Final Report

ANLY 506-50- R-2017/Fall - Exploratory Data Analytics

Rajbir Dadhiyala

Janice Lim

Aruna Vedula

Darpan Vyas

Harrisburg University

# **Table of Content**

[**Introduction**](#_hr48hp3x10l) **2**

[**Background**](#_boalstvlpaus) **2**

[**Data**](#_9apb7lp7ynyu) **3**

[Data Cleansing](#_w1xj16z3lf08) 5

[Data Preparation](#_kugxvo34zanm) 5

[Data Exploration and Visualization](#_faou0x4nji20) 6

[**Analysis Techniques**](#_3hah4kpb37em) **7**

[**Conclusion**](#_bynxmut1irrr) **9**

[**Appendix**](#_tam34mxgibc7) **11**

# 

# 

# **Introduction**

A key component in comprehending violent crime is understanding where and why it actually happens. Thus, understanding crime provides insights that is vital in developing policies that work.

The current political climate in the United States is highly inclined towards nationalist ideals. Authorities rose in power partly by instilling fear and invoking xenophobia. As budding data analyst and scientists, we wanted to test this fear. Fear that immigrants are increasing responsible for the crime committed in the United States.

Crime, both non-violent as well as violent, is a complex thing to predict. A combination of complex Demographical, Social and Environmental factors contribute to levels of crime in a society. Keeping our core question in mind, and considering the complexity involved, we want to understand the impact of immigrants on crime (violent and nonviolent combined).

The research question for our project is “***Is there a relationship between immigration and higher crimes rates in a given region?***”

# **Background**

We are relying on variety of sources to gain help and information for this project. This includes class lectures, past classes taken at Harrisburg University (Analytics II, Data Visualization, and Quant Decision Making), “R for Data Science” by Hadley Wickham, and finally the Cran-R Project.

# **Data**

The base data for this project is sourced from Communities and Crime Unnormalized Data Set, from UCI Machine Learning Repository. This dataset combines socio-economic data from the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR.

Since our research question deals with immigration and crime, we narrowed down our data frame to contains only the following columns that are relevant to our analysis.

Below are the categorical variables needed for our analysis:

1. **state**: US state
2. **communityname**: community name

Below are the continuous variables needed for our analysis:

1. **population**: population for community
2. **NumImmig**: total number of people known to be foreign born
3. **PctImmigRecent**: percentage of \_immigrants\_ who immigated within last 3 years
4. **PctImmigRec5**: percentage of \_immigrants\_ who immigated within last 5 years
5. **PctImmigRec8**: percentage of \_immigrants\_ who immigated within last 8 years
6. **PctImmigRec10**: percentage of \_immigrants\_ who immigrated within last 10 years
7. **PctRecentImmig**: percent of \_population\_ who have immigrated within the last 3 years
8. **PctRecImmig5**: percent of \_population\_ who have immigrated within the last 5 years
9. **PctRecImmig8**: percent of \_population\_ who have immigrated within the last 8 years
10. **PctRecImmig10**: percent of \_population\_ who have immigrated within the last 10 years
11. **ViolentCrimesPerPop**: total number of violent crimes per 100K population
12. **NonViolPerPop**: total number of nonviolent crimes per 100K population

Secondly, we source an additional data source that maps states to regions. This data was downloaded from the internet, which was in the form of an Excel file. The URL for the state-region reference dataset is <http://researchertools.blogspot.com/2012/09/excel-file-with-us-states-abbreviations.html>

## **Data Cleansing**

* The base dataset contained missing values, which were designated by “?”. These were removed and assigned NA while loading data in R dataframe.
* Headers were assigned to all the 147 variables.
* Over 300 instances of violent and non-violent crime variables were NA. Since these variables are part of what would later become our dependent variable, we decided to remove these rows from the data frame. We believed that this would not compromise the data integrity.

