Cole Woods Capstone 1 Consolidated Report

- What is the problem?
    - The problem I want to solve is to determine what statistics in the NBA are most highly valued.
- Who is my client?
    - My client is players in the nba and the respective organizations.
    - The reason my client cares is that this will allow them to determine what is currently valued in play.
    - Players and organizations won't be able to see the likelihood of obtaining exactly what they want just based on simple statistics about their play and the analysis I'm attempting to provide will greatly help.
- What data am I using?
    - The data I am using is as follows:
        - Player names
        - Salaries
        - Position played
        - Age
        - Team
        - Games played
        - Minutes played
        - Player Efficiency Rating
        - True shooting percentage

        - Total rebound percentage
        - Assist percentage
        - Steal percentage
        - Block percentage
        - Turnover percentage
        - Usage percentage
        - Offensive win shares
        - Defensive win shares
        - Total win shares
        - Box plus/minus score
        - Value over replacement player
    - I will obtain this data from kaggle for the most part, but extra digging will be done.
- How will I solve this problem?
    - First I will I gather all the data that I could potentially want to use.
    - Second I will organize the data from each set in a way that allows me to use it.
    - Third I will combine the data sets in order to pull from them
    - Fourth I will begin analysis.
- What are my deliverables?
    - Right now I plan on my deliverables being code and a paper.

Data Wrangling Steps

As far as gathering and organizing the data for this project, my steps were not too complicated. In order to prepare to wrangle the data I selected four sources:

- Salary based on Advanced NBA Metrics from kaggle

The wrangling part of things came next and went as follows:

1. I went through and deleted unnecessary columns.
2. I searched for extreme outliers and deleted them.
3. I also went through and deleted rows with zeros
4. I made sure each to determine value to base everything on. In this case it was average salary.
5. I reformatted salary to be in a US currency based format.
6. The final step was just reviewing steps 1-4 and making sure no mistakes were made.

There was only 1 outlier that I missed and it was very extreme so I removed that row of data. There were no a few missing values and as with the outliers I had those rows of data removed from the set.

After further looking into the data I have been working with my views and what data is and isn't related has changed. My more individualistic advanced statistics for the NBA would have a greater impact and affect win/loss records. My theory based on the correlation between various statistics and salary was that teams were overvaluing how good a player is holistically and not looking enough at individual statistics and what a team needs based on what they struggle with. After looking more into it though I noticed

that those individualistic statistics had no effect on winning or losing and it was really just the all encompassing stats such as value over replacement player that lead to more likely wins. The steps I took to reach this conclusion went as follows:

- First I created a new dataframe based on the original one I was working with, but this time it was grouped by team averages instead of by each individual player's statistics.
- Next I created several scatter plots. It was at this point I noticed that the trends that I noticed when comparing the in game statistics to salary were very similar to the ones I was currently looking at when comparing those statistics to win count.
- At this point I realized my initial hypothesis was not correct, but I wanted to do some testing, so I ran multiple t tests to try and see if I was just missing something.
- From the multiple tests I gained the following information:
    - The relationship between holistic statistics and win count is very significant.
    - The relationship between individualistic statistics and win count has essentially no significance except for a very, very slight significance with assist percentage.

For the in-depth analysis for my first capstone project I didn't too much and really tried to keep things simple. The following steps were taken in order to get the material that I really needed:

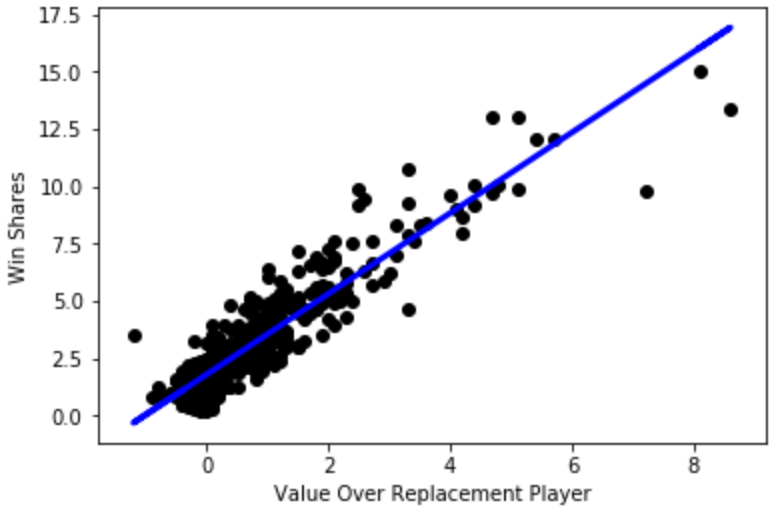
1. First I took the information I had previously gathered in order to determine what I wanted to do my regressions on.
    a. I found after looking further into the scatterplots I had previously created that not only did the wholistic statistics effect win rate more reliably as far a trends go, but the other individualistic statistics couldn't seem to fit any sort of regression reliably.
    b. The conclusion I came to was that I needed for now to only create regression models for win shares as the y value vs. value over replacement player, box plus/minus score, and player efficiency ratings as the x values.
2. The next step was to reshape the data in order to be able to set up my regressions. The data was in a 1d array so this was a necessary step for being to use regression models from scikit-learn.
3. The fourth step I took was preparing the different regression models for the three statistics I was looking at more in depth. I set up a linear model for the three major statistics as they relate to win shares.
    a. My findings were interesting for the regression models I created. I found the following:
        i. The first thing I found was that while all three of the statistics had a positive impact on win shares, player efficiency rating had by far the least significant impact of the three.

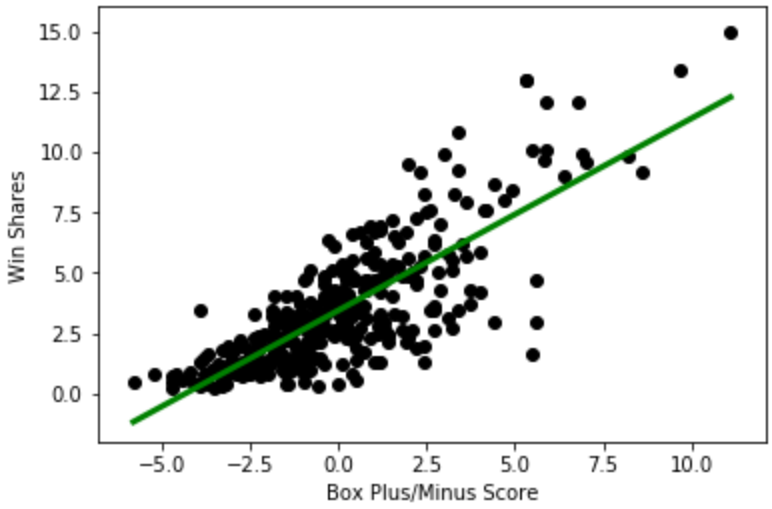
ii. Next I realized that the trend lines compared between box plus/minus score and value over replacement player while being similar were not nearly as similar as I thought they'd be. Those two statistics are similar in that value over replacement player is directly related to box plus/minus score as far as the formulas go. Value over replacement player was the statistic, however, that was clearly the most significant.

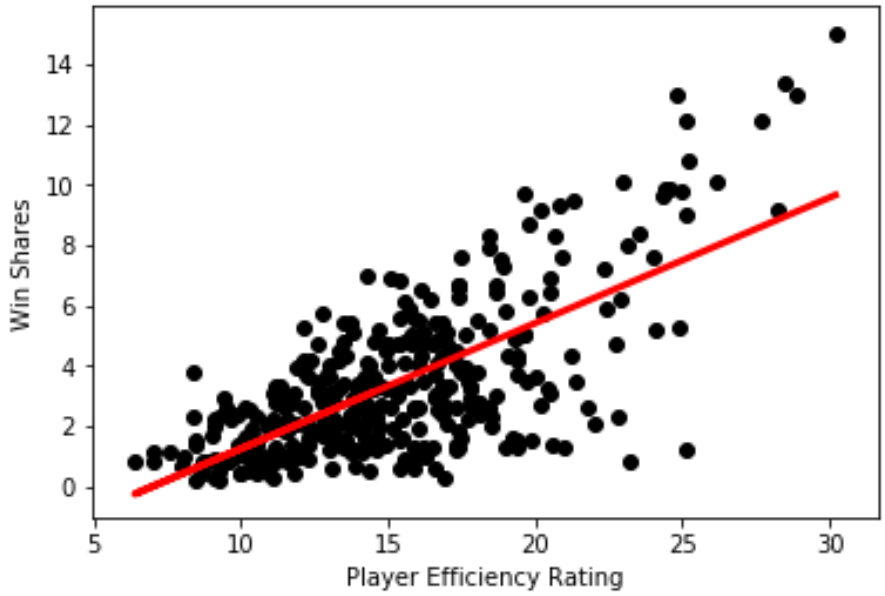
After completing the analysis and looking more into the data I was working with I further developed my hypothesis. Originally I had thought that teams were undervaluing individualistic statistics based on the correlation between those statistics and salary and that teams should look more into the needs of the team individually rather than just getting the best overall players possible. After doing more research I found that individualistic statistics had basically no effect on how much a player was contributing to a team's wins. I realized that wholistic statistics were much more important. After doing more research though I found that two of the wholistic statistics I was looking at in box plus/minus score and value over replacement player were based mostly on the individualistic statistics I looked at. This leads me to believe that not only are the wholistic statistics more important, but in addition how well rounded you are as a player also will have a relatively significant impact on how much you contribute to your team's win rate.

