

# Report

*Carlos Espino, Xavier Gonzalez, Diego Llarrull, Woojin Kim*

*December 14, 2015*

## Contents

<b>Introduction</b>	<b>2</b>
<b>Dataset</b>	<b>2</b>
<b>ANOVA models</b>	<b>8</b>
<b>Confidence Interval for the Median</b>	<b>10</b>
<b>Dependency analysis with predictive models</b>	<b>10</b>
<b>Linear analysis</b>	<b>12</b>
<b>Lasso Analysis</b>	<b>16</b>
<b>Non-linear modeling</b>	<b>17</b>
Linear model . . . . .	17
Polynomial model . . . . .	17
Splines . . . . .	18
Combined models . . . . .	19

# Introduction

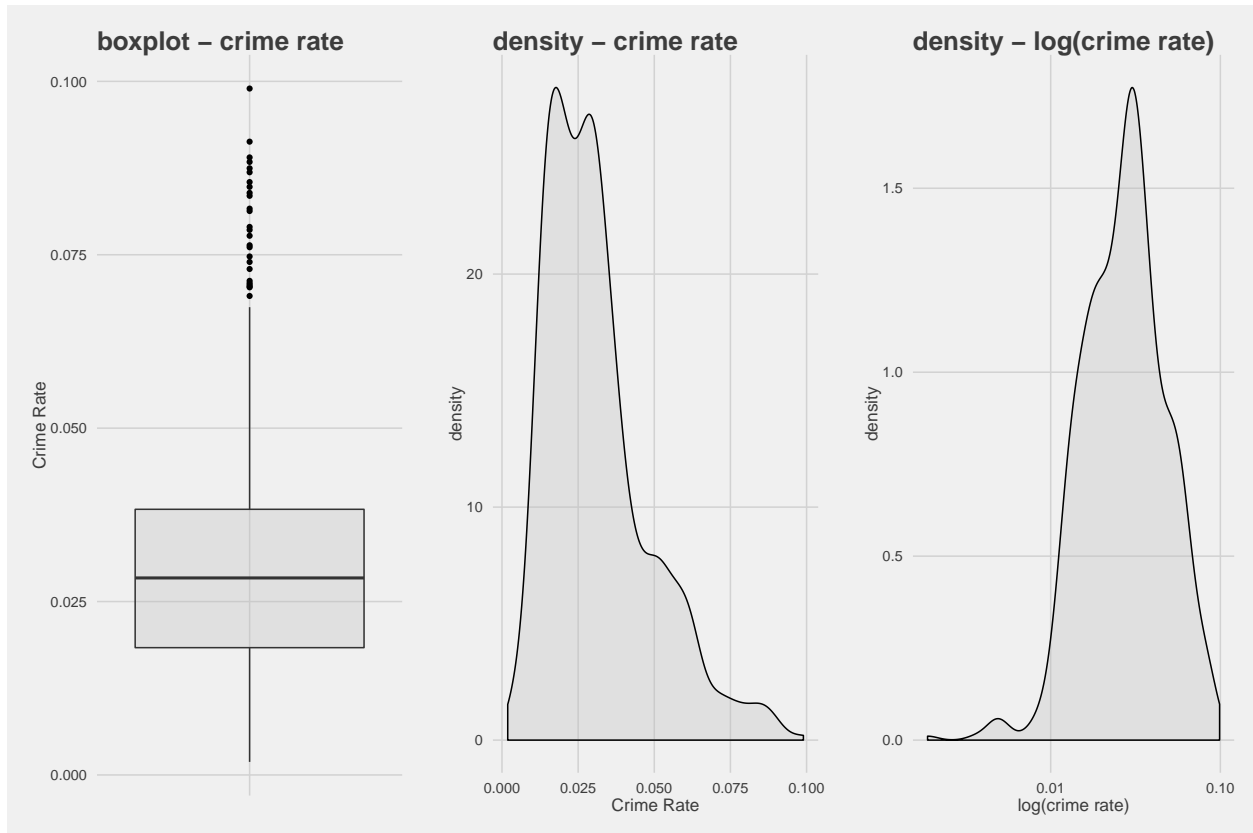
Understanding the factors behind criminal behaviour is one of the most crucial tasks for preventing and controlling future crime. In this report, we explore the potential factors affecting crime rates based on the demographics and econometrics data gathered from 197 counties in North Carolina from 1981 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 25% error using the selected parameters.

