{-7} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{TotalPermits}} \\
& - 1.599 \times 10^{-10} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \cdot X_{\text{property\_assessment\_median}} \\
& - 0.005405 \cdot X_{\text{male\_percentage}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& - 8.529 \times 10^{-8} \cdot X_{\text{male\_percentage}} \cdot X_{\text{property\_assessment\_median}}
\end{aligned}
$$

When sector is WEST, the equation would is

$$
\begin{aligned}
\hat{Y}_{\log(\text{crime\_rate})} = & -78870 + 78.4 \cdot X_{\text{Year}} - 0.01948 \cdot X_{\text{Year}^2} - 1.767 \cdot X_{\text{SectorsWEST}} \\
& - 0.00243 \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 7.361 \times 10^{-8} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^2} \\
& - 4.518 \times 10^{-12} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^3} + 1.04 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& - 8.586 \cdot X_{\text{male\_percentage}} + 0.2464 \cdot X_{\text{age\_75\_plus\_percentage}} - 0.9201 \cdot X_{\text{TotalPermits}} \\
& + 4.281 \times 10^{-6} \cdot X_{\text{property\_assessment\_median}} - 5.171 \cdot X_{\text{canada\_unemployment\_rate}} \\
& + 0.004294 \cdot X_{\text{Year}} \cdot X_{\text{male\_percentage}} + 0.0004562 \cdot X_{\text{Year}} \cdot X_{\text{TotalPermits}} \\
& + 0.0002448 \cdot X_{\text{SectorsWEST}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \\
& - 0.0001216 \cdot X_{\text{SectorsWEST}} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} + 0.02565 \cdot X_{\text{SectorsWEST}} \cdot X_{\text{male\_percentage}} \\
& + 0.01838 \cdot X_{\text{SectorsWEST}} \cdot X_{\text{age\_75\_plus\_percentage}} - 0.001597 \cdot X_{\text{SectorsWEST}} \cdot X_{\text{TotalPermits}} \\
& - 5.244 \times 10^{-7} \cdot X_{\text{SectorsWEST}} \cdot X_{\text{property\_assessment\_median}} \\
& + 3.873 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{male\_percentage}} \\
& + 1.727 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& + 4.484 \times 10^{-7} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{TotalPermits}} \\
& - 1.599 \times 10^{-10} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \cdot X_{\text{property\_assessment\_median}} \\
& - 0.005405 \cdot X_{\text{male\_percentage}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& - 8.529 \times 10^{-8} \cdot X_{\text{male\_percentage}} \cdot X_{\text{property\_assessment\_median}}
\end{aligned}
$$

When the sector is SOUTHWEST, the equation is

$$
\begin{aligned}
\hat{Y}_{\log(\text{crime\_rate})} = & -78870 + 78.4 \cdot X_{\text{Year}} - 0.01948 \cdot X_{\text{Year}^2} \\
& - 0.00243 \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 7.361 \times 10^{-8} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^2} \\
& - 4.518 \times 10^{-12} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}^3} + 1.04 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \\
& - 8.586 \cdot X_{\text{male\_percentage}} + 0.2464 \cdot X_{\text{age\_75\_plus\_percentage}} - 0.9201 \cdot X_{\text{TotalPermits}} \\
& + 4.281 \times 10^{-6} \cdot X_{\text{property\_assessment\_median}} - 5.171 \cdot X_{\text{canada\_unemployment\_rate}} \\
& + 0.004294 \cdot X_{\text{Year}} \cdot X_{\text{male\_percentage}} + 0.0004562 \cdot X_{\text{Year}} \cdot X_{\text{TotalPermits}} \\
& + 3.873 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{male\_percentage}} \\
& + 1.727 \times 10^{-5} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& + 4.484 \times 10^{-7} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} \cdot X_{\text{TotalPermits}} \\
& - 1.599 \times 10^{-10} \cdot X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} \cdot X_{\text{property\_assessment\_median}} \\
& - 0.005405 \cdot X_{\text{male\_percentage}} \cdot X_{\text{age\_75\_plus\_percentage}} \\
& - 8.529 \times 10^{-8} \cdot X_{\text{male\_percentage}} \cdot X_{\text{property\_assessment\_median}}
\end{aligned}
$$

When the sector is SOUTHWEST, the coefficient can be interpreted as below.

SHORTEST_DISTANCE_TO_POLICE_METERS: For each one units closer to the nearest police station, the log(crime rate) increase by ($1.04 \times 10^{-5} \pm 1.599 \times 10^{-10} X_{\text{property\_assessment\_median}}$) units.

male_percentage : For each one percentage increase in male, the log(crime rate) increase by (-8.586+$0.004294 X_{\text{male\_percentage}}$+$3.873 \times 10^{-5} X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}}$+$-0.005405 X_{\text{age\_75\_plus\_percentage}}$+$-8.529 \times 10^{-8} \cdot X_{\text{property\_assessment\_median}}$) units.

age_75_plus_percentage: For one percentage increase, the log(crime rate) increase by (0.2464+$1.727 \times 10^{-5} X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} - 0.005405 \cdot X_{\text{male\_percentage}}$) units.

TotalPermits : For each totalpermits units increase, the log(crime rate) increase by(-0.9201+$0.0004562 X_{\text{Year}}$++$4.484 \times 10^{-7} X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}}$) units.

property_assessment_median : For each one units of property_assessment_median increase, the log(crime rate) increase by (4.281×10^{-6}+$-1.599 \times 10^{-10} X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}}$+ $-8.529 \times 10^{-8} \cdot X_{\text{male\_percentage}}$) units.

canada_unemployment_rate: For each one units if Canada_unemployment_rate increase, the log(crime rate) increase by (- 5.171) units.

Between the sector WEST and CENTER, the difference in the value of log(crime rate) is

$-1.767+0.0002448X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} - 0.0001216X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} + 0.02565X_{\text{male\_percentage}}+0.01838X_{\text{age\_75\_plus\_percentage}} - 0.001597X_{\text{TotalPermits}} - 5.244 \times 10^{-7}X_{\text{property\_assessment\_median}}$

Between the SOUTHEAST and CENTER, the difference in the value of log(crime rate) is

$-4.992+0.0001219X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 0.0001316X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}}+0.04239X_{\text{male\_percentage}} + 0.1931X_{\text{age\_75\_plus\_percentage}} - 0.005772X_{\text{TotalPermits}} + 1.327 \times 10^{-6}X_{\text{property\_assessment\_median}}$

Between the SOUTH and CENTER, The difference in the value of log(crime rate) is

$-4.812-0.0001764X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 0.0001026X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}}+0.06792X_{\text{male\_percentage}} + 0.02133X_{\text{age\_75\_plus\_percentage}} - 0.0007758X_{\text{TotalPermits}} + 1.557 \times 10^{-6}X_{\text{property\_assessment\_median}}$

Between the NORTHWEST and CENTER, the difference of log(crime rate) is

$-7.486+0.0001077X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} - 3.832 \times 10^{-5}X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}}+0.1243X_{\text{male\_percentage}}+ 0.006047X_{\text{age\_75\_plus\_percentage}}+0.004583X_{\text{TotalPermits}}+7.823 \times 10^{-7}X_{\text{property\_assessment\_median}}$

Between the NORTHEAST and CENTER, the difference of log(crime rate) is

$6.831-2.455 \times 10^{-5}X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 3.835 \times 10^{-5}X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} - 0.1353X_{\text{male\_percentage}} + 0.101X_{\text{age\_75\_plus\_percentage}} - 0.00314X_{\text{TotalPermits}} - 1.763 \times 10^{-6}X_{\text{property\_assessment\_median}}$

Between the NORTH and CENTER, the difference of log(crime rate) is

$14.07+2.601 \times 10^{-5}X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} - 1.568 \times 10^{-5}X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} - 0.2949X_{\text{male\_percentage}} - 0.02659X_{\text{age\_75\_plus\_percentage}} - 0.003733X_{\text{TotalPermits}} + 7.487 \times 10^{-7}X_{\text{property\_assessment\_median}}$

Between the EAST and CENTER, the difference of log(crime rate) is

$-1.324-0.0004616X_{\text{SHORTEST\_DISTANCE\_TO\_LRT\_METERS}} + 0.0004969X_{\text{SHORTEST\_DISTANCE\_TO\_POLICE\_METERS}} + 0.001377X_{\text{male\_percentage}} + 0.0806X_{\text{age\_75\_plus\_percentage}} + 0.0008713X_{\text{TotalPermits}} + 3.807 \times 10^{-7}X_{\text{property\_assessment\_median}}$

In summary, our study found that, spatial and economic factors were significant predictors for community crime rate. Compared to center sector, the crime rate was lower in the other sectors (main effect: varying from 2 - 3 / 100,000 population). And sectors interacted with most of other predictors in the model, and the effect size and direction of the interaction differed according to the different geographical divisions. The other spatial variable

Distance between community and LRT was also found to be associated with crime rate, with main effect of as distance increases by 1km, crime rate decreased by 1 / 100,000 population. Canada's unemployment rate was found to be positively associated with crime rate: as 1% increase in Canada's unemployment rate, crime rate increases by 20/ 100,000 population (main effect). Number of building permit was also found to be positively associated with crime rate: 1 building permit increases, crime rate increases by 1/100,000 (main effect). These foundlings are valuable for understanding the nuanced relationship between various factors and crime rates, and it can help inform targeted strategies for crime prevention and intervention tailored to specific geographic areas.

Discussion

Using our best fit model, we focus on Calgary Communities that will be impacted by the future Green LRT Line(17). We apply the PREDICT function on our observed dataset and evaluate the results of the prediction plots.

After we will employ the proposed Green Leg LRT Transit Station locations to update our predictor variable "Shortest LRT distance in meters" and plot the variance in prediction values.

See Appendix A for sources of LRT Transit Station dataset, SQL query performed, distance calculation.

First figure for each Prediction Plot

- Dataset filtered to just the specific community for 2018-2023, with LRT locations as it is today
- Average: Mean of the total crime counts
- Deviation: Observed data point to Average line
- Predicted: Fitted values from applying PREDICT function with the Regression model
- Regression: Variance between Fitted value and Average line
- Residual: Variance between Observed Data Point and Fitted Value

Second figure for each Prediction Plot

- Dataset is updated using R code to go through the dataset and update the column for "Shortest Distance to LRT" with values that consider Green Line LRT locations. Table below shows the previous value used as "Current" in the dataset then updated value as "Green".
    - Then apply PREDICT function with our regression model but with this updated dataset

Table 4 Communities impacted by Green Line LRT

# Community Code	Community Name	Current	Green
SHI	SHEPARD INDUSTRIAL	5780.38	614.656
BLN	BELTLINE	252.401	252.401
MCT	MCKENZIE TOWNE	7682.4	350.192
OGD	OGDEN	3861.44	882.95
CRE	CRESCENT HEIGHTS	1445.83	80.9746
SET	SETON	8981.92	484.164
THO	THORNCLIFFE	4570.6	465.315
HAR	HARVEST HILLS	6795.57	1159.98
TUX	TUXEDO PARK	2448.26	517.127
AUB	AUBURN BAY	7785.31	1097.43
HPK	HIGHLAND PARK	3440.41	29.6261
BED	BEDDINGTON HEIGHTS	5936.87	1047
HUN	HUNTINGTON HILLS	5673.79	566.373
LIV	LIVINGSTON	10162.4	671.398
CAR	CARRINGTON	10042.8	756.349
CHV	COUNTRY HILLS VILLAGE	7926.13	638.325
RAM	RAMSAY	1051.63	584.051
DNC	DOWNTOWN COMMERCIAL CORE	76.8825	76.8825
EAU	EAU CLAIRE	752.986	373.508

- New Average: Average of the updated Fitted values
- Prediction Updated: Fitted values from the updated dataset (Green LRTs)
- Prediction Variance: Variance between the previous Prediction and Updated Prediction
- Previous Prediction: Fitted values from the previous dataset (Current LRTs)

Prediction Plots

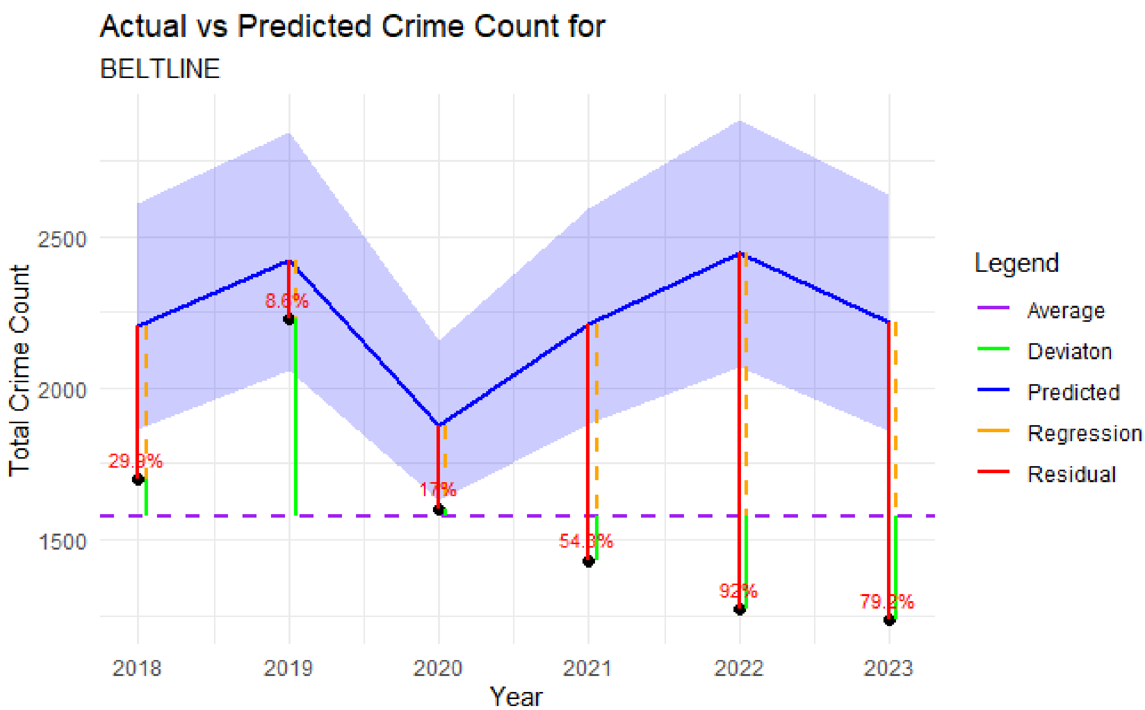


Figure 11 Actual vs Predicted for Beltline

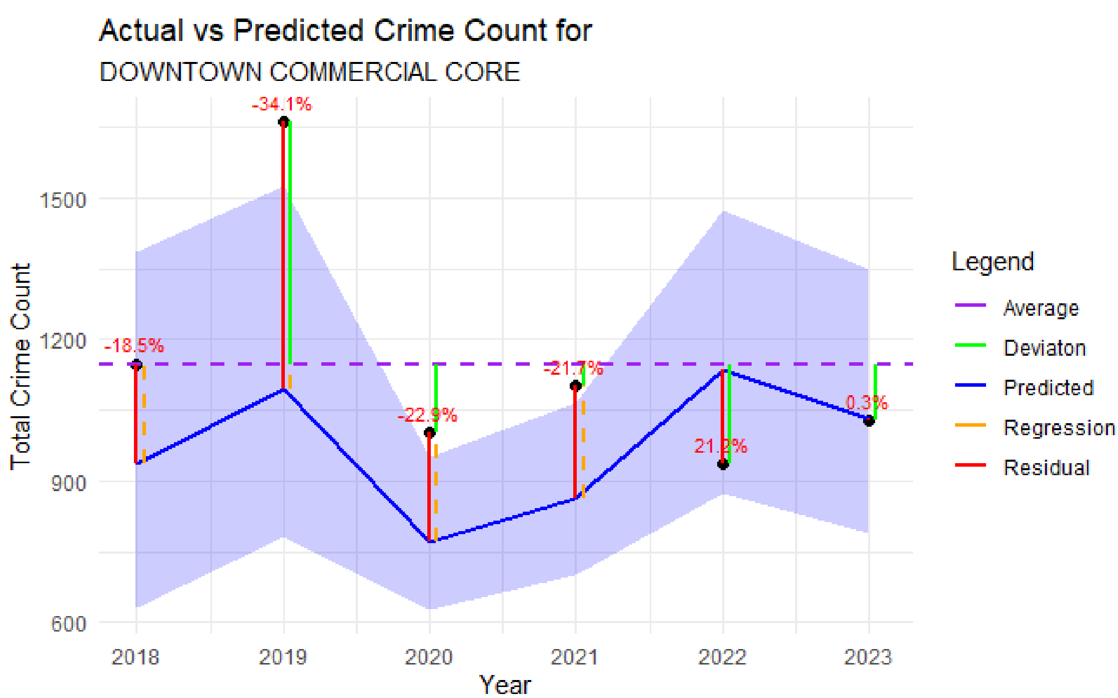


Figure 12 Actual vs Predicted Crime Counts Downtown Core

No change with Green LRT in terms of shortest distance to the central community point for Downtown Commerical Core as the current LRT is already the shortest and most central to the community.

