Project #2 stat319

Alan Lin

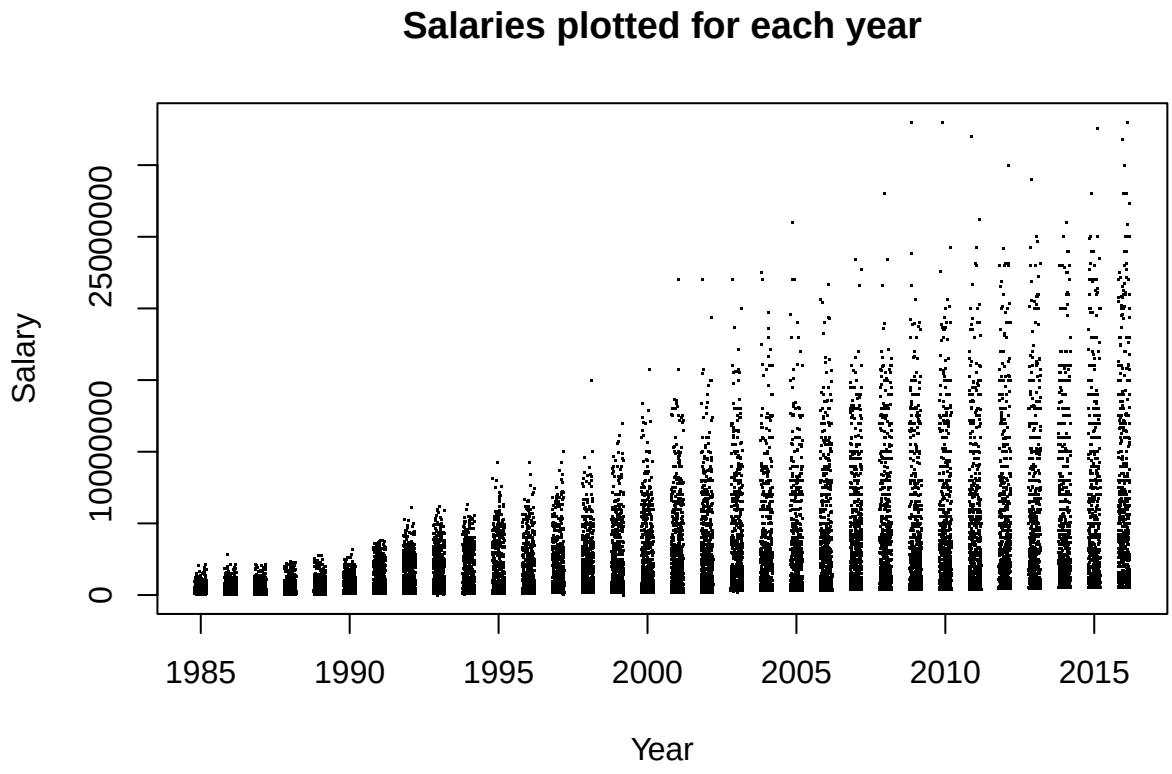
3/15/2022

Lecture 1.

1. How many observations is there salary information for? What range of years is there salary information for?

There are 26428 observations of salary information. The range of years is 1985 - 2016.

