University of Waterloo

Faculty of Engineering  
Department of Management Sciences

Status Update on 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

13 November 2018

1. Summary

* R programming language is used
* Scope of project reduced so instead of random unspecified blocks of places, we just do by community areas
* No dimension of time considered anymore (makes analysis harder)
* 12 predictors: average school rating of community area, average SSL rating of community area, total park area each community area has, number of hospitals that area has, teenage mother birth rate, infant mortality rate, proportion of Hispanic, black, white, Asian, other races, and percent of children in poverty
* Preprocessing took a lot of time to do (a lot of data was geographic data and we had to use polygon algorithms and do intersections of polygons to get e.g. total park area for each community area)
* EDA: some x’s were definitely related to the class variable (e.g. race, teen mom rate, etc.)
* Prediction:
* Clustering:
* Association:

1. Data Collection Progress

Table 1 displays all of the final chosen datasets, which increased from the original project proposal due to the inclusion of additional predictors such as public health data, and because some predictors required additional datasets to be able to transform them into usable data for prediction and clustering analysis.

Table 1. Information on Final Chosen Datasets.

|  |  |  |
| --- | --- | --- |
| Dataset | Number of Rows | What Does a Row Represent? |
| Crimes from 2001 [1] | 6,706,459 | Reported Crime |
| Strategic Subject List [2] | 398,684 | Person Likely to be Involved in a Shooting |
| Chicago Public Schools - School Profile Information SY1718 [3] | 661 | School |
| Population and Poverty Data by Chicago Community Area [4] | 77 | Community Area |
| Parks - Chicago Park District Park Boundaries (current) [5] | N/A (Shapely File of 597 Parks) | N/A (Shapely File of 597 Parks) |
| Boundaries - Community Areas (current) [6] | N/A (Shapely File of 77 Community Areas) | N/A (Shapely File of 77 Community Areas) |
| Hospitals – Chicago [7] | 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 [8] | 77 | Community Area |
| Public Health Statistics- Infant mortality in Chicago, 2005– 2009 [9] | 77 | Community Area |

The remainder of this section will describe how all the predictors and the class variable are obtained from the above datasets. Note that the code written in R to conduct these data transformations are available in the Appendix section.

Class Variable

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

*violentCrimeForCommunityArea \* 1000 / populationOfCommunityArea*

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 include: assault, battery, and homicide. Finally, the population of each community area is obtained from census data.

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” [3]. 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.

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 crime in the near future. 0 is extremely low risk and 500 is extremely high risk. 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 is from dataset, “Strategic Subject List” [2], which takes samples from the list of arrest data from August 1, 2012 to July 31, 2016.

Total Park Area

To obtain the total park area for each community area, all park shape files [5] and all community area shape files [6] 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.

Number of Hospitals

To obtain the number of hospitals for each community area from the hospital data [7] was simple, since each hospital data point included the community area it is located in.

Birth Rate by Teenage Mothers

The dataset of birth rate by teenage mothers has all the rates from 1999 to 2009 [8]. The average of all these birth rates for each community areas is used.

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 where data was available: from 2005 to 2009 [9]. Two values for one community area from two years had null values, so the average of the non-null values was calculated.

Proportion of Hispanic People

Nothing was needed to be transformed for this predictor [4].

Proportion of Black People

Nothing was needed to be transformed for this predictor [4].

Proportion of White People

Nothing was needed to be transformed for this predictor [4].

Proportion of Asians

Nothing was needed to be transformed for this predictor [4].

Proportion of Other Races

Nothing was needed to be transformed for this predictor [4].

Percent of Children in Poverty

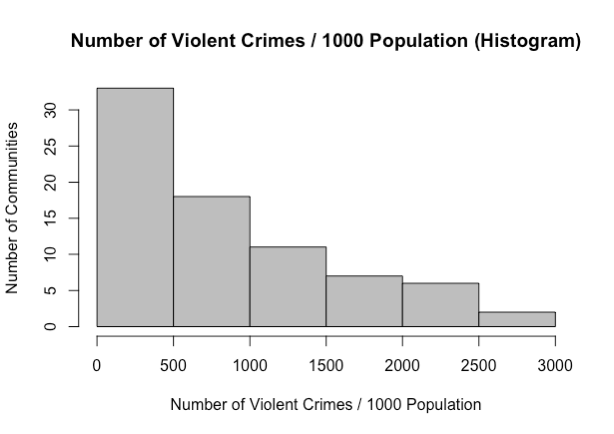
Note that actual poverty rate was unable to be obtained, but percentage of children in poverty was easily obtainable and thus is used instead.

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 [4]. 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.

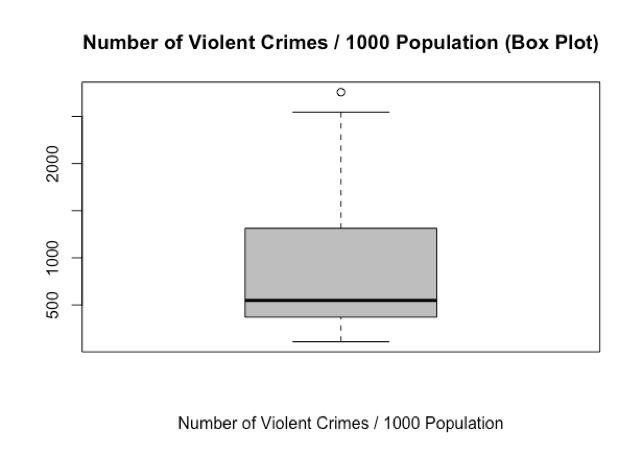
1. Exploratory Data Analysis Progress

Scatter plot, histogram, and box plot are used to analyse each variable.

Class Variable



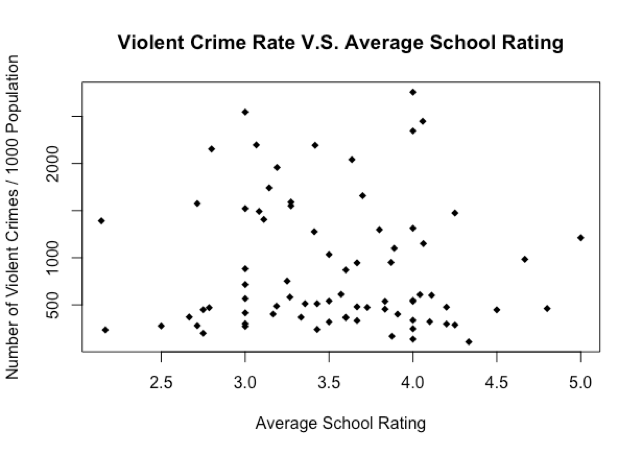
*Figure 1. Histogram of Class Variable*



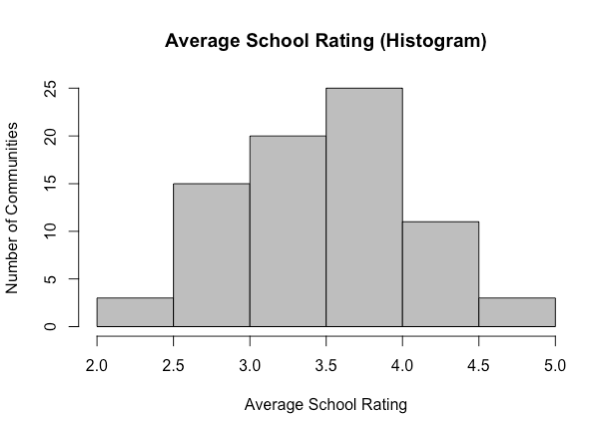
*Figure 2. Box Plot of Class Variable*

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 index and why the community area is described as one of the most dangerous places to live in Chicago.

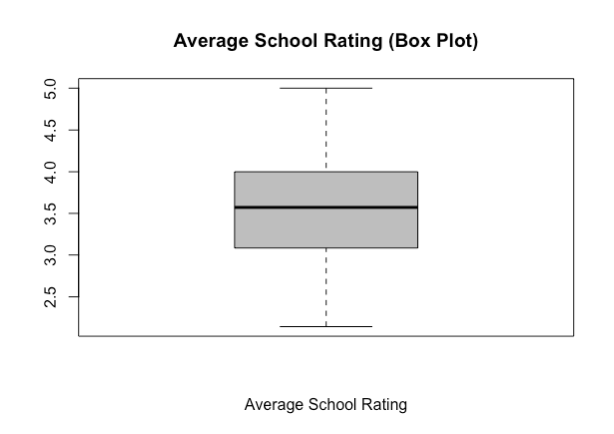
Average School Rating



*Figure 3. Scatter Plot for Average School Rating*



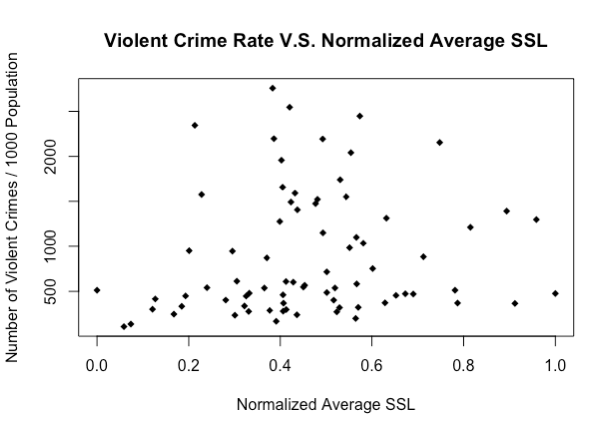
*Figure 4. Histogram for Average School Rating*



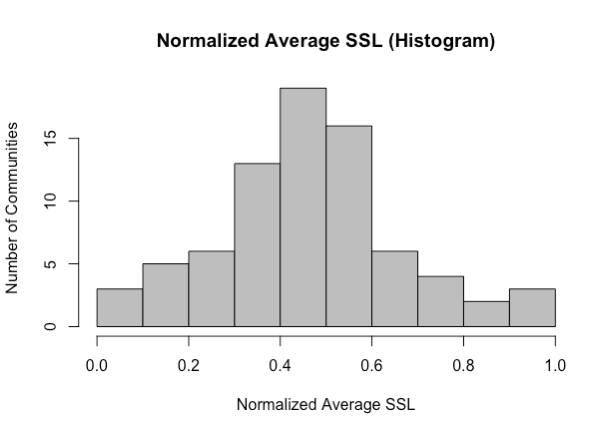
*Figure 5. Box Plot for Average School Rating*

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.

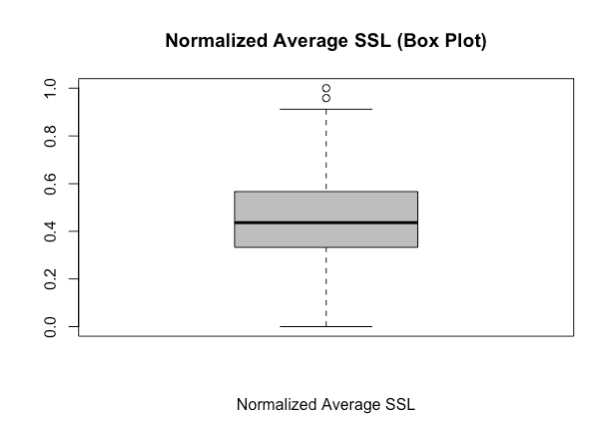
Average SSL Rating



*Figure 6. Scatter Plot for Normalized Average SSL Rating*



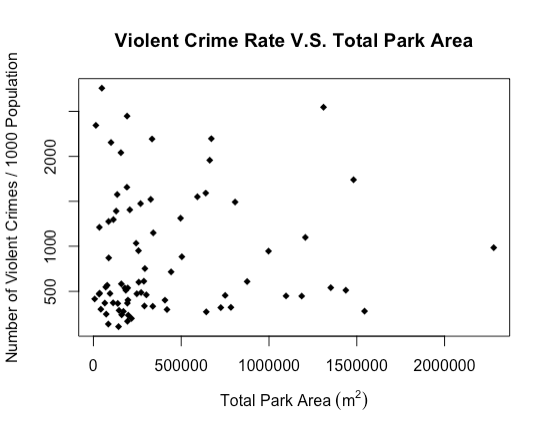
*Figure 7. Histogram for Normalized Average SSL Rating*



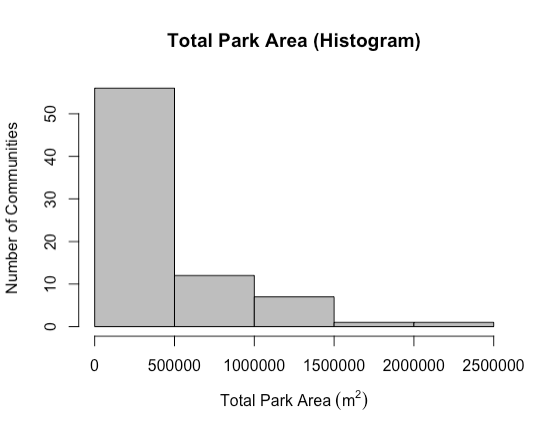
*Figure 8. Box Plot for Normalized Average SSL Rating*

For the purpose of clarity, the average SSL rating is 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.

