Introduction

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.

The problem that we addressed in our project is the influence of immigration statistics on the crime occurring in that community. We analyzed the contrapuntal impact on factors like ethnicity, poverty, median income, age and education with the increase or decrease in the number of immigrants and its impact on 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 is 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 we analyzed is crime rate in an area with more immigrant population and its impact on the national average rate of violent and non-violent crimes. We analyzed 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.

Related Work

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 were 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 used available analysis in finding out additional data sources to expand our research and performed analysis.

Data

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. After loading the data into R studio, we took the first glimpse of the data by running commands like summary (), str (), head () and glimpse () on it to get the nature of the dataset, and performed cleansing which included, removing the special characters from the column names. The next step we decided was the primary key of the dataset which we found by concatenating community name and state into a single column. Next, we reassigned data types and factorized the attributes as required. Because the dataset has a large number of attributes (148), we used sapply () to run the function on all the columns. Finally, we worked on handling the missing values by getting the mean, sum or any other aggregate function of a column by not considering the missing values.

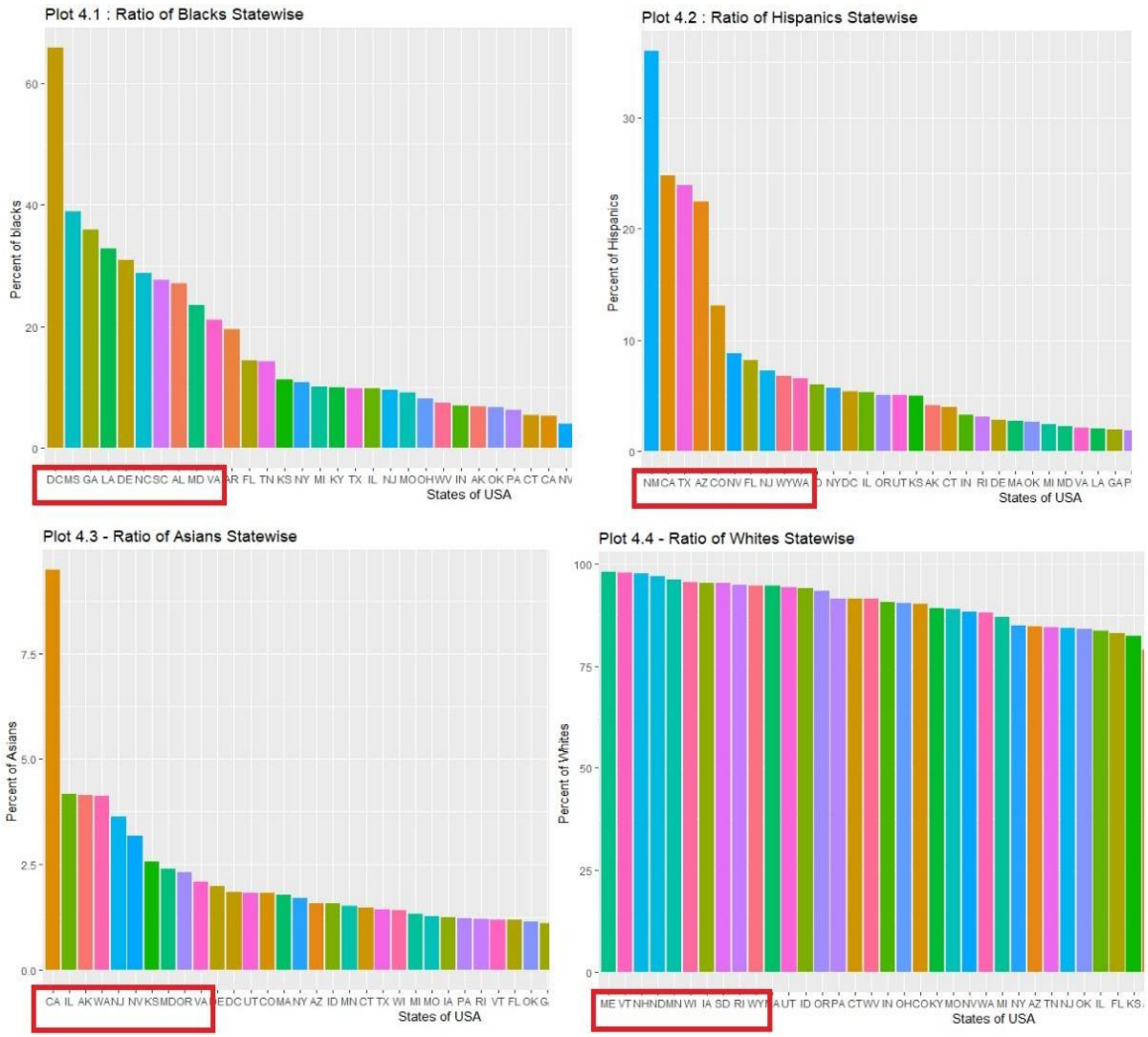
Technical Approach

We followed the generic process of analyzing data and the flow of our process is as below. The steps that we followed are – Preprocessing, Data Cleaning and exploratory data analysis.

1. Preprocessing –We transformed raw data into an understandable format that we used in the next steps.
2. Data Cleaning –We removed all missing values and checked for data types to use in our exploratory data analysis part.
3. Exploratory Data Analysis –We have used all the variables of our interest to see the trend patterns with crime. We also worked in expanding our analysis by including data from US census bureau to see trend patterns on a yearly basis to give our research an objective, a much clearer conclusion.
4. Models/Methods – We used the k-means clustering algorithm to cluster states based on the variables of our interest
5. Evaluation –We used the elbow method to evaluate our clusters and determined the optimal number of cluster to be used.

Exploratory Data Analysis

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



First, we averaged the racial population of each state for the four major ethnicities separately. Next, we ranked the state's population-wise, and derived the top 10 states for each ethnicity. Finally, we compared the crime rate of these states with national average of crime

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.