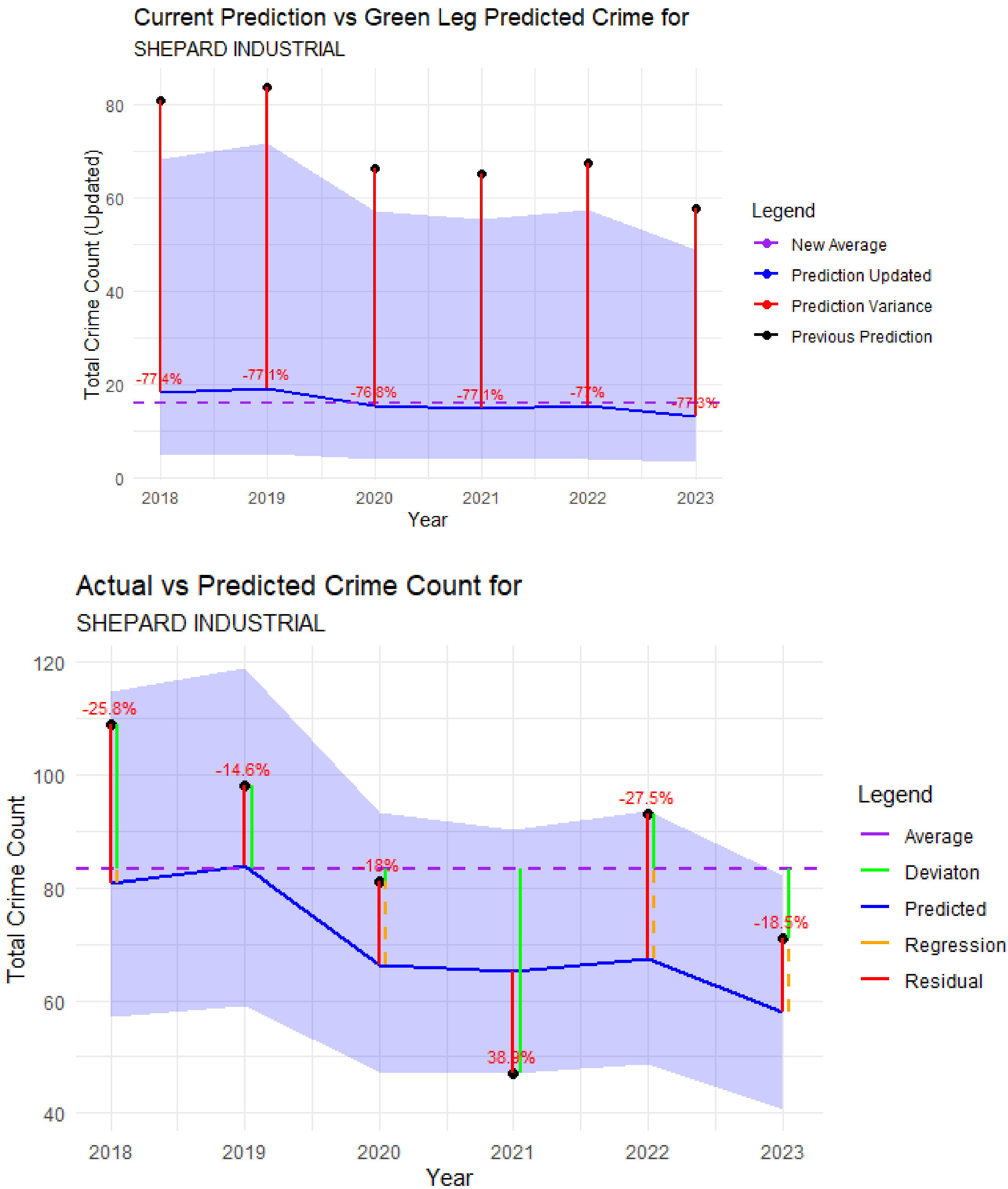


Figure 13 Actual vs Predicted for Shepard Industrial

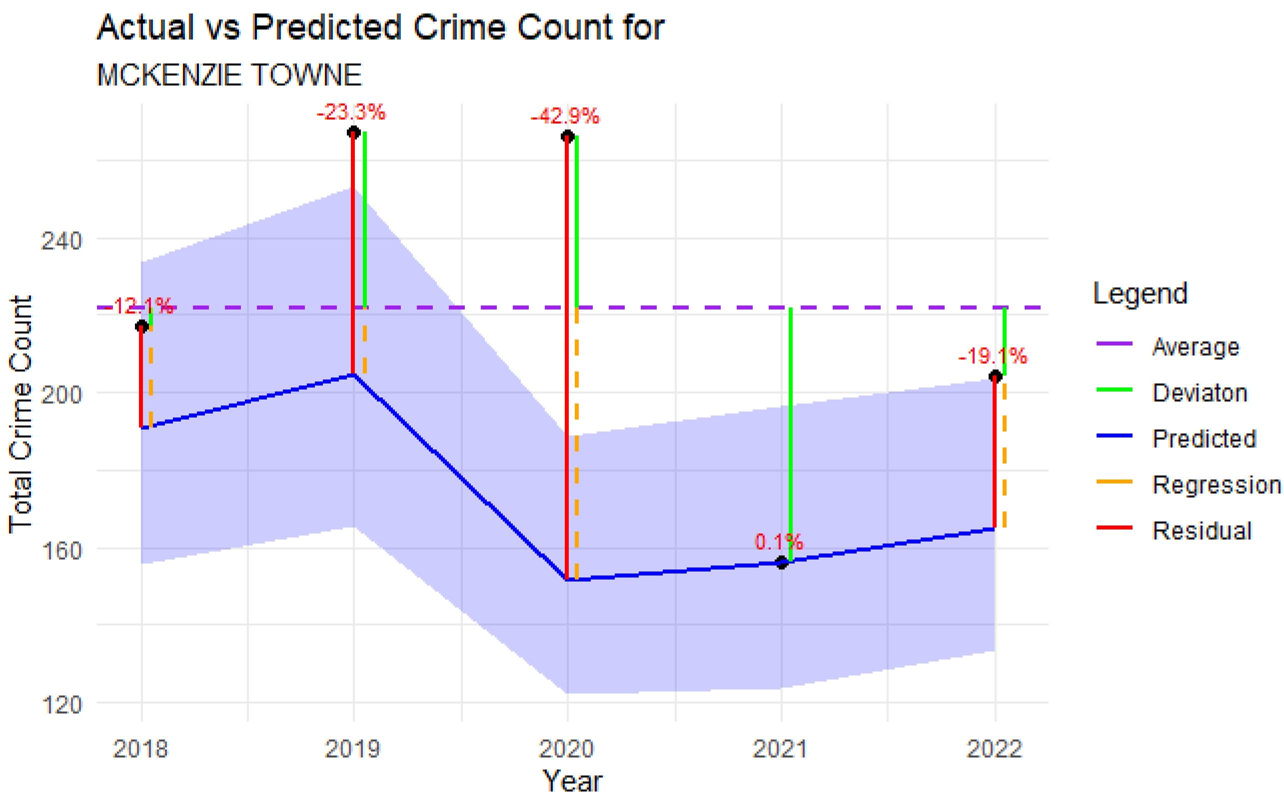
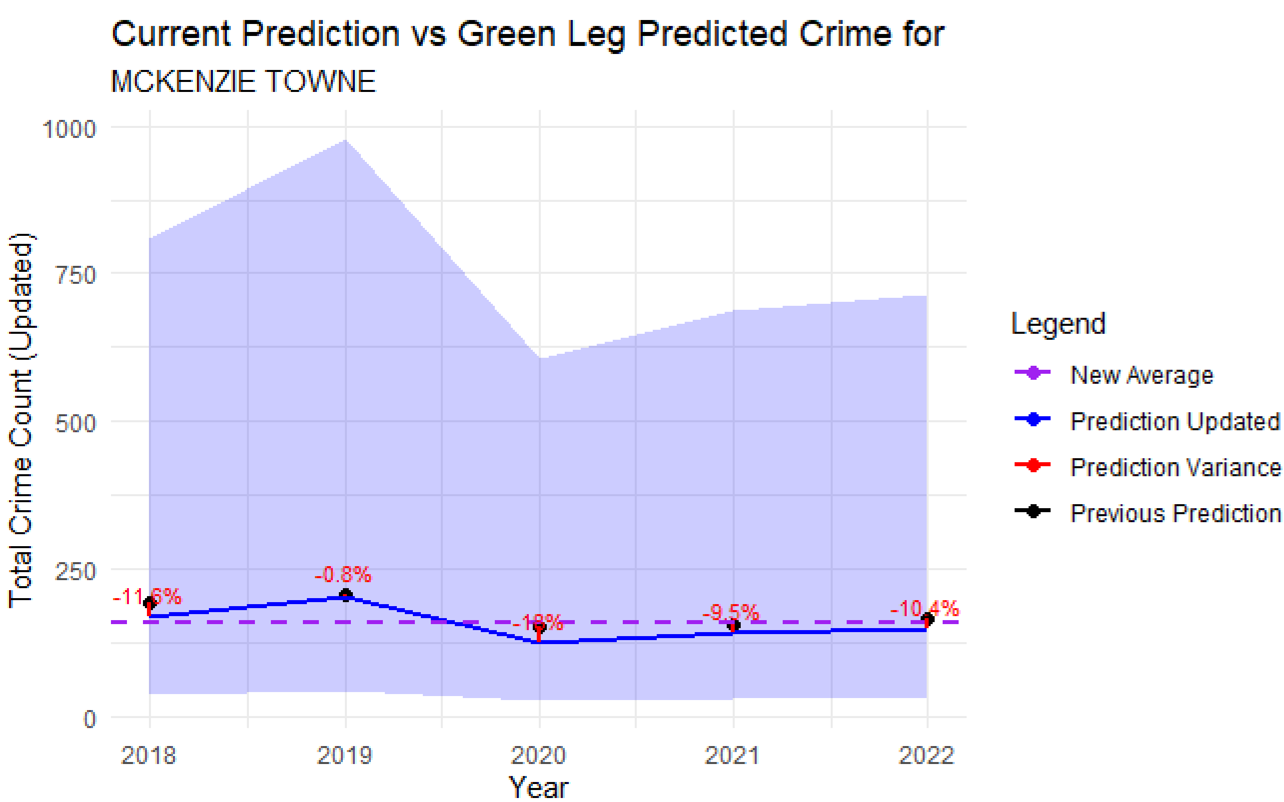


Figure 14 Actual vs Predicted for McKenzie Towne



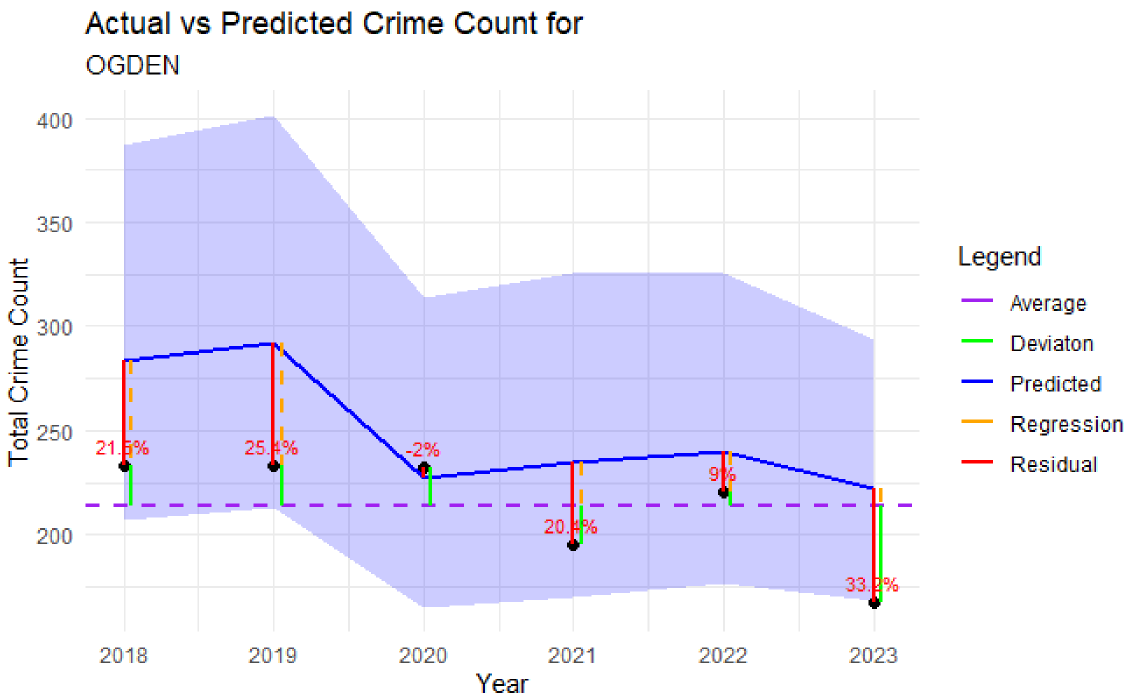
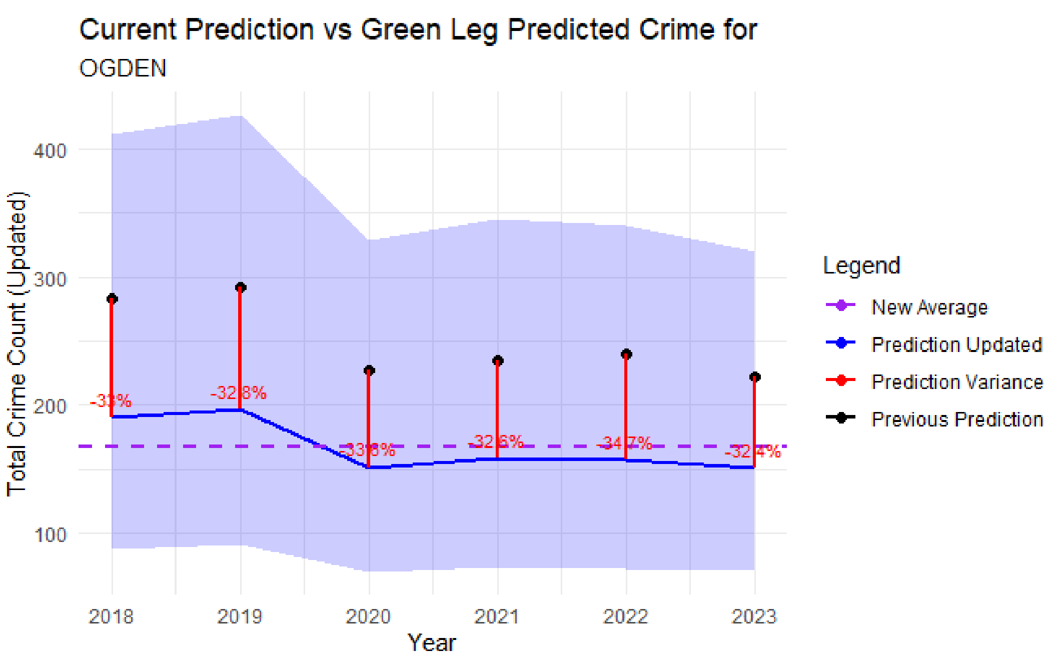


