Relationship Between High School Graduation and Robbery Rates

Armandeep Kaur, Joe Posillico, Aaron Spotts, Sarah Todd

University of North Carolina at Charlotte

Dr. Nadia Najjar, Dr. Kenzie Reed

October 7, 2024

Abstract

This study explores the potential for reducing robbery rates by increasing high school graduation rates and considering multiple different economic factors, using the "Communities and Crimes" dataset, which combines data from the 1990 U.S. Census and the FBI's Uniform Crime Reporting program. The analysis focuses on predicting a decrease in robbery rates as high school graduation rates rise. However, the baseline model indicated that graduation rates alone are not sufficient to drive a significant reduction in robberies. Therefore, additional factors, such as economic conditions, family structure, and public assistance, are introduced to better understand their combined impact on crime rates and to develop more effective crime prevention strategies for policymakers.

Introduction

The total annual cost of crime on society is estimated to be around $4.71–$5.76 trillion with robbery accounting for roughly 6% or $294 billion of this annual amount (Anderson, 2021, p877). To put this into context, the cost of robbery alone is equal to roughly 2.5% of the country's gross domestic income which is quite a significant amount. Robbery is defined by the FBI as “the taking or attempting to take anything of value from the care, custody, or control of a person or persons by force or threat of force or violence and/or by putting the victim in fear” (FBI, 2019). Along with the financial burden that robberies put on society, the violent nature associated with this type of crime also impacts the general well-being of all individuals in society, especially in areas with higher-than-average percentages of robberies being committed.

Because of the substantial financial impacts combined with the psychological and social toll on communities, robbery is a crime that must be addressed and ways of reducing this crime needs to be explored. This project will explore the notion that robbery is directly affected by high school graduation rates, with the hypothesis being that an increase in high school graduation rates will in turn directly lead to a decrease in robbery rates. Due to this question placing such high importance on high school graduation, it is important to understand what impacts graduation rates and therefore relationships between high school graduation rates and factors such as median household income, parental marital status, and housing costs. These factors will be explored in order to determine the best way to increase high school graduation rates which will in turn decrease robbery rates. These findings can be provided to policymakers to help them make decisions about crime reduction. While reforms can be expensive, if this study can show a strong relationship between these high school graduation rates and robbery it could help policymakers justify investments in increasing high school graduation.

Background

The current theories discuss higher dropout ages having an “incapacitation” effect where young people are not committing crimes because they are in school rather than having free time on the street. This incapacitation effect continues even after students graduate because it prevents them from building habits of crime, as well as prevents students from accumulating criminal histories in formative years of social development. Criminality at a formative age is highly predictive of criminality later in life. While this may be due to the incapacitation effect or other factors, it shows the important relationship between education and crime (Jacob & Lefgren, 2003).

Each year, an estimated 600,000 students drop out of U.S. high schools (Rumberger, 2011, p132). Although each student has individual circumstances that may influence their decision to drop out, examining the social, economic, and other factors that contribute to lower graduation rates can derive important results. One massive factor that may contribute to an individual's decision to drop out of high school is their family circumstances. Not all children get to live with both their parents, which could be due to divorce. This can have a big impact on a family's life financially. Children living with single parents are less likely to experience upward financial mobility (Anderson, 2014) and high school-aged children with divorced parents are more likely to drop out, regardless of the introduction of a step-parent (Sandefur et. al., 1992).

Therefore, children living with their married, biological parents consistently have better physical, emotional, and academic well-being (Anderson, 2014).