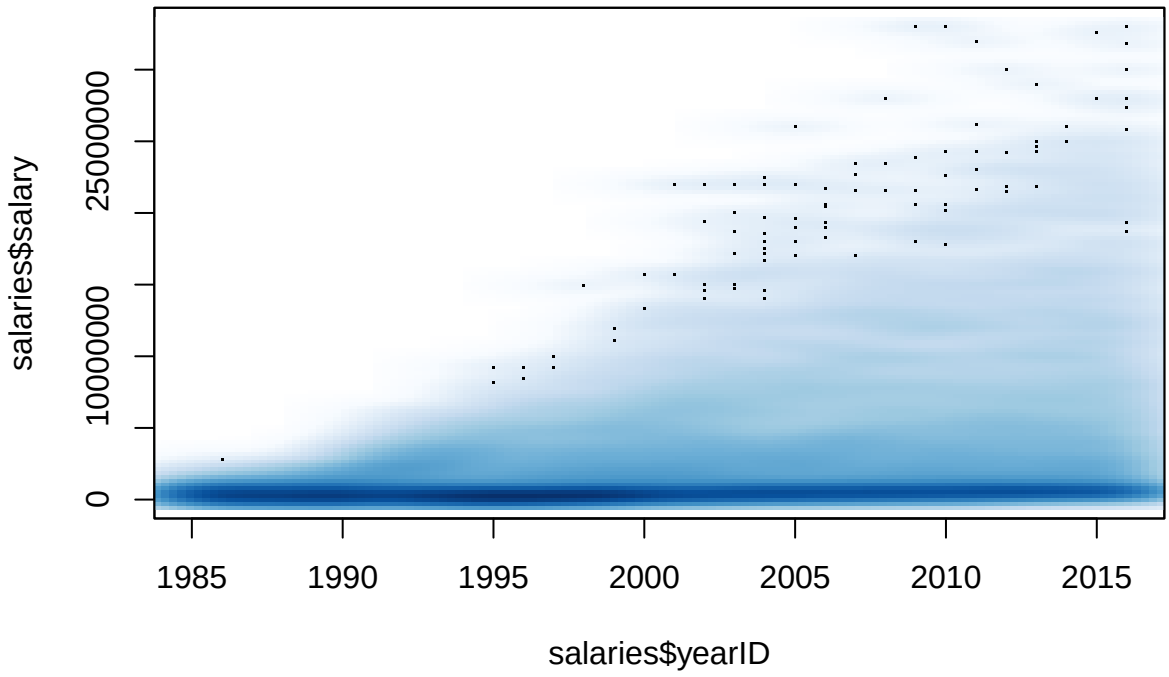
2. Create a scatter plot showing salaries plotted versus year.



3. Create a similar plot using smoothScatter

The high-density regions of points is the darkest shade of blue which indicates that the salaries are pretty low, though the scale would possibly indicate below 1 million a year, with several players that make more

than 30 million a year.



4. Fit a multiple linear regression model for salaries, using two predictor variables: year, and league, and interpret the coefficients.

If we don't count the years as factors, the model can be interpreted that, players in the year 0(really?) and the AL league can expect to have start with a salary of -271,424,769 that increases by 136,738 with each year, so that an AL player in the year 2000 would increase their starting salary by 273,476,000, which would be around 2 million dollars. A player in the NL league would just start at a lower salary(by 167213 dollars) in year 0.

5. Fit a similar model, except modelling salary on a log-scale. Once again, interpret the fitted model.

For every unit increase in the yearID, the salary will increase by 7.454%. In addition, being in the National League will decrease the salary by 4.834%,

6. Which of the previous two models appears to be a better fit? Explain

The model with salary on the log-scale seems to be a better fit, not only because the R-squared value is larger, but because when it comes to salaries, the log-scale tends to make the largest salaries more comparable to a lower salary.

7. Joining Teams table with Salaries table.

I first use a sub-query to join the Teams table with the Salaries table, where I also filter for data in the year 2016. Then I can use "GROUP BY" to group each team, SUM each team's salaries, and then "ORDER BY" to reorder the salaries in descending order. This produces the following table.

