Milestone Report:

NBA Player Value: Win Shares and Salary

Where should an organization put its money to generate the best chance at success? A question asked by businesses and professional sports teams alike; with the most accurate answer leading to favorable outcomes. I took a deeper dive into the NBA and analyzed players based on their productivity and their salary. Can we determine a player's true value to an organization in regards to specifically their on court production? What expectations do various salaries bring? Based upon production of a player, can we predict their value?

This analysis aims to answer those questions and more to provide insights for organizations to increase their rate of success.

Problem Statement: To predict a player’s Salary utilizing a player’s statistics from on court performance.

Sources of Data and Background Information

The Data:

It's common for NBA fans to research statistics and information regarding teams and players within the organization. The best part (and possibly the least appreciated) is the fact that there is SO MUCH data publicly available. I (and many others) find themselves on Basketball-Reference.com more frequent than others as it is a trusted and reliable source of statistical information for players and teams. For this analysis, I utilized NBA Advanced Player Statistics as well as their contract information. This data is available and kept up to date on a consistent basis.

More Info:

One caveat to my dataset is that is is only up to date through 4/6/2019. This means that it did not include all statistics through the entire regular season. This analysis is largely relying on one particular data point for NBA Players, Win Shares (WS). This number is something that is calculated by Basketball-Reference and you can find more information regarding that by following this link. In short, it is this: Win Shares is a player statistic which attempts to divvy up credit for team success to the individuals on the team.

Utilizing player statistics and contract information, we can determine a lot of information regarding a players value to an organization, in theory.

Cleaning the Data:

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.

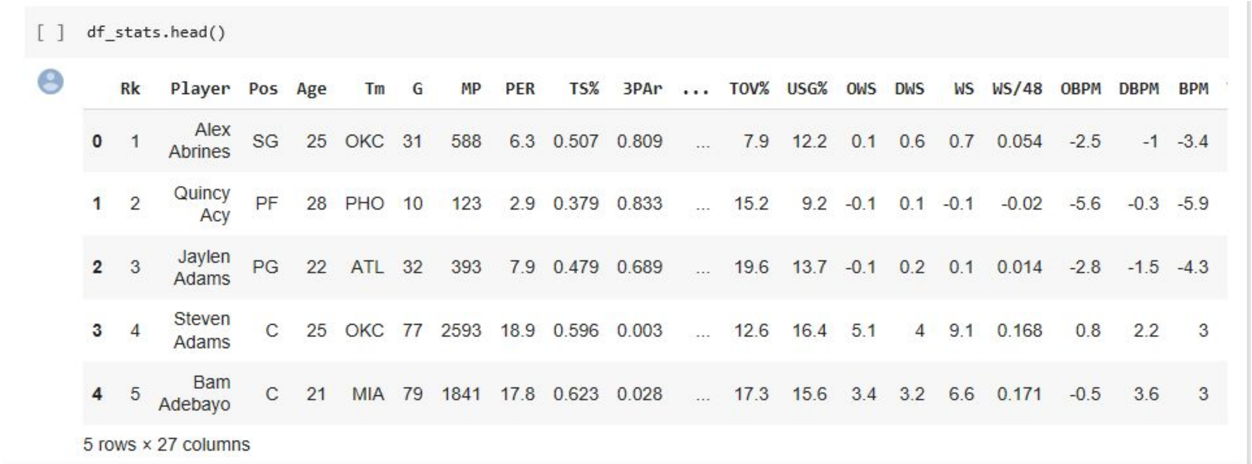
Through a process of writing a loop to read in each line of the data table one by one from the HTML code, and then saving the table in csv format, the goal was to recreate the data from basketball-reference.com whenever the analysis was needed. Using the help of my mentor, we created the loop to parse 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.

