STAT 512 Final Report

Spring 2019

Swati Banerjee, Priyanka Tiwari, Qizhen Ye, Jingya Wang

Purdue University

West Lafayette, Indiana

Introduction

Crime rates has been an interesting topic in all times. It is always scholar's and sociologist's interest to find out what factors can be used to explain crime rates. Community attributes are often considered related to the crime activities in the corresponding neighbourhood. And a more concrete question lies in which of these attributes can explain crime rates better.

In sociology, the social disorganization theory is a theory developed by the Chicago School, beginning with the social disorganization approach of Shaw and McKay (1969). They argued that socioeconomic status (SES), racial and ethnic heterogeneity, and residential stability account for variations in social disorganization and hence informal social control, which in turn account for the distribution of community crime. This theory directly links crime rates to neighborhood ecological characteristics; a core principle of social disorganization theory states that location matters. In other words, a person's residential location is a substantial factor shaping the likelihood that the person will become involved in illegal activities. The theory suggests that, among determinants of a person's later illegal activity, residential location is as significant as or more significant than the person's individual characteristics (e.g., age, gender, or race).

Sampson and W. Byron Groves (1989) extended the model of social disorganization by Shaw and McKay (1969). They performed empirical study on five dimensions related to community structure, which impacts social disorganization (refer Fig 1.)

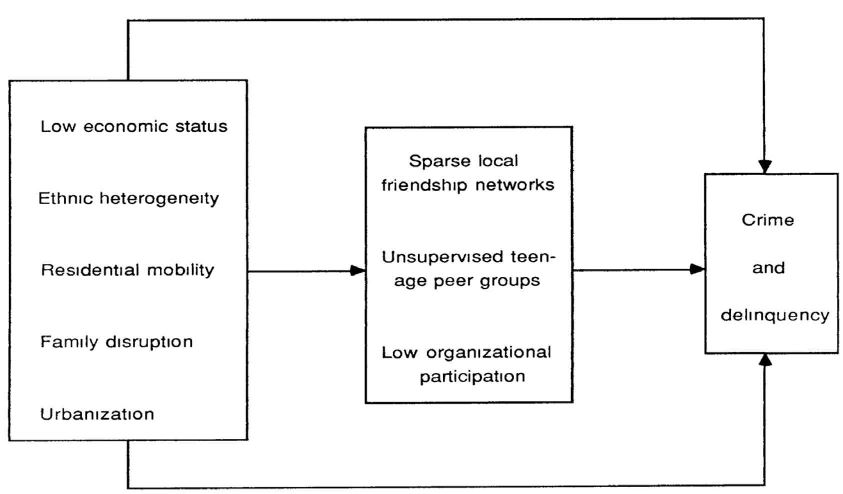


FIG. 1.—Causal model of extended version of Shaw and McKay's theory of community systemic structure and rates of crime and delinquency.

In this project, we utilize a dataset created by the University of California Irvine (UCI). The dataset consists of crime data of the US in 1990s and a group of socioeconomic attributes from that time, and we try to conclude which of the attributes are contributing the most to the five dimensions in the extended social disorganization model. A correlation model will be constructed to demonstrate the result.

Methods

Data Description

The dataset is a cross-sectional data acquired from UCI machine learning repository website. The title of the dataset is 'Crime and Communities'. It is prepared using real data from socio-economic data from 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. This dataset contains a total number of 147 attributes and 2216 instances. It is a large dataset and we had used two different selection methods to build a model with fewer predictor variable. The selection methods have been discussed later on. The response variable -violentPerPop, is the same for both the model.

Response Variable:

We have violentPerPop as our response variable. The per capita violent crimes variable i.e. total number of violent crimes per 100K population, is the aggregate of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault, using population values included in the 1995 FBI data (which differ from the 1990 Census values). Instead of a combined figure of violent and non - violent crime, we decided to restrict our response variables to violent crimes, because the nature of these two crime group are very different. The analysis of non-violent crime and it's relationship with community attributes needs a separate research.

Predictor Variables:

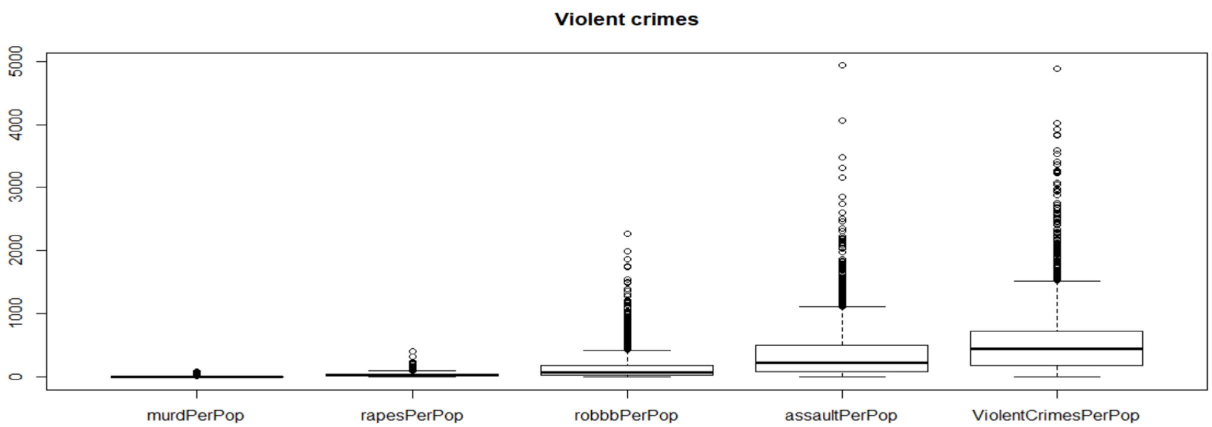
We have given the description of the respective predictor variables of the models in the model selection section.

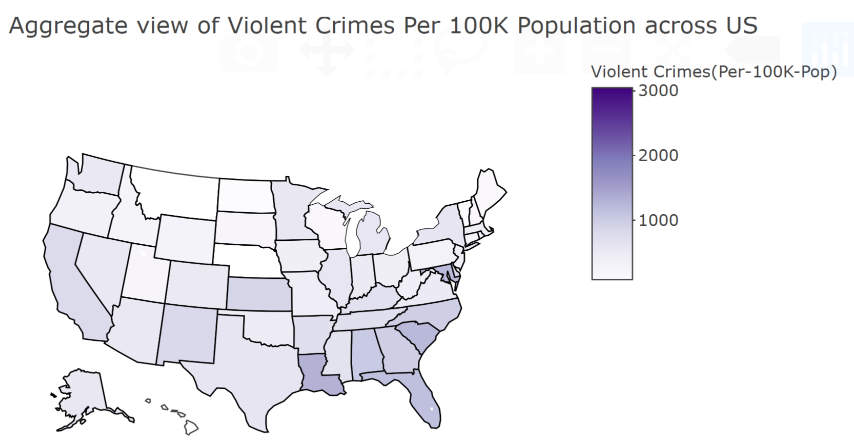
Research questions:

