Ben Ghouzi, Sasha DiVall, Emma Rabbath

Project Final Report

I. Problem Statement and Background

The FBI states that there is "definitely a correlation between the number of people who do (or do not) attain a high-school diploma and/or post-secondary education and higher violent crime rates". The project intends to investigate whether there is a correlation between the rate of crime in various states of the United States and the level of education in those states. Specifically, the research aims to ascertain whether a lower level of education correlates with a higher crime rate. This study is pertinent as understanding such correlations can help in informed policy-making and allocation of resources in sectors like education and law enforcement.

The motivation for undertaking this research is to highlight the importance of education. If a clear correlation is established, it can serve as a tool to show the role of education in reducing crime rates. Unfortunately, not everyone is able to receive an education, the main reason being financial. We hope to make education more accessible.

II. The Data

Education Data

The education data was sourced from the United States Census, 2020 from table S1501: Educational Attainment. This table includes total counts and percent values for each state for the following categories: educational attainment by age, educational attainment by race, and educational attainment by median household income. For this project, educational attainment by age was of primary interest because it gave the most general overview of educational attainment in the United

States. This data was accessed via a CSV file download that was made available on the Census website. Since this download came from a government website, there were no concerns over ethics and privacy. In regards to potential bias, although all Americans are sent a census every ten years, it is not guaranteed that every American responds to the census. This presents a nonresponse bias. In particular, those with a low educational attainment may be reluctant to respond to this particular question on the census, which may cause the *actual* rate of Americans who have a low educational attainment higher than presented in the final census data.

Crime Data

The crime data was sourced from the Federal Bureau of Investigation's National Incident-Based Reporting System (NIBRS) data on crimes against persons in the United States in 2022. This dataset also contained data on crimes against society and crimes against property, however crimes against persons was selected because it was of the opinion of the group that this type of crime affects the day-to-day life of Americans the most. The data was accessed from the FBI's Crime Data Explorer, an interactive website which provides dynamic maps with details regarding crime in each state as well as an ability to download files containing the data. Similar to the education data, the data was accessed via CSV file download, and since it came from a government website, there were no ethical or privacy concerns when accessing the data. The NIBRS depends on individual agencies across the United States to report the number of crimes that were reported to their agency in a given time. Therefore, a potential bias arises if there exist agencies in the United States that do not participate in the NIBRS or if agencies do not report crimes with perfect accuracy. If this bias exists, the actual number of crimes committed in a particular state may be higher than presented in the data.

Additional Datasets

Additional data was included in the creation of the K-nearest neighbors classifier to examine the potential cofactors contributing to crime in the United States. The first dataset included was Underlying Cause of Death from the Center for Disease Control and Prevention (CDC), 2018-2021. The CDC provides a querying interface called CDCWonder which allows users to group by particular attributes and focus on particular causes of death. Diseases are classified by ICD-10 codes, and for this query, the ICD-10 codes used are associated with unintentional drug overdose. In the query, data was grouped by state and age was filtered to include only ages 25 and over. Ages 18 to 24 were excluded because the education data included educational attainment for those aged 25 and over. The results of the query were stored via a tab-delimited TXT file. Since this data came directly from a government website, there were no ethical or privacy concerns. Similarly, the CDC does not collect data on a voluntary response basis, therefore there is limited concern over potential biases present in the data.

