Stories behind the NBA stats

Shuo Yang

Objectives

To explore stories behind the NBA stats from 1980 to 2009 including: a) overall trends; b) what are the key factors that make a team win a championship? c) correlation between regular season winning percentage and KPIs. d) predicting 2012 ~ 2013 season's NBA championship.

Data

1) Data source
Major source: DatabaseBasketball website [1]
Other sources: NBA official site [2], Wikipedia [3], ESPN [4]

2) Data description
2-1 Data from DatabaseBasketball website [1]
contents:

- Player regular season stats
- Player regular season career totals
- Player playoff stats
- Player playoff career totals
- Player all-star game stats
- Team regular season stats
- Complete draft history

description:

- provided as a zip file including seven csv files.
- ranges from 1946 ~ 1947 season to 2009 ~ 2010 season including both NBA and ABA league.
- contains cumulative stats of players and teams season by season.
- most data are accurate, however, there are some inconsistency between different files or in a single file.
- no information of championships
- missing stats of playoff teams of 2009 season

2-2 Data from NBA official website [2]
content:

- list of champions from 1980 season to 2009 season

description:

- copied and pasted into a csv file
- data are accurate

2-3 Data from Wikipedia [3]
content:

- a list of playoff teams of 2009 season

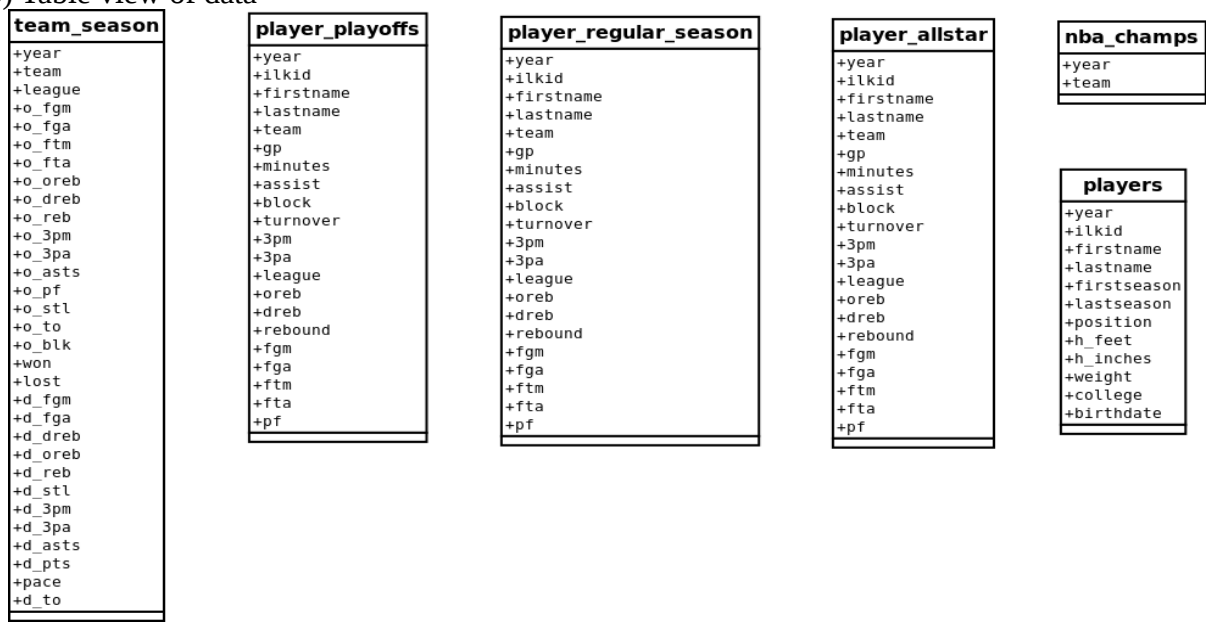
description:
- hardcoded into python script
- data are accurate

2-4 Data from ESPN [4]

content:
- stats of 2012-2013 NBA season

description:
- copied and pasted into a csv file
- data are accurate
- used for predication

3) Table view of data



4) Three types of data inconsistency
- inconsistency of abbreviation of team names. For example, San Antonio Spurs is coded as 'SAS' except in 2005 and 2008 season it was coded as 'SAN'.
- extra comma. For example, “University of California, Berkeley” instead of “University of California Berkeley”.

```
1ilkid,firstname,lastname,position,firstseason,lastseason,h_feet,h_inches,weight,college,birthdate
2ABDELAL01,Alaa,Abdelnaby,F,1990,1994,6,10,240,Duke University,1968-06-24 00:00:00
3ABDULKA01,Kareem,Abdul-jabbar,C,1969,1988,7,2,225,University of California - Los Angeles,1947-04-16 00:00:00
4ABDULMA01,Mahmo,Abdul-rauf,G,1990,2000,6,1,162,Louisiana State University,1969-03-09 00:00:00
5ABDULTA01,Tariq,Abdul-wahad,G,1997,2002,6,6,223,University of Michigan,1974-11-03 00:00:00
6ABDURSH01,Shareef,Abdur-rahim,F,1996,2007,6,9,225,University of California, Berkeley,1976-12-11 00:00:00
7ABERNTO01,Tom,Abernethy,F,1976,1980,6,7,220,Indiana University,1954-05-06 00:00:00
8ABLEFO01,Forest,Able,G,1956,1956,6,3,180,Western Kentucky University,1932-07-27 00:00:00
9ABRAMJO01,John,Abramovic,F,1946,1947,6,3,195,Salem College,1919-02-09 00:00:00
10ACKERAL01,Alex,Acker,G,2005,2008,6,5,185,Pepperdine University,1983-01-21 00:00:00
11ACKERDO01,Donald,Ackerman,G,1953,1953,6,0,183,Long Island University,1930-09-04 00:00:00
12ACRESMA01,Mark,Acres,C,1987,1992,6,11,220,Oral Roberts University,1962-11-15 00:00:00
13ACTONCH01,Charles,Acton,F,1967,1967,6,6,210,Hillsdale College,1942-01-11 00:00:00
14ADAMSAL01,Alvan,Adams,C,1975,1987,6,9,210,University of Oklahoma,1954-07-19 00:00:00
15ADAMSDO01,Don,Adams,F,1970,1976,6,6,210,Northwestern University,1947-11-27 00:00:00
16ADAMSGE01,George,Adams,F,1972,1974,6,5,210,Gardner-Webb University,1949-05-15 00:00:00
17ADAMSMI01,Michael,Adams,G,1985,1995,5,10,162,Boston College,1963-01-19 00:00:00
18AdamsHa01,Hassan,Adams,G,2006,2008,6,4,220,University of Arizona,1984-06-20 00:00:00
19ADDISRA01,Rafael,Addison,F,1986,1996,6,7,215,Syracuse University,1964-07-22 00:00:00
20ADELMRI01,Rick,Adelman,G,1968,1974,6,1,175,Loyola Marymount University,1946-06-16 00:00:00
```

