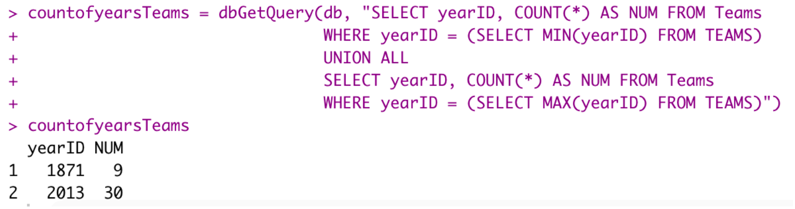
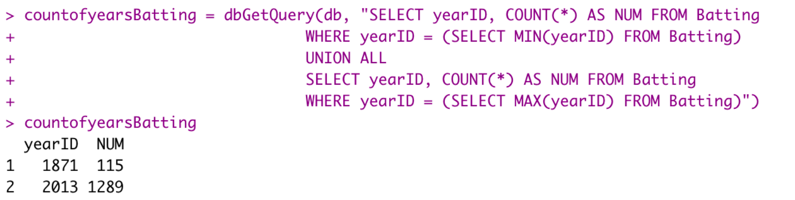
Jong Chan Park
SID # : 913274897
Assignment 3 Baseball
STA141B Fall 2020

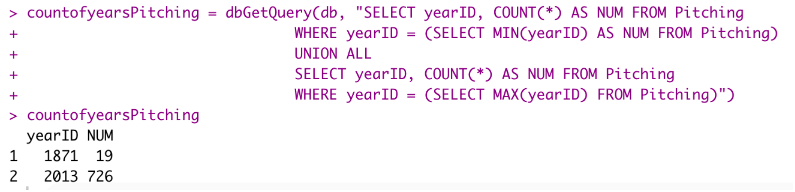
## 1. What years does the data cover? Are there data for each of these years?



```
> countofyearsTeams = dbGetQuery(db, "SELECT yearID, COUNT(*) AS NUM FROM Teams
+                                 WHERE yearID = (SELECT MIN(yearID) FROM TEAMS)
+                                 UNION ALL
+                                 SELECT yearID, COUNT(*) AS NUM FROM Teams
+                                 WHERE yearID = (SELECT MAX(yearID) FROM TEAMS)")
> countofyearsTeams
  yearID NUM
1   1871   9
2   2013  30
```



```
> countofyearsBatting = dbGetQuery(db, "SELECT yearID, COUNT(*) AS NUM FROM Batting
+                                 WHERE yearID = (SELECT MIN(yearID) FROM Batting)
+                                 UNION ALL
+                                 SELECT yearID, COUNT(*) AS NUM FROM Batting
+                                 WHERE yearID = (SELECT MAX(yearID) FROM Batting)")
> countofyearsBatting
  yearID  NUM
1   1871  115
2   2013 1289
```



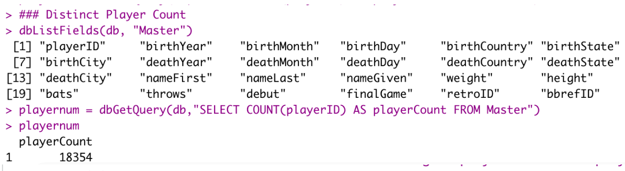
```
> countofyearsPitching = dbGetQuery(db, "SELECT yearID, COUNT(*) AS NUM FROM Pitching
+                                 WHERE yearID = (SELECT MIN(yearID) AS NUM FROM Pitching)
+                                 UNION ALL
+                                 SELECT yearID, COUNT(*) AS NUM FROM Pitching
+                                 WHERE yearID = (SELECT MAX(yearID) FROM Pitching)")
> countofyearsPitching
  yearID NUM
1   1871  19
2   2013 726
```

First when I did look at which tables contains year information. First table that I looked at was Master Table. However Master table had only the birth year of players, which we can't assume about years that data covers. Therefore I then looked at the Teams table. I thought Teams table is the table that I should look for since Teams are needed in order to league to be present. So I extracted the yearIDs and numbers of information with corresponding yearID where yearID is maximum and minimum value from Team table. Then I used UNION ALL to attach two tables together.
I then applied the code to Batting and Pitching tables just to confirm the maximum and minimum year for data base.

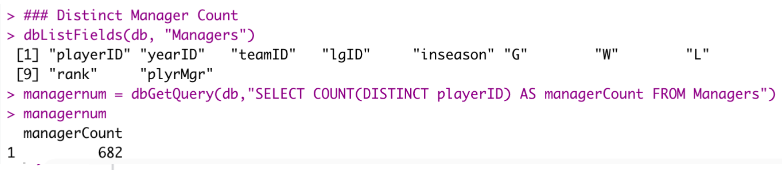
We can see that the maximum and minimum years are same for all three tables. We can see that the data covers from1871 to 2013. Also one thing that we found is that as year increase, the number of information increases. We can assume that as time goes, the league expanded.

## 2. How many (unique) people are included in the database? How many are players, managers, etc?

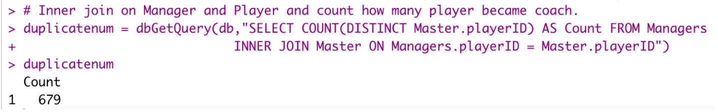
When I first looked at the question, the first variable that I thought of was playerID from Master table. Since Master table contains the all distinct player's information, I thought it would be great to count numbers of unique people in the database. However Master table only contains the number of players. So I used the manager table to find the number of managers. Unlike Master table, there were some duplicates on manager's playerID. So I used the Distinct function  to count them. Then I joined the Master table and Manager table on where playerIDs are present on both tables to find players who became managers. From Master table, I got 18354 distinct players, and from Manager table, I got 682 managers total. From inner joined table, I got 679 duplicates. From the calculation, I got 18357 unique people from data base. One thing that I found from the number is that there were only 3 managers that were not player.



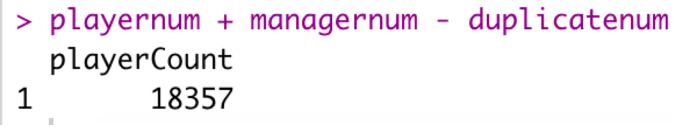
```
> ### Distinct Player Count
> dbListFields(db, "Master")
 [1] "playerID"     "birthYear"    "birthMonth"   "birthDay"     "birthCountry" "birthState"
 [7] "birthCity"    "deathYear"    "deathMonth"   "deathDay"     "deathCountry" "deathState"
[13] "deathCity"    "nameFirst"    "nameLast"     "nameGiven"    "weight"       "height"
[19] "bats"         "throws"       "debut"        "finalGame"    "retroID"      "bbrefID"
> playernum = dbGetQuery(db,"SELECT COUNT(playerID) AS playerCount FROM Master")
> playernum
  playerCount
1       18354
```



```
> ### Distinct Manager Count
> dbListFields(db, "Managers")
 [1] "playerID" "yearID"   "teamID"   "lgID"     "inseason" "G"        "W"        "L"
 [9] "rank"     "plyrMgr"
> managernum = dbGetQuery(db,"SELECT COUNT(DISTINCT playerID) AS managerCount FROM Managers")
> managernum
  managerCount
1          682
```



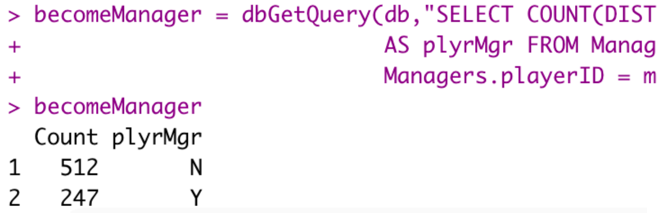
```
> # Inner join on Manager and Player and count how many player became coach.
> duplicatenum = dbGetQuery(db,"SELECT COUNT(DISTINCT Master.playerID) AS Count FROM Managers
+                           INNER JOIN Master ON Managers.playerID = Master.playerID")
> duplicatenum
  Count
1   679
```



