Violent Crime in Communities

Kevin Kuo

Department of Computer & Information Sciences   
College of Science & Mathematics   
Towson University

kkuo1@students.towson.edu

Mary Snyder

Department of Computer & Information Sciences   
College of Science & Mathematics   
Towson University

msnyde8@students.towson.edu

**ABSTRACT**

The Federal Bureau of Investigation Uniform Crime Report (UCR) of 1995 data set contains many attributes and statistics about the communities in which crime data is collected. Combined with Census data from 1990 as well as law enforcement data from the 1990 US LEMAS survey, the crime data set is one of the most extensive data sets available. We will analyze the data set and determine which indicators are associated with communities of high violent crime, and more importantly, how differences in indicators may predict an increase or decrease in per capita crime levels in other communities.

**Categories and Subject Descriptors**

H.2.8 **[Database Management]** Database Applications – *Data mining*

**Keywords**

Uniform Crime Report; Federal Bureau of Investigation; Crime; Violent Crime; Communities; Police; Investment; Socioeconomic; Census; LEMAS Survey; Racial; Distribution of Attributes; Histogram; Scatterplot; Scatterplot Matrix; Normalization; Binning; Equal Width; Regression Analysis; Classification; Multivariate; Categorical; Decision Tree Classification; Training and Testing; Holdout Method; Accuracy; Error Rate; Sensitivity; Specificity; Precision; Recall; F Measure; Frequent Itemset; Association Rule Mining; arules; Apriori Algorithm; Eclat Algorithm; Support; Confidence;

# INTRODUCTION

## Dataset

We chose a dataset from the UCI Machine Learning Repository whose attributes are categorical in nature. This Community and Crime dataset comprises socioeconomic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. The dataset contains information that describes or involves the community, such as the percent of the population considered urban and the median family income, as well as involve law enforcement, such as per capita number of police officers and percent of officers assigned to drug units. The issue of violent crime has been a contentious domestic political issue with sweeping implications. In this analysis, we will explore other indicators of communities along with its prevalence of violent crime to determine what relationships exist and suggest further areas of study.

### Data Attributes Selection and Values

The attributes selected for this dataset were from larger databases of census and crime information. They were included here if there was any plausible connection to crime and/or the attribute the dataset authors intended to be predicted, Per Capita Violent Crimes (ViolentCrimesPerPop). There are 1994 instances with some missing values (indicated by a ‘?’ value) and 128 attributes: 122 predictive, 5 non-predictive, and 1 goal.

Non-predictive values include:

*state* - US state, by number

*county* - numeric code for county[[1]](#footnote-1)

*community* - numeric code for community1

*communityname* - for information only

*fold* - fold number for non-random 10 fold cross validation, potentially useful for debugging, paired tests

Predictive values include:

*population* - population for community

*householdsize* - mean people per household

*racepctblack* - % of population that is African American

*racePctWhite* - % of population that is Caucasian

*racePctAsian* - % of population that is of Asian heritage

*racePctHisp* - % of population of Hispanic heritage

*agePct12t21* - % of population 12-21 in age

*agePct12t29* - % of population 12-29 in age

*agePct16t24* - % of population 16-24 in age

*agePct65up* (% of population 65+ in age

*numbUrban* - # of people living in areas classified as urban

pctUrban - % of people living in areas classified as urban

*medIncome* - median household income

*pctWWage* - % of households with wage or salary income[[2]](#footnote-2))

*pctWFarmSelf* - % of households w/farm or self-employment income2

*pctWInvInc* - % of households with investment/rent income2

*pctWSocSec* - % of households with social security income2

*pctWPubAsst* - % of households with public assistance income2

*pctWRetire* - % of households with retirement income2

*medFamInc* - median family income[[3]](#footnote-3); differs from household income for non-family households

*perCapInc* - per capita income

*whitePerCap* - per capita income for Caucasians

*blackPerCap* - per capita income for African Americans

*indianPerCap* - per capita income for Native Americans

*AsianPerCap* - per capita income for people with Asian heritage

*OtherPerCap* - per capita income for people with 'other' heritage

*HispPerCap* - per capita income for people w/Hispanic heritage

*NumUnderPov* - # of people under the poverty level

*PctPopUnderPov* - % of people under the poverty level

*PctLess9thGrade* - % of people 25+ with l< 9th grade education

*PctNotHSGrad* - % of people 25+ not high school graduates

*PctBSorMore* - % of people 25+ with bachelor’s degree or higher education

*PctUnemployed* - % of people 16 and older[[4]](#footnote-4) and unemployed

*PctEmploy* - % of people 16+ who are employed

*PctEmplManu* - % of people 16+ employed in manufacturing

*PctEmplProfServ* - % of people 16+ employed in professional services

*PctOccupManu* - % of people 16+ employed in manufacturing[[5]](#footnote-5)

*PctOccupMgmtProf* - % of people 16+ employed in management or professional occupations

*MalePctDivorce* - % of males who are divorced

*MalePctNevMarr* - % of males who have never married

*FemalePctDiv* - % of females who are divorced

*TotalPctDiv* - % of population who are divorced

*PersPerFam* - mean number of people per family

*PctFam2Par* - % of families (w/kids) headed by 2 parents

*PctKids2Par* - % of kids in family housing w/2 parents

*PctYoungKids2Par* - % of kids 4 or older in 2 parent households

*PctTeen2Par* - % of kids age 12-17 in 2 parent households

*PctWorkMomYoungKids* - % of moms of kids <=6 in labor force

*PctWorkMom* - % of moms of kids <18 in labor force

*NumIlleg* - # of kids born to never married

*PctIlleg* - % of kids born to never married

*NumImmig* - total number of people known to be foreign born

*PctImmigRecent* - % of immigrants immigrated w/in last 3 years

*PctImmigRec5* - % of immigrants immigrated w/in last 5 years

*PctImmigRec8* - % of immigrants immigrated w/in last 8 years

*PctImmigRec10* - % of immigrants immigrated w/in last 10 years

*PctRecentImmig* - % of population immigrated w/in last 3 years

*PctRecImmig5* - % of population immigrated w/in last 5 years

*PctRecImmig8* - % of population immigrated w/in last 8 years

*PctRecImmig10* - % of population immigrated w/in last 10 years

*PctSpeakEnglOnly* - % of people speak only English

*PctNotSpeakEnglWell* - % of people do not speak English well

*PctLargHouseFam* - % of family households that are large[[6]](#footnote-6)

*PctLargHouseOccup* - % of occupied households that are large6

*PersPerOccupHous* - mean persons per household

*PersPerOwnOccHous* - mean persons per owner occupied household

*PersPerRentOccHous* - mean persons per rental household

*PctPersOwnOccup* - % of people in owner occupied households

*PctPersDenseHous* - % of persons in dense housing[[7]](#footnote-7)

*PctHousLess3BR* - % of housing units with less than 3 bedrooms

*MedNumBR* - median number of bedrooms

*HousVacant* - # of vacant households

*PctHousOccup* - % of housing occupied

PctHousOwnOcc - % of households owner occupied

PctVacantBoarded - % of vacant housing that is boarded up

*PctVacMore6Mos* - % of vacant housing that has been vacant more than 6 months

*MedYrHousBuilt* - median year housing units built

*PctHousNoPhone* - % of occupied housing units without phone[[8]](#footnote-8)

*PctWOFullPlumb* - % of housing w/out complete plumbing facilities

*OwnOccLowQuart* - owner occupied housing, lower quartile value

*OwnOccMedVal* - owner occupied housing, median value

*OwnOccHiQuart* - owner occupied housing, upper quartile value

*RentLowQ* - rental housing, lower quartile rent

*RentMedian* - rental housing, median rent[[9]](#footnote-9)

*RentHighQ* - rental housing, upper quartile rent

*MedRent* - median gross rent[[10]](#footnote-10)

*MedRentPctHousInc* - median gross rent as % of household income

*MedOwnCostPctInc* - median owners cost as % of household income - for owners w/mortgage

