UNITED STATES MILITARY ACADEMY

PROJECT REPORT

MA478: GENERALIZED LINEAR MODELS

SECTION H2

COL NICHOLAS CLARK

BY

CADET SYDNEY WATSON '24, CO G1

WEST POINT, NEW YORK

07 MAY 2024

WATSON.SYDNEY.NICOLE.1246387717 Digitally signed by WATSON.SYDNEY.NICOLE.1246387717 Date: 2024.05.07 22:55:28 -04'00'

____ I CERTIFY THAT I HAVE COMPLETELY DOCUMENTED ALL SOURCES THAT I USED TO COMPLETE THIS ASSIGNMENT AND THAT I ACKNOWLEDGED ALL ASSISTANCE I RECEIVED IN THE COMPLETION OF THIS ASSIGNMENT.

____ I CERTIFY THAT I DID NOT USE ANY SOURCES OR RECEIVE ANY ASSISTANCE REQUIRING DOCUMENTATION WHILE COMPLETING THIS ASSIGNMENT.

SIGNATURE: WATSON.SYDNEY.NICOLE.1246387717 Digitally signed by WATSON.SYDNEY.NICOLE.1246387717 Date: 2024.05.07 22:55:47 -04'00'

Statistical Analysis of Burglaries in Chicago

Abstract

This research aims to uncover relationships between socioeconomic and temporal factors that impact the number of burglaries in each zip code in Chicago. We go on to answer these guiding questions: (1) Are socioeconomic factors associated with the number of daily burglaries? (2) Is there an effect of zip code on daily burglaries that the other features do not account for? (3) Is there an impact of day of the week on burglaries? Should additional officers be hired part time? We used data from 2022-2023 from the Chicago Data Portal and the U.S. Census Bureau. We used a Poisson Regression, a Mixed Effects Poisson Model, and a Generalized Linear Mixed Model to answer each question, respectively. We found that there is an association between all the socioeconomic factors, there is an effect of zip code on daily burglaries, and that Friday and Monday are significant days for a decrease in burglaries.

Key Words: Chicago, Burglaries, Chicago Police Department, 2022-2023, Census, Chicago Data Portal

Introduction

Crime in Chicago has been running rampant since the 90s. Police departments and policymakers are constantly seeking to gain more understanding on the factors that drive crime to create a safe environment for all its residents. This study aims to highlight relationships between burglaries in Chicago by zip code and factors like total population, median earnings, day of the week, and the ratio of males to females. We are looking to build models to answer the following questions: (1) Are socioeconomic factors associated with the number of daily burglaries? (2) Is there an effect of zip code on daily burglaries that the other features do not account for? (3) Is there an impact of day of the week on burglaries? Should additional officers be hired part time? We will use various statistical models to answer these questions and provide policy recommendations to the Chicago Police Department based on our findings.

Literature Review

When considering our data we pulled from the Chicago Police Department Data Portal and the U.S. Census. We concluded that there are many sources of variation that stem from this data

collection and might have an impact on how we interpret our results and provide our recommendations. In addition to this, one of the factors we wanted to explore was the day of the week that crimes in Chicago occurred. Due to the crime in Chicago being a prevalent topic across major cities and academia, statisticians and researchers have explored these topics. The two topics that furthered our understanding of our data and the crime in Chicago are the civilian sentiments around reporting and spatial-time analysis of crimes.

Based on our experiences with Chicago in the media and popular culture we know that there exists a general distrust between citizens and police officers' due to things like misuse of different law enforcement methods and inconsistent investigation practices. In and article by Maribeth L. Rezey and Janet L. Lauritsen, both professors of Criminology, they use the current reported crime rates and use the National Crime Victimization Survey (NCVS) to spot any potential inconsistencies between citizens in Chicago reporting or not reporting crimes and compare it to the crimes that the Chicago Police Department (CPD) have on record.[1] The primary findings from this study were that most Chicago residents will report burglaries and that their reasons for not reporting is what varies across other major cities in the United States and that the number of burglaries recorded in the Chicago Data Portal were on par with the number of burglaries reported in the NCVS.[2] This article helps put into perspective any potential holes in our data set. It also allows us to address any concerns of mistrust playing a role into crimes not being reported, ultimately impacting the data from the Chicago Data Portal.

