Name - vvgupta2

Homework 11

Homework Instructions

For questions that require code, please create a code chunk directly below the question and type your code there. Your knitted pdf will show both your code and your output. You are encouraged to knit your file as you work to check that your coding and formatting is done so appropriately.

For written responses or multiple choice questions, please bold your (selected) answer.

Grading Details

All questions will be graded full credit (1 point), half credit (0.5 point) or no credit (0 points).

Full credit responses should have the correct response and appropriate code (if applicable). Half credit responses will have a reasonable attempt (typically no more than one small error or oversight), and no credit responses will be either non-attempts or attempts with significant errors.

Exercise 1

Which statement is true regarding the predictors of a sensible multiple regression model? **bold** your answer

- Good models have predictors that are all highly correlated with each other

- Good models have predictors that are each highly correlated with at least one other predictor in the model

- Good models have predictors that rarely have high correlation with other predictors in the model

Exercise 2

Which statement correctly interprets what *adjusted* r^2 measures? **bold** your answer

- The percentage of variability in the response variable that is *not* explained by the predictors after adjusting for correlation likely explained due to random chance

- The percentage of variability in the response variable that *is* explained by the predictors after adjusting for correlation likely explained due to random chance

- The probability of seeing this much correlation explained by random chance if none of the predictors are correlated with the response variable
- The probability of seeing this much correlation explained by random chance if all of the predictors are correlated with the response variable

Exercise 3

Let's examine the `Boston` data from the `MASS` package for the remaining exercises. Each row in this data represents one community and summary results of that commnity. For example: `medv` represents the median home value of that community.

We'd like to try to predict the `medv` of a community based on other predictors in the model.

First, create a full model using all other predictors as linear predictors of `medv`. Report the summary of your model

```
#solution
library(MASS)
#?Boston
#names(Boston)
model1 = lm(medv~crim+rad+tax+zn+indus+ptratio+chas+black+nox+lstat+rm+age+dis, data = Boston)
summary(model1)
```

```
##
## Call:
## lm(formula = medv ~ crim + rad + tax + zn + indus + ptratio +
##     chas + black + nox + lstat + rm + age + dis, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Exercise 4

Compute the variance inflation factor for all predictors in the model by using the `vif` function in the `faraway` package.

Which predictor is *most* explained by the other predictors in this model?

```
#solution
library(faraway)
vif(model1)
```

```
##     crim      rad      tax       zn    indus  ptratio     chas    black 
## 1.792192 7.484496 9.008554 2.298758 3.991596 1.799084 1.073995 1.348521 
##      nox    lstat       rm      age      dis 
## 4.393720 2.941491 1.933744 3.100826 3.955945
```

Tax: is the most explained by the other predictors in this model

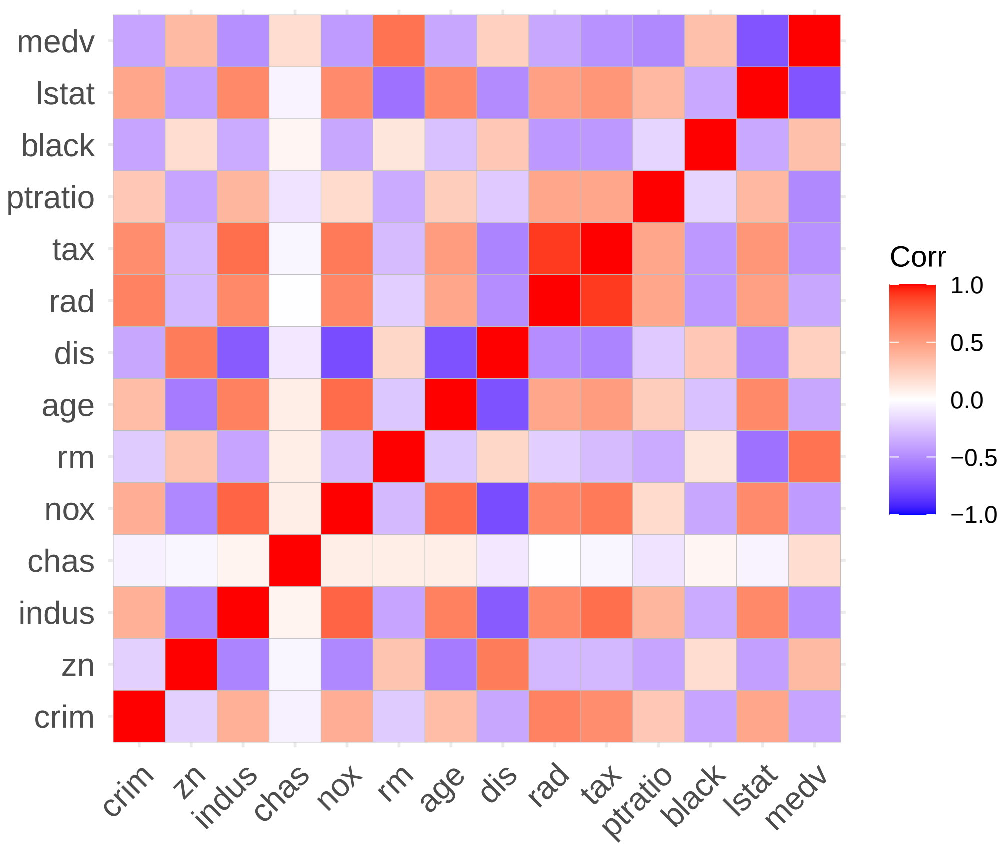
Exercise 5

Now compute a full correlation table of all variables in `Boston`. Round these correlations to 2 decimal places.

