Report

Carlos Espino

December 12, 2015

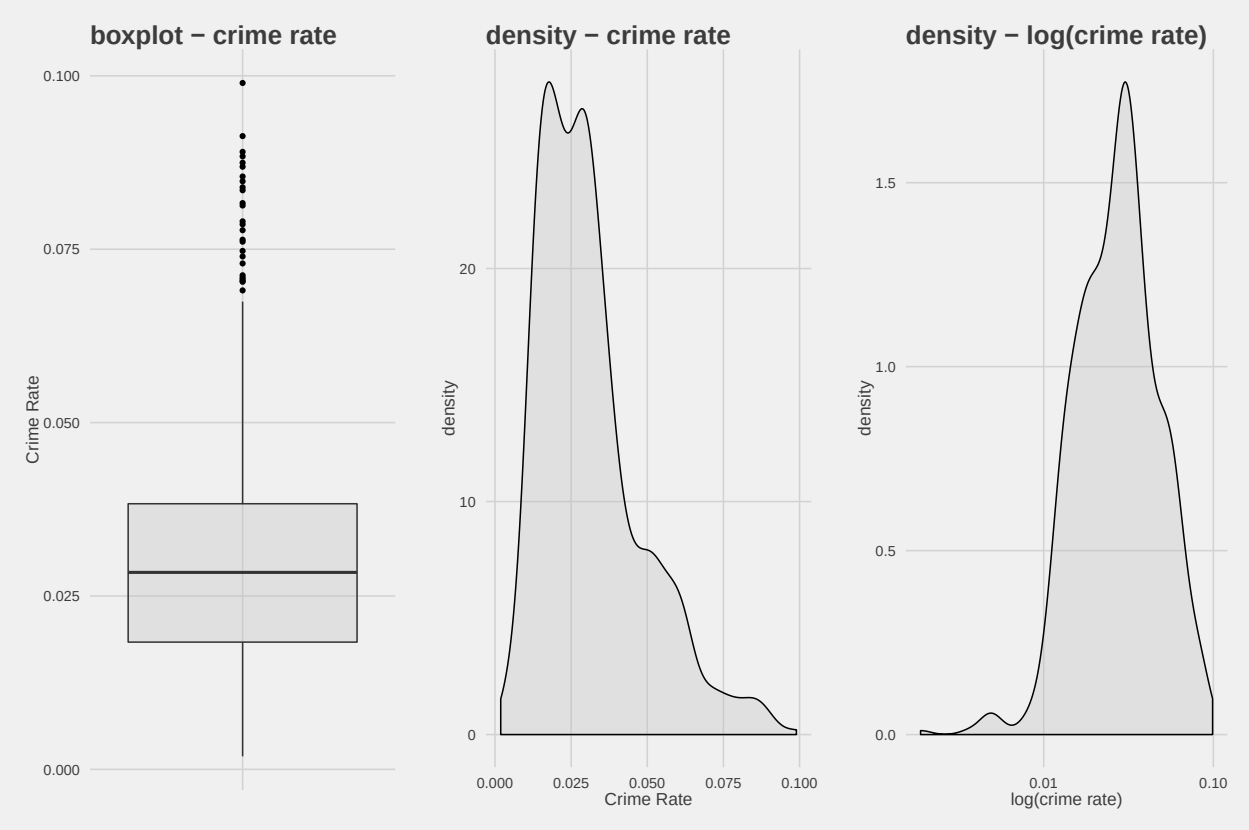
Introduction

Dataset

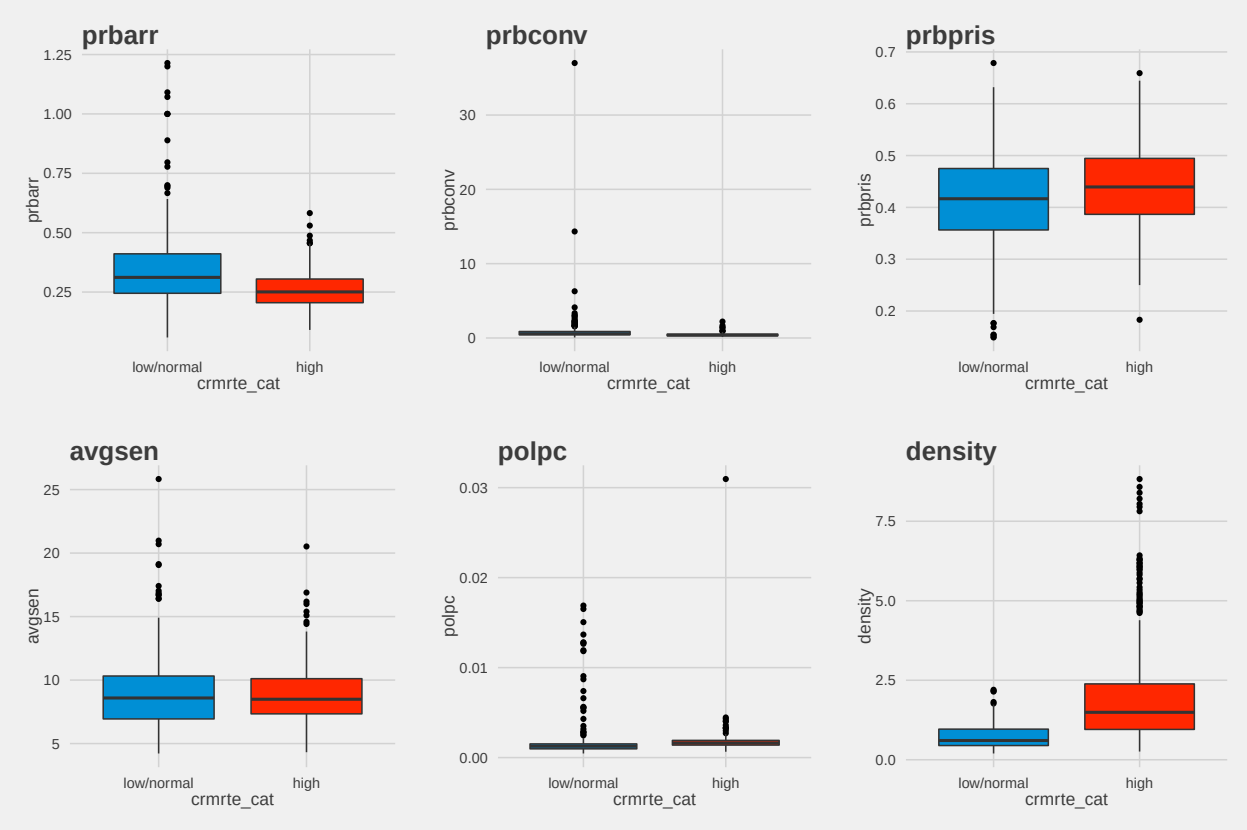
The dataset we will use as source as a study contains a total of 626 observations, 3 of them were considered as outliers and were removed according to the result of a residual analysis in a linear regression section. The columns of the data are showed below.

