# DRAFT: Lab 3: Reducing Crime

*C. Akkineni, A. Thorp, K. Hanna*

*November 27, 2018*

## Contents

**Stage 1: Draft Report** **2**

**Preliminary Infomations** **4**

Variables: . . . . . 4

Steps for evaluating variables . . . . . 6

**1. What do you want to measure? Make sure you identify variables that will be relevant to the concerns of the political campaign.** **7**

**2. What transformations should you apply to each variable? This is very important because transformations can reveal linearities in the data, make our results relevant, or help us meet model assumptions.** **7**

**3. Are your choices supported by EDA? You will likely start with some general EDA to detect anomalies (missing values, top-coded variables, etc.). From then on, your EDA should be interspersed with your model building. Use visual tools to guide your decisions.** **7**

**4. What covariates help you identify a causal effect? What covariates are problematic, either due to multicollinearity, or because they will absorb some of a causal effect you want to measure?** **7**

## Stage 1: Draft Report

```
crime = read.csv('crime_v2.csv')
# Delete the 6 empty rows at the end
crime[92:100,]
```

```
##      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
## NA       NA   NA     NA     NA      <NA>      NA     NA    NA       NA    NA
## NA.1     NA   NA     NA     NA      <NA>      NA     NA    NA       NA    NA
## NA.2     NA   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
## NA     NA       NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## NA.1   NA       NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## NA.2   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
## NA     NA  NA      NA
## NA.1   NA  NA      NA
## NA.2   NA  NA      NA
```

```
crime = crime[1:91, ]

# Convert columns to factors and logical.
crime$county = as.factor(crime$county)
crime$year = as.factor(crime$year)
crime$west = as.logical(crime$west)
crime$central = as.logical(crime$central)
crime$urban = as.logical(crime$urban)

# Fix prbconv, convert from factor to numeric
summary(crime$prbconv)
```

```
##                       ` 0.068376102 0.140350997 0.154451996 0.203724995
##           0           0           1           1           1           1
## 0.207830995    0.220339 0.226361006 0.229589999 0.248275995 0.259833008
##           1           1           1           1           1           1
## 0.267856985 0.271946996  0.28947401 0.300577998 0.308411002 0.314606994
```

```
##           1           1           1           1           1           1
## 0.322580993 0.325300992 0.327868998 0.328664005 0.334701002 0.340490997
##           1           1           1           1           1           1
## 0.343023002 0.347799987 0.352941006  0.36015299 0.364353001 0.371879011
##           1           1           1           1           1           1
##    0.381908 0.384236008 0.385495991 0.386925995 0.393413007    0.401198
##           1           1           1           1           1           1
## 0.403780013 0.406780005 0.410596013 0.412698001 0.426777989 0.436441004
##           1           1           1           1           1           1
## 0.438960999 0.443114012 0.443681002 0.449999988 0.450567007 0.452829987
##           1           1           1           1           1           1
## 0.457210004 0.459215999 0.468531013 0.476563007 0.477732986 0.492940009
##           1           1           1           1           1           1
## 0.493438005 0.495575011  0.50819701 0.515464008 0.520606995 0.520709991
##           1           1           1           1           1           1
## 0.522387981 0.525424004 0.527595997 0.528302014 0.548494995 0.549019992
##           1           1           1           1           1           1
## 0.559822977 0.571429014 0.573943973 0.588859022 0.589905024 0.595077991
##           1           1           1           2           1           1
##  0.62251699 0.722972989 0.736908972 0.739394009 0.763333023 0.769231021
##           1           1           1           1           1           1
## 0.781608999 0.793232977 0.909090996 0.972972989 1.015380025 1.068969965
##           1           1           1           1           1           1
## 1.182929993 1.225610018 1.234380007 1.358139992 1.481480002         1.5
##           1           1           1           1           1           1
## 1.670519948 2.121210098
##           1           1
```

