Stories Behind NBA Stats

-- Shuo Yang

Agenda

- Introduction
- Dataset & KPIs
- Data Analysis & Visualization
- Prediction Results
- Project Plan
- Lessons Learned & Future works

Introduction

- Objectives
  - Explore overall trends of the NBA.
  - What are the key factors that make a team win a championship?
  - Correlation between regular season winning percentage and KPIs.
  - Predicting 2012 ~ 2013 season's NBA championship

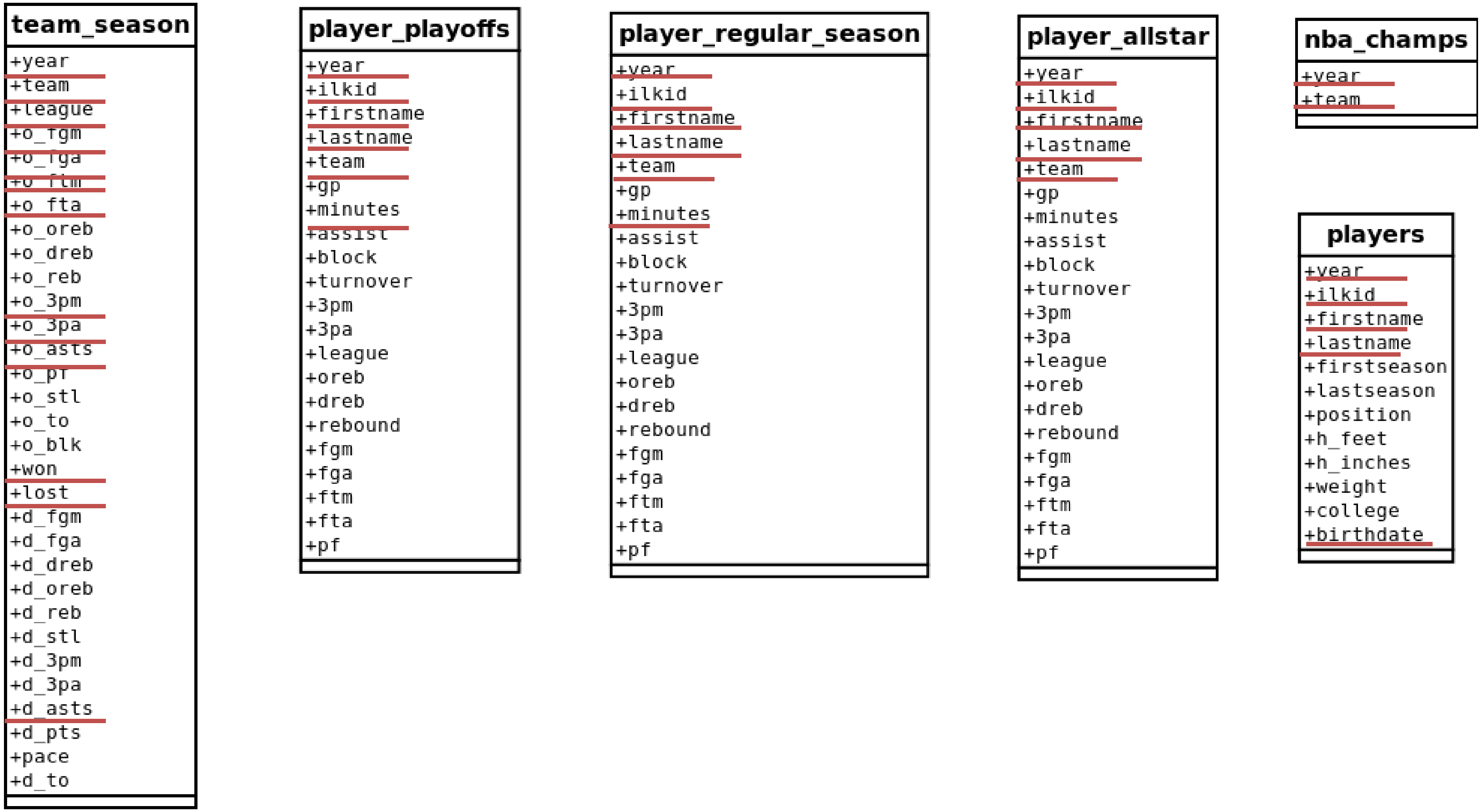
Introduction

- Motivation
  - I am a big fan of basketball.
  - The business of professional sports like NBA is a multi-billion dollar industry.

Dataset

- Main Data Source
  - DatabaseBasketball website [1]
- Other Sources
  - NBA official website
  - Wikipedia
  - ESPN

Table View of Data



team_season
+year
+team
+league
+o_fgm
+o_fga
+o_ftm
+o_fta
+o_oreb
+o_dreb
+o_reb
+o_3pm
+o_3pa
+o_asts
+o_pf
+o_stl
+o_to
+o_blk
+won
+lost
+d_fgm
+d_fga
+d_dreb
+d_oreb
+d_reb
+d_stl
+d_3pm
+d_3pa
+d_asts
+d_pts
+pace
+d_to

player_playoffs
+year
+ilkid
+firstname
+lastname
+team
+gp
+minutes
+assist
+block
+turnover
+3pm
+3pa
+league
+oreb
+dreb
+rebound
+fgm
+fga
+ftm
+fta
+pf

player_regular_season
+year
+ilkid
+firstname
+lastname
+team
+gp
+minutes
+assist
+block
+turnover
+3pm
+3pa
+league
+oreb
+dreb
+rebound
+fgm
+fga
+ftm
+fta
+pf

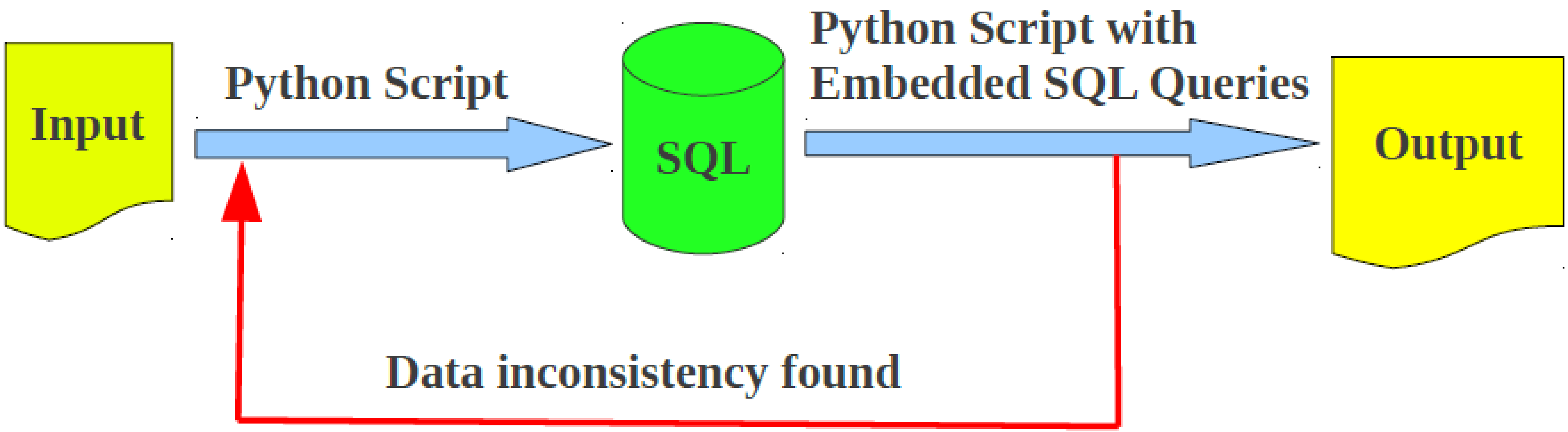
player_allstar
+year
+ilkid
+firstname
+lastname
+team
+gp
+minutes
+assist
+block
+turnover
+3pm
+3pa
+league
+oreb
+dreb
+rebound
+fgm
+fga
+ftm
+fta
+pf

nba_champs
+year
+team

players
+year
+ilkid
+firstname
+lastname
+firstseason
+lastseason
+position
+h_feet
+h_inches
+weight
+college
+birthdate

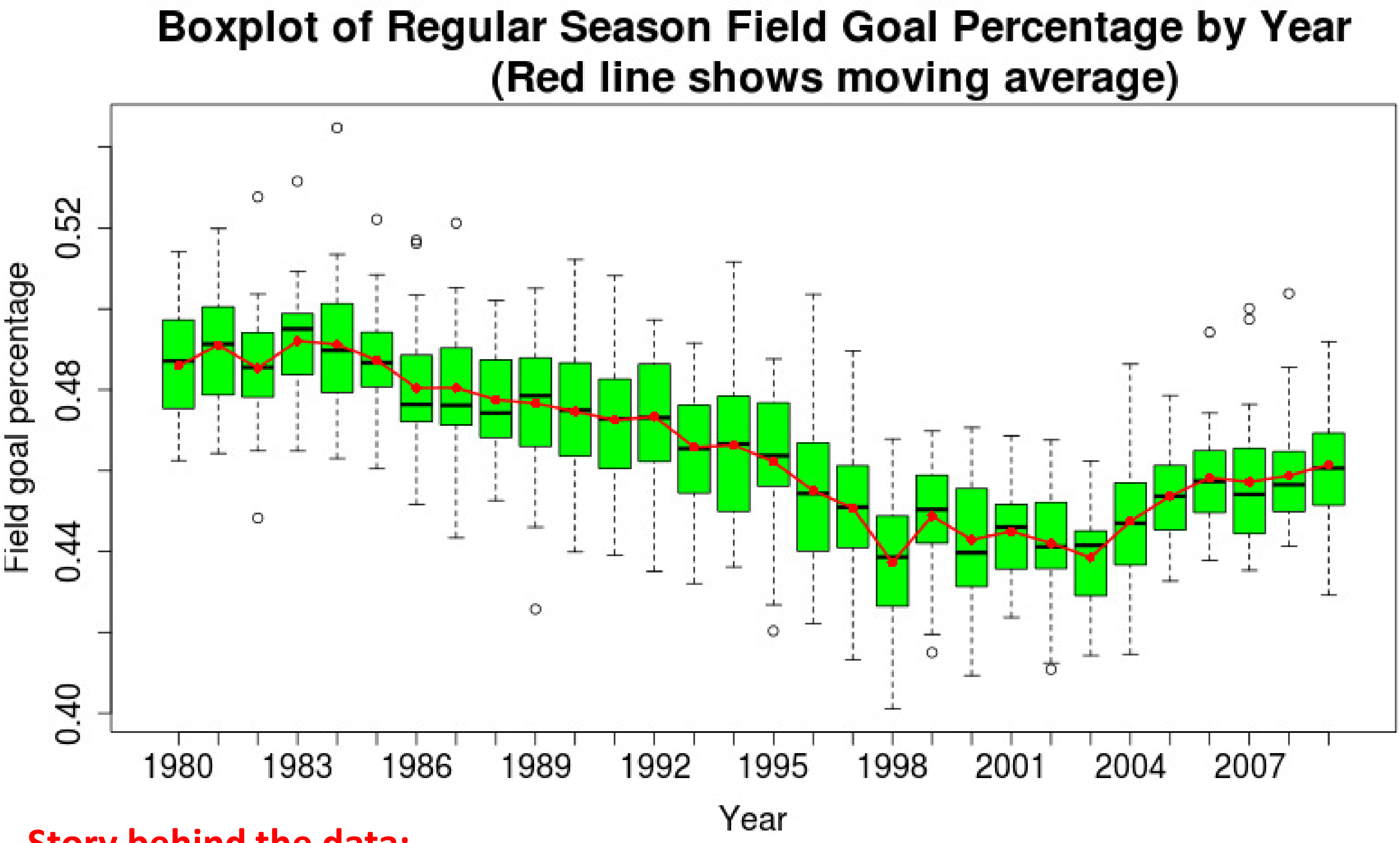
Preliminary KPIs		Refined KPIs
• ratio of assist made by team over assist made by opponents	→	• ratio of assist made by team over assist made by opponents
• regular season field goal percentage per team	→	• regular season field goal percentage per team
• regular season three-point percentage per team	→	• regular season three-point percentage per team
• average team age weighted by minutes played	→	• average team age weighted by minutes played
• number of all star per team	→	• number of all star per team
• regular season free throw percentage per team	→	**Dropped**