Figure 15 Actual vs Predicted for Ogden



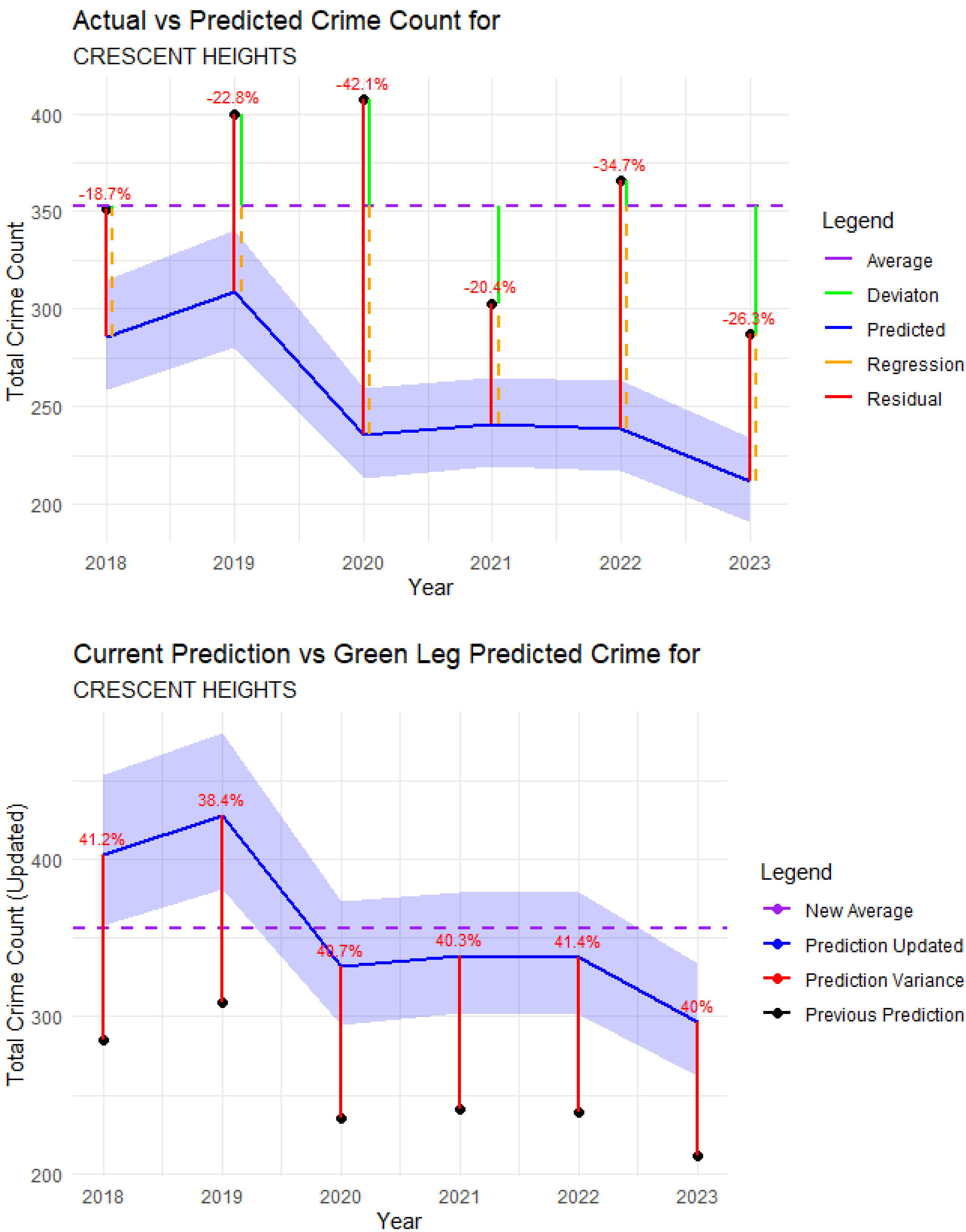


Figure 16 Actual vs Predicted for Crescent Heights

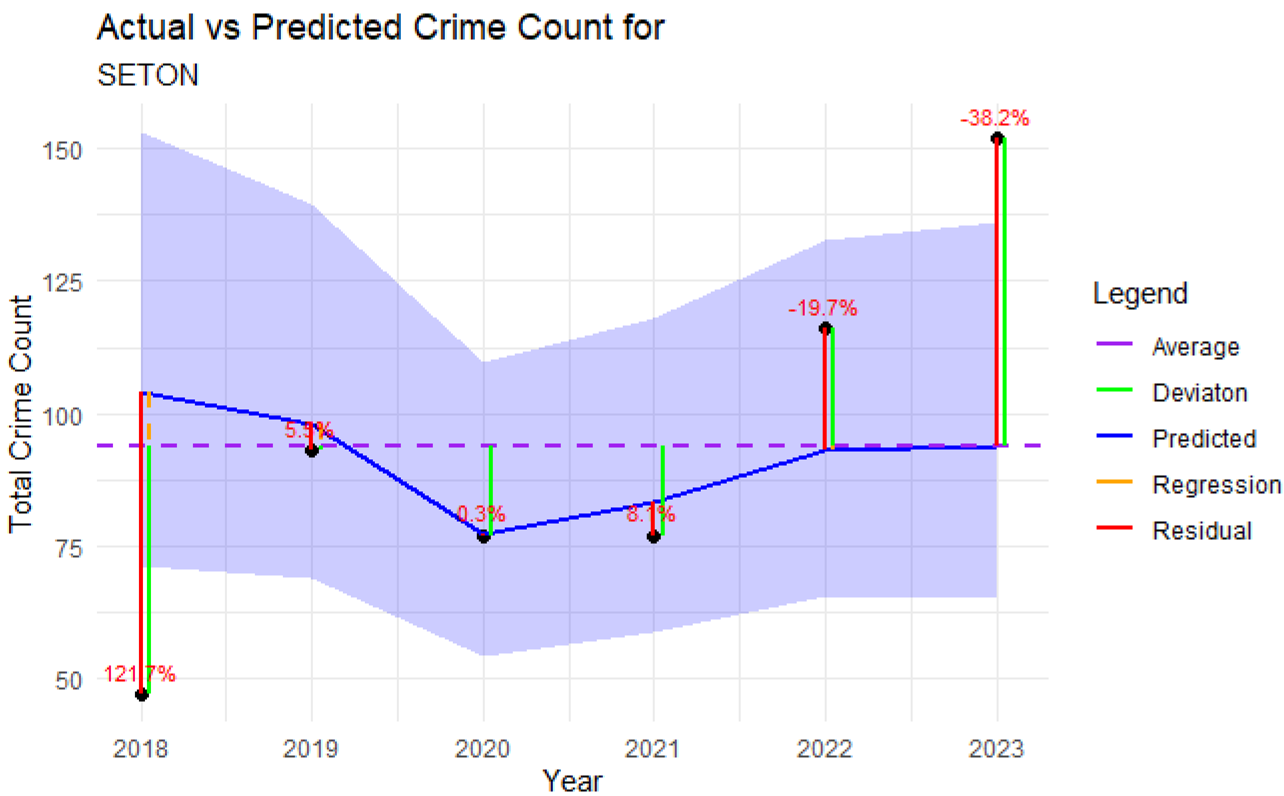
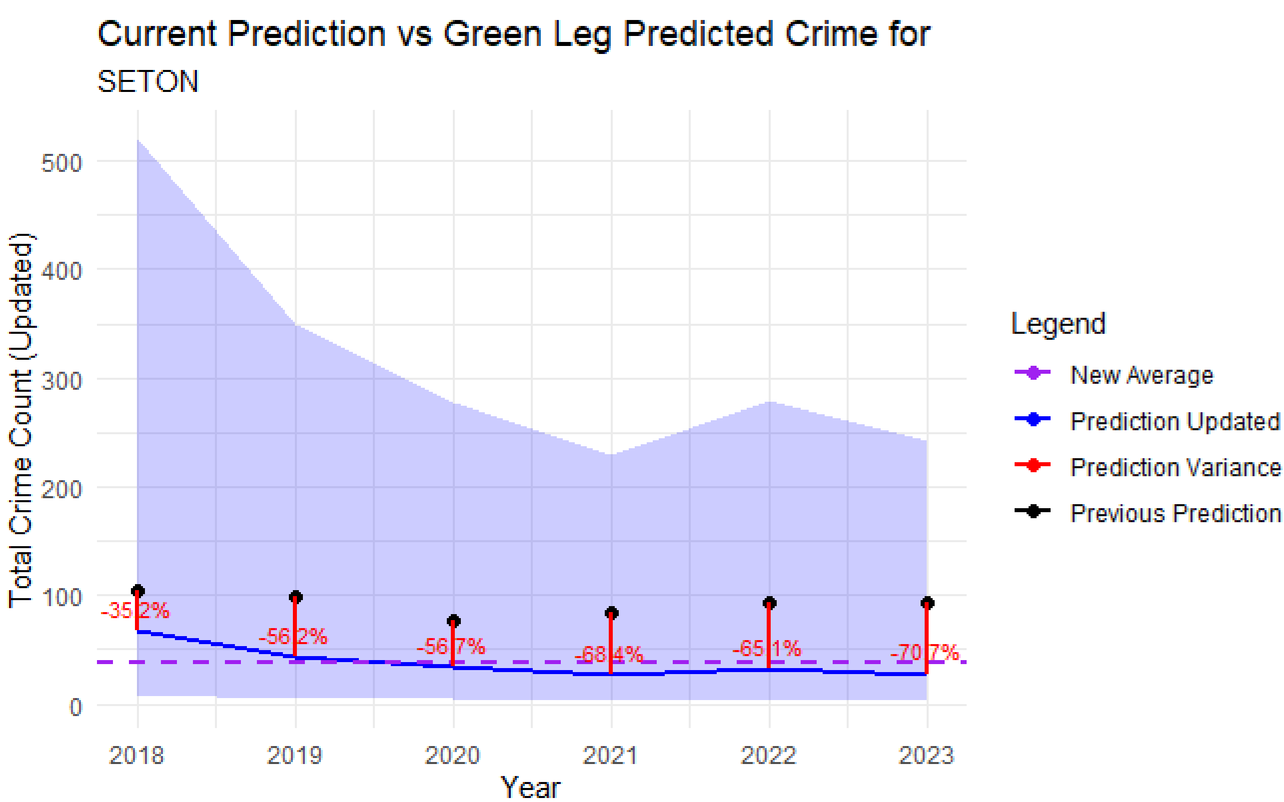


Figure 17 Actual vs Predicted for Seton



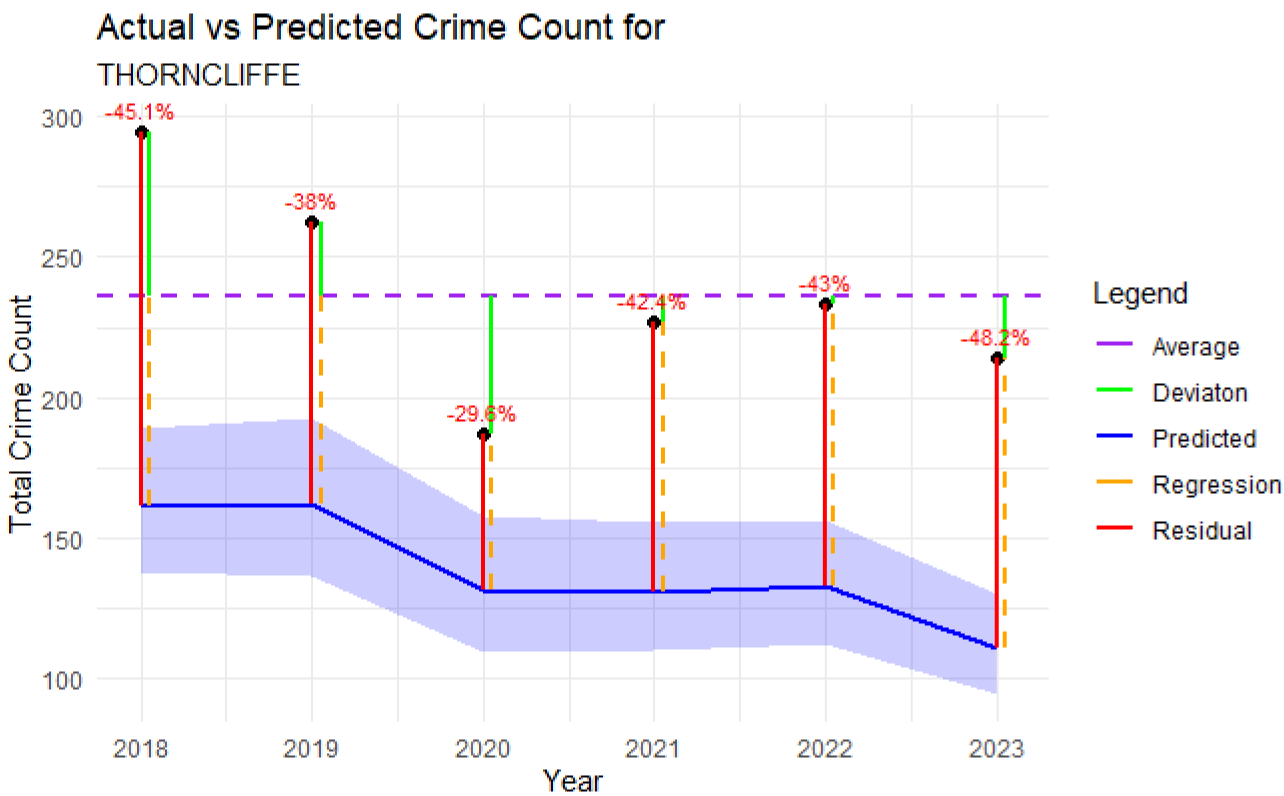
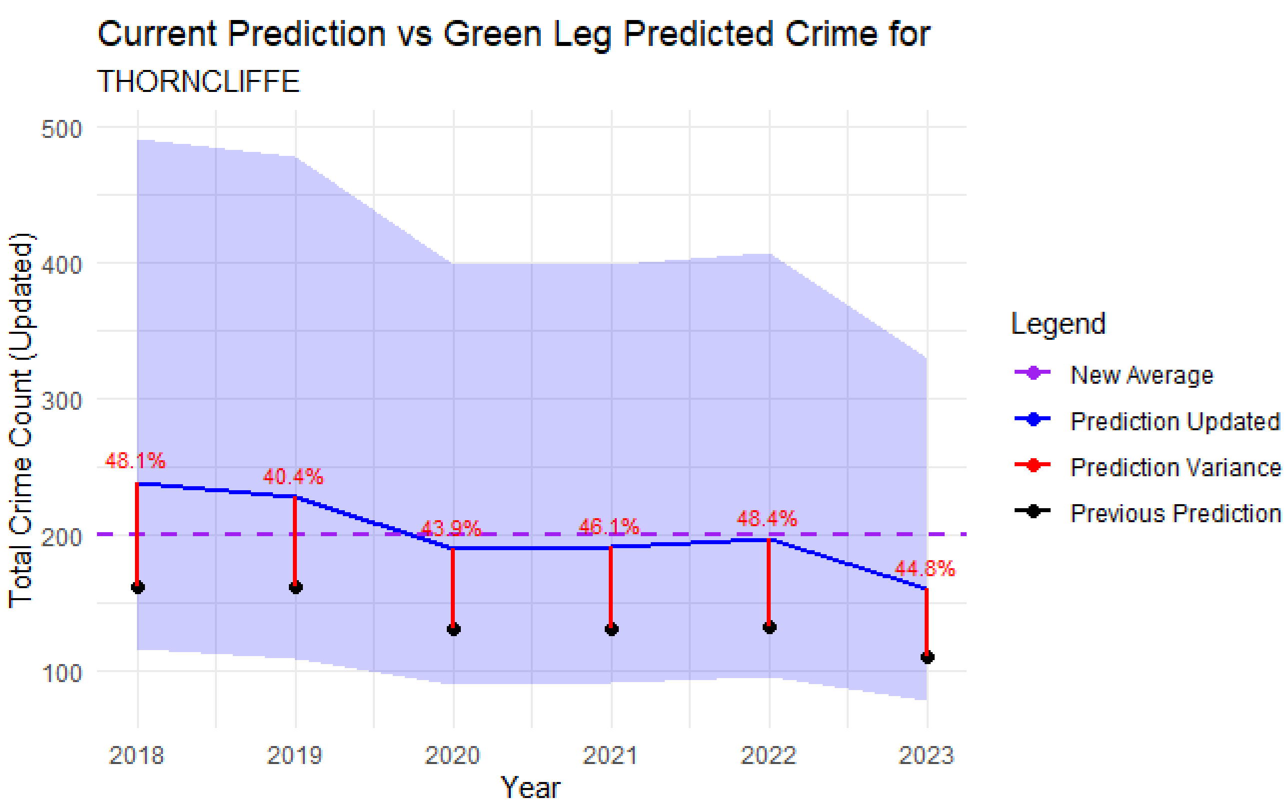


