Lab 3

David Hou, Scott Hungerfield, Irene Seo

March 20, 2018

Introduction

The purpose of this study is to provide information for political campaign in North Carolina. Specifically, we want to determine what variables contribute to crime rate and help the campaign propose policy suggestions to local governments. To accomplish this, we were given crime data from several North Carolina counties along with other variables. We will run ordinary least square regressions to help determine which of these are the best predictors of crime.

Data Cleaning

First we need to clean the data. In the raw data, we notice that that the last 6 rows are empty. The integer columns are probably more useful to us as factors. The prbconv is coded as a factor, so we turn it into a numeric.

We also notice that prbarr and prbconv have values that are greater than 1, which does not make much sense because they are probability variables. We assume that these values were coded incorrectly and filter those out.

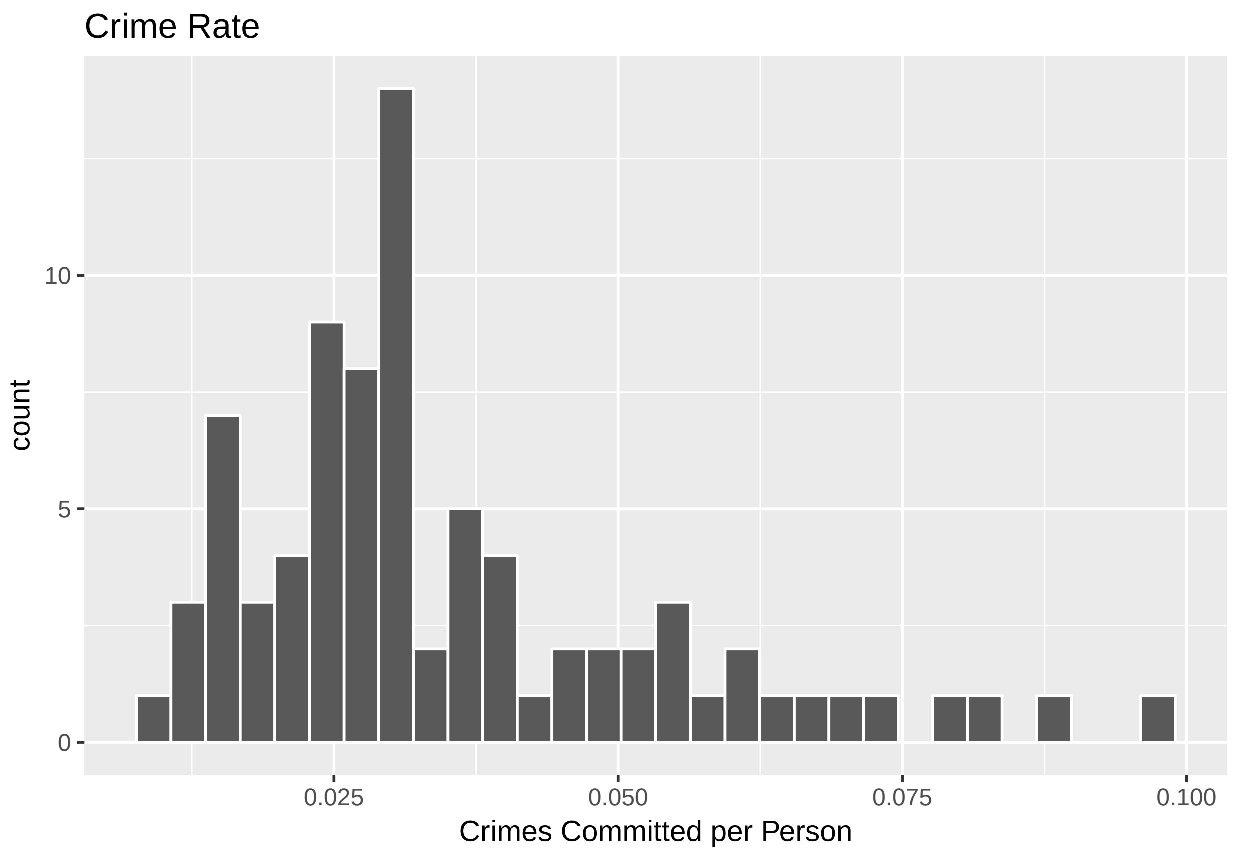
```r
raw = as_tibble(read.csv('crime_v2.csv'))
t = raw %>%
    filter(!is.na(county)) %>%
    mutate(prbconv = as.numeric(as.character(prbconv))) %>%
    mutate_if(is.integer, as.factor) %>%
    filter(prbarr < 1 & prbconv < 1)
levels(t$west) = c('East', 'West')
t$west = relevel(t$west, 'West') # Put West first so it appears on the left on facet plots
levels(t$central) = c('Outer', 'Central')
levels(t$urban) = c('Non-urban', 'Urban')
```

We also do not see an advantage to analyzing each wage individually. Thus, we create a new column that is the sum of all the wage columns.

```r
t = t %>% mutate(wage = wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc)
```

Examining Key Variables of Interest

```r
qplot(t$crmrte, col = I('white')) +
    labs(title = 'Crime Rate', x = 'Crimes Committed per Person')
```

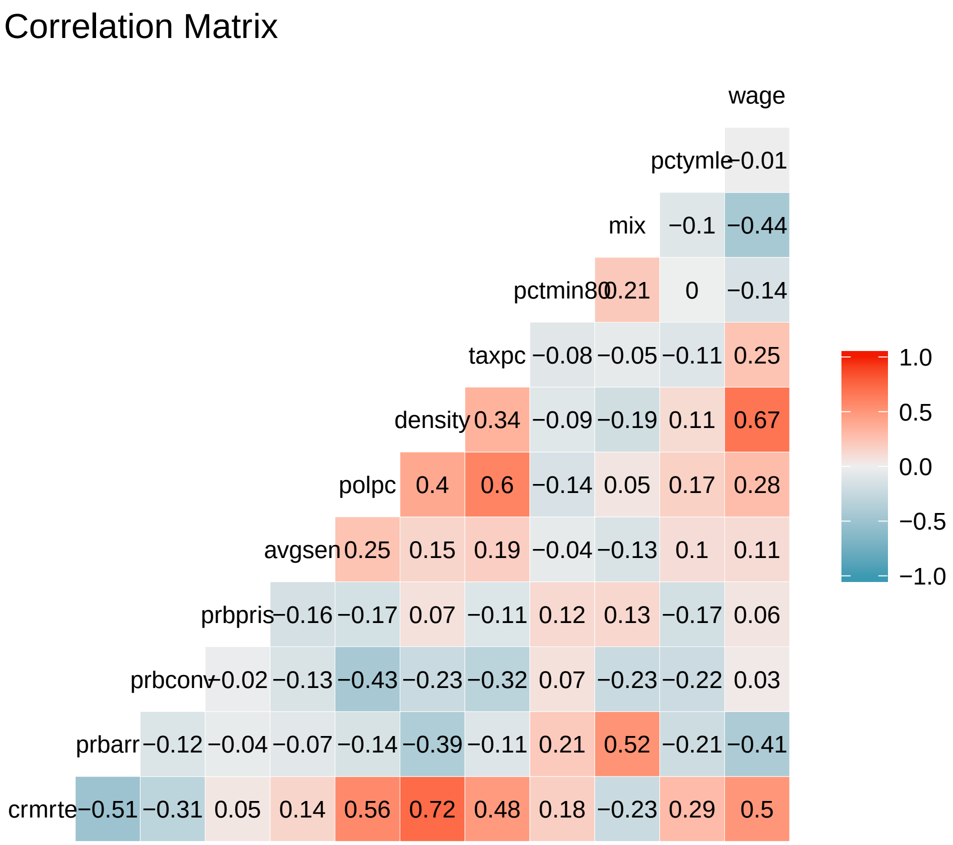


```
summary(t$crmrte)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.01062 0.02337 0.03043 0.03536 0.04374 0.09897
```

We see that the main variable of interest, crime rate, has some positive skew, but does not seem to have a very exotic distribution. To determine which variables are of interest to us when predicting crime rate, we look at the correlation matrix among the variables.