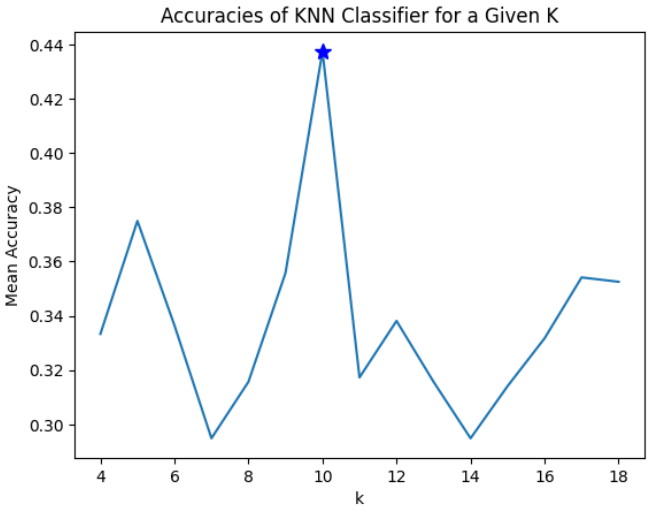
The second dataset included in the classifier algorithm is Prevalence of Any Mental Illness in the United States from Mental Health America, 2022. Mental Health America is a non-government organization whose mission is to improve mental health across the United States. The website provides state percentages for residents in each state reporting having any mental illness. Mental Health America draws from several surveys to compile their own annual report, including SAMHSA's National Survey of Drug Use and Health and CDC's Behavioral Risk Factor Surveillance System. The data was accessed via web scraping, which presents some concerns of privacy. Because the data used by Mental Health America comes from several other sources, there could be inconsistencies in data reporting across all the surveys. Additionally, if the surveys are offered on a voluntary basis, those who suffer from a mental

illness may be reluctant to report this. If these biases exist, the actual rate of mental illness in each state would be different than the rate that is reported by Mental Health America.

III. Data Science Approaches

K-Nearest Neighbors Classifier To Determine Crime Level

The K-nearest neighbors classifier technique is a supervised machine learning algorithm. Given a data point with an unknown class value, the algorithm compares the datapoint to the known class of its K-nearest neighbors. The term "neighbors" refers to the points the shortest Euclidean distance away from the unlabeled point. K refers to an integer value and can be any number up to but not including the number of total points in the dataset. Class values were assigned to each state to determine its crime level as "low" (0), "medium" (1), or "high" (2). A state was deemed to be low crime if the total crime offenses as a percentage of the covered population (crime rate) was lower or equal to the first tercile. A state was deemed to be medium crime if the crime rate was between the first and second tercile. A state was deemed to be high crime if its crime rate was higher than the second tercile. The data used to build the classifier included the percent of people in each state that met each educational attainment, the crude rate of unintentional drug overdose, and the prevalence of mental illness in that state. Due to the varying range of values in each of these datasets, these columns were normalized using a minimum-maximum normalization technique. The model is built by splitting the data into a training and testing set. The model is fit using the training set, and the model's accuracy is determined by applying the algorithm to the testing set. In this classifier, the



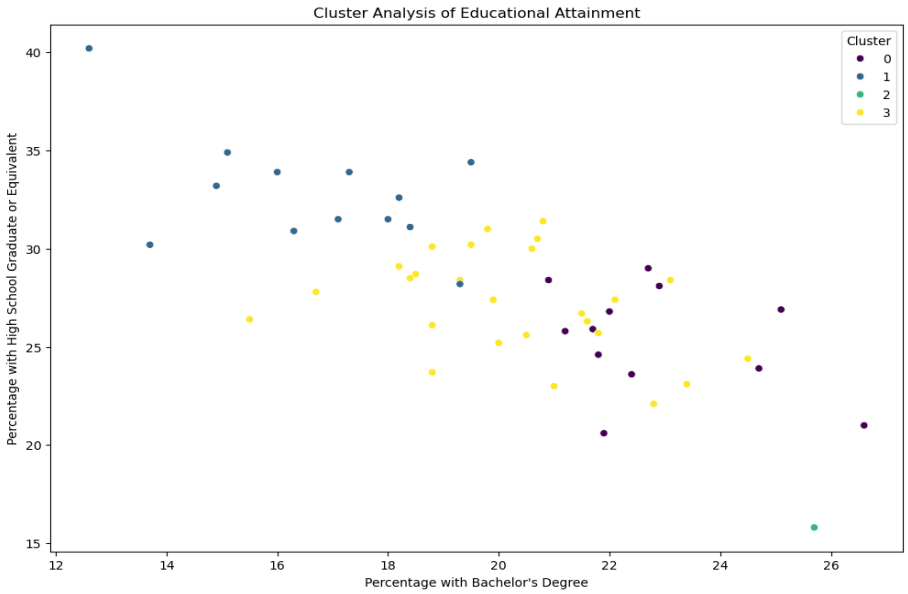
optimal value of K was chosen based on accuracy, which in this context is defined as the percent of testing set values whose class label was correctly predicted by the model. The plot above shows that the maximum accuracy of 43.75% occurs when K = 10. Thus, the classifier in this algorithm was built using a K value of 10, meaning each testing data point was compared to its ten nearest neighbors. After fitting the model and predicting the classes of the testing data, the final accuracy of the model was 46.15%.

