

Introduction to resources and data sets

Daniel J. Eck

Welcome

Welcome to STAT 430: Baseball Analytics!

Let's have a fun semester exploring the ability of statistics to quantify winning games and evaluating players.

These brief lecture slides are meant to introduce data sets used in class and supplement Chapter 1 in your textbook. They will also go over course logistics.

GitHub

GitHub is a a cloud-based service that implements a Git repository hosting system.

Course materials will be distributed from my GitHub organization which was built using CS's GitHub-repo-creator. Your repo can be created through this link

https://edu.cs.illinois.edu/create-gh-repo/sp25_stat430

See the `setup.md` file in the **stat430sp25** root directory for details on setting up GitHub in this course.

Software

- ▶ The R Project for Statistical Computing:
<https://www.r-project.org/>
- ▶ RStudio as an integrated development environment for R:
<https://www.rstudio.com/>

Install both R and RStudio and consider creating an [RStudio project](#) for better organization.

Data sets used in class

- ▶ Lahman package:

```
install.packages ("Lahman")
```

- ▶ Retrosheet data. Appendix A in your textbook provides an R script file for downloading and parsing all the game log files. A possible more convenient approach for obtaining retrosheets is included in this slide deck. Or you can occasionally use the `retrosheet` package for simple retrosheets:

```
install.packages ("retrosheet")
```

- ▶ Statcast data obtained from the `baseballr` package:

```
install.packages("baseballr")
```

- ▶ `baseball_R`. Coding scripts and data sets that supplement your textbook.
- ▶ Statcast data obtained from the `bbd` package:

```
# install.packages("devtools")  
devtools::install_github("daviddalpiaz/bbd")
```

The `bbd` package was created by UIUC CS professor [David Dalpiaz](#). [Here is a link](#) to the GitHub repo.

Statcast and Retrosheet data sets will be provided if past difficulties persist.

► Data scraped from [baseball reference](#):

```
bwar_bat = readr::read_csv("https://www.baseball-reference.com/data/war_daily_bat.txt",  
                           na = "NULL")  
bwar_pit = readr::read_csv("https://www.baseball-reference.com/data/war_daily_pitch.txt",  
                           na = "NULL")
```

► Era-adjusted data sets are accessible in the `fullhouse` R package:

```
#install.packages("devtools")  
install_github(repo = "DEck13/fullhouse")
```

This is beta version of an R package developed by myself and my students. It contains era-adjusted baseball statistics that appear on [our website](#). These stats are obtained through an application of what we call a “Full House Model,” a statistical approach presented in a paper that was recently accepted for publication.

Introduction to Lahman

The `Lahman` package contains several tables consisting of useful stat. We highlight a few tables below

```
#install.packages("Lahman")  
library(Lahman)  
data(Batting)  
data(Pitching)  
data(Fielding)  
data(Teams)
```


Lahman Batting table

```
head(Batting)
```

[illegible]

Lahman Pitching table

```
head(Pitching)
```

##	playerID	yearID	stint	teamID	lgID	W	L	G	GS	CG	SHO	SV	IP	Pouts	H	ER	HR	BB	SO
## 1	aardsda01	2004	1	SFN	NL	1	0	11	0	0	0	0	32	20	8	1	10	5	
## 2	aardsda01	2006	1	CHN	NL	3	0	45	0	0	0	0	159	41	24	9	28	49	
## 3	aardsda01	2007	1	CHA	AL	2	1	25	0	0	0	0	97	39	23	4	17	36	
## 4	aardsda01	2008	1	BOS	AL	4	2	47	0	0	0	0	146	49	30	4	35	49	
## 5	aardsda01	2009	1	SEA	AL	3	6	73	0	0	0	38	214	49	20	4	34	80	
## 6	aardsda01	2010	1	SEA	AL	0	6	53	0	0	0	31	149	33	19	5	25	49	
##	BAOpp	ERA	IBB	WP	HBP	BK	BFP	GF	R	SH	SF	GIDP							
## 1	0.417	6.75	0	0	2	0	61	5	8	0	1	1							
## 2	0.214	4.08	0	1	1	0	225	9	25	1	3	2							
## 3	0.300	6.40	3	2	1	0	151	7	24	2	1	1							
## 4	0.268	5.55	2	3	5	0	228	7	32	3	2	4							
## 5	0.190	2.52	3	2	0	0	296	53	23	2	1	2							
## 6	0.198	3.44	5	2	2	0	202	43	19	7	1	5							