## Dataset

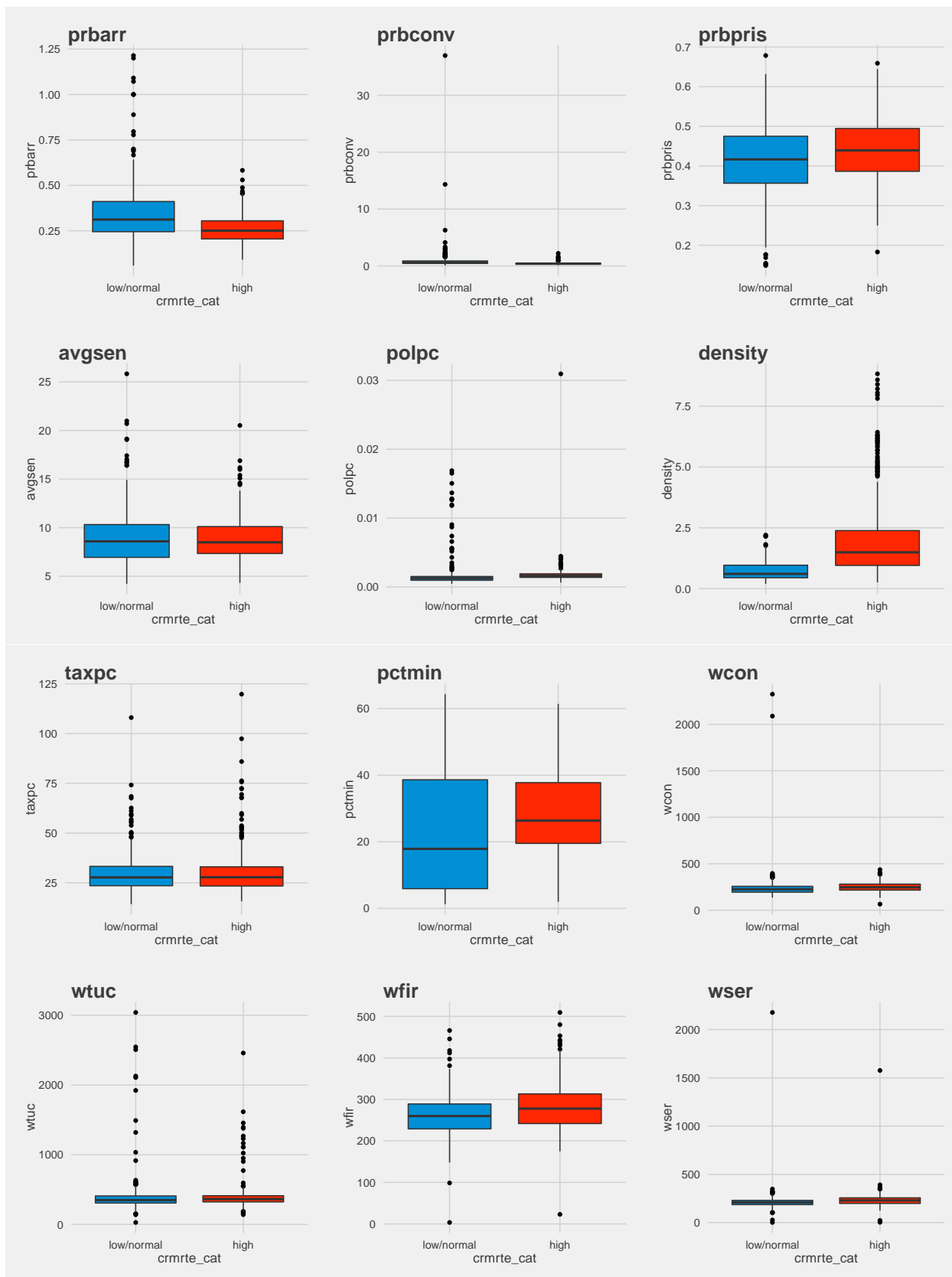
Predictor	Description
county	county identifier
year	year from 1981 to 1987
crmrte	crimes committed per person
prbarr	'probability' of arrest
prbconv	'probability' of conviction
prbpris	'probability' of prison sentence
avgsen	average sentence, days
polpc	police per capita
density	people per square mile
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

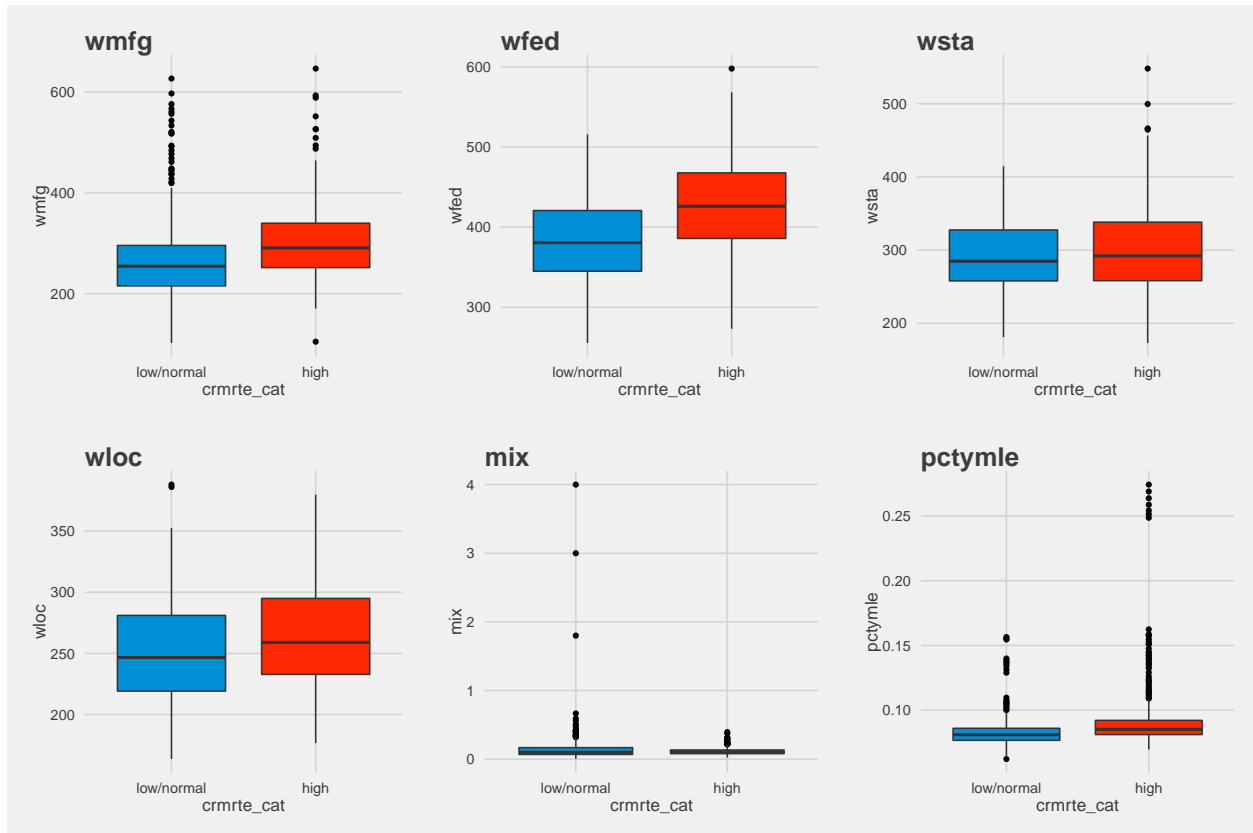
Table 1: Description of the predictors in the dataset

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 than 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 statistical 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.

