DIMENSIONLESS
TECHNOLOGY

“Moneyball-The Power of Sport Analytics”

Baseball Rules

MONEYBALL

The Power of Sports Analytics

Moneyball, a book by Michael Lewis in 2003 and a movie in 2011 starring Brad Pitt.

- Moneyball discusses how sports analytics changed baseball.

- Moneyball tells the story of the Oakland A's.

- They were once a rich team, but the team was purchased in 1995 by owners who enforced strict budget cuts.

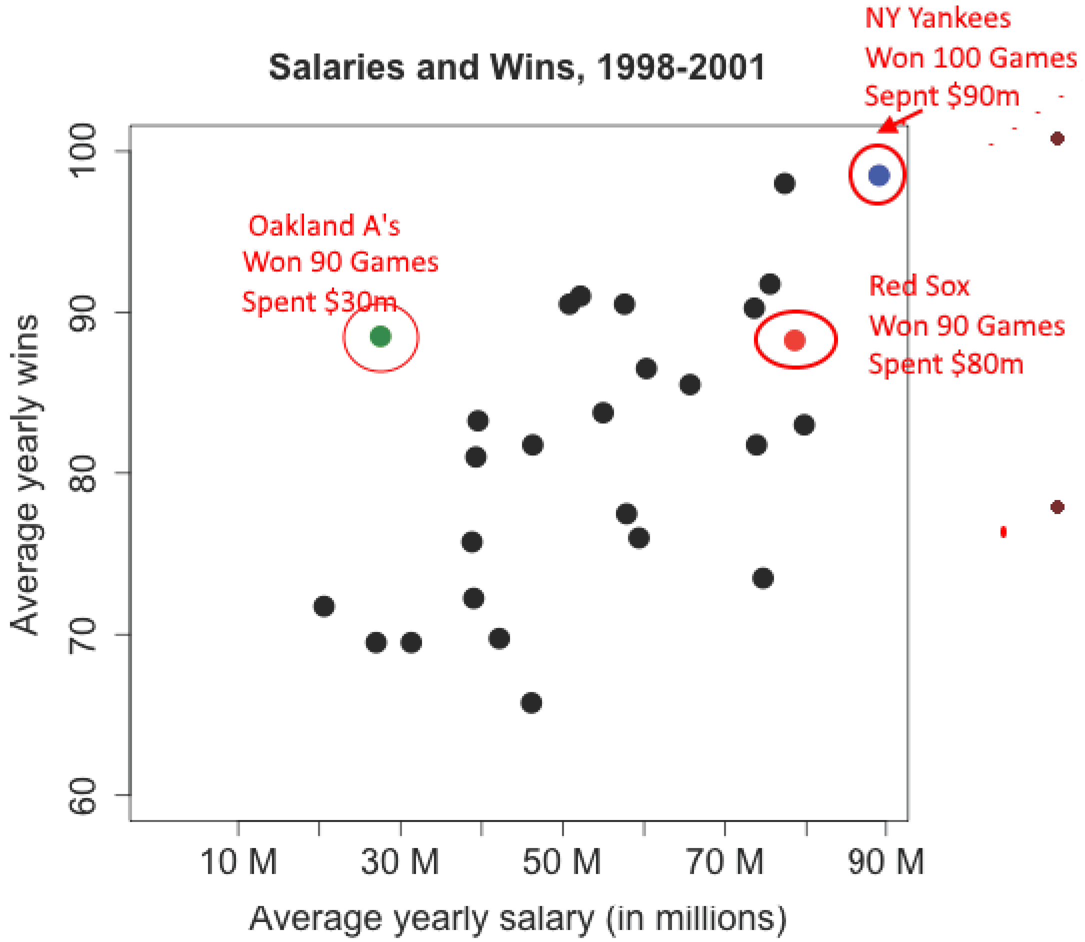
The Story

- *Moneyball* tells the story of the Oakland A's in 2002
  - One of the poorest teams in baseball
    - New ownership and budget cuts in 1995
  - But they were improving

Year	Win %
1997	40%
1998	46%
1999	54%
2000	57%
2001	63%

- How were they doing it?
  - Was it just luck?
- In 2002, the A's lost three key players
- Could they continue winning?

The Problem



- Rich teams can afford the all-star players
- How do the poor teams compete?

Competing as a poor team

- Competitive imbalances in the game
  - Rich teams have four times the salary of poor teams
- The Oakland A's can't afford the all-stars, but they are still making it to the playoffs. How?
- They take a quantitative approach and find undervalued players

A Different Approach

- The A's started using a different method to select players
- The traditional way was through scouting
  - Scouts would go watch high school and college players
  - Report back about their skills
  - A lot of talk about speed and athletic build
- The A's selected players based on their statistics, not on their looks

  "The statistics enabled you to find your way past all sorts of sight-based scouting prejudices."