Figure 18 Actual vs Predicted for Thorncliffe



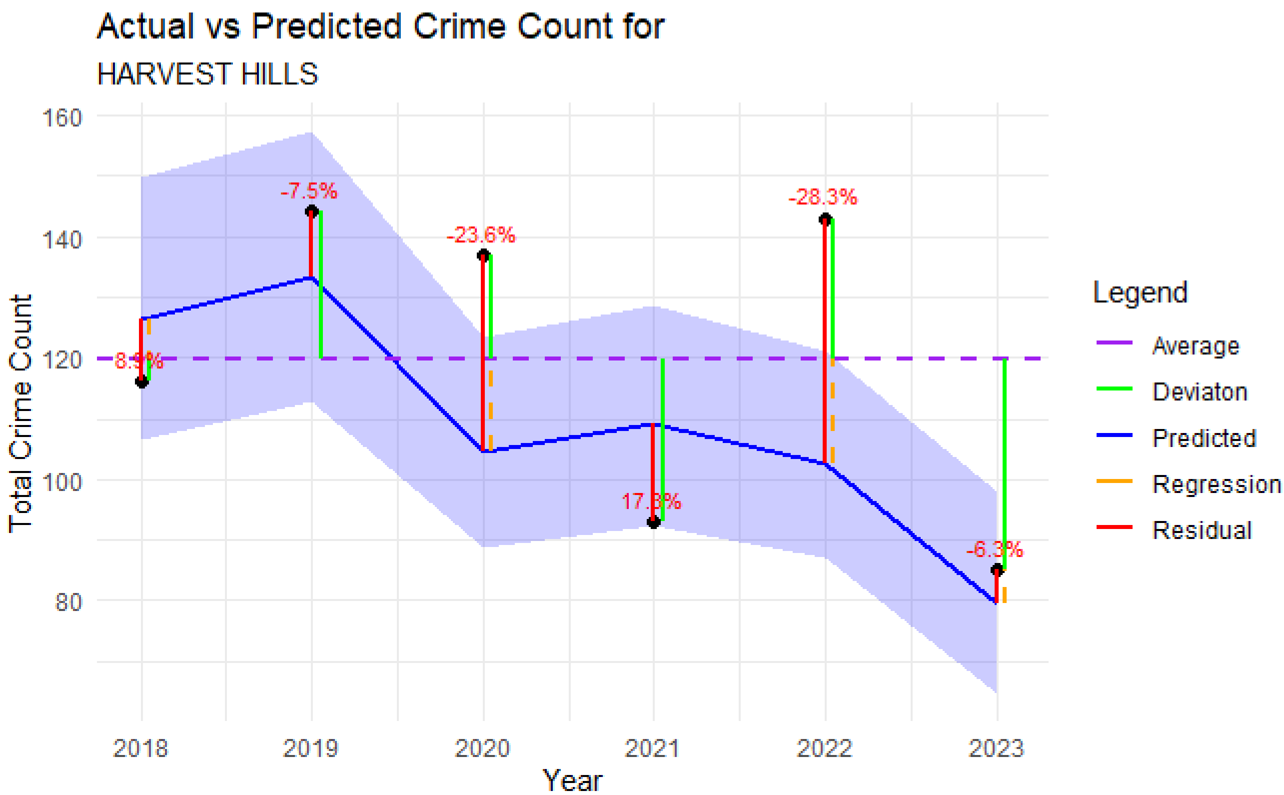
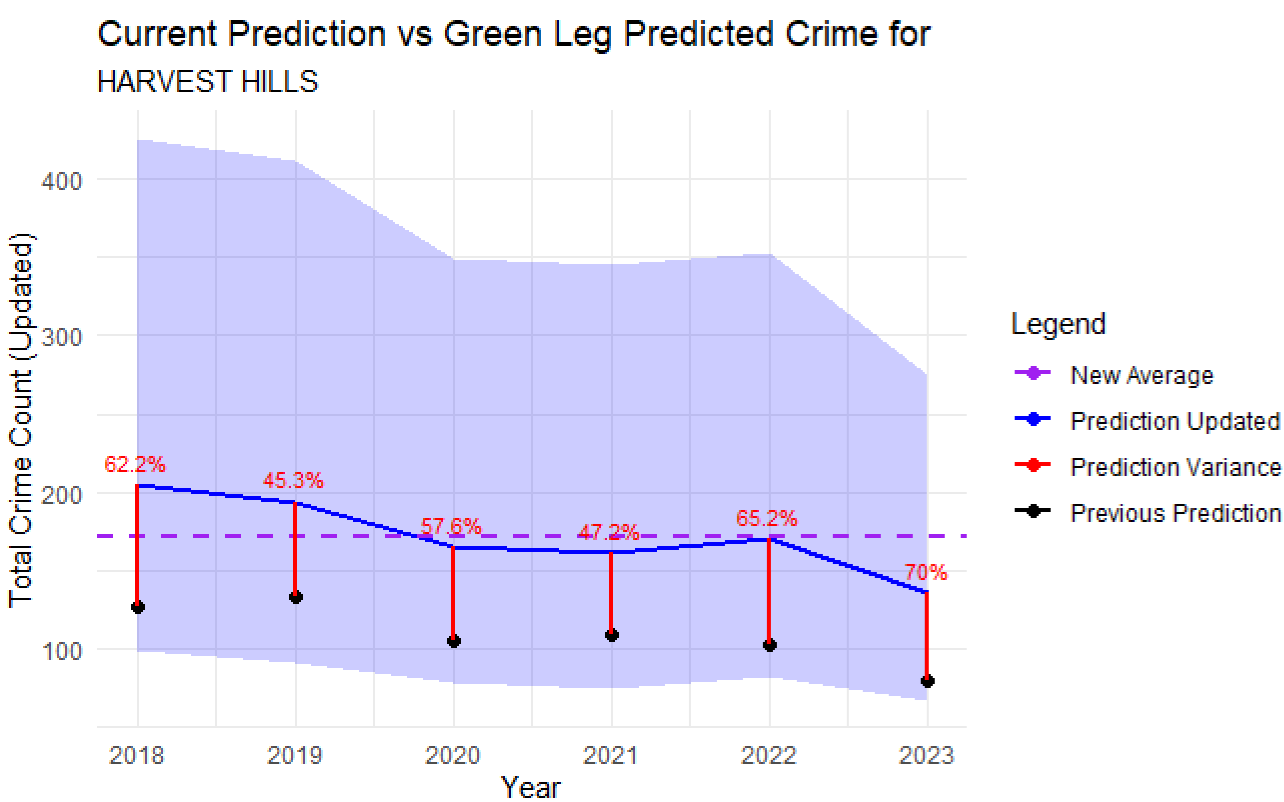


Figure 19 Actual vs Predicted for Harvest Hills



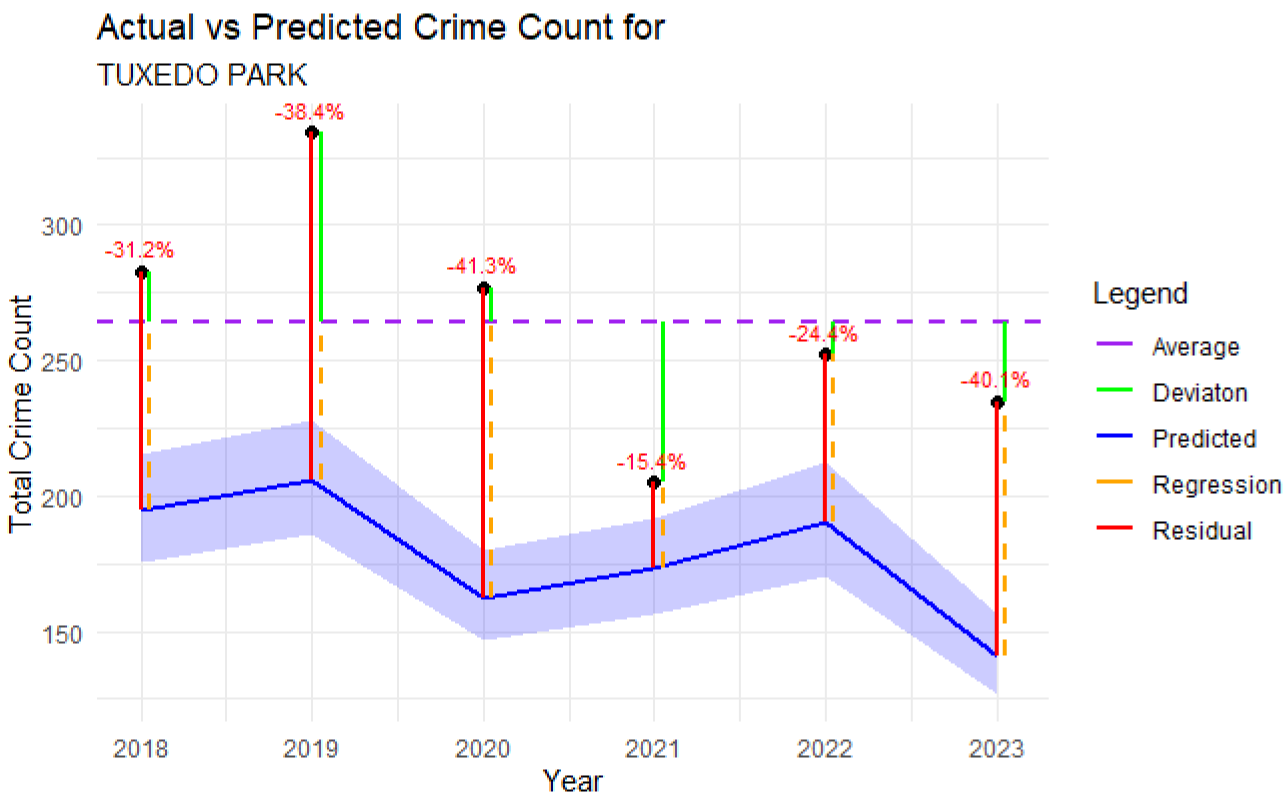
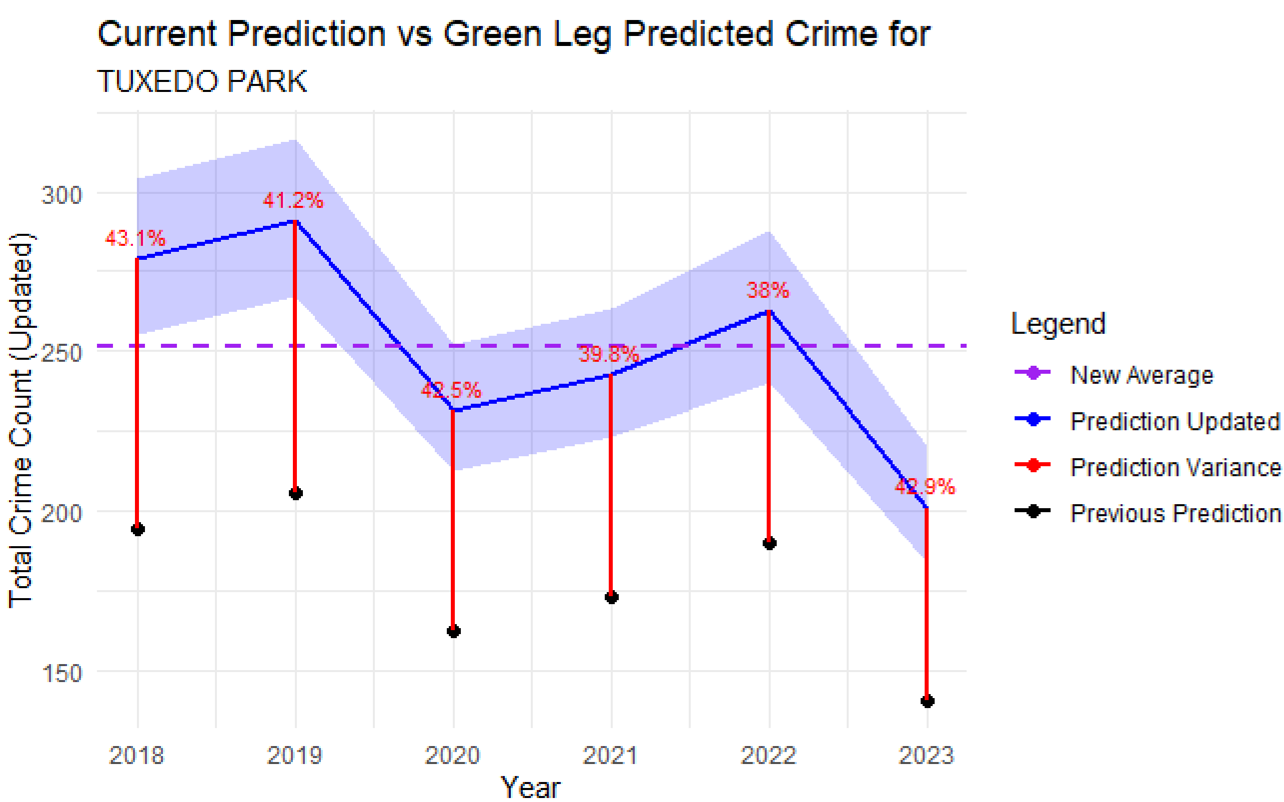


Figure 20 Actual vs Predicted for Tuxedo Park



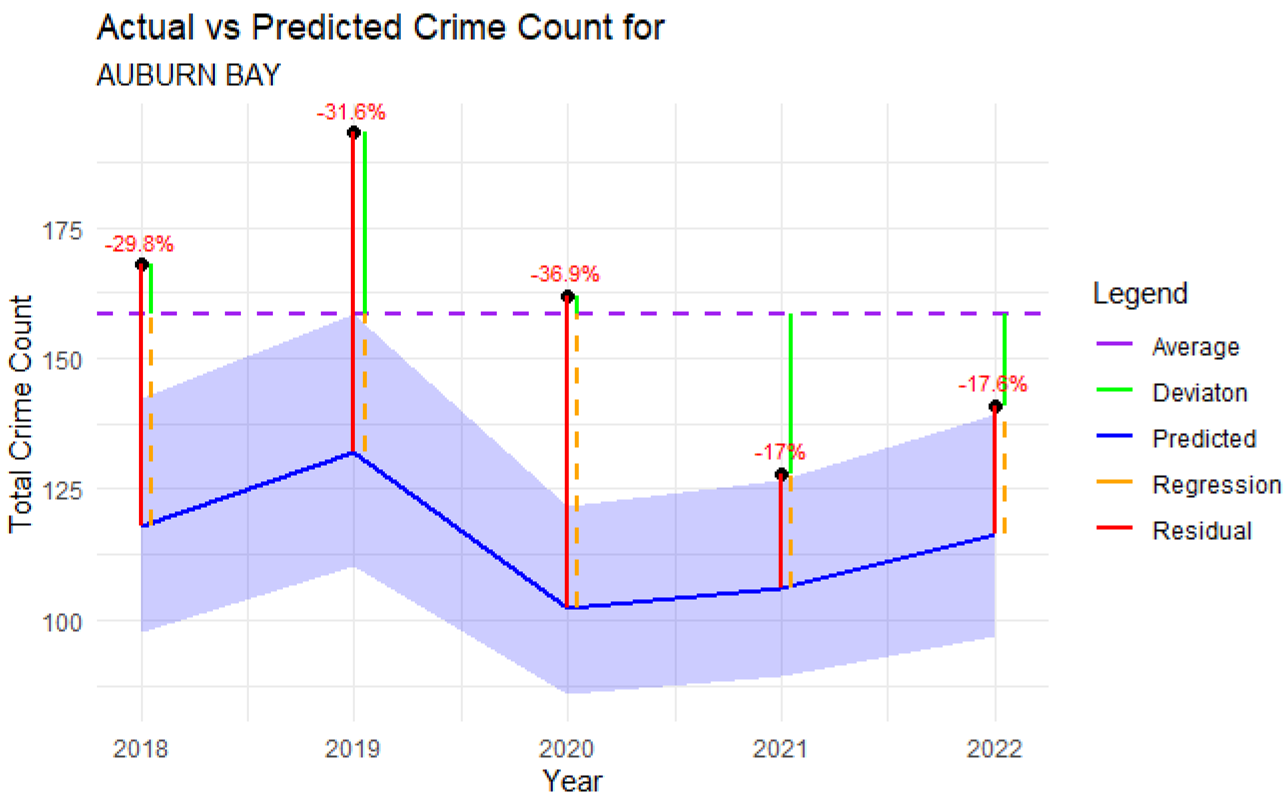
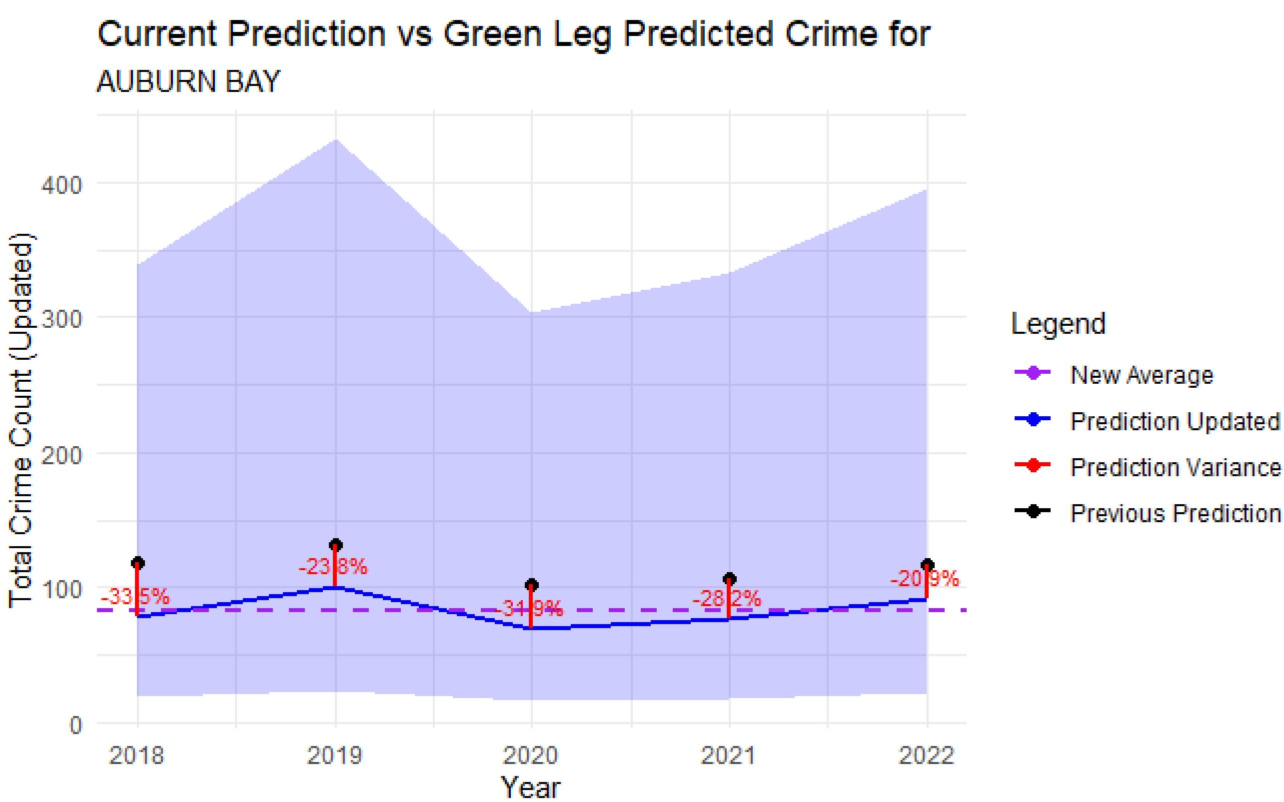


