Relationships between Specific Crime Rates and Crime Inducing Attributes In California Counties

Brevan Chun
University of California, Santa Cruz
Santa Cruz, CA, United States
bzchun@ucsc.edu

Eagle Huang
University of California, Santa Cruz
Santa Cruz, CA, United States
khuang53@ucsc.edu

Justin Cheng
University of California, Santa Cruz
Santa Cruz, United States



Abstract

This report studies the various relationships between household incomes, high school completion rates, and other data elements in every California county in order to identify patterns and relationships that contribute to area specific crime rates. This report utilizes data from government databases such as the CDE (California Department of Education), the California Census, and Federal House Index as well as other reputable sources.

Keywords: Data mining, Crime, Predictive model, California, Classification techniques, Naive Bayes Classification, Linear Regression, Decision Tree Induction, data integration and cleaning. Clustering techniques, Evaluation, ROC Curves. Accuracy measures.

1 Introduction

California is both the most populous, and the 3rd largest state in the country. It is no surprise that the state also holds one of the highest incarceration rates in the world. With the 2019 Coronavirus pandemic, the crime rates in the state have shifted in different aspects. Though overall property crime has decreased, homicide cases have increased by a troubling 30% and motor theft has increased by 20% (PPIC).

There are an immense number of factors that potentially contribute to this increase in crime. The impact of the pandemic is impossible to ignore, however, we are looking to analyze additional aspects, such as housing income, high school graduation rates, suspension rates that may or may have not been affected by coronavirus.

1.1 Our Goals and Expectations

Our goal was to find interesting and undiscovered relationships between specific crime rates and various indicators (e.g. socioeconomic conditions and aggravated assault). The relationship between population and population density and crime rates (or lack thereof) is already well documented so we wanted to examine several other factors. Discovering such relationships would help policy makers in making informed decisions in their efforts to reduce crime.

We expected to find interesting relationships between crime rates and income, age, race, county spending, and more. Not every attribute we use will correlate with a specific crime, but we expect to see several that do. It is important to remember that correlations and relationships outlined and discovered in this report do not translate to causation. Our goal was to discover relationships to potentially isolate for further study.

2 Data

We compiled data from several different and reliable sources. We separated the data by California county which meant we would have a rather small data set. To make up for the low number of examples (there are 58 counties in total) we had a lot of attributes. In total, we had 27 attributes for each object. The data attributes are general or related to crime or demographics.

- General
  - Total Population
  - Personal Income per Capita
  - Highschool Graduation Rate
  - Unemployment Rate
  - Square Mileage
  - Median Home Price
  - Active Voters
  - GDP
  - Public Assistance Spending
  - Education & Recreation Spending
  - Median Age

- Crime
  - Violent
  - Murder
  - Rape
  - Robbery
  - Aggravated Assault
  - Property
  - Burglary
  - Vehicle Theft
  - Larceny Theft
  - Total Crime Cases
- Demographic
  - White %
  - Asian %
  - African %
  - Hispanic or Latino %
  - Native American %
  - Others %

2.1 Data Selection and Preprocessing

We sourced data from many different sources and compiled it into a single file - there are 58 rows and 28 columns. Comparatively speaking, there is very little data to use for modeling and analysis. Most data mining projects will utilize many more examples (over 10x) but we were only interested in crime in California and there are only so many areas to look at.

Attributes such as "High School Graduation Rate" with a few missing values are handled by taking the average of the rest of the known values and using that as replacement. Redundant attributes like population, which was included in multiple data sets, were discarded.

Our primary goal in the selection of data attributes was to get factors that would be different in every county. Things like total population, income per capita, county GDP, and graduation rates will all differ greatly in each county. We hoped that these differences would be identifiable and useful in drawing relationships.

Preliminary data analysis shows us that there are often outliers for nearly every attribute - and this outlier is nearly always Los Angeles county. L.A. county has by far the largest population, highest GDP, most voters, highest expenditures, and more. This presents a problem in our models because even after normalizing, L.A. county remains a stark outlier. For example, after z-score normalization, L.A. county has a population value of 6.402 while the next closest county is San Diego with a value of 1.790 (and the average being 0.00).

On the other hand, counties such as Aline are so low in population compared to other areas, it has the same effect. In future studies and projects, it may be best to only look at reasonably average areas and study L.A. county individually due to its shear size. These outliers may have influenced the performance and reliability of our clustering and classification models.

2.2 Data Correlation

Looking at the correlation matrix in figure 1, it is evident that white population percentage negatively correlates with many other attributes, including demographic diversity and crime. On the other hand, Asian/African population are positively correlated with crimes such as robbery, property theft and vehicle theft while the native American population are positively correlated with violent crimes, rape, aggravated assault and burglary. Personal income per capita is also positively correlated with larceny, property theft and robbery. Further relationships can be found in cluster analysis.

3 Cluster Analysis

When performing clustering, we first normalized the data using range transformation so that each attribute is equally weighted but also maintains the same distribution. Attributes are transformed into the range 0 to 1. It should be noted that attributes that do not show notable correlation with crime or other attributes are discarded for this process as it could become noise in the data. As a result, attributes like life expectancy and GDP who did not show correlation were discarded. Additionally, choosing the number of clusters is an important step when performing clustering. We first used the k means++ algorithm in RapidMiner to determine the best number of clusters, which came out to be 5. However, this only yielded an average distance within centroid of 0.38. We believed this value could be lower so we tested each number of cluster in order to determine the most optimal value which will ensure a more accurate result.

