Study of the Effects of Immigrants on Crime Rate in USA (Status Report)

Pallava Arasu Pari

Madhu Shri Rajagopalan

Deepti Gupta

Shriya Deshmukh

Rahul Munot

Harrisburg University of Science and Technology

Work Split

Discussions and brainstorming around all the questions – Pallava, Madhu, Deepti

Study of the Effects of Immigrants on Crime Rate in USA (Status Report)

**The Problem Statement**

Question 1. What is the problem you are trying to solve or question you are trying to answer?

The crime dataset is formed by combining 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.

Looking at the dataset, the problem that we will address in our project is the influence of immigration statistics on the crime occurring in that community. We will be analyzing the contrapuntal impact on factors like ethnicity, poverty, median income, age and education with the increase or decrease in the number of immigrants and how it impacts the crime of the city. For example, for a particular city which has a high crime rate, what is the ratio of immigrants in that city with the current population, what is the ratio of the immigrants from different ethnicities as in African American, Caucasian , Asian, Hispanic heritage or others and which ratio affects the crime the most or least, what are the education statistics with respect to these ethnicities and how does the age differs in these communities. One example will be communities with higher crime rate and immigrant ratio might have higher poverty, low education rate, low median income and younger age bracket should have higher percentage.

The other area to analyze crime rate could be if an area with more immigrant population might have an impact on the national average rate of violent and non-violent crimes. We will be analyzing the percentage of immigrants who immigrated within last 3 years, 5 years, 8 years and 10 years to see if there is an impact during these years with these immigrants on the crime rate.

**Background Information**

Question 2. All relevant background information including any relevant literature you have/will use.

Being an immigrant in the country of United States and the recent thoughts of people on the increase in crime rate is caused by immigrants, stirred the interest as data analysts to use data to find out if there is any trend caused due to immigration and its relationship with crime. From the available studies in this area of crime in the United States, most of the studies show the relationship with ethnicity and crime rate. We are interested in finding out if in a state, irrespective of the number of immigrants, factors such as poverty, education, age and employment plays a major role in causing crime. We will be using available analysis on finding out additional data sources to expand our research and include rates in consecutive years to see if there is any trend flow.

**Process**

Question 3. The overall process you will follow for the entire project.

We will be following the generic process of analyzing data and the flow of our process is as below. We have completed the steps – Preprocessing, Data Cleaning and exploratory data analysis.

1. Preprocessing – The preprocessing of our data for proceeding with the analysis is complete. We completed transforming raw data into the understandable format that we will use in next steps.
2. Data Cleaning – The data wrangling process is also complete. We have removed all missing values and checked for data types to use the data as we require in our exploratory data analysis part.
3. Exploratory Data Analysis – The Exploratory data analysis part is 90 percent complete. We have used all the variables of our interest to see the trend patterns with crime. We are also working with expanding our analysis by including data from US census bureau to see trend patterns on a yearly basis to give our research objective, a much clearer conclusion.
4. Models/Methods – This part of our project will begin once we are complete with the exploratory data analysis. We have decided to work with k-means clustering algorithm and we might also work with using PCA to see the accuracy yields.
5. Evaluation – This will be the final part of the project once we have results from our models. We will use the confusion matrix and any other evaluation/prediction technique to assess the performance of our models.

**Exploratory Data Analysis**

Question 4. A description of any relevant, interesting exploratory data analyses.

In our exploratory data analysis part, we worked with our variables of interest, namely, age, ethnicity, poverty, employment and see how the crime rate was related with them. We plotted many statistical charts namely the box plot, bar plot and used maps with gradient scales for better visualizations.

Below are some interesting results we got from exploratory data analysis -

1. When comparing the national top 10 states with highest crime rate with the states where the ethnicity of Blacks, Asians, Hispanics and Whites is the highest, we found that the states where Asians have maximum occupancy are only 2 in contrast to other ethnicities where it matched to 8 out of 10 states.
2. When we saw the pattern of crime like percentage of rapes, burglaries, assaults and robberies etc. in all the four groups (asians, hispanics, blacks and whites) of top 10 states on a similar scaled barplot, we found that the graph for asians as compared to other groups in terms of number of crimes committed showed lesser number of crimes.

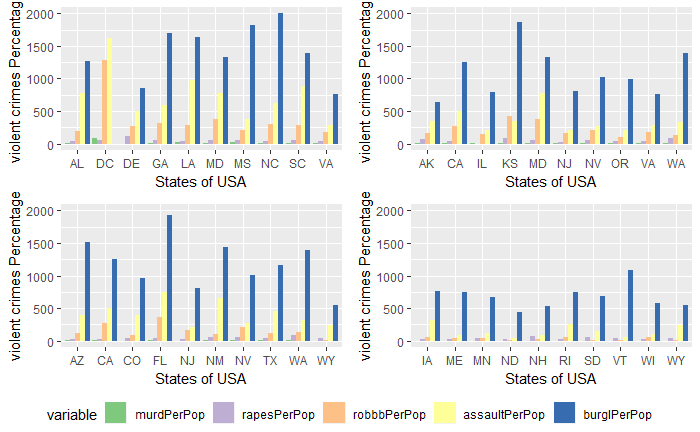


Figure 1: Bar plot showing the trend of crimes for ethnicities : Blacks, Asians, Whites and Hispanics (clockwise)

1. When we plotted top 10 poor states of USA and compared it with top 10 crime rated states, we found that four of them were in common.

