Baseball and R Markdown Introduction

<Kaitlyn Watson Group #9>

Module 1 In-class Assignment 4

Introduction

Sean Lahman's Baseball Database contains pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2019. It includes data from the two current leagues (American and National), the four other "major" leagues (American Association, Union Association, Players League, and Federal League), and the National Association of 1871-1875. The `Lahman` package in R contains a plethora of baseball data. This assignment will use a subset of data from the `Lahman` package to expose you to some basic descriptive statistical functions and data subsetting within the R Markdown environment.

A subset of the data is stored in a Rdata file. In order to read-in the data correctly, save the Rmd file in a folder with the rest of your course work. Place the file baseball.Rdata in the same folder. Go to Session > Set Working Directory > To Source File Location. To load the data run the code below:

```
load("baseball.Rdata")
```

Lahman data

Accessing the data

Using the `ls()` function to see the objects loaded from the baseball data set. Function `ls()` lists all the objects in the current R environment.

You may see other objects from previous instances of work in R.

```
ls()
```

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

To access the content of an object in `R` use the object's name. Keep in mind that `R` is case sensitive. Thus, we need to type an object's name exactly as it appears.

Above we see the object `CarltonFiskBA`. Run the code below to see the contents of `CarltonFiskBA`.

```
CarltonFiskBA
```

```
 1969  1971  1972  1973  1974  1975  1976  1977  1978  1979  1980  1981  1982
0.000 0.312 0.293 0.246 0.299 0.331 0.255 0.315 0.284 0.272 0.289 0.263 0.267
 1983  1984  1985  1986  1987  1988  1989  1990  1991  1992  1993
0.289 0.231 0.238 0.221 0.256 0.277 0.293 0.285 0.241 0.229 0.189
```

Baseball abbreviations

Abbreviation	Meaning
BA	Batting Average
HR	Home Runs
RBI	Runs Batted In

These are some measures of a batter's success.

Descriptive statistics

The names of many functions in `R` are self-explanatory. To compute the minimum, maximum, and mean for Carlton Fisk's career batting average we can use the corresponding functions given below.

```
min(CarltonFiskBA)
```

```
[1] 0
```

```
max(CarltonFiskBA)
```

```
[1] 0.331
```

```
mean(CarltonFiskBA)
```

```
[1] 0.2572917
```

To find the year in which Carlton Fisk had his lowest batting average and the year in which he had his highest batting average, we can make use of the functions `which.min()` and `which.max()`, respectively.

```
which.min(CarltonFiskBA)
```

```
1969
   1
```

```
which.max(CarltonFiskBA)
```

```
1975
   6
```

Let's examine how Carlton Fisk's batting average changed throughout his career. First, we compute year-over-year differences, then view the results. Second, we will look at which year he had the largest increase and which year he had the largest decrease.

```
# compute differences
CarltonFiskBA_diffs <- diff(CarltonFiskBA, lag = 1)
CarltonFiskBA_diffs
```

```
  1971   1972   1973   1974   1975   1976   1977   1978   1979   1980   1981 
 0.312 -0.019 -0.047  0.053  0.032 -0.076  0.060 -0.031 -0.012  0.017 -0.026 
  1982   1983   1984   1985   1986   1987   1988   1989   1990   1991   1992 
 0.004  0.022 -0.058  0.007 -0.017  0.035  0.021  0.016 -0.008 -0.044 -0.012 
  1993 
-0.040 
```

```
# find years
which.max(CarltonFiskBA_diffs)
```

```
1971 
   1 
```

```
which.min(CarltonFiskBA_diffs)
```

```
1976 
   6 
```

The # symbol was used to add comments. `R` does not execute anything following #. Use # for code documentation to explain to others why you are doing what you are doing with your code. Good code documentation is also beneficial for your future self.

Summary statistics with two variables

Recall that the correlation measures the linear strength between two quantitative variables. Let's look at the correlation between each pair of available variables for Jim Rice: batting average, home runs, and RBIs.

```
cor(CarltonFiskBA, CarltonFiskHR)
```

```
[1] 0.3403434
```

```
cor(CarltonFiskBA, CarltonFiskRBI)
```

