Lab 3 - Reducing Crime

Clayton G. Leach, Karl I. Siil, Timothy S. Slade

July 23, 2018

Introduction

Our client is running for office in the state of North Carolina (NC). Her campaign commissioned us to research the determinants of crime in NC to help her develop her platform regarding crime-related policy initiatives at the level of local government. This report explores a subset of the county-level data from Cornwell & Trumball's *Estimating the Economic Model of Crime with Panel Data (1994)* that provides various economic, demographic, and crime indicators for 1987. Our analysis describes the dataset, presents initial summary statistics, develops several linear regression models, and proposes additional research to inform policy recommendations.

Initial Exploratory Data Analysis (EDA)

We begin by exploring our dataset. We see that it has 97 records and 25 variables.

```
dim(crime_raw)
```

```
## [1] 97 25
```

```
names(crime_raw)
```

```
##  [1] "county"   "year"     "crmrte"   "prbarr"   "prbconv"  "prbpris"
##  [7] "avgsen"   "polpc"    "density"  "taxpc"    "west"     "central"
## [13] "urban"    "pctmin80" "wcon"     "wtuc"     "wtrd"     "wfir"
## [19] "wser"     "wmfg"     "wfed"     "wsta"     "wloc"     "mix"
## [25] "pctymle"
```

We generate a series of scatterplots that we use to examine our variables. Code is presented below, but results are omitted for parsimony.[1]

```
make_scatters <- function(df, var_list, y, trans) {

    if (!missing(trans)) {
        var_list <- append(var_list, str_glue("{trans}({var_list})"))
    }

    for (v in var_list) {
        print(ggplot(df, aes_string(x = v, y = y)) + geom_point() + geom_smooth(method = "lm",
            se = FALSE) + xlab(v) + ylab(y) + ggtitle(str_glue("{y} vs {v}")))
    }
}
```

The notes we receive provide the following insight into the variables:

Table 1: Available Variables from Cornwell & Trumball (1994)

[1] They can be viewed in the *Appendix*

#	Variable Name	Type	Description
1	`county`	integer	*Source county of data*
2	`year`	integer	*Source year of data*
3	`crmrte`	numeric	*crime rate*
4	`prbarr`	numeric	*'probability' of arrest*
5	`prbconv`	numeric	*'probability' of conviction*
6	`prbpris`	numeric	*'probability' of prison sentence*
7	`avgsen`	numeric	*average sentence, in days*
8	`polpc`	numeric	*police per capita*
9	`density`	numeric	*people per sq. mile*
10	`taxpc`	numeric	*tax revenue per capita*
11	`west`	dummy	*source county of data is in Western NC*
12	`central`	dummy	*source county of data is in Central NC*
13	`urban`	dummy	*source county of data is urban*
14	`pctmin80`	numeric	*percent minority in 1980*
15	`wcon`	numeric	*wages in the construction industry*
16	`wtuc`	numeric	*wages in the transportation, utilities, and communication industries*
17	`wtrd`	numeric	*wages in the construction industry*
18	`wfir`	numeric	*wages in the finance, insurance, real estate industries*
19	`wser`	numeric	*wages in the service industry*
20	`wmfg`	numeric	*wages in the manufacturing industry*
21	`wfed`	numeric	*wages among federal employees*
22	`wsta`	numeric	*wages among state employees*
23	`wloc`	numeric	*wages among local government employees*
24	`mix`	numeric	*mix of offenses; face-to-face v others*
25	`pctymle`	numeric	*percent young male*

We see some variables which *a priori* seem useful: the 'probability' variables, police per capita, tax revenue, wages, and youth and minority composition of a county. Before exploring them further, however, we search the entire dataset for missing values that may affect our analyses.

Missing Values

We find 6 rows that are missing data; further scrutiny shows 5 are entirely blank and 1 contains only a backtick. We eliminate those to generate our working dataset.

```
crime_na <- crime_raw %>% filter_all(any_vars(!is.na(.)))  # Rows with no data
crime_na %>% filter_all(any_vars(is.na(.))) %>% select(which(!is.na(.)))  # Formerly row with one back
```

```
## # A tibble: 0 x 0
```

```
crime_na <- crime_na %>% filter_all(all_vars(!is.na(.)))  # Verification
```

Erroneously Duplicated Records

Continuing our QC, we note that 1 of the counties' records has been duplicated exactly. We therefore drop the duplicate record from our dataset.

```
crime_na %>% count(county) %>% filter(n > 1)  # county 193 is an exact duplicate
```

```
## # A tibble: 1 x 2
##   county     n
##    <int> <int>
## 1    193     2
```

```
crime_na <- crime_na %>% filter(!duplicated(.))  # removed
```

Plausibility Checks for Variables

Three of our key variables of interest (`prbarr`, `prbconv`, and `prbpris`) represent probabilities and should therefore theoretically be in the range of 0:1.

```
# Examine 'probability' variables.
non_prob <- crime_na %>% filter(!between(prbarr, 0, 1) | !between(prbconv, 0,
    1) | !between(prbpris, 0, 1))
```

Examining the data[2], we find 10 counties have values for the "probability" variables that are outside of the expected range. In each case, it is either `prbconv` (10 records) or `prbarr` (1 record) that fall outside the range.

Per the notes accompanying our data, *The probability of conviction is proxied by the ratio of convictions to arrests...* Given that definition, if not all suspects arrested are convicted, `prbconv` will be below 1. However, it may also exceed 1 if the number of exonerated suspects is exceeded by the number of suspects convicted of multiple charges. (See **here** for examples of multiple charges stemming from a single arrest.)

The notes on `prbarr` indicate *the probability of arrest is proxied by the ratio of arrests to offenses....* If multiple suspects are arrested for a single offense, and this happens more frequently than offenses which do not lead to arrests, `prbarr` would indeed exceed 1.

In both cases, there are plausible explanations for a probability value in excess of 1. However, one of the observations appears to be an outlier. The county labeled `115` has the lowest crime rate by far (~50% lower than that of any other county), the highest 'probability' of arrest (>1 arrest per offense, nearly 58% greater than the county with the second-highest probability), the longest average sentence (20.7 days, ~15% higher than the second-longest), and the largest number of police per capita (9 officers per 1,000 residents, more than twice as many as the second-highest county). While those numbers appear unusual, they are also internally consistent: one would expect a very low crime rate from a county that has a very strong police presence, arrests a large proportion of suspects, and punishes convicted criminals severely.[3]

Examining the remainder of our data, we found no substantial evidence of *top-coded* or *bottom-coded* (i.e., truncated) variables which might bias our regression models. However, there is an extreme outlier in `wser`, the variable indicating the county's weekly wage in the service industry. To determine if this is valid we looked at the wage values for other sectors of the economy and did not see elevated values. It is improbable, but not impossible, that individuals in the service industry are making significantly more than anyone else in the county. We believe this data point should be scrutinized further before a determination is made to modify it.

Transformation Analysis

If the relationship between two variables is not linear, adding them to a linear regression model as-is (without a transformation) will generate inaccurate results and possibly result in an invalidation of our heteroskedasticity assumption. It is therefore important to explore whether the relationship between two variables shares some non-linear relationship and thus whether a transformation is required. As part of our EDA we explored this question for all of the variables in the dataset.

[2] See *Appendix* for the `non_prob` table

[3] Discussion with peers and further research reveals that the county labels are, in fact, FIPS (Federal Information Processing Standard) codes. A deep dive into additional contextual factors that could inform our analyses is beyond the scope of this report; we leave it as an exercise for the reader.

Our first step was to evaluate if any variables had significant skew in their distributions by checking whether they generally conformed to a normal distribution using R's qqplot. While this is not necessarily a reason to transform a variable, it can help us identify variables of interest.[4]

From our graphs we saw that probability of arrest (`prbarr`), probability of conviction (`prbconv`), police per capita (`polpc`), tax revenue per capita (`taxpc`), population density (`density`), proportion of young males in the population (`pctymle`), and the mix of face-to-face vs. impersonal crimes (`mix`) all deviated from normality. We will consider this in addition to other factors when deciding if a transformation would be beneficial.

Secondly, while not a perfect approach due to the possible interactions amongst variables, we also wanted to look at whether there is any obvious non-linearity when looking at crime rate and each variable independently. To do this we looked at a scatterplot of crime rate vs. each variable.[5]

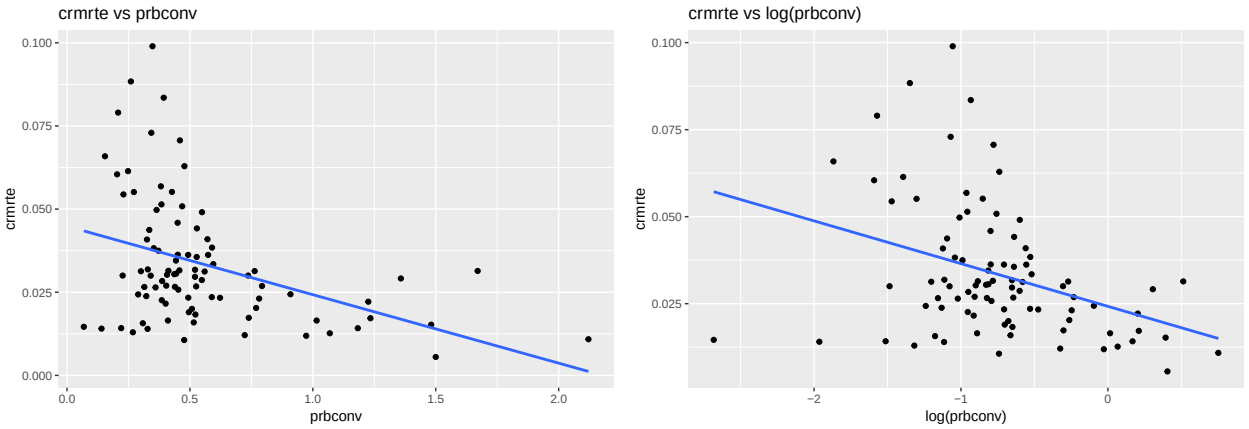
Reviewing these graphs yields five (5) variables which appear to have a non-linear relationship with crime rate: `prbarr`, `prbconv`, `polpc`, `density`, and `taxpc`. While it is not particularly important that predictor variables be normally distributed, we do want them to display both *symmetry* and *high variance*; to the extent applying a transformation advances those goals, it is worth considering provided it does not inhibit interpretability or run strongly counter to theory.

