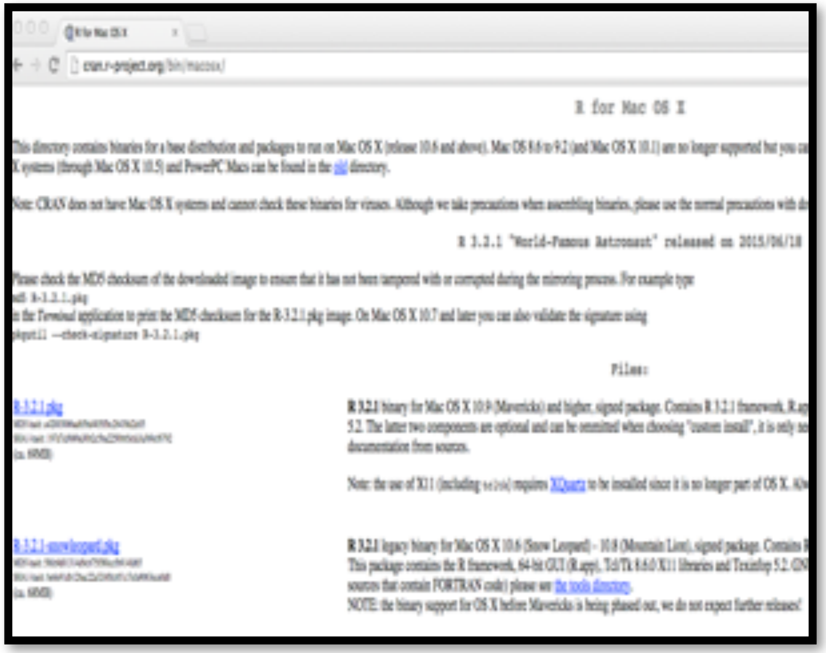
Putting the R in Sports

By: Daniel Willis

Installing R Easy



R for Mac OS X

This directory contains binaries for a base distribution and packages to run on Mac OS X (release 10.6 and above). Mac OS 8.6 to 9.2 (and Mac OS X 10.1) are no longer supported but you ca
X systems (through Mac OS X 10.5) and PowerPC Macs can be found in the old directory.

Note: CRAN does not have Mac OS X systems and cannot check these binaries for viruses. Although we take precautions when assembling binaries, please use the normal precautions with do

R 3.2.1 "World-Famous Astronaut" released on 2015/06/18

Please check the MD5 checksum of the downloaded image to ensure that it has not been tampered with or corrupted during the mirroring process. For example type
md5 R-3.2.1.pkg
in the *Terminal* application to print the MD5 checksum for the R-3.2.1.pkg image. On Mac OS X 10.7 and later you can also validate the signature using
pkgutil --check-signature R-3.2.1.pkg

Files:

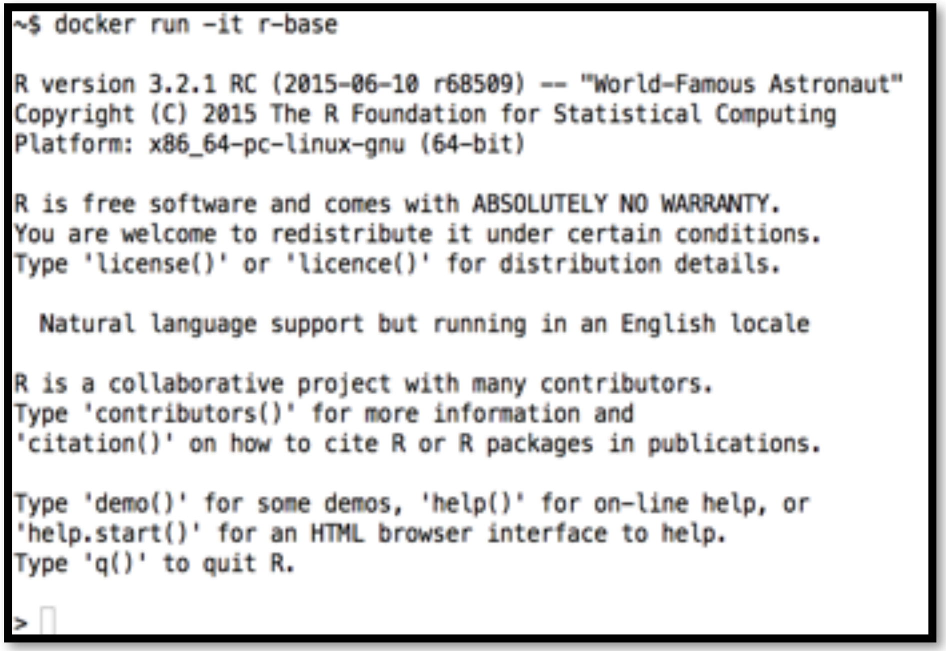
R-3.2.1.pkg
MD5-hash: [illegible]
SHA1-hash: [illegible]
(ca. 69MB)

R 3.2.1 binary for Mac OS X 10.9 (Mavericks) and higher, signed package. Contains R 3.2.1 framework, R.ap
5.2. The latter two components are optional and can be ommitted when choosing "custom install", it is only ne
documentation from sources.

Note: the use of X11 (including tcltk) requires XQuartz to be installed since it is no longer part of OS X. Alw

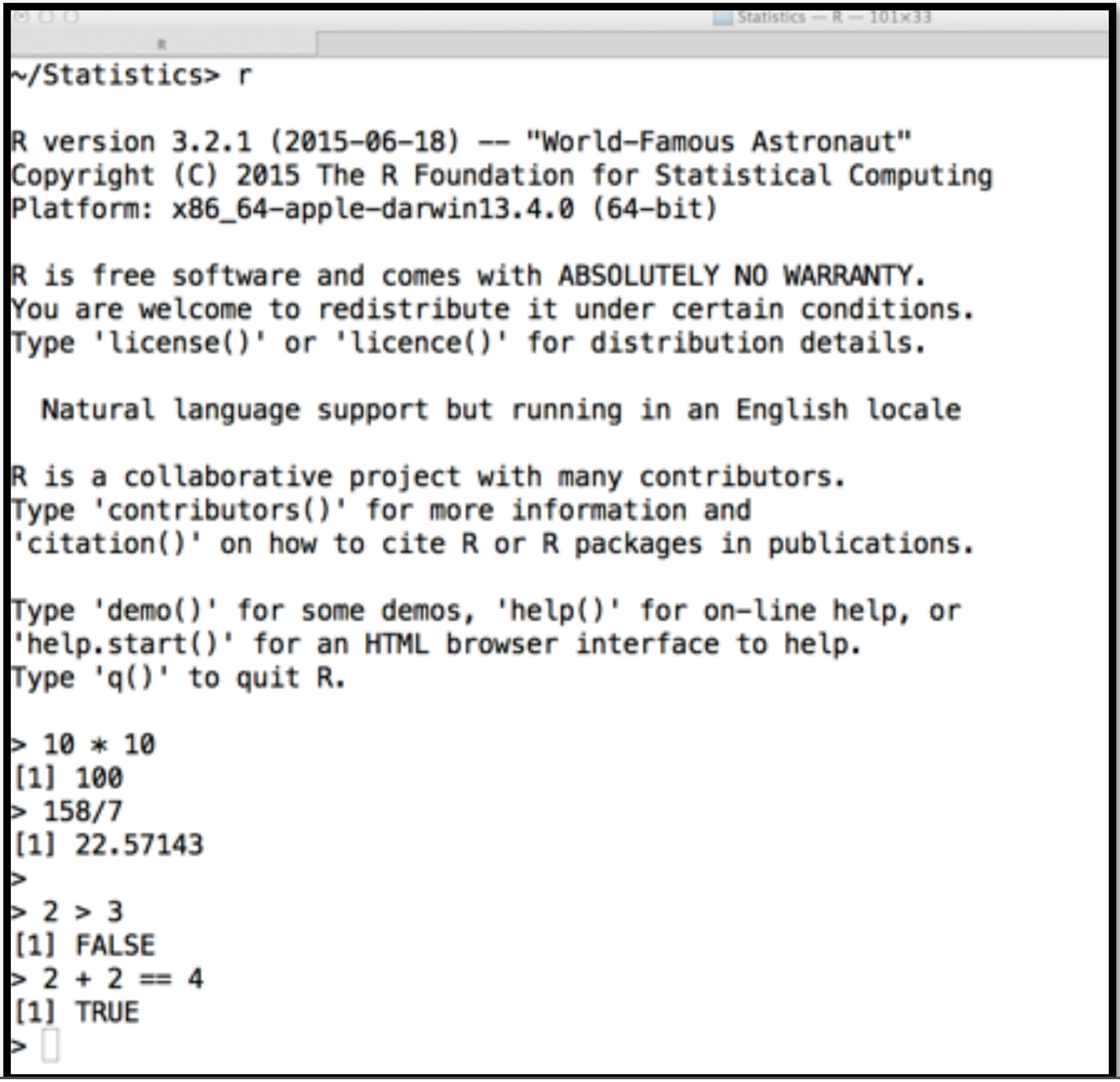
R-3.2.1-snowleopard.pkg
MD5-hash: [illegible]
SHA1-hash: [illegible]
(ca. 68MB)

R 3.2.1 legacy binary for Mac OS X 10.6 (Snow Leopard) - 10.8 (Mountain Lion), signed package. Contains R
This package contains the R framework, 64-bit GUI (R.app), Tcl/Tk 8.6.0 X11 libraries and Texinfo 5.2. GN
sources that contain FORTRAN code) please see the tools directory.
NOTE: the binary support for OS X before Mavericks is being phased out, we do not expect further releases!



