# Relationship between NBA Players and Compensation

*Ashley Liu, Chitwan Kaudan, Sheng Siong Ong, Karthik Nataraj*

*December 1, 2016*

## Abstract

The objective of this project is to analyze the relationship between the performance of a basketball player and his compensation. As a proxy for the performance of a player, a weighted EFF (EFF'), or "efficiency" statistic, was computed for each player. The "value" of each player was then computed as a ratio between the player's EFF' and salary. Through our analysis, we found that the player's EFF' and salary are not strongly correlated, due to both flaws in the EFF' that create bias by player position as well as potential omitted confounding factors.

## Introduction

When determining the relationship between professional basketball players and their salaries, our team was curious whether performance was a deterministic factor in determining how a player would be compensated. To answer this question, we were faced with the following analytical problems:

- Generating a single statistic that would provide a holistic measure of a player's performance, accounting for potential biases based on their position
- Comparing the generated performance statistic to the player's salary
- Drawing conclusions from the nature of the relationship between the player's performance and salary – for example, whether or not each player was overvalued or undervalued
- Evaluating the performance statistics for potential problems and locating next steps or possible areas of improvement in our analysis

In order to approach these analytical problems, our exploratory research questions included:

- How can the player's performance be evaluated?
- Are there undervalued professional players? If so, is there some confounding factor? (i.e. Do all defensive players tend to be undervalued, while all offensive players tend to be overvalued?)

## Data

We used data on NBA players for the 2015-2016 season for our analysis. The NBA stands for the National Basketball Association, which is the predominant professional basketball league in North America. The league has 30 teams in total (1 in Canada and 29 in the United States). While each team has at least 14 players on its roster, each team can only have 13 active players. This means that not every player on the roster may have played during the 2015-2016 season. These players were not included in our final dataset.

Our data comes from *Basketball Reference,* which is an online source for player statistics for a number of professional and college level sports. Although unaffiliated with the NBA, Basketball Reference is a trusted platform for players' performance statistics, salaries, rosters, and other such information, and it is referenced and highly praised by ESPN and former players alike.

For each of the 30 teams, we used three distinct datasets from Basketball Reference.com, titled Roster, Totals, and Salaries. Below are the variables contained in each table:

- *Roster* contains the player's name, position, height, weight, birth date, country of origin, years of experience, and college attended.
- *Totals* contains the player's name, age, games played, games started, minutes played, field goals, field goal attempts, and other statistics regarding their performance.
- *Salaries* contains the player's name and salary.

Each dataset is downloadable from the site as a CSV file. Each file contains a table where all values are text or characters.

# Methodology

## Tools

We used R and packages within Rstudio to scrape, clean, analyze and visualize our data. Some We also briefly used MS Excel to visualize some of our large csv files and data frames (from RStudio) but all computation was performed within RStudio using R.

## Main Stages

*Data Scraping & Merging:* The data tables that we had identified were located on webpages with largely similar URLs (http://www.basketball-reference.com/teams/CLE/2016.html (http://www.basketball-reference.com/teams/CLE/2016.html)). As such, we used a simple for-loop to parse and then extract the information from all 90 data tables (3 tables per team x 30 teams) that we had identified for our project from Basketball-Reference.com. We wrote each of these 90 tables to a csv file in the appropriate folder.

*Cleaning & Formatting:* One of the challenges of the project was cleaning up the data, as without this step, it would be impossible to perform any sort of analysis. Each team had 3 tables, and the data from each of the tables had to be cleaned up and merged. In the first step of the merging process, we combined the tables of each data type together (making a full roster table, a full salary table, and a full stats table - regardless of team). In this step, we were careful to remove any players who were missing either salary or statistics data. As the purpose of our project was to compare player efficiency with salaries, we had to ensure that our cleaned data for each player has statistics and salary data. If either information was missing, they cannot be compared. We also noted that some players appeared twice on the merged dataset. These players most likely changed teams during the season. They had different salary and statistics data for each team that they played for. We did not remove these 'duplicates' as there was no reason that we should have selected a particular dataset over the other. We then had to clean up the merged dataset. We wrote some functions to clean up certain issues with data formatting (converting height to inches, making dollars earned a numeric value, etc.) and also added more intuitive column names, and removed NA values. We then wrote the compiled data frame into "roster-salary-stats.csv", which is the foundation for our data analysis.

