Lab 4: Reducing Crime

Kim Vignola, Kiersten Hendersen, Aaron Yuen

August 17, 2017

Section 1: Introduction

The authors, Kim, Kiersten and Aaron, were hired to provide research for a political campaign in North Carolina to understand the determinants of crime (both correlational and causal) using exploratory data analysis and OLS regression. The end goal is to leverage the data to provide policy suggestions that are applicable to local government to reduce crime.

The provided cross-sectional dataset consists of statistics for a selection of counties for a given year. Data for 90 counties and 25 variables for each county were provided.

For this analysis, the following assumptions were made:

- The 90 counties provided were randomly sampled among the 100 counties in North Carolina.

Section 2: Exploratory Analysis

Data Load and Library Imports

Reading the data and loading the right libraries:

```r
library(car)
library(corrplot)
library(sandwich)
library(stargazer)

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

Univariate Variable Analysis

There are 90 data points and 25 variables

```r
nrow(data)
```

```
## [1] 90
```

```r
colnames(data)
```

```
##  [1] "X"        "county"   "year"     "crime"    "probarr"  "probconv"
##  [7] "probsen"  "avgsen"   "police"   "density"  "tax"      "west"
## [13] "central"  "urban"    "pctmin"   "wagecon"  "wagetuc"  "wagetrd"
## [19] "wagefir"  "wageser"  "wagemfg"  "wagefed"  "wagesta"  "wageloc"
## [25] "mix"      "ymale"
```

There doesn't seem to be any NAs in the dataset.

```r
apply(!is.na(data[,]), MARGIN = 2, mean)
```

```
##        X   county     year    crime  probarr probconv  probsen   avgsen
##        1        1        1        1        1        1        1        1
##   police  density      tax     west  central    urban   pctmin  wagecon
##        1        1        1        1        1        1        1        1
##  wagetuc  wagetrd  wagefir  wageser  wagemfg  wagefed  wagesta  wageloc
##        1        1        1        1        1        1        1        1
##      mix    ymale
##         1         1
```

The following summarizes the different variables types based on the variable desciptions and basic understanding of the data:

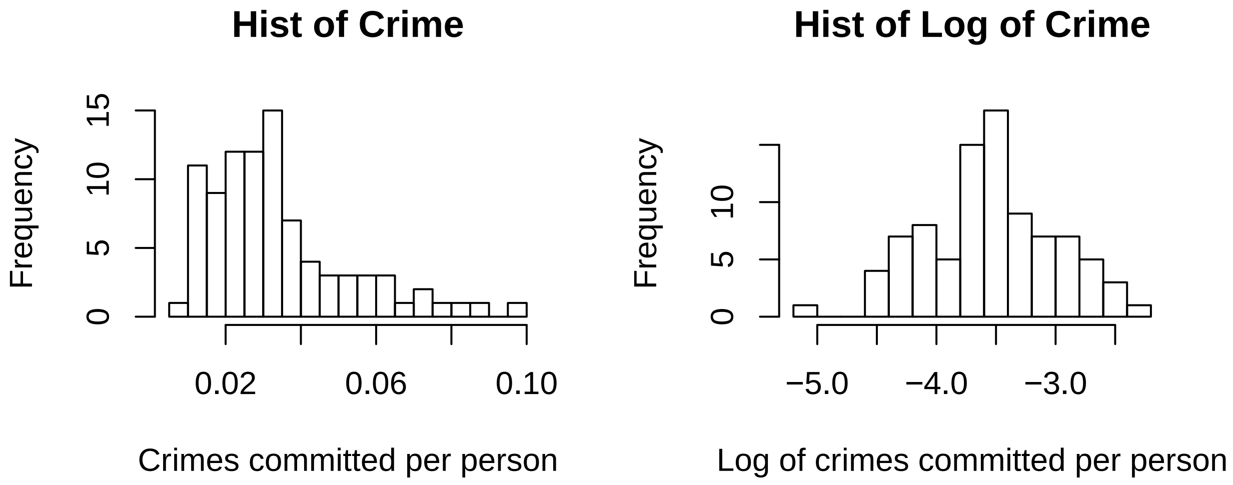
1. Rates, averages, and probabilities - crime, probarr, probconv, probsen, avgsen, police, density, pctmin, mix, ymale
2. $ variables - tax, wagecon, wagetuc, wagetrd, wagefir, wageser, wagemfg, wagefed, wagesta, wageloc
3. Indicator variables - west, central, urban. No base categories are in the dataset (e.g. non-west/central, rural)
4. Other miscellaneous variables - X, county, year

The remaining EDA focuses the analysis on only the key variables of interest.

Crimes Committed per Person (crime)

Crime is the main dependent variable of interest. Looking at the histogram of crime, the distribution tends to be right skewed. Taking the log of crime tends to make the histgram appear more normal. As a result, for our modeling, we will proceed with using log of crime as the dependent variable.

