Capstone 1: NBA Player Value Based Upon On-Court Performance

Exploratory and Machine Learning Techniques to Predict NBA Player Salary

Introduction

- NBA Players get contracts in varying amounts and timeframes
- How are those contract details determined?
  - Is it based upon their past and predicted stats?
- Win Share statistic a tell all?
  - This is a computed number based upon a players stats.
- Correlations between contracts and statistics
- Data obtained, wrangled, explored and used to construct linear regression models
  - Estimates of player salaries based upon on-court production

Data Wrangling

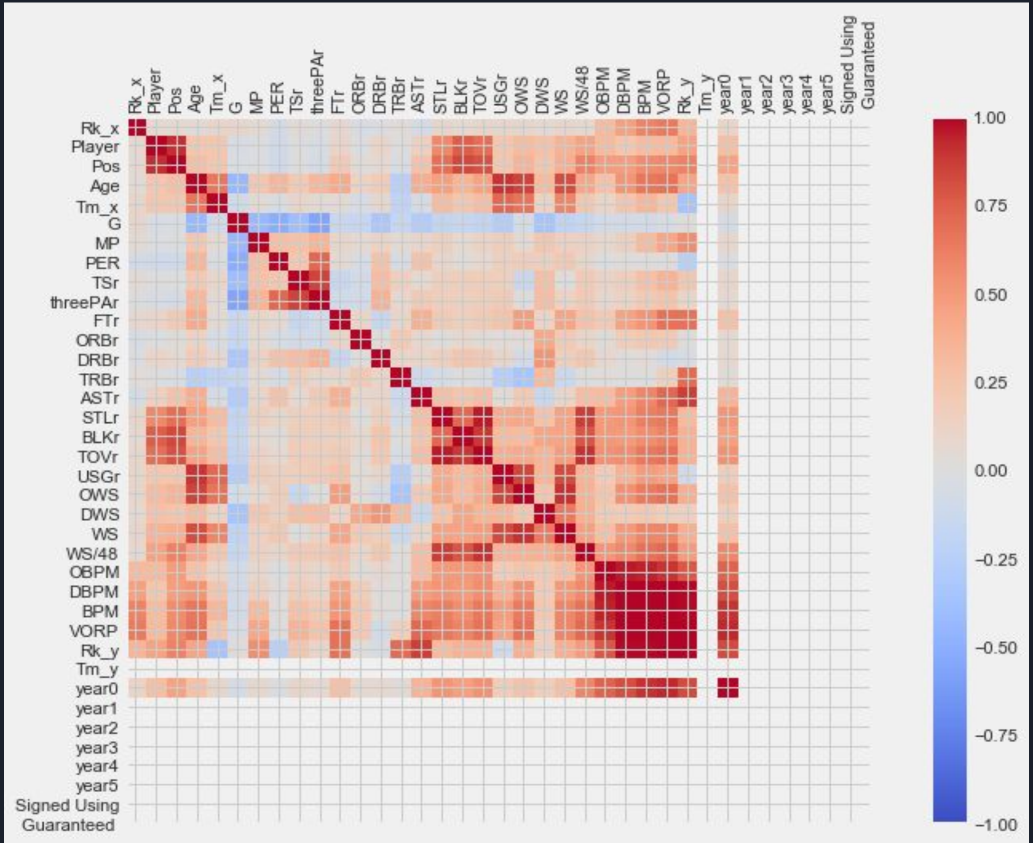
- Obtained data from Basketball-Reference.com
  - Reliable, up to date information on contracts and player stats.
- Merged together two dataframes
  - One with contract data and one with statistics
- 22 Features
- Converted data types from object to correct format
  - Float, integer, string

Features

Rk_x	object
Player	object
Pos	object
Age	float64
Tm_x	object
G	float64
MP	float64
PER	float64
TS%	float64
3PAr	float64
FTr	float64
ORB%	float64
DRB%	float64
TRB%	float64
AST%	float64
STL%	float64
BLK%	float64
TOV%	float64
USG%	float64

