Capstone 1 - Understanding the Relationship Between NBA Player Salary and Statistical Performance

•••

Statistical Analysis and Machine Learning

Statement of Purpose

The purpose of this project is to understand the relationship between NBA player performance and salary.

There are two potential groups of clients for this problem:

- The first, and primary, clients are NBA players themselves who can use the information to not only determine when their career performances may peak, but also to determine when to pursue maximum contracts during the course of their careers.

- The second group that can find this analysis useful would be NBA owners and teams who may be able to use it to determine when to best invest in a player based on their career performances in particular metrics.

Data Wrangling Process

The data comes from two different sources:

- The first is salary data between 1991 and 2017. It lists the yearly salary of every player in the league except for those players who were signed late in the season, cut early, or only on 10-day contracts.

- The second set of data is from Kaggle and lists statistics of every player between 1950 and 2017. The statistics include yearly totals as well as career totals in over 40 categories. It also includes when players' careers began and ended

That makes a total of three datasets - one for salary data, one for player statistics, and the third for career length. Our goal is to combine all three into one dataset.

Data Wrangling cont.

Several steps need to be taken to clean and combine the three datasets.

The first major step is to combine the salary and career datasets:

- The career dataset is cropped to include only players who began their careers after 1991. It is also cleared of unnecessary columns, duplicate data, and null values, and errors.
- The salary dataset contains several missing values, and because it would be difficult (and time consuming) to track down those values, it's easier to just drop the respective players from the dataset. Doing so shouldn't influence the analysis heavily.
- The salary and career datasets are then interjoined on the 'Player' column. Additional duplicate rows are dropped and the index is reset.

After the stats dataset is cleared of errors, it is inner joined with the career/salary dataset on the 'Player' and 'Year' columns:

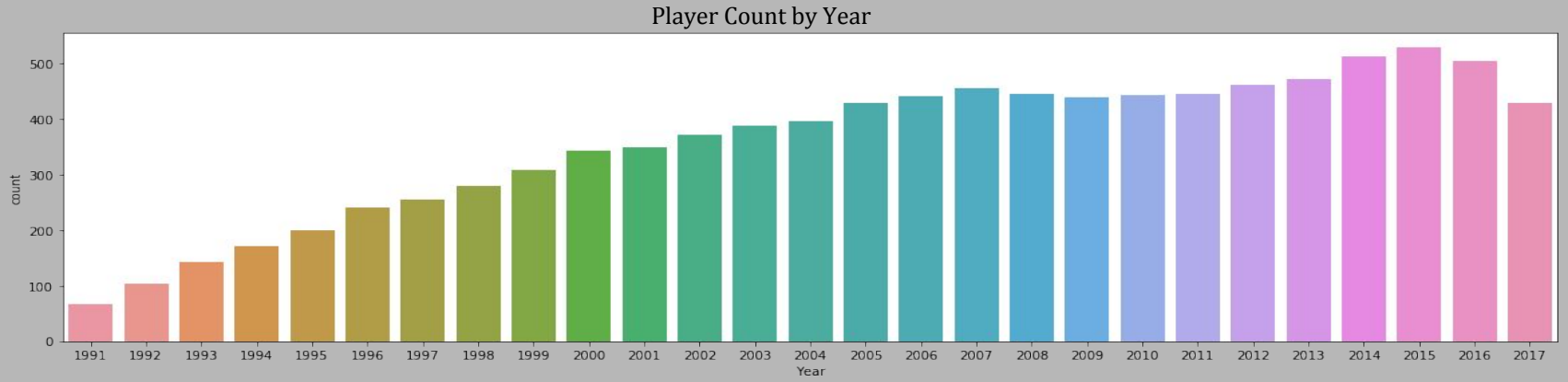
- Additional nulls are discovered in our final dataset, but they are due to division by zero, so they're simply replaced with 0.0 floats. Remaining duplicate rows are dropped, and our dataset is now ready for analysis.

Disclaimer

Before we move on, we need to make a quick note about bias in the dataset that could influence some of our results

Since we were originally interested in analyzing the relationship between statistics and salary over the course of players' entire careers, we didn't include any players whose careers began before 1991.

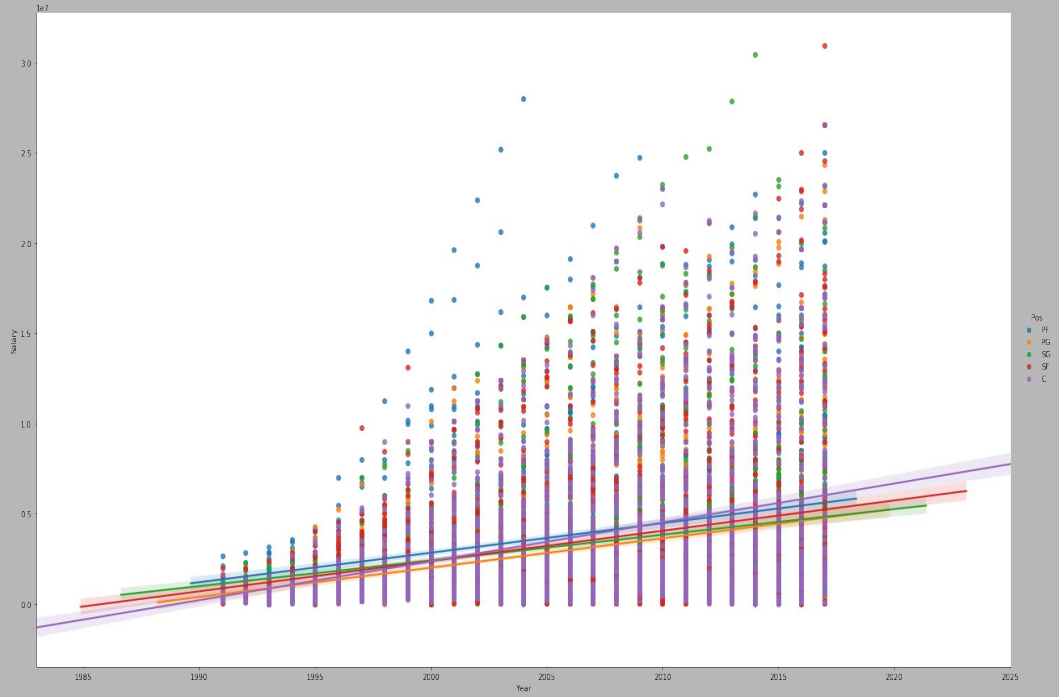
As a result, when we count the number of players in any given year, the stats in the 1990s contain fewer players than subsequent years. The population stabilizes around 2005. Consequently, there will be some bias, especially when we track relationships against years.



