STA2101 Draft Bibliography and Exploratory Data Analysis- Willem Atack

1 Summary of Previous Work

In existing literature, data-driven crime prevention is investigated from two different lenses. In the first, researchers attempt to predict offenders and victims of crime. In the second, the time and place of a crime is predicted, using environmental data. This project's dataset and proposed method falls into the latter category, though crime rates will be predicted on a broader scale (community-wide crime rates will be understood/predicted as opposed to individual crimes).

Due to the large number of environmental variables (covariates in the project dataset) involved with understanding and predicting crime, selecting a subset of important features is important to reduce model complexity, understand significant contributers to crime rates, and avoid overfitting when building predictive models. In past work, LASSO-based feature selection has proved effective at reducing the dimensionality of the data and selecting only important features [4]. In the same study, the environmental data was used to predict what type of crime would be most prevalent in different neighbourhoods. Using naive Bayes classification models was effective using recall, accuracy and precision to evaluate the model [4].

Most literature in the space of data-driven crime prevention studies machine learning techniques. In a systematic literature review, Dakalbab et al. survey 120 research papers which use AI approaches at predicting crime, and identified 64 different machine learning techniques which were used, mostly in the realm of supervised learning [1]. In this paper, they identify mean error (MSE) as the most popular performance metric, and the most appropriate for regression tasks [1]. Further, random forest machine learning methods tend to be used most frequently [1]. In this project, we will compare the performance of random forest methods to traditional regression techniques.

While advancements in data-driven crime prediction are promising, one gap in recent literature is that machine learning methods are typically not ”explainable”, i.e. they are ”blackbox” approaches. In contrast, traditional regression techniques are capable of revealing the significance and contribution of different features, but perform poorly at predicting crime. A recent study attempts to overcome this gap by taking advantage of the interpretability of advanced machine learning methods, like the XGBoost algorithm [5]. In this study 17 variables are selected to be used in the model, and Shapley additive explanation (SHAP) is used to isolate the contribution of individual variables (a variable with a higher SHAP value has a higher correlation with the model output) [5]. In this project, I will separately use classical regression techniques at explaining significant contributers to crime, and then compare the predictive performance of these models with machine learning approaches, though if time permits, exploring SHAP methods would be an interesting extension.

Though beyond the scope of this project, He et al. propose using GAN neural networks to build a prediction model of city floor plans and corresponding crime distribution maps [2]. In this model, a city map and corresponding environmental variables can be directly fed into a neural network which outputs a heat map displaying neighbourhoods with high likely crime distributions. The model was able to accurately predict crime concentration areas when untrained on Philadelphia city data [2]. The idea of this model is to use the results to adjust city layout to reduce crime rates. This project will not use similar techniques to this, but I thought it was interesting to read about the most recent developments in this realm.

Finally, it is worth noting that the previous policy studies have identified family structure, and poverty rates to be the two main root causes of crime [3]. This should serve as a launching point for an Exploratory Data Analysis.

2 Raw Dataset Description

The dataset in question combines socio-economic data from the 1990 U.S. census, law enforcement data from a 1990 Law Enforcement Management survey, and FBI crime data from 1995. It has 2,215 instances of data representing the demographics and crime rates in different communities across the United States.

Each instance has 5 non-predictive features which identify the community. They have 124 predictive features, which include detailed demographic data about the community. This includes, but is not limited to, information on population, population density, age distribution, income distribution, education levels, unemployment rates, family structure data, race, immigration, housing, homelessness, as well as information on the community's police force such as number of police cars, and police department budget. An exhaustive list is not included due to the number of features, though a link to the dataset is included at the end of this section. All of the features are unnormalized, and many instances contain one or more missing values.

It is worth noting that many features are redundant, and can be eliminated before beginning the analysis. In many cases, a feature representing an absolute value, and another feature representing the same quantity as a rate are included. For instance, ”number of people living in areas classified as urban” and ”percentage of people living in areas classified as urban” are separate features. There are many other subsets of features which seem likely to have very strong correlations with each other. For example, it is likely that including both ”percentage of people that have immigrated in the last 3 years” and ”percentage of people that have immigrated in the last 5 years” would add unnecessary model complexity. Of course, this will be formally tested before removing any covariates from the model.