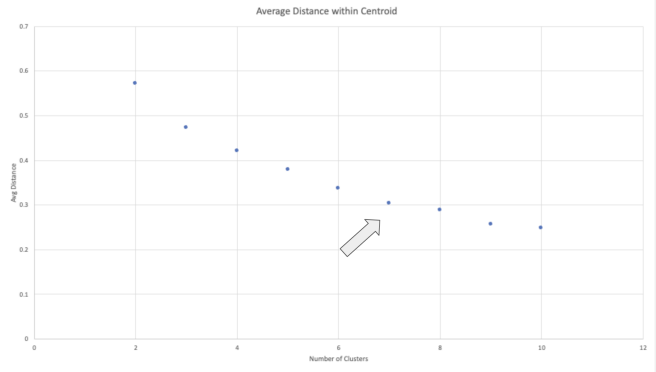

Figure 2. Average Distance Within Centroid for Different Clusters

Looking at Figure 2, it can be seen that the decrease of average distance between each subsequent cluster starts to level off at 7 clusters. Thus, we chose 7 as our most optimal number of cluster and ran our clustering model using 7 as our initial number of centroids.

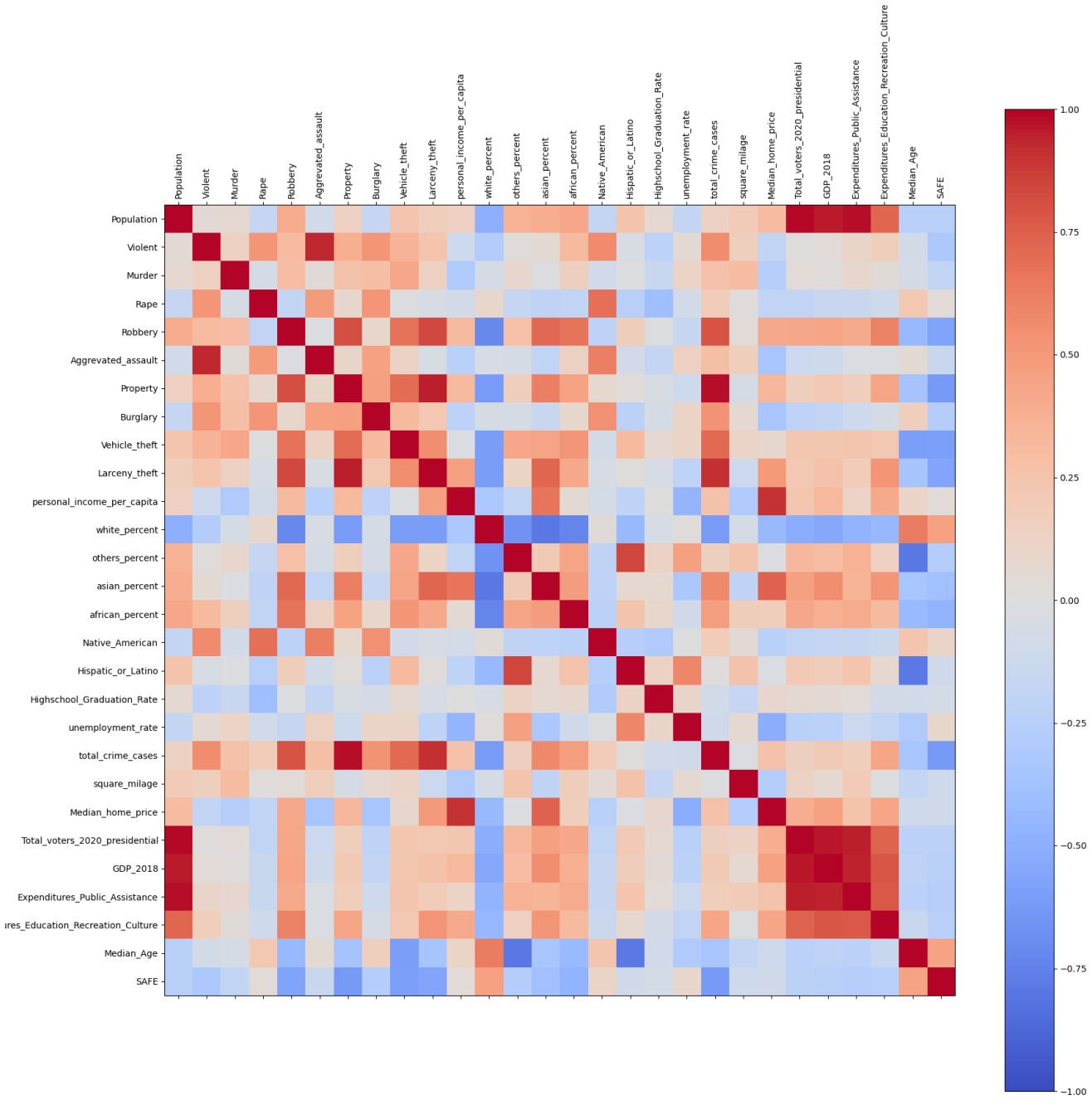

Figure 1. Correlation matrix with all the attributes in the data set. Red is positively correlated, blue is negatively correlated, white is no correlation.

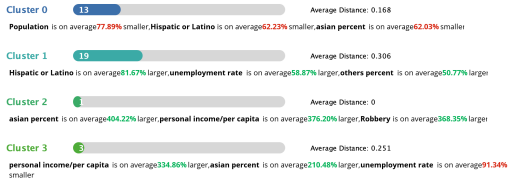

Figure 3. Cluster Analysis 1

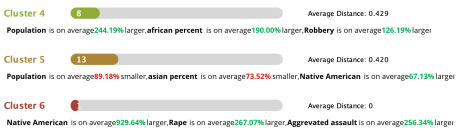

Figure 4. Cluster Analysis 2

Figures 3 and 4 each displays key characteristics of each cluster, most notable of which, are clusters 2,6 and 4. Although clusters 2 and 6 shows strong correlation between some of the attributes, they look to be potential outliers as they each only contain 1 item from the dataset. However, due to the fact that we are working with a relatively small number of data, outliers are being kept in the dataset for further analysis. On the other hand, cluster 4 shows a strong correlation between African population and robbery, which conforms with previous studies. Figure 5 is a visualization of the clusters in the form of a heat map. The dark green square shown in cluster 6 confirms that it is an outlier as the "native-American" attribute is 900% the average.