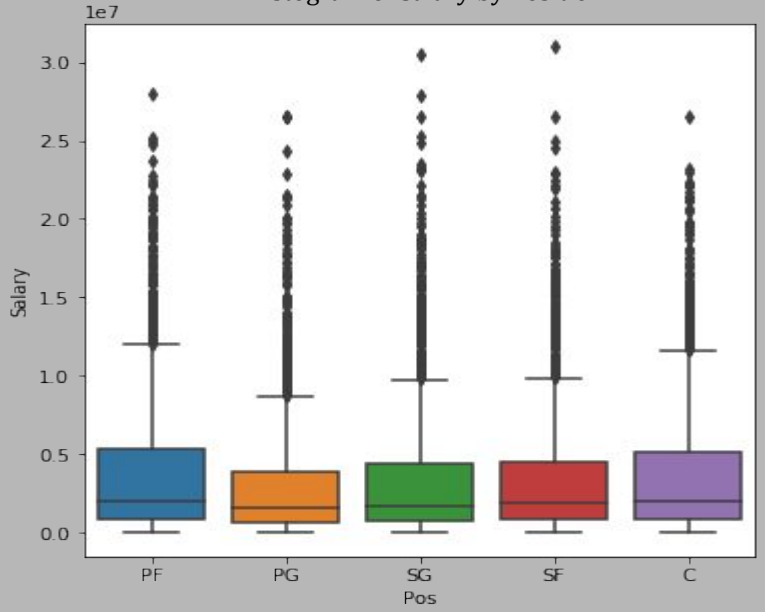
Initial Findings

Salary by Position

Salary Over Time (color coded by position)



Histogram of Salary by Position



Initial Findings cont.

Salary by Position cont.

From the previous graphs, we can see not only that salaries for all players (and all positions) have increased steadily since 1991, but that point guards are historically paid less than other positions, particularly centers and power forwards.

An ANOVA test, where our significance level is 0.05, finds that the difference in salary by position is statistically significant.

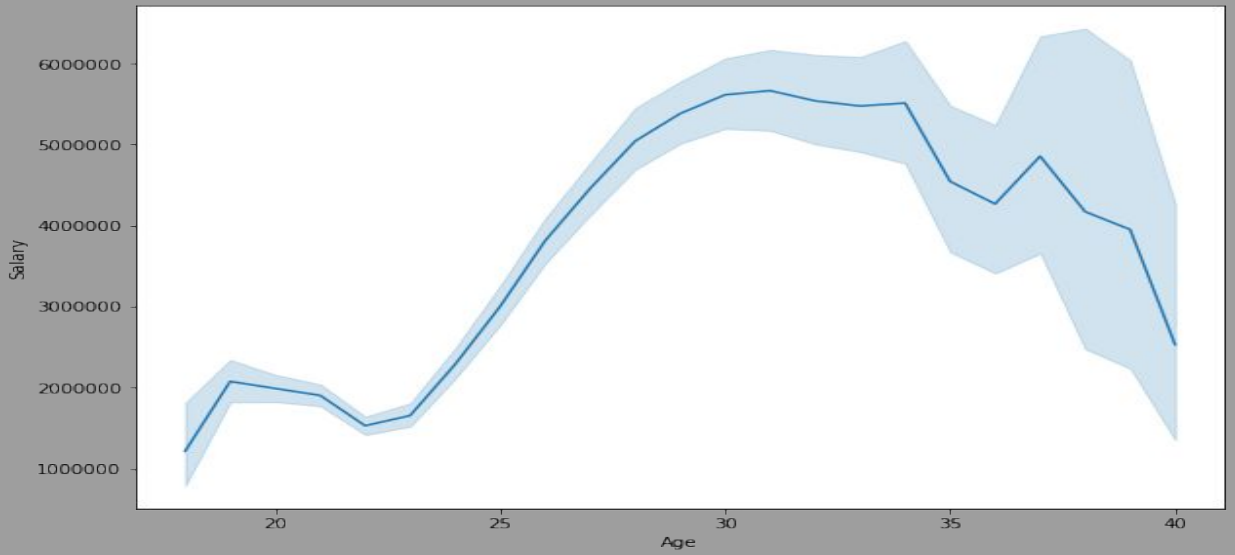
It's not immediately clear as to why point guards are paid so much less, and our dataset doesn't have the necessary data for us to develop a testable hypothesis.

The main takeaway for players is that being a center or power forward is, on average, more lucrative than being a point guard and that one's position does influence one's salary.

Initial Findings cont.

Salary, Age, and Performance

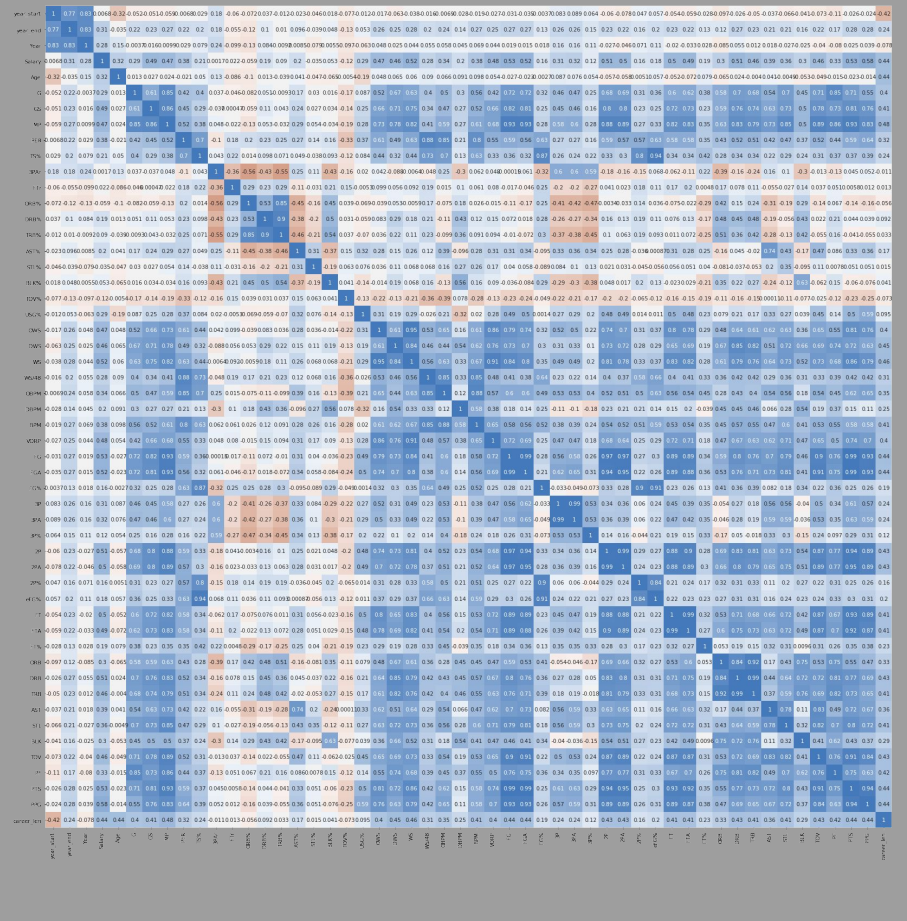
We know that salaries generally increase with age, though there is a drop as players get closer to the age of 40. Salaries peak in the late 20s to early 30s before dropping. Yet, the salaries of veterans are still far greater than those of very young players. This may partly be due to veterans having “proven their worth” so to speak, with years of evidence to demonstrate that they are worthy of high salaries. However, another reason is that veteran players are guaranteed greater pay due to rules negotiated with the NBA by their players’ union.



