# Predicting Violent Crimes per 100k Population according to 1990 US Census

**Author Names**: Varalakshmi Vakkalagadda

Sai Gaurav Daula

Vigna Shree Telukunta

Leela Sai Sankar Reddy Parvatham

**Professor Name**: Ranji

**University Name**: George Mason University

**INTRODUCTION**

## Why this Dataset?

What could be the reason behind the violent crime rate in US? The Violent crimes in US are mainly the murders, rapes, robberies, and assault. The violent crimes in US have seen a decline in the last two decades. Aggravated assault is the most common one in the various type of crimes reported in US. In 2018, the total crime rate was reported to be 382.9, in which 246.8 was for aggravated crime. It's important to note that, at times few violent crimes may not be reported, so we cannot assume the reported crime rate to be a very accurate.

## Source of the data set:

The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. The dataset contains 2216 records and 147 attributes.

The dataset has all the attributes that have a connection to the crime. The dataset contains attributes such as median family income, percent population considered urban, number of murders, rapes, assaults, etc. The per capita violent crimes variable was calculated using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault. Some of the important attributes in the dataset are

| Attribute Name | Description |
|---|---|
| **State, County, Community and Community Name** | |
| **Population** | Population of the community |
| **Household Size** | Mean people per household |
| **Racepctblack, RacepctWhite, RacepctAsian** | Percentage of population that are African American, Caucasians and Asian |
| **agePct12t21, agePct12t29, agePct16t24 and agePct65up** | Percentage of population in 12-21, 12-29, 16-24 and above 65 |
| **numUrban, pctUrban** | Number and Percentage of Urban people in the community |
| **medIncome** | Median Household income |
| **pctWWage, pctWFarmSelf, pctInvInc, pctWSocSec, pctWPubAsst, pctWRetire** | Percentage of households with wage, farm or self-employment, investment or rent, social security, public assistance, and retirement income. |
| **medFamInc** | Median Family Income differs from household income from non-family households |
| **whitePerCap, blackPerCap, indianPerCap, AsianPerCap, OtherPerCap and HispPerCap** | Per capita income of Caucasians, African and Native Americans, Asian, other and Hispanic heritage. |
| **NumUnderPov, PctPopUnderPov** | Number and Percentage of People under Poverty Level |
| **PctLess9thGrade, PctNotHSGrad, PctBSorMore** | Percentage of people 25 and over with less than 9th grade, not high school and bachelor's degree education |
| **PctUnemployed, PctEmploy, PctEmplManu, PctEmplProfServ, PctOccupManu and PctOccupMgmtProf** | Percentage of people 16 and over in labor force and Unemployed, Employed, Manufacturing, Professional Services and Management or Professional occupations |
| **MalePctDivorce, MalePctNevMarried, FemalePctDiv, TotalPctDiv** | Percentage of Males who are Divorced and Never Married, Percentage of Female Population Divorced. Total Percentage of Population who are Divorced |
| **PersPerFam** | Mean Number of People per Family |
| **PactFam2Par,PactKids2Par,PactYoungKids2Par, PctYoungKids2Par,PctYoungKids2Par, PctWorkMomYoungKids, PctWorkMom** | Percentage of kids per family, Percentage of kids with 2 parents, Percentage of teens and young kids, percentage of young working moms. |
| **NumIlleg, PctIlleg** | Number and percentage of kids born and never to be married. |

