W203 Lab 3 | Reducing Crime in North Carolina

Gordon Jack, Anna Jacobson, and Sarah Reed

08/07/18

1.0 | Introduction

It was once thought that criminality was inherited (Lombroso, 1911). Just as an individual is born with blue eyes, a person could also be born a criminal. We now understand that there are many external factors that influence the incidence of crime - factors that may be positively impacted by policy decisions. Our client requires data-driven insights regarding crime, race, and justice in preparation for the 1988 North Carolina gubernatorial race. We are prioritizing the areas of focus as 1) reducing violent crime and 2) reducing nonviolent crime.

Crime can be modeled economically as the outcome of benefit versus risk of an illegal activity (Becker, 1968). Crime may become more attractive to those that are poor and lack legal alternatives for economic gain, increasing the perceived benefit relative to the risk. To that end, we consider methods of increasing legitimate economic opportunity. The risks of illegal activity involve the certainty of getting caught and the severity of the resulting punishment. We explore potential policy measures, supported by our analysis, that can affect this part of the equation.

We assert that extrinsic factors, such as the opportunity for crime and cultural influences, are also drivers of crime rate. We will evaluate which of these factors contribute most to the variation in crime rate. We hypothesize that the motivations for violent and nonviolent crime are different and assess which proposals will impact each type of crime.

In summary, our research questions are:

1. Is there a true relationship between a county's legal, economic, and sociodemographic conditions and its crime rate in North Carolina?

2. Do these conditions influence violent crime differently than nonviolent crime?

2.0 | Initial Data Loading and Cleaning

```
Crime = read.csv("crime_v2.csv")
```

Data Description

The dataset used in this analysis includes North Carolina county-level statistics from 1987. The data was initially gathered by Cornwell and Trumbull from the University of Georgia and West Virginia University; it has `97` observations with `25` variables. Our primary outcome variable is the crime rate (`crmrte`) which is the number of crimes per person for each county. Our independent variables can be grouped into legal, economic, and demographic categories and are sourced from the FBI's Uniform Crime Reports, the North Carolina Department of Correction, North Carolina Employment Security Commission, and 1980 census data.

Preliminary Data Checks

To inform and refine our hypotheses, we first perform an exploratory data analysis. However, examining our data may introduce bias or cause problems with reproducibility when performing tests on the same dataset. Therefore, we partition the data for our exploratory data analysis (EDA) and reserve the remainder of the data for our confirmatory data analysis (CDA). This allows us to better determine if the model specifications we develop are robust. Before data partitioning, we perform a number of "blind" data checks:

- First, we check the data for completeness and find missing entries for our key outcome variable (`crmrte`). We see that 6/97 (6.2%) of entries have no data for crime rate and are unusable in our analysis. We exclude these rows.
- We remove the rows for which the crime rate (`crmrte`) data is not given, leaving `91` observations. Checking again for completeness, we now find that we have 100% complete data for all variables.
- Subtracting the number of unique county identifiers (`county`) from the total number of observations in the data set shows that there is `1` duplicate. An additional check using R's duplicated function confirms that all entries for that row are duplicated (rather than just a miscoded value for county number) which means we should remove one of them. This leaves `90` observations.
- When checking the classes of the variables, it is discovered that the `prbconv` variable is classed as a Factor. We convert this variable to numeric for analysis.
- There are `35` counties that are not categorized to be in the `west` or `central` regions. This may indicate that there may be missing region indicator variables, such as `north`, `east`, or `south`. There is `1` observation reported to belong to both the `west` and `central` regions, which we believe is an error. However, we do not expect this erroneous data to have a great impact on our analysis, so we do not alter or exclude this observation.

Partitioning the Data

To partition the dataset, we randomly select a subset of half of the observations (`45` rows) to create our EDA partition. We place the remaining observations into another subset, which we reserve for our CDA.

```
set.seed(2018)
C1 <- Crime[sample(1:nrow(Crime), 45, replace=FALSE),]
C2 <- anti_join(Crime, C1)
```

Data Quality: EDA Partition

- Only `6` counties are categorized as `urban`, which is approximately `13` percent of the EDA partition. We want to include a metric that describes the local environment, but believe that `density` captures this with more granularity.
- Interestingly, both the `prbarr` and `prbconv` variables have positive skews with values over 1. We reason that since the `prbconv` represents the number of convictions per arrest, it is possible that multiple convictions can result from the same arrest. However, it is harder to justify the maximum value for `prbarr` (`1.09`), since this would indicate that there were more arrests than offenses. We will consider the impact of this data point in the model building process.
- From an independent review of the 1980 Census, we know the overall population density in North Carolina is 120 people per square mile, ranging by county from 9 to 766. However, the values of the `density` variable ostensibly in the same units range from nearly 0 to just under 9, which is not a reasonable range. We believe the `density` variable should be scaled by 100 for units to match the variable description. Even with this change to the density scaling factor, there is one unreasonably low value of `0.002`. We will consider the impact of this data point in the model building process.

- The `wser` variable representing the average weekly wage for those employed in the service industry for each county has an extreme outlier at $2177.07/week, which is much greater than the third quartile value of $270.49/week. We are unable to determine if this outlier is due to misreporting of the data or is far outside the norm but correct (potentially a few very high earners in a small sample size). We look at this in more detail in Section 3.0, where we explore dimension reduction of the many wage variables.
- We notice that the `pctymle` variable has values between 0 and 1, suggesting that it is a ratio rather than a percentage. To facilitate interpretability, we multiply this variable by 100 so that its unit is given as a percent. It also has a heavy positive skew. Performing a log transformation of the variable reduces the skew, but the majority of the data points remain towards the low end of the range. A log transformation of `pctymle` may also hinder interpretability since the model results will be framed as a percent change in the percent of young males. Therefore, no transformation on this variable is included in our model specifications.
- The `crmrte`, `polpc`, `density`, and `taxpc` variables all show positive skews in their distributions and have no negative values. While it's not a necessity for the univariate distributions to be normal when modeling using OLS, these variables are good candidates for a potential log transformation. It would certainly help with interpretability of these variables, where a percentage change is easier to interpret across counties than an absolute change. This will be investigated further in the model building process.
- We assess that other variables have reasonable distributions **within the EDA partition.**

3.0 | Model Building Process

A General Model for Crime Rate

In 1968, economist Gary Becker proposed that the motivation for an individual to commit a crime is largely due to a rational assessment of the benefit from the illegal activity offset by the potential risk. For example, an individual may determine that the potential monetary gain from theft is worth the chance of being arrested and sentenced to jail. This can be conceptualized as:

(Crime Rate) = (Risk of Criminal Activity) + (Benefit of Criminal Activity)

However, we assert that this internal calculation is further complicated by the opportunity for crime and social or cultural influences. For example, a young male may have a greater appetite for risk than other demographics. Also, there are likely to be more opportunities for crime in a densely populated area compared to a rural area due to the proximity of potential perpetrators and common targets of crime, such as homes, businesses, or other individuals. This can be conceptualized as:

(Crime Rate) = (Risk of Criminal Activity) + (Benefit of Criminal Activity) + (Opportunity for Criminal Activity) + (Social Influence to Commit Criminal Activity)

To further extend our model, it is reasonable to assume that the incidence of crime can also vary by geographic region. This can be conceptualized as:

(Crime Rate) = (Risk of Criminal Activity) + (Benefit of Criminal Activity) + (Opportunity for Criminal Activity) + (Social Influence to Commit Criminal Activity) + (Regional Influence on Criminal Activity)

Relative Weight of Crimes

The overall crime rate does not distinguish between different types of crime; therefore a model using this outcome variable weighs the impact of all crimes equally, which is not ideal for guiding policy decisions.

In an extreme example, a 5% reduction in the overall crime rate could hypothetically come from a 5% reduction in jaywalking and loitering, or it could come from a 5% reduction in murder and rape. We contend that the latter is more damaging on a per-crime basis and that people generally prioritize on their own safety and the safety of their families over their protection of their property. If so, policies that reduce violent crimes would be more valuable than policies to reduce the overall crime rate. However, we find in this study that nonviolent crimes are much more prevalent, and therefore a far greater number of people are likely to be affected by them. From that perspective, policies that reduce nonviolent crimes have the potential to have a positive impact on more people. We believe that making the distinction between violent and nonviolent crime is important for determining policies that can best address each.

To this end, our "A" series of models isolate the face-to-face crime rate as the outcome variable and our "B" series of models isolate the non-face-to-face crime rate as the outcome variable. We use the same general form for the factors of each to allow a meaningful comparison of the relative impact to each type of crime. Though the specifications for these models are the same, we expect that the statistical and practical significance of each variable may differ. For example, the risk of getting caught may serve as a better deterrent for violent crime than nonviolent crime, if the punishment for violent crime is appropriately more severe.

Limitations of the Data

Another important consideration is the distinction between the measurable values we use as proxies for each component of our conceptual models and the true value we are trying to characterize. Consider the example of `prbarr` as a proxy for risk of criminal activity. A criminal's decision-making process is more likely to be based on their perception of the chance of being caught rather than the actual statistical probability. If `prbarr` goes up 10% overnight, will a criminal's perception of the risk of committing a crime immediately reflect this change? We believe not; there will likely be latency to the response of an increase in arrests as this information spreads through a community. Furthermore, the actual change in arrest rate (measurable) and the increase in perceived probability of being arrested (not measurable) may not be equal. A criminal will not pull out the local arrest records to weigh the risks of being caught. They are acting on their own personal proxy variables of the real metrics. We are illustrating the point using the `Risk` component of our model, but `Benefit`, `Opportunity` and `Social Influence` suffer from similar limitations in our ability to measure them as they are heavily dictated by the criminal's perceptions of these concepts.

Concern of Omitted Variable Bias

A correctly specified model is one that has no missing, redundant, or extraneous predictors. Unfortunately, it is not practical to obtain measurements on all factors that can contribute to crime. As a consequence, our analysis is impacted by omitted variable bias which we address in detail in Section 6.0. Redundant predictors exhibit multicollinearity which reduces our precision and limits our understanding of an individual predictor when multiple predictors are related. Extraneous predictors are factors that are not relevant to understanding the variation in crime rate. In selecting proxy variables from our dataset, our criteria include 1) the relevance to the components of our conceptual model and 2) the potential of the chosen variables to suggest actionable policy decisions.

3.1 | Key Variable Analysis

Outcome Variable A: Violent Crime Rate

To represent the "Violent Crime Rate" outcome variable in Models 1A, 2A, and 3A, we weight the "Overall Crime Rate" (`crmrte`) by the percentage of face-to-face crimes. Face-to-face crimes can be a proxy for violent crime since they are likely to be confrontational in nature. However, it is an imperfect proxy because some crimes that are included in this category may not be violent. We have interpreted the `mix` variable as the ratio of face-to-face crimes to other types of crime and other types of crime as a complement to face-to-face crimes. Based on these assumptions, we can manipulate the 'mix' variable to represent the percentage of all crimes that are face-to-face:

$$\frac{ftf\ crimes}{overall\ crimes} = \frac{mix}{1+mix}$$

Then, we multiply this value by the `crmrte` variable to create the `ftfrte` variable which is used as a proxy for the number of violent crimes per person. The intuition behind the creation of `ftfrte` is as shown:

$$ftfrte = \frac{ftf\ crimes}{overall\ crimes} * crmrte = \frac{ftf\ crimes}{overall\ crimes} * \frac{overall\ crimes}{person} = \frac{ftf\ crimes}{person}$$

```
C1$ftfrte <- (C1$mix/(1+C1$mix)) * C1$crmrte
```

The variable `ftfrte` has a positive skew with most of its points concentrated on the low end of its range. Calculating the log transformation of the `ftfrte` variable (`ftfrte_log`) results in a more normal distribution and allows the interpretation of model results to be in terms of an increase or decrease in the face-to-face crime rate percentage. Reference Appendix A.

Outcome Variable B: Nonviolent Crime Rate

To represent the "Nonviolent Crime Rate" outcome variable in Models 1B, 2B, and 3B, we isolate the `nvcrte` variable by subtracting the face-to-face crime rate from the overall crime rate, as follows:

$$(nvc\ crimes) = (overall\ crimes) - (ftf\ crimes)$$

As with `ftfrte`, calculating the log transformation of the `nvcrte` variable (`log_nvcrte`) results in a more normal distribution and allows the interpretation of model results to be in terms of an increase or decrease in the non-face-to-face crime rate percentage. Reference Appendix A.

```
C1$nvcrte <- C1$crmrte - C1$ftfrte
```

Explanatory Variables

Risk of Criminal Activity: Certainty of Punishment

In the original dataset, `prbarr`, `prbconv`, `prbpris` are variables that can serve as a proxy for the perceived likelihood of punishment. The variable `prbarr` is the ratio of arrests to offenses and represents the probability that a criminal will be caught. Once apprehended, the variable `prbconv` is the ratio of convictions per arrest. Of those convicted of a crime, the variable `prbpris` is the ratio of convictions resulting in a prison sentence to total convictions. Considering these three components separately may provide more detailed insight on which aspect of the justice pipeline serves as the greatest deterrent of crime, but for now we are using this as a method of dimension reduction for a set of similar variables. We exclude `avgsen` from this metric, as we hypothesize that criminals may not consider the severity of punishment (long-term consequence) as much as the more immediate factors of whether they're arrested and punished at all. Here we calculate the variable `pctpris` to serve as a proxy for the certainty of punishment, given an offense has occurred:

$$pctpris = probability\ of\ arrest\ *\ probability\ of\ conviction\ *\ probability\ of\ prison\ sentence * 100$$

$$\rightarrow pctpris = prbarr\ *\ prbconv * \ prbpris * 100$$

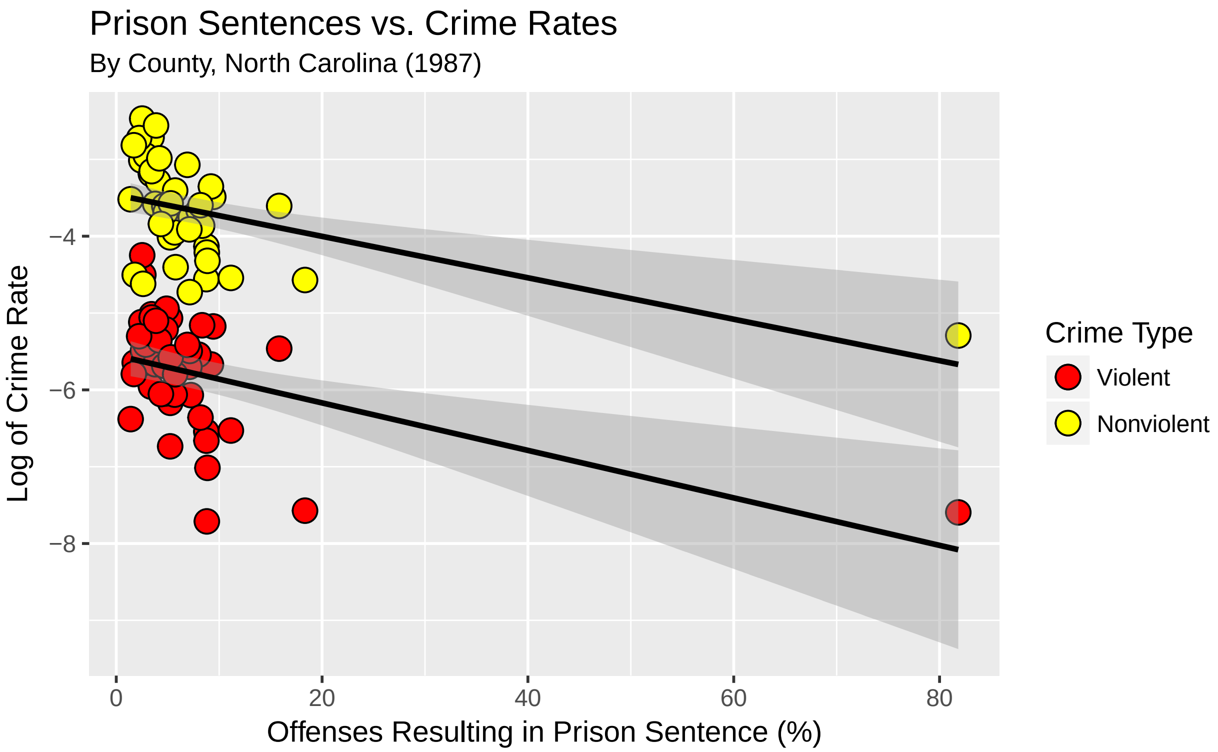
$$\rightarrow pctpris = \frac{arrests}{offenses} * \frac{convictions}{arrests} * \frac{convictions\ resulting\ in\ prison\ sentence}{convictions} * 100$$

$$\rightarrow pctpris = \frac{convictions\ resulting\ in\ prison\ sentence}{offenses} * 100$$