Initial Findings cont.

Our heatmap shows that there are several stats that correlate somewhat positively with salary including VORP (Value Over Replacement Player), PPG (Points Per Game), and PER (Player Efficiency Rating).

VORP and PER are often used as metrics by statisticians to determine a player's worth, while PPG is a popular stat amongst fans, but high PPG values often lead to higher salaries for players.

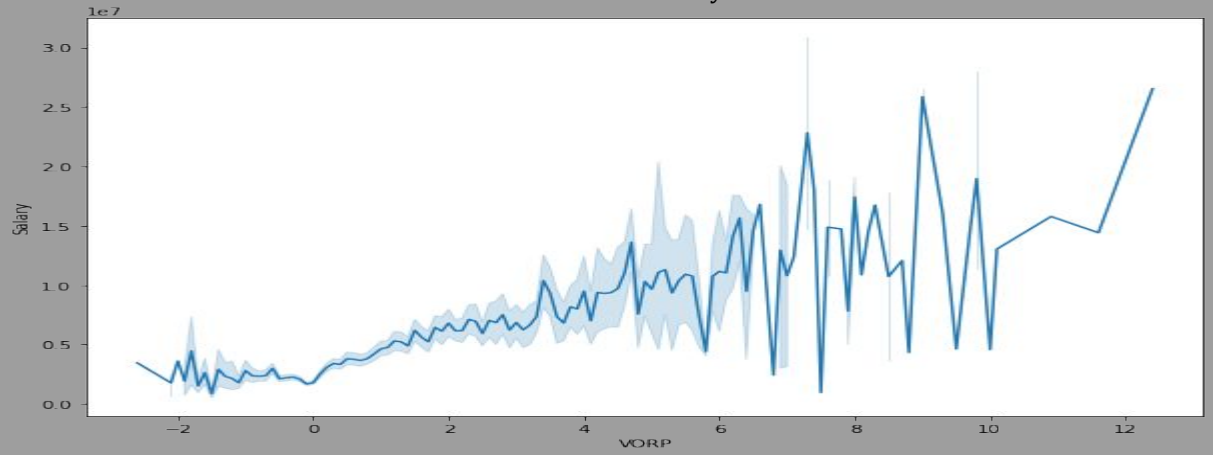


Initial Findings cont.

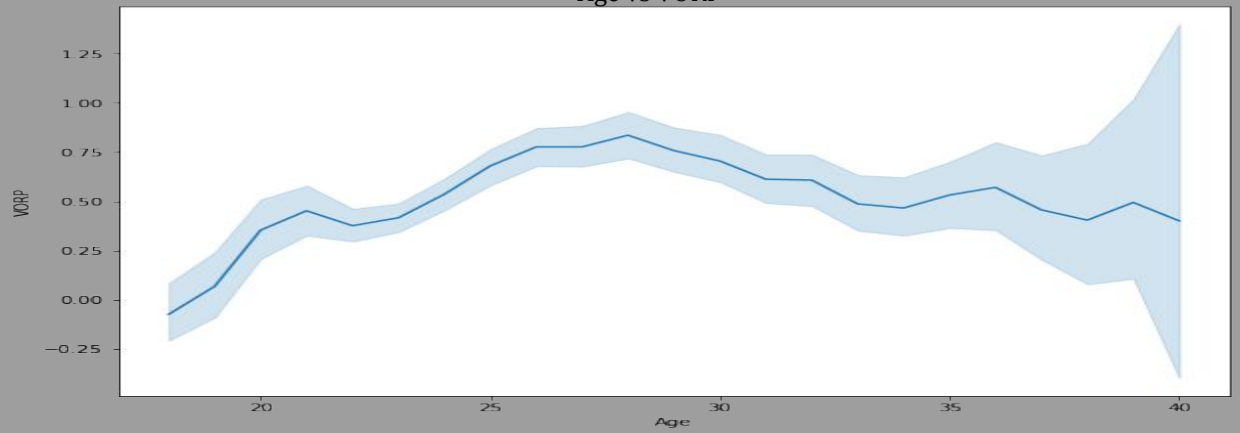
In terms of salary, VORP correlates fairly well with salary. With regard to age, VORP tends to peak in the mid to late 20's.

Since VORP is calculated using several offensive and defensive stats, the decrease in VORP values after 30 suggests that players' overall performances may begin to decrease in their 30's.

