Basketball Analytics with R

UConn Sports Analytics Symposium 2020

Jackson P. Lautier

10/10/2020

Workshop Outline

- Introduce the `R` package `BasketballAnalyzeR`
- Introduce Fundamental (Advanced) Basketball Statistics
- Recap 17-18' National Basketball Association (NBA) Season
- Statistical Case Studies
  - Is Kevin Durant 'good at basketball'?
  - Visualizing the value of Kyle Lowry
  - Understanding the Golden State Warriors (GSW) offense
  - Clustering NBA teams
  - The lost art of the mid-range
- Summary
- References and Further Reading

BasketballAnalyzeR

- Paola Zuccolotto and Marica Manisera (2020), *Basketball Data Science – with Applications in R*. Chapman and Hall/CRC. ISBN 9781138600799
- https://bdsports.unibs.it/
- https://bdsports.unibs.it/basketballanalyzer/

Figure 1: Buy Me! I'm a 'Slam Dunk'

BasketballAnalyzeR Cont.

- Includes many useful functions: `shotchart()`, `fourfactors()`, `assistnet()`, `expectedpts()`, and many more
- Includes preloaded datasets for the 2017-2018 NBA season:
  - `Obox`: GSW opponent's box scores
  - `PbP.BDP`: GSW play-by-play data
  - `PBox`: Players box score statistics
  - `Tadd`: Team Standings
  - `TBox`: Team box score statistics
- All figures and analysis in today's presentation compiled using `BasketballAnalyzer`. See associated `github` materials for code:
- https://github.com/jackson-lautier/UCSAS-Basketball-Analytics-R

Fundamental Basketball Statistics

Table 2.4 (Zuccolotto and Manisera)

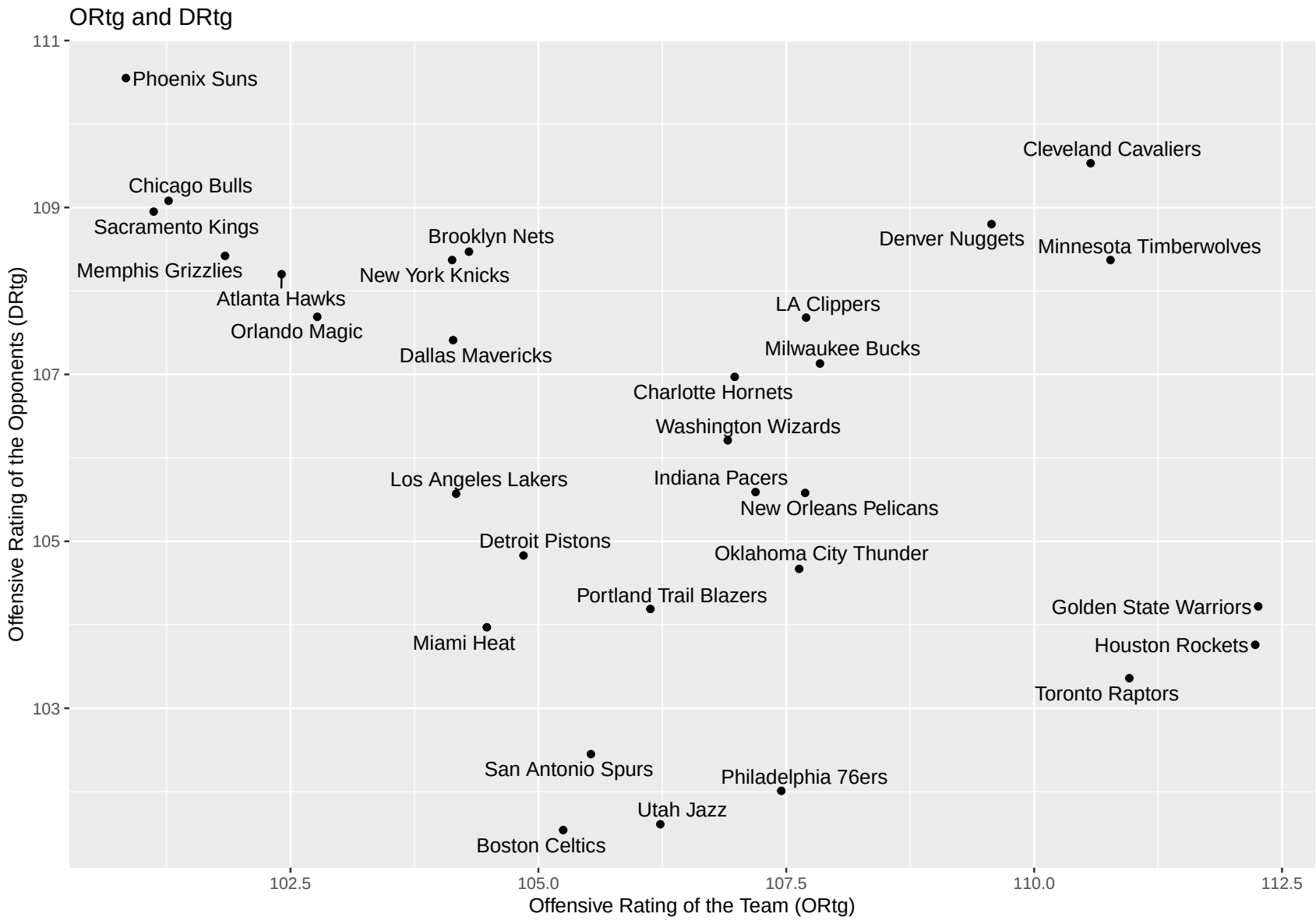
Factor	Offense	Defense
eFG%	$\frac{(2PM)_T+1.5\times(3PM)_T}{(2PA)_T+(3PA)_T}$	$\frac{(2PM)_O+1.5\times(3PM)_O}{(2PA)_O+(3PA)_O}$
TO Ratio	$\frac{TOV_T}{POSS_T}$	$\frac{TOV_O}{POSS_O}$
REB%	$\frac{OREB_T}{OREB_T+DREB_O}$	$\frac{DREB_T}{OREB_O+DREB_T}$
FT Rate	$\frac{FTM_T}{(2PA)_T+(3PA)_T}$	$\frac{FTM_O}{(2PA)_O+(3PA)_O}$

