Lab3

Determinants of crime

David Hou, Scott Hungerford, Irene Seo

```r
setwd("~/Desktop/W203.2/Assignments/Lab_3")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
```

Introduction

We are hired to examine the data to help the campaign understand the determinants of crime and to generate policy suggestions that are applicable to local government.

Cleansing Data

First we clean the data, using the dplyr package for its nice verbs. We remove NAs, change prbconv to numeric, and change all integer columns to factors.

```r
raw = as_tibble(read.csv('crime_v2.csv'))
t = raw %>%
    filter(!is.na(county)) %>%
    mutate(prbconv = as.numeric(prbconv) / 100) %>%
    mutate_if(is.integer, as.factor)
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')
```

As a data transformation, we sum up all of the wage types to make a single total wage.

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

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    91 obs. of  26 variables:
##  $ county  : Factor w/ 90 levels "1","3","5","7",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ year    : Factor w/ 1 level "87": 1 1 1 1 1 1 1 1 1 1 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : num  0.63 0.89 0.13 0.62 0.52 0.03 0.59 0.78 0.42 0.86 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : Factor w/ 2 levels "West","East": 2 2 1 2 1 1 2 2 2 2 ...
##  $ central : Factor w/ 2 levels "Outer","Central": 2 2 1 2 1 1 1 1 1 1 ...
##  $ urban   : Factor w/ 2 levels "Non-urban","Urban": 1 1 1 1 1 1 1 1 1 1 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
##  $ wage    : num  3055 2653 2554 2823 2759 ...
```

We have 91 observations from the data set to analyze.

Univariate Analysis

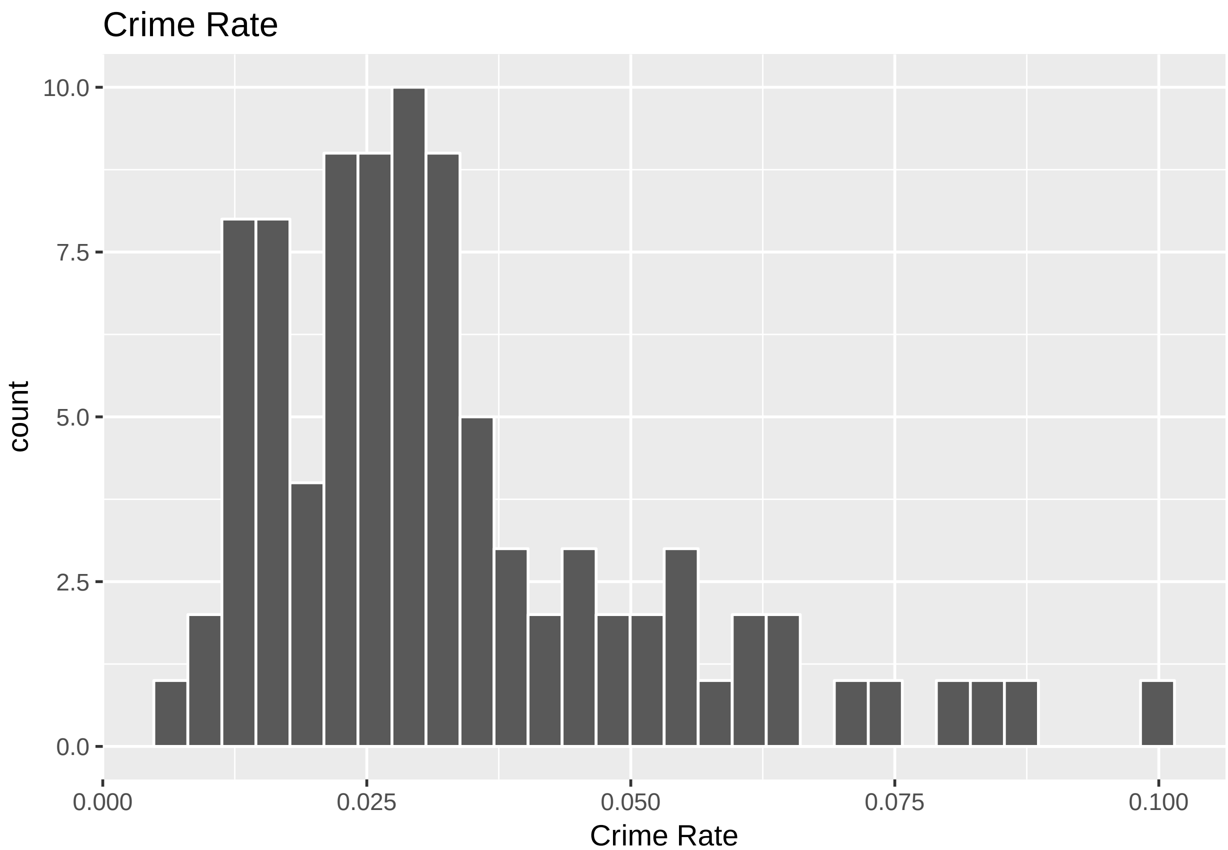
First we examine variables of interest to a politician with regards to changing policies. 1. Crime rate 2. Tax revenue per capita 3. Wages 4. Police per capita 5. Average sentences in days

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

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005533 0.020927 0.029986 0.033400 0.039642 0.098966
```

