Homework 5

Nicholas Nagle

March 30, 2015

For this assigment you will use a dataset of housing prices in Boston. These data were used in an early publication in environmental economics to study the effect of air quality on housing price. You can get a copy of the data in the spdep R package. Don't forget to use `install.packages` if you need to!

```
library(spdep)
```

```
## Warning: package 'spdep' was built under R version 3.1.3
```

```
data(boston)
```

There is a codebook in the help file for this dataset

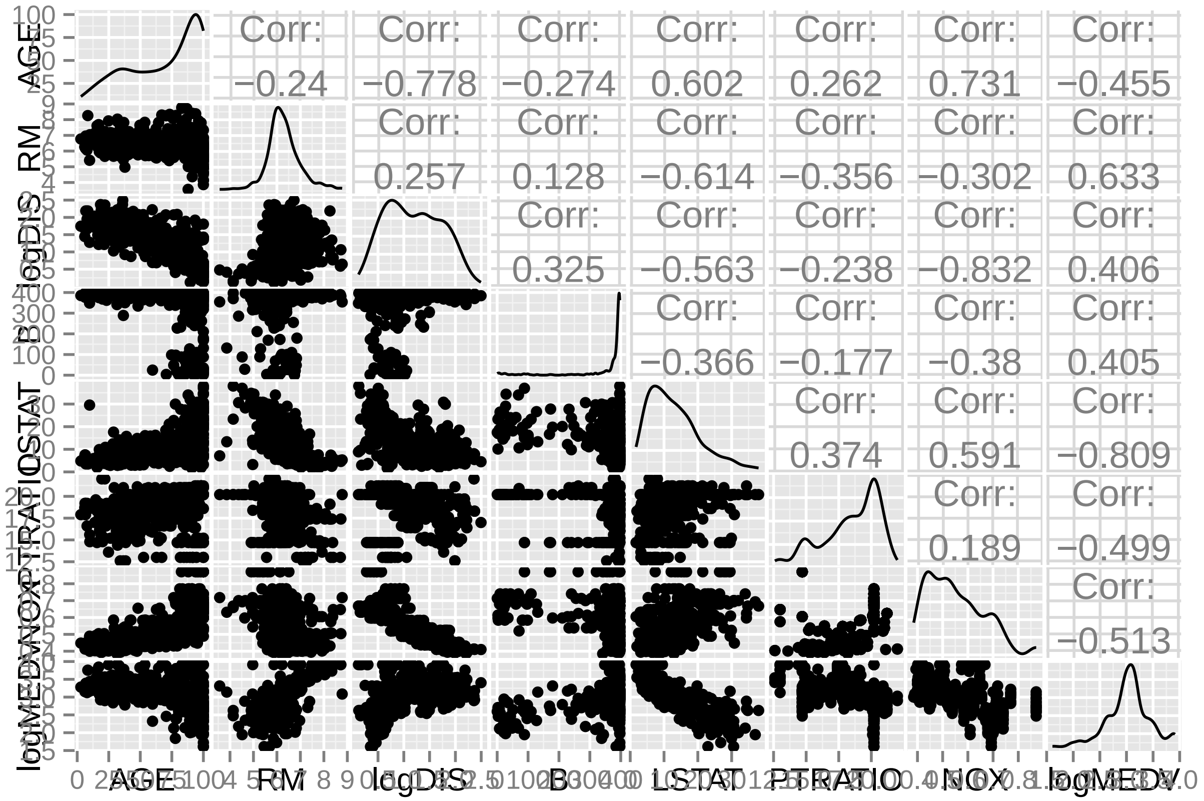
```
help(boston)
```

```
head(boston.c)
```

```
##         TOWN TOWNNO TRACT    LON   LAT MEDV CMEDV    CRIM ZN INDUS CHAS
## 1     Nahant      0  2011 -70.95 42.26 24.0  24.0 0.00632 18  2.31    0
## 2 Swampscott      1  2021 -70.95 42.29 21.6  21.6 0.02731  0  7.07    0
## 3 Swampscott      1  2022 -70.94 42.28 34.7  34.7 0.02729  0  7.07    0
## 4 Marblehead      2  2031 -70.93 42.29 33.4  33.4 0.03237  0  2.18    0
## 5 Marblehead      2  2032 -70.92 42.30 36.2  36.2 0.06905  0  2.18    0
## 6 Marblehead      2  2033 -70.92 42.30 28.7  28.7 0.02985  0  2.18    0
##     NOX    RM  AGE   DIS RAD TAX PTRATIO     B LSTAT
## 1 0.538 6.575 65.2 4.090   1 296    15.3 396.9  4.98
## 2 0.469 6.421 78.9 4.967   2 242    17.8 396.9  9.14
## 3 0.469 7.185 61.1 4.967   2 242    17.8 392.8  4.03
## 4 0.458 6.998 45.8 6.062   3 222    18.7 394.6  2.94
## 5 0.458 7.147 54.2 6.062   3 222    18.7 396.9  5.33
## 6 0.458 6.430 58.7 6.062   3 222    18.7 394.1  5.21
```