This is an estimate as the data comes from multiple sources (FBI's Uniform Crime Reports and NC Dept of Correction). We don't know if the metrics are time-paired, meaning number of arrests, convictions and prison sentences all logged for the same periods, or if they are measured on a case-by-case basis (each case tracked at each step).

```r
ftf_nv_plotter <- function(xvar, title_var="Title",
                           subtitle_var="By County, North Carolina (1987)",
                           xlab = "xlab", ylab="Log of Crime Rate") {
colors <- c("col1" = "black", "col2" = "black")
shapes <- c("s1" = 21, "s2" = 21)
fills <- c("f1"="red", "f2"="yellow")
ggplot(data = C1, aes(xvar)) +
  geom_point(aes(y=ftfrte_log, color="col1", shape="s1", fill="f1"), size=4) +
  geom_smooth(aes(y=ftfrte_log), method = "lm", se=TRUE,
              color="black", linetype="solid", size=1) +
  geom_point(aes(y=nvcrte_log, color="col2", shape="s2", fill="f2"), size=4) +
  geom_smooth(aes(y=nvcrte_log), method = "lm", se=TRUE,
              color="black", linetype="solid", size=1) +
  labs(title = title_var, subtitle = subtitle_var) +
  labs(x = xlab, y = ylab) +
  scale_color_manual(name = "Crime Type", breaks = c("col1", "col2"),
                     values = colors, labels = c("Violent", "Nonviolent")) +
  scale_shape_manual(name = "Crime Type", breaks = c("s1", "s2"),
                     values = shapes, labels = c("Violent", "Nonviolent")) +
  scale_fill_manual(name = "Crime Type", breaks = c("f1", "f2"),
                     values = fills, labels = c("Violent", "Nonviolent"))
}
```

```r
ftf_nv_plotter(C1$pctpris, t = "Prison Sentences vs. Crime Rates",
               xl = "Offenses Resulting in Prison Sentence (%)")
```



The plot shows a tight cluster of data points below 20% and an extreme outlier of each crime type above 80% which is causing standard errors to be much greater on the right end of the regression lines. Upon examination of the Residuals vs. Leverage plot, this outlier has a Cook's distance much greater than 1. The `prbarr` value for this county, which is used in the calculation of `pctpris`, is 1.09. As described in the EDA, it's not generally reasonable to expect a probability of arrest greater than 1. In addition, several other categories for this same county, such as `polpc` and `avgsen`, are also influential outliers relative to crime rate. Since the crime rate for this county is very low at 0.0055 crimes per capita, the extreme variability in the independent variables for this county is likely due to having less of an averaging effect from the very few crimes that occurred in this county. To reduce the noise introduced by this phenomenon, we impose a threshold and include data points above 0.01 crimes per capita in our analysis. Applying this cutoff to our dataset has the consequence of limiting the applicability of our models to only counties that have crime rates higher than this threshold.

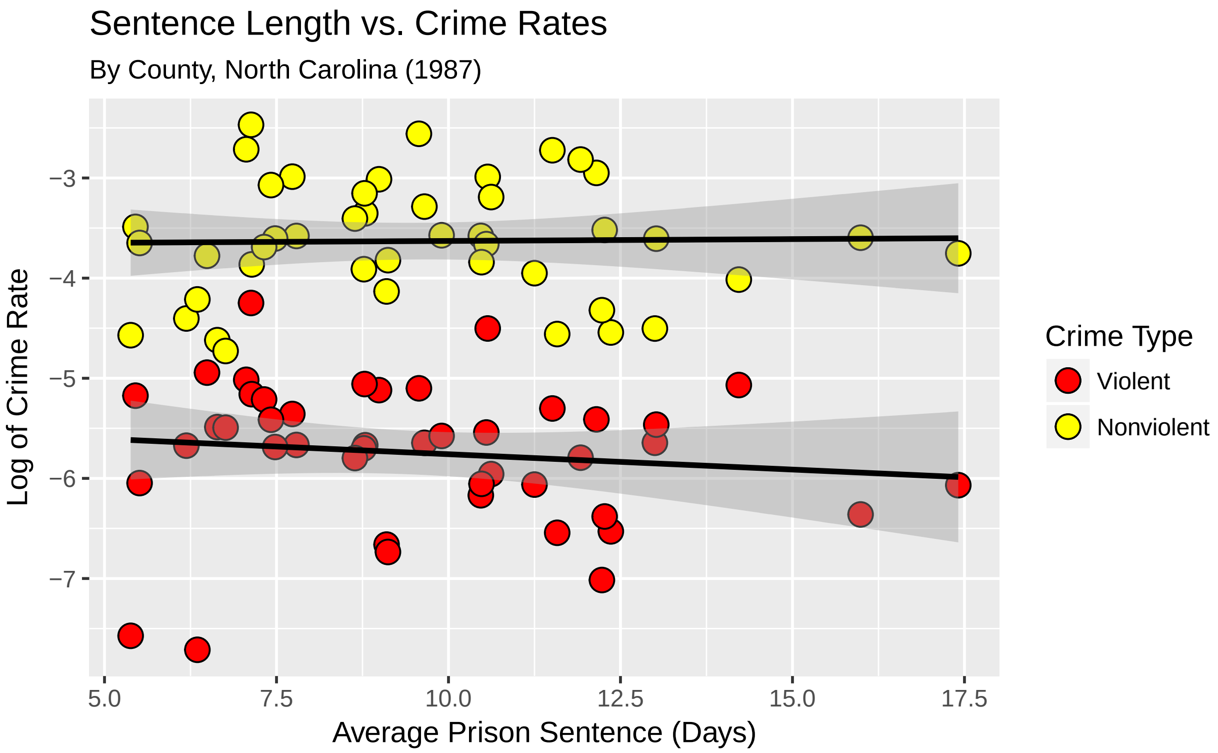
```
C1 <- C1[C1$crmrte > 0.01, ]
```

The new plot still shows a strong negative correlation for both outcome variables but a more even distribution of values (Appendix C); as expected, lower crime rate is associated with a higher likelihood of receiving a prison sentence.

Risk of Criminal Activity: Severity of Punishment

In addition to the certainty of punishment, the magnitude of the penalty will impact the motivation for committing a crime. In the dataset, the variable `avgsen`, which is the average prison sentence in days, serves as a proxy for the harshness of punishment.

```
ftf_nv_plotter(C1$avgsen, t = "Sentence Length vs. Crime Rates",
               xl = "Average Prison Sentence (Days)")
```

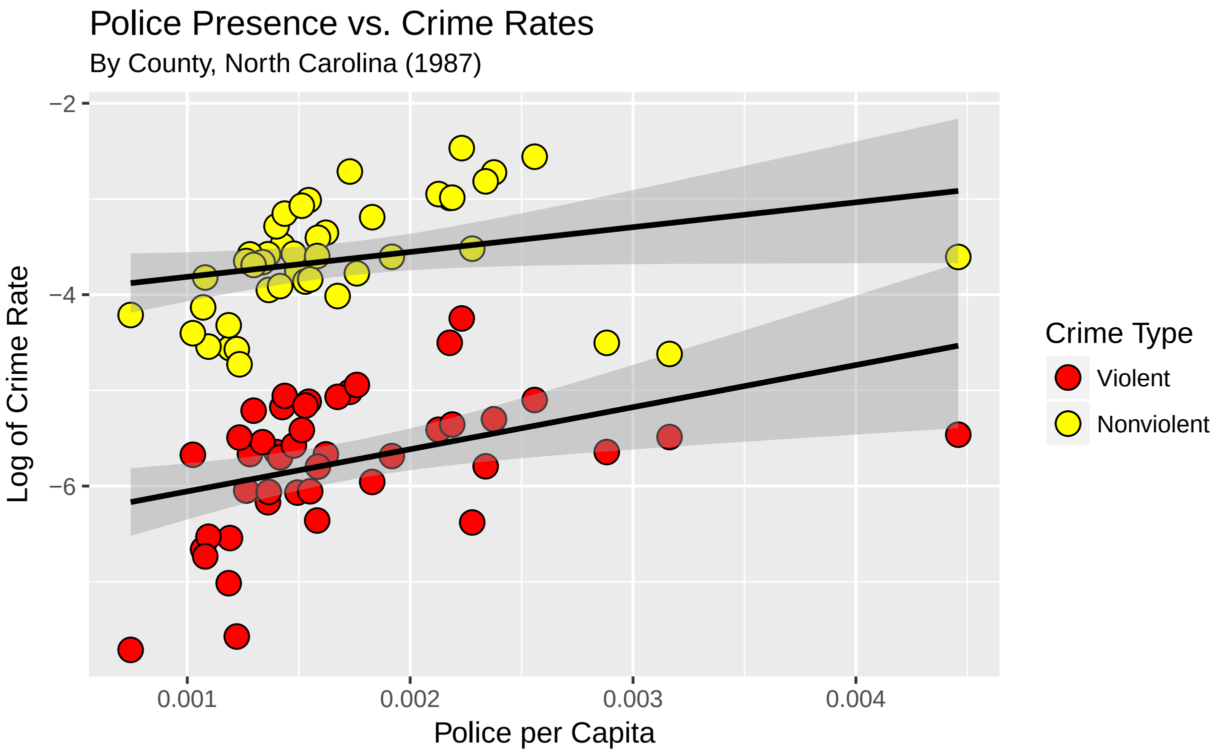


The plot shows almost no correlation between `avgsen` and `nvcrte_log`; we infer that at least without including covariates, the severity of punishment is not associated with any change in nonviolent crime rate. However, there is a slight but perceptible negative correlation between `avgsen` and `ftfrte_log`. The direction and degree are reasonable based on our expectation, in that severity of punishment is associated with a slight decrease in crime rate. Standard errors are greater on the right end of the regression lines where there are fewer data points.

Risk of Criminal Activity: Perceived Probability of Being Caught

The variable `polpc` is the number of police per capita which could be a potential proxy for the perceived opportunity of committing an illegal activity. When there is a greater police presence, there should be a higher perceived chance of getting caught, which should contribute to `Risk`. However, this effect is convoluted as communities will be motivated to increase their police force if they experience greater amounts of crime. In this case, we may also expect that a larger police force may be associated with a higher crime rate. Due to its ambiguous relationship with crime rate, it is difficult to infer causal insight from this variable, but we have included this variable in a subset of our model specifications to control for variation.

```r
ftf_nv_plotter(C1$polpc, t = "Police Presence vs. Crime Rates",
               xl = "Police per Capita")
```



The plot shows strong positive correlations for both outcome variables, especially with `ftfrte_log`, indicating that a greater number of police per capita is associated with a higher crime rate. This gives credence to the idea that police presence increases as a result of crime rate. The plot shows a clear cluster of data points on the left at or below 0.0025 and two obvious extreme outliers on the right at about 0.0045. Standard errors are much greater on the high end of the regression lines, where there is only one data point.

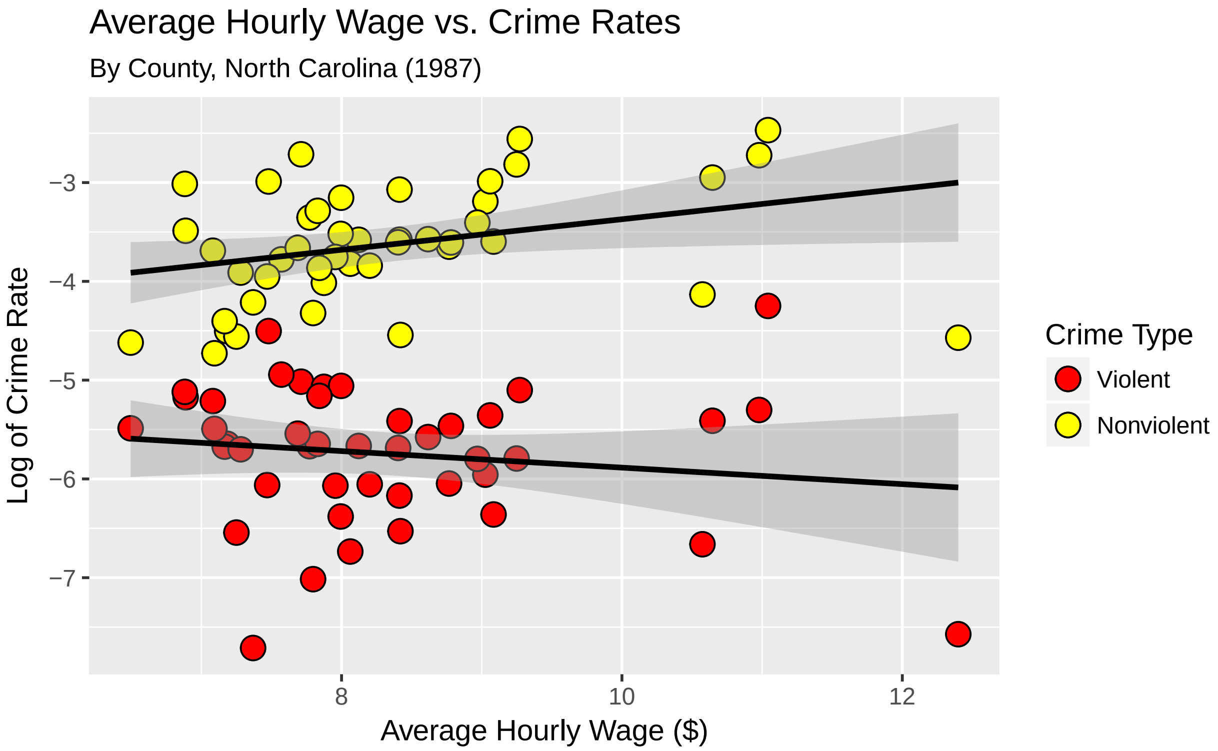
Benefit of Criminal Activity: Wages

As previously mentioned, there may be a greater perceived benefit of crime if there is a weak labor market resulting in low income and lack of opportunity for an individual to find work with a fair wage. There are nine variables representing the average weekly wage in various employment sectors: construction (`wcon`), transportation & utilities (`wtuc`), retail (`wtrd`), finance & real estate (`wfir`), service industry (`wser`), manufacturing (`wmfg`), federal government employees (`wfed`), state government employees (`wsta`), and local government employees (`wloc`). The various wage variables are generally correlated with one another, since all wages may be higher in one area than another due to cost of living, relationships between industries, competition for workers, or other factors.

To reduce the effect of multicollinearity in our models, we average the nine weekly wage categories to create a single composite wage variable. To more directly relate our interpretation of model results to potential policy changes regarding minimum wage, we convert from weekly wage to hourly wage assuming a 40 hour work week. The resulting variable is called `hrwg`. An important consideration when creating this composite wage variable is that each employment sector is weighted equally, since the dataset lacks information regarding the number of employees in each industry. We must keep in mind that this variable represents the average of the industry average wages in the county, not the overall average wage, which would be a better proxy for the economic strength of a county.