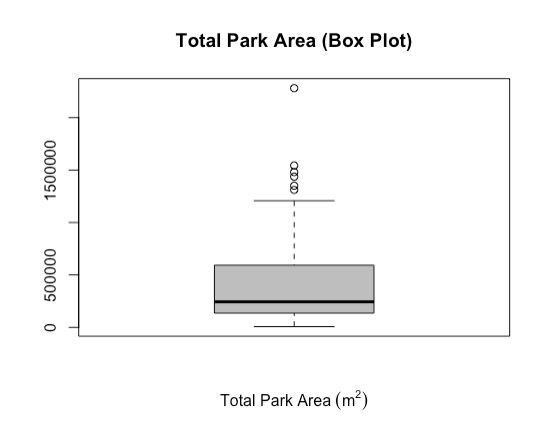
Total Park Area



*Figure 9. Scatter Plot for Total Park Area*



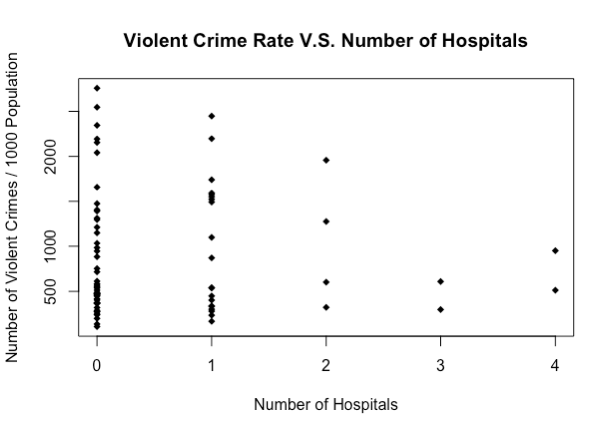
*Figure 10. Histogram for Total Park Area*



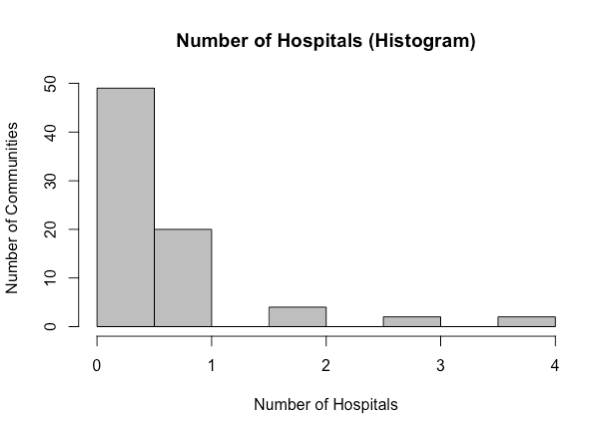
*Figure 11. Box Plot for Total Park Area*

Over 50 communities only have less than 500000 m2 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 real correlation.

Number of Hospitals



*Figure 12. Scatter Plot for Number of Hospitals*

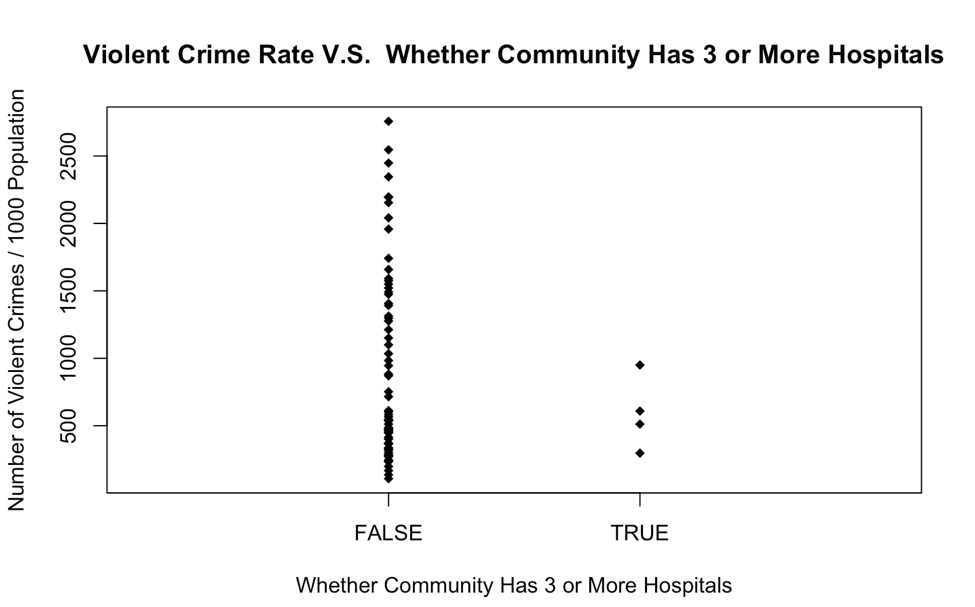


*Figure 13. Histogram for Number of Hospitals*



*Figure 14. Box Plot for Number of Hospitals*

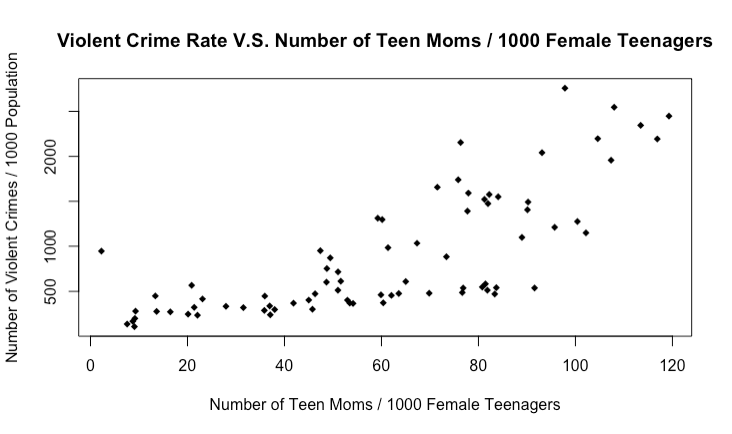
As shown in all the plots for number of hospitals, the majority of the communities does not have a hospital. There are 4 outliers, Uptown and West Town with 3 hospitals, and Lincoln Square and New West Side with 4 hospitals.



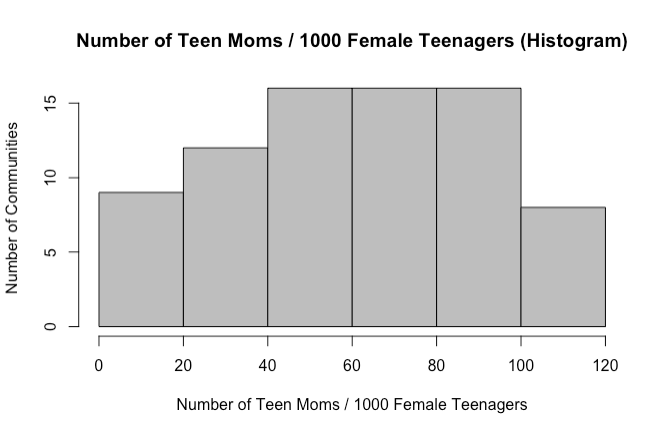
*Figure 15. Scatter Plot for Whether Community Has 3 or More Hospitals*

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.

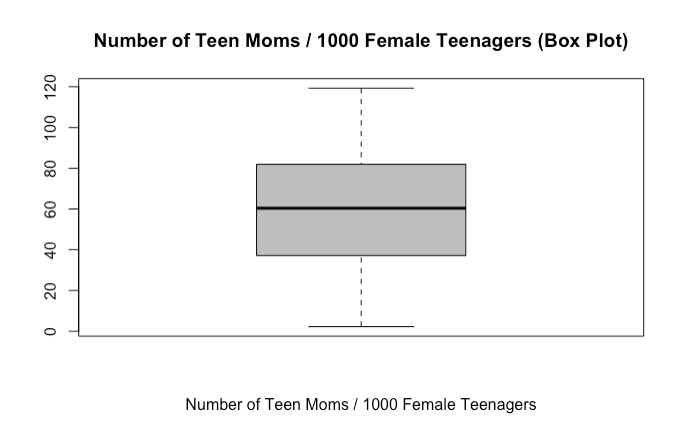
Birth Rate by Teenage Mothers



*Figure 16. Scatter Plot for Birth Rate by Teenage Mothers*



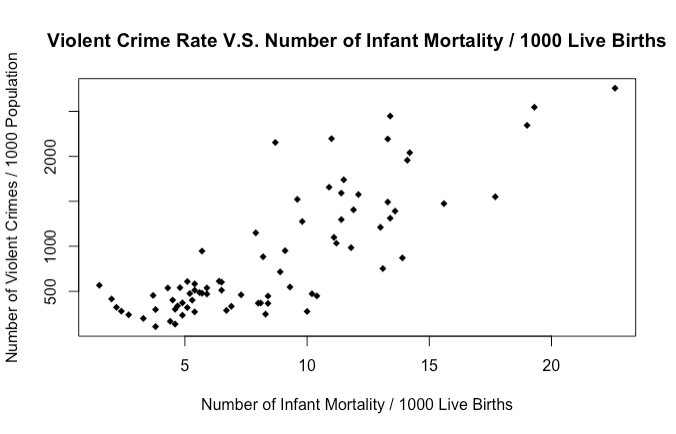
*Figure 17. Histogram Plot for Birth Rate by Teenage Mothers*



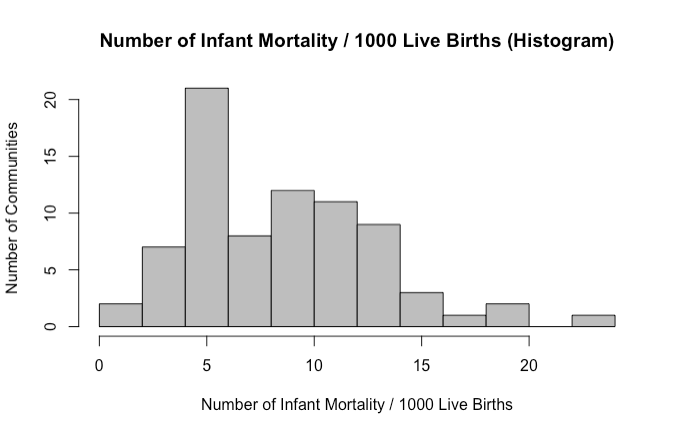
*Figure 18. Box Plot for Birth Rate by Teenage Mothers*

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.

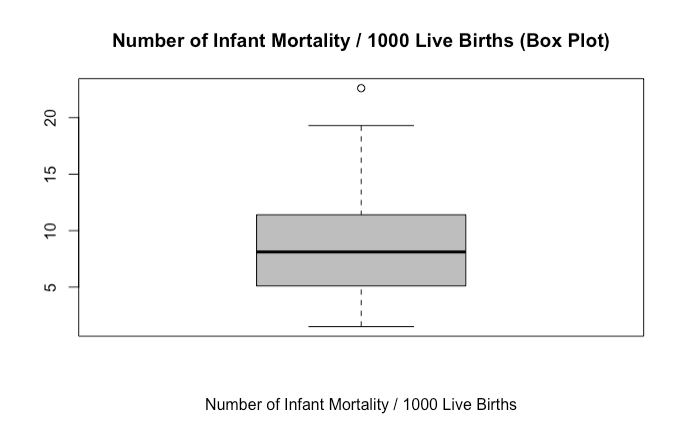
Infant Mortality Rate



*Figure 19. Scatter Plot for Infant Mortality Rate*



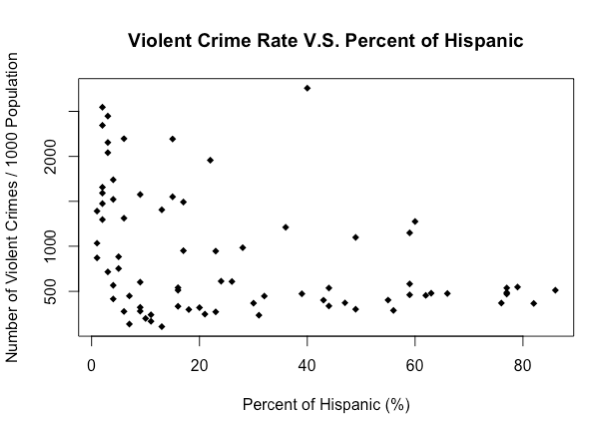
*Figure 20. Histogram for Infant Mortality Rate*



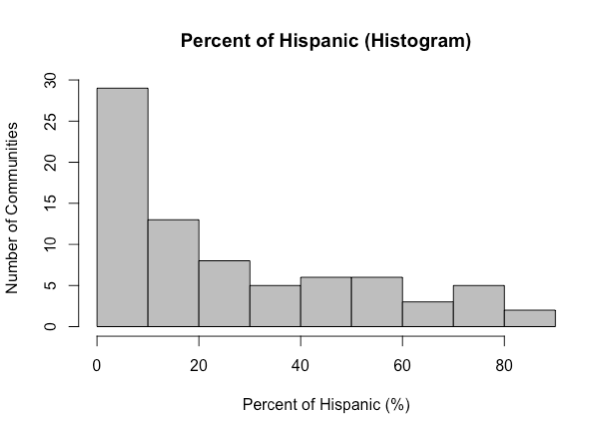
*Figure 21. Box Plot for Infant Mortality Rate*

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 population. The only outlier indicated in the box plot is Fuller Park, which has 22.6 infant mortalities per 1000 live births. This result matches what is observed from the plotting of the class variable.

