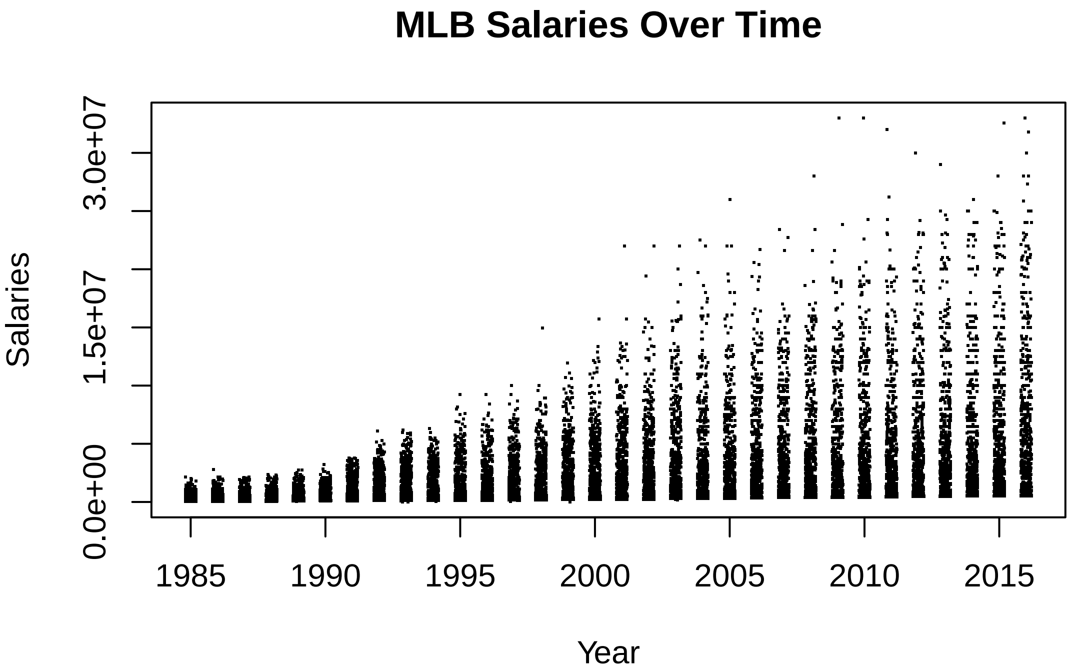
SQL: Baseball Statistics Project

Isabel Arvelo

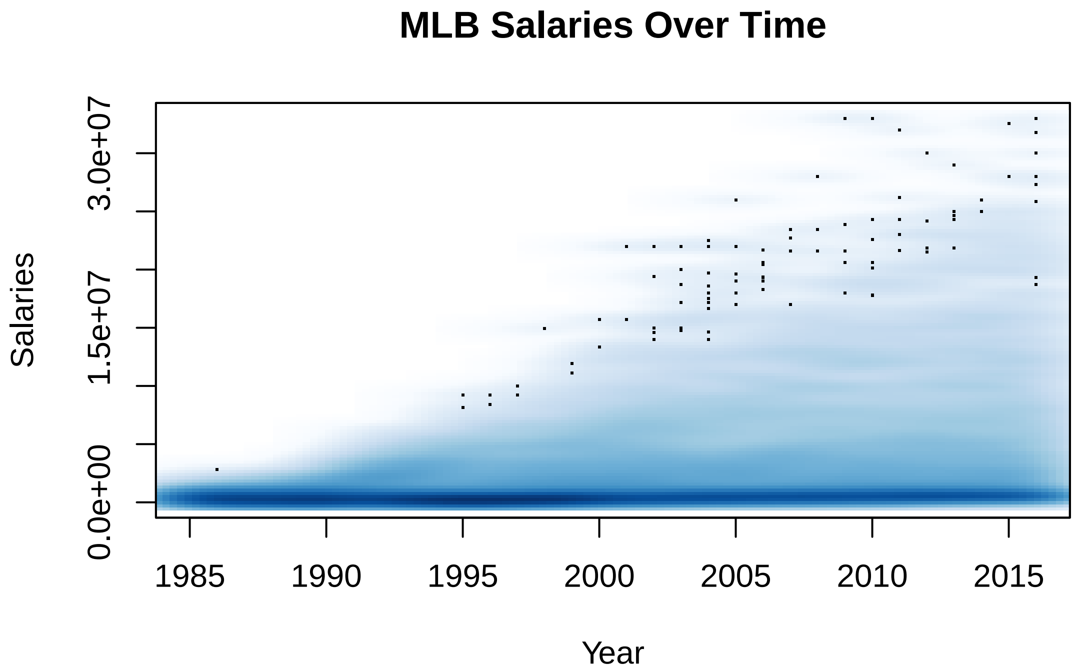
2022-04-05

1. There are 26,428 observations in the SQL Salaries table. There is salary information for the years 1985 to 2016.

2.