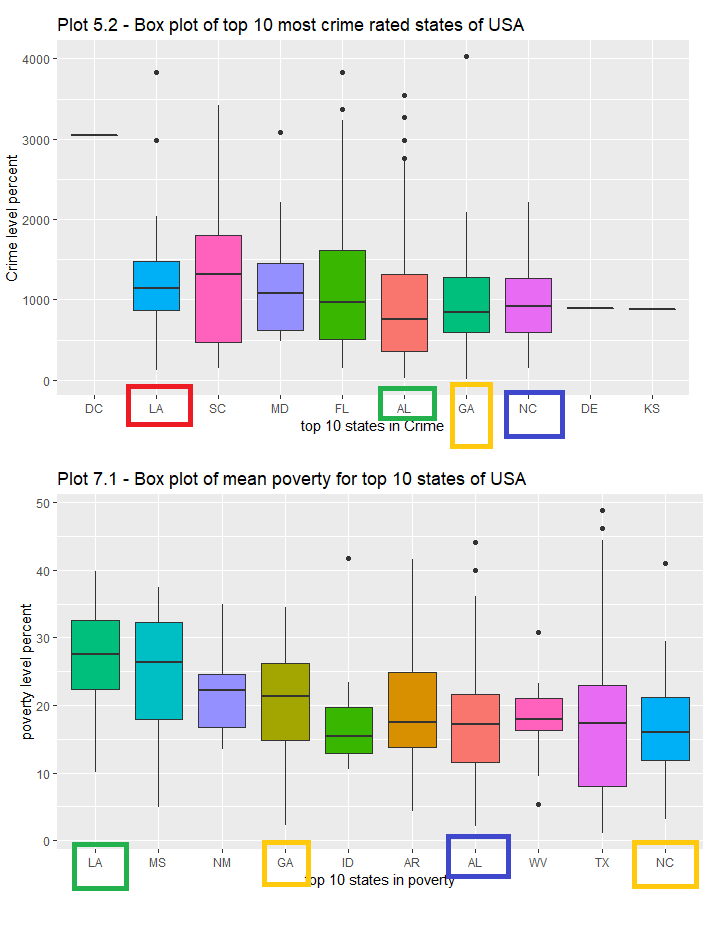


Figure 2: Comparison of top 10 crime rated states with top 10 poor states

1. When we compared top 10 poor states of USA with top 10 most Asian occupied states, we found no state was common
2. When we plotted top 10 less educated (below 9th grade) states of USA and compared it with top 10 crime rated states, we found that four of them were in common.

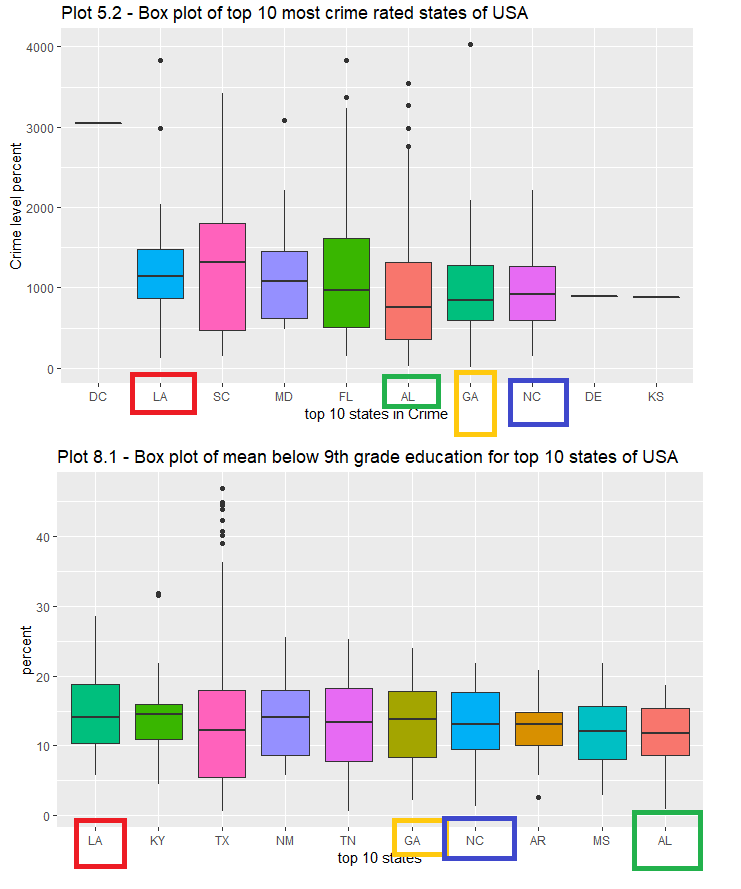


Figure : Comparison of top 10 crime rated states with top 10 less educated states

1. When we compared top 10 less educated (below 9th grade) states of USA with top 10 most Asian occupied states, we found no state was common.