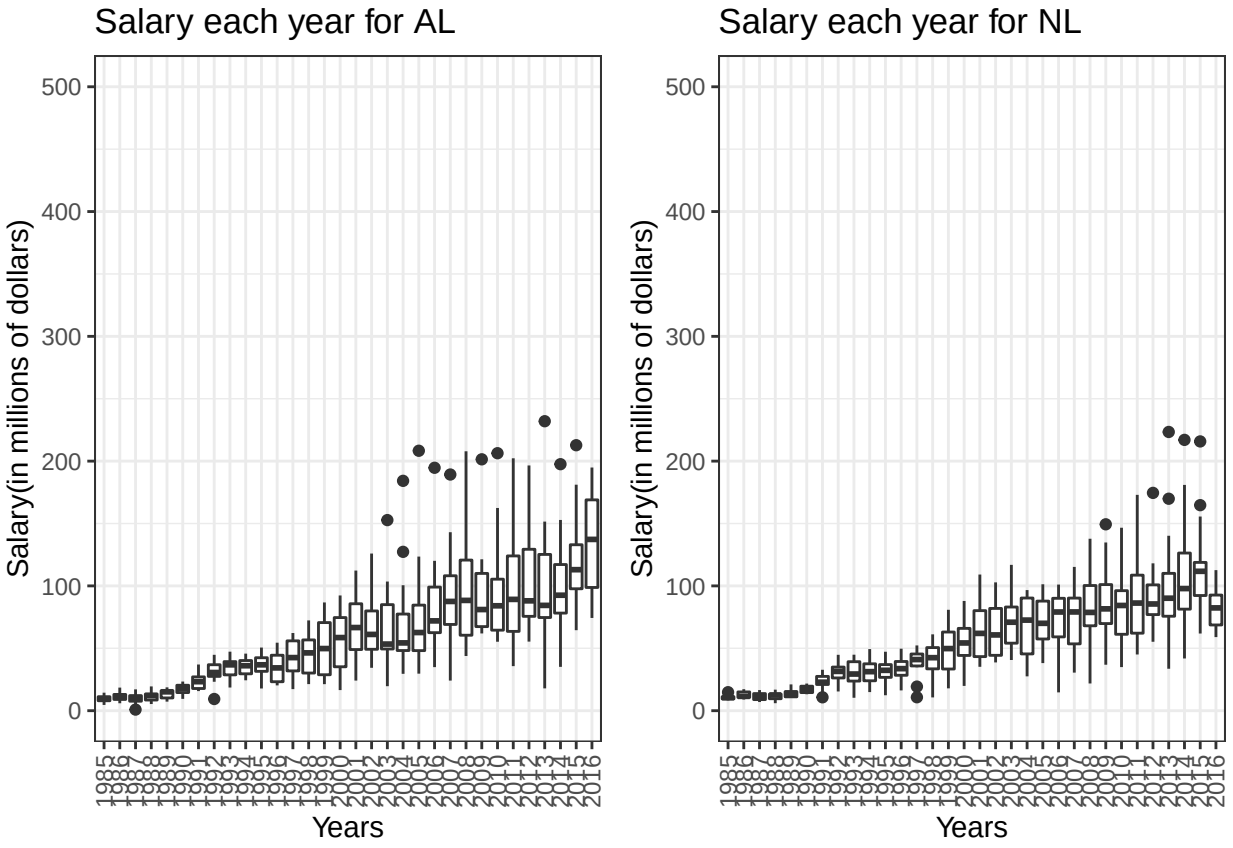
```
##    teamID       sum
## 1     DET 194876481
## 2     BOS 188545761
## 3     TEX 176038723
## 4     BAL 161863456
## 5     TOR 138701700
## 6     LAA 137251333
## 7     SEA 135683339
## 8     COL 112645071
## 9     PIT 103778833
## 10    MIN 102583200
## 11    HOU  94893700
## 12    CIN  88940059
## 13    ARI  87439063
## 14    OAK  86806234
## 15    MIA  77314202
## 16    CLE  74311900
## 17    MIL  68775237
## 18    ATL  68498291
## 19    PHI  58980000
```

8. Extract a dataframe that has, for each combination of yearID and teamID, the total salary of that team in that year, and also the league the team played in. Find the number of rows in this dataframe.

We can use a similar query as the previous one but instead selecting for distinct columns, removing the filter for 2016, and adding the lgID column. There are 907 rows in this dataframe.

```
##    teamID lgID yearID       sum
## 1     ANA   AL   1997  31135472
## 2     ANA   AL   1998  41281000
## 3     ANA   AL   1999  55388166
## 4     ANA   AL   2000  51464167
## 5     ANA   AL   2001  47535167
## 6     ANA   AL   2002  61721667
## 7     ANA   AL   2003  79031667
## 8     ANA   AL   2004 100534667
## 9     ARI   NL   1998  32347000
## 10    ARI   NL   1999  68703999
```

9. For each league, create a plot with vertical boxplots that show the distribution of the total team salaries of one year(in millions of dollars).