- extra space. Below is an example of extra space.
  Player id without extra space:

```
ilkid,year,firstname,lastname,team,leag,gp,minutes,pts,dreb,oreb,reb,asts,stl,
ARMSTPA01,1949,Paul,Armstrong,FTW,N,3,0,9,0,0,0,6,0,0,0,6,22,4,4,1,0,0
BARKECL01,1949,Cliff,Barker,INI,N,6,0,34,0,0,0,13,0,0,0,10,31,12,15,10,0,0
BARNHLE01,1949,Leo,Barnhorst,CH1,N,2,0,22,0,0,0,4,0,0,0,10,25,8,6,6,0,0
BEARDRA01,1949,Ralph,Beard,INI,N,5,0,66,0,0,0,22,0,0,0,11,70,22,28,22,0,0
BLACKCH01,1949,Charlie,Black,AND,N,8,0,57,0,0,0,17,0,0,0,38,61,18,29,21,0,0
BOBBNE01,1949,Nelson,Bobb,PH1,N,2,0,2,0,0,0,1,0,0,0,3,3,1,0,0,0,0
BORNHJA01,1949,Jake,Bornheimer,PH1,N,2,0,2,0,0,0,0,0,0,0,2,3,1,0,0,0,0
BORYLVI01,1949,Vince,Boryla,NYK,N,5,0,75,0,0,0,7,0,0,0,25,52,23,32,29,0,0
```

player id with extra space:

```
ilkid,firstname,lastname,position,firstseason,lastseason,h_feet,h_inches,weight,college,birthdate
ABDELAL01 ,Alaa,Abdelnaby,F,1990,1994,6,10,240,Duke University,1968-06-24 00:00:00
ABDULKA01 ,Kareem,Abdul-jabbar,C,1969,1988,7,2,225,University of California - Los Angeles,1947-04-16 00:00:00
ABDULMA01 ,Mahmo,Abdul-rauf,G,1990,2000,6,1,162,Louisiana State University,1969-03-09 00:00:00
ABDULTA01 ,Tariq,Abdul-wahad,G,1997,2002,6,6,223,University of Michigan,1974-11-03 00:00:00
ABDURSH01 ,Shareef,Abdur-rahim,F,1996,2007,6,9,225,University of California Berkeley,1976-12-11 00:00:00
ABERNTO01 ,Tom,Abernethy,F,1976,1980,6,7,220,Indiana University,1954-05-06 00:00:00
```

KPIs

- ratio of assist made by team over assist made by opponents
- regular season field goal percentage
- regular season three-point percentage
- average team age weighted by minutes played

ETL

1) Challenge of ETL
   - information needed is separated in different csv files.
   - there are data inconsistencies between csv files or inside a single csv file.

2) input & output

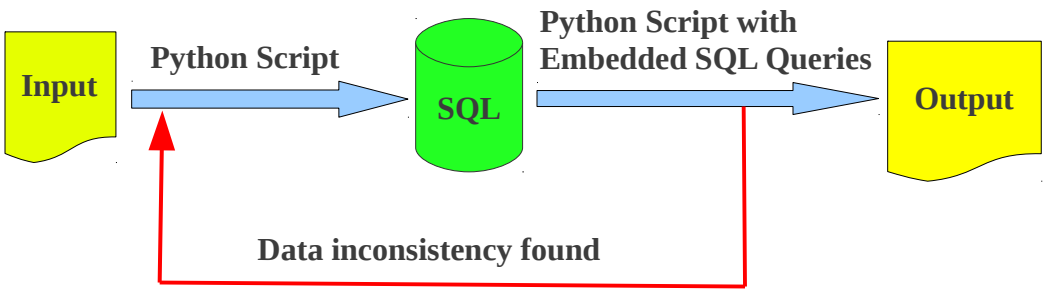
input (6 csv files):

- Player regular season stats
- Player playoff stats
- Player playoff career totals
- Player all-star game stats
- Team regular season stats
- NBA championships from 1980 to 2009

output (1 csv file):

- team stats from 1980 to 2009 including :
  - ratio of assist made by team over assist made by opponents
  - regular season field goal percentage
  - regular season three-point percentage
  - average team age weighted by minutes played
  - regular season winning percentage
  - is a championship or not

3) ETL iteration

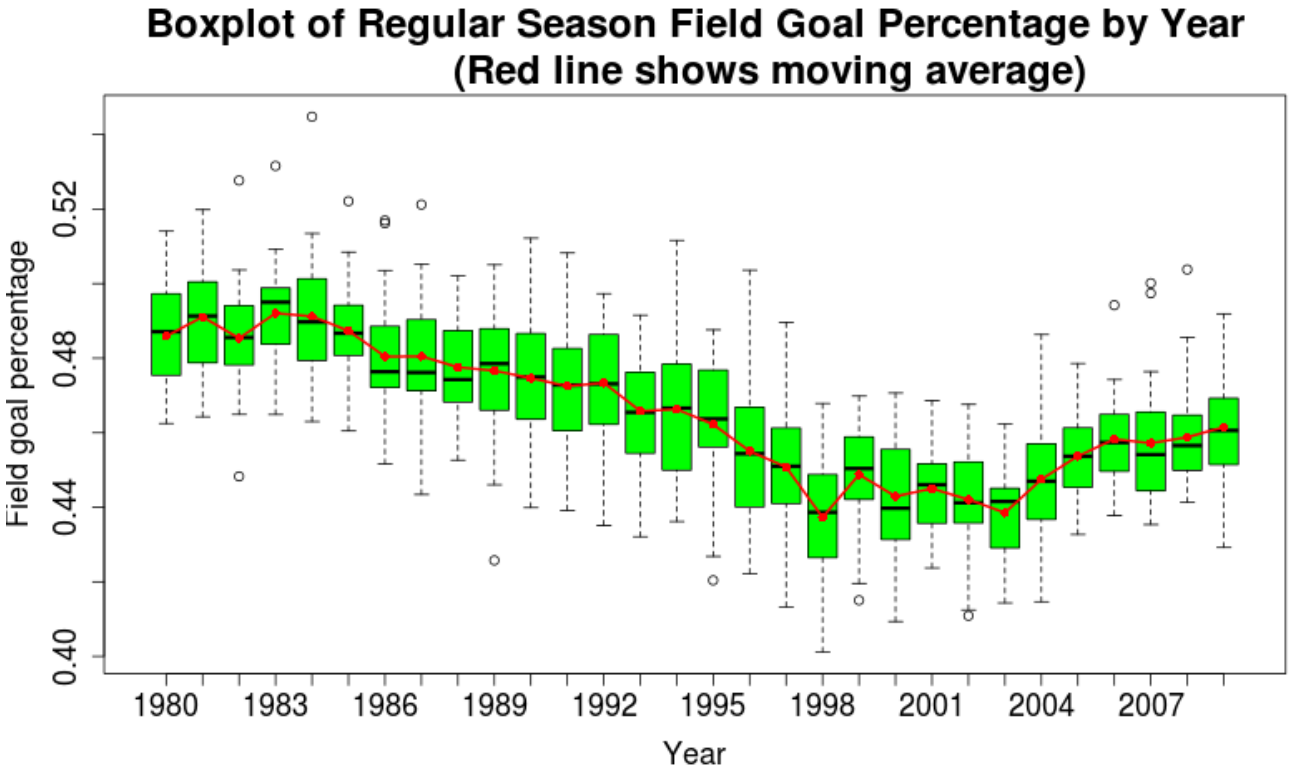


4) Benefits of using SQL

- The data lends itself to relational database.
- SQL makes tables join and merge easy and quick.
- SQL can be embedded and executed in Python.

Data Analysis & Visualization

How did field goal percentage change over time ?



Story behind the data:
Overall decreasing of field goal percentage

- League has become more and more physically demanding.
- Players' and teams' defensive skills have been improving
- Scoring inside the paint has become more and more difficult.
  Many teams now rely on jump shot which drags down the field goal percentage.