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



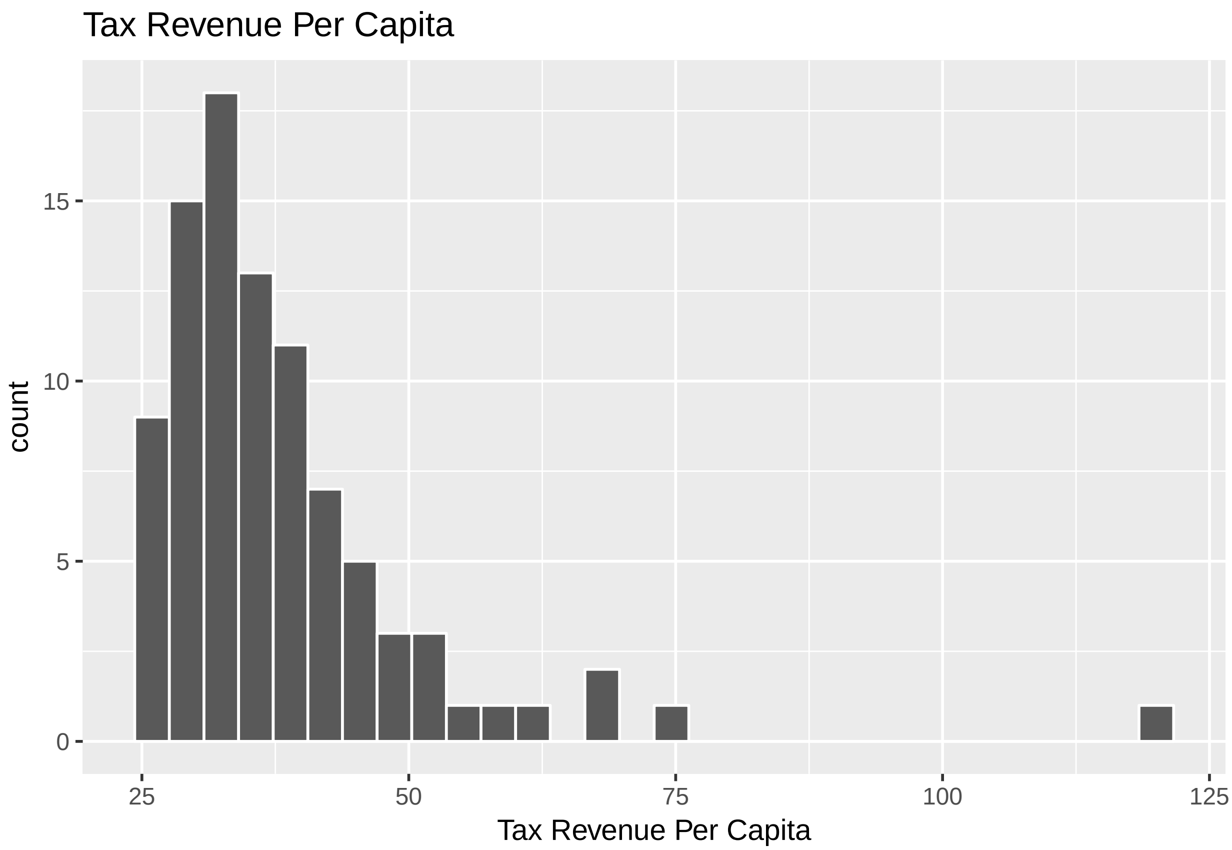
Crime rate shows skewed distribution with mean at 3.3%.

```r
summary(t$taxpc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.69   30.66   34.87   38.06   40.95  119.76
```

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



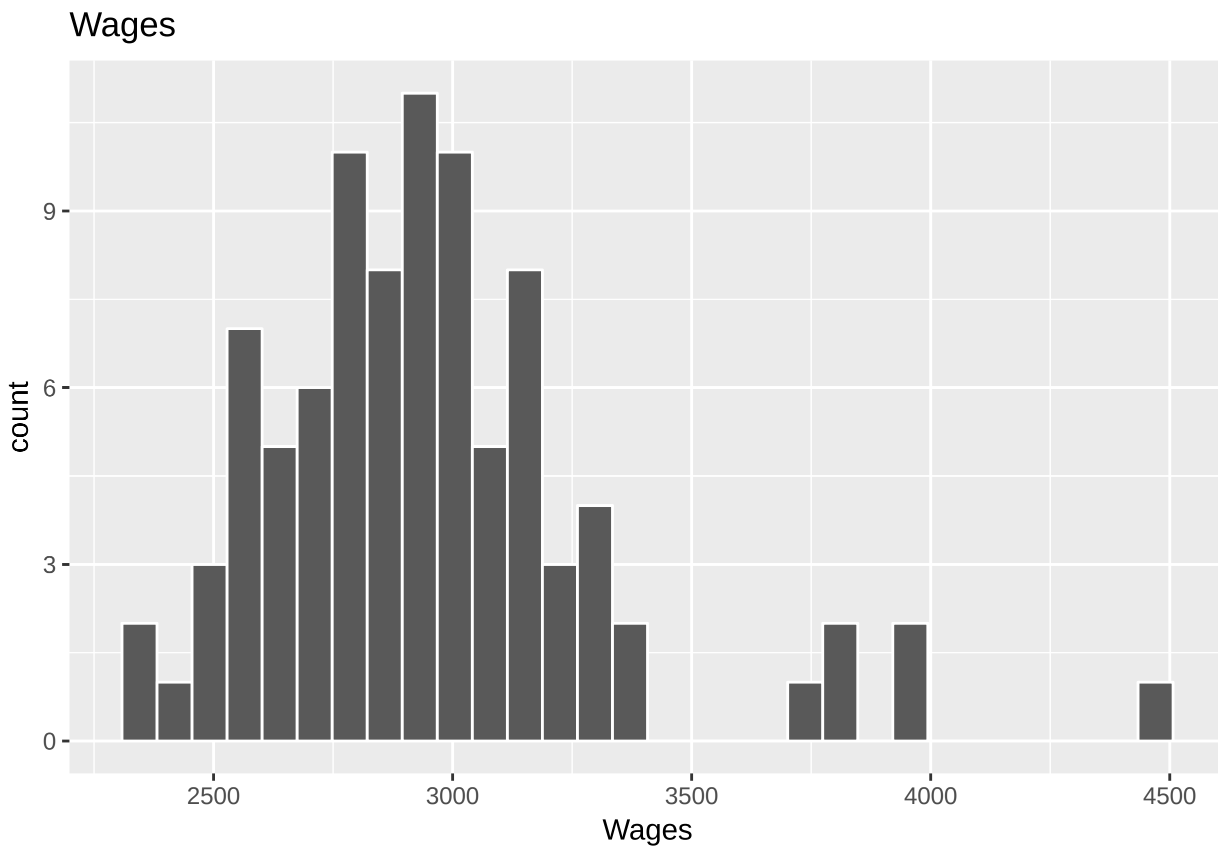
Tax revenue per capita also shows skewed distribution with mean value at 38 thousand dollars. (?)

```r
summary(t$wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2338    2722    2910    2955    3119    4464
```