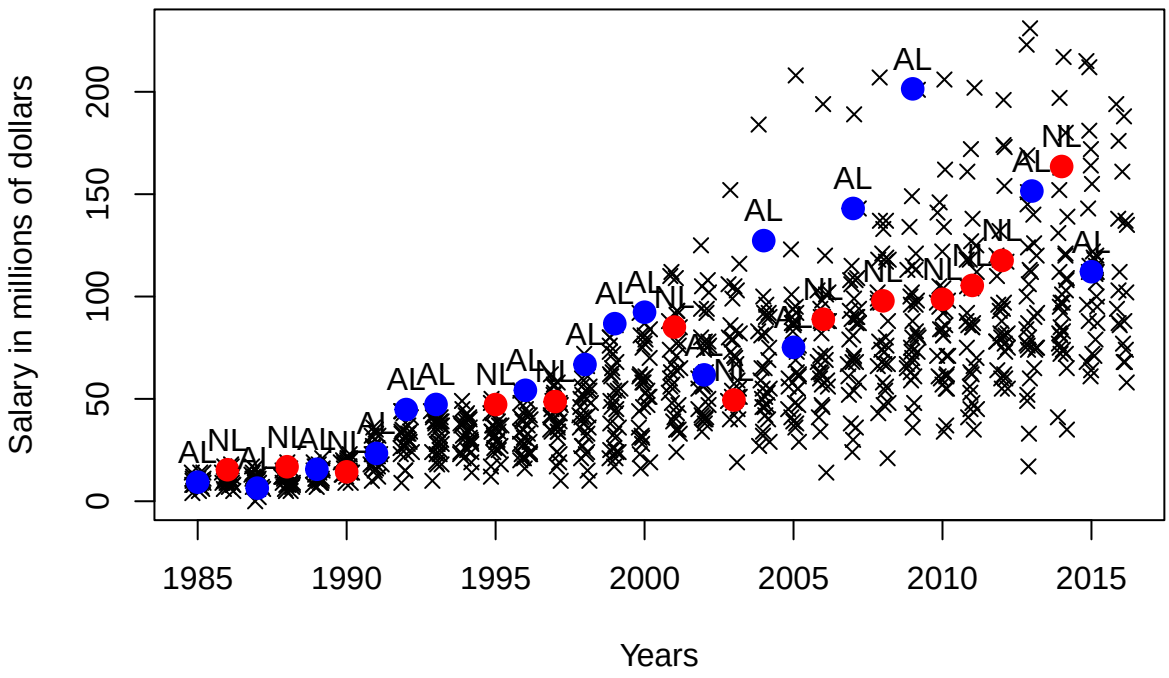


10. How many times has the World Series winner been from the American League, and how many times from the National League? Find the average salary of the World Series winning team within each league.

17 times from the American League, 13 times from the National League. The average salary of the WS winning team in the AL is 77,596,664 dollars and 73,004,014 in the NL.

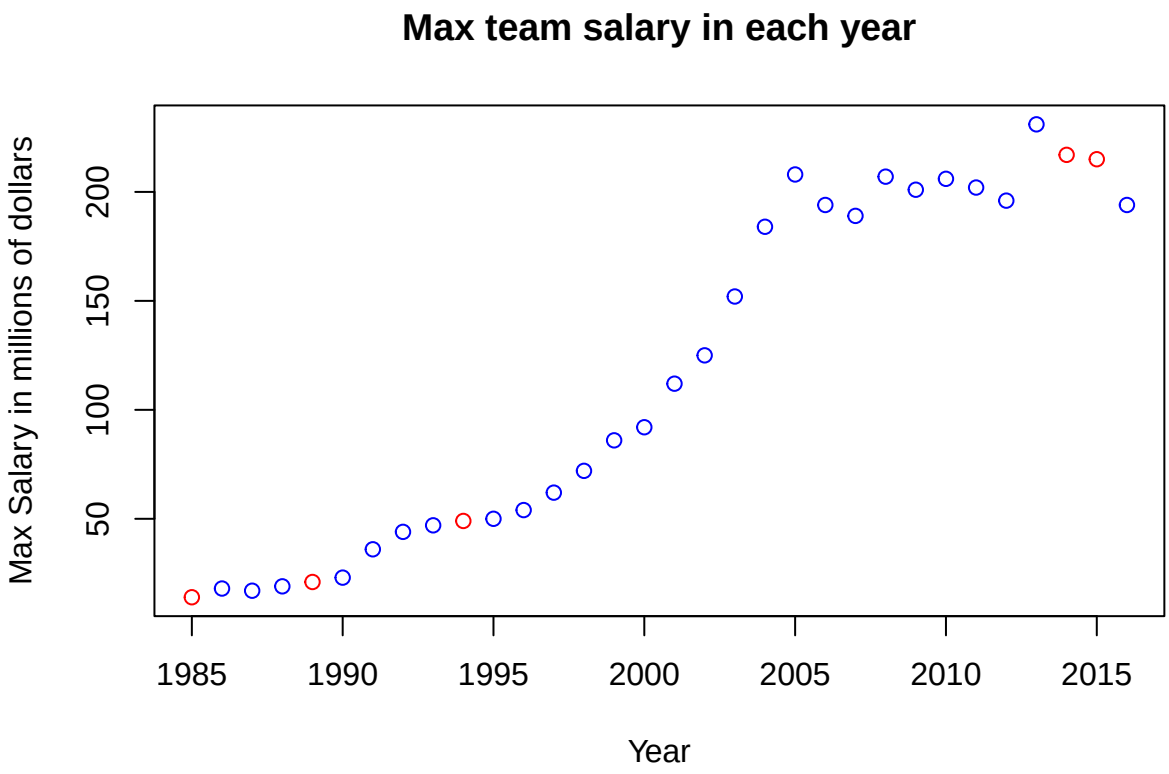
11. Create a plot of total salaries of all teams in each and salary of the World Series winning team each year. Draw any conclusions about the salaries of the winning team.

