W203.1_Lab3

Mary Boardman, Alan Tan

3/24/2018

Initial cleanse of the data

We will load the data and take an initial look to start our EDA.

Load data, and take a first look

```
#load("crime_v2.RData")
df_crime_orig = read.csv("crime_v2.csv")
print('Datatype of values of each column')
```

```
## [1] "Datatype of values of each column"
```

```
str(df_crime_orig)
```

```
## 'data.frame':    97 obs. of  25 variables:
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ 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 : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52 3 59 78 42 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    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 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 ...
```

```
#cat('\n\n')
#print('Summary of each column')
#summary(df_crime_orig)
```

Most columns are “int” or “num”, but “probconv” is an exception, it is a “Factor” type. We will deal with this column later.

Process NA values

With the exception of $prbconv (which is a factor instead of numeric data type), all other columns got 6 NA values, in detail:

```
# whereisna <- function(x) which(is.na(x))
# apply(df_crime_orig, 2,  whereisna)
```

Turns out all the NA are in the last 6 lines (line 92 to line 97).

```
tail(df_crime_orig)
```

```
##    county year crmrte prbarr prbconv prbpris avgsen polpc density taxpc
## 92     NA   NA     NA     NA       `      NA     NA    NA      NA    NA
## 93     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 94     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 95     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 96     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 97     NA   NA     NA     NA              NA     NA    NA      NA    NA
##    west central urban pctmin80 wcon wtuc wtrd wfir wser wmfg wfed wsta
## 92   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 93   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 94   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 95   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 96   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 97   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
##    wloc mix pctymle
## 92   NA  NA      NA
## 93   NA  NA      NA
## 94   NA  NA      NA
## 95   NA  NA      NA
## 96   NA  NA      NA
## 97   NA  NA      NA
```

Because these last 6 lines contain no actual data, they can be safely dropped.

```
df_crime_v2 <- df_crime_orig[1:91,]
```

After these observations were dropped, there seem to be no additional missing data.

Convert prbconv to numeric from Factor

Remember, $probconv was the one with different data type - it was a factor instead of numeric data type. So we will convert it from factor to numeric.

```
df_crime_v2$prbconv <- as.numeric(as.character(df_crime_orig$prbconv[1:91]))
print("New probconv column data type:")
```

```
## [1] "New probconv column data type:"
```

```
str(df_crime_v2$prbconv)
```

```
##  num [1:91] 0.528 1.481 0.268 0.525 0.477 ...
```

```
print("New summary of column prbconv")
```

```
## [1] "New summary of column prbconv"
```

```r
summary(df_crime_v2$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34541 0.45283 0.55128 0.58886 2.12121
```

Now we have 91 valid observations, with data type fixed for next step.

Variables we are interested in

We want to measure crime rate (crmrte) as dependent variable and several independent variables We hope that Policy per capita will have an impact to crime rate, especially from a policy perspective. We also expect probability of arrest and probability of conviction would represent effectiveness of police and prosecution, which in turn should have an impact to crime rate.

On the other hand, probability of prison time and average sentence can be linked to the nature of the crimes committed, and is more likely to be a result of the crime in addition to influence, we anticipate these to be less effective policy tools.

We are interested in demographic correlation with crime rate, however, we are conscious that these variables are not easy to manipulate by policies, and is more for information and reference (i.e. campaign focus neighborhoods, or even tailor messages for different neighborhoods).

We will consider log transformation for some of the variables, mainly because social behavior tent to follow marginal effect reduction principle - for policy to keep yield desire effect as it increase, the unit increase in resource will need to increase to yield similar result.

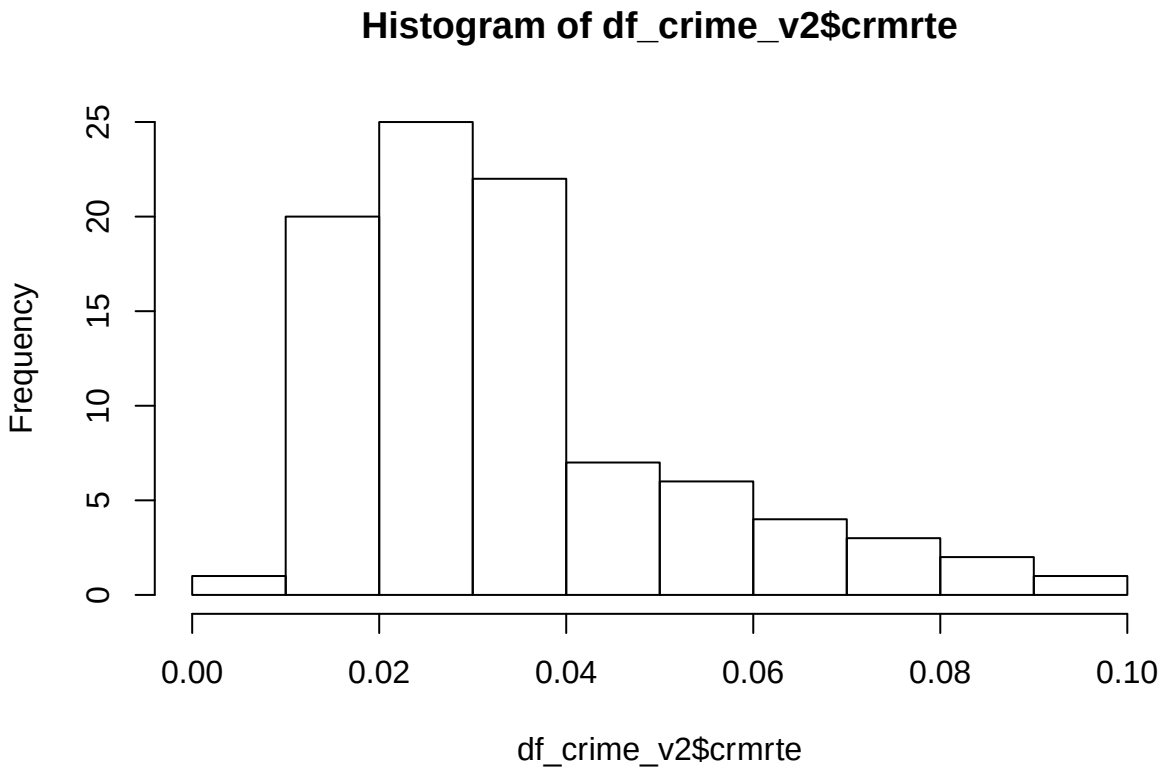
Details are provided here:

The dependent variable is crmrte, or crimes committed per person.

```r
summary(df_crime_v2$crmrte)
```

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

```r
hist(df_crime_v2$crmrte,breaks=13)
```



The histogram shows a clear left skew, which is expected - because crime rate has a bottom of "0", and unlimited upper end, so it should have a left skew. The sharp drop on the left is intriguing, let's take a quick look at the extreme left where crime rate is less than 0.01.

```
df_crime_v2[which(df_crime_v2$crmrte < 0.01),]
```

```
##    county year    crmrte  prbarr prbconv prbpris avgsen      polpc
## 51    115   87 0.0055332 1.09091     1.5     0.5   20.7 0.00905433
##      density   taxpc west central urban pctmin80     wcon     wtuc
## 51 0.3858093 28.1931    1       0     0  1.28365 204.2206 503.2351
##        wtrd     wfir     wser   wmfg  wfed   wsta   wloc mix    pctymle
## 51 217.4908 342.4658 245.2061 448.42 442.2 340.39 386.12 0.1 0.07253495
```

Looks like line 51, county #115 is uniquely low in crime. It is the only one which has a crime rate lower than 0.01, while 20 other counties have crime rate between 0.01 and 0.02.

Within the context of political campaign, it is interesting to note that high crime rate actually affect less counties, where most countries have relatively low crime rate. We will keep this in mind in our further analysis: should we reduce already low crime rate (do what we did better), or help high crime rate neighborhood achieve more dramatic crime reductions?