| | |
|---|---|
| **NumImmig, PctImmigRecent, PctImmigRec5, PctImmigRec8, PctImmigRec10, PctRecentImmig, PctRecImmig5, PctRecImmig8, PctRecImmig10** | These fields contain data of people born outside the country and percentage of population migrated 3-10 years back |
| **PctSpeakEnglOnly, PctNotSpeakEnglWell** | Percentage of people who do not speak English and people who do not speak English well. |
| **PctLargHouseFam,** | Percentage of large family households. |
| **PersPerOccupHous,PersPerOwnOccHous, PersPerRentOccHous,PctPersOwnOccup, PctPersDenseHous, PctHousLess3BR** | Percentage of people who do not own a household, have rented a house and percentage of people per household. |
| **MedNumBR, HousVacant** | Median number of bedrooms, percentage of vacant house. |
| **PctHousOccup, PctHousOwnOcc** | Percentage of house occupied and owned. |
| **PctVacantBoarded, PctVacMore6Mos** | Percentage of vacant houses and houses vacant for more than 6 months. |
| **MedYrHousBuilt,** | Median year housing units built |
| **PctHousNoPhone, PctWOFullPlumb** | Percentage of house without phone and without plumbing |
| **OwnOccLowQuart, OwnOccMedVal, OwnOccHiQuart, RentLowQ, RentMedian, RentHighQ** | Owned and rented house: low, medium and high quartile. |
| **MedRent, MedRentPctHousInc, MedOwnCostPctInc, MedOwnCostPctIncNoMtg** | Median of rent and houses owned of household income |
| **NumInShelters, NumStreet** | Number of people in shelters and on street |
| **PctForeignBorn, PctBornSameState** | Percentage of people foreign born and born in the same state. |
| **PctSameHouse85, PctSameCity85, PctSameState85** | Percentage of people living in the same city, state, house for 85 years. |
| **LemasSwornFT,LemasSwFTPerPop, LemasSwFTFieldOps,LemasTotReqPerPop, PolicReqPerOffic, PolicPerPop** | Information about police officers. |
| **RacialMatchCommPol, PctPolicWhite, PctPolicBlack, PctPolicHisp, PctPolicAsian, PctPolicMinor** | Racial information about the police |
| **OfficAssgnDrugUnits, NumKindsDrugsSeiz** | Number of officers assigned to the drug unit and number of drugs sized. |
| **PolicAveOTWorked** | Average overtime work of a police officer |
| **LandArea** | Land area in square miles. |
| **PopDens,** | Population density |
| **PctUsePubTrans, PolicCars,** | Percentage of people using public transport and number of police cars |
| **PolicOperBudg** | Police operation budget |
| **LemasPctPolicOnPatr, PolicBudgPerPop,** | Total police budget and police budget per population |
| **ViolentCrimesPerPop** | Violent crimes per population (PREDICTIVE VARIABLE) |
| **LemasGangUnitDeploy,LemasPctOfficDrugUn, LemasPctPolicOnPatr** | Number of full-time officers on patrol, unit gang deployed, officers per gang unit. |

Table 1. Description of Attributes

## OBJECTIVE

The main areas where I wanted to do my analysis was to predict what factors contribute to violent crimes in a society. Firstly, I wanted to analyze if the number of vacant houses in a locality or the period the house was vacant had any significant contribution to the crime rate. Unemployment is another factor which contribute proportionally to crime rate, therefore I wanted to analyze if the unemployment rate directly affects the crime rate in a community. The next thing I wanted to concentrate on was age. As the age of the defaulter describes the state of mind I wanted to analyze if a certain age group was more vulnerable to crime. It is believed that people from certain ethnicity are more prone to commit and crime than others thus I wanted to analyze and confirm if people from a ethnicity are prone to crime. Education plays and important role in shaping a child's mind, thus with the information in the dataset I have analyzed if education helps bring down the crime rate.

## Hypothesis testing:

First let's discuss what is meant by hypothesis, it is an informed supposition about something in your general surroundings. It ought to be testable, either by test or perception. In other words it is a statistical way to test the results of a survey to see If we have meaningful results, basically we are testing whether our results are valid by understanding the odds of occurring. In hypothesis testing we perform test to prove our Null hypothesis is accepted or rejected.

**$H_0$: The greatest number of people who commit crimes are in the age range of 12 to 24, that of age 12 to 29.**

$H_0$ is rejected. As the number of people in the age range of 12 to 24 are 31997 and the number of people in the range of 12 to 29 are 61233. Thus, the greatest number of people in the age range of 12 to 29 are more.

**$H_0$: The total number of crimes are more in the states of CA, MA, TX, NJ compared to that of any other states.**

$H_0$ is accepted. The total number of crimes in CA are 279, MA are 123, TX are 162 and NJ are 211. Thus we can say that the above listed states have the most number of crimes.

**$H_0$: The number of crimes that took place include individuals from all type of races.**

$H_0$ is rejected. The total number of crimes committed by Whites are 186015, Black are 20677, Asian are 5914 and Hispanic are 17609. Looking at the results we can conclude that people from the race White and Black are more compared to any other race.