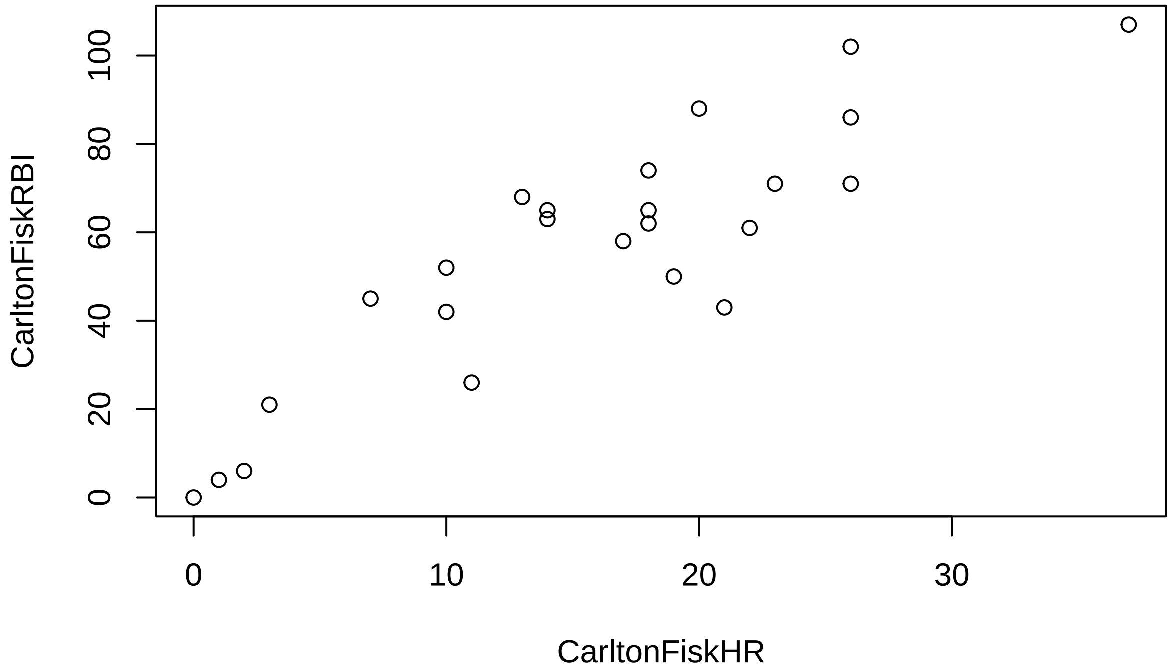
```
[1] 0.4379065
```

```
cor(CarltonFiskHR, CarltonFiskRBI)
```

```
[1] 0.8871394
```

To view a simple plot of Fisk's home runs versus his RBIs we can use the `plot()` function.

```
plot(CarltonFiskHR, CarltonFiskRBI)
```



Exercises

Answer parts a-i below. Use a separate code chunk for each part that requires code. You will examine data on Ted Williams.

To remind yourself of the variable names use the function `ls()`.

a. Use the `length` function to determine how many seasons Ted Williams played.

```
length(TedWilliamsBA)
```

```
[1] 19
```

b. Which season did Ted Williams have his highest batting average?

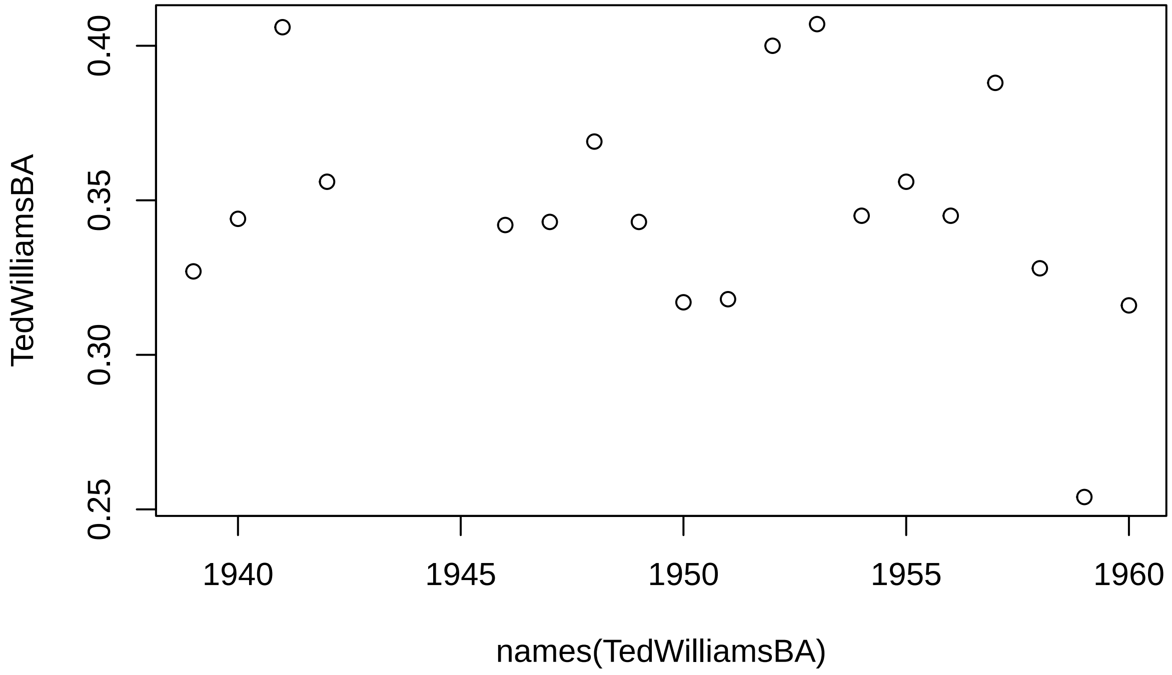
```
which.max(TedWilliamsBA)
```

```
1953
  12
```

```
#1953
```

