WIH

Dustin Fontaine, Dylan Fontaine,
Kate Larsen, Kristin Meier,
Eric Pawlakos

Agenda

- Problem Overview
- Key Factors
- Player Clusters
- Salary Analysis

Problem Overview

Key Questions:

- How do we categorize different types of players?
- How do salaries differ by player characteristics?
- Can we identify over and underpaid players?

Methodology

Define Factors   ⟶   Find Clusters

Incorporate Salary Data

Analyze Salary Data across Clusters

Key Factors

The following factors were derived as the main variable groupings from the box score statistics. All statistics, besides games and minutes played, were normalized with per minute of playtime.

Factor	Statistics
Big Man	Rebounds, Blocks, Fouls, FG Percentage
Scoring	Points, Free Throws, Field Goals
Ball Handling	Assists, Steals, Turnovers
Play Time	Minutes Played, Games Played

Player Clusters

Using the derived variable factors, six different types of players were found in the data.

Factors

Clusters	Big Man	Scoring	Ball Handling
Offensive Stars	Average	High	High
Defensive Big Men	High	Low	Average
Offensive Big Men	High	High	Average
Scorers	Average	High	Average
Ball Handlers	Low	Average	High
Non-Producers	Low	Low	Low

Legend: Low | Average | High

Team and Cluster Matchups

Starting lineups can be visualized by cluster instead of position (PG, SG, SF, PF, C).

	Offensive Stars	Defensive Big Men	Offensive Big Men	Scorers	Ball Handlers	Non Producers
Cleveland Cavaliers	LeBron James, Kyrie Irving	Tristan Thompson	Kevin Love			J.R. Smith
Golden State Warriors	Steph Curry	Andrew Bogut	Draymond Green	Klay Thompson, Harrison Barnes		

Different Clusters have Different Salaries?

Scatter Plots appear to show trends in salary.
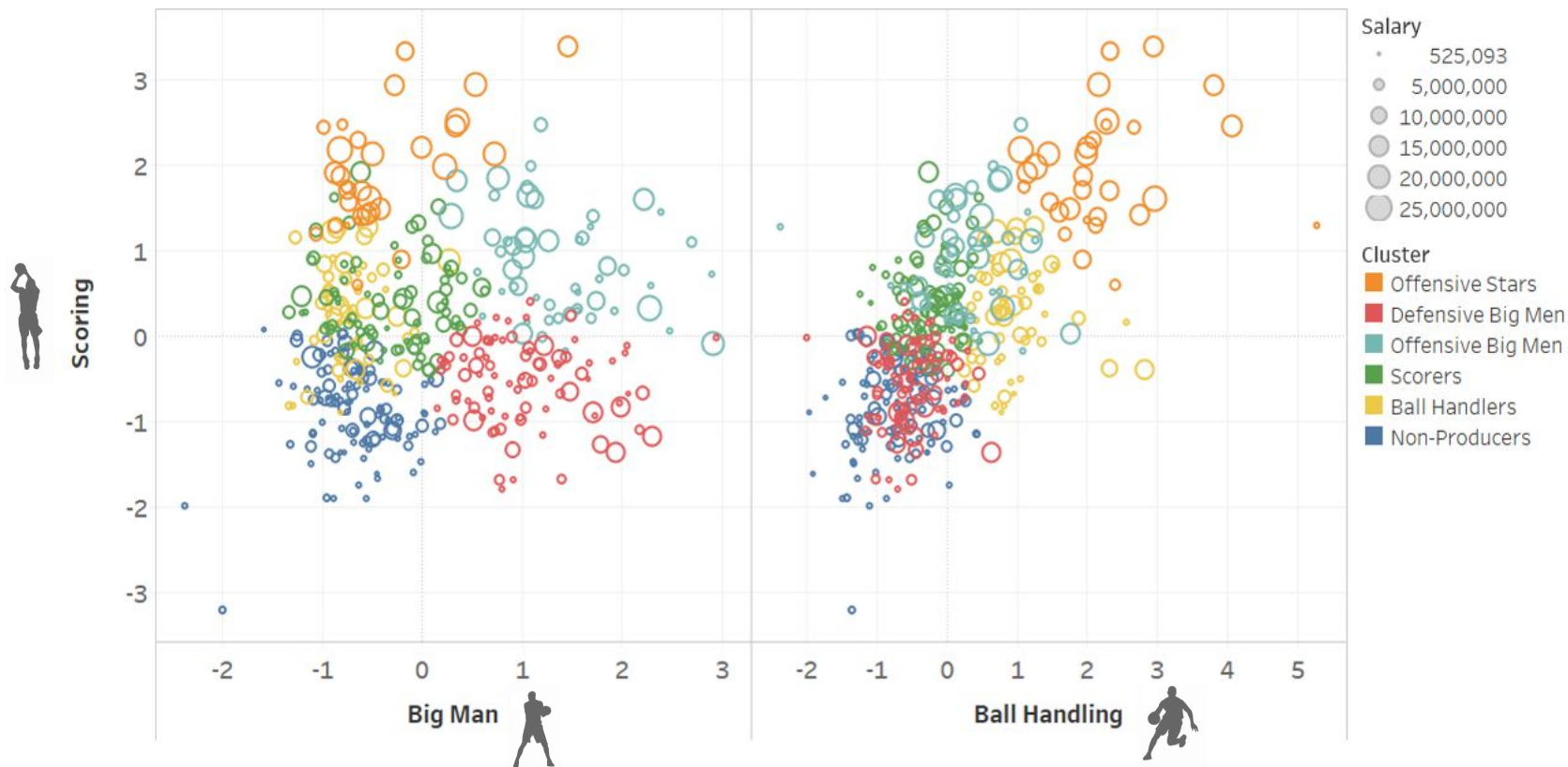
Salary Distribution by Cluster