## **Data Preparation**

After deleting NA rows, we performed the following actions to our data frame:

* Since our research question deals with crime, both nonviolent and violent, we created a new column ‘Total Crime per 100K pop’ by summing the aforementioned two columns. Thus, you can imagine, having NA values in the two variables created an issue, which we resolved by deleting the rows.
* Secondly, we wanted to analyze data on a different granularity. The base dataset was on city level. In order to analyze data at region level, we needed to downloaded the state-region mapping from the internet, and joined this dataset to our base dataset to bring in region.
* Finally, we identified that DC was not being tagged to any region as it is not considered a state. We had to manually assign it to Northeast region.

## **Data Exploration and Visualization**

* To identify outliers, we created box plots of Number of Immigrants and Total Crimes per pop. Refer Fig. 1 in the Appendix.
* From Figure 2, we have the density plots to see if the response variable is close to normality. The first plot was “numimmig” or Number of Immigrants and second plot was “totcrimeperpop” of Total Crimes per 100k of population.
* A scatter plot, Figure 3 was also created to look at the number of immigrants and total population. Given the density of data, this step did not reveal much additional useful information to draw any conclusion.
* Since we couldn't find any resourceful result from Figure 3, we then created another scatter plot (Figure 4) of Total Crime and Population by Region and Figure 5; scatter plot of Total Crime and Population by Immigrants by Region.
* From the above scatter plots (Figure 4 and Figure 5), we can see that there were some degree of positive correlation between Crime and immigrants population.
* To avoid any sort of biases, we have also created the same scatter plot but on a percentage of population instead of total number of population. From this plot (Figure 6), it shows a negative correlation.
* After all the plotting, we will look at some correlation between the variables related to the research question. Refer to Figure 7.
* We were also interested in looking at the correlation of all related variables between the four different regions. Results shown on Figure 8, Figure 9, Figure 1o, and Figure 11.
* Last but not least, we have also created linear regressions (Figure 12, Figure 13, Figure 14, Figure 15 and Figure 16) to check on the P-value in the national level and all four regions.

# **Analysis Techniques**

After having performed data exploration and generating some graphs to visualize and understand the hidden trends in the data, it was time to validate the findings from the visualizations. This was done by checking Correlations and performing Linear Regression, which we did at a national as well as on the regional level:

* Correlation: This is a measure of how two variables are linked. So we created a correlation matrix of Population, Number of Immigrants and Total Crimes per 100K of population. Then we also created correlation matrices for the above but of each region separately to see the differences by Region.
* Regression: We also built a linear regression model using the “Number of Immigrants” and “Total Crimes per 100K of population” variables nationally and regionally to see the differences by Region.

Once the correlation matrix is created and the linear regression model is build, it is important to interpret the results of the correlation matrix and the linear model to ensure our model holds any statistical significance. Since Correlation is a measure of how two variables are linked, there is no linkage if the correlation is closer to zero. And there is a good linkage as we move closer to 1 or -1 (negative values mean inverse relation).

The correlation matrix when created on the basis of percentage of population, it was found there was as little as negligible correlation between the variable “Number of Immigrants” and “Total Crimes per 100K of population”. To be precise, the figures were, 0.09,0.33,0.23 and 0.09 for the Northeast, Midwest, South and West regions respectively.

With Correlation taken care of, we proceeded to analyze and interpret the results of the Linear Regression. This is done by carefully reading and interpreting the summary statistics of the linear model like the p-Value, the Residual Standard Error, and Adjusted R-Squared values.

Now the general rule of thumb of criteria to follow to validate the regression model and to be able to conclusively say that it holds statistical significance is:

1. The p-value should be less than 0.05 to be able to confidently **reject** the Null Hypothesis, that there is a relationship between the number of immigrants and the crimes taking place.

2. The Residual Standard error is the average distance of the observed values from the regression line. The smaller the value, the better it is, as it signifies a better fit of the regression line.