**$H_0$: Crimes occur in accordance to high Income individuals i.e., the individuals who have more income are targeted more.**

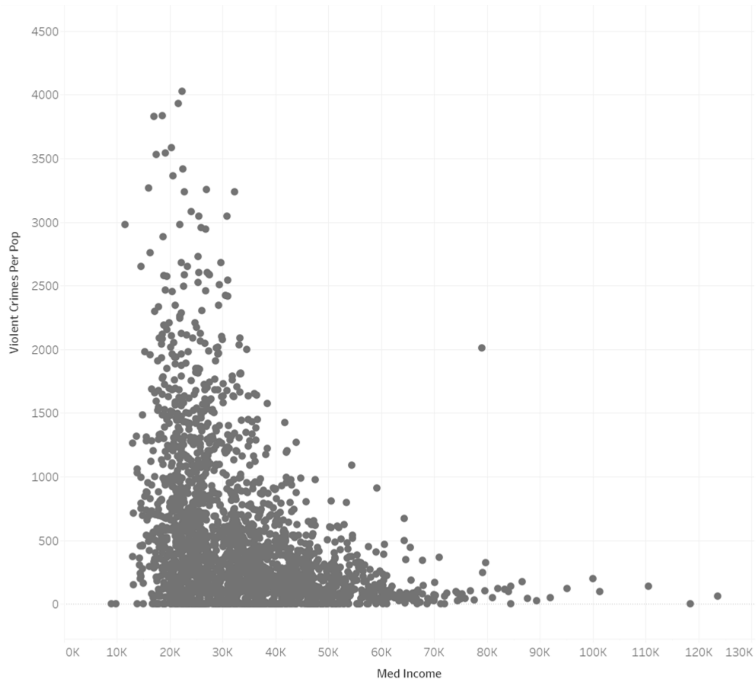


Fig 1. Correlation plot of Median Income of Family the Community and the Rate of Violent Crimes Per Population in the Communities

From the above graph we can say that $H_0$ is rejected. The crimes recorded are more on people who fall under a median income range less than 60K. We can say that due to low income and less security in low income houses they can be assumed as an easy bait or an easy target. Thus, we can conclude the crimes occur on population whose income is less than the average income.

**$H_0$ : The crimes occur more in the region where there is more poverty.**

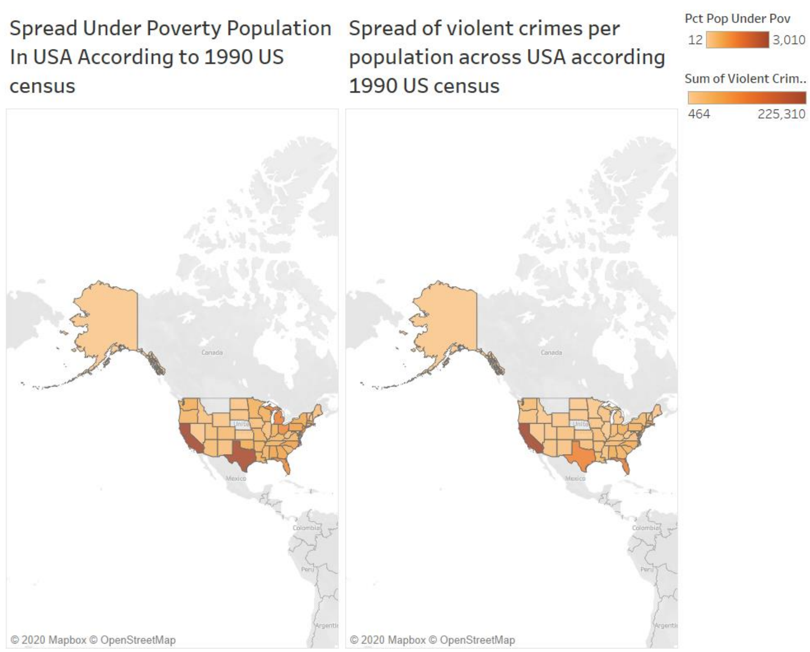


**Fig 2.** The relation between Poverty Rate and Violent Crime Rate in the community across various states.

The map highlights the poverty regions in the country. We can assume that people who live in areas where the rate of poverty is high live with people with almost no or very little income and this leads to more crime. Thus, from the above graph we can say that $H_0$ is accepted. Here we can say that the crimes are more and directly proportional to that of poverty of the state.

## PREPROCESSING

Data preprocessing is an essential step in data mining, because if the data is sent unclean to the model training, it would deteriorate the accuracy and quality of results. Data has to be cleaned, i.e, missing values, noise and outliers should be handled properly.