The other factor within our data we wanted to explore was the impact of month and day of the week on the burglaries in Chicago. In our initial exploration of the data, we found that month over years, exhibited a cyclical pattern and we also assumed that there must be a similar pattern for day of the week. In an article by Jun Luo, a professor in the Department of Geography at Missouri State, he uncovers the spatial-time relationship of residential burglaries in Chicago.[3] He conducted analysis by applying a Getis-Ord G* statistic to each spatial-temporal time instance and found that crimes between 2006 and 2016 were most common in the month of December, on the day of the week Monday, and on the time of day from 0000 to 0200.[4] While we will not be focusing on this analysis method, this research helps us confirm that month and day of the week will provide additional clarity into burglaries in Chicago and that it is worth uncovering in our

own analysis. This research will also serve as a baseline and guide our findings and interpretations of how we interpret day of the week to impact burglaries in Chicago.

These articles provided us better understanding of our data and the potential ways we can explore and expand on the literature surrounding burglaries in Chicago. Knowing that each entry in the Chicago Data Portal accounts for almost all crimes in Chicago informs us that we have the most updated and comprehensive dataset of crimes in Chicago. In addition to this, seeing that there has been research done surrounding the spatial temporal aspects of crimes in Chicago informs us that time of year, month, and day is a key factor to consider in the exploration of Chicago burglaries.

Methodology

We obtained our data from the Chicago Data Portal and collected all crimes from 2022 to 2023, we then found corresponding socioeconomic factors to supplement the data set from the U.S. Census. Once we obtained our data, we cleaned the data types, deleted any entries that were missing the location, and then queried for all crimes that were burglaries. We transformed the population of men column to be the ratio of men to 100 women.

To find the socioeconomic factors associated with the number of daily burglaries we built a spatial model with Census covariate (Appendix A). We specifically used a Poisson model. To compare the first model, we create a null model as the baseline and then used the dispersion parameter, log likelihood, and the residuals v. fitted values.

To answer the second question to see if there was an effect of zip code on daily burglaries that the Census factors did not account for, we used a spatial model that accounted for the random effect of the zip code (Appendix B). We used a Poisson Mixed-Effects Model to accomplish this. To compare this model, we created another null model and looked at the deviance and residuals v. fitted values.

Finally, to see the relationship between day of the week and number of burglaries we created a spatial and temporal model that included zip code as a random effect and each day of the week was a fixed effect with Friday being the baseline day (Appendix C). We decided to use the Generalized Linear Mixed Model (GLMM) to handle the spatial and temporal factors. We looked at just the residuals v. fitted values for the final model.

Data Exploration

During our data exploration we wanted to visualize trends and confirm any assumptions we have that may impact our interpretation of the model. We started our exploration with burglaries over time. We saw there was a cyclical relationship between the month and number of burglaries (Appendix D). Burglaries would consistently increase during the summer months and were lower during the winter months. When looking at burglaries over day of the week we noticed that number of burglaries was consistent across all days and did not expect for day of the week to be associated with number of burglaries (Appendix E). We then looked at the number of burglaries by zip code and noticed that burglaries that were next to each other had similar numbers of burglaries (Appendix F).

Next, we plotted the number of burglarized by median earnings and saw that there was a higher number of burglaries when the median income was around $35,000 (Appendix G). We initially believed that population would impact the number of burglaries because as population increases, there should be a higher number of burglaries (Appendix H). With this knowledge we knew we had to include population as an offset in our models. Finally, we created a correlation matrix to see if there was any correlation between the census features and the total number of burglaries (Appendix I). We found that there was a possible correlation between population and median earnings. Seeing a correlation between population and total number of burglaries prompted us to include population as an offset in our models.

Modeling

For our first question we create a baseline Poisson Model and then a Negative Binomial model to be able to compare if one model is preferrable to another (Appendix A). In this model we were looking for number or burglaries in each zip code on a day of the week. We considered the median earnings in dollars, the percentage of the population in each zip code has a degree above high school, and the ration of males to 100 females. We included an offset for population because from our data exploration we know that there is an effect of population on number of burglaries. We then compared the models using the dispersion parameter, log likelihood, and the residuals v. fitted plot (Appendix J).