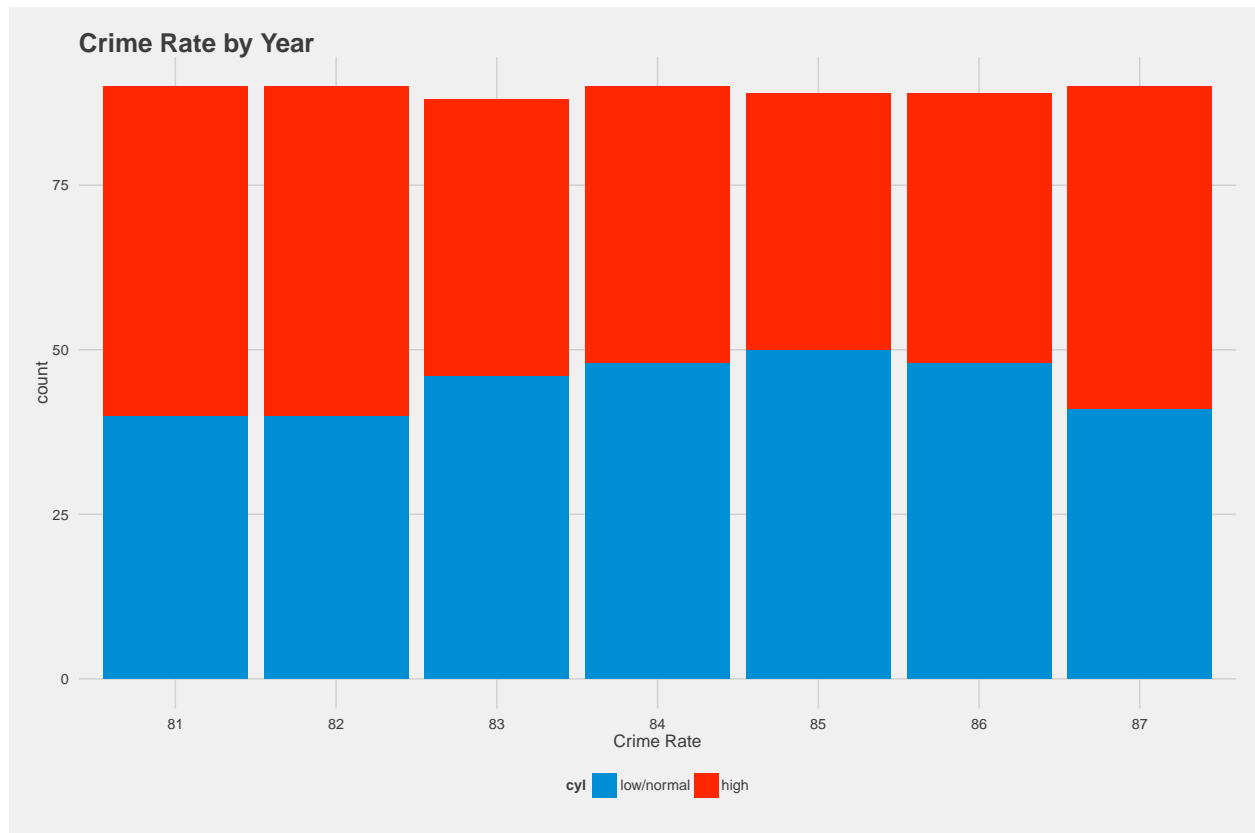


Besides the target variable, the dataset contains other 21 variables we used as predictors. Two of them have categorical values. The **region** variable can have 3 possible values: **other**, **west** or **central** and the **smsa** can have **yes** or **no**. The dataset also contains the **year** variable 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.