Each instance has 18 possible response variables. These represent the absolute number, and the rate of 9 different types of crime. The types of crime are ”murder”, ”rape”, ””robberies”, ”assaults”, ”burglaries”, ”larcenies”, ”autoTheft”, ”arsons” and ”non-violent”. In the analysis, the rates of each of these different crimes will be considered the response variable in different models.

Link to dataset: https://archive.ics.uci.edu/dataset/211/communities+and+crime+unnormalized

3 Exploration of Missing Values and Sampling Bias

The biggest challenge with this dataset is the number of covariates and the number of possible predictors. To make meaningful insights or predictions with the dataset, finding ways to appropriately reduce the number of predictors is necessary. To do so, first the number of missing values was investigated. Among the 124 covariates, more than 500 missing values (of a total 2,215 instances) was found in 22 of them. Interestingly, there were no missing values in the other 102 columns. Since such a high proportion of instances had missing values in these covariates, they were removed from the dataset.

Another major challenge with the dataset is the number of redundant/correlated columns. With 102 covariates, using a correlation matrix was overwhelming, however many of the covariates could reasonably be removed from the dataset when doing a manual pass over the data. To use the previous example, ”number of people living in rural areas” was removed, and ”percentage of people in rural areas” was kept. By using knowledge of demographic metrics (subject matter knowledge), an additional 61 potential covariates were eliminated as they were very clearly redundant. These 41 covariates will be used in the EDA phase, and more formal feature selection methods like LASSO and criterion tests will later be used to further decrease the dataset size.

In terms of the potential targets, the number and rate of arsons and rapes had over 200 missing values, while no other potential target had more than 3 missing values. Therefore, the analysis will focus on predicting the other 7 target variables, and fill missing values with the column mean.

In the dataset, there are 2,215 instances of communities were checked. Sampling bias was checked in three separate ways. First, the number of communities in each state were checked to ensure that the instances are geographically spread out. In general this was not a huge issue, as 48/50 states have communities included in the dataset (there are no communities included in Vermont or Montana, which are small states). The states with the most communities are California (279), New Jersey (211), and Texas (162). Though New Jersey is likely over-represented, California and Texas are two of the largest states, so this makes sense. Notably, New York has just 46 communities in the dataset despite being the third largest state. Overall, the geographical distribution isn’t perfect, but shouldn’t cause an issue in the analysis. A table is included in the appendix showing the count by state.

Second, the size of the communities were investigated to make sure that both small towns and large cities are included. The community with the largest population is New York City (7mm) while the smallest is Lake City (10k). The median population is 22k, while the mean is 53k. this shows that the dataset is mostly made up of small communities, with a small number of large cities skewing the mean population. This is in line with what is expected across the USA, so there does not seem to be a sampling bias problem here. See Figure 1 to see a histogram of the community populations.

Finally, the distribution of total violent crime rates was investigated to ensure that both areas with high and low crime rates are included. As viewed in the histogram in Figure 2, there appears to be a wide range of crime rates in the dataset, with a total range of 241 to 30,202 crimes/100k people, and quartiles of [3167, 6968]. There does not appear to be any sampling bias here.

