

Overview of “Boston” Dataset from “MASS” Package

Julian Hatwell

January 20, 2016

This document provides a brief overview of the Boston dataset in the MASS R package.

	crim	zn	indus	chas
##	Min. : 0.00632	Min. : 0.00	Min. : 0.46	Min. : 0.00000
##	1st Qu.: 0.08204	1st Qu.: 0.00	1st Qu.: 5.19	1st Qu.: 0.00000
##	Median : 0.25651	Median : 0.00	Median : 9.69	Median : 0.00000
##	Mean : 3.61352	Mean : 11.36	Mean : 11.14	Mean : 0.06917
##	3rd Qu.: 3.67708	3rd Qu.: 12.50	3rd Qu.: 18.10	3rd Qu.: 0.00000
##	Max. : 88.97620	Max. : 100.00	Max. : 27.74	Max. : 1.00000
	nox	rm	age	dis
##	Min. : 0.3850	Min. : 3.561	Min. : 2.90	Min. : 1.130
##	1st Qu.: 0.4490	1st Qu.: 5.886	1st Qu.: 45.02	1st Qu.: 2.100
##	Median : 0.5380	Median : 6.208	Median : 77.50	Median : 3.207
##	Mean : 0.5547	Mean : 6.285	Mean : 68.57	Mean : 3.795
##	3rd Qu.: 0.6240	3rd Qu.: 6.623	3rd Qu.: 94.08	3rd Qu.: 5.188
##	Max. : 0.8710	Max. : 8.780	Max. : 100.00	Max. : 12.127
	rad	tax	ptratio	black
##	Min. : 1.000	Min. : 187.0	Min. : 12.60	Min. : 0.32
##	1st Qu.: 4.000	1st Qu.: 279.0	1st Qu.: 17.40	1st Qu.: 375.38
##	Median : 5.000	Median : 330.0	Median : 19.05	Median : 391.44
##	Mean : 9.549	Mean : 408.2	Mean : 18.46	Mean : 356.67
##	3rd Qu.: 24.000	3rd Qu.: 666.0	3rd Qu.: 20.20	3rd Qu.: 396.23
##	Max. : 24.000	Max. : 711.0	Max. : 22.00	Max. : 396.90
	lstat	medv		
##	Min. : 1.73	Min. : 5.00		
##	1st Qu.: 6.95	1st Qu.: 17.02		
##	Median : 11.36	Median : 21.20		
##	Mean : 12.65	Mean : 22.53		
##	3rd Qu.: 16.95	3rd Qu.: 25.00		
##	Max. : 37.97	Max. : 50.00		

From the summary, and the associated help (not shown), the following observations can be made:

The dataframe contains 506 rows and 14 columns.

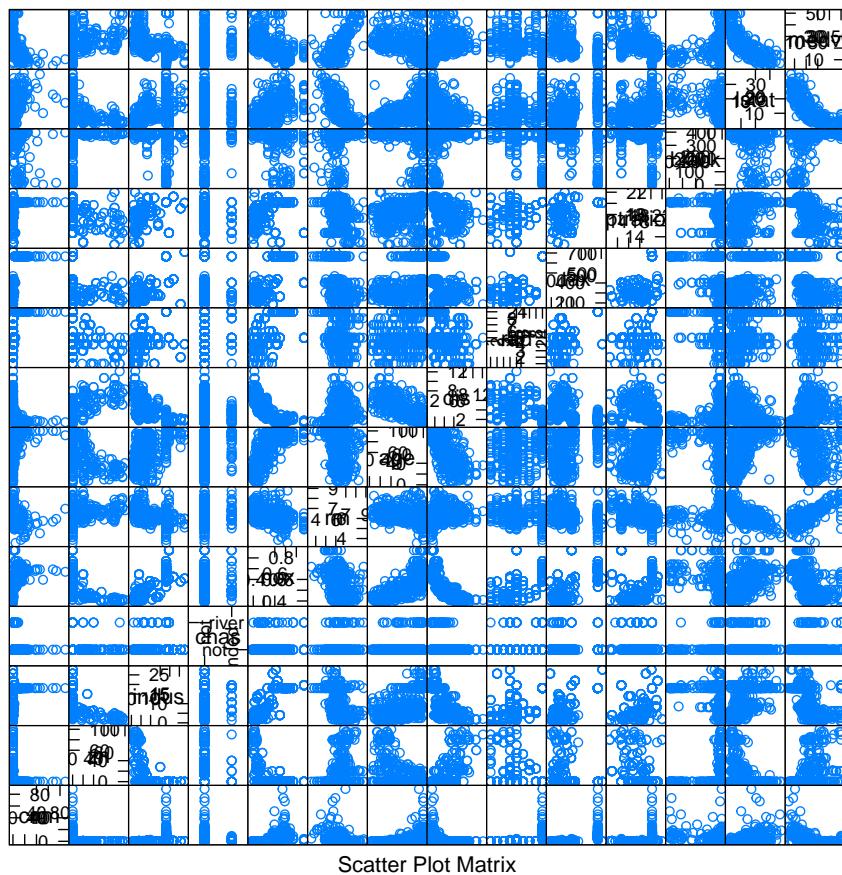


Figure 1: multi-variate comparisons

```

## 
## Welch Two Sample t-test
## 
## data:  crim by chas
## t = 3.2224, df = 120.42, p-value = 0.001636
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7298305 3.0557226
## sample estimates:
## mean in group not mean in group river bounded
## 3.744447 1.851670

