# Shell and SQL

**YIBING LUO**

*" I certify that I have acknowledged any code that I used from any other person in the class, from Piazza or any Web site or book or other source. Any other work is my own. "*

## 1. UNIX Shell Tools

### *a. compute the numbers of airports and sort*

*(1) SHELL*

```
wget http://eeyore.ucdavis.edu/stat141/Data/Airline2012_13.tar.gz
$ tar -xzf Airline2012_13.tar.gz -C e://sta141/assignment/6/hw/airpline12_13
$ time for file in 201*.csv
> do
> cut -d , -f 15 $file | egrep 'SFO|SMF|OAK|LAX|JFK' >>orig
> done

real    2m32.892s
user    2m18.751s
sys     0m1.543s


$ time sort orig | uniq -c |sort -r
 214275 "LAX"
 162443 "SFO"
 104228 "JFK"
  41560 "OAK"
  40748 "SMF"

real    0m1.312s
user    0m3.546s
sys     0m0.076s
```

below is the result of the sorting and the running time.

*(2)R*

```
setwd("E:\\STA141/ASSIGNMENT/6/hw/airpline12_13")
dir = list.files()
pattern = c('SFO','SMF','OAK','LAX','JFK')


#caculate the number of th five airports
```

```
stat =

function(file, pattern){
   #
   #file: the file name
   #pattern: the name of the airports we want to find out
   #
   line = read.csv(file)
   aa = sapply(1:5, function(i) sum(grepl(pattern[i] , line[,15])))
   aa
}

System.time( result = sapply(1:12, function(i) stat(dir[i],pattern)) )
tt = sapply(1:12, function(i) stat(dir[i],pattern))
result = sapply(1:5, function(i) sum(tt[i,]))
names(result) = c('SFO','SMF','OAK','LAX','JFK')
sort(result , decreasing = TRUE)
```

```
LAX     SFO     JFK     OAK     SMF
214275 162443 104228   41560   40748
```

```
> system.time(sapply(1:12, function(i) stat(dir[i],pattern)))
   用户    系统    流逝
391.06    9.46 455.67
```

***Conclusion:***

We use R and SHELL separately to rank the airports which have the most number of the airplanes. And according to the same results, the two programs achieve the same goals.

After that, the time consuming (user time) is recorded. In shell, the time is 2m18s, and in R the time is 391.06s. As a result, processing the data in shell can save a lot of time comparing with the same procedure in R.

## *b. compute the total number of flights of the 5 airports*

```
subset =
   #
   #file:the name of the file
   #get the lines which involve any of these five airports
   #
function(file){
   cmd = paste('E:/STA141/bin/grep    -e SFO -e SMF -e OAK -e LAX -e JFK', file ,'>> pairs.txt')
   shell(cmd)
}


sapply(1:12, function(i) subset(dir[i])) #get pairs.txt

shell("E:/STA141/bin/cut -d , -f 15,25 pairs.txt >> pairs1.txt")
pair = read.table('pairs1.txt',sep = ",")
```

```
num_out = apply(sapply(1:5,function(i) grepl(pattern[i], pair[[1]])), 2, sum)
num_in = apply(sapply(1:5,function(i) grepl(pattern[i], pair[[2]])), 2, sum)

num_in_out = num_out+num_in
names(num_in_out) =    c('SFO','SMF','OAK','LAX','JFK')
sort(num_in_out, decreasing = TRUE)
```

**result:**

| LAX | SFO | JFK | OAK | SMF |
|---|---|---|---|---|
| 428611 | 324818 | 208373 | 83119 | 81483 |

# 2. Basebal, Databases and SQL

```
setwd("E:\\STA141/ASSIGNMENT/6/hwsql")
library(RSQLite)
db = dbConnect(SQLite(), dbname = 'lahman2013.sqlite')
table_name = dbListTables(db)    #get the names of all tables
```

**1. What years does the data cover? are there data for each of these years?**

```
year.range =
      #
      #pattern: is the string we want to find
      #
function(pattern, name){
   if(sum(grepl(pattern,dbListFields(db,name))) == 1){
      cmd = paste('SELECT MAX(yearID), MIN(yearID) FROM',name)
      dbGetQuery(db, cmd)
   }
}

year.p =

   function(pattern, name){
      if(sum(grepl(pattern,dbListFields(db,name))) == 1){
         cmd = paste('SELECT COUNT(DISTINCT (yearID)) FROM',name)
         dbGetQuery(db, cmd)
      }
   }


pattern = 'yearID'
year = sapply(1:25 , function(i) list(year.range(pattern,table_name[i]),year.p(pattern,table_name[i])))
```

**ANSWER:**

$AllstarFull

| | MAX(yearID) | MIN(yearID) | yearinterval | numberyear |
|---|---|---|---|---|
| 1 | 2013 | 1933 | 81 | 81 |

$Appearances

| | MAX(yearID) | MIN(yearID) | yearinterval | numberyear |
|---|---|---|---|---|
| 1 | 2013 | 1871 | 143 | 143 |

$AwardsManagers

| | MAX(yearID) | MIN(yearID) | yearinterval | numberyear |
|---|---|---|---|---|
| 1 | 2013 | 1936 | 78 | 78 |

$AwardsPlayers

| | MAX(yearID) | MIN(yearID) | yearinterval | numberyear |
|---|---|---|---|---|
| 1 | 2013 | 1877 | 137 | 115 |

There are only 4 tables showed here for the reason of space. The rest will be involved in code.
The data cover the year from 1871 to 2013.

