University of Waterloo
Faculty of Engineering
Department of Management Sciences

Exploring and Predicting Violent Crime in Chicago

University of Waterloo
200 University Ave W, Waterloo, ON N2L 3G1
Waterloo, Ontario, Canada

Prepared by

Yingzi Zhang
20515934
4A Mechatronics Engineering

And

Xiang Li
20574900
4A Mechatronics Engineering

27 November 2018

1 Abstract

Chicago has historically had a consistently higher rate of violent crimes compared to the U.S. average. As a result, there is a demand for better police resource allocation to community areas where violent crimes are likely to occur. The authors of this report hypothesize that there may be multiple city-related predictors such as demographic, educational, health, and urban planning data that may be correlated with the rate of violent crime. 12 explanatory variables are collected: average school rating, average SSL rating, total park area, number of hospitals, teenage pregnancy rate, infant mortality rate, proportion of Hispanics, blacks, whites, Asians, 'other' races, and finally, percent of children in poverty.

Exploratory data analysis shows that many of these variables are correlated with the class variable, such as: teenage mother birth rate, infant mortality rate, percentage of children in poverty, and percentage of black people, Hispanics, whites, and Asians.

Regression, clustering analysis, and association rule mining were conducted. With both ordinary least squares and elastic nets linear regression, a relatively high R^2 value of 0.82 was achieved, which means that the 12 mentioned predictors can, in fact, accurate predict violent crime rate of a community area. In addition, by using the k-means algorithm and looking at the corresponding elbow curve, it seems clear that the data is naturally grouped into 3 clusters, and that these clusters

generally seem to make sense. For example, low teen birth rate is clustered with low infant mortality rate; the vice versa is also clustered together. Finally, association rule mining revealed that a low percentage of black people in a community area is highly correlated with a low rate of violent crime. The vice versa is also highly associated with each other. In addition, it was discovered that a high percentage of black people is associated with a low percentage of all other races.

2 Introduction

Crime in Chicago has been recorded since the last century by the Chicago Police Department, and throughout this entire period, Chicago has had a consistently higher rate of violent crimes compared to the U.S. average. As a result, the city has been under intense scrutiny for its high violent crime rates. In 2016, the murder rate in the U.S. rose about 13%, and almost half of that increase is solely due to violence in Chicago [1]. In fact, in recent years, Chicago had recorded more murders and victims of shooting than both Los Angeles and New York City combined [2]. Such high rates of violent crime naturally lead to a need for better police resource allocation to areas where violent crimes are most likely to occur.

The authors of this report hypothesize that there may be multiple city-related predictors such as demographic, educational, health, and urban planning data that may be correlated with the rate of violent crime in Chicago. These include: the quality of nearby schools, how close the crimes are to any parks or recreation areas, the demographics of the neighborhood that the crime took place in, teenage pregnancy rate, and many other variables that will be discussed. Each of these predictors will be grouped by ‘community area’. Chicago has 77 official community areas, and these areas have been historically used in sociological research of Chicago, and is thus deemed an appropriate geographic unit of choice for this project [3].

The authors of this report hope that a data science approach can prove to be beneficial in predicting where and when violent crimes are more likely to occur. These predictions can then aid police officers by helping them efficiently allocate required resources to crime hotspots.

Firstly, a regression algorithm will be made in an attempt to accurately predict the rate of violent crime for each community area of Chicago via the city-related explanatory variables. Secondly, both a clustering algorithm and an association rule mining algorithm will be used to see if the explanatory variables are related in any way with each other. By reviewing extant academic literature, a specific set of variables were determined that the authors' of this report hypothesize may be important predictors of crime.

The rest of the format of this report is as follows. The *Related Works and Hypotheses* section describes related works in academic literature, and the sources of inspiration for the explanatory variables. The *Data Collection Progress* section describes the source of datasets and how the explanatory variables were collected for the clustering and predictive analysis. The *Exploratory Data Analysis* section summarizes noteworthy insights obtained from plotting various plots of each explanatory variable. Finally, the *Numeric Regression Results*, *Clustering Analysis Results*, and *Association Rule Mining Results* sections explain the supervised and unsupervised learning results using the collected data.

3 Related Works and Hypotheses

There are several related works that involve analyzing Chicago's crime data. Firstly, a noteworthy prediction model called the "Strategic Subject List" (SSL) was made by the Illinois Institute of Technology for the Chicago Police Department. This list attempts to identify 'subjects' who are most at risk of being involved in a shooting [4]. The authors hypothesize that a high SSL score may be correlated with a high violent crime rate for a community area.

The authors also hypothesize that the quality and amount of schools and parks may have an effect on violent crime. The effect of education on crime has already been represented broadly in literature, such as in Lochner and Moretti's paper [5]. In addition, there has been a paper by Schusler et al. that examined the association between the amount of tree canopy area and crime in each census tract boundary in Chicago, which led to the idea that the amount and size of city parks could also be relevant predictors. [6].

Finally, demographic and health data are also hypothesized to be relevant factors, as described in Brantingham's book, "Patterns in Crime". In addition, there are also prominent health-related crime theories, such as the lead-crime hypothesis that claims blood lead levels in children have a direct correlation with crime [7]. These led the authors of this report to hypothesize that race, poverty rate, and available health information from Chicago's public health datasets, such as teenage pregnancy rate and infant mortality rate, may be relevant predictors.

4 Data Collection Progress

Table 1 displays all of the final chosen datasets. From these datasets, explanatory variables are obtained.

Table 1. Information on Final Chosen Datasets.

Dataset	Number of Rows	What Does a Row Represent?
Crimes from 2001 [8]	6,706,459	Reported Crime
Strategic Subject List [4]	398,684	Person Likely to be Involved in a Shooting
Chicago Public Schools - School Profile Information SY1718 [9]	661	School
Population and Poverty Data by Chicago Community Area [10]	77	Community Area
Parks - Chicago Park District Park Boundaries (current) [11]	N/A (Shapely File of 597 Parks)	N/A (Shapely File of 597 Parks)
Boundaries - Community Areas (current) [12]	N/A (Shapely File of 77 Community Areas)	N/A (Shapely File of 77 Community Areas)
Hospitals – Chicago [13]	N/A (Shapely File of 42 Hospitals)	N/A (Shapely File of 42 Hospitals)
Public Health Statistics - Births to mothers aged 15-19 years old in Chicago, by year, 1999-2009 [14]	77	Community Area
Public Health Statistics- Infant mortality in Chicago, 2005– 2009 [15]	77	Community Area

4.1 Class Variable

The class variable is the rate of violent crime in an area. In order to scale the results so that communities with higher populations are not overrepresented, the

class variable is calculated as the percent of violent crime per 1000 people in the specified community area. This is obtained via the following formula:

*violentCrimeForCommunityArea * 1000 / populationOfCommunityArea*

Note that the multiplier of 1000 will not change prediction results, as with it, all the constants of a regression model will simply be scaled up. It will also not have any effect in clustering or association rule mining, as the former normalizes all of its explanatory variables, and the latter bins them. The reason this multiplier is included is because it is a standard way to assess rate by population in social sciences, and is thus more easily interpretable for people [16].

In order to obtain the total number of violent crime in a community area, the "Crimes from 2001" dataset is used. In this dataset, each crime is given a type and the community area the crime occurred in, and each crime's type was used to filter for violent types only. Examples of violent crime types include: assault, battery, and homicide. Finally, the population of each community area is obtained from census data.

4.2 Predictors

Table 2 lists all 12 predictors included in this project for both supervised and unsupervised learning. Note that all 12 predictors are numerical variables.

Table 2. Explanatory Variables.

Explanatory Variable	Description
avgSchoolRating	The average rating of all schools per community area. Rating from 1 to 5, where 5 is the best. Based on the

	Chicago Public School Board's official "School Quality Rating Policy" [17]
avgSSLRating	The average "Strategic Subject List" score for all strategic subjects who committed a crime in a specific community area. Range between 0 to 500, where 500 represents that subjects in this community area are most likely to be involved in a future shooting.
totalParkArea	The total park area of a community area in m^2.
numHospitals	The number of hospitals for a community area.
teenMomRate	The teenage pregnancy rate for a community area.
infantMortalityRate	The infant mortality rate for a community area.
hispanic	The percentage of people who are Hispanic in a community area.
black	The percentage of people who are black in a community area.
white	The percentage of people who are white in a community area.
asian	The percentage of people who are Asian in a community area.
other	The percentage of people who are designated as "other race" in a community area.
percentChildrenInPov	The percentage of children who are in poverty for a community area. Note that the actual poverty rate per community area could not found by the authors of this report. However, the authors believe percentage of children in poverty is an appropriate substitute.

The Appendix section at the end of this report will describe in detail how all the predictors are obtained and preprocessed from the aforementioned datasets.

5 Exploratory Data Analysis

Scatter plots, histograms, and box plots are used to analyse each variable, and detailed comments on each are included in the Appendix. This section will summarize noteworthy results.

It is worth noting that some variables are not normally distributed, but rather exponentially distributed, including the class variable of violent crime rate. Explanatory variables that have such non-normal distributions include: total park area, number of hospitals, percentage of Hispanics, percentage of whites, and percentage of Asians. Interestingly, as shown in Figure 1, the distribution of the percentage of black people has an upside-down bell curve shape, which implies that community areas generally either have few to no black people, or have a large percentage of black people. Very few seem to have a moderate percentage of black people.

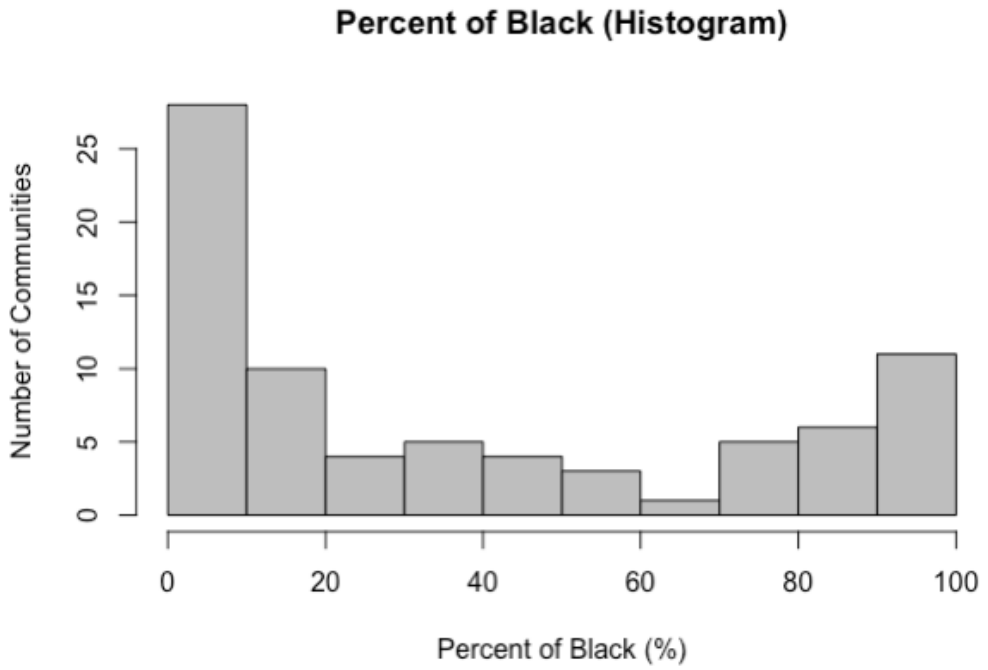

Figure 1. Histogram for Percentage of Black People.

Finally, out of the 12 explanatory variables, 7 of them visually have a clear correlation with the class variable, as seen from scatter plots. The rest do not seem to have as strong of a correlation. Teenage pregnancy rate, infant mortality rate, percentage of black people, and percentage of children in poverty had a strong positive correlation with the class variable. In contrast, the percentage of Hispanics, whites, and Asians had a strong negative correlation. Some of these correlations also seemed to be non-linear. These include: total park area, teenage mother birth rate, infant mortality rate, percentage of Hispanics, percentage of whites, percentage of Asians, and percentage of children in poverty. An example of a positive non-linear correlation can be seen in Figure 2 for percent of children in poverty versus rate of violent crime.

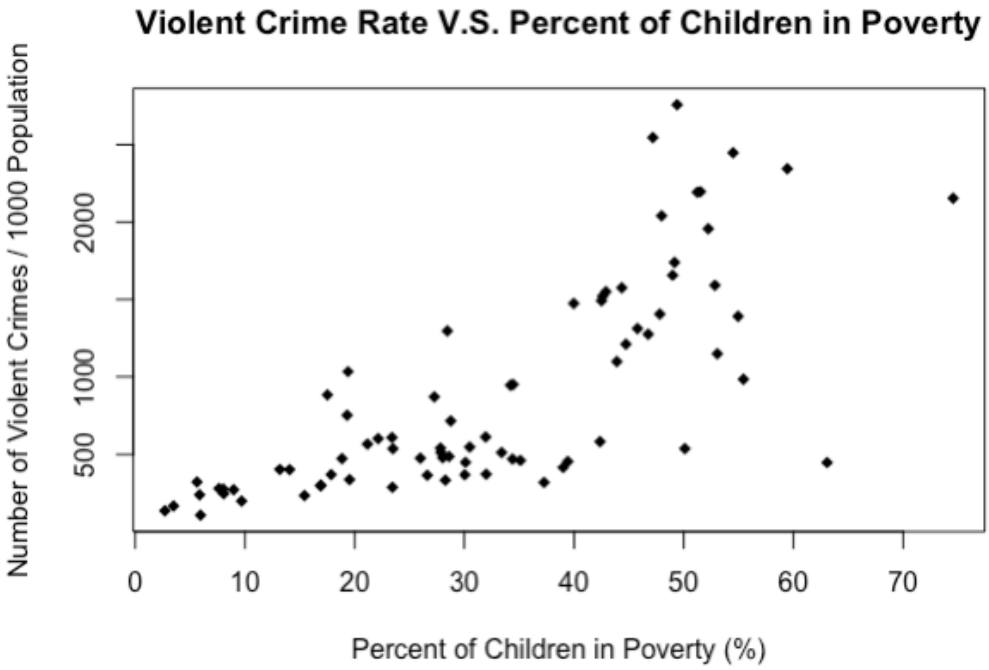

Figure 2. Scatter Plot of Child Poverty Rate versus Class Variable.

6 Numeric Regression

The general hypothesis that will be tested is that the aforementioned city-related predictors can predict violent crime rate well. Table 3, Table 4, and Table 5 illustrate sample data of these predictors and the class variable that will be inputted into the various regression models that will be tested. Table 3 also includes each row's community area number and name. However, these two columns are not inputted into the regression models.

Table 3. Class Variable for Numeric Regression.

Community	Community Number	Violent Crime Rate
ROGERS PARK	1	612.3729
WEST RIDGE	2	318.6595
UPTOWN	3	512.4375
LINCOLN SQUARE	4	297.2172
NORTH CENTER	5	239.0875

Table 4. Predictors for Numeric Regression (Part 1 of 2).

Average School Rating	Average SSL Rating	Total Park Area (m^2)	Number of Hospitals	Teen Mom Rate	Infant Mortality Rate
3.571429	277.6664	287364.9	0	51.61818	6.4
3.5	286.1914	725446.2	0	31.52727	5.1
3.428571	266.0711	1437616	4	51.02727	6.5
4.2	281.7773	419080.8	3	37.98182	3.8
3.428571	282.6642	161026.1	0	37.07273	2.7

Table 5. Predictors for Numeric Regression (Part 2 of 2).

Hispanic	Black	White	Asian	Other	Percent Children in Poverty
0.24	0.24	0.45	0.05	0.03	0.319509
0.2	0.13	0.41	0.21	0.04	0.372624
0.16	0.19	0.51	0.11	0.03	0.278424
0.18	0.06	0.62	0.1	0.04	0.168878
0.11	0.09	0.73	0.04	0.03	0.058669

Three types of regression algorithms were used: ordinary least squares (OLS) linear regression, elastic nets regularized regression, and generalized additive linear model (GAM) regression. Elastic nets is seen as a better version of OLS with the added bonus of feature selection, which is why it is being tested to see if it can provide better predictions. It is a mixture of lasso and ridge regression, both of which penalize the coefficients in regular OLS regression in order to prevent overfitting. Ridge regression is where the square of each coefficient for each explanatory variable is penalized, and lasso regression is where the absolute value of each coefficient for each explanatory variable is penalized. Lasso regression can force coefficients of features uncorrelated with the class variable to be zero (which can be seen positively as feature selection), though this may lead to information loss. Thus, elastic nets are seen as the "best of both worlds".

However, as explained in the previous section of this report, many of the relationships between the explanatory variables and the class variable seem to be nonlinear in nature. Thus, the authors hypothesize GAM may also give fruitful results, since the GAM algorithm attempts to fit nonlinear polynomial splines on the explanatory variables, and then conduct an additive linear regression on these transformed predictors.

10-fold cross validation was used to obtain the optimal alpha and lambda hyperparameters for elastic nets regression. Note that alpha is the mixing parameter with a range from 0 to 1. A value of 0 represents 100% ridge

regression, and a value of 1 represents 100% lasso regression. Also note that lambda is the amount of penalization with increasing coefficients for the explanatory variables in the final model.

10-fold cross validation was also used to obtain three performance metrics for each regression type, which include: root mean squared error, (RMSE), coefficient of determination (R^2), and mean absolute error (MAE).

6.1 OLS Linear Regression

Figure 3 shows the predicted class variables versus the actual class variables for OLS linear regression. Ideally, the points should be uniformly distributed around the line $y = x$. As shown, the predictions have this trend, which means the predictions seem to be roughly accurate. However, the general trend has somewhat of a small nonlinear concave curve.

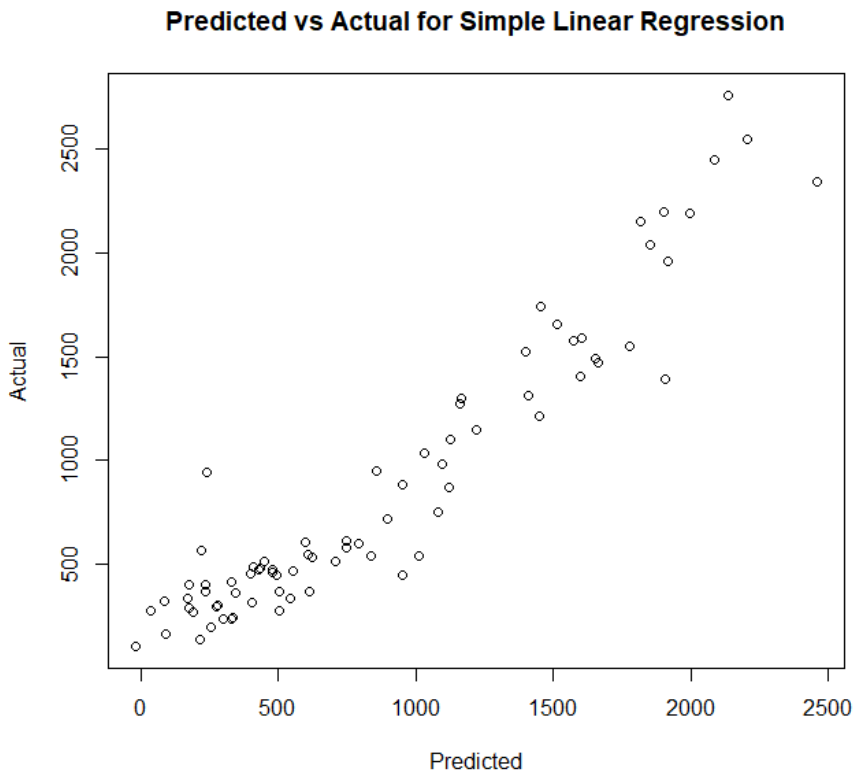

Figure 3. Predicted Class Values vs Actual Values for OLS Linear Regression

This concavity is more pronounced in the residual plot shown in Figure 4. Evidently, this plot shows that the data's uniformity around the line $y = 0$ is not independent of the value of x (the predicted values), which suggests a non-linear regression algorithm may be more suitable.

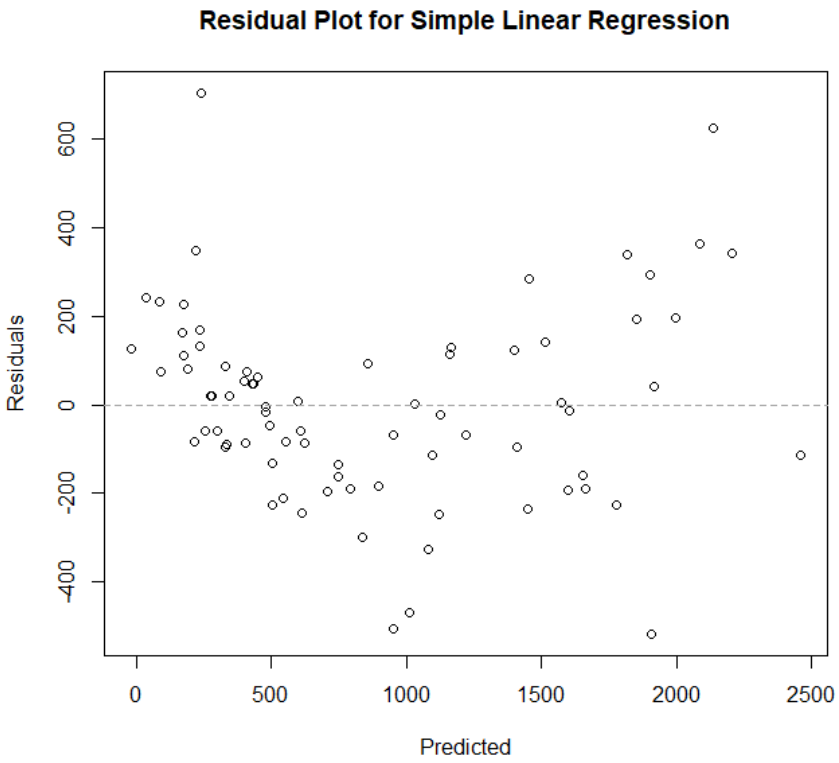

Figure 4. Residual Plot for OLS Linear Regression.

Figure 5 shows the residual histogram for OLS linear regression. While it shows a roughly uniform distribution centred around 0, it does not show the concavity pattern from the previous figure, which is one weakness of histogram plots.