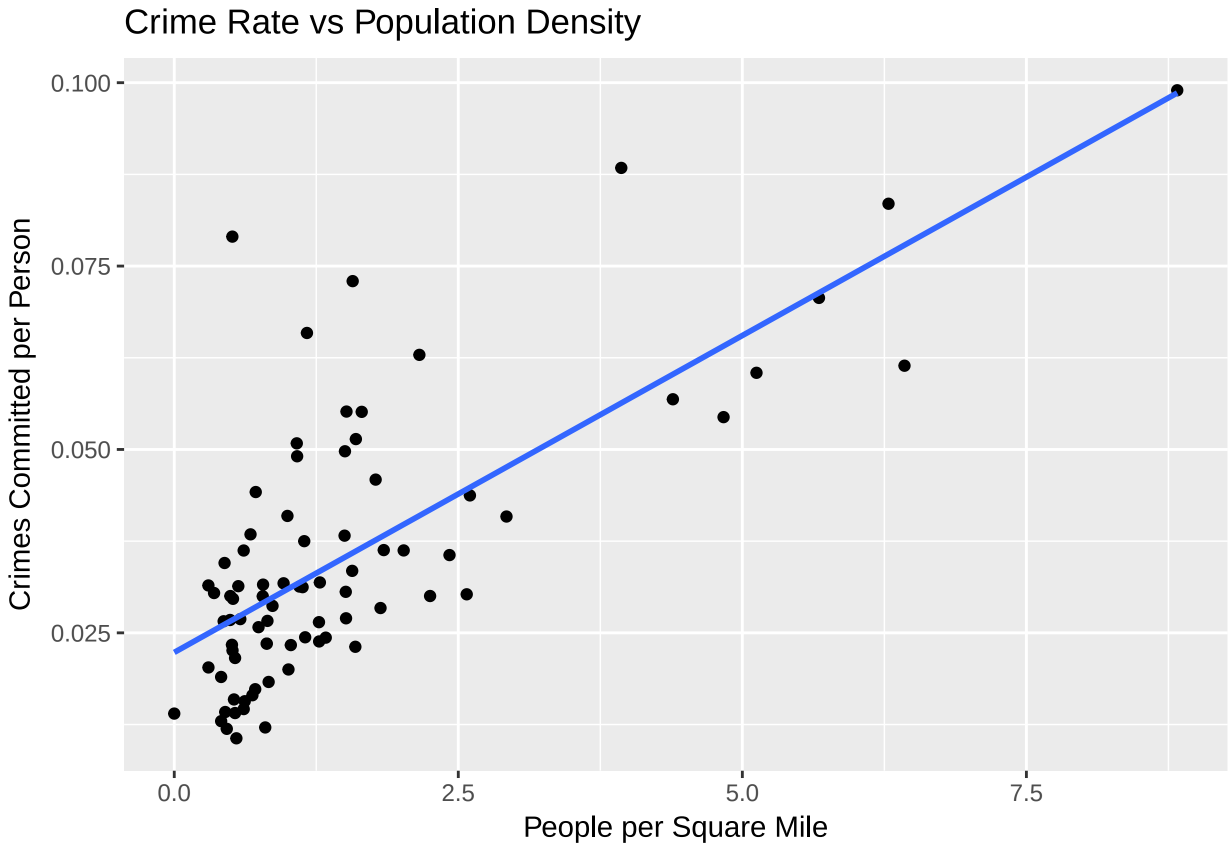
```
t2 = t %>% select(crmrte, prbarr, prbconv, prbpris, avgsen, polpc, density, taxpc, pctmin80, mix, pctyml
ggcorr(t2, label = TRUE, label_round = 2, label_size = 3, size = 3) + ggtitle('Correlation Matrix')
```



	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	pctmin80	mix	pctymle	wage
wage											
pctymle											−0.01
mix										−0.1	−0.44
pctmin80									0.21	0	−0.14
taxpc								−0.08	−0.05	−0.11	0.25
density							0.34	−0.09	−0.19	0.11	0.67
polpc						0.4	0.6	−0.14	0.05	0.17	0.28
avgsen					0.25	0.15	0.19	−0.04	−0.13	0.1	0.11
prbpris				−0.16	−0.17	0.07	−0.11	0.12	0.13	−0.17	0.06
prbconv			−0.02	−0.13	−0.43	−0.23	−0.32	0.07	−0.23	−0.22	0.03
prbarr		−0.12	−0.04	−0.07	−0.14	−0.39	−0.11	0.21	0.52	−0.21	−0.41
crmrte	−0.51	−0.31	0.05	0.14	0.56	0.72	0.48	0.18	−0.23	0.29	0.5

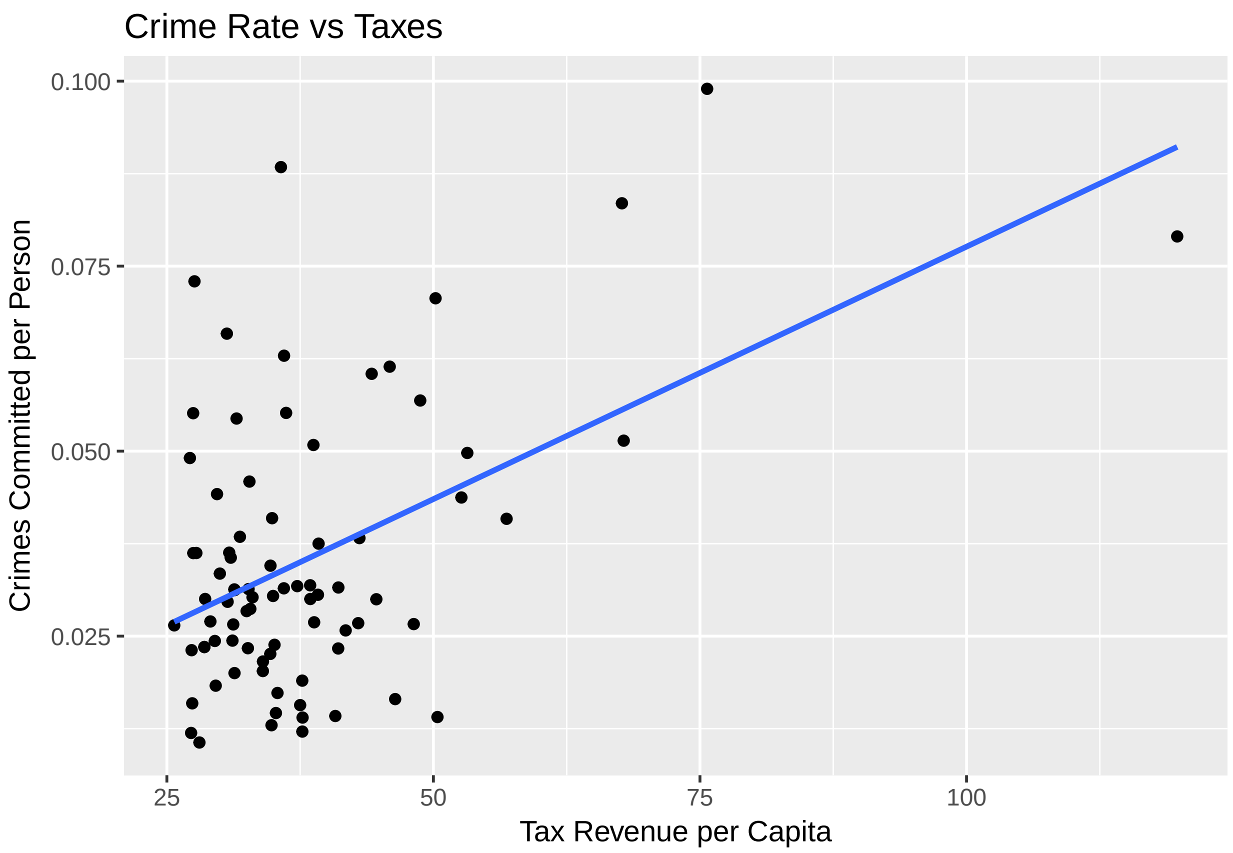
From the correlation matrix, we see that population density stands out as being highly correlated with crime rate (r = 0.72). This variable looks like a good candidate as a causal predictor for crime rate. One explanation could be that as more people move into an area, the increased number of interactions give opportunity for more crime.

```
qplot(t$density, t$crmrte) +
    labs(title = 'Crime Rate vs Population Density', x = 'People per Square Mile', y = 'Crimes Committe
    geom_smooth(method = 'lm', se = FALSE)
```