  "We're not selling jeans here"

The Perfect Batter

The A's

A catcher who couldn't throw
Gets on base a lot

The Yankees

A consistent shortshop
Leader in hits and stolen bases

The Perfect Pitcher

The A's

Unconventional delivery
Slow speed

The Yankees

Conventional delivery
Fast speed

Billy Beane

- The general manager since 1997
- Played major league baseball, but never made it big
  - Sees himself as a typical scouting error
- Billy Beane succeeded in using analytics
  - Had a management position
  - Understood the importance of statistics – hired Paul DePodesta (a Harvard graduate) as his assistant
  - Didn't care about being ostracized

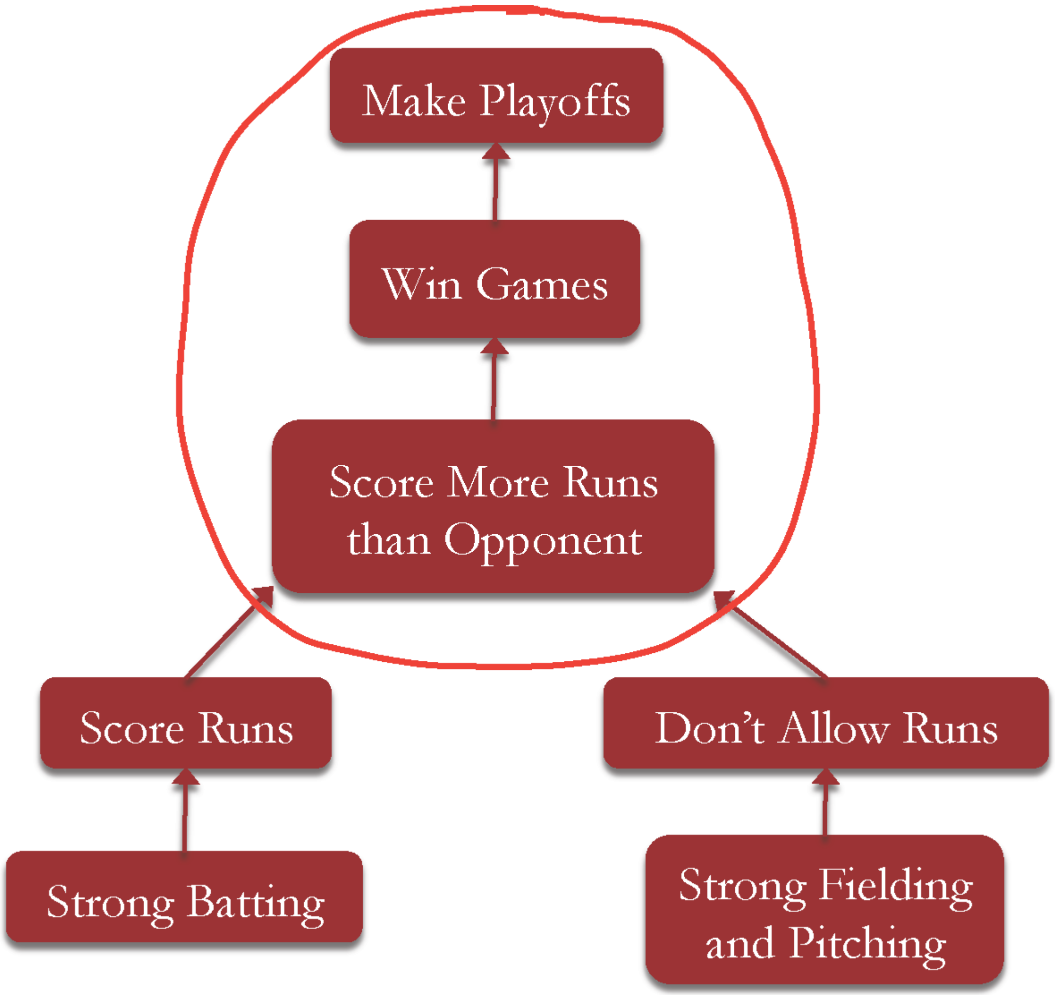
Taking a Quantitative View

- Paul DePodesta spent a lot of time looking at the data
- His analysis suggested that some skills were undervalued and some skills were overvalued
- If they could detect the undervalued skills, they could find players at a bargain

