Understanding and Predicting Crime Rates Using Statistical Methods

Carlos Espino, Xavier Gonzalez, Diego Llarrull, Woojin Kim

December 15, 2015

Contents

1 Introduction

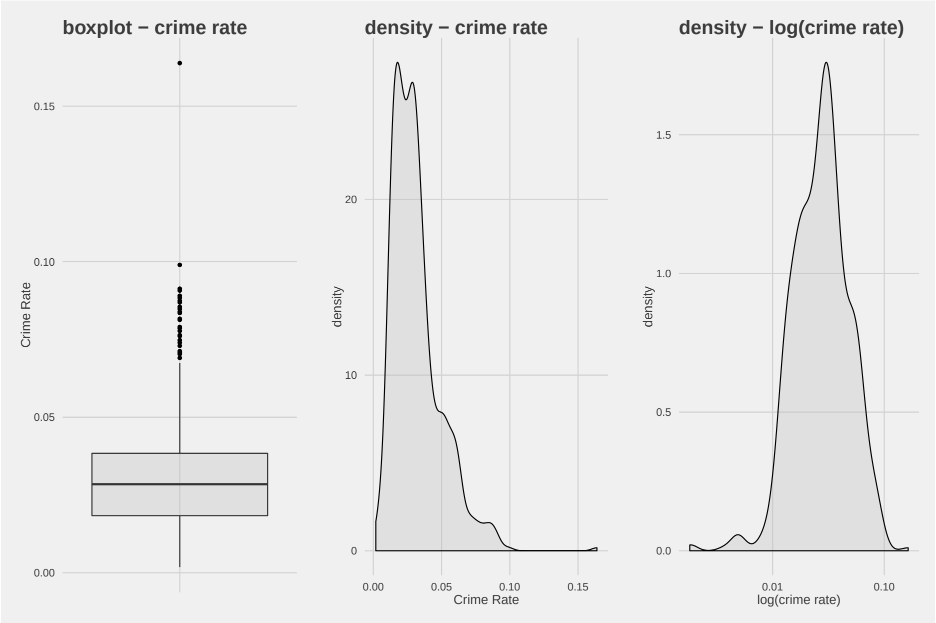
Understanding the factors behind criminal behaviour is one of the most crucial tasks for preventing and controlling future crime. In this report, we explore the potential factors affecting crime rates based on the demographics and econometrics data gathered from 197 counties in North Carolina from 1981 to 1987. Using various statistical methods and modeling techniques, we analyze and identify the most important factors and metrics tied to crime rates. We also present a predictive model capable of estimating the crime rate with under 20.4% error using the selected parameters.

2 Dataset

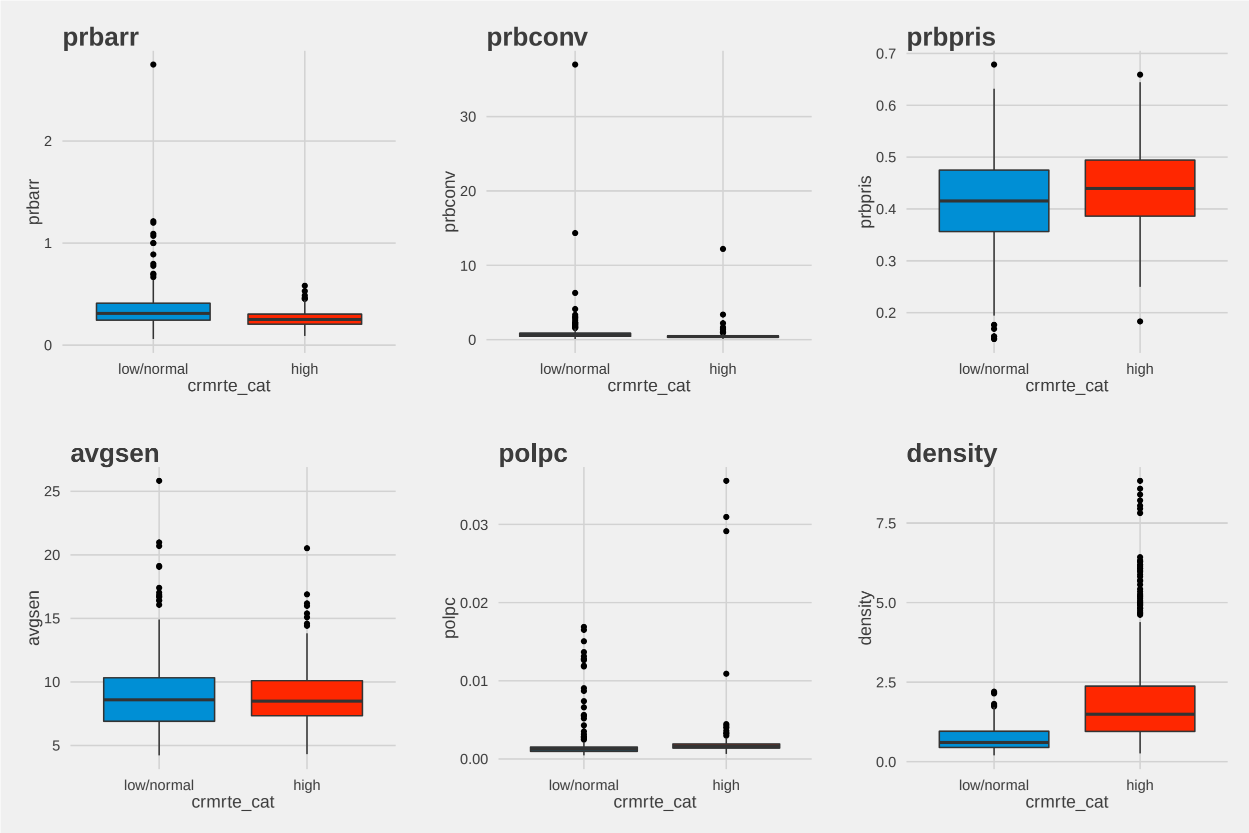
Predictor	Description
county	county identifier
year	year from 1981 to 1987
crmrte	crimes committed per person
prbarr	'probability' of arrest
prbconv	'probability' of conviction
prbpris	'probability' of prison sentence
avgsen	average sentence, days
polpc	police per capita
density	people per square mile
taxpc	tax revenue per capita
region	one of 'other', 'west' or 'central'
smsa	'yes' or 'no' if in SMSA
pctmin	percentage minority in 1980
wcon	weekly wage in construction
wtuc	weekly wage in trns, util, commun
wtrd	weekly wage in whole sales and retail trade
wfir	weekly wage in finance, insurance and real estate
wser	weekly wage in service industry
wmfg	weekly wage in manufacturing
wfed	weekly wage of federal employees
wsta	weekly wage of state employees
wloc	weekly wage of local governments employees mix offence mix: face-to-face/other
pctymle	percentage of young males