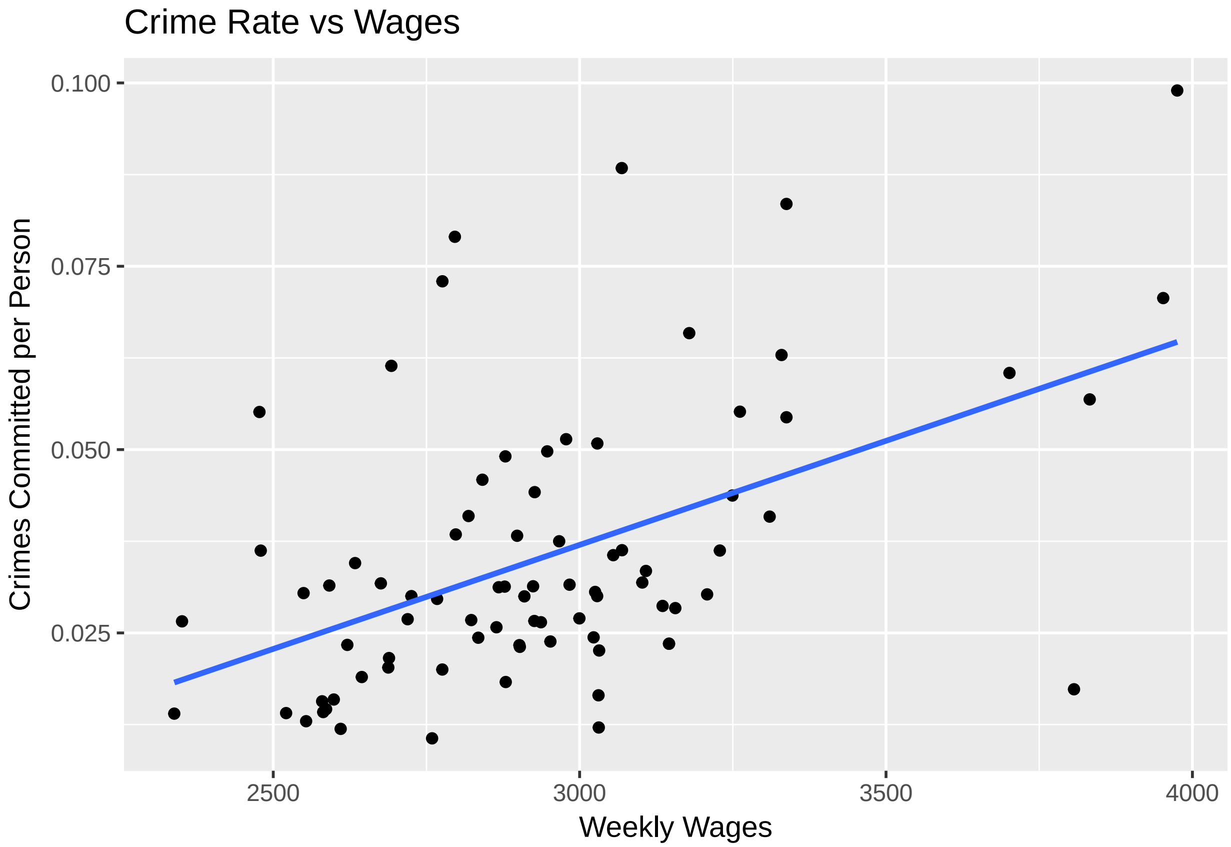


The other two variables with moderately positive correlation are tax per capita (r = 0.48) and wages (r = 0.5). It is interesting to note that taxes and wages are not very correlated with themselves (r = 0.25). This finding is surprising, as one would expect that wages and taxes would go up very closely with each other. Also note that population density is weakly correlated with taxes (r = 0.34) and moderately correlated with wages (r = 0.67). We believe that taxes and wages are not directly causing higher crime rates but could be good indirect indicators.

```
qplot(t$taxpc, t$crmrte) +
    labs(title = 'Crime Rate vs Taxes', x = 'Tax Revenue per Capita', y = 'Crimes Committed per Person'
    geom_smooth(method = 'lm', se = FALSE)
```

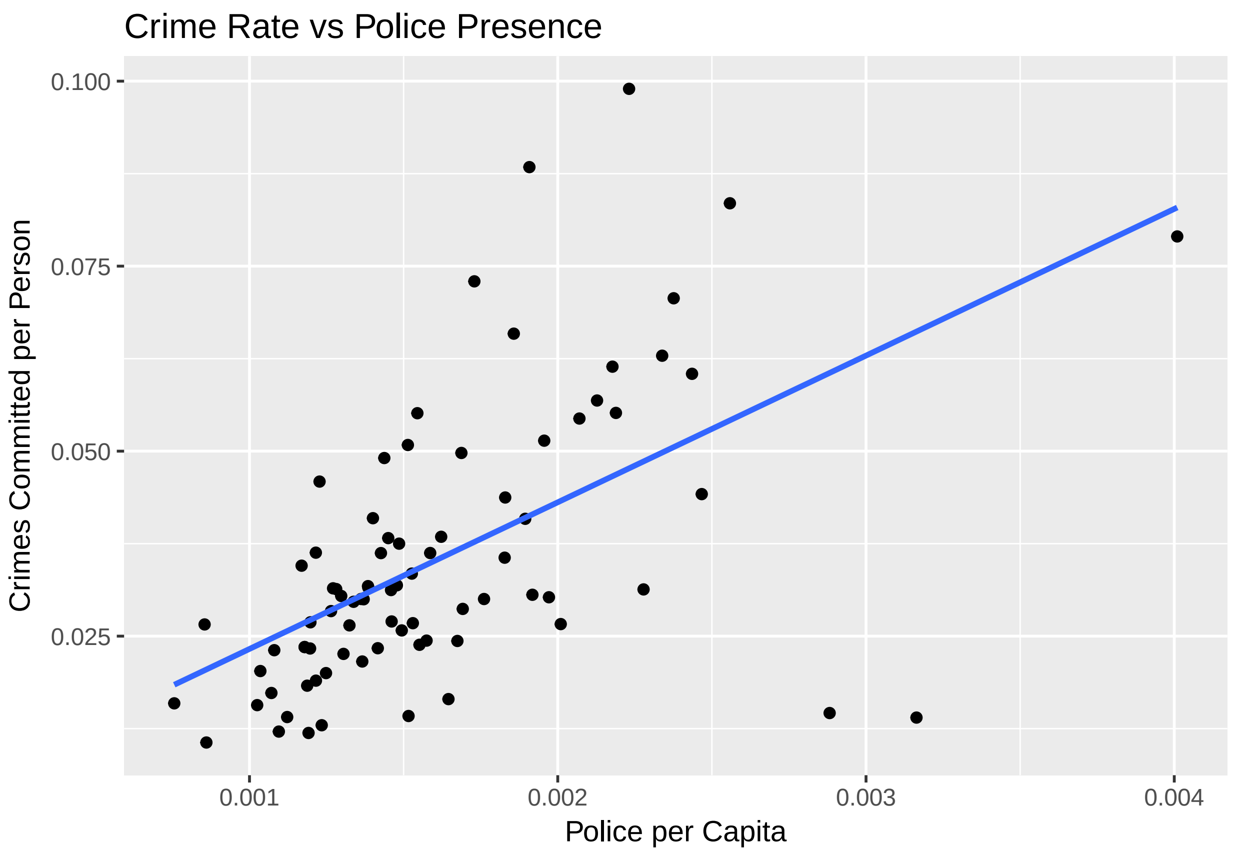


```r
qplot(t$wage, t$crmrte) +
    labs(title = 'Crime Rate vs Wages', x = 'Weekly Wages', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
```



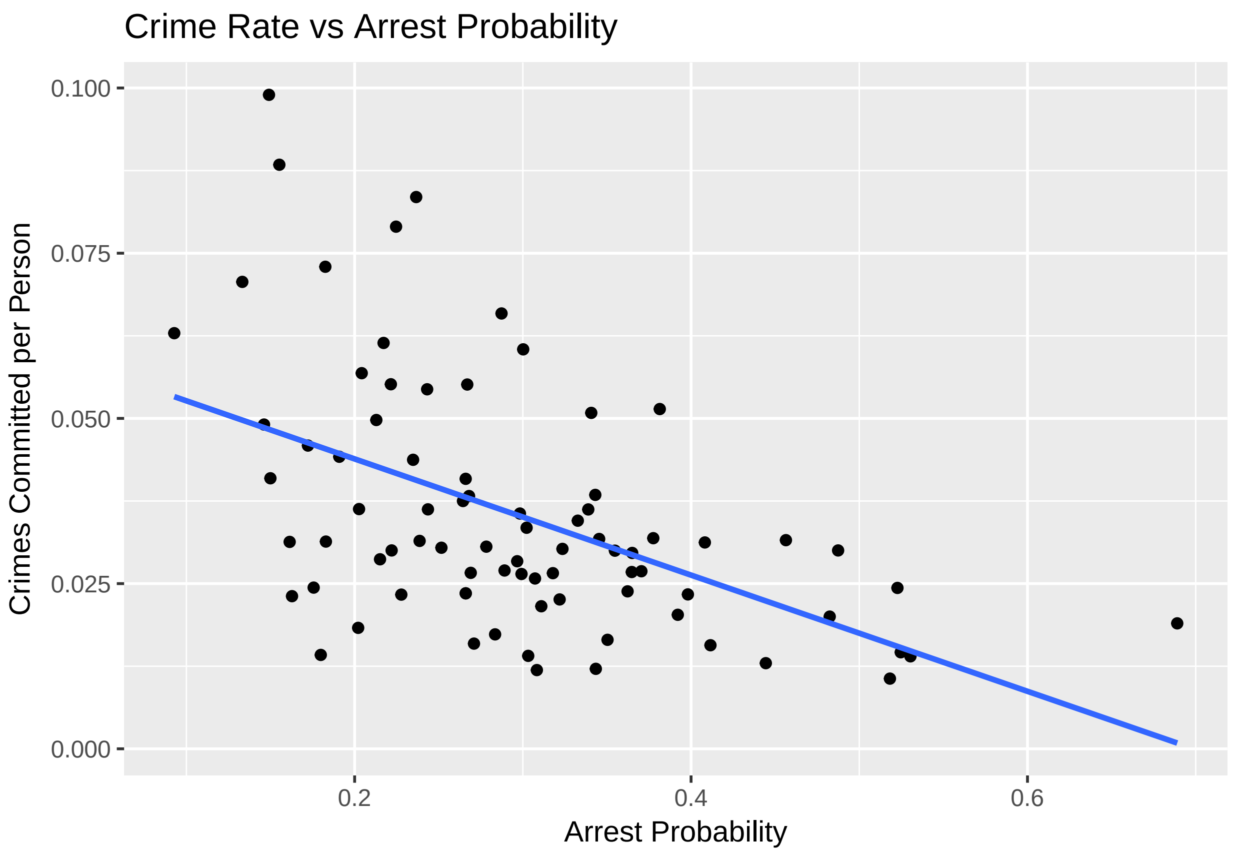
Interestingly, the relationship between police per capita and crime rate is positive and moderately large (r = 0.56). This means that either increasing police presence makes crime rate worse or that crime is causing an increase in police presence rather than vice versa. The latter explanation seems much more logical.

```
qplot(t$polpc, t$crmrte) +
    labs(title = 'Crime Rate vs Police Presence', x = 'Police per Capita', y = 'Crimes Committed per Pe
    geom_smooth(method = 'lm', se = FALSE)
```

