Overview of "prostate" Dataset from "faraway" Package

Julian Hatwell

January 3, 2016

This document provides a brief overview of the prostate dataset in the faraway R package.

```
##      lcavol           lweight          age              lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004  
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581  
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263  
##       svi              lcp             gleason          pgg45       
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
##  3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00  
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
##       lpsa        
##  Min.   :-0.4308  
##  1st Qu.: 1.7317  
##  Median : 2.5915  
##  Mean   : 2.4784  
##  3rd Qu.: 3.0564  
##  Max.   : 5.5829
```

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

The dataframe contains 97 rows and 9 columns.

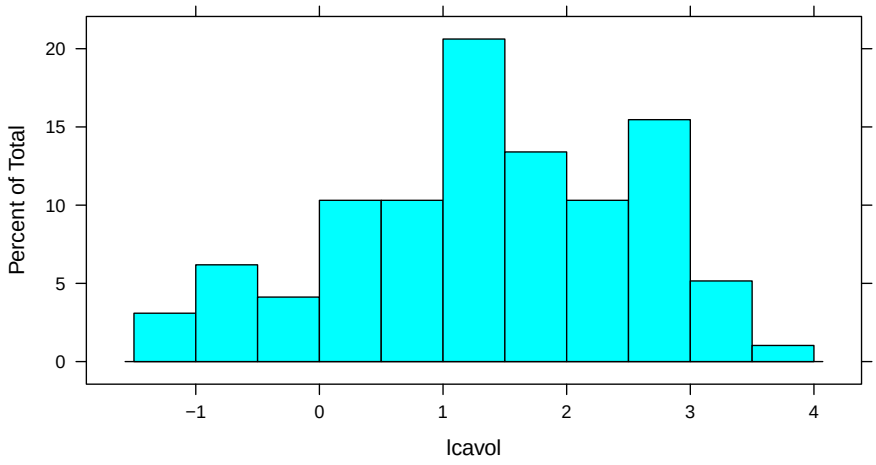

Figure 1: Histogram of the lcavol variable

```
##
##  Welch Two Sample t-test
##
## data:  lcavol by svi
## t = -8.0351, df = 51.172, p-value = 1.251e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.917326 -1.150810
## sample estimates:
##  mean in group no mean in group yes
##          1.017892          2.551959
```

```
# carry out any required transformations
# for tidy data standardisation, e.g. set
# factors correctly
df <- mutate(df, svi = factor(svi, labels = c("no",
    "yes")), gleason = factor(gleason))
```