For our second model we create a null model and then a Poisson Mixed Effects Model to see if there was any difference (Appendix B). In this model we were looking for the number of burglaries in each zip code by day of the week and included zip code as a random effect in out model and maintained the offset of population. We chose to include zip code as random effect because we are assuming that burglaries within a given zip code are more similar than comparing burglaries from two different zip codes. This random effect accounts for any cultural similarities that exist in a specific zip code. To determine which model to use we used the deviance and the residuals v. fitted values (Appendix K).

For the final model we chose to do a GLMM to consider both fixed and random effects in one model (Appendix C). In this final model, each day of the week (Sunday to Saturday) is a fixed effect because we wanted to see if any specific day influenced the number of burglaries in each zip code. We kept zip code as a random effect and the population as an offset for the same reasons stated in the first two models. To determine model effectiveness, we looked at the plot of residuals v. fitted values (Appendix L).

Results

In Model 1 (Appendix A) we found that while the Negative Binomial model is preferred, that the original Poisson model was more interpretable, and the residuals v. fitted plot did not suggest that the Poisson model was insufficient for the data. We also uncovered that median earning in dollars, the ratio of men for every 100 women, and the percentage of a zip code having a degree above high school are associated with the number of burglaries in each zip code on a given day.

In Model 2 (Appendix B), this model is preferrable to the null model based on the deviances (Appendix K). The residuals v. fitted plot does not suggest that the model was insufficient. From this we concluded that there is an effect of zip code on burglaries that were not captured by previous factors in Model 1.

In Model 3 (Appendix C) we only relied on the plot of residuals v. fitted values. Based on this model we observed that each day of the week suggested a decrease in the average number of burglaries. We found that Friday and Monday were statistically significant days and had the largest coefficients in the model: -7.204 for Friday and .144 on Monday.

Discussion and Conclusion

Based on our results we found that there is an association between all the socioeconomic factors, there is an effect of zip code on daily burglaries, and that Friday and Monday are significant days for a decrease in burglaries. Based on our data exploration our initial assumptions of population and median earnings are associated with the number of burglaries given a zip code. Our findings on day of the week contradicted Professor Luo's spatial temporal research on burglaries in Chicago when he claims that Monday's and Friday's see an influx of burglaries.

Our model only considers socioeconomic factors and many of these factors are also associated with demographic factors like race. If we were to consider race as a factor in our model, we would have to be cognizant of pre-existing dynamics that exist between certain races and the police in the area. Making decisions based solely off our models and not considering these factors could have negative repercussions on neighborhoods if policymakers create changes that do not consider the current social climate of certain neighborhoods. Using our model to highlight increase manning may put a financial strain on current police precincts. With any changes made from this research, it will be difficult to attribute any singular policy to a decrease in burglaries with policies always being enacted. To mitigate any potential ethical considerations, it is crucial to integrate any new policies gradually and frequently touch base with the communities experiencing increased police presence or the effects of a certain policy.

Some limitations to our study include lacking intersectionality with demographics, considering the time of day, and fully building out each model to correctly account for different variations. Including factors like race, age, and religion distributions across zip codes may better prepare police officers for the environment they will be operating in. Conversely having a better understanding of these demographics may perpetuate biases in certain communities and worsen stigmas surrounding racial profiling. In the study by Professor Luo, he considers the time of day that burglaries occur. This data was included in the Chicago Data Portal and adding this information would provide another level that would increase our understanding of what burglaries in Chicago look like. While we created a null model for the first two models, we did not create a baseline for the final model and used the residuals v. fitted plot to compare it to the

other two models. Creating a base model would allow us to compare it effectively and include additional temporal factors we had in the data set.