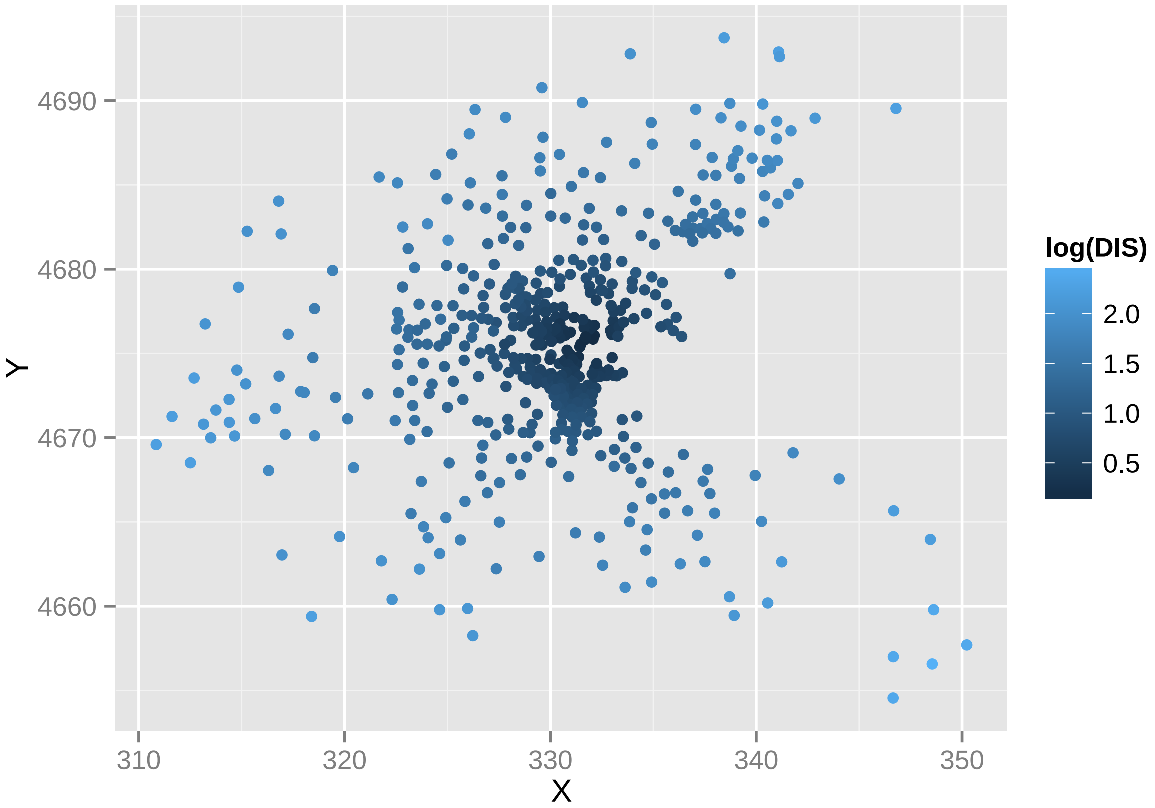
Most of these variables were selected because Economic theory suggests that each should impact median value. A scatterplot matrix is a helpful to quickly visualize many bivariate relations. I like the scatterpot matrix function in the GGally package called `ggpairs`. Sorry it looks so bad printed out. It's better on a big screen.

```
library(ggplot2)
library(GGally)
library(dplyr)
boston.c %>% mutate(logMEDV = log(CMEDV), logDIS=log(DIS)) %>%
  select(AGE, RM, logDIS, B, LSTAT, PTRATIO, NOX, logMEDV) %>% ggpairs()
```



We are trying to understand the various determinants of house price, including air pollution. One of the most important aspects of house price in the US is suburbanization. From the scatterplots, we see a significant relationship between value and distance. It may be helpful to map this out. You could use Latitude and Longitude to map it out, but it is better to use projected coordinates, which are in units of meters, rather than in units of geographic degrees. Fortunately, these have already been calculated for you.
(Note, a GIS course would teach you more about projections.
You could do the projection using GIS software like Quantum GIS or ArcGIS, or you could do it in R using the spTransform function in the sp package.)

```
boston.c$X <- boston.utm[,1]
boston.c$Y <- boston.utm[,2]
ggplot(boston.c) + geom_point(aes(x=X, y=Y, color=log(DIS)))
```



Homework Assignment:

1. Using the scatterplot matrix (`ggpairs`),

a. Describe the correlates of house price.
b. Describe the correlates of NOX.

2. Imagine the multivariate regression of log(CMEDV) on NOX, AGE, log(DIS), RM, CRIM, PTRATIO, B, LSTAT, and CHAS. DO NOT RUN THE REGRESSION YET. For each of these variables, predict whether you think the regression coefficient will be positive or negative, and why. Remember, the multivariate regression relationship is the relationship AFTER you hold the other values fixed. So, for instance to think about the relationship between Distance and value, you should think like: "Imagine two houses that have the same age, same number of rooms, same racial and ethnic neighborhood, same tax rate, etc. Now move one of those houses farther from workplaces. Should that change increase or decrease housing value." Answer: Economic theory suggest that everything else equal, being far from work is a bad thing. Note, the Charles River is a particulurly industrial part of town.
3. One of the relationships is a negative relationship between Distance from Work (primarily Boston) and House Value. Fit a bivariate regression between log CMEDV and log DIS

a. Report the slope of this regression and interpret it's value.
b. Report approximate 95% confidence intervals for the slope.

4. Fit the linear regression from question 2.

a. Report the coefficient of log Distance. Interpret it's value and report it's 95% confidence interval.

b. Explain why the coefficient on log Distance changed so dramatically from in question 3.
c. Come to a conclusion regarding the relationship of air quality (measured by NOX). Is there evidence that NOX has a relationship on house value? Be sure to describe both the value of that relationship and the possible range of values.