VORP vs Salary



Age vs VORP

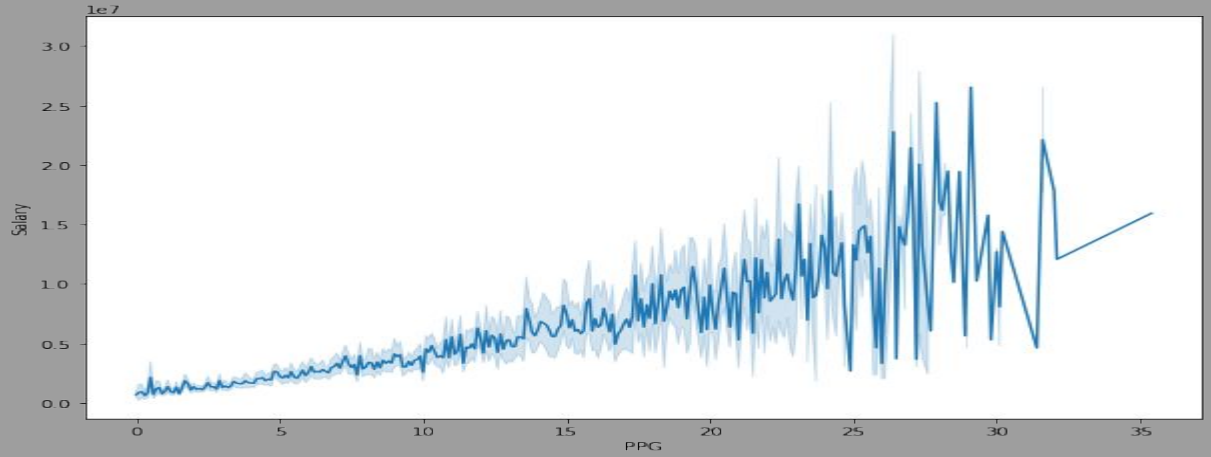


Initial Findings cont.

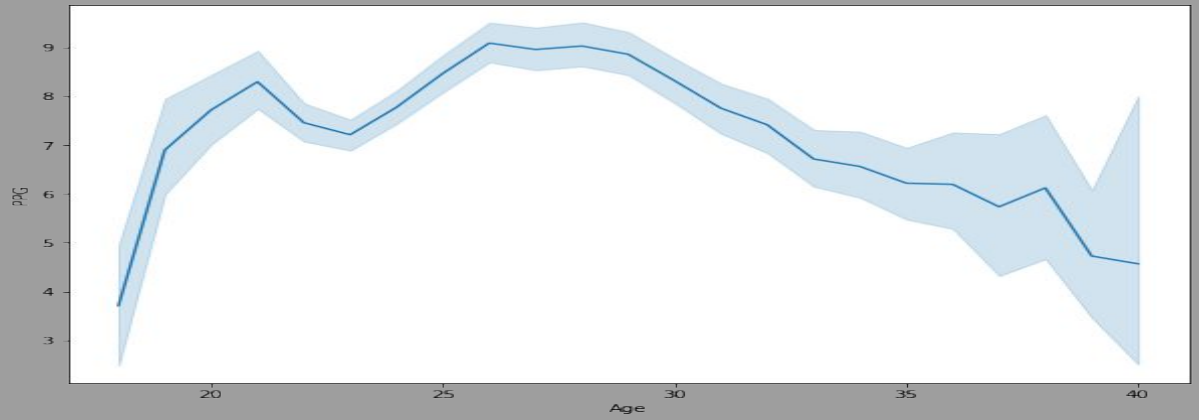
PPG correlates rather positively with salary. As with VORP, PPG peaks during players' mid to late 20's. However, there is a much sharper drop after 30.

If a player who is primarily known for their PPG prowess, wishes to continue making a high salary into their 30s, it would be wise to diversify their skillset due to how sharply PPG drops for older players.

PPG vs Salary



Age vs PPG

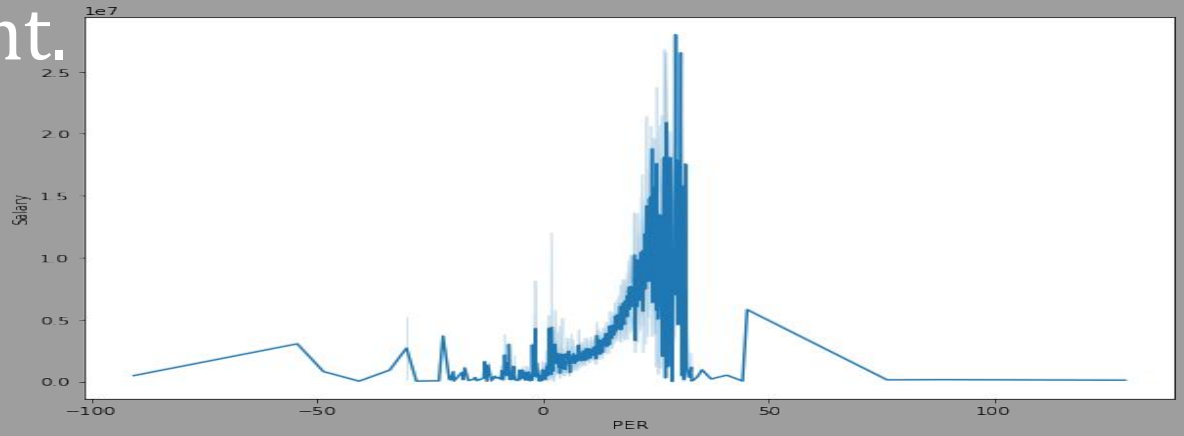


Initial findings cont.

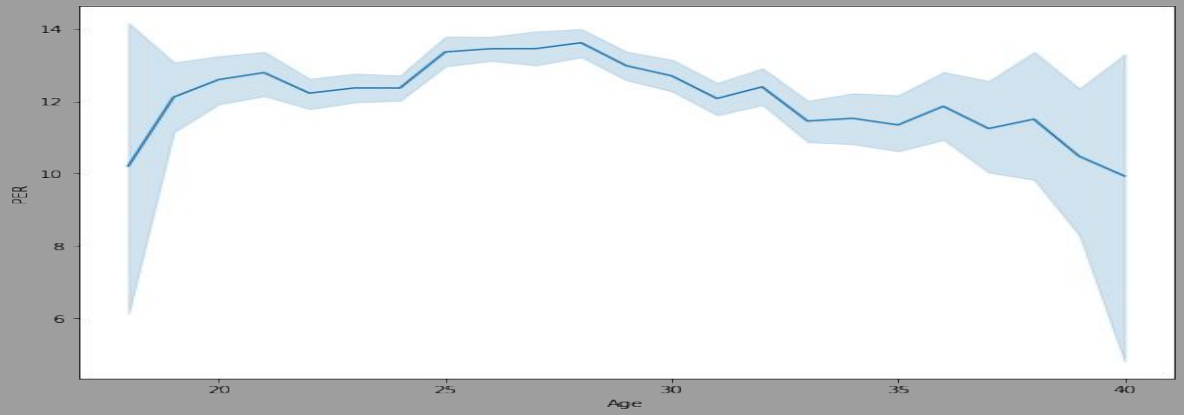
PER is a rather odd metric, as we'll discuss in more detail shortly. However, just in terms of salary, PER seems to cluster around 15-20 for a broad range of salaries. This particular graph doesn't seem to suggest much of a correlation between PER salary.