*MedOwnCostPctIncNoMtg* - median owners cost as % of household income - for owners w/out mortgage

*NumInShelters* - # of people in homeless shelters

*NumStreet* - # of homeless people counted in the street

*PctForeignBorn* - % of people foreign born

*PctBornSameState* - % of people born in same state as currently living

*PctSameHouse85* - % of people living in same house as 1985[[11]](#footnote-11)

*PctSameCity85* - % of people living in the same city as 198511

*PctSameState85* - % of people living in the same state as 198511

*LemasSwornFT* - # sworn full time police officers

*LemasSwFTPerPop* - sworn full time police officers per 100K population

*LemasSwFTFieldOps* - # sworn full time police officers in field operations12

*LemasSwFTFieldPerPop* - sworn full time police officers in field operations[[12]](#footnote-12) per 100K population

*LemasTotalReq* - total requests for police

*LemasTotReqPerPop* - total requests for police per 100K popuation

*PolicReqPerOffic* - total requests for police per police officer

*PolicPerPop* - police officers per 100K population

*RacialMatchCommPol* - a measure of the racial match between the community and the police force[[13]](#footnote-13); High values indicate proportions in community and police force are similar

*PctPolicWhite* - % of police that are Caucasian

*PctPolicBlack* - % of police that are African American

*PctPolicHisp* - % of police that are Hispanic

*PctPolicAsian* - % of police that are Asian

*PctPolicMinor* - % of police that are minority of any kind

*OfficAssgnDrugUnits* - # officers assigned to special drug units

*NumKindsDrugsSeiz* - # of different kinds of drugs seized

*PolicAveOTWorked* - police average overtime worked

*LandArea* - land area in square miles

*PopDens* - population density in persons per square mile

*PctUsePubTrans* - % of people using public transit[[14]](#footnote-14)

*PolicCars* - # of police cars

*PolicOperBudg* - police operating budget

*LemasPctPolicOnPatr* - % of sworn full time police officers on patrol

*LemasGangUnitDeploy* - gang unit deployed (0 means NO, 1 means YES, 0.5 means Part Time)

*LemasPctOfficDrugUn* - % of officers assigned to drug units

*PolicBudgPerPop* - police operating budget per population

Goal values:

*ViolentCrimesPerPop* - total number violent crimes per 100K popuation[[15]](#footnote-15)

### Limitations

Predictive and goal attributes (the numeric data in the dataset) was normalized into the decimal range 0.00-1.00 using an equal-interval binning method. Through this process, attributes retain their distribution and skew. Normalization also preserves the approximate ratios values within an attribute; however, it does not preserve relationships between attribute values. Hence, it is not possible/meaningful to compare the values for attributes such as whitePerCap against blackPerCap.

The values for the Per Capita Violent Crimes attribute was calculated using population data as well as the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault. A few states, many of which are in the Midwest, do not count rapes in the same manner as the others. This resulted in missing values for rapes and a skewed calculation of Per Capita Violent Crimes. These cities were therefore not included in the final dataset.

Information used for the LEMAS related attributes was limited to police departments with a minimum of 100 officers. A few smaller police departments were used as a random sample.

Communities that could not be found in both census and crime dataset were not included in this dataset. Since many communities in the census were missing LEMAS data, this eliminated those communities, limiting the data selection.

## Objective of Analysis

If changes in crime rates can be predicted based on changes of highly correlated indicators from the community, resources could be better allocated to help local governments in addressing crime occurring in their jurisdiction. The objective of this analysis is to explore other indicators of communities along with its prevalence of violent crime to determine what relationships exist among these indicators and suggest further areas of study. We will use utilize techniques such as classification, regression analysis, frequency item sets, and association rules, to analyze the data set.

## Risks

With a dataset this large, there is a high risk that we will not be able to filter out the “noise” to find a definite answer to our question. There are such a large number of factors contributing to any trends in the data that it may be hard to isolate which is actually the actual cause and what may be red herrings. To be clear, we are aiming to determine correlation and are not confident we will be able to determine causation without a time-lapse analysis. For example, without also utilizing FBI Uniform Crime Report datasets from previous years, we would not be able to analyze how the year-to-year changes in attributes may have resulted in a change to the violent crimes result.

Another risk we may run into is expectations of trends. While people strive to be bias free in their research there are certain ideas that seem intuitive about crime and criminal activity in general. For example, some people believe there is more crime in areas of lower income, but this may not be true among all races, geographical areas, etc. We will need to be diligent in minimizing assumptions and avoiding trying to get the data to match an assumption when no correlation exists. Another significant risk is the accuracy of the data set we will be using. The FBI uses the UCR Statistics; However, that data is first collected by local law enforcement agencies, which may not have the same reporting criteria or standards. This possible discrepancy is a point of discussion among many crime studies.

# RELATED WORK

There has been many studies and papers on the importance of sharing information among law enforcement agencies as well as why the sharing of data is lagging. However, past efforts to solve this issue have failed and the problem still plagues law enforcement. The challenge in the past has simply been how to collect, share, aggregate, and standardize data. This challenge remains today as law enforcement agencies continue to use different standards in reporting crime. The Federal Bureau of Investigation’s Uniform Crime Report data has been the more extensive effort to solve this problem to date. It is not without flaws and oversights, and the interpretation/analysis of this data is highly contentious making all the more difficult for law enforcement agencies to benefit from it. There is also a significant effort to interpret the data into the more standard UCR, which smaller stations may not have the manpower to support. Since there are still large obstacles to overcome in the sharing of data, the analysis of the data becomes cumbersome when looking in several location for the type of data represented in this dataset. Other studies have looked at single types of data or data for a limited number of communities, but not as extensive research could be accomplished for data such as that in this dataset.

# METHODOLOGY

We will begin by taking an exploratory data analysis approach. We will first visualize the data we have gathered and then begin comparing multiple variables. By comparing multiple variables through histograms, scatter plots, and various other visualization trends, we will note any obvious patterns by examining their two-way interactions. We will try to describe this relationship with a type of relationship that best fits. Once we have completed most of our exploratory data analysis, we will begin to perform testing and cross-validation to determine if classification can be successfully completed on the data. Finally, we will attempt to analyze frequent itemsets to obtain more information about trends in the dataset.

## Preprocessing

### Normalization

As previously mentioned, the numeric data in the dataset was normalized before being presented for consumption by others. The attributes retained their distribution and skew as well as preserved the approximate ratios for values within an attribute.

### Support

To complete the frequent itemset analysis we also determined the best value to use for support. With support of 0.01, the relative item frequency graph (see Figure 1) showed there were that were near 0 (zero) and were not frequent. Increasing the support to 0.2 would have eliminated many of the possible combinations and left us with very few itemsets to work with. Therefore, we chose to check the relative item frequency graph for a support value of 0.1 (see Figure 2). This second graph shows a variety of itemsets without including many with low frequency. Hence, we used the support value of 0.1 in our analysis.