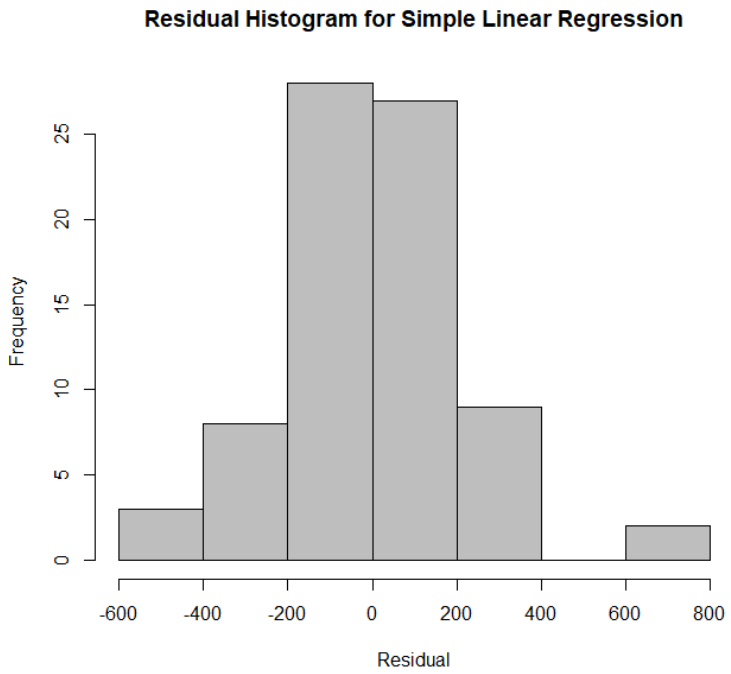

Figure 5. Residual Histogram for OLS Linear Regression.

The first row in Table 6 shows the performance metrics of this algorithm using 10-fold cross validation. With an R^2 of 0.82 using cross validation, it is clear that the general hypothesis is correct: city-related predictors can, in fact, predict the rate of violent crime in a community area. The second row in the same table shows the performance metrics when the model is built using all 77 community area data points. A roughly 0.07 increase in R^2 indicates that OLS does not overfit by much.

Table 6. Performance Metrics for OLS Linear Regression

	RMSE	R^2	MAE
Cross Validation	280.188301046522	0.828664642419824	210.877542633719
Entire Dataset	217.015546090911	0.896063782960191	165.112765159068

The coefficients obtained from OLS are displayed in Table 7. Concordant with exploratory data analysis results, the total park area has a very small coefficient as it did not seem to be correlated with violent crime rate. The percentage of all the different races and child poverty rate have much higher coefficient magnitudes than the other variables. This makes sense though as these predictors were all less than 1 in magnitude due to them being percentages. However, it is surprising to see that that the coefficient for the percentage of black people is not as high as the coefficient for white people. However, this makes sense when considering the limitations of OLS, and by looking at the scatter plots for each versus the class variable. The relationship between the percentage of white people versus the class variable resembles a negative exponential function, which means that when this explanatory variable is very small, the rate of violent crime drastically increases; hence why the coefficient is large. This also suggests a non-linear regression algorithm may be more suitable and can characterize this relationship better than OLS.

Table 7. Coefficients for OLS Linear Regression.

Coefficient Description	Value
Intercept	-4770.107
avgSchoolRating	-52.83907
avgSSLRating	0.6859105
totalParkArea	-0.00004059856
Has3OrMoreHospitals	-21.89742
teenMomRate	10.66917
infantMortalityRate	48.85435
Hispanic	3483.318
black	4396.733
white	4884.87
asian	4459.988
other	-1979.584
percentChildrenInPov	1298.906

6.2 Elastic Net Regression and Feature Selection

Using 10-fold cross validation, the best alpha and lambda values were determined to be 0.6 (60% lasso regression; 40% ridge regression) and 98.693, respectively. Figure 6, Figure 7, and Figure 8 show similar types of plots as ones shown for OLS regression in the last subsection. It seems that there is not a large visual difference between elastic net and OLS, though the histogram distribution slightly changed.

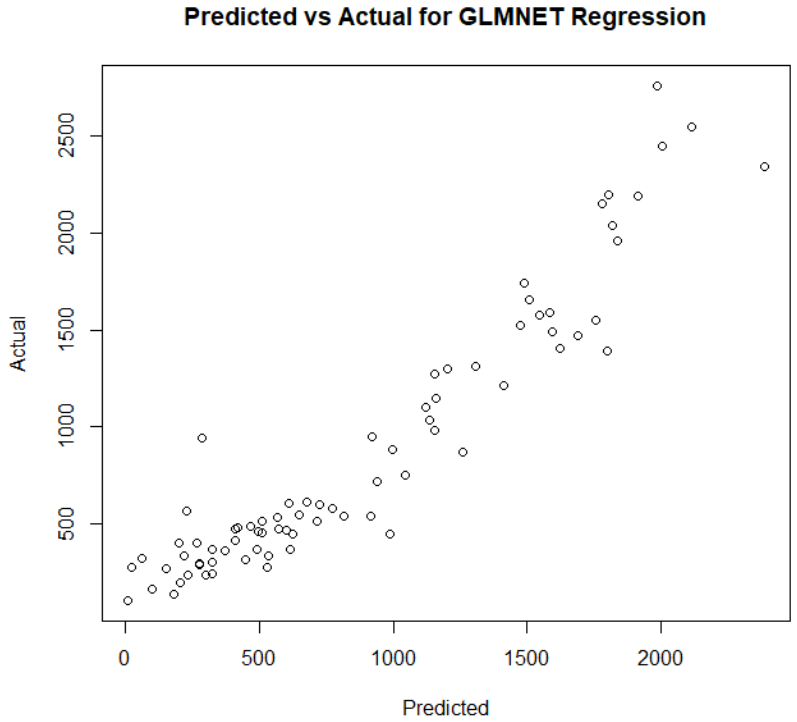

Figure 6. Predicted Class Values vs Actual Values for Elastic Net Regression.

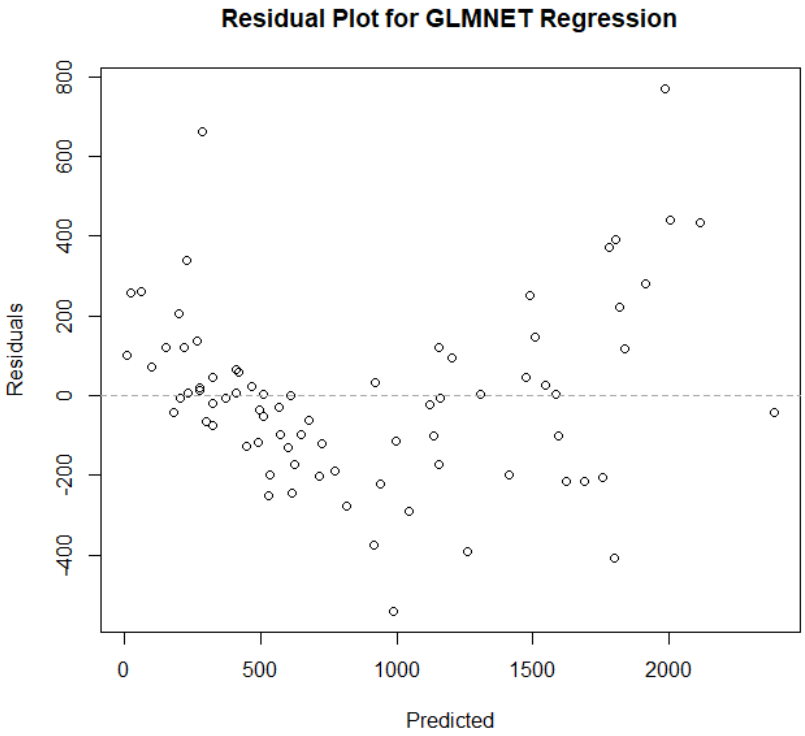

Figure 7. Residual Plot for Elastic Net Regression.

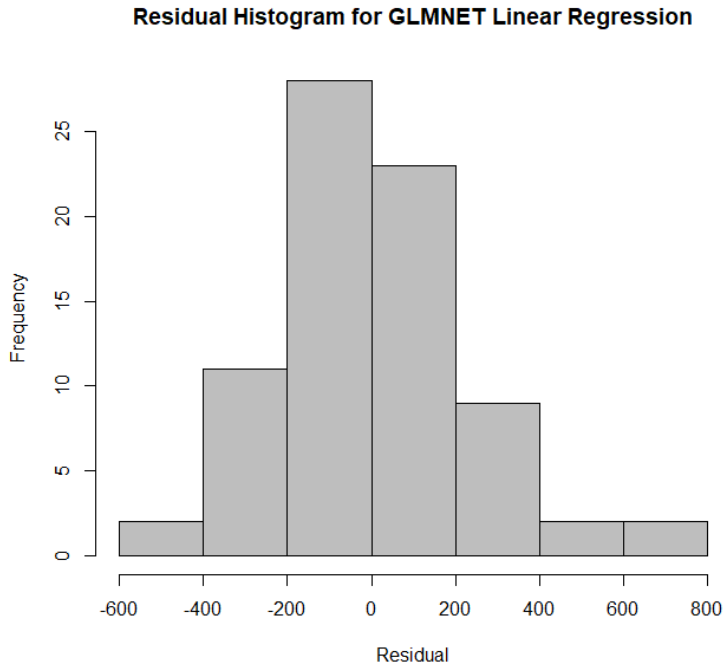

Figure 8. Residual Histogram for Elastic Net Regression.

Ultimately, both rows in Table 8 show that the performance metrics improved slightly from OLS, which is concordant with the authors' earlier hypothesis that elastic nets should perform better. There is also slightly less overfitting when using elastic nets regression, as the difference between the first row and second row in the table is less than the difference seen in the equivalent table for OLS.

Table 8. Performance Metrics for Elastic Net Linear Regression

	RMSE	R^2	MAE
Cross Validation	282.621148416023	0.830021595997032	204.969583822849
Entire Dataset	246.600369438578	0.875267084873369	185.117571612978

Table 9 shows the coefficients of the elastic nets model. These values actually make more tangible sense for humans than the previous OLS model. For example, explanatory variables that did not show a clear correlation with the class variable from exploratory data analysis were removed via elastic nets regression's built-in feature selection via lasso regression. In addition, variables that seem to have a stronger correlation with the class variable have appropriately high coefficient values, such as the percent of children in poverty and the percentage of blacks. This make senses when looking at the slope of the general trends between each explanatory variable and the class variable back in the exploratory data analysis section.

Table 9. Coefficients for Elastic Nets Linear Regression.

Coefficient Description	Value
Intercept	-155.46933
avgSchoolRating	-
avgSSLRating	-
totalParkArea	-

Has3OrMoreHospitals	-
teenMomRate	7.21632
infantMortalityRate	47.57234
Hispanic	-483.31170
black	178.03394
white	-
asian	-
other	-
percentChildrenInPov	804.43955

Note that other differences in magnitudes between, say, teenage pregnancy rate and percent of children in poverty, are magnified due to the difference in units amongst the predictors. To clarify though, fixing and normalizing the units did not make any significant difference in the performance metrics.

6.3 GAM Regression

Using the scatter plots of each explanatory variable versus the class variable in the *Exploratory Data Analysis* section of this report, the relevant explanatory variables that may have nonlinear relationships with the class variable were identified. These include: total park area, birth rate for teenage mothers, infant mortality rate, percentage of Hispanics, percentage of white people, percentage of Asian people, and percentage of children in poverty. Thus, the resulting formula used is the following, where s is a function that fits a nonlinear spline between the input and the outcome:

$$\hat{y} = avgSchoolRating + avgSSLRating + s(totalParkArea) + has3OrMoreHospitals + s(teenMomRate) + s(infantMortalityRate) + s(hispanic) + black + s(white) + s(asian) + other + s(percentChildrenInPoverty)$$

Figure 9 seems to indicate that the GAM model built on all 77 community area data points reduces the problematic nonlinear concavity. In addition, Figure 10 and Figure 11 shows that the range of the residual decreased to $(-300,400)$. These all seem to indicate that the GAM model fits the data better.

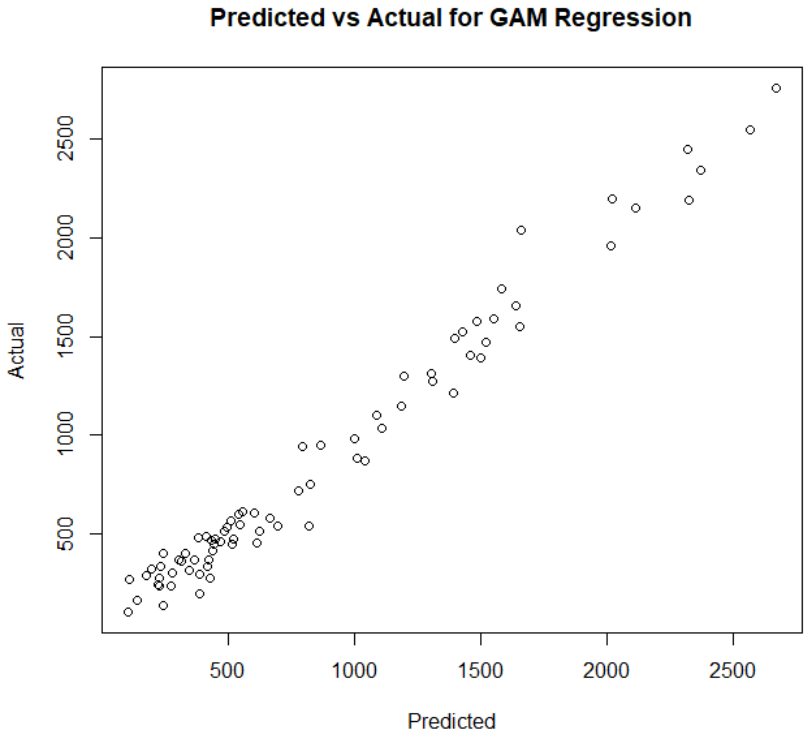

Figure 9. Predicted Class Values vs Actual Values for GAM Regression.

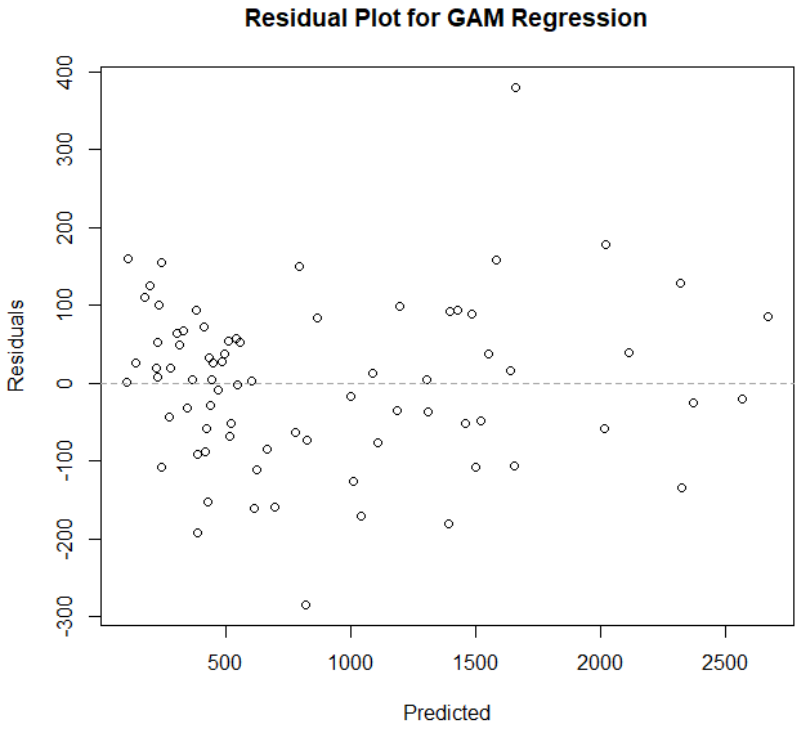

Figure 10. Residual Plot for GAM Regression.

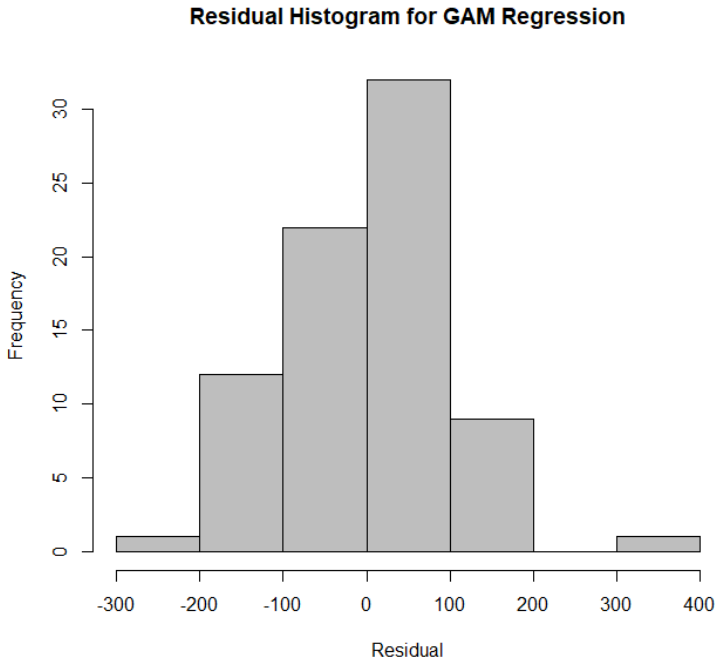

Figure 11. Residual Histogram for GAM Regression.

However, the performance metrics obtained from cross validation shown in the first row of Table 10 indicates that the GAM model performance on unseen data is actually worse than the previous two regression algorithms.

Table 10. Performance Metrics for GAM Regression.

	RMSE	R^2	MAE
Cross Validation	308.050825959162	0.796551460938803	188.925268679306
Entire Dataset	103.30301646473	0.976500401871001	79.55816922112

The second row displays the performance metric of the GAM model built with all 77 community area data points. It is clear that by comparing these two rows that leaving even 7 data points out of the model, which is what 10-fold cross validation does for each fold, significantly changes the performance metrics. Thus, the GAM algorithm heavily overfits. In addition, the fact that GAM uses complex splines to get the best fit for each predictor means that the authors also cannot analyze the exact coefficients of the model.

6.4 Discussion

In conclusion, it seems that out of all three regression algorithms, elastic net regression performs the best, though it is only marginally better than OLS regression. Future recommendations include using simpler nonlinear regression algorithms other than GAM, as its fitted splines may use polynomials with higher than necessary orders, which can cause unnecessary overfitting to occur.

7 Clustering Analysis Results

The k-means clustering analysis via Euclidean distance is performed on the 12 explanatory variables using R Language's function, "kmeans". Since Euclidean distance-based clustering works the best under isotropic conditions, all of the explanatory variables are normalized and scaled to the range of 0 to 1. Table 11 and Table 12 display sample data of these 12 normalized variables.

Table 11. Normalized Explanatory Variables for Clustering (Part 1 of 2).

Violent crime Rate	Average School Rating	Normalized Average SSL	Total Park Area (m^2)	Number of Hospitals	Teen Mom Rate
0.5	0.3048546	0.1232683	0	0.42178033	0.232227488
0.475	0.5289858	0.3160378	0	0.25011651	0.170616114
0.45	0	0.6294148	1	0.41673139	0.236966825
0.72	0.4129330	0.1812274	0.75	0.30526642	0.109004739
0.45	0.4362516	0.0676753	0	0.29749883	0.056872038

Table 12. Normalized Explanatory Variables for Clustering (Part 2 of 2).

Hispanic	Black	White	Asian	Other	Percent Children in Poverty
0.270588	0.247423	0.535714	0.102041	0.6	0.407143
0.223529	0.134021	0.488095	0.428571	0.8	0.481074
0.176471	0.195876	0.607143	0.22449	0.6	0.349956
0.2	0.061856	0.738095	0.204082	0.8	0.197479
0.117647	0.092784	0.869048	0.081633	0.6	0.044078

7.1 Analysis of Number of Centers

For the first part of this analysis, the optimal number of cluster centers from 1 to 25 is analyzed, with the total within-cluster sum of squares data being collected for each. The result is shown in Figure 12.

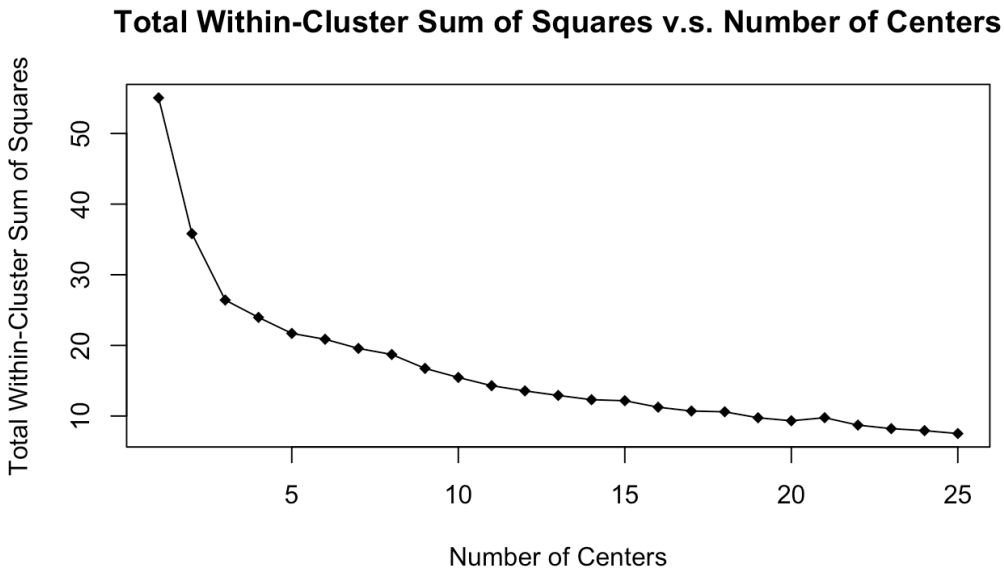

Figure 12. Total within-cluster sum of squares versus the number of centers

As shown in Figure 12, there is a significant drop in the sum of squares until the number of centers reaches 3. The decrease in the sum of squares substantially lowers as the number of centers increases from 3 to 25.

7.2 Analysis of Different Starting Points for Centers

Another analysis is performed on the resulting sum of squares when different starting points are randomly chosen for the clustering center. Two trials are executed, and the result is shown in Figure 13.