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

```
players = dbGetQuery(db , 'SELECT DISTINCT playerID FROM Master')$playerID
managers = dbGetQuery(db, 'SELECT DISTINCT PlayerID FROM Managers')$playerID

length(players)
length(managers)
```

**ANSWER:**
There are total 18354 unique people included in the database. There are 682 managers of them, and 17672 players.

**3. What team won the World Series in 2000?**

```
winner = dbGetQuery(db, 'SELECT name FROM Teams WHERE    yearID = 2000 AND WSWin = "Y" ')
```

**ANSWER:**

| name |
|---|
| New York Yankees |

**4. What team lost the World Series each year?**

```
lose =dbGetQuery(db, 'SELECT name,yearID, WSWin, LgWIN FROM Teams WHERE WSWin = "N" AND LgWIN = "Y"')
```

**ANSWER:**

| | name | yearID | WSWin | LgWin |
|---|---|---|---|---|
| **1** | New York Metropolitans | 1884 | N | Y |
| **2** | St. Louis Browns | 1885 | N | Y |

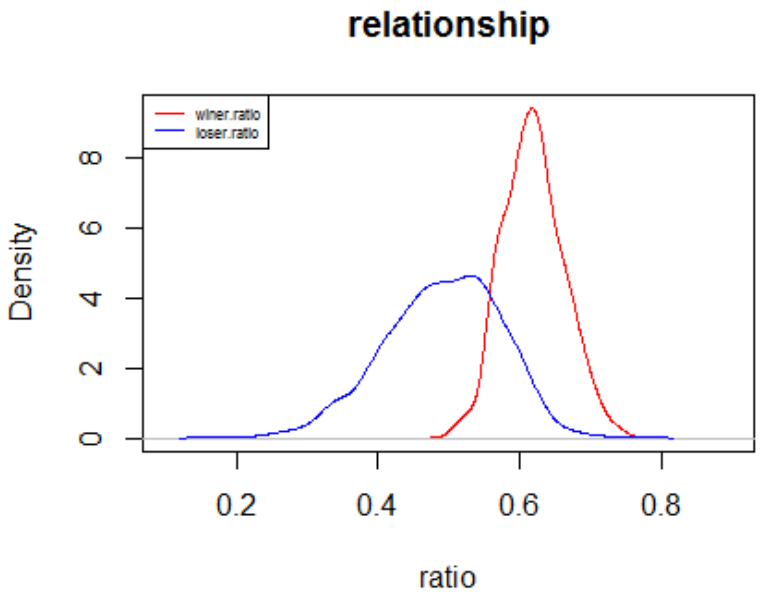
| | name | yearID | WSWin | LgWin |
|---|---|---|---|---|
| **3** | Chicago White Stockings | 1885 | N | Y |
| **4** | Chicago White Stockings | 1886 | N | Y |
| **5** | St. Louis Browns | 1887 | N | Y |
| **6** | St. Louis Browns | 1888 | N | Y |
| **7** | Brooklyn Bridegrooms | 1889 | N | Y |
| **8** | Louisville Colonels | 1890 | N | Y |
| **9** | Brooklyn Bridegrooms | 1890 | N | Y |
| **10** | Pittsburgh Pirates | 1903 | N | Y |
| **11** | Philadelphia Athletics | 1905 | N | Y |
| **12** | Chicago Cubs | 1906 | N | Y |
| **13** | Detroit Tigers | 1907 | N | Y |
| **14** | Detroit Tigers | 1908 | N | Y |
| **15** | Detroit Tigers | 1909 | N | Y |

There are 15 teams there because of the space. And the rest are in CODE.

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

```
winner.win = dbGetQuery(db,'SELECT yearID , W, G FROM Teams WHERE WSWin == "Y"')
winner.win.ratio =round( winner.win$W / winner.win$G,3)

loser.win = dbGetQuery(db,'SELECT yearID , W, G FROM Teams WHERE WSWin == "N"')
loser.win.ratio = round(loser.win$W / loser.win$G,3)

plot(density(winner.win.ratio),main = c("relationship"),col = 'red',xlim = c(0.1,0.9))
lines(density(loser.win.ratio),col = 'blue')
legend("topleft",legend = c("winer.ratio","loser.ratio"),col = c("red","blue"),lty = 1,cex=0.5)
```

**ANSWER:**




Based on the plot, it can be concluded that the team which won the world series won a relatively big number of games in a season.

**6. In 2003, what were the three highest salaries? (We refer here to unique salaries, i.e., more than one player might be paid one of these salaries.)**

```
salary_1 = dbGetQuery(db, 'SELECT DISTINCT salary FROM Salaries WHERE yearID =2003 ORDER BY salary')
-sort(-salary_1$salary)[1:3]
```

**ANSWER:**

The three highest salaries are 22,000,000, 20,000,000, 18,700,000.