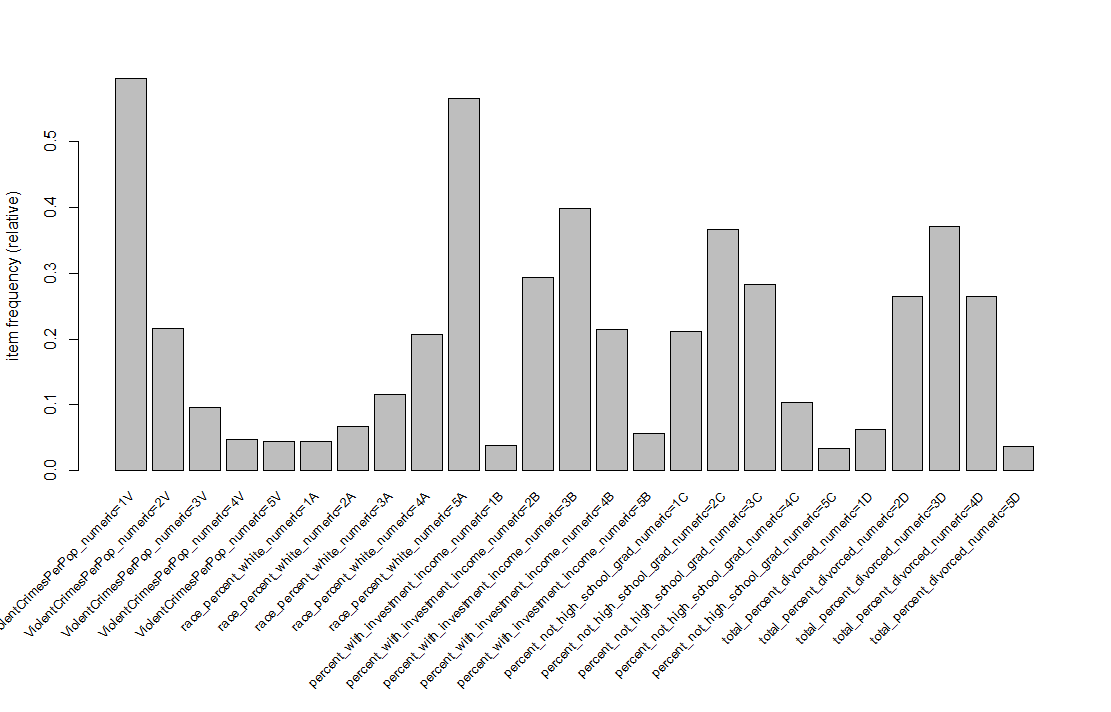


Figure 1. Item Frequency Support of 0.01

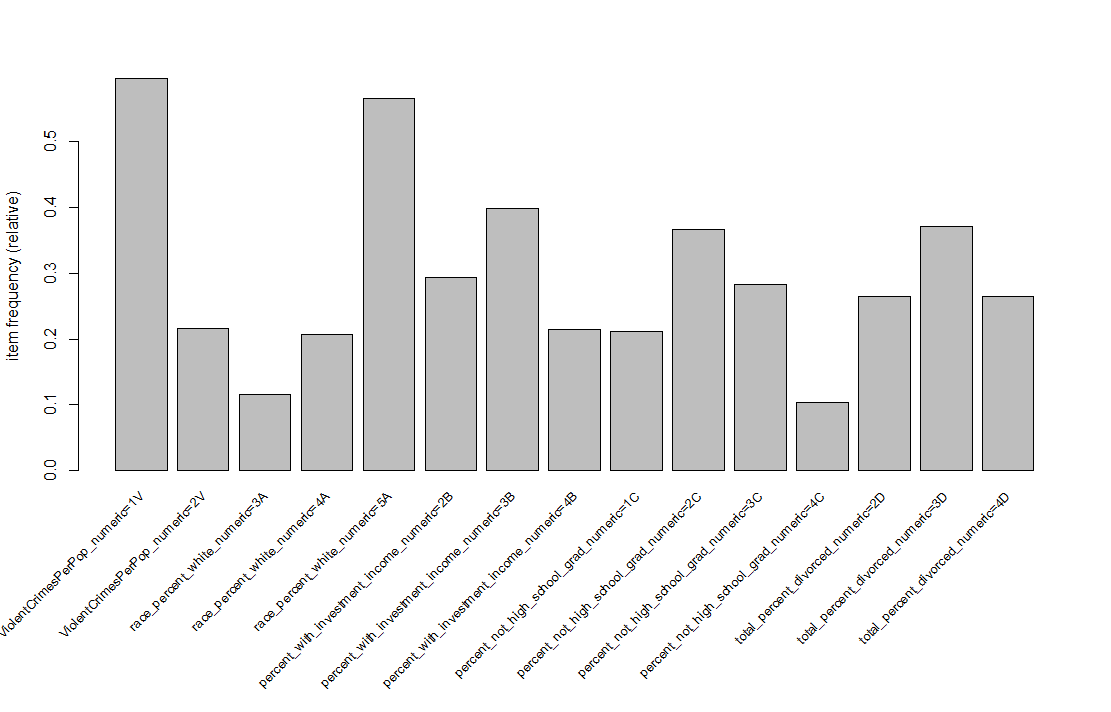


Figure 2. Item Frequency Support of 0.1

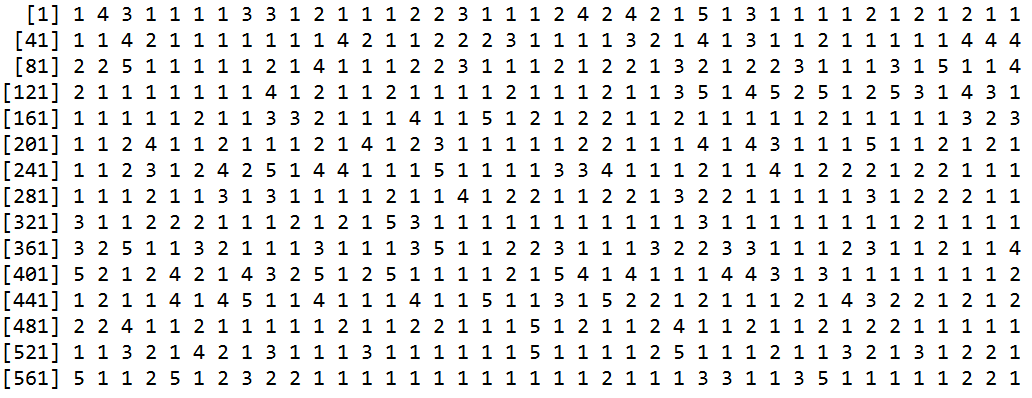
### Binning

Data binning, sometimes referred to as bucketing, is a technique used in data pre-processing to accommodate algorithms that use categorical rather than continuous variables. Since in our case we wish to not only perform classification, but also frequent itemset analysis, binning will be necessary for the attributes involved. When binning, the values for the attributes are each categorized into a group or “bin” which represents that field. There are four common methods to bin field values: equal width binning, equal frequency binning, binning by clustering, and binning based on predictive value. Since many of the values we are working with are percentages and have been normalized to values between 0.00 and 1.00, we felt the most appropriate technique was equal width.

#### Binning

In equal width binning, the attribute values are divided into *k* categories of equal width. While this is not one of the preferred methods of binning since outliers may influence the width of the bins, we thought it would work well with our data since our attribute values are representative of percentages. We binned the attribute values into groups of 10% (*k* of 10) as well as 20% (*k* of 5). Using the attribute ViolentCrimesPerPop as an example, with *k* of 5 (each bin representing 20%) Table 1 shows a sampling of the post-binned data and Figure 3 shows a histogram of the values binned. In addition, with *k* of 10 (each bin representing 10%) Table 2 shows a sampling of the post binned data and Figure 4 shows a histogram the values binned.

