Justin Byun
Professor Daniel Turek
Statistics 319
5 April 2022

SQL: Baseball Statistics Project Report

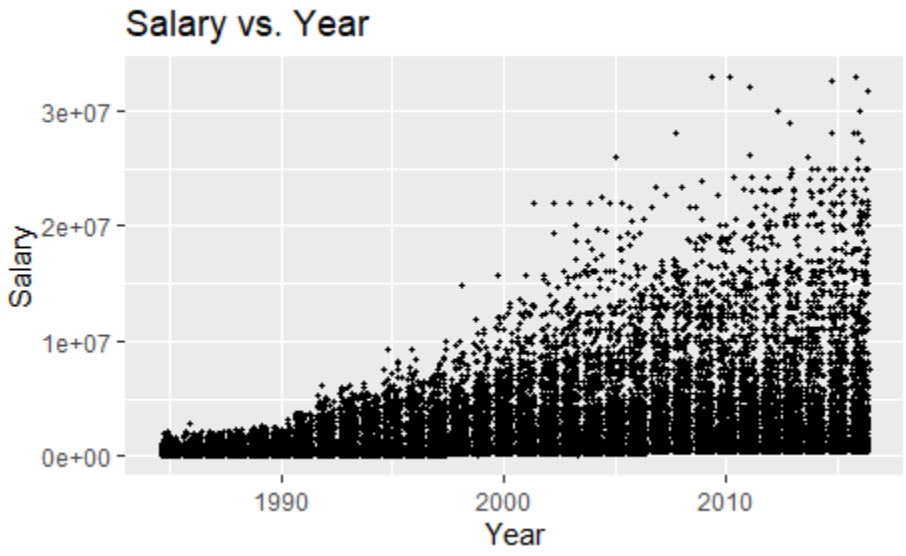
Introduction:

For this project, we used SQLite to investigate the world of Sabermetrics, also known as the statistics of baseball. Through various SQL queries and the RSQLite package in RStudio, we found answers to numerous statistical questions based on data available from Sean Lahman's Baseball SQLite database. The statistical questions are labeled as "deliverables" below, and the answers to these deliverables, along with the code used to acquire them, are obtained just below the actual question. Furthermore, these deliverables (and their answers) will be accompanied by figures related to them.

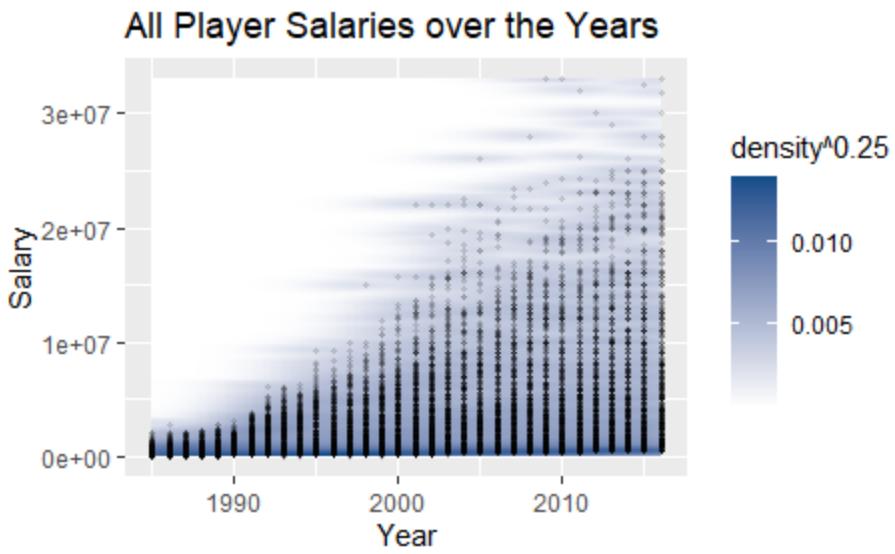
Deliverables:

#1: Using a query to access the yearID, salary, and lgID from the Salaries table and using the nrow(), the min(), and the max() functions, we found that there is salary information for 26428 observations, which span from the years 1985 to 2016.

