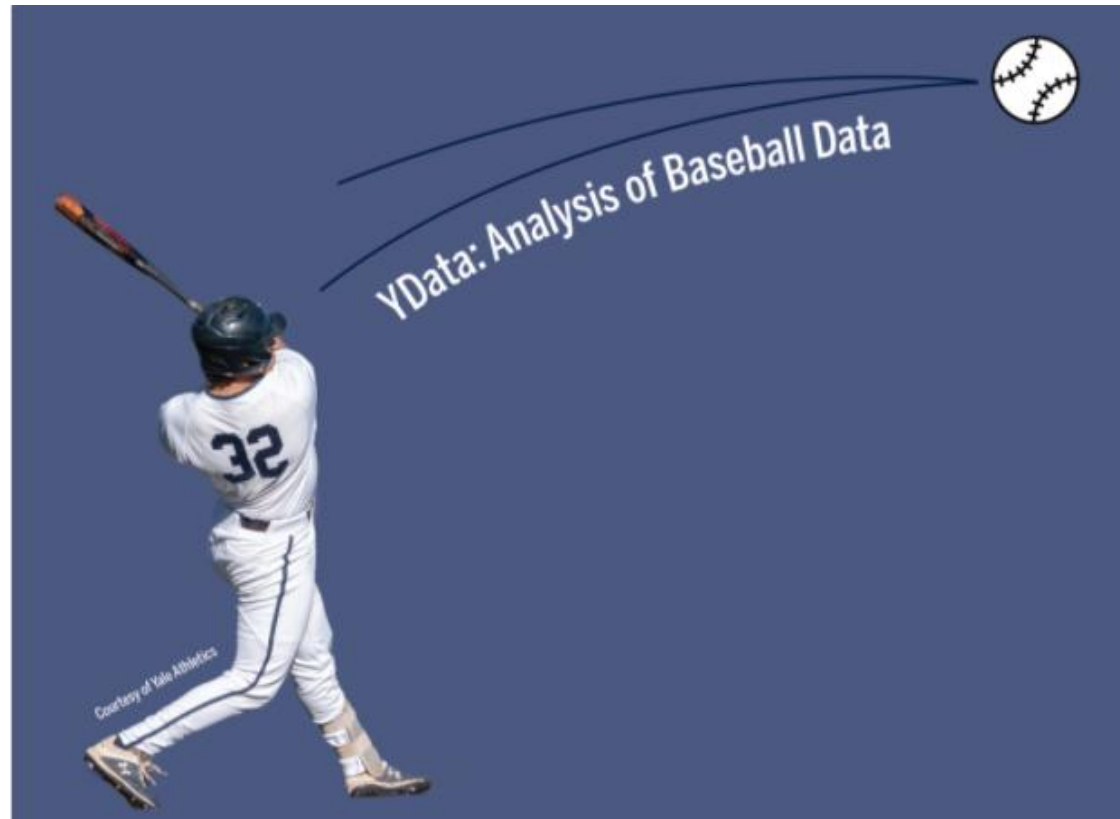


Class project presentations and a few last analyses



Overview

Discussion on Astroball chapter 9 and Ben Reiter's visit

Final project presentations

A few last topics/my final project

Wrap up

Thoughts on Ben Reiter's visit and the end of Astrobball?

Is there anything you found particularly interesting?

Astroball chapter 9

Discussion of game 7 of the world series:

"There was no doubt that the Astros' hitters were talented. On this night, it was almost as if they *knew* what was coming.

In fact they did."

- Astroball page 209

[Correa proposing](#)

[You all know David Ortiz right?](#)



Check out Ben Reiter's [podcast](#)








[David Ortiz selfie scandal rocks White House](#)

The MLB season has finished the first month...

Is it still a fluke that the Red Sox are in first place?

[This play was interesting](#)

American League					National League		
AL East							
Team	W	L	Pct	GB	Home	Away	L10
 Red Sox	18	12	.600	-	9-8	9-4	6-4
 Yankees	15	14	.517	2.5	8-7	7-7	7-3
 Rays	16	15	.516	2.5	7-10	9-5	5-5
 Blue Jays	14	14	.500	3.0	7-4	7-10	6-4
 Orioles	14	16	.467	4.0	4-10	10-6	6-4

Final projects presentations

Presentations should be ~5 minutes with 2-3 minutes of questions.

Take notes when others are presenting to give them suggestions on their projects.

- E.g., additional analyses/directions, things to improve the analyses, etc.

Presentation order?

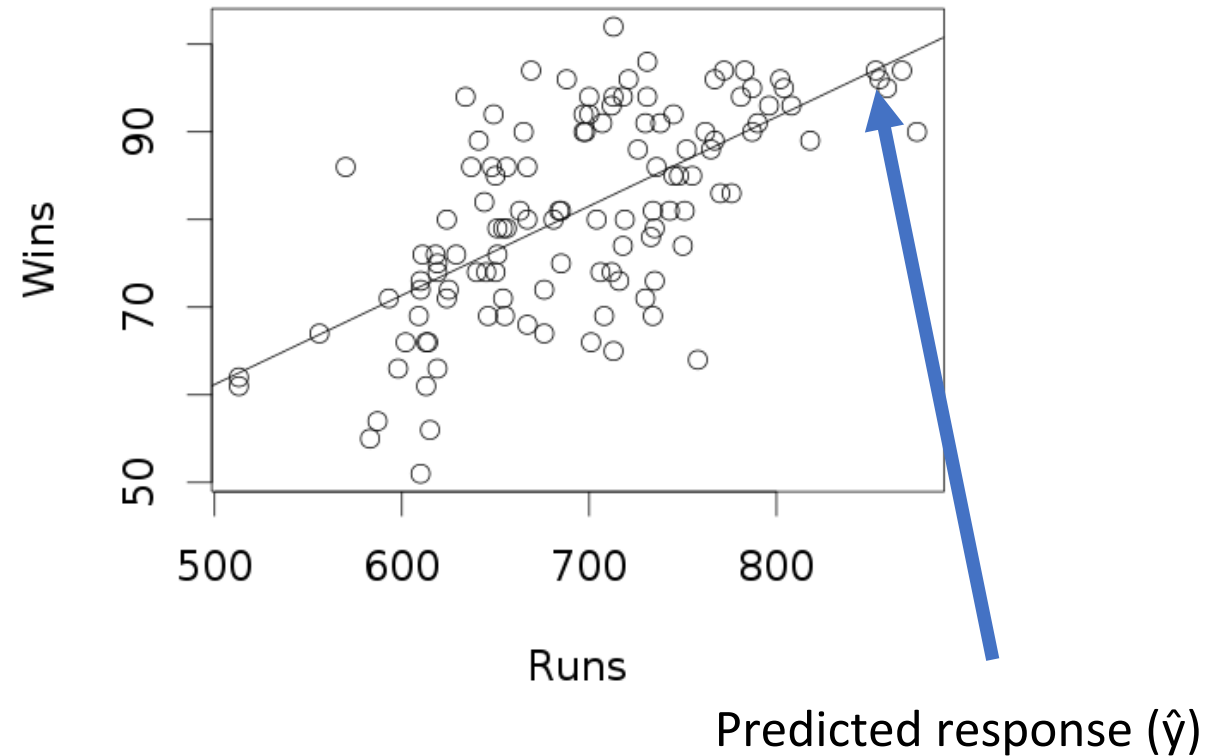
Reminder:

- Written reports are due on Thursday May 13th (last day of reading period)
- Report should be 8-10 pages long

Review regression

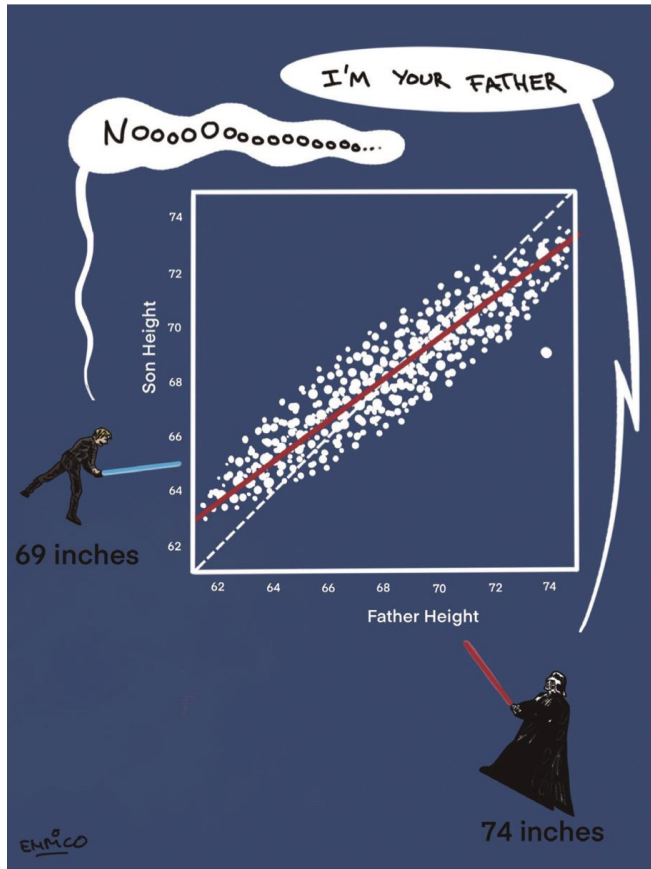
Regression is method of using one variable to predict the value of a second variable

In **linear regression** we fit a line to the data, called the **regression line**