3. Adjusted RSquared values should be as high as possible and at least higher than 0.07

# **Conclusion**

Reading the values of the three parameters we primarily are looking at (the p-value, the Residual standard error and the Adjusted R squared values) to interpret the results of the regression analysis, it is clear beyond doubt that we can very confidently reject the assumption of the Null Hypothesis, that there is any evidence that the immigrants are affecting the crime rate in the country, regionally as well as nationally.

**Nationally**- The p-value nationally, is 2e-16 the residual standard error is 3220, and the Adjusted R Squared value is 0.005128. All the three parameters indicate to the same fact that the Null Hypothesis should be rejected and there is no relationship between the crimes being committed and the Number of immigrants.

The results were the same, with every region taken into account individually as well. The results of the Linear Regressions run regionally as follows:  
**Northeast**- p-value = 2e-16, the Residual Standard Error = 2777, and the Adjusted R Squared Values = 0.006982  
**Midwest**- p-value = 2e-16, the Residual Standard Error = 2125, and the Adjusted R Squared Values = 0.001025  
**South**- p-value = 2e-16, the Residual Standard Error = 3373, and the Adjusted R Squared Values = 0.04973  
**West-** p-value = 2e-16, the Residual Standard Error = 2778, and the Adjusted R Squared Values = 0.006006

From the data that we have in hand and looking at all the results, regionally and nationally, it can be safely concluded that Immigrants were not affecting the crime rates in the USA. Having said that, it is also important to understand that crime is dependent on a number of factors, and a combination of all those factors need to be taken into consideration to be able to reach and conclusive judgement.