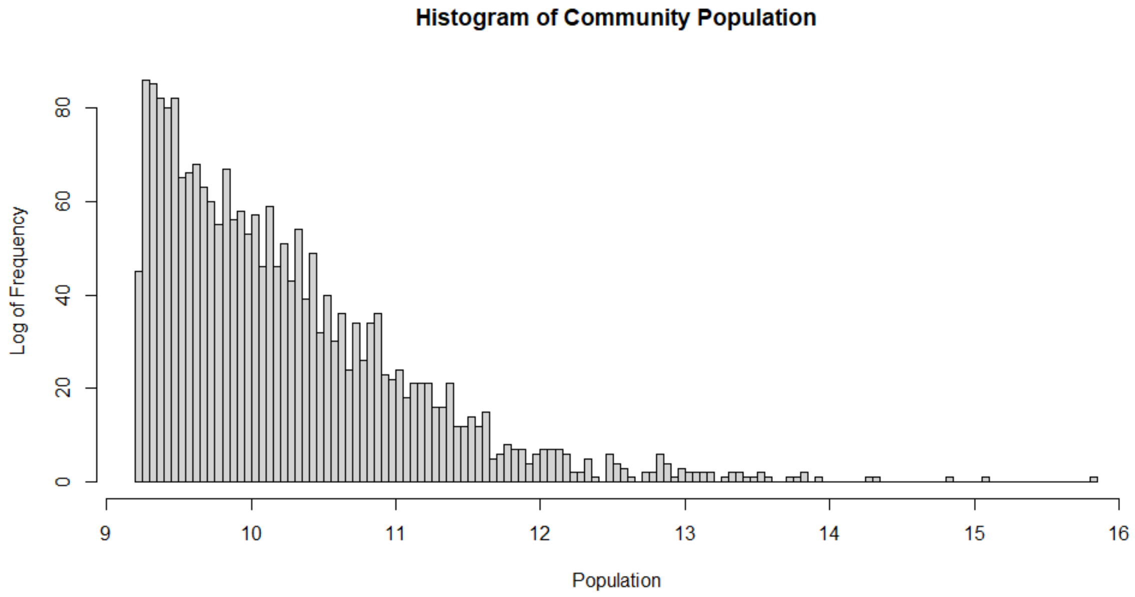


Figure 1: Histogram with the log-frequency of community populations

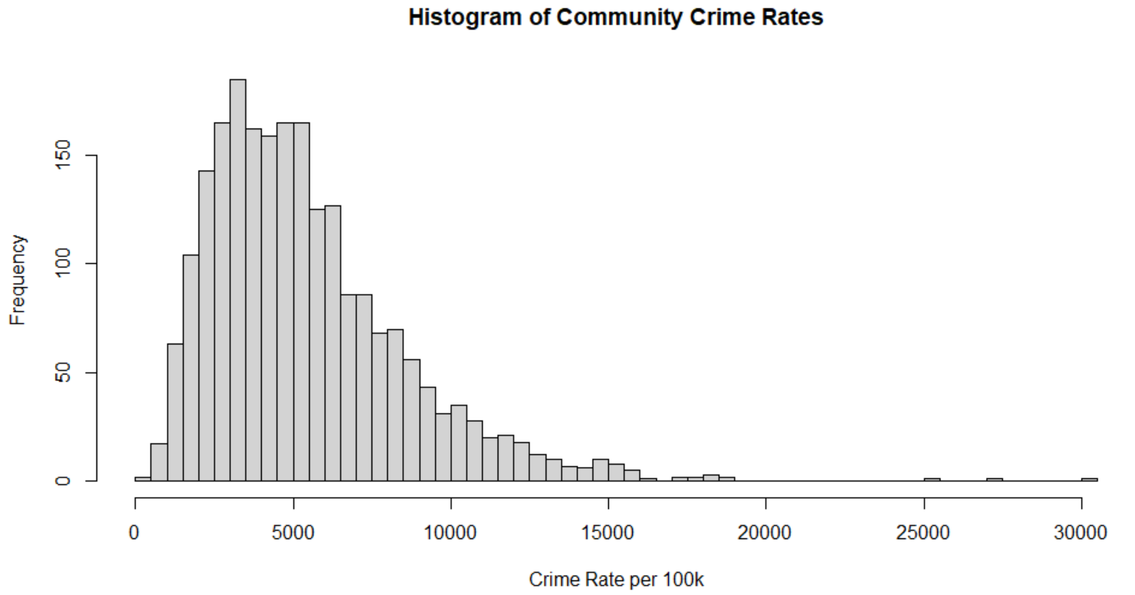


Figure 2: Histogram with the frequency of community total crime rates

4 Exploratory Data Analysis

While the distribution of total crime rates was already explored, one of the research questions was to determine if all types of crimes have the same contributors. Therefore, the correlation between the different types of crimes was checked, by creating the correlation matrix shown in Figure 3. Clearly, the correlation between all types of crime is positive, and in many cases is close to 1. If one type of crime is prevalent in a community, it is likely other types are also prevalent. The least correlated types of crime appear to be between Larcenies and Auto Thefts, with Larcenies also being less correlated to murder.