The service industry wage category (`wser`) has an extreme outlier. Therefore, the influence of the outlier on the composite wage variable (`hrwg`) is evaluated by examining the effect of removing the outlier from `wser` before calculating `hrwg`. There is minimal difference in our composite wage variable, `hrwg`, before and after removing the `wser` outlier, so the outlier is kept in the dataset.

```r
ftf_nv_plotter(C1$hrwg, t = "Average Hourly Wage vs. Crime Rates",
               xl = "Average Hourly Wage ($)")
```



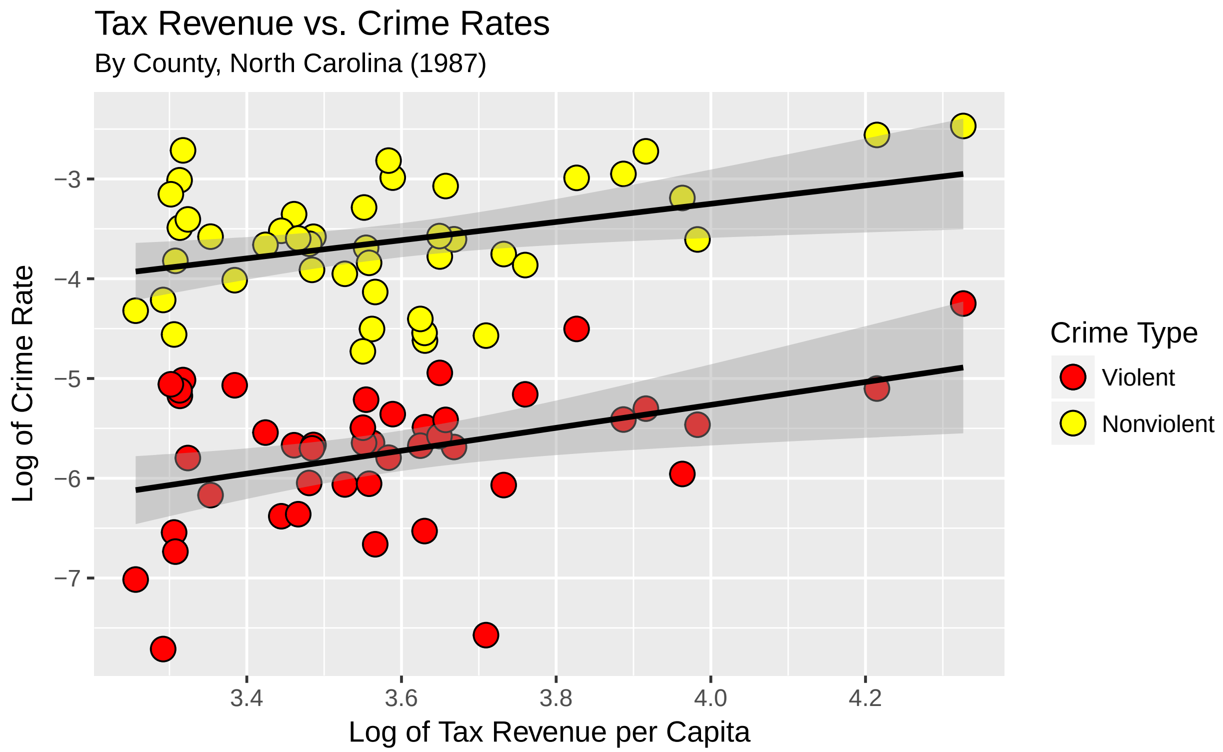
The plot of `hrwg` and `ftfrte_log` shows a negative correlation, while `nvcrte_log` shows a positive correlation, which somewhat surprisingly indicates that higher average wages are associated with higher nonviolent crime rates. This might be due to higher wages being associated with more densely populated areas where crime rate is also higher. It might also be that higher wages allow for more acquisition of property, which in turn provides more targets for nonviolent crimes such as theft. The plot shows highly diffuse data points concentrated on the left side of the plot, between about $6 and $9/hour. The extreme outlier described in the EDA can be clearly seen at the far right side at over $12/hour. There are four additional, less extreme outliers between $10 and $12/hour. Standard errors are greater on the high end of the regression lines.

Benefit of Criminal Activity: Tax Revenue

Another proxy for economic strength is the `taxpc` variable which is the tax revenue per capita, since higher incomes and more valuable property relate to prosperity. A county's tax revenue can also have implications on the social influence to commit crime, since lower tax revenue may result in fewer social services and a lack of institutional support for vulnerable individuals.

The distribution of `taxpc` shows a positive skew with most of the data concentrated on the lower end (Appendix B). There is still a moderate positive skew after performing a log transformation of the `taxpc` variable (`taxpc_log`). However, transforming the variable by taking a log of `taxpc` allows a more meaningful interpretation of model results by framing the outcome as a percentage increase or decrease in the amount of tax revenue instead of a dollar magnitude.

```r
ftf_nv_plotter(C1$taxpc_log, t = "Tax Revenue vs. Crime Rates",
               xl = "Log of Tax Revenue per Capita")
```



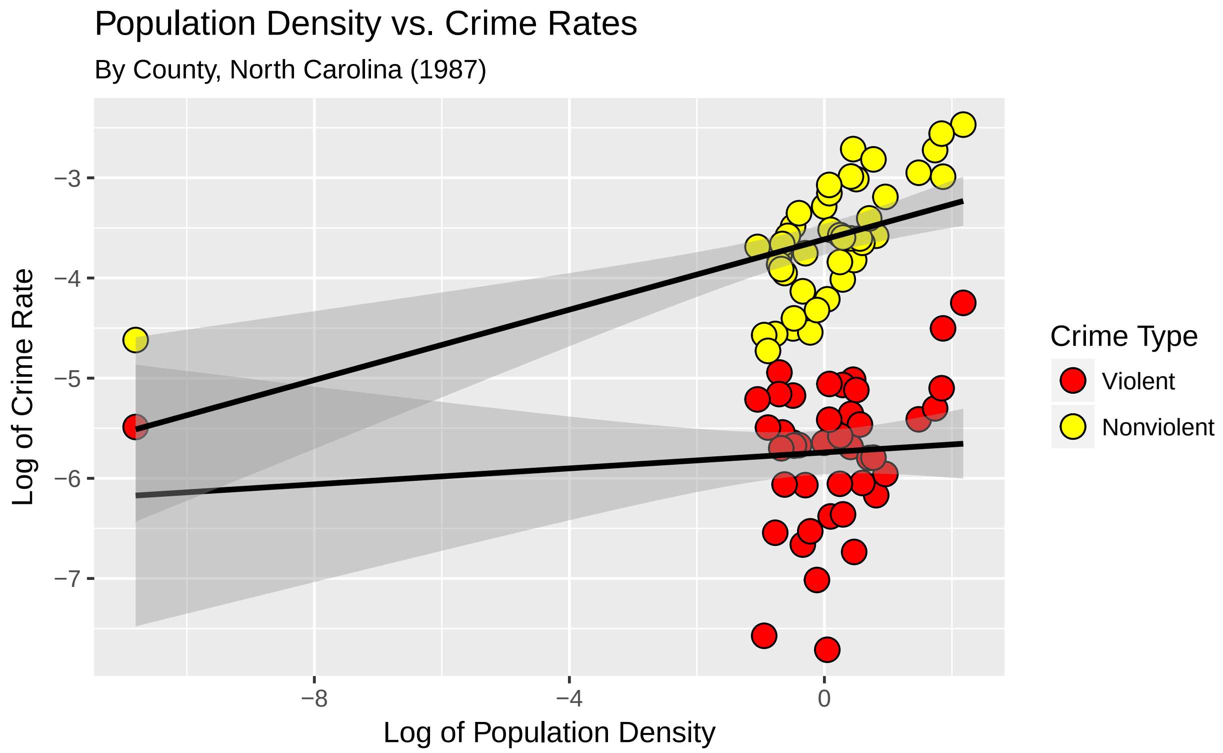
Counter to our intuition, but consistent with the concepts in the average wage discussion above, higher crime rate is correlated with increasing tax revenue. This may be because higher tax revenue is associated with higher population density (Appendix C) which in turn, is also correlated with crime rate.

Opportunity for Criminal Activity: Population Density

The primary proxy for opportunity to commit crime is `density`, which gives population density. We reason that opportunities for crime are more likely when interactions with others and their property are more frequent. If you live in the country and your nearest neighbor is five miles away, you'll have a much different profile of opportunities for crime than someone who lives on the 10th floor of a high-rise, just steps from tens or hundreds of neighbors.

The distribution of `density` has a heavy positive skew with most of the data concentrated on the lower end (Appendix B). After taking the log of the variable, the data is distributed more normally, though it now has a negative skew. However, the transformation also allows us to frame the model results in terms of a percentage change of density rather than a change in absolute magnitude, which is helpful for interpretability.

```
ftf_nv_plotter(C1$density_log, t = "Population Density vs. Crime Rates",
                xl = "Log of Population Density")
```



As noted in the EDA, there is an extremely low population density outlier which is far below the 1980 Census minimum value. Upon examining the Residuals vs. Leverage plot, this outlier has a Cook's distance much greater than 1. Since this value is not consistent with other sources of information and exerts a disproportionate amount of influence on the model, the anomalous `density_log` value for this county is replaced with the average of `density_log`.

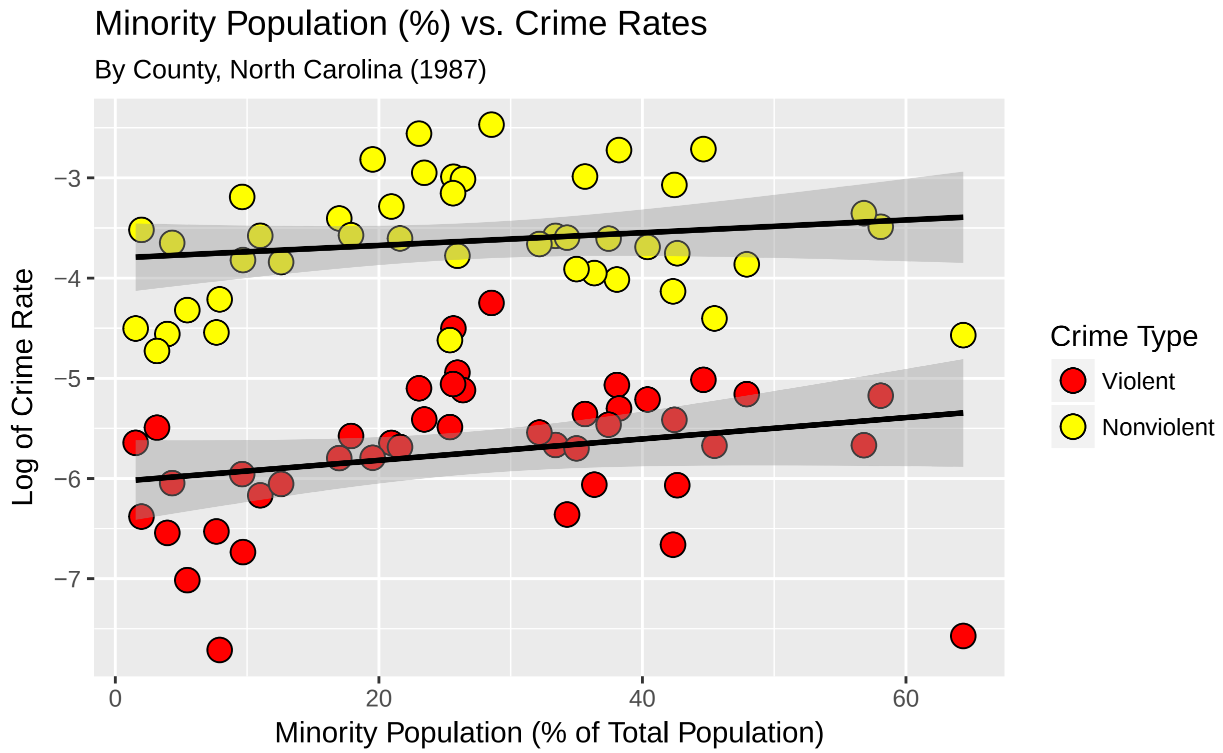
```r
C1["79", "density_log"] <- mean(C1$density_log)
```

The new plot of population density versus crime rates clearly shows positive correlations for both outcome variables (Appendix C); an increase in population density is associated with increased crime rates. While there are not many policy options that can address the issue of population density directly, it is worth noting that urban areas may have different needs than more rural areas.

Social Influence to Commit Criminal Activity

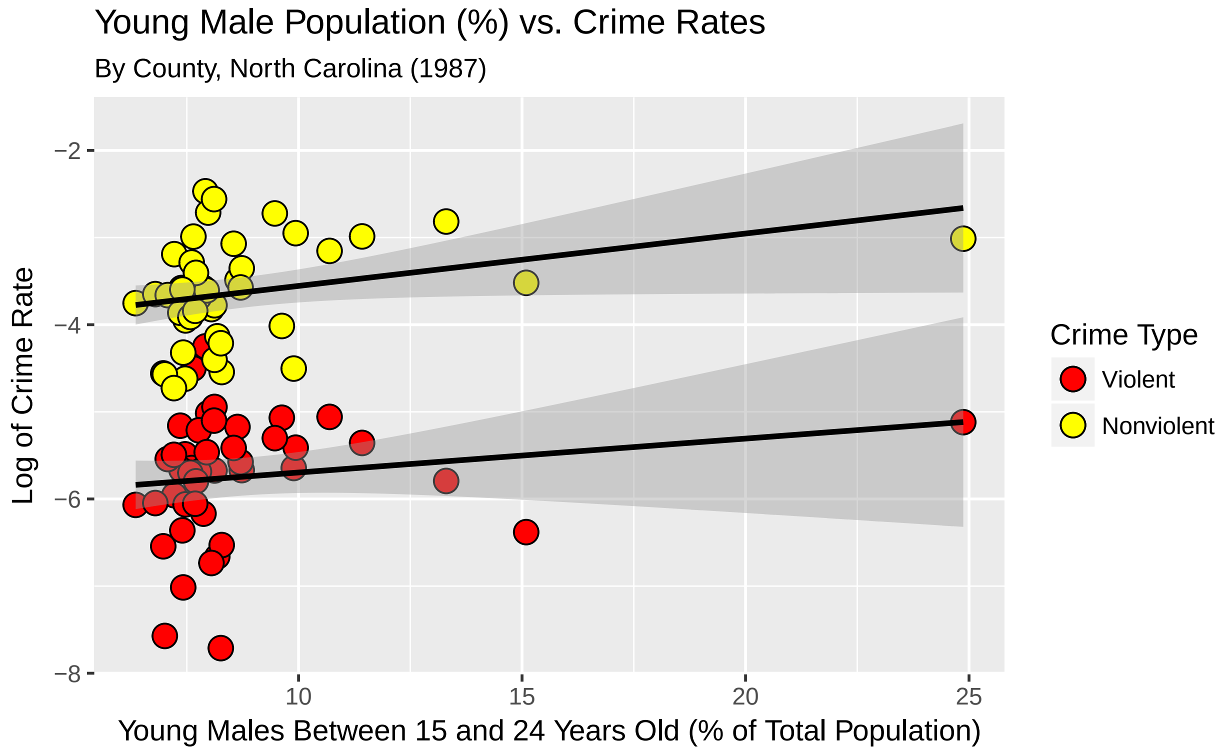
The two variables that represent the social influence to commit criminal activity are `pctmin80`, which is the percent of the population that are racial minorities per the 1980 Census, and `pctymle`, which is the proportion of the county's population that is male between the ages of 15 and 24. The proportion of minorities in a population may impact the apparent incidence of crime, because some minorities are over-represented in the criminal justice system. Also, there may be more social pressure on young males to participate in illegal activity, as well as more acceptance of such activity among their peers. Notably, minorities and young men are also more likely to be the victims of many crimes than other demographic groups.

```r
ftf_nv_plotter(C1$pctmin80, t = "Minority Population (%) vs. Crime Rates",
                xl = "Minority Population (% of Total Population)")
```



There appears to be a moderate correlation between the proportion of minorities in a county and the county's crime rate, where the crime rate increases as the percentage of minorities increase. This correlation is stronger with `ftfrte_log` than `nvcrte_log`, but there is also more dispersion in these data points. An extreme outlier can be seen in the bottom right corner; this point represents a very high minority population and very low violent crime rate (high leverage/high influence).

```
ftf_nv_plotter(C1$pctymle, t = "Young Male Population (%) vs. Crime Rates",
               xl = "Young Males Between 15 and 24 Years Old (% of Total Population)")
```



Similar to the proportion of minorities in a county, there appears to be a moderate correlation between the proportion of young males and crime rate, where crime rate increases as the number of young males increases. An extreme outlier can be seen at the right end of the regression line, representing a very high young male population but fairly average crime rate (high leverage/low influence).