Table 1. Sample of *k*=5 Bin of ViolentCrimesPerPop Data



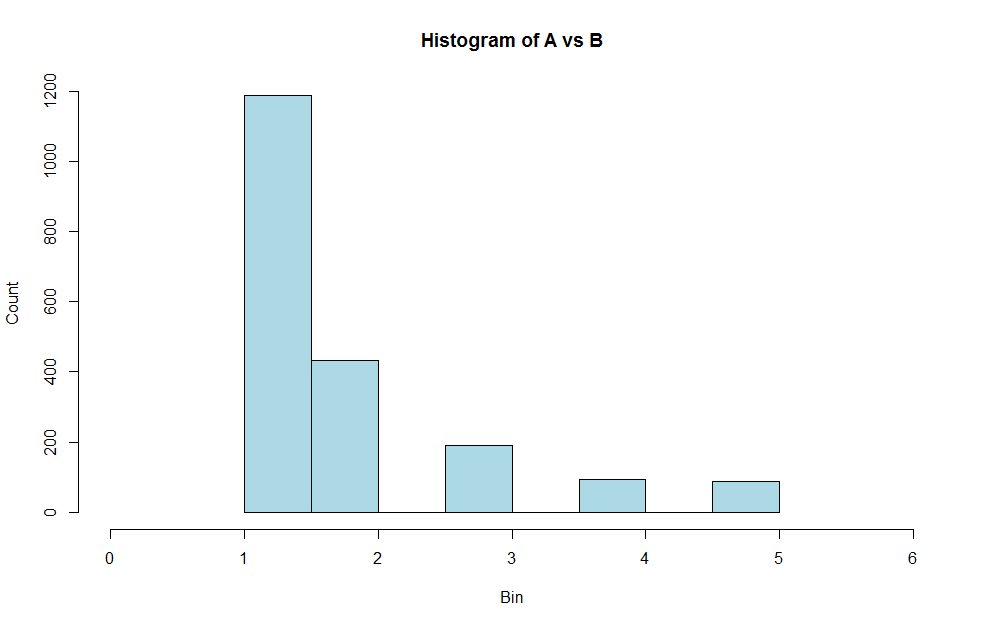
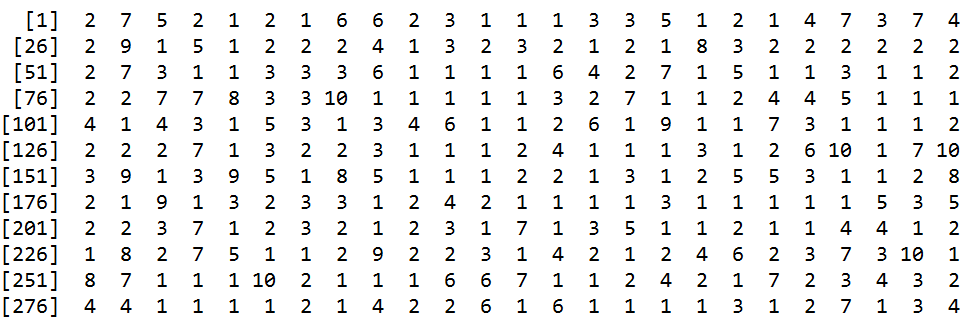


Figure 3. Histogram of Equal Width Binned Weight *k* = 5

Table 2. Sample of *k*=10 Bin of ViolentCrimesPerPop Data



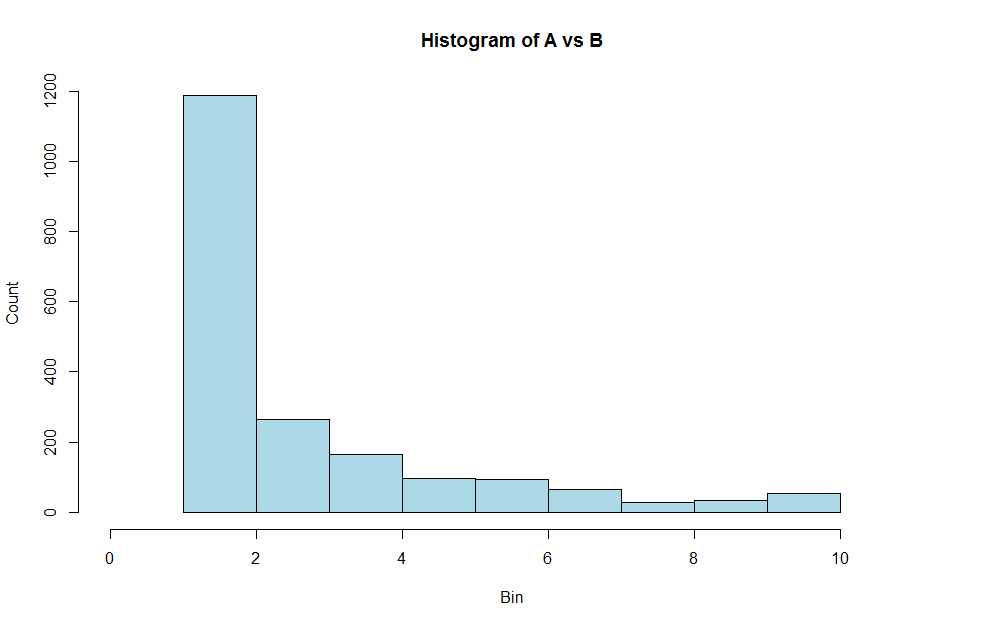


Figure 4. Histogram of Equal Width Binned Weight *k* = 10

## Experiment Design, Tools, & Approaches

### Classification

#### Holdout Method

The Holdout Method is a type of cross validation for classification. In the method, the entire data set is randomly partitioned into two independent sets of specified size: the training set and the test set. The training set is used for model construction and the test set is used to evaluate the accuracy of the constructed model. In the case of our analysis, the data was partitioned with 70% in the training set and 30% in the test set. Both sets contain examples of each classification type.

#### Decision Tree

Decision Tree classification uses a flowchart-like tree structure for classification. In the tree, each test on an attribute is represented by an internal tree node, each outcome of the attribute test is represented by a tree branch, and each tree leaf node has a classification label. A path from the root to a leaf node is a representation of a classification rule.

The tree is constructed in a divide-and-conquer manner with no backtracking. The training examples are all at the root at the start of tree construction and are partitioned reclusively based on the provided selected attributes as the construction proceeds. Partitioning ends when all the samples for a given node belong to the same class, there are no remaining attributes for further partitioning, and there are no samples left to partition.

### Frequent Itemsets – Arules

The objective of frequent pattern analysis is to find inherent regularizes in the data, which reveal intrinsic and important properties of the dataset. Frequent pattern analysis is achieved through use of itemsets, a set of one or more items. The relative support of an itemset is the fraction of transactions that contain the itemset out of the total number of itemsets. If the support for an itemset is not less than a specified threshold, the itemset is considered frequent.

The statistical computing environment and language R contains a package ‘arules’ to assist in frequent itemset or association rule mining. The two specific tools in the arules package we used in our analysis: Apriori and Eclat.

#### Apriori

Apriori in the R arules package uses the Apriori algorithm to mine frequent datasets or association rules. The algorithm conducts level-wise searches for frequent itemsets as follows:

Ck: Candidate itemset of size k

Lk : frequent itemset of size k

L1 = {frequent items};

for (k = 1; Lk != Ø; k++) do begin

Ck+1 = candidates generated from Lk;

for each transaction *t* in database do

increment the count of all candidates in Ck+1 that are contained in *t*

Lk+1 = candidates in Ck+1 with min\_support

end

return Uk Lk;

While there are a few disadvantages of the Apriori algorithm, it requires multiple scans of the database and the candidate generation can result in very large candidate sets, we felt it would still produce acceptable results for our analysis.

#### Eclat

Eclat in the R arules package uses the Eclat algorithm to mine frequent datasets. The algorithm uses bottom-up lattice traversal and simple intersection operations for equivalence class clustering. The algorithm is defined recursively with the initial call using all the single items with their transaction ids (tids) and each recursive call examines the intersections of pairs of tids to generate new candidates. Support is determined for any k-itemset by intersecting tid-lists of two of its (k-1) subsets.

While the Eclat algorithm may take more time for computing intersections and may require large amounts of space to store candidate sets, it is very fast in support counting. Since our analysis would involve large amounts of data, but not necessarily large number of candidate sets or intersections, we felt it would aid in our analysis.

# EVALUATION METHODOLOGY

## Evaluation Metrics

### Classification

#### Accuracy and Error Rate

Accuracy is calculated as the percentage of test samples correctly calculated (TP is true positive, TN is true negative):

Error rate is calculated as the opposite, or 1 - accuracy (FP is false positive, FN is false negative):

#### Sensitivity and Specificity

Sensitivity is calculated as the true positive (TP) recognition rate:

Specificity is calculated as the true negative (TN) recognition rate:

Accuracy can be written as a function of both sensitivity and specificity:

#### Precision and Recall

There is an inverse relationship between precision and recall.