### HANDLING MISSING VALUES

1. **Replacing with user-defined constant**

The missing data values in the dataset is are represented with '?', we have replaced it with 'NA'.

```
> fNA_counts <- sapply(data, function(x) sum(is.na(x)))
> fNA_counts
      communityname               state          countyCode       communityCode                fold          population 
                  0                   0                1221                1224                   0                   0 
      householdsize        racepctblack        racePctWhite        racePctAsian         racePctHisp         agePct12t21 
                  0                   0                   0                   0                   0                   0 
        agePct12t29         agePct16t24         agePct65up           numbUrban            pctUrban           medIncome 
                  0                   0                   0                   0                   0                   0 
           pctWWage        pctWFarmSelf          pctWInvInc          pctWSocSec         pctWPubAsst          pctWRetire 
                  0                   0                   0                   0                   0                   0 
          medFamInc           perCapInc         whitePerCap         blackPerCap        indianPerCap         AsianPerCap 
                  0                   0                   0                   0                   0                   0 
        OtherPerCap          HispPerCap         NumUnderPov       PctPopUnderPov     PctLess9thGrade        PctNotHSGrad 
                  1                   0                   0                   0                   0                   0 
         PctBSorMore       PctUnemployed           PctEmploy          PctEmplManu      PctEmplProfServ        PctOccupManu 
                  0                   0                   0                   0                   0                   0 
     PctOccupMgmtProf      MalePctDivorce       MalePctNevMarr        FemalePctDiv         TotalPctDiv          PersPerFam 
                  0                   0                   0                   0                   0                   0 
          PctFam2Par         PctKids2Par    PctYoungKids2Par          PctTeen2Par  PctWorkMomYoungKids          PctWorkMom 
                  0                   0                   0                   0                   0                   0 
    NumKidsBornNeverMar  PctKidsBornNeverMar           NumImmig      PctImmigRecent        PctImmigRec5        PctImmigRec8 
                  0                   0                   0                   0                   0                   0 
       PctImmigRec10      PctRecentImmig        PctRecImmig5        PctRecImmig8       PctRecImmig10     PctSpeakEnglOnly 
                  0                   0                   0                   0                   0                   0 
   PctNotSpeakEnglWell     PctLargHouseFam    PctLargHouseOccup     PersPerOccupHous    PersPerOwnOccHous   PersPerRentOccHous 
                  0                   0                   0                   0                   0                   0 
      PctPersOwnOccup     PctPersDenseHous      PctHousLess3BR            MedNumBR          HousVacant          PctHousOccup 
                  0                   0                   0                   0                   0                   0 
       PctHousOwnOcc     PctVacantBoarded      PctVacMore6Mos        MedYrHousBuilt      PctHousNoPhone        PctWOFullPlumb 
                  0                   0                   0                   0                   0                   0 
      OwnOccLowQuart         OwnOccMedVal       OwnOccHiQuart         OwnOccQrange            RentLowQ          RentMedian 
                  0                   0                   0                   0                   0                   0 
           RentHighQ          RentQrange             MedRent    MedRentPctHousInc    MedOwnCostPctInc MedOwnCostPctIncNoMtg 
                  0                   0                   0                   0                   0                   0 
        NumInShelters           NumStreet      PctForeignBorn     PctBornSameState       PctSameHouse85        PctSameCity85 
                  0                   0                   0                   0                   0                   0 
       PctSameState85         LemasSwornFT      LemasSwFTPerPop    LemasSwFTFieldOps  LemasSwFTFieldPerPop       LemasTotalReq 
                  0                1872                1872                1872                1872                1872 
    LemasTotReqPerPop      PolicReqPerOffic          PolicPerPop    RacialMatchCommPol        PctPolicWhite        PctPolicBlack 
               1872                1872                1872                1872                1872                1872 
        PctPolicHisp       PctPolicAsian        PctPolicMinor   OfficAssgnDrugUnits     NumKindsDrugsSeiz      PolicAveOTWorked 
               1872                1872                1872                1872                1872                1872 
            LandArea             PopDens       PctUsePubTrans            PolicCars         PolicOperBudg    LemasPctPolicOnPatr 
                  0                   0                   0                1872                1872                1872 
   LemasGangUnitDeploy  LemasPctOfficDrugUn      PolicBudgPerPop              murders           murdPerPop               rapes 
               1872                   0                1872                   0                   0                 208 
          rapesPerPop          robberies          robbbPerPop             assaults        assaultPerPop           burglaries 
                208                   1                   1                  13                  13                   3 
          burglPerPop           larcenies          larcPerPop           autoTheft      autoTheftPerPop              arsons 
                  3                   3                   3                   3                   3                  91 
          arsonsPerPop  ViolentCrimesPerPop         nonViolPerPop 
                 91                 221                  97 
```

2. **Deleting Records**

The columns: LemasSwornFT, LemasSwFTPerPop, LemasSwFTFieldOps, LemasSwFTFieldPerPop, LemasTotalReq, LemasTotReqPerPop, PolicReqPerOffic, PolicPerPop, RacialMatchCommPol, PctPolicWhite, PctPolicBlack, PctPolicHisp, PctPolicAsian, PctPolicMinor, OfficAssgnDrugUnits, NumKindsDrugsSeiz, PolicAveOTWorked, PolicCars, PolicOperBudg, LemasPctPolicOnPatr, LemasGangUnitDeploy, PolicBudgPerPop had 80% of the data values missing. So, we dropped these attributes.