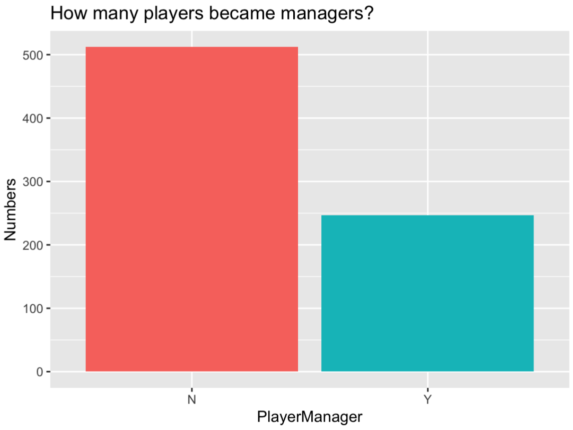
```
> playernum + managernum - duplicatenum
  playerCount
1       18357
```

### 3. How many players became managers?

This question is kind of related to previous question. In order to get number of players that became the manager, I first looked at the plyrMgr variable from Managers table. I thought the plyrMgr columns are representing the player who became manager. However from the piazza post 245, I found out that they are not. However I thought that plyrMgr with Y value is still counts towards players who became managers. So I decided to count the numbers of player who became managers with plyrMgr value. So I inner joined the Managers and Master tables once again and count the number of distinct playerID with grouping by plyrMgr.



```
> becomeManager = dbGetQuery(db,"SELECT COUNT(DISTINCT Master.playerID) AS Count, Managers.plyrMgr
+                            AS plyrMgr FROM Managers INNER JOIN Master ON
+                            Managers.playerID = master.playerID GROUP by Managers.plyrMgr")
> becomeManager
  Count plyrMgr
1   512       N
2   247       Y
```

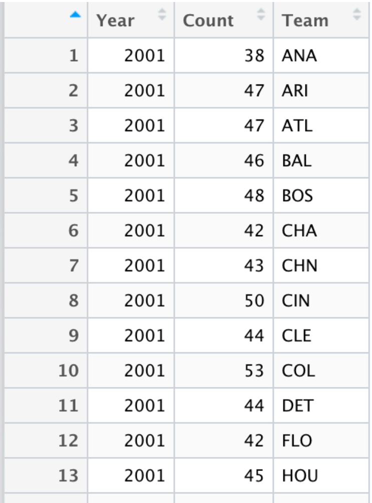


We found that total numbers of player who became manager is 759, 247 with plyrMgr and 512 not plyrMgr.

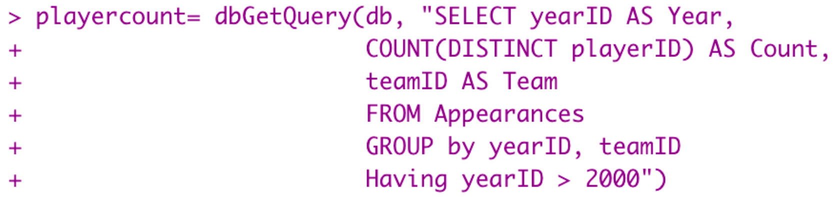
We found out from the table and plot that there are nearly half more managers who are not plyrMgr.

### 4. How many players are there in each year, from 2000 to 2013? Do all teams have the same number of players?

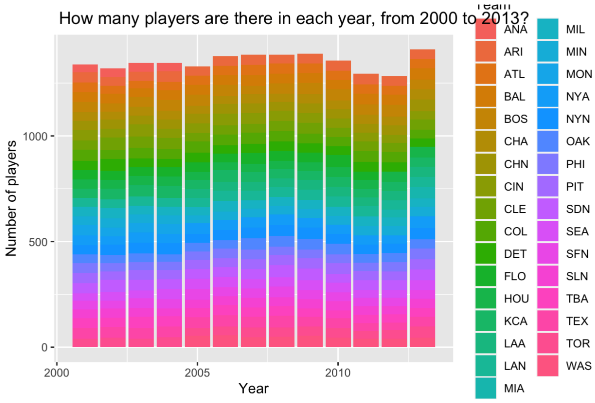
When I first looked at the question, I thought I should use the appearance table, not Master table. Master table contains the all player's information however it doesn't contain player's teamID or yearID. Therefore I decided to use the Appearance table since it contains both TeamID of that year. So from SQL, I selected year and count on distinct player ID and group them by yearID and teamID. Since from question 1, we saw that data only covers until 2013, I decided to just use one having statement. Then I got the table on the right Then I plotted them using ggplot.

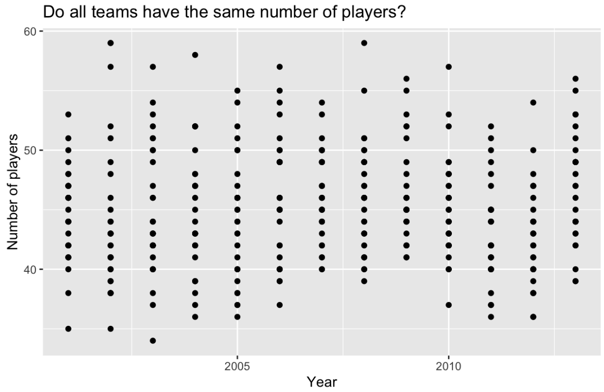


| ▲ | Year | Count | Team |
|---|---|---|---|
| 1 | 2001 | 38 | ANA |
| 2 | 2001 | 47 | ARI |
| 3 | 2001 | 47 | ATL |
| 4 | 2001 | 46 | BAL |
| 5 | 2001 | 48 | BOS |
| 6 | 2001 | 42 | CHA |
| 7 | 2001 | 43 | CHN |
| 8 | 2001 | 50 | CIN |
| 9 | 2001 | 44 | CLE |
| 10 | 2001 | 53 | COL |
| 11 | 2001 | 44 | DET |
| 12 | 2001 | 42 | FLO |
| 13 | 2001 | 45 | HOU |



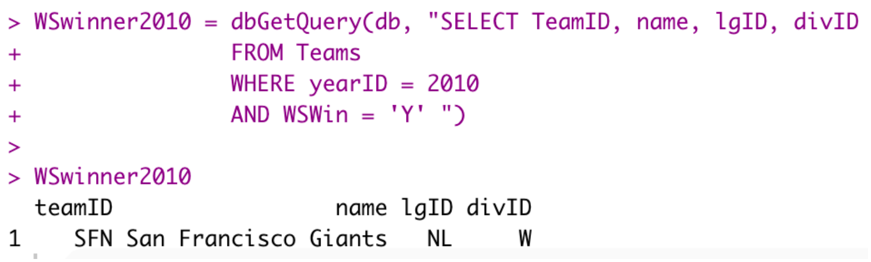
```
> playercount= dbGetQuery(db, "SELECT yearID AS Year,
+                         COUNT(DISTINCT playerID) AS Count,
+                         teamID AS Team
+                         FROM Appearances
+                         GROUP by yearID, teamID
+                         Having yearID > 2000")
```





We can see from bar chart that there is not much difference in number of player by year. Since 2010, the number of players decreased but in 2013, the number of player increased to maximum value. From the dot plot, we can see that the number of players in each team differ. Although there are few teams that have nearly 60 players in a team and few teams that have below 40 players, most of teams have 40 to 50 players in a team.

5. **What team won the World Series in 2010? Include the name of the team, the league and division.**

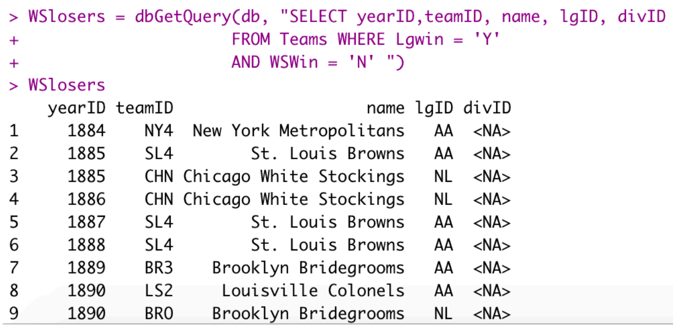


```
> WSwinner2010 = dbGetQuery(db, "SELECT TeamID, name, lgID, divID
+                 FROM Teams
+                 WHERE yearID = 2010
+                 AND WSWin = 'Y' ")
>
> WSwinner2010
  teamID                 name lgID divID
1    SFN San Francisco Giants   NL     W
```

After reading the question, I first looked at SeriesPost table and decided to join it with Team table since it has league and division ID. However then I found that Team table contains WSwin variable which allowed me to use Team table instead.