The landscape of crime in Chicago would benefit from additional research that considers more factors to highlight what conditions encourage burglars to burglarize. Some things in addition to our Census data could be the type of alarm systems people use or if they use alarms at all, what is the dollar amount stolen in a burglary, and whether a zip code or census block is primarily residential or commercial. Looking at alarm systems may help policymakers consider ways to find deter burglars. The dollar among of items stolen and whether a zip code is primarily residential may have an association that impacts a burglar's decision to burglarize that building or may serve as a deterrent if the perceived dollar amount is not high enough. In addition to further statistical analysis, obtaining anecdotes from burglar's could provide additional insight that confirms some of the findings in this or future studies.

References

[3, 4] Luo, J. “MULTI-SPATIOTEMPORAL PATTERNS OF RESIDENTIAL BURGLARY CRIMES IN CHICAGO: 2006-2016.” *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences* IV-4/W2 (2017): 193–198. Web.

[1,2] Rezey, Maribeth L, and Janet L Lauritsen. “Crime Reporting in Chicago: A Comparison of Police and Victim Survey Data, 1999–2018.” *The journal of research in crime and delinquency* 60.5 (2023): 664–699. Web.

Appendix

Appendix A: Model 1 – Poisson Regression Model

$$\eta_{i,j} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \log(population)$$
$$i = zip\ code$$
$$j = day\ of\ week$$
$$y_{i,j} = burglaries\ in\ zip\ code\ by\ day\ of\ week$$
$$y_{i,j} \sim pois(\lambda_{i,j})$$
$$\log(\lambda_{i,j}) = \eta_{i,j}$$
$$x_{1i} = median\ earnings$$
$$x_{2i} = \%\ degree$$
$$x_{3i} = male\ ratio$$

Appendix B: Model 2 – Poison Mixed Effect Model

$$\eta_{i,j} = \beta_0 + \psi_i + \log(population)$$
$$i = zip\ code$$
$$j = day\ of\ week$$
$$y_{i,j} = burglaries\ in\ zip\ code\ by\ day\ of\ week$$
$$y_{i,j} \sim pois(\lambda_{i,j})$$
$$\log(\lambda_{i,j}) = \eta_{i,j}$$
$$(zip\ code\ r.e.)\ \psi_i \sim N(0, \sigma_u^2)$$

Appendix C: Model 3 – Generalized Linear Mixed Model

$$\eta_{i,j} = \beta_0 + \beta_1 x_{mon} + \beta_2 x_{tues} + \beta_3 x_{wed} + \beta_4 x_{thur} + \beta_5 x_{fri} + \beta_6 x_{sat} + \psi_i + \log(population)$$