Story behind the data:
The lowest average in 30 years shows the impact of 1998 ~ 1999 season long time (over six month) lockout.

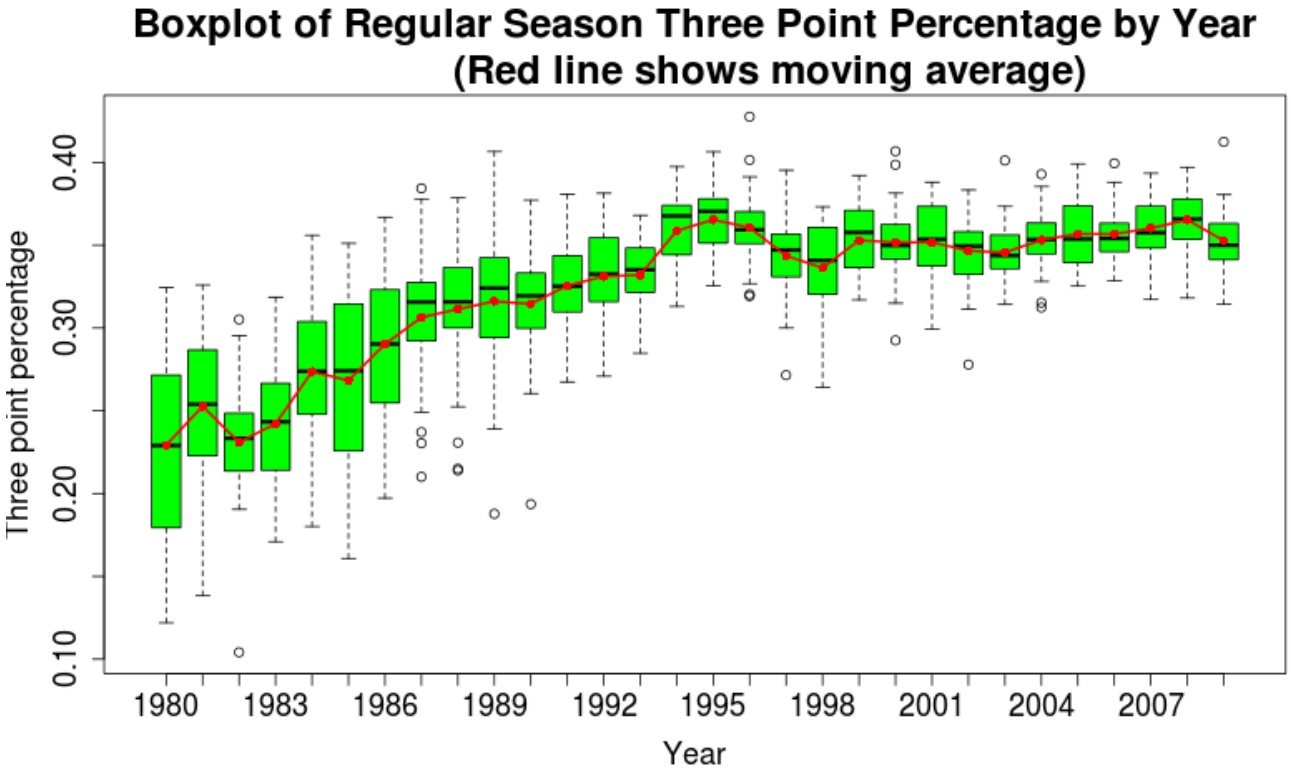
- 1998 – 1999 regular season was shorted to 50 games.
- Tightened schedule resulted in many 'ugly games'.
- All-Star Game was canceled.

Story behind the data:
Field goal percentage has been improving since 2003-2004 season.

- Many talented players emerged from colleges or other countries, like LeBron James, Dirk Nowitzki.

- The competitive environment pushes players to practice and play hard in order to stay on contract.

How did three point percentage change over time ?



Stories behind the data:
The overall increasing of three-point percentage

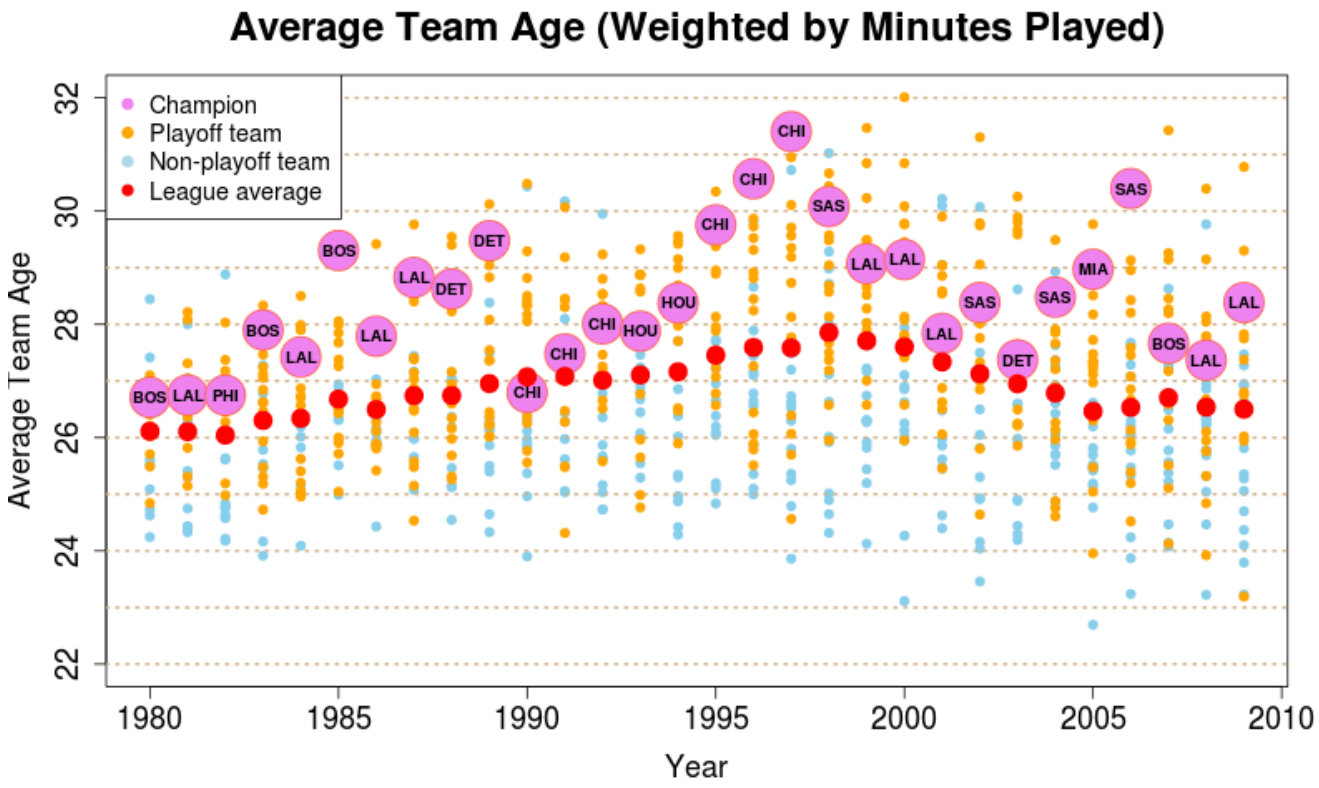
- There is growing importance of three-point.
- There is a growing popularity of the three-pointer, like Ray Allen.
- Sometimes three point is crucial to win the game.