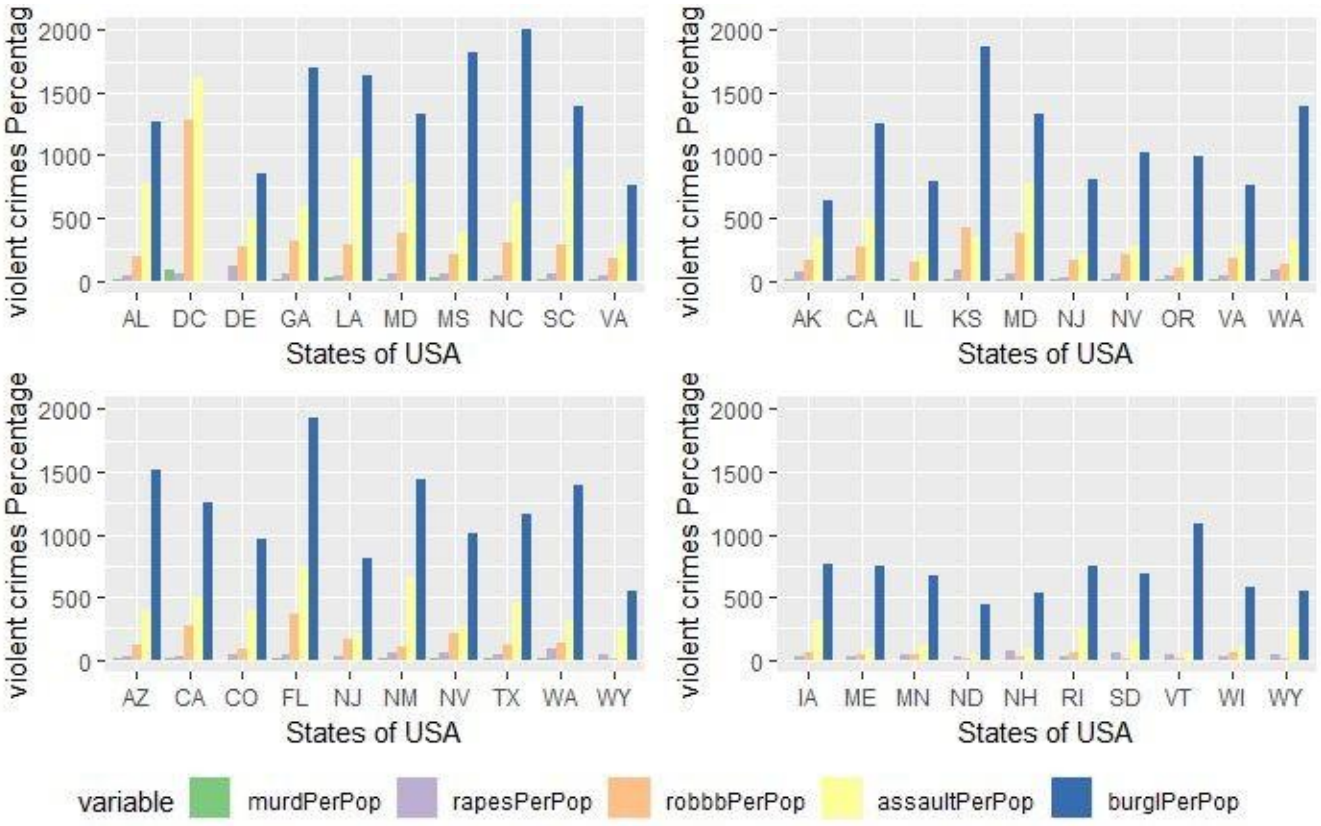


Figure 1: Bar plot showing the trend of crimes for ethnicities : Blacks, Asians, Whites and Hispanics (clockwise)

3. 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.

