Capstone 1: Predicting the Salaries of NBA players based on their stats

By Alvaro Hu

The Problem

- What makes a player more valuable than another?
  - Are “overpaid” players really overpaid?

- How much money should a player be making depending on the stats they produced the previous season?
  - Are certain players worth the money? Is the team getting their best ‘bang for their buck?’

Why is this useful?

- Team General Managers (GM)
  - How much is a certain player worth?
  - Can we afford a player of this caliber, and If not, what is the best deal we can make
- Players
  - Is there anything specific I can do to earn more money?
  - What change in my game will result in a definite increase in pay?

The Data

- Three datasets from basketball-reference.com
  - 2018-2019 Salaries for each player

  - 2017-2018 Common Statistics
    - Per game stats, field goal percentages, etc.

  - 2017-2018 Advanced Statistics
    - Win Shares, Player Efficiency Rating, Rebound Percentage, etc.
- Merged these together into one

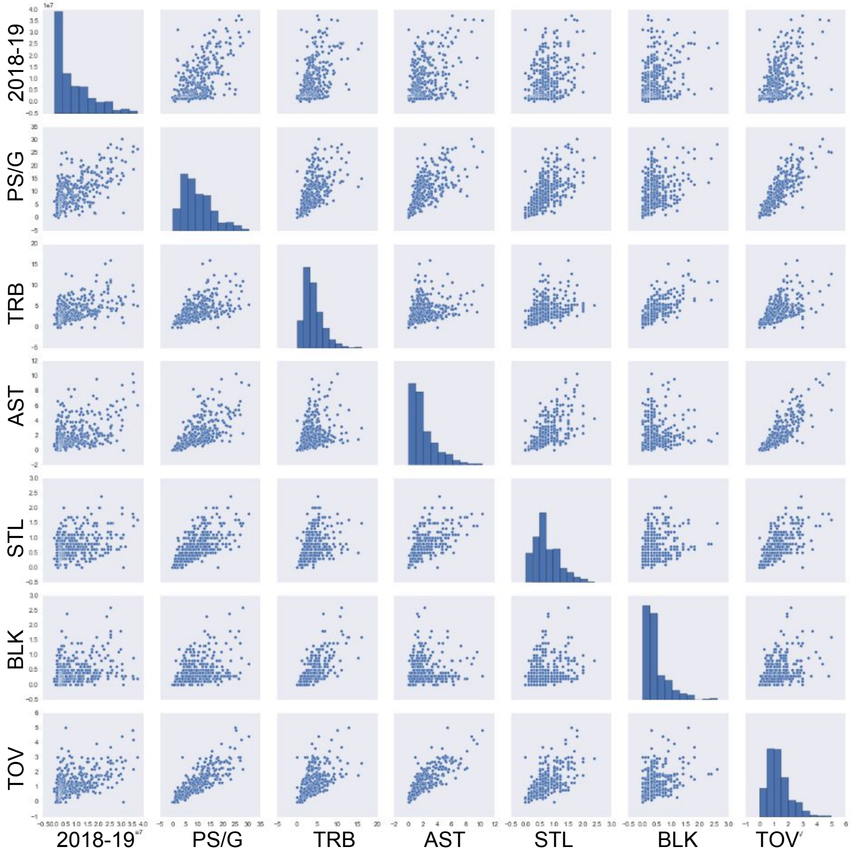
Final DataFrame

Tm	2018-19	Name	Pos	G	GS	FG	FG%	3P	3P%	eFG%	FT%	ORB	TRB	AST	STL	BLK	PS/G	TOV	Age	WS	PER	MP	TRB%	AST%	TOV%	OWS	DWS
GSW	37457154	Stephen Curry	PG	51	51	8.4	0.495	4.2	0.423	0.618	0.921	0.7	5.1	6.1	1.6	0.2	26.4	3.0	29	9.1	28.2	1631	9.0	30.3	13.3	7.2	1.8
HOU	35654150	Chris Paul	PG	58	58	6.3	0.460	2.5	0.380	0.550	0.919	0.6	5.4	7.9	1.7	0.2	18.6	2.2	32	10.2	24.4	1847	9.5	40.9	12.5	7.5	2.7
LAL	35654150	LeBron James	PF	82	82	10.5	0.542	1.8	0.367	0.590	0.731	1.2	8.6	9.1	1.4	0.9	27.5	4.2	33	14.0	28.6	3026	13.1	44.4	16.1	11.0	3.0
OKC	35350000	Russell Westbrook	PG	80	80	9.5	0.449	1.2	0.298	0.477	0.737	1.9	10.1	10.3	1.8	0.3	25.4	4.8	29	10.1	24.7	2914	15.3	49.8	16.4	5.5	4.5
DET	31873932	Blake Griffin	PF	58	58	7.5	0.438	1.9	0.345	0.493	0.785	1.3	7.4	5.8	0.7	0.3	21.4	2.8	28	4.9	19.6	1970	12.0	28.1	12.6	3.2	1.8

