Data 8 Connector: Sports Analytics

...

What is this course?

Data Science + Sports = Sports Analytics

Why this course?

Data Scientist + Sports Enthusiast = Teach Sports Analytics

Why this course?

Data Science Student + Sports Enthusiast = Learn Sports Analytics

What is this course about?

Demystifying sports analytics

Lots of talk, lots of jargon, lots of sites and blogs.

Most is actually quite simple: collecting data that wasn't there before and basic summarizations.

Ex: Pick and roll pairs efficiency. The challenge is *getting* the data on pick and roll plays

There's still a hefty chunk of quite interesting work that isn't so simple.

What is this course about?

Not just the what, but the how and the why

Could plow through a lot of concepts in two weeks and give a coarse overview.
Ex: What is a park factor? Not how it's computed or how to use it or why.

The point is to get a feel for how this kind of work is done and why it's done the way it is.

Also a great way to connect with the DS/Stats/Math field at-large through sports and data

What is this course about?

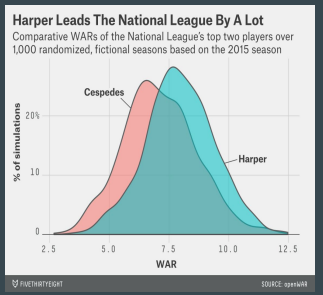
Hands-on with data and Python

We're going to try to be as hands on as possible.

You don't want to hear me just talk. And it's pretty fun developing this stuff.

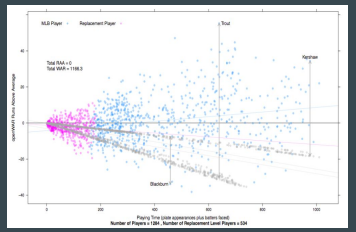
What do we want to explore?

Measuring performance

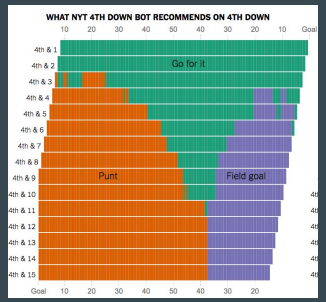


Advanced Metrics

When Stats Go Bad

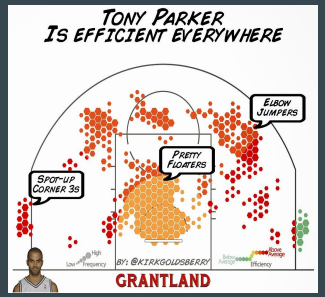


Regression modeling



Inference

Assorted Topics



What we hope to learn

What goes into producing advanced metrics

Read and discuss data oriented analysis (and ignore the taeks)

Begin your own path of getting into the world of sports analytics

Course Details

➔ Your Instructor

Alex Papanicolaou

Formerly a Data Scientist in FX trading Researcher in risk for finance Sports enthusiast

➔ Your Course Assistant

Anastasia Vela

Sophomore in CS

Course Details

Prerequisites

- ➔ Data 8: we're going to move fairly fast
- ➔ Expected value/average
- ➔ Variance/Standard Deviation/Dispersion
- ➔ Simple linear regression/line of best fit

Course Details

- Participation
  - Show up and participate. In class and online discussions
- HW: “Weekly”
  - Not supposed to be hard. We’ll ease up if it’s too hard. We’re experimenting
- Grading policy
  - Checks and plus/minus for participation and HW
- Course site:
  - Piazza: you should already be enrolled

Course Details

Textbooks, resources, and other reading material (nothing required)

“Textbooks”

- ➔ The Book (Tom Tango and Mitchel Lichtman)
- ➔ Analyzing Baseball Data with R (Jim Albert and Max Marchi)
- ➔ Mathletics (Wayne L. Winston)
- ➔ Basketball on Paper (Dean Oliver)

Course Details

Blogs

- ➔ The Hardball Times (FanGraphs Blog)
- ➔ 538
- ➔ Beyond the Box Score
- ➔ Nylon Calculus