Four of the variables appear to benefit from a log transform, while the fifth `density` appears to be related to crime rate via the square root function. The pre- and post-transformation scatterplots and q-q plots for `prbconv` are presented here for illustrative purposes.[6] In addition to checking whether the relationship appeared to improve in linearity, we also checked whether this transformation helped with variable skew. In every case the distribution of our variable moved closer to normality. **Given the improvement in linearity and normality we will use these transformations moving forward. <- TS: Given what we now know about not needing to transform dependent variables unless it helps with our need for symmetry and high variability, do we still want to do this?**

Linearity Graphs:

```
# make_scatters(df = crime_na, var_list = c('prbarr'), y = 'crmrte', trans =
# 'log')
```

```
make_scatters(df = crime_na, var_list = c("prbconv"), y = "crmrte", trans = "log")
```




```
# make_scatters(df = crime_na, var_list = c('polpc'), y = 'crmrte', trans =
# 'log')
```

[4] See *Appendix* for detailed code.
[5] See *Appendix* for detailed code
[6] See *Appendix* for the others.

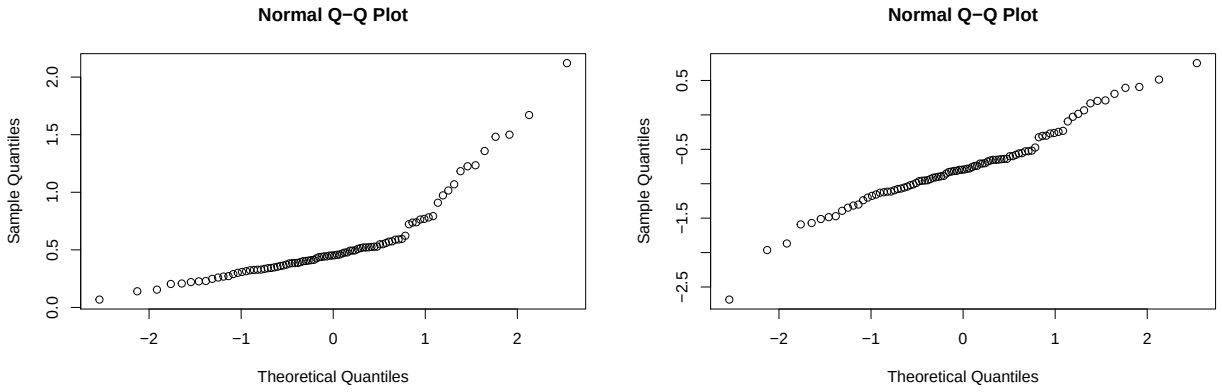
```
# make_scatters(df = crime_na, var_list = c('density'), y = 'crmrte', trans
# = 'sqrt')
```

```
# make_scatters(df = crime_na, var_list = c('taxpc'), y = 'crmrte', trans =
# 'log')
```

Normality/Skew With and Without Transformation:

```
# Normality/skew with and without transformations
# print(c(qqnorm(crime_na$prbarr), qqnorm(log(crime_na$prbarr))))
```

```
print(c(qqnorm(crime_na$prbconv), qqnorm(log(crime_na$prbconv))))
```




```
# print(c(qqnorm(crime_na$polpc), qqnorm(log(crime_na$polpc))))
```

```
# print(c(qqnorm(crime_na$density), qqnorm(sqrt(crime_na$density))))
```

```
# print(c(qqnorm(crime_na$taxpc), qqnorm(log(crime_na$taxpc))))
```

Finally, we examined collinearity between various variables that we thought would comprise our models. For the sake of narrative flow, discussion of those findings is presented alongside the models in question.

Research Question and Model-Building

Our **research question** is the following: **Should our candidate support a traditional "Tough on Crime" platform? ** <- TS: I still think we should reframe this somehow. Let's discuss.****

We face a key limitation: our data does not give us visibility into the crimes themselves or changes in crime, but rather provides only the official *crime rate*. The crime rate is a function not only of crimes committed but also of various additional factors, some of which may be unobservable. For instance, poor community-police relations may bias crime rates downward if an area's residents **do not report all the crimes they observe or experience**. Conversely, those poor relations may also bias crime rates upward if police officers engage in **predatory policing practices** and the community lacks the wherewithal to fight back. A full discussion of omitted variable bias will occur later in this analysis, but we preface our model-building section with this note to as a practical way to invite the reader to critically examine the models we propose.

In order to answer our research question we created several models which included variables related to key crime policy decisions. While there are 25 variables in our raw dataset, our model-building proceeded systematically. First, we grouped variables together *thematically*. Next, we built a model with the variables comprising the theme we believed *a priori* would be most likely to have high predictive value. After evaluating the model's performance, we added the variables for the next theme, and so forth.

Thematic grouping of variables

Theme 1: Crime & Punishment

Our first theme revolves around crime and the likelihood of punishment resulting from the act of committing a crime. We believe that the variables police per capita (`polpc`), probability of arrest (`prbarr`), probability of conviction (`prbconv`), probability of incarceration (`prbpris`), and average sentence length (`avgsen`) fit well together. Taken comprehensively, they help frame the observed crime rate in economic terms as a sort of "risk-reward" proposition for would-be criminals: is the presence of a well-functioning criminal justice system[7]—in which crimes consistently lead to arrests (of the actual perpetrators, and not innocent bystanders), and arrests consistently lead to (warranted) convictions, and convictions lead to prison sentences, and sentences are indeed punitive—predictively associated with the crime rate?

The variables in this theme are particularly valuable because they provide a solid foundation for policy development. If the size of the police force is associated with changes in the crime rate, it can be scaled up or scaled down appropriately. If the probability of arrest makes a difference, the county could invest in improving community-police relationships with the goal **of improving investigations and... <- TS: sort of lost my train of thought here.** If the probability of conviction is important, funds could be invested in further training the police and District Attorneys on the collection and presentation of evidence. If the probability of incarceration and average sentence are important, legislators could advocate for changes to sentencing guidelines.

Theme 2: Demography & Economics

Our second theme focuses on issues of community composition, economic resources, and opportunity. Population density (`density`) may influence the crime rate in complex ways: from a sociological perspective, the social ecology approach to studying criminality considers population density to be one of several so-called *criminogenic* ("crime-causing") factors. From a psychological perspective, a larger population may weaken community bonds[8] and thus the pressure to exhibit prosocial behavior. Where significant wealth or income disparities exist within close proximity, relative deprivation theory suggests population density may drive greater criminality. On the other hand, large, sparsely-populated swathes of land are difficult for police to monitor and may present opportunities for various non-violent crimes.[9]

Social control theory suggests that young people with weaker bonds to their communities may be more likely to commit crime. Indeed, economic and criminological research over the last several decades has argued young men are disproportionately involved in crime. There has also been widespread discussion of the role that race and ethnicity play in criminality. We thus include the variables `pctymle` and `pctmin80` in this model.

The police force is funded by the local government, which in turn is funded by the local taxpayers. We might thus expect that the taxes paid per capita (`taxpc`) might have an influence on the size and/or effectiveness of the police force, either reinforcing or undermining any effect `polpc` might have on the crime rate.

The dataset contains nine variables related to wage: `wfed`, `wsta`, `wloc`, `wmfg`, `wser`, `wcon`, `wfir`, `wtrd`, `wtuc`. While much of their cumulative influence on a community may be captured in the `taxpc` variable - and thus we must beware excessive collinearity - different groupings of those wage variables may provide insight into well- or under-funded segments of the local economy.

We believe that many of these variables may exhibit some collinearity with the predictors we included in **Model 1**, but they are sufficiently distinct as to warrant their own model. Specifically, we suspect `taxpc` and `density` may be collinear with `polpc`; `pctymle` and `pctmin80` may be collinear with `prbarr`; and government

[7] Using an admittedly naïve formulation in which socioeconomic status, racial bias, and predatory policing are assumed to be non-issues.

[8] Consider the various tiers of Dunbar's number

[9] Consider, for instance, the prevalence of marijuana cultivation in rural northern California or moonshine production in the rural South. Both are industries in which lots of land with very few people living on it is a desirable feature.

wages (`wfed`, `wsta`, `wloc`) may be collinear with `prbconv`.[10] We theorize that both `prbpris` and `avgsen` may be collinear with some underlying omitted variable (for instance, local sentencing guidelines), but lack an alternative variable that might serve as a reasonable proxy.

Theme 3: 'The Kitchen Sink, or, everything else'

Of the 25 variables in the dataset, we incorporated five into **Model 1** and 13 into **Model 2**. Of the remaining seven, `year` is meaningless (it is a constant, 87, for all observations in the dataset) and `crmrte` is our outcome of interest. That leaves five we can incorporate into a final model that attempts to capture everything that might plausibly predict crime rate: three dummy variables describing a county's location within NC (`west`, `central`) and its degree of urbanization (`urban`), one describing the ratio of face-to-face vs. 'other' crimes (`mix`), and the county identifier (`county`), which we know is a FIPS code. The unifying theme for these variables is essentially the absence of a unifying theme: these are the leftovers.

Building of Initial Models

Model 1: Crime & Punishment

Drawing upon the transformations suggested by our exploratory data analysis and the discussion above, the first model we develop and examine is thus as follows:

$$\texttt{crmrte} = \beta_0 + \beta_1 \log(\texttt{polpc}) + \beta_2 \texttt{prbarr} + \beta_3 \texttt{prbconv} + \beta_4 \texttt{prbpris}^2 + \beta_5 \texttt{avgsen} + u \tag{1}$$

```
# Linear Regression model using only our key variables of interest,
# transformed as needed
model1_trans <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) + log(prbconv) +
    prbpris + avgsen))

# Adding AIC to our model to help us compare models in the future.
model1_trans$AIC <- AIC(model1_trans)

# Output model results in nice format using tidy and kable
kable(tidy(model1_trans))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.1264077	0.0330279	3.8273045	0.0002485
log(polpc)	0.0204502	0.0043081	4.7469395	0.0000084
log(prbarr)	-0.0243821	0.0036130	-6.7483637	0.0000000
log(prbconv)	-0.0137211	0.0026509	-5.1761024	0.0000015
prbpris	0.0088527	0.0177443	0.4989026	0.6191520
avgsen	-0.0006839	0.0005680	-1.2040260	0.2319614

Comments on Model 1:

The log of police per capita, the log of the probability of arrest, and the log of probability of conviction all have high levels of significance. The overall model has an adjusted R^2 of 0.486. There are three main points to highlight:

1) Our coefficient for log police per capita (`polpc`) is positive and highly significant. If we inaccurately

[10] The expected collinearity of government wages with `prbconv` is premised on a theorized relationship between `prbconv` as an outcome variable and the quality or effectiveness of police investigators, court-appointed advocates, judges, and other government officials *as proxied* loosely *by wage*.

assumed this model was causal the best mechanism to reduce crime rates would be to sunset the police force! However, a more plausible interpretation is that a higher number of police per capita is a response to higher levels of criminal activity, rather than a cause of it.

2) Both log probability of arrest (`prbarr`) and log probability of conviction (`prbconv`) have highly significant and negative coefficients. This fits with what we would expect: the more likely an individual is to be caught and convicted the less likely they are to commit crime.

3) Neither the probability of incarceration of the average sentence length have high levels of significance. Furthermore, the probability of prison has a positive coefficient, indicating that criminal activity increases are correlated with higher incarceration rates (ceteris perebus). This is highly counterintuitive if we were to interpret this causally.

Model 2

Model 2 builds upon Model 1 by including the additional covariates discussed earlier.

$$\begin{aligned}\texttt{crmrte} = \beta_0 &+ \beta_1 \log(\texttt{polpc}) + \beta_2 \texttt{prbarr} + \beta_3 \texttt{prbconv} + \beta_4 \texttt{prbpris}^2 + \beta_5 \texttt{avgsen} \\ &+ \beta_6 \log(\texttt{taxpc}) + \beta_7 \texttt{density} + \beta_8 \texttt{pctymle} + \beta_9 \texttt{pctmin80} + \beta_{10} \\ &+ \beta_{11} + \beta_{12} + \beta_{13} + \beta_{14} \\ &+ \beta_{15} + \beta_{16} + \beta_{17} + \beta_{18} + u\end{aligned} \tag{Model 1}$$

(2)

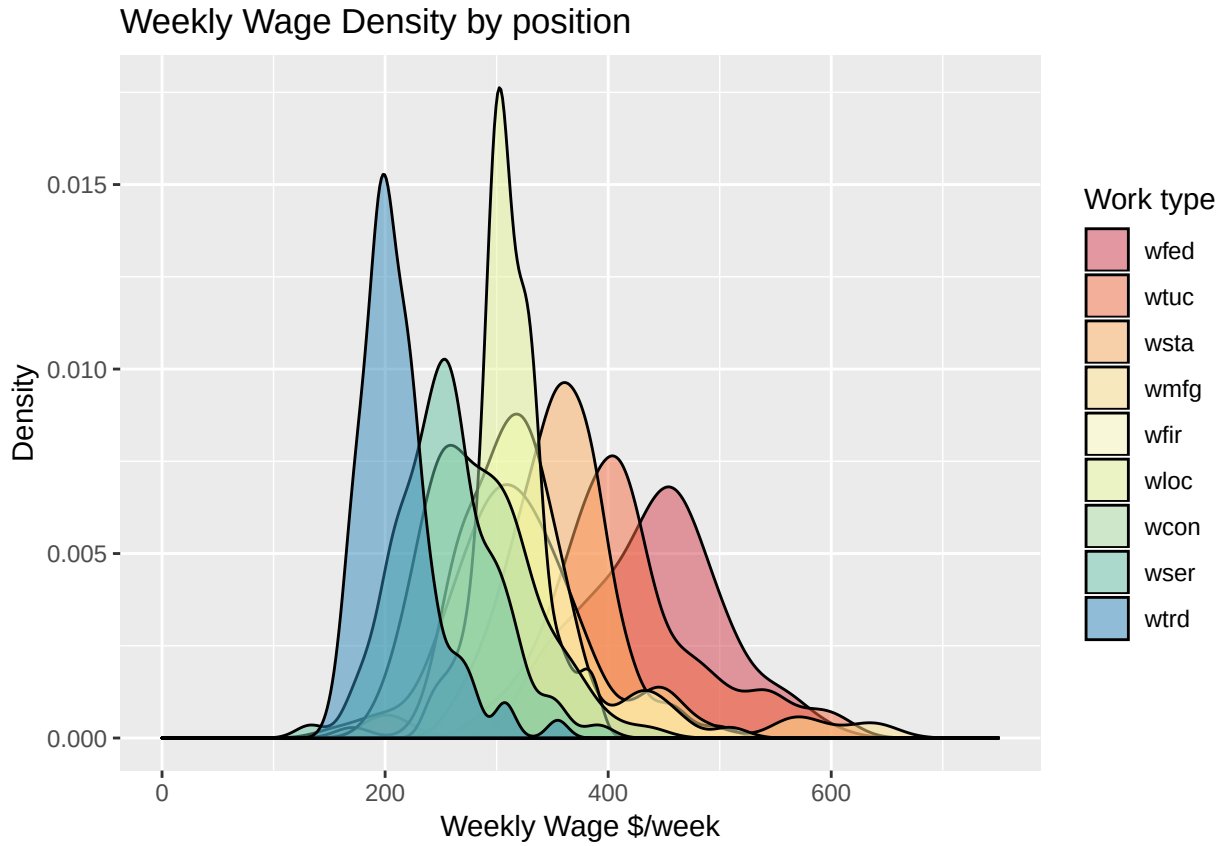
↓ TS: This section still needs to find a new home. I just haven't found it yet.

Creation of New Variables To Simplify Wage Metrics

Our dataset contains 9 wage variables, each representing a different sector or group of industries. We do not have *a priori* justification to believe *a single industry* might contribute disproportionately to crime, but we can assume that low wages in general might create an environment of economic scarcity in which crime incidence would increases. Including all 9 variables when our dataset only contains 90 observations would be extremely limiting, but excluding them entirely prohibits us from understanding how microeconomic conditions may be contributing to observed crime rates. Researching the composition of each county's economy and weighting each variable accordingly might be a fruitful strategy, but it lies outside the scope of this report. The solution we ultimately implemented was to create three new composite variables:

1) Government Wage: Average of `wfed` (federal government wage), `wsta` (state government wage), and `wloc` (local government wage)
2) Blue-Collar Wage: Average of `wmfg` (manufacturing), `wser` (service), `wcon` (construction)
3) Professional Wage: Average of `wfir` (Finance/Investment/Real Estate), `wtrd` (Wholesale/Retail Trade), and `wtuc` (Transportation, Utilities, Communication)

```
crime_na %>% gather(wfed, wtuc, wsta, wmfg, wfir, wloc, wcon, wser, wtrd, key = "work_type",
    value = "weekly_wage") %>% select(county, work_type, weekly_wage, everything()) %>%
    mutate(work_type = factor(work_type, c("wfed", "wtuc", "wsta", "wmfg", "wfir",
        "wloc", "wcon", "wser", "wtrd"))) %>% ggplot(aes(x = weekly_wage, fill = work_type)) +
    geom_density(alpha = 0.5) + scale_fill_brewer(palette = "Spectral") + ggtitle("Weekly Wage Density
    xlab("Weekly Wage $/week") + ylab("Density") + labs(fill = "Work type") +
    xlim(0, 750)
```




```
crime_na$govt_wg <- (crime_na$wfed + crime_na$wsta + crime_na$wloc)/3
crime_na$physical_wg <- (crime_na$wmfg + crime_na$wser + crime_na$wcon)/3
crime_na$industry_wg <- (crime_na$wfir + crime_na$wtrd + crime_na$wtuc)/3
```

```
# Model with our key explanatory variables, and what we suspect to be key
# covariates
model2_trans <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) + log(prbconv) +
    prbpris + avgsen + log(taxpc) + sqrt(density) + govt_wg + pctymle + pctmin80))

model2_trans_no_minvar <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) +
    log(prbconv) + prbpris + avgsen + log(taxpc) + sqrt(density) + govt_wg +
    pctymle))

added_adj_r_squared <- summary(model2_trans)$adj.r.squared - summary(model2_trans_no_minvar)$adj.r.squa

# Adding AIC to our model to help us compare models in the future.
model2_trans$AIC <- AIC(model2_trans)

# Output model results in nice format using tidy and kable
kable(tidy(model2_trans))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0479040	0.0407873	1.1744813	0.2437305
log(polpc)	0.0130076	0.0034467	3.7739224	0.0003093
log(prbarr)	-0.0145504	0.0030098	-4.8343414	0.0000065
log(prbconv)	-0.0094333	0.0020048	-4.7053652	0.0000106

term	estimate	std.error	statistic	p.value
prbpris	-0.0077756	0.0122379	-0.6353700	0.5270233
avgsen	-0.0003789	0.0003877	-0.9771417	0.3314818
log(taxpc)	0.0057067	0.0044022	1.2963339	0.1986346
sqrt(density)	0.0188890	0.0028515	6.6241504	0.0000000
govt_wg	-0.0000111	0.0000419	-0.2645387	0.7920538
pctymle	0.0461039	0.0467308	0.9865841	0.3268586
pctmin80	0.0003688	0.0000607	6.0707261	0.0000000

Comments on Model 2:

1) One item of interest was the extreme degree of significance we see associated with our percent minority variable (`pctmin80`). Interestingly, when using only this variable to predict crime rates our R^2 is very low; after controlling for other factors, however, this variable becomes extremely important. One way to measure this importance is by calculating the difference between the adjusted R^2 values for a model that includes the variable and one that excludes it. When including the minority variable in our model, the adjusted R^2 is 0.773. When excluding it, the adjusted R^2 is 0.671, a difference of 0.102

2) The square root of the population density is also highly significant. Given the transformation the correct interpretation is that crime activity increases quickly when moving from very small to medium sized population densities, but requires increasingly large levels of population increase to have an effect on crime rate.