```
#solution
#install.packages("ggcorrplot")
library(ggplot2)
library(ggcorrplot)
corr = round ( cor(Boston), 2)
corr
```

```
##          crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio
## crim     1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58    0.29
## zn      -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31   -0.39
## indus    0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72    0.38
## chas    -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04   -0.12
## nox      0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67    0.19
## rm      -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29   -0.36
## age      0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51    0.26
## dis     -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53   -0.23
## rad      0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91    0.46
## tax      0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00    0.46
## ptratio  0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46    1.00
## black   -0.39  0.18 -0.36  0.05 -0.38  0.13 -0.27  0.29 -0.44 -0.44   -0.18
## lstat    0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54    0.37
## medv    -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47   -0.51
##         black lstat  medv
## crim    -0.39  0.46 -0.39
## zn       0.18 -0.41  0.36
## indus   -0.36  0.60 -0.48
## chas     0.05 -0.05  0.18
## nox     -0.38  0.59 -0.43
## rm       0.13 -0.61  0.70
## age     -0.27  0.60 -0.38
## dis      0.29 -0.50  0.25
## rad     -0.44  0.49 -0.38
## tax     -0.44  0.54 -0.47
## ptratio -0.18  0.37 -0.51
## black    1.00 -0.37  0.33
## lstat   -0.37  1.00 -0.74
## medv     0.33 -0.74  1.00
```

