Capstone 1 Exploratory Data Analysis

Contents

Introduction

In the previous portion of the Capstone 1 project, we noticed several interesting trends based on a graphical analysis of our NBA dataset. First, salaries have increased fairly steadily since the 1990s. Second, in the 2000s, as the NBA play style has evolved to incorporate a higher number of 3-point shots, certain 3-point-associated statistics have increased including 3-point percentage and 3PAr, which is a ratio of 3-point attempts divided by field goal attempts. Third, usage percentage appears to be higher among older players. Fourth, salaries tend to differ significantly among each of the five positions, with centers and power forwards making the highest average salary and point guards the lowest. Finally, PER or player efficiency rating, is often touted as an important statistic that determine the “best” players in the NBA, partly because it incorporates a variety of offensive and defensive variables. However, our data story suggested that there is only a small correlation between PER and player salary. It’s worth testing all of these statistics against salary to determine whether there is any correlation between them and salary. The code for all of the calculations in this report can be found in a GitHub repository at:
https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/exploratory_data_analysis.ipynb

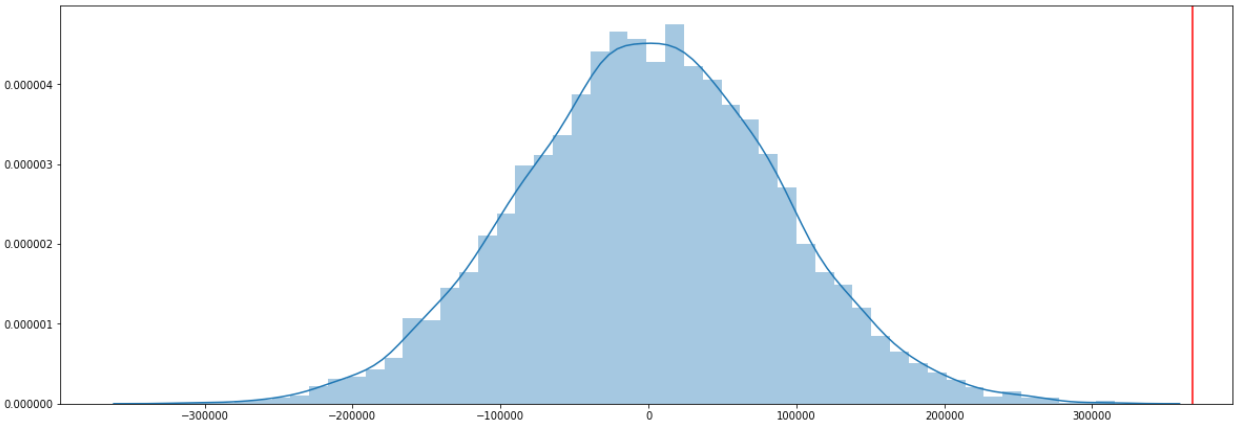
In our data story, we found that two 3-point related statistics have increased fairly steadily over close to the last 15 years - 3PAr and 3P%. In particular the average of these two variables correlated well with average salary. So, we would like to determine if there is a correlation between higher values in these 3-point statistics and higher salaries. For both, we'll perform a two sample bootstrap test two different ways. First, we'll divide groups by 3PAr and 3P%, respectively and compute the mean salary difference between those above and below the average 3PAr and 3P%. Second, we'll divide players by average salary and compute the difference in mean 3PAr and 3P%. For all of these tests, our null hypothesis will be that there is no difference between the two groups of players, regardless

of how their divided. Our alternative hypothesis will be that there is a difference. For all of our tests, our significance level will be set at 0.05.

Salary difference of means above/below average 3PAr

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.

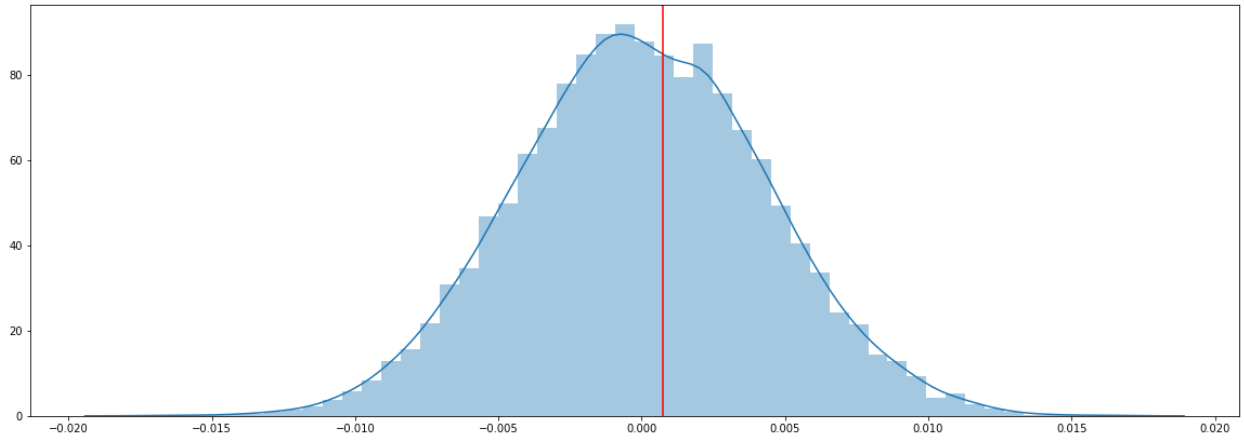
Our null hypothesis is that there is no difference in salary between these two groups. Thus, the alternative hypothesis is that We run a two sample bootstrap analysis. Our difference of means in salary is $368,052.38. Below is the distribution of the bootstrap sampling with our difference of means included 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.