*Exploratory Data Analysis:* In our exploratory data analysis, we first examined some basic summary statistics. This is not very useful, but it told us what kind of data we were working with. Then we compiled a data table that detailed salary statistics by team, rather than by player. This allowed us to see which teams had the highest payroll, how the average pay of each team varied, and generally giving us a sense of teams' salaries. We also generated plots and histograms to visualize and understand the data better.

*Calculating adjusted player efficiency:* We then performed a Principal Components Analysis (PCA) to determine the efficiency of the players. We standardized the variables, made sure to adjust for number of games played and added negative weights for "bad" statistics (i.e. turnovers, missed field goals, missed free throws). After obtaining the weights (taking its absolute value) and the standard deviation for the PCA by using prcomp(), we

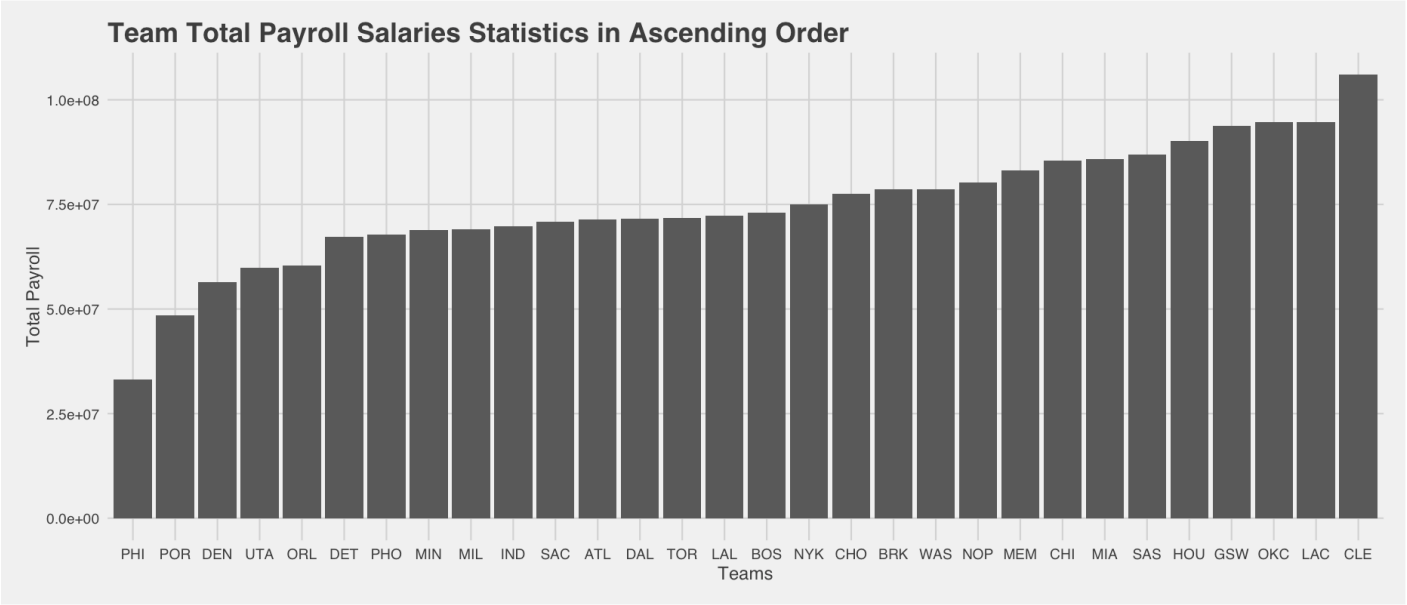
wrote a function to calculate the efficiency statistics for each player, stratified by position. Finally, we compiled a new data frame that included the efficiency of the player, along with other game statistics. We then wrote this data frame into “eff-salary-stats.csv” for further data analysis.

*Calculating Value of Player:* Once we had the efficiency of each player, it was a simple vector calculation to compute the value of each player.

*Two Shiny Apps:* We built two Shiny Apps to interact with the data in real time, and observe how the statistics, salaries, and efficiencies are distributed and correlated. The first Shiny App allowed us to compare various salary statistics between teams in ascending or descending order. The second Shiny App allowed us to compare various player statistics and showed us the correlation coefficient of the compared statistics. In particular, we were interested to see how efficiency compares with salary. We also allowed the second Shiny App to be subsetted by player position to see whether there were significant trends within each position. These allowed us to gain some insights on how to move forward with the analysis.