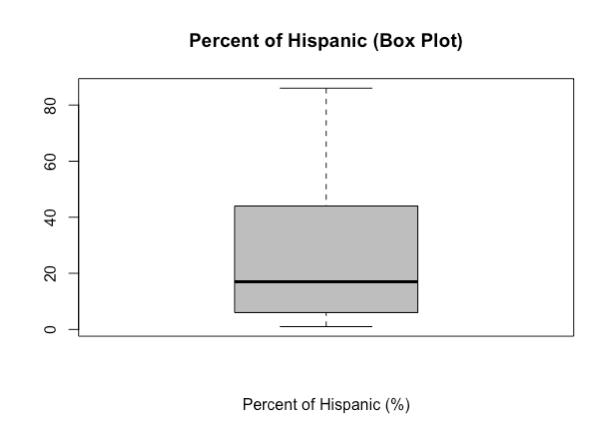
Proportion of Hispanic People



*Figure 22. Scatter Plot for Proportion of Hispanic People*



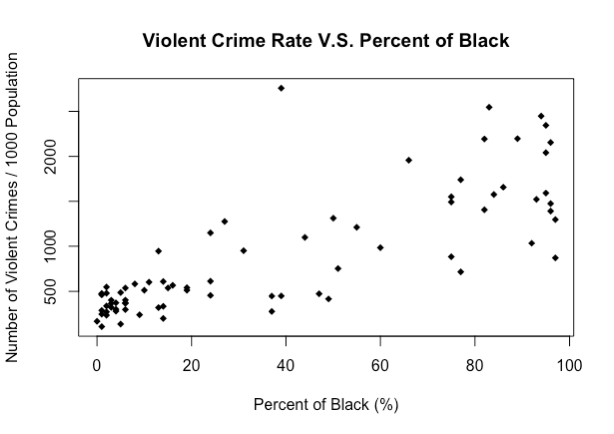
*Figure 23. Histogram for Proportion of Hispanic People*



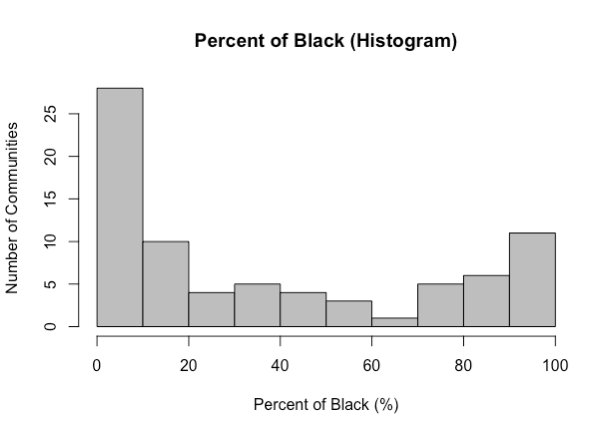
*Figure 24. Box Plot for Proportion of Hispanic People*

The number of communities exponentially decreases while the percentage of hispanic people increases from 0 to 90%. The negative exponential relationship can be seen from the scatter plot.

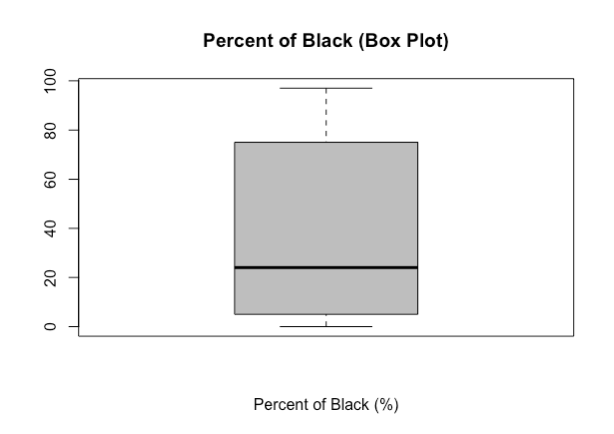
Proportion of Black People



*Figure 25. Scatter Plot for Proportion of Black People*



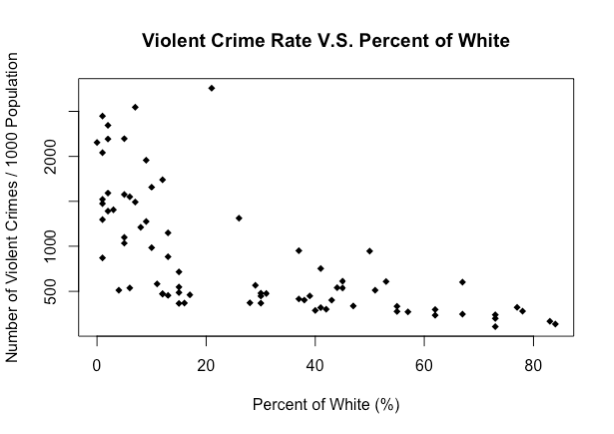
*Figure 26. Histogram for Proportion of Black People*



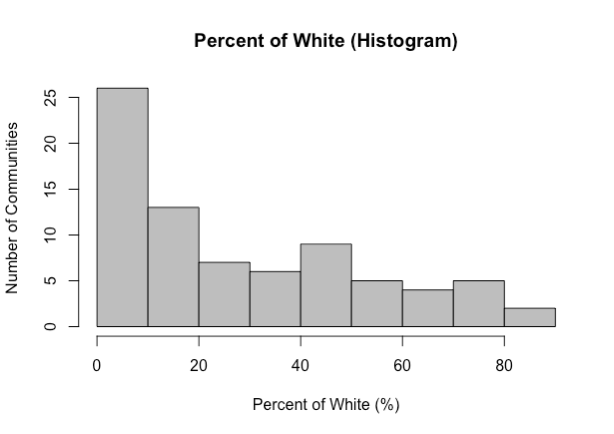
*Figure 27. Box Plot for Proportion of Black People*

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

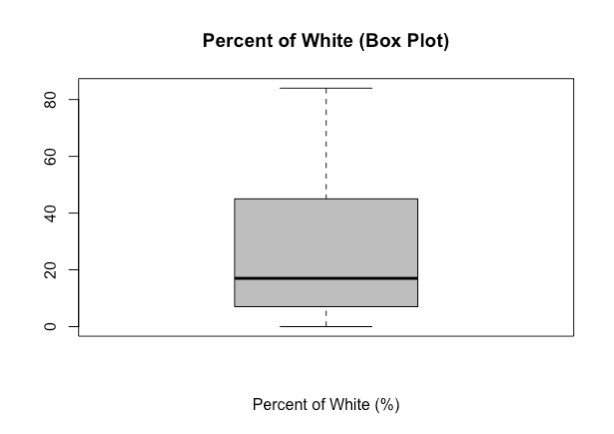
Proportion of White People



*Figure 28. Scatter Plot for Proportion of White People*



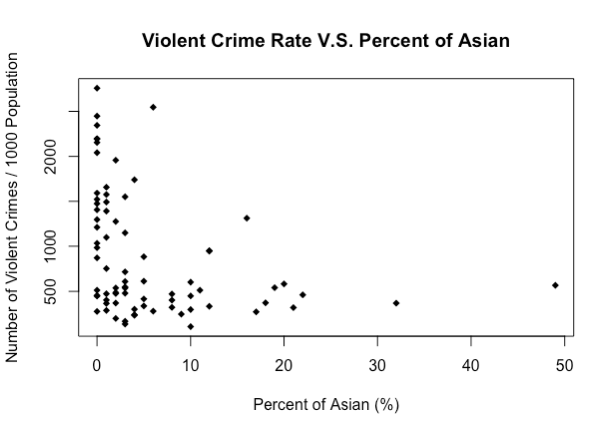
*Figure 29. Histogram for Proportion of White People*



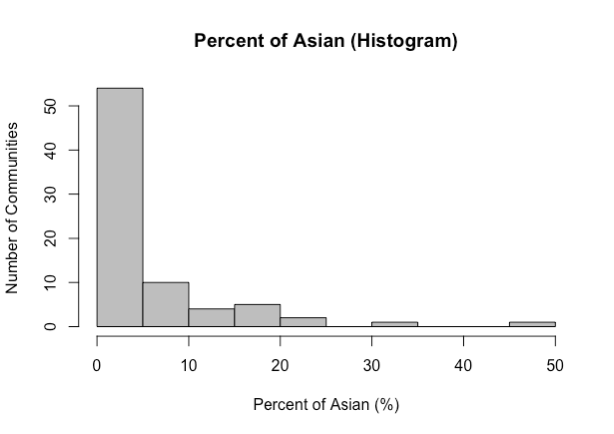
*Figure 30. Box Plot for Proportion of White People*

The number of communities exponentially decreases as the percentage of white people increases. As shown in the scatter plot, there is a negative exponential relationship.

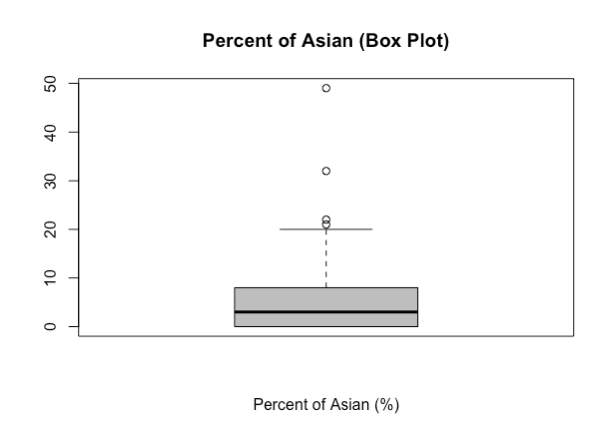
Proportion of Asians



*Figure 31. Scatter Plot for Proportion of Asian People*



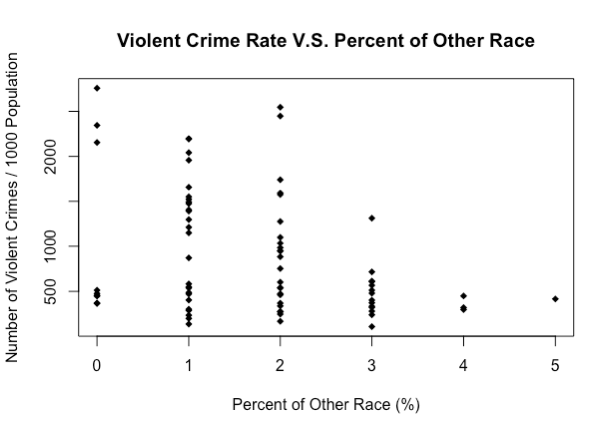
*Figure 32. Histogram for Proportion of Asian People*



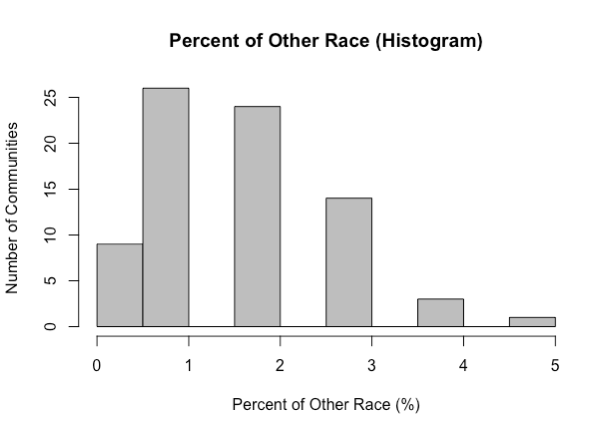
*Figure 33. Box Plot for Proportion of Asian People*

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, however, it is not clear due to the low number of samples at high percentage.

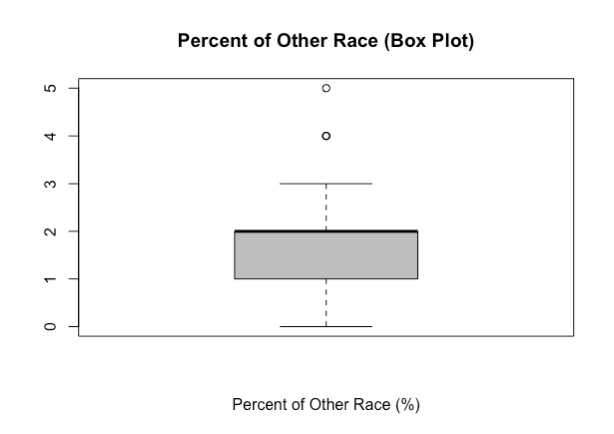
Proportion of Other Races



*Figure 34. Scatter Plot for Proportion of Other Races*



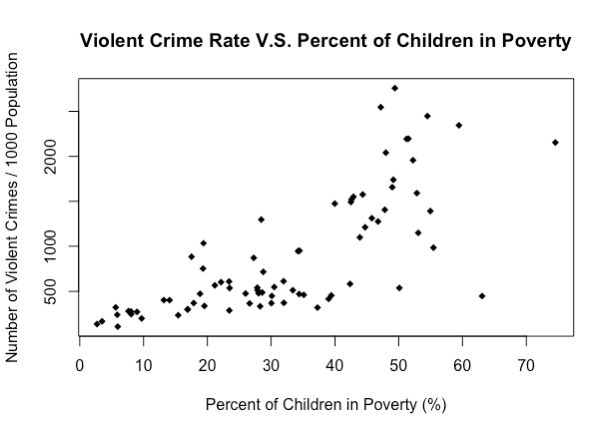
*Figure 35. Histogram for Proportion of Other Races*



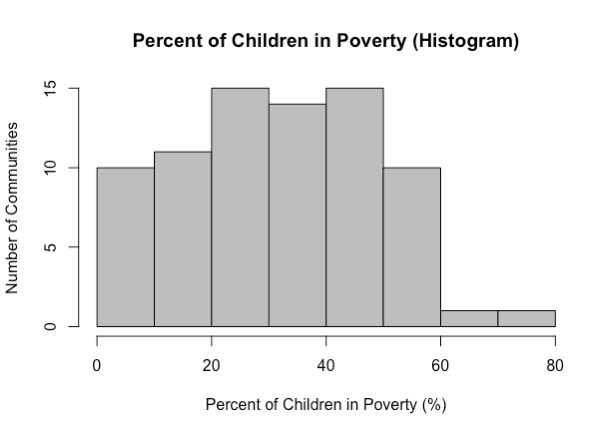
*Figure 36. Box Plot for Proportion of Other Races*

The percentage of other races does not indicate any correlation in particular.

