# Introduction to ddply

## Cleaning and Summarizing Data

Iowa State University

# Outline

- conditionals & subsets
- `for` loops
- avoiding `for` loops with `ddply`

# Baseball Data

- The `plyr` package contains the data set `baseball`
- seasonal batting statistics of all major league players (through 2007)

```
library(plyr)
help(baseball)
head(baseball)
```

```
           id year stint team lg  g  ab  r  h X2b X3b hr rbi sb cs bb so ibb hbp sh sf gidp
4   ansonca01 1871     1  RC1    25 120 29 39  11   3  0  16  6  2  2  1  NA  NA NA NA   NA
44  forceda01 1871     1  WS3    32 162 45 45   9   4  0  29  8  0  4  0  NA  NA NA NA   NA
68  mathebo01 1871     1  FW1    19  89 15 24   3   1  0  10  2  1  2  0  NA  NA NA NA   NA
99  startjo01 1871     1  NY2    33 161 35 58   5   1  1  34  4  2  3  0  NA  NA NA NA   NA
102 suttoez01 1871     1  CL1    29 128 35 45   3   7  3  23  3  1  1  0  NA  NA NA NA   NA
106 whitede01 1871     1  CL1    29 146 40 47   6   5  1  21  2  2  4  1  NA  NA NA NA   NA
```

# Baseball Data

- We would like to create career summary statistics for each player
- Plan: subset on a player, and compute statistics

```r
ss <- subset(baseball, id=="sosasa01")
head(ss)
```

```
            id year stint team lg   g  ab  r   h X2b X3b hr rbi sb cs bb  so ibb hbp sh sf gidp
66822 sosasa01 1989     1  TEX AL  25  84  8  20   3   0  1   3  0  2  0  20   0   0  4  0    3
66823 sosasa01 1989     2  CHA AL  33  99 19  27   5   0  3  10  7  3 11  27   2   2  1  2    3
67907 sosasa01 1990     1  CHA AL 153 532 72 124  26  10 15  70 32 16 33 150   4   6  2  6   10
69018 sosasa01 1991     1  CHA AL 116 316 39  64  10   1 10  33 13  6 14  98   2   2  5  1    5
70599 sosasa01 1992     1  CHN NL  67 262 41  68   7   2  8  25 15  7 19  63   1   4  4  2    4
71757 sosasa01 1993     1  CHN NL 159 598 92 156  25   5 33  93 36 11 38 135   6   4  0  1   14
```

```r
mean(ss$h/ss$ab)
## [1] 0.2681506
```

# Baseball Data

- We would like to create career summary statistics for each player
- Plan: subset on a player, and compute statistics

```
ss <- subset(baseball, id=="sosasa01")
head(ss)
```

```
            id year stint team lg   g  ab  r   h X2b X3b hr rbi sb cs bb  so ibb hbp sh sf gidp
66822 sosasa01 1989     1  TEX AL  25  84  8  20   3   0  1   3  0  2  0  20   0   0  4  0    3
66823 sosasa01 1989     2  CHA AL  33  99 19  27   5   0  3  10  7  3 11  27   2   2  1  2    3
67907 sosasa01 1990     1  CHA AL 153 532 72 124  26  10 15  70 32 16 33 150   4   6  2  6   10
69018 sosasa01 1991     1  CHA AL 116 316 39  64  10   1 10  33 13  6 14  98   2   2  5  1    5
70599 sosasa01 1992     1  CHN NL  67 262 41  68   7   2  8  25 15  7 19  63   1   4  4  2    4
71757 sosasa01 1993     1  CHN NL 159 598 92 156  25   5 33  93 36 11 38 135   6   4  0  1   14
```

```
mean(ss$h/ss$ab)
## [1] 0.2681506
```

We need an automatic way to calculate this

# for loops

- Idea: repeat the same (set of) statement(s) for each element of an index set
- Setup:
  - Introduce counter variable (sometimes named i)
  - Reserve space for results
- Generic Code:

```
result <- rep(NA, length(indexset))
for(i in indexset){
  ... some statments ...
  result[i] <- ...
}
```