Course Details

Other books

- ➔ Moneyball
- ➔ The Undoing Project
- ➔ Big Data Baseball

There’s no shortage of resources and these lists are by no stretch of the imagination complete

Course Details

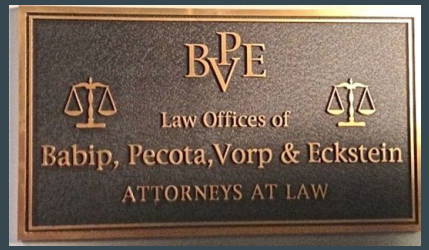
- Group Project
  - Groups of about 3-4
  - Final written report at the end (3-5 pages). Presentations last week.
  - First milestone: form a group and submit a proposal by end of Week 3.
  - More details on Piazza

We want effort devoted to attendance, participation and this project.

Let’s Begin!

Analytical Thinking, Runs, and Wins

Analytical Thinking



What is Sabermetrics about?

Bill James: “the search for objective knowledge about baseball.”

Developing hypotheses, using logic and reasoning, and grounding it in data

Ex: Is a sacrifice bunt a worthwhile strategy?

(Go to the data and see if a bunt increases your chances to score)

Ex: Pitching is X% of baseball.

(Not directly evaluated but examined through the consequences of such a claim)

Analytical Thinking

What's the process?

Analytical Thinking

What's the process?

No, not that process.

Analytical Thinking

What’s the process?

Learn about the games, how they're played, what matters (e.g. "grit" doesn't matter)

Then measure what matters (getting hits, getting on base, hitting for power, scoring)

Then start making deductions and inferences (Player A is good at X, strategy Y is not worth it)

Analytical Thinking

Challenges and also critical thinking

- ➔ What can we answer?
- ➔ What are the caveats?
- ➔ When can we draw conclusions? And what kind?
- ➔ What do we need to do to make improvements?

Analytical Thinking

What can we answer?

- ➔ Data determines what you can answer and how well
- ➔ We have performance data so we can ask...
  ...How much should we value a player?
  ...Which strategy should we use?

- ➔ We cannot address inane questions like...
  ...Who plays the “right” way?
- ➔ But we also cannot address important questions like...
  ...Who’s shooting form is better? (Unless you have video and advanced computer vision methods)
  ...How is a certain player “feeling” or if that matters? (Orgs would probably keep this under wraps)

Analytical Thinking

What are the caveats?

- ➔ What are the limits of the data we are using?
- ➔ In baseball...
  Do we have the data on where the ball was hit or where the defenders were located?
  Do we have the pitch types?
- ➔ In basketball...
  Do we have player tracking data for fine grain analysis? Who set screens or made cuts?
- ➔ How much data do we have?

This is universal:

We can only analyze what we measure
And our certainty is limited by the amount of data

Analytical Thinking

When can we draw conclusions?

- ➔ This is a question of robustness
- ➔ Nothing is certain
- ➔ But we can poke, prod, and test our analyses for issues
- ➔ And we can quantify uncertainty
- ➔ Ex:

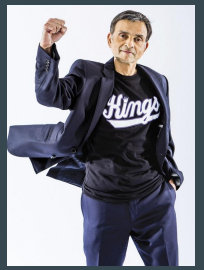
  Does an analysis/method show some no-name player is better than Lebron?

  The 2018 Giants are expected to be X wins better. What is the uncertainty in that projection?

Analytical Thinking

What kind of conclusions are we looking for?

We aren't trying to produce non-stop Gladwellian counterintuitive conclusions



Often, we just want to put numbers to things we already know (i.e., match the "eye test")

Analytical Thinking

What do we need to do to make improvements?

- ➔ This is part of the poking and prodding process
- ➔ Did you fail to account for something?
- ➔ Are the modeling assumptions valid enough?
  For example, basketball lineups (5 man units played by the coach) are not built as randomized controlled trials. Does that ruin our analyses?
- ➔ Do we even have the data to address the question?
  What is the value of a ball distributor who doesn't fill up the stat sheet due to secondary assists?

