Lab 1: Exploratory Data Analysis

This Bitter Earth

Ben Thomas

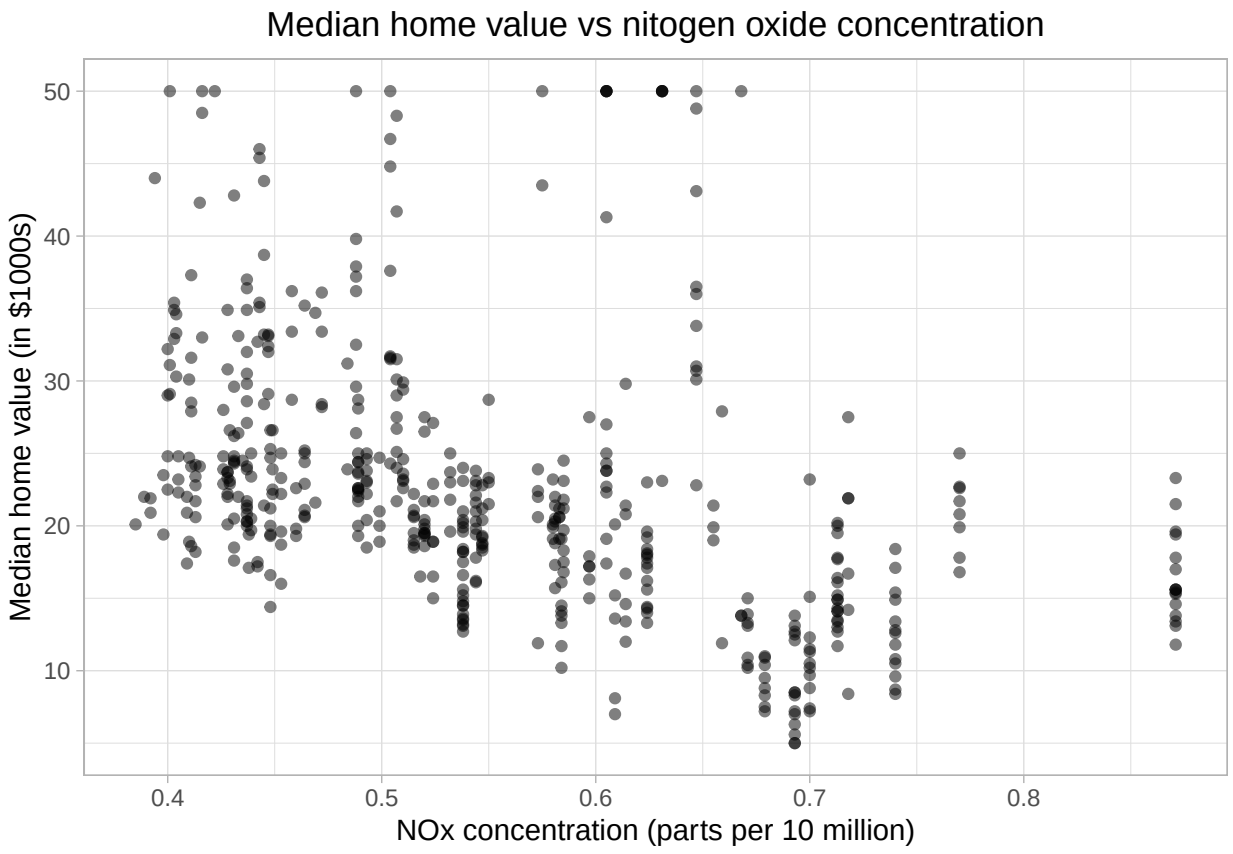
Exercise 1

```
library(MASS)
?Boston
Boston <- Boston
```

There are 506 rows and 14 columns in this data set. The columns in this data set represent different variables, such as per-capita crime rates and accessibility to radial highways. The rows in this data set represent different towns, presumably within the Boston metropolitan area (though it doesn't explicitly say).