Tukey's HSD test shows a statistically significant difference in average salary between Non-Producers and Offensive Stars and all other clusters.

	1	2	3	4	5	6
	Non-Producers	Offensive Stars	Defensive Big Men	Offensive Big Men	Scorers	Ball Handlers
	$1.05M	$6.43M	$1.58M	$3.13M	$2.11M	$1.87M

Cluster
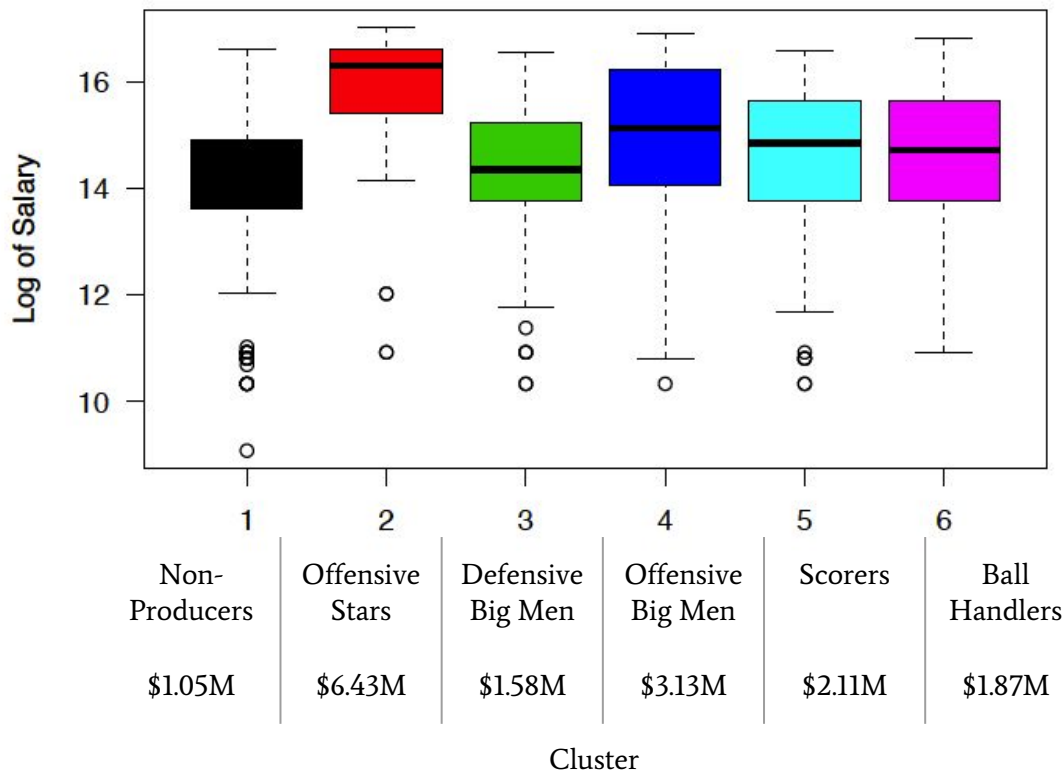
Identifying Overpaid and Underpaid Players

In reality:

   Salary = f(stats, years experience, popularity, …) + Error

Our model:

   Salary = f(stats) + Error

We fit linear regression, neural network, and tree based models to the boxscore data in order to predict salary.

