Baseball Eras

Authors: Samantha Attar, Hannah Dineen, Andy Fullerton, Nora Hanson, Cam Kelso, Katie McLaughlin, and Caitlyn Nolan

Introduction:

The following analysis displays batting performance data from baseball's historical eras to challenge whether or not there are statistical connections between the two. The eras we looked at are the Dead Ball era (1901-1920), the Live ball era (1921-1942), the Integration era (1943-1961), The Expansion era (1962-1977), the Free Agent era (1978-1994), the Steroid era (1995-2004), and the Contemporary era (2005-2014). The three specific batting performance statistics compared between time periods were home runs, runs batted in, and on-base percentage. Our group hypothesized that baseball eras would not tie directly to baseball statistics.

Methods:

We relied on the Batting data frame to make comparisons between baseball's historical eras and three statistics: home runs, runs batted in, and on-base percentage. In order to identify if there was a connection between baseball's historical eras and batting statistics, we separated data into the seven different baseball eras mentioned in the introduction. We used dplyr to identify the data from each era with a distinct color and then used ggplot to make images displaying the data in order to view each era simultaneously from 1901-2014. By looking at each graphic, we looked to see if there is actually a distinction in each batting statistic between the historical baseball eras.

```
head(Batting)
```

```
##    playerID yearID stint teamID lgID  G  AB  R  H X2B X3B HR RBI SB CS BB
## 1 abercda01   1871     1    TRO   NA  1   4  0  0   0   0  0   0  0  0  0
## 2  addybo01   1871     1    RC1   NA 25 118 30 32   6   0  0  13  8  1  4
## 3 allisar01   1871     1    CL1   NA 29 137 28 40   4   5  0  19  3  1  2
## 4 allisdo01   1871     1    WS3   NA 27 133 28 44  10   2  2  27  1  1  0
## 5 ansonca01   1871     1    RC1   NA 25 120 29 39  11   3  0  16  6  2  2
## 6 armstbo01   1871     1    FW1   NA 12  49  9 11   2   1  0   5  0  1  0
##   SO IBB HBP SH SF GIDP
## 1  0  NA  NA NA NA   NA
## 2  0  NA  NA NA NA   NA
## 3  5  NA  NA NA NA   NA
## 4  2  NA  NA NA NA   NA
## 5  1  NA  NA NA NA   NA
## 6  1  NA  NA NA NA   NA
```

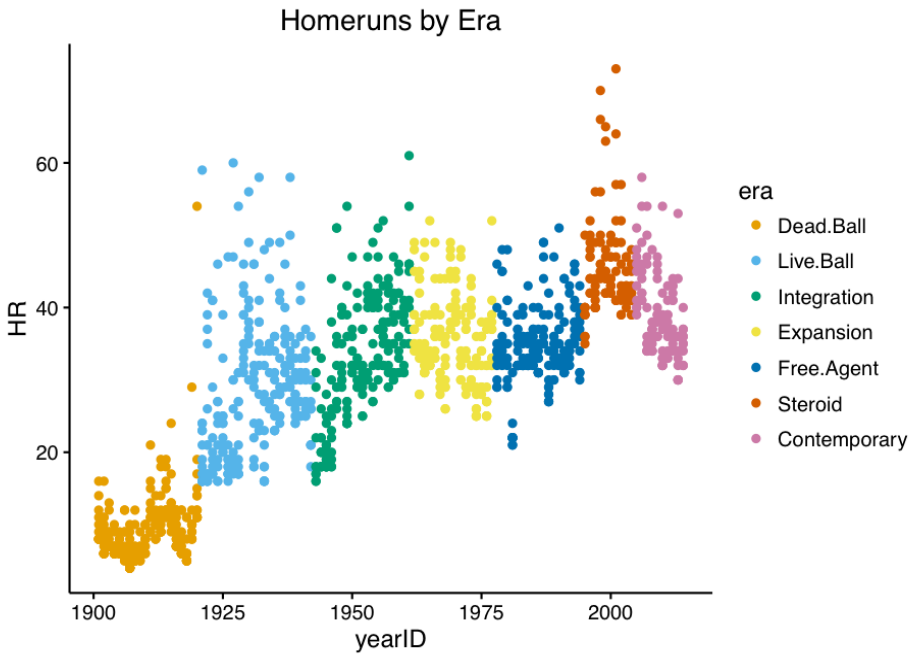
```
batting <- Batting %>% filter(yearID>1900)
batting$era <- cut(batting$yearID,c(1900,1920,1942,1961,1977,1994,2004,2014),
                   labels=c("Dead.Ball","Live.Ball","Integration","Expansion",
                            "Free.Agent","Steroid","Contemporary" ))
levels(batting$era)
```

```
## [1] "Dead.Ball"    "Live.Ball"    "Integration"  "Expansion"
## [5] "Free.Agent"   "Steroid"      "Contemporary"
```

