INVESTIGATING IMMIGRATION AND CRIME

William Davis, Emily Foley, Madeline Baillie,

Kaila DeChristofaro, Megan Long, and Dr. Dil Singhabahu

Mathematics and Statistics Department  
Slippery Rock University

Slippery Rock, PA, 16057

*Objective* — The purpose of this study is to see whether immigrants affect crime by state, looking specifically at ones with traditionally higher immigration or crime rates. With the states separately and as a whole, we want to look at other factors that could contribute to crime rates and explore forecasting crime rate via other factors.

# Introduction

With recent political capitalization about immigration within the United States being broadcasted in the media, there is much debate as to whether an increase in the amount of immigrants causes an increase in the country’s crime levels. While it has become a popular topic of moral debate, we can investigate this using statistical methods and analysis.

# Previous Work

With poverty in the United States being one of the most enduring explanations for crime, recent studies have been finding that immigrants have lower levels of criminal involvement than their native-born counterparts. Christa Polzcynski’s report, “Immigration and Violent Crime: Citizenship Status and Social Disorganization,” analyzes the impact of immigration on homicide and other violent crimes in Orange County, Orlando, Florida. It was found that percentages of people living below poverty level and of female headed households positively influenced homicide rate, and these factors are also significant predictors of robbery. There was no support found for the idea that immigrant concentration impacts violent crimes of Orange County.

A similar study done by Scott Atkins in Austin, Texas studied immigration effects on homicide rates. From 1980 to 2000 Austin, TX experienced a 580% increase in immigrant population making it a favorable location for testing. Atkins set the dependant variable to be the homicide rate per 10,000 people and created a stability index before performing tests. Using mainly poisson and negative binomial estimation, Atkins concluded that there was no significant increase in homicide rates as a result of immigration.

Lastly, we researched a study conducted by Graham Ousey and Charis Kubrin that further looked into the topic of immigration and crime correlation. Upon completion of several statistical tests, they found an r-value of -.031 and a p-value of .032. From this they were able to say that immigration had a slightly negative correlation, but were not able to definitively conclude that immigration rates affect crime rates.

# Methodology

3.1 Datasets

The datasets were acquired from 4 different accredited United States platforms. The Department of Homeland Security is the source of the “dependent variable” of our investigation. The Department of Homeland Security records the number of immigrants per year by the state than they plan to reside in.

The United States Census is the source of the poverty data used. The data is within the Census’s own study of Historical Poverty from 1959 to 2017.

The United States Bureau of Labor Statistics was the source of the workforce data. From the U.S. Bureau of Labor Statistics, employment rate, unemployment rate, population in the labor force, and overall population were all used when investigating immigration.

The United States Department of Justice, specifically their Uniform Crime Reporting database, is the source of all of the data related to crime rates. Both general statistics, i.e. Violent Crime Rate and Property Crime Rate, and more specific statistics, i.e. Murder Rate, Aggravated Assault, Larceny Theft Rate, etc. were sourced from the Department of Justice.

A list of all attributes considered follows:

Population, Violent Crime, Murder and Manslaughter, Rape, Robbery, Aggravated Assault, Property Crime, Burglary, Larceny Theft, Motor Vehicle Theft, Labor Force, Employment Rate, Unemployment Rate, Poverty Rate, Admitted Immigrant Rate.

All of these attributes were collected at a state level yearly, that is each record for each attribute is associated with a state and a year.

3.2.1 Tools

All methods spoken of in Section 3 were done using Python. Python libraries used include:

Scikit-learn: a popular machine learning library built on NumPy, SciPy, and matplotlib. Sub-libraries used include cluster, decomposition, and manifold.

Seaborn: a statistical data visualization library.