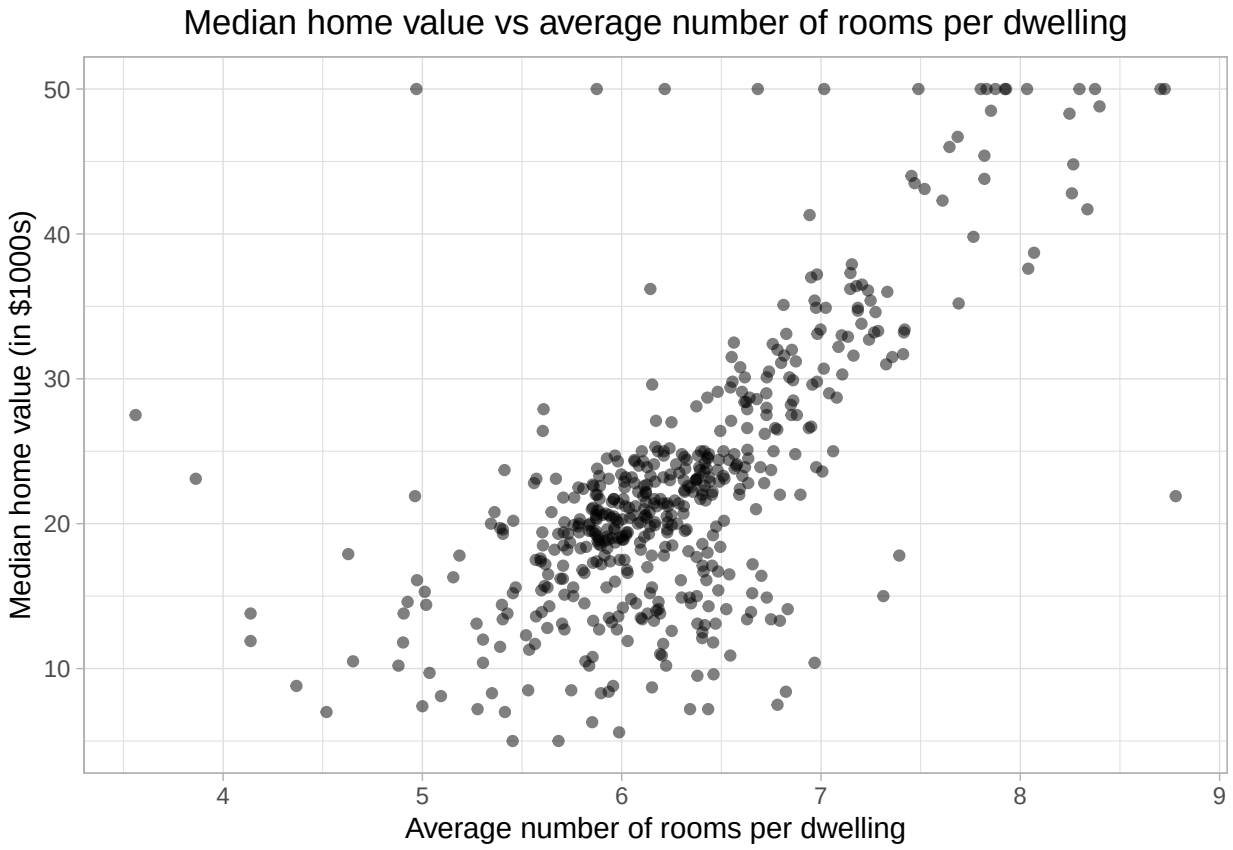
Exercise 2

```
# Create the first scatterplot
library(ggplot2)
plot1 <- ggplot(Boston, aes(x=nox, y=medv)) + geom_point(alpha=.5)
plot1 + labs(title="Median home value vs nitogen oxide concentration",
          y="Median home value (in $1000s)", x="NOx concentration (parts per 10 million)") +
  theme_light() + theme(plot.title = element_text(hjust = 0.5))
```



As we see in this first scatterplot, there seems to be a slight negative relationship between a town's median home value and its concentration of NOx. This makes sense, as all else equal, a home in an area with more pollution should be less desirable–and therefore less valuable–than a home in a less-polluted area.

```
# Create the second scatterplot
library(ggplot2)
plot2 <- ggplot(Boston, aes(x=rm, y=medv)) + geom_point(alpha=.5)
plot2 + labs(title="Median home value vs average number of rooms per dwelling",
         y="Median home value (in $1000s)", x="Average number of rooms per dwelling") +
  theme_light() + theme(plot.title = element_text(hjust = 0.5))
```



As we see in this second scatterplot, there is a strong, positive correlation between a town's median home value and the average number of rooms in its dwellings. This, again, makes a great deal of sense, as we'd expect larger homes to be more valuable than smaller ones.

Exercise 3

```r
# Create the correlation matrix
correlationMatrix <- cor(Boston)

# Display and round the correlations with stargazer
library(stargazer)
stargazer(correlationMatrix[2:14,1], digits = 2, title="Correlations with crime")
```

Table 1: Correlations with crime

zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
-0.20	0.41	-0.06	0.42	-0.22	0.35	-0.38	0.63	0.58	0.29	-0.39	0.46	-0.39

After creating a simple correlation matrix for all variables in the "Boston" data set, rounding these correlation coefficients to the hundredths place, and displaying the correlations with crime rates, many of the predictors are positively correlated with crime rates:

- the proportion of non-retail business acres per town
- the NOx concentration
- the proportion of units built prior to 1940
- the index of accessibility to radial highways
- the property tax rate per $10,000
- the pupil-teacher ratio
- the percent of the population of "lower status"

