STA141B HW3

Sitong Qian

11/10/2020

#1:What years does the data cover? are there data for each of these years?

To solve this, first I read over the description of baseball and found out the table have yearID as variable, since Teams probably is the most conclusive one, so I focus on Teams Table, and with SQL function like this. To find if there is data for each of these years, I simply use distinct function to find how many distince years covered in the dataset and do some alegebra, to test if it is covered for all the range.

```
yearID = dbGetQuery(baseball,'SELECT yearID FROM Teams')
distinct_yearID <- distinct(yearID)
```

The years covered from 1871 to 2013. 2013-1871+1 = 143, Since there are 143 rows in the unique yearID. thus there are data for each of these years.

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

To solve this, I first looked for playerID in Master table by using SQLite select functions, for players, I union join three tables, Batting, Fielding and Pitching to find if individuals are actually involved in playing games, and in case there are repetitive count in three tables, I used distinct function here. Then I use SQLite function extract number of managers in managers table.

```
#first group three sets by player id
#union (search )
#inner join the manager table with this outer join
listFields = dbListFields(baseball,'Master')
playeridplayer = dbGetQuery(baseball,'SELECT playerID From Master')
playeridmaster = dbGetQuery(baseball,'SELECT playerID From Managers')
playeridplyear_row = nrow(table(playeridplayer)) #count the individual player occured
times, to find how many unique players are there
playeridmaster_row =nrow(table(playeridmaster)) #count the individual player occured
times, to find how many unique managers are there
playerjoin = dbGetQuery(baseball,'SELECT playerID FROM Batting UNION SELECT playerID
FROM Fielding UNION SELECT playerID FROM Pitching')
playerjoin_row = nrow(table(distinct(playerjoin)))
```

Based on the output, there are 18354 unique people, about 682 managers, and 18170 players recorded in this data.

#3: How many players became managers?

To find this, I simply use R program, by rbind two dataset, and use table function on playerID.Since I precheck each ID only appear once in one dataset. Thus, If I use table function on combined dataset and have frequency is 2 means the same ID appeared in both dataset, which indicates, the players became managers.

```
unqiuemasterid = unique(playeridmaster)
combinedid = rbind(playerjoin,unqiuemasterid)
combinedidfreq = as.data.frame(table(combinedid))
rep = subset(combinedidfreq,combinedidfreq$Freq == 2)
rep_row = nrow(rep)
```

There are 561 players become managers.

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

I first used SQLlite approached for selecting playerID with time period greater than 2000 in three position information table, and used pipeline functions to group the processed dataset by yearID. Since I knew from previous questions tone playerID will only occur once for each position information, I just need to count the yearID, to find out how many players in each year. Then, I apply the same method for teamID.

```
playeryearID = dbGetQuery(baseball,'SELECT yearID,playerID FROM Batting Where yearID
>= 2000 UNION SELECT yearID,playerID FROM Fielding Where yearID >= 2000 UNION SELECT
yearID,playerID FROM Pitching Where yearID >= 2000')
playeryearIDcount <- playeryearID  %>%
  group_by(yearID) %>%
  count(yearID) %>%
  rename(year_players_count = n)
playeryearIDcount
```

```
## # A tibble: 14 x 2
## # Groups:   yearID [14]
##    yearID year_players_count
##     <int>              <int>
##  1   2000               1230
##  2   2001               1220
##  3   2002               1218
##  4   2003               1230
##  5   2004               1247
##  6   2005               1237
##  7   2006               1242
##  8   2007               1278
##  9   2008               1291
## 10   2009               1266
## 11   2010               1249
## 12   2011               1295
## 13   2012               1284
## 14   2013               1305
```

```
playeryearteamID = dbGetQuery(baseball,'SELECT yearID,playerID,teamID FROM Batting Wh
ere yearID >= 2000 UNION SELECT yearID,playerID,teamID FROM Fielding Where yearID >=
2000 UNION SELECT yearID,playerID,teamID FROM Pitching Where yearID >= 2000')
playeryearteamIDcount <- playeryearteamID  %>%
  group_by(teamID) %>%
  count(teamID) %>%
  rename(team_players_count = n)
playeryearteamIDcount
```

```
## # A tibble: 33 x 2
## # Groups:   teamID [33]
##    teamID team_players_count
##    <chr>               <int>
##  1 ANA                   204
##  2 ARI                   637
##  3 ATL                   612
##  4 BAL                   665
##  5 BOS                   686
##  6 CHA                   577
##  7 CHN                   644
##  8 CIN                   647
##  9 CLE                   677
## 10 COL                   683
## # … with 23 more rows
```

No,as obtained from table, different teams have different number of players.

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

I used SQlite function to extract information in SeriesPost and left join with Teams table on both with same teamID, and also set the round to be the final round, and yearID = 2010. When doing this question, I found divisionID changed with repesct to time for same team, so take this into account, I also included another condition, joining information when two tables with same yearID information.

```
WorldSerieswin2010 = dbGetQuery(baseball,'SELECT SeriesPost.yearID,SeriesPost.teamIDw
inner,SeriesPost.lgIDwinner,TeamsHalf.divID,Teams.name FROM SeriesPost LEFT JOIN Team
sHalf ON SeriesPost.teamIDwinner = TeamsHalf.teamID LEFT JOIN Teams on SeriesPost.tea
mIDwinner = Teams.teamID WHERE SeriesPost.yearID == 2010 AND SeriesPost.round = "WS"
AND SeriesPost.yearID == Teams.yearID')
distinct(WorldSerieswin2010)
```

```
##   yearID teamIDwinner lgIDwinner divID                name
## 1   2010          SFN         NL     W San Francisco Giants
```

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

I used SQlite function to extract information in SeriesPost and left join with Teams table with same teamID, and also set the round to be the final round, and order by yearID

```
WorldSeriesloss = dbGetQuery(baseball,'SELECT SeriesPost.yearID,SeriesPost.teamIDlose
r,SeriesPost.lgIDloser,Teams.divID,Teams.name FROM SeriesPost LEFT JOIN Teams ON Seri
esPost.teamIDloser = Teams.teamID WHERE SeriesPost.yearID = Teams.yearID AND SeriesPo
st.round = "WS" ORDER BY SeriesPost.yearID')
distinct(WorldSeriesloss)
```

```
##     yearID teamIDloser lgIDloser divID                   name
## 1     1903         PIT        NL  <NA>     Pittsburgh Pirates
## 2     1905         PHA        AL  <NA> Philadelphia Athletics
## 3     1906         CHN        NL  <NA>           Chicago Cubs
## 4     1907         DET        AL  <NA>         Detroit Tigers
## 5     1908         DET        AL  <NA>         Detroit Tigers
## 6     1909         DET        AL  <NA>         Detroit Tigers
## 7     1910         CHN        NL  <NA>           Chicago Cubs
## 8     1911         NY1        NL  <NA>        New York Giants
## 9     1912         NY1        NL  <NA>        New York Giants
## 10    1913         NY1        NL  <NA>        New York Giants
## 11    1914         PHA        AL  <NA> Philadelphia Athletics
## 12    1915         PHI        NL  <NA>  Philadelphia Phillies
## 13    1916         BRO        NL  <NA>        Brooklyn Robins
## 14    1917         NY1        NL  <NA>        New York Giants
## 15    1918         CHN        NL  <NA>           Chicago Cubs
## 16    1919         CHA        AL  <NA>      Chicago White Sox
```

```
## 17    1920         BRO        NL  <NA>        Brooklyn Robins
## 18    1921         NYA        AL  <NA>       New York Yankees
## 19    1922         NYA        AL  <NA>       New York Yankees
## 20    1923         NY1        NL  <NA>        New York Giants
## 21    1924         NY1        NL  <NA>        New York Giants
## 22    1925         WS1        AL  <NA>    Washington Senators
## 23    1926         NYA        AL  <NA>       New York Yankees
## 24    1927         PIT        NL  <NA>     Pittsburgh Pirates
## 25    1928         SLN        NL  <NA>    St. Louis Cardinals
## 26    1929         CHN        NL  <NA>           Chicago Cubs
## 27    1930         SLN        NL  <NA>    St. Louis Cardinals
## 28    1931         PHA        AL  <NA> Philadelphia Athletics
## 29    1932         CHN        NL  <NA>           Chicago Cubs
## 30    1933         WS1        AL  <NA>    Washington Senators
## 31    1934         DET        AL  <NA>         Detroit Tigers
## 32    1935         CHN        NL  <NA>           Chicago Cubs
## 33    1936         NY1        NL  <NA>        New York Giants
## 34    1937         NY1        NL  <NA>        New York Giants
## 35    1938         CHN        NL  <NA>           Chicago Cubs
## 36    1939         CIN        NL  <NA>        Cincinnati Reds
## 37    1940         DET        AL  <NA>         Detroit Tigers
## 38    1941         BRO        NL  <NA>       Brooklyn Dodgers
## 39    1942         NYA        AL  <NA>       New York Yankees
## 40    1943         SLN        NL  <NA>    St. Louis Cardinals
## 41    1944         SLA        AL  <NA>       St. Louis Browns
## 42    1945         CHN        NL  <NA>           Chicago Cubs
## 43    1946         BOS        AL  <NA>         Boston Red Sox
## 44    1947         BRO        NL  <NA>       Brooklyn Dodgers
## 45    1948         BSN        NL  <NA>          Boston Braves
## 46    1949         BRO        NL  <NA>       Brooklyn Dodgers
## 47    1950         PHI        NL  <NA>  Philadelphia Phillies
## 48    1951         NY1        NL  <NA>        New York Giants
## 49    1952         BRO        NL  <NA>       Brooklyn Dodgers
## 50    1953         BRO        NL  <NA>       Brooklyn Dodgers
## 51    1954         CLE        AL  <NA>      Cleveland Indians
## 52    1955         NYA        AL  <NA>       New York Yankees
## 53    1956         BRO        NL  <NA>       Brooklyn Dodgers
## 54    1957         NYA        AL  <NA>       New York Yankees
## 55    1958         ML1        NL  <NA>       Milwaukee Braves
## 56    1959         CHA        AL  <NA>      Chicago White Sox
## 57    1960         NYA        AL  <NA>       New York Yankees
## 58    1961         CIN        NL  <NA>        Cincinnati Reds
## 59    1962         SFN        NL  <NA>   San Francisco Giants
## 60    1963         NYA        AL  <NA>       New York Yankees
## 61    1964         NYA        AL  <NA>       New York Yankees
## 62    1965         MIN        AL  <NA>        Minnesota Twins
## 63    1966         LAN        NL  <NA>    Los Angeles Dodgers
```

```
## 64    1967         BOS        AL  <NA>         Boston Red Sox
## 65    1968         SLN        NL  <NA>    St. Louis Cardinals
## 66    1969         BAL        AL     E      Baltimore Orioles
## 67    1970         CIN        NL     W        Cincinnati Reds
## 68    1971         BAL        AL     E      Baltimore Orioles
## 69    1972         CIN        NL     W        Cincinnati Reds
## 70    1973         NYN        NL     E          New York Mets
## 71    1974         LAN        NL     W    Los Angeles Dodgers
## 72    1975         BOS        AL     E         Boston Red Sox
## 73    1976         NYA        AL     E       New York Yankees
## 74    1977         LAN        NL     W    Los Angeles Dodgers
## 75    1978         LAN        NL     W    Los Angeles Dodgers
## 76    1979         BAL        AL     E      Baltimore Orioles
## 77    1980         KCA        AL     W     Kansas City Royals
## 78    1981         NYA        AL     E       New York Yankees
## 79    1982         ML4        AL     E      Milwaukee Brewers
## 80    1983         PHI        NL     E  Philadelphia Phillies
## 81    1984         SDN        NL     W       San Diego Padres
## 82    1985         SLN        NL     E    St. Louis Cardinals
## 83    1986         BOS        AL     E         Boston Red Sox
## 84    1987         SLN        NL     E    St. Louis Cardinals
## 85    1988         OAK        AL     W      Oakland Athletics
## 86    1989         SFN        NL     W   San Francisco Giants
## 87    1990         OAK        AL     W      Oakland Athletics
## 88    1991         ATL        NL     W         Atlanta Braves
## 89    1992         ATL        NL     W         Atlanta Braves
## 90    1993         PHI        NL     E  Philadelphia Phillies
## 91    1995         CLE        AL     C      Cleveland Indians
## 92    1996         ATL        NL     E         Atlanta Braves
## 93    1997         CLE        AL     C      Cleveland Indians
## 94    1998         SDN        NL     W       San Diego Padres
## 95    1999         ATL        NL     E         Atlanta Braves
## 96    2000         NYN        NL     E          New York Mets
## 97    2001         NYA        AL     E       New York Yankees
## 98    2002         SFN        NL     W   San Francisco Giants
## 99    2003         NYA        AL     E       New York Yankees
## 100   2004         SLN        NL     C    St. Louis Cardinals
## 101   2005         HOU        NL     C          Houston Astros
## 102   2006         DET        AL     C         Detroit Tigers
## 103   2007         COL        NL     W       Colorado Rockies
## 104   2008         TBA        AL     E         Tampa Bay Rays
## 105   2009         PHI        NL     E  Philadelphia Phillies
## 106   2010         TEX        AL     W          Texas Rangers
## 107   2011         TEX        AL     W          Texas Rangers
## 108   2012         DET        AL     C         Detroit Tigers
## 109   2013         SLN        NL     C    St. Louis Cardinals
```

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

I used SQlite function to extract information in SeriesPost and left join with Teams table with same teamID, and also set the round to be the final round, and order by yearID.

```
WorldSerieswin = dbGetQuery(baseball,'SELECT SeriesPost.yearID,SeriesPost.teamIDwinne
r,SeriesPost.lgIDwinner,Teams.divID,Teams.name FROM SeriesPost LEFT JOIN Teams ON Ser
iesPost.teamIDwinner = Teams.teamID WHERE SeriesPost.yearID = Teams.yearID AND Series
Post.round = "WS" ORDER BY SeriesPost.yearID')
distinctwin = distinct(WorldSerieswin)
distinctwin
```

```
##     yearID teamIDwinner lgIDwinner divID                   name
## 1     1884          PRO         NL  <NA>       Providence Grays
## 2     1887          DTN         NL  <NA>     Detroit Wolverines
## 3     1890          BRO         NL  <NA>   Brooklyn Bridegrooms
## 4     1903          BOS         AL  <NA>       Boston Americans
## 5     1905          NY1         NL  <NA>        New York Giants
## 6     1906          CHA         AL  <NA>      Chicago White Sox
## 7     1907          CHN         NL  <NA>           Chicago Cubs
## 8     1908          CHN         NL  <NA>           Chicago Cubs
## 9     1909          PIT         NL  <NA>     Pittsburgh Pirates
## 10    1910          PHA         AL  <NA> Philadelphia Athletics
## 11    1911          PHA         AL  <NA> Philadelphia Athletics
## 12    1912          BOS         AL  <NA>         Boston Red Sox
## 13    1913          PHA         AL  <NA> Philadelphia Athletics
## 14    1914          BSN         NL  <NA>          Boston Braves
## 15    1915          BOS         AL  <NA>         Boston Red Sox
## 16    1916          BOS         AL  <NA>         Boston Red Sox
## 17    1917          CHA         AL  <NA>      Chicago White Sox
## 18    1918          BOS         AL  <NA>         Boston Red Sox
## 19    1919          CIN         NL  <NA>        Cincinnati Reds
## 20    1920          CLE         AL  <NA>      Cleveland Indians
## 21    1921          NY1         NL  <NA>        New York Giants
## 22    1922          NY1         NL  <NA>        New York Giants
## 23    1923          NYA         AL  <NA>       New York Yankees
## 24    1924          WS1         AL  <NA>    Washington Senators
## 25    1925          PIT         NL  <NA>     Pittsburgh Pirates
## 26    1926          SLN         NL  <NA>    St. Louis Cardinals
## 27    1927          NYA         AL  <NA>       New York Yankees
## 28    1928          NYA         AL  <NA>       New York Yankees
## 29    1929          PHA         AL  <NA> Philadelphia Athletics
## 30    1930          PHA         AL  <NA> Philadelphia Athletics
## 31    1931          SLN         NL  <NA>    St. Louis Cardinals
## 32    1932          NYA         AL  <NA>       New York Yankees
```

```
## 33    1933          NY1         NL  <NA>        New York Giants
## 34    1934          SLN         NL  <NA>    St. Louis Cardinals
## 35    1935          DET         AL  <NA>         Detroit Tigers
## 36    1936          NYA         AL  <NA>       New York Yankees
## 37    1937          NYA         AL  <NA>       New York Yankees
## 38    1938          NYA         AL  <NA>       New York Yankees
## 39    1939          NYA         AL  <NA>       New York Yankees
## 40    1940          CIN         NL  <NA>        Cincinnati Reds
## 41    1941          NYA         AL  <NA>       New York Yankees
## 42    1942          SLN         NL  <NA>    St. Louis Cardinals
## 43    1943          NYA         AL  <NA>       New York Yankees
## 44    1944          SLN         NL  <NA>    St. Louis Cardinals
## 45    1945          DET         AL  <NA>         Detroit Tigers
## 46    1946          SLN         NL  <NA>    St. Louis Cardinals
## 47    1947          NYA         AL  <NA>       New York Yankees
## 48    1948          CLE         AL  <NA>      Cleveland Indians
## 49    1949          NYA         AL  <NA>       New York Yankees
## 50    1950          NYA         AL  <NA>       New York Yankees
## 51    1951          NYA         AL  <NA>       New York Yankees
## 52    1952          NYA         AL  <NA>       New York Yankees
## 53    1953          NYA         AL  <NA>       New York Yankees
## 54    1954          NY1         NL  <NA>        New York Giants
## 55    1955          BRO         NL  <NA>       Brooklyn Dodgers
## 56    1956          NYA         AL  <NA>       New York Yankees
## 57    1957          ML1         NL  <NA>       Milwaukee Braves
## 58    1958          NYA         AL  <NA>       New York Yankees
## 59    1959          LAN         NL  <NA>    Los Angeles Dodgers
## 60    1960          PIT         NL  <NA>     Pittsburgh Pirates
## 61    1961          NYA         AL  <NA>       New York Yankees
## 62    1962          NYA         AL  <NA>       New York Yankees
## 63    1963          LAN         NL  <NA>    Los Angeles Dodgers
## 64    1964          SLN         NL  <NA>    St. Louis Cardinals
## 65    1965          LAN         NL  <NA>    Los Angeles Dodgers
## 66    1966          BAL         AL  <NA>      Baltimore Orioles
## 67    1967          SLN         NL  <NA>    St. Louis Cardinals
## 68    1968          DET         AL  <NA>         Detroit Tigers
## 69    1969          NYN         NL     E          New York Mets
## 70    1970          BAL         AL     E      Baltimore Orioles
## 71    1971          PIT         NL     E     Pittsburgh Pirates
## 72    1972          OAK         AL     W      Oakland Athletics
## 73    1973          OAK         AL     W      Oakland Athletics
## 74    1974          OAK         AL     W      Oakland Athletics
## 75    1975          CIN         NL     W        Cincinnati Reds
## 76    1976          CIN         NL     W        Cincinnati Reds
## 77    1977          NYA         AL     E       New York Yankees
## 78    1978          NYA         AL     E       New York Yankees
## 79    1979          PIT         NL     E     Pittsburgh Pirates
```

```
## 80    1980          PHI         NL     E  Philadelphia Phillies
## 81    1981          LAN         NL     W    Los Angeles Dodgers
## 82    1982          SLN         NL     E    St. Louis Cardinals
## 83    1983          BAL         AL     E      Baltimore Orioles
## 84    1984          DET         AL     E         Detroit Tigers
## 85    1985          KCA         AL     W     Kansas City Royals
## 86    1986          NYN         NL     E          New York Mets
## 87    1987          MIN         AL     W        Minnesota Twins
## 88    1988          LAN         NL     W    Los Angeles Dodgers
## 89    1989          OAK         AL     W      Oakland Athletics
## 90    1990          CIN         NL     W        Cincinnati Reds
## 91    1991          MIN         AL     W        Minnesota Twins
## 92    1992          TOR         AL     E      Toronto Blue Jays
## 93    1993          TOR         AL     E      Toronto Blue Jays
## 94    1995          ATL         NL     E         Atlanta Braves
## 95    1996          NYA         AL     E       New York Yankees
## 96    1997          FLO         NL     E        Florida Marlins
## 97    1998          NYA         AL     E       New York Yankees
## 98    1999          NYA         AL     E       New York Yankees
## 99    2000          NYA         AL     E       New York Yankees
## 100   2001          ARI         NL     W   Arizona Diamondbacks
## 101   2002          ANA         AL     W         Anaheim Angels
## 102   2003          FLO         NL     E        Florida Marlins
## 103   2004          BOS         AL     E         Boston Red Sox
## 104   2005          CHA         AL     C      Chicago White Sox
## 105   2006          SLN         NL     C    St. Louis Cardinals
## 106   2007          BOS         AL     E         Boston Red Sox
## 107   2008          PHI         NL     E  Philadelphia Phillies
## 108   2009          NYA         AL     E       New York Yankees
## 109   2010          SFN         NL     W   San Francisco Giants
## 110   2011          SLN         NL     C    St. Louis Cardinals
## 111   2012          SFN         NL     W   San Francisco Giants
## 112   2013          BOS         AL     E         Boston Red Sox
```