I selected the teamID, name, laid, and divID where yearID is 2010 and won world series. The winner of World Series in 2010 was San Francisco Giants from Western division of National League

6. **What team lost the World Series each year? Again, include the name of the team, league and division**

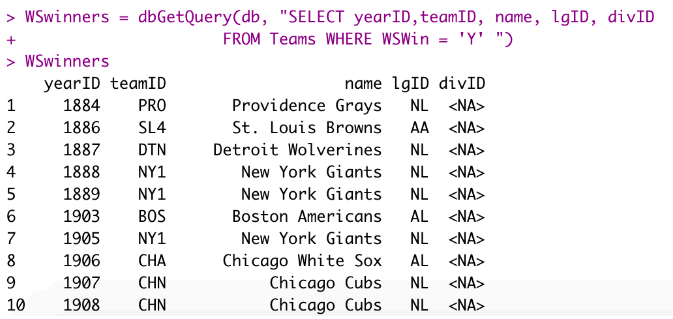


```
> WSlosers = dbGetQuery(db, "SELECT yearID,teamID, name, lgID, divID
+                     FROM Teams WHERE Lgwin = 'Y'
+                     AND WSWin = 'N' ")
> WSlosers
  yearID teamID                 name lgID divID
1   1884    NY4  New York Metropolitans   AA  <NA>
2   1885    SL4        St. Louis Browns   AA  <NA>
3   1885    CHN Chicago White Stockings   NL  <NA>
4   1886    CHN Chicago White Stockings   NL  <NA>
5   1887    SL4        St. Louis Browns   AA  <NA>
6   1888    SL4        St. Louis Browns   AA  <NA>
7   1889    BR3     Brooklyn Bridegrooms   AA  <NA>
8   1890    LS2      Louisville Colonels   AA  <NA>
9   1890    BRO     Brooklyn Bridegrooms   NL  <NA>
```

Just like previous question, I used the Teams table. However unlike previous question, it was which team LOST the World Series each year. So I selected the same variables but have different condition. First we didn't have a year condition so I deleted the year condition, but added a new where team that won the league but not World Series.

One thing that I learn from table is that teams didn't have division before 1969. It was after 1968 when divisionID is created.

7. **Compute the table of World Series winners for all years, again with the name of the team, league and division.**



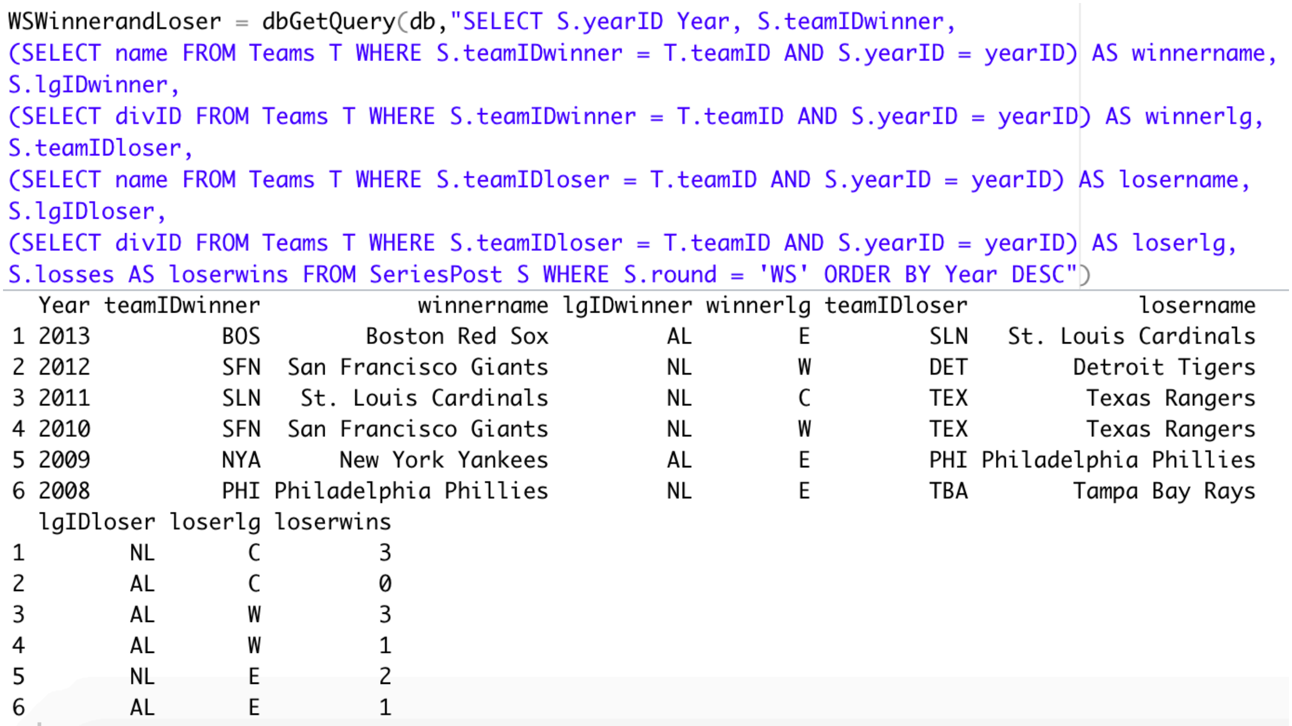
```
> WSwinners = dbGetQuery(db, "SELECT yearID,teamID, name, lgID, divID
+                     FROM Teams WHERE WSWin = 'Y' ")
> WSwinners
   yearID teamID                name lgID divID
1    1884    PRO     Providence Grays   NL  <NA>
2    1886    SL4     St. Louis Browns   AA  <NA>
3    1887    DTN   Detroit Wolverines   NL  <NA>
4    1888    NY1      New York Giants   NL  <NA>
5    1889    NY1      New York Giants   NL  <NA>
6    1903    BOS     Boston Americans   AL  <NA>
7    1905    NY1      New York Giants   NL  <NA>
8    1906    CHA    Chicago White Sox   AL  <NA>
9    1907    CHN         Chicago Cubs   NL  <NA>
10   1908    CHN         Chicago Cubs   NL  <NA>
```

Just like question 5 and 6, I selected the yearID, teamID, name, lgID, and divID but with different condition.

I deleted the Lgwin condition and changed the WSWin condition to 'Y' so that it contains the only World Series winner's information.

**8. Compute the table that has both the winner and runner-up for the World Series in each tuple/row for all years, again with the name of the team, league and division, and also the number games the losing team won in the series.**

This question seemed very simple when I first read the question because It was similar to what I've done in previous few questions. However, this question was most time consuming. First I decided to inner join SeriesPost and Team tables because we need the number of win that losers had. However I struggled on which variable I should join the tables on. After struggling for few hours, I found the concept of subquery where we can use query in select statement. So I applied to my code below



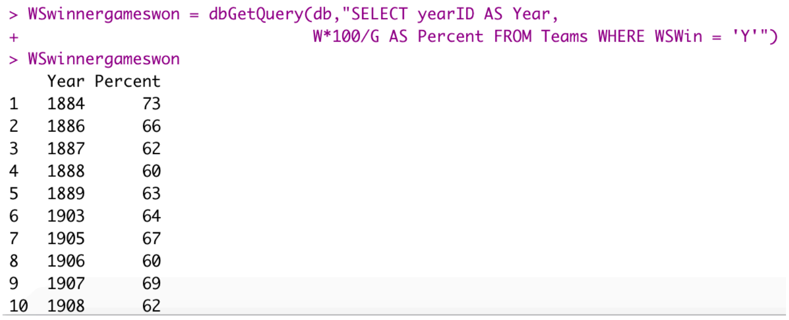
```
WSWinnerandLoser = dbGetQuery(db,"SELECT S.yearID Year, S.teamIDwinner,
(SELECT name FROM Teams T WHERE S.teamIDwinner = T.teamID AND S.yearID = yearID) AS winnername,
S.lgIDwinner,
(SELECT divID FROM Teams T WHERE S.teamIDwinner = T.teamID AND S.yearID = yearID) AS winnerlg,
S.teamIDloser,
(SELECT name FROM Teams T WHERE S.teamIDloser = T.teamID AND S.yearID = yearID) AS losername,
S.lgIDloser,
(SELECT divID FROM Teams T WHERE S.teamIDloser = T.teamID AND S.yearID = yearID) AS loserlg,
S.losses AS loserwins FROM SeriesPost S WHERE S.round = 'WS' ORDER BY Year DESC")
```

```
  Year teamIDwinner            winnername lgIDwinner winnerlg teamIDloser            losername
1 2013          BOS        Boston Red Sox         AL        E         SLN  St. Louis Cardinals
2 2012          SFN  San Francisco Giants         NL        W         DET       Detroit Tigers
3 2011          SLN   St. Louis Cardinals         NL        C         TEX        Texas Rangers
4 2010          SFN  San Francisco Giants         NL        W         TEX        Texas Rangers
5 2009          NYA      New York Yankees         AL        E         PHI Philadelphia Phillies
6 2008          PHI Philadelphia Phillies         NL        E         TBA       Tampa Bay Rays
  lgIDloser loserlg loserwins
1        NL       C         3
2        AL       C         0
3        AL       W         3
4        AL       W         1
5        NL       E         2
6        AL       E         1
```