3.



4.

Holding league constant, each one year increase in year is associated with an expected average increase of $136,738 in salary. Since the p-value (< 2e-16) is less than .05,we can reject the null hypothesis and conclude that year is a statistically signifcant predictor variable for salary. In other words, changes in the year are associated with changes in salary at the population level.

Holding year constant, an individual in the 'NL' league is expected to earn, on average, $167,212 less than an individual in the 'AL' league. Since the p-value (2.68e-05) is less than .05, we can reject the null hypothesis to conclude that league is a statistically signiifcant predictor variable for salary. Changes in league are associated with changes in salary at the population level.

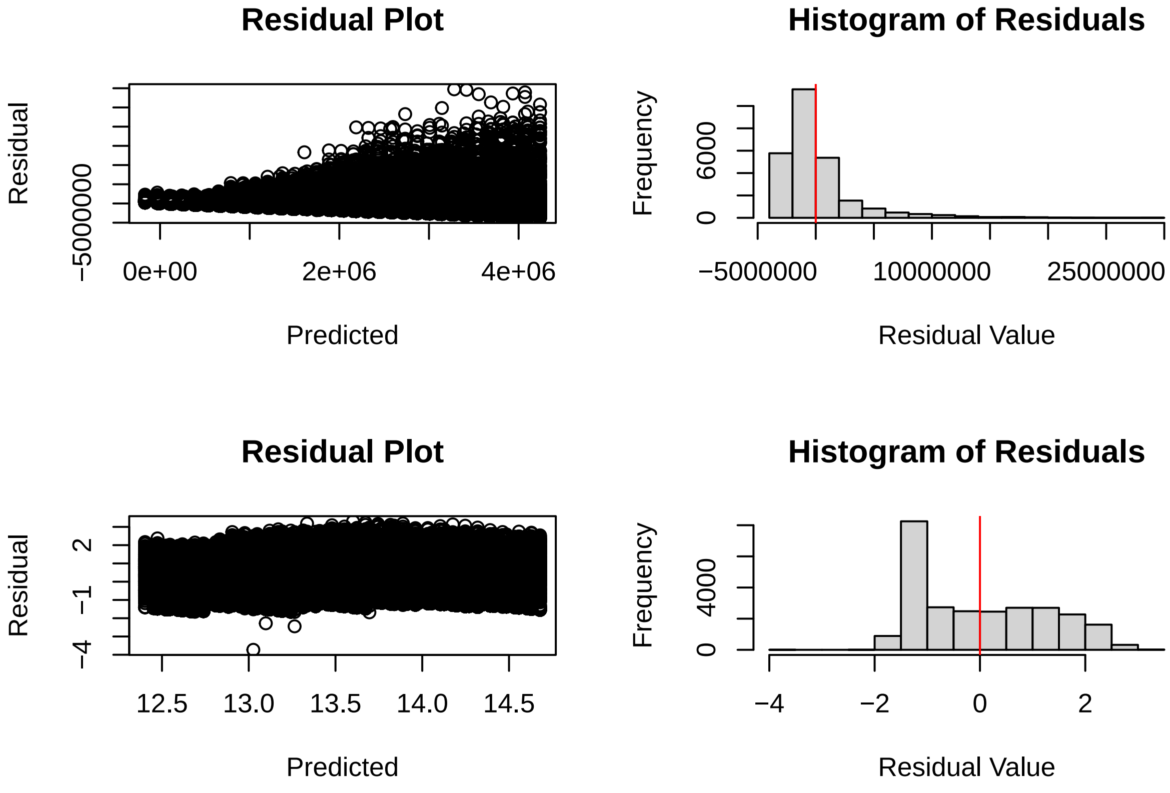
5.

```
##
## Call:  glm(formula = log(salary) ~ yearID + league_NL2, data = Salaries_2)
##
## Coefficients:
## (Intercept)       yearID   league_NL2
##  -130.14268      0.07184     -0.04953
##
## Degrees of Freedom: 26425 Total (i.e. Null);  26423 Residual
## Null Deviance:       51220
## Residual Deviance: 40410     AIC: 86230
```

Holding league constant, each one year increase in year is associated with an expected average increase of 0.07184 dollars in log(salary). Since the p-value (< 2e-16) is less than .05,we can reject the null hypothesis to conclude that year is a statistically signifcant predictor variable for log(salary). In other words, changes in the year are associated with changes in log(salary) at the population level.