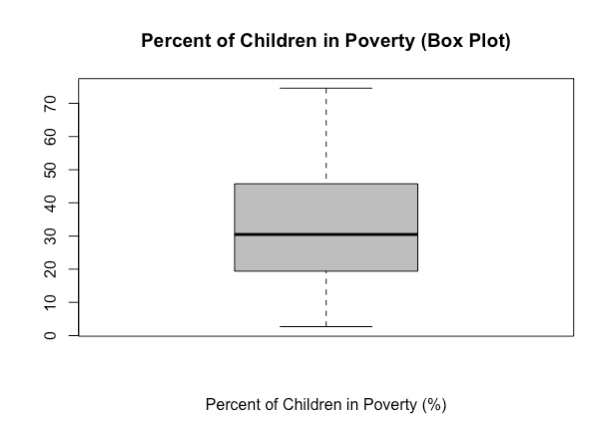
Percent of Children in Poverty



*Figure 37. Scatter Plot for Percent of Children in Poverty*



*Figure 38. Histogram for Percent of Children in Poverty*



*Figure 39. 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 number of violent crimes. The percentage of children in poverty ranges from 0 to 80%, concentrated at the range of 0 to 60%.

1. Supervised and Unsupervised Learning Progress

Numeric regression and Clustering analysis were successfully conducted as of current. However, association rule mining is not completed yet.

Numeric Regression Analysis

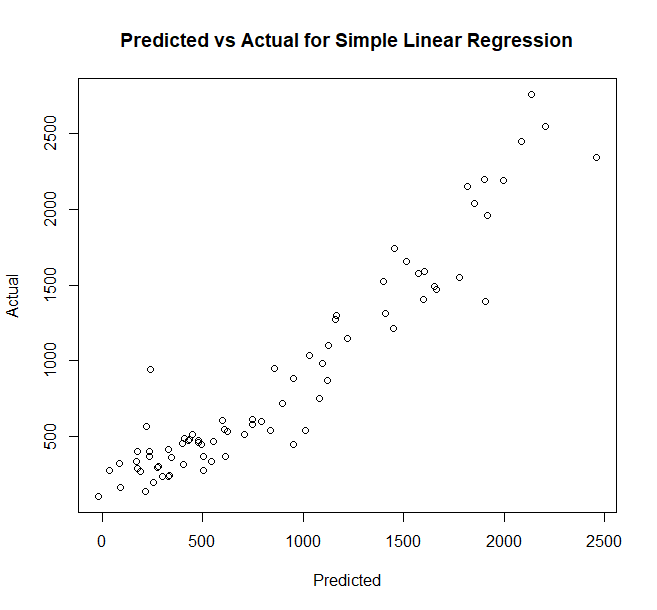
Three types of regressions were used: ordinary least squares (OLS) linear regression, elastic net regularized regression, and generalized additive linear model. Elastic nets is a mixture of lasso and ridge regression, both of which penalize the coefficients in regular OLS regression in order to prevent 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 all the coefficients for each explanatory variable is penalized. Lasso regression can force coefficients of uncorrelated features to be zero (which can be seen as feature selection), though this may lead to information loss. Thus, elastic nets are seen as “best of both worlds”. However, as seen from 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 believed GAM might be proven to give fruitful results, since the GAM algorithm attempts to fit polynomial splines on the explanatory variables and conduct an additive linear regression on these transformed predictors.

10-fold cross validation was used to obtain the alpha and lambda parameters for the elastic nets regression. Note that alpha is 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.

Leave-one-out cross validation was used to obtain three performance metrics for each regression type, which include: root mean squared error, (RMSE), coefficient of determination (R2), and mean absolute error (MAE).

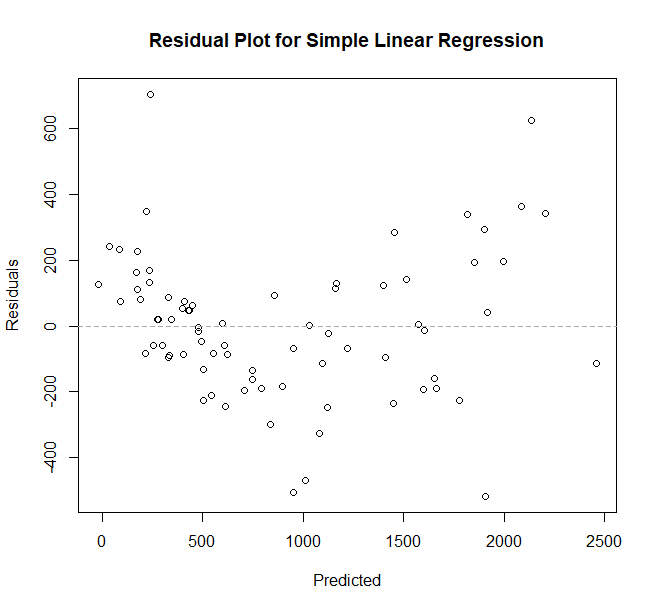
OLS Linear Regression

Figure 1 shows the predicted class variables versus the actual class variables for OLS linear regression. Ideally, the points should be uniformly distributed around the line . 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.



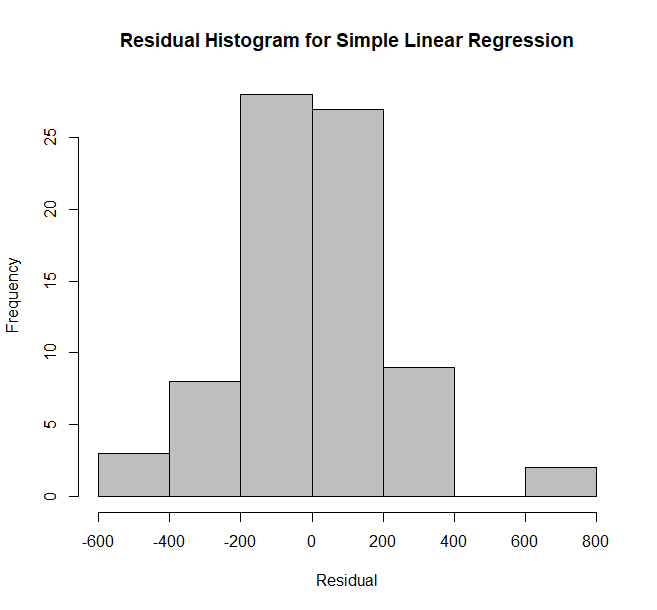
*Figure 40. Predicted Class Values vs Actual Values for OLS Linear Regression*

This concavity is more pronounced in the residual plot shown in Figure 2. Evidently, the scatter plots show that the data’s uniformity around the line is not independent of the value of x (the predicted values), which suggests linear regression may not be the best algorithm.



*Figure 41. Residual Plot for OLS Linear Regression.*

Figure 3 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 a histogram.



*Figure 42. Residual Histogram for OLS Linear Regression.*

Table 2 shows the performance metrics of this algorithm.

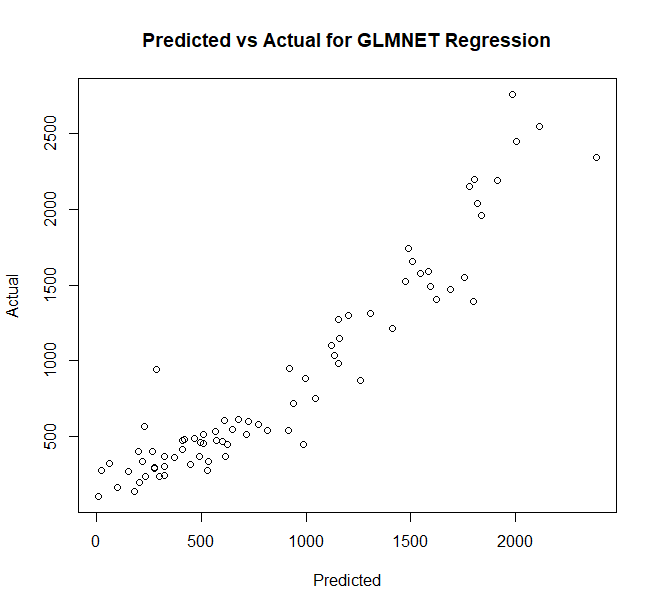
*Table 2. Performance Metrics for OLS Linear Regression*

|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 285.243497339416 | 0.822397725231063 | 206.537526304179 |

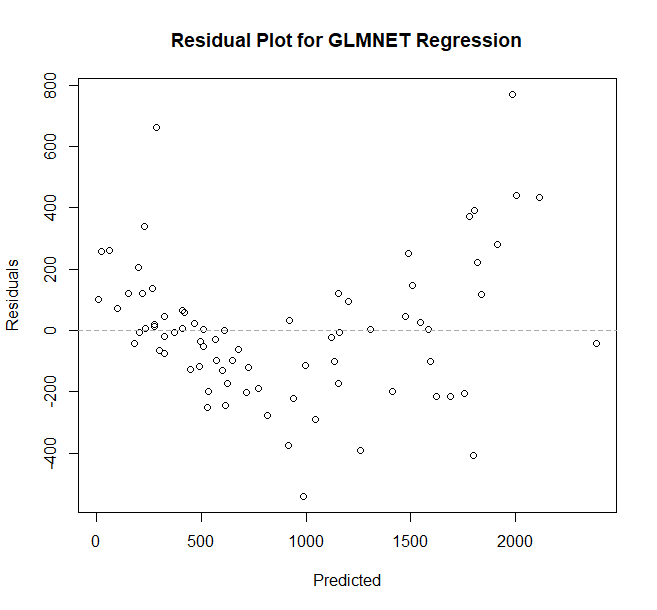
Elastic Net Regression

Using 10-fold cross validation, the best alpha and lambda values were determined to be 0 (100% ridge regression) and 1.

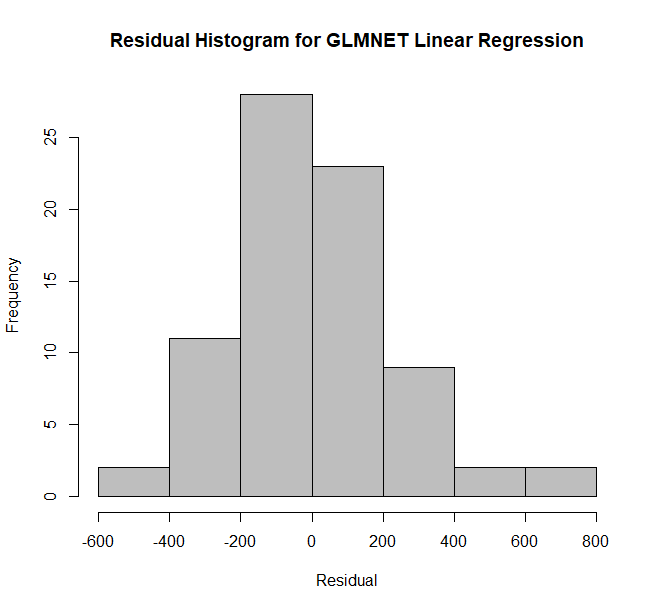
Figure 4, Figure 5, and Figure 6 show similar types of plots as ones shown for OLS regression in the last subsection. It seems that there is not a large visual difference between elastic net and OLS, though the histogram distribution slightly changed.



*Figure 43. Predicted Class Values vs Actual Values for Elastic Net Regression.*



*Figure 44. Residual Plot for Elastic Net Regression.*



*Figure 45. Residual Histogram for Elastic Net Regression.*

Ultimately, Table 3 shows that the performance metrics only improved very slightly from OLS. The changed digits are bolded for ease of reading.