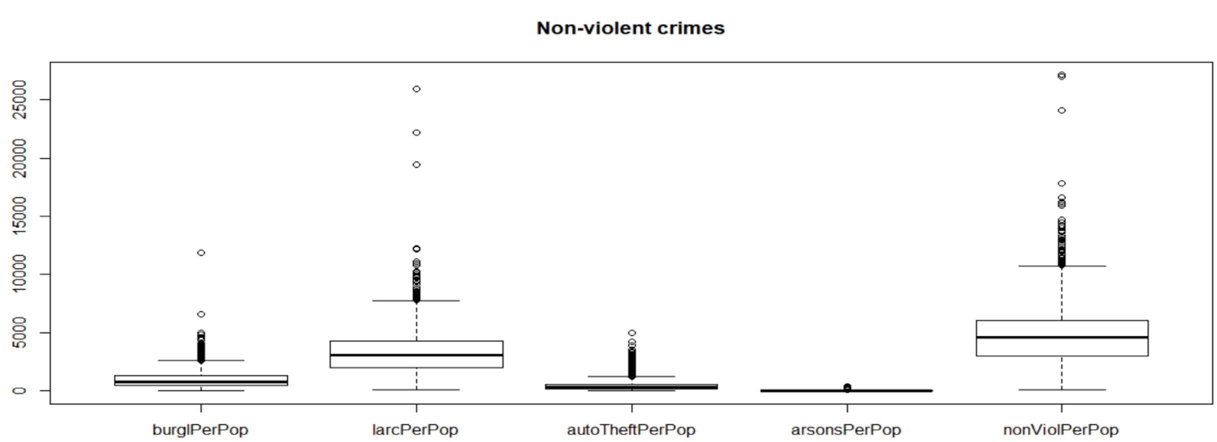
- Can the crime rate of a community be explained by certain attributes/variables of the community?
- To empirically test the explanatory power of extended model of Shaw and MaKay's theory in explaining the crime level in the community.

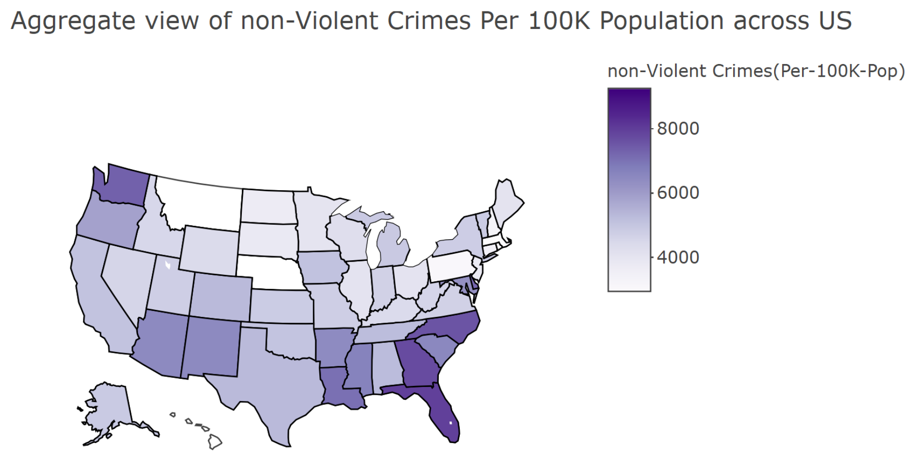
Exploratory Analysis

From the boxplot, it is clear that there are outliers in the data. We can see from the map that some states are much more crime prone than the rest –









Even though, we see the outliers in the data, as we can see some states are much more crime prone than the others in case of both violent and non-violent crimes. It was decided not to drop the outliers. The reason being, some communities in USA are actually much more crime prone areas than the others, and if we drop them, we'll not be able to look into the real problems of the socio-economic issues that leads to high crime.

From correlation matrix in Table III, Appendix B, it is clear that strong correlation exists between quite a few predictor variables. agepct12-21, agepct12-29, agepct16-24 are highly correlated (r>0.85). Since they represent the overlapping age group. There are other highly correlated variables such as pctKidsBornNevrMarr and pctBlack. Having these variables may lead to the problem of multicollinearity in the regression model. But, at this stage, instead of dropping any variable, we decide to proceed to the model selection process to choose variables with best features. So, meanwhile, we'll proceed without dropping any variable.

Model Selection

Predicting the level of Violent Crime in a USA community is a numerical prediction, and therefore requires a model that can provide a quantitative output. The measures of performance was MSE for initial model comparison, while both RMSE and R^2 were used for the comparison of the optimized models.
We came to an inference that we need a model which can minimize/avoid multicollinearity in the dataset and give us true relationships with the response variable. This led to the choose the following two models -

Model 1

Linear Regression : Subset Selection

In order to reduce the number of inputs, the Subset selection method, to choose the best set of features, was tested. The regsubsets function was used to find the combination of features which provided the best MSE result. However, using more than six features required a large amount of processing and was not feasible given the timeline. The function was run with the maximum number of variables set to six.

Output -

```
## [1] "Number of Coefficients in Best Model=        6"
##      (Intercept)       population      agePct12t29       pctWInvInc
##       0.74124622       0.08751896      -0.18608212      -0.07565421
##      PctKids2Par PctPersDenseHous   racepctblackBC
##      -0.27393525       0.16452897       0.16950682
## [1] "Subset Selection Linear Regression MSE=    0.00476"
```

The MSE result from the Feature Subset Selection for Linear Regression is 0.00476 which is worse than the result for Simple Linear Regression Model. A reduced subset of Features therefore does not improve the performance of the linear model. Increasing the variables to 8 or 10 would have given better result.

However, the features chosen as the best predictors are of interest. It can be seen that these include:

population	population for community
agepct12-29	percentage of population between the ages of 12 and 29
pctWdiv	percentage of households with investment / rent income in 1989
pctBlack	percentage of the population that is african american
pctKids2Par	percentage of kids in family housing with two parents
pctPopDenseHous	percent of persons in dense housing (more than 1 person per room)

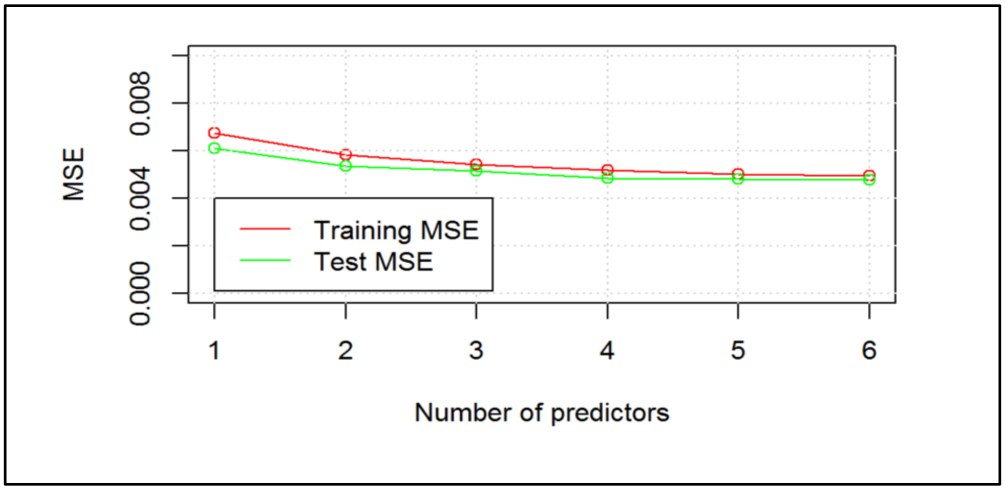


Fig 2 - Effect of the Number of Predictors on Subset Selection MSE

The plot of the MSE vs the number of predictors (Fig.2) shows that the error reduces with the number of features added to the dataset. It can also be seen that the test error is less than the training error which is not expected. This could be a result of the sample selection.