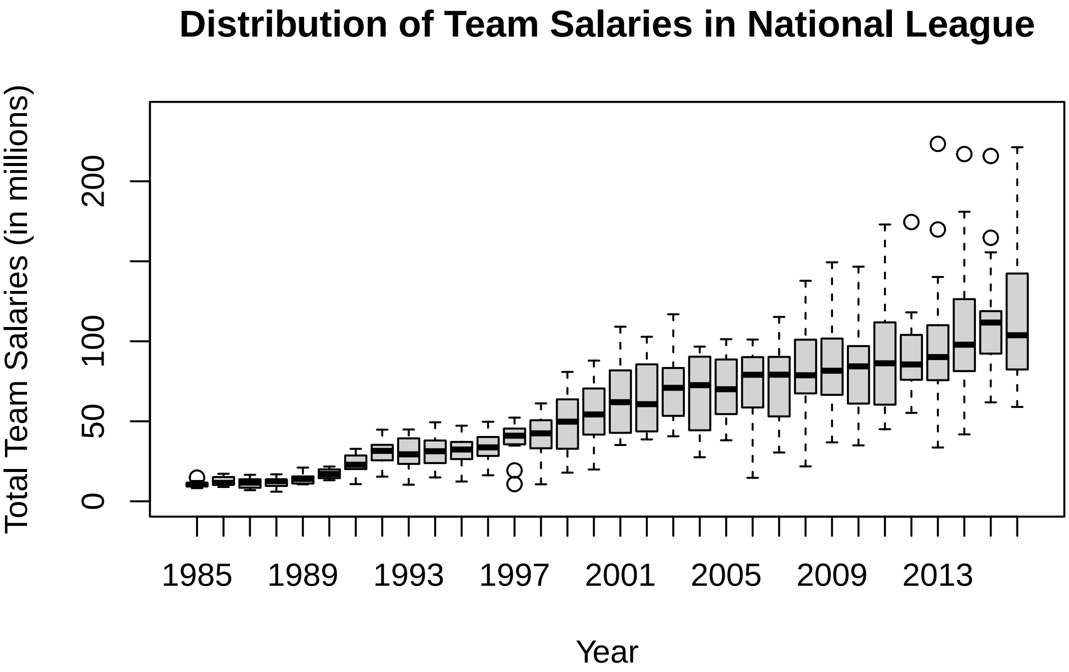
Holding year constant, an individual in the 'NL' league is expected to, on average, have a log(salary) $0.49 less than an individual in the 'AL' league. Since the p-value (0.00115)is less than .05, we can reject the null hypothesis to conclude that league is a statistically signifcant predictor variable for log(salary). In other words, changes in the year are associated with changes in log(salary) at the population level.

6. . Modelling salary on a log scale appears to be a better fit because it has a smaller residual deviance, a lower AIC value (which penalizes for fit and complexity) and it better meets the assumption of a linear model. The top two plots correspond to the original model and the bottom two to the model with a log transformation of salary.

