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 1994 dataset from Cornwell & Trumball that provides county-level economic, demographic, and crime data. Our analysis describes the dataset, presents initial summary statistics, and develops three linear regression models.

Initial Exploratory Data Analysis (EDA)

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 2)
```

```
## Warning: 1 parsing failure.
## row # A tibble: 1 x 5 col     row col     expected actual file           expected   <int> <chr>   <ch
```

Missing Values

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

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

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

Upon loading the data, we examine the 6 rows that are missing data, finding that 5 are entirely blank and 1 contains only a backtick. We eliminate those to generate our working dataset.

Erroneous Duplicate Records

```
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 %>% filter(county == 193)
```

```
## # A tibble: 2 x 25
##   county  year crmrte prbarr prbconv prbpris avgsen   polpc density taxpc
##    <int> <int>  <dbl>  <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl> <dbl>
```

```
## 1    193    87 0.0235  0.266   0.589   0.423   5.86 0.00118   0.814  28.5
## 2    193    87 0.0235  0.266   0.589   0.423   5.86 0.00118   0.814  28.5
## # ... with 15 more variables: 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>
```

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

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

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.

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

Examining the data, 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 the values we observe. Therefore we will not drop these records from our dataset. We will, however, subject them to further scrutiny.

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 that individuals in the service industry are making 5-10x more than anyone else in the county, and therefore we will replace this value with an NA.

–Talk about 1 row with probabiltiy of arrest, conviction, and police as outliers.

One of the observations, however, appears to be an outlier for some of our variables of interest. 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 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. **We will test the robustness of our models by verifying how they change if this observation is retained instead of being treated as an outlier.**

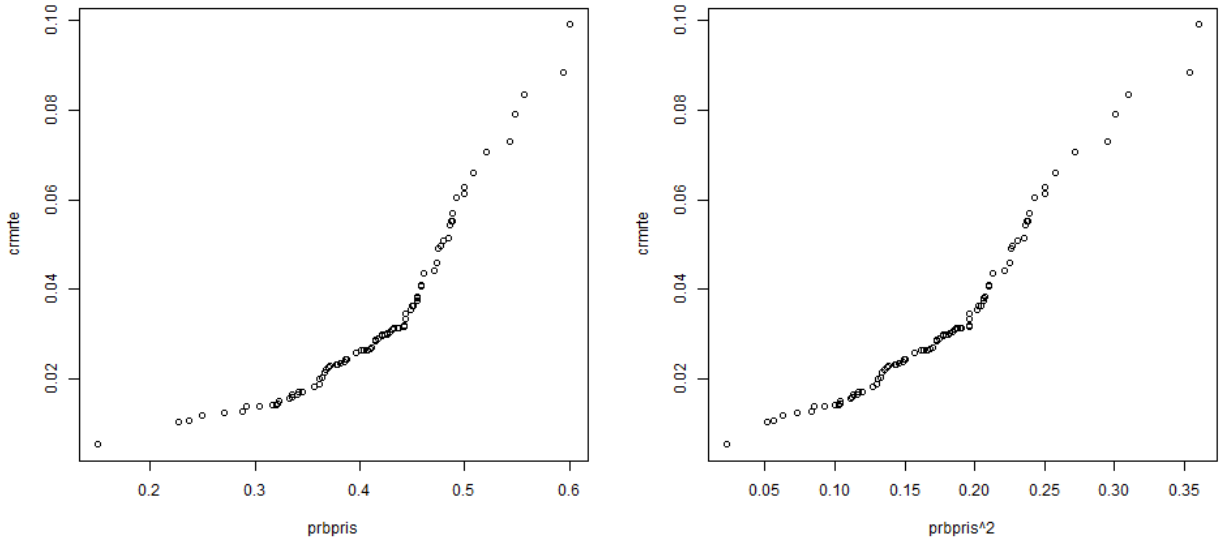
Transformation Analysis

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. It is therefore important to explore whether the relationship between two variables is logarithmic, exponential or otherwise and thus whether a transformation is required. As part of our EDA we explored this question for all of the variables in the dataset. For the sake of parsimony we discuss below only those variables which required a transformation.

–Evaluate the scatter of each of the 5 variables to see if a transform is necessary. –Final table with new transforms. Police should be log transformed. probability of prison should be squared

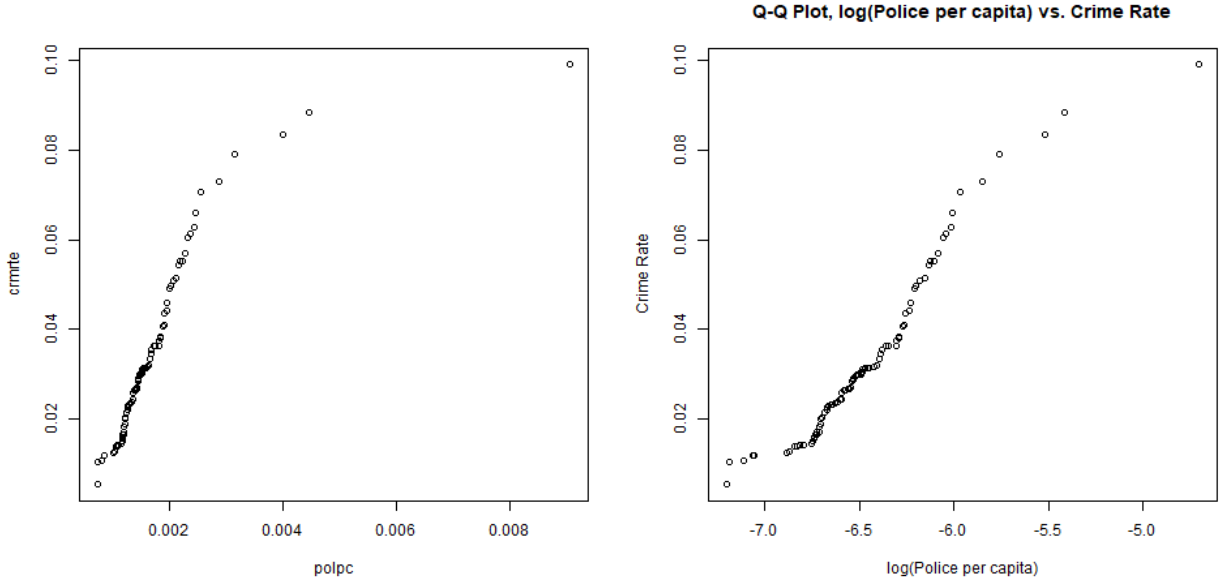
As can be seen below, the q-q plot of `prbpris` and `crmrte` appears to display some curvature, suggesting a parabola. Transforming the `prbpris` variable by squaring it appears to account for some of that curvature, so we will apply that transformation before including it in our models.

```
knitr::include_graphics(c("qq_prbpris_base.png", "qq_prbpris_sq.png"))
```



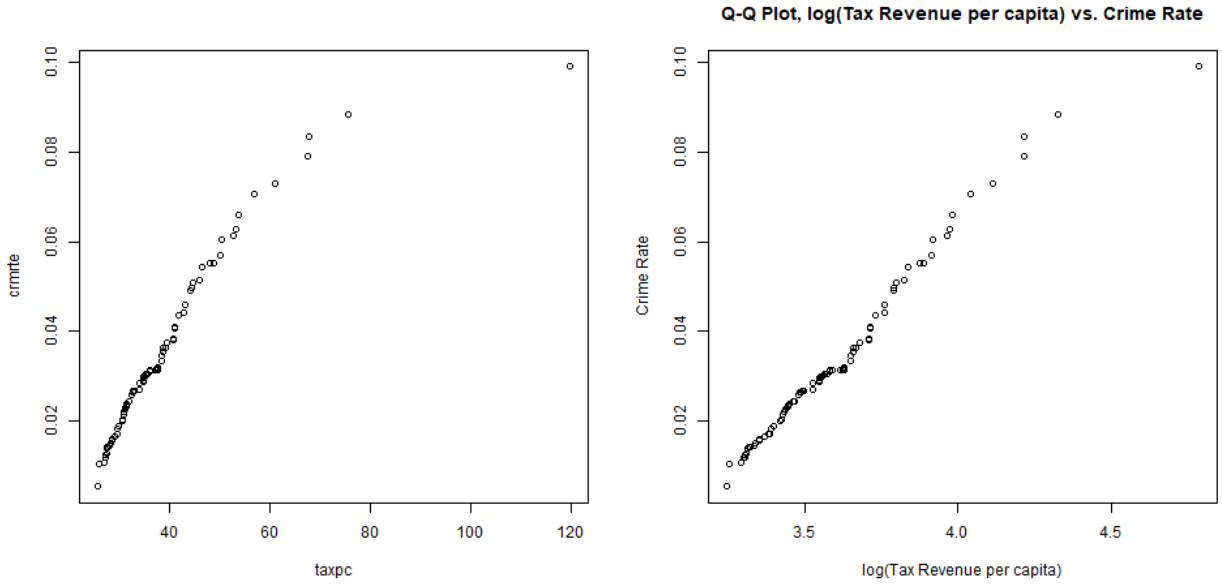
The q-q plot of `polpc` and `crmrte` indicates that the distributions from which they were sampled deviate at higher values of `polpc`. A log transformation appears to improve the fit.

```
knitr::include_graphics(c("qq_polpc_base.png", "qq_polpc_log.png"))
```



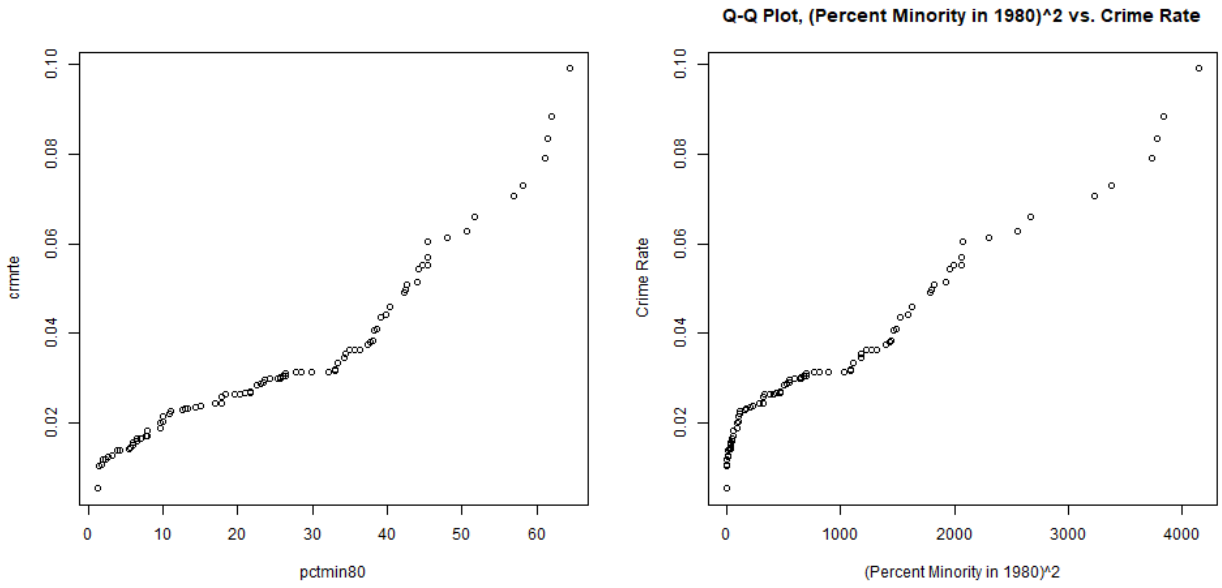
The q-q plot of tax revenue per capita and crime rate appears to be highly affected by the presence of an outlier. Applying a log transformation to the tax revenue variable improves the fit without requiring the outlier to be dropped. **Once we begin constructing models we will verify whether they are robust to retaining this outlier. [TS: <- To-do during cleanup / QC...]**

```
knitr::include_graphics(c("qq_taxpc_base.png", "qq_taxpc_log.png"))
```



Discuss percent minority transformation here...GUYS: this is obviously the wrong transformation, but it's here as a placeholder. Take a look at the qq plot down at the end of the document. pctmin80 seems to have a clear sigmoid shape...what would be the right transformation there? Both log and square have been bad, so I'm thinking maybe we just don't transform at all? See below...

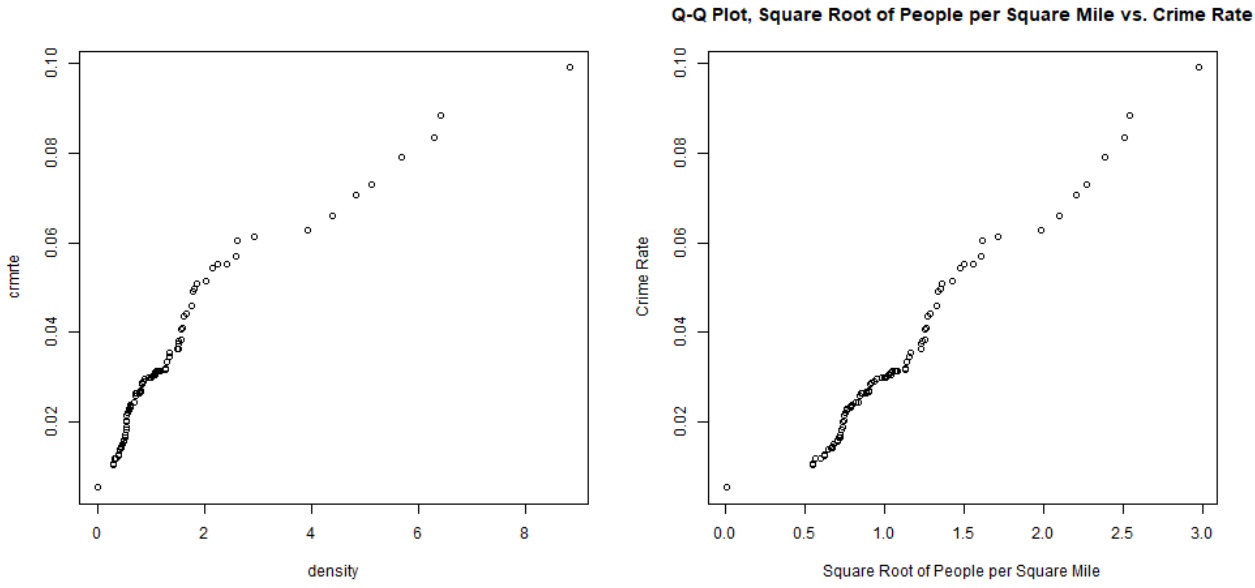
```
knitr::include_graphics(c("qq_pctmin80_base.png", "qq_pctmin80_log.png"))
knitr::include_graphics(c("qq_pctmin80_base.png", "qq_pctmin80_sq.png"))
```



The q-q plot of population density (persons per square mile) and crime rate appears to be an ill fit. Transforming the population density by taking the square appears to correct the misalignment, albeit with an outlier that appears to be very rural and have a low crime rate.

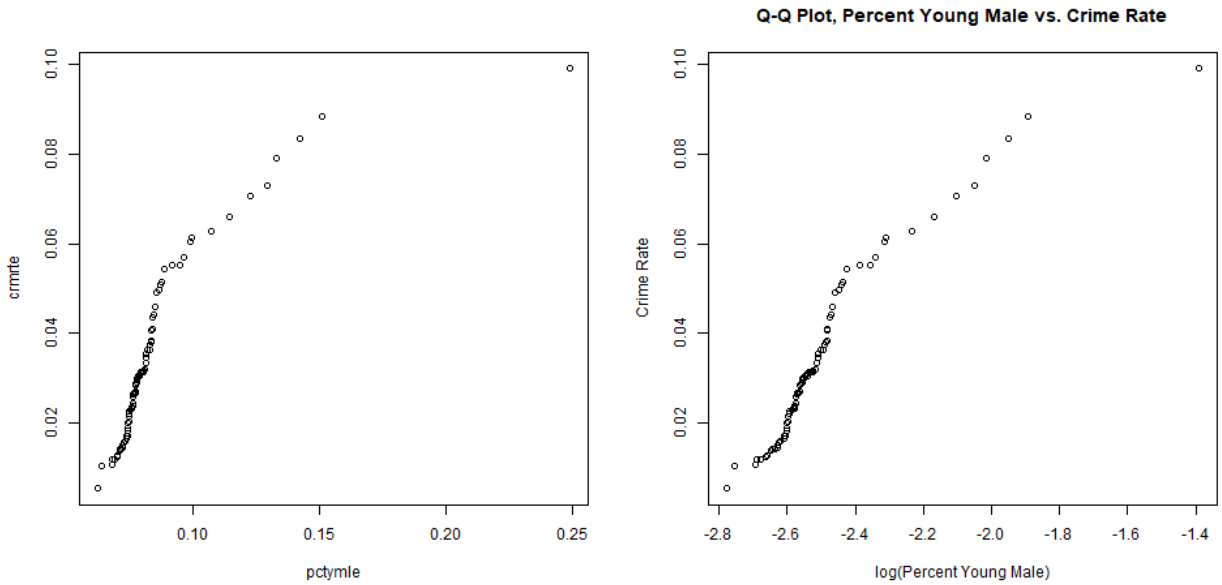
Guys: I'm not sure what the square root of persons per square mile means. Is this one where we forego the transformation because of the lack of interpretability? Also, I think the untransformed version seems to have more "outliers" than the square root version, but I'm def. not married to it.

