Capstone Project 1 Data Story - NBA Salaries and Player Performance

This Capstone 1 project is concerned with analyzing the relationships between NBA players' salaries and performances between the years of 1991 and 2017 through graphing. A previous data wrangling project took three datasets - salaries, career length, and performance statistics - and combined them into one dataset which contained no null values or duplicate data. For this stage, we'll be taking the dataset and exploring potential relationships graphically.

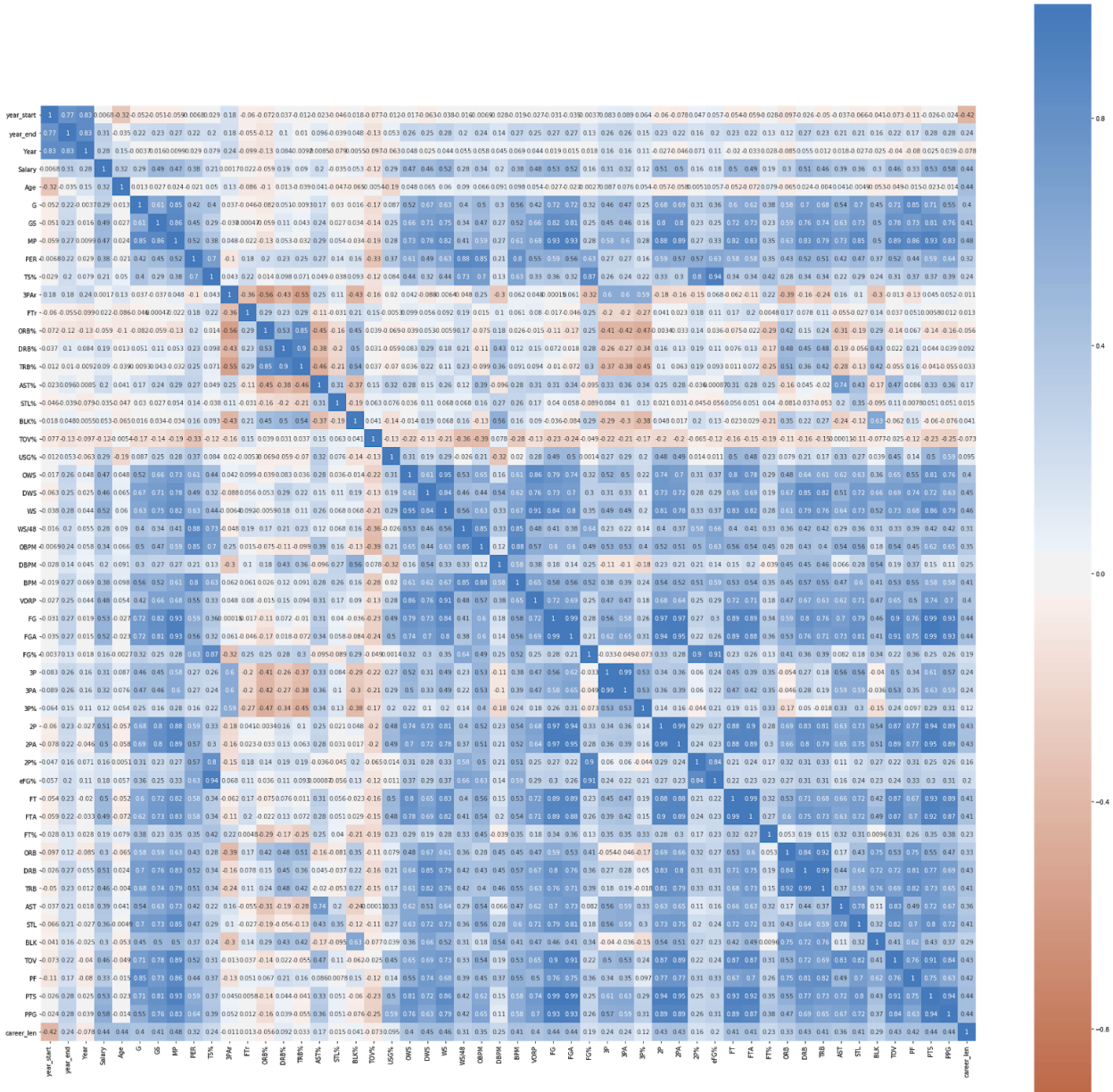
The Jupyter Notebook file of this data story, which includes all of the code used to generate the following graphs can be found on GitHub:
https://github.com/kjd999/Springboard-files/blob/master/Capstone%20Project/nba_data_story.ipynb

There are several questions that can be asked :
1) Is there a trend for salaries over the course of years, and is that trend upward?
2) Is there a connection between salary and position?
3) What does a line plot that shows individual player salaries over time or in relation to specific stats indicate?
4) What connections are there between the various stats such as PPG, VORP, PER, etc and salary?
5) Do salaries increase/decrease as performance increases/decreases?
6) Have average player statistics risen over the years at the same rate as salaries or even at all?

The expectation is that average salaries have increased from 1991 to 2017, not only due to inflation but also because the league has increased the salary cap over the years, allowing teams to spend more money on players. However, it's likely not the case that player statistics have increased at the same rate as salaries. Despite rule changes over time that favor offensive performance, average player statistics may not have increased as dramatically as salaries. This in part may be due to obvious physical limitations and built-in limitations to certain advanced metrics such as PER (player efficiency rating).

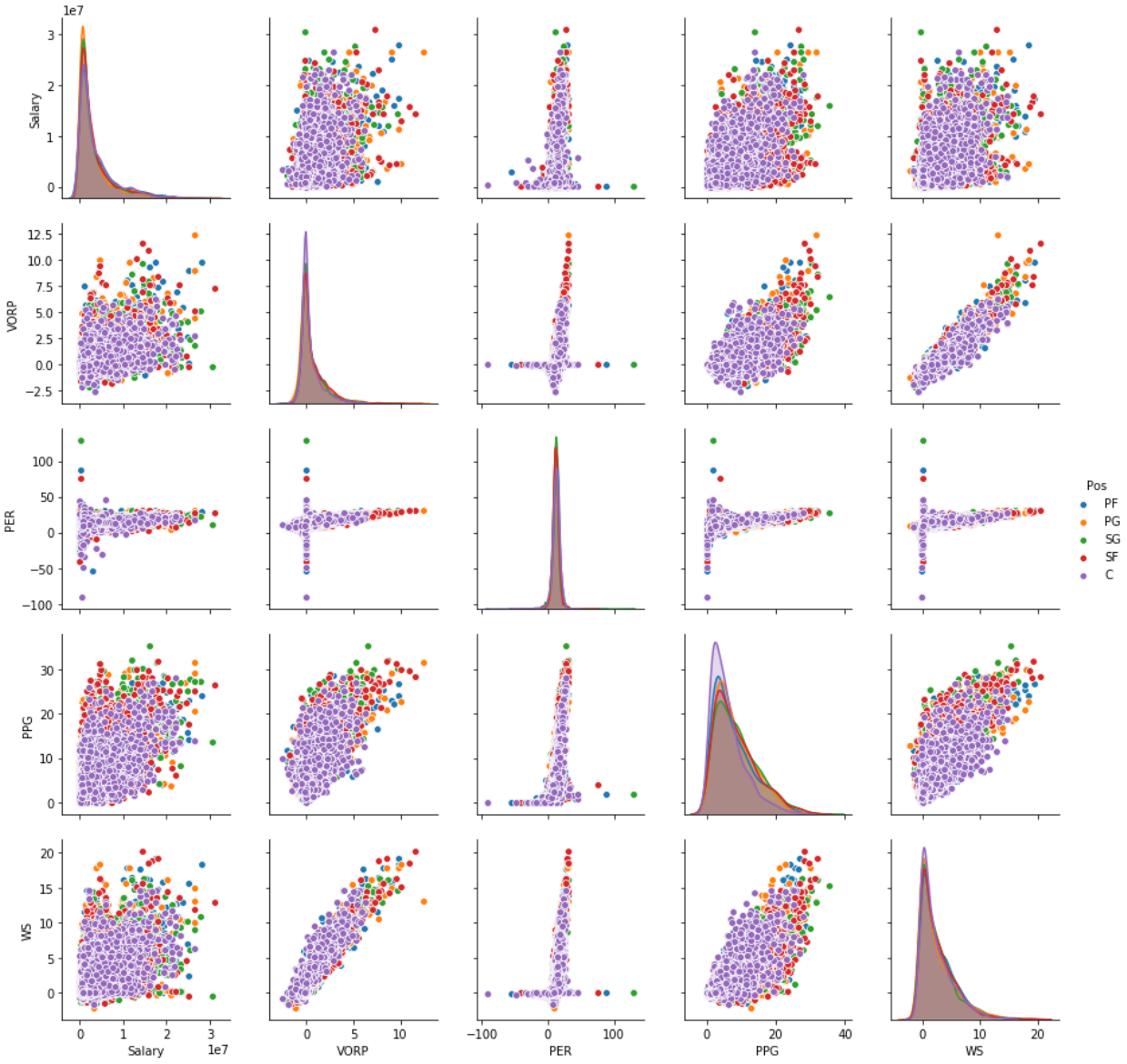
More importantly, however, is what happens to player salaries over the span of their careers. Here, the expectation is that salaries will start low as a player first enters the league, peak at a certain age as performance continues to improve, and then drop as a player's skill gradually diminishes. This expectation does run counter to headlines that often point out how veteran players who are in the so-called twilight of their careers appear to be signing to ever high salaries despite their decreased performance. Such players are likely the exception rather than the rule.