Model 3

Model 2 yielded greater predictive value than **Model 1** by including a few extra covariates. **Model 3** throws in essentially every variable in the dataset.

$$\begin{aligned}\texttt{crmrte} = \beta_0 &+ \beta_1 \log(\texttt{polpc}) + \beta_2 \texttt{prbarr} + \beta_3 \texttt{prbconv} + \beta_4 \texttt{prbpris}^2 + \beta_5 \texttt{avgsen} && \text{(Model 1)}\\ &+ \beta_6 \log(\texttt{taxpc}) \beta_7 \texttt{density} + \beta_8 \texttt{pctymle} + \beta_9 \texttt{pctmin80} + \beta_{10}\\ &+ \beta_{11} + \beta_{12} + \beta_{13} + \beta_{14}\\ &+ \beta_{15} + \beta_{16} + \beta_{17} + \beta_{18} && \text{(Model 2)}\\ &+ \beta_{19} \texttt{mix} + \beta_{20} \texttt{central} + \beta_{21} \texttt{west} + \beta_{22} \texttt{urban} + \beta_{23} \texttt{county} + u && (3)\end{aligned}$$

Given the countless ways behavioral issues are interconnected, we wondered whether every variable we had data on might be correlated with either the crime rate or an already included variable in some fashion. Our focus was to determine if including all our variables substantially changed the significance or coefficient of any of our previously included variables. Additionally, we wanted to understand if any of the variables we had previously left out were in fact predictive overall.

```
# Linear model including our key explanatory variables, suspected
# covariates, and most other variables
model3_trans <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) + log(prbconv) +
    prbpris + avgsen + log(taxpc) + sqrt(density) + govt_wg + pctymle + pctmin80 +
    west + central + urban + physical_wg + industry_wg + mix))

# Adding AIC to our model to help us compare models in the future.
model3_trans$AIC <- AIC(model3_trans)

# Output model results in nice format using tidy and kable
kable(tidy(model3_trans))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0855755	0.0457433	1.8707776	0.0653840
log(polpc)	0.0155349	0.0035528	4.3725722	0.0000401
log(prbarr)	-0.0134649	0.0030555	-4.4067304	0.0000354
log(prbconv)	-0.0089994	0.0021761	-4.1355031	0.0000936
prbpris	-0.0043629	0.0121923	-0.3578438	0.7214928
avgsen	-0.0009032	0.0004166	-2.1679559	0.0334227
log(taxpc)	0.0037515	0.0049573	0.7567610	0.4516286
sqrt(density)	0.0174525	0.0040678	4.2904401	0.0000539
govt_wg	-0.0000155	0.0000445	-0.3485367	0.7284403
pctymle	0.0320931	0.0471316	0.6809257	0.4980728
pctmin80	0.0002928	0.0000948	3.0872972	0.0028548
west	-0.0059240	0.0039528	-1.4986898	0.1382663
central	-0.0048302	0.0028514	-1.6939831	0.0945314
urban	0.0035046	0.0057624	0.6081836	0.5449534
physical_wg	-0.0000259	0.0000157	-1.6443824	0.1043973
industry_wg	0.0000335	0.0000317	1.0581013	0.2934980
mix	-0.0185578	0.0155657	-1.1922224	0.2370353

Model 3 Notes:

Similarly to model 2, we find that log police per capita, log probability of arrest, log probability of conviction, square root of density, and percent minority are significant. While model 3 coefficients do change relative to model 2, there are not any drastic changes, or sign (+/-) changes indicating a reversal of effect.

Our adjusted R squared value for this model is 0.785 which represents a negligible improvement over our previous model.

Model 4

One item we wished to investigate is whether we could build a more parsimonious model by selectively removing variables from our second model which did not appear to be significant.

```
# Linear model including our key explanatory variables, suspected
# covariates, and most other variables
model4_trans <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) + log(prbconv) +
    sqrt(density) + pctmin80))

# Adding AIC to our model to help us compare models in the future.
model4_trans$AIC <- AIC(model4_trans)

# Output model results in nice format using tidy and kable
kable(tidy(model4_trans))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0611285	0.0181177	3.373970	0.0011230
log(polpc)	0.0134516	0.0027352	4.918024	0.0000043
log(prbarr)	-0.0162011	0.0027616	-5.866582	0.0000001
log(prbconv)	-0.0108138	0.0018482	-5.850964	0.0000001
sqrt(density)	0.0184354	0.0023000	8.015235	0.0000000
pctmin80	0.0003842	0.0000572	6.719658	0.0000000

Model Comparison

Below is a comparison table for our three initial models, as well as the fourth model that excluded all non-significant predictors. The table reports key statistics related to each model, including the coefficients for each predictor, the R^2 and adjusted R^2, and the *Akaike information criterion*, or AIC.

Our findings match with what we would hope to see from a model building perspective: Our AIC and adjusted R^2 numbers suggest that of our three original models, *Model 2* (which includes both our key explanatory variables and plausible covariates) performs the best. Our model which excludes key covariates has significantly less predictive power (as measured by adjusted R^2), and our model which includes everything–despite having the highest R^2 values–performs slightly less well on the AIC (a measure of explanatory power and parsimony).

```
stargazer(model1_trans, model2_trans, model3_trans, model4_trans, type = "latex",
    report = "vc", header = FALSE, title = "Linear Models Predicting Crime Rate",
    keep.stat = c("aic", "rsq", "adj.rsq", "n"), omit.table.layout = "n")
```

Omitted Variables

Despite the promising results from our three models it is difficult to ascribe causality to the variables of interest. One issue with causal inference in general is ommitted variable bias, which can invalidate our ability to assume each explanatory variable is uncorrelated with the error term. While there are infinite variables which exist, there are several which deserve commentary:

1) Political Party in Control: Traditionally the issue of crime policy is a highly partisan issue, with each party having very different approaches to crime reduction. All else being equal, we might expect police levels and average sentence lengths to be correlated with the party that is crafting legislation. Assuming we construct our variable as a boolean indicator ("is_republican"), we might expect the coefficient for police per capita to diminish as we assume a priori that higher police levels are correlated with conservative crime policies. We could include this data by appending public records to our dataset which would be more appropriate than trying to find some other variable to proxy.

2) Unemployment Rate: While we have data on weekly wages, this does nothing to tell us what percentage of the population was actually earning those wages. It is likely that a higher unemployment rate would be correlated with higher rates of crime as people who may not normally commit criminal activity are pushed to their limits. We might also wish to be more granular, and include both minority and majority unemployment rates to help control for racial inequality. We realistically could obtain this data from the Bearue of Labor Statistics and append it to our dataset; we leave this next step to future researchers.

3) Concentrated/Siloed Urban Blight: Our data is at the county level and therefore may obscure differences within the county. We would expect there to be a difference in crime rates between a county which is relatively homogenous with respect to the variables, and one which has drastic differences (e.g. a very poor area and a very nice area). One way to proxy this might be to calculate a normalized standard deviation of housing prices which could help capture if this phenomenom exists.

4) Policing Methodology: In recent years there has been a growing focus on "warrior" versus "guardian" mindsets in policing. Depending on the type of methodology we might see very different rates of arrest and conviction.

5) Police Representation: A community's relationship with the police can vary drastically from place to place. In recent years certain cities have had significant unrest, with a key element being a majority white police force existing in a largely minority community. Similar to police methodology, we might expect very different rates of arrest and conviction depending on whether the community feels the police represent them and their interests.

Table 6: Linear Models Predicting Crime Rate

	Dependent variable:			
	crmrte			
	(1)	(2)	(3)	(4)
log(polpc)	0.020	0.013	0.016	0.013
log(prbarr)	−0.024	−0.015	−0.013	−0.016
log(prbconv)	−0.014	−0.009	−0.009	−0.011
prbpris	0.009	−0.008	−0.004	
avgsen	−0.001	−0.0004	−0.001	
log(taxpc)		0.006	0.004	
sqrt(density)		0.019	0.017	0.018
govt_wg		−0.00001	−0.00002	
pctymle		0.046	0.032	
pctmin80		0.0004	0.0003	0.0004
west			−0.006	
central			−0.005	
urban			0.004	
physical_wg			−0.00003	
industry_wg			0.00003	
mix			−0.019	
Constant	0.126	0.048	0.086	0.061
Observations	90	90	90	90
R^2	0.524	0.798	0.824	0.787
Adjusted R^2	0.495	0.773	0.785	0.775
Akaike Inf. Crit.	−512.807	−580.148	−580.222	−585.459

6) Criminals' Perception of Risk-Reward Ratios: Of particular relevance to the Model 1 ("Crime and Punishment") analysis is the issue of criminals' perception of the "effectiveness'"of the justice system. The model makes two critical assumptions: first, that criminality is driven by rational behavior; second, that criminals have sufficient insight into the outcomes of criminal justice proceedings for their decisions (which we are assuming to be rational) to be informed by them. Obviously if criminals are not of the genus *Homo economicus*, a major theoretical underpinning of the model is removed. However, even if criminals *are* rational actors, they may be operating with imperfect information: they may make the "incorrect" risk-reward calculation if they are unaware of how risky their criminality truly is. For instance, a thief who does not know the District Attorney always 'gets her man' may not be deterred by a high `prbconv` as our model might assume. Similarly, the salience of the potential punishment **<- TS: Need to finish this thought...**

7)

Conclusion: Findings and Policy Recommendations

Unfortunately, we only have a single cross section of data at our disposal and therefore it is nearly impossible to determine whether our variables are causes or effects. In reality it is most likely a combination of both, with changes in crime rate driving changes in policy, which in turn impact crime rates, ad infinitum. Because our ability to conduct causal inference is restricted, it would be unwise to make specific policy reccomendations based on our model.

We can recommend to our candidate that she advocate for increased investment in research that will investigate the relationships between crime rates and the following factors:

1) The number of police per capita
2) The likelihood that crimes, once reported, result in arrests
3) The likelihood that arrests result in convictions
4) The population density in a given area
5) The demographic composition in a given area

The relationships suggested by our models are intuitive: an increase in the number of police per capita and the population density in an area are associated with (predictive of, in our model) increases in the crime rate; increases in the probability of arrest and probability of conviction are associated with decreases in the crime rate. An increase in the percentage of minority residents is associated with an increase in the crime rate.

Four of them are simple to interpret: a 1% increase in the number of police is associated with an increase in the crime rate of 11.1 crimes per 100,000 people. Increasing by 1% the probability that a crime report leads to an arrest is associated with a decrease in the crime rate of 15.9 crimes per 100,000 people; a similar increase in the probability of conviction is associated with a decrease in the crime rate of 10.2 crimes per 100,000 people. A 1% increase in the minority share of the population is associated with an increase of 3.8 crimes per 100,000 people. The relationship with the population density is more complex, however, and bears further investigation