```
crime$prbconv = as.numeric(crime$prbconv)

# county 193 is duplidated, remove one
crime[crime$county == 193, ]
```

```
##    county year    crmrte   prbarr prbconv  prbpris avgsen      polpc
## 88    193   87 0.0235277 0.266055      70 0.423423   5.86 0.00117887
## 89    193   87 0.0235277 0.266055      70 0.423423   5.86 0.00117887
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 88 0.8138298 28.51783 TRUE   FALSE FALSE  5.93109 285.8289 480.1948
## 89 0.8138298 28.51783 TRUE   FALSE FALSE  5.93109 285.8289 480.1948
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc       mix
## 88 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.1105016
## 89 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.1105016
##       pctymle
## 88 0.07819394
## 89 0.07819394
```

```
crime = crime[-c(89), ]

summary(crime)
```

```
##      county   year        crmrte             prbarr           prbconv
##  1      : 1   87:90   Min.   :0.005533   Min.   :0.09277   Min.   : 3.00
##  3      : 1           1st Qu.:0.020604   1st Qu.:0.20495   1st Qu.:25.25
##  5      : 1           Median :0.030002   Median :0.27146   Median :47.50
##  7      : 1           Mean   :0.033510   Mean   :0.29524   Mean   :47.50
```

```
##  9       : 1           3rd Qu.:0.040249   3rd Qu.:0.34487   3rd Qu.:69.75
##  11      : 1           Max.   :0.098966   Max.   :1.09091   Max.   :92.00
##  (Other):84
##     prbpris           avgsen           polpc              density
##  Min.   :0.1500   Min.   : 5.380   Min.   :0.0007459   Min.   :0.00002
##  1st Qu.:0.3642   1st Qu.: 7.375   1st Qu.:0.0012378   1st Qu.:0.54718
##  Median :0.4222   Median : 9.110   Median :0.0014897   Median :0.97925
##  Mean   :0.4106   Mean   : 9.689   Mean   :0.0017080   Mean   :1.43567
##  3rd Qu.:0.4576   3rd Qu.:11.465   3rd Qu.:0.0018856   3rd Qu.:1.56926
##  Max.   :0.6000   Max.   :20.700   Max.   :0.0090543   Max.   :8.82765
##
##      taxpc           west          central          urban
##  Min.   : 25.69   Mode :logical   Mode :logical   Mode :logical
##  1st Qu.: 30.73   FALSE:68        FALSE:56        FALSE:82
##  Median : 34.92   TRUE :22        TRUE :34        TRUE :8
##  Mean   : 38.16
##  3rd Qu.: 41.01
##  Max.   :119.76
##
##     pctmin80           wcon            wtuc            wtrd
##  Min.   : 1.284   Min.   :193.6   Min.   :187.6   Min.   :154.2
##  1st Qu.:10.024   1st Qu.:250.8   1st Qu.:374.3   1st Qu.:190.7
##  Median :24.852   Median :281.2   Median :404.8   Median :203.0
##  Mean   :25.713   Mean   :285.4   Mean   :410.9   Mean   :210.9
##  3rd Qu.:38.183   3rd Qu.:315.0   3rd Qu.:440.7   3rd Qu.:224.3
##  Max.   :64.348   Max.   :436.8   Max.   :613.2   Max.   :354.7
##
##       wfir            wser            wmfg            wfed
##  Min.   :170.9   Min.   : 133.0   Min.   :157.4   Min.   :326.1
##  1st Qu.:285.6   1st Qu.: 229.3   1st Qu.:288.6   1st Qu.:398.8
##  Median :317.1   Median : 253.1   Median :321.1   Median :448.9
##  Mean   :321.6   Mean   : 275.3   Mean   :336.0   Mean   :442.6
##  3rd Qu.:342.6   3rd Qu.: 277.6   3rd Qu.:359.9   3rd Qu.:478.3
##  Max.   :509.5   Max.   :2177.1   Max.   :646.9   Max.   :598.0
##
##       wsta            wloc            mix            pctymle
##  Min.   :258.3   Min.   :239.2   Min.   :0.01961   Min.   :0.06216
##  1st Qu.:329.3   1st Qu.:297.2   1st Qu.:0.08060   1st Qu.:0.07437
##  Median :358.4   Median :307.6   Median :0.10095   Median :0.07770
##  Mean   :357.7   Mean   :312.3   Mean   :0.12905   Mean   :0.08403
##  3rd Qu.:383.2   3rd Qu.:328.8   3rd Qu.:0.15206   3rd Qu.:0.08352
##  Max.   :499.6   Max.   :388.1   Max.   :0.46512   Max.   :0.24871
##
```

