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

Data on the batting average, number of home runs, and number of RBIs, 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 will 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.)

```r
> con <- url("http://www.stt.msu.edu/~melfi/STT301/Day2/Day2Batting.gz")
> load(con)
> close(con)
> rm(con)
> ls()
```

Two functions that are quite useful are `rm` and `ls`. The first removes an object (in this case, `con`) from the current R workspace. The second lists all the objects in the current R workspace. If you had created other objects in an earlier R session, they might be listed along with the objects read in from the external file.

Initially we’ll be interested in the nine objects which contain the batting averages, 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.

Initially we’ll be interested in the nine objects which contain the batting averages, 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.

```r
> JimRiceBA

0.269 0.309 0.282 0.320 0.315 0.325 0.294 0.284 0.309 0.305
0.280 0.291 0.324 0.277 0.264 0.234
```

```r
> mean(JimRiceBA)
```
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.

```r
> JimRiceBAdiffs <- diff(JimRiceBA, lag = 1)
> JimRiceBAdiffs
  0.040  -0.027  0.038  -0.005  0.010  -0.031  -0.010  0.025
 -0.004  -0.025  0.011  0.033  -0.047  -0.013  -0.030
> max(JimRiceBAdiffs)
[1] 0.04
> which.max(JimRiceBAdiffs)
  1975
     1
```
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.

```r
> cor(JimRiceBA, JimRiceRBI)
[1] 0.7942272
> cor(JimRiceBA, JimRiceHR)
[1] 0.7576232
> cor(JimRiceHR, JimRiceRBI)
[1] 0.9305225
> plot(JimRiceBA, JimRiceRBI)
```
> plot(JimRiceBA, JimRiceHR)
Both the scatter plots and the correlation calculations show that the strongest relationship exists between home runs and RBIs.\footnote{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 \texttt{ggplot2} which provides an alternative to the base graphics.}

\textbf{Exercise 1.} Answer the following questions for Ted Williams. Use R functions to answer the questions rather than just reading off the answers by looking at the data.

(a) How many seasons did Ted Williams play? (Hint: Use the \texttt{length} function.)

(b) In which season did Ted Williams have his highest batting average?

(c) What was this highest batting average?

(d) What was Ted Williams’ mean batting average?

(e) For which pair of the variables representing home runs, RBIs, and batting average, is the correlation the highest? What is this correlation?

(f) What was the largest jump in Ted Williams’ RBIs from one season to the next? In which season did this jump occur?
The R objects like `JimRiceBA` or `JimRiceRBI` that we worked with above are vectors, which typically contain values for one variable. Data frames provide a way to collect many variables together. The data frame `batting_stats` contains data on several variables for many baseball players. Later in the course we will learn a lot about data frames, which provide a way to represent data sets with many variables.

Variables in data frames are represented by columns. We can find out the names of the variables using the `names` function. Also the `dim` function yields the number of rows and columns of the data frame. We can also select particular rows and/or columns as shown below.

```
> names(batting_stats)
[1] "playerID" "yearID"  "stint"    "teamID"
[5] "lgID"    "G"       "G_batting" "AB"
[9] "R"       "H"       "X2B"      "X3B"
[13] "HR"      "RBI"     "SB"       "CS"
[17] "BB"      "SO"      "IBB"      "HBP"
[21] "SH"      "SF"      "GIDP"     "G_old"
[25] "BA"      "PA"      "TB"       "SlugPct"
[29] "OBP"     "OPS"     "BABIP"    "playerName"

> dim(batting_stats)
[1] 97889   32

> batting_stats[90000, c(32, 25, 13, 14)]

   playerName BA HR RBI
90000 Tim Corcoran 0.214 0 22

> batting_stats[c(90000, 91000, 92000), c(32, 25, 13, 14)]

   playerName  BA  HR RBI
90000 Tim Corcoran 0.214 0 22
91000 Todd Zeile 0.273 18 87
92000 Tom Paciorek 0.284 1 11

The `subset` function provides a way to select rows of a data frame that satisfy a condition of interest (rather than just specifying the row number(s) as above). We will use `subset` to select the rows corresponding to a particular player, Rocky Colavito, and also will select all the rows for which the batting average is greater than 0.400, and for which there were at least 100 games played.
```r
> RockyColavito <- subset(batting_stats, playerName == "Rocky Colavito")
> RockyColavito[, c(32, 25, 13, 14)]