ETL Iteration



Data Analysis & Visualization

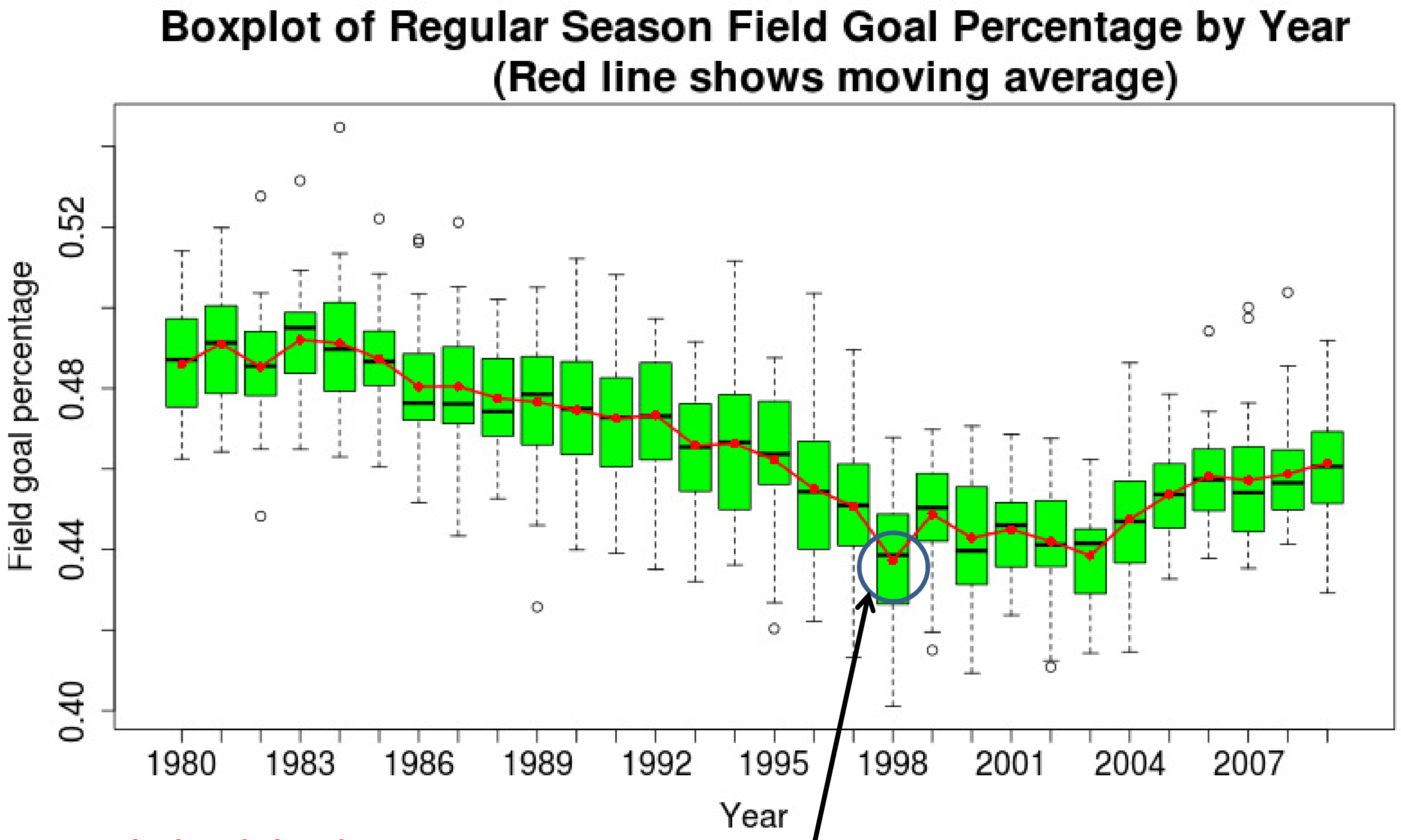
How did field goal percentage change over time ?



Story behind the data:

Overall decreasing of field goal percentage

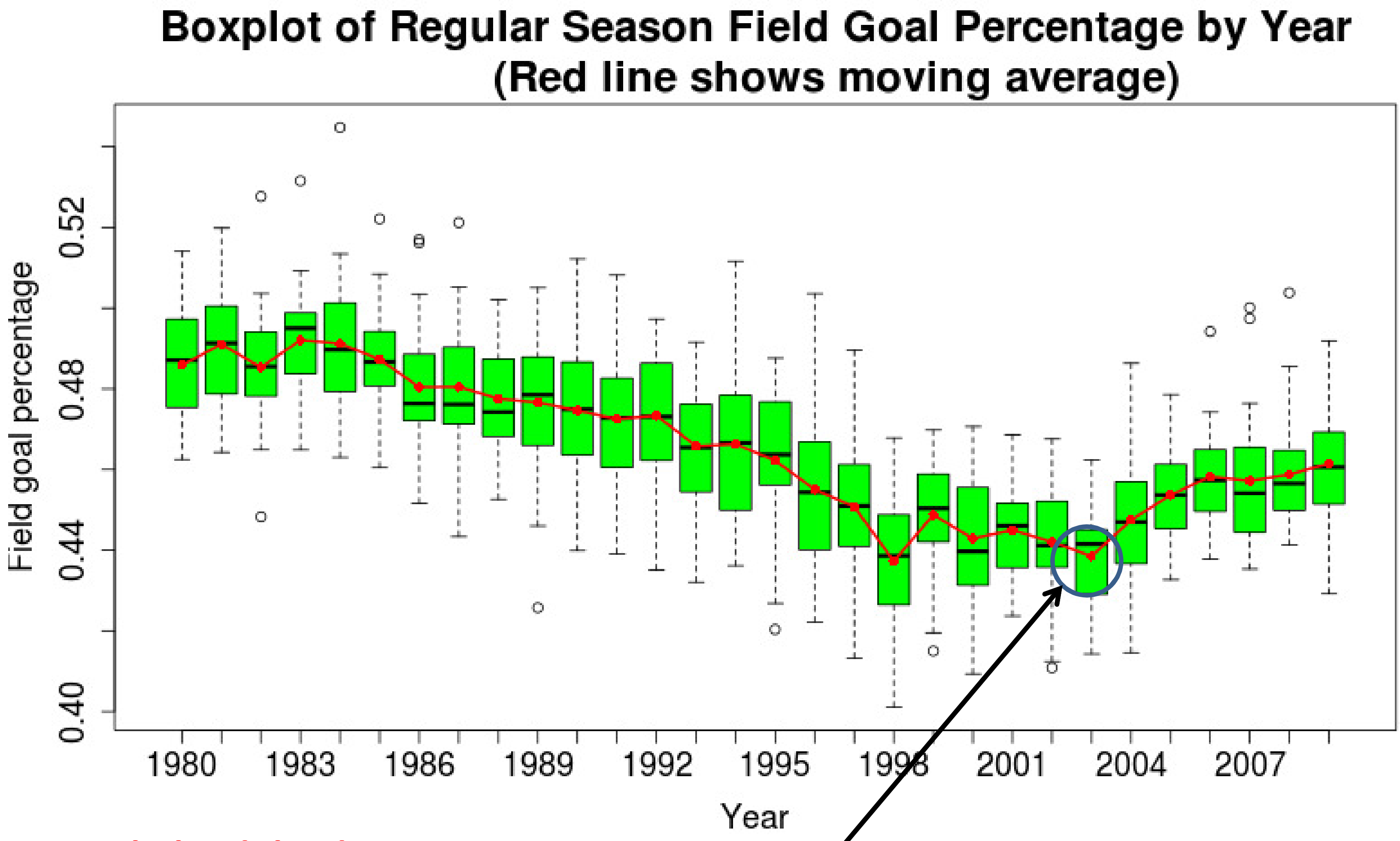
- League has become more and more physically demanding.
- Players' and teams' defensive skills have been improving
- Scoring inside the paint has become more and more difficult.
- Many teams now rely on jump shot which drags down the field goal percentage.



Story behind the data:

The lowest average in 30 years shows the impact of 1998 ~ 1999 season long time which lasted over six month.

