MA388 Sabermetrics: Lesson 2

Introduction to R

```r
library(Lahman)
library(tidyverse)
library(knitr)
```

Compiling an .qmd file

You'll submit most assignments in this course as PDFs via Canvas. Let's make sure you can produce a PDF from Quarto Markdown. If you already know how to do this, sit tight for a minute…

If this is new for you, you'll need to install a TeX distribution. This is easy with `TinyTeX`.

```r
install.packages("tinytex")
```

With that package loaded, you're ready to install TeX.

```r
tinytex::install_tinytex()
```

Now you should be able to download the source file for these class notes (see "lsn_2_intro_to_R.qmd") and "Render to PDF." Make sure this works before the end of class today.

R Syntax Best Practices

Variable Assignemnt

A lot of folks like to use the `=` sign to assign variable names. While it generally does work, the more appropriate syntax is `->` or `<-`, depending on which side of the argument your variable name is. `=` is used more for

Piping

The pipe passes the output of the previous argument to the first argument of the following function. It's incredibly useful for data munging. You are probably most familiar with the pipe that looks like `%>%`. However, that pipe is specific to the `dplyr` package. There is now a pipe that is native to base R that is probably better to use `|>`.

Review

- You are investigating whether the hit-and-run is a good strategy. Which data set should you use?
- You are investigating whether pitchers throw more fastballs to hitters at the bottom of the batting order. Which data set would you use?
- Use R to find career batting average (H/AB) leaders of your favorite team. More specifically, select a team and report a table of the 10 players with at least 2500 at bats (AB) with the highest batting averages while playing for that team.

Table 1: Red Sox Career Batting Average Leaders

playerID	H	AB	AVG
willite01	2654	7706	0.344
boggswa01	2098	6213	0.338
speaktr01	1327	3935	0.337
garcino01	1281	3968	0.323
runnepe01	825	2578	0.320
foxxji01	1051	3288	0.320
peskyjo01	1277	4085	0.313
ramirma02	1232	3953	0.312
lynnfr01	944	3062	0.308
goodmbi01	1344	4399	0.306

Merging Data Frames (pg 41)

When using a relational database (what does that mean again?), a common task is to add information from one table to another. For example, consider our table above from the review. Add the player's name and when they played their final game.

```
# I called my table from the review redSox.

redSox <- redSox |>
  left_join(select(People, playerID, nameLast, nameFirst, finalGame), by = "playerID" )
```

Your turn…the `HallofFame` data frame contains data on every Hall of Fame ballot. Add to your table above whether the player is in the Hall of Fame. (hint: `filter(inducted == "Y")`)

Table 2: Red Sox career batting average leaders

playerID	H	AB	AVG	nameLast	nameFirst	finalGame	yearID	inducted
willite01	2654	7706	0.344	Williams	Ted	1960-09-28	1966	Y
boggswa01	2098	6213	0.338	Boggs	Wade	1999-08-27	2005	Y
speaktr01	1327	3935	0.337	Speaker	Tris	1928-08-30	1937	Y

playerID	H	AB	AVG	nameLast	nameFirst	finalGame	yearID	inducted
garcino01	1281	3968	0.323	Garciaparra	Nomar	2009-10-04	NA	NA
runnepe01	825	2578	0.320	Runnels	Pete	1964-05-14	NA	NA
foxxji01	1051	3288	0.320	Foxx	Jimmie	1945-09-23	1951	Y
peskyjo01	1277	4085	0.313	Pesky	Johnny	1954-09-24	NA	NA
ramirma02	1232	3953	0.312	Ramirez	Manny	2011-04-06	NA	NA
lynnfr01	944	3062	0.308	Lynn	Fred	1990-10-03	NA	NA
goodmbi01	1344	4399	0.306	Goodman	Billy	1962-09-30	NA	NA

What cleaning would you still want to do to this table to make it more readable?

What is the difference between left_join(), right_join(), inner_join(), and outer_join()?