Pandas: a data analysis library with multiple data structure options

Matplotlib: a 2D plotting library

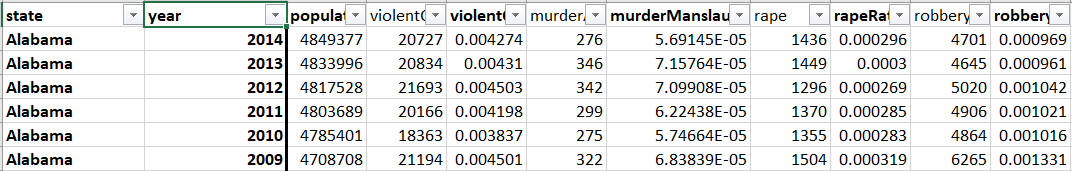
StatsModels: A statistics centered library for modeling purposes

3.2.2 Preprocessing

The data from each of the four previously mentioned sources was compiled into one aggregated excel document. Since the number of murders in Alabama is vastly uncomparable to the number of murders in California due to the differences in population. The raw number of each attribute was divided by the population of that state during that year to create a rate for each attribute. The only variable that this was not applied to was population itself.

Similarly to how some of the original sources were formatted, the final aggregated data was formatted into a multi-index table where each record is represented a State-Year combination. Next, it would be extremely beneficial to be able to investigate the relationship between attributes at a state-level instead of just a country level. In order to accommodate for this, the entire excel sheet was separated into 50 sub-sheets where each sheet was only comprised of an individual state. The final document has both the initial aggregated data, as well as the state-separated data leaving a total of 51 sheets.

This sheet is now read for upload into Python for analysis. Lastly, simply for ease of access, once this sheet was uploaded into Python, all of the columns that represent raw numbers, instead of rates, were dropped and thus not considered at all in the analysis process. Figure 1 depicts a sample of the final sheet imported into Python.



**Figure 1 - Final Data Excel Sheet**

3.3 Analysis Techniques

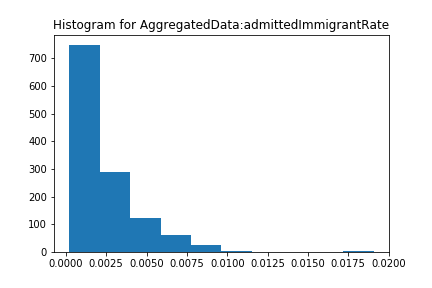
Many techniques were considered or used in the process of analyzation. Initially, descriptive statistics were used to obtain a better understanding of the overall data. The descriptive analysis did bring things to light that slightly changed the course of the analysis later down the road, and thus proved to be extremely beneficial.

After this, correlation matrices, simple regression, multiple regression, and clustering were all implement in order to understand the relationship between immigration and standard of living factors.

3.3.1 Pre-Transformation Descriptive Statistics

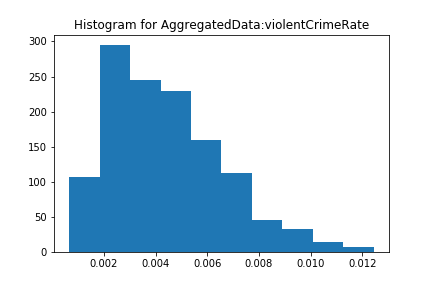
Upon initial investigation of some of the histograms of the variables which were being considered, it was clear that the majority of the data was not normally distributed, rather heavily right skewed. This is logical and should be expected, as the majority of states have crime rates closer to 0%, with a few states containing crime rights much higher than others.

The following figure shows the right-skewed distribution of Admitted Immigrant Rate.

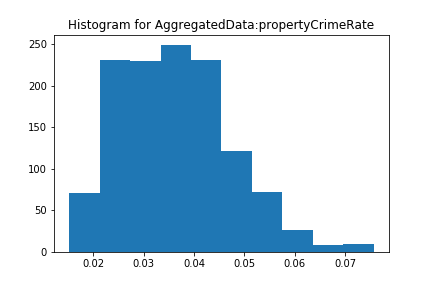
****

**Figure 2 - Admitted Immigrant Rate Histogram**

The following two figures also show the right-skewed distribution of the two general crime categories.