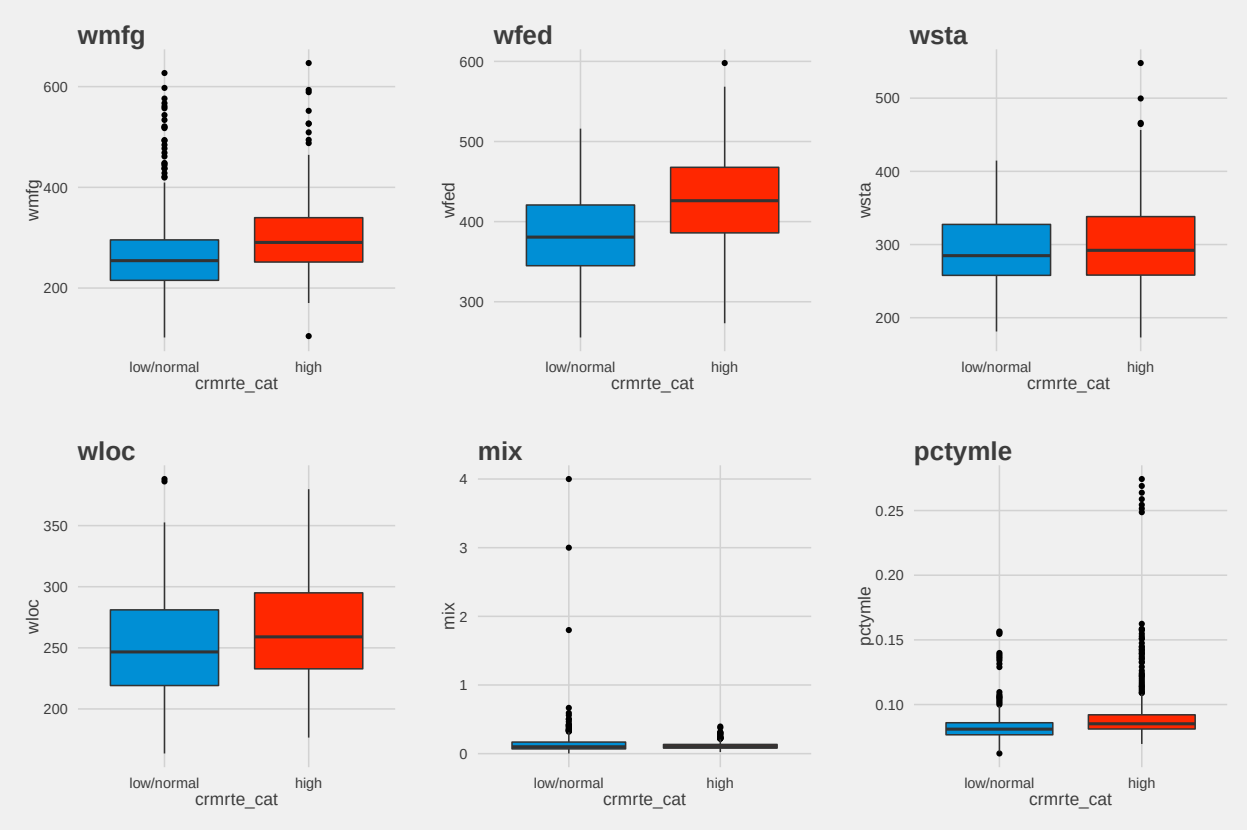
county	county identifier
year	year from 1981 to 1987
crmrte	crimes committed per person
prbarr	'probability' of arrest
prbconv	'probability' of conviction
prbpris	'probability' of prison sentenc
avgsen	average sentence, days
polpc	police per capita
density	people per square mile
taxpc	tax revenue per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

We analized the variables in the dataset starting with the target variable: 'crmrte', the crime rate. Along this study, we will use this variable in diferent 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. Also, we will use the natural logarithm of the variable to adequately transform it be able to apply diferent 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 and a softened histogram of the logarithm of the variable.