Appendix

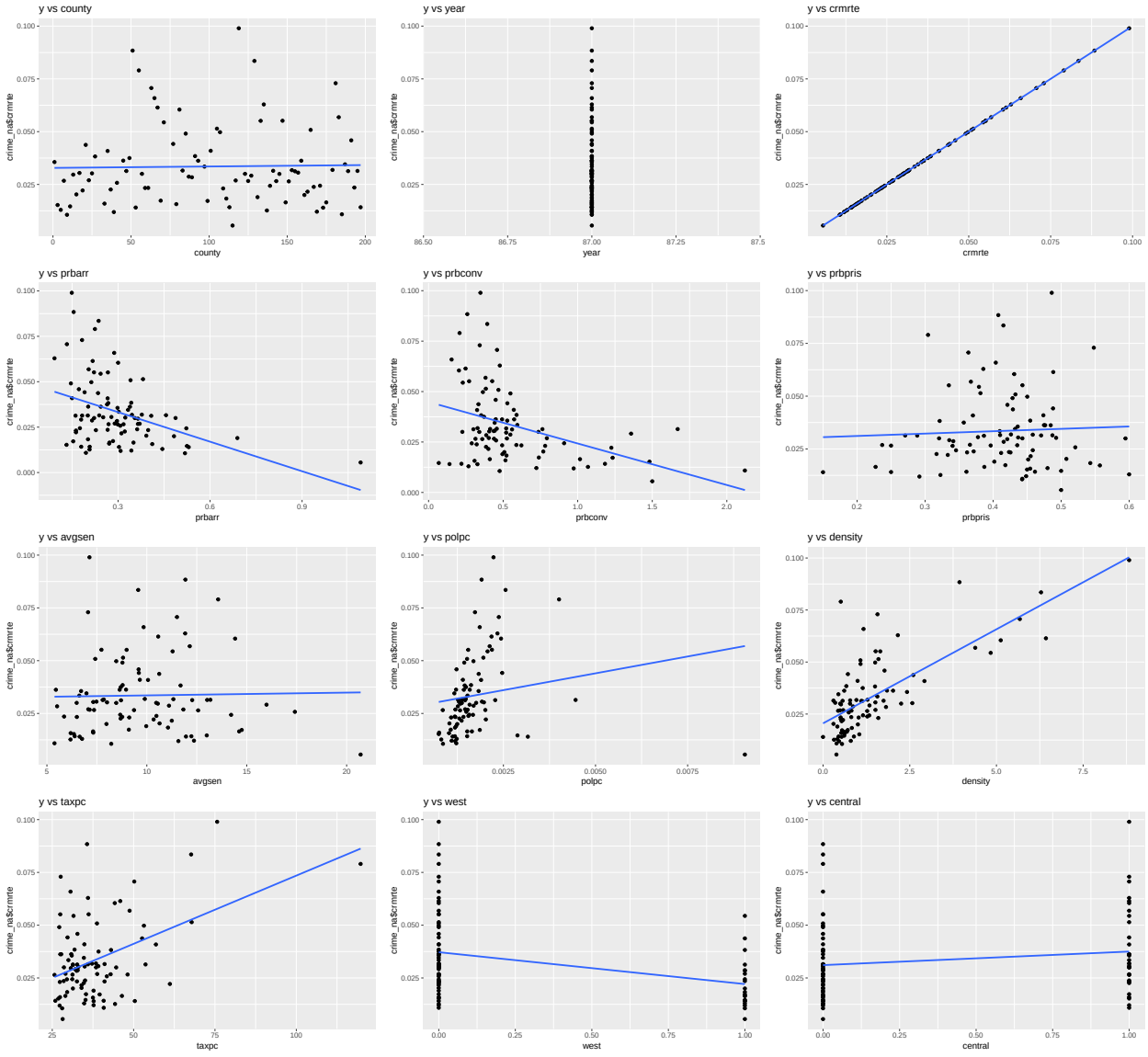
Detail: scatterplots of variables generated during EDA.

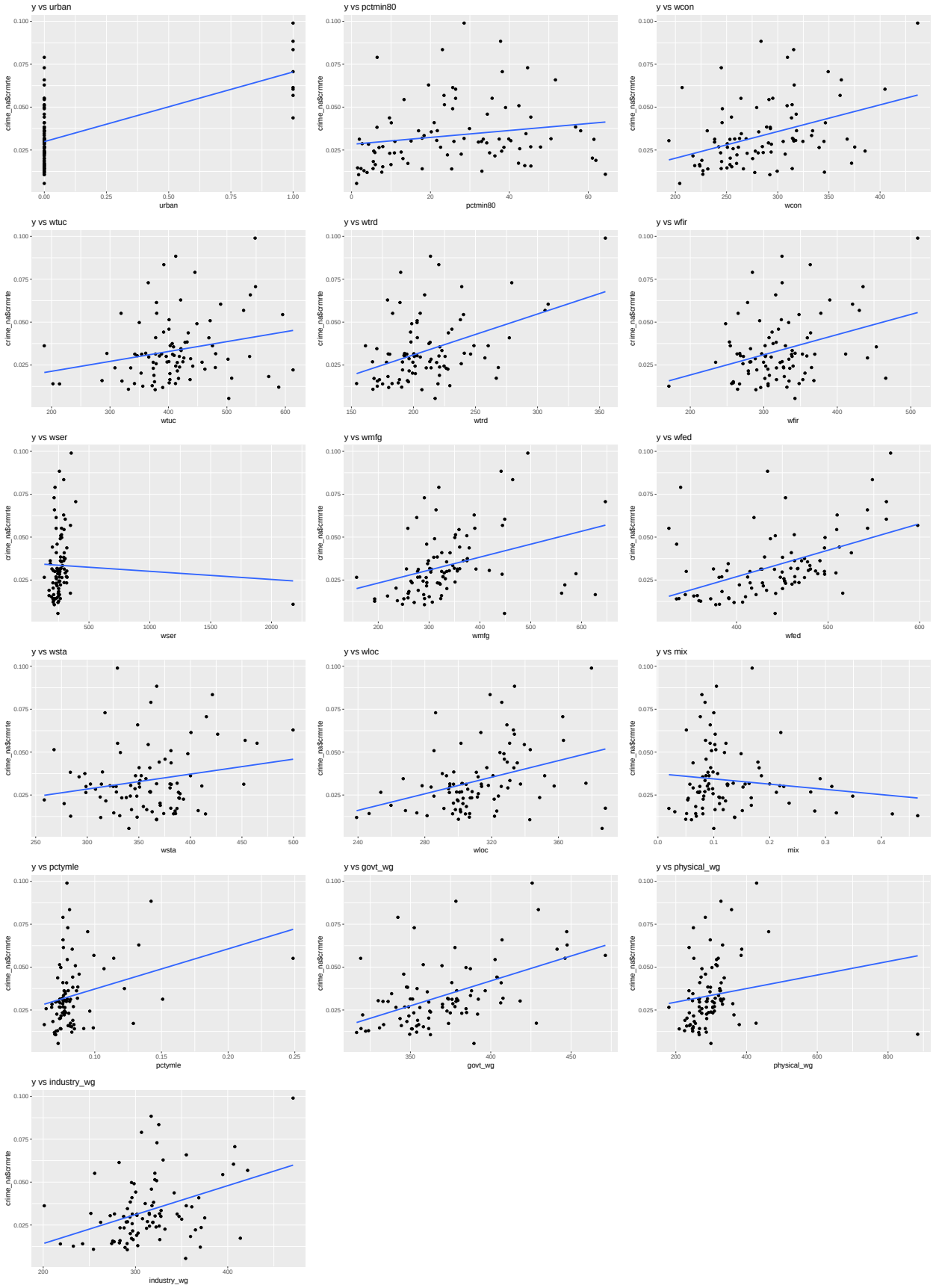
```
make_scatters <- function(df, var_list, y, trans) {

    if (!missing(trans)) {
        var_list <- append(var_list, str_glue("{trans}({var_list})"))
    }

    for (v in var_list) {
        print(ggplot(df, aes_string(x = v, y = y)) + geom_point() + geom_smooth(method = "lm",
            se = FALSE) + xlab(v) + ylab(deparse(substitute(y))) + ggtitle(str_glue("{deparse(substitut
    }
}

make_scatters(crime_na, var_list = names(crime_na), crime_na$crmrte)
```




y vs urban
y vs pctmin80
y vs wcon
y vs wtuc
y vs wtrd
y vs wfir
y vs wser
y vs wmfg
y vs wfed
y vs wsta
y vs wloc
y vs mix
y vs pctymle
y vs govt_wg
y vs physical_wg
y vs industry_wg
crime_na$crmrte
urban
pctmin80
wcon
wtuc
wtrd
wfir
wser
wmfg
wfed
wsta
wloc
mix
pctymle
govt_wg
physical_wg
industry_wg

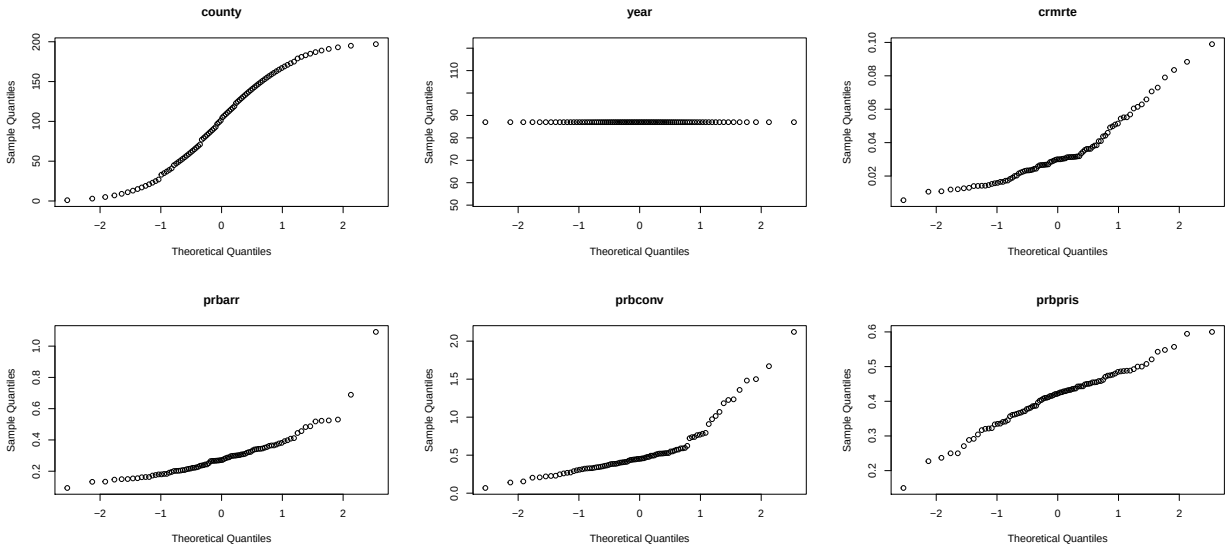
Detail: observations in which the 'probability' variables did not behave as probabilities insofar as they fell outside the range of 0:1.