It's hoped that through this analysis, we can get a better sense of how salaries increase and decrease over a player's career and how that's related to player performance.



We begin by creating heatmap of the original dataset to see if there are any strong correlations between the various columns, particularly in relation to salary. It doesn't seem to indicate salary having a high correlation with any particular variable, though there does appear

to be some correlation with G, GS, WS, VORP, FG, FG%, PTS, and PPG. It makes sense that players that start more games will have higher salaries. Good performances lead to both increased play time and increased pay. And, in turn, if a player is playing more often, that leads to increased scoring - both overall and per game.

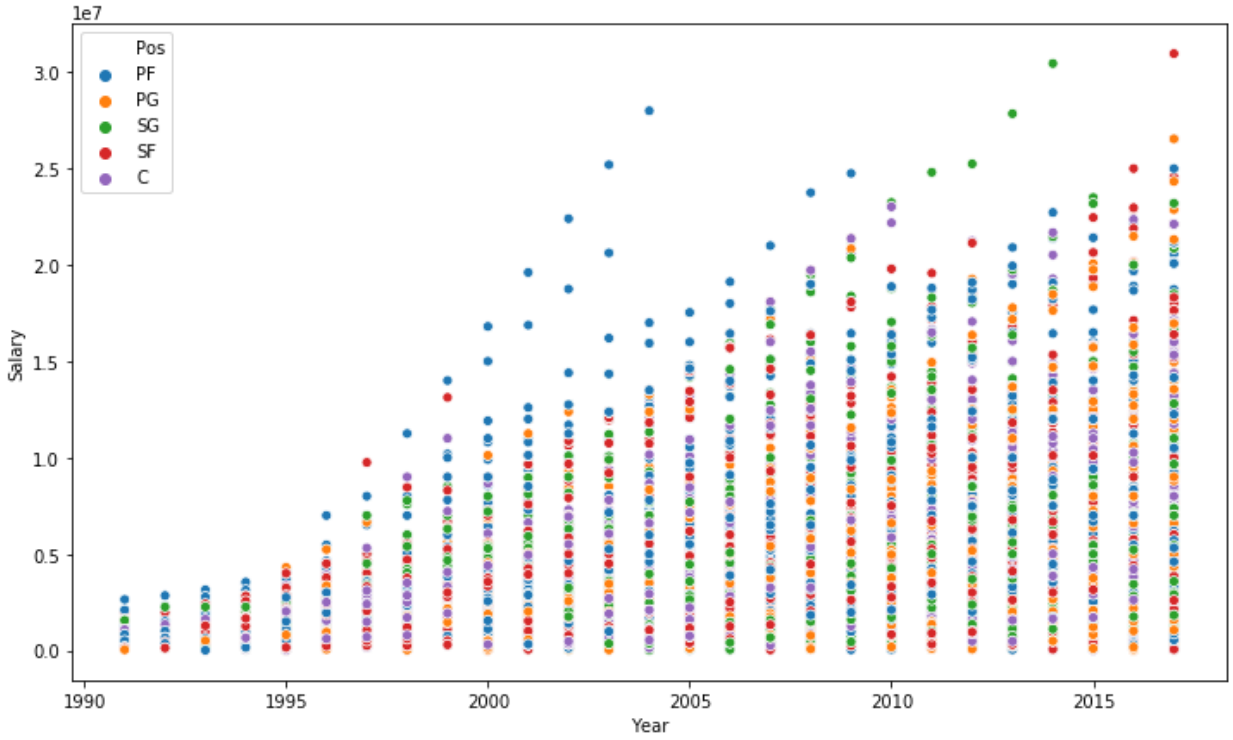


Creating a pairplot with salaries, VORP, PPG, PER, and WS, we see several interesting correlations. Centers tend to be more clustered than other positions for most of the stats, indicating a smaller range. All of the stats, other than PER are fairly right-skewed, which suggests that most players fall below the average and median. PER, meanwhile, seems not to

correlate well with salary or any other stat. That may indicate that it isn't a strong consideration when teams decide how much to pay their players.

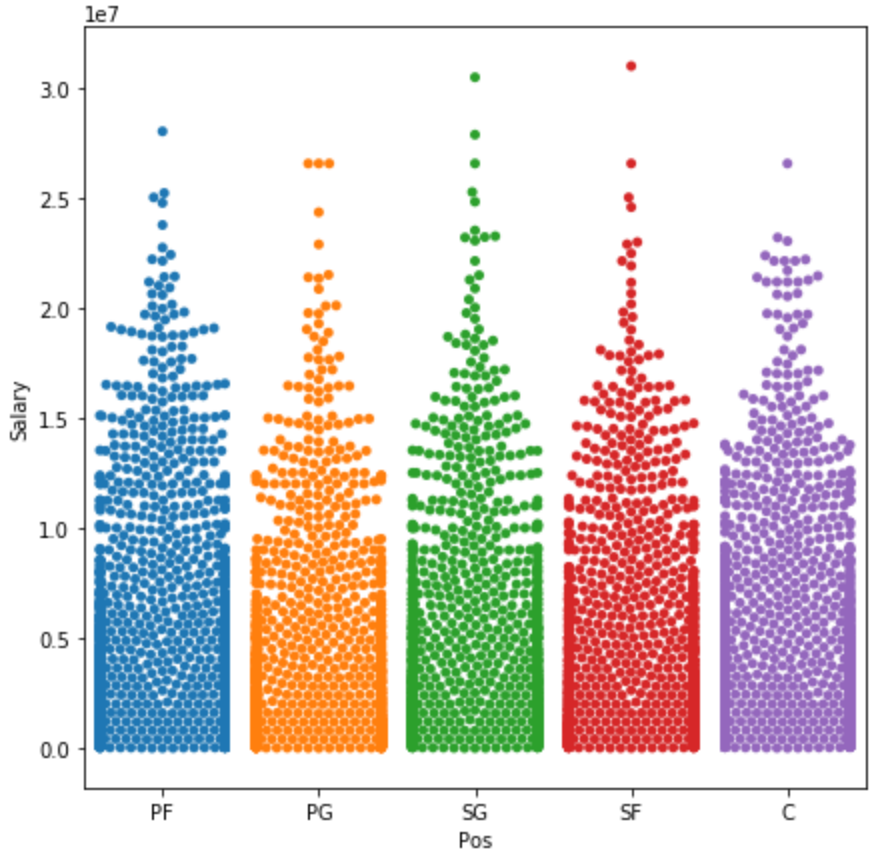
1) Is there a trend for salaries over the course of years, and is that trend upward?