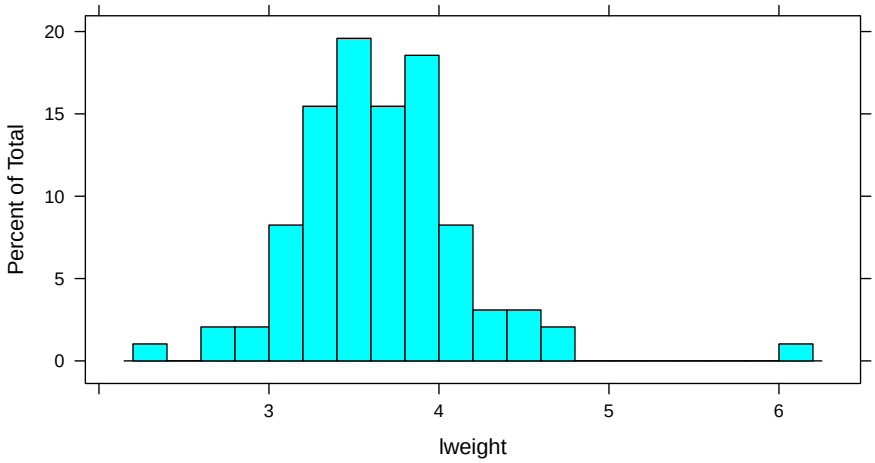

Figure 2: Histogram of the lweight variable

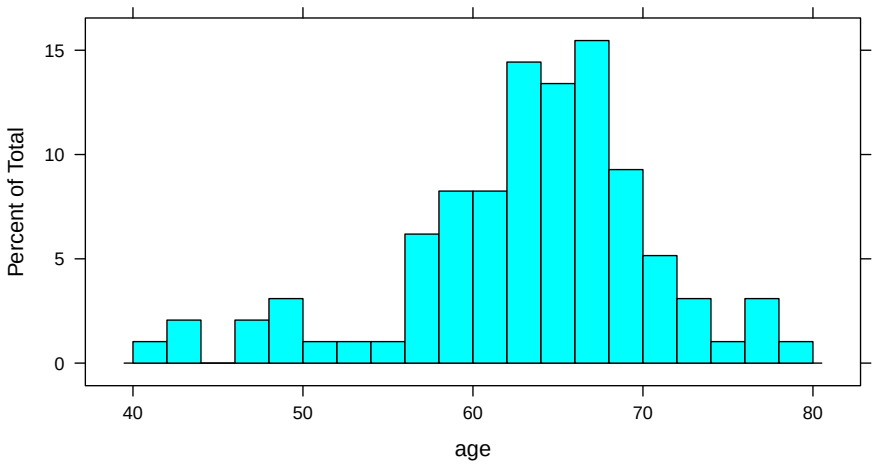

Figure 3: Histogram of the age variable

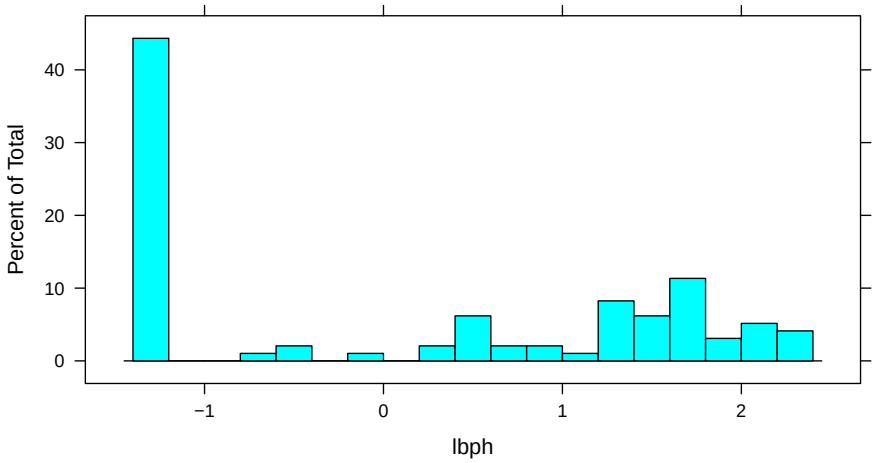

Figure 4: Histogram of the lbph variable