```
par(mfrow=c(1,2))
hist(data$crime, breaks=20,
     main="Hist of Crime",
     xlab="Crimes committed per person", cex=0.7)
hist(log(data$crime), breaks=20,
     main="Hist of Log of Crime",
     xlab="Log of crimes committed per person", cex=0.7)
```



Probability of Conviction (probconv)

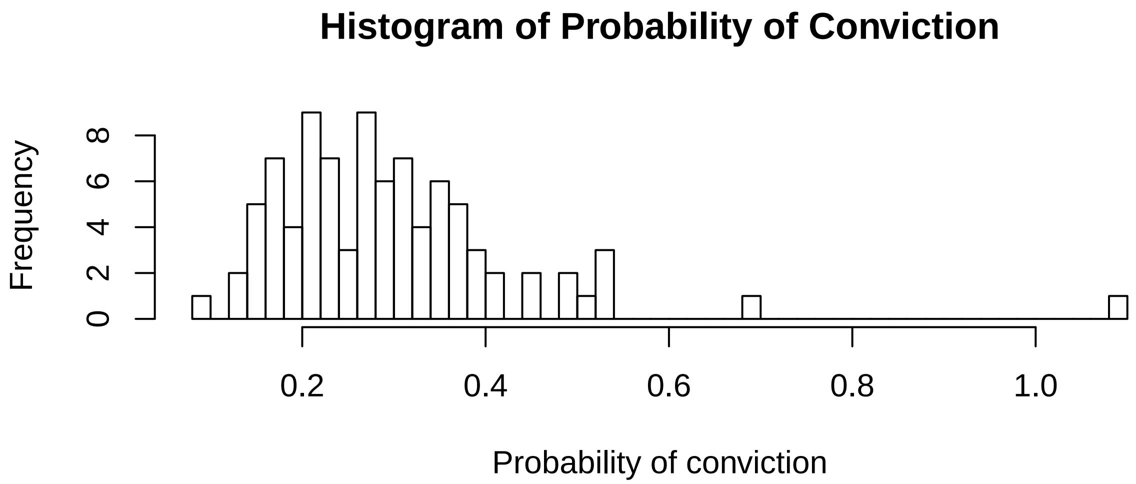
Probconv seems to have values greater than 1.0, which is unexpected given that probability values are supposed to be between 0.0 to 1.0. The variable description does not seem to indicate how the variable is calculated. After discussing this in the lectures and office hours, we will assume that probconv values above 1.0 is okay, and that the higher the value, the higher the probabilty of conviction.

The histogram of probconv does not seem to look very normal. That said, taking the log of a probabilty does not seem to make sense from an interpretability perspective. For example, an "increase is 10% of probability" is not intuitively interpretable. As a result, for probconv no log transformation will be applied. It also does not look like other transforms would be suitable for this variable.

```
summary(data$probconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.09277 0.20495 0.27146 0.29524 0.34487 1.09091
```