#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.

I basically used the same logic as the above question, but just including more information from two tables. Here I have noticed that there are problems exsited in the loss information for world series, thus the table start from 1903, otherwise, it should start from 1884.

```
WorldSeriesloser = dbGetQuery(baseball,'SELECT SeriesPost.yearID,SeriesPost.teamIDlos
er,SeriesPost.lgIDloser,Teams.divID,Teams.name,SeriesPost.losses FROM SeriesPost
                              LEFT JOIN Teams ON SeriesPost.teamIDloser = Teams.teamID
                              WHERE SeriesPost.yearID = Teams.yearID AND SeriesPost.ro
und = "WS"
                              ORDER BY SeriesPost.yearID')
distinct = distinct(WorldSeriesloser)
merge_WorldSeriesloser <- merge(WorldSerieswin,WorldSeriesloser,by = c('yearID'))
merge_WorldSeriesloser
```

```
##     yearID teamIDwinner lgIDwinner divID.x               name.x teamIDloser
## 1     1903          BOS         AL    <NA>       Boston Americans         PIT
## 2     1905          NY1         NL    <NA>        New York Giants         PHA
## 3     1906          CHA         AL    <NA>      Chicago White Sox         CHN
## 4     1907          CHN         NL    <NA>           Chicago Cubs         DET
## 5     1908          CHN         NL    <NA>           Chicago Cubs         DET
## 6     1909          PIT         NL    <NA>     Pittsburgh Pirates         DET
## 7     1910          PHA         AL    <NA> Philadelphia Athletics         CHN
## 8     1911          PHA         AL    <NA> Philadelphia Athletics         NY1
## 9     1912          BOS         AL    <NA>         Boston Red Sox         NY1
## 10    1913          PHA         AL    <NA> Philadelphia Athletics         NY1
## 11    1914          BSN         NL    <NA>          Boston Braves         PHA
## 12    1915          BOS         AL    <NA>         Boston Red Sox         PHI
## 13    1916          BOS         AL    <NA>         Boston Red Sox         BRO
## 14    1917          CHA         AL    <NA>      Chicago White Sox         NY1
## 15    1918          BOS         AL    <NA>         Boston Red Sox         CHN
## 16    1919          CIN         NL    <NA>        Cincinnati Reds         CHA
## 17    1920          CLE         AL    <NA>      Cleveland Indians         BRO
## 18    1921          NY1         NL    <NA>        New York Giants         NYA
## 19    1922          NY1         NL    <NA>        New York Giants         NYA
## 20    1923          NYA         AL    <NA>       New York Yankees         NY1
## 21    1924          WS1         AL    <NA>    Washington Senators         NY1
## 22    1925          PIT         NL    <NA>     Pittsburgh Pirates         WS1
## 23    1926          SLN         NL    <NA>    St. Louis Cardinals         NYA
## 24    1927          NYA         AL    <NA>       New York Yankees         PIT
## 25    1928          NYA         AL    <NA>       New York Yankees         SLN
## 26    1929          PHA         AL    <NA> Philadelphia Athletics         CHN
## 27    1930          PHA         AL    <NA> Philadelphia Athletics         SLN
## 28    1931          SLN         NL    <NA>    St. Louis Cardinals         PHA
## 29    1932          NYA         AL    <NA>       New York Yankees         CHN
## 30    1933          NY1         NL    <NA>        New York Giants         WS1
## 31    1934          SLN         NL    <NA>    St. Louis Cardinals         DET
## 32    1935          DET         AL    <NA>         Detroit Tigers         CHN
## 33    1936          NYA         AL    <NA>       New York Yankees         NY1
## 34    1937          NYA         AL    <NA>       New York Yankees         NY1
## 35    1938          NYA         AL    <NA>       New York Yankees         CHN
```

```
## 36    1939          NYA         AL    <NA>       New York Yankees         CIN
## 37    1940          CIN         NL    <NA>        Cincinnati Reds         DET
## 38    1941          NYA         AL    <NA>       New York Yankees         BRO
## 39    1942          SLN         NL    <NA>    St. Louis Cardinals         NYA
## 40    1943          NYA         AL    <NA>       New York Yankees         SLN
## 41    1944          SLN         NL    <NA>    St. Louis Cardinals         SLA
## 42    1945          DET         AL    <NA>         Detroit Tigers         CHN
## 43    1946          SLN         NL    <NA>    St. Louis Cardinals         BOS
## 44    1947          NYA         AL    <NA>       New York Yankees         BRO
## 45    1948          CLE         AL    <NA>      Cleveland Indians         BSN
## 46    1949          NYA         AL    <NA>       New York Yankees         BRO
## 47    1950          NYA         AL    <NA>       New York Yankees         PHI
## 48    1951          NYA         AL    <NA>       New York Yankees         NY1
## 49    1952          NYA         AL    <NA>       New York Yankees         BRO
## 50    1953          NYA         AL    <NA>       New York Yankees         BRO
## 51    1954          NY1         NL    <NA>        New York Giants         CLE
## 52    1955          BRO         NL    <NA>       Brooklyn Dodgers         NYA
## 53    1956          NYA         AL    <NA>       New York Yankees         BRO
## 54    1957          ML1         NL    <NA>       Milwaukee Braves         NYA
## 55    1958          NYA         AL    <NA>       New York Yankees         ML1
## 56    1959          LAN         NL    <NA>    Los Angeles Dodgers         CHA
## 57    1960          PIT         NL    <NA>     Pittsburgh Pirates         NYA
## 58    1961          NYA         AL    <NA>       New York Yankees         CIN
## 59    1962          NYA         AL    <NA>       New York Yankees         SFN
## 60    1963          LAN         NL    <NA>    Los Angeles Dodgers         NYA
## 61    1964          SLN         NL    <NA>    St. Louis Cardinals         NYA
## 62    1965          LAN         NL    <NA>    Los Angeles Dodgers         MIN
## 63    1966          BAL         AL    <NA>      Baltimore Orioles         LAN
## 64    1967          SLN         NL    <NA>    St. Louis Cardinals         BOS
## 65    1968          DET         AL    <NA>         Detroit Tigers         SLN
## 66    1969          NYN         NL       E          New York Mets         BAL
## 67    1970          BAL         AL       E      Baltimore Orioles         CIN
## 68    1971          PIT         NL       E     Pittsburgh Pirates         BAL
## 69    1972          OAK         AL       W      Oakland Athletics         CIN
## 70    1973          OAK         AL       W      Oakland Athletics         NYN
## 71    1974          OAK         AL       W      Oakland Athletics         LAN
## 72    1975          CIN         NL       W        Cincinnati Reds         BOS
## 73    1976          CIN         NL       W        Cincinnati Reds         NYA
## 74    1977          NYA         AL       E       New York Yankees         LAN
## 75    1978          NYA         AL       E       New York Yankees         LAN
## 76    1979          PIT         NL       E     Pittsburgh Pirates         BAL
## 77    1980          PHI         NL       E  Philadelphia Phillies         KCA
## 78    1981          LAN         NL       W    Los Angeles Dodgers         NYA
## 79    1982          SLN         NL       E    St. Louis Cardinals         ML4
## 80    1983          BAL         AL       E      Baltimore Orioles         PHI
## 81    1984          DET         AL       E         Detroit Tigers         SDN
## 82    1985          KCA         AL       W     Kansas City Royals         SLN
```

```
## 83    1986          NYN         NL       E          New York Mets         BOS
## 84    1987          MIN         AL       W        Minnesota Twins         SLN
## 85    1988          LAN         NL       W    Los Angeles Dodgers         OAK
## 86    1989          OAK         AL       W      Oakland Athletics         SFN
## 87    1990          CIN         NL       W        Cincinnati Reds         OAK
## 88    1991          MIN         AL       W        Minnesota Twins         ATL
## 89    1992          TOR         AL       E      Toronto Blue Jays         ATL
## 90    1993          TOR         AL       E      Toronto Blue Jays         PHI
## 91    1995          ATL         NL       E         Atlanta Braves         CLE
## 92    1996          NYA         AL       E       New York Yankees         ATL
## 93    1997          FLO         NL       E        Florida Marlins         CLE
## 94    1998          NYA         AL       E       New York Yankees         SDN
## 95    1999          NYA         AL       E       New York Yankees         ATL
## 96    2000          NYA         AL       E       New York Yankees         NYN
## 97    2001          ARI         NL       W   Arizona Diamondbacks         NYA
## 98    2002          ANA         AL       W         Anaheim Angels         SFN
## 99    2003          FLO         NL       E        Florida Marlins         NYA
## 100   2004          BOS         AL       E         Boston Red Sox         SLN
## 101   2005          CHA         AL       C      Chicago White Sox         HOU
## 102   2006          SLN         NL       C    St. Louis Cardinals         DET
## 103   2007          BOS         AL       E         Boston Red Sox         COL
## 104   2008          PHI         NL       E  Philadelphia Phillies         TBA
## 105   2009          NYA         AL       E       New York Yankees         PHI
## 106   2010          SFN         NL       W   San Francisco Giants         TEX
## 107   2011          SLN         NL       C    St. Louis Cardinals         TEX
## 108   2012          SFN         NL       W   San Francisco Giants         DET
## 109   2013          BOS         AL       E         Boston Red Sox         SLN
##     lgIDloser divID.y                 name.y losses
## 1          NL    <NA>     Pittsburgh Pirates      3
## 2          AL    <NA> Philadelphia Athletics      1
## 3          NL    <NA>           Chicago Cubs      2
## 4          AL    <NA>         Detroit Tigers      0
## 5          AL    <NA>         Detroit Tigers      1
## 6          AL    <NA>         Detroit Tigers      3
## 7          NL    <NA>           Chicago Cubs      1
## 8          NL    <NA>        New York Giants      2
## 9          NL    <NA>        New York Giants      3
## 10         NL    <NA>        New York Giants      1
## 11         AL    <NA> Philadelphia Athletics      0
## 12         NL    <NA>  Philadelphia Phillies      1
## 13         NL    <NA>        Brooklyn Robins      1
## 14         NL    <NA>        New York Giants      2
## 15         NL    <NA>           Chicago Cubs      2
## 16         AL    <NA>      Chicago White Sox      3
## 17         NL    <NA>        Brooklyn Robins      2
## 18         AL    <NA>       New York Yankees      3
## 19         AL    <NA>       New York Yankees      0
```