c. Plot Williams' batting average over time. To put the years on the x-axis, use `names(TedWilliamsBA)`.

```
plot(names(TedWilliamsBA), TedWilliamsBA)
```



d. What was Williams' highest batting average?

```
which.max(TedWilliamsBA)
```

```
1953
  12
```

```
#12
```

e. What was Williams' career mean batting average?

```
mean(TedWilliamsBA)
```

```
[1] 0.3475789
```

f. What was the correlation between Williams' home runs and RBIs? Was it higher than Jim Rice's correlation?

```
cor(TedWilliamsHR, TedWilliamsRBI)
```

```
[1] 0.8422571
```

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

```
[1] 0.9305225
```

```
#Jim Rice had a higher correlation between homeruns and RBIs than Williams
```

g. What was the largest absolute change in Williams' RBIs year-over-year?

```
TedWilliamsRBI_diffs <- diff(TedWilliamsRBI, lag = 1)
TedWilliamsRBI_diffs
```

```
1940 1941 1942 1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 
 -32    7   17  -14   -9   13   32  -62   29 -123   31   55   -6   -1    5   -2 
1959 1960 
 -42   29 
```

```
# 1952 had the largest absolute change dropping from 29 to -123 and then rising to 31 a year later.
```

h. Why does Ted Williams not have any statistics from 1943 - 1945? Was he hurt?

```
#Ted Williams was on active duty for the Marine corps during 1943-1945, but he was sent back to play in
```

i. Which of the three players (Fisk, Rice, Williams) was most consistent year-over-year with regards to the batting average metric? How did you define consistency?

```
CarltonFiskBA_diffs <- diff(CarltonFiskBA, lag = 1)
CarltonFiskBA_diffs
```

```
  1971   1972   1973   1974   1975   1976   1977   1978   1979   1980   1981 
 0.312 -0.019 -0.047  0.053  0.032 -0.076  0.060 -0.031 -0.012  0.017 -0.026 
  1982   1983   1984   1985   1986   1987   1988   1989   1990   1991   1992 
 0.004  0.022 -0.058  0.007 -0.017  0.035  0.021  0.016 -0.008 -0.044 -0.012 
  1993 
-0.040 
```

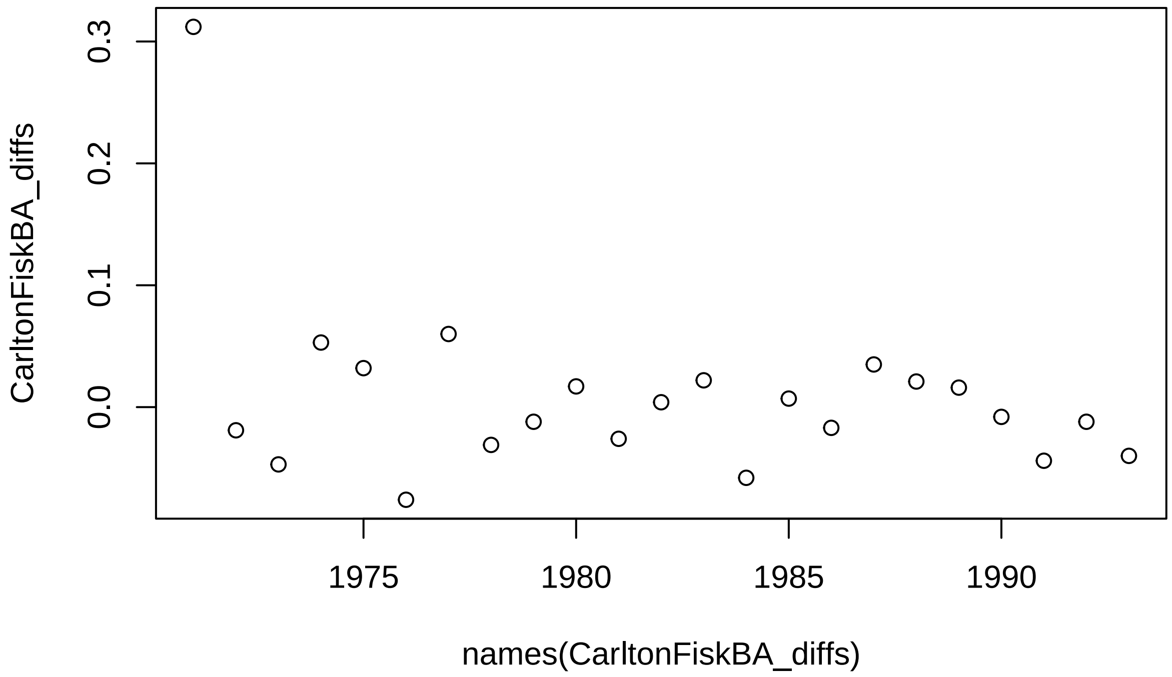
```
TedWilliamsBA_diffs <- diff(TedWilliamsBA, lag = 1)
TedWilliamsBA_diffs
```

```
  1940   1941   1942   1946   1947   1948   1949   1950   1951   1952   1953 
 0.017  0.062 -0.050 -0.014  0.001  0.026 -0.026 -0.026  0.001  0.082  0.007 
  1954   1955   1956   1957   1958   1959   1960 
-0.062  0.011 -0.011  0.043 -0.060 -0.074  0.062 
```