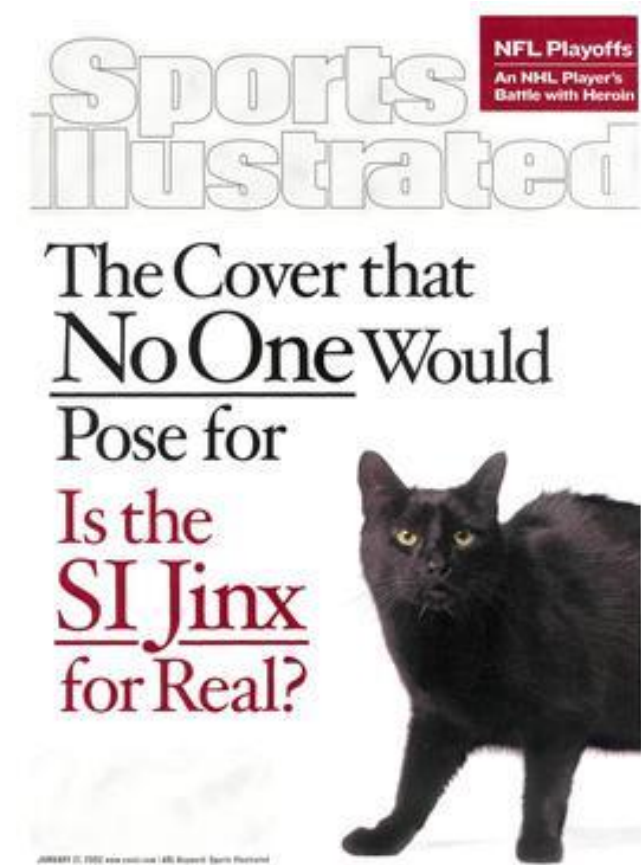


$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \dots + \hat{\beta}_k \cdot x_k$$

Regression to the mean



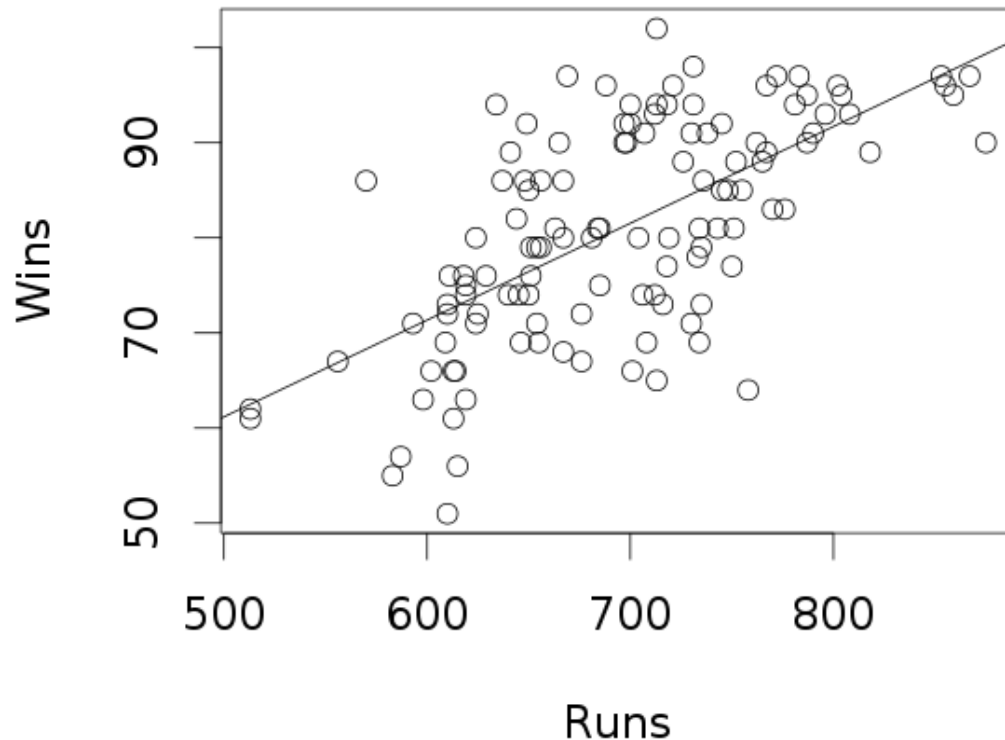
Original data from Galton, 1886



- Sports Illustrated Cover Jinx
- Rookie of the year curse

Bill James' "Pythagorean Expectation"

Recall that our equation for predicting the number of **wins** a team would score as a **function of the number of runs** they produced had some issues...



$$\hat{w} = 14.47 + .088 \cdot \text{Runs}$$

What happens when 0 runs are scored all season?

Bill James' "Pythagorean Method"

Bill James came up with a formula that he called the "Pythagorean Method" that relates:

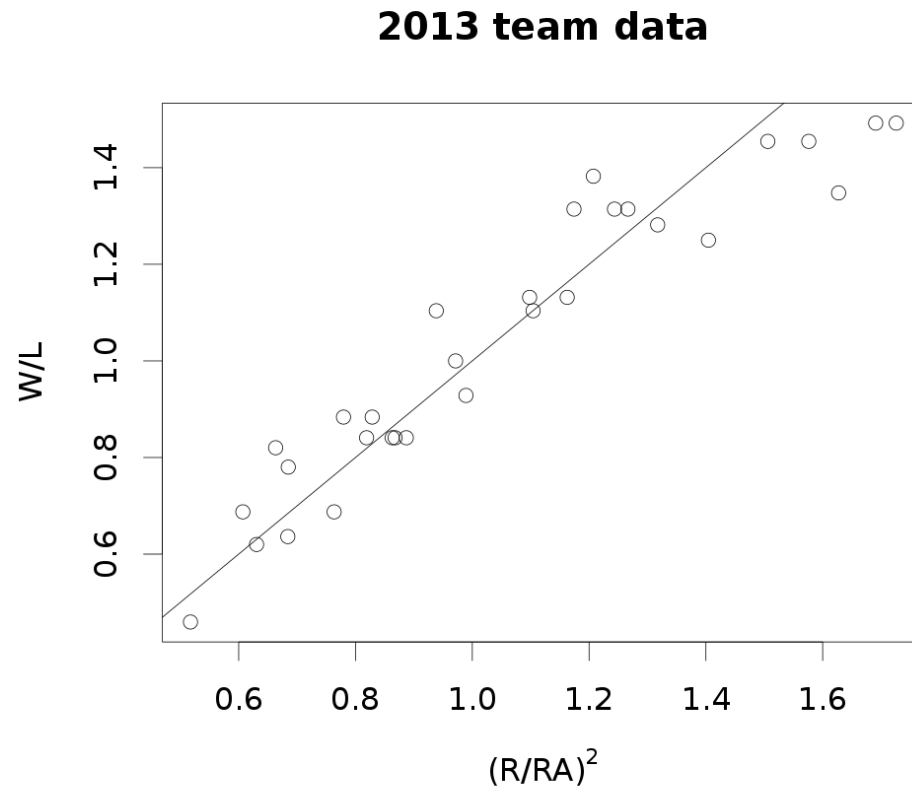
- wins (W) and losses (L) to
- runs scored (R) and runs allowed (RA)

$$\frac{W}{L} = \left(\frac{R}{RA} \right)^2$$

What happens when a team scores 0 runs with this formula?

How can we tell how good this formula is?

An answer: look at a scatter plot of W/L ratio predicted by $(R/RA)^2$ and the actual W/L ratio



How can we tell how good this formula is?

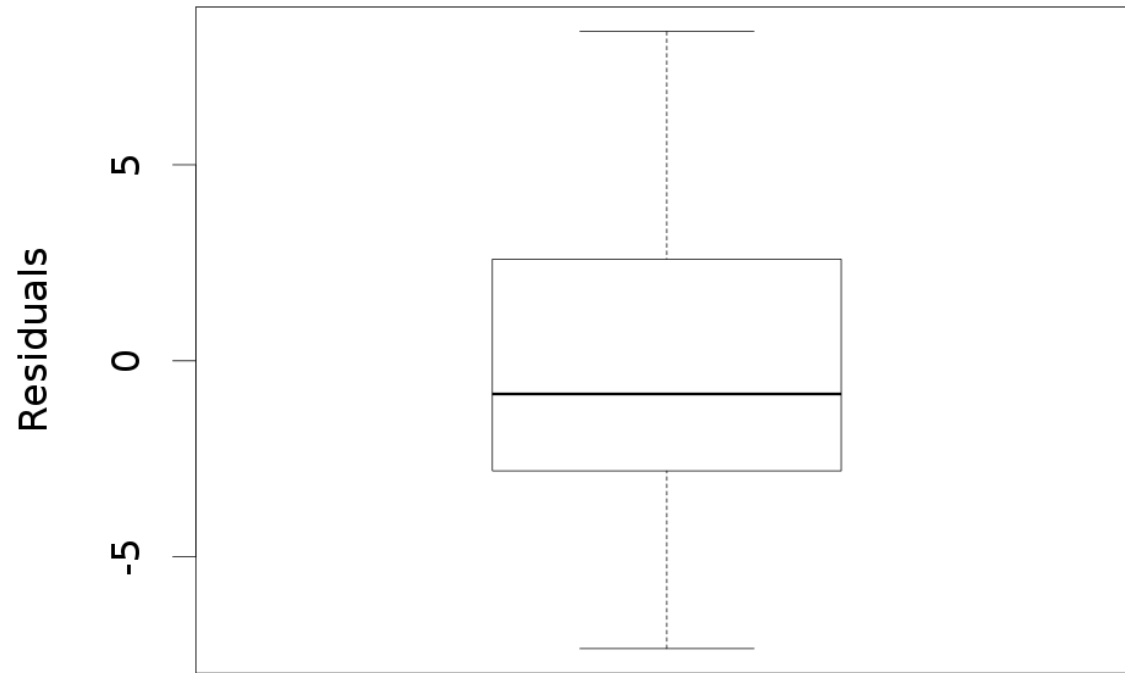
An answer: compare the number of wins predicted by the R and RA values, to the number of wins actually achieved by each team

$$(W/L)_{\text{pred}} = (R/RA)^2$$

$$(W_{\text{pred}} / (162 - W_{\text{pred}})) = (R/RA)^2 \quad \dots \text{ some algebra } \dots$$

$$W_{\text{pred}} = (162 \cdot (R/RA)^2) / (1 + (R/RA)^2)$$

How can we tell how good this formula is?



