5-10-2018

Data Challenges in Grouping Criminals by Past Behavior

Benjamin J. Blanco
University of Pennsylvania

Ayya A. Elzarka
University of Pennslyvania

Data Challenges in Grouping Criminals by Past Behavior

Keywords
"censoring, clustering, k-means, criminal behavior"

Disciplines
Applied Statistics

Data Challenges in Grouping Criminals by Past Behavior

By

Ayya Elzarka and Ben Blanco

An Undergraduate Thesis submitted in partial fulfillment of the requirements for the

WHARTON RESEARCH SCHOLARS

Faculty Advisor:

Abba Krieger

David Steinberg Professor; Statistics, Marketing, Operations

THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA

MAY 2018

Introduction

The field of criminology has long sought to understand heterogeneity among criminals. Qualitatively, distinctions between non-violent, violent, serial, and opportunist criminals seem to emerge. However, these descriptions are not uniformly supported across the criminology literature.[1] This analysis provides clarity on criminal behavior by quantitatively grouping individuals based on past criminal behavior with supervised and unsupervised methods. Resulting from the analysis, groups do emerge. However, the results of the analysis are constrained by idiosyncrasies in the data. Overall, the importance of understanding data before analysis becomes clear.

This paper begins by motivating the questions and methods of analysis from the criminology and statistics literature. The research question is fleshed out and the data generation process is described. The following data challenges are presented: ascertainment bias, censoring, many Zeroes, duplicates, missing juvenile data for groups of individuals, use of convictions rather than offenses, extreme skewness. Following, the nature of the variables (univariate analysis) and relationships between variables (bivariate analysis) in the dataset are discussed. Finally, unsupervised and supervised methods are described and applied to the data.

Literature Review

As this paper seeks to find insight regarding criminal behavior through means of employing various statistical procedures, it is important to explore literature around both areas:

[1] Nagin and Paternoster 1991

criminology and clustering analysis.

Beginning with criminology, the study of crime specialization examines the relationship between criminal careers, longitudinal sequences of crimes and interventions around a given offender, and key crime statistics. The two main lines of inquiry are the progression of criminal careers and the patterns of criminal behavior among offenders. This research focuses on the second question and seeks to characterize individuals based on counts of their prior criminal arrests.

This research is further motivated by the debate between criminal specialization and criminal versatility, or whether criminals offend in specific crimes or not. In addition to the evidence presented above, supporters of criminal specialization argue that individual psychological characteristics inform specific crimes. This is presented in a framework where individuals' underlying psychologies feed into their interpersonal relationships with others and the violent strategies of their actions.[2] This paper hopes to further this debate by showing that significant patterns of criminal priors exist among offenders.

Clustering techniques have been examined in a criminology context. Since patterns of criminal behavior are likely to be heterogeneous, the focus of clustering is to group individuals' behavior into a manageable number of groups. Individuals in the same group tend to have similar patterns. One study discussed the process to move from raw criminal data to an interpretable mapping of individuals to clusters and using these clusters for predictions of future behavior.[3] In the system, the criminal profile (based patterns of crimes committed), the severity of crimes, the frequency of crimes, and the duration of criminal careers were used to group individuals.

[2]Horning, Salfati, and Crawford 2010
[3]De Bruin et al. 2006

A key focus of the research was calculating differences between groups. Researchers also focused on methods to best represent the final results of the clustering. Additionally, two key problems to address emerged from this research. One problem was the effect of a large number of one-time offenders on the distance calculations. A second problem was the heavy computation required to run the analysis.

Next, this paper looks to a similar analysis in which clustering has been employed within the context of supervised and unsupervised learning. While clustering analysis is utilized frequently across many disciplines, its new extension, in the context of supervised learning is quite distinct from typical practices. There have been many novel methods tried: penalized regression-based clustering, clustering based on proximity values in a random forest, clustering within the context of neural networks based on Graph-based Activity Regularization (GAR) techniques as well as many others. Supervised and semi-supervised methods are known to improve the performance of unsupervised methods through labeling data or the addition of constraints.[4]

Problem/Question

This research seeks to provide empirical evidence for variations in criminal behavior across gender and age. This paper also seeks to understand whether a juvenile criminal career can serve as a precursor to adult criminal behavior. Clustering on criminal history data and analyzing patterns will allow us to make claims about the validity of these differences. As results appear, we seek to measure the robustness of the statistical methods employed.

[4]Pan, Shen, and Liu 2013; Berk 2016; Kilinc and Uysal 2017; Basu, Bilenko, and Mooney 2004

Data

The dataset is of individuals who are being considered for probation within the Pennsylvania criminal justice system. The data were provided by Richard Berk. There were 31 variables available for each instance. The demographic variables, which were unfortunately limited in scope were: age and gender. There were outcome variables on prior juvenile offenses of various types: gun, weapons, sex assault, drug usage and distribution as well as violence charges. A set of variables exists for adult crimes that parallel the information available on juvenile crimes. In addition, there are two supplemental variables: failure to appear in court and absconding charges. Moreover, an indicator variable that specified whether an adult was charged with murder is also available. A snapshot of the dataset with some of the entries appears in Table 21:

	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors
4697	24.22	1	0	0	0	0	FALSE	0	16	3
5596	26.54	1	0	0	0	4	FALSE	0	1	4
5944	22.35	1	0	0	0	0	FALSE	0	11	0
10033	24.12	1	0	0	0	6	FALSE	0	22	9
12571	28.54	1	0	0	0	0	FALSE	0	3	0
13245	23.13	1	0	0	0	0	FALSE	0	6	0

Table 1: Subset of Data

Ideally, one would want these data to either be a census of all individuals with information at each stage in the criminal justice process (initial booking, charges, final sentencing, parole, etc.) or a random sample of such individuals. Instead, the dataset describes a subset of parolees within the criminal population at the time of their parole. In other words, these data are observational and as such this poses some critical issues which this paper will discuss in turn: ascertainment bias, censoring, many zeroes, duplicates, missing juvenile data for groups of individuals, do not have convictions but rather only offenses- there are many citations that

are possibly many accounts for one previous appearance, and extreme skewness.

Ascertainment Bias

Ascertainment bias is "a systematic distortion in measuring the true frequency of a phenomenon due to the way in which the data are collected" and thus affects who is eligible to be in the dataset.[5] This decreases the utility of the analysis as the individuals in the dataset may not represent the general criminal population. Ascertainment bias emerges in the data as only a subset of criminals qualify for probation. Individuals who are extremely "bad" are presumably already jailed and not observed in the dataset. Additionally, first-time offenders who commit extremely severe crimes may also not be eligible for probation and not observed in the dataset. Failing to understand this bias results in spurious conclusions about the nature of criminal behavior. For example, in the data, there is a negative relationship between age and criminality. At first glance, this implies that older individuals are less likely to be criminals. However, given ascertainment bias, it may be the case that older individuals have more time to commit crimes. Additionally, as age increases, it becomes clearer that a given individual is less crime-oriented. This removes older individuals with high levels of criminality, creating a dataset where older individuals are those less prone to crime and creating the negative relationship observed. Failing to recognize ascertainment bias would have led to false claims indicating that a propensity to commit crime decreases with age.

This paper partially deals with the issues by looking at a subset of the population of individuals in the 20 to 30 range to limit the window of time an individual would have to commit crimes.

[5] "Ascertainment Bias" n.d.

This restricts the analysis to individuals before they self-select out of the dataset. Beginning the analysis at age 20 controls for the removal of individuals with severe juvenile records who are not eligible for parole and removes juveniles in the dataset who have not realized the full array of juvenile crimes.

Censoring

Censoring occurs because this dataset takes a snapshot of an individual at a particular point in time. Future behavior is unknown. Specifically, one might expect that younger individual would have fewer crimes, not because of their propensity to commit crimes, but rather because data are only available for a relatively short period of time. This reduces the robustness of conclusions drawn about criminal behavior. Ascertainment bias and age drive censoring in the dataset. Ascertainment bias removes individuals not eligible for parole from the dataset, blinding us to individuals with high propensities to commit crimes. Age is only collected at the time of individuals probation hearing. For a given individual, no information is provided about their current age or the age at which they committed their previous crimes.