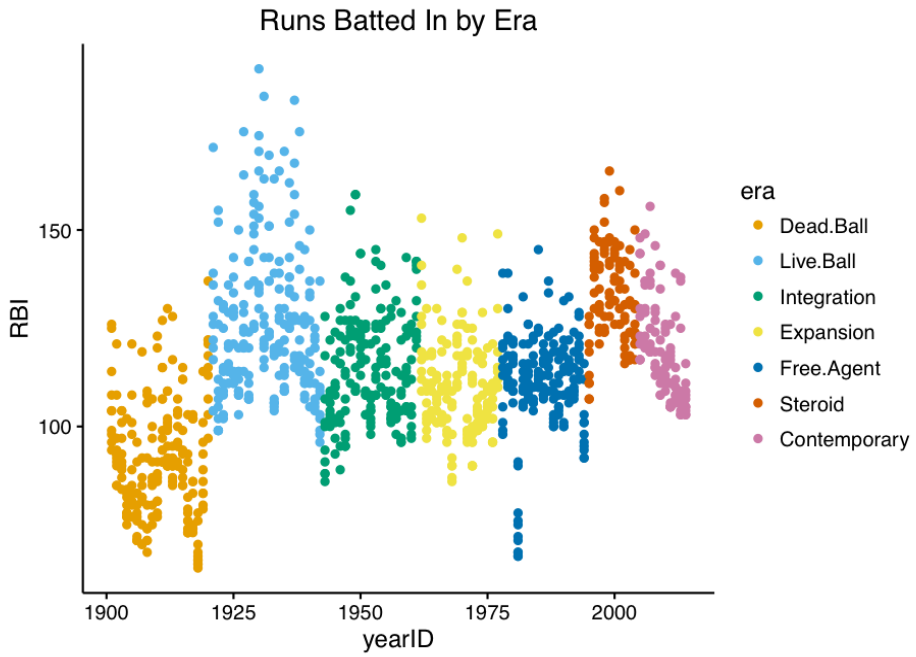
```
head(batting)
```

```
##    playerID yearID stint teamID lgID   G  AB   R   H X2B X3B HR RBI SB CS
## 1 anderjo01   1901     1    MLA   AL 138 576  90 190  46   7  8  99 35 NA
## 2 bakerbo01   1901     1    CLE   AL   1   4   0   0   0   0  0   0  0 NA
## 3 bakerbo01   1901     2    PHA   AL   1   3   0   1   0   0  0   1  0 NA
## 4 barreji01   1901     1    DET   AL 135 542 110 159  16   9  4  65 26 NA
## 5 barrysh01   1901     1    BSN   NL  11  40   3   7   2   0  0   6  1 NA
## 6 barrysh01   1901     2    PHI   NL  67 252  35  62  10   0  1  22 13 NA
##   BB SO IBB HBP SH SF GIDP       era
## 1 24 NA  NA   3  4 NA   NA Dead.Ball
## 2  0 NA  NA   0  0 NA   NA Dead.Ball
## 3  0 NA  NA   0  0 NA   NA Dead.Ball
## 4 76 NA  NA   5  7 NA   NA Dead.Ball
## 5  2 NA  NA   1  0 NA   NA Dead.Ball
## 6 15 NA  NA   2 12 NA   NA Dead.Ball
```