# Preliminary Infomations

## Variables:

1. Target

- crmrte

2. Label

- county

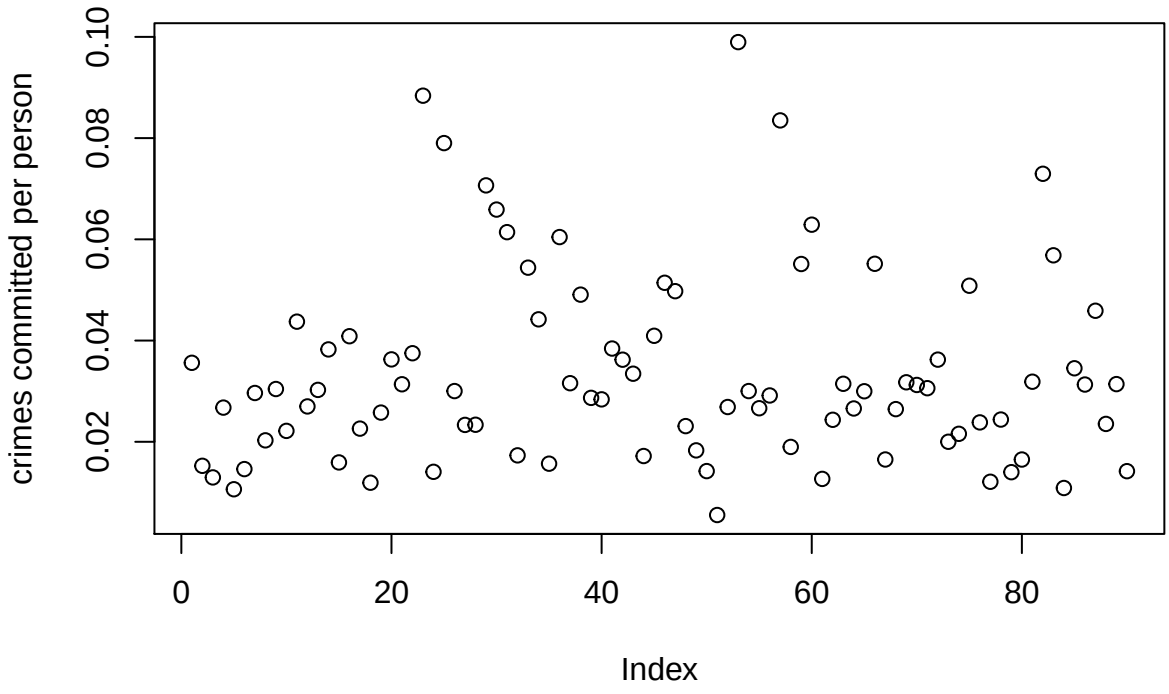
3. Segregates: (my own word, just things that can segregate the data). Counties can belong to 0 or more from west, central and urban.

