Capstone 1 Final Report - NBA Salaries and Statistics

Contents

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. If a player can better understand when he will be likely to experience a drop in performance, he can use that information to his advantage to sign more lucrative contracts at an appropriate point in his career. Once machine learning is incorporated, a player may be able to use the information for contract negotiations.For example, given certain career accomplishments, and projected accomplishments, a player may be able to use this analysis to better understand how to negotiate for a better salary based on how his peers have been compensated for similar performances.

The second group than 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. For an NBA team, maximizing their investments in their players is crucial to their financial success. This project would allow them to have a better idea of how much it's worth to pay a player at different points in his career

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 data was acquired from a basketball statistics website.

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.

The first major step is to combine the salaries and career datasets. Since the salaries data begins in 1991, the career data also has to be cropped so that the only players who are included in the final data set are those who began their careers in 1991 or after. Unnecessary columns are dropped from both columns and they are both reindexed to the 'Player' column. The career set requires further cleaning, however, because there is duplicate player data, partly because of data entry errors and partly because several pairs of players shared the same name. All of those players are dropped because there are only a few instances and the cleanup required would have been time consuming. Their removal should influence the final results only minimally.

In the salary dataset, the 'Salary' column's numbers are string values that needed to be formatted properly to convert into float. There are also several 'Unknown" values in the 'Salary' column. It proves difficult to find all of the missing salary values, especially since they were from the 1998-1999 season, which was shortened due to a strike. Instead, all players that have unknown salary values are dropped from the dataset. As for the remaining numerical values, they have to be stripped of extra spaces on both the left and the right sides.

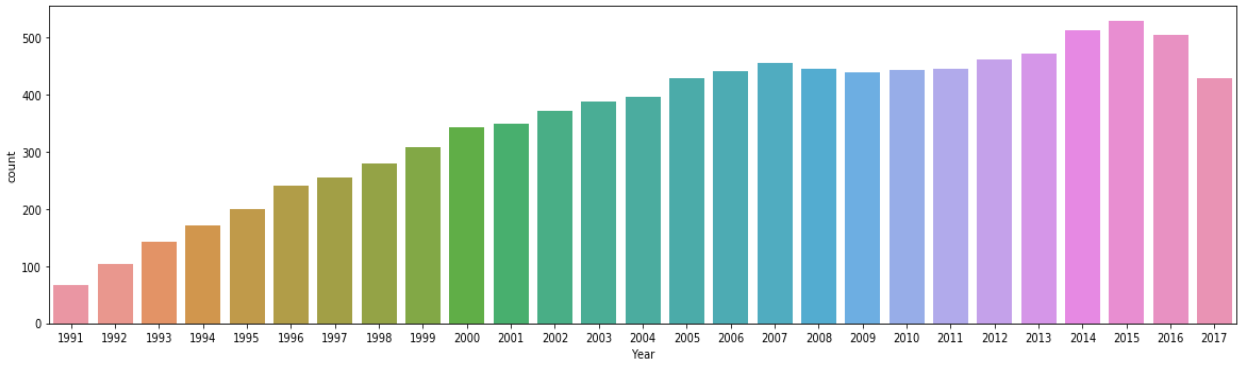
The salary and career sets are then inner joined by the 'Player' column. The new combined salary/career set is checked for null values and none are found. For a few players, there are duplicate rows when taking into account player names, year, and salary. The second instance of these rows are dropped because they are redundant - they are acting as both the total paid to a player by a particular tem and as the yearly total. For these particular players, they changed teams during the course of the season, but their yearly total was paid just by one team, not both teams that they were members of. In contrast, looking for duplicates just based on player name and year indicates that several players who played for multiple teams in a year were paid different salaries by each team they played for that specific year. Those salaries are summed using groupby on player and year columns. The index is then reset so that set is no longer multiindex.

The stats data is then cleaned, first by looking at the player/year duplicate rows. In the set, when players played for multiple teams in a given year, the totals for that year are listed in the first row, while subsequent rows for the same year divided the totals by team. Those subsequent rows are dropped while the yearly totals are kept.

The new salary/career data is then combined with the stats data through inner join on the 'Player' and 'Year', and columns. A search for null values reveals several in multiple columns. Three players have null values for stats that would not be calculable with the given data, so those players were dropped. A closer look at the remaining null values, and the formulas required to calculate each columns respective stats, indicate that they are due to division by zero. Those null values are replaced by 0.0 floats. An additional column for points per game is added to the set by dividing the number of points scored in a year divided by the number of games played in that same year. It is a surprising omission for the set, given that it's a widely used and cited stat when assessing player value. A final inspection for null values and duplicates finds nothing. At that point the final dataset is considered fully cleaned.

Disclaimer

Before we move on, we need to make a quick note about bias in the dataset that could influence some of our results. As detailed above, when the current dataset was being created, the original salary dataset began in 1991, while the player statistics dataset began much earlier. 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. This includes some of the top players of the 1990s, including Michael Jordan, Charles Barkley, and Magic Johnson. 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. As a result, there will be some bias, especially when we track relationships against years. The population difference is evident in the graph below.

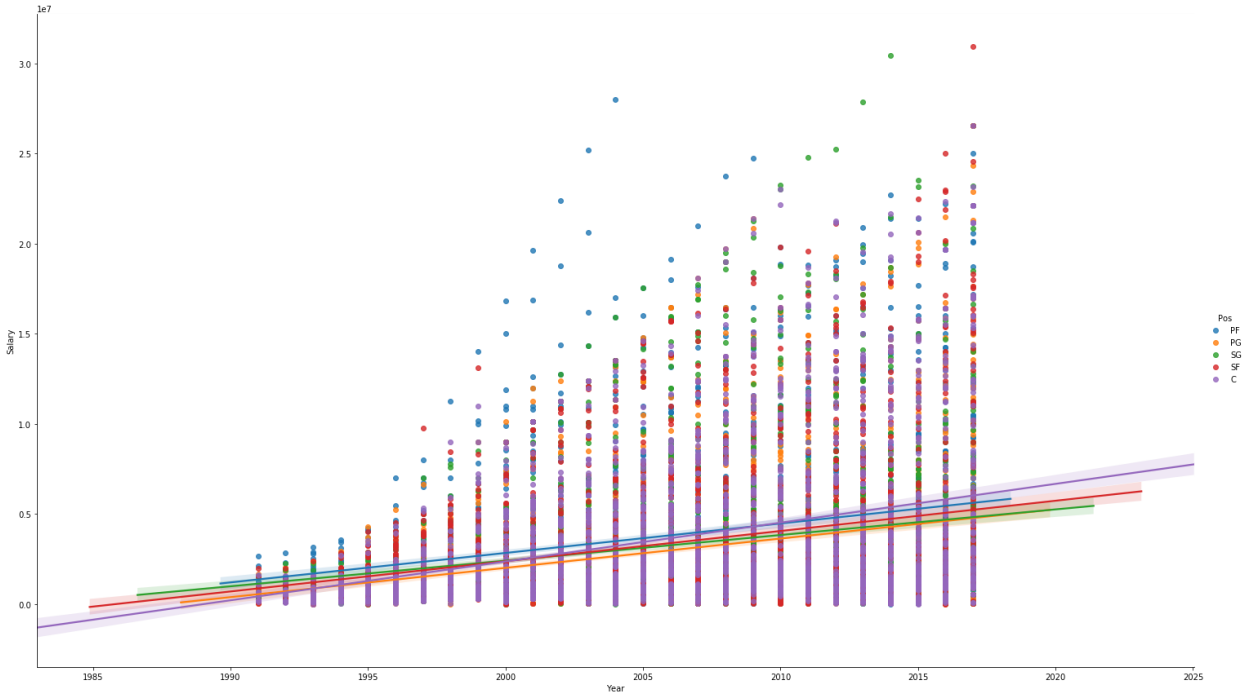


Ideally, we would like to redo the data to also include all players who were playing in the 1990s, regardless of whether their career began before 1991. However, due to the time constraints of the course, that's just not possible at the moment. At a later time, we may consider revising the dataset appropriately.