$$i = zip\ code$$

$$j = day\ of\ week$$

$$y_{i,j} = burglaries\ in\ zip\ code\ by\ day\ of\ week$$

$$\beta_0 = \text{Sunday}$$

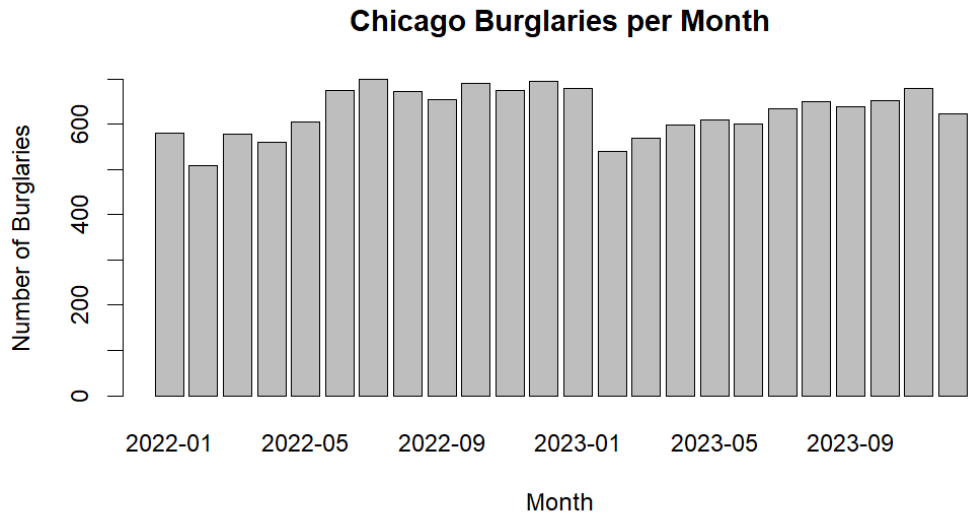
$$y_{i,j} \sim pois(\lambda_{i,j})$$

$$\log(\lambda_{i,j}) = \eta_{i,j}$$

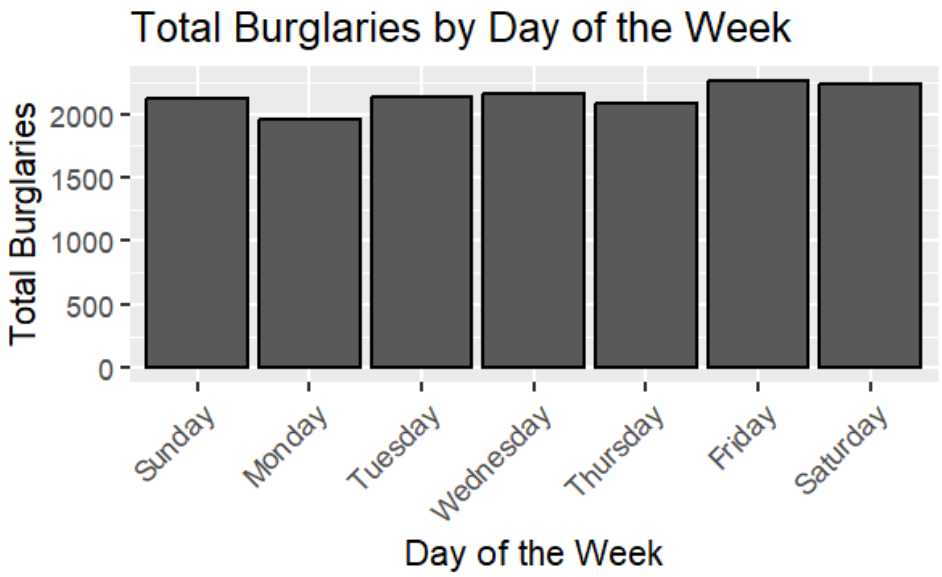
$$x_{day} = \begin{cases} 1: if\ it\ is\ respective\ day \\ 0: otherwise \end{cases}$$

$$(zip\ code\ r.e.)\ \psi_i \sim N(0, \sigma_u^2)$$

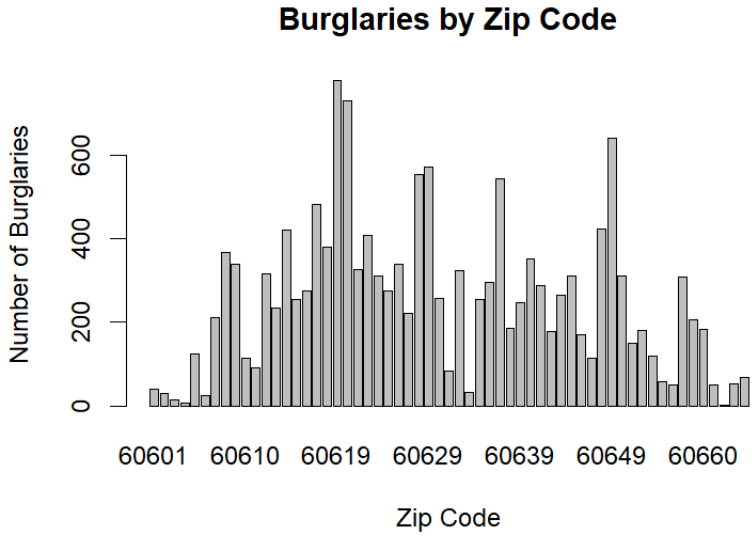
Appendix D: Bar Chart of Burglaries per Month