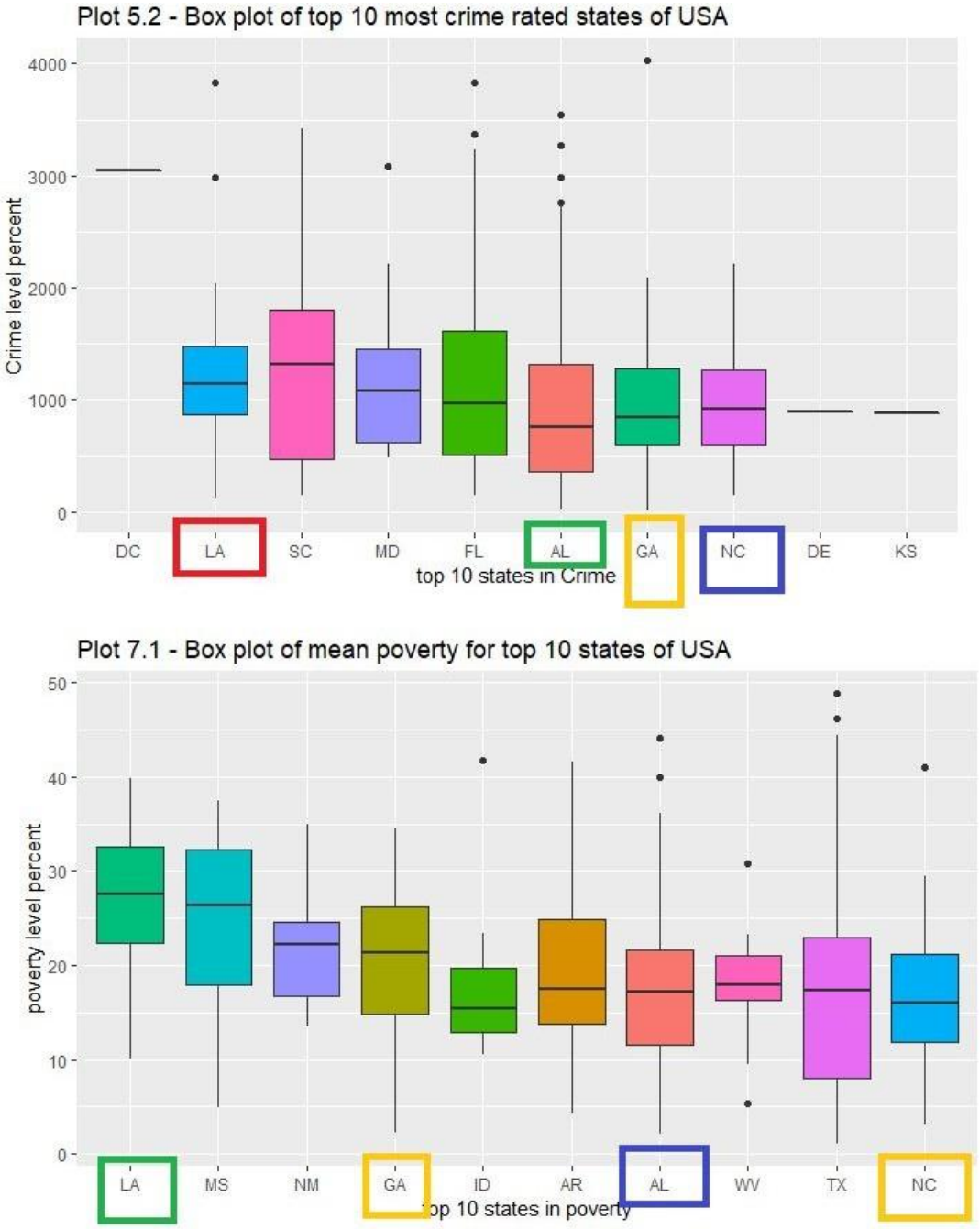


Figure 2: Comparison of top 10 crime rated states with top 10 poor states

4. When we compared top 10 poor states of USA with top 10 most Asian occupied states, we found no state was common

