NBA Analytics Project

Springboard Capstone Project #1

Milestone Report

Presented by Brian Zamkotowicz

January 2019

Problem Description

The purpose of the following project is to analyze readily available data from the National Basketball Association (NBA) in the hopes of gaining insight into statistics that may have predictive value for determining wins and losses in an upcoming season. From a business perspective this information could be useful to NBA general managers in the construction of their team. If a relationship between certain statistics and a positive change in wins in the coming season can be proven, then the predictive value of those statistics can be used as basis for off-season player personnel decisions. Essentially, what will be determined is which statistics are most relevant when making business decisions regarding team construction.

Data Acquisition and Wrangling

The first thing necessary before proceeding with the analysis was to find relevant data. Because basketball, and sports in general, are very statistics-oriented much of the information was readily available on the internet. In this case, the website basketballreference.com provided a wealth of information. In addition to well known statistics like points, rebounds, assists, locks and steals, which even a casual basketball fan is thoroughly familiar with, they also provide a number of advanced metrics. Because these advanced metrics are intended to provide a clearer picture of how players perform relative to one another, I chose to focus on several of these advanced statistics as a jumping off point.

In choosing the stats to examine I decided to focus on PER, WORP, VORP, and Box Plus Minus. I chose these particular stats for a number of reasons. Each of those stats is supposed to effectively gauge players against one another so if a player has a higher number he is supposedly "better." Also each of these statistics has been used for a number of years, and are therefore at least somewhat familiar to some sports fans. Each of these advanced metrics tries to combine more basic statistics by weighting them in a way that hopes to determine who the best players are.

One metric examined is PER (Player Efficiency Rating),developed by ESPN's John Hollinger. It seeks to determine how effective a player is in minutes spent on the court. One potential issue with this metric is that there is no way to separate players who play well in limited minutes (often older players) from players who apply that same efficiency over a large portion of the game. WORP translates to wins above replacement player   . In theory if Player X has a WORP of three, he should contribute to his team winning 3 more games than if they had a league average player at his position instead of him. VORP, or Value Over Replacement Player, also seeks to compare player's to the league average in a meaningful way. Finally, I chose to examine Box Plus Minus (BPM). Box Plus Minus attempts to quantify a player's contribution per 100 possessions. A box plus minus of +5 means that a team with player X on the floor instead of a league average, would should score 5 extra points over 100 possessions. More information on BPM can be found here (https://www.basketball-reference.com/about/bpm.html).

