Lab 3: Reducing Crime

Zach Merritt

March 31, 2018

I. Introduction

I have been hired to provide research for a political campaign regarding crime in North Carolina in 1988. In this report, I identify the determinants of crime to be used for future policy decisions and campaign platforms.

```r
#set up
rm(list = ls())

#libraries
library(tidyverse)
library(forcats)
library(ggthemes)
library(maps)
library(moments)
library(corrplot)
library(gridExtra)
library(purrr)
library(stargazer)

#data
crime <- read.csv(file="../01_raw_data/crime_v2.csv", stringsAsFactors = FALSE, na = c(""," ", "<NA>", "NA","`"))
schema <- read.csv(file="../01_raw_data/crime_schema.csv", stringsAsFactors = FALSE, na = c(""," ", "<NA>", "NA","`"))
```

II. Discuss variables

The crime data set I have been provided contained 97 observations with 25 variables. However, not each observation was complete which I discuss in more detail in section III.

Please see the table below, which shows each variable in the data set and the provided description of each variable. For my eventual model creation, I treat 'crmrte' (or 'crimes committed per person') as the dependent variable and all other variables as independent variables (except for county and year). I do not examine county in the model because there is nothing inherently identifying about the county number (the county is only important for what data points it contains). Finally, I do not examine year because it is always the same on 1987.

```
schema
```

```
##         var                             desc
## 1    county                county identifier
## 2      year                             1987
## 3    crmrte      crimes committed per person
## 4    prbarr          ‘probability’ of arrest
## 5   prbconv      ‘probability’ of conviction
## 6   prbpris ‘probability’ of prison sentence
## 7    avgsen              avg. sentence, days
## 8     polpc                police per capita
## 9   density              people per sq. mile
## 10    taxpc           tax revenue per capita
## 11     west            =1 if in western N.C.
## 12  central            =1 if in central N.C.
## 13    urban                    =1 if in SMSA
## 14 pctmin80             perc. minority, 1980
## 15     wcon        weekly wage, construction
## 16     wtuc     wkly wge, trns, util, commun
## 17     wtrd  wkly wge, whlesle, retail trade
## 18     wfir     wkly wge, fin, ins, real est
## 19     wser       wkly wge, service industry
## 20     wmfg          wkly wge, manufacturing
## 21     wfed          wkly wge, fed employees
## 22     wsta        wkly wge, state employees
## 23     wloc         wkly wge, local gov emps
## 24      mix  offense mix: face-to-face/other
## 25  pctymle               percent young male
```

III. Identify Missing Data

In the data set there are six observations which are completely missing information for *every* variable. Please see the code below which identifies these observations and removes them from the data set. After removing these observations, my data set contains 91 observations.

```
#A. Count missing data
crime %>%
  summarise_all(funs(sum(is.na(.))))
```

```
##   county year crmrte prbarr prbconv prbpris avgsen polpc density taxpc
## 1      6    6      6      6       6       6      6     6       6     6
##   west central urban pctmin80 wcon wtuc wtrd wfir wser wmfg wfed wsta wloc
## 1    6       6     6        6    6    6    6    6    6    6    6    6    6
##   mix pctymle
## 1   6       6
```

```
#B. Subset to any missing data
crime %>%
  filter_all(any_vars(is.na(.)))
```

```
##   county year crmrte prbarr prbconv prbpris avgsen polpc density taxpc
## 1     NA   NA     NA     NA      NA      NA     NA    NA      NA    NA
## 2     NA   NA     NA     NA      NA      NA     NA    NA      NA    NA
## 3     NA   NA     NA     NA      NA      NA     NA    NA      NA    NA
## 4     NA   NA     NA     NA      NA      NA     NA    NA      NA    NA
## 5     NA   NA     NA     NA      NA      NA     NA    NA      NA    NA
## 6     NA   NA     NA     NA      NA      NA     NA    NA      NA    NA
##   west central urban pctmin80 wcon wtuc wtrd wfir wser wmfg wfed wsta wloc
## 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   NA
## 3   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 4   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 5   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 6   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##   mix pctymle
## 1  NA      NA
## 2  NA      NA
## 3  NA      NA
## 4  NA      NA
## 5  NA      NA
## 6  NA      NA
```

```
  #Note: Clearly this data is completely erroneous and useless, so I'm going to delete the six rows

#C. Subset crime to not contain six rows

crime <- crime %>%
  filter(!is.na(county))

#Just that one filter takes care of everything, see:
```

```r
crime %>%
  summarise_all(funs(sum(is.na(.))))
```

```
##   county year crmrte prbarr prbconv prbpris avgsen polpc density taxpc
## 1      0    0      0      0       0       0      0     0       0     0
##   west central urban pctmin80 wcon wtuc wtrd wfir wser wmfg wfed wsta wloc
## 1    0       0     0        0    0    0    0    0    0    0    0    0    0
##   mix pctymle
## 1   0       0
```

IV. Create choropleth Maps

As each observation is a county, I thought it a valuable exercise to create choropleth maps to generally explore North Carolina and the prevalence of crime and crime related variables around the state. The provided data set identifies counties using a numerical code. Because I have some experience with United States county-related data, I hypothesized that these codes were official 'FIPS' ('Federal Information Processing Standard') codes. I used the R-data package "maps" to bring in mapping data for North Carolina (which contains FIPS codes). I was able to confirm my hypothesis by create a choropleth-type map with the location variables in the data set, West and Central. In this process, I learned that our data set is incomplete and does not cover 10 counties in North Carolina (I output the list of these 10 counties below).

```r
#A. Load Data

data("county.fips")
nc <- map_data("county") %>%
  filter(region == "north carolina")

#B. Structure nc_fips
nc_fips <- county.fips %>%
  separate(polyname, c("state", "subregion"), sep =",") %>%
  separate(subregion, c("subregion", "sub_subregion"), sep =":", fill="right") %>%
  select(-sub_subregion) %>%
  distinct() %>%
  filter(state == "north carolina") %>%
  full_join(
    nc,
    by = "subregion"
  ) %>%
  mutate(county = as.numeric(substr(fips, 3, 5)))
```

```
#C. Create mapping data frame

crime_map <- left_join(
  crime,
  nc_fips,
  by = "county"
) %>%
  select(-year)

#Note: there are 10 counties which are not in our crime data
anti_join(
  nc_fips,
  crime,
  by = "county"
) %>%
  distinct(county, subregion)
```

```
##    subregion county
## 1     camden     29
## 2   carteret     31
## 3       clay     43
## 4      gates     73
## 5     graham     75
## 6       hyde     95
## 7      jones    103
## 8   mitchell    121
## 9    tyrrell    177
## 10    yancey    199
```

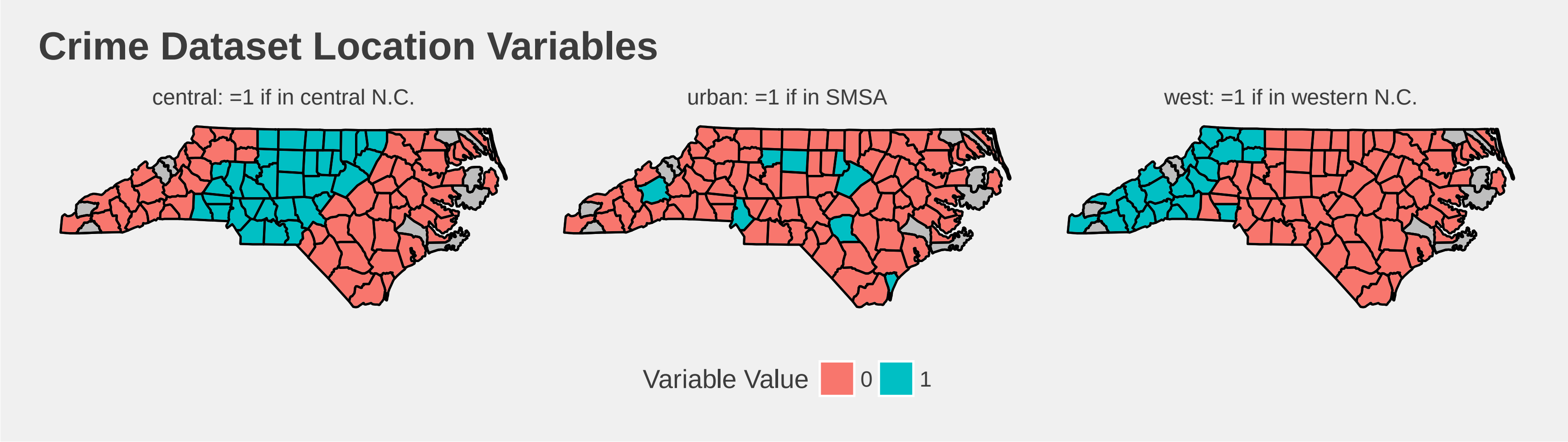
```
#D. Create Crime Map Long (strictly for facet_wrap capabilities)

crime_map_long <- crime_map %>%
  gather(var, val, -county, -fips, -state, -subregion,-long, -lat, -group, -order, -region) %>%
  group_by(var) %>%
  mutate(`Variable Percentile` = percent_rank(val)) %>%
  left_join(
    schema,
    by = "var"
  )
```

```
#E. Create Percentil Density Maps

#1. Set up
  #a. base
  nc_base <- ggplot(data = nc, mapping = aes(x = long, y = lat, group = group)) +
    coord_fixed(1.3) +
    geom_polygon(color = "black", fill = "gray")

  #b. mapping neccesities
  ditch_the_axes <- theme(
    axis.text = element_blank(),
    axis.line = element_blank(),
    axis.ticks = element_blank(),
    panel.border = element_blank(),
    panel.grid = element_blank(),
    axis.title = element_blank()
  )
```