With regard to age, as with VORP and PPG, it peaks in the mid to late 20's. Though, it doesn't drop as sharply as PPG after the age of 30.

PER vs Salary



Age vs PER



Initial Findings cont.

It should be noted that while VORP, PPG, and PER performance tends to peak in the mid to late 20's, salaries don't generally peak until the early 30's. There may be some possible explanations:

- Young players often have to demonstrate to teams that they are worthy of a larger investment. As such, once a player has shown that they can play well for a sustained amount of time, they usually end up signing either with their current teams or with new teams for larger contracts. But in this relationship, salary is almost always playing catch up to statistical performance.

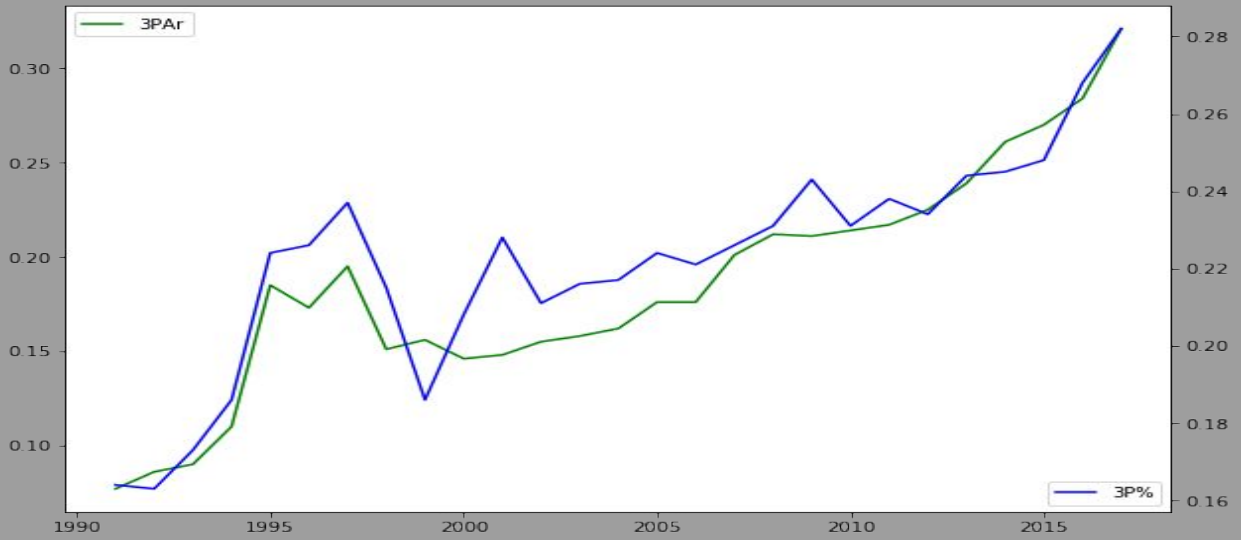
- On the other hand, veterans have "proven their worth" so to speak, with years of evidence to demonstrate that they are worthy of high salaries. In addition, veteran players are guaranteed greater pay due to rules negotiated with the NBA by their players' union.

Initial Findings cont.

3-Point Trends and Salary

While various 3-point metrics didn't correlate highly with salary in our initial dataset, they did show a high correlation when we averaged our data by year. As we see below, both 3PAR and 3P% have increased rather rapidly over the last 10 years. It's indicative of the recent NBA trend in which teams are shifting to schemes that are dedicated to making 3-pointers a greater part of their offense.

Average 3PAr and 3P% vs Time

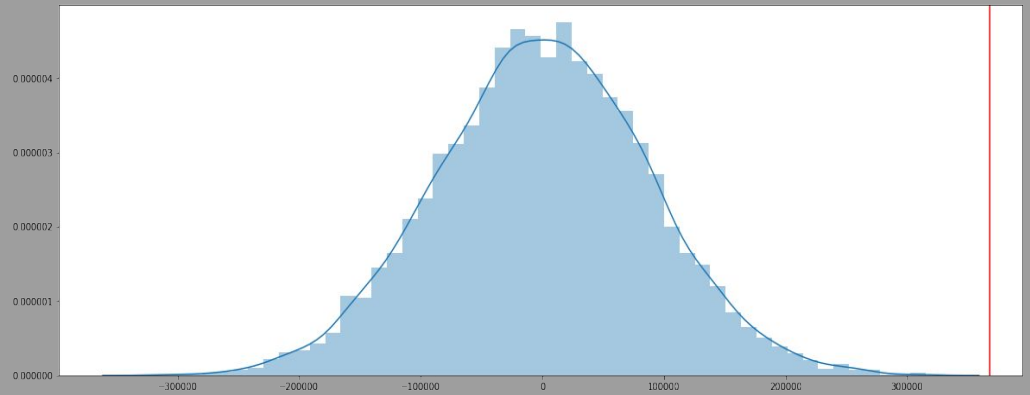


Initial Findings cont.

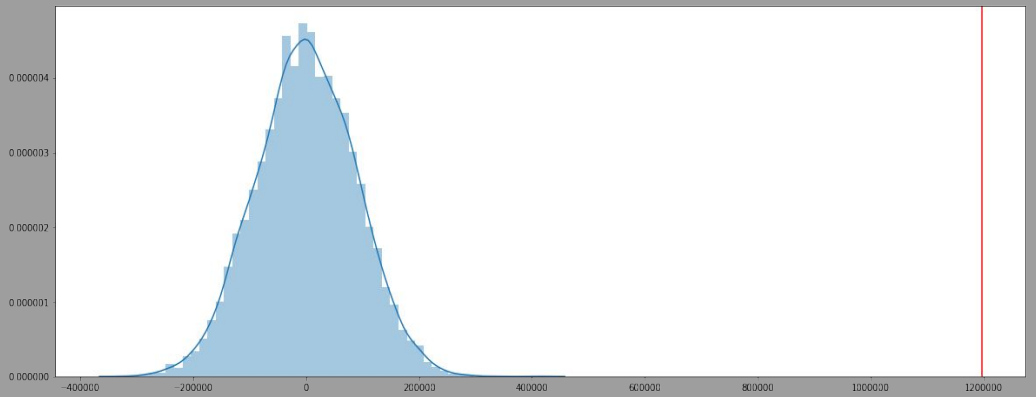
Despite the yearly increase in 3PAr and 3P%, neither stat seems to correlate well with salary. Regardless, we decide to run a bootstrap analysis using both stats in comparison to salary to determine whether the connection between both 3-point stats and salary are statistically significant. For both stats, we find the difference of means in salary of groups above and below the average of each respective stat.