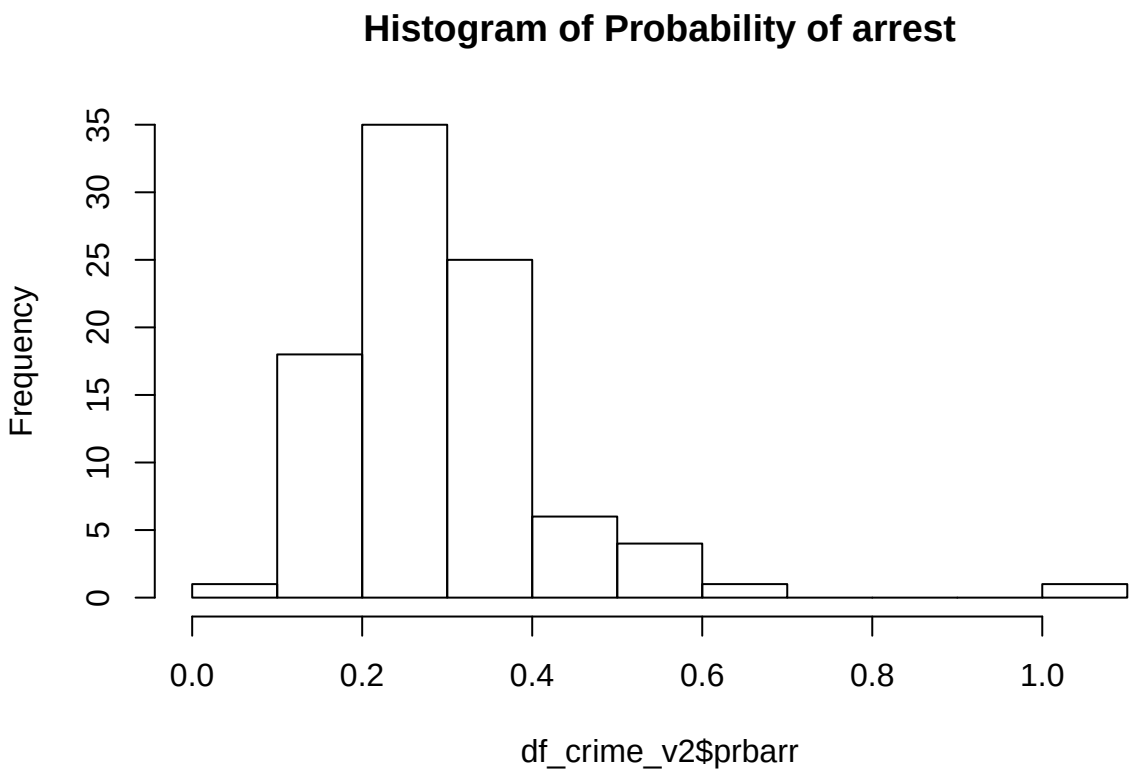
Independent Variables

For independent variables, we start with several groups. First, there is the probability of punishment. These are captured by probability of arrest (prbarr), probability of conviction (prbconv), probability of prison sentence (prbpris), Secondly, the average sentence in days (avgsen) is used to represent severity of punishment. Third, police per capita (polpc) is used. These 3 groups are mainly policy affecting areas that are most relevant to a policy debate during political campaign. We anticipate these to be (highly) col-linear.
The other variables, mostly demographic will be used to guide the political campaign messaging.

Probability of punishment variables (policy related)

Probability of arrest

```
hist(df_crime_v2$prbarr, main="Histogram of Probability of arrest")
```



Also, there is a heavy left skew. Looks like most crimes are not arrested. Most counties arrest less than 1/3 of the crime happened.

```
summary(df_crime_v2$prbarr)
```

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

```
cat('\nThe line contains the outlier (prbarr > 1) is:' , which(df_crime_v2$prbarr>1))
```

```
##
## The line contains the outlier (prbarr > 1) is: 51
```

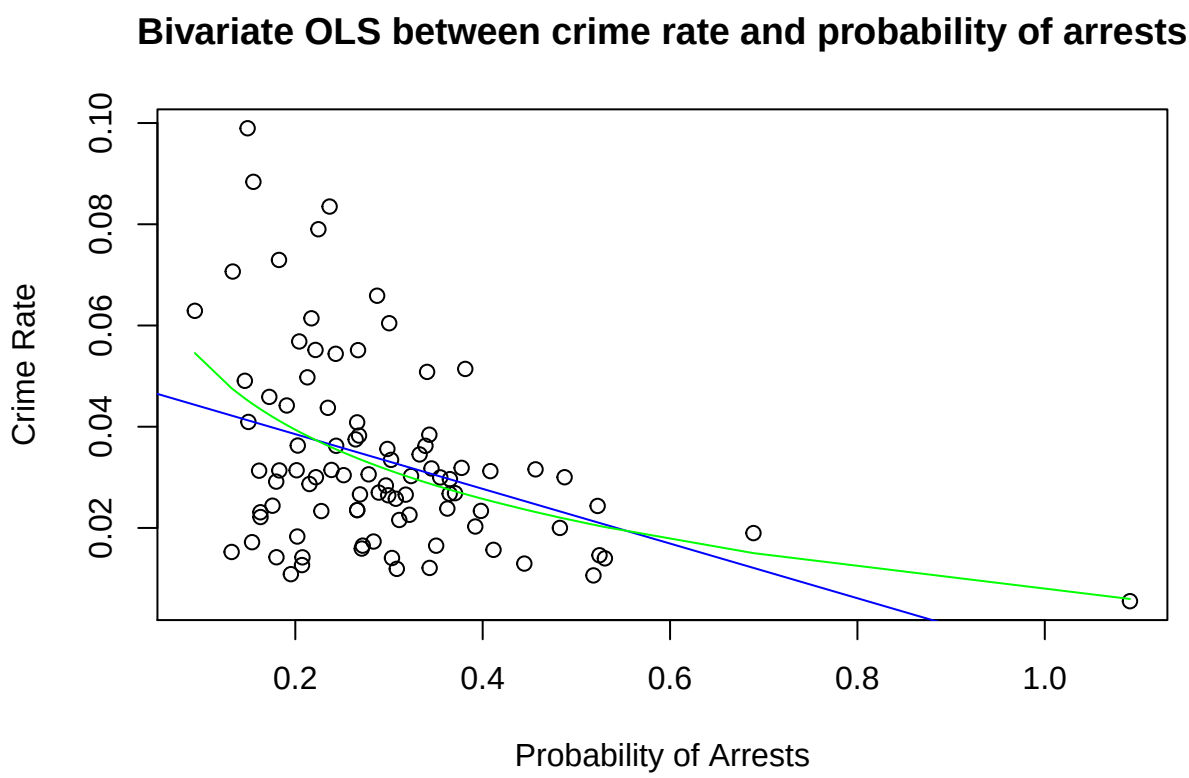
With probability of arrest, there does appear to be an outlier. It's interesting that there is a value greater than "1". Looks like a county has very high arrest rate, it is actually capturing more suspects than crimes happening within its boarder.

```
df_crime_v2[which(df_crime_v2$prbarr>1),]
```

```
##    county year    crmrte  prbarr prbconv prbpris avgsen      polpc
## 51    115   87 0.0055332 1.09091     1.5     0.5   20.7 0.00905433
##      density   taxpc west central urban pctmin80     wcon     wtuc
## 51 0.3858093 28.1931    1       0     0  1.28365 204.2206 503.2351
##        wtrd     wfir     wser   wmfg  wfed   wsta   wloc mix    pctymle
## 51 217.4908 342.4658 245.2061 448.42 442.2 340.39 386.12 0.1 0.07253495
```

Yes, it is indeed county #115 which we saw has pretty low crime, and maybe it capture suspects from other counties...

```r
plot(df_crime_v2$prbarr , df_crime_v2$crmrte, xlab = "Probability of Arrests", ylab = "Crime Rate", mai
  probarr_model <- lm (crmrte ~ prbarr, data = df_crime_v2)
  trn_prbarr = log(df_crime_v2$prbarr)
  probarr_model_trn <- lm (crmrte ~ trn_prbarr , data = df_crime_v2)
  abline(probarr_model, col="blue")
  lines(df_crime_v2$prbarr[order(df_crime_v2$prbarr)],probarr_model_trn$fitted.values[order(df_crime_v2
```



Looks like there is a (weak) inverse correlation between probability of arrest and crime rate.
Blue line is the direct bivariate OLS, green line represents the bivariate OLS between Crime Rate and log(Probability of Arrests)

```r
  probarr_model$coefficients
```

```
## (Intercept)      prbarr
##  0.04933460 -0.05403004
```

```r
  cat ("r^2 for linear model with $prbarr is:" , summary(probarr_model)$r.squared, "\n")
```

```
## r^2 for linear model with $prbarr is: 0.1547083
```

```r
  probarr_model_trn$coefficients
```

```
##  (Intercept)   trn_prbarr
##  0.007681533 -0.019713974
```