Regional Influence to Commit Criminal Activity

The indicator variables that categorize the county as western (`west`) or central (`central`) region may also serve as a proxy for varying social norms across geography.

4.0 | Regression Models

In Section 3.0, we established our 'complete' conceptual model for crime rate as:

(Crime Rate) = (Risk of Criminal Activity) + (Benefit of Criminal Activity) + (Opportunity for Criminal Activity) + (Social Influence to Commit Criminal Activity) + (Regional Influence on Criminal Activity)

This model specification risks being overly elaborate, and there is value in having a simpler one. Following a review of our EDA dataset, we selected proxy variables that we think best represent the 'Risk', 'Benefit', 'Opportunity', 'Social Influence', and 'Regional Influence' categories from our available data described in Section 3 and summarized in the table below.

Concept	Proxy Variable
Risk	$pctpris$
Risk	$avgsen$
Benefit	$hrwg$
Opportunity	$density\_log$
Social Influence	$pctmin80$
Social Influence	$pctymle$
Opportunity	$urban$
Regional Influence	$central$
Regional Influence	$west$
Control for Variation	$taxpc\_log$
Control for Variation	$polpc$

From our EDA, the variables `polpc` and `taxpc_log` are more tightly coupled to trends in population density than with crime rates (Appendix C). However, including these two variables may potentially allow us to control for additional variation when evaluating our other variables of interest. We will evaluate various configurations of these variables to establish a model that is both parsimonious and adequately captures the important aspects of our conceptual model.

General Model Assumptions (Applicable to All Models)

When employing OLS regression, we must consider its foundational assumptions that allow us to make the conclusions about the unbiasedness of its predictors, efficiency of the OLS method and its predictive power for determining significance. These assumptions are not strict requirements for OLS to be useful, but we should consider them when performing this type of analysis. We assess the first three assumptions in general for all models that follow.

Assumption MLR.1 (Linear in Parameters): This states that our parameters should form a linear equation based on each of our included variables. In other words, the beta coefficients should be scalars of each variable included in the regression and a constant term or intercept is also permitted. This assumption is honored in each of our model specifications.

Assumption MLR.2 (Random Sampling): While we used a semi-random method for sampling our subset of data from the larger crime dataset, we are still at risk of violations of the random sampling assumption. Because there are 100 counties in North Carolina and we are working with a total of 90 observations, we know that we have a sample; however, it is likely not a random sample. We would need more details to determine the level of potential clustering or non-randomness present in the dataset. For this analysis, we proceed as though this assumption is not violated, but should bear in mind that our results could be biased.

Assumption MLR.3 (No Perfect Collinearity): We can't include one variable that is a linear combination of another (or multiple others) included in the same regression. For this reason, we would want to avoid using both `hrwg` and `wkwg` in the same model.

We will assess the three other model assumptions on a model chosen for its performance and parsimony after exploring various model specifications.

4.1 | Regression Models: Model 1

We start with the simplest conceptual model (Becker's economic model) and apply it to both of our proxy variables for violent and nonviolent crime.

Conceptual Model: (Crime Rate) = (Risk) + (Benefit)

To capture the risk associated with criminal activity, we use two variables which capture some of the elements a criminal might consider before committing a crime. We want to distinguish between the probability of punishment (if caught, will a punishment occur) from the severity of punishment (if caught, how bad will it be). This distinction is important, as criminals may not think things through to their ultimate conclusion. They may only evaluate a subjective measure of risk like whether they think they'll be caught/punished and not the true probability of punishment or potential length of prison term. Our more specific hypothesis is as follows:

(Crime Rate) = (Prob of Punishment) + (Severity of Punishment) + (Benefit)

After substituting in proxy variables, the model is as follows:

$$ftfrte\_log\ or\ nvcrte\_log = \beta_0 + \beta_1 * pctpris + \beta_2 * avgsen + \beta_3 * hrwg$$

```r
mod_1A_1 = lm(C1$ftfrte_log ~ C1$pctpris + C1$avgsen + C1$hrwg)
mod_1B_1 = lm(C1$nvcrte_log ~ C1$pctpris + C1$avgsen + C1$hrwg)

se.mod_1A_1 = sqrt(diag(vcovHC(mod_1A_1)))
se.mod_1B_1 = sqrt(diag(vcovHC(mod_1B_1)))

stargazer(mod_1A_1, mod_1B_1,
          type = "latex", title = "Linear Models Predicting Crime Rates",
          dep.var.caption = "Model 1", omit.stat = c("f", "ser"),
          se = list(se.mod_1A_1, se.mod_1B_1),
          star.cutoffs = c(0.05, 0.01, 0.001),
          add.lines=list(c("AIC", round(AIC(mod_1A_1),1), round(AIC(mod_1B_1),1)))
)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Aug 07, 2018 - 14:39:03

Table 1: Linear Models Predicting Crime Rates

	Model 1	
	ftfrte_log	nvcrte_log
	(1)	(2)
pctpris	-0.100	-0.095^{*}
	(0.052)	(0.039)
avgsen	-0.035	-0.006
	(0.042)	(0.032)
hrwg	-0.027	0.205^{*}
	(0.112)	(0.090)
Constant	-4.593^{***}	-4.719^{***}
	(0.919)	(0.831)
AIC	90.3	65.2
Observations	44	44
R^2	0.268	0.406
Adjusted R^2	0.213	0.362
Note:	*p<0.05; **p<0.01; ***p<0.001	

For both the violent and nonviolent crime models, the results show that avgsen is not statistically significant, and its effect size is particularly small for nonviolent crime. We anticipated that average sentence length might be a weaker determinant of crime rate than the more immediate "certainty of punishment" when proposing our independent variables of interest, and this seems to agree. It's important to recognize one assumption we are making here, which is regarding the distribution of crime types. We're assuming that changes in avgsen and pctpris are due to other characteristics of a county besides a difference in the distribution of crimes committed. This is likely not a fair assumption, which we will discuss more in Section 6.0 regarding omitted variable bias.

In the violent crime model, no predictors are statistically significant. Also, the adjusted R^2 values for both the violent and nonviolent crime models are relatively low at 0.213 and 0.362 respectively which indicate that only a small portion of the variation in both types of crime rates are explained by this simplistic model. For these reasons, it is likely that these models are underspecified and require additional predictors to produce less biased estimates.

4.2 | Regression Models: Model 2

For our second model, we want to add in elements that capture the 'Social Influence' and 'Opportunity' categories of our conceptual model, particularly since we concluded that our Model 1 series was underspecified. Since we suspect from Model 1 results that the avgsen variable does not make much of a contribution to explaining the variation in our two types of crime rates, we evaluate the impact of including or removing the avgsen variable in our Model 2 specifications.

Conceptual Model: (Crime Rate) = (Risk) + (Benefit) + (Opportunity) + (Social Infl)

After substituting in proxy variables, the model is as follows:

$$ftfrte\_log\ or\ nvcrte\_log = \beta_0 + \beta_1 * pctpris + \beta_2 * avgsen$$

$$+\beta_3 * hrwg + \beta_4 * density\_log + \beta_5 * pctymle + \beta_6 * pctmin80$$

```
mod_2A_1 = lm(C1$ftfrte_log ~ C1$pctpris + C1$avgsen + C1$hrwg + C1$density_log + C1$pctymle
              + C1$pctmin80)
mod_2A_2 = lm(C1$ftfrte_log ~ C1$pctpris + C1$hrwg + C1$density_log + C1$pctmin80)
mod_2B_1 = lm(C1$nvcrte_log ~ C1$pctpris + C1$avgsen + C1$hrwg + C1$density_log + C1$pctymle
              + C1$pctmin80)
mod_2B_2 = lm(C1$nvcrte_log ~ C1$pctpris + C1$hrwg + C1$density_log + C1$pctmin80)

se.mod_2A_1 = sqrt(diag(vcovHC(mod_2A_1)))
se.mod_2B_1 = sqrt(diag(vcovHC(mod_2B_1)))
se.mod_2A_2 = sqrt(diag(vcovHC(mod_2A_2)))
se.mod_2B_2 = sqrt(diag(vcovHC(mod_2B_2)))

stargazer(mod_2A_1, mod_2B_1, mod_2A_2, mod_2B_2,
          type = "latex", title = "Linear Models Predicting Crime Rates",
          dep.var.caption = "Model 2", omit.stat = c("f", "ser"),
          se = list(se.mod_2A_1, se.mod_2B_1, se.mod_2A_2, se.mod_2B_2),
          star.cutoffs = c(0.05, 0.01, 0.001),
          add.lines=list(c("AIC", round(AIC(mod_2A_1),1), round(AIC(mod_2B_1),1),
                           round(AIC(mod_2A_2),1), round(AIC(mod_2B_2),1)))
)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Aug 07, 2018 - 14:39:03

Removing the `avgsen` and `pctymle` variables, which are not statistically significant in either of our Model 2 specifications and have standard errors greater than the value of their coefficients, produces smaller standard errors leading to more precise coefficient estimates for the remaining variables. The AIC metrics improve when removing these variables, indicating that the model without `avgsen` and `pctymle` is the more parsimonious choice for both outcome variables. Also, the improvement in precision allows the coefficient for the `hrwg` variable to become statistically significant for the violent crime model, as well as increasing the significance of `pctpris` in the nonviolent crime model. Therefore, we decide not to include the `avgsen` or `pctymle` variables in our models going forward.

There is a considerable improvement in this set of models compared to our Model 1 specifications. More of the variation in both types of crime rates are explained by these models as indicated by the improvement in adjusted R^2 from `0.21` to about `0.52` in the violent crime rate model and from `0.36` to about `0.69` in the nonviolent crime rate model. In addition, the Akaike Information Criterion (AIC) improves from `90.3` to `69.7` in the violent crime rate model and from `65.2` to `34.5` in the nonviolent crime rate model.

4.3 | Regression Models: Model 3

For our third model, we continue to introduce complexity by adding proxies representing the regional influence on crime as well as most other available variables in our dataset to assess if there are additional sources of variation for which we may need to control.

Conceptual Model: (Crime Rate) = (Risk) + (Benefit) + (Opportunity) + (Social Infl) + (Regional Infl) + (Other Sources of Variation)

After substituting in proxy variables, the model is as follows:

$$ftfrte\_log \ or \ nvcrte\_log = \beta_0 + \beta_1 * pctpris + \beta_2 * avgsen$$

$$+\beta_3 * hrwg + \beta_4 * density\_log + \beta_5 * pctymle + \beta_6 * pctmin80$$

Table 2: Linear Models Predicting Crime Rates

	Model 2			
	ftfrte_log	nvcrte_log	ftfrte_log	nvcrte_log
	(1)	(2)	(3)	(4)
pctpris	-0.092	-0.052^{*}	-0.091	-0.060^{**}
	(0.064)	(0.023)	(0.054)	(0.020)
avgsen	-0.020	-0.007		
	(0.040)	(0.021)		
hrwg	-0.214	0.009	-0.211^{*}	0.005
	(0.119)	(0.090)	(0.105)	(0.075)
density_log	0.399	0.482^{***}	0.382	0.484^{***}
	(0.219)	(0.107)	(0.203)	(0.104)
pctymle	-0.012	0.026		
	(0.019)	(0.027)		
pctmin80	0.024^{***}	0.015^{***}	0.025^{***}	0.016^{***}
	(0.007)	(0.003)	(0.006)	(0.003)
Constant	-3.839^{***}	-4.044^{***}	-4.180^{***}	-3.814^{***}
	(0.990)	(0.859)	(0.761)	(0.614)
AIC	72.8	36	69.7	34.5
Observations	44	44	44	44
R^2	0.571	0.733	0.562	0.718
Adjusted R^2	0.501	0.690	0.517	0.689
Note:		$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$		

$$+\beta_7 * urban + \beta 8 * central + \beta_9 * west + \beta_{10} * polpc + \beta_{11} * taxpc\_log$$

```r
mod_3A_1 = lm(C1$ftfrte_log ~ C1$pctpris + C1$avgsen + C1$hrwg + C1$density_log
              + C1$pctymle + C1$pctmin80 + C1$urban  + C1$central + C1$west + C1$taxpc_log
              + C1$polpc)
mod_3B_1 = lm(C1$nvcrte_log ~ C1$pctpris + C1$avgsen + C1$hrwg + C1$density_log
              + C1$pctymle + C1$pctmin80 + C1$urban + C1$central + C1$west + C1$taxpc_log
              + C1$polpc)

se.mod_3A_1 = sqrt(diag(vcovHC(mod_3A_1)))
se.mod_3B_1 = sqrt(diag(vcovHC(mod_3B_1)))

stargazer(mod_1A_1, mod_1B_1, mod_2A_2, mod_2B_2, mod_3A_1, mod_3B_1,
          type = "latex", title = "Linear Models Predicting Crime Rates",
          dep.var.caption =
            "---------- Model 1 ------------------ Model 2 ------------------ Model 3 ----------",
          omit.stat = c("f", "ser"), column.sep.width = "-2pt",
          se = list(se.mod_1A_1, se.mod_1B_1, se.mod_2A_2, se.mod_2B_2, se.mod_3A_1,
                    se.mod_3B_1),
          star.cutoffs = c(0.05, 0.01, 0.001),
          add.lines=list(c("AIC", round(AIC(mod_1A_1),1), round(AIC(mod_1B_1),1),
                           round(AIC(mod_2A_2),1), round(AIC(mod_2B_2),1),
                           round(AIC(mod_3A_1),1), round(AIC(mod_3B_1),1)))
)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Aug 07, 2018 - 14:39:03

The adjusted R^2 for both the violent and nonviolent crime rate models improve from Model 2 specifications and the AIC for the violent crime rate model is slightly lower. However, the improvement in these metrics is relatively modest considering the high number of additional predictors introduced in this specification. The standard errors also increase for most of the variable coefficients in Model 3.

To confirm if these additional variables have predictive power, we perform a joint significance test of the null that the coefficients for `avgsen`, `pctymle`, `urban`, `central`, `west`, `taxpc_log`, and `polpc` are zero. The p-value for the F-test in the violent crime model is 0.22 and is 0.39 for the nonviolent crime model. In both cases, the p-values are far greater than the significance threshold of 0.05, so we fail to reject the null that these variables have coefficients greater than 0. This further validates our suspicion that Model 3 is overspecified, and we select Model 2 for confirmation of reproducibility and interpretation of results on our reserved CDA dataset.