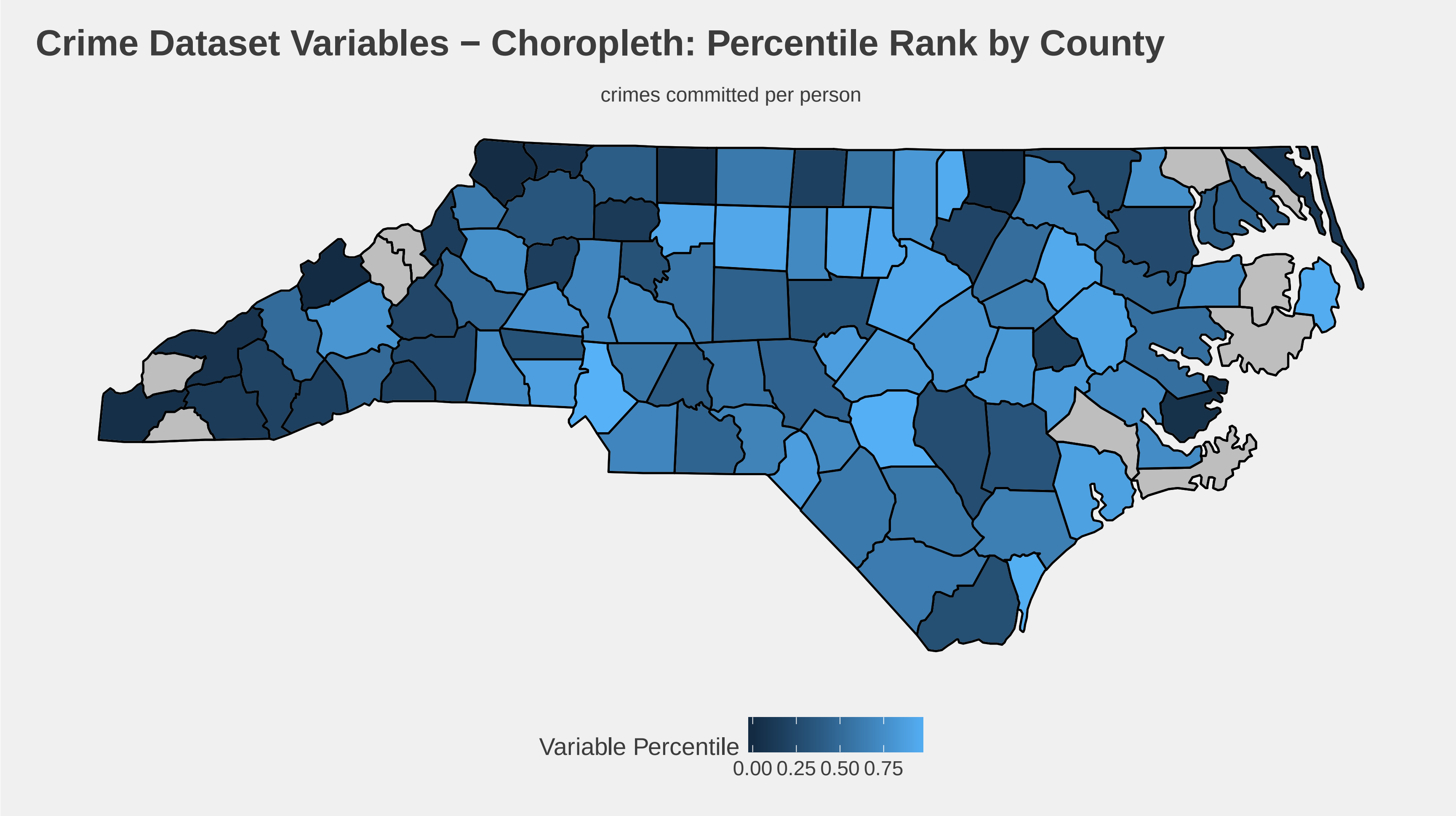
Map 1: Choropleth of Crime Rate and the Location Variables

See the maps below which highlight where the West, Central, and Urban areas of North Carolina are. In addition, see the choropleth of the crime rate. There are higher rates of crime in the Urban areas, but this is a concept we will analyze in greater detail later.

```
#2. Create Plots
 a <- nc_base +
    geom_polygon(data = crime_map_long %>%
                   mutate(desc = paste0(var,": ",desc)) %>%
                   filter(var %in% c("west","central","urban")) %>%
                   mutate(`Variable Value` = factor(val)), aes(fill = `Variable Value`), color = "white") +
    geom_polygon(color = "black", fill = NA)  +
    theme_bw() +
    facet_wrap(~desc) +
    theme_fivethirtyeight() +
    ditch_the_axes +
    ggtitle("Crime Dataset Location Variables")

  b<- nc_base +
    geom_polygon(data = crime_map_long %>%
                   filter(var %in% c("crmrte")), aes(fill = `Variable Percentile`), color = "white") +
    geom_polygon(color = "black", fill = NA) +
    theme_bw() +
    facet_wrap(~desc) +
    theme_fivethirtyeight() +
    ditch_the_axes +
    ggtitle("Crime Dataset Variables - Choropleth: Percentile Rank by County")

  a
```


Crime Dataset Location Variables
central: =1 if in central N.C.
urban: =1 if in SMSA
west: =1 if in western N.C.
Variable Value
0
1

```
b
```