The *Four Factors* by Kubatko, J., Oliver, D., Pelton, K., and Rosenbaum, D. T. (2007). *A starting point for analyzing basketball statistics*. Journal of Quantitative Analysis in Sports, 3(3):1–22

2017-2018 NBA Season Review

- NBA Champions: Golden State Warriors def. Cleveland Cavaliers (4-0); (CLE def. BOS, GSW def. HOU)
- Finals MVP: Kevin Durant (GSW)
- MVP: James Harden (HOU)
- ROY: Ben Simmons (PHI)
- DPOY: Rudy Gobert (UTA)
- All-NBA First Team: F - Kevin Durant (GSW), F - LeBron James (CLE), C - Anthony Davis (NOP), G - James Harden (HOU), G - Damian Lillard (POR)

2017-2018 NBA Season Review (Cont.)



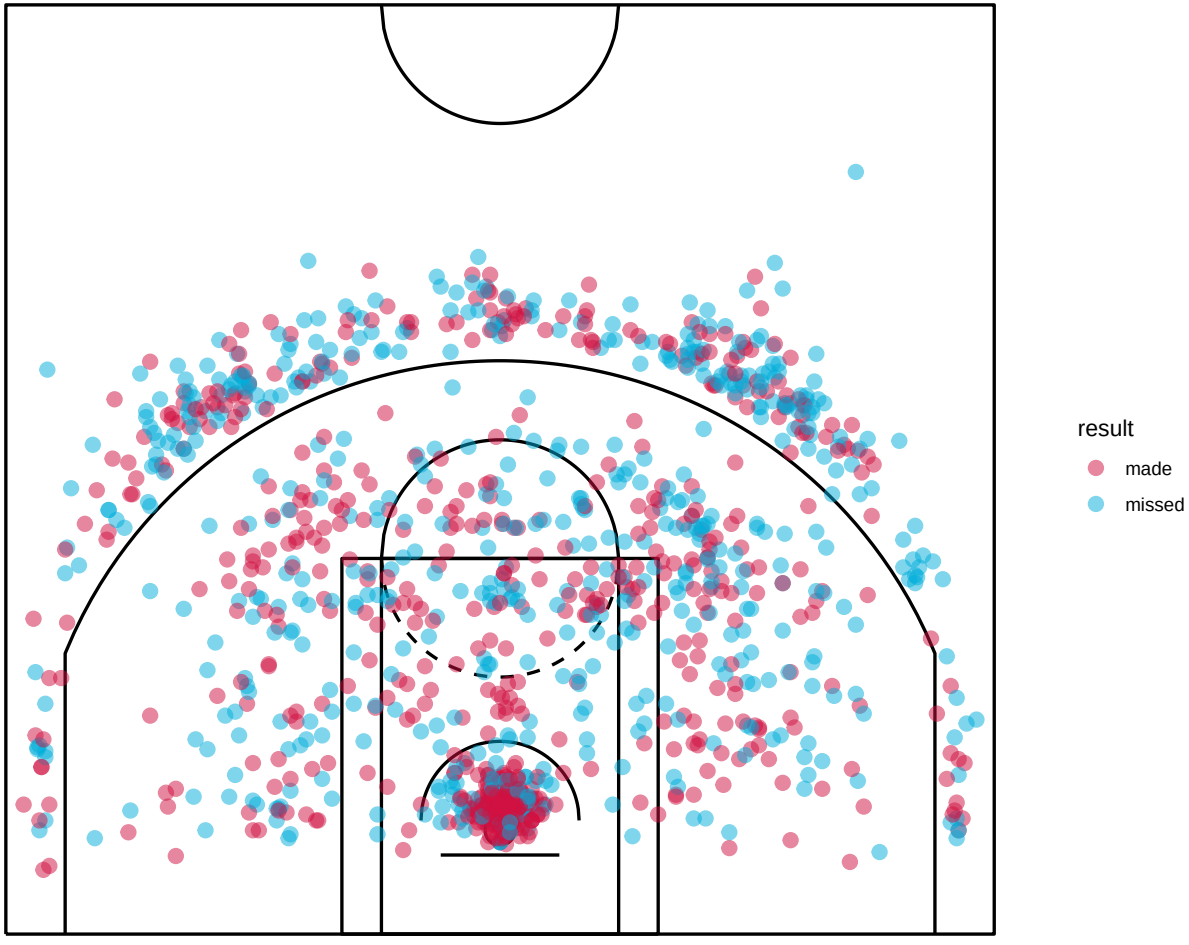
CS1: Kevin Durant

Figure 2: “I don’t need analytics to tell me LeBron James is good at basketball." -Jeff Van Gundy

Warm-up: Is Kevin Durant ‘good at basketball’?

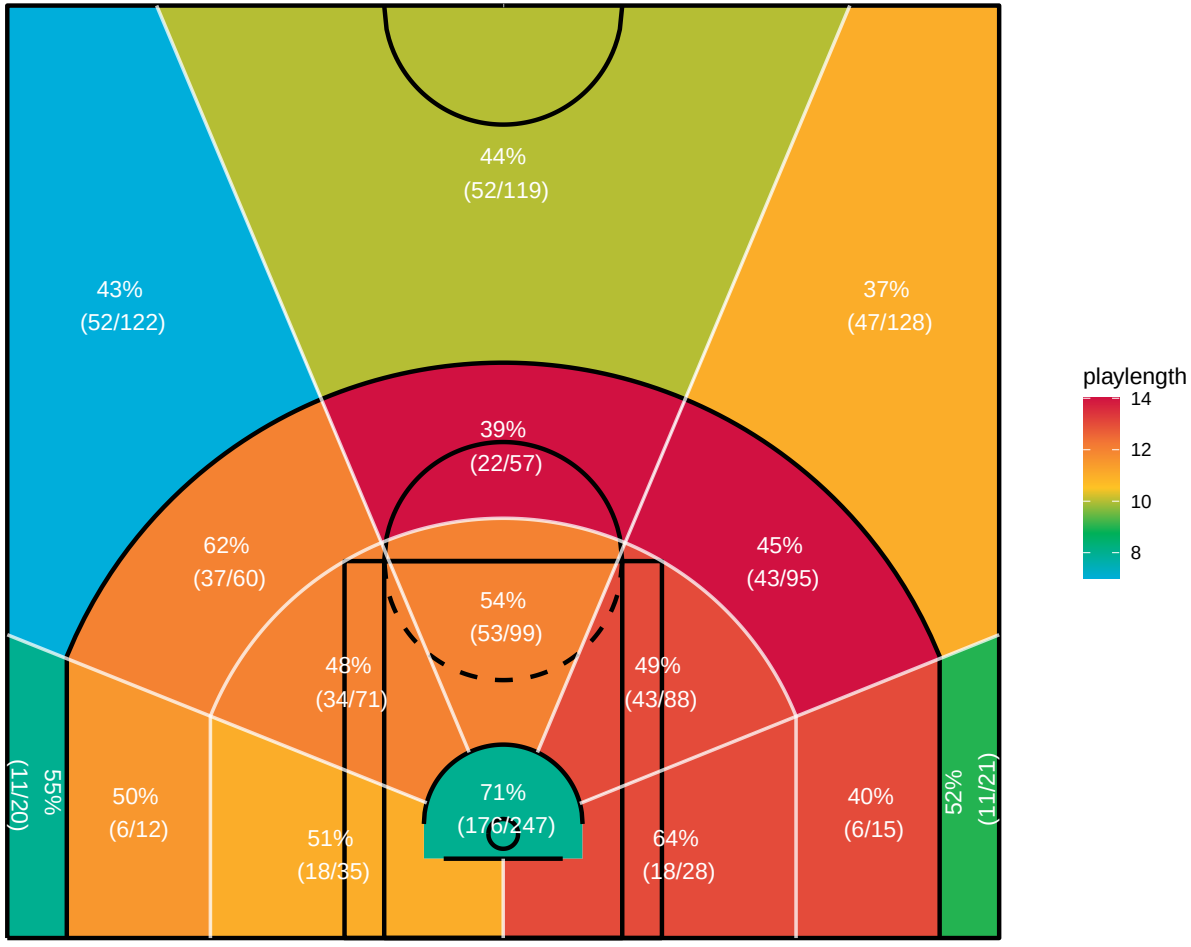
C1: Kevin Durant (Cont.)