RMSE = 3.9

Five number summary of the residuals:
(-7.34, -2.81, -0.85, 2.59, 8.30)

Can we do better the James' formula?

As you saw on homework 10, we can potentially improve on James' formula by finding a better exponent on R/RA rather than just assuming it is 2

$$\frac{W}{L} = \left(\frac{R}{RA} \right)^2$$

$$\frac{W}{L} = \left(\frac{R}{RA} \right)^k$$

How can we do this?

Can we do better the James' formula?

If we take the logarithm of James' formula, it becomes a linear equation

$$\log\left(\frac{W}{L}\right) = 2 \cdot \log\left(\frac{R}{RA}\right)$$

$$\log\left(\frac{W}{L}\right) = k \cdot \log\left(\frac{R}{RA}\right)$$

Since this equation is linear we can find k with linear regression!

Can we do better the James' formula?

On homework 10 you found a better exponent k using Python

You also assess whether the new k was better in terms of the ability of the ability to make predictions on new data by:

1. Finding the optimal k using data from 2000 to 2009
2. Making predictions using this optimal k using data from 2010 to 2018

Q: What is the most important question of the semester?




















A: How are the Red Sox going to do this year?

FiveThirtyEight

FiveThirtyEight gives the Red Sox:

- 31% chance of making the playoffs
- 2% chance of winning the World Series
- Predict they will win 83 games

Let's make our own predictions based on their start of the season performance

TEAM ↕	DIVISION ↕	TEAM RATING ↕	1-WEEK CHANGE ↕	AVG. SIMULATED SEASON		POSTSEASON CHANCES		
				RECORD ↕	RUN DIFF. ↕	MAKE PLAYOFFS ↕	WIN DIVISION ↕	WIN WORLD SERIES ↕
 Dodgers 17-12	NL West	1599	-1	102-60	+220	96%	75%	28%
 Yankees 14-14	AL East	1563	+4	92-70	+115	74%	49%	12%
 Padres 17-13	NL West	1559		93-69	+107	78%	22%	8%
 Astros 15-13	AL West	1550	+3	91-71	+116	71%	54%	9%
 Mets 11-12	NL East	1531	-3	85-77	+26	45%	34%	4%
 Rays 15-15	AL East	1527	-1	86-76	+33	43%	18%	3%
 Twins 11-16	AL Central	1526	+3	84-78	+51	38%	25%	3%
 Blue Jays 14-13	AL East	1525	+4	86-76	+59	44%	19%	3%
 Braves 12-16	NL East	1523	-4	82-80	+13	33%	23%	3%
 White Sox 15-12	AL Central	1522	-1	87-75	+66	52%	37%	4%
 Brewers 17-12	NL Central	1520	-1	88-74	+38	60%	44%	4%
 Nationals 12-12	NL East	1518	+5	83-79	+5	38%	27%	2%
 Athletics 18-12	AL West	1516	-3	87-75	+18	50%	29%	3%
 Indians 14-13	AL Central	1513	+2	84-78	+28	41%	26%	3%
 Cardinals 17-12	NL Central	1511	+7	86-76	+38	46%	31%	2%
 Red Sox 17-12	AL East	1510		83-79	+15	31%	13%	2%
 Angels 13-14	AL West	1503	-4	80-82	-23	22%	10%	1%

Confidence intervals on Red Sox winning ability

The Red Sox record is 18-12

We can create a confidence interval for the Red Sox true winning percentage

- $CI_{95} = \text{statistic} \pm 2 \cdot SE$
- $CI_{95} = \hat{p} \pm 2 \cdot SE$

We could estimate the SE using the bootstrap, but instead we will use a parametric formula:

$$\hat{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Could you create a confidence interval for the Red Sox true winning percentage?

Confidence intervals on Red Sox winning ability

The Red Sox record is 18-12

- $\hat{p} = 18/30 = .600$
- $CI_{95} = \hat{p} \pm 2 \cdot SE$

$$\hat{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Our SE estimate is: 0.0894

Our confidence interval is: [0.421, 0.779]

Question: how can we interpret this confidence interval?

Predicting final winning percentage based on current number of wins and losses

As mentioned, the Red Sox record currently is:

- Wins = 18
- Losses = 12

Any ideas how we could use this to predict Red Sox final winning percentage at the end of the season?

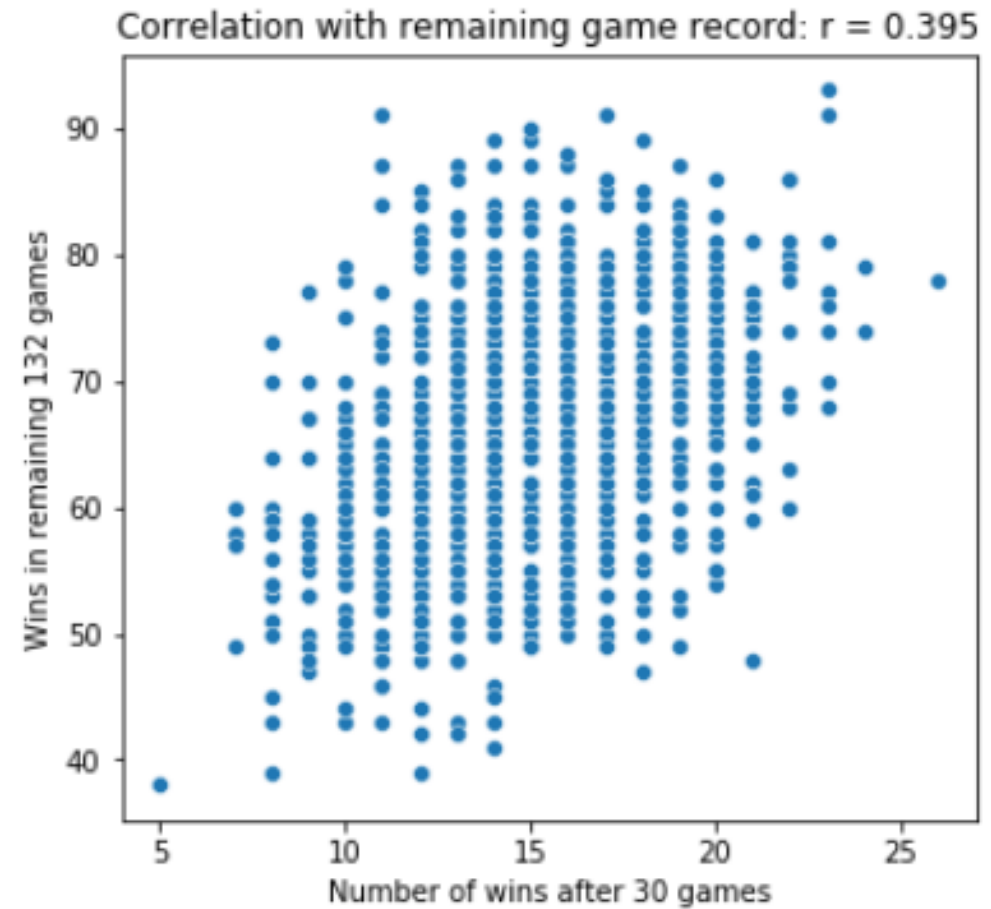
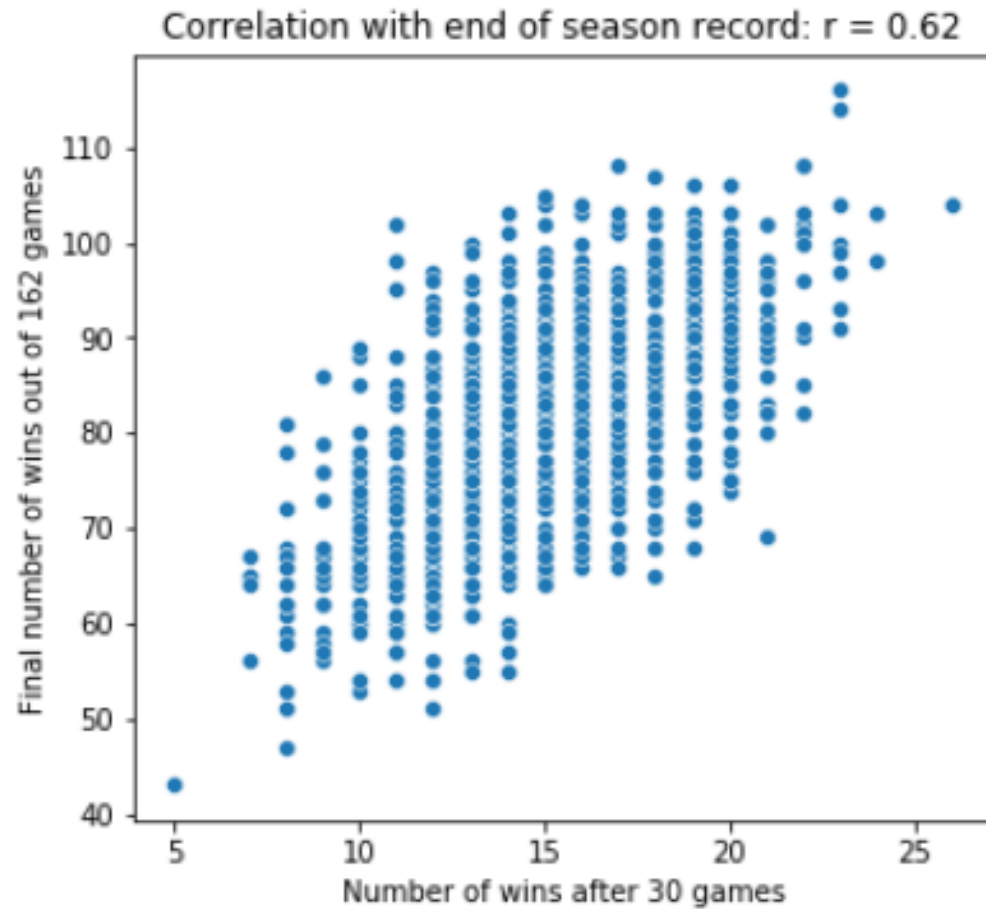
Correlation between current and final number of wins