```r
hist(data$probconv, breaks=50,
     main="Histogram of Probability of Conviction",
     xlab="Probability of conviction")
```



Probability of Prison Sentence (probsen)

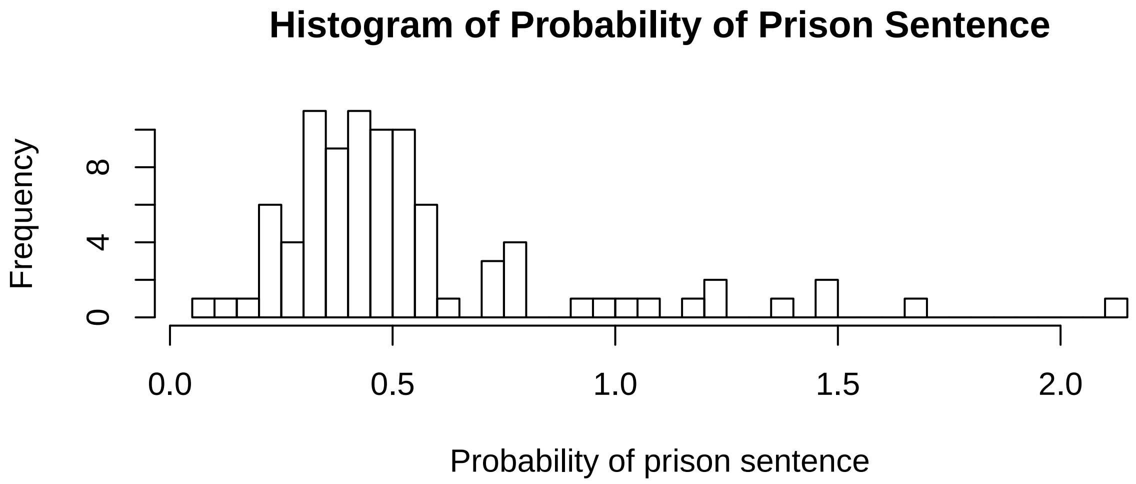
Similarly, probsen seems to have values greater than 1.0, which is unexpected given that probability values are supposed to be between 0.0 to 1.0. After discussing this in the lectures and office hours, we will assume that probsen values above 1.0 is okay, and that the higher the value, the higher the probabilty of prison sentence.

Also, similar to probconv, for probsen no log transformation will be applied. It also does not look like other transforms would be suitable for this variable.

```r
summary(data$probsen)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.06838 0.34422 0.45170 0.55086 0.58513 2.12121
```

```r
hist(data$probsen, breaks=50,
     main="Histogram of Probability of Prison Sentence",
     xlab="Probability of prison sentence")
```



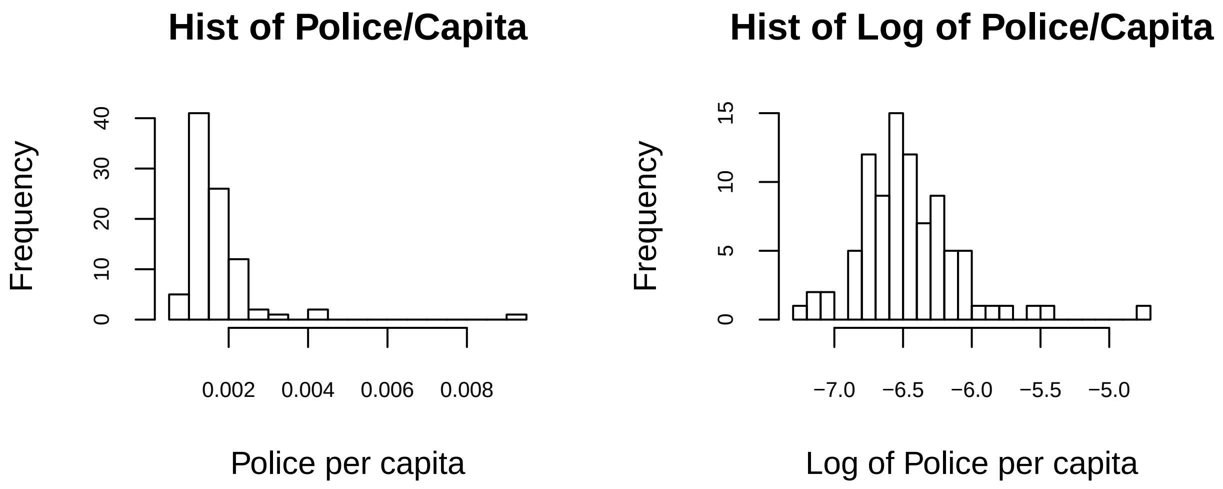
Police per Capita (police)

Looking at the histogram of police, the distribution tends to be right skewed. Taking the log of police tends to make the histgram appear more normal. As a result, for our modeling, we will proceed with using log of police.