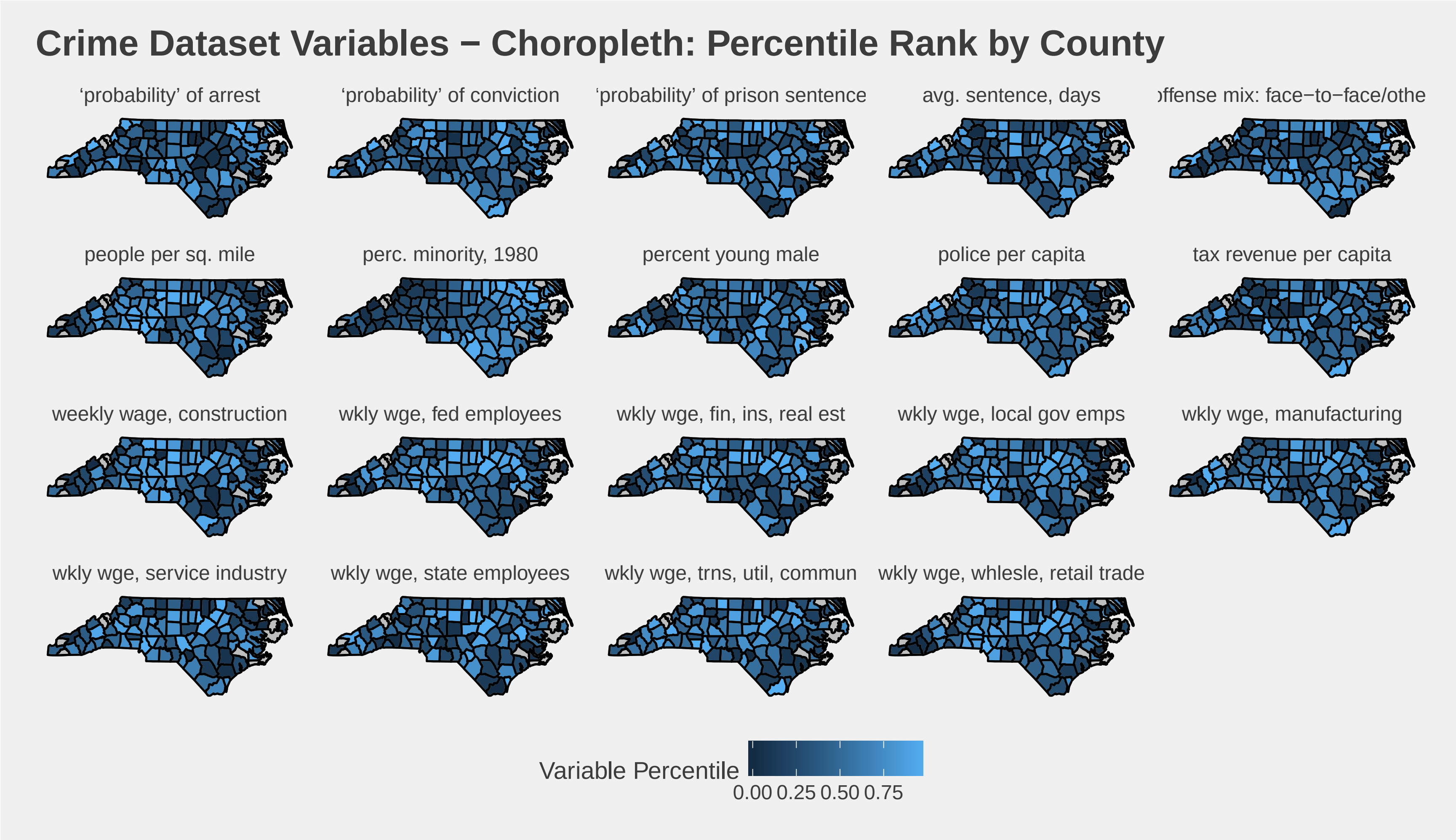
Crime Dataset Variables − Choropleth: Percentile Rank by County

crimes committed per person

Map 2: Choropleth of all Independent Variables (besides location variables)

This series of maps is an interesting exploration of each variable. At an initial glance, it may be interpreted as nothing more than an art project. However, this series serves as a valuable tool to refer back too as we explore certain determinants of crime in more detail. In addition, if you are a stranger to North Carolina this map may seem a tad overwhelming (imagine how much more informational it would be if it was of your home state). My client is very familiar with North Carolina and this will serve as valuable tool.

```r
#b. show variable percentiles
nc_base +
  geom_polygon(data = crime_map_long %>%
                 filter(!var %in% c("west","central","urban", "crmrte")), aes(fill = `Variable Percentile`), color = "white") +
  geom_polygon(color = "black", fill = NA) +
  theme_bw() +
  facet_wrap(~desc) +
  theme_fivethirtyeight() +
  ditch_the_axes +
  ggtitle("Crime Dataset Variables - Choropleth: Percentile Rank by County")
```



Crime Dataset Variables − Choropleth: Percentile Rank by County
‘probability’ of arrest
‘probability’ of conviction
‘probability’ of prison sentence
avg. sentence, days
offense mix: face−to−face/othe
people per sq. mile
perc. minority, 1980
percent young male
police per capita
tax revenue per capita
weekly wage, construction
wkly wge, fed employees
wkly wge, fin, ins, real est
wkly wge, local gov emps
wkly wge, manufacturing
wkly wge, service industry
wkly wge, state employees
wkly wge, trns, util, commun
wkly wge, whlesle, retail trade
Variable Percentile
0.00
0.25
0.50
0.75

V. Examine Distribution of Variables

In this section, I examine the distribution of the independent variables and crime rate.

First, I log transform every variable (including crime rate) and I measure the skewness and kurtosis of the distribution of the variable before and after the log transformation (‘applied_no_trans’ and ‘log_trans’, respectively). I then output a table of the results which adds a flag of TRUE/FALSE if the log transformation decreased the absolute value of the skewness (i.e., if the log transformation, roughly speaking, made the distribution of the variable *more* normal). While I will not discuss the benefits of normally distributed variables in this draft report, this is an important step in multiple linear regression.

Second, I output three series of plots: * 1) Crime Data set Variables - Univariate Distribution for Variables with Decreased Skewness from Log Transformation * 2) Crime Data set Variables - Univariate Distribution for Variables with Increased Skewness from Log Transformation * 3) Crime Rate - Univariate Distribution with and without Log Transformation

Finally, I modified the crime data set and log transformed any variables which experienced a decreased in skewness (or an increase in normality) from the log transformation, I hold on to a copy of the completely non-transformed data set that I use for one of linear models in section VII.

```
#1. Create Crime_long (for facet_wrap capabilties)
  library(moments)
  crime_plus_log_long <- crime %>%
    gather(var, val, -county, -year) %>%
    mutate(var_type = "applied_no_trans") %>%
    group_by(var) %>%
    mutate_at(vars(val), c("skewness","kurtosis")) %>%
    mutate_at(vars(skewness, kurtosis), funs(round(.,digits = 2))) %>%
    ungroup() %>%
    bind_rows(
      crime %>%
        gather(var, val, -county, -year) %>%
        mutate(var_type = "log_trans",
               val = log(val)) %>%
        group_by(var) %>%
        mutate_at(vars(val), c("skewness","kurtosis")) %>%
        mutate_at(vars(skewness, kurtosis), funs(round(.,digits = 2))) %>%
        ungroup()
    )

#2. Create Distribution Plots (order from most to least skewed -- when no transformation is applied)

  #a. create levels for future factoring of create 'var' variable
  plot_levels <- crime_plus_log_long %>%
    filter(var_type == "applied_no_trans") %>%
    arrange(-abs(skewness)) %>%
    distinct(var)

  plot_levels <- plot_levels$var

  #b. Which variables do not see a reduction in skew when log transformed?
  skew_improv_results <- crime_plus_log_long %>%
    filter(!var %in% c("west","central","urban","crmrte")) %>% #filter out binary variables and independent variable
    distinct(var, var_type, skewness) %>%
    mutate(var_type = paste0(var_type,"_abs_skew"),
           skewness = abs(skewness)) %>%
    spread(var_type, skewness) %>%
    mutate(improvement_in_skewness_byLOG = log_trans_abs_skew < applied_no_trans_abs_skew) %>%
    arrange(improvement_in_skewness_byLOG)
```

A. Table of Skewness and Kurtosis with and without Log Transformation

A table of the results which adds a flag ('improvement__in__skewness__byLOG') of TRUE/FALSE if the log transformation decreased the absolute value of the skewness (i.e., if the log transformation, roughly speaking, made the distribution of the variable *more* normal)

```
  skew_improv_results
```