```
knitr::include_graphics(c("qq_density_base.png", "qq_density_sqrt.png"))
```



The q-q plot of the percentage of young males and the crime rate appears to reveal a non-linear relationship. Applying a log transformation to the percent young male variable appears to improve the fit.

```
knitr::include_graphics(c("qq_pctymle_base.png", "qq_pctymle_log.png"))
```



Variables Available for Analysis

The table below details the variables available to us, which model(s) we included them in, and any transformations we applied before including them in our model.

To do: remove the 'used in models...' column, just add that information in the discussion before each model

Table 1: Hypothesized Primary Determinants of Observed Crime Rate

Variable Name	Description	Transformation Applied	Models Using
`county`	*Source county of data*	-	-
`year`	*Source year of data*	-	-
`crmrte`	*crime rate*	-	1, 2, 3, 4, 5
`prbarr`	*'probability' of arrest*	-	1, 2, 3, 4, 5
`prbconv`	*'probability' of conviction*	-	1, 2, 3, 4, 5
`prbpris`	*'probability' of prison sentence*	square ($prbpris^2$)	1, 2, 3
`avgsen`	*average sentence, in days*	-	2, 3
`polpc`	*police per capita*	log($polpc$)	1, 2, 3, 4, 5
`density`	*people per sq. mile*	$\sqrt{(density)}$	2, 3, 4
`taxpc`	*tax revenue per capita*	log($taxpc$)	2, 3
`west`	*Dummy: source county of data is in Western NC*	-	3
`central`	*Dummy: source county of data is in Central NC*	-	3
`urban`	*Dummy: source county of data is urban*	-	3
`pctmin80`	*percent minority in 1980*	TBD-Guys??	2, 3, 4
`wcon`	*wages in the construction industry*	-	3
`wtuc`	*wages in the transportation, utilities, and communication industries*	-	3
`wtrd`	*wages in the construction industry*	-	3
`wfir`	*wages in the finance, insurance, real estate industries*	-	3
`wser`	*wages in the service industry*	-	3
`wmfg`	*wages in the manufacturing industry*	-	3
`wfed`	*wages among federal employees*	-	3
`wsta`	*wages among state employees*	-	3
`wloc`	*wages among local government employees*	-	3
`mix`	*mix of offenses; face-to-face v others*	-	3
`pctymle`	*percent young male*	log($pctymle$)	2, 3

Research Question and Model-Building

Our **research question** is the following: **Should our candidate support a traditional "Tough on Crime" platform?**

We face a key limitation: our data does not give us visibility into crime, it only gives us insight into the official *crime rate*. The crime rate is a function not only of crimes committed but also of various 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. As we report our findings we will make note of potential bias that results from our inability to observe and analyze critical variables.

In order to answer our research question we created a model which included variables related to key crime policy decisions. We only included variables which would allow our canadidate to make concrete policy proposals that lie within her purview.

Using the aforementioned criteria we choose the variables police per capita (`polpc`), probability of arrest (`prbarr`), probability of conviction (`prbconv`), probability of incarceration (`prbpris`), and average sentence length (`avgsen`) for our first model. Understanding how these items relate to crime rate can help shape her position, and understand whether a "Tough on Crime" stance does in fact acheive a reduction in crime.

Specifically, we can help her understand which departement/function should receive additional funding given a limited budget (e.g. if conviction rates are highly correlated perhaps we invest more in our District Attorneys).

Table 2: Model 1: Hypothesized Key Determinants of Observed Crime Rate

Variable Name	Description	Transformation Applied
`prbarr`	*'probability' of arrest*	-
`prbconv`	*'probability' of conviction*	-
`prbpris`	*'probability' of prison sentence*	square ($prbpris^2$)
`avgsen`	*average sentence, in days*	-
`polpc`	*police per capita*	log(*polpc*)

Transformation Analysis

–Evaluate the scatter of each of the 5 variables to see if a transform is necessary. –Final table with new transforms.
Police should be log transformed. probability of prison should be squared

```
#Linear Regression model using only our key variables of interest, transformed as needed
model1_trans <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) +
                                    prbconv + prbpris^2 + 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) %>% select(-std.error),
      caption = 'Model 1: Hypothesized Key Determinants of Observed Crime Rate')
```

Table 3: Model 1: Hypothesized Key Determinants of Observed Crime Rate

term	estimate	statistic	p.value
(Intercept)	0.1601083	5.3360888	0.0000008
log(polpc)	0.0224302	5.6021152	0.0000003
log(prbarr)	-0.0247760	-7.2938710	0.0000000
prbconv	-0.0250464	-6.3898742	0.0000000
prbpris	0.0105302	0.6296984	0.5306010
avgsen	-0.0004725	-0.8768939	0.3830452

Comments on Model 1:

The log of police per capita, the log of the probability of arrest, and the probability of conviction all have high levels of significance. The overall model has an adjusted R^2 of .5232. There are three main points to highlight:

1) Our coefficient for police per capita (`polpc`) is positive, large, and highly significant. If we inaccurately assumed this model was causal the best mechanism to reduce crime rates would be to sunset the police force! This model is not causal, however, and a more plausible interpretation is that a higher number of police per capita is a response to higher levels of criminal activity.

2) Both probability of arrest (`prbarr`) and 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 square of the probability of incarceration (`prbpris`) nor average sentence length (`avgsen`) were statistically significant. Furthermore, the coefficient for the squared probability of incarceration rate is positive, which is counterintuitive.

Model 2

While our first model showed promise, there are several other factors which might be correlated with these explanatory variables; this would lead to multicollinearity issues. When multiple independent variables exhibit collinearity it becomes increasingly difficult to untangle their individual effects on the crime rate. This is troublesome because the model coefficients may misrepresent the impact of a variable, leading to a policy which has underwhelming impact. In order to control for this we created a second model which includes variables we believe to be highly correlated with our three key variables.

Collinearity with Police Per Capita: We would expect policing practices in urban areas to differ substantially from those in suburban or rural areas; including density and the urban dummy variable can help control for this. Additionally, the police force is funded by taxpayers, and therefore we might expect a larger police force per capita to be correlated with higher tax revenues. *Given the assumed diminishing marginal returns of money* **[TS: <- Could you unpack that a bit more, Clay?]** it makes sense to include the log transformation so that we can interpret our coefficient as the change given a 1% increase.

Creation of variables to proxy tax revenue

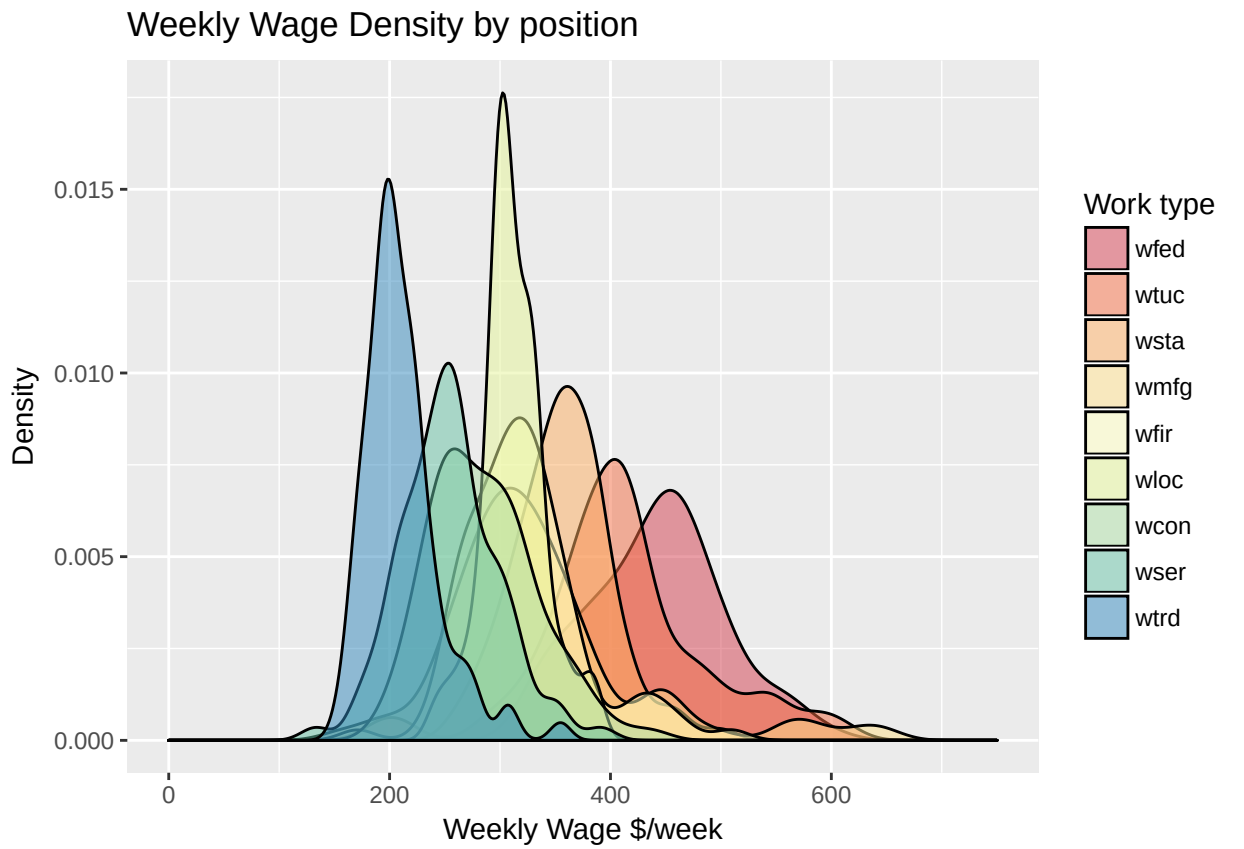
Our dataset contains 9 wage variables, each representing a different sector or group of industries. We do not have *a priori* justification to believe *a single industry* **[TS: <- This feels a bit awk; to wordsmith...]** might contribute disproportionately to crime, but we can assume that low wages 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, **[TS: <- Poss. to add a phrase explaining why?]** 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 **[TS: <- Okay edit? Service industry not really physical labor in same way as manufacturing and construction...]** 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) **[TS: <- I'd advocate for grouping 'wtuc' with blue-collar; maybe retail as well. It's likely generate incomes closer to the blue-collar range than the professional wage. That would leave wfir alone, so no need for a composite var...]**

```
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 by position') +
  xlab('Weekly Wage $/week') +
```

```
  ylab('Density') +
  labs(fill = 'Work type') +
  xlim(0, 750)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_density).
```




```
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
```

Collinearity with Probability of Arrest: Sociological research suggests that men - especially young, minority men - are at an increased risk of arrest. Therefore we will include both the `pctymle` (percentage of young male, under a log transformation) and `pctmin80` (percentage of minorities in 1980) variables.

Collinearity with Probability of Conviction: The likelihood of an arrest leading to a conviction depends not only on the culpability of the suspect, but also on the quality or effectiveness of the police department's investigative team, the district attorney, court-appointed advocates, judges, and other government officials. Quality is an unobservable (omitted) variable, but it might be loosely correlated with the government wage. We will thus include our government wage variable as a proxy.

Collinearity with Probability of Incarceration and Average Sentencing: While there may be unobservable covariates for these two variables, we cannot identify any reasonable proxies to include in a revised model to address the issue.

Table 4: Model 2: Hypothesized Key Determinants of Observed Crime Rate with Additional Covariates

Variable Name	Description	Transformation Applied
prbarr	*'probability' of arrest*	-
prbconv	*'probability' of conviction*	-
prbpris	*'probability' of prison sentence*	square ($prbpris^2$)
polpc	*police per capita*	log($polpc$)
avgsen	*average sentence, in days*	-
taxpc	*tax revenue per capita*	log($taxpc$)
density	*persons per square mile*	$\sqrt{density}$
pctymle	*percentage of young males*	log($pctymle$)

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

model2_trans_no_minvar <- with(crime_na, lm(crmrte ~ polpc + prbarr +
                                            prbconv + prbpris + avgsen + log(taxpc) +
                                            sqrt(density) + govt_wg + log(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)	0.1466529	0.0395261	3.7102823	0.0003835
log(polpc)	0.0171836	0.0035645	4.8207274	0.0000068
prbarr	-0.0421132	0.0086252	-4.8825660	0.0000054
prbconv	-0.0153394	0.0029234	-5.2471302	0.0000013
prbpris	-0.0034123	0.0118676	-0.2875321	0.7744579
avgsen	-0.0001293	0.0003816	-0.3387918	0.7356655
log(taxpc)	0.0046694	0.0042822	1.0904010	0.2788513
sqrt(density)	0.0186672	0.0027664	6.7478855	0.0000000
govt_wg	-0.0000276	0.0000400	-0.6892903	0.4926602
log(pctymle)	0.0059100	0.0052787	1.1195953	0.2662787
pctmin80	0.0003594	0.0000585	6.1442897	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 0.786. When excluding it, the adjusted R^2 is 0.689, a difference of 0.097

2) The square root of the population density is also highly significant. However, it is unclear what the interpretation may be. Absent the square root transformation, the coefficient is ~.0055, indicating that an additional person per square mile increases the crime rate by .55%. **TS: Is that right? I'm not so sure about it...**

Model 3

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) + prbarr + prbconv + prbpris^2 +
                            avgsen + taxpc + sqrt(density) + govt_wg + log(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)	0.1527653	0.0359997	4.2435120	0.0000638
log(polpc)	0.0163089	0.0036955	4.4132032	0.0000346
prbarr	-0.0347328	0.0086535	-4.0137409	0.0001432
prbconv	-0.0141941	0.0034683	-4.0925614	0.0001088
prbpris	0.0027770	0.0117160	0.2370283	0.8132987
avgsen	-0.0005200	0.0004105	-1.2665774	0.2093329
taxpc	0.0001843	0.0000982	1.8780075	0.0643742
sqrt(density)	0.0176515	0.0039232	4.4992318	0.0000252
govt_wg	-0.0000160	0.0000433	-0.3703152	0.7122199
log(pctymle)	0.0063302	0.0053595	1.1811138	0.2413913
pctmin80	0.0002966	0.0000913	3.2472176	0.0017617
west	-0.0041638	0.0038415	-1.0839028	0.2819749
central	-0.0043397	0.0027540	-1.5757408	0.1194097
urban	0.0023929	0.0054640	0.4379385	0.6627240
physical_wg	-0.0000128	0.0000163	-0.7829644	0.4361808
industry_wg	0.0000149	0.0000304	0.4895079	0.6259499
mix	-0.0192487	0.0146214	-1.3164731	0.1921337

Model 3 Notes:

Model Comparison

Below is a comparison table for our three models, along with key statistics related to the model. Our findings match with what we would hope to see from a model building perspective: Our AIC and adjusted R squared numbers suggest that the model 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 squared), and our model which includes everything–despite having the highest unadjusted R squared–performed worse as we arbitrarily added additional variables.