Figure 21 Actual vs Predicted for Auburn Bay



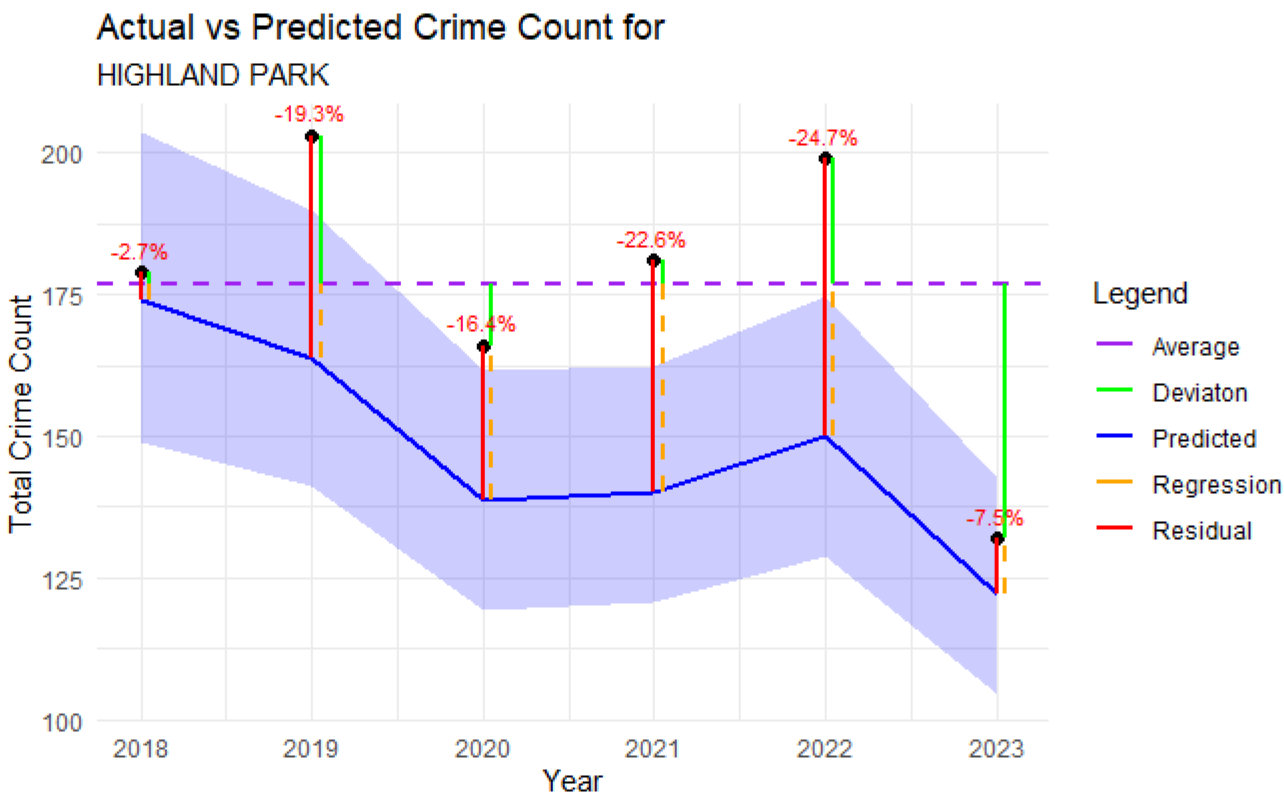
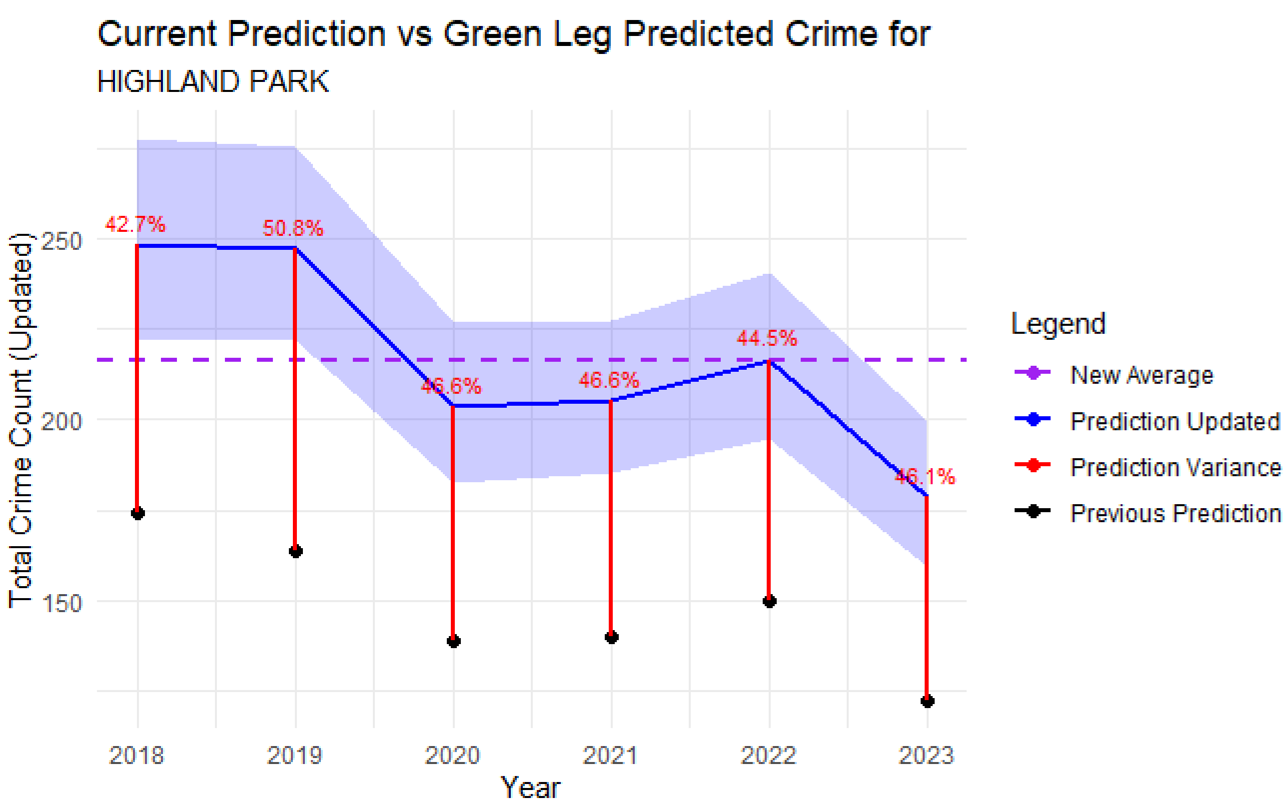


Figure 22 Actual vs Predicted for Highland Park



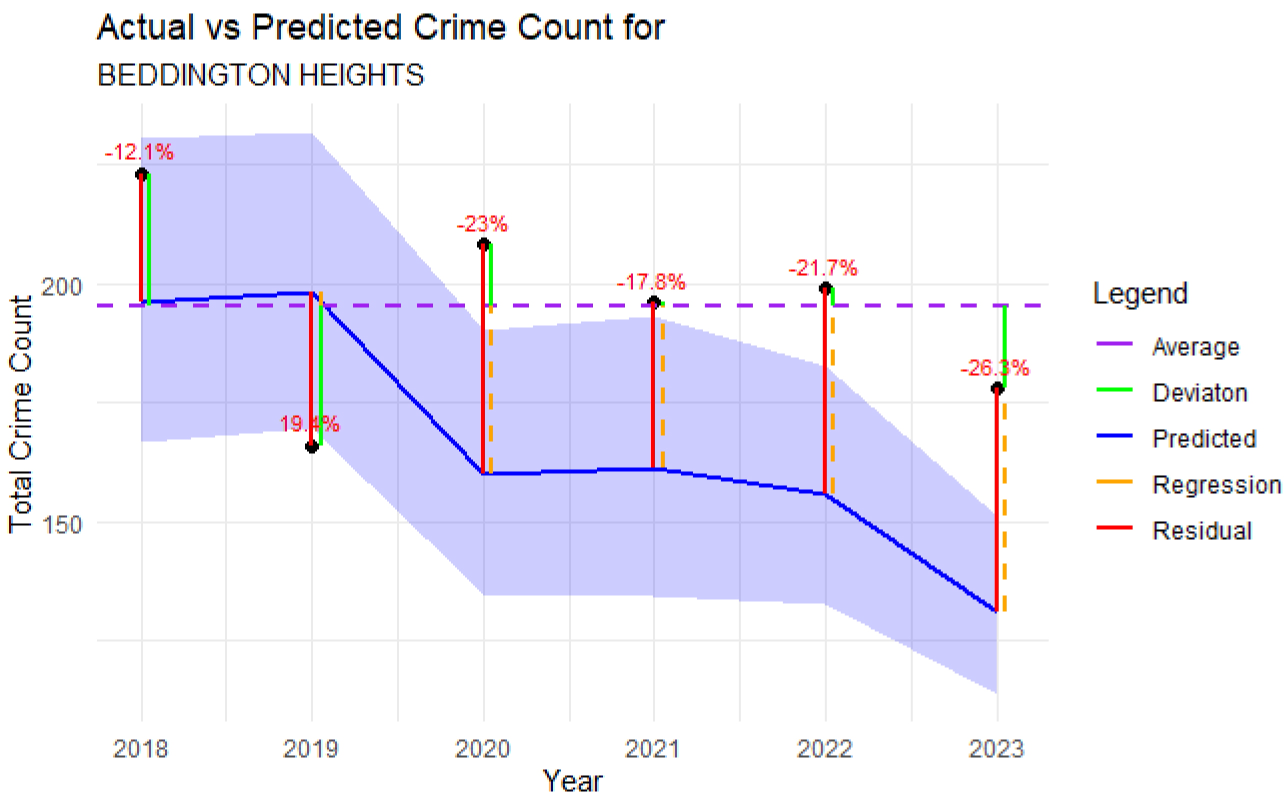
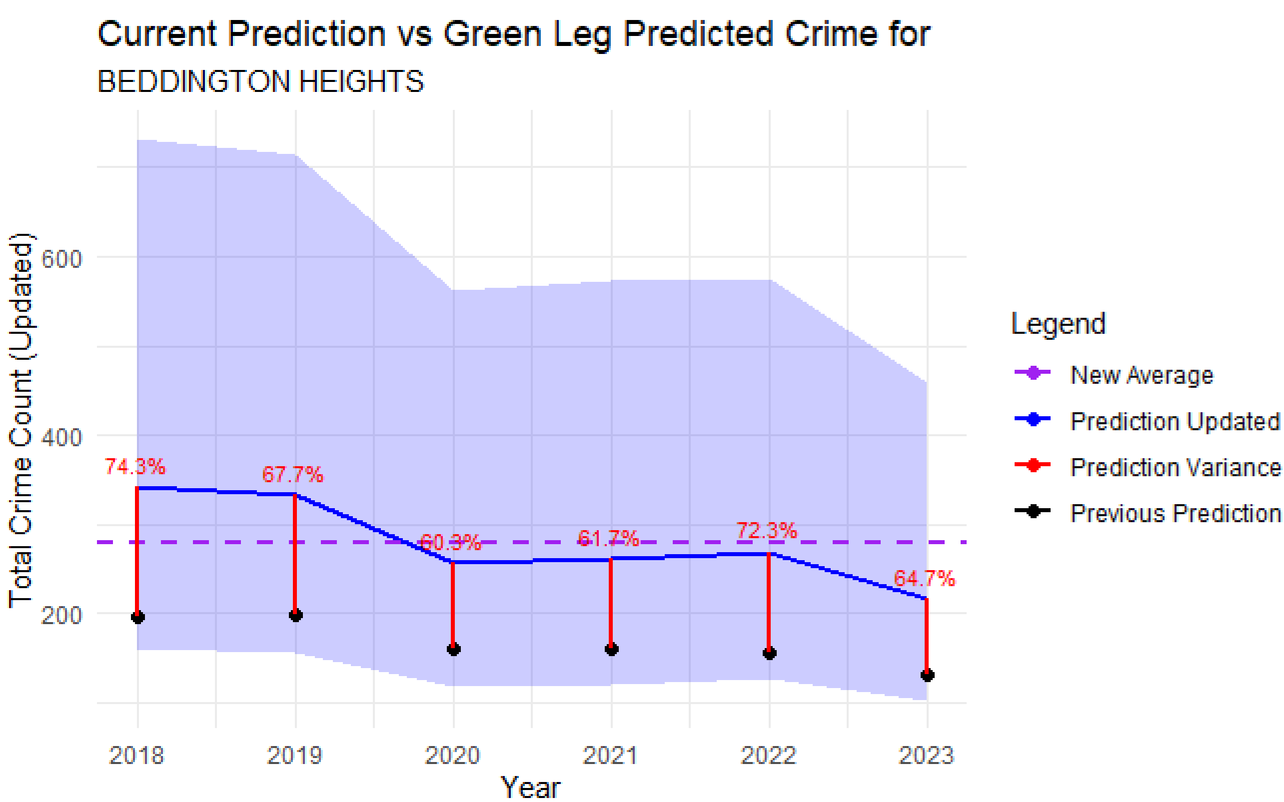


Figure 23 Actual vs Predicted for Beddington Heights



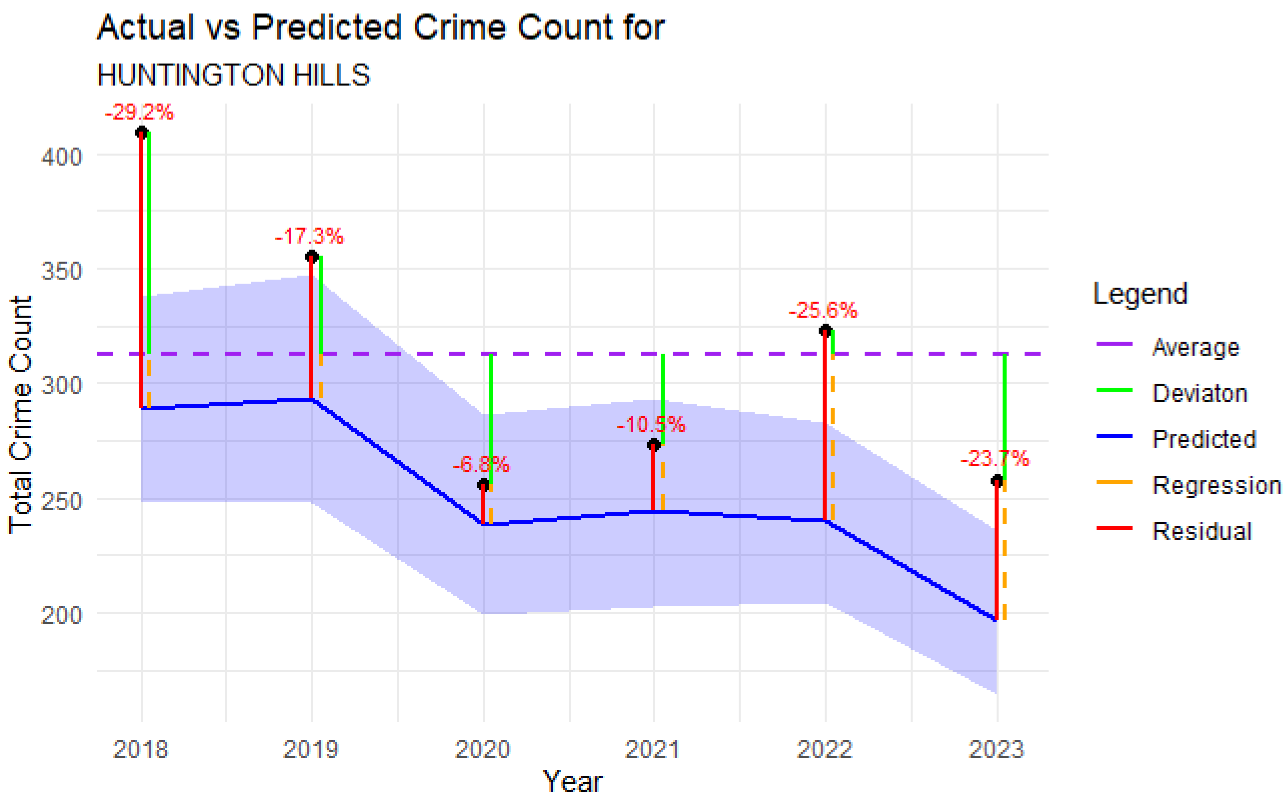
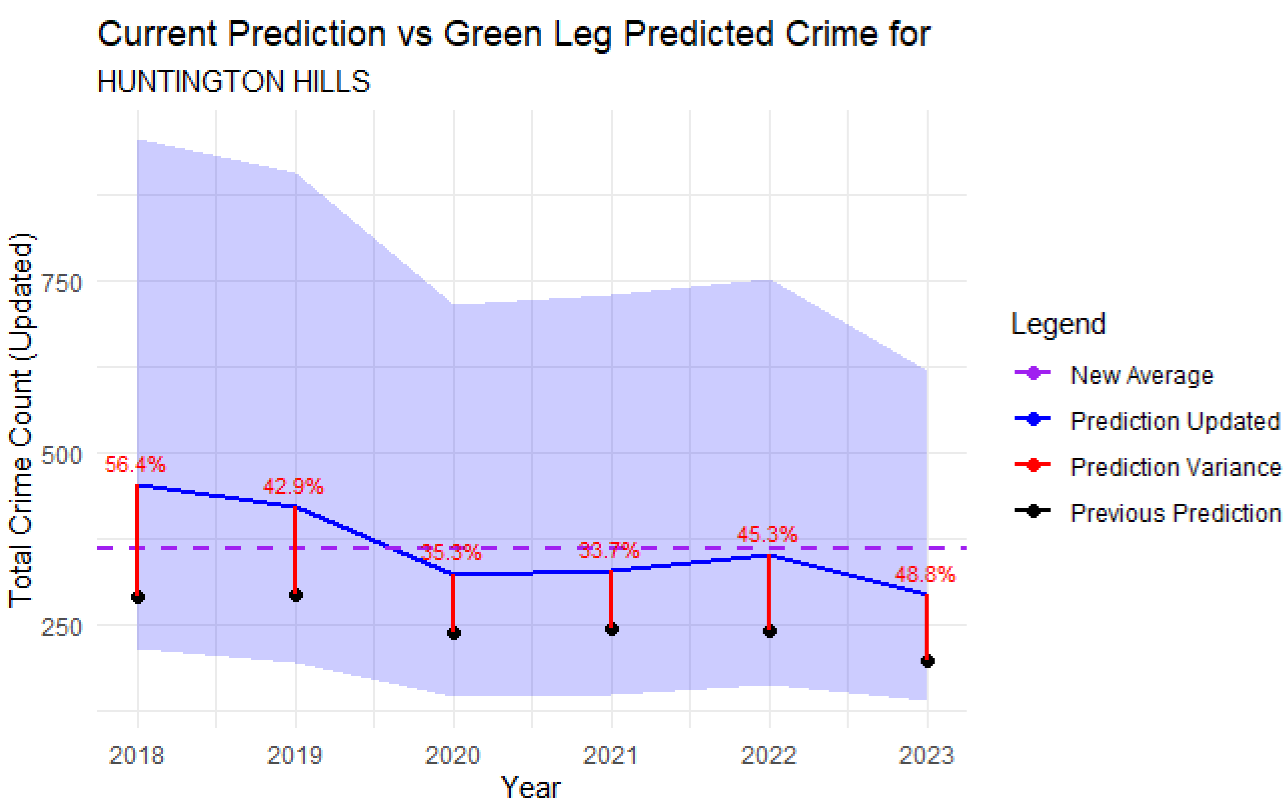