Thus, clustering results showed us that robbery and African population does have a strong correlation between different counties in California. In addition, going forward, we kept in mind that our data consists of some outliers so any results should be double checked, especially correlations related to the "native-American" attribute.

Figure 5. Cluster Visualization using Heat Map

4 Classification Models

A binary attribute called “Safe” is constructed based on information from other attributes like total crime and the type of crimes. Attribute construction is utilized in this study so that classification models can be used it as a target attribute for prediction. A binary attribute was chosen to maintain simplicity.

To predict the target attribute, we implemented several classification models: predictive trees, linear regression, K-nearest neighbor, and Naive-Bayes. We chose to use many classification models simply because it was not clear how these models would behave given the relatively small number of examples and high number of attributes.

4.1 Predictive Trees

Figure 6. One of the decision trees generated by the random forest tree model.

We used three types of predictive trees in total: a standard decision tree, a random forest tree, and a gradient boosted tree. The tree models were implemented in RapidMiner and all used cross validation to create and test the models. Additionally, all the models used a filtered attribute list and normalized data.

The tree models did not utilize the full list of attributes, instead only used a subset of them. The attributes that were not used included nearly all the crime data. This is because a model that predicts crime but also itself needs that crime data is not very useful. A model that can predict crime levels based on other information may be more useful and insightful. Thus, each tree model only utilized 17 attributes.

In the end, the tree models performed similarly. It is hard to say which model is “best” because they are all relatively close in terms of accuracy, recall, and class precision. It would be interesting to see how they perform when more data is introduced.

4.1.1 Decision Tree. The standard decision tree exhibited the highest overall accuracy among the two other tree models. We found the highest accuracy while using pre and post pruning and gain ratio as the indicator. These are close to the standard settings for decision trees in RapidMiner. The performance of the standard decision tree model was better than expected.

Decision Tree		Class Precision	
Overall Accuracy	Recall	Yes	No
88.03% +/- 4.30%	88.81% +/- 6.49%	91.43%	82.61%

Table 1. Standard decision tree model performance.

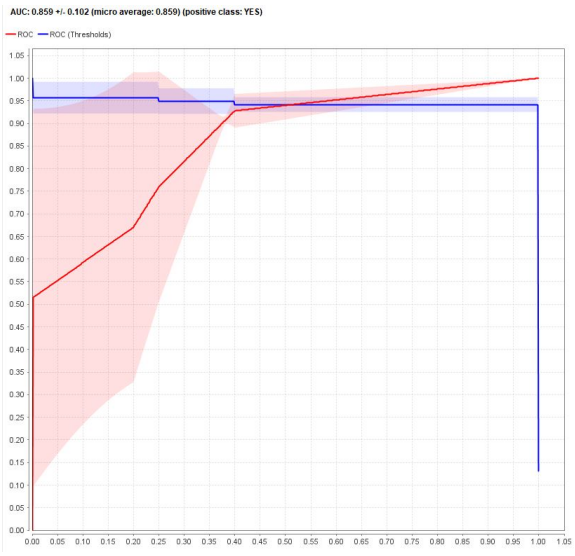

Figure 7. The ROC curve of the standard decision tree model.

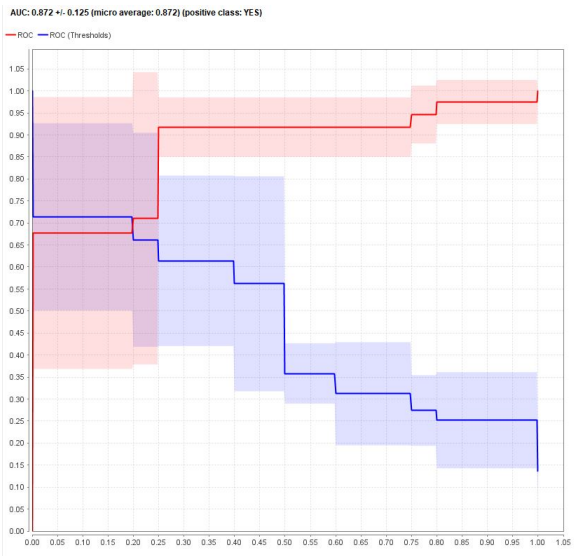

Figure 8. Random forest tree ROC curve.

4.1.2 Random Forest Tree.

Random Forest Trees attempt to improve upon standard decision trees by generating more trees and using the average prediction of the trees. This usually leads to greater accuracy and precision than standard decision trees. For our random forest tree model, we chose to generate 100 trees and information gain. Information gain gave us better accuracy than gini index.

The performance of this model is marginally worse than our standard decision tree. There could be many reasons for this, one of which is the data itself. Having so few examples makes it difficult to create high quality models.

4.1.3 Gradient Boosted Tree.

Similar to random forest trees, gradient boosted trees usually have improved performance over standard decision trees. Gradient boosted trees are also generally considered to better perform random forest trees as well using a much more complex method of optimizing. Like our random forest tree model, we set the model to generate 100 trees and a cross validation of 3 folds.

This model only seemed to improve marginally in "yes" class precision. However, this came at the cost of a lower "no" class precision. Overall, this model provides no significant improvement or loss over the previous.

Random Forest Tree		Class Precision	
Overall Accuracy	Recall	Yes	No
82.73% +/- 5.95%	82.74% +/- 15.69%	88.24%	75.00%

Table 2. Random forest tree model performance.

Gradient Boosted Tree		Class Precision	
Overall Accuracy	Recall	Yes	No
82.81% +/- 5.80%	78.17% +/- 12.17%	93.33%	71.43%

Table 3. Gradient boosted tree model performance.