Histograms and Clustering as an Analysis of Education

The analysis of educational levels across different states in the United States was approached through the development of a series of histograms. These histograms were designed to categorize the population's educational attainment into distinct levels: 'Less than 9th Grade', '9th to 12th Grade, No Diploma', 'High School Graduate', 'Some College, No Degree', 'Associate Degree', 'Bachelor Degree', and 'Graduate or Professional Degree'. This method offered a visual representation of the distribution of educational levels across various states. In constructing the histograms, the y-axis was utilized to represent all the states in the U.S., while the x-axis depicted the percentage of the population in each state achieving a specific educational level. This configuration facilitated a comparative analysis of educational attainment across states.

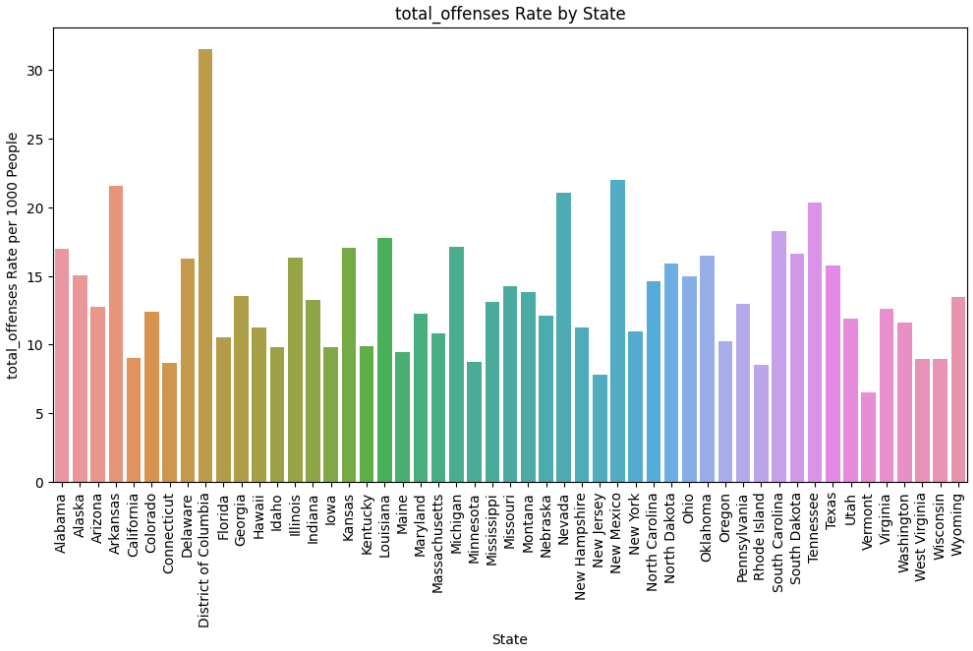
The deployment of histograms enabled an effective visualization of the educational spectrum in each state. This approach allowed for the identification of patterns and trends in educational attainment, such as highlighting states with higher percentages of higher education degrees or those with significant portions of the population having lower educational levels. The comparative nature of these histograms made it easier to discern regional educational disparities and similarities.

