### An Analysis for Valuations of Team Members in Relation to Actual Contribution to Team Success

Are businesses receiving the value they placed on employees through salary? Where is the best ROI and can contribution to success be measured in a way to determine the best value from a team member? Utilizing current, NBA individual player stats and their respective salaries, this project will analyze the valuations placed on a player based upon their pay and the contribution to the team’s success (win-share). This analysis will provide a foundation for businesses to gauge where they are getting the best ROI on employees, strictly from a cost vs. return perspective. Upon review, a client could utilize the outcome from this analysis to determine a more accurate valuation of a team member and therefore identify any gaps in pay vs performance. The problem to solve: can an organization accurately place a value on a team member for their contribution to success based on performance?

The data to be utilized in this analysis will involve a variety of aspects. First of all, the individual stats of a player will be analyzed to determine their respective “win-share,” or their individual contribution to a team’s wins throughout a season. This win-share variable is something that has already been created through basketball-reference.com utilizing the inidividual stats and their offensive and defensive efficiency. Second, the collection of each individual player’s salary will provide the value that an organization placed on a player in regards to their hopeful contribution to the team’s success. This data is publicly available on espn.com. A deep analysis of individual salaries and their respective win-share will determine a players true value to an organization while they are on the court.

There are limitations to this analysis as this will not factor in any unique performance based bonuses for players nor their off-court value to an organization. These are two very important aspects to keep in mind as this analysis progresses as a player can provide much more to a team than just what they contribute while on the court. Such examples include advertising, popularity, ticket and jersey sales, and more.

My approach to solving the primary problem will involve analyzing the win-share contribution of a player based upon their individual stats on the court in relation to their salary. This provides a baseline value for each player respective to their performance. Answering the primary question does open doors to other questions that will also be explored. Other questions that could be examined include:

* Are the highest valued players contributing their fair share?
* Where are the biggest gaps in current value vs contribution to success?
* What could a successful team look like if an organization was looking for the best value at each position?
* Are role players more valuable than “Superstars?”
* What is the true value of a “Superstar” to a team?

The deliverables for the client in this analysis will be broken down into various pieces; code for any further analysis needed, a written paper with an in depth explanation to the project, the approach and results, and a visual representation of the results. These final pieces will provide insight for the client to make data-driven decisions to maximize efficiency of spending in relation to performance of a team member.

### Data Wrangling

The analysis of an NBA player’s value based upon salary and win-share contribution brought my search for relevant data to the web, more specifically basketball-reference.com. This site provides up to date statistical information regarding player data as well as contract information for professional athletes. It is a verified and trusted source with the data available on the site for free, in table format.

Given this initial, available data, the process of pulling the information to make it usable for analysis became the only true hurdle. The process could be completed in a variety of ways, from web scraping to copy and pasting the table. One of the factors to consider was how up to date the data should be for this project and for analysis purposes given that the NBA season was still in progress when I began this project. Basketball Reference updated their data daily, and therefore a web scrape would be the best option as it would always be up to date for future use.

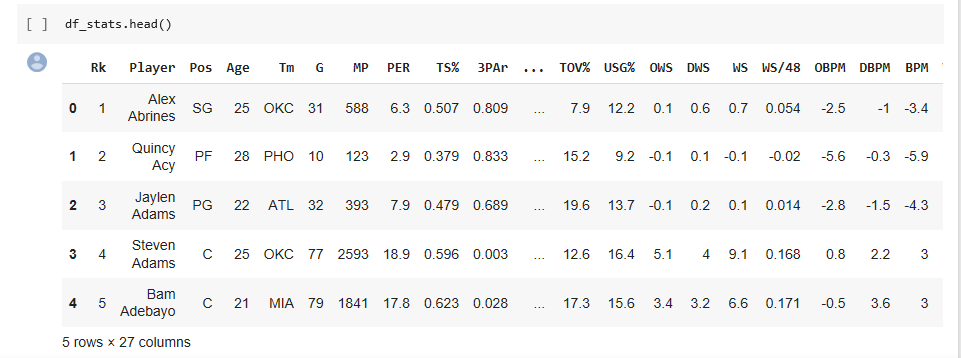
Through a process of writing a loop to read in each line of the data table one by one from the HTML code, and then saving the table in csv format, the goal was to recreate the data from basketball-reference.com whenever the analysis was needed. Using the help of my mentor, we created the loop to parse the HTML code from the site to write the csv. Through this process, a few issues arose such as compatibility between different versions of the software we were using. This led to a bigger realization that the code would have to be uniform to work with various versions and an overall problem that at any given time, Basketball Reference could update and the entire project would no longer be useful.