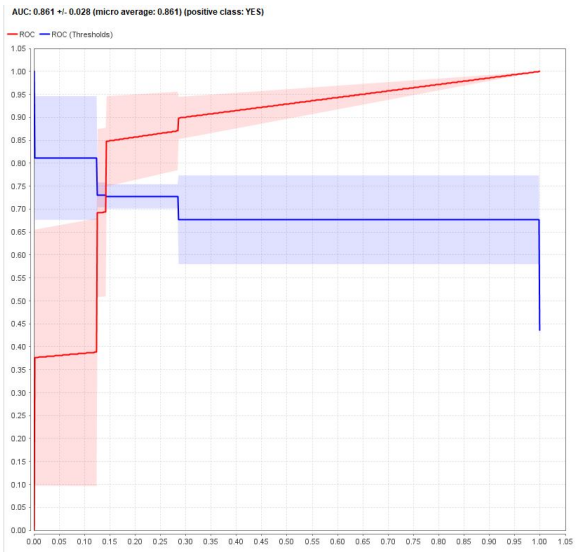

Figure 9. Gradient boosted tree ROC curve.

4.2 Linear Regression

In linear regression, we want to predict the weights of the independent attributes to estimate the value of our target variable. We convert SAFE classes NO and YES into zero and one. The remaining 27 attributes are selected as dependent variables (feature variables), while SAFE is our dependent variable (target variable). Our linear regression model will then predict the probability that a county is safe or not. Python allows us to quickly apply linear regression to the data matrix, through the use of the sklearn linear model library. The data is divided into 80/20 training and test sets, where the regression model calculates weights of the dependent variables based on the training set. Then our model takes the test set as an input to determine the accuracy of the linear model. The accuracy of a linear regression model applied to this data averaged to 55.56 percent.

4.3 K-Nearest Neighbor

For a KNN classifer, we continue utilizing our data as training and test sets. In this method, the test set is classified as SAFE or UNSAFE based its distances from the training sets. We chose k = 5, since our data is not large. For each test vector, we calculate the five closest training vectors to determine whether the testing data is SAFE or UNSAFE. Though KNN is considered a "lazy" classifier, the method is intuitive and simple for our small dataset. On the other hand, our data has high dimensionality, where KNN is not the best at handling multiple attributes. The resulting accuracy of our KNN classifier was 57.41 percent.

4.4 Naive-Bayes

Another approach to this classification problem, is with the use of a Naive Bayesian model, as this classifier is best suited for data with high dimensionality and size. We must first assume that all attributes in the data are independent of each other. Python's sklearn library provides a variety of functions to apply the Naive Bayes model to all of our data. We specifically chose a Gaussian Naive Bayesian model over a multinomial Naive Bayesian model, due to frequent zero values in our crime rates data.

First, prior probabilities are computed, as well as the histograms for each attribute column to determine mean and variance values. We can then utilize SAFE and UNSAFE Gaussian distributions for each county attribute. Furthermore, logarithmic functions are applied to the prior probability and gaussian likelihoods for each test sets to determine the scores of whether the training set is class 0 or class 1. This provided a much higher average accuracy of 78 percent.

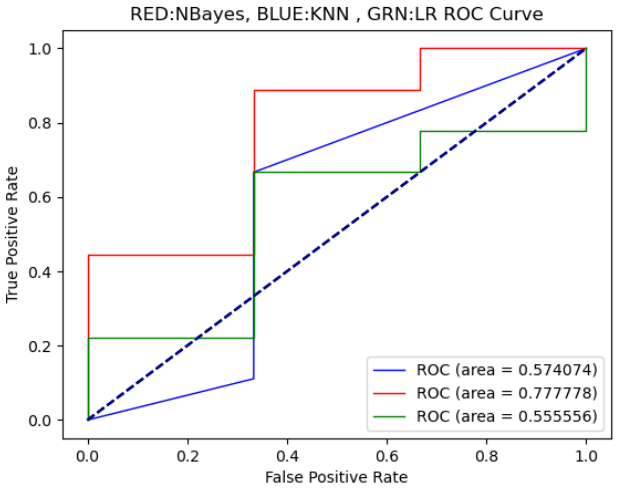

Figure 10. ROC curves for out linear regression, K-nearest neighbor, and Naive-Bayes models. (Red represents Naive Bayes, Blue represents K Nearest Neighbors, and Green represents Linear Regression).

5 Conclusions

Our classification models were successful in predicting the "safe" attribute. Some models performed better than others but adding more data may change that. For now, our models allow us to predict whether or not a California county is safe or not based on several data attributes not necessarily related to crime.

Finding such patterns enables us to hone in on certain aspects of these geographic areas that may contribute to crime levels. Further study in these aspects could enable us to make strong recommendations to policy makers for reducing crime. For example, unemployment correlated with general crime so we suggest greater support for businesses so more jobs are available. Additionally, Latino, Hispanic, and "other" ethnicity correlated with unemployment so we

suggest greater focus on employing these groups and increasing diversity in businesses overall. Similarly, Asian, African, and high income per capita correlates with higher levels of robbery. Counties with such demographics should focus resources on this type of crime in particular. It is important to recognize that these are correlations and not causation. We only provide these conclusions as a stepping point from which to study further. Doing so would require less data mining and more sociology and psychology.

To improve upon this project we would require much more data. The obvious next step is to expand the data set nationally - gathering data from every county in the United States. Still, there are only about 3,000 counties in the US so we would also gather more attributes - perhaps also replacing some of the ones used in this project. There are so many factors that influence crime, it may be difficult to gather all of them.

Another improvement could be in our classification attribute. Making the "safe" attribute a numeric value would make it more practical for real life applications. A binary attribute, safe or not safe, offers very little depth and insight into the characteristics of the county. A numeric value would enable much greater nuance in our model predictions.