# **Appendix**

Fig 1. Box plot to identify outliers

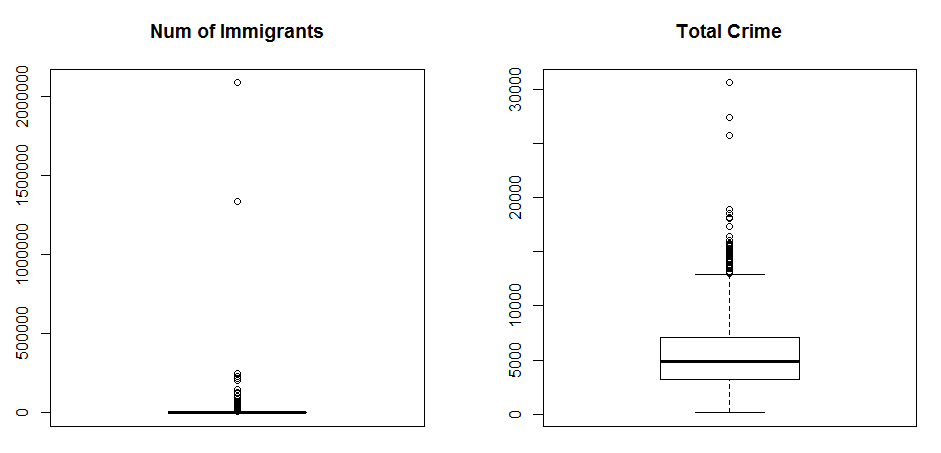


Fig 2. Density plots to identify normality in the variables

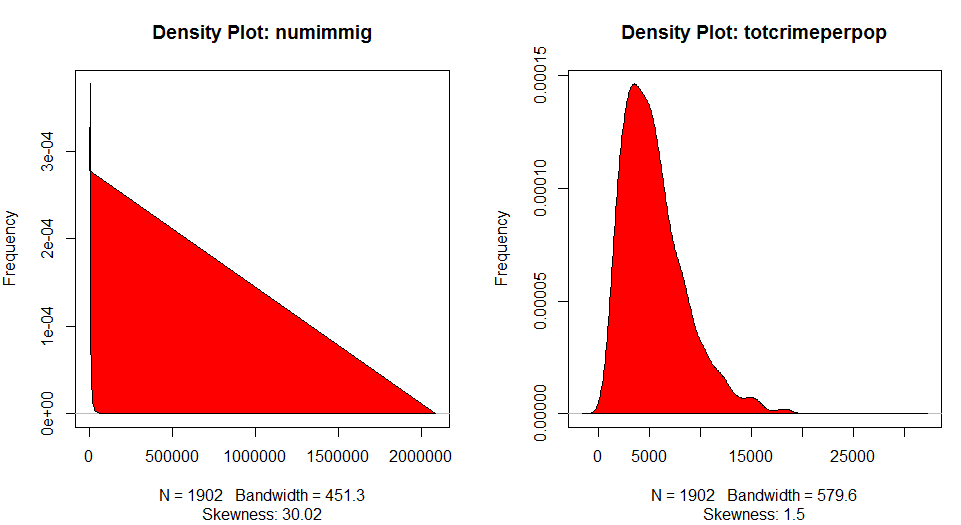


Fig 3. Scatter plot of total population, total immigrants, and total crime (national level)

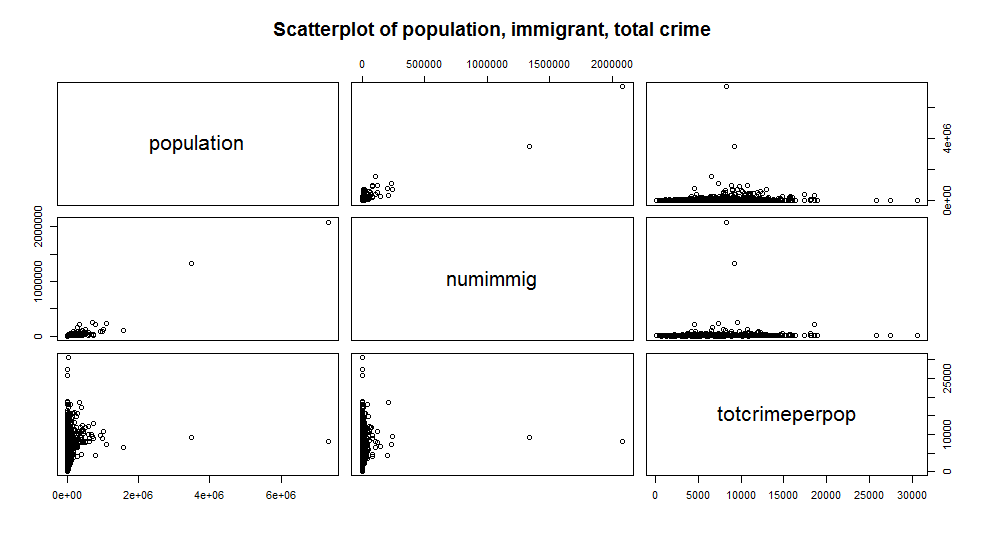


Fig 4. Scatter plot of total crime and total population, by region (y axis: log2 scale)