Being unable to fully observe the criminal observation or understand the ages at which individuals committed crimes make drawing conclusions difficult. It may be the case that younger individuals in the dataset have the capacity to commit more and different crimes than is observed in the dataset and older individuals have already demonstrated their full capacity. Censoring makes it difficult to compare these individuals as they are at different points in their criminal history. Robust claims could be made if the future crimes of younger individuals and the criminal histories of older individuals were observed. Censoring limits

the data to the number of crimes committed at the time of probation.

Limiting the scope of the analysis to individuals 20-30 years of age addresses this issue by narrowing the window of crime committed to smaller time period than is initially observed in the dataset. This equalizes the time window individuals in the dataset had to commit a crime, regardless of age at probation.

Zeros

The dataset contains first-time offenders with no criminal history and no prior charges. Since the current crime is not observed, in the dataset first-time offenders have values of zero for all crime variables. Seventeen percent of individuals are first-time offenders. Several criminal priors variables contain primarily zero counts. For example, only 0.01% of the criminal in the dataset had a value in anymurder. This can be attributed once again to the ascertainment bias in the dataset as well as the nature of criminal behavior. Following the discussion in the literature review, criminals have a tendency to specialize within classes of crimes and do not broadly commit every single crime. This heterogeneity in criminal behavior lends itself to many zeros across the priors. Moreover, for individuals who do have prior counts among all crimes, they are likely to be considered "severe" criminals and thus not have the opportunity for probation.

The appearance of zeros makes it difficult to find meaningful variation among criminal based on prior criminal charges. For example, given two first-time offenders whose age and gender is the only information provided, it is difficult to form groups that speak to patterns in criminal behavior. To reduce the number of zeros, first-time offenders are removed. The

number of zeros in specific variables is not explicitly addressed but is considered when drawing conclusions.

Duplications

There are 15,369 duplicate rows in the dataset. This comprises twelve percent of the dataset. While there is uncertainty regarding why these observations appear in the dataset, the inclusion of these duplicates would arbitrarily increase the weight of these observations. To remove the additional weight these duplicates would they were simply excluded from the analysis.

Missing Data

The dataset also suffers from systematically missing data which, unaddressed, would have lead to false conclusions. An example of this is the fact that no juvenile criminal records exist for any individual above the age of 47. Looking closer, for certain juvenile crimes, there is no information available for individuals above the age of 30. Similar to the issues with ascertainment bias, this would lead to false conclusions about the relationship between juvenile and adult crimes. As discussed before, the resolution was to proceed with analysis only on individuals between ages 20-30 to remove the effects of this systemic missingness.

Ambiguity of Criminal Charges

For a given crime, multiple charges may be pursued. As a result, individuals can receive multiple charges for the same incident. The dataset only records the number of charges and

this measure is the only proxy of criminality in the dataset. This limits the ability to form claims regarding criminal behavior. Take the following scenario as an example: the dataset observed ten charges of sexual priors for two individuals. One individual committed sexual assault once a year for ten years and the other accumulates ten sexual assault charges in one incident. Acute differences in frequency and recency are not captured in the dataset, making it difficult to analyze the criminal "career" of an individual. This impacts the claims made about the severity of criminals. There is no explicit solution to this problem. Thus, claims about criminal careers are not made following the analysis.

Skewness

Both demographic and criminal prior variables are right-tail skewed towards younger people, men, and lower crime counts which affects the usefulness of the clustering techniques employed. There were a few outliers of particular interest in the data: a 22-year-old male with 444 total charges, a 16-year-old male with 102 prior charges, and a woman aged 69 who is on probation trial with 80 prior criminal charges. The grouping methods used are sensitive to skewness and variation among the variable. For example, K-means clustering places larger weight on highly-skewed variables, changing the resulting groupings. This is problematic clustering techniques will form separate clusters to address the handful of outliers in the data because of their obvious differences instead of creating clusters to distinguish between the average criminal in the dataset. To address this skewness, the data are log-transformed.

Univariate Analysis

As mentioned above, the variables surrounding priors in the dataset are extremely skewed as shown numerically in Tables 2, 3, and 14 and visually in Figure /ref{yrshist}. The overwhelming mode within each of these criminal priors is zero and there are many people with a low number of charges in each criminal domain with a few outliers who possess many charges for a given crime. To reiterate, the source of the skewness is attributable to ascertainment bias and the nature of criminal behavior. Because criminals in the dataset have the option of probation, it is much more likely to see non-severe criminals who have low counts within priors and not many charges across priors as criminals who are "very bad" and have high counts within each prior and across multiple priors are already in jail. Moreover, literature seems to indicate that criminals specialize and tend not to be involved in a large breadth of crimes thus one can expect a large number of zeroes within each crime. At an aggregate level, amongst adult crimes, drug possession is the most prevalent among criminals with murder being the rarest. These observations seem logical in context as drug crimes have lighter sentences and thus individuals with drug charges are more likely to be candidates for probation rather than individuals with charges in the other more serious offenses (i.e.weapons, sexual assault).[6] Amongst juveniles, property theft is the most common crime and sexual assault is the rarest.

Regarding demographics, age is almost normally distributed across ages 20-30 which is the age group the analysis is limited to. The dataset is comprised of 86 % men and 14% women.

[6]Howard 2017

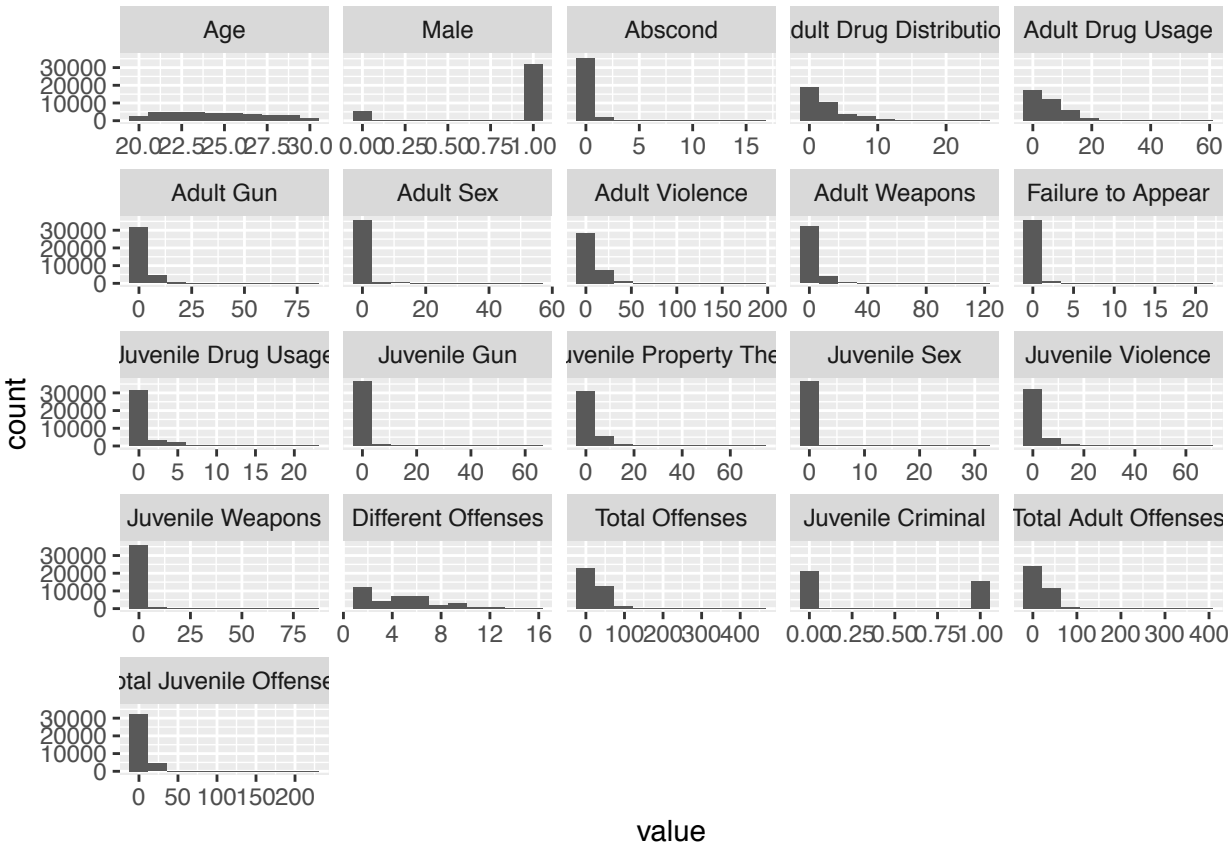

Figure 1: Histograms of all Data

This is an artifact of the fact that the criminal population is mostly male.[7]