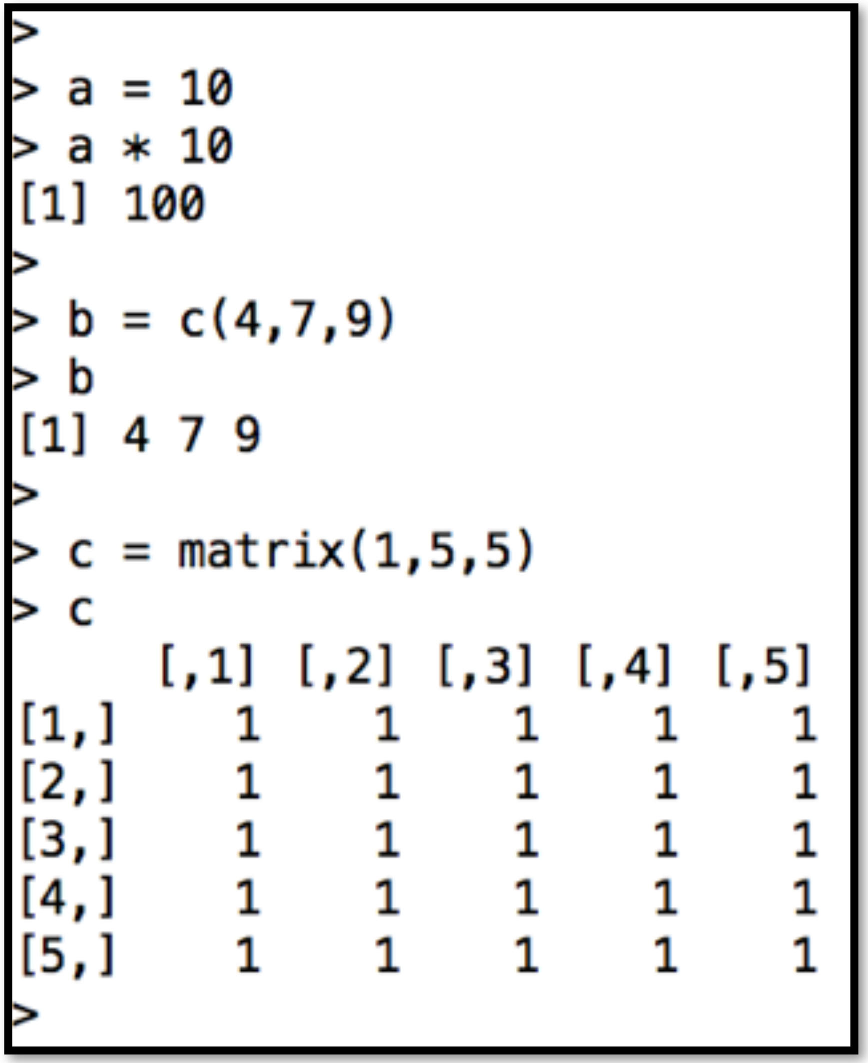
```
~$ docker run -it r-base

R version 3.2.1 RC (2015-06-10 r68509) -- "World-Famous Astronaut"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Type in R to start R



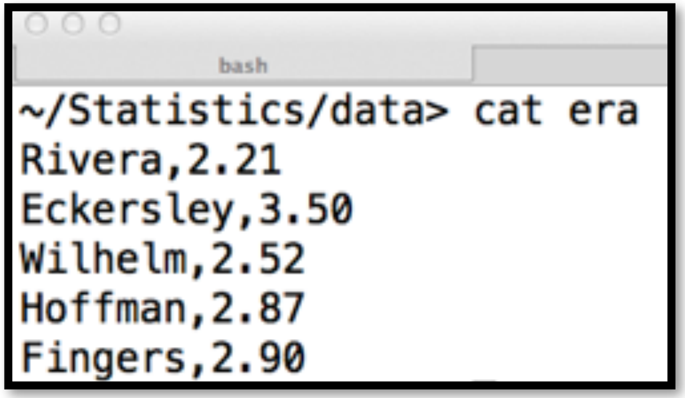
```
~/Statistics> r

R version 3.2.1 (2015-06-18) -- "World-Famous Astronaut"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 10 * 10
[1] 100
> 158/7
[1] 22.57143
>
> 2 > 3
[1] FALSE
> 2 + 2 == 4
[1] TRUE
> 
```

I Learned R Variables

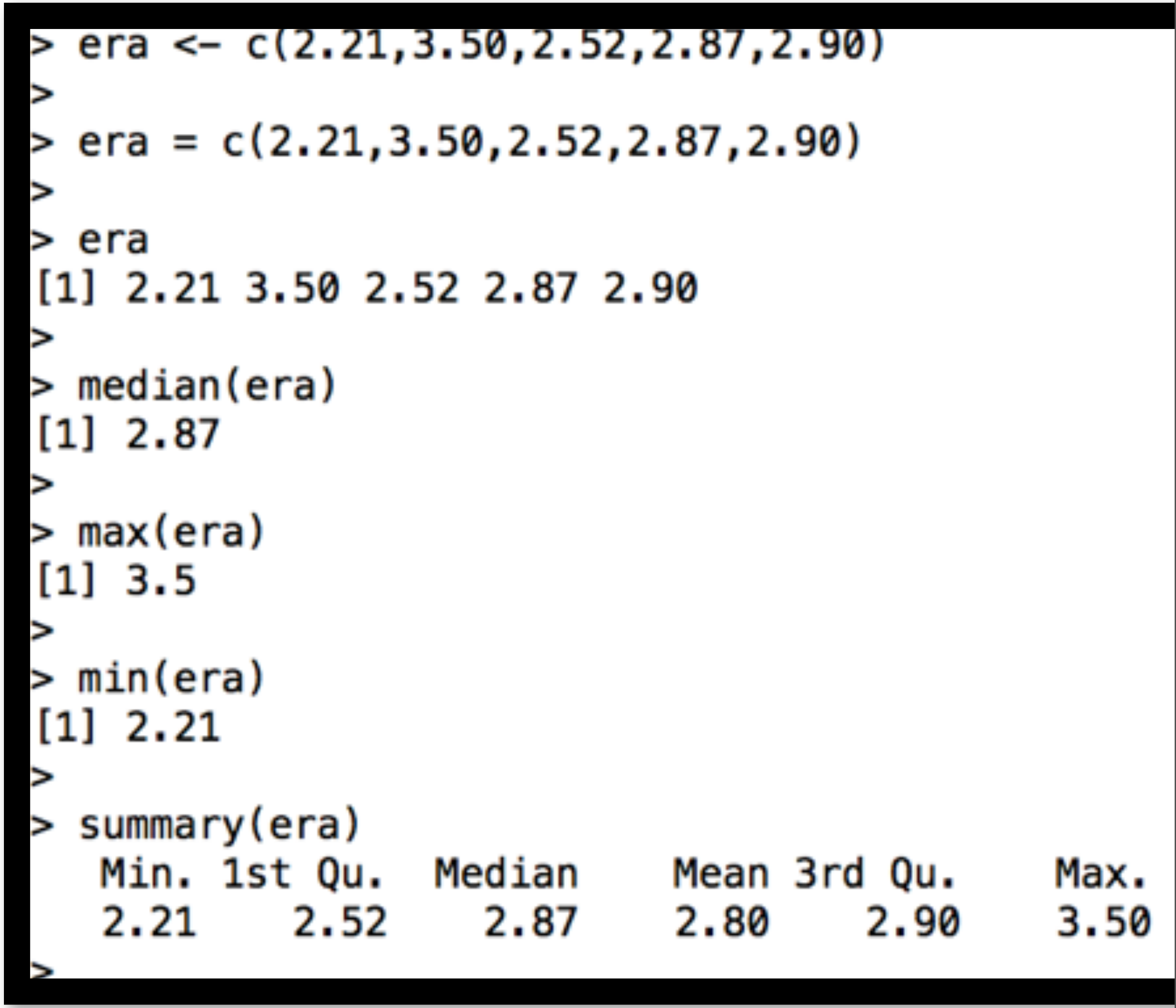


```
>
> a = 10
> a * 10
[1] 100
>
> b = c(4,7,9)
> b
[1] 4 7 9
>
> c = matrix(1,5,5)
> c
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    1    1    1    1
[2,]    1    1    1    1    1
[3,]    1    1    1    1    1
[4,]    1    1    1    1    1
[5,]    1    1    1    1    1
>
```

Let’s Look at an Example

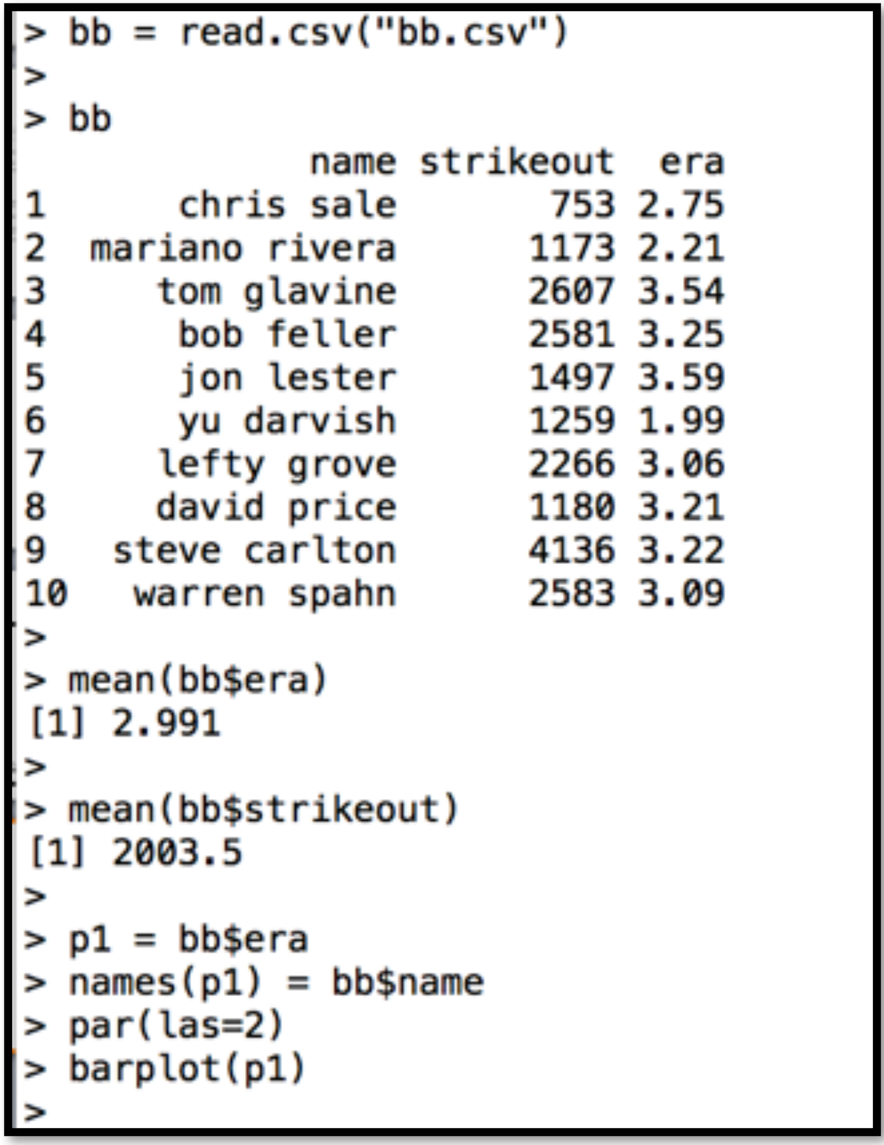