```
# Code from here: https://stackoverflow.com/questions/47494761/show-akaike-criteria-in-stargazer (using
stargazer(model1, model2, model3,
```

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

Table 7: Linear Models Predicting Crime Rate

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
polpc	10.587	6.549	6.824
prbarr	−0.090	−0.055	−0.049
prbconv	−0.027	−0.019	−0.020
prbpris	0.012	0.001	0.004
avgsen	−0.0003	−0.0002	−0.0004
taxpc		0.0002	0.0002
density		0.005	0.006
govt_wg		0.00002	−0.00000
pctymle		0.093	0.078
pctmin80		0.0003	0.0003
west			−0.004
central			−0.004
urban			−0.002
physical_wg			−0.00000
industry_wg			0.00002
mix			−0.022
Constant	0.055	0.012	0.021
Observations	90	90	90
R^2	0.550	0.824	0.839
Adjusted R^2	0.523	0.802	0.804
Akaike Inf. Crit.	−517.931	−592.541	−588.550

Omitted Variables

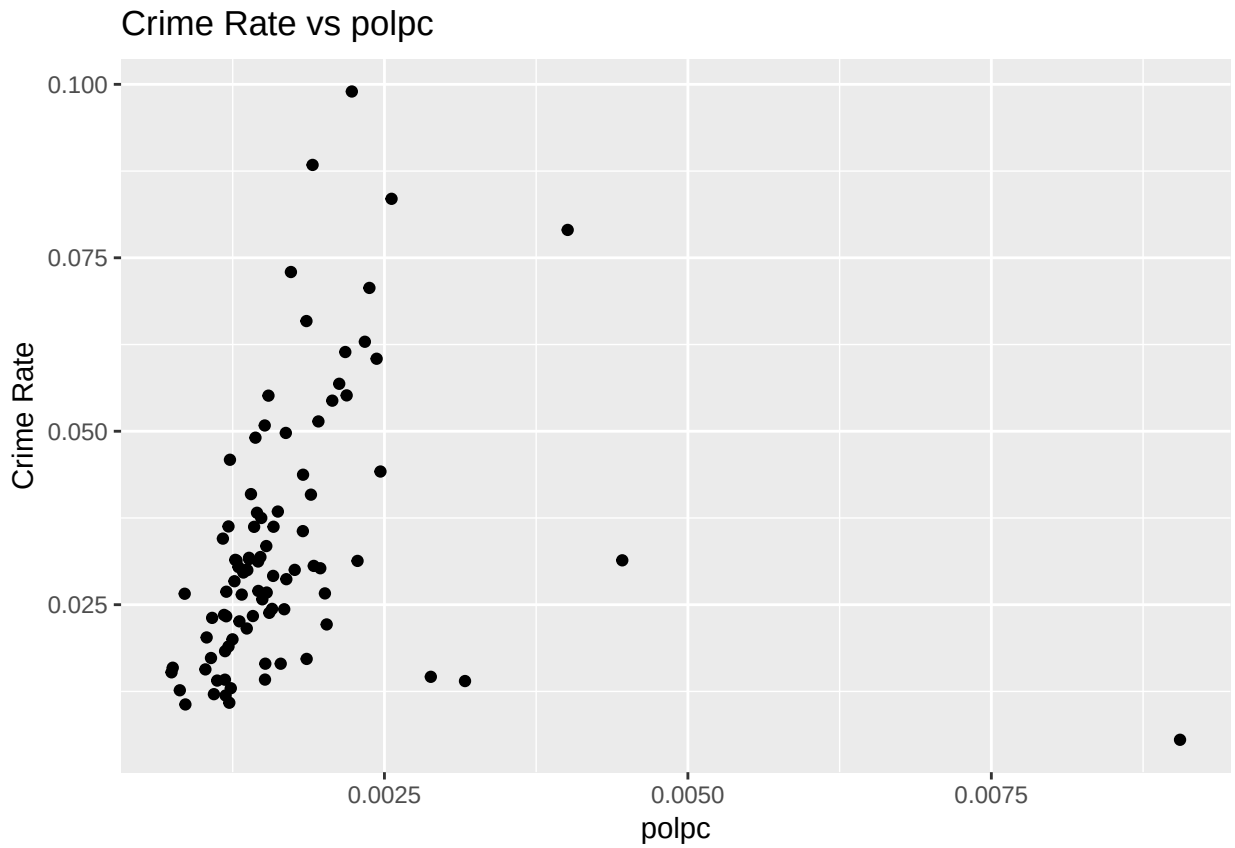
Despite the promising results from our three models it is difficult to ascribe causality to the variables of interest. One issue with causal inference in general is ommitted variable bias, which can invalidate our ability to assume each explanatory variable is uncorrelated with the error term. While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