```r
ggcorrplot(corr)
```



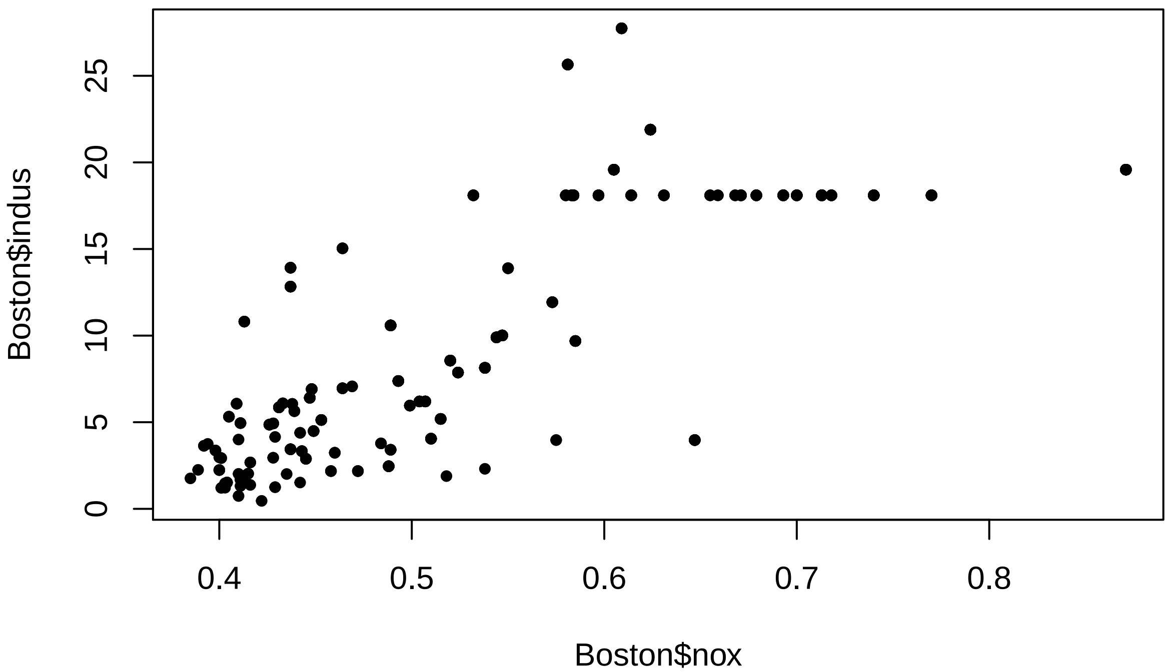
Exercise 6

There are three predictor variables that have fairly high correlations with one another: `indus`, `nox`, and `dis`. (`rad` and `tax` are also high with each other, but we'll revisit later).

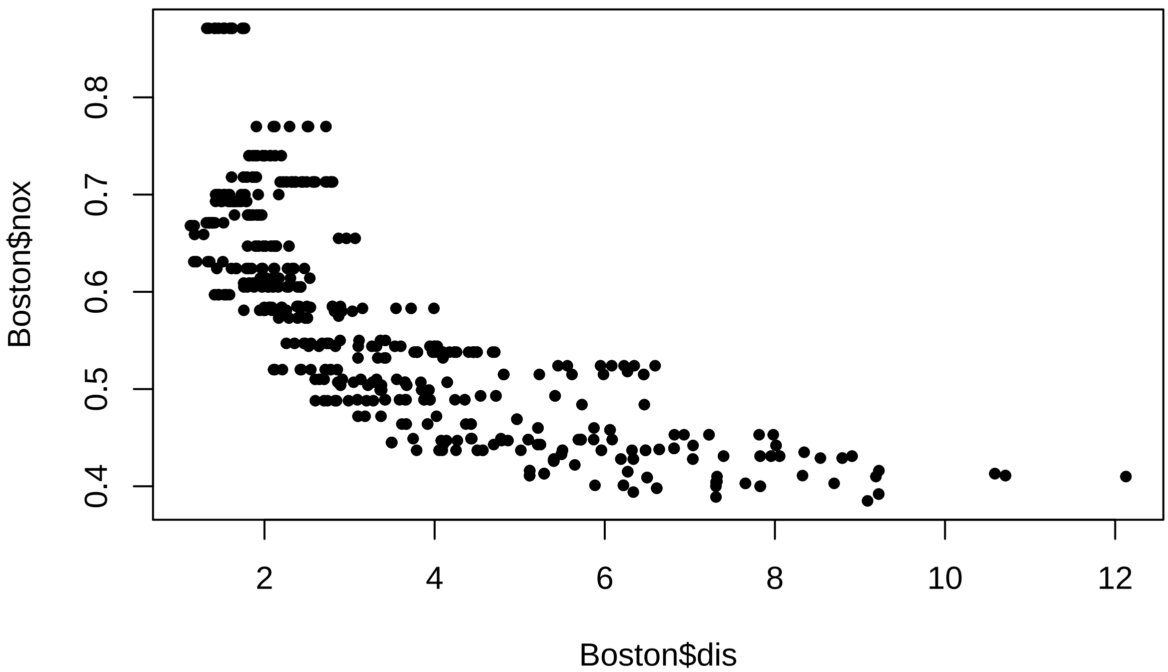
Take a look at the documentation for the data to learn what each of these variables represents.

