Report

Carlos Espino

December 12, 2015

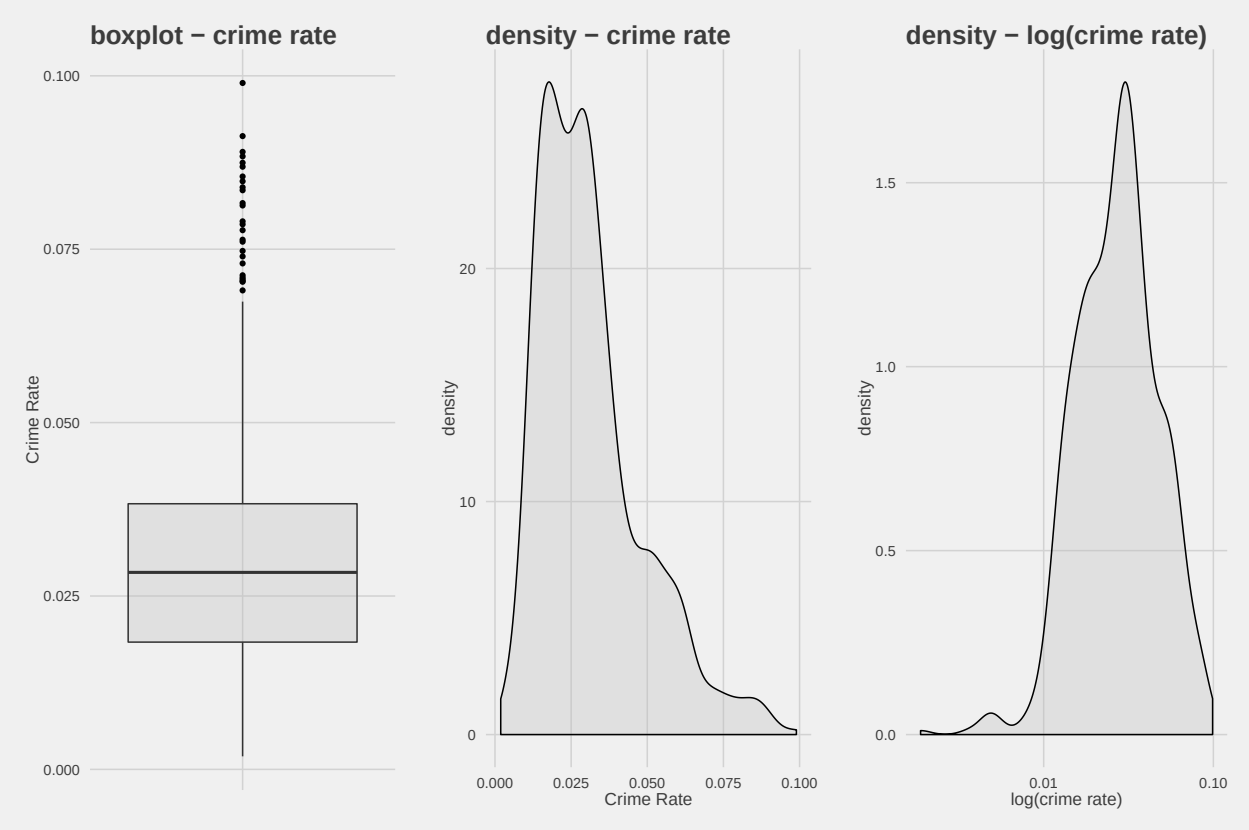
Introduction

Dataset

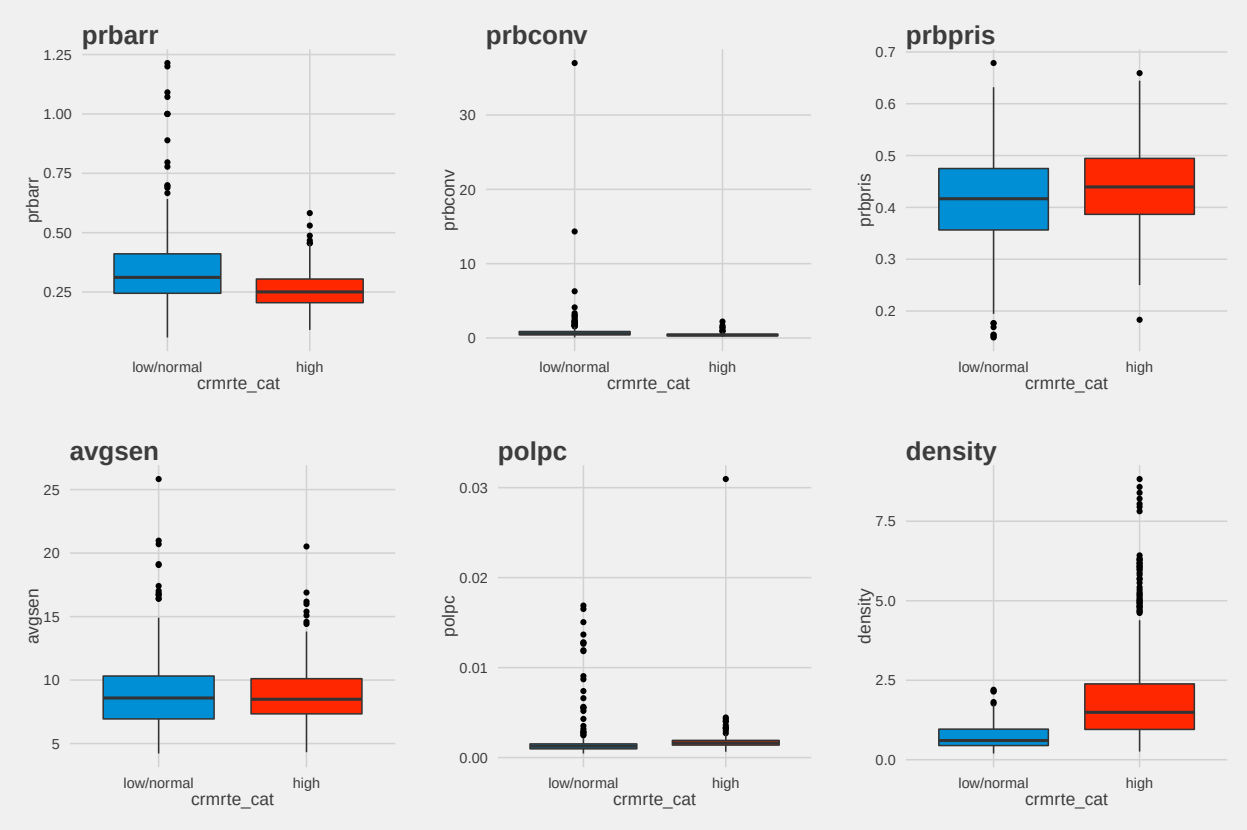
The dataset

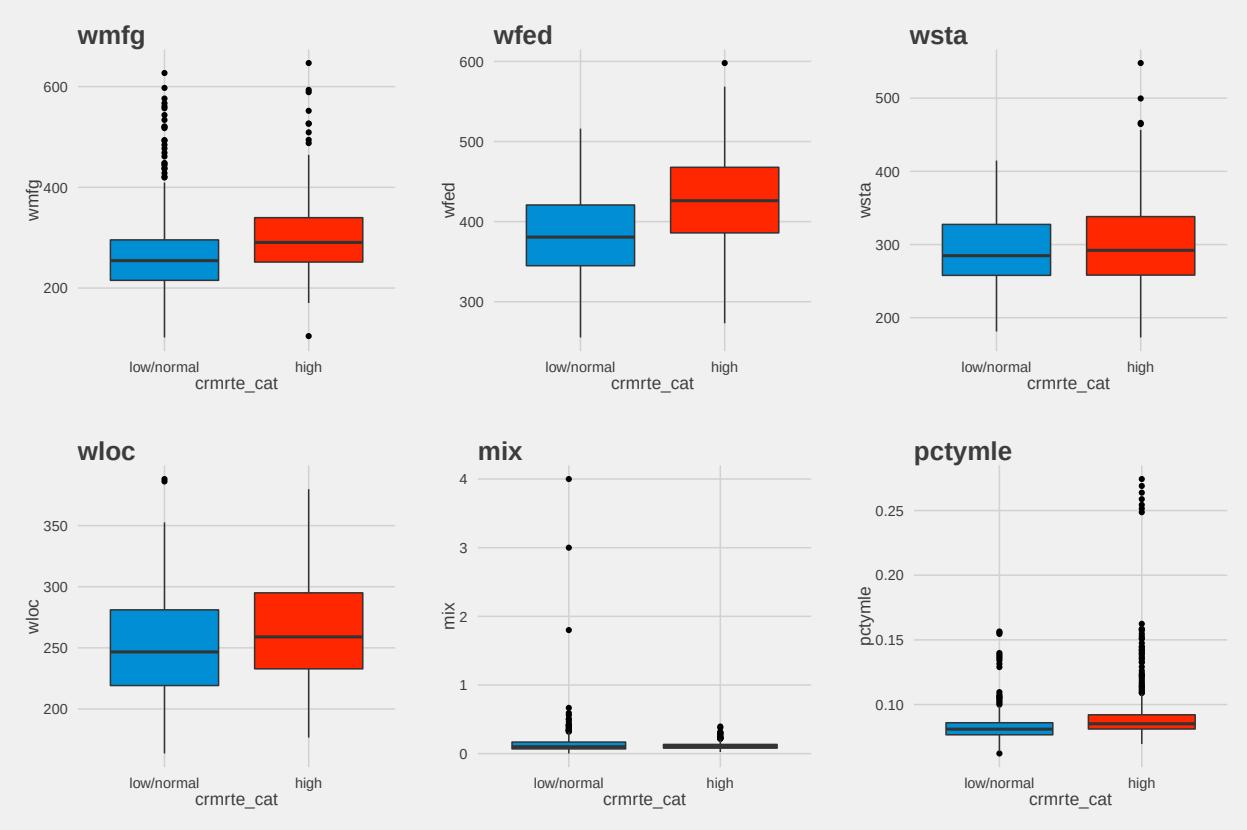
county	county identifier
year	year from 1981 to 1987
crmrte	crimes committed per person
prbarr	'probability' of arrest
prbconv	'probability' of conviction
prbpris	'probability' of prison sentenc
avgsen	average sentence, days
polpc	police per capita
density	people per square mile
taxpc	tax revenue per 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. The target variable behaviour is represented with a boxplot, a softened histogram and a softened histogram of the logarithm of the variable.