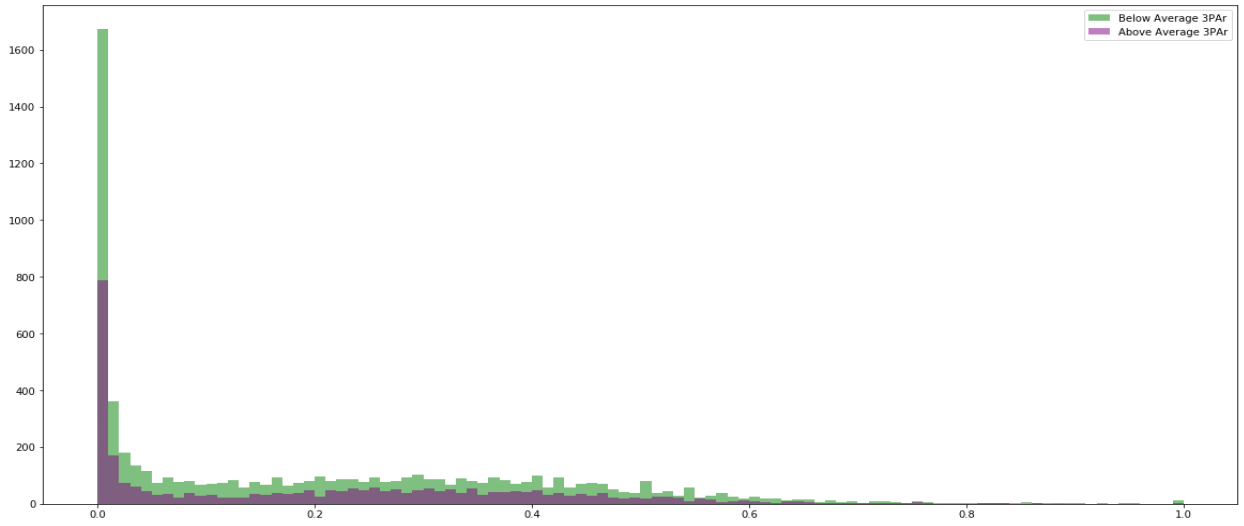
3PAr difference of means above/below average salary

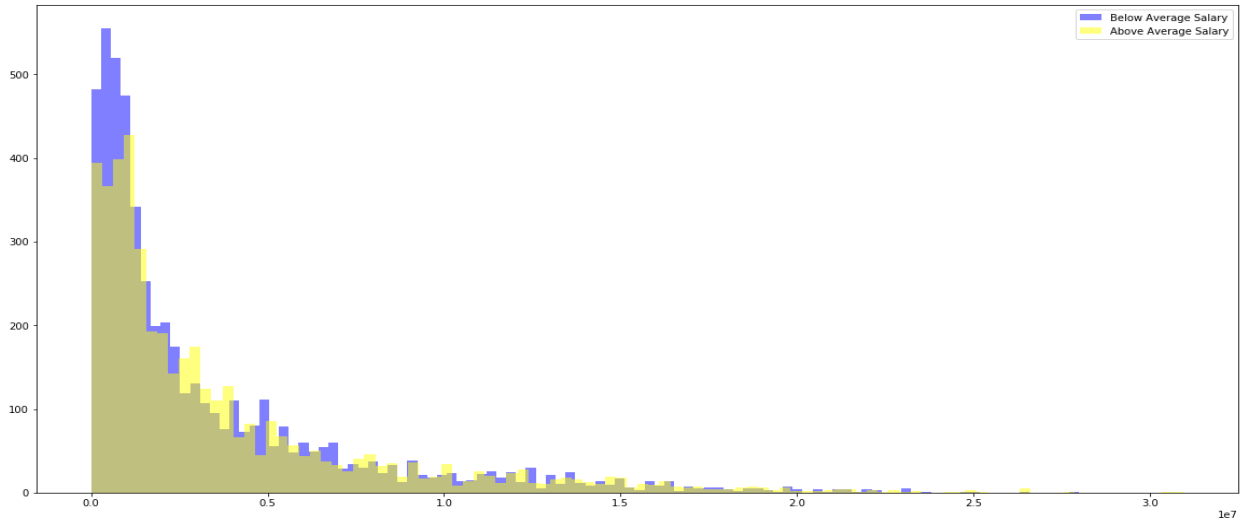
Out of curiosity, we decide to also take the 3PAr data and divide it based on average salary. The average league salary from 1991-2017 was $3,535,028. Again our null hypothesis is that there is nor difference in means between the two groups. Our result, however, is unexpected given the above result. Our difference in means is 0.000747. Below is the distribution from the bootstrap analysis:



Our p-value is 0.4293, well above the significance level of 0.05. We decide to run an independent t-test on on both sets of data and get similar results. It's not clear why the data split one way return a significant result and fails to do so if split another way.

Looking at the histograms of both groups allows for a better understanding of the discrepancy:





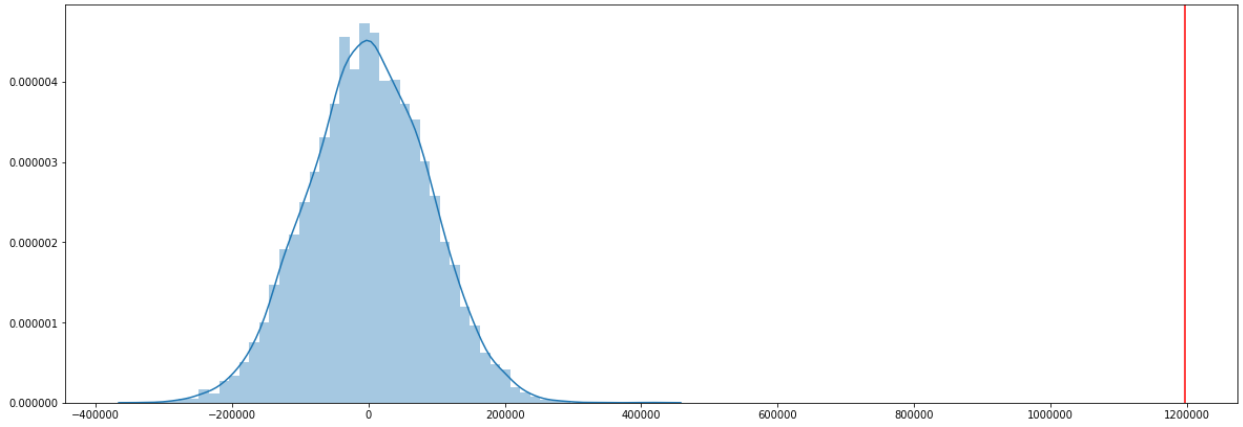
In the top graph, the two groups have a very similar shape, despite the separation in numbers. This suggests that it's less likely that there would be a significant difference in 3PAR means. In the bottom graph, there doesn't appear to be a large difference in salary between groups above and below average 3PAr. However, the shapes of the groups are quite a bit different and that difference is supported by our tests, which indicate that the difference is statistically significant. While these histograms offer a partial explanation, we're currently at a loss to fully explain why this difference in p-values occurs. Perhaps it has to do with the fact that the difference in means between the 3PAr data is very small. Overall, we can at least conclude that there is a significant difference in salary between players who are above and below average 3PAr.

Given the unusual results for 3PAr, we'll split the data two different ways for 3P% as well. 3PAr and 3P% are connected in the sense that 3-point attempts are a factor in both statistics. They have also both increased similarly over the last 5-10 years of the NBA's history, as evidenced in our data story.

Salary difference of means above/below 3P%

Three point percentage 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.

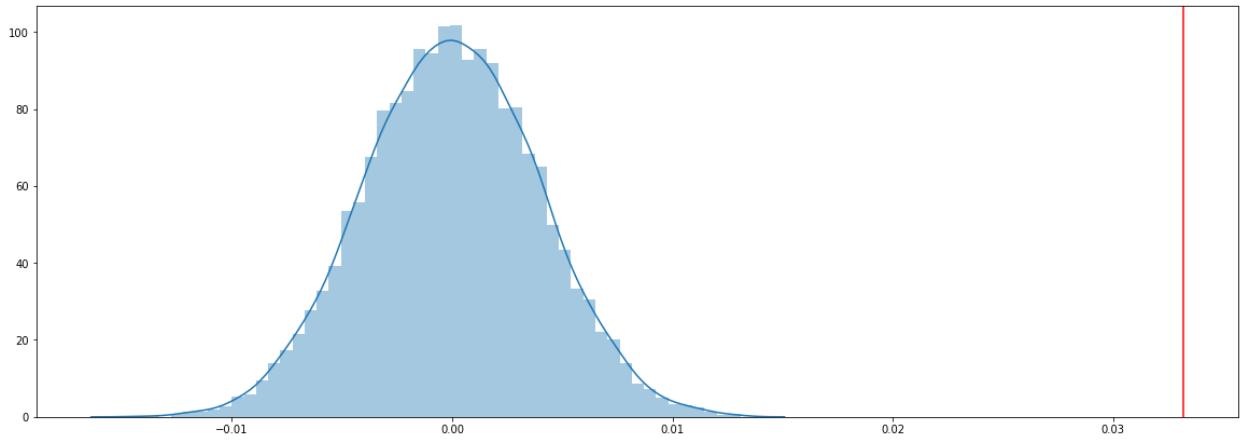
Our null hypothesis is that there is no difference in salary between the two groups, while our alternate hypothesis is that there is a significant difference. Our difference of means in salary is $1196259.38. The bootstrap analysis results in the following distribution:



Our resulting p-value is 0.0, allowing us to reject the null hypothesis.

3P% difference above/below average salary

Our null hypothesis is that there is no difference in means in 3P% for players above and below the average salary. The difference of means in 3P% is 0.0332. This difference is much larger than the difference in 3PAr for players above and below the average salary. The following is the distribution of the 10,000 bootstrap replicates:



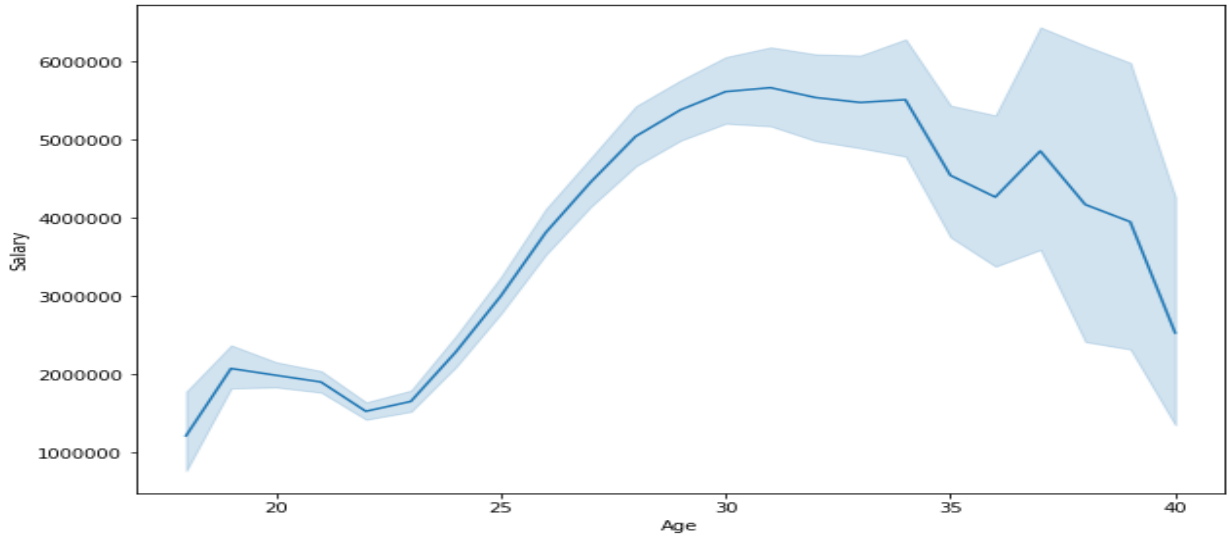
Again, our resulting p-value is 0.0. Thus, we reject the null hypothesis. Regardless of how we divide the data, our results are statistically significant, and we can say that there is a connection between being a strong 3-point shooter and making an above average salary.