Initial Findings

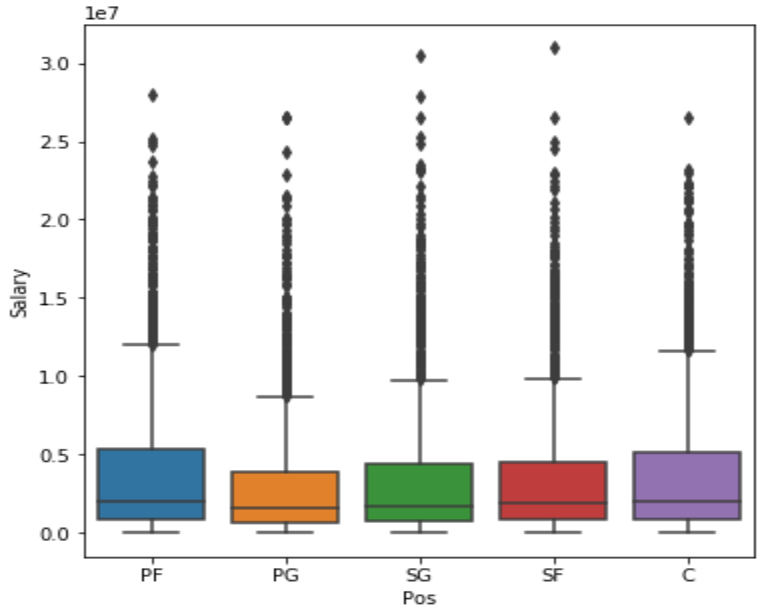
Salary by Position

We'll begin by plotting salaries over time as a scatterplot.



The scatterplot, hued based on positions, makes it clear that salaries have generally trended upward since 1991, regardless of our aforementioned data bias. Before 2005, it appears that most of the outliers were PFs. More recently, it appears that shooting guards are the outliers. However, as the trendlines indicate, salaries for all positions have risen since 1991.

A boxplot of position and salary better demonstrates the relationship between salary and position.



The boxplot indicates that point guards have been paid lower salaries as their plot is more condensed than the other positions. It also has less extreme outliers. We can use an ANOVA test to determine if the difference in salary by position is significant.

For an ANOVA test, it's crucial that data be normally distributed. A log transformation of the data showed a fairly normal distribution allowing us to proceed. For our analysis, our null hypothesis is that there is no difference in salary between positions, while our alternative hypothesis is that the difference is significant. We'll use a significance level of 0.05.

An initial analysis confirms that centers and power forwards have the highest average salary.

Pos	N	Mean	SD	SE	95% Conf.	
C	1979	3.796417e+06	4.288673e+06	96405.112966	3.607416e+06	3.985419e+06
PF	2030	3.869948e+06	4.513690e+06	100180.615818	3.673546e+06	4.066351e+06
PG	1836	3.100985e+06	3.916515e+06	91403.639404	2.921785e+06	3.280185e+06
SF	1864	3.451256e+06	4.121353e+06	95458.990769	3.264106e+06	3.638406e+06
SG	1912	3.407347e+06	4.191853e+06	95865.439075	3.219402e+06	3.595292e+06

Running an ANOVA test confirms that we should reject the null hypothesis:

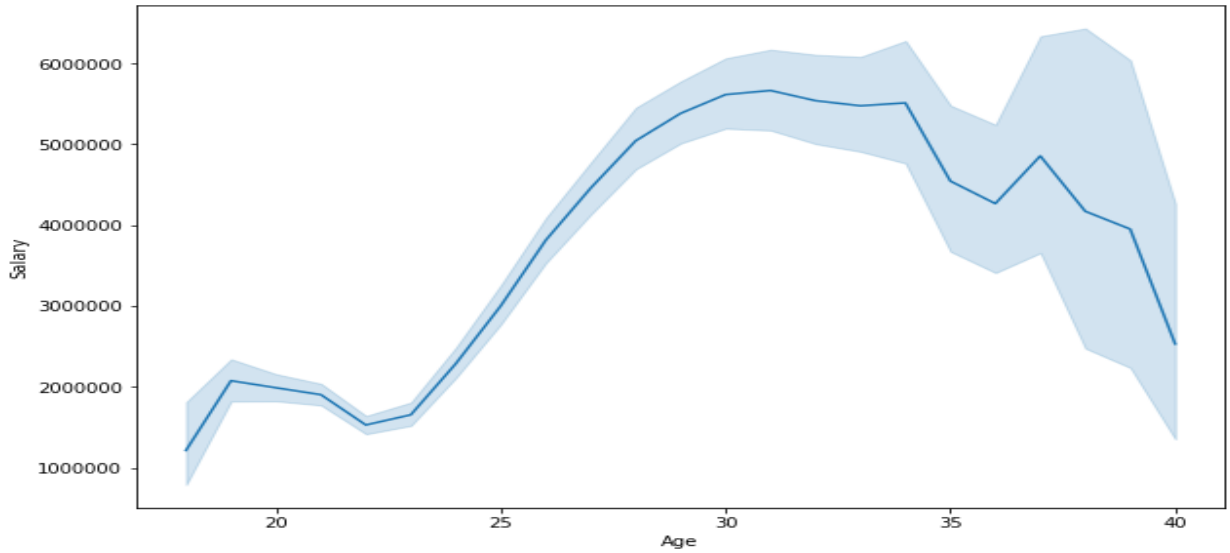
	sum_sq	df	F	PR(>F)
C(Pos)	7.530635e+14	4.0	10.581414	1.489611e-08

The sum of squares gives the amount of variance in the data. In this case, there is a great amount of variance. Paired with the extremely low p-value and the F-stat, the ANOVA test indicates that we should reject our null hypothesis. We can conclude that there is a significant difference in salary between positions. It's likely the case that position does influence salary. Centers and power

forwards are paid more than other positions, while point guards are paid the least, by a wide margin. 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.

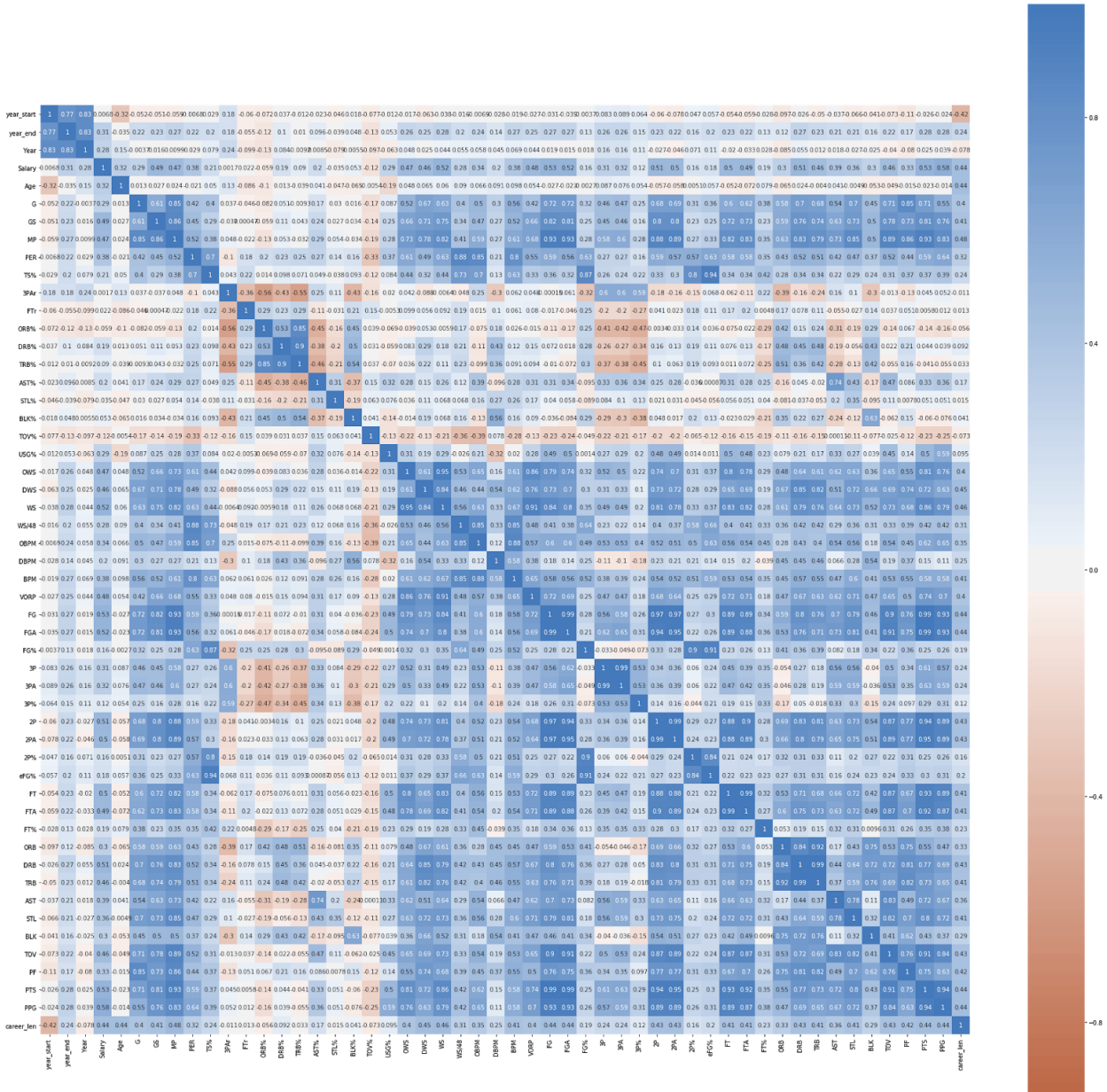
Salary, Age, and Performance

We know that salaries generally increase with age, though there is a drop as players get closer to the age of 40, as evident in the graph below:

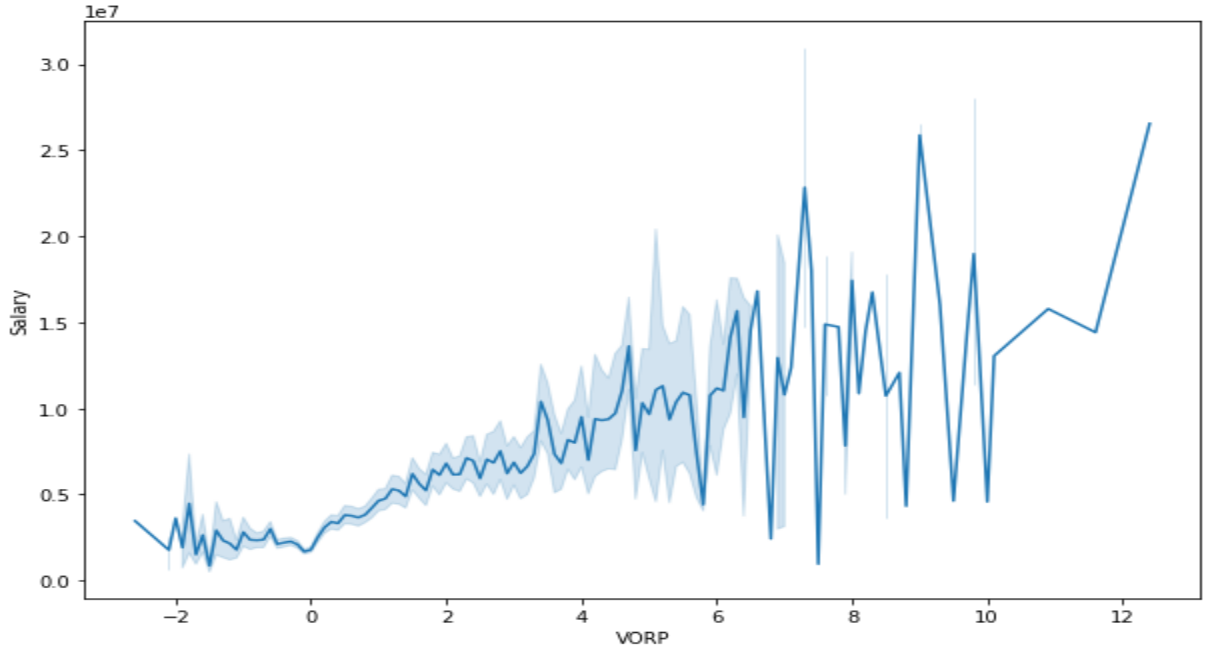


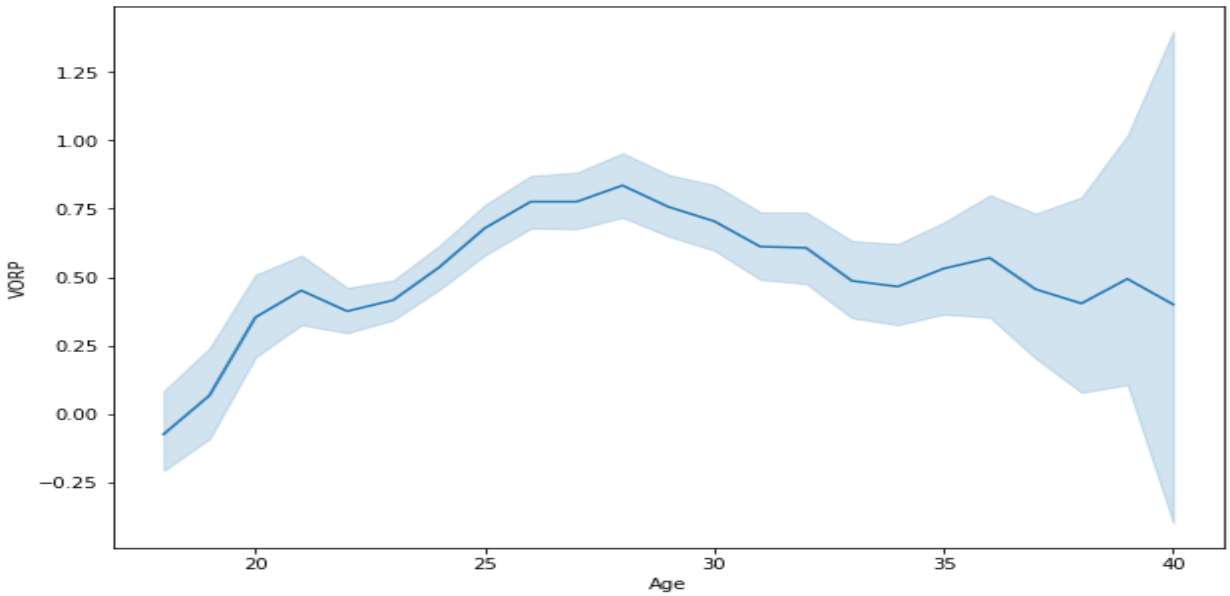
Salaries peak in the late 20s to early 30s before dropping. Yet, the salaries of veterans are still far greater than those of very young players. This may partly be due to veterans having “proven their worth” so to speak, with years of evidence to demonstrate that they are worthy of high salaries. However,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

In a previously created heatmap, we notice that several statistics correlate positively with salary, including VORP, WS, PPG, and, to a lesser extent, PER.



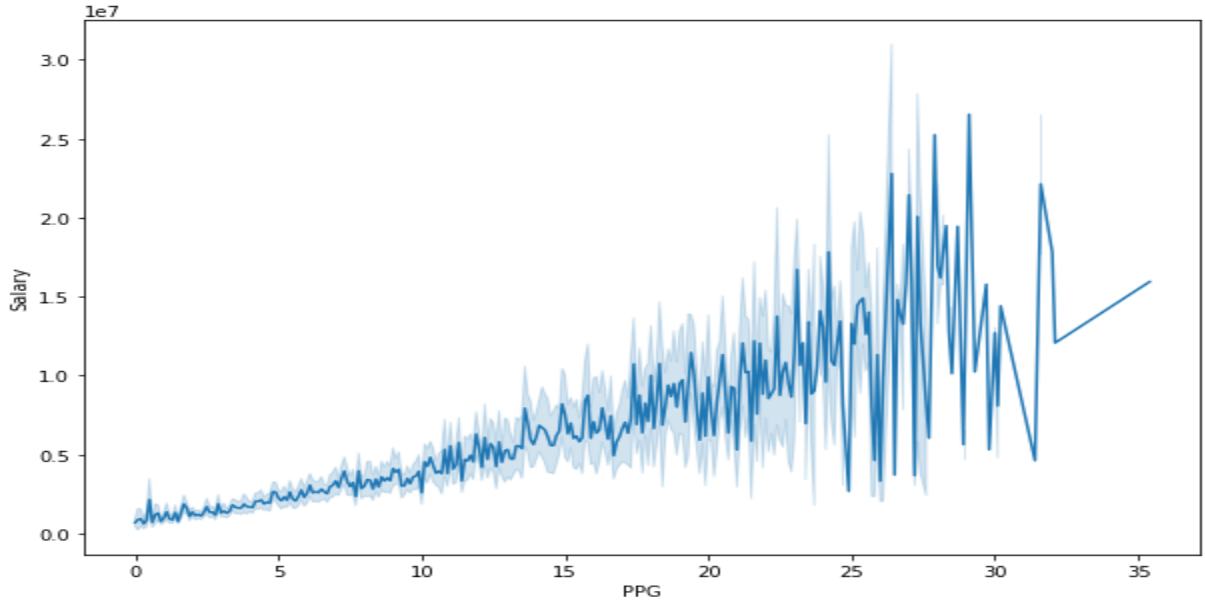
VORP, or Value Over Replacement Player, is measured by converting BPM (Box Plus Minus) in terms of each player's contribution in a given game. BPM itself is calculated using a variety of offensive and defensive metrics. Plotting VORP against salary and age, we get the following results:

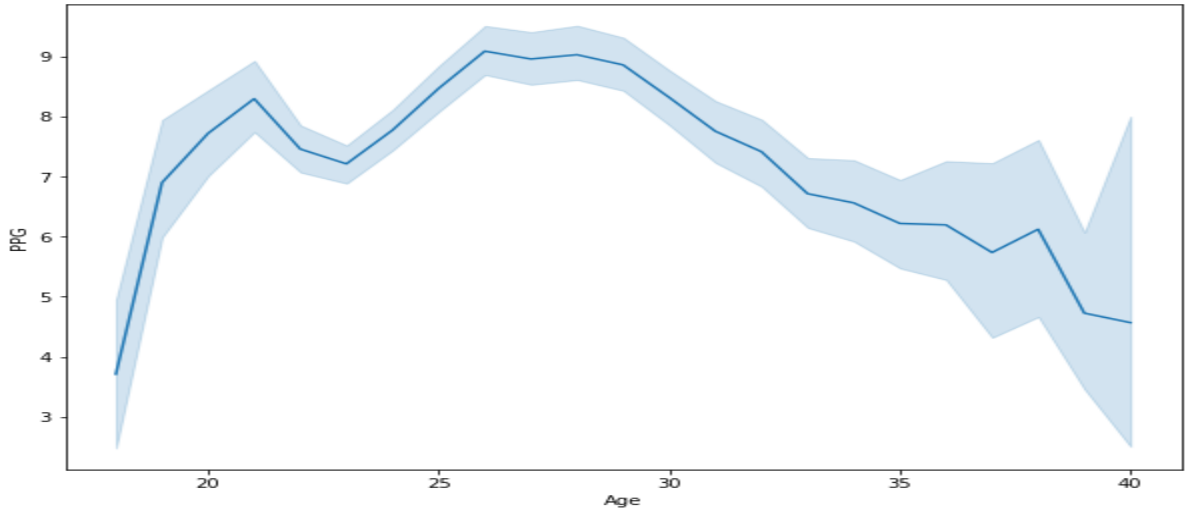




In terms of salary, VORP appears to correlate with an increase in salary. With regard to age, it peaks in the late 20s and continues to fall through the rest of most players' careers. It suggests that having a strong VORP rating does contribute to a higher salary, but as various offensive and defensive stats drop, so does one VORP rating. WS, or Win Shares, are very highly correlated with VORP and have very similar trends for both salary and age.

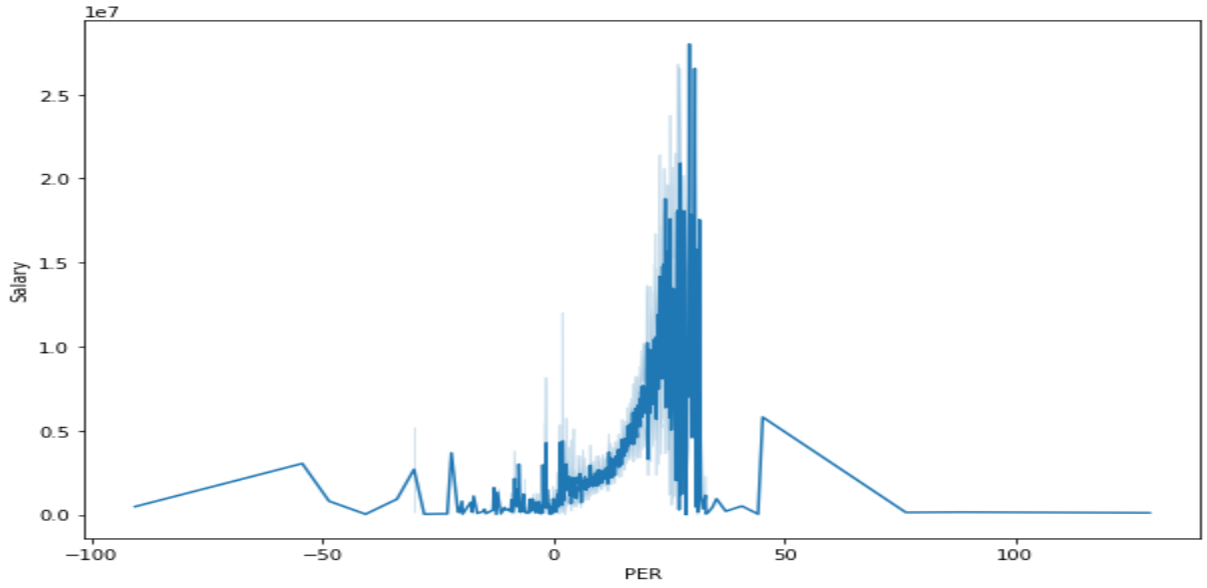
PPG, or Points Per Game, is stat that is perhaps more valued by fans than statisticians. Yet, that popularity with fans can often lead to higher salaries for players. If we graph it opposite both salary and age, we get the following result:

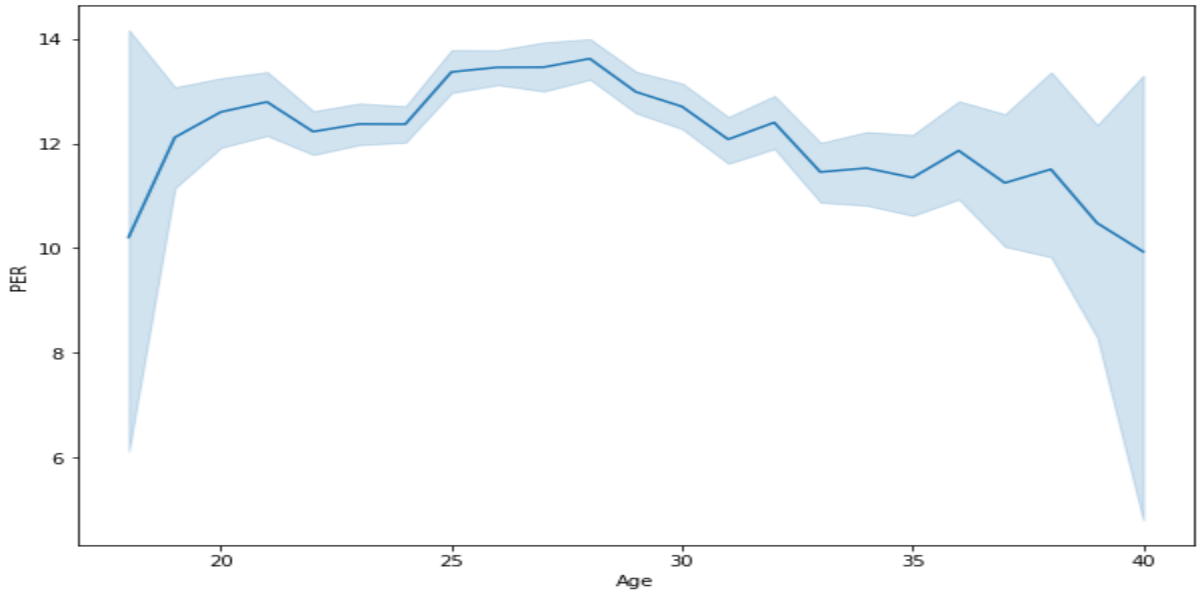




An increase in PPG correlates well with an increase in salary, but compared to WS and VORP, it peaks even earlier than the previous stats and falls at an even steeper rate after age 30. If a player who is primarily known for their PPG prowess, wishes to continue making a high salary into their 30s, it would be wise to diversify their skillset due to how sharply PPG drops for older players.

Finally, PER, or Player Efficiency Rating, will be discussed in much more detail later in this paper because it's a bit unusual compared to many other NBA stats and metrics. However, here is how it trends in comparison to both salary and age:





PER remains a bit more steady than the other stats, but it still seems to peak in the mid to late 20s and drops off after age 30. However, there seems to be little correlation between PER and salary. Players seem to concentrate somewhere between 0 and 50 PER regardless of salary, suggesting that PER may have little to do with a salary increase. As we'll see shortly, understanding PER is a bit more complicated than it initially appears.

An additional item of note is that salaries tend to peak after many of the above statistics do rather than at the same time. The likely reason is that young players 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.

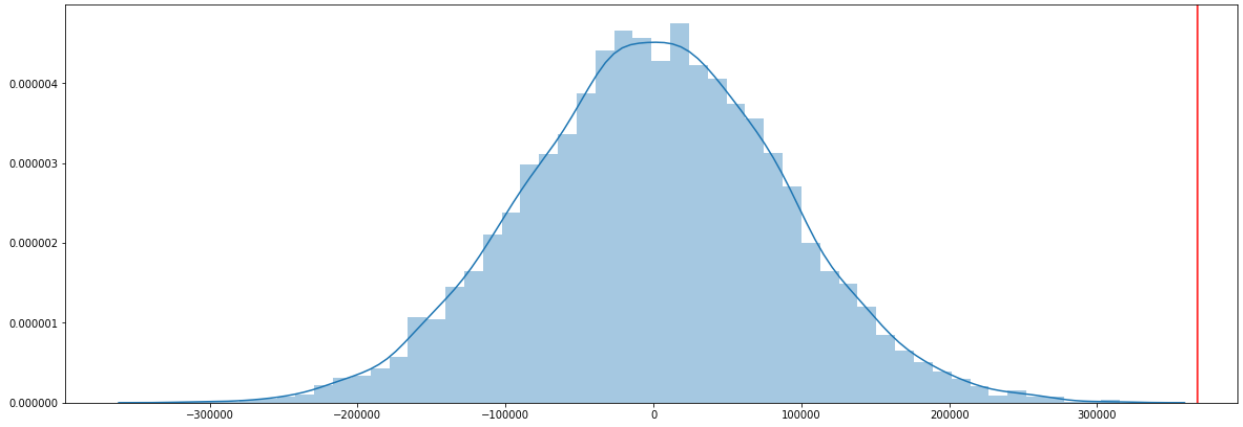
3-Point Trends and Salary

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.