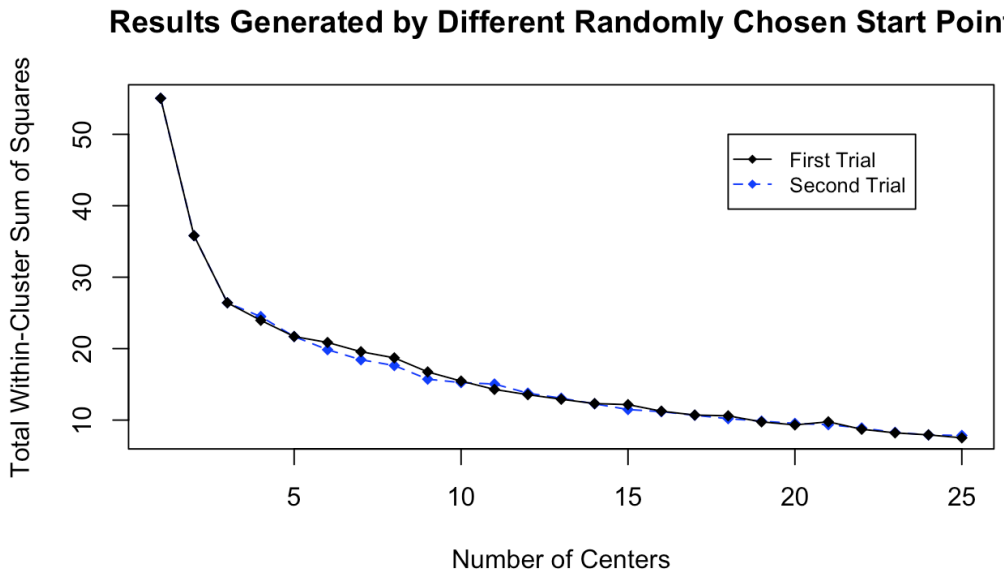

Figure 13. Sum of squares by different randomly chosen start point of center

As shown in the result, different choices of starting points of the clustering centers will result in different sum of squares, which is reasonable due to the randomized nature of the k-means algorithm. Another analysis is then conducted, with 50 different sets of starting points chosen. Figure 14 indicates the average sum of squares obtained from the 50 sets of centers for each number of centers.

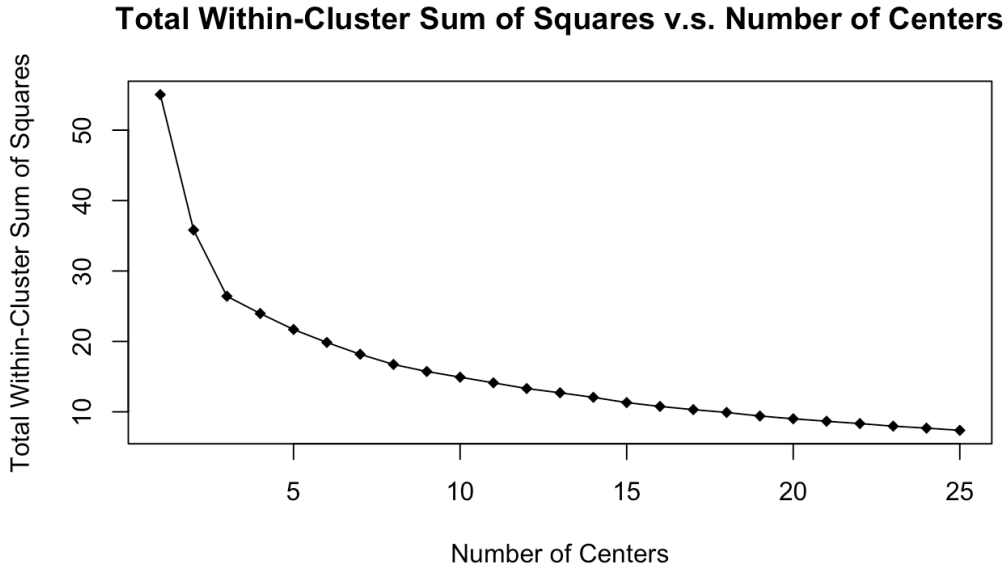

Figure 14. Average sum of squares with 50 sets of starting points

By looking at the above figure, it becomes evident that the significant drop in sum of squares still stops when 3 clustering centers is selected. Therefore, for the clustering analysis, the number of clustering centers is set to 3. The average total within-cluster sum of squares for 3 clustering centers is 26.416911. Table 13 shows the three centroids obtained from averaging the 50 centroids acquired from the 50 different random initial starting points.

Table 13. K-Means Clustering Centroids.

	Centroid 1	Centroid 2	Centroid 3
School Rating	0.4710328	0.54229084	0.45844453
SSL Rating	0.32036502	0.55741364	0.52679244
Park Area	0.20544814	0.12255941	0.19737737
Num. Hospital	0.24107143	0.04761905	0.09821429
Teen Mom	0.24212283	0.57094032	0.68752913
Infant Mort.	0.18009479	0.25321598	0.55196344
Hispanic	0.20546218	0.71876751	0.09411765
Black	0.13291605	0.13451154	0.82916053
White	0.66071429	0.22959184	0.0922619
Asian	0.21428571	0.09232264	0.03206997

Other	0.50714286	0.22857143	0.27142857
Child Poverty	0.22904627	0.43306172	0.57932275

It can be seen from the above table that the cluster centroids of many of the explanatory variables are close together, indicating that the data is heavily concentrated in a small region or that variance is low in these dimensions. These explanatory variables include: school rating, SSL, total park area, and number of hospitals. In contrast, the cluster centers of other explanatory variables are more useful because they are further apart, which suggests natural clusters in the data for these dimensions. For example, there are cluster centroids at low, medium, and high values of teenage pregnancy rate and infant mortality rate.

7.3 Discussion

After the clustering centroids are determined, a plot is made to indicate the clustering relationship between each two explanatory variables from these clustering centroids, as shown in Figure 15.

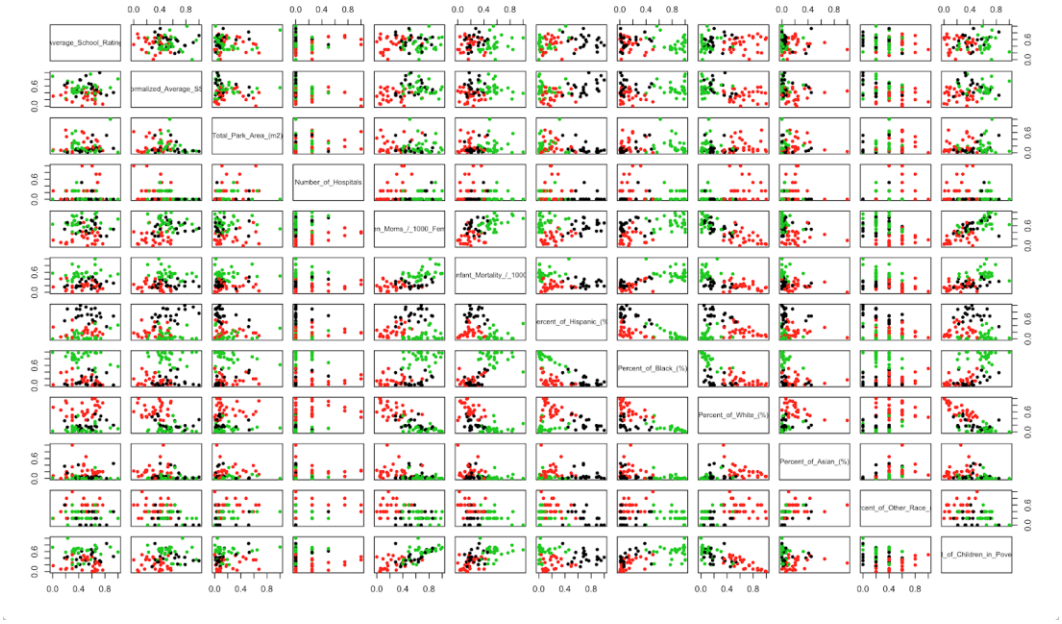

Figure 15. Clustering Relationship between Each Two Explanatory Variables

Figure 15 shows that most of the plots are cleanly divided into the three clusters from the k-means algorithm, specifically at columns 5, 6, 7, 8, 9, and 12, which represent teen pregnancy rate, infant mortality rate, percent of Hispanic, percent of black, percent of white, and percent of children in poverty, respectively. For example, it can be observed from the plot in column 5 and row 6 (the plot which shows the clustering between teenage mom rate and infant mortality rate), that low teenage mom rate seems to be associated with low infant mortality rate, and high teenage mom rate is associated with high infant mortality rate.

Since the plots in Figure 15 are small and thus hard to interpret, the clustering relationship between specific and noteworthy variables will be discussed and visualized. For example, a 3D plot between teen pregnancy rate, infant mortality rate, and child poverty rate is made, where each data point is colored based on which of the 3 clusters it is in. This plot is shown below in Figure 16.

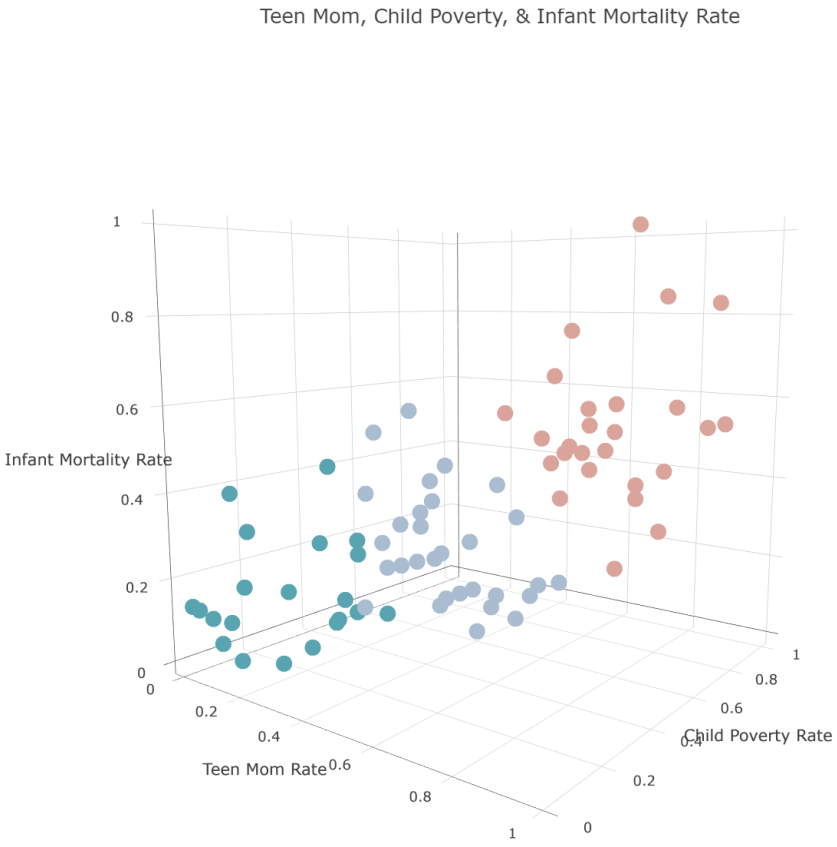

Figure 16. 3D Cluster Plot of Teenage Pregnancy, Infant Mortality, and Child Poverty Rate.

As clearly shown above, there is a cluster when all three values are low, when all three are medium-range, and when all are high. This seems to make sense, as a higher rate of all of them generally indicate poor public health in a community, while a lower rate of all of them indicate good public health. This 3D scatter plot is re-plotted as a choropleth in Figure 17, in which the color now represents the rate of violent crimes. It is evident that by comparing the two plots, the violent crime rate is naturally associated with the clusters formed by k-means. To summarize, when a community area has high violent crime, it also generally has

poor public health. Conversely, when a community area has low violent crime, it also generally has good public health.

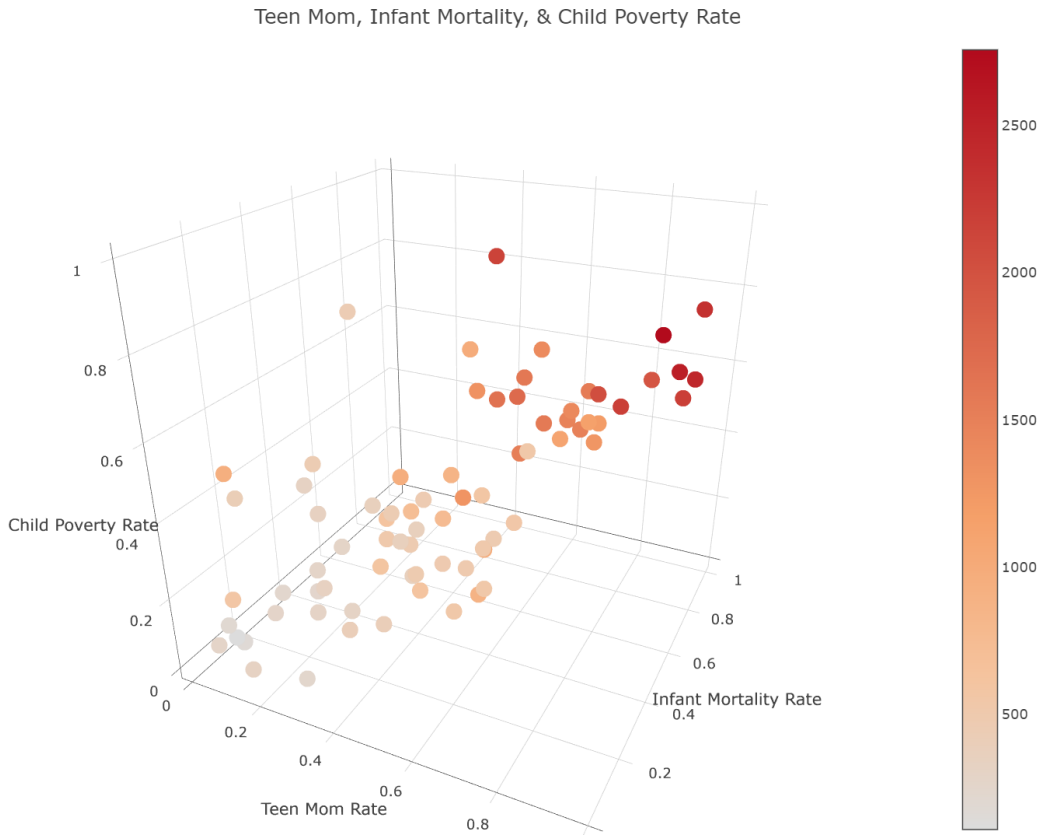

Figure 17. 3D Choropleth Plot of Teenage Pregnancy, Infant Mortality, and Child Poverty Rate.

A second noteworthy example of the clustering results includes the percentage of Hispanic, black, and white people, as shown in Figure 18. Clearly, the clusters indicate that when there is a high percentage of one race, there is a decrease in the proportion of other races.

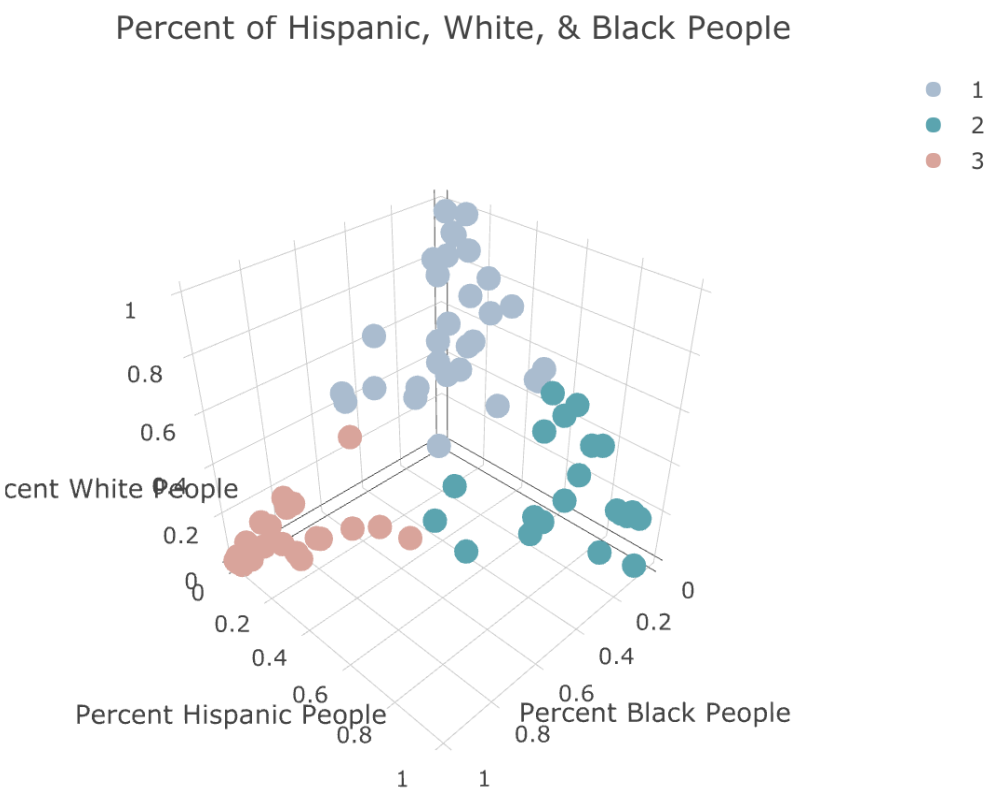

Figure 18. 3D Cluster Plot of Percentage of Hispanic, White, and Black People.

However, an interesting result can be seen when comparing this plot with the choropleth version of it in Figure 19, where the intensity of color indicates violent crime rate.

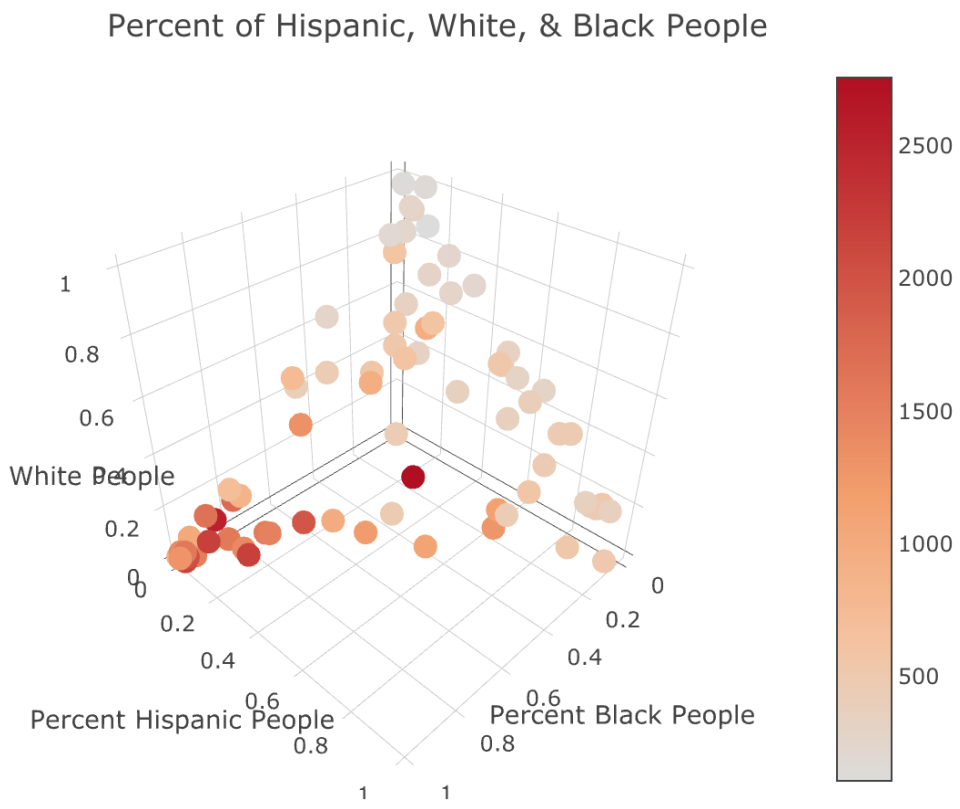

Figure 19. 3D Choropleth Plot of Percentage of Hispanic, White, and Black People.

As shown in the two figures, the region of communities that have high violent crime rates on the choropleth correspond to the cluster with pink points, which represent communities where there is a high percentage of black people. This also serves as an example of how the clustering results are providing appropriate clusters that are consistent with previously discussed exploratory data analysis and regression results.

8 Association Rule Mining Analysis

The Apriori algorithm, an algorithm for association rule mining, is applied to both the class variable and the explanatory variables to obtain the most frequent item sets in the collected data. Since most of the explanatory variables are numerical and continuous, data binning is required to divide the numerical values into buckets.

8.1 Binning

8.1.1 Number of Bins

To determine the optimal number of bins efficiently for each explanatory variable, the Freedman–Diaconis rule is utilized, which is a commonly used rule to select an appropriate number of bins based on the size and variance of the data (specifically, the interquartile range and number of data points) [18]. The equation for the rule is

$$bin\ size = \frac{2IQR(x)}{\sqrt[3]{n}}$$

where IQR(x) is the interquartile range of the data and n is number of observations of sample data x. To get a wholistic result of the association rules that can be obtained from the data, two different binning methods are tested and compared: fixed width and adaptive width binning. Note that although the Freedman-Diaconis rule is designated for fixed width binning, for consistency, the same numbers of bins are used for each variable in both methods.

8.1.2 Fixed Width Binning

Fixed width binning divides data such that all bins correspond to intervals of the same size. The advantages of this strategy include that it is simple to implement, and that it produces a straight-forward and reasonable abstraction of data. On the other hand, this method is unsupervised, and as a consequence, it is hard to know just from the result of this binning method if the data is divided in a 'proper' way.