```

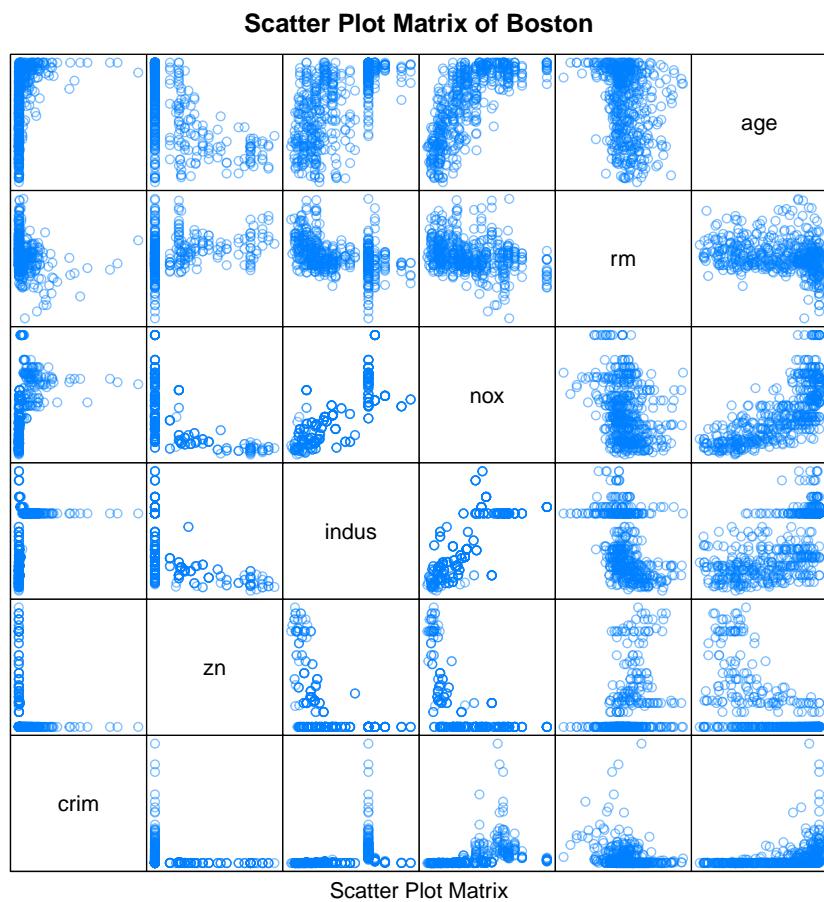


Figure 2: multi-variate comparisons

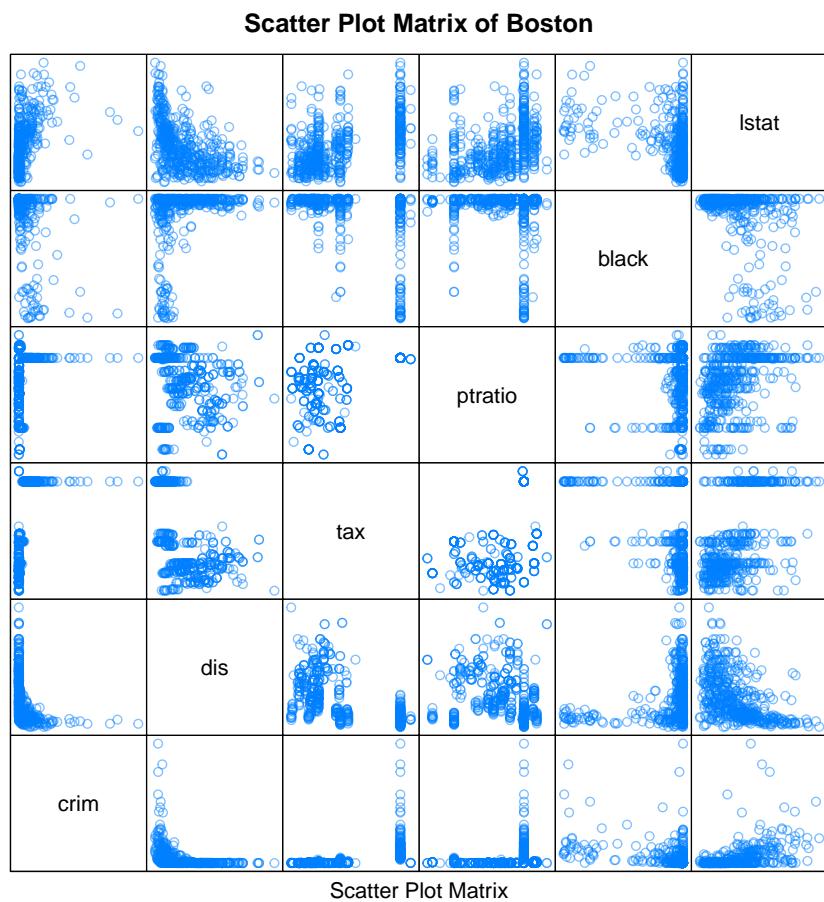


Figure 3: multi-variate comparisons

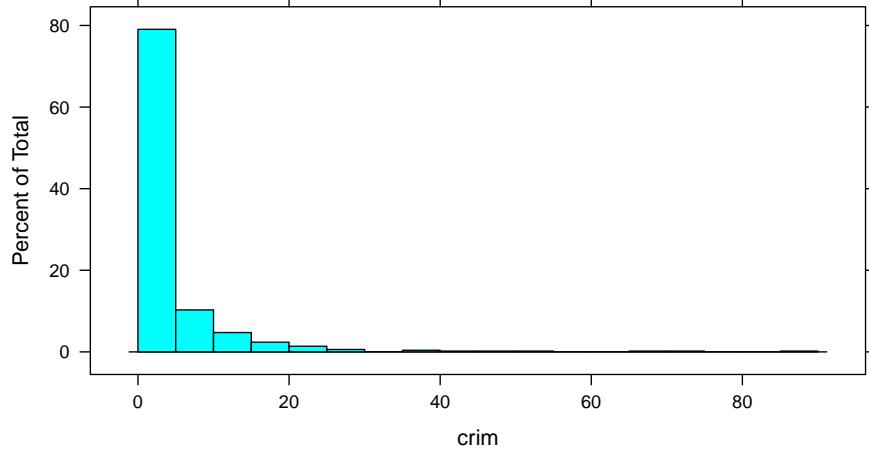


Figure 4: Histogram of the crim variable

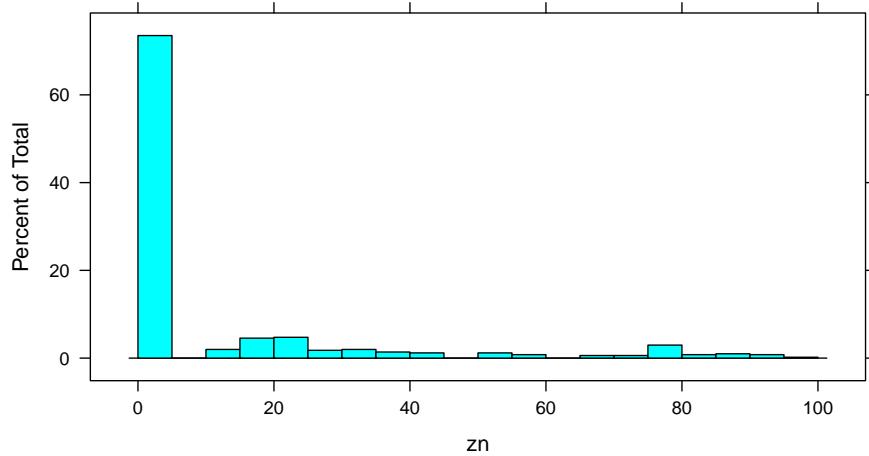


Figure 5: Histogram of the zn variable

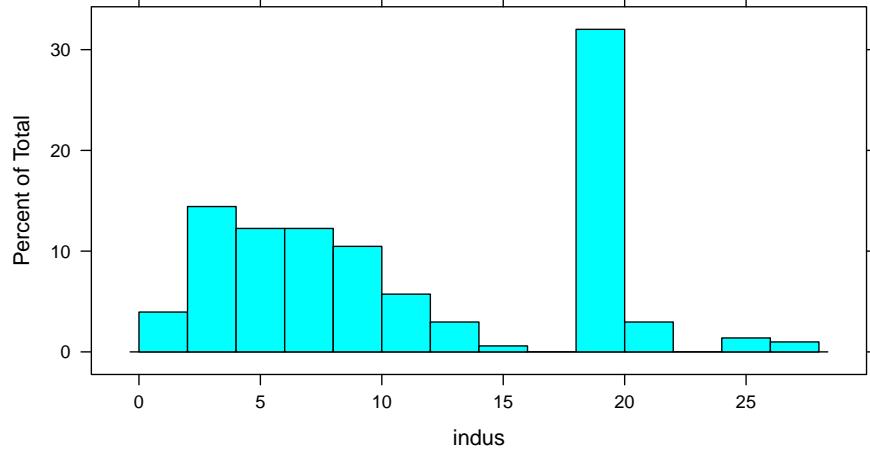


Figure 6: Histogram of the indus variable

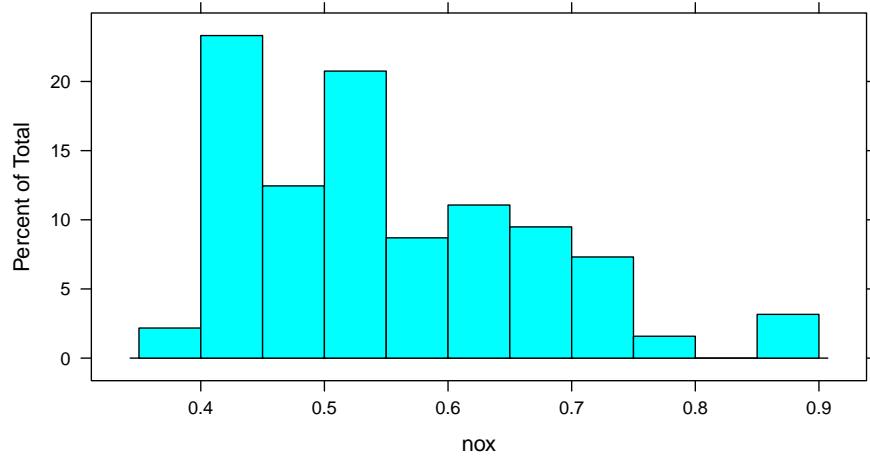


Figure 7: Histogram of the nox variable

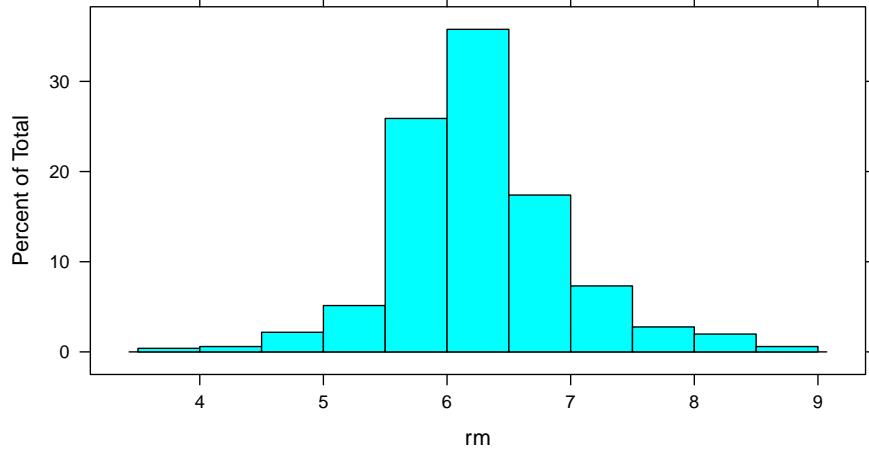


Figure 8: Histogram of the `rm` variable

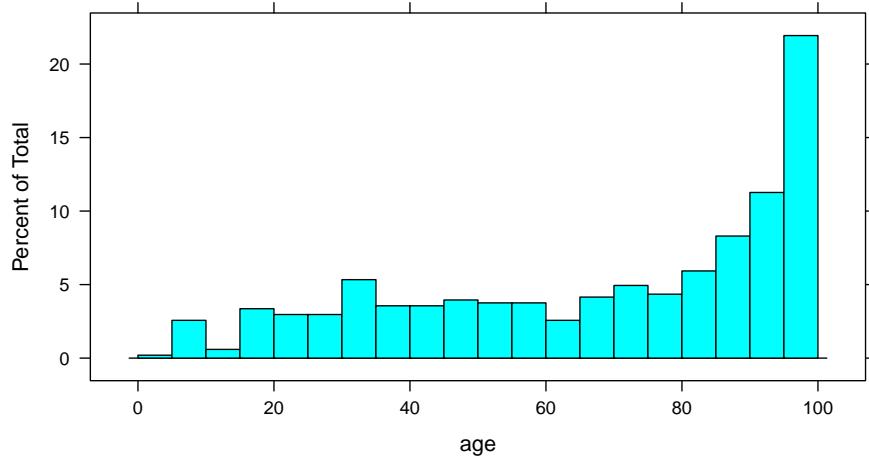


Figure 9: Histogram of the `age` variable

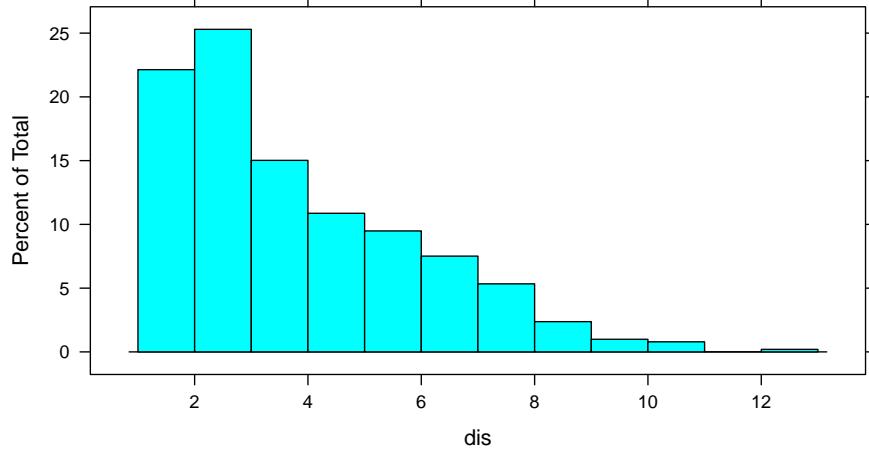


Figure 10: Histogram of the dis variable