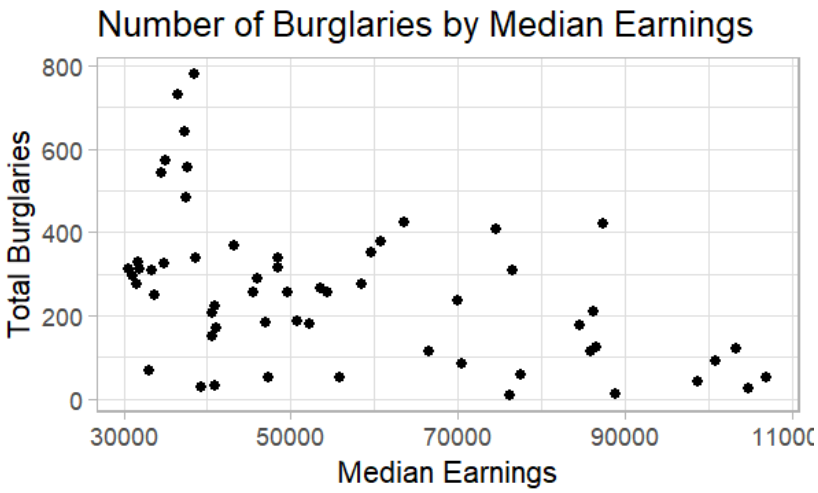
Appendix E: Count of Burglaries by Day of the Week



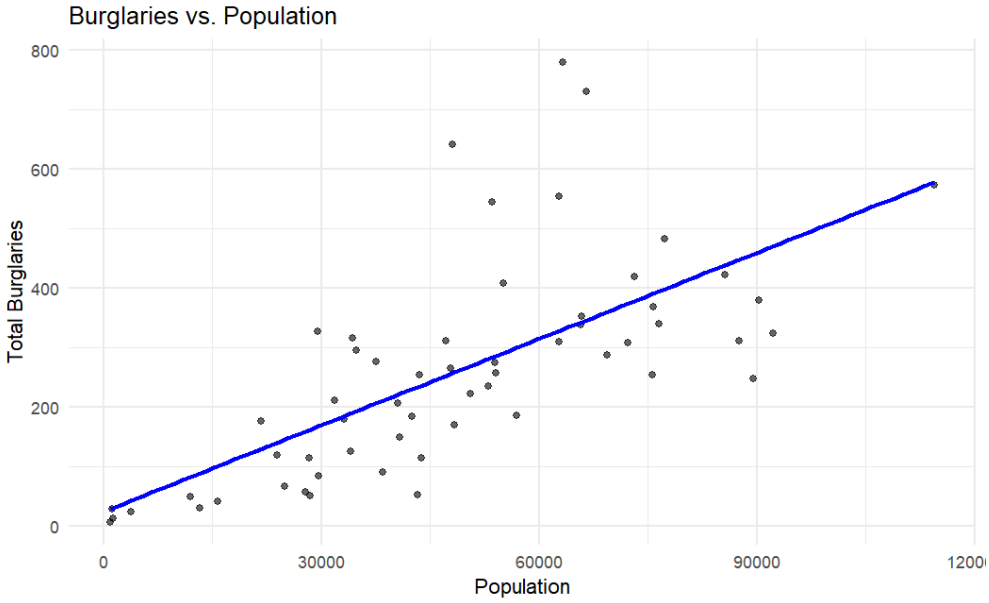
Appendix F: Bar Chart Burglaries by Zip Code