Looking at the position of the red and blue dots indicating WS winning teams, the salaries of these teams aren't always the highest paid teams for each year. Many times, the team that wins the World Series are actually one of the average paid teams, with the exception of the early 1990s to the early 2000s, where the winning teams were some of the top paid teams of that year.



12. Create a plot showing the maximum team salaries for each year. Draw conclusions about the highest team salaries over time, and salaries within each league.

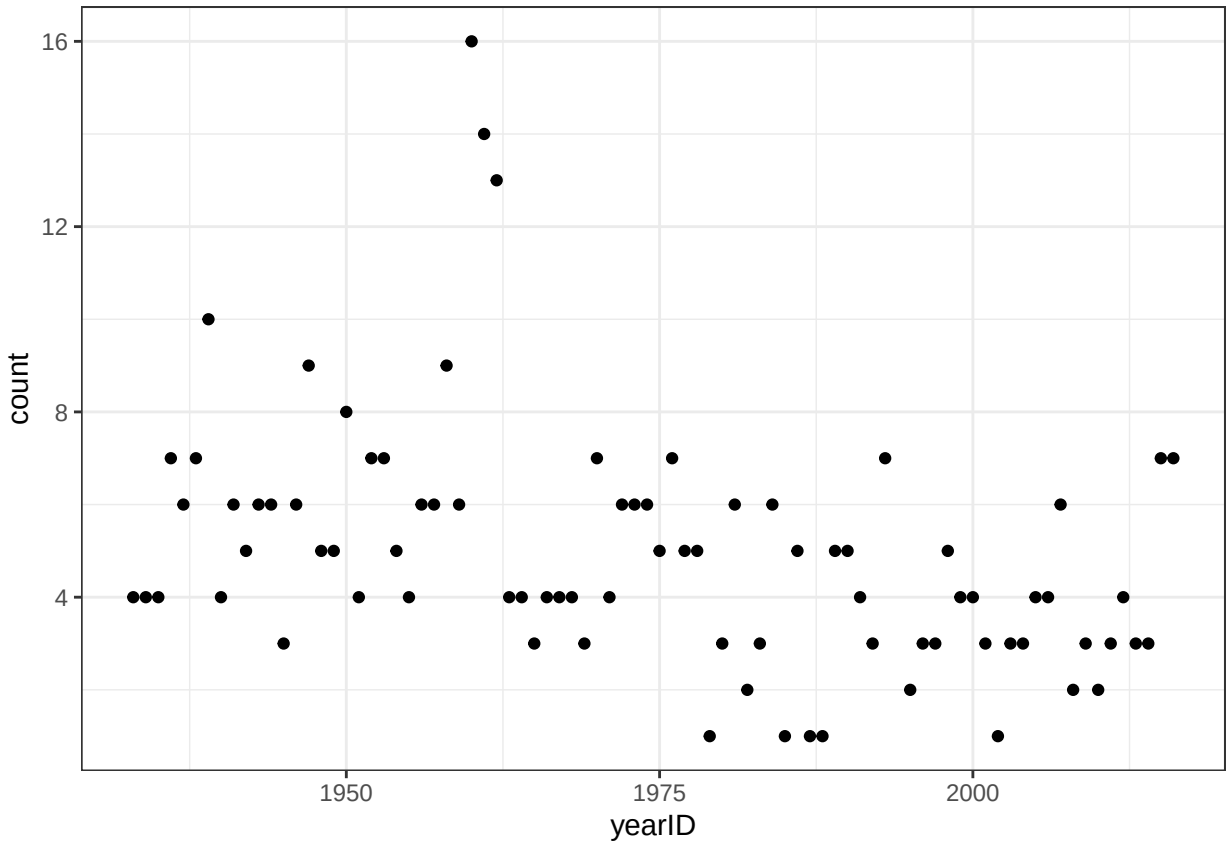
There seems to be a positive trend in the max team salaries over time. More and more players are being paid higher salaries, which makes sense as the sport is more popularized and commercialized than ever before. The vast majority of the points in the plot indicates that the American League has most of the highest team salaries compared to the National League, with only 5 teams from the National League being the max salary team in the MLB from 1985 to 2016.