Beginning with the first of the questions mentioned above, we plot salaries over time as scatterplot, which is hued by position.

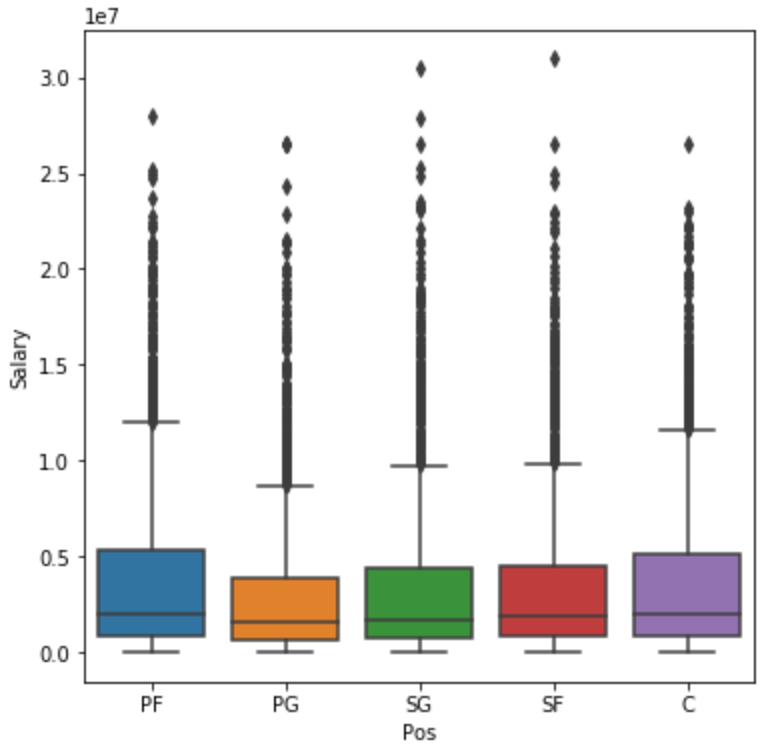


It indicates that salaries have increased steadily from 1991 to 2017, with a slight dip after 2010 and a large increase in 2017. That increase is likely due to increased spending cap limits. Before 2005, it appears that most of the outliers were PFs. More recently, it appears that shooting guards are the outliers. We'll need to plot out different graphs to answer our next question about positions.

2) Is there a connection between salary and position?

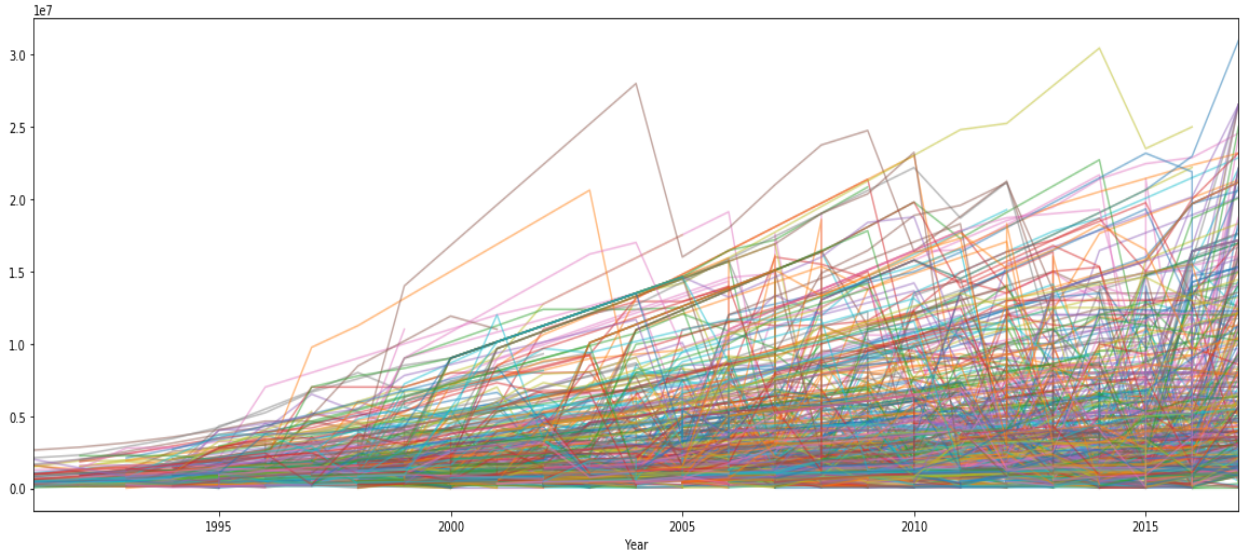


Using a swarmplot, we can also see how salaries are distributed by position, answering our second question. It indicates that for every position, the majority of yearly salaries have been below $15 million, with point guards have fewer and less extreme outliers.



A box plot of the same situation clarifies matter further, indicating that point guards have been paid lower salaries as their boxplot is more condensed than the other positions. It also has less extreme outliers. It's not immediately clear why this is the case. Perhaps it is because teams carry more guards than any other position, driving down their salaries, or it may be due to other positions being less expendable. It's likely beyond the scope of this data to determine the actual answer.

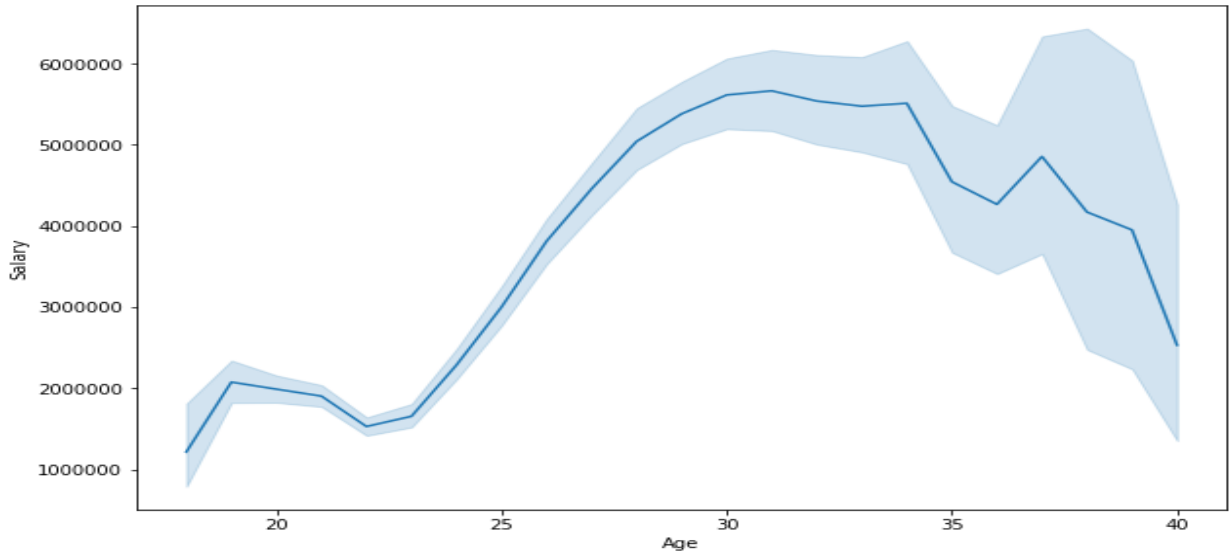
3) What does a line plot that shows individual player salaries over time indicate?