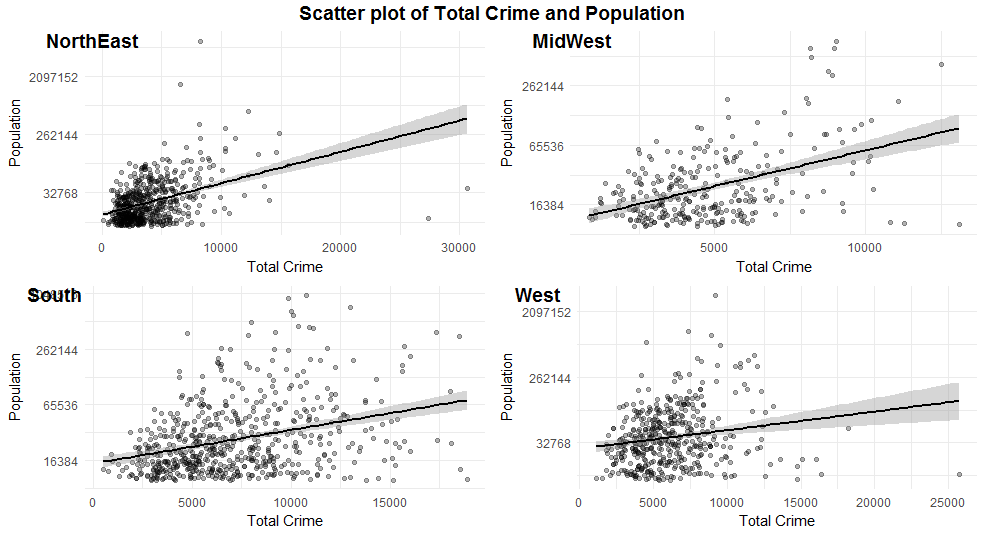


Fig 5. Scatter plot of total crime and total immigrant population, by region (y axis: log2 scale)

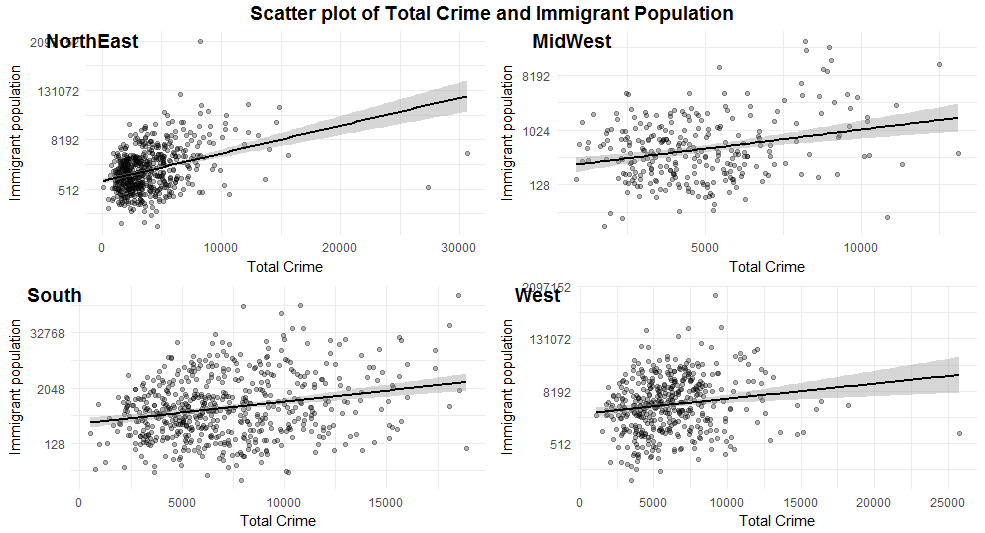


Fig 6. Scatter plot of total crime and total immigrant population on percentage basis compared to total population, broken down by region (y axis: log2 scale)