6 Appendix

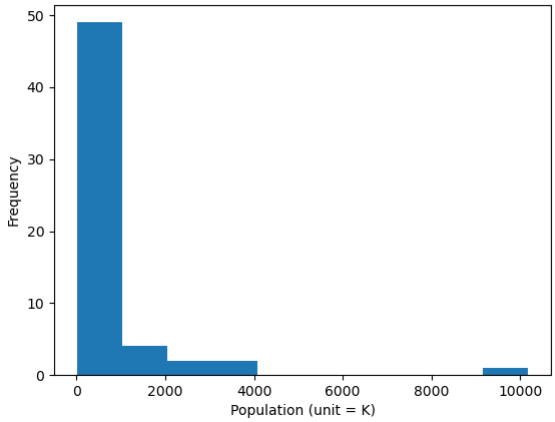

Figure 11. Population Histogram

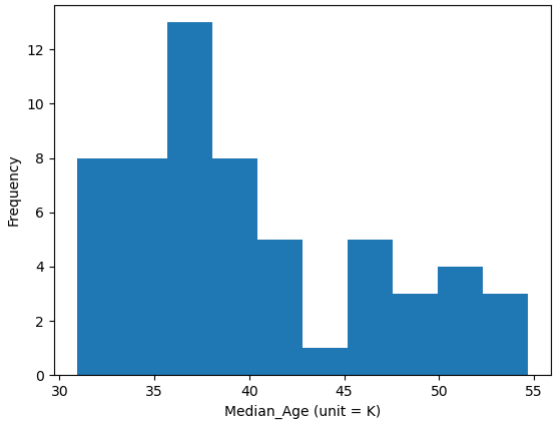

Figure 12. Median Age Histogram

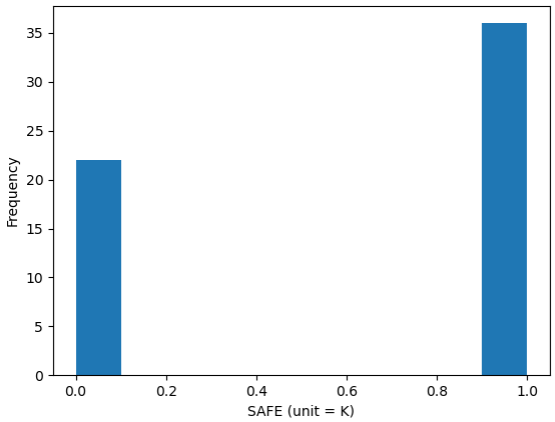

Figure 13. SAFE attribute Histogram

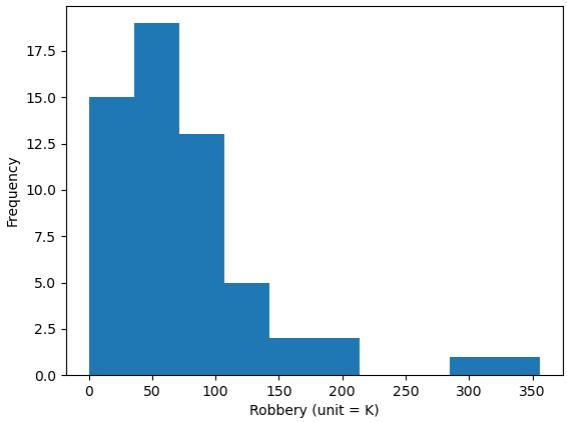

Figure 15. Robbery Cases Histogram

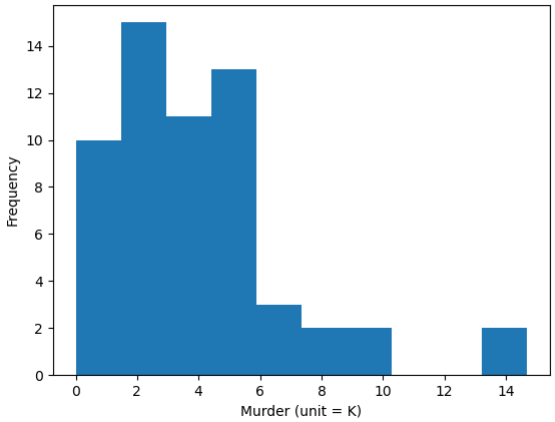

Figure 14. Murder Cases Histogram

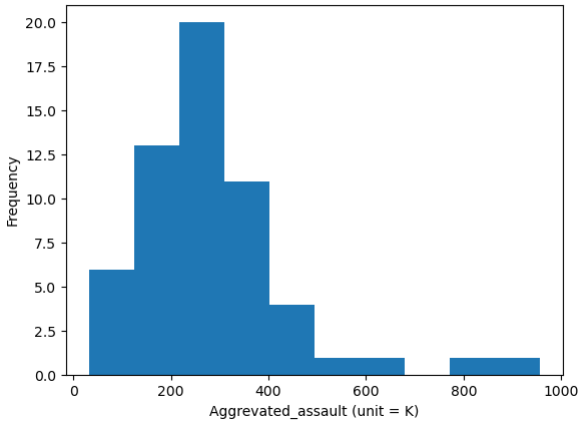

Figure 16. Assault Cases Histogram

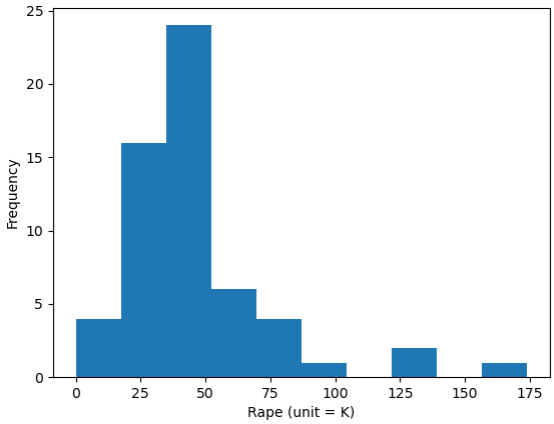

Figure 17. Rape Cases Histogram

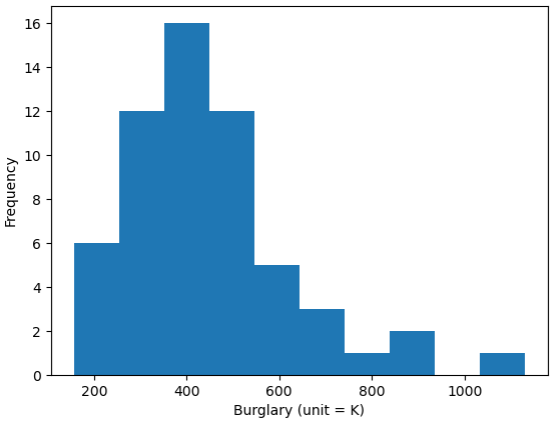

Figure 19. Burglary Cases Histogram

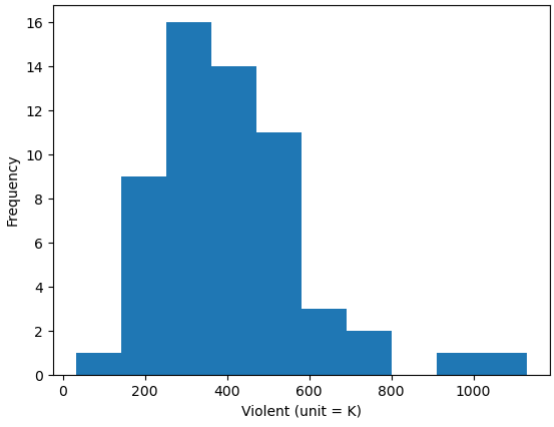

Figure 18. Total Violent Cases Histogram (murder, robbery, assault, rape)