```
summary(data$police)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## 0.0007459 0.0012378 0.0014897 0.0017080 0.0018856 0.0090543
```

```
par(mfrow=c(1,2))
hist(data$police, breaks=20,
     main="Hist of Police/Capita",
     xlab="Police per capita", cex.axis = 0.7)
hist(log(data$police), breaks=20,
     main="Hist of Log of Police/Capita",
     xlab="Log of Police per capita", cex.axis = 0.7)
```



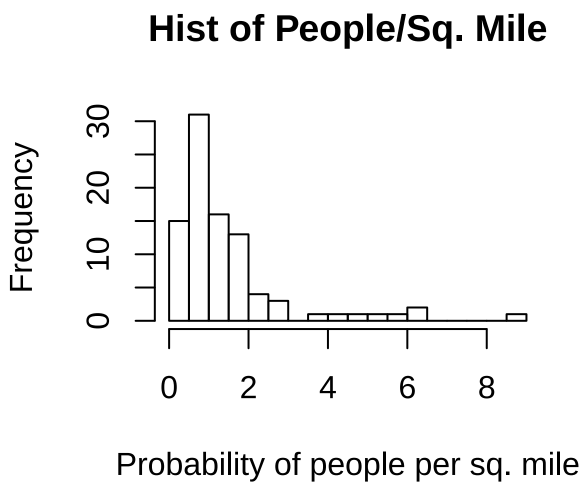
People per Sq. Mile (density)

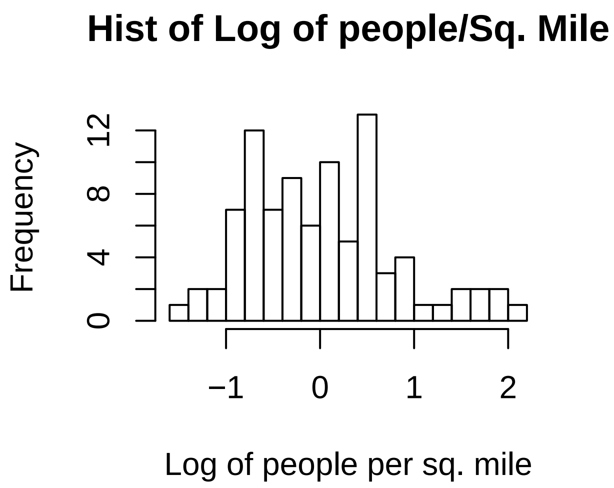
Looking at the histogram of density, the distribution tends to be right skewed. Taking the log of density tends to make the histgram appear more normal. As a result, for our modeling, we will proceed with using log of density.

```
summary(data$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.2034  0.5472  0.9792  1.4379  1.5693  8.8277
```

```
par(mfrow=c(1,2))
hist(data$density, breaks=20,
     main="Hist of People/Sq. Mile",
     xlab="Probability of people per sq. mile")
hist(log(data$density), breaks=20,
     main="Hist of Log of people/Sq. Mile",
     xlab="Log of people per sq. mile")
```





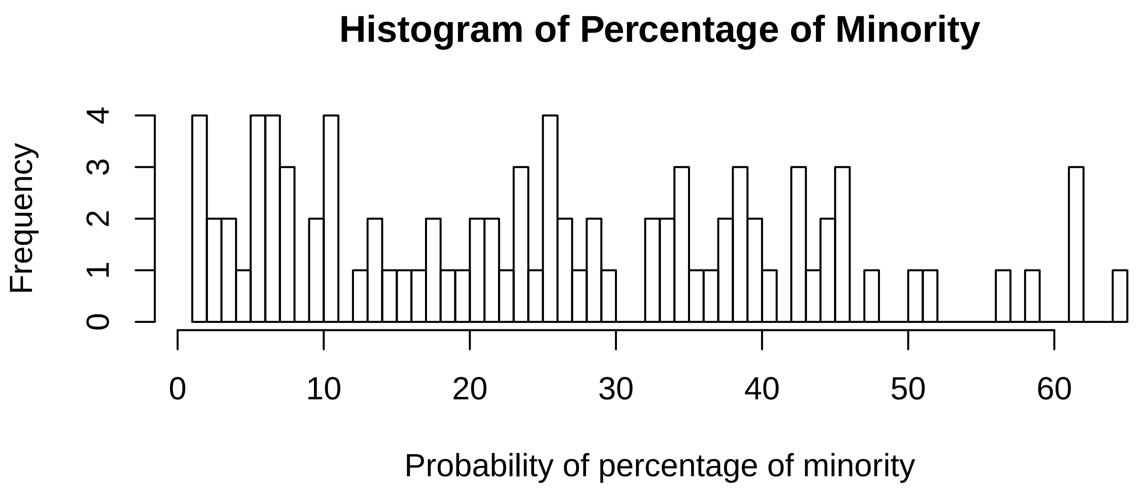
Percentage of Minority, 1980 (pctmin)

The data and histogram for pctmin looks expected. However, the data is between 0 - 100 whereas other percentage variables is between 0.0 and 1.0; We may want to transform this variable to keep things consistent from an interpretability perspective. This will be taken care of in the data transformation section.