13. Create a data frame with the number of All Star players on the World Series winning team for each year. Find the 5 years with the most All Star players on the winning team, and create a plot of the data frame.

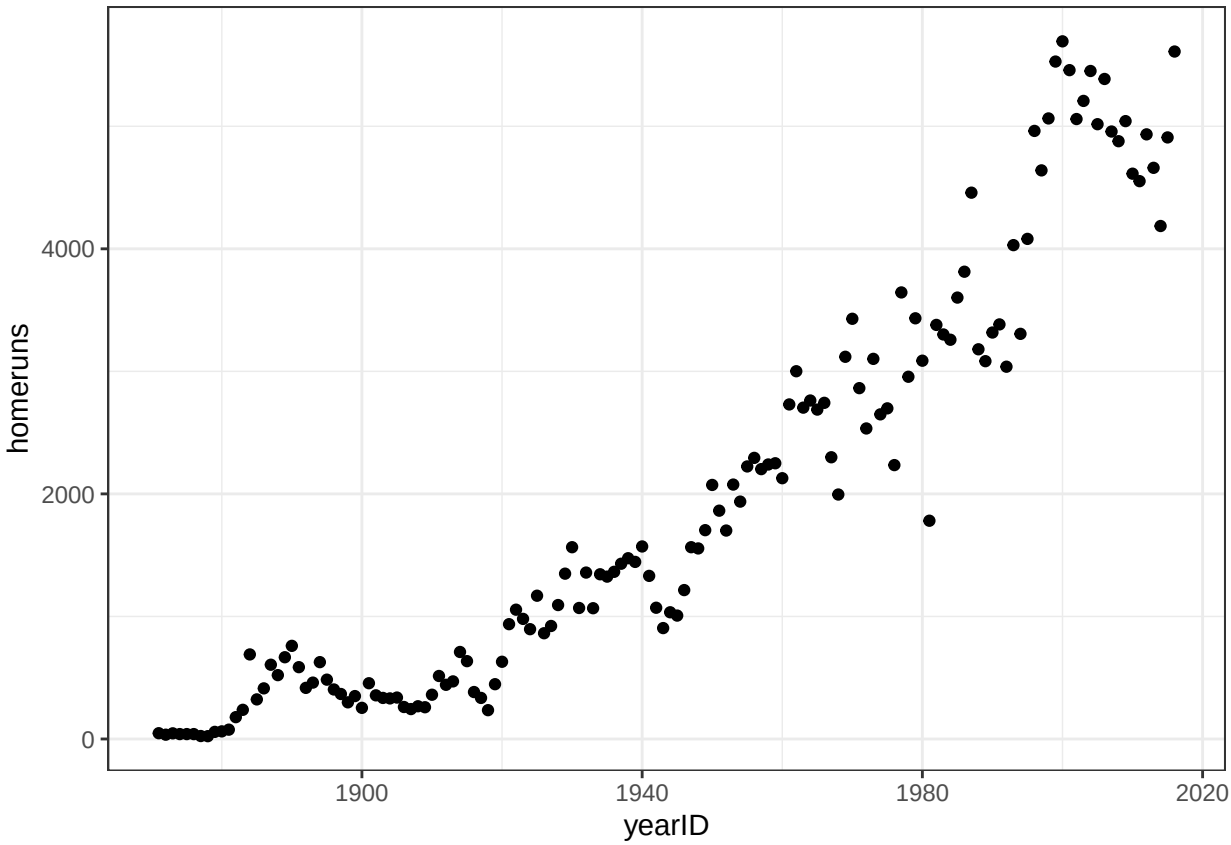
The five years with the most All Star players on the winning team is 1939, 1947, 1960, 1961, 1962.

```
##   yearID teamID count
## 1   1960    PIT    16
## 2   1961    NYA    14
## 3   1962    NYA    13
## 4   1939    NYA    10
## 5   1947    NYA     9
```