# Results

The first information that struck us was the huge disparity in salaries. Salary ranges from $8000 to $25 million USD. This disparity exists between teams where the Cleveland Cavaliers spent around a total of $100 million USD while the Philadelphia 76ers spent around a total of $33 million USD. This is in spite of efforts by the league in recent years to balance the competitiveness within the league by implementing a salary cap. This suggests that the disparity in salaries between teams prior to the salary cap must have been even larger, with even less competition between teams.




Salary statistics by team in ascending order

Moreover, there is huge disparity within teams as well. When we look at the graph of comparison of standard deviations in salaries between teams, there are 4 teams with about $7 million USD standard deviation in salaries. Given that the highest salary is about $25 million USD, the standard deviation is highly substantial.


Salary Standard Deviation statistics by team in ascending order

While it seems logical that the better players receive higher salaries, it is difficult to quantify what makes a “better” player. As such, the rest of the project seeks to quantify the “value” of a player through primarily the player efficiency formula EFF’.

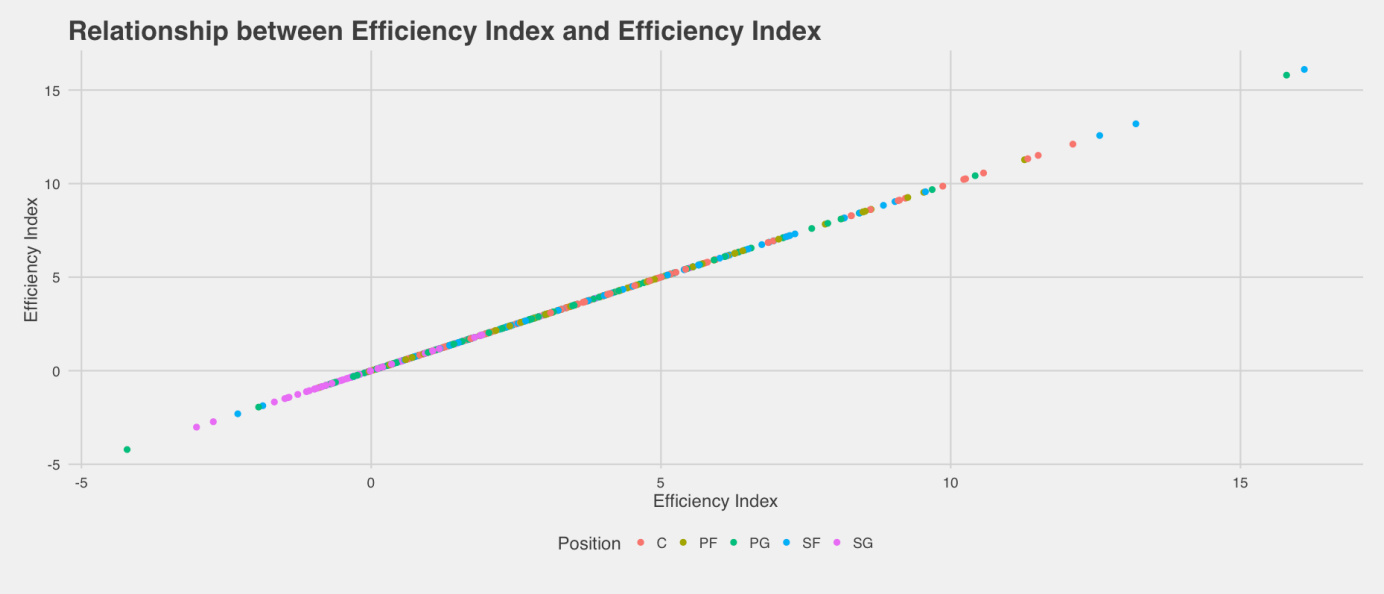
| player | efficiency |
|---|---|
| Tony Wroten | -4.212008861 |
| Orlando Johnson | -3.014160998 |
| Orlando Johnson | -2.724677008 |
| Duje Dukan | -2.302275054 |
| Bryce Cotton | -1.942933455 |
| Bruno Caboclo | -1.87026037 |
| Devyn Marble | -1.671244629 |
| Jordan Adams | -1.490343153 |
| Nick Young | -1.443380509 |
| Rashad Vaughn | -1.418211063 |
| Luis Montero | -1.267098619 |
| Dion Waiters | -1.117109662 |
| Jared Cunningham | -1.066388723 |
| O.J. Mayo | -0.981391515 |
| Marco Belinelli | -0.975829 |
| Gerald Green | -0.934120364 |
| Alex Stepheson | -0.898734854 |
| Aaron Harrison | -0.872519538 |
| Randy Foye | -0.855541484 |
| Alec Burks | -0.841229166 |