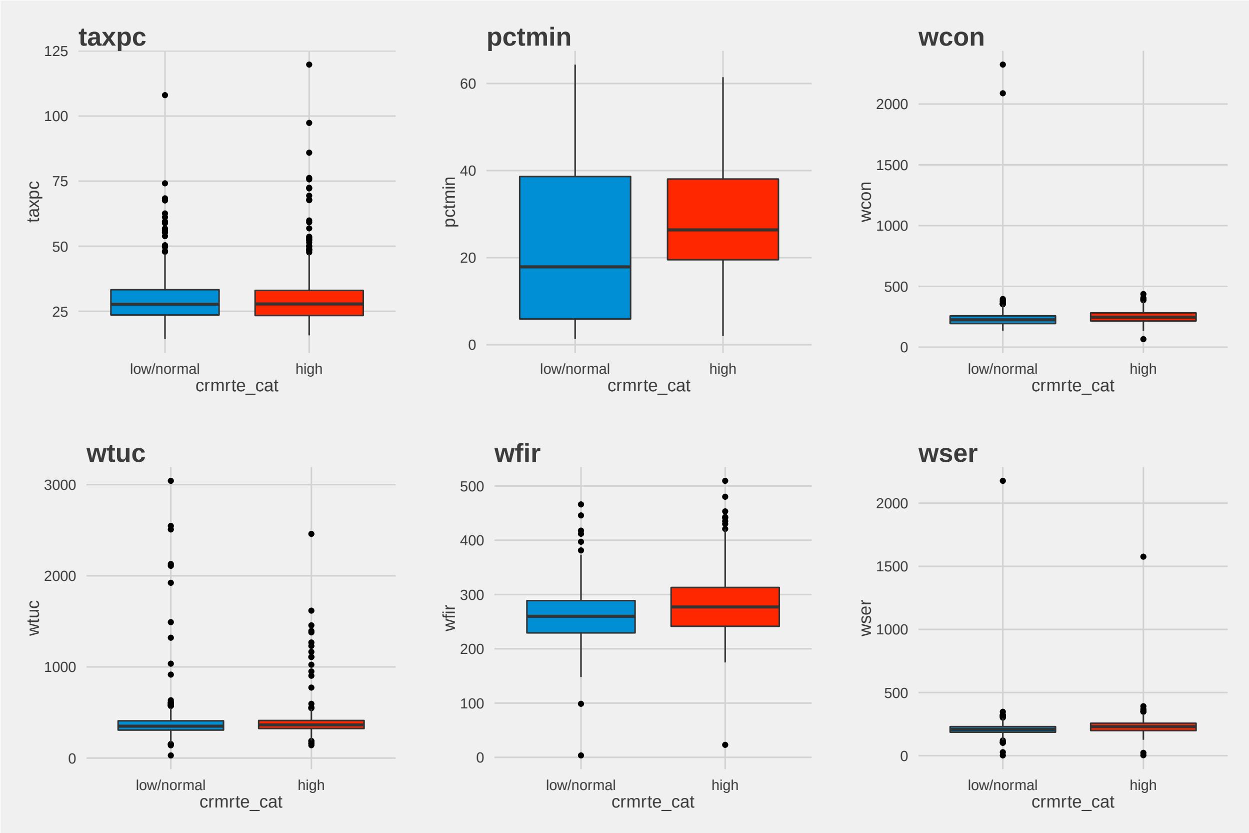
Table 1: Description of the predictors in the dataset

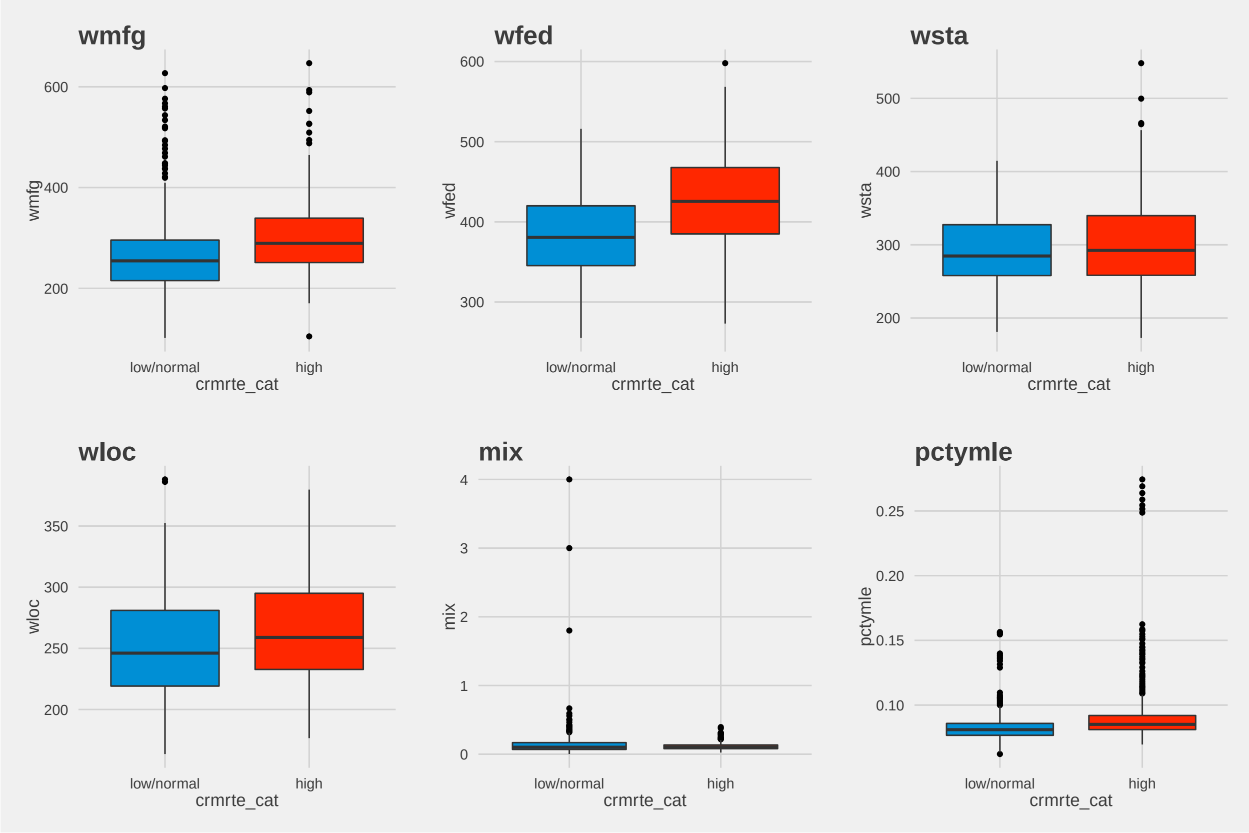
We analyzed the variables in the dataset starting with the target variable: `crmrte`, the crime rate. Along this study, we will use this variable in different forms. We define a categorical value equal to one representing high crime rate, when the value of the target variable is higher that its median value. We called this variable `crmrte_cat`. Also, we will use the natural logarithm of the variable to adequately transform it to apply different statisticals models to predict and describe the data. We assume that the target value depends on the other variables. The behaviour of the target is represented with a boxplot, a softened histogram of the variable, and a softened histogram of the logarithm of the variable.