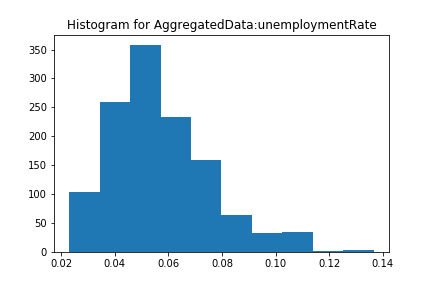


**Figure 3 - Violent Crime Rate Histogram**



**Figure 4 - Property Crime Rate Histogram**

Other standard of living metrics also shared this pattern, such as unemployment rate.



**Figure 5 - Unemployment Rate Histogram**

Section 3.3.2 will discuss the measures taken to make these attributes more normal, and thus able to be used in the different models implemented.

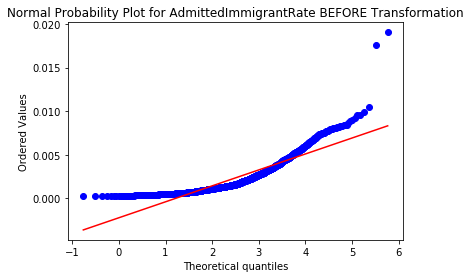
In addition, since much of the data will numerically change due to section 3.3.2, a further analysis of the descriptives will be conducted in section 3.3.3.

3.3.2 Transformation

Due to the issues of normality brought up by the original histograms, a standard cube root transformation was applied to all of the crime rate attributes, standard of living attributes, and Admitted Immigrant Rate.

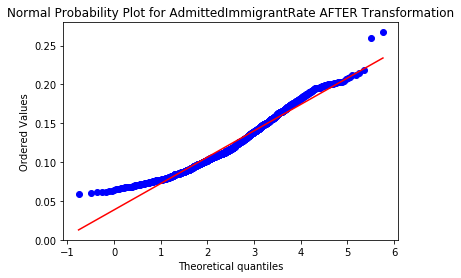
The cube root transformation is a commonly used transformation when attempting to turn right skewed data into more normal data. A cube root transformation was chosen over a logarithm transformation, a similar style in terms of turning right skewed dataset into a more normal one, due to the fact that much of the data is extremely close to 0, and applying a logarithm would not change the values as close to zero as it would other values. Thus a cube root transformation was applied.

Figure 6 depicts the normal probability plot for Admitted Immigrant Rate before the cube root transformation was applied.



**Figure 6 - Normal Probability Plot Before Transformation**

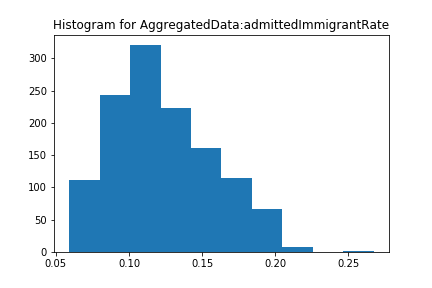
Figure 7 depicts the normal probability plot for Admitted Immigrant Rate after the cube root transformation was applied. While the lower order values are still a bit away from the linear trend, overall the data fits the plot significantly better than before the transformation.



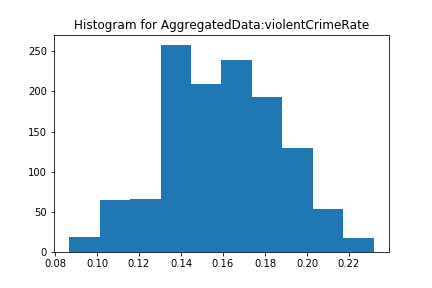
**Figure 7 - Normal Probability Plot After Transformation**

3.3.3 Post-Transformation Descriptive Statistics

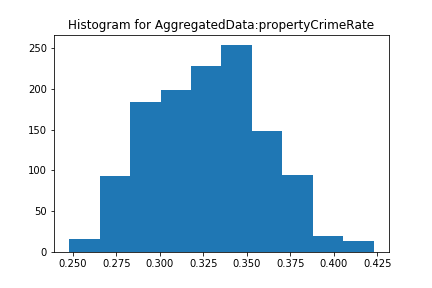
After the cube root transformation was applied, the histograms were again investigated to see if a normality assumption would be reasonable at this stage. The following figures depict the same variables mentioned in section 3.3.1 to demonstrate the change in distribution after the transformation.