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



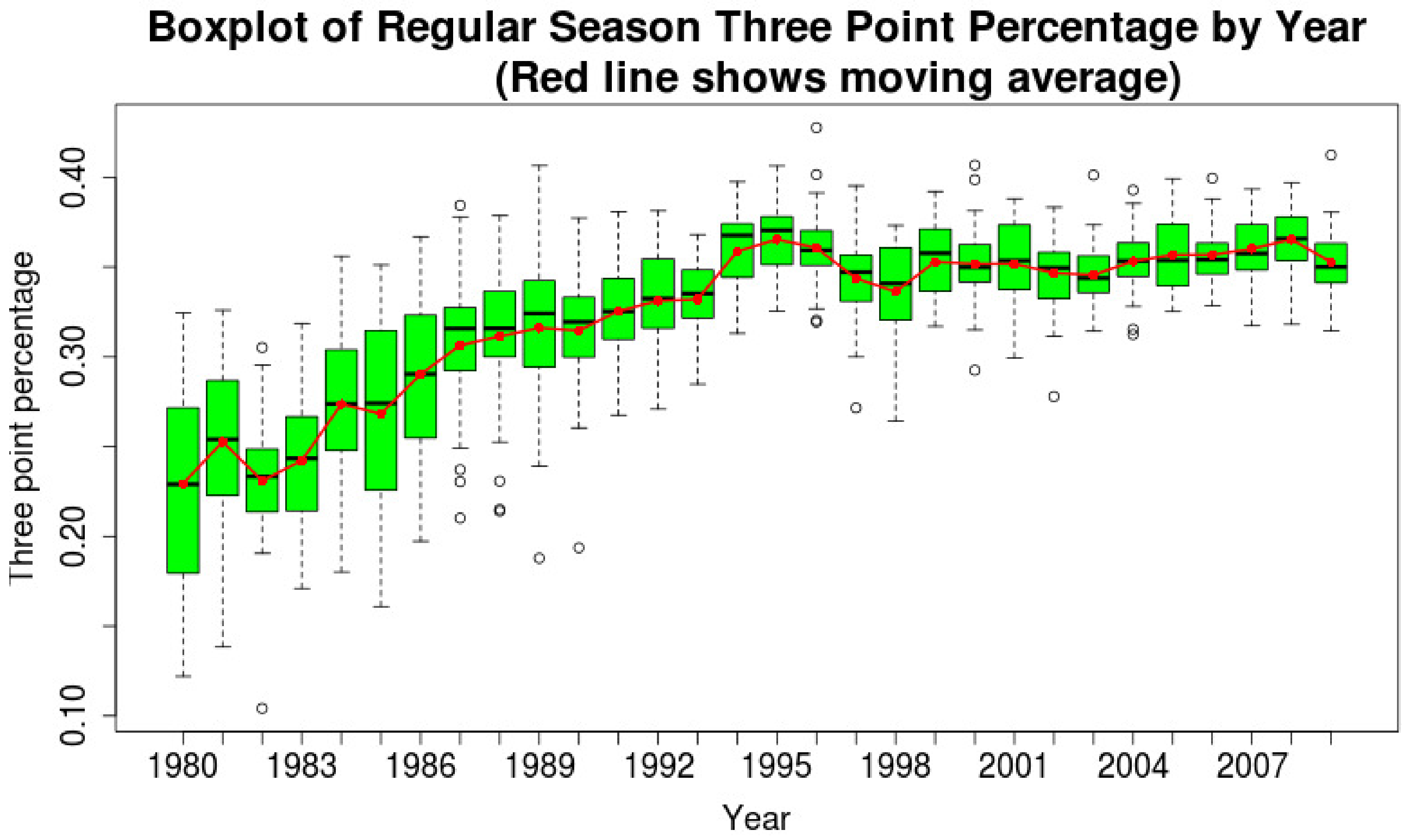
Story behind the data:

Field goal percentage has been improving since 2003-2004 season

- Many talented players emerged from colleges or other countries, like LeBron James who entered NBA in 2003, Dirk Nowitzki.
- The competitive environment pushes players to practice and play hard in order to stay on contract.

Data Analysis & Visualization

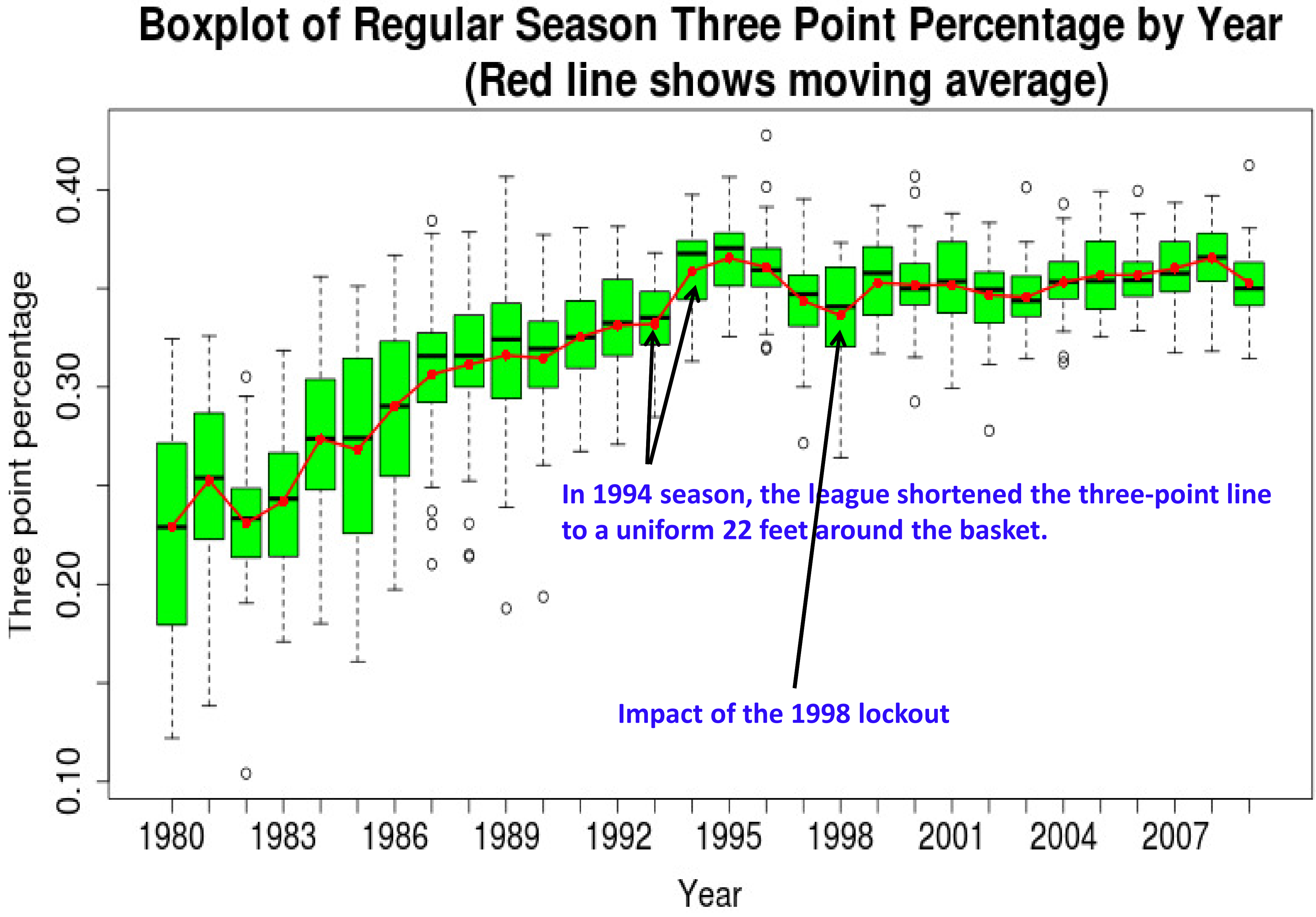
How did three-point percentage change over time ?



Stories behind the data:

The overall increasing of three-point percentage

- There is a 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.

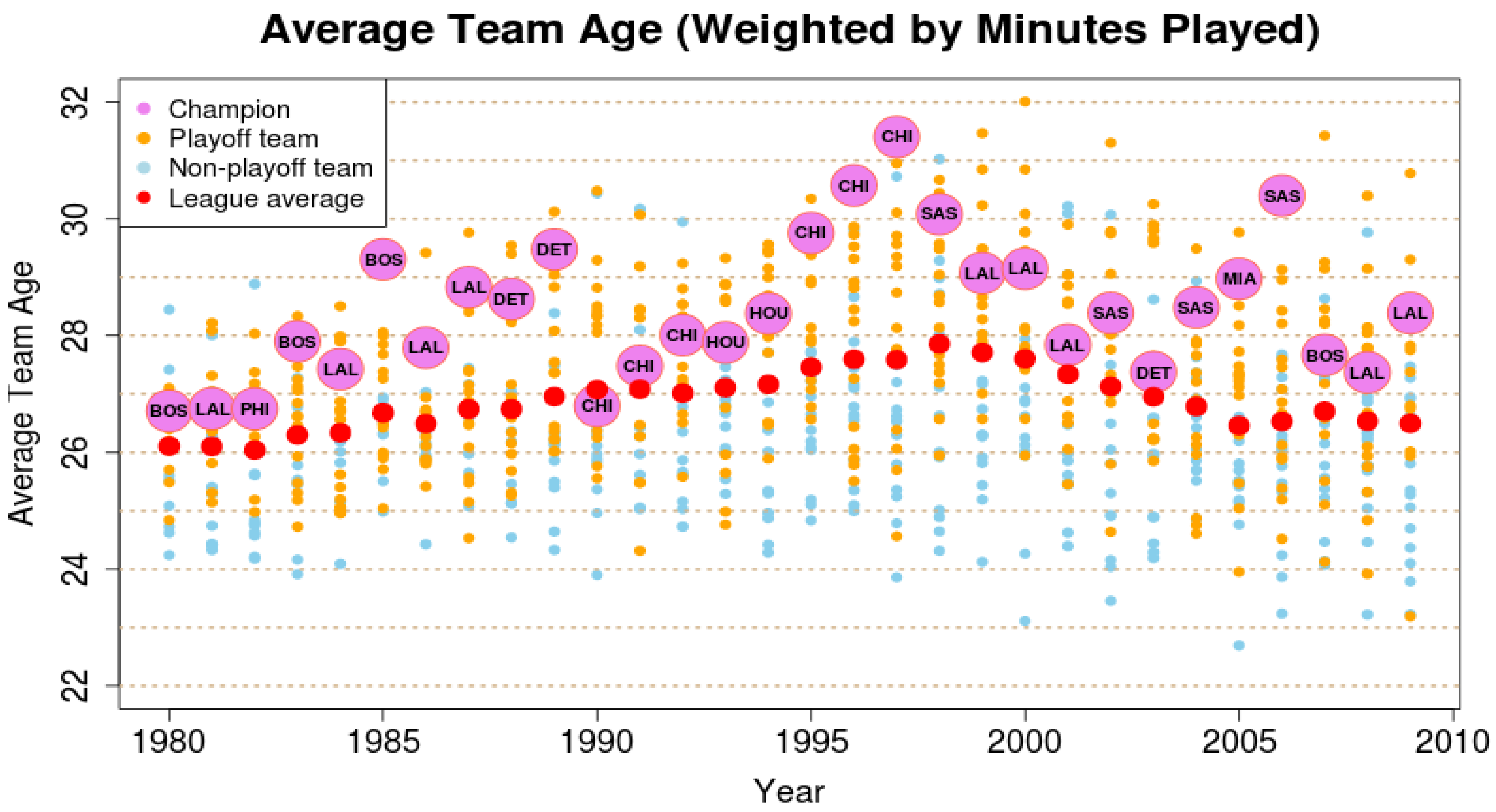


In 1994 season, the league shortened the three-point line to a uniform 22 feet around the basket.