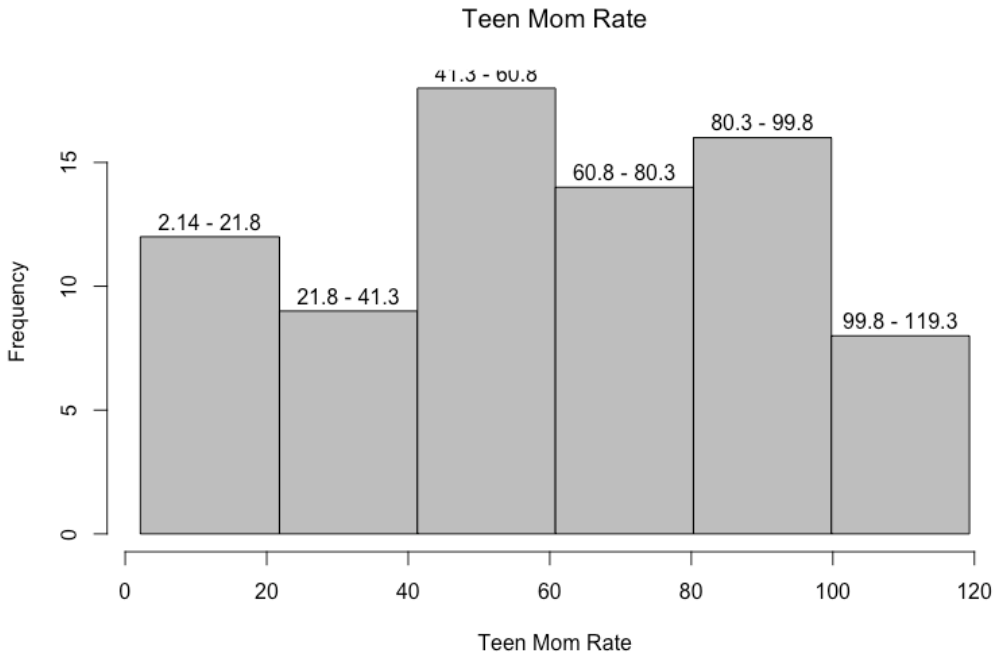

Figure 20 shows an example of the resulting bins for one of the explanatory variables, teenage pregnancy rate, using fixed width binning. For comparison, Figure 21 shows the scatter plot of this variable versus the class variable. Clearly, the number of bins and choice of divisions between the bins seem to make sense, based on the scatter plot.

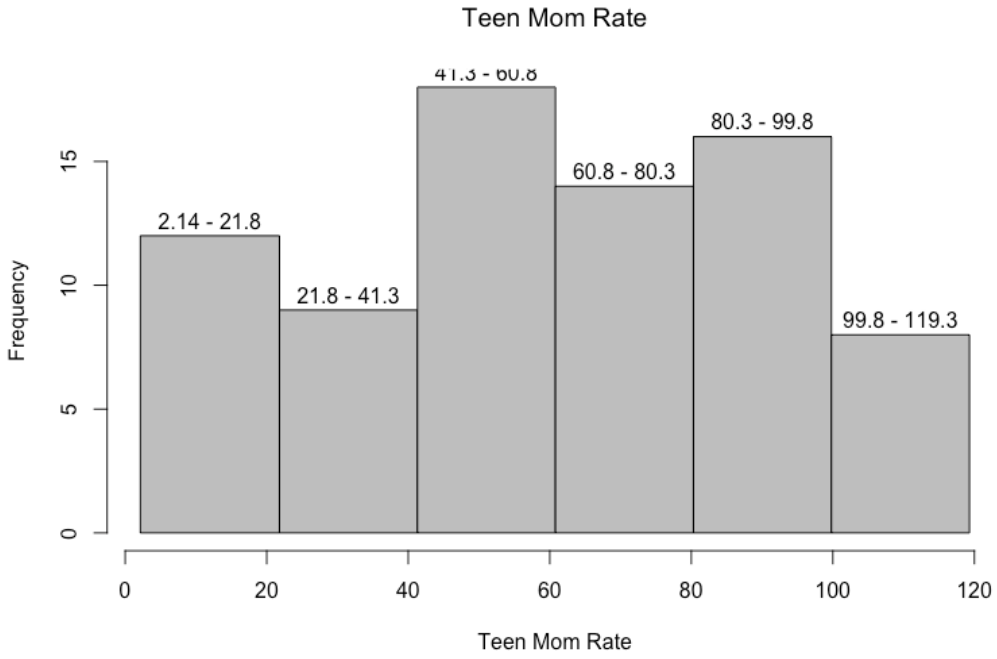

Figure 20. Histogram of Teen Pregnancy Rate with Fixed Binning Method.

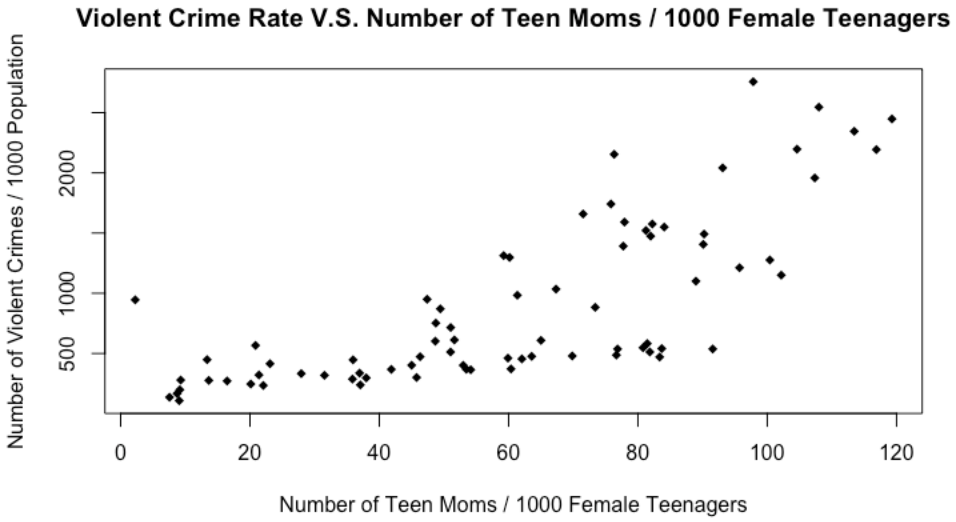

Figure 21. Scatter Plot of Teen Pregnancy Rate versus Violent Crime Rate.

8.1.3 Adaptive Width Binning

Adaptive width binning divides data into intervals of equal number of samples via the interquartile range of the data. Depending on the data distribution, this

strategy may give a better classification of data. However, there are several pitfalls of this method. Uniformly distributing the number of samples in each interval may lead to an over-weighting of outliers, and data points with very similar values may fall in different groups [19].

Similar to the previously shown histogram, Figure 22 shows the histogram of teen pregnancy rate for the adaptive binning method. As clearly seen, this method tries to make the number of samples more uniformly distributed between all the bins.

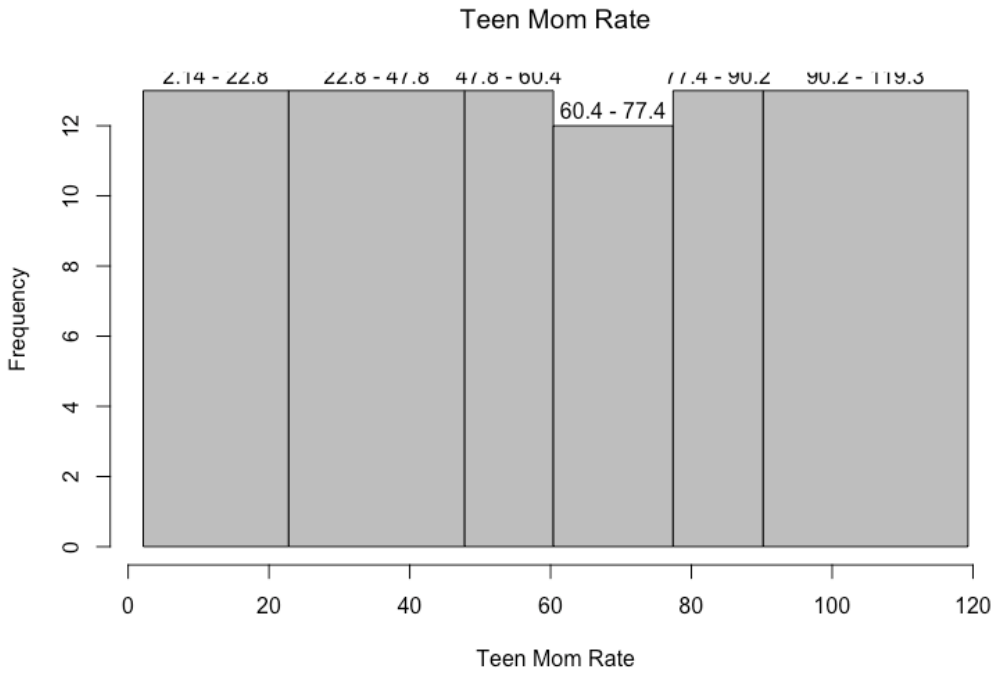

Figure 22. Histogram of Teen Pregnancy Rate with Adaptive Binning Method.

8.2 Results

Since the dataset only contains 77 samples (77 community areas), in order to obtain rules that are more general, a higher support is required compared to larger datasets. The starting point is set to be 0.5 support and 0.8 confidence. The

support value is lowered by steps of 0.1 or 0.05 through the process, to help with the selection of a proper support value used in the final analysis. The confidence is unchanged through the process since it is decided that 80% is the lowest acceptable confidence, due to the same previous reason of the number of samples being relatively small. Table 14 and Table 15 show sample data using the fixed width binning method.

Table 14. Binned Explanatory Variables for Association (Part 1 of 2).

Violent crime Rate	Average School Rating	Normalized Average SSL	Total Park Area (m^2)	Number of Hospitals	Teen Mom Rate	Infant Mortality Rate
crimes (550,991]	school (3.57,4.05]	SSL (276,280]	park (1.97e+05, 3.86e+05]	hospital 0	teenMoms (41.3,60.8]	infactMortality (5.02,6.78]
crimes (105,550]	school (3.57,4.05]	SSL (285,290]	park (5.75e+05, 7.65e+05]	hospital 0	teenMoms (21.8,41.3]	infactMortality (5.02,6.78]
crimes (105,550]	school (3.1,3.57]	SSL (266,271]	park (1.33e+06, 1.52e+06]	hospital 4	teenMoms (41.3,60.8]	infactMortality (5.02,6.78]
crimes (105,550]	school (4.05,4.52]	SSL (280,285]	park (3.86e+05, 5.75e+05]	hospital 3	teenMoms (21.8,41.3]	infactMortality (3.26,5.02]
crimes (105,550]	school (3.1,3.57]	SSL (280,285]	park (4.96e+03, 1.97e+05]	hospital 0	teenMoms (21.8,41.3]	infactMortality (1.48,3.26]

Table 15. Binned Explanatory Variables for Association (Part 2 of 2).

Hispanic	Black	White	Asian	Other	Percent Children in Poverty
hispanic (18,35]	black (0,48.5]	white (33.6,50.4]	asian (4.9,9.8]	other 3	childPoverty (29.6,38.6]
hispanic (18,35]	black (0,48.5]	white (33.6,50.4]	asian (19.6,24.5]	other 4	childPoverty (29.6,38.6]
hispanic (0.915,18]	black (0,48.5]	white (50.4,67.2]	asian (9.8,14.7]	other 3	childPoverty (20.7,29.6]
hispanic (0.915,18]	black (0,48.5]	white (50.4,67.2]	asian (9.8,14.7]	other 4	childPoverty (11.7,20.7]

hispanic (0.915,18]	black (0,48.5]	white (67.2,84.1]	asian (0,4.9]	other 3	childPoverty (2.63,11.7]

8.2.1 Fixed Width Binning

With the fix width binning method, every itemset with higher than 0.5 support has only one item, so the support was then lowered. Table 16 shows the result generated by Apriori algorithm with 0.4 support and 0.8 confidence.

Table 16. Association Rules with Fixed Width Binning (0.4 Support and 0.8 Confidence).

	Antecedents	**Consequents**	**Support**	**Confidence**
1	{crimes(105,550]}	{black(0,48.5]}	0.487	0.974
2	{white(0,16.8]}	{asian(0,4.9]}	0.436	0.895

The first association rule in Table 16 indicates that communities that have low violent crime rates also a have low percentage of black people. The second rule indicates that communities with a low percentage of white people also have a low percentage of Asian people.

As the support value is decreased to 0.3, the following list in Table 17 is appended to the previous list in Table 16.

Table 17. Association Rules with Fixed Width Binning (0.3 Support and 0.8 Confidence).

	Antecedents	Consequents	Support	Confidence
1	{hospital 0, crimes (105,550]}	{black (-0.097,48.5]}	0.321	0.962
2	{black (48.5,97.1]}	{hispanic (0.915,18]}	0.321	0.893
3	{black (48.5,97.1]}	{white (0,16.8]}	0.321	0.893
4	{hospital 0, black (0,48.5]}	{crimes (105,550]}	0.321	0.806

5	{other 1}	{asian (0,4.9]}	0.308	0.923
6	{black (48.5,97.1]}	{asian (0,4.9]}	0.308	0.857

The appended list illustrates the relationship between the percentages of races in communities, such as the fact that communities with a high percentage of black people usually have a low percentage of Hispanic, white and Asian people. These particular association rules are similar to the clustering results of different races found in the previous section of this report. Note that the rules with 0 hospitals do not indicate much, since around two thirds of the communities have no hospitals.

8.2.2 Adaptive Width Binning

With the adaptive width binning method, every itemset with a support of higher than 0.3 has only one item. Table 18 shows the result of using adaptive width binning with a decreased support value of 0.2 and a confidence of 0.8, such that there is more than one item in each itemset.

Table 18. Association Rules with Adaptive Width Binning (0.2 Support and 0.8 Confidence).

	Antecedents	Consequents	Support	Confidence
1	{asian (0, 4.9]}	{black (24,97.1]}	0.231	0.900
2	{hispanic (0.915,4]}	{black (24,97.1]}	0.205	0.941
3	{white (0,5]}	{black (24,97.1]}	0.205	0.941

Unlike the result obtained from the fixed width binning method, the association rules with high support from the adaptive width binning method do not involve the class variable. However, similar to the result obtained from the fixed width method, the adaptive binning method also reveals a relationship between the

percentages of races – that there is usually a high percentage of black people if there is a low percentage of Asian, Hispanic, or white people.

The support is decreased to 0.15 and the following list in Table 19 is appended to the previous table.

Table 19. Association Rules with Adaptive Width Binning (0.15 Support and 0.8 Confidence).

	Antecedents	Consequents	Support	Confidence
1	{hispanic (53.8,86.1]}	{black (0,24]}	0.192	0.938
2	{white (50.8,84.1]}	{black (0,24]}	0.192	0.938
3	{hospital 0, hispanic (53.8,86.1]}	{black (0,24]}	0.179	1.000
4	{black (0,24], hispanic (53.8,86.1]}	{hospital 0}	0.179	0.933
5	{hispanic (53.8,86.1]}	{hospital 0}	0.179	0.875
6	{hispanic (53.8,86.1]}	{hospital 0, black (0,24]}	0.179	0.875
7	{white (0,5]}	{asian (0,4.9]}	0.179	0.824
8	{crimes (1560,2760]}	{black (24,97.1]}	0.167	1.000
9	{white (13,30.6]}	{hospital 0}	0.167	1.000
10	{white (0,5], asian (0,4.9]}	{black (24,97.1]}	0.167	0.929
11	{hospital 0, asian (0,4.9]}	{black (24,97.1]}	0.167	0.867
12	{white (0,5], black (24,97.1]}	{asian (0,4.9]}	0.167	0.813
13	{crimes (105,313]}	{black (0,24]}	0.154	0.923
14	{crimes (451,548]}	{black (0,24]}	0.154	0.923
15	{other 3}	{black (0,24]}	0.154	0.857

As explained in the previous section, association rules with item “hospital 0” can be ignored in this analysis since most of the communities do not have a hospital. Row 8, 13, and 14 in Table 19 indicate a relationship between the class variable and the percentage of black people. To conclude, communities that have high violent crime rates also have a high percentage of black people. Consequently,

communities that have relatively low violent crime rates also have a low percentage of black people. The rest of the rules all pertain to the percentage of races, most of which illustrate the segregation between the percentage of black people and all the other races. In other words, when there is a low percentage of Hispanic, Asian, and white people, there is usually a high percentage of black people.

8.3 Discussion

Fixed width binning provides better rules in this case, as association rules with higher support are obtained with the Apriori algorithm, which means that the generated rules are more general for the obtained community data.

Results from both strategies indicate that there is a strong relationship between the class variable and the percentage of black people. Both include rules that indicate ‘where there is low violent crime rate, there is a low percentage of black people’, while adaptive width binning also includes the rule of ‘where there is a high violent crime rate, there is a high percentage of black people’. In addition, results from both binning methods reveal a zero-sum-like relationship between the percentage of black people and the percentage of all other races in a community.

9 Conclusion

A data science approach was conducted in order to see if various city-related predictors were related with one another, and if they could help predict the rate of violent crime for a specified community area in Chicago. Results of exploratory data analysis show that indeed, many of these variables are correlated with the rate of violent crime, such as: teenage pregnancy rate, infant mortality rate, percentage of children in poverty, and finally, the percentage of black people, Hispanics, whites, and Asians.

Regression, clustering analysis, and association rule mining were conducted. With both ordinary least squares and elastic nets linear regression, a relatively high R^2 value of 0.82 was achieved, which suggests that these predictors can effectively predict violent crime rate – a result useful for both city officials and the police department. It was also discovered that OLS performed almost as well as elastic nets, despite it being the simpler algorithm. In addition, despite the fact that exploratory data analysis indicated that some of the predictors had a nonlinear relationship with the class variable, it was discovered that using one nonlinear regression called GAM caused massive overfitting to occur. This suggests that an algorithm to conduct nonlinear regression with simpler polynomials may be more effective, and can be seen as possible future improvement for this project.

By using the k-means algorithm and looking at the corresponding elbow curve, it seems clear that the data is naturally grouped into 3 clusters, and that these

clusters generally seem to make sense for some of the explanatory variables. For example, low teen birth rate is clustered with low infant mortality rate and low poverty rate; the vice versa is also clustered together.

Finally, association rule mining revealed that a low percentage of black people in a community area is highly correlated with a low rate of violent crime. The vice versa is also highly associated with each other. In addition, it was discovered that a high percentage of black people is associated with a low percentage of all other races. This latter finding may be correlated with various factors such as the historical segregation of African American populations in America – however, there is the obvious need for further research to be able to make such a claim.

10 Bibliography

[1] J. Sanburn, "Chicago Is Responsible for Almost Half of the Increase in U.S. Homicides," Time, 19 September 2016. [Online]. Available: http://time.com/4497814/chicago-murder-rate-u-s-crime/. [Accessed 23 September 2018].

[2] J. Gorner, P. Nickeas and E. Malagon, "August most violent month in Chicago in nearly 20 years," Chicago Tribune, 29 August 2016. [Online]. Available: http://www.chicagotribune.com/news/local/breaking/ct-august-most-violent-shootings-chicago-20160829-story.html. [Accessed 23 September 2018].

[3] Newberry Library, "Chicago's Community Areas," The Encyclopedia of Chicago, [Online]. Available: http://www.encyclopedia.chicagohistory.org/pages/1760.html. [Accessed 23 September 2018].

[4] City of Chicago, "Strategic Subject List," 7 December 2017. [Online]. Available: https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np. [Accessed 23 September 2018].

[5] L. Lochner and E. Moretti, "The effect of education on crime: Evidence from prison inmates, arrests, and self-reports," *American economic review*, vol. 94, no. 1, pp. 155-189, 2004.

[6] T. Schusler, L. Weiss, D. Treering and E. Balderama, "Research note: Examining the association between tree canopy, parks and crime in Chicago," *Landscape and Urban Planning*, vol. 170, pp. 309-313, 2018.

[7] P. B. Stretesky and a. M. J, "The relationship between lead and crime," *Journal of Health and Social Behavior*, vol. 45, no. 2, pp. 214-229, 2004.

[8] City of Chicago, "Crimes - 2001 to present," [Online]. Available: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2. [Accessed 23 September 2018].

[9] City of Chicago, "Chicago Public Schools - School Profile Information SY1718," 19 October 2018. [Online]. Available: https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Profile-Information-/w4qj-h7bg. [Accessed 26 October 2018].

[10] Illinois Action for Children, "Population and Poverty Data by Chicago Community Area," [Online]. Available: http://www.actforchildren.org/wp-content/uploads/2018/01/Census-Data-by-Chicago-Community-Area-2017.pdf. [Accessed 23 September 2018].

[11] City of Chicago, "Parks - Chicago Park District Park Boundaries (current)," 28 September 2018. [Online]. Available: https://data.cityofchicago.org/Parks-Recreation/Parks-Chicago-Park-District-Park-Boundaries-curren/ej32-qgdr. [Accessed 26 October 2018].

[12] City of Chicago, "Boundaries - Community Areas (current)," 11 July 2018. [Online]. Available: https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6. [Accessed 26 October 2018].

[13] City of Chicago, "Hospitals - Chicago," 28 August 2011. [Online]. Available: https://data.cityofchicago.org/Health-Human-Services/Hospitals-Chicago/ucpz-2r55.

[Accessed 4 November 2018].

[14] City of Chicago, "Public Health Statistics - Births to mothers aged 15-19 years old in Chicago, by year, 1999-2009," 28 March 2013. [Online]. Available: https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Births-to-mothers-aged-15/9kva-bt6k. [Accessed 4 November 2018].

[15] City of Chicago, "Public Health Statistics- Infant mortality in Chicago, 2005– 2009," 11 April 2014. [Online]. Available: https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Infant-mortality-in-Chica/bfhr-4ckq. [Accessed 4 November 2018].

[16] C. Taylor, "How to Calculate Prevelance Rates Per Thousand," 13 March 2018. [Online]. Available: https://sciencing.com/calculate-prevalence-rates-per-thousand-7533277.html. [Accessed 17 November 2018].

[17] Chicago Public Schools, "School Quality Rating Policy," 25 October 2018. [Online]. Available: https://cps.edu/Performance/Pages/PerformancePolicy.aspx. [Accessed 23 September 2018].

[18] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L 2 theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete,* vol. 57, no. 4, pp. 453-476, 1981.

[19] City of Chicago, "Parks - Chicago Park District Park Boundaries (current)," [Online]. Available: https://data.cityofchicago.org/Parks-Recreation/Parks-Chicago-Park-District-Park-Boundaries-curren/ej32-qgdr. [Accessed 23 September 2018].

[20] C. Schnell, A. A. Braga and E. L. Piza, "The Influence of Community Areas, Neighborhood Clusters, and Street Segments on the Spatial Variability of Violent Crime in Chicago," *Journal of quantitative criminology,* vol. 33, no. 3, pp. 469-496, 2017.

