USA Crime Demographics Study:
Violent Versus Non-Violent Case Study

The analysis was undertaken in a Jupyter notebook which can be found in the refences section.

Introduction

There are many things that contribute to the motive of why a person commits a crime. There have been studies exploring personal factors including personality types and mental health (Mundia 2016). This report looks at demographic factors across the USA and seeks to investigate if there are links to different types of crime.
Understanding this relationship is critical for using the police resources effectively and efficiently within each county's jurisdiction.

1 Data, Research Questions and Analytical Approach

1.1 Data

There are multiple sources of crime data for the USA, but many only give detail at a state level.
For this investigation four datasets were used that give county level data:

1. USA Demographic Data 2015
2. USA Crime Data 2015
3. Shape file USA
4. Regional breakdown of the USA

2015 was selected for further investigation because there is data at a county level.
The sources of the datasets are given in the references section below.

The crime dataset gives records of all crimes committed across the USA during 2015 and categorises them. For the purpose of this report, crimes are classed as either violent or non-violent. This is shown in Table 1 below:

Table 1

Violent Crimes	Non-Violent Crimes
Murder	Burglary
Rape	Larceny
Robbery	Motor Vehicle Theft
Aggravated Assault	Arson

The crime data, demographic and shape files will be merged together by using the county FIPS code (area code).

Dona Anna county data was missing from the Crime Data. This was added to the dataset from another source: https://catalog.data.gov/dataset?tags=crime.

Alaska was not included in the demographic dataset so was removed from this study.

1.2 Research Questions

When the combined data for violent and non-violent crimes are shown visually on a map (figures 1 & 2), human reasoning can identify there are patterns across the USA.

Figure 1 - log scaled volume of Violent Crime in USA by County

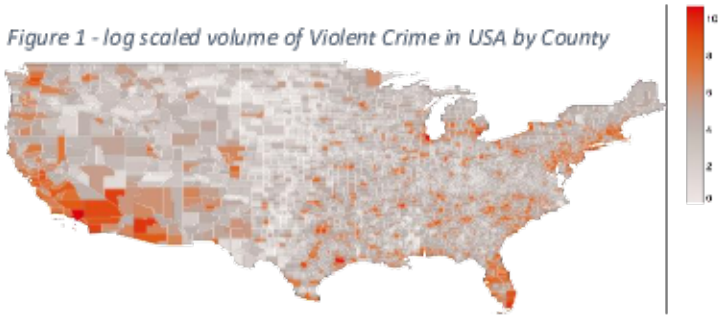
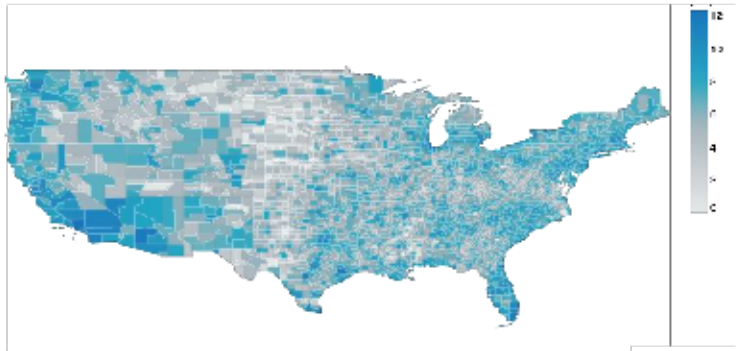


Figure 2 - log scaled volume of Non-Violent Crime in the USA by County



Figures 1 & 2 appear to illustrate that there are high levels of crime on both coasts and the major cites of the USA. It is with these patterns in mind that the report derived the following research questions to help define the scope of the analysis.

1. Ascertain the key demographic features that contribute to violent and non-violent crime and explore whether they are a substantial difference in demographics between the two?
2. Investigate where the importance of these key features changes geographically across the USA?

The results aim to understand the patterns seen in figures 1 and 2 and can then be used to implement targeted counter crime strategies and allocate police resources effectively and efficiently.

1.3 Analytical Approach

To investigate the research questions the following plan was derived. While it may be laid out in a linear sequence in this report, many of the steps were revisited and serval iterations were taken to find the most comprehensive results.

