STA141B HW3

Kwasie Agbemadon

11/6/2020

This homework assignment is about collecting and analyzing data about baseball teams, records, winners, etc.

Before I get to the questions, I will start by loading the database, using RSQLite and connecting it to the file. I will also load some of the tables in the database.

## Warning: package 'RSQLite' was built under R version 4.0.3

## Warning: package 'ggplot2' was built under R version 4.0.3

I used the head function to answer most of these questions, since there is too much data in the answers.

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

For this one, I used dbGetQuery and had it obtain the years from the yearID column. I grouped it by year to avoid any duplicate elements.

## yearID  
## 1 1871  
## 2 1872  
## 3 1873  
## 4 1874  
## 5 1875  
## 6 1876

This data covers the years 1871-2013. There is data for each year.

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

I used the master and managers tables to answer this question, and selected the playerID from both tables to get the answer. I used the unique function as it removes all duplicate elements, and then enclose it inside the length function to return the amount of unique players and managers.

## [1] 18354

## [1] 682

There are 18,354 players and 682 managers in this database.

1. How many players become managers?

I used the length function to count each player. I used the %in% function as it extracts the elements of the Manager’s playerID column that is also in the playerID column from the Master table.

## [1] 682

There are 682 players that became managers.

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

To answer the first sub-question, I used dbGetQuery and had it select the year and the amount of players from that year using the COUNT function. These variables come from the Appearances dataset. I then subsetted the data ranging from 2000 to 2013, grouping the data by year and then ordering the data by year. I did the same thing for the second sub-question but replaced yearID with teamID.

## yearID COUNT(playerID)  
## 1 2000 1381  
## 2 2001 1339  
## 3 2002 1319  
## 4 2003 1346  
## 5 2004 1345  
## 6 2005 1329

## teamID COUNT(playerID)  
## 1 ANA 204  
## 2 ARI 638  
## 3 ATL 611  
## 4 BAL 663  
## 5 BOS 684  
## 6 CHA 574

All teams do not have the same number of players.

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

I selected teamID, lgID, and divID from the Teams dataset then subsetted the yearID to 2010 and WSWin to Y (Yes). The same thing could also be done using the series\_post dataset, and subsetting round to “WS”, and yearID to 2010. Then, I could select lgIDwinner.

## teamID lgID divID  
## 1 SFN NL W

The SFN team won the World Series in 2010. Their league is NL and their division is W.

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

I used the series\_post dataset and selected the first three columns. I then added the divID column from the Teams dataset and used JOIN to merge the columns. “Round” was subsetted to “WS” and the data is grouped by year.

Notice how in the data that all values in the divID column are “NA.” The sport didn’t start using divisions until 1969, but they are named on the full table,

## yearID teamIDloser lgIDloser divID  
## 1 1884 NYP AA <NA>  
## 2 1885 STL AA <NA>  
## 3 1886 CHC NL <NA>  
## 4 1887 STL AA <NA>  
## 5 1888 STL AA <NA>  
## 6 1889 BRO AA <NA>

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

I simply did the same thing as the previous question, but I replaces the loser teams with the winner teams.

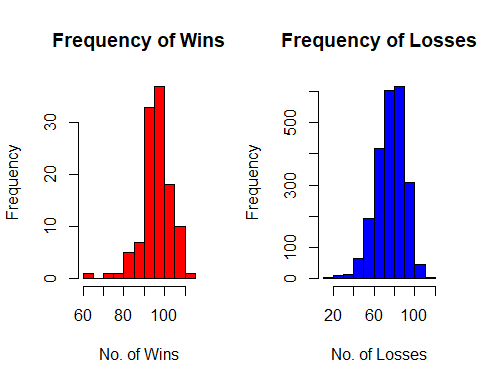
## yearID teamIDwinner lgIDwinner divID  
## 1 1884 PRO NL <NA>  
## 2 1885 CHC NL <NA>  
## 3 1886 STL AA <NA>  
## 4 1887 DTN NL <NA>  
## 5 1888 NYG NL <NA>  
## 6 1889 NYG NL <NA>

1. 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.