#2: The scatter plot showing salaries versus year is shown below:



#3: A plot similar to the one above but with smoothScatter() is shown below:



#4: After fitting a multiple linear regression model for salaries on year and league, we found that the coefficients for yearID and lgIDNL were 136738 and -167213, respectively. This means that for every year that passes, we expect the salary of a baseball player to increase by 136738 dollars, on average. Furthermore, The difference in the average salary of a baseball player between the National League and the American League is 167213 dollars (AL - NL).

#5: The coefficients for yearID and lgIDNL, after we modeled salary on a log-scale, were 0.07190 and 0.04955, respectively. This means that For every year that passes, we expect the log(salary) to increase by around 0.07190, on average. Furthermore, 0.04955 is the difference between log(salary) in the National League and the American League (AL - NL).

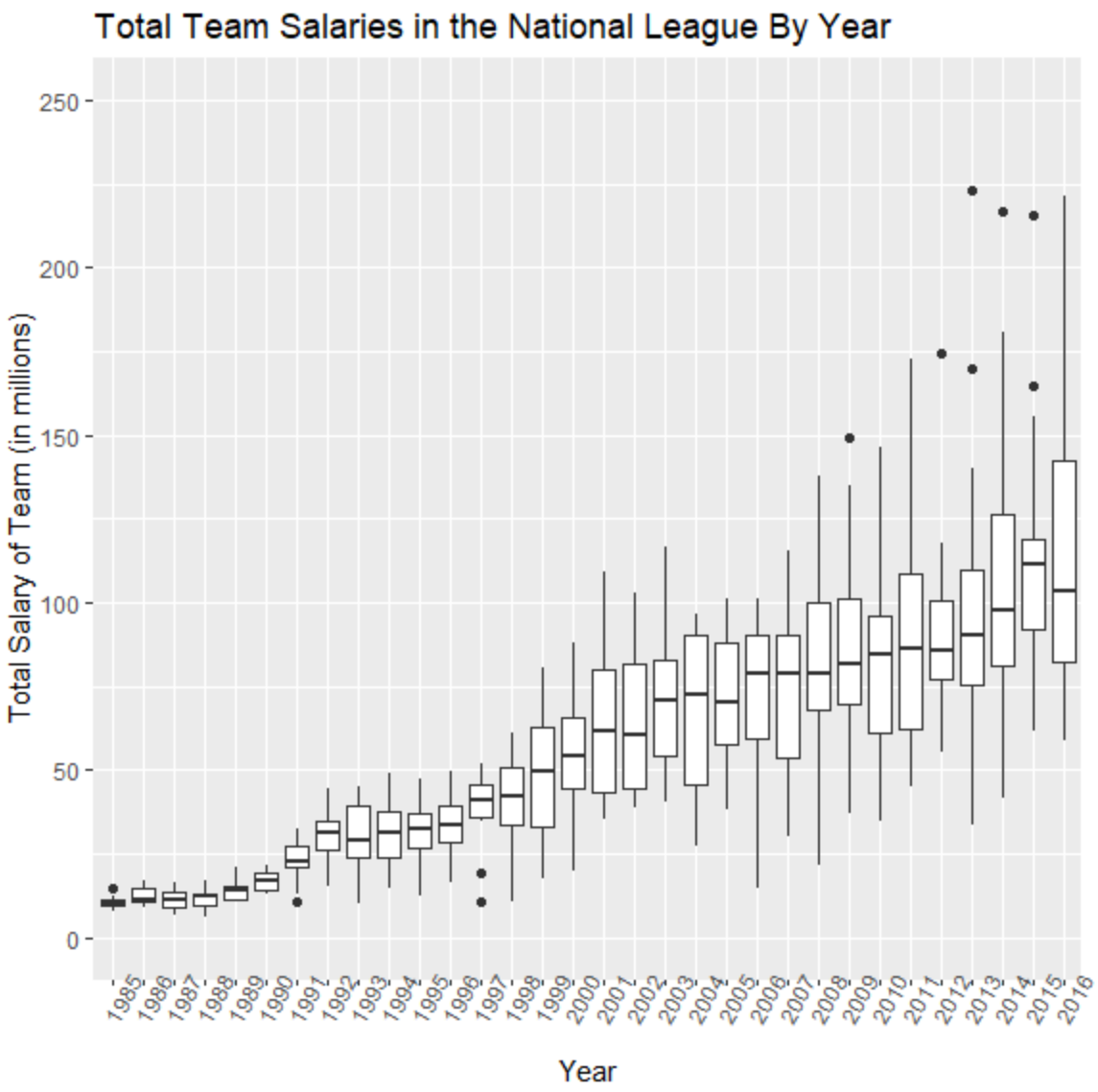
#6. The multiple regression model with the log-scale appears to be a better fit, most because of its r^2 values and the residual standard error, compared to the model without the log-scale. While the latter has an r^2 value of 0.1243, the former has a r^2 value of 0.2099. Furthermore, the model with the log-scale for salary has a significantly lower residual standard error of 1.242, compared to 3234000 for the model without the log-scale.

