FOR/STT 875, Exercise 1

Learning objectives

- Practice setting up a working directory and reading in data
- Explore the workspace within RStudio and associated commands
- Produce basic descriptive statistics and graphics

Deliverables

Create a R script that follows the code format of the script I created for the first part of this exercise (see exercise_1.R in the D2L Exercise 1 directory). Your answer script should follow the format of the template (see template_1.R). The Exercise 1 video provides some guidance for completing this exercise.

After you check that your script runs and produces the correct answers to the questions listed at the end of this document, upload it to the Exercise 1 D2L dropbox.

Grading

You will receive full credit if your script: 1) runs without error; 2) correctly answers the six questions below (with code and commented answer); and 3) is neatly formatted with sufficient commenting. Here, code comments should identify the lines of code used to answer a given question. Use RStudio's code sections (see template_1.R) to divide your script by question.

Introduction

The `Lahman` R package contains a wealth of data related to baseball. As the semester progresses we will make use of this data set in a variety of ways. For this exercise we'll perform some very simple analyses using the data.

Data on the batting average (BA), number of home runs (HR), and number of runs batted in (RBI) for three players, Jim Rice, Carlton Fisk, and Ted Williams, were extracted from the larger data set.

First we will read these data into our R session. At this point, just mimic the statements below to read in the data, without worrying about what they mean. Later in the course we'll cover reading data into R in some detail. (For those who are curious, the `url` function creates a connection to the url containing the data. The `load` function loads the objects that were saved into R and `close` closes the connection to my web server where the data are located. Finally, just to keep my workspace tidy, I use the `rm` function to remove the `con` object.)

```
con <- url("http://blue.for.msu.edu/FOR875/data/batting.RData")
load(con)
close(con)
rm(con)
```

Two useful functions are `rm` and `ls` . The first removes an object (e.g., the `con` object above) from the current workspace. The second lists all the objects in the current R workspace.

```
ls()
```

```
 [1] "batting_stats"  "CarltonFiskBA"  "CarltonFiskHR"  "CarltonFiskRBI"
 [5] "JimRiceBA"      "JimRiceHR"      "JimRiceRBI"     "TedWilliamsBA"
 [9] "TedWilliamsHR"  "TedWilliamsRBI"
```

Initially we'll be interested in the nine objects that contain the batting data set, number of home runs, and number of RBIs for three players, Jim Rice, Carlton Fisk, and Ted Williams. The other object we read in, `batting_stats`, contains such data and more for a wide variety of players.

First let's look at the batting average data for Jim Rice. This *vector* holds his batting average for the years 1974-1989. *Vectors* typically contain values for one variable, e.g, Jim Rice's batting average with the vector named `JimRiceBA`.

```
JimRiceBA
```

```
 1974  1975  1976  1977  1978  1979  1980  1981  1982  1983  1984  1985
0.269 0.309 0.282 0.320 0.315 0.325 0.294 0.284 0.309 0.305 0.280 0.291
 1986  1987  1988  1989
0.324 0.277 0.264 0.234
```

```
str(JimRiceBA)
```

```
 Named num [1:16] 0.269 0.309 0.282 0.32 0.315 0.325 0.294 0.284 0.309 0.305 ...
 - attr(*, "names")= chr [1:16] "1974" "1975" "1976" "1977" ...
```

Right off the bat, we notice that this *vector* contains years and their batting averages. What kind of structure is this? The `str` function tells us this is a "Named num," which is a numeric vector whose members have names. Notice how the descriptive statistic functions run on the numeric values and not the names (the years).

```
mean(JimRiceBA)
```

```
[1] 0.292625
```

```
max(JimRiceBA)
```

```
[1] 0.325
```

```
which.max(JimRiceBA)
```

```
1979
   6
```

```
min(JimRiceBA)
```

```
[1] 0.234
```

```
which.min(JimRiceBA)
```

```
1989
  16
```

First we display the data. Next we compute the mean of the batting averages. Next we find the maximum. Next we determine which of the values (in this case, the value in the 6th position, from the year 1979) contains the maximum. Then we do the same for the minimum batting average.