	Age	Male	Abscond	Adult Drug Distribution	Adult Drug Usage	Adult Gun	Murder	Adult Sex	Adult Violence	Adult Weapons
1	Min. :20.00	Min. :0.0000	Min. : 0.00000	Min. : 0.000	Min. : 0.000	Min. : 0.000	Mode :logical	Min. : 0.0000	Min. : 0.000	Min. : 0.000
2	1st Qu.:22.17	1st Qu.:1.0000	1st Qu.: 0.00000	1st Qu.: 0.000	1st Qu.: 1.000	1st Qu.: 0.000	FALSE:36365	1st Qu.: 0.0000	1st Qu.: 0.000	1st Qu.: 0.000
3	Median :24.35	Median :1.0000	Median : 0.00000	Median : 1.000	Median : 4.000	Median : 0.000	TRUE :737	Median : 0.0000	Median : 3.000	Median : 0.000
4	Mean :24.57	Mean :0.8597	Mean : 0.09687	Mean : 2.324	Mean : 5.465	Mean : 2.035		Mean : 0.3825	Mean : 7.055	Mean : 2.635
5	3rd Qu.:26.86	3rd Qu.:1.0000	3rd Qu.: 0.00000	3rd Qu.: 4.000	3rd Qu.: 8.000	3rd Qu.: 3.000		3rd Qu.: 0.0000	3rd Qu.: 10.000	3rd Qu.: 4.000
6	Max. :30.00	Max. :1.0000	Max. :16.00000	Max. :25.000	Max. :58.000	Max. :81.000		Max. :54.0000	Max. :188.000	Max. :117.000

Table 2: Summary of Adult Predictors

	Failure to Appear	Juvenile Drug Usage	Juvenile Gun	Juvenile Property Theft	Juvenile Sex	Different Offenses	Total Offenses
1	Min. : 0.0000	Min. : 0.0000	Min. : 0.0000	Min. : 0.000	Min. : 0.0000	Min. : 1.000	Min. : 1.00
2	1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: 0.000	1st Qu.: 0.0000	1st Qu.: 2.000	1st Qu.: 8.00
3	Median : 0.0000	Median : 0.0000	Median : 0.0000	Median : 0.000	Median : 0.0000	Median : 4.000	Median : 18.00
4	Mean : 0.1663	Mean : 0.7044	Mean : 0.2244	Mean : 1.583	Mean : 0.0756	Mean : 4.647	Mean : 24.69
5	3rd Qu.: 0.0000	3rd Qu.: 0.0000	3rd Qu.: 0.0000	3rd Qu.: 2.000	3rd Qu.: 0.0000	3rd Qu.: 7.000	3rd Qu.: 33.00
6	Max. :21.0000	Max. :22.0000	Max. :63.0000	Max. :71.000	Max. :31.0000	Max. :15.000	Max. :444.00

Table 3: Summary of Juvenile Predictors

In order to proceed with the analysis, the log transformation of the data is taken to mediate the extent of the skewness. The results are displayed in Figure 2. While it was impossible to

[7]Rowe, Vazsonyi, and Flannery 1995

	Juvenile Violence	Juvenile Weapons	Juvenile Criminal	Total Adult Offenses	Total Juvenile Offenses
1	Min. : 0.000	Min. : 0.0000	Min. :0.0000	Min. : 0.00	Min. : 0.000
2	1st Qu.: 0.000	1st Qu.: 0.0000	1st Qu.:0.0000	1st Qu.: 7.00	1st Qu.: 0.000
3	Median : 0.000	Median : 0.0000	Median :0.0000	Median : 15.00	Median : 0.000
4	Mean : 1.439	Mean : 0.5082	Mean :0.4219	Mean : 20.18	Mean : 4.535
5	3rd Qu.: 0.000	3rd Qu.: 0.0000	3rd Qu.:1.0000	3rd Qu.: 27.00	3rd Qu.: 6.000
6	Max. :67.000	Max. :83.0000	Max. :1.0000	Max. :388.00	Max. :219.000

Table 4: Summary of Total Offenses

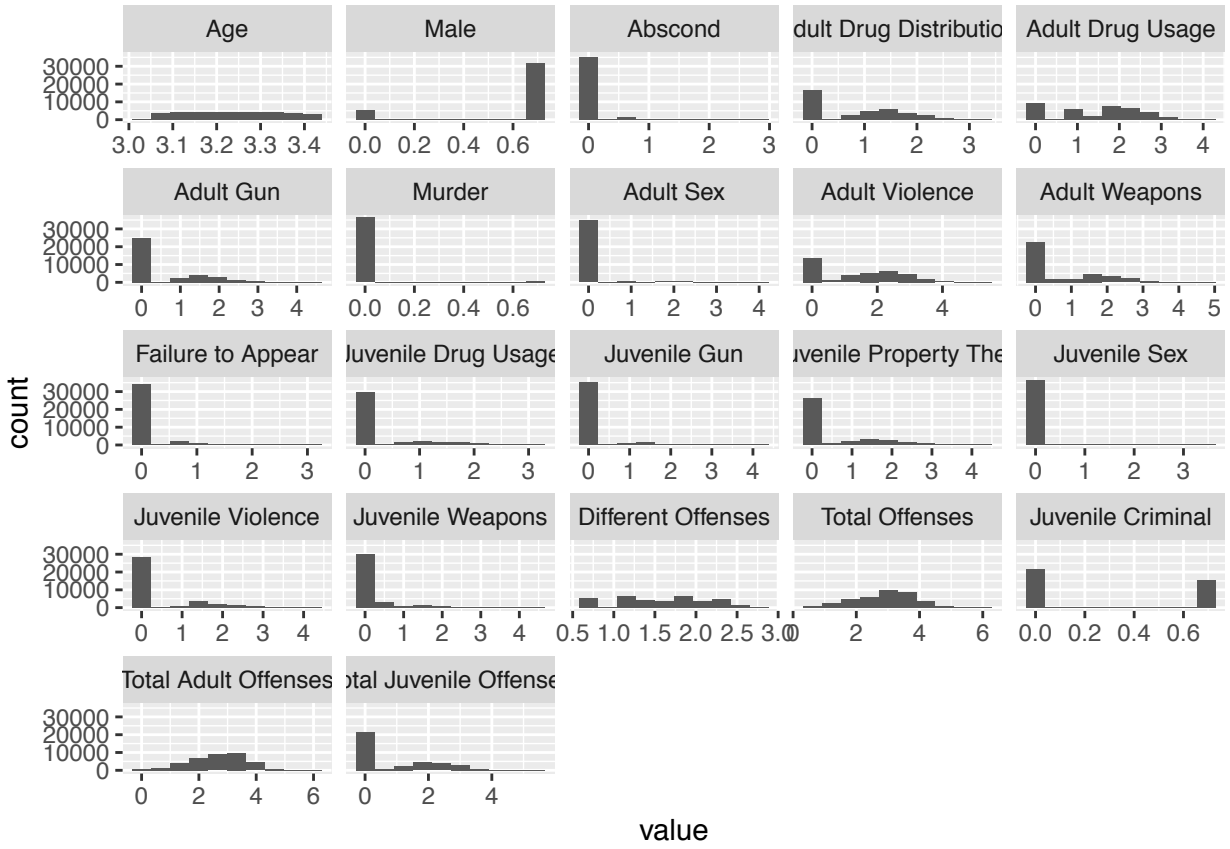

Figure 2: Log Transformed Histrograms of All Data

completely remove the skewness because of the high frequency of zeroes among the priors, the transformation eliminated the tail and normalized the remaining non-zero observations within each category.