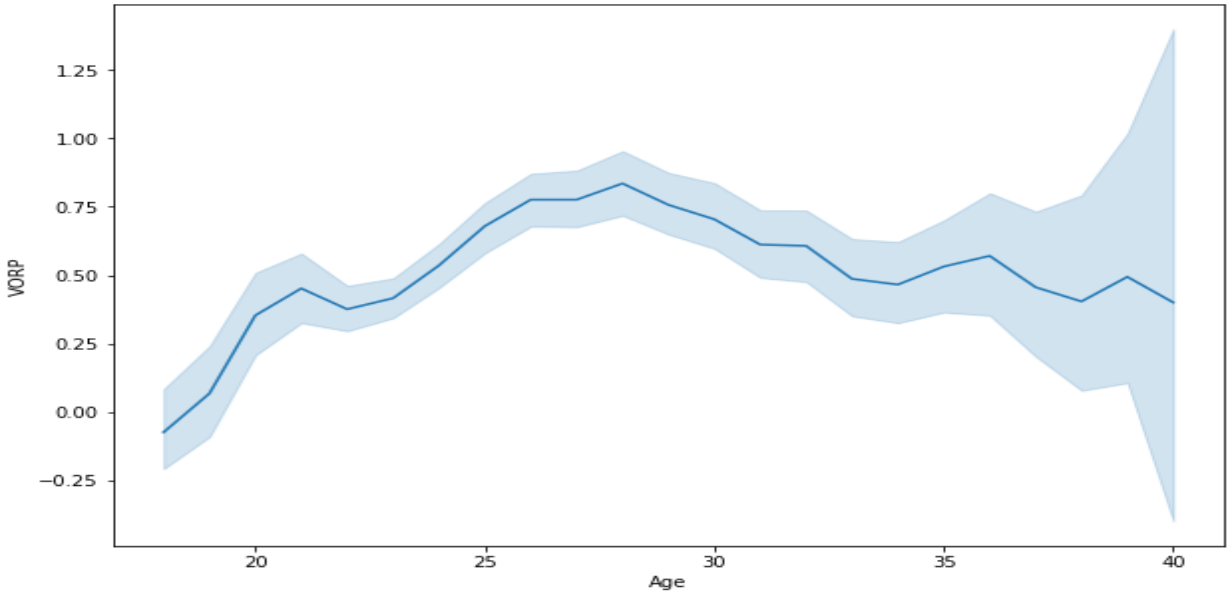
To answer our third question about individual salaries over the course of a career, we graph a line plot of salaries against years. Though it is a bit hard to decipher, due to so many overlapping lines, we can see that player's salaries do rise and fall over the course of a player's career. That is, it isn't always the case that a player's salary will continue to increase over the course of their career. As they age, and their skills diminish, many players are likely going to be signed to smaller contracts than those when their skills were at their peak. However, it also indicates, as previous graphs have indicated, that salaries do rise overall closer to 2017.

4) What connections are there between the various stats such as PPG, VORP, PER, etc and salary? 5) Do salaries increase/decrease as performance increases/decreases?

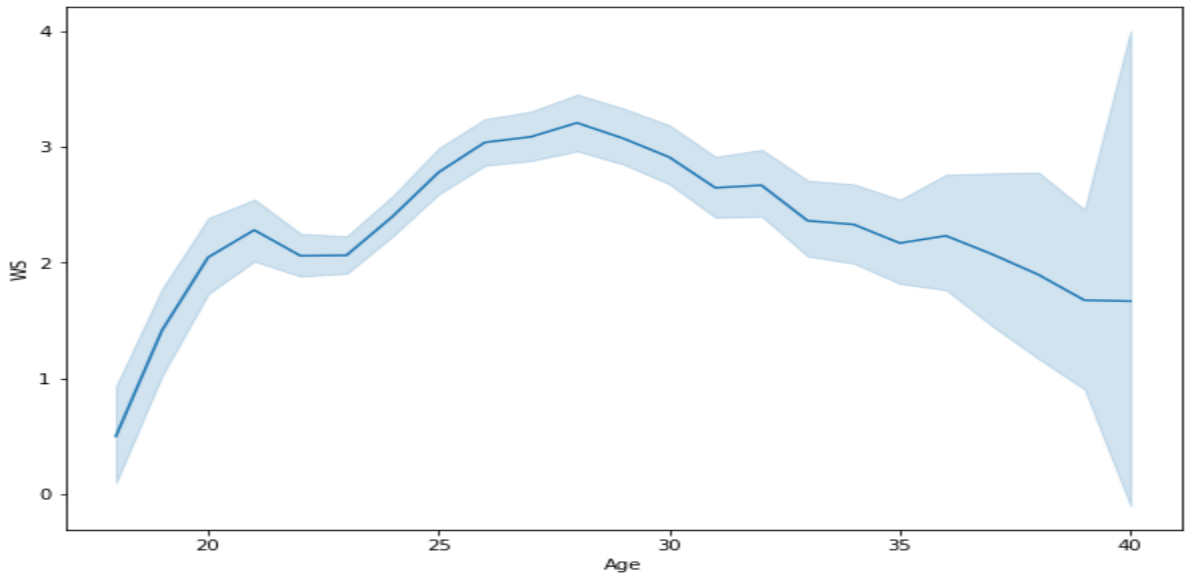
Getting back to some of the other statistics mentioned earlier, such as VORP, PER, WS, and PPG, we've seen from our heatmap and pairplot, that there does seem to be some correlation between those stats and salary. However, it may be more informative to plot those stats and salary against age to get a better idea of the connection between them and salary.



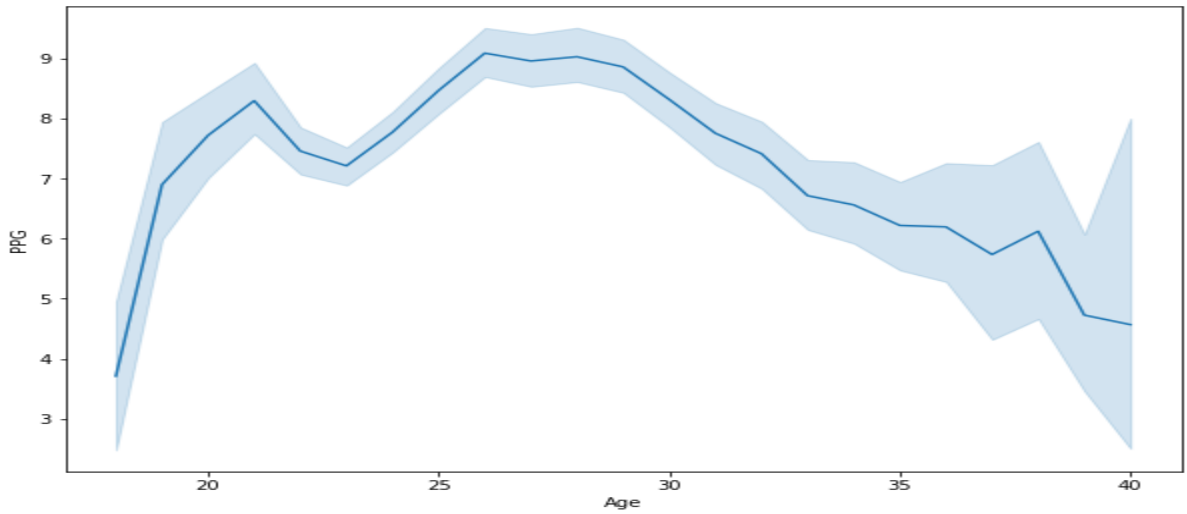
Graphing age vs salary reveals that salaries increase quite a bit as players reach their late 20s and early 30s. With the exception of a few outliers, salaries then begin to drop after about the age of 34 and continue to do so as players approach 40.