DIMENSIONLESS
TECHNOLOGY

Making it to the Playoffs

The Goal of a Baseball team

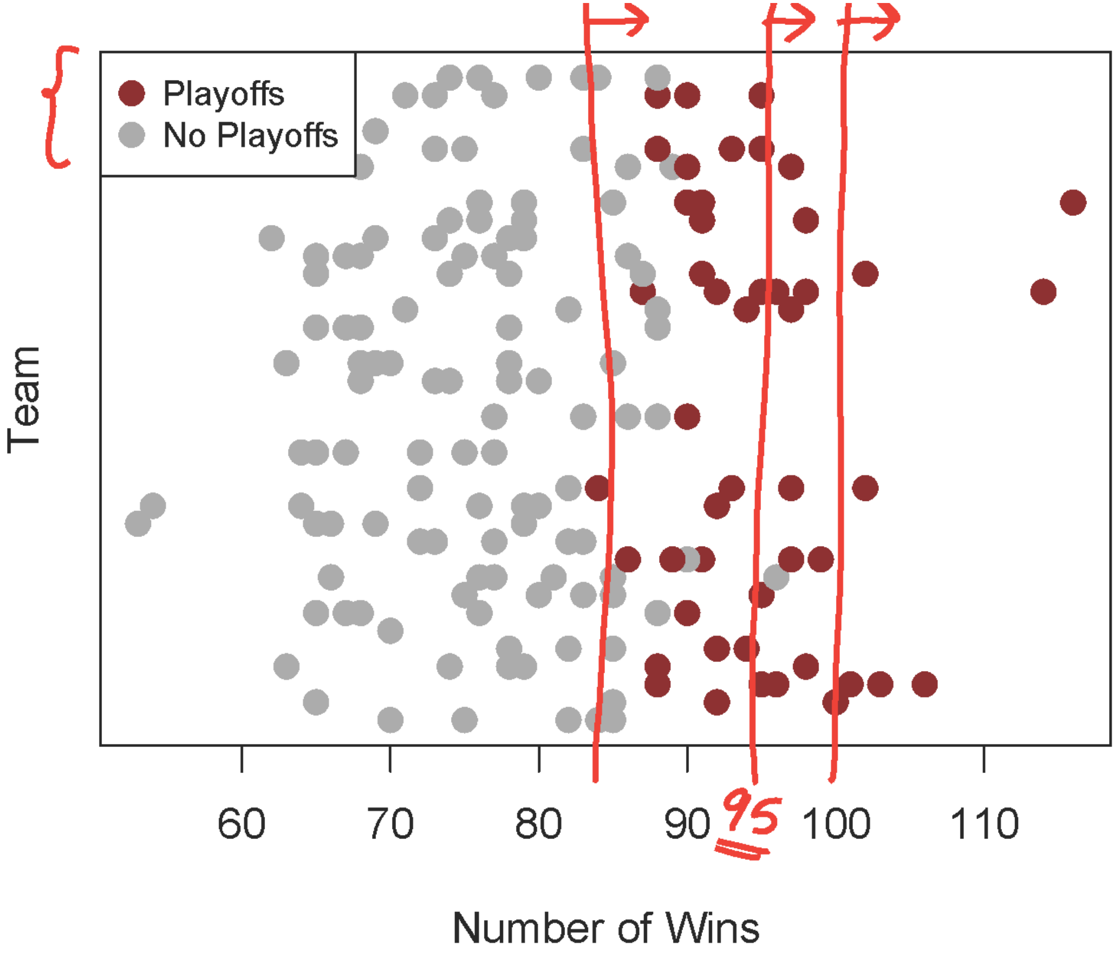


Making it to the Playoffs

- How many games does a team need to win in the regular season to make it to the playoffs?

- "Paul DePodesta reduced the regular season to a math problem. He judged how many wins it would take to make it to the playoffs: 95."

Making it to the Playoffs



Data from all teams 1996-2001

Winning 95 Games

- How does a team win games?
- They score more runs than their opponent
- But how many more?
- The A's calculated that they needed to score 135 more runs than they allowed during the regular season to expect to win 95 games
- Let's see if we can verify this using linear regression

Linear Regression in R

- Read the file “baseball.csv” and save into a dataframe “baseball”
  - https://storage.googleapis.com/dimensionless/Analytics/baseball.csv
- Look at the structure of the dataframe.
- This data set includes all observation for every team and year pair from 1962 to 2012.
- 15 Variables
  - RS:- Runs Scored
  - RA:- Runs Allowed
  - W:- Wins

Loading the data

- We want to build models using data Paul DePodesta had in 2002, so let's start by sub setting our data.
- Subset the data for Year<2002, and save it in data frame "moneyball"
- See the structure of "moneyball", you can find 902 observations for 15 variables.
- Build a linear regression equation to predict wins using the difference between runs scored and runs allowed.
- Create a new variable moneyball$rd
  - moneyball$RD = moneyball$RS - moneyball$RA

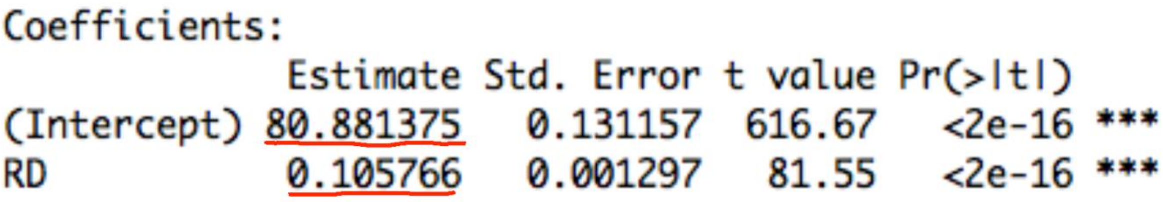
Building the model

- Visually check to see if there's a linear relationship between Run Difference and Wins.
- Create a scatter plot with the plot function.
- On the x-axis, put RD, Run Difference, and on the y-axis, put W, Wins.
- This scatter plot shows us that there's a very strong linear relationship between these two variables, which is a good sign for our linear regression equation.

- Create linear regression model, which we'll call WinsReg.
  - WinsReg = lm(W ~ RD, data=moneyball)
  - summary(WinsReg)