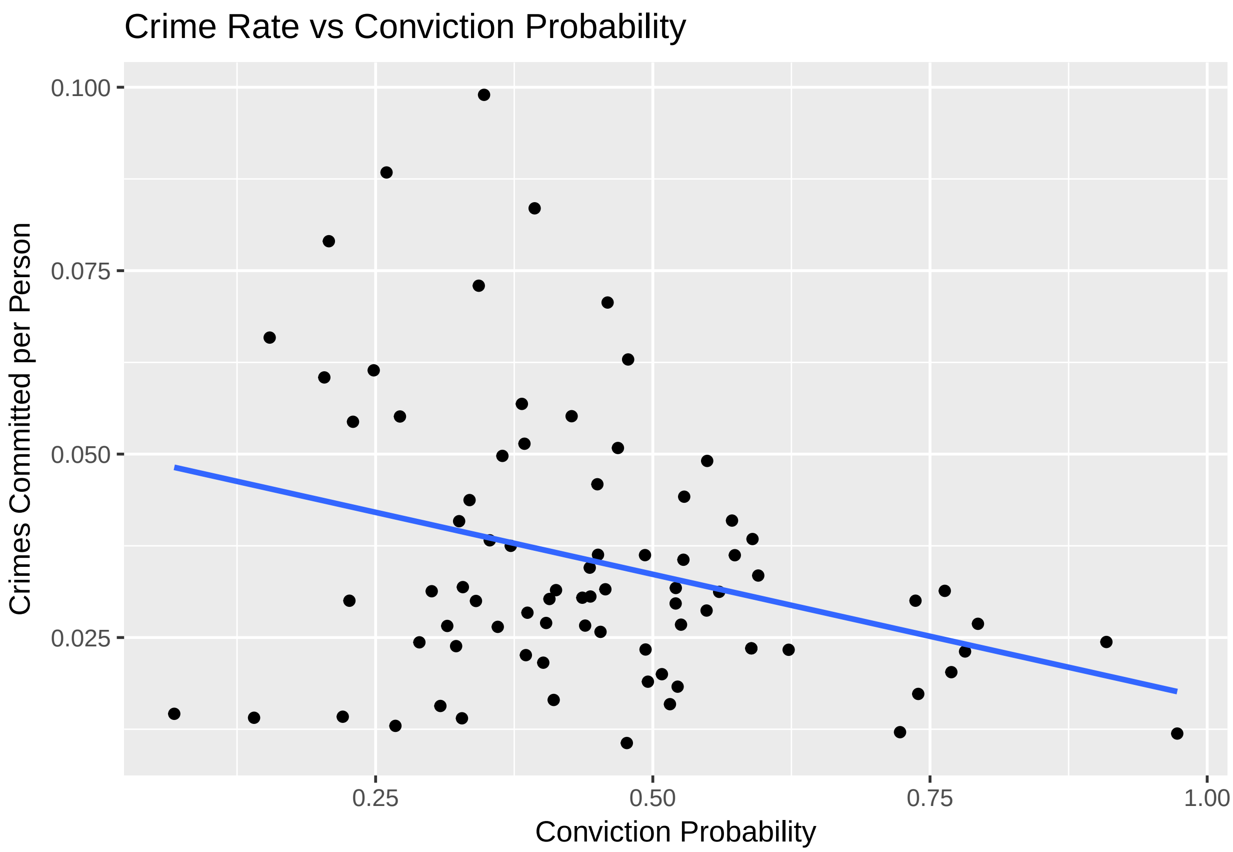


Of the three "certainty of punishment" variables, it looks like arrest probability has a moderate effect (r = -0.51) and conviction probability has a weak effect (r = -0.31), but probability of prison sentence has almost no effect (r = 0.05). It is important to note that these three probabilities seem uncorrelated with one another, so we will include multiple ones in our regression without fear of multicolinearity. The "severity of punishment" variable, average prison sentence length, does not seem to be correlated with crime rate (r = 0.14).

```
qplot(t$prbarr, t$crmrte) +
    labs(title = 'Crime Rate vs Arrest Probability', x = 'Arrest Probability', y = 'Crimes Committed pe
    geom_smooth(method = 'lm', se = FALSE)
```



```
qplot(t$prbconv, t$crmrte) +
    labs(title = 'Crime Rate vs Conviction Probability', x = 'Conviction Probability', y = 'Crimes Comm
    geom_smooth(method = 'lm', se = FALSE)
```



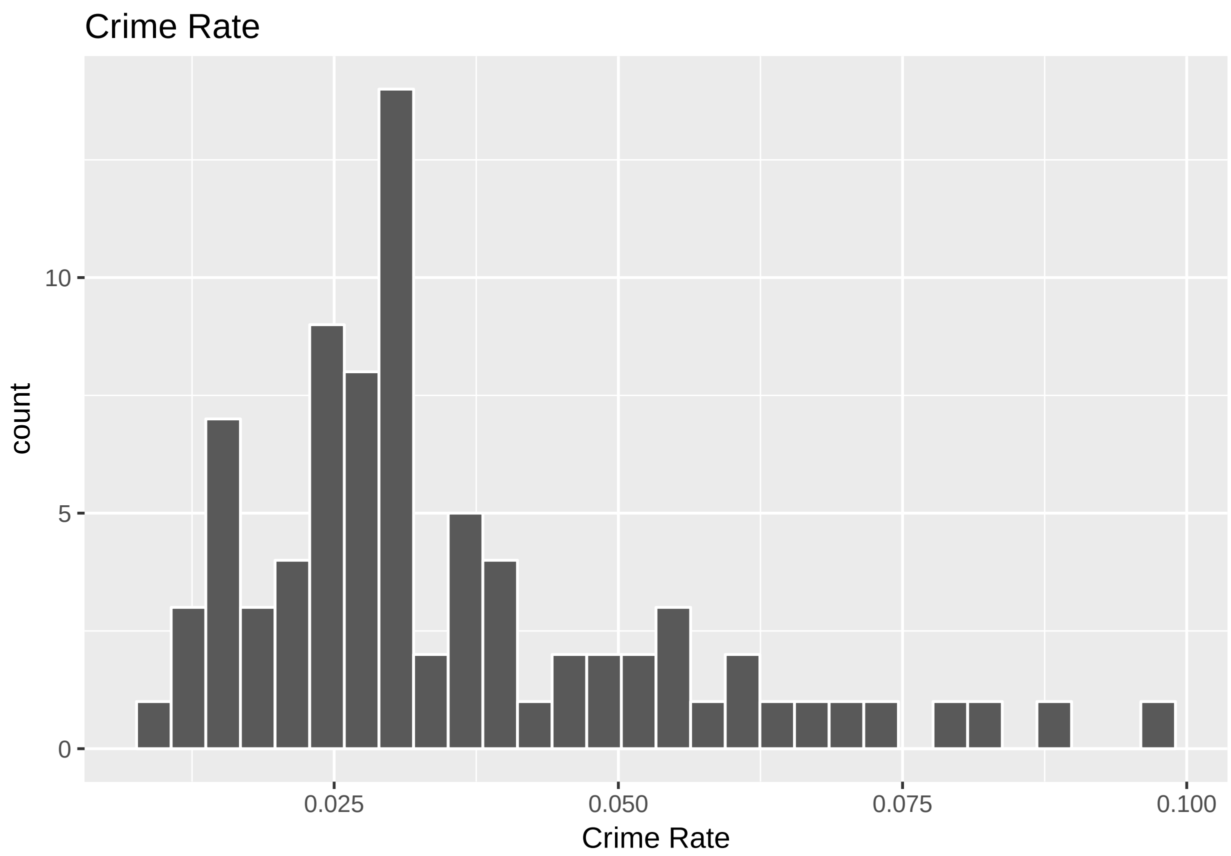
Model Building

```
m1 = lm(t$crmrte ~ t$density)
m2 = lm(t$crmrte ~ t$density + t$prbarr + t$prbconv)
m3 = lm(t$crmrte ~ t$density + t$prbarr + t$prbconv + t$taxpc + t$wage)
m4 = lm(t$crmrte ~ t$prbarr + t$prbconv + t$prbpris + t$avgsen + t$density + t$taxpc + t$pctmin80 + t$mi
m5 = lm(t$crmrte ~ t$prbarr + t$prbconv + t$density + t$taxpc + t$pctmin80 + t$pctymle)
stargazer(m1, m2, m3, m4, m5, type = 'text')
```