1. Upload datasets and merge on FIPS code
2. Clean data: reshape, impute missing values, handle outliers
3. Data transformations
4. Feature engineering
5. Investigate scatter plots relationships
6. Correlations between features
7. Regression of key features to quantify the relationships found
8. Perform clustering to classify counties together across the USA
9. Repeat 6-7 on the different clusters found in step 8 to see if there are differences in feature importance across the USA

2 Findings and Critical Reflection

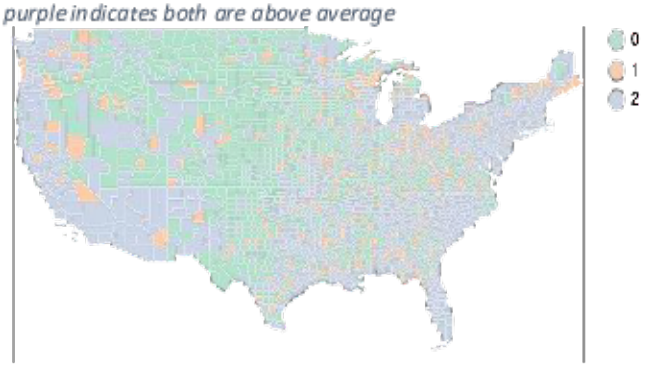
2.1 Findings

2.1.1 Distributions of high crime areas across the USA

The first simple analysis performed was to look for areas with above average rates of both violent and non-violent crime. This can be seen in Figure i & ii in appendix 1. They show that areas with high levels of crimes are located near the main cites and the coasts.

However, when looking at the differences between the figures, it is visible that there are not many counties (coloured in orange) that have high violent crime and low non-violent crime or vice versa (figure 3). This is the first indication that there are similarities between the demographics of both types of crime. It is also noted there is no obvious pattern in the location of these differences.

Figure 3 - areas of high or low crime. Green indicates areas that have violent and non-violent crime below average, orange indicates areas where one of the types of crime is above average, purple indicates both are above average



2.1.2 Scatter plot analysis

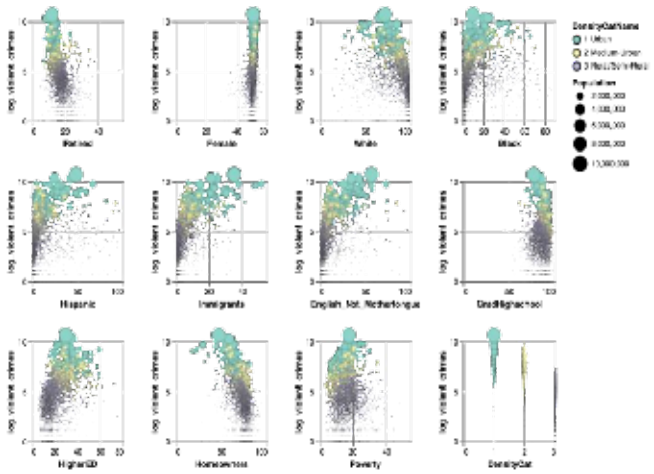
The initial features selected for investigation are seen in table 2.

Table 2

Variable	Description
Female	% of the population that are Female
White	% of the population that are White ethnicity
Black	% of the population that are Black ethnicity
Hispanic	% of the population that are Hispanic ethnicity
Immigrant	% of the population that are immigrants
English not mother tongue	% of the population that don't speak English as their main language
Graduated High School	% of the population that graduated high school
Higher Educated	% of the population that have higher education degrees
Homeowners	% of the population that own a home
Poverty	% of the population that live under the poverty line
Density Category	Three categories of how densely populated a county is, Rural, Suburban and Urban

Below is Figure 4 showing the scatter plots against Violent Crime (Non-Violent Crime plots can be seen in Appendix 2).