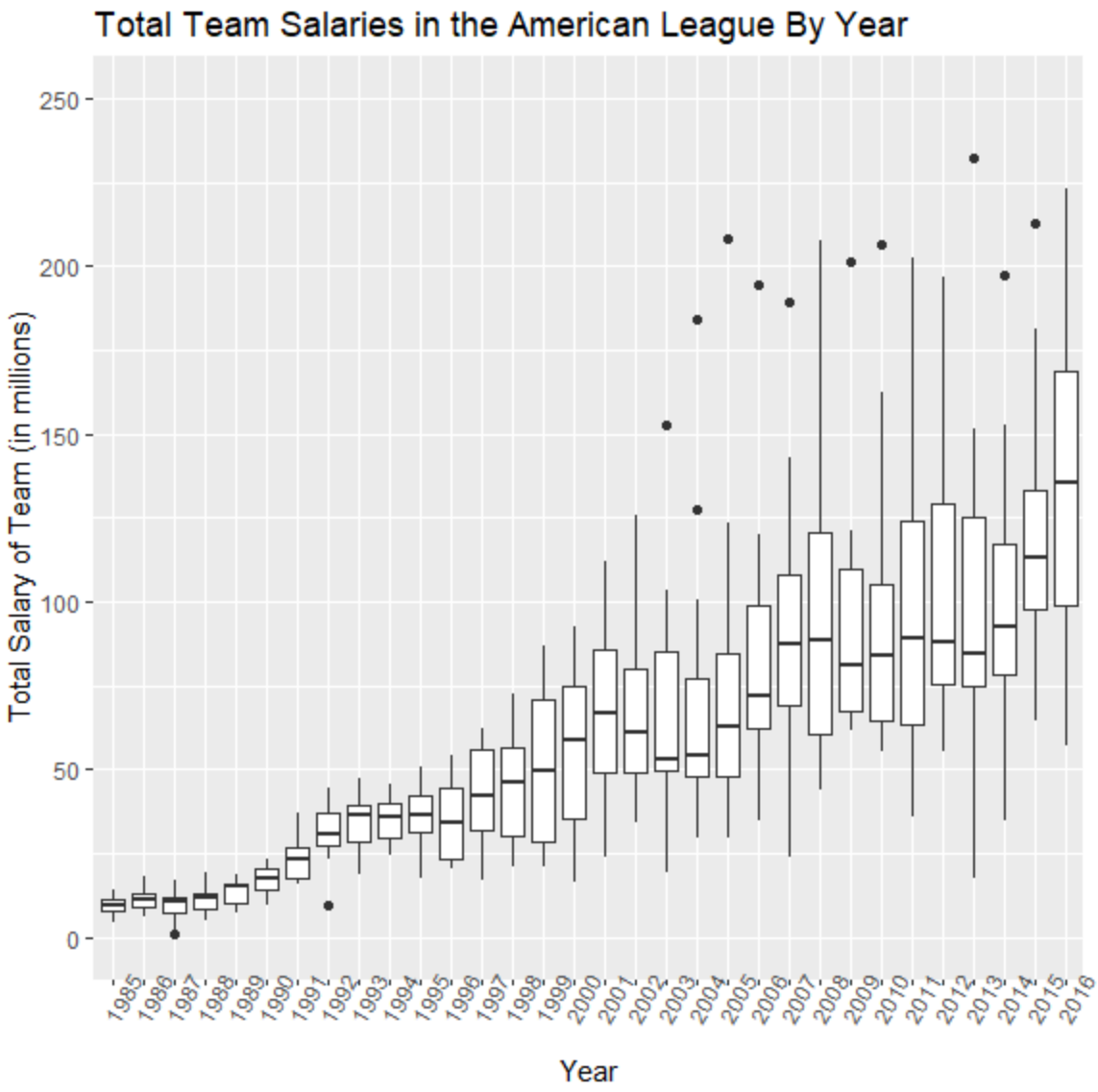
#7. The teams with the highest and lowest salaries in the year 2016 were the Detroit Tigers (DET) and the Philadelphia Phillies (PHI), respectively. DET's total salary was $194876481, while PHI's was $58980000. The total salary for each team in the year 2016 is shown below:

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

#8. The number of rows in the data frame containing every combination of yearID and teamID, as well as the team's total salary that year and the league they played in, is 918.

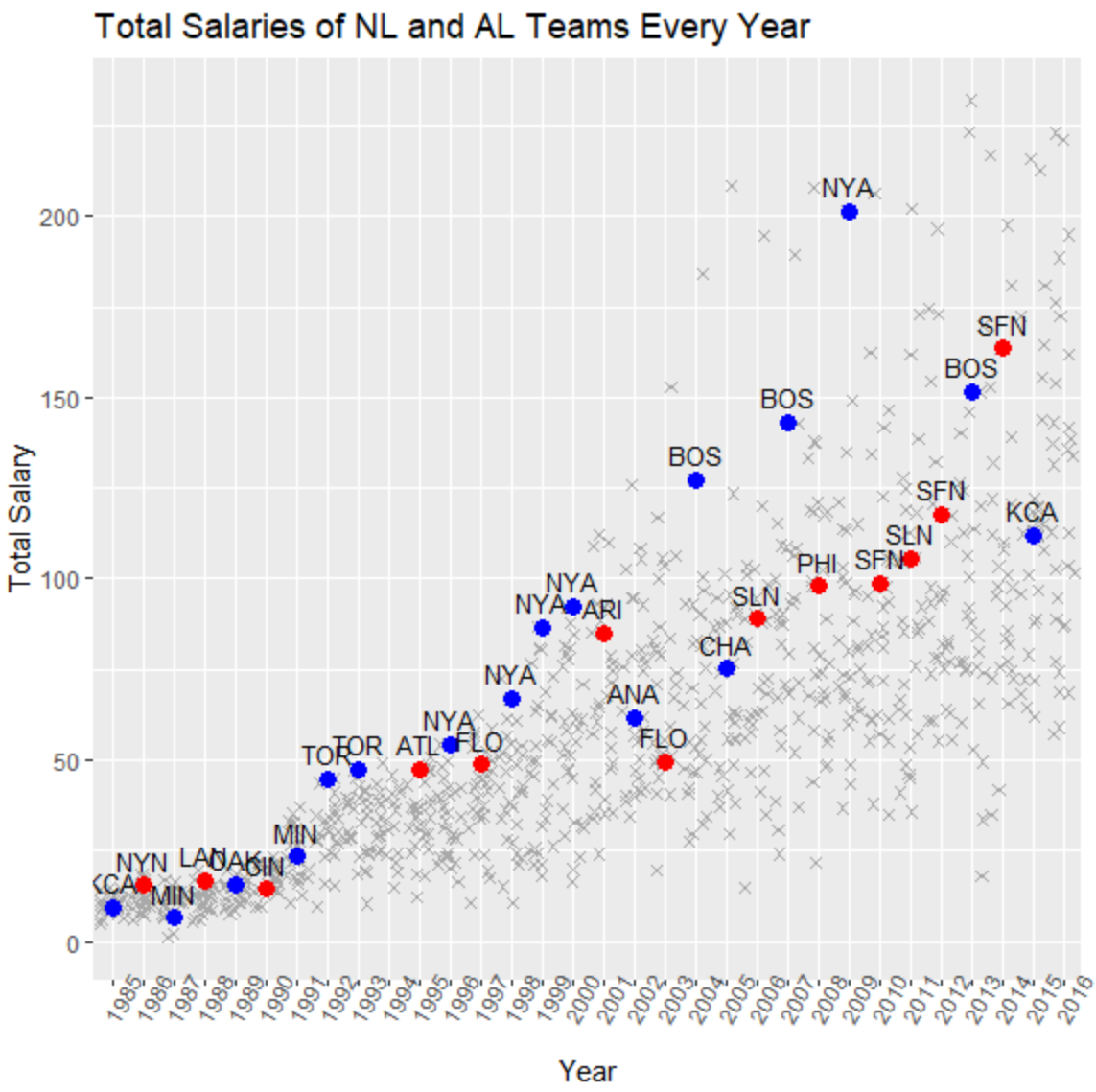
#9. A box plot displaying the distribution of the total team salaries of each year (in millions of dollars) is shown below for each league:


Total Team Salaries in the National League By Year