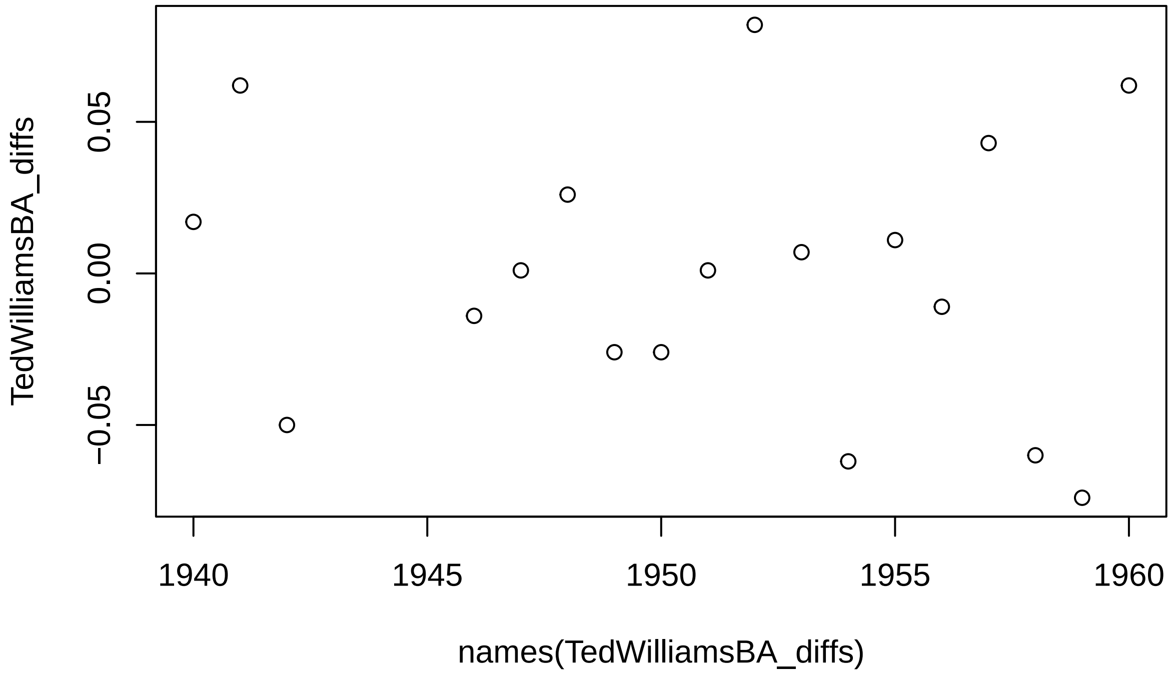
```
JimRiceBA_diffs <- diff(JimRiceBA, lag = 1)
JimRiceBA_diffs
```