```
summary(data$pctmin)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.284  10.024  24.852  25.713  38.183  64.348
```

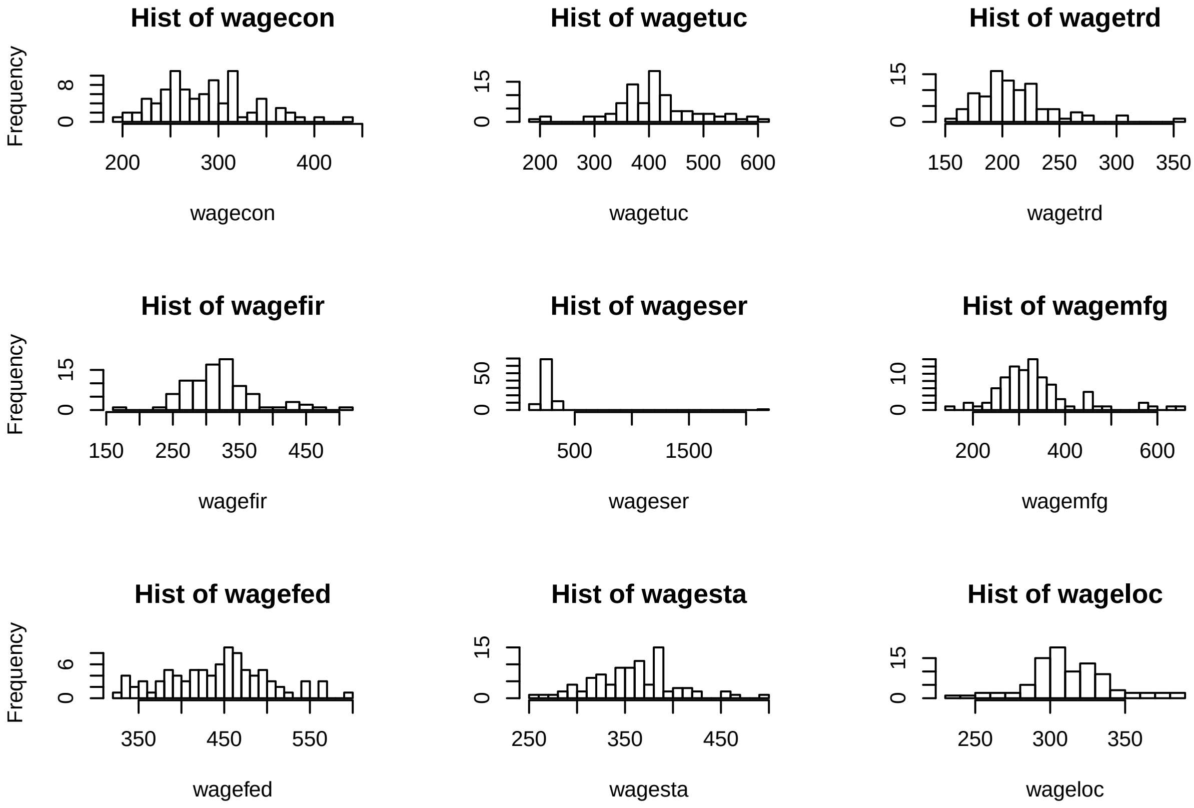
```
hist(data$pctmin, breaks=50,
     main="Histogram of Percentage of Minority",
     xlab="Probability of percentage of minority")
```



Wage variables (wage*)

Looking at the histogram of the wage variables, there is no obvious case for supporting any transforms on the variables, since for most cases the distribution does not look very skewed.