```
~/Statistics/data> cat era
Rivera,2.21
Eckersley,3.50
Wilhelm,2.52
Hoffman,2.87
Fingers,2.90
```

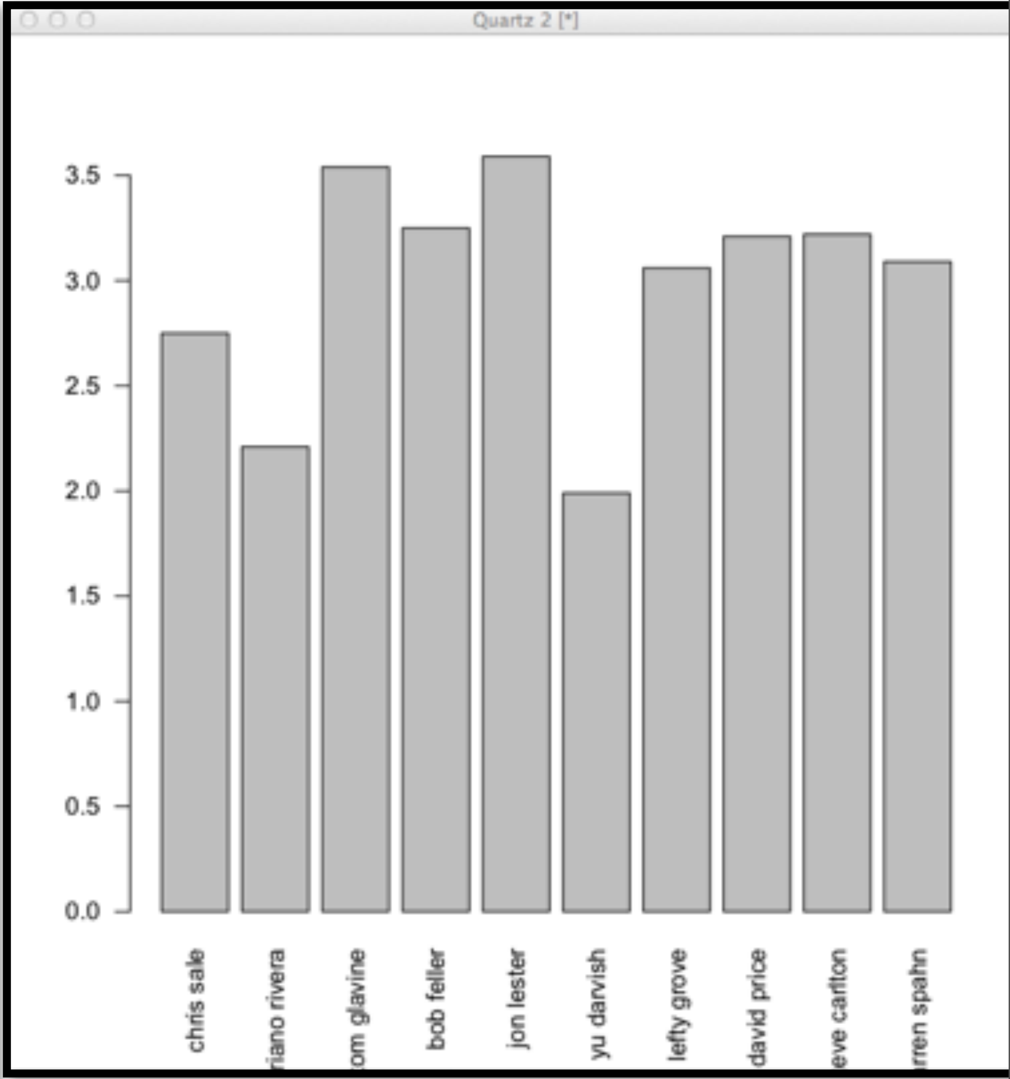


```
> era <- c(2.21,3.50,2.52,2.87,2.90)
>
> era = c(2.21,3.50,2.52,2.87,2.90)
>
> era
[1] 2.21 3.50 2.52 2.87 2.90
>
> median(era)
[1] 2.87
>
> max(era)
[1] 3.5
>
> min(era)
[1] 2.21
>
> summary(era)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.21    2.52    2.87    2.80    2.90    3.50
>
```

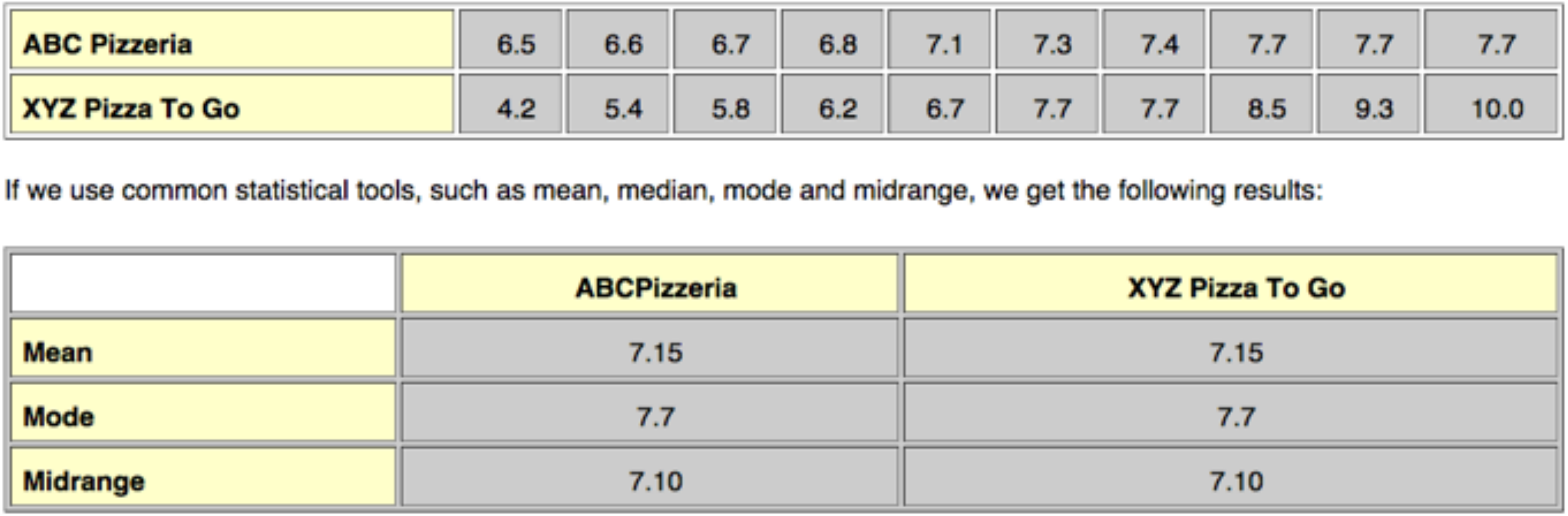
You can read files with R