Impact of the 1998 lockout

Data Analysis & Visualization

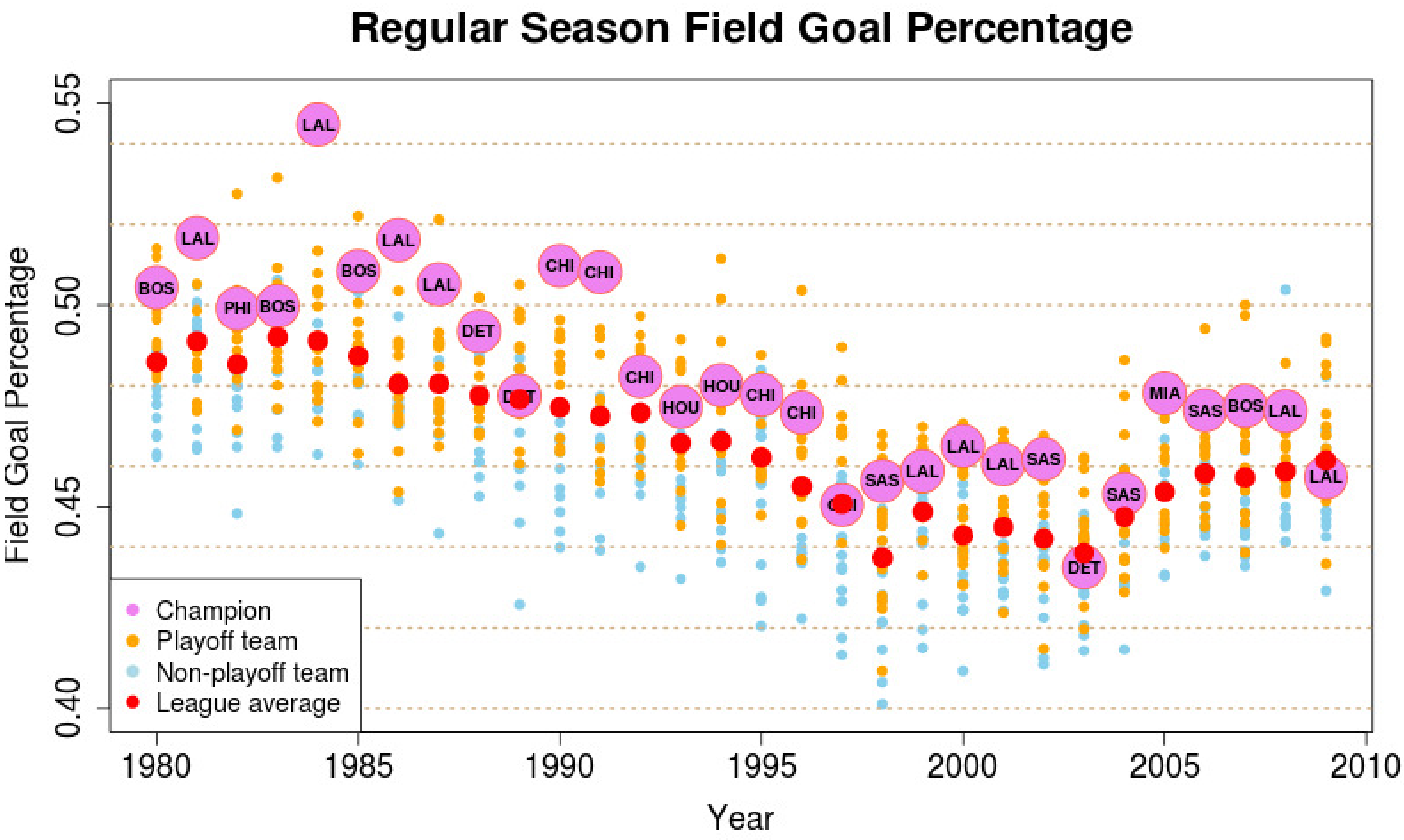
What makes a team a championship ?



Stories behind the data:

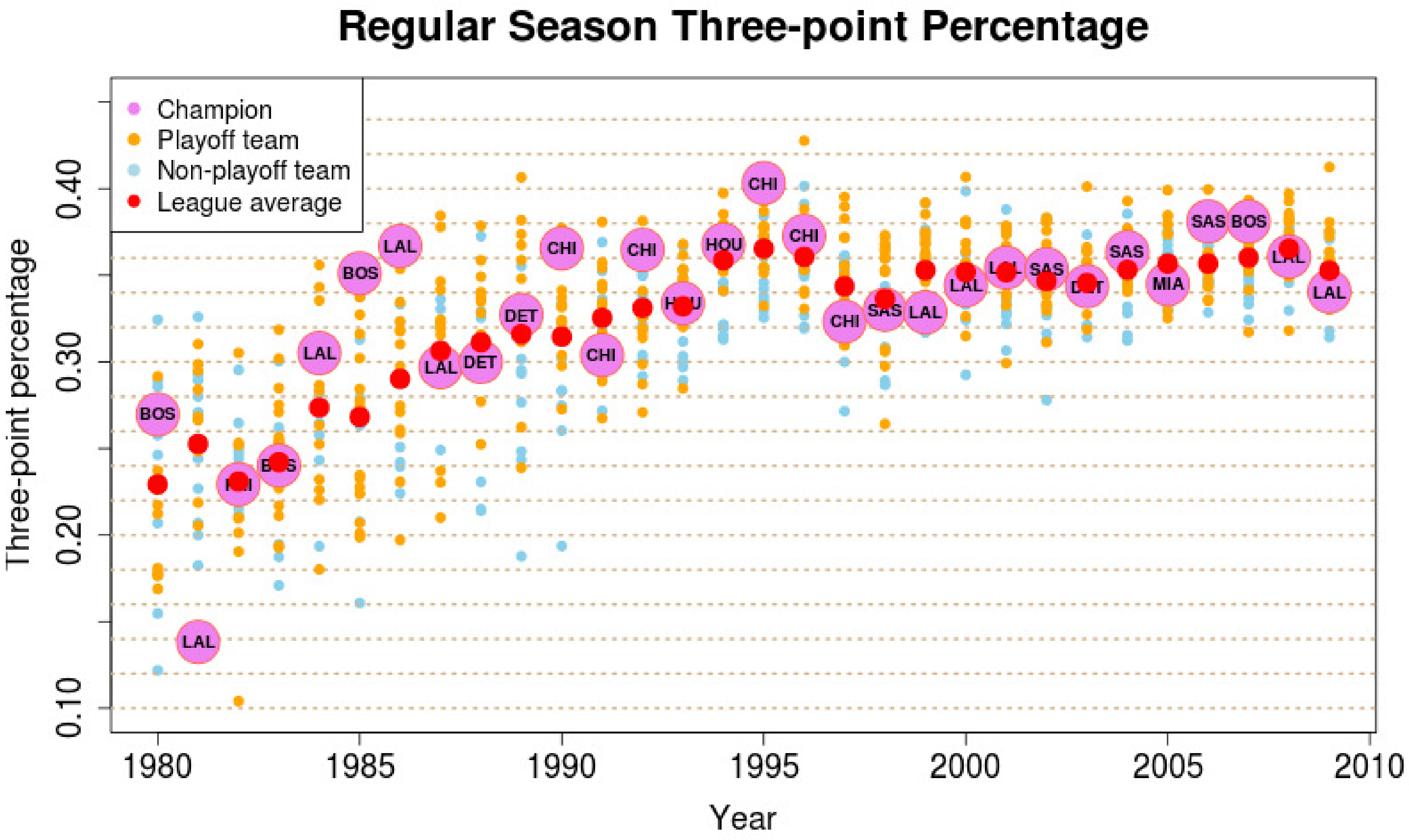
Championship teams are significantly older than league average, and they continue to win as they get older [6].

- 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.



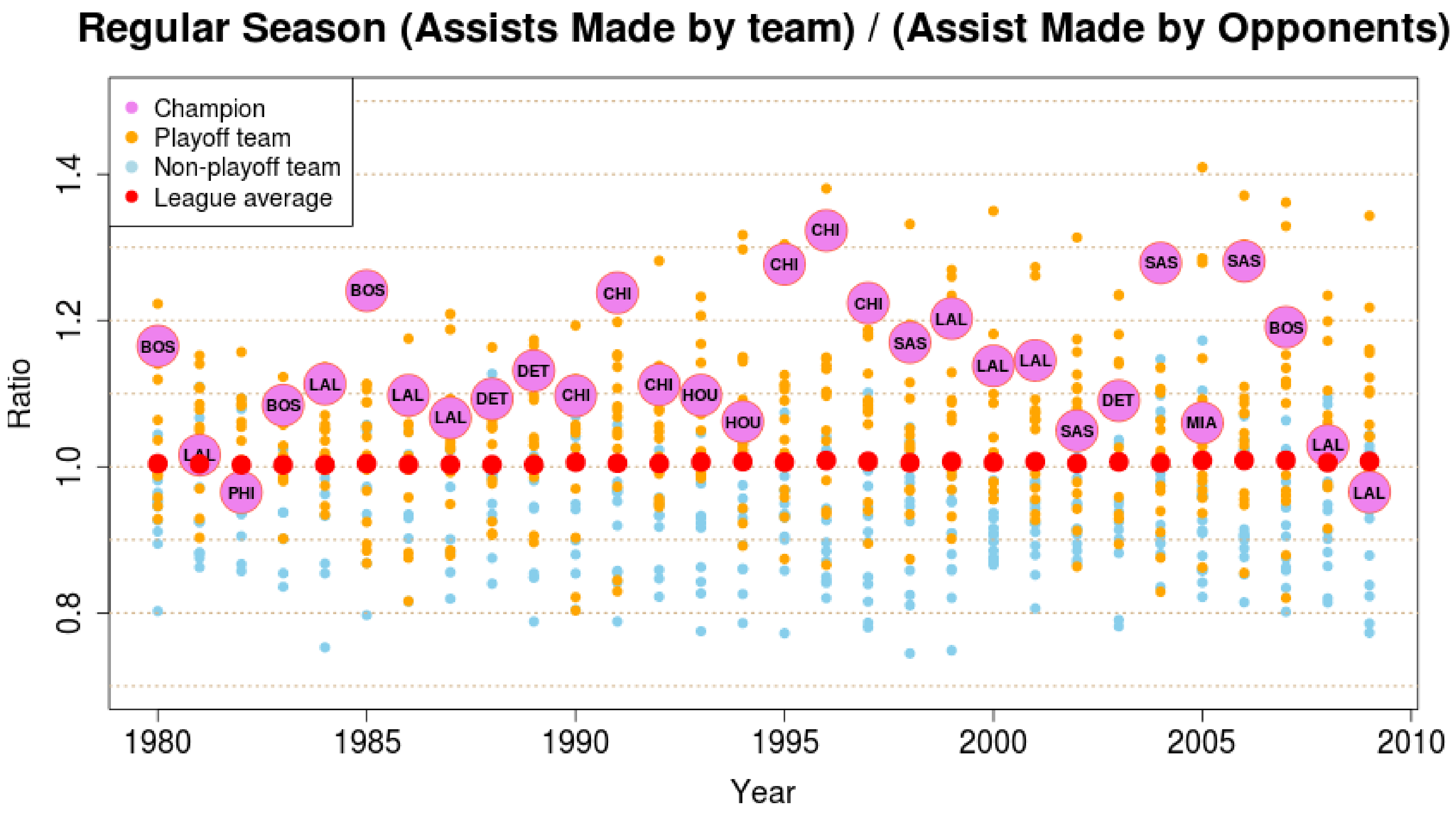
Stories behind the data:

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



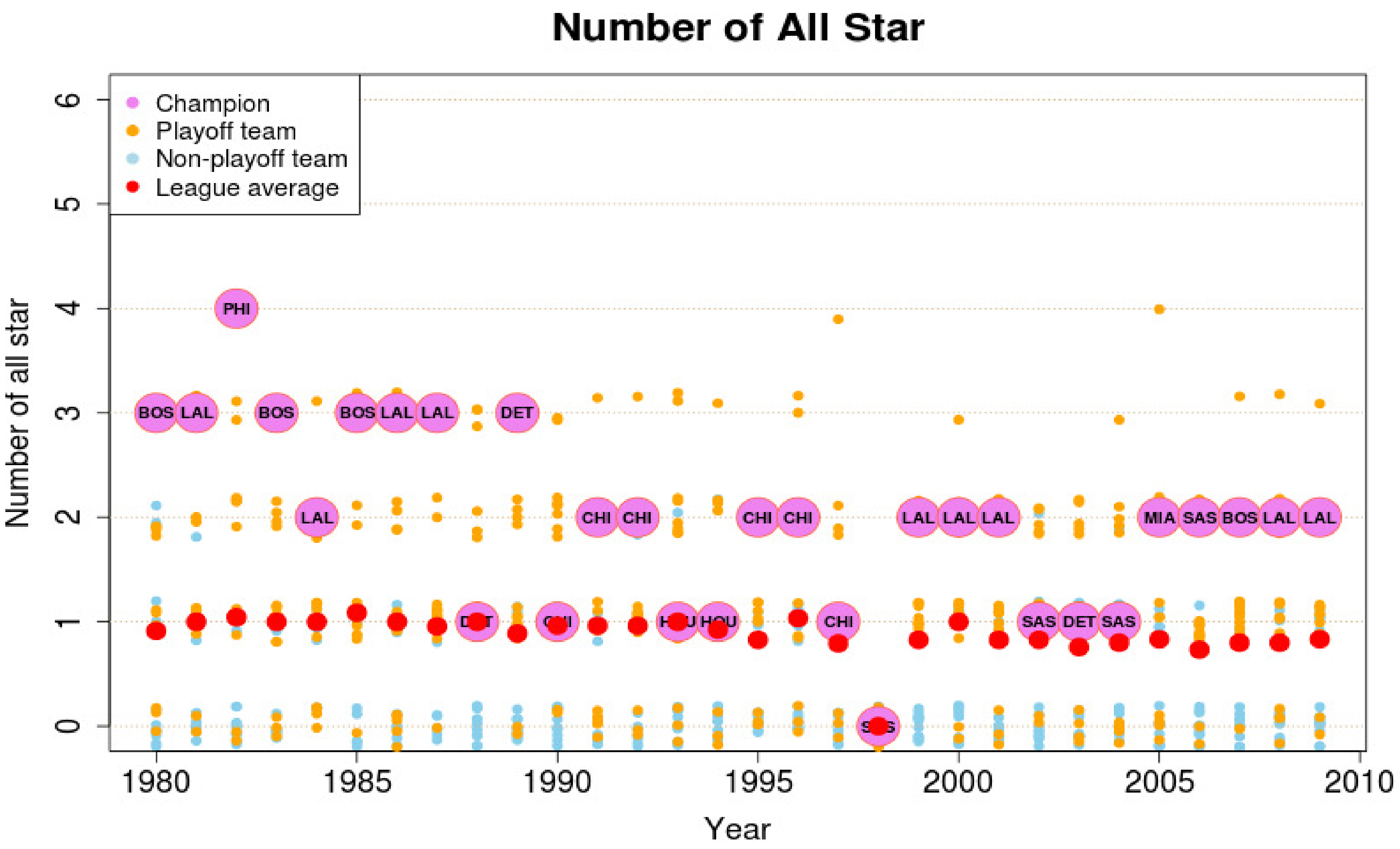
Stories behind the data:

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.

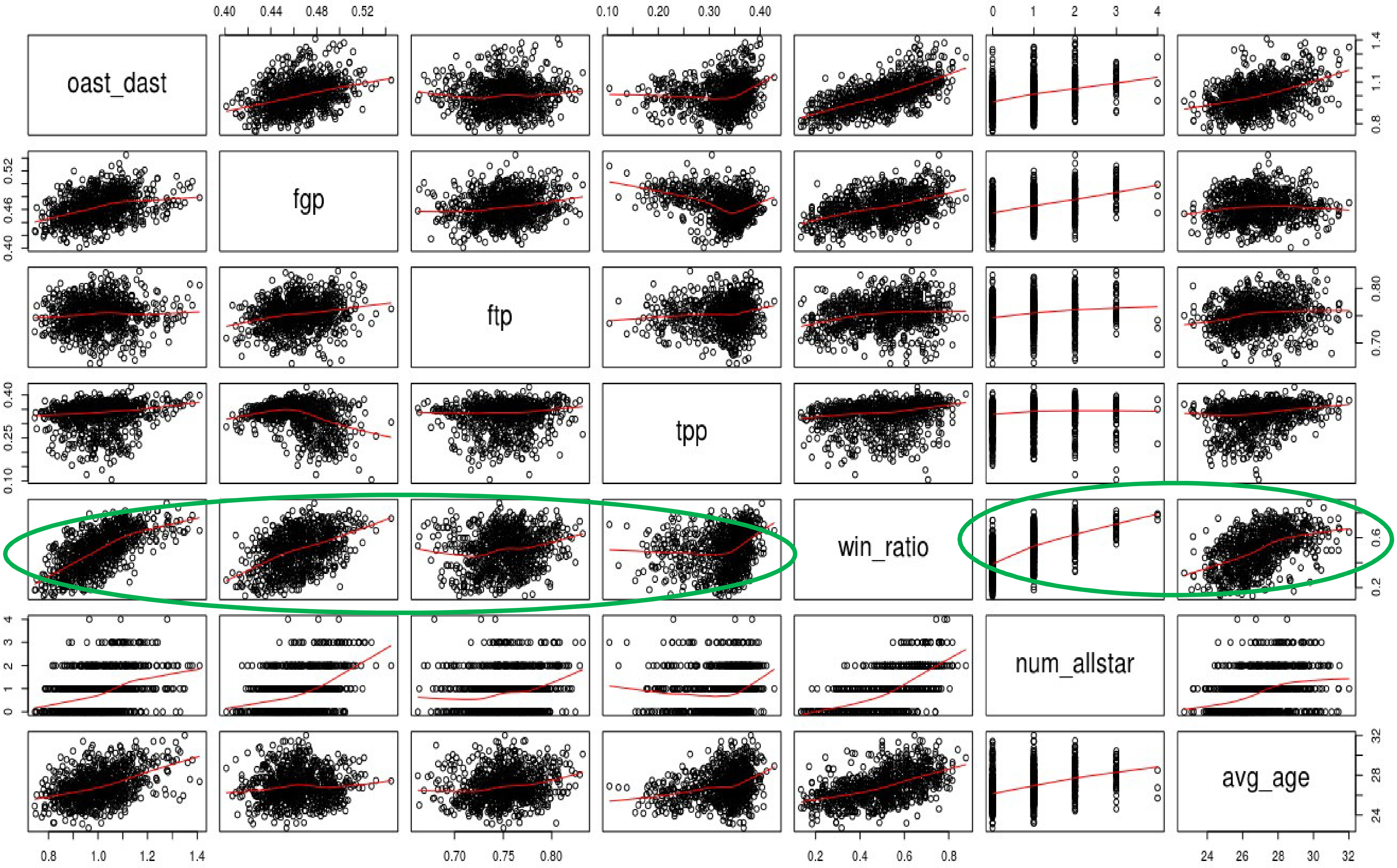


Stories behind the data:

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

Data Analysis & Visualization

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



win_ratio (regular season winning ratio)
avg_age (average team age)
fgp (field goal percentage)

tfp : free throw percentage
tpp : three-point percentage
num_allstar : number of all-star
oast_dast : assists made by team / assists by opponents

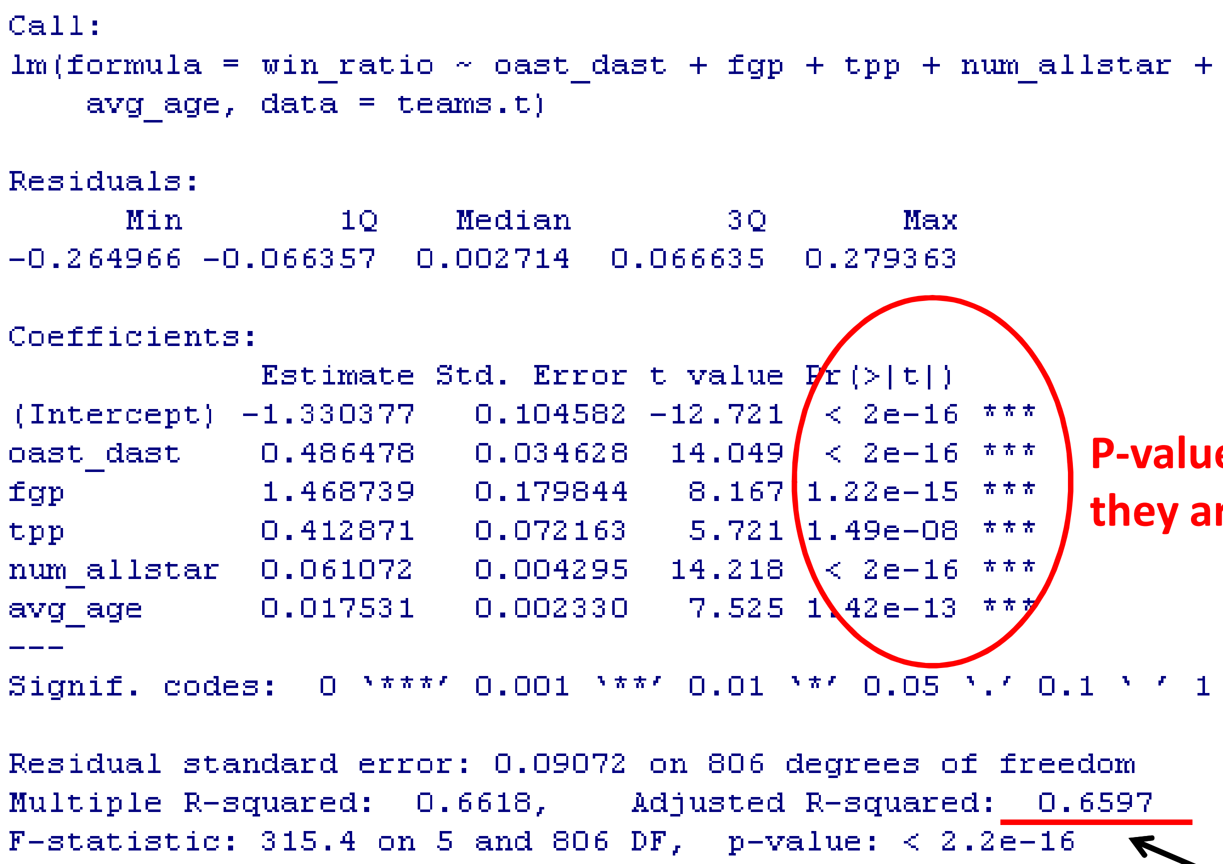
Prediction

Applying Linear Regression to predict 2012 ~ 2013 regular season winning percentage for each team.