<table>
<thead>
<tr>
<th>playerName</th>
<th>BA</th>
<th>HR</th>
<th>RBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rocky Colavito</td>
<td>0.444</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.276</td>
<td>21</td>
<td>65</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.252</td>
<td>25</td>
<td>84</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.303</td>
<td>41</td>
<td>113</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.257</td>
<td>42</td>
<td>111</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.249</td>
<td>35</td>
<td>87</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.290</td>
<td>45</td>
<td>140</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.273</td>
<td>37</td>
<td>112</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.271</td>
<td>22</td>
<td>91</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.274</td>
<td>34</td>
<td>102</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.287</td>
<td>26</td>
<td>108</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.238</td>
<td>30</td>
<td>72</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.221</td>
<td>3</td>
<td>29</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.241</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.220</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>Rocky Colavito</td>
<td>0.204</td>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>

> Over400 <- subset(batting_stats, BA > 0.4 & G >= 100)
> dim(Over400)

[1] 25 32

> Over400[, c("playerName", "BA", "yearID")]

<table>
<thead>
<tr>
<th>playerName</th>
<th>BA</th>
<th>yearID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Terry</td>
<td>0.401</td>
<td>1930</td>
</tr>
<tr>
<td>Billy Hamilton</td>
<td>0.404</td>
<td>1894</td>
</tr>
<tr>
<td>Ed Delahanty</td>
<td>0.407</td>
<td>1894</td>
</tr>
<tr>
<td>Ed Delahanty</td>
<td>0.404</td>
<td>1895</td>
</tr>
<tr>
<td>Ed Delahanty</td>
<td>0.410</td>
<td>1899</td>
</tr>
<tr>
<td>Fred Dunlap</td>
<td>0.412</td>
<td>1884</td>
</tr>
<tr>
<td>George Sisler</td>
<td>0.407</td>
<td>1920</td>
</tr>
<tr>
<td>George Sisler</td>
<td>0.420</td>
<td>1922</td>
</tr>
<tr>
<td>Harry Heilmann</td>
<td>0.403</td>
<td>1923</td>
</tr>
<tr>
<td>Hugh Duffy</td>
<td>0.440</td>
<td>1894</td>
</tr>
<tr>
<td>Hughie Jennings</td>
<td>0.401</td>
<td>1896</td>
</tr>
<tr>
<td>Jesse Burkett</td>
<td>0.409</td>
<td>1895</td>
</tr>
<tr>
<td>Jesse Burkett</td>
<td>0.410</td>
<td>1896</td>
</tr>
<tr>
<td>Nap Lajoie</td>
<td>0.426</td>
<td>1901</td>
</tr>
</tbody>
</table>
```
To specify that a variable is equal to a given value, we use `==`. If, as in `RockyColavito[, , 25, 13, 14]`, we leave the row specification blank, R returns all rows. To specify both of two conditions, we use `&`. And we can specify columns either by the column number, or by the column name.

**Exercise 2.** Use the `batting_stats` data frame to answer these questions.

(a) Extract all the rows giving data on the baseball player Lou Gehrig and assign the result to an object called `LouGehrig`.

(b) Display the player name, batting average, number of home runs, number of RBIs, and year for Lou Gehrig.

(c) Display the same data only for the years when Lou Gehrig hit at least 30 home runs.

(d) Extract all the rows where the number of home runs is 50 or more and the batting average is over 0.300. Assign the result to an object called `PowerAndAverage`.

(e) Which players had years with 50 or more home runs and an average over 0.300?

(f) In which years did these players achieve this result?

(g) Display the player name, number of home runs, number of RBIs, batting average, year, and number of walks (in the column called `BB`) for these seasons.

(h) Try to extract the rows when the number of home runs is greater than 90. What happens?

★