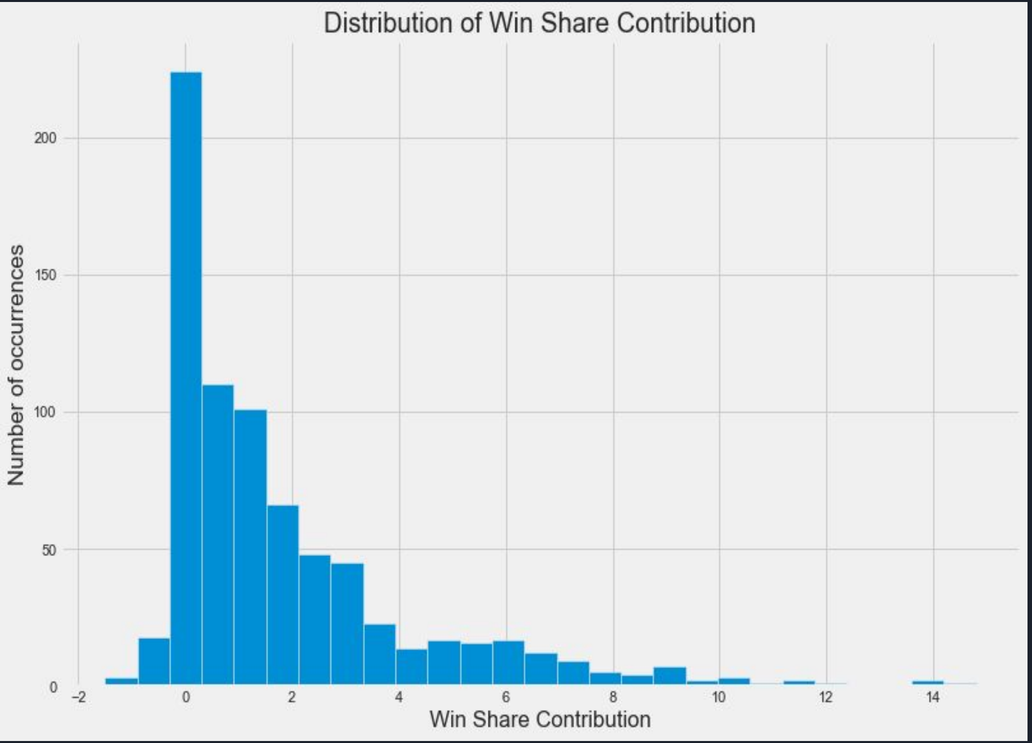
OWS	float64
DWS	float64
WS	float64
WS/48	float64
OBPM	float64
DBPM	float64
BPM	float64
VORP	float64
Rk_y	object
Tm_y	object
year0	float64
year1	float64
year2	float64
year3	float64
year4	float64
year5	float64
Signed Using	object
Guaranteed	float64

Feature Correlations



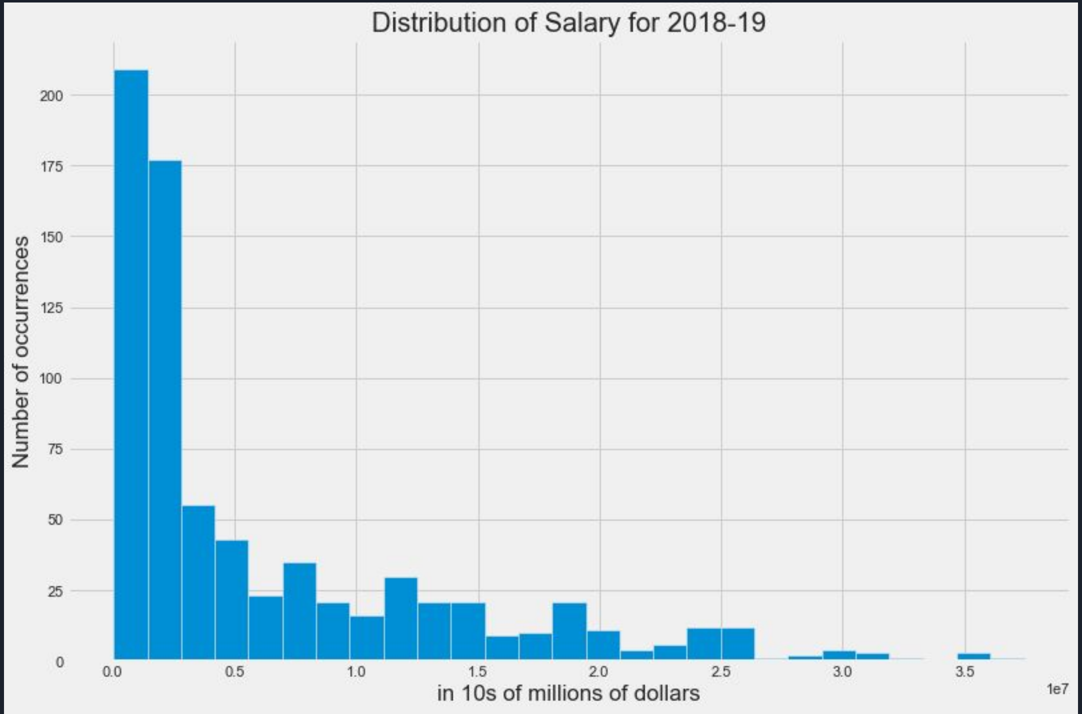
- This shows correlations between our features.
- Red is a positive correlation
- Blue is a negative correlation
- The darker the color, the greater the correlation.