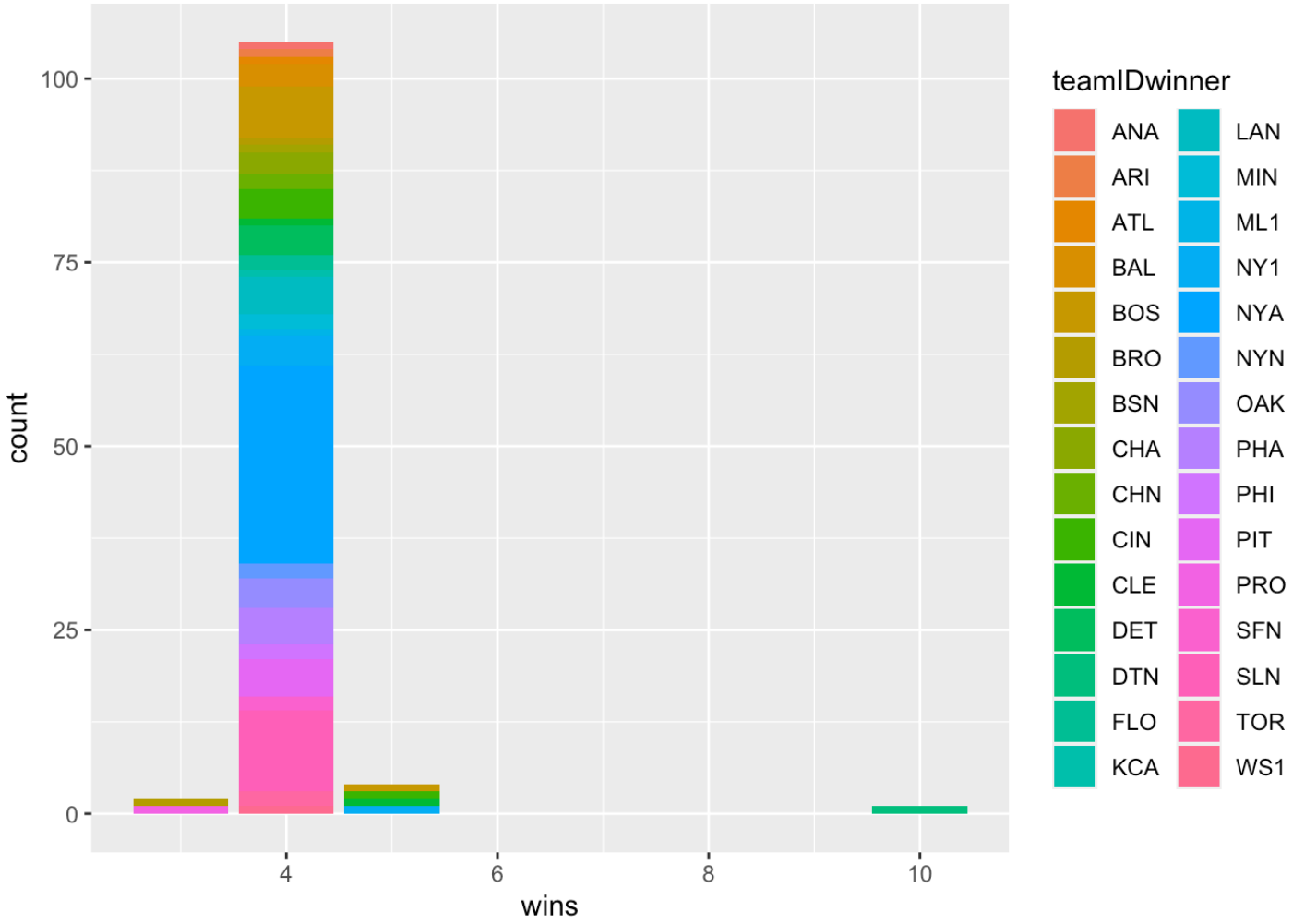
```
## 20          NL    <NA>         New York Giants      2
## 21          NL    <NA>         New York Giants      3
## 22          AL    <NA>     Washington Senators      3
## 23          AL    <NA>        New York Yankees      3
## 24          NL    <NA>      Pittsburgh Pirates      0
## 25          NL    <NA>     St. Louis Cardinals      0
## 26          NL    <NA>            Chicago Cubs      1
## 27          NL    <NA>     St. Louis Cardinals      2
## 28          AL    <NA> Philadelphia Athletics      3
## 29          NL    <NA>            Chicago Cubs      0
## 30          AL    <NA>     Washington Senators      1
## 31          AL    <NA>          Detroit Tigers      3
## 32          NL    <NA>            Chicago Cubs      2
## 33          NL    <NA>         New York Giants      2
## 34          NL    <NA>         New York Giants      1
## 35          NL    <NA>            Chicago Cubs      0
## 36          NL    <NA>         Cincinnati Reds      0
## 37          AL    <NA>          Detroit Tigers      3
## 38          NL    <NA>        Brooklyn Dodgers      1
## 39          AL    <NA>        New York Yankees      1
## 40          NL    <NA>     St. Louis Cardinals      1
## 41          AL    <NA>        St. Louis Browns      2
## 42          NL    <NA>            Chicago Cubs      3
## 43          AL    <NA>          Boston Red Sox      3
## 44          NL    <NA>        Brooklyn Dodgers      3
## 45          NL    <NA>           Boston Braves      2
## 46          NL    <NA>        Brooklyn Dodgers      1
## 47          NL    <NA>  Philadelphia Phillies      0
## 48          NL    <NA>         New York Giants      2
## 49          NL    <NA>        Brooklyn Dodgers      3
## 50          NL    <NA>        Brooklyn Dodgers      2
## 51          AL    <NA>       Cleveland Indians      0
## 52          AL    <NA>        New York Yankees      3
## 53          NL    <NA>        Brooklyn Dodgers      3
## 54          AL    <NA>        New York Yankees      3
## 55          NL    <NA>        Milwaukee Braves      3
## 56          AL    <NA>       Chicago White Sox      2
## 57          AL    <NA>        New York Yankees      3
## 58          NL    <NA>         Cincinnati Reds      1
## 59          NL    <NA>   San Francisco Giants      3
## 60          AL    <NA>        New York Yankees      0
## 61          AL    <NA>        New York Yankees      3
## 62          AL    <NA>         Minnesota Twins      3
## 63          NL    <NA>    Los Angeles Dodgers      0
## 64          AL    <NA>          Boston Red Sox      3
## 65          NL    <NA>     St. Louis Cardinals      3
## 66          AL       E       Baltimore Orioles      1
```

```
## 67          NL       W        Cincinnati Reds      1
## 68          AL       E      Baltimore Orioles      3
## 69          NL       W        Cincinnati Reds      3
## 70          NL       E          New York Mets      3
## 71          NL       W    Los Angeles Dodgers      1
## 72          AL       E         Boston Red Sox      3
## 73          AL       E       New York Yankees      0
## 74          NL       W    Los Angeles Dodgers      2
## 75          NL       W    Los Angeles Dodgers      2
## 76          AL       E      Baltimore Orioles      3
## 77          AL       W     Kansas City Royals      2
## 78          AL       E       New York Yankees      2
## 79          AL       E      Milwaukee Brewers      3
## 80          NL       E  Philadelphia Phillies      1
## 81          NL       W       San Diego Padres      1
## 82          NL       E    St. Louis Cardinals      3
## 83          AL       E         Boston Red Sox      3
## 84          NL       E    St. Louis Cardinals      3
## 85          AL       W      Oakland Athletics      1
## 86          NL       W   San Francisco Giants      0
## 87          AL       W      Oakland Athletics      0
## 88          NL       W         Atlanta Braves      3
## 89          NL       W         Atlanta Braves      2
## 90          NL       E  Philadelphia Phillies      2
## 91          AL       C      Cleveland Indians      2
## 92          NL       E         Atlanta Braves      2
## 93          AL       C      Cleveland Indians      3
## 94          NL       W       San Diego Padres      0
## 95          NL       E         Atlanta Braves      0
## 96          NL       E          New York Mets      1
## 97          AL       E       New York Yankees      3
## 98          NL       W   San Francisco Giants      3
## 99          AL       E       New York Yankees      2
## 100         NL       C    St. Louis Cardinals      0
## 101         NL       C         Houston Astros      0
## 102         AL       C         Detroit Tigers      1
## 103         NL       W       Colorado Rockies      0
## 104         AL       E         Tampa Bay Rays      1
## 105         NL       E  Philadelphia Phillies      2
## 106         AL       W          Texas Rangers      1
## 107         AL       W          Texas Rangers      3
## 108         AL       C         Detroit Tigers      0
## 109         NL       C    St. Louis Cardinals      3
```

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

I extract the wins information by using almost the identity method that I used for the above one, and used ggplot to visualize it, namely, how many rounds did each World Series winner winned in the history.

```
WorldSerieswins = dbGetQuery(baseball,'SELECT SeriesPost.yearID,SeriesPost.teamIDwinner,SeriesPost.lgIDwinner,Teams.divID,Teams.name,SeriesPost.wins FROM SeriesPost LEFT JOIN Teams ON SeriesPost.teamIDwinner = Teams.teamID WHERE SeriesPost.yearID = Teams.yearID AND SeriesPost.round = "WS" ORDER BY SeriesPost.yearID')

graph <- ggplot(WorldSerieswins,aes(x = wins, fill = teamIDwinner)) + geom_bar(position = 'stack')
graph
```



Based on the graph, if the team win four games or more, it's highly likely this team will win the whole series.

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

To find the three highest salaries, I extract information from Master and Salaries, left join both on same playerID with yearID = 2003,and then have is order in the descendant way and take the first three rows.

```
salarytopthree  = dbGetQuery(baseball,'SELECT Salaries.*,Master.* FROM Salaries LEFT
JOIN Master on Salaries.playerID = MASTER.playerID WHERE Salaries.yearID = "2003" ORD
ER BY Salaries.salary DESC LIMIT 3')
salarytopthree
```

```
##   yearID teamID lgID  playerID   salary  playerID birthYear birthMonth birthDay
## 1   2003    TEX   AL rodrial01 22000000 rodrial01      1975          7       27
## 2   2003    BOS   AL ramirma02 20000000 ramirma02      1972          5       30
## 3   2003    TOR   AL delgaca01 18700000 delgaca01      1972          6       25
##   birthCountry birthState     birthCity deathYear deathMonth deathDay
## 1          USA         NY      New York        NA         NA       NA
## 2         D.R.         Di Santo Domingo        NA         NA       NA
## 3         P.R.       <NA>     Aguadilla        NA         NA       NA
##   deathCountry deathState deathCity nameFirst  nameLast          nameGiven
## 1         <NA>       <NA>      <NA>      Alex Rodriguez Alexander Emmanuel
## 2         <NA>       <NA>      <NA>     Manny   Ramirez   Manuel Aristides
## 3         <NA>       <NA>      <NA>    Carlos   Delgado        Carlos Juan
##   weight height bats throws        debut     finalGame  retroID   bbrefID
## 1    225     75    R      R 773643600000 1380085200000 rodra001 rodrial01
## 2    225     72    R      R 746946000000 1302066000000 ramim002 ramirma02
## 3    215     75    L      R 749451600000 1241931600000 delgc001 delgaca01
```

#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.