Using a significance level of 0.05, our p-value (indicated by a red line on the graphs) falls well below that level for both analyses. That suggests that there is a connection between 3-point ability and salary, such that players who have strong 3-point abilities are likely to be paid a higher salary.

Bootstrap distribution of 3PAr



Bootstrap distribution of 3P%

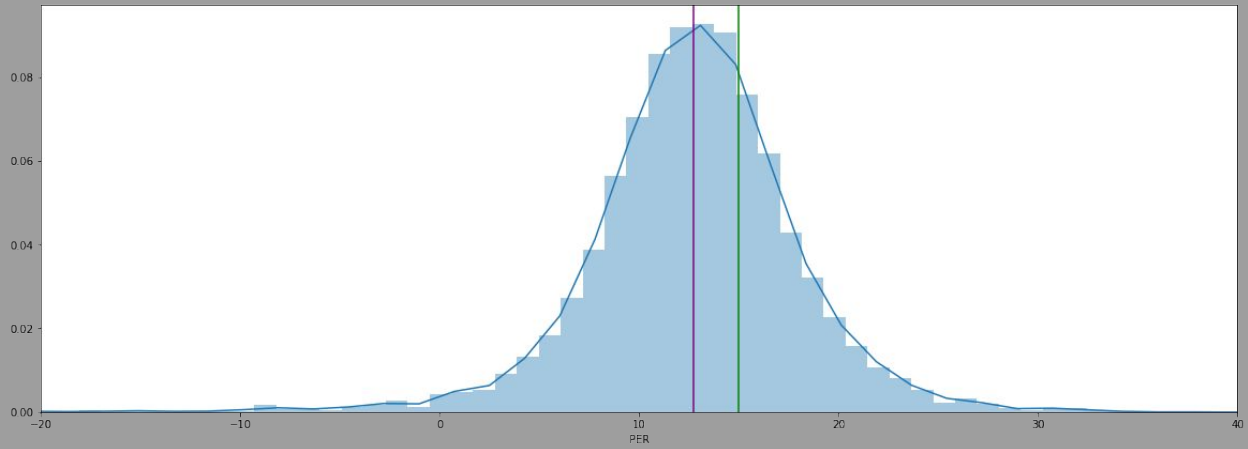


Initial Findings cont.

Returning to PER, it’s a rather curious metric. It's calculated using a fairly elaborate equation that takes into account at least a dozen other player stats. Each season, the league average is set to 15 in order to allow comparisons between seasons.

It's often cited as a valuable stat in determining the 'best' players in the league, partly because it incorporates so many different variables, both offensive and defensive (though it favors offense over defense). However, our graphs seemed to indicate that there was very little difference in PER between players of all different salary ranges. In fact, many players with high salaries often had PER values similar to players near the bottom of the salary scale.

So, a question arises as to why such a valuable stat seems to have such a low correlation (0.378) with salary. We’ll begin by looking at the distribution, where the green line is the NBA standardized average of 15 and the purple line is the average of our dataset - 12.7 :

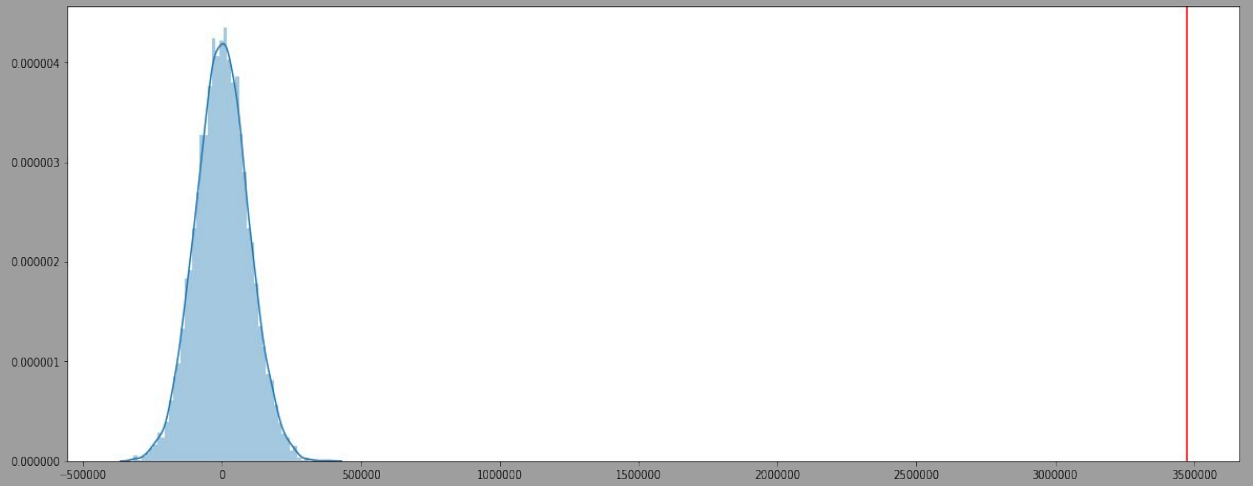


Initial Findings cont.

We run a bootstrap analysis for the difference of means in salary of those players above and below the designated average of 15. We set our significance level at 0.05.

Our p-value falls well below 0.05, as indicated by the red line in the graph below. Thus, we have to conclude that there is a statistically significant correlation between salary and PER, despite what our previous analysis had suggested.

Bootstrap distribution of PER



Initial Findings cont.

PER From both a Player and Team's Perspective

- The main takeaway for players is that they should strive to have high PER values. But given the range of variables that go into such a calculation, a player needs to be fairly well rounded and do well in a number of offensive and defensive categories. It's not to say that players can't be one dimensional and still make high salaries, but being multidimensional can significantly increase one's chances of becoming well compensated financially.

- In our previous Age vs PER graph, we noticed that there was not a dramatic rise and fall in PER, perhaps due to the fact that the average is set to 15 every year. Yet, we know that older players get paid more than younger players for a variety of factors, including player union agreements. But is it worth it for a team to invest more money in older players if they could get similar PER ratings from younger players? If PER was the sole factor in determining wins and losses, then perhaps. However, veterans tend to have less statistically quantifiable assets, such as experience, that may be more important for a successful season.