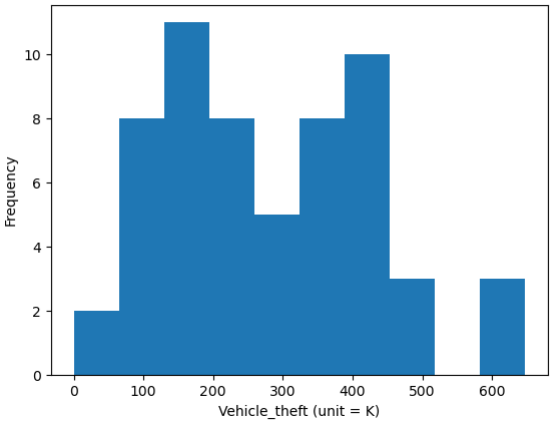

Figure 20. Vehicle Theft Cases Histogram

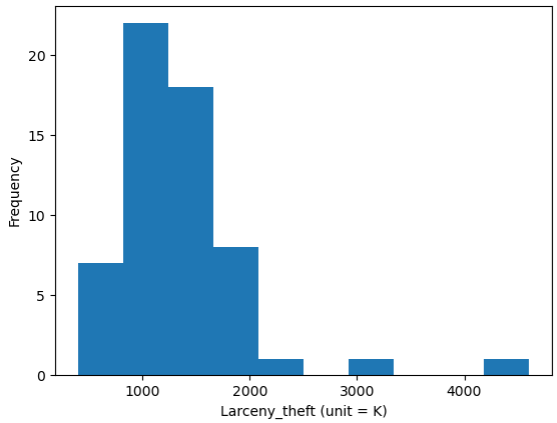

Figure 21. Larceny Cases Histogram

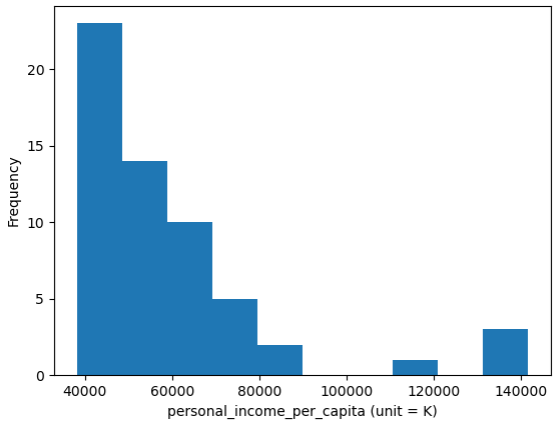

Figure 23. Personal Income per Capita Histogram

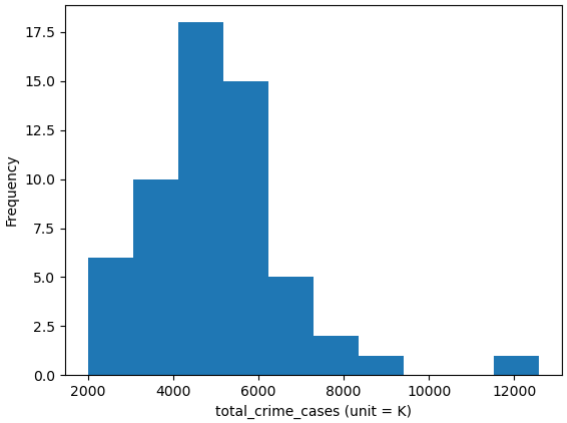

Figure 22. Total Crime Cases Histogram

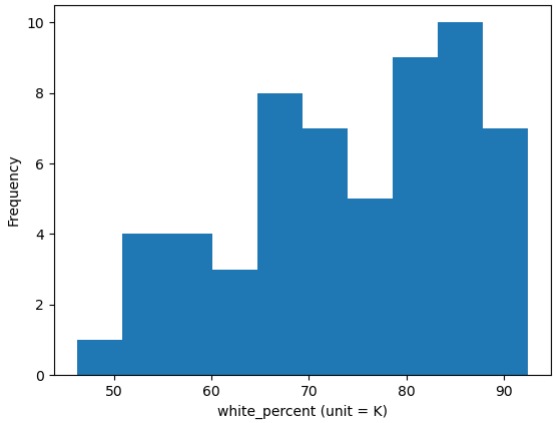

Figure 24. White Percent per County Histogram

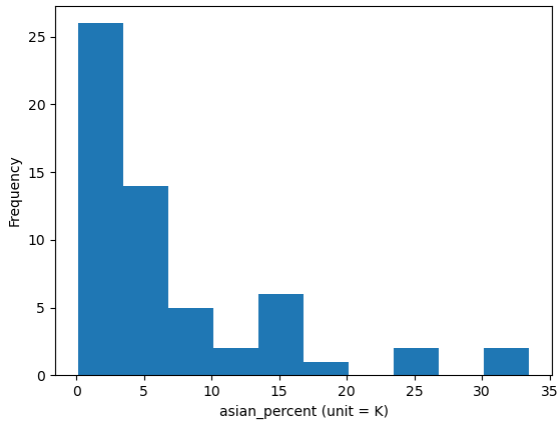

Figure 25. Asian Percent per County Histogram

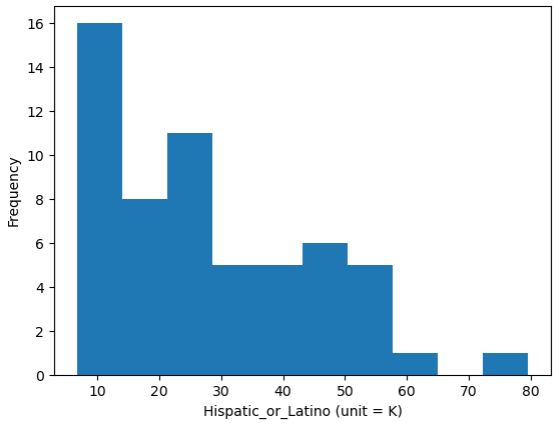