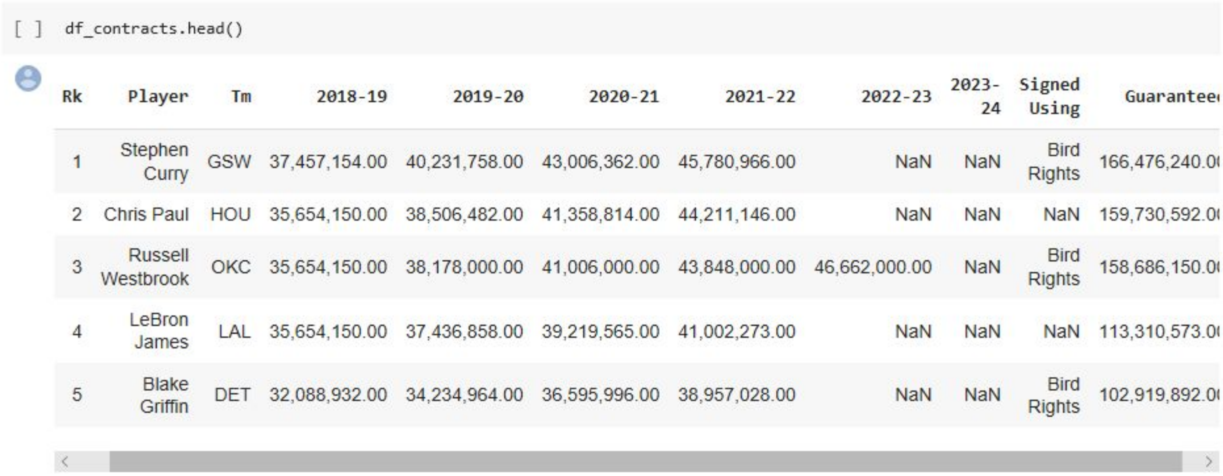
The most difficult part of my data wrangling process was trying to find a way where the dataset was up to date as the website itself updated. However, after many attempts and roadblocks, the best decision was to just save the data in its most current state to utilize for my project.



```
[ ]  df_stats.head()
```

	Rk	Player	Pos	Age	Tm	G	MP	PER	TS%	3PAr	...	TOV%	USG%	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM
0	1	Alex Abrines	SG	25	OKC	31	588	6.3	0.507	0.809	...	7.9	12.2	0.1	0.6	0.7	0.054	-2.5	-1	-3.4
1	2	Quincy Acy	PF	28	PHO	10	123	2.9	0.379	0.833	...	15.2	9.2	-0.1	0.1	-0.1	-0.02	-5.6	-0.3	-5.9
2	3	Jaylen Adams	PG	22	ATL	32	393	7.9	0.479	0.689	...	19.6	13.7	-0.1	0.2	0.1	0.014	-2.8	-1.5	-4.3
3	4	Steven Adams	C	25	OKC	77	2593	18.9	0.596	0.003	...	12.6	16.4	5.1	4	9.1	0.168	0.8	2.2	3
4	5	Bam Adebayo	C	21	MIA	79	1841	17.8	0.623	0.028	...	17.3	15.6	3.4	3.2	6.6	0.171	-0.5	3.6	3

5 rows × 27 columns



```
[ ]  df_contracts.head()
```

Rk	Player	Tm	2018-19	2019-20	2020-21	2021-22	2022-23	2023-24	Signed Using	Guarantee
1	Stephen Curry	GSW	37,457,154.00	40,231,758.00	43,006,362.00	45,780,966.00	NaN	NaN	Bird Rights	166,476,240.0
2	Chris Paul	HOU	35,654,150.00	38,506,482.00	41,358,814.00	44,211,146.00	NaN	NaN	NaN	159,730,592.0
3	Russell Westbrook	OKC	35,654,150.00	38,178,000.00	41,006,000.00	43,848,000.00	46,662,000.00	NaN	Bird Rights	158,686,150.0
4	LeBron James	LAL	35,654,150.00	37,436,858.00	39,219,565.00	41,002,273.00	NaN	NaN	NaN	113,310,573.0
5	Blake Griffin	DET	32,088,932.00	34,234,964.00	36,595,996.00	38,957,028.00	NaN	NaN	Bird Rights	102,919,892.0

The Process:

Initially, I began with two datasets; one for the statistics for the active players in the 2018-19 NBA season and one for their contracts and salary data. Cleaning the data mostly included getting the columns in their respective data type for analysis purposes. An additional step included removing a header row that was in the tables from Basketball Reference. To identify and remove those rows that were headers, I found a unique identifier to the header row and called upon that to remove those. The next step was to combine the datasets. This was completed by using the NBA Players' names as a reference point to combine the statistics with contract data.

One of the interesting things to make note of here is that there are times where players are traded during the regular season and this resulted in the same player having statistics for that particular team. I treated these instances as they were, additional data points. Sometimes a player may contribute and play differently on a new team, and so there may be more than one data point for a player if he were traded midseason.