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)

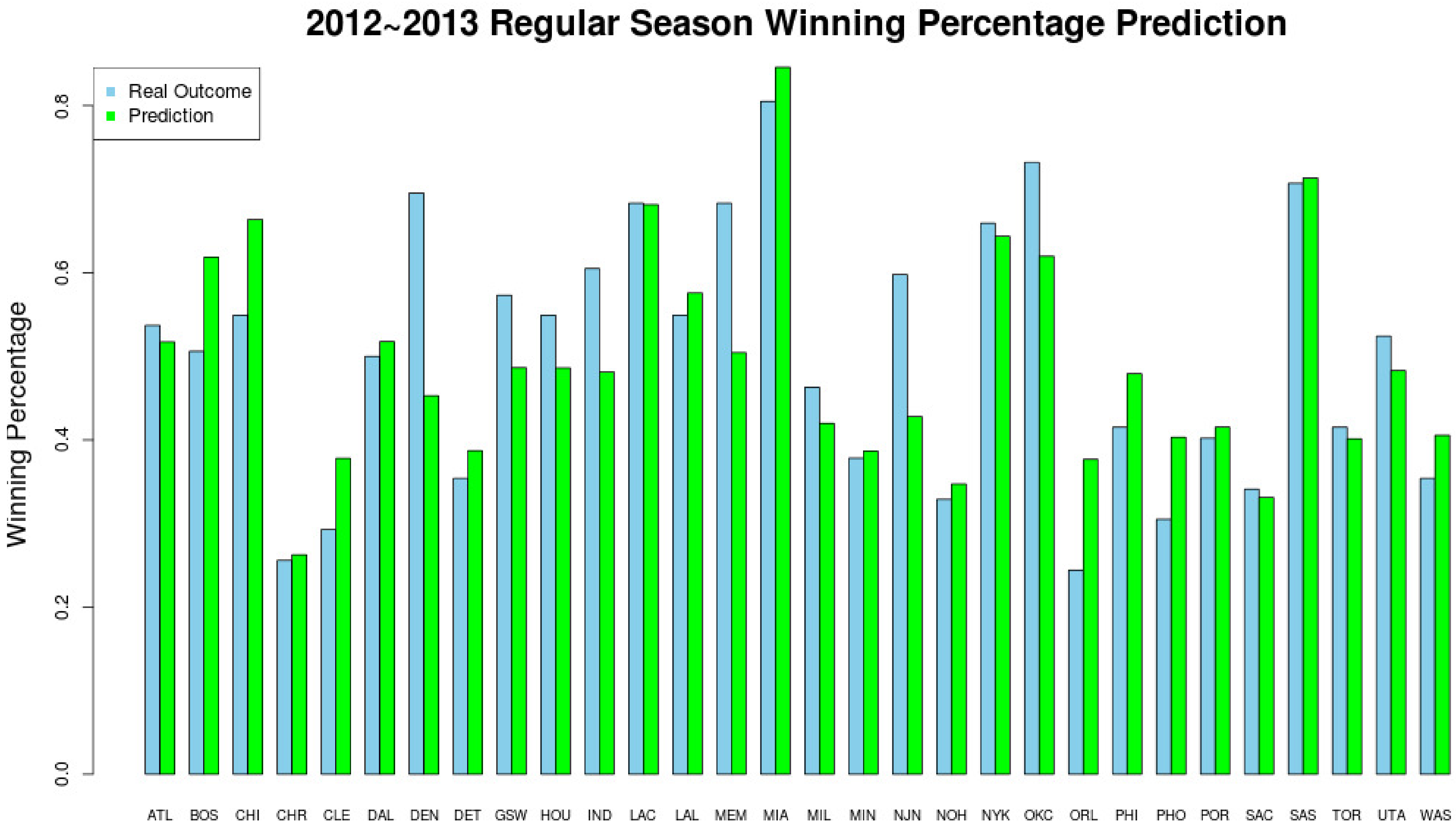
Summary of the Linear Regression Model



```
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
```

P-value shows that
they are all statistical significant

Indicates the model fits well

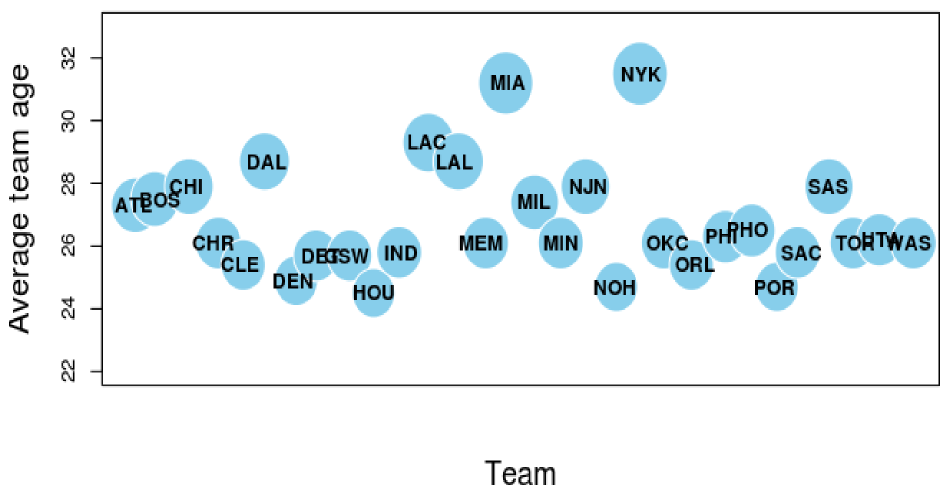


On average, there is a 6.5 percent deviation. More factors need to be considered !

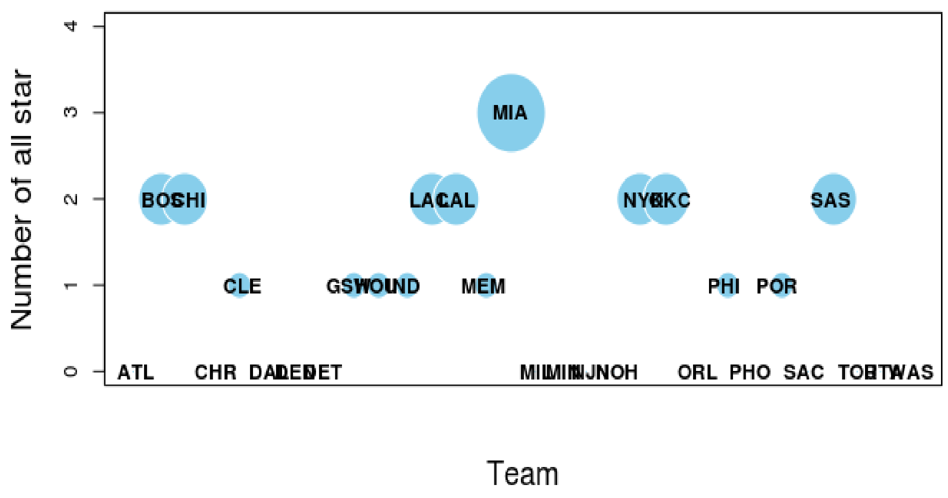
Prediction

Which team is going win 2012 ~ 2013 NBA Championship ?

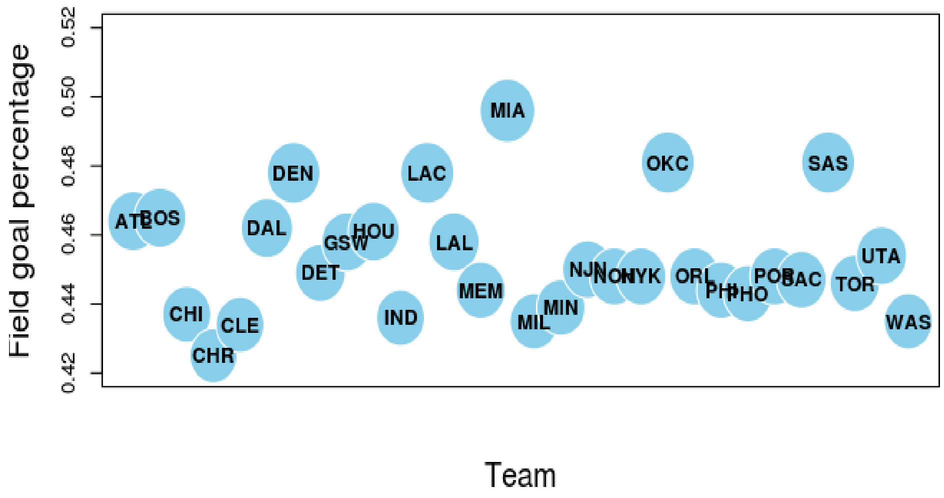
Comparison of average age between teams



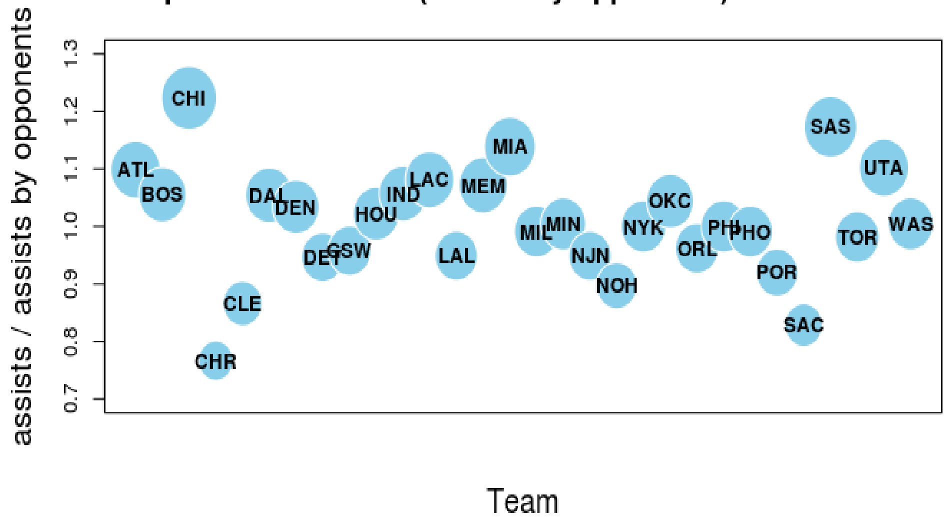
Comparison of number of all star between teams



Comparison of field goal percentage between teams



Comparison of assists/(assists by opponents) between teams



MIA (Miami Heat) is outstanding in all 4 factors. Followed by SAS (San Antonio Spurs).