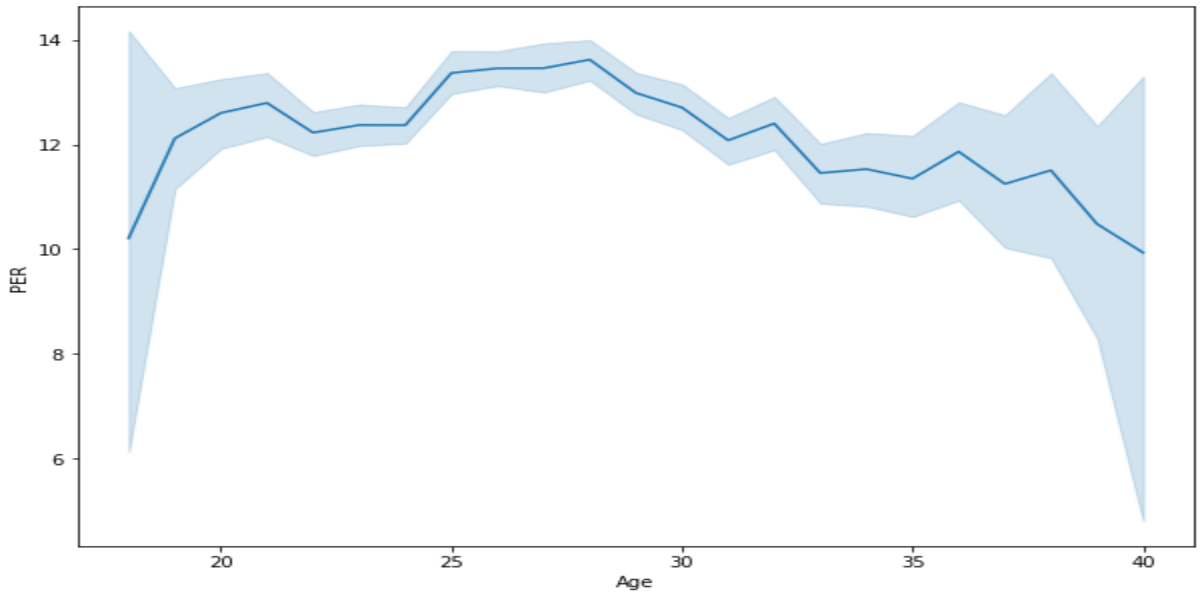
VORP (value over replacement player) peaks in the late 20s and continues to fall through the rest of most players' careers.



WS (win shares) is highly correlated with VORP, so it's not surprising that their graphs look very similar.



PPG (points per game) peaks even earlier than the previous stats and falls at an even steeper rate after 30.



PER (player efficiency rating) 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 30.

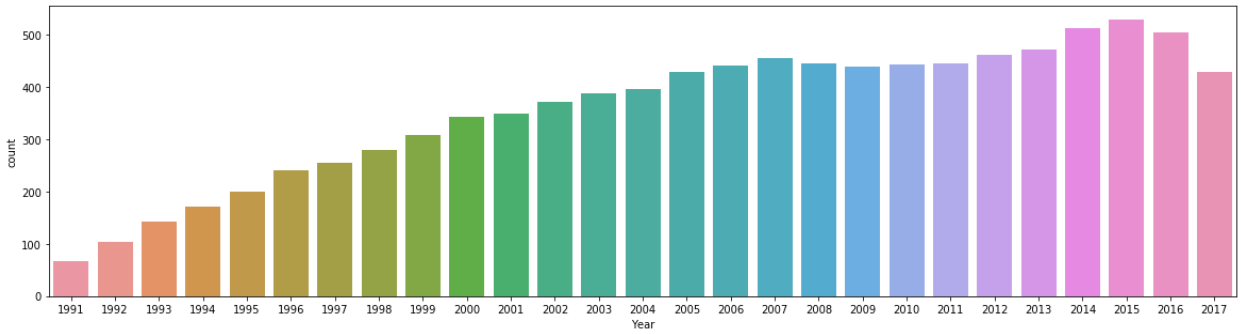
It appears, based on the above graphs of salary and stats against age, that careers peak in the mid to late 20s, but that salaries peak in the early 30s. That likely indicates that players get paid higher salaries once they have proven to be strong performers. The lag between performance increase and salary increase is likely due to older, less lucrative contracts needing to expire before more lucrative contracts can be signed, as well as teams needing to reassure themselves that players are proven performers.

There are a couple of possible takeaways here for both players and teams. For players, it may be better to sign shorter contracts in their early to mid 20s while their skills are at their peak in order for them to have a chance at more lucrative contracts as early as possible. And, then, to sign onto long term lucrative contracts knowing that their skills will soon begin to diminish.

For teams, it's important to realize how skills diminish over time. Looking at some of the above graphs, younger players perform almost as well, and in some cases better than older players. Given the fact that younger players are more affordable and that performance does not seem to be drastically lower, it may be smarter for most teams to invest in younger players rather than veterans.

Before we move onto answering our last question, we need to address a bias in the dataset. In a previous data wrangling project, we created our current dataset by dropping any players that had not begin their careers in 1991 or after. This was done due to the limitations of a salary dataset that only began in 1991. The consequence of this is that our populations for the

1990s are far smaller than subsequent years because older players are not included in that decade. As a result, many prominent players such as Michael Jordan, Magic Johnson, and Patrick Ewing are not on this list. We can graph a count plot to see the difference in yearly population size.



As previous graphs have indicated, most veterans (up to a certain age) make more money than their younger counterparts. As the graph above indicates, the sample sizes of players in the 1990s are far smaller than subsequent years. Thus, it's one of the reasons that average salaries in the 1990s are lower than in the 2000s. However, due to inflation and increased salary caps, salaries in the 1990s would be lower regardless, just not as dramatically as these graphs suggest. By 2005, the yearly population appears to have stabilized, and we can see that salaries are still lower in most years compared to the most current years in the dataset. So, it's safe to assume that salaries in the 1990s were still lower than those in this decade.