```r
vcovHC_3A_1 <- vcovHC(mod_3A_1)
vcovHC_3B_1 <- vcovHC(mod_3B_1)

test_3A_1 <- linearHypothesis(mod_3A_1, c("C1$avgsen = 0", "C1$pctymle = 0", "C1$urban = 0",
                                          "C1$west = 0", "C1$central = 0", "C1$taxpc_log = 0",
                                          "C1$polpc = 0"), vcov=vcovHC_3A_1)
test_3B_1 <- linearHypothesis(mod_3B_1, c("C1$avgsen = 0", "C1$pctymle = 0", "C1$urban = 0",
                                          "C1$west = 0", "C1$central = 0", "C1$taxpc_log = 0",
                                          "C1$polpc = 0"), vcov=vcovHC_3B_1)
test_3A_1$`Pr(>F)`[2]
```

```
## [1] 0.2190028
```

Table 3: Linear Models Predicting Crime Rates

	Model 1		Model 2		Model 3	
	ftfrte_log	nvcrte_log	ftfrte_log	nvcrte_log	ftfrte_log	nvcrte_log
	(1)	(2)	(3)	(4)	(5)	(6)
pctpris	−0.100 (0.052)	−0.095* (0.039)	−0.091 (0.054)	−0.060** (0.020)	−0.092* (0.044)	−0.055 (0.029)
avgsen	−0.035 (0.042)	−0.006 (0.032)			−0.039 (0.040)	−0.024 (0.023)
hrwg	−0.027 (0.112)	0.205* (0.090)	−0.211* (0.105)	0.005 (0.075)	−0.222* (0.111)	0.080 (0.087)
density_log			0.382 (0.203)	0.484*** (0.104)	0.097 (0.208)	0.411* (0.160)
pctymle					−0.0003 (0.034)	0.014 (0.032)
pctmin80			0.025*** (0.006)	0.016*** (0.003)	0.014 (0.011)	0.005 (0.009)
urban					0.168 (0.516)	0.135 (0.348)
central					−0.111 (0.206)	−0.310 (0.173)
west					−0.348 (0.442)	−0.552 (0.334)
taxpc_log					0.812 (0.773)	−0.297 (0.471)
polpc					249.781 (218.533)	63.720 (200.369)
Constant	−4.593*** (0.919)	−4.719*** (0.831)	−4.180*** (0.761)	−3.814*** (0.614)	−6.629* (2.806)	−2.901* (1.471)
AIC	90.3	65.2	69.7	34.5	67.3	36.6
Observations	44	44	44	44	44	44
R^2	0.268	0.406	0.562	0.718	0.699	0.785
Adjusted R^2	0.213	0.362	0.517	0.689	0.595	0.711

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

```
test_3B_1$`Pr(>F)`[2]
```

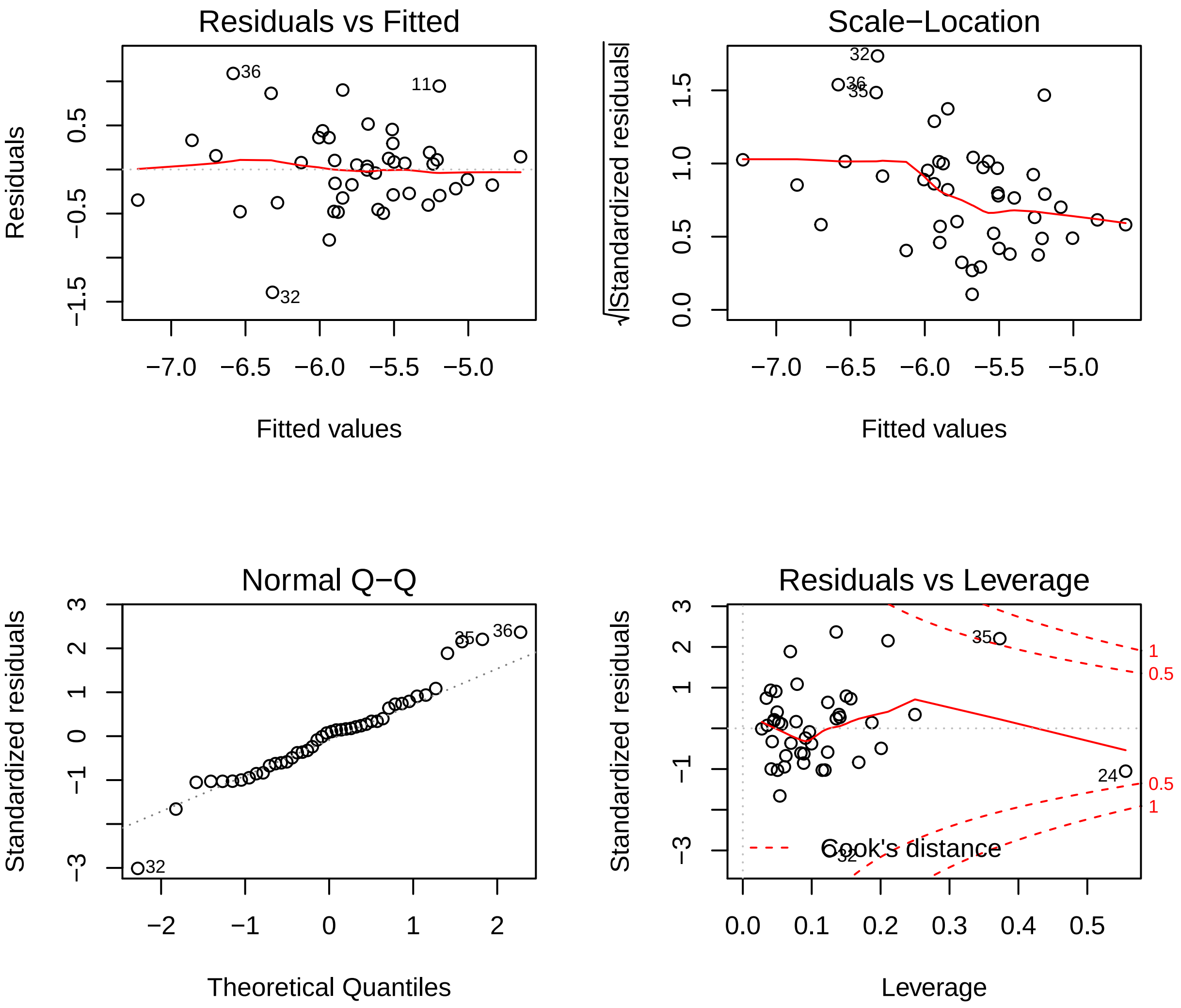
```
## [1] 0.3889972
```

4.4 | Regression Models: Evaluation of CLM Assumptions

In comparing our Models 1, 2, and 3 specifications for violent and nonviolent crime, the Model 2 specifications appear to achieve an effective balance between parsimony and including enough predictors to explain the variation in our dependent variables to reduce bias. We proceed in examining model assumptions in greater detail for Model 2.

Discussion of CLM Assumptions 4-6: Violent Crime Rate Model

```
par(mfrow=c(2,2))
plot(mod_2A_2, which=1)
plot(mod_2A_2, which=3)
plot(mod_2A_2, which=2)
plot(mod_2A_2, which=5)
```



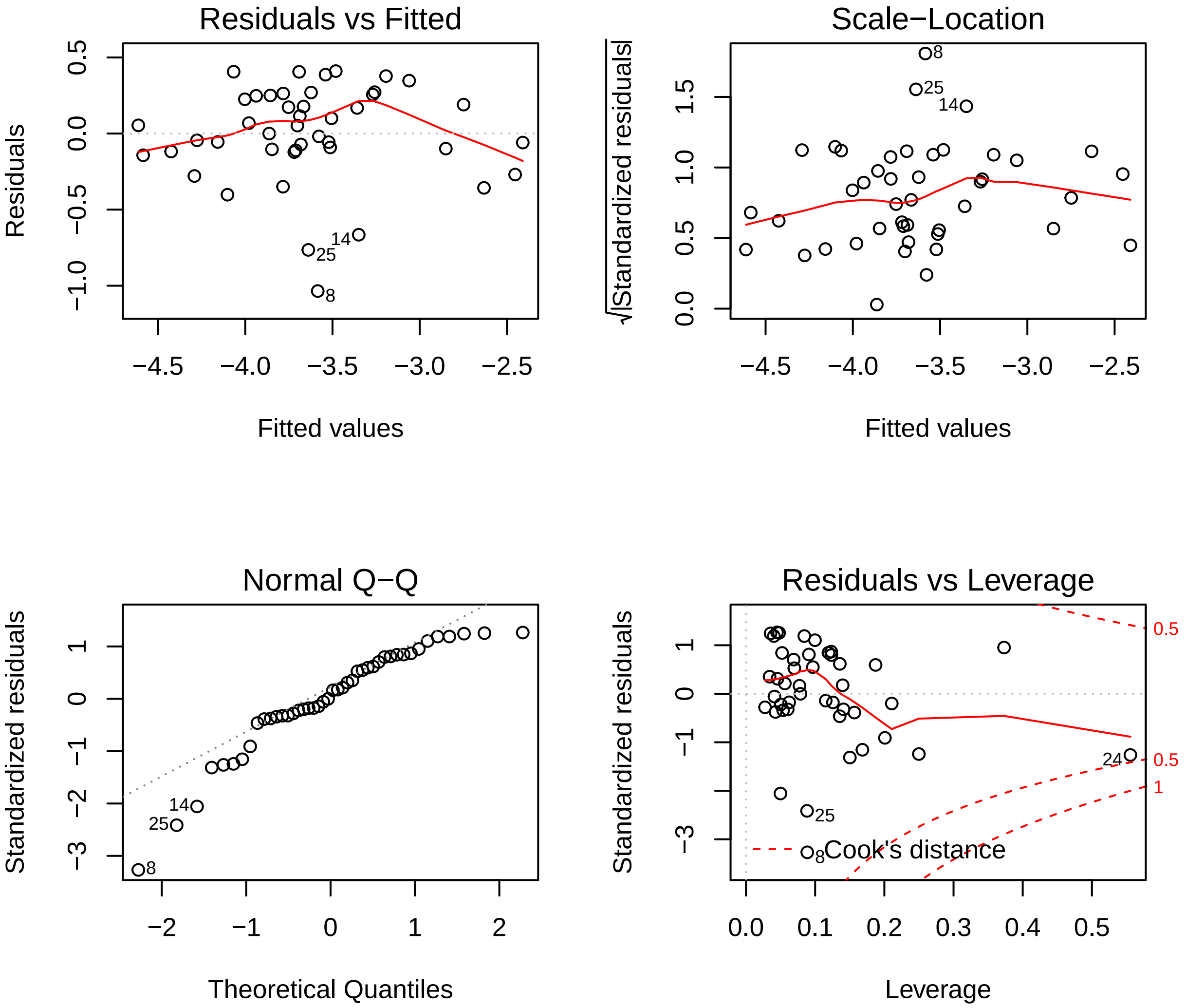
Assumption MLR.4 (Zero Conditional Mean): In the residuals vs fitted plot, we see a relatively flat line, upholding the zero conditional mean assumption.

Assumption MLR.5 (Homoskedasticity): This condition is not met as evident from looking at the Scale-Location plot. We would expect a uniform band, but we observe a range in variances when moving across the fitted values on the x-axis. Therefore, we use standard errors that are robust to heteroskedasticity in our analysis of model results.

Assumption MLR.6 (Normality of Errors): The normal Q-Q plot is fairly well-behaved except at very low and very high values. Since we have more than 30 samples, we can use the Central Limit Theorem to assume normality.

Discussion of CLM Assumptions 4-6: Nonviolent Crime Rate Model

```
par(mfrow=c(2,2))
plot(mod_2B_2, which=1)
plot(mod_2B_2, which=3)
plot(mod_2B_2, which=2)
plot(mod_2B_2, which=5)
```



Assumption MLR.4 (Zero Conditional Mean): In the residuals vs fitted plot, we observe a minor violation of the zero-conditional mean assumption. This is evident in the slight bend observed in the spline curve.

We would expect a flat line in the case of zero-conditional mean. However, the deviation is relatively mild and not particularly troubling.

Assumption MLR.5 (Homoskedasticity): This condition is not met as evident from looking at the Scale-Location plot. We would expect a uniform band, but instead, we see a narrow-wide-narrow range in variances when moving across the fitted values on the x-axis. This may be due to fewer data points at the low and high ends of the distribution. However, we will use standard errors that are robust to heteroskedasticity in our analysis of model results to be conservative.

Assumption MLR.6 (Normality of Errors): The Q-Q plot indicates a negative skew with a high slope on the left. This shows evidence of non-normality. However, we can use the Central Limit Theorem to assume normality.

Outliers

While technically not a classical linear model assumption, it is important to assess the impact of outliers on our models. For the violent crime model, all data points are within a Cook's distance of 0.5 and do not warrant concern. For the nonviolent crime model, there is one point that is at Cook's distance of 0.5. Generally, outliers with a Cook's distance greater than 1 may be exerting too much influence on a model. Since none of these outliers are above 1, no treatment is required.

5.0 | Confirmatory Data Analysis

In the exploratory phase, we identified patterns and features of the data and test different model specifications, finding that Model 2 most closely fulfills our original goal of identifying determinants of crime. In the confirmatory phase, we evaluate the robustness of the selected model using a separate subset of the data (the CDA partition), and interpret the model results.

Data Quality: CDA Partition

- Only `2` counties are categorized as `urban`, which is approximately `4` percent of the CDA partition. This confirms our assessment from the EDA partition that this indicator variable will not be very useful due to the extreme class imbalance.
- In this partition, the `prbconv` variable again has a positive skew with values over 1. We continue to believe that it is not unreasonable for the `prbconv` variable to exceed 1, as described in the EDA. Therefore, we do not manipulate these variables. The other "probability" variables (`prbarr` and `prbpris`) are well within the expected range of 0 and 1.
- There are two outliers on the high end and one outlier at the low end of the 'wfir' variable, representing the average weekly wage of those employed in finance, insurance, and real estate. Since compensation in these industries is often commission-based, we believe it reasonable to see outliers representing either very high or very low performance.
- As in the EDA partition, the `crmrte`, `polpc`, `density`, and `taxpc` show a positive skew in their distribution. To be consistent with the model specifications developed from the EDA partition, we apply a log transformation to `crmrte`, `density`, and `taxpc`.
- We know from the EDA partition that we must multiply the `pctymle` variable by 100 so that its unit is %. As in the EDA, we decide against transforming this variable, though it does have a positive skew.
- All other variables are within a reasonable range and distribution *within the CDA partition*.

Statistical Tests: CDA Partition

Now that we have evaluated several model specifications on the EDA partition and judged Model 2 to have the appropriate balance of parsimony and performance, we apply the Model 2 specs on the reserved CDA partition to determine if the models are reproducible, and then interpret the results.

```
CDA_mod_2A_2 = lm(C2$ftfrte_log ~ C2$pctpris + C2$hrwg + C2$density_log + C2$pctmin80)
CDA_mod_2B_2 = lm(C2$nvcrte_log ~ C2$pctpris + C2$hrwg + C2$density_log + C2$pctmin80)

se.CDA_mod_2A_2 = sqrt(diag(vcovHC(CDA_mod_2A_2)))
se.CDA_mod_2B_2 = sqrt(diag(vcovHC(CDA_mod_2B_2)))

names(mod_2A_2$coefficients) <- names(CDA_mod_2A_2$coefficients)
names(mod_2B_2$coefficients) <- names(CDA_mod_2B_2$coefficients)

se.mod_2A_2 = sqrt(diag(vcovHC(mod_2A_2)))
se.mod_2B_2 = sqrt(diag(vcovHC(mod_2B_2)))

stargazer(mod_2A_2, CDA_mod_2A_2, mod_2B_2, CDA_mod_2B_2,
          type = "latex", title = "Linear Models Predicting Crime Rates",
          dep.var.caption = "Model 2", column.labels  = c("EDA", "CDA", "EDA", "CDA"),
          omit.stat = c("f", "ser"),
          se = list(se.mod_2A_2, se.CDA_mod_2A_2, se.mod_2B_2, se.CDA_mod_2B_2),
          star.cutoffs = c(0.05, 0.01, 0.001),
          add.lines=list(c("AIC", round(AIC(mod_2A_2),1), round(AIC(CDA_mod_2A_2),1),
                           round(AIC(mod_2B_2),1), round(AIC(CDA_mod_2B_2),1))))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Aug 07, 2018 - 14:39:12

Applying Model 2 specifications on the CDA partition yields similar results to our EDA. The main difference is that `hrwg` is no longer significant for the violent crime rate model. For nonviolent crime, `pctpris` is highly statistically significant with a 1% increase in the convictions resulting in prison sentence per offense associated with a 6% reduction in nonviolent crime. Population density is highly statistically significant for nonviolent crime, but just below the threshold for statistical significance for face-to-face crime. While a 1% increase in population density seems to have a small effect of 0.3-0.5% increase in either type of crime, there are orders of magnitude between the densities of the least and most dense counties, so this is practically significant. The proportion of minorities in a county's population as of 1980 is highly statistically significant for both violent and nonviolent crime rate models. A 1% increase in the proportion of minorities in a county is associated with about a 2.5% increase in crime rate for violent crime and a 1.2% increase in crime rate for nonviolent crime.