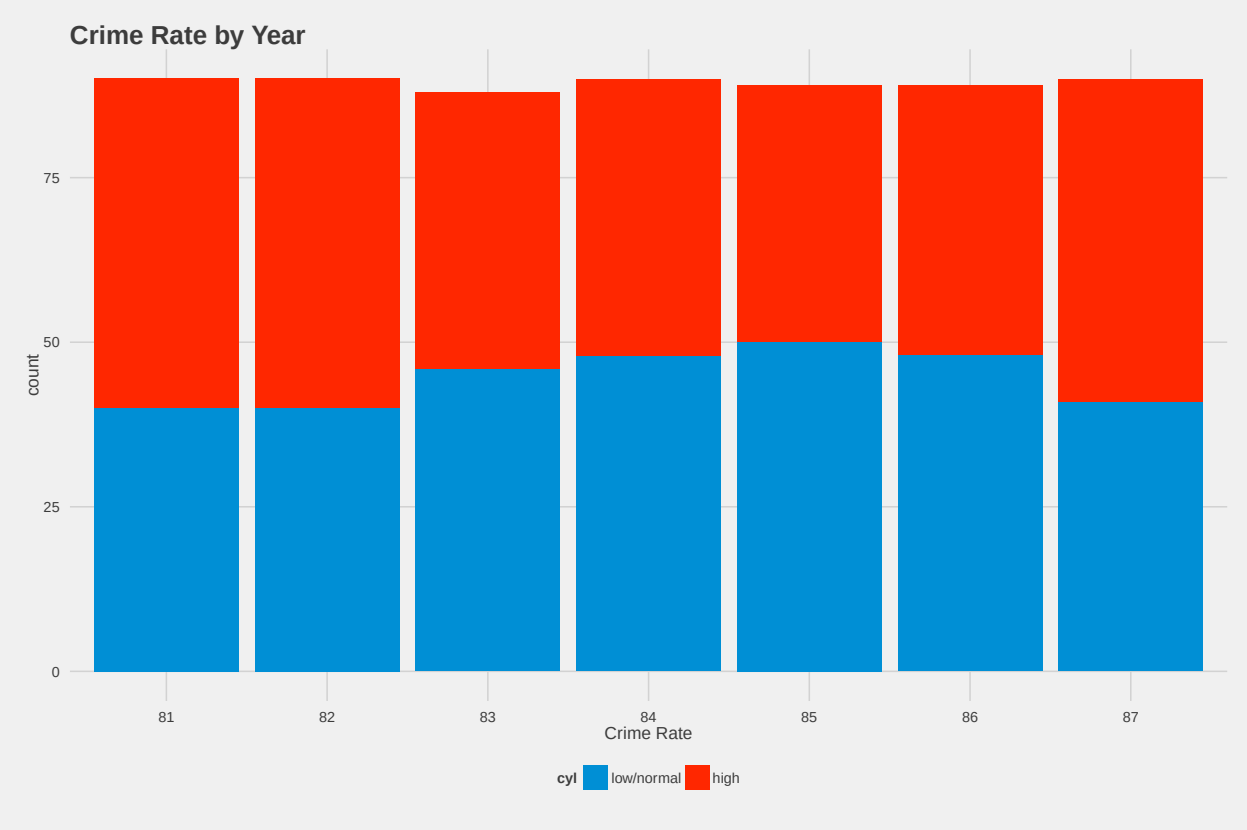


Besides the variable target, the dataset contains other 20 variables we used as predictors. Two of them have categorical values. The 'region' variable can have 3 possible values: 'other', 'west' or 'central' and the 'smsa' can have 'yes' or 'no'. The dataset also contains the 'year' variable which can 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 relation with the target.