**Figure 8 -Admitted Immigrant Histogram after Transformation**

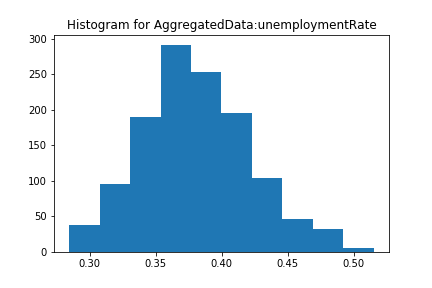
Admitted Immigrant Rate is significantly more normally distributed after the transformation, but is the only attribute that remains slightly right-skewed.

**Figure 9 -Violent Crime Rate after Transformation**

Violent Crime Rate can be assumed to be normally distributed.

**Figure 10 - Property Crime Rate after Transformation**

Property Crime Rate can be assumed to be normally distributed.

**Figure 11 -Unemployment Rate after Transformation**

Unemployment Rate can be assumed to be normally distributed.

Since these histograms all showed considerable improvement due to the cube root transformation applied, a further descriptive analysis follows.

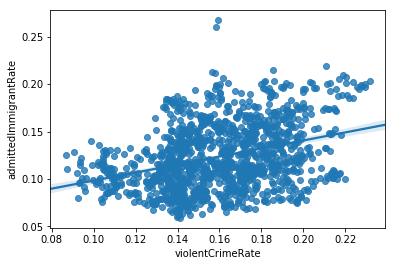
3.3.4 Regression Overview

Since the “Alternate Hypothesis” for this project is technically that there is a statistically significant relationship between immigration and crime, the “Null Hypothesis” follows to be that there is no statistically significant relationship between immigration and crime. Thus during the regression analysis, there are two goals in mind: 1) Obtain a low R-Squared for the model, meaning that the model does not accurately fit the data, 2) Obtain a high R-Squared for the model, but either have no crime variables left in the model, or the crime variables left in the model have negative constants, meaning that there is a negative correlation between that variable and immigration.

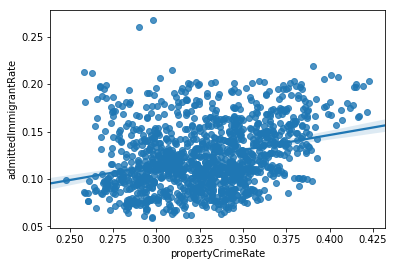
For most of the Simple Regression models and all of the Multiple Regression models, Admitted Immigrant Rate will act as the dependant variable.

3.3.4.1 Simple Regression

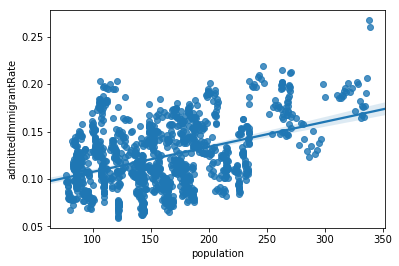
The first Simple model considered consisted of Admitted Immigrant Rate vs. Violent Crime Rate at the national level. The R-Squared for this model was 0.116, meaning the the model does not fit the data particularly well. The following figure depicts the linear relationship between the two variables.