Win Share Distribution



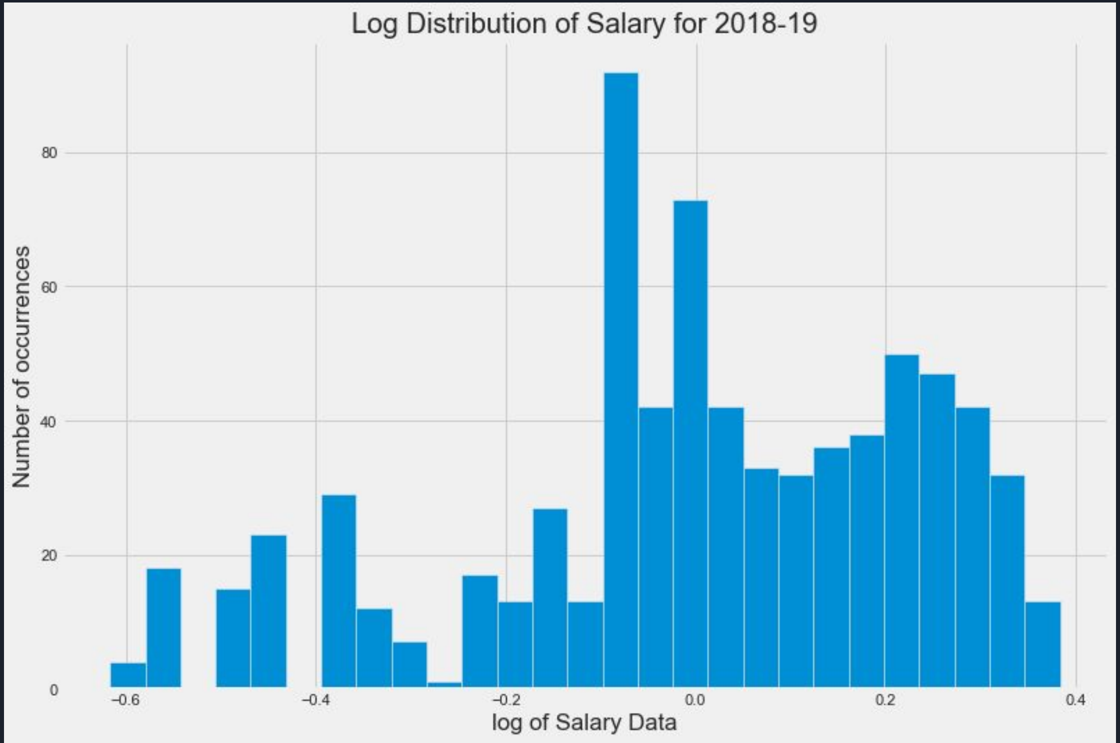
- This shows the distribution of players based upon their Win Share statistic.
- We can see that this is skewed and many players are contributing close to 0 or to just a small portion of the teams success.