Another large factor that contributes to a student's decision to drop out of high school is their school and learning environment. When students fail courses, get low test scores, have poor attendance, or are disinterested in graduation, they are more likely to drop out (McDermott et. al., 2019). One of the most important factors, however, is a student's financial situation. There are many different ways that this can affect their education or home life. There is a negative correlation between a state's child poverty rate and student performance (Ladd, 2012), which could be due to factors such as teacher behavior. A child attending school in a high-poverty area is more likely to receive academic punishment, as the teachers are less trained and tend to report children for their behavior more than teachers in higher-income areas (McDermott et. al., 2019). Thus, it is extremely important to develop solutions that can counteract the effects of these factors on high school graduation rates. One shortcoming is that the previously discussed literature only looks at crime data with precisions such as violent, nonviolent, property, or drug crimes. This paper aims to fill this gap by looking at a specific type of crime, robbery, to see if trends present in broader categories of crime hold for this specific type of crime. Robbery is a crime where the motivations are property-related but the methods are violent. This puts it in a liminal state Another limitation of the previously discussed literature is the lack of control for economic factors.

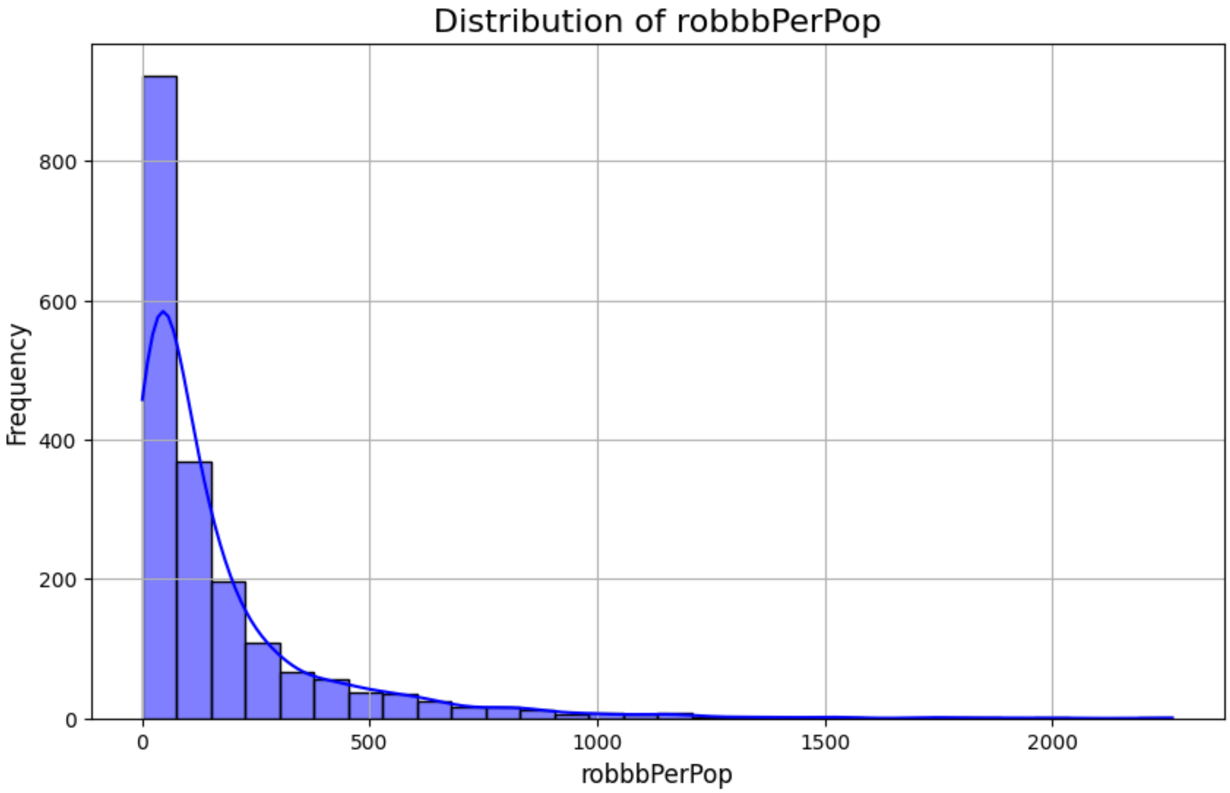
Dataset Description

Prior to creating models and exploring the data set it is important to fully understand the data and where it is coming from. The dataset that will be used to test the question this project explores is called "Communities and Crimes" which was published in 1999 on the University of California, Irvine