5. 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.

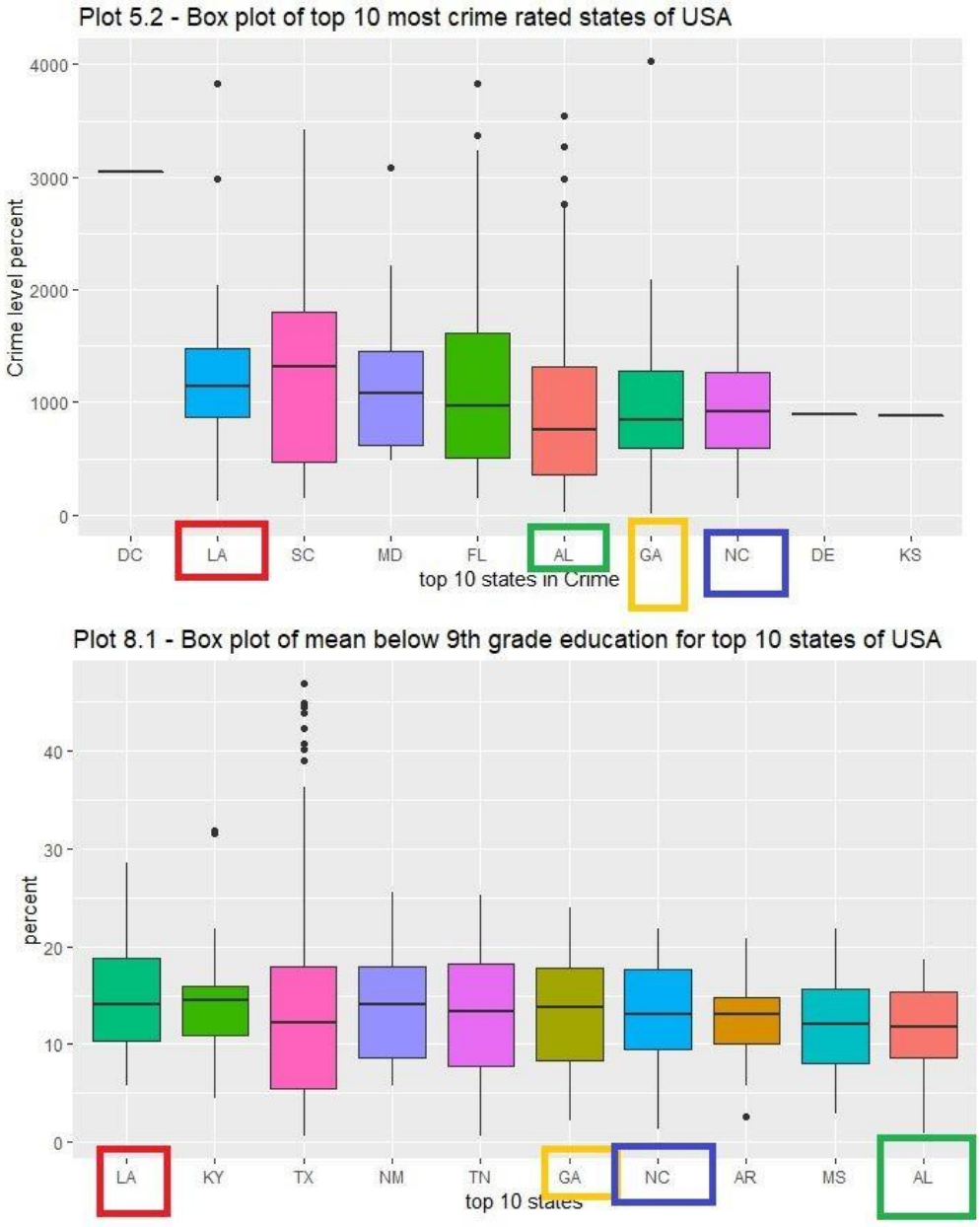


Figure 3: Comparison of top 10 crime rated states with top 10 less educated states