*Some features were selected because of supposed importance, but others were chosen out of curiosity to see their impact on salary

Exploring the Data

- Pairplot of the per-game stat categories
- Positive Correlations across the board
- Non-normal distributions

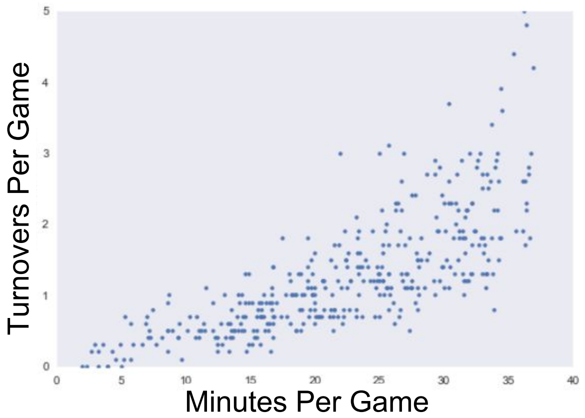


Takeaways From EDA

- Strong Correlation Coefficients with Salary
  - Points per game, games started, minutes per game, rebounds, and…. **Turnovers**?

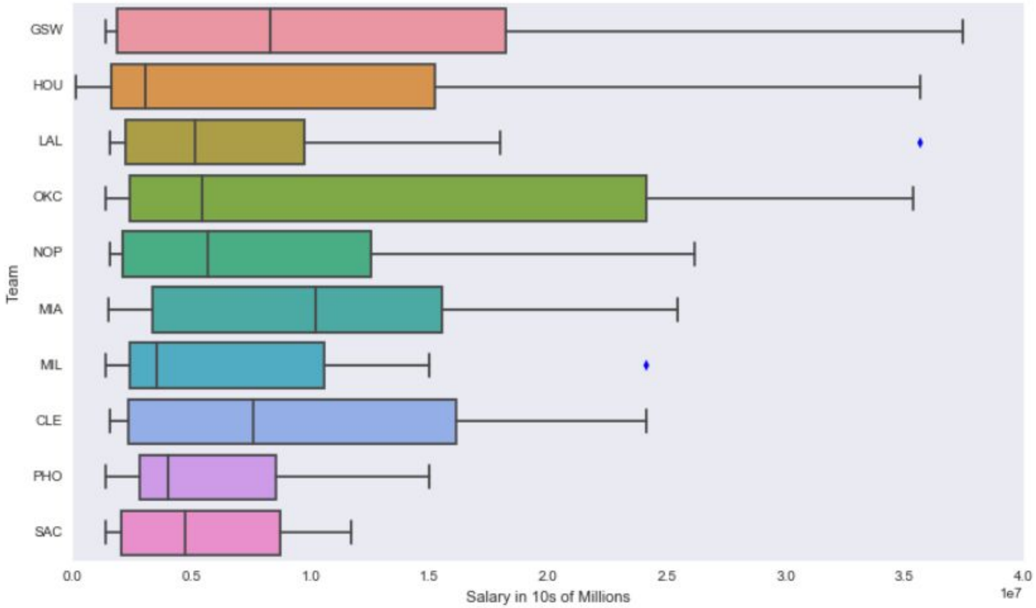
- Turnovers with 2nd strongest correlation of the per-game statistics
  - Quick analysis of minutes vs. turnover reveals

  - Draw back from intercorrelation of variables
    - Question of correlation vs. causation



Takeaways (cont.)

- Different teams seem to be making different amounts of money.
- “Soft” salary cap enforced



Takeaways (cont.)

