DA607 Week 02 Assignment

Dan Fanelli

February 4, 2016

Set your working directory to the same location as this Rmd file.

If you don't want to keep re-downloading the data files, put a # in front of the 2 download statements below:

```
#download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data", "./hou
#download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names", "./ho
```

Read that file into a data frame and show its str and dim:

```
initial.data.frame <- read.table("./housing.data", header=FALSE)

str(initial.data.frame)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ V1 : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ V2 : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ V3 : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ V4 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V5 : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ V6 : num  6.58 6.42 7.18 7 7.15 ...
##  $ V7 : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ V8 : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ V9 : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ V10: num  296 242 242 222 222 222 311 311 311 311 ...
##  $ V11: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ V12: num  397 397 393 395 397 ...
##  $ V13: num  4.98 9.14 4.03 2.94 5.33 ...
##  $ V14: num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```
dim(initial.data.frame)
```

```
## [1] 506  14
```

The following came from the HEADER file, but we'll manually rename these to be more user friendly:

7. Attribute Information:

    1. CRIM per capita crime rate by town
    2. ZN proportion of residential land zoned for lots over 25,000 sq.ft.
    3. INDUS proportion of non-retail business acres per town
    4. CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
    5. NOX nitric oxides concentration (parts per 10 million)
    6. RM average number of rooms per dwelling
    7. AGE proportion of owner-occupied units built prior to 1940
    8. DIS weighted distances to five Boston employment centres

9. RAD index of accessibility to radial highways
10. TAX full-value property-tax rate per $10,000
11. PTRATIO pupil-teacher ratio by town
12. B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT % lower status of the population
14. MEDV Median value of owner-occupied homes in $1000's

Now, lets set the col names exactly as they were in that header file:

```
colnames(initial.data.frame) <- c("CRIM", "ZN", "INDUS", "CHAS", "NOX", "RM", "AGE", "DIS", "RAD", "TAX
```

Now, lets subset the data by column name, and take the interesting ones:

```
cols.subset.data.frame <- initial.data.frame[,c("CRIM","AGE","TAX","MEDV")]

head(cols.subset.data.frame)
```

```
##      CRIM  AGE TAX MEDV
## 1 0.00632 65.2 296 24.0
## 2 0.02731 78.9 242 21.6
## 3 0.02729 61.1 242 34.7
## 4 0.03237 45.8 222 33.4
## 5 0.06905 54.2 222 36.2
## 6 0.02985 58.7 222 28.7
```

Finally, lets only look at the "20-Somethings" crowd:

```
rows.subset.data.frame <- cols.subset.data.frame[ which(cols.subset.data.frame$AGE>=20 & cols.subset.da

head(rows.subset.data.frame)
```

```
##       CRIM  AGE TAX MEDV
## 17 1.05393 29.3 307 23.1
## 40 0.02763 21.8 252 30.8
## 53 0.05360 21.1 243 25.0
## 54 0.04981 21.4 243 23.4
## 56 0.01311 21.9 226 35.4
## 59 0.15445 29.2 284 23.3
```

The End.