To do this question, I used the SeriesPost dataset and the Teams dataset (for the names and divisions). I initialized those datasets as sp (SeriesPost), t1 (for the name of the winning team), and t2 (for the name of the losing team) I selected the needed columns then LEFT JOINED the Teams table based on the TeamIDs twice, because we need the winning and losing IDs to make it work.

## yearID teamIDwinner name lgIDwinner divID teamIDloser  
## 1 2013 BOS Boston Americans AL <NA> SLN  
## 2 2012 SFN San Francisco Giants NL <NA> DET  
## 3 2011 SLN St. Louis Browns NL <NA> TEX  
## 4 2010 SFN San Francisco Giants NL <NA> TEX  
## 5 2009 NYA New York Highlanders AL <NA> PHI  
## 6 2008 PHI Philadelphia Blue Jays NL <NA> TBA  
## name lgIDloser divID losses  
## 1 St. Louis Browns NL <NA> 3  
## 2 Detroit Tigers AL <NA> 0  
## 3 Texas Rangers AL W 3  
## 4 Texas Rangers AL W 1  
## 5 Philadelphia Blue Jays NL <NA> 2  
## 6 Tampa Bay Devil Rays AL E 1

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

I created two histograms: One is for World Series Winners and the other is for World Series Losers. The data for the histograms contained the amount of wins for all teams, where they both won and lost the World Series for every year. This was done by assigning the W (win) column from the Teams dataset, then subsetted the World Series wins using vector syntax and the which() function. Both histograms are color coded. 

From both of the histograms, there are no relationships to winning games and winning the World Series, as the histograms are nearly the same.

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

I used the Salaries dataset and selected all columns where the year was 2003. I wanted to include the player’s full name and their origin, so I grabbed that information from the master table and used the JOIN clause based on playerID. I then ordered the data by Salary, used DESC to start from the highest values, and limited the table to just 3 entries using LIMIT.

## yearID teamID lgID playerID nameFirst nameLast birthCountry salary  
## 1 2003 TEX AL rodrial01 Alex Rodriguez USA 22000000  
## 2 2003 BOS AL ramirma02 Manny Ramirez D.R. 20000000  
## 3 2003 TOR AL delgaca01 Carlos Delgado P.R. 18700000

Alex Rodriguez (rodrial01) from USA was the richest baseball player with $22M. He played for the Texas Rangers.

Manny Ramirez (ramirma02) from Dominican Republic was the second richest with $20M. He played for the Boston Red Stockings.

Carlos Delgado (delgaca01) from Puerto Rico was the third richest with $18.7M. He played for the Toronto Blue Jays.

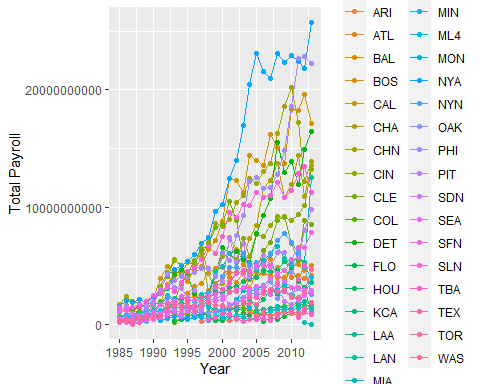
1. For 2010, compute the total payroll of each of the different teams. Next compute the team payrolls for all years in the database for which we have salary information. Display these in a plot.

For a, I simply selected all columns of the Salaries table and added up the salaries for each team. I then subsetted yearID to only show information from 2010.

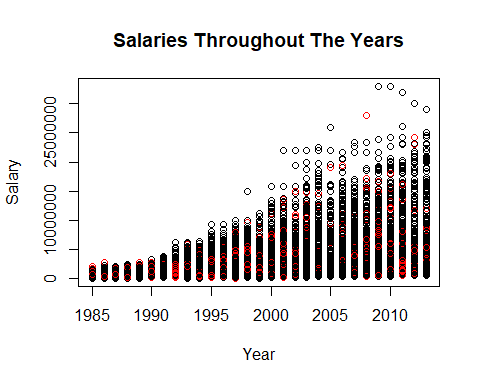
For b, I made a new variable and added divisions from the Teams table, then joined Teams using LEFT JOIN and grouped the data by both the year and the team.