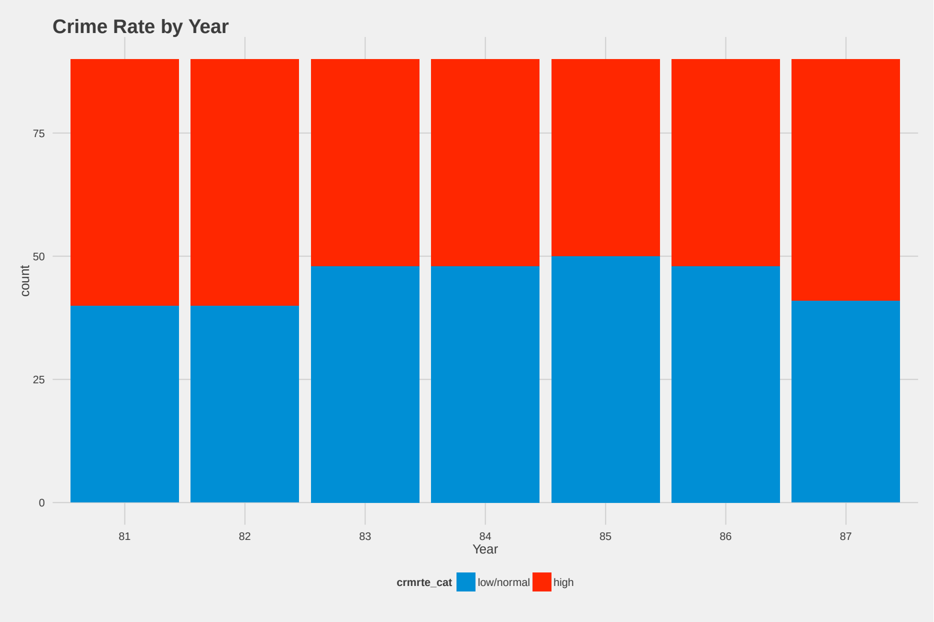
Besides the target variable, the dataset contains other 21 variables we used as predictors. Two of them have categorical values. The `region` variable can have 3 possible values: `other`, `west` or `central` and the `smsa` can have `yes` or `no`. The dataset also contains the `year` variable which can be considered as a time reference. A short description of each variable can be found in the table above. Next, we plot some charts to explore the behaviour of the variables and their relationships with the target.



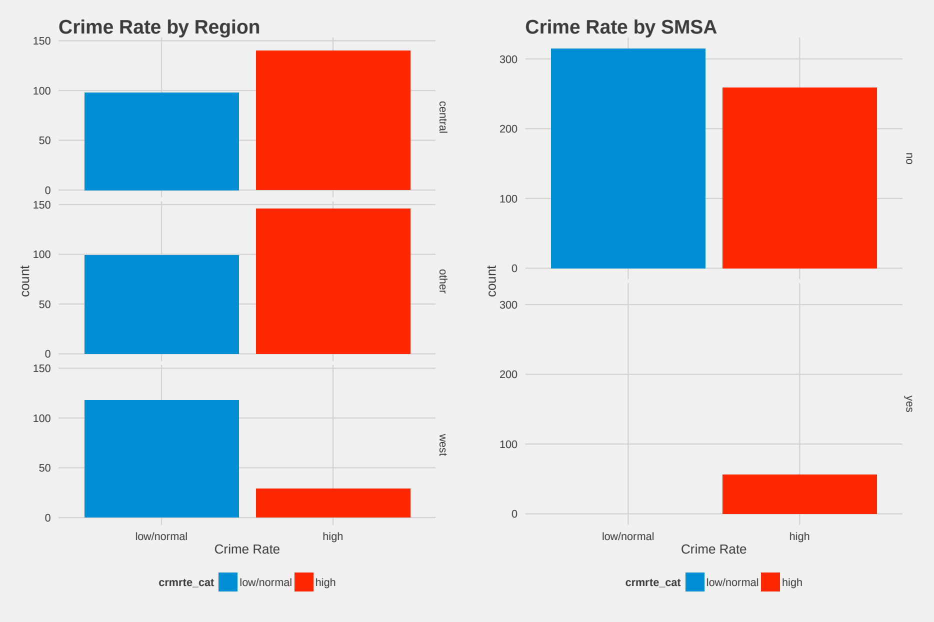




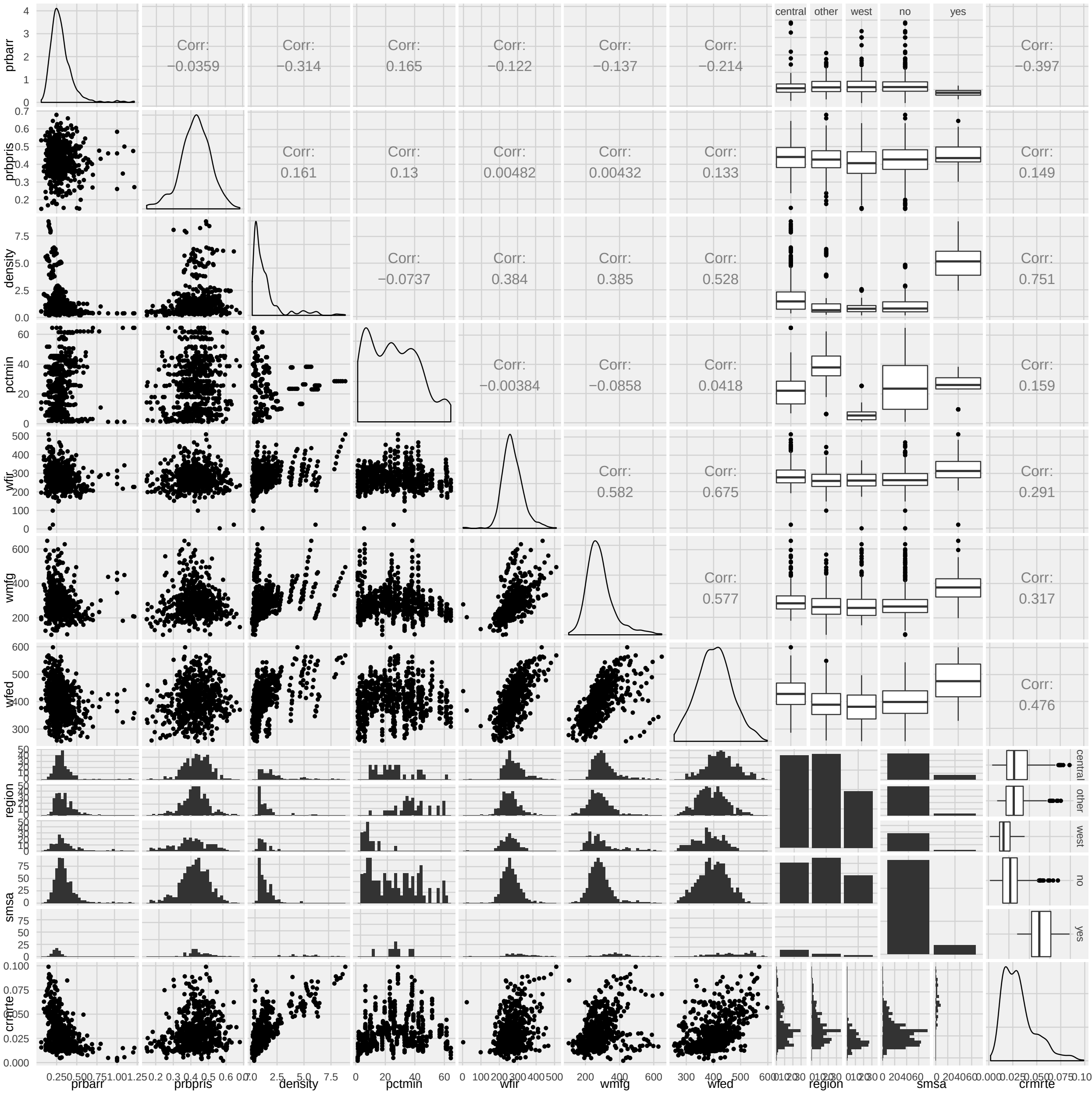
In the boxplots above, we can see that the variables that may have a predictive value with the target are variables prbarr, density, pctmin, wfed, wmfg and pctymle as they separate the population partially by the value of the defined target variable. We explore the rest of the predictors by tracing them on the following charts, starting with the variable year.



From the above plot, we notice that there is no significant trend on the crime rate along the timeline being considered. The other two variables with categorical values are region and smsa.