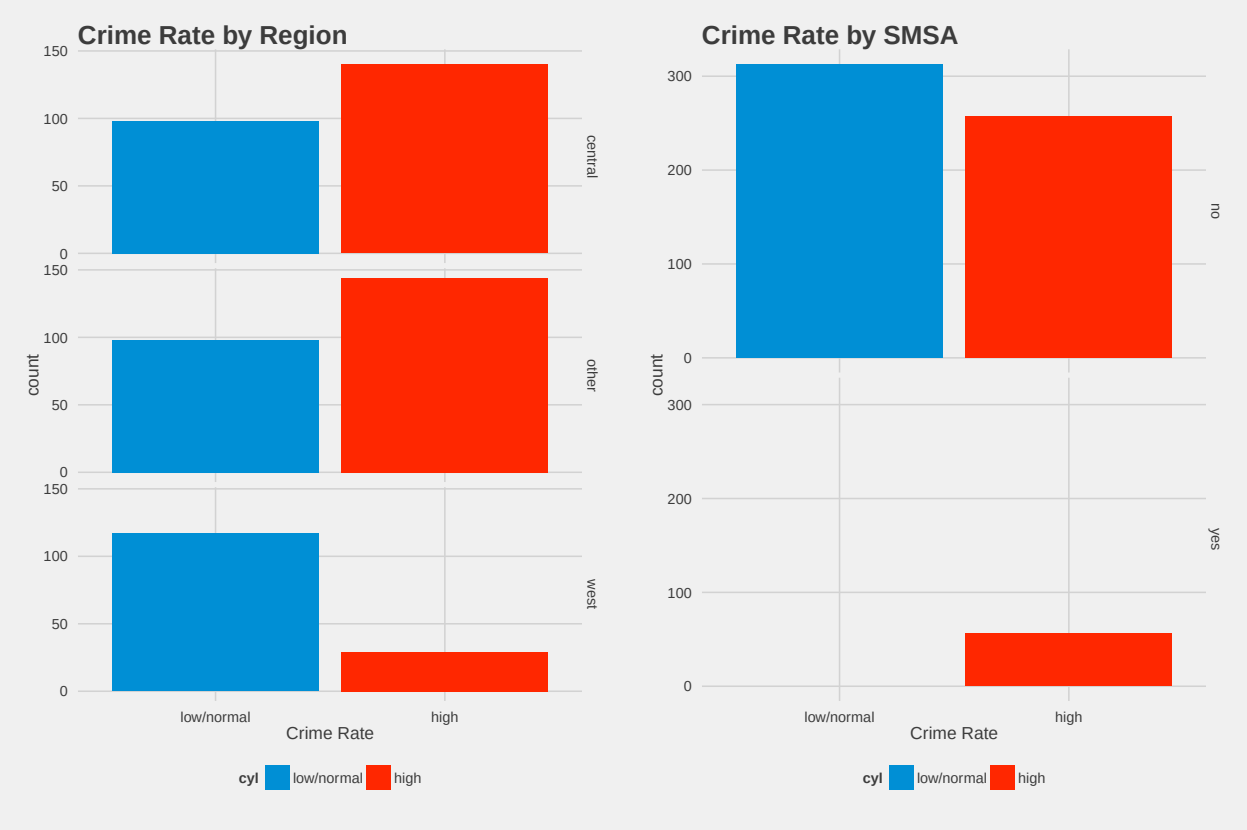

prbarr
prbarr
1.25
1.00
0.75
0.50
0.25
low/normal
high
crmrte_cat
prbconv
prbconv
30
20
10
0
low/normal
high
crmrte_cat
prbpris
prbpris
0.7
0.6
0.5
0.4
0.3
0.2
low/normal
high
crmrte_cat
avgsen
avgsen
25
20
15
10
5
low/normal
high
crmrte_cat
polpc
polpc
0.03
0.02
0.01
0.00
low/normal
high
crmrte_cat
density
density
7.5
5.0
2.5
0.0
low/normal
high
crmrte_cat
taxpc
taxpc
125
100
75
50
25
low/normal
high
crmrte_cat
pctmin
pctmin
60
40
20
0
low/normal
high
crmrte_cat
wcon
wcon
2000
1500
1000
500
0
low/normal
high
crmrte_cat
wtuc
wtuc
3000
2000
1000
0
low/normal
high
crmrte_cat
wfir
wfir
500
400
300
200
100
0
low/normal
high
crmrte_cat
wser
wser
2000
1500
1000
500
0
low/normal
high
crmrte_cat