To find the three highest salaries, I extract information from Master and Salaries, left join both on same playerID with yearID = 2010,and then have is order in the descendant way. For all years, I used the same way except, not setting the yearID, and then, I applied ggplot, and have it converted to numeric display instead of scientific notation.

```
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(ggthemes)
salarypayteam2010 = dbGetQuery(baseball,'SELECT SUM(Salaries.salary) AS teamSalary, S
alaries.yearID,Salaries.teamID FROM Salaries WHERE Salaries.yearID = 2010 GROUP BY Sa
laries.yearID,Salaries.teamID ORDER BY teamSalary DESC ')
salarypayteam2010
```

```
##    teamSalary yearID teamID
## 1   206333389   2010    NYA
## 2   162447333   2010    BOS
## 3   146609000   2010    CHN
## 4   141928379   2010    PHI
## 5   134422942   2010    NYN
## 6   122864928   2010    DET
## 7   105530000   2010    CHA
## 8   104963866   2010    LAA
## 9    98641333   2010    SFN
## 10   97559166   2010    MIN
## 11   95358016   2010    LAN
## 12   93540751   2010    SLN
## 13   92355500   2010    HOU
## 14   86510000   2010    SEA
## 15   84423666   2010    ATL
## 16   84227000   2010    COL
## 17   81612500   2010    BAL
## 18   81108278   2010    MIL
## 19   71923471   2010    TBA
## 20   71761542   2010    CIN
## 21   71405210   2010    KCA
## 22   62234000   2010    TOR
## 23   61400000   2010    WAS
## 24   61203966   2010    CLE
## 25   60718166   2010    ARI
## 26   57029719   2010    FLO
## 27   55254900   2010    OAK
## 28   55250544   2010    TEX
## 29   37799300   2010    SDN
## 30   34943000   2010    PIT
```

```
salarypayteamall = dbGetQuery(baseball,'SELECT SUM(Salaries.salary) AS teamSalary, Sa
laries.yearID,Salaries.teamID FROM Salaries GROUP BY Salaries.yearID,Salaries.teamID
ORDER BY yearID  ')
salarypayteamall
```

```
##     teamSalary yearID teamID
```

```
## 1     14807000   1985    ATL
## 2     11560712   1985    BAL
## 3     10897560   1985    BOS
## 4     14427894   1985    CAL
## 5      9846178   1985    CHA
## 6     12702917   1985    CHN
## 7      8359917   1985    CIN
## 8      6551666   1985    CLE
## 9     10348143   1985    DET
## 10     9993051   1985    HOU
## 11     9321179   1985    KCA
## 12    10967917   1985    LAN
## 13     5764821   1985    MIN
## 14    11284107   1985    ML4
## 15     9470166   1985    MON
## 16    14238204   1985    NYA
## 17    10834762   1985    NYN
## 18     9058606   1985    OAK
## 19    10124966   1985    PHI
## 20     9227500   1985    PIT
## 21    11036583   1985    SDN
## 22     4613000   1985    SEA
## 23     8221714   1985    SFN
## 24    11817083   1985    SLN
## 25     7676500   1985    TEX
## 26     8812550   1985    TOR
## 27    17102786   1986    ATL
## 28    13001258   1986    BAL
## 29    14402239   1986    BOS
## 30    14427258   1986    CAL
## 31    10418819   1986    CHA
## 32    17208165   1986    CHN
## 33    11906388   1986    CIN
## 34     7809500   1986    CLE
## 35    12335714   1986    DET
## 36     9873276   1986    HOU
## 37    13043698   1986    KCA
## 38    14913776   1986    LAN
## 39     8748167   1986    MIN
## 40     9943642   1986    ML4
## 41    11103600   1986    MON
## 42    18494253   1986    NYA
## 43    15393714   1986    NYN
## 44     9779421   1986    OAK
## 45    11590166   1986    PHI
## 46    10843500   1986    PIT
## 47    11380693   1986    SDN
```

```
## 48     5958309   1986    SEA
## 49     8947000   1986    SFN
## 50     9875010   1986    SLN
## 51     6743119   1986    TEX
## 52    12611047   1986    TOR
## 53    16544560   1987    ATL
## 54    13900273   1987    BAL
## 55    10144167   1987    BOS
## 56    12843499   1987    CAL
## 57    10641843   1987    CHA
## 58    14307999   1987    CHN
## 59     9281500   1987    CIN
## 60     8513750   1987    CLE
## 61    12122881   1987    DET
## 62    12608371   1987    HOU
## 63    11828056   1987    KCA
## 64    13675403   1987    LAN
## 65     6397500   1987    MIN
## 66     7293224   1987    ML4
## 67     6942052   1987    MON
## 68    17099714   1987    NYA
## 69    13846714   1987    NYN
## 70    11680839   1987    OAK
## 71    11514233   1987    PHI
## 72     7652000   1987    PIT
## 73    11065796   1987    SDN
## 74     2263500   1987    SEA
## 75     7290000   1987    SFN
## 76    11758000   1987    SLN
## 77      880000   1987    TEX
## 78    10479501   1987    TOR
## 79    12728174   1988    ATL
## 80    13532075   1988    BAL
## 81    13896092   1988    BOS
## 82    11947388   1988    CAL
## 83     6390000   1988    CHA
## 84    13119198   1988    CHN
## 85     8888409   1988    CIN
## 86     8936500   1988    CLE
## 87    12869571   1988    DET
## 88    12286167   1988    HOU
## 89    14556562   1988    KCA
## 90    16850515   1988    LAN
## 91    12462666   1988    MIN
## 92     8402000   1988    ML4
## 93     9603333   1988    MON
## 94    19441152   1988    NYA
```

```
## 95    15269314   1988    NYN
## 96     9690000   1988    OAK
## 97    13838000   1988    PHI
## 98     5998500   1988    PIT
## 99     9561002   1988    SDN
## 100    7342450   1988    SEA
## 101   12380000   1988    SFN
## 102   12880000   1988    SLN
## 103    5342131   1988    TEX
## 104   12241225   1988    TOR
## 105   11112334   1989    ATL
## 106    8275167   1989    BAL
## 107   17481748   1989    BOS
## 108   15097833   1989    CAL
## 109    7265410   1989    CHA
## 110   10668000   1989    CHN
## 111   11072000   1989    CIN
## 112    9094500   1989    CLE
## 113   15146404   1989    DET
## 114   15029500   1989    HOU
## 115   18683568   1989    KCA
## 116   21071562   1989    LAN
## 117   15531666   1989    MIN
## 118   11533000   1989    ML4
## 119   13807389   1989    MON
## 120   17114375   1989    NYA
## 121   19885071   1989    NYN
## 122   15613070   1989    OAK
## 123   10604000   1989    PHI
## 124   12737500   1989    PIT
## 125   14195000   1989    SDN
## 126    9779500   1989    SEA
## 127   14962834   1989    SFN
## 128   16078833   1989    SLN
## 129   11893781   1989    TEX
## 130   16261666   1989    TOR
## 131   14555501   1990    ATL
## 132    9680084   1990    BAL
## 133   20558333   1990    BOS
## 134   21720000   1990    CAL
## 135    9491500   1990    CHA
## 136   13624000   1990    CHN
## 137   14370000   1990    CIN
## 138   14487000   1990    CLE
## 139   17593238   1990    DET
## 140   18330000   1990    HOU
## 141   23361084   1990    KCA
```

```
## 142   21318704   1990    LAN
## 143   14602000   1990    MIN
## 144   19719167   1990    ML4
## 145   16586388   1990    MON
## 146   20912318   1990    NYA
## 147   21722834   1990    NYN
## 148   19887501   1990    OAK
## 149   13173667   1990    PHI
## 150   15556000   1990    PIT
## 151   17588334   1990    SDN
## 152   12553667   1990    SEA
## 153   19335333   1990    SFN
## 154   20523334   1990    SLN
## 155   14874372   1990    TEX
## 156   17756834   1990    TOR
## 157   18403500   1991    ATL
## 158   17519000   1991    BAL
## 159   35167500   1991    BOS
## 160   33060001   1991    CAL
## 161   16919667   1991    CHA
## 162   23175667   1991    CHN
## 163   26305333   1991    CIN
## 164   17635000   1991    CLE
## 165   23838333   1991    DET
## 166   12852500   1991    HOU
## 167   26319834   1991    KCA
## 168   32790664   1991    LAN
## 169   23361833   1991    MIN
## 170   23115500   1991    ML4
## 171   10732333   1991    MON
## 172   27344168   1991    NYA
## 173   32590001   1991    NYN
## 174   36999167   1991    OAK
## 175   22487332   1991    PHI
## 176   23634667   1991    PIT
## 177   22150001   1991    SDN
## 178   15691833   1991    SEA
## 179   30967666   1991    SFN
## 180   21860001   1991    SLN
## 181   18224500   1991    TEX
## 182   19902417   1991    TOR
## 183   34625333   1992    ATL
## 184   23780667   1992    BAL
## 185   43610584   1992    BOS
## 186   34749334   1992    CAL
## 187   30160833   1992    CHA
## 188   29829686   1992    CHN
```

```
## 189   35931499   1992    CIN
## 190    9373044   1992    CLE
## 191   27322834   1992    DET
## 192   15407500   1992    HOU
## 193   33893834   1992    KCA
## 194   44788166   1992    LAN
## 195   28027834   1992    MIN
## 196   31013667   1992    ML4
## 197   15822334   1992    MON
## 198   37543334   1992    NYA
## 199   44602002   1992    NYN
## 200   41035000   1992    OAK
## 201   24383834   1992    PHI
## 202   33944167   1992    PIT
## 203   26854167   1992    SDN
## 204   23179833   1992    SEA
## 205   33163168   1992    SFN
## 206   27583836   1992    SLN
## 207   30128167   1992    TEX
## 208   44788666   1992    TOR
## 209   41641417   1993    ATL
## 210   29096500   1993    BAL
## 211   37120583   1993    BOS
## 212   28588334   1993    CAL
## 213   39696166   1993    CHA
## 214   39386666   1993    CHN
## 215   44879666   1993    CIN
## 216   18561000   1993    CLE
## 217   10353500   1993    COL
## 218   38150165   1993    DET
## 219   19330545   1993    FLO
## 220   30210500   1993    HOU
## 221   41346167   1993    KCA
## 222   39331999   1993    LAN
## 223   28217933   1993    MIN
## 224   23806834   1993    ML4
## 225   18899333   1993    MON
## 226   42624900   1993    NYA
## 227   39043667   1993    NYN
## 228   37812333   1993    OAK
## 229   28538334   1993    PHI
## 230   24822467   1993    PIT
## 231   25511333   1993    SDN
## 232   32696333   1993    SEA
## 233   35050000   1993    SFN
## 234   23367334   1993    SLN
## 235   36376959   1993    TEX
```

```
## 236   47279166   1993    TOR
## 237   49383513   1994    ATL
## 238   38849769   1994    BAL
## 239   37859084   1994    BOS
## 240   25156218   1994    CAL
## 241   39183836   1994    CHA
## 242   36287333   1994    CHN
## 243   40961833   1994    CIN
## 244   30490500   1994    CLE
## 245   23887333   1994    COL
## 246   41446501   1994    DET
## 247   21633000   1994    FLO
## 248   33126000   1994    HOU
## 249   40541334   1994    KCA
## 250   38000001   1994    LAN
## 251   28438500   1994    MIN
## 252   24350500   1994    ML4
## 253   19098000   1994    MON
## 254   45731334   1994    NYA
## 255   30956583   1994    NYN
## 256   34172500   1994    OAK
## 257   31599000   1994    PHI
## 258   24217250   1994    PIT
## 259   14916333   1994    SDN
## 260   29228500   1994    SEA
## 261   42638666   1994    SFN
## 262   29275601   1994    SLN
## 263   32973597   1994    TEX
## 264   43433668   1994    TOR
## 265   47235445   1995    ATL
## 266   43942521   1995    BAL
## 267   32455518   1995    BOS
## 268   31223171   1995    CAL
## 269   46961282   1995    CHA
## 270   29505834   1995    CHN
## 271   43144670   1995    CIN
## 272   37937835   1995    CLE
## 273   34154717   1995    COL
## 274   37044168   1995    DET
## 275   24515781   1995    FLO
## 276   34169834   1995    HOU
## 277   29532834   1995    KCA
## 278   39273201   1995    LAN
## 279   25410500   1995    MIN
## 280   17798825   1995    ML4
## 281   12364000   1995    MON
## 282   48874851   1995    NYA
```

```
## 283   27674992   1995    NYN
## 284   37739225   1995    OAK
## 285   30555945   1995    PHI
## 286   18355345   1995    PIT
## 287   26382334   1995    SDN
## 288   36481311   1995    SEA
## 289   36462777   1995    SFN
## 290   37101000   1995    SLN
## 291   34581451   1995    TEX
## 292   50590000   1995    TOR
## 293   49698500   1996    ATL
## 294   54490315   1996    BAL
## 295   42393500   1996    BOS
## 296   28738000   1996    CAL
## 297   45139500   1996    CHA
## 298   33081000   1996    CHN
## 299   42526334   1996    CIN
## 300   48107360   1996    CLE
## 301   40179823   1996    COL
## 302   23438000   1996    DET
## 303   31022500   1996    FLO
## 304   28487000   1996    HOU
## 305   20281250   1996    KCA
## 306   35355000   1996    LAN
## 307   23117000   1996    MIN
## 308   21730000   1996    ML4
## 309   16264500   1996    MON
## 310   54191792   1996    NYA
## 311   24479500   1996    NYN
## 312   21243000   1996    OAK
## 313   34314500   1996    PHI
## 314   23017500   1996    PIT
## 315   28348172   1996    SDN
## 316   41328501   1996    SEA
## 317   37144725   1996    SFN
## 318   40269667   1996    SLN
## 319   39041528   1996    TEX
## 320   29555083   1996    TOR
## 321   31135472   1997    ANA
## 322   52278500   1997    ATL
## 323   58516400   1997    BAL
## 324   43558750   1997    BOS
## 325   57740000   1997    CHA
## 326   42155333   1997    CHN
## 327   49768000   1997    CIN
## 328   56802460   1997    CLE
## 329   43559667   1997    COL
```

```
## 330   17272000   1997    DET
## 331   48692500   1997    FLO
## 332   34777500   1997    HOU
## 333   34655000   1997    KCA
## 334   45380304   1997    LAN
## 335   34072500   1997    MIN
## 336   23655338   1997    ML4
## 337   19295500   1997    MON
## 338   62241545   1997    NYA
## 339   39800400   1997    NYN
## 340   24018500   1997    OAK
## 341   36656500   1997    PHI
## 342   10771667   1997    PIT
## 343   37363672   1997    SDN
## 344   41540661   1997    SEA
## 345   35592378   1997    SFN
## 346   45456667   1997    SLN
## 347   53448838   1997    TEX
## 348   47079833   1997    TOR
## 349   41281000   1998    ANA
## 350   32347000   1998    ARI
## 351   61186000   1998    ATL
## 352   72355634   1998    BAL
## 353   56757000   1998    BOS
## 354   38335000   1998    CHA
## 355   50838000   1998    CHN
## 356   23005000   1998    CIN
## 357   60800166   1998    CLE
## 358   50484648   1998    COL
## 359   24065000   1998    DET
## 360   41322667   1998    FLO
## 361   42374000   1998    HOU
## 362   36862500   1998    KCA
## 363   48820000   1998    LAN
## 364   33914904   1998    MIL
## 365   27927500   1998    MIN
## 366   10641500   1998    MON
## 367   66806867   1998    NYA
## 368   52077999   1998    NYN
## 369   21303000   1998    OAK
## 370   36297500   1998    PHI
## 371   15065000   1998    PIT
## 372   46861500   1998    SDN
## 373   54087036   1998    SEA
## 374   42565834   1998    SFN
## 375   54672521   1998    SLN
## 376   27280000   1998    TBA
```

```
## 377   56572095   1998    TEX
## 378   51376000   1998    TOR
## 379   55388166   1999    ANA
## 380   68703999   1999    ARI
## 381   73140000   1999    ATL
## 382   80605863   1999    BAL
## 383   63497500   1999    BOS
## 384   25620000   1999    CHA
## 385   62343000   1999    CHN
## 386   33962761   1999    CIN
## 387   72978462   1999    CLE
## 388   61935837   1999    COL
## 389   36489666   1999    DET
## 390   21085000   1999    FLO
## 391   54914000   1999    HOU
## 392   26225000   1999    KCA
## 393   80862453   1999    LAN
## 394   43377395   1999    MIL
## 395   21257500   1999    MIN
## 396   17903000   1999    MON
## 397   86734359   1999    NYA
## 398   65092092   1999    NYN
## 399   24431833   1999    OAK
## 400   31692500   1999    PHI
## 401   24697666   1999    PIT
## 402   49768179   1999    SDN
## 403   54125003   1999    SEA
## 404   46595057   1999    SFN
## 405   49778195   1999    SLN
## 406   38870000   1999    TBA
## 407   76709931   1999    TEX
## 408   45444333   1999    TOR
## 409   51464167   2000    ANA
## 410   81027833   2000    ARI
## 411   84537836   2000    ATL
## 412   81447435   2000    BAL
## 413   77940333   2000    BOS
## 414   31133500   2000    CHA
## 415   60539333   2000    CHN
## 416   46867200   2000    CIN
## 417   75880771   2000    CLE
## 418   61111190   2000    COL
## 419   58265167   2000    DET
## 420   19872000   2000    FLO
## 421   51289111   2000    HOU
## 422   23433000   2000    KCA
## 423   87924286   2000    LAN
```

```
## 424   36505333   2000    MIL
## 425   16519500   2000    MIN
## 426   32994333   2000    MON
## 427   92338260   2000    NYA
## 428   79509776   2000    NYN
## 429   31971333   2000    OAK
## 430   47308000   2000    PHI
## 431   28928334   2000    PIT
## 432   54821000   2000    SDN
## 433   58915000   2000    SEA
## 434   53737826   2000    SFN
## 435   61453863   2000    SLN
## 436   62765129   2000    TBA
## 437   70795921   2000    TEX
## 438   44838332   2000    TOR
## 439   47535167   2001    ANA
## 440   85082999   2001    ARI
## 441   91936166   2001    ATL
## 442   67599540   2001    BAL
## 443  110035833   2001    BOS
## 444   65653667   2001    CHA
## 445   64715833   2001    CHN
## 446   48986000   2001    CIN
## 447   93152001   2001    CLE
## 448   71541334   2001    COL
## 449   53416167   2001    DET
## 450   35762500   2001    FLO
## 451   60612667   2001    HOU
## 452   35422500   2001    KCA
## 453  109105953   2001    LAN
## 454   43886833   2001    MIL
## 455   24130000   2001    MIN
## 456   35159500   2001    MON
## 457  112287143   2001    NYA
## 458   93174428   2001    NYN
## 459   33810750   2001    OAK
## 460   41663833   2001    PHI
## 461   57760833   2001    PIT
## 462   39182833   2001    SDN
## 463   74720834   2001    SEA
## 464   63280167   2001    SFN
## 465   78538333   2001    SLN
## 466   56980000   2001    TBA
## 467   88633500   2001    TEX
## 468   76895999   2001    TOR
## 469   61721667   2002    ANA
## 470  102819999   2002    ARI
```

```
## 471   92870367   2002    ATL
## 472   60493487   2002    BAL
## 473  108366060   2002    BOS
## 474   57052833   2002    CHA
## 475   75690833   2002    CHN
## 476   45050390   2002    CIN
## 477   78909449   2002    CLE
## 478   56851043   2002    COL
## 479   55048000   2002    DET
## 480   41979917   2002    FLO
## 481   63448417   2002    HOU
## 482   47257000   2002    KCA
## 483   94850953   2002    LAN
## 484   50287833   2002    MIL
## 485   40425000   2002    MIN
## 486   38670500   2002    MON
## 487  125928583   2002    NYA
## 488   94633593   2002    NYN
## 489   40004167   2002    OAK
## 490   57954999   2002    PHI
## 491   42323599   2002    PIT
## 492   41425000   2002    SDN
## 493   80282668   2002    SEA
## 494   78299835   2002    SFN
## 495   74660875   2002    SLN
## 496   34380000   2002    TBA
## 497  105526122   2002    TEX
## 498   76864333   2002    TOR
## 499   79031667   2003    ANA
## 500   80657000   2003    ARI
## 501  106243667   2003    ATL
## 502   73877500   2003    BAL
## 503   99946500   2003    BOS
## 504   51010000   2003    CHA
## 505   79868333   2003    CHN
## 506   59355667   2003    CIN
## 507   48584834   2003    CLE
## 508   67179667   2003    COL
## 509   49168000   2003    DET
## 510   49450000   2003    FLO
## 511   71040000   2003    HOU
## 512   40518000   2003    KCA
## 513  105572620   2003    LAN
## 514   40627000   2003    MIL
## 515   55505000   2003    MIN
## 516   51948500   2003    MON
## 517  152749814   2003    NYA
```

```
## 518  116876429   2003    NYN
## 519   50260834   2003    OAK
## 520   70780000   2003    PHI
## 521   54812429   2003    PIT
## 522   45210000   2003    SDN
## 523   86959167   2003    SEA
## 524   82852167   2003    SFN
## 525   83786666   2003    SLN
## 526   19630000   2003    TBA
## 527  103491667   2003    TEX
## 528   51269000   2003    TOR
## 529  100534667   2004    ANA
## 530   69780750   2004    ARI
## 531   90182500   2004    ATL
## 532   51623333   2004    BAL
## 533  127298500   2004    BOS
## 534   65212500   2004    CHA
## 535   90560000   2004    CHN
## 536   46615250   2004    CIN
## 537   34319300   2004    CLE
## 538   65445167   2004    COL
## 539   46832000   2004    DET
## 540   42143042   2004    FLO
## 541   75397000   2004    HOU
## 542   47609000   2004    KCA
## 543   92902001   2004    LAN
## 544   27528500   2004    MIL
## 545   53585000   2004    MIN
## 546   40897500   2004    MON
## 547  184193950   2004    NYA
## 548   96660970   2004    NYN
## 549   59425667   2004    OAK
## 550   92919167   2004    PHI
## 551   32227929   2004    PIT
## 552   55384833   2004    SDN
## 553   81515834   2004    SEA
## 554   82019166   2004    SFN
## 555   83228333   2004    SLN
## 556   29556667   2004    TBA
## 557   55050417   2004    TEX
## 558   50017000   2004    TOR
## 559   62329166   2005    ARI
## 560   86457302   2005    ATL
## 561   73914333   2005    BAL
## 562  123505125   2005    BOS
## 563   75178000   2005    CHA
## 564   87032933   2005    CHN
```

```
## 565   61892583   2005    CIN
## 566   41502500   2005    CLE
## 567   47839000   2005    COL
## 568   69092000   2005    DET
## 569   60408834   2005    FLO
## 570   76779000   2005    HOU
## 571   36881000   2005    KCA
## 572   94867822   2005    LAA
## 573   83039000   2005    LAN
## 574   39934833   2005    MIL
## 575   56186000   2005    MIN
## 576  208306817   2005    NYA
## 577  101305821   2005    NYN
## 578   55425762   2005    OAK
## 579   95522000   2005    PHI
## 580   38133000   2005    PIT
## 581   63290833   2005    SDN
## 582   87754334   2005    SEA
## 583   90199500   2005    SFN
## 584   92106833   2005    SLN
## 585   29679067   2005    TBA
## 586   55849000   2005    TEX
## 587   45719500   2005    TOR
## 588   48581500   2005    WAS
## 589   59684226   2006    ARI
## 590   90156876   2006    ATL
## 591   72585582   2006    BAL
## 592  120099824   2006    BOS
## 593  102750667   2006    CHA
## 594   94424499   2006    CHN
## 595   60909519   2006    CIN
## 596   56031500   2006    CLE
## 597   41233000   2006    COL
## 598   82612866   2006    DET
## 599   14671500   2006    FLO
## 600   88694435   2006    HOU
## 601   47294000   2006    KCA
## 602  103472000   2006    LAA
## 603   98447187   2006    LAN
## 604   57568333   2006    MIL
## 605   63396006   2006    MIN
## 606  194663079   2006    NYA
## 607  101084963   2006    NYN
## 608   62243079   2006    OAK
## 609   88273333   2006    PHI
## 610   46717750   2006    PIT
## 611   69896141   2006    SDN
```

```
## 612   87959833   2006    SEA
## 613   90056419   2006    SFN
## 614   88891371   2006    SLN
## 615   34917967   2006    TBA
## 616   68228662   2006    TEX
## 617   71365000   2006    TOR
## 618   63143000   2006    WAS
## 619   52067546   2007    ARI
## 620   87290833   2007    ATL
## 621   93174808   2007    BAL
## 622  143026214   2007    BOS
## 623  108671833   2007    CHA
## 624   99670332   2007    CHN
## 625   68524980   2007    CIN
## 626   61673267   2007    CLE
## 627   54041000   2007    COL
## 628   94800369   2007    DET
## 629   30507000   2007    FLO
## 630   87759000   2007    HOU
## 631   67116500   2007    KCA
## 632  109251333   2007    LAA
## 633  108454524   2007    LAN
## 634   70986500   2007    MIL
## 635   71439500   2007    MIN
## 636  189259045   2007    NYA
## 637  115231663   2007    NYN
## 638   79366940   2007    OAK
## 639   89428213   2007    PHI
## 640   38537833   2007    PIT
## 641   58110567   2007    SDN
## 642  106460833   2007    SEA
## 643   90219056   2007    SFN
## 644   90286823   2007    SLN
## 645   24123500   2007    TBA
## 646   68318675   2007    TEX
## 647   81942800   2007    TOR
## 648   36947500   2007    WAS
## 649   66202712   2008    ARI
## 650  102365683   2008    ATL
## 651   67196246   2008    BAL
## 652  133390035   2008    BOS
## 653  121189332   2008    CHA
## 654  118345833   2008    CHN
## 655   74117695   2008    CIN
## 656   78970066   2008    CLE
## 657   68655500   2008    COL
## 658  137685196   2008    DET
```