```
par(mfrow=c(3,3))
hist(data$wagecon, breaks=20, main="Hist of wagecon", xlab="wagecon", ylab="Frequency")
hist(data$wagetuc, breaks=20, main="Hist of wagetuc", xlab="wagetuc", ylab="")
hist(data$wagetrd, breaks=20, main="Hist of wagetrd", xlab="wagetrd", ylab="")
hist(data$wagefir, breaks=20, main="Hist of wagefir", xlab="wagefir", ylab="Frequency")
hist(data$wageser, breaks=20, main="Hist of wageser", xlab="wageser", ylab="")
hist(data$wagemfg, breaks=20, main="Hist of wagemfg", xlab="wagemfg", ylab="")
hist(data$wagefed, breaks=20, main="Hist of wagefed", xlab="wagefed", ylab="Frequency")
hist(data$wagesta, breaks=20, main="Hist of wagesta", xlab="wagesta", ylab="")
hist(data$wageloc, breaks=20, main="Hist of wageloc", xlab="wageloc", ylab="")
```



Focusing on wageser, there seems to be one data point that looks to be an extreme outlier. Looking further, it seems to be coming from data point 84. This data point has an extremely high value for probsen and wageser. For our model, we will remove this datapoint, as we have found that this data point tend to have high Cook's distance if we were to include it in our models.

```
data[data$X == 84,c("probsen", "wageser")]
```

```
##    probsen  wageser
## 84 2.12121 2177.068
```

New Variables (X, Y, Z)

MORE HERE LATER

Transformed and Filtered Dataset

Based on the univariate analysis, the following transformations were proposed, as well as removal of one data point per analysis in the previous section.

```
data$log_crime = log(data$crime)
data$tot_wages = (data$wagecon + data$wagetuc + data$wagetrd + data$wagefir + data$wageser + data$wagemfg + d
data$log_police = log(data$police)
data$log_density = log(data$density)
data$log_wagedensity = log(data$tot_wages/data$density)
data$pctmin = data$pctmin/100
data = data[data$X != 84,]