Total Salary of Team (in millions)
250
200
150
100
50
0
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Year



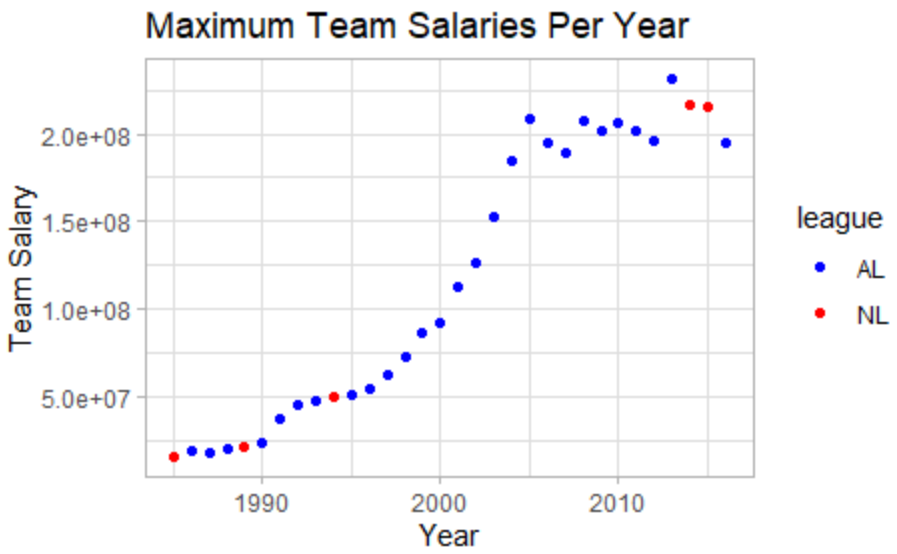
#10. The World Series Winner has been from the American League 17 times, and it has been from the National League 13 times. The average salary of the World Series winning team for the National League is 73004014 dollars. For the American League, the average salary is 77596664 dollars.

