W203 Statistics for Data Science

Section 2 - Lab 3

Cristopher Benge
cris.benge@berkeley.edu

Joy First
jfirst@berkeley.edu

Kevin Kory
knkory92@berkeley.edu

A study of criminal statistics in North Carolina for the honorable Terry Sanford (D-NC)

School of Information, Graduate Studies
University of California, Berkeley
United States

December 10, 2019

Contents

Section 1

Introduction

1.1 Abstract

We have been hired to provide sound criminal reform and policy research for Terry Sanford (D-NC), junior Senator representing North Carolina for the 100th U.S. Congress. We have obtained a single cross-section of crime statistics for a selection of counties in North Carolina from calendar year 1987 from which to construct our analysis. We endeavor to help the Sanford re-election campaign understand the determinants of crime and generate policy suggestions that are applicable to local [North Carolina] government agencies.

1.2 Code Book

Our crime statistics data was provided in a mysteriously sourced *crime_v2.csv* file for which we were provided only the following variable descriptions:

Pos	Variable	Description
1	county	county identifier
2	year	1987
3	crmrte	crimes committed per person
4	prbarr	'probability' of arrest
5	prbconv	'probability' of conviction
6	prbpris	'probability' of prison sentence
7	avgsen	avg. sentence, days
8	polpc	police per capita
9	density	people per sq. mile
10	taxpc	tax revenue per capita
11	west	=1 if in western N.C.
12	central	=1 if in central N.C.
⋮	⋮	⋮

Pos	Variable	Description
⋮	⋮	⋮
13	urban	=1 if in SMSA
14	pctmin80	perc. minority, 1980
15	wcon	weekly wage, construction
16	wtuc	wkly wge, trns, util, commun
17	wtrd	wkly wge, whlesle, retail trade
18	wfir	wkly wge, fin, ins, real estx1
19	wser	wkly wge, service industry
20	wmfg	wkly wge, manufacturing
21	wfed	wkly wge, fed employees
22	wsta	wkly wge, state employees
23	wloc	wkly wge, local gov emps
24	mix	offense mix: face-to-face/other
25	pctymle	percent young male

Table 1.1: Crime_V2 Code Book

In the literature on crime, researchers often distinguish between the certainty of punishment (do criminals expect to get caught and face punishment) and the severity of punishment (for example, how long prison sentences are). The former concept is the motivation for the 'probability' variables. The probability of arrest is proxied by the ratio of arrests to offenses, measures drawn from the FBI's Uniform Crime Reports. The probability of conviction is proxied by the ratio of convictions to arrests, and the probability of prison sentence is proxied by the convictions resulting in a prison sentence to total convictions. The data on convictions is taken from the prison and probation files of the North Carolina Department of Correction.

The percent young male variable records the proportion of the population that is male and between the ages of 15 and 24. This variable, as well as percent minority, was drawn from census data. The number of police per capita was computed from the FBI's police agency employee counts. The variables for wages in different sectors were provided by the North Carolina Employment Security Commission.

Section 2

Exploratory Data Analysis

2.1 Missing & Duplicate Observations

We observe an initial 97 samples in the *crime_v2.csv* file, as well as all 25 columns listed in the code book above. An initial pass through the data reveals two obvious data-collection errors: (**a**) 6 empty rows at the tail of the file (see 2.1a), and (**b**) one row that has been duplicated for county 193 (see 2.1b):

(a) 6 rows with missing values at tail of file

Show 10 entries Search:

	county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polp
	All	All	All	All	All	All	All	All
88	193	87	0.0235277	0.266054988	0.588859022	0.423422992	5.860000134	0.001178
89	193	87	0.0235277	0.266054988	0.588859022	0.423422992	5.860000134	0.001178
90	195	87	0.031397302	0.201397002	1.670519948	0.470587999	13.02000046	0.004459
91	197	87	0.0141928	0.207595006	1.182929993	0.360825002	12.22999954	0.001185
92								
93								
94								
95								
96								
97								

Showing 1 to 10 of 10 entries Previous 1 Next

(b) Row for county 193 duplicated

	county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc
	All	All	All	All	All	All	All	All
88	193	87	0.0235277	0.266054988	0.588859022	0.423422992	5.860000134	0.00117887
89	193	87	0.0235277	0.266054988	0.588859022	0.423422992	5.860000134	0.00117887

Figure 2.1: EDA : Duplicated and Missing Rows

2.2 Descriptive Statistics

Following removal of six empty rows and one of the duplicate rows, we are left with 90 observations useful for analysis. The descriptive statistics for the variables of interest were captured:

Table 2.1: EDA : Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
county	90	100.60	58.32	1	51.5	150.5	197
year	90	87.00	0.00	87	87	87	87
crmrte	90	0.03	0.02	0.01	0.02	0.04	0.10
prbarr	90	0.30	0.14	0.09	0.20	0.34	1.09
prbconv	90	0.55	0.35	0.07	0.34	0.59	2.12
prbpris	90	0.41	0.08	0.15	0.36	0.46	0.60
avgsen	90	9.69	2.83	5.38	7.38	11.47	20.70
polpc	90	0.002	0.001	0.001	0.001	0.002	0.01
density	90	1.44	1.52	0.0000	0.55	1.57	8.83
taxpc	90	38.16	13.11	25.69	30.73	41.01	119.76
west	90	0.24	0.43	0	0	0	1
central	90	0.38	0.49	0	0	1	1
urban	90	0.09	0.29	0	0	0	1
pctmin80	90	25.71	16.98	1.28	10.02	38.18	64.35
wcon	90	285.35	47.75	193.64	250.75	314.98	436.77
wtuc	90	410.91	77.36	187.62	374.33	440.68	613.23
wtrd	90	210.92	33.87	154.21	190.71	224.28	354.68
wfir	90	321.62	54.00	170.94	285.56	342.63	509.47
wser	90	275.34	207.40	133.04	229.34	277.65	2,177.07
wmfg	90	336.03	88.23	157.41	288.60	359.89	646.85
wfed	90	442.62	59.95	326.10	398.78	478.26	597.95
wsta	90	357.74	43.29	258.33	329.27	383.15	499.59
wloc	90	312.28	28.13	239.17	297.23	328.78	388.09
mix	90	0.13	0.08	0.02	0.08	0.15	0.47
pctymle	90	0.08	0.02	0.06	0.07	0.08	0.25

From the descriptive statistics, we note several potential areas of concern and interest. First, the *county* variable appears to be the EPA FIPS code for North Carolina counties. The values are odd numbered only and from Figure 2.8 (below) we can see that the Central/West/East indicators provided in the data geographically aligns using these values as FIPS codes.

Additionally, the variables *wser*, *density*, *polpc*, *taxpc*, and *pctymle* all appear to have a distribution that suggest potential outliers. We will explore these variables to see if there may be more issues in the data collection that we can address.

2.3 Outlier Analysis

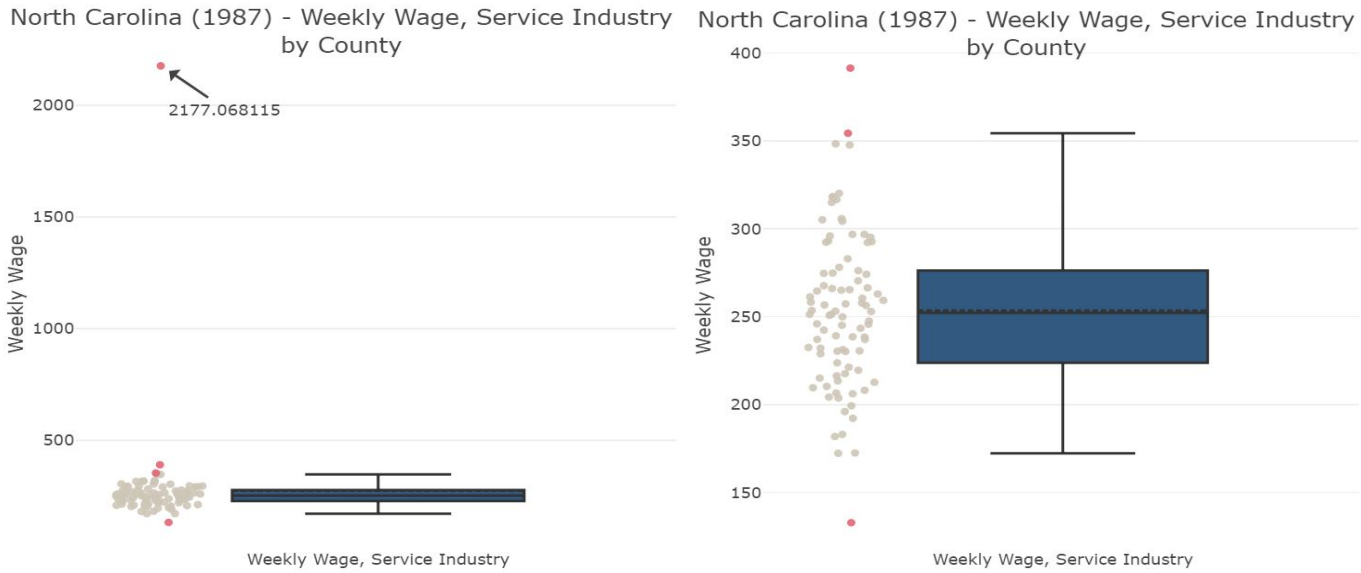
2.3.1 Weekly Wage, Service Industry [WSER]

We start with *wser*, which appears to be the *average weekly income for service industry workers.* There exists a single large maximum value of 2177.07 which appears to be well outside the distribution of the other values (see 2.2a). The remaining values appear to be in the range of 133-348, so it seems very unlikely that only one county has a value in the 2,000+ weekly range ($104,000 / year) for the service industry. The county tied to this record is 185, which is the FIPS code for Warren County, North Carolina.

This value appears to be the result of a decimal placement issue, where the likely real value is 217.7068, based on a survey of the counties surrounding Warren County: Vance County (347.6609), Franklin County (239.2233), Nash County (305.7612), Halifax County (172.6281), and Northampton County (213.5822). Given these surrounding county wages for service industry professionals, we come to a regional mean average of $255.7711:

$$\mu_{regional\_wser} = \frac{Vance + Franklin + Nash + Halifax + Northampton}{n}$$
$$= \frac{347.6609 + 239.2233 + 305.7612 + 172.6281 + 213.5822}{5} = \frac{1278.856}{5}$$
$$\therefore = \mathbf{255.7711}$$

Given these results, we elect to remediate the large outlier in *wser* by multiplying the value by 0.1. The impact to distribution of values is depicted in 2.2b below:



(a) WSER with uncorrected, large outlier

(b) WSER with large outlier corrected

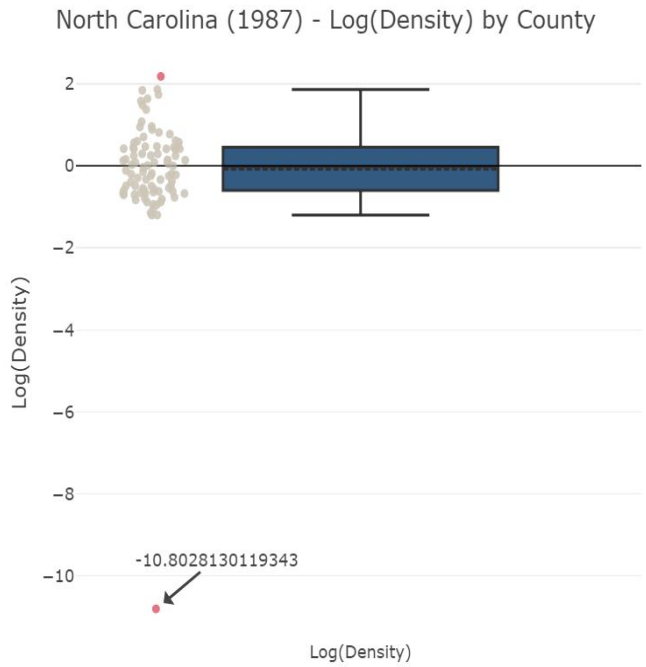
Figure 2.2: Outliers : Weekly Wage, Service Industry (WSER)

2.3.2 People per Square Mile [DENSITY]

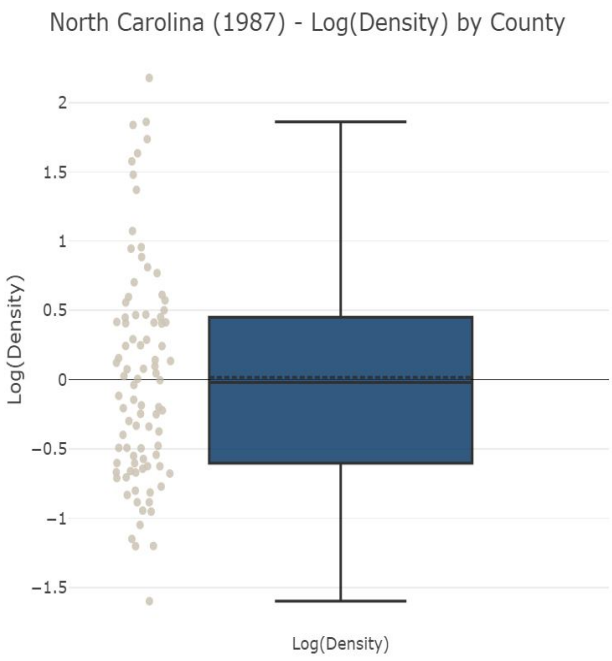
Observation 79 (county = 173, Swain County) is currently listed with a *density* of of 0.0000203422 people per square mile. By a wide-margin, this is the lowest value in the dataset (see 2.3a) and appears to be a potential mistake. According to Wikipedia, Swain County has a landmass of 541 square miles, which would equal 0.011 people living in the entire county - an impossibility.

According to U.S. Census Bureau records, Swain County North Carolina had a population of 10,932 in 1987. Upon reviewing *density* more closely along with the Census Bureau records for population and the square mile landmass reported on Wikipedia, it appears that *this variable is actually in units of 100 people per square mile*. With that adjustment, the data for Swain County would still equal only 1.1 person for the entire county; which is still clearly incorrect.

Based on the adjusted amount of 109.32 persons (in units of 100), the correct *density* value for Swain County in 1987 should be 0.202070. We adjust accordingly (see 2.3b):



(a) log(density) with uncorrected, small outlier



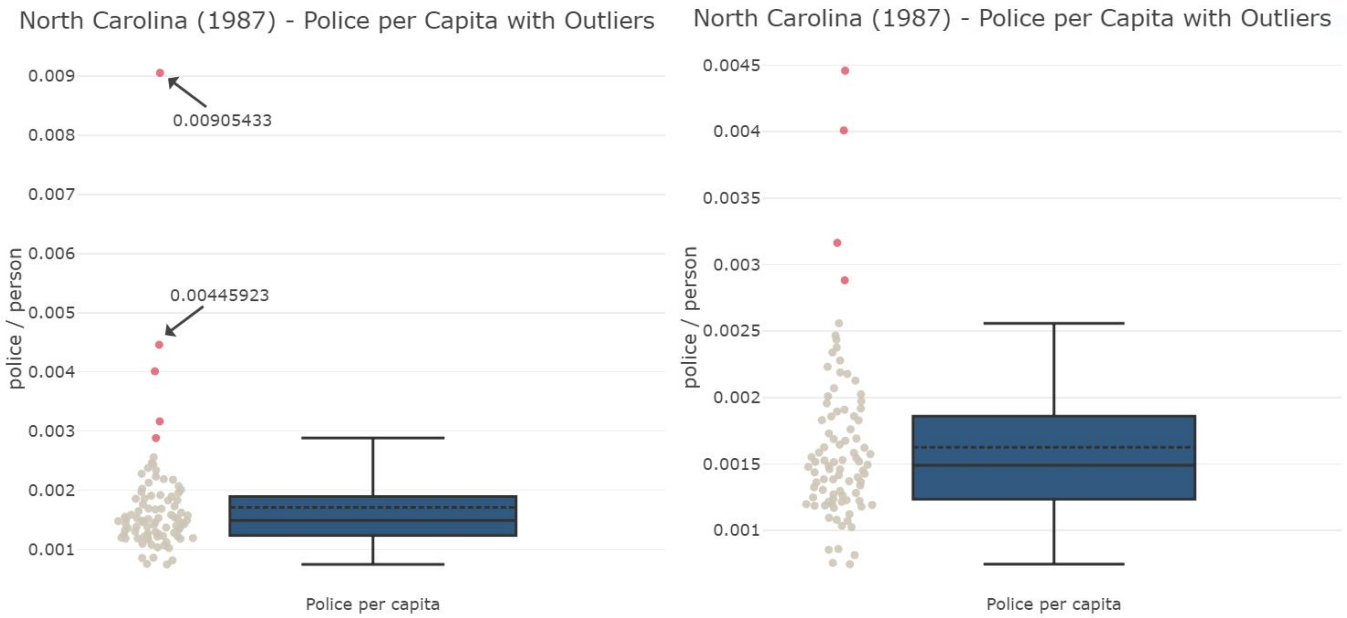
(b) log(density) with small outlier, corrected

Figure 2.3: Outliers : People per Square Mile (DENSITY)

2.3.3 Police per Capita [POLPC]

There are a total of five data points in the *polpc* variable that qualify as anomalous (IQR Rule), but one stands well above the others and warrants additional scrutiny. The entry for county 115 (Madison County) has a value of 0.00905433 (see 2.4a), significantly higher than the other values in all other counties. According to the U.S. Census, Madison County, NC had a population size of only 17,051 residents in 1987 making it one of the smaller counties in the state overall. Madison County covers only 452 square miles of geography and is located in the Northwest portion of the state, directly bordering Tennessee.