3. **Replacing missing values with Mean or Median**

The missing values in rape, robbery, burglaries, larcenies, auto thefts, arsons and assault have been replaced by the median, as all the data in these attributes is right or left skewed, it has been replaced with Median. Only rape, robbery, murder and assault are considered as violent crimes according to the 1990 US Census.

The attribute rapesPerPop was replaced using the formula:

df$rapesPerPop <- ifelse(is.na(df$rapesPerPop),((100000)/df$population) * df$rapes , df$rapesPerPop)

The attribute had to be replaced using this formula, as the outcome variable, i.e, violent crimes is calculated per population.

**4. Replacing Missing values in State attribute**

The value in State are character type, so this cant be used in modelling. We replaced the character type values in State attribute with FIPS codes (FIPS codes are numerical values that uniquely defines each state).

## Regression Trees

The Regression trees partition the dataset into smaller groups and fit a model on to each subgroup an association tree is built and developed incrementally. It is built by a process called binary recursive portioning. The regression trees have root, leaf, decision, and child nodes. The main advantage of Regression trees is we can visualize the tree in each step and decision making is easy. From Fig 3. We can see that Assaults per population the root node stands important in making decision. The leaf values represent the value of Violent crimes per population. On whole the algorithm used Assaults per population, Rapes, States, Robberies Per population and Assaults Per population in generating the Regression tree. The RMSE value **of R-square is 36.72%** which is very low.

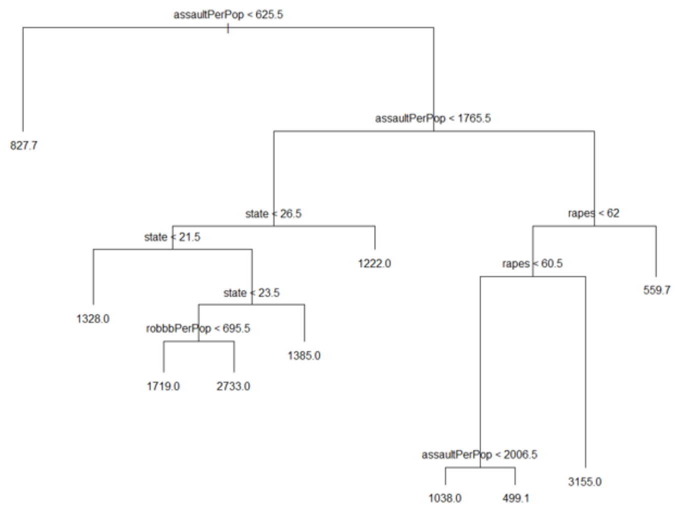


Fig 3. Regression Tree

## Random Forest Classifier

Random Forest is a supervised Learning Algorithm that can be used for both Regression and Classification. It aggregates many Decision trees. Each tree Randomly draws samples from original dataset and generated splits. Adding this randomness will prevent from overfitting. From Random Forest Model we can generate Variable importance plot to identify the most important variables that are used in predicting the final attribute. The R-square value of the Random Forest Model is 51.93%. From the Fig 4 we can see that Assault per population, Rapes, States, Robberies and Rapes per population are important in predicting the violent attributes.

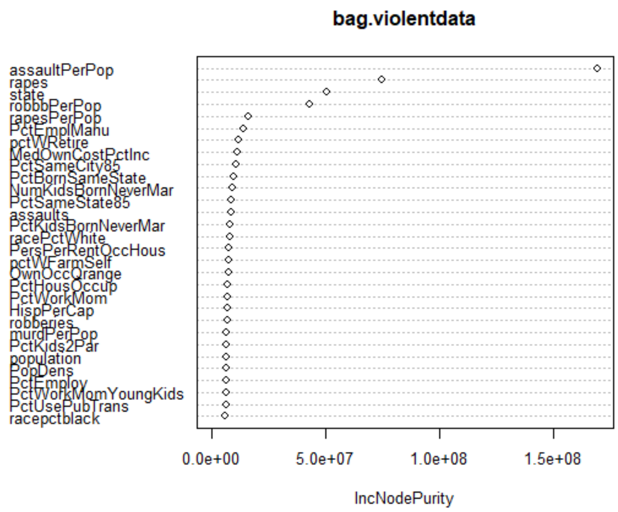


**Fig 4.** Variable Importance Plot generated from Random Forest Model

| Classifier | RMSE | R-Square |
|---|---|---|
| **Regression Trees** | 573.358 | 36.72 |
| **Random Forest** | 357.934 | 51.93 |

**Table 2**: Model Performance on data before performing PCA

The R-square values generated from Regression trees and Random Forest are very low. So, we further analyzed the models by removing the skewness using Box-Cox Transformation and performed Principal Component Analysis to remove the over lapping information.