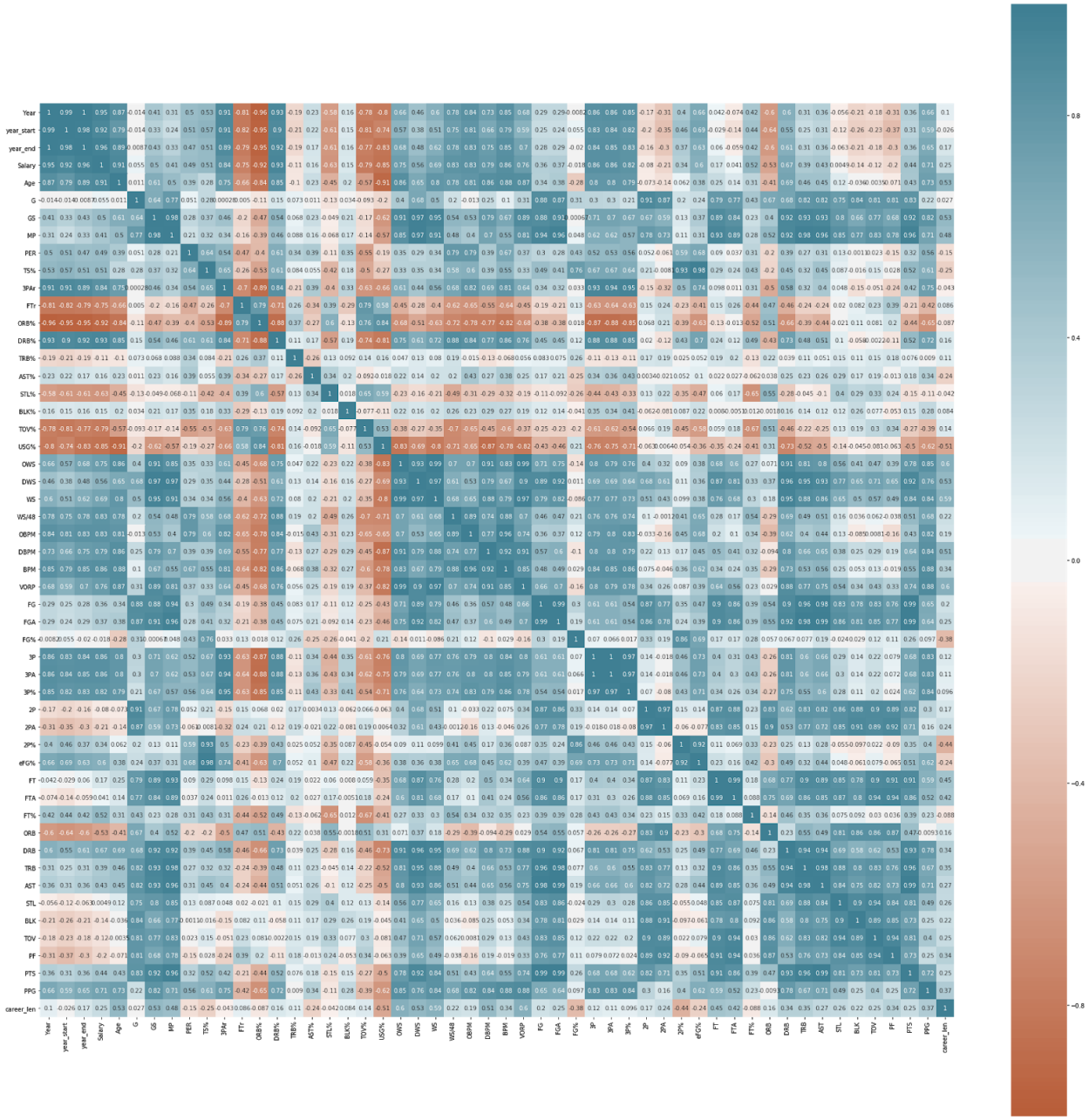
6) Have average player statistics risen over the years at the same rate as salaries or even at all?

Thus far, we've seen that salaries have increased quite a bit since 1991, in spite of the aforementioned biased in this dataset. However, have player performances increased at the same rate? There are, of course, human limitations that only allow for a certain amount of improvement. At the same time, athletes have become more athletic, our training techniques have improved, and the way that the game is played has changed. Regarding this last point, due to rule changes, the current style of play does favor offenses rather than defenses and 3 point shots are much more common than they used to be. So, it'll be curious to see whether certain metrics have increased over the years, and if they've increased similarly to salary.

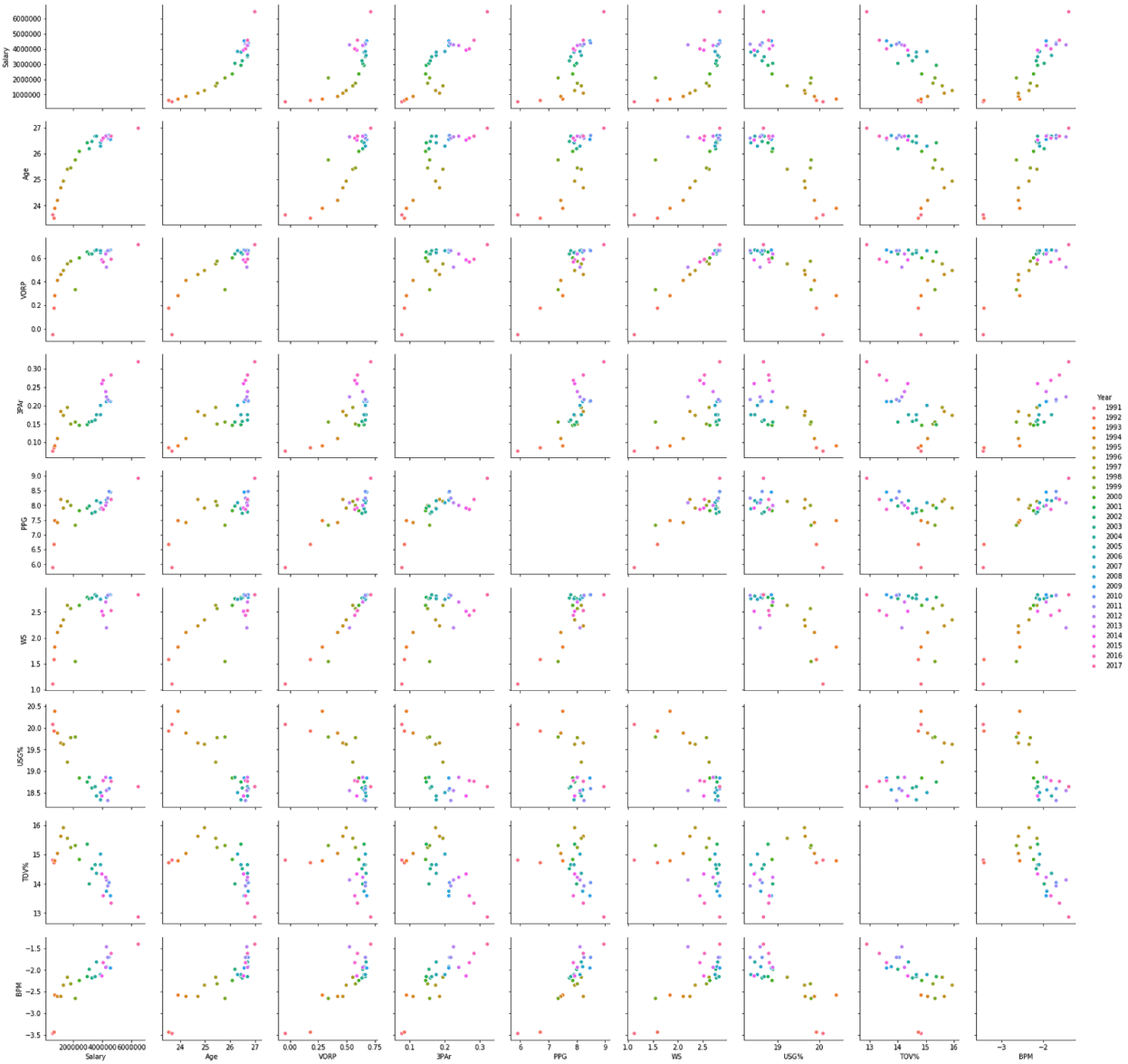
One reason to ask this question is because players are being paid more than ever before. Are they playing far better than players of the past to warrant such a salary increase or are they being paid more for other reasons (such as better union-negotiated benefits and the

fact that the league continues to increase revenue every year). There are likely a great number of factors involved, but with this dataset we can at least get a sense of whether players are outperforming their past counterparts on the court to warrant higher salaries.