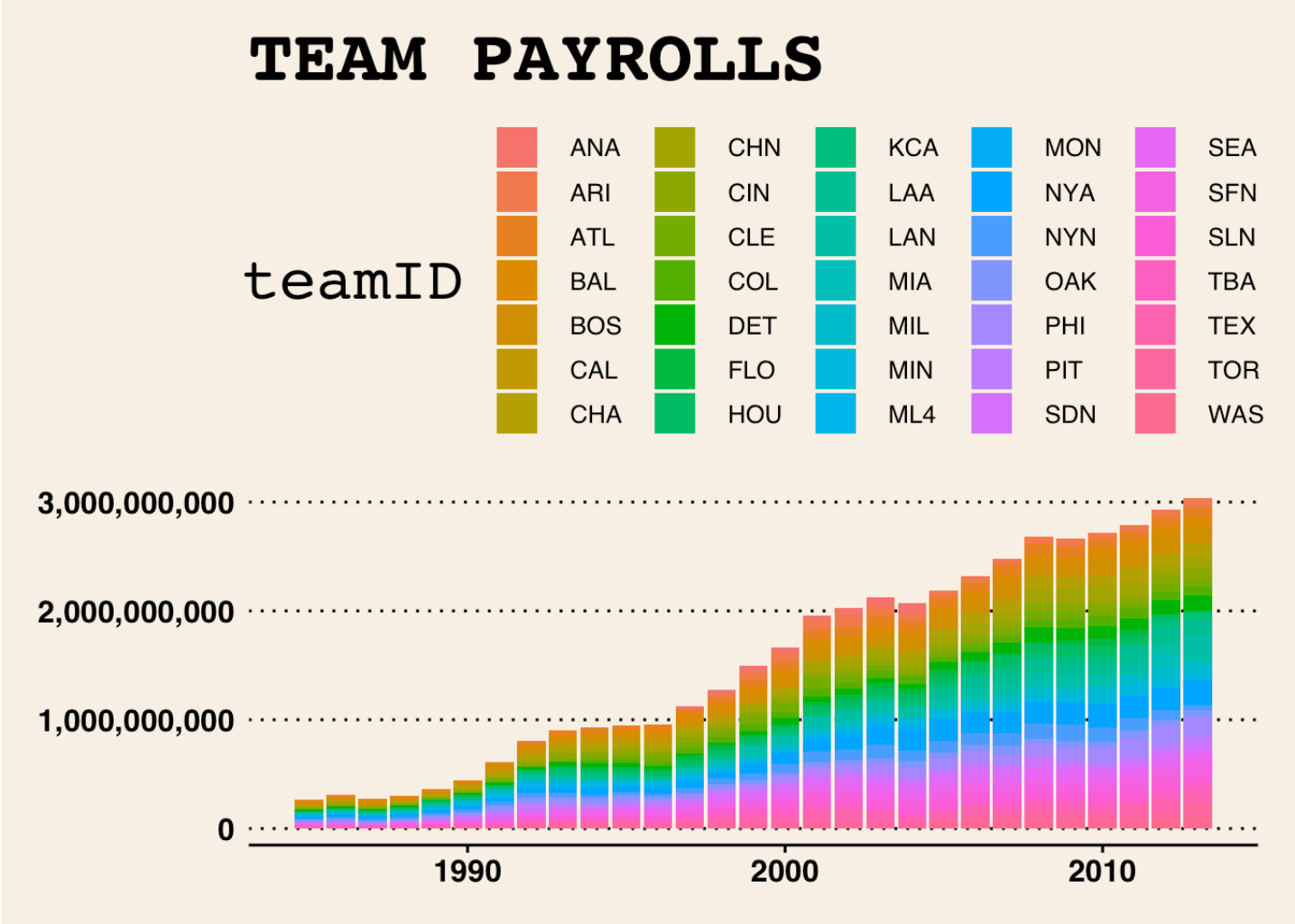
```
## 659   21811500   2008    FLO
## 660   88930414   2008    HOU
## 661   58245500   2008    KCA
## 662  119216333   2008    LAA
## 663  118588536   2008    LAN
## 664   80937499   2008    MIL
## 665   56932766   2008    MIN
## 666  207896789   2008    NYA
## 667  137793376   2008    NYN
## 668   47967126   2008    OAK
## 669   97879880   2008    PHI
## 670   48689783   2008    PIT
## 671   73677616   2008    SDN
## 672  117666482   2008    SEA
## 673   76594500   2008    SFN
## 674   99624449   2008    SLN
## 675   43820597   2008    TBA
## 676   67712326   2008    TEX
## 677   97793900   2008    TOR
## 678   54961000   2008    WAS
## 679   73115666   2009    ARI
## 680   96726166   2009    ATL
## 681   67101666   2009    BAL
## 682  121345999   2009    BOS
## 683   96068500   2009    CHA
## 684  134809000   2009    CHN
## 685   73558500   2009    CIN
## 686   81579166   2009    CLE
## 687   75201000   2009    COL
## 688  115085145   2009    DET
## 689   36834000   2009    FLO
## 690  102996414   2009    HOU
## 691   70519333   2009    KCA
## 692  113709000   2009    LAA
## 693  100414592   2009    LAN
## 694   80182502   2009    MIL
## 695   65299266   2009    MIN
## 696  201449189   2009    NYA
## 697  149373987   2009    NYN
## 698   61910000   2009    OAK
## 699  113004046   2009    PHI
## 700   48693000   2009    PIT
## 701   43333700   2009    SDN
## 702   98904166   2009    SEA
## 703   83026450   2009    SFN
## 704   88528409   2009    SLN
## 705   63313034   2009    TBA
```

```
## 706   68178798   2009    TEX
## 707   80538300   2009    TOR
## 708   59928000   2009    WAS
## 709   60718166   2010    ARI
## 710   84423666   2010    ATL
## 711   81612500   2010    BAL
## 712  162447333   2010    BOS
## 713  105530000   2010    CHA
## 714  146609000   2010    CHN
## 715   71761542   2010    CIN
## 716   61203966   2010    CLE
## 717   84227000   2010    COL
## 718  122864928   2010    DET
## 719   57029719   2010    FLO
## 720   92355500   2010    HOU
## 721   71405210   2010    KCA
## 722  104963866   2010    LAA
## 723   95358016   2010    LAN
## 724   81108278   2010    MIL
## 725   97559166   2010    MIN
## 726  206333389   2010    NYA
## 727  134422942   2010    NYN
## 728   55254900   2010    OAK
## 729  141928379   2010    PHI
## 730   34943000   2010    PIT
## 731   37799300   2010    SDN
## 732   86510000   2010    SEA
## 733   98641333   2010    SFN
## 734   93540751   2010    SLN
## 735   71923471   2010    TBA
## 736   55250544   2010    TEX
## 737   62234000   2010    TOR
## 738   61400000   2010    WAS
## 739   53639833   2011    ARI
## 740   87002692   2011    ATL
## 741   85304038   2011    BAL
## 742  161762475   2011    BOS
## 743  127789000   2011    CHA
## 744  125047329   2011    CHN
## 745   75947134   2011    CIN
## 746   48776566   2011    CLE
## 747   88148071   2011    COL
## 748  105700231   2011    DET
## 749   56944000   2011    FLO
## 750   70694000   2011    HOU
## 751   35712000   2011    KCA
## 752  138543166   2011    LAA
```

```
## 753  104188999   2011    LAN
## 754   85497333   2011    MIL
## 755  112737000   2011    MIN
## 756  202275028   2011    NYA
## 757  118847309   2011    NYN
## 758   66536500   2011    OAK
## 759  172976379   2011    PHI
## 760   45047000   2011    PIT
## 761   45869140   2011    SDN
## 762   86110600   2011    SEA
## 763  118198333   2011    SFN
## 764  105433572   2011    SLN
## 765   41053571   2011    TBA
## 766   92299264   2011    TEX
## 767   62567800   2011    TOR
## 768   63856928   2011    WAS
## 769   73804833   2012    ARI
## 770   82829942   2012    ATL
## 771   77353999   2012    BAL
## 772  173186617   2012    BOS
## 773   96919500   2012    CHA
## 774   88197033   2012    CHN
## 775   82203616   2012    CIN
## 776   78430300   2012    CLE
## 777   78069571   2012    COL
## 778  132300000   2012    DET
## 779   60651000   2012    HOU
## 780   60916225   2012    KCA
## 781  154485166   2012    LAA
## 782   95143575   2012    LAN
## 783  118078000   2012    MIA
## 784   97653944   2012    MIL
## 785   94085000   2012    MIN
## 786  196522289   2012    NYA
## 787   93353983   2012    NYN
## 788   55372500   2012    OAK
## 789  174538938   2012    PHI
## 790   62951999   2012    PIT
## 791   55244700   2012    SDN
## 792   81978100   2012    SEA
## 793  117620683   2012    SFN
## 794  110300862   2012    SLN
## 795   64173500   2012    TBA
## 796  120510974   2012    TEX
## 797   75009200   2012    TOR
## 798   80855143   2012    WAS
## 799   90132000   2013    ARI
```

```
## 800   87871525   2013    ATL
## 801   84393333   2013    BAL
## 802  151530000   2013    BOS
## 803  120065277   2013    CHA
## 804  100567726   2013    CHN
## 805  106404462   2013    CIN
## 806   75771800   2013    CLE
## 807   74409071   2013    COL
## 808  145989500   2013    DET
## 809   17890700   2013    HOU
## 810   80091725   2013    KCA
## 811  124174750   2013    LAA
## 812  223362196   2013    LAN
## 813   33601900   2013    MIA
## 814   76947033   2013    MIL
## 815   75337500   2013    MIN
## 816  231978886   2013    NYA
## 817   49448346   2013    NYN
## 818   60132500   2013    OAK
## 819  169863189   2013    PHI
## 820   77062000   2013    PIT
## 821   65585500   2013    SDN
## 822   74005043   2013    SEA
## 823  140180334   2013    SFN
## 824   92260110   2013    SLN
## 825   52955272   2013    TBA
## 826  112522600   2013    TEX
## 827  126288100   2013    TOR
## 828  113703270   2013    WAS
```

```
graph_aggregatebyyear <- ggplot(salarypayteamall,aes(x = yearID, y = teamSalary, fill
= teamID)) + geom_bar(stat="identity")+scale_y_continuous(labels = comma)+theme_wsj()
+ scale_colour_wsj("colors6")+ggtitle('TEAM PAYROLLS')
graph_aggregatebyyear
```



#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.

I added two another dataset, one for world Series win teams from 1985 to 2013 since I found out the salary information only recorded from 1985, one for salary information. Then, I joined two datasets with same teamID and yearID by using function from dplyr package in R, and graph it. To combined two graph together, one for total teams add-up payrolls and one for winner team payrolls, I created a new dataset, and included both tables information, and graph it as it from two categorical for each year.

```
library("ggpubr")
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

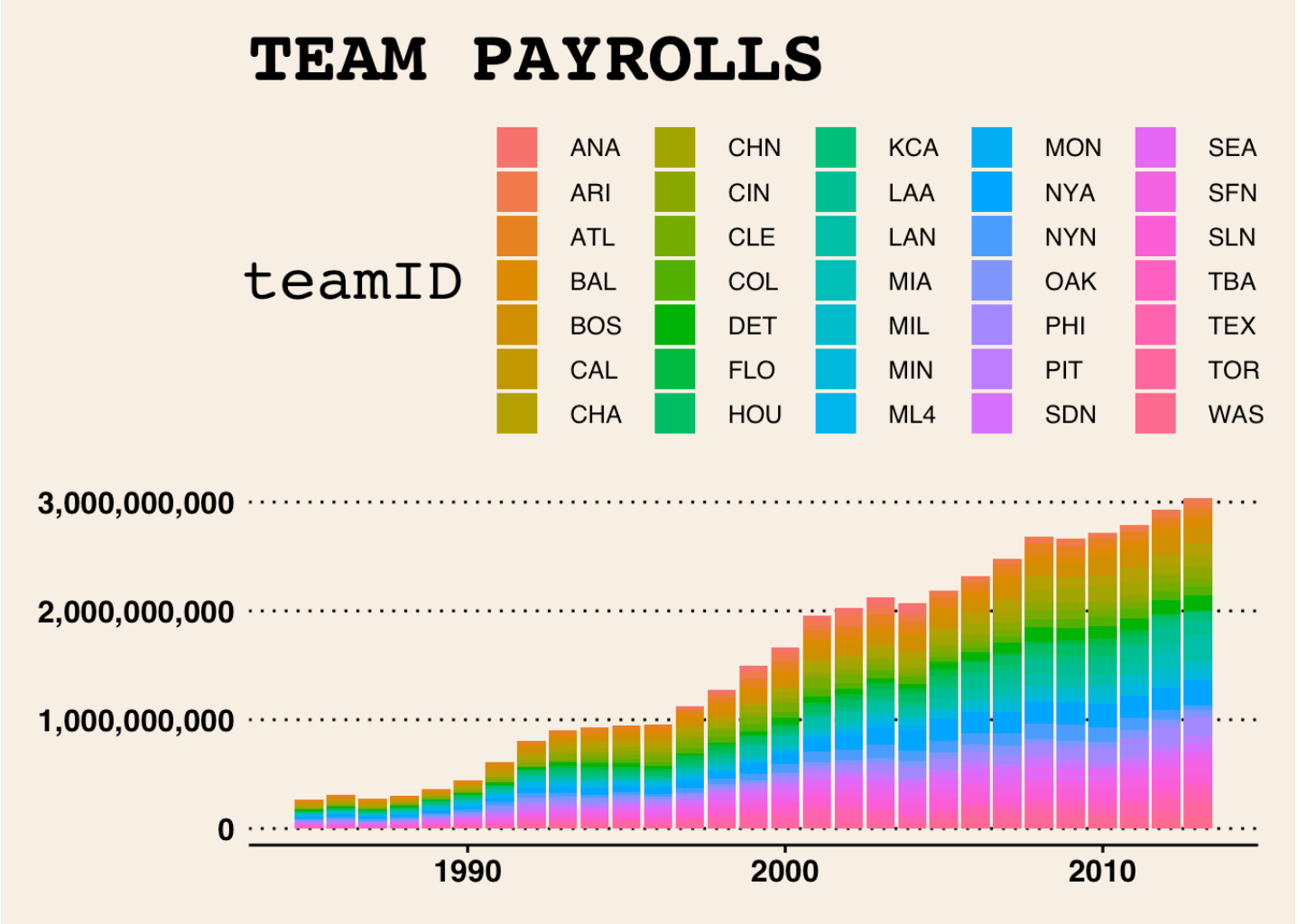
```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

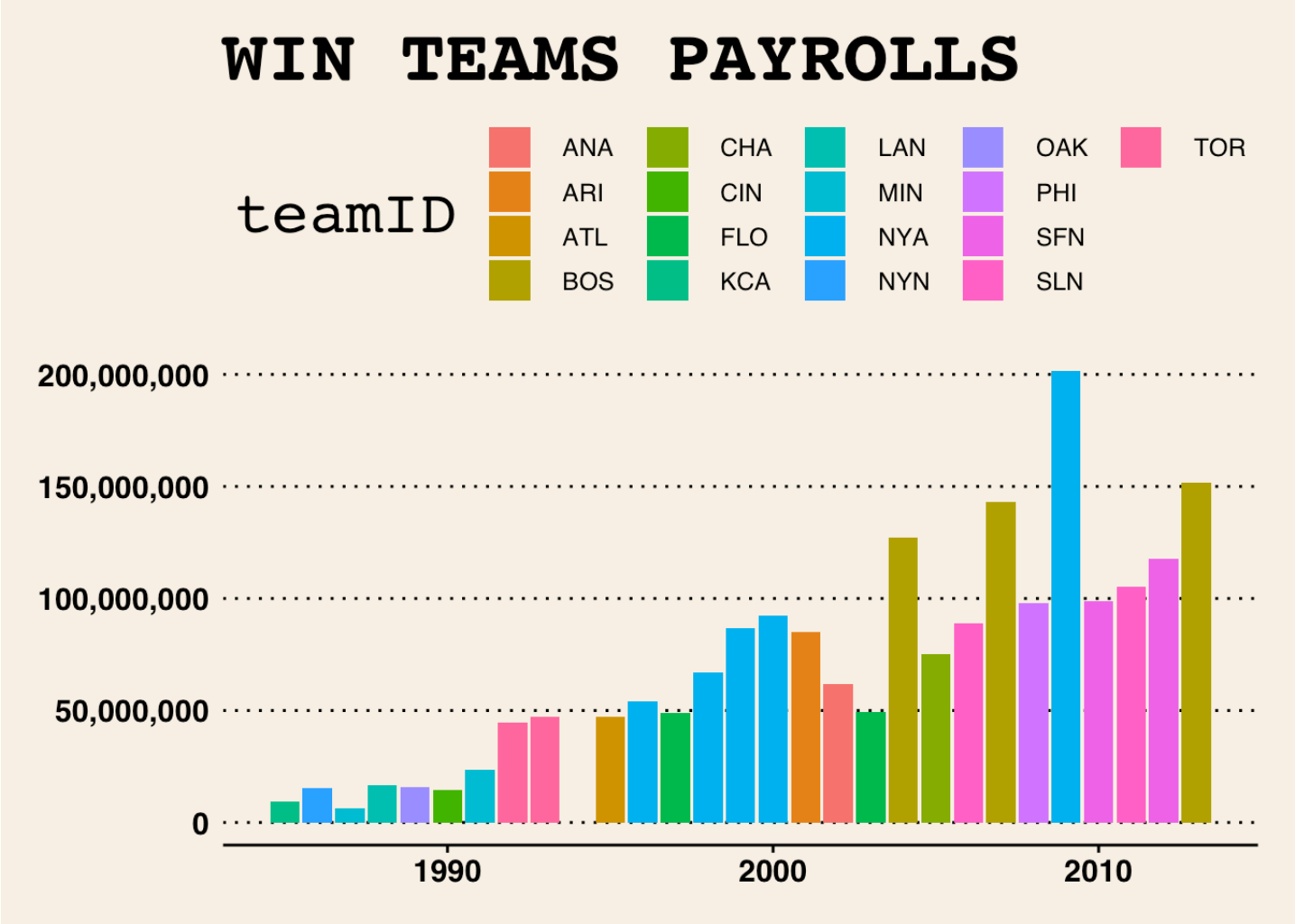
```
##
## Attaching package: 'forecast'
```

```
## The following object is masked from 'package:ggpubr':
##
##     gghistogram
```