[21] R. Kaur and S. S. Sehra, "Analyzing and Displaying of Crime Hotspots," *International Journal of Computer Applications,* vol. 103, no. 1, pp. 25-28, 2014.

[22] D. McDowall, C. Loftin and M. Pate, "Seasonal cycles in crime, and their variability," *Journal of Quantitative Criminology,* vol. 28, no. 3, pp. 389-410, 2012.

[23] P. J. Brantingham and P. L. Brantingham, Patterns in crime, New York: Macmillan, 1984.

11 Appendix

11.1 Data Collection and Preprocessing Method for Predictors

11.1.1 Average School Rating

The average school rating is the average rating of schools in a certain community area, and it is obtained by formula:

Sum(schoolRatingForCommunityArea) / numSchoolsInCommunityArea

The ratings for each school are from dataset, “Chicago Public Schools - School Profile Information SY1718” [9]. In this dataset, the general information about schools is given, such as names, location, ratings, and student count. The strings representing school levels in the “Overall_Rating” column is translated to numerical scores from 1 to 5 according to the level, where 1 is the worst and 5 is the best.

11.1.2 Average SSL Rating

Recall that the SSL is defined as a numerical score with a range of 0 to 500, representing the likelihood of an offender to be involved in a shooting in the near future. 0 is extremely low risk and 500 is extremely high risk. Thus, the average SSL rating predictor is simply calculated as the average SSL rating of all strategic subject people in a certain community area. The SSL ratings are from the dataset, “Strategic Subject List” [4], which takes samples from the list of arrest data from August 1, 2012 to July 31, 2016.

11.1.3 Total Park Area

To obtain the total park area for each community area, all park shape files [11] and all community area shape files [12] were obtained. Then, using the Raster library in R, the intersection area of each park with each community area is calculated. These intersection areas are then summated for each community area.

11.1.4 Number of Hospitals

Obtaining the number of hospitals for each community area from the hospital data [13] was simple, since each hospital data point included the community area it is located in.

11.1.5 Birth Rate by Teenage Mothers

The dataset of birth rate by teenage mothers has all the rates from 1999 to 2009 [14]. The average of all these birth rates for each community areas is used to try to filter out birthrates which may be outliers.

11.1.6 Infant Mortality Rate

Similar to teenage mother birth rate, infant mortality rate for each community area is calculated as the average infant mortality rate from all years in which data was available: from 2005 to 2009 [15]. Two values for one community area from two years had null values, so the average of the non-null values was calculated.

11.1.7 Proportion of Hispanic People

Nothing was needed to be transformed for this predictor [10].

11.1.8 Proportion of Black People

Nothing was needed to be transformed for this predictor [10].

11.1.9 Proportion of White People
Nothing was needed to be transformed for this predictor [10].

11.1.10 Proportion of Asians
Nothing was needed to be transformed for this predictor [10].

11.1.11 Proportion of Other Races
Nothing was needed to be transformed for this predictor [10].

11.1.12 Percent of Children in Poverty
Note that actual poverty rate was unable to be obtained, but percentage of children in poverty was easily obtainable and thus is used instead.

The original dataset of children poverty rates in 2018 has children separated between ages 0 to 5 and ages 6 to 12, and each of these two groups had a poverty rate percentage [10]. A weighted average based on the total population of these two groups was used to obtain the average poverty rate of all children across these two age ranges. This was done in Excel rather than in R, so no code for this data preprocessing exists in the Appendix section.

11.2 Exploratory Data Analysis Plots and Detailed Comments

Note that, when appropriate, some of the scatter plots and bar plots in this section are normalized and thus are percentage-based plots.

11.2.1 Class Variable

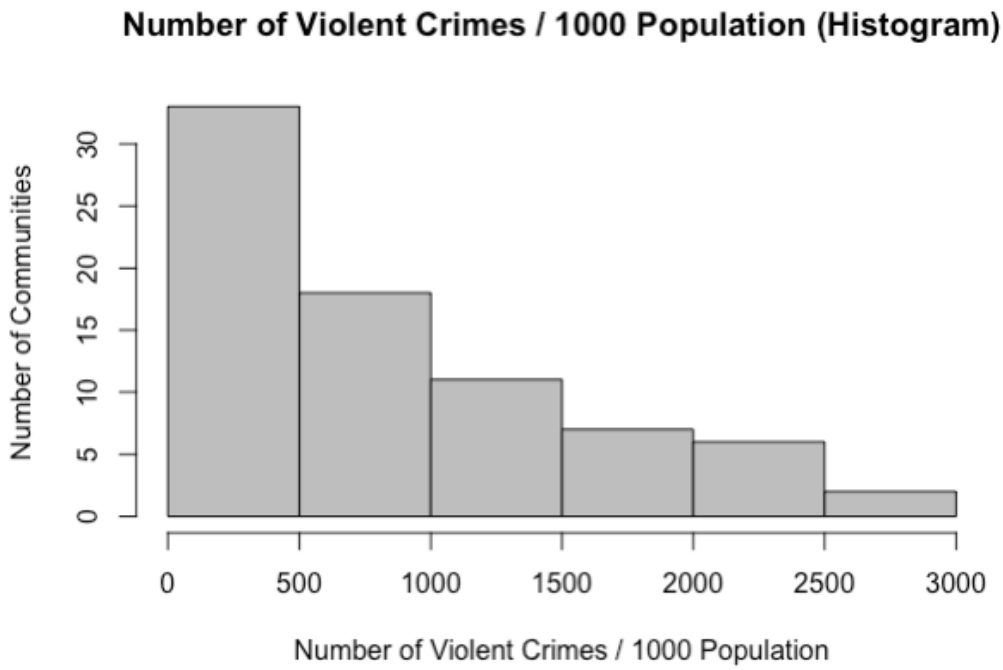

Figure 23. Histogram of Class Variable

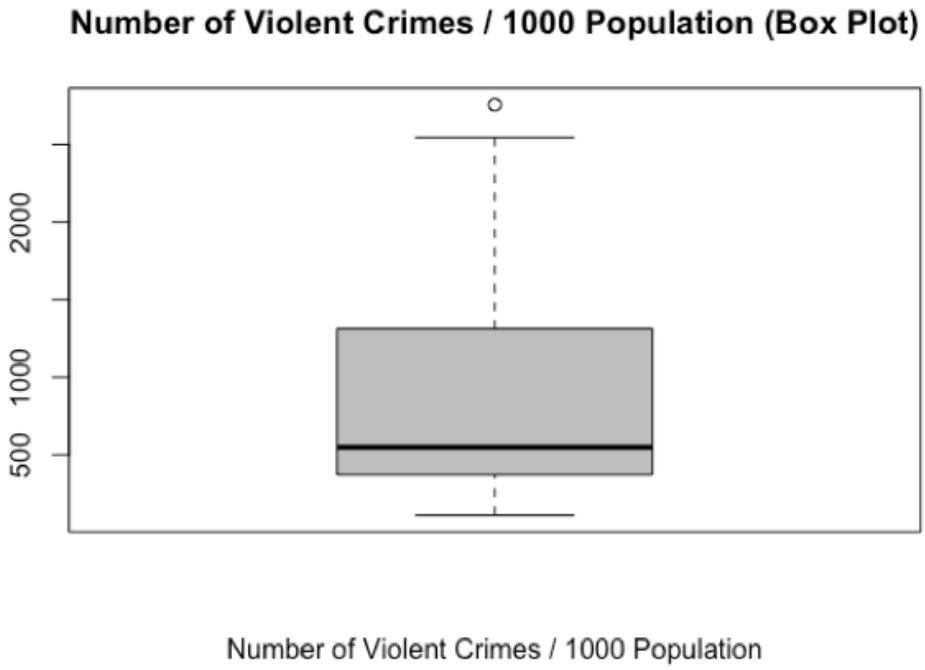

Figure 24. Box Plot of Class Variable

From the histogram, the number of communities exponentially decreases from low to high number of violent crimes. About half of the communities have under 500 cumulative violent crimes per 1000 people since 2001. Two communities, Washington Park and Fuller Park, have more than 2500 cumulative violent crimes per 1000 people. The only one outlier shown in the box plot is also Fuller Park, with 2757 calculated cumulative violent crimes per 1000 people and only 2876 population in 2010. This explains why it does not take too many criminal acts to boost its violent crime rate and why the community area is frequently in the news as one of the most dangerous places to live in Chicago.

11.2.2 Average School Rating

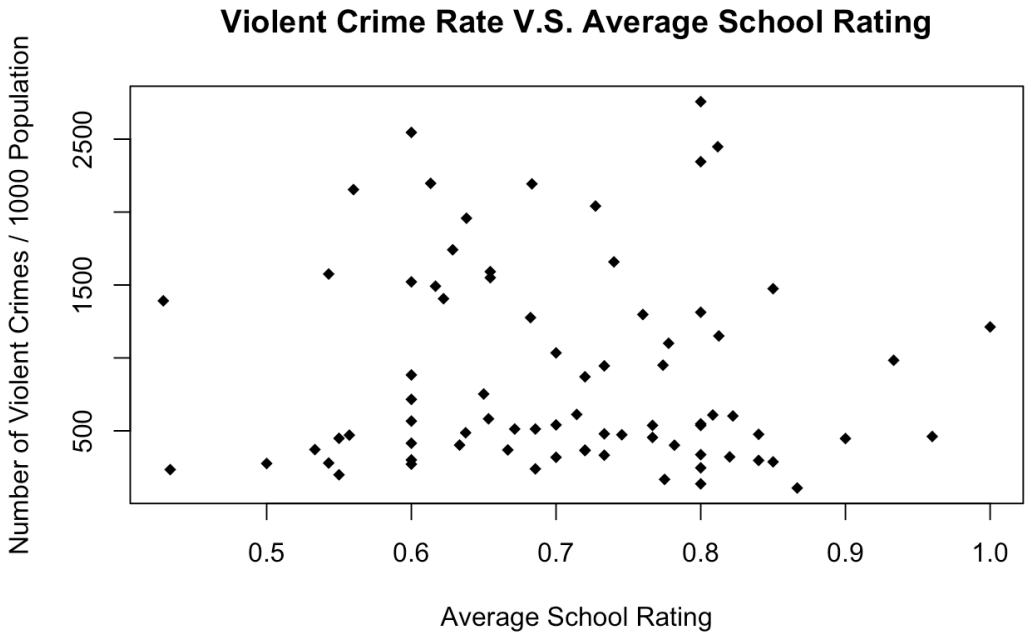

Figure 25. Scatter Plot for Normalized Average School Rating

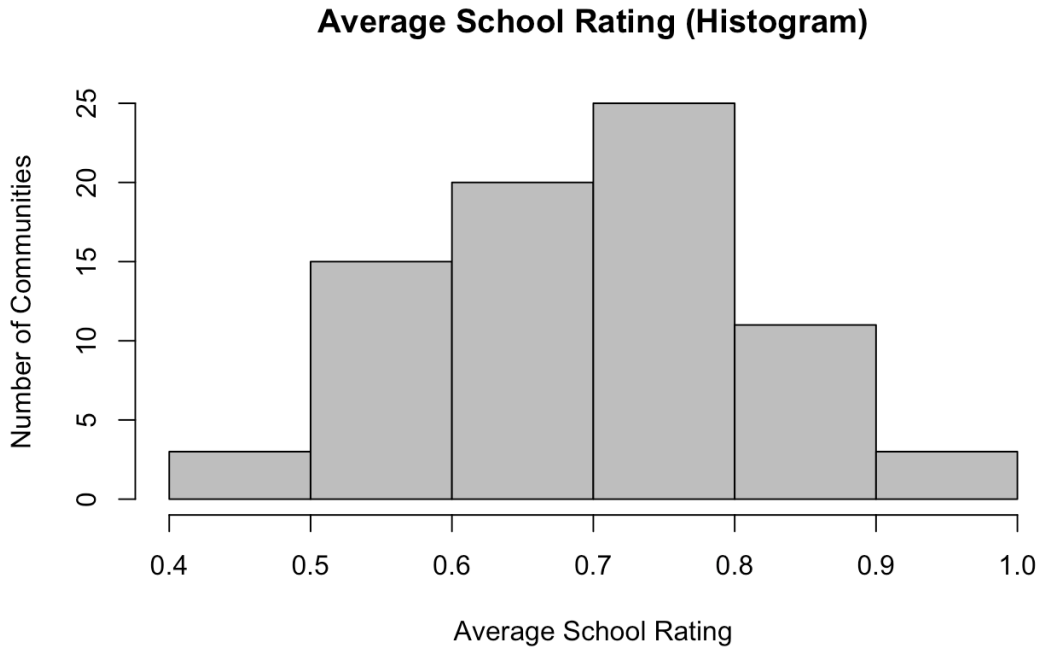

Figure 26. Histogram for Normalized Average School Rating

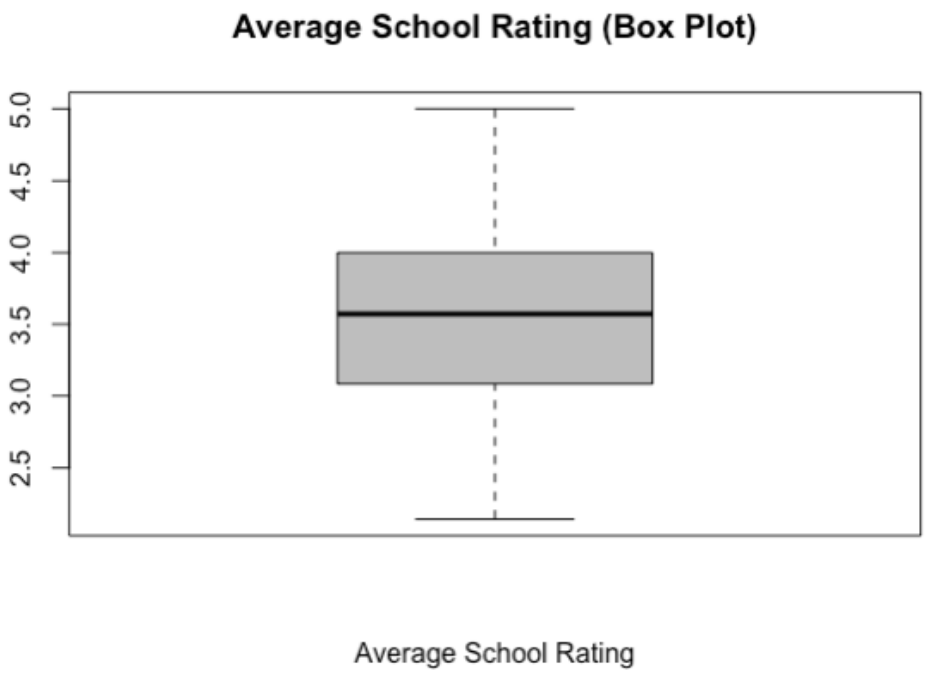

Figure 27. Box Plot for Average School Rating

The histogram and the box plot show that the average school rating is normally distributed. No clear correlation is indicated in the scatter plot, therefore, the average school rating should have no or a very small weight in the prediction model.

11.2.3 Average SSL Rating

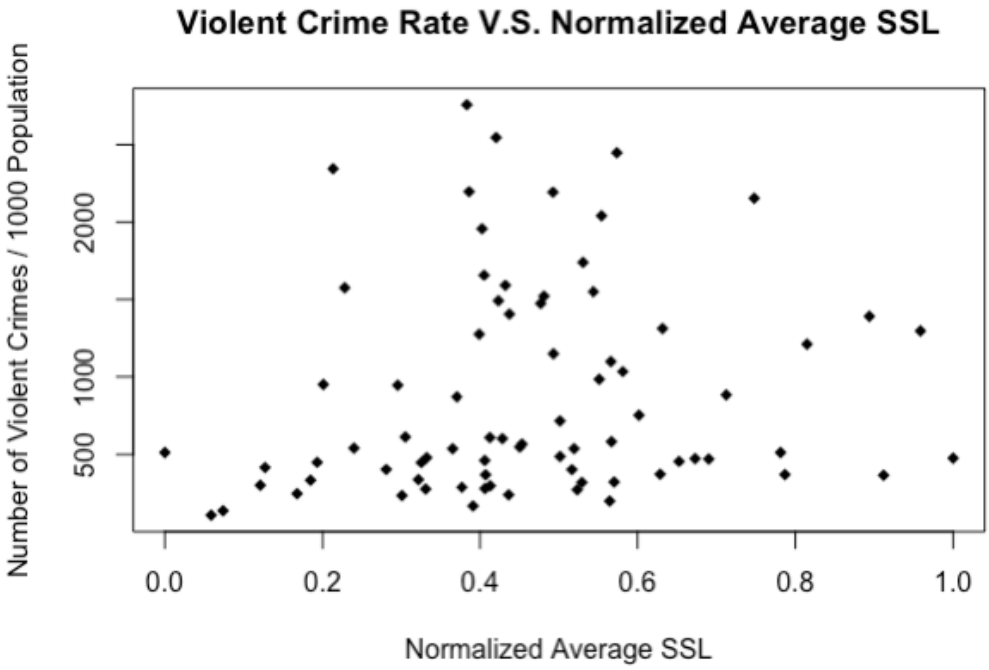

Figure 28. Scatter Plot for Normalized Average SSL Rating

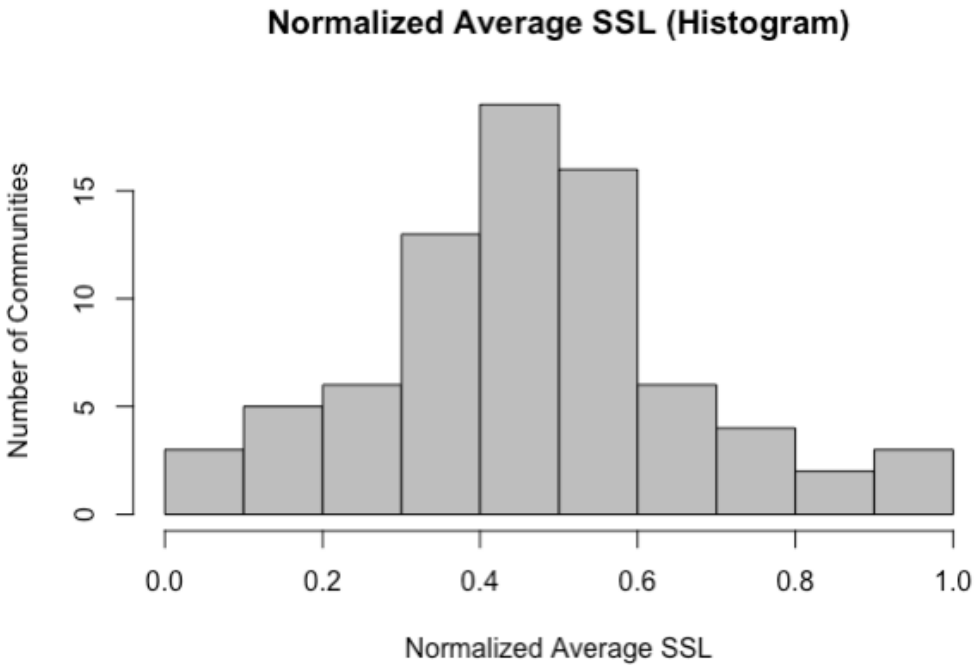

Figure 29. Histogram for Normalized Average SSL Rating

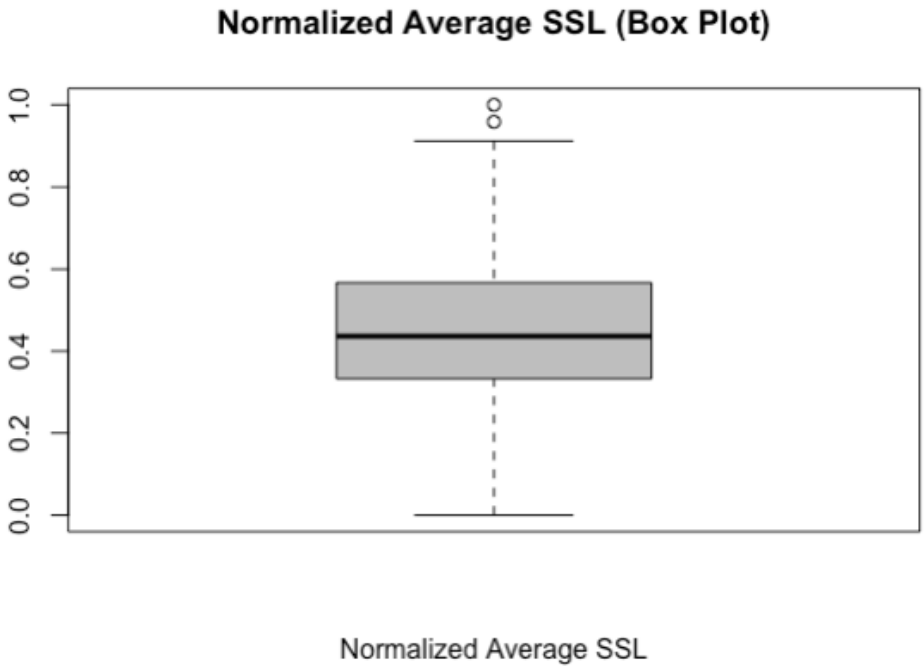

Figure 30. Box Plot for Normalized Average SSL Rating

Note that the average SSL rating is normalized and scaled down from the original 266 to 304 range to 0 to 1 range. It is normally distributed, and there is no evident correlation between the SSL rating and the number of violent crimes.