Bivariate Analysis

Bivariate analysis is conducted to understand relationships within criminal behavior. Correlations amongst the variables are examined through the Association Navigator, a tool developed by Andreas Buja, Abba Krieger, and Ed George.[8] Mean differences of criminal prior counts are examined across age and gender. Given the large sample size of the dataset, measures of statistical significance are less meaningful than measures of substantive significance as any even a minuscule difference can be found to be significant. Effect sizes are used to quantify these differences. Differences based on demographics emerge and missing data problems are uncovered.

Association Navigator produces a correlation matrix to visually evaluate correlations across all variables.[9] Examining correlations between variables prompts initial hypotheses about grouping individuals. Additionally, surprising correlations uncover idiosyncrasies in the data. However, the sparsity of the data discussed above makes interpreting correlations difficult. It may be the case that certain variables have many data points near zero and few data points on the extreme ends, leading to instability in correlation measurements. To address this, correlations are used for data exploration rather than for drawing final conclusions.

The correlation matrix reveals a negative relationship between age and juvenile crimes and age and total crimes. This goes against the intuition that older individuals have more opportunities to commit crimes and should have more overall criminal priors and juvenile priors. Ascertainment bias may drive this unexpected relationship, with individuals with

[8]"A Visualization Tool for Mining Large Correlation Tables: The Association Navigator | Handbook of Big Data | Taylor & Francis Group" n.d.

[9]See Appendix A

high criminal counts not eligible for parole and not in the dataset.

Association Navigator reveals a strong, positive relationship between adult gun and weapons priors. Within juvenile criminal priors, there is a positive relationship between drug, gun, and property priors and violent and weapon priors. Being male is positively correlated with all juvenile and adult crimes, different types of crimes, and total crimes.[10] These results align with prior expectations around gender and criminality and prior expectations of violent crimes.

The difference in means across age groups reveals missing juvenile data in the dataset. Age buckets are as follows: Adolescent - 0 to 18 years old, Young Adult - 19 to 35 years old, Adult - 36 to 55 years old, Elder - 56+ years old. Mean differences of juvenile crimes are found to be zero (See Table 2). This seems highly unlikely across ~48,000 Adult observations and ~4,000 Elder observations, indicating missing juvenile data for older individuals. Digging deeper, juvenile data is missing for individuals older than 33 years old. Addressing this, the dataset used for analysis only looks at 20 to 30-year-olds. See Tables 5 and 10.

	Group.1	Age	Male	Abscond	Adult Drug Distribution	Adult Drug Usage	Adult Gun	Murder	Adult Sex	Adult Violence	Adult Weapons	Failure to Appear
1	Adolescent	17.26	0.95	0.00	0.36	0.77	6.13	0.04	0.57	15.87	7.36	0.07
2	Young Adult	26.04	0.82	0.09	2.11	5.04	1.68	0.02	0.34	6.55	2.28	0.17
3	Adult	43.10	0.74	0.12	1.41	4.28	0.95	0.00	0.57	8.47	2.06	0.19
4	Elder	60.32	0.85	0.06	1.04	3.76	1.18	0.00	0.73	6.50	2.53	0.11

Table 5: Mean differences for adult crimes across age groups. Adolescent - 0 to 18, Young Adult - 18 to 35, Adult - 35 to 55, Elder - 55+

	Group.1	Juvenile Drug Usage	Juvenile Gun	Juvenile Property Theft	Juvenile Sex	Juvenile Violence	Juvenile Weapons
1	Adolescent	2.33	1.00	3.13	0.30	5.03	1.93
2	Young Adult	0.57	0.16	1.22	0.06	1.10	0.38
3	Adult	0.00	0.00	0.00	0.00	0.00	0.00
4	Elder	0.00	0.00	0.00	0.00	0.00	0.00

Table 6: Mean differences for juvenile crimes across age groups. Adolescent - 0 to 18, Young Adult - 18 to 35, Adult - 35 to 55, Elder - 55+

The hypothesis around age and criminal priors described using Association Navigator appears

[10]Individuals are coded 1 in the male variable if they are male.

to be driven by missing data rather than the nature of the data-generating process. This speaks to the difficulty of overcoming idiosyncrasies in the dataset to draw conclusions.

Mean differences between men and women affirm the positive relationship between being male and criminality. Men outperform women in the frequency of crimes within every category other than Absconding and Failure to Appear in Court. These differences are substantively significant based on effect sizes (greater than 0.8). See Tables 7, 8, and 9.

	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	a.cluster
1	0.9567	0.0520	2.3659	5.6080	10.7814	0.0440	0.6378	22.8167	13.8050	0.0731	56.1839	0.0000	1.0000
2	0.7662	0.0730	0.7161	2.3867	0.6897	0.0072	0.4340	5.1075	0.9247	0.1677	10.5066	0.0000	2.0000
3	0.9375	0.0883	6.6680	13.6247	1.1501	0.0114	0.1872	3.7620	1.5300	0.1249	27.1467	0.0000	3.0000

Table 7: Adult Priors: Differences in Means and Effect Sizes Across Gender

	Juvenile Drug Usage	Juvenile Gun	Juvenile Property Theft	Juvenile Sex	Juvenile Violence	Juvenile Weapons
1	0.1736	0.0856	0.5553	0.019592	0.9074	0.2376
2	0.79103	0.2470	1.7506	0.0847	1.5258	0.5523
3	-1.4886	0.2276	-0.2752	-0.2692	-2.7353	-0.7044
4	Large	Small	Small	Small	Large	Medium

Table 8: Juvenile Priors: Differences in Means and Effect Sizes Across Gender

	Different Offenses	Total Offenses	Juvenile Criminal	Total Adult Offenses	Total Juvenile Offenses
1	2.8325	12.7228	0.2395	10.7486	1.9793
2	4.9435	26.6489	0.45166	21.7194	4.9518
3	-1.2992	-2.6369	-1.8022	0.6845	1.1122
4	Large	Large	Large	Medium	Large

Table 9: Total Offenses: Differences in Means and Effect Sizes Across Gender

Methodology

The main question focuses on subsetting criminals based on their prior behavior. Separating criminals into groups allows us to respond to questions raised by the criminology literature such as: Is there a difference between violent and non-violent offenders? How do crimes differ on gender? What is the relationship between adult and juvenile crimes?

Several methods exist to group data points. These methods are divided into supervised and

unspecified methods. Supervised methods group individuals based on preset, subject-matter relevant criteria. Unsupervised methods group individuals without clear reasoning behind the groups. In this analysis, unsupervised methods are used to determine patterns between adult criminal priors and supervised methods are used to explore groupings based on the total number of adult crimes and the relationship between adult and juvenile criminal priors.

The unsupervised clustering methods used are K-means clustering and Principal Components Analysis (PCA). Since there is no way to know how the groupings of observations or variables will be created, these methods are used to create groupings of adult criminal priors where it is unknown if criminals can be grouped on criminal priors.

Decision trees and conditional associations using PCA are used for supervised clustering. For this analysis, decision trees are created with the estimation target of adult total offenses. Groups are created based on the rules around committing adult crimes. Conditional association builds on this analysis by conditioning on a measure of overall criminality to see if there are differences in crimes between individuals with the same propensity to commit crimes. Conditional associations are also used to understand if the relationship between adult and juvenile crimes changes for a given level of criminality.[11]

[11]Holland and Rosenbaum 1986

Analysis

Can criminals be grouped by adult criminal priors?

K-means clustering

Individuals are heterogeneous and commit different crimes at different propensities, thus when clustering individuals into groups based on their behavior, one would expect to see different types of criminals emerge. To explore whether clusters based on adult prior crimes exist, the K-means clustering algorithm is run as described above on the adult criminal priors of 20 to 30-year-olds. The centroids of the three clusters reveal three distinct groups: high-frequency criminals; drug-using and drug-distributing criminals; and low violence criminals. These groupings are consistent across individuals with and without juvenile criminal histories.