## Principal Component Analysis

After filtering and cleaning the data from the missing values, outliers and inconsistent data, We have transform the data if there is any multicollinearity in the data for the numerical predictors. PCA is a dimensionality reduction algorithm, that reduces the number of variables in the dataset into various principal components by removing the overlapping information. The most important use of PCA is to represent multi variate data in tabled to smaller set of variables. From Fig 3, we can see that by applying Principal Components Analysis the 113 attributes have been reduced to 43 components with 95% variance explained.

Here, I used Preprocess function of Caret package to remove skewness in the data, perform Box-Cox transformation and Principal Component Analysis. From the below results we can see that Box-Cox transformation is performed on 83 attributes and 113 attributes are centered and scaled.

```
Created from 2214 samples and 113 variables

Pre-processing:
  - Box-Cox transformation (83)
  - centered (113)
  - ignored (0)
  - principal component signal extraction (113)
  - scaled (113)

Lambda estimates for Box-Cox transformation:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.0000 -0.2000  0.4000  0.3855  0.9500  2.0000

PCA needed 43 components to capture 95 percent of the variance
> 
```

**Fig 5**. The Results of Principal component Analysis

## Linear Regression

Linear Regression is a linear approach of modelling the relationship between scalar response and one or more explanatory variables. The overall idea of linear regression is to examine two things does it set the predictor variables and do a god job in predicting the outcome variable and to identify which variables are significant predictors of the outcome variable. Our predicting variable is ViolentCrimesPerPop and our response variables are burglaries, burglPerPop, larcenies, larcPerPop, autoTheft, autoTheftPerPop, arsons, arsonsPerPop, nonViolPerPop. Initially when tried to fit a Linear Regression model on the data the R square is very 42%. The R-square is very low because of many variables variance has impacted on the prediction of the final attribute. Later, after removing skewness and applying PCA the R-square value has improved drastically to 96% with Residual error as 125.5. From this we can say that the data has lots of overlapping information. Using "sigma.default <- function (object, use.fallback = TRUE, ...) *sqrt( deviance(object, ...) / (NN - PP) )" this function we could calculate the Residual Standard Deviation and we know the error rate in the model from which we could calculate the Residual standard error which is 125.5 on 2170 degree of freedom. All the response variables that we choose are important to obtain good accuracy.

## Diagnostic Plots:

The diagnosis plots can be created using plot() function of ggplot2 package. These plots explain residuals in 4 different ways.

Residual vs Fitted: The main aim of this plot is to check the linear Relationship assumptions. From our plot we can say that residuals are spread across horizontal line without any distinct patterns. This a good indication and proves we don't have any non-linear relationships.

Normal Q-Q plot: Used to explain whether the residuals are normally distributed or not. From the plot we can say that there is slight deviation in residuals and point 359 on top right corner is deviated and far away from the regression line.

Scale-Location plot: This plot is used to check the homogeneity of variance of the residuals. From the plot we can see that the residuals are equally spread across the fitted line. There is slight deviation on the end of right side of the fitted line.

Residual vs Leverage: This plot helps us to find if any influential points in the data. Not all outliers influence the linear regression Analysis and create problem in generating the regression line. But few points are very influential and alter the results. From this plot we can see that are no points on upper right corner or lower right corner. Hence, we can say that there are no points in the data that influence in regression line.