For c, I used ggplot and plotted the year and the payroll together, and then grouped and colored the team names. ggplot automatically color codes the teams. I also added lines and points which represents the teams and years, respectfully.

## yearID teamID lgID SUM(salary)  
## 1 2010 ARI NL 60718166  
## 2 2010 ATL NL 84423666  
## 3 2010 BAL AL 81612500  
## 4 2010 BOS AL 162447333  
## 5 2010 CHA AL 105530000  
## 6 2010 CHN NL 146609000



1. 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 plotted the the salaries based on the year, and color-coded the dots using ifelse() on the col parameter of plot() where it’s only red when a team was won the World Series and/or the League at that year.  Salary relates to winning the World Series and/or league as it is based on the performance and progress of a team. Chances are there are more incoming players entering teams as well as skilled players that factor in the increase of salaries for those teams.

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

## Year Team League Division Total Payroll  
## 2 1985 BAL AL <NA> 693642720  
## 3 1985 BOS AL <NA> 1231424280  
## 4 1985 CAL AL <NA> 461692608  
## 5 1985 CHA AL <NA> 1112618114  
## 8 1985 CLE AL <NA> 740338258  
## 9 1985 DET AL <NA> 1169340159

## Year Team League Division Total Payroll  
## 1 1985 ATL NL <NA> 710736000  
## 6 1985 CHN NL <NA> 1753002546  
## 7 1985 CIN NL <NA> 1036629708  
## 10 1985 HOU NL <NA> 519638652  
## 12 1985 LAN NL <NA> 614203352  
## 15 1985 MON NL E 340925976

## Year Team League Division Total Payroll  
## 11 1985 KCA AL C 419453055  
## 14 1985 ML4 AL C 315954996  
## 37 1986 KCA AL C 586966410  
## 40 1986 ML4 AL C 278421976  
## 63 1987 KCA AL C 532262520  
## 66 1987 ML4 AL C 204210272

## Year Team League Division Total Payroll  
## 15 1985 MON NL E 340925976  
## 26 1985 TOR AL E 326064350  
## 41 1986 MON NL E 399729600  
## 52 1986 TOR AL E 466608739  
## 67 1987 MON NL E 249913872  
## 78 1987 TOR AL E 387741537

## Year Team League Division Total Payroll  
## 21 1985 SDN NL W 496646235  
## 22 1985 SEA AL W 170681000  
## 25 1985 TEX AL W 322413000  
## 47 1986 SDN NL W 512131185  
## 48 1986 SEA AL W 220457433  
## 51 1986 TEX AL W 283210998

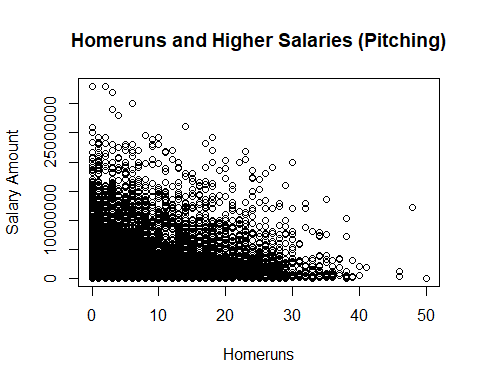
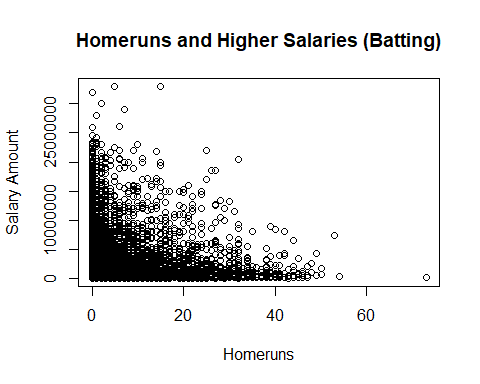
East Coast and Midwest baseball players in the AL league have higher salaries than teams from other regions of the country. Same applies to the NL league, except they are little bit behind on the AL league’s salary progress.