Precision is measured as a percentage of the samples classified with a positive label that are actually positive, or exactness:

Recall is measured as a percentage of positive samples actually classified with a positive label, or completeness.

A perfect score would be 1.0 or 100%.

#### F-Measures

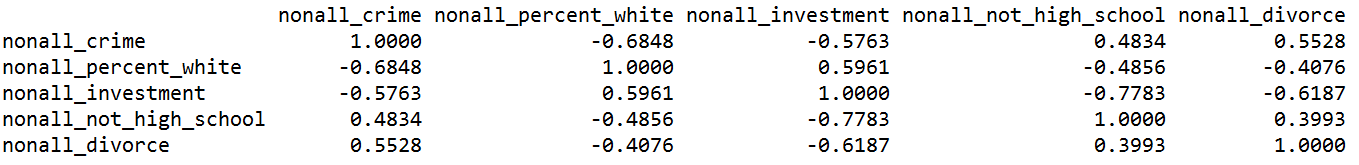
F-measure is a type of accuracy measurement, which takes into account both precision and recall, with the resulting score assigned is between 0 and 1.

F-measure can also be a weighted measurement as follows:

#### Correlation Values

Using the p-value correlation values, the probability of observing the same result as were observed in this dataset in the future can be determined. Since p-value is a probability the value will fall between 0 and 1, with a smaller value favoring rejection of the hypothesis and a larger value favoring excepting the hypothesis. The calculated correlation values for this analysis can be found in Table 3.

Table 3. Correlation Values



### Frequent Itemsets

#### Support

The support value of an itemset is defined as the proportion of transactions in the database that contain the itemset. It can be written as supp(*X*), where *X* is an itemset.

#### Confidence

The confidence value of a rule is the proportion of the transactions that contain an itemset, that also contain another non-intersecting itemset of the same dataset. It can be written as conf(*X* => *Y*) = supp(*X* U *Y*) / supp(*X*), where *X* and Y are itemsets such that *X* ∩ *Y* = Ø and *X* => *Y* is a rule.

# RESULTS

## Classification

#### Decision Tree

A Decision Tree was constructed from the training set using binning of *k* of 10 (see Figure 5) the tree seemed simplistic compared to others we had seen. When examining the below Figure 6 of the relative error and complexity point (CP) the simplistic nature of the tree made sense. After the size of the tree reached three, the CP stabilized to a degree with respect to relative error. The resulting Decision Tree had very few leaf nodes despite numerous examples of multiple classification types in the training data. The Decision Tree confusion matrix shown in Table 3 also confirms the trouble the Decision Tree classification had trouble classifying half of the types, showing no classifications of 6 through 9 and few classifications of 10 for any of the test data.

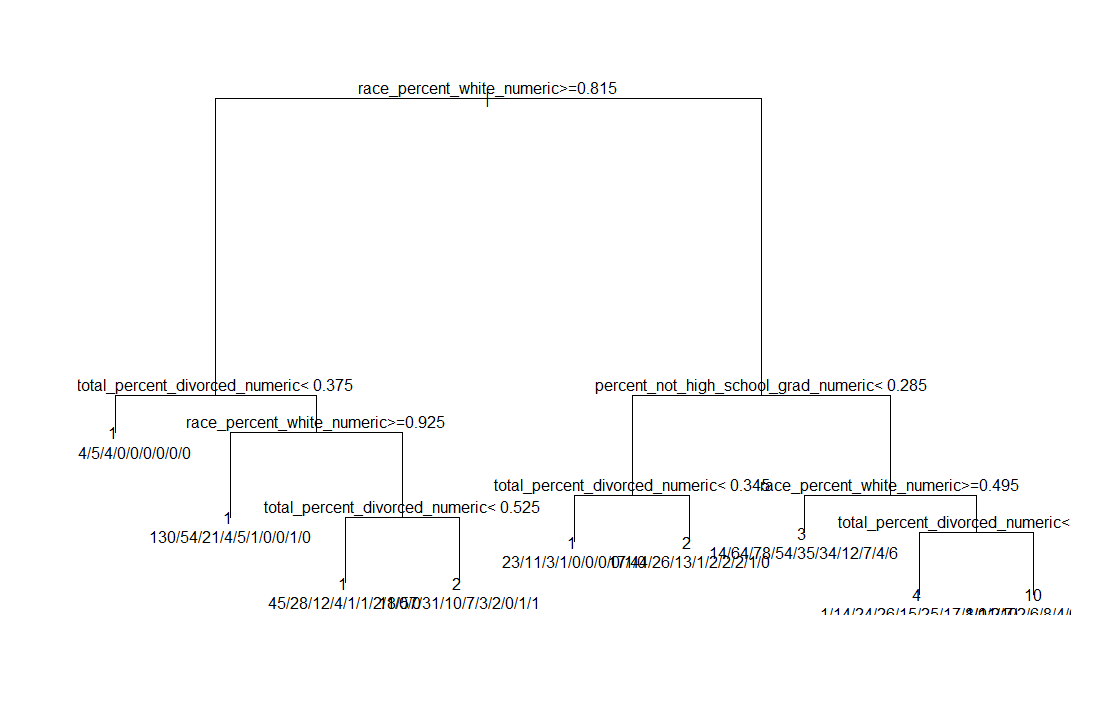


Figure 5. Classification Decision Tree – Binning *k* = 10

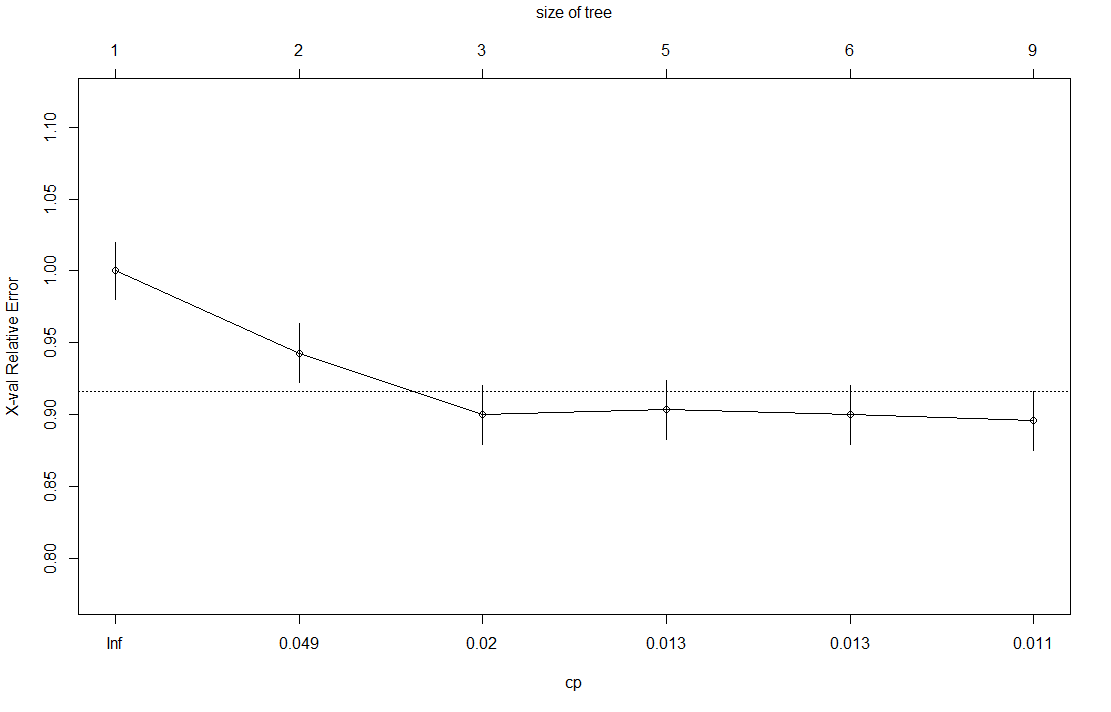
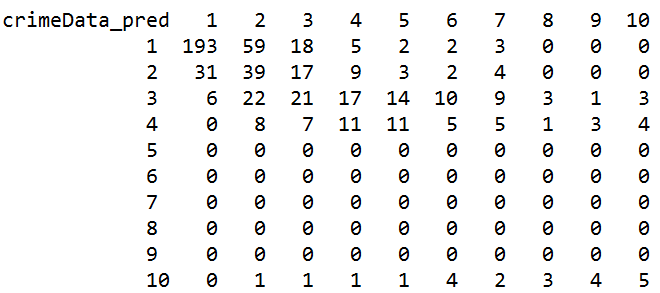