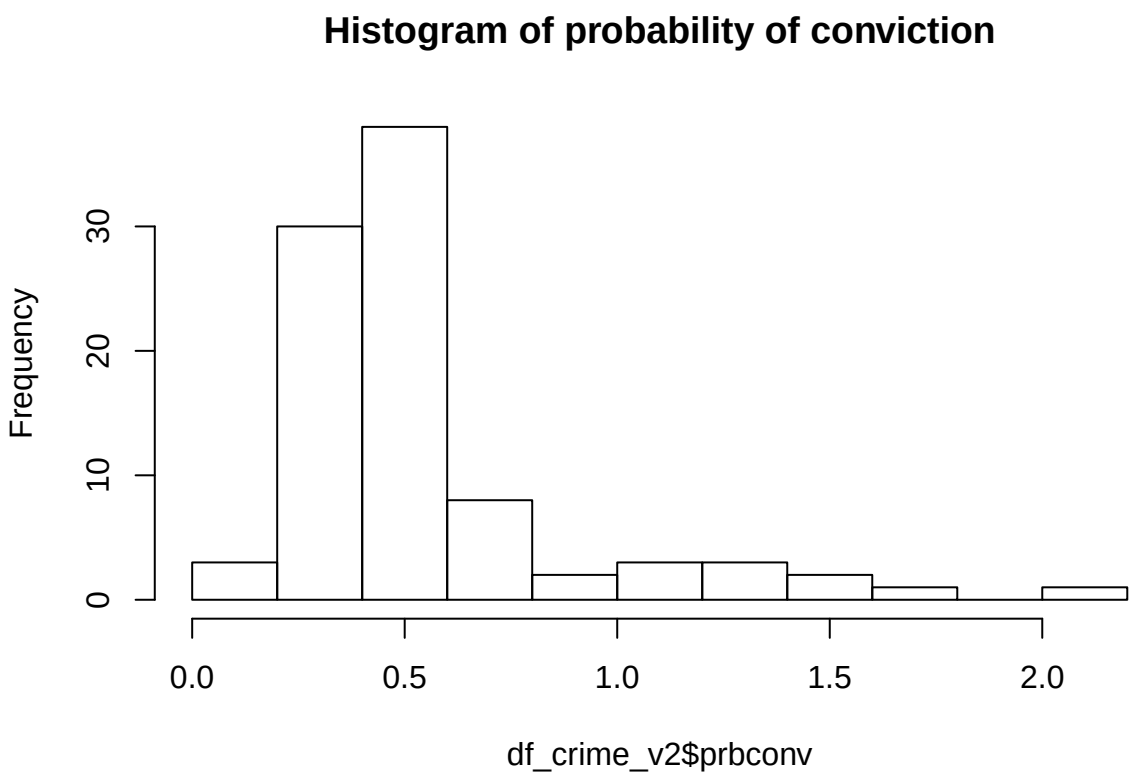
```r
  cat ("r^2 for linear model with log($prbarr) is:" , summary(probarr_model_trn)$r.squared, "\n")
```

```
## r^2 for linear model with log($prbarr) is: 0.176251
```

So as expected, the r^2 for log(prbarr) is higher than that of probarr directly.

Probability of conviction

```
hist(df_crime_v2$prbconv,breaks=13, main="Histogram of probability of conviction")
```



Similar to other variables, there is a left skew. Also there is outlier:

```
summary(df_crime_v2$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34541 0.45283 0.55128 0.58886 2.12121
```

```
cat('\nThe line contains the outlier (prbconv > 2) is:' ,which(df_crime_v2$prbconv > 2) )
```

```
##
## The line contains the outlier (prbconv > 2) is: 84
```

So the outlier is line 84, which has a conviction probability over 2.

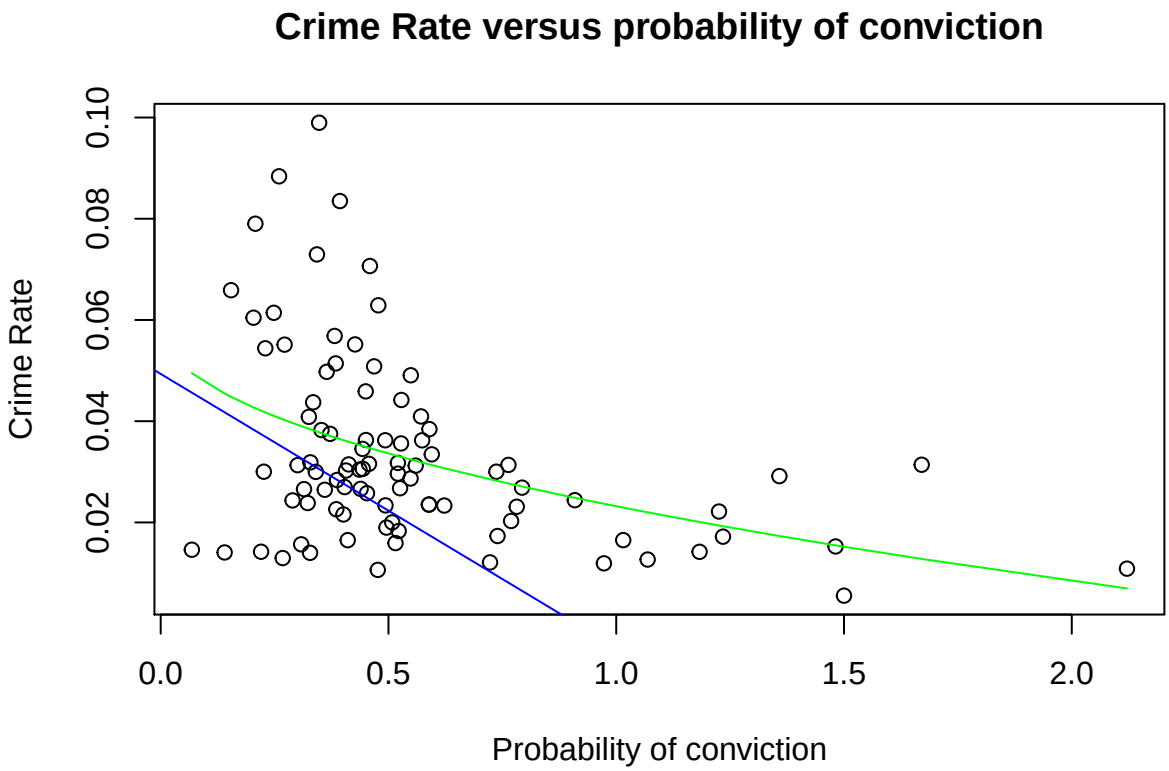
```
df_crime_v2[84,]
```

```
##    county year    crmrte   prbarr prbconv  prbpris avgsen     polpc
## 84    185   87 0.0108703 0.195266 2.12121 0.442857   5.38 0.0012221
##      density    taxpc west central urban pctmin80     wcon    wtuc
## 84 0.3887588 40.82454    0       1     0  64.3482 226.8245 331.565
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 84 167.3726 264.4231 2177.068 247.72 381.33 367.25 300.13 0.04968944
##       pctymle
## 84 0.07008217
```

Take a look at county #185, everything else seems "normal".

```
  plot(df_crime_v2$prbconv , df_crime_v2$crmrte, , xlab = "Probability of conviction", ylab = "Crime Ra
  probconv_model <- lm (crmrte ~ prbconv, data = df_crime_v2)
  trn_prbconv = sqrt(df_crime_v2$prbconv)
  probconv_model_trn <- lm (crmrte ~ trn_prbconv , data = df_crime_v2)
```

```
  abline(probarr_model, col="blue")
  lines(df_crime_v2$prbconv[order(df_crime_v2$prbconv)],probconv_model_trn$fitted.values[order(df_crime_
```



According to the scatterplot, there seems to be a (weak) inverse correlation.

```
  probconv_model$coefficients
```

```
## (Intercept)     prbconv
##  0.04476325 -0.02061212
```

```
  cat ("r^2 for linear model with $prbconv is:" , summary(probconv_model)$r.squared, "\n")
```

```
## r^2 for linear model with $prbconv is: 0.1489747
```

```
  probconv_model_trn$coefficients
```

```
## (Intercept) trn_prbconv
##  0.05878176 -0.03558465
```

```
  cat ("r^2 for linear model with sqrt($prbconv) is:" , summary(probconv_model_trn)$r.squared, "\n")
```