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

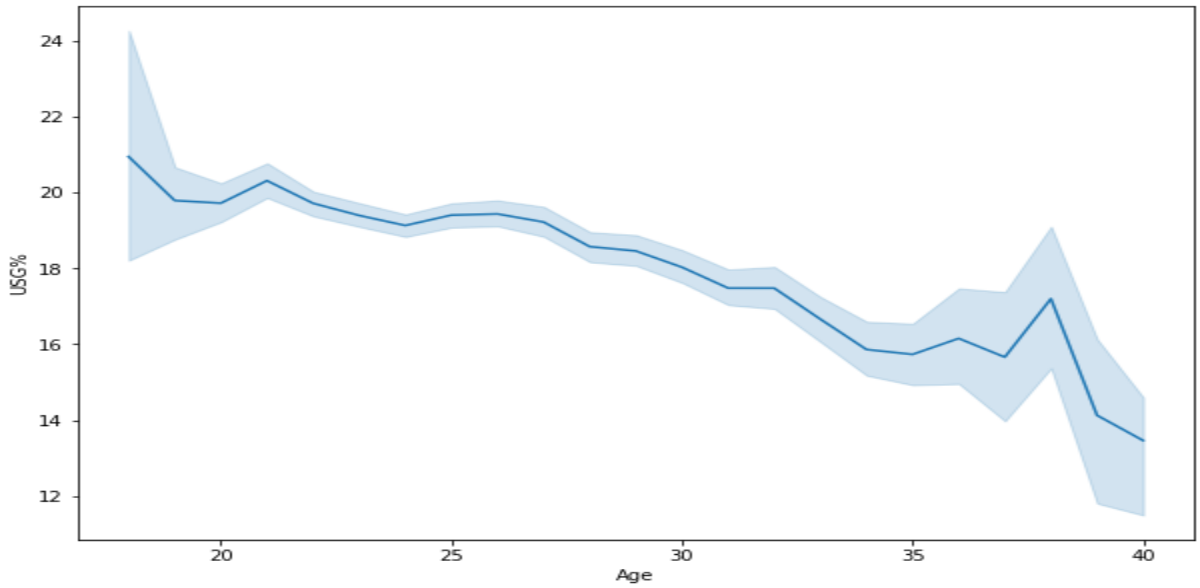
important the 3-point shot has now become, even centers and power forwards would benefit from becoming better 3-point shooters. As we'll see with our PER analysis, the versatile and multi-dimensional a player becomes, the more they benefit financially.

Salary difference above/below average USG%

Our data story also suggested that usage percentage is lower among older players. Older players, having been in the league longer, get paid a higher average salary than younger players, implying, curiously, that usage rate should be lower among players who get paid higher salaries. As we can see from our graphical analysis, older players command higher salaries:

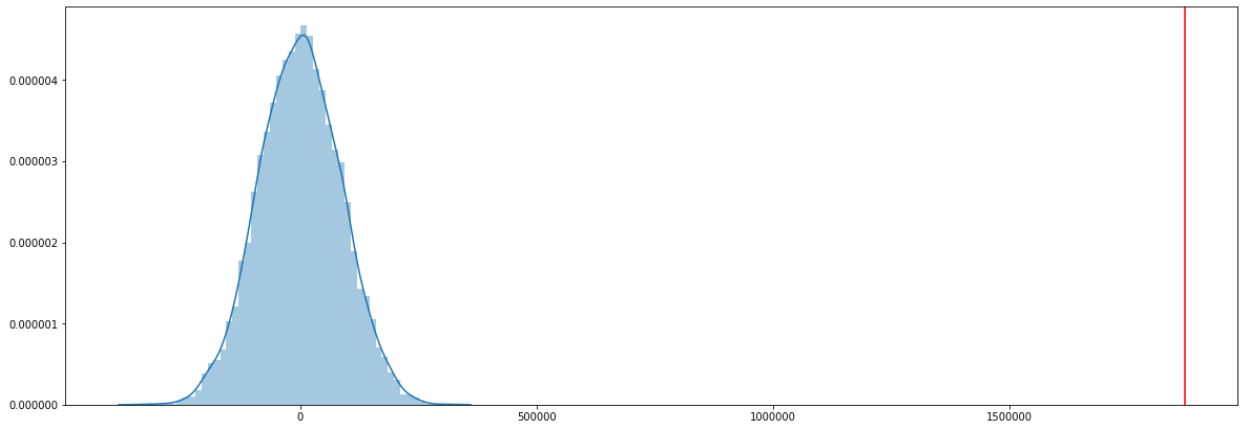


Yet, their USG% is far lower than younger players:



USG% is a fairly complex formula that takes into account field goals and free throws attempted, turnovers, and minutes played. For all variables, both a player's and team's values are included. In essence, USG% attempts to determine how involved a player is while they are actually on the court.

For our test, our null hypothesis is that there is no difference between USG% and salary. Our alternative hypothesis is that there is a significant difference. We divide our data based on players who are above and below the average USG%. Our difference of means between the two groups is $1,870,136.92. Running our bootstrap analysis of 10,000 replicates results in the following distribution:



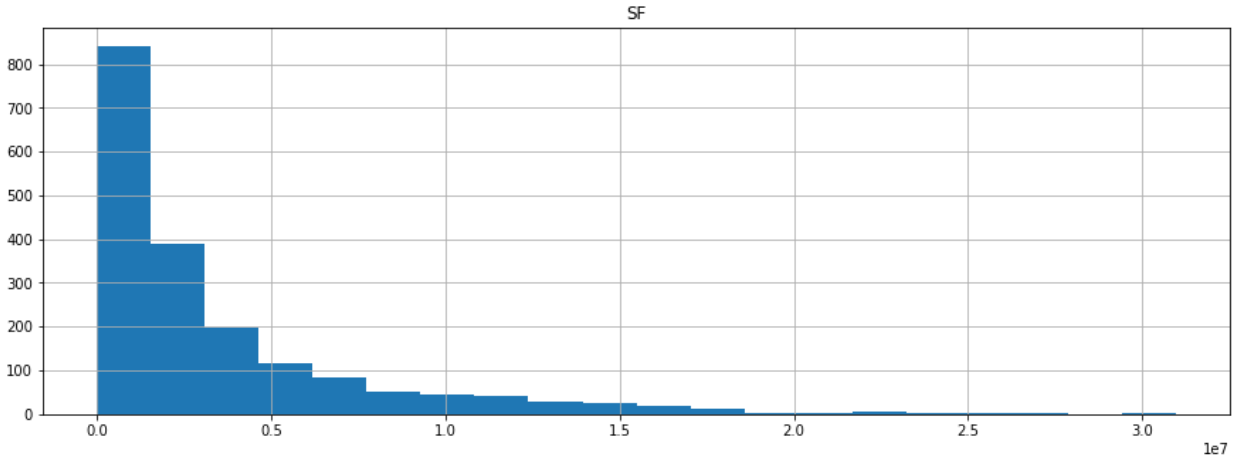
Our p-value is 0.0, allowing us to reject our null hypothesis. Despite our initial graphical analysis, our difference of means indicates that average salary is higher among players who have a higher USG%. There is a connection between USG% and salary, just not the connection that our original analysis had lead us to believe. It makes sense that players who receive higher salaries have higher usage percentage values. USG% takes into account not only minutes played, but also how involved a player is while they are on the court. More experienced players are involved at a higher percentage, play more minutes, and have larger salaries. Obviously, teams want to get the most out of a player if they are going to be paid higher salaries. But it's also true that strong players benefit their teams more while on the floor.

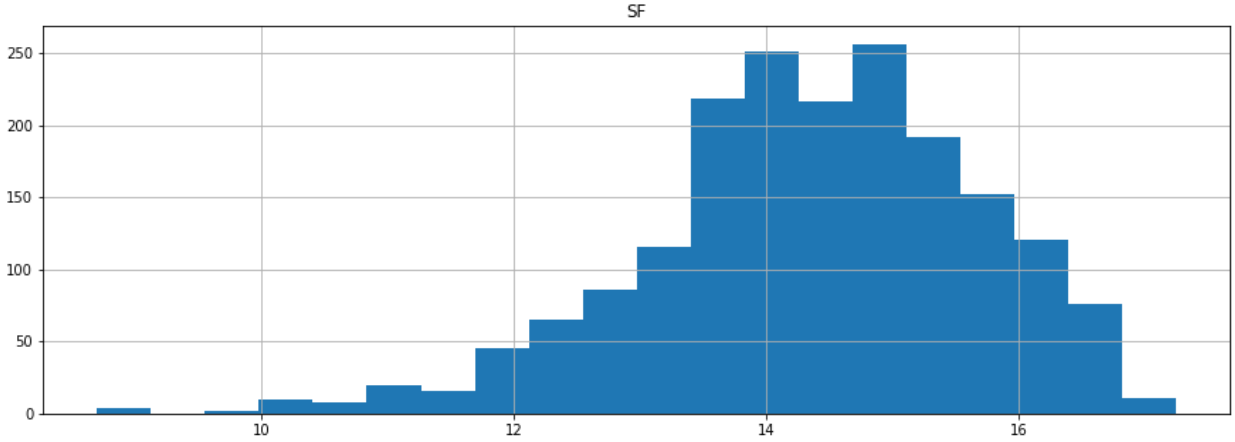
ANOVA test of salary versus position

We're also interested in determining whether there is a significant difference in salary based on position. A graphical analysis in the data story had indicated that point guards appeared to be paid the least and had the lowest amount of variation, while centers and power forwards had the largest variation. In order to compare all 5 positions together, we'll be conducting an ANOVA analysis.

For an ANOVA test, data has normally distributed. An initial graphing of the salary distribution of each position showed histograms that were heavily right skewed, which is to be expected given that the overall salary distribution of the entire league is also right skewed. A log transformation of each data set showed more normally distributed data, allowing us to proceed with the ANOVA test. Below

are histograms of the small forward salary distribution before and after the log distribution, respectively:





An initial description of the five positions gave the following results:

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

Confirming our initial analysis, power forwards and centers have the highest average salaries as well as the largest standard deviations. Point guards have the lowest values in each of those two categories. Running a regression results in the following:

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	3.796e+06	9.48e+04	40.039	0.000	3.61e+06	3.98e+06
C(Pos)[T.PF]	7.353e+04	1.33e+05	0.552	0.581	-1.88e+05	3.35e+05
C(Pos)[T.PG]	-6.954e+05	1.37e+05	-5.088	0.000	-9.63e+05	-4.28e+05
C(Pos)[T.SF]	-3.452e+05	1.36e+05	-2.535	0.011	-6.12e+05	-7.83e+04
C(Pos)[T.SG]	-3.891e+05	1.35e+05	-2.876	0.004	-6.54e+05	-1.24e+05

With the center position acting as a control, we see that the p-values in relation to it and the other positions, except power forward, are significant, giving evidence that there is a connection between salary and position.