Figure 6. Relative Error and CP – Binning *k* = 10

Table 4. Confusion Matrix Results – Binning *k* = 10



Another Decision Tree was constructed from the training set using binning of *k* of 5 (see Figure 7) the tree again seemed simplistic. When we examined graph of the relative error and CP (Figure 8) the simplicity of the tree again made sense. After the size of the tree reached three, the CP stabilized with respect to relative error. The resulting Decision Tree had less leaf nodes than expected despite numerous examples of multiple classification types in the training data. The Decision Tree confusion matrix shown in Table 4 also confirms the trouble the Decision Tree classification had difficulty classifying many types, showing no classifications of 3 or 4 and few classifications of 5 for any of the test data.

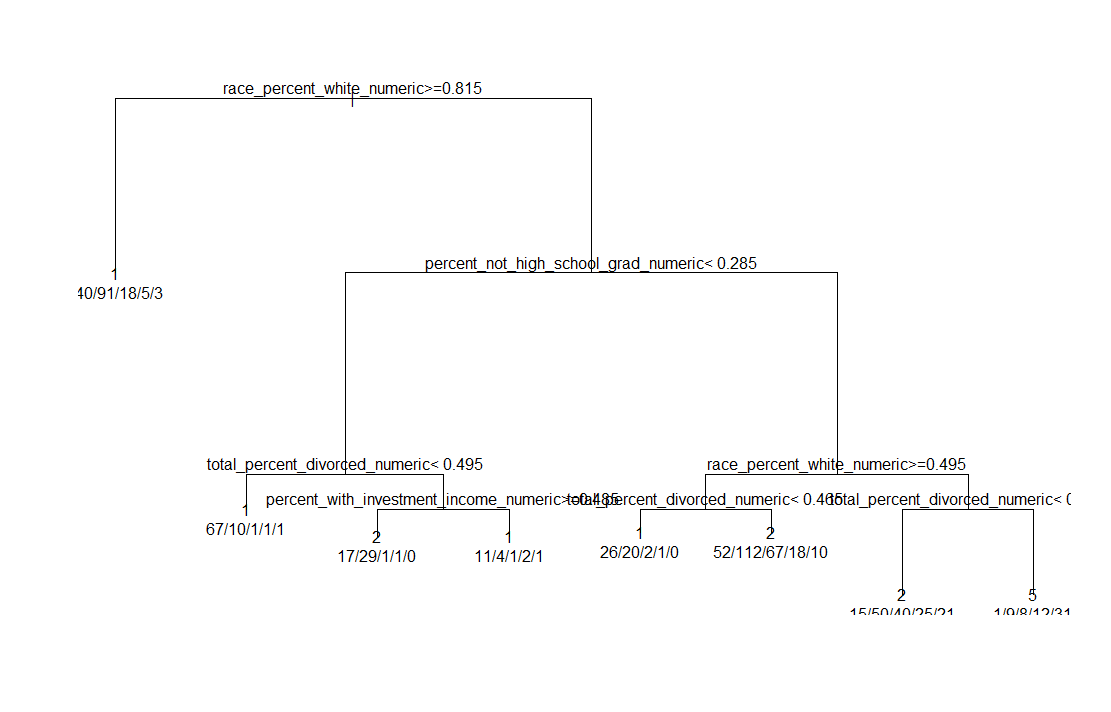


Figure 7. Classification Decision Tree – Binning *k* = 5

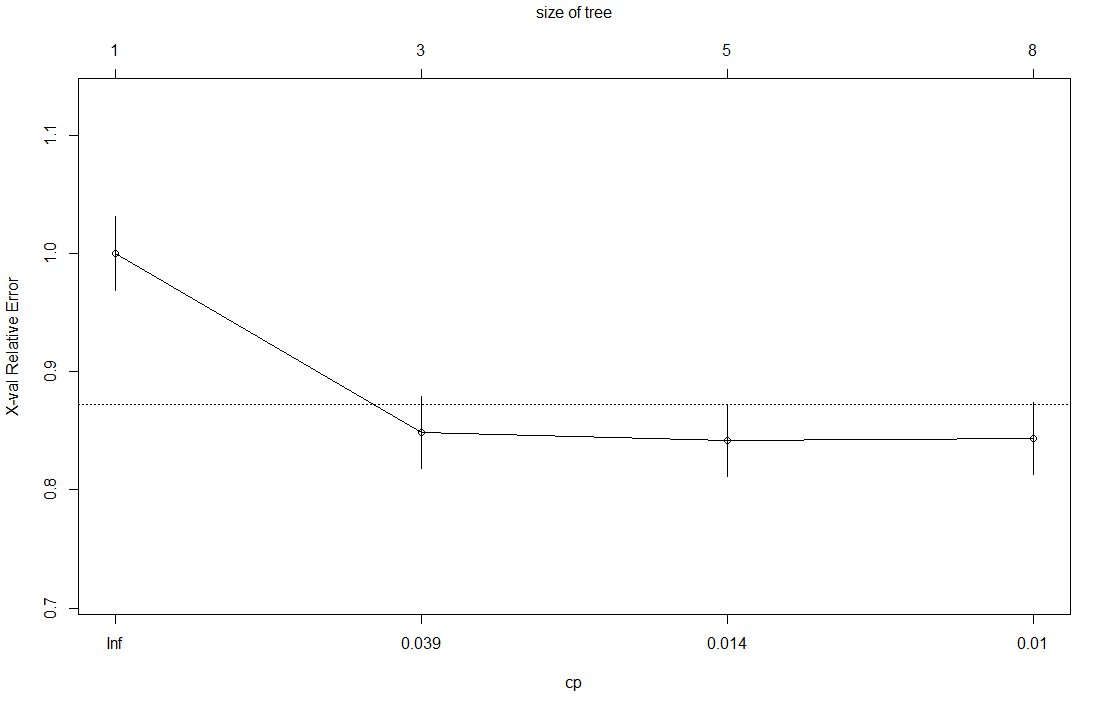
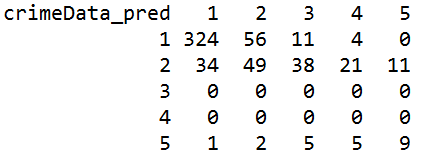


Figure 8. Relative Error and CP – Binning *k* = 5

Table 5. Confusion Matrix Results – Binning *k* = 5



#### Accuracy and Error Rate

The overall accuracy for the *k* of 10 binned data could be calculated using Table 3 as follows:

(193 + 39 + 21+ 11 + 0 + 0 + 0 + 0 + 0 + 5) / 570

= 0.467 ≈ 47%

The overall error rate for the *k* of 10 binned data could be calculated using Table 3 as follows:

(37 + 90 + 43 + 32 + 31 + 23 + 23 + 7 + 8 + 7) / 570

= 0.528 ≈ 53%

With an error rate greater than the accuracy, we determined it was futile to continuing to calculate the rest of the classification evaluation metrics for *k* of 10 binned data. As inaccurate and error prone as the *k* of 10 binned data was obviously at determining classification, it would not provide us useful results. However, we continued the evaluation with the *k* of 5 binned data.