Next let's find out when Jim Rice's batting average increased and decreased the most. (Take a look at the manual page for the `diff` either using `?diff` on the console or using the Help tab and search in the lower right RStudio window. As you see new functions being used, it is helpful to read their associated manual pages.)

```
JimRiceBAdiffs <- diff(JimRiceBA, lag = 1)
JimRiceBAdiffs
```

```
  1975   1976   1977   1978   1979   1980   1981   1982   1983   1984
 0.040 -0.027  0.038 -0.005  0.010 -0.031 -0.010  0.025 -0.004 -0.025
  1985   1986   1987   1988   1989
 0.011  0.033 -0.047 -0.013 -0.030
```

```
max(JimRiceBAdiffs)
```

```
[1] 0.04
```

```
which.max(JimRiceBAdiffs)
```

```
1975
   1
```

```
min(JimRiceBAdiffs)
```

```
[1] -0.047
```

```
which.min(JimRiceBAdiffs)
```

```
1987
  13
```

Next we look at the relationship between batting average, RBIs, and home runs for Jim Rice. Specifically we'll draw three scatter plots, and compute three correlation coefficients.

```
cor(JimRiceBA, JimRiceRBI)
```

```
[1] 0.7942272
```

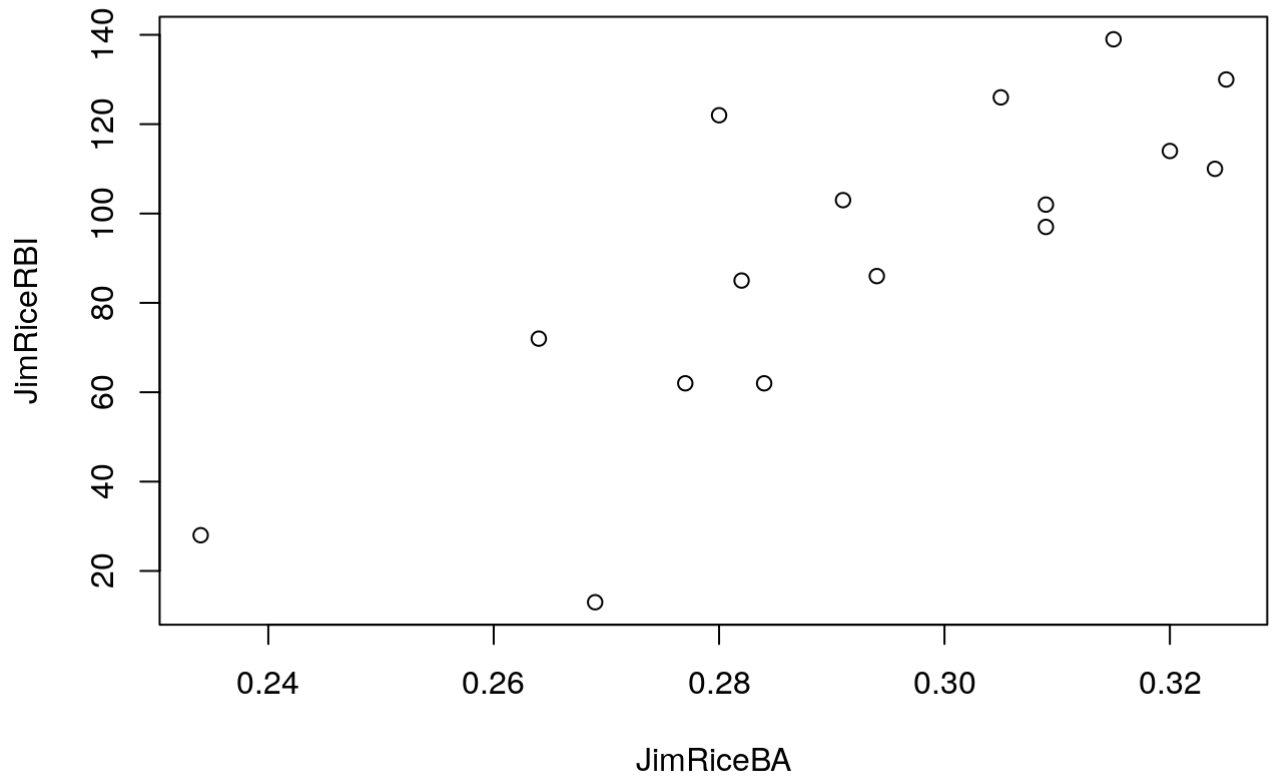
```
cor(JimRiceBA, JimRiceHR)
```

```
[1] 0.7576232
```