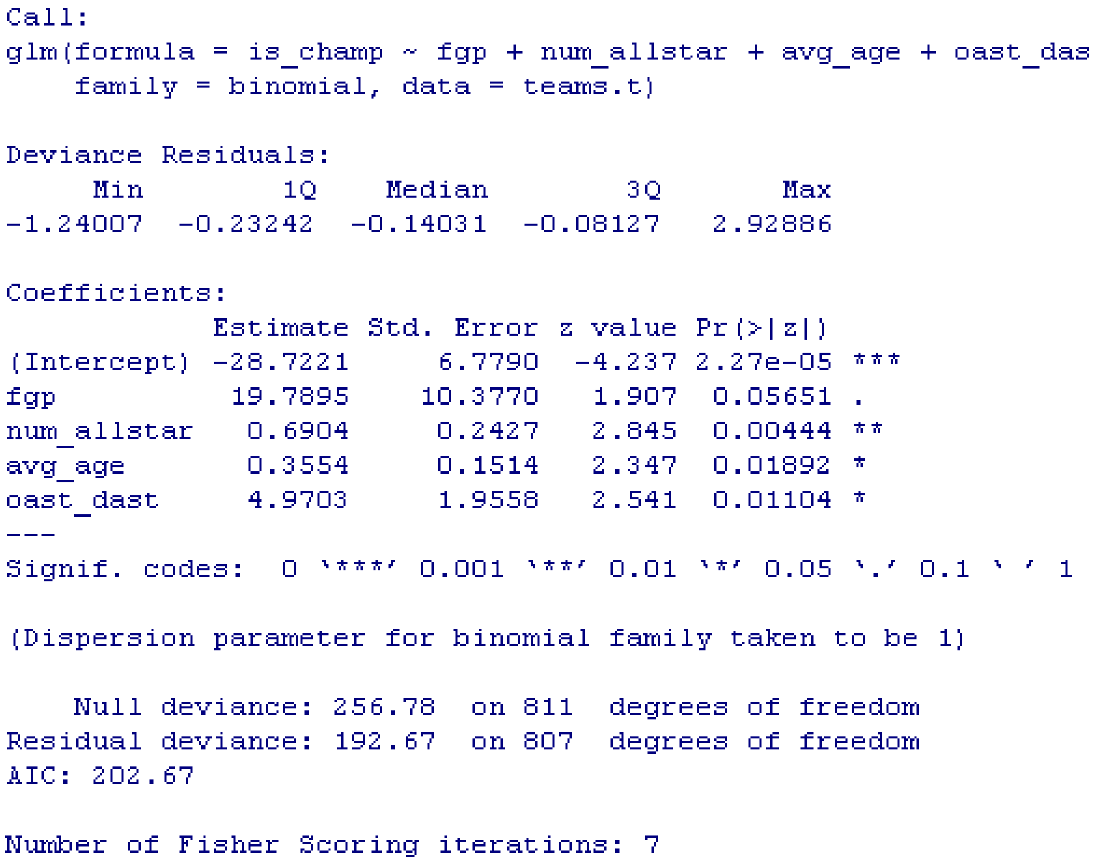
Prediction

Applying Logistic Regression to predict which team is going win 2012 ~ 2013 NBA Championship ?

R code:

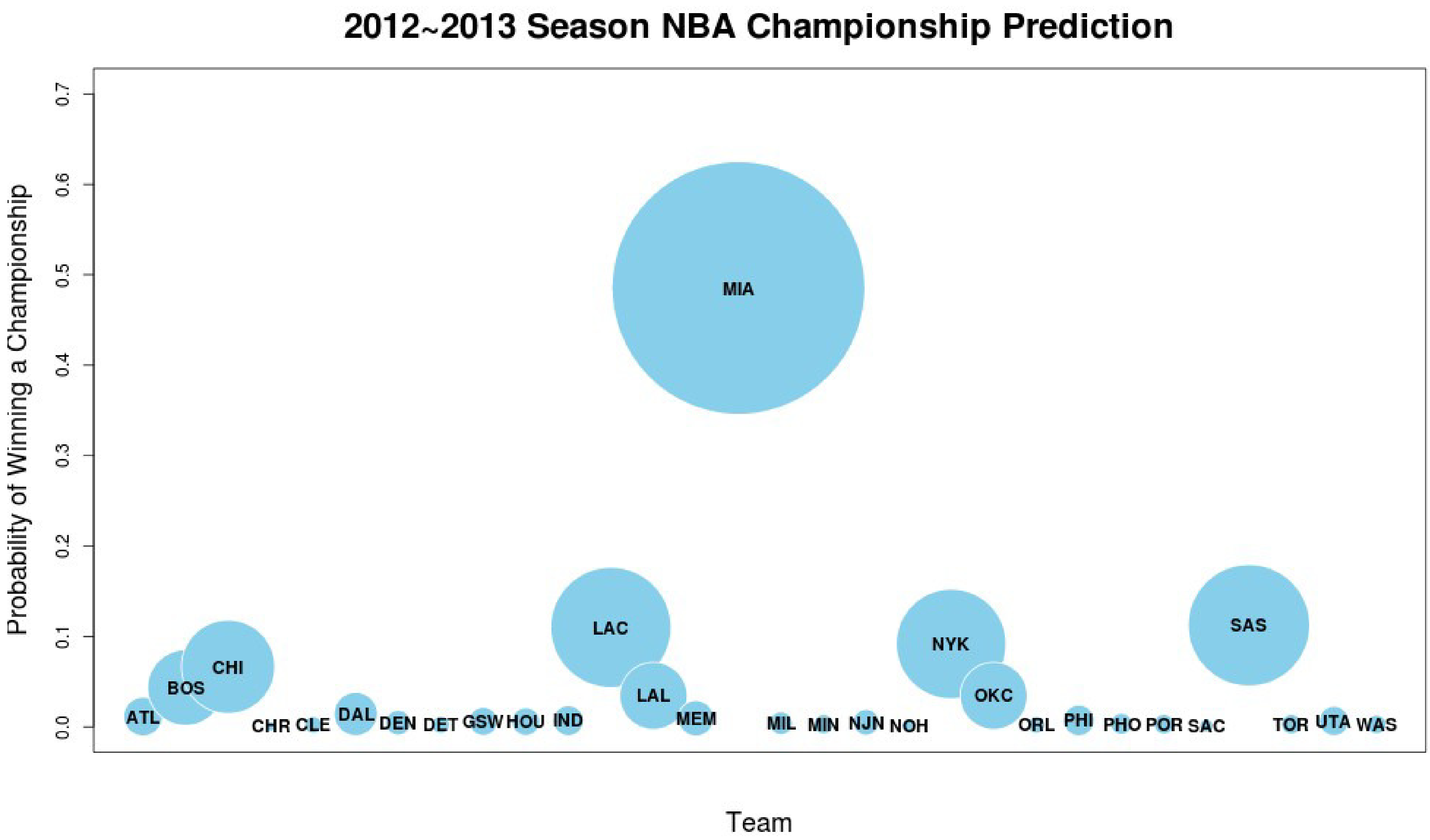
```
teams.t = read.csv('team_season_1980-2009.csv', header=TRUE)
tt.logm = glm(is_champ~fgp+num_allstar+avg_age+oast_dast, data=teams.t, family=binomial)
```