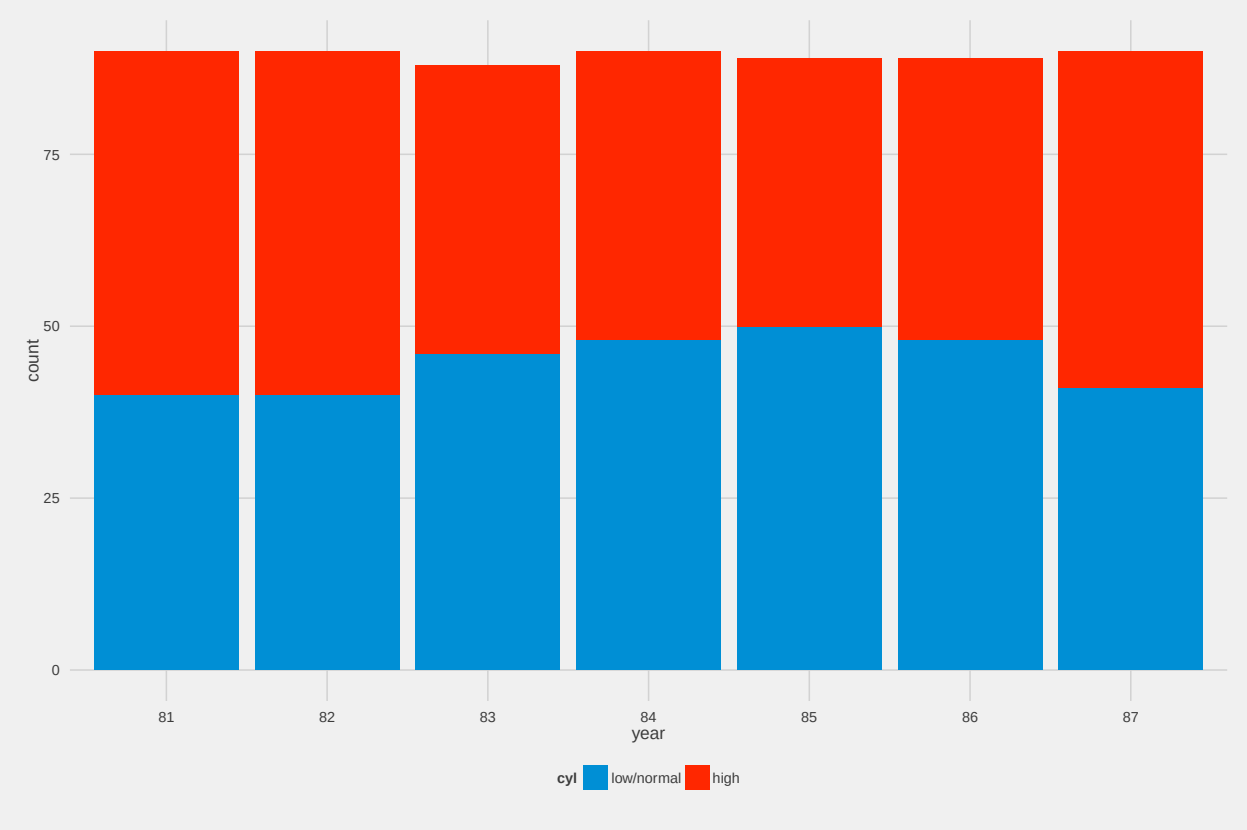


Besides the variable target, the dataset contains other 20 variables we used as predictors. Two of them have categorical values. The 'region' variable can have 3 possible values: 'other', 'west' or 'central' and the 'smsa' can have 'yes' or 'no'. The dataset also contains the 'year' variable which can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