11.2.4 Total Park Area

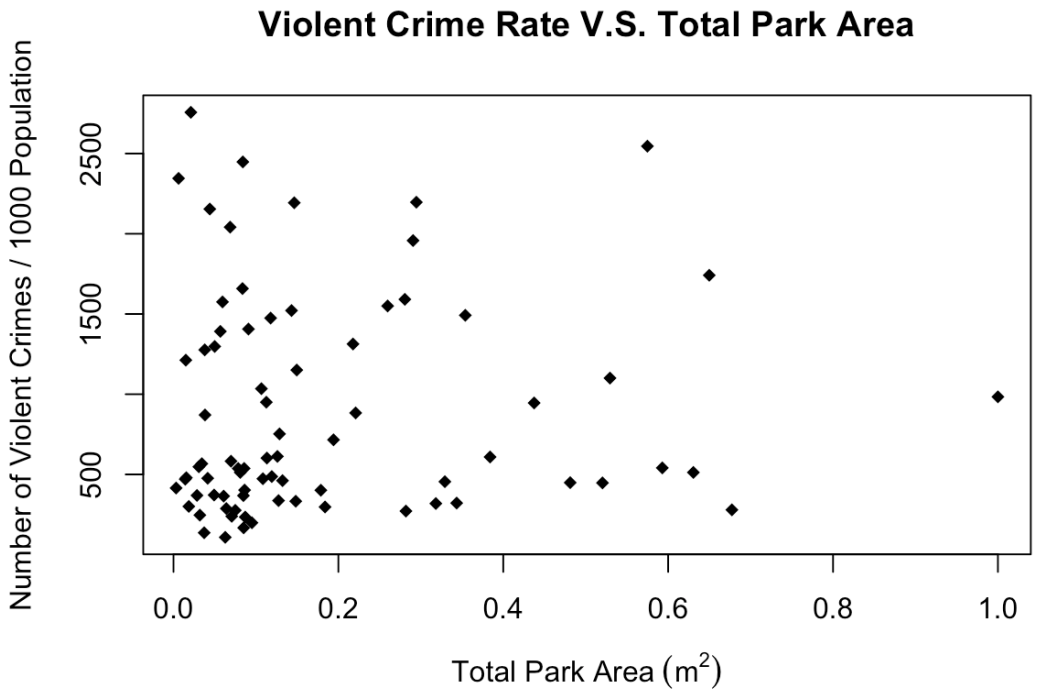

Figure 31. Scatter Plot for Normalized Total Park Area

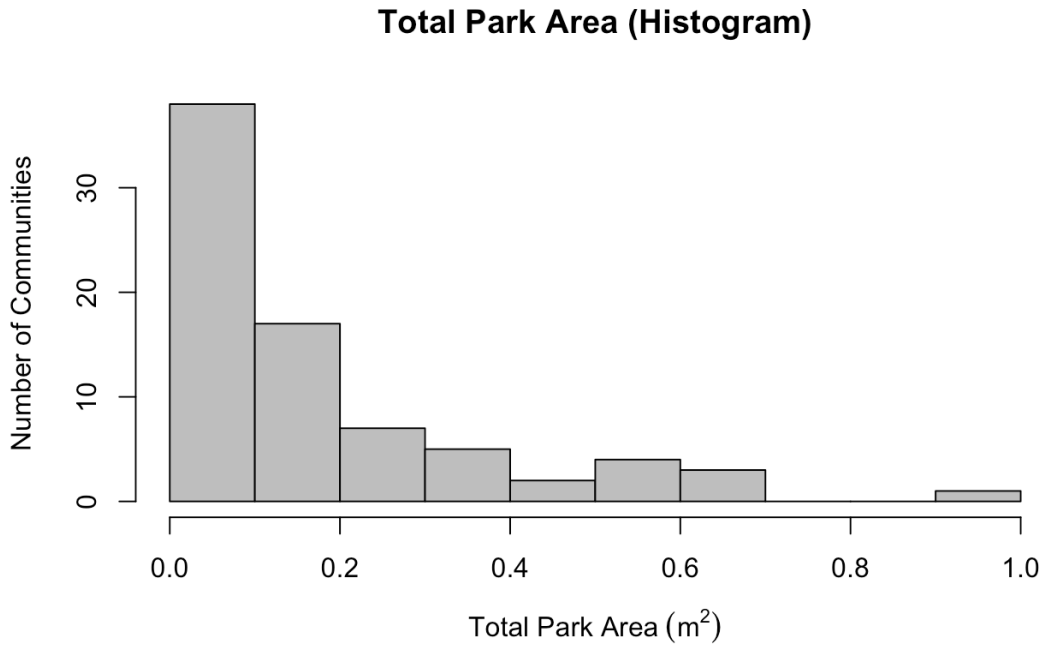

Figure 32. Normalized Histogram for Total Park Area

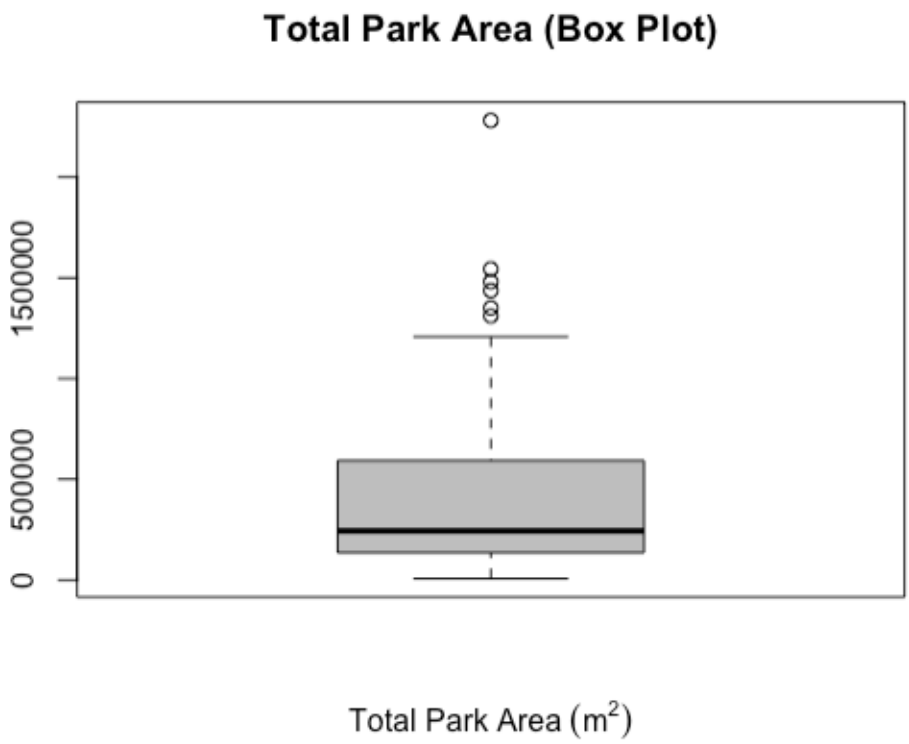

Figure 33. Box Plot for Total Park Area

Over 50 communities have less than 500000 m^2 of park area. A negative exponential relationship can be observed from the scatter plot. However, since the number of samples with large park areas is low, it requires more analysis when building the prediction model to see if there is a definite correlation.

11.2.5 Number of Hospitals

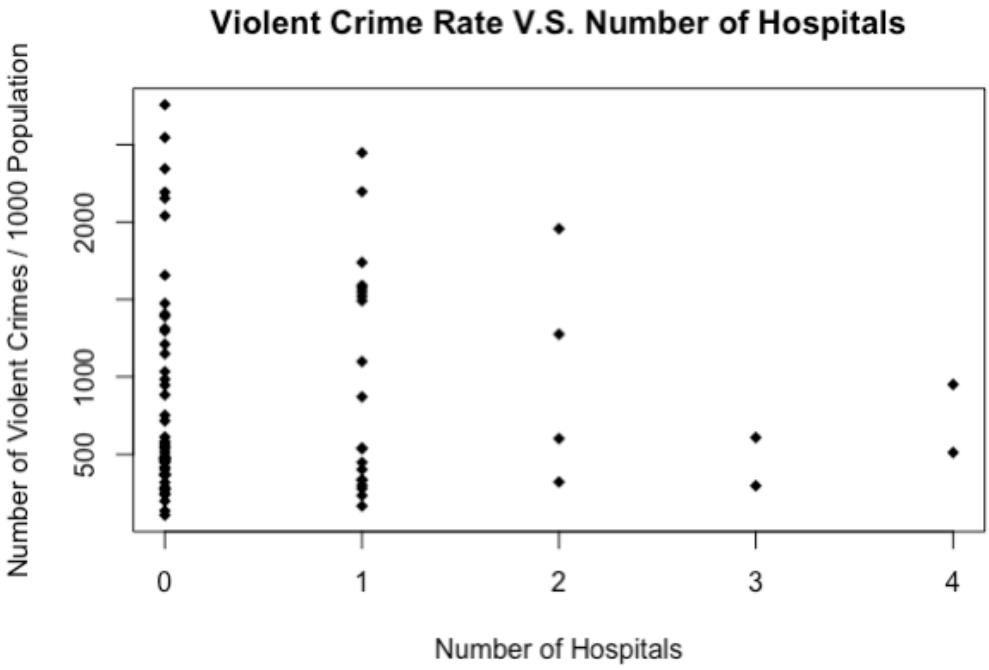

Figure 34. Scatter Plot for Number of Hospitals

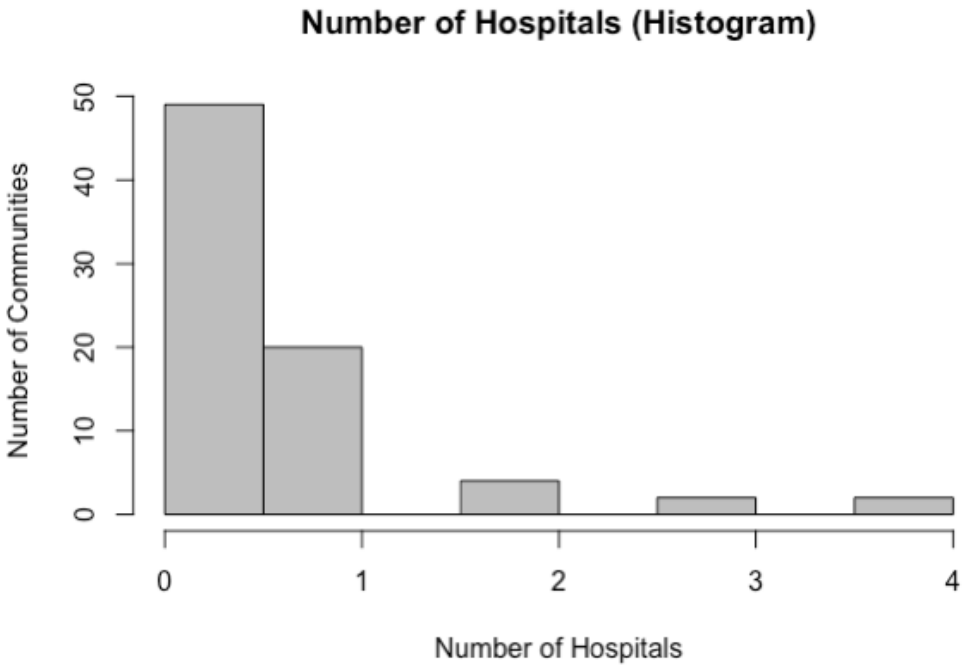

Figure 35. Histogram for Number of Hospitals

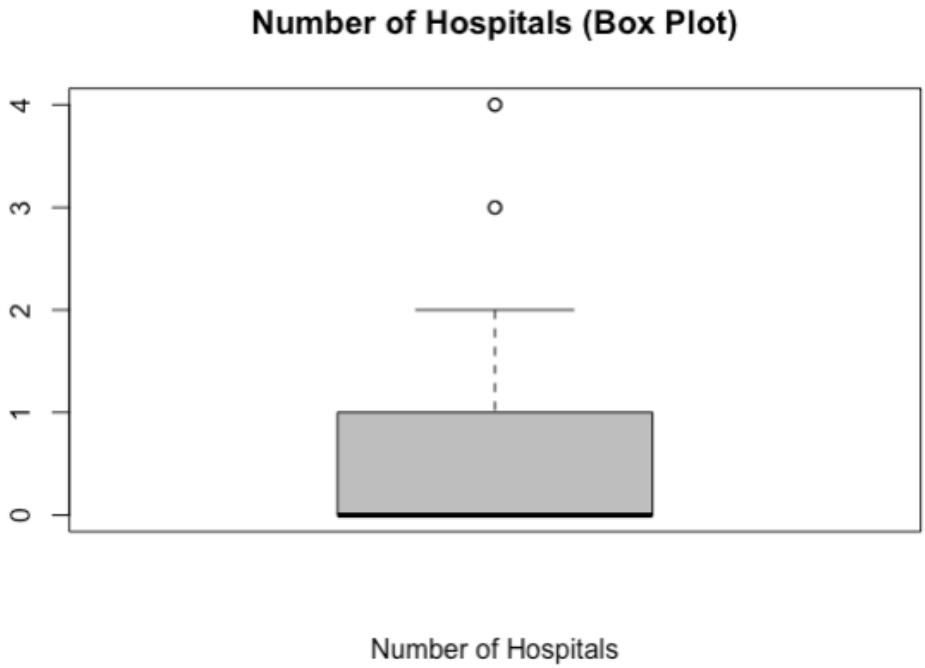

Figure 36. Box Plot for Number of Hospitals

As shown in all the above plots, a majority of the communities do not have hospitals. There are 4 outliers: Uptown and West Town with 3 hospitals, and Lincoln Square and New West Side with 4 hospitals.

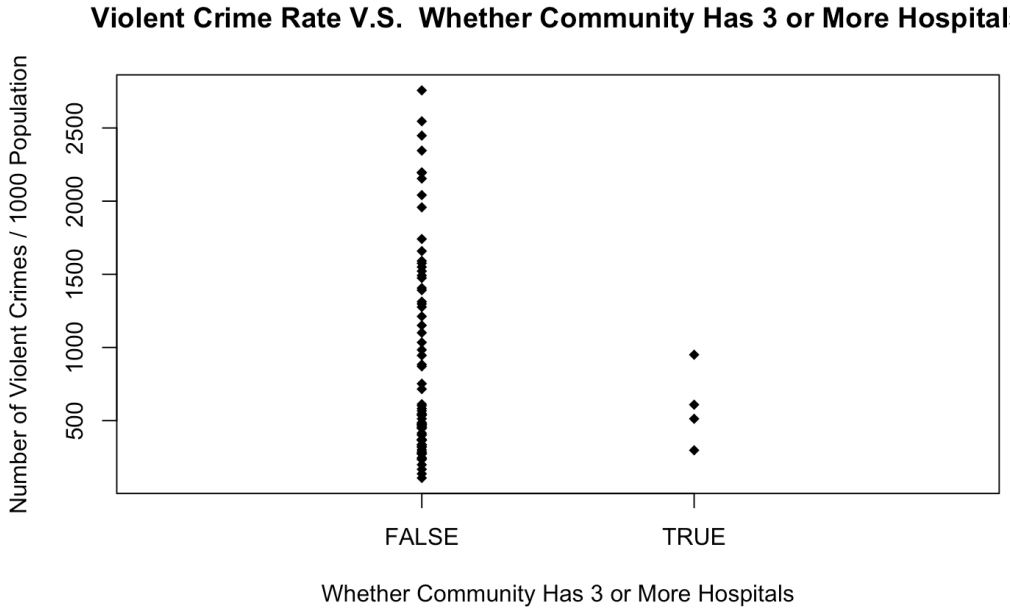

Figure 37. Scatter Plot for Whether Community Has 3 or More Hospitals

Another plot is made with a Boolean of whether the community has 3 or more hospitals as the x axis. All four communities which have 3 or 4 hospitals have lower than 1000 violent crimes per 1000 population. This transformed Boolean variable, which will be referred to as 'has3OrMoreHospitals', may produce better prediction results and is therefore used instead of the more general 'numOfHospitals' variable when conducting regression.

11.2.6 Birth Rate by Teenage Mothers

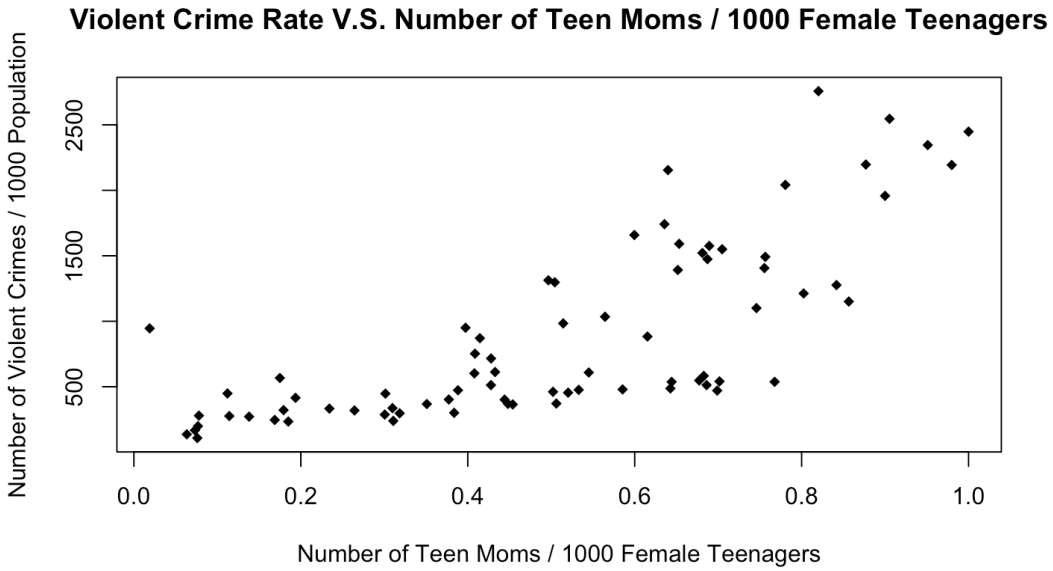

Figure 38. Scatter Plot for Normalized Teenage Pregnancy Rate

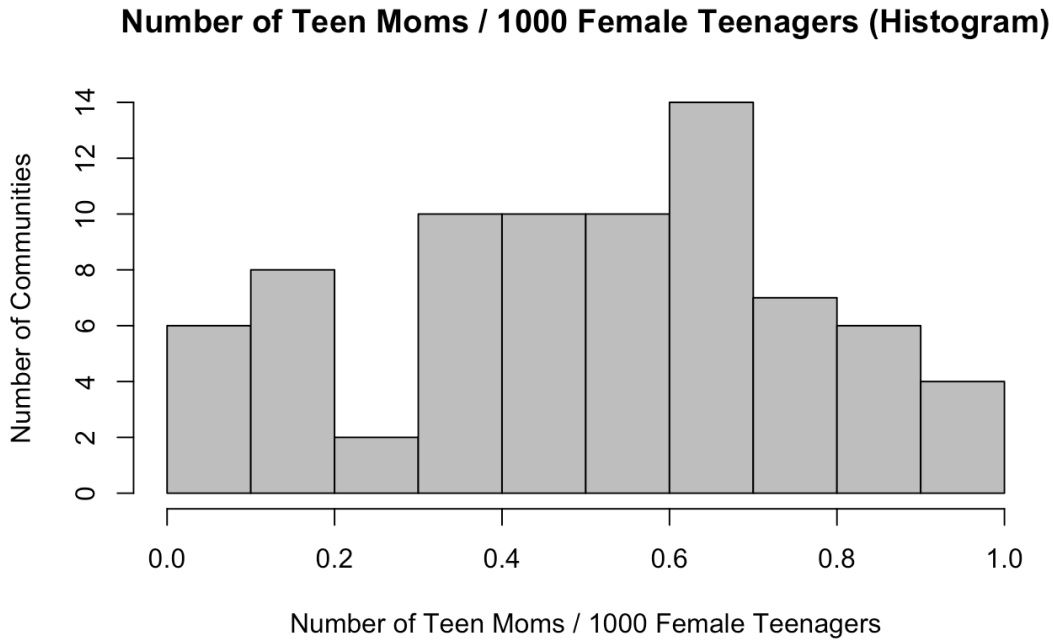

Figure 39. Normalized Histogram Plot for Birth Rate by Teenage Mothers

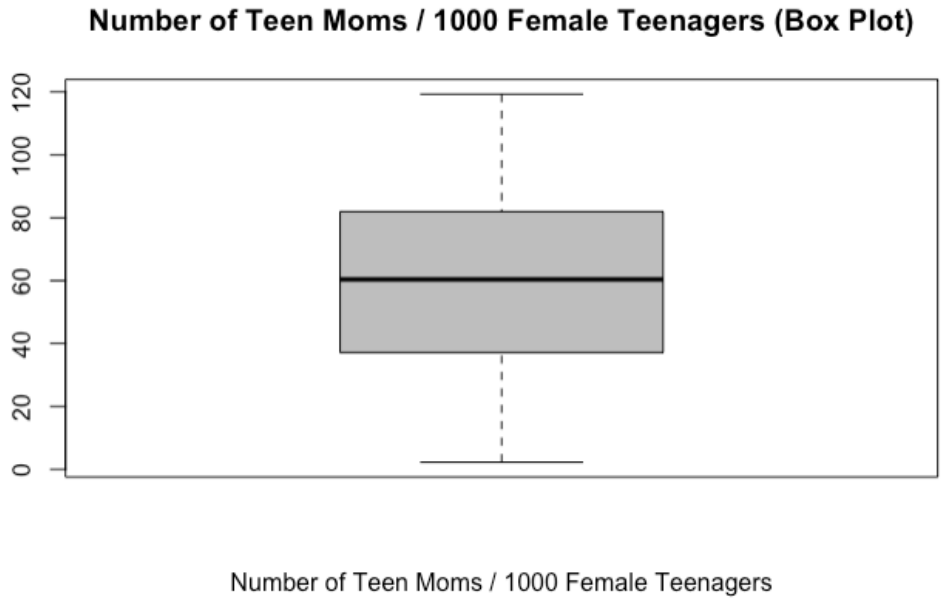

Figure 40. Box Plot for Birth Rate by Teenage Mothers

The number of teenage moms per 1000 female teenagers is also normally distributed in a 0 to 120 range. As indicated in the scatter plot, there is a clear positive relationship between the number of teenage moms and the number of violent crimes.