6. 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

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

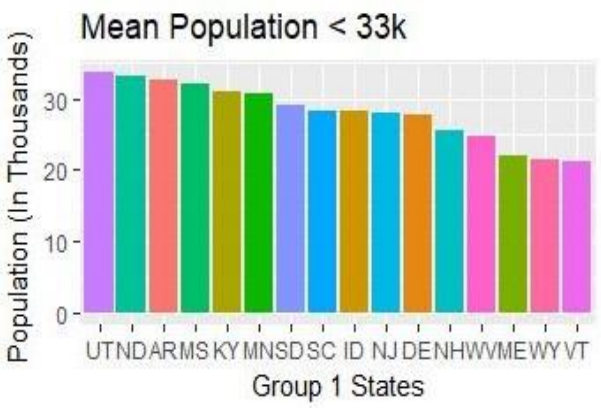
To choose the right model to develop, we began with descriptive statistics and correlation analysis to find out the comparison between the crime rates on a national level and crime rates on a state level based on ethnicity since our focus on the analysis is inclined towards finding out the association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

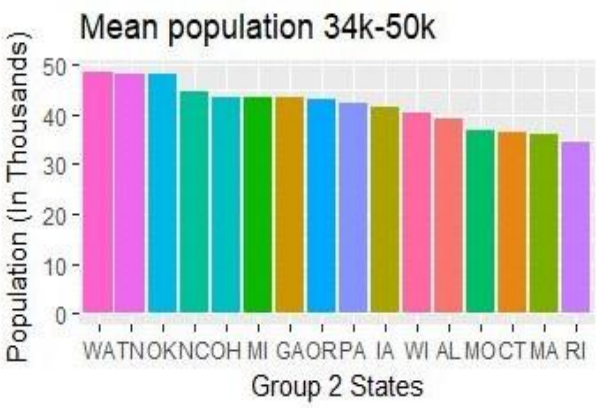
The data grouping into clusters and initial analysis helped us understand the probability of distributions of our dataset with respect to the variables of our interest, thereby letting us choose the best correlating variables to be included in developing our final model. We used an unsupervised machine learning algorithm, the k-means clustering method for partitioning the data into clusters based on ethnicity. We then worked with the several available k-means methods to calculate the cluster variations with the centroid repeatedly until the cluster assignment remained the same. We finally evaluated the model developed to determine its effectiveness in supporting our hypothesis using plotting and confusion matrix to find its accuracy and efficiency.

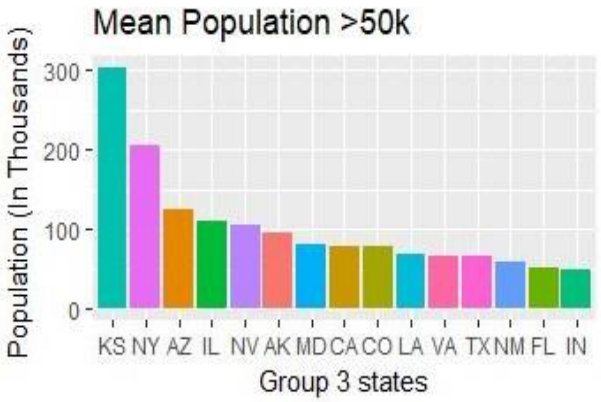
K – Means Clustering

For our analysis, we decided to work with k-means clustering technique. In order to proceed with this machine learning algorithm, we followed the steps below to complete the analysis.

Before performing k-means clustering on the entire data, for obtaining better clusters depending on our variables of interest, we grouped states into three based on mean population, namely, group 1 with states that have a mean population of less than 33,000, group 2 with states that have a mean population between 34,000 and 50,000, group 3 with states that have a population greater than 50,000.