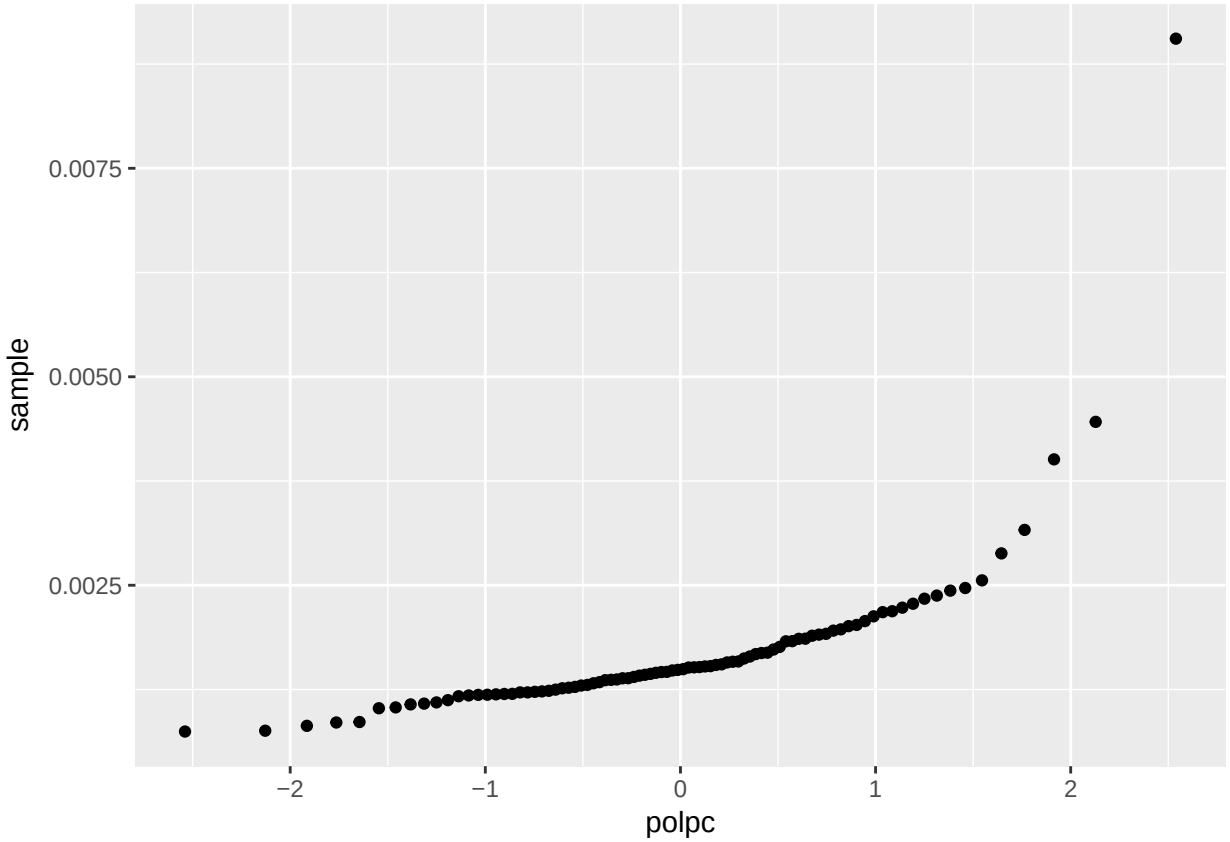
1) Political Party in Control: Traditionally the issue of crime policy is a highly partisan issue, with each party having very different approaches to crime reduction. All else being equal, we might expect police levels and average sentence lengths to be correlated with the party that is crafting legislation. 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:
5) Police Representation:
6)

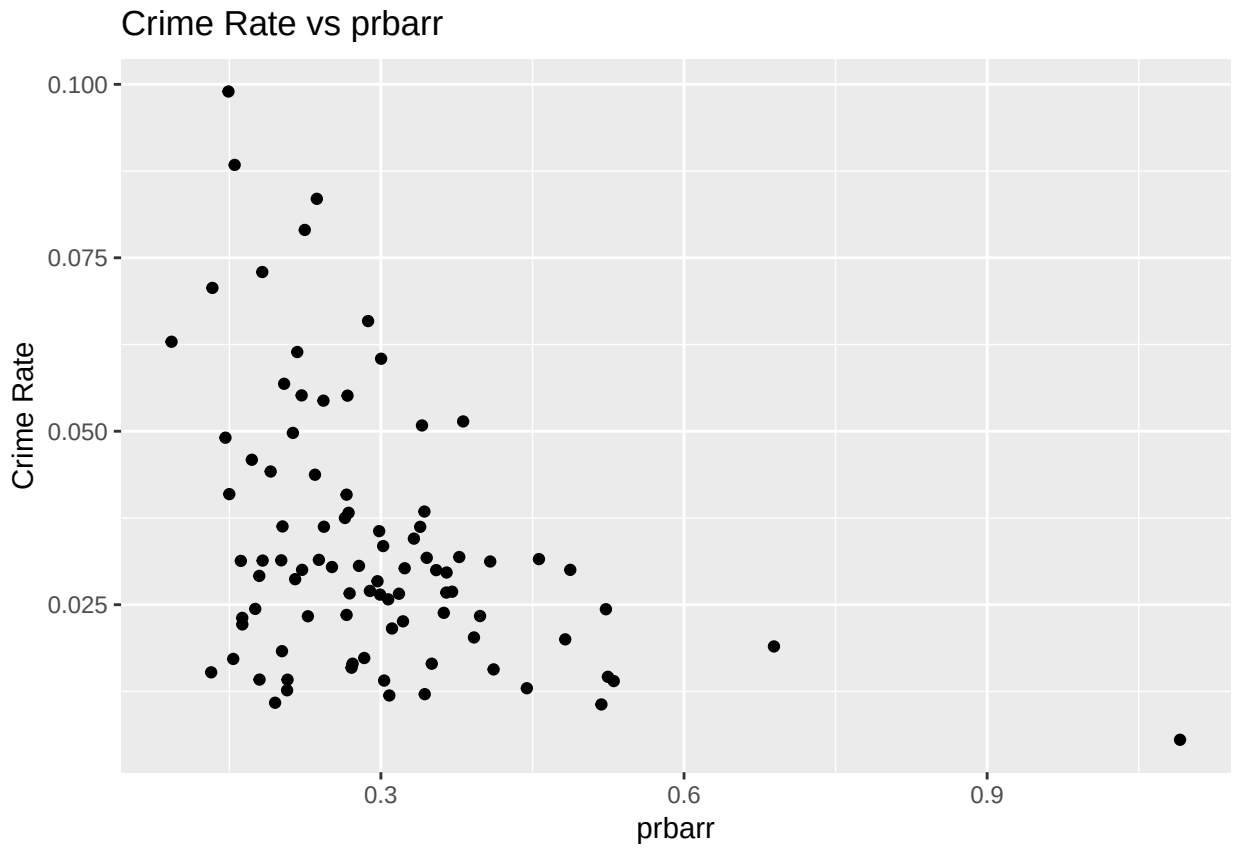
Findings and Policy Recommendations

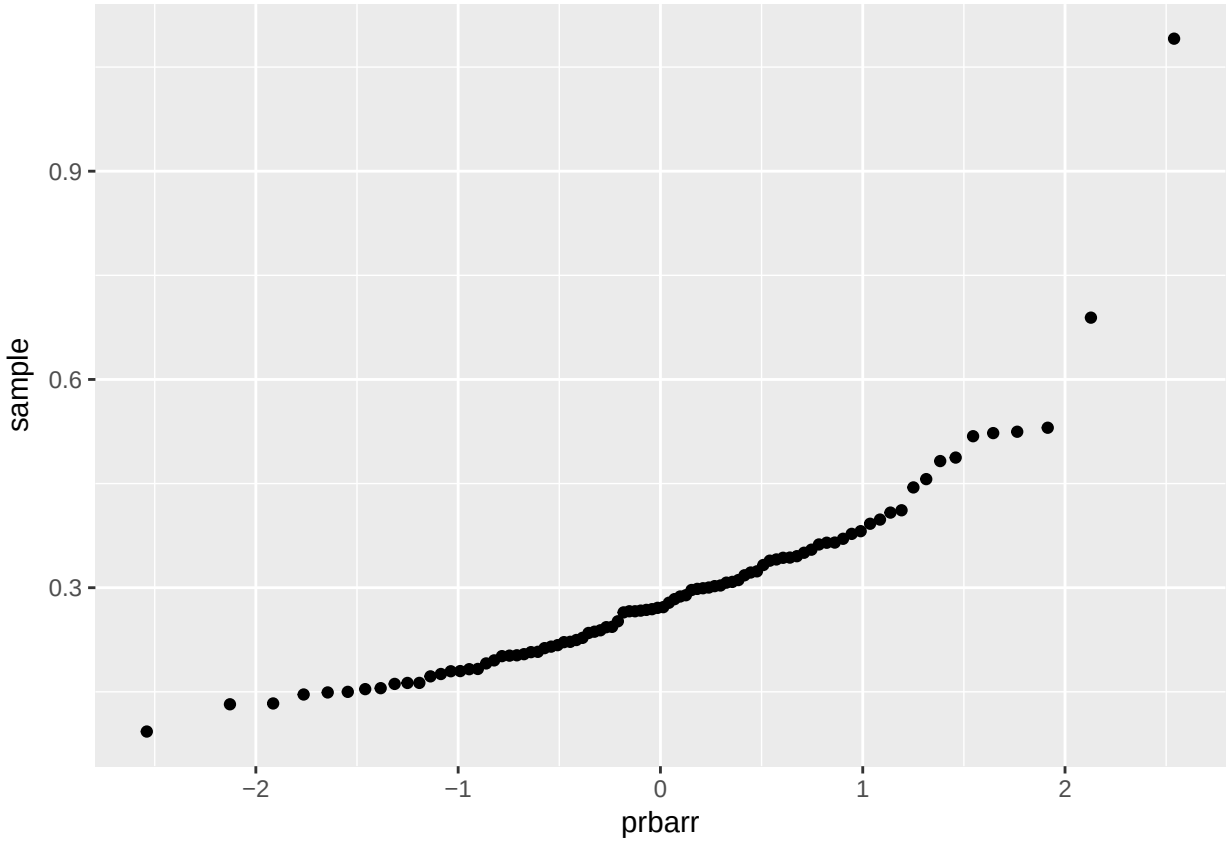
Conclusion

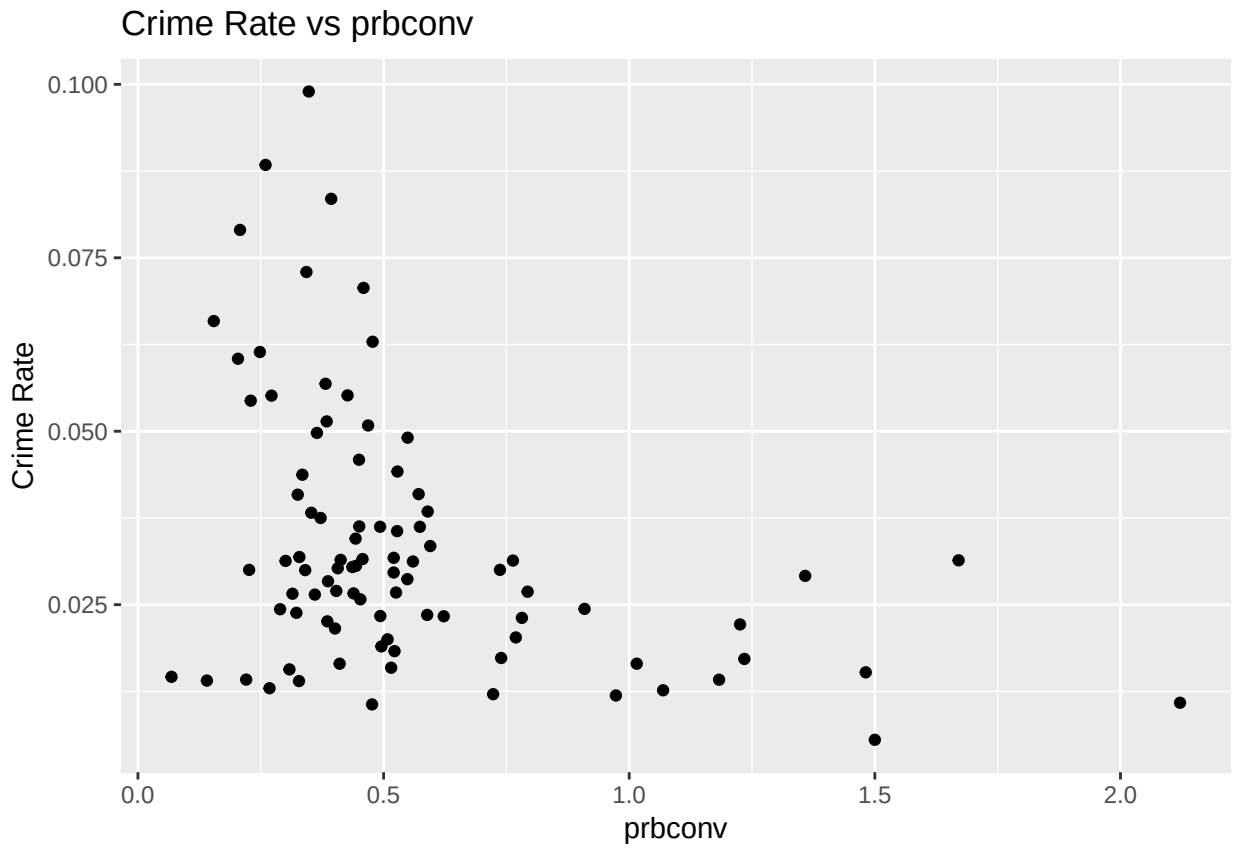
```
mod1vars <- c("polpc","prbarr","prbconv","prbpris","avgsen")
mod2vars <- c("taxpc","density","govt_wg","pctymle","pctmin80")
mod3vars <- c("")
for (v in mod1vars){
  print(ggplot(crime_na, aes(y = crmrte)) +
          geom_point(aes_string(x = v)) +
          xlab(v) +
          ylab('Crime Rate') +
          ggtitle(str_glue('Crime Rate vs {v}'))
          )