machine learning repository. This dataset was created by gathering data through different Government resources, which includes the 1990 United States Census which collects data on the demographic, social, and economic characteristics of the U.S. population, also from the FBI UCR or Uniform Crime Reporting program which is used to collect, compile, and publish crime statistics from law enforcement agencies across the United States. Data was also collected through LEMAS, the Law Enforcement Management and Administrative Statistics survey which collects all types of information about law enforcement such as management, operations, and administrative functions of state and local law enforcement agencies. It is worth noting that there is a known limitation with the LEMAS survey due to the fact that it was only taken from police departments that had at least 100 police officers which means that smaller counties that do not have as large of a police force did not partake in this survey which could cause biases and skew in the data.

The "Communities and Crimes" data set itself contains 2215 individual rows as well as 147 different columns which all pertain to different types of metrics regarding policing and community information. With 147 different variables, there are several variables that can be used to help understand the question at hand. Although at first glance there are no null values inside this dataset, the data set still contains a lot of missing data which is represented by a "?" in the missing fields. In total, there are 2104 rows out of the 2215 total rows that contain at least one column with a missing value. Although this amount of missing data indicates that 95% of the rows have some form of missing data it is important to note once again that there are 147 different columns so although 95% of the rows have some columns missing data there is still valuable data that can be used.