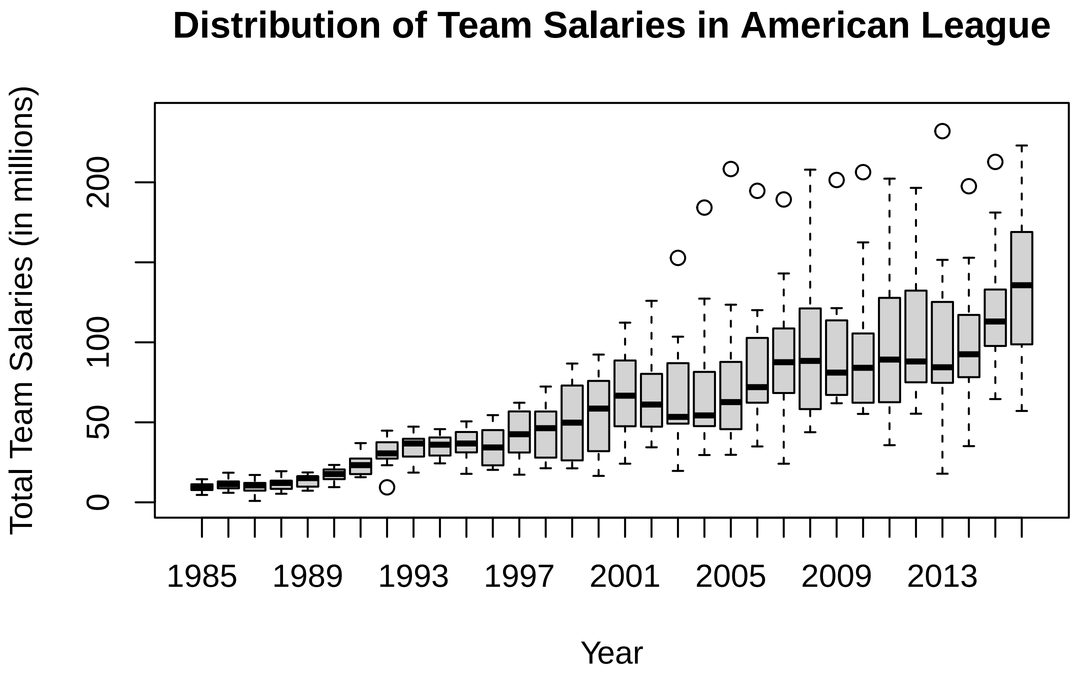


The normal probability plot indicates that the residuals for the second model are more normally distributed because they fall more approximately along a straight line. The histogram of the residuals of both models are skewed right, but the residuals of the transformed model are less widely dispersed. The errors of the second model all have approximately the same variance σ . This is called homoscedastic. For the most part, the points appear to be about the same distance from the regression line.This homogeneity of variance or homoskedascity is seen in the residual plot because the points are equally spread out indicating that that the points have the same scatter or finite variance. However, in the first model there's an observable pattern in the residuals and the fanning out shape in the residual plot indicates heteroskedasticity.

7. The team with the highest salary in 2016 was the Detroit Tigers, with a total sum of 194,876,481 dollars paid in salaries. The team with the lowest salary was the Philadelphia Phillies with a total sum of 58,980,000 dollars paid in salaries.
8. There are 918 rows in the data frame that has the total salary and the league (lgID) for each combination of yearID and teamID.

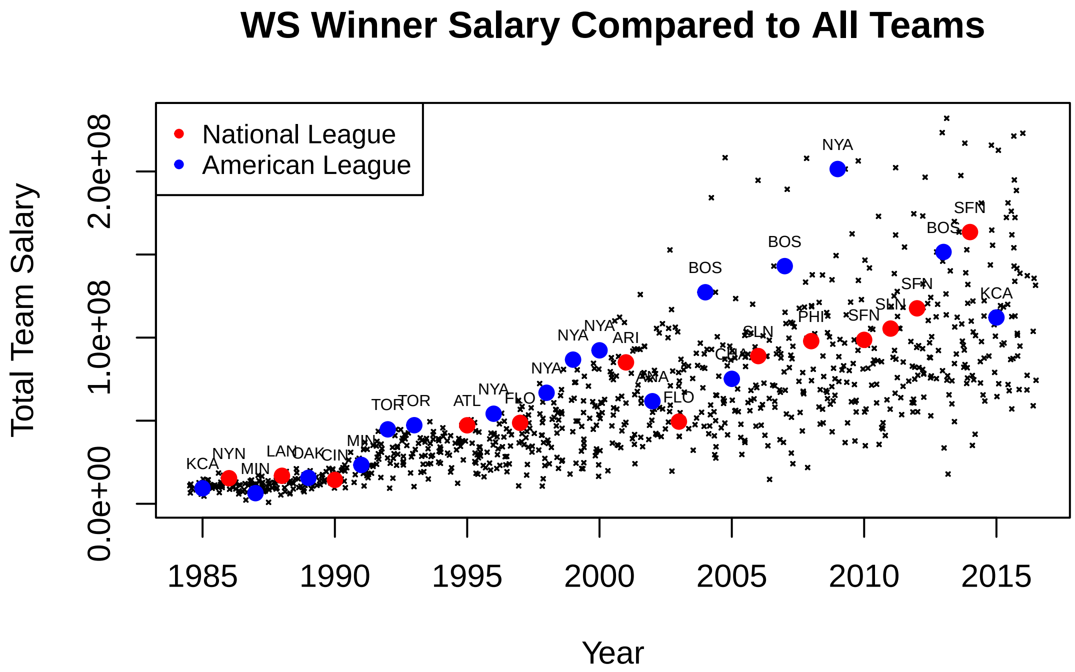


9.