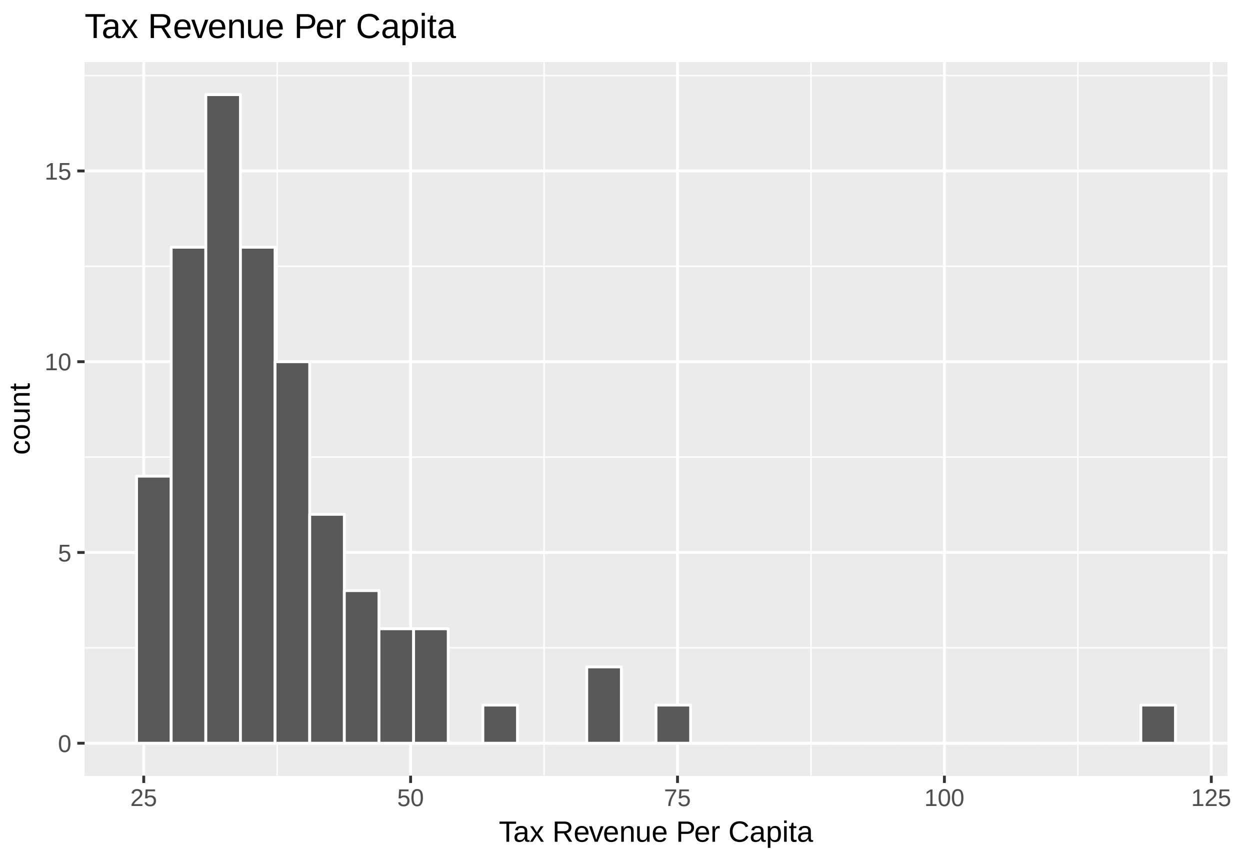
```
##
## ====================================================================================================
##                                                                     Dependent variable:
##                     --------------------------------------------------------------------------------
##                                                                           crmrte
##                               (1)                    (2)                   (3)                     (4
## ----------------------------------------------------------------------------------------------------
## density                     0.009***               0.007***               0.006***               0.005
##                             (0.001)                (0.001)                (0.001)                 (0.0
##
## prbarr                                            -0.055***              -0.054***               -0.05
##                                                    (0.014)                (0.013)                 (0.0
##
## prbconv                                           -0.024***               -0.017**               -0.01
##                                                    (0.008)                (0.008)                 (0.0
##
## prbpris                                                                                           0.0
##                                                                                                  (0.0
##
## avgsen                                                                                           -0.0
##                                                                                                 (0.00
##
## taxpc                                                                   0.0003***                0.000
##                                                                          (0.0001)                (0.00
##
## pctmin80                                                                                        0.000
##                                                                                                 (0.00
##
## mix                                                                                              -0.0
##                                                                                                  (0.0
##
## pctymle                                                                                          0.140
##                                                                                                  (0.0
##
## wage                                                                     -0.00000                 0.00
##                                                                         (0.00001)               (0.00
##
## Constant                   0.022***               0.053***               0.040**                  0.0
##                            (0.002)                (0.007)                (0.016)                 (0.0
##
## -----------------------------------------------------------------------------------------------------
## Observations                  81                     81                     81                     81
## R2                          0.525                  0.626                  0.670                   0.8
## Adjusted R2                 0.519                  0.612                  0.648                   0.7
## Residual Std. Error    0.013 (df = 79)        0.012 (df = 77)        0.011 (df = 75)         0.009 (
## F Statistic         87.187*** (df = 1; 79) 43.051*** (df = 3; 77) 30.502*** (df = 5; 75) 28.861*** (
## =====================================================================================================
## Note:
```

Here is some single variate EDA.

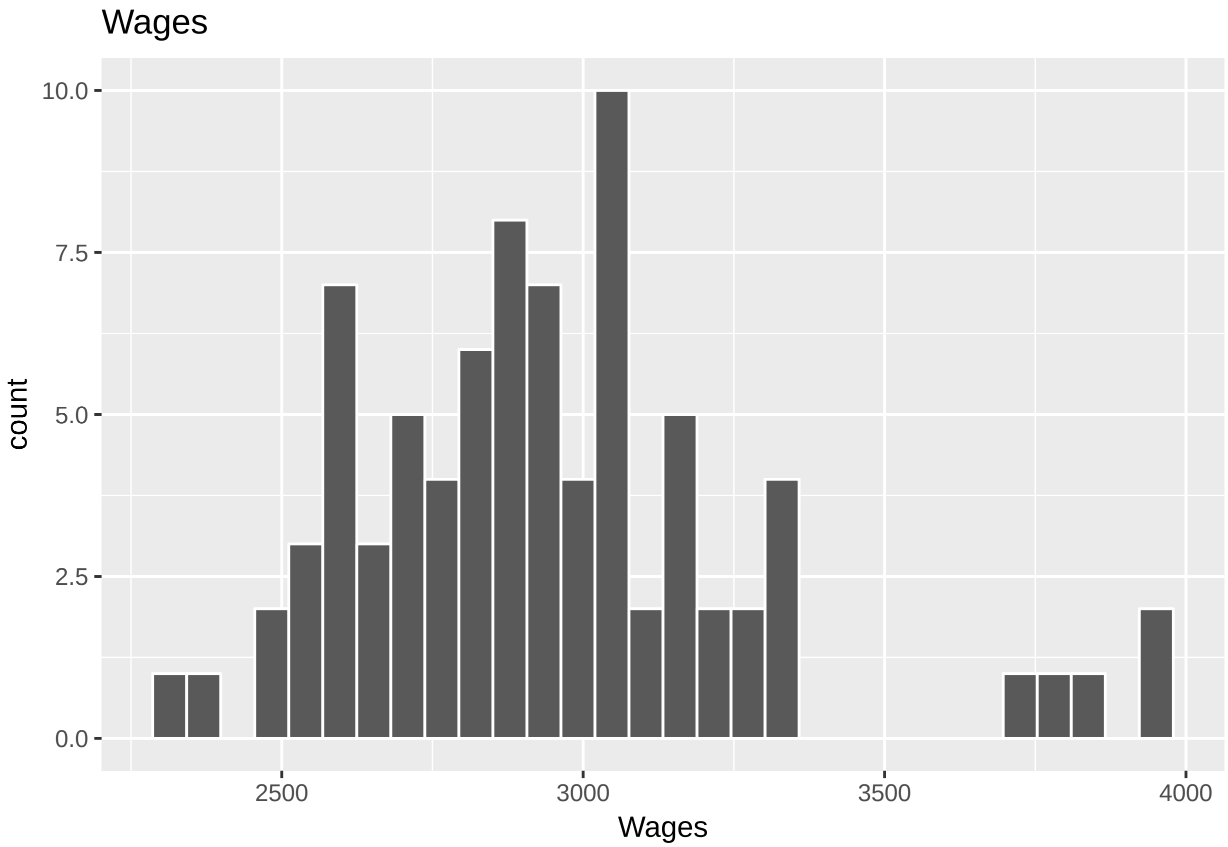
```
qplot(t$crmrte, geom = 'histogram', col = I('white'), main = 'Crime Rate', xlab = 'Crime Rate')
```



```
qplot(t$taxpc, geom = 'histogram', col = I('white'), main = 'Tax Revenue Per Capita', xlab = 'Tax Revenu
```