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



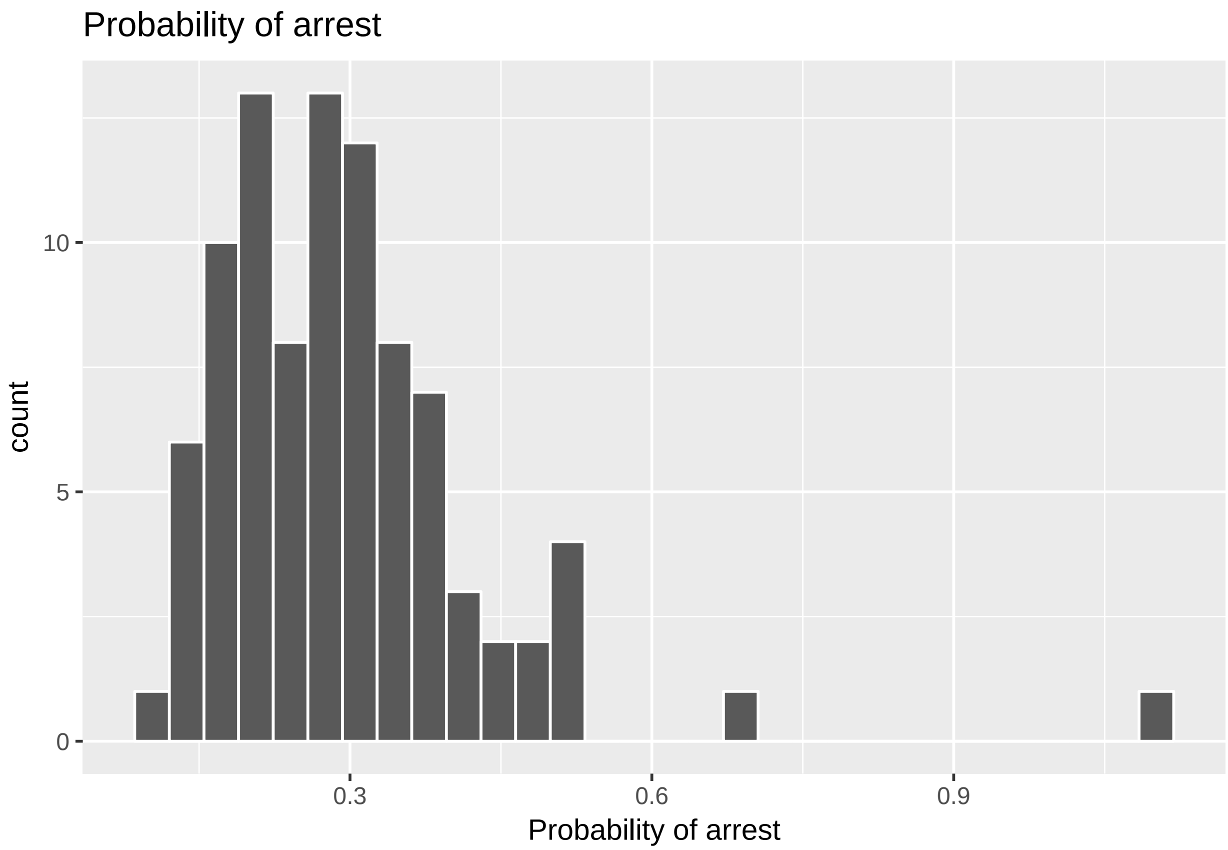
Wage also shows slightly skewed distribution similar to that of crime rate, with mean value at 2,955 and with some outliers to the right side of distribution.

```
summary(t$prbarr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09277 0.20568 0.27095 0.29492 0.34438 1.09091
```

```
qplot(t$prbarr, geom = 'histogram', col = I('white'), main = 'Probability of arrest', xlab = 'Probabili
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



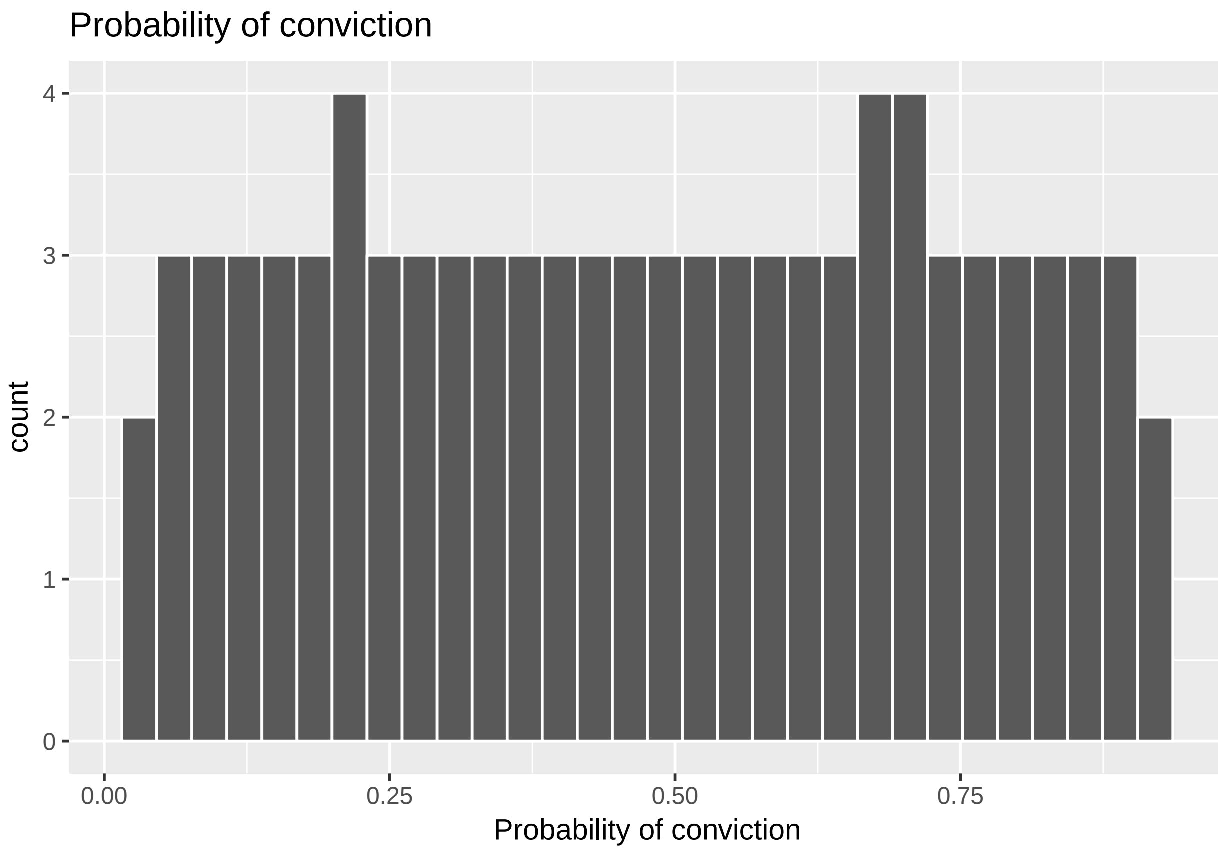
Probability of arrest has a left skewed distribution around mean at 29%.