The Red Sox have played 30 games

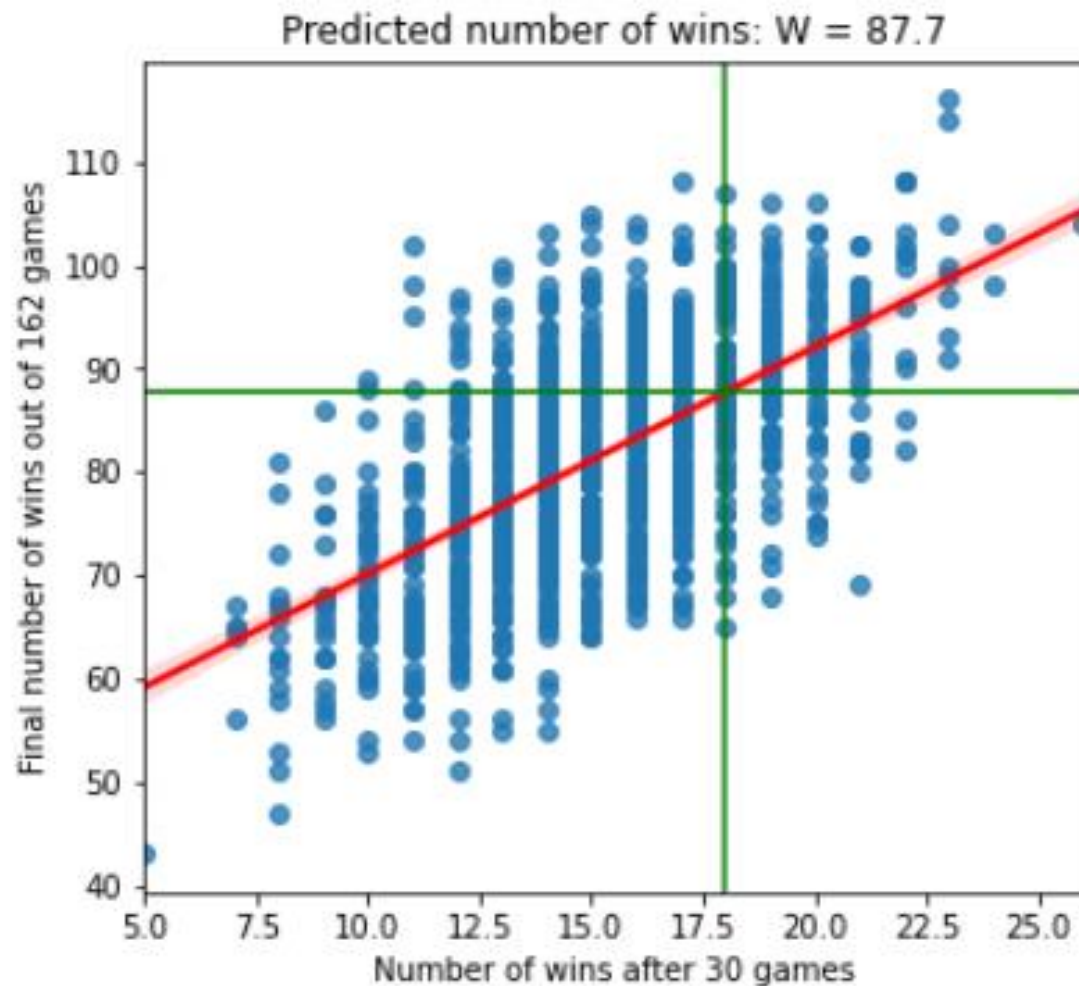
What is the final winning percentage for teams that have played this many games?