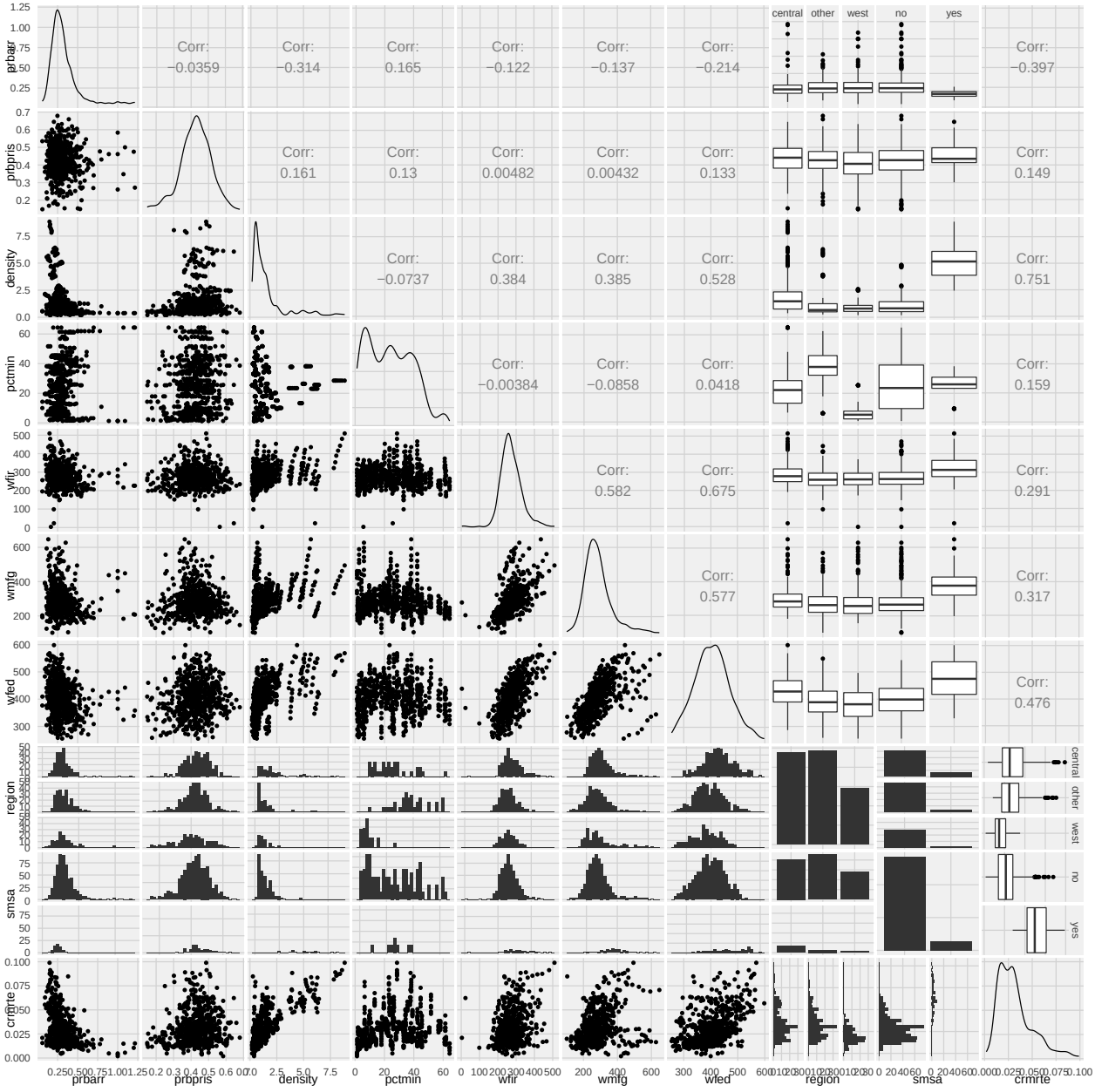
From the boxplot above, we can see that the variables that may have a predictive value with the target are variable 'prbarr', 'density', 'pctmin', 'wfed', 'wmfg' and 'pctymle' as they separate partialy the populations by the value of the target variable defined. In regards to the rest of the predictors, we explore them tracing these folowing charts. We start from the variable 'year'.



We comment on avobe plot that there is no sifnificance trend on the crime rate along the time line considered. The other two variables with categorical values are 'region' and 'smsa'.