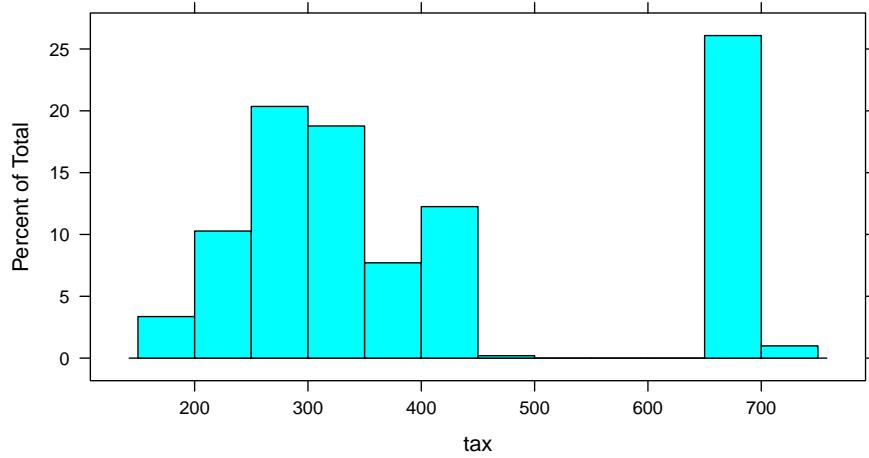


Figure 11: Histogram of the tax variable

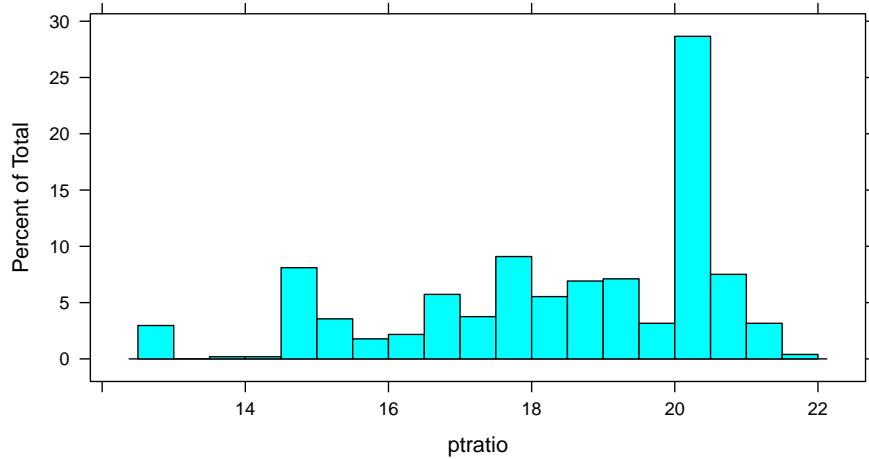


Figure 12: Histogram of the pptratio variable

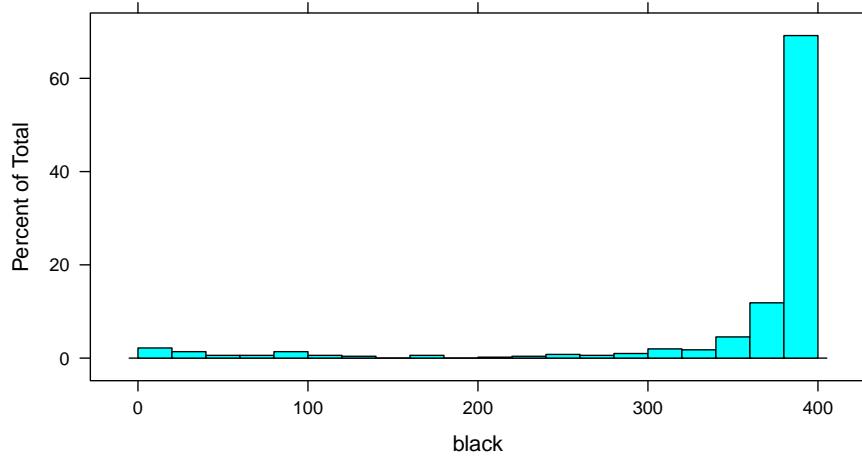


Figure 13: Histogram of the black variable

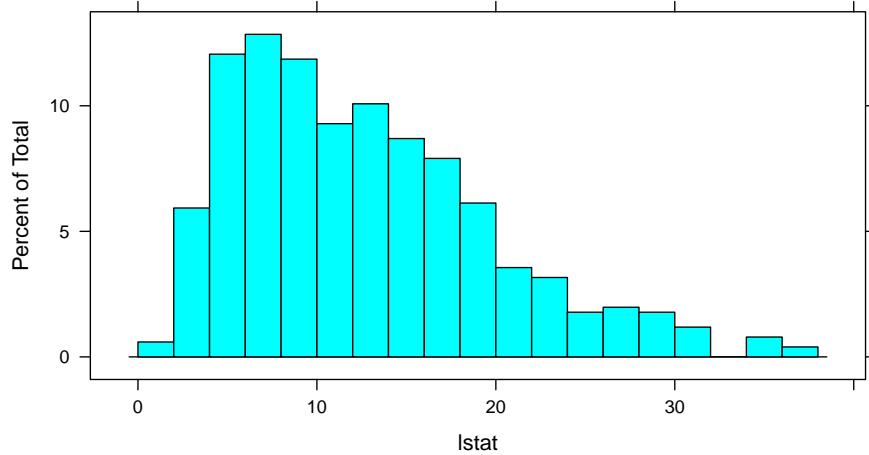


Figure 14: Histogram of the `lstat` variable

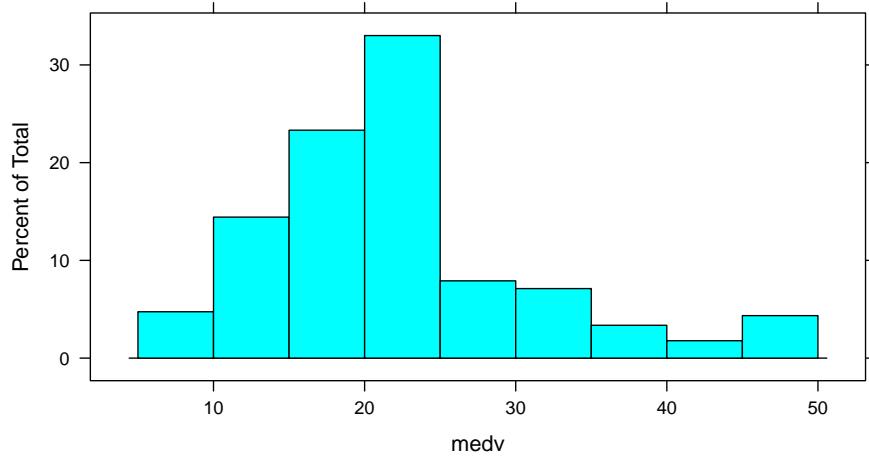


Figure 15: Histogram of the `medv` variable

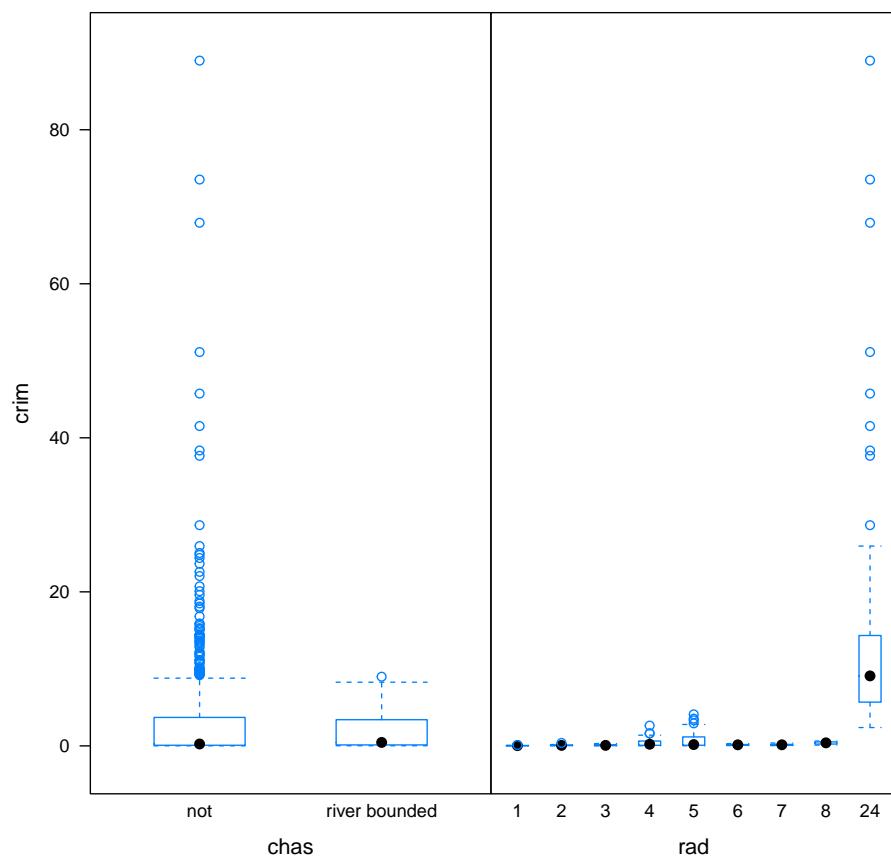


Figure 16: Boxplot of the dependent variable `crim` by each factor variable

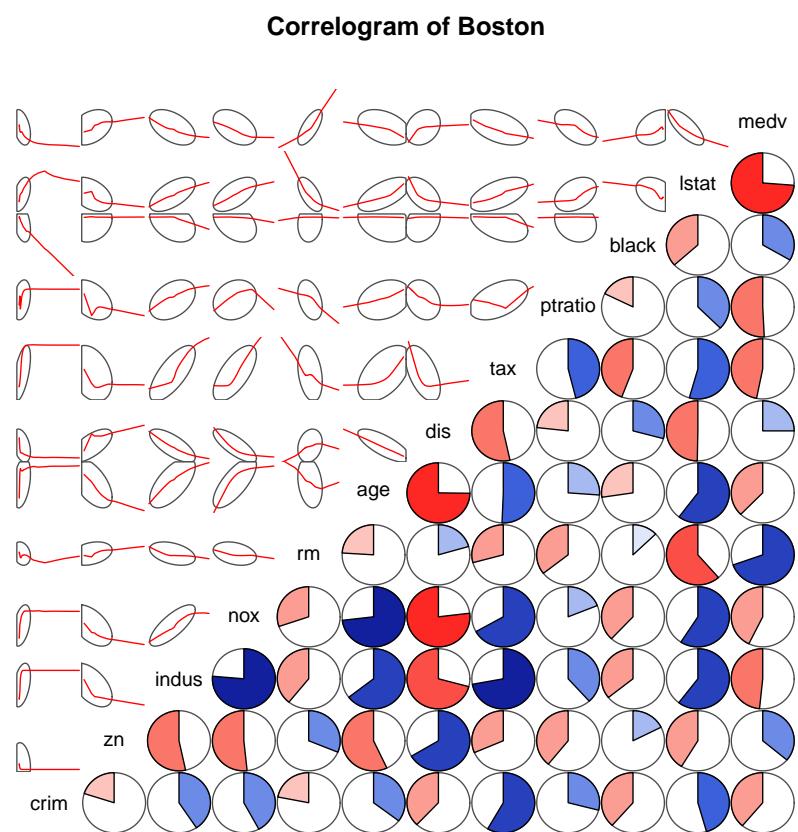


Figure 17: Correlogram

```

## 
## Call:
## lm(formula = fmla1, data = df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -4.429 -4.222 -2.620  