Salary for 2018-19 Distribution



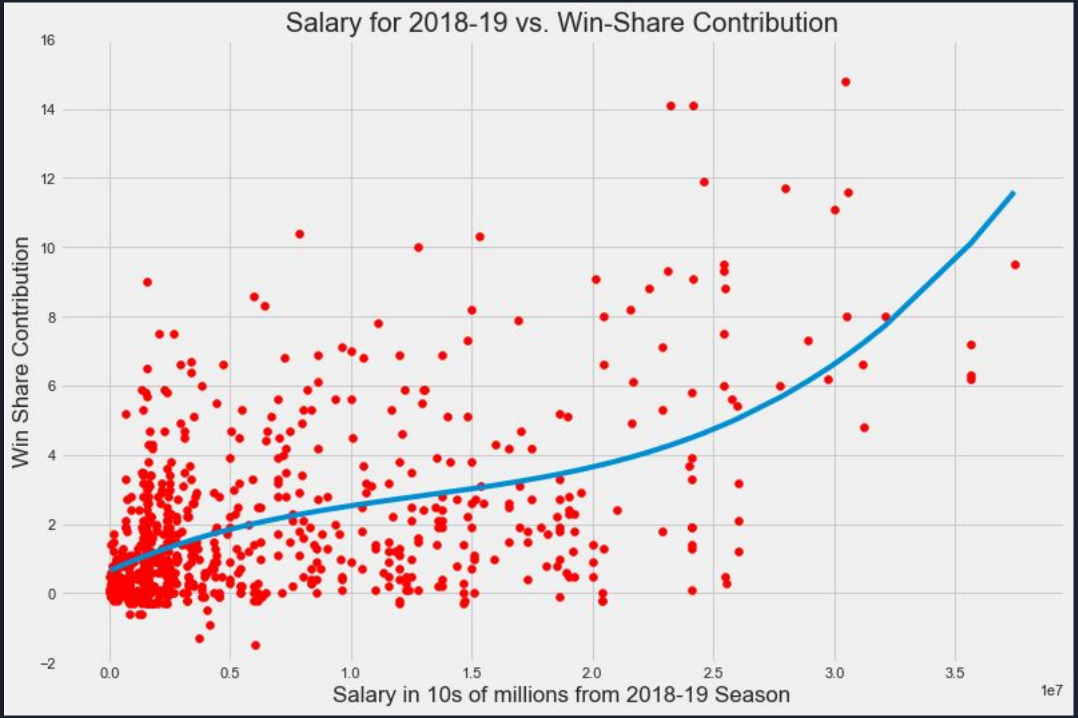
- This is a distribution of players' salaries for 2018-19 season.
- This is also skewed and shows that many players are receiving salaries less than $2.5 million per year.

Log Histogram of Salary Data



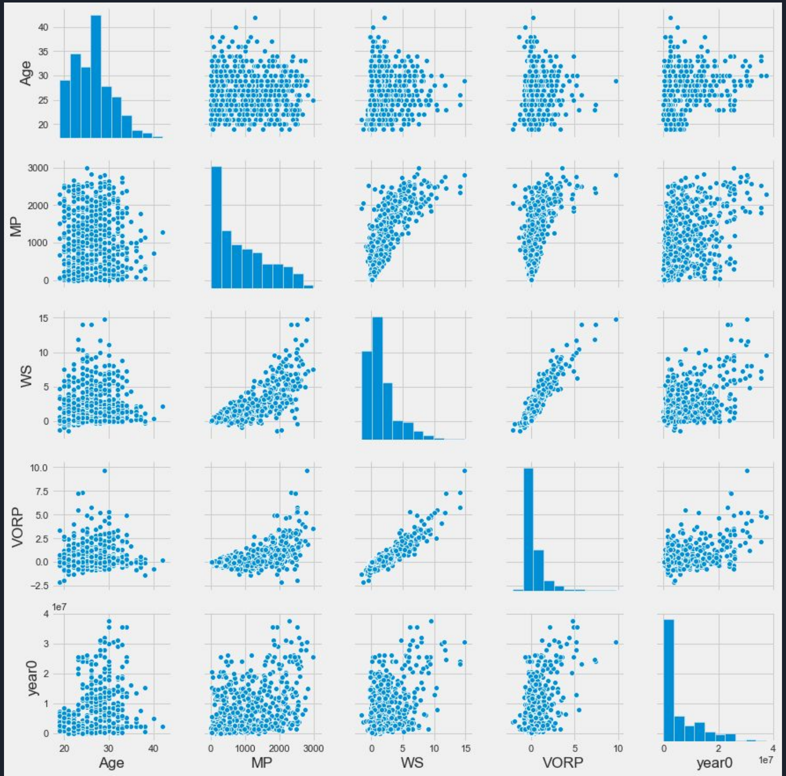
- This shows the salary data from the previous slide after taking the log of the salaries.
- You can see that this helped to normalize the data slightly to show relative change vs. absolute change in the amounts.