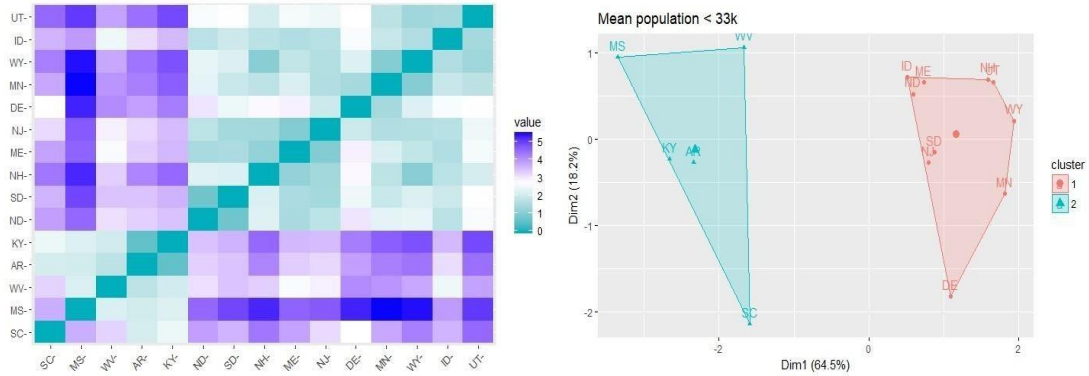




1. We learnt the technique and measured differences in our observations. We clustered states based on level of poverty, number of high school and middle school dropouts, number unemployed and overall violent crime rate on our prepared data using the K-means clustering technique.

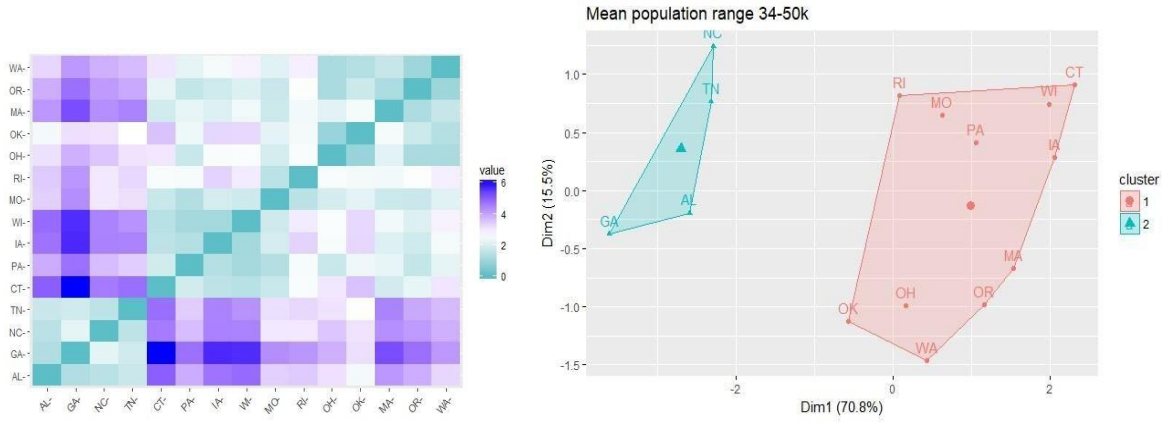
2. We determined the number of sub groups (K) in which we grouped our data for cluster analysis.

Clustering States with Mean Population < 33k:



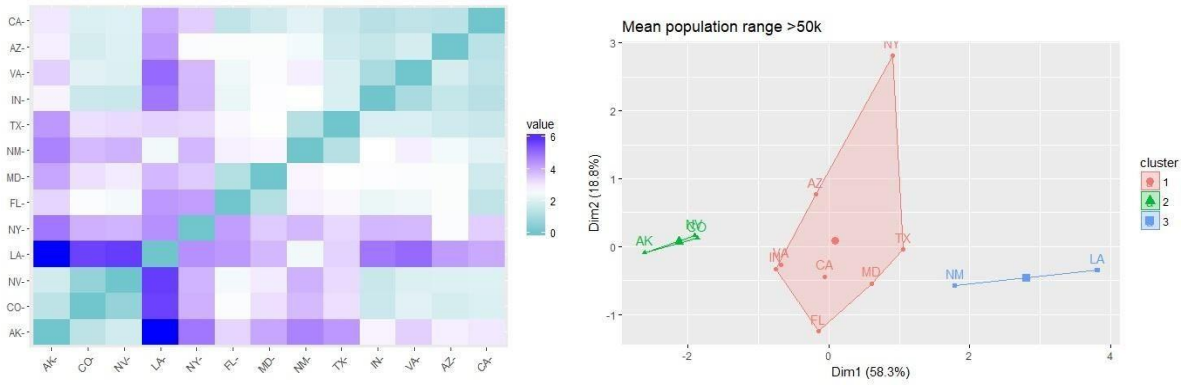
From the distance matrix above, the gradient scale highlighted three different dominant colors. Performing clustering with k input as 2/3 grouped the states of similar characteristics together. In performing k means clustering with k=2, Cluster 1 consisted of states MS, WV, KY, AR, SC that show similar trend. Whereas, cluster 2 consisted of states ID, ME, ND, NJ, SD, MN, WY, NH, UT, DE that show similar trend.

