The Grammar of Graphics and ggplot2

2023-06-07

Anscombe's Quartet: demonstrates that simple summary statistics do not cut it. We need visualizations to better understand the distributions and corroborate our inferences. Data Viz: Florence Nightingale

```
Batting <- as_tibble(Batting)

year_batting_summary <- Batting %>%
  filter(lgID %in% c("AL", "NL")) %>%
  group_by(yearID) %>%
  summarize(across(c("H", "HR", "SO", "BB", "AB"), \(x) sum(x, na.rm = TRUE))) %>%
  mutate(batting_avg = H/AB)

year_batting_summary
```

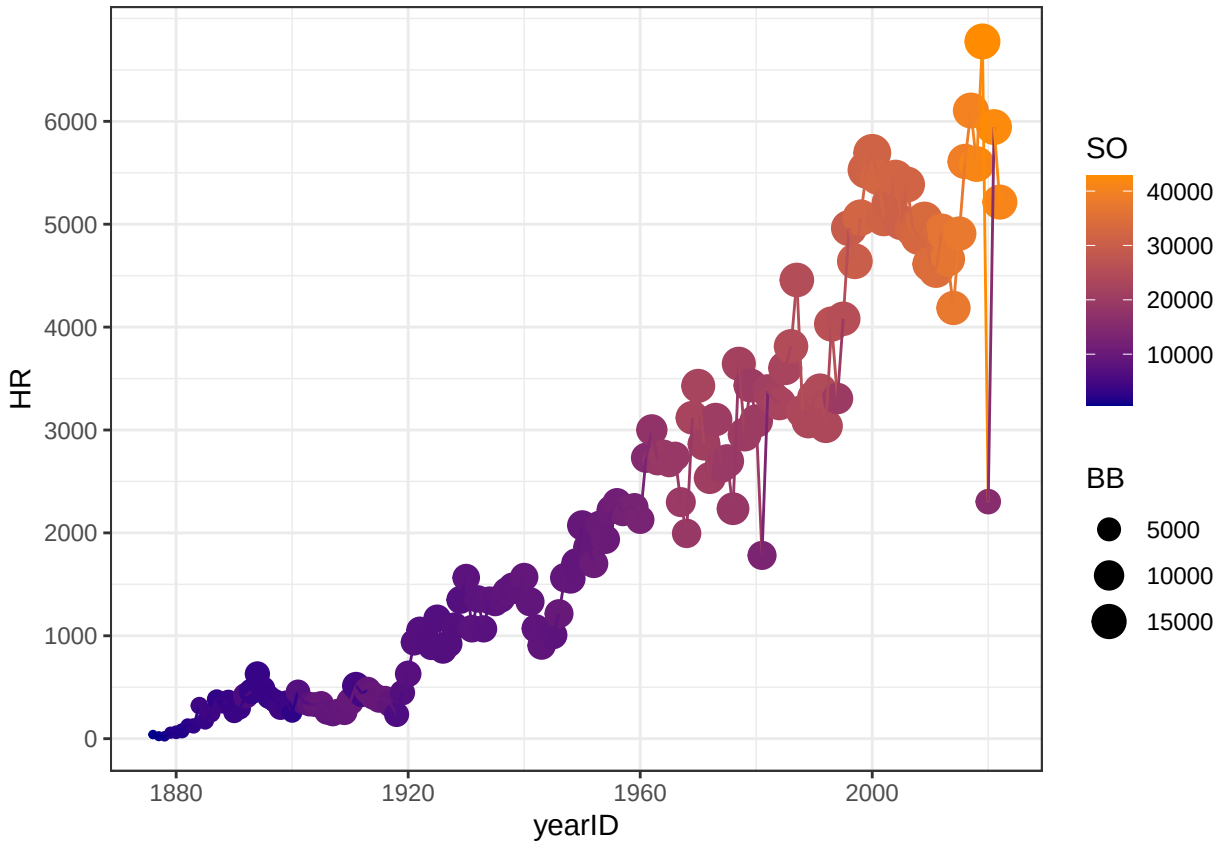
```
## # A tibble: 147 x 7
##    yearID     H    HR    SO    BB    AB batting_avg
##     <int> <int> <int> <int> <int> <int>       <dbl>
##  1   1876  5338    40   589   336 20121       0.265
##  2   1877  3705    24   726   345 13667       0.271
##  3   1878  3539    23  1081   364 13644       0.259
##  4   1879  6171    58  1843   508 24155       0.255
##  5   1880  5946    62  1993   740 24301       0.245
##  6   1881  6339    76  1784  1033 24377       0.260
##  7   1882  6225   126  2159   960 24769       0.251
##  8   1883  7611   124  2877  1121 29012       0.262
##  9   1884  8071   321  4335  1821 32687       0.247
## 10   1885  7516   174  3337  1845 31123       0.241
## # i 137 more rows
```