The KCA team has excelled on income given they were in the C division. The W division has one of the highest payrolls.

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

## playerID HR yearID  
## 1 bondsba01 73 2001  
## 2 mcgwima01 70 1998  
## 3 sosasa01 66 1998  
## 4 mcgwima01 65 1999  
## 5 sosasa01 64 2001  
## 6 sosasa01 63 1999

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

I created 2 separate graphs, one for Batters and one for Pitchers using the plot() function. I used both the batting and pitching datasets and used which() to select the players and salaries from the salary dataset that were in both said datasets. I then then plotted then both against each other.  From the graphs, it can be concluded that more homeruns do not result in higher salaries, regardless if they are a batter or a pitcher.

1. Are certain baseball parks better for hitting home runs?

Here, I used the Teams table and selected the park and the maximum amount of homeruns (using the max() function). I then grouped the data by the amount of maximum homeruns and ordered it descending.

## park sum(HR)  
## 1 Fenway Park II 4025  
## 2 Shibe Park 3500  
## 3 Yankee Stadium I 3318  
## 4 Shibe Park 3268  
## 5 Braves Field 3256  
## 6 Baker Bowl 3060

Fenway Park II, Shibe Park, Yankee Stadium I, Braves Field, and Baker Bowl are best for home runs.

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

I made separate tables for both types of plays. I used the FieldingPost table and grabbed the playerID, team, league, year, and summed up the amount of double/triple plays each player received.

## playerID teamID lgID sum(DP) yearID  
## 1 jeterde01 NYA AL 101 1996  
## 2 martiti02 SEA AL 73 1995  
## 3 olerujo01 TOR AL 55 1991  
## 4 garvest01 LAN NL 45 1974  
## 5 pujolal01 SLN NL 45 2001  
## 6 vizquom01 CLE AL 42 1995

## playerID teamID lgID sum(TP) yearID  
## 1 wambsbi01 CLE AL 1 1920  
## 2 aaronha01 ML1 NL 0 1957  
## 3 aasedo01 CAL AL 0 1979  
## 4 abbotgl01 OAK AL 0 1975  
## 5 abbotje01 CHA AL 0 2000  
## 6 abbotku01 FLO NL 0 1997

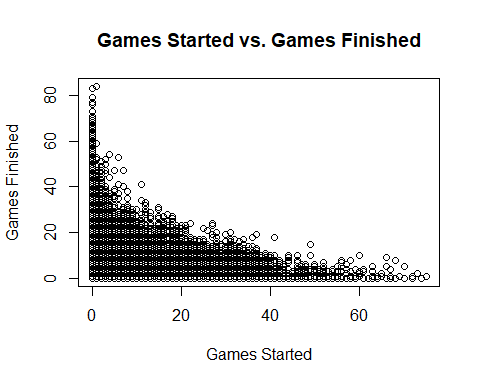
jeterde01, martiti02, olerujo01, garvest01, and pujolal01, and vizquom01 have large numbers of double plays.

Only wambsbi01 has one triple play. Nobody else obtained any triple plays.

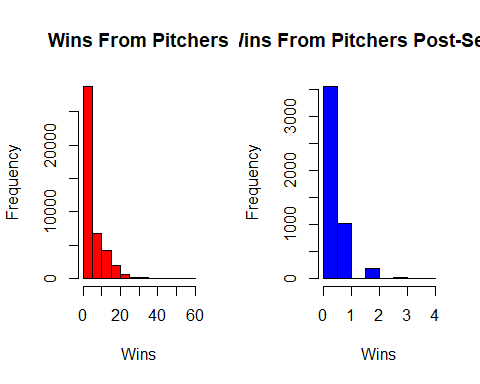
1. How many games do pitchers start in a season? Plot this against games finished in a season.

I created a new query called “games,” used the Pitching dataset, then selected the playerID, GS (Games Started), and GF (Games Finished) column. I then plotted the amount of games started against the amount of games finished.