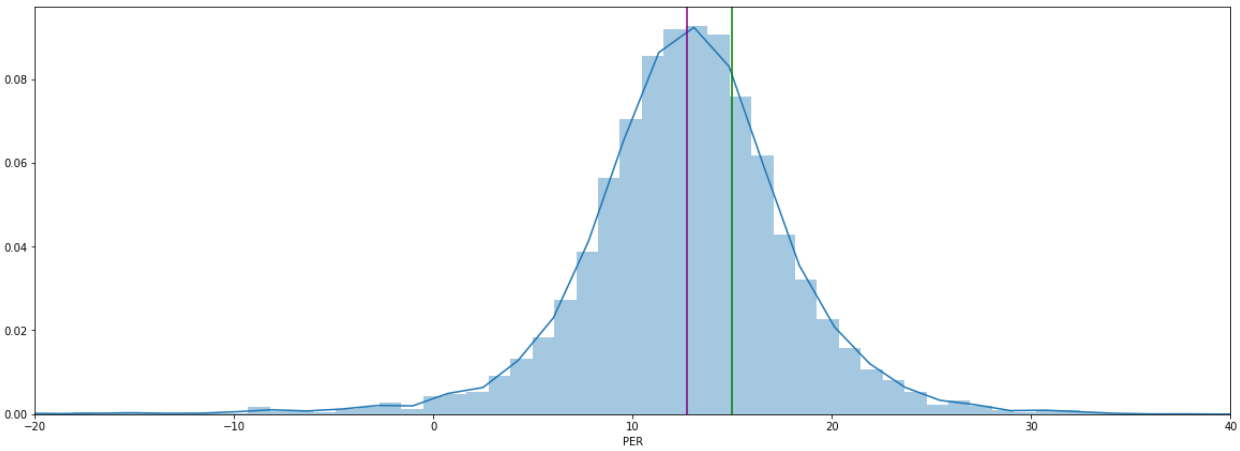
Running an ANOVA test confirms that we should reject the null hypothesis:

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

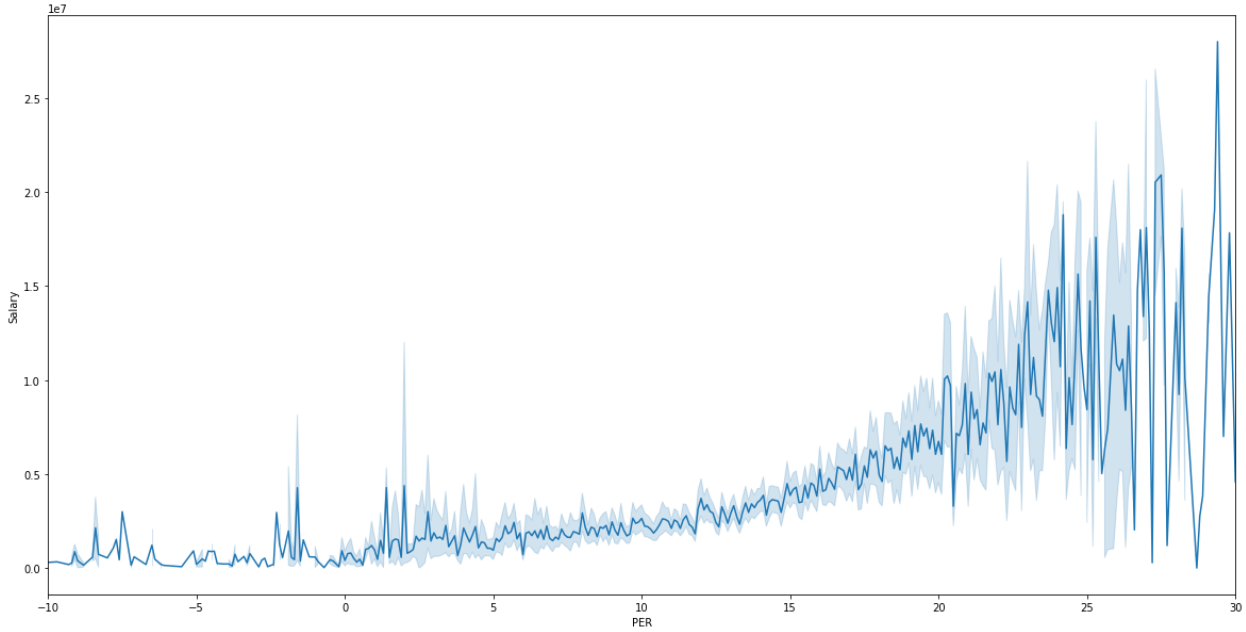
The sum of squares gives the amount of variance in the data. In this case, there is a great amount of variance. Paired with the extremely low p-value and the F-stat, the ANOVA test indicates that we should reject our null hypothesis. We can conclude that there is a significant difference in salary between positions. It's likely the case that position does influence salary (in addition to other variables, as we've seen). Centers and power forwards are paid more than other positions, while point guards are paid the least, by a wide margin. As we stated in our data story, 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.