Stories behind the data:
A sharp increase from 1993 season to 1994 season reflects the change of the three-point rule [5].

- In 1994 season, the league shortened the three-point line to a uniform 22 feet around the basket.

We can also see the impact of the lockout in 1998 through low average and uneven distribution.

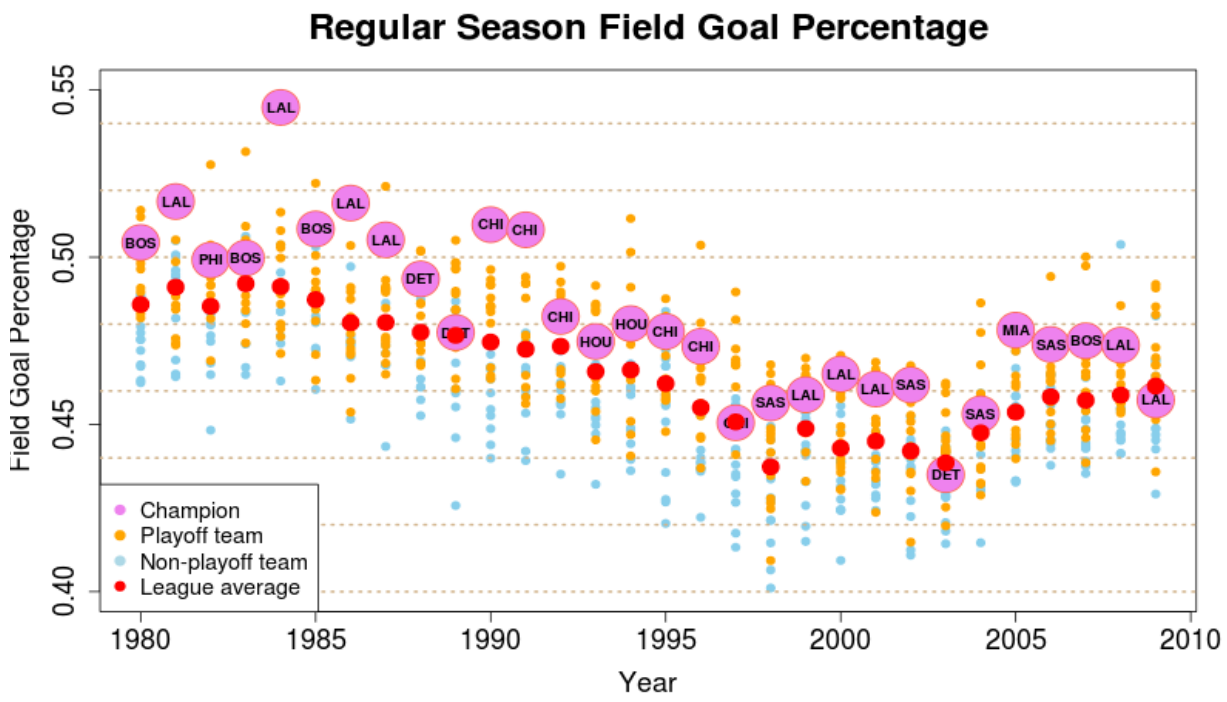
What makes a team a championship ?



<u>Stories behind the data:</u>

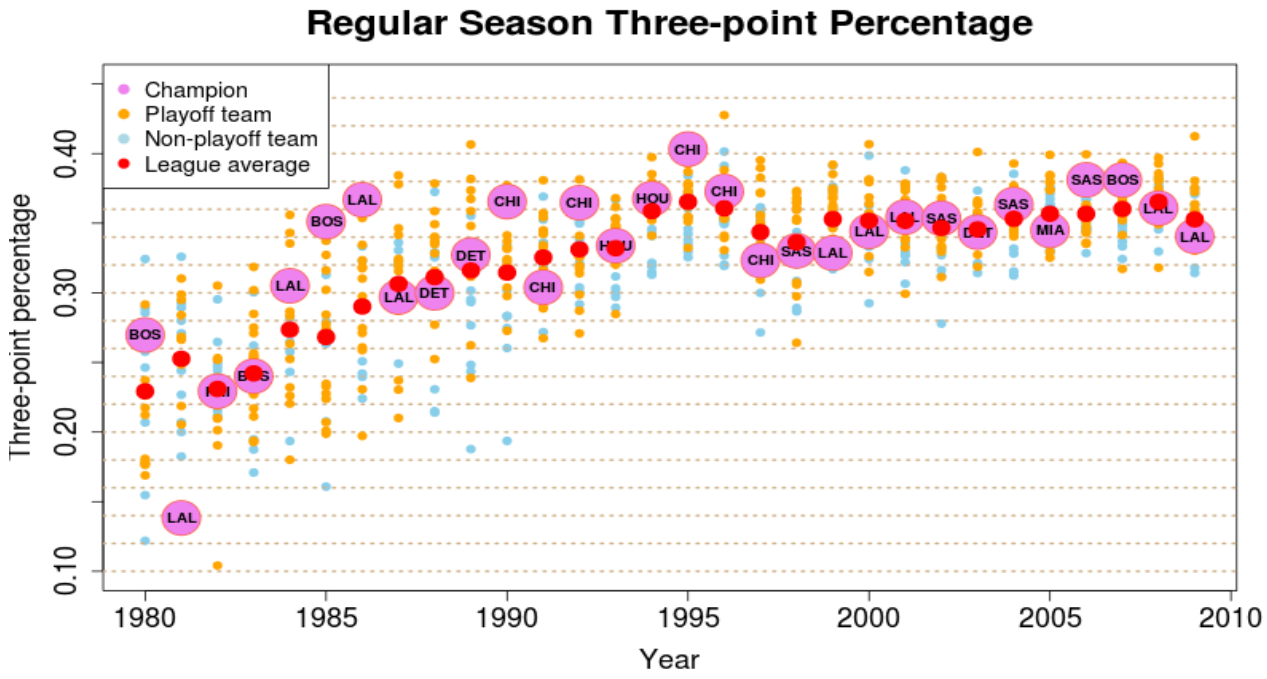
Championship teams are significantly older than league average, and they continue to win as they get older [6]. (LAL, BOS in 80's, CHI in 90's, LAL and SAS in 00's)

- Older team age implies better team cohesion.
- They know how to play together.
- They have built excellent leadership.
- They share experiences and make each other a better player.
- Good teams are kept in one piece. They build their teams around 2-4 core players. As these players get older, they are more likely to win a championship.