  print(ggplot(crime_na) +
          geom_qq(aes_string(sample = v)) +
            xlab(v))
}
```

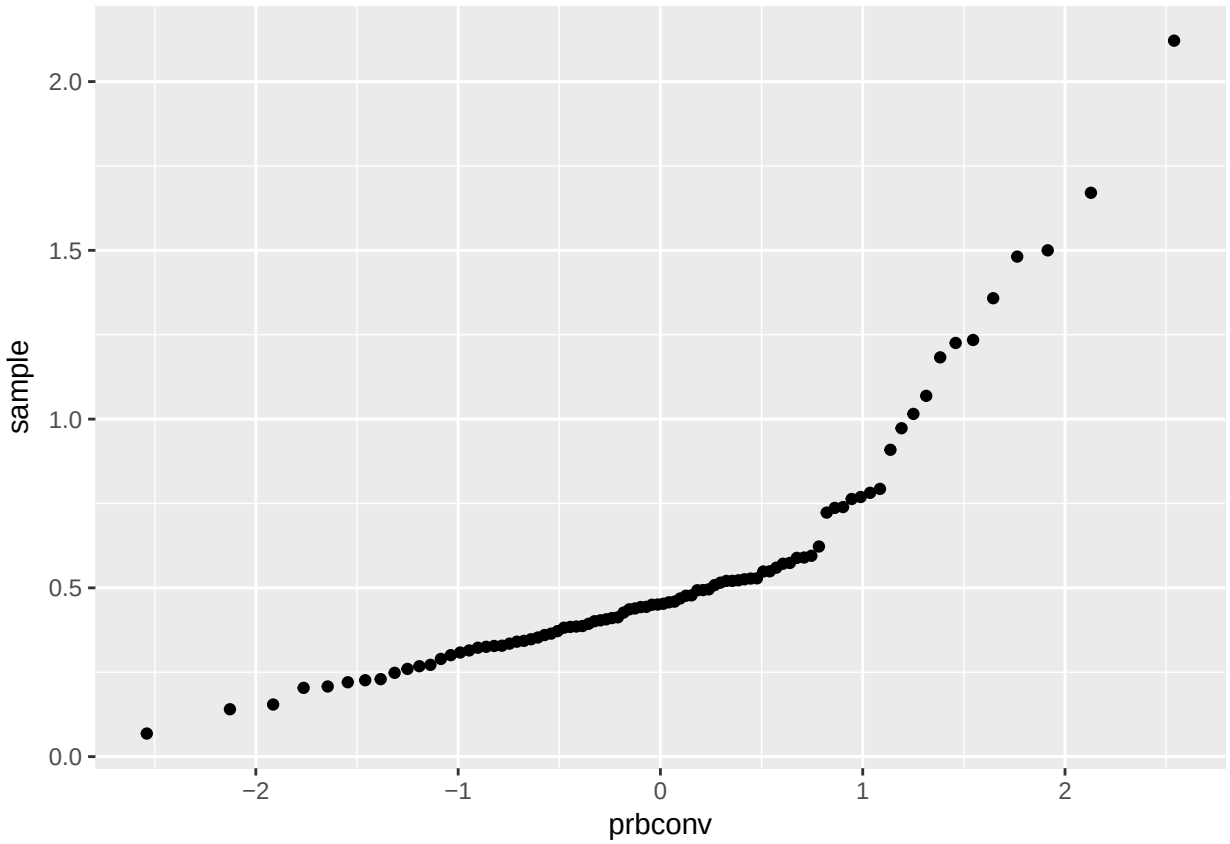

Crime Rate vs polpc
0.100
0.075
0.050
0.025
Crime Rate
0.0025
0.0050
0.0075
polpc

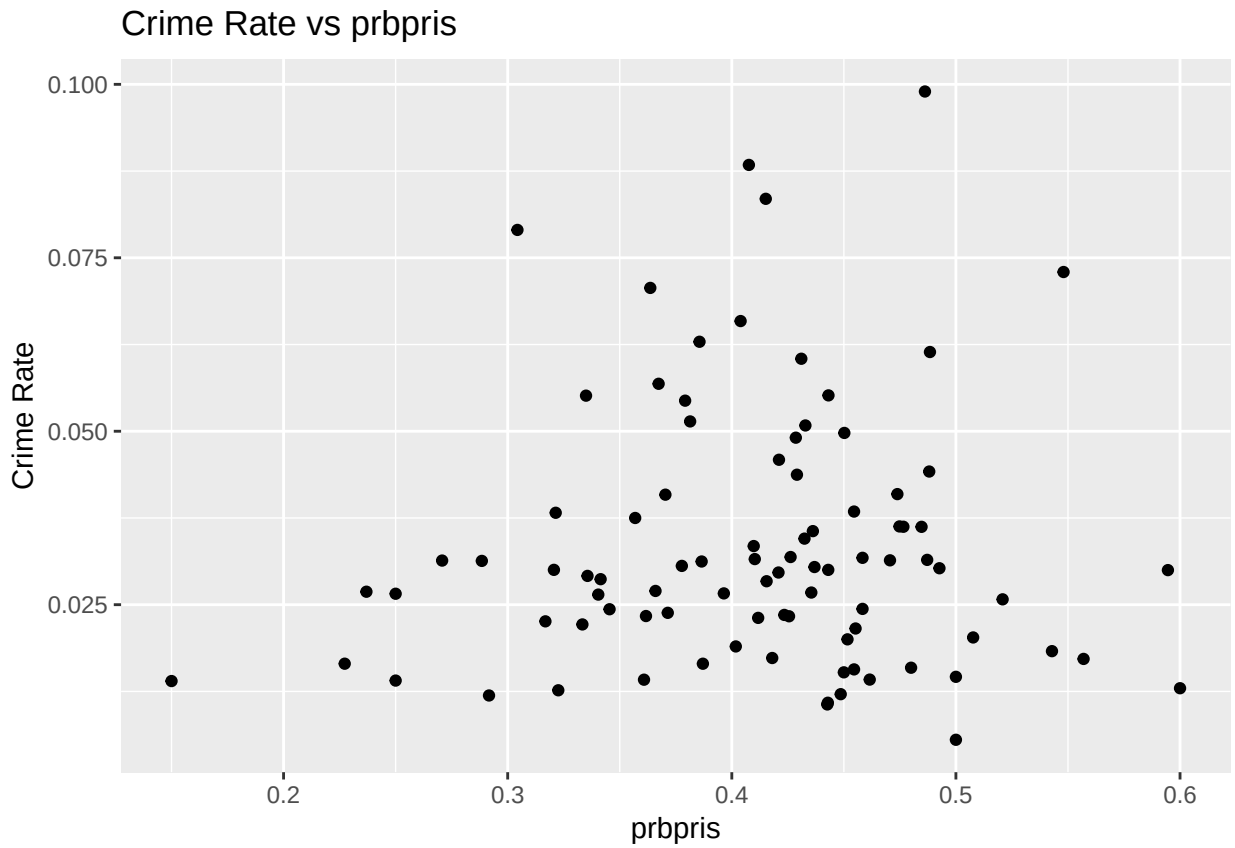

0.0075
0.0050
0.0025
sample
−2
−1
0
1
2
polpc

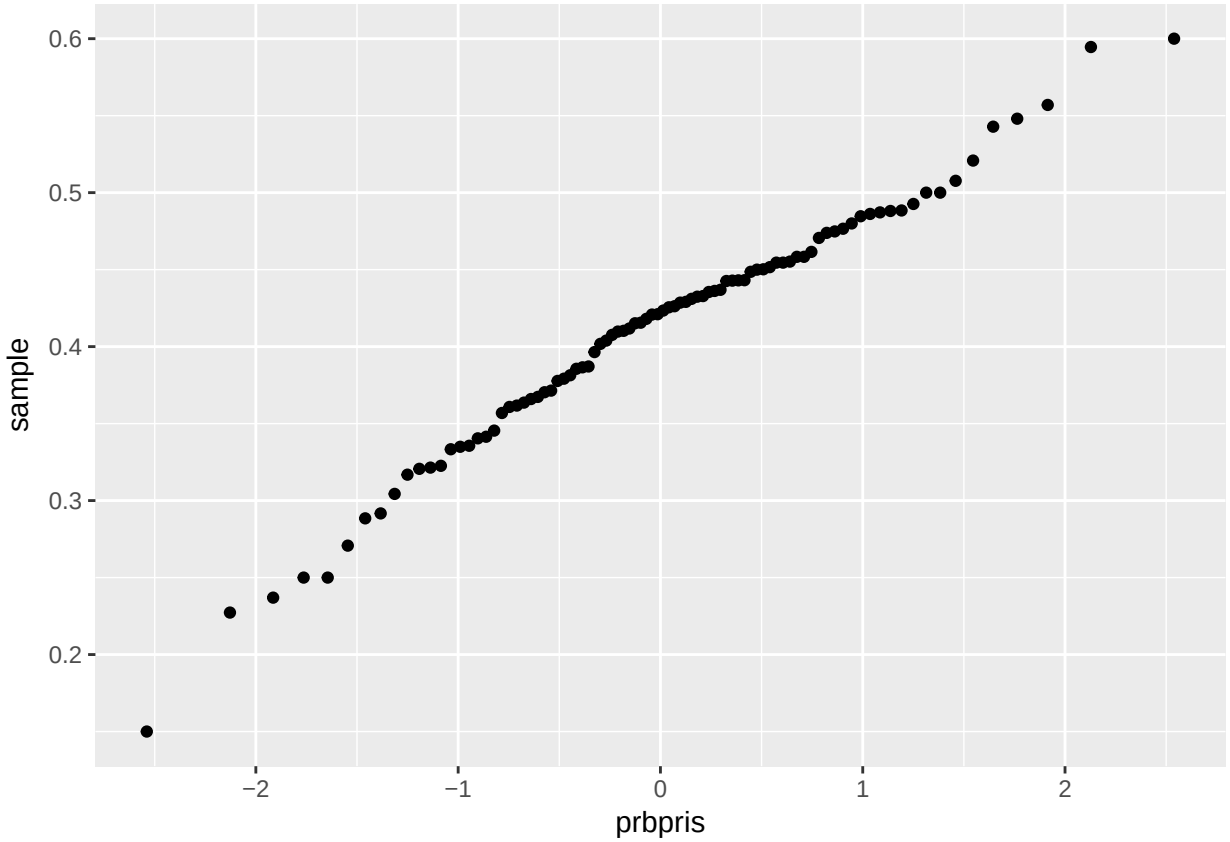

Crime Rate vs prbarr
0.100
0.075
0.050
0.025
Crime Rate
0.3
0.6
0.9
prbarr

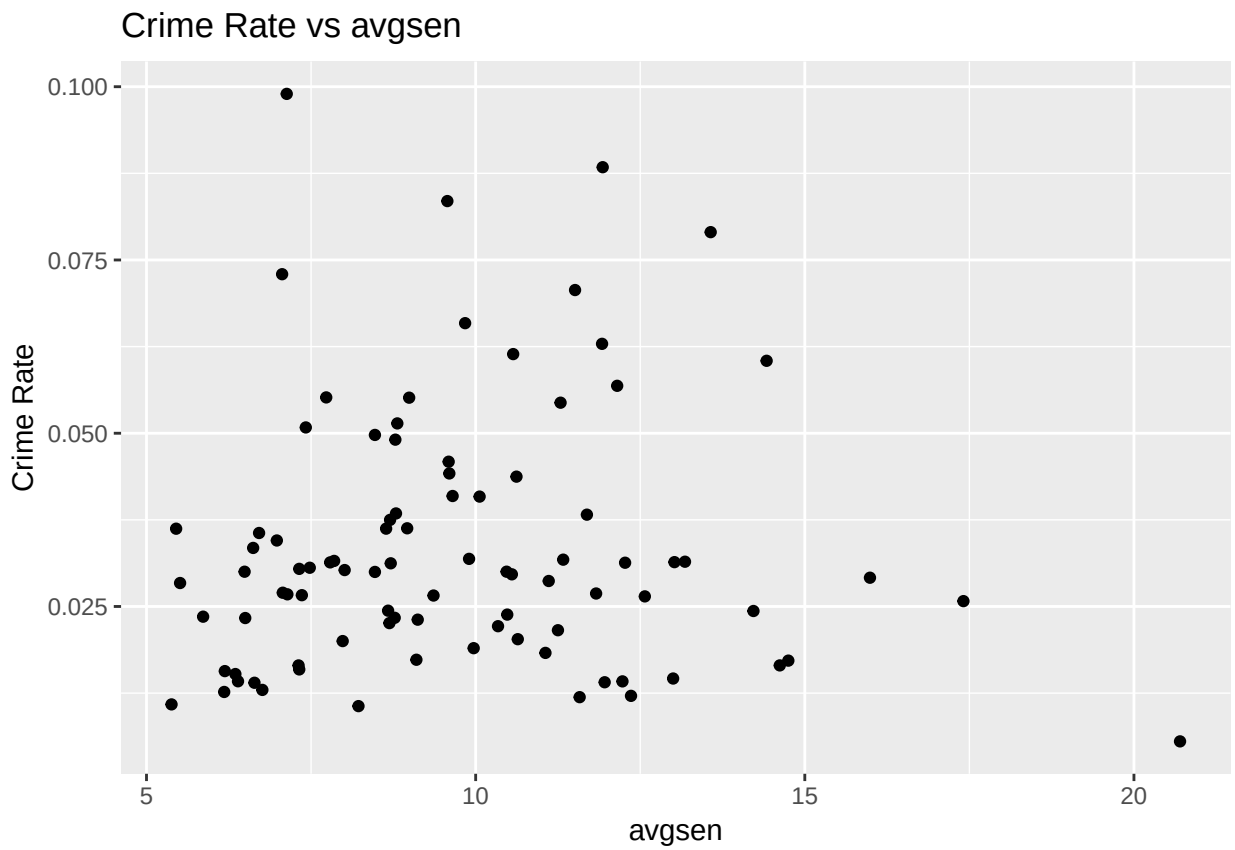

sample
0.9
0.6
0.3
−2
−1
0
1
2
prbarr

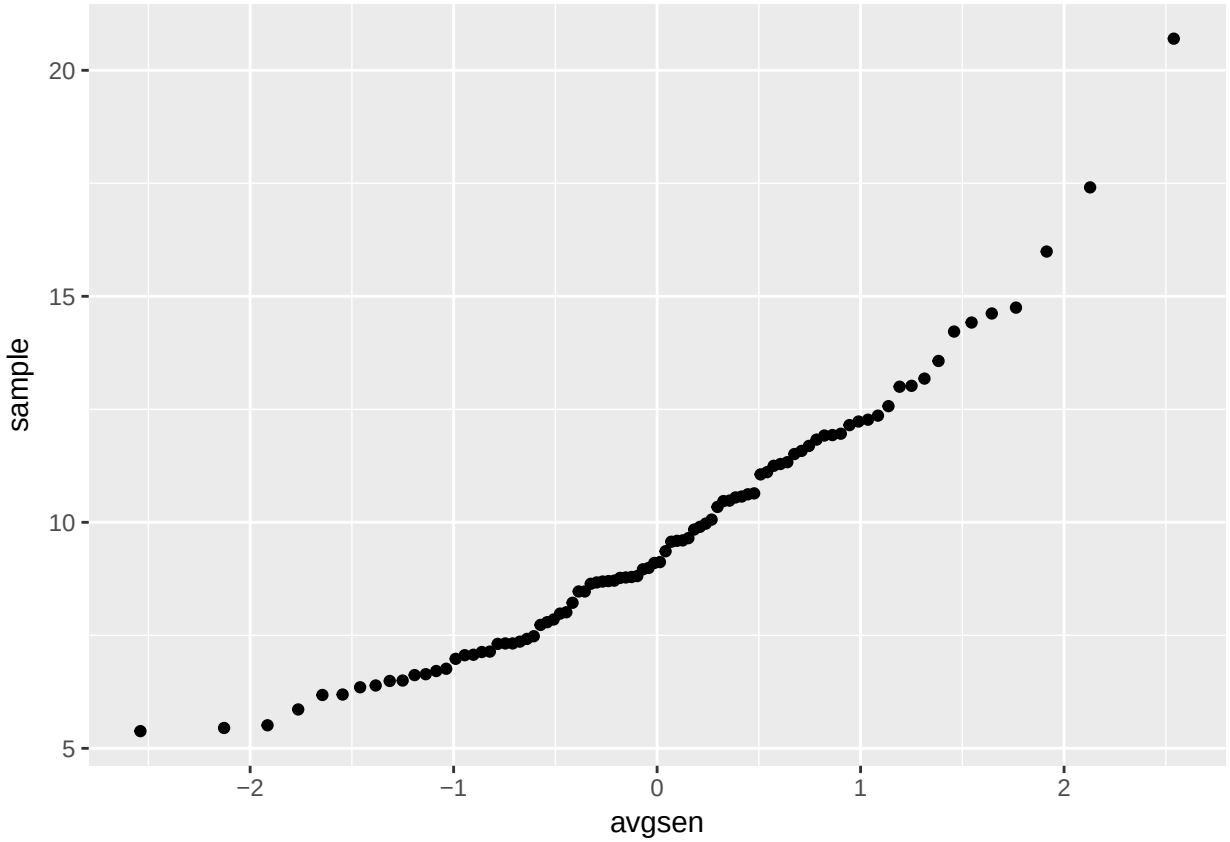

Crime Rate vs prbconv
Crime Rate
0.100
0.075
0.050
0.025
0.0
0.5
1.0
1.5
2.0
prbconv


2.0
1.5
1.0
0.5
0.0
sample
−2
−1
0
1
2
prbconv

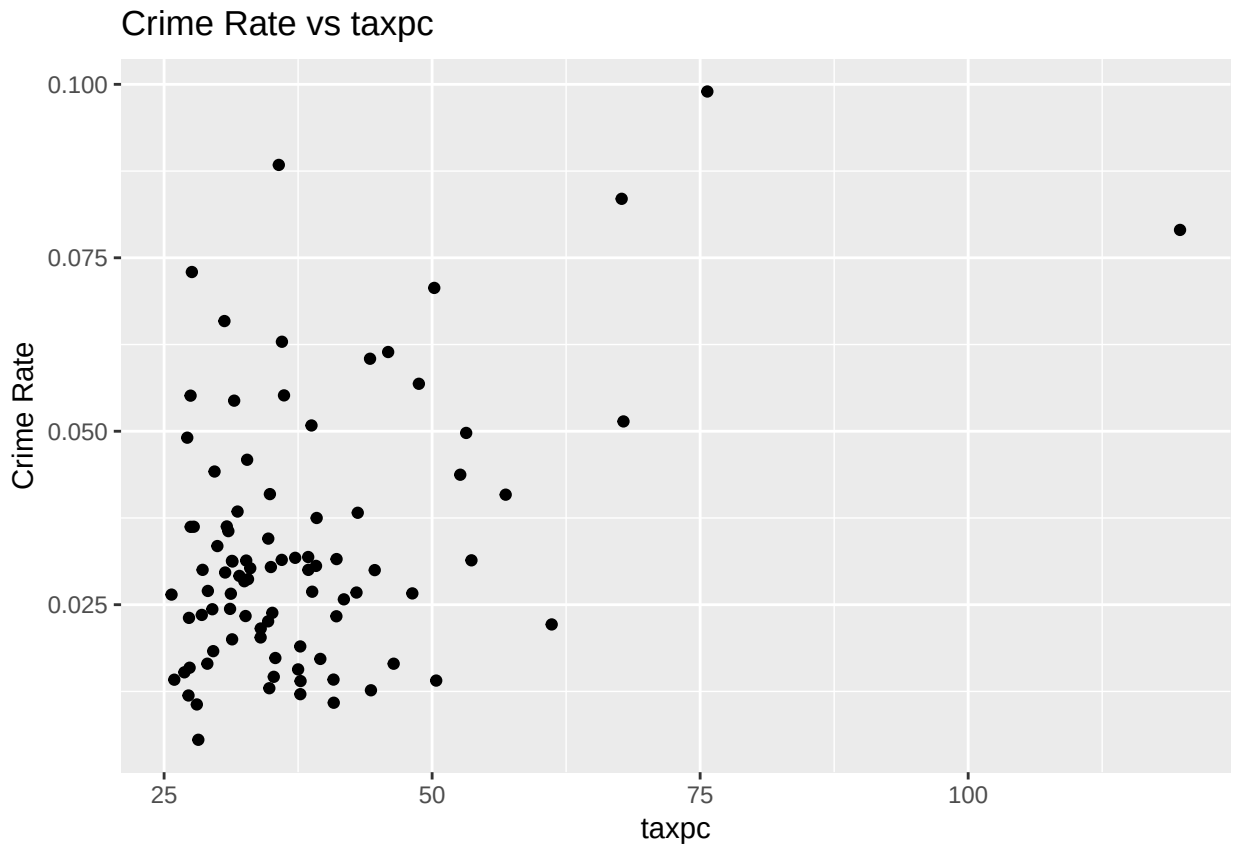

Crime Rate vs prbpris
Crime Rate
0.100
0.075
0.050
0.025
0.2
0.3
0.4
0.5
0.6
prbpris

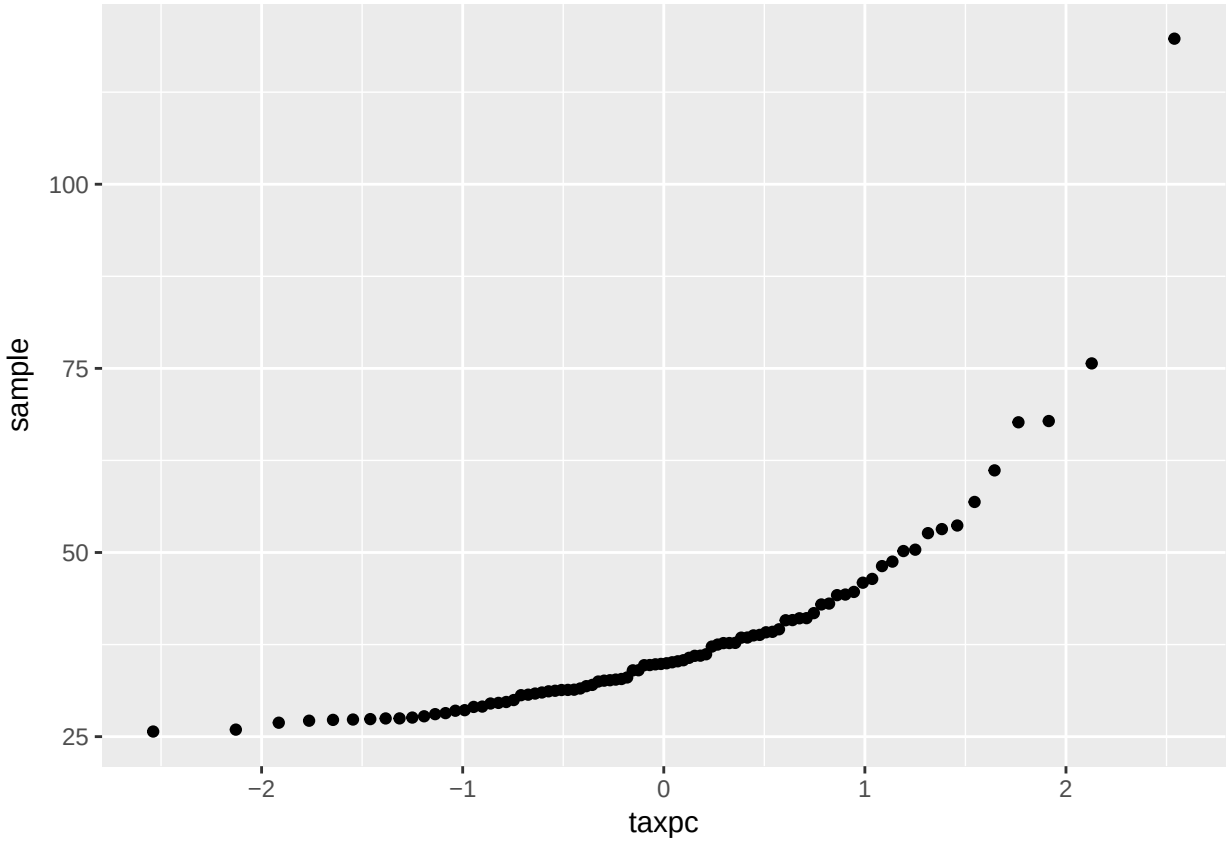

0.6
0.5
0.4
0.3
0.2
sample
−2
−1
0
1
2
prbpris

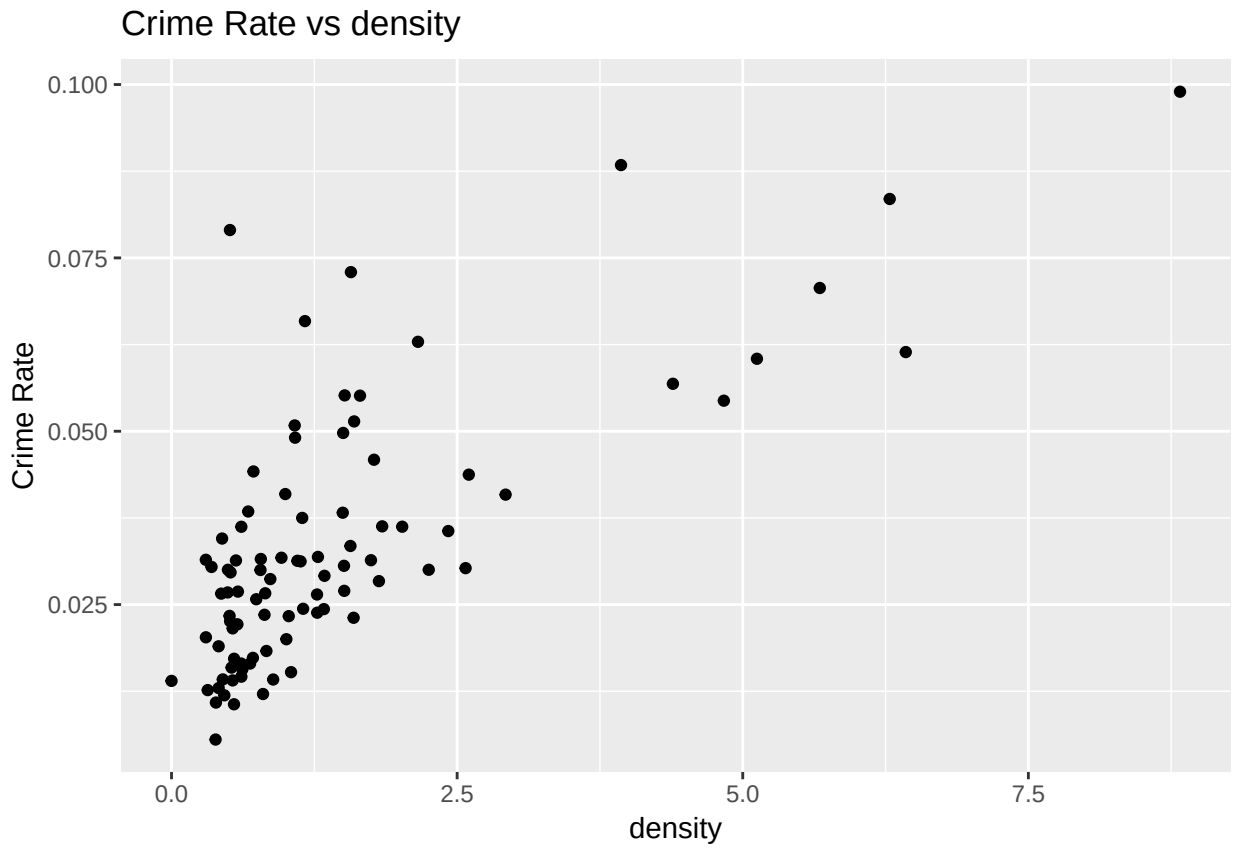

Crime Rate vs avgsen
Crime Rate
0.100
0.075
0.050
0.025
5
10
15
20
avgsen

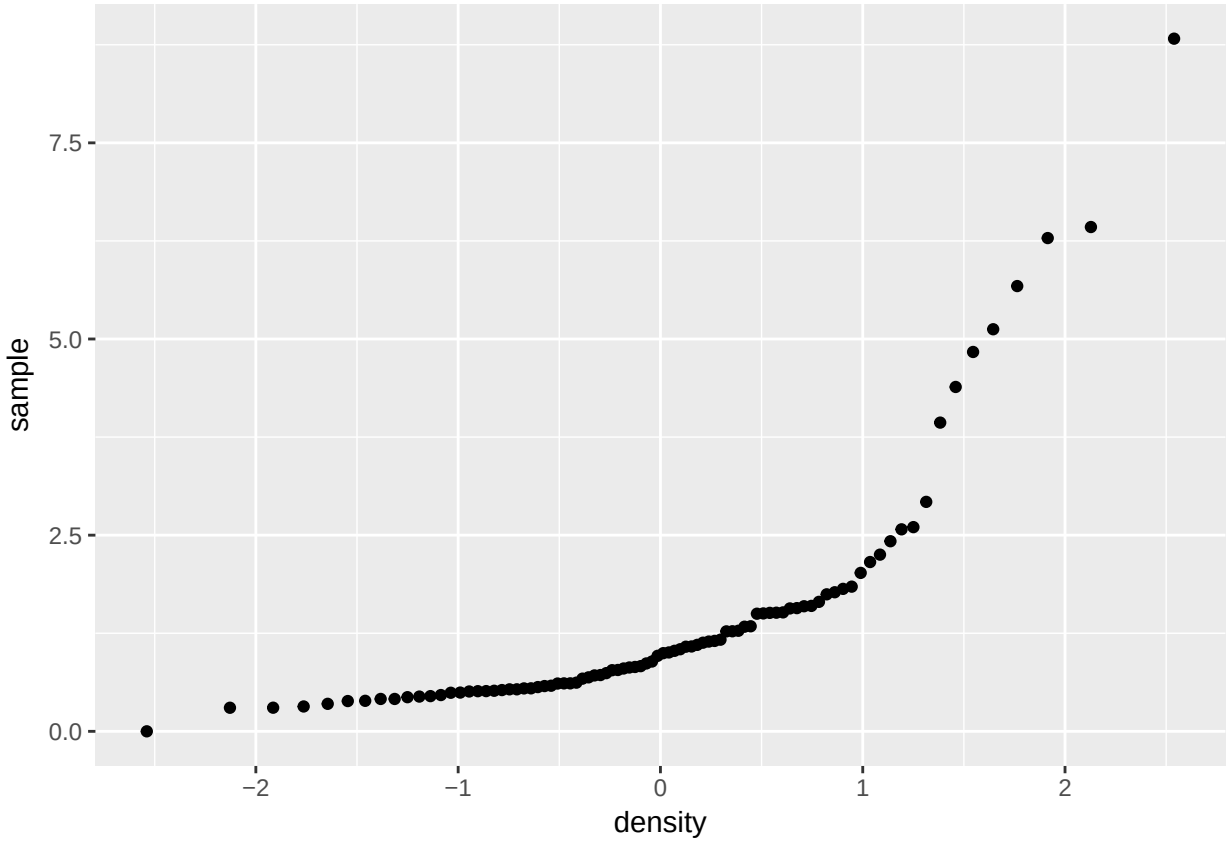


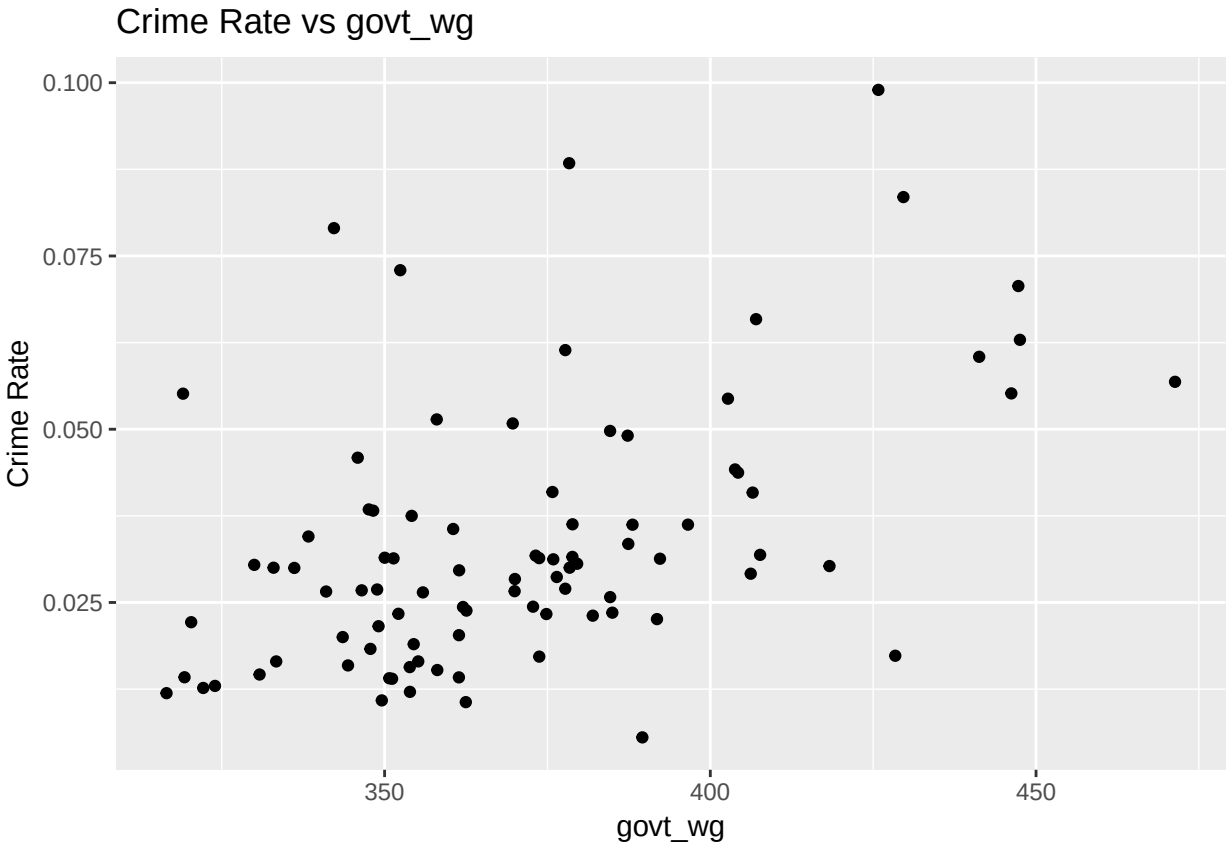