At this per capita level, Madison County would have 154.38538 officers covering just 17,051 people. The mean of the *polpc* variable is 0.00162543 (excluding Madison County); with this value substituted for Madison, we would have a more realistic level of ≈ 27.715 law enforcement officers, which is in-line with other counties in the 20k and below population range. We'll substitute with the mean for Madison County to address this apparent mistake; 2.4b reflects the data distribution following the adjustment:



(a) POLPC with uncorrected, large outlier

(b) POLPC with large outlier corrected

Figure 2.4: Outliers : Police per Capita (POLPC)

2.3.4 Tax Revenue per Capita [TAXPC]

Next, we analyze a single large outlier found in the *taxpc* variable. According to the code book (ref: 1.1), *taxpc* is the tax revenue per capita and while the typical range is from 25 - 75, county 55 (Dare County) has a large value of 119.76 per person. In 1987, Dare County had a population of 19,580 according to U.S. Census records. The tax rate in Dare County is roughly the same as other counties at 2% with a total state + county combined rate of 6.75%.

Based on the historical tax rates and the increasing burden we see in NC taxes from 1981-1987 (reference), it is not clear if the value reported in the data is incorrect. As such, we elect not to treat this value and leave it as-is for the purposes of our analysis.

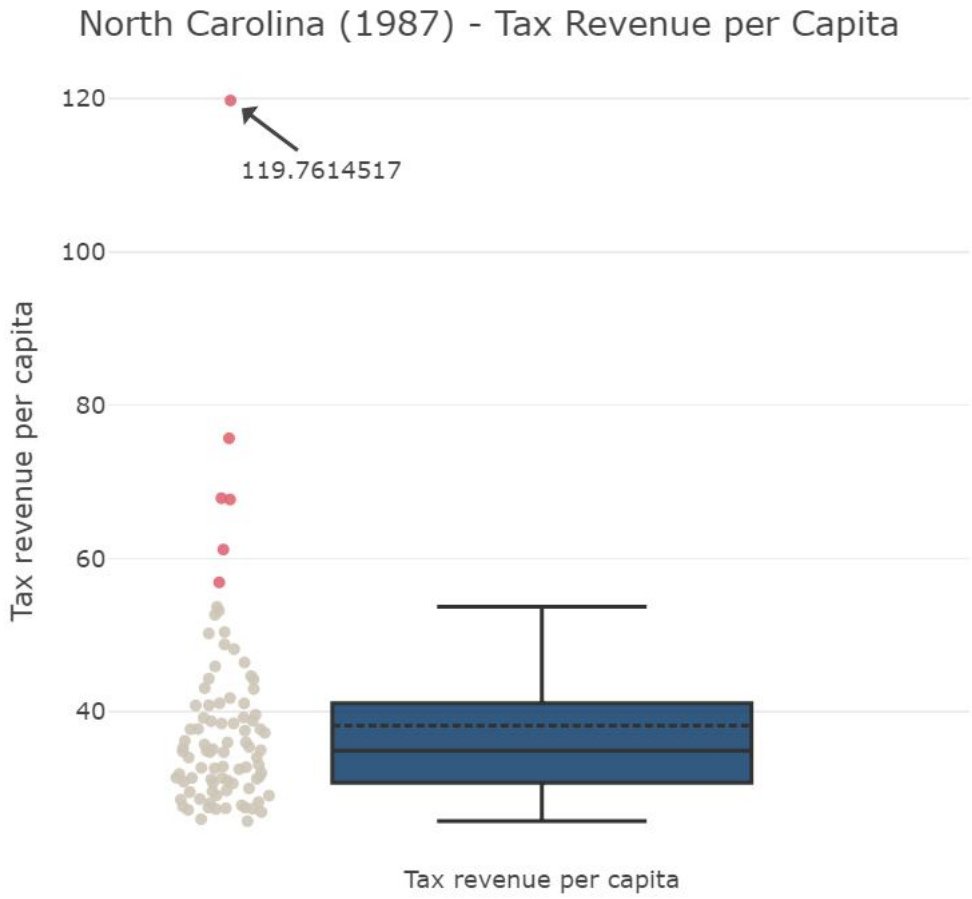


Figure 2.5: Outliers : Tax Revenue per Capita (TAXPC)

2.3.5 Percentage of Demographic as Young Males [PCTYMLE]

We analyze a single large outlier found in the *pctymle* variable. According to the code book (ref: 1.1), *pctymle* is the percent of young males representing the county population. County 133 (Onslow County) has a relatively large value of almost 25%, warranting further investigation. According to Wikipedia, Onslow County is home to the U.S. Marine Corps Base Camp Lejeune.

Though women have been permitted to join the U.S. Marines since 1918, historically they have made up less than 10% of all U.S. Marines roles (reference). It was not until calendar year 2016 that women were allowed to serve in all roles. In addition to this, contemporary demographics statistics report that the median age of Onslow County is 25 (reference). Given these data, *we conclude that this large value is accurate* and elect to leave it as-is for the purposes of this analysis.

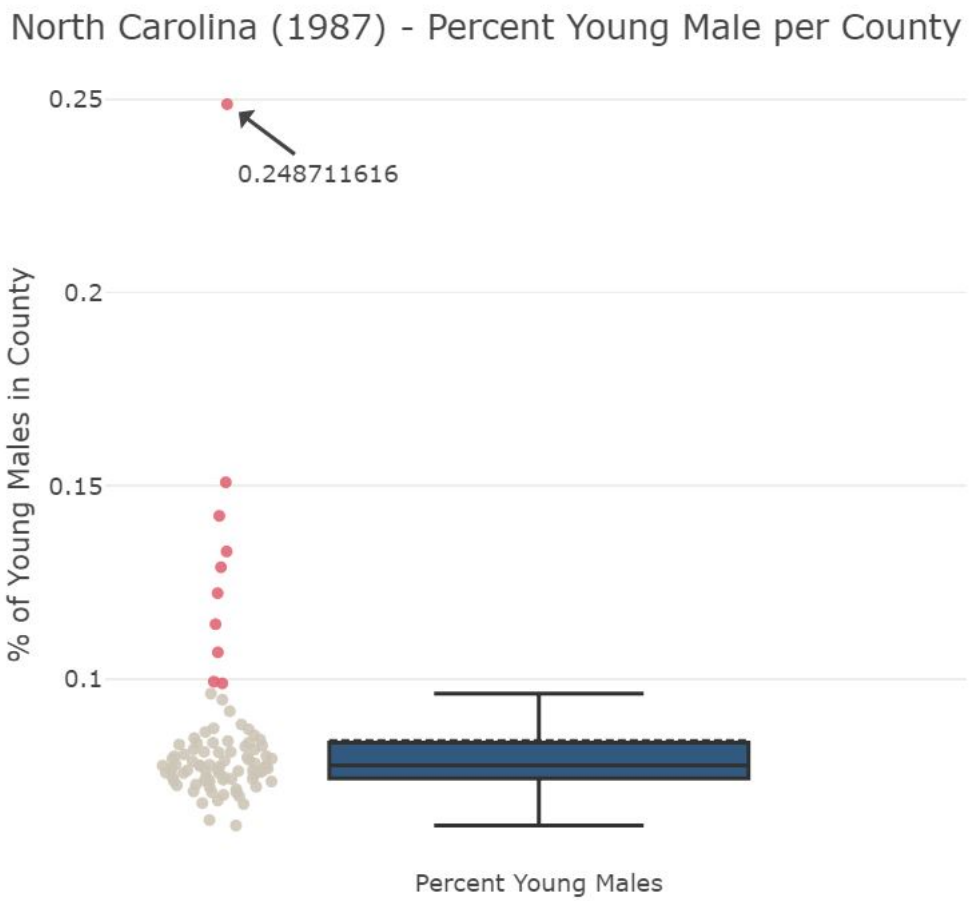


Figure 2.6: Outliers : Percentage of Demographic as Young Males (PCTYMLE)

2.4 Location Errata

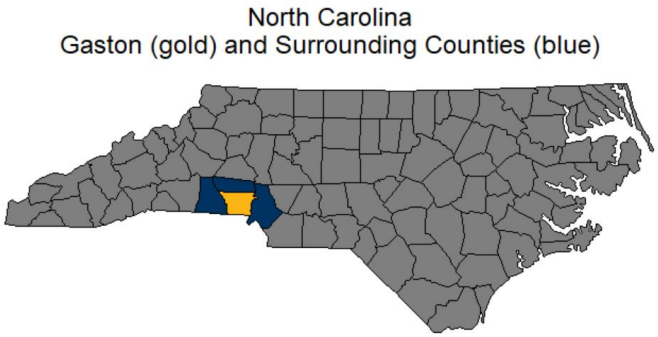
The use of one-hot encoding for location variables *west* and *central* is potentially problematic as it allows for the possibility of insert/update anomalies. That is, the form of the data allows for impossible assignments into more than one location. The dataset we were provided with contains only one such anomaly for Gaston County (FIPS code 71). This variable has inadvertently been assigned to both the "Central" and "West" groups (see figures 2.7a and 2.7b).

Show 5 entries Search:

county	west	central	urban	crmrte
All	All	All	All	All
45	0	1	0	0.036280699
71	1	1	0	0.054406099
109	0	1	0	0.023099501
119	0	1	1	0.098965898

Showing 1 to 4 of 4 entries Previous 1 Next

(a) Gaston County assigned to two location codes



(b) Gaston and surrounding counties

Figure 2.7: EDA : Location Category of Gaston County

Gaston County is surrounded by only three North Carolina counties: Cleveland County to the West, Lincoln County to the North, and Mecklenburg County to the East. In this case, all of the surrounding counties are labeled as members of the "Central" category, so we correct the value for Gaston by removing it from the "West" category, leaving it assigned to "Central". Following this correction, all counties appear to be distributed uniformly (see 2.8)

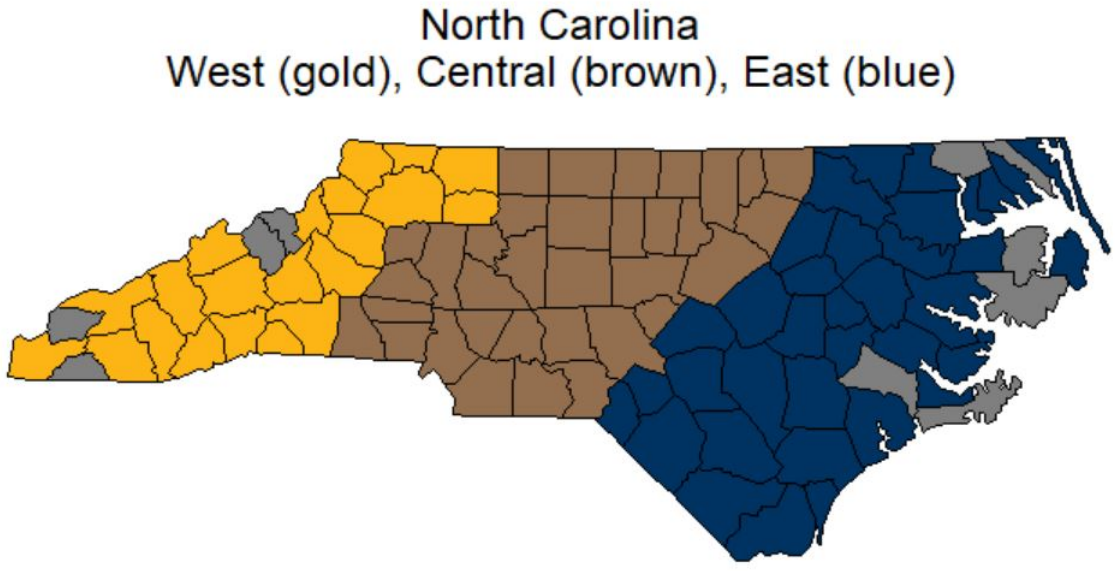


Figure 2.8: EDA : North Carolina Geographic Boundaries (West, Central, East)

2.5 Crime Rate by County

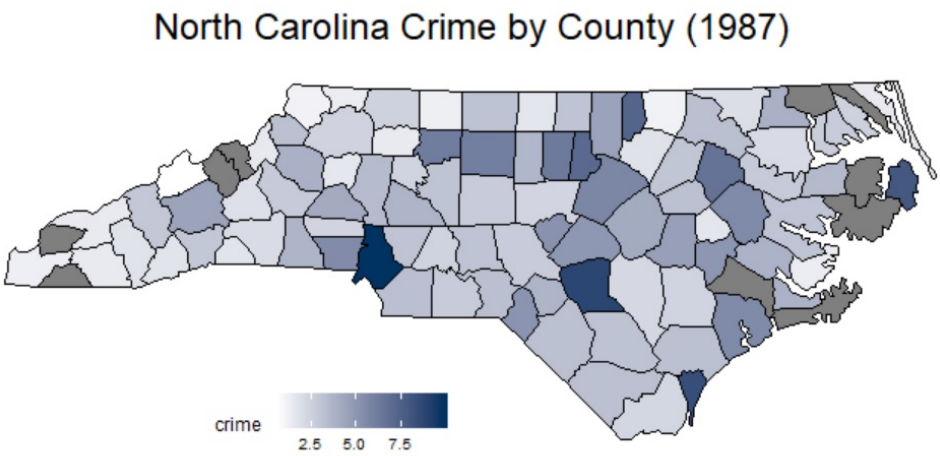
Earlier, we identified the *county* variable as the FIPS code for North Carolina counties (ref 2.2). We can use this information to identify counties missing from our dataset as well as plot crime rates at a county level to see if there are any overt geographical signals of crime rate.

We received data for only 90 of the 100 counties in North Carolina; the missing counties are shown in Figure 2.9, and are identified geographically in Figure 2.10 [in grey].

Figure 2.9: EDA : Missing Counties

FIPS	County
29	Camden
31	Carteret
43	Clay
73	Gates
75	Graham
95	Hyde
103	Jones
121	Mitchell
177	Tyrrell
199	Yancey

Figure 2.10: EDA : North Carolina Crime by County, 1987



No apparent strong crime rate patterns exist from a purely visual geographic positioning perspective; however, it is noteworthy that counties with high crime rate per West/Central/East grouping *in general* tends to increase moving West to East.

Figure 2.11: EDA : Top 10 Counties by Crime Rate

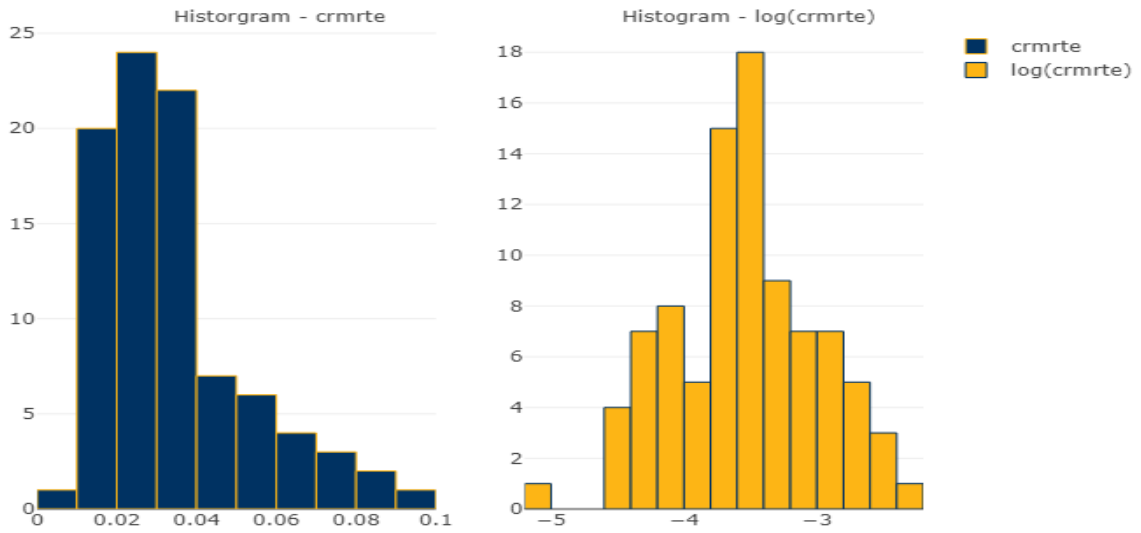
FIPS	County	Crime Rate
119	Mecklenburg	9.897%
51	Cumberland	8.838%
129	New Hanover	8.350%
55	Dare	7.902%
181	Vance	7.295%
63	Durham	7.066%
65	Edgecombe	6.588%
135	Orange	6.290%
67	Forsyth	6.142%
81	Guilford	6.045%