We Checked our model 1 for non-constant variance by conducting Breusch–Pagan test:

Null Hypothesis for : Variance is constant
Non-constant Variance Score Test -
Chisquare = 1655.388, Df = 1, p = < 2.22e-16

We clearly reject the null hypothesis because p value is negligible. Hence, Model 1 has non-constant variance, we can also see in the residual plot (Table IV, Appendix B) that the variance is increasing. As a diagnostics, we conducted Weighted Least Square regression of Model 1.

Model 2

Linear regression : VIF Selection

Models like Forward or backward selection of variables could produce inconsistent results, variance partitioning analyses may be unable to identify unique sources of variation, or parameter estimates may include substantial amounts of uncertainty. Collinearity, or excessive correlation among explanatory variables can complicate or prevent the identification of an optimal set of explanatory variables for a statistical model.

Analytical limitations related to collinearity requires to think carefully about the variables we choose to model, rather than adopting a naive approach where we blindly use all information to understand complexity.

A simple approach to identify collinearity among explanatory variables is the use of variance inflation factors (VIF).

VIF calculations are straightforward - the higher the value, the higher the collinearity. A VIF for a single explanatory variable is obtained using the r-squared value of the regression of that variable against all other explanatory variables.

The correlation matrix for the random variables can be calculated using a threshold [refer Table V, Appendix B].

Using VIF, we check on variables which have VIF higher than threshold should be dropped from the model [refer Table VII and Table VIII]. We can then simply run the linear regression model using the variables which we got after dropping.

This results in an R2 of about 0.822, which is pretty good.

The Model 2 is as follows -

violentPerPop~HousVacant + PctHousOccup + PctHousOwnOcc + PctVacantBoarded + PctVacMore6Mos + PctEmploy + murdPerPop + rapesPerPop + assaultPerPop +larcPerPop + autoTheftPerPop + arsonsPerPop

The in detailed description of the predictors of models 2 are in Table I, Appendix B.

We Checked our model 2 for non-constant variance by conducting Breusch–Pagan test:

Null Hypothesis for : Variance is constant
Non-constant Variance Score Test -
Chisquare = 8231.892, Df = 1, p = < 2.22e-16

We clearly reject the null hypothesis because p value is negligible. Hence, Model 1 has non-constant variance, we can also see in the residual plot (Table VI, Appendix B) that the variance is increasing. Like Model 1, we conducted Weighted Least Square regression of Model 2 as a diagnostics effort for the non constant variance.

Inference: Based on the WLS regression of the both models, we looked into the t-stats of the parameter estimates to verify if the selected parameters are helping us to understand crime in the community or the selected attributes are not significant. The F-stats of both the model are very high, and p value negligible, which means both the models are significant.

Results

Model 1 WLS

violentPerPop ~ population + agepct12-29 + pctWdiv + pctBlack + pctKids2Par + pctPopDenseHous

Regression Summary (Parameter Estimate Table) -

```
Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-3.7950 -0.9862 -0.3819  0.5960 15.9370 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.368e+03  7.564e+01  18.091  < 2e-16 ***
population       6.989e-04  1.080e-04   6.473 1.18e-10 ***
racepctblack     1.233e+01  1.135e+00  10.865  < 2e-16 ***
agePct12t29     -4.801e+00  5.970e-01  -8.041 1.44e-15 ***
PctKids2Par     -1.284e+01  8.417e-01 -15.259  < 2e-16 ***
PctPersDenseHous  2.541e+01  2.394e+00  10.611  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.431 on 2209 degrees of freedom
Multiple R-squared:   0.39,	Adjusted R-squared:  0.3886 
F-statistic: 282.5 on 5 and 2209 DF,  p-value: < 2.2e-16
```

95% CI of Parameters estimates

```
                        2.5 %        97.5 %
(Intercept)       1.220112e+03  1.516796e+03
population        4.871604e-04  9.105854e-04
racepctblack      1.010180e+01  1.455145e+01
agePct12t29      -5.971279e+00 -3.629761e+00
PctKids2Par      -1.449442e+01 -1.119318e+01
PctPersDenseHous  2.071122e+01  3.010234e+01
```

ANOVA Model 1

```
Response: ViolentCrimesPerPop
                   Df Sum Sq Mean Sq  F value   Pr(>F)    
population          1  501.8  501.85 244.9346 < 2e-16 ***
racepctblack        1 1323.6 1323.64 646.0203 < 2e-16 ***
agePct12t29         1    6.5    6.47   3.1583 0.07568 .  
PctKids2Par         1  831.3  831.32 405.7403 < 2e-16 ***
PctPersDenseHous    1  230.7  230.68 112.5893 < 2e-16 ***
Residuals        2209 4526.0    2.05                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 2WLS

violentPerPop~HousVacant + PctHousOccup + PctHousOwnOcc + PctVacantBoarded + PctVacMore6Mos + PctEmploy + murdPerPop + rapesPerPop + assaultPerPop +larcPerPop + autoTheftPerPop + arsonsPerPop

```
Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-5.0600 -0.8608 -0.2757  0.4067 13.2661 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -2.437e+02  6.229e+01  -3.912 9.43e-05 ***
HousVacant        2.622e-03  1.539e-03   1.703  0.08868 .  
PctHousOccup      1.517e+00  6.267e-01   2.420  0.01561 *  
PctHousOwnOcc     8.201e-01  1.731e-01   4.738 2.29e-06 ***
PctVacantBoarded -7.603e-01  1.004e+00  -0.758  0.44877    
PctVacMore6Mos   -4.106e-01  1.839e-01  -2.233  0.02566 *  
PctEmploy         7.479e-01  2.504e-01   2.987  0.00285 ** 
murdPerPop        3.623e+00  6.866e-01   5.277 1.44e-07 ***
rapesPerPop       3.603e+00  1.459e-01  24.692  < 2e-16 ***
assaultPerPop     9.966e-01  2.006e-02  49.670  < 2e-16 ***
larcPerPop        6.026e-03  2.024e-03   2.977  0.00294 ** 
autoTheftPerPop   1.295e-01  1.280e-02  10.117  < 2e-16 ***
arsonsPerPop     -9.706e-02  9.762e-02  -0.994  0.32021    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.623 on 2202 degrees of freedom
Multiple R-squared:  0.7945,	Adjusted R-squared:  0.7934 
F-statistic: 709.3 on 12 and 2202 DF,  p-value: < 2.2e-16
```

95% C.I of Parameters of Model 2