```
for (v in mod2vars){
  print(ggplot(crime_na, aes(y = crmrte)) +
          geom_point(aes_string(x = v)) +
          xlab(v) +
          ylab('Crime Rate') +
          ggtitle(str_glue('Crime Rate vs {v}'))
          )
  print(ggplot(crime_na) +
          geom_qq(aes_string(sample = v)) +
            xlab(v))
}
```

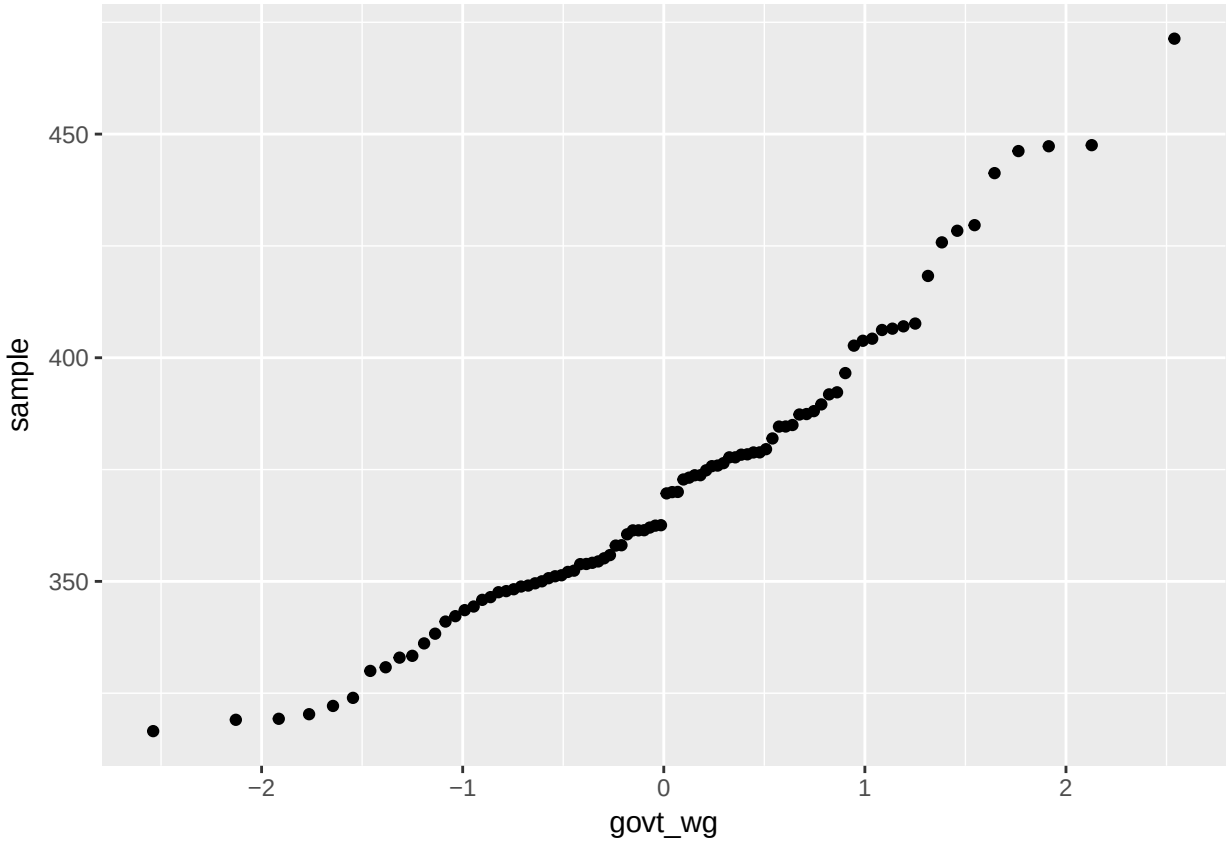

Crime Rate vs taxpc
Crime Rate
0.100
0.075
0.050
0.025
25
50
75
100
taxpc

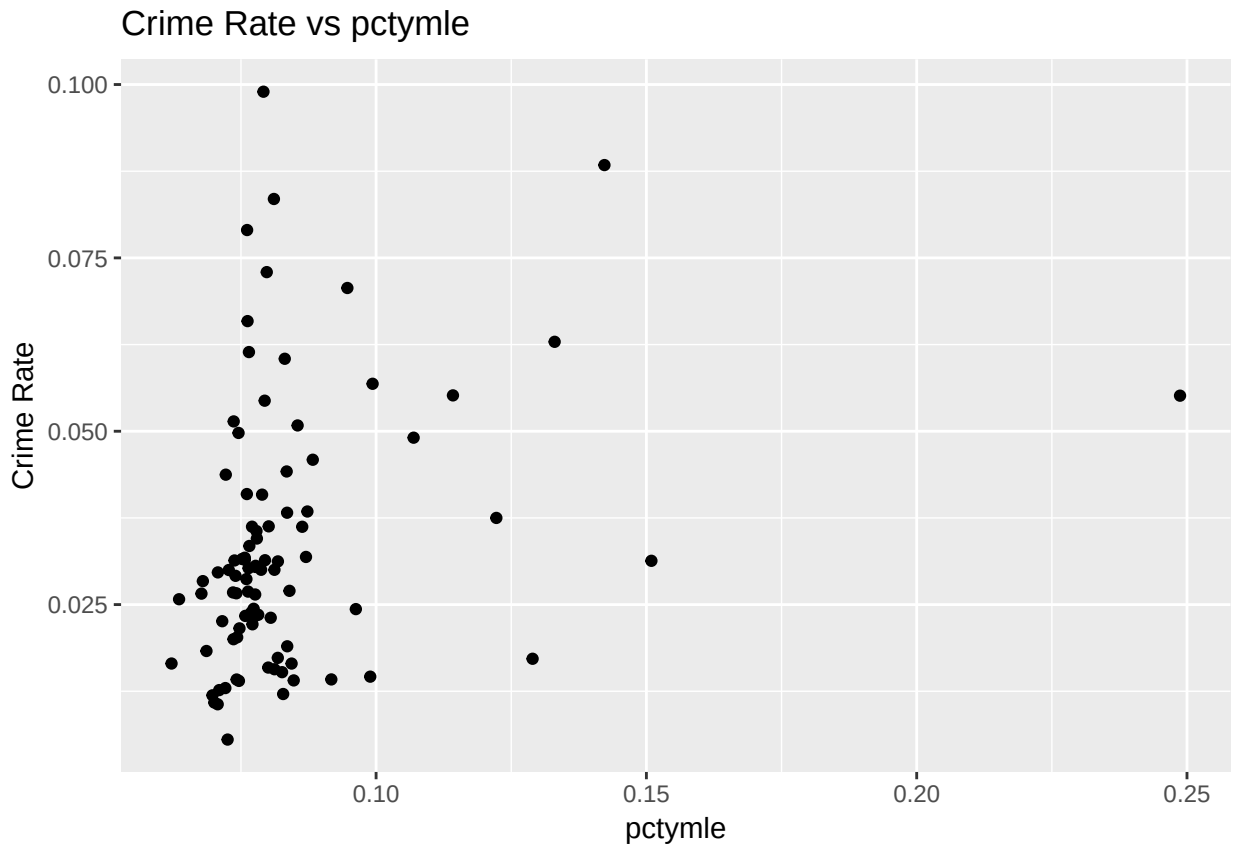

sample
100
75
50
25
−2
−1
0
1
2
taxpc

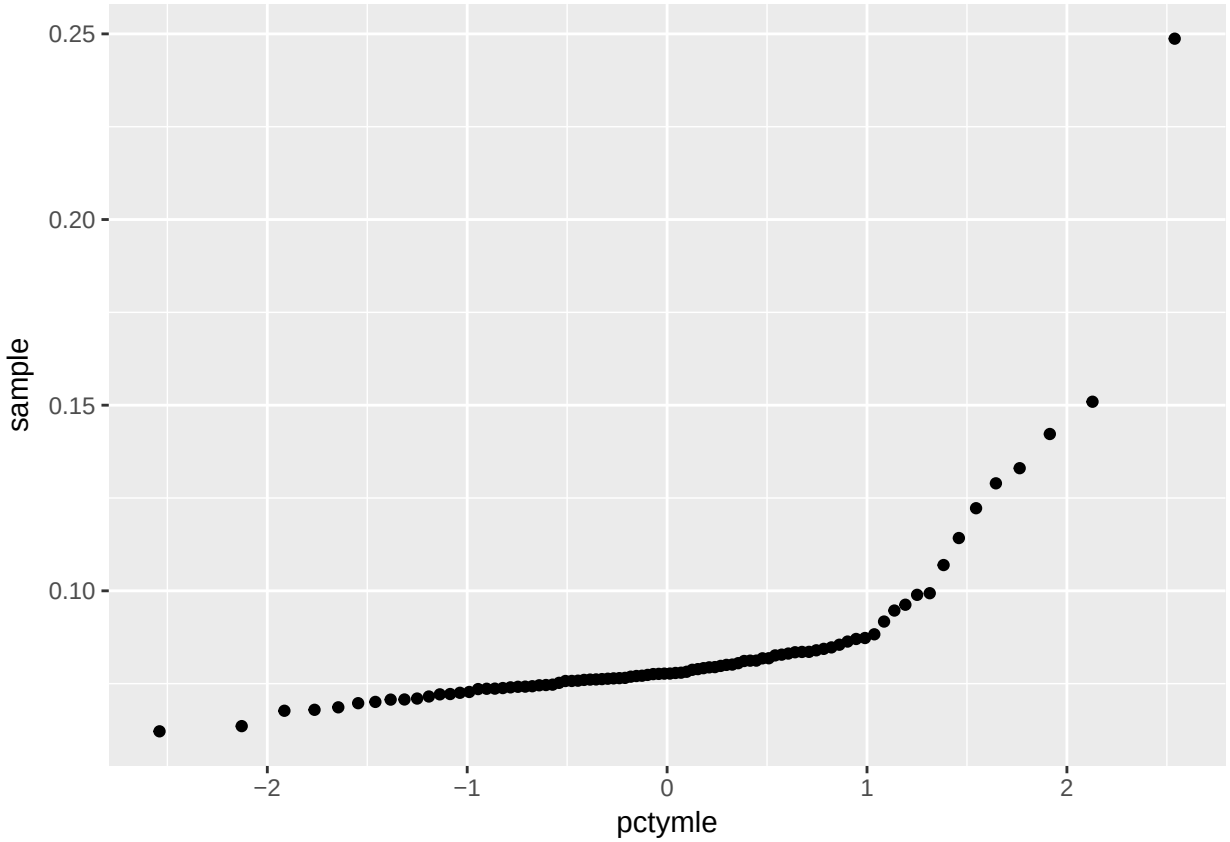

Crime Rate vs density
Crime Rate
0.100
0.075
0.050
0.025
0.0
2.5
5.0
7.5
density