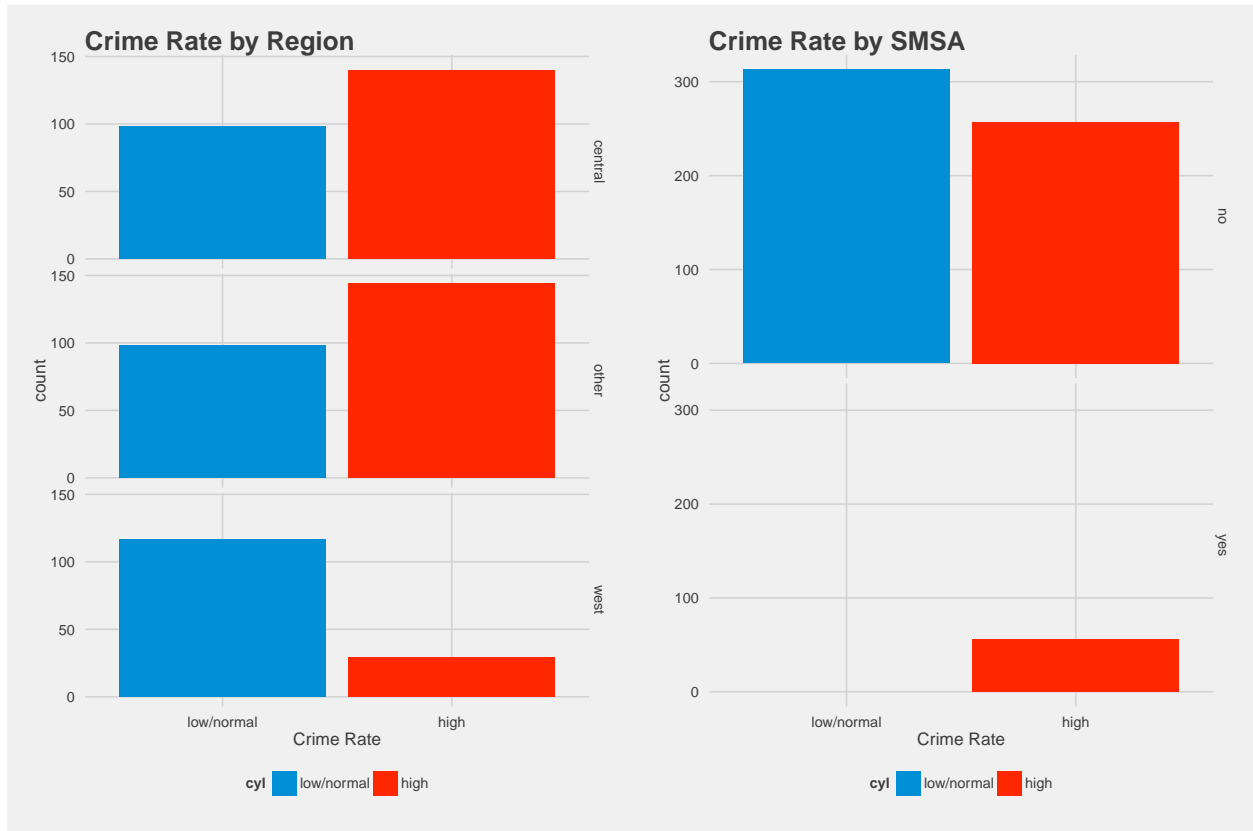




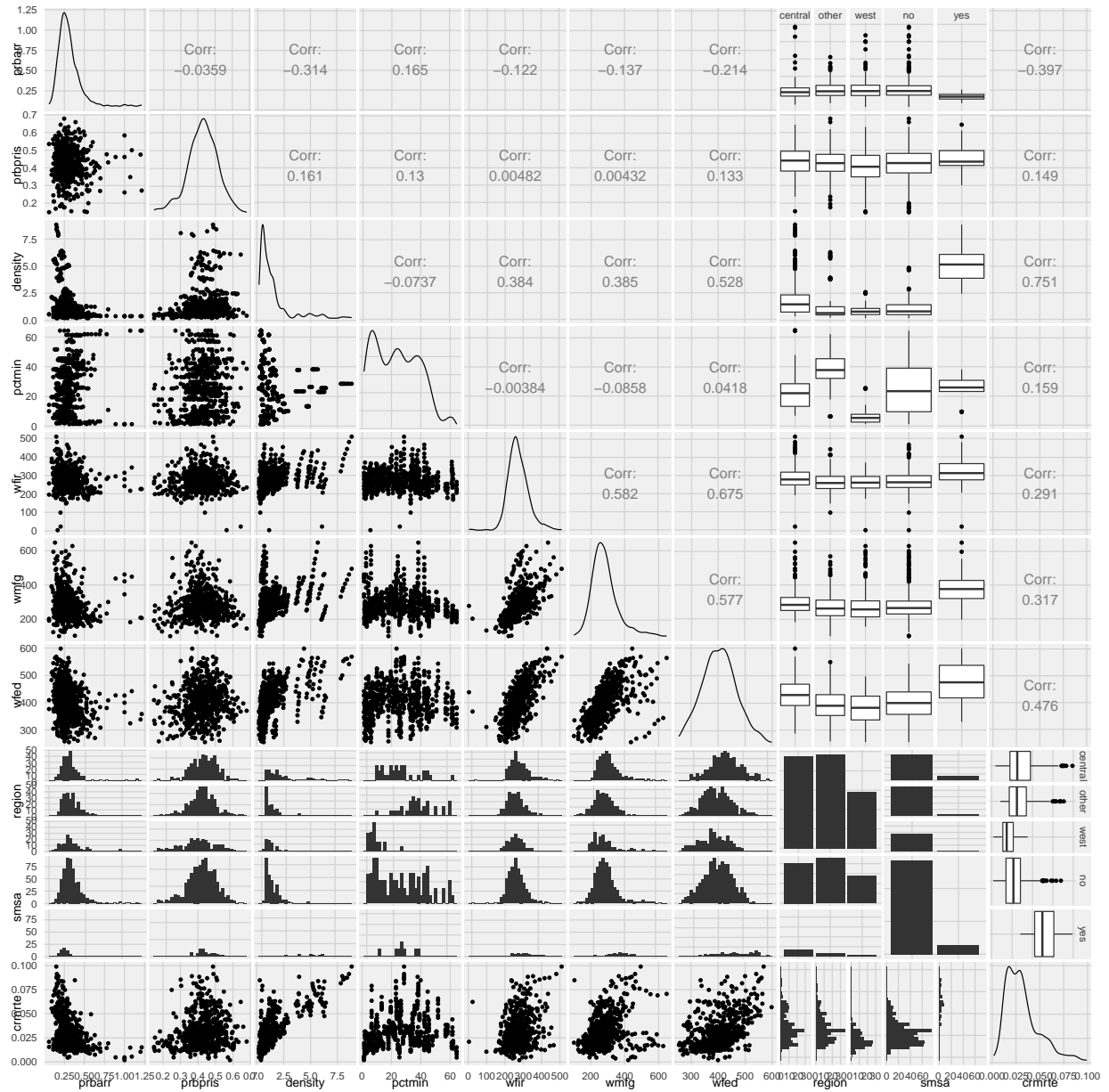
In the boxplots above, we can see that the variables that may have a predictive value with the target are variables `prbarr`, `density`, `pctmin`, `wfed`, `wmfg` and `pctmle` as they separate the population partially by the value of the defined target variable. We explore the rest of the predictors by 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 **wmg**, **wfed** and **density**. We will later discuss whether these variables are significant for modeling.

## 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 0.02476 0.024761   97.18 <2e-16 ***
```



```
## Residuals    624 0.15899 0.000255
## ---
## 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
## 0.03475108 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 returned in the package.

```
## (Intercept) region_w_nww
## 0.03475108 -0.01487221
```

The model is given by

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

Now, we repeat the same analysis considering two factors. We also 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 0.02476 0.02476   155.8 <2e-16 ***
## smsa         1 0.05999 0.05999   377.5 <2e-16 ***
## Residuals    623 0.09900 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 hypothesis (i.e., the means are equal) is rejected. The coefficients in this case are:

```
## (Intercept) region_w_nww      smsayes
## 0.03123831 -0.01300927 0.03441082
```

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 0.02476 0.02476 157.927 < 2e-16 ***
## smsa         1 0.05999 0.05999 382.613 < 2e-16 ***
## region_w_nw:smsa  1 0.00147 0.00147   9.404 0.00226 **
## Residuals      622 0.09752 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.