The cluster analysis was conducted using a K-Means clustering algorithm - an unsupervised machine learning tool. The cluster analysis presented in the plot categorizes states into four distinct clusters based on the percentages of high school graduates and Bachelor's degree holders, each distinguished by a unique color. Cluster 0 (purple) represents states with high percentages of both high school and Bachelor's degree graduates, indicating strong overall educational attainment. Cluster 1 (blue) includes states with moderate to high high school graduation rates but lower Bachelor's degree holders, suggesting a gap in higher education completion. Cluster 2 (Green) is characterized by states with notably high high school graduation rates, yet a lower proportion of Bachelor's degree holders, implying a focus on general education over higher education. Cluster 3 (Yellow) consists of states with moderate levels of high school graduates and fewer Bachelor's degree holders, possibly indicating a general trend with potential for growth in higher education.



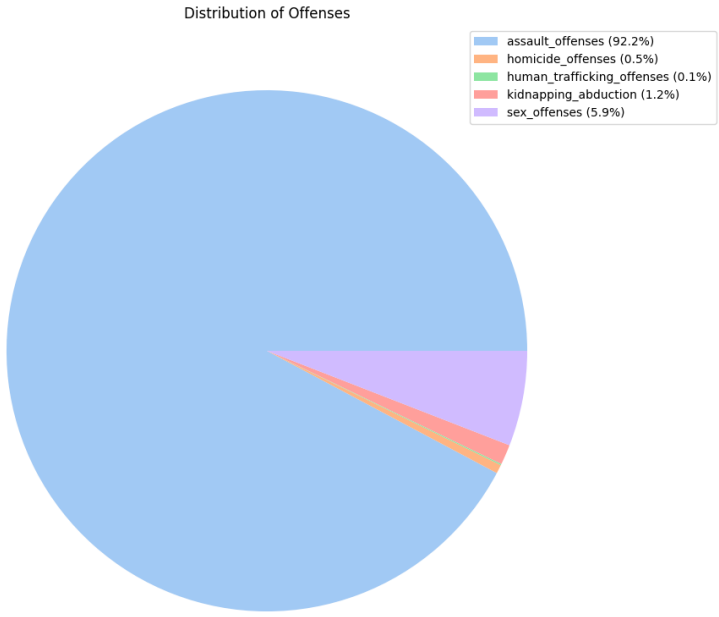
These clusters offer insights into educational strategies. States in Cluster 0 likely have effective educational policies promoting both high school and higher education completion. Conversely, states in Clusters 1 and 3 might benefit from policies that encourage higher education attainment. The outlier state in Cluster 2 warrants further investigation to understand its unique educational landscape.

Conducting an Analysis of Crime Data in the U.S.



We first look at the distribution of total offenses per 1000 people per state. The data does not fluctuate a lot, we are able to distinguish one outlier which is the District of Columbia, which stands at 31.5 offenses per thousand of people. However, after conducting an analysis, we can see that the Southwest and Southeast regions have a higher average crime rate per 1,000 people than other regions. It would be interesting to dig deeper into the underlying factors contributing to these disparities. If education is not necessarily the answer, are there socioeconomic factors, demographic differences, or specific policies that may explain the variations in crime rates.