The "boosted tree" model performed best, with a cross-validation $R^2 \cong 0.5$

Overpaid Players

Player	Actual Salary	Predicted Salary	Residual
Kobe Bryant	$25.0M	$8.5M	$16.5M
Dwight Howard	$22.4M	$8.7M	$13.7M
Roy Hibbert	$15.6M	$2.1M	$13.5M

Underpaid Players

| Player	Actual Salary	Predicted Salary	Residual	2016-17 Salary
Damian Lillard	$4.2M	$14.3M	-$10.1M	⬆ $24.3M
Andre Drummond	$3.3M	$12.5M	-$9.2M	⬆ $22.1M
Karl-Anthony Towns	$5.7M	$14.5M	-$8.8M	$6.0M

Appendix

Factor Calculations

- Exploratory Factor Analysis
- Varimax Rotation
- 4 Factors
  - F1: Big Man
  - F2: Scoring
  - F3 Ball handling
  - F4: Play time
- The first three factors (F1, F2, F3) are normalized by minute of play time.

Loading Vectors

	1	2	3	4
GAME_PLAYED				0.91
MINUTES_PLAYED				0.87
FG_per	0.54			
FT_per				0.46
FG_min		0.87		
FGatt_min		0.87		
FT_min		0.71	0.43	
FTatt_min	0.37	0.66	0.41	
P_min		0.95		
OR_min	0.83			
DR_min	0.74			
R_min	0.86			
AST_min			0.75	
PF_min	0.46			
S_min			0.57	
TO_min			0.57	
B_min	0.7			
AT_ratio				0.38

