**INTRODUCTION**

Social security has been a serious issue in United State for decades and the crime rate is relatively high among developed countries. According to a report about *Crime Index for Country 2018,* the crime index of United States is 49.58 and the safety index is 50.42, which is ranked 35th for its crime index and unfortunately wins the first among developed countries. The dataset that is related to the total number of violent crimes per 100,000 population contains information from 319 communities located in 55 different American states and consists of a series of factors associated with crime statistics. Generally, these factors be divided into six categories including states, portion of ethnicity minorities, population age, socioeconomic status, immigration numbers and police enforcement.

The 7 factors among 28 measurements are found to have an important association with the violent crime rate based on a regression analysis conducted on this dataset. These factors include percentage of population of that is African American, percentage of population of that is 12 -21 in age, percentage of people who are 16 and over and unemployed, percentage of kids in family housing with two parents, percentage of people foreign born, total requests for police and total requests for police per police officer. Utilizing the benefits of the predicting model will help to measure how much the violent crime rate is likely to be among 100,000 population with available figures related to these factors in a specific area in America.

**DATA ANALYSIS**

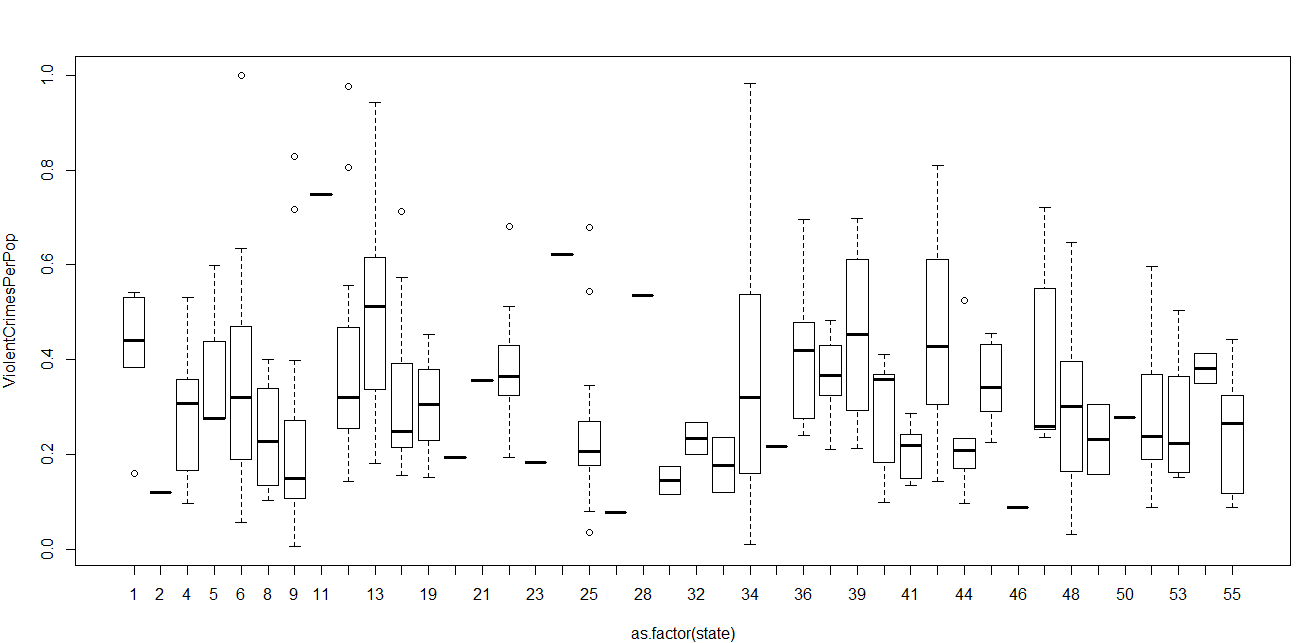
Both of covariates state and communityname are nominal. However, the communityname is quite unique among all covariate, and we would get null results if trying to conduct any univariate analysis linked with it since it has no degrees of freedom or anything for comparison purpose. Besides, I did the ANOVA to test the null hypothesis that whether there is any difference among different states.

*Covariate: state Variable: Nominal Test: ANOVA*

Df Sum Sq Mean Sq F value Pr(>F)

as.factor(state) 41 1.975 0.04818 1.428 0.0517

Residuals 277 9.344 0.03373

Consequently, we fail to reject the null hypothesis and the conclusion would be that there are no significant differences of crime rate in different states at level of significance α = 0.05. Just out of some curiosity, since the total numbers of violent are based on large figure - per 100,000 population, we may check the little differences that exit among the states. 