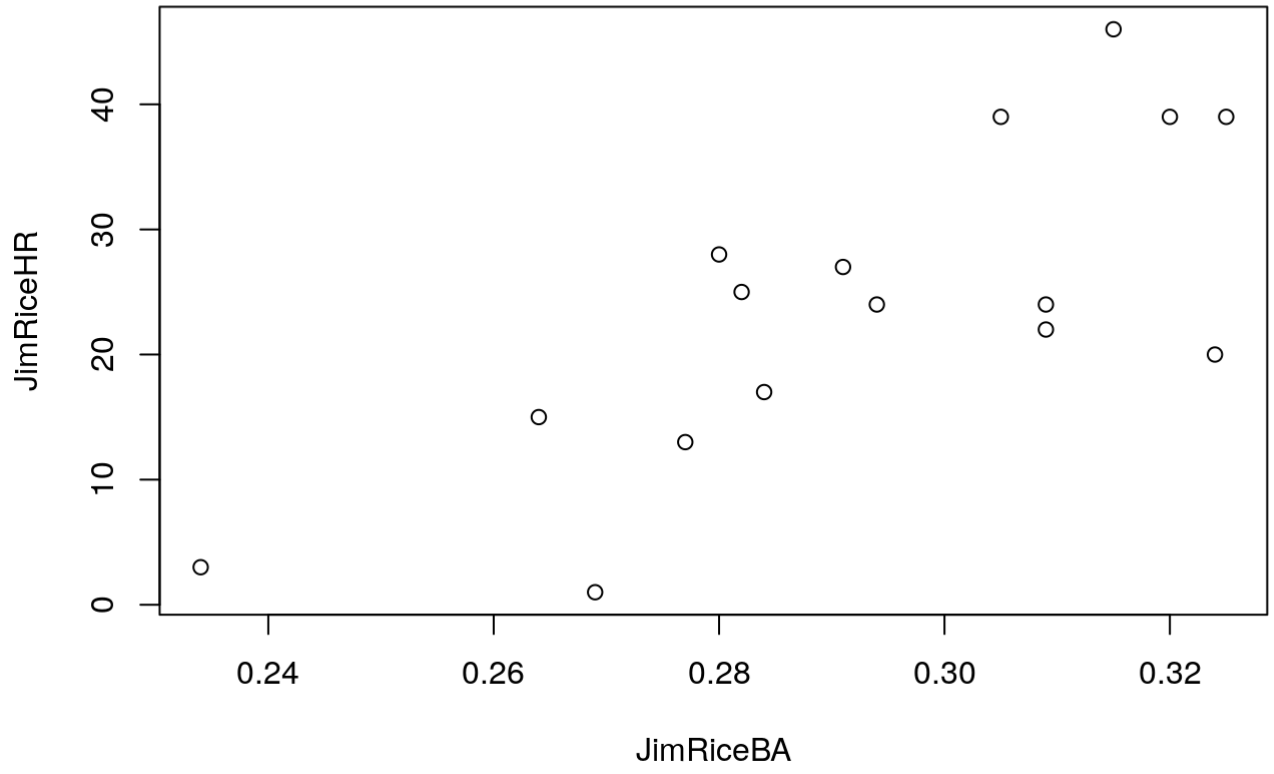
```
cor(JimRiceHR, JimRiceRBI)
```

```
[1] 0.9305225
```