Figure 27. Hispanic Percent per County Histogram

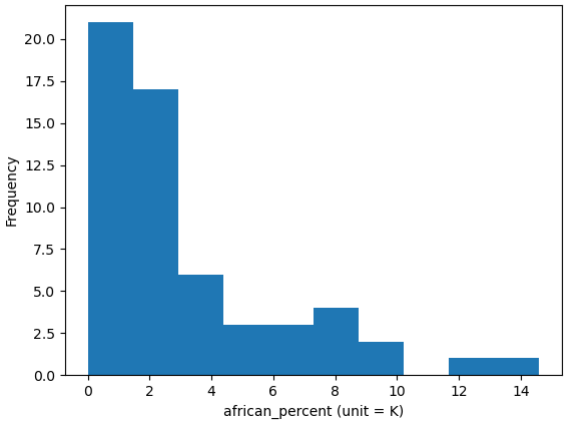

Figure 26. African Percent per County Histogram

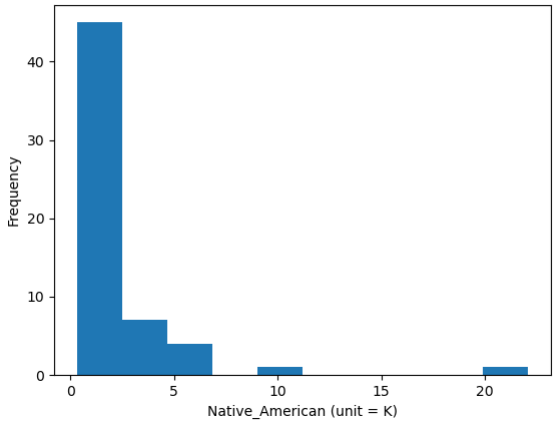

Figure 28. Native American Percent per County Histogram

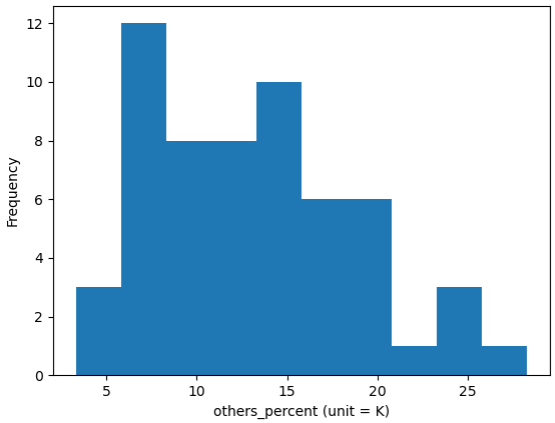

Figure 29. Others Percent per County Histogram

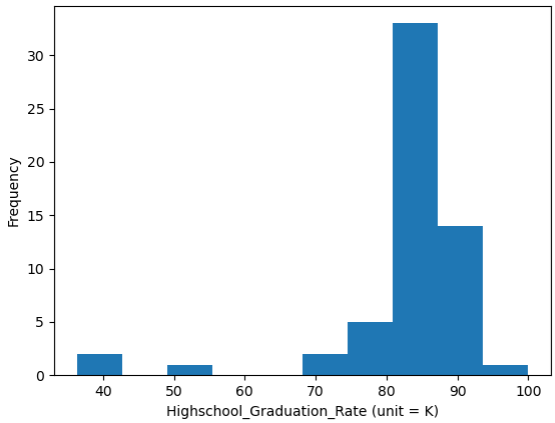

Figure 31. High School Graduation Rates Histogram

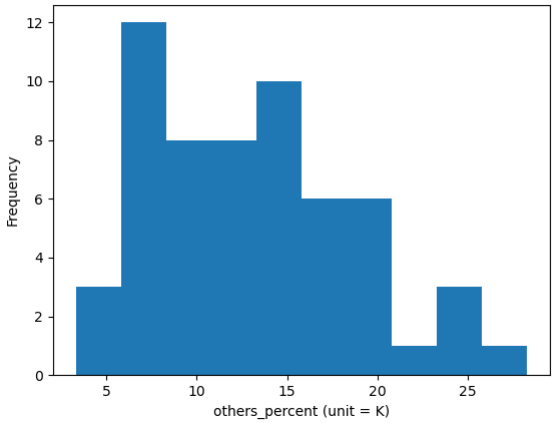

Figure 30. Others Percent per County Histogram

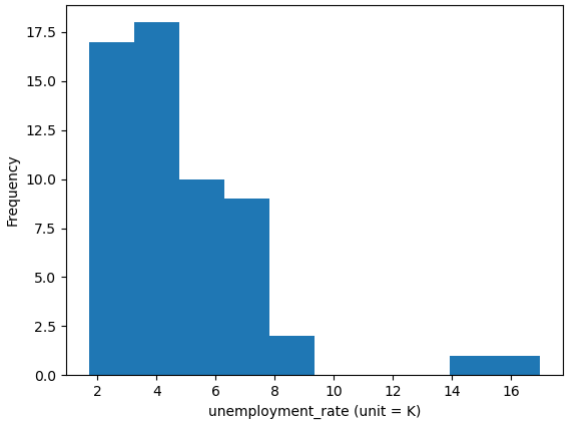

Figure 32. Median Home Prices Histogram

