Data Scraping Lab

Liam Niehus-Staab, Ishaan Tibrewal, Thomas Leistikow

11/12/2018

## Clean Data

baseball = read\_html("https://en.wikipedia.org/wiki/Triple\_Crown\_(baseball)")  
  
#pull out table elements from wiki page  
ball = html\_nodes(baseball, "table")  
  
#choose pitching table  
pitching = html\_table(ball, header = TRUE, fill = TRUE)[[3]]  
  
#rename columns  
names(pitching)[5:7] = c("EarnedRunAverage", "Wins", "StrikeOuts")  
  
#get rid of asterix in the 3 cols above  
pitching$EarnedRunAverage = as.numeric(str\_replace\_all(pitching$EarnedRunAverage,  
 "\\\*", ""))  
pitching$Wins = as.numeric(str\_replace\_all(pitching$Wins,  
 "\\\*", ""))  
pitching$StrikeOuts = as.numeric(str\_replace\_all(pitching$StrikeOuts,  
 "\\\*", ""))  
#remove extra useless refs col  
pitching$`Ref(s)` <- NULL  
  
#get batting data  
batting = html\_table(ball, header = TRUE, fill = TRUE)[[2]]  
  
#delete useless col  
batting$`Ref(s)` <- NULL  
  
#rename columns  
names(batting)[6:8] = c("HomeRuns", "RunsBattedIn", "BattingAvg")  
  
#get rid of asterix in the 3 cols above  
batting$HomeRuns = as.numeric(str\_replace\_all(batting$HomeRuns,  
 "\\\*", ""))  
batting$RunsBattedIn = as.numeric(str\_replace\_all(batting$RunsBattedIn,  
 "\\\*", ""))  
batting$BattingAvg = as.numeric(str\_replace\_all(batting$BattingAvg,  
 "\\\*", ""))

## Show Cleaned Data

#show tables  
head(batting)

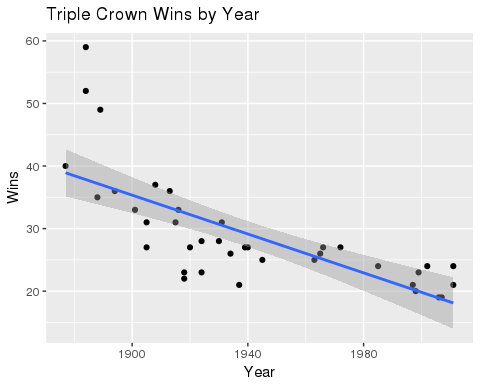
## Year Player Position Team League  
## 1 1878 Paul Hines Center fielder Providence Grays NL  
## 2 1887 Tip O'Neill Left fielder St. Louis Browns AA  
## 3 1894 Hugh Duffy Outfielder Boston Beaneaters NL  
## 4 1901 Nap Lajoie Second baseman Philadelphia Athletics AL  
## 5 1909 Ty Cobb Right fielder Detroit Tigers AL  
## 6 1922 Rogers Hornsby Second baseman St. Louis Cardinals NL  
## HomeRuns RunsBattedIn BattingAvg  
## 1 4 50 0.358  
## 2 14 123 0.435  
## 3 18 145 0.440  
## 4 14 125 0.426  
## 5 9 107 0.377  
## 6 42 152 0.401

head(pitching)

## Year Player Team League EarnedRunAverage Wins  
## 1 1877 Tommy Bond Boston Red Caps NL 2.11 40  
## 2 1884 Guy Hecker Louisville Colonels AA 1.80 52  
## 3 1884 Charles Radbourn Providence Grays NL 1.38 59  
## 4 1888 Tim Keefe New York Giants NL 1.74 35  
## 5 1889 John Clarkson Boston Beaneaters NL 2.73 49  
## 6 1894 Amos Rusie New York Giants NL 2.78 36  
## StrikeOuts  
## 1 170  
## 2 385  
## 3 441  
## 4 335  
## 5 284  
## 6 195

## Plot Data

#make plot  
ggplot(data = pitching, aes(x = Year, y = Wins)) +   
 geom\_point() + stat\_smooth(method = lm) + labs(title = "Triple Crown Wins by Year")

 It can be seen that as time has gon one, winners of the pitching tripple crown in baseball have had a clearly decreasing ratio of wins. Because having the triple crown means that each player lead the league in wins, run average, and strike outs, that means that ever single pitcher in the whole league has been progressively getting fewer wins to their name as the years go on. This could say something about teams being more balanced as time goes on, or it may have something to do with there being more pitchers in each team (causing the wins to be spread out over a greater number of players) over time.