- Using data since 1970 for teams that had 162 game season

Correlation between current and final number of wins

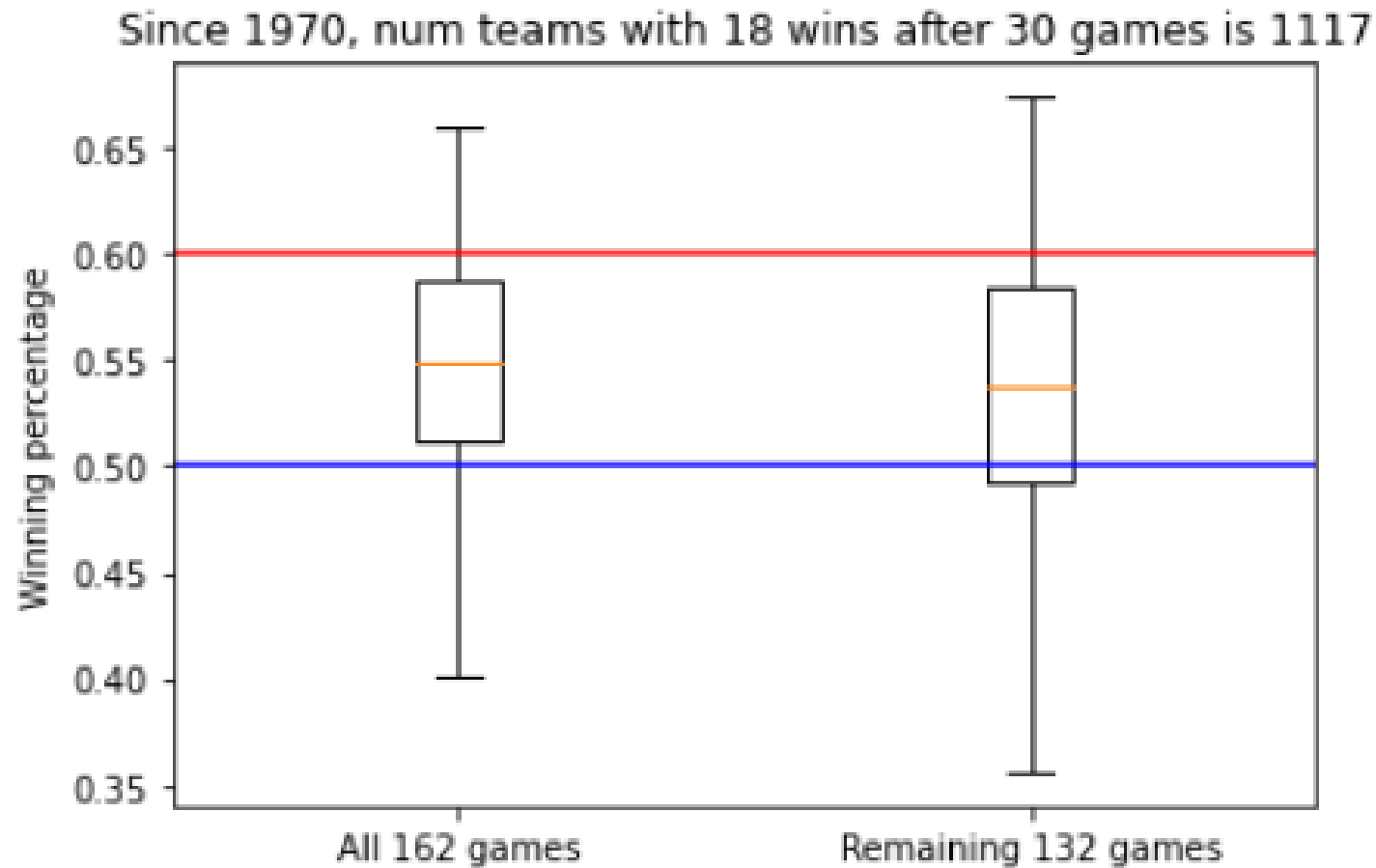


Using regression to predict final number of wins based on current number of wins

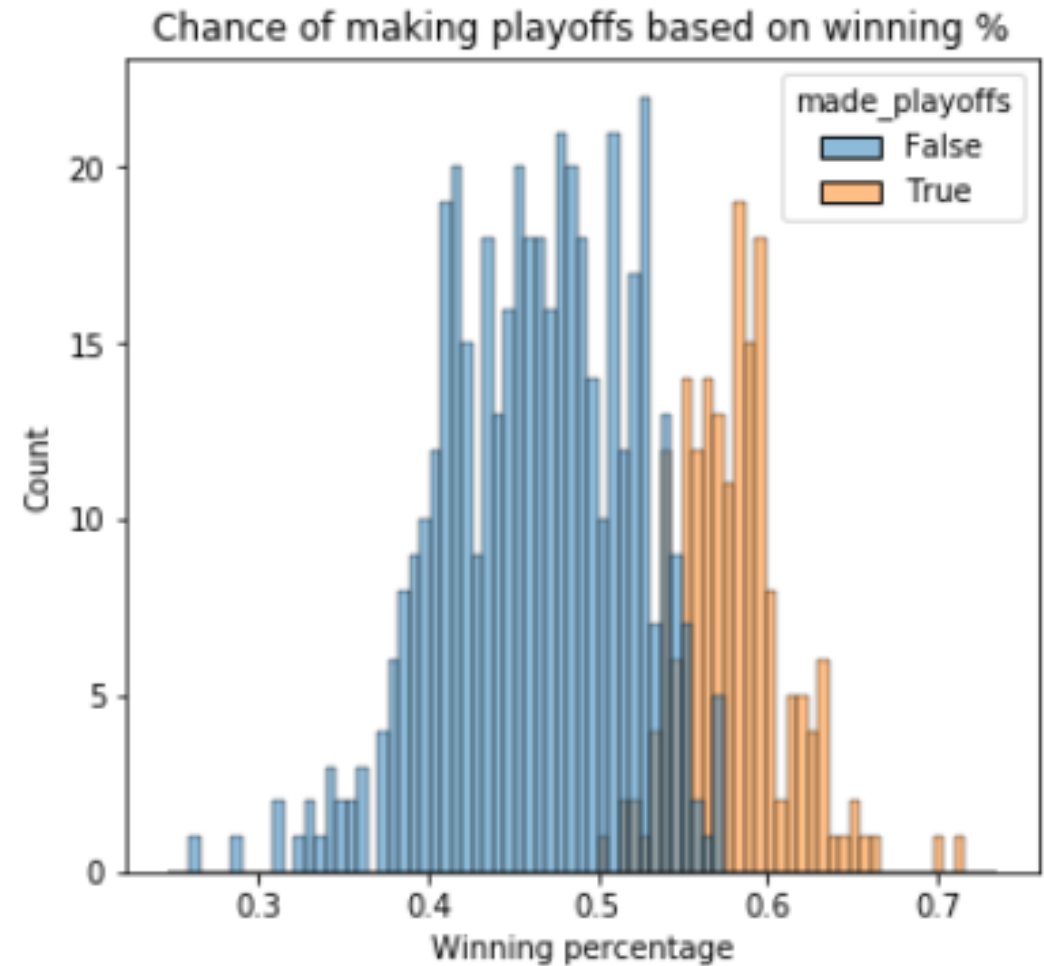
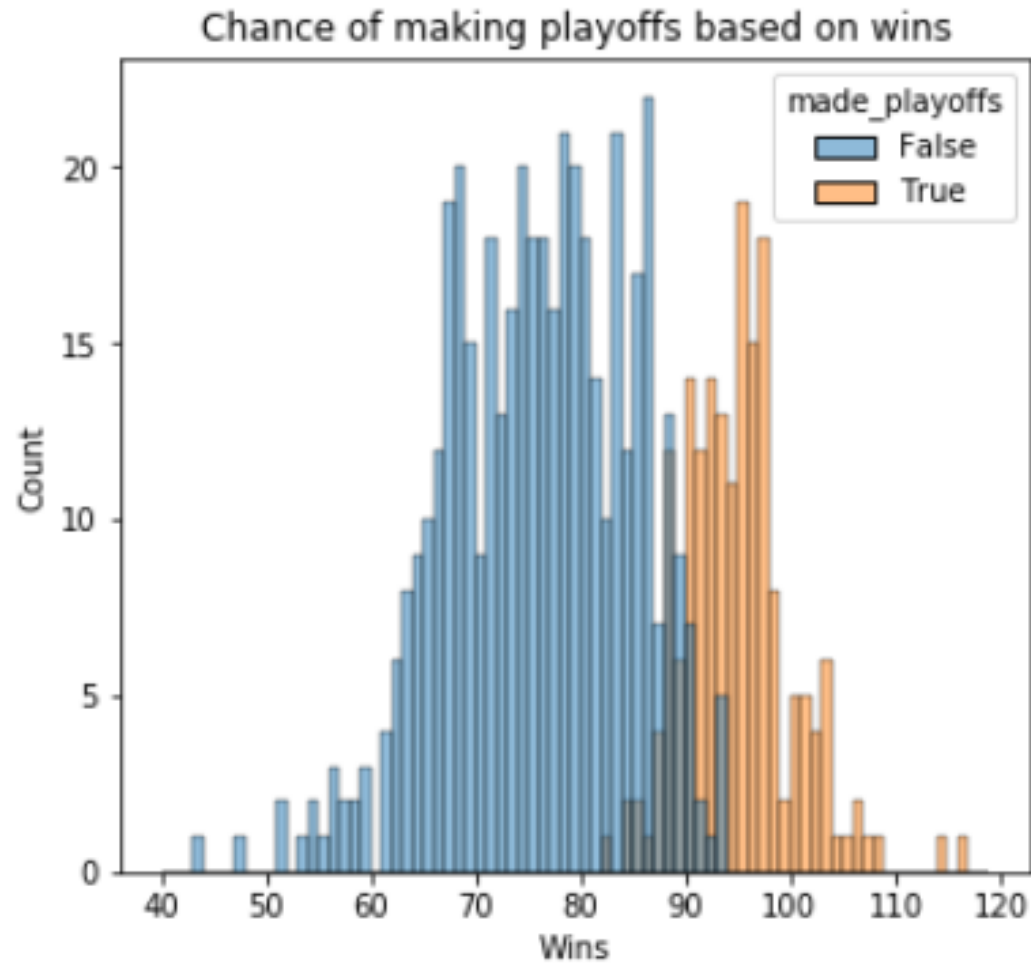


Predicted number of wins is 87.7

Looking at distribution of final winning percentage based on current number of wins

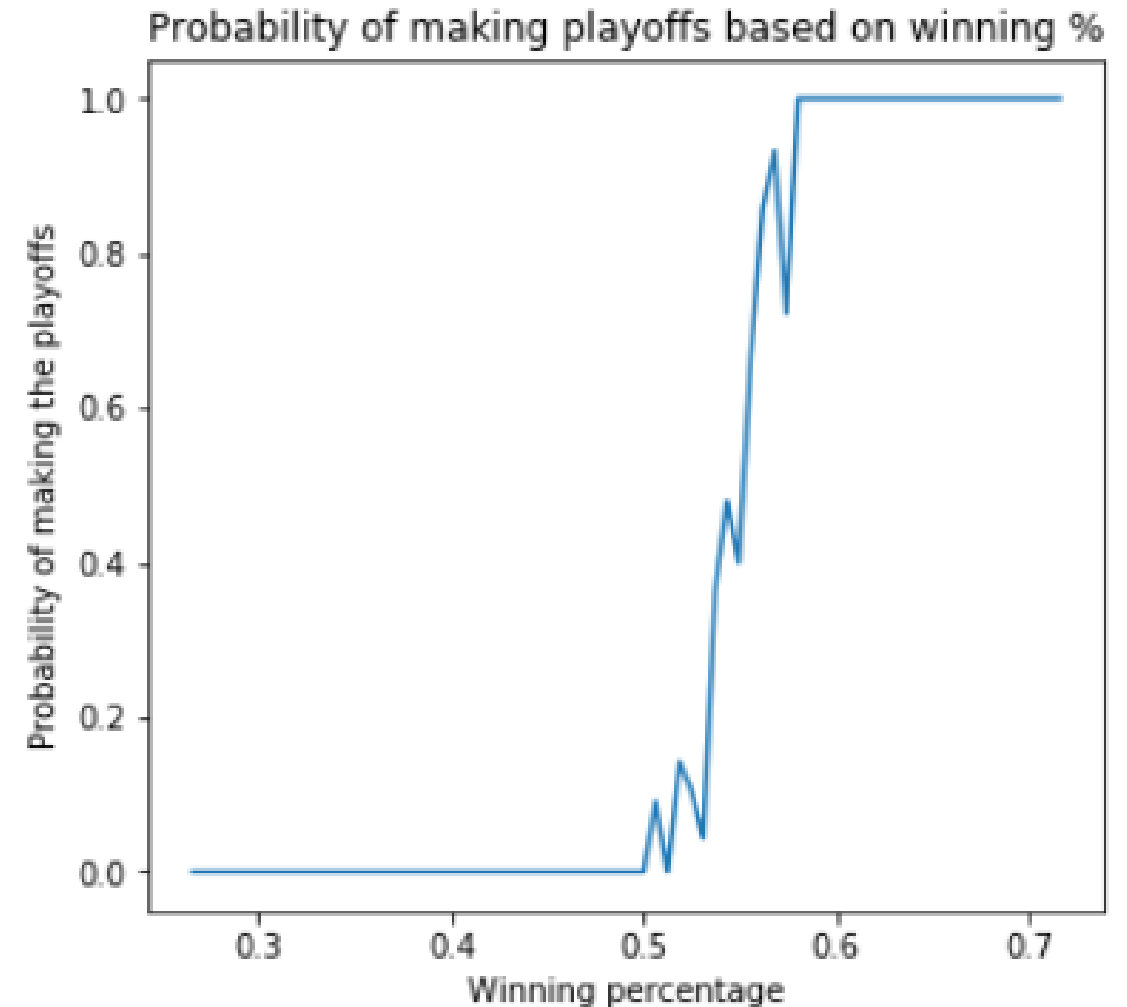
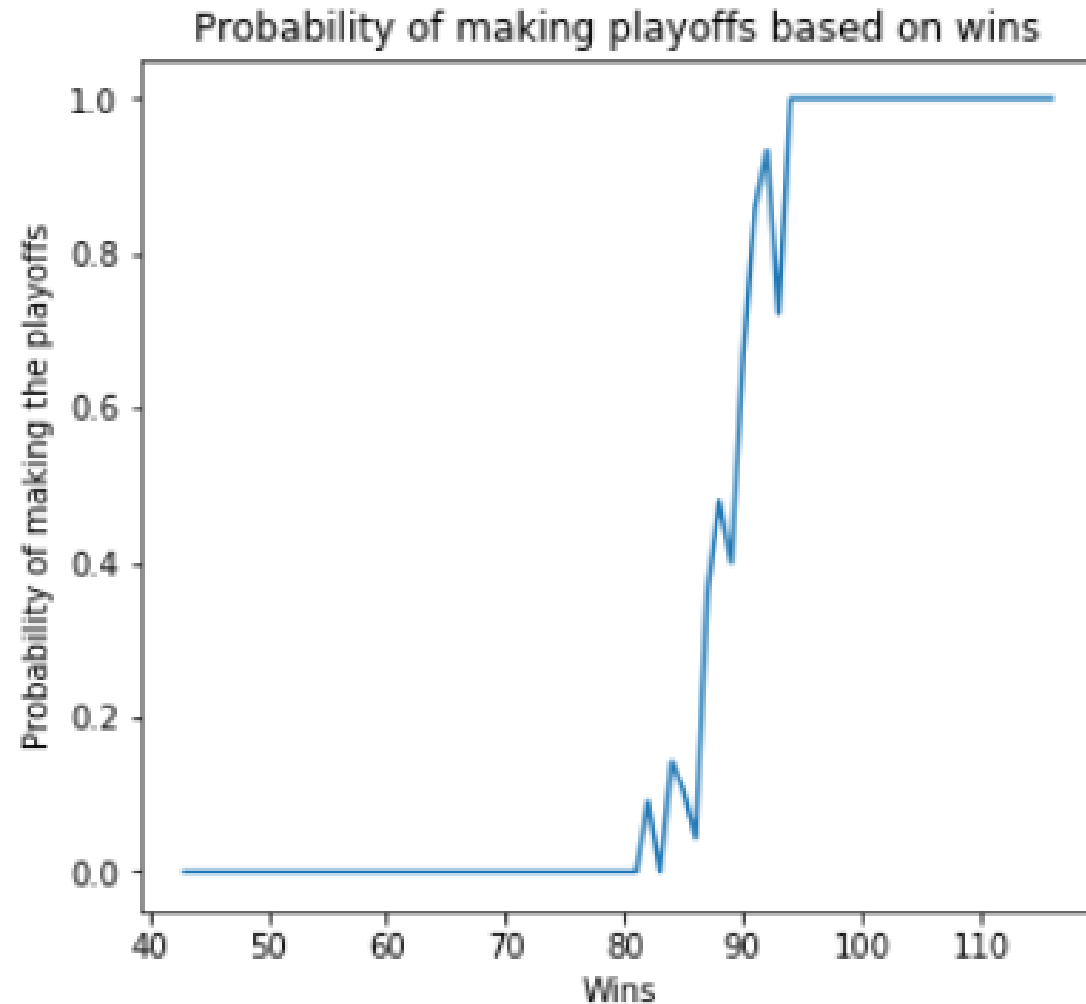