```
summary(t$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0300  0.2550  0.4800  0.4775  0.7000  0.9200
```

```
qplot(t$prbconv, geom = 'histogram', col = I('white'), main = 'Probability of conviction', xlab = 'Proba
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



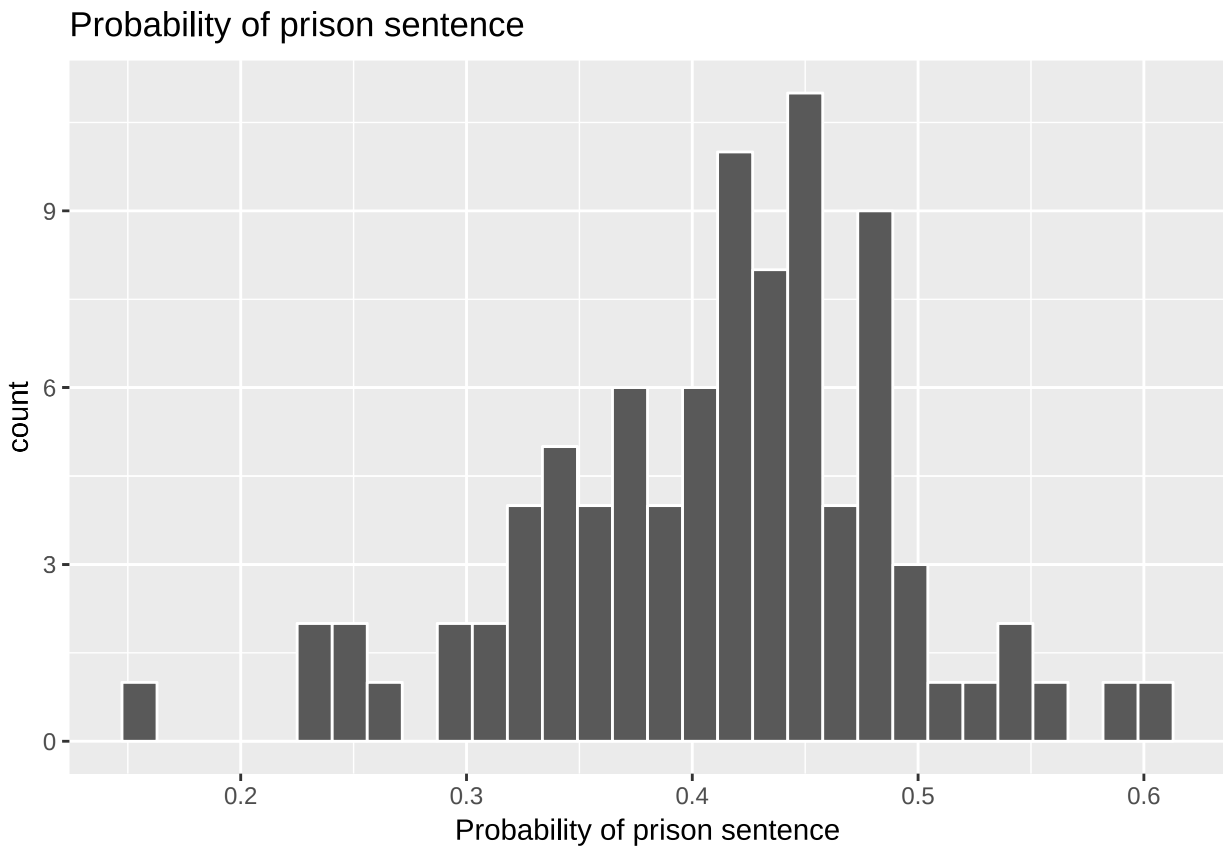
Probability of conviction has a uniform distribution.

```
summary(t$prbpris)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1500  0.3648  0.4234  0.4108  0.4568  0.6000
```

```
qplot(t$prbpris, geom = 'histogram', col = I('white'), main = 'Probability of prison sentence', xlab =
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



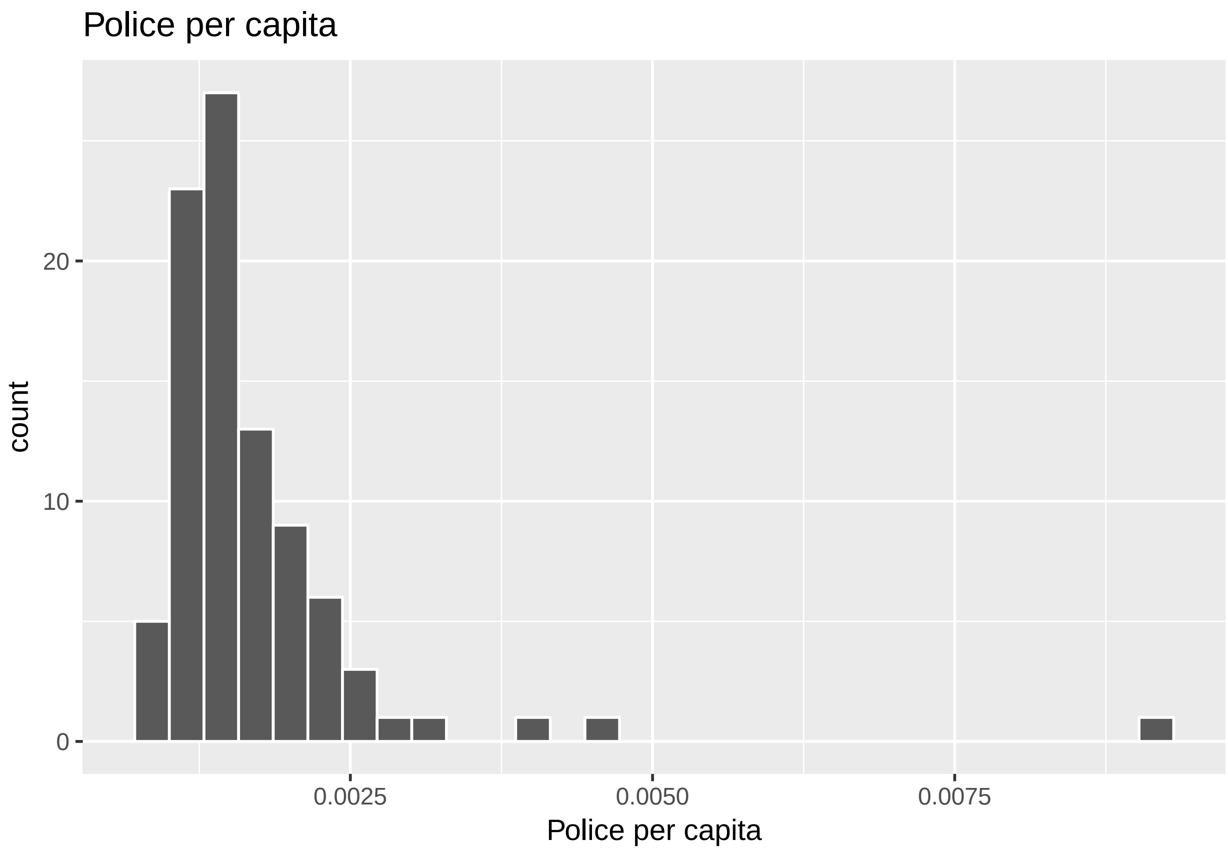
Probability of prison sentence has a right skewed distribution around mean at 41%.