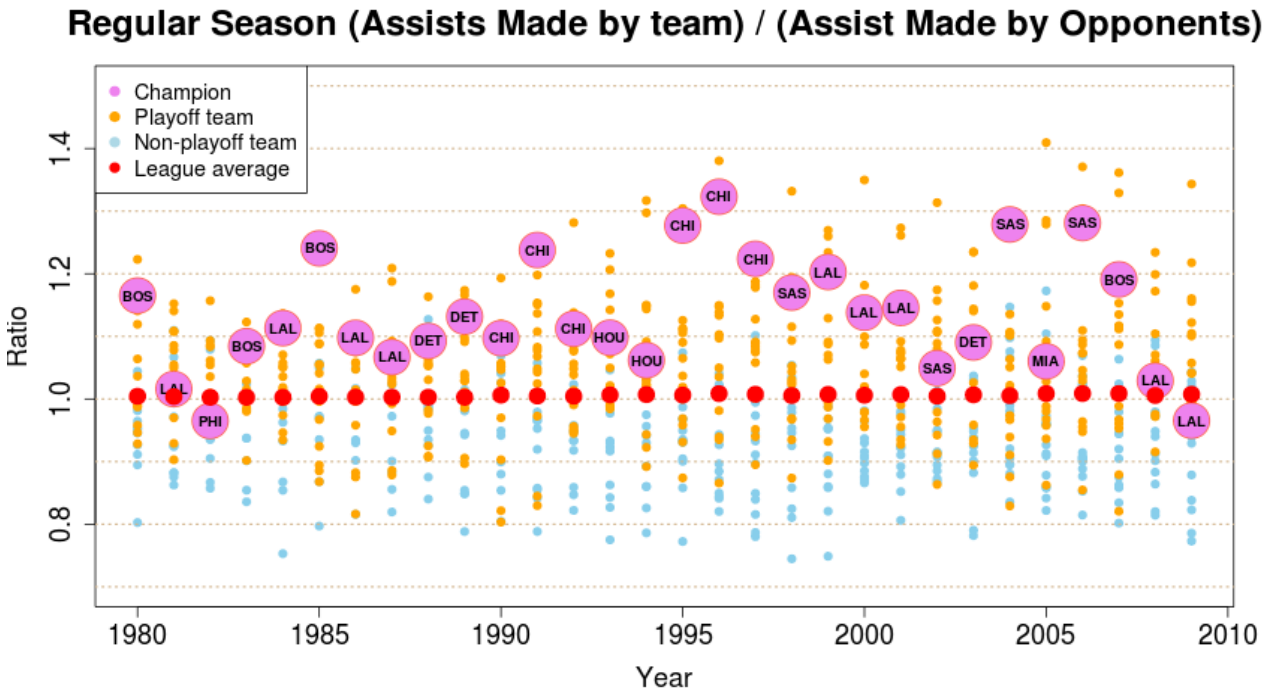
Stories behind the data:

Championship teams usually have above league average field goal percentage which implies that a key to win championship is consistency.



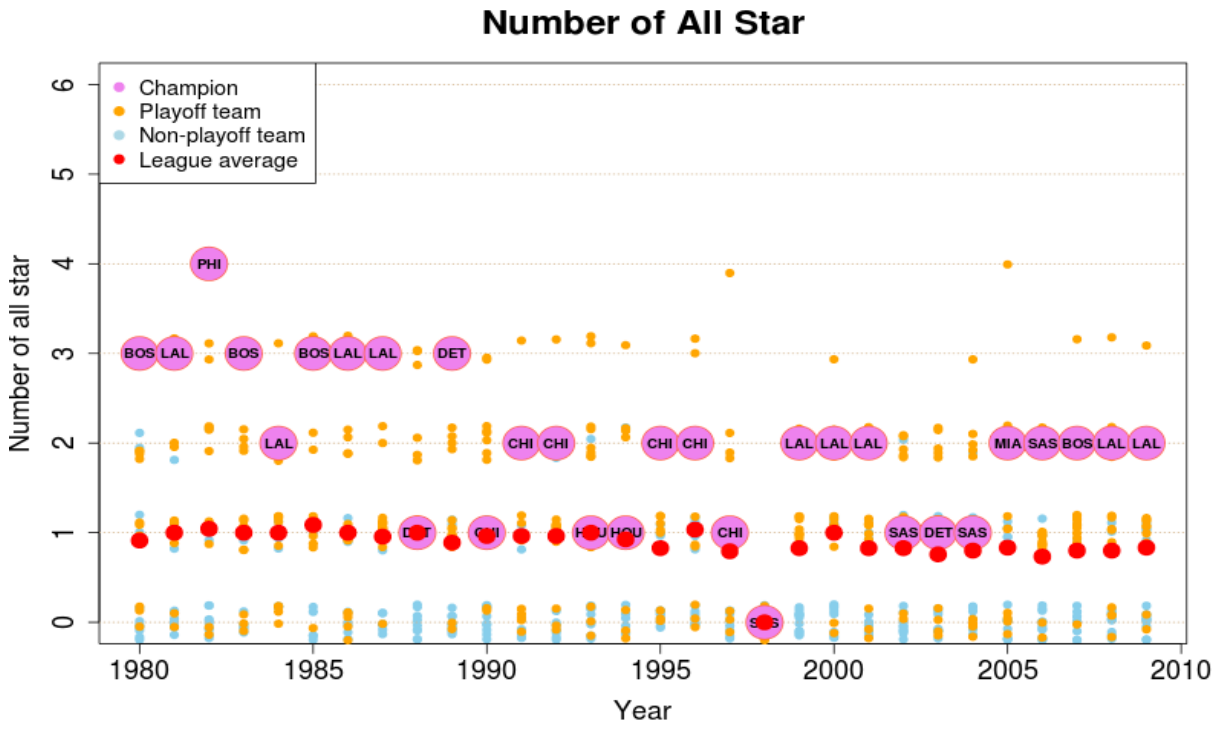
Stories behind the data:

Three-point percentage seems not correlated with winning a championship. But a championship should at least achieve the league average as shown by the graph.



Stories behind the data:

A championship team usually gives more assists than its opponents. Higher ratio here means a better offense team and a better defensive team.



Stories behind the data:

A championship team should have at least one super star. 1998 season's all star game was canceled due to lockout.

Correlation Analysis: regular season winning percentage Vs. (KPIs)

Analysis technique used: Linear Regression

R code:

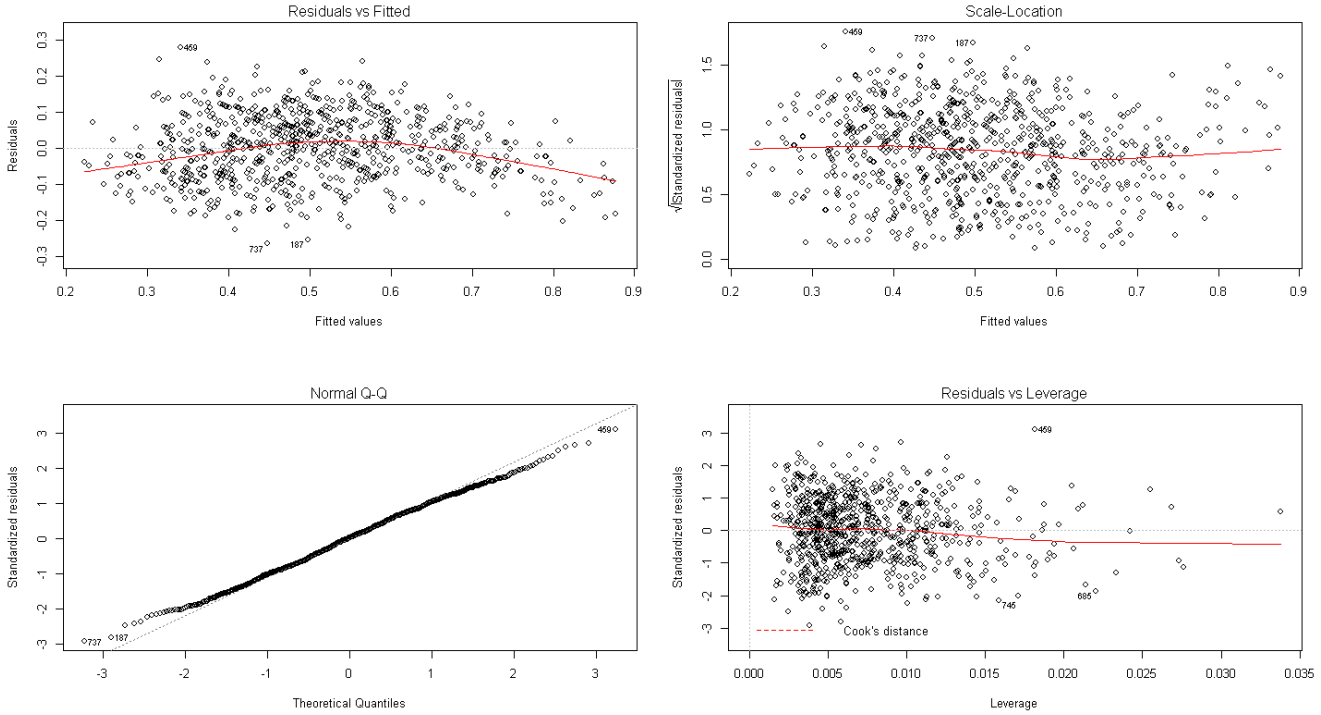
```
teams.train = read.csv('team_season_1980-2009.csv', header=TRUE)
tt.lm = lm(win_ratio ~ oast_dast+fgp+tpp+num_allstar+avg_age, data=teams.train)
```