Figure 24 Actual vs Predicted for Huntington Hills



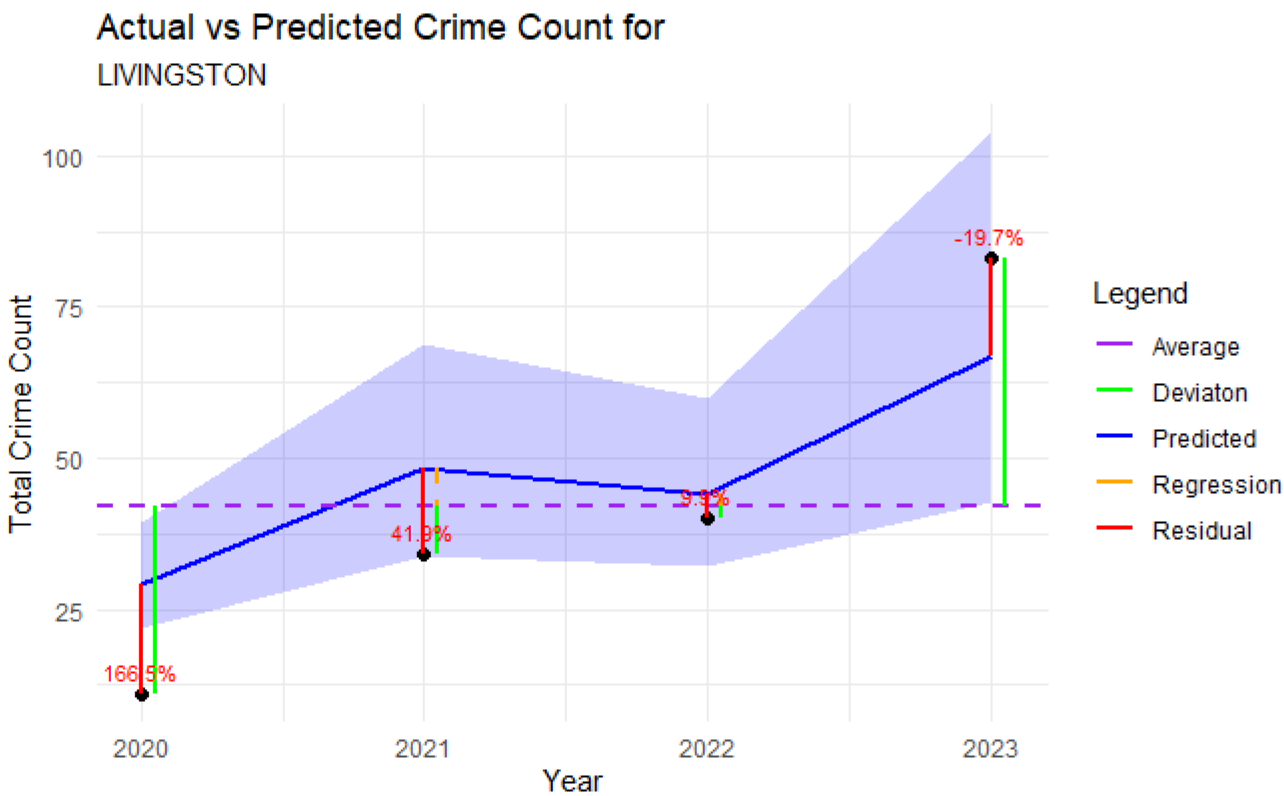
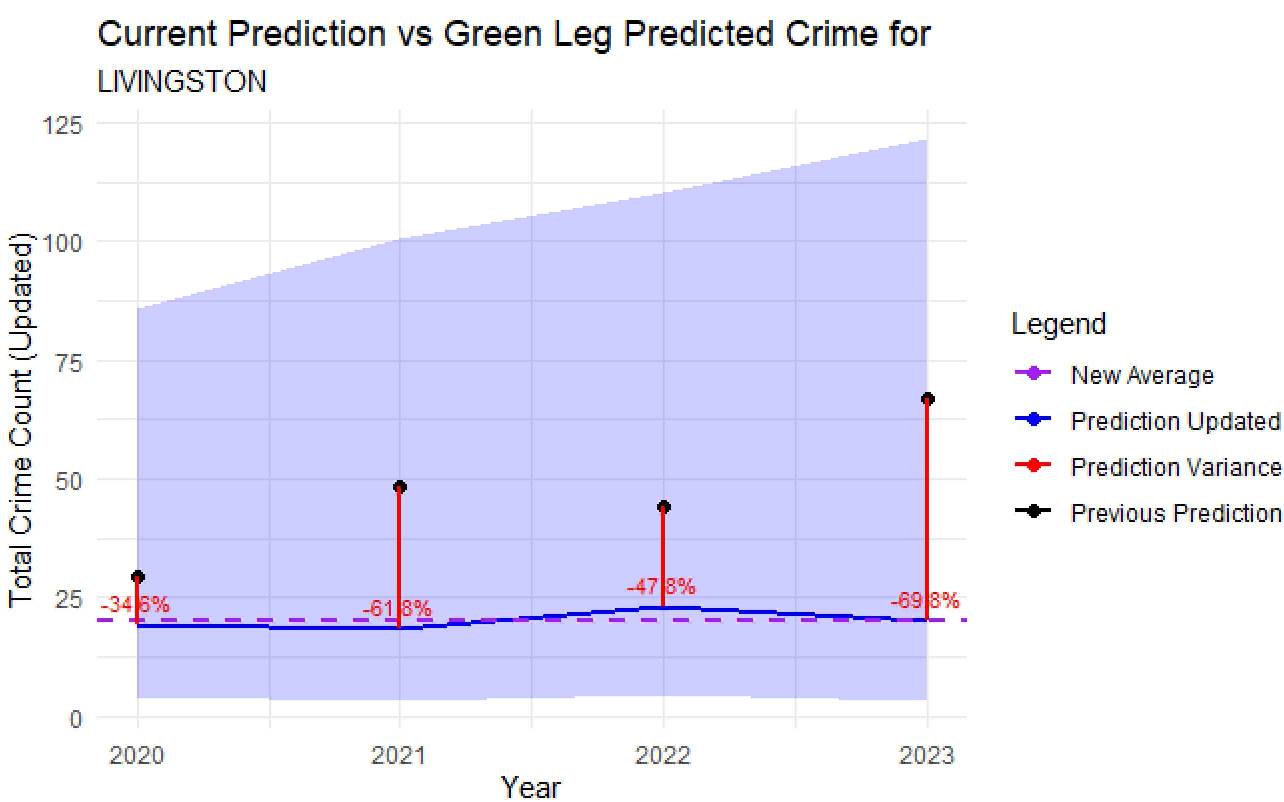


Figure 25 Actual vs Predicted for Livingston



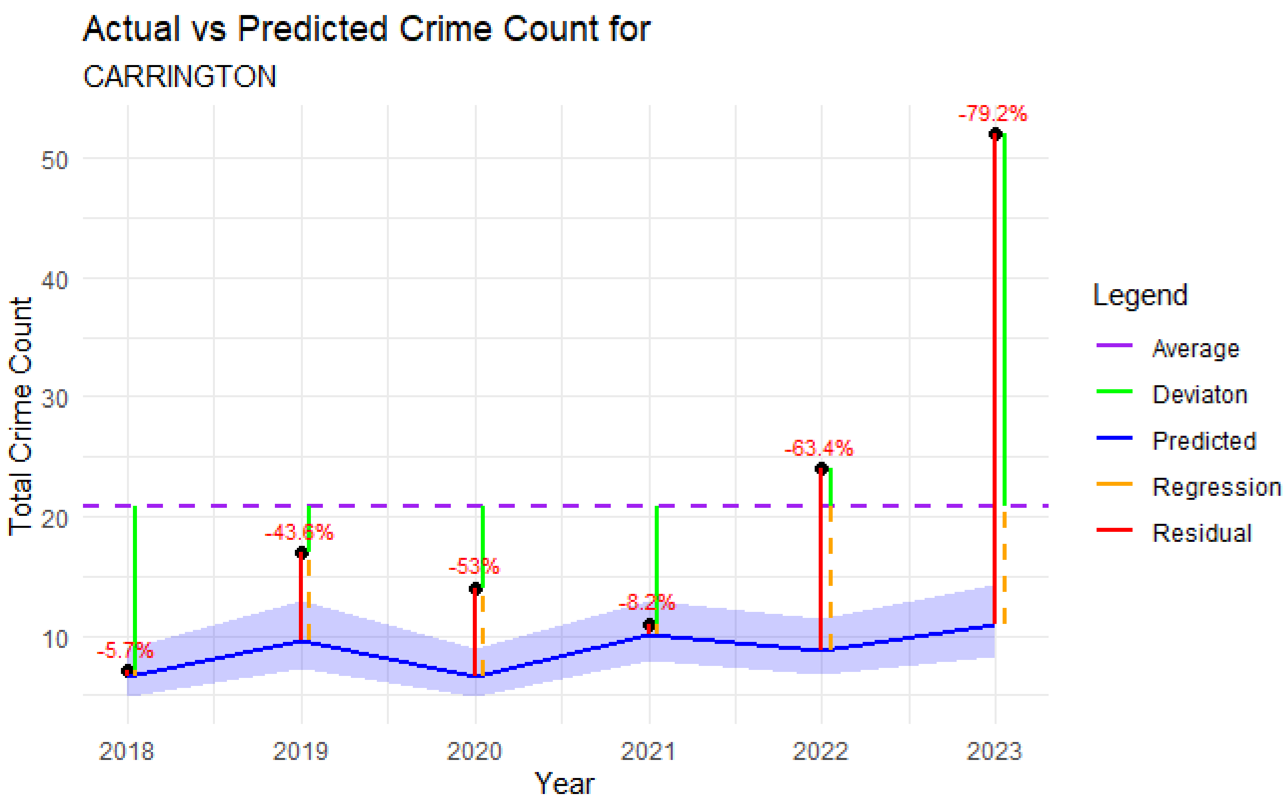
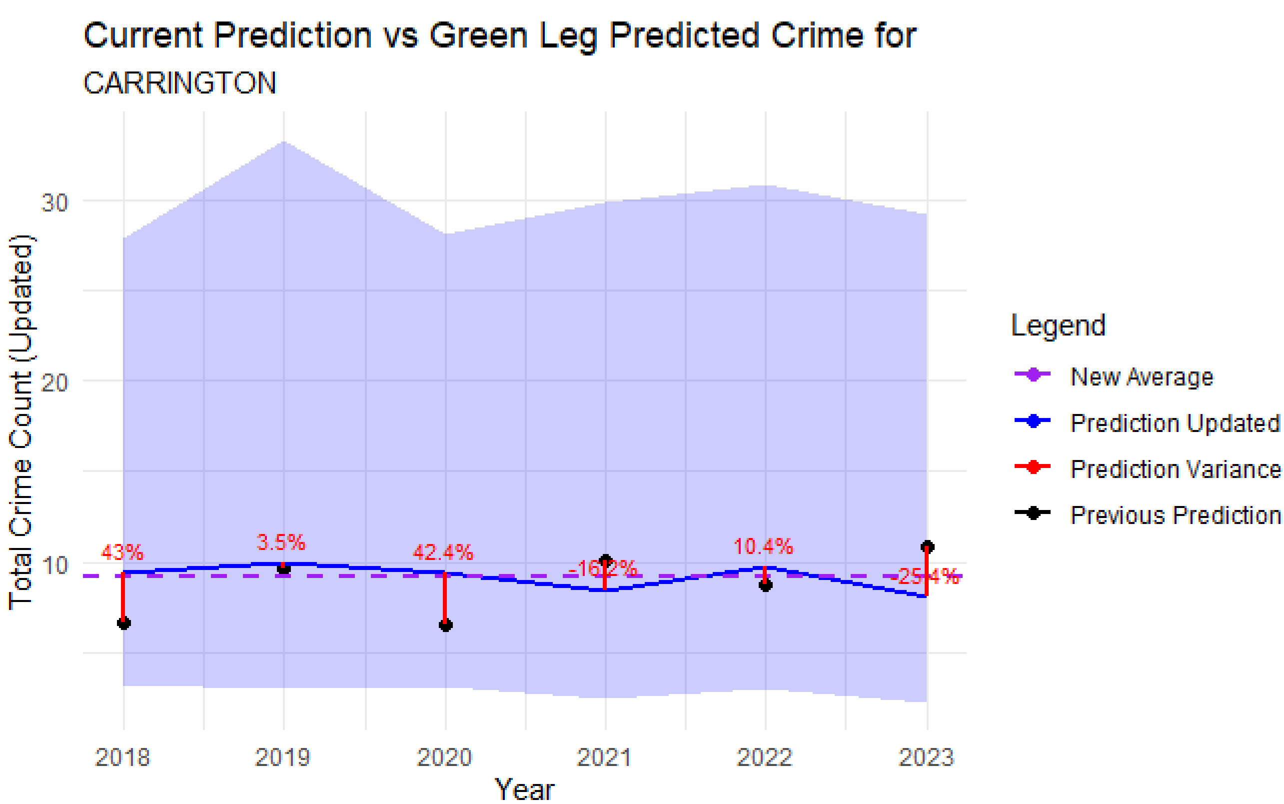