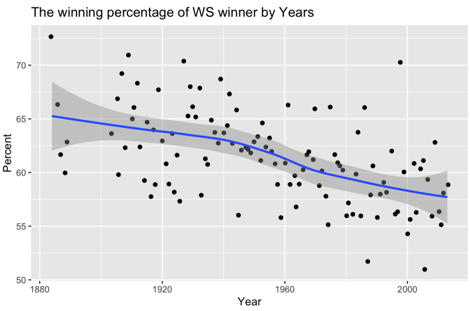
Then I finally got the table I wanted. One thing that I realize is that the post in 1800s, have more than 4 wins by losers. We can see that the post series for 1800s were kind of different system from post series today.

**9. Do you see a relationship between the number of games won in a season and winning the World Series?**

First when I read the question, I thought about how to show relationship between number of games won in a season and wining world series. I first decided to use linear regression but when I thought about it, I thought it was better to show the winning percentage of season and compare it by years.



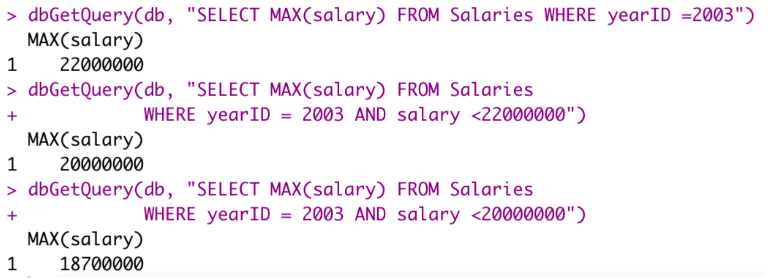
```
> WSwinnergameswon = dbGetQuery(db,"SELECT yearID AS Year,
+                                W*100/G AS Percent FROM Teams WHERE WSWin = 'Y'")
> WSwinnergameswon
   Year Percent
1  1884      73
2  1886      66
3  1887      62
4  1888      60
5  1889      63
6  1903      64
7  1905      67
8  1906      60
9  1907      69
10 1908      62
```



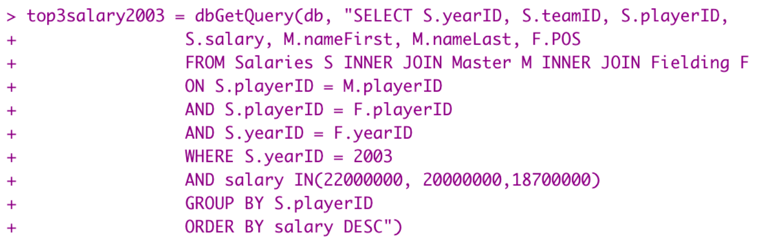
The winning percentage of WS winner by Years

We can see that in early years, the high winning percentage in season won the world series compare to world series winner in later years. However most of teams with World Series win have at least 50% of winning percentage.

**10. In 2003, what were the three highest salaries? (We refer here to unique salaries, i.e., there maybe several players getting the exact same amount.)Find the players who got any of these 3 salaries with all of their details?**

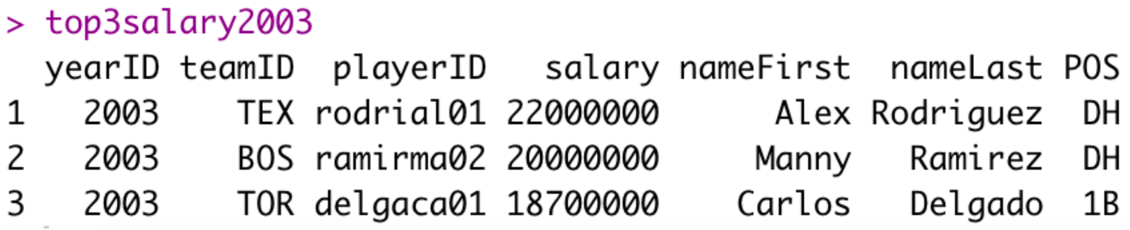


```
> dbGetQuery(db, "SELECT MAX(salary) FROM Salaries WHERE yearID =2003")
  MAX(salary)
1    22000000
> dbGetQuery(db, "SELECT MAX(salary) FROM Salaries
+            WHERE yearID = 2003 AND salary <22000000")
  MAX(salary)
1    20000000
> dbGetQuery(db, "SELECT MAX(salary) FROM Salaries
+            WHERE yearID = 2003 AND salary <20000000")
  MAX(salary)
1    18700000
```



```
> top3salary2003 = dbGetQuery(db, "SELECT S.yearID, S.teamID, S.playerID,
+                  S.salary, M.nameFirst, M.nameLast, F.POS
+                  FROM Salaries S INNER JOIN Master M INNER JOIN Fielding F
+                  ON S.playerID = M.playerID
+                  AND S.playerID = F.playerID
+                  AND S.yearID = F.yearID
+                  WHERE S.yearID = 2003
+                  AND salary IN(22000000, 20000000,18700000)
+                  GROUP BY S.playerID
+                  ORDER BY salary DESC")
```

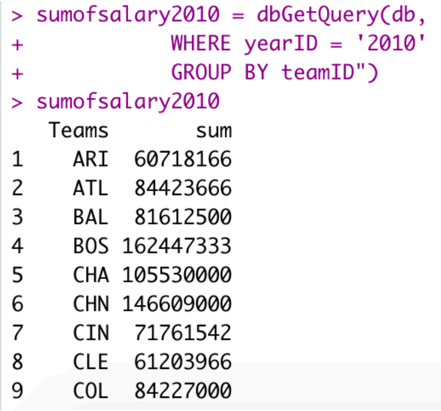
First when I read the question, I thought we need the Salaries table and Master table. One more thing that I wanted to add was the position. I wanted to know which player got highest salary and which position he had. So I used fielding table to know the position. First, I got Top 3 salaries with yearID = 2003, which were 22000000, 20000000, 18700000. Then I decided to apply in the code as below. I joined the three tables on playerID of each and set condition where yearID is 2003 and salary is in top 3 salary.  We can see from the result that top 2 salary in 2003 are the Designated Hitter, which can assume that the Designated Hitters get paid more.



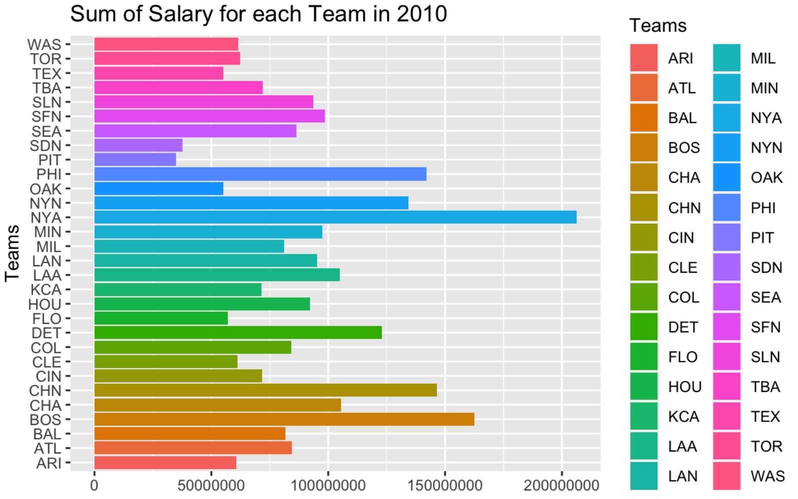
```
> top3salary2003
  yearID teamID  playerID   salary nameFirst  nameLast POS
1   2003    TEX rodrial01 22000000      Alex Rodriguez  DH
2   2003    BOS ramirma02 20000000     Manny   Ramirez  DH
3   2003    TOR delgaca01 18700000    Carlos   Delgado  1B
```