We begin by creating a heatmap of all of our previous stats averaged by year:

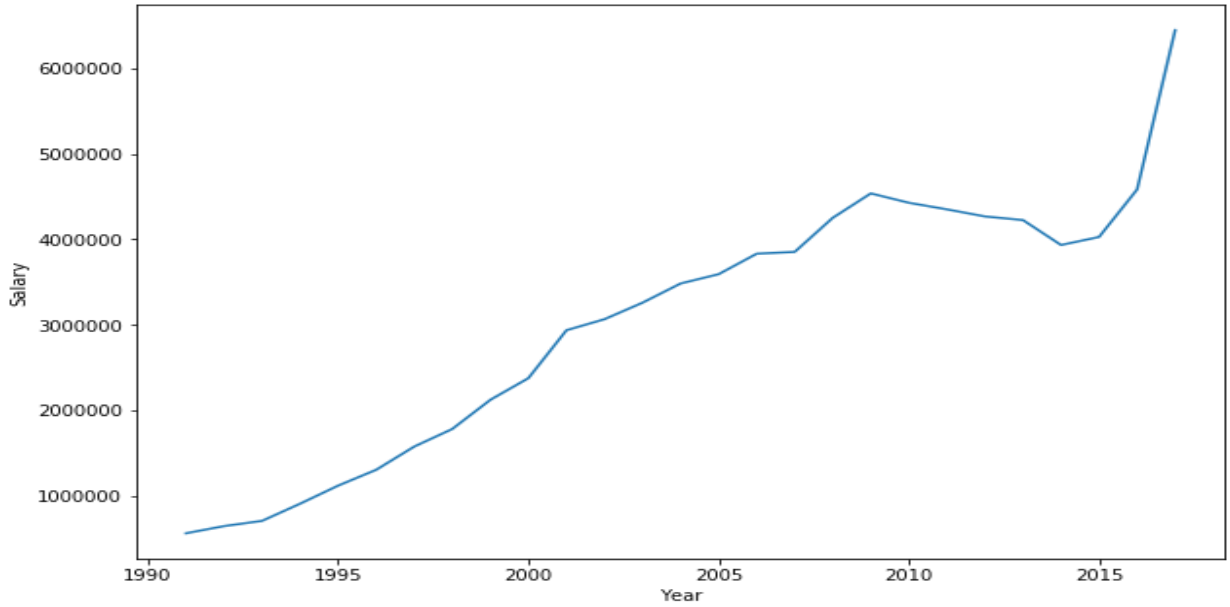


This heat map of the average values of the original stats indicates a high correlation between several categories. We'll use this as a guide as to which relationships to investigate further. For example, salary seems to have a high positive correlation with age, 3PAr, BPM, VORP, 3PA, 3PM, 3P%, and PPG. So, those are worth investigating. However, it may also be worth investigating why there's a strong negative correlation with TOV% and USG%. We can begin the process with a pairplot.

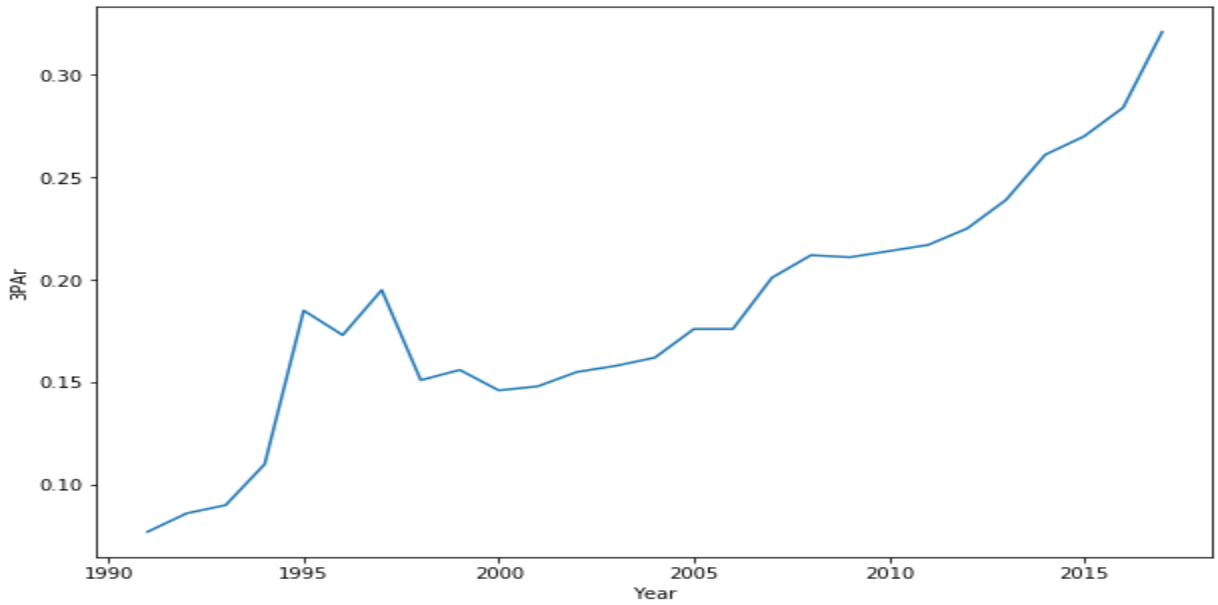


Looking at the above pairplot of various stats that are positively and negatively correlated, there are some obvious trends (such as salary increasing with age, as we've already established, or VORP and WS being highly correlated), but there are some interesting trends as well - salary and BPM appear to be highly correlated and both seem to be increasing as the data approaches 2017. The same can be said of salary and both VORP and WS. However, salary is fairly negatively correlated with USG% and TOV%. We can look at some of these a bit closer below.

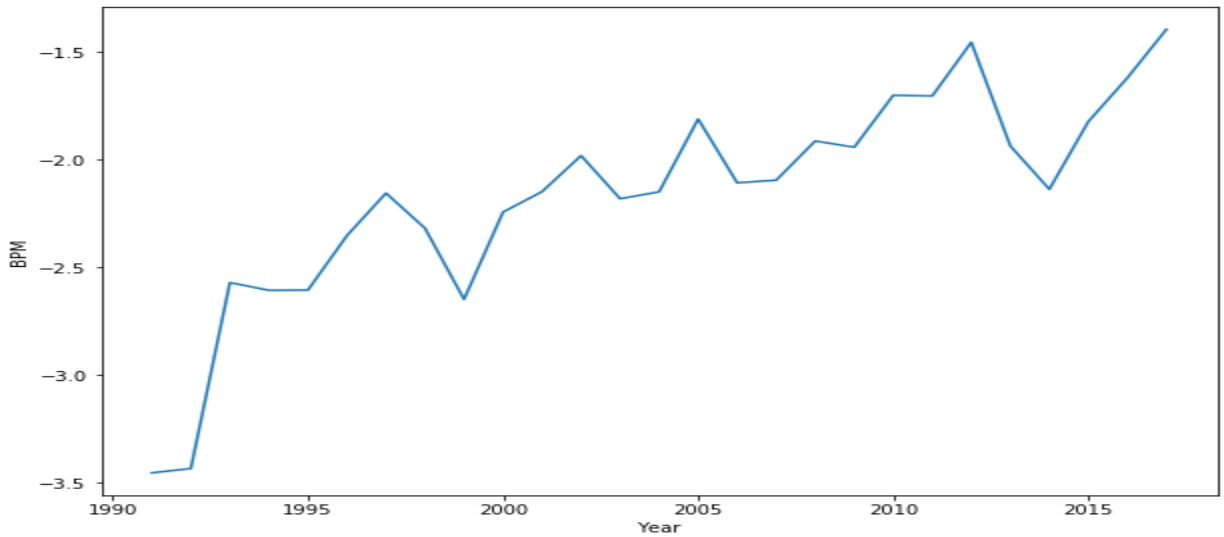
There are several other interesting correlations that can be explored as well (age and VORP, TOV% and BPM, for example), but since our main concern is statistical relationships with salary, we'll have to leave those unexplored for now.
We'll begin by graphing several stats and salary against years. Below is salary vs. years:



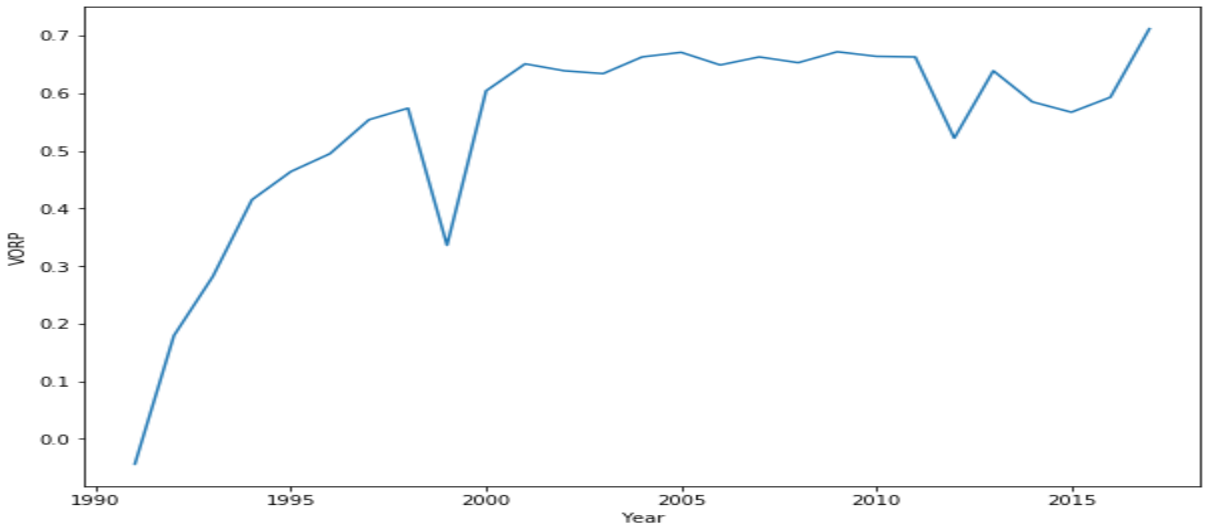
It's important to keep in mind, in light of the aforementioned discussion, that salaries are lower in the 1990s than they should be because of the smaller and younger player population. Yet, even as the population stabilizes by 2005, salary is still lower than the last few years.



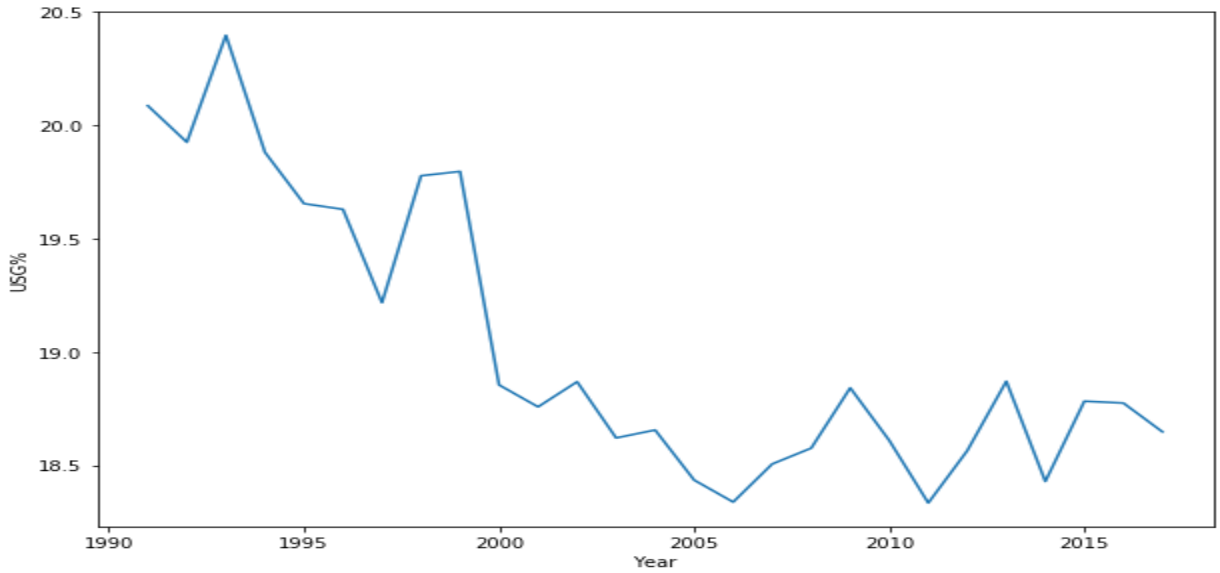
3PAr, which is calculated by dividing 3 point attempts by field goal attempts have increased quite dramatically since about 2007. It's due in part to a change in strategy in which teams are much more likely to attempt 3 point shots than in any point in NBA history. As a result, top 3 point shooters, such as Stephen Curry and Klay Thompson have become quite valuable and have signed lucrative contracts.



BPM appears to fluctuate quite a bit, but the overall general trend has been upward.



VORP is relatively stable since 2005. The pre 2005 values can be explained by higher concentration of young players in each of those years (again due to the bias in our dataset). So, it appears VORP is generally higher for more experienced players. Yet, it also appears that there have not been any significant increases in VORP over the years.



Usage percentage, which measures the percentage of the number of offensive plays that a player is involved in, appears to have dropped quite significantly since 1991. But since 2005, it's been relatively stable, which suggests that USG% is higher among younger players. Thus, if there was less bias in our data, we might see a more stable USG% over our entire timeline.

Of the aforementioned stats, the only one that we can safely say has increased/decreased as significantly as salary would be 3APr. However, there isn't necessarily a causal relationship between the two variables. As we mentioned previously, the NBA's play style has changed in recent years, with more 3-pointers being attempted and the league overall becoming better at making 3-pointers, as the graph immediately above shows. Thus, good 3-point shooters have become more valuable and are being paid better as a whole. However, we can't necessarily say that their increased salaries contribute to an overall increase since most players in the modern NBA seem to be paid better than their past counterparts.

So, to answer our last question, 3-point play making has gotten better over the years, but most stats averages have not changed significantly over the years, at least not in correlation with salary. As a result, players are likely not being paid more because they play "better" than players of the past (despite the fact that they may be more athletic), but it's more likely due to factors outside of this dataset, such as better union-negotiated benefits and the fact that there is just more money in the NBA than ever before.