Create three scatterplots, one for each pair, using whatever plotting option of your choice. *No particular formatting required.* Then briefly describe in words how these three variables seem to relate to one another. Please write your response in the white space below your code chunk for the plots.

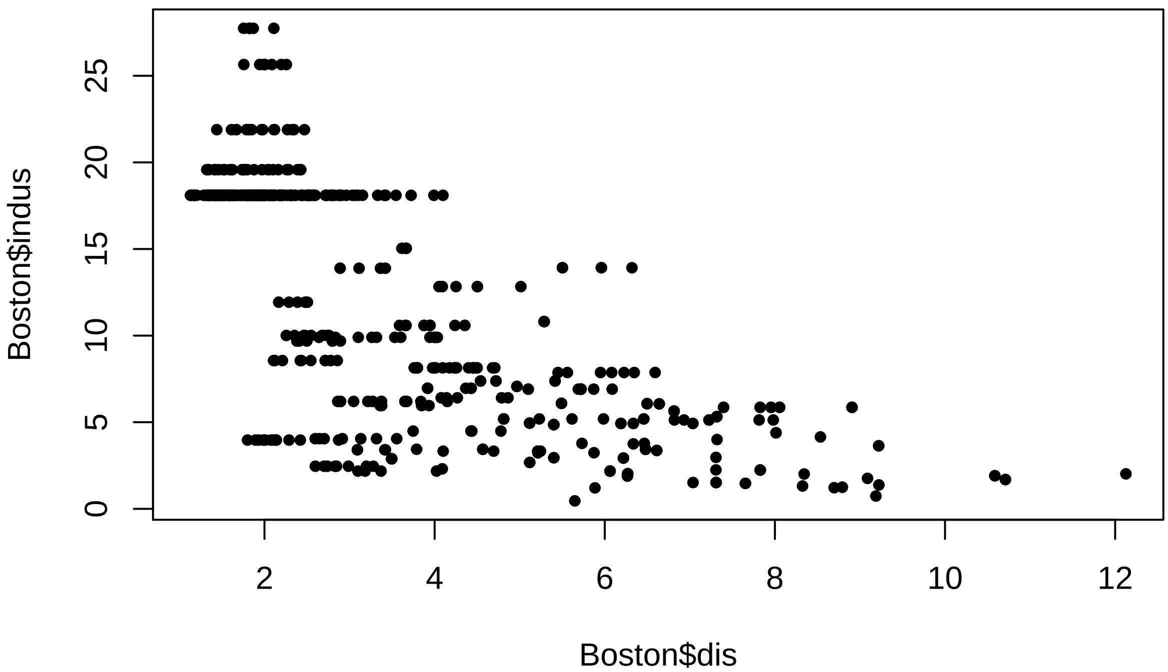
```r
#solution
plot(Boston$indus~Boston$nox, pch =20)
```