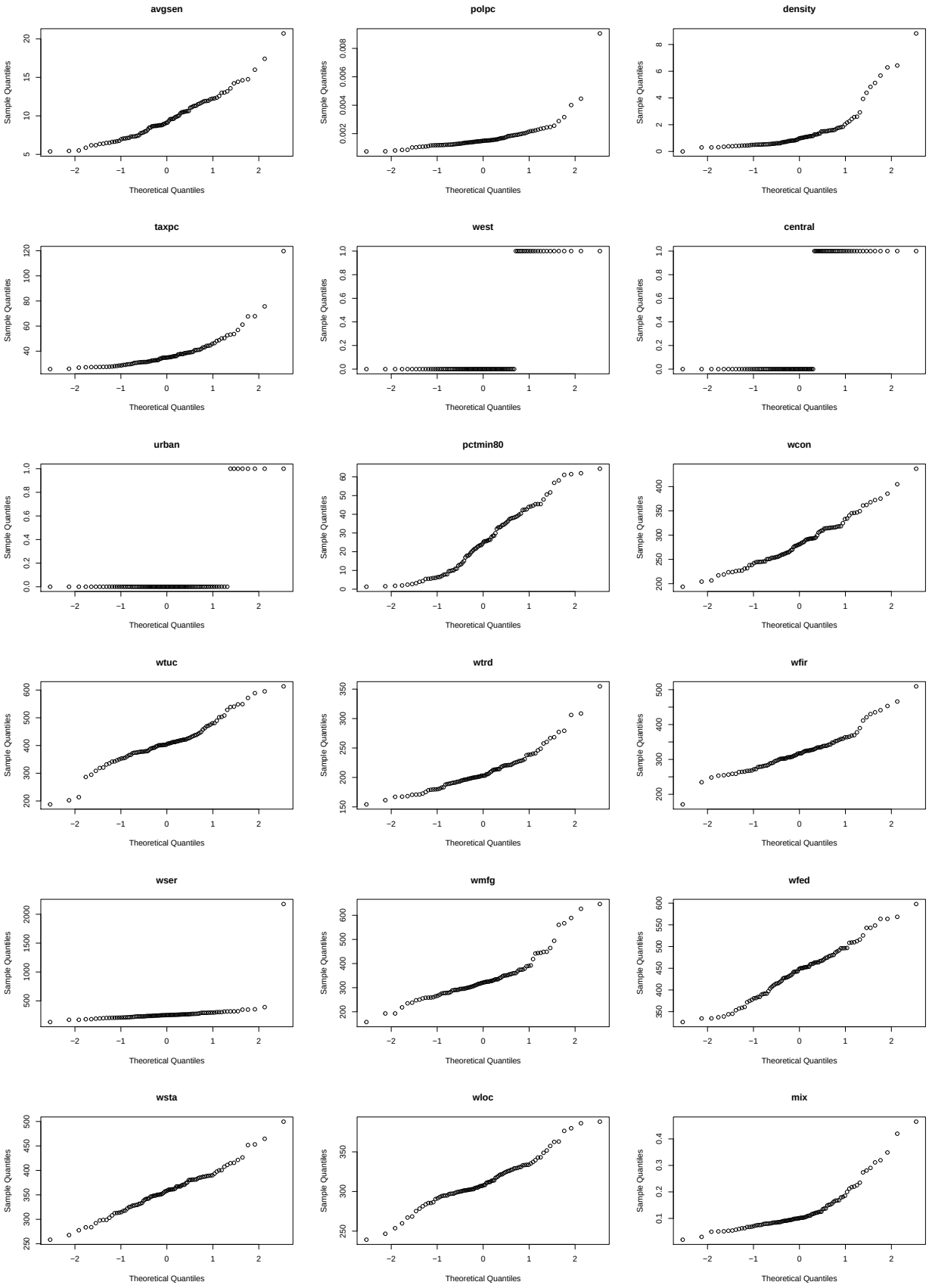
```
non_prob
```

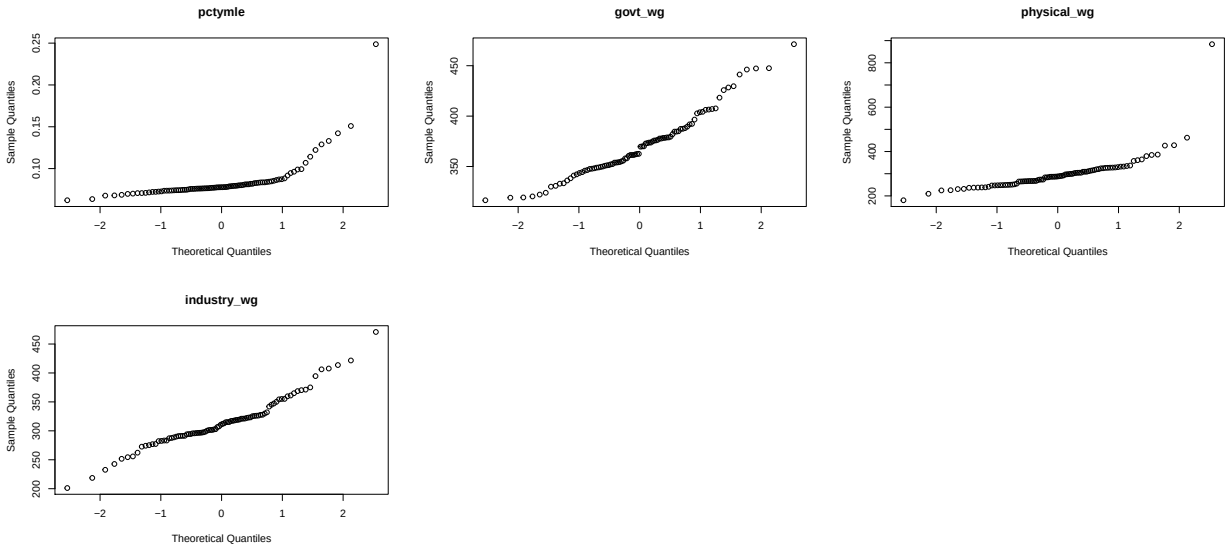
```
## # A tibble: 10 x 25
##    county  year  crmrte prbarr prbconv prbpris avgsen    polpc density
##     <int> <int>   <dbl>  <dbl>   <dbl>   <dbl>  <dbl>    <dbl>   <dbl>
##  1      3    87 0.0153   0.132    1.48   0.450   6.35 0.000746   1.05
##  2     19    87 0.0222   0.163    1.23   0.333  10.3  0.00202    0.577
##  3     99    87 0.0172   0.154    1.23   0.557  14.8  0.00186    0.548
##  4    115    87 0.00553  1.09     1.5    0.5    20.7  0.00905    0.386
##  5    127    87 0.0291   0.180    1.36   0.336  16.0  0.00158    1.34
##  6    137    87 0.0127   0.207    1.07   0.323   6.18 0.000814   0.317
##  7    149    87 0.0165   0.272    1.02   0.227  14.6  0.00152    0.609
##  8    185    87 0.0109   0.195    2.12   0.443   5.38 0.00122    0.389
##  9    195    87 0.0314   0.201    1.67   0.471  13.0  0.00446    1.75
## 10    197    87 0.0142   0.208    1.18   0.361  12.2  0.00119    0.890
## # ... with 16 more variables: taxpc <dbl>, west <int>, central <int>,
## #   urban <int>, pctmin80 <dbl>, wcon <dbl>, wtuc <dbl>, wtrd <dbl>,
## #   wfir <dbl>, wser <dbl>, wmfg <dbl>, wfed <dbl>, wsta <dbl>,
## #   wloc <dbl>, mix <dbl>, pctymle <dbl>
```

Detail: code to generate qqplots for evaluation of normality in EDA step

```
for (i in 1:length(colnames(crime_na))) {
    column_interest <- paste("crime_na$", colnames(crime_na)[i], sep = "")
    qqnorm(eval(parse(text = column_interest)), main = colnames(crime_na)[i])
}
```