In these two charts above we see the crime rate decrease when the variable region takes the value west and when the smsa variable takes the value yes. Consequently, we continue to further explore the relationship between these two categorical variables and the target variable by implementing *ANOVA* in the next section, but first we analyze the variances and covariances between all predictors. We trace a paired graph with some selected variables in order to explore the correlation between the variables.

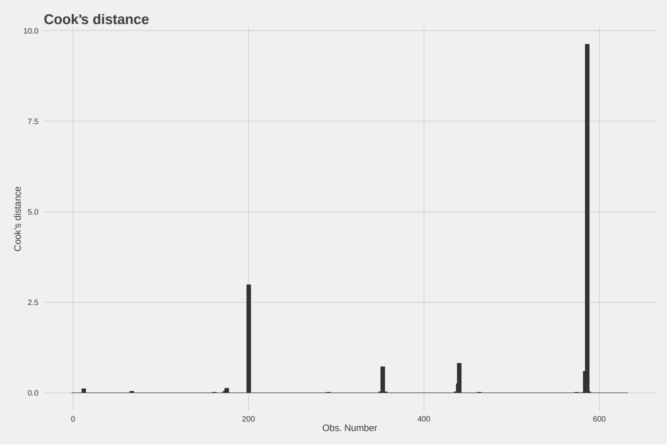


In the graphs above we show the correlation between the selected predictors and between the predictors and the target. The highest value of correlation is between the target variable and `density`. Other high values of correlation involve variables `wmfg`, `wfed` and `density`. We will later discuss whether these variables are significant for modeling.

3 Analysis

3.1 Influencial Observation Detection

Just in order to identify influencial points, we run a linear model with all the continous variables as predictors. Then, we calculated the cook distance of each observation and traced a plot. We detected the observations that have a cook distance value greater than 0.5. This treshold value was calculated as the average of two methods.



Clearly, there are 5 points that are highly influential. These points, showed in the table below, have a value greater than the treshold and they are consequently eliminated from the dataset. In a real context, this analysis would lead to a deeper reaserch about the reasons under this high leverages.

	Cooks.Dist
584	0.6037403
353	0.7179964
440	0.8217147
200	2.9935603
586	9.6199316

3.2 ANOVA models

In the first analysis, we model the mean of the target variable using a two-level factor. We aggregate all values of `region` (`west`, `central` and `other`) into `w` and `nw`, whether they take value equal to `west` or not. Running *ANOVA*, we obtain the following output:

```
##              Df  Sum Sq  Mean Sq F value Pr(>F)    
## region_w_nw   1 0.02541 0.025408   98.97 <2e-16 ***
## Residuals   623 0.15994 0.000257                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results show a very low $p-$value for the variable, which means that the model is accurate. The null hypothesis (i.e., means are equal for both regions) is rejected. Then, we compare the means of the crime rate between the `west` and other regions.

```
##         nw          w 
## 0.03494763 0.01987887
```

Considering the above analysis, we can assume that the model can correctly fit the value of the mean in each region: (`west`, `other`). The coefficients of the model can be extracted from the fit value retuned in the package.

```
##  (Intercept) region_w_nww 
##   0.03494763  -0.01506876
```

The model is given by

$$\mu_{crmrte} = 0.0347 - 0.0148 I_{\{region='w'\}}$$

Now, we repeat the same analysis considering two factors. We alse include the other categorical variable: `smsa`. We fit an *ANOVA* model and we get the following output:

```
##               Df  Sum Sq Mean Sq F value Pr(>F)    
## region_w_nw    1 0.02541 0.02541   154.4 <2e-16 ***
## smsa           1 0.05761 0.05761   350.2 <2e-16 ***
## Residuals    622 0.10233 0.00016                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, we obtain a good $p-$value for each of the two variables, which means that both factors have a strong relationship with the response variable. The null hypotesis (i.e., the means are equal) is rejected. The coefficients in this case are:

```
##  (Intercept) region_w_nww      smsayes 
##   0.03154125  -0.01329217   0.03399280
```

Besides the two categorical variables, we include in the analysis of the variance the interaction effect between the two variables.

```
##                    Df  Sum Sq Mean Sq F value  Pr(>F)    
## region_w_nw         1 0.02541 0.02541 156.310 < 2e-16 ***
## smsa                1 0.05761 0.05761 354.405 < 2e-16 ***
## region_w_nw:smsa    1 0.00139 0.00139   8.553 0.00357 ** 
## Residuals         621 0.10094 0.00016                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The $p-$values indicate that the two factors and the interaction between them are significant.

3.3 Confidence Interval for the Median

As discussed above, we defined a categorical target variable: `high`, if the value of the crime rate was higher that the median, and `low/normal` otherwise. This was done in order to be able to run models that require such variables, as we will see in the following sections. Therefore, it would be very useful to have confidence intervals about the median. Computing the median value, we get:

```
## [1] 0.0284567
```

We can simply obtain a confidence interval around that value. Considering the binomial distribution with $n = 626$ observations and a probability of 0.5, we want to obtain the $k - th$ observation that returns the 98%, the 95%, the 88% and the 81% of the probability by doing the following:

$$1 - 2 \times p_{binom}(k, n = 626, p = 0.5)$$

the $k - th$ values corresponding to those intervals are

```
## [1] 283 290 295 298
```

From the vector of sorted values for `crmrte`, we select the $k - th$ elements of the vector and the $n - k + 1$ elements corresponding to the four confidence intervals:

signif.%	low.level	up.level
0.99	0.0265877	0.0296451
0.95	0.0267532	0.0294232
0.88	0.0269621	0.0292244
0.81	0.0271005	0.0291268

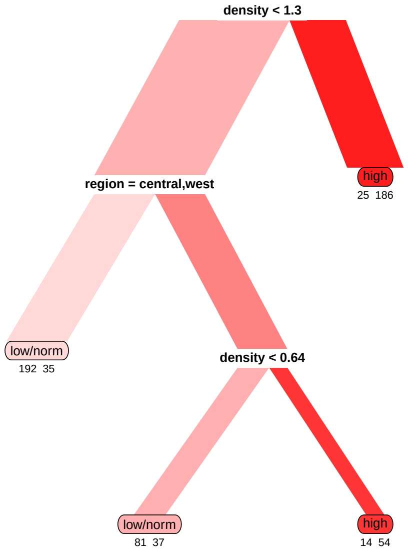
A more sofisticated method to obtain the confidence interval for the median is the *Wilcoxon Signed Rank Test*. As this test assumes symmetry of the variable's distribution, we applied it to the logarithm of the target variable, as discussed above. The results were then transformed to return the values to the original scale by applying the exponential function to the intervals obtained.

signif.%	low.level	up.level
0.99	0.0258289	0.0290536
0.95	0.0262102	0.0286561
0.88	0.0264442	0.0284015
0.81	0.0265839	0.0282435

The results are similar to the simpler sign test.

3.4 Dependency Analysis with Predictive Models

Continuing with the analysis we fitted a decision tree model. To do so, we considered the target variable in the categorical format. The purpose of this model is to further explore the data and understand which variables are relevant to the response.