```
summary(t$polpc)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0007459 0.0012308 0.0014853 0.0017022 0.0018768 0.0090543
```

```
qplot(t$polpc, geom = 'histogram', col = I('white'), main = 'Police per capita', xlab = 'Police per capi
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



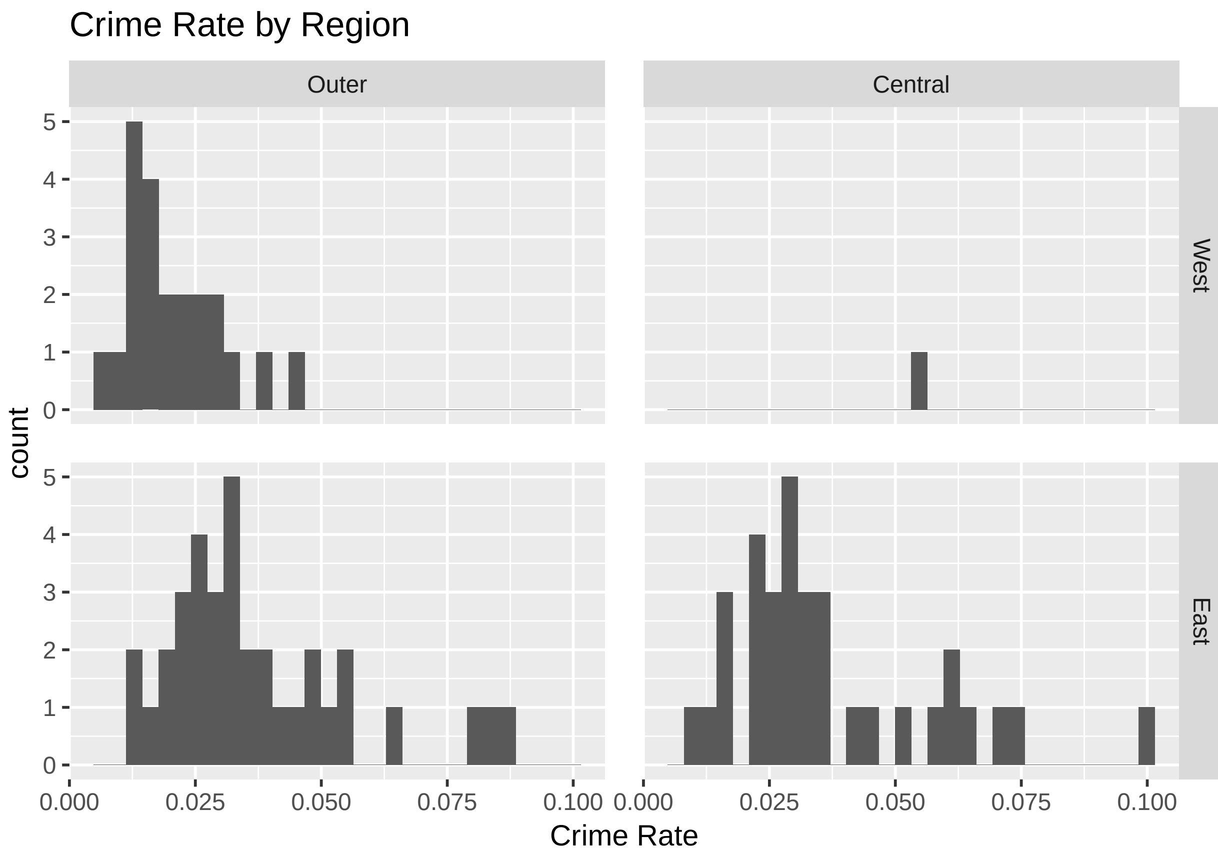
Police per capita has very skewed distribution with mean at .0017. It also has an extreme outlier to the right.

Crime rates by regions

Next we examine crime rates by regions and by density of population.

```r
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')
```

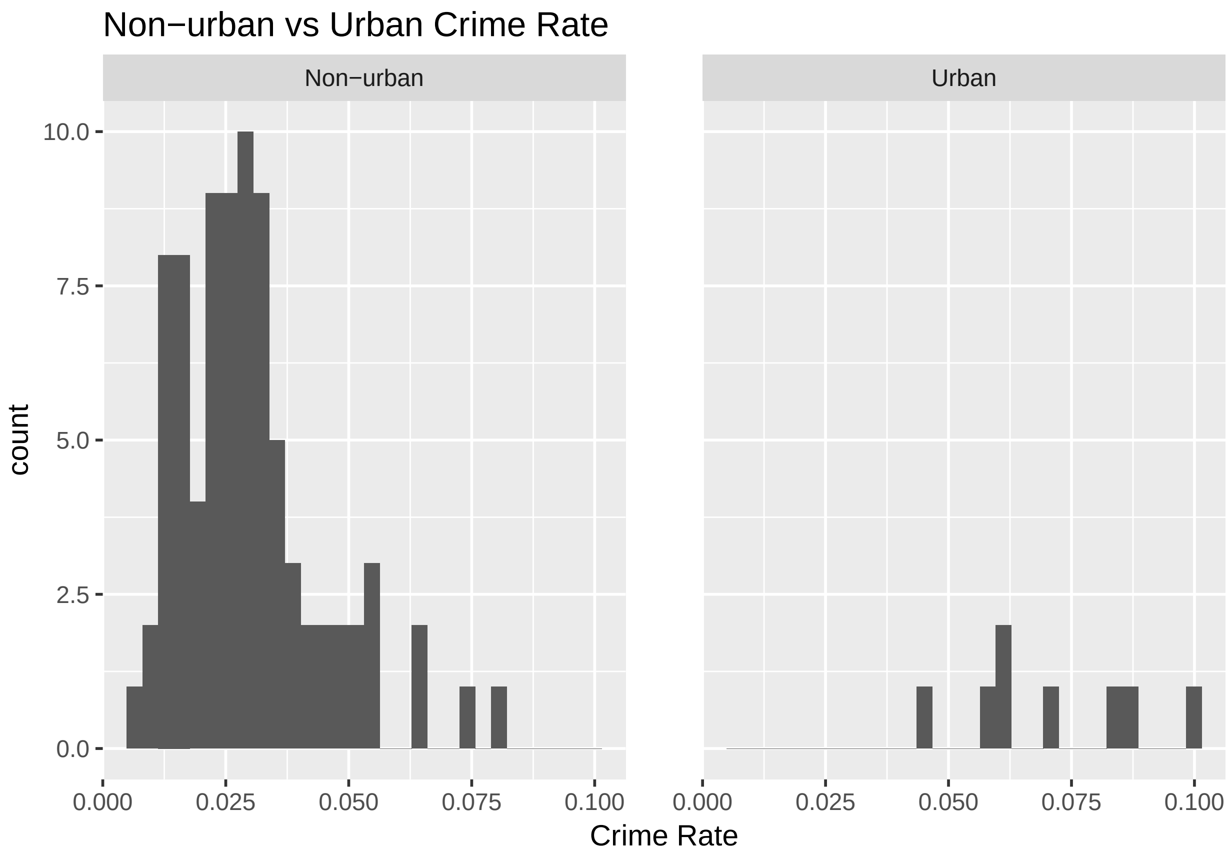
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Our data points come more from Eastern N.C. than Western N.C., and more from outer region than central region. Except for Western central region, they all show left skewed distributions. Eastern N.C. has crime rate spikes around 2.5%, whereas crime rate peaks at 1% in Western outer region.

```
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')
```