```
plot(Boston$nox~Boston$dis, pch =20)
```




```
plot(Boston$indus~Boston$dis, pch = 20)
```



There is a negative correlation between indus and dis as well as nox and dis but there is a positive correlation between indus and nox.

Exercise 7

Let's consider how our model fit changes when we remove one or two of these three highly correlated predictors. Of these three predictors, it appears that `indus` has the weakest contribution to the full model. Followed by `nox`.

Create two new models–one named `ind_mod` that excludes `indus`, and one named `indnox_mod` that excludes `indus` and 'nox. All other predictors still included.

Now run summaries of these two models.

Hint: You can use a minus sign, like "-pred_name" to exclude a variable. This is very handy to pair with "." when simply removing one or two variables.

```
#solution
ind_mod = lm(medv~crim+rad+tax+zn+ptratio+chas+black+nox+lstat+rm+age+dis, data = Boston)
summary(ind_mod)
```

```
##
## Call:
## lm(formula = medv ~ crim + rad + tax + zn + ptratio + chas +
##     black + nox + lstat + rm + age + dis, data = Boston)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -15.587  -2.737  -0.506   1.742  26.212
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.636e+01  5.091e+00   7.143 3.30e-12 ***
## crim        -1.084e-01  3.281e-02  -3.304 0.001022 **
## rad          2.999e-01  6.367e-02   4.710 3.22e-06 ***
## tax         -1.178e-02  3.378e-03  -3.489 0.000529 ***
## zn           4.593e-02  1.364e-02   3.368 0.000816 ***
## ptratio     -9.471e-01  1.296e-01  -7.308 1.10e-12 ***
## chas         2.716e+00  8.562e-01   3.173 0.001605 **
## black        9.282e-03  2.682e-03   3.461 0.000586 ***
## nox         -1.743e+01  3.681e+00  -4.735 2.87e-06 ***
## lstat       -5.235e-01  5.052e-02 -10.361  < 2e-16 ***
## rm           3.797e+00  4.158e-01   9.132  < 2e-16 ***
## age          6.971e-04  1.320e-02   0.053 0.957898
## dis         -1.490e+00  1.948e-01  -7.648 1.08e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.741 on 493 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7343
## F-statistic: 117.3 on 12 and 493 DF,  p-value: < 2.2e-16
```

```r
indnox_mod = lm(medv~crim+rad+tax+zn+ptratio+chas+black+lstat+rm+age+dis, data = Boston)
summary(indnox_mod)
```

```
##
## Call:
## lm(formula = medv ~ crim + rad + tax + zn + ptratio + chas +
##     black + lstat + rm + age + dis, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.140  -2.839  -0.797   1.575  27.238
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.068295   4.337556   5.318 1.59e-07 ***
## crim        -0.097230   0.033430  -2.909 0.003795 **
## rad          0.277528   0.064853   4.279 2.25e-05 ***
## tax         -0.015292   0.003366  -4.543 6.98e-06 ***
## zn           0.050232   0.013899   3.614 0.000332 ***
## ptratio     -0.757567   0.125912  -6.017 3.47e-09 ***
## chas         2.459729   0.872837   2.818 0.005025 **
## black        0.010384   0.002729   3.805 0.000160 ***
## lstat       -0.545575   0.051385 -10.617  < 2e-16 ***
## rm           4.038353   0.421496   9.581  < 2e-16 ***
## age         -0.016500   0.012960  -1.273 0.203569
## dis         -1.159319   0.185765  -6.241 9.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.843 on 494 degrees of freedom
## Multiple R-squared:  0.7288, Adjusted R-squared:  0.7227
## F-statistic: 120.7 on 11 and 494 DF,  p-value: < 2.2e-16
```

Exercise 8

Now, let's compare the full model to the two models we just created.

First, extract the adjusted r squared value of each of these three models.

Also complete F-tests to compare the full model to `ind_mod`, and another to compare `ind_mod` to `indnox_mod`.

```
#solution
summary(model1)$adj.r.squared
```

```
## [1] 0.7337897
```

```
summary(ind_mod)$adj.r.squared
```

```
## [1] 0.7342694
```

```
summary(indnox_mod)$adj.r.squared
```

```
## [1] 0.7227467
```

```
anova(model1,ind_mod)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ crim + rad + tax + zn + indus + ptratio + chas + black +
##     nox + lstat + rm + age + dis
## Model 2: medv ~ crim + rad + tax + zn + ptratio + chas + black + nox +
##     lstat + rm + age + dis
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    492 11079
## 2    493 11081 -1   -2.5167 0.1118 0.7383
```

```
anova(indnox_mod,ind_mod)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ crim + rad + tax + zn + ptratio + chas + black + lstat +
##     rm + age + dis
## Model 2: medv ~ crim + rad + tax + zn + ptratio + chas + black + nox +
##     lstat + rm + age + dis
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    494 11585
## 2    493 11081  1    503.96 22.421 2.867e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 9

Based on your results from Exercise 8, which seems to be the most predictive model for `medv` without being a redundant model?