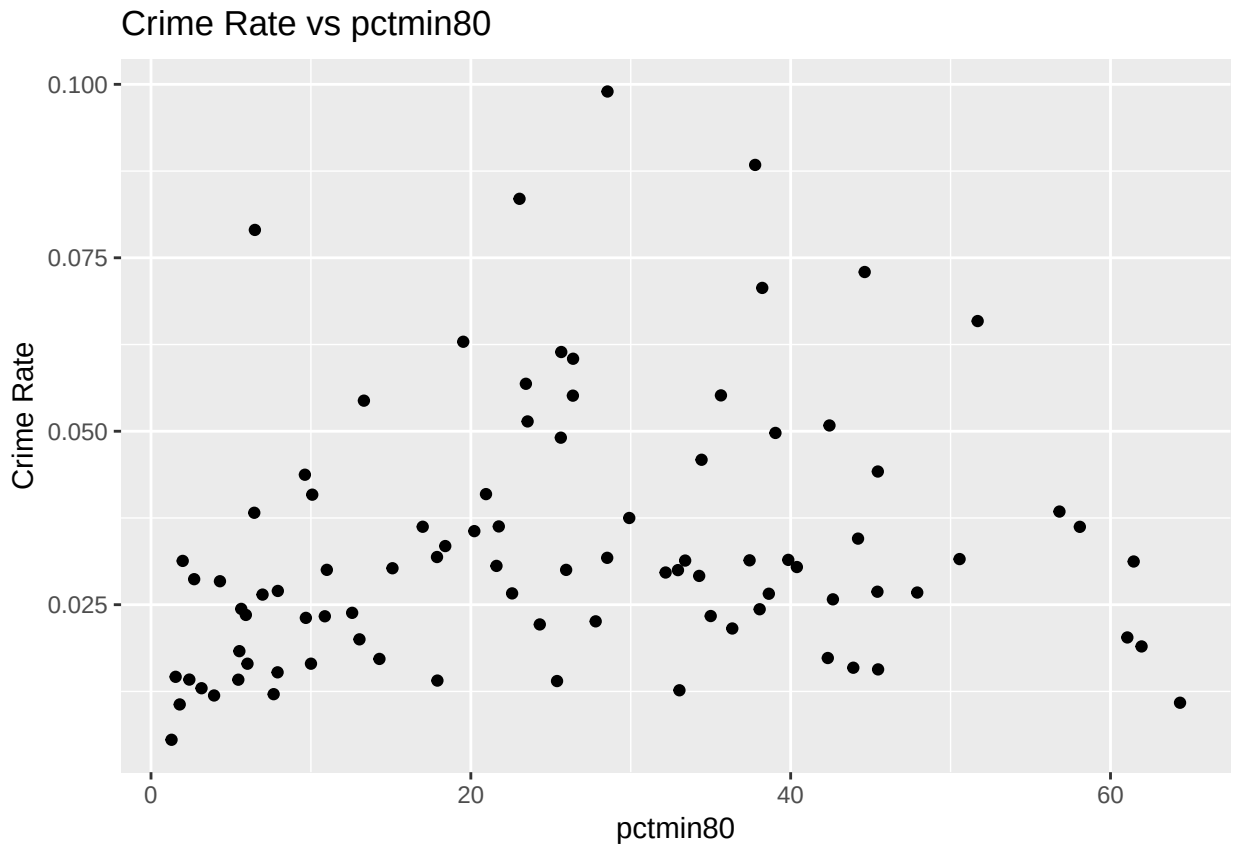

7.5
5.0
2.5
0.0
sample
−2
−1
0
1
2
density

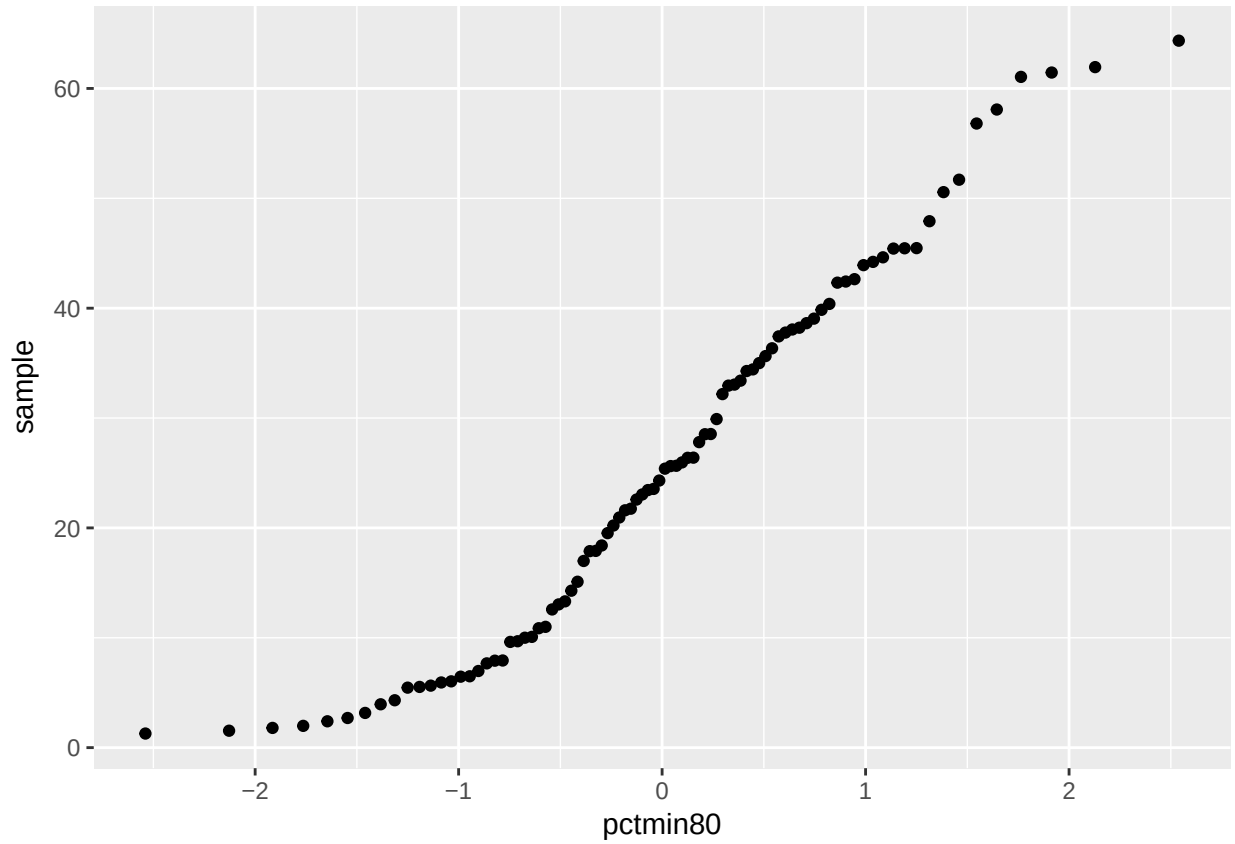

Crime Rate vs govt_wg
Crime Rate
0.100
0.075
0.050
0.025
350
400
450
govt_wg


sample
450
400
350
−2
−1
0
1
2
govt_wg

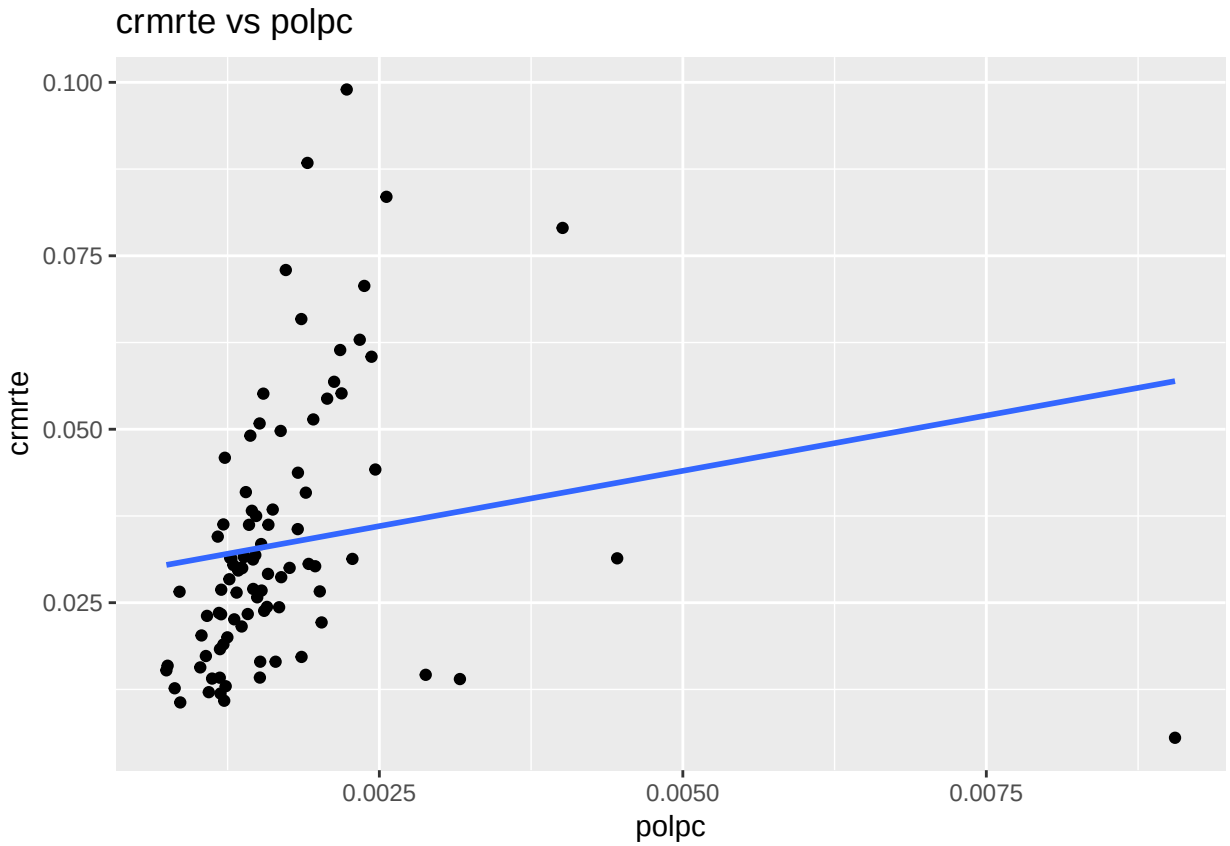

Crime Rate vs pctymle
0.100
0.075
0.050
0.025
Crime Rate
0.10
0.15
0.20
0.25
pctymle

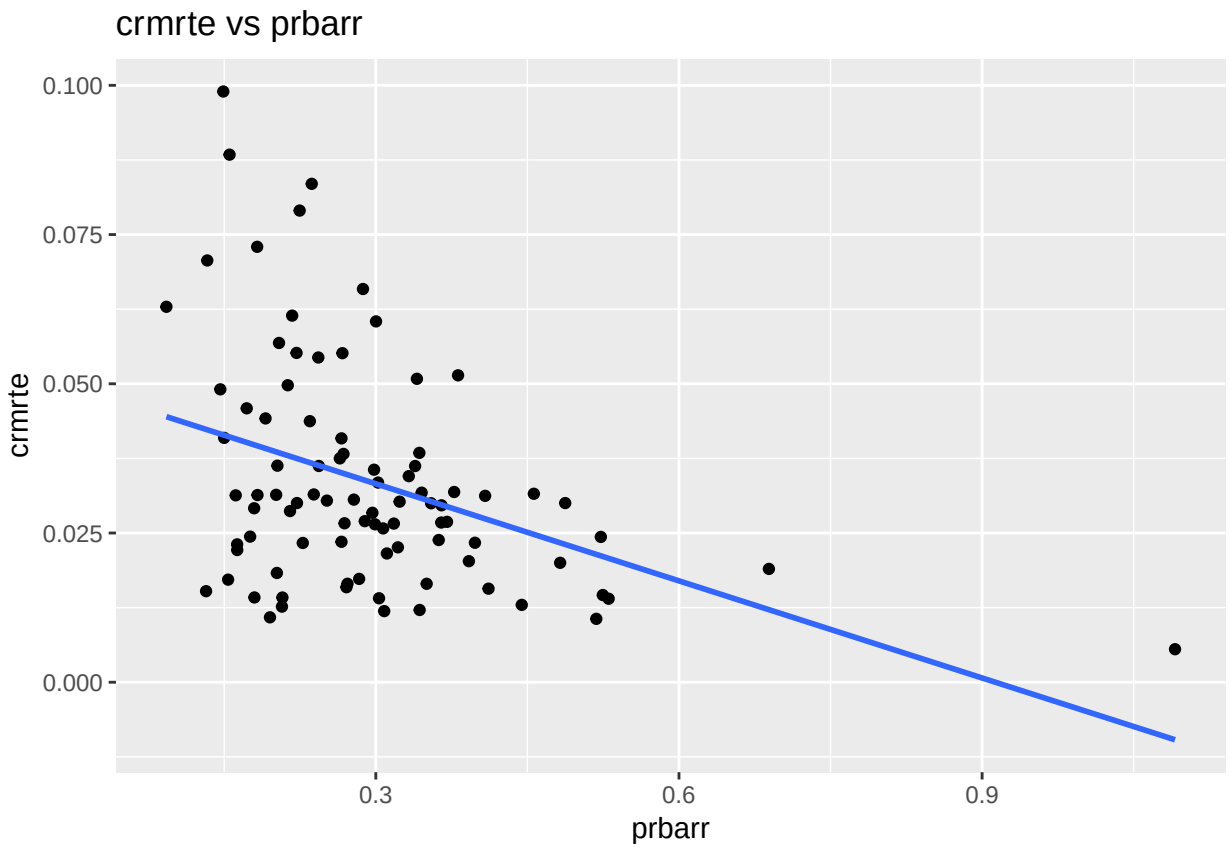

0.25
0.20
0.15
0.10
sample
−2
−1
0
1
2
pctymle

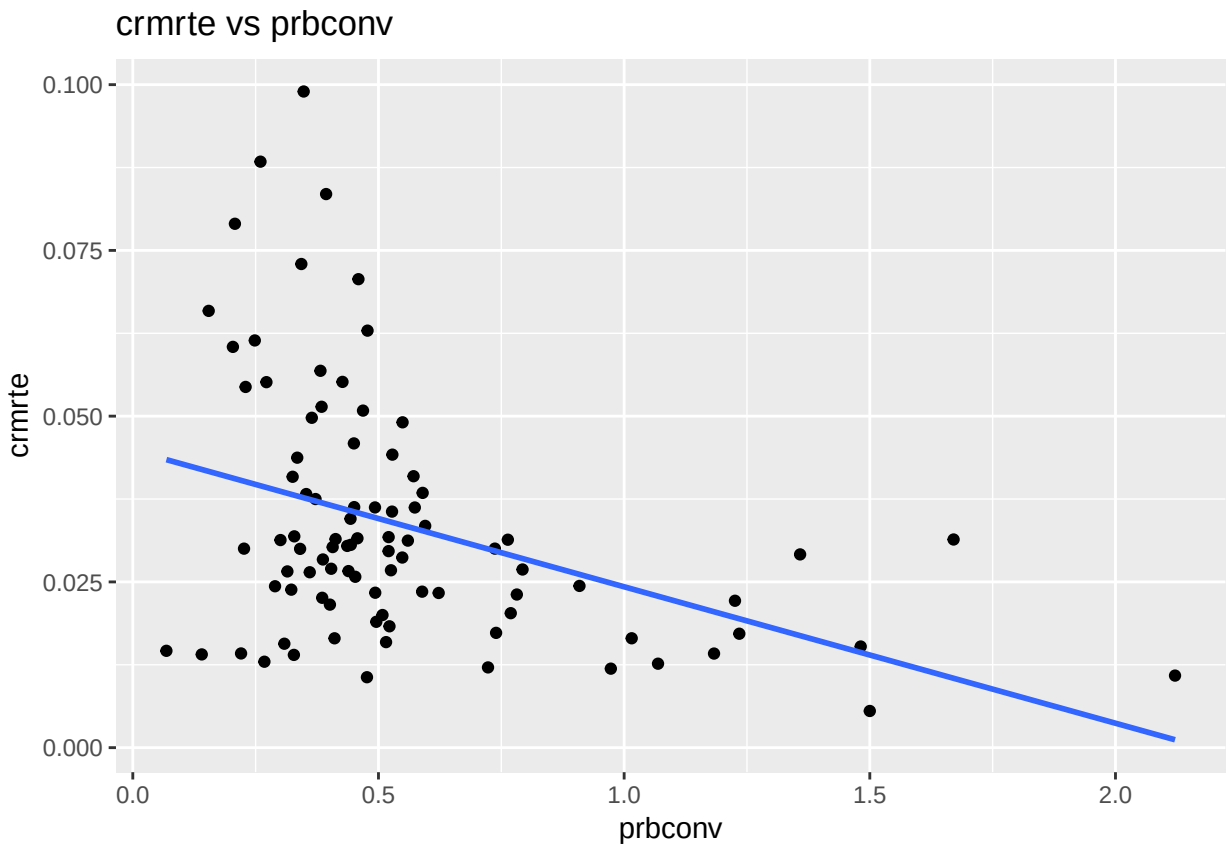

Crime Rate vs pctmin80
Crime Rate
0.100
0.075
0.050
0.025
0
20
40
60
pctmin80

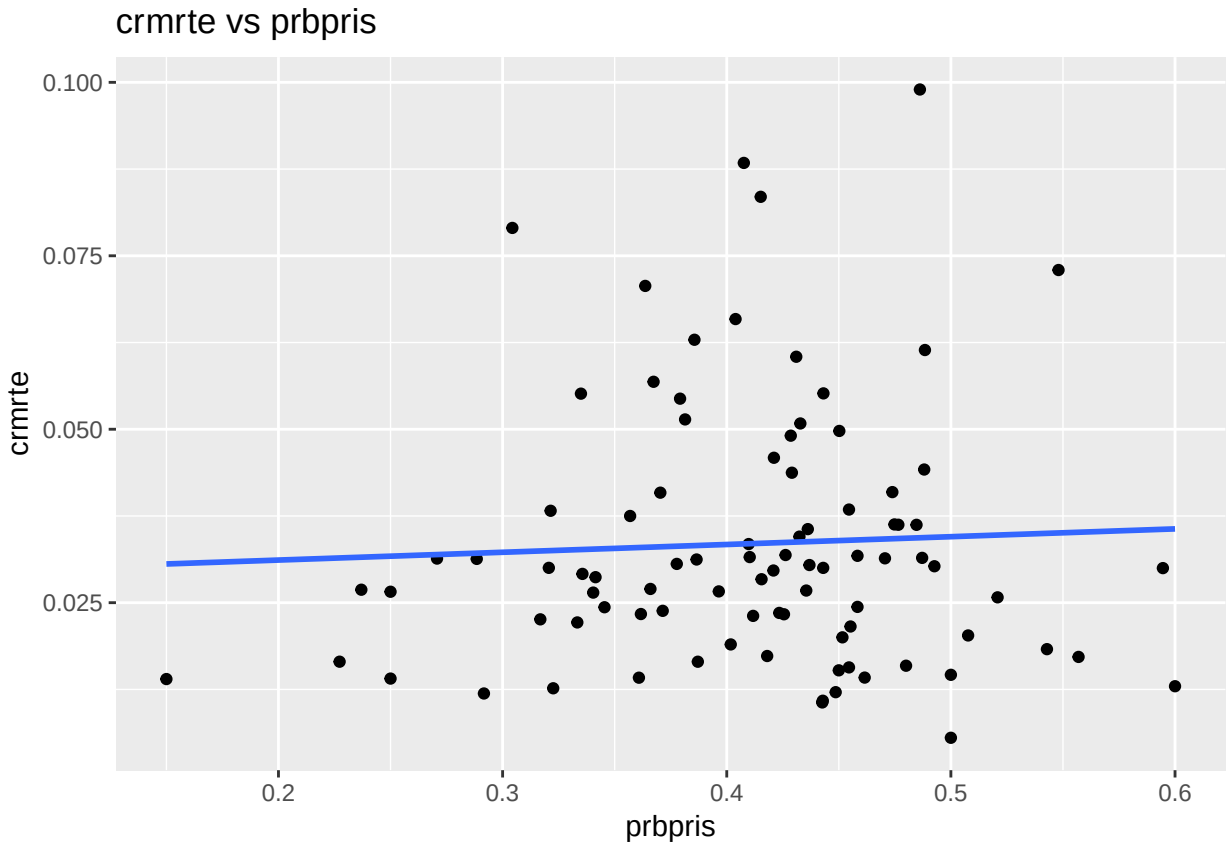


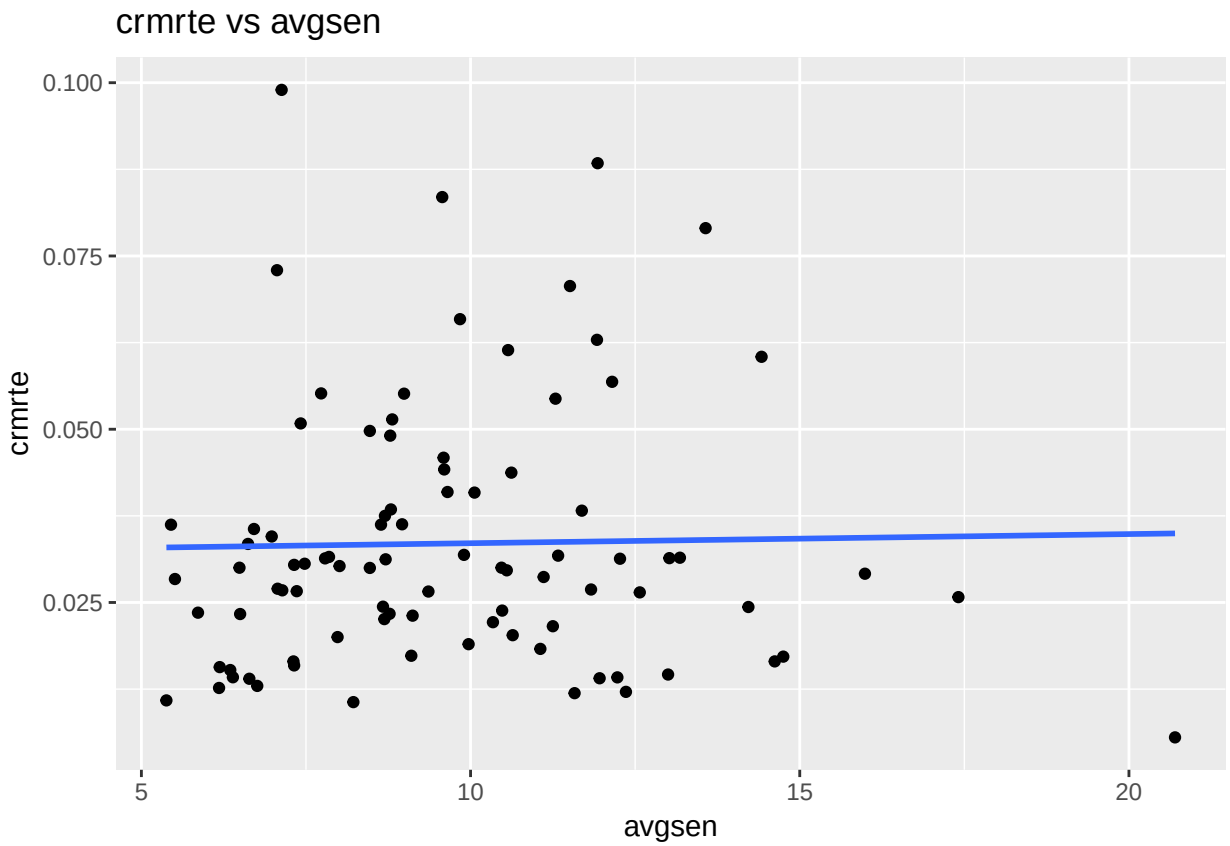