**7. For 1999, 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.**

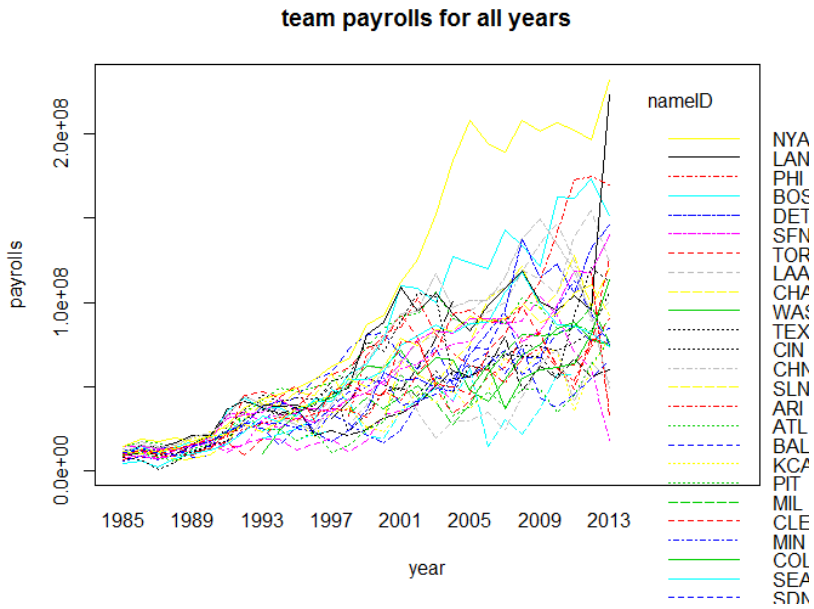
```
payroll = dbGetQuery(db, 'SELECT teamID, SUM(salary) AS SUM FROM Salaries WHERE yearID = 1990
                          GROUP BY teamID')

payroll.allyear = dbGetQuery(db, 'SELECT teamID,yearID, SUM(salary) AS payroll
                                  FROM Salaries GROUP BY teamID,yearID')

interaction.plot(payroll.allyear$yearID,payroll.allyear$teamID,
                 payroll.allyear$payroll,col = c(1:length(payroll.allyear$teamID)),trace.label = "nameID")
```

ANSWER:

For 1999, the total payroll of each of the different teams is as below:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| teamID | "ATL" | "BAL" | "BOS" | "CAL" | "CHA" | "CHN" | "CIN" | "CLE" | "DET" |
| SUM | "14555501" | " 9680084" | "20558333" | "21720000" | " 9491500" | "13624000" | "14370000" | "14487000" | "17593238" |
| teamID | "HOU" | "KCA" | "LAN" | "MIN" | "ML4" | "MON" | "NYA" | "NYN" | "OAK" |
| SUM | "18330000" | "23361084" | "21318704" | "14602000" | "19719167" | "16586388" | "20912318" | "21722834" | "19887501" |
| teamID | "PHI" | "PIT" | "SDN" | "SEA" | "SFN" | "SLN" | "TEX" | "TOR" | |
| SUM | "13173667" | "15556000" | "17588334" | "12553667" | "19335333" | "20523334" | "14874372" | "17756834" | |

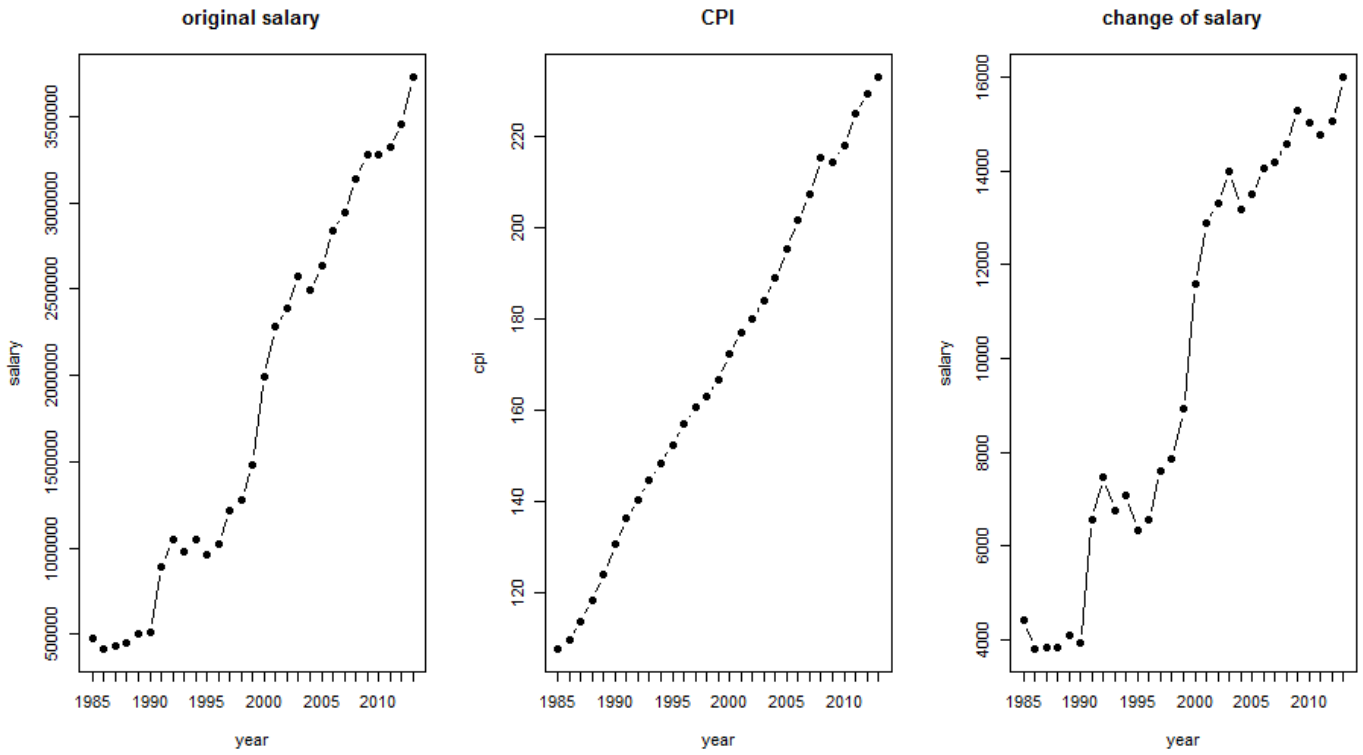
The next plot is about computing the team payrolls for all years in the database for which we have salary information:




**8. Study the change in salary over time. Have salaries kept up with inflation, fallen behind, or grown faster?**

```
salary.years = dbGetQuery(db, 'SELECT yearID,AVG(salary) AS salary FROM Salaries GROUP BY yearID')

setwd('E://STA141/ASSIGNMENT/6/hwsql')
inf = read.table('inflation.txt', head = TRUE)

salary.change = salary.years$salary /inf$Annual
#PLOT
par(mfrow = c(1,3))
plot(salary.years$salary , main = 'original salary' , type = 'b',xaxt = 'n',
        pch =19 , xlab = 'year',ylab = 'salary')
axis(1,at=1:length(salary.years$yearID),labels=salary.years$yearID)
plot(inf$Annual,main = 'CPI' , type = 'b',xaxt = 'n',pch =19 , xlab = 'year',ylab = 'cpi')
axis(1,at=1:length(salary.years$yearID),labels=salary.years$yearID)
plot(salary.change,main = 'change of salary' , type = 'b',xaxt = 'n',pch =19 , xlab = 'year',ylab = 'salary')
axis(1,at=1:length(salary.years$yearID),labels=salary.years$yearID)
```

**ANSWER:**




The set of plots below show what we want. The first plot shows the change of the original salary which involved inflation. The second plot shows the change of CPI, and it makes sense keep increasing. The third one reflection the change of salary which depart the factor of inflation. And from the third plot, it can be concluded that the salaries grow faster.

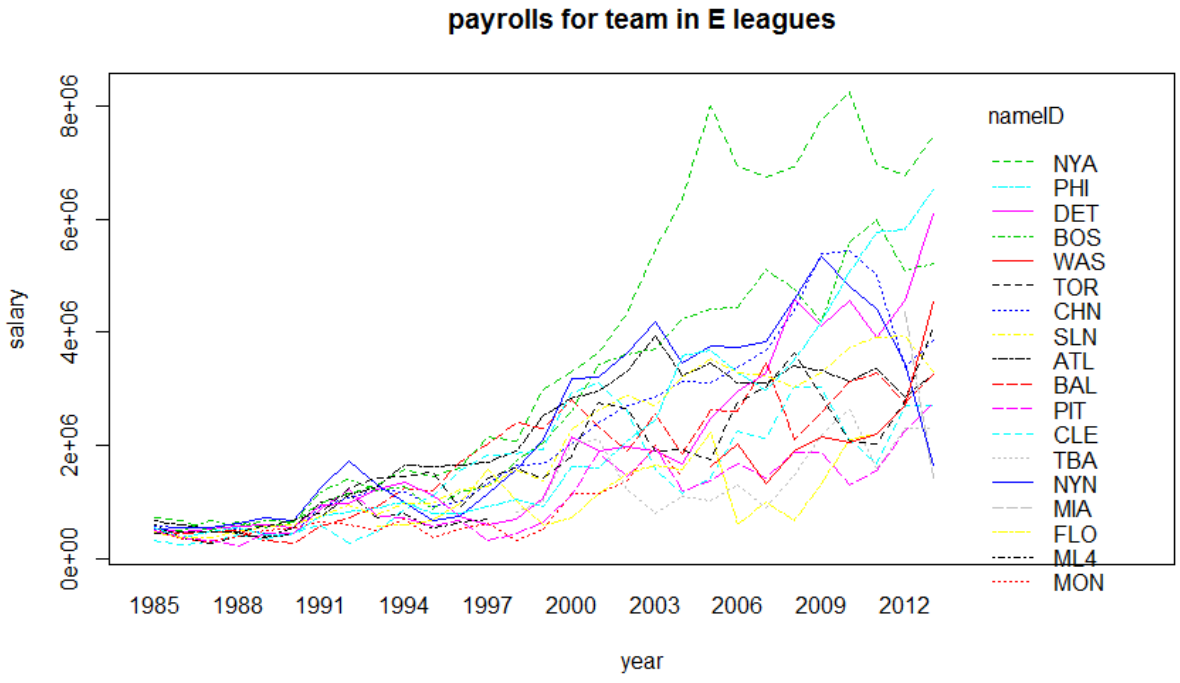
**9. Compare payrolls for the teams that are in the same leagues, and then in the same divisions. Are there any interesting characteristics? Have certain teams always had top payrolls over the years? Is there a connection between payroll and performance?**