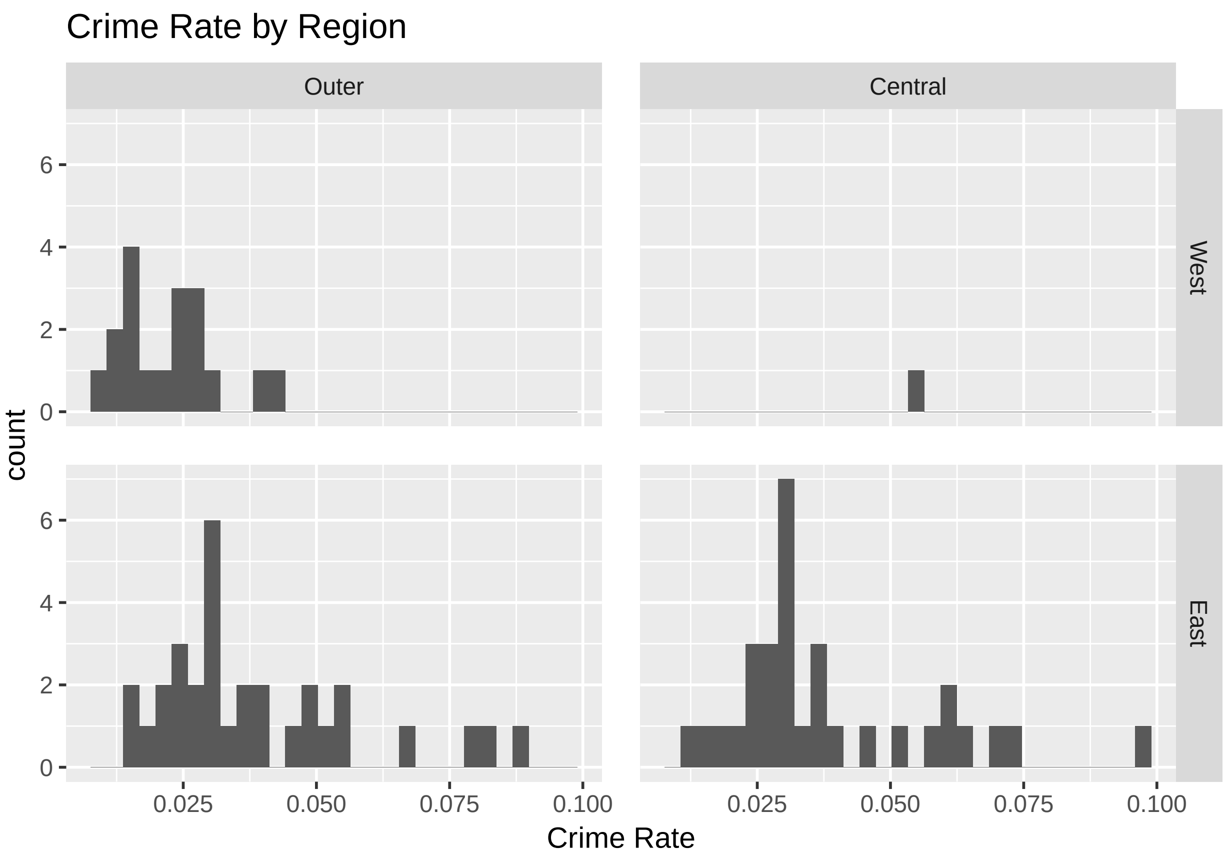


```r
qplot(t$wage, geom = 'histogram', col = I('white'), main = 'Wages', xlab = 'Wages')
```

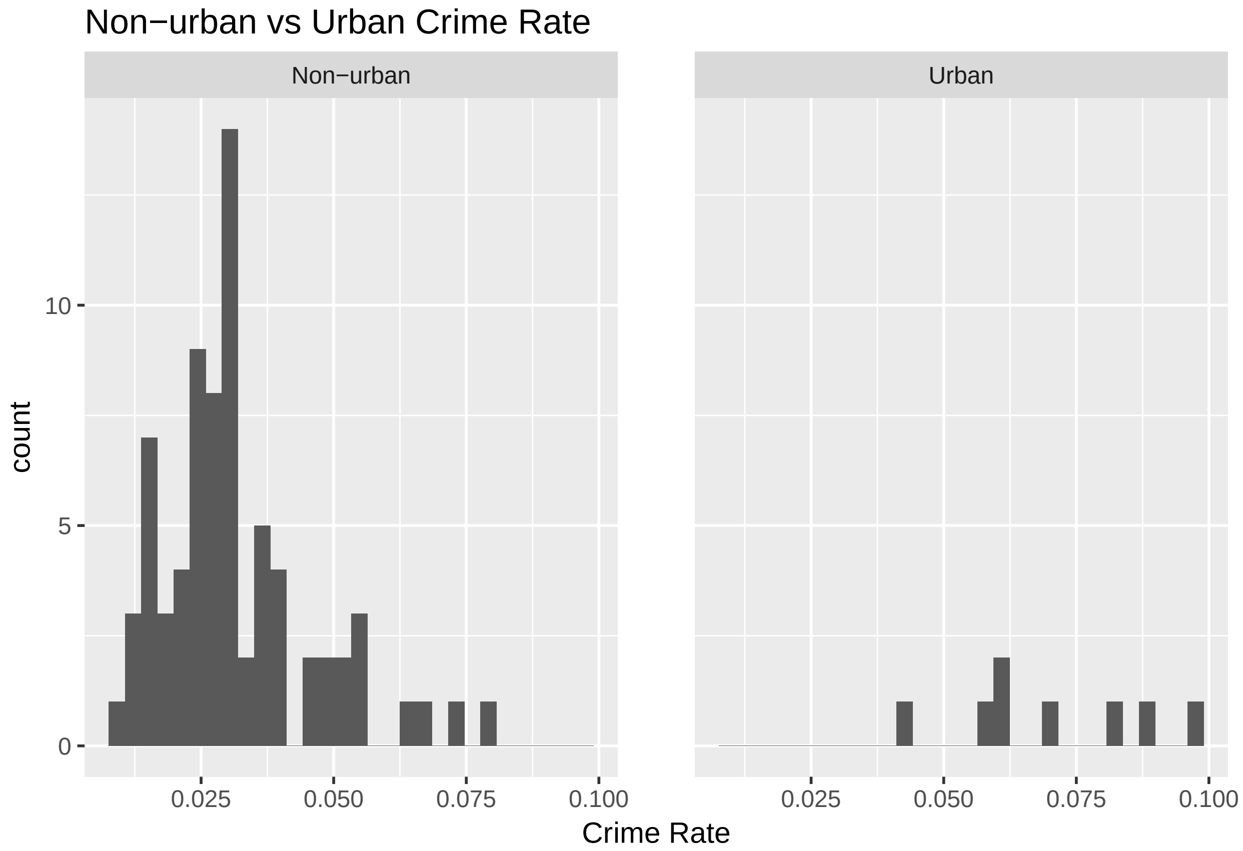


Here are some facet plots.

```
ggplot(t, aes(crmrte)) +
    geom_histogram() +
    facet_grid(west ~ central) +
    theme(panel.spacing = unit(1, "lines")) +
    labs(title = 'Crime Rate by Region', x = 'Crime Rate')
```