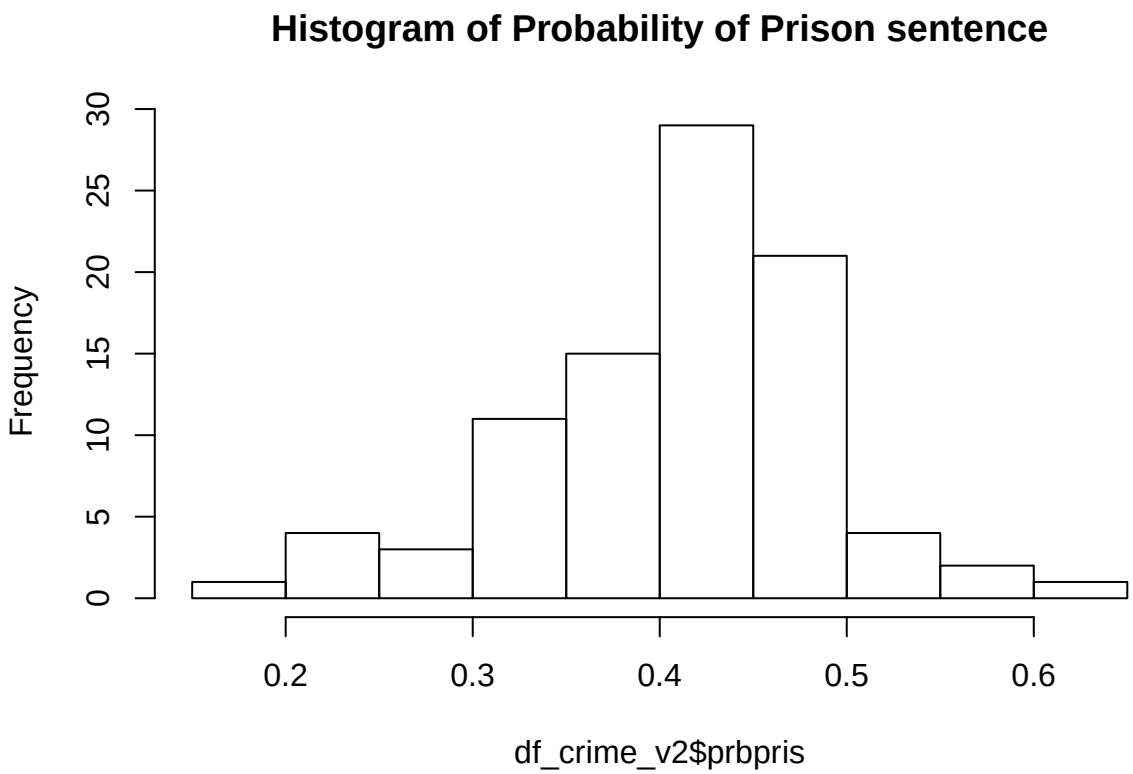
```
## r^2 for linear model with sqrt($prbconv) is: 0.1538574
```

The r^2 for $\sqrt{prbconv}$ is higher than that of probconv directly.

For probability of conviction, a different transformation yield better result - square root, which also demonstrate a feature of "tapering off" effect, that marginal increase of x has reduced effect on y.

probability of Prison time

```
hist(df_crime_v2$prbpris,breaks=13, main="Histogram of Probability of Prison sentence")
```

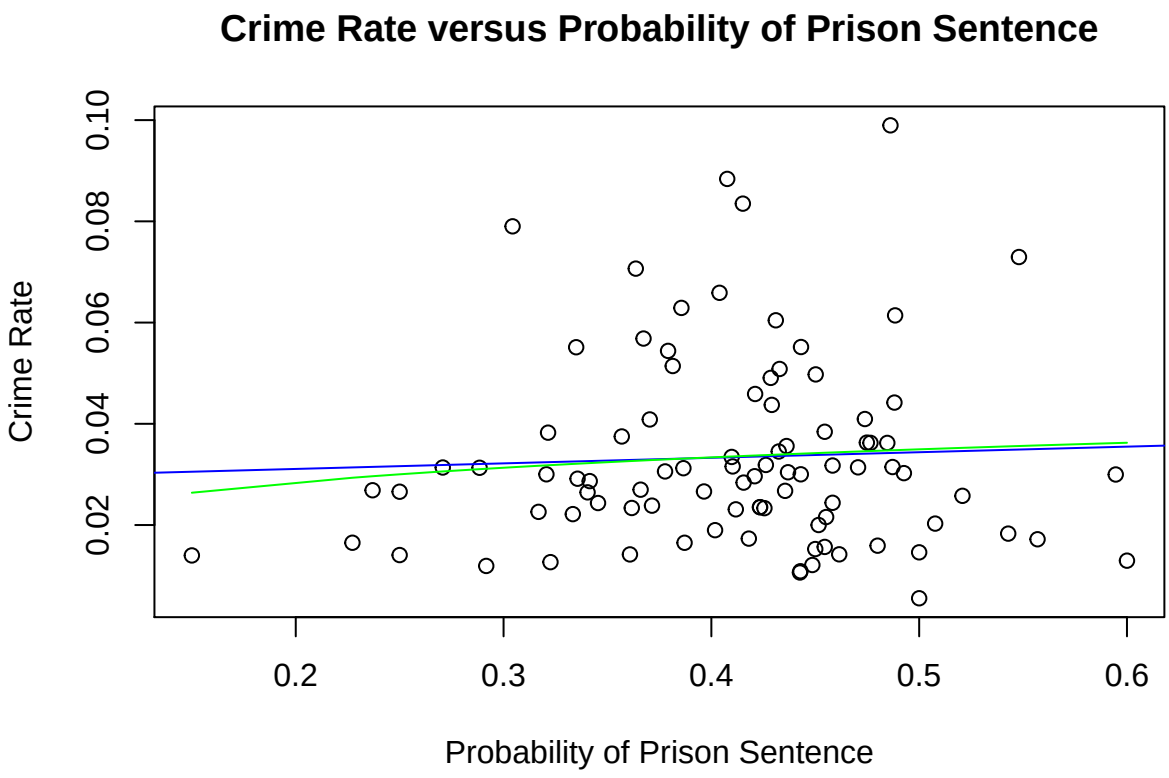


From the histogram, prbpris almost approximates normal. No obvious outlier shown

```
summary(df_crime_v2$prbpris)
```

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

```
plot(df_crime_v2$prbpris , df_crime_v2$crmrte, xlab = "Probability of Prison Sentence", ylab = "Crime Ra
  probpris_model <- lm (crmrte ~ prbpris, data = df_crime_v2)
  trn_prbpris = log(df_crime_v2$prbpris)
  probpris_model_trn <- lm (crmrte ~ trn_prbpris , data = df_crime_v2)
  abline(probpris_model, col="blue")
  lines(df_crime_v2$prbpris[order(df_crime_v2$prbpris)],probpris_model_trn$fitted.values[order(df_crime_
```



There *might* be a positive correlation here between probpris and crime rate. This could be explained by the possibility that in higher crime rate area, repeated offend are more likely, and repeated offence are known to be less tolerated, less forgiven, and more likely to trigger prison time.

```
  probpris_model$coefficients
```

```
## (Intercept)     prbpris
##  0.02887553  0.01101526
```

```
  cat ("r^2 for linear model with $prbarr is:" , summary(probpris_model)$r.squared, "\n")
```

```
## r^2 for linear model with $prbarr is: 0.002207522
```

```
  probpris_model_trn$coefficients
```

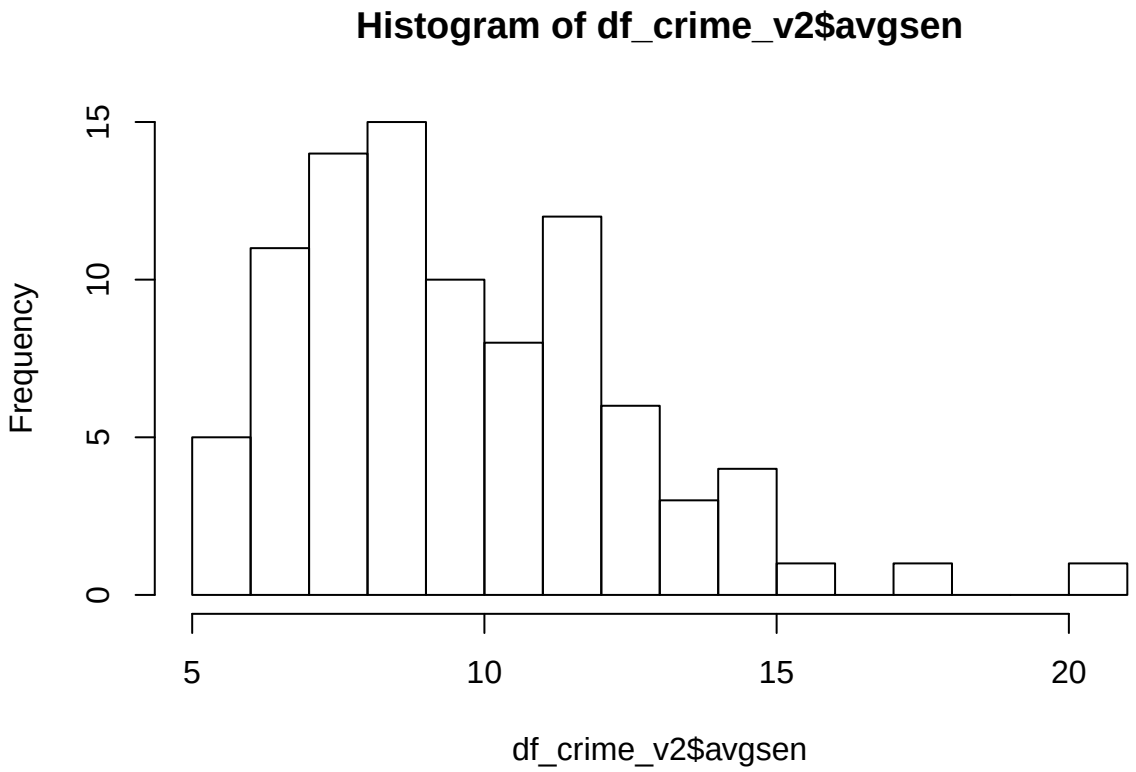
```
## (Intercept) trn_prbpris
## 0.039891917 0.007120793
```

```
  cat ("r^2 for linear model with log($prbarr) is:" , summary(probpris_model_trn)$r.squared, "\n")
```

```
## r^2 for linear model with log($prbarr) is: 0.006993036
```