- density (likely related to others, especially urban)
- west
- central
- urban

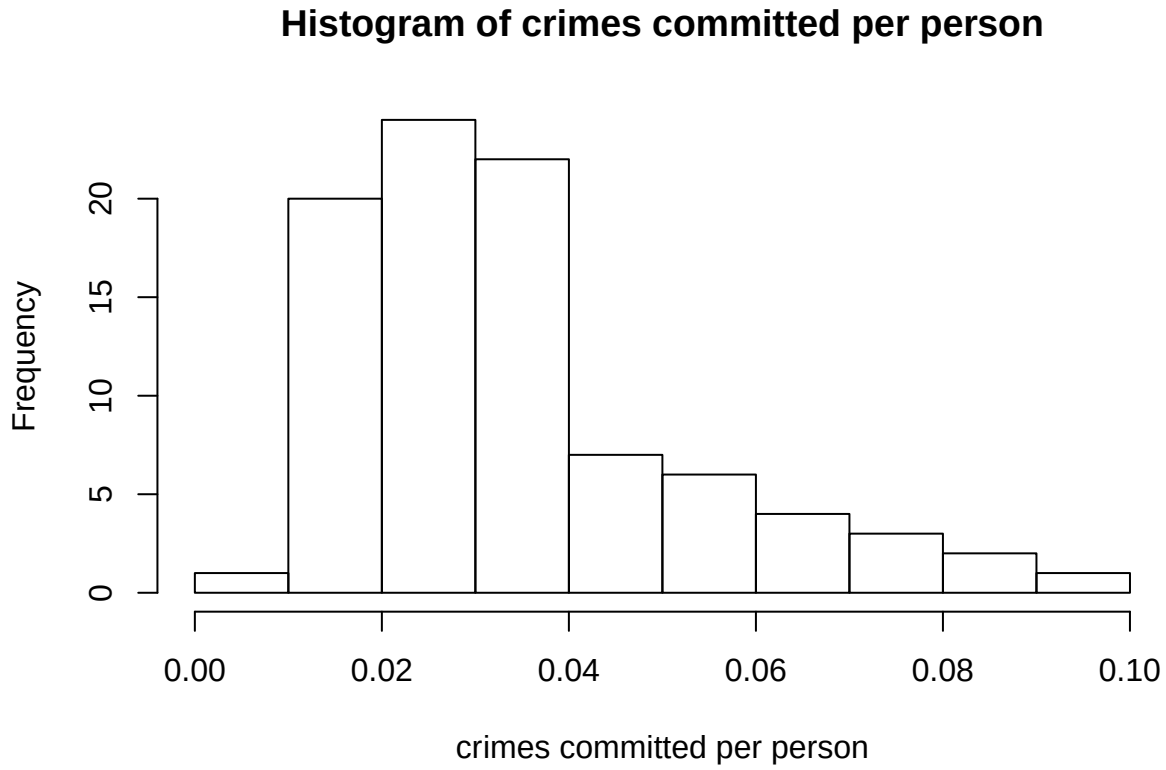
4. Cost of doing crime:

- prbconv
- prbpris
- avgsen
- prbarr
- polpc (likely related to prbconv)

```
plot(crime$crmrte, ylab = 'crimes committed per person')
```



```r
hist(crime$crmrte, xlab = 'crimes committed per person', main = 'Histogram of crimes committed per perso
```




```r
model1 = lm(crmrte ~ prbarr + polpc + density, data = crime)
(model1$coefficients)
```

```
##  (Intercept)       prbarr        polpc      density
##  0.028231799 -0.040443974  3.738309135  0.007546142
```

### Steps for evaluating variables

Leverage (and Influence if required)
Goodness-of-Fit : AIC
Endoginaity
Omitted variable bias
MSE
E[theta hat] = theta

```r
sum(c(crime$urban, crime$west))
```

```
## [1] 30
```

**1. What do you want to measure? Make sure you identify variables that will be relevant to the concerns of the political campaign.**

**2. What transformations should you apply to each variable? This is very important because transformations can reveal linearities in the data, make our results relevant, or help us meet model assumptions.**

**3. Are your choices supported by EDA? You will likely start with some general EDA to detect anomalies (missing values, top-coded variables, etc.). From then on, your EDA should be interspersed with your model building. Use visual tools to guide your decisions.**

**4. What covariates help you identify a causal effect? What covariates are problematic, either due to multicollinearity, or because they will absorb some of a causal effect you want to measure?**