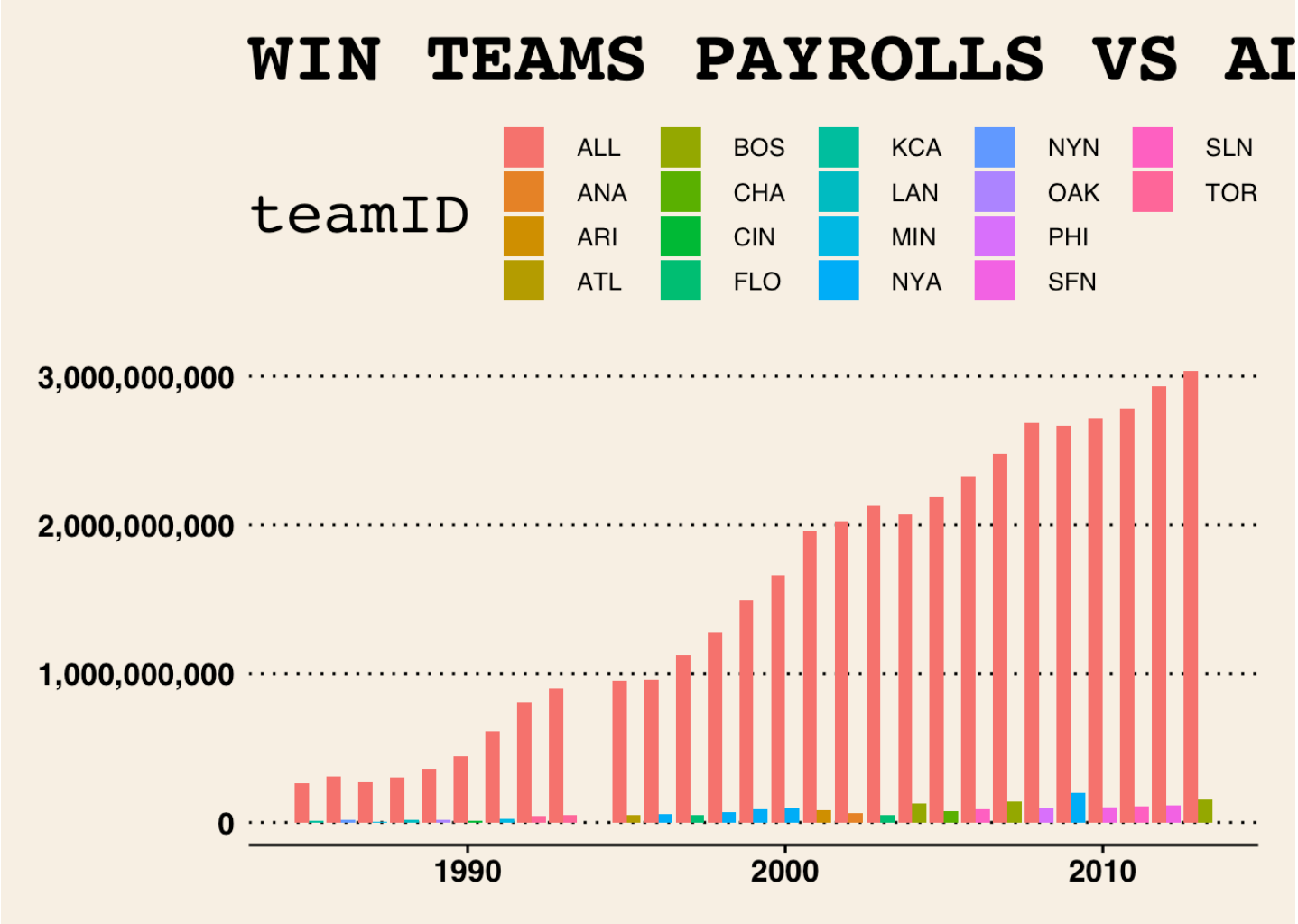
```
graph_aggregatebyyear <- ggplot(salarypayteamall,aes(x = yearID, y = teamSalary, fill
= teamID)) + geom_bar(stat="identity")+scale_y_continuous(labels = comma)+theme_wsj()
+ scale_colour_wsj("colors6")+ggtitle('TEAM PAYROLLS')


WorldSerieswinsteam <- dbGetQuery(baseball,'SELECT SeriesPost.yearID,SeriesPost.teamI
Dwinner,SeriesPost.lgIDwinner FROM SeriesPost WHERE round = "WS"')

WorldSerieswinsteam1985_2013 <- subset(WorldSerieswinsteam,WorldSerieswinsteam$yearID
>= '1985')
rename_WorldSerieswinsteam1985_2013 <- rename(WorldSerieswinsteam1985_2013, teamID =
teamIDwinner)


joined <- inner_join(salarypayteamall, rename_WorldSerieswinsteam1985_2013, by = c("t
eamID" = "teamID", "yearID" = "yearID"))


joined_graph = ggplot(joined,aes(x = yearID, y = teamSalary, fill = teamID)) + geom_b
ar(stat="identity")+scale_y_continuous(labels = comma)+theme_wsj()+ scale_colour_wsj(
"colors6")+ggtitle('WIN TEAMS PAYROLLS')

aggregatebyyear = aggregate(teamSalary ~ yearID, data = salarypayteamall, FUN = sum)
joined2 <- inner_join(joined, aggregatebyyear, by = c("yearID" = "yearID"))


joined22 <- select(joined2,c(yearID, teamSalary.y))
joined22$teamID <- c('ALL')
joined33 <- select(joined2,c(yearID, teamID,teamSalary.x))
renamed_joined33 <- rename(joined33,teamSalary.y = teamSalary.x)
combinedjoined <- rbind(joined22,renamed_joined33)

joined_graph2 = ggplot(combinedjoined,aes(x = yearID, y = teamSalary.y, fill = teamID
))+geom_bar(position="dodge", stat="identity")+scale_y_continuous(labels = comma) + t
heme_wsj()+ scale_colour_wsj("colors6")+ggtitle('WIN TEAMS PAYROLLS VS ALL TEAMS PAYR
OLLS')
graph_aggregatebyyear
```




```
joined_graph
```




```
joined_graph2
```



As it can be seen from the graph, if the team win the world series, it's will have a good amount of salary, but still not that impressive compared to payrolls for all team.

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

First, invetigated on the description, home runs are indicated in Batting and Pitching table. Using SQLite to extract infomration from these two tables and left join with Master tables for player names.Then, I found out find the max for each year is diffcult to apply in SQLite, I used aggregate function in R and inner_join two dataset one same values.

```
homerunsbatting <- dbGetQuery(baseball, 'SELECT Batting.playerID,Batting.yearID,Batti
ng.HR,Master.nameFirst,Master.nameLast FROM Batting LEFT JOIN Master on Batting.playe
rID = Master.playerID GROUP BY Batting.playerID,Batting.yearID ORDER BY Batting.HR DE
SC')

homerunspitching <- dbGetQuery(baseball, 'SELECT Pitching.playerID,Pitching.yearID,Pi
tching.HR,Master.nameFirst,Master.nameLast FROM Pitching LEFT JOIN Master on Pitching
.playerID = Master.playerID GROUP BY Pitching.playerID,Pitching.yearID ORDER BY Pitch
ing.HR DESC')


aggregate_max_batting <- aggregate(HR ~ yearID, data =homerunsbatting , max)
joined_aggregate_max_batting <- inner_join(aggregate_max_batting, homerunsbatting , b
y = c("HR" = "HR", "yearID" = "yearID"))

aggregate_max_pitching <- aggregate(HR ~ yearID, data =homerunsbatting , max)
joined_aggregate_max_pitching <- inner_join(aggregate_max_pitching, homerunspitching
, by = c("HR" = "HR", "yearID" = "yearID"))

joined_aggregate_max_batting
```

```
##     yearID HR  playerID nameFirst    nameLast
## 1     1871  4 meyerle01      Levi     Meyerle
## 2     1871  4  pikeli01       Lip        Pike
## 3     1871  4 treacfr01      Fred     Treacey
## 4     1872  6  pikeli01       Lip        Pike
## 5     1873  4  pikeli01       Lip        Pike
## 6     1874  5 orourji01       Jim    O'Rourke
## 7     1875  6 orourji01       Jim    O'Rourke
## 8     1876  5  hallge01    George        Hall
## 9     1877  4  pikeli01       Lip        Pike
## 10    1878  4 hinespa01      Paul       Hines
## 11    1879  9 jonesch01   Charley       Jones
## 12    1880  6 orourji01       Jim    O'Rourke
## 13    1880  6 stoveha01     Harry      Stovey
## 14    1881  8 broutda01       Dan   Brouthers
## 15    1882  7 walkeos01     Oscar      Walker
## 16    1882  7  woodge01    George        Wood
## 17    1883 14 stoveha01     Harry      Stovey
## 18    1884 27 willine01       Ned  Williamson
## 19    1885 13 stoveha01     Harry      Stovey
## 20    1886 11 broutda01       Dan   Brouthers
## 21    1886 11 richaha01     Hardy  Richardson
## 22    1887 19 obriebi01     Billy     O'Brien
## 23    1888 16  ryanji01     Jimmy        Ryan
## 24    1889 20 thompsa01       Sam    Thompson
```

```
## 25    1890 14 connoro01     Roger      Connor
## 26    1891 16 stoveha01     Harry      Stovey
## 27    1891 16 tiernmi01      Mike     Tiernan
## 28    1892 13 hollibu01       Bug    Holliday
## 29    1893 19 delahed01        Ed   Delahanty
## 30    1894 18 duffyhu01      Hugh       Duffy
## 31    1895 18 thompsa01       Sam    Thompson
## 32    1896 13 delahed01        Ed   Delahanty
## 33    1897 11 duffyhu01      Hugh       Duffy
## 34    1898 15 colliji01     Jimmy     Collins
## 35    1899 25 freembu01      Buck     Freeman
## 36    1900 12  longhe01    Herman        Long
## 37    1901 16 crawfsa01       Sam    Crawford
## 38    1902 16 seyboso01     Socks     Seybold
## 39    1903 13 freembu01      Buck     Freeman
## 40    1904 10 davisha01     Harry       Davis
## 41    1905  9 odwelfr01      Fred      Odwell
## 42    1906 12 davisha01     Harry       Davis
## 43    1906 12 jordati01       Tim      Jordan
## 44    1907 10 brainda01      Dave       Brain
## 45    1908 12 jordati01       Tim      Jordan
## 46    1909  9  cobbty01        Ty        Cobb
## 47    1910 10  beckfr02      Fred        Beck
## 48    1910 10 schulfr01     Frank     Schulte
## 49    1910 10 stahlja01      Jake       Stahl
## 50    1911 21 schulfr01     Frank     Schulte
## 51    1912 14 zimmehe01    Heinie   Zimmerman
## 52    1913 19 cravaga01     Gavvy     Cravath
## 53    1914 19 cravaga01     Gavvy     Cravath
## 54    1915 24 cravaga01     Gavvy     Cravath
## 55    1916 12  pippwa01     Wally        Pipp
## 56    1916 12 roberda01      Dave   Robertson
## 57    1916 12 willicy01        Cy    Williams
## 58    1917 12 cravaga01     Gavvy     Cravath
## 59    1917 12 roberda01      Dave   Robertson
## 60    1918 11  ruthba01      Babe        Ruth
## 61    1918 11 walketi01    Tillie      Walker
## 62    1919 29  ruthba01      Babe        Ruth
## 63    1920 54  ruthba01      Babe        Ruth
## 64    1921 59  ruthba01      Babe        Ruth
## 65    1922 42 hornsro01    Rogers     Hornsby
## 66    1923 41  ruthba01      Babe        Ruth
## 67    1923 41 willicy01        Cy    Williams
## 68    1924 46  ruthba01      Babe        Ruth
## 69    1925 39 hornsro01    Rogers     Hornsby
## 70    1926 47  ruthba01      Babe        Ruth
## 71    1927 60  ruthba01      Babe        Ruth
```

```
## 72    1928 54  ruthba01      Babe        Ruth
## 73    1929 46  ruthba01      Babe        Ruth
## 74    1930 56 wilsoha01      Hack      Wilson
## 75    1931 46 gehrilo01       Lou      Gehrig
## 76    1931 46  ruthba01      Babe        Ruth
## 77    1932 58  foxxji01    Jimmie        Foxx
## 78    1933 48  foxxji01    Jimmie        Foxx
## 79    1934 49 gehrilo01       Lou      Gehrig
## 80    1935 36  foxxji01    Jimmie        Foxx
## 81    1935 36 greenha01      Hank   Greenberg
## 82    1936 49 gehrilo01       Lou      Gehrig
## 83    1937 46 dimagjo01       Joe    DiMaggio
## 84    1938 58 greenha01      Hank   Greenberg
## 85    1939 35  foxxji01    Jimmie        Foxx
## 86    1940 43  mizejo01    Johnny        Mize
## 87    1941 37 willite01       Ted    Williams
## 88    1942 36 willite01       Ted    Williams
## 89    1943 34  yorkru01      Rudy        York
## 90    1944 33 nichobi01      Bill   Nicholson
## 91    1945 28 holmeto01     Tommy      Holmes
## 92    1946 44 greenha01      Hank   Greenberg
## 93    1947 51 kinerra01     Ralph       Kiner
## 94    1947 51  mizejo01    Johnny        Mize
## 95    1948 40 kinerra01     Ralph       Kiner
## 96    1948 40  mizejo01    Johnny        Mize
## 97    1949 54 kinerra01     Ralph       Kiner
## 98    1950 47 kinerra01     Ralph       Kiner
## 99    1951 42 kinerra01     Ralph       Kiner
## 100   1952 37 kinerra01     Ralph       Kiner
## 101   1952 37 sauerha01      Hank       Sauer
## 102   1953 47 matheed01     Eddie     Mathews
## 103   1954 49 kluszte01       Ted  Kluszewski
## 104   1955 51  mayswi01    Willie        Mays
## 105   1956 52 mantlmi01    Mickey      Mantle
## 106   1957 44 aaronha01      Hank       Aaron
## 107   1958 47 bankser01     Ernie       Banks
## 108   1959 46 matheed01     Eddie     Mathews
## 109   1960 41 bankser01     Ernie       Banks
## 110   1961 61 marisro01     Roger       Maris
## 111   1962 49  mayswi01    Willie        Mays
## 112   1963 45 killeha01    Harmon   Killebrew
## 113   1964 49 killeha01    Harmon   Killebrew
## 114   1965 52  mayswi01    Willie        Mays
## 115   1966 49 robinfr02     Frank    Robinson
## 116   1967 44 killeha01    Harmon   Killebrew
## 117   1967 44 yastrca01      Carl Yastrzemski
## 118   1968 44 howarfr01     Frank      Howard
```

```
## 119   1969 49 killeha01    Harmon   Killebrew
## 120   1970 45 benchjo01    Johnny       Bench
## 121   1971 48 stargwi01    Willie    Stargell
## 122   1972 40 benchjo01    Johnny       Bench
## 123   1973 44 stargwi01    Willie    Stargell
## 124   1974 36 schmimi01      Mike     Schmidt
## 125   1975 38 schmimi01      Mike     Schmidt
## 126   1976 38 schmimi01      Mike     Schmidt
## 127   1977 52 fostege01    George      Foster
## 128   1978 46  riceji01       Jim        Rice
## 129   1979 48 kingmda01      Dave     Kingman
## 130   1980 48 schmimi01      Mike     Schmidt
## 131   1981 31 schmimi01      Mike     Schmidt
## 132   1982 39 jacksre01    Reggie     Jackson
## 133   1982 39 thomago01    Gorman      Thomas
## 134   1983 40 schmimi01      Mike     Schmidt
## 135   1984 43 armasto01      Tony       Armas
## 136   1985 40 evansda01   Darrell       Evans
## 137   1986 40 barfije01     Jesse    Barfield
## 138   1987 49 dawsoan01     Andre      Dawson
## 139   1987 49 mcgwima01      Mark     McGwire
## 140   1988 42 cansejo01      Jose     Canseco
## 141   1989 47 mitchke01     Kevin    Mitchell
## 142   1990 51 fieldce01     Cecil     Fielder
## 143   1991 44 cansejo01      Jose     Canseco
## 144   1991 44 fieldce01     Cecil     Fielder
## 145   1992 43 gonzaju03      Juan    Gonzalez
## 146   1993 46 bondsba01     Barry       Bonds
## 147   1993 46 gonzaju03      Juan    Gonzalez
## 148   1994 43 willima04      Matt    Williams
## 149   1995 50 belleal01    Albert       Belle
## 150   1996 52 mcgwima01      Mark     McGwire
## 151   1997 56 griffke02       Ken     Griffey
## 152   1998 70 mcgwima01      Mark     McGwire
## 153   1999 65 mcgwima01      Mark     McGwire
## 154   2000 50  sosasa01     Sammy        Sosa
## 155   2001 73 bondsba01     Barry       Bonds
## 156   2002 57 rodrial01      Alex   Rodriguez
## 157   2003 47 rodrial01      Alex   Rodriguez
## 158   2003 47 thomeji01       Jim       Thome
## 159   2004 48 beltrad01    Adrian      Beltre
## 160   2005 51 jonesan01    Andruw       Jones
## 161   2006 58 howarry01      Ryan      Howard
## 162   2007 54 rodrial01      Alex   Rodriguez
## 163   2008 48 howarry01      Ryan      Howard
## 164   2009 47 pujolal01    Albert      Pujols
## 165   2010 54 bautijo02      Jose    Bautista
```

```
## 166   2011 43 bautijo02      Jose    Bautista
## 167   2012 44 cabremi01    Miguel     Cabrera
## 168   2013 53 davisch02     Chris       Davis
```

```
joined_aggregate_max_pitching
```

```
##    yearID HR  playerID nameFirst  nameLast
## 1    1871  4 brainas01       Asa  Brainard
## 2    1871  4 mcmuljo01      John  McMullin
## 3    1871  4 paborch01   Charlie     Pabor
## 4    1872  6 brittji01       Jim     Britt
## 5    1873  4 cummica01     Candy  Cummings
## 6    1874  5 mcbridi01      Dick   McBride
## 7    1875  6 cassijo01      John   Cassidy
## 8    1875  6 fishech01  Cherokee    Fisher
## 9    1877  4 bradlge01    George   Bradley
## 10   1877  4 devliji01       Jim    Devlin
## 11   1878  4 larkite01     Terry    Larkin
## 12   1880  6 corcola01     Larry  Corcoran
## 13   1880  6 coreyfr01      Fred     Corey
## 14   1882  7 goldsfr01      Fred Goldsmith
## 15   1882  7 welchmi01    Mickey     Welch
## 16   1883 14 goldsfr01      Fred Goldsmith
## 17   1886 11 atkinal01        Al  Atkinson
## 18   1886 11 baldwla01      Lady   Baldwin
## 19   1886 11 ferguch01   Charlie  Ferguson
## 20   1886 11 mullato01      Tony   Mullane
## 21   1886 11 stemmbi01      Bill  Stemmyer
## 22   1886 11 wiedmst01     Stump   Wiedman
## 23   1888 16 portehe01     Henry    Porter
## 24   1890 14 clarkjo01      John  Clarkson
## 25   1890 14 kilroma01      Matt    Kilroy
## 26   1890 14 lovetto01       Tom    Lovett
## 27   1890 14 stiveja01      Jack  Stivetts
## 28   1897 11 frasech01     Chick    Fraser
## 29   1897 11 lewiste01       Ted     Lewis
## 30   1900 12 kitsofr01     Frank    Kitson
## 31   1904 10 cronija01      Jack    Cronin
## 32   1905  9 gibsono01   Norwood    Gibson
## 33   1907 10 lindavi01      Vive  Lindaman
## 34   1909  9 wiltsho01     Hooks    Wiltse
## 35   1910 10 cranddo01       Doc  Crandall
## 36   1986 40 morrija02      Jack    Morris
```

The player hit most run called Barry Bonds, with 73 runs in year 2001.