Figure 26 Actual vs Predicted for Carrington



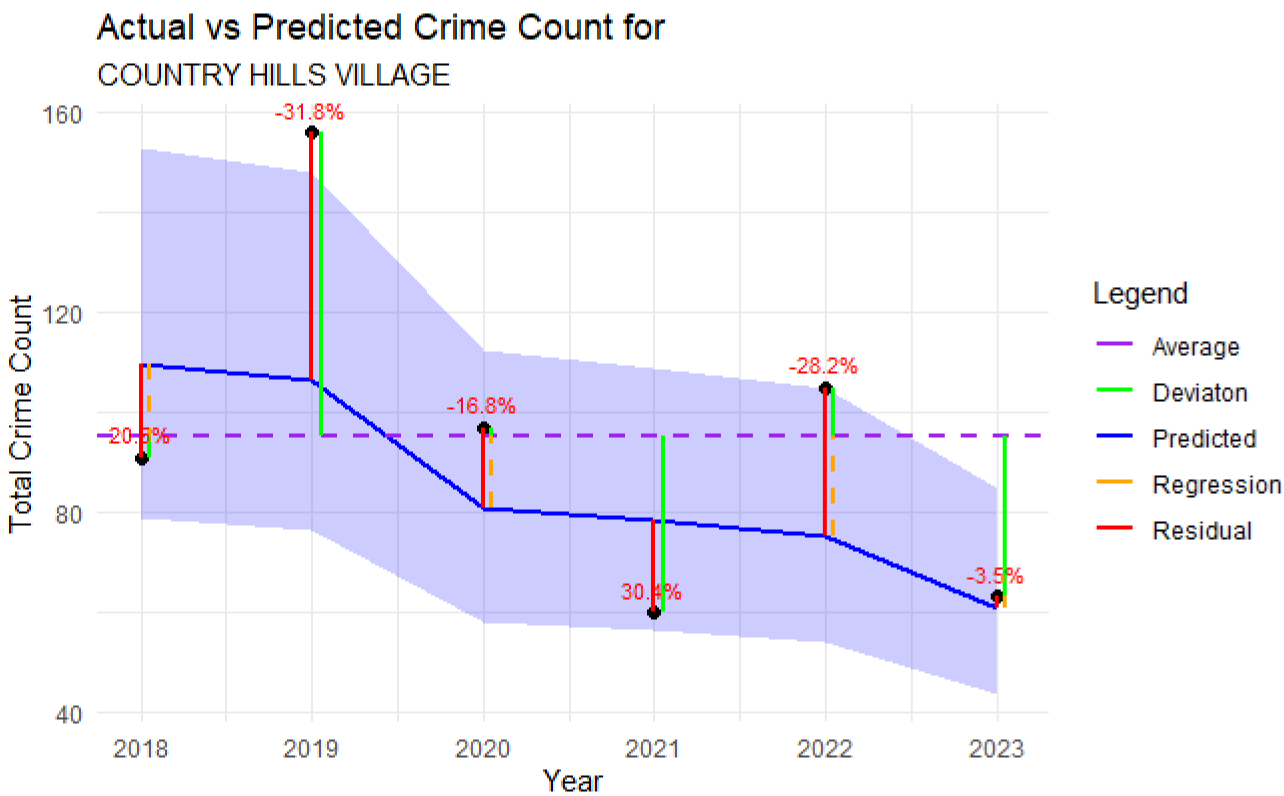
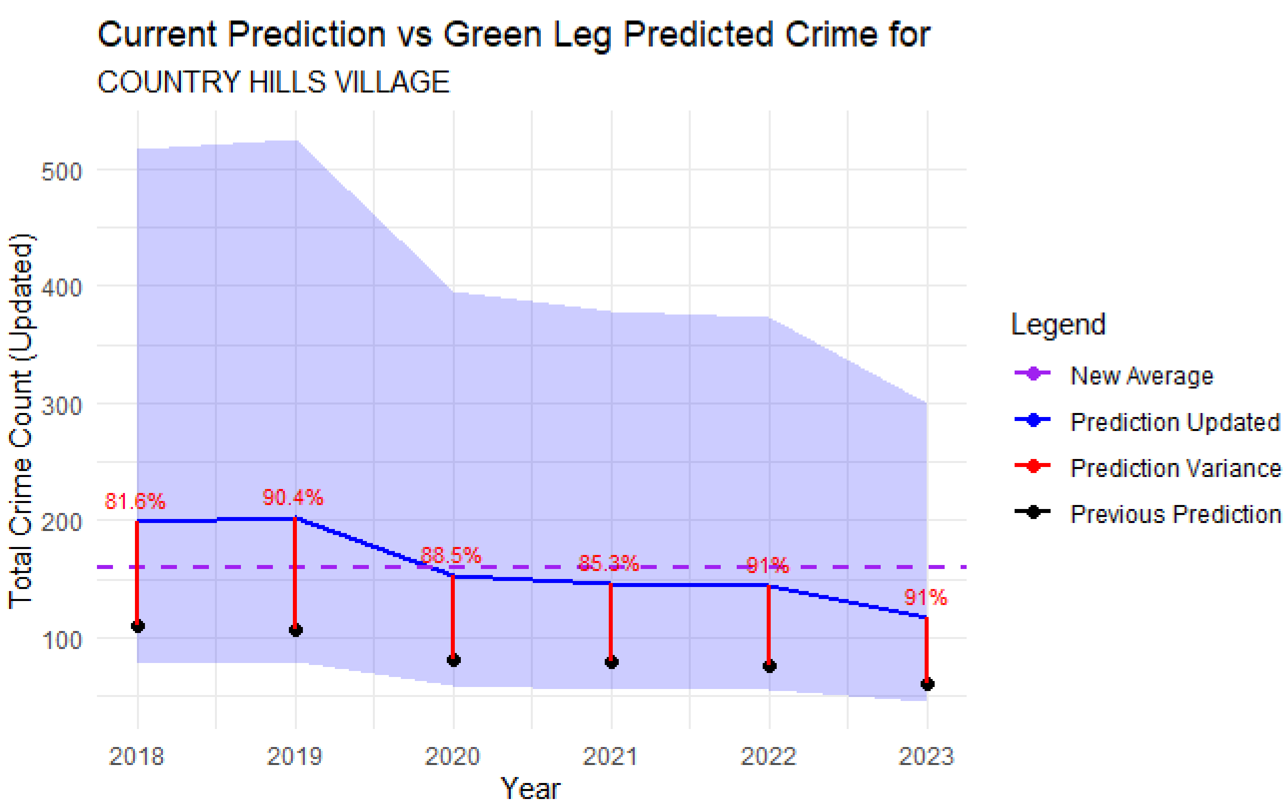


Figure 27 Actual vs Predicted for Country Hills Village



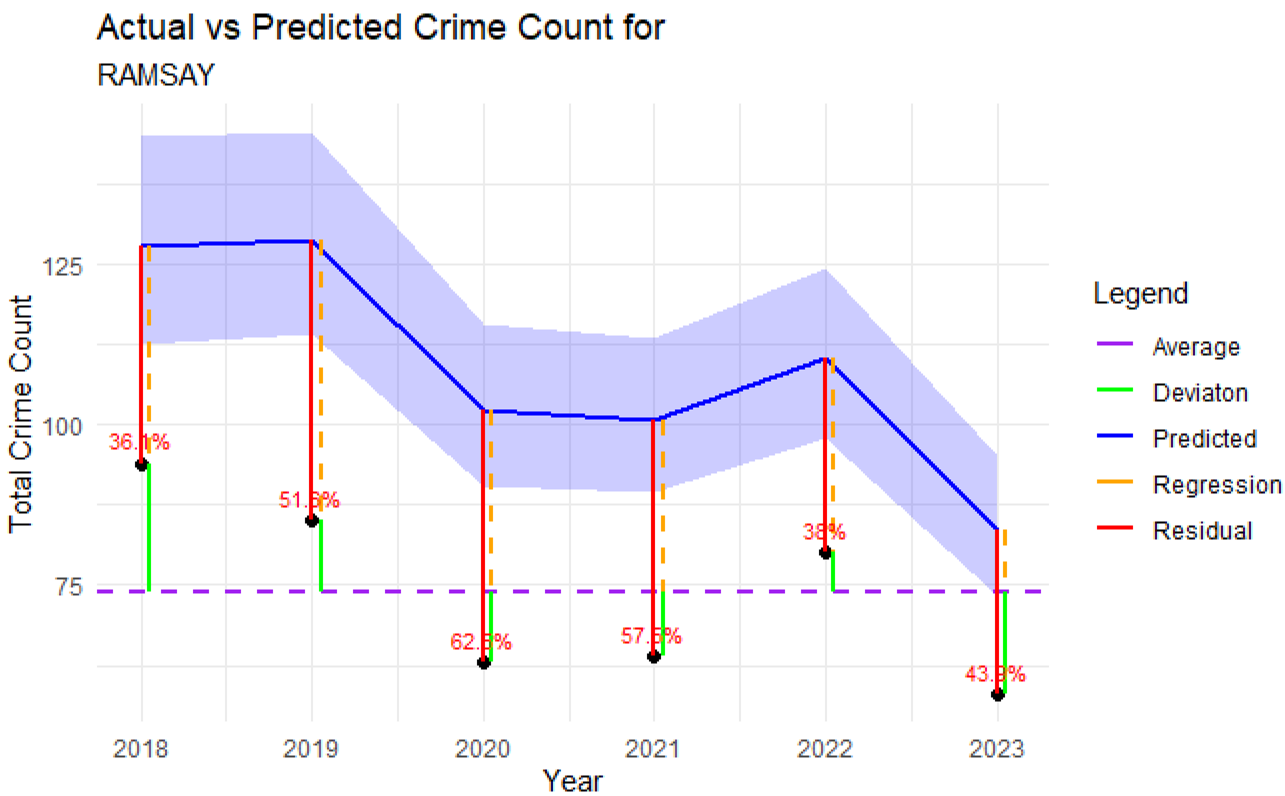
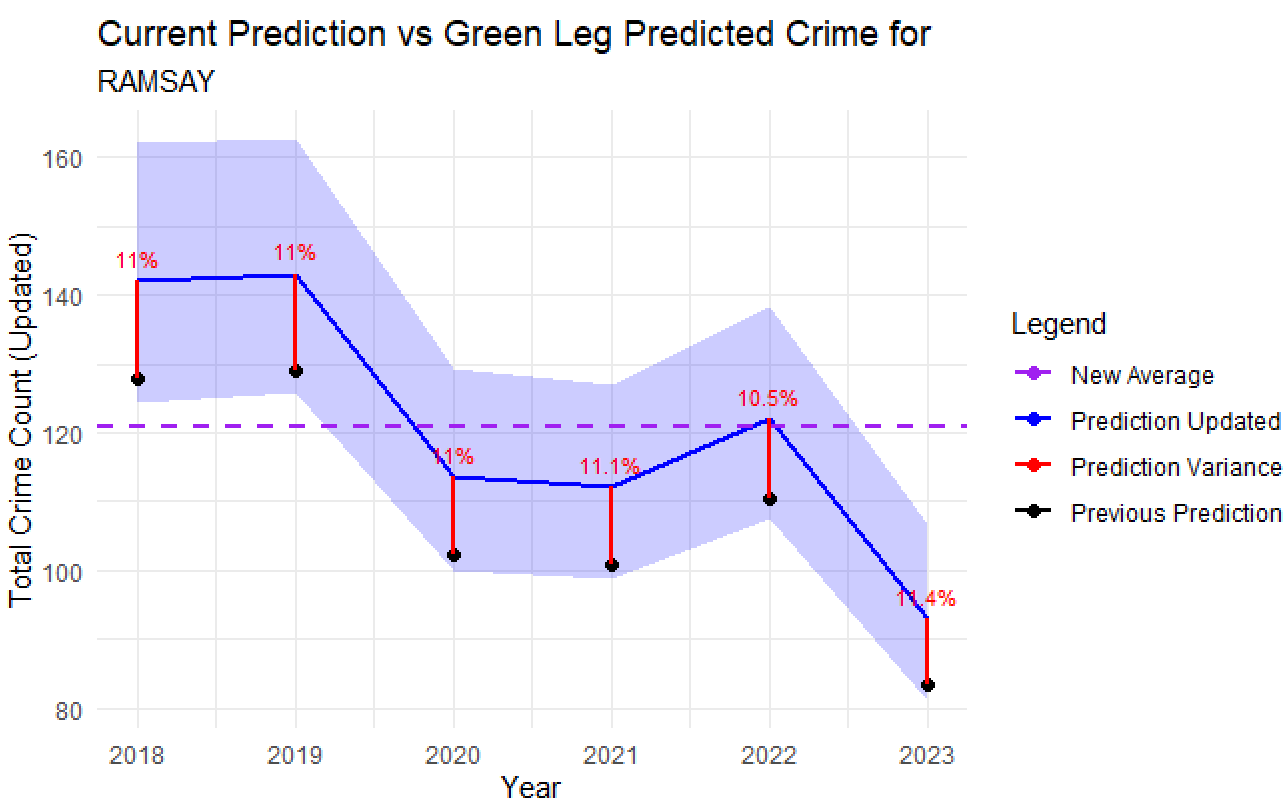


Figure 28 Actual vs Predicted Crime Count Ramsay



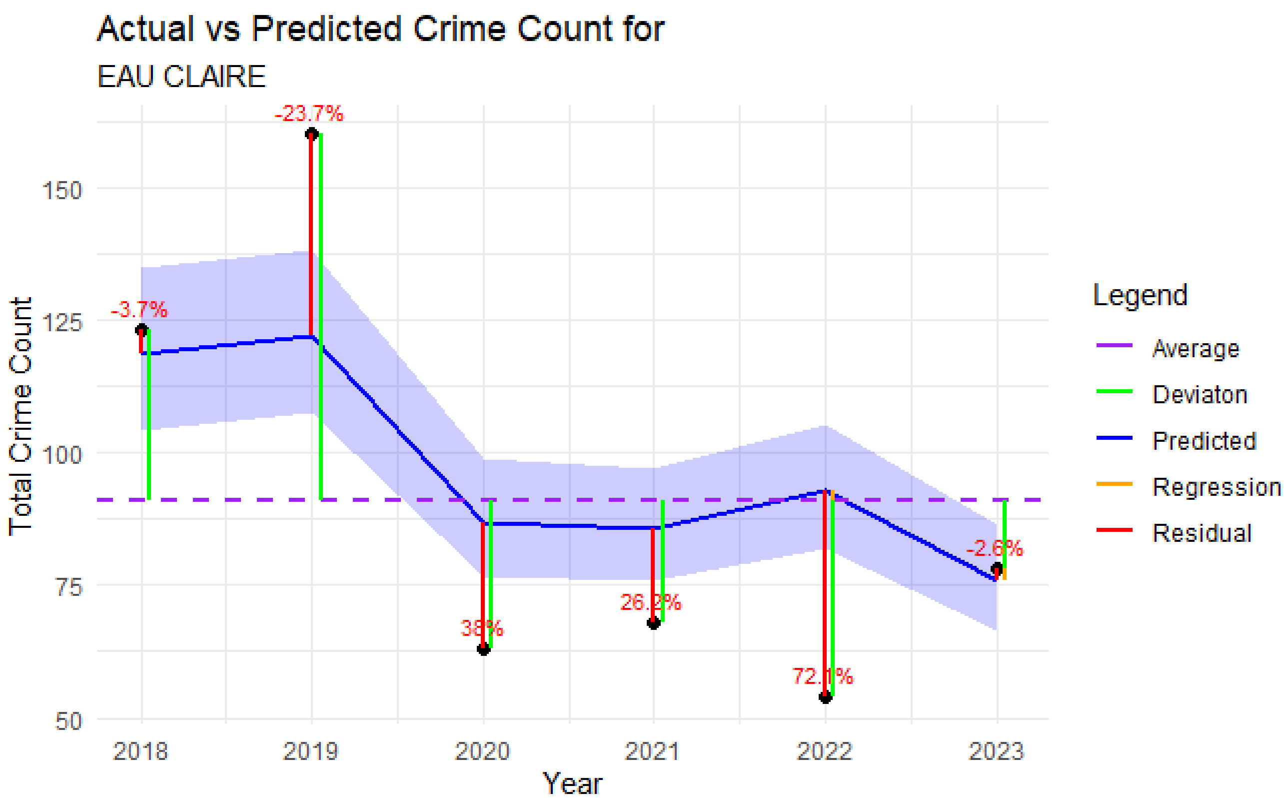
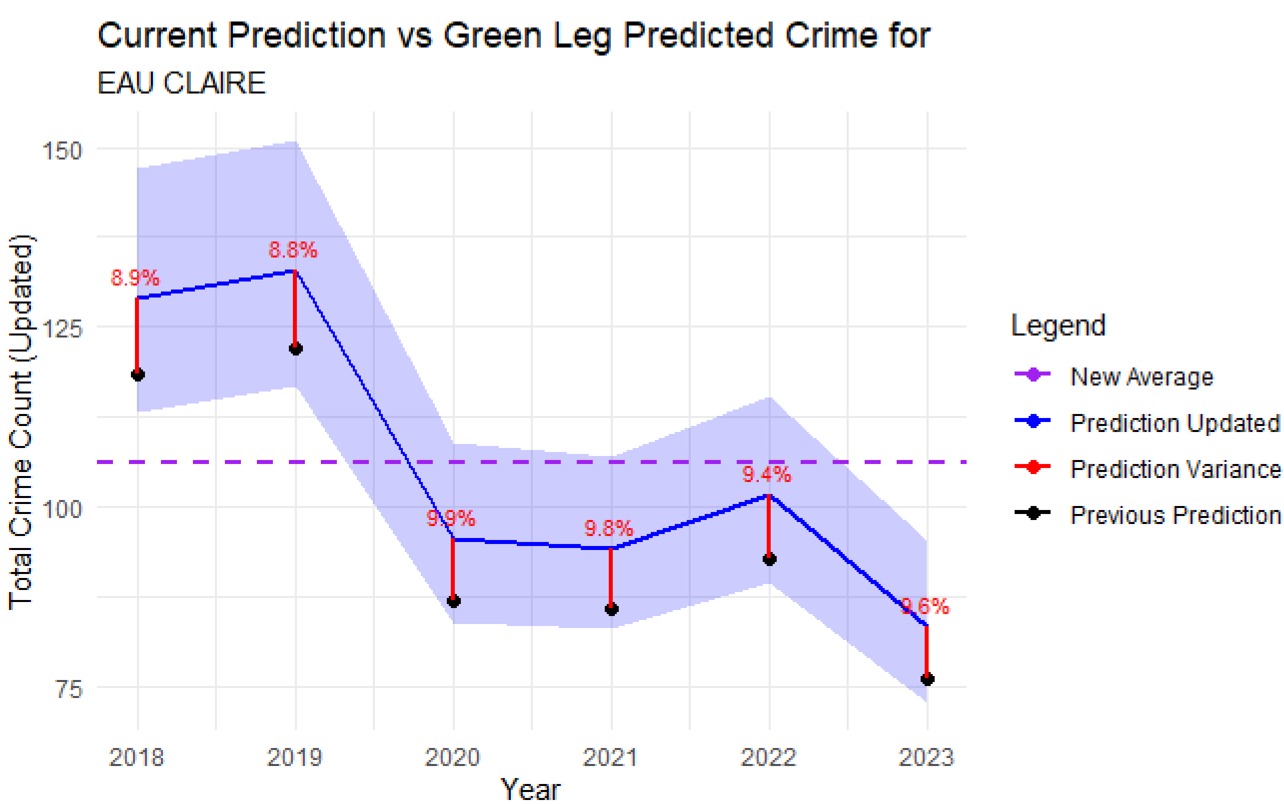


Figure 29 Actual vs Predicted Crime Counts Eau Claire



Conclusion

Summary of Findings:

Our regression model has demonstrated a surprising level of accuracy in predicting community responses with respect to their distance to the closest Light Rail Transit Station. With an adjusted R-squared value of 76%, the model accounts for a substantial portion of variability. The F-statistics, standing at 57.81 with 68 and 1152 degrees of freedom, further supports the model.