Clustering States with Mean Population between 33k-50k:



From the distance matrix above, and k=2/3, cluster 1 showed states GA, AL, TN, NC with similar trends. Cluster 2 showed states RI, MO, PA, WI, CT, IA, MA, OR, WA, OH, OK with similar trends.

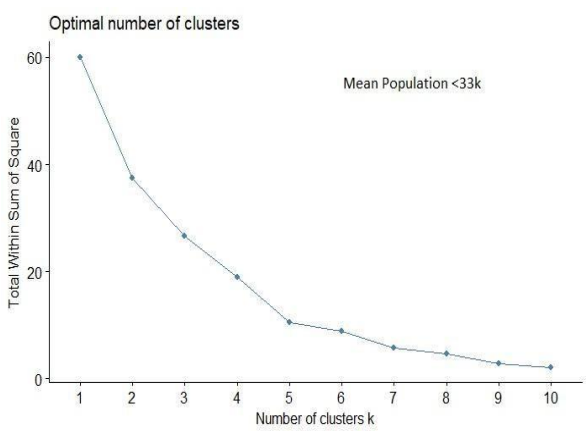
Clustering States with Mean Population between >50k:

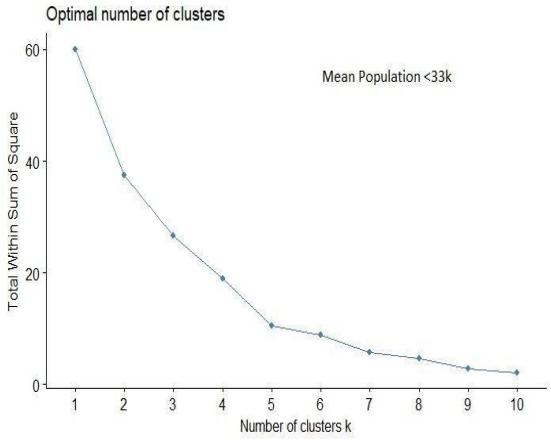


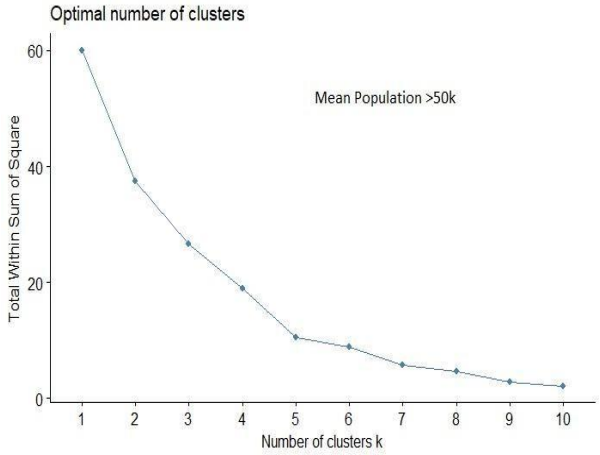
From the distance matrix above, the gradient scale still highlighted three different dominant colors. With k=2/3, cluster 1 showed states Ak, CO, NV with similar trends. Cluster 2 showed states IN, VA, AZ, NY, CA, FL, MD, TX with similar trends and cluster 3 showed states NM, LA with similar trends.

Test and Evaluation

Evaluation of the result was done to compare the accuracy of the clusters which finally helped us identify the right number of clusters for grouping our data. Using the elbow method, the below charts showed the optimal number of clusters as 5 for all the groups.