- Although some positions have higher max salaries, they all seem to have relatively the same median salaries
- Many outliers in the PG position
  - PG are the “face” of the team
  - Salary dependent on teams performance?

Inferential Statistics

- How reliable are these strong correlations?
  - Bootstrapping while shuffling 10% of the data
    - $H_0 : \rho = \rho_{\text{From Data}}$
    - $H_a : \rho \neq \rho_{\text{From Data}}$
  - Tells me with a 95% confidence interval the following:
    - $\rho_{PS/G} = 0.637$, $\rho_{WS} = 0.591$, $\rho_{GS} = 0.569$, $\rho_{M/G} = 0.583$
    - These all fall in the category of "moderate positive correlations"
  - Rechecked with Mann-Whitney U test for Non-Normal distributions

Are all teams paid differently?

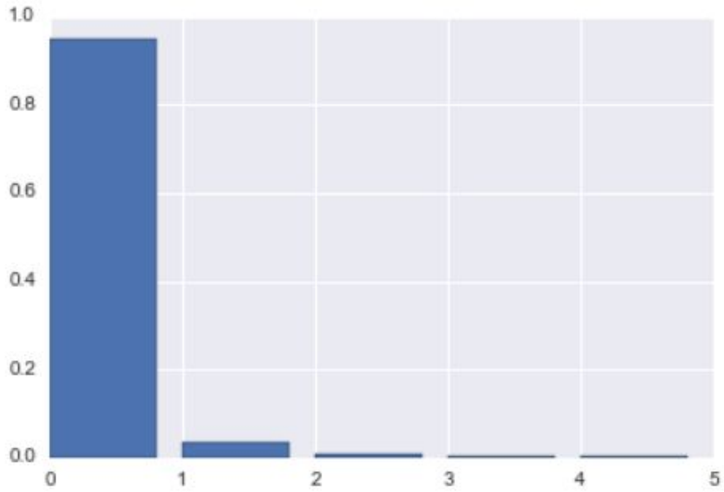
- Bootstrapping technique to test the difference in means
  - Shuffle salaries from two teams and redistribute to teams
  - Check the mean difference and repeat many times
    - $H_o : \mu_{\text{Team 1}} = \mu_{\text{Team 2}}$
    - $H_a : \mu_{\text{Team 1}} > \mu_{\text{Team 2}}$
- Results
  - No difference between GSW and OKC
  - Significant difference between GSW and SAC

Are positions paid differently?

- Ran the same bootstrapping test but with positions
  - $H_0 : \mu_{\text{Position 1}} = \mu_{\text{Position 2}}$
  - $H_a : \mu_{\text{Position 1}} \neq \mu_{\text{Position 2}}$
- Result
  - There is no difference in mean pay between positions.

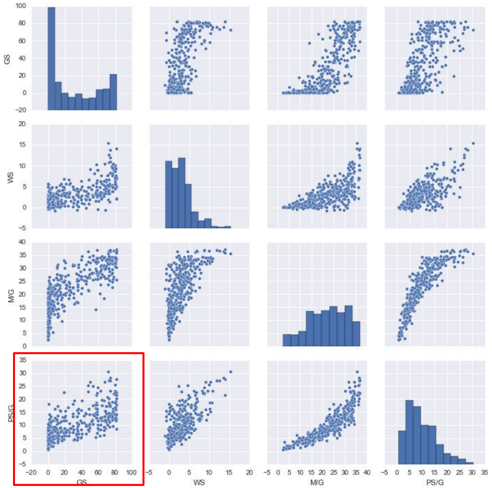
Principal Component Analysis

- Using Games Started, Win Shares, Points per Game, Minutes per Game, and Rebounds per Game
    - Only 1 Principal Component
    - Could be caused by the intercorrelation of the features



Ordinary Least Squares (OLS) Regression

- Run twice
  - Choosing least correlated features with correlations with salary
    - Points per game and Games Started
  - Results
    - R-square: 0.434
    - F-statistic: 143.5
    - P-value: 0.000 for Both



2nd OLS Regression

- Started using entire dataset and removed one at a time
- Ended with Games Played, Assists per Game, Points per Game, Win Shares, and Assist percentage
- Results
  - R-squared: 0.504 (slightly better)
  - F-Statistic: 75.38 (not as good of model)
  -