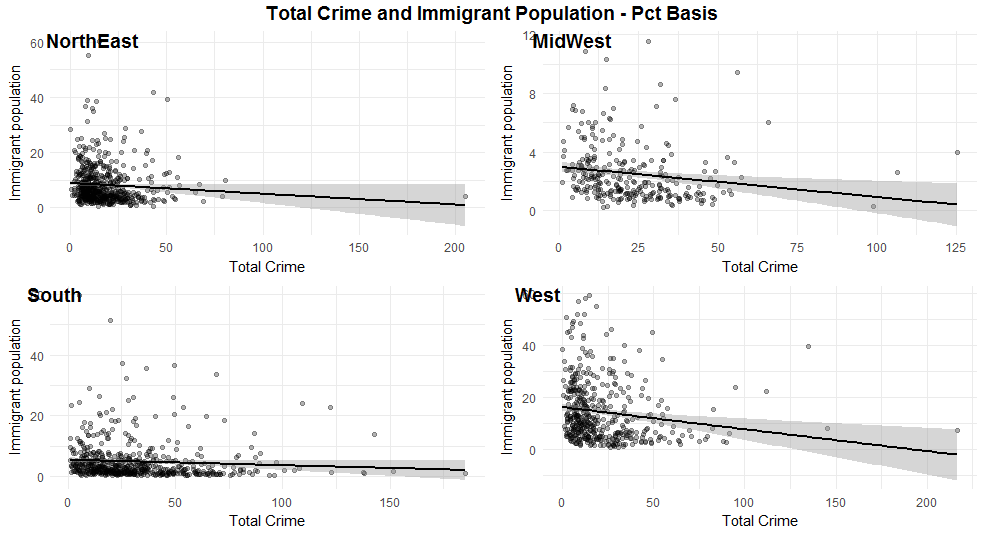


Fig 7. Correlation plot of various variables related to the research question

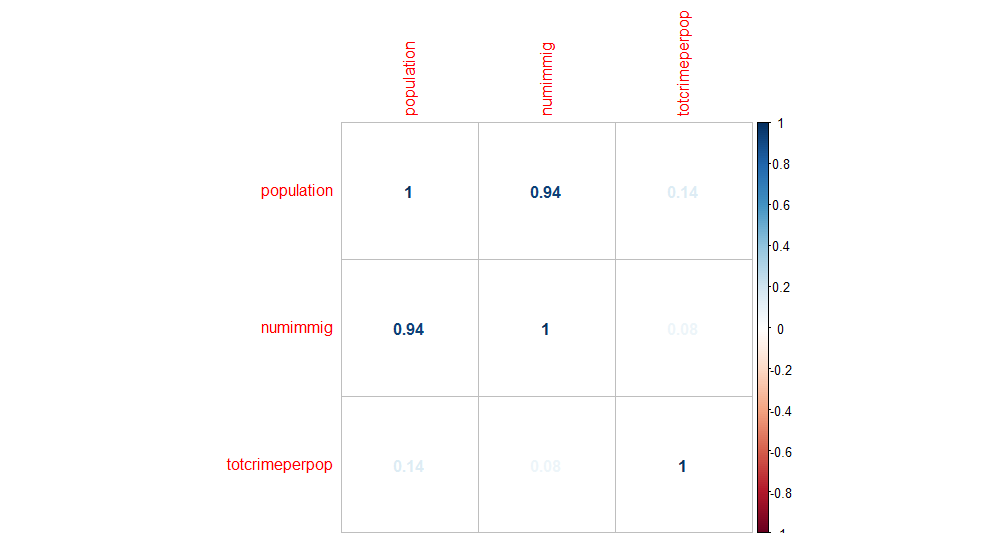


Fig 8. Correlation plot of various variables related to the research question for North East region

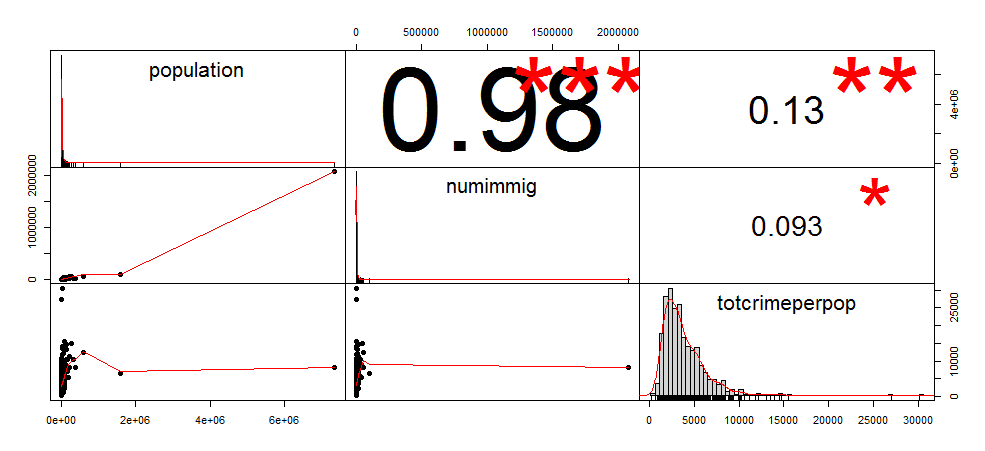


Fig 9. Correlation plot of various variables related to the research question for Mid West region