Lahman Fielding table

```
head(Fielding)
```

##	playerID	yearID	stint	teamID	lgID	POS	G	GS	InnOuts	PO	A	E	DP	PB	WP	SB	CS	ZR
## 1	aardsda01	2004	1	SFN	NL	P	11	0	32	0	0	0	0	NA	NA	NA	NA	NA
## 2	aardsda01	2006	1	CHN	NL	P	45	0	159	1	5	0	1	NA	NA	NA	NA	NA
## 3	aardsda01	2007	1	CHA	AL	P	25	0	97	2	4	1	0	NA	NA	NA	NA	NA
## 4	aardsda01	2008	1	BOS	AL	P	47	0	146	3	6	0	0	NA	NA	NA	NA	NA
## 5	aardsda01	2009	1	SEA	AL	P	73	0	214	2	5	0	1	NA	NA	NA	NA	NA
## 6	aardsda01	2010	1	SEA	AL	P	53	0	149	2	3	1	0	NA	NA	NA	NA	NA

Lahman Teams table

```
head(Teams, 3)
```

```
##   yearID lgID teamID franchID divID Rank  G  Ghome  W  L DivWin WCWin LgWin
## 1  1871   NA   BS1      BNA  <NA>    3 31    NA 20 10  <NA>  <NA>    N
## 2  1871   NA   CH1      CNA  <NA>    2 28    NA 19  9  <NA>  <NA>    N
## 3  1871   NA   CL1      CFC  <NA>    8 29    NA 10 19  <NA>  <NA>    N
##   WSWin  R  AB  H X2B X3B HR BB SO SB CS HBP SF  RA  ER  ERA CG SHO SV
## 1  <NA> 401 1372 426  70  37  3 60 19 73 16  NA NA 303 109 3.55 22  1  3
## 2  <NA> 302 1196 323  52  21 10 60 22 69 21  NA NA 241  77 2.76 25  0  1
## 3  <NA> 249 1186 328  35  40  7 26 25 18  8  NA NA 341 116 4.11 23  0  0
##   IPouts  HA HRA BBA SOA  E DP  FP      name
## 1    828 367  2  42  23 243 24 0.834  Boston Red Stockings
## 2    753 308  6  28  22 229 16 0.829  Chicago White Stockings
## 3    762 346 13  53  34 234 15 0.818  Cleveland Forest Citys
##   park attendance BPF PPF teamIDBR teamIDlahman45
## 1      South End Grounds I      NA 103  98      BOS      BS1
## 2      Union Base-Ball Grounds      NA 104 102      CHI      CH1
## 3 National Association Grounds      NA  96 100      CLE      CL1
##   teamIDretro
## 1      BS1
## 2      CH1
## 3      CL1
```

Retrosheets

There is a lot of box score information contained in a retrosheet.

Basic retrosheets can be obtained from the `retrosheet` package (the following code chunk has `eval = FALSE` because the retrosheet will not fit on a single slide).

```
library(retrosheet)  
getRetrosheet(type = "game", year = 2012)
```

More comprehensive retrosheets can be obtained from the `baseballr` package (the following code chunk has `eval = FALSE` because the retrosheet will take awhile to load and will be stored locally).

```
library(baseballr)
retrosheet_data(path_to_directory = "~/Desktop/baseball_course/retrosheet",
                 years_to_acquire = 1998)
```

Obtaining retrosheets via `baseballr` requires some work outlined by Bill Petti [here](#).

The steps in the hyperlink above require one to first download and install files from [the Chadwick Bureau](#).

Follow the instructions in the `INSTALL` file in the downloaded Chadwick tarball (this course used `chadwick-0.9.5`).