The overall accuracy for the *k* of 5 binned data could be calculated using Table 4 as follows:

(324 + 49 + 0 + 0 + 9) / 570 = 0.67 ≈ 67%

The overall error rate for the *k* of 5 binned data could be calculated using Table 4 as follows:

(35 + 58 + 54 + 30 + 11) / 570 = 0.33 ≈ 33%

With an accuracy near 2/3 and an error rate closer to 1/3, the results for *k* of 5 binned data were much better than those of the *k* of 10 binned data.

#### Sensitivity and Specificity

The sensitivity for each value for *k* of 5 binned data classification can be calculated using Table 4 as follows:

1: 324 / 395 = 0.82 ≈ 82%

2: 49 / 103 = 0.32 ≈ 32%

3 & 4: 0 / 0 so they cannot be calculated

5: 9 / 22 = 0.41 ≈ 41%

The specificity for each value for *k* of 5 binned data classification can be calculated using Table 4 as follows:

1: 140 / 175 = 0.8 ≈ 80%

2: 359 / 417 = 0.861 ≈ 86.1%

3: 516 / 570 = 0.905 ≈ 90.5%

4: 540 / 570 = 0.95 ≈ 95%

5: 537 / 570 = 0.98 ≈ 98%

#### Precision and Recall

The precision for each value for the *k* of 5 classification can be calculated using Table 4 as follows:

1: 324 / (324 + 106) = 0.75 ≈ 75%

2: 49 / (49 + 162) = 0.23 ≈ 23%

3: 0 / (0 + 54) = 0%

4: 0 / (0 + 30) = 0%

5: 9 / (9 + 24) = 0.273 ≈ 27.3%

The recall for each value for the *k* of 5 classification can be calculated using Table 4 as follows:

1: 324 / (324 + 71) = 0.82 ≈ 82%

2: 49 / (49 + 104) = 0.32 ≈ 32%

3: 0 / (0 + 0) so it cannot be calculated

4: 0 / (0 + 0) so it cannot be calculated

5: 9 / (9 + 13) = 0.41 ≈ 41%

#### F-Measures

The F-measure for each value for the *k* of 5 classification can be calculated using Table 4 as follows:

1: (2 \* 0.75 \* 0.82) / (0.75 + 0.82) = 0.78 ≈ 78%

2: (2 \* 0.23 \* 0.32) / (0.23 + 0.32) = 0.27 ≈ 27%

3: cannot be calculated since recall could not be calculated

4: cannot be calculated since recall could not be calculated

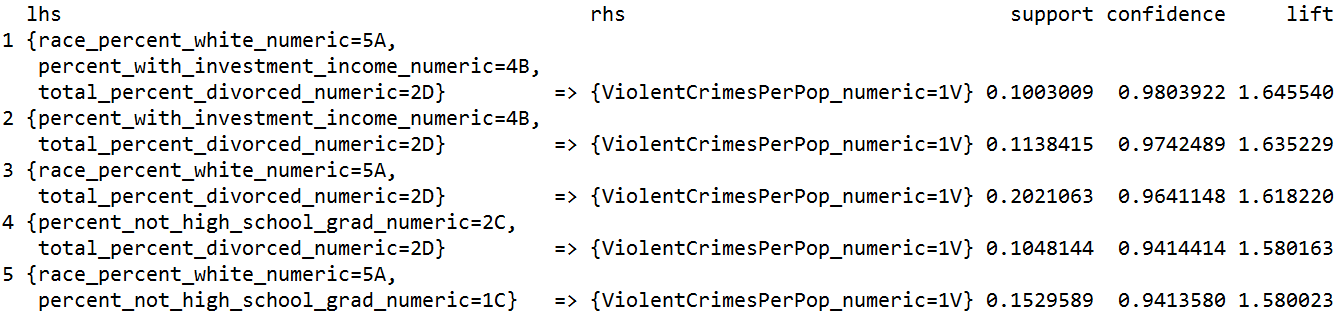
5: (2 \* 0.828 \* 0.41) / (0.828 + 0.41) = 0.33 ≈ 33%

## Frequent Itemsets

### Apriori

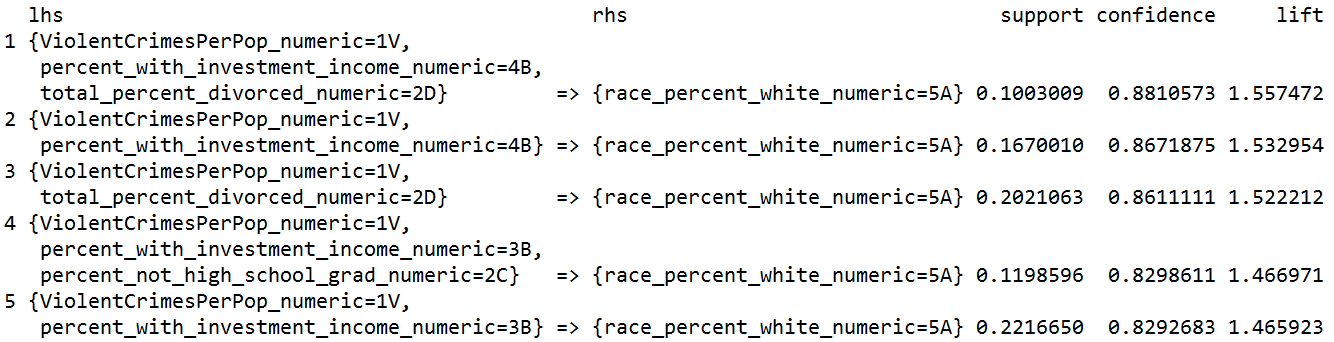
Filtering the Apriori results for any ViolentCrimesPerPop values in contained in the right-hand side (RHS) only produced results for only 1V or the bin containing 0 to 20% (see Table 5). All other values for ViolentCrimesPerPop failed to produce any frequent itemsets. We also tried running the algorithm with smaller bins (*k* of 10) to see if better results were produced. Unfortunately, this produced fewer results with no frequent itemsets for any ViolentCrimesPerPop bins.

Table 6. Apriori ViolentCrimesPerPop RHS Confidence Sort



Since there were so few values of ViolentCrimesPerPop that produced results for the RHS, we decided to try filtering the Apriori results for any ViolentCrimesPerPop values in contained in the left-hand side (LHS). Unfortunately, the only results produced were for 1V or the bin containing 0 to 20% (see Table 6). All other values for ViolentCrimesPerPop failed to produce any frequent itemsets. Re-running the algorithm with smaller bins (*k* of 10) as before also produced no frequent itemset results.

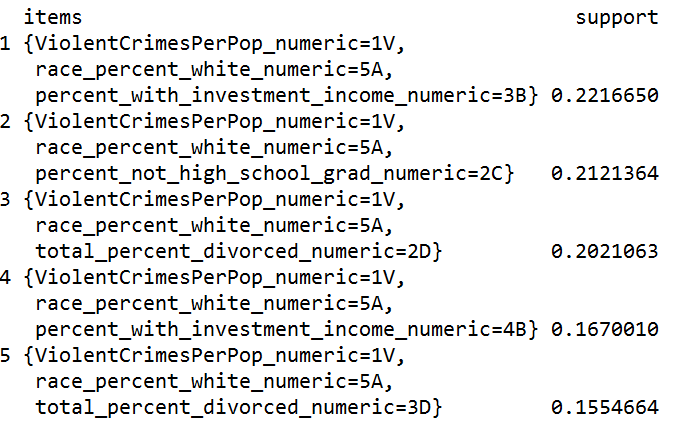
Table 7. Apriori ViolentCrimesPerPop LHS Confidence Sort



### Eclat

Filtering the Eclat results for ViolentCrimesPerPop values was a little more challenging with Eclat. There is no confidence value as in Apriori, so support was used instead. The Eclat results were filtered for the top five support values regardless of value. As with Apriori, all the results were for 1V or the bin containing 0 to 20% (see Table 7). Re-running the results using *k* of 10 bins as we did with Apriori produced no better results.