While the coefficients shows a positive correlation, the effect seems really small, and the r^2 is very very low both for the direct bivariate OLS and the log transformed version.

```
hist(df_crime_v2$avgsen, breaks=13)
```



Looks like a typical left skew with some outliers on the right.

```
print("Summary of Average Sentence:")
```

```
## [1] "Summary of Average Sentence:"
```

```
summary(df_crime_v2$avgsen)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.380   7.340   9.100   9.647  11.420  20.700
```

```
cat("The outliers (average sentence > 20) is:", which(df_crime_v2$avgsen > 20))
```

```
## The outliers (average sentence > 20) is: 51
```

```
df_crime_v2[which(df_crime_v2$avgsen > 20),]
```

```
##    county year    crmrte  prbarr prbconv prbpris avgsen      polpc
## 51    115   87 0.0055332 1.09091     1.5     0.5   20.7 0.00905433
##      density   taxpc west central urban pctmin80     wcon     wtuc
## 51 0.3858093 28.1931    1       0     0  1.28365 204.2206 503.2351
##        wtrd     wfir     wser   wmfg  wfed   wsta   wloc mix    pctymle
## 51 217.4908 342.4658 245.2061 448.42 442.2 340.39 386.12 0.1 0.07253495
```

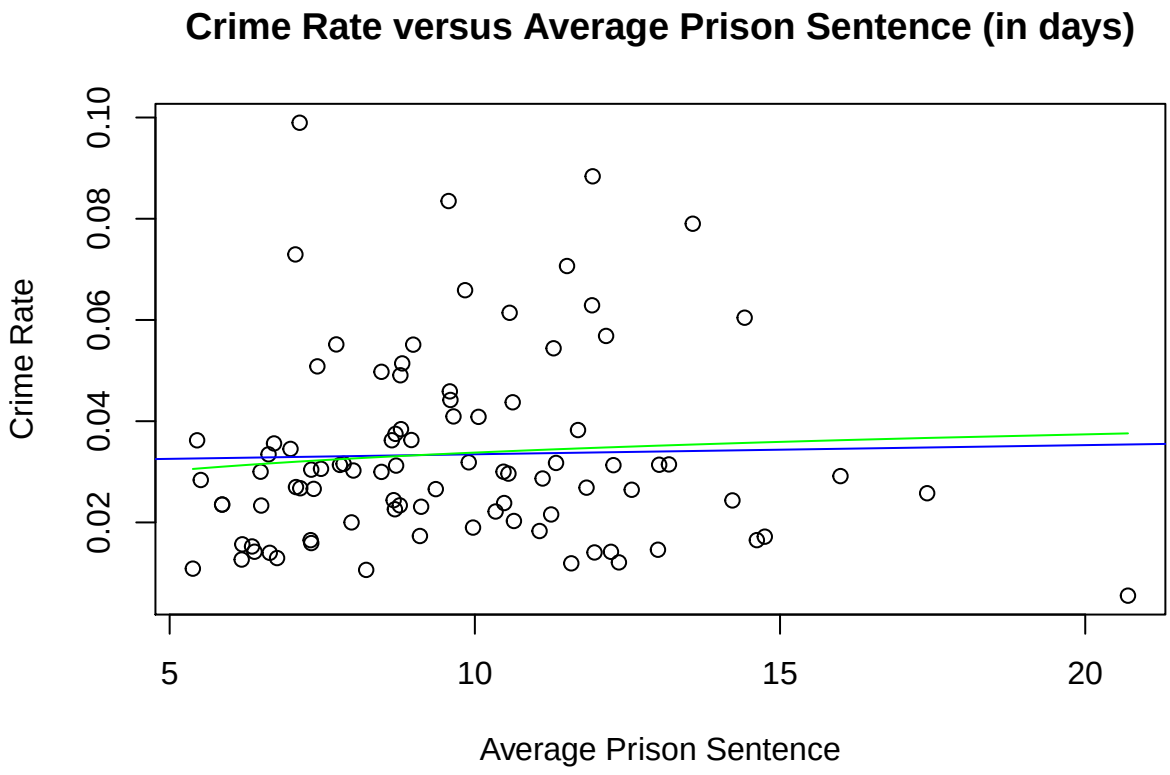
Hmm, if you remember, line 51, county 115 was the one with really low crime rate, and really high probability of arrest, now we found it also has the highest sentence length.

So, it looks like this one has more than 1 features that are different than others.

Take a look at plot of avgsen .vs. crime rate:

```
plot (crmrte ~ avgsen , data = df_crime_v2, xlab = "Average Prison Sentence", ylab = "Crime Rate", main
  avgsen_model <- lm (crmrte ~ avgsen, data = df_crime_v2)
  trn_avgsen = log(df_crime_v2$avgsen)
  avgsen_model_trn <- lm (crmrte ~ trn_avgsen , data = df_crime_v2)
```

```
  abline(avgsen_model, col="blue")
  lines(df_crime_v2$avgsen[order(df_crime_v2$avgsen)],avgsen_model_trn$fitted.values[order(df_crime_v2$a
```



There appears to have a rough positive correlation.

```
  avgsen_model$coefficients
```

```
##  (Intercept)       avgsen
## 0.0316530042 0.0001811194
```

```
  cat ("r^2 for linear model with $prbarr is:" , summary(avgsen_model)$r.squared, "\n")
```

```
## r^2 for linear model with $prbarr is: 0.0007513802
```

```
  avgsen_model_trn$coefficients
```

```
## (Intercept)  trn_avgsen
## 0.021787114 0.005216242
```

```
  cat ("r^2 for linear model with log($prbarr) is:" , summary(avgsen_model_trn)$r.squared, "\n")
```

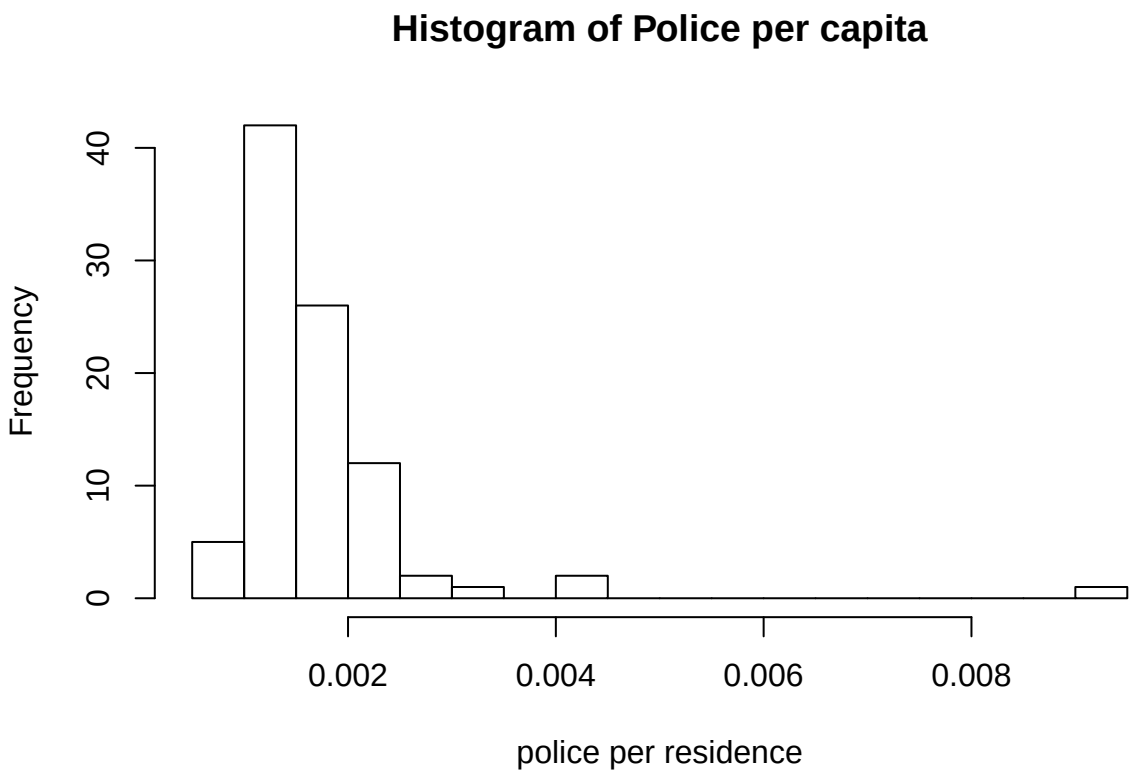
```
## r^2 for linear model with log($prbarr) is: 0.006158416
```

Like expected, although the slope is slightly positive, the r^2 is very small, and r^2 of the log version although is 8 times "better" still is negligible.