Figure 4 - Scatter plots of violent crime versus independent variables



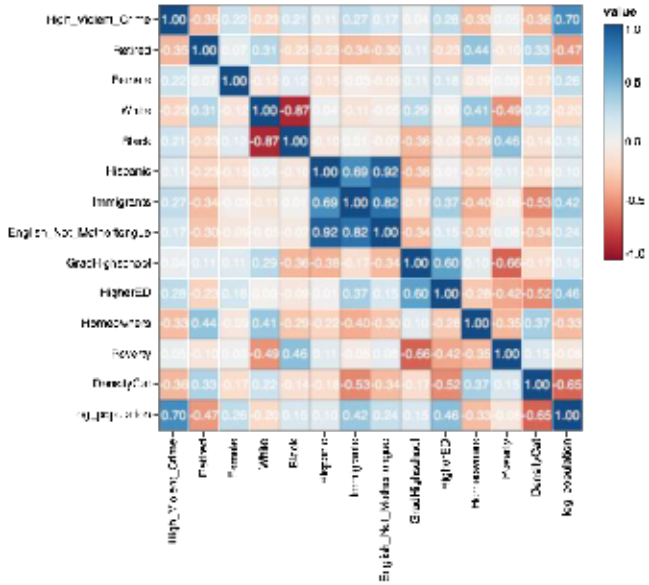
Firstly, there was negligible different between the plots of violent and non-violent crime.
White and Homeowners can be seen to have negative relationship with violent crime, while Black, Hispanic, Immigrants and English Not Mother Tongue have positive relationships.
The colour of the density categories makes it obvious there is a relationship here to investigate further.
The population sizes of counties are striking, it highlights that all counites with a large population are areas with higher crime rates. Therefore, it was concluded that population will have to be added as a feature for the ongoing analysis.

The rest of this report will focus solely on violent crime because of the different between violent and non-violent are negligible.

2.1.3 Correlation matrix

A Pearson's correlation plot can be seen in Figure 5.

Figure 5 - Correlation matrix between all features



This correlation matrix, used alongside the scatter plots, is used to guide the next part of the analysis. A reason for each selection is included in the commentary on the Jupyter notebook.

Selected features:

- Density category
- Retired
- White
- Immigrants
- Higher education
- Homeowners
- Log of population

2.1.4 Regression Analysis

Logistic regression was carried out to inspect the coefficients of the key features as a tool to determine feature importance. This is done by looking at how the features predict if a county has a higher than average level of violent crime.

Figure 6 - Confusion matrix

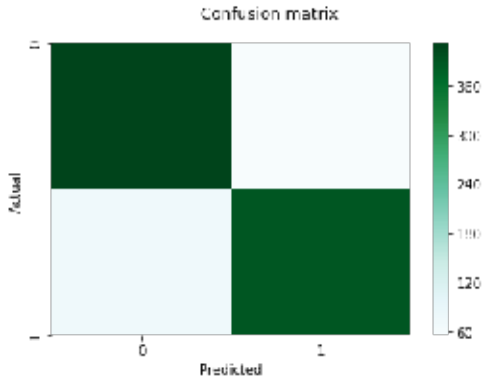
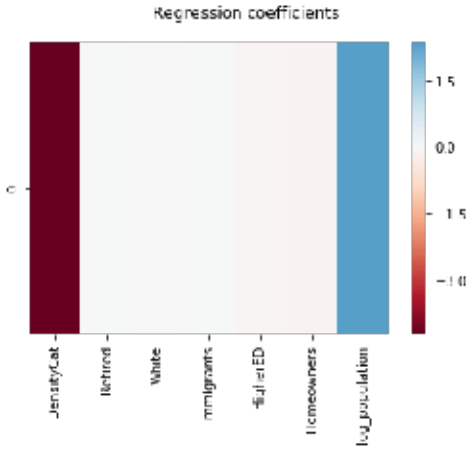


Table 3

	Precision	Recall	F1-Score
0	*0.84*	*0.88*	*0.86*
1	*0.87*	*0.83*	*0.85*
Weighted Avg	*0.86*	*0.86*	*0.86*

Accuracy – 86%

Figure 7 - Regression coefficient Importance



All the performance statistics show encouraging results with every metric being 86%. These results compared with class balance shown in Appendix 3 of approximately 50% is impressive. This shows that if needed the model could be additionally used as a prediction tool.

The coefficients show that the most important features in the model are density category and population size. While they are the most important without the other five features the results of the regression dropped dramatically. Hence, they are still deemed important to high levels of violent crime but to a lesser extent.