```
                         2.5 %        97.5 %
(Intercept)      -3.658439e+02 -1.215331e+02
HousVacant       -3.969448e-04  5.640558e-03
PctHousOccup      2.874549e-01  2.745604e+00
PctHousOwnOcc     4.807132e-01  1.159547e+00
PctVacantBoarded -2.728447e+00  1.207775e+00
PctVacMore6Mos   -7.711705e-01 -4.998545e-02
PctEmploy         2.568297e-01  1.238986e+00
murdPerPop        2.276631e+00  4.969672e+00
rapesPerPop       3.316501e+00  3.888733e+00
assaultPerPop     9.572363e-01  1.035929e+00
larcPerPop        2.056591e-03  9.995682e-03
autoTheftPerPop   1.044211e-01  1.546364e-01
arsonsPerPop     -2.884864e-01  9.437444e-02
```

ANOVA Model 2

```
Response: ViolentCrimesPerPop
                   Df Sum Sq Mean Sq   F value    Pr(>F)    
HousVacant          1 1876.1  1876.1  712.6544 < 2.2e-16 ***
PctHousOccup        1   92.2    92.2   35.0400 3.737e-09 ***
PctHousOwnOcc       1 1166.6  1166.6  443.1370 < 2.2e-16 ***
PctVacantBoarded    1  127.4   127.4   48.4012 4.564e-12 ***
PctVacMore6Mos      1 1346.6  1346.6  511.5085 < 2.2e-16 ***
PctEmploy           1   28.2    28.2   10.7157  0.001079 ** 
murdPerPop          1 3223.9  3223.9 1224.6064 < 2.2e-16 ***
rapesPerPop         1 6003.6  6003.6 2280.4929 < 2.2e-16 ***
assaultPerPop       1 8199.4  8199.4 3114.5772 < 2.2e-16 ***
larcPerPop          1   74.5    74.5   28.2841 1.154e-07 ***
autoTheftPerPop     1  267.2   267.2  101.4851 < 2.2e-16 ***
arsonsPerPop        1    2.6     2.6    0.9885  0.320208    
Residuals        2202 5797.0     2.6                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The dataset contains a large amount of information collected from each community which can be summarized in the broad categories of race, age, employment, marital status, immigration data and home ownership.

In respect to research questions, crime rates can certainly be explained by certain attributes of the community.

In Model 1, the explanatory variables included in the linear regression analysis explains only 35% variation in the response variable "ViolentCrimesPerPop". The linear regression model developed will be incomplete without considering the other 120+ explanatory variables available in the dataset.

In Model 2, the features which were most frequently were used to predict the level of Violent Crime were:

- percentage of kids in family housing with two parents
- percentage of kids born to parents who never married
- number of kids born to parents who never married
- percent of persons in dense housing (more than 1 person per room)
- percentage of population that is Caucasian
- percentage of the population that is African American

Discussion

Model 1 - From, the parameter estimates t-stats, we find that all the attributes are significant in explaining the violent crime rates in the community. We find that as the population is increasing, the violent crime is also predicted to increase, while parameter estimates of population is very small, we should keep in mind that we have not changed the unit of population, i.e. if the population increases by ten thousand in the community, the community is predicted to have seven more violent crimes per 100k population, which is significant given the nature of the crime. While, we see that as the percent of children with two parents decreases, violent crime also increases, which proves the model theory that family disruption leads to crime in the community. Similarly, we see that as the percent of persons in dense housing (more than 1 person per room) increases, the crime also increases, we may say that people with low economic status are may be the ones who having dense household population. Hence, the results of the regression model supports the model of Shaw and MaKay's theory in explaining the crime level in the community, and proves that the community socio-economic attributes have good explanatory power when it comes to the crime in the community.

Model 2 - In order to reduce the number of inputs, the Subset selection method to choose the best set of features was tested. The regsubsets function was used to find the combination of features which provided the best MSE result. However, the using more than 6 features required a large amount of processing and was not feasible. The function was run with the maximum number of variables set to 6.

During the time in 1990s, US was going through Crack epidemic, which mainly affected underprivileged African American communities. In addition, between 1990 to 1995, the number of 15-24 year olds increased by roughly 20%, and the share of the population between the age of 15 to 24 was increased from 13.7% to 14.6%. Their impact may explain the strength of the predictive performance of features relating to race and children.

There is some similarity between these features, which indicates actions to reduce correlated features should have been taken in the data preparation phase.

The large number of features required to obtain a good performance with the Linear Models indicates that it could have been worthwhile exploring polynomial or recursive feature selection if additional processing capability was availible.

In regards to models, we can train and test the model 1 with more than 6 features which could increase the prediction levels. Model 2 provides a good fit to the data with the features.

References

Dataset

- Communities and Crime Unnormalized Data Set
  http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized?fbclid=IwARWuEZ9HO_tpY33I2w1DQ6uoCLikq5D8qXxWmkH6Af9MCrAe_OJxI15fXM

Citation

- Community Structure and Crime: Testing Social-Disorganization Theory
  https://www.journals.uchicago.edu/doi/abs/10.1086/229068