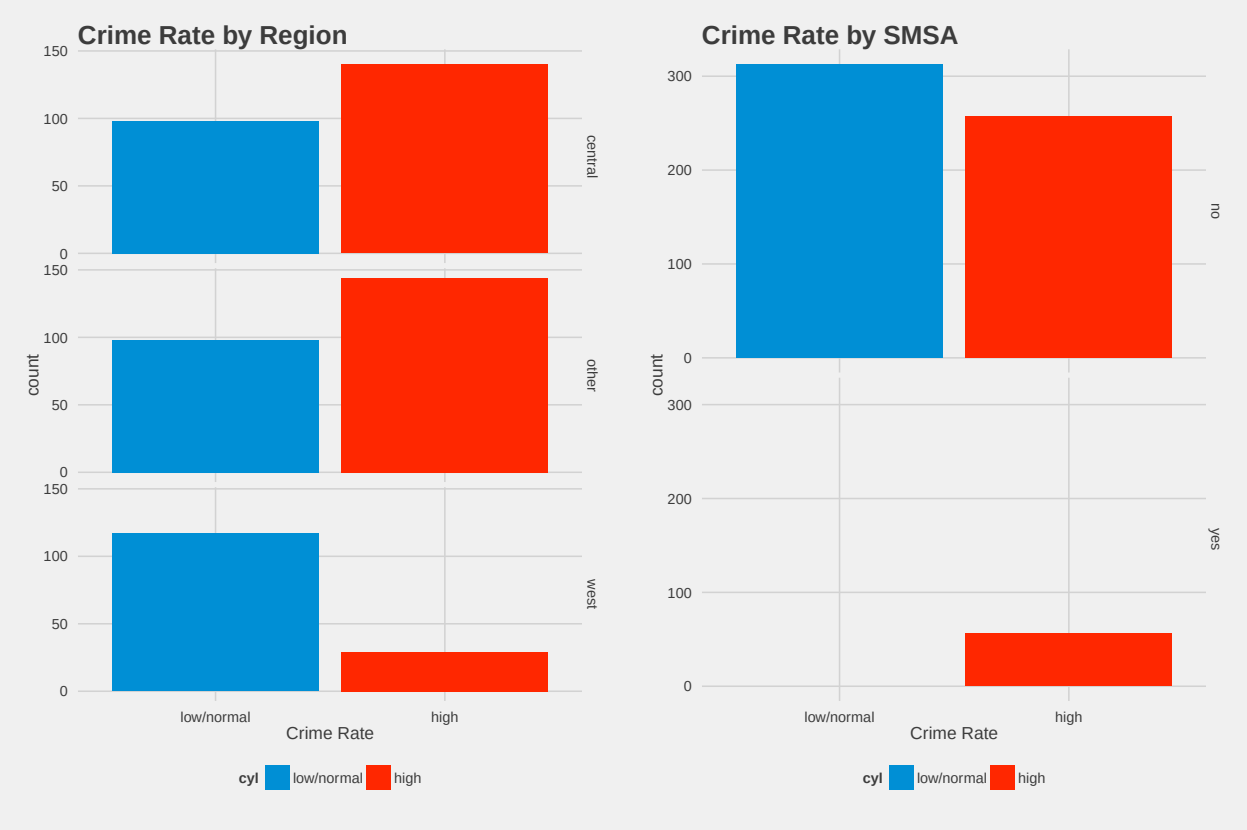

prbarr
prbarr
1.25
1.00
0.75
0.50
0.25
low/normal
high
crmrte_cat
prbconv
prbconv
30
20
10
0
low/normal
high
crmrte_cat
prbpris
prbpris
0.7
0.6
0.5
0.4
0.3
0.2
low/normal
high
crmrte_cat
avgsen
avgsen
25
20
15
10
5
low/normal
high
crmrte_cat
polpc
polpc
0.03
0.02
0.01
0.00
low/normal
high
crmrte_cat
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 chart 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. We fit an analysis of variance model by call to a linear regression for each stratum.

In the first analisys, we consider the variable 'region' and separate 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.

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

Now, we can repeat the same analisis considering two factors. We include the other categorical variable: 'smsa'

```
##              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 variables.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.