List of 20 lowest 20 EFF indexes

| player | efficiency |
|---|---|
| Kevin Durant | 16.10174941 |
| Stephen Curry | 15.79534505 |
| LeBron James | 13.19517192 |
| Kawhi Leonard | 12.57067175 |
| Karl-Anthony Towns | 12.10792725 |
| Brook Lopez | 11.50945574 |
| DeMarcus Cousins | 11.33030914 |
| Anthony Davis | 11.27394017 |
| Nikola Vucevic | 10.56633031 |
| Chris Paul | 10.42216919 |
| Al Horford | 10.25642578 |
| Hassan Whiteside | 10.22475717 |
| Pau Gasol | 9.86307631 |
| Russell Westbrook | 9.680095226 |
| Paul George | 9.563892012 |
| Blake Griffin | 9.534872373 |
| LaMarcus Aldridge | 9.259696911 |
| Greg Monroe | 9.219312268 |
| Marc Gasol | 9.116205729 |
| Marcin Gortat | 9.092731214 |

List of 20 highest 20 EFF indexes

As we anticipated, the EFF' metric produced a ranking of players that appeared to be reasonably consistent with their respective skill levels and team contributions. Compared to the non-adjusted EFF, which would favor offensive players, the EFF' does a better job of leveling the playing field between offensive and defensive players. This is confirmed by the fact that the Top 20 EFF' contains 6 players that can be reasonable considered defensive players: Chris Paul, Al Horford, Hassan Whiteside, Kawhi Leonard, Paul George, Marc Gasol.