Since the certainty of punishment (`pctpris`) is a significant component impacting the variation in crime rate, we study the components of this variable more closely to assess if the probability of arrest, conviction, or receiving a prison sentence are equal deterrents of crime.

```
model_deter <- lm(C2$nvcrte_log ~ C2$prbarr + C2$prbconv + C2$prbpris + C2$hrwg +
                    C2$density_log + C2$pctmin80 +C2$pctymle)
vcovHC <- vcovHC(model_deter)
coeftest(model_deter, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      -4.3998030  0.8368919 -5.2573 6.354e-06 ***
```

Table 4: Linear Models Predicting Crime Rates

	Model 2			
	ftfrte_log	ftfrte_log	nvcrte_log	nvcrte_log
	EDA	CDA	EDA	CDA
	(1)	(2)	(3)	(4)
pctpris	−0.091	−0.072	−0.060**	−0.063**
	(0.054)	(0.038)	(0.020)	(0.022)
hrwg	−0.211*	0.084	0.005	0.147
	(0.105)	(0.193)	(0.075)	(0.094)
density_log	0.382	0.343	0.484***	0.286*
	(0.203)	(0.200)	(0.104)	(0.123)
pctmin80	0.025***	0.025***	0.016***	0.012***
	(0.006)	(0.005)	(0.003)	(0.002)
Constant	−4.180***	−6.810***	−3.814***	−4.753***
	(0.761)	(1.611)	(0.614)	(0.777)
AIC	69.7	77.7	34.5	21
Observations	44	45	44	45
R^2	0.562	0.500	0.718	0.680
Adjusted R^2	0.517	0.450	0.689	0.648
Note:		*p$<$0.05; **p$<$0.01; ***p$<$0.001		

```
## C2$prbarr      -1.5691805  0.4291626 -3.6564 0.0007903 ***
## C2$prbconv     -0.5911589  0.3201363 -1.8466 0.0728178 .
## C2$prbpris     -0.7863812  0.6427516 -1.2235 0.2288906
## C2$hrwg         0.1781583  0.0979724  1.8185 0.0770971 .
## C2$density_log  0.2477333  0.1473908  1.6808 0.1012277
## C2$pctmin80     0.0117261  0.0028211  4.1566 0.0001834 ***
## C2$pctymle      0.0125050  0.0288903  0.4328 0.6676412
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
test <- linearHypothesis(model_deter, "C2$prbarr = C2$prbconv", vcov = vcovHC)
test$`Pr(>F)`[2]
```

```
## [1] 0.003073178
```

```
test <- linearHypothesis(model_deter, "C2$prbarr = C2$prbpris", vcov = vcovHC)
test$`Pr(>F)`[2]
```

```
## [1] 0.2201627
```

```
test <- linearHypothesis(model_deter, "C2$prbconv = C2$prbpris", vcov = vcovHC)
test$`Pr(>F)`[2]
```

```
## [1] 0.68721
```

From examining the individual t-tests of the coefficients in the nonviolent crime model, the probability of arrest is statistically significant while the probabilities of conviction or receiving a prison sentence upon being convicted are not statistically significant.

When testing to evaluate if two coefficients are different, we cannot reject the null hypothesis that `prbconv` and `prbpris` or `prbpris` and `prbarr` have the same slopes, since the p-values for both of these F-tests are far above the threshold of 0.05. However, the p-value of the F-test when comparing the coefficients of `prbarr` with the coefficient of `prbconv` is statistically significant with a value of 0.02. Therefore, the probability of arrest has a slope different from the probability of conviction. The probability of arrest also has the greatest practical significance of the three variables as a 1% (absolute) increase in the probability of arrest associated with a ~1.6% decrease in nonviolent crime.

```
model_deter <- lm(C2$ftfrte_log ~ C2$prbarr + C2$prbconv + C2$prbpris + C2$hrwg +
                    C2$density_log + C2$pctmin80 + C2$pctymle)
vcovHC <- vcovHC(model_deter)
coeftest(model_deter, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.3583824  1.6762181 -3.1967 0.002844 **
## C2$prbarr       -1.5816694  0.7868651 -2.0101 0.051755 .
## C2$prbconv      -1.1965111  0.5025370 -2.3809 0.022529 *
## C2$prbpris      -0.2615575  1.1939778 -0.2191 0.827804
## C2$hrwg          0.0973075  0.1820042  0.5346 0.596094
## C2$density_log   0.3031787  0.2176431  1.3930 0.171931
## C2$pctmin80      0.0245818  0.0056842  4.3246 0.000111 ***
## C2$pctymle      -0.0987799  0.0822835 -1.2005 0.237582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
test <- linearHypothesis(model_deter, "C2$prbarr = C2$prbconv", vcov = vcovHC)
test$`Pr(>F)`[2]
```

```
## [1] 0.5924528
```

```
test <- linearHypothesis(model_deter, "C2$prbarr = C2$prbpris", vcov = vcovHC)
test$`Pr(>F)`[2]
```

```
## [1] 0.3404286
```

```
test <- linearHypothesis(model_deter, "C2$prbconv = C2$prbpris", vcov = vcovHC)
test$`Pr(>F)`[2]
```

```
## [1] 0.4399302
```

From examining the individual t-tests of the coefficients in the violent crime model, the probability of conviction is statistically significant, and the probability of arrest is borderline statistically significant. An increase in `prbarr` appears to have the largest effect of the three, as a 1% (absolute) change in `prbarr` would correspond to a ~1.6% decrease in violent crime (same as with the nonviolent crime model). `prbconv` shows a deterrent effect of similar magnitude. When testing to evaluate if two coefficients are different, we cannot reject the null hypothesis that for any of the tests comparing the slopes of `prbarr`, `prbconv`, and `prbpris` to one another.

6.0 | Omitted Variables

Omitted variable bias (OVB) can occur in OLS regression when a variable that is correlated with both the outcome variable and one or more of the explanatory variables is omitted from the regression equation. This omission can introduce bias into the model, resulting in under- or over-estimation of the explanatory variables' coefficients.

The variables discussed in this section could be considered to be omitted variables in the models described in the preceding sections. Specifically, the analysis in this section is based on our selected violent crime model (Model 2_2). Note that we choose to analyze the violent crime model over the nonviolent crime model because its variables show less statistical significance, which indicates a greater possibility of omitted variable bias. This model can be conceptualized as follows:

Violent Crime Rate = β_0 *+ (*β_1 *Percent of Offenses Resulting in Prison Sentence) + (*β_2 *Hourly Wage (Average Across Industries)) + (*β_3 *Population Density) + (*β_4 *Minority Population) + u*

The model's formula is as follows:

ftfrte_log = -4.180 + (-0.091 pctpris) + (-0.211 hrwg) + (0.382 density_log) + (0.025 pctmin80)

Illegal Drugs

The connection between crime and illegal drugs is strong and complex; the sale and use of such drugs are crimes in themselves, and serious drug use can intensify other criminal activity. Drug crimes range from use-related (where a crime is committed as a result of the perpetrator being on drugs) to economic-related (where a crime is committed to fund a drug habit) to system-related (where a crime is committed to produce, transport, or sell drugs).

Crime Rate = β_0 *+ (*β_1 *Percent of Offenses Resulting in Prison Sentence) + (*β_2 *Hourly Wage (Average Across Industries)) + (*β_3 *Population Density) + (*β_4 *Minority Population) + (*β_5 *Drugs) + u*

Drugs = α_0 *+ (*α_1 *Percent of Offenses Resulting in Prison Sentence) + (*α_2 *Hourly Wage (Average Across Industries)) + (*α_3 *Population Density) + (*α_4 *Minority Population) + v*

We would expect the correlation between crime rate and illegal drugs to be positive, therefore $\beta_5 > 0$. We would also expect a positive correlation with the percentage of offenses resulting in prison sentences (certainty of punishment) and with the proportion of minorities in the population, so α_1 and $\alpha_4 > 0$. Therefore, for certainty of punishment, $OVB > 0$ and the OLS coefficient on this explanatory variable would be scaled toward zero (less negative), losing statistical significance. For minority population, $OVB > 0$ and the OLS coefficient on this explanatory variable would be scaled away from zero (more positive), gaining statistical significance. Though we would expect the effects on each of the two explanatory variables to be large, because they are in opposite directions, the net effect of the bias from drugs could be small.

Proxy variables for prevalence of drugs could include the following:

- Reported drug use (commonly ascertained through surveys such as the HSA's National Survey on Drug Use and Health). However, this data could fail to accurately represent the prevalence of drugs due to respondents' reluctance to admit to illegal drug use.
- Arrests for drug offenses or drug seizure rates (tracked by FBI and DEA, as well as other law enforcement agencies). However, these variables are not a pure measure of the prevalence of drugs since they are highly dependent on the effectiveness of law enforcement. In addition, arrests for drug offenses are already included in the probability of arrests, which is included in the calculation for probability of prison sentence.

Alcohol

Because alcohol is legal and widely available, it plays a particularly strong role in crimes. Alcohol is estimated to be a factor in 40% of all violent crimes, and according to the Department of Justice, 37% of almost two million convicted offenders currently incarcerated report that they were drinking at the time of their arrest (Wilcox 2015). In cases of domestic violence, not only is the perpetrator likely to have been drinking, the victims are more likely to abuse alcohol themselves.

Crime Rate = β_0 + (β_1 Percent of Offenses Resulting in Prison Sentence) + (β_2 Hourly Wage (Average Across Industries)) + (β_3 Population Density) + (β_4 Minority Population) (β_5 Alcohol) + u

Alcohol = α_0 + (α_1 Percent of Offenses Resulting in Prison Sentence) + (α_2 Hourly Wage (Average Across Industries)) + (α_3 Population Density) + (α_4 Minority Population) + v

We would expect the correlation between crime rate (particularly face-to-face crime rate) and alcohol to be positive, therefore $\beta_5 > 0$. We would also expect a positive correlation with the percentage of offenses resulting in prison sentences (certainty of punishment), so $\alpha_1 > 0$. Therefore, $OVB > 0$ and the OLS coefficient on this explanatory variable would be scaled toward zero (less negative), losing statistical significance. We would expect the bias from prevalence of alcohol to be large.

Proxy variables for prevalence of alcohol could include the following:

- Alcohol sales. However, this is an imperfect proxy because alcohol sales would not directly correspond to alcohol consumption, consumption would not directly correspond to abuse, and abuse would not directly correspond to crime.
- Reported alcohol use. However, this data could fail to accurately represent the prevalence of alcohol due to respondents' under-reporting their alcohol consumption.

Education

Research has found that there is a strong relationship between education and crime. For example, educational attainment has been found to to lower the probability of incarceration significantly, and differences in educational attainment between black and white men can explain a large portion of the black-white gap in male incarceration rates (Lochner, Lance, & Moretti 2004). Education also increases economic opportunity, which in this model is proxied by earnings.

Crime Rate = β_0 + (β_1 Percent of Offenses Resulting in Prison Sentence) + (β_2 Hourly Wage (Average Across Industries)) + (β_3 Population Density) + (β_4 Minority Population) + (β_5 Education) + u

Education = α_0 + (α_1 Percent of Offenses Resulting in Prison Sentence) + (α_2 Hourly Wage (Average Across Industries)) + (α_3 Population Density) + (α_4 Minority Population) + v

We would expect the correlation between crime rate and education to be negative, therefore $\beta_5 < 0$. We would also expect negative correlation with the percentage of offenses resulting in prison sentences (certainty of punishment) and minority population, and positive correlation with average hourly wage, so α_1 and $\alpha_4 < 0$ and $\alpha_2 > 0$. Therefore, for certainty of punishment, $OVB > 0$ and the OLS coefficient on this explanatory variable would be scaled toward zero (less negative), losing statistical significance. For average wage, $OVB < 0$ and the OLS coefficient on this explanatory variable would be away from zero (more negative), gaining statistical significance. For minority population, $OVB > 0$ and the OLS coefficient on this explanatory variable would be scaled away from zero (more positive), gaining statistical significance. Though some muting might occur because of opposite signs, we would expect the net effect of the bias from education to be large.

Proxy variables for education could include the following:

- High school attendance rate. Truancy has been shown in other research to be linked to crime.

- High school graduation rate.
- College degree attainment rate.

Unemployment

Unemployment itself has not been found to have a strong link to crime. However, sustained unemployment can represent a systemic lack of economic opportunity, which among specific groups (particularly young people) can lead to criminality (particularly property crimes). It can increase the perceived benefit of criminal activity by both increasing levels of criminal motivation as well as the vulnerability of criminal targets. Unemployment statistics have long illustrated racial disparities in the economy, with black unemployment nearly double the national average this year (Ell 2018).

Crime Rate = β_0 *+ (*β_1 *Percent of Offenses Resulting in Prison Sentence) + (*β_2 *Hourly Wage (Average Across Industries)) + (*β_3 *Population Density) + (*β_4 *Minority Population) + (*β_5 *Unemployment) + u*

Unemployment = α_0 *+ (*α_1 *Percent of Offenses Resulting in Prison Sentence) + (*α_2 *Hourly Wage (Average Across Industries)) + (*α_3 *Population Density) + (*α_4 *Minority Population) + v*

We would expect the correlation between crime rate and unemployment to be positive, therefore $\beta_5 > 0$. We would expect a negative correlation with average hourly wage and a positive correlation with minority population, so $\alpha_2 < 0$ and $\alpha_4 > 0$. Therefore, for average wage, $OVB < 0$ and the OLS coefficient on this explanatory variable would be toward zero (less positive), losing statistical significance. For minority population, $OVB > 0$ and the OLS coefficient on this explanatory variable would be scaled away from zero (more positive), gaining statistical significance. Though we would expect the effects on each of the two explanatory variables to be large, because they are in opposite directions, we would expect the net effect of the bias from unemployment to be small.

General unemployment should not require a proxy variable, as the statistic is easily measurable and closely tracked by the United States Bureau of Labor Statistics. However, it is important to note that this data does not include the unemployed who have ceased to actively seek new employment, which is the group of most interest to this study since it is likely to include a higher proportion of those who have turned to crime to earn a living. A more meaningful measurement might be proportion of offenders (arrested or convicted) who were unemployed at the time of their crime. Poverty rate and income inequality, while different from unemployment, might have similar relationships with crime.

Distribution of Crime Types (Crime Severity)

More severe crimes should be more likely to result in incarceration and longer prison sentences. Our analysis is missing a key factor in that we're applying general incarceration and sentencing rates to the more specific outcome variables of violent and nonviolent crime. Ideally, we would parse the independent variables of how these cases are treated to their appropriate categories, meaning felony violent crime rates are viewed as a function of felony violent crime incarceration rates.

Felony Violent Crime Rate = β_0 *+ (*$\cancel{\beta_1}$ *Percent of Offenses Resulting in Prison Sentence) + (*β_2 *Hourly Wage (Average Across Industries)) + (*β_3 *Population Density) + (*β_4 *Minority Population) + (*β_5 *Percent of Felony Violent Offenses Resulting in Prison Sentences) + u*

In this case, we do not consider the standard OVB form, as we are replacing one variable completely with a more appropriate one. We did not see a strong relationship for average sentence in our regression, but it would be worth revisiting after separating our crimes into appropriate categories.