1. Add more features to the model
2. Or, get more data
   For example: without access to advanced tracking data, get humans to gather information

Analytical Thinking

Let’s push this into practice

What matters in sports? What is our outcome? Wins

Analytical Thinking

> Wins and losses are the currency of baseball. They're the only things that count in the standings, so we want to develop statistics and metrics that align with that reality. Teams are trying to win.
>
> - FanGraphs

Great! That's easy to measure.

Analytical Thinking

Okay, so now what?

- ➔ How do we measure a player's effect on wins?
- ➔ How do we evaluate one strategy versus another?
- ➔ Walk-off home runs and game winning shots seem to have direct effects on wins

  We say "Player A won the game"
- ➔ Pinch-hits show the manager "pushed the right buttons"

Analytical Thinking

> [“Teams are trying to win” is] a rather obvious statement, but you have to build your framework for evaluating players from the ground up. You win games by scoring more runs than your opponent, meaning that run scoring and prevention are the building blocks of wins and losses. So when we say we care about winning and losing, we really care about scoring and preventing runs and the things players do toward those ends.
>
> - Continued from FanGraphs

Runs and points lead to wins. That's a fundamental feature of sports.

You need to score runs/points and prevent runs/points for the other team.

Analytical Thinking

The essence of it all:

- Runs, points, events: this is what's rich and data dense
- Collect as much as you can about what is happening that produced runs
- The questions expand beyond reductive win-loss binary outcomes

For another read on this, see *The Sabermetric Manifesto* by David Grabiner

http://seanlahman.com/baseball-archive/sabermetrics/sabermetric-manifesto/

We'll explore this simple relationship in our first lab...

Data

Before we begin...

Let’s cover a bit of the data that’s out there to help us answer questions.

Data

What type of datasets are out there?

- ➔ Raw: game box scores; play-by-play; player tracking
- ➔ Extracted Events: hits, runs, points, rebounds, assists, etc
- ➔ Stats: batting avg, total bases, RBI, shooting %, etc

Data

Where can we find this data?

- Websites:
  - Leagues: MLB.com, NBA.com
  - General: ESPN, Baseball/Basketball/Football Reference, FanGraphs
- API/Published
  - PitchF/X, Statcast
  - NBA Stats
- Curated (not necessarily free)
  - Lahman Database, Retrosheet, armchairanalysis.com (cheap with .edu email)
- Other
  - API tools and scrapers published on GitHub (LOTS of repos out there)
  - Data Collectives: Kaggle, data.world

Data

Focus on MLB/NBA: other data not as prominent or collected as well: e.g. NFL

Lab 1: Runs, Points, and Wins

Let's begin with a simple lab

The lab explores

- ➔ The empirical relationship between runs and wins
- ➔ Derives the Pythagorean Expectation formula
- ➔ Explores some of its consequences
- ➔ Does the same for NBA data

Lab 1: Runs, Points, and Wins

Let’s begin with a simple lab that will explore the relationship between runs and wins

Lab 1: Runs, Points, and Wins

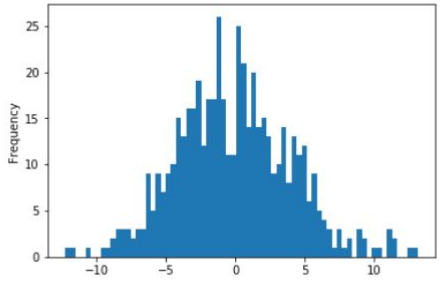
What can we take away from the lab on Pythagorean Expectation?