Figure 2.12: EDA : Bottom 10 Counties by Crime Rate

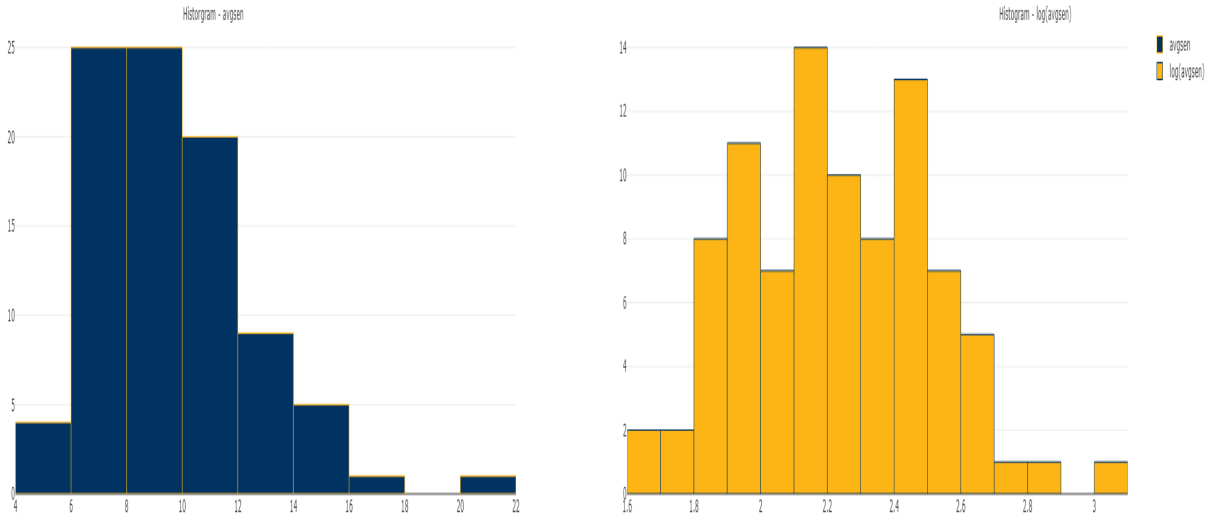
FIPS	County	Crime Rate
117	Martin	0.553%
9	Ashe	1.062%
185	Warren	1.087%
39	Cherokee	1.192%
169	Stokes	1.210%
137	Pamlico	1.267%
5	Alleghany	1.296%
173	Swain	1.399%
53	Currituck	1.407%
197	Yadkin	1.419%

2.6 Frequency Distribution (Natural & Log)

After identifying and, when appropriate, treating outliers, we move to consider the distribution of our data. Here, we provide a Histogram view for all raw and log transformed data only for variables that are non-binary or of no regression value (i.e. *west*, *central*, *urban*, *county*, and *year*). We evaluate log transformations for their common utility in improving explanatory power and natural tendency to normalize distribution:



(a) EDA : Histogram of CRMRTE and log(CRMRTE)

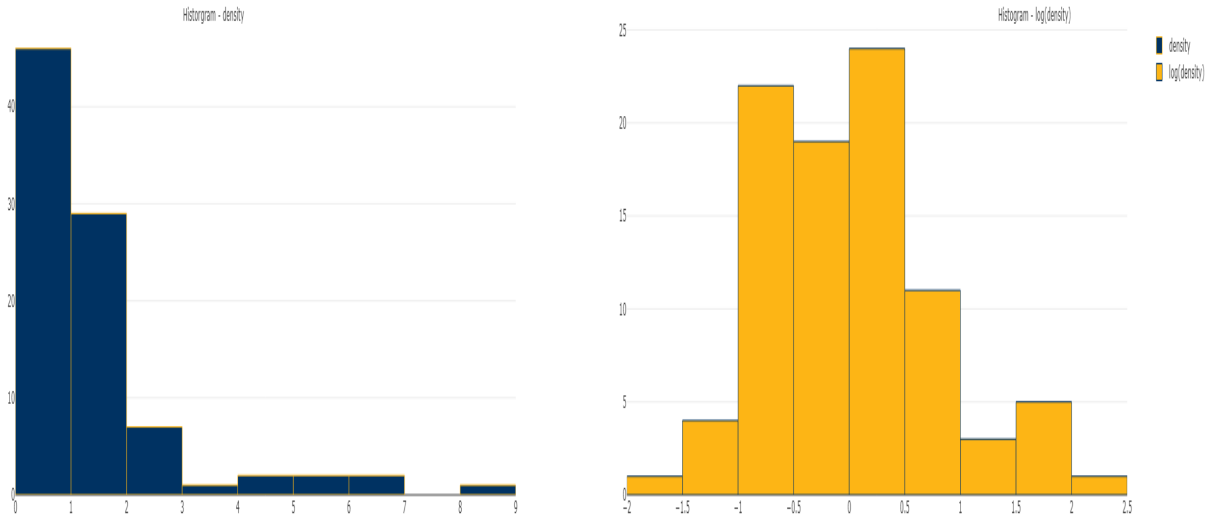


(b) EDA : Histogram of AVGSEN and log(AVGSEN)

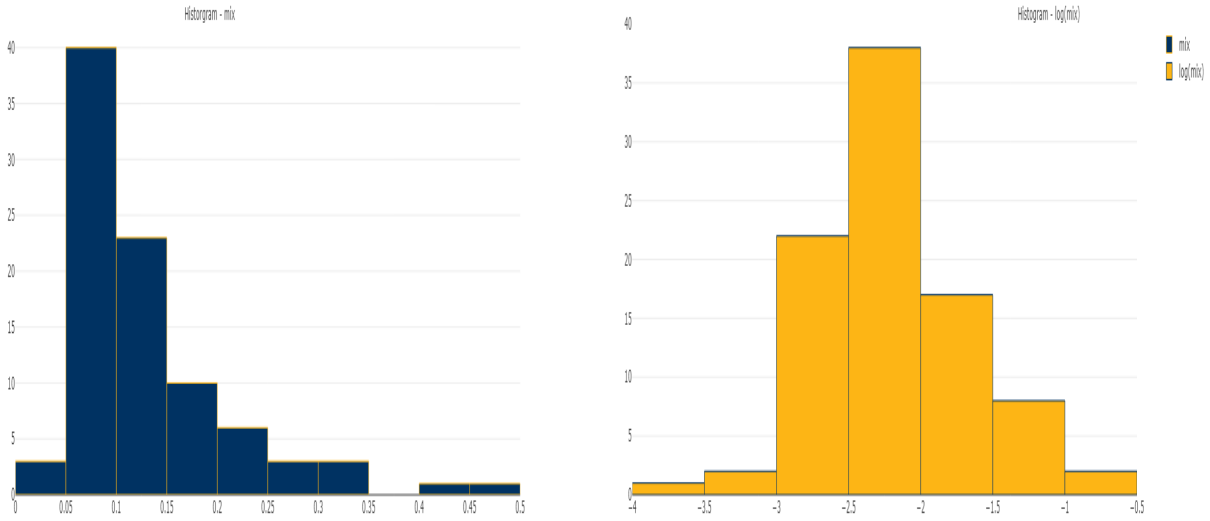
Figure 2.13: EDA : Distribution of Variables CRMRTE and AVGSEN

EXPLORATORY DATA ANALYSIS - CONTD.

Histograms Continued.



(a) EDA : Histogram of DENSITY and log(DENSITY)

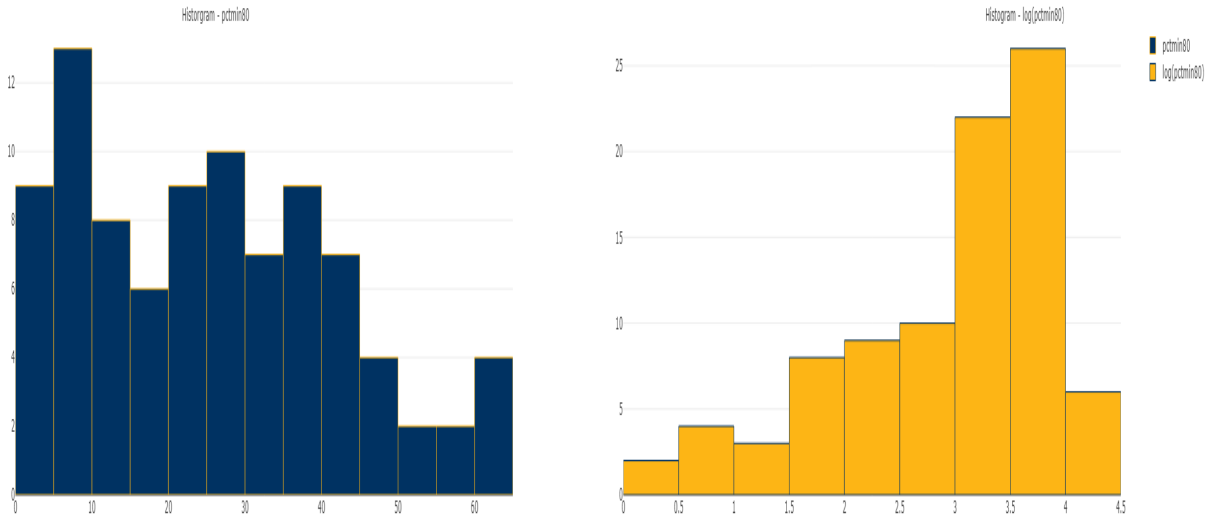


(b) EDA : Histogram of MIX and log(MIX)

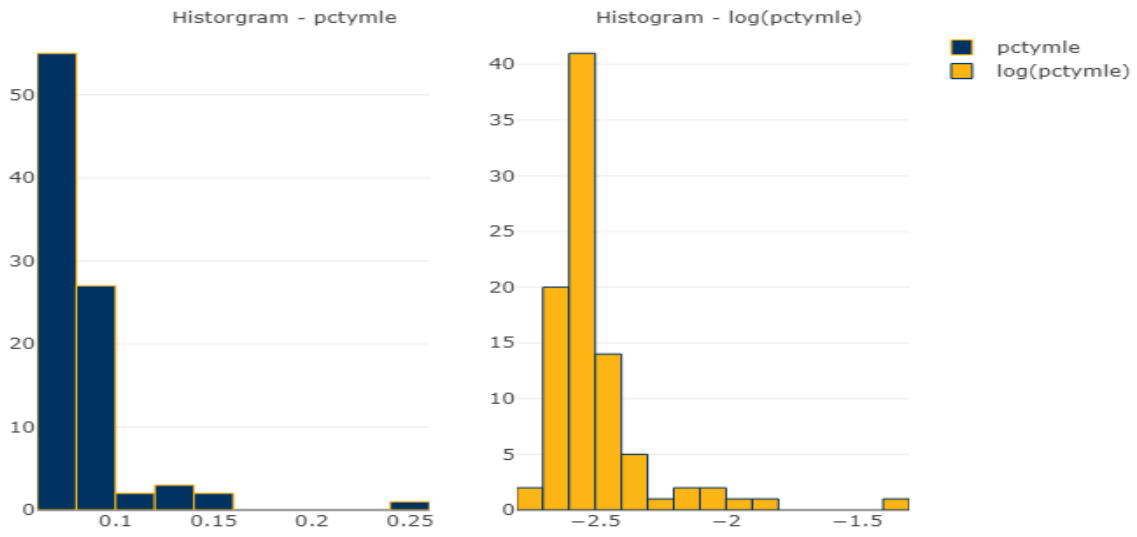
Figure 2.14: EDA : Distribution of Variables DENSITY and MIX

EXPLORATORY DATA ANALYSIS - CONTD.

Histograms Continued.



(a) EDA : Histogram of PCTMIN80 and log(PCTMIN80)

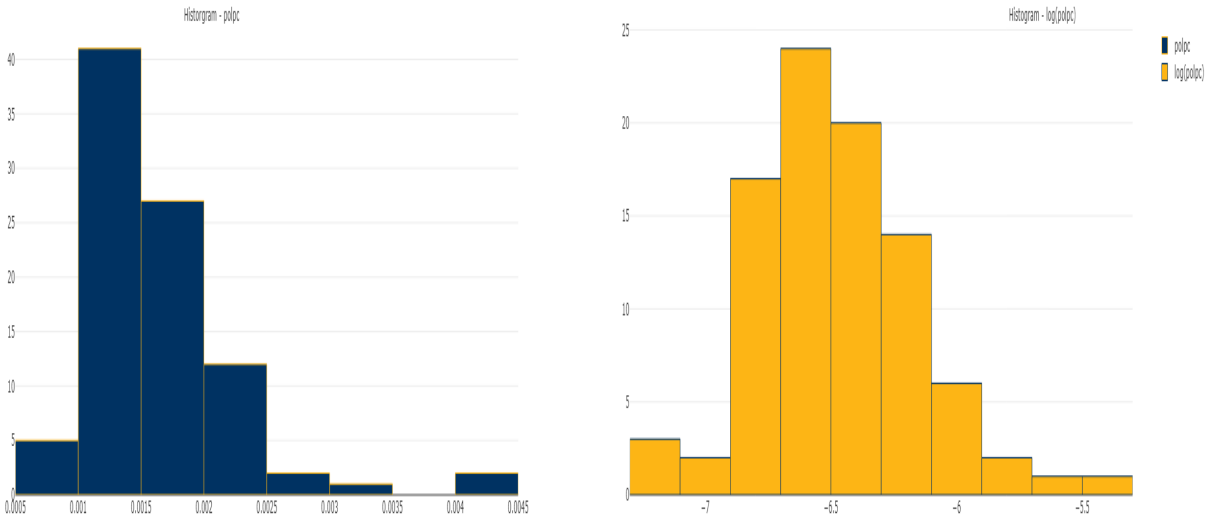


(b) EDA : Histogram of PCTYMLE and log(PCTYMLE)

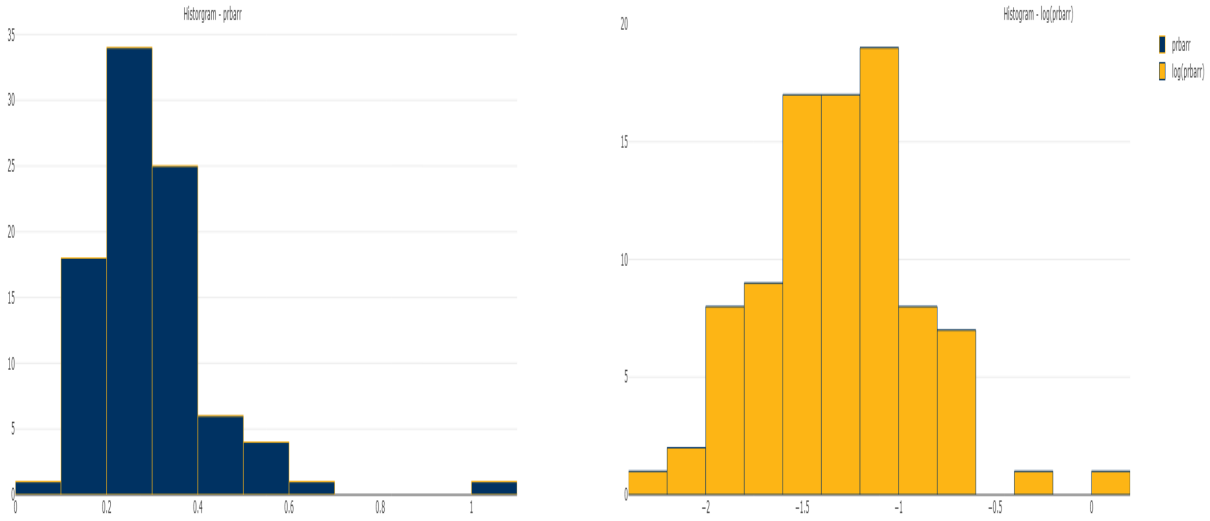
Figure 2.15: EDA : Distribution of Variables PCTMIN80 and PCTYMLE

EXPLORATORY DATA ANALYSIS - CONTD.

Histograms Continued.