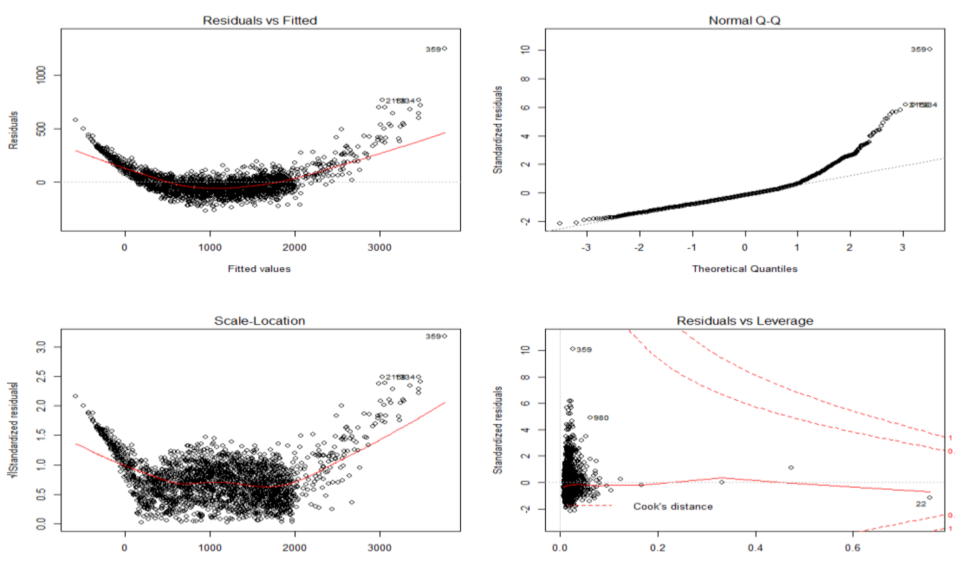


**Fig 6.** Diagnostic plots of Linear Regression

### Lasso and Ridge Regression:

Lasso and Ridge Regression are used to find and reduce the set of variables to result a optimal performance model. In Ridge Regression the variables with minor contribution have their coefficients close to zero, whereas in Lasso regression the coefficients of less contributive variables are forced to be exactly zero. Only the most significant value are kept on the model. The glmnet package does cross-validation to identify the best lambda that fits the model and minimizes the test error. The plot shows Mean Square error on y-axis and log(lambda) on x-axis. The vertical dotted line explains the lowest Mean Square error and the second dotted line explains log of lambda value with one standard error. We can see that with little increase in value we got 42 predictors (principal components) compared to lowest mean-squared error which has 35 predictors. Fig 7(a) visually describes this. The plot in Fig 7(b) chooses best lambda that minimizes the test error. The plot shows log of lambda on x axis and coefficients on y-axis. Every colored line corresponds each predictor. The big the lambda values get the more the coefficients are shrunk to zero. From Fig 7(b) we can see that the principal component 25 and 15 are highly impacting on the final predictor variable.

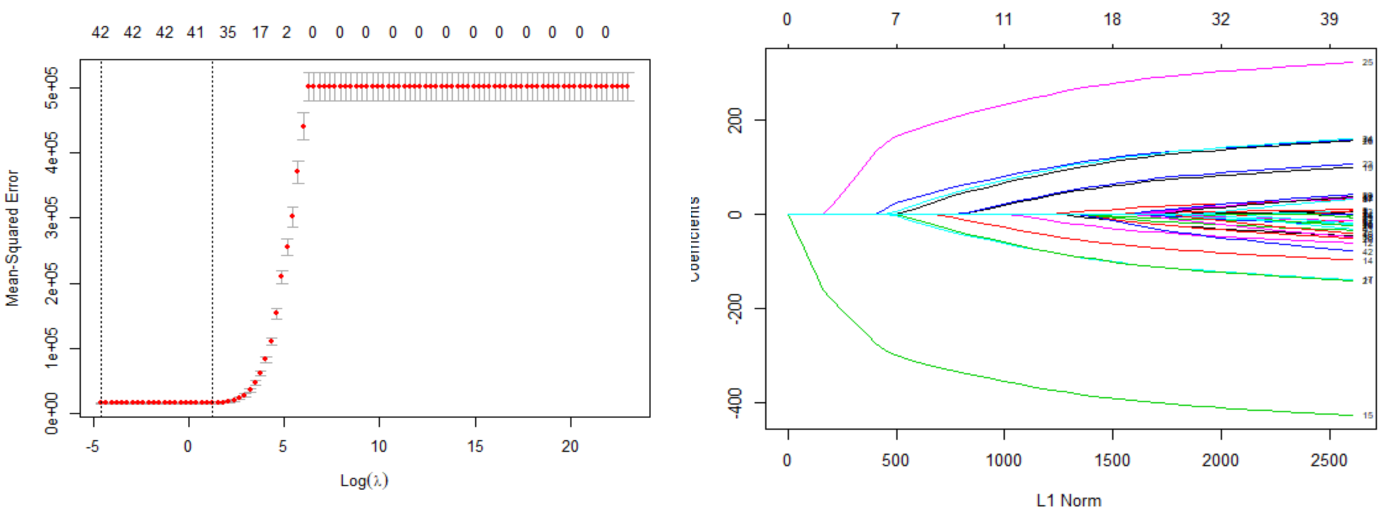


**Fig 7(a) and 7(b)** : Plots generated from Lasso Regression

The Fig 8 explains the plots generated from Ridge Regression. From Fig 8(a) we can see that all 43 predictors are included in the model for best lambda and also lambda with one standard error. Where as we can see in Lasso Regression only 35 are considered for lambda with 1 standard error. From Fig 8(b) we can say that Principal component 25 and 15 are highly varaying and impacting on the final predictor variable.