Kevin Durant Shot Chart (interesting but not definitive)



CS1: Kevin Durant (Cont.)

Kevin Durant Shot Chart by Play-Length (much more definitive)



CS2: Kyle Lowry

Kyle Lowry (TOR) was a polarizing player in 2017-2018. Some felt he was 'overrated', while other felt he was 'under appreciated'. His traditional counting statistics are not conclusive. Can we use more advanced data analytics to better understand Kyle Lowry's value?

Yr	GP	FG%	3P%	FT%	RPG	APG	PPG
2017-18	78	0.427	0.399	0.854	5.6	6.9	16.2

CS2: Kyle Lowry (Cont.)

Multidimensional Scaling (MDS) is a "nonlinear dimensionality reduction tool that allows [us] to plot a map visualizing the level of similarity of individual cases [within] a dataset".

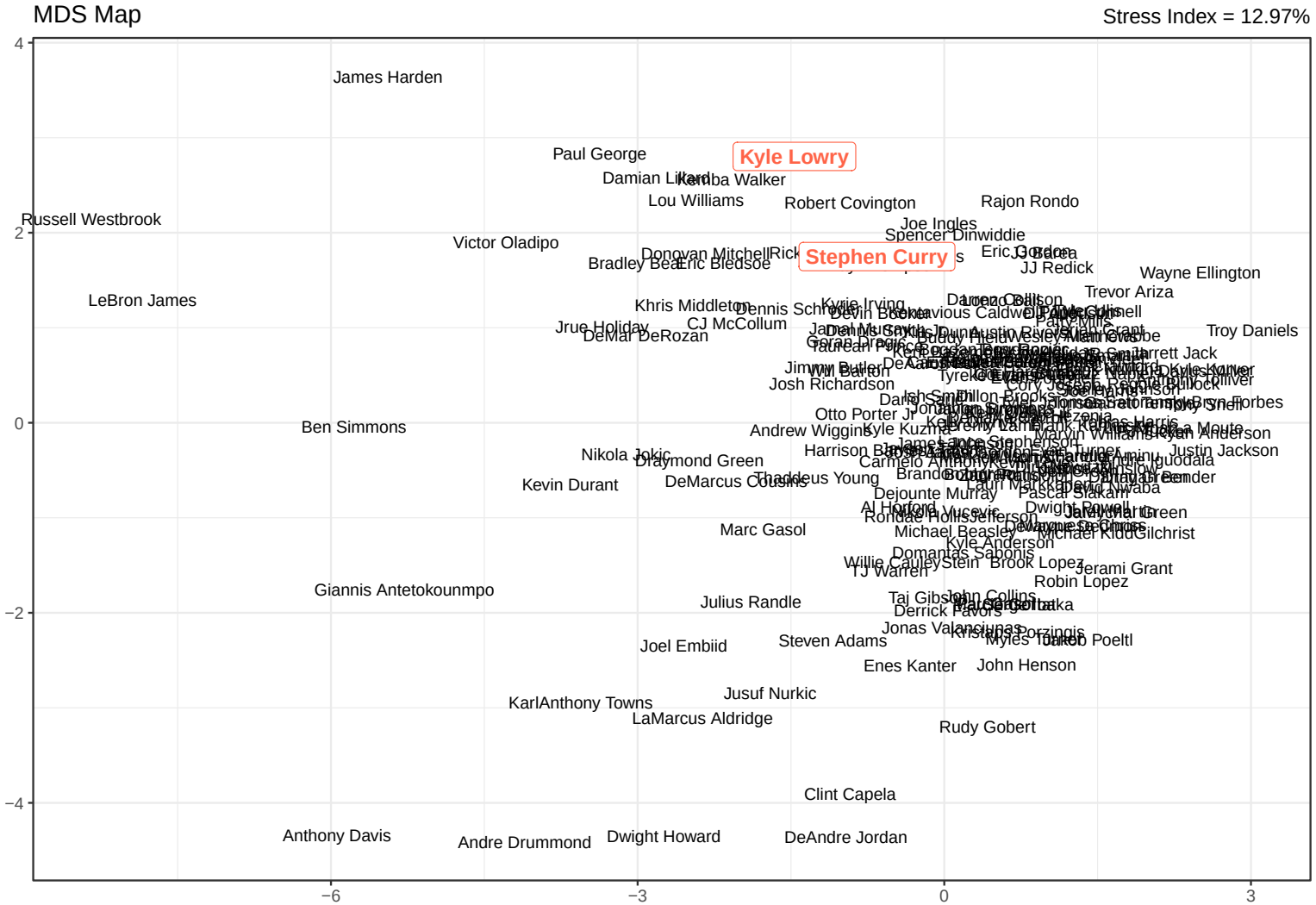
We start with a distance matrix $\mathbf{D}^p = (d_{ij})_{i,j=1,\ldots,N}$ based on all p variables, $X_1, \ldots, X_p$ and attempt to find $q << p$ such that $\mathbf{D}^q$ fits as closely as possible to $\mathbf{D}^p$.