Police per capita

Police per capita is usually a community investment decision during political campaign.

```
hist(df_crime_v2$polpc, breaks = 13, xlab = "police per residence", main="Histogram of Police per capita
```



Looks like a left skew with an outlier again.

```
summary(df_crime_v2$polpc)
```

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

```
cat("The outlier (polpc > 0.008) is: ",which(df_crime_v2$polpc > 0.008))
```

```
## The outlier (polpc > 0.008) is:  51
```

Ah, it's that line 51 again.

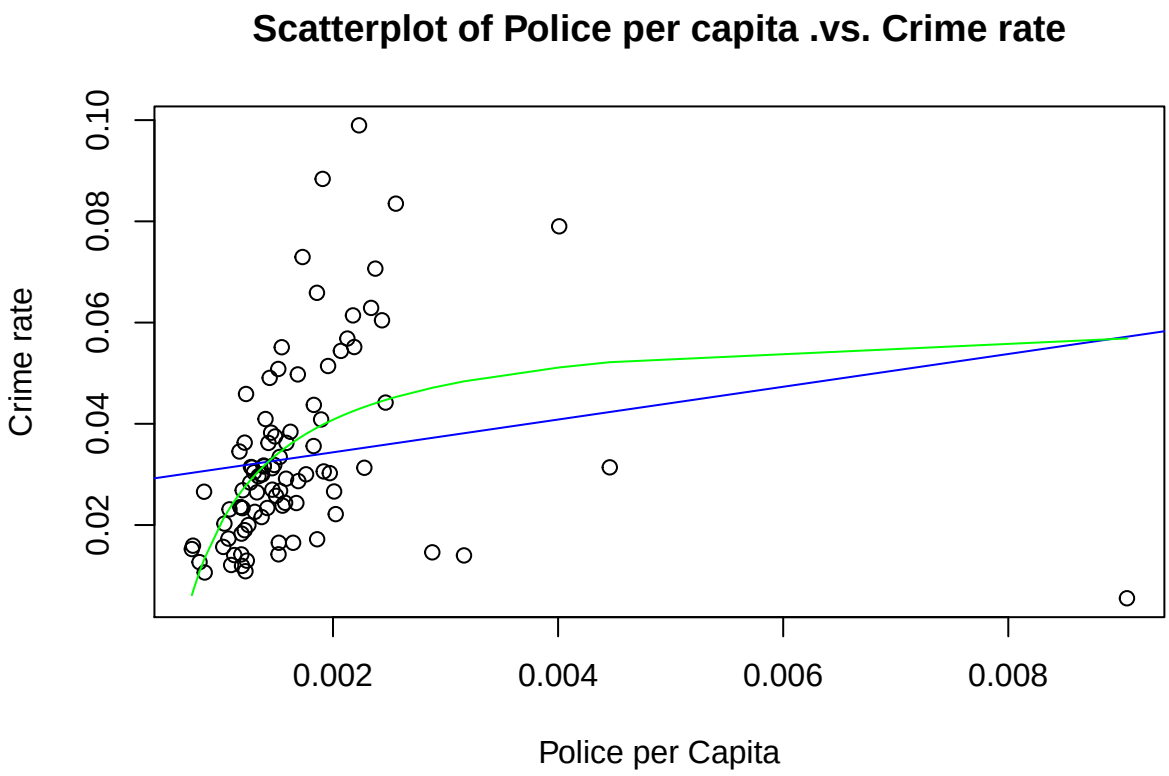
```
df_crime_v2[which(df_crime_v2$polpc > 0.008),]
```

```
##    county year    crmrte  prbarr prbconv prbpris avgsen      polpc
## 51    115   87 0.0055332 1.09091     1.5     0.5   20.7 0.00905433
##      density   taxpc west central urban pctmin80     wcon     wtuc
## 51 0.3858093 28.1931    1       0     0  1.28365 204.2206 503.2351
##        wtrd     wfir     wser   wmfg  wfed   wsta   wloc mix    pctymle
## 51 217.4908 342.4658 245.2061 448.42 442.2 340.39 386.12 0.1 0.07253495
```

Apparently the police to residents ratio in county #115 is way higher than other counties. Don't forget, it also has the highest prison time, highest arrest rate, and the lowest crime rate.

Take a look at plot of police per capita .vs. crime rate:

```
plot (crmrte ~ polpc, data = df_crime_v2, xlab="Police per Capita", ylab="Crime rate", main="Scatterplo
  polpc_model <- lm (crmrte ~ polpc, data = df_crime_v2)
  trn_polpc = 1/(df_crime_v2$polpc*1000)
  polpc_model_trn <- lm (crmrte ~ trn_polpc , data = df_crime_v2)
  abline(polpc_model, col="blue")
  lines(df_crime_v2$polpc[order(df_crime_v2$polpc)],polpc_model_trn$fitted.values[order(df_crime_v2$pol
```



Looks like a rough positive correlation between crime rate and police per capita. What is noteworthy is the outlier- it appears the highest polpc (to a lesser degree, actually 4 out of 5 highest polpc) shows (relatively) low crime rate.

```
  polpc_model$coefficients
```

```
## (Intercept)       polpc
##   0.0278888   3.2379068
```

```
  cat ("r^2 for linear model with $polpc is:" , summary(polpc_model)$r.squared, "\n")
```

```
## r^2 for linear model with $polpc is: 0.02886086
```

```
  polpc_model_trn$coefficients
```

```
## (Intercept)   trn_polpc
##  0.06141701 -0.04122766
```

```
  cat ("r^2 for linear model with 1/($polpc) is:" , summary(polpc_model_trn)$r.squared, "\n")
```

```
## r^2 for linear model with 1/($polpc) is: 0.2333191
```

Because Police per Capita is in reverse relationship with crime per capita, so we did a transformation to use 1/polpc*1000 as our transformed variable. The reason to multiple polpc by 1000 was because it was a police per 1000 residence value, without first multiple it by 1000, its inverse would be in the thousands, and skew the slop too much.

We can see that the inverse of polpc has the strongest r^2 so far, and is way better than polpc used directly.

One thing worth noting though is comparing the transformed (green) line and straight (blue) line, it is very likely that the outlier (county #115) is having a big influence on the curve of the green line.