Salary and Win Share



- This shows the correlation between salary and Win Share.
- We can see, visually, that there is a correlation between these two features.

Paired Plot of Most Relevant Features



- After looking closely at our feature correlation plot, these were the features that had the greatest correlation to 2018-19 salary.
- This plot gives us more insight into the correlation of these features with salary as well as with one another.

Statistical Inference

- Salary correlation to multiple features, including Win Share, is significant.
- Win Share is correlated to Salary, and a Pearson's Correlation of: 0 .53543
- 8 features are statistically significant with p-values less than .05
  - Age,
  - Games Played (G),
  - Minutes Played (MP),
  - Usage Rate (USGr),
  - Offensive Box +/- (OBPM),
  - Defensive Box +/- (DBPM),
  - Box +/- (BPM), and
  - Value Over Replacement Player (VORP)

	p-value
Age	0.000
G	0.000
MP	0.000
USGr	0.000
OBPM	0.003
DBPM	0.003
BPM	0.003
VORP	0.000

Analysis

- The difference in models using all features (22) vs. our best model (8) is below:
- This reduction in features helps make the model more accurate. The R2 remained relatively similar while we increased the F stat and slightly decreased the AIC.
- This means that we are accounting for roughly the same amount of variance in salary as all the features did with just a fraction of the features while simultaneously increasing the confidence that these variables are related to the target.

OLS Regression Results

Dep. Variable:	year0	**R-squared:**	0.480
Model:	OLS	**Adj. R-squared:**	0.465
Method:	Least Squares	**F-statistic:**	32.92
Date:	Tue, 09 Jul 2019	**Prob (F-statistic):**	6.32e-96
Time:	10:23:27	**Log-Likelihood:**	-13675.
No. Observations:	809	**AIC:**	2.740e+04
Df Residuals:	786	**BIC:**	2.750e+04
Df Model:	22		

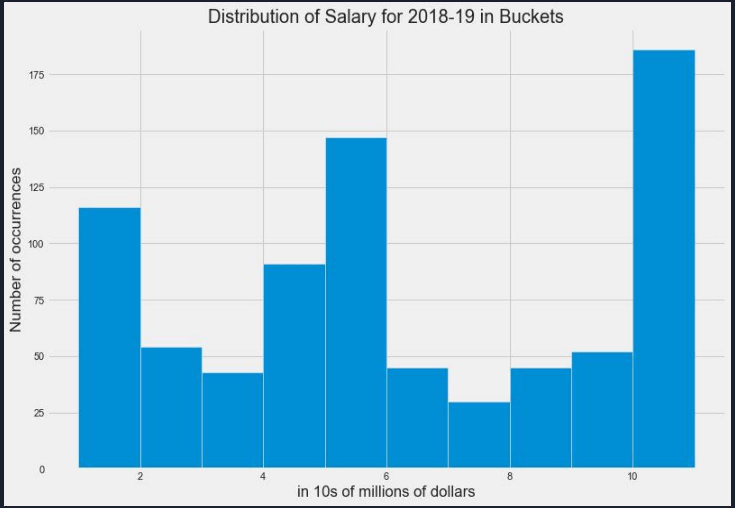
OLS Regression Results

Dep. Variable:	year0	**R-squared:**	0.465
Model:	OLS	**Adj. R-squared:**	0.459
Method:	Least Squares	**F-statistic:**	86.85
Date:	Tue, 09 Jul 2019	**Prob (F-statistic):**	2.74e-103
Time:	10:23:27	**Log-Likelihood:**	-13686.
No. Observations:	809	**AIC:**	2.739e+04
Df Residuals:	800	**BIC:**	2.743e+04
Df Model:	8		

Analysis (cont.)

- It is clear that our current model for prediction isn't very accurate in predicting exact player salaries using player statistics.
- What if we created buckets of specified ranges to estimate instead of absolute values?

- Bucket Ranges (USD):
- 1: <= 500,000
- 2: 500,001 - 1,000,000
- 3: 1,000,001 - 1,500,000
- 4: 1,500,001 - 2,000,000
- 5: 2,000,001 - 3,000,000
- 6: 3,000,001 - 4,000,000
- 7: 4,000,001 - 5,000,000
- 8: 5,000,001 - 7,000,000
- 9 - 7,000,001 - 10,000,000
- 10: 10,000,001 - 15,000,000
- 11: > 15,000,001



Results From Bucket Attempt

- We can see that this model was less accurate than the model built for exact values given that our R^2 value is lower than before and our F-stat is as well.