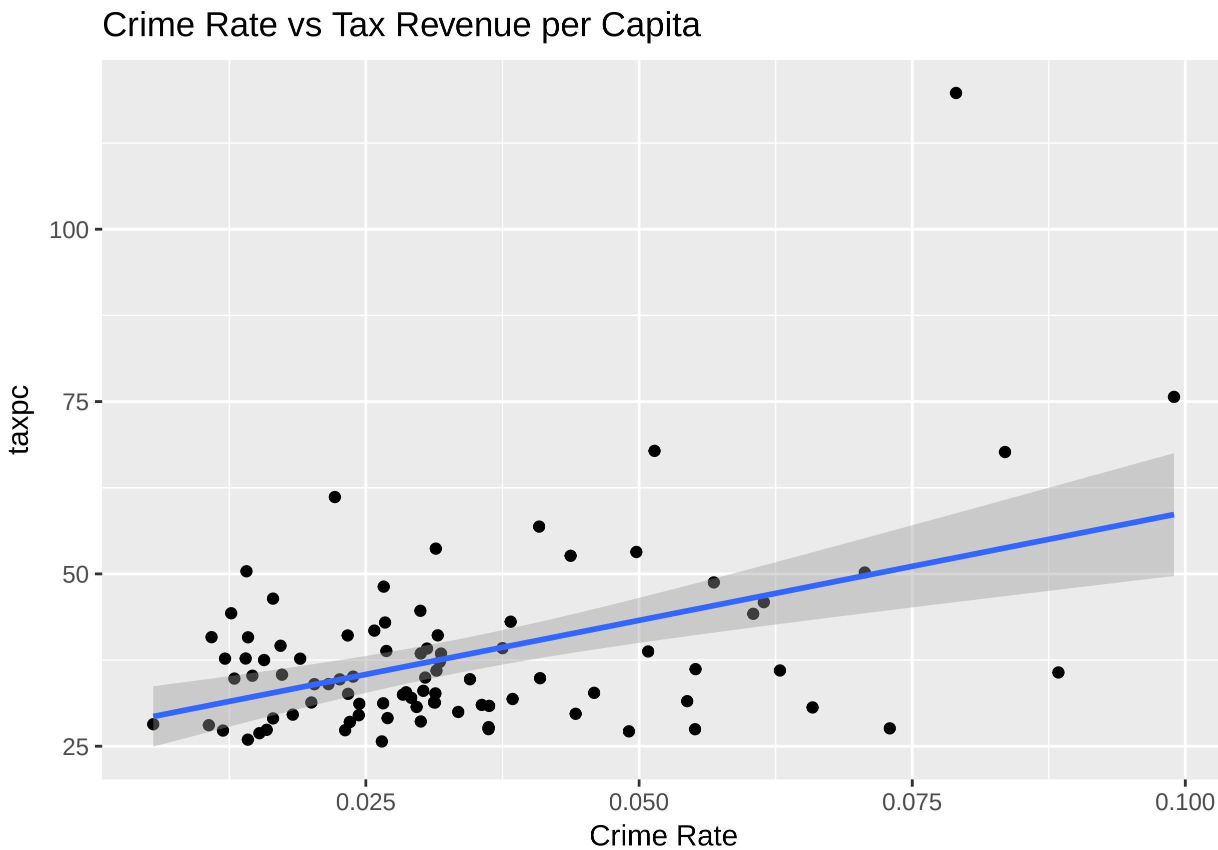
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We have far more data points of crimes in non-urban area than urban area. (?) Crime rates are densely populated around 3% in non-urban area with a left skewed distribution, whereas urban crime rates are dispersed around higher value of 5% to 10%.

Bivariate Analysis

```
ggplot(t, aes(crmrte, taxpc)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    labs(title = 'Crime Rate vs Tax Revenue per Capita', x = 'Crime Rate')
```




```r
lm(crmrte ~ taxpc, data=t)
```

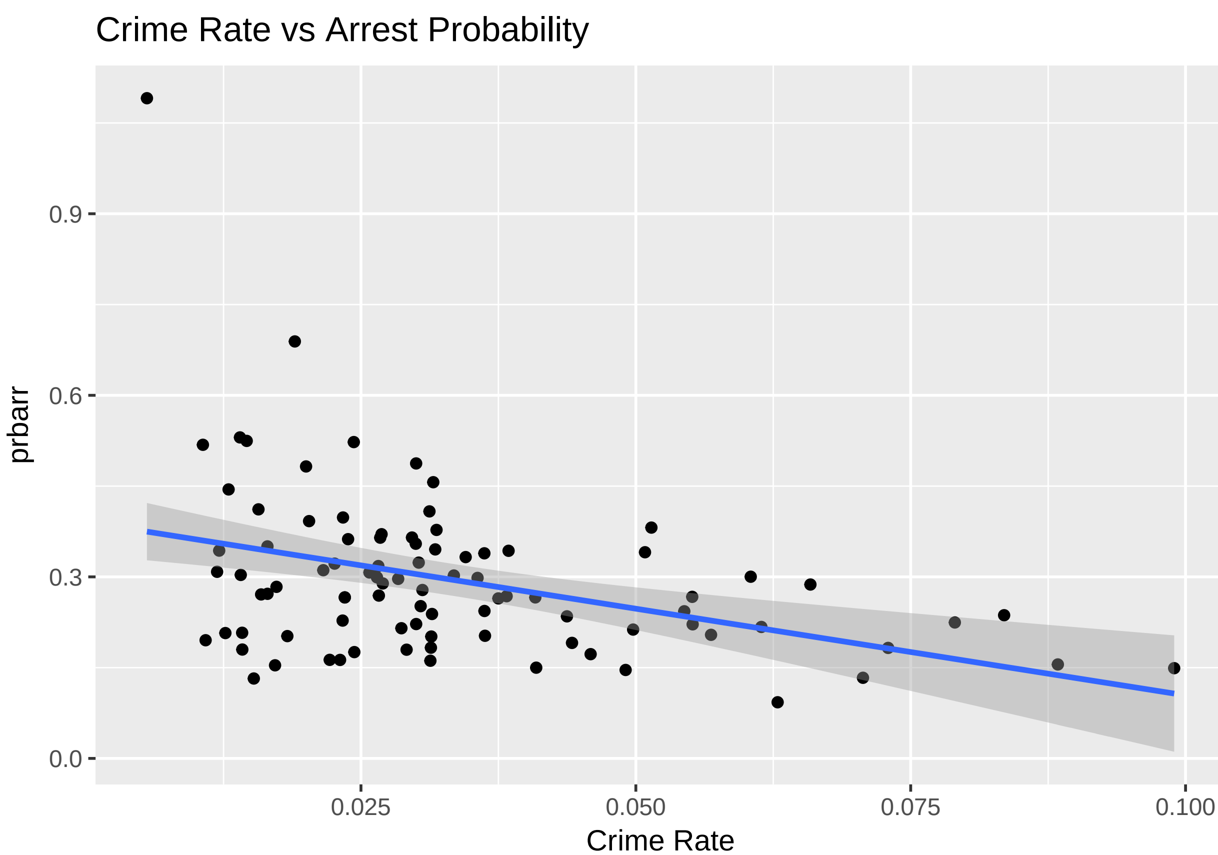
```
##
## Call:
## lm(formula = crmrte ~ taxpc, data = t)
##
## Coefficients:
## (Intercept)        taxpc
##   0.0087148    0.0006487
```