Hadley Wickham PhD thesis ggplot2

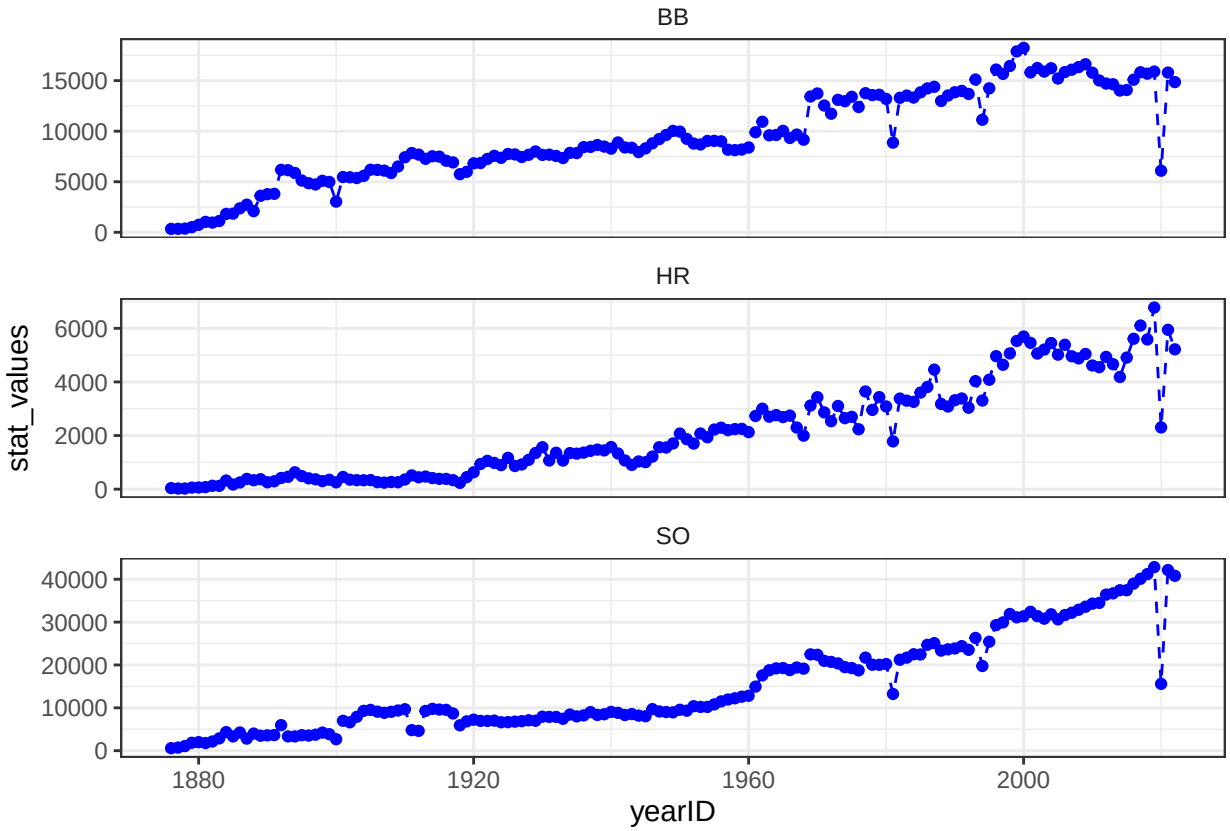
Grammar of Graphics:

1. data
2. geom
3. aes
4. scale
5. facet
6. stat
7. coord
8. labs
9. theme

```
year_batting_summary %>%
  ggplot(aes(x = yearID, y = HR, color = SO))+
  geom_point(aes(size = BB))+
  geom_line()+
  scale_y_continuous(breaks = seq(0,6000, by = 1000))+
  scale_color_gradient(low = "darkblue", high = "darkorange")+
  theme_bw()+
  labs(xlab = "Homeruns", "")
```




```
year_batting_summary %>%
  select(yearID, HR, SO, BB) %>%
  pivot_longer(HR:BB, names_to = "stat", values_to = "stat_values") %>%
  ggplot(aes(x = yearID, y = stat_values))+
  geom_point(color = "blue")+
  geom_line(color = "blue", linetype = "dashed")+
  facet_wrap(~stat, scales = "free_y", nrow = 3)+
  theme_bw()+
  theme(strip.background = element_blank())+
  labs(xlab = "Year")
```


BB
HR
SO
stat_values
15000
10000
5000
0
6000
4000
2000
0
40000
30000
20000
10000
0
1880
1920
1960
2000
yearID