*Table 3. Performance Metrics for Elastic Net Linear Regression*

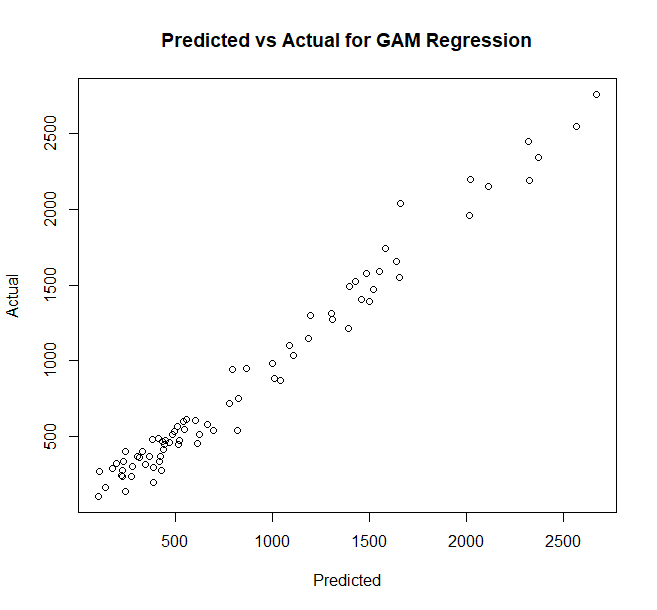
|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 284.**112757633739** | 0.82**3577225374136** | 206.53**3355885329** |

GAM Regression

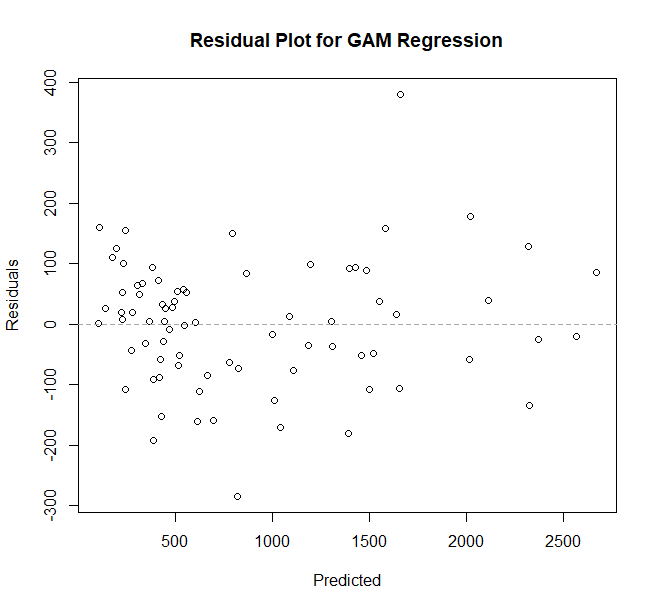
Using the scatter plots of each explanatory variable versus the class variable in the “Exploratory Data Analysis” section of this report, the relevant explanatory variables that may have nonlinear relationships with the class variable were identified. These include: total park area, birth rate for teenage 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 the function is a function that fits a spline between the input and the outcome.

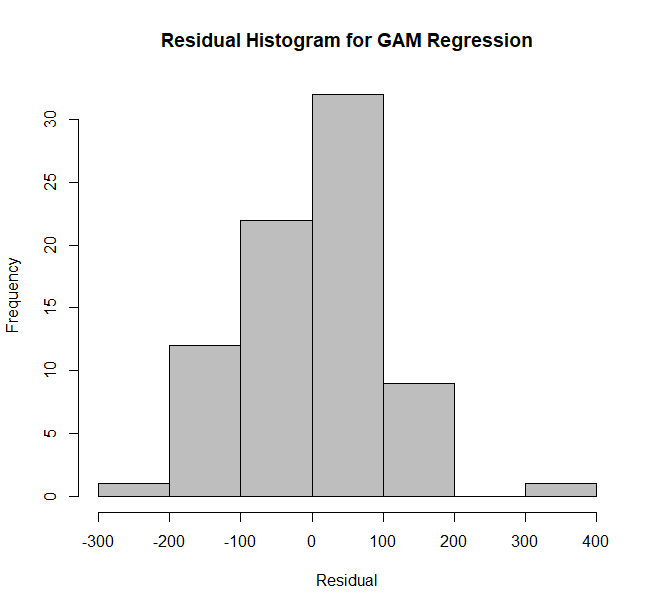
Figure 7 seems to indicate that the GAM model built on all 77 community area data points reduces the nonlinear concavity. In addition, Figure 8 and Figure 9 shows that the range of the residual decreased to . These all seem to indicate that the GAM model fits the data better.



*Figure 46. Predicted Class Values vs Actual Values for GAM Regression.*



*Figure 47. Residual Plot for GAM Regression.*



*Figure 48. Residual Histogram for GAM Regression.*

However, the performance metrics obtained from leave-one-out cross validation shown in Table 4 indicates that the GAM model performance on unseen data is actually worse than the previous two regression algorithms. With leave-one-out cross validation, 76 GAM models are each built from randomly selected 76 community area data points and tested on the single remaining community area.

*Table 4. Performance Metrics for GAM Regression.*

|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| **308.050825959162** | 0.**796551460938803** | **188.925268679306** |

Table 5 displays the performance metric of the GAM model built with all 77 community area data points. It is clear that by comparing Table 5 and Table 4, leaving even one data point out of the model significantly changes the performance metrics. Thus, the GAM algorithm heavily overfits.

*Table 5. Performance Metrics for GAM Regression (Built With All Data Points)*

|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| **103.30301646473** | 0.**976500401871001** | **79.55816922112** |

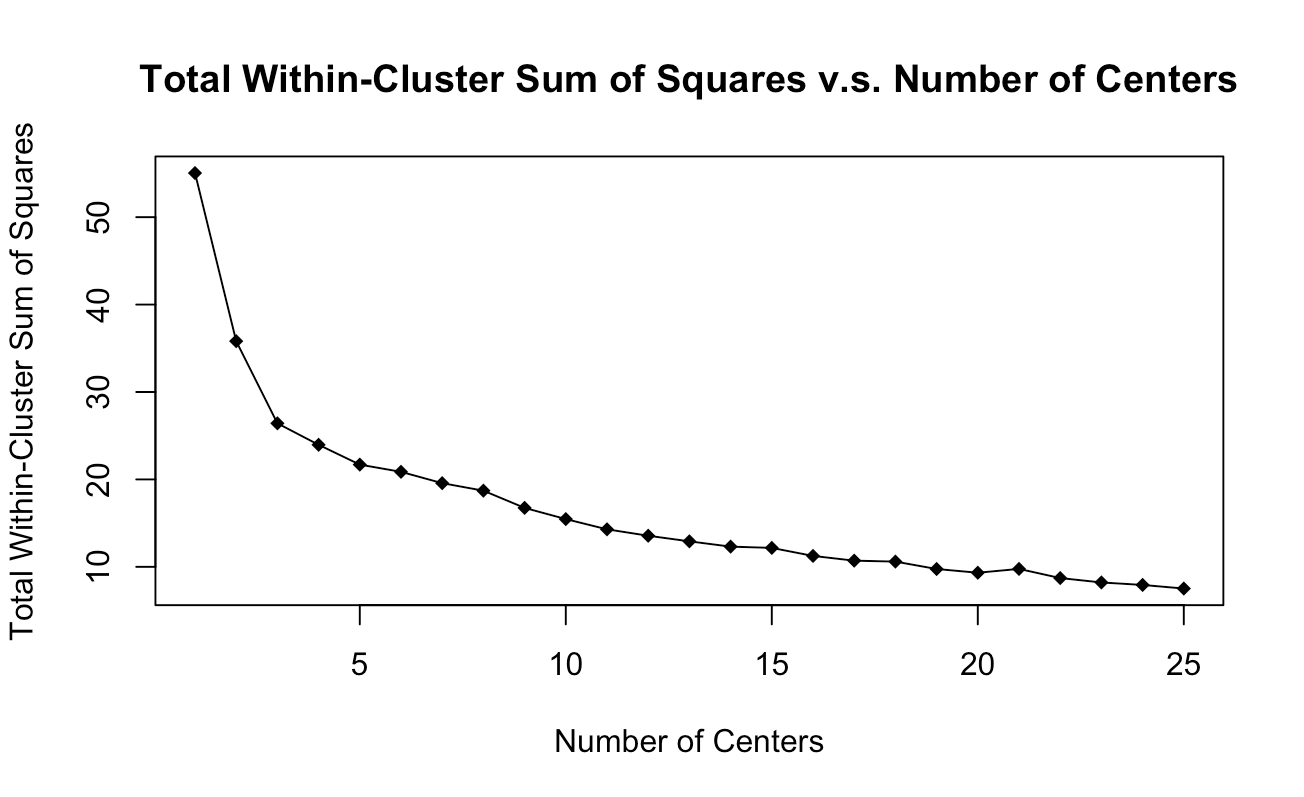
It seems that out of all three regression algorithms, elastic net regression performs the best, though it is only marginally better than OLS regression. Future recommendations include using other nonlinear regression algorithms other than GAM, as its fitted splines may use polynomials with higher than necessary orders.

Clustering Analysis

The K-means clustering analysis is performed on the 12 explanatory variables using R Language’s  method, “kmeans”. Since clustering works the best under isotropic condition, all of the explanatory variables are scaled to the range of 0 to 1.

Analysis of Number of Centers

In the first part, as changing the number of clustering centers from 1 to 25, the total within-cluster sum of squares data is collected, and the result is shown in Figure 49.

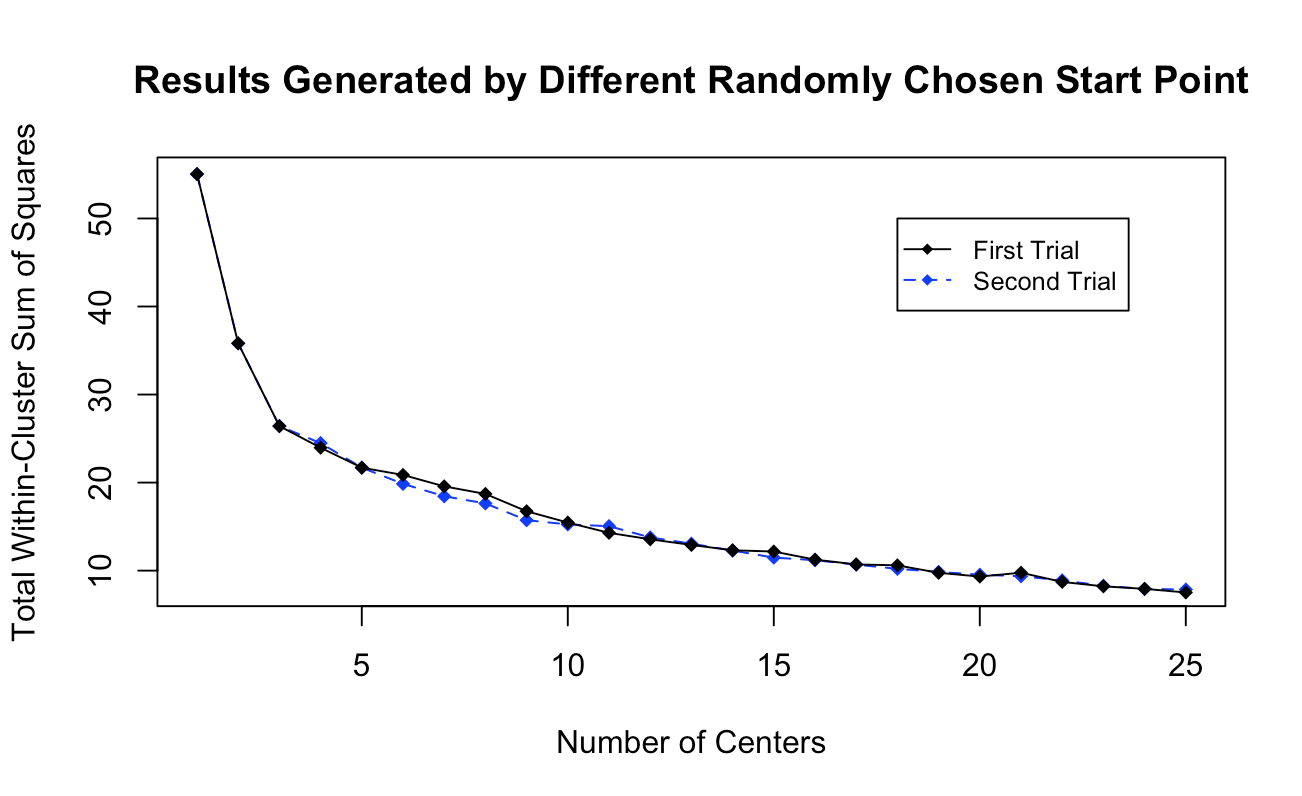


*Figure 49. Total within-cluster sum of squares versus the number of centers*

As shown in Figure 49, 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.