## Confidence Interval for the Median

As discussed above, we defined a categorical target variable: **high**, if the value of the crime rate was higher than 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] 0.02840405
```

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	0.0265029	0.0296409
0.95	0.0266806	0.0293452
0.88	0.0268714	0.0292154
0.81	0.0269836	0.0290823

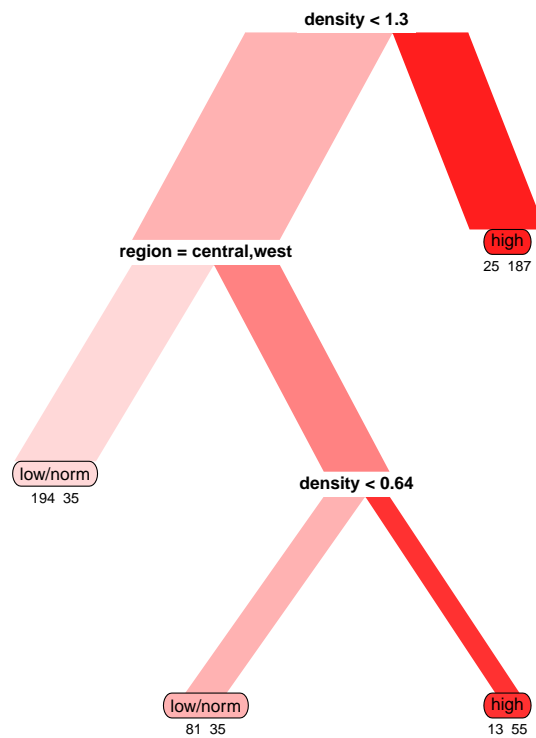
A more sophisticated method to obtain the confidence interval for the median is the *Wilcoxon Signed Rank Test*. As this test assumes symmetry of the variable's distribution, we applied it to the logarithm of the target variable, as discussed above. The results were then transformed to return the values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

signif. %	low.level	up.level
0.99	0.0257032	0.0289377
0.95	0.0260857	0.0285302
0.88	0.0263252	0.0282818
0.81	0.0264716	0.0281065

The results are similar to the simpler sign test.

## 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    102    28
## high          17    84
##
##           Accuracy : 0.8052
##           95% CI : (0.7482, 0.8542)
##       No Information Rate : 0.5152
##       P-Value [Acc > NIR] : <2e-16
##
##           Kappa : 0.6089
##  Mcnemar's Test P-Value : 0.136
##
##           Sensitivity : 0.8571
##           Specificity : 0.7500
##       Pos Pred Value : 0.7846
##       Neg Pred Value : 0.8317
##           Prevalence : 0.5152
##       Detection Rate : 0.4416
##       Detection Prevalence : 0.5628
##       Balanced Accuracy : 0.8036
##
##       '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`.

## 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 as shown in the code below:

We then run a simple linear fit with 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)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.0212397 -0.0043024  0.0000726  0.0037447  0.0309866
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.111e-02  2.185e-02   3.254 0.001201 **
## county        1.288e-05  5.086e-06   2.532 0.011593 *
## year         -7.737e-04  2.987e-04  -2.591 0.009814 **
## prbarr        -2.182e-02  2.754e-03  -7.923 1.13e-14 ***
## prbconv       -1.339e-03  2.330e-04  -5.746 1.45e-08 ***
## prbpris       -3.512e-03  3.451e-03  -1.018 0.309238
## avgscen        6.224e-05  1.096e-04   0.568 0.570397
## polpc          1.092e+00  1.769e-01   6.172 1.25e-09 ***
## density        5.530e-03  4.551e-04  12.151 < 2e-16 ***
## taxpc          1.901e-04  3.175e-05   5.987 3.69e-09 ***
## regionother    2.501e-03  8.242e-04   3.034 0.002520 **
## regionwest     7.219e-05  9.990e-04   0.072 0.942419
## smsayes       -4.186e-04  1.981e-03  -0.211 0.832697
## pctmin         1.137e-04  2.760e-05   4.120 4.32e-05 ***
## wcon           2.487e-07  2.437e-06   0.102 0.918757
## wtuc          -5.245e-08  1.073e-06  -0.049 0.961010
## wtrd           5.217e-06  3.435e-06   1.519 0.129384
## wfir          -6.563e-06  8.139e-06  -0.806 0.420382
## wser          -4.666e-06  2.846e-06  -1.640 0.101613
## wmf           -5.262e-06  4.818e-06  -1.092 0.275165
## wfed           2.252e-05  7.793e-06   2.889 0.004002 **
## wsta          -8.173e-06  8.061e-06  -1.014 0.311055
## wloc           2.040e-05  1.446e-05   1.411 0.158871
## mix            2.807e-03  1.791e-03   1.567 0.117698
## pctymle        4.517e-02  1.319e-02   3.426 0.000655 ***
## crmrte_cat.L    9.456e-03  5.249e-04  18.016 < 2e-16 ***
## region_w_nww           NA           NA           NA           NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.006987 on 600 degrees of freedom
## Multiple R-squared:  0.8406, Adjusted R-squared:  0.8339
## F-statistic: 126.6 on 25 and 600 DF,  p-value: < 2.2e-16
```

as the  $R^2$  value is sufficiently high (0.8405914), 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:

After getting all best subsets with size  $k = 1 \dots p$ , 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

```
## [1] 0.0004752172
```

The cross-validation error is

```
## [1] 0.0005478937
```

The actual training and test errors for this subset, on the original datasets, are

```
## [1] 0.001248871
```

```
## [1] 0.001365305
```