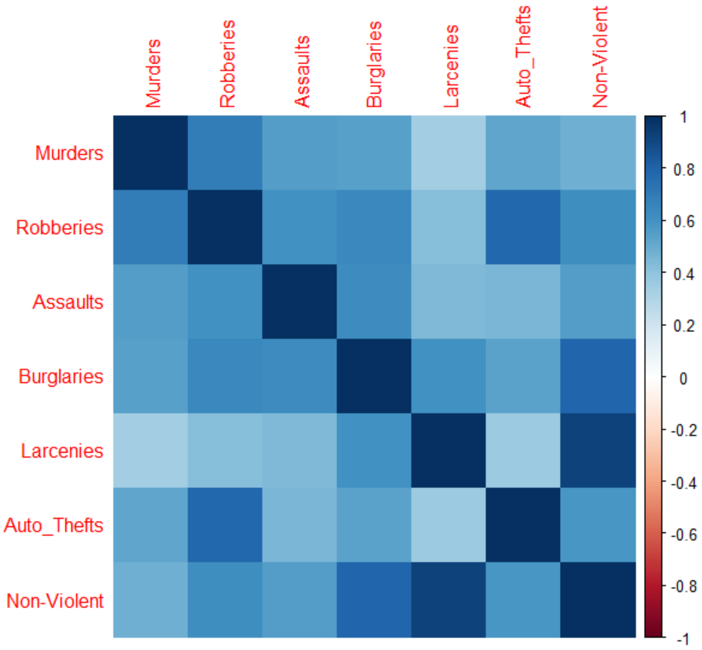


Figure 3: Correlation matrix between the different crime types

The distribution and relative frequency of each of the crime types was also plotted using a boxplot, shown in Figure 4. This boxplot shows us first that non-violent and larcenies are the most frequent crime types, and that murders and robberies are much less common. Further it shows us that for all crime types, the median values are closer to the minimum than the maximum values, and that there are many more outliers on the high end of crime rates. This tells us that in general, crime rates have a denser mass on the lower end of the range of possible rates, but the distribution is more fat-tailed when we exceed the median number of crimes. In other words, a small proportion of communities have a crime rate which is much higher than the median rate.

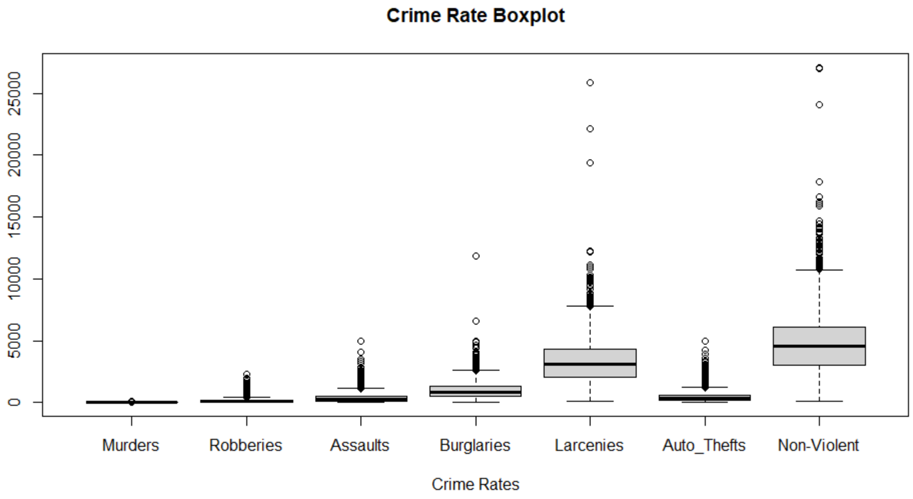


Figure 4: BoxPlot of Crime Types

In the next section of the EDA, the relationship between poverty levels and total violent and non-violent crime was investigated. Median house-hold income was plotted against violent and non-violent crime rates, shown in Figure 5. Clearly, lower incomes cause higher crime rates, and it appears as though this would be a good predictor of crime. Interestingly, non-violent crime rates seem to be less sensitive to a change in income.