11.2.7 Infant Mortality Rate

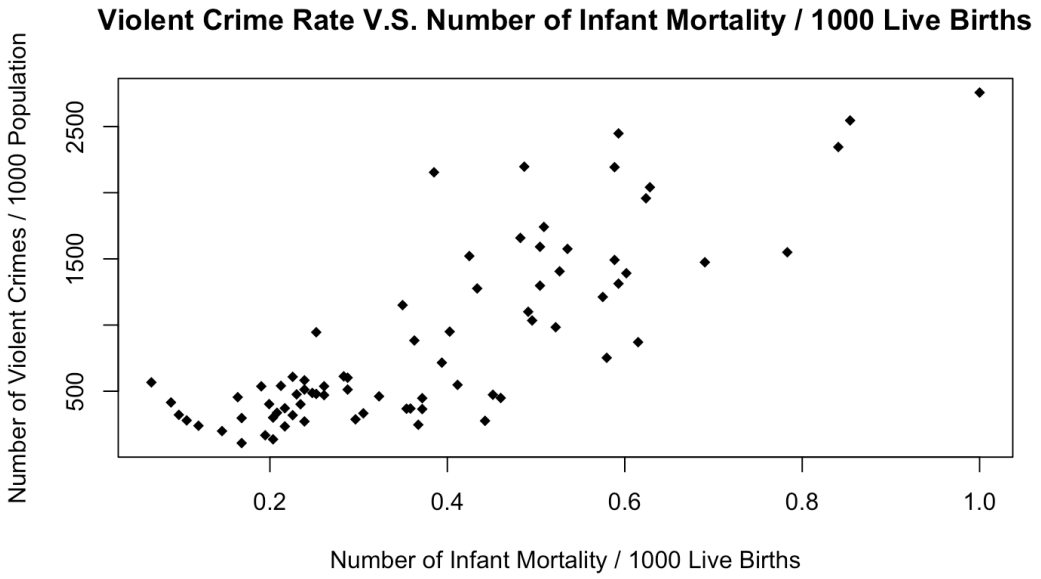

Figure 41. Scatter Plot for Normalized Infant Mortality Rate

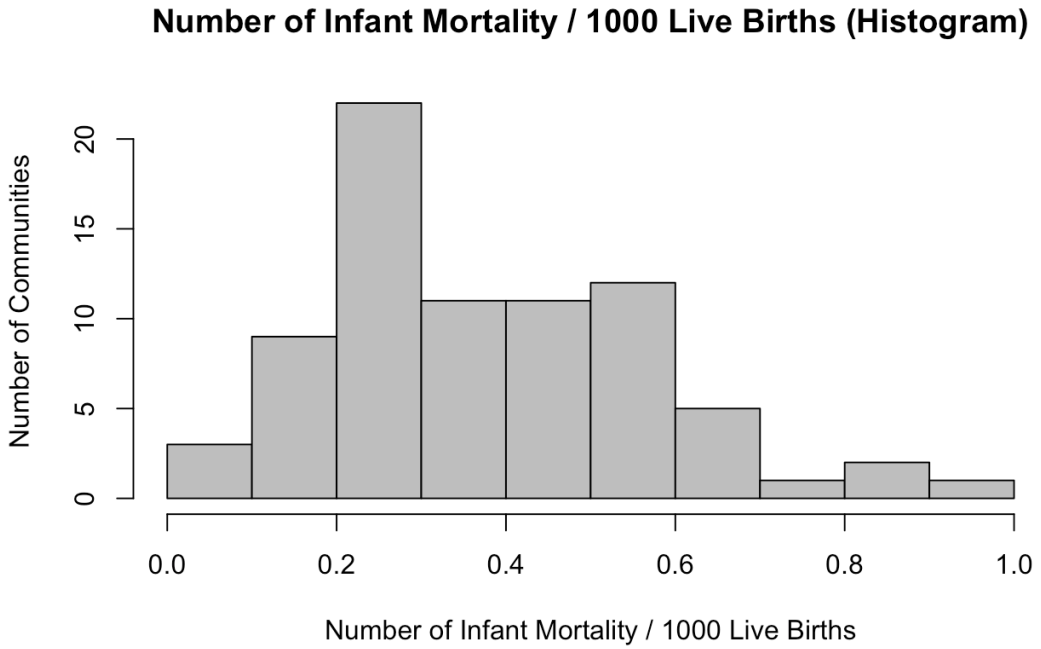

Figure 42. Normalized Histogram for Infant Mortality Rate

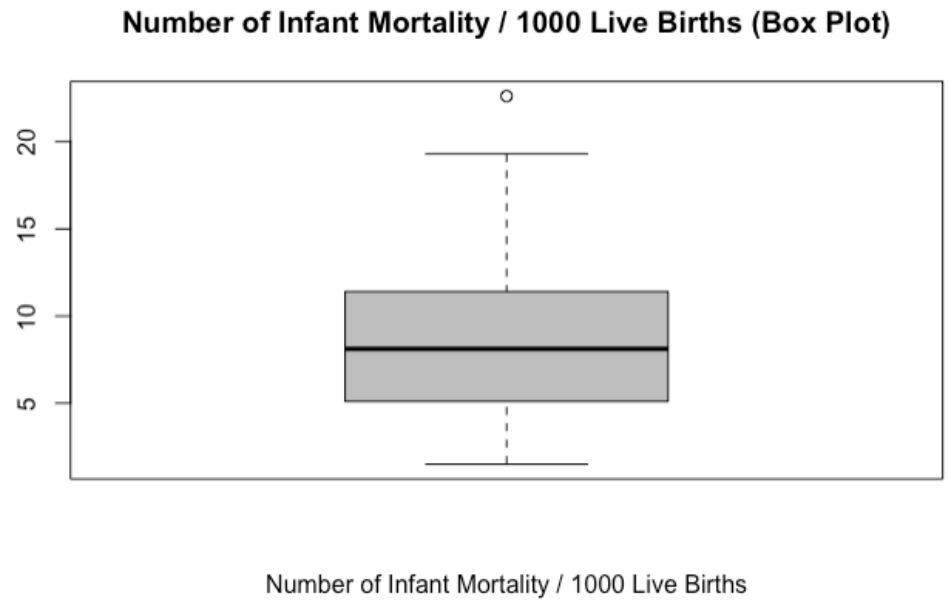

Figure 43. Box Plot for Infant Mortality Rate

From the scatter plot, it can be observed that there is a clear positive relationship between the number of infant mortalities per 1000 live births and the number of violent crimes per 1000 people. The only outlier indicated in the box plot is Fuller Park, which has 22.6 infant mortalities per 1000 live births.

11.2.8 Proportion of Hispanic People

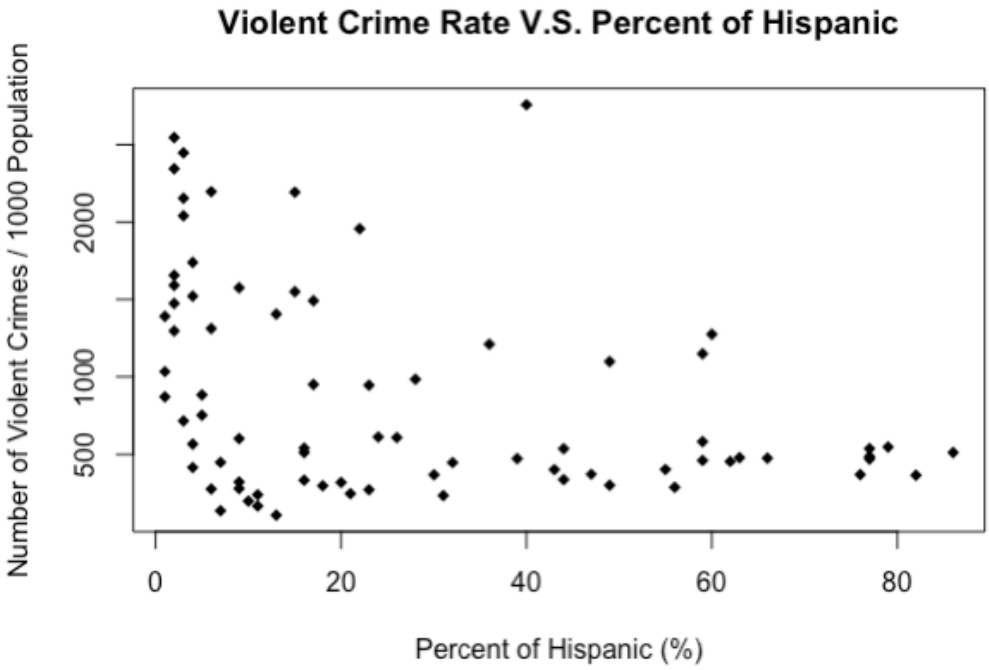

Figure 44. Scatter Plot for Proportion of Hispanic People

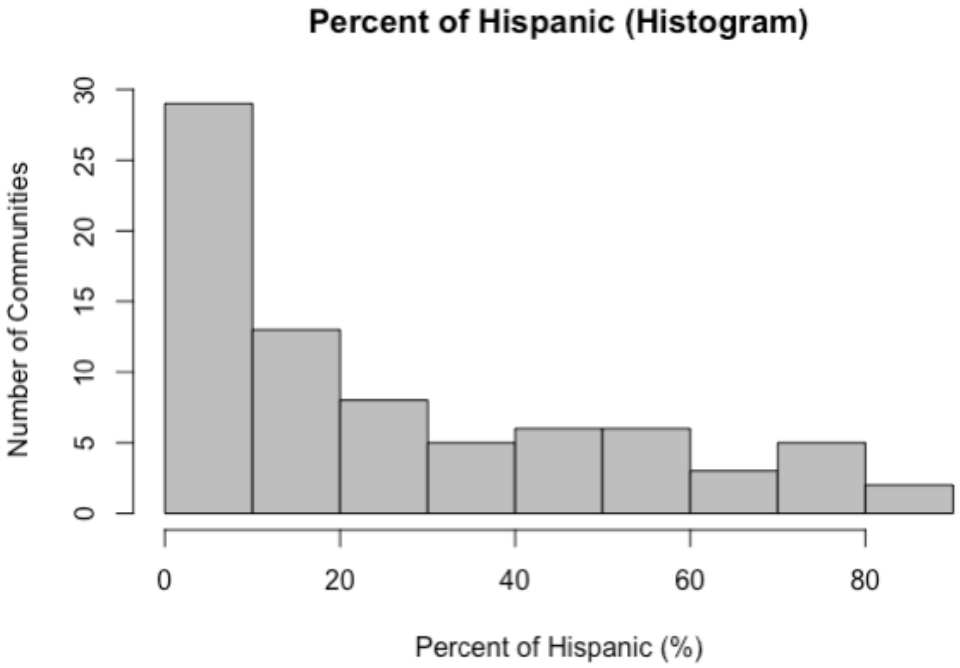

Figure 45. Histogram for Proportion of Hispanic People

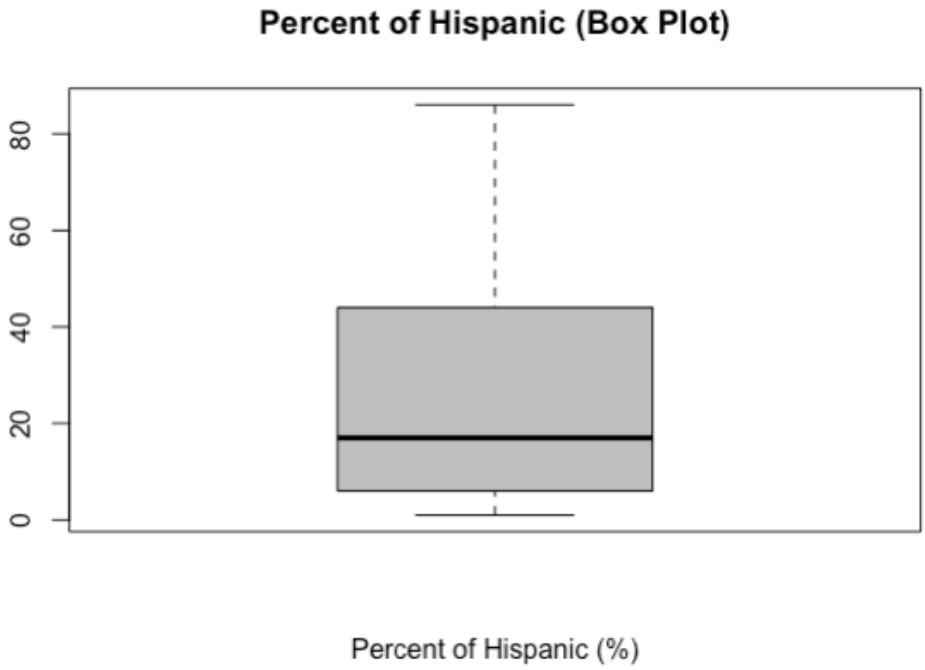

Figure 46. Box Plot for Proportion of Hispanic People

The violent crime rate exponentially decreases as the percentage of Hispanic people increases from 0 to 90%, as shown in the scatter plot. The histogram also indicates that the distribution is not normal, but rather exponential as well.

11.2.9 Proportion of Black People

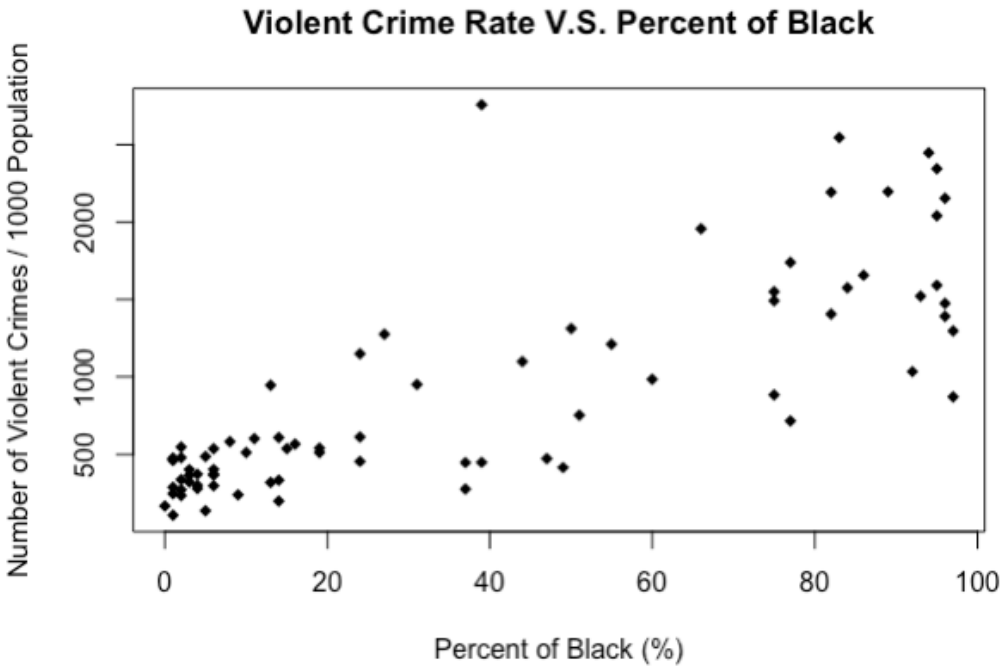

Figure 47. Scatter Plot for Proportion of Black People

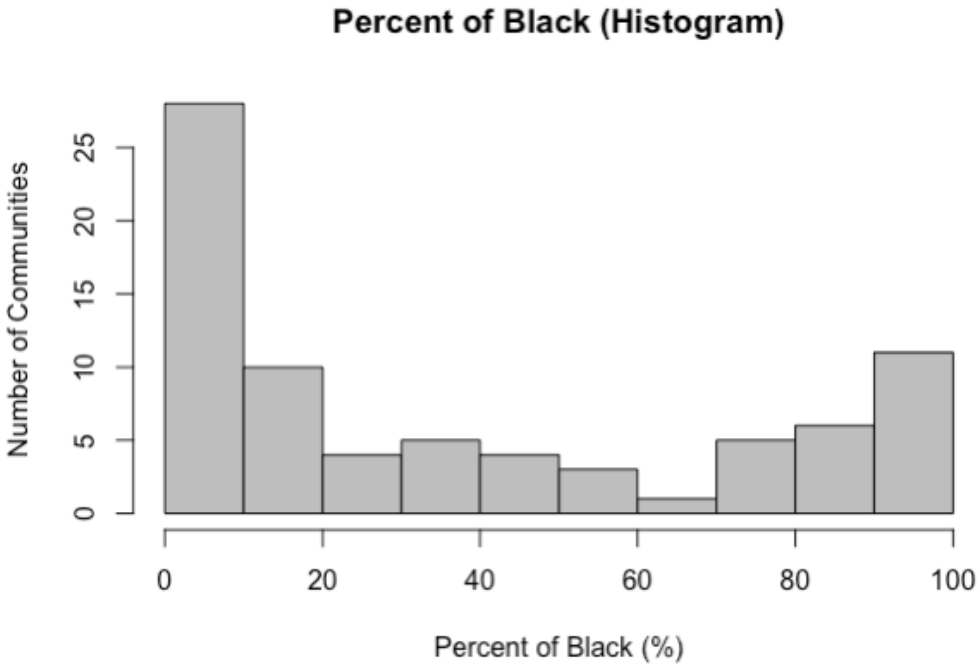

Figure 48. Histogram for Proportion of Black People

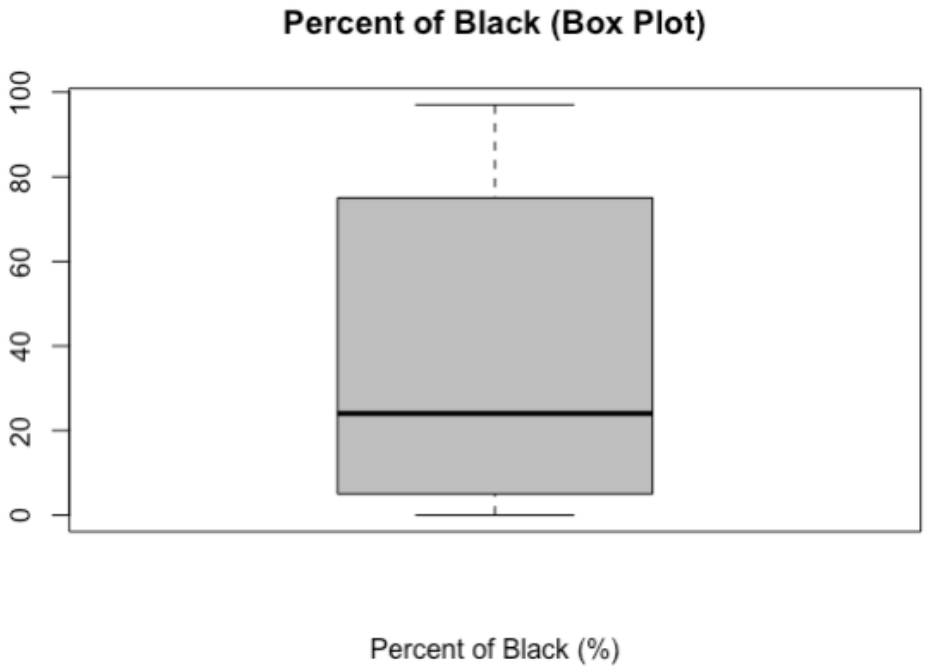

Figure 49. Box Plot for Proportion of Black People

The number of communities decreases and then increases as the percentage of black people increases from 0 to 100%. About 28 communities have 0 to 10% black people; about 10 communities have 10% - 20% and about 10 communities have 90% - 100%. A very low number of communities have 20% - 90% of black people. In addition, there is a clear positive relationship that can be observed with the class variable from the scatter plot.

11.2.10 Proportion of White People

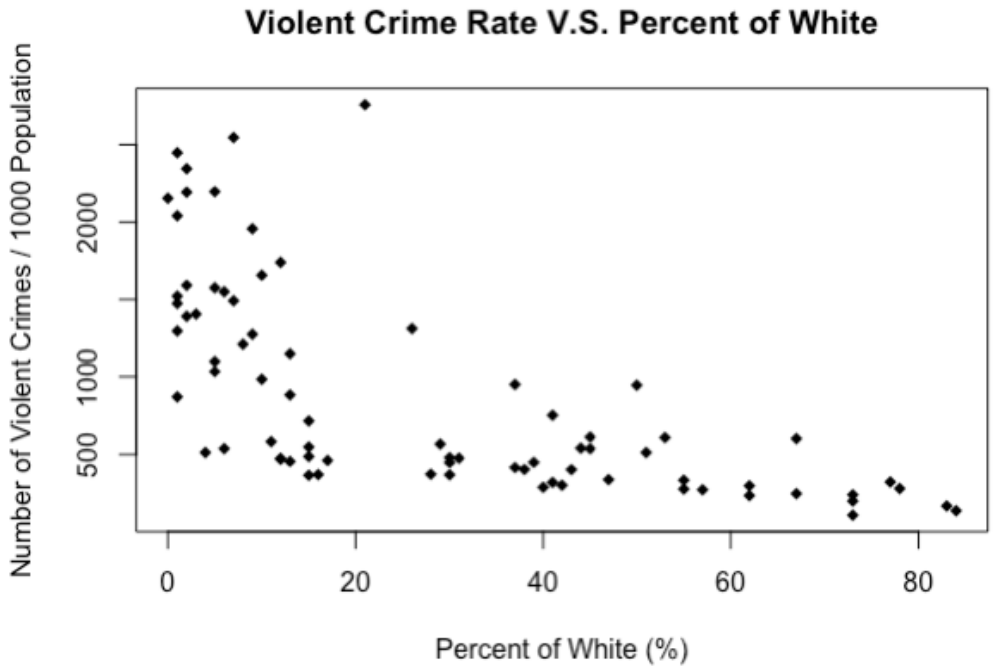

Figure 50. Scatter Plot for Proportion of White People

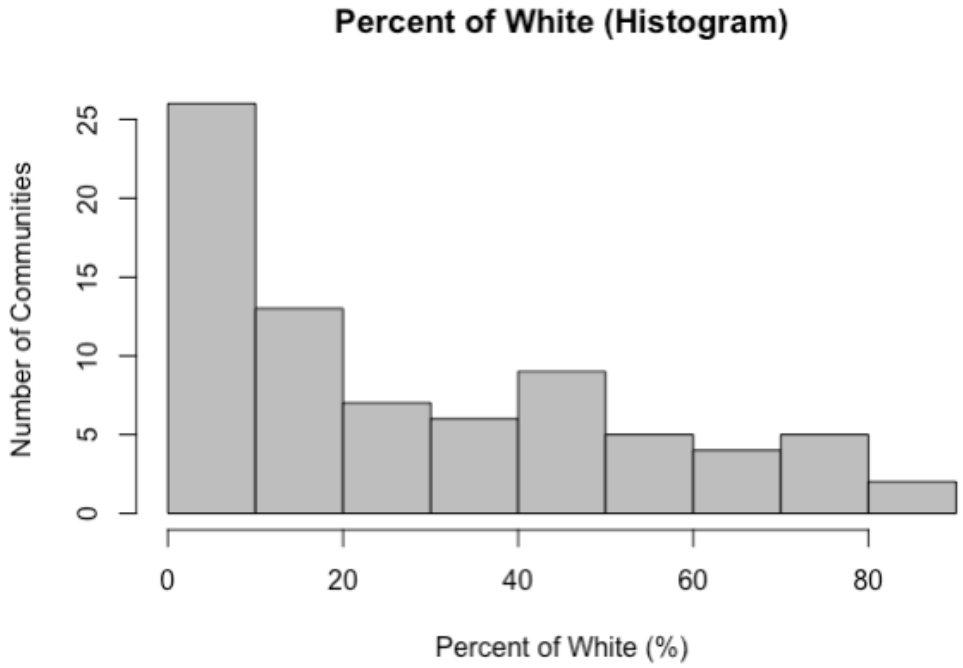

Figure 51. Histogram for Proportion of White People

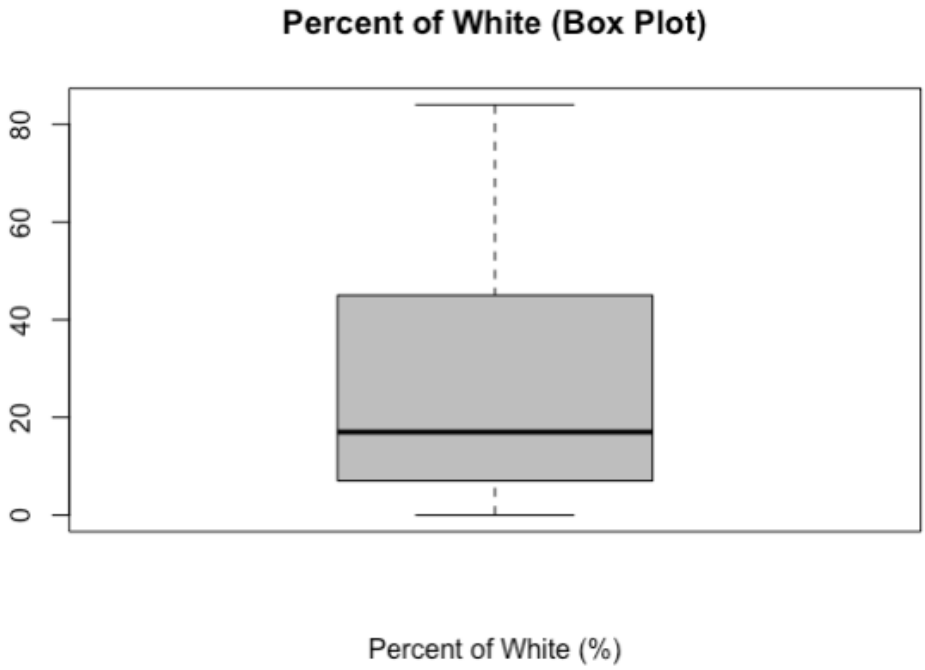

Figure 52. Box Plot for Proportion of White People

The distribution of the histogram is an exponential, in that the number of communities exponentially decreases as the percentage of white people increases. Furthermore, as shown in the scatter plot, there is a negative exponential relationship with the class variable.