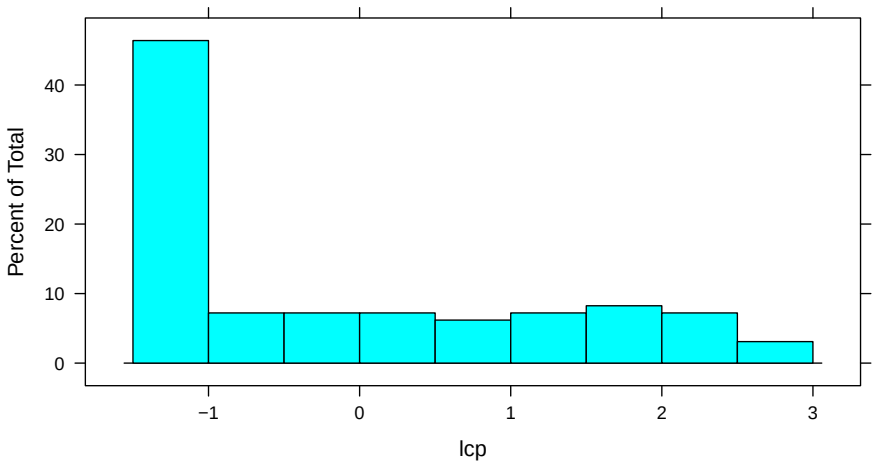

Figure 5: Histogram of the lcp variable

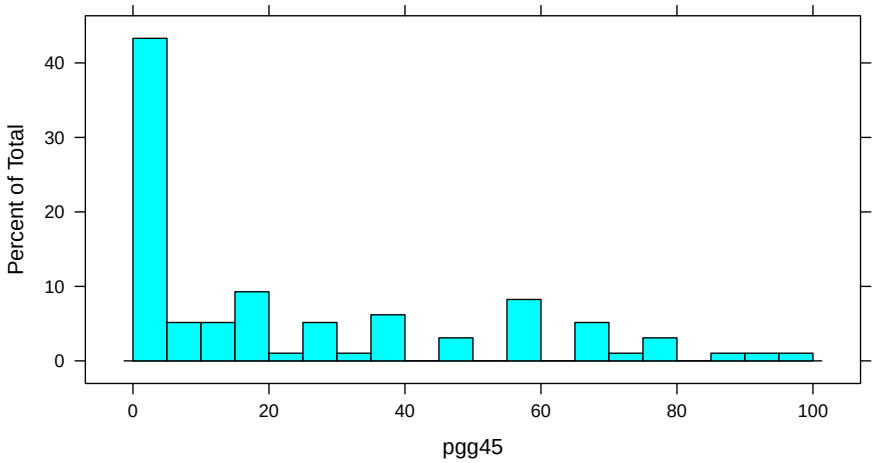

Figure 6: Histogram of the pgg45 variable

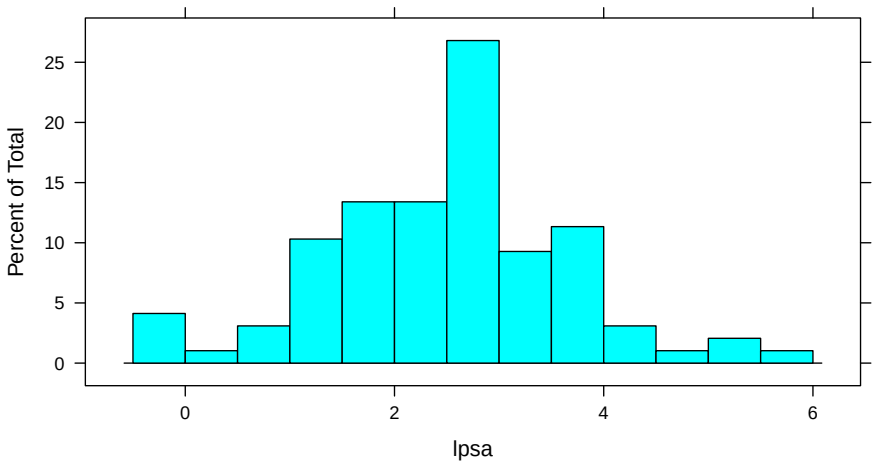

Figure 7: Histogram of the lpsa variable