**11. For 2010, compute the total payroll of each of the different teams. Next compute the team payrolls for all years in the database for which we have salary information. Display these in a plot.**

Similar to previous question, I used the Salaries table. For total payroll of each different team in year 2010, I selected teamID, sum of salary and set condition where yearID is 2010. Then I used ggplot to plot



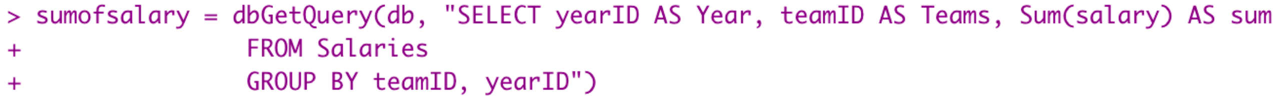
```
> sumofsalary2010 = dbGetQuery(db, "SELECT teamID AS Teams, Sum(salary) AS sum FROM Salaries
+                  WHERE yearID = '2010'
+                  GROUP BY teamID")
> sumofsalary2010
  Teams       sum
1   ARI  60718166
2   ATL  84423666
3   BAL  81612500
4   BOS 162447333
5   CHA 105530000
6   CHN 146609000
7   CIN  71761542
8   CLE  61203966
9   COL  84227000
```



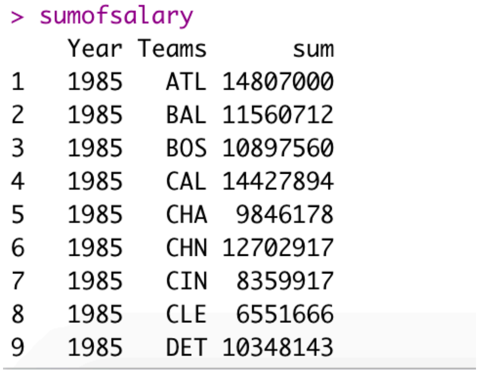
We can see from the graph that NYA (New York Yankees) spend the most salary in 2010. Also NYN (New York Mets) and

BOS (Boston Red Sox), PHI (Philadelphia Phillies) spent much money on salary.

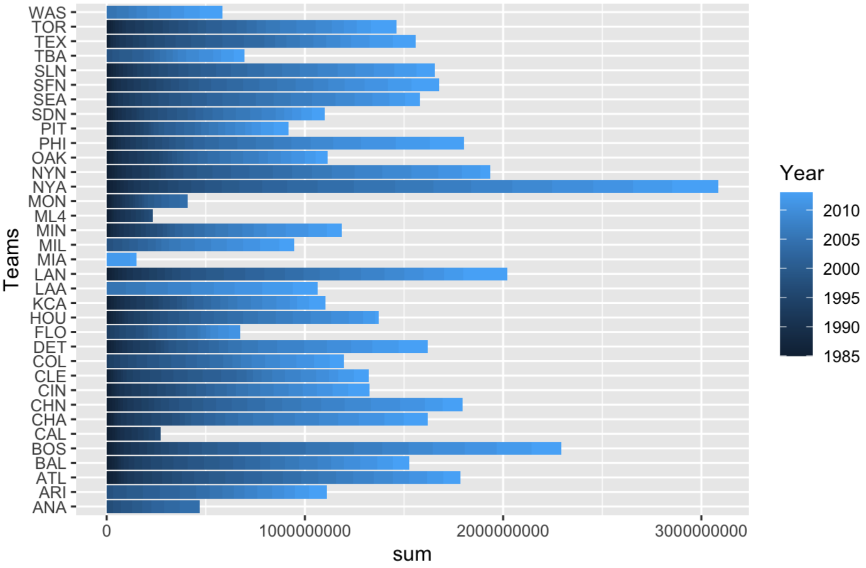
For each year, I’ve used same code except instead of setting yearID to 2010, I group by teamID and yearID.



```
> sumofsalary = dbGetQuery(db, "SELECT yearID AS Year, teamID AS Teams, Sum(salary) AS sum
+               FROM Salaries
+               GROUP BY teamID, yearID")
```



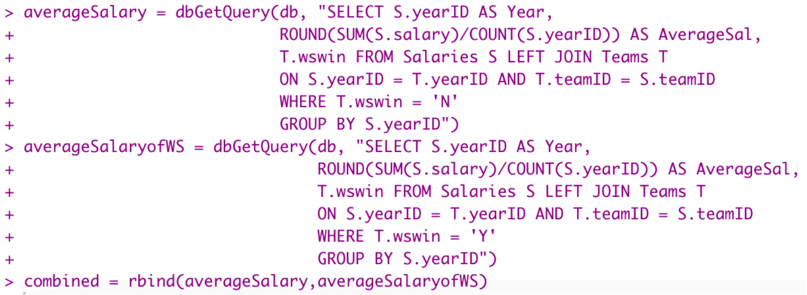
```
> sumofsalary
  Year Teams      sum
1 1985   ATL 14807000
2 1985   BAL 11560712
3 1985   BOS 10897560
4 1985   CAL 14427894
5 1985   CHA  9846178
6 1985   CHN 12702917
7 1985   CIN  8359917
8 1985   CLE  6551666
9 1985   DET 10348143
```



We again can see that NYA, BOS used most money on salary. We can assume that these teams use much money on salary.

**12. Explore the change in salary Over time. Use a plot. Identify the teams that won the world series or league on the plot. How does salary relate to winning the league and/or world series.**

Once I read the question, I decided to use bar chart to compare average salary of WS winning team and other teams. So I got tables of average salary of World Series winning team and all other teams, and combined two tables by using rbind function. . Then I plotted them with ggplot



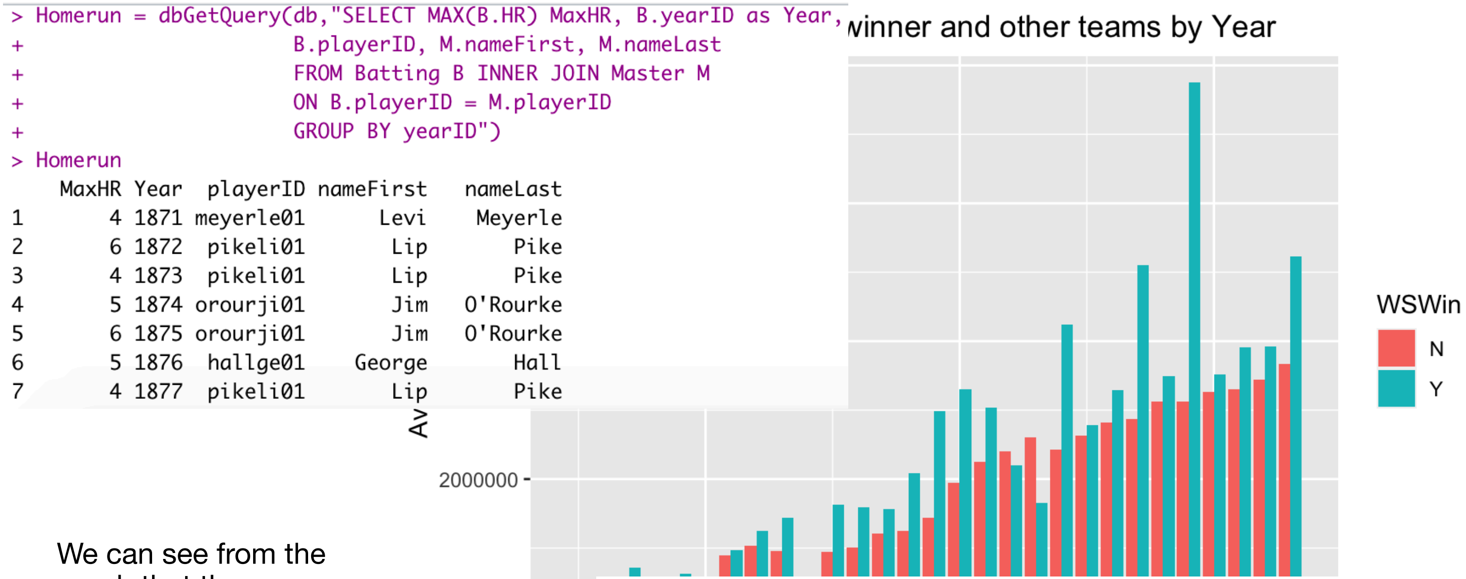
```
> averageSalary = dbGetQuery(db, "SELECT S.yearID AS Year,
+                           ROUND(SUM(S.salary)/COUNT(S.yearID)) AS AverageSal,
+                           T.wswin FROM Salaries S LEFT JOIN Teams T
+                           ON S.yearID = T.yearID AND T.teamID = S.teamID
+                           WHERE T.wswin = 'N'
+                           GROUP BY S.yearID")
> averageSalaryofWS = dbGetQuery(db, "SELECT S.yearID AS Year,
+                               ROUND(SUM(S.salary)/COUNT(S.yearID)) AS AverageSal,
+                               T.wswin FROM Salaries S LEFT JOIN Teams T
+                               ON S.yearID = T.yearID AND T.teamID = S.teamID
+                               WHERE T.wswin = 'Y'
+                               GROUP BY S.yearID")
> combined = rbind(averageSalary,averageSalaryofWS)
```

We can see from the plots that average salary increases as year increases.