PER analysis

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.

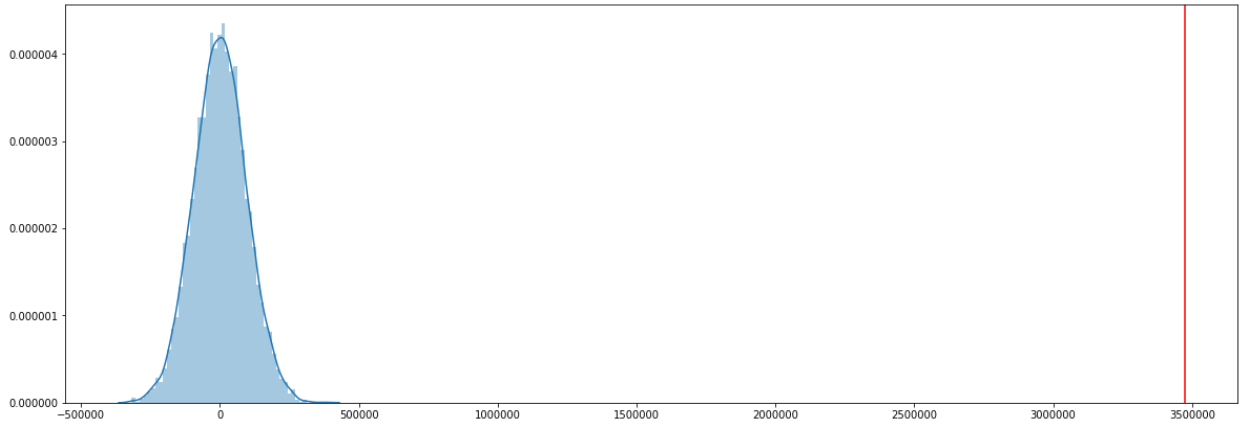


For most PER values salary is fairly stable. There is large increase after 15. Yet, the most noticeable aspect is the drastic increase in variance after 20. It's not clear why there is such a huge variance. It's likely due to a small group of players with high variance having large salaries well above 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:



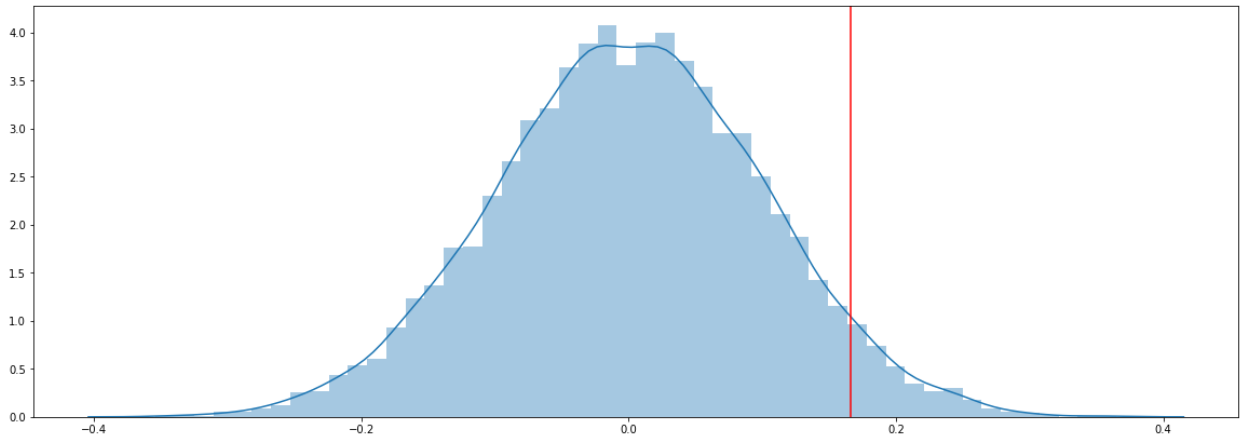
Our test returns a p-value of 0.0. Our p-value is statistically significant, thus we have to reject the null hypothesis and conclude that there is a correlation between PER and salary. In fact, it's very rare that a player below average PER will be paid a high salary. However, there are still several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

PER difference of means above/below average age

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. Luckily, for older players, this isn't the only factor that teams 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, 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.