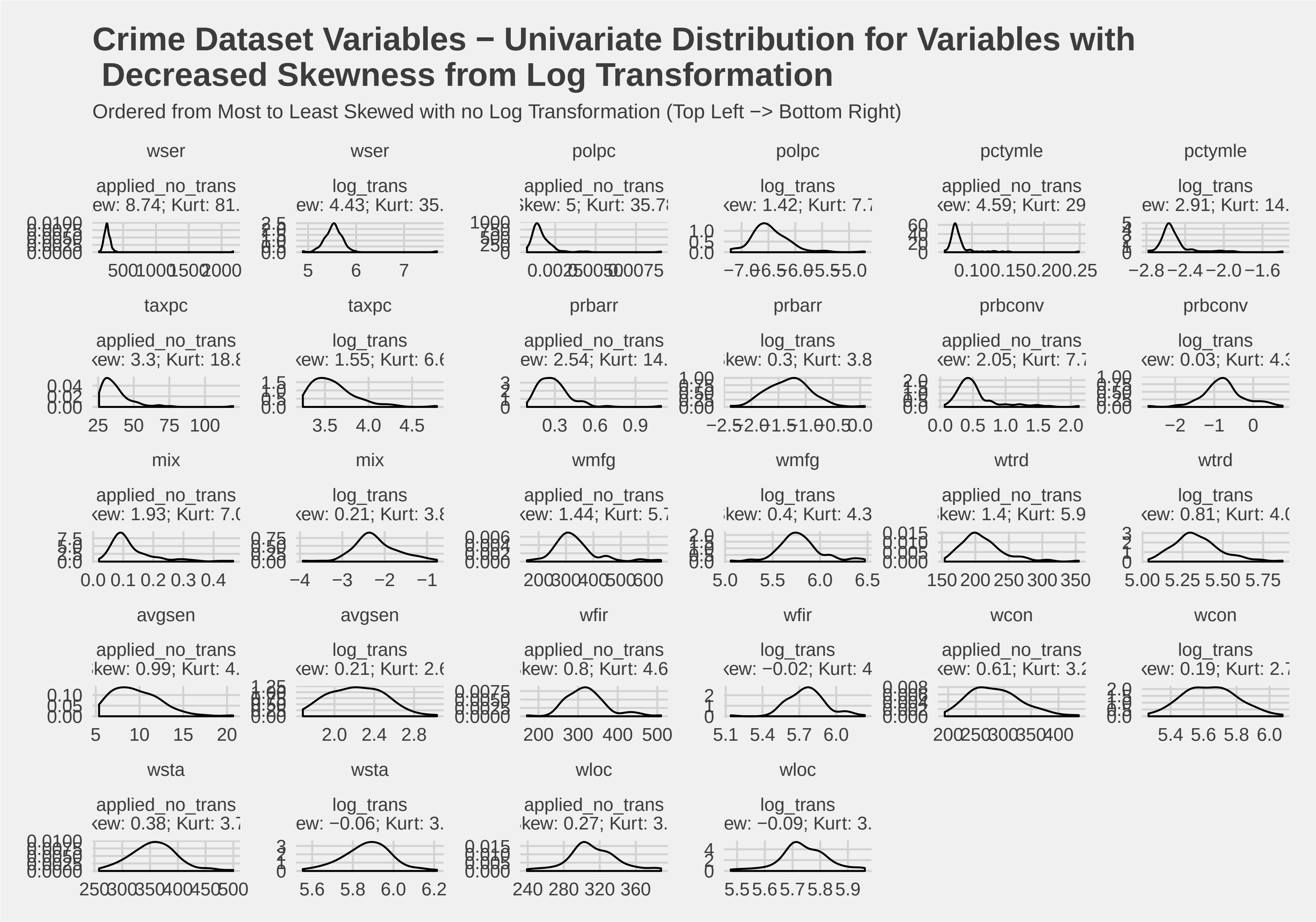
```
## # A tibble: 19 x 4
##         var applied_no_trans_abs_skew log_trans_abs_skew
##       <chr>                     <dbl>              <dbl>
##  1  density                      2.66               5.19
##  2 pctmin80                      0.38               0.92
##  3  prbpris                      0.46               1.43
##  4     wfed                      0.12               0.21
##  5     wtuc                      0.05               1.05
##  6   avgsen                      0.99               0.21
##  7      mix                      1.93               0.21
##  8  pctymle                      4.59               2.91
##  9    polpc                      5.00               1.42
## 10   prbarr                      2.54               0.30
## 11  prbconv                      2.05               0.03
## 12    taxpc                      3.30               1.55
## 13     wcon                       0.61               0.19
## 14     wfir                       0.80               0.02
## 15     wloc                       0.27               0.09
## 16     wmfg                       1.44               0.40
## 17     wser                       8.74               4.43
## 18     wsta                       0.38               0.06
## 19     wtrd                       1.40               0.81
## # ... with 1 more variables: improvement_in_skewness_byLOG <lgl>
```

B. Series of Distribution Plots - with and without Log Transformation

1) Crime Data set Variables - Univariate Distribution for Variables with Decreased Skewness from Log Transformation

Note that each variable shows two distribution plots (one with and one without Log Transformation)

```
  #c. plot univariate distriubtion for each variable, log and no log transformation
    #i. for variables which improved by transformation
    ggplot(data = crime_plus_log_long %>%
             filter(var %in% skew_improv_results$var[
               skew_improv_results$improvement_in_skewness_byLOG == TRUE
             ]) %>%
             mutate(var_label = paste0(var_type,"\n[Skew: ", skewness,"; Kurt: ",kurtosis,"]")) %>%
             mutate(var = factor(var, levels = plot_levels)), aes(val)) +
      geom_density(kernel = "gaussian") +
      facet_wrap(var~var_label, scales = "free", ncol = 6) +
      theme_fivethirtyeight() +
      labs(title = "Crime Dataset Variables - Univariate Distribution for Variables with \n Decreased Skewness from Log Transformation",
           subtitle ="Ordered from Most to Least Skewed with no Log Transformation (Top Left -> Bottom Right)")
```



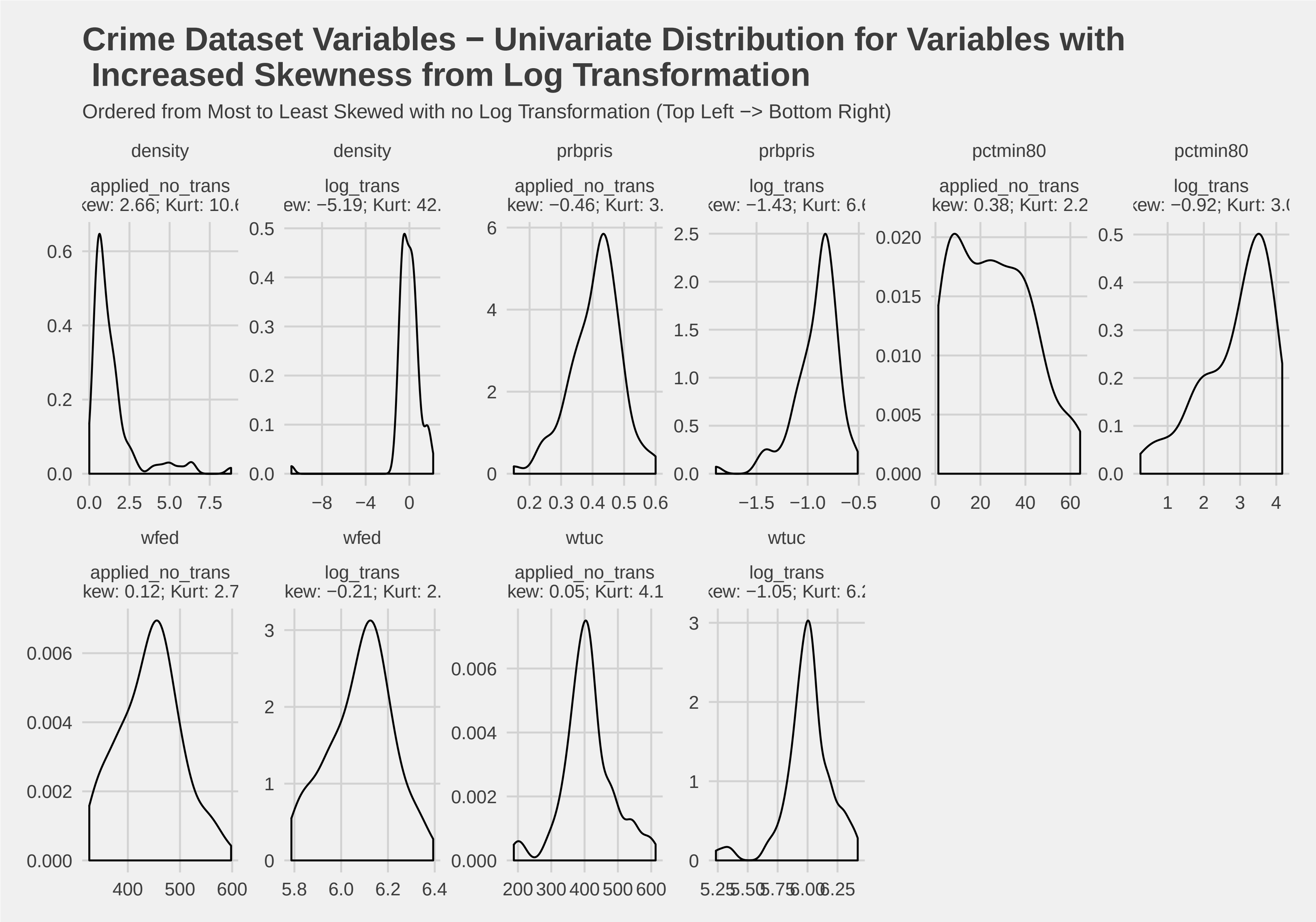
Crime Dataset Variables − Univariate Distribution for Variables with Decreased Skewness from Log Transformation