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	a.cluster
1	1	24.82	0.96	0.10	2.97	6.85	7.10	0.04	0.44	15.34	9.01	0.14	41.99	7.21	1.00
2	2	24.67	0.76	0.09	0.11	1.03	0.31	0.01	0.58	6.37	0.51	0.18	9.20	3.17	2.00
3	3	24.30	0.90	0.10	4.19	9.13	0.32	0.02	0.14	2.01	0.42	0.17	16.50	4.10	3.00

Table 10: Log Clusters: Adult Priors with All Criminals

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	a.cluster
1	1	23.55	0.85	0.14	0.56	1.84	0.56	0.02	0.85	10.84	0.89	0.18	15.88	10.20	1.00
2	2	23.09	0.92	0.12	3.45	7.64	0.29	0.02	0.04	1.09	0.37	0.20	13.23	9.38	2.00
3	3	23.88	0.98	0.13	3.44	7.89	7.70	0.05	0.39	15.76	9.75	0.18	45.28	13.03	3.00

Table 11: Log Clusters: Adult Priors with Juvenile Criminals

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	a.cluster
1	1	25.91	0.93	0.06	2.67	6.17	6.38	0.02	0.42	13.23	8.04	0.09	37.09	0.00	1.00
2	2	25.30	0.73	0.07	0.13	0.86	0.22	0.01	0.70	7.26	0.44	0.18	9.86	0.00	2.00
3	3	25.19	0.83	0.08	3.49	7.99	0.20	0.01	0.11	1.42	0.28	0.16	13.74	0.00	3.00

Table 12: Log Clusters: Adult Priors with Non-Juvenile Criminals

Looking across all individuals, three criminal "types" emerge. The first criminal type describes individuals with high overall counts of crime, particularly in crimes involving violence, guns, weapons, and drugs (Table 10: Cluster 1, Table 11: Cluster 3, Table 12: Cluster 1). This

group tends to be male. The second criminal type describes individuals with frequent drug usage and drug distribution. Compared to the first group, the second group has lower occurrences of violent crimes (Table 10: Cluster 3, Table 11: Cluster 2, Table 12: Cluster 3). The third criminal type describes non-violent criminals (Table 10: Cluster 2, Table 11: Cluster 2, Table 12: Cluster 2). This group has a higher proportion of females and a higher proportion of sex offenders.

These groups hold regardless of whether or not individuals committed juvenile crimes (See Table 11 and 12). As expected, there are higher counts of crimes amongst individuals who were juvenile criminals compared to those who weren't. Effect sizes quantify whether there are meaningful differences between the groups. Across most predictors, the groups differ quite significantly between one another indicating the formation of distinct clusters. (See Tables 25, 26,27, and 28 in Appendix).

While these cohesive groups form on log-transformed data, the clusters change quite drastically when one switches to performing analysis on scaled data. (See Tables 22, 23, and 24 in Appendix) Because the clusters are formed based off of distances between centroids, as the units are transformed, the distances will change and thus so will the clusters. The analysis is limited to clusters formed from the log-transformed data as it is the appropriate transformation to handle issues of skewness mentioned prior. However, it is important to note the sensitivity of k-means to data transformation as a finding regarding the stability and consistency of unsupervised methods.

Principal Components Analysis (PCA)

The components that emerge from PCA align with the criminal types observed through K-means clustering: violent criminals, non-violent drug users and distributors, violent drug users and distributors, and non-violent sexual offenders. See Table 13. To interpret the components, each variable is given a score. The magnitude of the score indicates the importance of the variable and the direction of the loading indicates the relationship between variables. For example, in the first component, weapons, violence, and gun priors are of similar magnitude and direction. One group of criminals is defined by high levels of these crime types.

There are patterns in criminal priors for adult crimes for 20 to 30-year-olds.

	Comp.1	Comp.2	Comp.3	Comp.4
Abscond	-0.00	-0.00	0.01	-0.01
Adult Drug Distribution	-0.04	-0.59	0.23	0.02
Adult Drug Usage	-0.02	-0.73	0.24	-0.02
Adult Gun	-0.51	-0.14	-0.45	0.07
Murder	-0.01	-0.00	-0.00	0.00
Adult Sex	-0.04	0.05	0.14	0.99
Adult Violence	-0.64	0.27	0.70	-0.14
Adult Weapons	-0.57	-0.13	-0.42	0.03
Failure to Appear	0.01	-0.00	-0.01	-0.04

Table 13: Significant Variables in First Four Clusters

Conditional Association

Controlling for individuals' propensities to commit crimes, there is groupings do not emerge based on drug and distribution priors, sex and violent priors, and drugs and violent priors. are correlations between different criminal priors after breaking individuals into groups based on their propensity to commit crimes. Negative relationships between the variables indicate

that groups may be formed based on the variables. The only distinction that emerges is one between adult drug and violence priors. This distinction is not significant based on a 0.05 significance level for individuals with the highest propensity to commit crimes. Tables 14, 15 and 16 show this for individuals with juvenile crime histories. The results are the same across the total population and for individuals without juvenile criminal histories. The only difference is that there is a significant distinction between drug and violent criminal priors for individuals without juvenile criminal who have the highest propensity to commit crimes. There is no clear story that emerges from these results. Given the nature of the data, there is no way to determine the drivers of this observation and whether controlling for criminality allows groupings of adult crimes.

	Percentile	Correlations	P-values
cor	4th	-0.02	0.18
cor	3rd	-0.39	9.45e-197
cor	2nd	-0.39	1.29e-199
cor	1st	-0.48	4.24e-313

Table 14: Conditional associations between Drugs and Violence only for individuals with juvenile history

	Percentile	Correlations	P-values
cor	4th	0.21	1.01e-53
cor	3rd	0.31	2.30e-120
cor	2nd	0.27	4.25e-89
cor	1st	0.26	7.36e-84

Table 15: Conditional associations between Sex and Violence only for individuals with juvenile history

	Percentile	Correlations	P-values
cor	4th	0.92	0
cor	3rd	0.90	0
cor	2nd	0.88	0
cor	1st	0.85	0

Table 16: Conditional associations between Drugs and Dist only for individuals with juvenile history

Is there a relationship between adult and juvenile criminal priors?

Decision Tree

The results varied drastically between data transformations within standard unsupervised clustering techniques. Following the typical criticisms of unsupervised methods, because clusters are formed based off of distances between centroids, as the units are transformed, the distances will change and thus so will the clusters. One would expect that with supervised methods because there is a response variable, the clusters formed will have more stability. Thus, the intention of using decision trees to form clusters based on the nodes individuals fall was to create clusters with more meaning or stability.

The analysis is limited to clusters formed based off of 4 trees (See Figures /ref{y5}, /ref{y6}, /ref{y7}, and /ref{y8} in the Appendix): log-transformed and normalized trees with 3 nodes and log-transformed and normalized trees with 4 nodes.

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	cluster
1	3	24.93	0.83	0.09	2.03	4.86	1.68	0.01	0.42	6.65	2.19	0.16	18.10	2.25	3.00
2	4	23.17	0.95	0.13	3.40	7.65	2.57	0.03	0.23	7.11	3.29	0.20	24.62	8.97	4.00
3	5	23.85	0.96	0.10	3.15	7.15	5.12	0.05	0.34	12.13	6.53	0.14	34.72	21.76	5.00

Table 17: Scaled Juvenile Priors with Decision Tree Clusters

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	cluster
1	3	24.93	0.83	0.09	2.03	4.86	1.68	0.01	0.42	6.65	2.19	0.16	18.10	2.25	3.00
2	4	24.11	0.96	0.10	2.72	6.22	5.15	0.05	0.46	12.88	6.55	0.14	34.27	20.86	4.00
3	5	23.24	0.96	0.12	3.42	7.70	2.97	0.04	0.23	7.80	3.80	0.19	26.29	11.13	5.00

Table 18: Log Juvenile Priors with Decision Tree Clusters

The three clusters correspond to three distinct groups: high violence criminals; drug-using and drug-distributing criminals; and low violence criminals. These groupings are consistent across individuals with and without juvenile criminal histories.