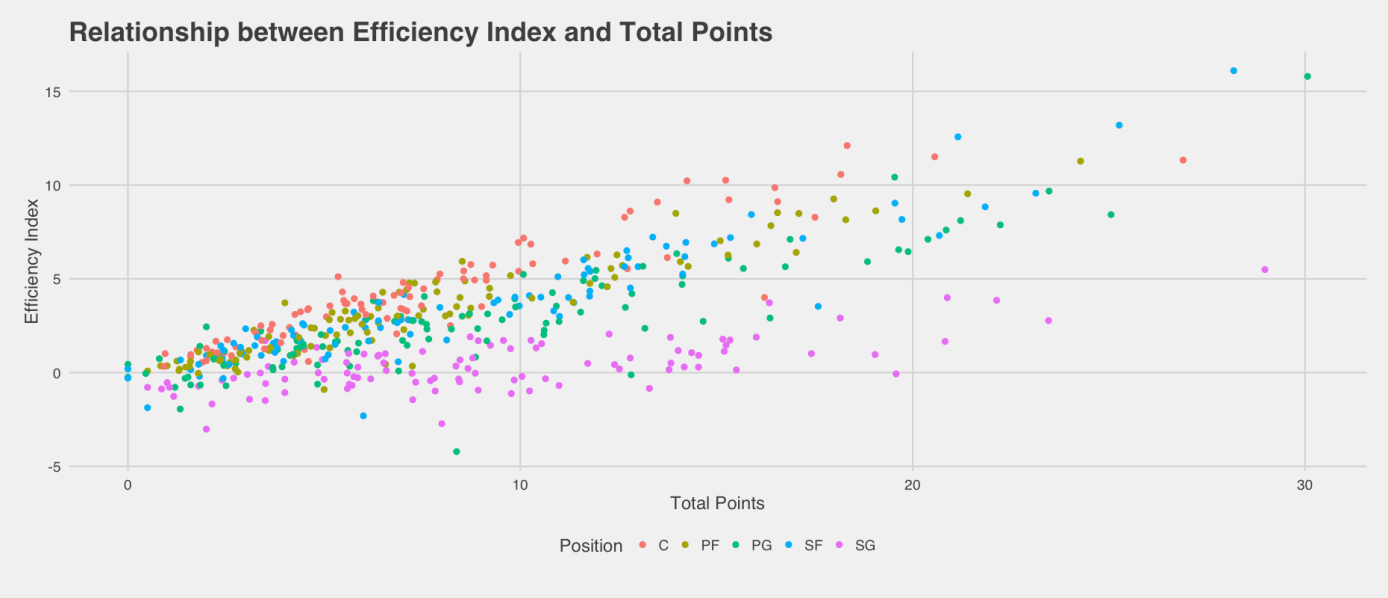



Correlation Coefficient: 1

Shooting guards have lower EFF Indexes

One thing we found interesting was that shooting guards tend to have lower EFF' scores compared to other positions, while Centers and Power Forwards have higher EFF' scores. As you see from the scatter plot above, most purple dots representing shooting guards are centered around an EFF Index of 0. Whereas if you remove the two outliers from the plot, the orange dots have the highest average EFF centered near an EFF Index of 5.

This seems plausible given that Centers and Power Forwards often operate close to the rim, leading to higher percentage shots while shooting guards often operate further away, near the 3-point line.




Correlation Coefficient: 0.725409668748172

Shooting guards have lower Efficiency per Total Points

This can be viewed from the scatterplot of the relationship between efficiency and points. Most of the purple points (shooting guards) are clustered near the bottom of the graph, suggesting that shooting guards have lower efficiency for the same amount of points. This seems to suggest that the efficiency of shooting guards are lower than those of other positions. When we account for efficiency as compared to salaries earned, shooting guards earn much higher salaries for their level of efficiency. There are several ways that we can view this trend.

Firstly, the EFF' formula that we use may inadvertently favor certain positions over the others. While it does account for offensive and defensive statistics, it does not measure the full "value" of a player. A player may put up low statistical number but contribute to the team through "energy" and "hustle", which may be difficult to quantify. Shooting guards are often tasked to guard along the perimeter, and may provide defensive value through hassling the other team without earning themselves a steal. As such, shooting guard's efficiency value may be higher than what the formula yields.

Secondly, this may also suggest that there is a dearth in good shooting guards in the NBA currently. When we view the list of top 20 efficient players in the NBA, there are no shooting guards on the list. This reaffirms what we previously said with regard to the low efficiency of shooting guards. However, rather than suggesting that shooting guards are always inefficient, we can merely conclude that in the year that we conducted our data analysis, shooting guards are not as efficient as compared to other positions. This may change in subsequent years. Moreover, there is increasingly a blurring of positions in the NBA. While a player may be classified as a shooting guard, he can similarly be playing the point guard or small forward position.

Lastly, we can view this trend at face level: That shooting guards are indeed just more inefficient as compared to other positions. With the exception of Michael Jordan and Kobe Bryant, shooting guards just do not often lead teams to championships. Shooting guards, as their name suggests, are primarily shooters that stretch defenses. As such, they are just not as likely to be as efficient as the other positions.