- What if we use a more evenly distributed bucketing system instead of pre-defined ranges?

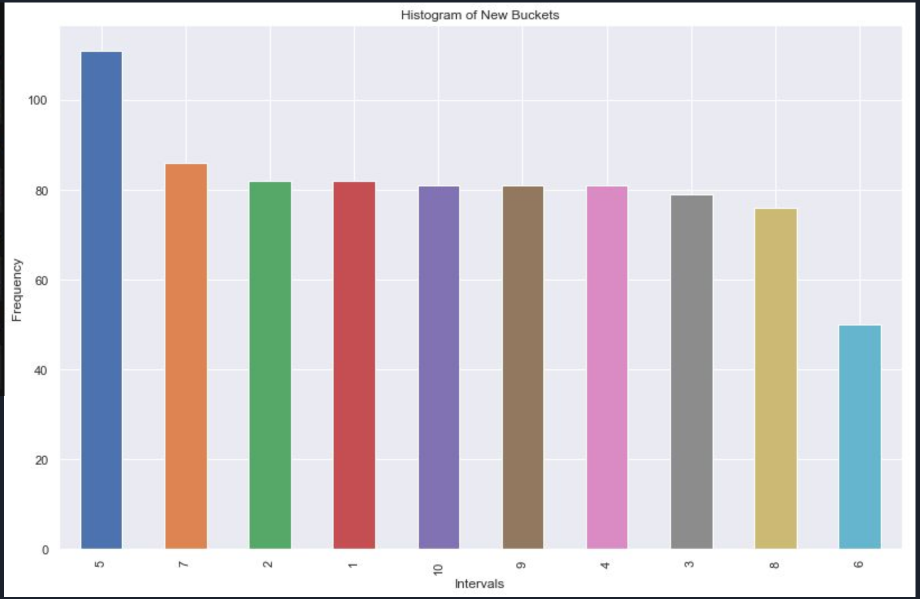
OLS Regression Results

Dep. Variable:	bucket	**R-squared:**	0.414
Model:	OLS	**Adj. R-squared:**	0.408
Method:	Least Squares	**F-statistic:**	70.52
Date:	Tue, 09 Jul 2019	**Prob (F-statistic):**	1.51e-87
Time:	10:23:28	**Log-Likelihood:**	-1916.8
No. Observations:	809	**AIC:**	3852.
Df Residuals:	800	**BIC:**	3894.
Df Model:	8		

Quantile Cut Results

OLS Regression Results

Dep. Variable:	bucket_even	**R-squared:**	0.405
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	68.09
Date:	Tue, 09 Jul 2019	**Prob (F-statistic):**	4.43e-85
Time:	10:39:30	**Log-Likelihood:**	-1791.8
No. Observations:	809	**AIC:**	3602.
Df Residuals:	800	**BIC:**	3644.
Df Model:	8		



- A more even distribution of the data, but a less accurate model than the previous binning attempt.

Linear Models

Model Name and Target	Number of Features	R^2 Value	F-Stat Value
m, Salary	1 (Win Shares)	.287	301.0
m_all, Salary	22	.480	32.92
m_sel, Salary	10	.468	70.13
m_1, Salary	8	.465	86.85
m_new, bucket (first)	22	.432	27.20
m_new1, bucket (first)	8	.414	70.52
m1, bucket (second)	22	.422	26.07
m2, bucket (second)	8	.405	68.09

- This is a breakdown of the different models that were created and their primary details.
- We can see that the model that was used for absolute prediction was more accurate than any of our binning models.
- Even our best model accounted for just ~46.5% of the variance in Salary.

Conclusion

- Our best model resulted in an R^2 value of .465.
  - This is relatively poor, but provided additional insight.
- Given that the use of player statistics for predicting a player's salary results in such a low accuracy, it is clear that a player's value comes from much more than just on-court production.
- Organizations take into consideration the value a player brings to them on the court as well as off the court. Meaning, a player that is more recognized or popular may be given a greater salary because of the added benefit to the organization through the potential of increased ticket sales, merchandise, apparel and value to the organization's name itself.