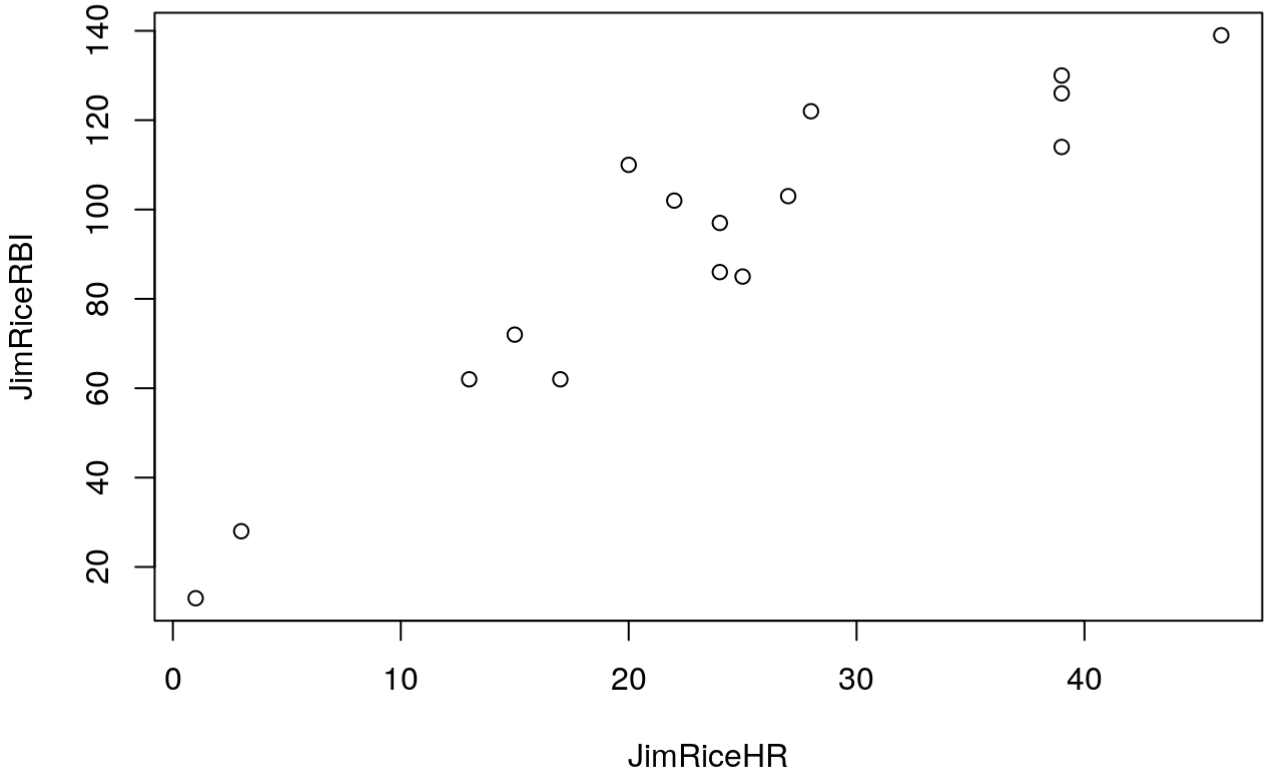
```
plot(JimRiceBA, JimRiceRBI)
```




```
plot(JimRiceBA, JimRiceHR)
```




```
plot(JimRiceHR, JimRiceRBI)
```



Both the scatter plots and the correlation calculations show that the strongest relationship exists between home runs and RBIs (At this point we are using the `base` graphics functions in R to create graphics. Later in the course we will learn to use a package called `ggplot2` which provides an alternative to the base graphics.)

Questions to answer

Write an R script to answer the following questions for Ted Williams using `TedWilliamsBA`, `TedWilliamsHR`, and `TedWilliamsRBI` vectors.

1. How many seasons did Ted Williams play? (Hint: Use the `length` function.)
2. In which season did Ted Williams have his highest batting average?
3. What was this highest batting average?
4. What was Ted Williams' mean batting average?
5. For which pair of the variables representing home runs, RBIs, and batting average, is the correlation the highest? What is this correlation?
6. What was the largest jump in Ted Williams' RBIs from one season to the next? In which season did this jump occur?