(a) EDA : Histogram of POLPC and log(POLPC)

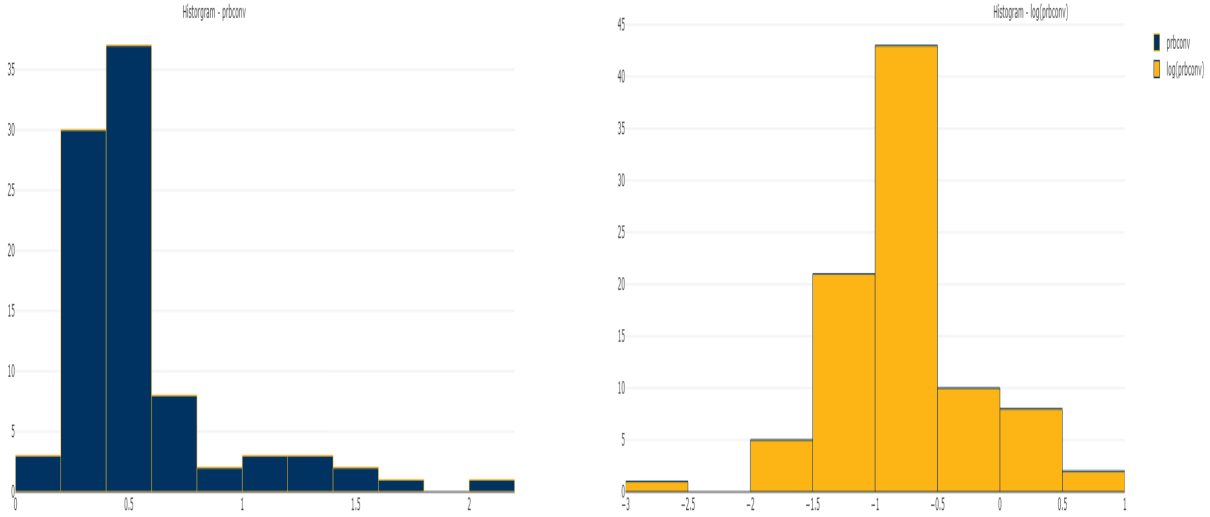


(b) EDA : Histogram of PRBARR and log(PRBARR)

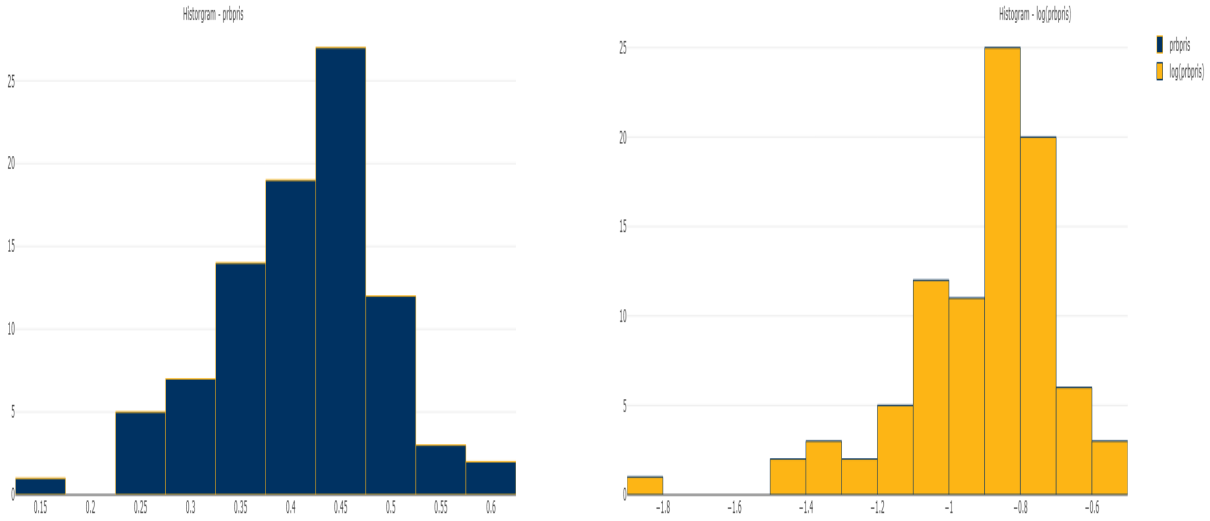
Figure 2.16: EDA : Distribution of Variables POLPC and PRBARR

EXPLORATORY DATA ANALYSIS - CONTD.

Histograms Continued.



(a) EDA : Histogram of PRBCONV and log(PRBCONV)

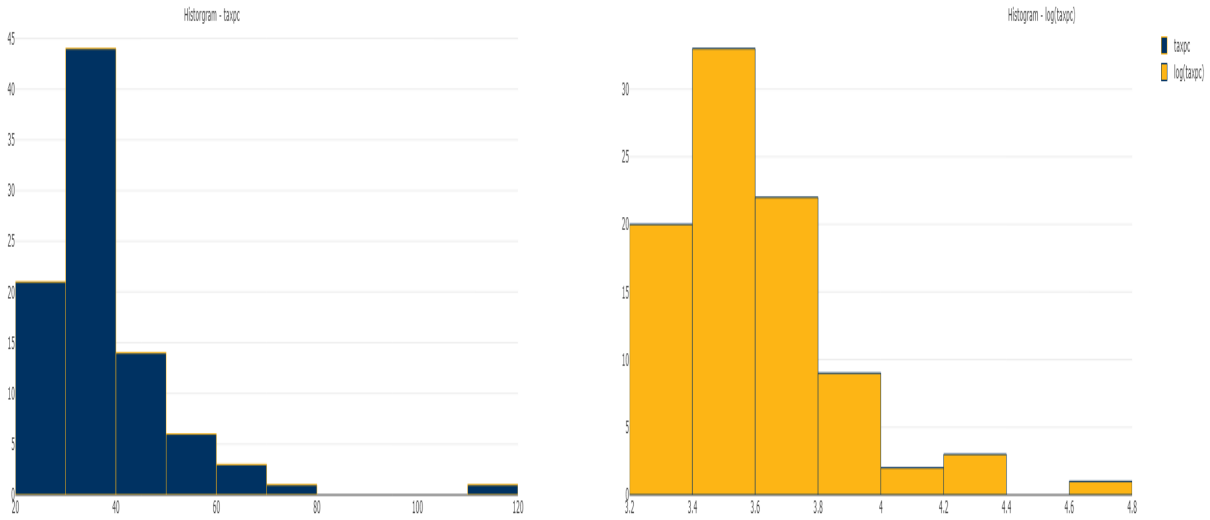


(b) EDA : Histogram of PRBPRIS and log(PRBPRIS)

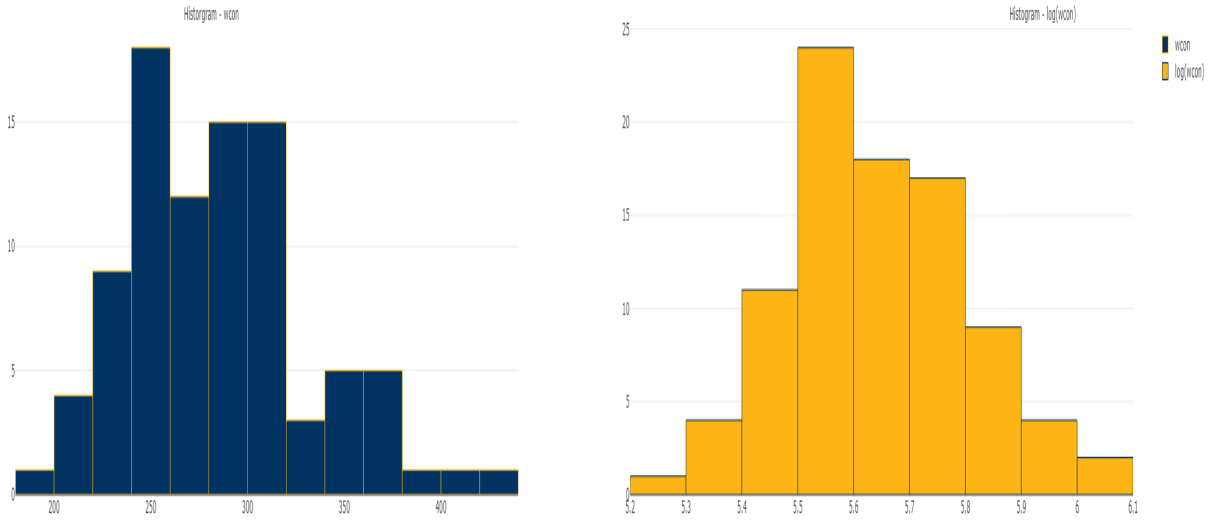
Figure 2.17: EDA : Distribution of Variables PRBCONV and PRBPRIS

EXPLORATORY DATA ANALYSIS - CONTD.

Histograms Continued.



(a) EDA : Histogram of TAXPC and log(TAXPC)

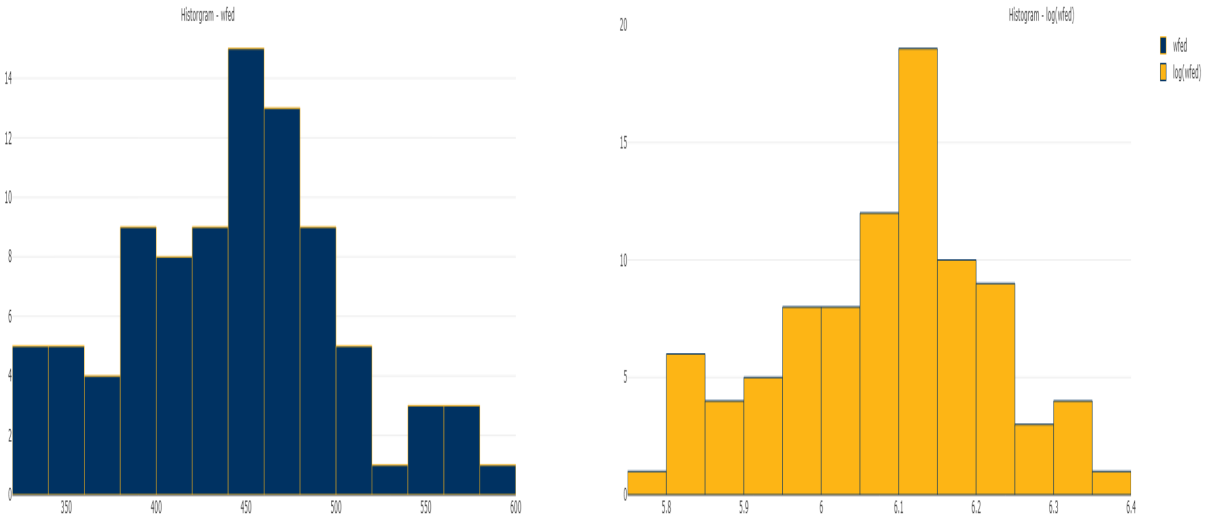


(b) EDA : Histogram of WCON and log(WCON)

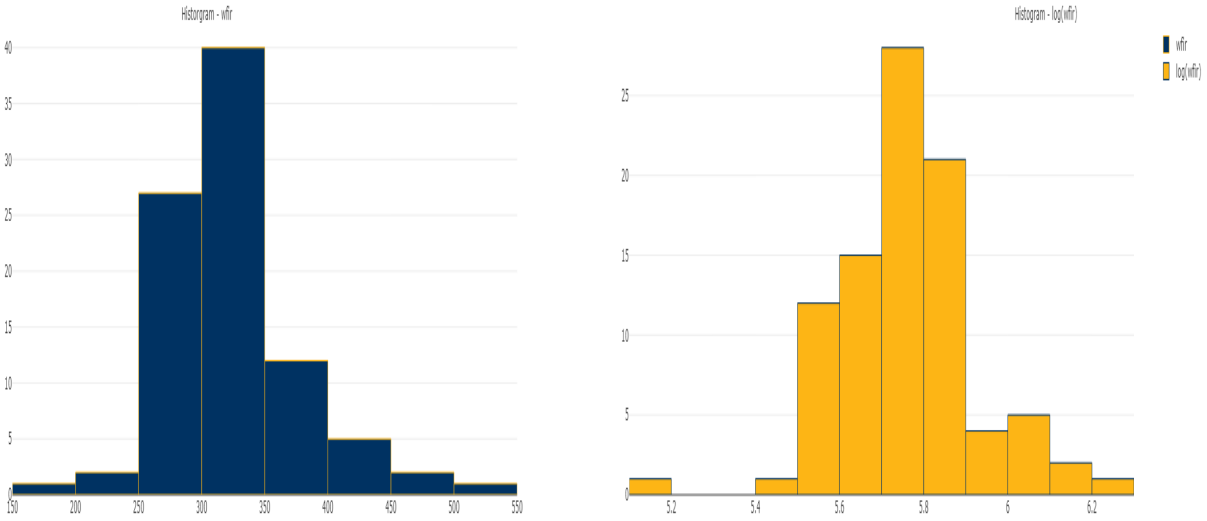
Figure 2.18: EDA : Distribution of Variables TAXPC and WCON

EXPLORATORY DATA ANALYSIS - CONTD.

Histograms Continued.



(a) Histogram of WFED

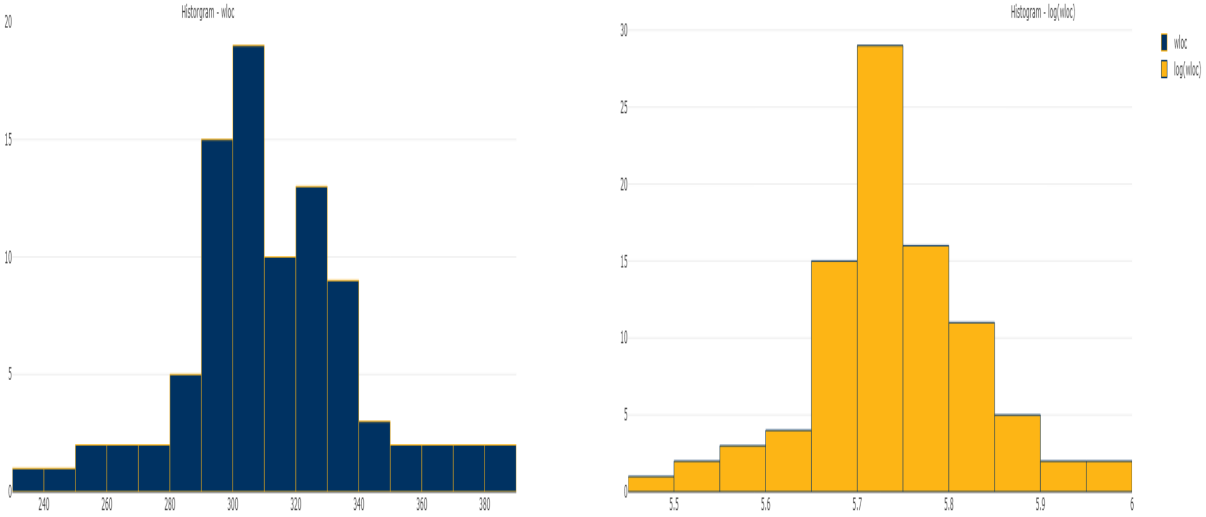


(b) Histogram of WFIR

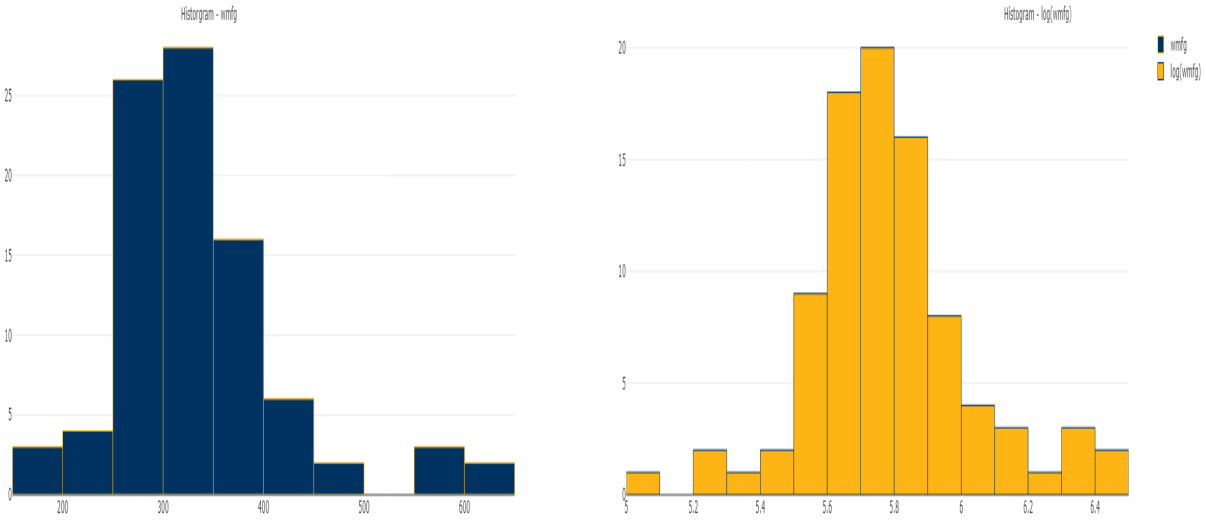
Figure 2.19: EDA : Distribution of Variables WFED and WFIR

EXPLORATORY DATA ANALYSIS - CONTD.

Histograms Continued.