```
make_scatters <- function(df, var_list, y, trans) {

  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}'))
        )
    }
}


make_scatters(crime_na, mod1vars, y = 'crmrte')
```


crmrte vs polpc
0.100
0.075
0.050
0.025
crmrte
0.0025
0.0050
0.0075
polpc


crmrte vs prbarr
0.100
0.075
0.050
0.025
0.000
crmrte
0.3
0.6
0.9
prbarr


crmrte vs prbconv
0.100
0.075
0.050
0.025
0.000
crmrte
0.0
0.5
1.0
1.5
2.0
prbconv


crmrte vs prbpris
0.100
0.075
0.050
0.025
crmrte
0.2
0.3
0.4
0.5
0.6
prbpris




```
#Linear model retaining only significant variables from mod_2
model3_trans <- with(crime_na, lm(crmrte ~ log(polpc) + prbarr + prbconv + prbpris^2 +
                            avgsen + taxpc + sqrt(density) + govt_wg + log(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) %>% select(-std.error))
```

term	estimate	statistic	p.value
(Intercept)	0.1527653	4.2435120	0.0000638
log(polpc)	0.0163089	4.4132032	0.0000346
prbarr	-0.0347328	-4.0137409	0.0001432
prbconv	-0.0141941	-4.0925614	0.0001088
prbpris	0.0027770	0.2370283	0.8132987
avgsen	-0.0005200	-1.2665774	0.2093329
taxpc	0.0001843	1.8780075	0.0643742
sqrt(density)	0.0176515	4.4992318	0.0000252
govt_wg	-0.0000160	-0.3703152	0.7122199
log(pctymle)	0.0063302	1.1811138	0.2413913
pctmin80	0.0002966	3.2472176	0.0017617
west	-0.0041638	-1.0839028	0.2819749
central	-0.0043397	-1.5757408	0.1194097

term	estimate	statistic	p.value
urban	0.0023929	0.4379385	0.6627240
physical_wg	-0.0000128	-0.7829644	0.4361808
industry_wg	0.0000149	0.4895079	0.6259499
mix	-0.0192487	-1.3164731	0.1921337