```
> bb = read.csv("bb.csv")
>
> bb
             name strikeout  era
1      chris sale       753 2.75
2  mariano rivera      1173 2.21
3     tom glavine      2607 3.54
4      bob feller      2581 3.25
5      jon lester      1497 3.59
6      yu darvish      1259 1.99
7     lefty grove      2266 3.06
8     david price      1180 3.21
9   steve carlton      4136 3.22
10   warren spahn      2583 3.09
>
> mean(bb$era)
[1] 2.991
>
> mean(bb$strikeout)
[1] 2003.5
>
> p1 = bb$era
> names(p1) = bb$name
> par(las=2)
> barplot(p1)
>
```



Explaining Standard Deviation is Hard



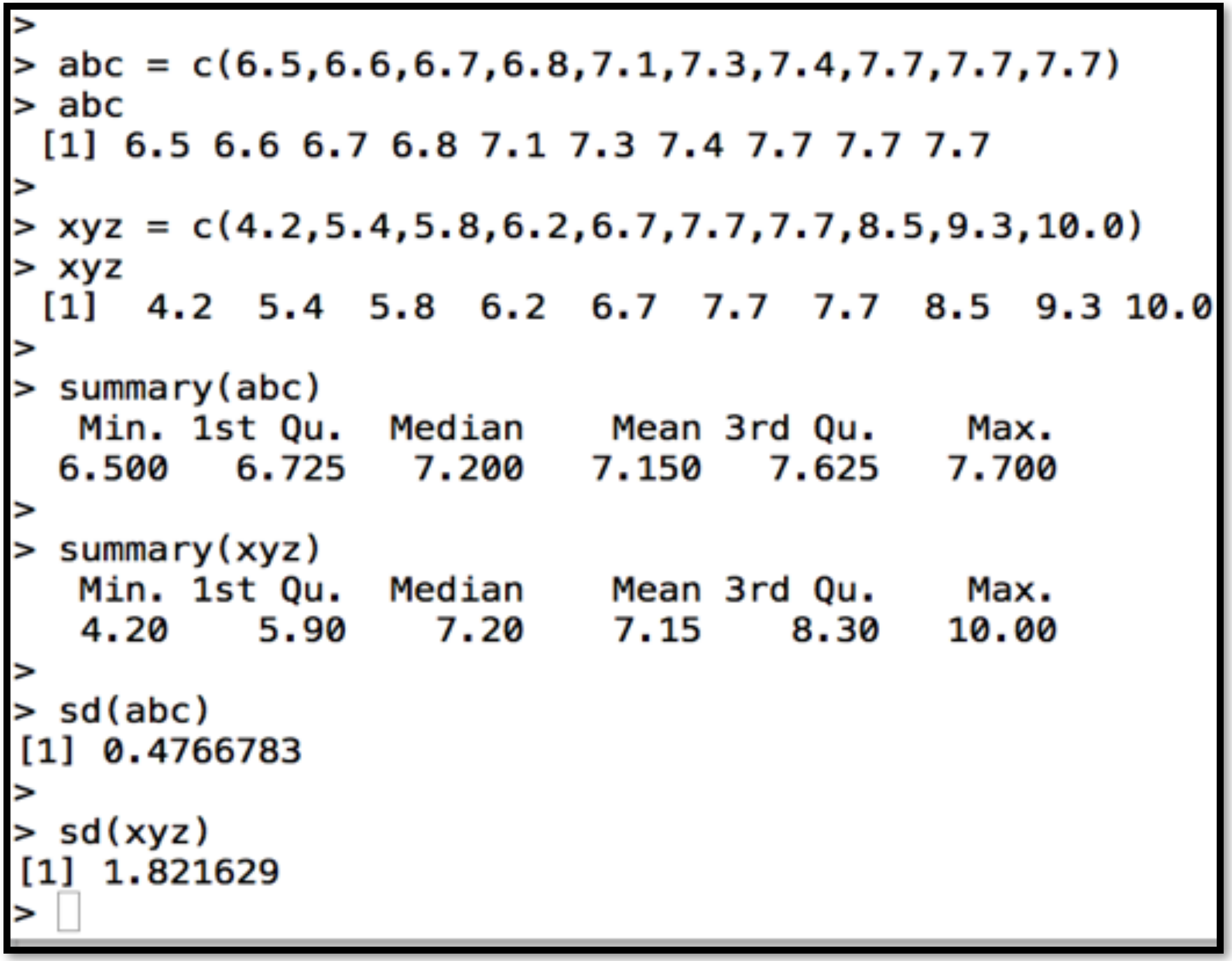
ABC Pizzeria	6.5	6.6	6.7	6.8	7.1	7.3	7.4	7.7	7.7	7.7
XYZ Pizza To Go	4.2	5.4	5.8	6.2	6.7	7.7	7.7	8.5	9.3	10.0

If we use common statistical tools, such as mean, median, mode and midrange, we get the following results:

	ABCPizzeria	**XYZ Pizza To Go**
Mean	7.15	7.15
Mode	7.7	7.7
Midrange	7.10	7.10

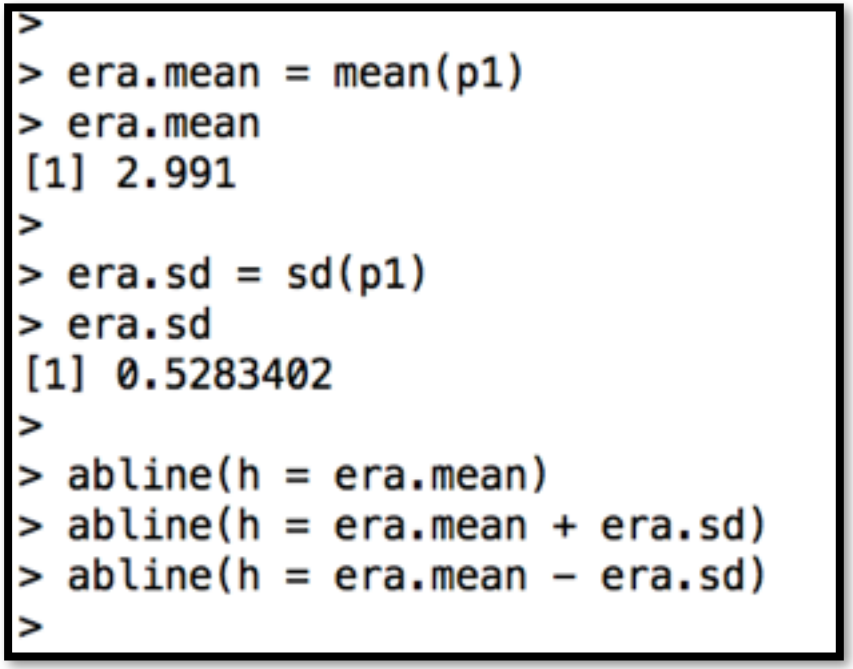
http://www.isixsigma.com/tools-templates/variation/variation-root-all-process-evil/

We put the Pizza Times into Vectors

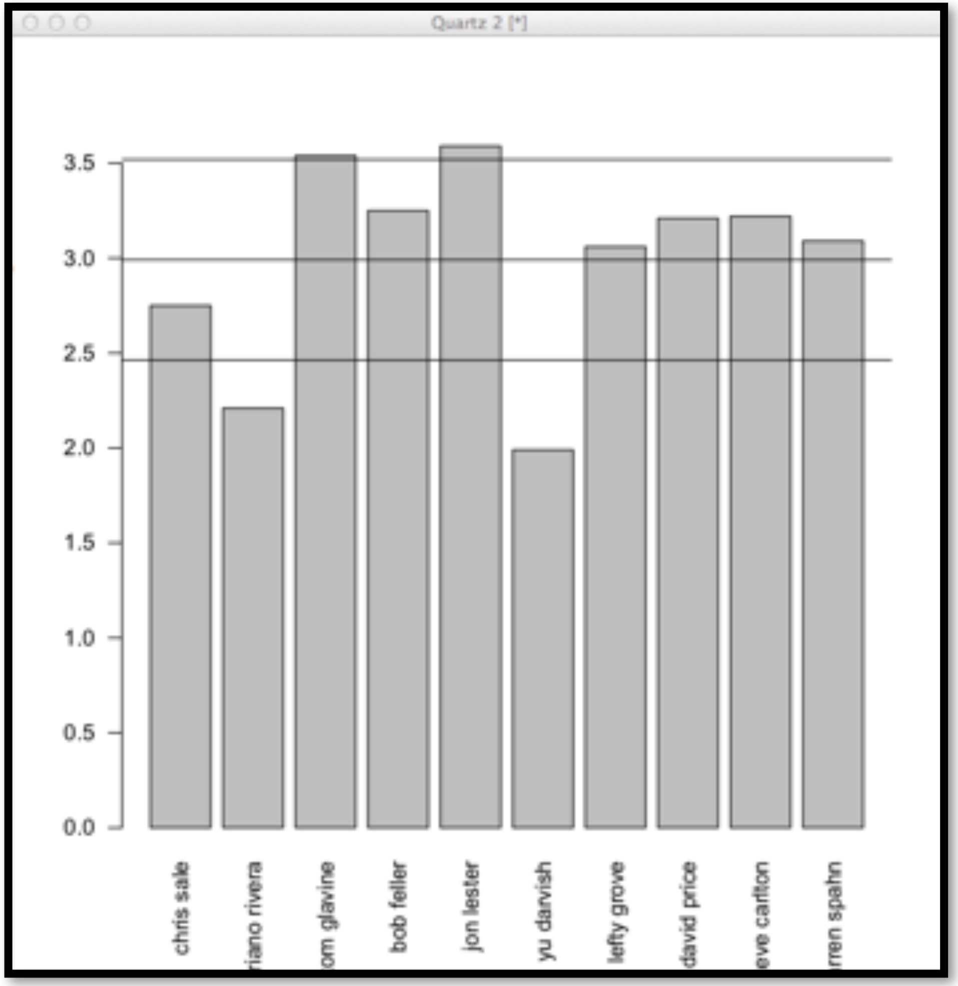