(a) Histogram of WLOC

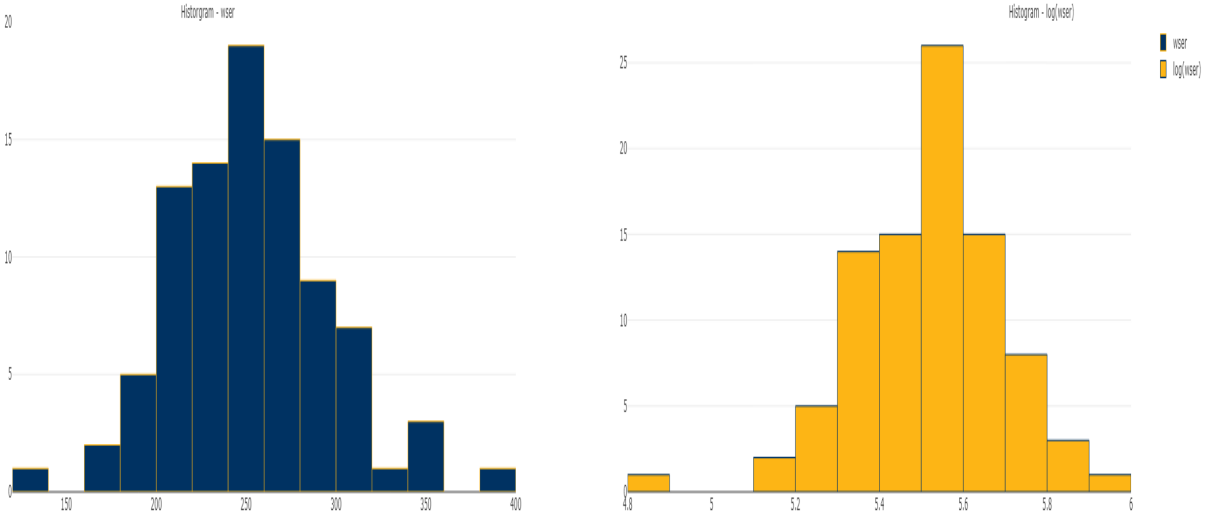


(b) Histogram of WMFG

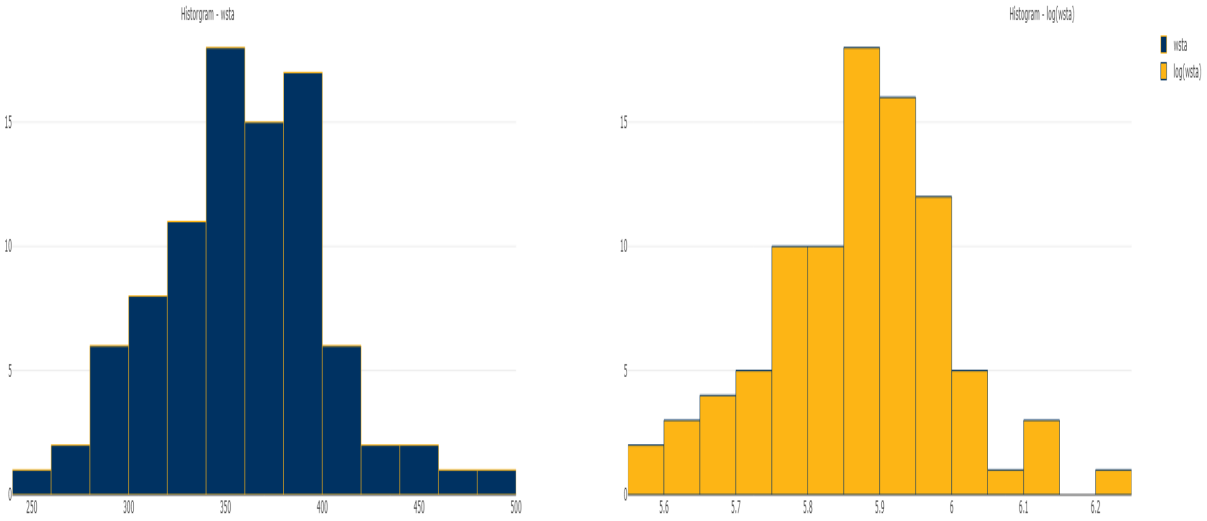
Figure 2.20: EDA : Distribution of Variables WLOC and WMFG

EXPLORATORY DATA ANALYSIS - CONTD.

Histograms Continued.



(a) Histogram of WSER

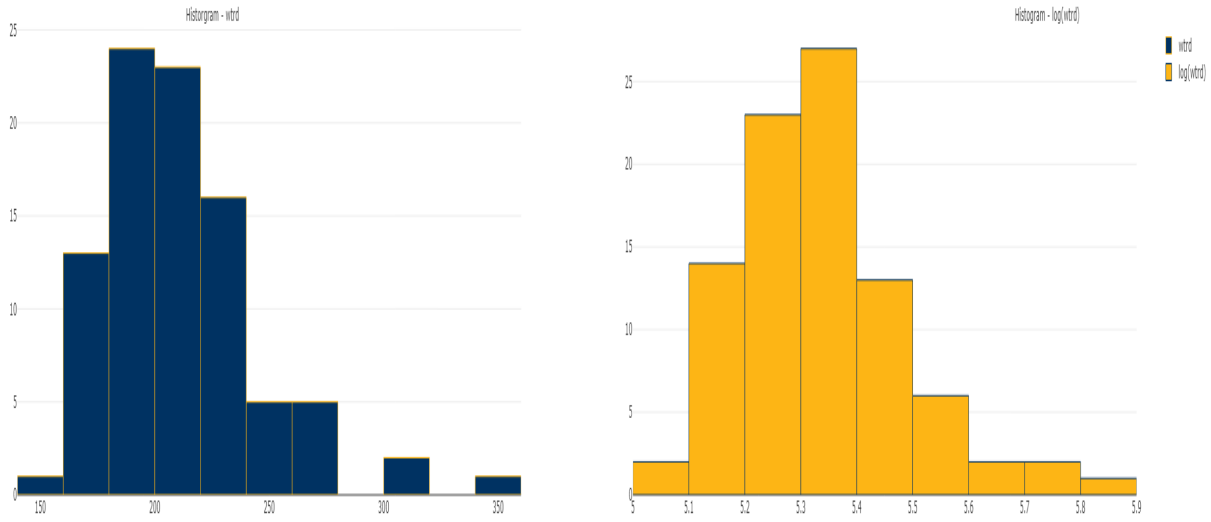


(b) Histogram of WSTA

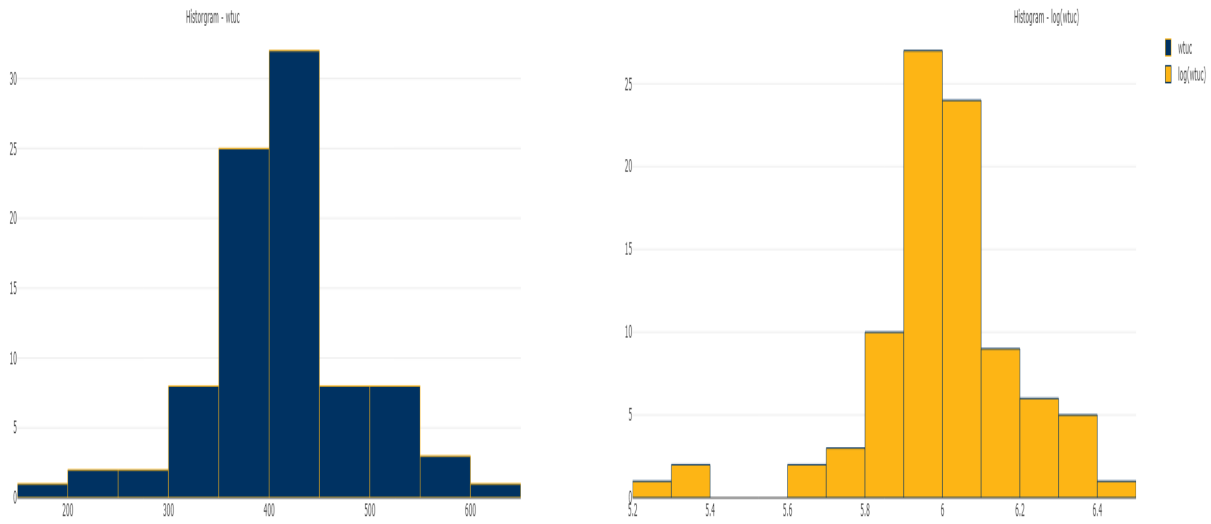
Figure 2.21: EDA : Distribution of Variables WSER and WSTA

EXPLORATORY DATA ANALYSIS - CONTD.

Histograms Continued.



(a) Histogram of WTRD



(b) Histogram of WTUC

Figure 2.22: EDA : Distribution of Variables WTRD and WTUC

2.7 Correlation

We assess the correlation between dependent and independent variables, excluding all binary and identifier attributes. Intersections where there exists a low statistical significance are removed from the plot to improve visibility of significant relationships. There appears to be the strongest Pearson r correlation of the dependent variable $crmrte$ with potential regressors $density$ (0.73), $wfed$ (0.49), and $polpc$ (0.48).

Superficially, the relationship between $polpc$ and $crmrte$ seems as though it might be causal in the opposite direction - that is, the more crime present, the more police the county hires. In this sense, $polpc$ might be better thought of as a dependent variable.

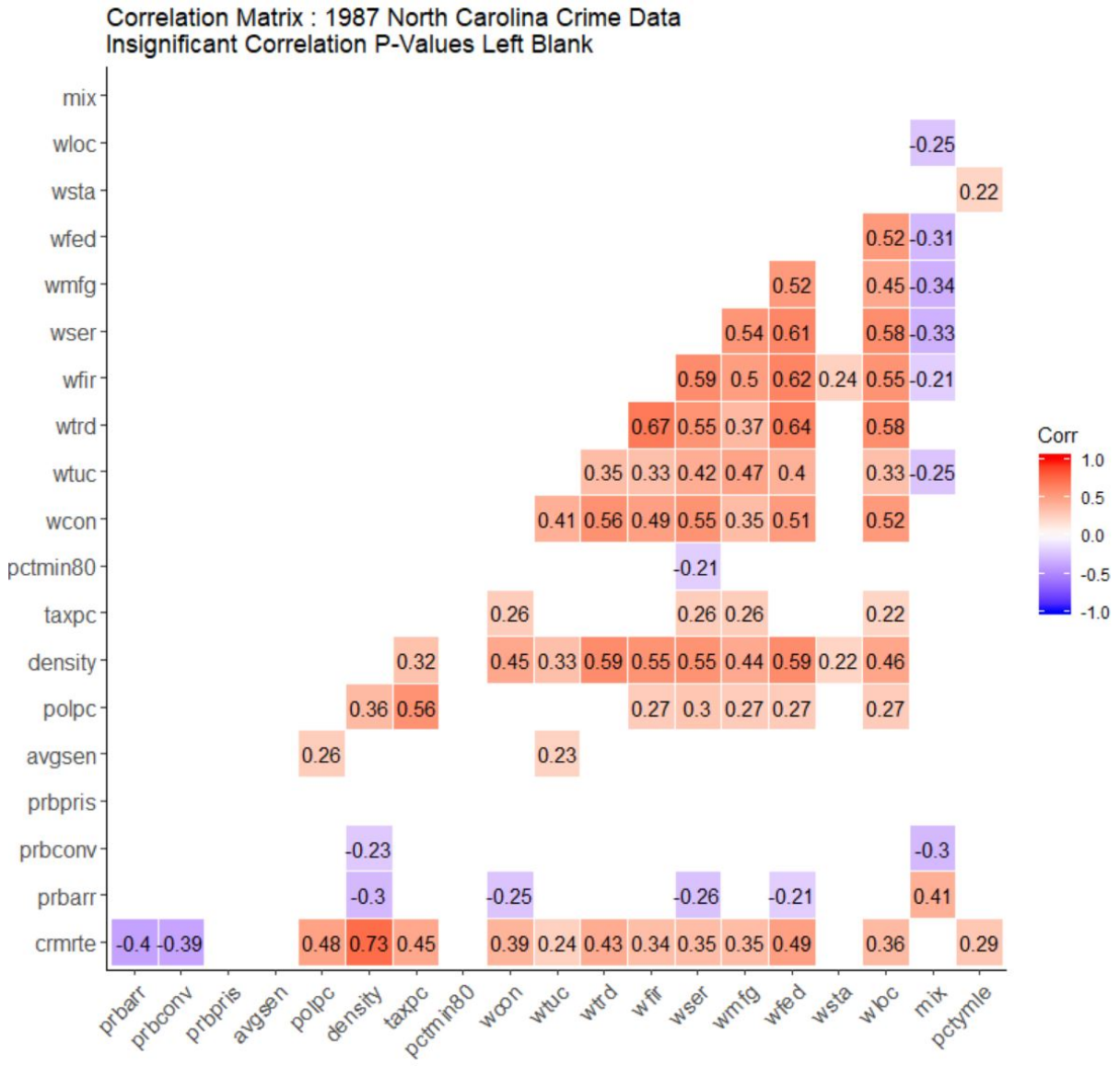


Figure 2.23: EDA : Correlation of Independent and Dependent Variables

EXPLORATORY DATA ANALYSIS - CONTD.

We also assess the correlation between log transforms of the variables of interest to assess the impact. Feature intersections where there exists a low statistical significance are also removed from the plot for better visibility. Generally, the impact is minimal to correlation between log and non-log transformed variables.

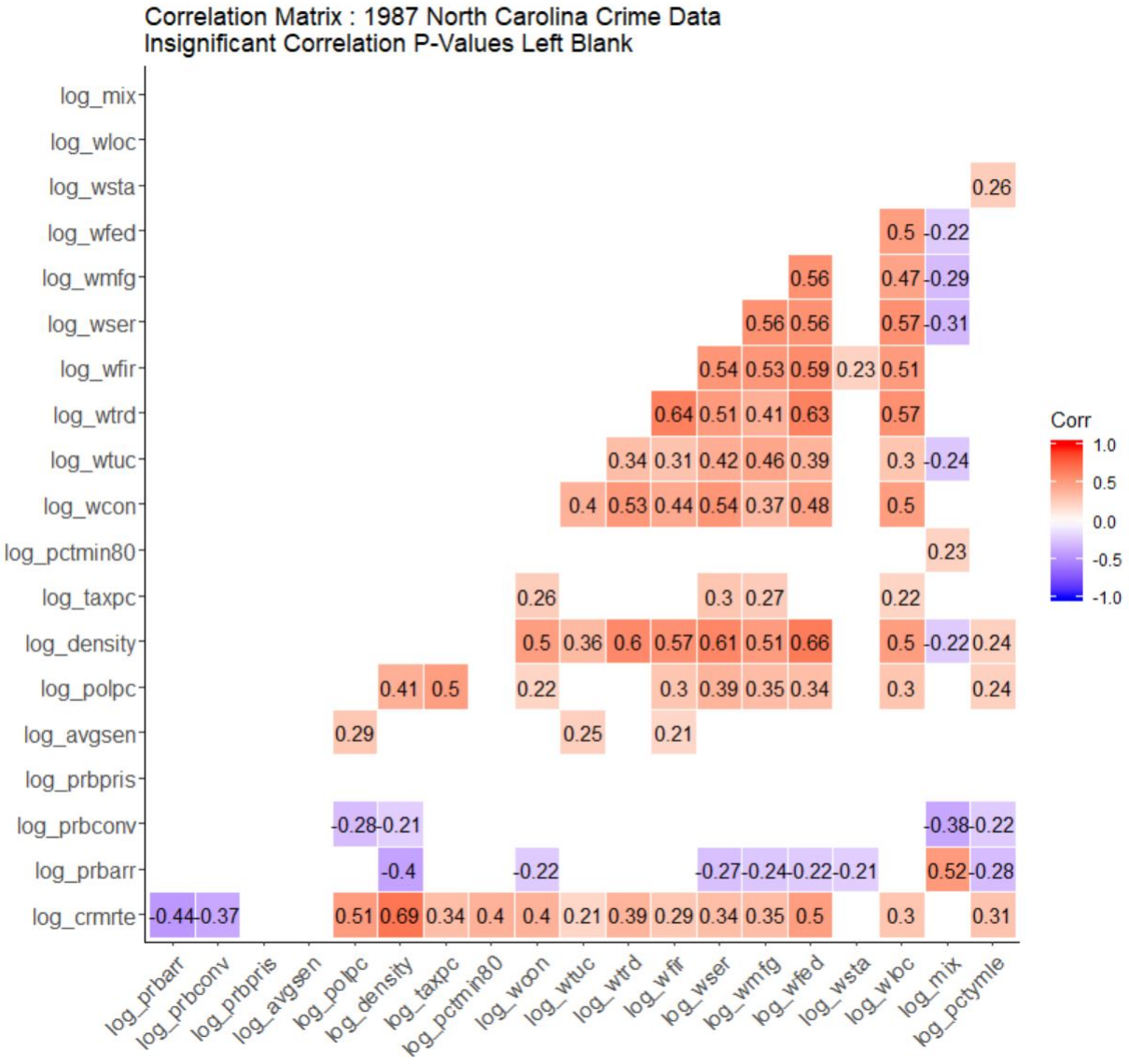


Figure 2.24: EDA : Correlation of Independent and Dependent Variables

Section 3

Analysis & Models

3.1 Analysis of Variables

There are a number of open questions regarding the determinants of crime. Why do people commit crime? What socioeconomic, demographic, and other factors are associated with increased crime rates? More importantly, what levers does the government have to reduce crime, both in the short and long term? For this campaign research, we are particularly interested in what policy recommendations we can make that can help reduce crime at the local and state government level.

From a public policy perspective, the levers we can most use to address crime rate are criminal justice related. The probability of arrest, conviction, and prison sentence are all deterrents than can help reduce the crime rate. Prison sentence length also serves as a deterrent. Having a larger and more robust police presence is also likely to deter crime; however as mentioned earlier this may be harder to determine as there is potentially a reverse causality relationship that areas with higher crime will then have a higher police presence. With that context, we are particularly interested in the following variables:

- PRBARR : 'probability' of arrest.
- PRBCONV : 'probability' of conviction.
- PRBPRIS : 'probability' of prison sentence.
- AVGSEN : average sentence in days.
- POLPC : police per capita.

ANALYSIS CONTD.

There are a myriad of other socioeconomic and societal factors that influence crime rates, but are less under the control of local government. Some of the observations here could have less applicable policy recommendations. They can range from the trite (i.e. "we should reduce poverty to then reduce crime rates") to potentially challenging (i.e. "we should reduce the number of young males or minorities in the state to reduce crime"). That said, the following factors could help strengthen the casual effect of the levers that we do have more control over. Many of the variables are collinear and we will ultimately not want to include each of these variables. In particular, the wage variables between industries are highly correlated.

- DEMOGRAPHIC :
  - DENSITY 100 people per square mile.
  - PCTMIN80 percent minority, circa 1980.
  - PCTYMLE percent of young males.
  - TAXPC tax revenue per capita.
  - URBAN =1 in SMSA.
- ECONOMIC :
  - WTUC wkly wge, trns, util, commun.
  - WTRD wkly wge, whlesle, retail trade.
  - WFIR wkly wge, fin, ins, real estx1.
  - WSER wkly wge, service industry.
  - WMFG wmfg wkly wge, manufacturing.
  - WFED wkly wge, fed employees.
  - WSTA wkly wge, state employees.
  - WLOC wkly wge, local gov emps.