**Methods and Techniques**

Question 5. A description of the methods/techniques/tools/algorithms you have/will use to complete the project.

We were able to find the most influential variables with respect to crime via EDA. We will be using k means algorithm to cluster the data with similar characteristics with respect to crime.

To choose the right model to develop, we will begin with descriptive statistics and correlation analysis to find out the comparison between the crime rates on a national level and crime rates on a state level based on ethnicity since our focus on the analysis is inclined towards finding out the association of crime rate with the variation in ethnicity and the effects of other factors like poverty, age, income and education on it.

The data grouping into clusters and initial analysis thus far will help us understand the probability of distributions of our dataset with respect to the variables of our interest, thereby letting us choose the best corelating variables to be included in developing our final model. We will be using an unsupervised machine learning algorithm, the k-means clustering method for partitioning the data into clusters based on ethnicity. We will then work with the several available k-means methods to calculate the cluster variations with the centroid repeatedly until the cluster assignment stays the same. Since we will be working with R, we can presume that the number of iterations will be twenty approximately in our analysis. We will finally evaluate the model developed to determine its effectiveness in supporting our hypothesis using plotting and confusion matrix to find its accuracy and efficiency.

**Challenges Encountered**

Question 6. A description of the challenges you have had working on the project so far

The interesting part of our analysis is dealing with the challenges encountered when trying to think of our objective and work towards it. Due to group effort, we were successfully able to overcome the following challenges.

* 1. When trying to make our exploratory data analysis plots much more meaningful, we decided to add maps with gradient scale in addition to our regular statistical plots such as the box, bar plots. We were learning about the ways in R to use ggplot to use the geom-polygon to plot the map of the united states. This was challenging since we had to use the state abbreviation data and merge it with our crime data set in order to use the state names to match the ones in a R package- fiftystater. Also, since our crime dataset had only 47 states, we added in states – Hawaii, Montana and Nebraska for plotting on map.
  2. With the percentage of population data for ethnicity, plotting with the percentage grouping under state names did not give us a meaningful insight. After several discussions, we came up with the idea to convert the population percentage into approximate actual population based on race using the overall population of the state. This helped us visualize data in a much more meaningful sense.

**Sections Completed**

Question 7. A discussion of the parts of the project that have been completed

On a high level, we have covered the 90% part of exploratory data analysis and shall start with data modelling as the next stage of our analysis.

Stage 1 consisted of loading the data into R studio, taking the first glimpse of it by running commands like summary(), str(), head() and glimpse() on it to get the nature of the dataset and first level of cleansing which included removing the special characters from the column names. The next step was deciding on the primary key of the dataset which we found by concatenating community name and state into a single column. Stage 2 is dedicated to reassigning datatypes and factorizing the attributes as required. Because the dataset has huge 148 attributes so we used sapply() to run the function on all the columns in one go. Stage 3 works on handling the missing values and NA’s by getting the mean, sum or any other aggregate function of a column by not considering the NA or missing values.

Stage 4-9 are all created to do distinctive tasks of exploratory data analysis. In exploratory data analysis, we formulated questions around our hypothesis to have an insight into the data by plotting it on bar charts, scatter plots, box plots and geological maps. Our main areas of analysis lie in understanding the association of different ethnicities populated in the various states of USA with the respective crime rate in that state and the impacts of the underlying factors like age, poverty level, income level and education on it.

**Sections Remaining**

Question 8. A discussion of the parts of the project that remain to be completed.

The parts that still needs to work upon include data modelling and model evaluation. For our analysis, we have decided to work with k-means clustering technique. In order to proceed with this machine learning algorithm, we will be following the remaining steps below to complete the analysis.

* 1. We will need to prepare our data for the cluster analysis
  2. We will be learning on understanding the technique and how to measure differences in our observations
  3. We will determine the number of sub groups (K) in which we will group our data for cluster analysis
  4. Evaluation the result will be done to compare the accuracy of the model which will finally identify the right number of clusters for grouping our data.

**About Project Report and Presentation**

Question 9. A discussion of how you will finish the final project report and presentation.

The project report will comprise of the complete introduction of the crime dataset followed by a step-by-step briefing of the procedure followed in the various stages of the project. The project report will give a complete sense of understanding on why and how we formulated our research question and the significance of choosing it specifically. It will include the findings and conclusions formed about the data from exploratory data phase and furthermore, the summary of data modeling performed by clustering the data using k-means clustering and conclusions by reading the summary of k-means using confusion matrix and plotting it. The summary report will enclose R markdown document file (rmd file) and MS word knit report generated by R for reference.

In the similar fashion, the presentation will include all the key points from our research and data analysis to be presented in a simple manner to our audience in order to make them understand our research from start to end.

References

<http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized>

<https://en.wikipedia.org/wiki/K-means_clustering>

<https://www.datascience.com/blog/k-means-clustering>

<https://searchdatamanagement.techtarget.com/definition/data-modeling>