#11. The plot demonstrating the team salaries as small gray x's and the total salary of the World Series winner for each year as larger dots is shown below. Blue points represent American League winners, while red points represent National League winners.



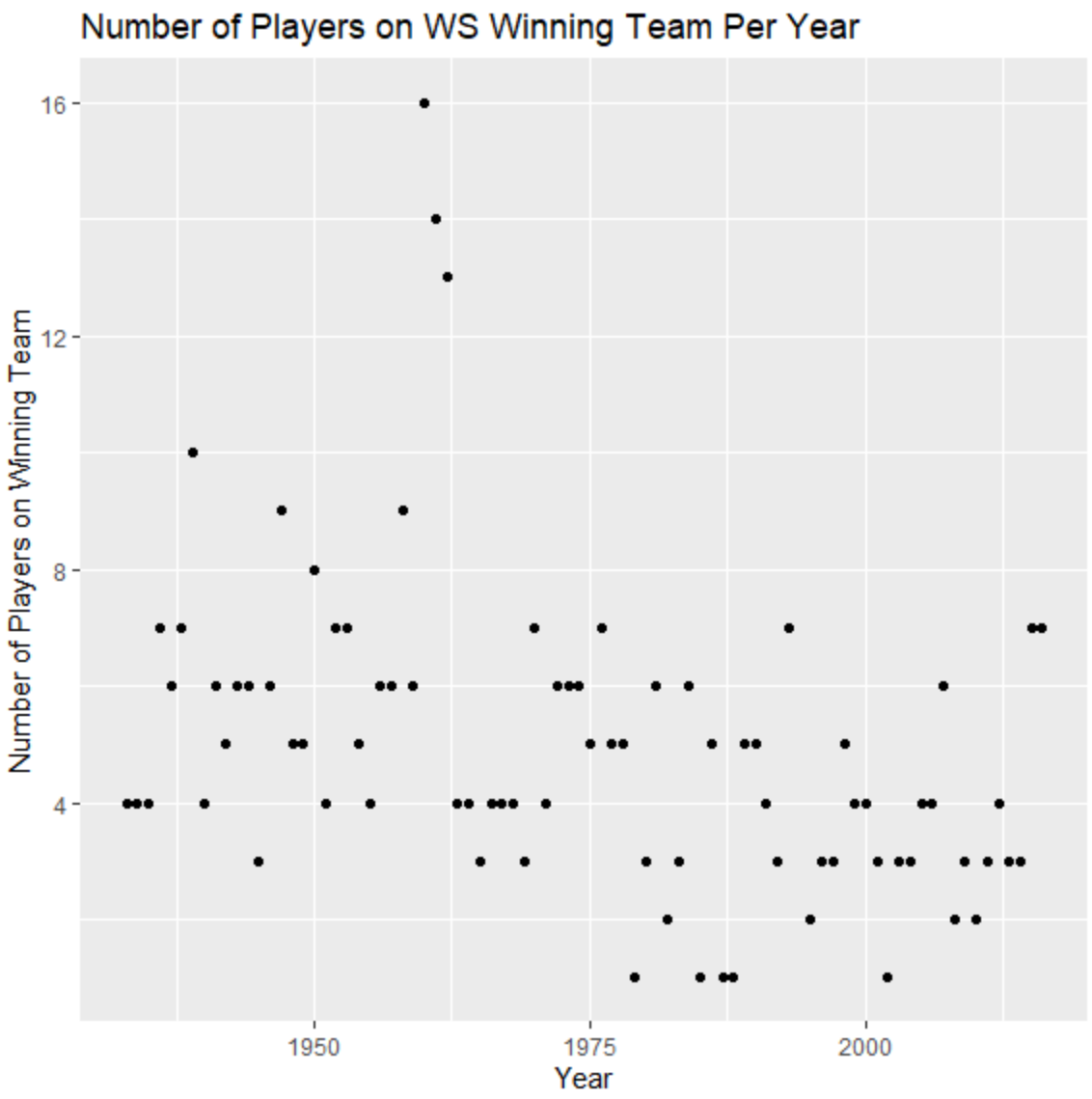
Overall, we see that the total salaries of the winning team every year in the World Series are often higher than most of the other teams in a given year, regardless of whether the team belongs to the National League or the American League.