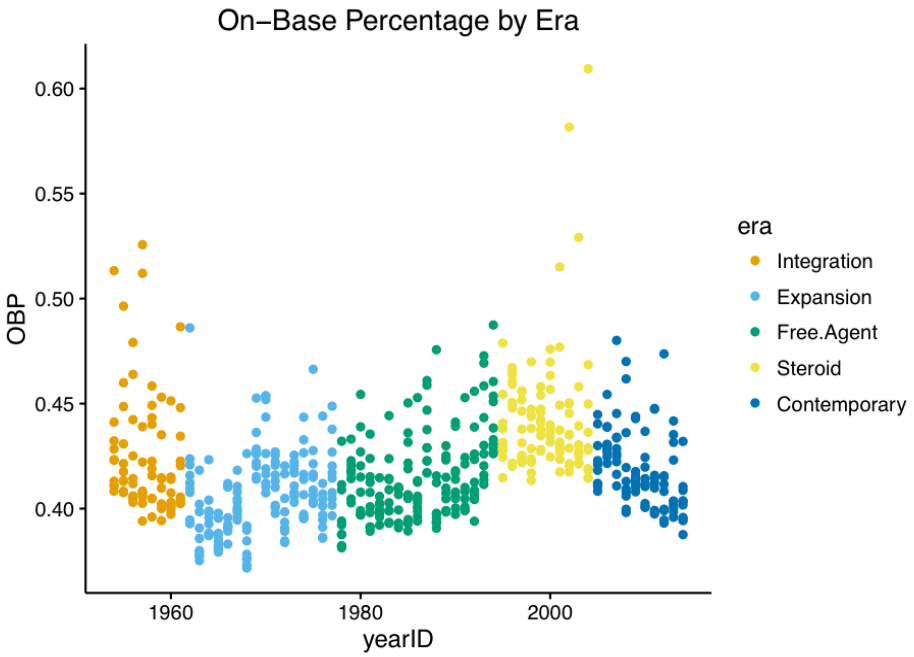
```
cbPalette <- c("#E69F00", "#56B4E9", "#009E73", "#F0E442",
               "#0072B2", "#D55E00", "#CC79A7", "#999999")

top_HR <- batting %>%
  group_by(yearID) %>%
  top_n(10,HR) %>%
  arrange(desc(HR)) %>%
  ggplot(.,aes(x=yearID,y=HR)) + geom_point(aes(colour=era),
  pch=19) + scale_colour_manual(values=cbPalette) + theme_classic() + ggtitle("Homeruns by Era")
top_HR
```




```
top_RBI <- batting %>%
  group_by(yearID) %>%
  top_n(10,RBI) %>%
  arrange(desc(RBI)) %>%
  ggplot(.,aes(x=yearID,y=RBI)) + geom_point(aes(colour=era),pch=19) +
  scale_colour_manual(values=cbPalette) + theme_classic() + ggtitle("Runs Batted In by Era")
top_RBI
```




```
top_OBP <- batting %>%
  filter(AB>250) %>%
  mutate(PA=AB+BB+HBP+SF,
         OBP=(H+BB+HBP)/PA) %>%
  group_by(yearID) %>%
  top_n(10,OBP) %>%
  arrange(desc(OBP)) %>%
  ggplot(.,aes(x=yearID,y=OBP)) + geom_point(aes(colour=era),pch=19) +
  scale_colour_manual(values=cbPalette) + theme_classic() + ggtitle("On-Base Percentage by Era")
top_OBP
```



Findings

Baseball's eras make sense historically, but when we view them statistically, they can be divided differently. When looking at home runs, there is a distinct difference in home run numbers between the Dead and Live ball eras, however the distribution of home runs is similar throughout the Integration, Expansion, and Free Agent eras, which span from 1943 to 1994. There is a dramatic increase in home runs hit during the Steroid era which then drops off in the Contemporary era once steroids became a more publicized issue and drug testing became common.

When examining the eras by runs batted in (RBI's), a trend similar to that found in home runs can be seen. The difference between the Dead ball and Live ball eras is very obvious, most likely explained by an increase in base hits following the addition of a better ball. Again, the Integration, Expansion, and Free Agent eras are very similar in numbers, followed by a spike in RBI's when entering the Steroid Era and a drop off in the Contemporary Era.

Finally, we examined on base percentage (OBP). OBP could only be measured after 1954 because it was not recorded prior to that year. The lack of data on the Dead and Live ball eras aside, a similar trend continues when compared to home runs and RBI's. The most obvious change that could occur in the labeling if it were based on statistics is the combination of the Integration, Expansion, and Free Agent eras.

Discussion, Overview, and Implications

By viewing home runs, runs batted in, and on-base percentage, it is clear that the values of each of these statistics are not distinct based on which baseball era one is looking at. The first major shift in batting statistics occurred following the introduction of a better baseball, represented by the transition from the

Dead Ball era to the Live ball era. The disparity between the three statistics during this period is likely a result of a ball that came off the bat with more force. The balls being hit harder would result in better batting statistics.

The Steroid era (1995-2004) is clearly distinct from the others in regards to the three batting statistics we examined. Players taking steroids were likely stronger than those not taking steroids in the pre-Steroid Era, resulting in more home runs, runs batted in, and a higher on base percentage. These higher numbers could be explained by the ability to hit the ball harder and farther. The only statistic in which the Steroid era isn't staggeringly different when compared to the others is on base percentage. While still higher than the other eras, one explanation for the less drastic difference would be that when hitting a higher number of home runs per at bat, the on base percentage wouldn't be as high.

Statistically, the Integration, Expansion, and Free Agent eras are essentially identical because these chronological delineations are a result of historical changes in the game. While these may have had some impact on style of play, the major changes occurred in how teams were managed financially and how the league was structured. These differences would have had no real impact on the hitting statistics we were examining.