## playerID GS GF  
## 1 aardsda01 0 5  
## 2 aardsda01 0 9  
## 3 aardsda01 0 7  
## 4 aardsda01 0 7  
## 5 aardsda01 0 53  
## 6 aardsda01 0 43

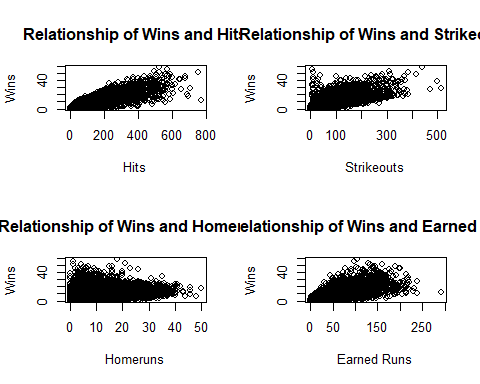
 Equally, there are pitchers that have started games without finishing them, and pitchers that didn’t start games but finished alot of them. Some pitchers lean in-between the amounts.

1. How many games do pitchers win in a season?

I used two histograms: One for the Pitching dataset and the another for pitching during post-seasons. I used the hist() function to do this. I then selected the W (Win) column from both datasets and color-coded them. 

It turns out that there are only a small portion of Pitchers that win at least 5 games. For the post-season, the majority don’t win any games.

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

I made datasets and graphs for each baseball asset using the Pitching dataset, because that set had all of the needed information. I then plotted each one. 

Each one is related as more of these baseball tactics increases the chances of winning. Homeruns are the exception, as players achieved alot of wins with only a few homeruns and vice versa.

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

I used the SchoolsPlayers dataset and selected both schoolID and counted schoolID (meaning the amount of players) using dbGetQuery, then grouped the data by schoolID and ordered the data by the amount of players starting from the highest amount. I then subsetted the table to only include the first 10 results.

The second sub-question is easy, I just used the length() function to count the rows as all of the tuples are unique.

## schoolID COUNT(schoolID)  
## 1 usc 102  
## 2 texas 100  
## 3 arizonast 98  
## 4 stanford 82  
## 5 michigan 77  
## 6 holycross 75  
## 7 notredame 70  
## 8 illinois 68  
## 9 ucla 66  
## 10 arizona 66

## [1] 749

For the amount of schools that are in the database, there are 749 represented.

1. What players have pitched in the post season and also hit a home run in their career?

Using the PitchingPost dataset, I grabbed the playerID’s and made sure it took unique names using the DISTINCT clause. I added up the amount of homeruns each player hit during the post season and subsetted the data to include nonzero amounts of homeruns. I grouped the players and sorted the data by the amount of total homeruns starting by the highest amount.

## playerID SUM(HR)  
## 1 pettian01 31  
## 2 glavito02 21  
## 3 hunteca01 21  
## 4 mussimi01 19  
## 5 clemero02 17  
## 6 smoltjo01 17

## [1] 871

About 871 players pitched and hit at least one homerun during the post season.

The whole report is concluded by disconnecting the database.

Code appendix:

library(RSQLite)  
library(ggplot2)  
baseball = dbConnect(SQLite(), 'lahman2013.sqlite')  
tables = dbListTables(baseball)  
batting = dbGetQuery(baseball, "SELECT \* FROM Batting")  
teamdata = dbGetQuery(baseball, "SELECT \* FROM Teams")  
master = dbGetQuery(baseball, "SELECT \* FROM MASTER")  
managers = dbGetQuery(baseball, "SELECT \* FROM Managers")  
salaries = dbGetQuery(baseball, "SELECT \* FROM Salaries")  
appearances = dbGetQuery(baseball, "SELECT \* FROM Appearances")  
fielding\_post = dbGetQuery(baseball, "SELECT \* FROM FieldingPost")  
pitching = dbGetQuery(baseball, "SELECT \* FROM Pitching")  
pitching\_post = dbGetQuery(baseball, "SELECT \* FROM PitchingPost")  
schools = dbGetQuery(baseball, "SELECT \* FROM Schools")  
collegiates = dbGetQuery(baseball, "SELECT \* FROM SchoolsPlayers")  
series\_post = dbGetQuery(baseball, "SELECT \* FROM SeriesPost")  
#1  
head(dbGetQuery(baseball, "SELECT yearID FROM teams GROUP BY yearID"))  
#2  
master\_ = unique(master$playerID)  
managers\_ = unique(managers$playerID)  
length(master\_)  
length(managers\_)  
#3  
length(managers\_ %in% master$playerID)  
#4a  
head(dbGetQuery(baseball, "SELECT yearID, COUNT(playerID) FROM APPEARANCES WHERE yearID BETWEEN 2000 and 2013 GROUP BY yearID ORDER BY yearID"))  
  