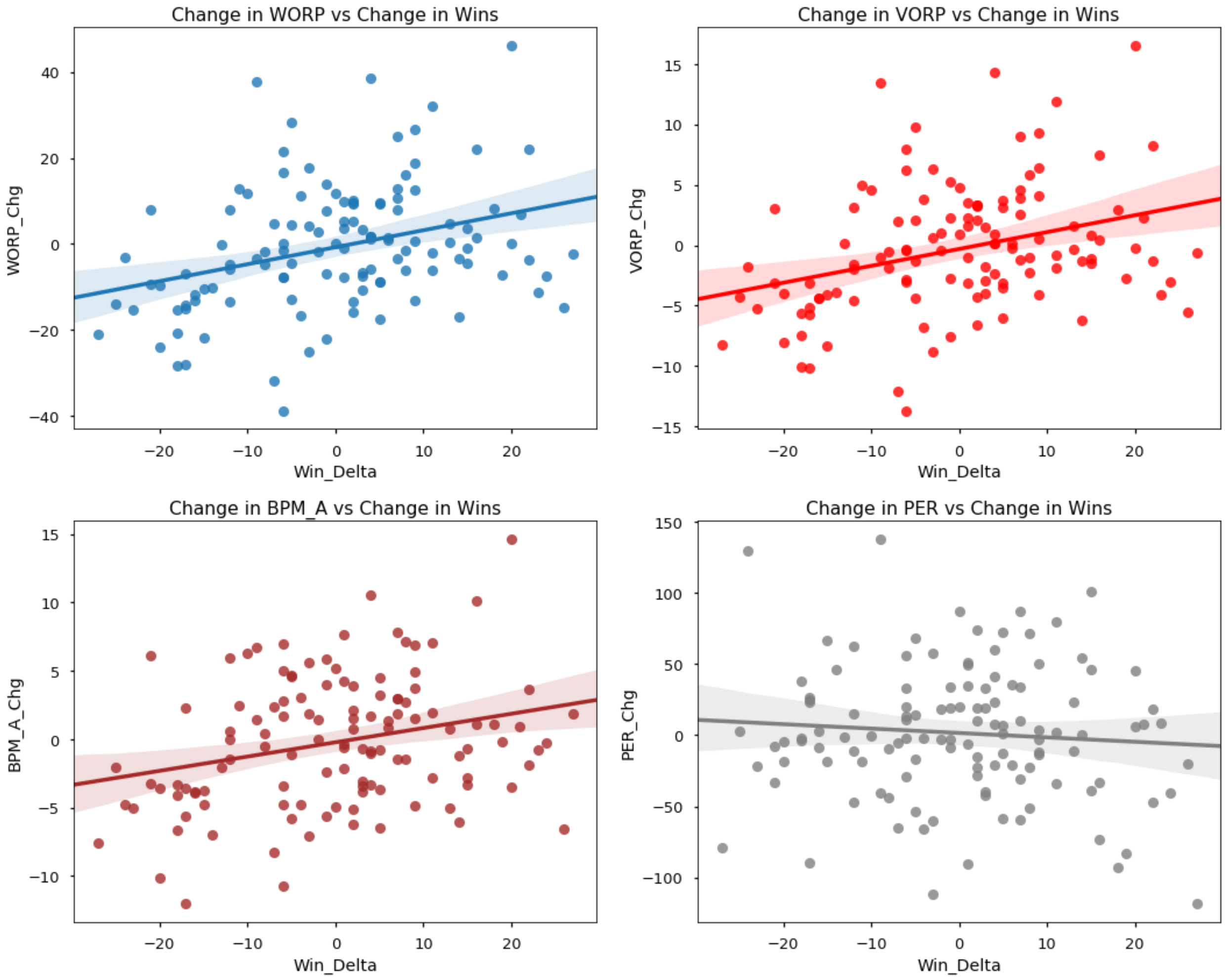
Thanks to the ready availability of sports statistics, finding a data set to begin working with was relatively simple. The following spreadsheet found on basketballreference.com contained all of the advanced metrics I had chosen, as well as many other pieces of useful information BPM Excel Spreadsheet. Once the data was found it was necessary to organize it in a way that would be useful for my specific examination. The first issue became obvious when doing a simple count of team names. I had chosen to explore team data from the period of 2011 to 2015, and in this time several teams had moved or changed names. Specifically, there were issues with the Nets, Hornets and Pelicans. These were resolved by appropriately grouping the new teams back to their original moniker. I then extracted the categories I planned to work with into a new Pandas dataframe. At this point I also chose to make an adjustment to the BPM statistic. Since BPM was a per 100 possession measure, and I thought it was important to look at how many minutes (per season) the player contributed at that level, I created a new stat BPM_A (Box Plus Minus Adjusted) to account for playing time.

The next decision dealt with rookies. College statistics do not translate well into the NBA. In truth, a lifetime of statistical analysis could probably be dedicated determining how college players will perform in the NBA, but that is outside the scope of this project. I chose to address this problem by adding a new entry for each rookie in the year before they entered the league with the same statistics they generated in their rookie year. In this way I made it appear that each rookie basically performed as expected, since this was not a variable that could be addressed in the offseason as part of our business problem anyway. Rookies essentially became a sort of control group within the examination.

Once I was confident that I had workable data I set to work on aggregating it into the form needed. Since what needed to be examined was changes in WORP, VORP, BPM_A and PER of teams from year to year I aggregated the data by team and year. I then wrote a function that could figure the difference in each stat (or any category) from the previous year and add it as a new column of the dataframe. This was also useful for compiling the change in wins from the previous year (hereafter referred to as Win_Delta.) With appropriately clean data on NBA season 2012 through 2015, that now included information on changes from the previous season, it was time to start exploring the information contained within the data set.