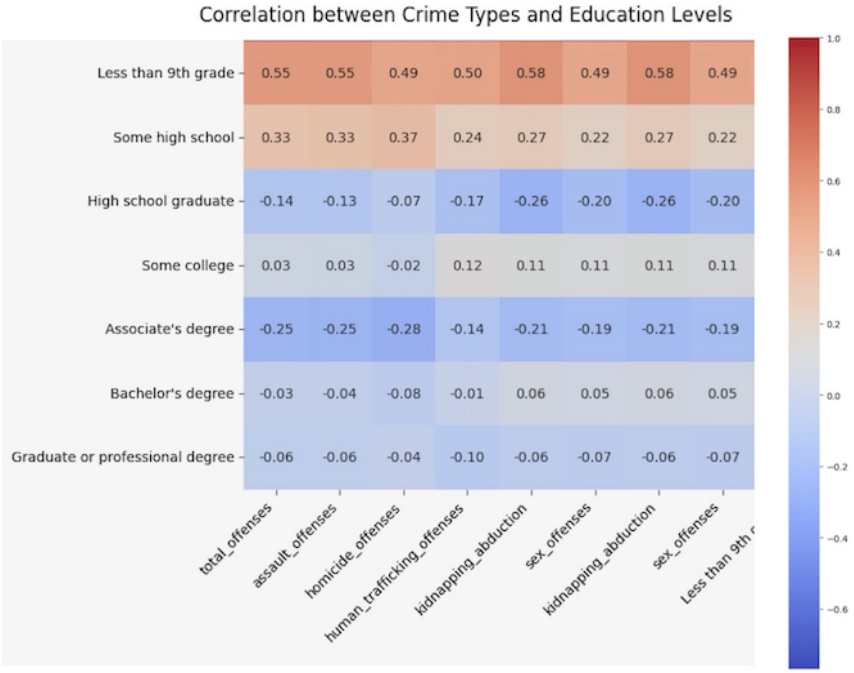
The pie chart illustrates a focus on assault offenses, which represents 92.2% of total offenses. Even though this should urge the importance of establishing a strategy, we cannot overlook the presence of other offenses. Sex Offenses and Kidnapping/Abduction represent 5.9% and 1.2% respectively. While Homicide and Human



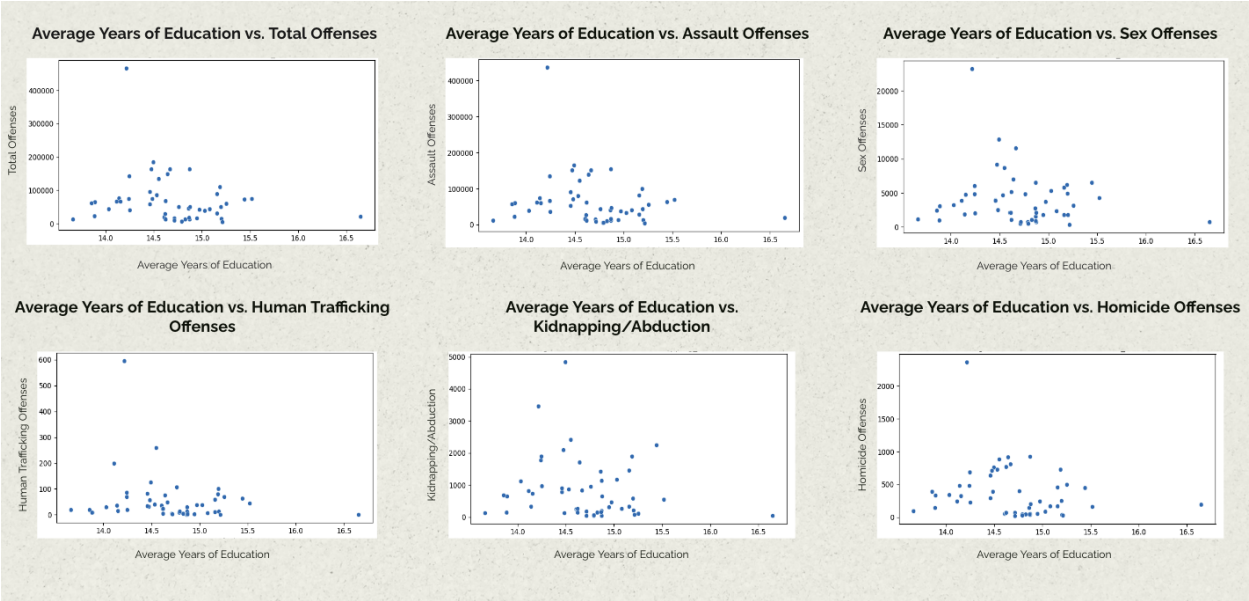
Trafficking Offenses are on the lower end, the severity of these offenses still demands interventions.