Also for early years, the difference in average salary of World Series winning team and non winning teams didn’t differ to much. However as year increase, we can see that they starts to differ.

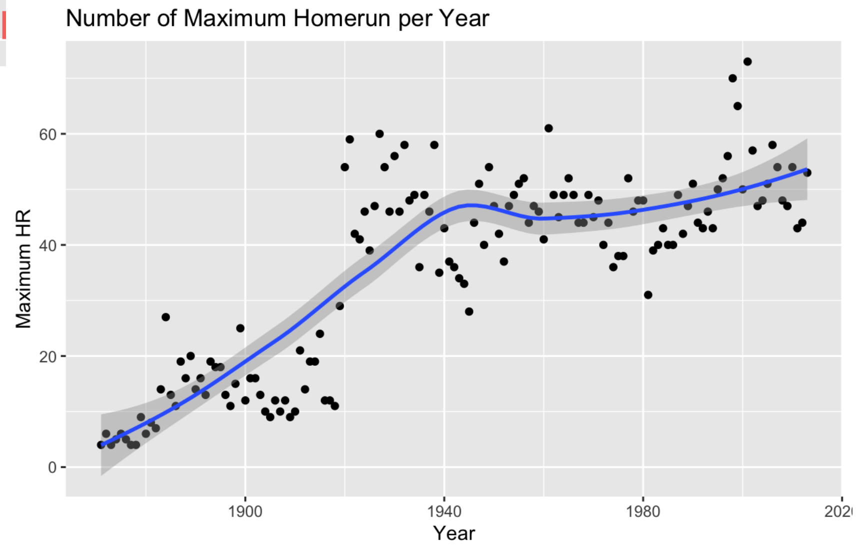
**13. Which player has hit the most home runs? Show the number per year.**

Since it’s related to Home run, I looked at the Batting table. Then I found the variable HR. So I decided to use Batting table. Then I joined the table with Master On playerID so that I can get their name.



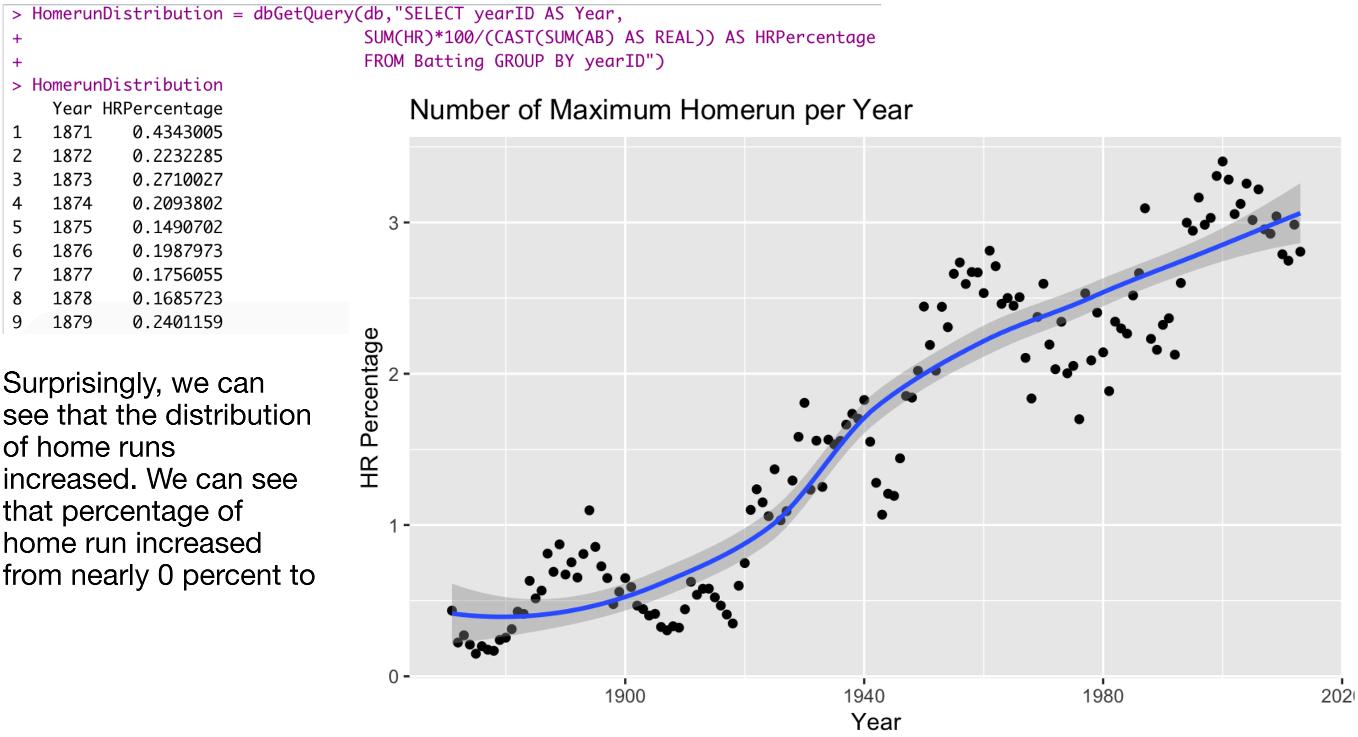
```
> Homerun = dbGetQuery(db,"SELECT MAX(B.HR) MaxHR, B.yearID as Year,
+                       B.playerID, M.nameFirst, M.nameLast
+                       FROM Batting B INNER JOIN Master M
+                       ON B.playerID = M.playerID
+                       GROUP BY yearID")
> Homerun
   MaxHR Year  playerID nameFirst   nameLast
1      4 1871 meyerle01      Levi    Meyerle
2      6 1872  pikeli01       Lip       Pike
3      4 1873  pikeli01       Lip       Pike
4      5 1874 orourji01       Jim   O'Rourke
5      6 1875 orourji01       Jim   O'Rourke
6      5 1876  hallge01    George       Hall
7      4 1877  pikeli01       Lip       Pike
```

We can see from the graph that the maximum Home runs increases as year increases. I think this is due to increase in number of games.



### 14. Has the distribution of home runs for players increased over the years?

From previous question, we found that the maximum Home runs increased as year increases. I thought it was due to the increase in number of games. So I'll take a look at the distribution of home runs over years. So I wanted to take a percentage of Home runs. So I divided the Homerun by at bat variable. It is similar to the previous question but without joining the tables and select Home run percentage instead.



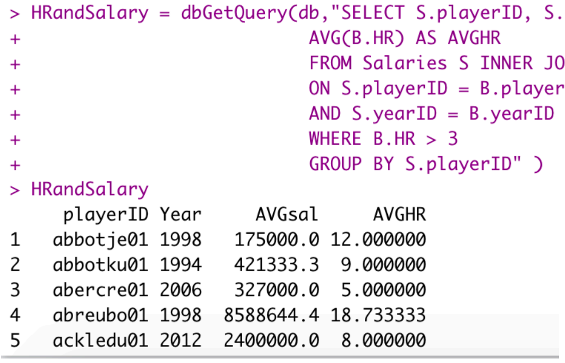
```
> HomerunDistribution = dbGetQuery(db,"SELECT yearID AS Year,
+                                  SUM(HR)*100/(CAST(SUM(AB) AS REAL)) AS HRPercentage
+                                  FROM Batting GROUP BY yearID")
> HomerunDistribution
    Year HRPercentage
1   1871    0.4343005
2   1872    0.2232285
3   1873    0.2710027
4   1874    0.2093802
5   1875    0.1490702
6   1876    0.1987973
7   1877    0.1756055
8   1878    0.1685723
9   1879    0.2401159
```

Surprisingly, we can see that the distribution of home runs increased. We can see that percentage of home run increased from nearly 0 percent to

3 percent. We may assume that batter's ability to make home run increased over time.