Appendix A: Code

```
#Violent and non-Violent crimes by state - Aggregate view
#group Violent crime and nonViolent crime by state

crimedatafile <- read.csv("~/crimedatafile.csv", na.strings=c("NA", "-", "?"), header = T,
stringsAsFactors = FALSE)
install.packages("magrittr")
crimedata_state=aggregate(newdata[,c('ViolentCrimesPerPop','nonViolPerPop')],
by=list(crimedatafile$state), FUN=mean)
crimedata_state=aggregate(crimedatafile[,c('ViolentCrimesPerPop','nonViolPerPop')],
by=list(crimedatafile$state), FUN=mean)
library(magrittr)
library("plotly")
l <- list(color = toRGB("white"), width = 2)
g <- list(
+    scope = 'usa',
+    projection = list(type = 'albers usa'),
+    showlakes = TRUE,
+    lakecolor = toRGB('white')
+ )
g <- list(
scope = 'usa',
projection = list(type = 'albers usa'),
showlakes = TRUE,
lakecolor = toRGB('white')
)
plot_geo(crimedata_state, locationmode = 'USA-states') %>%
add_trace(
z = ~crimedata_state$nonViolPerPop,locations = ~crimedata_state$Group.1,
color = ~crimedata_state$nonViolPerPop, colors = 'Purples'
) %>%
colorbar(title = "non-Violent Crimes(Per-100K-Pop)") %>%
layout(
title = 'Aggregate view of non-Violent Crimes Per 100K  Population across US',
geo = g
)
```

```
plot_geo(crimedata_state, locationmode = 'USA-states') %>%
add_trace(
z = ~crimedata_state$ViolentCrimesPerPop,locations = ~crimedata_state$Group.1,
color = ~crimedata_state$ViolentCrimesPerPop, colors = 'Purples'
) %>%
colorbar(title = "Violent Crimes(Per-100K-Pop)") %>%
layout(
title = 'Aggregate view of Violent Crimes Per 100K  Population across US',
geo = g
)


#Boxplots - Exploratory Data Analysis of Response Variables
#Boxplot of non violent crime variables
Violent=
crimedatafile[,c('murdPerPop','rapesPerPop','robbbPerPop','assaultPerPop','ViolentCrimesPerP
op')]
nonViolent                                                                    =
crimedatafile[,c('burglPerPop','larcPerPop','autoTheftPerPop','arsonsPerPop','nonViolPerPop')]
nonViolentxLabels                                                             =
c('burglPerPop','larcPerPop','autoTheftPerPop','arsonsPerPop','nonViolPerPop')
violentxLabels                                                                =
c('murdPerPop','rapesPerPop','robbbPerPop','assaultPerPop','ViolentCrimesPerPop')
boxplot(Violent,main="Violent crimes",violentxLabels)
boxplot(nonViolent,main="Non-violent crimes",nonViolentxLabels)

###Linear Regression (Sub-Selection) [Model 1] ###

df.working <- crimedatafile
drops <- c("X", "fold","Êcommunityname","state")
df.working <- df.working[, !(names(df.working) %in% drops)]
set.seed(1)
train_ind = sample(1:nrow(df.working), 0.7 * nrow(df.working))
normalize <- function(x) {
 +    return((x - min(x))/(max(x) - min(x)))
 + }
df.working_dt <- df.working
notneededFeatures <- c("PctSpeakEnglOnlyCat", "PctNotSpeakEnglWellCat",
```

```
+ "PctHousOccupCat", "RentQrange")
possible_predictors = colnames(df.working)[!(colnames(df.working) %in%
+notneededFeatures)]
df.working = df.working[, names(df.working) %in% possible_predictors]
df.norm <- as.data.frame(lapply(df.working, normalize))
install.packages("leaps")
library(leaps)
regfit.full = regsubsets(ViolentCrimesPerPop ~ ., data = df.norm[train_ind,], really.big = T, nvmax
= 6)
training.mat = model.matrix(ViolentCrimesPerPop ~ ., data = df.norm[train_ind,])
training.errors = rep(NA, 6)
for (ii in 1:6) {
  coefi = coef(regfit.full, id = ii)
  pred = training.mat[, names(coefi)] %*% coefi
  training.errors[ii] = mse(df.norm[train_ind, 97], pred)
}
test.mat = model.matrix(ViolentCrimesPerPop ~ ., data = df.norm[-train_ind,])
test.errors = rep(NA, 6)
for (ii in 1:6) {
  coefi = coef(regfit.full, id = ii)
  pred = test.mat[, names(coefi)] %*% coefi
  test.errors[ii] = mse(df.norm[-train_ind, 97], pred)
}
k = which.min(test.errors)
MSE_SLM = test.errors[k]



###Linear Regression (VIF Selection) [Model 2] ###

#Correlaions
crimedata.fourth <- crimedatafile
cols =
c('HousVacant','PctHousOccup','PctHousOwnOcc','PctVacantBoarded','PctVacMore6Mos','PctU
nemployed','PctEmploy','murdPerPop','rapesPerPop','robbbPerPop','assaultPerPop','ViolentCri
mesPerPop','burglPerPop','larcPerPop','autoTheftPerPop','arsonsPerPop')
head(crimedata.fourth)
crimedata.fourth[,cols]
```

```
crimedata.study = crimedata.fourth[,cols]
library(dplyr)
correl <- round(cor(crimedata.study),2)

library(ggcorrplot)
ggcorrplot(correl)

cor_df <- as.data.frame(as.table(correl))
cor_df <- cor_df[cor_df$Freq != 1,]
cor_df %>%  arrange(desc(abs(Freq))) %>% filter(abs(Freq)>0.5)

#there exists multicollinearity between variables.  We will use VIF to remove multicollinearity

library(car)
library(plyr)

fit=lm(ViolentCrimesPerPop ~ . , data=crimedata.study)

vif(fit)

# Set a VIF threshold. All the variables having higher VIF than threshold are dropped from the model
threshold=2.5

# Sequentially drop the variable with the largest VIF until all variables have VIF less than threshold
drop=TRUE

aftervif=data.frame()
while(drop==TRUE) {
  vfit=vif(fit)
  aftervif=rbind.fill(aftervif,as.data.frame(t(vfit)))
  if(max(vfit)>threshold) { fit=
    update(fit,as.formula(paste(".","~",".","-",names(which.max(vfit))))) }
  else { drop=FALSE }}


# Model after removing correlated Variables
print(fit)
```

```
# How variables removed sequentially
t_aftervif= as.data.frame(t(aftervif))
edit(t_aftervif)

# Final (uncorrelated) variables with their VIFs
vfit_d= as.data.frame(vfit)

set.seed(1)
row.number <- sample(1:nrow(crimedata.study), 0.9*nrow(crimedata.study))
train = crimedata.study[row.number,]
test = crimedata.study[-row.number,]
dim(train)
dim(test)
New_Fit=lm(ViolentCrimesPerPop ~ HousVacant + PctHousOccup + PctHousOwnOcc +
PctVacantBoarded + PctVacMore6Mos + PctEmploy + murdPerPop + rapesPerPop +
assaultPerPop +larcPerPop + autoTheftPerPop + arsonsPerPop , data=train)
summary(New_Fit)
pred1 <- predict(New_Fit, newdata = test)
library(Metrics)
c(RMSE = rmse, R2=summary(New_Fit)$r.squared)
anova(New_Fit)
```

Appendix B: Output

Table I : Variable Description

Variable	Description
HousVacant	Number of vacant households (numeric - expected to be integer)
PctHousOccup	Percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
PctHousOwnOcc	Percent of households owner occupied (numeric - decimal)
PctHousOwnOcc	Percent of vacant housing that is boarded up (numeric - decimal)
PctVacantBoarded	Percent of vacant housing that has been vacant more than 6 months (numeric - decimal)
PctEmploy	Percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal)