Other Variables

Other potential omitted variables could include family structure, mental illness, access to social services, prevalence of firearms, and rate of recidivism. However, although each of these variables would seem to have a relationship with crime rate, none appear to have strong relationships with any of the other explanatory variables.

7.0 | Conclusion

The purpose of this study is to identify relationships between crime rate and other county-level characteristics and to generate policy suggestions that are supported by the research findings.

Our analysis of the isolated face-to-face crime rate provides interesting insights into factors related specifically to violent crime. Compared to the nonviolent crime rate, average hourly wage in our EDA model has much greater statistical significance and is far more impactful (a $1 increase corresponds to a 21% decrease in the face-to-face crime rate, more than four times higher than the decrease in nonviolent crime with the same change) based on our EDA model results. We were not able to confirm a similar effect in our CDA model, and the effect changed directions, so we are cautious about moving forward with this recommendation. We find that the proportion of offenses resulting in prison sentences and population density did not have statistical significance at $p<0.05$. While we do not have a scientifically publishable result, we can still advise that the threat of incarceration (particularly the probability of arrest) appears to have a deterrent effect on violent crime in many North Carolina communities.

Our analysis of the isolated non-face-to-face crime rate provides interesting insights about factors related specifically to nonviolent crime. Contrary to the violent crime rate, we find that the proportion of offenses resulting in prison sentences is highly significant. Population density also has strong statistical significance. Unlike our EDA model results for violent crime, average hourly wage is not significant and has a very small impact in the nonviolent crime models.

Our analysis of both violent and nonviolent crime rates shows that minority population has a statistically significant positive relationship with both types of crime, although the effect sizes are relatively small. The issues surrounding race and crime are numerous and complex; the findings of this research suggest many avenues for further study, including of related socioeconomic factors such as poverty, drug use, and education, as well as institutionalized racial discrimination in law enforcement and the judicial system.

We set out to find the different determinants in nonviolent and violent crime, but for the most part, found the same deterrent and driving factors for each. Based on our findings, we recommend the following positions for the candidate's platform:

Reducing Crime in General

Platform Position 1: Reduce crime by increasing the criminal's perception of probability of arrest
Our analysis shows that a 1% increase in a criminal's chance of receiving a prison sentence results in a ~6% decrease in the overall crime rate. Parsing this composite metric into its constituent parts, we find that the probability of arrest is a particularly strong deterrent (significantly stronger than `prbconv` or `prbpris` in the nonviolent model) meaning what happens after arrest appears to have a less deterrent effect. Our theory for this, is that a criminal considering the risk associated with a crime may only consider the more immediate threat of being arrested, rather than the less immediate consequences. This solution is aimed at increasing the perception of this threat to would be criminals.

We recommend introducing a high-profile and heavily publicized effort that shows North Carolina is making a change and taking a stand against crime. One method of doing this is to establish a special state-wide task force that is focused on finding headline-worthy crimes. This task force will create commercials announcing their presence, push stories to local news outlets with pictures of them taking down perpetrators and attempt to stay in the media as much as possible. The primary goal of the unit is to become a household name, making a constant impression on potential criminals about the state of law enforcement in North Carolina. This unit should not need to make a meaningful change to arrest rates, which we are aware can have deleterious effects on a community. We believe that a small amount of very public police activity could make a much larger impact on the real metric, the criminal's perception of the risk of being arrested.

Platform Position 2: Reduce crime through community-oriented policing. A 10% decrease in population density is associated with a 5% decrease in the nonviolent crime rate. However, it is impractical to try to decrease population density through policy, and any attempt to do so would potentially undo the benefits of density, such as increased affordability of housing and decreased commute times. We believe that population density itself is not a problem; however, as discussed elsewhere in this report, density is a proxy for opportunity to commit crimes. Community-oriented policing seeks to reduce this opportunity through building relationships between communities and the law enforcement agencies that serve them. One of its main goals is to prevent criminal activity instead of responding to it, in part through assigning officers to a defined area ("beat") in which they become familiar with its residents and the specific crimes they experience. Highly dense areas lend themselves to this approach, as they allow for foot patrols and face-to-face interactions.

Reducing Violent Crime Specifically

Platform Position 3: Reduce violent crime by increasing the minimum wage. Many people believe that an increase in the minimum wage can lift working citizens out of poverty, increase worker productivity and job satisfaction, strengthen local economies, and reduce the need for government assistance. However, the public is largely unaware that higher wages can reduce violent crime. Based on our EDA regression, raising North Carolina's current minimum wage of $2.90/hour to $3.35/hour (as per the current proposal to the General Assembly of North Carolina in House Bill 229) could help reduce the rate of violent crime by approximately 10%. Unfortunately, we weren't able to confirm a similar effect in our CDA sample, so we must temper this recommendation with that result.

As governor of North Carolina, our client will be tasked with distributing millions of state and federal dollars to agencies and organizations across the state for programs to reduce crime. The governor's office has a civic duty to ensure that it is selecting and funding initiatives that address the state's greatest needs and most effective solutions. The reasons for their selection should be based on objective facts, not tradition, emotion, or political expediency. The data-driven insights in this report will allow the candidate to make well-substantiated, practical policy proposals that will benefit North Carolina today and for decades to come.

References

Becker, G. S. (1968). “Crime and Punishment: An Economic Approach,” Journal of Political Economy 76, 169-217.

Ell, K. (2018, May 04). April jobs report shows racial disparities in unemployment rates continue. Retrieved August 7, 2018, from https://www.cnbc.com/2018/05/04/aprils-jobs-report-shows-racial-inequalities-in-unemployment-rate.html

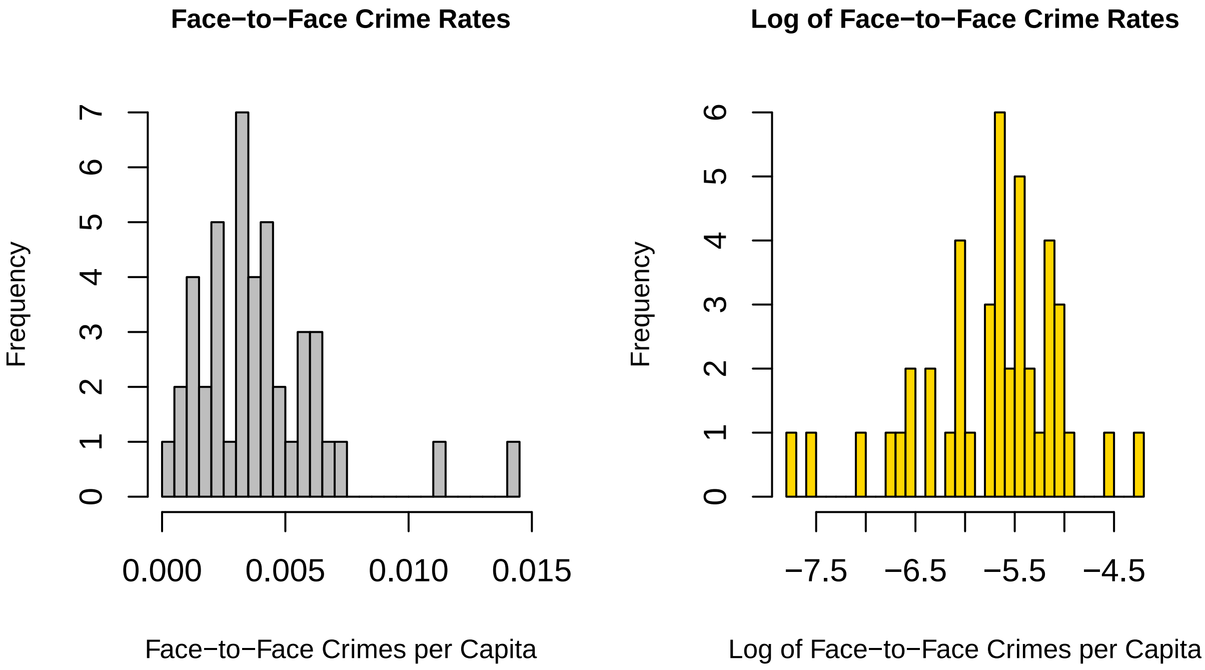
Lochner, Lance, and Enrico Moretti. (2004). “The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports.” American Economic Review, 94 (1): 155-189.

Lombroso-Ferrero, G., & Lombroso, C. (1911). *Criminal man, according to the classification of Cesare Lombroso*. New York: Putnam.

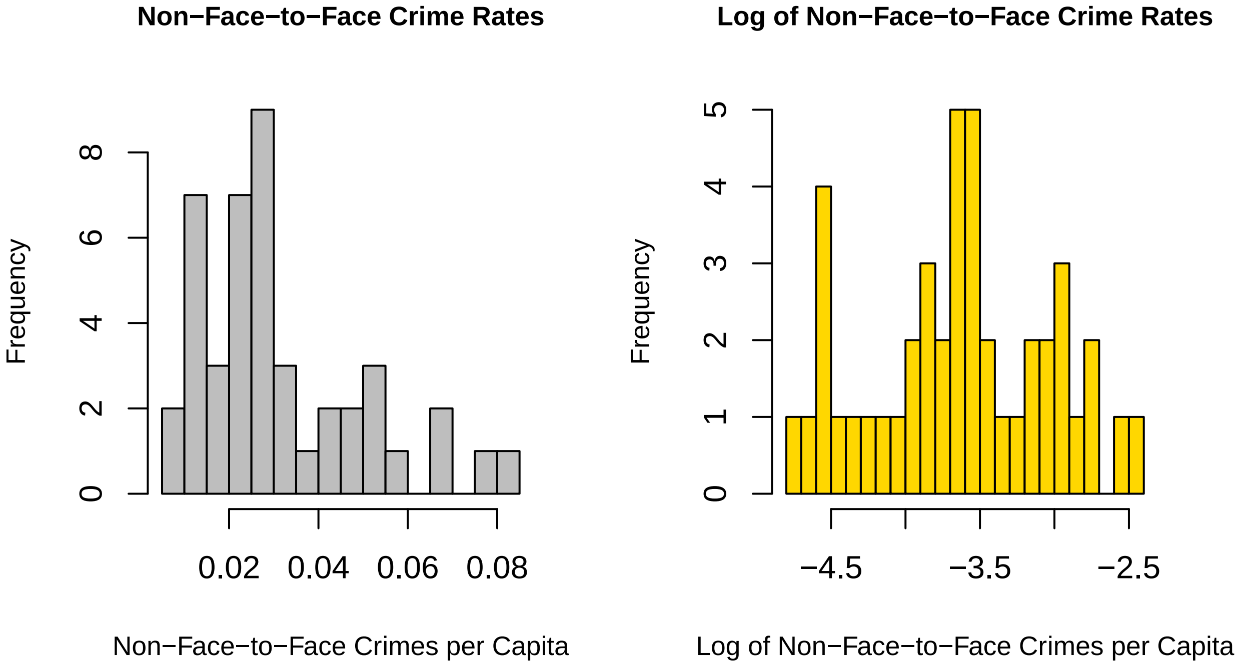
Wilcox, S. (2015, June 27). Alcohol, Drugs and Crime. Retrieved August 7, 2018, from https://www.ncadd.org/about-addiction/alcohol-drugs-and-crime

Appendix A | Histograms of Outcome Variables (EDA Partition)

```r
par(mfrow=c(1,2))
hist(C1$ftfrte, breaks = 25, col="Gray",
     main = "Face-to-Face Crime Rates", cex.main = .8,
     xlab = "Face-to-Face Crimes per Capita", cex.lab=0.8)
hist(C1$ftfrte_log, breaks = 25, col="Gold",
     main = "Log of Face-to-Face Crime Rates", cex.main = .8,
     xlab = "Log of Face-to-Face Crimes per Capita", cex.lab=0.8)
```

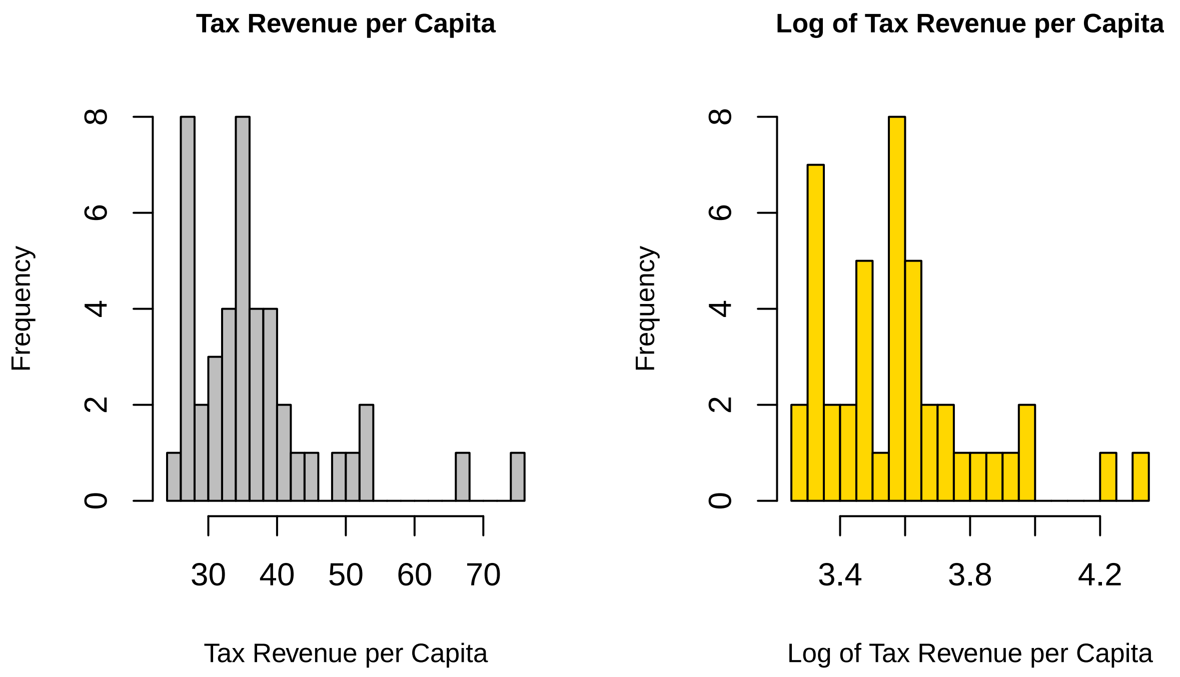


```r
par(mfrow=c(1,2))
hist(C1$nvcrte, breaks = 25, col="Gray",
     main = "Non-Face-to-Face Crime Rates", cex.main = .8,
     xlab = "Non-Face-to-Face Crimes per Capita", cex.lab=0.8)
hist(C1$nvcrte_log, breaks = 25, col="Gold",
     main = "Log of Non-Face-to-Face Crime Rates", cex.main = .8,
     xlab = "Log of Non-Face-to-Face Crimes per Capita", cex.lab=0.8)
```

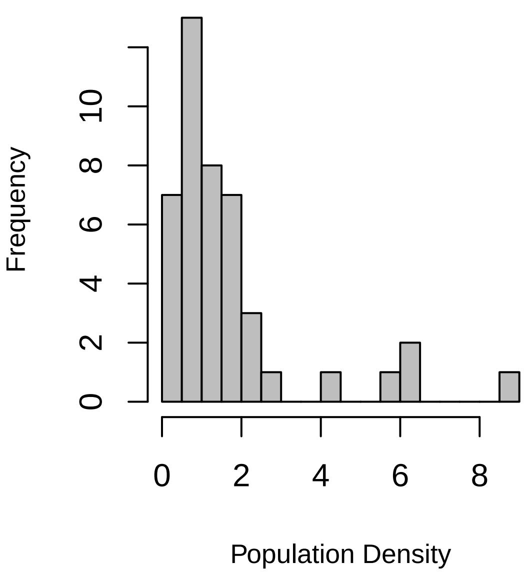

Non−Face−to−Face Crime Rates
Frequency
0
2
4
6
8
0.02
0.04
0.06
0.08
Non−Face−to−Face Crimes per Capita
Log of Non−Face−to−Face Crime Rates
Frequency
0
1
2
3
4
5
−4.5
−3.5
−2.5
Log of Non−Face−to−Face Crimes per Capita