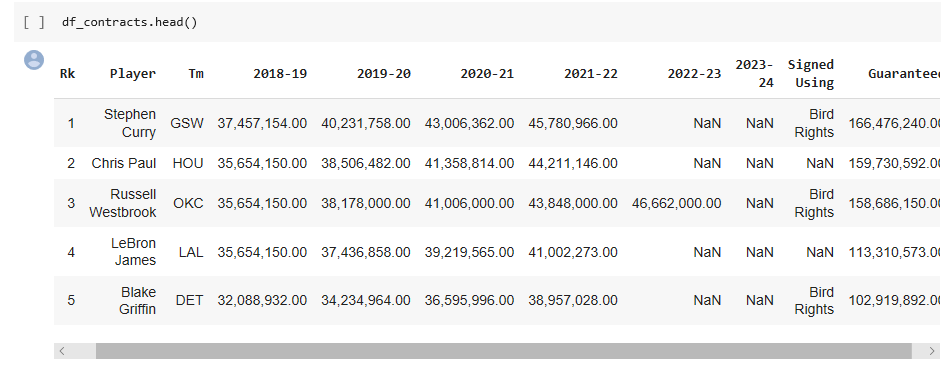
Given this newfound information, I resorted to copying and pasting the data table from basketball-reference.com into an excel spreadsheet. From here, the data file would be unchanged without conscious changes to the file itself, rather than worrying about a third party ruining the project. The steps required after having the Excel file created were fairly straight forward to cleaning up the data for analysis purposes.

The data cleaning steps involved creating a saved CSV from the Excel file for easier importing. The next steps involved cleaning title rows that were in the table to make the data easier to read for users. This accounted for quite a few of the null values of the data. Another step in the cleaning process was making sure that the columns were able to be referenced for analysis purposes, so I made sure to rename the “years” columns to a better format while still maintaining the readability. The final steps to the cleaning process involve changing the data type of the columns to their appropriate type. Initially, when the csv is read into the notebook using Pandas, the data types are all “object,” which can make for issues when analyzing the data later on.

After reviewing the data, I was fortunate enough to not find any outliers that had to be dealt with or explained from my data outside of short term contracts for a few NBA players.

The most difficult part of my data wrangling process was trying to find a way where the dataset was up to date as the website itself updated. However, after many attempts and roadblocks, the best decision was to just save the data in its most current state to utilize for my project.





### Applying Inferential Statistics:

Understanding what a normal NBA players Win Share contribution to an organization would be a huge benefit to an organization. What this could allow for is advanced insight towards expectations of your players as well as how to properly compensate these players for their on court presence. This should result in a better understanding of what average players in the NBA are contributing towards the teams’ success, then a better understanding of what the pay range for certain contributions should be and therefore help an organization that employs a maximum of 15 players make the best decisions when negotiating contracts.

To do this, we’re looking closely at the Win Share statistic from our data and learning more about the correlation to Salary. The hypothesis is as follows:

H0 : There is no correlation between Win Shares and Salary

H1 : Salary is correlated to Win Shares.

I conducted multiple correlation tests: Spearman’s, Pearson’s and Kendall’s. A Spearman Correlation is a nonparametric measure of rank correlation(statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).

Kendall's tau coefficient (after the Greek letter [τ](https://en.wikipedia.org/wiki/%CE%A4)), is a statistic used to measure the ordinal association between two measured quantities. A tau test is a non-parametric hypothesis test for statistical dependence based on the tau coefficient.

The results are as follows:

Spearman’s correlation: 0.51514

Samples are correlated (reject H0) p = 0.000

Pearson’s correlation: 0.53543

Samples are correlated (reject H0) p= 0.000

Kendall correlation coefficient: 0.362

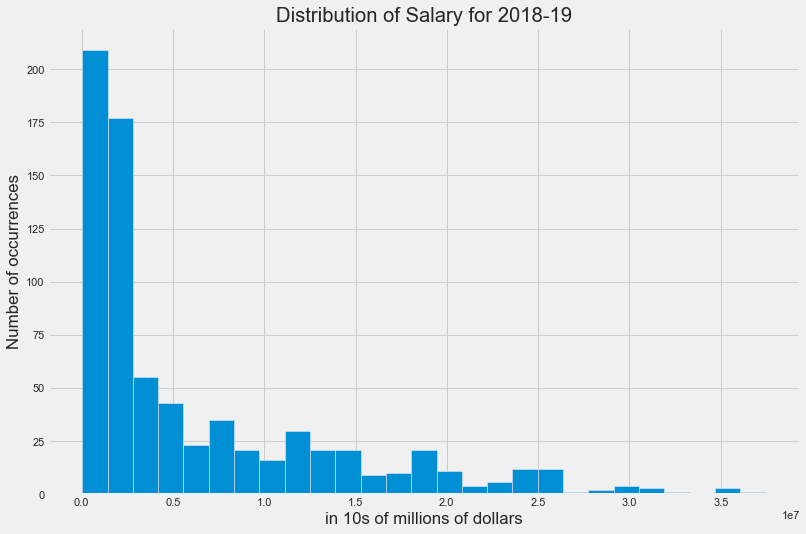
Samples are correlated (reject H0) p= 0.000

The most important takeaway here is that in all three tests, we rejected the null hypothesis.

#### Summarizing:

After initial analysis, we can see that using ONLY the Win Shares as a feature for prediction yields a relatively low accuracy. A strong positive that came from this testing was statistically proving that there is a correlation between Win Shares and Salary. One thing to note that was extremely important is determining what value to use to represent the average of the data.

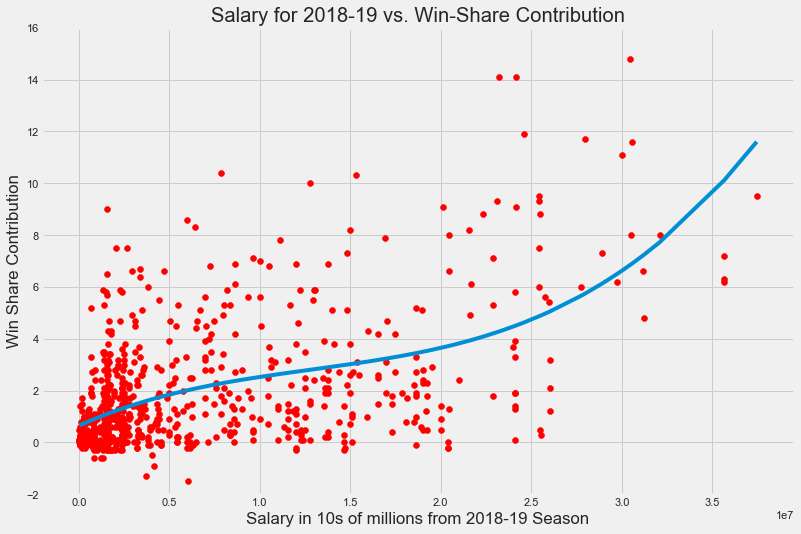
Initially, I used the mean, but this value was likely skewed because of the distribution we saw earlier. From there, I trimmed the data and removed the top and bottom 15% to analyze that new average value. The trimmed mean was reduced significantly, which was expected. The median is the value I chose to represent an average data point for both WS and Salary moving forward.



Once this was completed, I tested for correlation between WS and Salary using a variety of tests. In each case, we rejected the null hypothesis and can say with confidence that there is a correlation between WS and Salary. To be even more accurate, there is a positive correlation, and Pearson's test of correlation, a .53543 correlation between WS and 2018-19 Salary.

The goal moving forward is to utilize the data and Win Shares stat to predict the upcoming Salary of NBA players.

The goal moving forward is to utilize the data and Win Shares stat to predict the future salaries of NBA Players.



### Machine Learning:

This analysis was conducted to determine an NBA Player’s Value (salary) based upon their on court performance (statistics). An organization would see a lot of value in being able to accurately pay their players for their performance. This process was broken down into a few steps to try and build the most accurate model. First, and likely most important, was choosing the best learning technique for our data and in this case a linear regression model was chosen. I chose this model because the goal was to predict a continuous value (salary) using one or more independent variables (players’ stats).