As 3-pointers have become more important in the ever-changing NBA playstyle, more 3-point shots are being taken and made. So, it's worth determining if there is a connection between salary and these two 3-point stats.

We begin by determining the difference of means in salary between groups divided by the average 3PAr value. 3PAr is calculated by taking the ratio 3-point attempts to field goal attempts. The larger that a percentage of a player's total shots are 3-pointers, the larger their resulting 3PAr. This statistic doesn't take into account shots made.

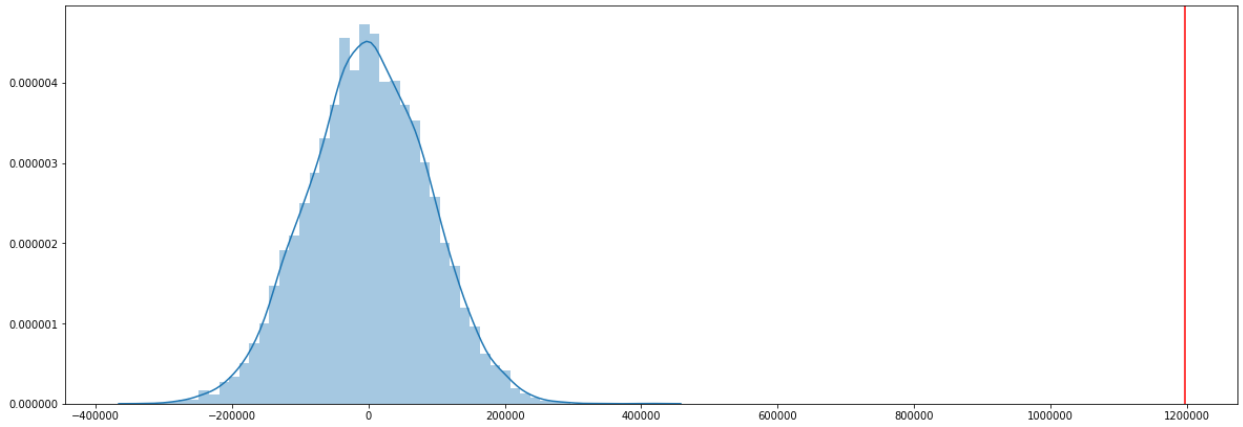
Our null hypothesis is that there is no difference in salary between these two groups. Thus, the alternative hypothesis is that there is a significant difference between players above and below average 3PAr. We run a two sample bootstrap analysis. Our difference of means in salary is $368,052.38. Below is the distribution of the bootstrap sampling with our difference of means included as a red line:



Our p-value is 0.0. That allows us to reject the null hypothesis and conclude that there is a significant salary difference between groups above and below average 3PAr.

Next, we'll take a look at three-point percentage, which is a simple calculation that involves taking the ratio of 3-pointers made to 3-point attempts. As the league has started to shift towards a more 3-point oriented game, more 3-point shots are being attempted. As a result, it seems reasonable to assume that players who are good at making such shots would be highly valued and paid well as a result.

Our null hypothesis is that there is no difference in salary between the two groups, while our alternate hypothesis is that there is a significant difference. Our difference of means in salary is $1196259.38. The bootstrap analysis results in the following distribution:

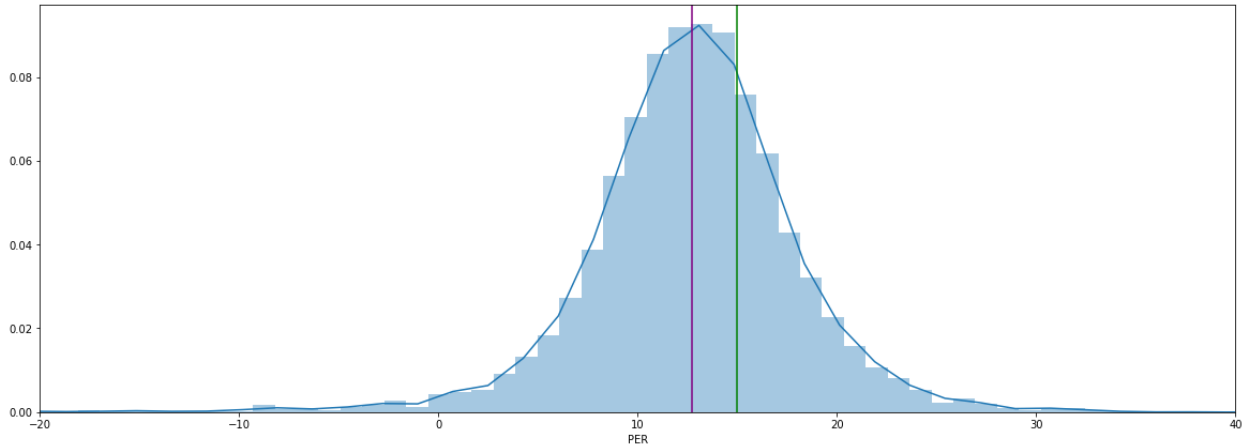


Again our resulting p-value is 0.0, allowing us to reject the null hypothesis. The main takeaway for players, especially in the modern NBA, is that in order to increase one's salary, it's advisable to to become a competent 3-point shooter. In past decades, this advice may not have been relevant for every position (centers and power forwards, for example). But given how important the 3-point shot has now become, even centers and power forwards would benefit from becoming better 3-point

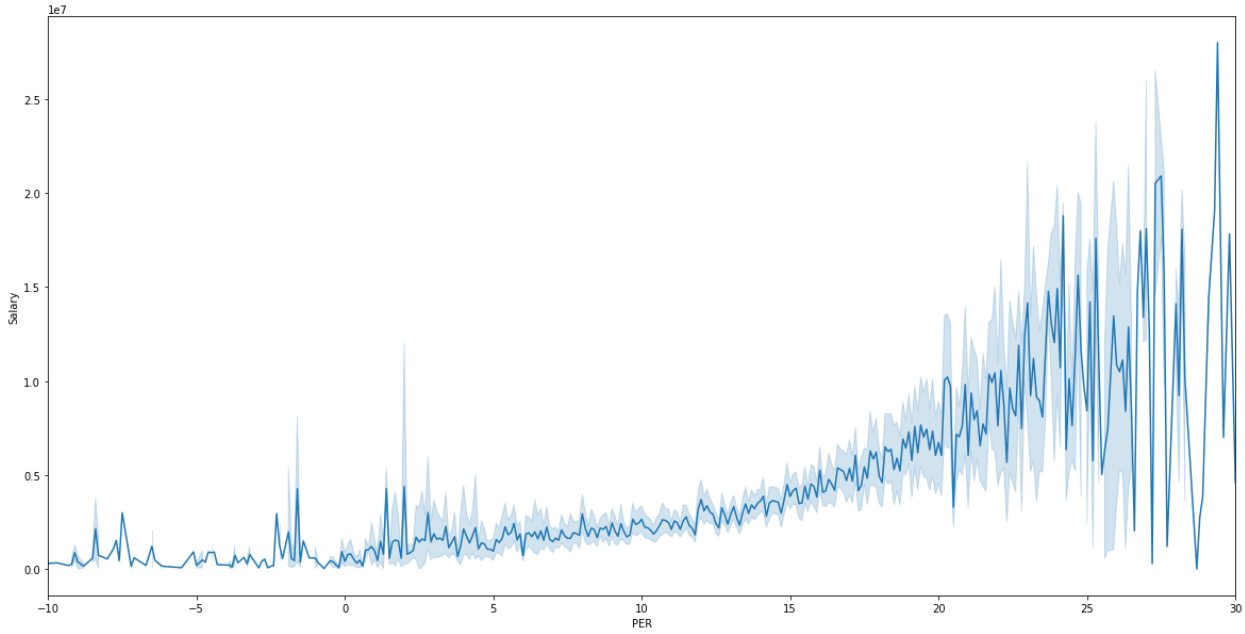
shooters. As we'll see with our PER analysis, the more versatile and multi-dimensional a player becomes, the more they benefit financially.