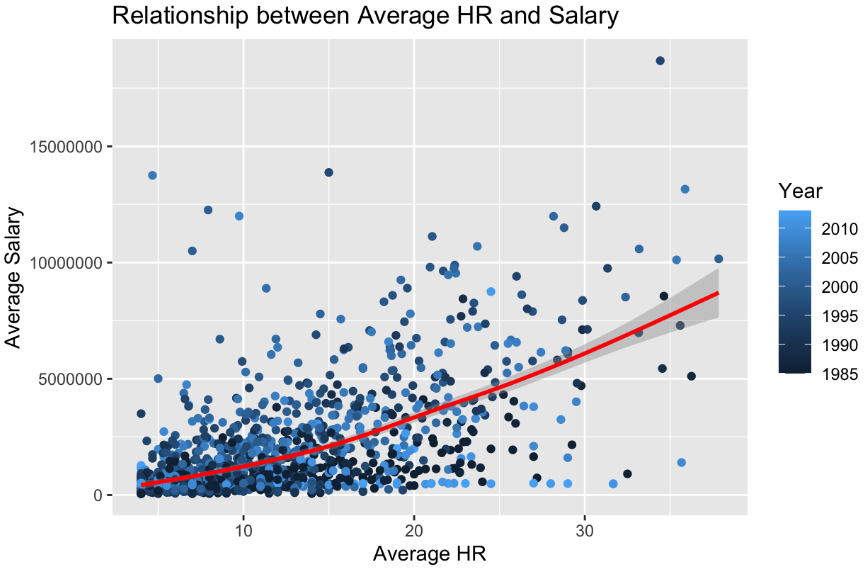
15. Do players who hit more home runs receive higher salaries?

In order to compare the number of Home runs and salary, I had to join the Batting table and Salaries table. So I joined them on playerID and yearID because we have to compare them with home runs and salary on same year. I set the condition that HR has to be greater than 3 since there are many pitchers that have few home runs.

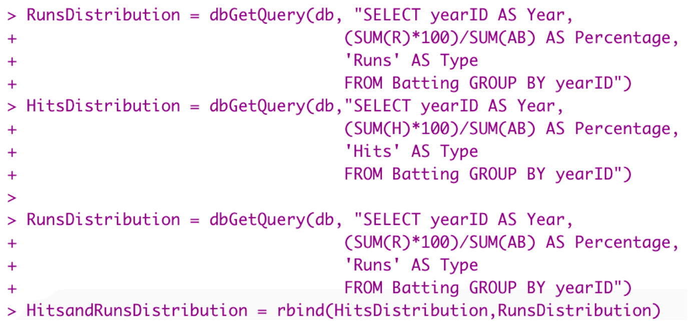


```
> HRandSalary = dbGetQuery(db,"SELECT S.playerID, S.yearID as Year, AVG(S.Salary) As AVGsal,
+                           AVG(B.HR) AS AVGHR
+                           FROM Salaries S INNER JOIN Batting B
+                           ON S.playerID = B.playerID
+                           AND S.yearID = B.yearID
+                           WHERE B.HR > 3
+                           GROUP BY S.playerID" )
> HRandSalary
   playerID Year     AVGsal     AVGHR
1  abbotje01 1998   175000.0 12.000000
2  abbotku01 1994   421333.3  9.000000
3  abercre01 2006   327000.0  5.000000
4  abreubo01 1998  8588644.4 18.733333
5  ackledu01 2012  2400000.0  8.000000
```

We can see the positive relationship with average salary and average home runs. We now know that players with home runs gets higher salary.



**16. What's the distribution of Runs and Hits over years?**

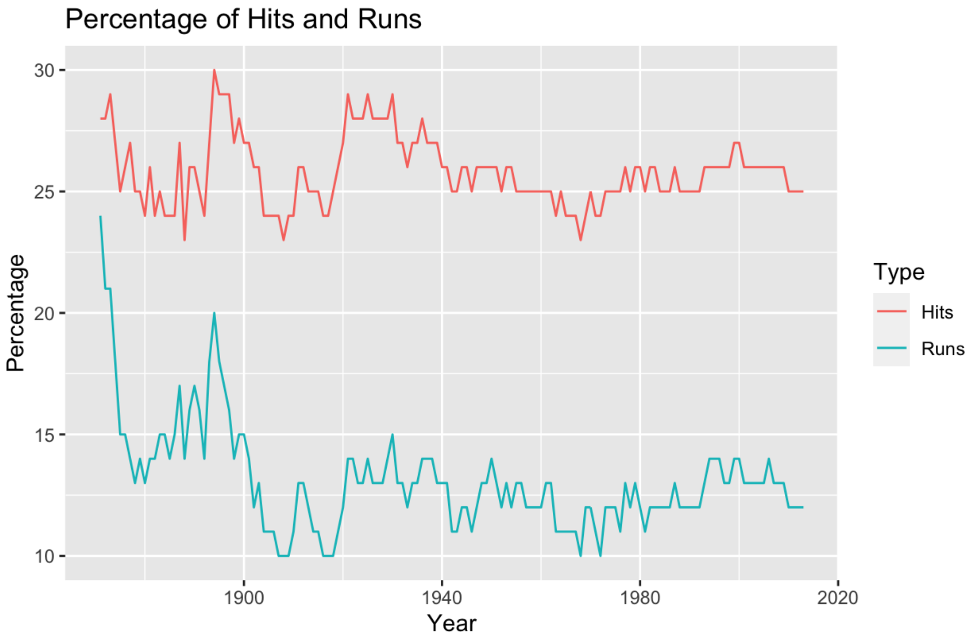


```
> RunsDistribution = dbGetQuery(db, "SELECT yearID AS Year,
+                               (SUM(R)*100)/SUM(AB) AS Percentage,
+                               'Runs' AS Type
+                               FROM Batting GROUP BY yearID")
> HitsDistribution = dbGetQuery(db,"SELECT yearID AS Year,
+                               (SUM(H)*100)/SUM(AB) AS Percentage,
+                               'Hits' AS Type
+                               FROM Batting GROUP BY yearID")
>
> RunsDistribution = dbGetQuery(db, "SELECT yearID AS Year,
+                               (SUM(R)*100)/SUM(AB) AS Percentage,
+                               'Runs' AS Type
+                               FROM Batting GROUP BY yearID")
> HitsandRunsDistribution = rbind(HitsDistribution,RunsDistribution)
```

Then I was curious about the Runs and Hits since we have went over the home runs. I decided to look for distribution of Runs and Hits. Hits and Runs informations are located in Batting table, I selected Hits and Runs percentage for each tables and combined them by using rbind, and plotted.

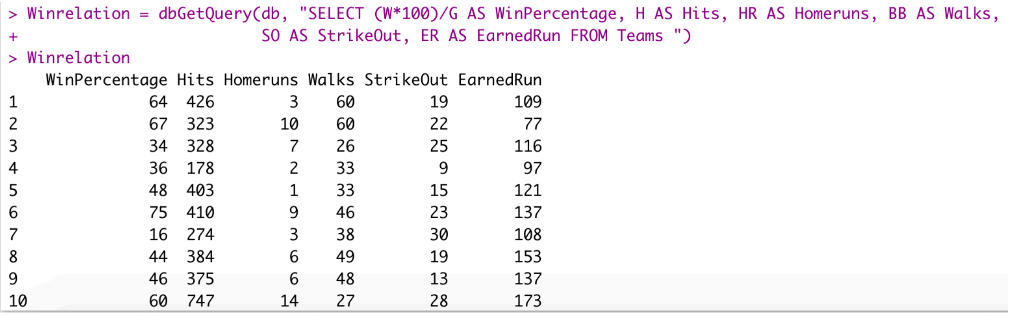
Although Hits percentage has ups and downs, it ranges between 20 to 30 percentage.

However the runs percentage, we can see that it has decreased as year increases. We can see that pitcher and defense's ability to stop runs has increased over years.