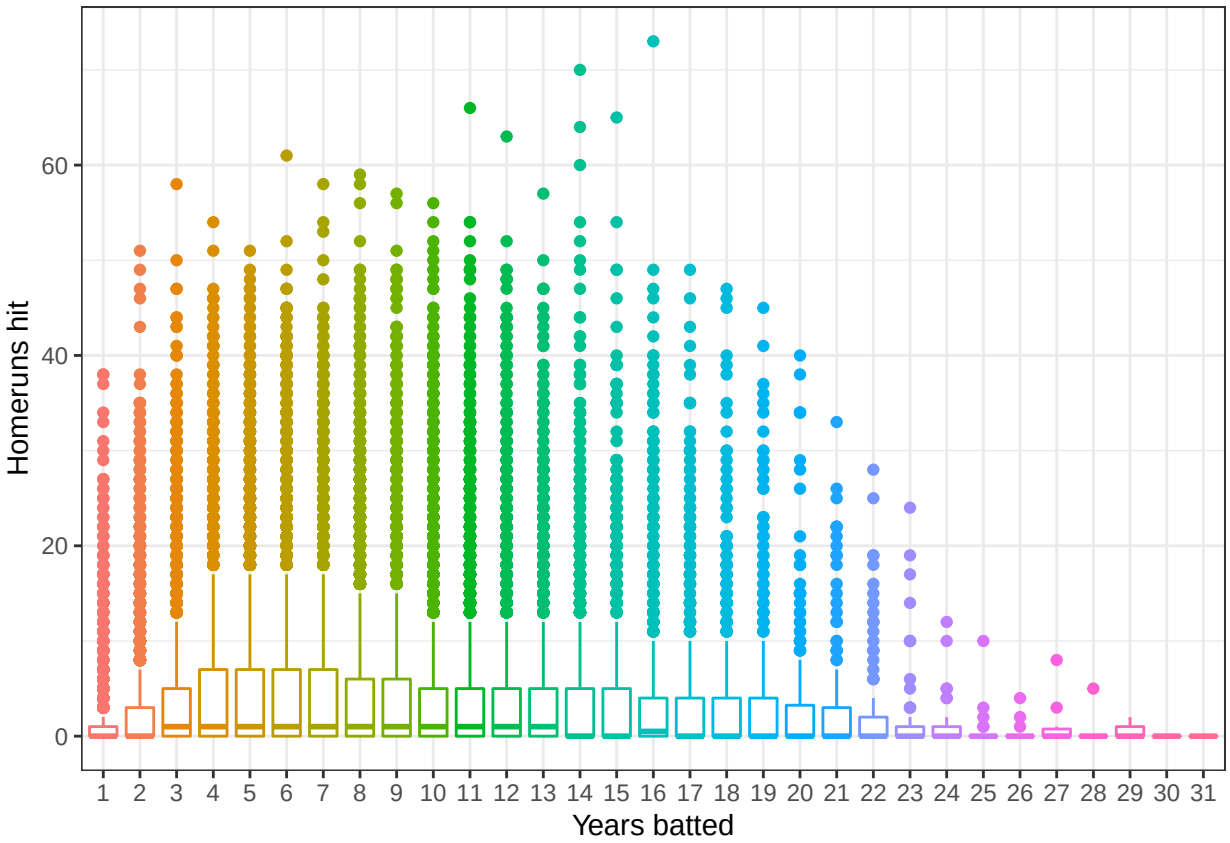


14. Has the distribution of the number of home runs hit by players in a given year changed over time? Now consider individual players over time. Do players tend to hit progressively more home runs throughout their career? If not, then what trends do you notice over time? Only consider players with at least a 10 year batting history.