10. Between the years 1985 and 2016, the World Series winner has been from the American League 17 times and from the National League 13 times. The average salary of the World Series winning teams from the American League is 77,596,664 dollars. The average salary of the World Series winning teams from the National League is 73,004,014 dollars.

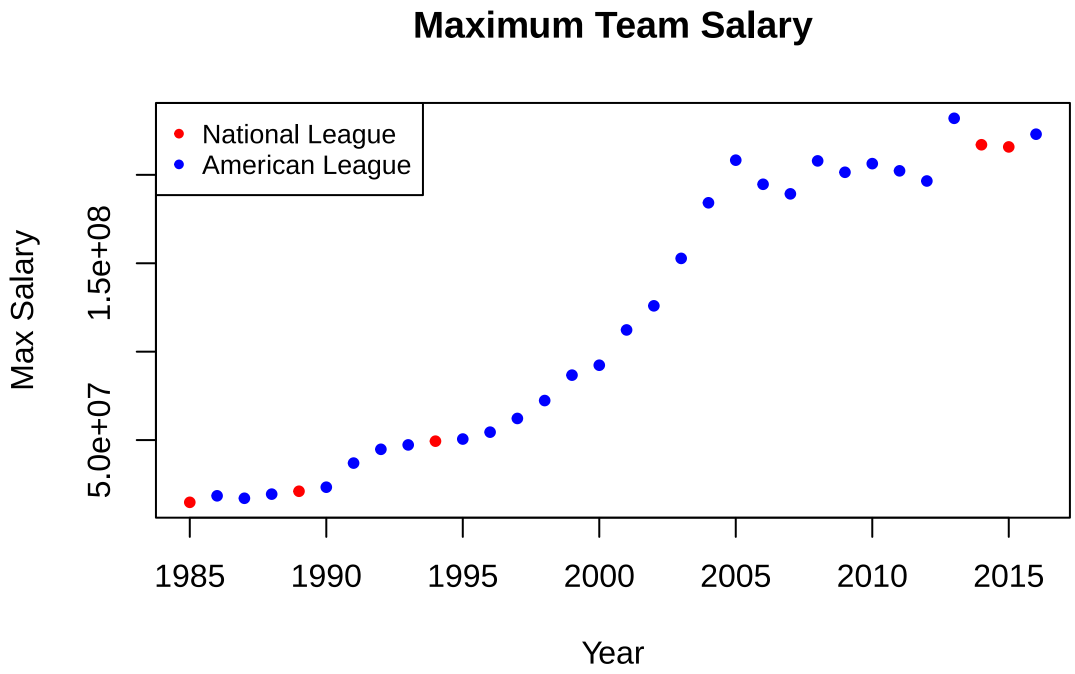
11.



The maximum, as well as the range of salaries in the MLB, has increased over time. Similarly, the salary of the World Series winner generally increased over time and fell within the upper half of all team salaries. Up until about 2000, the salary of the team that won the world series tended to be very close to the maximum team salary for that year. In the remaining years, this pattern was not as consistent and there were some years where there were several teams with higher team salaries than the world series winner. This could suggest that after a certain threshold, higher salaries don't necessarily translate to world series success.