Looking across all individuals, three criminal "types" emerge. The first criminal type describes individuals with high overall counts of violent crimes: violence, guns, weapons, and drugs (Table 17: Cluster 3, Table 18: Cluster 2). This group tends to be male. The second criminal type describes individuals with frequent drug usage and drug distribution. Compared to the first group, the second group has lower occurrences of violent crimes (Table 17: Cluster 2, Table 18: Cluster 3). The third criminal type describes non-violent criminals (Table 17: Cluster 1, Table 18: Cluster 2). This group has a higher proportion of females and a higher proportion of sex offenders. These groups are almost exactly identical to those created when using k-means clustering on juvenile groups.

As expected, there is consistency between the groups formed from each of the two data transformations (scaled and log-transformed) unlike the dissimilarities observed with unsupervised methods.

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	cl
1	3	24.93	0.83	0.09	2.03	4.86	1.68	0.01	0.42	6.65	2.19	0.16	18.10	2.25	
2	4	23.17	0.95	0.13	3.40	7.65	2.57	0.03	0.23	7.11	3.29	0.20	24.62	8.97	
3	6	23.84	0.97	0.10	3.18	7.22	4.68	0.05	0.34	11.31	5.99	0.14	33.00	19.33	
4	7	24.01	0.93	0.11	2.66	6.19	11.77	0.07	0.44	24.54	14.73	0.23	60.74	58.52	

Table 19: Scaled Juvenile Priors with Decision Tree Clusters

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	cl
1	3	24.93	0.83	0.09	2.03	4.86	1.68	0.01	0.42	6.65	2.19	0.16	18.10	2.25	
2	4	24.11	0.96	0.10	2.72	6.22	5.15	0.05	0.46	12.88	6.55	0.14	34.27	20.86	
3	6	23.17	0.95	0.13	3.40	7.65	2.58	0.03	0.23	7.11	3.30	0.20	24.63	8.98	
4	7	23.62	0.97	0.10	3.55	8.02	5.10	0.05	0.24	11.56	6.53	0.15	35.29	22.81	

Table 20: Log Juvenile Priors with Decision Tree Clusters

When clustering based on trees with four nodes, four distinct groups form: violent criminals,

non-violent drug users and distributors, violent drug users and distributors, and non-violent sexual offenders.

The first criminal type describes individuals with high overall counts of violent crimes: violence, guns, weapons, and drugs (Table 19: Cluster 4, Table 20: Cluster 2). This group tends to be male. The second criminal type describes individuals with frequent drug usage and drug distribution and high violence (Table 19: Cluster 3, Table 20: Cluster 4). The third criminal type describes individuals with frequent drug usage and drug distribution (Table 19: Cluster 2, Table 20: Cluster 3). Compared to the second group, the third group maintains high levels of drug use but with lower occurrences of violent crimes. The fourth criminal type describes non-violent criminals (Table 19: Cluster 1, Table 20: Cluster 1). This group has a higher proportion of females and a higher proportion of sex offenders.

Once again the consistency between the two clusterings despite the data transformations is maintained. Comparing the clustering analysis between when 3 groups are 4 formed, it appears that high drug users get broken out into two seperate groups - high violence drug users and low violence drug users.

Conditional Association

Controlling for criminality, there is a positive relationship between adult and juvenile crimes for individuals with the highest propensity for crime. This relationship is true across the total population and for individuals with juvenile records. The correlations between adult and juvenile weapons, violent, and drugs crimes are strongest for individuals with juvenile records.

Men are overrepresented in the subset of the population with juvenile records.

	Percentile	Correlations	P-values
cor	4th	0.92	0
cor	3rd	0.90	0
cor	2nd	0.88	0
cor	1st	0.85	0

Table 21: Relationships between Adult and Juvenile Gun Priors for individuals with junvenile criminal histories

Conclusion

Supervised and unsupervised methods reveal groupings among criminals based on past behavior. K-means clustering, PCA, and decision trees reveal groupings based on violent and weapons-related crimes, drug-related crimes, and sex-related crimes.

There are also differences between individuals in the dataset with and without juvenile crimes. No relationship is found between adult and juvenile crimes when criminality is controlled for except for individuals with the highest level of criminality. Finally, men are found to be overrepresented in violent crimes and in the subset of the dataset with juvenile criminal histories.

Using supervised and unsupervised methods produce different results; however, supervised methods produce more stable results across data transformation. Examining unsupervised methods, groups formed from K-means clustering were similar across different specifications of the model that sets the number of groups found. Groupings formed from unsupervised methods are not sensitive to model specifications, supporting the use of unsupervised methods. Unfortunately, there are large differences between the groups formed from the log-transformed

and scaled data. This difference is not present for supervised methods. This indicates that supervised methods create more stable and reliable results relative to unsupervised methods.

The idiosyncrasies from the data make it difficult to expand on the relationships observed. While groupings emerge from the dataset, these cannot be generalized to the criminal population because the analysis only focuses on a subset of criminals selected in an idiosyncratic manner. For example, expanding the groupings to the total population, unobserved individuals because of ascertainment bias may not be separable into sexual and drug-related offenders. Missing data issues also narrows the view of the criminal population to individuals age 20 to 30. Since there is no way to know if the groups found are a function of the type of individuals in the dataset or of criminal behavior, no strong results are found.

Even assuming that the analysis is only concerned with understanding criminal behavior for individuals in the dataset to reduce generalization concerns, the nature of the data also presents challenges to drawing conclusions. Complications around the sentencing process, like multiple charges being assigned to a single crime, make it unclear how well the variables in the dataset represent criminal behavior. The number of zeros and outliers also make it difficult to form meaningful distinct groups. For example, are individuals with hundreds of criminal priors an important group to consider or should they be removed so that groups are formed on other factors? From this, drawing simple conclusions around variable relationships is non-trivial.

Next Steps

The analysis struggled to overcome data limitations. Data that allows a full view of individuals' criminal history across the entire criminal population would address these challenges. This data includes current criminal charge, age and time of past criminal charges, longitudinal data for individuals across a time period that allows for "full" criminal histories.

There are also many additional avenues of analysis for grouping the criminal population. Within unsupervised methods, extensions of k-means clustering include varying the method of distance measures within k-means or conducting k-medians clustering. One could explore the variability in the results as a form of sensitivity analysis. To understand whether the frequency of crimes or occurrence of crimes categorizes criminal behavior, one could repeat this analysis utilizing dichotomized data. A dichotomized analysis would control for issues of skewness and the incomplete nature of the data.

As another extension of the analysis, one could redo the above procedures with a focus on the propensity to commit crime rather than the counts of crimes. This could be done by dividing the counts in each category by the age of the individual to understand their criminal behavior within the scope of time they had to commit crimes. This would address the issue of censoring due to age in the data.

Extending the supervised analysis, one could explore clustering of proximity values from a random forest. One would expect similar results as those produced by the decision tree clustering methods employed in this paper.

If given more demographic information, one could perform more intensive clustering analysis.

First, if the data were sufficiently rich one could cluster on criminal priors for different subsets of individuals based on these demographics. Alternatively, one could cluster on the outcome variable and see how the groupings differ by demographic. One could then move individuals from one cluster to another reducing the clustering on the outcome ever so slightly but resulting in a greater distinction on the demographic covariates.

Final Thoughts

While the analysis set out to answer questions about criminal behavior, the most important finding is that one cannot jump into analysis mode. Understanding where the data come from and "looking" at the data carefully is critical. The conclusions one draws is potentially very sensitive to data issues as in this case. The data could provide serious limitations and caveats on the conclusions that are drawn. Given data where software is so readily available making analysis easy to perform, it is tempting to take data and access this software without getting an understanding of the data beforehand. This conclusion is not trivial in this day and age.