- The full model

- The `ind_mod`

- The `indnox_mod`

Exercise 10

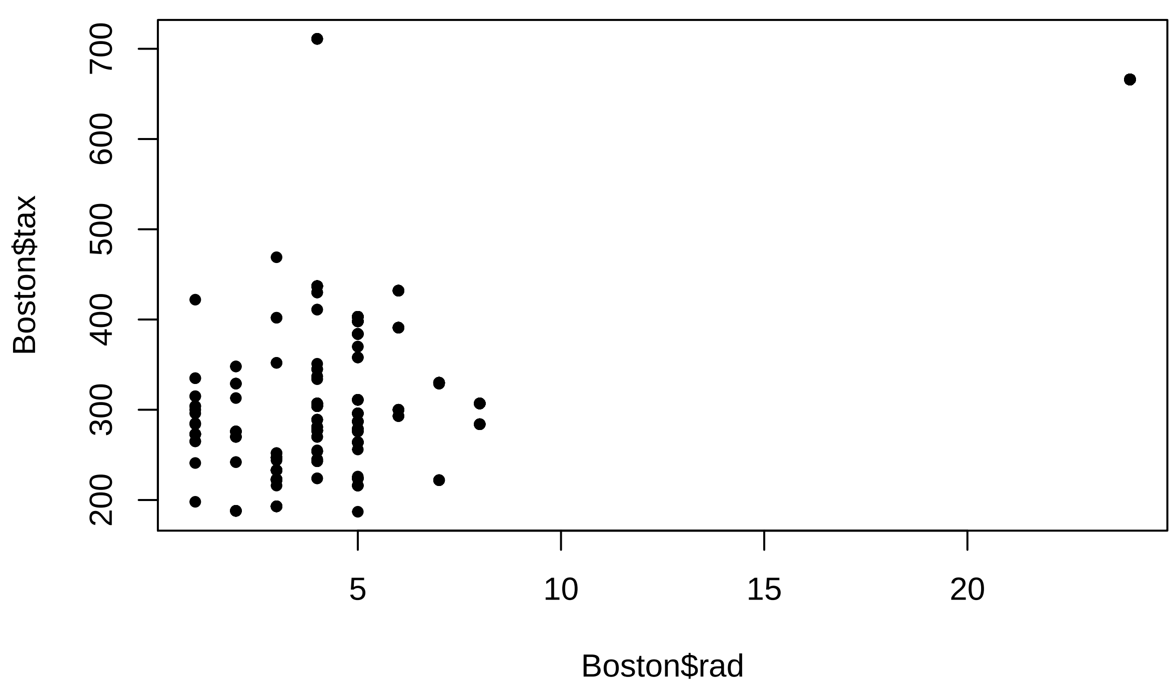
You might remember that `rad` and `tax` are two other predictors that are highly correlated with one another. Check the documentation to learn what these variables represent.

You are going to make TWO scatterplots. One with base R plotting. No additional formatting required.