Ordered from Most to Least Skewed with no Log Transformation (Top Left −> Bottom Right)

2) Crime Data set Variables - Univariate Distribution for Variables with Increased Skewness from Log Transformation

Note that each variable shows two distribution plots (one with and one without Log Transformation)

```r
#ii. for variables which were not improved by transformation
ggplot(data = crime_plus_log_long %>%
         filter(var %in% skew_improv_results$var[
           skew_improv_results$improvement_in_skewness_byLOG == FALSE
           ]) %>%
         mutate(var_label = paste0(var_type,"\n[Skew: ", skewness,"; Kurt: ",kurtosis,"]")) %>%
         mutate(var = factor(var, levels = plot_levels)), aes(val)) +
  geom_density(kernel = "gaussian") +
  facet_wrap(var~var_label, scales = "free", ncol = 6) +
  theme_fivethirtyeight() +
  labs(title = "Crime Dataset Variables - Univariate Distribution for Variables with \n Increased Skewness from Log Transformation",
       subtitle ="Ordered from Most to Least Skewed with no Log Transformation (Top Left -> Bottom Right)")
```

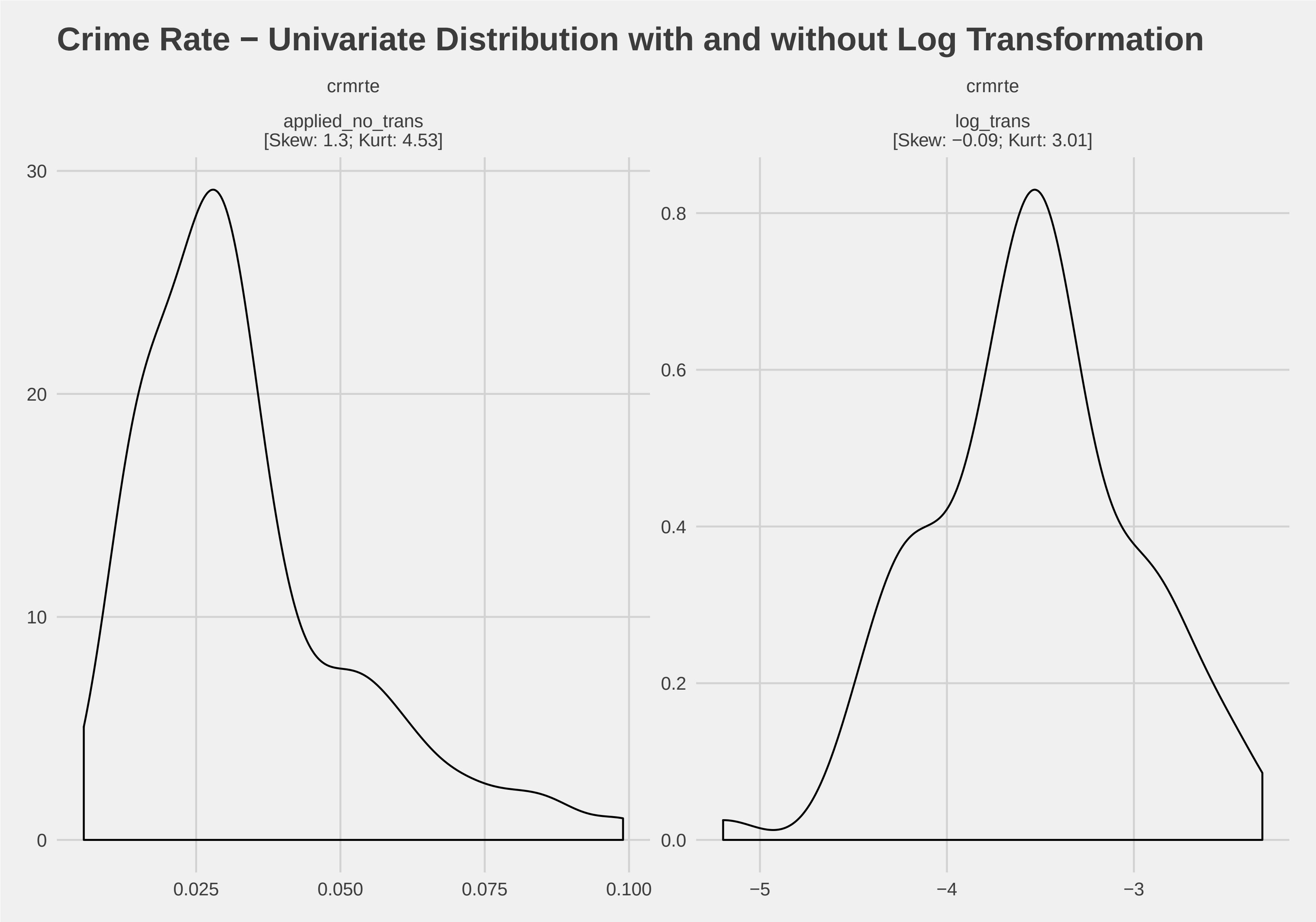


Crime Dataset Variables − Univariate Distribution for Variables with Increased Skewness from Log Transformation

Ordered from Most to Least Skewed with no Log Transformation (Top Left −> Bottom Right)

3) Crime Rate - Univariate Distribution with and without Log Transformation

```r
#3. Create Distribution Plots for log-transformation of CrimeRate (order from most to least skewed)

    ggplot(data = crime_plus_log_long %>%
             filter(var %in% "crmrte") %>%
             mutate(var_label = paste0(var_type,"\n[Skew: ", skewness,"; Kurt: ",kurtosis,"]")), aes(val)) +
      geom_density(kernel = "gaussian") +
      facet_wrap(var~var_label, scales = "free") +
      theme_fivethirtyeight() +
      labs(title = "Crime Rate - Univariate Distribution with and without Log Transformation")
```


Crime Rate − Univariate Distribution with and without Log Transformation
crmrte
applied_no_trans
[Skew: 1.3; Kurt: 4.53]
30
20
10
0
0.025
0.050
0.075
0.100
crmrte
log_trans
[Skew: −0.09; Kurt: 3.01]
0.8
0.6
0.4
0.2
0.0
−5
−4
−3

C. Log Transformation within the Crime Data Set

In the steps below, I modified the crime data set and log transformed any variables which experienced a decreased in skewness (or an increase in normality) from the log transformation. In addition to the crime rate variable, these include all variables in the V.B.1 series of distributions.

```
#4. If the variable had reduced skewness from log transformation (improvement_in_skewness_byLOG == TRUE),
    # then transform it, otherwise, keep it

    #a. create vector of variables to log transform
    log_transform_these_variables <- skew_improv_results$var[
      skew_improv_results$improvement_in_skewness_byLOG == TRUE
      ]
    #b. add crmrte to transformation vector
    log_transform_these_variables <- c(log_transform_these_variables, "crmrte")

    #c. transform the variables
    crime_noTrans <- crime

    crime <- crime %>%
      mutate_at(vars(log_transform_these_variables), funs(log = log))

    #d. remove the variables that were log transformed
    crime[,log_transform_these_variables] <- NULL
```

VI. Examine Correlation Among Variables

In this section I examine the correlation between the independent variables and between the independent variables and crime rate.