A standard measure of distance is Euclidean distance,

$$d_{ij} = \sqrt{\sum_{h=1}^{p} (x_{ih} - x_{jh})^2}$$

The "Stress Index" (S) allows us to assess how close $\mathbf{D}^q$ approximates $\mathbf{D}^p$; 0.00% is a perfect fit, and we should avoid $S > 20\%$.

CS2: Kyle Lowry (Cont.)



Original variable dimension (8): PTS, P3M, P2M, REB, AST, TOV, STL, BLK reduced to two dimensions. Restricted to players with over 1,500 minutes.

CS3: Golden State Warriors Offense

The Warriors offense was famous for its passing and 'free-flowing' ball movement. Can we use statistics to better understand and assess player roles?

Figure 3: 2017-2018 Golden State Warriors

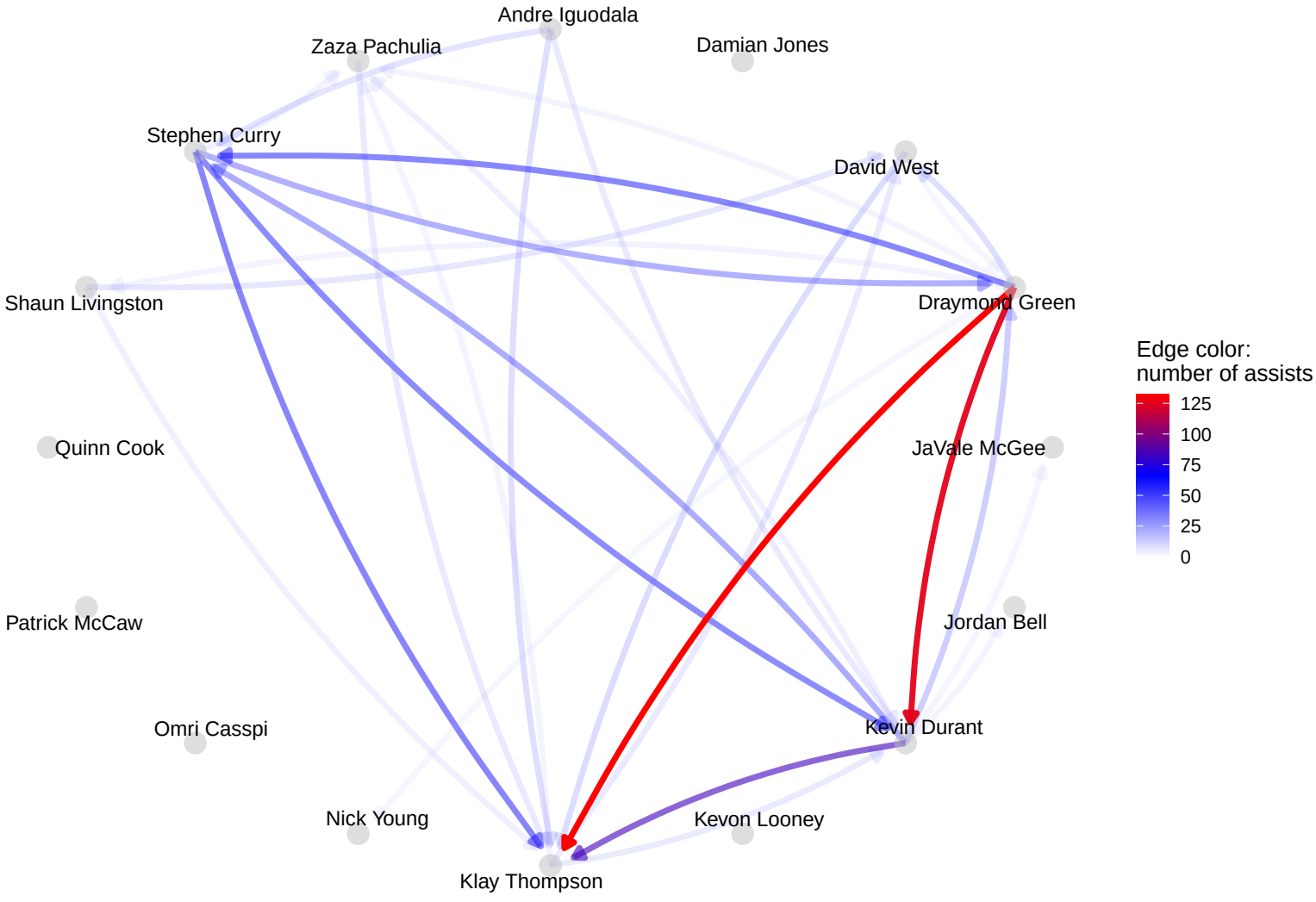
CS3: Golden State Warriors Offense (Cont.)

We can employ *network analysis*, in which we construct and analyzes graphs consisting of nodes related to each other by a set of attributes. This will allow us to find symmetric or asymmetric relationships between discrete objects.

Our nodes/discrete objects will be players, and we will build an assist-network in hopes of better understanding the roles of each player with Golden State's offense.

Note: the underlying data will be "play-by-play" data.

CS3: Golden State Warriors Offense (Cont.)



CS4: Cluster Analysis of NBA teams

The NBA using a weighted lottery system to determine draft selection order. The worse a team's record from the previous season, the higher its odds at receiving a high draft pick. To take advantage of this, some teams have employed a 'tanking strategy', in which a team purposefully employs a weak roster in hopes of getting a high draft pick in the upcoming draft. Can we use statistics to help a team determine its strategy?