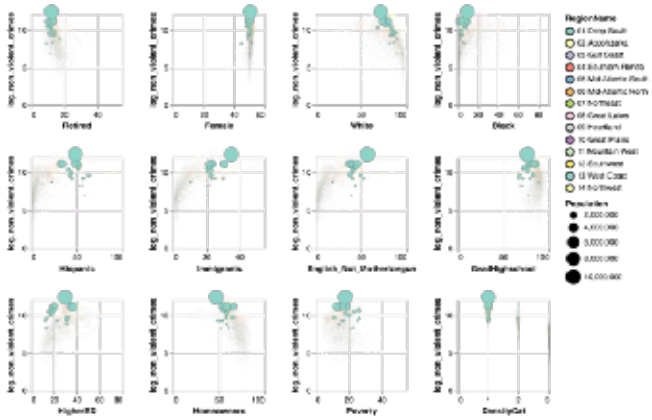
2.1.5 Regions differences

Regions referred to in this report are shown on a map in Appendix 4.

To inspect the merit of analysis into regional differences in feature importance scatter plots were used. but this time with the ability to look at different regions of the USA.

It was obvious that there was some regional variation in relationships of certain features and level of violent crime. Therefore, further investigation is required to assess the extent. This can be seen using the West Coast as an example in figure 8.

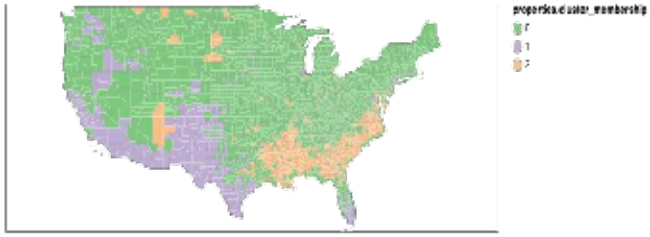
Figure 8 - Scatter plots for the West Coast



K- Means clustering was undertaken to classify similar counties together into clusters before looking at feature importance of each cluster. The use of three clusters had the best results, this selection process is explained in the Jupyter Notebook. Figure 9 shows the clusters on a map. This is a positive result as counties grouped together to divide up the USA. The clusters

appear to be combinations of serval regions highlighted by the scatter plots as having different correlations. The West Coast and South West form one cluster and the Deep South and Mid Atlantic South forming another.

Figure 9 – Clusters



When steps 4-5 from the plan were repeated on the clusters. It showed there was indeed different feature importance to each region. Different features were selected to perform the logistic regression on each cluster. The Jupyter notebook describes this decision process. Table 4 shows the features selected for each cluster.

Table 4

Cluster 0	Cluster 1	Cluster 2
Density Category	Immigrants	English not Mother Tongue
Retired	Retired	Retired
White	White	White
Female	Female	Female
Higher Education	Graduated High School	Poverty
Homeowners	Homeowners	Log of Population
Log of Population	Log of Population	

Results of regression can be seen in Appendix 5. The results are positive and show that the features selected in table 4 are the most important for each cluster. Therefore, concluding that there are regional differences in the factors that results in high violent crime counties, although population is always an important factor.

2.2 Critical Reflections

Clustering could have been improved by transforming the data to Geo-Weighted statistics before doing the clustering or by using a different clustering technique like Algometric clustering.

There were slight class imbalances when performing the regional regression and obviously there are different numbers of counties in each cluster. This is due to there being no set number of counties assigned to a cluster. Therefore, not all performance statistics should be compared to the results of the regression from the USA as a whole.

An alternative modelling method like Random Forests may have leaded to improved results but the results from the models had satisfactory results so this was not deemed necessary.

4 Refences

[1] Jupyter Notebook: file:///Users/sebkirk/Downloads/SebastienKirk_Notebook_PDS.html

[2] Contributions of sociodemographic factors to criminal behavior: Lawrence Mundia, Rohani Matzin, Salwa Mahalle, Malai Hayati Hamid, Ratna Suriani OsmanPsychol Res Behav Manag. 2016; 9: 147–156. Published online 2016 Jun 22. doi: 10.2147/PRBM.S95270

[3] Demographic Data and shape files data: kaggle.com/benhamner/2016-us-election

[4] Crime Data: https://www.kaggle.com/mikejohnsonjr/united-states-crime-rates-by-county