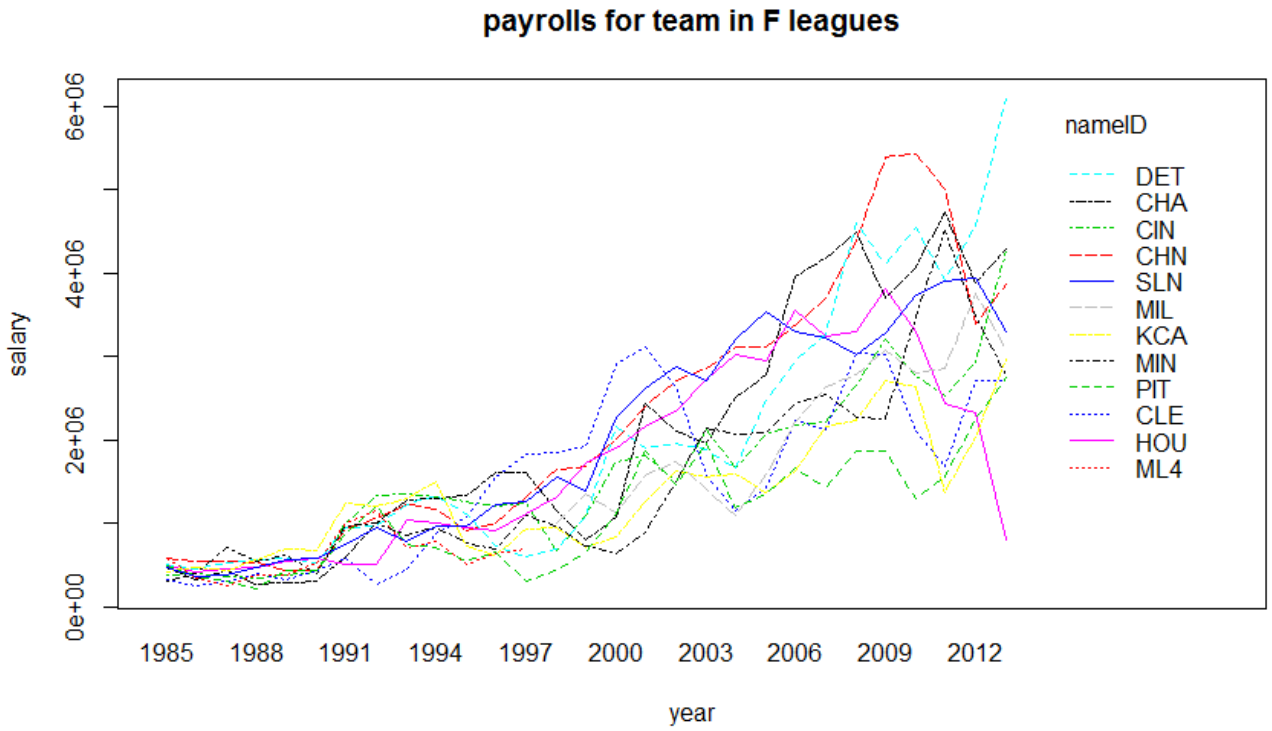
```
salary.info = dbGetQuery(db,'SELECT s.yearID, s.teamID, s.lgID, SUM(s.salary),
                t.DivID
                FROM Salaries AS s,
                Teams AS t
                WHERE s.teamID = t.teamID AND s.lgID =t.lgID AND s.yearID = t.yearID AND s.lgID = t.lgID
                        GROUP BY s.teamID, s.yearID;')

leagues = split(salary.info , salary.info$divID)
divisions = split(salary.info , salary.info$lgID)

#teams in the same leagues
interaction.plot(leagues$E[,1],leagues$E[,2],leagues$E[,4],type = "l",
                col = c(1:length(table(leagues$E[,1]))),trace.label = "nameID")
interaction.plot(leagues$C[,1],leagues$C[,2],leagues$C[,4],type = "l",
                col = c(1:length(table(leagues$E[,1]))),trace.label = "nameID")
interaction.plot(leagues$W[,1],leagues$W[,2],leagues$W[,4],type = "l",
                col = c(1:length(table(leagues$E[,1]))),trace.label = "nameID")

#teams in the same divisions
interaction.plot(divisions$AL[,1],divisions$AL[,2],divisions$AL[,4],type = "l",
                col = c(1:length(table(divisions$AL[,1]))),trace.label = "nameID")
```

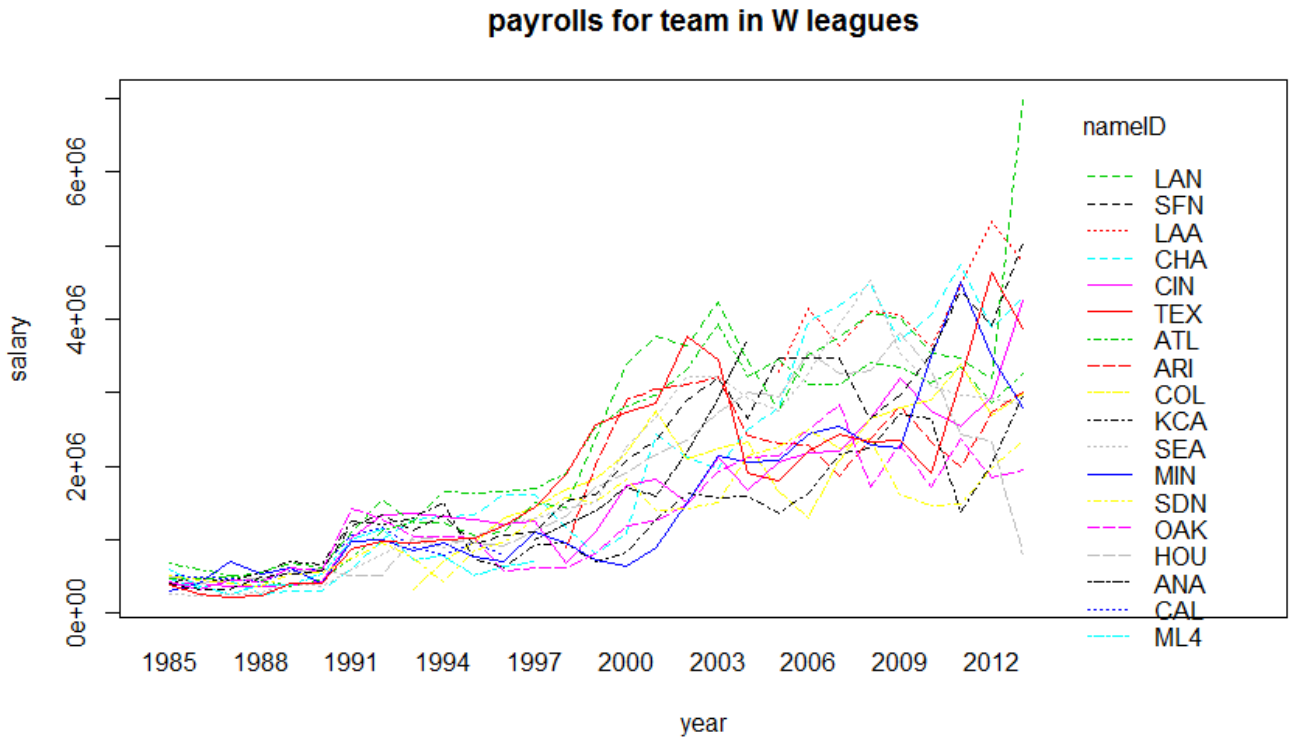
```
interaction.plot(divisions$NL[,1],divisions$NL[,2],divisions$NL[,4],type = "l",
                 col = c(1:length(table(divisions$NL[,1]))), trace.label = "nameID")

team.info = dbGetQuery(db,'SELECT s.yearID,s.teamID,SUM(s.salary) AS salary,
                       t.teamID,t.yearID, t.G,t.W,t.DivWin,t.WCWin,t.LgWin,t.WsWin
                       FROM Salaries AS s,
                       Teams AS t
                       ON s.teamID = t.teamID
                         GROUP BY s.yearID, s.teamID;')
#performance
tt = split(team.info, team.info$yearID)
sapply(1:29, function(i) any(tt[[i]][order(-tt[[i]][,3]),][,8:11][1:3,]== 'Y'))

plot(team.info$salary,team.info$W/team.info$G,pch=19,cex=.5)
cor(team.info$salary,team.info$W/team.info$G)
```