1.250 84.523 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.45369   0.41722 10.675 < 2e-16 ***
## zn          -0.07393   0.01609 -4.594 5.51e-06 ***  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 8.435 on 504 degrees of freedom
## Multiple R-squared:  0.04019, Adjusted R-squared:  0.03828 
## F-statistic: 21.1 on 1 and 504 DF,  p-value: 5.506e-06 
## 
## 
## Call:
## lm(formula = fmla1, data = df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -11.972 -2.698 -0.736  0.712 81.813 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.06374   0.66723 -3.093  0.00209 **  
## indus        0.50978   0.05102  9.991 < 2e-16 ***  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 7.866 on 504 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1637 
## F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16 
## 
## 
## Call:
## lm(formula = fmla1, data = df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -12.371 -2.738 -0.974  0.559 81.728 

```

```

## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -13.720     1.699  -8.073 5.08e-15 ***
## nox          31.249     2.999  10.419 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.81 on 504 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.1756 
## F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16 
## 
## 
## Call:
## lm(formula = fmla1, data = df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max  
## -6.604 -3.952 -2.654  0.989 87.197 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  20.482     3.365   6.088 2.27e-09 ***
## rm           -2.684     0.532  -5.045 6.35e-07 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.401 on 504 degrees of freedom
## Multiple R-squared:  0.04807, Adjusted R-squared:  0.04618 
## F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07 
## 
## 
## Call:
## lm(formula = fmla1, data = df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max  
## -6.789 -4.257 -1.230  1.527 82.849 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
## age          0.10779    0.01274   8.463 2.85e-16 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

## 
## Residual standard error: 8.057 on 504 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.1227
## F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.708 -4.134 -1.527  1.516 81.674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  9.4993    0.7304 13.006 <2e-16 ***
## dis         -1.5509    0.1683 -9.213 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.965 on 504 degrees of freedom
## Multiple R-squared:  0.1441, Adjusted R-squared:  0.1425
## F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.513 -2.738 -0.194  1.065 77.696
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -8.528369   0.815809 -10.45 <2e-16 ***
## tax          0.029742   0.001847   16.10 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.997 on 504 degrees of freedom
## Multiple R-squared:  0.3396, Adjusted R-squared:  0.3383
## F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
##
##
## Call:

```

```

## lm(formula = fmla1, data = df)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -7.654 -3.985 -1.912  1.825 83.353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.6469    3.1473 -5.607 3.40e-08 ***
## ptratio       1.1520    0.1694  6.801 2.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.24 on 504 degrees of freedom
## Multiple R-squared:  0.08407, Adjusted R-squared:  0.08225
## F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -13.756 -2.299 -2.095 -1.296 86.822
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.553529  1.425903 11.609 <2e-16 ***
## black        -0.036280  0.003873 -9.367 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.946 on 504 degrees of freedom
## Multiple R-squared:  0.1483, Adjusted R-squared:  0.1466
## F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -13.925 -2.822 -0.664  1.079 82.862
##
## Coefficients:

```

```

##             Estimate Std. Error t value Pr(>|t|) 
## (Intercept) -3.33054   0.69376 -4.801 2.09e-06 ***
## lstat       0.54880   0.04776 11.491 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.664 on 504 degrees of freedom
## Multiple R-squared:  0.2076, Adjusted R-squared:  0.206 
## F-statistic: 132 on 1 and 504 DF, p-value: < 2.2e-16
## 
## 
## Call:
## lm(formula = fmla1, data = df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -9.071 -4.022 -2.343  1.298 80.957 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) 
## (Intercept) 11.79654   0.93419 12.63   <2e-16 ***
## medv       -0.36316   0.03839 -9.46   <2e-16 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.934 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491 
## F-statistic: 89.49 on 1 and 504 DF, p-value: < 2.2e-16
## 
## Call:
## lm(formula = fmla, data = df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -12.902 -2.565 -0.337  1.172 76.819 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) 
## (Intercept) 1.996632   7.156885   0.279 0.780376 
## zn          0.031478   0.019410   1.622 0.105502 
## indus      -0.231197   0.082681  -2.796 0.005371 ** 
## nox         -4.736583   5.411823  -0.875 0.381875 
## rm          0.752370   0.636657   1.182 0.237873 
## age        -0.007888   0.018609  -0.424 0.671846 
## dis        -0.979380   0.293553  -3.336 0.000913 *** 

```

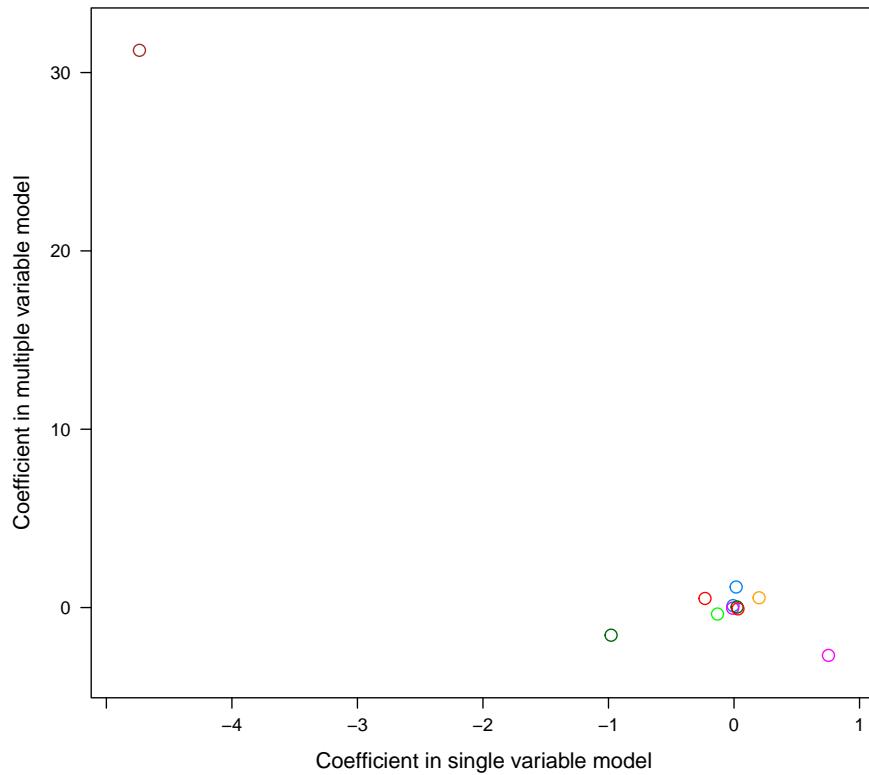
```

## tax          0.024448   0.003079   7.941 1.37e-14 ***
## ptratio      0.017028   0.189022   0.090 0.928255
## black       -0.010845   0.003793  -2.859 0.004426 **
## lstat        0.199854   0.078074   2.560 0.010770 *
## medv        -0.131571   0.061354  -2.144 0.032482 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.711 on 494 degrees of freedom
## Multiple R-squared:  0.4045, Adjusted R-squared:  0.3912
## F-statistic: 30.5 on 11 and 494 DF,  p-value: < 2.2e-16

```

Single vs Multivariate model parameters

age	○	indus	○	nox	○	tax	○
black	○	lstat	○	ptratio	○	zn	○
dis	○	medv	○	rm	○		



```

# recoded these 2 vars as factors
df <- df %>% mutate(chas = factor(chas, labels = c("not",
  "river bounded")), rad = factor(rad))

```