```
>
> abc = c(6.5,6.6,6.7,6.8,7.1,7.3,7.4,7.7,7.7,7.7)
> abc
 [1] 6.5 6.6 6.7 6.8 7.1 7.3 7.4 7.7 7.7 7.7
>
> xyz = c(4.2,5.4,5.8,6.2,6.7,7.7,7.7,8.5,9.3,10.0)
> xyz
 [1]  4.2  5.4  5.8  6.2  6.7  7.7  7.7  8.5  9.3 10.0
>
> summary(abc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.500   6.725   7.200   7.150   7.625   7.700
>
> summary(xyz)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.20    5.90    7.20    7.15    8.30   10.00
>
> sd(abc)
[1] 0.4766783
>
> sd(xyz)
[1] 1.821629
>
```

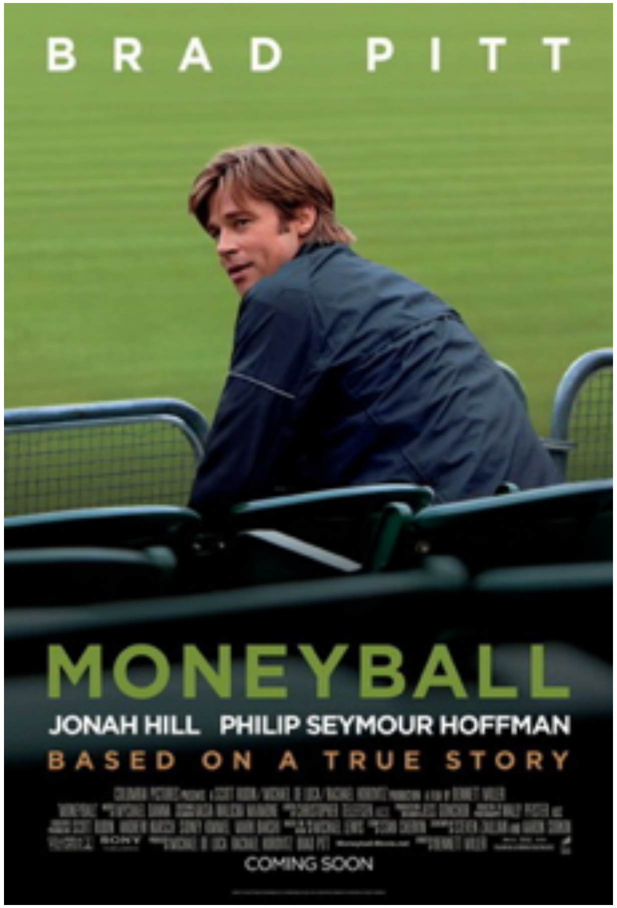
Let’s go back to the earlier ERA Plot

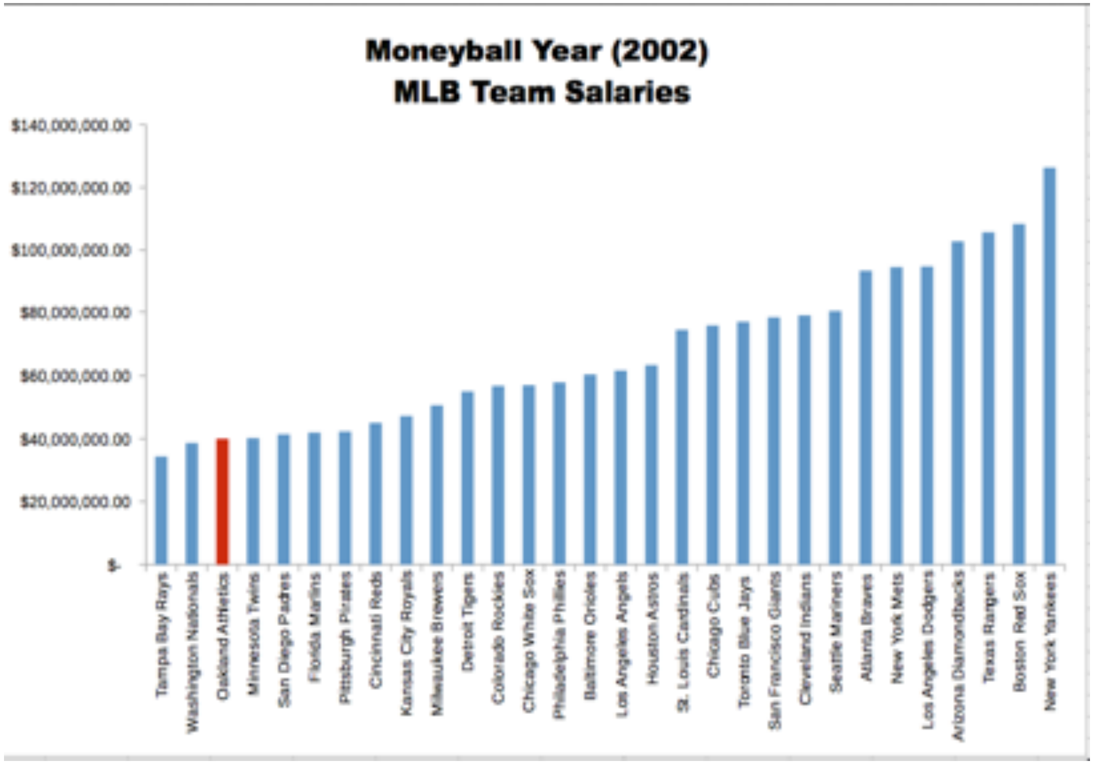