Machine Learning Analysis

Predicting Player Salary from Statistical Performance

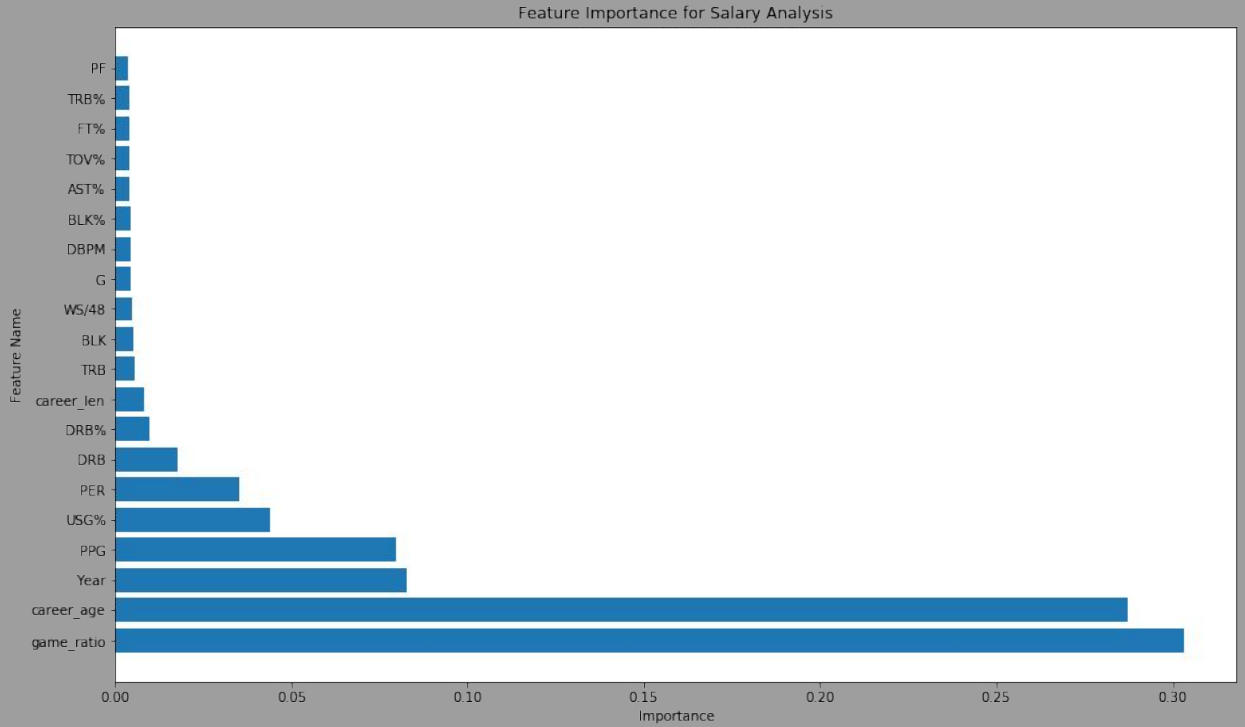
An initial, basic Linear Regression model, which did not have hyperparameter tuning or feature engineering, gave an accuracy of 57%

We then did a more detailed analysis involving:

- Creating two new variables - career age (difference of current year and year player's career began) and game ratio (games started divided by games played).
- Using OneHotEncoder and ColumnTransformer, we transform our numerical and categorical features.
- Testing several initial models including linear regression, random forest, ridge, lasso, and SVR. We find random forest to be the most promising.
- We begin by tuned the model using PCA, which results in a decrease in accuracy. Though it does tell us that we only need to use 70 of our 80+ features.
- Next, we tune the hyperparameters using GridSearchCV.
- Our new, tuned model returns an accuracy of 71%

Machine Learning Analysis cont.

In order to determine how our model was able to increase its accuracy, we find the importance of each feature. Interestingly, our two new variables - game ratio and career age were far more helpful at predicting salary than any of the existing variables. If we revisit this analysis in the future, perhaps we may try to create more variables.



Machine Learning Analysis cont.

There may be several reasons why our accuracy was not higher:

- Stats haven't increased at the same rate as salaries. In other words, while average salaries have increased substantially since the 1990s, many statistical values have not increased at the same rate. As such, the model would have a difficult time projecting stats to salary values because players are being paid substantially higher wages for only slightly better statistical output.
- The inflation rate in the U.S. has varied throughout the years, but salaries in the 1990s would be larger if the inflation rate were taken into consideration. However, this approach still may not make our model very accurate. In the U.S., salaries have never kept pace with the rate of inflation. Even as inflation has increased, salaries have remained largely stagnant. Yet, there's little reason to believe that NBA salaries have much to do with inflation at all.
- There may also be a variety of factors related to players being underpaid or overpaid. For example, teams competing with one another over particular players in a given situation drive up the price of a player in such a way that such an increase may not be warranted outside of that context (such as playoff contenders needing to pick up last minute help, a team looking to rebuild during the offseason with very few strong free agents to choose from, etc).
- There are also considerations of labor union influenced league minimums, which have lead to average salary increases, regardless of player skill. Player popularity may also play a role as teams may be eager to cash in on advertising and ticket sales from a popular player, often resulting in a player getting paid more that his statistical output may warrant. Additionally, as the NBA has become a global brand over the last decade, more money has flowed into the organization, resulting in much of that money being distributed to players.

Machine Learning Analysis cont.