#14. Has the distribution of home runs for players increased over the years? To solve this, I basically applied aggregate function from R and have the function set to SUM. Then applied ggplot2 to this dataset.

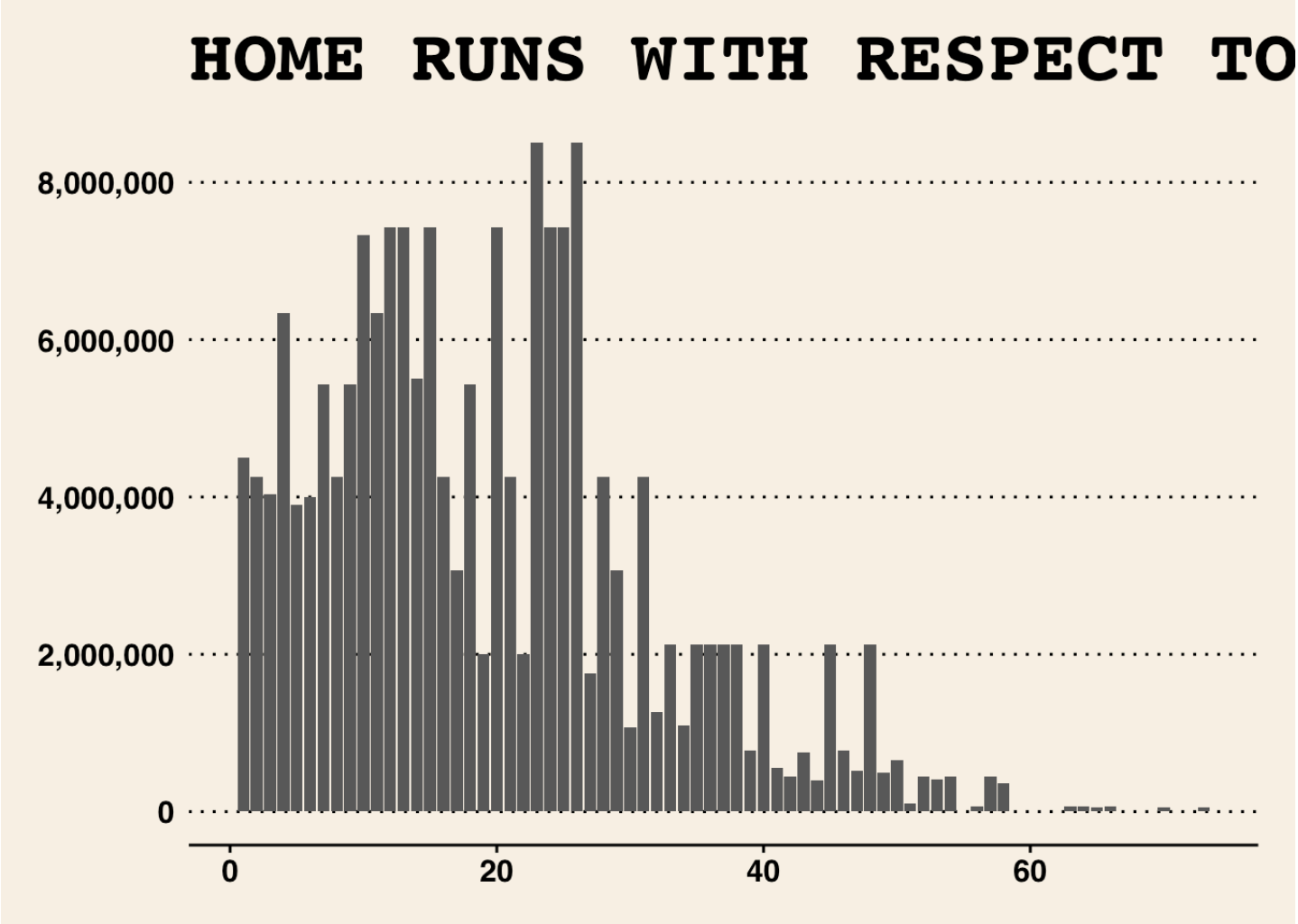
```
aggregate_sum_batting <- aggregate(HR ~ yearID, data =homerunsbatting , sum)
aggregate_sum_pitching <- aggregate(HR ~ yearID, data =homerunspitching , sum)


aggregate_sum_batting$type = c('batting')
aggregate_sum_pitching$type = c('pitching')

combinedsumhomeruns <- rbind(aggregate_sum_batting,aggregate_sum_pitching)

aggregate_sum_batting_picture = ggplot(combinedsumhomeruns,aes(x = yearID, y = HR, fi
ll = type)) +geom_bar(position="dodge", stat="identity")+scale_y_continuous(labels =
comma) + theme_wsj()+ scale_colour_wsj("colors6")+ggtitle('HOME RUNS WITH RESPECT TO
YEARS')
```

From the graph, it can be seen that the distributions for both batting and pitching increased over time.

#15 Do players who hit more home runs receive higher salaries? I used SQLite function to extract infomration from three tables,and I have noticed that there are some overlap information between batting and pitching table, For which if HR indicated as 0, it probably store information in another table. Thus, I combined two table and subset the dataset as if HR is not 0. Then applied ggplot on it, to see the overall distribution of change of salary with respect to home runs.

```
homerunsbatting_salary <- dbGetQuery(baseball, 'SELECT Batting.playerID,Batting.yearI
D,Batting.HR,Master.nameFirst,Master.nameLast,Salaries.salary FROM Batting LEFT JOIN
Master on Batting.playerID = Master.playerID LEFT JOIN Salaries ON Batting.playerID =
Salaries.playerID GROUP BY Batting.playerID,Batting.yearID ORDER BY Salaries.salary D
ESC')


homerunspitching_salary <- dbGetQuery(baseball, 'SELECT Pitching.playerID,Pitching.ye
arID,Pitching.HR,Master.nameFirst,Master.nameLast,Salaries.salary FROM Pitching LEFT
JOIN Master on Pitching.playerID = Master.playerID LEFT JOIN Salaries ON Pitching.pla
yerID = Salaries.playerID GROUP BY Pitching.playerID,Pitching.yearID ORDER BY Salarie
s.salary DESC')


cc <- rbind(homerunspitching_salary,homerunsbatting_salary)

combinedtable_home_run_salary <- subset(cc, cc$HR != 0 )

aggregate_home_run_salary = ggplot(combinedtable_home_run_salary,aes(x = HR, y = sala
ry)) +geom_bar(position="dodge", stat="identity")+scale_y_continuous(labels = comma)
+ theme_wsj()+ scale_colour_wsj("colors6")+ggtitle('HOME RUNS WITH RESPECT TO SAlARY'
)

aggregate_home_run_salary
```

```
## Warning: Removed 38434 rows containing missing values (geom_bar).
```



As it shown in the graph, there is even a negative association between number of home runs and salary individual got.

#16 Are certain baseball parks better for hitting home runs?

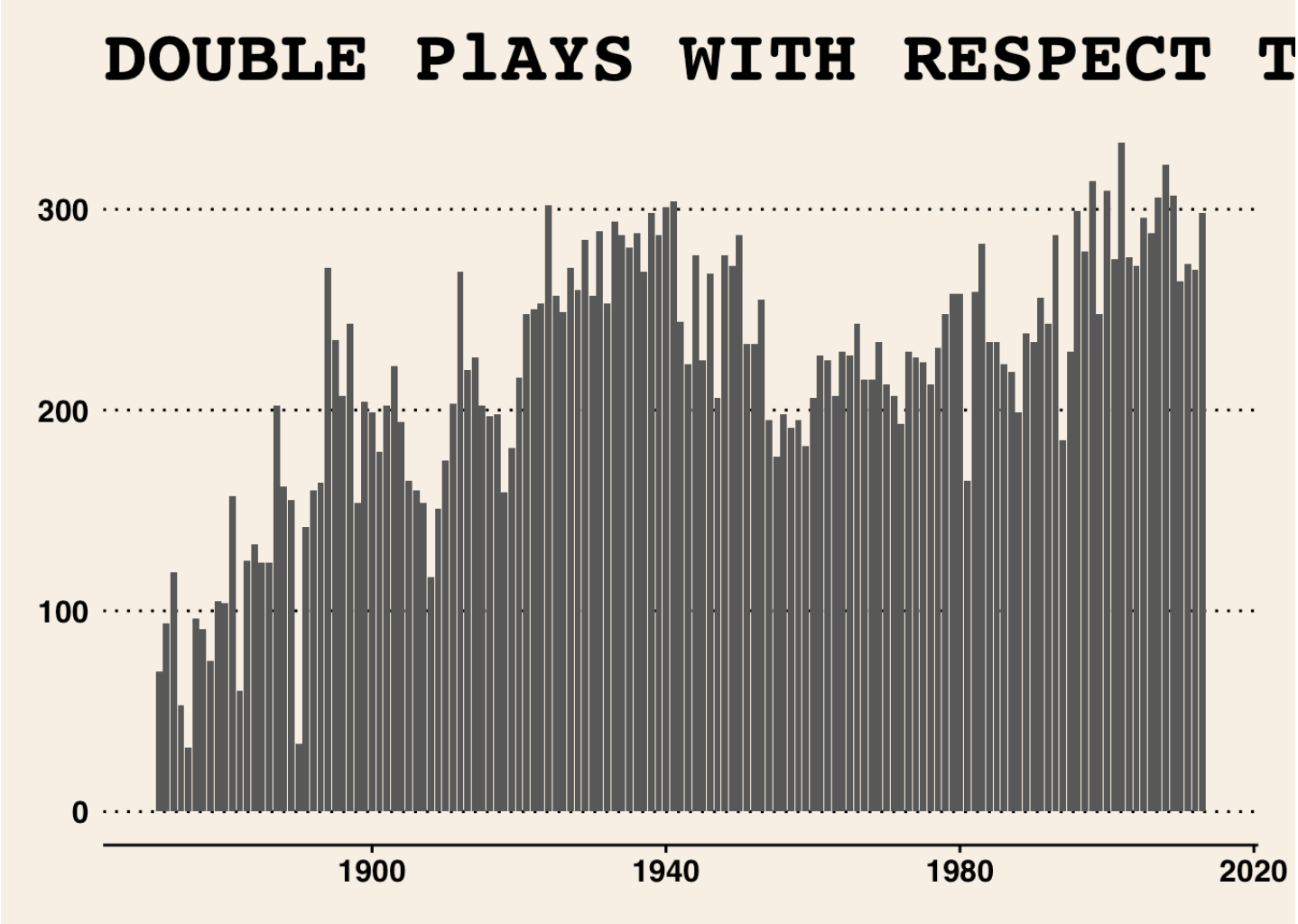
First, I found out the information for baseball parks stored in Teams table. Thus, I first extract information from Teams table. I take the first 20 parks as recommendation for hitting home runs.

```
park_home_run <- dbGetQuery(baseball,'SELECT Teams.HR, Teams.park FROM  Teams Group B
y Teams.park ORDER BY HR DESC LIMIT 20')
park_home_run
```

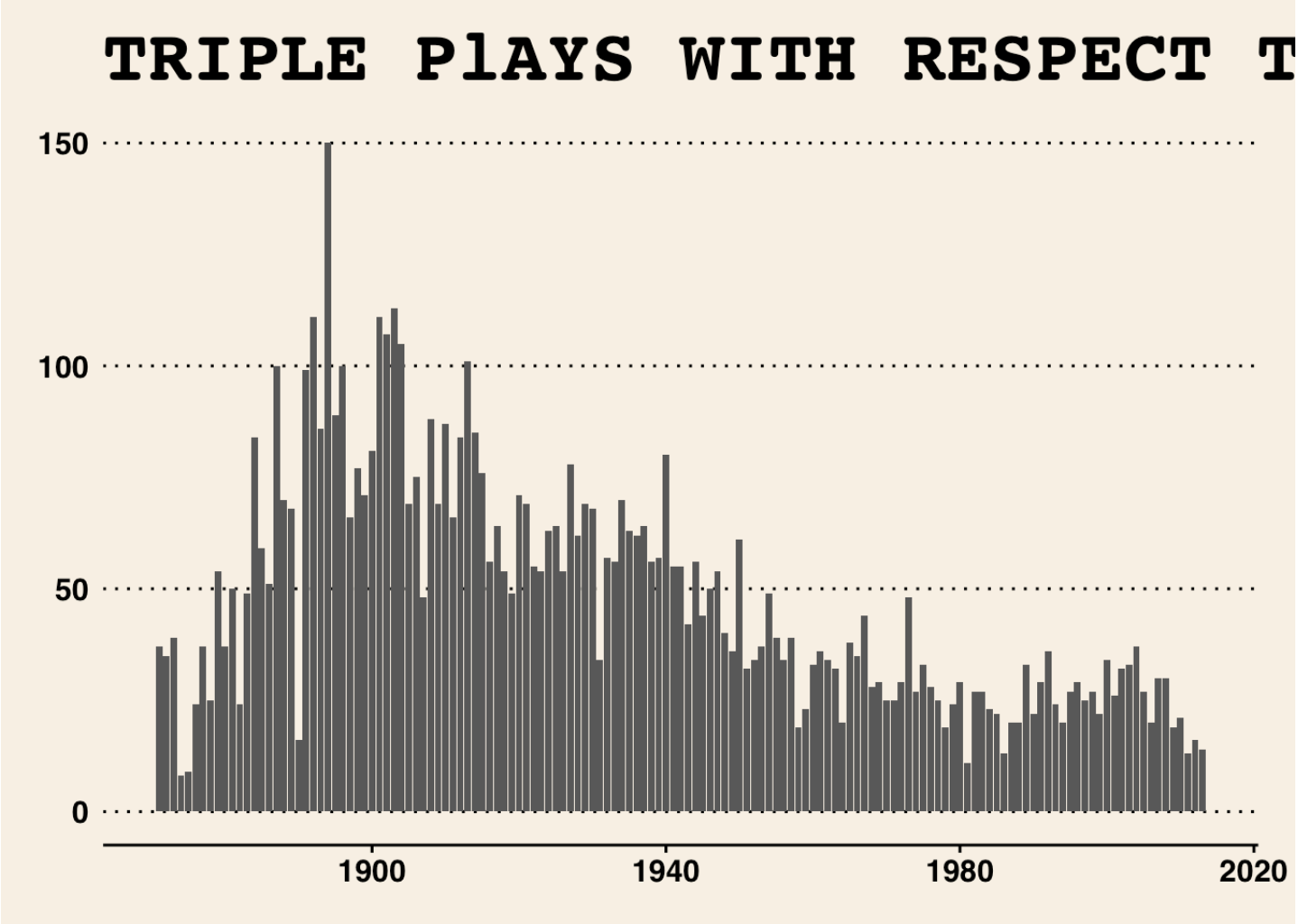
```
##     HR                              park
## 1  260                  Ameriquest Field
## 2  249                       Enron Field
## 3  244                Yankee Stadium III
## 4  244           Kingdome / Safeco Field
## 5  226                      PacBell Park
## 6  220               U.S. Cellular Field
## 7  215                Citizens Bank Park
## 8  209                       Miller Park
## 9  207    Atlanta-Fulton County Stadium
## 10 200                       Coors Field
## 11 198                      Safeco Field
## 12 195                     O.co Coliseum
## 13 191 Crosley Field/Riverfront Stadium
## 14 189                Wrigley Field (LA)
## 15 189      Network Associates Coliseum
## 16 187         Angel Stadium of Anaheim
## 17 184                 Busch Stadium III
## 18 183                          SBC Park
## 19 182         Great American Ball Park
## 20 180                     Tiger Stadium
```

#17. What's the distribution of double plays? triple plays? I simply extract information from Teams table, and found out there is diffculty calling 2B and 3B, so I used R to process the later code.

```
plays <- dbGetQuery(baseball,"SELECT Teams.* FROM Teams GROUP BY yearID ")
Double_plays_graph = ggplot(plays,aes(x = yearID, y = `2B`)) +geom_bar(position="dodg
e", stat="identity")+scale_y_continuous(labels = comma) + theme_wsj()+ scale_colour_w
sj("colors6")+ggtitle('DOUBLE PlAYS WITH RESPECT TO YEAR')
Double_plays_graph
```




```
Triple_plays_graph = ggplot(plays,aes(x = yearID, y = `3B`)) +geom_bar(position="dodg
e", stat="identity")+scale_y_continuous(labels = comma) + theme_wsj()+ scale_colour_w
sj("colors6")+ggtitle('TRIPLE PlAYS WITH RESPECT TO YEAR')
Triple_plays_graph
```



As it is showed in the graph, the number for Double Plays in a increased trend until 1940, and in a decreased trend until 1980, but again, follow a increased trend from 1980 to 2020.

However, for triple play, it is steady decreased trend.

#18. What pitchers have a large number of double or triple plays? Again, give their details (names,team, year, …).