PER

PER, or player efficiency rating, is a rather curious statistic. 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 data story 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 :



It appears that the vast majority of PER values fall just under 15. This may be due to the league average being adjusted to 15 every season as a way to better allow comparisons between seasons. Thus, most players fall under the standardized average. Let's take a look at the how PER and salary compare graphically.

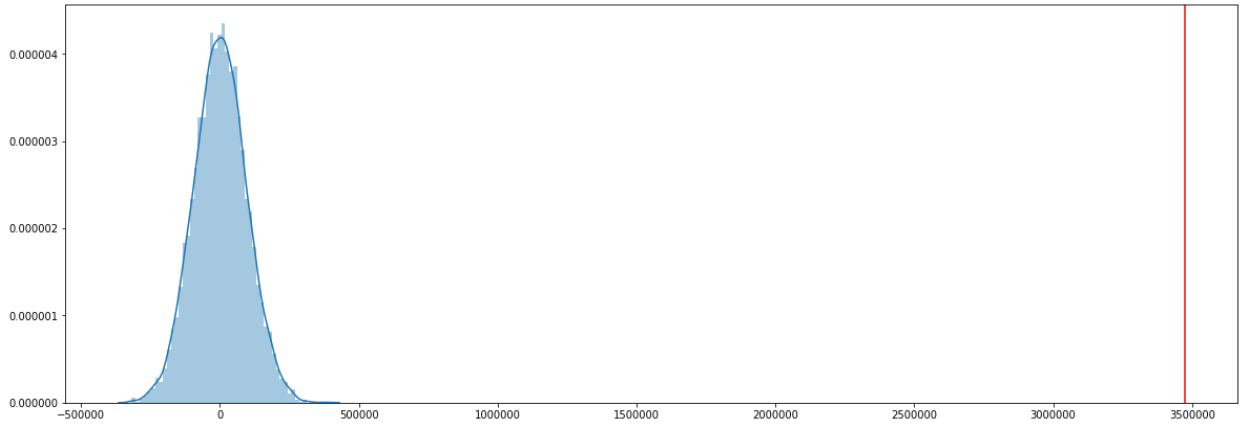


For most PER values salary is fairly stable. There is large increase after 15. Yet, the most noticeable aspect is the drastic increase in variance after 20. It's not clear why there is such a huge variance. It's likely due to a small group of players with high variance having large salaries well above the league average. Players above 20 are considered elite, with the majority of them being all-star and MVP candidates.

In the entire data set, from 1991 to 2017, there are only 614 cases of a PER higher than 20, so it's quite rare. Below, by graphing all PER values above 20, we see how drastic the variance in the data is. While the correlation between salary and PER is quite low, because there are many extreme salaries associated with strong PER performance, the weight of those salaries may be enough to lead to the conclusion that the salary difference between players above and below the average PER is statistically significant, despite the fact that a majority of all players' PER values fall in a very narrow range.

We'll run a bootstrap analysis to see if there is a difference in salary between players who are above the designated average of 15 and those below. For our null hypothesis, we'll assume that there is no correlation between PER and salary. Our alternate hypothesis will be that there is a correlation.

The difference of means between the two groups is $3,472,748.92. The bootstrap analysis gives the following distribution:

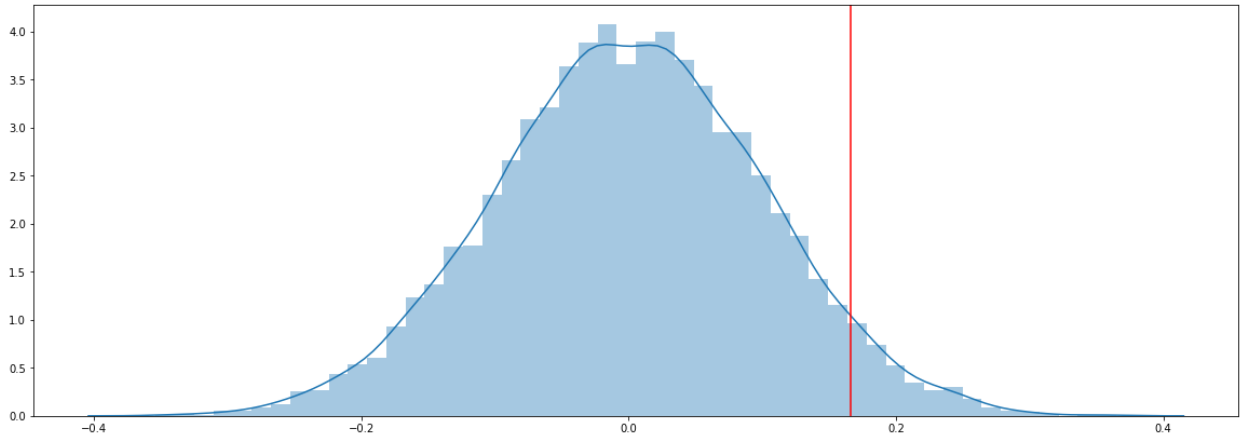


Our test returns a p-value of 0.0. Our p-value is statistically significant, thus we have to reject the null hypothesis and conclude that there is a correlation between PER and salary. In fact, it's very rare that a player below average PER will be paid a high salary. However, there are still several cases in which a strong PER performer will receive a salary comparable to players well below the average PER. 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.

So, from a player's perspective, it's worth having a high PER rating. What about PER from a team's perspective? In our data story, we noticed that for several statistics, player performances rose steadily and peaked in their late 20s and early 30s, before declining. The same trend was apparent with PER, though neither the rise or decline was as sharp as with other variables. That may partly be due to PER being averaged to 15 every year. It's worth testing whether PER is significantly different between players above and below the average age.

We know that older players get paid more than younger players for a variety of factors, including player union agreements. But is it worth for a team to invest more money in older players if they could get similar PER ratings from younger players? There's no statistic that encompasses every aspect of NBA play or can give a complete picture of team success. But PER, because of its multi-stat approach, maybe the best indicator in our dataset of how valuable a player may be for overall team success. As such, we can use the results of our test to make some sort of recommendation regarding PER.

For our null hypothesis, we assume that there is no difference in PER between players above and below the average age. The alternate is that there is a statistically significant difference. The difference of means between the groups is 0.166, which is rather small given the range and standard deviation (5.77) of PER values. The bootstrap analysis gives the following distribution:



Our resulting p-value is 0.0489. It is technically significant given our alpha value of 0.05. However, it's not overwhelmingly convincing. As such, while we can reject our null hypothesis, we do need to consider the practical implications of doing so. Since there is not a great degree of difference in PER values between older and younger players, but there is a large difference in salary, it may be more cost effective for a team to have more younger rather than older players on their roster. There are, of course, many other factors to consider, including the performance history of each player. But from a statistical perspective, younger players can offer similar PER performance to older players at a much lower cost.

Finally, there is a concerning question from these two PER analyses - salary/PER and PER/age - why is there such a large difference in salary above and below average PER, but such a small difference in PER above and below average age, especially given that older players are paid much higher salaries than younger players? There are a couple of possible explanations:

One, younger players with PER values slightly above the average may be getting paid higher for potential rather than actual performance adding to the effect that higher PER and higher salaries are connected, but not adding much to the connection between older players and higher PER values. The other possibility may be that older players with high PER and extremely high salaries, as we saw in an earlier graph, are skewing the salary/PER results to a greater extent than the PER/age results. The difference in their salary in relation to the rest of the league is far greater than the difference in their PER in relation to the rest of the league. As a result, we see a much lower p-value for the salary/PER analysis than we do for the PER/age analysis.

Machine Learning Analysis

In the previous portions of this project, we had established statistically significant correlations between player salaries and several statistics, including VORP (Value Over Replacement Player), PER (Player Efficiency Rating), and 3P% (three-point percentage), among others. We also noticed smaller correlations between salary and a few other stats. Thus, for the first part of this project, we'll attempt to predict player salary through the best regression model. However, we're not particularly

optimistic that we'll be able to develop a highly predictable model given that there are many other factors that help determine salary than just stats, as we'll discuss in more detail later.