**Figure 12- Admitted Immigrant Rate vs. Violent Crime Rate**

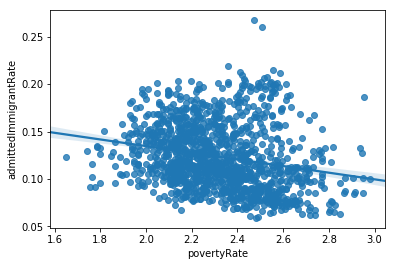
The next Simple model considered was Admitted Immigrant Rate vs. Property Crime Rate at the national level. The R-Squared for this model was 0.089, again meaning the model does not fit the data well. The following figure depicts the linear relationship between Admitted Immigrant Rate and Property Crime Rate.

**Figure 13- Admitted Immigrant Rate vs. Property Crime Rate**

The next Simple model considered was Admitted Immigrant Rate vs. Population at the national level. The R-Squared for this model was 0.175, again meaning the model does not fit the data well. The following figure depicts the linear relationship between Admitted Immigrant Rate and Population.

**Figure 14- Admitted Immigrant Rate vs. Population**

The next Simple model considered was Admitted Immigrant Rate vs. Poverty Rate. The R-Squared for this model was 0.049 again meaning that the model does not fit the data well. The following figure depicts the linear relationship between Admitted Immigrant Rate and Poverty Rate. It is interesting, however, that the weak relationship between these two variables was negative, meaning that as Immigration Rises, Poverty Rate decreases. Due to the R-Squared however, this conclusion should not be accepted to be credible, just something to note moving forward.

**Figure 15- Admitted Immigrant Rate vs. Poverty Rate**

Lastly, the final Simple model considered was Admitted Immigrant Rate vs. Unemployment Rate. The R-Squared was 0 for this so it can be concluded that there is no fit whatsoever between the two variables. Due to this, there is no figure associated with this model.

3.3.4.2 Multiple Regression

For Multiple Regression, there will be two categories of models considered: One being an “Outer Model”, where the independent variables considered include the general crime rates, i.e. Violent Crime Rate and Property Crime Rate as well as the other standard of living variables, and an “Inner Model”, where the independent variables considered include the specific crime rates, i.e. Murder Rate, Larceny Theft Rate, as well as the same standard of living variables as the Outer Model.

In addition, both nation-level and state-level versions of the Outer and Inner models will be considered. Since the goal is to be completely exhaustive, multiple regression, stepwise regression, variance inflation factors, and significance levels will all be considered in order to interrogate the worst-case scenario.

Since Multiple Regression can be considered “multi-dimensional simple regression” there will be are no 2-Dimensional plots associated with the Multiple Regression Models.

The four models implemented for Multiple Regression follow:

Model 1) Outer Model of Aggregated States (i.e. National)

Model 2) Inner Model of Aggregated States (i.e. National)

Model 3) Outer Model of Individual States

Model 4) Inner Model of Individual States

Model 1’s Multiple Regression implementation produced an R-Squared of 0.379, and no VIFs over 4. In addition, all variables considered in Model 1 were deemed significant, so the Stepwise Regression model produced the same model as the Multiple Regression procedure. The low R-Squared value obtained leads us to conclude that again we cannot say that there is significant relationship between immigration and crime.

Model 2’s Multiple Regression implementation produced an R-Squared of .495, however not only were 3 variables deemed not significant at the 0.10 level, but there were also 3 variables, namely Robbery, Murder, and Burglary that contained VIF values over 4.

When investigating Model 2’s Stepwise Regression Model that contained an R-Squared value of .491, the variables that were deemed not significant in the Multiple Regression Model were thrown out, however the 3 variables with VIF values over 4 were all deemed significant and thus remained in the model. As a result, a final model was created to correct this flaw.

The final model for Model 2 was an implementation of the stepwise model, however the 3 variables with high variance inflation factors, Murder, Burglary, and Robbery were manually thrown out before the stepwise procedure. This final model produced an R-Squared of .429.

At all stages during Model 2, none of the R-Squared were significant enough to “reject” the null hypothesis in consideration.