#4b  
head(dbGetQuery(baseball, "SELECT teamID, COUNT(playerID) FROM APPEARANCES WHERE yearID BETWEEN 2000 and 2013 GROUP BY teamID ORDER by teamID"))  
#5  
dbGetQuery(baseball, "SELECT teamID, lgID, divID FROM Teams WHERE yearID == 2010 AND WSWin == 'Y'")  
#6  
head(dbGetQuery(baseball, "SELECT sp.yearID, sp.teamIDloser, sp.lgIDloser, t.divID FROM SeriesPost as sp JOIN Teams as t ON sp.yearID = t.yearID WHERE sp.round == 'WS' GROUP BY sp.yearID"))  
#7  
head(dbGetQuery(baseball, "SELECT sp.yearID, sp.teamIDwinner, sp.lgIDwinner, t.divID FROM SeriesPost as sp JOIN Teams as t ON sp.yearID = t.yearID WHERE sp.round == 'WS' GROUP BY sp.yearID"))  
#8  
head(dbGetQuery(baseball, "SELECT sp.yearID, sp.teamIDwinner, t1.name, sp.lgIDwinner, t1.divID, sp.teamIDloser, t2.name, sp.lgIDloser, t2.divID, sp.losses FROM SeriesPost as sp LEFT JOIN Teams as t1 ON sp.teamIDwinner = t1.teamID LEFT JOIN Teams as t2 ON sp.teamIDloser = t2.teamID WHERE sp.round == 'WS' GROUP BY sp.yearID ORDER by sp.yearID DESC"))  
#9  
par(mfrow = c(1,2))  
hist(teamdata$W[which(teamdata$WSWin == 'Y')], xlab = "No. of Wins", main = "Frequency of Wins", col = "red")  
hist(teamdata$W[which(teamdata$WSWin == 'N')], xlab = "No. of Losses", main = "Frequency of Losses", col = "blue")  
#10  
dbGetQuery(baseball, "SELECT s.yearID, s.teamID, s.lgID, s.playerID,m.nameFirst, m.nameLast, m.birthCountry, s.salary FROM Salaries as s JOIN Master as m ON m.playerID = s.playerID WHERE yearID == 2003 ORDER BY salary DESC LIMIT 3")  
#11a  
head(dbGetQuery(baseball, "SELECT yearID, teamID, lgID, SUM(salary) FROM Salaries WHERE yearID == 2010 GROUP BY teamID"))  
  
#11b  
total\_payroll = dbGetQuery(baseball, "SELECT s.yearID, s.teamID, s.lgID, t.divID, SUM(s.salary) FROM Salaries as s LEFT JOIN Teams as t ON s.teamID = t.teamID GROUP BY s.yearID, s.teamID")  
colnames(total\_payroll) = c("Year", "Team", "League", "Division", "Total Payroll")  
  
#11c  
options(scipen = 9)  
ggplot(total\_payroll, aes(Year, `Total Payroll`, group = Team, colour = Team)) + geom\_line() + geom\_point()  
#12  
#color code the dot where the team won  
options(scipen = 4)  
plot(salaries$yearID, salaries$salary, xlab = "Year", ylab = "Salary", main = "Salaries Throughout The Years", col = ifelse(teamdata$WSWin == 'Y' | teamdata$LgWin == 'Y', "red", "black"))  
#13  
#Leagues  
head(total\_payroll[which(total\_payroll$League == "AL"),])  
head(total\_payroll[which(total\_payroll$League == "NL"),])  
  