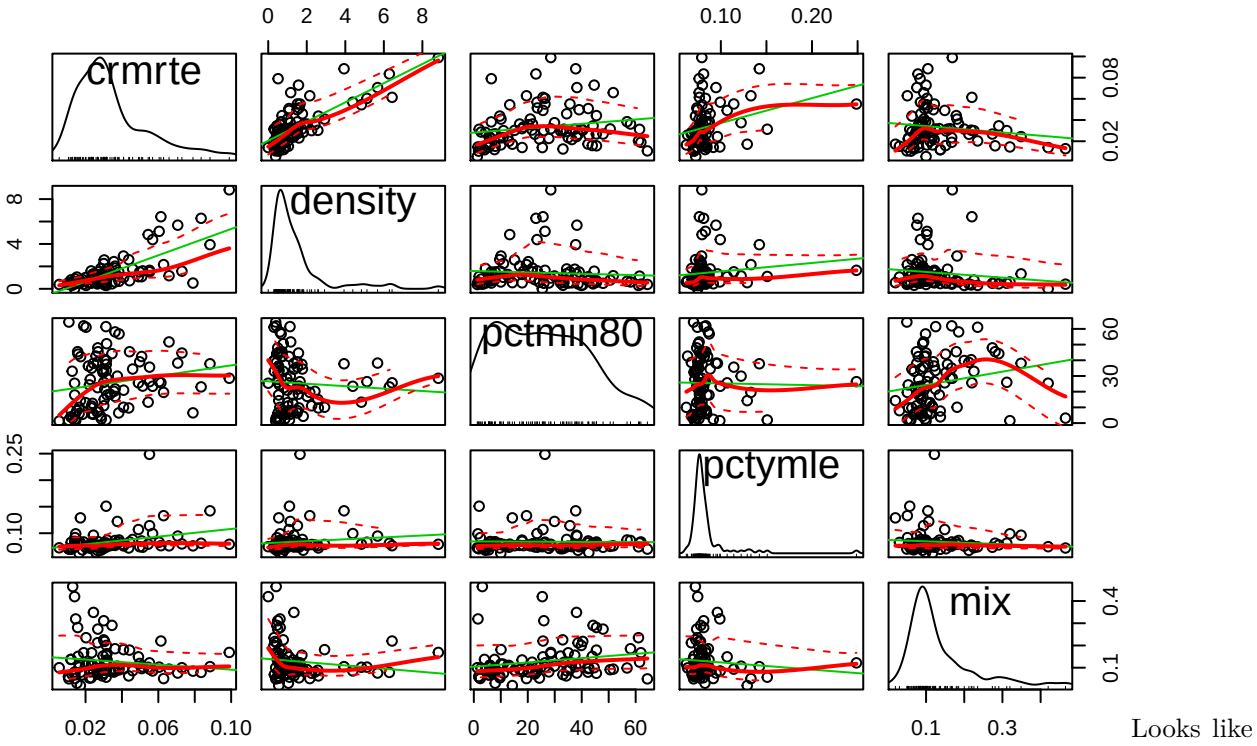
Demographic Variables

There are other demographic variables that at first glance are less related to policy making, but nevertheless maybe interesting to a political campaign.

population variables

We analyze the population density, male percentage, and minority variables here:

```
scatterplotMatrix(~ crmrte + density + pctmin80 + pctymle + mix, data=df_crime_v2 )
```



Looks like density is the only one shows a relatively positive trend. pctmin80 on the other hand looks like a non-linear relationship with center raises up, but two ends are lower. There seems to be an outlier for the young male data:

```
cat("The outlier for young male (pctymle > 0.2) is line#: " , which(df_crime_v2$pctymle > 0.2))
```

```
## The outlier for young male (pctymle > 0.2) is line#:  59
```

```
  density_model<-lm(crmrte ~ density, data = df_crime_v2)
  density_model$coefficients
```

```
## (Intercept)     density
## 0.020463225 0.009054221
```

```
  cat("r.square for density model is:", summary(density_model)$r.square, "\n\n")
```

```
## r.square for density model is: 0.5313873
```

```
  pctmin80_model<-lm(crmrte ~ log(pctmin80), data = df_crime_v2)
  pctmin80_model$coefficients
```

```
##   (Intercept) log(pctmin80)
##   0.015864267   0.006044989
```

```
  cat("r.square for log(minority) model is:",summary(pctmin80_model)$r.square, "\n\n")
```

```
## r.square for log(minority) model is: 0.09541619
```

```
  pctymle_model<-lm(crmrte~ log(pctymle), data = df_crime_v2)
  pctymle_model$coefficients
```

```
##  (Intercept) log(pctymle)
##   0.11086202   0.03097077
```

```
  cat("r.square for log(young male) model is:",summary(pctymle_model)$r.square, "\n\n")
```

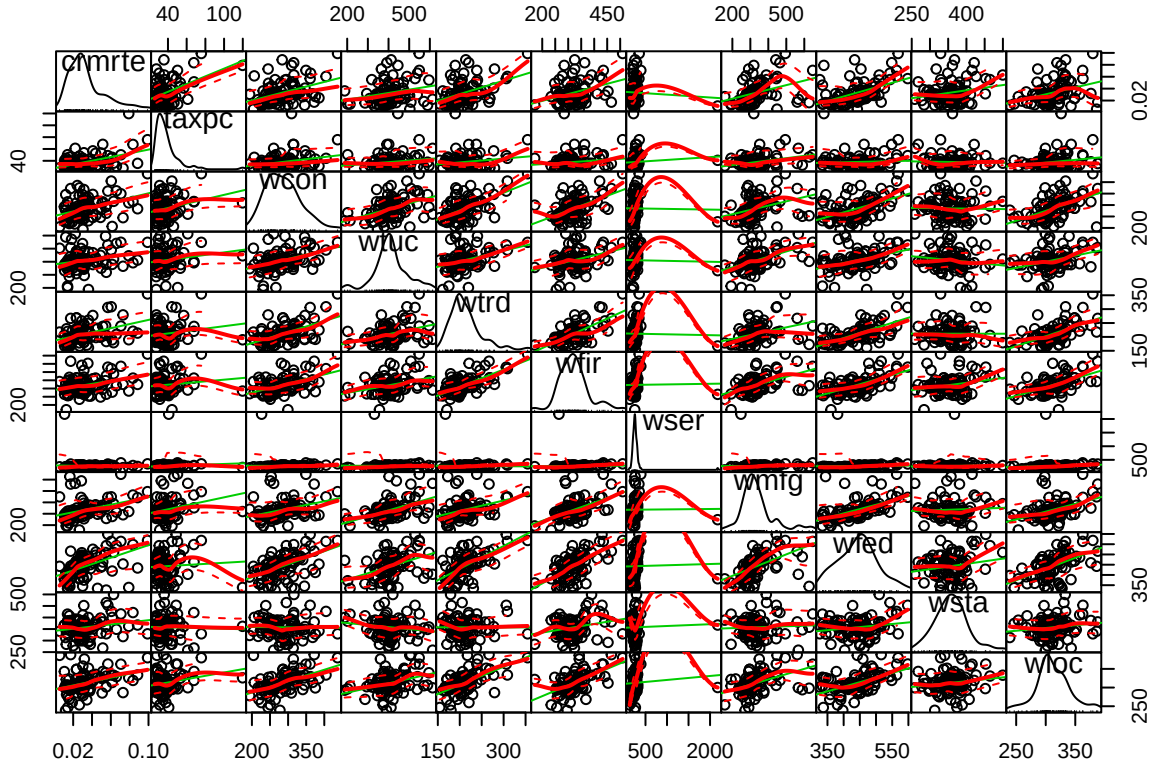
```
## r.square for log(young male) model is: 0.1056072
```

It looks like density has the highest correlation with crime rate so far.

Economical development variables

Here, we plot economical variables

```
scatterplotMatrix(~ crmrte + taxpc + wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc, data=
```



Interestingly, total tax per capita, wage trade, and wage Fed have obvious upward trend, with the wage fed showing clearer pattern. There seems an outlier for tax per capita:

```
cat("The outlier for tax (taxpc > 100) is line#: " , which(df_crime_v2$taxpc > 100))
```

```
## The outlier for tax (taxpc > 100) is line#:  25
```

```
cat("\nThe outlier for services wage (wser > 100) is line#: " , which(df_crime_v2$wser > 600))
```

```
##
## The outlier for services wage (wser > 100) is line#:  84
```

```
  taxpc_model<-lm(crmrte ~ taxpc, data = df_crime_v2)
  taxpc_model$coefficients
```

```
##  (Intercept)        taxpc
## 0.0087148195 0.0006486761
```

```
  cat("r.square for tax/person model is:", summary(taxpc_model)$r.square, "\n\n")
```

```
## r.square for tax/person model is: 0.2033828
```

```
  wfed_model<-lm(crmrte ~ wfed, data = df_crime_v2)
  wfed_model$coefficients
```

```
##   (Intercept)          wfed
## -0.0344698359  0.0001532399
```

```
  cat("r.square for fed wage model is:",summary(wfed_model)$r.square, "\n\n")
```

```
## r.square for fed wage model is: 0.2363474
```

```
  wtrd_model<-lm(crmrte ~ (wtrd), data = df_crime_v2)
  wtrd_model$coefficients
```

```
##   (Intercept)          wtrd
## -0.0142970959  0.0002254629
```

```
  cat("r.square for trade wage model is:",summary(wtrd_model)$r.square, "\n\n")
```

```
## r.square for trade wage model is: 0.1681866
```

Interestingly, the wage of federal workers in a county is correlated to crime rate! (correlation higher than police per capita).