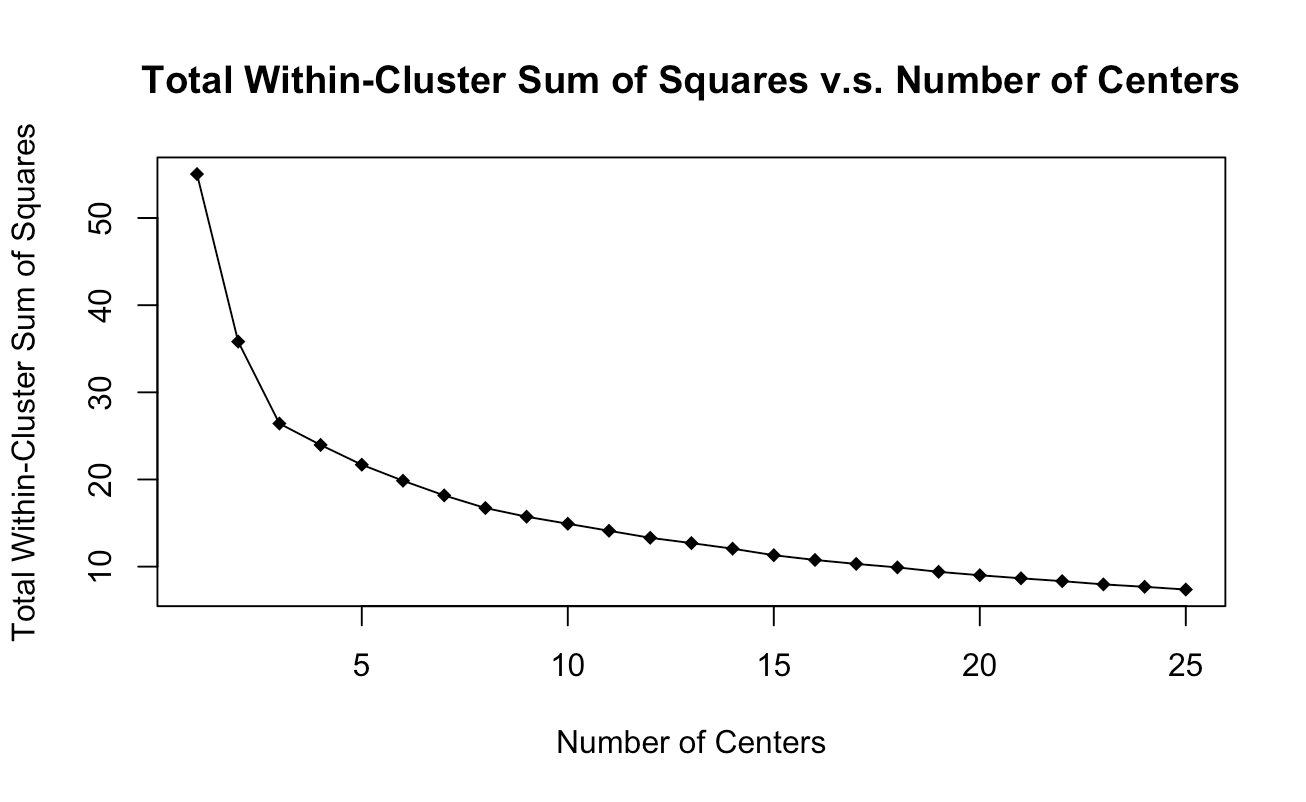
Analysis of Different Starting Points for Centers

After the relationship between sum of squares and number of clustering centers is explored, 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 50.



*Figure 50. 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 51 indicates the average sum of squares obtained from the 50 sets of centers for each number of centers.

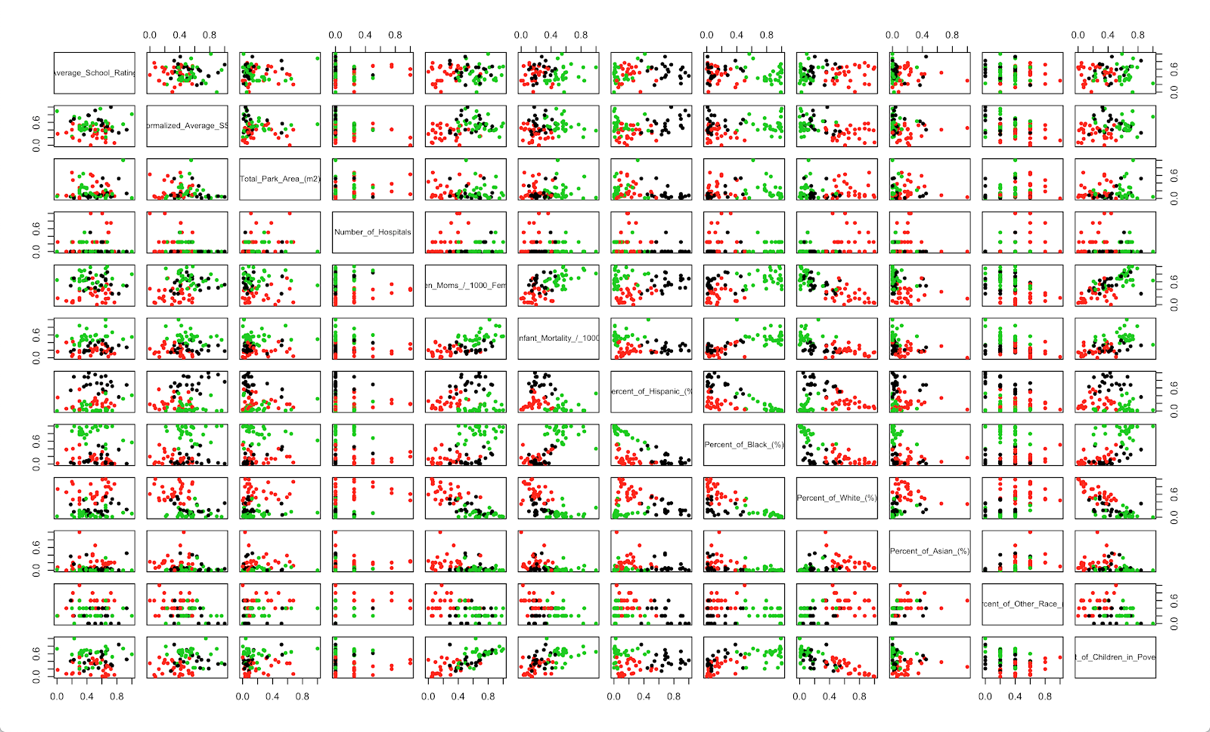


*Figure 51.* Average sum of squares with 50 sets of starting points

By looking at the above figure, it becomes evident that the significant drop in sum of squares still stops at when 3 clustering centers is selected. Therefore, for the clustering analysis, the number of clustering centers is set to 3.

Clustering between Each Two Explanatory Variables

After the number of clustering centers is determined, a plot is made to indicate the clustering between each two explanatory variables with 3 clustering centers, as shown in Figure 52.



*Figure 52. Clustering Relationship between Each Two Explanatory Variables*

Figure 52 shows that most of the plots are clearly divided into clusters at column 5, 6, 7, 8, 9, and 12, respectively representing the teen mom birth rate, infant mortality rate, percent of Hispanic, percent of black, percent of white, and percent of children in poverty. For example, it can be observed in column 5, row 6, the clustering between teenage mom rate and infant mortality rate, that usually low teenage mom rate is associated with low infant mortality rate, and high teenage mom rate is associated with high infant mortality rate. More details will be discussed and included in the final report of the project.

Association Rule Mining Analysis

Association rule mining is still in progress as of current. Current plans are to preprocess and create a CSV file of the explanatory variables in R. This CSV file will contain data that is interpretable with Python’s Apriori’s library. The reason this Python will be used is because the authors of this report are not familiar with association rule mining libraries in R, and because the authors believe they should gain more experience in Python.

1. Bibliography

|  |  |  |
| --- | --- | --- |
| [1] | 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]. |  |
| [2] | 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]. |  |
| [3] | 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]. |  |
| [4] | 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]. |  |
| [5] | 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]. |  |
| [6] | 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]. |  |
| [7] | 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]. |  |
| [8] | 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]. |  |
| [9] | 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]. |  |
| [10] | 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]. |  |

1. Appendix

Data Collection and Preprocessing Code

Class Variable

library(tidyr)

library(dplyr)

library(plyr)

# crime data from 2001

crime <- data.frame(crimes2001$X...ID,

crimes2001$Date,

crimes2001$Primary.Type,

crimes2001$Description,

crimes2001$Location.Description,

crimes2001$Community.Area)

names(crime) <- c('id', 'date', 'type', 'description', 'location', 'community')

crime <- crime[!(is.na(crime$community) | crime$community=='' | crime$community=='0'), ]

# dividing offense involving children into violent vs non-violent

crime$type <- ifelse(grepl(('CRIM SEX ABUSE BY FAM MEMBER'), crime$description)

| grepl(('CHILD ABUSE'), crime$description)

| grepl(('AGG SEX ASSLT OF CHILD FAM MBR'), crime$description)

| grepl(('CHILD ABDUCTION'), crime$description)

| grepl(('AGG CRIM SEX ABUSE FAM MEMBER'), crime$description)

| grepl(('SEX ASSLT OF CHILD BY FAM MBR'), crime$description)

| grepl(('CRIM SEX ABUSE BY FAM MEMBER'), crime$description),

gsub('OFFENSE INVOLVING CHILDREN', 'VIOLENT OFFENSE INVOLVING CHILDREN', crime$type),

gsub('OFFENSE INVOLVING CHILDREN', 'NON-VIOLENT OFFENSE INVOLVING CHILDREN', crime$type))

# generating counts

commu\_crime <- data.frame(crime$type, crime$community)

names(commu\_crime) <- c('type', 'community')

count\_commu\_crime <- ddply(commu\_crime, .(commu\_crime$community, commu\_crime$type), nrow)

names(count\_commu\_crime) <- c("community", "type", "count")

count <- spread(count\_commu\_crime, key = type, value = count)

count[is.na(count)] <- 0

# population

population <- data.frame(population\_chicago$GeogKey,

population\_chicago$Geog,

population\_chicago$`Total Population`)

names(population) <- c('community', 'community name', 'population(2010)')

# sum violent crimes and total crimes

names(sum\_crime) <- c('community')

sum\_crime$violent\_crime <- NA

sum\_crime$total\_crime <- NA

sum\_crime$violent\_crime <- rowSums(count[, c('ASSAULT', 'BATTERY', 'CRIM SEXUAL ASSAULT',

'HOMICIDE', 'KIDNAPPING',

'VIOLENT OFFENSE INVOLVING CHILDREN',

'PUBLIC PEACE VIOLATION', 'RITUALISM',

'ROBBERY', 'SEX OFFENSE', 'WEAPONS VIOLATION')])

sum\_crime$total\_crime <- rowSums(count[, !(colnames(count) == "community")])

#merge sum\_crime and population

sum\_crime <- merge(x = sum\_crime, y = population, by.x = 'community', by.y = 'community', all = TRUE)

sum\_crime <- sum\_crime[, c(1, 4, 5, 2, 3)]

Average School Rating

# Creating new dataframe for school info

schooldf <- data.frame(School\_Profile\_Information$School\_ID,

School\_Profile\_Information$Short\_Name,

School\_Profile\_Information$Overall\_Rating,

School\_Profile\_Information$Rating\_Status,

School\_Locations$COMMAREA,

School\_Locations$WARD\_15)

community <- data.frame(census\_data\_by\_community\_area$communityAreaNumber,

census\_data\_by\_community\_area$Community)

names(schooldf) <- c('id', 'name', 'rating', 'status', 'community', 'ward')

names(community) <- c('number', 'name')

# Converting ratings to numbers

schooldf[, 'rating'] = toupper(schooldf[, 'rating'])

schooldf$rating\_num <- NA

schooldf$rating\_num[schooldf$rating=='INABILITY TO RATE'] <- -1

schooldf$rating\_num[is.na(schooldf$rating)] <- -1

schooldf$rating\_num[schooldf$rating=='LEVEL 3'] <- 1

schooldf$rating\_num[schooldf$rating=='LEVEL 2'] <- 2

schooldf$rating\_num[schooldf$rating=='LEVEL 2+'] <- 3

schooldf$rating\_num[schooldf$rating=='LEVEL 1'] <- 4

schooldf$rating\_num[schooldf$rating=='LEVEL 1+'] <- 5

# Calculating average rating by community area

avg <- aggregate(schooldf$rating\_num, list(schooldf$community), mean)

names(avg) <- c('community', 'avg\_rating')

# Add community number

community[, 'name'] = toupper(community[, 'name'])

community[, 'name'] <- gsub('[^[:alnum:][:space:]]', '', community[, 'name'])

# Join two data frames

avg\_commu <- merge(x = community, y = avg, by.x = 'name', by.y = 'community', all = TRUE)

avg\_commu\_ordered <- avg\_commu[, c(2, 1, 3)]

avg\_commu\_ordered <- avg\_commu\_ordered[with(avg\_commu\_ordered, order(number)), ]

Average SSL Rating

# Create data frame for ssl and community

ssl <- data.frame(Strategic\_Subject\_List$`SSL SCORE`,

Strategic\_Subject\_List$`COMMUNITY AREA`)

community <- data.frame(census\_data\_by\_community\_area$communityAreaNumber,

census\_data\_by\_community\_area$Community)

names(ssl) <- c('score', 'community')

names(community) <- c('number', 'name')

# Eliminate rows with blank community

ssl <- ssl[!(is.na(ssl$community) | ssl$community==''), ]

# Standardize community names

ssl[, 'community'] = toupper(ssl[, 'community'])

ssl[, 'community'] <- gsub('[^[:alnum:][:space:]]', '', ssl[, 'community'])

# Calculate the average

avg <- aggregate(ssl$score, list(ssl$community), mean)

names(avg) <- c('community', 'avg\_rating')

# Add community number

community[, 'name'] = toupper(community[, 'name'])

community[, 'name'] <- gsub('[^[:alnum:][:space:]]', '', community[, 'name'])