What winning percentage is needed to make the playoffs?

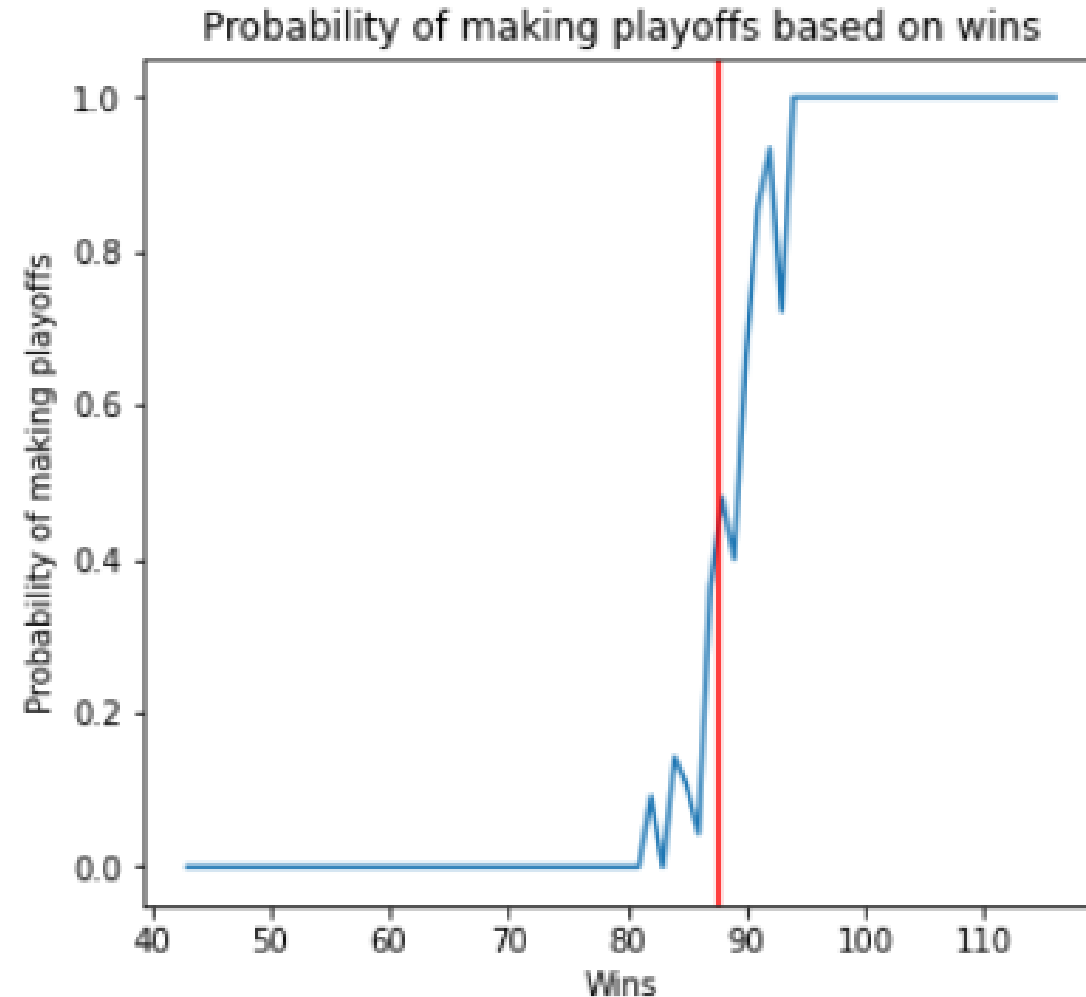


Using only data since the wildcard was introduced (since 1994)

What winning percentage is needed to make the playoffs?



Red Sox probability of making the playoffs based on their current record



Based on the fact that the predicted number of wins is 87.7, the Red Sox have a 48% of making the playoffs

Predicting Red Sox playoff potential based on the number of runs scored and runs allowed

Wins and losses are a pretty crude measure of performance

Perhaps it would be better to predict winning percentage based on the number of runs scored and runs allowed

We can use Bill James' Pythagorean formula to do this!

Applying the Pythagorean formula

The Red Sox have scored 149 runs

The Red Sox have allowed 125 runs

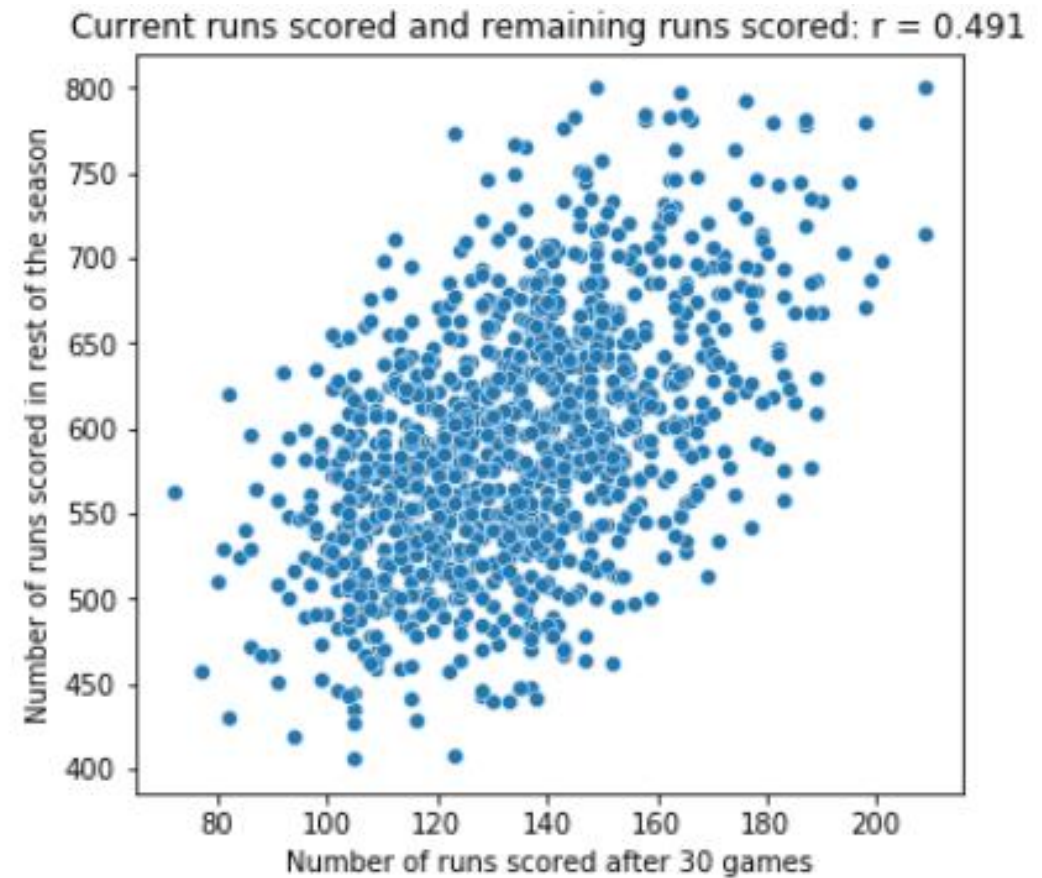
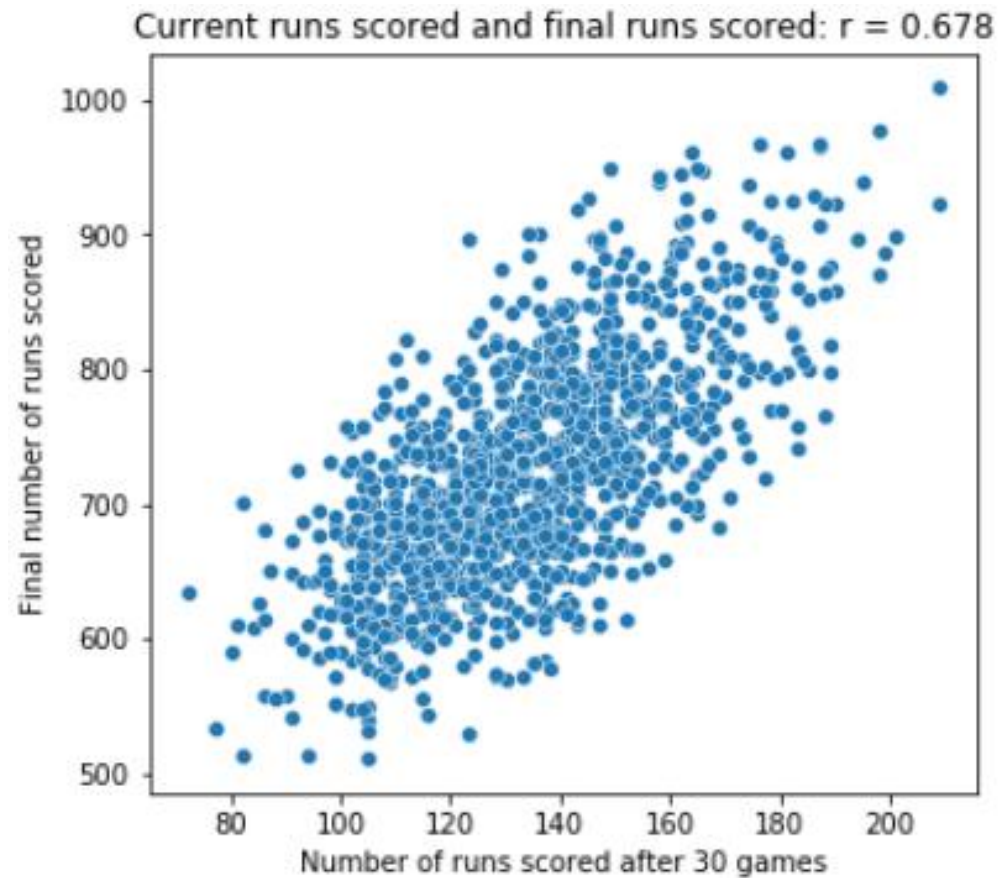
Bill James' Pythagorean formula is: $\frac{W}{L} = \left(\frac{R}{RA} \right)^2$