# for loops for Baseball

- Index set: player id
- Setup:

```
# Index set
players <- unique(baseball$id)
n <- length(players)

# Place to store data
ba <- rep(NA, n)

# Loop
for(i in 1:n){
  career <- subset(baseball, id==players[i])
  ba[i] <- with(career, mean(h/ab, na.rm=T))
}

# Results
summary(ba)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.1831  0.2459  0.2231  0.2699  0.5000       6
```

# for loops for Baseball

- Index set: player id
- i=0

```
# Index set
players <- unique(baseball$id)
n <- length(players)

# Place to store data
ba <- rep(NA, n)

head(ba)
## [1] NA NA NA NA NA NA
```

# for loops for Baseball

- Index set: player id
- i=1

```
# Index set
players <- unique(baseball$id)
n <- length(players)

# Place to store data
ba <- rep(NA, n)

# Loop
for(i in 1:1){
  career <- subset(baseball, id==players[i])
  ba[i] <- with(career, mean(h/ab, na.rm=T))
}
i
## [1] 1

head(ba)
## [1] 0.3371163        NA        NA        NA        NA        NA
```

# for loops for Baseball

- Index set: player id
- i=2

```
# Index set
players <- unique(baseball$id)
n <- length(players)

# Place to store data
ba <- rep(NA, n)

# Loop
for(i in 1:2){
  career <- subset(baseball, id==players[i])
  ba[i] <- with(career, mean(h/ab, na.rm=T))
}
i
## [1] 2

head(ba)
## [1] 0.3371163 0.2489226        NA        NA        NA        NA
```

# Your Turn

- MLB rules for the greatest all-time hitters are that players have to have played at least 1000 games with at least as many at-bats in order to be considered

- Extend the for loop above to collect the additional information Introduce and collect data for two new variables: `games` and `atbats`

# How did it go? What was difficult?

- household chores (declaring variables, setting values each time) distract from real work
- indices are error-prone
- loops often result in slow code because R can compute quantities using entire vectors in an optimized way

# Summarise

- A special function: `summarise` or `summarize`

```
library(dplyr)
baseball <- read.csv("../data/baseball.csv")
summarise(baseball, ab=mean(h/ab, na.rm=T))
##          ab
## 1 0.2339838

summarise(baseball,
          ba = mean(h/ab, na.rm=T),
          games = sum(g, na.rm=T),
          hr = sum(hr, na.rm=T),
          ab = sum(ab, na.rm=T))
##          ba   games     hr      ab
## 1 0.2339838 1580070 113577 4891061

summarise(subset(baseball, id=="sosasa01"),
          ba = mean(h/ab, na.rm=T),
          games = sum(g, na.rm=T),
          hr = sum(hr, na.rm=T),
          ab = sum(ab, na.rm=T))
##          ba games  hr   ab
## 1 0.2681506  2354 609 8813
```

# dplyr + Summarize

A powerful combination to create summary statistics

```
careers <- summarise(group_by(baseball, id),
                     ba = mean(h/ab, na.rm=T),
                     games = sum(g, na.rm=T),
                     homeruns = sum(hr, na.rm=T),
                     atbats = sum(ab, na.rm=T))
```

```
head(careers)
## Source: local data frame [6 x 5]
##
##          id        ba games homeruns atbats
## 1 aaronha01 0.3010752  3298      755  12364
## 2 abernte02 0.1824394   681        0    181
## 3 adairje01 0.2363071  1165       57   4019
## 4 adamsba01 0.2096513   482        3   1019
## 5 adamsbo03 0.2378073  1281       37   4019
## 6 adcocjo01 0.2751690  1959      336   6606
```

# Your Turn

- Find some summary statistics for each of the teams (variable `team`)
  - How many different (`unique`) players has the team had?
  - What was the team's first/last season?
- Challenge:
  Find the number of players on each team over time. Does the number change?