# Join two data frames

avg\_commu <- merge(x = community, y = avg, by.x = 'name', by.y = 'community', all = TRUE)

avg\_commu\_ordered <- avg\_commu[, c(2, 1, 3)]

avg\_commu\_ordered <- avg\_commu\_ordered[with(avg\_commu\_ordered, order(number)), ]

Total Park Area

##########################################

#parks by community area

##########################################

library**(**rgdal**)**

library**(**sp**)**

library**(**dplyr**)**

library**(**sf**)**

library**(**tidyverse**)**

library**(**raster**)**

#import the shape files

chicagoparks **<-** readOGR**(**'4A/MSCI 446/R/dataToPreprocess/chicagoparksshapefile', 'geo\_export\_287c1e81-adfc-4076-bbd4-7ac4b1ca62c2'**)**

chicagocommunityareas **<-** readOGR**(**'4A/MSCI 446/R/dataToPreprocess/communityareashapefile', 'geo\_export\_f2c553e7-eb62-4773-9655-8037a1bdd109', stringsAsFactors **=** **FALSE)**

#RUN THIS CODE FOR TOTAL PARK AREA FOR EACH COMMUNITY AREA#

totalParkAreaForCommunityAreas **<-** rep**(**0, nrow**(**chicagocommunityareas**))**

**for(**i **in** 1**:**nrow**(**chicagocommunityareas**))** **{**

totalArea **<-** 0

**for** **(**j **in** 1**:**nrow**(**chicagoparks**))** **{**

#get intersection of community area & park

intersect **<-** intersect**(**chicagocommunityareas**[**i, **]**, chicagoparks**[**j, **])**

**if** **(!**is.null**(**intersect**))** **{**

#if intersection!=null, add to totalArea for current community area

totalArea **<-** totalArea **+** area**(**intersect**)**

**}**

**}**

totalParkAreaForCommunityAreas**[**i**]** **<-** totalArea

**}**

#create the dataframe consisting of three columns: communityArea, communityAreaNumber, and totalParkArea

library**(**readxl**)**

censusdata **<-** read\_excel**(**"4A/MSCI 446/R/dataToPreprocess/Census-Data-by-Chicago-Community-Area-2017 (2).xlsx"**)**

censusdata **<-** data.frame**(**censusdata**$**Community, censusdata**$**CommunityAreaNumber**)**

names**(**censusdata**)** **<-** c**(**'Community', 'communityAreaNumber'**)**

censusdata**$**Community **<-** toupper**(**censusdata**$**Community**)**

chicagocommunityareas@data**$**community**[**75**]** **<-** 'O\'HARE' #naming difference

communityAreaNumber **<-** rep**(**0, nrow**(**chicagocommunityareas**))**

#Need to match totalParkAreaForCommunityAreas to each communityAreaNumber

**for(**i **in** 1**:**nrow**(**chicagocommunityareas**))** **{**

communityAreaNumber**[**i**]** **<-** censusdata**[**which**(**censusdata**$**Community**==**chicagocommunityareas@data**$**community**[**i**])**,2**]**

**}**

totalParkAreaDF **<-** data.frame**(**chicagocommunityareas@data**$**community, communityAreaNumber, totalParkAreaForCommunityAreas**)**

names**(**totalParkAreaDF**)** **<-** c**(**'Community', 'communityAreaNumber', 'totalParkArea'**)**

#save the totalParkArea by Community Area Number

write.csv**(**totalParkAreaDF, 'totalParkAreaByCommunityArea.csv'**)**

Number of Hospitals, Teen Mom Birth Rate, Infant Mortality Rate

##########################################

#Public Safety Data

##########################################

library**(**rgdal**)**

library**(**sp**)**

library**(**dplyr**)**

library**(**sf**)**

library**(**tidyverse**)**

library**(**raster**)**

#Code for numHospitalsPerCommunityArea#################################

hospitals **<-** readOGR**(**'4A/MSCI 446/R/dataToPreprocess/Hospitals', 'Hospitals', stringsAsFactors **=** **FALSE)**

numHospitalsPerCommunityArea **<-** as.data.frame**(**table**(**hospitals@data**$**AREA\_NUMBE**))**

names**(**numHospitalsPerCommunityArea**)** **<-** c**(**'communityAreaNum', 'numHospitals'**)**

numHospitalsPerCommunityArea**$**communityAreaNum **<-** as.numeric**(**levels**(**numHospitalsPerCommunityArea**$**communityAreaNum**))**

**for** **(**i **in** 1**:**77**)** **{**

**if** **(**sum**(**numHospitalsPerCommunityArea**$**communityAreaNum **==** i**)** **==** 0**)** **{**

newDF **<-** data.frame**(**i,0**)**

names**(**newDF**)<-**c**(**'communityAreaNum', 'numHospitals'**)**

numHospitalsPerCommunityArea **<-** rbind**(**numHospitalsPerCommunityArea, newDF**)**

**}**

**}**

numHospitalsPerCommunityArea **<-** numHospitalsPerCommunityArea**[**order**(**numHospitalsPerCommunityArea**$**communityAreaNum**)**,**]**

var**(**numHospitalsPerCommunityArea**$**numHospitals**)**

#Code for teenMomRatePerCommunityArea#################################

teenMomsData **<-** read.csv**(**'4A/MSCI 446/R/dataToPreprocess/Public\_Health\_Statistics\_-\_Births\_to\_mothers\_aged\_15-19\_years\_old\_in\_Chicago\_\_by\_year\_\_1999-2009.csv'**)**

teenBirthRates **<-** data.frame**(**teenMomsData**$**Teen.Birth.Rate.1999,

teenMomsData**$**Teen.Birth.Rate..2000,

teenMomsData**$**Teen.Birth.Rate.2001,

teenMomsData**$**Teen.Birth.Rate.2002,

teenMomsData**$**Teen.Birth.Rate.2003,

teenMomsData**$**Teen.Birth.Rate.2004,

teenMomsData**$**Teen.Birth.Rate.2005,

teenMomsData**$**Teen.Birth.Rate.2006,

teenMomsData**$**Teen.Birth.Rate.2007,

teenMomsData**$**Teen.Birth.Rate.2008,

teenMomsData**$**Teen.Birth.Rate.2009**)**

teenBirthRates **<-** teenBirthRates**[**1**:**nrow**(**teenBirthRates**)-**1,**]**

teenBirthRatesTransposed **<-** t**(**teenBirthRates**)**

rownames**(**teenBirthRatesTransposed**)** **<-** **NULL**

colnames**(**teenBirthRatesTransposed**)** **=** seq**(**1**:**77**)**

#find the mean teenMomBirthRate for years 1999-2009. Use this as each community area's "teenMomBirthRate"

teenMomRatePerCommunityAreaVec**=**c**()**

**for(**i **in** 1**:**ncol**(**teenBirthRatesTransposed**)){**

teenMomRatePerCommunityAreaVec**[**i**]** **=** mean**(**teenBirthRatesTransposed**[**,i**]**, na.rm **=** **FALSE)**

**}**

teenMomRatePerCommunityArea **<-** data.frame**(**teenMomsData**[**1**:**nrow**(**teenMomsData**)-**1,1**]**, teenMomRatePerCommunityAreaVec**)**

names**(**teenMomRatePerCommunityArea**)** **<-** c**(**'communityAreaNum', 'teenMomRate'**)**

#Code for infantMortalityRatePerCommunityArea#################################

infantMortalityData **<-** read.csv**(**'4A/MSCI 446/R/dataToPreprocess/Public\_Health\_Statistics-\_Infant\_mortality\_in\_Chicago\_\_2005\_\_2009.csv'**)**

infantMortalityData **<-** infantMortalityData**[**1**:**nrow**(**infantMortalityData**)-**1,**]**

infantMortalityRatePerCommunityArea **<-** data.frame**(**infantMortalityData**$**ﮮCommunity.Area, infantMortalityData**$**Average.Infant.Mortality.Rate.2005...2009**)**

names**(**infantMortalityRatePerCommunityArea**)** **<-** c**(**'communityAreaNum', 'infantMortalityRate'**)**

remove**(**infantMortalityData**)**

#write all three to csv

publicHealthData **<-** data.frame**(**infantMortalityRatePerCommunityArea**$**communityAreaNum,

numHospitalsPerCommunityArea**$**numHospitals,

teenMomRatePerCommunityArea**$**teenMomRate,

infantMortalityRatePerCommunityArea**$**infantMortalityRate**)**

names**(**publicHealthData**)** **<-** c**(**'communityAreaNum',

'numHospitals',

'teenMomRate',

'infantMortalityRate'**)**

write.csv**(**publicHealthData, 'publicHealth.csv'**)**

Proportion of Different Races, and Percent of Children in Poverty

##########################################

#poverty & race by community area

##########################################

#did most of the conversion in excel, and using R to just create a csv of it.

library**(**readxl**)**

censusdata **<-** read\_excel**(**"4A/MSCI 446/R/dataToPreprocess/Census-Data-by-Chicago-Community-Area-2017 (2).xlsx"**)**

censusdata **<-** data.frame**(**censusdata**$**Community,

censusdata**$**CommunityAreaNumber,

censusdata**$**Hispanic,

censusdata**$**Black,

censusdata**$**White,

censusdata**$**Asian,

censusdata**$**Other,

censusdata**$**PercentChildrenInPoverty**)**

names**(**censusdata**)** **<-** c**(**'Community', 'communityAreaNumber', 'Hispanic', 'Black', 'White', 'Asian', 'Other', 'PercentChildrenInPoverty'**)**

write.csv**(**censusdata, 'censusdataByCommunityArea.csv'**)**

Combining all Datasets into One

##########################################

#Combining all the data

##########################################

avgSchoolRating **<-** read.csv**(**"4A/MSCI 446/R/explanatoryvariables/avg\_school\_rating\_by\_community.csv"**)**

avgSSLscore **<-** read.csv**(**"4A/MSCI 446/R/explanatoryvariables/avg\_ssl\_score\_by\_community.csv"**)**

censusData **<-** read.csv**(**"4A/MSCI 446/R/explanatoryvariables/censusdataByCommunityArea.csv"**)**

typesOfCrimes **<-** read.csv**(**"4A/MSCI 446/R/explanatoryvariables/crime\_count\_in\_community.csv"**)**

predictedVarDF **<-** read.csv**(**"4A/MSCI 446/R/explanatoryvariables/total\_crime\_by\_community.csv"**)**

totalParkArea **<-** read.csv**(**"4A/MSCI 446/R/explanatoryvariables/totalParkAreaByCommunityArea.csv"**)**

publicHealth **<-** read.csv**(**"4A/MSCI 446/R/explanatoryvariables/publicHealth.csv"**)**

#because totalParkArea dataframe is not sorted by ascending community area number:

totalParkArea **<-** totalParkArea**[**order**(**totalParkArea**$**communityAreaNumber**)**,**]**

predTable **<-** data.frame**(**totalParkArea**$**Community,

totalParkArea**$**communityAreaNumber,

predictedVarDF**$**violent\_crime **\*** 1000 **/** **(**predictedVarDF**$**population.2010.**)**,

avgSchoolRating**$**avg\_rating,

avgSSLscore**$**avg\_rating,

totalParkArea**$**totalParkArea,

publicHealth**$**numHospitals,

publicHealth**$**teenMomRate,

publicHealth**$**infantMortalityRate,

censusData**[**,4**:**ncol**(**censusData**)])**

names**(**predTable**)** **<-** c**(**

"community",

"communityAreaNum",

"percentViolentCrimePer1000Population",

"avgSchoolRating",

"avgSSLRating",

"totalParkArea",

"numHospitals",

"teenMomRate",

"infantMortalityRate",

"hispanic",

"black",

"white",

"asian",

"other",

"percentChildrenInPov"**)**

**)**

#write to csv

write.csv**(**predTable, 'predTable.csv'**)**

Explanatory Data Analysis Code

# gather useful columns

explanatory <- data.frame(predTable$communityAreaNum,

predTable$percentViolentCrimePer1000Population,

predTable$avgSchoolRating,

predTable$avgSSLRating,

predTable$totalParkArea,

predTable$numHospitals,

predTable$teenMomRate,

predTable$infantMortalityRate,

100\*predTable$hispanic,

100\*predTable$black,

100\*predTable$white,

100\*predTable$asian,

100\*predTable$other,

100\*predTable$percentChildrenInPov)

names(explanatory) <- c('community', 'number\_of\_violent\_crimes\_per\_1000\_population',

'Average\_School\_Rating', 'Normalized\_Average\_SSL', 'Total\_Park\_Area\_(m2)',

'Number\_of\_Hospitals', 'Number\_of\_Teen\_Moms\_/\_1000\_Female\_Teenagers', 'Number\_of\_Infant\_Mortality\_/\_1000\_Live\_Births',

'Percent\_of\_Hispanic\_(%)', 'Percent\_of\_Black\_(%)', 'Percent\_of\_White\_(%)',

'Percent\_of\_Asian\_(%)', 'Percent\_of\_Other\_Race\_(%)', 'Percent\_of\_Children\_in\_Poverty\_(%)')

# normalize SSL (266.0711 - 304.1068)

explanatory$Normalized\_Average\_SSL <- (explanatory$Normalized\_Average\_SSL-min(explanatory$Normalized\_Average\_SSL))/(max(explanatory$Normalized\_Average\_SSL) - min(explanatory$Normalized\_Average\_SSL))