The last two models were more difficult to analyze since each of Model 3 and Model 4 consists of 50 sub-models that each represent an individual state. As a result, in each case, the sub-models with the highest R-Squared are interrogated, and the sub-models with the lowest R-Squared are considered just for reference.

Model 3’s Multiple Regression implementation produced a maximum R-Squared of 0.85 and was that of Tennessee, however, 4 of the 5 variables in the model were deemed not significant. In addition, there were 2 variables with VIF values over 4. The minimum R-Squared for Model 3’s Multiple Regression was 0.242.

Model 3’s Stepwise Regression produced a maximum R-Squared of 0.817 and was again that of Tennessee. Since the previous model demonstrated the 4 of the 5 variables we not significant, the final stepwise model in this scenario only contained 1 variable, being population. This result says that the stepwise model for Tennessee that only contained population is a very good fit for the data. The minimum R-Squared for Model 3’s Stepwise Regression was 0, due to the fact that all the variables were deemed not significant at the 0.10 level.

Model 4’s Multiple Regression produced a maximum R-Squared of 0.924, again that of Tennessee. For this model, 0 of the 11 variables were deemed significant, and 10 of the 11 contained VIF values over 4.

Due to this lack of fit, Model 4’s Stepwise Regression model produced a maximum R-Squared of 0.915, again that of Tennessee. This model only left Population, Labor Force Rate, and Aggravated Assault Rate in the model. However, Aggravated Assault Rate contained a negative coefficient of -1.4032, showing that Aggravated Assault Rate and Admitted Immigrant Rate have a negative linear relationship. The minimum R-Squared for Model 4’s Stepwise Regression model was 0.347.

Any physical regression outputs are available upon request.

3.3.5 Clustering

The approach to clustering was similar to that of Regression. Both Outer Models and Inner Models of variables were considered in clustering. In addition, both 2-Dimensional Clustering and Multi-Dimensional Clustering were considered as well.

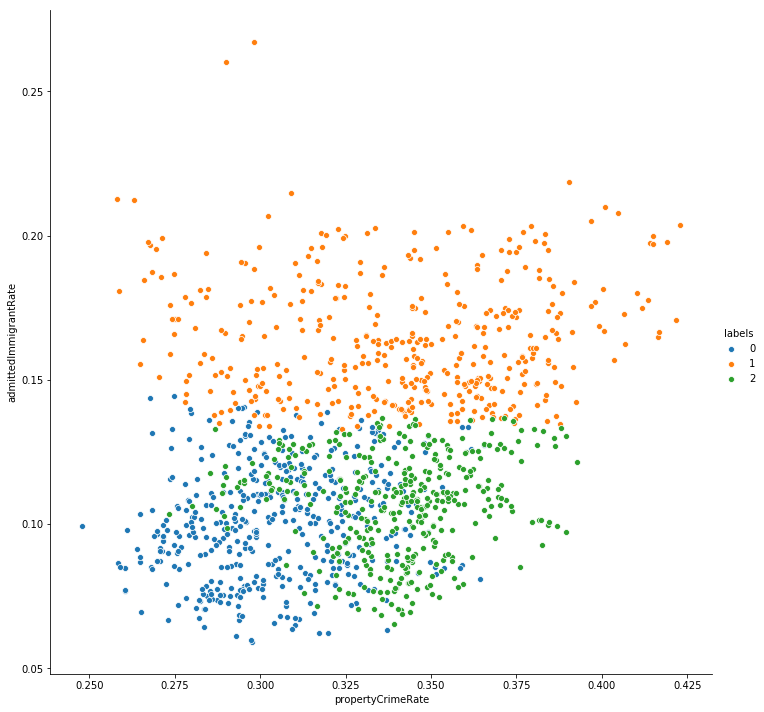
The first implementation of 2-Dimensional clustering was a simple 2-D K-Means. The generic input parameter of 3 clusters was used. For this specific problem, there are no “class labels” so implemented clustering is very much done as a blind approach. In the future, K-Means could be investigated with a variety of input number of cluster parameters, or even Mean-Shift, a K-Means alternative that solves the issue of inputting the number of parameters, could be implemented.