11.2.11 Proportion of Asians

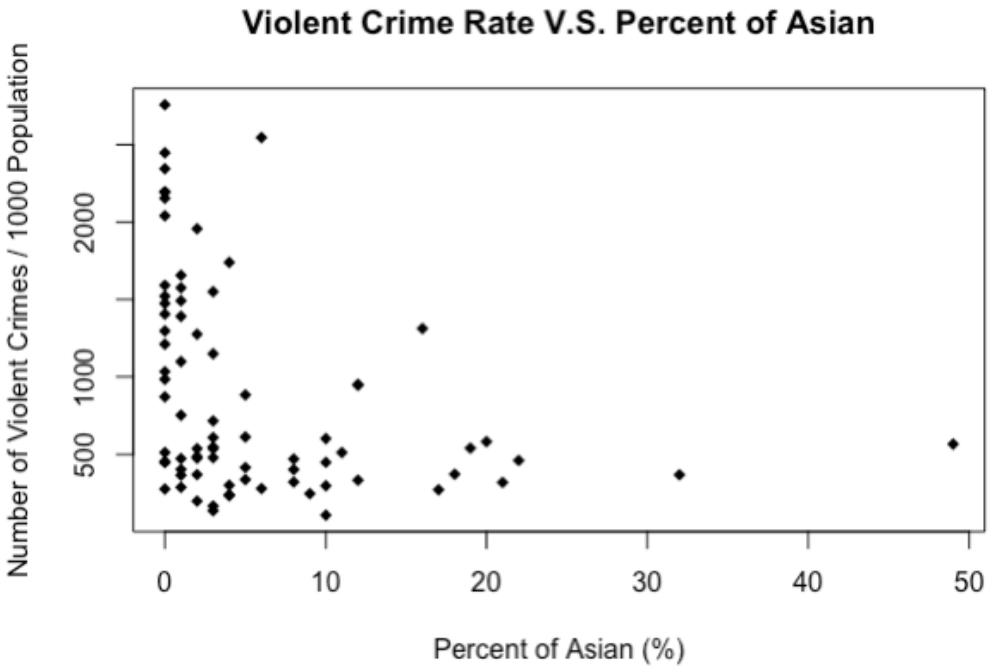

Figure 53. Scatter Plot for Proportion of Asian People

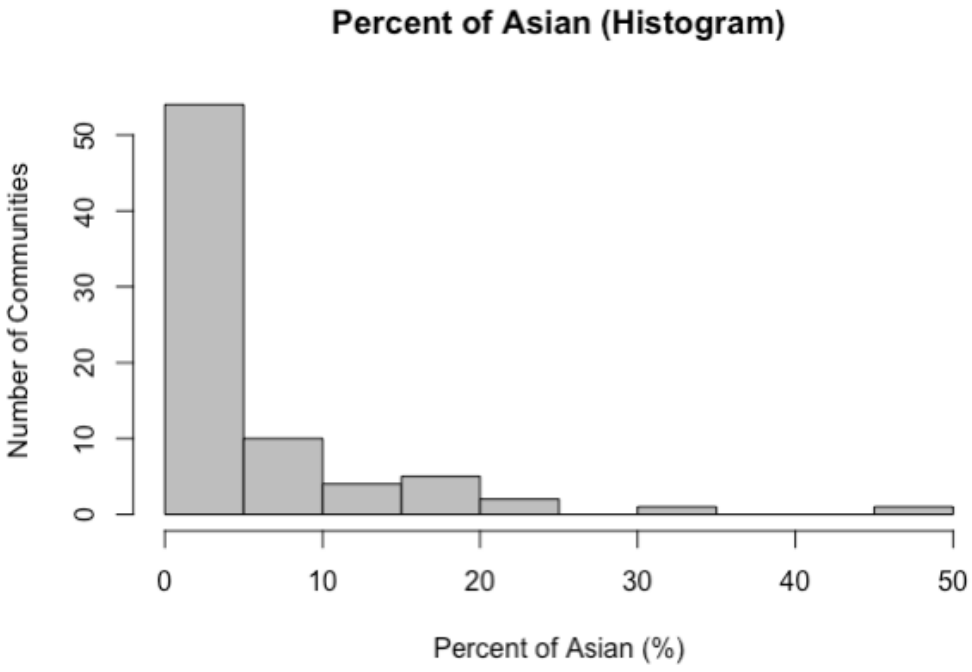

Figure 54. Histogram for Proportion of Asian People

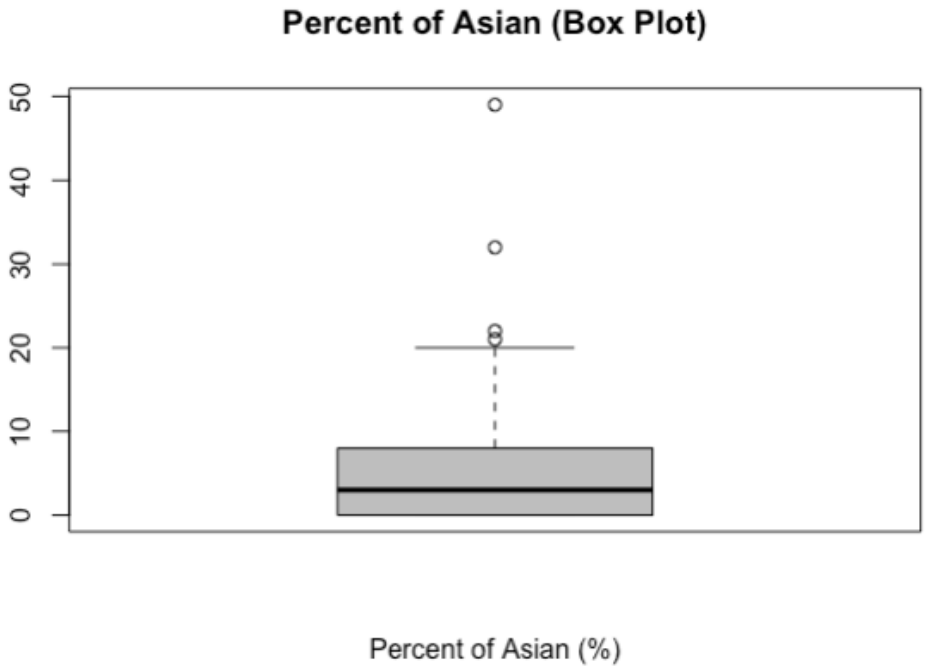

Figure 55. Box Plot for Proportion of Asian People

Similar to percentage of white people, the number of communities exponentially decreases as the percentage of Asian people increases from 0 to 50%. As shown in the scatter plot, there is a negative exponential relationship with the class variable. However, it is not clear due to the low number of samples with a high Asian percentage rate.

11.2.12 Proportion of Other Races

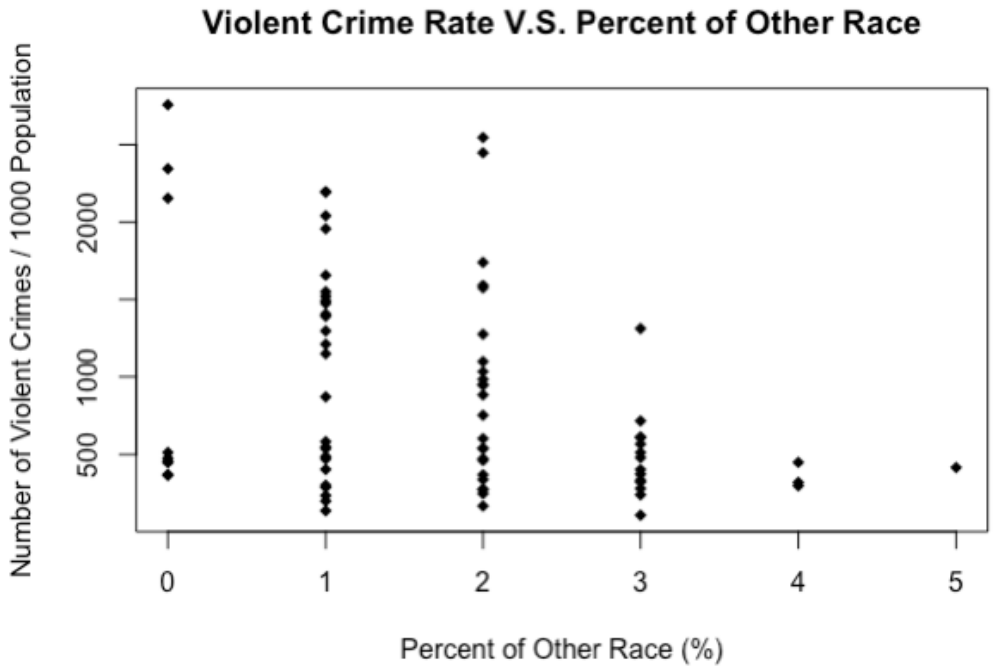

Figure 56. Scatter Plot for Proportion of Other Races

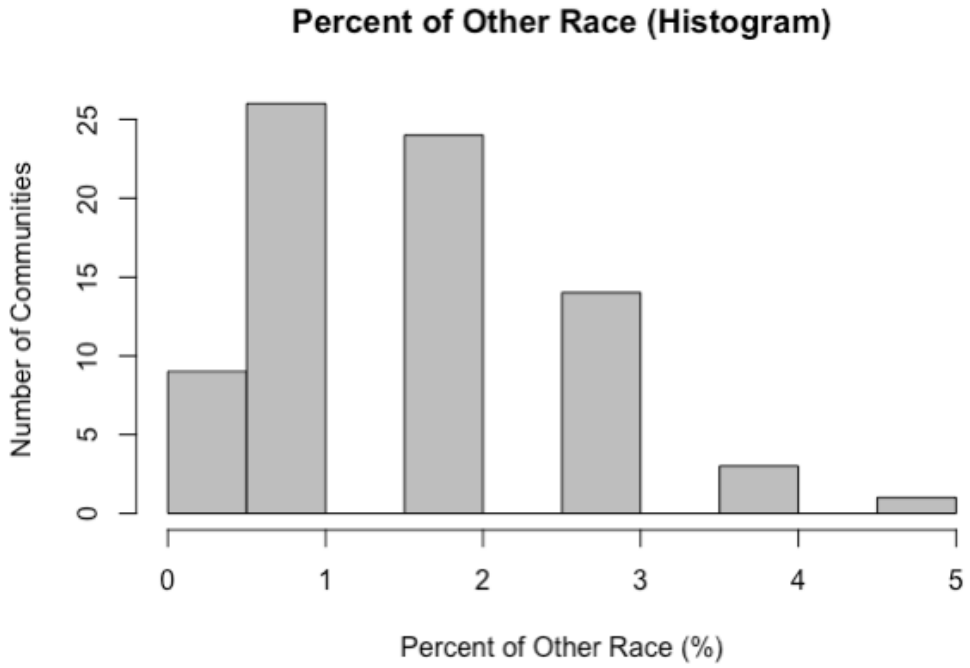

Figure 57. Histogram for Proportion of Other Races

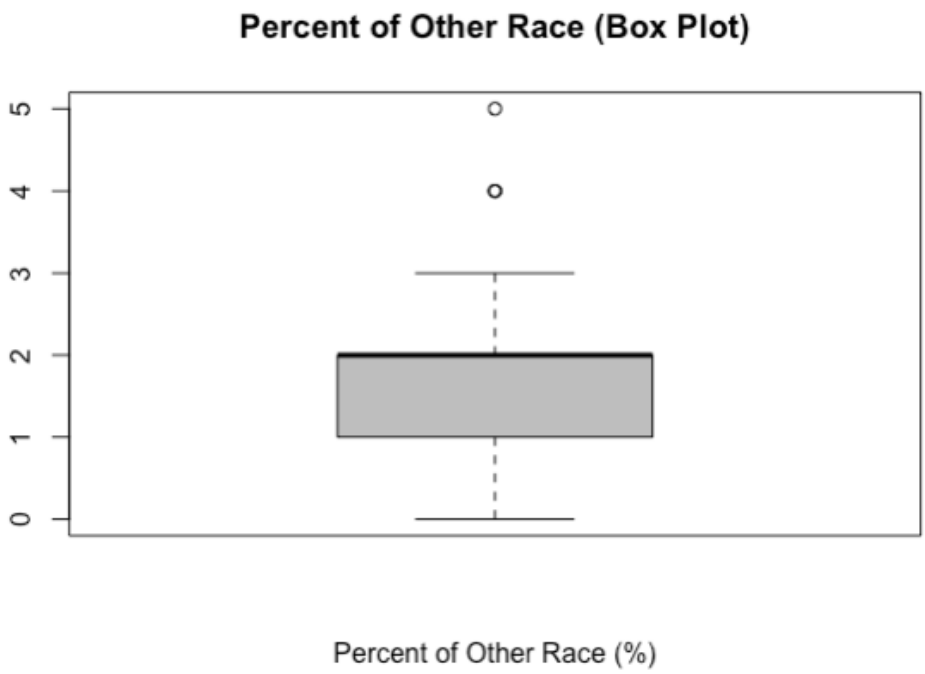

Figure 58. Box Plot for Proportion of Other Races

The percentage of other races does not seem indicate any strong correlation with the class variable. There are also very few communities with a percentage of people of "other races" of 4% or more.

11.2.13 Percent of Children in Poverty

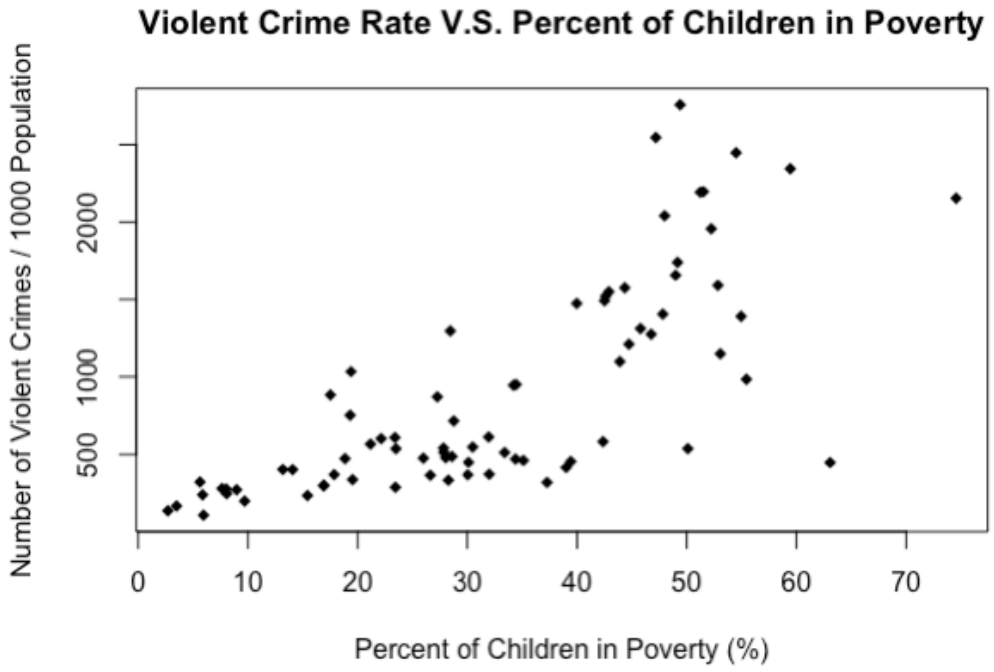

Figure 59. Scatter Plot for Percent of Children in Poverty

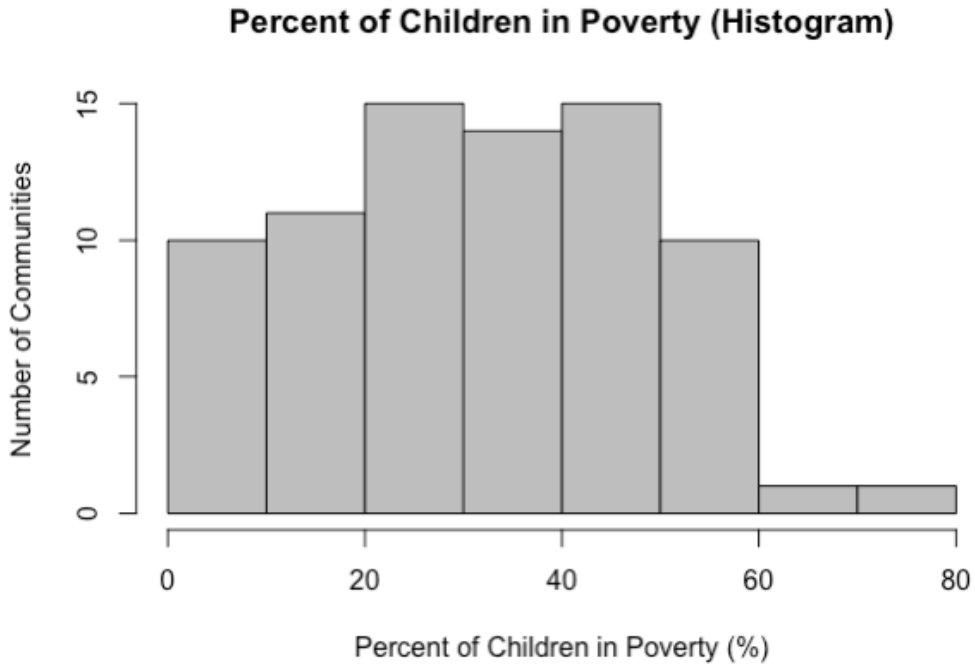

Figure 60. Histogram for Percent of Children in Poverty

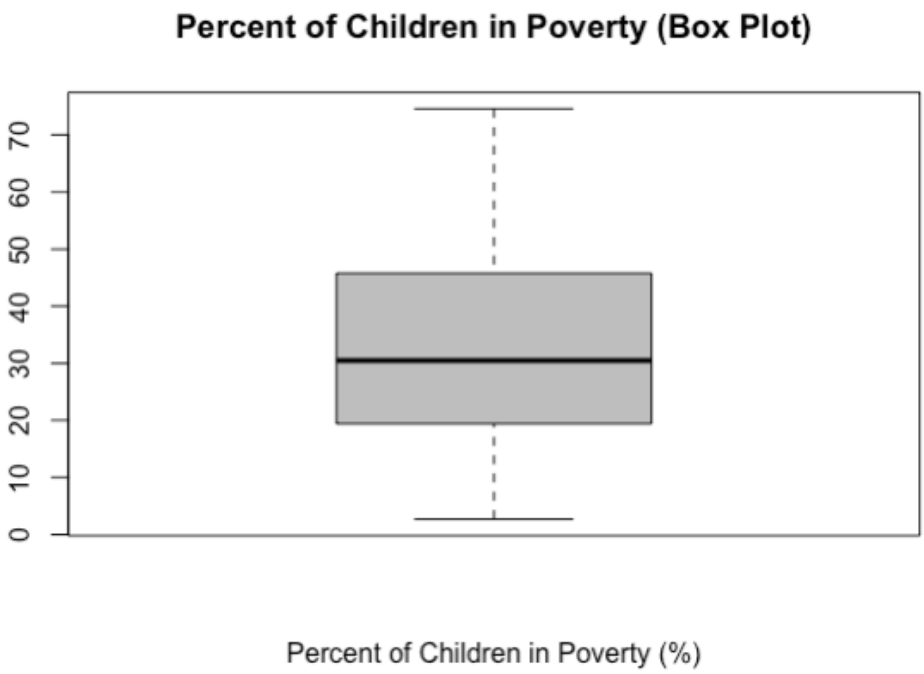

Figure 61. Box Plot for Percent of Children in Poverty

From the scatter plot, it can be observed that there is a clear positive relationship between the percentage of children in poverty and the rate of violent crimes. The percentage of children in poverty ranges from 0 to 80%, concentrated between the range of 0 to 60%.