#Divisions  
head(total\_payroll[which(total\_payroll$Division == 'C'),])  
head(total\_payroll[which(total\_payroll$Division == 'E'),])  
head(total\_payroll[which(total\_payroll$Division == 'W'),])  
#14  
head(dbGetQuery(baseball, "SELECT playerID, HR, yearID from Batting ORDER BY HR DESC"))  
#16  
plot(batting$HR[which(salaries$playerID %in% batting$playerID)], salaries$salary[which(salaries$playerID %in% batting$playerID)], xlab = "Homeruns", ylab = "Salary Amount", main = "Homeruns and Higher Salaries (Batting)")  
plot(pitching$HR[which(salaries$playerID %in% batting$playerID)], salaries$salary[which(salaries$playerID %in% batting$playerID)], xlab = "Homeruns", ylab = "Salary Amount", main = "Homeruns and Higher Salaries (Pitching)")  
#17  
head(dbGetQuery(baseball, "SELECT park, sum(HR) FROM Teams GROUP BY HR ORDER BY sum(HR) DESC"))  
#19  
head(dbGetQuery(baseball, "SELECT playerID, teamID, lgID, sum(DP), yearID from FieldingPost GROUP BY PlayerID ORDER BY sum(DP) DESC"))  
head(dbGetQuery(baseball, "SELECT playerID, teamID, lgID, sum(TP), yearID from FieldingPost GROUP BY PlayerID ORDER BY sum(TP) DESC"))  
#20  
games = dbGetQuery(baseball, "SELECT playerID, GS, GF FROM Pitching")  
head(games)  
plot(games$GS, games$GF, xlab = "Games Started", ylab = "Games Finished", main = "Games Started vs. Games Finished")  
#21  
par(mfrow = c(1,2))  
hist(pitching$W, xlab = "Wins", ylab = "Frequency", main = "Wins From Pitchers", col = 'red')  
hist(pitching\_post$W, xlab = "Wins", ylab = "Frequency", main = "Wins From Pitchers Post-Season", col = 'blue')  
#22  
par(mfrow = c(2,2))  
win\_hits = dbGetQuery(baseball, "SELECT playerID, W, H FROM Pitching")  
win\_strikeout = dbGetQuery(baseball, "SELECT playerID, W, SO FROM Pitching")  
win\_walks = dbGetQuery(baseball, "SELECT playerID, W, BB FROM Pitching")  
win\_hr = dbGetQuery(baseball, "SELECT playerID, W, HR FROM Pitching")  
win\_earnedruns = dbGetQuery(baseball, "SELECT playerID, W, ER FROM Pitching")  
plot(win\_hits$H, win\_hits$W, xlab = "Hits", ylab = "Wins", main = "Relationship of Wins and Hits")  
plot(win\_strikeout$SO, win\_strikeout$W, xlab = "Strikeouts", ylab = "Wins", main = "Relationship of Wins and Strikeouts")  
plot(win\_hr$HR, win\_hr$W, xlab = "Homeruns", ylab = "Wins", main = "Relationship of Wins and Homeruns")  
plot(win\_earnedruns$ER, win\_earnedruns$W, xlab = "Earned Runs", ylab = "Wins", main = "Relationship of Wins and Earned Runs")  
  
#23  
dbGetQuery(baseball, "SELECT schoolID, COUNT(schoolID) FROM SchoolsPlayers GROUP BY schoolID ORDER BY COUNT(schoolID) DESC LIMIT 10")  
length(schools$schoolName)  
#24  
ps\_players = dbGetQuery(baseball, "SELECT DISTINCT playerID, SUM(HR) FROM PitchingPost WHERE HR > 0 GROUP BY playerID ORDER BY SUM(HR) DESC")  
head(ps\_players)  
dim(ps\_players)[1]  
dbDisconnect(baseball)