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


```
lm(crmrte ~ wage, data=t)
```

```
##
## Call:
## lm(formula = crmrte ~ wage, data = t)
##
## Coefficients:
## (Intercept)         wage
##  -2.442e-02    1.957e-05
```

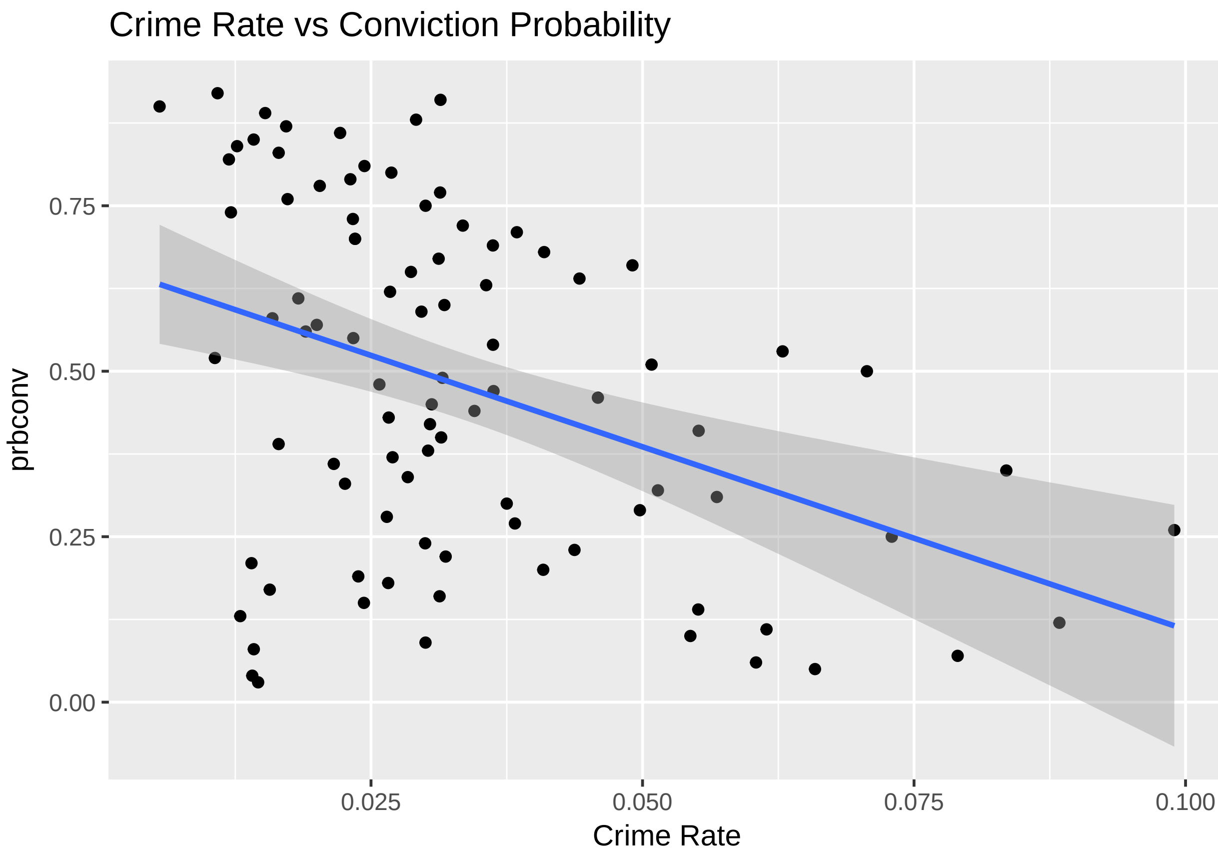
```
ggplot(t, aes(crmrte, prbarr)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    labs(title = 'Crime Rate vs Arrest Probability', x = 'Crime Rate')
```




```r
lm(crmrte ~ prbarr, data=t)
```

```
##
## Call:
## lm(formula = crmrte ~ prbarr, data = t)
##
## Coefficients:
## (Intercept)       prbarr
##     0.04933     -0.05403
```

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




```r
lm(crmrte ~ prbconv, data=t)
```