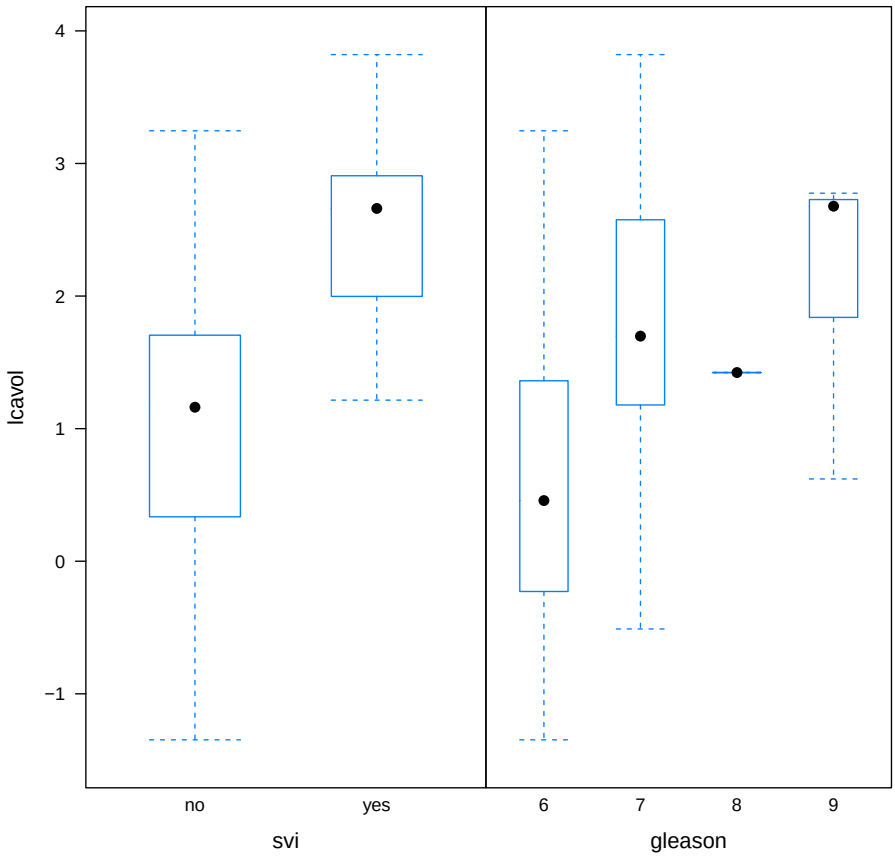

Figure 8: Boxplot of the dependent variable lcavol by each factor variable

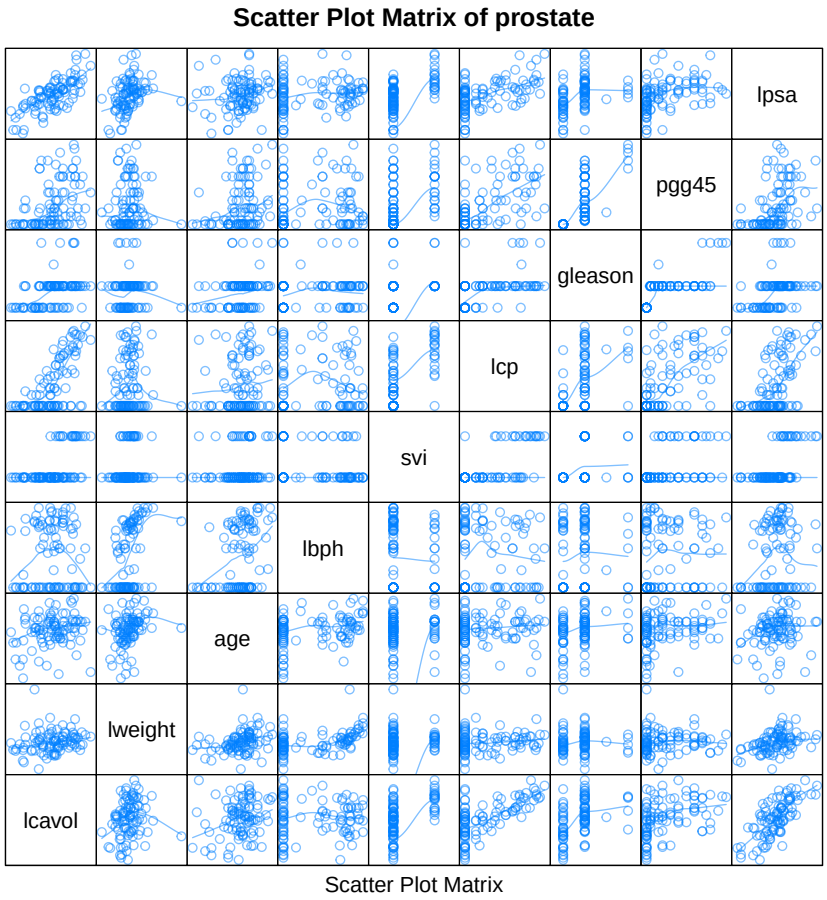

Figure 9: multi-variate comparisons

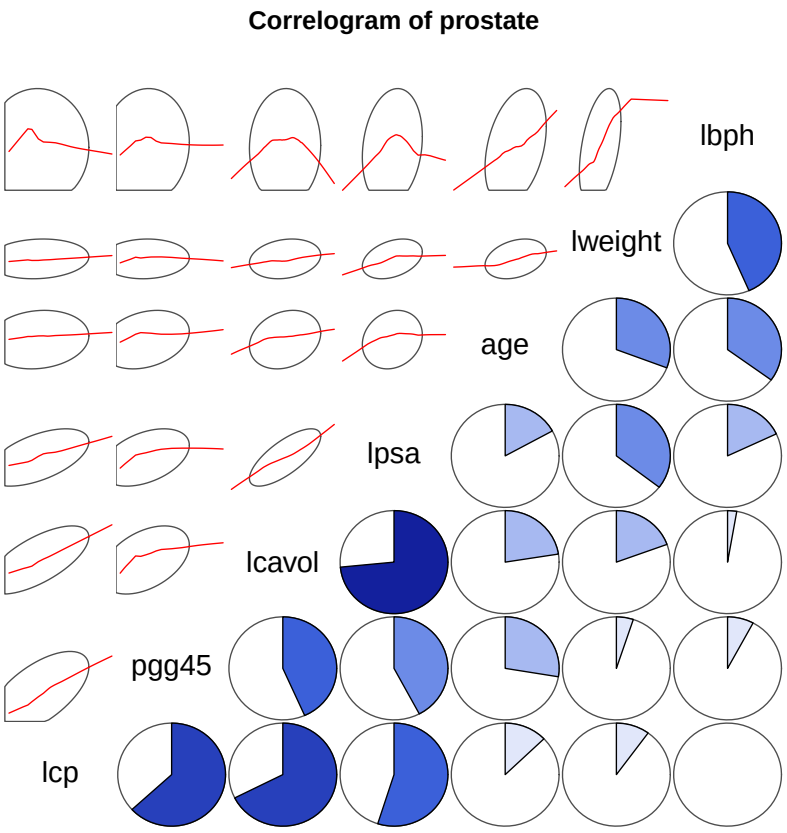

Figure 10: Correlogram