```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction   low/normal high
##   low/normal        108   27
##   high               12   88
```

```
##
##                Accuracy : 0.834
##                  95% CI : (0.7802, 0.8792)
##     No Information Rate : 0.5106
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.667
##  Mcnemar's Test P-Value : 0.02497
##
##             Sensitivity : 0.9000
##             Specificity : 0.7652
##          Pos Pred Value : 0.8000
##          Neg Pred Value : 0.8800
##              Prevalence : 0.5106
##          Detection Rate : 0.4596
##    Detection Prevalence : 0.5745
##       Balanced Accuracy : 0.8326
##
##        'Positive' Class : low/normal
##
```

We get a testing accuracy of 83% and verify that the most relevant variables to the target are `region` and `density`.

3.5 Linear Analysis

Additionally, we considered a standard *linear regression* model involving all predictors, as an alternative means to view the significance of each predictor. Note that in this context, performing *k-fold cross-validation* or *bootstrapping* isn't necessary as we are only interested in significant predictors, hence we performed an ordinary 80/20 splitting of the data into a training and a testing sets. We then run a simple linear fit will all predictors, in order to analyse the significance levels of the parameters, provided that the linear test itself has a significant R^2 value.

```
##
## Call:
## lm(formula = crmrte ~ ., data = Crime_data[-test, ])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.037646 -0.005344 -0.000880  0.003851  0.068660
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.752e-02  3.364e-02   2.899 0.003916 **
## X           -1.541e-04  8.400e-05  -1.835 0.067134 .
## county       4.889e-04  2.635e-04   1.855 0.064163 .
## year        -1.311e-03  4.574e-04  -2.866 0.004338 **
## prbarr      -3.038e-02  3.030e-03 -10.027  < 2e-16 ***
## prbconv     -3.041e-03  4.090e-04  -7.436 4.83e-13 ***
## prbpris     -1.002e-04  5.289e-03  -0.019 0.984888
## avgsen      -6.643e-05  1.696e-04  -0.392 0.695478
## polpc        2.638e+00  1.839e-01  14.348  < 2e-16 ***
## density      7.363e-03  6.737e-04  10.929  < 2e-16 ***
## taxpc        6.050e-05  4.840e-05   1.250 0.211957    
## regionother  5.576e-03  1.242e-03   4.489 8.96e-06 ***
## regionwest  -1.634e-03  1.560e-03  -1.047 0.295497    
## smsayes     -3.922e-03  3.004e-03  -1.305 0.192390    
## pctmin       1.003e-04  4.184e-05   2.398 0.016880 *  
## wcon         5.094e-06  4.879e-06   1.044 0.297018    
## wtuc         2.210e-07  1.582e-06   0.140 0.888942    
## wtrd         2.656e-06  4.716e-06   0.563 0.573616    
## wfir        -1.865e-05  1.234e-05  -1.512 0.131201    
## wser        -1.752e-06  3.952e-06  -0.443 0.657605    
## wmfg         3.271e-06  7.280e-06   0.449 0.653385    
## wfed         4.331e-05  1.162e-05   3.727 0.000217 ***
## wsta         3.128e-06  1.245e-05   0.251 0.801661    
## wloc         3.819e-05  2.224e-05   1.717 0.086584 .  
## mix          1.676e-02  4.716e-03   3.553 0.000418 ***
## pctymle      7.692e-02  1.978e-02   3.888 0.000115 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.009484 on 479 degrees of freedom
## Multiple R-squared:  0.742,  Adjusted R-squared:  0.7286 
## F-statistic: 55.12 on 25 and 479 DF,  p-value: < 2.2e-16
```

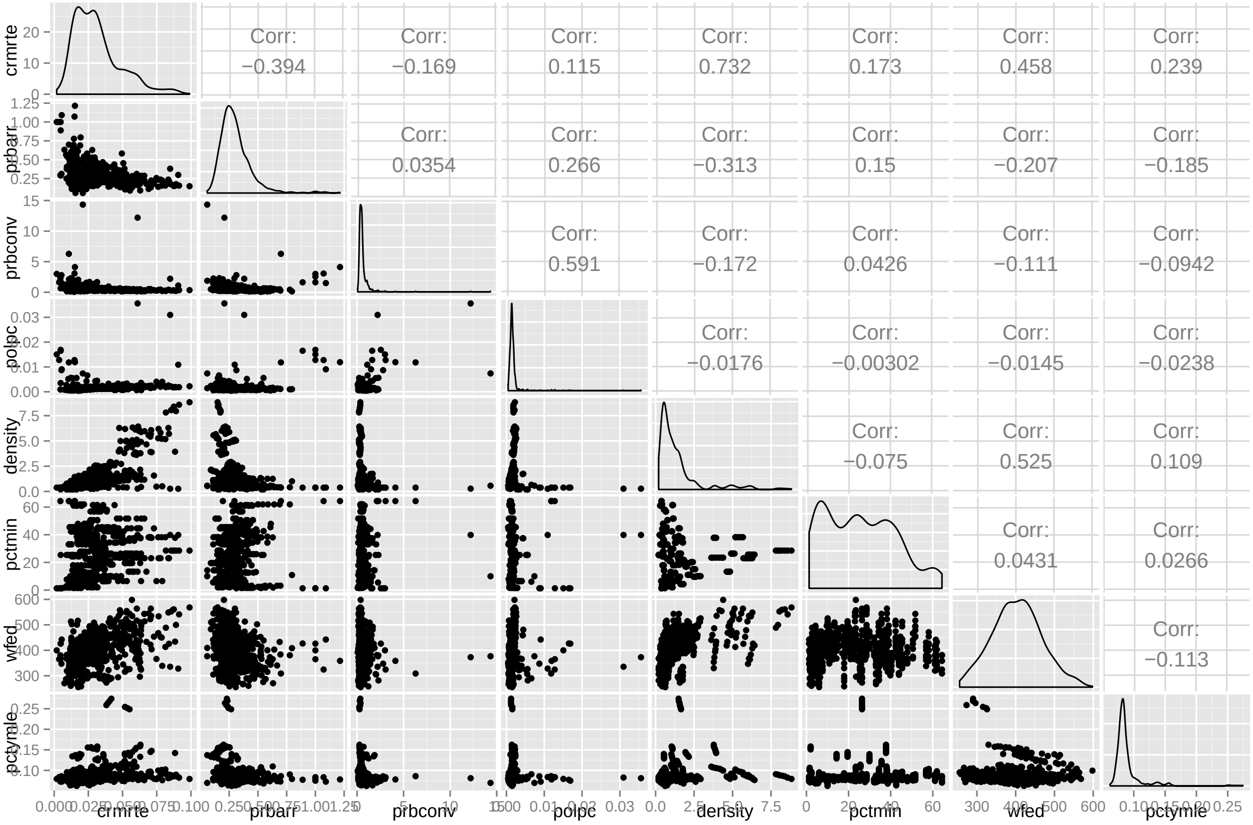
as the R^2 value is sufficiently high (0.7420498), we decided to perform *best subset selection* on the set of predictors. Although we are aware of the performance penalties of doing this for $p = 23$, the running times were considerably short and hence we decided to stick to this approach. Finally, after getting all best subsets with size $k = 1...p$, we analysed both *training* and *testing* errors by performing *k-fold cross validation* with $k = 10$ and then getting the minimum errors on all iterations.

The cross-validation estimate of the training error is 5.1449641×10^{-4} and the cross-validation error is 4.280511×10^{-4}. The actual training and test errors for this subset, on the original datasets, are 0.0012464 and 0.0014357, respectively. Both were obtained when using *best subset* with $k = 8$ predictors. The ratio between *testing* and *training* errors is (1.1519269). Consequently, we can conclude that the predictors yielded by the subset generated using *best subset selection* belong to a consistent model and, hence, can be used as a basis for non linear models. Nevertheless, we decided to run a linear fit with these predictors in order to check our conclusions:

```
## 
## Call:
## lm(formula = crmrte ~ prbarr + prbconv + polpc + density + as.factor(region) + 
##     pctmin + wfed + pctymle, data = Crime)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.023988 -0.005533 -0.000721  0.003915  0.050795 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             1.220e-02  3.593e-03   3.395 0.000729 ***
## prbarr                 -3.642e-02  3.080e-03 -11.823  < 2e-16 ***
## prbconv                -5.182e-03  5.463e-04  -9.486  < 2e-16 ***
## polpc                   2.630e+00  1.986e-01  13.243  < 2e-16 ***
## density                 6.867e-03  3.176e-04  21.623  < 2e-16 ***
## as.factor(region)other  4.487e-03  9.724e-04   4.614 4.80e-06 ***
## as.factor(region)west  -3.067e-03  1.150e-03  -2.667 0.007855 **
## pctmin                  1.443e-04  3.211e-05   4.494 8.36e-06 ***
## wfed                    2.199e-05  6.868e-06   3.202 0.001433 **
## pctymle                 6.235e-02  1.536e-02   4.058 5.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008779 on 615 degrees of freedom
## Multiple R-squared:  0.7442, Adjusted R-squared:  0.7405
## F-statistic: 198.8 on 9 and 615 DF,  p-value: < 2.2e-16
```

We can note that all coefficients are significant and the R^2, as expected, was reduced but only marginally (0.7442437 versus 0.7420498), which confirms that the model with this subset is indeed a good model.

Next, we proceeded to graphically analyse any nonlinearities between these predictors and the response, by looking at all pairwise plots:



It can be seen that the relationship between `crmrte` and `prbrarr`, `prbconv` and `polpc`, respectively, could be better explained by applying a *log* to these predictors. Additionally, `wfed` and `pctmin` seem to have a nonlinear relationship with the response, which makes them suitable as polynomial regression predictors. Consequently, we run a new, nonlinear model with these modified predictors:

```
##
## Call:
## lm(formula = crmrte ~ log(prbarr) + log(prbconv) + log(polpc) +
##     density + as.factor(region) + poly(pctmin, 4) + poly(wfed,
##     3) + pctymle, data = Crime_data[-test, ])
##
```

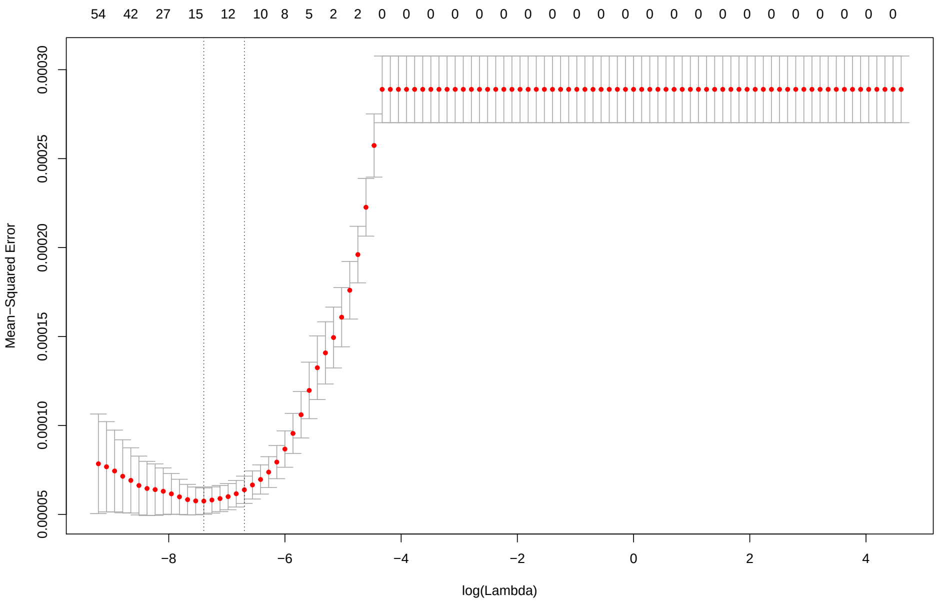
```
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.021877 -0.004627 -0.000597  0.003989  0.097203 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             0.0856338  0.0058694  14.590  < 2e-16 ***
## log(prbarr)            -0.0157783  0.0012325 -12.802  < 2e-16 ***
## log(prbconv)           -0.0106568  0.0008499 -12.538  < 2e-16 ***
## log(polpc)              0.0138227  0.0008799  15.710  < 2e-16 ***
## density                 0.0053045  0.0004291  12.363  < 2e-16 ***
## as.factor(region)other  0.0038644  0.0011286   3.424 0.000669 ***
## as.factor(region)west  -0.0067882  0.0016100  -4.216 2.96e-05 ***
## poly(pctmin, 4)1        0.0604510  0.0145040   4.168 3.63e-05 ***
## poly(pctmin, 4)2        0.0121701  0.0113810   1.069 0.285447    
## poly(pctmin, 4)3       -0.0312023  0.0100986  -3.090 0.002117 ** 
## poly(pctmin, 4)4       -0.0141515  0.0099306  -1.425 0.154785    
## poly(wfed, 3)1          0.0017264  0.0115936   0.149 0.881684    
## poly(wfed, 3)2         -0.0054095  0.0096429  -0.561 0.575066    
## poly(wfed, 3)3         -0.0096365  0.0091300  -1.055 0.291725    
## pctymle                 0.0093259  0.0183221   0.509 0.610983    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.008842 on 490 degrees of freedom
## Multiple R-squared:  0.7707, Adjusted R-squared:  0.7641 
## F-statistic: 117.6 on 14 and 490 DF,  p-value: < 2.2e-16
```

The lack of significance of the polynomials for `wfed` and the increase in R^2 suggests that this model could possibly overfit. Hence, we removed the polynomials related to `wfed`, but kept the *log* predictors as they have shown to be very significant. `pctmin` has a special behaviour, where we see that only the $3rd$ degree polynomial is used and seems less significant that the linear approach. Consequently, we decided to remove both the polynomial and the linear coefficients and re attach `pctmin` as a spline in Non-linear Modeling. This updated model yielded a *testing MSE* of 6.3680458×10^{-5}, and an R^2 value of 0.7706655.

Next, we analysed all possible interactions between all original predictors and plugged them to our previous model, yielding two new models. The number of interactions that we added to them is 325 and the R^2 values for each fit are 0.9885782 and 0.9896239, respectively. Consequently, we can affirm that both models are seriously overfitting because the number of predictors has skyrocketed due to all interaction combinations. Even though it is tempting to keep only those interactions with a relevant significance value, since the removal of each of these predictors affects the overall model, we chose instead to refine it by using a *Stepwise Algorithm* applying *AIC* to decide. The resulting fit has 235 coefficients, which means a reduction on the number of predictors by 27.6923077%. We then proceeded to calculate both *training MSE* (2.0005836×10^{-4}) and *testing MSE* (0.0020185). Now that we obtained a complex model consisting of linear variables, *log* variables and *interaction* variables, we will perform *Lasso* in order to remove all interaction terms that are not significant, so that we arrive to a model easy to understand.

3.6 Lasso Analysis

With the resulting model from all our previous steps, we performed *k-fold cross validation* using Lasso, in order to obtain the optimum value of λ for our model. The plot showing the *cross-validation error* as λ increases is the following:



We then chose a value of λ within 1 standard deviation from the optimum value, as this is a commonly established good practice.

Lasso yielded the following 13 non-zero predictors:

log(prbconv)
log(polpc)
poly(wfed, 3)1
prbarr
wfed
pctmin:county
polpc:regionother
density:regionother
density:pctmin
density:pctymle
pctmin:regionwest
regionwest:wsta
pctymle:regionwest

This means a reduction of 94% with respect to the number of predictors returned by *stepwise AIC*. However, some of the interaction terms that appear involve predictors that were not in the original model (`wsta`, `pctymle`, `county`, `region`, `pctmin` and `density`). Hence, we added these predictors (except `pctmin` which, as mentioned beforme, will be added as a spline) to the model in order to provide the final predictor set for this section. The *testing MSE* for this model is 6.263589×10^{-5} and this error is 4.3626443% of the *testing MSE* of our original model using *best subset selection*. Finally, we analysed the $L1$-norm between our estimated responses and the true model, also called *approximation error*, as:

$$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_{lasso_i}}{y_i}\right|$$

which yielded a value of 0.2036028, which is more than reasonable since it's below 30%. Consequently, the final set obtained in this section, after performing *linear*, *best subset selection*, *log*, *polynomial*, *stepwise AIC* and *Lasso* yielded a model that will be used in the following sections for more complex fits that will derive in our final model.

3.7 Non-linear Modeling

From both Linear Analysis and the pairs analysis in Dataset, we identified a predictor `pctmin` shown that could benefit from a more flexible modeling using poylnomial regression and splines and improve the overall prediction.

3.7.1 Linear Model