sort(colnames(data))
```

```
##  [1] "avgsen"          "central"         "county"
##  [4] "crime"           "density"         "log_crime"
##  [7] "log_density"     "log_police"      "log_wagedensity"
## [10] "mix"             "pctmin"          "police"
## [13] "probarr"         "probconv"        "probsen"
## [16] "tax"             "tot_wages"       "urban"
## [19] "wagecon"         "wagefed"         "wagefir"
## [22] "wageloc"         "wagemfg"         "wageser"
## [25] "wagesta"         "wagetrd"         "wagetuc"
## [28] "west"            "X"               "year"
## [31] "ymale"
```

Bi-variate analysis

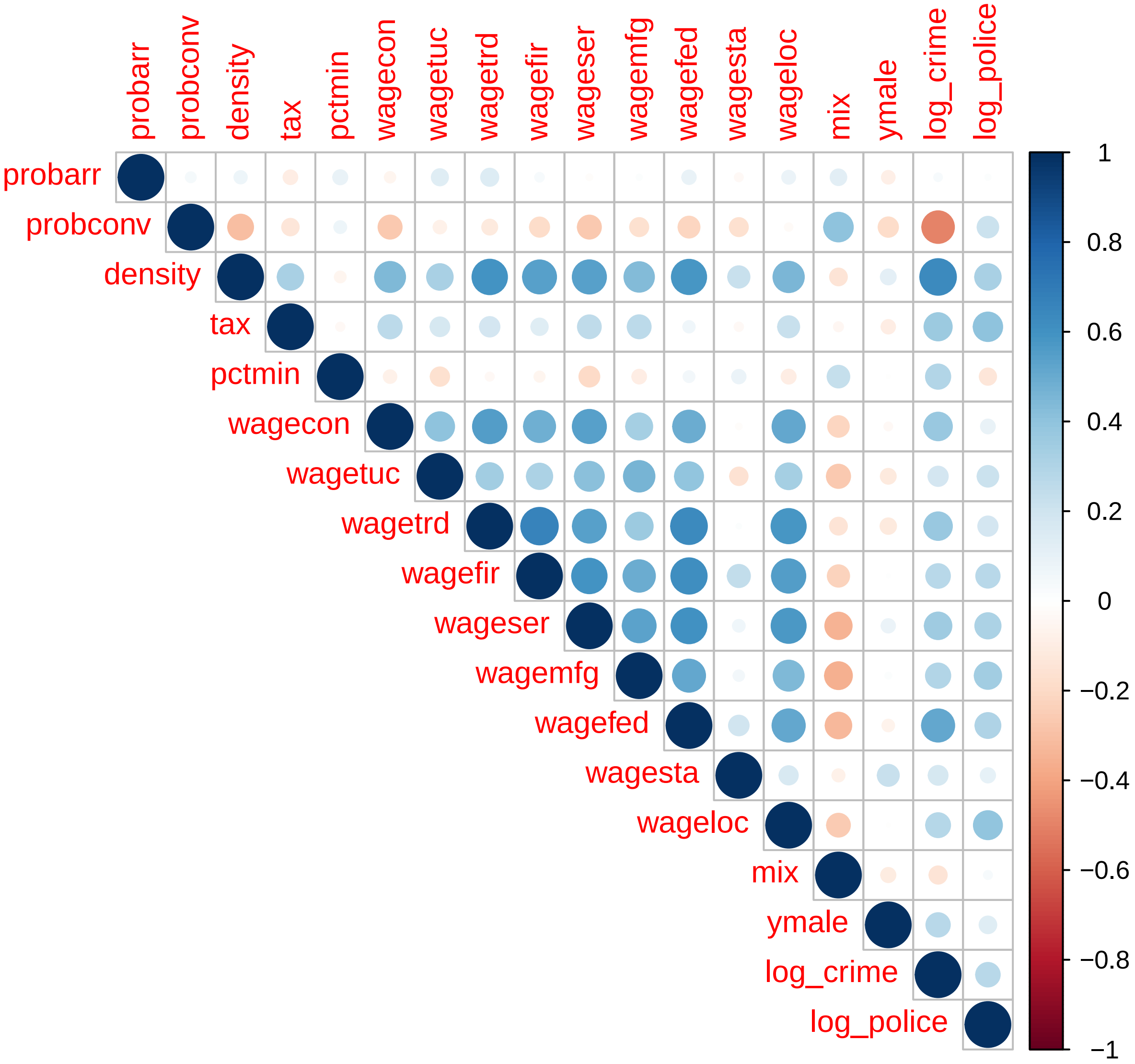
Looking at the correlation plot of the non-indicator variables, it looks like the following variables correlate with log of crime (last column in the plot) highly and positively:

1. density - This makes sense since the crime rate tends to increase in more populated areas.
2. tax - This makes sense since crime rate could increase in areas where there are more tax revenues collected.
3. wage variables - This somewhat makes sense since as wages go up, there may be higher likelihood for crime.

It looks like the following variables correlate with log of crime (second-last column in the plot) highly and negatively:

1. probconv - Makes sense since if probability of convictions go down, then there are more (potential) criminal out on the streets.

```
corrplot(cor(data[ , (names(data) %in%
                        c("log_crime", "probarr", "probconv", "probpris", "tax",
                          "log_police", "density", "pctmin", "ymale","taxpc",
                          "wagecon", "wagetuc", "wagetrd", "wagefir", "wageser", "wagemfg",
                          "wagefed", "wagesta", "wageloc", "mix"))]),
                                 method="circle", type="upper")
```



Section 3: Model Building

TBD

Section 4: Model Assumptions

TBD

Section 5: Model Specifications

TBD

Section 6: Model Summary

TBD

Section 7: Causality

TBD

Section 8: Conclusion

TBD