[5] Dona anna data: https://catalog.data.gov/dataset?tags=crime

[6] Regions of the USA: https://jeremyposadas.org/regions/

5 Word count

Introduction – 115 words

Data, Research Questions and Analytical Approach – 385

Findings and Critical Refelections – 938 words

6 Appendices

Appendix 1

Figure i - High violent crime counties

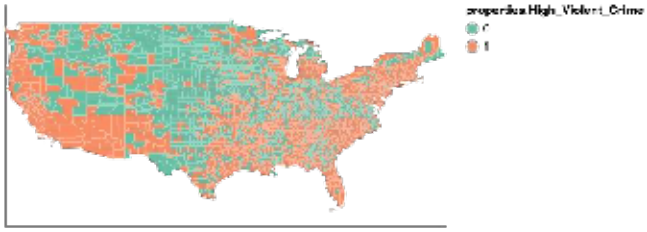
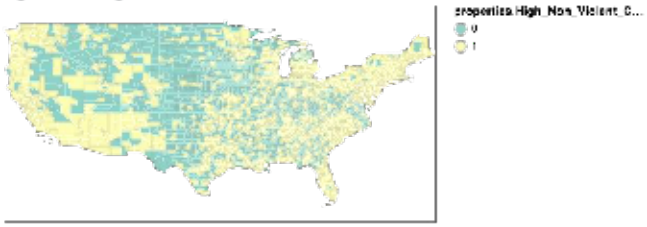
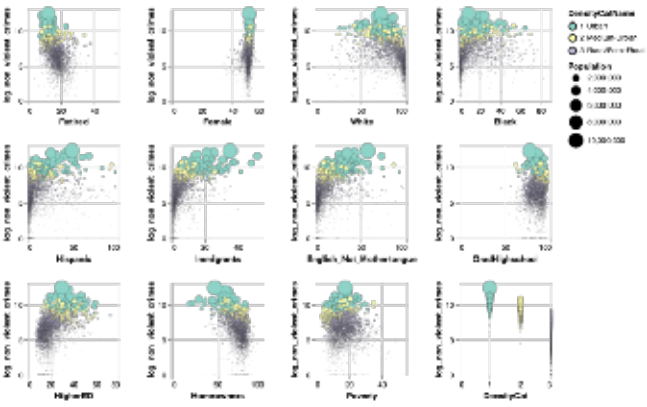