**ANSWER:**

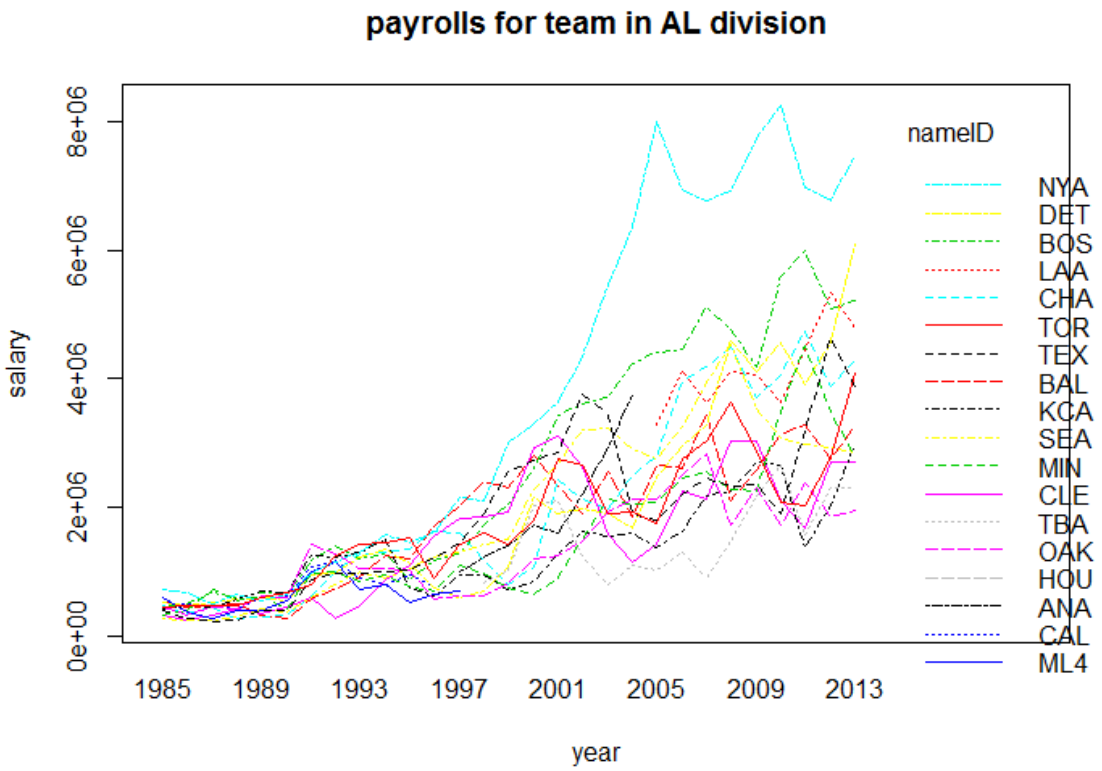
In the same leagues:

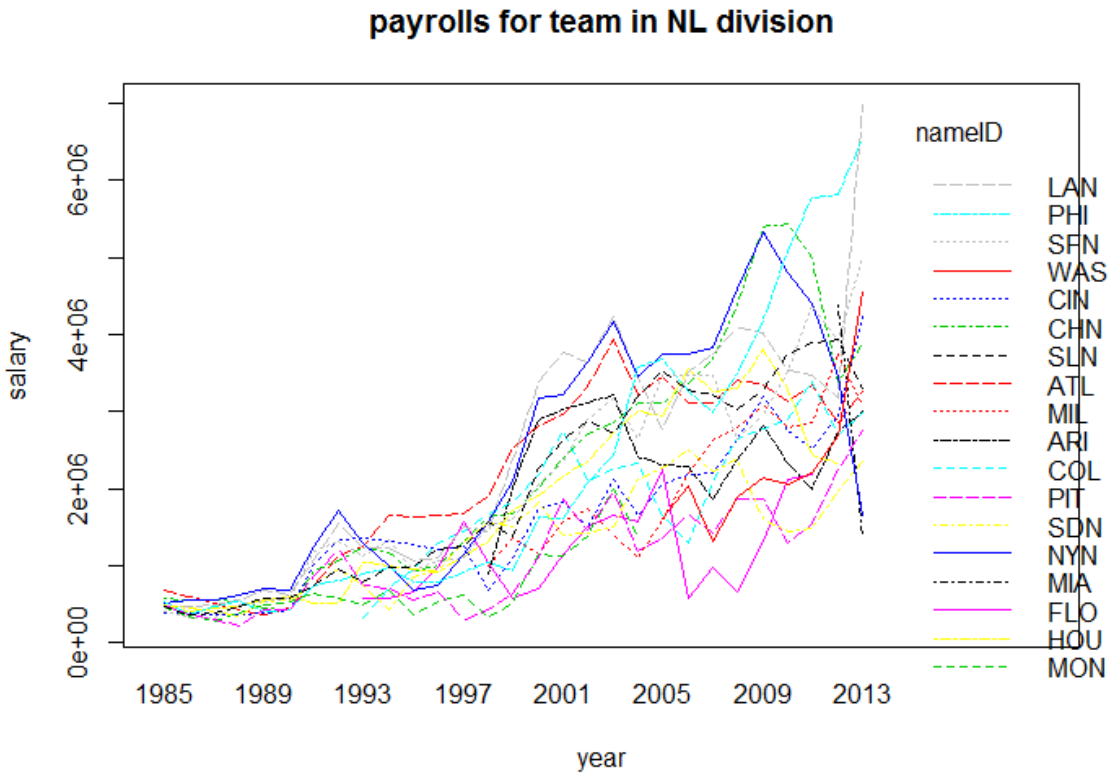





In the same divisions:




From the plots we can see the payrolls of the teams were almost the same at the beginning (1985), and remain increasing. However, when the time goes by, the difference between each teams can be different and somehow tremendous.

One interesting characteristic is that: teams which paid a low salary at the beginning period would remain a low salary in the future or just dismissed. And the teams who pay a high salary in the 2013, usually have a long history.

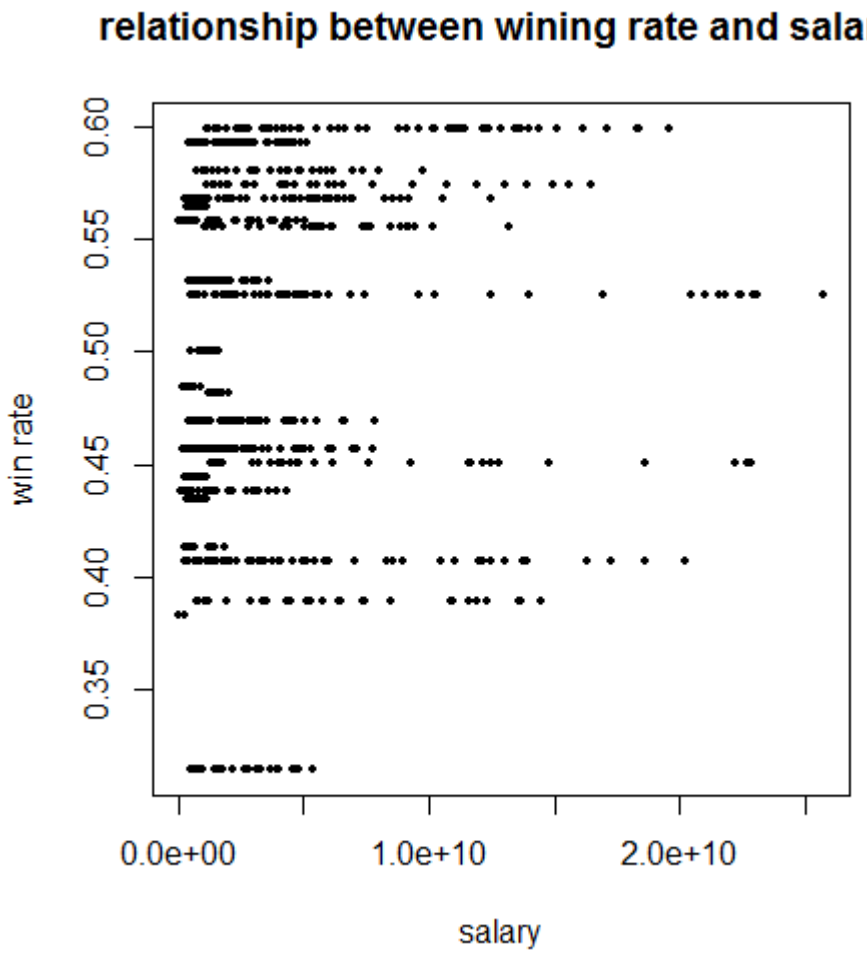
NYA(New York Yankees) always had top payrolls from 1999~2013. But others were not to be so.

The table above shows that whether the top three highest payrolls teams of each year won any kind of championship.

[1] TRUE TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

[17] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE TRUE TRUE TRUE

It seems that there is a relationship between salary and performance. But when considering about wining rate, it goes differently.




There is no direct pattern shows there is a relationship between salary and win rate. And the correlation coefficient confirm it ( corr(salary, winrate) = 0.1237777).