Confirming the Claims

- Can we confirm the claim made in Moneyball that a team needs to score at least 135 more runs than they allow to win at least 95 games.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 80.881375   0.131157  616.67   <2e-16 ***
RD           0.105766   0.001297   81.55   <2e-16 ***
```

$$W = 80.8814 + 0.1058(RD)$$

$$W \geq 95$$

$$80.8814 + 0.1058(RD) \geq 95$$

$$RD \geq \frac{95 - 80.8814}{0.1058} = 133.4 \sim 135$$

Quick Question

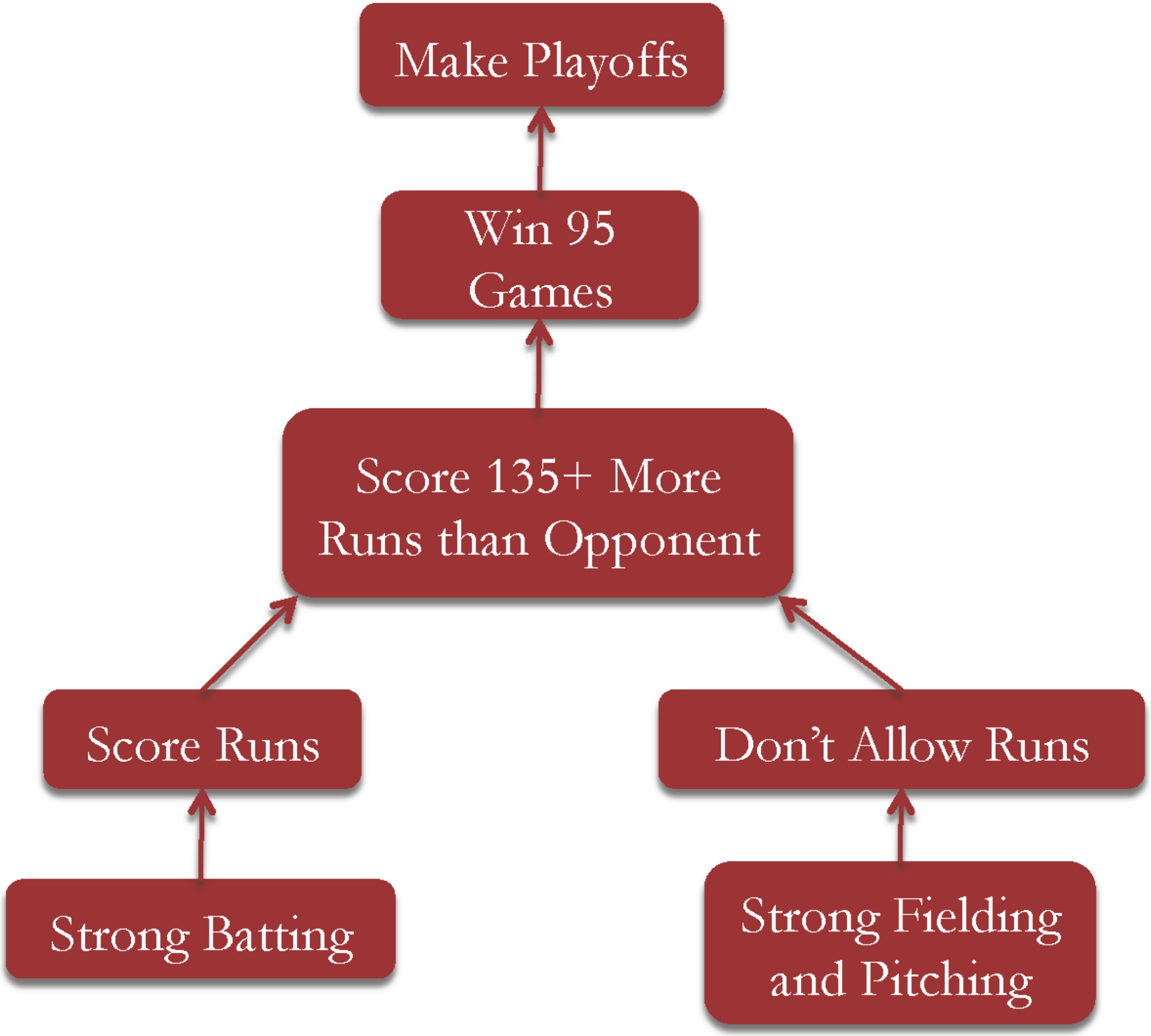
- If a baseball team scores 713 runs and allows 614 runs, how many games do we expect the team to win?
- Using the linear regression model constructed, find the number of games we expect the team to win:

Ans:-91

DIMENSIONLESS
TECHNOLOGY

Predicting Runs

The Goal of a Baseball team



Make Playoffs

Win 95 Games

Score 135+ More Runs than Opponent

Score Runs

Don’t Allow Runs

Strong Batting

Strong Fielding and Pitching

Scoring Runs

- How does a team score more runs?
- The A's discovered that two baseball statistics were significantly more important than anything else
  - On-Base Percentage (OBP)
    - Percentage of time a player gets on base (including walks)
  - Slugging Percentage (SLG)
    - How far a player gets around the bases on his turn (measures power)

Regression in R

- Let's take a look at the structure of our data, “moneyball” again, using the str function.
- OBP :- on-base percentage
- SLG :- Slugging percentage
- BA :- Batting Average
- Build a linear regression equation.
- We'll call it RunsReg.
  - RunsReg = lm(RS ~ OBP + SLG + BA, data=moneyball)
- Look at the summary of RunsReg
  - summary(RunsReg)
- You can see all the 3 variables are significant
- R^2 is 0.93
- Coefficient of BA is negative, which is counter-intuitive.

Selecting the Variables

- Case of multicollinearity.
- These three hitting statistics are highly correlated.
- Try removing batting average and check.
  - RunsReg = lm(RS ~ OBP + SLG, data=moneyball)
  - summary(RunsReg)
- R2 is still 0.93.
- So this model is simpler, with only two independent variables, and has about the same R-squared. Overall a better model.
- Experiment and see what if we'd removed on-base percentage or slugging percentage instead of batting average.
- Coefficient of OBP is greater than SLG.

Allowing Runs

- We can use pitching statistics to predict runs allowed
  - Opponents On-Base Percentage (OOBP)
  - Opponents Slugging Percentage (OSLG)

- We get the linear regression model