There does seem to be a positive trend in the number of home runs hit by players each year, with more and more home runs hit on average compared to previous years.



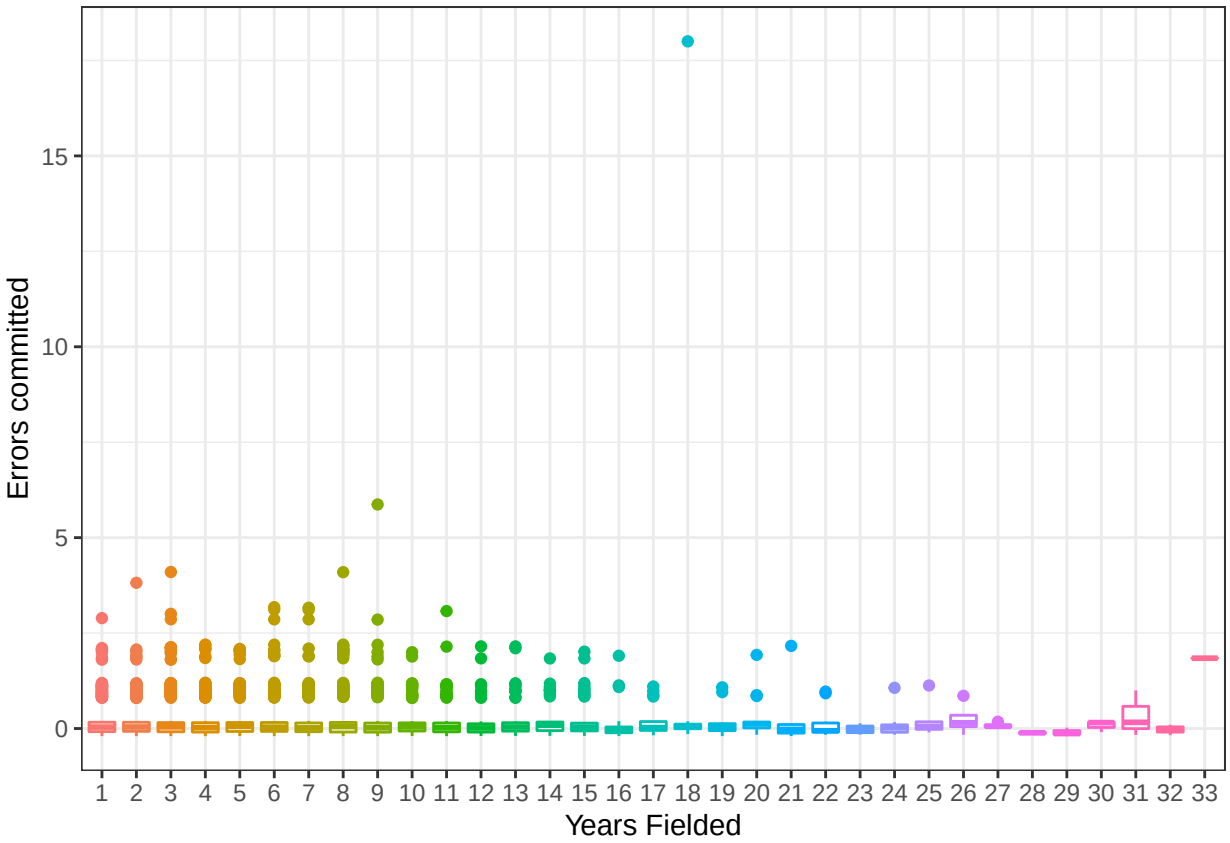
I plotted the home runs that each player that had at least 10 entries in the Batting table as boxplots with the x-axis being the number of years they've been in the league. That way, we can see the distribution of home runs as a player progresses through their career. There aren't any indications that players progressively hit more home runs as they stay in the league longer. However, it looks like players do improve in the first few seasons before they start to decline.



15. Pose another question of your own, of comparable difficulty.

I was interested in whether players make less errors as they progress throughout their career. As players become veterans, do they make less mistakes while out in the field? The boxplot below shows players that have at least 8 years worth of fielding experience and the errors they committed.

It doesn't look like there is a very noticeable trend in making less errors as players progress through their career, although it does looke slightly like a downward trend.


Errors committed
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
Years Fielded