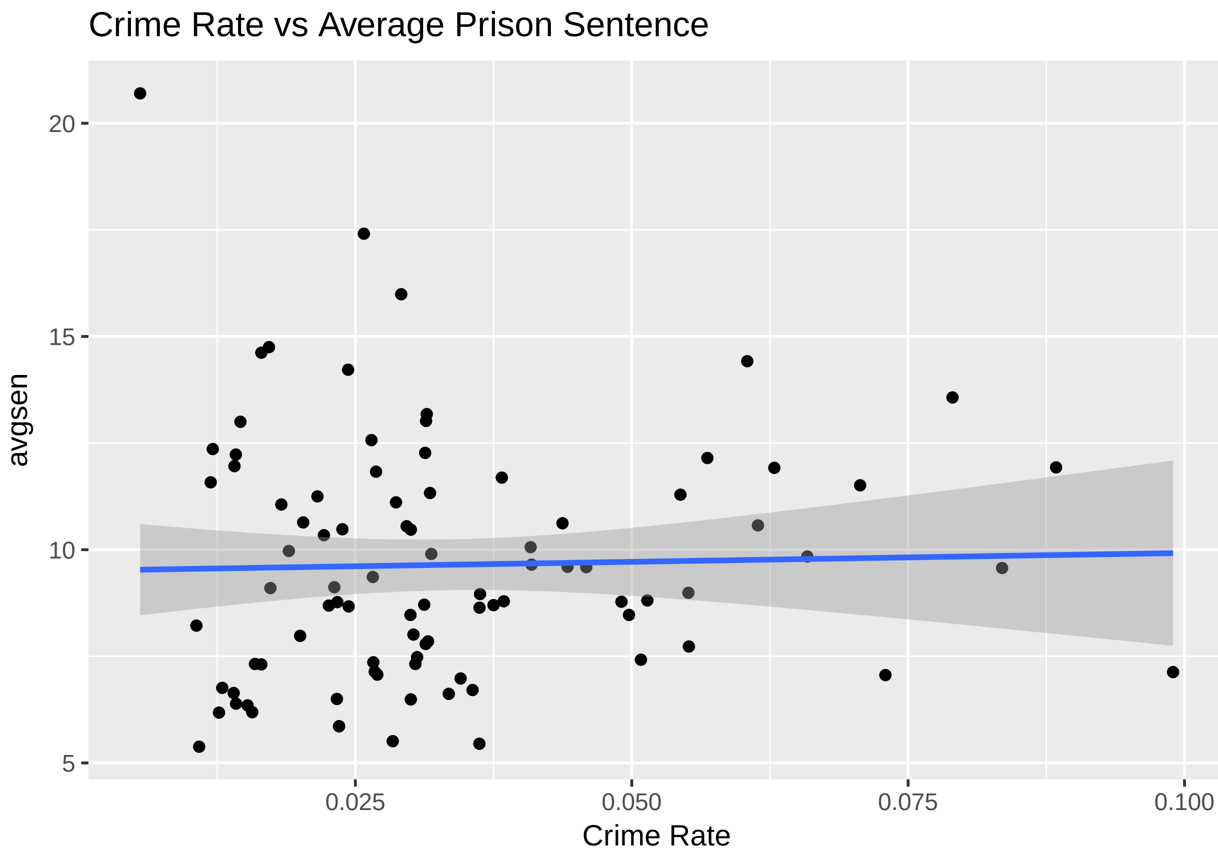
```
##
## Call:
## lm(formula = crmrte ~ prbconv, data = t)
##
## Coefficients:
## (Intercept)      prbconv
##     0.04711     -0.02872
```

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


```r
lm(crmrte ~ prbpris, data=t)
```

```
##
## Call:
## lm(formula = crmrte ~ prbpris, data = t)
##
## Coefficients:
## (Intercept)      prbpris
##     0.02888      0.01102
```

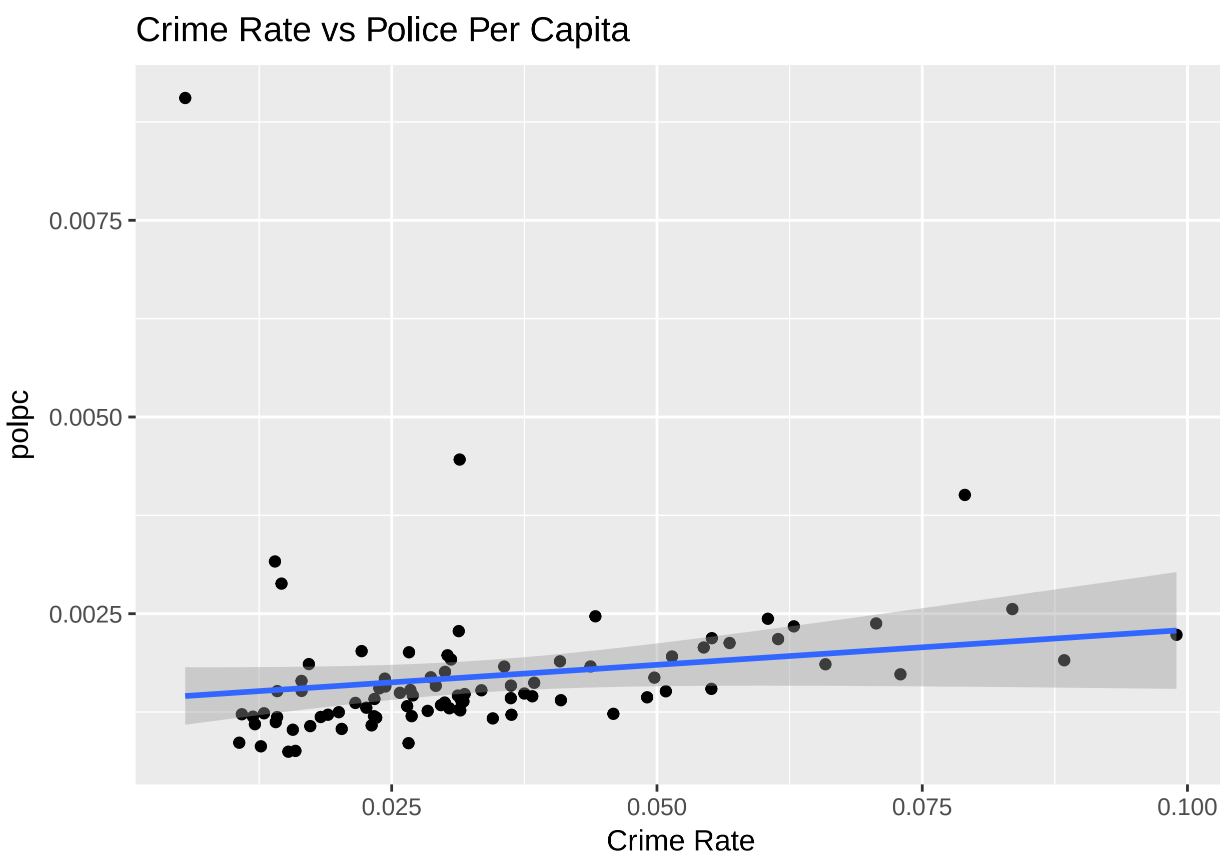
```r
ggplot(t, aes(crmrte, avgsen)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    labs(title = 'Crime Rate vs Average Prison Sentence', x = 'Crime Rate')
```




```r
lm(crmrte ~ avgsen, data=t)
```

```
##
## Call:
## lm(formula = crmrte ~ avgsen, data = t)
##
## Coefficients:
## (Intercept)       avgsen
##   0.0316530    0.0001811
```

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




```r
lm(crmrte ~ polpc, data=t)
```

```
##
## Call:
## lm(formula = crmrte ~ polpc, data = t)
##
## Coefficients:
## (Intercept)        polpc
##     0.02789      3.23791
```