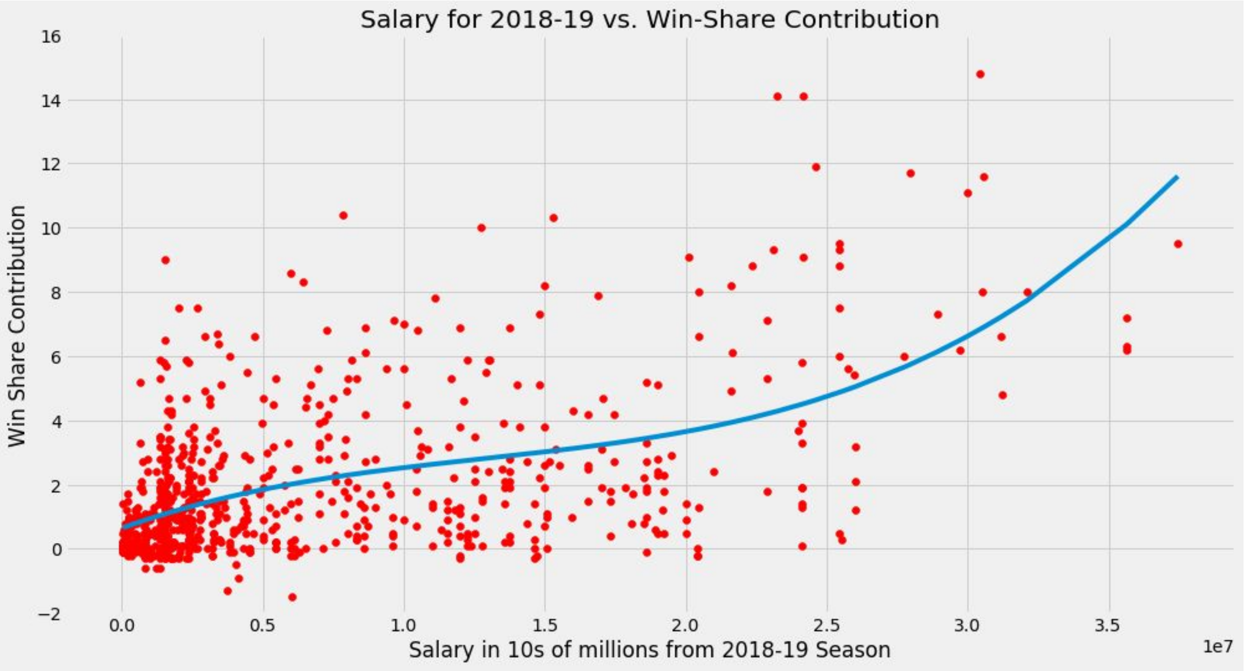
Another notable piece of information regarding the data here is that there were some null values initially after the merging of the two datasets. The number one reason here is that there were players who had short term, 10 day contracts, and may have never played a game. This would result in "contract" data but no statistical information. For these situations, and to make the results as accurate as possible, I removed the players who didn't have an actual contract (more than 10 days) AND didn't play any games to accumulate any stats.

What was left was a clean dataset that had null values in the contract section more than anything. There is a simple explanation here: 40% of the NBA Players will be free agents after this season ended. Therefore, there are many players who did not have contract data after this current season.

