FBI UCR Crime Data Analysis: Clustering and Regression

Scout Oatman-Gaitan
Data Analytics – ITWS 6600
Rensselaer Polytechnic Institute, Troy, NY

Abstract & Introduction

The goal of my project is to understand crime trends in the United States both at a national level and also at a state level. I will explore these rates and crime totals over time and see how the data is changing to try to gain some insight on which types of crimes are committed most in the United States.
At the state level, I want to understand what areas are more prone to crime and what areas aren't. I want to explore where states rank in crime totals versus crime rates. Obviously states with high populations like CA, NY, and TX will have high crime totals because they have more people.

Looking at crime rates will give us better insight into what states have 'worse' crime because we can see that scaled data. For these purposes I plan on developing both a clustering model and a regression model.

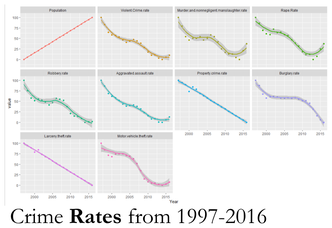
Data Description

I used data gathered by the FBI, specifically the Uniform Crime Reporting Program (UCR). I used two forms of this data.
- First, I used data at a national level that reports the year (1997-2016), population, crime rates for various categories, and crime totals for various categories. Note that crime rates are per 100,000 inhabitants.
- Second, I used a bigger dataset that included a breakdown by state. This dataset included information for each state from 1997-2014 and the same crime rates and totals by category. The first dataset was available in full from FBI UCR but the second dataset required joining for 50 tables. Both tables included a "legacy definition rape" category and a "revised definition rape" category. The revised definition was composed of mostly null values so I ignored this category. I had planned to also look at a county level but struggled to find good data for this.

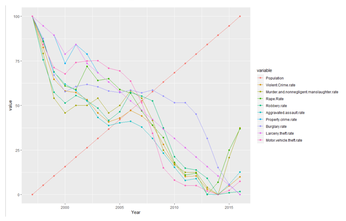
In addition to this main data, I also utilized a dataset from the FCC with the corresponding state Federal Information Processing System (FIPS) codes. This would help with geographical visualizations.

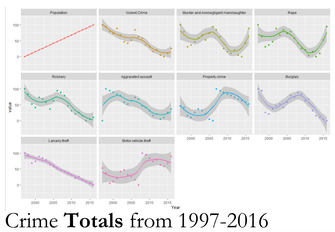
Time-Based Analysis

Major Takeaways: Every year the population is steadily increasing while every crime rate category is decreasing. Some categories are seeing an uptick in recent years. Crime rates have less variation than crime totals. Just looking at totals doesn't account for the increased population.



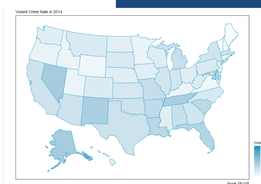
Crime **Rates** from 1997-2016



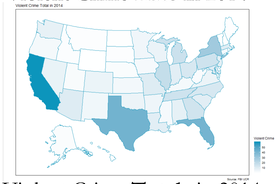


Crime **Totals** from 1997-2016

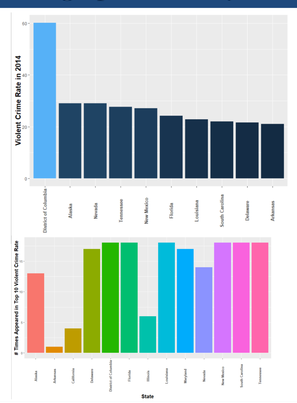
Geographical Analysis

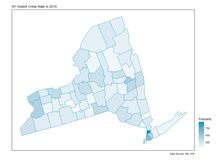


Violent Crime **Rate** in 2014



Violent Crime **Totals** in 2014

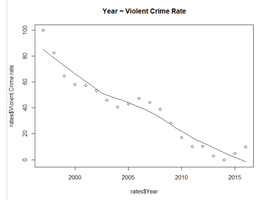




Major Takeaways: Rates are more telling than totals so we will use them. DC has very high rates and has every year since 1997. Other troubled states that have been in the top ten every year are FL, LA, NM, SC, and TN.