Figure 4: "The quickest way to win is to lose." - Sam Hinkie

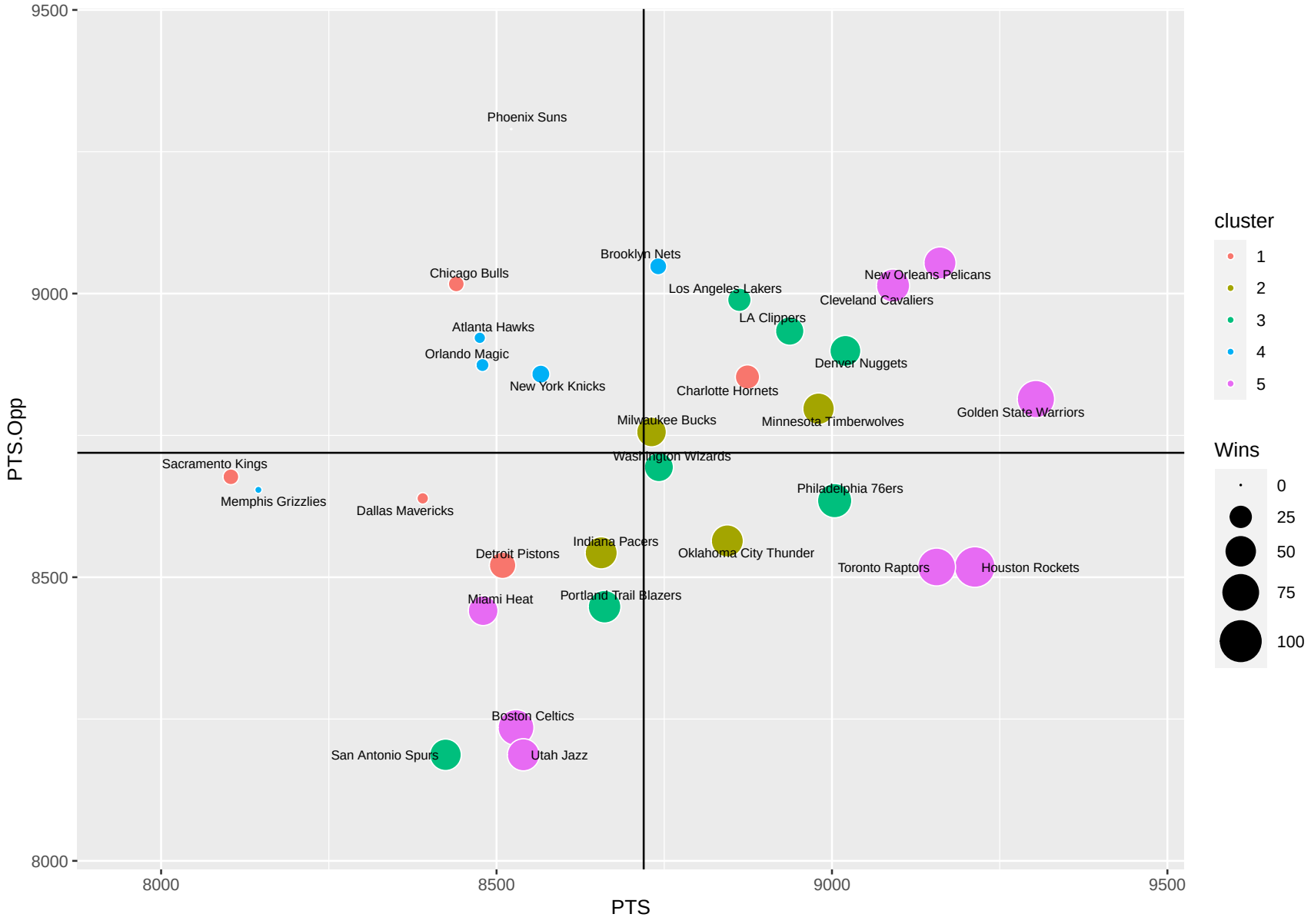
CS4: Cluster Analysis of NBA teams (Cont.)

Cluster Analysis is a classification technique used to divide individual cases into groups (clusters) such that each case within a cluster is "similar" (according to a given criterion) yet "different" from the cases in other clusters. Cluster Analysis is an *unsupervised* classification technique.

Here we employ a specific technique of Cluster Analysis, k-means clustering, to NBA teams based on the "four factors".

[see Ch. 4 of *Basketball Data Science* for details]

CS4: Cluster Analysis of NBA teams (Cont.)



CS5: The Lost Art of the Mid-Range

Figure 5: Stat Nerds Example of a "Bad Shot"

CS5: The Lost Art of the Mid-Range (Cont.)