12.

```
maxSalaries <- dbGetQuery(con, '
SELECT sub.year, sub.league, MAX(sub.salary_sum) as max_salary
FROM (
      SELECT SUM(Salaries."salary") AS salary_sum,
             Salaries."teamID" AS team,
             Salaries."lgID" AS league,
             Salaries."yearID" AS year
      FROM Salaries
      GROUP BY year, team
) sub
GROUP BY year
')
```

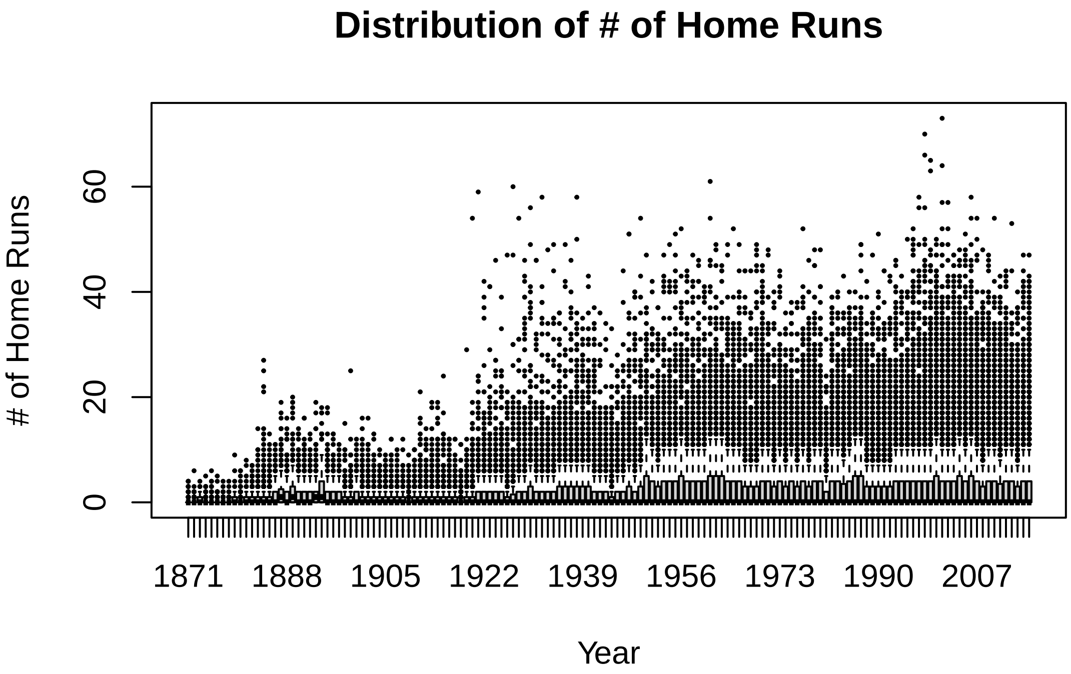


The maximum team salary has consistently increased each year between 1985 to 2005, with the greatest year-to-year increases occurring in the decade between 1995 and 2005. Since then it has followed a positive trend, but has not increased every year. It appears that the maximum salary in the American League tends to be higher than the maximum salary in the National League because most years the team that pays the greatest amount in salary is from the American League.

```
allstarplayers <- dbGetQuery(con, '
SELECT AllstarFull."yearID" AS year,
       AllstarFull."teamID" AS teamID,
      COUNT(AllstarFull."playerID") as n
FROM AllstarFull
JOIN SeriesPost
ON  SeriesPost."teamIDwinner" = AllstarFull."teamID" AND SeriesPost."yearid" = AllstarFull."yearID" AND
    SeriesPost.round = "WS"
GROUP BY year
ORDER by n DESC
')
```

The 5 years with the most All Star players on the winning team were 1939, 1947, 1960, 1961, and 1962.

14.