As what shows in the boxplot above, we may see that differences on violent crime rates for each state did exit. And those outliers may affect my final model of predicting violent crime rate. Proceed from the reality, violent crime rates would be more likely to associated with other detailed factors within communities or states. Therefore, the covariates - communityname and states would be excluded from the regression analysis for the final model.

The result of univariate analysis of all the other 26 covariates is presented on the Table 1 and it includes the mean, standard deviation and median for each covariate. Set the level of significance α = 0.01 since majority of the p-values are less than 0.01, which means they are highly correlated with the violent crime rates. For this standard, the candidate covariates for my final model would be racepctblack, racePctHisp, agePct12t21, medIncome, pctWPubAsst, PctPopUnderPov, PctUnemployed , TotalPctDiv, PctKids2Par, PctImmigRec5, PctImmigRec10, PctNotSpeakEnglWell, NumInShelters, PctUsePubTrans, LemasSwornFT, LemasTotalReq, PolicReqPerOffic, OfficAssgnDrugUnits.

For the further selection, I conducted the full model that includes all the covariates. The R-squared value is 0.7153, Adjusted R-squared value is 0.69 and p-value for the full model is less than 2.2e-16. Covariates including racepctblack, agePct12t21, PctKids2Par, NumInShelters, PctForeignBorn, LemasTotalReq, PolicReqPerOffic stand out at significance level of α = 0.10. However, the multicollinearity is definitely an issue after testing the full model with vif() function. The VIF (Variation Inflation Factors) scores of PctImmigRec5, PctImmigRec10, PctRecImmig5, PctRecImmig10, PctNotSpeakEnglWell, PctForeignBorn, LemasSwornFT are greater than 10. There is no doubt that covariates that measure the population who immigrated within last 5 years or 10 years, population who cannot speak English well and population who foreign born would highly correlated with each other. And figures to measure police enforcement are highly correlated with each other as other as well.

Table 1 – Summary of Univariate Analysis of Violent Crime Rate VS Covariates

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Covariate | Mean | Standard Deviation | | Median | p-value |
| Ethnicity Minorities | | | | | |
| racepctblack | 0.3371 |  | 0.3114 | 0.22 | < 2.2e-16 |
| racePctHisp | 0.2192 |  | 0.2611 | 0.10 | 3.99E-06 |
|  |  |  |  |  |  |
| Population Age | | | | | |
| agePct12t21 | 0.4205 |  | 0.1226 | 0.41 | 1.31E-08 |
|  |  |  |  |  |  |
| Socioeconomic Status | | | | | |
| pctUrban | 0.9921 |  | 0.0795 | 1.00 | 0.4756 |
| medIncome | 0.3051 |  | 0.1547 | 0.26 | < 2.2e-16 |
| pctWPubAsst | 0.4118 |  | 0.2222 | 0.39 | < 2.2e-16 |
| PctPopUnderPov | 0.3952 |  | 0.1991 | 0.40 | < 2.2e-16 |
| PctUnemployed | 0.4451 |  | 0.1907 | 0.41 | < 2.2e-16 |
| TotalPctDiv | 0.6039 |  | 0.1521 | 0.61 | < 2.2e-16 |
| PctKids2Par | 0.4661 |  | 0.2053 | 0.48 | < 2.2e-16 |
| NumInShelters | 0.1400 |  | 0.219 | 0.05 | 7.40E-13 |
| PctUsePubTrans | 0.3012 |  | 0.2955 | 0.18 | 5.26E-08 |
|  |  |  |  |  |  |
| Immigration | | | | | |
| PctImmigRec5 | 0.4411 |  | 0.1548 | 0.44 | 0.007397 |
| PctImmigRec10 | 0.5168 |  | 0.1517 | 0.53 | 0.0005642 |
| PctRecImmig5 | 0.3003 |  | 0.2845 | 0.18 | 0.051 |
| PctRecImmig10 | 0.3023 |  | 0.2876 | 0.18 | 0.02165 |
| PctNotSpeakEnglWell | 0.2460 |  | 0.2697 | 0.13 | 8.35E-07 |
| PctForeignBorn | 0.3180 |  | 0.2765 | 0.21 | 0.1413 |
|  |  |  |  |  |  |
| Police Enforcement | | | | | |
| LemasSwornFT | 0.0697 |  | 0.1382 | 0.02 | 3.76E-11 |
| LemasTotalReq | 0.0980 |  | 0.1619 | 0.04 | 3.70E-13 |
| PolicReqPerOffic | 0.3436 |  | 0.1972 | 0.29 | 3.26E-04 |
| OfficAssgnDrugUnits | 0.0755 |  | 0.1202 | 0.04 | 2.45E-12 |
| NumKindsDrugsSeiz | 0.5561 |  | 0.2032 | 0.57 | 0.2216 |
| PolicAveOTWorked | 0.3060 |  | 0.2269 | 0.26 | 0.1734 |
| LemasGangUnitDeploy | 0.4404 |  | 0.4058 | 0.50 | 0.6511 |
| LemasPctOfficDrugUn | 0.5879 |  | 0.2660 | 0.56 | 0.1108 |