Bibliography

"A Visualization Tool for Mining Large Correlation Tables: The Association Navigator | Handbook of Big Data | Taylor & Francis Group." n.d. Accessed April 26, 2018. https://proxy.library.upenn.edu:4137/books/e/9781482249088/chapters/10.1201%2Fb19567-8.

"Ascertainment Bias." n.d. Accessed May 5, 2018. https://www.mun.ca/biology/scarr/Ascertainment_bias.html.

Basu, Sugato, et al. "A Probabilistic Framework for Semi-Supervised Clustering." A Probabilistic Framework for Semi-Supervised Clustering, Aug. 2004, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.3.2964&rep=rep1&type=pdf.

Berk, Richard A. 2016. Statistical Learning from a Regression Perspective. 2nd ed. Springer Texts in Statistics. Springer International Publishing. //www.springer.com/us/book/9783319440477.

Blumstein, Alfred, Jacqueline Cohen, and David P. Farrington. 1988. "CRIMINAL CAREER RESEARCH: ITS VALUE FOR CRIMINOLOGY*." Criminology 26 (1): 1-35. https://doi.org/10.1111/j.1745-9125.1988.tb00829.x.

De Bruin, Jeroen, Tim Cocx, Walter Kosters, Jeroen F. J. Laros, and Joost N. Kok. 2006. "Data Mining Approaches to Criminal Career Analysis." In Proceedings - IEEE International Conference on Data Mining, ICDM, 171-77. https://doi.org/10.1109/ICDM.2006.47.

Holland, Paul W., and Paul R. Rosenbaum. 1986. "Conditional Association and Unidimensionality in Monotone Latent Variable Models." The Annals of Statistics 14 (4): 1523-43.

Horning, Amber M., C. Gabrielle Salfati, and Kristan Crawford. 2010. "Prior Crime

Specialization and Its Relationship to Homicide Crime Scene Behavior Type." Homicide Studies 14 (4): 377-99. https://doi.org/10.1177/1088767910382833.

Howard, Marc Morj?. 2017. Unusually Cruel: Prisons, Punishment, and the Real American Exceptionalism. Oxford University Press.

Kilinc, Ozsel, and Ismail Uysal. 2017. "Auto-Clustering Output Layer: Automatic Learning of Latent Annotations in Neural Networks." ArXiv:1702.08648 [Cs], February. http://arxiv.org/abs/1702.08648.

Nagin, Daniel S., and Raymond Paternoster. 1991. "On the Relationship of Past to Future Participation in Delinquency." Criminology 29 (2): 163-89. https://doi.org/10.1111/j.1745-9125.1991.tb01063.x.

Pan, Wei, Xiaotong Shen, and Binghui Liu. 2013. "Cluster Analysis: Unsupervised Learning via Supervised Learning with a Non-Convex Penalty." Journal of Machine Learning Research 14: 1865-89.

Rowe, David C., Alexander T. Vazsonyi, and Daniel J. Flannery 1995. "Sex Differences In Crime: Do Means and Within-Sex Variation Have Similar Causes?" Journal of Research in Crime and Delinquency 32 (1): 84-100. https://doi.org/10.1177/0022427895032001004.

Appendix

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	a
1	1	25.02	0.97	0.09	2.55	6.12	12.41	0.07	0.64	26.36	15.85	0.15	64.25	9.32	
2	2	24.48	0.81	0.09	0.81	2.55	0.93	0.01	0.43	5.60	1.22	0.17	11.80	3.73	
3	3	24.69	0.96	0.13	6.50	13.42	1.51	0.02	0.17	4.38	1.99	0.16	28.28	5.12	

Table 22: Normalized Clusters: Adult Priors with All Criminals

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	a
1	1	23.24	0.89	0.12	1.27	3.40	1.21	0.00	0.38	6.23	1.57	0.18	14.35	10.11	
2	2	24.01	0.98	0.16	5.81	12.32	5.76	0.00	0.29	12.34	7.37	0.20	44.27	11.92	
3	3	22.75	0.98	0.07	3.03	6.99	5.02	1.00	0.52	12.27	6.50	0.22	35.61	12.96	

Table 23: Normalized Clusters: Adult Priors with Juvenile Criminals

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	a
1	1	26.28	0.96	0.05	2.37	5.61	10.78	0.04	0.64	22.82	13.80	0.07	56.18	0.00	
2	2	25.21	0.77	0.07	0.72	2.39	0.69	0.01	0.43	5.11	0.92	0.17	10.51	0.00	
3	3	25.70	0.94	0.09	6.67	13.62	1.15	0.01	0.19	3.76	1.53	0.12	27.15	0.00	

Table 24: Normalized Clusters: Adult Priors with Non-Juvenile Criminals

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	a.cluster
1	1	26.2820	0.9567	0.0520	2.3659	5.6080	10.7814	0.0440	0.6378	22.8167	13.8050	0.0731	56.1839	0.0000	1.0000
2	2	25.2054	0.7662	0.0730	0.7161	2.3867	0.6897	0.0072	0.4340	5.1075	0.9247	0.1677	10.5066	0.0000	2.0000
3	3	25.7049	0.9375	0.0883	6.6680	13.6247	1.1501	0.0114	0.1872	3.7620	1.5300	0.1249	27.1467	0.0000	3.0000

Table 25: Juvenile Criminals: Differences in Means and Effect Sizes Across K-means Clusters

	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total
1	0.0439628482972136	0.637770897832817	22.8167182662539	13.8049535603715	0.0730650154798762	56.1839009287926	0
2	0.00717796171753751	0.433975685463011	5.10747542679772	0.924728401448526	0.167679772374547	10.5065959648215	0
3	0.011441647597254	0.187185354691076	3.76201372997712	1.52997711670481	0.124942791762014	27.1466819221968	0
Effect Size Cluster 1 and 2	-0.540881631562933	0.497910729108362	0.116604848307573	-1.98168553200518	0.89036136749379	-310.677485505468	1.22645879843159
Magnitude Cluster 1 and 2	Medium	Small	Negligible	Large	Large	Large	Large
Effect Size Cluster 1 and 3	-0.786515810162877	0.444564128774434	0.00402066871282512	-2.71644430615798	0.862249508848193	-207.118323670312	0.250635712840459
Magnitude Cluster 1 and 3	Medium	Small	Negligible	Large	Large	Large	Small
Effect Size Cluster 2 and 3	-1.0321632642629	0.391217882347502	-0.108563496625952	-3.45161353075949	0.834137840815943	-517.795809175779	-0.725120619763387
Magnitude Cluster 2 and 3	Large	Small	Negligible	Large	Large	Large	Medium

Table 26: Juvenile Criminals: Differences in Means and Effect Sizes Across K-means Clusters

	Group.1	age	male	abscpriors	Adistpriors	Adrugpriors	Agunpriors	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total	a.cluster
1	1	26.2820	0.9567	0.0520	2.3659	5.6080	10.7814	0.0440	0.6378	22.8167	13.8050	0.0731	56.1839	0.0000	1.0000
2	2	25.2054	0.7662	0.0730	0.7161	2.3867	0.6897	0.0072	0.4340	5.1075	0.9247	0.1677	10.5066	0.0000	2.0000
3	3	25.7049	0.9375	0.0883	6.6680	13.6247	1.1501	0.0114	0.1872	3.7620	1.5300	0.1249	27.1467	0.0000	3.0000

Table 27: Non- Juvenile Criminals: Differences in Means and Effect Sizes Across K-means Clusters