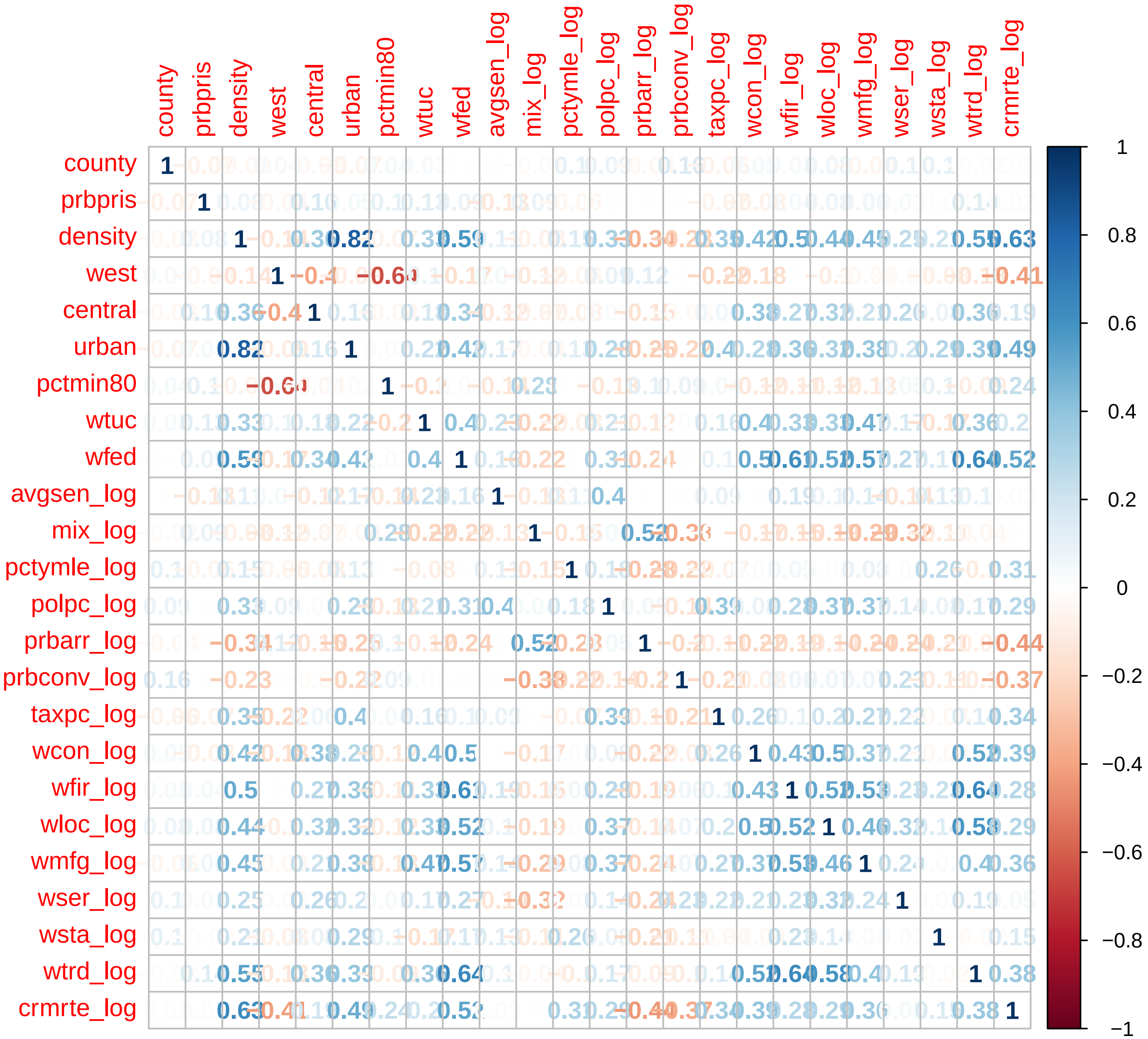
First, I output a correlation table which shows the correlation among all variables.

Second, I output a series of scatter plots for nine pairs of independent variables which have the highest correlation among all independent variables. As a reminder, some of these variables have been log transformed. If the variables have been log transformed, they are now called '[variable name]_log.'

Finally, I output a similar series of scatter plots, but for the nine variables with the highest correlation with the dependent variable, crime rate.

A. Correlation among all variables: heat table

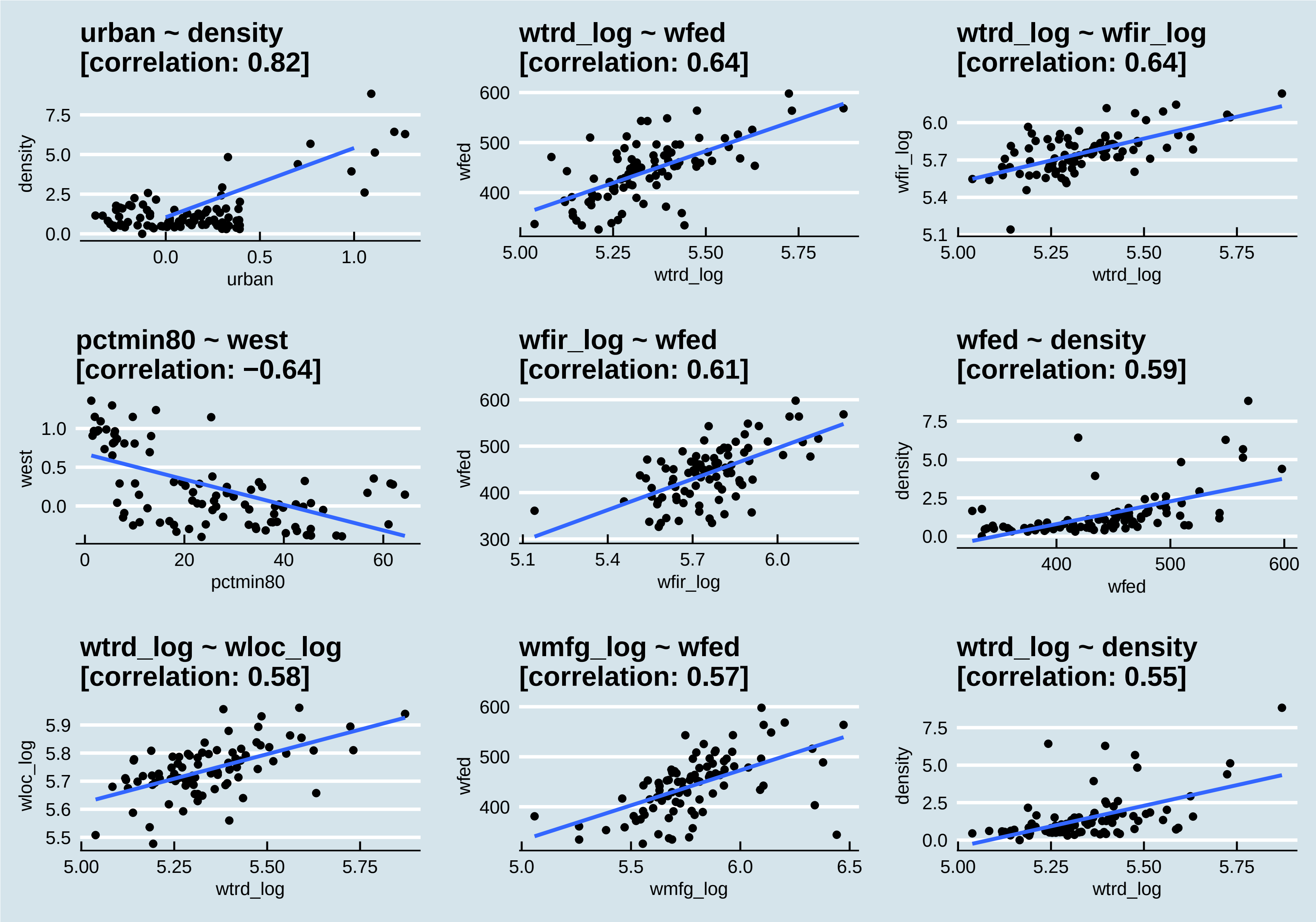
```
#1. Create corplot
  corrplot(cor(crime %>% select(-year)),method="number")
```