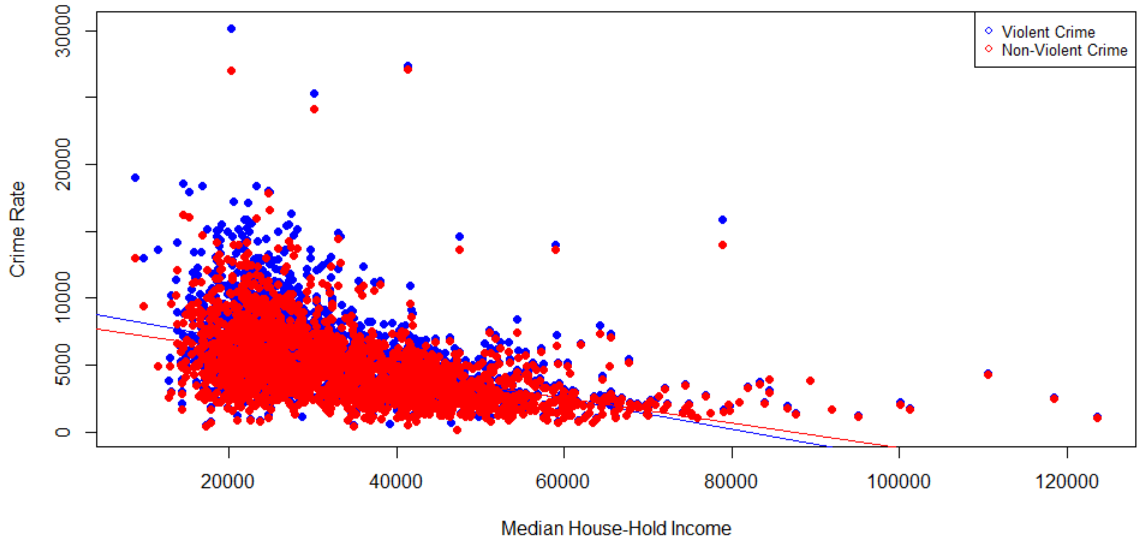


Figure 5: Relationship between crime rates and income

There are many covariates in the dataset that are likely to be correlated with household income. To better understand some of these relationships, a pair plot between median household income, poverty rates, unemployment rates and percentage of non-high school graduates was generated, as seen in Figure 6. It is clear that lower incomes correspond to a higher poverty and unemployment rate, and a less educated community.

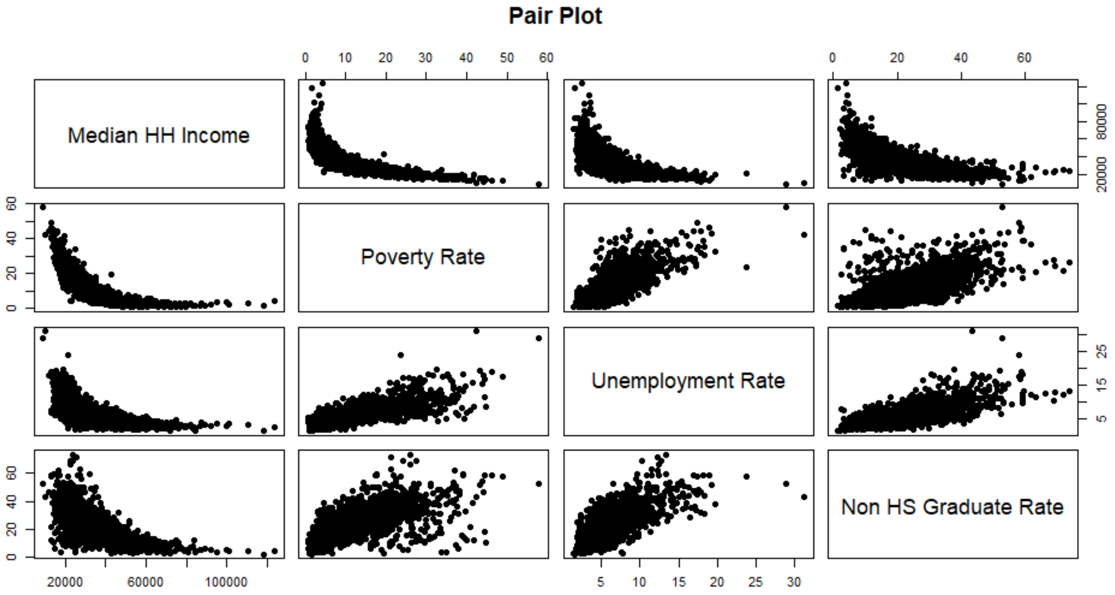


Figure 6: Relationship between crime rates and income

Next, the relationship between family structure and crime rates was investigated. Figure 7 shows that the higher the percentage of children who grew up with a two-parent household, the lower crime rates are. Once again this relationship is more sensitive for violent crime rates, as opposed to non-violent crime rates. Interestingly, Figure 8 shows that there is not a clear relationship between crime rates and the percentage of children under 6 with mothers in the labor force.