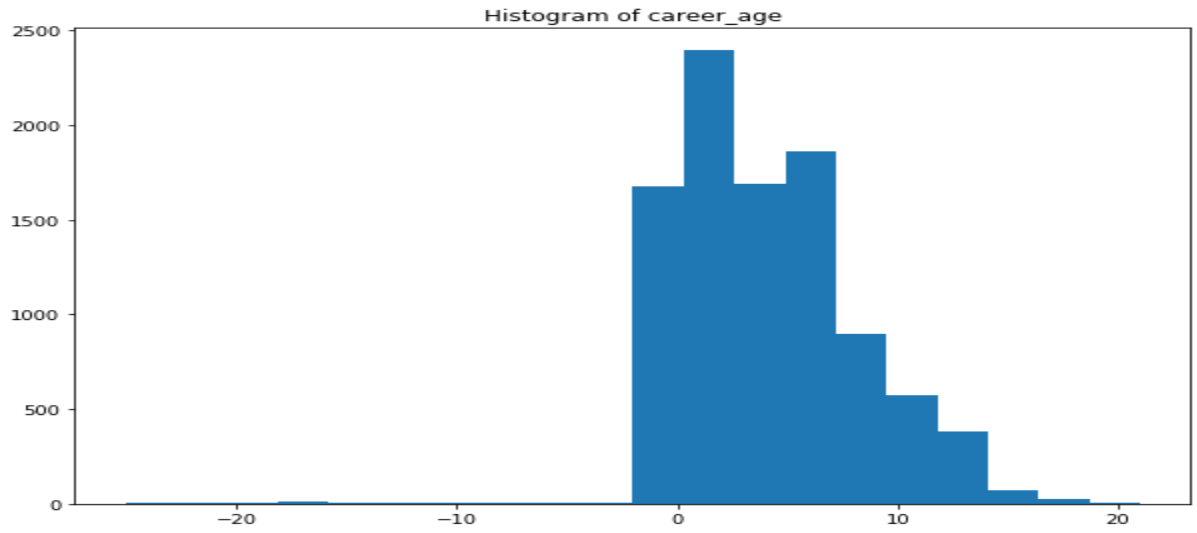
We had also established that there is a statistically significant difference between the salaries of different player positions, with point guards historically making less money on average than every other position, with centers and power forwards typically making the most. However, there are other differences between positions as well. For example, point guards tend to have more assists on average than other positions, while centers and power forwards have a higher block rate. Shooting guards and small forwards may have more balanced stats overall, while also averaging higher points per game. As such, for our second analysis, we'll attempt to predict player position from statistical performance.

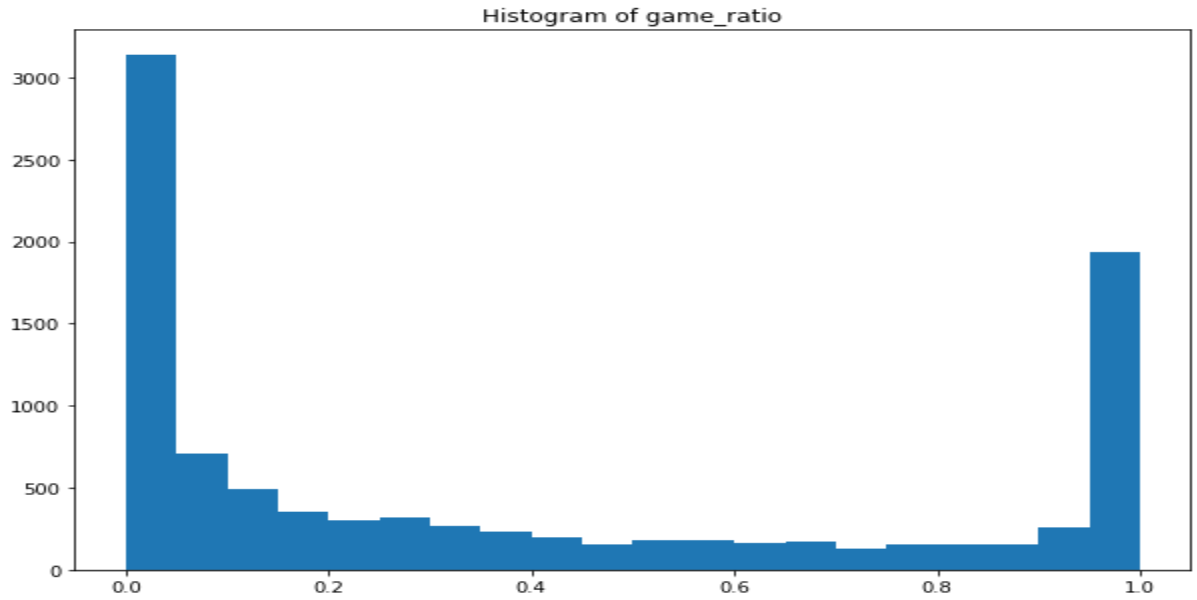
Predicting Player Salary from Statistical Performance

We should also mention that anytime we refer to 'accuracy' during our salary analysis, what were are referring to is actually the coefficient of determination, r-squared. For our purposes, and especially if this analysis is partially aimed toward a client, it is easier to understand the concept of accuracy than r-squared.

We're going to engage in a bit of feature engineering and add a couple of new columns to our dataset in hopes that they may improve our predictive abilities. The first is career_age gives us how long a player has been in the league for a given year. Players who have more experience tend to get paid more. The second column is game_ratio, which is the ratio of games started to games played. Typically, the best paid players start the most games for their respective teams.

Computing the correlations of these features to salary, we find that career_age correlates 45.7% and game_ratio correlates 51.6%. While not strong correlations, they may still help our model.





It should be noted that there are a small number of values in the career age histogram. This is due to a slight error in our data in which a very small number of players had an incorrect starting-year value. It should make a negligible difference to our results. Both histograms display an uneven distribution of the histograms suggesting that there is a good deal of variance in the data.

We have two categorical features that we would like to include in our analysis - Team and Pos. To include them, we need to innumerate them using OneHotEncoder and ColumnTransformer. We also pass our numeric variables through ColumnTransformer and then develop a pipeline that concatenates both sets and applies linear regression to training data in order to predict test data. We use a similar pipeline for several other regression models - random forest, ridge, lasso, and SVR - to determine which seem the most promising. We find that random forest is the most successful, with a test accuracy of around 65%. Yet, it also has a training accuracy of 93%, which suggests overfitting. As a result, not only do we need to increase our test accuracy, but decrease overfitting at the same time.

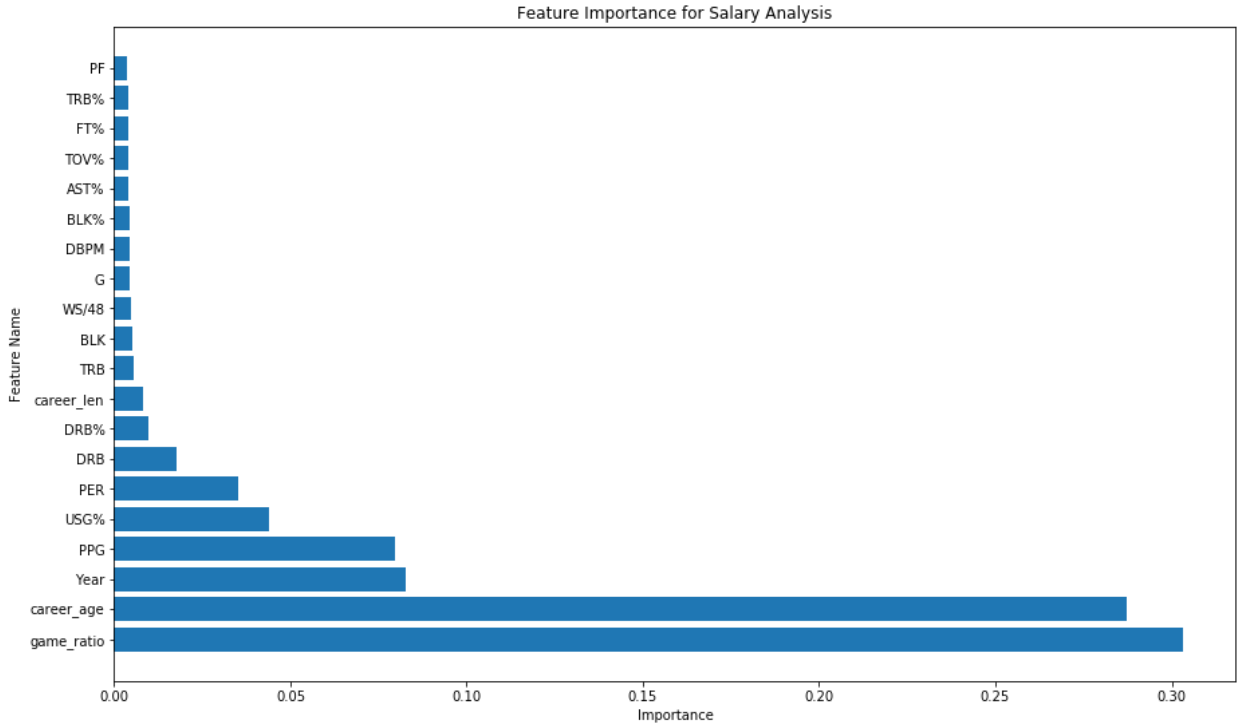
We attempt to use PCA to reduce the number of components, which have ballooned to over 80 from the original 42. We find that 70 components allow us to explain almost 100% of the data variance. However, applying PCA actually results in a lower accuracy than our initial RF model suggested. We drop PCA altogether and attempt to tune RF through trial and error, increasing our accuracy to 68%

Finally, we apply GridSearchCV to find the best combinations of parameters to maximize accuracy. It yields the following RF parameter values:

n_estimators = 50 , max_depth = 15, min_leaf_samples = 10, min_samples_split = 6.