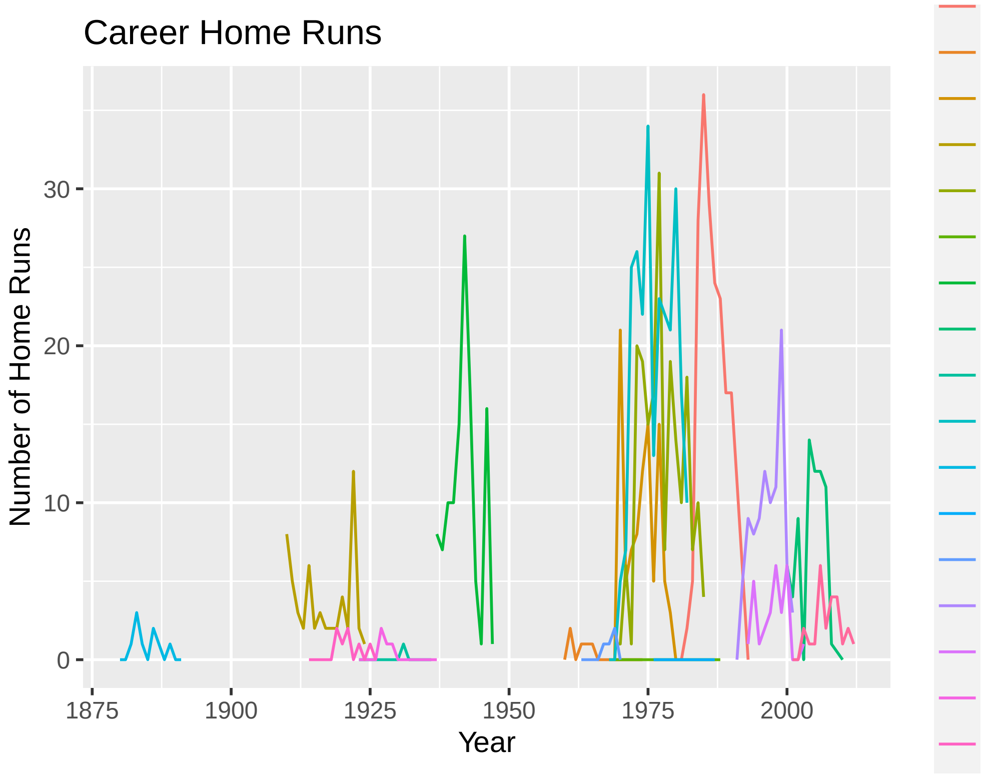


The distribution of the number of Home Runs has changed over time. The range has greatly increased, because while the the minimum has obviously stayed at 0, individual players are hitting more home runs each year, leading to a wider distribution.

The period between 1900 to 1920, has been referred to as the “Dead Ball Era” because games were typically low scoring and there was a greater emphasis placed on stolen bases and hit-and-run than on home runs. The boxplot indicates that even the players hitting the highest number of home runs were not hitting more than 20 per season. After the “Dead Ball Era”, ending around 1920, the median and maximum number of home runs increased.

Throughout the last 95 years, the IQR and median have stayed around the same values, but the upper outliers have continued to increase over time, causing the range to also become larger over time.

```
df27 <- dbGetQuery(con,'
      SELECT sub1.player,
              sub2.total_hr,
              sub2.year AS year
      FROM (
          SELECT COUNT(Batting."yearID") AS num_years,
           Batting."playerID" AS player
      FROM Batting
      GROUP BY player
      HAVING num_years > 10
      ) sub1
      JOIN (
          SELECT Batting."yearID" AS year,
           SUM(Batting."HR") AS total_hr,
           Batting."playerID" AS player
    FROM Batting
    GROUP BY year, player
    ORDER BY player
      ) sub2
      ON sub1.player = sub2.player
      ORDER BY sub1.player
    ')
```



Players tend to hit progressively more home runs throughout the beginning of their career and then peak between typically around halfway to three quarters of the way through their career. Then, the number of home runs tends to decline towards the end of individual players' careers.

15. Personal Question

In 2010, players from which alma mater were making up the largest share of salaries being paid in the MLB? In other words, if all MLB players were to donate their entire salary in 2010 to the college they went to, which school would receive the largest donation?

```
school_sals <- dbGetQuery(con, '
    SELECT  SUM(sub2.total_salary) as total_school_salary,

            sub1."schoolID"
    FROM (
      SELECT *
      FROM CollegePlaying
      GROUP BY CollegePlaying."playerid"
    ) sub1
    JOIN (
          SELECT Salaries."playerID" as player,
                 SUM(Salaries."salary") as total_salary,
                 Salaries."yearID" as year
           FROM Salaries
           WHERE year = 2010
           GROUP by player
    ) sub2
    ON sub1."playerID" = sub2.player
    GROUP BY schoolID
    ORDER BY total_school_salary DESC
    ')

head(school_sals)
```

```
##   total_school_salary  schoolID
## 1            32974974      ucsb
## 2            27870000    gatech
## 3            22535714      ucla
## 4            20830000   txsjjcn
## 5            20650000 floridast
## 6            20251000     utarl
```

In 2010, players from University of California, Santa Barbara were earning the greatest amount of money in the MLB, compared to other groups of players that went to the same school. Players that went there earned a total of $32,974,974 in 2010. Other groups of players with the same alma mater making up a large share of the salaries being paid in the MLB were Georgia Tech, UCLA, and San Jacinto College, North Campus.