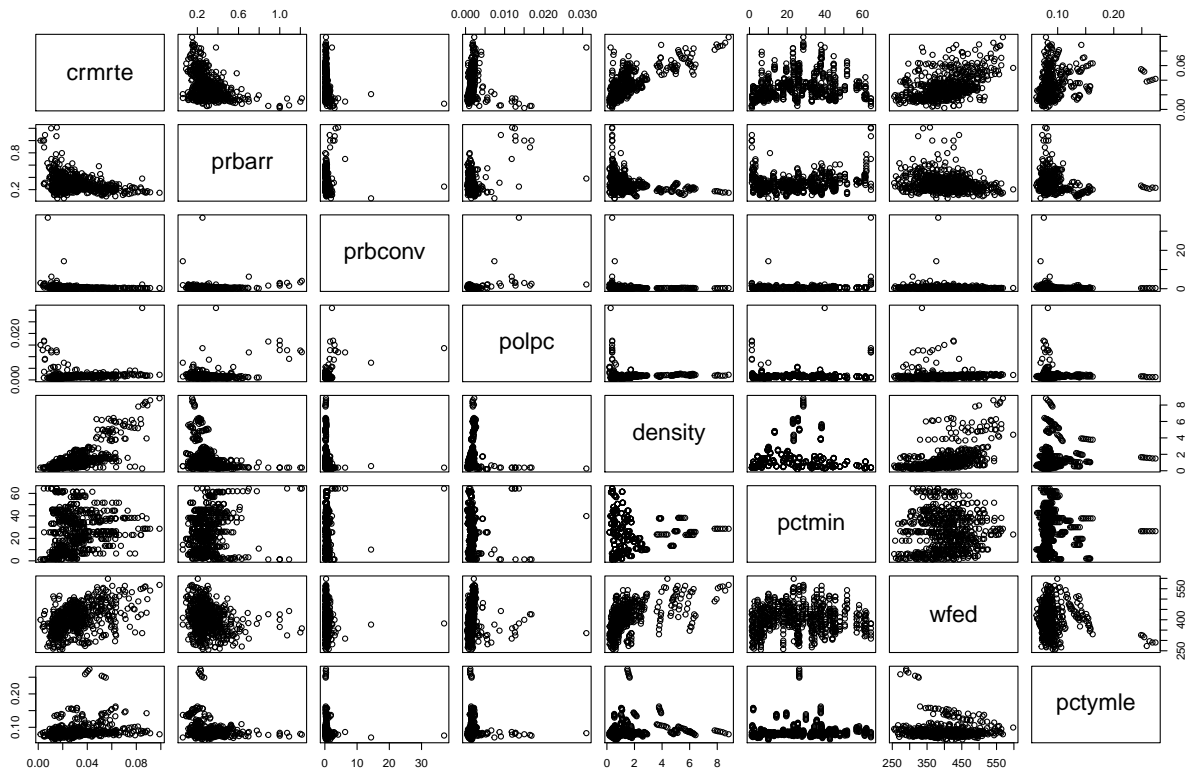
both obtained when using the best subset with  $k = 8$  predictors. The ratio between *testing* and *training* errors is (1.0932317). Consequently, we can conclude that the predictors yielded by the subset generated using *best subset selection* belong to a consistent model and, hence, can be used as 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
## -0.021247 -0.005769 -0.000725  0.004217  0.047783
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.205e-03  3.586e-03   2.009 0.044942 *
## prbarr         -3.202e-02  3.098e-03 -10.336 < 2e-16 ***
## prbconv        -1.807e-03  2.502e-04  -7.222 1.51e-12 ***
## polpc           1.944e+00  2.112e-01   9.203 < 2e-16 ***
## density         7.193e-03  3.171e-04  22.684 < 2e-16 ***
## as.factor(region)other 4.681e-03  9.785e-04   4.784 2.15e-06 ***
## as.factor(region)west -3.471e-03  1.161e-03  -2.990 0.002898 **
## pctmin          1.243e-04  3.220e-05   3.861 0.000125 ***
## wfed            2.641e-05  6.936e-06   3.808 0.000154 ***
## pctymle         7.466e-02  1.546e-02   4.828 1.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.00887 on 616 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.7324
## F-statistic: 191.1 on 9 and 616 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 (0.7362415 versus 0.8405914), 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` has a nonlinear relationship with the response, which makes it suitable as a polynomial regression predictor. 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) + pctmin + poly(wfed, 3) + pctymle,
##     data = Crime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.024202 -0.004384 -0.000246  0.003672  0.045724
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          0.0564773  0.0046897  12.043 < 2e-16 ***
## log(prbarr)          -0.0144423  0.0009070 -15.923 < 2e-16 ***
## log(prbconv)         -0.0113128  0.0006276 -18.024 < 2e-16 ***
## log(polpc)           0.0099603  0.0006892  14.452 < 2e-16 ***
## density              0.0052809  0.0003094  17.068 < 2e-16 ***
## as.factor(region)other 0.0039568  0.0008168   4.844 1.61e-06 ***
## as.factor(region)west -0.0042023  0.0009729  -4.319 1.82e-05 ***
## pctmin               0.0001813  0.0000271   6.691 5.01e-11 ***
## poly(wfed, 3)1        0.0155450  0.0094692   1.642  0.101
## poly(wfed, 3)2        0.0006990  0.0079938   0.087  0.930
## poly(wfed, 3)3        -0.0101414  0.0076411  -1.327  0.185
## pctymle              0.0073072  0.0135778   0.538  0.591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007408 on 614 degrees of freedom
## Multiple R-squared:  0.8166, Adjusted R-squared:  0.8133
## F-statistic: 248.6 on 11 and 614 DF,  p-value: < 2.2e-16

## [1] 4.763493e-05
```

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.

Next, we analysed all significant interactions between all original predictors and plugged them to our previous model. The number of interactions that we added to the model are:

```
## [1] 324 4
```

The  $R^2$  values for each fit are, respectively:

```
## [1] 0.9801254
```

```
## [1] 0.9806373
```

We can see that both models are seriously overfitting, while 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, which has the following number of coefficients:

```
## [1] 196
```

which means a reduction on the number of interactions by 30%, we proceeded to calculate both *training MSE* and *testing MSE*:

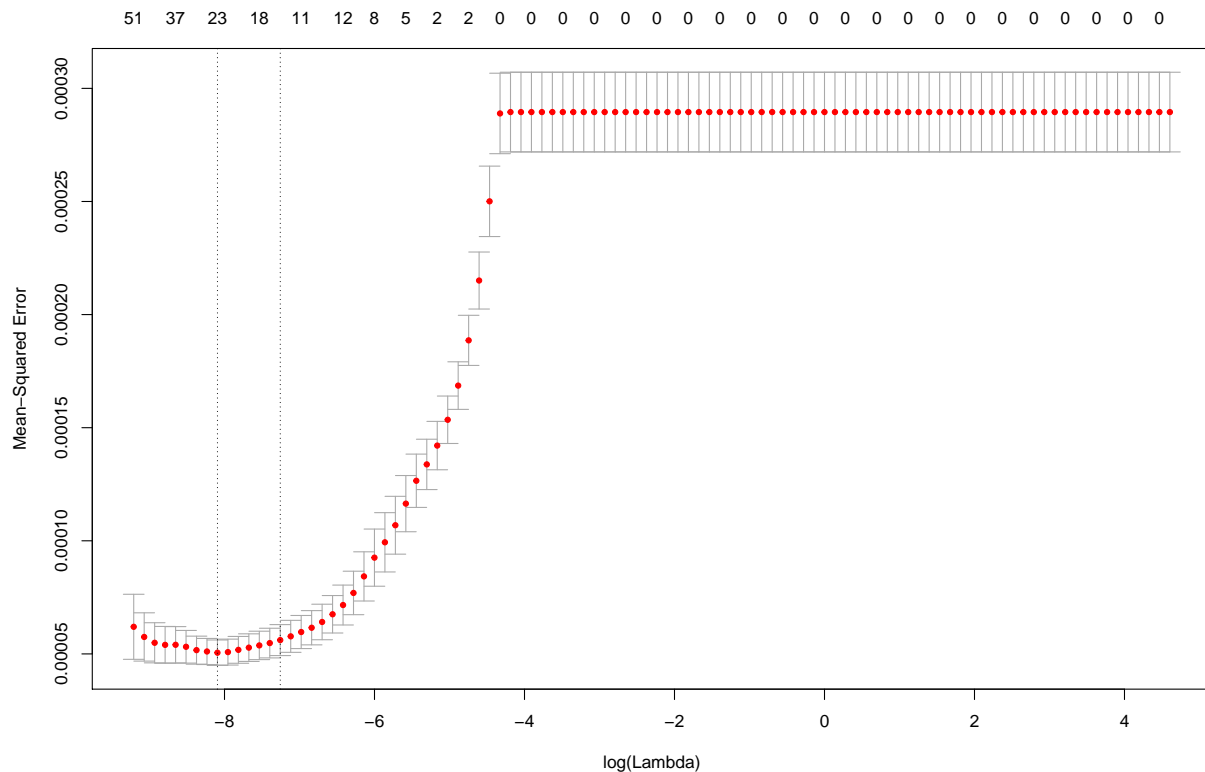
```
## [1] 0.001248871
```

```
## [1] 0.001365305
```

Here, we can see that the error rate remains close to one (1.0932317) which still proves that this model holds. 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.

## 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  $\lambda$  for our model. The plot showing the *cross-validation error* as  $\lambda$  increases is the following:



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

Lasso yielded 16 predictors with nonzero values, a reduction of 92% with respect to the *stepwise AIC*. The *test error* for this model is obtained using the same *k-folds* generated for the first model:

```
## [1] 5.608461e-05
```

and this error is 4.1078442% of the training error of our original model using *best subset selection*.

Finally, we analyse the  $L1$ -norm between our estimated responses and the true model:

```
## [1] 0.2026294
```

which yields a more than reasonable value, 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.



## Non-linear modeling

From the pairwise plot we generated shown in (), we identified a predictor `pctmin`, corresponding to the proportion of minorities in the region, showing a clear non-linear relationship with the crime rate that can benefit from higher order polynomial regression/splines and also help with predicting the overall crime rate.

### Linear model

First we evaluated the regression model generated using only using a linear model:

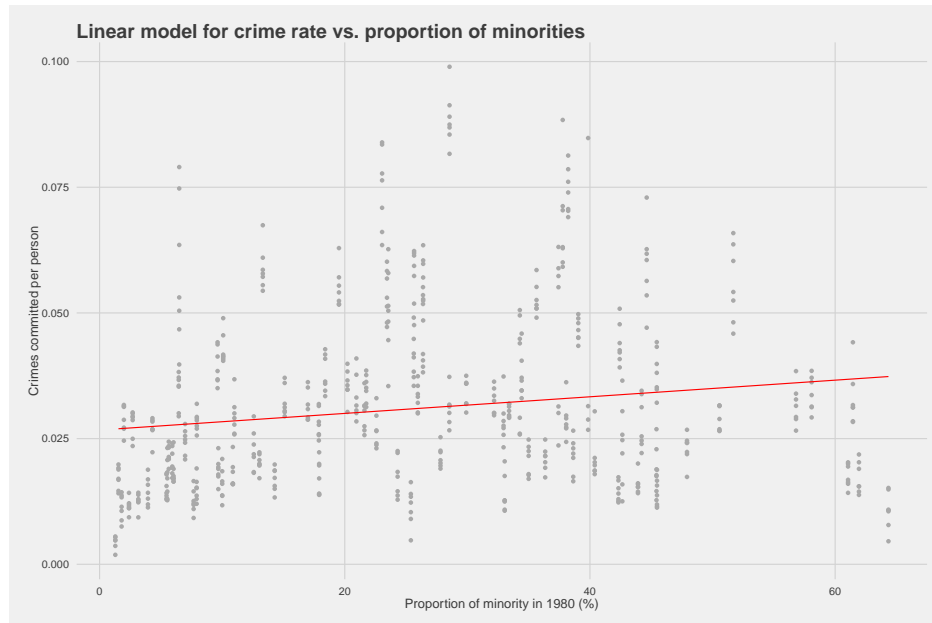


Figure 1: Linear model for crime rate vs. proportion of minorities

This naive model results in an error rate of 0.4808 for the testing set. From the plot it is very clear that the relationship between the crime rate and the proportion of minorities in the area is not linear.

### Polynomial model

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

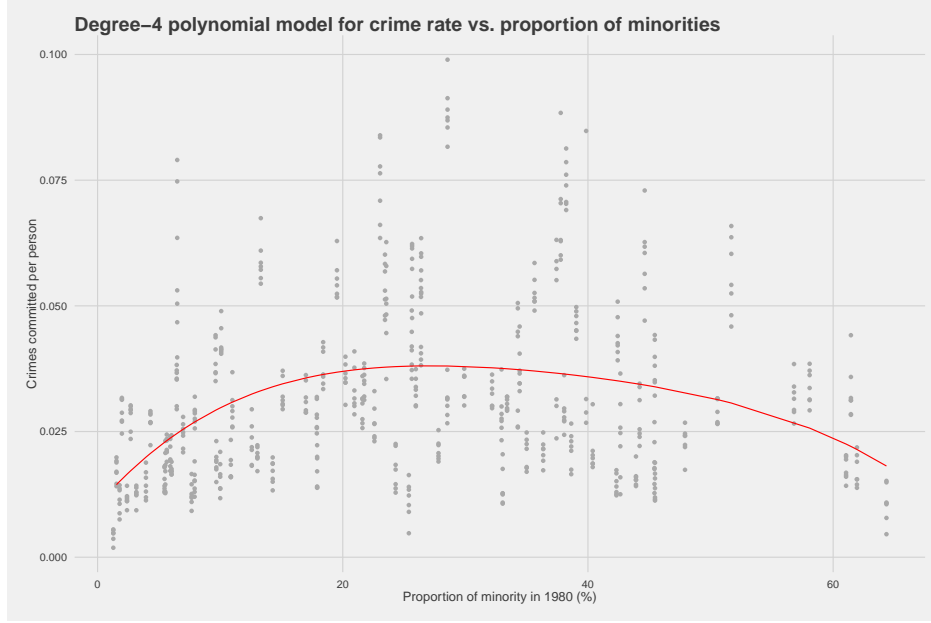


Figure 2: Degree-4 polynomial model for crime rate vs. proportion of minorities

The error rate was improved by reducing the bias of the model. This fit resulted in an error rate of 0.4755.

## Splines

We further attempt to reduce the bias, introducing a more flexible piecewise polynomial by using knots. Using a cubic spline, the fitted curves are constrained to be continuous. As splines often leads to high variance at the outern ranges of the predictors, we fit natural cubic spline, which forces the function to be linear at the boundary. `ns()` function was used to generate natural cubic knots with 6 degrees of freedom, with matrix of basis functions for splines and knots at 6.03%, 14.28%, 24.31%, 34.28%, and 42.64% of `pctmin`.

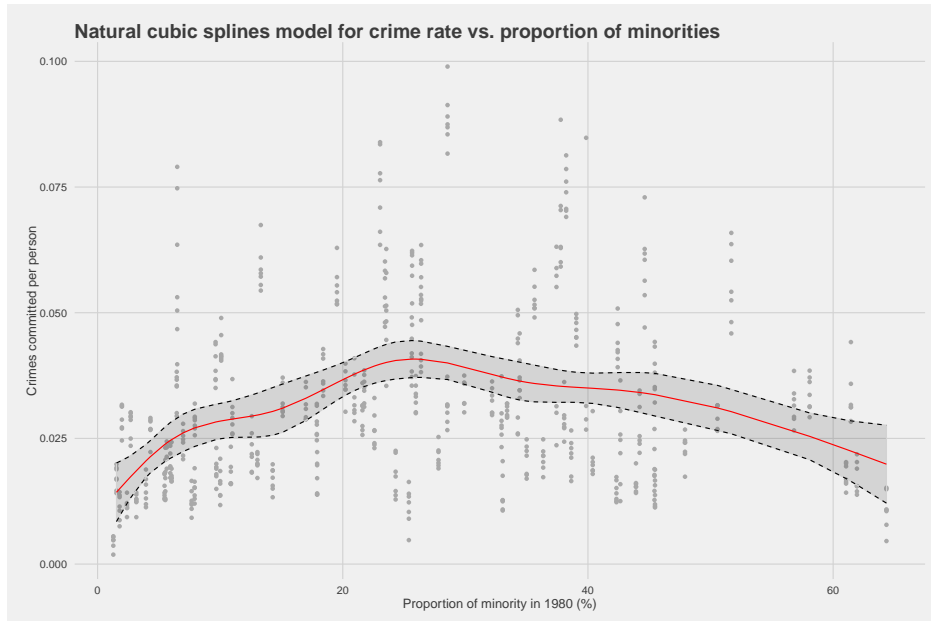


Figure 3: Natural cubic splines model for crime rate vs. proportion of minorities

Consequently, we see a modest improvement in the mean error: 0.468.

We also attempted to fit a smoothing spline with a value of  $\lambda$  chosen cross-validation. This resulted in a model very similar to the polynomial fit and failed to improve the mean  $k$ -folds cross-validation error.

## Combined models

Combining the predictors from the previous section with just the linear `pctmin` results in a mean error of 0.2385.

Finally, we included the splined version of `pctmin` predictor alongside the selected predictors, which resulted in a slight improvement in mean error to 0.2355.