Since it is asked for pitchers, I extract information from pitching table using sqlite function, and found out there probably a typo in the question, since there are doubles information stored for pitcher table, thus, I extract it from batting table, and then use R code to sort it.

```
Batting_plays  = dbGetQuery(baseball,'SELECT Master.*,Batting.* FROM Batting LEFT JOI
N Master on Batting.playerID = MASTER.playerID')
sorted_batting_plays_double <- Batting_plays[order(-Batting_plays$`2B`),]
sorted_batting_plays_triple <- Batting_plays[order(-Batting_plays$`3B`),]
head(sorted_batting_plays_double,1)
```

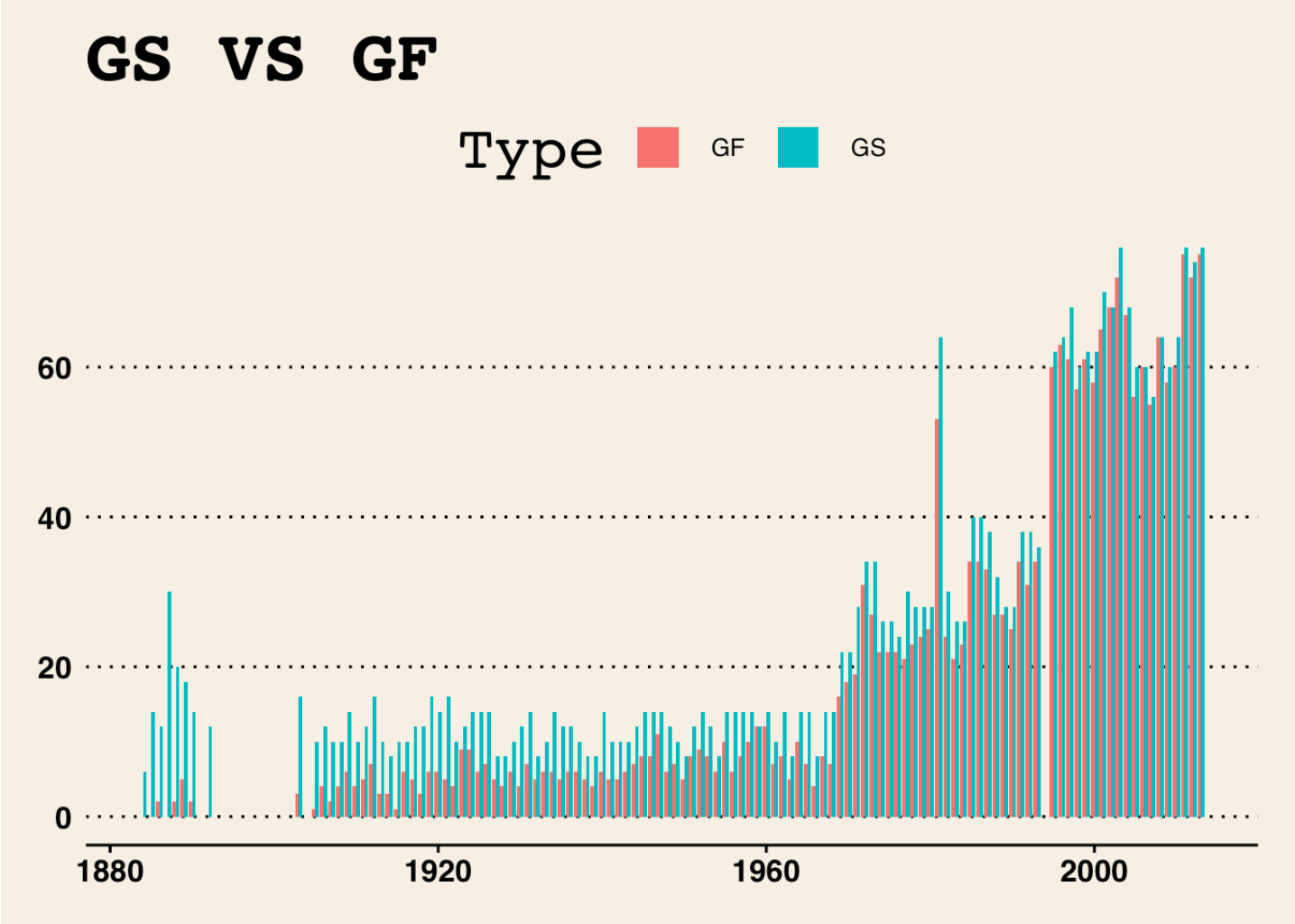
```
##       playerID birthYear birthMonth birthDay birthCountry birthState
## 92845 webbea01      1897          9       17          USA         TN
##          birthCity deathYear deathMonth deathDay deathCountry deathState
## 92845 White County      1965          5       23          USA         TN
##       deathCity nameFirst nameLast    nameGiven weight height bats throws
## 92845 Jamestown      Earl     Webb William Earl    185     73    L      R
##                debut      finalGame  retroID  bbrefID playerID yearID stint
## 92845 -1400698800000 -1144000800000 webbe101 webbea01 webbea01   1931     1
##       teamID lgID   G G_batting  AB  R   H 2B 3B HR RBI SB CS BB SO IBB HBP SH
## 92845    BOS   AL 151       151 589 96 196 67  3 14 103  2  2 70 51  NA   0  1
##       SF GIDP G_old
## 92845 NA   NA   151
```

```
head(sorted_batting_plays_triple,1)
```

```
##        playerID birthYear birthMonth birthDay birthCountry birthState birthCity
## 95250 wilsoch01      1883          8       21          USA         TX    Austin
##       deathYear deathMonth deathDay deathCountry deathState deathCity nameFirst
## 95250      1954          2       22          USA         TX   Bertram     Chief
##       nameLast nameGiven weight height bats throws          debut
## 95250   Wilson John Owen    185     74    L      R -1947520800000
##            finalGame  retroID   bbrefID  playerID yearID stint teamID lgID   G
## 95250 -1680458400000 wilsc102 wilsoch01 wilsoch01   1912     1    PIT   NL 152
##       G_batting  AB  R   H 2B 3B HR RBI SB CS BB SO IBB HBP SH SF GIDP G_old
## 95250       152 583 80 175 19 36 11  95 16 NA 35 67  NA   2 23 NA   NA   152
```

#19. How many games do pitchers start in a season? Plot this against games finished in a season. To do this, I first selected game started, game finished from pitching post table since it recorded statistics after a season.

```
pitchers_game <- dbGetQuery(baseball,'SELECT GS,GF,yearID FROM PitchingPost')
pitchers_game[is.na(pitchers_game)] <- 0
aggregate_pitcher_game_GS <- aggregate(GS ~ yearID,data = pitchers_game,sum)
aggregate_pitcher_game_GF <- aggregate(GF ~ yearID,data = pitchers_game,sum)
aggregate_pitcher_game_GS$Type = c('GS')
aggregate_pitcher_game_GF$Type = c('GF')
aggregate_pitcher_game_GS <- rename(aggregate_pitcher_game_GS,Game = GS)
aggregate_pitcher_game_GF <- rename(aggregate_pitcher_game_GF,Game = GF)
combined_GS_GF <- rbind(aggregate_pitcher_game_GS,aggregate_pitcher_game_GF )

combined_GS_GF_graph = ggplot(combined_GS_GF,aes(x = yearID, y = Game, fill = Type))
+geom_bar(position="dodge", stat="identity")+scale_y_continuous(labels = comma) + the
me_wsj()+ scale_colour_wsj("colors6")+ggtitle('GS VS GF')
combined_GS_GF_graph
```



As it shown in the graph,typically players started with more games at the begining of a seasons. but the rate for completing games have increaser over year. As for year after 2000, the two number are become very close.

#20. How many games do pitchers win in a season? I basically follow the same procedure as the previous one.Since if not win, the W column have number 0, thus it is not interupt with the final results.

```
pitchers_game_wins <- dbGetQuery(baseball,'SELECT W,yearID FROM PitchingPost')
aggregate_pitchers_game_wins <- aggregate(W ~ yearID,data = pitchers_game_wins,sum)
aggregate_pitchers_game_wins
```

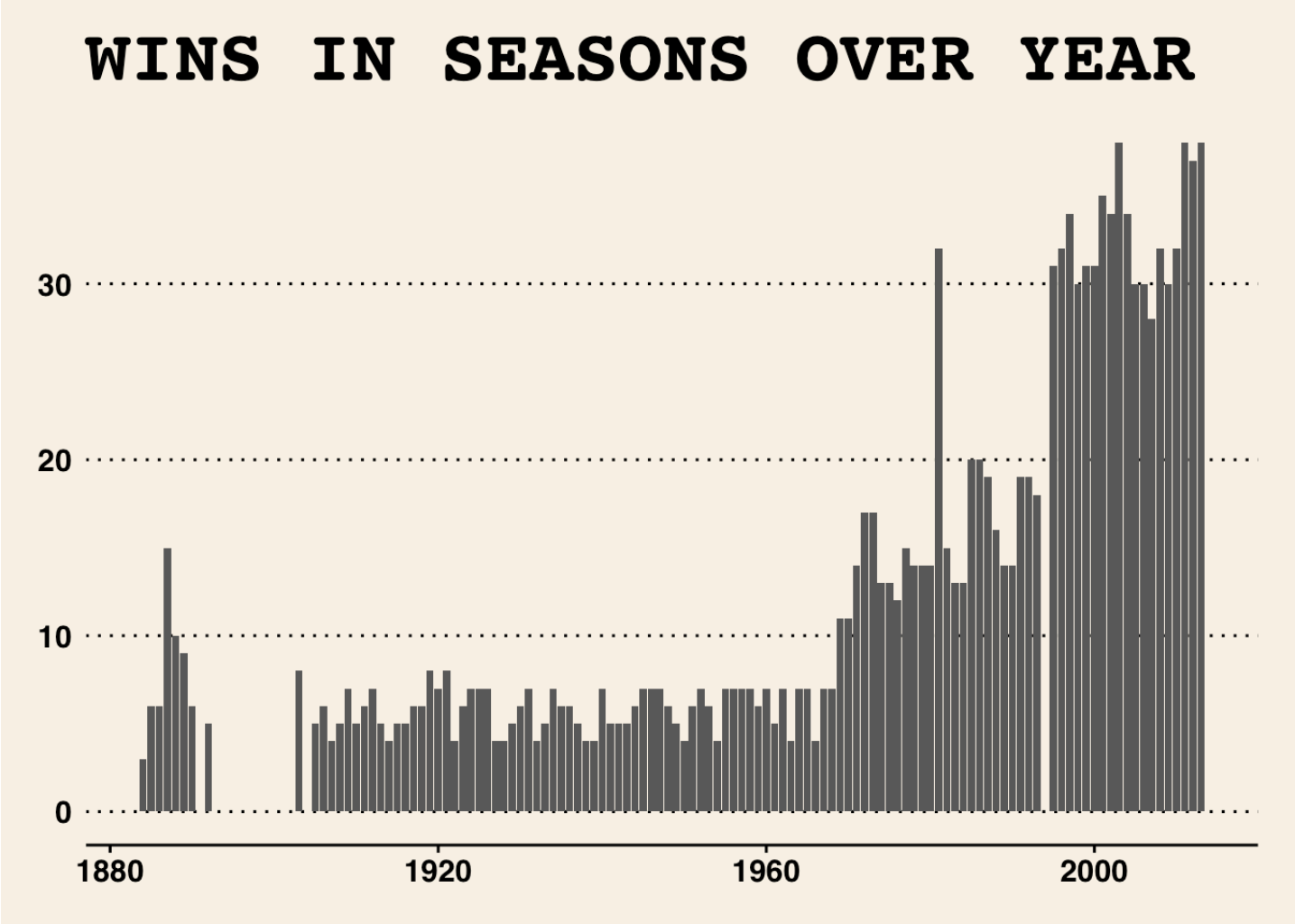
```
##     yearID  W
## 1     1884  3
## 2     1885  6
## 3     1886  6
## 4     1887 15
## 5     1888 10
## 6     1889  9
## 7     1890  6
## 8     1892  5
```

```
## 9      1903  8
## 10     1905  5
## 11     1906  6
## 12     1907  4
## 13     1908  5
## 14     1909  7
## 15     1910  5
## 16     1911  6
## 17     1912  7
## 18     1913  5
## 19     1914  4
## 20     1915  5
## 21     1916  5
## 22     1917  6
## 23     1918  6
## 24     1919  8
## 25     1920  7
## 26     1921  8
## 27     1922  4
## 28     1923  6
## 29     1924  7
## 30     1925  7
## 31     1926  7
## 32     1927  4
## 33     1928  4
## 34     1929  5
## 35     1930  6
## 36     1931  7
## 37     1932  4
## 38     1933  5
## 39     1934  7
## 40     1935  6
## 41     1936  6
## 42     1937  5
## 43     1938  4
## 44     1939  4
## 45     1940  7
## 46     1941  5
## 47     1942  5
## 48     1943  5
## 49     1944  6
## 50     1945  7
## 51     1946  7
## 52     1947  7
## 53     1948  6
## 54     1949  5
## 55     1950  4
```

```
## 56    1951  6
## 57    1952  7
## 58    1953  6
## 59    1954  4
## 60    1955  7
## 61    1956  7
## 62    1957  7
## 63    1958  7
## 64    1959  6
## 65    1960  7
## 66    1961  5
## 67    1962  7
## 68    1963  4
## 69    1964  7
## 70    1965  7
## 71    1966  4
## 72    1967  7
## 73    1968  7
## 74    1969 11
## 75    1970 11
## 76    1971 14
## 77    1972 17
## 78    1973 17
## 79    1974 13
## 80    1975 13
## 81    1976 12
## 82    1977 15
## 83    1978 14
## 84    1979 14
## 85    1980 14
## 86    1981 32
## 87    1982 15
## 88    1983 13
## 89    1984 13
## 90    1985 20
## 91    1986 20
## 92    1987 19
## 93    1988 16
## 94    1989 14
## 95    1990 14
## 96    1991 19
## 97    1992 19
## 98    1993 18
## 99    1995 31
## 100   1996 32
## 101   1997 34
## 102   1998 30
```

```
## 103   1999 31
## 104   2000 31
## 105   2001 35
## 106   2002 34
## 107   2003 38
## 108   2004 34
## 109   2005 30
## 110   2006 30
## 111   2007 28
## 112   2008 32
## 113   2009 30
## 114   2010 32
## 115   2011 38
## 116   2012 37
## 117   2013 38
```

```
aggregate_pitchers_game_wins_graph = ggplot(aggregate_pitchers_game_wins,aes(x = yearID, y = W)) +geom_bar(position="dodge", stat="identity")+scale_y_continuous(labels = comma) + theme_wsj()+ scale_colour_wsj("colors6")+ggtitle('WINS IN SEASONS OVER YEAR')
aggregate_pitchers_game_wins_graph
```



As it shown in the graph, the number of games win after seasons have grow with respect to year.

#21. How are wins related to hits, strikeouts, walks, homeruns and earned runs? I first selected information from Pitchingpost with all respected columns. Then, I subset the data with Wins not equal to 0, and then, for each category, I subset for categorical indicator is not 0, and finally have it recorded in a dataframe.

```
pitchers_game_wins_inv <- dbGetQuery(baseball,'SELECT W,yearID,H,SO,BB,HR,ER FROM Pit
chingPost')
pitchers_game_wins_inv <- subset(pitchers_game_wins_inv,pitchers_game_wins_inv$W != 0
)
pitchers_game_wins_inv_hits <- subset(pitchers_game_wins_inv,pitchers_game_wins_inv$H
!= 0)
pitchers_game_wins_inv_strikeouts <- subset(pitchers_game_wins_inv,pitchers_game_wins
_inv$SO != 0)
pitchers_game_wins_inv_walks <- subset(pitchers_game_wins_inv,pitchers_game_wins_inv$
BB != 0)
pitchers_game_wins_inv_homeruns <- subset(pitchers_game_wins_inv,pitchers_game_wins_i
nv$HR != 0)
pitchers_game_wins_inv_earnedruns <- subset(pitchers_game_wins_inv,pitchers_game_wins
_inv$ER != 0)

SUM <- data.frame('TYPE' = c('hits','strikeouts','walks','homeruns','earnedruns'),'Nu
mber_of_Wins' = c(1187,1186,1071,489,929))
SUM
```

```
##         TYPE Number_of_Wins
## 1       hits           1187
## 2 strikeouts           1186
## 3      walks           1071
## 4   homeruns            489
## 5 earnedruns            929
```

#22. What are the top ten collegiate producers of major league baseball players? How many colleges are represented in the database?

First,I used unique playerID, to count number of students for each collegiate producers by using pipeline function to count the number of appearance of schoolID and extracted another table by only selecting schoolID and schoolName to use the distince functions.

```
school_player <- dbGetQuery(baseball,'SELECT SchoolsPlayers.playerid,Schools.schoolNa
me,SchoolsPlayers.schoolID FROM SchoolsPlayers LEFT JOIN Schools On Schools.schoolID
= SchoolsPlayers.schoolID')

schools <- dbGetQuery(baseball,'SELECT Schools.schoolName,SchoolsPlayers.schoolID FRO
M SchoolsPlayers LEFT JOIN Schools On Schools.schoolID = SchoolsPlayers.schoolID')

school_player_sort <- sort(table(school_player$schoolID))
school_player_sort <- school_player %>%
  count(schoolID) %>%
  rename(number_of_players = n)
school_player_sorted<- school_player_sort[order(-school_player_sort$number_of_players
),]
school_player_sort_10 <- head(school_player_sorted,10)
Joined_information_school <- inner_join(school_player_sort_10,schools, by = c('school
ID'='schoolID'))
distinct(Joined_information_school)
```

```
## # A tibble: 10 x 3
##    schoolID  number_of_players schoolName
##    <chr>                <int> <chr>
##  1 usc                    102 University of Southern California
##  2 texas                  100 University of Texas at Austin
##  3 arizonast               98 Arizona State University
##  4 stanford                82 Stanford University
##  5 michigan                77 University of Michigan
##  6 holycross               75 College of the Holy Cross
##  7 notredame               70 University of Notre Dame
##  8 illinois                68 University of Illinois at Urbana-Champaign
##  9 arizona                 66 University of Arizona
## 10 ucla                    66 University of California, Los Angeles
```

```
nrow(distinct(schools))
```

```
## [1] 713
```

There are 713 distinct schools documented in this dataset.