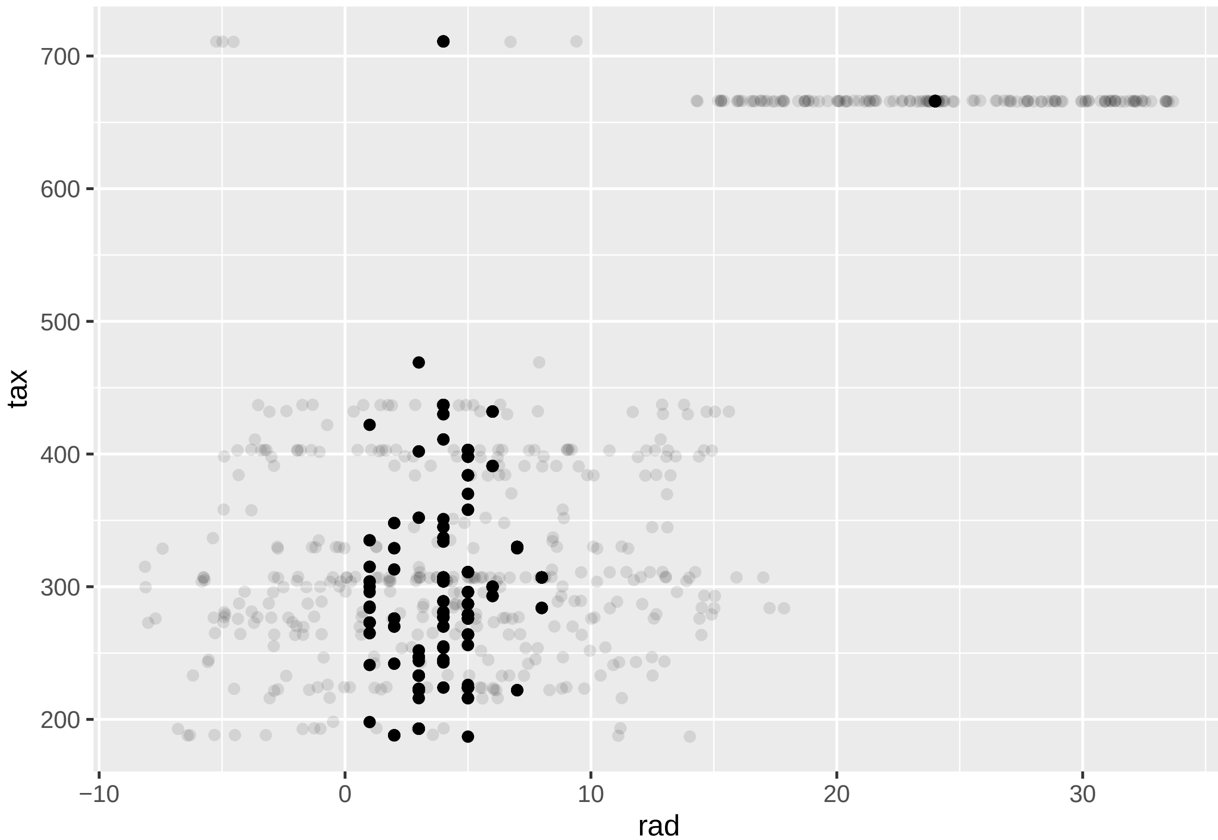
Then make a scatterplot using ggplot2. **Specifically, you should use the `geom_jitter` option, set the width argument to 10, and set alpha to a low value, like 0.1.** *All other formatting optional.*

Then briefly discuss what the ggplot2 graph reveals that may not be obvious from the base R plot. View the Boston data if needed!

```
#solution
plot(Boston$tax~Boston$rad, pch = 20)
```



```r
ggplot(Boston,aes(x=rad, y=tax)) +
  geom_point()+
  geom_jitter(width=10, alpha=0.1)
```



Exercise 11

Since the correlation is not particularly consistent between these two predictor variables, it may make sense to hold both in the model.

Create a new model: `indtax_mod`, that does not include `indus` or `tax`, but includes all other predictors. Then complete an F-test to against `ind_mod` to determine if there is a statistically significant improvement by keeping `tax`, or whether the difference is statistically negligible.

```r
#solution
indtax_mod = lm(medv~crim+rad+zn+ptratio+chas+black+nox+lstat+rm+age+dis, data = Boston)
anova(indtax_mod, ind_mod)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ crim + rad + zn + ptratio + chas + black + nox + lstat +
##     rm + age + dis
## Model 2: medv ~ crim + rad + tax + zn + ptratio + chas + black + nox +
##     lstat + rm + age + dis
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
```

```
## 1    494 11355
## 2    493 11081  1    273.59 12.172 0.0005287 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Tax does not add any statistically significant improvement to the model

- Tax does add a statistically significant improvement to the model

Exercise 12

Just kidding. There is no exercise 12. This is just 1 free point. :)