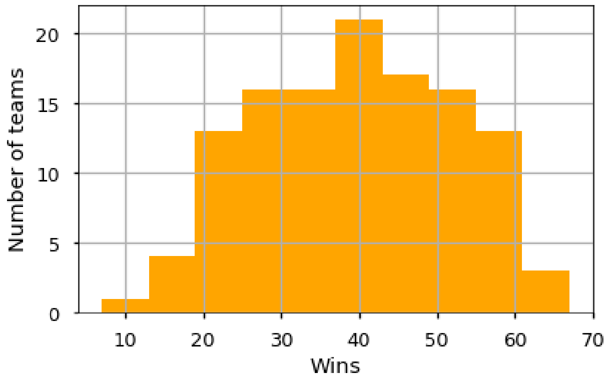
Data Exploration and Storytelling:

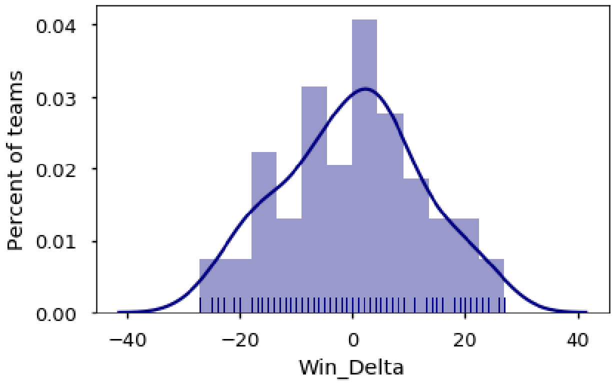
The next step once the data was in order, was to begin to see if a relationship could be established between player statistical measures and change in wins in the next season. In order to establish that relationship, I chose to look at each of the aforementioned statistics and how it relates to Win_Delta in a regression model:



In the scatterplots changes in WORP, VORP, and BPM_A show a positive correlation to change in wins, while PER showed no, or even a slightly negative relationship to Win_Delta.

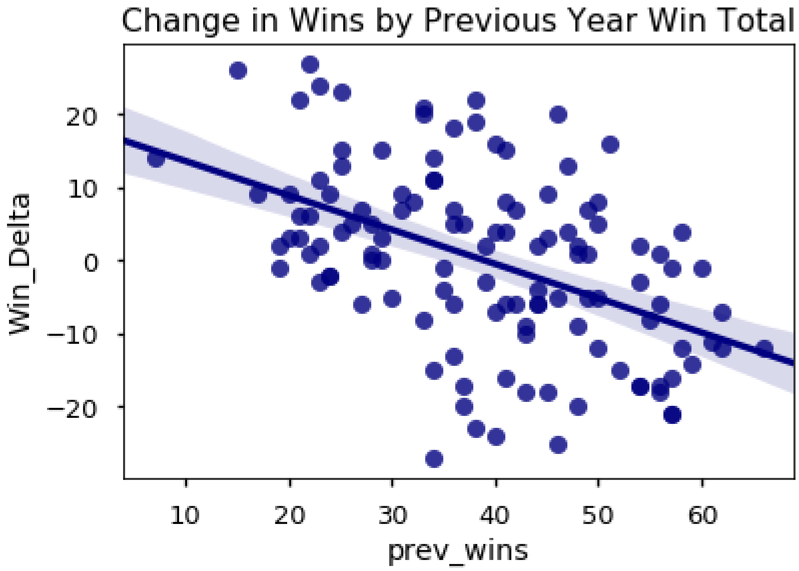
Another factor explored was the distribution of wins and change in wins throughout the NBA:



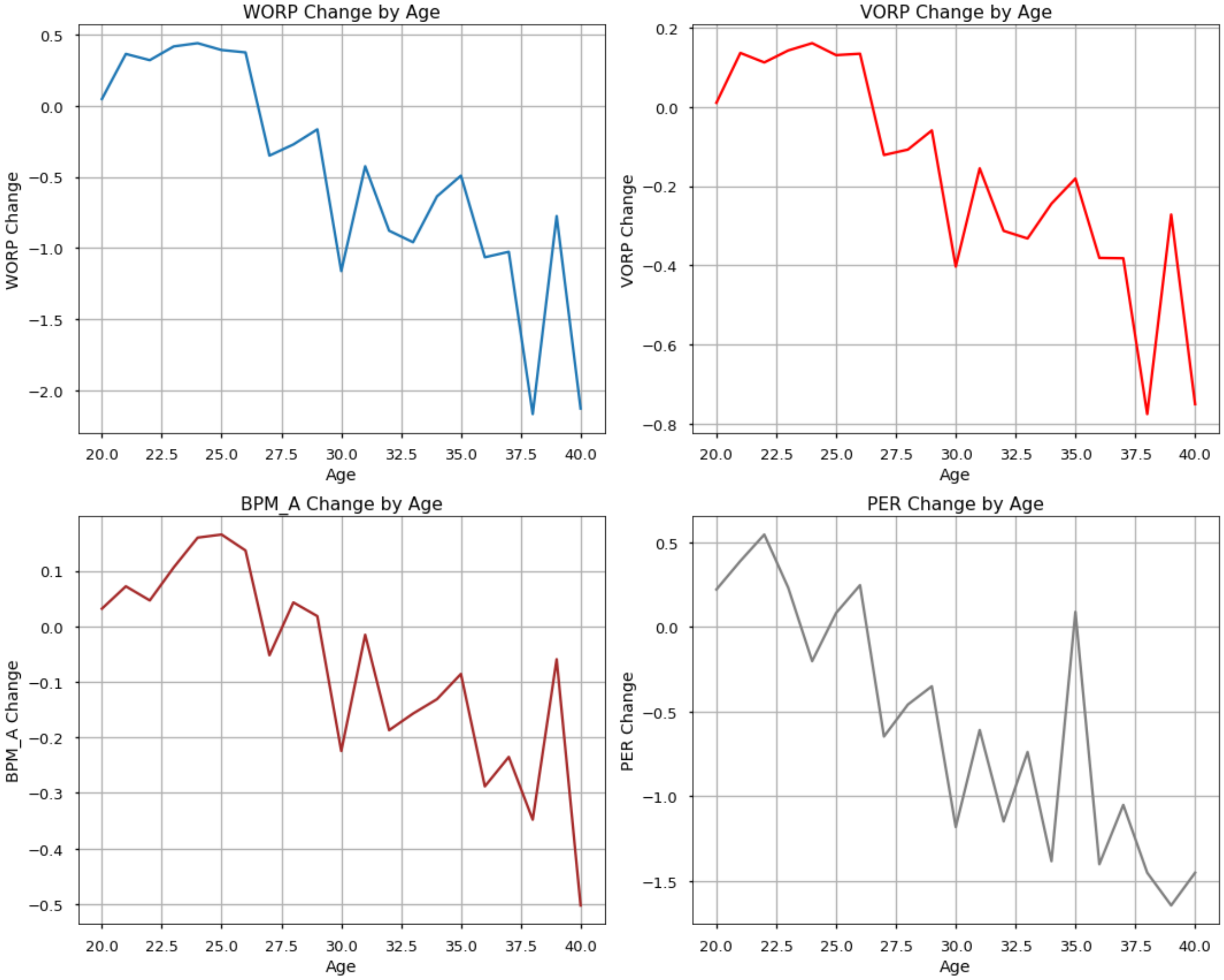


The first graphic explains that the majority of NBA teams win about half of there games and the distribution plot on the right shows that large changes in win totals from year to year are not the norm.