Model Predictions and Confidence Intervals:

However, following the prediction update considering data from Green Line LRT station locations, we observed wider confidence intervals at the 95% level on the updated PREDICT model. This implies greater uncertainty in the point estimates for these updated predictions, indicating there are areas in the modeling process that may require further examination.

Potential Causes for Increased Uncertainty:

Several factors may contribute to this increased uncertainty:

Multicollinearity: the distance to LRT was highlighted with its 4.19 upon performing a VIF test, the change in values could be exacerbated and lead to wider confidence intervals

Non-Linearity: The relationship between the distance to LRT and the response variable is non-linear, complicating the predictive accuracy of the model.

Variable Transformation: Adjustments made to variables, perhaps to meet model assumptions, could introduce additional variability.

Predictor Significance:

Contrary to our initial assumptions, the proximity to LRT was not the strongest predictor of community crime rate. While distance to LRT was significant, it did not hold the highest test statistic value nor was it deemed the most significant by its p-value.

Stronger Predictive Variables:

The analysis highlighted that demographic factors, particularly gender composition and the percentage of individuals aged 75 and above, had stronger predictive power. Additionally, the sector of Calgary in which a community is located was also found to have a substantial influence on the model's predictions. Reviewing the plots, there is potential patterns when comparing communities by sector with the prediction plots.

Recommendations for Model Improvement:

To refine our model's accuracy, we suggest acquiring more precise crime location data rather than using an approximation based on community center points as I assume City of Calgary does not want to violate privacy, a compromise could be postal code or forward sortation area.

An interesting next step would be implementing blocking techniques to group communities. This would allow us to isolate and control for variables, better understanding their individual and collective effects. When reviewing the prediction plots, with only a handful of communities from each sector there was a hint of patterns by geography and we know we have interaction terms with the Sector predictor variable.

Appendix A

From Data 604 our project was focused on Calgary Crime and LRT locations
We loaded into an SQL table LRT station data from these two datasets

- https://data.calgary.ca/Transportation-Transit/Transit-LRT-Stations/2axz-xm4q/about_data
- https://data.calgary.ca/Transportation-Transit/Green-Line-Stations/4y6b-yvdc/about_data

```
1 •   SELECT * FROM `l01-4`.transit_lrt_stations;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

STATION_ID	STATION_NAME	LEG	DIRECTION	DIST_NB	DIST_SB	DIST_EB	DIST_WB	ROUTE	STATUS	LRT_POINT
1	45 Street SW Station	West	West/East	NULL	NULL	NULL	NULL	202	Current	BLOB
2	Sirocco Station	West	West/East	NULL	NULL	NULL	NULL	202	Current	BLOB
3	City Hall Station	DTWestbnd	West	NULL	NULL	200.0	NULL	201/202	Current	BLOB
4	1st Street SW Station	DTWestbnd	West	0.0	0.0	467.0	439.0	201/202	Current	BLOB
5	Dalhousie Station	NW	North/South	3966.0	2732.0	0.0	0.0	201	Current	BLOB
6	Southland Station	SW	North/South	1654.0	1064.0	0.0	0.0	201	Current	BLOB
7	Fish Creek - Lacombe Station	SW	North/South	1505.0	1464.0	0.0	0.0	201	Current	BLOB
8	Centre Street Station	DTEastbnd	East	0.0	0.0	319.0	544.0	201/202	Current	BLOB
9	Bridgeland Station	NE	East/West	0.0	0.0	1100.0	1236.0	202	Current	BLOB
10	8th Street SW Station	DTEastbnd	East	1199.0	0.0	205.0	287.0	201/202	Current	BLOB
11	Marlborough Station	NE	North/South	1845.0	1722.0	0.0	0.0	202	Current	BLOB
12	69 Street SW Station	West	West/East	NULL	NULL	NULL	NULL	202	Current	BLOB
13	Sunalta Station	West	West/East	NULL	NULL	NULL	NULL	202	Current	BLOB
14	Lions Park Station	NW	North/South	1220.0	923.0	0.0	0.0	201	Current	BLOB
15	Sunnyside Station	NW	North/South	1026.0	1147.0	0.0	0.0	201	Current	BLOB
16	Canyon Meadows Station	SW	North/South	2041.0	1505.0	0.0	0.0	201	Current	BLOB
17	Barlow/Max Bell Station	NE	East/West	0.0	0.0	934.0	1379.0	202	Current	BLOB
18	City Hall Station	DTEastbnd	East	1392.0	1027.0	0.0	319.0	201/202	Current	BLOB
19	Downtown West - Kerby Sta...	DTWestbnd	West/East	0.0	0.0	NULL	NULL	202	Current	BLOB
20	Whitehorn Station	NE	North/South	2640.0	1276.0	0.0	0.0	202	Current	BLOB
21	Martindale Station	NE	North/South	NULL	NULL	NULL	NULL	202	Current	BLOB
22	Erlton/Stampede Station	SW	North/South	725.0	1669.0	0.0	0.0	201	Current	BLOB

Status column is our indicator that the LRT station is currently built and in service.
Green Line LRT stations will have a status of ‘Future’.

STATION_ID	STATION_NAME	LEG	DIRECTION	DIST_NB	DIST_SB	DIST_EB	DIST_WB	ROUTE	STATUS	LRT_POINT
48	Douglas Glen	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
49	Ogden	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
50	144 Avenue N	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
51	Eau Claire	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
52	4 Street SE	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
53	40 Avenue N	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
54	26 Avenue SE	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
55	Prestwick	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
56	South Hill	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
57	Auburn Bay / Mahogany	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
58	Mcknight Boulevard	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
59	Hospital	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
60	28 Avenue N	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
61	McKenzie Towne	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
62	9 Avenue N	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
63	North Pointe	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
64	Lynnwood / Millican	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
65	160 Avenue N	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
66	Quarry Park	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
67	Seton	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
68	96 Avenue N	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB
69	7 Avenue SW	Green	NULL	NULL	NULL	NULL	NULL	NULL	Future	BLOB

We also load Communities into an SQL table, the COMMUNITY_POINT column holds the central location of the community.

```
1 ● SELECT * FROM `l01-4`.communities;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

COMM_CODE	CLASS_CODE	COMMUNITY_NAME	SECTOR	SRG	COMM_STRUCTURE	LONGITUDE	LATITUDE	COMMUNITY_POINT
01B	4	01B	NORTHWEST	NULL	UNDEVELOPED	-114.2424553106718	51.102837628962696	BLOB
01C	4	01C	WEST	FUTURE	OTHER	-114.2371340658742	51.08677646102224	BLOB
01F	4	01F	NORTHWEST	FUTURE	UNDEVELOPED	-114.26336586209555	51.119618566466585	BLOB
01H	4	01H	WEST	FUTURE	UNDEVELOPED	-114.28068484035994	51.09104170938791	BLOB
01I	4	01I	WEST	FUTURE	UNDEVELOPED	-114.26084798313484	51.08016787165521	BLOB
01K	4	01K	NORTHWEST	NULL	UNDEVELOPED	-114.22275942835957	51.168724389792594	BLOB
02B	4	02B	NORTH	FUTURE	UNDEVELOPED	-114.19939955328874	51.17603193831223	BLOB
02C	4	02C	NORTH	FUTURE	UNDEVELOPED	-114.17667459134884	51.175901659020575	BLOB
02E	4	02E	NORTHWEST	FUTURE	OTHER	-114.19943703174916	51.161434401104025	BLOB
02F	4	02F	NORTHWEST	NULL	OTHER	-114.17507478941316	51.160012522429504	BLOB
02K	4	02K	NORTH	FUTURE	UNDEVELOPED	-114.19945296242446	51.19055316346715	BLOB
02L	4	02L	NORTH	FUTURE	UNDEVELOPED	-114.1179455753532	51.19773367470223	BLOB
03W	4	03W	NORTH	FUTURE	OTHER	-114.02476962843926	51.19796795834797	BLOB
05D	4	05D	NORTHEAST	NULL	UNDEVELOPED	-113.95866183876556	51.17959764644064	BLOB
05E	4	05E	NORTHEAST	FUTURE	UNDEVELOPED	-113.92920994202798	51.17990789589725	BLOB
05F	4	05F	NORTHEAST	FUTURE	UNDEVELOPED	-113.9164930819886	51.14703525408642	BLOB
05G	4	05G	NORTHEAST	NULL	UNDEVELOPED	-113.9165003221115	51.13619284855831	BLOB
06A	4	06A	WEST	FUTURE	UNDEVELOPED	-114.22953869988052	51.05887315839214	BLOB
06B	4	06B	WEST	FUTURE	UNDEVELOPED	-114.22950383050404	51.07789620483404	BLOB
06C	4	06C	WEST	FUTURE	UNDEVELOPED	-114.22718136194828	51.08439996156163	BLOB
09D	4	09D	CENTRE	NULL	UNDEVELOPED	-114.01242825668491	51.0118625678248	BLOB

We then perform a SQL query for each Community and determine the minimum distance from the Community Point to the LRT Point and capture the distance using a custom function based on the HAVERSINE formula(18, 19).

This SQL query to determine the shortest distance from the community to closest LRT station

```
WITH lrt_shortest_distances AS (
  SELECT
    `comm`.`COMM_CODE` AS `COMM_CODE`,
    `lrt`.`STATION_ID` AS `LRT_STATION_ID`,
    `HAVERSINE_DISTANCE_GEO`(`comm`.`COMMUNITY_POINT`, `lrt`.`LRT_POINT`) AS `DISTANCE_TO_LRT_METERS`,
    ROW_NUMBER() OVER (
      PARTITION BY `comm`.`COMM_CODE`
      ORDER BY `HAVERSINE_DISTANCE_GEO`(`comm`.`COMMUNITY_POINT`, `lrt`.`LRT_POINT`)
    ) AS `rn`
  FROM `communities` `comm`
  JOIN `transit_lrt_stations` `lrt` ON (`lrt`.`STATUS` = 'Current')
),
police_shortest_distances AS (
  SELECT
```

```
        `comm`.`COMM_CODE` AS `COMM_CODE`,
        `police`.`STATION_ID` AS `POLICE_STATION_ID`,
        `HAVERSINE_DISTANCE_GEO`(`comm`.`COMMUNITY_POINT`, `police`.`STATION_POINT`) AS `DISTANCE_TO_POLICE_METERS`,
        ROW_NUMBER() OVER (
          PARTITION BY `comm`.`COMM_CODE`
          ORDER BY `HAVERSINE_DISTANCE_GEO`(`comm`.`COMMUNITY_POINT`, `police`.`STATION_POINT`)
        ) AS `rn`
    FROM `communities` `comm`
    JOIN `police_service_stations` `police`
)
SELECT
    `c`.`COMM_CODE` AS `COMM_CODE`,
    `c`.`COMMUNITY_NAME` AS `COMMUNITY_NAME`,
    `lrt`.`LRT_STATION_ID` AS `NEAREST_LRT_STATION_ID`,
    `lrt`.`DISTANCE_TO_LRT_METERS` AS `SHORTEST_DISTANCE_TO_LRT_METERS`,
    `police`.`POLICE_STATION_ID` AS `NEAREST_POLICE_STATION_ID`,
    `police`.`DISTANCE_TO_POLICE_METERS` AS `SHORTEST_DISTANCE_TO_POLICE_METERS`
FROM `communities` `c`
LEFT JOIN `lrt_shortest_distances` `lrt` ON (`c`.`COMM_CODE` = `lrt`.`COMM_CODE` AND `lrt`.`rn` = 1)
LEFT JOIN `police_shortest_distances` `police` ON (`c`.`COMM_CODE` = `police`.`COMM_CODE` AND `police`.`rn` = 1);
```

This function is used to calculate the distance in meters between two GEOMETRY POINTs.

Function HAVERSINE_DISTANCE_GEO –

```
DELIMITER $$
CREATE DEFINER=`l01-4`@`%` FUNCTION `HAVERSINE_DISTANCE_GEO`(point1 GEOMETRY,
    point2 GEOMETRY
) RETURNS float
    DETERMINISTIC
BEGIN
    DECLARE lat1 FLOAT;
    DECLARE lon1 FLOAT;
    DECLARE lat2 FLOAT;
    DECLARE lon2 FLOAT;
    DECLARE R INT DEFAULT 6371000; -- Earth radius in meters
    DECLARE phi1 FLOAT;
    DECLARE phi2 FLOAT;
    DECLARE delta_phi FLOAT;
    DECLARE delta_lambda FLOAT;
```

```
    DECLARE a FLOAT;
    DECLARE c FLOAT;
    DECLARE d FLOAT;

    -- Extract lat and lon from the points
    SET lon1 = ST_X(point1);
    SET lat1 = ST_Y(point1);
    SET lon2 = ST_X(point2);
    SET lat2 = ST_Y(point2);

    -- Convert degrees to radians
    SET phi1 = radians(lat1);
    SET phi2 = radians(lat2);
    SET delta_phi = radians(lat2 - lat1);
    SET delta_lambda = radians(lon2 - lon1);

    -- Haversine formula
    SET a = sin(delta_phi / 2) * sin(delta_phi / 2) +
            cos(phi1) * cos(phi2) *
            sin(delta_lambda / 2) * sin(delta_lambda / 2);
    SET c = 2 * atan2(sqrt(a), sqrt(1-a));
    SET d = R * c;

    RETURN d;
END$$
DELIMITER ;
```

REFERENCES

1. Police-reported crime statistics in Canada, 2022. "Statistics Canada." Retrieved from https://www150.statcan.gc.ca/n1/daily-quotidien/230727/dq230727b-eng.htm?HPA=1. Date accessed: 2024, March 3.
2. Crime severity index and weighted clearance rates, Canada, provinces, territories and Census Metropolitan Areas. "Statistics Canada." Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3510002601. Date accessed: 2024, March 3.
3. The Corporate Research Team, Customer Service and Communications, The City of Calgary. 2023 Perspectives on Calgary Safety Perceptions, Final Report. June 2023.
4. Wilkinson, D. L., and Fagan, J. A theory of violent events. In R. F. Meier, L. W. Kennedy, & V. F. Sacco (Eds.), The process and structure of crime: Ciminal events and crime analysis (pp. 169–195). New York: Transaction Publishers; 2001.
5. The Chicago School And Cultural/Subcultral Theories of Crime. Retrieved from https://www.sagepub.com/sites/default/files/upm-binaries/29411_6.pdf. Date accessed: 2024, March 3.2009.
6. Community Crime Statistics. "City of Calgary's Open Data Portal." Retrieved from https://data.calgary.ca/Health-and-Safety/Community-Crime-Statistics/78gh-n26t/about_data. Date accessed: 2024, March 3.
7. Unemployment rates. "City of Calgary's Open Data Portal." Retrieved from https://data.calgary.ca/Business-and-Economic-Activity/Unemployment-rates/wzpt-744u. Date accessed: 2024, March 3.
8. Building Permits by Community. "City of Calgary's Open Data Portal." Retrieved from https://data.calgary.ca/Business-and-Economic-Activity/Building-Permits/c2es-76ed/about_data. Date accessed: 2024, March 3.
9. Assessments by Community. "City of Calgary's Open Data Portal." Retrieved from https://data.calgary.ca/Government/Assessments-by-Community/p84b-7zbi/about_data. Date accessed: 2024, March 3.
10. Civic Census by Community. "City of Calgary's Open Data Portal." Retrieved from https://data.calgary.ca/Demographics/Civic-Census-by-Community/s7f7-3gjj/data. Date accessed: 2024, March 3.
11. Open Government Licence. "City of Calgary's Open Data Portal." Retrieve from https://data.calgary.ca/stories/s/Open-Calgary-Terms-of-Use/u45n-7awa. Date accessed: 2024, March 3.
12. Bursik, R. J., & Grasmick, H. G. (1993). Neighborhoods and crime: The dimensions of effective community control. New York: Lexington Books.
13. Bridgeland, W. M. (1979). Ruth R. Kornhauser. Social Sources of Delinquency. Pp. ix, 277. Chicago: University of Chicago Press, 1978. The Annals of the American Academy of Political and Social Science, 442(1), 172-173.
14. Daday. JK, Broidy. LM, Crandall. CS, Sklar DP. Individual, Neighborhood, and Situational Factors Associated with Violent Victimization and Offending. Criminal Justice Studies. 2005;18(2):215-35.
15. Field A. Discovering statistics using SPSS. 3 ed. London: SAGE publications Ltd; 2009. p. 822. .
16. D.Oztuna, AH.Elhan, E.Tuccar. Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. Turkish Journal of Medical Sciences. 2006;3(63):171-6.
17. Green Line Stations. "City of Calgary's Open Data Portal." Retrieved from https://data.calgary.ca/Transportation-Transit/Green-Line-Stations/4y6b-yvdc/about_data. Date accessed: 2024, March 3.

18. Kyle Dempsey on Jan 17, 2024. Calculating distance between two points in SQL. https://www.airops.com/blog/calculating-distance-between-two-points-in-sql#:~:text=The%20Haversine%20formula%20allows%20you,using%20latitudinal%20and%20longitudinal%20coordinates. Date accessed: 2024, March 30.
19. Haversine Calculation in User Defined Function on 2003-09-25. Retrieved from https://www.sqlservercentral.com/scripts/haversine-calculation-in-user-defined-function. Date accessed: 2024, March 30.