Based on results from the full model, univariate analysis and significance level α = 0.10, I would include listed covariates - racepctblack, agePct12t21, PctKids2Par, NumInShelters, PctForeignBorn, LemasTotalReq, PolicReqPerOffic, and form a possibly selected model 1(Table 2). The R-squared value and Adjusted R-squared value for model 1 are 0.6838 and 0.6766, and both of them dropped a little bit when comparing the full model. But the model 1 has no issue of multicollinearity, which is a better possible choice.

The model 2 takes advantage of the stepwise method to selected significant covariates from the full model. Model 2 includes racepctblack, racepctHisp, agePct12t21, pctWPubAsst, PctUnemployed, PctKids2Par, PctRecImmig10, NumInShelters, PctForeignBorn, LemasSwornFT, LemasTotalReq, PolicReqPerOffic, OfficAssgnDrugUnits. However, it has issue of collinearity even though it has R-squared values of 0.7123, which is better than the model 1 and the full model. The VIF scores for PctRecImmig10, PctForeignBorn, LemasSwornFT are greater than 10.

To improve the model 2 and remove its issue of collinearity, I made the model 3. Combining the information from the model 1 and the full model and setting significance level α =0.05, I removed the PctRecImmig10 and keep the PctForeignBorn in the category of Immigration since PctForeignBorn has the p-value of 0.013964, indicating that it is more significant than PctRecImmig10 (0.077605) in the model 2. In the category of Police Enforcement, I removed the OfficAssgnDrugUnits first (p-value = 0.123956 in the model2) while it is not significant (p-value = 0.190035) in the full model as well. In addition, the p-values of the OfficeAssgnDrugUnits are larger than those of LemasSwornFT in both full model and model 2. Hence, I built model 3, then checked the VIF scores again. However, VIF scores for LemasSwornFT and LemasTotalReq are 7.191171 and 7.560016 which are not quite ideal in this category. LemasSwornFT (p-value = 0.171465 in the full model and 0.11793 in the model2) is less significant than other two covariates – LemasTotalReq and PolicReqPerOffic in both full model and model 2. In model 4 excluded the LemasSwornFT, and consequently came with lower VIF scores.

To establish a more accurate model with less predictors, I would eliminate some factors from the model 4 by the ranking of p-values. Firstly, I set the significance level of α = 0.10, removed the pctwPubAsst and got the model 5. Just very subtle changes happened to R-square and Adjusted R-square values, indicating that pctwPubAsst is not very significant in the model 4.