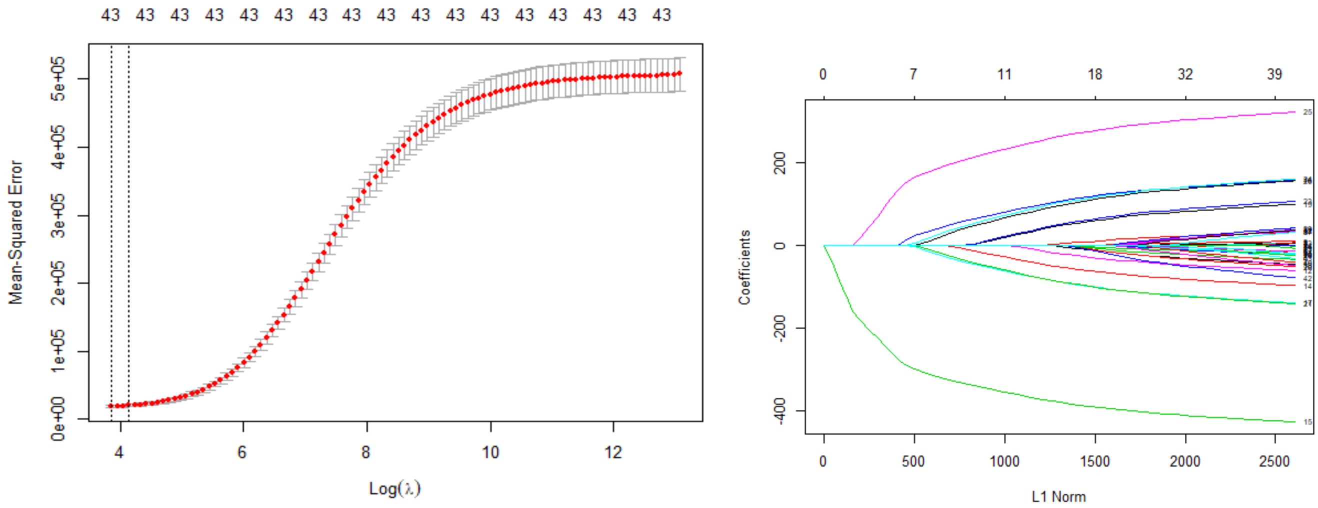


**Fig 8(a) and 6(b)** : Plots generated from Ridge Regression

**Model Performance Comparision**

| Classifier | Root Mean Square Error | R-Square |
|---|---|---|
| **Linear Regression** | 125.25 | 96.92 |
| **Lasso Regression** | 120.95 | 96.94 |
| **Ridge Regression** | 122.13 | 96.92 |

**Table 3**. Comparision of performance of Models

From the above table, we can see that all the models have same R square value in predicting the final attribute violent crime per population.

## Conclusion

Intially the performance of the models is very low, later on after removing the skewness and performing Principal Component Analysis the Rsquare value has improved drastically from 51% to 96%. This implied that the data has lot of over lapping information. From the variable importance plot we can say the Rapes and Assaults are important in predicting the violent crime rate.

## Instructions to Run Code

**Step1**: Open the "Daula_Leela_telukunta_Vakkalagadda.R" file

**Step 2:** Replace the file directory of the datasets. The datasets are included in the zip file or can be found here.

**Step 3:** All libraries and dependencies are listed in the code. The code can be run until the line 12.

**Step 4:** Each model can be found in the respective section labeled with the comments.

**Step 5:** From line 14, the preprocessing begins. The preprocessing is done separated for all models and the models begins at line 221.

**References**

[1] Francis, & Don. (2018, March 11). Penalized Logistic Regression Essentials in R: Ridge, Lasso and Elastic Net. Retrieved from http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/

[2] Francis, & Don. (2018, March 11). Penalized Logistic Regression Essentials in R: Ridge, Lasso and Elastic Net. Retrieved from http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/

[3] Schmidt, P. (2020, May 3). The Lasso - R Tutorial (Part 3). Retrieved from http://thatdatatho.com/2018/05/07/the-lasso-r-tutorial-part-3/