Though it is not strictly necessary, we elected to apply a log transformation for each variable as it has the benefits of best visualization of distriubtion and of making the results easier to explain. By taking a log transformation of the independent and dependent variables, we can make more direct apples-to-apples comparisons and say a 10% increase in X leads to a 10% increase in Y.

For the rest of the analysis, we have imputed outliers for which there is evidence of entry or data collection error (each of which are individually documented in the Outlier Analysis section). We removed observations only under the condition that the row was duplicate or contained missing values.

3.2 Models

3.2.1 Naive Model (Model 0)

One model with only the explanatory variables of key interest (possibly transformed, as determined by your EDA), and no other covariates.

We elected to generate a naive, univariate linear model to establish a baseline for model development (technically, not required). Based on our EDA, the *density* variable offers the highest Pearson correlation with the dependent variable. Additionally, we elect to leverage the logarithm transformation for both dependent and predictor variable based on Histogram EDA results providing a smoother, normal-like distribution for both variables.

Our naive model suggests a strong linear relationship between the density and crime rate. Increasing density by 1% increases the crime rate by 0.48%. The model has an R-squared of 47.9%. This result makes sense as there is a known relationship that denser, more urban areas have higher crime rates. Reviewing the residual plots, this appears to satisfy all 6 CLM assumptions. There appears to be some heteroscedasticity as the residuals are slightly wider towards the left, but it is not concerning. Looking at the residuals vs leverage plot, these same residuals are low leverage points and do not skew the regression.

Table 3.1: Model : Naive

	Dependent variable:
	log(crmrte)
log(density)	0.486^{***} (0.380, 0.592)
Constant	-3.550^{***} (-3.632, -3.467)
Observations	90
R^2	0.479
Adjusted R^2	0.473
Residual Std. Error	0.398 (df = 88)
F Statistic	80.837^{***} (df = 1; 88)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3.2: Model : RESET test and Breusch-Pagan test p-values (Naive)

RESET (power=2)	Breusch-pagan
0.636	0.017

MODEL 0 - NAIVE MODEL CONTD.

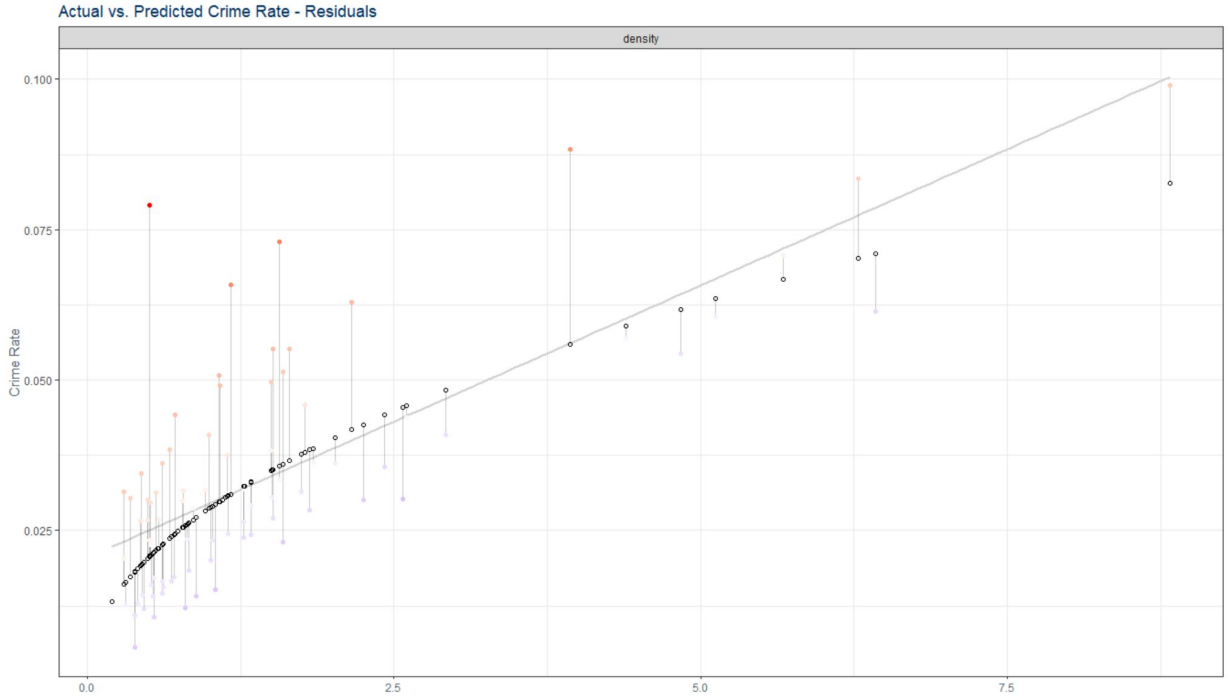


Figure 3.1: Model : Prediction Error, Naive Model

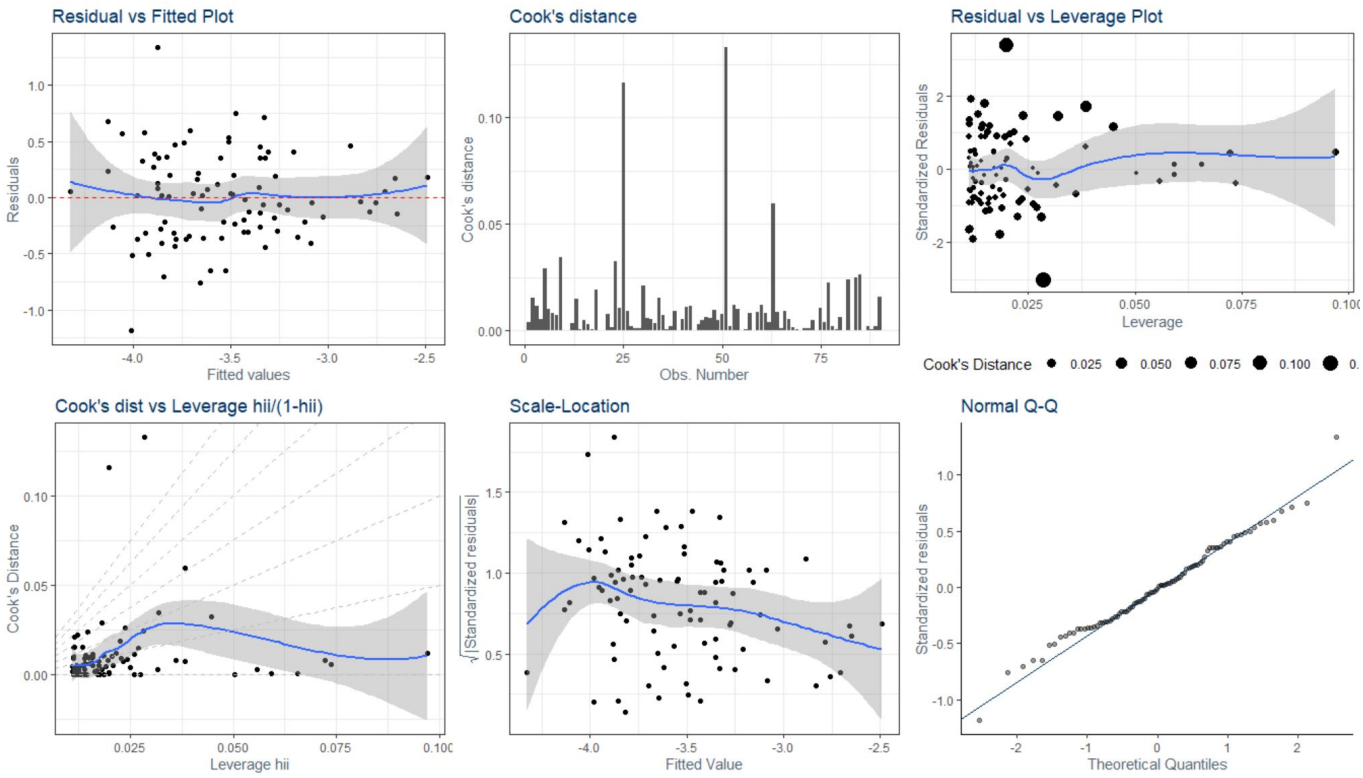


Figure 3.2: Model : Diagnostics Plots, Naive Model

3.2.2 Manually Tuned Model (Model 1)

One model with only the explanatory variables of key interest (possibly transformed, as determined by your EDA), and no other covariates.

Our manually tuned model was built from reasoning through EDA variables and common-sense understanding of the indicators of crime; this resulted in selection of four key variables of interest: $log(density)$, $log(pctmin80)$, $prbconv$, and $log(polpc)$, which resulted in an adjusted R^2 of **0.760**:

Table 3.3: Model : Manually Tuned

	Dependent variable:
	log(crmrte)
log(density)	0.363^{***} (0.283, 0.442)
log(pctmin80)	0.226^{***} (0.168, 0.284)
prbconv	-0.433^{***} (−0.596, −0.270)
log(polpc)	0.439^{***} (0.249, 0.628)
Constant	-1.126^{*} (−2.357, 0.104)
Observations	90
R^2	0.771
Adjusted R^2	0.760
Residual Std. Error	0.269 (df = 85)
F Statistic	71.538^{***} (df = 4; 85)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3.4: Model : RESET test and Breusch-Pagan test p-values (Manually Tuned)

RESET (power=2)	Breusch.pagan
0.346	0.0001

Table 3.5: Model : VIF Scores (Manually Tuned)

log_density	log_pctmin80	prbconv	log_polpc
1.241	1.001	1.065	1.208

MODEL 1 - MANUALLY TUNED CONTD.

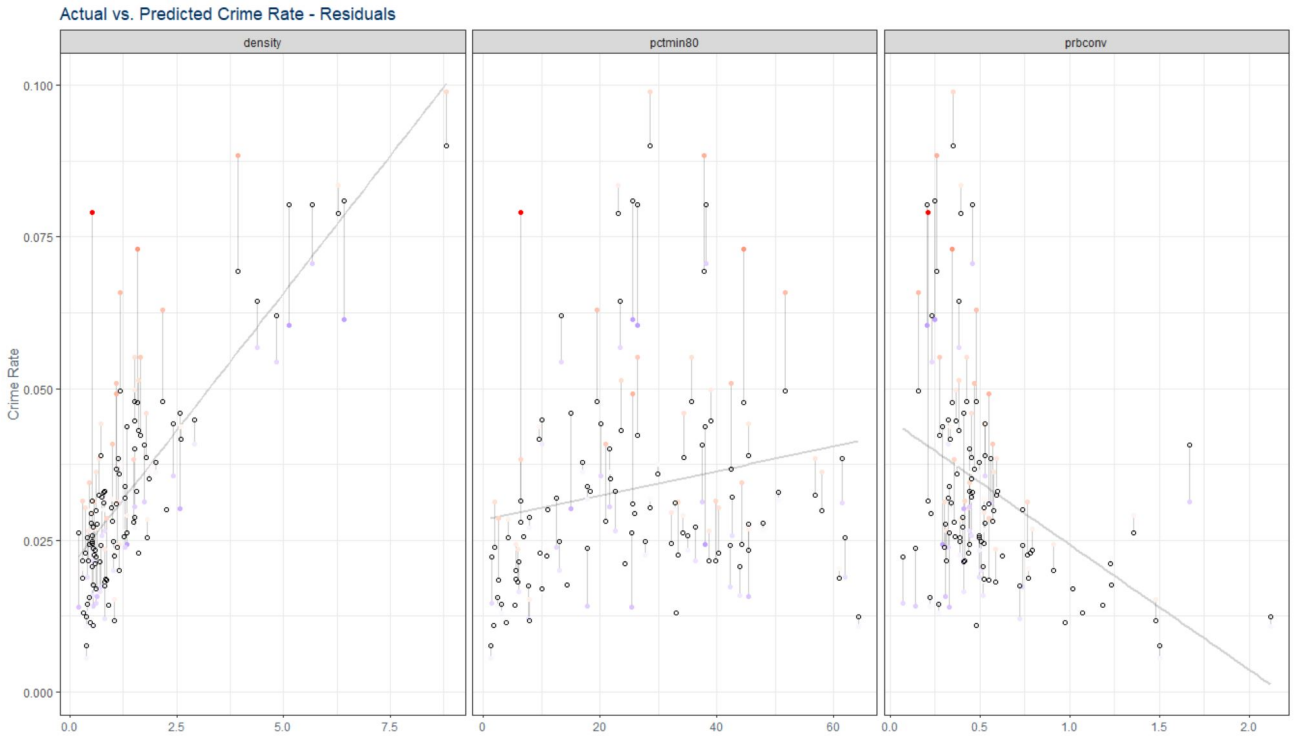


Figure 3.3: Model : Prediction Error, Manually Tuned Model

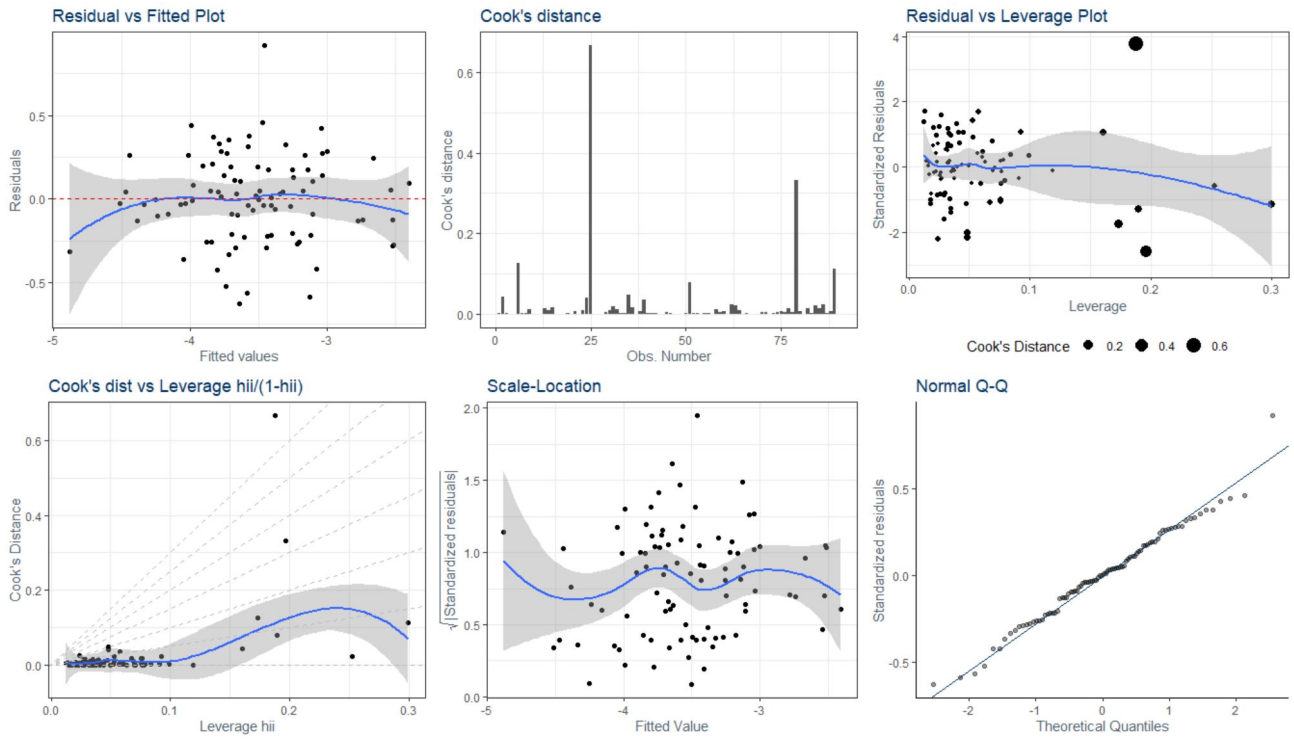


Figure 3.4: Model : Diagnostic Plots, Manually Tuned Model

MODEL 1 - MANUALLY TUNED CONTD.

Model 1 suggests a strong linear relationship between the density (0.36%), percent minority (0.27%), police per capita (0.44%), and crime rate. There is a negative relationship between probability of conviction and the crime rate (increase probability of conviction by 1 lowers the crime rate by 0.43%) The model has an R-squared of 77.1% and adjusted R-squared of 76.0%.

Adding additional variables suggests that there is a less strong relationship between density and the crime rate, which makes sense as there is a medium correlation between the additional variables and there is clear intuition for why each factor impacts the crime rate. Reviewing the residual plots, this model also appears to satisfy all 6 CLM assumptions. There appears to be more heteroscedasticity than in the niave model as the residuals are slightly wider towards the middle than the ends. The Breusch-Pagan test confirms there is heteroscedasticity. We correct for this by using a heteroskedasticity consistent (HC) variance covariance to compute the standard errors.

CLM assumption analysis for *MANUALLY TUNED MODEL*:

- CLM 1 - Linear in parameters : As the linear model is constructed in such a way that that the parameters are linear with error term u , we assess the Linear Assumption as affirmed by definition.