Plugging into Bill James' Pythagorean we get:

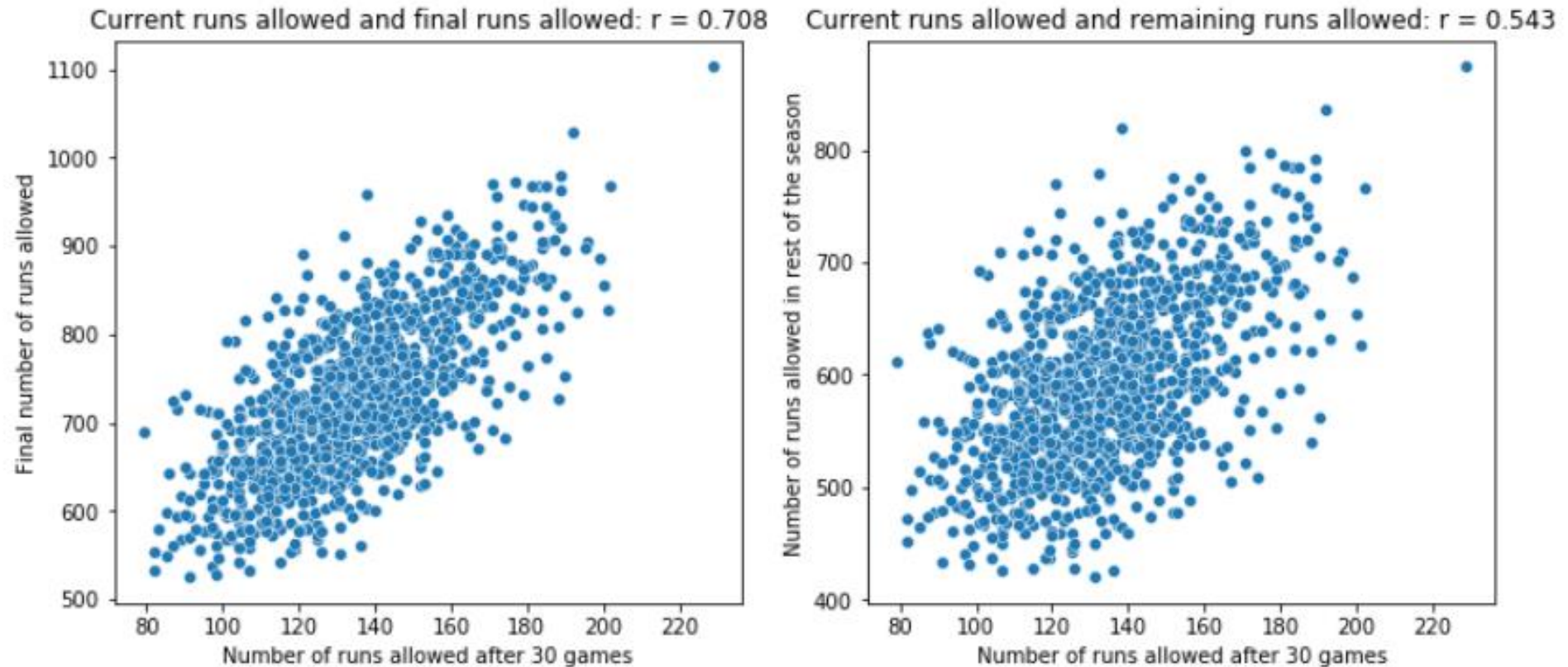
- 95.08 wins
- .587 winning percentage

Question: Is this a realistic estimate of the Red Sox final number of wins?

Final number of runs scored based on runs scored after 30 games



Final number of runs allowed based on runs allowed after 30 games



Applying the Pythagorean formula

The Red Sox are predicted to have scored 764.64 runs

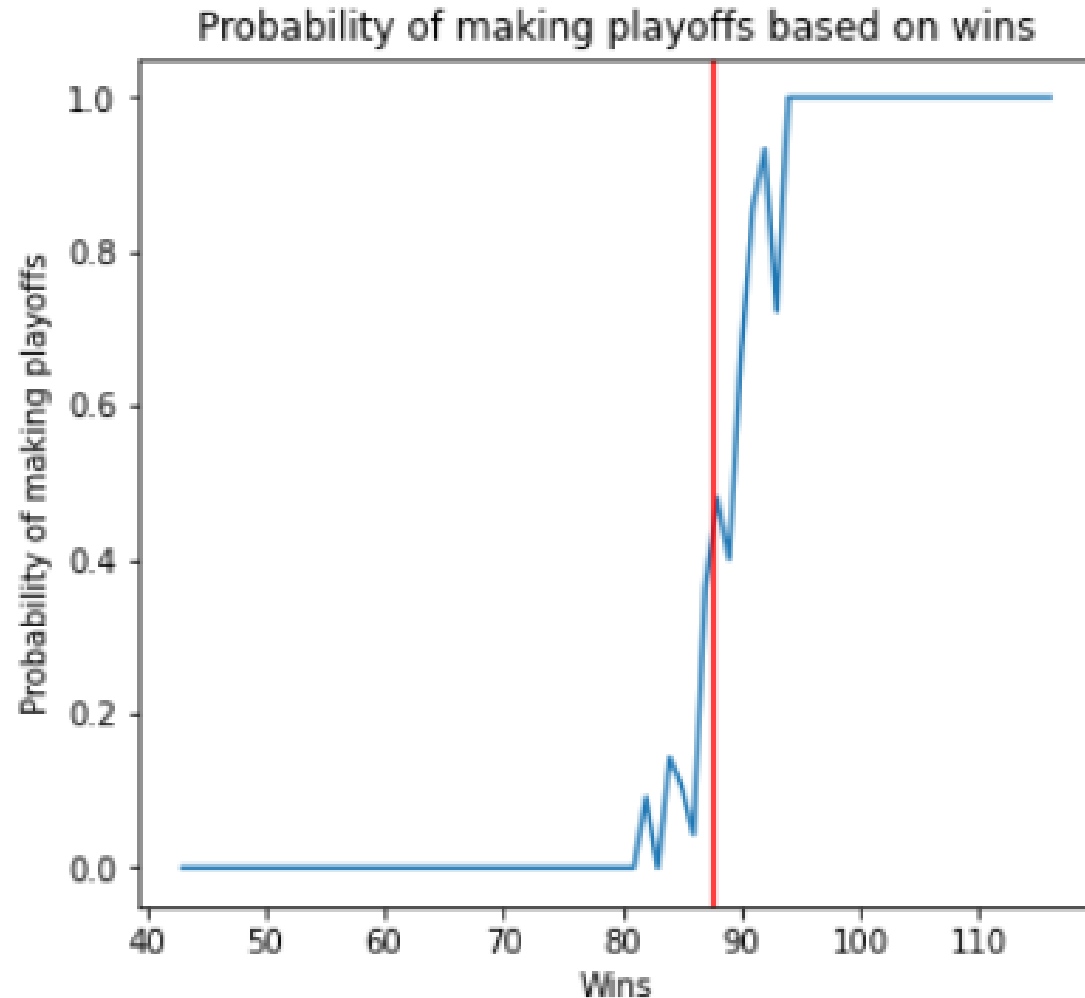
The Red Sox are predicted to have allowed 699.44 runs

Bill James' Pythagorean formula is: $\frac{W}{L} = \left(\frac{R}{RA} \right)^2$

Plugging into Bill James' Pythagorean formula we get:

- 88.2 Wins
- 0.544 winning percentage

Applying the Pythagorean formula



Based on the fact that the predicted number of wins is 88.2, the Red Sox have a 48% of making the playoffs

Does predicting final records from early win/loss records or runs scored/allowed seem reasonable?

What are strengths/weakness of this approach?


















What are ways we could improve the accuracy of the predictions?