In these two charts above we see an decrease in the crime rate when the variable 'region' takes value 'west' and when the 'smsa' variable takes the value 'yes'. In this sense, we continue to explore further the relationtship between these two categorical variables and the target variable by implementing method ANOVA in next section, but first we analize the variances and covariances between all predictors. We trace a paired graph with some selected variables to preliminary explore the correlation between the variables.



In the graphs avobe we can find the correlations between the selected predictors and the correlations between the predictors and the target. The highest value of correlation is given by the target and 'density' variable, which is very positive for building predicting models. Another high values of correlation are among variables 'wmfg', 'wfed' and 'density'. These high correlations will be analized further when we cover the variables selection for modelling.

ANOVA models

In the first analisys, we consider the variable 'region' and agregate their possible values ('west', 'central' and 'other') into ('w' and 'nw'). 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 result show a very low p-value, which means that the model considerign diferent populations is acurate. We can compare the means of the crime rate between the 'west' region and other regions. The hypotesis null that the median are equals is rejected.

```
##         nw          w
## 0.03475108 0.01987887
```

Considerign above analysis, we can assume that a model can shape correclty the value of the median in each region: ('west', 'other'). The coefficient of the model can be extacted from the fit value retuned in the package.

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

The model is given by

Mean(crmrte) = 0.03475108 - 0.01487221*I (region='w')

Now, we repeat the same analisis considering two factors. We include also the other categorical variable: 'smsa'. We fit an ANOVA model and we get the followign 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-valio for each of the two variables, which means that both factors have a strong relation with the variable. The hypotesis null that the medians are equals is rejected. The coeficients in this case are

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

Besides the two categorical variables, we can include in the analysis of the variance the interaction efect 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 indicates that the two factors and the interaction between them are significant.

Confidence Interval for the Median

As discussed above, we defined a target variable in a categorical format ('high', 'low/normal') if the value of the crime rate was higher that the median to be able to run, in next sections, models that require it that way. Therefore, it would be very usefull to have confidence intervals about the median. The median value for the crime rate is calculated as follows.

```
## [1] 0.02840405
```

We can simply obtain a confidence interval arround that value. Considering the binomial probabilty with n = 626 observations and a probability of 0.5, we want to obtain the kth than returns the 98%, the 95%, the 88% and the 81% of the probability by doing the following

1- 2*pbinom(k, 626, 0.5)

the kth values that returns that interval are k=

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

From the vector of sorted values of crime rate, we select the kth elements of the vector and the n-k+1 elements correspondigly to have the four confidence intervals correspondingly.