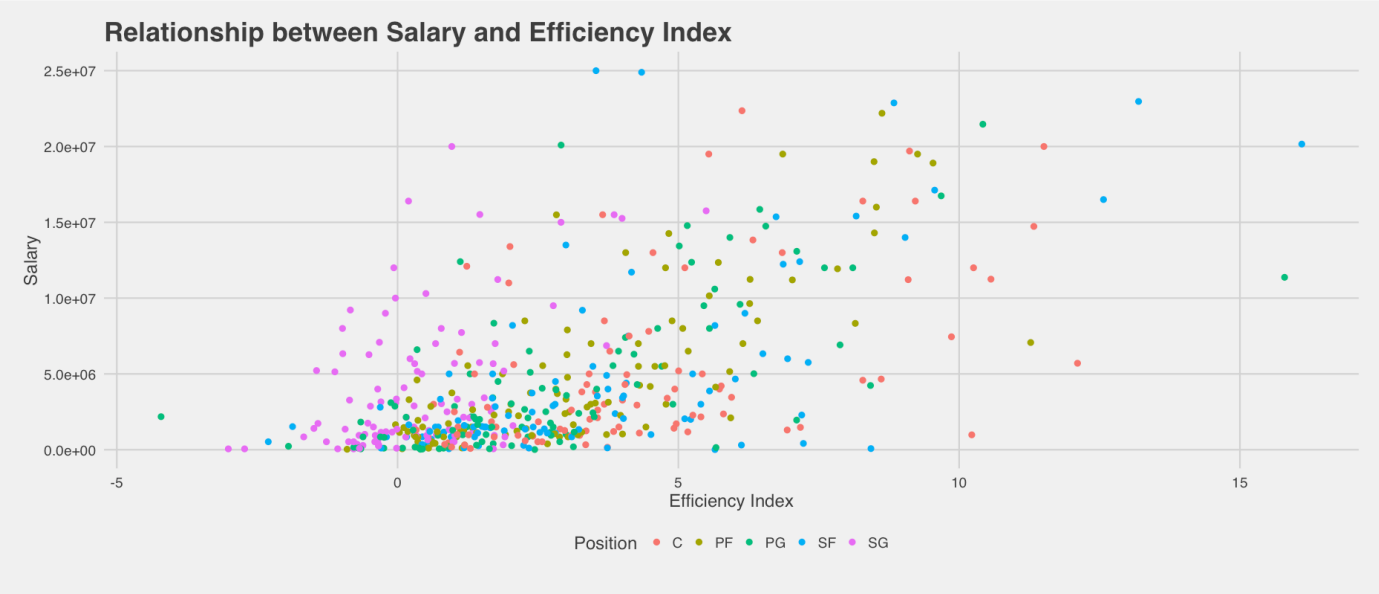
```
================================
List of 20 most valuable players
================================
                    player team player_value
476          Alan Williams  PHO 1.004404e-05
477          Nate Robinson  NOP 1.006019e-05
478         Jason Thompson  TOR 1.018331e-05
479       Hassan Whiteside  MIA 1.041909e-05
480           Bryce Cotton  MEM 1.146614e-05
481          Briante Weber  MEM 1.315649e-05
482            Coty Clarke  BOS 1.480499e-05
483            Chuck Hayes  HOU 1.550913e-05
484          Jordan Farmar  MEM 1.635279e-05
485            Joe Johnson  MIA 1.743501e-05
486        Michael Beasley  HOU 1.997489e-05
487 Thanasis Antetokounmpo  NYK 2.208904e-05
488             Henry Sims  BRK 2.312232e-05
489           Jordan McRae  PHO 2.649021e-05
490        Jimmer Fredette  NYK 2.856866e-05
491        Jordan Hamilton  NOP 3.126788e-05
492            Tim Frazier  NOP 3.933540e-05
493            James Ennis  NOP 1.130464e-04
494          Briante Weber  MIA 1.976460e-04
495          Dahntay Jones  CLE 6.408706e-04
================================
List of 20 least valuable players
================================
                player team  player_value
1     Orlando Johnson  NOP -5.409284e-05
2     Orlando Johnson  PHO -4.889769e-05
3      Alex Stepheson  MEM -2.909657e-05
4    Jared Cunningham  MIL -1.847328e-05
5     Jimmer Fredette  NOP -1.429720e-05
6     Elliot Williams  MEM -1.175213e-05
7        Bryce Cotton  PHO -8.496930e-06
8      Xavier Munford  MEM -7.569540e-06
9       Lorenzo Brown  PHO -6.289433e-06
10      Keith Appling  ORL -4.828975e-06
11         Duje Dukan  SAC -4.384509e-06
12       J.J. O'Brien  UTA -2.773076e-06
13        Erick Green  DEN -2.442392e-06
14       Luis Montero  POR -2.413094e-06
15 Bryce Dejean-Jones  NOP -2.191591e-06
16     Elijah Millsap  UTA -1.987838e-06
17       Devyn Marble  ORL -1.977666e-06
18        Tony Wroten  PHI -1.932687e-06
19     Aaron Harrison  CHO -1.661648e-06
20    Pat Connaughton  POR -1.489441e-06
```

Top 20 Most and Least Valued Players

The individuals with the highest “value” ratios have relatively high EFF’ scores compared to their salaries. In contrast, the individuals with the lowest “value” ratios have relatively low EFF’ scores compared to their salaries. However, we were surprised that the “value” metric generated a list of players that did not showcase any particularly recognizable or famous players. As such, the ranking generated by the “value” metric does not appear to describe players by their performance or usefulness to their respective teams.

