Lesson 2 - Intro to R tidyverse

Kevin Cummiskey

1/9/2020

Research Question: *How (and why) has the rate of walks changed over the history of baseball?*

Which Datafile from the Lahman package would you use to answer this question?

```
library(tidyverse)
library(Lahman)

Teams %>% head(2)
```

```
##   yearID lgID teamID franchID divID Rank  G Ghome  W  L DivWin WCWin LgWin
## 1   1871   NA    BS1      BNA  <NA>    3 31    NA 20 10   <NA>  <NA>     N
## 2   1871   NA    CH1      CNA  <NA>    2 28    NA 19  9   <NA>  <NA>     N
##   WSWin   R   AB   H X2B X3B HR BB SO SB CS HBP SF  RA  ER  ERA CG SHO SV
## 1  <NA> 401 1372 426  70  37  3 60 19 73 16  NA NA 303 109 3.55 22   1  3
## 2  <NA> 302 1196 323  52  21 10 60 22 69 21  NA NA 241  77 2.76 25   0  1
##   IPouts  HA HRA BBA SOA   E DP    FP                    name
## 1    828 367   2  42  23 243 24 0.834    Boston Red Stockings
## 2    753 308   6  28  22 229 16 0.829 Chicago White Stockings
##                      park attendance BPF PPF teamIDBR teamIDlahman45
## 1     South End Grounds I         NA 103  98      BOS            BS1
## 2 Union Base-Ball Grounds         NA 104 102      CHI            CH1
##   teamIDretro
## 1         BS1
## 2         CH1
```

On page 35 of the text, there is a list of the five main dplyr verbs (and group_by). Below, briefly describe what each does.

- `select`
- `filter`
- `arrange`
- `mutate`
- `summarize`
- `group_by`

Discuss how you could use these verbs and the Teams database to calculate the walk rate (per nine innings). Write your code below.

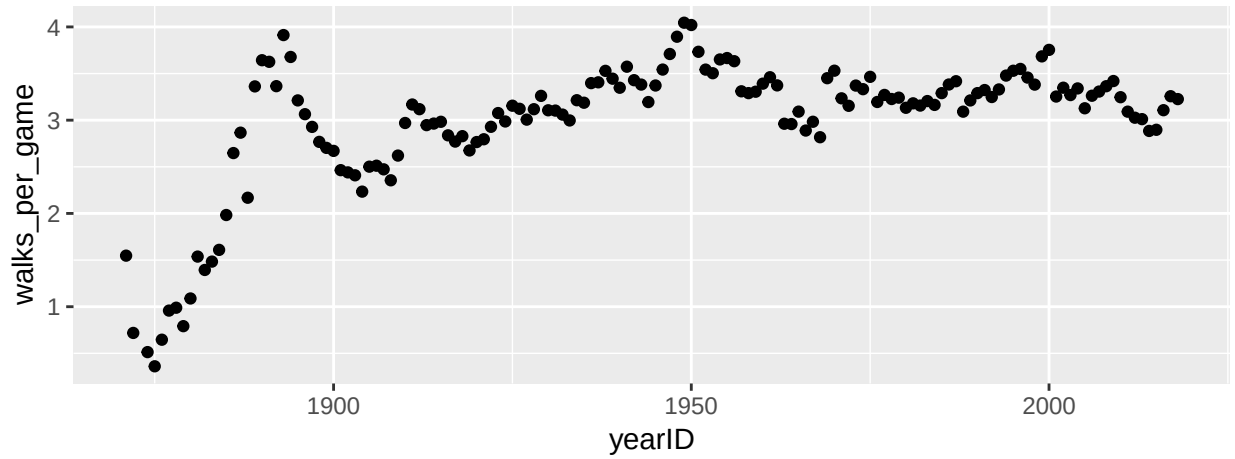
```
#Calculate the walk rate per 9 innings by season
Teams %>%
  group_by(yearID) %>%
  summarize(walks_per_game = sum(BB)/sum(G),
            SO_per_game = sum(SO)/sum(G)) -> walks

walks %>% head(3)
```

```
## # A tibble: 3 x 3
##   yearID walks_per_game SO_per_game
##    <int>          <dbl>       <dbl>
## 1   1871          1.55        0.689
## 2   1872          0.719       0.724
## 3   1873         NA           0.698
```

Let's look at a plot of the results.

```
walks %>%
  ggplot(aes(x = yearID, y = walks_per_game)) +
  geom_point()
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



What trends do you observe in walks?

The previous plot tells us nothing about `why` we observe changes. Briefly explain two other analyses we might conduct to better understand trends in walks. Identify the appropriate Datafiles for these analyses.

Let's see if the ratio of walks to strikeouts tells us anything interesting. Without looking on the next page, write the code you would use to calculate the ratio of walks to strikeouts by season.

```r
walks = walks %>%
  mutate(walks_SO_ratio = walks_per_game/SO_per_game)
```

```r
walks %>%
  ggplot(aes(x = yearID, y = walks_SO_ratio)) +
  geom_point()
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```