Summary of the model:

```
Call:
lm(formula = win_ratio ~ oast_dast + fgp + tpp + num_allstar +
    avg_age, data = teams.t)

Residuals:
      Min        1Q    Median        3Q       Max
-0.264966 -0.066357  0.002714  0.066635  0.279363

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.330377   0.104582 -12.721  < 2e-16 ***
oast_dast    0.486478   0.034628  14.049  < 2e-16 ***
fgp          1.468739   0.179844   8.167 1.22e-15 ***
tpp          0.412871   0.072163   5.721 1.49e-08 ***
num_allstar  0.061072   0.004295  14.218  < 2e-16 ***
avg_age      0.017531   0.002330   7.525 1.42e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09072 on 806 degrees of freedom
Multiple R-squared:  0.6618,    Adjusted R-squared:  0.6597
F-statistic: 315.4 on 5 and 806 DF,  p-value: < 2.2e-16
```

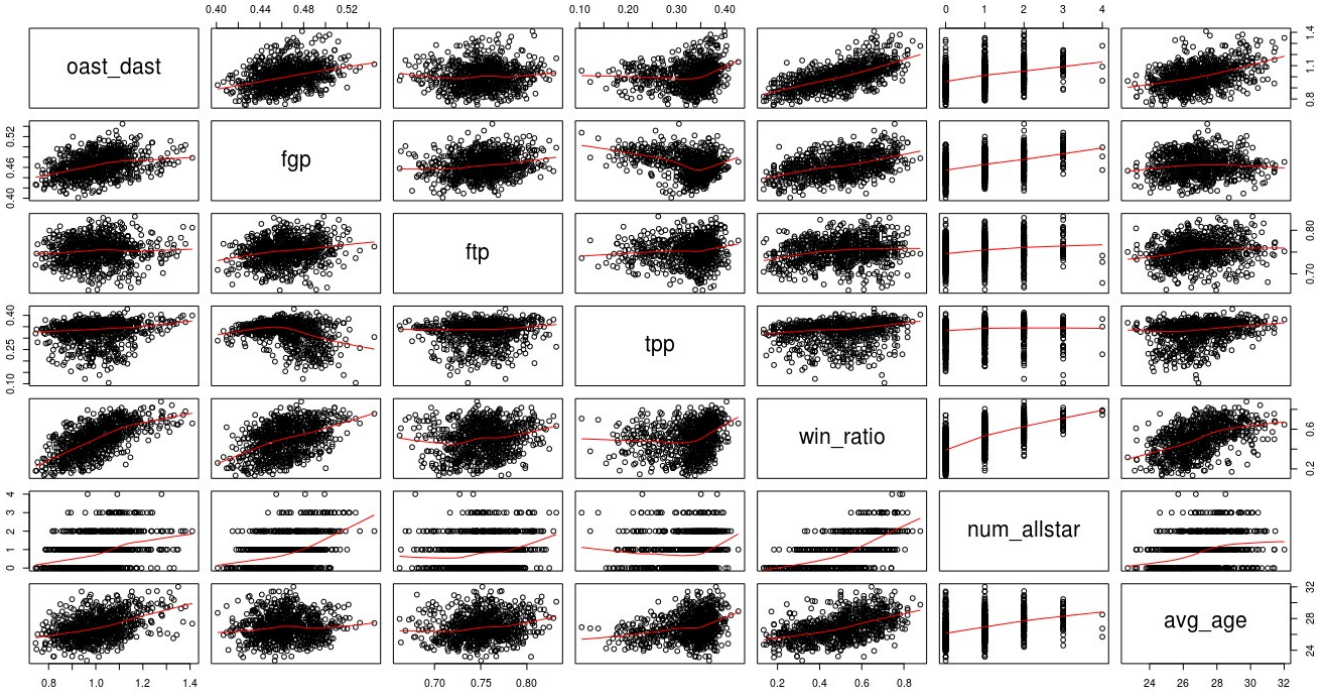
As shown above, all five factors are statistical significant by looking at P value (last column).
Also, the value of multiple R-squared indicates that the model is effective. (1 is the ideal value)

Plot of the model



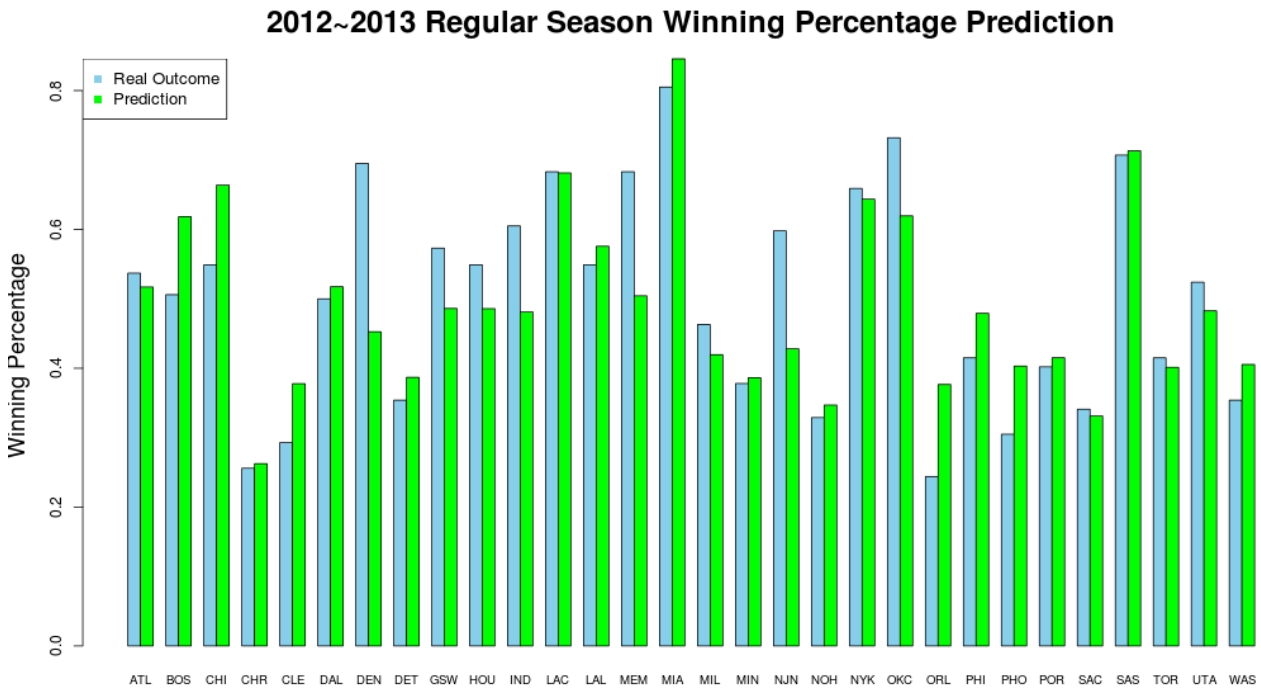
Basically, they are all in good shape.