Table 8. Eclat ViolentCrimesPerPop Support Sort

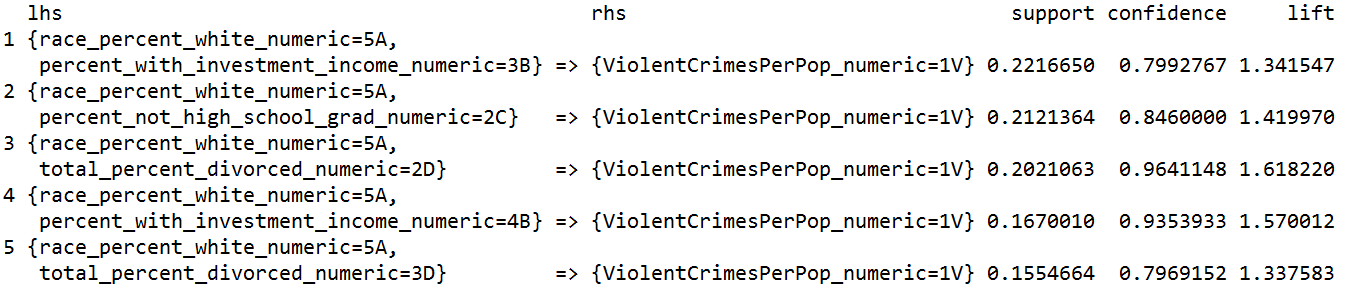


### Apriori vs Eclat

For a better comparison between the Apriori and Eclat algorithm results, we re-ran the Apriori algorithm sorting the results by support as was done with the Eclat algorithm.

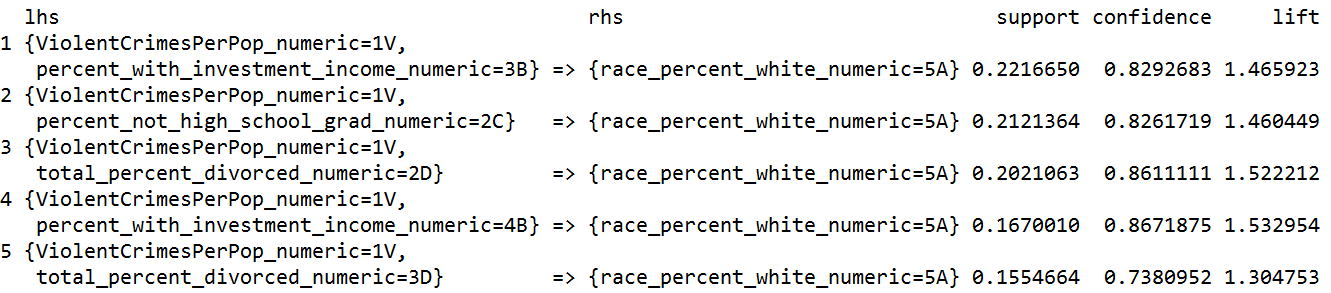
Filtering the Apriori results for any ViolentCrimesPerPop values contained in the RHS gave results very similar to the Eclat algorithm (see Table 8).

Table 9. Apriori ViolentCrimesPerPop RHS Support Sort



Filtering the Apriori results for any ViolentCrimesPerPop values contained in the LHS also produced results very similar to the Eclat algorithm (see Table 9).

Table 10. Apriori ViolentCrimesPerPop LHS Support Sort



# DISCUSSION

We had originally hypothesized that factors such as percentage of police per capita, per capita income, and percentage of population with Bachelor’s degrees or higher would be significant in terms of violent crimes per capita. It turned out that these are indirectly related to the attributes that we found most significant.

Also in our original hypothesis was the percentage of police per capita. This too was a poor predictor of the incidents of violent crime per capita. We believe this result is biased in that the communities with police departments that had less than 100 police officers were not included in the data set minus a few randomly sampled examples.

Investment income was one of the attributes we determined to a factor in violent crimes per capita. It is indirectly related to per capita income in that investment income is usually thought of as the difference between income and expenses. It is also common for areas with high incomes per capita to be areas with a very high cost of living. Therefore, if we were to use purely per capita income the perceived wealth/income would be skewed in this scenario. Investment income is a better indicated of how well a community is doing relative to its cost of living.

In addition, there was not a significant separation between communities with a more educated resident base beyond a high school degree. It is possible that a high school degree was sufficient for a comfortable standard of living in the 1980s and 1990s. It would be interesting to explore this today and see if the threshold has increased to a bachelor’s degree.

# CONCLUSIONS

In our analysis, we saw a slight proportional relationship between high violent crime per capita and high percentage of divorcees as well as a high percentage of the population without a high school degree. We also saw a slight inverse proportional relationship between high violent crime per capita and communities with higher percentage of investment income as well as largely Caucasian communities.

When there was lower percentage of the population without a high school degree, a lower percentage of divorcees and higher percentage of Caucasians were found frequently. In addition, a higher percentage of Caucasians in the community was frequently found associated with a higher percentage of the population with investment income. High percentages of Caucasians in a community were also many times found with low to mid percentages of divorcees in the community.

The error rate for the classification was very high. There were also very few results in our frequent itemset analysis with the low percentage of violent crime per capita the only bin producing results. Even with the large size of the dataset, the fact that there was so much error and so little results does not give us confidence that the attributes we selected do have a strong case for influencing violent crime per capita.

We feel it would be useful to extend this study to analyze how change in any specific attribute is correlated with an increase or decrease in violent crime per capita. This would require additional datasets in preceding and/or succeeding years. This additional data would also help in determining outliers in the data, which could be used to improve the results of any analysis.

# REFERENCES

1. Larose, D. and Larose, C.D. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience, 2014.
2. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
3. U. S. Department of Commerce, Bureau of the Census, Census Of Population And Housing 1990 United States: Summary Tape File 1a & 3a (Computer Files), U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992).
4. U.S. Department of Justice, Bureau of Justice Statistics, Law Enforcement Management And Administrative Statistics (Computer File), U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992)
5. U.S. Department of Justice, Federal Bureau of Investigation, Crime in the United States (Computer File) (1995)
6. Redmond, M. and Baveja, A. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research,* 141 (2002), 660-678.
7. Law Enforcement Information Sharing, *Information Sharing Environment*, Retrieved February 28, 2016 from Information Sharing Environment: https://www.ise.gov/law-enforcement-information-sharing.
8. Mitchell, R. L. It’s criminal: Why data sharing lags among law enforcement agencies, *Computerworld*. Retrieved February 28, 2016, from Computerworld: http://www.computerworld.com/article/2486359/government-it/it-s-criminal--why-data-sharing-lags-among-law-enforcement-agencies.html
9. Miller, P. How Can We Imporve Information Sharing Among Local Law Enforcement Agencies?, *Naval Postgraduate School, Monterey, CA*. Retrieved February 28, 2016: http://www.dtic.mil/dtic/tr/fulltext/u2/a439576.pdf

1. Many missing values [↑](#footnote-ref-1)
2. Income from year 1989 [↑](#footnote-ref-2)
3. Median family income differs from household income for non-family households [↑](#footnote-ref-3)
4. Limited to people 16+ who are in the labor force [↑](#footnote-ref-4)
5. Possible duplicate of information [↑](#footnote-ref-5)
6. Households with six or more people are considered large [↑](#footnote-ref-6)
7. Housing is considered dense with greater than 1 person per room [↑](#footnote-ref-7)
8. Authors noted in 1990 this was rare. [↑](#footnote-ref-8)
9. Census variable H32B from file STF1A [↑](#footnote-ref-9)
10. Census variable H43A from file STF3A - includes utilities [↑](#footnote-ref-10)
11. Five years beforehand [↑](#footnote-ref-11)
12. Police officers in field operations are considered on the street as opposed to administrative etc. [↑](#footnote-ref-12)
13. High values for RacialMatchCommPol indicate racial proportions in community and police force are similar [↑](#footnote-ref-13)
14. Use of public transit limited to commuting purposes [↑](#footnote-ref-14)
15. This is the goal attribute, to be predicted [↑](#footnote-ref-15)