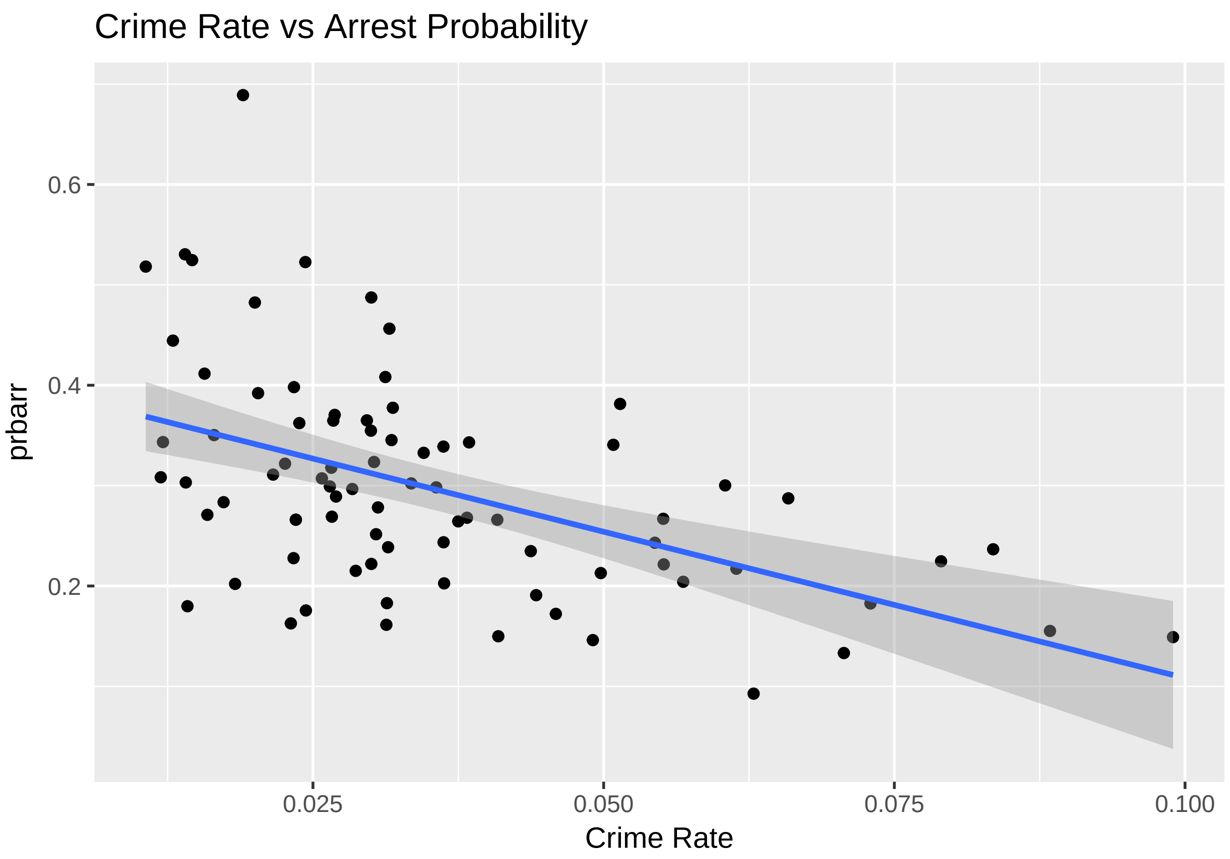


```r
ggplot(t, aes(crmrte)) +
    geom_histogram() +
    facet_grid(. ~ urban) +
    theme(panel.spacing = unit(2, "lines")) +
    labs(title = 'Non-urban vs Urban Crime Rate', x = 'Crime Rate')
```

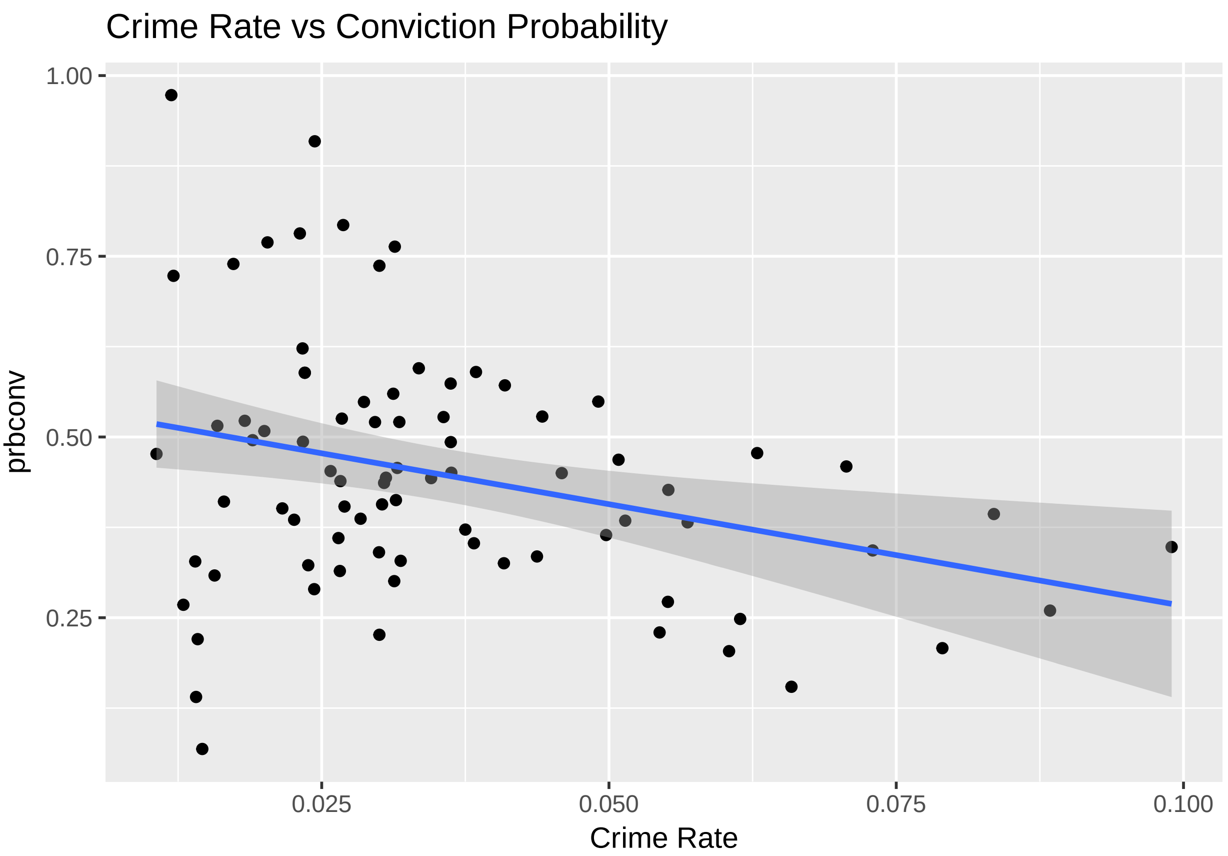


Here is some bivariate EDA.

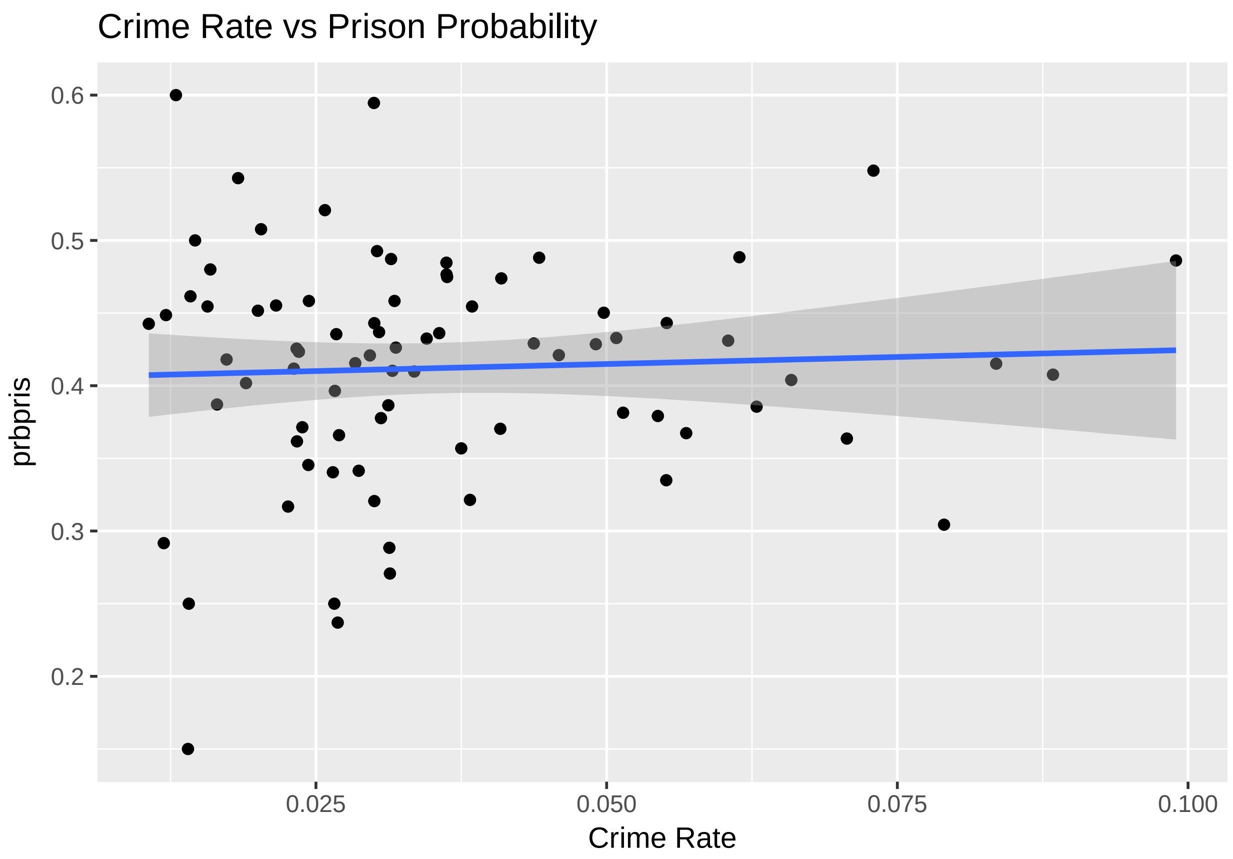
```
ggplot(t, aes(crmrte, prbarr)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    labs(title = 'Crime Rate vs Arrest Probability', x = 'Crime Rate')
```