Predicting Player Position from Statistical Performance

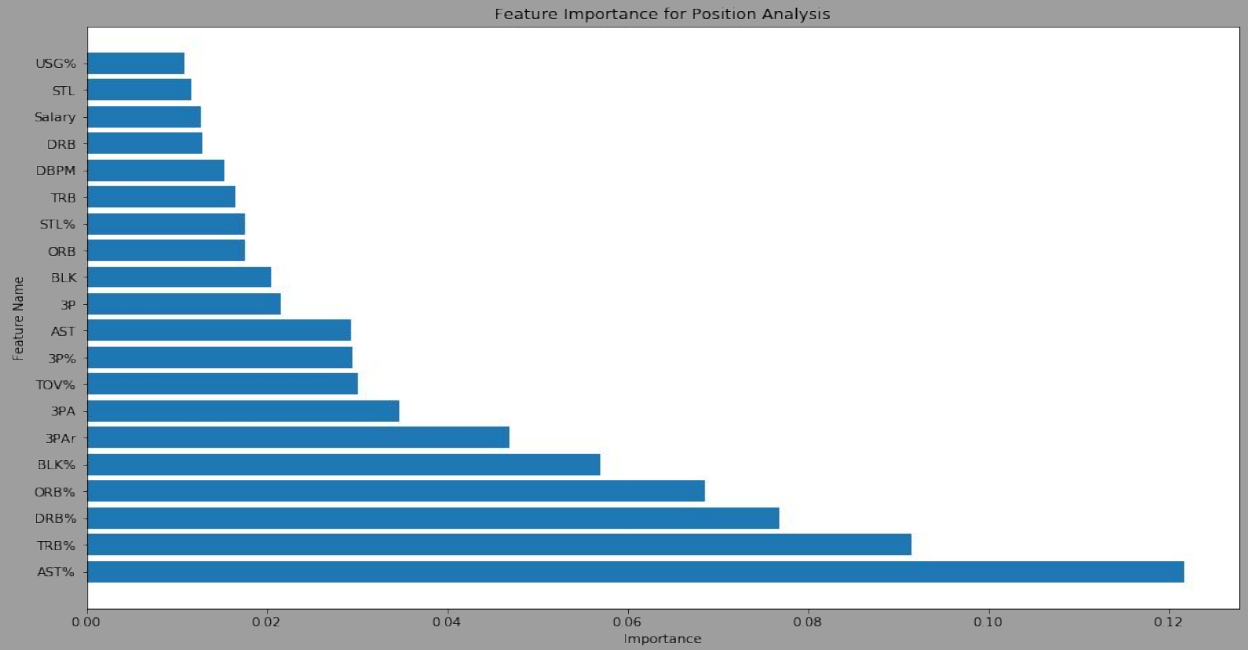
Out of curiosity, we attempt a short side analysis to predict player position. There are five positions in the NBA - center, power forward, small forward, shooting guard, and point guard.

Our process is similar to that used to predict player salary:

- We include our new variables - career age and game ratio. We also transform our data using OneHotEncoder and ColumnTransformer.
- We do an initial analysis using several simple categorical models including logistic regression, random forest, decision trees, naive Bayesian, Gaussian NB, linear discriminant, KNN, and SVC. Logistic regression, random forest, and SVC are found to be the most promising, all at around 65% accuracy.
- After a bit of hyperparameter tuning of all models using GridsearchCV, SVC returns the best accuracy at 69%, while the other two models return 67%.

Machine Learning Analysis cont.

Using our random forest model we can find the feature importance (though SVC had greater accuracy, the random forest code is already available due to our salary analysis). Several of the features that are the most important for the model are those that help differentiate the different NBA positions - assists, rebounds, blocks, and 3-point related stats. Point guards tend to average higher assists than other positions, while center and power forwards average higher blocks and rebounds.



Machine Learning Analysis cont.

While we could likely get our accuracy above 70%, it may be difficult to achieve greater than 80% accuracy for several reasons:

- Our models are looking for accuracies for all positions rather than running a binary analysis to confirm whether a given player is a center or point guard or any of the other three positions.
- While there are considerable differences between positions, such as centers and power forwards averaging higher salaries and blocks or point guards averaging the lowest salaries and assists, the small forward and shooting guard positions muddle things up a bit. These latter two positions tend to have more well-rounded stats than the other three positions. However, if we were using a binary model, we could likely get high accuracy for the three most distinguishable positions. But since our search isn't binary, the model is probably getting confused by, for example, small forwards with high assist values or shooting guards with a high number of blocks.
- Additionally, there's one other factor that we weren't able to take into account - height. Height varies greatly between positions, with point guards typically being the shortest players and centers and power forwards being the tallest.

Takeaways for Players and Teams

- The NBA isn’t as position specific as it used to be. Players at all positions have to be more versatile than in the past, and the more multidimensional a player is, the more valuable they become. It’s crucial for players to be aware of that for the sake of their career longevity and earnings. As we’ve seen, player performance peaks in the late 20s and salaries peak accordingly in the late 20s and early 30s.

- On the other side of the coin, NBA teams can also benefit from this analysis. By realizing when player performance tends to peak and by understanding the cost/benefit relationship between salary, performance, and age, teams can make better decisions about how much to pay players.

- This analysis seems a bit dismal for older players - they get paid high salaries, and in many cases, they don’t necessarily play much better than their younger, more affordable counterparts. So why shouldn’t teams just load up their rosters with younger players who cost less and can potentially offer similar results? Luckily, for older players, this isn’t the only factor that teams consider, or should consider, when putting together a roster. In addition to a variety of other statistics, there are many other factors, such as experience, temperament, the ability to perform under pressure, and the ability to get along with teammates, that aren't as easily quantifiable and that may be in greater abundance in older players.

Takeaways for Players and Teams cont.

- Finally, our machine learning analysis also has important insights for players.The fact that we weren't able to predict players' salaries anywhere close to 90% accuracy based on their statistical performance is worth noting. What it suggests is that there's more to how much a player gets paid than just their on-court performance.

- They, for example, can learn to better market themselves. Even if they happen to have mediocre stats, by marketing themselves properly, they can make themselves more financially lucrative to a team. This could be in the form of advertising deals, both for the team and player. The increased popularity can also lead to increased ticket sales and, of course, a greater salary for that player than his on-court performance warrants.

Links to References

Final dataset:
https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/nba_final_dataset.csv

Project Proposal:
https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/Documents/Capstone%201%20Project%20Proposal.pdf

Data Wrangling Report and Notebook:
https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/Documents/Capstone%201%20Data%20Wrangling%20Summary.pdf

https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/nba_data_cleanup%20.ipynb

Data Story Report and Notebook:
https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/Documents/Capstone%201%20Data%20Story.pdf

https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/nba_data_story.ipynb

Exploratory Data Analysis Report and Notebook:
https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/Documents/Capstone%201%20Exploratory%20Data%20Analysis.pdf

https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/exploratory_data_analysis.ipynb

Machine Learning Analysis Report and Notebook:
https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/Documents/Capstone%201%20In-Depth%20Analysis%20.pdf

https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/nba%20machine%20learning%20analysis.ipynb

Final Report:
https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/Documents/Capstone%201%20Final%20Report.pdf

Slide Deck:
https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/Documents/Capstone%201%20Slide%20Deck.pdf