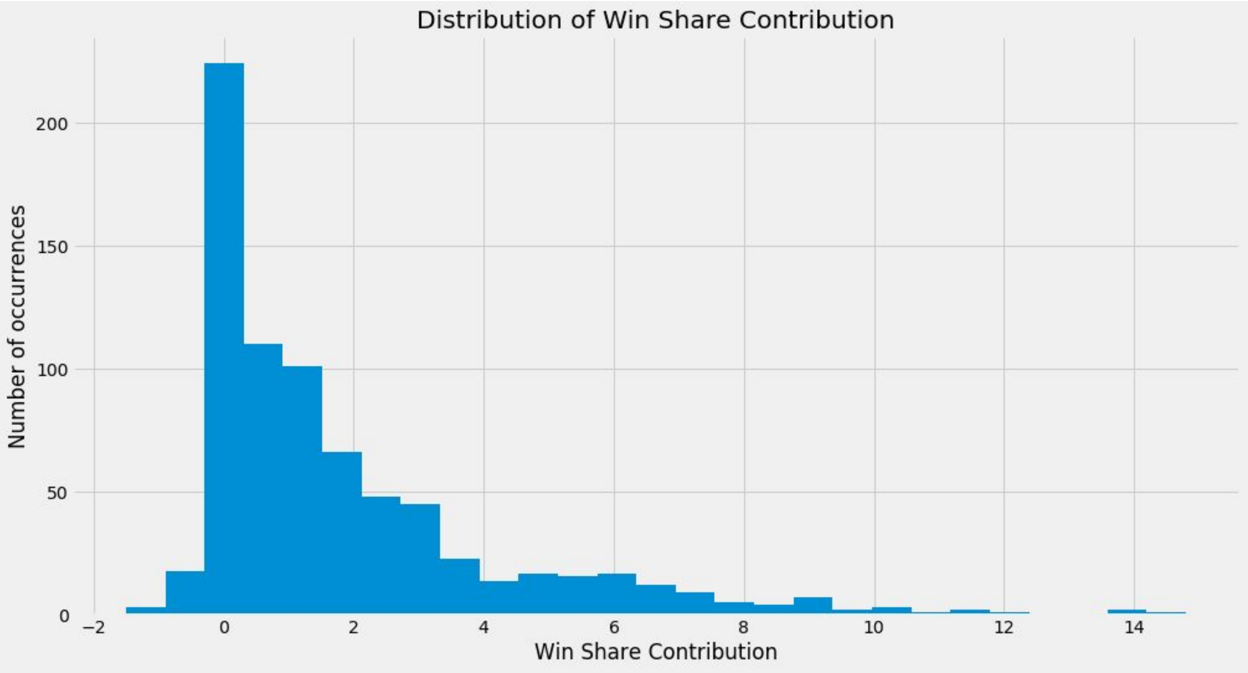
Initial Analysis:

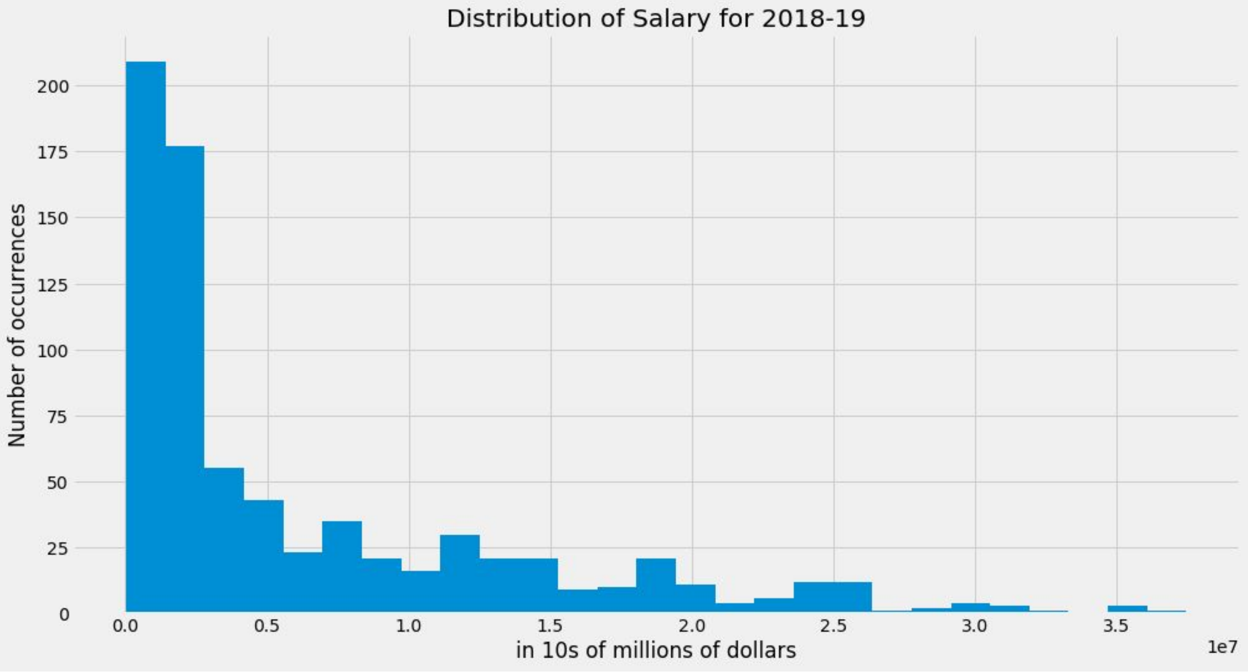
The goal is to utilize the players statistics, primarily Win Share contribution, to predict the salary that these players should have. Utilizing the information in this dataset to train and test a linear regression model to predict the 2019-20 contracts that players will have is going to be able to be tested on a hold out set, the current 2020 salary information.