Table II: Summary of few Important Community Attributes

```
  population        perHoush        pctBlack         pctWhite        pctAsian        pctHisp        agepct12-21
Min.   :  10005   Min.   :1.600   Min.   : 0.000   Min.   : 2.68   Min.   : 0.03   Min.   : 0.12   Min.   : 4.58
1st Qu.:  14366   1st Qu.:2.500   1st Qu.: 0.860   1st Qu.:76.32   1st Qu.: 0.62   1st Qu.: 0.93   1st Qu.:12.25
Median :  22792   Median :2.660   Median : 2.870   Median :90.35   Median : 1.23   Median : 2.18   Median :13.62
Mean   :  53118   Mean   :2.707   Mean   : 9.335   Mean   :83.98   Mean   : 2.67   Mean   : 7.95   Mean   :14.45
3rd Qu.:  43024   3rd Qu.:2.850   3rd Qu.:11.145   3rd Qu.:96.22   3rd Qu.: 2.67   3rd Qu.: 7.81   3rd Qu.:15.36
Max.   :7322564   Max.   :5.280   Max.   :96.670   Max.   :99.63   Max.   :57.46   Max.   :95.29   Max.   :54.40

 agepct12-29     agepct16-24      agepct65up       pctUrban          medIncome        medFamIncome      pctPoverty
Min.   : 9.38   Min.   : 4.64   Min.   : 1.66   Min.   :  0.00   Min.   :  8866   Min.   : 10447   Min.   : 0.64
1st Qu.:24.41   1st Qu.:11.32   1st Qu.: 8.75   1st Qu.:  0.00   1st Qu.: 23817   1st Qu.: 29538   1st Qu.: 4.51
Median :26.78   Median :12.54   Median :11.73   Median :100.00   Median : 31441   Median : 36678   Median : 9.33
Mean   :27.64   Mean   :13.98   Mean   :11.84   Mean   : 70.47   Mean   : 33985   Mean   : 39857   Mean   :11.62
3rd Qu.:29.20   3rd Qu.:14.35   3rd Qu.:14.41   3rd Qu.:100.00   3rd Qu.: 41481   3rd Qu.: 46999   3rd Qu.:16.91
Max.   :70.51   Max.   :63.62   Max.   :52.77   Max.   :100.00   Max.   :123625   Max.   :139008   Max.   :58.00

 pctNotHSgrad    pctUnemploy      pctAllDivorc     pctKidsBornNevrMarr pctLargHousFam   pctPopDenseHous
Min.   : 1.46   Min.   : 1.320   Min.   : 2.830   Min.   : 0.000      Min.   : 0.960   Min.   : 0.050
1st Qu.:13.92   1st Qu.: 4.045   1st Qu.: 8.575   1st Qu.: 1.070      1st Qu.: 3.390   1st Qu.: 1.290
Median :21.38   Median : 5.450   Median :10.900   Median : 2.040      Median : 4.280   Median : 2.340
Mean   :22.31   Mean   : 6.045   Mean   :10.813   Mean   : 3.115      Mean   : 5.387   Mean   : 4.132
3rd Qu.:29.20   3rd Qu.: 7.440   3rd Qu.:12.985   3rd Qu.: 3.910      3rd Qu.: 5.870   3rd Qu.: 4.730
Max.   :73.66   Max.   :31.230   Max.   :22.230   Max.   :27.350      Max.   :34.870   Max.   :59.490

persEmergShelt      persHomeless       pctSameCounty-5 pctSameState-5  violentPerPop    nonViolPerPop
Min.   :    0.00   Min.   :    0.00   Min.   :27.95   Min.   :32.83   Min.   :   0.0   Min.   :  116.8
1st Qu.:    0.00   1st Qu.:    0.00   1st Qu.:72.06   1st Qu.:85.20   1st Qu.: 161.7   1st Qu.: 2918.1
Median :    0.00   Median :    0.00   Median :79.49   Median :90.03   Median : 374.1   Median : 4425.4
Mean   :   66.95   Mean   :   17.82   Mean   :77.41   Mean   :88.11   Mean   : 589.1   Mean   : 4908.2
3rd Qu.:   22.00   3rd Qu.:    1.00   3rd Qu.:85.14   3rd Qu.:93.01   3rd Qu.: 794.4   3rd Qu.: 6229.3
Max.   :23383.00   Max.   :10447.00   Max.   :96.59   Max.   :99.90   Max.   :4877.1   Max.   :27119.8
                                                                      NA's   :221     NA's   :97
```

Table III: Multicollinearity

	pctPoverty	pctNotHSgrad	pctUnemploy	pctAllDivorc	pctKidsBornNevrMa	pctLargHousFam	pctPopDenseHous	persEmergShelt	persHomeless	pctSameCounty-5	pctSameState-5	violentPerPop	nonViolPerPop
population	0.09	0.05	0.08	0.11	0.19	0.10	0.11	0.93	0.92	0.02	-0.03	0.21	0.12
perHoush	0.08	0.14	0.17	-0.39	0.04	0.69	0.55	-0.03	-0.01	-0.12	-0.05	-0.02	-0.19
pctBlack	0.46	0.34	0.37	0.43	0.81	0.14	0.11	0.11	0.06	0.06	0.00	0.62	0.47
pctWhite	-0.53	-0.49	-0.51	-0.40	-0.80	-0.54	-0.60	-0.14	-0.10	-0.02	0.03	-0.68	-0.48
pctAsian	-0.14	-0.18	-0.12	-0.07	-0.05	0.22	0.30	0.06	0.07	-0.13	-0.14	0.04	-0.03
pctHisp	0.36	0.51	0.48	0.09	0.23	0.76	0.87	0.05	0.06	0.05	0.02	0.26	0.17
agepct12-21	0.49	0.07	0.23	-0.21	0.14	0.21	0.18	-0.01	-0.01	-0.43	-0.20	0.02	0.02
agepct12-29	0.47	0.08	0.20	-0.03	0.25	0.22	0.26	0.03	0.01	-0.55	-0.33	0.11	0.11
agepct16-24	0.46	0.02	0.16	-0.14	0.17	0.09	0.13	0.01	0.00	-0.52	-0.29	0.05	0.07
agepct65up	0.07	0.23	0.10	0.10	-0.01	-0.29	-0.24	-0.01	-0.01	0.32	0.21	0.05	0.13
pctUrban	-0.33	-0.25	-0.24	-0.05	0.01	0.06	0.05	0.07	0.05	0.08	-0.04	0.07	0.00
medIncome	-0.76	-0.66	-0.62	-0.56	-0.45	-0.15	-0.23	-0.04	-0.02	0.01	-0.03	-0.40	-0.47
medFamIncome	-0.73	-0.70	-0.65	-0.56	-0.46	-0.23	-0.30	-0.04	-0.03	-0.02	-0.05	-0.41	-0.46
pctPoverty	1.00	0.66	0.77	0.42	0.61	0.35	0.42	0.08	0.05	-0.07	0.01	0.50	0.51
pctNotHSgrad	0.66	1.00	0.74	0.39	0.55	0.49	0.57	0.05	0.03	0.34	0.30	0.47	0.37
pctUnemploy	0.77	0.74	1.00	0.37	0.56	0.49	0.51	0.07	0.05	0.18	0.17	0.47	0.39
pctAllDivorc	0.42	0.39	0.37	1.00	0.51	0.03	0.17	0.09	0.05	0.02	-0.03	0.54	0.61
pctKidsBornNevrMa	0.61	0.55	0.56	0.51	1.00	0.37	0.40	0.18	0.12	0.08	0.01	0.74	0.55
pctLargHousFam	0.35	0.49	0.49	0.03	0.37	1.00	0.88	0.07	0.07	0.13	0.08	0.34	0.16
pctPopDenseHous	0.42	0.57	0.51	0.17	0.40	0.88	1.00	0.08	0.08	0.03	-0.03	0.40	0.24
persEmergShelt	0.08	0.05	0.07	0.09	0.18	0.07	0.08	1.00	0.95	0.02	-0.02	0.19	0.10
persHomeless	0.05	0.03	0.05	0.05	0.12	0.07	0.08	0.95	1.00	0.01	-0.01	0.14	0.06
pctSameCounty-5	-0.07	0.34	0.18	0.02	0.08	0.13	0.03	0.02	0.01	1.00	0.74	0.07	-0.02
pctSameState-5	0.01	0.30	0.17	-0.03	0.01	0.08	-0.03	-0.02	-0.01	0.74	1.00	-0.01	-0.08
violentPerPop	0.50	0.47	0.47	0.54	0.74	0.34	0.40	0.19	0.14	0.07	-0.01	1.00	0.68
nonViolPerPop	0.51	0.37	0.39	0.61	0.55	0.16	0.24	0.10	0.06	-0.02	-0.08	0.68	1.00