# num\_hospital to binary

explanatory$Whether\_Community\_Has\_3\_or\_More\_Hospitals <- NA

explanatory$Whether\_Community\_Has\_3\_or\_More\_Hospitals <- explanatory$Number\_of\_Hospitals >= 3

col\_names <- colnames(explanatory)

# scatter plot

for(i in 3:14) {

plot(explanatory[,i], explanatory$number\_of\_violent\_crimes\_per\_1000\_population,

main=paste('Violent Crime Rate V.S.', gsub( '\\s\*\\([^\\)]+\\)', '', gsub('\_', ' ', col\_names[i]))),

xlab=gsub('\_', ' ', col\_names[i]), ylab='Number of Violent Crimes / 1000 Population', pch=18)

}

plot(explanatory[,15], explanatory$number\_of\_violent\_crimes\_per\_1000\_population,

main=paste('Violent Crime Rate V.S. ', gsub( '\\s\*\\([^\\)]+\\)', '', gsub('\_', ' ', col\_names[15]))),

xaxt='n', xlim=c(-1,2), xlab=gsub('\_', ' ', col\_names[15]), ylab='Number of Violent Crimes / 1000 Population', pch=18)

axis(1, at=0:1, labels=c('FALSE', 'TRUE'))

# histograms

hist(explanatory$number\_of\_violent\_crimes\_per\_1000\_population,

col='grey',

main='Number of Violent Crimes / 1000 Population (Histogram)',

xlab='Number of Violent Crimes / 1000 Population', ylab='Number of Communities')

for(i in 3:14) {

hist(explanatory[,i],

col='grey',

main=paste(gsub( '\\s\*\\([^\\)]+\\)', '', gsub('\_', ' ', col\_names[i])), '(Histogram)'),

xlab=gsub('\_', ' ', col\_names[i]), ylab='Number of Communities')

}

# box plots

boxplot(explanatory$number\_of\_violent\_crimes\_per\_1000\_population, data=explanatory,

col='grey',

main="Number of Violent Crimes / 1000 Population (Box Plot)",

xlab="Number of Violent Crimes / 1000 Population")

for(i in 3:14) {

boxplot(explanatory[,i], data=explanatory,

col='grey',

main=paste(gsub( '\\s\*\\([^\\)]+\\)', '', gsub('\_', ' ', col\_names[i])), '(Box Plot)'),

xlab=gsub('\_', ' ', col\_names[i]))

}

Numeric Regression Code

##########################################

#Numeric Regression

##########################################

library**(**caret**)**

library**(**dplyr**)**

#Import data: includes explanatory variables AND class variable but also other columns (e.g. community area name)

data **<-** read.csv**(**"4A/MSCI 446/R/explanatoryvariables/predTable.csv"**)**

#remove extraneous columns (e.g. community area name)

dataForPred **<-** dplyr**::**select**(**data, **-**X, **-**community, **-**communityAreaNum**)**

dataForPred **<-** dataForPred**[**,1**:**13**]**

dataForPred**$**numHospitals **<-** ifelse**(**dataForPred**$**numHospitals **>=** 3, 1, 0**)**

colnames**(**dataForPred**)[**5**]** **<-** "has3OrMoreHospitals"

#Use 10-fold cross-validation for getting alpha/lambda values for glmnet

tControlObj **<-** caret**::**trainControl**(**

method **=** "cv", number **=** 10,

verboseIter **=** **TRUE**,

summaryFunction **=** defaultSummary

**)**

k **<-** 77

#leave-1-out cross validation for performance metrics (RMSE, Rsquared, MAE)

splitPlan **<-** kWayCrossValidation**(**nrow**(**dataForPred**)**, k, **NULL**, **NULL)**

#initialization of the dataframe that will store the performance metrics

metricsDF **<-** as.data.frame**(**matrix**(**nrow **=** 3, ncol **=** 4**))**

names**(**metricsDF**)** **<-** c**(**"Model", "RMSE", "Rsquared", "MAE"**)**

OLS Linear Regression

##########################################

#train using linear regression#

modelLM **<-** train**(**

x **=** dataForPred**[**,2**:**13**]**,

y **=** dataForPred**[**,1**]**,

method **=** "lm",

trControl **=** tControlObj

**)**

#Predicted vs Actual Plot

plot**(**modelLM**$**finalModel**$**fitted.values, dataForPred**[**,1**]**, main**=**'Predicted vs Actual for Simple Linear Regression', xlab**=**'Predicted', ylab**=**'Actual'**)**

#Residual Plot

plot**(**modelLM**$**finalModel**$**fitted.values, modelLM**$**finalModel**$**residuals, main**=**'Residual Plot for Simple Linear Regression', xlab**=**'Predicted', ylab**=**'Residuals'**)**

abline**(**h **=** 0, col **=** "darkgrey", lty **=** 2**)**

#Residual Histogram Plot

hist**(**modelLM**$**finalModel**$**residuals,

col**=**'grey',

main**=**'Residual Histogram for Simple Linear Regression',

xlab**=**'Residual', ylab**=**'Frequency'**)**

#Leave-1-out Cross Validation to get OLS Performance Metrics

lmPredValues **<-** data.frame**(**"predicted" **=** rep**(**0, nrow**(**dataForPred**)))**

**for(**i **in** 1**:**k**)** **{**

split **<-** splitPlan**[[**i**]]**

model **<-** lm**(**percentViolentCrimePer1000Population **~** ., data **=** dataForPred**[**split**$**train,**])**

lmPredValues**$**predicted**[**split**$**app**]** **<-** predict**(**model, newdata **=** dataForPred**[**split**$**app,**])**

**}**

metricsDF**[**1,**]** **<-** c**(**"Linear Regression CV", postResample**(**lmPredValues**$**predicted, dataForPred**[**,1**]))**

metricsDF**[**4,**]** **<-** c**(**"Linear Regression", postResample**(**predict**(**modelLM, dataForPred**[**2**:**13**])**, dataForPred**[**,1**]))**

Elastic Net Regression

##########################################

#train using glmnet#

modelGLMNET **<-** train**(**

x **=** dataForPred**[**,2**:**13**]**,

y **=** dataForPred**[**,1**]**,

method **=** "glmnet",

metric **=** "RMSE",

tuneGrid **=** expand.grid**(**alpha **=** 0**:**10**/**10**)**, #lambda = seq(0.0001, 1, length = 20)

trControl **=** tControlObj

**)**

#obtain the predicted values

predictionGLMNET **<-** predict**(**modelGLMNET, dataForPred**[**, 2**:**13**])**

#plots RMSE over different alpha and lambda values.

plot**(**modelGLMNET, main**=**'Alpha and Lambda Values for GLMNET'**)**

#Predicted vs Actual Plot

plot**(**predictionGLMNET, dataForPred**[**,1**]**, main**=**'Predicted vs Actual for GLMNET Regression', xlab**=**'Predicted', ylab**=**'Actual'**)**

#Residual Plot

plot**(**predictionGLMNET, **(**dataForPred**[**,1**]-**predictionGLMNET**)**, main**=**'Residual Plot for GLMNET Regression', xlab**=**'Predicted', ylab**=**'Residuals'**)**

abline**(**h **=** 0, col **=** "darkgrey", lty **=** 2**)**

#Histogram Plot

hist**((**dataForPred**[**,1**]-**predictionGLMNET**)**,

col**=**'grey',

main**=**'Residual Histogram for GLMNET Linear Regression',

xlab**=**'Residual', ylab**=**'Frequency'**)**

#Leave-1-Out Cross Validation to get performance metrics

glmnetPredValues **<-** data.frame**(**"predicted" **=** rep**(**0, nrow**(**dataForPred**)))**

**for(**i **in** 1**:**k**)** **{**

split **<-** splitPlan**[[**i**]]**

model **<-** glmnet**(**as.matrix**(**dataForPred**[**split**$**train,2**:**13**])**, dataForPred**[**split**$**train,1**]**, alpha **=** modelGLMNET**$**bestTune**$**alpha, lambda **=** modelGLMNET**$**bestTune**$**lambda**)**

glmnetPredValues**$**predicted**[**split**$**app**]** **<-** predict**(**model, s **=** modelGLMNET**$**bestTune**$**lambda, newx **=** as.matrix**(**dataForPred**[**split**$**app,2**:**13**]))**

**}**

metricsDF**[**2,**]** **<-** c**(**"Elastic Net CV", postResample**(**glmnetPredValues**$**predicted, dataForPred**[**,1**]))**

metricsDF**[**5,**]** **<-** c**(**"Elastic Net", postResample**(**predictionGLMNET, dataForPred**[**,1**]))**

GAM Regression

##########################################

#GAM model#

#since caret can only do standard GAM model of y = s(x1) + s(x2) + etc. we will not be using caret

library**(**mgcv**)**

library**(**vtreat**)**

#GAM formula, based off scatter plots of each explanatory variable vs class variable from EDA

GAMformula **<-** percentViolentCrimePer1000Population **~**

avgSchoolRating **+**

avgSSLRating **+**

s**(**totalParkArea**)** **+**

has3OrMoreHospitals **+**

s**(**teenMomRate**)** **+**

s**(**infantMortalityRate**)** **+**

s**(**hispanic**)** **+**

black **+**

s**(**white**)** **+**

s**(**asian**)** **+**

other **+**

s**(**percentChildrenInPov**)**

#Leave-1-Out Cross Validation To get Performance Metrics

gamPredValues **<-** data.frame**(**"predicted" **=** rep**(**0, nrow**(**dataForPred**)))**

**for(**i **in** 1**:**k**)** **{**

split **<-** splitPlan**[[**i**]]**

model **<-** gam**(**GAMformula, data **=** dataForPred**[**split**$**train,**]**, family **=** gaussian**)**

gamPredValues**$**predicted**[**split**$**app**]** **<-** predict**(**model, newdata **=** dataForPred**[**split**$**app,**])**

**}**

metricsDF**[**3,**]** **<-** c**(**"GAM cv", postResample**(**gamPredValues**$**predicted, dataForPred**[**,1**]))**

#Building the final model

gamModel **<-** gam**(**GAMformula, data **=** dataForPred, family **=** gaussian**)**

finalPredictions **<-** predict**(**gamModel, dataForPred**[**, 2**:**13**])**

metricsDF**[**6,**]** **<-** c**(**"GAM", postResample**(**finalPredictions, dataForPred**[**,1**]))**

#Predicted vs Actual Plot

plot**(**finalPredictions, dataForPred**[**,1**]**, main**=**'Predicted vs Actual for GAM Regression', xlab**=**'Predicted', ylab**=**'Actual'**)**

#Residual Plot

plot**(**finalPredictions, **(**dataForPred**[**,1**]-**finalPredictions**)**, main**=**'Residual Plot for GAM Regression', xlab**=**'Predicted', ylab**=**'Residuals'**)**

abline**(**h **=** 0, col **=** "darkgrey", lty **=** 2**)**

#Residual Histogram

hist**((**dataForPred**[**,1**]-**finalPredictions**)**,

col**=**'grey',

main**=**'Residual Histogram for GAM Regression',

xlab**=**'Residual', ylab**=**'Frequency'**)**

postResample**(**predict**(**gamModel, dataForPred**[**, 2**:**13**])**, dataForPred**[**,1**])**

Clustering Code

# Clustering

# Normalize explanatory variables

normalized <- explanatory

for (col in 1:ncol(normalized)) {

normalized[,col] <- (normalized[,col]-min(normalized[,col]))/(max(normalized[,col]) - min(normalized[,col]))

}

# Clustering Euclidean Distance

euclidean <- matrix(, nrow = 25, ncol = 2)

for(round in 1:2) {

for(n in 1:25) {

cl <- kmeans(normalized[, 3:14], n)

euclidean[n, round] <- cl$tot.withinss

}

}

# plot the euclidean distance vs number of centers

plot(euclidean[,2], type='o', col=1, pch=18, lty=1,

main='Total Within-Cluster Sum of Squares v.s. Number of Centers',

xlab='Number of Centers', ylab='Total Within-Cluster Sum of Squares')

# plot trials with different sets of starting points

plot(euclidean[,1], type='o', col=4, pch=18, lty=2,

main='Results Generated by Different Randomly Chosen Start Point',

xlab='Number of Centers', ylab='Total Within-Cluster Sum of Squares')

lines(euclidean[,2], type="o", pch=18, lty=1, col=1)

legend(18, 50, c("First Trial","Second Trial"), cex=0.8,

col=c(1,4), pch=18:18, lty=1:2)

# perform k-means again with 50 sets of starting points and take average

for(n in 1:25) {

cl <- kmeans(normalized[, 3:14], n, nstart=50)

euclidean[n, round] <- cl$tot.withinss

}

plot(euclidean[,2], type='o', col=1, pch=18, lty=1,

main='Sum of Squares v.s. Number of Centers',

xlab='Number of Centers', ylab='Total Within-Cluster Sum of Squares')

# Clustering plots with 3 centres, with 50 sets of randomly chosen starting points

cl <- kmeans(normalized[, 3:14], 3, nstart=50)

plot(normalized[, 3:14], col = cl$cluster, pch=20)

points(cl$centers, col = cl$cluster, pch = 8, cex = 2)