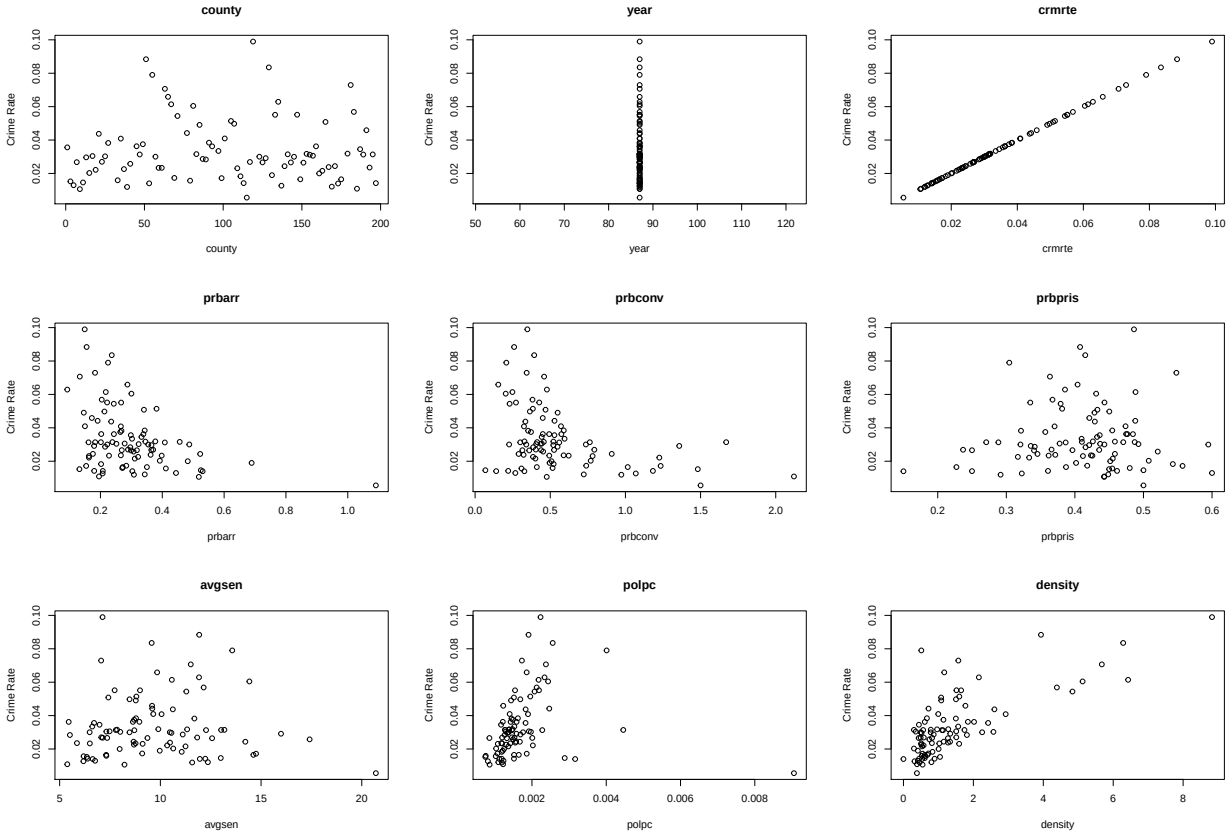


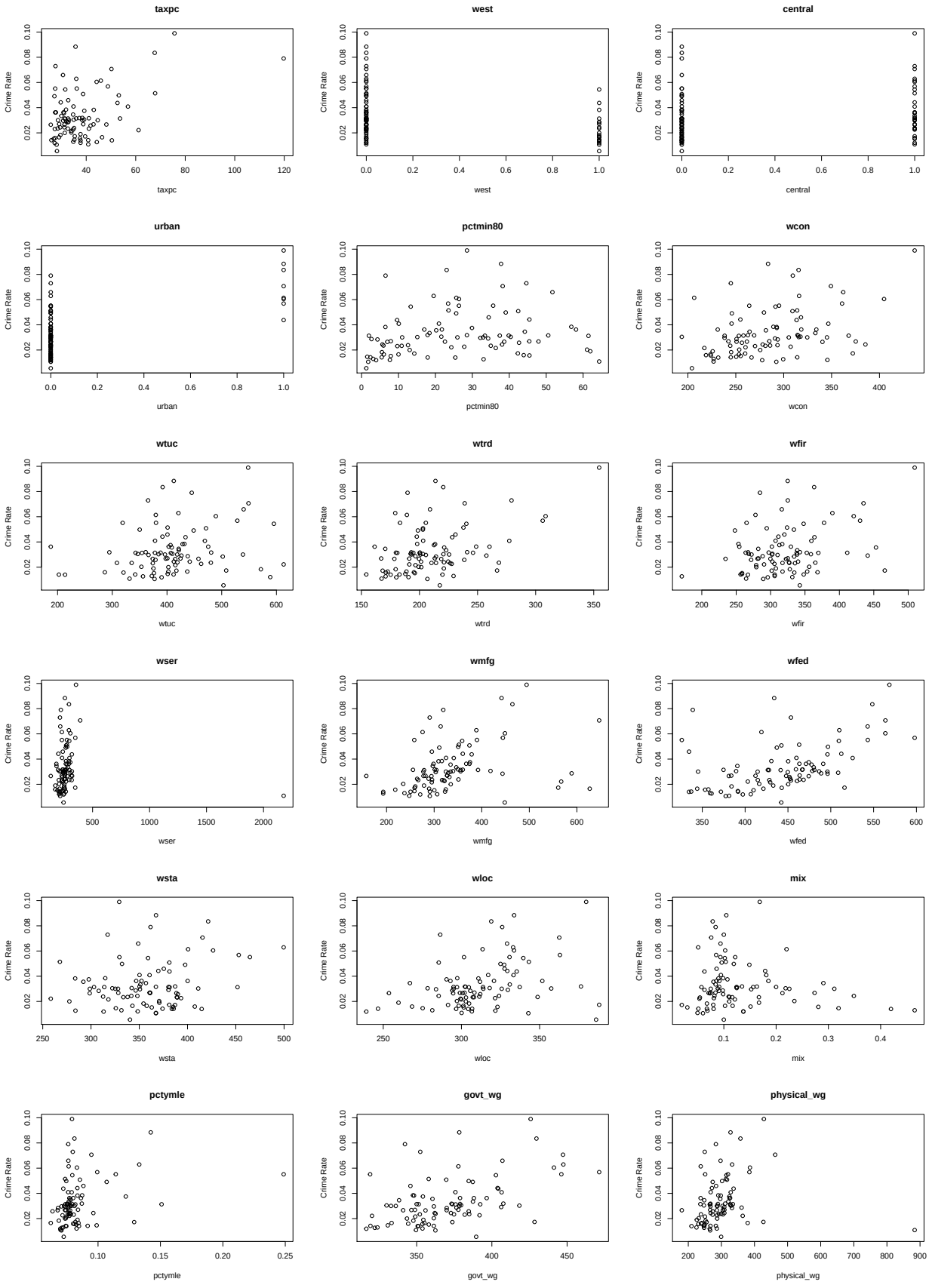

avgsen
polpc
density
taxpc
west
central
urban
pctmin80
wcon
wtuc
wtrd
wfir
wser
wmfg
wfed
wsta
wloc
mix
Sample Quantiles
Theoretical Quantiles

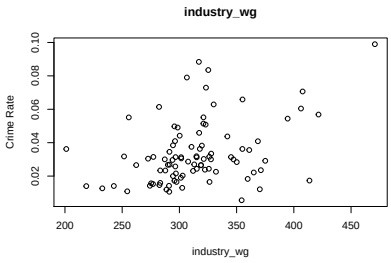


Detail: code to explore 'crime rate' vs. specific variables of interest

```
for (i in 1:length(colnames(crime_na))) {
    column_interest <- paste("crime_na$", colnames(crime_na)[i], sep = "")
    plot(eval(parse(text = column_interest)), crime_na$crmrte, main = colnames(crime_na)[i],
        ylab = "Crime Rate", xlab = colnames(crime_na)[i])
}
```

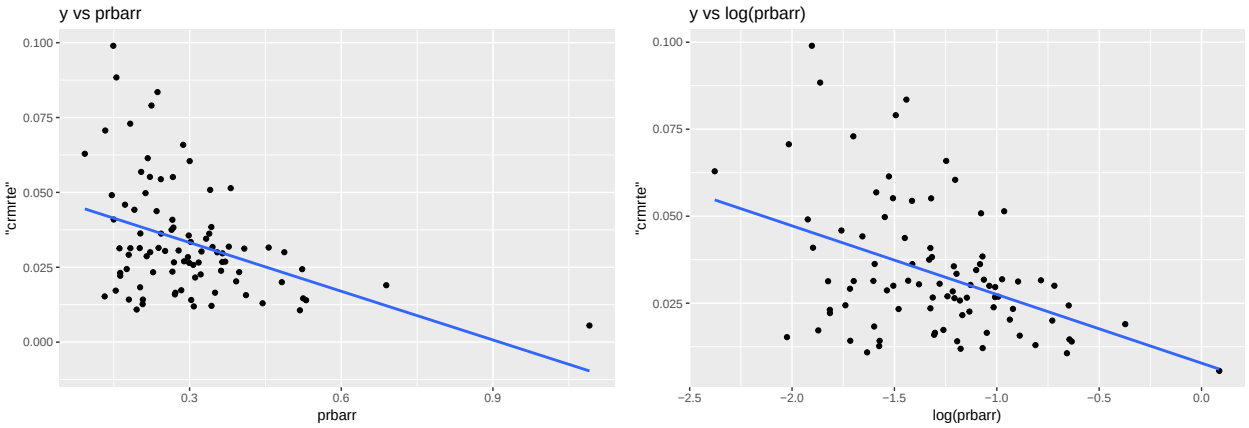



taxpc
Crime Rate
taxpc
west
Crime Rate
west
central
Crime Rate
central
urban
Crime Rate
urban
pctmin80
Crime Rate
pctmin80
wcon
Crime Rate
wcon
wtuc
Crime Rate
wtuc
wtrd
Crime Rate
wtrd
wfir
Crime Rate
wfir
wser
Crime Rate
wser
wmfg
Crime Rate
wmfg
wfed
Crime Rate
wfed
wsta
Crime Rate
wsta
wloc
Crime Rate
wloc
mix
Crime Rate
mix
pctymle
Crime Rate
pctymle
govt_wg
Crime Rate
govt_wg
physical_wg
Crime Rate
physical_wg

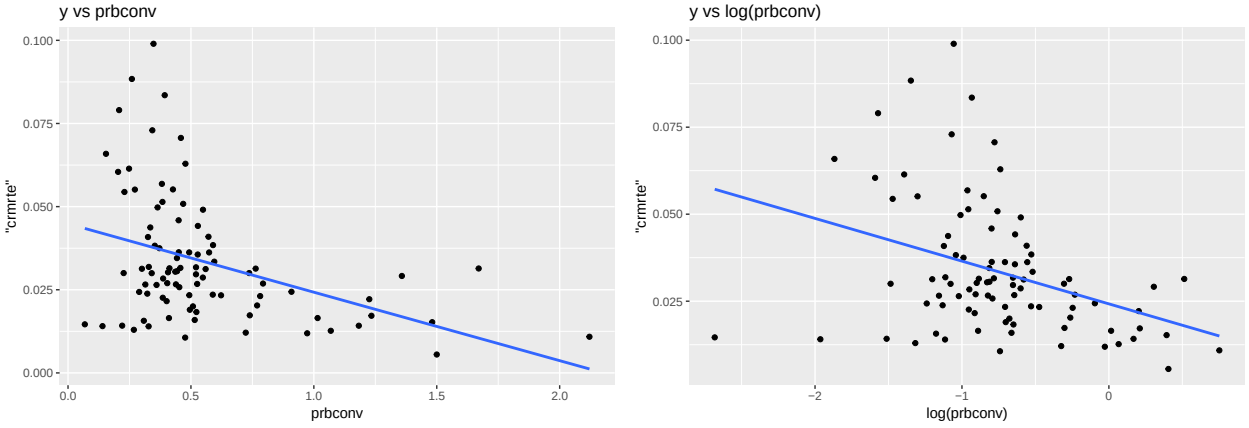


Detail: pre- and post-transformation scatterplots and q-q plots

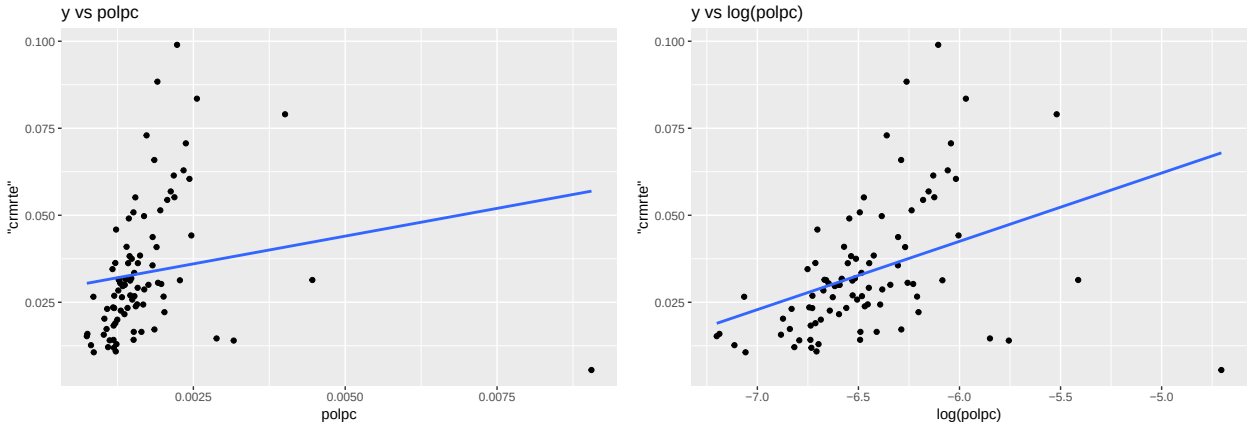
```
make_scatters(df = crime_na, var_list = c("prbarr"), y = "crmrte", trans = "log")
```