	pctPoverty	pctNotHSgrad	pctUnemploy	pctAllDivorc	pctKidsBornNevrMa	pctLargHousFam	pctPopDenseHous	persEmergShelt	persHomeless	pctSameCounty-5	pctSameState-5	violentPerPop	nonViolPerPop
population	0.09	0.05	0.08	0.11	0.19	0.10	0.11	0.93	0.92	0.02	-0.03	0.21	0.12
perHoush	0.08	0.14	0.17	-0.39	0.04	0.69	0.55	-0.03	-0.01	-0.12	-0.05	-0.02	-0.19
pctBlack	0.46	0.34	0.37	0.43	0.81	0.14	0.11	0.11	0.06	0.06	0.00	0.62	0.47
pctWhite	-0.53	-0.49	-0.51	-0.40	-0.80	-0.54	-0.60	-0.14	-0.10	-0.02	0.03	-0.68	-0.48
pctAsian	-0.14	-0.18	-0.12	-0.07	-0.05	0.22	0.30	0.06	0.07	-0.13	-0.14	0.04	-0.03
pctHisp	0.36	0.51	0.48	0.09	0.23	0.76	0.87	0.05	0.06	0.05	0.02	0.26	0.17
agepct12-21	0.49	0.07	0.23	-0.21	0.14	0.21	0.18	-0.01	-0.01	-0.43	-0.20	0.02	0.02
agepct12-29	0.47	0.08	0.20	-0.03	0.25	0.22	0.26	0.03	0.01	-0.55	-0.33	0.11	0.11
agepct16-24	0.46	0.02	0.16	-0.14	0.17	0.09	0.13	0.01	0.00	-0.52	-0.29	0.05	0.07
agepct65up	0.07	0.23	0.10	0.10	-0.01	-0.29	-0.24	-0.01	-0.01	0.32	0.21	0.05	0.13
pctUrban	-0.33	-0.25	-0.24	-0.05	0.01	0.06	0.05	0.07	0.05	0.08	-0.04	0.07	0.00
medIncome	-0.76	-0.66	-0.62	-0.56	-0.45	-0.15	-0.23	-0.04	-0.02	0.01	-0.03	-0.40	-0.47
medFamIncome	-0.73	-0.70	-0.65	-0.56	-0.46	-0.23	-0.30	-0.04	-0.03	-0.02	-0.05	-0.41	-0.46
pctPoverty	1.00	0.66	0.77	0.42	0.61	0.35	0.42	0.08	0.05	-0.07	0.01	0.50	0.51
pctNotHSgrad	0.66	1.00	0.74	0.39	0.55	0.49	0.57	0.05	0.03	0.34	0.30	0.47	0.37
pctUnemploy	0.77	0.74	1.00	0.37	0.56	0.49	0.51	0.07	0.05	0.18	0.17	0.47	0.39
pctAllDivorc	0.42	0.39	0.37	1.00	0.51	0.03	0.17	0.09	0.05	0.02	-0.03	0.54	0.61
pctKidsBornNevrMa	0.61	0.55	0.56	0.51	1.00	0.37	0.40	0.18	0.12	0.08	0.01	0.74	0.55
pctLargHousFam	0.35	0.49	0.49	0.03	0.37	1.00	0.88	0.07	0.07	0.13	0.08	0.34	0.16
pctPopDenseHous	0.42	0.57	0.51	0.17	0.40	0.88	1.00	0.08	0.08	0.03	-0.03	0.40	0.24
persEmergShelt	0.08	0.05	0.07	0.09	0.18	0.07	0.08	1.00	0.95	0.02	-0.02	0.19	0.10
persHomeless	0.05	0.03	0.05	0.05	0.12	0.07	0.08	0.95	1.00	0.01	-0.01	0.14	0.06
pctSameCounty-5	-0.07	0.34	0.18	0.02	0.08	0.13	0.03	0.02	0.01	1.00	0.74	0.07	-0.02
pctSameState-5	0.01	0.30	0.17	-0.03	0.01	0.08	-0.03	-0.02	-0.01	0.74	1.00	-0.01	-0.08
violentPerPop	0.50	0.47	0.47	0.54	0.74	0.34	0.40	0.19	0.14	0.07	-0.01	1.00	0.68
nonViolPerPop	0.51	0.37	0.39	0.61	0.55	0.16	0.24	0.10	0.06	-0.02	-0.08	0.68	1.00