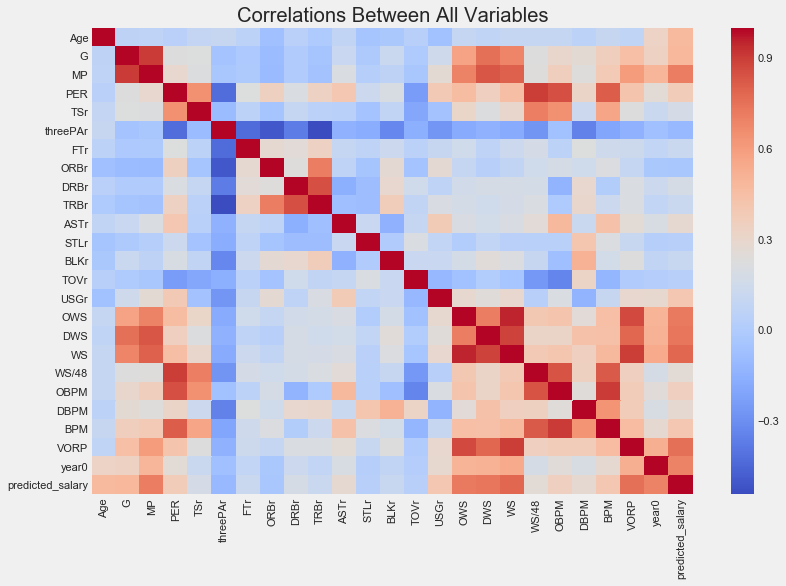
Initially, I looked at using only one statistic to predict the salary, Win Shares. Win Shares is a calculated value from the players’ on court performance. This resulted in a very low R2 value of .287, and an F-stat of 301.0. Clearly, this calculated stat was not accurate in predicting how a player would get paid. The next steps involved taking a broader look at the data to determine if there was a correlation between a player’s salary and their on court performance. This led to my hypothesis:

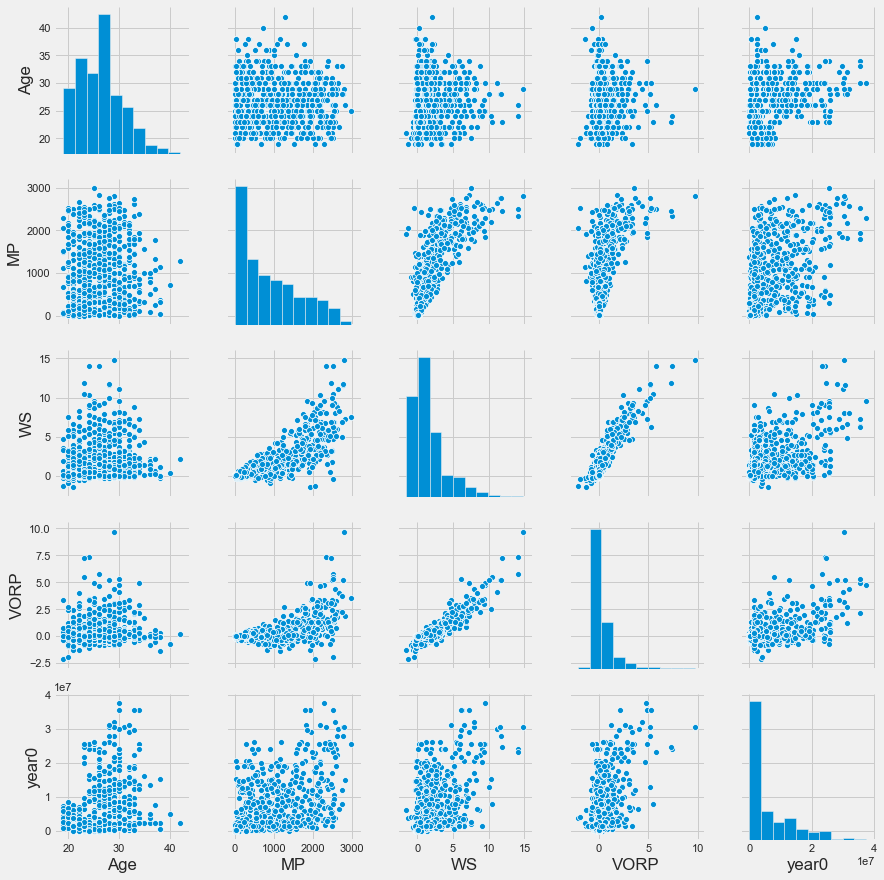
Hypothesis:

H0 : There is no correlation between a player's on court performance and Salary.

H1 : Salary is correlated to a player's on court performance.

A quick glance at a correlation heatmap provided insight into which features were correlated to one another.