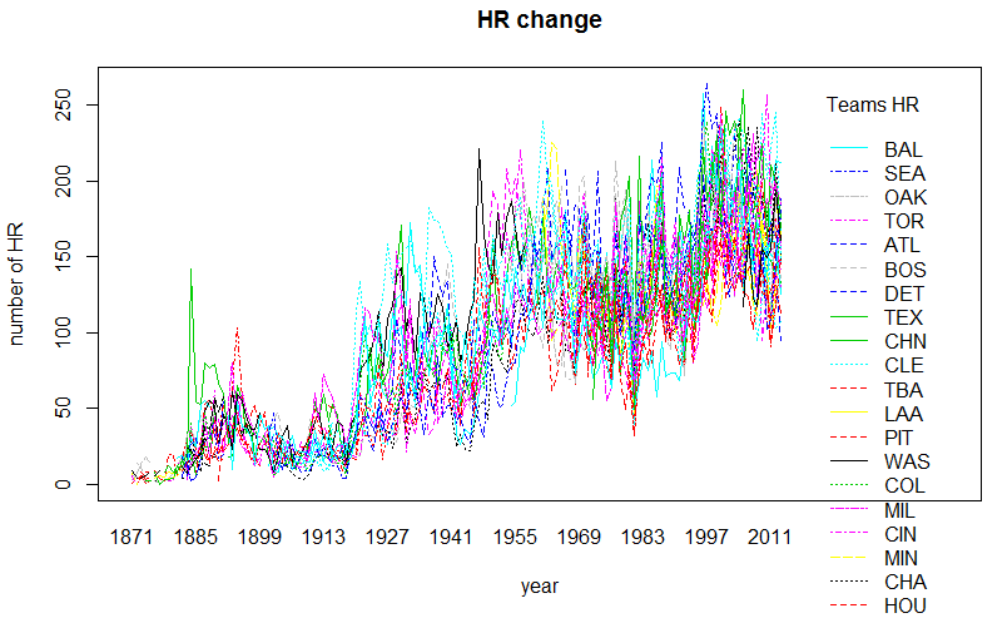
**10. Has the distribution of home runs for players increased over the years? When answering the questions, try to summarize the results in convenient and informative form (e.g. tables and/or plots) that illustrate the key features.**

```
#consider about HR in each team changing by the time
teams.HR =dbGetQuery(db,'SELECT teamID , yearID, SUM(HR) AS HR FROM Teams GROUP BY yearID, teamID')

teams.HR.year = split(teams.HR, teams.HR$teamID)

interaction.plot(teams.HR[,2],teams.HR[,1],teams.HR[,3],col=c(1:length(table(teams.HR[,1]))))


#consider about HR for every player changing by the time
hr.info = dbGetQuery(db,'SELECT playerID, yearID, HR FROM Batting order BY playerID;')
bat = split(batting.hr.info, batting.hr.info$playerID)
year.mean = sapply(1:length(bat), function(i) median(bat[[i]][[2]]))
```

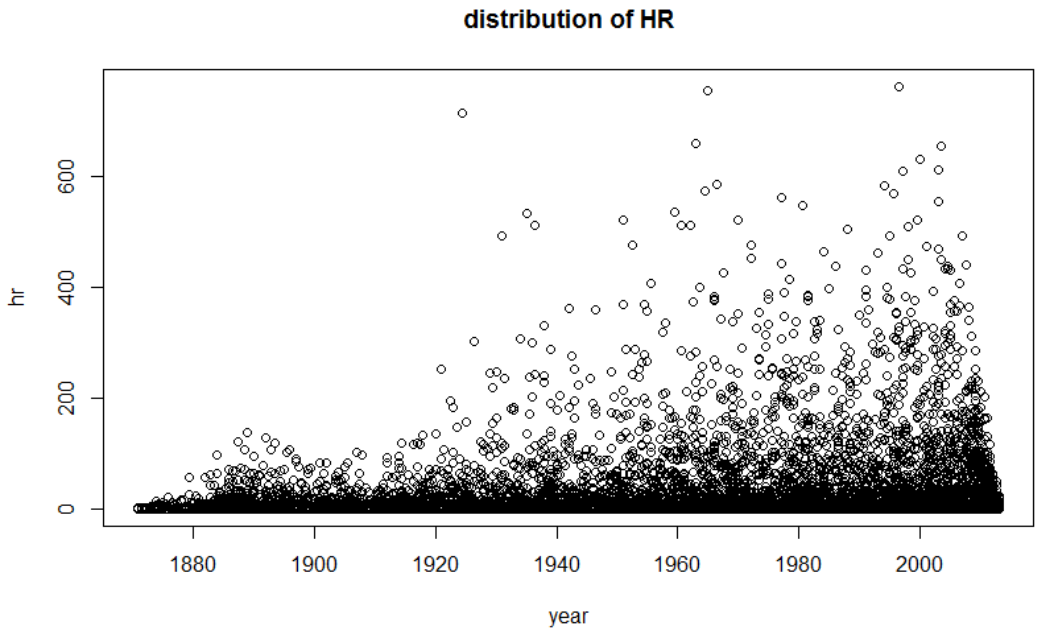
bat.hr = sapply(1:length(bat), function(i) sum(bat[[i]][[3]], na.rm =TRUE))

plot(year.mean, bat.hr, main = 'distribution of HR',xlab = 'year', ylab = 'hr')

**ANSWER:**




From the plot below, we can find that the number of HR in each team is increasing over the years even there are some




The points in the above plot represent the players. Year is the median of the year player has been experiencing. And the HR is the total number of the players homeruns time.

From this plot, we can find that the distribution of home runs for players increased over the years.