- CLM 2 - Random sampling : From the EDA : North Carolina Crime by County, 1987 graph in our EDA section, we note that North Carolina has 100 counties, of which we have data for 90; therefore, we have nearly the entire population available for our analysis. The dispersion of missing counties does not appear to fall along any conceivable pattern, save for a cluster of missing counties near the Eastern border that are geographically connected. Regardless, we have a sufficient percentage of the population from which we can draw reliable statistical inference and no reason to believe that the sample taken is biased or violates the tenants of IID sampling.

- CLM 3 - No perfect multi-colinearity : From Model : VIF Scores (Manually Tuned), we can review the Variance inflation factor scores for each coefficient to evaluate whether there exists a degree of multi-colinearity worth worrying over. Typically, scores above 4-5 are signal for concern - we see from our results that we do not have significant multi-colinearity in this model.

- CLM 4 - Zero-conditional mean : To meet this condition, we expect the error term u to be ≈ 0 for all variables, such that $E(u|x_1, x_2, ..., x_n) = 0$. We can verify this condition by reviewing the **Residual vs Fitted Plot** (see 3.4) and looking for an approximate straight loess line. Unfortunately, our line is slightly upward-convex, but is being pulled down from the zero-line by only a very few data points. In order to correct this, we will need to capture more of the variation in the model by adding appropriate variables currently omitted.

- CLM 5 - Homoskedasticity : Ideally, variance of the error term u in our model remains uniform across all fitted values. We can assess compliance with this condition via review of the loess line in the **Scale-Location** plot (see 3.4), or by executing a Breusch-pagan test and evaluating the p-values. From our plot, we see an oscillating pattern for variance, implying that our variance is not sufficiently uniform. From our BP test (see 3.4), we receive a p-value of 0.0001 which implies we can easily reject the null hypothesis $H_0 : Homoskedasticity$.

- CLM 6 - Normality : The assumption here is that the error population is independent [of the regressors] and that the error term u is normally distribution with $\mu = 0$. We can review this expectation in the **Normal Q-Q** plot (see 3.4); here, we see some slight back-and-forth on the plot, indicating the presence of kurtosis and a possible multi-modal distribution. In general, our error term is likely non-normally distributed; however, since our sample size is 90 we benefit from asymptotics and the assurance that our coefficients are approximately normal.

3.2.3 Best Fit Model (Model 2)

One model that includes key explanatory variables and only covariates that you believe increase the accuracy of your results without introducing substantial bias (for example, you should not include outcome variables that will absorb some of the causal effect you are interested in). This model should strike a balance between accuracy and parsimony and reflect your best understanding of the determinants of crime.

After achieving an R^2 of 0.76 from our best-educated manual selection, we employed the use of regSubsets to assist in feature selection that would perform well and avoid over-fitting. Our resulting model incorporated eight of the input features, including a mix of native and log-transformed values, and achieved an adjusted R^2 of **0.855**:

Table 3.6: Model : Best Fit

	Dependent variable:
	log(crmrte)
prbconv	-1.053^{***} $(-1.339, -0.766)$
pctmin80	0.014^{***} $(0.011, 0.016)$
wfir	-0.001^{**} $(-0.002, -0.0002)$
log(prbarr)	-0.426^{***} $(-0.541, -0.311)$
log(polpc)	0.464^{***} $(0.307, 0.621)$
log(density)	0.363^{***} $(0.290, 0.436)$
log(taxpc)	-1.322^{***} $(-1.901, -0.743)$
log(wsta)	-0.397^{**} $(-0.742, -0.052)$
Constant	5.698^{***} $(2.803, 8.594)$
Observations	90
R^2	0.869
Adjusted R^2	0.855
Residual Std. Error	0.184 (df = 79)
F Statistic	71.006^{***} (df = 10; 79)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3.7: Model : RESET test and Breusch-Pagan test p-values (Best Fit)

RESET (power=2)	Breusch-pagan
0.00004	0.00002

Table 3.8: Model : VIF Scores (Best Fit)

prbconv	pctmin80	wfir	log_prbarr	log_polpc	log_density	log_taxpc	log_wsta
1.318	1.070	1.654	1.426	1.570	2.176	1.392	1.182

MODEL 2 - BEST FIT CONTD.

Our feature selection algorithm selection was made using the lowest BIC score, which had a slightly lower overall R^2 value, but provides us with a more robust model:

(a) Input Features by BIC score *(b) Adjusted R-Squared by BIC score*

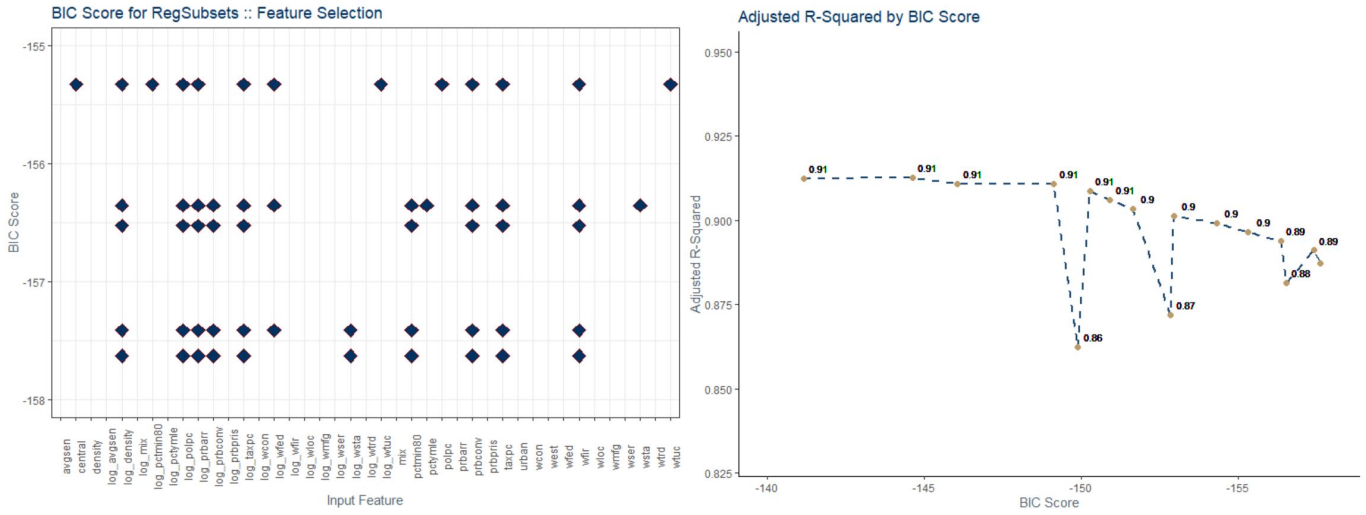


Figure 3.5: Model : Feature Selection, Best Fit Model

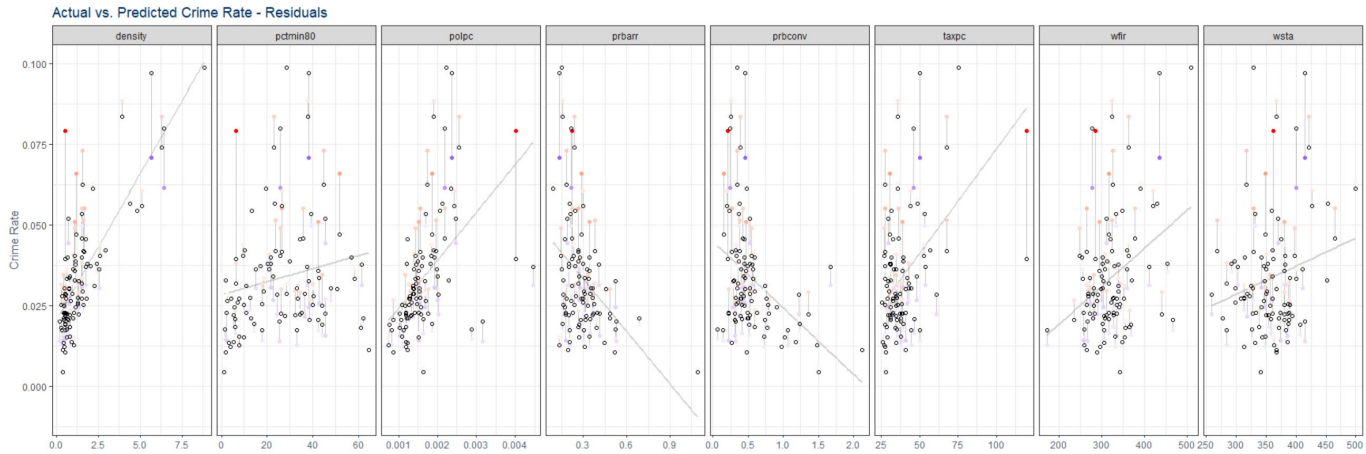


Figure 3.6: Model : Prediction Error, Best Fit Model

MODEL 2 - BEST FIT CONTD.

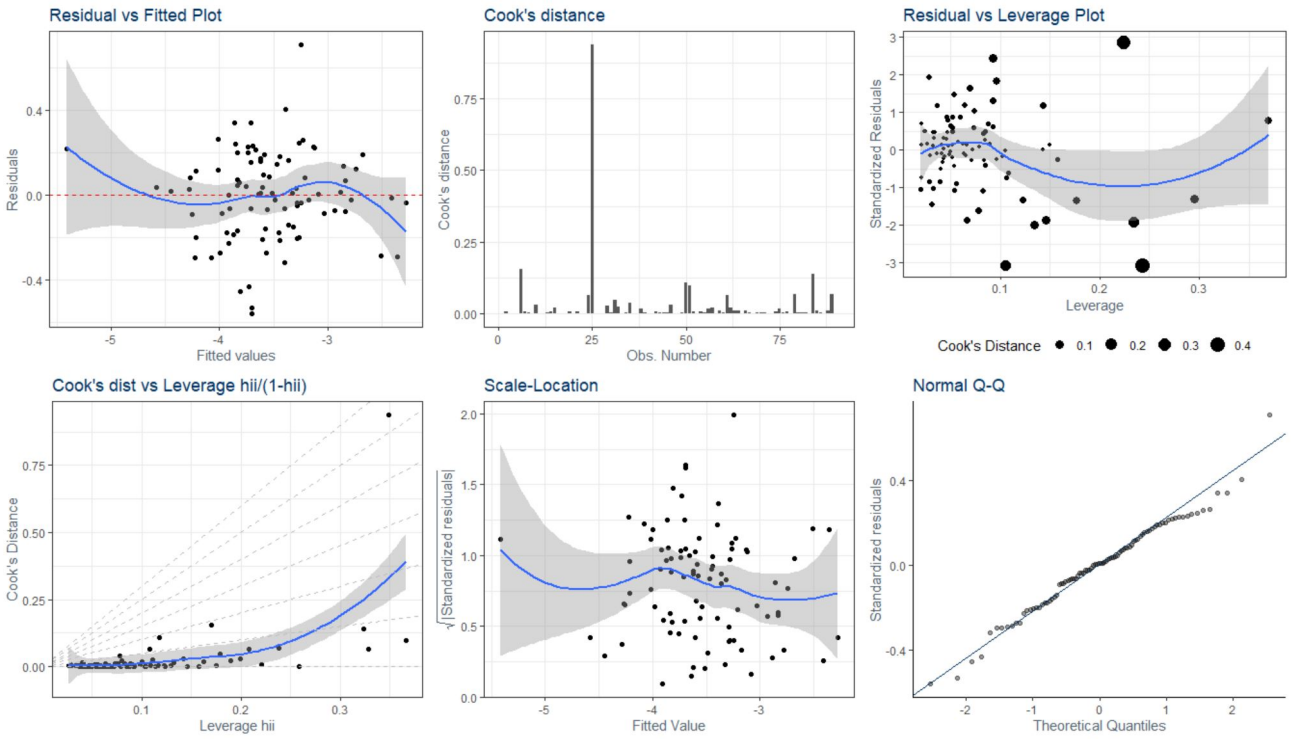


Figure 3.7: Model : Diagnostic Plots, Best Fit Model

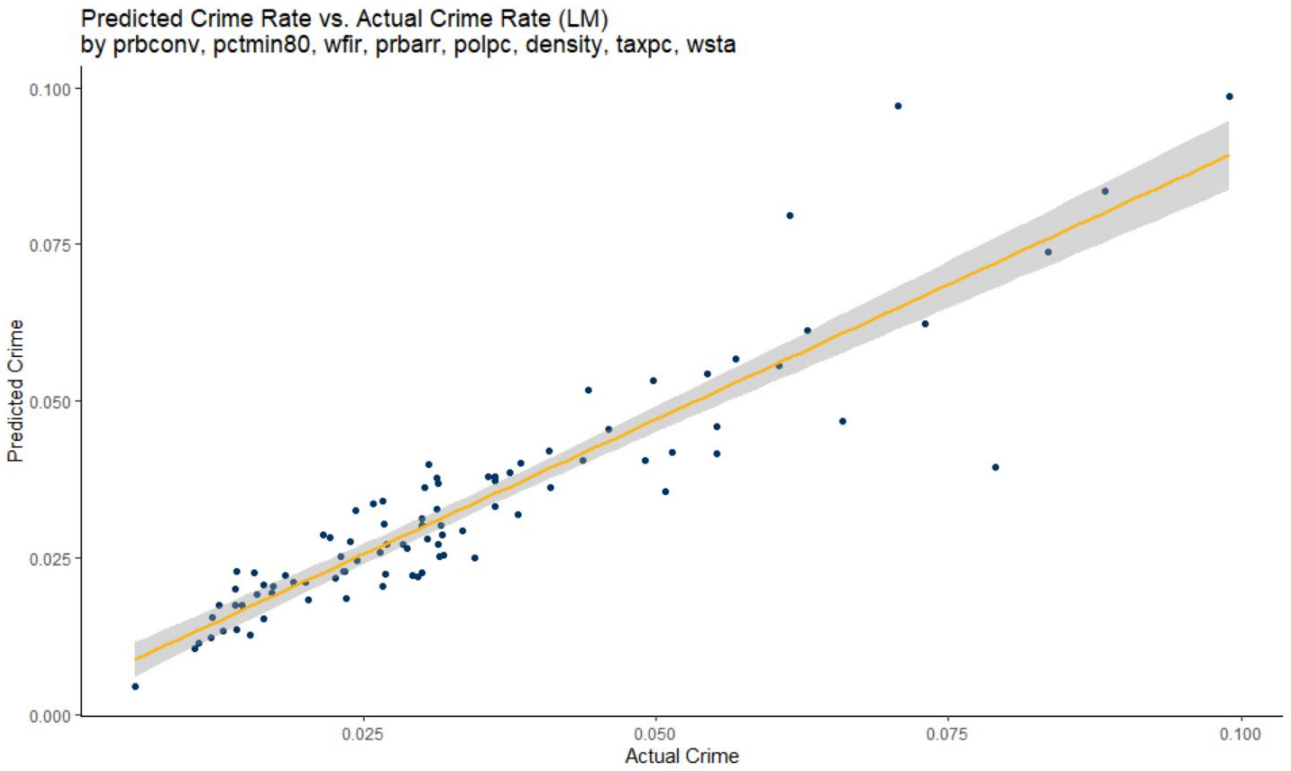


Figure 3.8: Model : Predicted vs. Actual, Best Fit Model

MODEL 2 - BEST FIT CONTD.

CLM assumption analysis for *BEST-FIT MODEL*:

- CLM 1 - LINEAR IN PARAMETERS : *Same as CLM 1 for Manually Tuned Model (see Manually Tuned Model (Model 1)).*
- CLM 2 - RANDOM SAMPLING : *Same as CLM 2 for Manually Tuned Model (see Manually Tuned Model (Model 1)).*
- CLM 3 - NO PERFECT MULTI-COLINEARITY : From Model : VIF Scores (Best Fit), we can review the Variance inflation factor scores for each coefficient to evaluate whether there exists a degree of multi-colinearity worth worrying over. Typically, scores above 4-5 are signal for concern - we see from our results that we do not have significant multi-colinearity in this model.
- CLM 4 - ZERO-CONDITIONAL MEAN : To meet this condition, we expect the error term u to be ≈ 0 for all variables, such that $E(u|x_1, x_2, ..., x_n) = 0$. We can verify this condition by reviewing the **Residual vs Fitted Plot** (see 3.7) and looking for an approximate straight loess line. Unfortunately, our line is S-curved, but is being pulled away from the zero-line by only a very few data points. In order to correct this, we will need to capture more of the variation in the model by adding appropriate variables currently omitted and, potentially, eliminate one or more noise variables.
- CLM 5 - HOMOSKEDASTICITY : Ideally, variance of the error term u in our model remains uniform across all fitted values. We can assess compliance with this condition via review of the loess line in the **Scale-Location** plot (see 3.7), or by executing a Breusch-pagan test and evaluating the p-values. From our plot, we see an oscillating pattern for variance, implying that our variance is not sufficiently uniform. From our BP test (see 3.7), we receive a p-value of 0.00002 which implies we can easily reject the null hypothesis $H_0 : Homoskedasticity$.
- CLM 6 - NORMALITY : The assumption here is that the error population is independent [of the regressors] and that the error term u is normally distribution with $\mu = 0$. We can review this expectation in the **Normal Q-Q** plot (see 3.7); here, we see some slight back-and-forth on the plot, indicating the presence of kurtosis and a possible multi-modal distribution. In general, our error term is likely non-normally distributed; however, since our sample size is 90 we benefit from asymptotics and the assurance that our coefficients are approximately normal.