As was previously stated, there are no class labels for this problem, so the goal is to find any micro-patterns within the clustering.

The following figure depicts the application of 2-Dimensional clustering on Admitted Immigrant Rate vs. Violent Crime Rate. The implementation of this algorithm on these variables produces and very cut and clear separation of the 3 cluster labels. Because of this, it is imperative to investigate the cluster to see whether there are any patterns within the separated clusters. Unfortunately, apart from a specific state appearing in close proximity to itself in other years, there is no significant pattern discovered as a result of this implementation.

**Figure 16- 2-Dimensional K-Means Clustering: Admitted Immigrant Rate vs Violent Crime Rate**

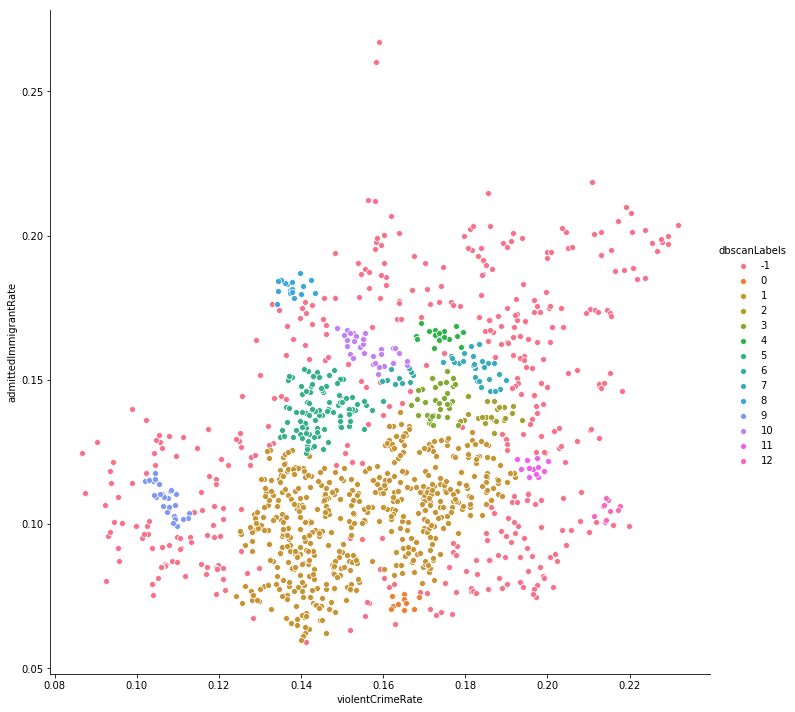
The next clustering model was very similar to the last, except investigated Admitted Immigrant Rate vs. Property Crime Rate. While the clustering output contained very clear and separate labels for the Violent Crime Rate, the implementation on Property Crime Rate did not have the same clarity in distinction between clusters. The following figure shows the output of the K-Means implementation with parameter number of cluster of 3 between Admitted Immigrant Rate vs. Violent Crime Rate.

**Figure 17- 2-Dimensional K-Means Clustering: Admitted Immigrant Rate vs Property Crime Rate**

In addition to implementing K-Means, DBSCAN was applied to the same variables in order to check multiple clustering methodologies on the dataset. DBSCAN is a density-based clustering that uses the minimum distance between nodes and the minimum number of connected nodes to be considered a cluster as parameters. Even with these parameters, the DBSCAN implementations are more resilient to bias in terms of the parameters used. In addition, DBSCAN is valuable due to the lack of cluster labels, since the goal is to find micro-clusters within the large dataset.

The first implementation of DBSCAN was on Admitted Immigrant Rate vs. Violent Crime Rate, and the figure showing the cluster assignments is listed below. Since the number of clusters is not a parameter, that value can fluctuate much more in DBSCAN implementations and thus the number of clusters in the first model was much higher than that of K-Means.