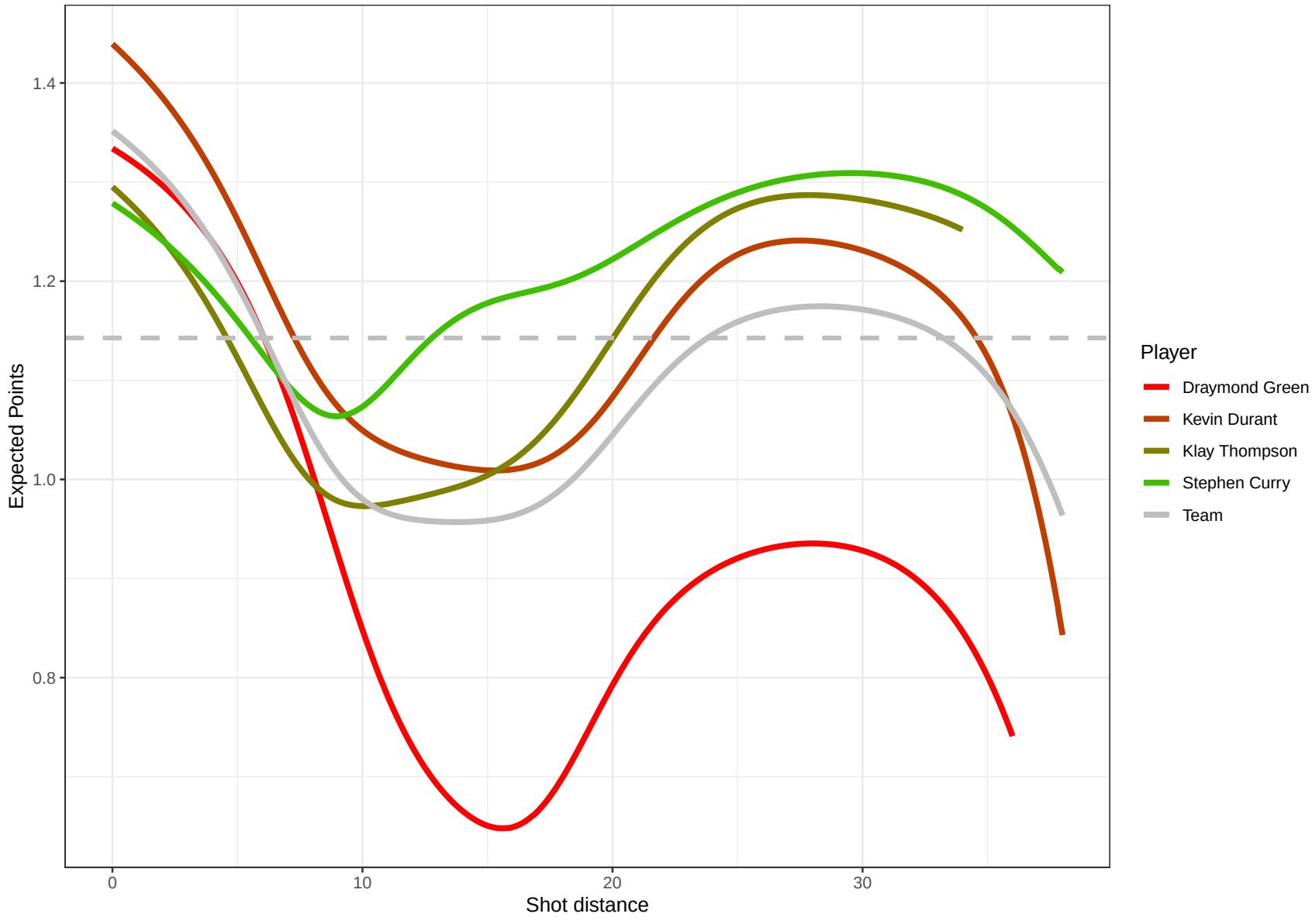
We briefly introduce the concept of *expected value*. Suppose we have a discrete random variable, X over a sample space, $\mathcal{X}$. We may define the expected value, $E(X)$, as

$$E(X) = \sum_{\mathcal{X}} x * P(X = x)$$

For example, if a player shoots 45% on 2-point FG's, his expected value per 2-point shot is

$$P(X = \text{Make}) * 2 + P(X = \text{Miss}) * 0 = (45\%)(2) + (55\%)(0) = 0.9$$

CS5: The Lost Art of the Mid-Range (Cont.)



Summary

- `BasketballAnalyzeR`
- Case studies to review
  - creating shot charts
  - dimension reduction techniques
  - network analysis
  - expected points per shot distance

Further Introductory Reading/Listening

- *The Lowe Post* with Zach Lowe (podcast, ESPN)
- *SprawlBall: A Visual Tour of the New Era of the NBA* by Kirk Goldsberry
- *Basketball Analytics: Objective and Efficient Strategies for Understanding How Teams Win* by Stephen M. Shea, Christopher E. Baker
- *Basketball on Paper: Rules and Tools for Performance Analysis* by Dean Oliver

Acknowledgements

Jeff Van Gundy Image: https://sports.yahoo.com/sources-jeff-van-gundy-coach-u-s-world-cup-qualifying-010607886.html

Kevin Durant Image: https://www.complex.com/tag/kevin-durant

Kyle Lowry Image: https://sportstar.thehindu.com/basketball/kyle-lowry-toronto-raptors-lebron-james-larry-bird-nba-finals-nba-playoffs/article32359075.ece

Golden State Warriors Image: https://www.nytimes.com/2017/01/24/sports/basketball/golden-state-warriors-kevin-durant-steph-curry.html

Michael Jordan Image: https://www.totalprosports.com/2017/01/29/utah-judge-rules-michael-jordan-pushed-off-on-bryon-russell-in-1998-nba-finals/