#12. The plot showing the maximum team salaries for each year is shown below. Blue points represent teams from the American League, while red points represent teams from the National League.

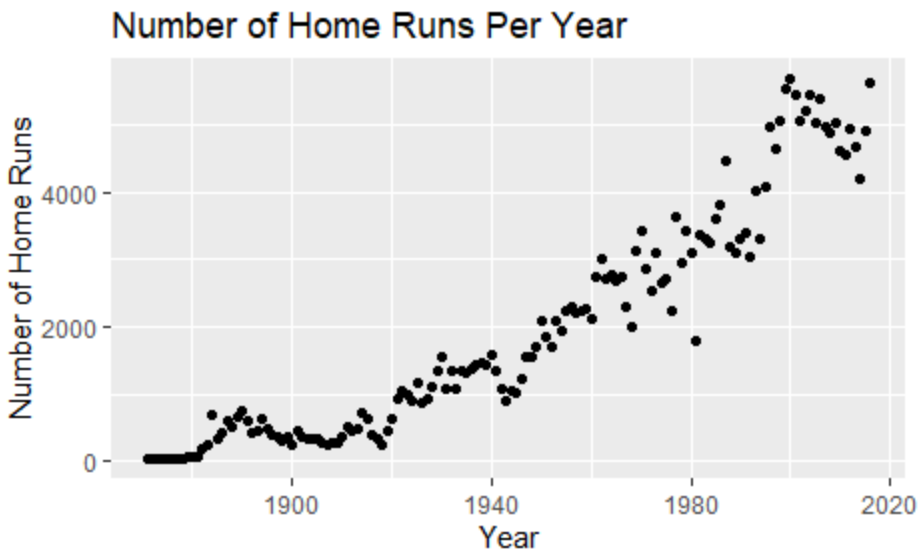


Over the years, we see that the highest team salaries per year slowly get higher and higher, with no exceptions from 1985 until 2005. Starting from 2005, we see that the highest team salary every year starts to stagnate and not increase steadily like it did before 2005. It decreases at some points and increases at other times. It seems that the American League tends to spend much more on player salaries than the National League, given that from 1985 to 2016, a majority of the highest salary teams per year have been from the American League.

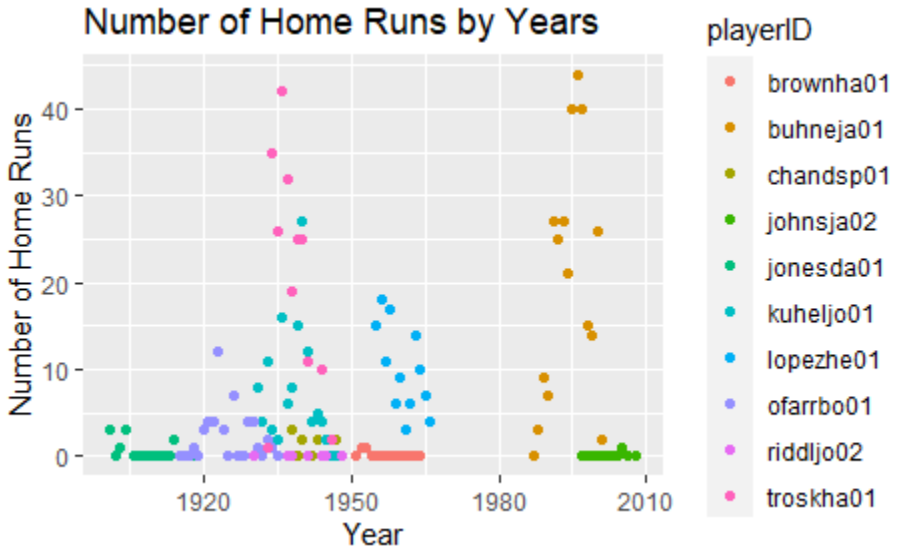
#13. The five years with the most All Star players on the winning team are 1960, 1959, 1962, 1961, and 2011. The plot showing the number of All Star players on the winning team for each year is shown below:



#14. A plot showing the total number of home runs scored per year is shown below:



As the years passed, baseball players gradually made more and more home runs. The scatterplot above shows a generally steady increase in the number of homeruns over the years. In order to find out if players tend to hit progressively more home runs as they advance in their career, we took a random sample of 10 players from the dataset and plotted the number of home runs they made each year they played. The plot can be found below:

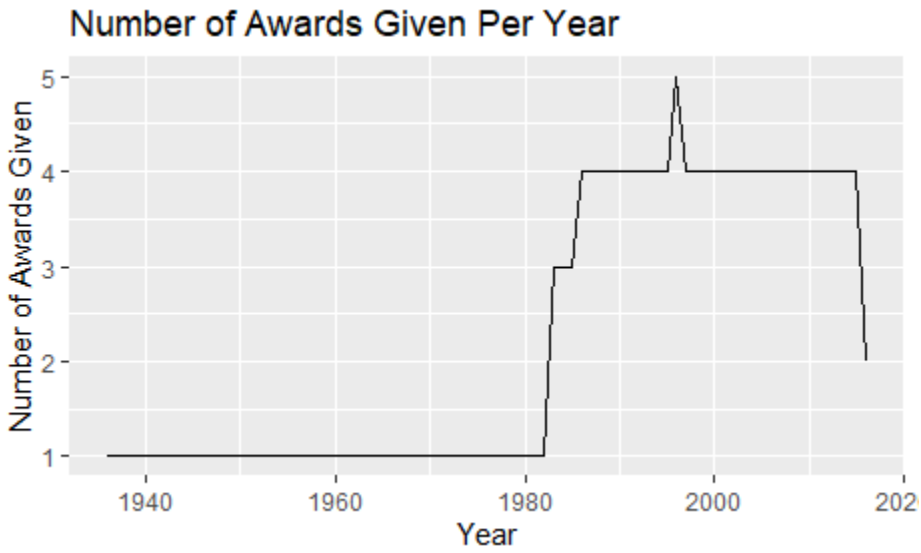


Based on the ten random players we chose who have at least a 10 year batting history, it appears that players generally do hit progressively more home runs throughout their career. Some players, however, don't see much improvements in terms of how often they score home runs. But others can see a dramatic increase in the number of home runs they hit in a single year. However,

these players who do see an improvement eventually reach a peak, after which the number of home runs they hit starts to gradually decrease further on in their career.

#15. For this question, we decided to create a dataframe containing the number of awards that each award-winning manager has received every year. We were interested in finding out the proportion of all managers who have won at least one award, along with how many (and what percentage) have won more than one. We also wanted to create a plot demonstrating how many awards were given out each year, and observe if the distribution of that number changed over the years.

We found that 3.78% of all managers have won at least one award, and that 49 managers (or 1.43% of all managers) have won more than one award. The plot showing the number of awards given out each year is below:



According to the line graph above, for much of the 1900's, only one award was given out every year to managers. Starting in the mid-1980's, however, more and more awards were given out each year. The number of awards given out in a given year peaked in the mid-1990's, when 5 awards were given to managers in just a single year. Since then, 4 awards were given out, and the number dropped to 2 in the last recorded year, 2016.