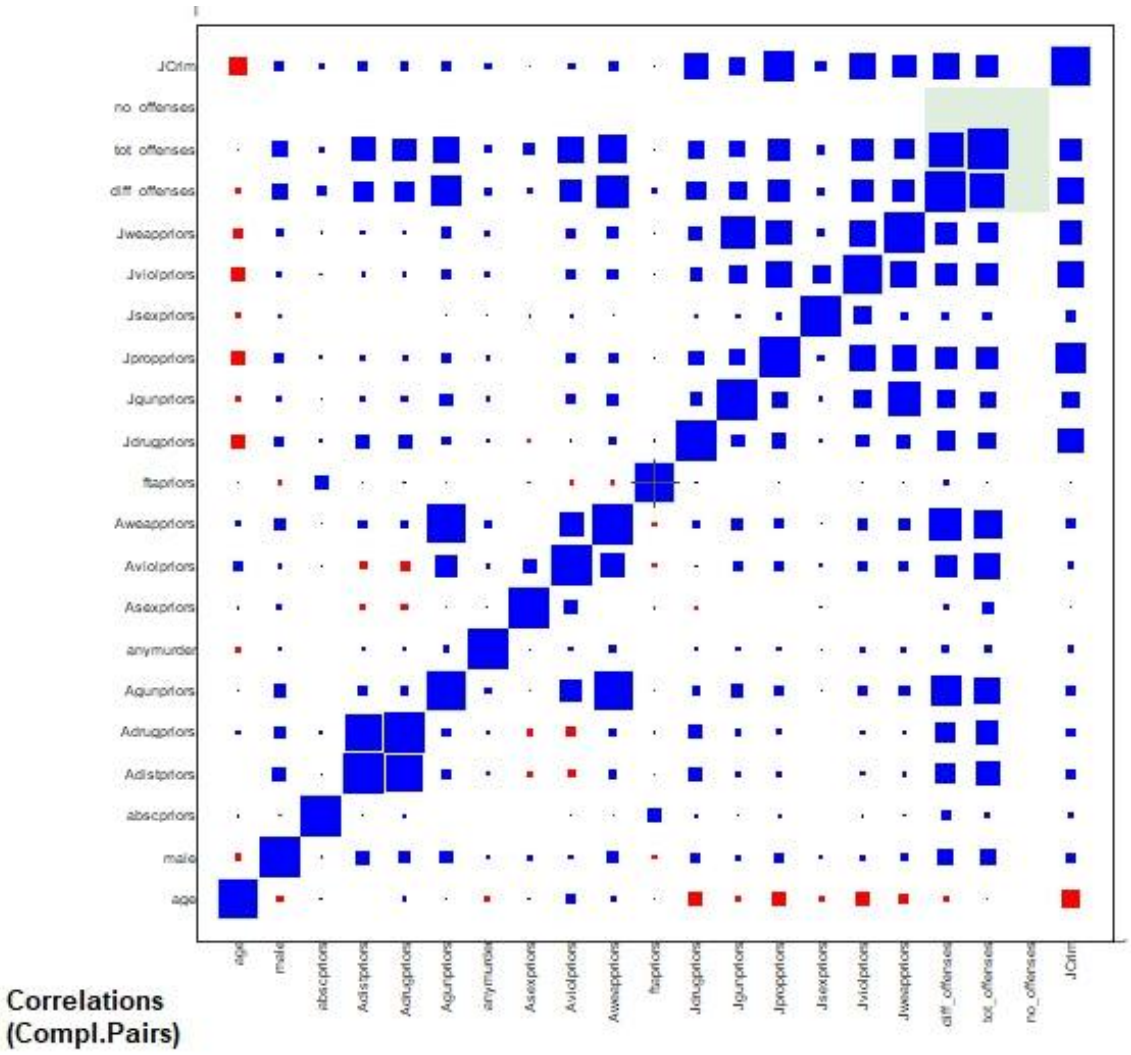

Figure 3: Appendix A: Association Navigator.

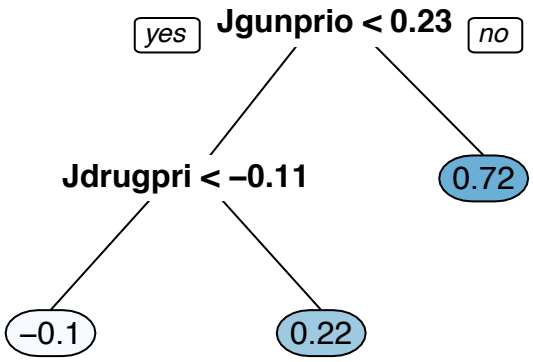

Figure 4: Tree for Scale Juvenile Priors

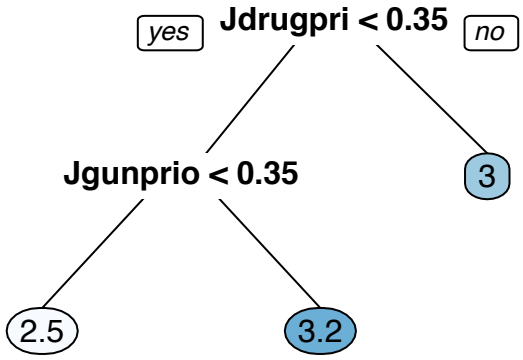

Figure 5: Tree for Log Juvenile Priors

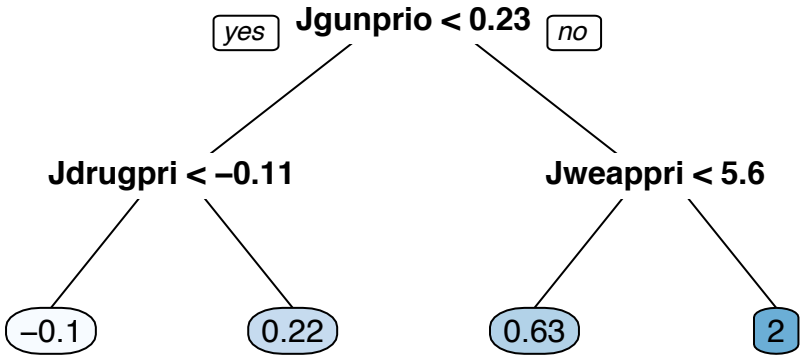

Figure 6: Tree for Scale Juvenile Priors

	anymurder	Asexpriors	Aviolpriors	Aweappriors	ftapriors	adult_total	juv_total
1	0.0439628482972136	0.637770897832817	22.8167182662539	13.8049535603715	0.0730650154798762	56.1839009287926	0
2	0.00717796171753751	0.433975685463011	5.10747542679772	0.924728401448526	0.167679772374547	10.5065959648215	0
3	0.011441647597254	0.187185354691076	3.76201372997712	1.52997711670481	0.124942791762014	27.1466819221968	0
Effect Size Cluster 1 and 2	-0.540881631562933	0.497910729108362	0.116604848307573	-1.98168553200518	0.89036136749379	-310.677485505468	1.22645879843159
Magnitude Cluster 1 and 2	Medium	Small	Negligible	Large	Large	Large	Large
Effect Size Cluster 1 and 3	-0.786515810162877	0.444564128774434	0.00402066871282512	-2.71644430615798	0.862249508848193	-207.118323670312	0.250635712840459
Magnitude Cluster 1 and 3	Medium	Small	Negligible	Large	Large	Large	Small
Effect Size Cluster 2 and 3	-1.0321632642629	0.391217882347502	-0.108563496625952	-3.45161353075949	0.834137840815943	-517.795809175779	-0.725120619763387
Magnitude Cluster 2 and 3	Large	Small	Negligible	Large	Large	Large	Medium

Table 28: Non-Juvenile Criminals: Differences in Means and Effect Sizes Across K-means Clusters

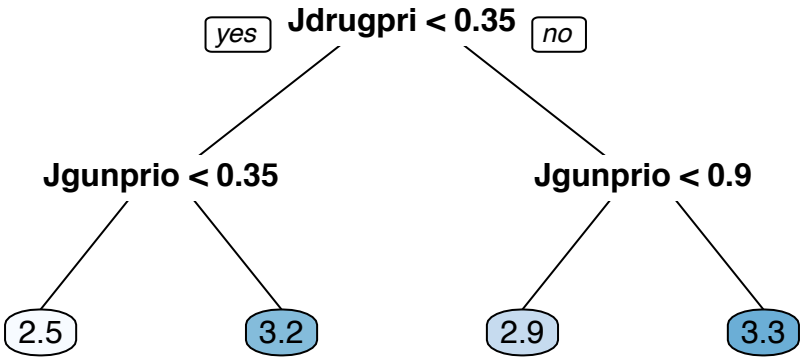

Figure 7: Tree for Log Juvenile Priors