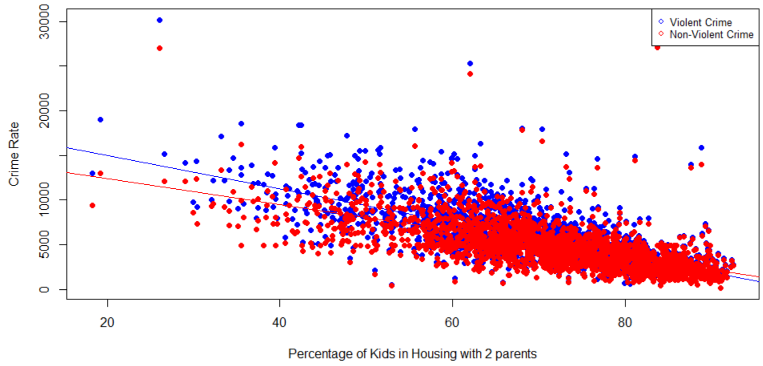


Figure 7: Relationship between crime rates and percentage of children growing up with two parents in their household.

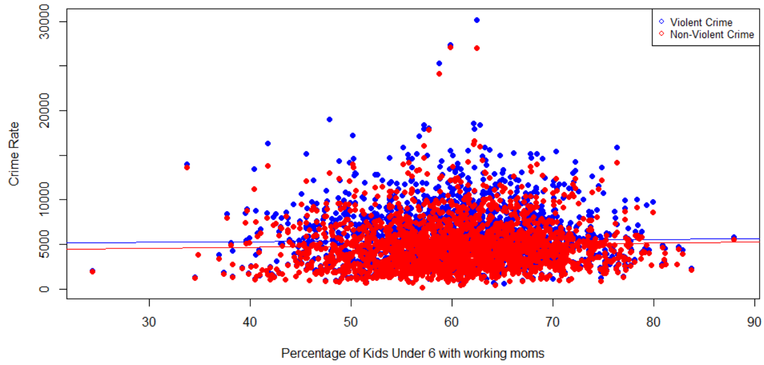


Figure 8: Relationship between crime rates and percentage of children under 6 with mothers in the labor force.

Finally, to understand how three variables relating to family structure are related, a pairs plot was generated between a community's divorce rate, percentage of children living with both parents, and the percentage of children born to never-married parents, shown in Figure 9. Again, a clear relationship is shown that higher divorce rates or percentage of children born to never-married parents decreases the likelihood that children live with both of their parents. Also, communities with high divorce rates tend to have more children born to non-married parents.

This EDA has exposed that the data represents a realistic distribution of communities in terms of geography, size, and the amount of crime experienced. We determined what types of crime are prevalent, understood the distribution of crime frequencies, and that there is a strong correlation between crime types. Finally, we explored the relationship between different covariates relating to poverty, as well as those relating to family structure, which are seen as two of the main root causes of crime.