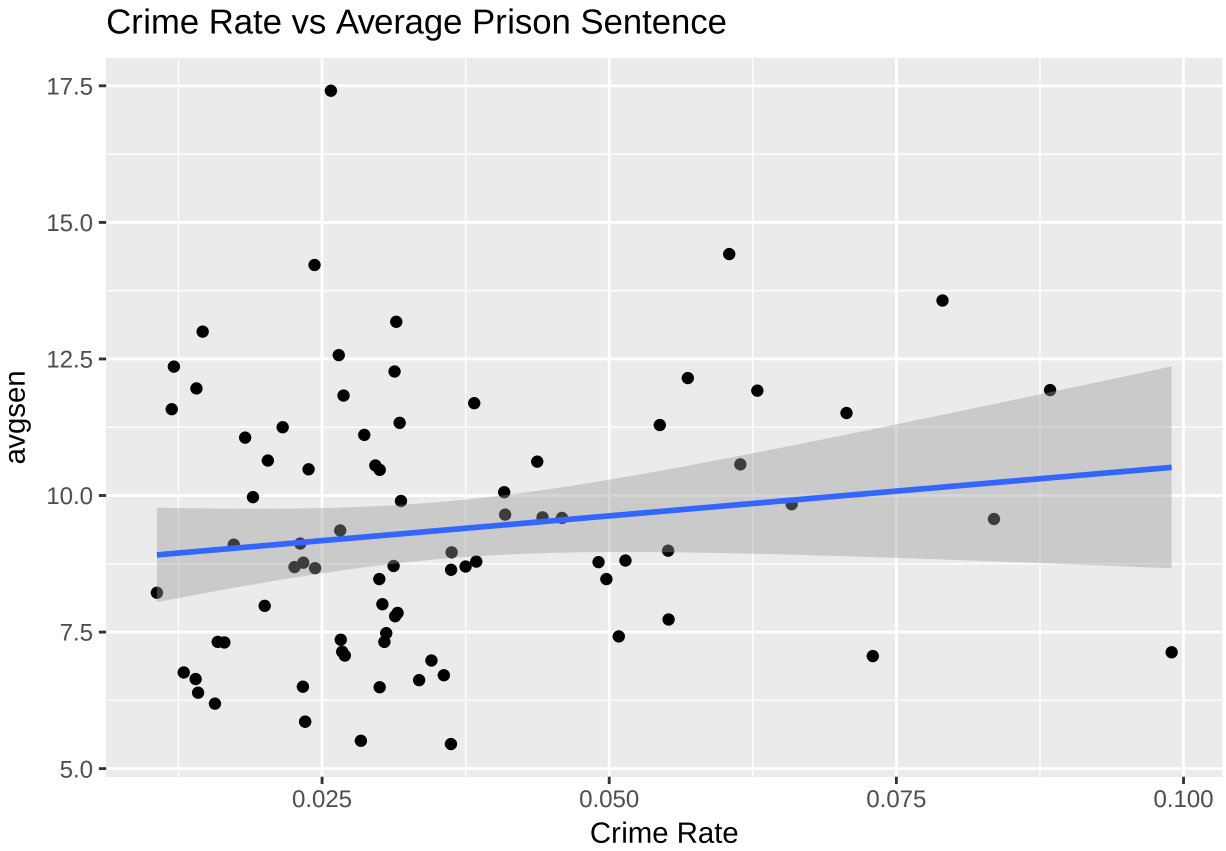
```
ggplot(t, aes(crmrte, prbconv)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    labs(title = 'Crime Rate vs Conviction Probability', x = 'Crime Rate')
```



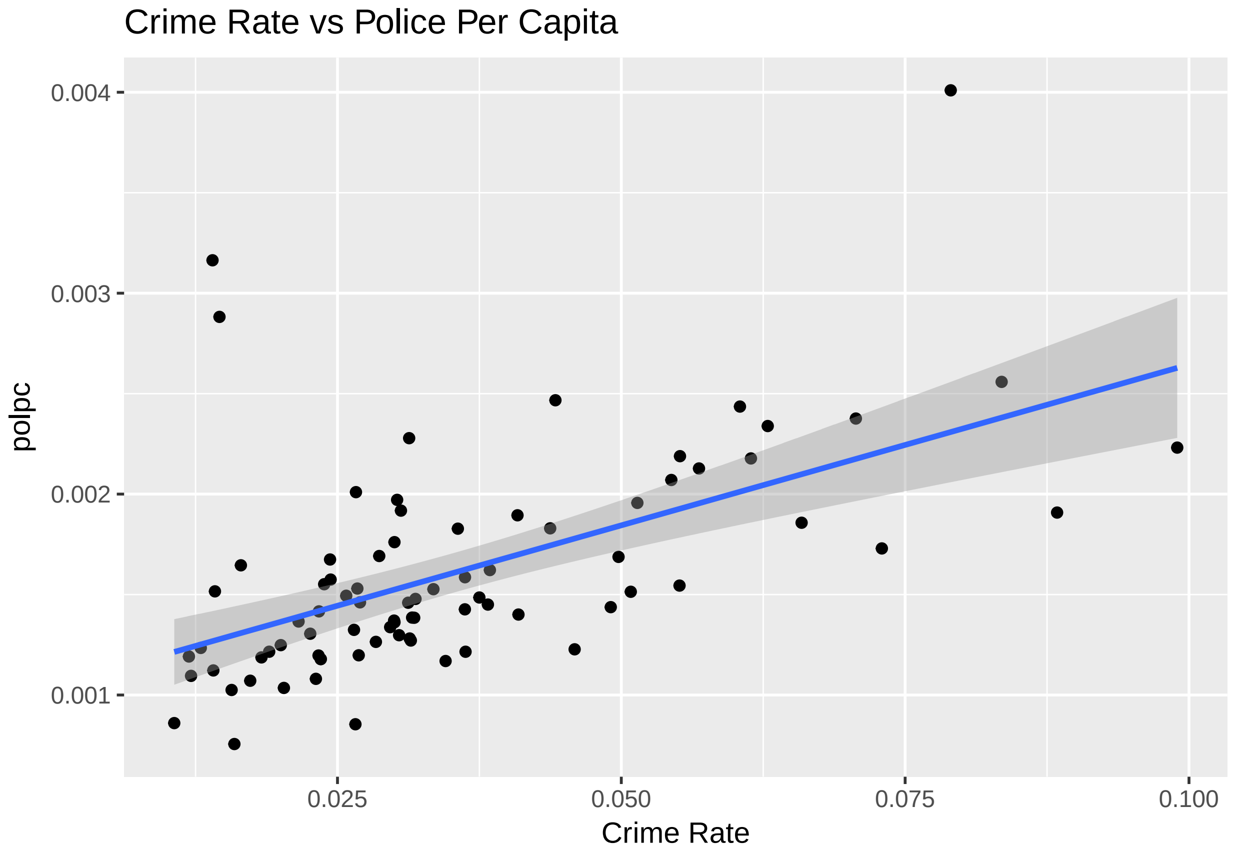
```r
ggplot(t, aes(crmrte, prbpris)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    labs(title = 'Crime Rate vs Prison Probability', x = 'Crime Rate')
```



```
ggplot(t, aes(crmrte, avgsen)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    labs(title = 'Crime Rate vs Average Prison Sentence', x = 'Crime Rate')
```



```
ggplot(t, aes(crmrte, polpc)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    labs(title = 'Crime Rate vs Police Per Capita', x = 'Crime Rate')
```


Crime Rate vs Police Per Capita
polpc
0.004
0.003
0.002
0.001
Crime Rate
0.025
0.050
0.075
0.100