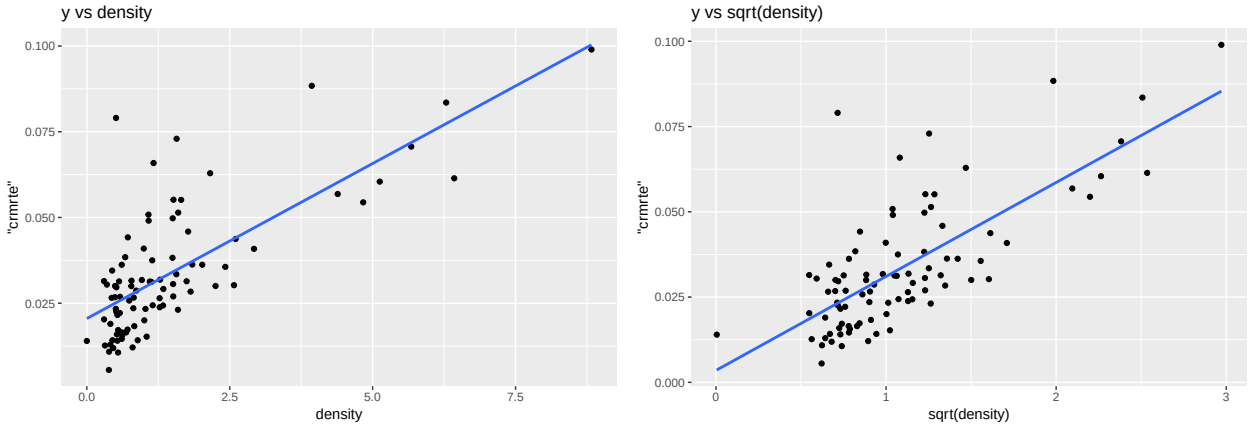

```
make_scatters(df = crime_na, var_list = c("prbconv"), y = "crmrte", trans = "log")
```



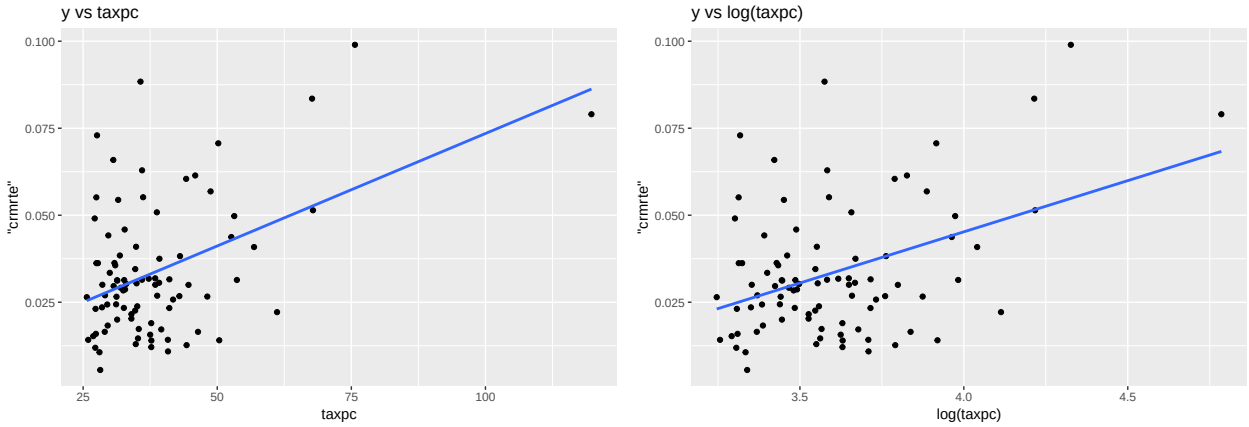

```
make_scatters(df = crime_na, var_list = c("polpc"), y = "crmrte", trans = "log")
```



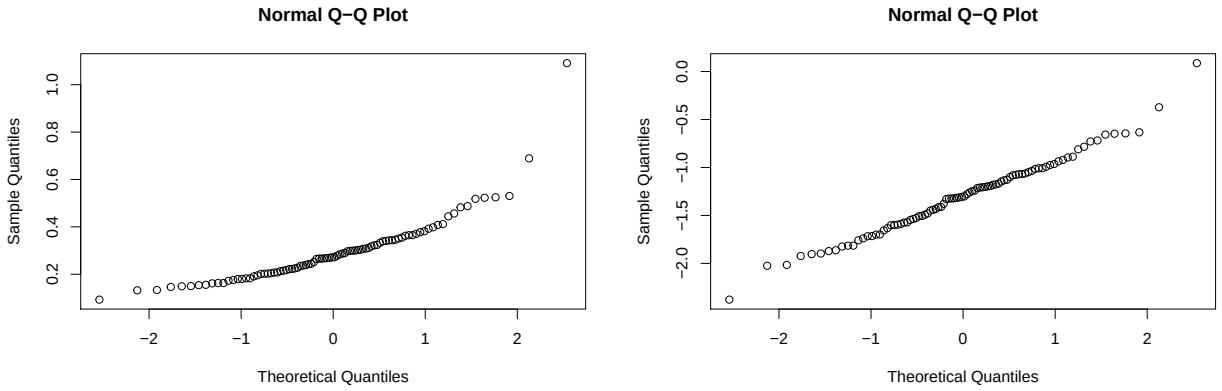

```
make_scatters(df = crime_na, var_list = c("density"), y = "crmrte", trans = "sqrt")
```



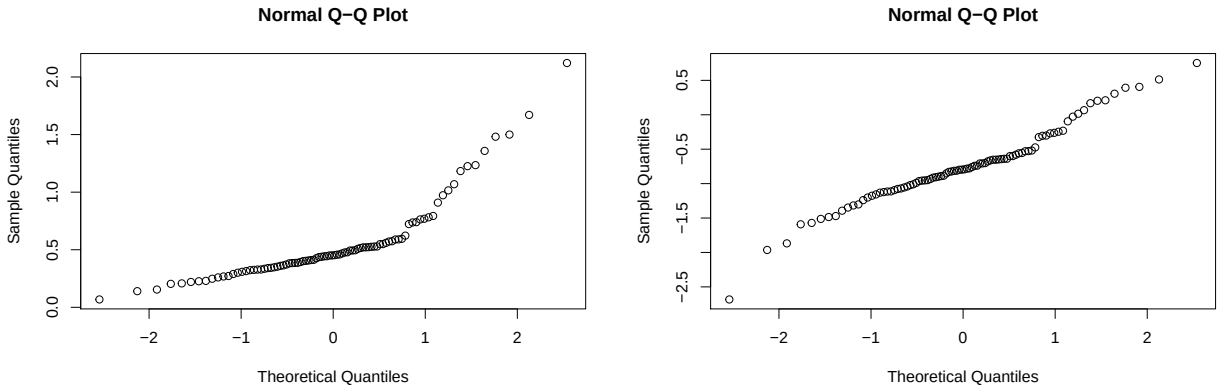

```
make_scatters(df = crime_na, var_list = c("taxpc"), y = "crmrte", trans = "log")
```



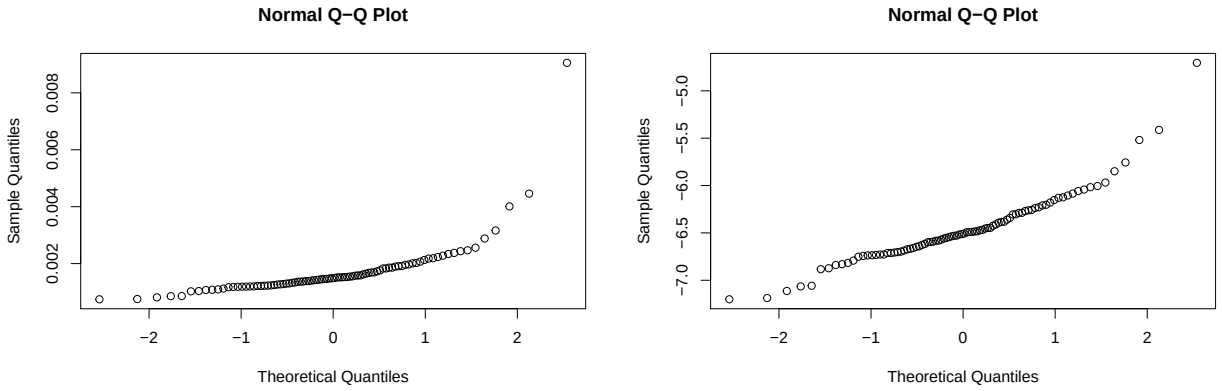

```
# Normality/skew with and without transformations
print(c(qqnorm(crime_na$prbarr), qqnorm(log(crime_na$prbarr))))
```



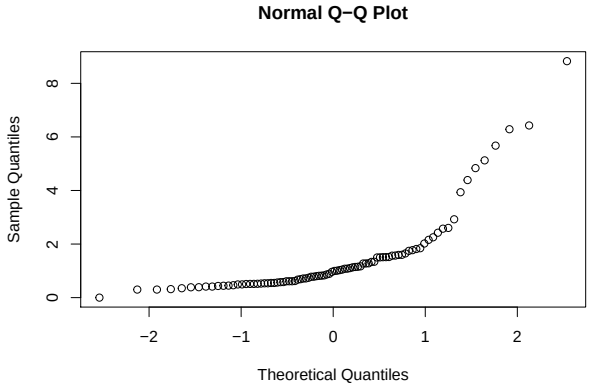

```
print(c(qqnorm(crime_na$prbconv), qqnorm(log(crime_na$prbconv))))
```

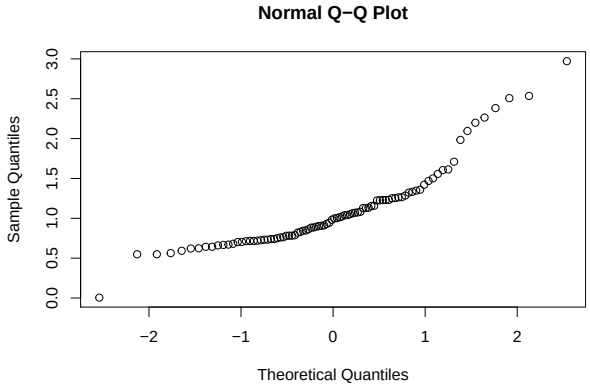



```
print(c(qqnorm(crime_na$polpc), qqnorm(log(crime_na$polpc))))
```




```
print(c(qqnorm(crime_na$density), qqnorm(sqrt(crime_na$density))))
```






```
print(c(qqnorm(crime_na$taxpc), qqnorm(log(crime_na$taxpc))))
```