Plot of correlation



As we can see, there are obvious correlation between win_ratio (regular season winning ratio) and avg_age (average team age), win_ratio and fgp (field goal percentage), win_ratio with num_allstar (number of all star), win_ratio with oast_dast (assists made by team / assists by opponents).

Prediction of regular season winning percentage of 2012 season using the linear regression model.

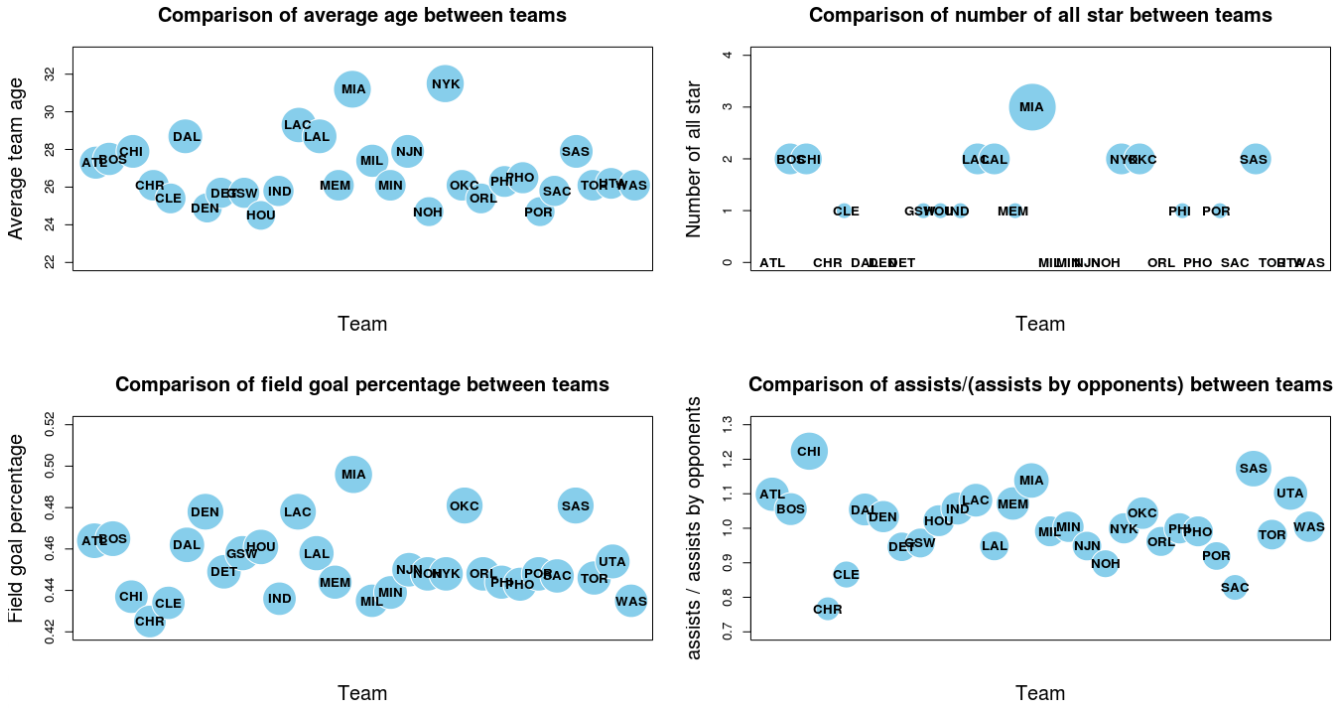


As we see from the above graph, the model gave very accurate results on teams like LAC, MIA, MIN and SAS. However, the deviation for DEN and NJN is quite large, suggesting that the model is not complete or fit on some teams.

On average, there is a 6.5 percent deviation.

Championship Prediction: Who's going to win this year's NBA championship ???

First, let's take a look at comparison of average team age, number of all star, field goal percentage and ratio of team's assists over opponents' assists among teams for 2012~2013 season.



Looks like MIA (Miami Heat) is more likely to win the championship, followed by SAS (San Antonio spurs).

Let's see what is the result given by the model.
The model used here is Logistic Regression which lends itself to binomial outcome prediction.
Logistic regression is used for predicting the outcome of a categorical dependent variable.
In this case, the categorical dependent variable is winning a championship or not. [7]

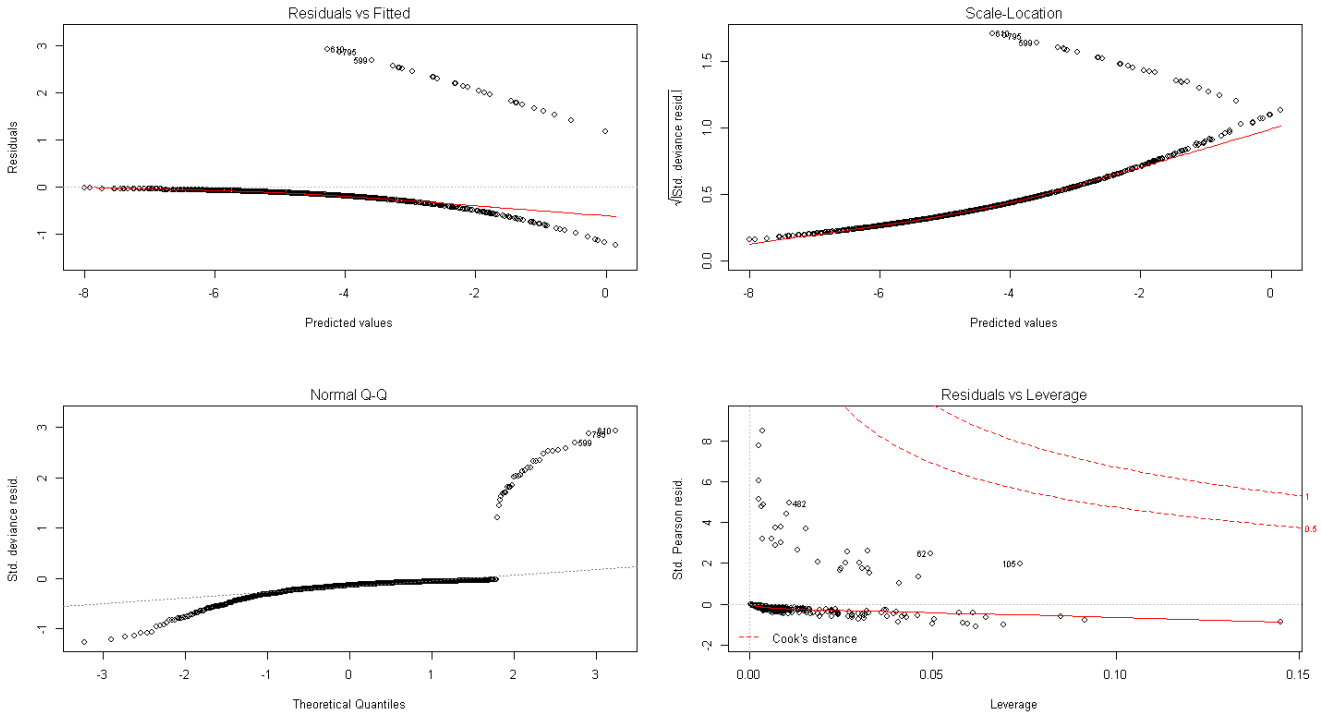
R code:

```
teams.t = read.csv('team_season_1980-2009.csv', header=TRUE)
tt.logm = glm(is_champ~fgp+num_allstar+avg_age+oast_dast, data=teams.t, family=binomial)
```

