Advanced Data Wrangling Using dplyr

Yogindra Raghav

October 18, 2018

```
library(Lahman)

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

1. Number of rows in data frame:

```
manny = filter(Batting, playerID == "ramirma02")

## Warning: package 'bindrcpp' was built under R version 3.4.4

nrow(manny)

## [1] 21
```

2. Manny Ramirez records grouped by team.

```
manny %>% group_by(teamID) %>% summarize(span = paste(min(yearID), max(yearID
), sep = "-"), numYears = n_distinct(yearID), numTeams = n_distinct(teamID),
BA = sum(H)/sum(AB), tH = sum(H), tHR = sum(HR), tRBI = sum(RBI))

## # A tibble: 5 x 8
##   teamID span      numYears numTeams     BA    tH   tHR  tRBI
##   <fct>  <chr>        <int>    <int>  <dbl> <int> <int> <int>
## 1 BOS    2001-2008        8        1 0.312   1232   274   868
## 2 CHA    2010-2010        1        1 0.261     18     1     2
## 3 CLE    1993-2000        8        1 0.313   1086   236   804
## 4 LAN    2008-2010        3        1 0.322    237    44   156
## 5 TBA    2011-2011        1        1 0.0588     1     0     1
```

3. Number of rows occuring if inner_join() is used in place of left_join().

```
mannyBySeason <- Batting %>%
  filter(playerID == "ramirma02") %>%
  inner_join(Master, by = c("playerID"  = "playerID")) %>%
  group_by(yearID) %>%
  summarize(
    Age = max(yearID - birthYear), numTeams = n_distinct(teamID),
    BA = sum(H)/sum(AB), tH = sum(H), tHR = sum(HR), tRBI = sum(RBI),
    OBP = sum(H + BB + HBP) / sum(AB + BB + SF + HBP),
    SLG = sum(H + X2B + 2*X3B + 3*HR) / sum(AB)
  ) %>%
  mutate(OPS   = OBP + SLG) %>%
  arrange(desc(OPS))

mannyAllStar = AllstarFull %>% filter(playerID == "ramirma02")

mannyBySeason %>% inner_join(mannyAllStar, by = c("yearID" = "yearID")) %>% s
elect(yearID, Age, OPS, GP, startingPos) %>% nrow()

## [1] 12
```

4. Barry Bonds has record for most home runs and Manny Rodriguez is in the top 20.

```
Batting %>% group_by(playerID) %>% summarize(HR = sum(HR)) %>% arrange(desc(H
R)) %>% head(20) %>% inner_join(Master, by = c("playerID" ="playerID")) %>% s
elect(nameFirst, nameLast, HR)

## # A tibble: 20 x 3
##    nameFirst nameLast     HR
##    <chr>     <chr>     <int>
##  1 Barry     Bonds       762
##  2 Hank      Aaron       755
##  3 Babe      Ruth        714
##  4 Alex      Rodriguez   696
##  5 Willie    Mays        660
##  6 Ken       Griffey     630
##  7 Jim       Thome       612
##  8 Sammy     Sosa        609
##  9 Albert    Pujols      591
## 10 Frank     Robinson    586
## 11 Mark      McGwire     583
## 12 Harmon    Killebrew   573
## 13 Rafael    Palmeiro    569
## 14 Reggie    Jackson     563
## 15 Manny     Ramirez     555
## 16 Mike      Schmidt     548
## 17 David     Ortiz       541
## 18 Mickey    Mantle      536
## 19 Jimmie    Foxx        534
## 20 Willie    McCovey     521
```

5. Every pitcher that accumulated 300 or more wins and 3,000 or more strikeouts.

```
Pitching %>% group_by(playerID) %>% summarize(W = sum(W), SO = sum(SO)) %>% f
ilter(W>=300 & SO >=3000) %>% inner_join(Master, by = "playerID") %>% select(
nameFirst, nameLast, W, SO)
```

```
## # A tibble: 10 x 4
##    nameFirst nameLast     W    SO
##    <chr>     <chr>    <int> <int>
##  1 Steve     Carlton    329  4136
##  2 Roger     Clemens    354  4672
##  3 Randy     Johnson    303  4875
##  4 Walter    Johnson    417  3509
##  5 Greg      Maddux     355  3371
##  6 Phil      Niekro     318  3342
##  7 Gaylord   Perry      314  3534
##  8 Nolan     Ryan       324  5714
##  9 Tom       Seaver     311  3640
## 10 Don       Sutton     324  3574
```

6. Most recent World Series MVP winners.

```
AwardsPlayers %>% filter(awardID == "World Series MVP") %>% arrange(desc(year
ID)) %>% head(10) %>% inner_join(Master, by ="playerID") %>% mutate(age = yea
rID - birthYear) %>% select(nameFirst, nameLast, age, awardID)
```

```
##    nameFirst  nameLast age          awardID
## 1        Ben   Zobrist  35 World Series MVP
## 2   Salvador     Perez  25 World Series MVP
## 3    Madison Bumgarner  25 World Series MVP
## 4      David     Ortiz  38 World Series MVP
## 5      Pablo  Sandoval  26 World Series MVP
## 6      David    Freese  28 World Series MVP
## 7      Edgar  Renteria  34 World Series MVP
## 8     Hideki    Matsui  35 World Series MVP
## 9       Cole    Hamels  25 World Series MVP
## 10      Mike    Lowell  33 World Series MVP
```