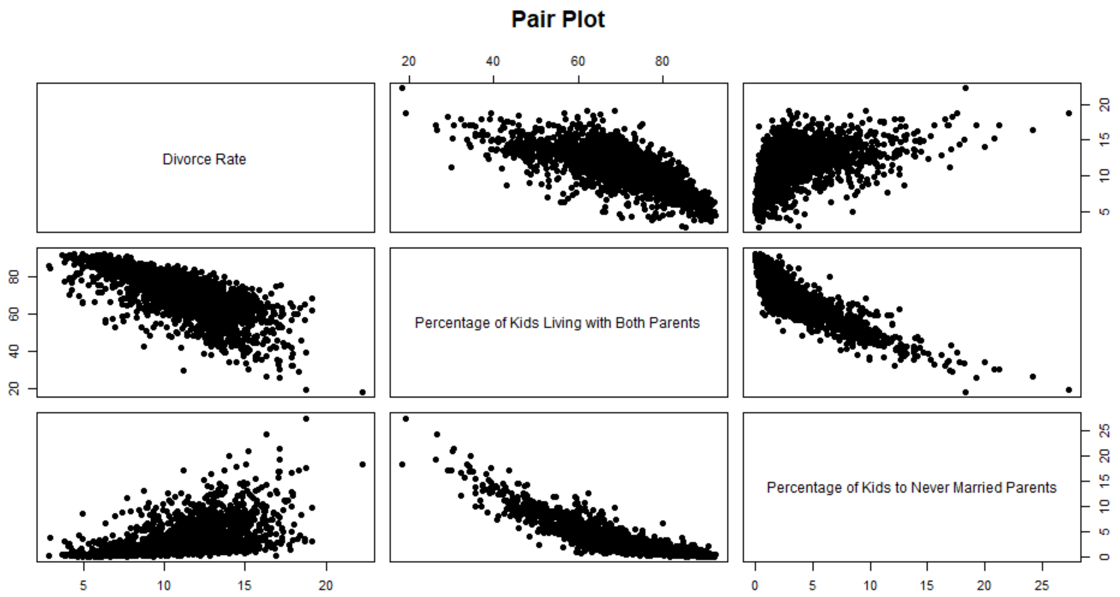
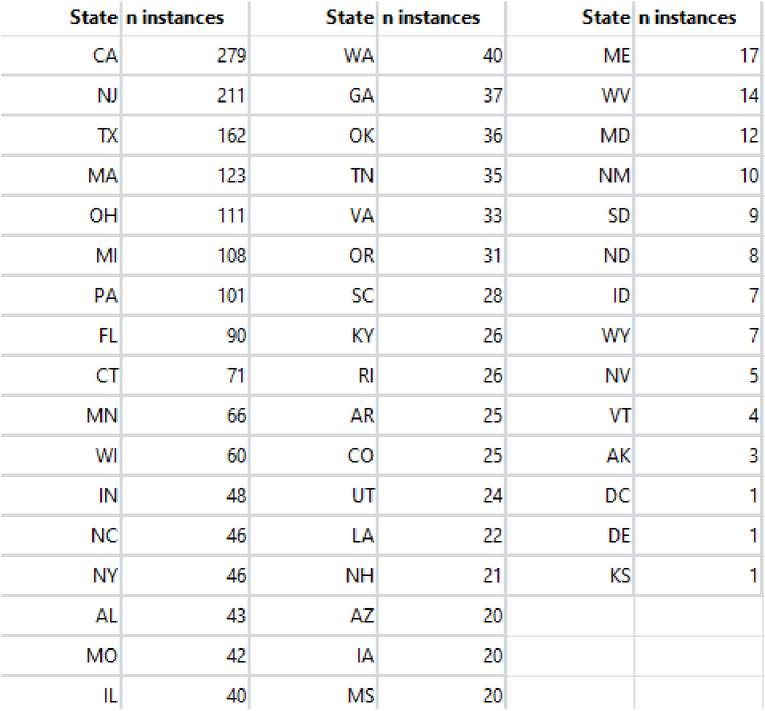


Figure 9: Pair plot between family structure variables

5 Appendix

Total Instances by State:



State	n instances	State	n instances	State	n instances
CA	279	WA	40	ME	17
NJ	211	GA	37	WV	14
TX	162	OK	36	MD	12
MA	123	TN	35	NM	10
OH	111	VA	33	SD	9
MI	108	OR	31	ND	8
PA	101	SC	28	ID	7
FL	90	KY	26	WY	7
CT	71	RI	26	NV	5
MN	66	AR	25	VT	4
WI	60	CO	25	AK	3
IN	48	UT	24	DC	1
NC	46	LA	22	DE	1
NY	46	NH	21	KS	1
AL	43	AZ	20		
MO	42	IA	20		
IL	40	MS	20		

Figure 10: State Counts

References

[1] Dakalbab, F., Abu Talib, M., Abu Waraga, O., Bou Nassif, A., Abbas, S., and Nasir, Q. (2022). Artificial intelligence crime prediction: A systematic literature review. *Social Sciences Humanities Open*, 6(1):100342.

[2] He, J. and Zheng, H. (2021). Prediction of crime rate in urban neighborhoods based on machine learning. *Engineering Applications of Artificial Intelligence*, 106:104460.

[3] Klimczuk, A. (2015). *Causes of Crime*, pages 308–311.

[4] Nitta, G., Rao, B., Sravani, T., Ramakrishiah, N., and Muthu, B. (2019). Lasso-based feature selection and naïve bayes classifier for crime prediction and its type. *Service Oriented Computing and Applications*, 13.

[5] Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L., and Chen, J. (2022). Interpretable machine learning models for crime prediction. *Computers, Environment and Urban Systems*, 94:101789.