Exercise 4

```
library(stargazer)
stargazer(Boston, summary.stat = c("min", "max", "median"),
          title="Summary Statistics", align=TRUE,
          covariate.labels=c("Crime rate","Proportion large zones",
                             "Proportion industrial","Charles River",
                             "NOx concentration","Average rooms",
                             "Proportion built pre-1940", "Distance to employment",
                             "Highway accessibility", "Property tax rate",
                             "Student-teacher ratio", "Black", "Lower status"))
```

Table 2: Summary Statistics

Statistic	Min	Max	Median
Crime rate	0.006	88.976	0.257
Proportion large zones	0	100	0
Proportion industrial	0.460	27.740	9.690
Charles River	0	1	0
NOx concentration	0.385	0.871	0.538
Average rooms	3.561	8.780	6.208
Proportion built pre-1940	2.900	100.000	77.500
Distance to employment	1.130	12.127	3.207
Highway accessibility	1	24	5
Property tax rate	187	711	330
Student-teacher ratio	12.600	22.000	19.050
Black	0.320	396.900	391.440
Lower status	1.730	37.970	11.360
medv	5	50	21.2

After looking at summary statistics of the "Boston" data set, we find the following:

- per-capita crime rates vary from 0.01 to 88.98 (a huge amount!)
- the proportion of residential land zoned for lots greater than 25k sq. ft. ranges from 0 to 100
- the proportion of industrial acres per town ranges from 0.46 to 27.74
- the Charles River dummy variable ranges from 0 to 1 (as it should, given that it's a dummy variable)
- the NOx concentration ranges from 0.38 to 0.87 parts per 10 million
- the average number of rooms per dwelling ranges from 3.56 to 8.78
- the percentage of housing built before 1940 ranges from 3 to 100 percent
- the mean of distances to different employment centers ranges from 1.13 to 12.13 (miles?)
- the index of accessibility to radial highways ranges from 1 to 24
- the property tax rate per $10,000 ranges from $187 to $711–a significant variation
- the pupil-teacher ratio by town ranges from 12.6 to 22.0
- the measure of blacks per town ranges from 0.32–a rather low amount–to 396.90
- the percent of population of a lower status ranges from 1.73% to 37.97%
- the median home value ranges from $5,000 to $50,000

Exercise 5

```
sum(Boston[,"chas"])
```

```
## [1] 35
```

Since "chas" = 1 when the town bounds the Charles River, we can easily find out how many towns bound the river by summing the "chas" column of the "Boston" data set. After doing so, we find that 35 towns bound the Charles River.

Exercise 6

```
median(Boston[,"ptratio"])
```

```
## [1] 19.05
```

The median pupil-teacher ratio among the towns in our data set is 19.05.

Exercise 7

In order to build a model predicting the average value of a home based on other variables, the ideal response variable would be value of each home in our data set. Given the Boston data set, which does not give the value for each home in the Boston area, but instead the median home value in each town, the closest approximation is the median home value. Potential inputs would be: (1) crime rates; (2) NOx concentration; (3) average number of rooms per dwelling; (4) proportion of units built prior to 1940; etc. Arguments could be made to include every single predictor in the Boston dataset in our inputs.

One attempt at a model predicting home values is as shown below, which uses a linear regression model. Since I could think of theoretical reasons that each predictor would be related to the home value, I've included all.

```
# Create the linear model
library(stargazer)
linearModel <- lm(medv ~ ., data = Boston)
#Display the linear model with stargazer
stargazer(linearModel, title="Regression Results", align=TRUE,
          dep.var.labels =c("Median home value"),
          covariate.labels=c("Crime rate","Proportion large zones",
                             "NOx concentration","Average rooms",
                              "Proportion built pre-1940", "Distance to employment",
                             "Student-teacher ratio", "Black", "Lower status"),
          omit.stat=c("LL","ser","f"), single.row =TRUE)
```

Table 3: Regression Results

	Dependent variable:
	Median home value
Crime rate	-0.108^{***} (0.033)
Proportion large zones	0.046^{***} (0.014)
Proportion industrial	0.021 (0.061)
Charles River	2.687^{***} (0.862)
NOx concentration	-17.767^{***} (3.820)
Average rooms	3.810^{***} (0.418)
Proportion built pre-1940	0.001 (0.013)
Distance to employment	-1.476^{***} (0.199)
Highway accessibility	0.306^{***} (0.066)
Property tax rate	-0.012^{***} (0.004)
Student-teacher ratio	-0.953^{***} (0.131)
Black	0.009^{***} (0.003)
Lower status	-0.525^{***} (0.051)
Constant	36.459^{***} (5.103)
Observations	506
R^2	0.741
Adjusted R^2	0.734
Note:	*p<0.1; **p<0.05; ***p<0.01

As we can see, this simple linear regression model does a fairly good job of predicting median home values, and accounts for 73% of the variation in median home values. Additionally, nearly all of the predictor variables are significant at either the .1% or 1% level, supporting our initial thought that nearly all of these predictors have theoretical justification for causing variation in home values.