Cluster Analysis

- K-means
- 6 clusters from 3 main factors
  - F1: Big Man
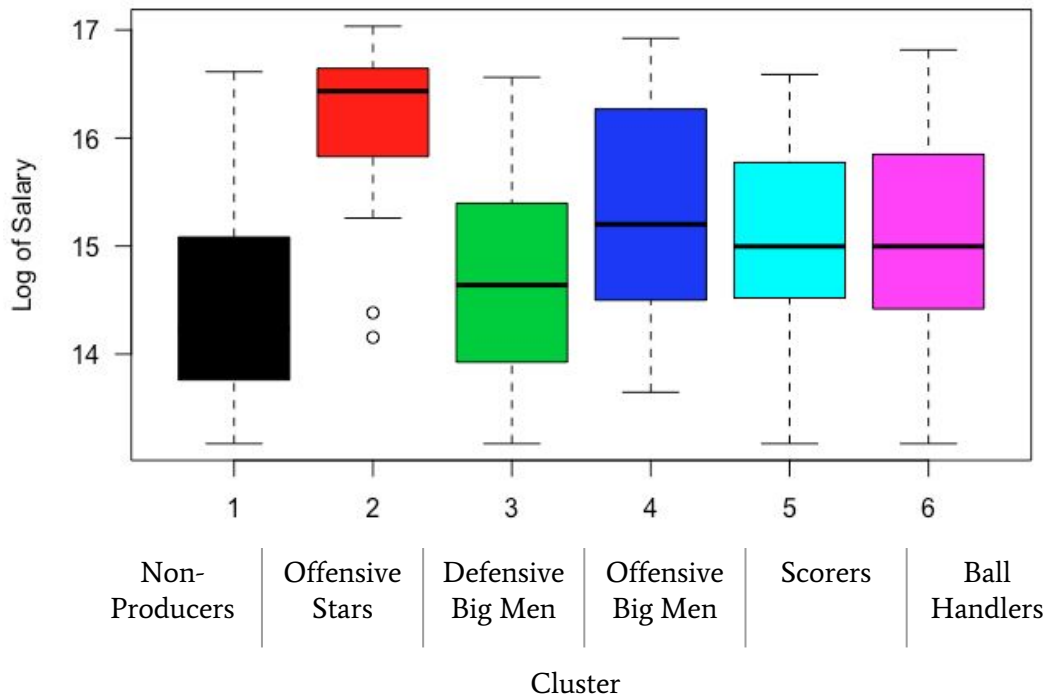 - F2: Scoring
  - F3 Ball handling
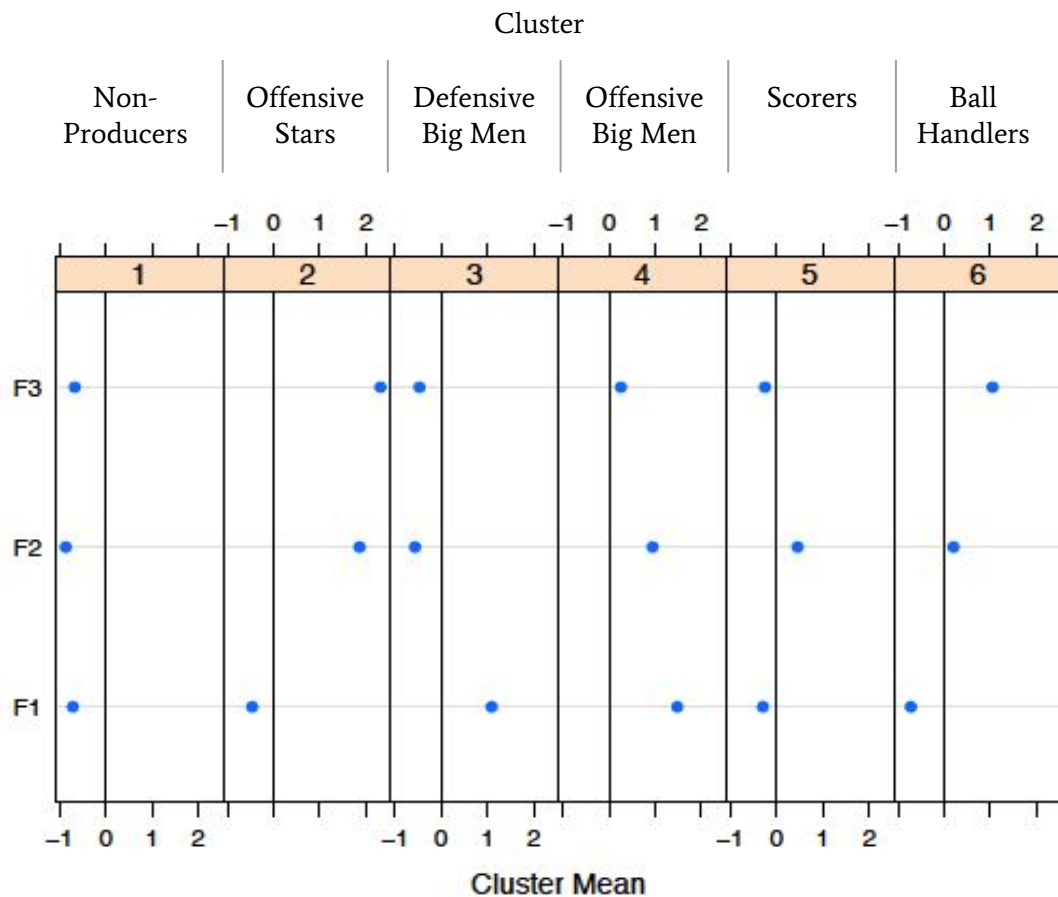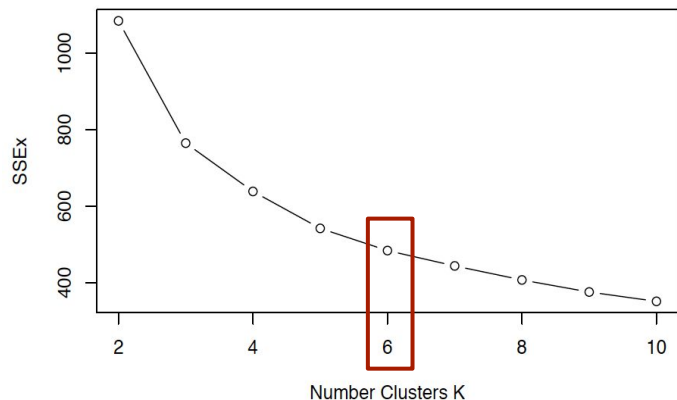
Salary Distribution by Cluster

This plot excludes players with salary below the league minimum.
Cluster 2 outliers: Dennis Schroder and Jordan Adams
- Both players are in our top 30 list of most underpaid players

Regression

- Methods: Linear regression and random forests
- CV R^2 is around .5 using only player stats and cluster assignments
- Including additional player data such as draft year and draft number, CV R^2 increases to .8 for random forest methods
- Variables with largest influence on salary (full model):
  - Minutes Played
  - Games Played
  - Draft Number
  - Draft Year
  - Points per minute
  - Blocks per minute
- Variables with largest influence on salary (individual model for each cluster):
  - Non-Producers: Plus_Minus, Games Played
  - Offensive Stars: Plus_Minus, FG Made, FG Att, FT Made, FT Att
  - Defensive Big Men: Plus_Minus, Games Played, Steals, Turnovers, Assists, Rebound
  - Offensive Big Men: Plus_Minus, Blocks, FG Att, Disqualifications
  - Scorers: Plus_Minus, Games Played
  - Ball Handlers: Plus_Minus, Assists, F3, Games Played