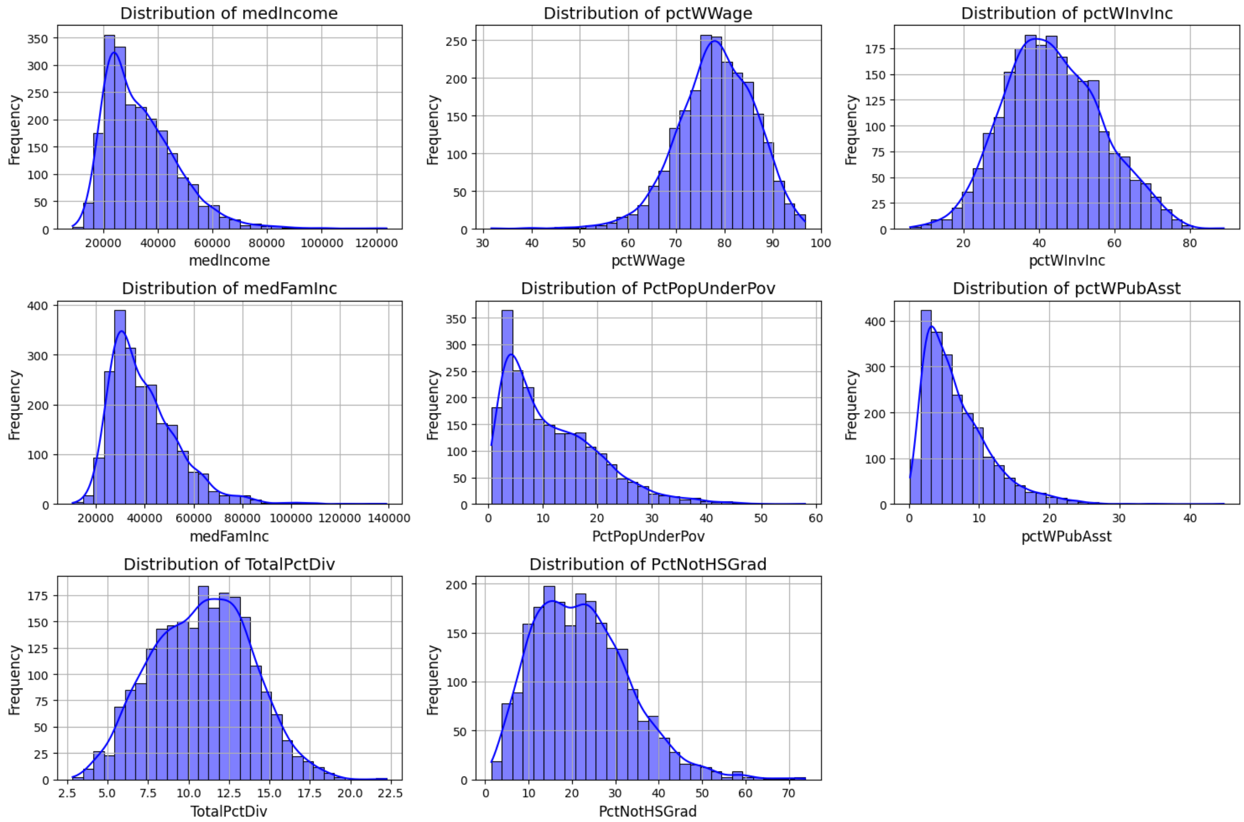
Methodology

The target variable for this research is robberies per population. The reason this was chosen as the target variable is because the amount of robberies per population is exactly what this project is trying to predict. The objective is to see a decrease in this target variable and determine which predictor variables have the greatest effect on reducing this type of crime. It is important to access the distribution of data for this variable, which is done with the following histogram:



This histogram shows the frequency of robberies per population and the main takeaway from this is that the distribution is quite heavily skewed rightward with a long tail. This is something that will possibly need to be addressed in order to avoid potentially biased predictions, as well as the fact that most machine learning algorithms work under the assumption that the variables are normally distributed.

The predictor variables chosen are as follows: percent of students without a high school degree, medium income, percentage with wage, percentage of people with investment income, median household income, number of people under the poverty line, percentage of population living under the poverty line, marital status and percentage of people acquiring public assistance. The predictor variables chosen for this analysis reflect significant social and economic factors that influence crime rates. The percentage of students without a high school diploma is particularly important, as low educational attainment has consistently been linked to higher rates of crime (Machin, 2012). Additionally, economic indicators such as median income, the percentage of individuals with wages, and those relying on public assistance provide insights into the financial well-being of a community, which correlates with crime levels (Jin, 2008). By analyzing these variables, this project seeks to investigate how enhancing education and economic conditions might lead to a decrease in robbery rates. The primary aim is to determine if targeted interventions focusing on these aspects can successfully reduce robbery rates, thus providing valuable insights for communities, law enforcement, and educational institutions in their decision-making processes. Similarly to how the distribution visualization was created for the target variable the same was done for the predictor variables as follows.



This shows once again that the data being used is not always normally distributed and perhaps will have to be addressed prior to modeling. Due to a lot of these variables being rightward skewed it may be necessary to eliminate outliers to address this prior to modeling.