  Runs Allowed = -837.38 + 2913.60(OOBP) + 1514.29(OSLG)

- $R^2 = 0.91$
- Both variables significant

Quick Question

If a baseball team's OBP is 0.311 and SLG is 0.405, how many runs do we expect the team to score?

Using the linear regression model constructed (the one that uses OBP and SLG as independent variables), calculate the number of runs we expect the team to score:

Ans:- 689

Quick Question

If a baseball team's opponents OBP (OOBP) is 0.297 and opponents SLG (OSLG) is 0.370, how many runs do we expect the team to allow?

Using the linear regression model discussed. Calculate the number of runs we expect the team to allow:

Ans:- 588

DIMENSIONLESS
TECHNOLOGY

Making Predictions

Predicting Runs and Wins

- Can we predict how many games the 2002 Oakland A's will win using our models?
- The models for runs use team statistics
- Each year, a baseball team is different
- We need to estimate the new team statistics using past player performance
  - Assumes past performance correlates with future performance
  - Assumes few injuries
- We can estimate the team statistics for 2002 by using the 2001 player statistics

Predicting Runs Scored

- At the beginning of the 2002 season, the Oakland A's had 24 batters on their roster
- Using the 2001 regular season statistics for these players
  - Team OBP is 0.339
  - Team SLG is 0.430
- Our regression equation was

$$RS = -804.63 + 2737.77(OBP) + 1584.91(SLG)$$

- Our 2002 prediction for the A's is

$$RS = -804.63 + 2737.77(0.339) + 1584.91(0.430) = 805$$

Predicting Runs Allowed

- At the beginning of the 2002 season, the Oakland A's had 17 pitchers on their roster
- Using the 2001 regular season statistics for these players
  - Team OOBP is 0.307
  - Team OSLG is 0.373
- Our regression equation was

$$RA = -837.38 + 2913.60(OOBP) + 1514.29(OSLG)$$

- Our 2002 prediction for the A's is

$$RA = -837.38 + 2913.60(0.307) + 1514.29\ (0.373) = 622$$

Predicting Wins

- Our regression equation to predict wins was

$$\text{Wins} = 80.8814 + 0.1058(\text{RS} - \text{RA})$$

- We predicted
  - RS = 805
  - RA = 622

- So our prediction for wins is

$$\text{Wins} = 80.8814 + 0.1058(805 - 622) = 100$$

The Oakland A's

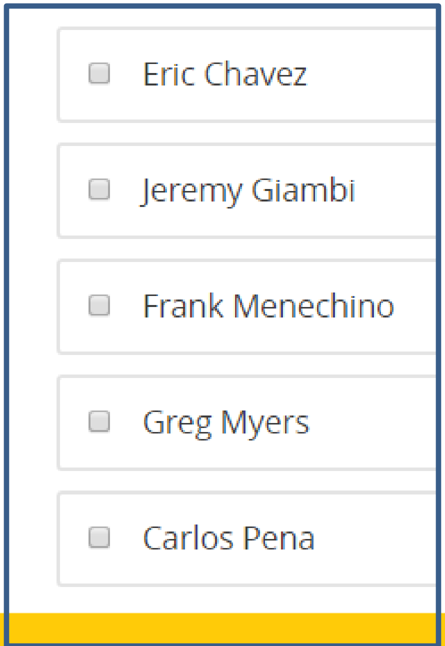
- Paul DePodesta used a similar approach to make predictions
- Predictions closely match actual performance

	Our Prediction	Paul's Prediction	Actual
Runs Scored	805	800 – 820	800
Runs Allowed	622	650 – 670	653
Wins	100	93 – 97	103

- The A's set a League record by winning 20 games in a row