Summary of the model:

```
Call:
glm(formula = is_champ ~ fgp + num_allstar + avg_age + oast_dast,
    family = binomial, data = teams.t)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.24007  -0.23242  -0.14031  -0.08127   2.92886

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -28.7221     6.7790  -4.237 2.27e-05 ***
fgp          19.7895    10.3770   1.907  0.05651 .
num_allstar   0.6904     0.2427   2.845  0.00444 **
avg_age       0.3554     0.1514   2.347  0.01892 *
oast_dast     4.9703     1.9558   2.541  0.01104 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 256.78  on 811  degrees of freedom
Residual deviance: 192.67  on 807  degrees of freedom
AIC: 202.67

Number of Fisher Scoring iterations: 7
```

Default plot of the model

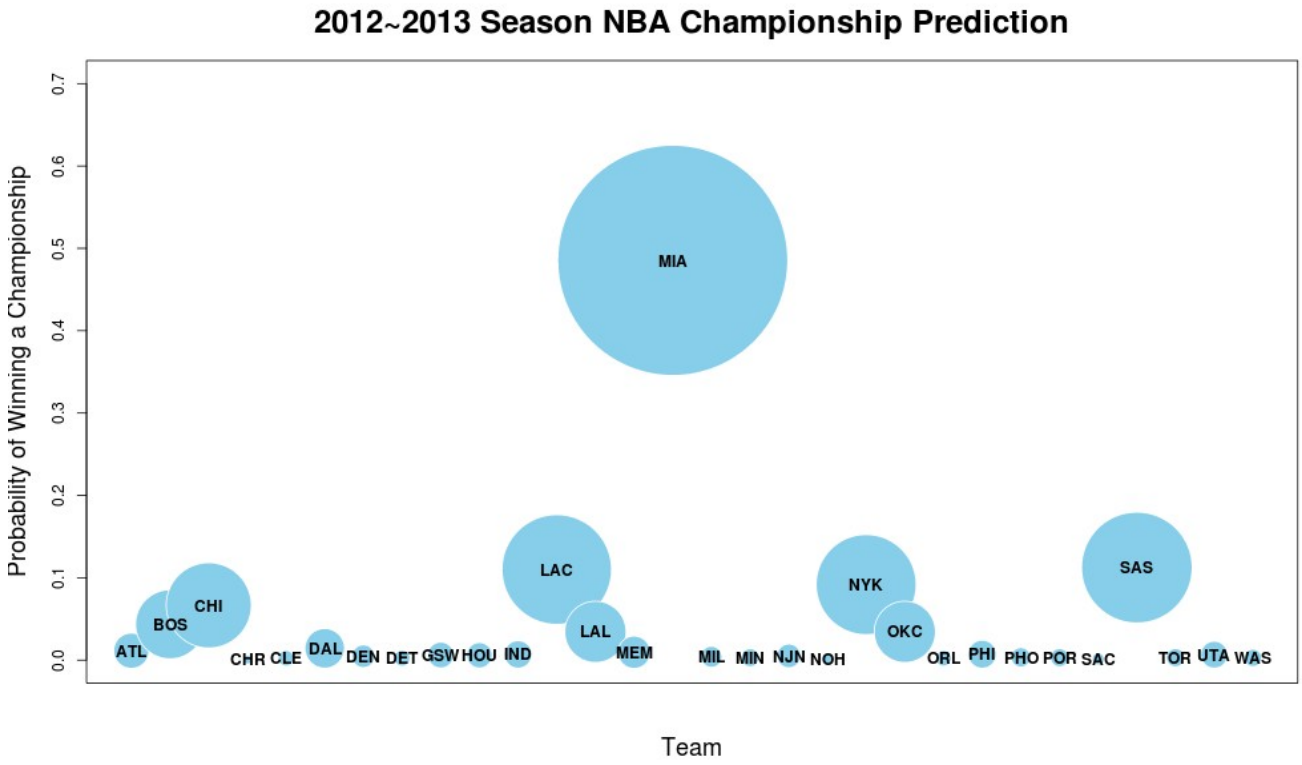


2013 NBA championship prediction

R code:

```
teams.p = read.csv('team_season_2010-2012.csv', header=TRUE)
p = predict(tt.logm, type="response", newdata=teams.p)
```

result:



Bigger circle means larger chance of winning a championship.

The prediction based on the model also indicates that MIA is more likely to win, its chance of winning a championship is about 48%.

Lessons learned

- Knowing the objective of the class is essential to the success of the class.
- Getting data ready for use was harder than I thought.
- Python + SQL is a good way of performing ETL.
- Solid statistical knowledge is the foundation of the data analysis and visualization.
- R rocks! Learning and using R was easier than I thought.
- I should have planned and started early.
- Project plan is also an iteration. It should be ready for change at any time.

Future works

- Performing more thoroughly check on data to make sure data is clean and consistent.
- Taking more factors into consideration since basketball is a complicated game, many factors can impact the game results.
- Using R's googlevis package to do animation.
- Applying more models on data and comparing the results given by each model.

References

[1] databaseBasketball [web] Available: http://www.databasebasketball.com
[2] NBA. (May 2 2008). NBA Rules History. [web]
Available: http://www.nba.com/history/finals/champions.html
[3] wikipedia. (April 23 2013) 2010 NBA Playoffs [Web]
Available: http://en.wikipedia.org/wiki/2010_NBA_Playoffs
[4] ESPN. NBA Player Scoring Per Game Statistics – 2012-13. [web] Available:
http://espn.go.com/nba/statistics/player/_/stat/scoring-per-game/sort/avgPoints/seasontype/2
[5] NBA. (May 2 2008). NBA Rules History. [web]
Available: http://www.nba.com/analysis/rules_history.html
[6] Paul Van Slembrouck. (May 08, 2011). I Know Who's Going to Win the NBA Finals. [web]
Available: http://www.paulvanslembrouck.com/2011/i-know-whos-going-to-win-the-nba-finals/
[7] wikipedia. (April 28 2013) Logistic regression [Web]
Available: http://en.wikipedia.org/wiki/Logistic_regression
[8] Matthew Beckler, Hongfei Wang, Michael Papamichael. (Spring 2009). NBA Oracle [web].
Available: http://www.mbeckler.org/coursework/2008-2009/10701_report.pdf