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

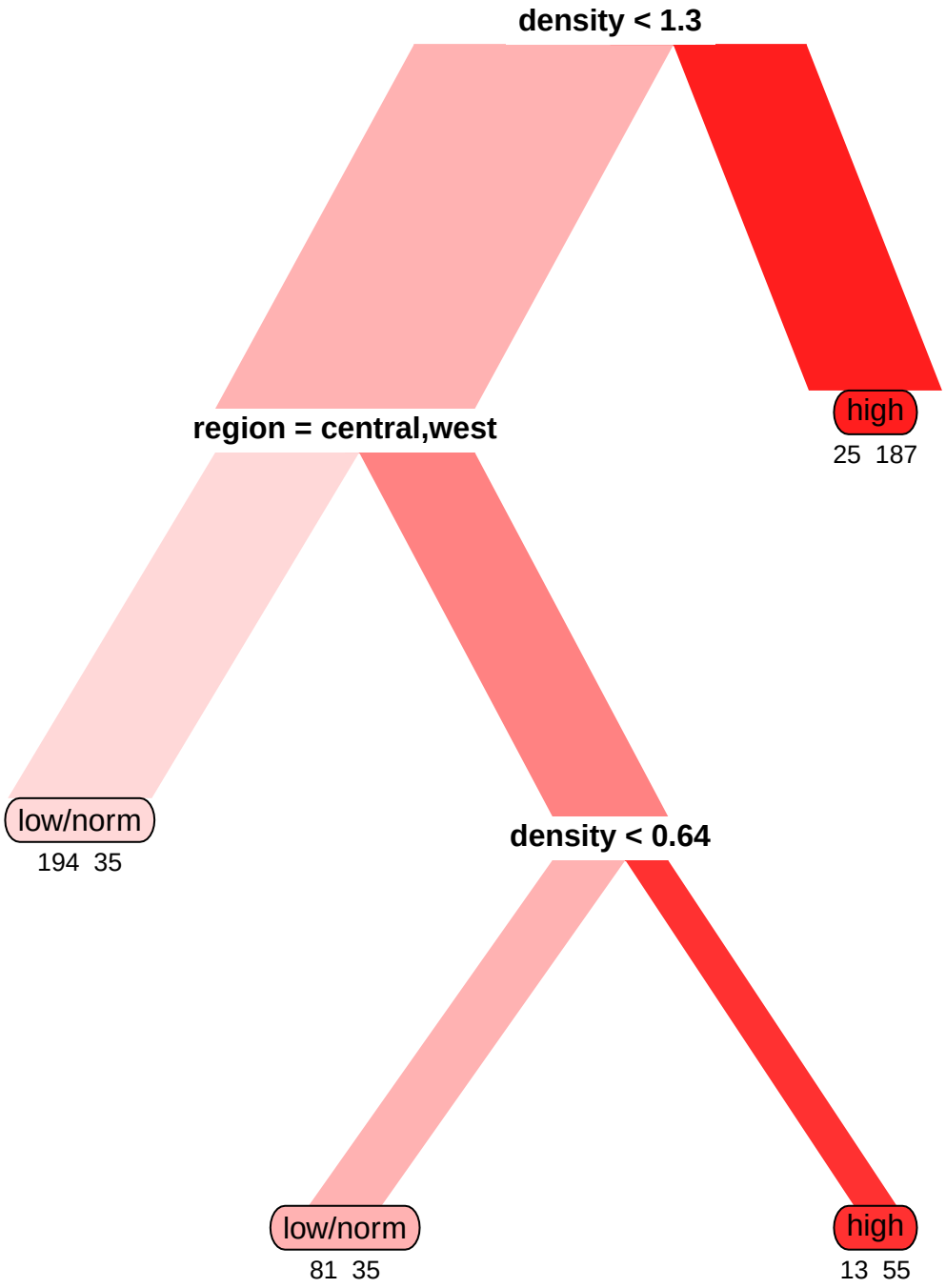
A more sofisticated method to obtain the confidence interval for the median is the Wilcoxon Signed Rank Tests for the same confidence levels to get the confidence intervals. As the Wilcoxon test assumes symetry in the distribution of the variable we applied it to the logarithm of the variable target, as discussed avobe.

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.

Predictive Models to Analize the dependence between the variables

To complete the analysis of the variables and their relation with the target we run two predictive models: a decision tree and a logistic regression. We consider the variable target in the categorical format. The aim of these modeling is explore further the data and understand which variables are relevant.




```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   low/normal high
##   low/normal         96   29
##   high               13  111
##
##                Accuracy : 0.8313
##                  95% CI : (0.7789, 0.8756)
##     No Information Rate : 0.5622
##     P-Value [Acc > NIR] : < 2e-16
```

```
##
##                   Kappa : 0.6628
##  Mcnemar's Test P-Value : 0.02064
##
##             Sensitivity : 0.8807
##             Specificity : 0.7929
##          Pos Pred Value : 0.7680
##          Neg Pred Value : 0.8952
##              Prevalence : 0.4378
##          Detection Rate : 0.3855
##    Detection Prevalence : 0.5020
##       Balanced Accuracy : 0.8368
##
##        'Positive' Class : low/normal
##
```

We get an accuracy of 81% in testing and we verify that the variables most relevant to the target are: region and density.