Heatmap Visualizing the Correlation Between Education and Crime Rate in the U.S.

In the project's exploration of the relationship between educational attainment and crime rates across various U.S. states, heat maps were employed as a pivotal analytical tool. This technique involved juxtaposing the data on educational levels with crime statistics to discern any underlying patterns or correlations. The heat map was constructed with axes representing different variables: one axis displayed the states, another illustrated various educational levels (ranging from 'Less than 9th Grade' to 'Graduate or Professional Degree'), and a third dimension, indicated through color intensity, represented crime rates. This multi-dimensional approach allowed for a nuanced analysis of how educational attainment and crime rates interrelate across states. The color gradients in the heat map provided a visual representation of the intensity of crime rates in relation to educational levels. This visualization enabled the identification of trends, such as whether states with higher percentages of higher educational attainment experienced lower crime rates, or vice versa. The heat map's ability to represent complex multi-variable data in a comprehensible and visually engaging manner was essential for drawing meaningful conclusions about the correlation between education and crime.



Scatter Plot Visualizing the Relationship Between Education and Crime Rate in the U.S.



Through our analysis of scatter plots, we seek to understand the relationship between crime rates and the levels of education. If there is a relationship, as the average years of education gets higher, the crime rate should get lower. In each scatter plot, every data points signifies the relationship between average years of education and the corresponding total crime rate for a specific state. Globally, most dots lie between 14 and 15.5 years of education, suggesting some college experience across states. Interestingly, beyond the 15.5 average years of education, only one state remains, the District of Columbia (DC), with an average education exceeding 16.5. Contrarily to our initial assumption, the crime rate in DC, despite its high average of education years, is not the lowest. This leads us to conclude that there isn't a strong correlation between education and crime rates.

IV. Results and Conclusions

The low accuracy of the KNN classifier raised some eyebrows in the group about how well education predicted crime rates. Upon further investigation, the correlation heatmap matrix shows

that although there is a positive correlation between low educational attainment and high crime rate, and a negative correlation between high educational attainment and low crime rate, neither of these correlations are very strong. The highest correlation in the matrix is between an educational attainment of Less than 9th Grade and kidnapping abductions, with a value of 0.58. The lowest correlation values occur when comparing the percentage of Americans with a Bachelor's degree and each crime type, each having a correlation value of almost zero. Therefore, we conclude that education is not a key determinant of crime in the United States.

V. Future Work

The results of our project, delving into the correlation between education levels and crime rates in the United States, have provided valuable insights into American societal structures. However, the discovery of a weak correlation in this area suggests the need to explore other potential influences on crime rates. A promising next step would be to investigate the relationship between crime rates and income levels across various states. This future work would involve collecting and analyzing comprehensive data on income metrics such as average household income and poverty rates. Employing statistical tools for correlation analysis and utilizing visualization techniques like heat maps, the study aims to examine whether lower income levels correlate with higher crime rates. Additionally, a comparative study with our previous findings on education and crime could offer a broader understanding of the socio-economic factors affecting crime. Such an analysis is anticipated to provide deeper insights into the socio-economic dynamics influencing criminal activities and could significantly contribute to developing more targeted and effective crime prevention strategies.

VI. Works Cited

1. U.S. Census Bureau. "EDUCATIONAL ATTAINMENT." *American Community Survey, ACS 5-Year Estimates Subject Tables, Table S1501*, 2021, https://data.census.gov/table/ACSST5Y2021.S1501?g=010XX00US$0400000. Accessed on December 4, 2023.
2. Federal Bureau of Investigation. “NIBRS Offenses By Agency. ” 2022. https://cde.ucr.cjis.gov/LATEST/webapp/# . Accessed on December 4, 2023.
3. “The Clear Correlation between Education and Crime.” *Esfandi Law Group*, 24 June 2022, esfandilawfirm.com/correlation-between-education-and-crime/.