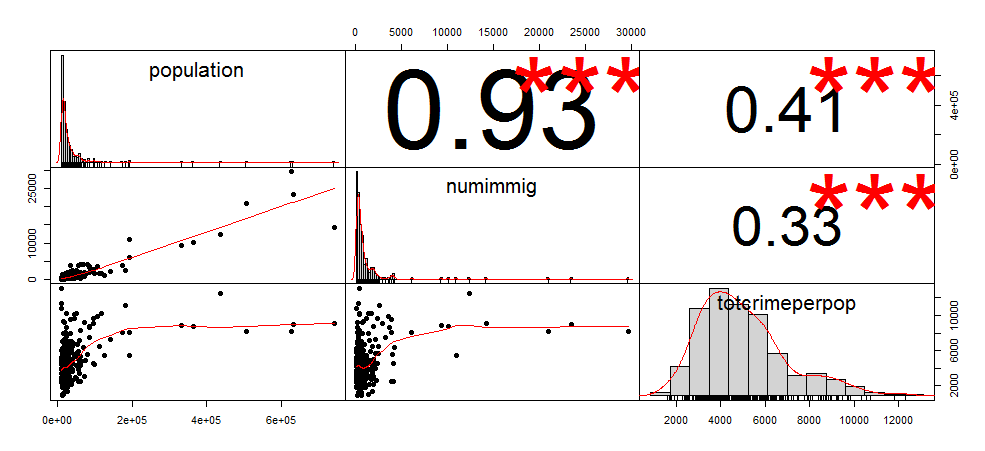


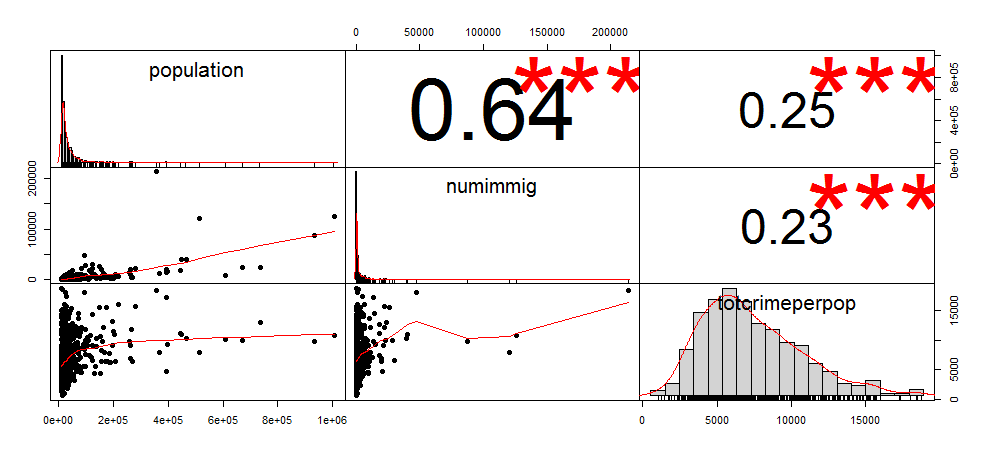
Fig 10. Correlation plot of various variables related to the research question for South region  


Fig 11. Correlation plot of various variables related to the research question for West region