3.2.4 Overfit Model (Model 3)

One model that includes the previous covariates, and most, if not all, other covariates. A key purpose of this model is to demonstrate the robustness of your results to model specification.

For our final model, we ask stepAIC to consider all input parameters, including those that are perfectly multi-colinear, and generate the best performing model. Here, we achieve a deceiving adjusted R^2 of **0.913** but critical CLM assumptions are likely violated in the process.

Table 3.9: Model : Overfit

	Dependent variable:
	log(crmrte)
prbconv	-0.786^{***} $(-1.086, -0.487)$
polpc	-344.709^{**} $(-619.648, -69.770)$
taxpc	0.027^{***} $(0.015, 0.040)$
central	-0.135^{***} $(-0.219, -0.051)$
pctmin80	0.007^{**} $(0.001, 0.013)$
wtuc	-0.003^{**} $(-0.005, -0.001)$
wfir	-0.001^{***} $(-0.002, -0.0004)$
wser	0.009^{**} $(0.001, 0.017)$
wmfg	-0.004^{***} $(-0.006, -0.001)$
log(prbarr)	-0.419^{***} $(-0.529, -0.309)$
log(prbconv)	0.181^{*} $(-0.013, 0.375)$
log(polpc)	1.089^{***} $(0.573, 1.606)$
log(density)	0.310^{***} $(0.233, 0.387)$
log(taxpc)	-1.195^{***} $(-1.800, -0.591)$
log(pctmin80)	0.111^{*} $(-0.007, 0.228)$
log(wcon)	0.280^{*} $(-0.008, 0.567)$
log(wtuc)	1.161^{**} $(0.228, 2.094)$
log(wser)	-2.715^{***} $(-4.690, -0.739)$
log(wmfg)	1.421^{***} $(0.483, 2.359)$
log(wsta)	-0.348^{*} $(-0.690, -0.006)$
log(wloc)	0.479^{*} $(-0.072, 1.029)$
Constant	4.905 (−5.987, 15.796)
Observations	90
R^2	0.933
Adjusted R^2	0.913
Residual Std. Error	0.162 (df = 68)
F Statistic	45.280^{***} (df = 21; 68)
Note:	*p<0.1; **p<0.05; ***p<0.01

MODEL 3 - OVERFIT CONTD.

Table 3.10: Model : RESET test and Breusch-Pagan test p-values (Overfit)

RESET (power=2)	Breusch.pagan
0.942	0.800

Table 3.11: Model : VIF Scores (Overfit)

prbconv	polpc	taxpc	central	pctmin80	wtuc	wfir
9.899	24.550	22.879	1.489	10.560	31.481	2.409

Table 3.12: Model : VIF Scores Contd. 1/2 (Overfit)

wser	wmfg	log_prbarr	log_prbconv	log_polpc	log_density	log_taxpc
108.478	43.072	1.731	10.309	24.688	3.182	22.632

Table 3.13: Model : VIF Scores Contd. 2/2 (Overfit)

log_pctmin80	log_wcon	log_wtuc	log_wser	log_wmfg	log_wsta	log_wloc
11.199	1.960	31.227	107.342	45.781	1.506	2.161

MODEL 3 - OVERFIT CONTD.

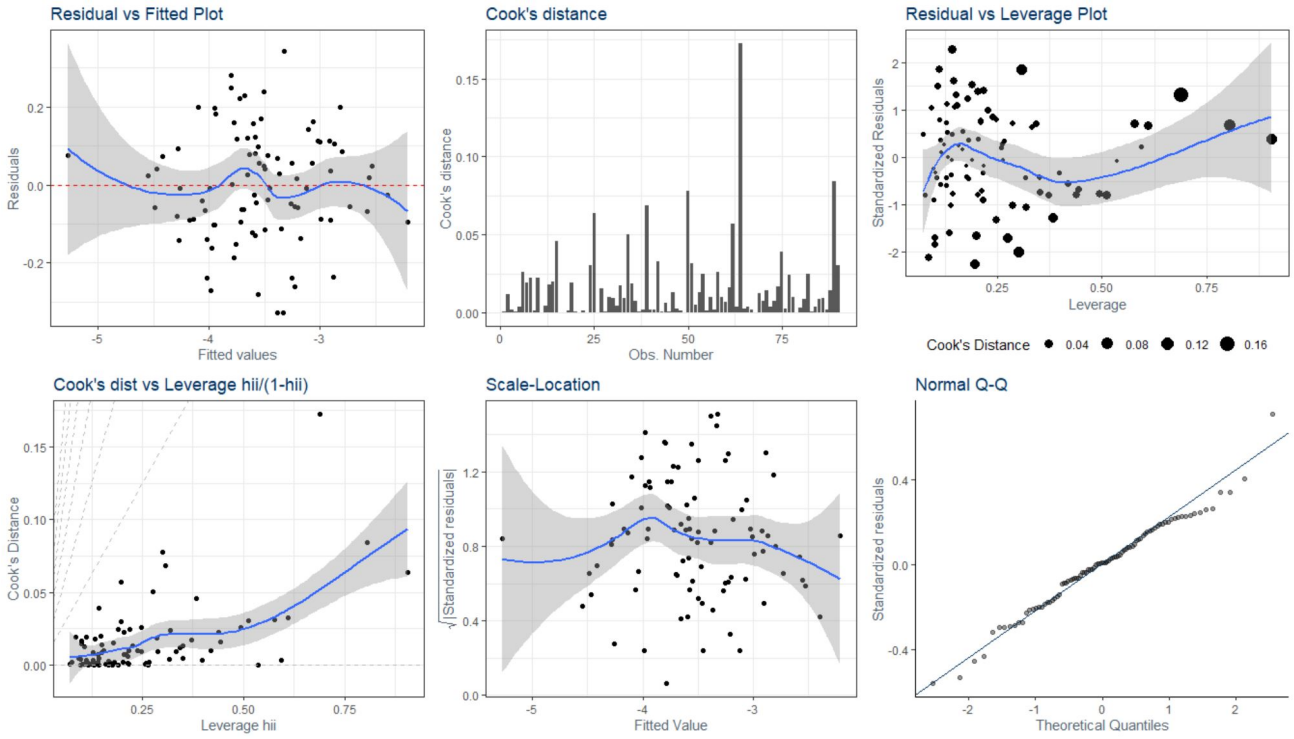


Figure 3.9: Model : Diagnostic Plots, Overfit Model

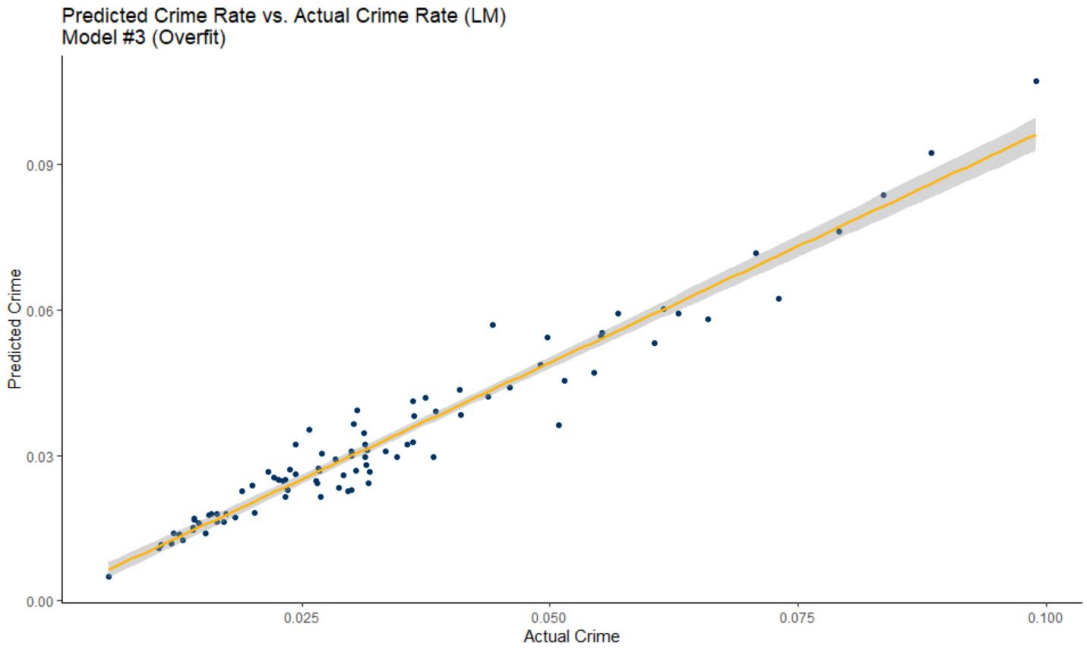


Figure 3.10: Model : Predicted vs. Actual, Overfit Model

MODEL 3 - OVERFIT CONTD.

CLM assumption analysis for *OVERFIT MODEL*:

- CLM 1 - LINEAR IN PARAMETERS : *Same as CLM 1 for Manually Tuned Model (see Manually Tuned Model (Model 1)).*
- CLM 2 - RANDOM SAMPLING : *Same as CLM 2 for Manually Tuned Model (see Manually Tuned Model (Model 1)).*
- CLM 3 - NO PERFECT MULTI-COLINEARITY : From **??**, we can review the Variance inflation factor scores for each coefficient to evaluate whether there exists a degree of multi-colinearity worth worrying over. Typically, scores above 4-5 are signal for concern - we see from our results that we highly significant multi-colinearity problems with the 'Overfit' model due primarily to the overlap between natural and log-transformed variables (i.e. wser + log(wser)).
- CLM 4 - ZERO-CONDITIONAL MEAN : To meet this condition, we expect the error term u to be ≈ 0 for all variables, such that $E(u|x_1, x_2, ..., x_n) = 0$. We can verify this condition by reviewing the **Residual vs Fitted Plot** (see 3.9) and looking for an approximate straight loess line. Unfortunately, our line is S-curved, but is being pulled away from the zero-line by only a very few data points. In order to correct this, we will need to capture more of the variation in the model by adding appropriate variables currently omitted and, potentially, eliminate one or more noise variables.
- CLM 5 - HOMOSKEDASTICITY : Ideally, variance of the error term u in our model remains uniform across all fitted values. We can assess compliance with this condition via review of the loess line in the **Scale-Location** plot (see 3.9), or by executing a Breusch-pagan test and evaluating the p-values. From our plot, we see an oscillating pattern for variance, implying that our variance may or may not be sufficiently uniform. From our BP test (see **??**), we receive a p-value of 0.8 which implies we cannot reject the null hypothesis $H_0 : Homoskedasticity$.
- CLM 6 - NORMALITY : The assumption here is that the error population is independent [of the regressors] and that the error term u is normally distribution with $\mu = 0$. We can review this expectation in the **Normal Q-Q** plot (see 3.9); here, we see some slight back-and-forth on the plot, indicating the presence of kurtosis and a possible multi-modal distribution. In general, our error term is likely non-normally distributed; however, since our sample size is 90 we benefit from asymptotics and the assurance that our coefficients are approximately normal.

3.3 Omitted Variables

In regression analysis we will encounter the problem of omitted variable bias because either (a) we ignored other determinants of the dependent variable that correlate with the explanatory variables, or (b) the data was not available for analysis. This problem will result in a bias effect of one or more regressors. There are 2 conditions must hold for an omitted variable bias to exist:

1. The omitted variable must be correlated with the dependent variable.
2. he omitted variable must be correlated with one or more other explanatory, independent variables.

There are a number of variables that contribute to crime rate that are omitted from this report. Below are a subset of variables that may impact the regression output (with the caveat that each variable may be difficult to measure)

- POVERTY RATE : average wage is typically correlated with poverty, but it does not fully capture the nuance that the poverty rate captures. There can be two counties with equal average wages, but a much higher poverty rate if there is higher inequality. Higher poverty is correlated with higher crime rates so would have a positive effect on the crime rate. It is inversely correlated with wages, so by not having it in the regression it decreases the effect size for wage variables..
- DRUG USE : drug use is highly correlated with crime. From a causal factor, drug use increases crime both through organized crime of selling drugs and the associated violence as well as an increase in crime for drug users who want access to drugs. This is likely inversely correlated with wages, so by not having it in the regression it decreases the effect size for wage variables.
- TRUST IN POLICE/INSTITUTIONS : this a broad category of items that are likely highly correlated so we would only need one of these variables. People who have higher trust in police, government, and institutions are less likely to commit crime. Increased trust leads to a lower crime rate. Trust is likely associated with wages. By not including trust, this increases the effect size for wages.
- EDUCATION LEVELS : education is inversely correlated with the crime rate (notwithstanding the potential increase in white-collar crime). This is correlated with wages, By not including education, it increases the effect size for wages.
- CITIZEN'S ATTITUDE TOWARDS CRIME : If people are more accepting of crime (i.e. don't view certain crimes as unethical/immoral) then they are more likely to commit crime. This variable is likely correlated with wages, but has unclear regional and density correlations. By not including attitudes towards crime, this increases the effect size for wages.
- CRIME RATE REPORTED : partly due to trust in police, different jurisdictions will report different levels crime. The probability of arrest captures this data, but that is based on the percentage of crimes actually reported. If certain counties have the same population level of crime, but one has much higher levels of reported crime, it will skew the results compared to the true population model. It is unclear how this would impact each variable, but likely has a large effect if there is a large variation between counties.

3.4 Summary

In this study, we try to understand how demagogical, social, and economic factors impact the crime rate by analyzing the crime statistics for a selection of counties in North Carolina from the calendar year 1987. We first address outliers and apparent mistakes due to human or data collection error. Further, we log transform some crucial variables to make them easier to interpret. After cleaning up the data, one naive and three linear regression models, increasing in complexity, are developed and evaluated.

We found that Best Fit Model (Model 2), the model including explanatory variables selected by our knowledge in criminology, is the best fit of the data even though Overfit Model (Model 3), the model suggested by R package stepAIC, has a higher adjusted R^2 and thus explains more of the variance. Overfit Model (Model 3) is not our choice of regression model primarily due to over-fitting the data; the model has 21 variables, but there are only 90 observations in the data! Lastly, we discuss the Omitted Variables bias and suggest what factors that researchers can look into in their future studies.

In conclusion, we consider our key research questions as follow:

What are the best signals for predicting a change in crime rate?

From our Best Fit Model (Model 2), we learned that the ideal independent variables for predicting crime are probability of conviction, percent minority demographic, weekly wage in finance, probability of arrest rate, police per capita rate, population density rate, tax rate, and weekly wage of state employees rate.

What policy considerations should be focused on to reduce crime?

We believe a higher probability of conviction will significantly affect a reduction in the crime rate. Improving the quality of police work, forensic analysis, and legal follow-through of the state are the best platform positions for addressing crime, at least in North Carolina.

SUMMARY CONTD.

Table 3.14: Model : Comparison of Naive, Manually Tuned, and Best Fit

	Dependent variable:		
	log(crmrte)		
	Model 0	Model 1	Model 2
log(density)	0.486***	0.363***	0.310***
	(0.054)	(0.041)	(0.039)
log(taxpc)			−1.195***
			(0.309)
log(pctmin80)		0.226***	0.111*
		(0.030)	(0.060)
log(wcon)			0.280*
			(0.147)
log(wtuc)			1.161**
			(0.476)
log(wser)			−2.715***
			(1.008)
log(wmfg)			1.421***
			(0.479)
log(wsta)			−0.348*
			(0.174)
log(wloc)			0.479*
			(0.281)
prbconv		−0.433***	−0.786***
		(0.083)	(0.153)
polpc			−344.709**
			(140.278)
taxpc			0.027***
			(0.006)
central			−0.135***
			(0.043)
pctmin80			0.007**
			(0.003)
wtuc			−0.003**
			(0.001)
wfir			−0.001***
			(0.0005)
wser			0.009**
			(0.004)
wmfg			−0.004***
			(0.001)
log(prbarr)			−0.419***
			(0.056)
log(prbconv)			0.181*
			(0.099)
log(polpc)		0.439***	1.089***
		(0.097)	(0.263)
Constant	−3.550***	−1.126*	4.905
	(0.042)	(0.628)	(5.557)
Observations	90	90	90
R^2	0.479	0.771	0.933
Adjusted R^2	0.473	0.760	0.913
Residual Std. Error	0.398 (df = 88)	0.269 (df = 85)	0.162 (df = 68)
F Statistic	80.837*** (df = 1; 88)	71.538*** (df = 4; 85)	45.280*** (df = 21; 68)
Note:			*p<0.1; **p<0.05; ***p<0.01

Bibliography

Baltagi, B. H. (2006). Estimating an economic model of crime using panel data from north carolina. *Journal of Applied Econometrics*, *21*(4), 543–547. doi:10.1002/jae.861

Carolina, N. (2019). State code 15a, article 20 - arrest. Retrieved from https://www.ncleg.net/EnactedLegislation/Statutes/PDF/ByArticle/Chapter%5C_15A/Article%5C_20.pdf

Census, U. (2019). County intercensal tables 1980-1990. Retrieved from https://www.census.gov/data/tables/time-series/demo/popest/1980s-county.html

Cornwell, C., & Trumbull, W. N. (1994). Estimating the economic model of crime with panel data. *Review of Economics and Statistics*, *76*(2), 360–366. doi:10.2307/2109893

Wikipedia. (2019). List of counties in north carolina. Retrieved from https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina

Appendix

List of Figures

List of Tables