- Won one more game than the previous year, and made it to the playoffs

Quick Question

- Suppose you are the General Manager of a baseball team, and you are selecting TWO players for your team. You have a budget of $1,500,000, and you have the choice between the following players:
- Given your budget and the player statistics, which TWO players would you select?



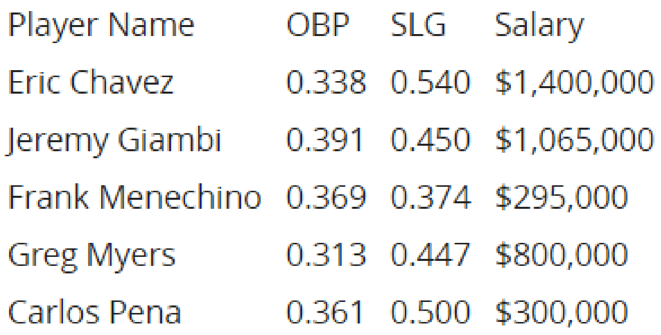
☐ Eric Chavez

☐ Jeremy Giambi

☐ Frank Menechino

☐ Greg Myers

☐ Carlos Pena

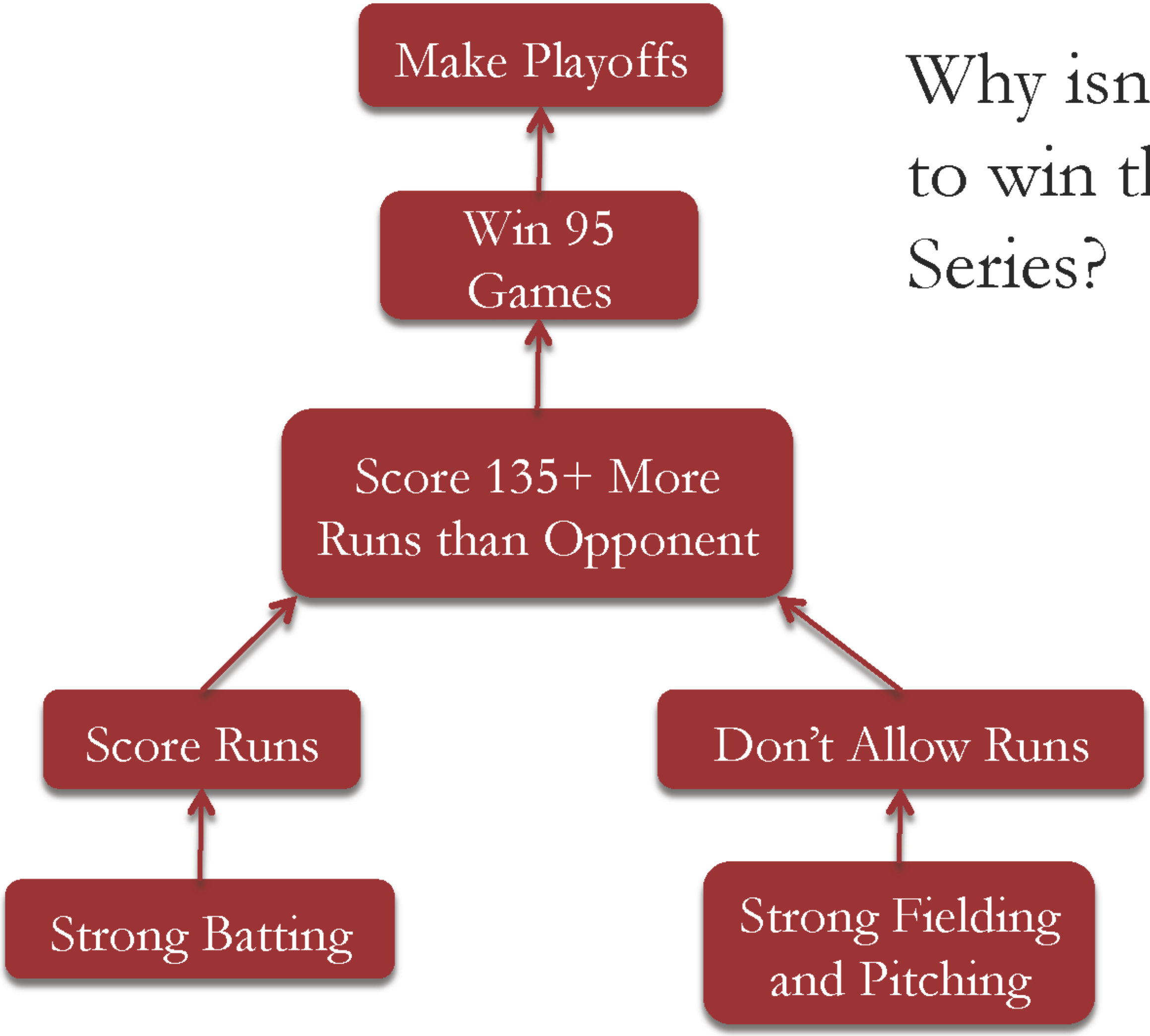


Player Name	OBP	SLG	Salary
Eric Chavez	0.338	0.540	$1,400,000
Jeremy Giambi	0.391	0.450	$1,065,000
Frank Menechino	0.369	0.374	$295,000
Greg Myers	0.313	0.447	$800,000
Carlos Pena	0.361	0.500	$300,000

DIMENSIONLESS TECHNOLOGY

Winning the World Series

The Goal



Why isn't the goal to win the World Series?

Luck in the Playoffs

- Billy and Paul see their job as making sure the team makes it to the playoffs – after that all bets are off
  - The A's made it to the playoffs in 2000, 2001, 2002, 2003
  - But they didn't win the World Series
- Why?
- "Over a long season the luck evens out, and the skill shines through. But in a series of three out of five, or even four out of seven, anything can happen."

Is Playoff Performance Predictable ?

- Using data 1994-2011 (8 teams in the playoffs)
- Correlation between winning the World Series and regular season wins is 0.03
- Winning regular season games gets you to the playoffs
- But in the playoffs, there are too few games for luck to even out
- *Logistic regression* can be used to predict whether or not a team will win the World Series

Quick Question

In 2012 and 2013, there were 10 teams in the MLB playoffs: the six teams that had the most wins in each baseball division, and four "wild card" teams. The playoffs start between the four wild card teams - the two teams that win proceed in the playoffs (8 teams remaining). Then, these teams are paired off and play a series of games. The four teams that win are then paired and play to determine who will play in the World Series.

We can assign rankings to the teams as follows:

Rank 1: the team that won the World Series

Rank 2: the team that lost the World Series

Rank 3: the two teams that lost to the teams in the World Series

Rank 4: the four teams that made it past the wild card round, but lost to the above four teams

Rank 5: the two teams that lost the wild card round

Quick Question

In your R console, create a corresponding rank vector by typing