Summary of the Logistic Regression 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
```

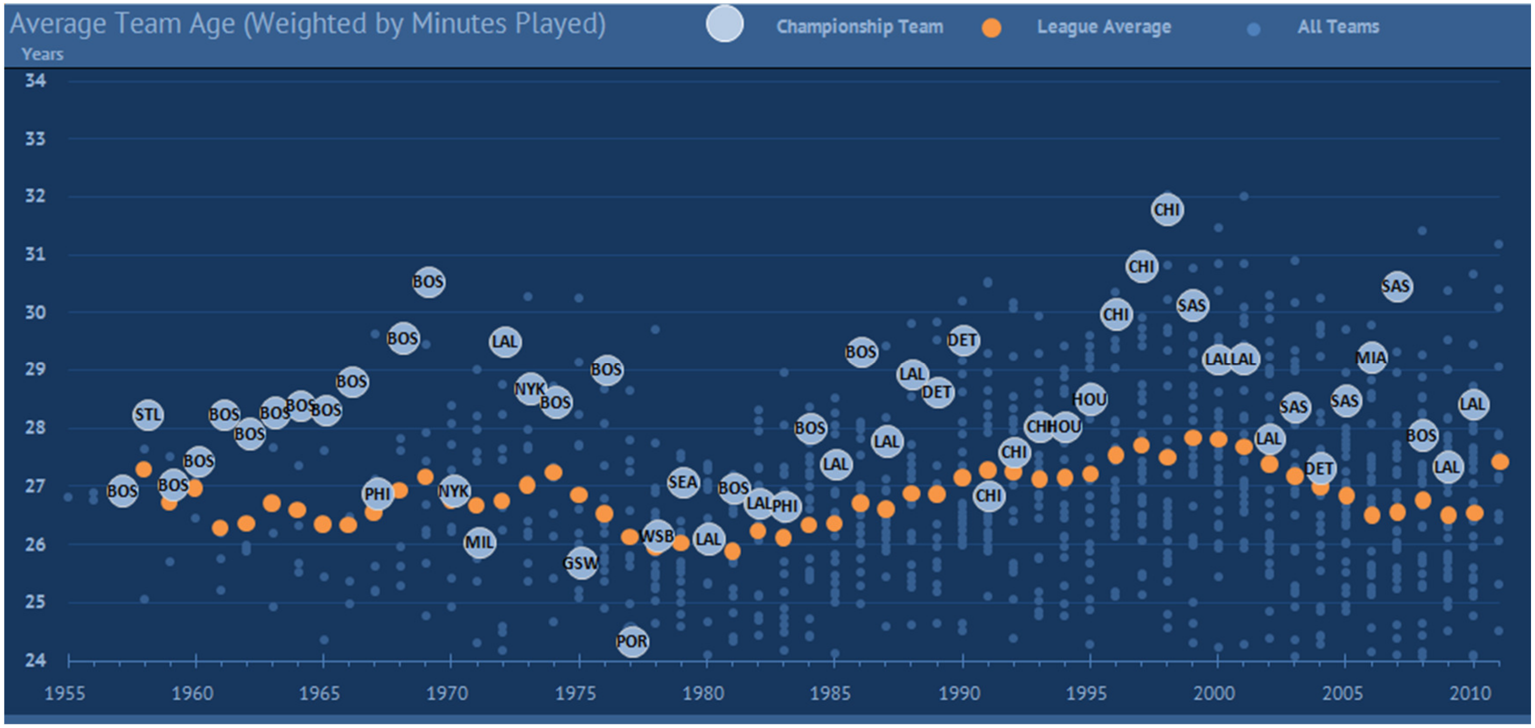
Prediction Result Generated by the Logistic Regression Model

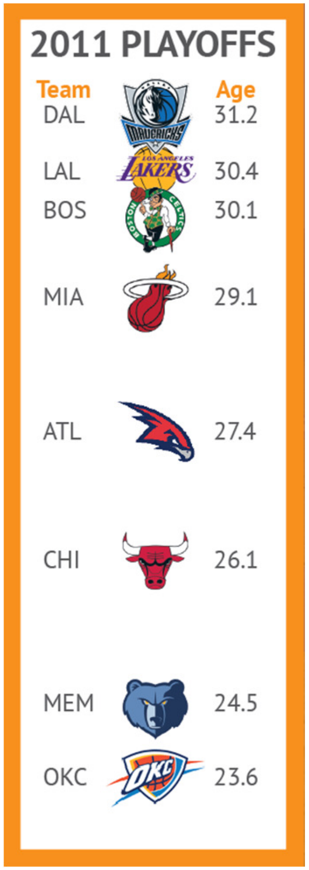


The chance of Miami Heat (MIA) winning a championship calculated by the model is about 48% which is way ahead of other teams !

Comparing my work with works done by others

Works done by Paul Van Slembrouck [6]





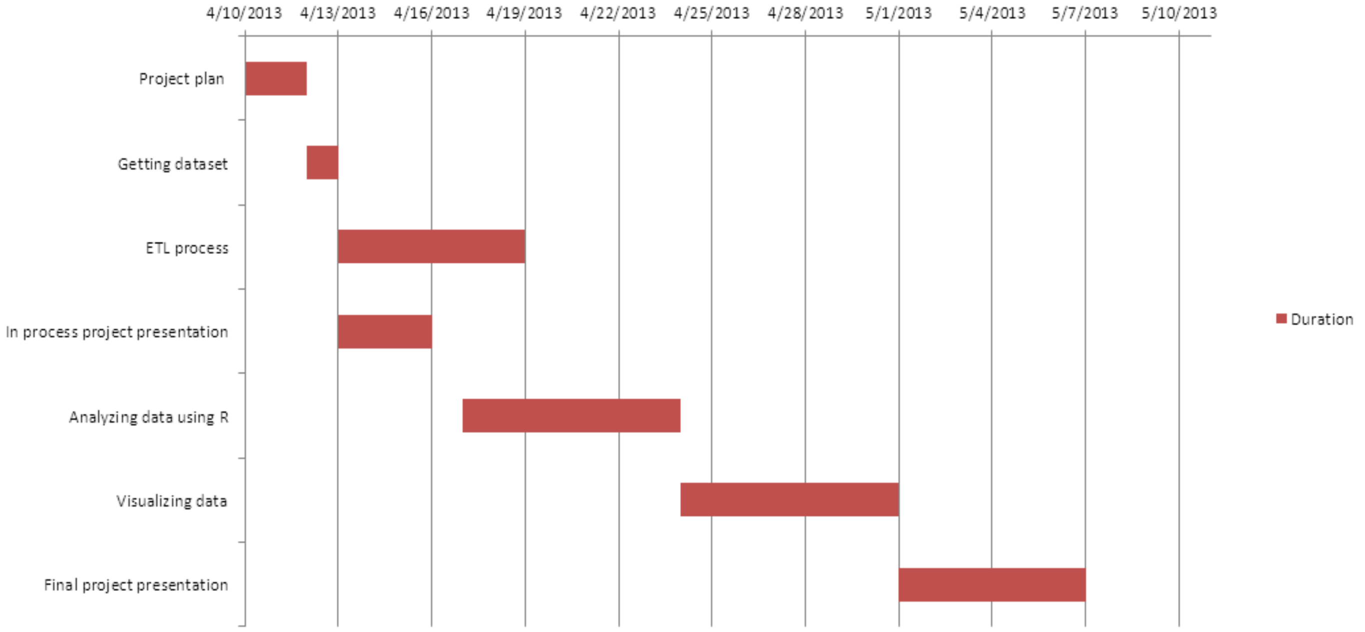
2011 PLAYOFFS

Team	Age
DAL	31.2
LAL	30.4
BOS	30.1
MIA	29.1
ATL	27.4
CHI	26.1
MEM	24.5
OKC	23.6

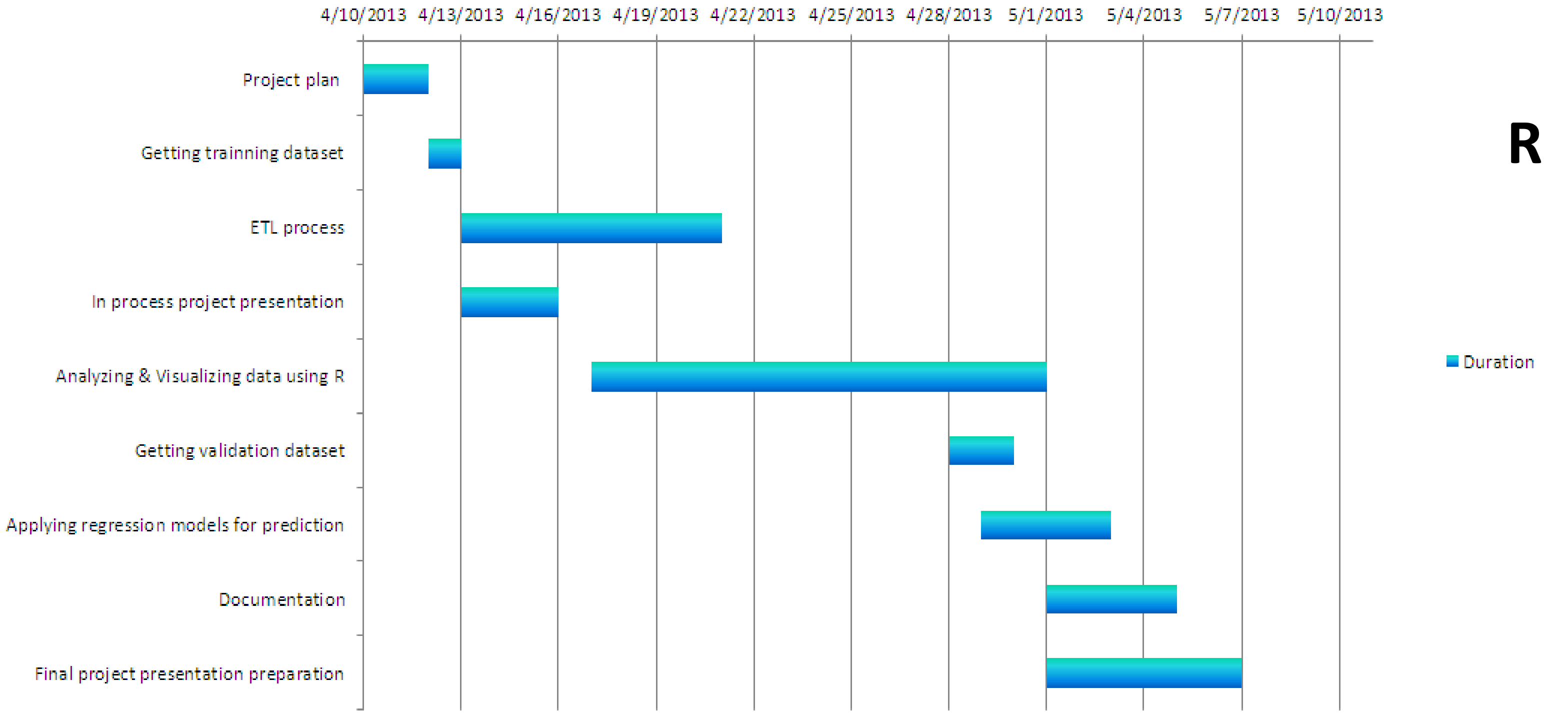
Comparison:

- He did better visualization.
- I separated playoff teams with non-playoff teams.
- He only used average team age to predict 2011 NBA championship.
- I took 4 factors into consideration and applied logistic regression model.

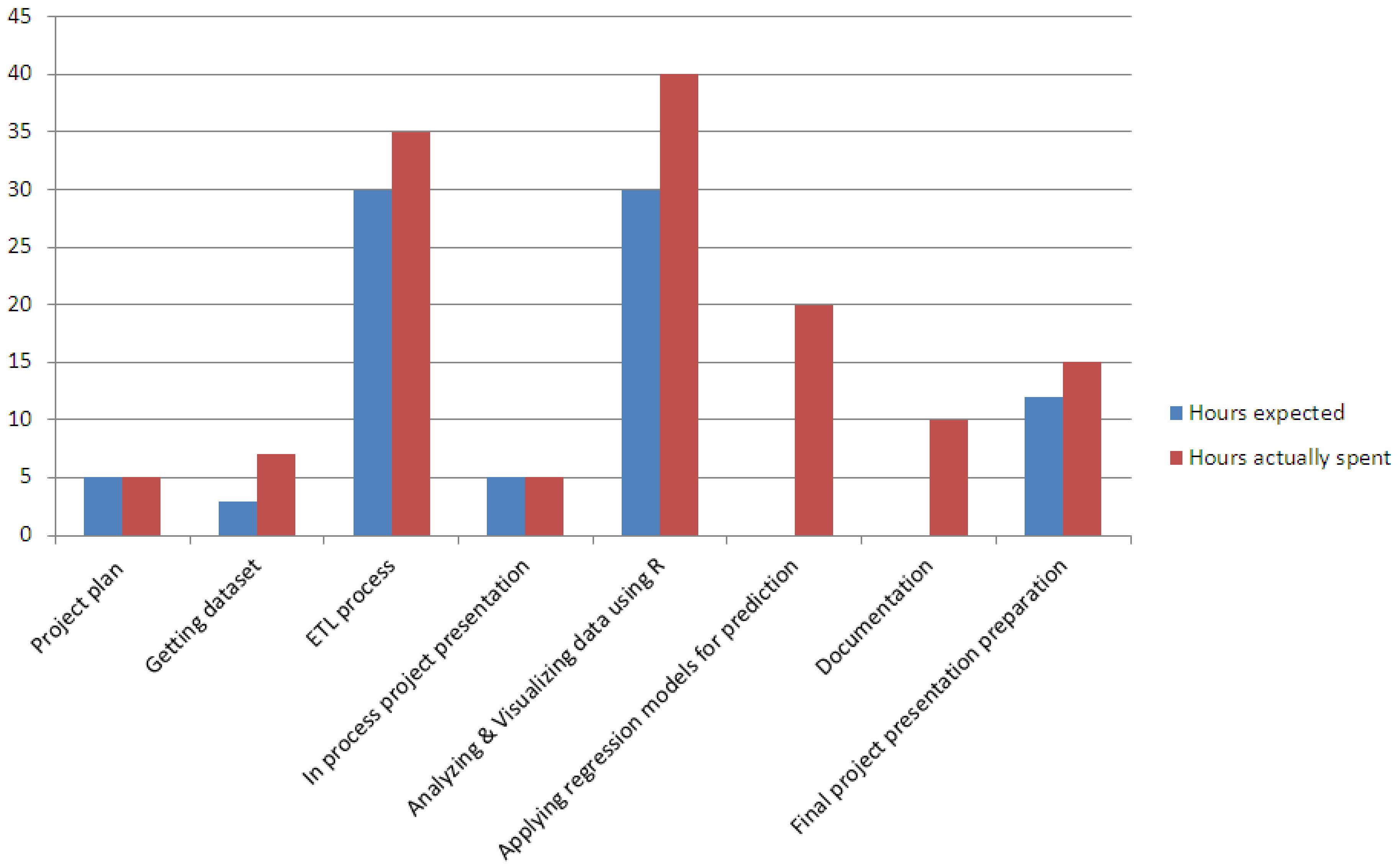
Initial Plan



Real Plan



Hours Spent


45
40
35
30
25
20
15
10
5
0
Project plan
Getting dataset
ETL process
In process project presentation
Analyzing & Visualizing data using R
Applying regression models for prediction
Documentation
Final project presentation preparation
🟦 Hours expected
🟥 Hours actually spent

Scope Change

- Originally, I only wanted to find out the key factors that make a NBA championship.

- As I dived into the data, I found there were more interesting stories to explore, so I changed the project title from "How do teams win NBA championship?" to "Stories behind the NBA stats".

Lesson 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 earlier.
- 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

Thanks !!!

Questions ???