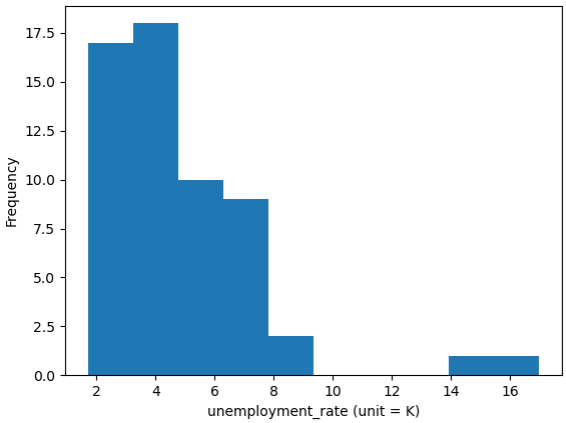

Figure 33. Unemployment Rate Histogram

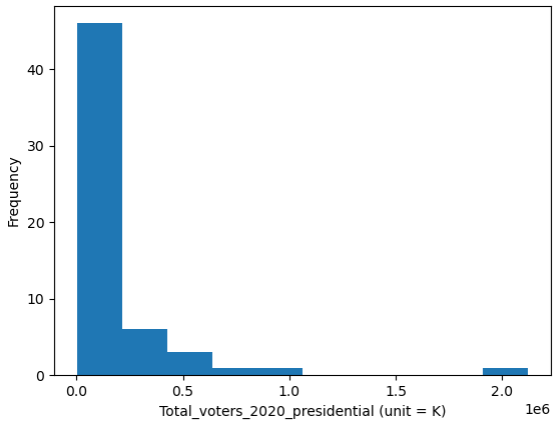

Figure 35. Total presidential votes per County (2020) Histogram

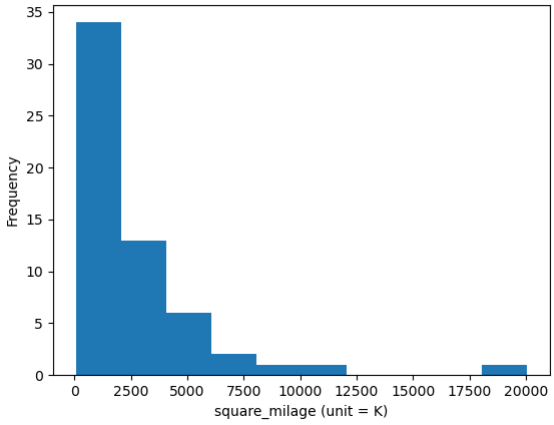

Figure 34. Square Mileage of County Histogram

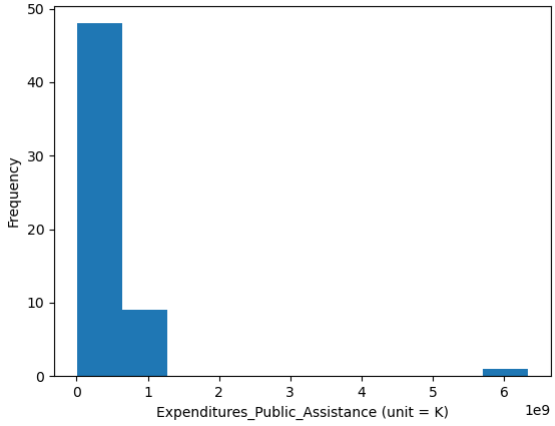

Figure 36. Public Assistance Expenditures Histogram

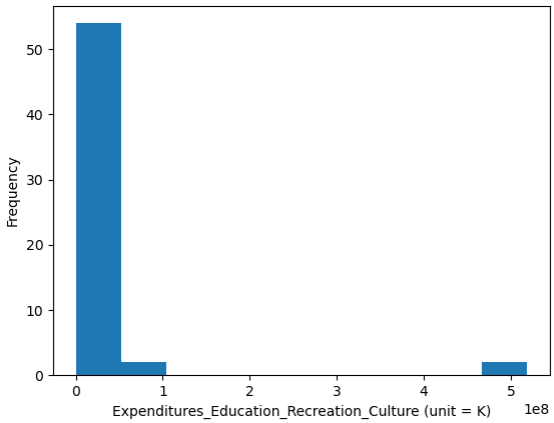

Figure 37. Education, Recreation, Culture Expenditures Histogram

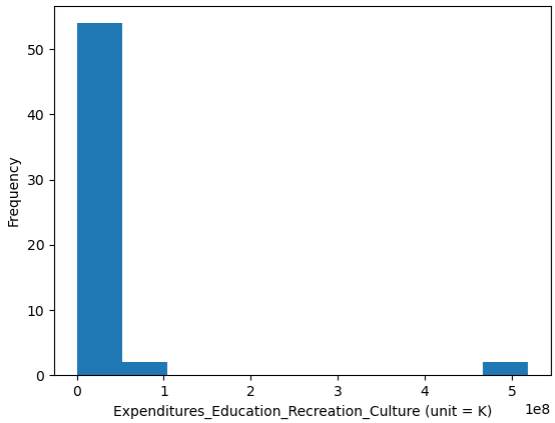

Figure 38. Gross Domestic Product per County Histogram