```
>
> era.mean = mean(p1)
> era.mean
[1] 2.991
>
> era.sd = sd(p1)
> era.sd
[1] 0.5283402
>
> abline(h = era.mean)
> abline(h = era.mean + era.sd)
> abline(h = era.mean - era.sd)
>
```



How Many People Have Seen Moneyball?

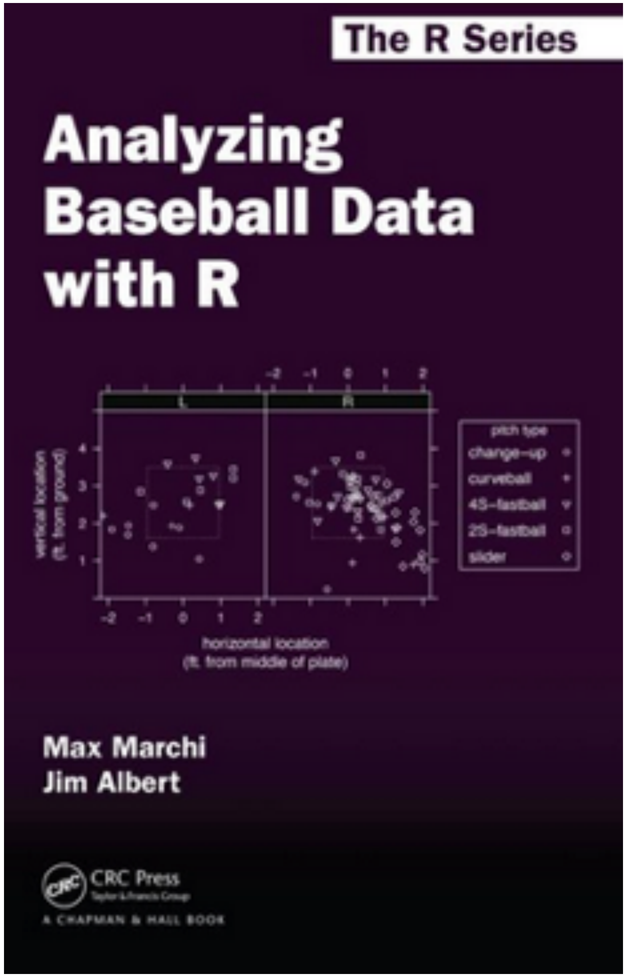




Sabermetrics

- Society for American Baseball Research
- Bill James is the father of Sabermetrics
- He joined the Redsox in 2003
- In 2004 Redsox broke the curse

My Dad and I Found This



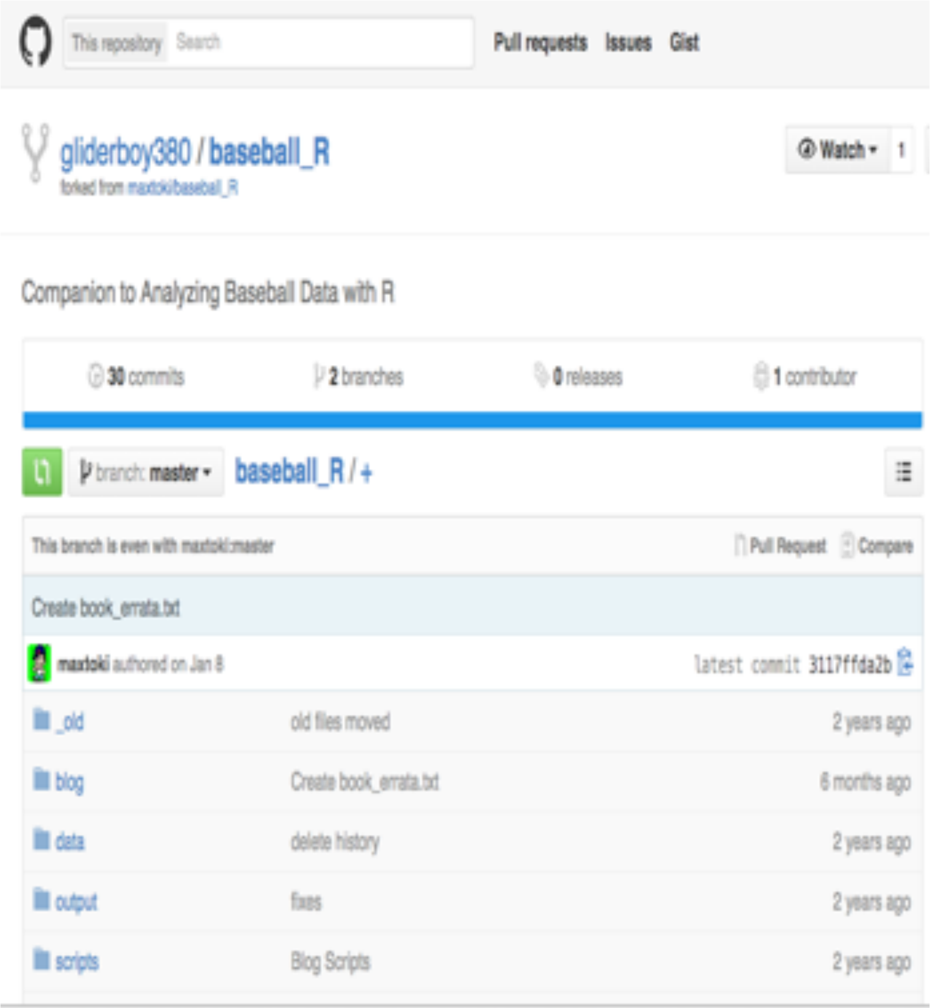
The R Series

Analyzing Baseball Data with R

Max Marchi
Jim Albert

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

L
R
vertical location (ft. from ground)
horizontal location (ft. from middle of plate)
pitch type
change-up
curveball
4S-fastball
2S-fastball
slider



gliderboy380 / baseball_R
forked from maxtoki/baseball_R

Companion to Analyzing Baseball Data with R

30 commits · 2 branches · 0 releases · 1 contributor

branch: master · baseball_R / +

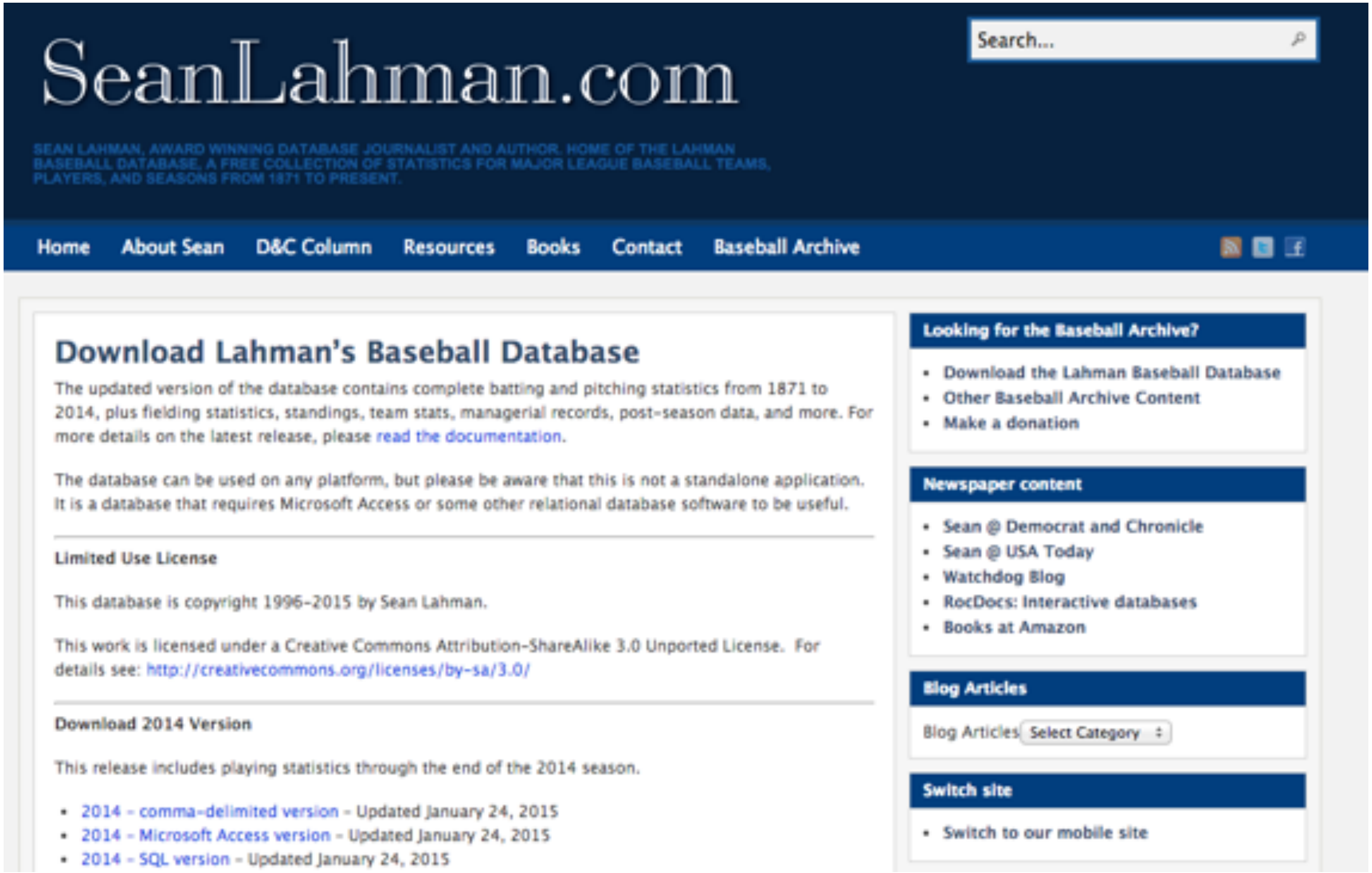
This branch is even with maxtoki:master · Pull Request · Compare

Create book_errata.txt

maxtoki authored on Jan 8 · latest commit 3117ffda2b

_old	old files moved	2 years ago
blog	Create book_errata.txt	6 months ago
data	delete history	2 years ago
output	fixes	2 years ago
scripts	Blog Scripts	2 years ago

Lahman Baseball Database



SEAN LAHMAN, AWARD WINNING DATABASE JOURNALIST AND AUTHOR. HOME OF THE LAHMAN BASEBALL DATABASE, A FREE COLLECTION OF STATISTICS FOR MAJOR LEAGUE BASEBALL TEAMS, PLAYERS, AND SEASONS FROM 1871 TO PRESENT.

Download Lahman's Baseball Database

The updated version of the database contains complete batting and pitching statistics from 1871 to 2014, plus fielding statistics, standings, team stats, managerial records, post-season data, and more. For more details on the latest release, please read the documentation.

The database can be used on any platform, but please be aware that this is not a standalone application. It is a database that requires Microsoft Access or some other relational database software to be useful.

Limited Use License

This database is copyright 1996-2015 by Sean Lahman.

This work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License. For details see: http://creativecommons.org/licenses/by-sa/3.0/

Download 2014 Version

This release includes playing statistics through the end of the 2014 season.

- 2014 - comma-delimited version - Updated January 24, 2015
- 2014 - Microsoft Access version - Updated January 24, 2015
- 2014 - SQL version - Updated January 24, 2015

Looking for the Baseball Archive?

- Download the Lahman Baseball Database
- Other Baseball Archive Content
- Make a donation

Newspaper content

- Sean @ Democrat and Chronicle
- Sean @ USA Today
- Watchdog Blog
- RocDocs: Interactive databases
- Books at Amazon

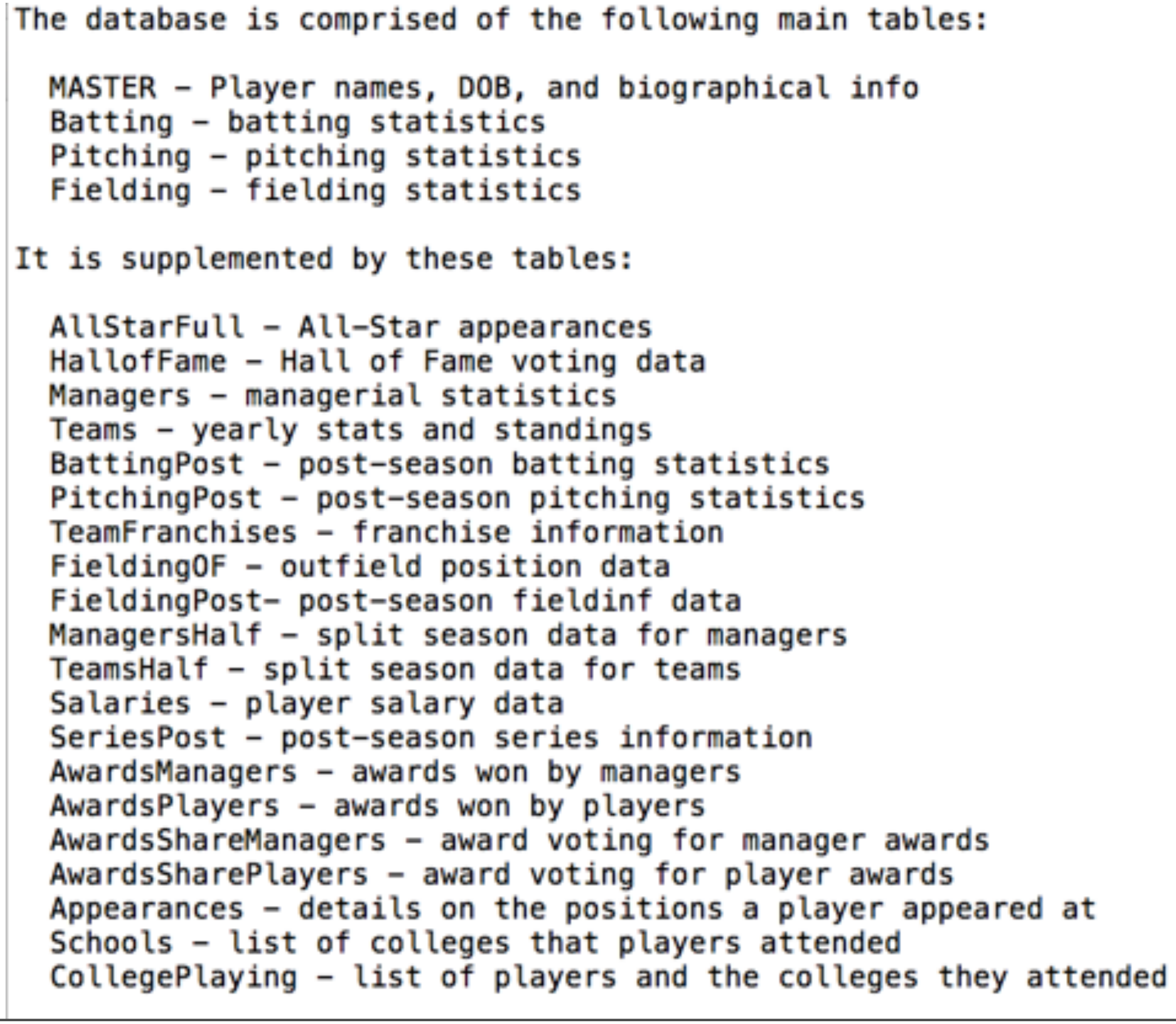
Blog Articles

Blog Articles Select Category

Switch site

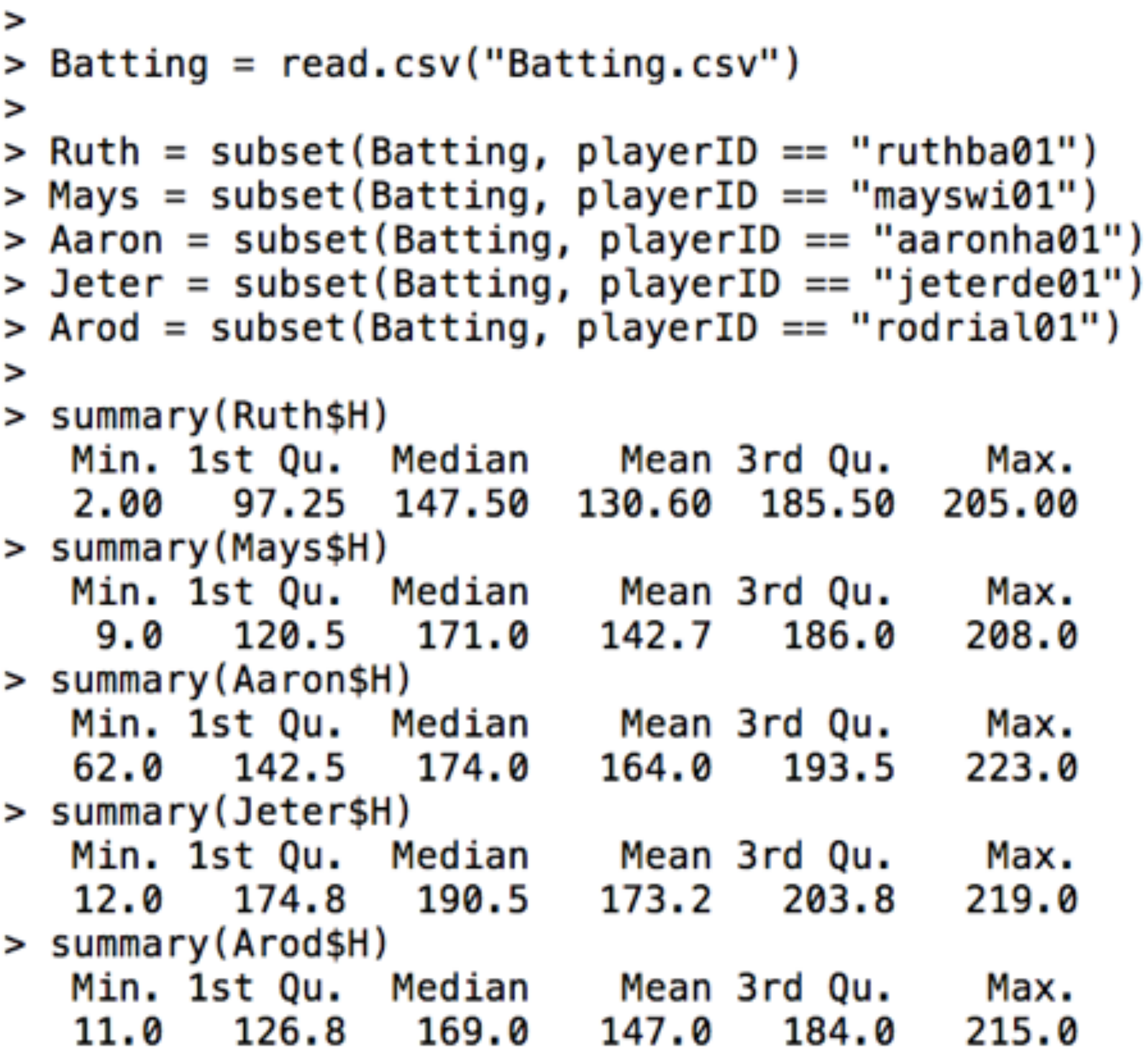
- Switch to our mobile site

The Lahman Database Has Great Examples



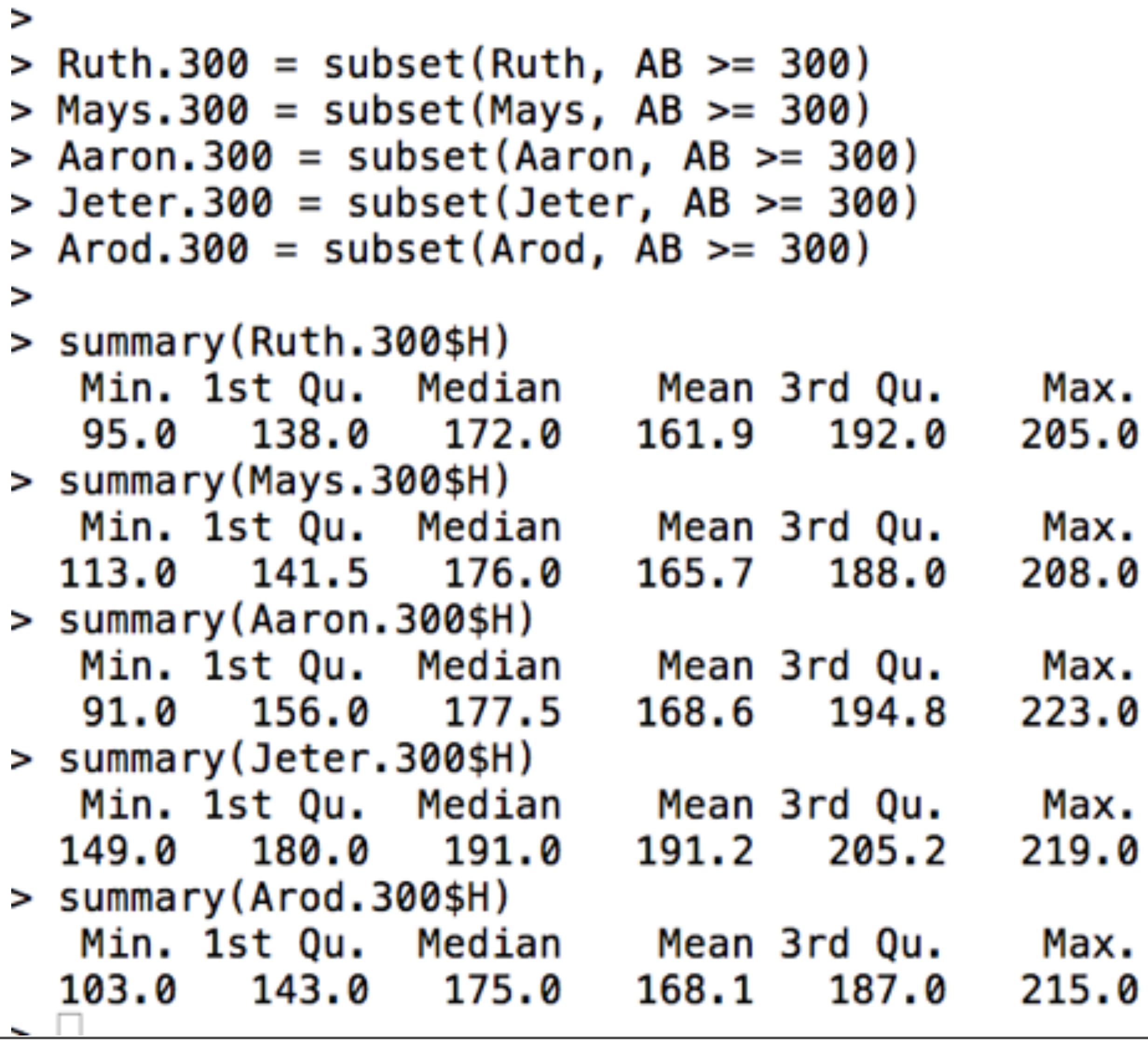
```
The database is comprised of the following main tables:

  MASTER - Player names, DOB, and biographical info
  Batting - batting statistics
  Pitching - pitching statistics
  Fielding - fielding statistics

It is supplemented by these tables:

  AllStarFull - All-Star appearances
  HallofFame - Hall of Fame voting data
  Managers - managerial statistics
  Teams - yearly stats and standings
  BattingPost - post-season batting statistics
  PitchingPost - post-season pitching statistics
  TeamFranchises - franchise information
  FieldingOF - outfield position data
  FieldingPost- post-season fieldinf data
  ManagersHalf - split season data for managers
  TeamsHalf - split season data for teams
  Salaries - player salary data
  SeriesPost - post-season series information
  AwardsManagers - awards won by managers
  AwardsPlayers - awards won by players
  AwardsShareManagers - award voting for manager awards
  AwardsSharePlayers - award voting for player awards
  Appearances - details on the positions a player appeared at
  Schools - list of colleges that players attended
  CollegePlaying - list of players and the colleges they attended
```

In this example we load 5 players all time career hits



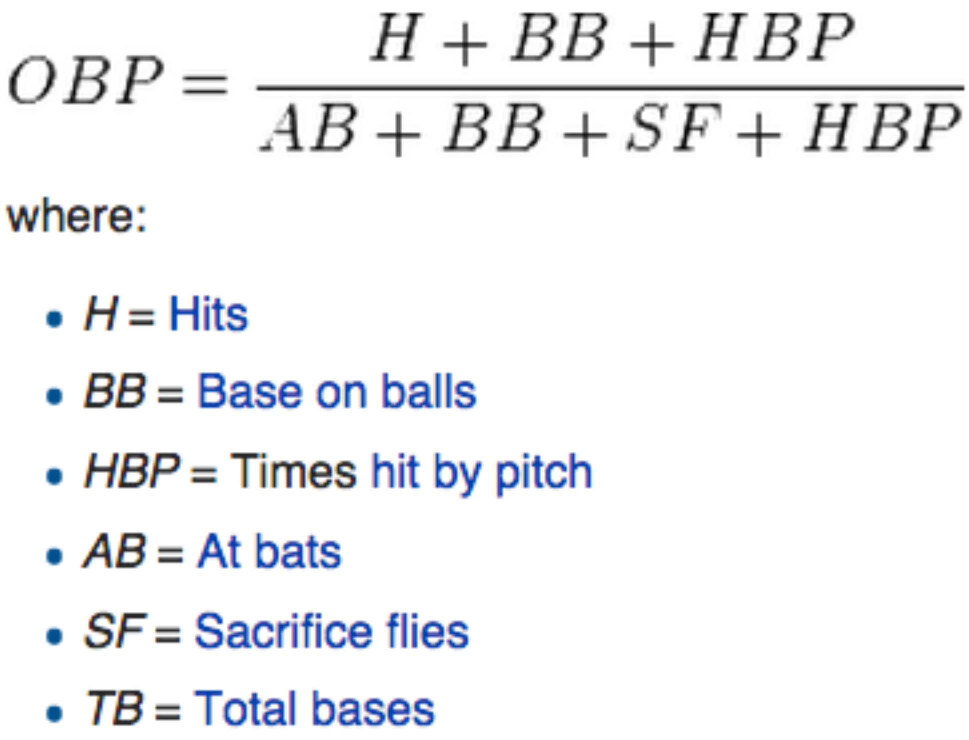
```
>
> Batting = read.csv("Batting.csv")
>
> Ruth = subset(Batting, playerID == "ruthba01")
> Mays = subset(Batting, playerID == "mayswi01")
> Aaron = subset(Batting, playerID == "aaronha01")
> Jeter = subset(Batting, playerID == "jeterde01")
> Arod = subset(Batting, playerID == "rodrial01")
>
> summary(Ruth$H)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   97.25  147.50  130.60  185.50  205.00
> summary(Mays$H)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    9.0   120.5   171.0   142.7   186.0   208.0
> summary(Aaron$H)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   62.0   142.5   174.0   164.0   193.5   223.0
> summary(Jeter$H)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   12.0   174.8   190.5   173.2   203.8   219.0
> summary(Arod$H)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   11.0   126.8   169.0   147.0   184.0   215.0
```

In this example we only take years where they had at least