It also should be noted that any nodes with a -1 cluster label are determined “noise” by the algorithm and thus not assigned to any clusters.

**Figure 18- 2-Dimensional DBSCAN Clustering: Admitted Immigrant Rate vs. Violent Crime Rate**

There are a number of notable micro-clusters that appear in the plot that would suggest that there should be more investigation into the potential patterns that are appearing. Unfortunately, there were no notable patterns within each of these densely packed nodes.

The next implementation of DBSCAN was to investigate Admitted Immigrant Rate and Property Crime Rate. Similarly to K-Means, the implementation on Property Crime Rate was less revealing than that of Violent Crime Rate, although the figure of the implementation follows.

**Figure 19- 2-Dimensional DBSCAN Clustering: Admitted Immigrant Rate vs Property Crime Rate**

The only implementation of Multi-Dimensional Clustering was with the K-Means algorithm. In the future, this could be expanded upon.

Since there is no 2-Dimensional Plot available for Multi-Dimensional Clustering, the algorithm run on the Outer Model, was compared against the algorithm run on the Inner Model. The cluster label assignments were then compared against each other and 0 out of the 1250 nodes changed label assignments between the Outer Model and the Inner Model. This suggests that investigating the Outer Model is comparable to the Inner Model, specifically for clustering, but potentially for other statistical analyses.

# Results

Since the format of this project is a little unconventional, through the implementation of various types of regression, and clustering, it can be concluded that there is not enough statistical evidence to say that there is any relationship between Immigration Rates and Crime Rates.

# Conclusion

As we saw from the correlation matrix and the regression, none of the variables contained high R-squared values. There were similar results for the inner and outer multiple and stepwise regression. While some originally contained higher values, none of them showed positive correlation between immigration and crime rates after accounting for the variables that were not considered to be significant and the high VIF values. The only significant variables we found for any models were Population, Labor Force Rate, and Aggravated Assault, which was actually negatively correlated.

Thus, we were not able to find evidence of a significant relationship between immigration and crime. There could be possible underlying causes that we could test if we had more variables to compare against the immigration rate. More clustering could be implemented to look at any potential micro-patterns in the data. In the future, if applicable, we could look at any hypothesis testing that could possibly be carried out.

##### References

1. Akins, Scott, et al. “Immigration, Economic Disadvantage, and Homicide: A Community-Level Analysis of Austin, Texas.” Homicide Studies, vol. 13, no. 3, Aug. 2009, pp. 307–314, doi:[10.1177/1088767909336814](https://doi.org/10.1177/1088767909336814).
2. Martinez, Ramiro, and Matthew T Lee. “On Immigration and Crime.” The Nature of Crime: Continuity and Change, vol. 1, pp. 485–515., www.ncjrs.gov/criminal\_justice2000/vol\_1/02j.pdf. National Criminal Justice Reference Service.
3. Mears, Daniel P. The Immigration-Crime Nexus: Towards an Analitical Framework for Assesing and Guiding Theory, Research, and Policy. vol. 44, Pacic Sociological Association, 2001, pp. 1–19.
4. Ousey, Graham C, and Charis Kubrin. “Immigration and Crime: Assessing a Contentious Issue.” Annual Review of Criminology, vol. 1, no. 1, 2008, pp. 63–84, doi:10.1146/annurev-criminol-032317-092026.
5. Polczynski Olson, Christa, et al. “Immigration and Violent Crime: Citizenship Status and Social Disorganization.” *Homicide Studies*, vol. 13, no. 3, Aug. 2009, pp. 227–241, doi:10.1177/1088767909336202
6. Disorganization.” Homicide Studies, vol. 13, no. 3, Aug. 2009, pp. 227–241, doi:[10.1177/1088767909336202](https://doi.org/10.1177/1088767909336202).