As a result, our accuracy increases to 71%. Given that we initially started at 57%, this 14 percentile point increase through the use of feature engineering and parameter fitting is significant. And yet, 71% is still not accurate enough to predict salaries consistently.

Finally, since we passed our data through transformers and changed the parameters it's difficult to know exactly how model managed to achieve the accuracy that it did. By capturing the feature importance of our model, we can get some sense of the most important variables.



Interestingly, the two variables that we created, game_ratio and career_age, had far greater importance for our model than our preexisting variables.It suggests that the amount of time that a player spends in the league and how many games a player starts are fairly strong indicators of how well he'll be paid. Some of the other variables that we found important in our exploratory data analysis, such as PPG (another variable that we created in a previous part of this project), Year, USG%, and PER also had a noticeable difference. While we had previously found that 3PA and 3P% both had a statistically significant correlation with salary, they both barely cracked the top 30 in terms of importance.

There may be several reasons why we may never achieve high accuracy values given this dataset:

1. Recall from our previous work on this dataset that there is some bias. Namely, the dataset only includes players whose careers began in 1991 and after. So, many players who played in their early '90s but began their careers in the '80s were not included. This may slightly decrease our accuracy. However, this is probably the least likely reason for low accuracy.

2. 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 we saw in a heatmap in a previous portion of this project, most correlations with salary were fairly low, rarely being greater than 60%. 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.

3. One may suggest that another possible reason may be inflation. 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. As suggested below, the NBA often lives inside it's own little bubble where salaries get dictated by a number of factors not necessarily related to current economic climates.

4. 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.

All of these various factors make it difficult to predict salary simply from stats. In fact, without datasets that took into account many of the factors mentioned above, we likely couldn't ever make an accurate model (anything greater than 95% accuracy) for this particular question.

Predicting Player Salary from Statistical Performance

This next section on predicting player position is an interesting side analysis, but we won’t go into as much depth as with the salary analysis due to time constraints.

There are five different positions in the NBA - center, power forward, small forward, shooting guard, and point guard. There are noticeable differences between positions. For instance, point guards average more assists, while centers and power forwards have higher numbers of blocks. Also, as we found in a previous project, point guards average a lower salary than any other position, particularly centers and guards. However, we are also conducting a multi-classification analysis rather than a binary classification. As a result, matching five different positions will be more difficult than simply confirming whether a given player is a particular position.

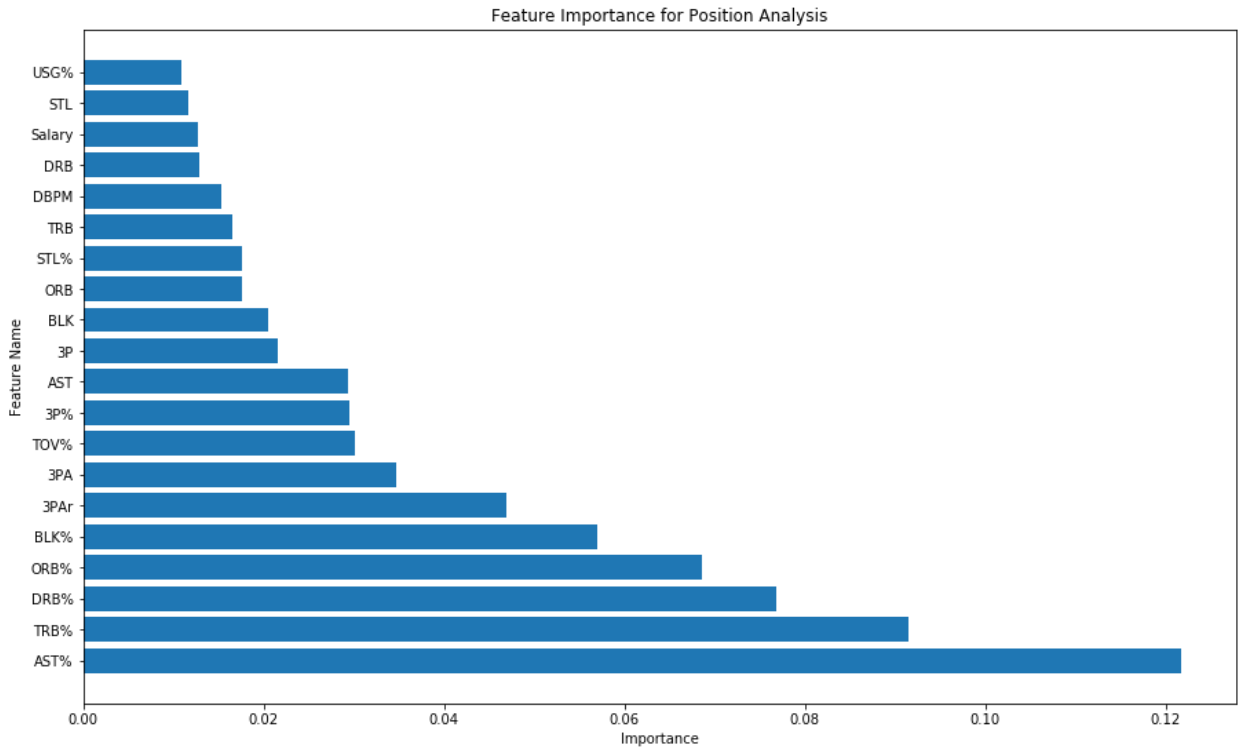
The overall procedure is similar to that used during our salary analysis. We add career_age and game_ratio variables to our dataset. We also run our data through OneHotEncoder and ColumnTransformer. We then import several classification models - logistic regression, decision trees, random forest, linear discriminant, Gaussian NB, KNN, and SVC.

Running each model analysis results in random forest, logistic regression, and SVC being our most promising models, each with around 65% accuracy. We then apply GridSearchCV to each model. For both logistic regression and SVC, due to time constraints, we only tune C. GridSearch finds that for both models, 10 is the best value for C. When we apply GridSearch to the RF model; it recommends: n_estimators=50, min_samples_split=6, and min_samples_leaf=8. After fitting our models with our new parameters, we get following accuracy scores:

Logistic regression - 67%
Random forest - 67%
SVC - 69%

SVC performs the best, but not by much. Our new random forest model also allowed us to decrease overfitting, with a 14 percentile point decrease (99 to 85) in the training score. While the test accuracies are still relatively low, the increases are promising given that we only did a small amount of parameter tuning.

Similar to our salary model, we also find the most important features for our random forest position model (due to our time constraints, it was more efficient to apply the template we used for our salary model rather than create a new process for our more successful SVC model).



It appears that many of the stats that help distinguish players of different positions are the ones with the highest importance levels. AST%, which measures the percentage of plays that a player assists o when on the court, is likely to be higher among point guards. The next three stats - TRB%, DRB%, and ORB% (all rebounding stats) - are all likely to be higher among centers and power forwards.

3-point related stats are more likely to be higher for point and shooting guards. Salary, which we had previously established as making a statistically significant difference among positions, is also one of the top 20 factors.

While we could tune our models further, it's doubtful that their accuracies would be significantly better. We could likely get our accuracies into the 70% range, but it's unlikely that we could get into the 80% range. There are several reasons why:

1. 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. If we were testing for just one position, the accuracy would be much higher, likely in the 80% range. We'll elaborate further on this point below.

2. As we mentioned at the start of this analysis, 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 last 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. In addition, there may be certain outliers, such as point guards with high salaries or centers with below average block values. All of these various factors, combined with a multi-position analysis result in lower accuracy levels.

3. 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. Had we had access to that data, our model may have been able to better distinguish between positions.

Takeaways for Players and Teams

Based on the analysis in this report, one can joke that to be paid a high salary in today's NBA, a player should be a 3-point shooting, multidimensional, center or power forward who should be near the top in a variety of different statistical measures. But this isn't that far from the truth. 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. Knowing this, a player can be more strategic about requesting a new contract versus a contract extension and in deciding whether to sign a short term versus long term contract.

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. 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.

Finally, our machine learning analysis also has important insights for players.The fact that we weren't able to predict players' salaries above 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. As mentioned above, there are a whole host of salary-determining factors that may not directly be connected to statistical output. Having that insight can prove useful for players. 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