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

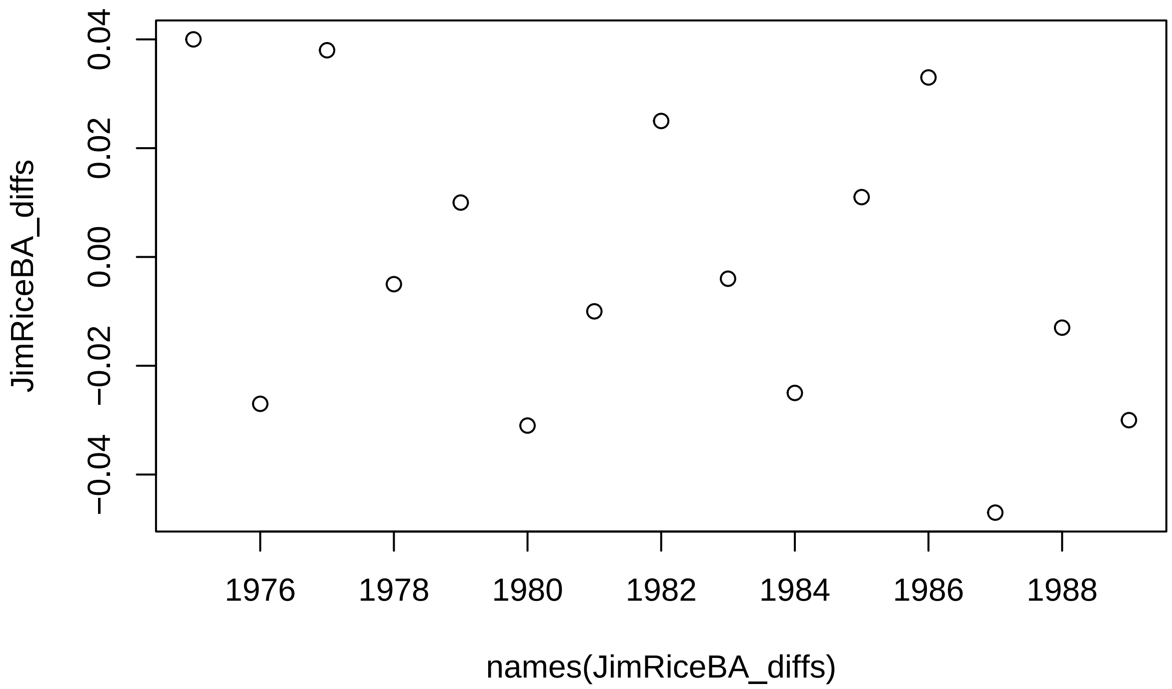
```
plot(names(CarltonFiskBA_diffs),CarltonFiskBA_diffs)
```



```
plot(names(TedWilliamsBA_diffs), TedWilliamsBA_diffs)
```



```
plot(names(JimRiceBA_diffs),JimRiceBA_diffs)
```



```
#Carlton had the most consistency because there is least year-to-year difference between his batting av
```

R Markdown practice

Exercises

Create the R Markdown file that produced the HTML file `RMarkdown practice`. All formatting should match, but you may replace my name with your group name. Below are some helpful hints.

1. YAML header should be

```
---
title: "RMarkdown practice"
author: " "
date: "Septmeber 15, 2021"
output:
  html_document:
    toc: true
    number_sections: true
    toc_float: true
    df_print: paged
---
```

2. In the Narration section you will use the `summary()`, `apply` functions. Check the function by using `?funtion_name`. Before using the `apply` function you will need to subset the dataset so that only numerical variables are included.

3. To create the Plot 2 use the below code. You will need to **install the `ggplot2` package** before you can load them with the `library` function.

```r
library(ggplot2)
ggplot(iris, aes(Petal.Length, Petal.Width, colour = Species)) +
    geom_point(aes(size = Sepal.Length), alpha = 0.7) +
  scale_size(range = c(2, 8))
```

Essential details

Deadline and submission

The M1 ICA4 is an in-class **group** assignment. However, each member of the group who worked together in class will submit their in-class assignment at the end of the class time. Submit your work by uploading both your RMD and HTML/PDF files through D2L. Although the expectation is to get it done within the class period, if a group needs a little bit of extra time to finish, they can submit the assignment on or before **11 PM of Sept 15,2021**. Late work after this deadline will not be accepted except under certain extraordinary circumstances.

Grading

ICAs are graded for participation and completion. Each team member who was present in class and worked on the ICA should submit their files. To grade the ICAs, the instructional team will pick one submission at random from each group. Thus it is important that there is good intra-group communication and teamwork. Each group should ensure that everyone in the team understood, worked through and completed the assignment.

References

1. Lahman, S. (2017) Lahman's Baseball Database, 1871-2016, Main page, http://www.seanlahman.com/baseball-archive/statistics/

2. https://en.wikipedia.org/wiki/Ted_Williams