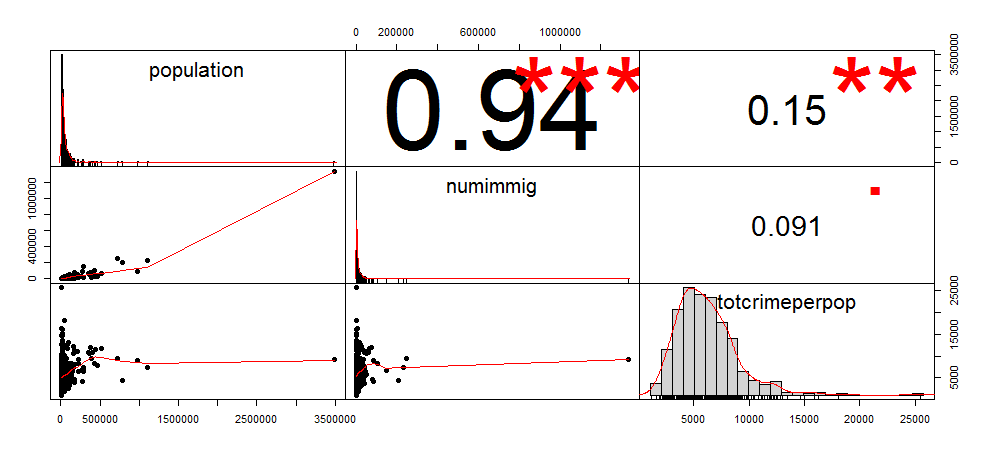


Fig 12. Linear Regression Model Summary, National level

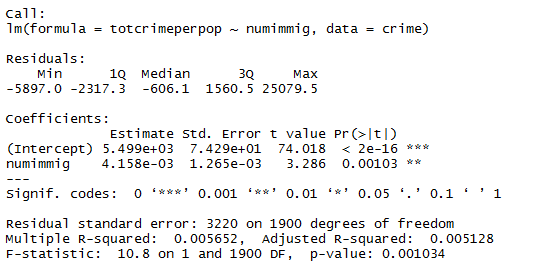


Fig 13. Linear Regression Model Summary, North East level

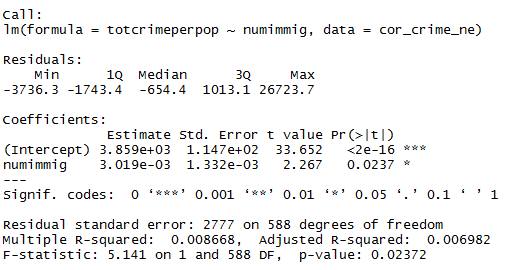


Fig 14. Linear Regression Model Summary, Midwest level

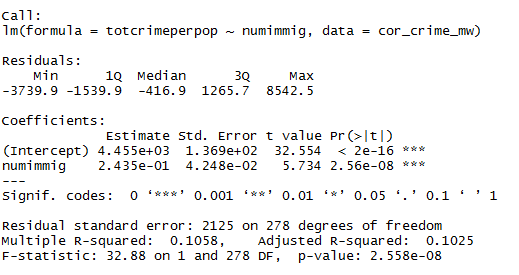


Fig 15. Linear Regression Model Summary, South level

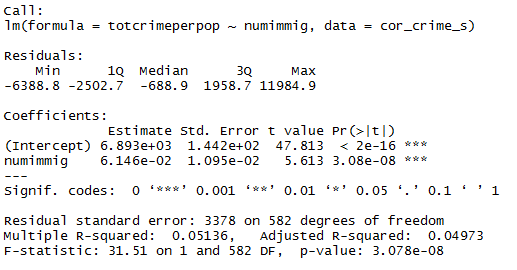
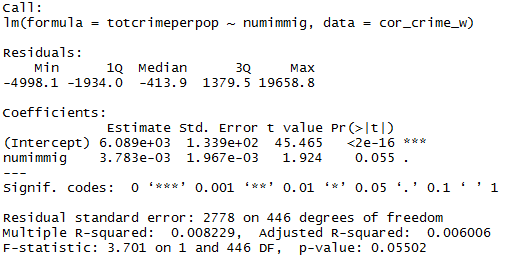


Fig 16. Linear Regression Model Summary, West level



# **Code**

R Code and supporting files for this project is uploaded online at **https://github.com/rd85/crime\_uci\_ml**