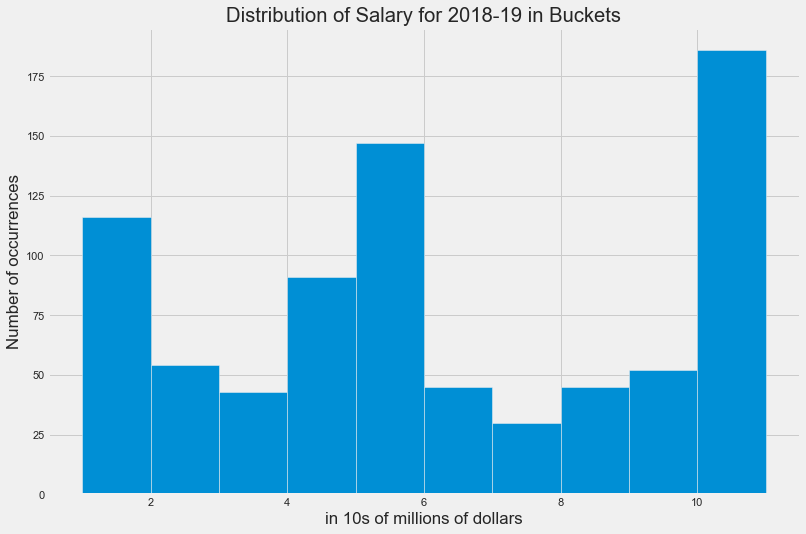




The next step was creating a linear regression model to determine which features contributed to the salary. I began with the model using all features, and still resulted with an R2 value of .480 and an F-Stat of 32.92. This R2 value tells us that the model doesn’t fit the model very well, and that’s including all of the features. From here, I aimed to fine tune the model to increase the F-stat and at least stay close if not improve the R2 value. I removed the features from the model that were not statistically significant using a p-value of .05 as my threshold. What this resulted in was a model with just 10 features instead of 22. This updated model resulted in an R2 value of .468 with the F-statistic more than doubling to 70.13. This is definitely an improvement as we reduced noise while maintaining similar R2 value. However, this model is still not very accurate.

This model ‘m\_1’ accounted for approximately the same amount of variance in salary as the model that included all of the features while simultaneously increasing the confidence that these features are related to the target (salary).

After discussing more with my mentor about the lack of accuracy with my model, the suggestion was made to try and predict salary ranges instead of absolute values based upon their on court production. I created 11 bins with custom ranges because the data for salaries is skewed due to over half of the players receiving salaries less than the mean and median. After the bins were created, I built another linear regression model to predict which bin players would fall into based upon their statistics.



The very first model created (which used all features) to predict which bin a player would fall into resulted in an R2 of .432 and an F-stat of 27.2. This was less accurate than the first model. I tried fine tuning the model using the same method as before, by eliminating features based upon their p-value significance. The tuned model resulted in an R2 value of .414 and F-stat of 70.52. We were able to increase our level of confidence of which features were related to the salary, but still have an inaccurate model.

I tried one last route, still utilizing bins, but utilized pandas’ qcut function to create my bins instead. What this does is discretize my variable into equal-sized buckets based on rank or based on sample quantiles.

The results were similar and even a little less accurate than the first binning method. The model using all features resulted in an R2 value of .422 and an F-stat of 26.07. Reducing the features again based upon their p-values resulted in just 8 features instead of the original 22. This second model ended up being the least accurate of all.

A summary of all models is below:

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name and Target | Number of Features | R2 Value | F-Stat Value |
| m, Salary | 1 (Win Shares) | .287 | 301.0 |
| m\_all, Salary | 22 | .480 | 32.92 |
| m\_sel, Salary | 10 | .468 | 70.13 |
| m\_1, Salary | 8 | .465 | 86.85 |
| m\_new, bucket (first) | 22 | .432 | 27.20 |
| m\_new1, bucket (first) | 8 | .414 | 70.52 |
| m1, bucket (second) | 22 | .422 | 26.07 |
| m2, bucket (second) | 8 | .405 | 68.09 |

#### Final Thoughts:

In conclusion, it is obvious that it is very difficult to predict a players salary strictly based upon their on court performance. At best, we are able to account for just under half of the variance in a player’s salary from their on court statistics. As a reminder to my hypothesis:

Hypothesis:

H0 : There is no correlation between a player's on court performance and salary.

H1 : Salary is correlated to a player's on court performance.

We are able to say, with confidence, that there is a correlation between a player’s on-court performance and salary, therefore rejecting our null hypothesis.

This analysis was able to provide insight that a player’s salary is based on much more than just their on court performance, and potentially, just half of their salary is based upon their performance. This analysis shifted in focus slightly after discovering that the Win Shares stat was not a quality predictor of salary. This led to analyzing and testing to find out what performance metrics from a player contributed to a player’s salary and if we could predict salaries using only a player’s on-court production. In the end, the model’s were not very accurate for prediction, but provided insight that the salaries were formed based upon much more than just their performance on the court.