FiveThirtyEight

FiveThirtyEight gives the Red Sox:

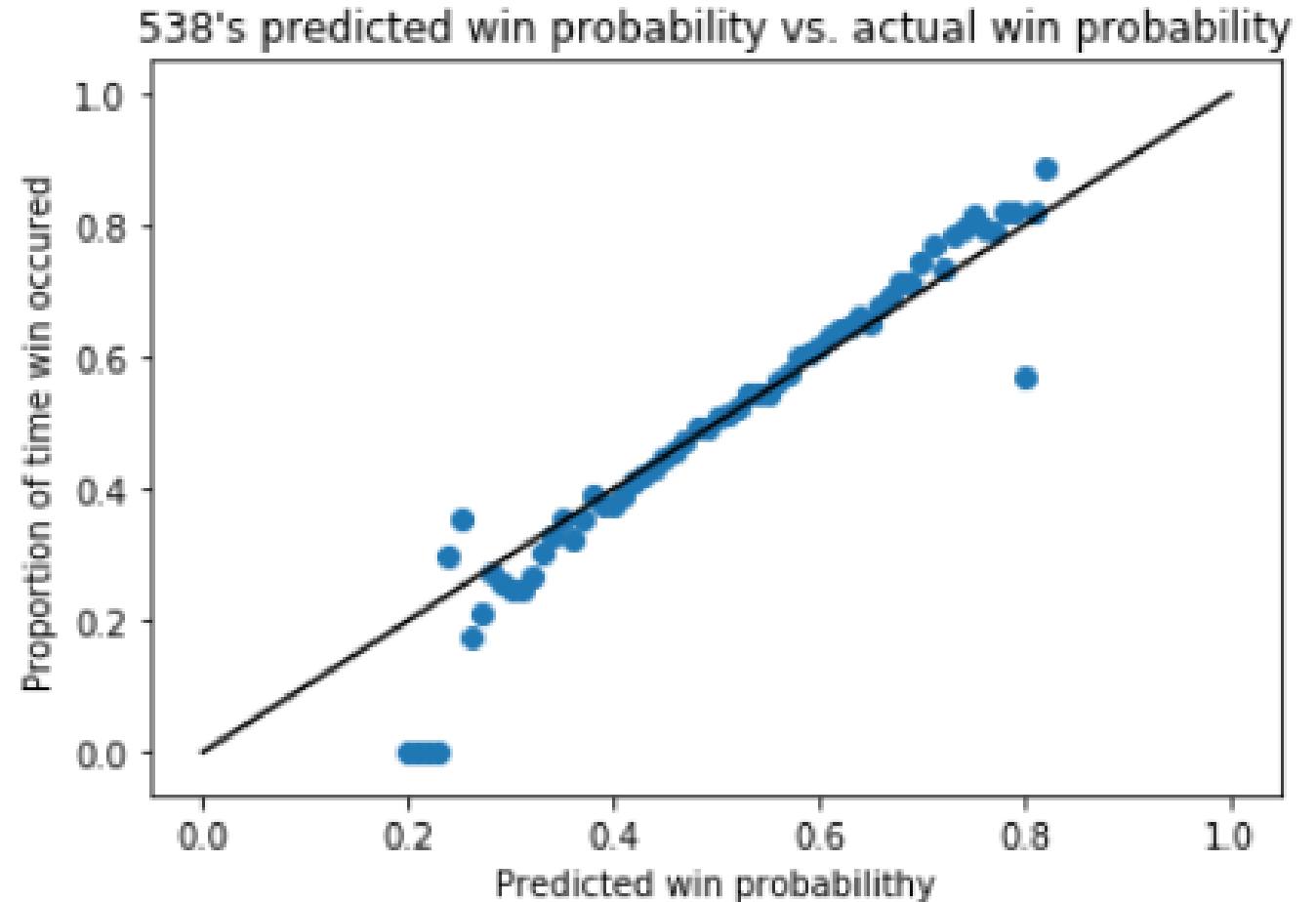
- 31% chance of making the playoffs
- 2% chance of winning the World Series
- Predict they will win 83 games

Whose predictions should we trust?

TEAM ↕	DIVISION ↕	TEAM RATING ↕	1-WEEK CHANGE ↕	AVG. SIMULATED SEASON		POSTSEASON CHANCES		
				RECORD ↕	RUN DIFF. ↕	MAKE PLAYOFFS ↕	WIN DIVISION ↕	WIN WORLD SERIES ↕
 Dodgers 17-12	NL West	1599	-1	102-60	+220	96%	75%	28%
 Yankees 14-14	AL East	1563	+4	92-70	+115	74%	49%	12%
 Padres 17-13	NL West	1559		93-69	+107	78%	22%	8%
 Astros 15-13	AL West	1550	+3	91-71	+116	71%	54%	9%
 Mets 11-12	NL East	1531	-3	85-77	+26	45%	34%	4%
 Rays 15-15	AL East	1527	-1	86-76	+33	43%	18%	3%
 Twins 11-16	AL Central	1526	+3	84-78	+51	38%	25%	3%
 Blue Jays 14-13	AL East	1525	+4	86-76	+59	44%	19%	3%
 Braves 12-16	NL East	1523	-4	82-80	+13	33%	23%	3%
 White Sox 15-12	AL Central	1522	-1	87-75	+66	52%	37%	4%
 Brewers 17-12	NL Central	1520	-1	88-74	+38	60%	44%	4%
 Nationals 12-12	NL East	1518	+5	83-79	+5	38%	27%	2%
 Athletics 18-12	AL West	1516	-3	87-75	+18	50%	29%	3%
 Indians 14-13	AL Central	1513	+2	84-78	+28	41%	26%	3%
 Cardinals 17-12	NL Central	1511	+7	86-76	+38	46%	31%	2%
 Red Sox 17-12	AL East	1510		83-79	+15	31%	13%	2%
 Angels 13-14	AL West	1503	-4	80-82	-23	22%	10%	1%

How good are FiveThirtyEight's predictions?

Individual game
winning probability
vs. how frequently
the team won



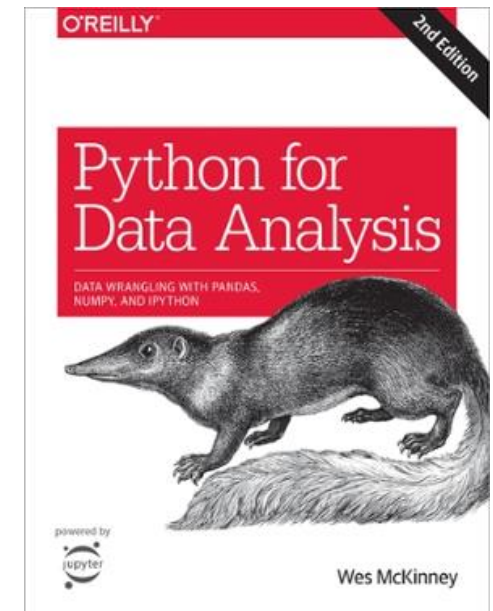
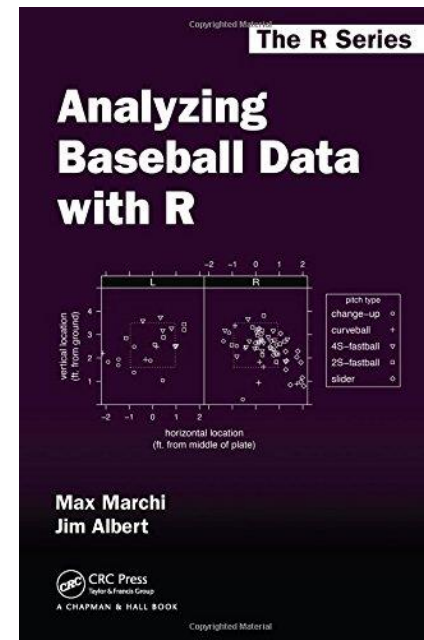
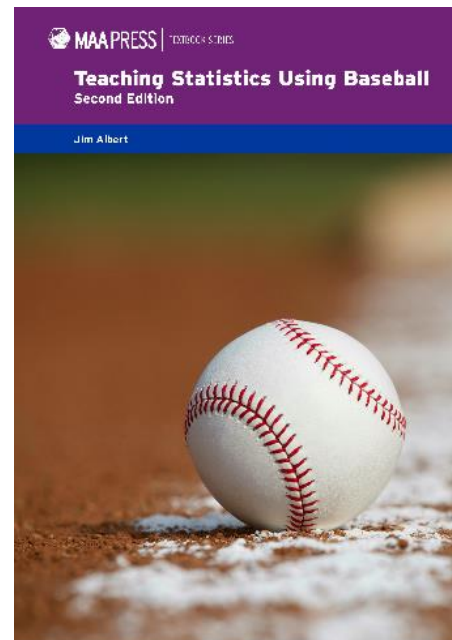
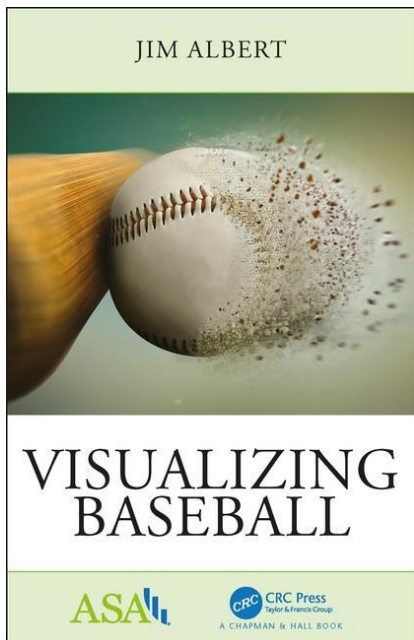
Future directions

If you want to continue to learn more about using Python for Data Science I recommend you read more about the following packages:



Future directions: other relevant books

- Visualizing Baseball
- Teaching Statistics Using Baseball
- Analyzing Baseball Data with R
- Python for Data Analysis



Have a good summer!

