Baseball Analysis: Homerun Records from 1871 to 2014

Tam Tran The

```r
hr_sum <- function(x) {
  x %>%
    select(playerID, yearID, HR) %>%
    summarize(homeruns = sum(as.numeric(HR), na.rm = TRUE))
}
```
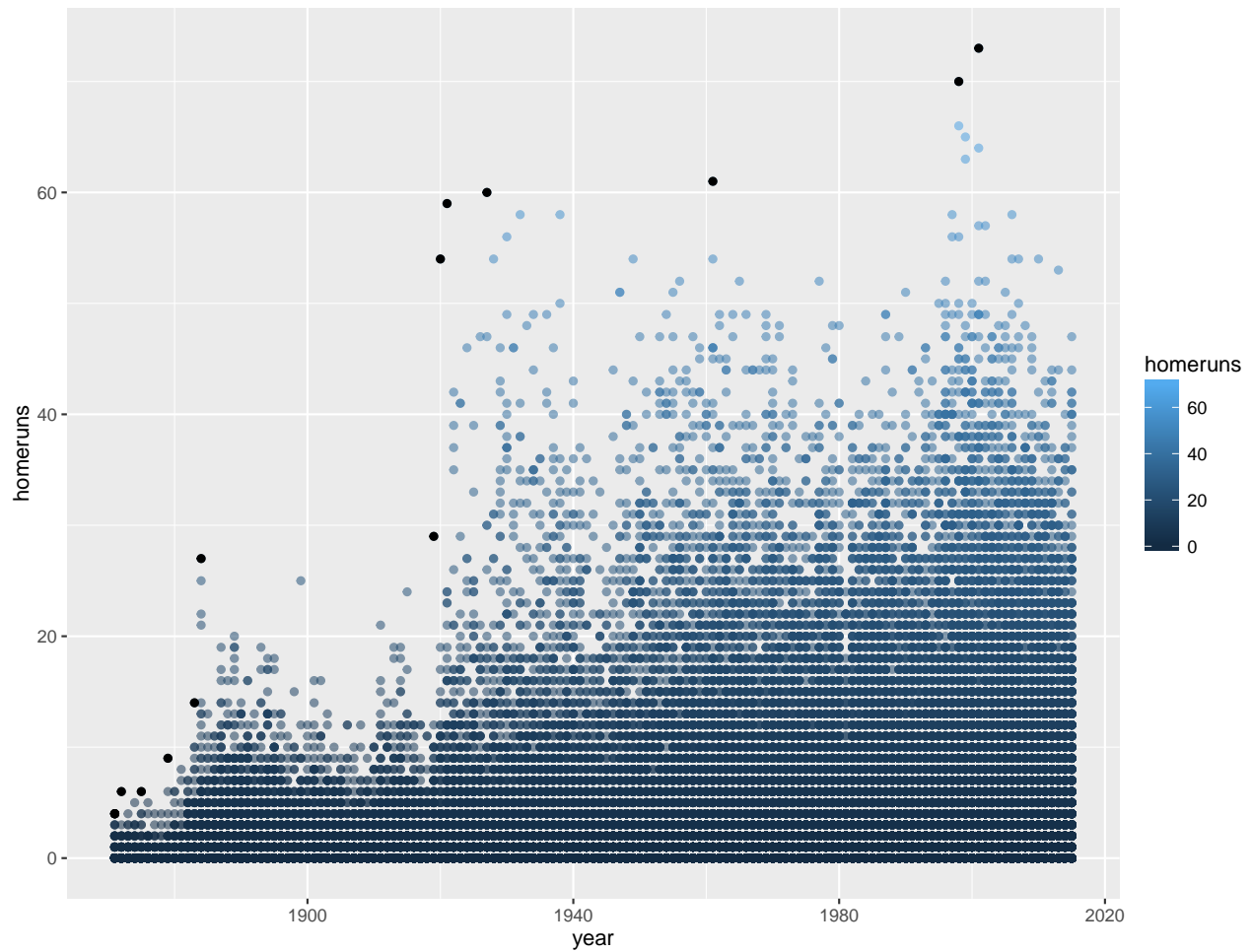
```r
 total_hr <- Batting %>%
  group_by(yearID, playerID) %>%
  do(hr_sum(.))
total_hr
```

```
## Source: local data frame [93,898 x 3]
## Groups: yearID, playerID [93,898]
##
##     yearID  playerID homeruns
##      <int>     <chr>    <dbl>
## 1    1871 abercda01        0
## 2    1871  addybo01        0
## 3    1871 allisar01        0
## 4    1871 allisdo01        2
## 5    1871 ansonca01        0
## 6    1871 armstbo01        0
## 7    1871 barkeal01        0
## 8    1871 barnero01        0
## 9    1871 barrebi01        0
## 10   1871 barrofr01        0
## # ... with 93,888 more rows
```

```r
record <- total_hr %>%
  group_by(yearID) %>%
  mutate(highest_hr = ifelse(homeruns == max(homeruns), 1, 0))
record <- record[record$highest_hr == 1,]
record <- record %>%
  ungroup() %>%
  mutate(new_record = ifelse(homeruns == cummax(homeruns), 1, 0))
record <- record[record$new_record == 1,]
```

```r
ggplot(total_hr, aes(x = yearID, y = homeruns)) +
  geom_point(aes(color = homeruns, position = "jitter", alpha = 0.4)) +
  geom_point(data=record, mapping=aes(x = yearID, y = homeruns, fill="red", position = "jitter")) +
  labs(title="Homerun Records from 1871-2014", x="year", y="homeruns") +
  guides(alpha=FALSE, fill=FALSE)
```

Homerun Records from 1871–2014


```r
# Check to see if the current highest record was set by Barry Bonds in 2001
master_info <- Master %>%
  select(playerID, nameFirst, nameLast) %>%
  unite(Name, nameFirst, nameLast, sep=" ", remove = TRUE)
total_hr <- left_join(total_hr, master_info, by = c("playerID", "playerID"))
total_hr[which.max(total_hr$homeruns),]
```

```
## Source: local data frame [1 x 4]
## Groups: yearID, playerID [1]
##
##   yearID  playerID homeruns        Name
##    <int>     <chr>    <dbl>       <chr>
## 1   2001 bondsba01       73 Barry Bonds
```