Table IV: Residual Plot Model 1

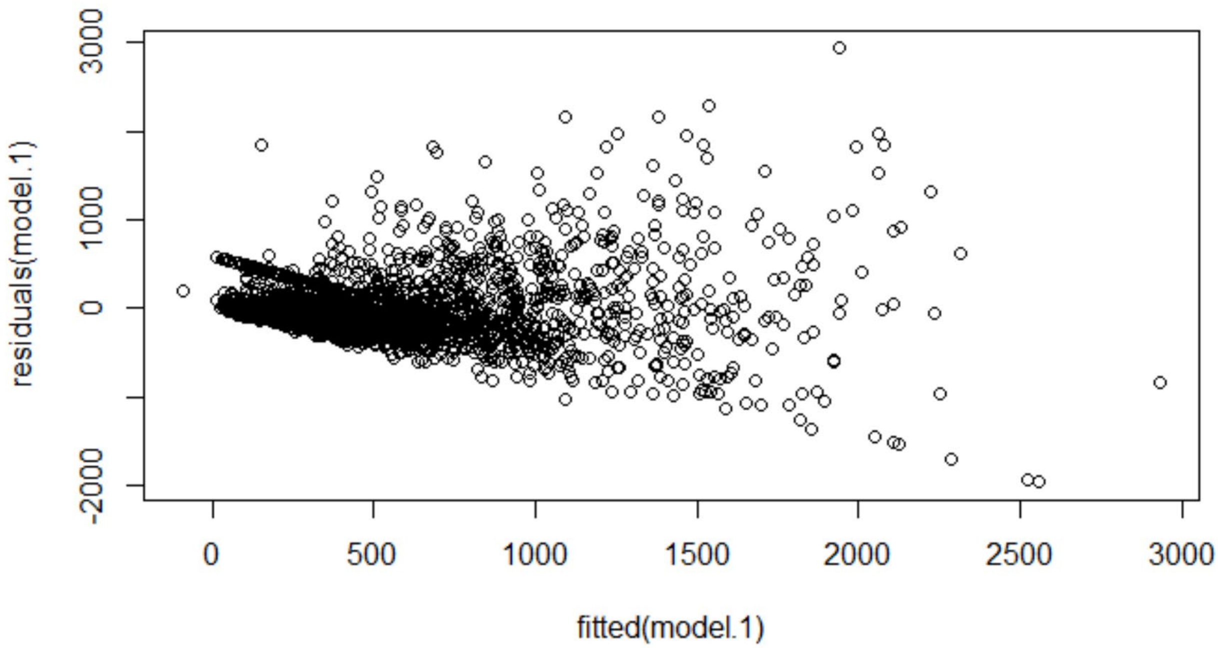


Table V: Correlation Matrix of Model 2

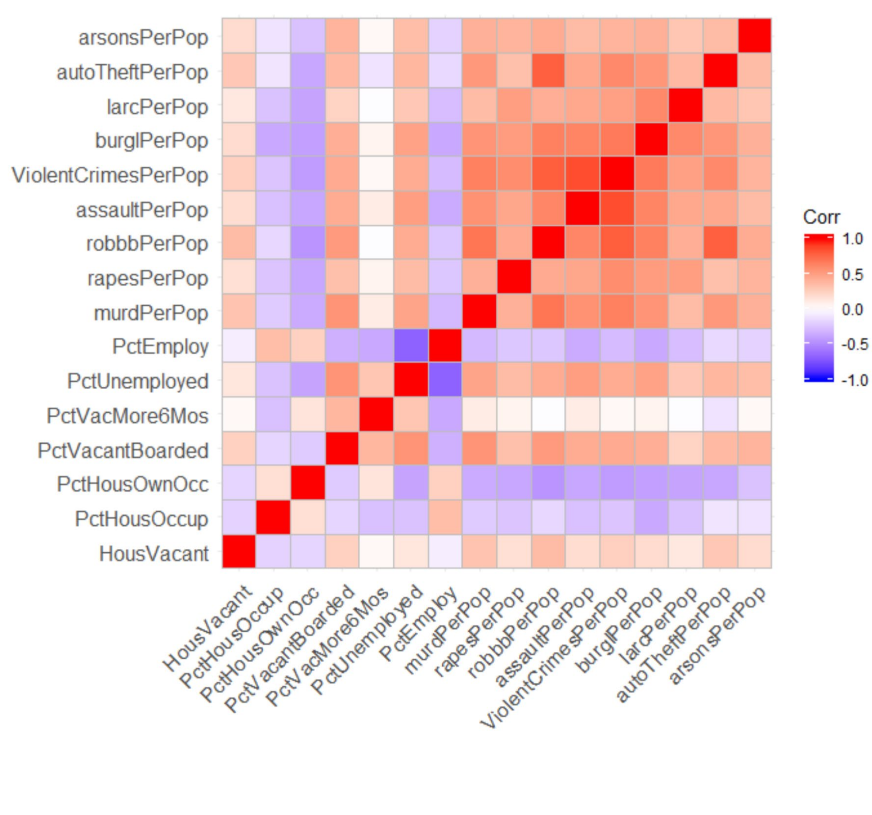


Table VI: Residual Plot Model 2

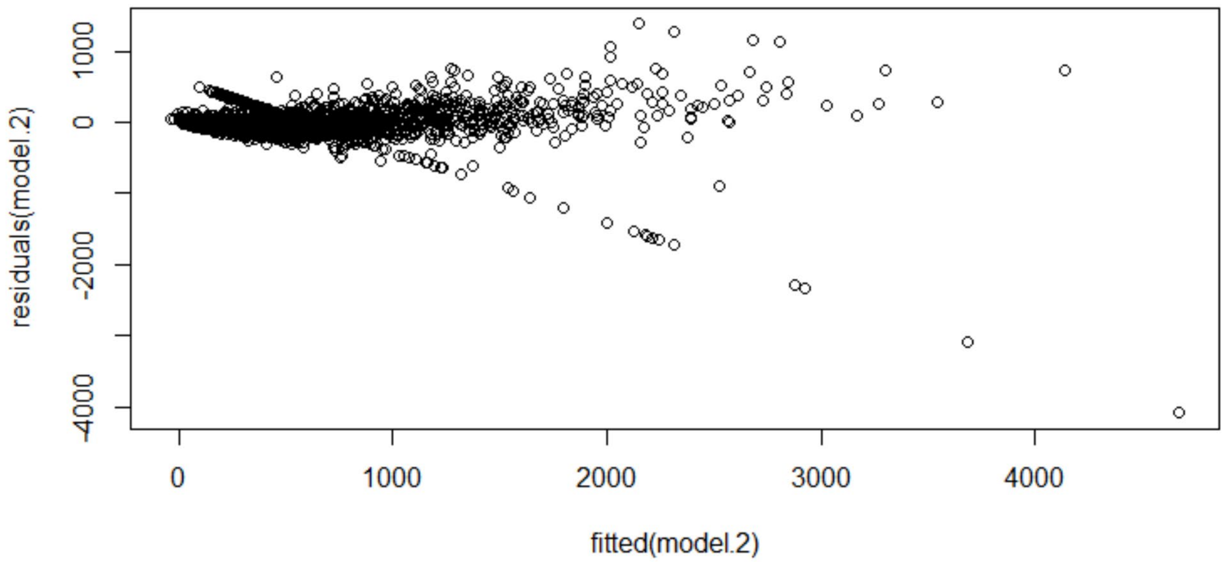


Table VII: Multicollinearity VIF Selection

```
vif(fit)

HousVacant      PctHousOccup      PctHousOwnOcc PctVacantBoarded     PctVacMore6Mos     PctUnemployed
1.213614         1.362877          1.562571         2.097157          1.557524          2.805008
PctEmploy         murdPerPop        rapesPerPop      robbbPerPop      assaultPerPop       burglPerPop
2.120294         2.295644          1.626940         4.546894          2.098244          2.762653
larcPerPop    autoTheftPerPop      arsonsPerPop
1.795462         2.785349          1.398063
```

Type VIII: Model 2 after removing correlated variables

```
Call:
  lm(formula = ViolentCrimesPerPop ~ HousVacant + PctHousOccup +
       PctHousOwnOcc + PctVacantBoarded + PctVacMore6Mos + PctEmploy +
       murdPerPop + rapesPerPop + assaultPerPop + larcPerPop + autoTheftPerPop +
       arsonsPerPop, data = crimedata.study)

Coefficients:
  (Intercept)         HousVacant       PctHousOccup      PctHousOwnOcc  PctVacantBoarded     PctVacMore6Mos
-1.459e+02          1.302e-03        -7.570e-02         -3.732e-01        -2.766e+00         -3.355e-01
PctEmploy          murdPerPop        rapesPerPop       assaultPerPop        larcPerPop    autoTheftPerPop
2.876e+00          7.758e+00         3.317e+00         8.371e-01          7.895e-03          2.239e-01
arsonsPerPop
-3.281e-01
```