```
teamRank = c(1,2,3,3,4,4,4,4,5,5)
```

In this quick question, we'll see how well these rankings correlate with the regular season wins of the teams. In 2012, the ranking of the teams and their regular season wins were as follows:

Rank 1: San Francisco Giants (Wins = 94)

Rank 2: Detroit Tigers (Wins = 88)

Rank 3: New York Yankees (Wins = 95), and St. Louis Cardinals (Wins = 88)

Rank 4: Baltimore Orioles (Wins = 93), Oakland A's (Wins = 94), Washington Nationals (Wins = 98), Cincinnati Reds (Wins = 97)

Rank 5: Texas Rangers (Wins = 93), and Atlanta Braves (Wins = 94)

Create a vector in R called wins2012, that has the wins of each team in 2012, in order of rank (the vector should have 10 numbers).

In 2013, the ranking of the teams and their regular season wins were as follows:

Rank 1: Boston Red Sox (Wins = 97)

Rank 2: St. Louis Cardinals (Wins = 97)

Rank 3: Los Angeles Dodgers (Wins = 92), and Detroit Tigers (Wins = 93)

Rank 4: Tampa Bay Rays (Wins = 92), Oakland A's (Wins = 96), Pittsburgh Pirates (Wins = 94), and Atlanta Braves (Wins = 96)

Rank 5: Cleveland Indians (Wins = 92), and Cincinnati Reds (Wins = 90)

Create another vector in R called wins2013, that has the wins of each team in 2013, in order of rank (the vector should have 10 numbers).

Quick Question

- What is the correlation between teamRank and wins2012?
- What is the correlation between teamRank and wins2013?

- Ans:- 0.35,-0.66

www.dimensionless.in

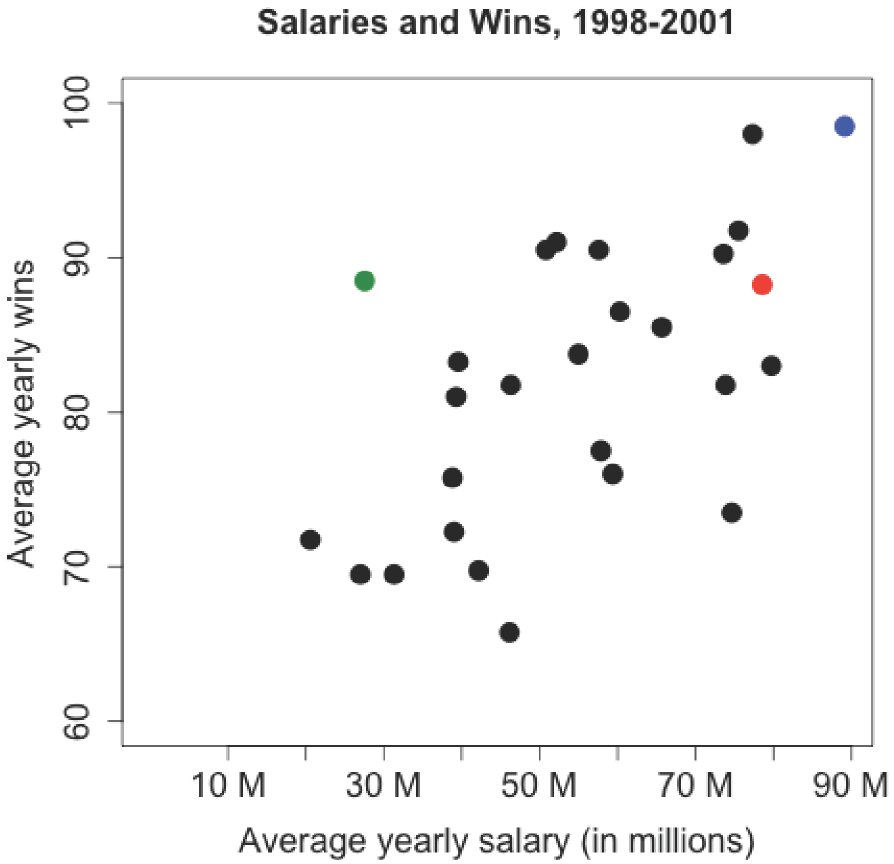
DIMENSIONLESS
TECHNOLOGY

The Analytics
Edge in Sports

Other Moneyball Strategies

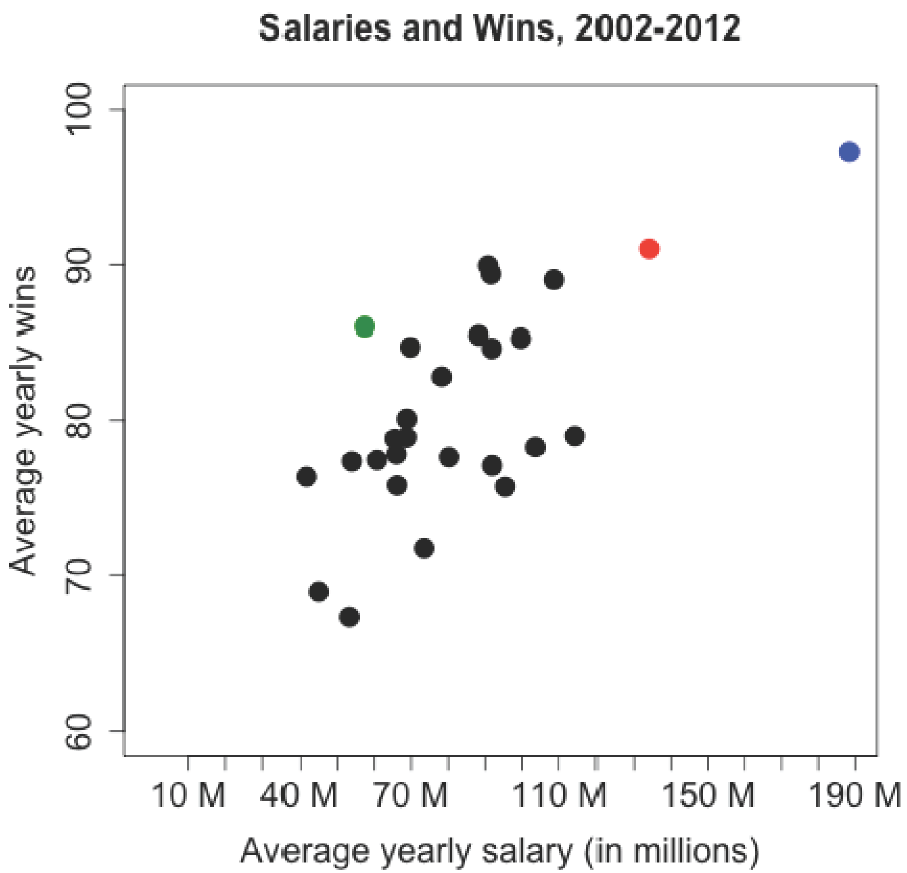
- *Moneyball* also discusses:
  - How it is easier to predict professional success of college players than high school players
  - Stealing bases, sacrifice bunting, and sacrifice flies are overrated
  - Pitching statistics do not accurately measure pitcher ability – pitchers only control strikeouts, home runs, and walks

Where was Baseball in 2002



- Before Moneyball techniques became more well-known, the A's were an outlier
