Lesson 4

Daniel Baller

1/13/2022

## Last Class

Key concepts

* five tidyverse verbs
* relational data bases

Review Questions

* Using the Teams data frame, calculate average wins (W) per season for each MLB team from 2000-present. Order your results from the best teams to the worst.
* Using the Teams data frame, calculate how many World Series wins (WSWin) each team had during this period and add this information to your table in the previous question.

Win percentage and World Series titles by team (2000-present).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| teamID | n | W | G | Wpct |
| ANA | 1 | 425 | 810 | 0.525 |
| ARI | 1 | 1623 | 3300 | 0.492 |
| BOS | 4 | 1816 | 3299 | 0.550 |
| CHA | 1 | 1635 | 3300 | 0.495 |
| CHN | 1 | 1658 | 3299 | 0.503 |
| FLO | 1 | 963 | 1942 | 0.496 |
| HOU | 1 | 1650 | 3300 | 0.500 |
| KCA | 1 | 1456 | 3300 | 0.441 |
| LAN | 1 | 1824 | 3300 | 0.553 |
| NYA | 2 | 1919 | 3298 | 0.582 |
| PHI | 1 | 1665 | 3299 | 0.505 |
| SFN | 3 | 1705 | 3298 | 0.517 |
| SLN | 2 | 1842 | 3297 | 0.559 |
| WAS | 1 | 1248 | 2488 | 0.502 |
| ATL | NA | 1770 | 3298 | 0.537 |
| OAK | NA | 1765 | 3298 | 0.535 |
| LAA | NA | 1323 | 2490 | 0.531 |
| CLE | NA | 1706 | 3298 | 0.517 |
| MIN | NA | 1664 | 3301 | 0.504 |
| TEX | NA | 1641 | 3301 | 0.497 |
| NYN | NA | 1634 | 3298 | 0.495 |
| TOR | NA | 1631 | 3299 | 0.494 |
| SEA | NA | 1622 | 3300 | 0.492 |
| TBA | NA | 1594 | 3298 | 0.483 |
| MIL | NA | 1594 | 3301 | 0.483 |
| CIN | NA | 1557 | 3302 | 0.472 |
| COL | NA | 1547 | 3302 | 0.469 |
| SDN | NA | 1545 | 3301 | 0.468 |
| DET | NA | 1534 | 3295 | 0.466 |
| MON | NA | 368 | 810 | 0.454 |
| PIT | NA | 1492 | 3296 | 0.453 |
| BAL | NA | 1478 | 3300 | 0.448 |
| MIA | NA | 586 | 1354 | 0.433 |

## Intro to R (Day 2)

### Split, apply, and combine data (pg 58)

It’s important in data science to be able to write procedures and apply them over a data set. The Marchi text refers to this as “splitting, applying, and combining data”. In other words, we split a data frame into pieces, apply a procedure or function to each piece, and then combine the results into a new data frame.

Ways to “split, apply, and combine”…

* group\_by() and summarize() functions
* split() and map\_df() functions

In general, these coding tasks involve three steps:

1. write a function to perform the task on one split of the data
2. split the data on a variable
3. apply the function to each split of the data

### Example

Two steps:

1. write a function that takes a data frame and returns the team with the most wins and whether they won the World Series
2. split the data frame on yearID
3. apply the function to each split

#step 1 - write a function  
  
#first write what you want the function to do on a small piece of the data  
Teams %>%   
 filter(yearID==2000) %>% #Select only entries for the year 2000  
 arrange(-W) %>% #arrange in decending order by number of wins  
 select(teamID, W, WSWin) %>% #only select these three columns  
 head(1) #pick the top one by number of wins

## teamID W WSWin  
## 1 SFN 97 N

#next generalize function to work with whatever data you provide  
mostWins <- function(data){  
 data %>%   
 arrange(-W) %>%   
 select(teamID, W, WSWin) %>%   
 head(1)  
}  
  
#step 2 and 3 - split and apply  
winLeaders <- Teams %>%   
 filter(yearID >= 2000) %>% #filter for entries in the year 2000 and later  
 split(pull(.,yearID)) %>% #turn the dataframe into a list where each entry is split by the yearID  
 map\_dfr(mostWins, .id = "yearID") #apply the function each entry of the list and provide the results as a dataframe  
  
winLeaders %>%   
 kable(caption = "Major League win leaders by season and whether they won the World Series.")

Major League win leaders by season and whether they won the World Series.

|  |  |  |  |
| --- | --- | --- | --- |
| yearID | teamID | W | WSWin |
| 2000 | SFN | 97 | N |
| 2001 | SEA | 116 | N |
| 2002 | NYA | 103 | N |
| 2003 | ATL | 101 | N |
| 2004 | SLN | 105 | N |
| 2005 | SLN | 100 | N |
| 2006 | NYA | 97 | N |
| 2007 | BOS | 96 | Y |
| 2008 | LAA | 100 | N |
| 2009 | NYA | 103 | Y |
| 2010 | PHI | 97 | N |
| 2011 | PHI | 102 | N |
| 2012 | WAS | 98 | N |
| 2013 | BOS | 97 | Y |
| 2014 | LAA | 98 | N |
| 2015 | SLN | 100 | N |
| 2016 | CHN | 103 | Y |
| 2017 | LAN | 104 | N |
| 2018 | BOS | 108 | Y |
| 2019 | HOU | 107 | N |
| 2020 | LAN | 43 | Y |

### Your turn