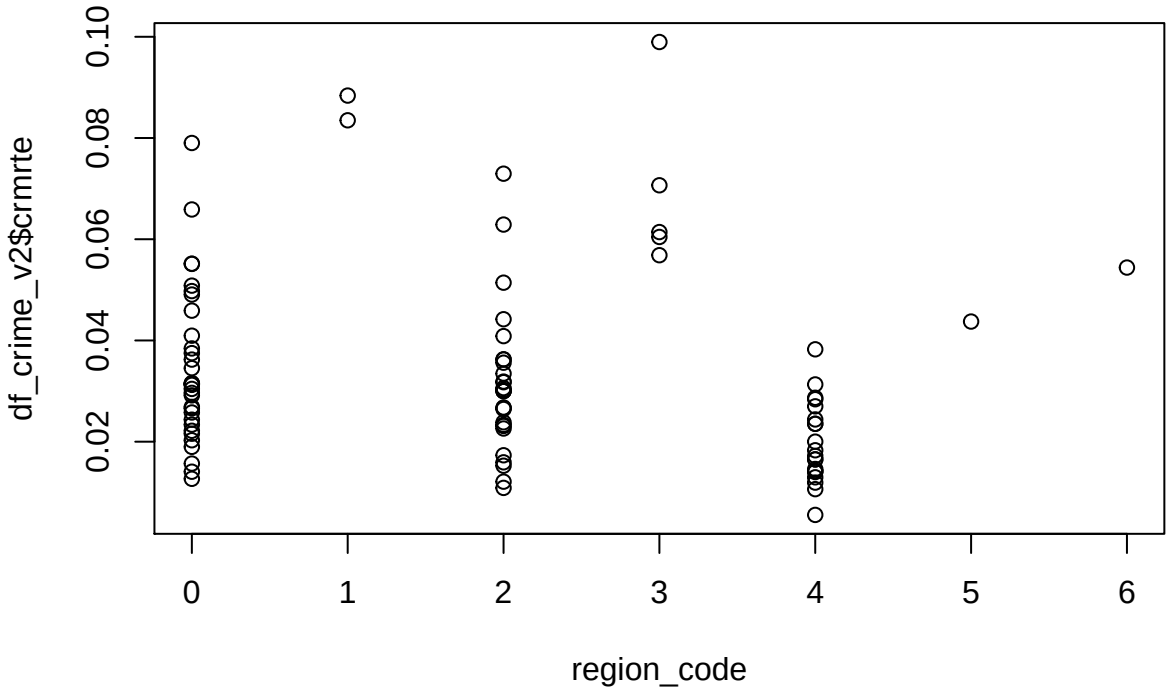
Geographical variables

We will also plot the geographical location information on a matrix. But before we do it, we will do a little transformation.

The geographical location was given as categorical multiple Boolean values. To plot them in one chart, we will combine them this way: 1) we will multiple urban by 1, 2) add central multiplied by 2, 3) add west multiplied by 4.

If we had more regions we would use logarithm, but since we have only 3, we will leave them as is.
This means we should only see counties that fall under either 1, or 2, or 4, if a county falls under 0, it would indicate it was not classified as any of the regions, and if a county falls under 3, it means it was classified as both urban and central. Similarly, if a county is under 5, 6, or 7, it means it was classified as both urban and west, both central and west, or all three.

```
region_code = (df_crime_v2$urban + 2*df_crime_v2$central + 4 * df_crime_v2$west)
plot (df_crime_v2$crmrte ~ region_code)
```



As expected, there are some encoding issues with geographical information. Lots of counties are not label as any of the 3 regions (hence shown on "0" in the above graph). And most urban counties are also labelled either central (on "3") or even west (on "5").

It looks like Urban has higher crime rate, not-any-of-the-3 has relatively higher rate than central, while west has obviously lower than all other categories.

Multicollinearity

There is no obvious col-linearity in the analysis so far (didn't see a scatter plot on perfect straight line). Nevertheless, we want to analyze if police per capita has much correlation with probability of arrest (cause more police means more likelihood to catch suspects?)

```
polpc_arr_model <- lm(prbarr ~ polpc, data = df_crime_v2)
summary(polpc_arr_model)$r.square
```

```
## [1] 0.1818519
```

So the r.square of bivariate OLS between prbarr and polpc is only 18%. This is far from perfect collinearity.

Multi variate OLS

```
inv_polpc = 1 / df_crime_v2$polpc
multi_var_model1 = lm(crmrte ~ inv_polpc+sqrt(prbconv)+log(prbarr) , data = df_crime_v2)
multi_var_model1
```

```
##
## Call:
## lm(formula = crmrte ~ inv_polpc + sqrt(prbconv) + log(prbarr),
##     data = df_crime_v2)
##
## Coefficients:
```

```
##   (Intercept)      inv_polpc  sqrt(prbconv)    log(prbarr)
##     5.276e-02     -3.356e-05     -3.740e-02     -2.309e-02
```

```
summary(multi_var_model1)$r.square
```

```
## [1] 0.5580921
```

```
multi_var_model2 = lm(crmrte ~ polpc+prbconv+prbarr+density+pctymle+pctmin80+taxpc, data = df_crime_v2)
multi_var_model2
```

```
##
## Call:
## lm(formula = crmrte ~ polpc + prbconv + prbarr + density + pctymle +
##     pctmin80 + taxpc, data = df_crime_v2)
##
## Coefficients:
## (Intercept)        polpc      prbconv       prbarr      density
##   0.0180024    6.5103236   -0.0188310   -0.0557366    0.0054600
##     pctymle     pctmin80        taxpc
##   0.0875755    0.0003518    0.0001843
```

```
summary(multi_var_model2)$r.square
```

```
## [1] 0.8230153
```

```
multi_var_model3 = lm(crmrte ~ polpc+prbconv+prbarr+density+pctmin80, data = df_crime_v2)
multi_var_model3
```

```
##
## Call:
## lm(formula = crmrte ~ polpc + prbconv + prbarr + density + pctmin80,
##     data = df_crime_v2)
##
## Coefficients:
## (Intercept)        polpc      prbconv       prbarr      density
##   0.0335711    8.1113972   -0.0216026   -0.0659367    0.0055514
##    pctmin80
##   0.0003705
```

```
summary(multi_var_model3)$r.square
```

```
## [1] 0.805648
```

We built 3 Multivariate OLS models.
The first one has the 3 originally conceived variables: Police per capita, probability of arrest and probability of conviction (certainty of punishment).

The second and third model, on the other hand, removed transformations, because some of the non-linear effect would be shown through multiple relationship across variables, for example, the probability of arrest should have some relationship to police per capita, and density also influence police per capita and also probability of arrest.

The second model includes demographic and economical variables. The third one we removed less influential variables, and kept only density and pctmin80 in the model.

The first model has an r.square of 54.5% which means it explains more than half of the variate.

The second model has an r.square of 82.3%.

The third model has an r.square of 80.6%. We found pctymle and taxpc combined only contributed to less than 1.7% of the increase in r.square. On the other hand, if we remove pctmin80, there would be a more

than 10% drop in r.squared. ## comparing models

```
stargazer(multi_var_model1,multi_var_model2, multi_var_model3, header=FALSE, type='latex', title="Regres
```

Table 1: Regression Results

	Dependent variable:		
	Crime Rate		
	(1)	(2)	(3)
1/PolicePerCap	-0.00003^{***}		
	(0.00001)		
sqrt(Prob of Conviction)	-0.037^{***}		
	(0.007)		
log(Prob of Arrest)	-0.023^{***}		
	(0.003)		
Police/Capita		6.510^{***}	8.111^{***}
		(1.260)	(1.164)
Prob of Conviction		-0.019^{***}	-0.022^{***}
		(0.003)	(0.003)
Prob of Arrest		-0.056^{***}	-0.066^{***}
		(0.009)	(0.008)
Population Density		0.005^{***}	0.006^{***}
		(0.001)	(0.001)
Young Male		(0.041)	
Minority		(0.0001)	(0.0001)
Tax/Capita		0.0002^{**}	
		(0.0001)	
Constant	0.053^{***}	0.018^{***}	0.034^{***}
	(0.007)	(0.006)	(0.003)
Observations	91	91	91
R^2	0.558	0.823	0.806
Adjusted R^2	0.543	0.808	0.794
Note:			*p<0.1; **p<0.05; ***p<0.01

Casual effect analysis

So it turns out in high density and high minority counties, crime is a more pressing issue, and while increase probability of arrest and conviction has a damping effect on crime, the police per capita increases as crime rate rises. This was not a surprise, although it did not work as independent variable as we hoped.

While crime is not a controlled experiment, not all policy related covariates are independent variables - because previous policies have already adjusted to previous crimes, so the causal effect is not as straight forward as one would hope. For example, we expected police per capita to be a cause of dropping crime, but it turns out it was more (likely) a result of higher crime rates.

Our model supported the hypothesis that probability of conviction and probability of arrest have an impact to crime rate.