county
prbpris
density
west
central
urban
pctmin80
wtuc
wfed
avgsen_log
mix_log
pctymle_log
polpc_log
prbarr_log
prbconv_log
taxpc_log
wcon_log
wfir_log
wloc_log
wmfg_log
wser_log
wsta_log
wtrd_log
crmrte_log
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

```
#2. Create Correlation Table
crime_cor <- as.data.frame(
  cor(crime %>%
        select(-county, - year))) %>%
    rownames_to_column("var1") %>%
    gather(var2, cor, -var1) %>%
    filter(var1 != var2) %>%
    arrange(cor) %>%
    filter(!(var1 == lag(var2, default = "") & var2 == lag(var1, default = ""))) #var 1 - var2 vice versa gets repeated

  #Top correlated independent variables - possible issues of multicollinearity
  top_cor_among_Ivars <- crime_cor %>%
    filter(var1 != "crmrte_log",
           var2 != "crmrte_log") %>%
    mutate(cor_abs = abs(cor)) %>%
    arrange(-cor_abs) %>%
    head(9)

  #Top correlated independent variables with Crime Rate
  top_cor_among_crmrte_log <- crime_cor %>%
    filter(var1 == "crmrte_log" |
           var2 == "crmrte_log") %>%
    mutate(cor_abs = abs(cor),
           temp = var1 == "crmrte_log",
           var1 = ifelse(temp, var2, var1),
           var2 = ifelse(temp, "crmrte_log", var2)) %>%
    arrange(-cor_abs) %>%
    head(9)

#3. Plot Top Correlated Independent Variables

  #a. create plotting function

  cor_plot <- function(x_in,y_in){
    crime$x <- crime[[x_in]]
    crime$y <- crime[[y_in]]

    cor <- cor(crime$x, crime$y)
    ggplot(crime, aes(x,y)) +
      geom_jitter() +
      geom_smooth(method = "lm", se = FALSE) +
      theme_economist() +
      labs(title = paste0(x_in, " ~ ", y_in, "\n[correlation: ",round(cor,2),"]"),
           x = x_in,
           y = y_in)
  }

  #b. create list of plots
    #i. Top correlated independent variables - possible issues of multicollinearity
    Ivars_plots <- pmap(list(as.list(top_cor_among_Ivars$var1),
                       as.list(top_cor_among_Ivars$var2)),
                  cor_plot)
    #ii. Top correlated independent variables with Crime Rate
    crmrte_log_plots <- pmap(list(as.list(top_cor_among_crmrte_log$var1),
                                  as.list(top_cor_among_crmrte_log$var2)),
                             cor_plot)
```

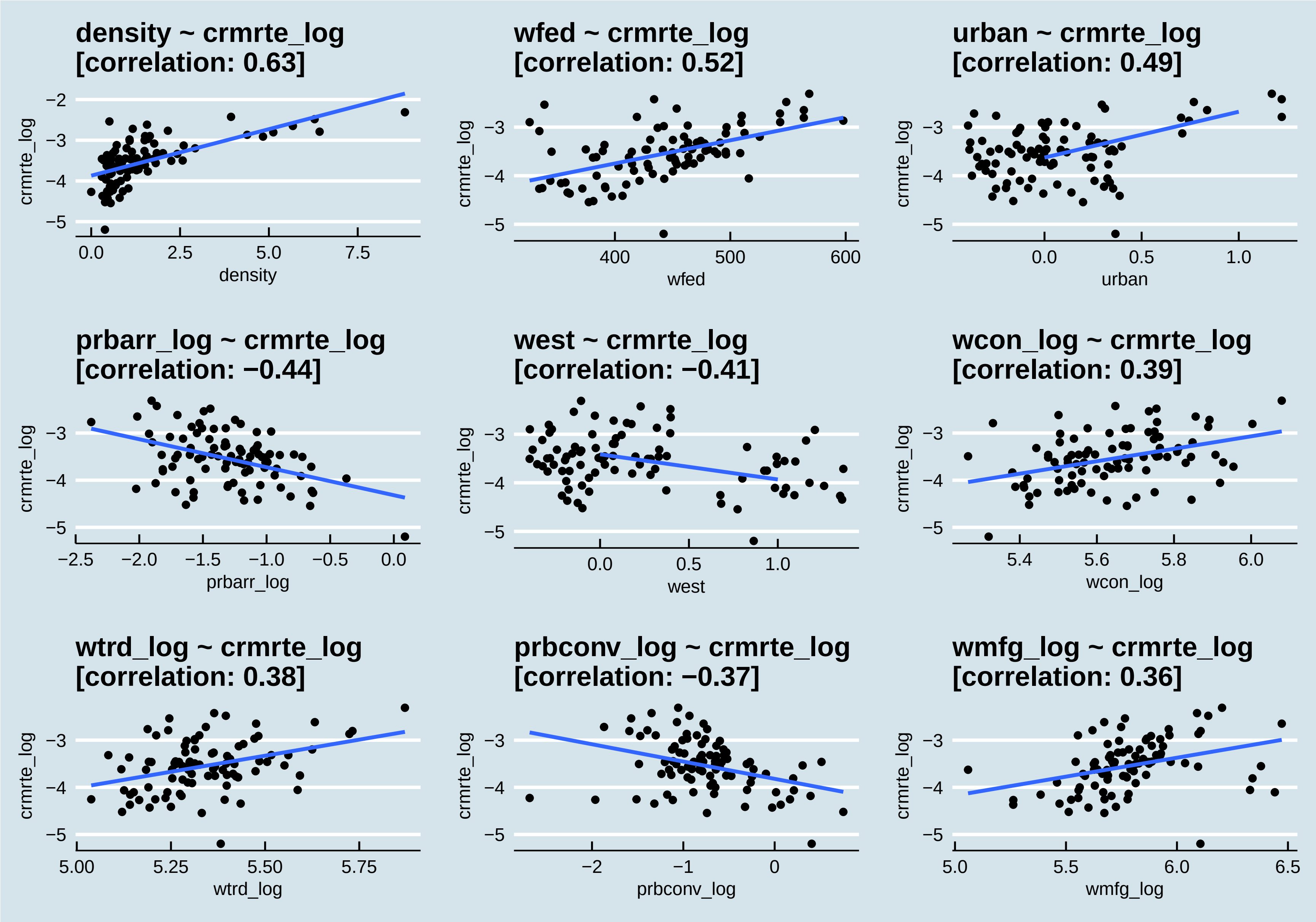
B. Scatter plots for nine pairs of independent variables which have the highest correlation among all independent variables.

```
#c. output grid of plots
  #i. Top correlated independent variables - possible issues of multicollinearity
  grid.arrange(grobs = Ivars_plots, ncol = 3)
```


urban ~ density
[correlation: 0.82]
density
urban
0.0
2.5
5.0
7.5
0.0
0.5
1.0
wtrd_log ~ wfed
[correlation: 0.64]
wfed
wtrd_log
400
500
600
5.00
5.25
5.50
5.75
wtrd_log ~ wfir_log
[correlation: 0.64]
wfir_log
wtrd_log
5.1
5.4
5.7
6.0
5.00
5.25
5.50
5.75
pctmin80 ~ west
[correlation: −0.64]
west
pctmin80
0.0
0.5
1.0
0
20
40
60
wfir_log ~ wfed
[correlation: 0.61]
wfed
wfir_log
300
400
500
600
5.1
5.4
5.7
6.0
wfed ~ density
[correlation: 0.59]
density
wfed
0.0
2.5
5.0
7.5
400
500
600
wtrd_log ~ wloc_log
[correlation: 0.58]
wloc_log
wtrd_log
5.5
5.6
5.7
5.8
5.9
5.00
5.25
5.50
5.75
wmfg_log ~ wfed
[correlation: 0.57]
wfed
wmfg_log
400
500
600
5.0
5.5
6.0
6.5
wtrd_log ~ density
[correlation: 0.55]
density
wtrd_log
0.0
2.5
5.0
7.5
5.00
5.25
5.50
5.75

C. Scatter plots for nine variables which have the highest correlation with crime rate.

```r
#ii. Top correlated independent variables with Crime Rate
grid.arrange(grobs = crmrte_log_plots, ncol = 3)
```


density ~ crmrte_log
[correlation: 0.63]
crmrte_log
−2
−3
−4
−5
0.0
2.5
5.0
7.5
density
wfed ~ crmrte_log
[correlation: 0.52]
crmrte_log
−3
−4
−5
400
500
600
wfed
urban ~ crmrte_log
[correlation: 0.49]
crmrte_log
−3
−4
−5
0.0
0.5
1.0
urban
prbarr_log ~ crmrte_log
[correlation: −0.44]
crmrte_log
−3
−4
−5
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
prbarr_log
west ~ crmrte_log
[correlation: −0.41]
crmrte_log
−3
−4
−5
0.0
0.5
1.0
west
wcon_log ~ crmrte_log
[correlation: 0.39]
crmrte_log
−3
−4
−5
5.4
5.6
5.8
6.0
wcon_log
wtrd_log ~ crmrte_log
[correlation: 0.38]
crmrte_log
−3
−4
−5
5.00
5.25
5.50
5.75
wtrd_log
prbconv_log ~ crmrte_log
[correlation: −0.37]
crmrte_log
−3
−4
−5
−2
−1
0
prbconv_log
wmfg_log ~ crmrte_log
[correlation: 0.36]
crmrte_log
−3
−4
−5
5.0
5.5
6.0
6.5
wmfg_log

VII. Make Models

After the extensive exploratory data analysis described above, I have constructed three linear models to help identify the determinants of the crime rate in North Carolina counties.