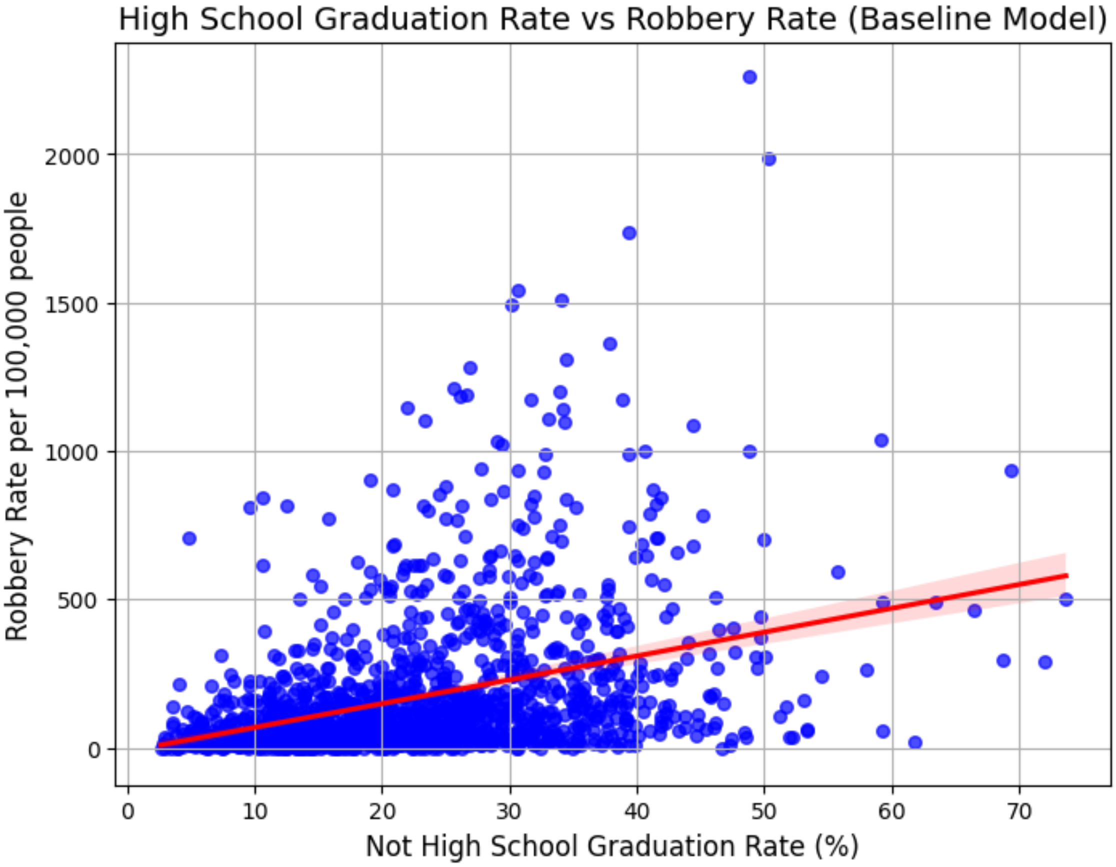
For the last step of the data exploration process, it is important to understand what types of data are being worked with. The entirety of the variables that have been previously presented are of integer or float values, meaning that they are all numeric. This points to linear regression as the most useful model. With this being said, minor manipulation was made to one predictor variable which is the percentage of non-high school graduates. This variable was turned from integer values to categorical values, this was done by splitting this variable into percentage ranges such as 0-10%, 10-20% all the way to 90-100%. This was done because transforming the variable into categorical ranges allows for a clearer interpretation of its impact on robbery rates and can help identify specific thresholds where changes in educational attainment may significantly influence crime levels.Categorizing the data allows for more effective analysis of trends and patterns related to educational attainment and its correlation with robbery rates.

The final step in the data preparation process was to explore the two data frames to make sure that no outliers were present. The reason this is needed to ensure no outliers are present is because the presence of outliers could negatively affect modeling and potentially skew hypothesis testing due to inaccurate data. The approach to cleaning up the outliers was to use a z-score in order to determine whether outliers are present or not. Z-score is a measure of how far a data point is from the mean of the data set with respect to standard deviation. The formula for z-score is as follows: $z = (X-\mu) / \sigma$, with X representing the data point, μ representing the mean of the data set, and σ representing the standard deviation of the data set. The goal of using the z-score is to determine whether the data point being checked is greater than 3 or less than -3, if so this indicates that the data point is clearly different from the other data points and therefore would be dropped due to being an outlier. Upon conducting z-scores on both of the two datasets; crimes and income, it was clear there were 164 outliers to be removed in the crimes data frame and 111 outliers to be removed in the income data frame.