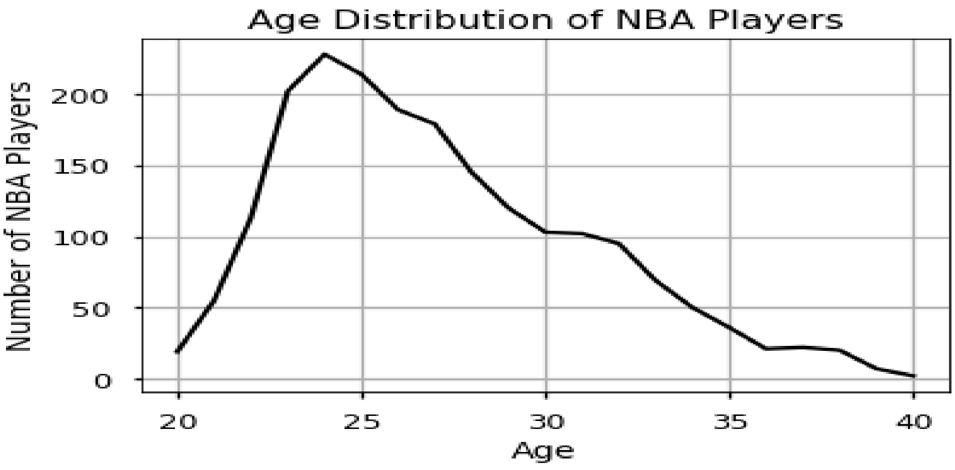
It is also worth noting, that perhaps the biggest factor in Win_Delta may have been wins in the previous year. A regression model shows that NBA teams tend to regress back towards the mean of winning approximately half of their games.



I also explored the relationship between age and the advanced stats in question, in the hopes that perhaps age could be a criteria for team construction, or at least be of predictive use. The results were surprising. I expected statistical prowess to peak and then drop off in a bell curves fashion, but this was not the case. As some of the oldest players in the league seemed to be some of the best statistical performers. I then examined the yearly change in those metrics by age, again thinking that I would see a peak and roughly bell shaped curve.



This was more along the lines of what was expected but did show some anomalous spikes. Further examination of the data showed that there were so few players at advanced ages that one exceptional performance could really affect the data. This was shown by the age distribution of players in the sample.



Applications of Inferential Statistics

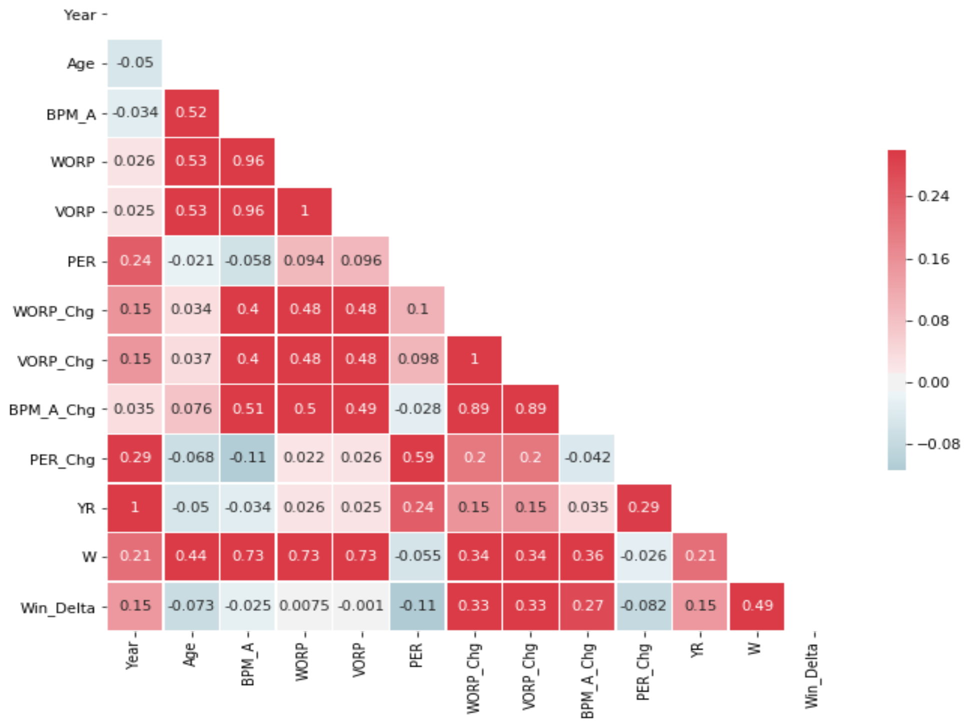
After a visual examination of the data, it was dig a little deeper into some of the information gained through that analysis. I previously had discovered through a regression plot that WORP, VORP, and BPM_A appeared to be correlated to Win_Delta. An examination of their Pearson Correlation Coefficients further explored that relationship:

WORP_Pearson = (0.3340651515935038, 0.00019230714271517446)
VORP_Pearson = (0.3270537732683617, 0.0002663617564581372)
BPM_A_Pearson = (0.27241342726842066, 0.0026120697699333027)
PER_Pearson = (-0.0815333961917619, 0.37600556347916425)

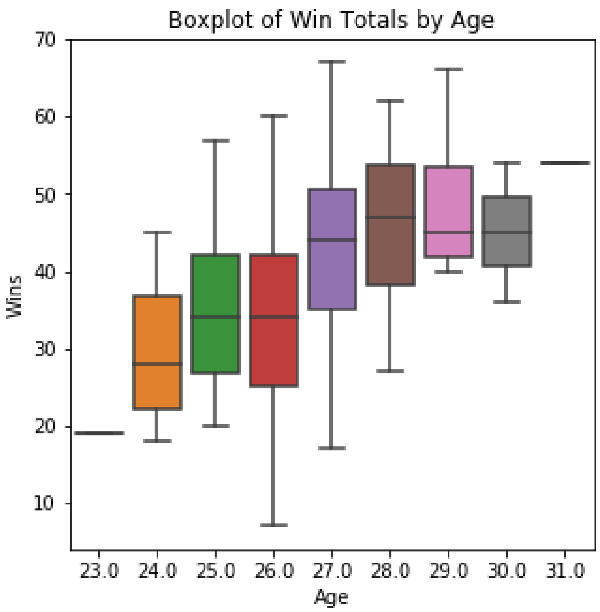
I then broke the league into thirds based on wins to see if I could determine statistical significance between these stats and win delta. On initial testing I was able to rule out the null hypothesis that WORP did not have a statistically significant relationship to Win_Delta, but tests of the other statistics came in above my α of 0.05, meaning the null hypothesis could not be ruled out.

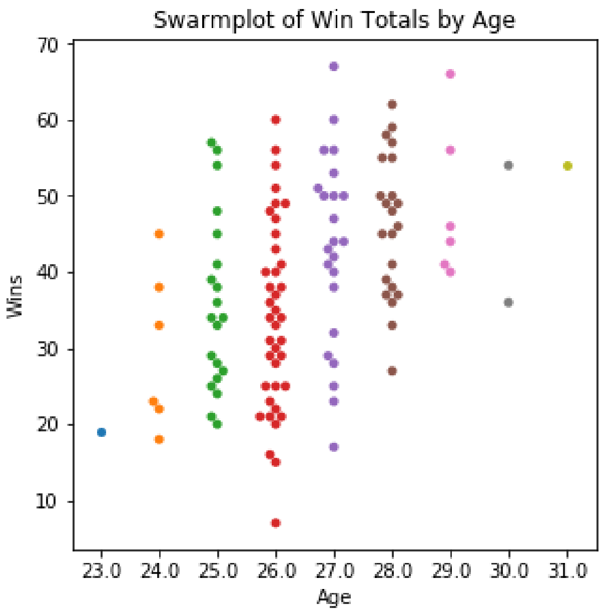
I believed that the result may have come from dividing the league into thirds, leaving me with groups that were too similar to each other. I divided the league into quarters and retried the test. This time I was able to invalidate the null hypothesis for WORP, VORP, and BPM_A. PER once again did not show a significant relationship.

I also created a heatmap showing the Pearson correlation coefficient of the stats contained in my dataframe:



Since average team age and wins appeared to show a positive correlation on the heatmap, I decided to take my exploration of age a bit further. I thought that there might be a peak age that NBA general managers could aim for when building a team, and once again I felt that the distribution of wins by age might be bell curved. The following plots show win distribution by age.





I was once again surprised not to see any drop off amongst older teams. I did discover that teams with average ages under 27 tend to lose more games than they win though.

Next Steps

Since there seems to be a quantifiable relationship between changes in WORP, VORP and BPM_A and change in wins, that relationship needs to be explored further. Using linear regression I hope to build a model that can look at multiple factors and predict a team's change in wins. This model could provide a tangible solution to the business problem of optimizing an NBA team through free-agency in the off-season in order to win as many games as possible.