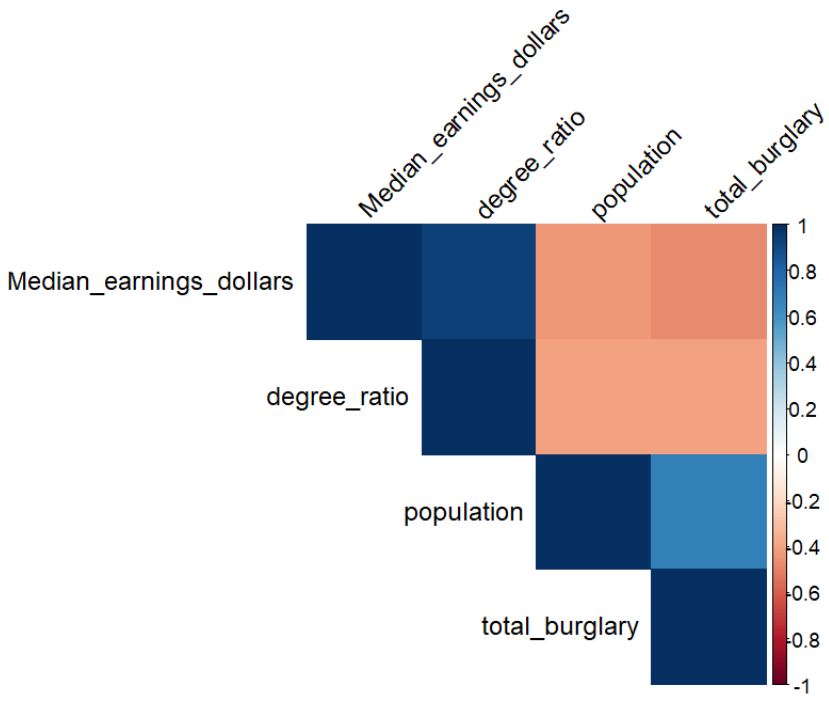
Appendix G: Scatter Plot of Burglaries by Median Earning ($)



Appendix H: Scatter Plot of Total Burglaries by Population

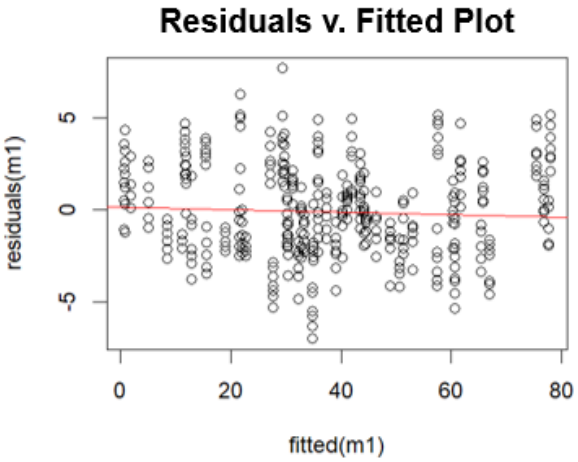


Appendix I: Correlation Matrix of Selected Features



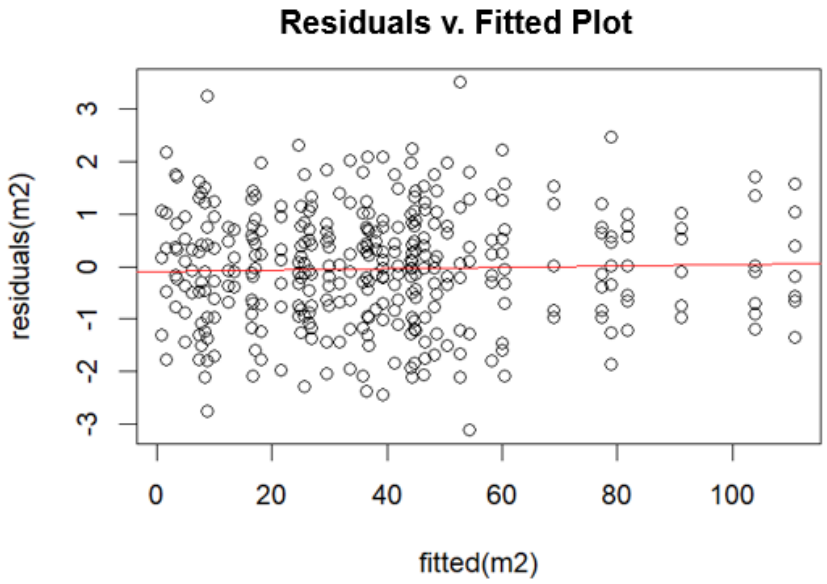
Appendix J: Model 1 Results

	Poisson	Negative Binomial
Dispersion Parameter	6.346	5.655
Log Likelihood	-2304.64	-1626.09



Appendix K: Model 2 Results

	Deviance
Null Model	3560.3
Model 2	2783.7
AIC	2787.7



Appendix L: Model 3 Results