First we evaluated the regression model generated using only using a linear model between `crmrte` and `pctmin`:

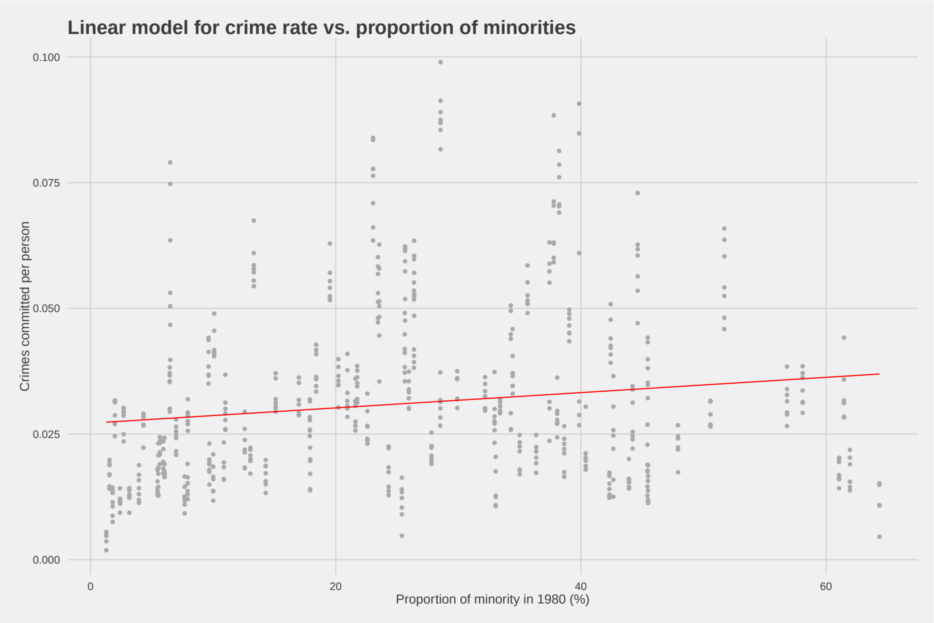


Figure 1: Linear model for crime rate vs. proportion of minorities

This naive model results in a MSE of 3.173×10^{-4} and a mean approximation error rate of 0.497 for the testing set. From the plot, it is clear that the relationship between the crime rate and the proportion of minorities in the area is not linear.

3.7.2 Polynomial Model

Next, we obtained a degree-4 polynomial function for a smooth fit over the `pctmin` data:

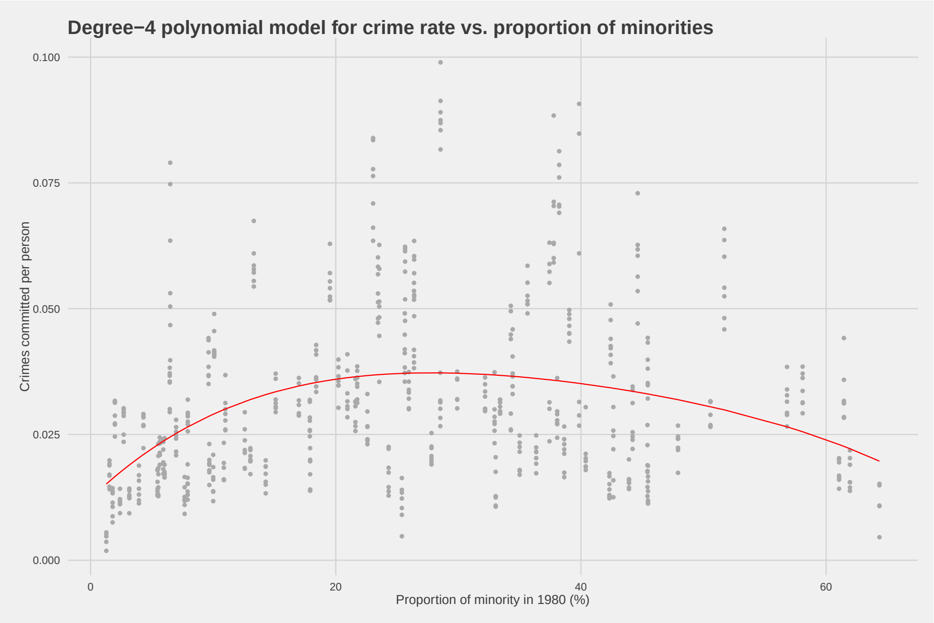


Figure 2: Degree-4 polynomial model for crime rate vs. proportion of minorities

This fit resulted in a MSE of 2.484×10^{-4} and a mean approximation error rate of 0.3931. The model was improved by reducing the bias of the model.

3.7.3 Splines

We further attempt to reduce the bias, introducing a more flexible piecewise polynomial by using knots. With a cubic spline, the fitted curves and their first and second derivatives are constrained to be continuous at the knots. As splines often lead to high variance at the outer ranges of the predictors, we fit a natural cubic spline, which forces the function to be linear at the boundary. `ns()` function was used to generate natural cubic knots with 8 degrees of freedom, with matrix of basis functions for splines and knots at 5.6%, 10.1%, 18%, 25.4%, 33%, 38.3%, and 45.4% of `pctmin`.

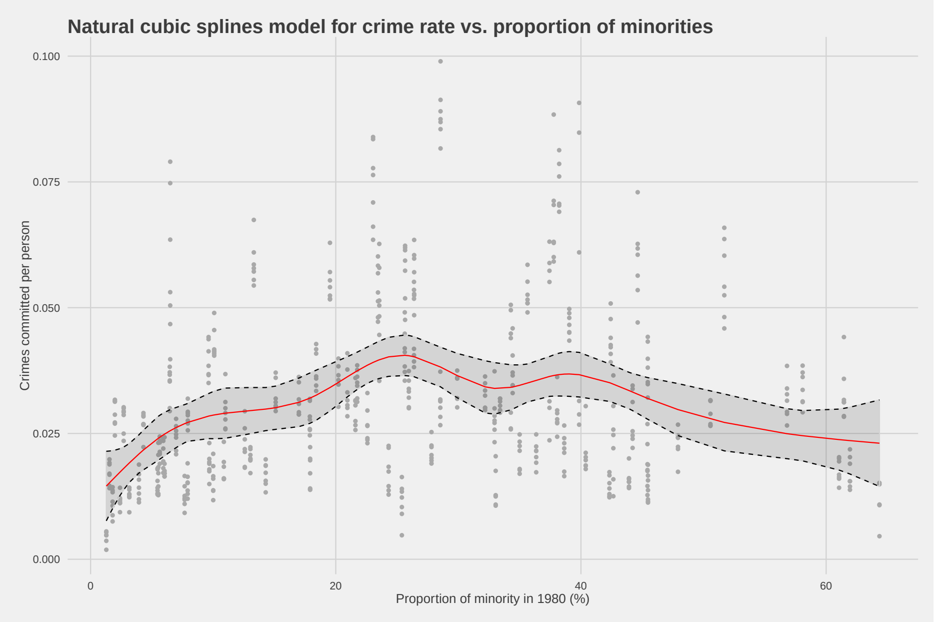


Figure 3: Natural cubic splines model for crime rate vs. proportion of minorities

Consequently, we see a modest improvement in the MSE (2.34×10^{-4}) and the mean approximation error (0.386).

We also attempted to fit a smoothing spline with a value of λ chosen using cross-validation. This resulted in a model very similar to the polynomial fit and failed to improve the mean testing error.

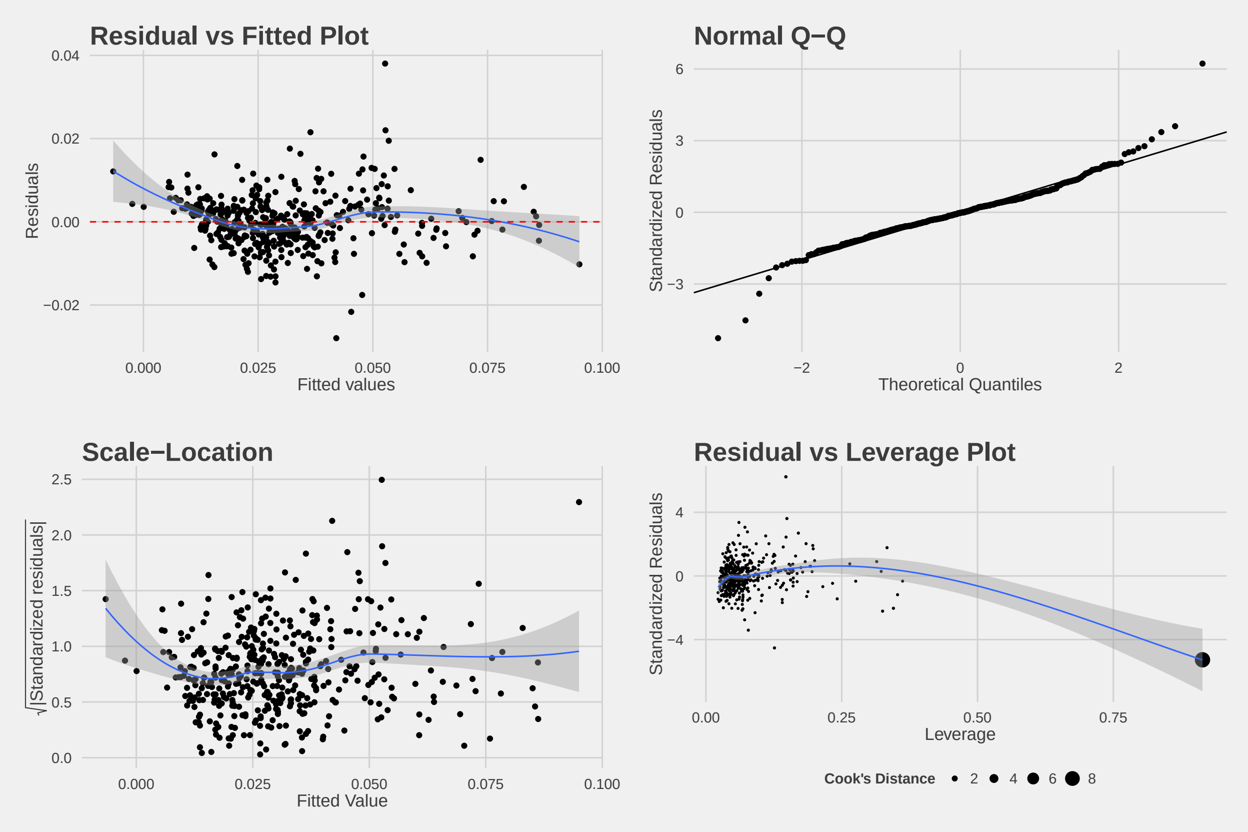
3.7.4 Combined models

Combining the predictors from the previous section with just the linear `pctmin` results in a MSE of 6.158×10^{-5} and a mean approximation error of 0.2018.

Finally, we included the splined version of `pctmin` predictor alongside the selected predictors, which resulted in a slight improvement in the MSE to 6.02×10^{-5} and the mean approximation error to 0.204.

3.8 Residual Analysis

To perform the residual analysis on our final project, we produced the following plot:



We observe the following:

- First, we notice that the residual variance is homoscedastic from the *Residual vs Fitted* and the *Scale-Location* plots , this means that it is constant. However the residuals are slighlty lower with smaller and larger values of the response variable. Additionally, you can see that the mean of the residuals is 0.
- The *Normal Q-Q Plot* shows that the quantiles of the standardized residuals versus the quantiles of a standard normal are similiar, and meet the line reference almost perfectly except for the extreme points. This indicates that our residuals follow the normality hypothesis.
- The *Residual vs Leverage Plot* exhibits that one of the residuals is highly leveraged and has a large Cook's distance, this corresponds to an observation which is both influential and an outlier. The rest of the residuals have a more regular behaviour.

4 Conclusion

In this project, we wanted to apply methods and knowledge learned in the class to a real-world data. Using crime rate data in North Carolina from 1981 to 1987, we aimed to build a robust model that could be used predict the crime rate and understand some of the underlying factors behind the crime rate. Toward this goal, we followed a methodology that consists in a series of steps, each of this corresponding to a statistical learning topic covered along the semester.

First, we indentified our target variable and explored their median value. We examined the predictors and their relationship with the response. Then, we noticed which variables were significant and correlated. We detected influential observations and removed them accordingly. After cleaning the dataset, we run an ANOVA model and a decision tree algorithm to confirm their interdependece.

We created a linear regression model containing all the predictors, performed best subset selection to determine the best combination of predictors to use for our predictive model. We further removed more predictors using lasso regression, which yielded 13 non-zero predictors. During this analysis, we also identified a predictor with a non-linear relationship that could benefit from a polynomial/splines modeling instead. Finally, we arrive at a predictive model that includes the predictors from the shrinkage analysis and the splines of the non-linear predictor. The progression of the mean squared errors is shown below:

model	trainingMSE	testingMSE
linear+best subset	0.0012464	0.0014357
linear+interactions (stepwise selection)	0.0002001	0.0020185
linear+interactions (LASSO)	0.0000414	0.0000626
lasso results + splines	0.0000393	0.0000602

Some of the variables identified through the methods include the percentage of minorities in the area (`pctmin`), interactions with population density (`density`) and region (`region`), weekly federal wages (`wfed`), as well as the probability of arrest (`prbarr`) and conviction (`prbconv`). Considering there were many more variables included in the dataset that one would assume to be related to incidence of crime, our model identified several statistically relevant predictors that might warrant a closer look in relation to crime rates.