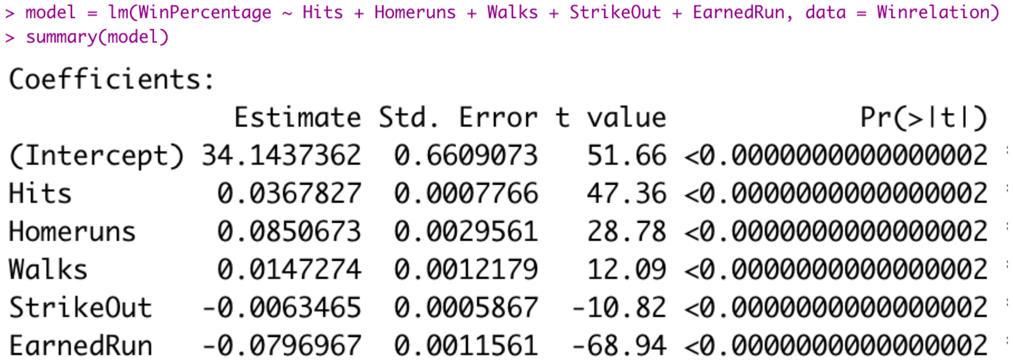


**17. How are wins related to hits, strikeouts, walks, homeruns and earned runs?**

For this question, I decided to use linear regression since there are multiple variables. For win, I decided to use winning percentage, number of hits, Home runs, walks, strike outs, and earned runs of each teams and use lm function to know correlation.



```
> Winrelation = dbGetQuery(db, "SELECT (W*100)/G AS WinPercentage, H AS Hits, HR AS Homeruns, BB AS Walks,
+                          SO AS StrikeOut, ER AS EarnedRun FROM Teams ")
> Winrelation
   WinPercentage Hits Homeruns Walks StrikeOut EarnedRun
1             64  426        3    60        19       109
2             67  323       10    60        22        77
3             34  328        7    26        25       116
4             36  178        2    33         9        97
5             48  403        1    33        15       121
6             75  410        9    46        23       137
7             16  274        3    38        30       108
8             44  384        6    49        19       153
9             46  375        6    48        13       137
10            60  747       14    27        28       173
```

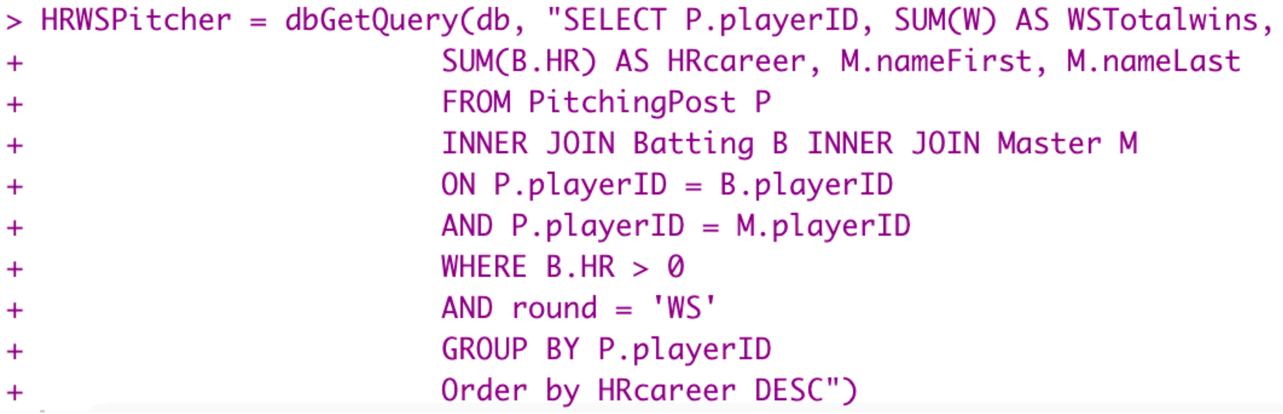


```
> model = lm(WinPercentage ~ Hits + Homeruns + Walks + StrikeOut + EarnedRun, data = Winrelation)
> summary(model)
Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept) 34.1437362  0.6609073   51.66 <0.0000000000000002
Hits         0.0367827  0.0007766   47.36 <0.0000000000000002
Homeruns     0.0850673  0.0029561   28.78 <0.0000000000000002
Walks        0.0147274  0.0012179   12.09 <0.0000000000000002
StrikeOut   -0.0063465  0.0005867  -10.82 <0.0000000000000002
EarnedRun   -0.0796967  0.0011561  -68.94 <0.0000000000000002
```

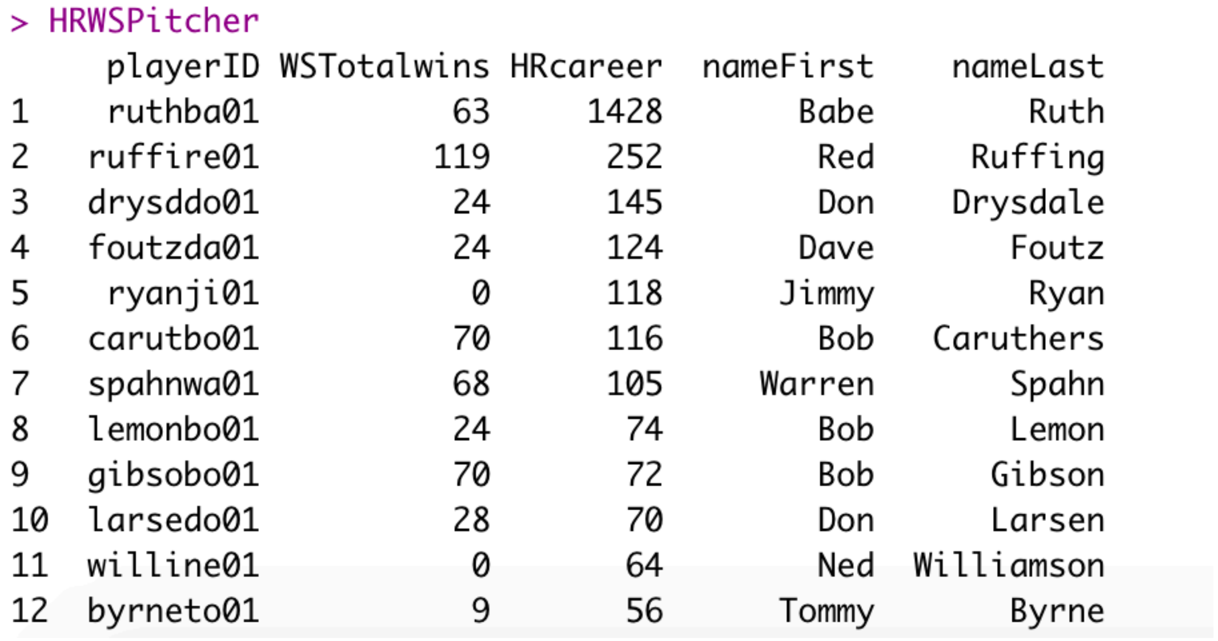
We can see from this linear regression model that Hits, Homeruns, Walks have positive relationship with winning percentage. However strikeouts, and earned runs have negative relationship with winning percentage.

**18. What players have pitched in the World Series and also hit a home run in their career**

Personally I found this question to be most interesting question because I thought there will be very few pitchers with very few home runs. So I joined the Batting tables and Pitching Post and find a player who have pitching record in World Series.



```
> HRWSPitcher = dbGetQuery(db, "SELECT P.playerID, SUM(W) AS WSTotalwins,
+                         SUM(B.HR) AS HRcareer, M.nameFirst, M.nameLast
+                         FROM PitchingPost P
+                         INNER JOIN Batting B INNER JOIN Master M
+                         ON P.playerID = B.playerID
+                         AND P.playerID = M.playerID
+                         WHERE B.HR > 0
+                         AND round = 'WS'
+                         GROUP BY P.playerID
+                         Order by HRcareer DESC")
```



```
> HRWSPitcher
    playerID WSTotalwins HRcareer nameFirst   nameLast
1   ruthba01          63     1428      Babe       Ruth
2  ruffire01         119      252       Red    Ruffing
3  drysddo01          24      145       Don   Drysdale
4  foutzda01          24      124      Dave      Foutz
5   ryanji01           0      118     Jimmy       Ryan
6  carutbo01          70      116       Bob  Caruthers
7  spahnwa01          68      105    Warren      Spahn
8  lemonbo01          24       74       Bob      Lemon
9  gibsobo01          70       72       Bob     Gibson
10 larsedo01          28       70       Don     Larsen
11 willine01           0       64       Ned Williamson
12 byrneto01           9       56     Tommy      Byrne
```

Then I found 462 players who have HR who have pitched in World Series. I was quite surprised with the result because I didn't expect this many pitchers that have HR.