This seems to be the result of how we calculated the “value” metric. We calculated player value by dividing a player’s EFF by his salary. However, EFF is a standardized metric that only ranges from -5 to 16. Salary is a non-standardized metric that ranges all the way from $8,000 to $25 million. This methodology seems to value players extremely low salaries and average EFF that only play a few games a season like Dahntay Jones and Briante Weber. This profile of a highly valued player doesn’t match our conceptual understanding of highly valued player.

As “value” metric produces a ranking of players that is not at all consistent with the ranking of players generated by the EFF’ metric, it suggests that players’ salaries and efficiencies are not well-correlated. The weak correlation between salaries and efficiencies is confirmed by the scatterplot below:




Correlation Coefficient: 0.58861248688803

A weak correlation between Efficiency and Salaries

From this figure, we observe that the correlation coefficient between salaries and EFF' is 0.589, which indicates that the strength of linear correlation is not strong. This finding was somewhat surprising, as one might initially expect players' compensation for a given season to be a direct function of (or at least, heavily based on) their current season's performance. However, in practice, the relationship between salary and EFF' may be flawed due to omitted variable bias – that is, the omission of other factors that may be influencing the player's salary.

Furthermore, we observe that different player positions yield different correlations with salary. For instance, a per-unit increase in the EFF' of a Shooting Guard will yield a larger expected increase in the player's salary for the same per-unit increase in the EFF' of a Center.

# Conclusions

The main finding of our analysis is that during the 2015-2016 season, a player's EFF' was weakly correlated with their salary. This is readily apparent in our correlation coefficient of 0.589. From this analysis, professional basketball team management could potentially benefit by using the "value" metric to identify players who are currently overvalued or undervalued with respect to their current performance. This would allow them to offer appropriate salary contracts to players based on their level of efficiency. Players with higher efficiency deserve better salary contracts while players with lower efficiency should get lower salary contracts. However, this is of course under the assumption that the efficiency formula is a good measure of the "value" of a player. Professional basketball team management may similarly decide to offer contracts based on other variables that they may consider valuable to their team, such as team experience or leadership. As such, the efficiency index should be used as only one of several other considerations when drawing up salary contracts for NBA players. Nonetheless, this could be used as reference to adjust current-year salaries in a more efficient manner in future years.

Moving on from our current analysis, there are several other factors that we can analyse in the future that can give our analysis greater breadth and depth. We can investigate sources of omitted variable bias in our current analysis. For example, a player's salary in the previous season tends to determine their salary in the current season (as a consequence of the rigid salary institution in professional basketball), reducing the importance of current performance as a determinant of salary. This relationship could be investigated by determining the correlation between player salary in time period **t** and player salary in time **t - 1** in order to gauge the effect of a

player’s salary in the previous year on their current salary. If this analysis were to yield a statistically significant relationship between the two, then we have reason to believe that the player’s salary in the previous year is a more important determinant of the player’s salary in the current year than their current performance.

Moreover, as we pointed out in the Results section, there is a disproportionate weightage given to salaries in the calculation of the value of the player in our data analysis. In the future, in order to better calculate the value of a player with respect to their player efficiency and salary, we could weigch the salaries to provide only cardinal rather than ordinal significance to the formula. This will provide a better depiction of the value of a player as compared to the current analysis which is highly skewed by extreme values of salaries.

Additionally, it is important to recall that although the adjusted EFF metric is much better than the original EFF metric at assessing defensive players, it still tends to favor offensive players. In future analyses, it may be helpful to adjust the formula used in the adjusted EFF metric in order to further reduce the bias towards offensive players. There are more variables rather than just blocks and steals that can impact the team defensively such as “number of shots affected” rather than “number of shots blocked”. These could be quantified and included into our efficiency index in the future.

In future analyses, it is thus most important to include as many variables that may possibly affect the “value” of a player, with the correct weightage. While it still may not yield the appropriate “value” of a player, it will allow us to get a better idea of how to draw up salary contracts for NBA players accurately. This is highly important given the exorbitant salaries NBA players earn. As such, there is a need to come up with an accurate valuation of individual NBA players.