```
>
> Ruth.300 = subset(Ruth, AB >= 300)
> Mays.300 = subset(Mays, AB >= 300)
> Aaron.300 = subset(Aaron, AB >= 300)
> Jeter.300 = subset(Jeter, AB >= 300)
> Arod.300 = subset(Arod, AB >= 300)
>
> summary(Ruth.300$H)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   95.0   138.0   172.0   161.9   192.0   205.0 
> summary(Mays.300$H)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  113.0   141.5   176.0   165.7   188.0   208.0 
> summary(Aaron.300$H)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   91.0   156.0   177.5   168.6   194.8   223.0 
> summary(Jeter.300$H)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  149.0   180.0   191.0   191.2   205.2   219.0 
> summary(Arod.300$H)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  103.0   143.0   175.0   168.1   187.0   215.0 
> 
```

On Base Percentage

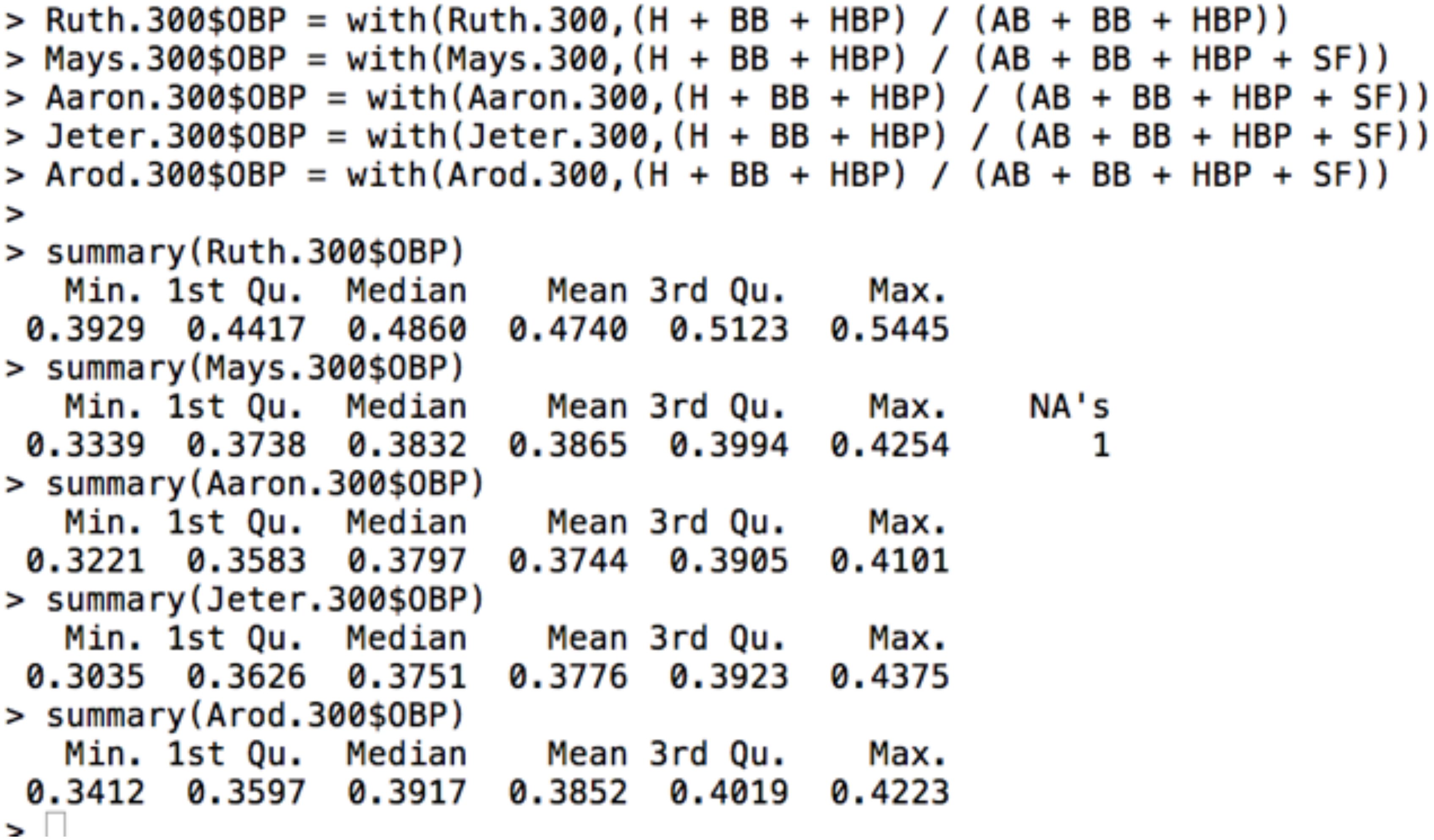


$$OBP = \frac{H + BB + HBP}{AB + BB + SF + HBP}$$

where:

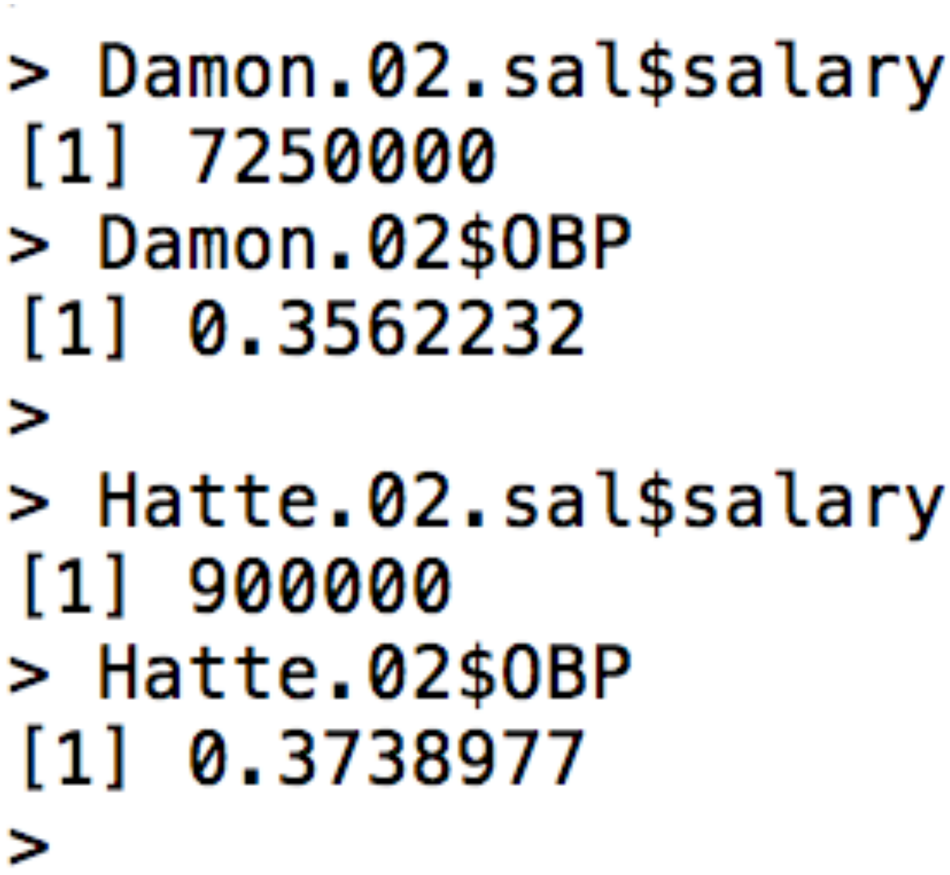
- H = Hits
- BB = Base on balls
- HBP = Times hit by pitch
- AB = At bats
- SF = Sacrifice flies
- TB = Total bases

We took 5 great players and looked at their career OBP’s



```
> Ruth.300$OBP = with(Ruth.300,(H + BB + HBP) / (AB + BB + HBP))
> Mays.300$OBP = with(Mays.300,(H + BB + HBP) / (AB + BB + HBP + SF))
> Aaron.300$OBP = with(Aaron.300,(H + BB + HBP) / (AB + BB + HBP + SF))
> Jeter.300$OBP = with(Jeter.300,(H + BB + HBP) / (AB + BB + HBP + SF))
> Arod.300$OBP = with(Arod.300,(H + BB + HBP) / (AB + BB + HBP + SF))
>
> summary(Ruth.300$OBP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3929  0.4417  0.4860  0.4740  0.5123  0.5445 
> summary(Mays.300$OBP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 0.3339  0.3738  0.3832  0.3865  0.3994  0.4254       1 
> summary(Aaron.300$OBP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3221  0.3583  0.3797  0.3744  0.3905  0.4101 
> summary(Jeter.300$OBP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3035  0.3626  0.3751  0.3776  0.3923  0.4375 
> summary(Arod.300$OBP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3412  0.3597  0.3917  0.3852  0.4019  0.4223 
> 
```

Moneyball in R



```
> Damon.02.sal$salary
[1] 7250000
> Damon.02$OBP
[1] 0.3562232
>
> Hatte.02.sal$salary
[1] 900000
> Hatte.02$OBP
[1] 0.3738977
>
```

New Stuff

- More work with Baseball and R
- Fantasy Football Analytics
- Stocks and R