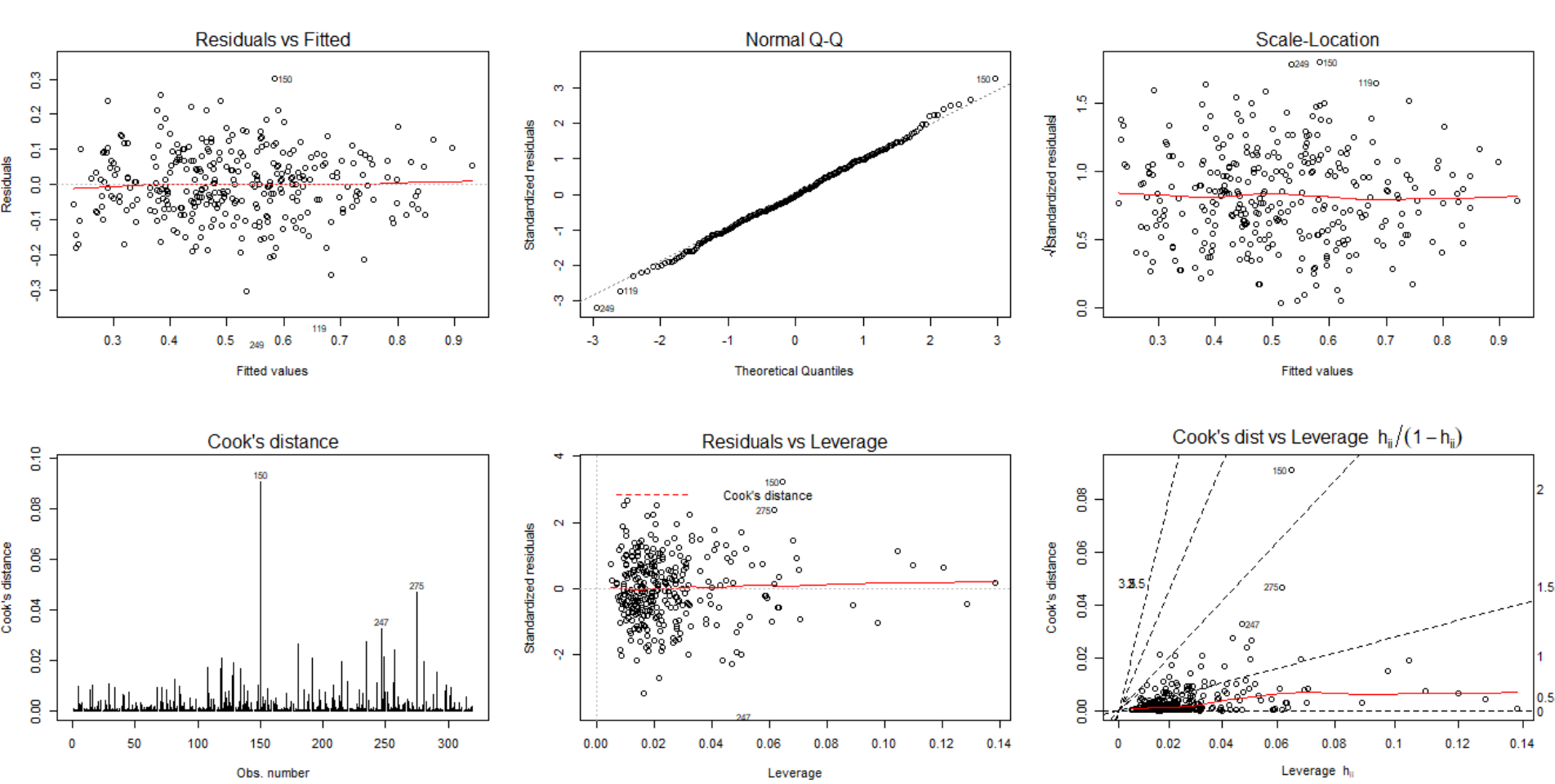
Secondly, the racePctHisp is eliminated from the model 5 to construct the model 6 for the significance level of 0.10. Thirdly, model 7 is based on model 6 by deleting the NumInShelters for the significance level of 0.05. As information from the table 2, from model 4 to model 7, changes on the performance are quite little. And predictors in model 7 are all significant since all the p-values are all less than 0.05.



The last step to get the final model based on model 7 is to check the demand of the power transformation and find the best lambda to transform ViolentCrimesPerPop(y). I tested the model with the transformation and decided to keep the transformation since the AIC value of the final model 8 is smaller than that of model 7.