- 20 more wins than teams with equivalent payrolls
- As many wins as teams with more than double the payroll

Where is Baseball now ?



- Now, the A's are still an efficient team, but they only have 10 more wins than teams with equivalent payrolls
- Fewer inefficiencies

Sabermetrics

- Sabermetrics is a more general term for Moneyball techniques
- There has been a lot of work done in this field
  - Baseball Prospectus (www.baseballprospectus.com)
  - Value Over Replacement Player (VORP)
  - Defense Independent Pitching Statistics (DIPS)
  - *The Extra 2%: How Wall Street Strategies Took a Major League Baseball Team from Worst to First*
    - A story of the Tampa Bay Rays
  - Game-time decisions: batting order, changing pitchers, etc.

Other Baseball Teams and Sports

- Every major league baseball team now has a statistics group
- The Red Sox implemented quantitative ideas and won the World Series for the first time in 86 years
- Analytics are also used in other sports, although it is believed that more teams use statistical analysis than is publically known

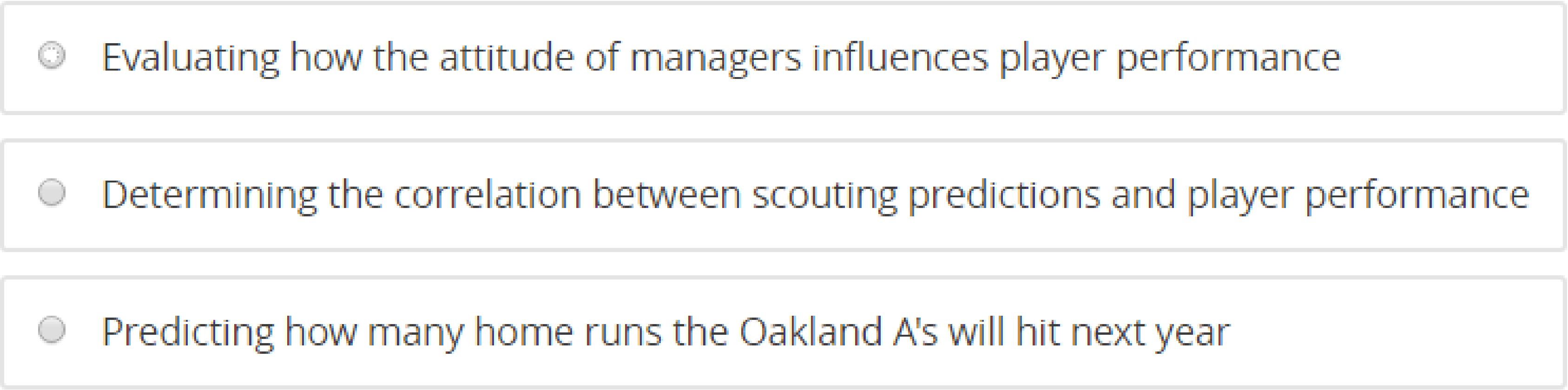
The Analytics Edge

- Models allow managers to more accurately value players and minimize risk
  - "In human behavior there was always uncertainty and risk. The goal of the Oakland front office was simply to minimize the risk. Their solution wasn't perfect, it was just better than … rendering decisions by gut feeling."

- Relatively simple models can be useful

Quick Question

Which of the following is MOST LIKELY to be a topic of Sabermetric research?



○ Evaluating how the attitude of managers influences player performance

○ Determining the correlation between scouting predictions and player performance

○ Predicting how many home runs the Oakland A's will hit next year

Analytics in various sports

- Basketball
  - 82games.com
- Soccer
  - socceranalysts.com, soccermetrics.net
- Cricket
  - Cricmetric.com, impactindexcricket.com
- Hockey
  - Hockeyanalytics.com, lighthousehockey.com