Lab 1: Runs, Points, and Wins

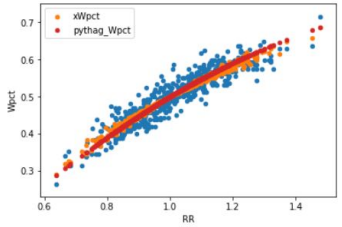
Baseball, and sports in general, lends itself to lots of great empirical work

	yearID	lgID	teamID	franchID	G	W	L	R	RA	name	Wpct	Lpct	RD	RDperG
2325	2000	AL	ANA	ANA	162	82	80	864	869	Anaheim Angels	0.506173	0.493827	-5	-0.030864
2326	2000	NL	ARI	ARI	162	85	77	792	754	Arizona Diamondbacks	0.524691	0.475309	38	0.234568
2327	2000	NL	ATL	ATL	162	95	67	810	714	Atlanta Braves	0.586420	0.413580	96	
2328	2000	AL	BAL	BAL	162	74	88	794	913	Baltimore Orioles	0.456790	0.543210	-119	
2329	2000	AL	BOS	BOS	162	85	77	792	745	Boston Red Sox	0.524691	0.475309	47	

```
In [12]: 1 mlb_df['pythag_luck'].plot.hist(bins=70)
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x113298358>
```



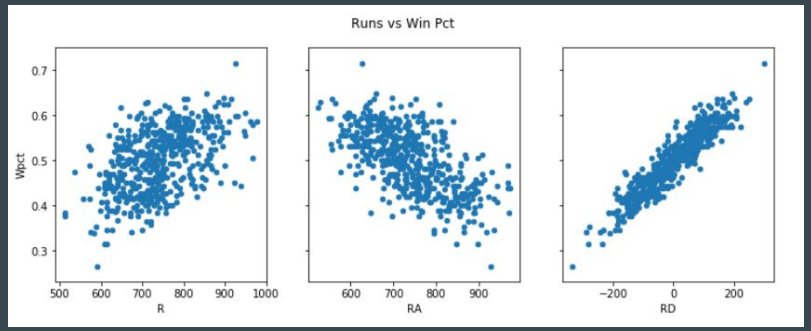

```
In [5]:
1 fig, ax = plt.subplots()
2 mlb_df.plot.scatter(ax=ax, x="RR", y='Wpct')
3 mlb_df.plot.scatter(ax=ax, x="RR", y='xWpct', color='C1', label='xWpct')
4 mlb_df.plot.scatter(ax=ax, x="RR", y='pythag_Wpct', color='C3', label='pythag_Wpct')
5 plt.legend()
6 ax.set_ylabel('Wpct');
```



Lab 1: Runs, Points, and Wins

The relationship between runs and wins is strong

$$\text{Pythagorean Win Pct} = \frac{\text{Runs Scored}^2}{\text{Runs Scored}^2 + \text{Runs Allowed}^2} = \frac{\text{Runs Scored / Runs Allowed}}{1 + (\text{Runs Scored / Runs Allowed})^2}$$



We can also use it to project second-half season performance!

Lab 1: Runs, Points, and Wins

We can specifically quantify about how many runs a win is equal to

Generally, we find that 10 runs equals a win

How does this help? Interpretation and contextualizing information

Ex: Player A is projected to be 20 runs better than Player B
Ex: Player A is projected to be 2 wins better than Player B

Which makes more sense?

Lab 1: Runs, Points, and Wins

This strong relationship also exists in the NBA and will apply more broadly.

But also! The nature of the sport is revealed in the Pythagorean Exponent.

Empirical results show that increases in point differential yield stronger responses in winning percentage

Hollinger's Pythagorean Expectation (Different exponents derived by Daryl Morey and Dean Oliver)

$$\text{Pythagorean Win Pct} = \frac{\text{Points For}^{16.5}}{\text{Points For}^{16.5} + \text{Points Against}^{16.5}} = \frac{(\text{Points For} \ / \ \text{Points Against})^{16.5}}{1 + (\text{Points For} \ / \ \text{Points Against})^{16.5}}$$

Summary

Analytical thinking

Empirical relationship betweens runs/points and wins

Summary

What do we want? Wins.

How do we do that? Scoring and preventing runs (or points)

No magic bullets!

But you can make headway towards improved solutions

And that's what this course is about

Summary

Next up:

- Measuring performance
- Run expectancy
- Run creation