I discuss each model below and then compare the three models. Finally, I offer a conclusion of my findings.

```
#1. Set - Up
  #a. Remove Variables
  #i. we don't need year it's always the same
  unique(crime$year)
```

```
## [1] 87
```

```
  #ii. we don't need county, it's always different and has no inherent numeric quality (we have region variables
  crime <- crime %>%
    select(-year, -county)

  crime_noTrans <- crime_noTrans %>%
    select(-year, -county)

  #b. create a vector of variables
  vars <- names(crime %>% select(-crmrte_log))

  vars_noTrans <- names(crime_noTrans %>% select(-crmrte))
```

A. Model One: Only Key Variables of Interest

In this model I examine the effect of density and pctmin80 ('perc. minority, 1980') on the log of crime rate.

- I chose to focus on density because it has the highest correlation with the log of crime rate. Furthermore, the variable seems like the most important variable, qualitatively, because it could be adjusted by policy. For example, a government could pass laws making it illegal for more than X amount of people to reside in a given square mile.
- I chose to focus on the minority percentage because of I believe this to serve as a valuable proxy of income (due to America's extreme racism). As proof of this racism, see that there is a positive correlation between the minority percentage of a given community and the probability that an individual in that community will be arrested, convicted, and/or be given a prison sentence.

In all, I believe that the income of a community in conjunction with how tightly packed a community is will serve as the biggest indicator of crime. Additionally, if this hypothesis is proven and a causal model can be established, these two factors could be directly targeted with policy.

```
#2. Make Models - with transformed variables
  #a. Only Key variables of interest
  mod_key_var <- lm(crmrte_log ~ density + pctmin80, data = crime)
  mod_key_var
```

```
##
## Call:
## lm(formula = crmrte_log ~ density + pctmin80, data = crime)
##
## Coefficients:
## (Intercept)      density     pctmin80
##   -4.110125     0.235450     0.009009
```

```
  summary(mod_key_var)$r.squared
```

```
## [1] 0.4799467
```

```
  AIC(mod_key_var)
```

```
## [1] 95.64913
```

B. Model Two: Only Key Variables of Interest

In this model I examine the key variables discussed above in addition to other covariates which increase the accuracy of my model without introducing substantial bias.

Weekly wage data for a variety of fields (such as federal employees and construction) has a large predictive impact in the model. This was foreshadowed by by high correlation between these variables and the log of crime rate (discussed in section VI.). For this model, I chose to exclude certain variables which had high collinearity with some key variables (such as the exclusion of ‘urban’ which is highly correlated with density at .84).

```
  #b. Key variables and covariates that increase the accruuacy of my results without introducing substantial bias
  mod_key_var_plus_covar <- lm(crmrte_log ~ density + pctmin80 +
                                 west + central + #I can get urban from density
                                 prbarr_log + prbconv_log +
                                 wfed+ wcon_log + wtrd_log + wmfg_log, data = crime)
  mod_key_var_plus_covar
```

```
##
## Call:
## lm(formula = crmrte_log ~ density + pctmin80 + west + central +
##     prbarr_log + prbconv_log + wfed + wcon_log + wtrd_log + wmfg_log,
##     data = crime)
##
## Coefficients:
## (Intercept)      density     pctmin80         west      central
##   -8.035888     0.096405     0.005960    -0.287434    -0.201708
##  prbarr_log  prbconv_log         wfed     wcon_log     wtrd_log
##   -0.484433    -0.397442     0.001998     0.296498    -0.041383
##    wmfg_log
##    0.187110
```

```
summary(mod_key_var_plus_covar)$r.squared
```

```
## [1] 0.7545365
```

```
AIC(mod_key_var_plus_covar)
```

```
## [1] 43.32789
```

C. Model Three: All Variables

In this model I examine all independent variables and their relationship to the log of the crime rate. As a reminder, many of these variables have been log transformed. You'll see that my r.squared is increased, but not drastically (r.squared will always increase with more variables). However, notably, the AIC of this model increases because of the use of so many variables. This is indicative of the accuracy and power of my second model.

```
#c. All variables (note: county and year have already been deleted)
mod_all_vars <- lm(paste0("crmrte_log ~ ",paste0(vars,collapse = " + ")),data=crime)
mod_all_vars
```

```
##
## Call:
## lm(formula = paste0("crmrte_log ~ ", paste0(vars, collapse = " + ")),
##     data = crime)
##
## Coefficients:
## (Intercept)      prbpris      density         west      central
##  -1.9634117   -0.2716276    0.1189759   -0.1955815   -0.1394382
##       urban     pctmin80         wtuc         wfed   avgsen_log
##  -0.1714518    0.0085456    0.0000808    0.0021012   -0.1884145
##     mix_log  pctymle_log    polpc_log   prbarr_log  prbconv_log
##   0.0154853    0.1935270    0.2686374   -0.5614187   -0.3360719
##   taxpc_log     wcon_log     wfir_log     wloc_log     wmfg_log
##   0.0285223    0.2802739   -0.2310584    0.1968565    0.1038259
##    wser_log     wsta_log     wtrd_log
##  -0.2873723   -0.3732983    0.1285550
```

```
  summary(mod_all_vars)$r.squared
```

```
## [1] 0.8067997
```

```
  AIC(mod_all_vars)
```

```
## [1] 45.54059
```

D. Omitted Variables

As briefly discussed, the largest omitted variables surround the idea of poverty. Certain variables in this data set like minority percentage and weekly wage of construction workers serve as proxies for this measure of poverty. In my first model, because poverty presumably has a positive relationship with the crime rate, density, and minority percentage this means that my omitted variable bias is positive. This means that, in my first model, I am overestimating the effect of density and minority percentage on the log of crime rate.

E. Model Summary and Report Conclusion

```
#4. Compare the six models
  stargazer(mod_key_var,
            mod_key_var_plus_covar,
            mod_all_vars,
            type = "latex",
            report = "vc", # Don't report errors, since we haven't covered them
            title = "Linear Models Predicting Crime Rate in North Carolina in 1988",
            keep.stat = c("rsq", "n"),
            font.size = "tiny",
            omit.table.layout = "n") # Omit more output related to errors
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Sun, Apr 01, 2018 - 3:55:51 PM

As can be seen in the summary table for models (1) and (2) above, the is a positive relationship between the crime rate in a community in North Carolina and density, the minority percentage, weekly wage for a) federal workers, b) construction workers, and c) manufacturing. In addition, there is

Table 1: Linear Models Predicting Crime Rate in North Carolina in 1988

	Dependent variable:		
	crmrte_log		crmrte_log ~
	(1)	(2)	(3)
prbpris			−0.272
density	0.235	0.096	0.119
pctmin80	0.009	0.006	0.009
wtuc			0.0001
west		−0.287	−0.196
central		−0.202	−0.139
urban			−0.171
prbarr_log		−0.484	−0.561
prbconv_log		−0.397	−0.336
taxpc_log			0.029
wfed		0.002	0.002
avgsen_log			−0.188
mix_log			0.015
pctymle_log			0.194
polpc_log			0.269
wcon_log		0.296	0.280
wfir_log			−0.231
wloc_log			0.197
wtrd_log		−0.041	0.129
wmfg_log		0.187	0.104
wser_log			−0.287
wsta_log			−0.373
Constant	−4.110	−8.036	−1.963
Observations	91	91	91
R^2	0.480	0.755	0.807

a negative relationship between the crime rate in a community in North Carolina and west/central location, the probability of arrest and/or conviction, and the weekly wage of workers in wholesale/retail trade.

Most notably, in model 1, a .235 increase in the amount of people per mile will increase the amount of crimes committed per person by 1 percent. I'd recommend focusing on improving highly dense areas (which are overwhelmingly urban) by making them less dense. No other variables can be determined to be causal. Postive and negative relationships may be used in other ways to *predict* crime in areas, but not adjust it.