Appendix B | Histograms of Transformed Explanatory Variables (EDA Partition)

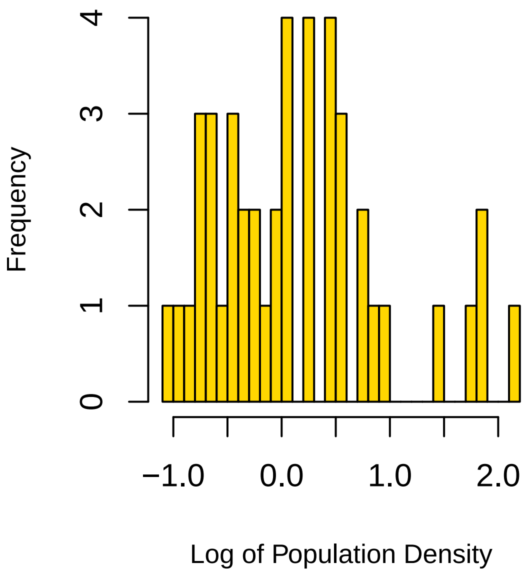
```
par(mfrow=c(1,2))
hist(C1$taxpc, breaks = 25, col="Gray",
     main = "Tax Revenue per Capita", cex.main = .8,
     xlab = "Tax Revenue per Capita", cex.lab=0.8)
hist(C1$taxpc_log, breaks = 25, col="Gold",
     main = "Log of Tax Revenue per Capita", cex.main = .8,
     xlab = "Log of Tax Revenue per Capita", cex.lab=0.8)
```



```
par(mfrow=c(1,2))
hist(C1$density, breaks = 25, col="Gray",
     main = "Population Density", cex.main = .8,
     xlab = "Population Density", cex.lab=0.8)
hist(C1$density_log, breaks = 25, col="Gold",
     main = "Log of Population Density", cex.main = .8,
     xlab = "Log of Population Density", cex.lab=0.8)
```



Population Density

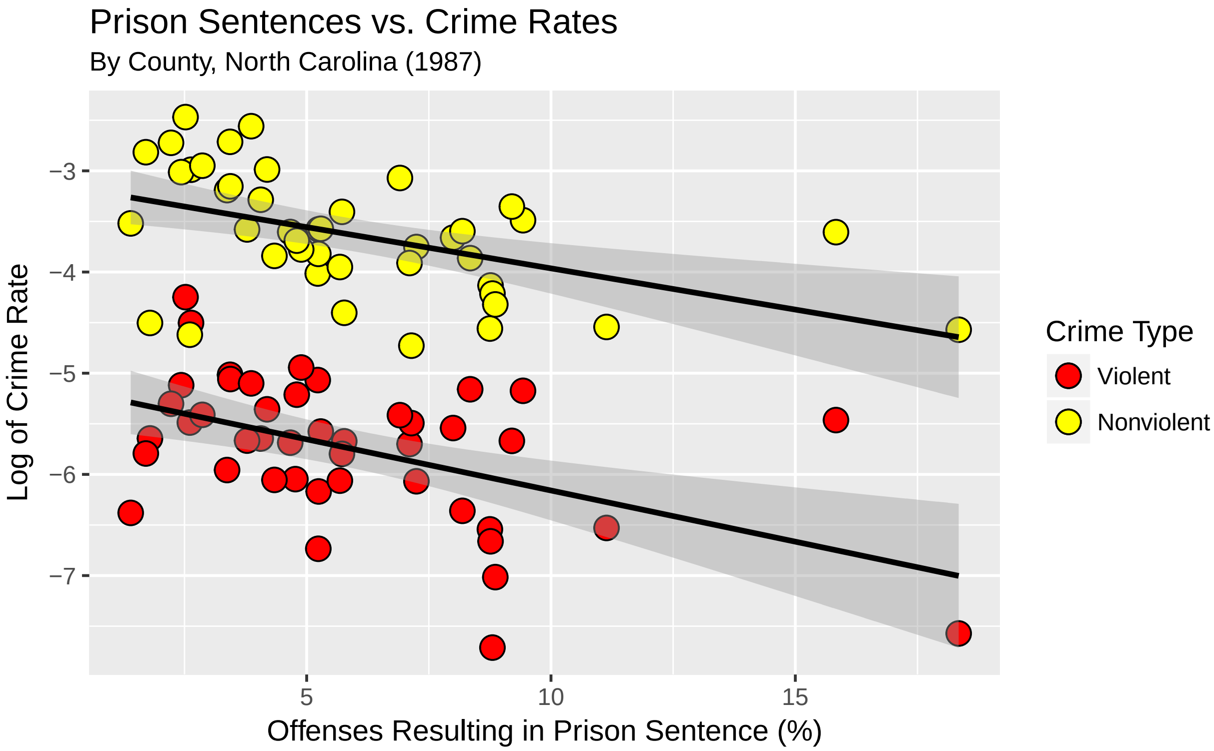


Log of Population Density

Appendix C | Additional Scatterplots (EDA Partition)

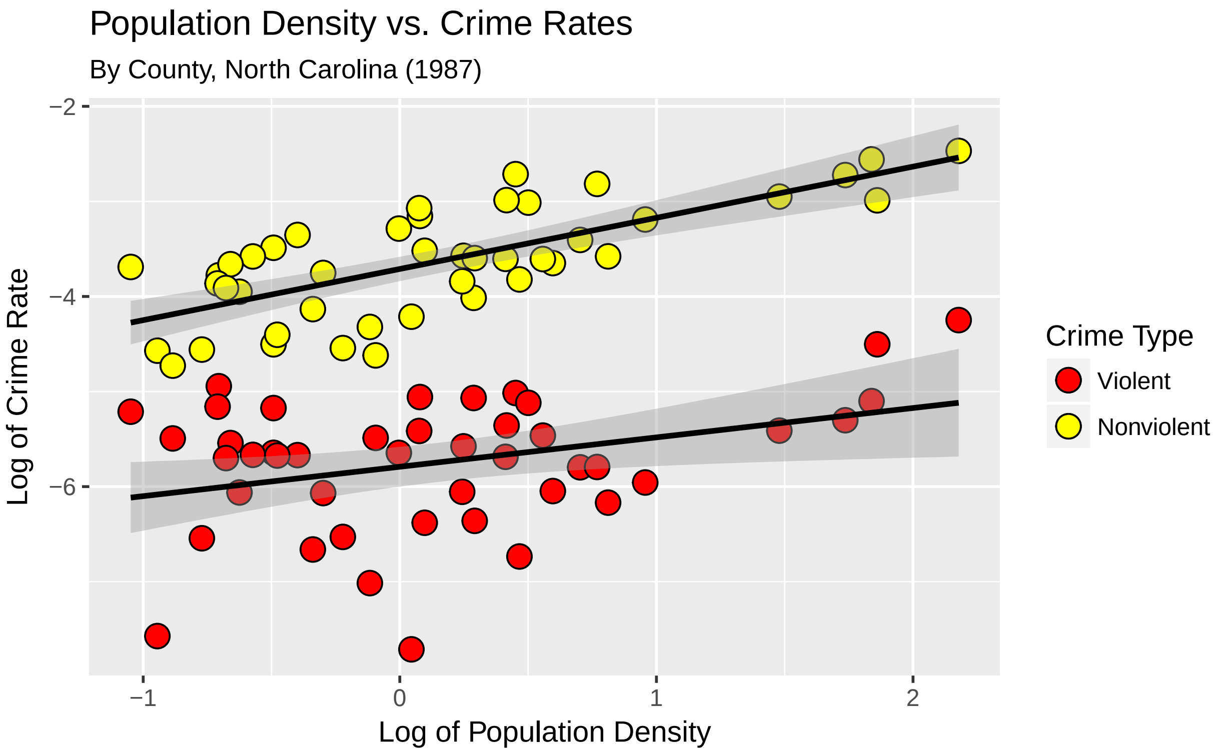
Prison sentences versus crime rates after removal of extreme outlier as described in Section 3.1 (Key Variable Analysis):

```
ftf_nv_plotter(C1$pctpris, t = "Prison Sentences vs. Crime Rates",
               xl = "Offenses Resulting in Prison Sentence (%)")
```

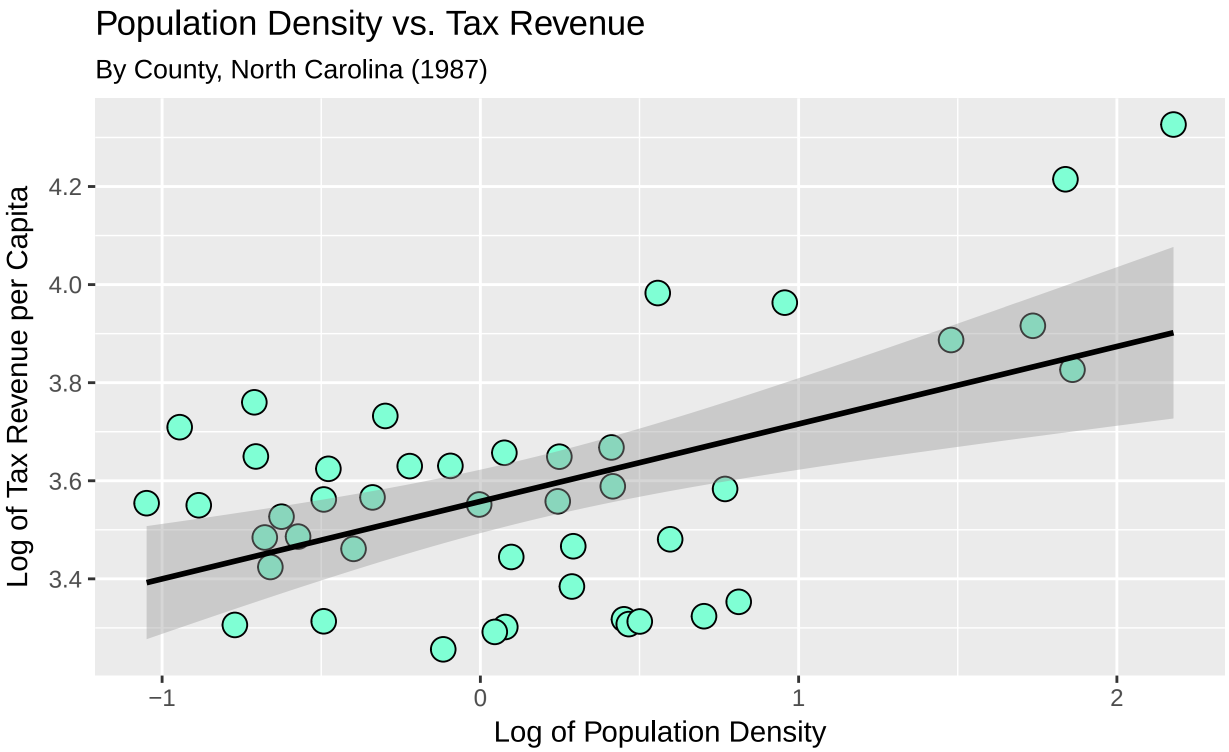


Prison sentences versus crime rates after imputation of extreme outlier as described in Section 3.1 (Key Variable Analysis):

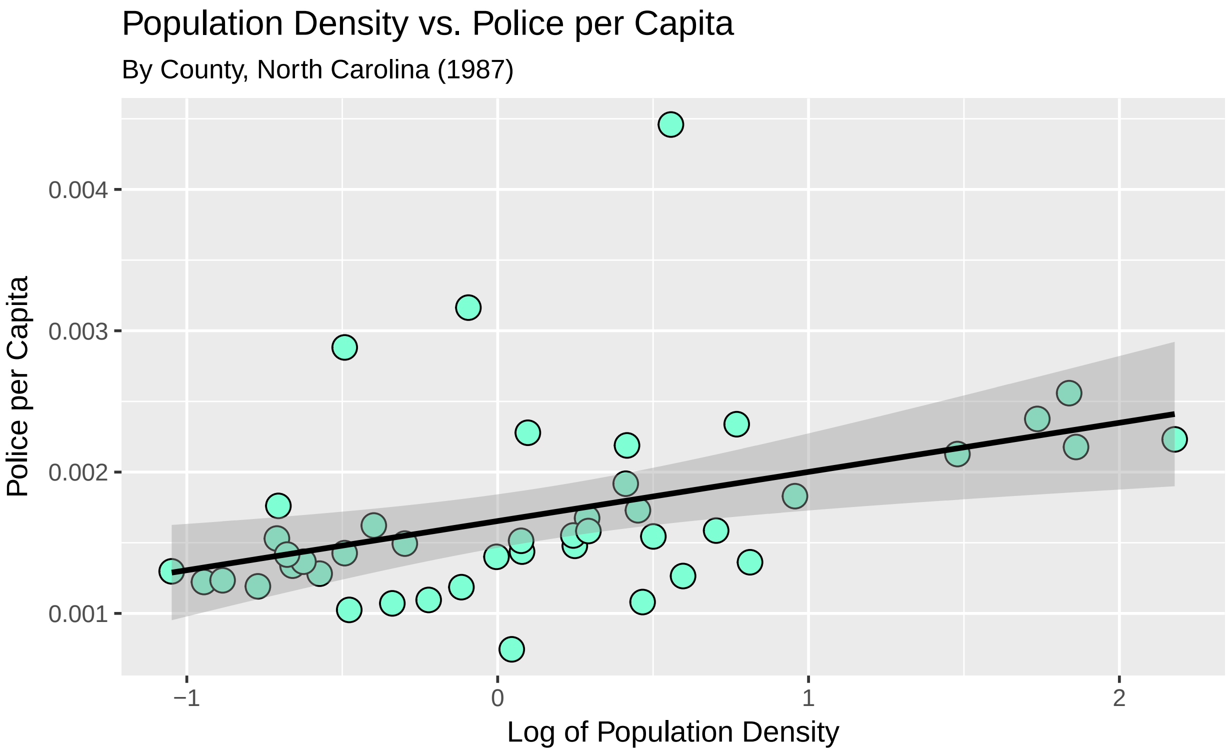
```
ftf_nv_plotter(C1$density_log, t = "Population Density vs. Crime Rates",
               xl = "Log of Population Density")
```



```
ggplot(C1) +
  geom_point(aes(density_log, taxpc_log, color=taxpc_log), size=4,
             color="black", fill="aquamarine", pch=21) +
  geom_smooth(aes(density_log, taxpc_log), method = "lm", se=TRUE,
              color="black", linetype="solid", size=1) +
  labs(title = "Population Density vs. Tax Revenue", subtitle = "By County, North Carolina (1987)") +
  labs (x = "Log of Population Density",
        y = "Log of Tax Revenue per Capita")
```



```
ggplot(C1) +
  geom_point(aes(density_log, polpc, color=taxpc_log), size=4,
             color="black", fill="aquamarine", pch=21) +
  geom_smooth(aes(density_log, polpc), method = "lm", se=TRUE,
              color="black", linetype="solid", size=1) +
  labs(title = "Population Density vs. Police per Capita", subtitle = "By County, North Carolina (1987)") +
  labs (x = "Log of Population Density",
        y = "Police per Capita")
```


Population Density vs. Police per Capita
By County, North Carolina (1987)
Police per Capita
0.004
0.003
0.002
0.001
−1
0
1
2
Log of Population Density