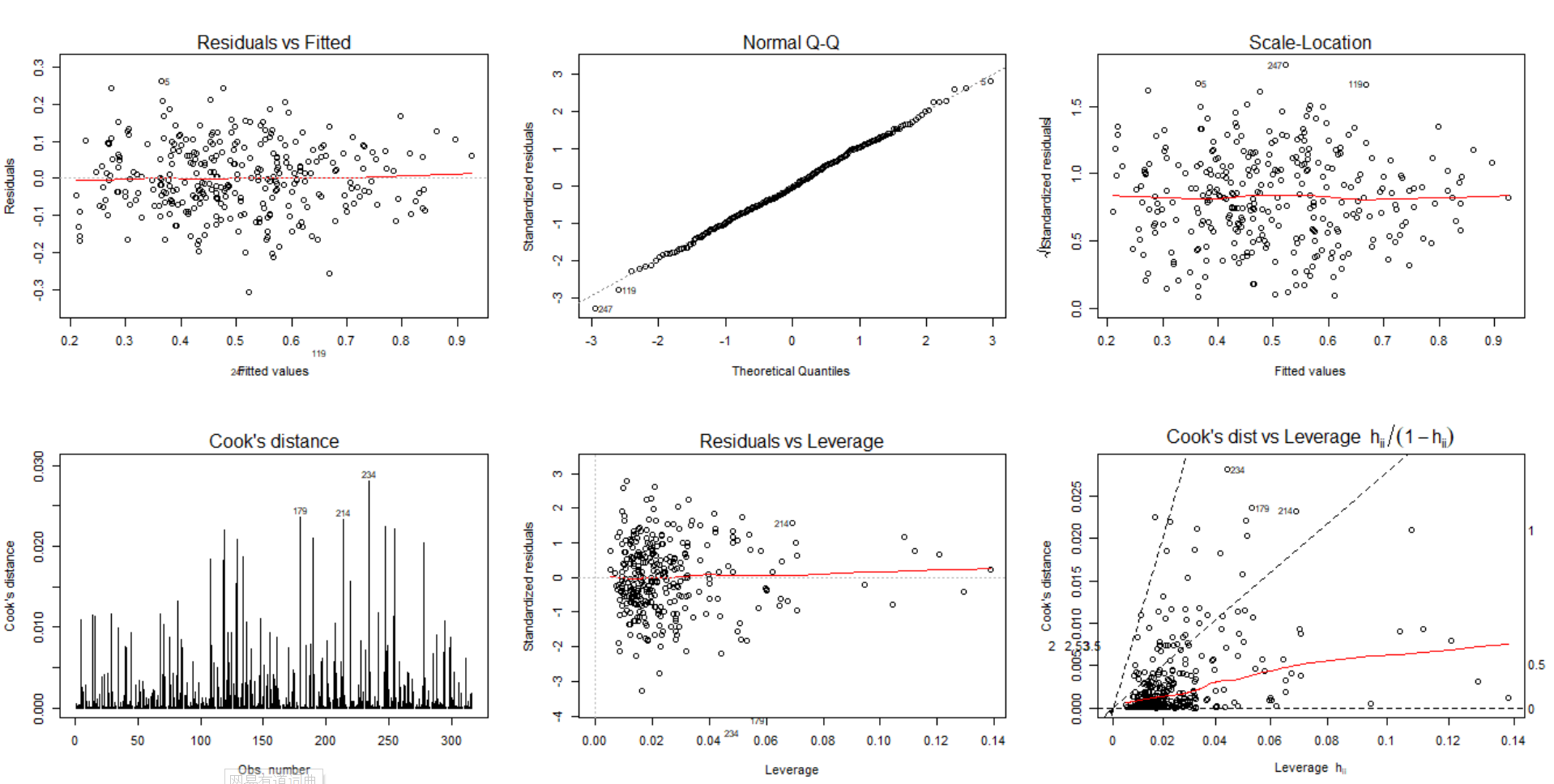
**MODEL DIAGNOSTICS**

Figure 1: Plots Summarizing Data from Model 8

Plot 1 shows a fairly straight line and few points including point 150, point 119 and point 249 stand out because of their relatively large residuals. Plot 2 shows there is no major deviation from the normality assumption except few points on the tail regions. The loess lines in plot 3 and plot 5 show the fairly straight lines as well. However, point 150, point 247and point 275 would be more possibly classified as outliers due to their relatively high Cook’s D values and low leverage values.

After removing the outliers, the final model has better performance as shown in the table 3.





The six plots are based on the model 9 which has excluded the outliers points (150, 247 and 275) from the original dataset. All of the Cook’s D are better now (< 0.030) although new points with high Cook’s values catch our eyes.

**CONCLUSION**

The regression analysis conducted on the dataset of violent crime rate helped us find out factors that have strong association with this severe social issue. The ideal final model would be helpful to predict the possible crime rate in the United States if corresponding data were available. I tried to remove the issue of multicollinearity by check the VIF values that are supposed to less than 10 in order to find the best mathematical model. Most of the part, the normality and linearity assumptions were not violated. Three observations were removed to check whether the model could have a little bit better performance. Though these factors that selected from 28 measurements may be highly significant from the prospective of regression analysis, real world situation would not exactly follow the pattern I predicted.