Regression Modeling

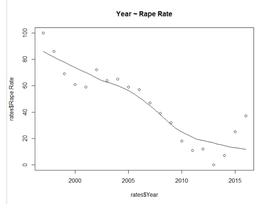
Violent Crime Rate Regression



Actual	Predicted
100	99.952
57.298	57.404
0.00	0.067
9.904	9.873

From the chart to the left we can see that the violent crime rate is decreasing with time. From the boxplot below we can see that there are no outliers. The correlation between violent crime rate and population is -0.954. I fit a linear regression model to the data and got a strong R^2 (rounded to 1), high f-statistic (2.59e+06) and low p-value (p<2.2e^-16). The strongest correlated variables are rape rate, robbery rate, and aggravated assault rate. After splitting the data, the model shows the same significant variables. The model is strong with a min-max accuracy of 74.86% and a MAPE Score: 8.49%

Rape Rate Regression



Actual	Predicted
100	99.952
57.298	57.404
0.00	0.067
9.904	9.873

My rape rate model ended up being similar to the violent crime rate model because of how correlated they are. Again, rape rate is decreasing and there are no outliers. The rape rate model has the same significant variables – violent crime, robbery, and aggravated assault. This model had a high R^2 (0.9997), a high f-statistic (4216), a low p-value (p<2.2e^-16). The model had a min-max accuracy of 92.11% and a MAPE score of 7.90%. These are better than the model above.

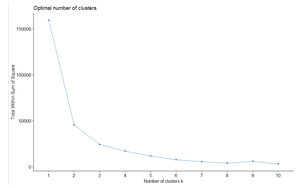
References

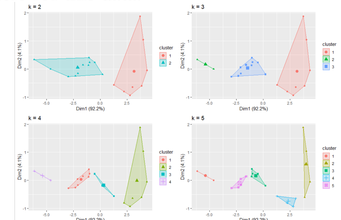
Data: FBI UCR:
https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/topic-pages/tables/table-1
State and County FIPS Code information:
https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt

Cluster Modeling

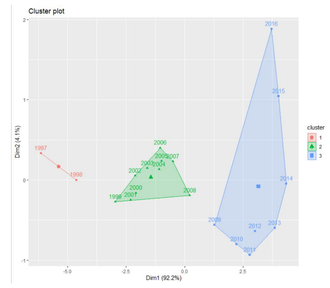
Clustered National-Level Data

Here, we can see the decision to use three clusters for this data:



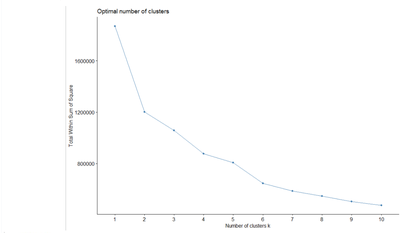


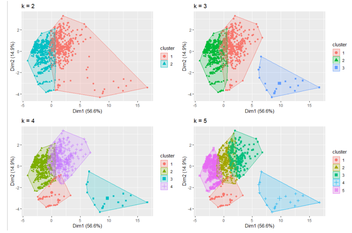
You can see that the years read chronologically from left to right. Cluster 1 has 1997-1998, Cluster 2 has 1998-2008, and Cluster 3 2009-2016. From left to right, these clusters are defined by an increasing population and decreasing crime rates. Most notably, vehicle theft and assault have decreased the most while population is increasing quite significantly.

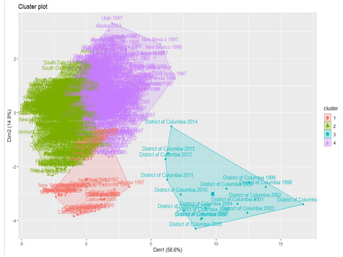


Clustered State-Level Data

Here, we can see the decision to use four clusters for this data:







Cluster 3 is composed entirely of DC from 1997-2014 and is defined by high rates in every category. Cluster 1 is composed of NY, CA, and TX is defined by high population. Cluster 2 and 4 are similar though 4 has higher rates and population whereas 2 has lower rates and populations. Cluster 3, DC, has the highest crime rates in every category but rape and burglary and it still comes in a close second.

Conclusions & Future Works

Conclusions

- ➔ Population is rising and crime rates are decreasing over time
- ➔ Washington DC is a very troubled areas with outlandishly high crime rates. FL, LA, NM, SC, and TN have been in the top 10 for violent crime rates every year since 1997.
- ➔ Violent crime rate, rape rate, robbery rate, and aggravated assault rates are all highly correlated

Future Work

- ➔ Get more access to county level data or crime categories and time to understand crime distribution in troubled areas and highly crime difference in urban and rural areas
- ➔ Drill down the data in other ways to answer more targeted questions. We can look at a specific crime category in a specific area at a specific time. Build a dashboard that allows a user to drill down how they see fit