Figure ii - High non-violent crime counties



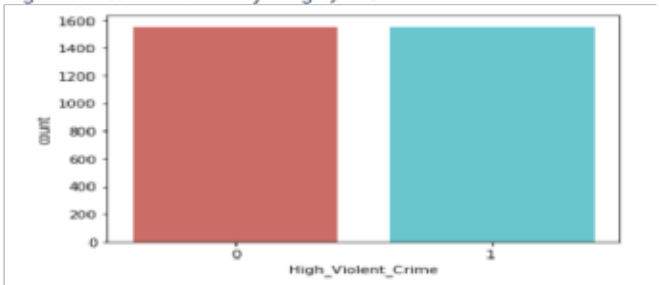
Appendix 2

Figure iii - Scatter plots of non-violent crime Vs independent variables



Appendix 3

Figure iv - Class balance of roughly 50%

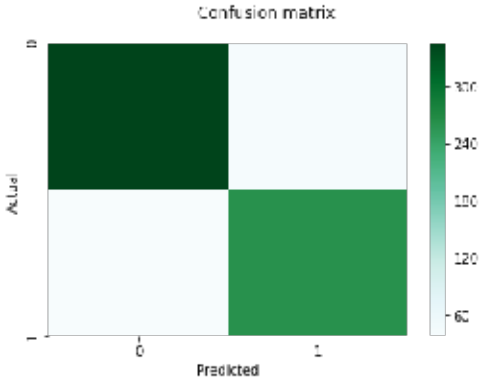


Appendix 4

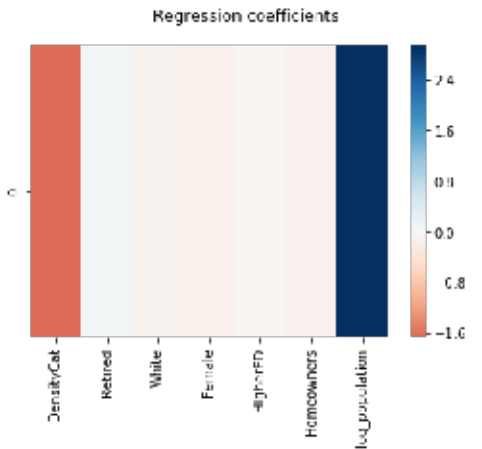
Figure v - regions of USA

Appendix 5

Cluster 0 results



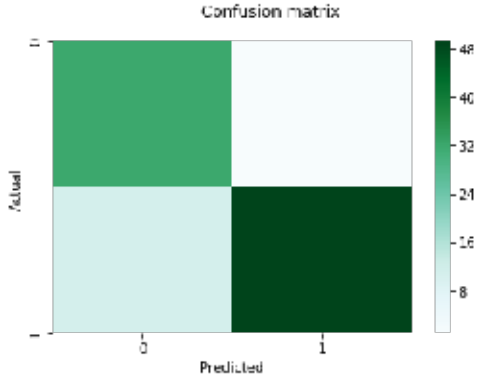
	precision	recall	f1-score	support
0	0.90	0.89	0.90	387
1	0.86	0.87	0.87	300
accuracy			0.88	687
macro avg	0.88	0.88	0.88	687
weighted avg	0.88	0.88	0.88	687



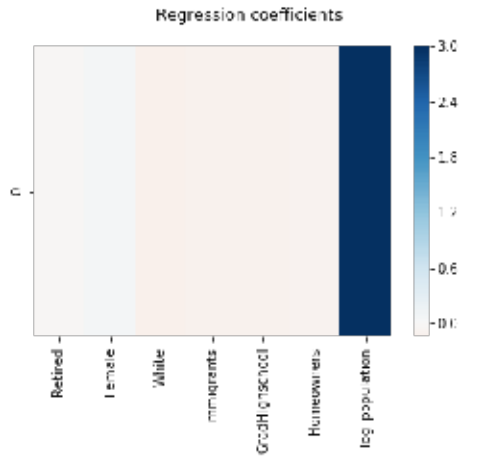
	DensityCat	Retired	White	Female	HigherED	Homeowners
0	-1.664307	0.074675	-0.090053	-0.14887	-0.069019	-0.127315

	log_population
0	2.921081

Cluster 1 results



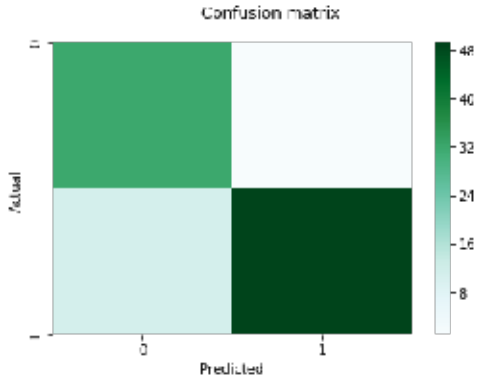
	precision	recall	f1-score	support
0	0.74	0.97	0.84	33
1	0.98	0.82	0.89	60
accuracy			0.87	93
macro avg	0.86	0.89	0.87	93
weighted avg	0.90	0.87	0.87	93



	Retired	Female	White	Immigrants	GradHighschool	Homeowners
0	-0.027316	0.055302	-0.150057	-0.119834	-0.130099	-0.108827

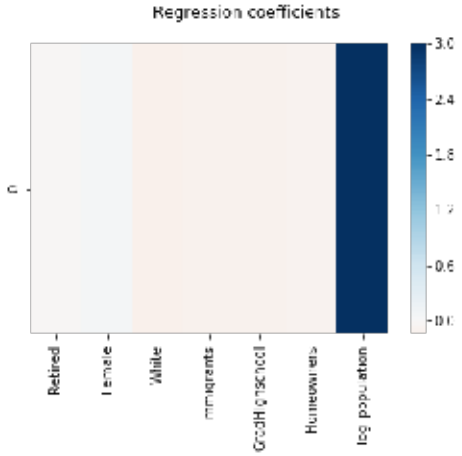
	log_population
0	3.006344

Cluster 2 results



	precision	recall	f1-score	support
0	0.74	0.97	0.84	33
1	0.98	0.82	0.89	60
accuracy			0.87	93

macro avg	0.86	0.89	0.87	93
weighted avg	0.90	0.87	0.87	93



	Retired	Female	White	Immigrants	GradHighschool	Homeowners
0	-0.027316	0.055302	-0.150057	-0.119834	-0.130099	-0.108827

	log_population
0	3.006344