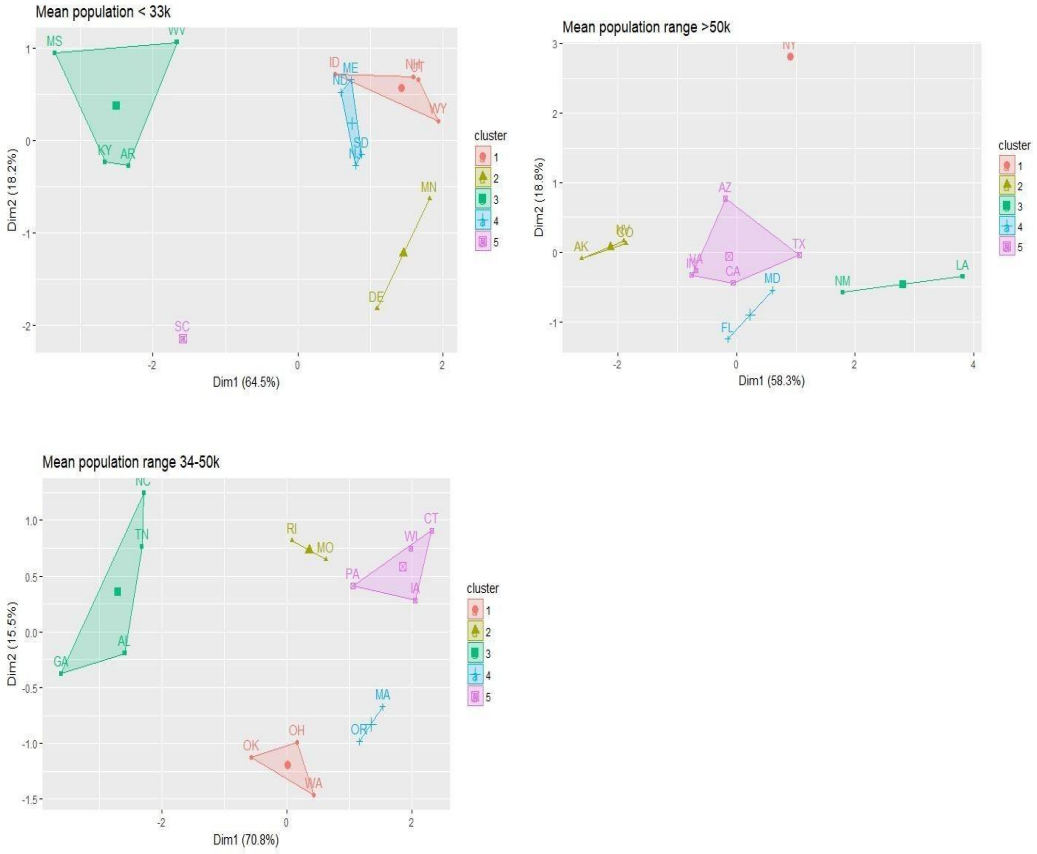






Updating K based on Optimal value:

From the elbow method, the states for all the groups were then clustered into 5 and the updated clusters group showed the sum of squared errors less than the previous clusters with k=3



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

When trying to make our exploratory data analysis plots much more meaningful, 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.

We also grouped data together for top 10 states and also based on population for clustering to help us take our approach to answering our research question in a much more sensible way.

Presentation

This final project report is to give a complete sense of understanding on why and how we formulated our research question and the significance of choosing it specifically. Our report includes the findings and conclusions formed about the data from exploratory data phase, the summary of data modeling performed by clustering the data using k-means clustering and conclusions. Also, enclosed along with the summary report is the R markdown document file (RMD file) and MS word knit report generated by R for reference. We presented our data analysis to our audience in a simple fashion which included all the key points of our research.

Conclusion

It is fair to assume that crime rates are high in places where the poverty rate is higher and literacy rate is lower. Our explanatory data analysis on the crime dataset helped us justify this assumption.

Poverty, Education, Ethnicity and Employment were the key variables on which we based our analysis. After performing EDA on the crime dataset, we hence conclude that crime rate is directly proportional to poverty rate and inversely proportional to literacy rate. From our analysis of top 10 states in terms of poverty, literacy, demographics and crime rate -- Louisiana, Alabama, Georgia and North Carolina are few states with highest crime rate. These states also have lower literacy rate and higher poverty rate. Lastly, states with highest Asian population seems to have higher literacy rate, lower poverty rate and thus in turn lower crime rates.

We performed K-means clustering algorithm to help us validate the variables for performing analysis on the crime data set. We used mean population across several states to form

group of states with similar attributes in terms of ethnicity and crime rate across United States. Results from this analysis would help new immigrants make key decisions in terms of location based on crime rate and ethnicity preference when they move to United States.

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