Setup and Models

The first step in setting up the data in order to have it ready to use for modeling is to partition said data into usable subsets that pertain to the problem at hand. The dataset was partitioned vertically by extracting the relevant fields into a separate data frame. This data frame consisted of the fields robbbPerPop, PctNotHSGrad, totalPctDiv, and the relevant economic factors (medIncome, pctWage, pctWInvInc, medFamInc, NumUnderPov, PctPopUnderPov, and pctWPubAsst). The partitioned dataset did not include any null values, as all empty entries were filled with the string "?" in the original data frame instead of leaving them null. This was investigated using the .isna() function, which returns the number of non-null values in each column. Searching for the string "?" ,however, it was found that there were 313 rows with missing data. This accounts for 14.13% of the data. It would be inaccurate to fill the missing data with pseudo-data, as the fields pertain to data about specific locations. Therefore, the rows that include missing data were dropped.

To evaluate the effectiveness of a more complex model, an initial baseline model was created. This baseline model is a linear regression between the percentage of people who did not graduate High School and the rate of robberies per population. The R-Squared value of this model was .1154. This means that there is an 11.54% variance in robbery rate based on the rate or people who did not graduate High School. The Mean Squared Error (MSE) is 47390.5758 and the Mean Absolute Error (MAE) is 143.6426, meaning that the predicted values are generally inaccurate and far from the actual values.



Conclusion

This research question aims to predict the relationship between educational attainment and robberies. More specifically, this question is trying to predict if an increase in high school graduation rates leads to a decrease in robbery rates. This information can help schools make decisions on how they can increase student graduation and help law enforcement understand what could be causing crime in areas. This information will also be useful to families when making decisions for their children’s futures. Limitations to this work include that the data was collected as a summary of whole counties and could have been more precise if collected on a smaller level or tracking outcomes of individuals. Additionally, while examining robberies provides a more precise view than prior research, robbery is not the only type of crime. Further research is needed to look at other types of crime and see if they hold similar trends.

Works Cited

Anderson, D. A. (2021). The aggregate cost of crime in the United States. *The Journal of Law and Economics,* 64(4), 857–885. https://doi.org/10.1086/715713

Anderson, J. (2014). The impact of family structure on the health of children: Effects of divorce. *The Linacre Quarterly,* 81(4), 378-387. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4240051/

Jacob, B. A., & Lefgren, L. (2003). Are idle hands the devil's workshop?: Incapacitation, concentration, and juvenile crime. *The American Economic Review*, 93(5), 1560–1577. https://doi.org/10.1257/000282803322655446

Jin, M. (2008). Does Unemployment Increase Crime? Journal of Human Resources, 43(2), 413-436. https://doi.org/10.3368/jhr.43.2.413

Ladd, H. F. (2012, March 8). Education and Poverty: Confronting the Evidence. *Journal of Policy Analysis and Management,* 31(2), 203-227. https://doi.org/10.1002/pam.21615

Machin, M., Marie, O., Vujic, S. (2012). Youth Crime and Education Expansion. *German Economic Review* 13(4), 366-384. https://doi.org/10.1111/j.1468-0475.2012.00576.x

McDermott, E. R., Donlan, A. E., & Zaff, J. F. (2019). Why do students drop out? Turning points and long-term experiences. *The Journal of Educational Research (Washington, D.C.),* 112(2), 270–282. https://doi.org/10.1080/00220671.2018.1517296

Rumberger, R. W. (2011). Dropping out: Why students drop out of high school and what can be done about it (1st ed.). *Harvard Univ. Press.* https://doi.org/10.4159/harvard.9780674063167

Sandefur, G., McLanahan, S., Wojtkiewicz, R. (1992) The Effects of Parental Marital Status during Adolescence on High School Graduation, *Social Forces*, 71(1), 103–121. https://doi.org/10.1093/sf/71.1.103

U. S. Federal Bureau of Investigation. (2019). *Robbery.* https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/robbery

Code:

https://github.com/joepo95/DTSC-Project-Group-3