An initial plot (below) of the Win Share and 2018-19 salary data can show that there is a correlation between the two.



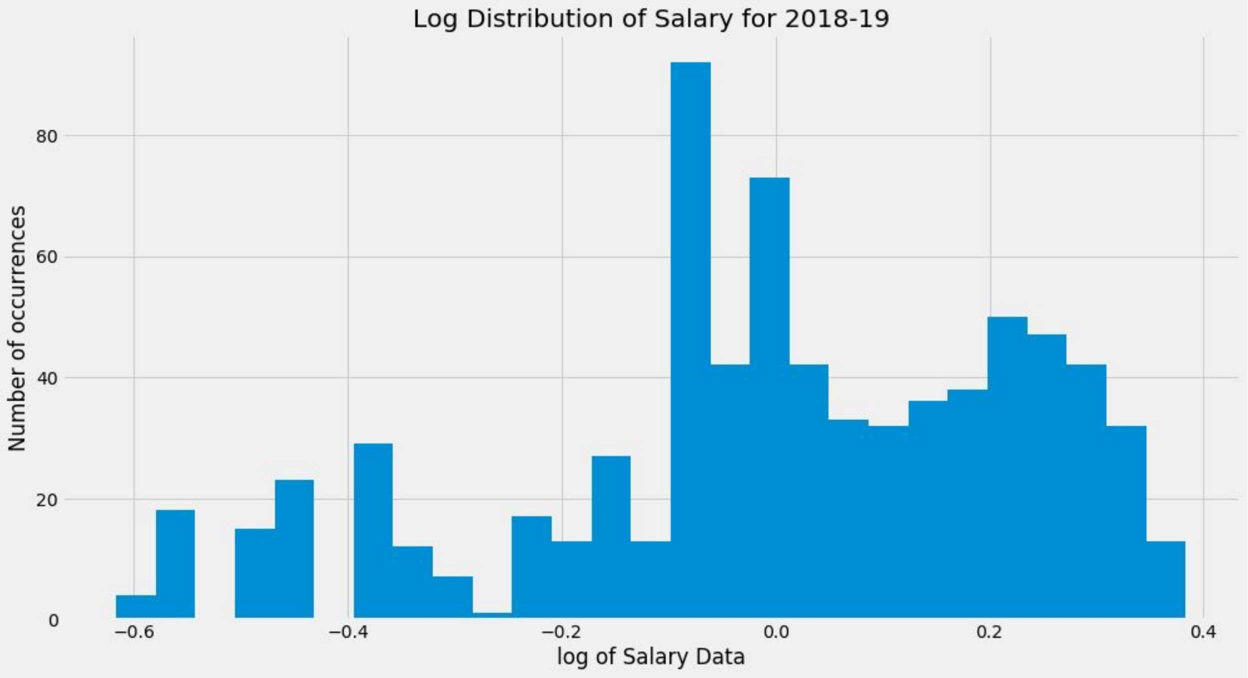
Something to note here is the amount of the data that is in the early portion of the scatter plot above. To further emphasize the skew, the following histogram plots show the data for Win Shares and Salary.





These plots are showing a skew as many of the players are not receiving contracts greater than $5 million/year and have a Win Share contribution of less than 2.

The skew in this Salary histogram plot can become more normalized through a transformation of the Salary data. This helps with showing the incremental change instead of the absolute value change in salaries.



A further dive into testing and machine learning will follow.

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:

H_0 : There is no correlation between Win Shares and Salary
H_1 : 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), 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 H_0)	p= 0.000
Pearson's correlation:	0.53543
Samples are correlated (reject H_0)	p= 0.000
Kendall correlation coefficient:	0.362
Samples are correlated (reject H_0)	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.