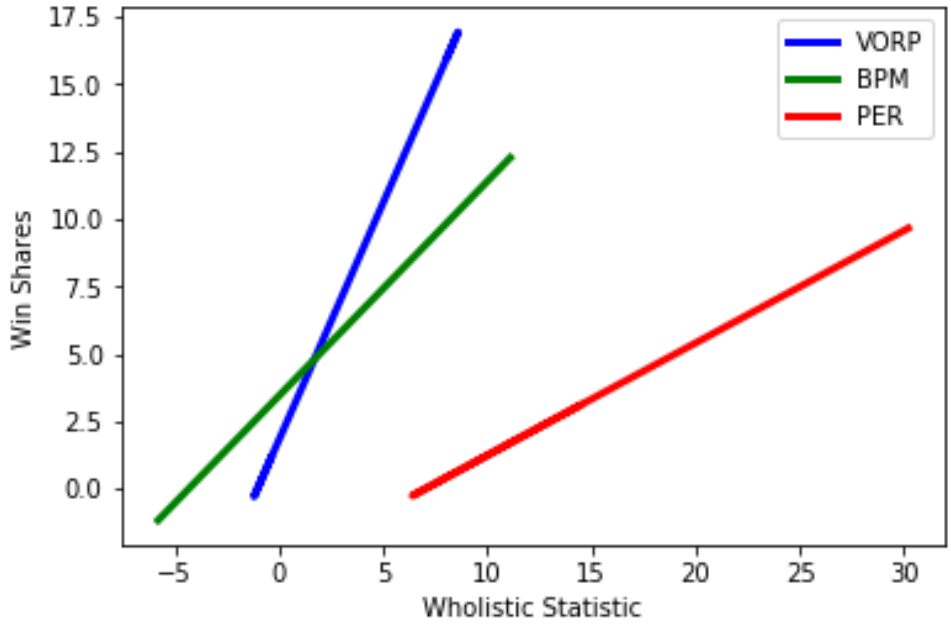
As a follow up to all the information above I must now say that new discoveries were made which didn't change anything directly, but did add to my discoveries. In

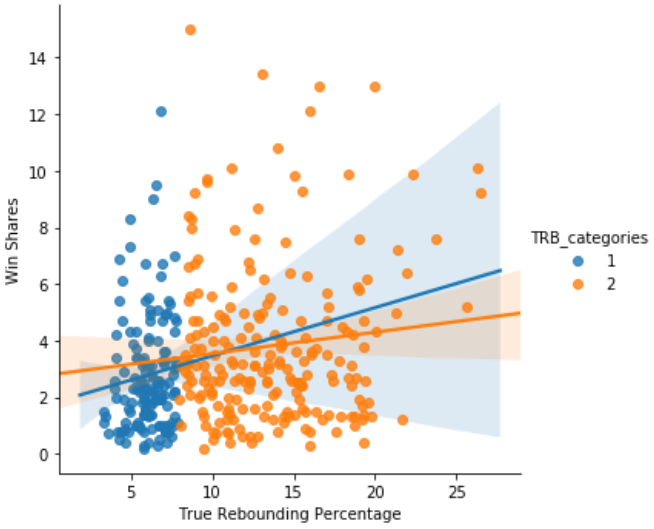
addition the visualizations have updated. The visualizations now depict linear regressions for the most relevant statistics. Those being the wholistic ones. As well, I noticed a trend where there were clusters in the bottom left corner of the individualistic statistics meaning that although the individualistic statistics don't show trends they do indicate that being on the lower end of any stat does lead to a much better chance at lower win shares. Below are the visualizations:

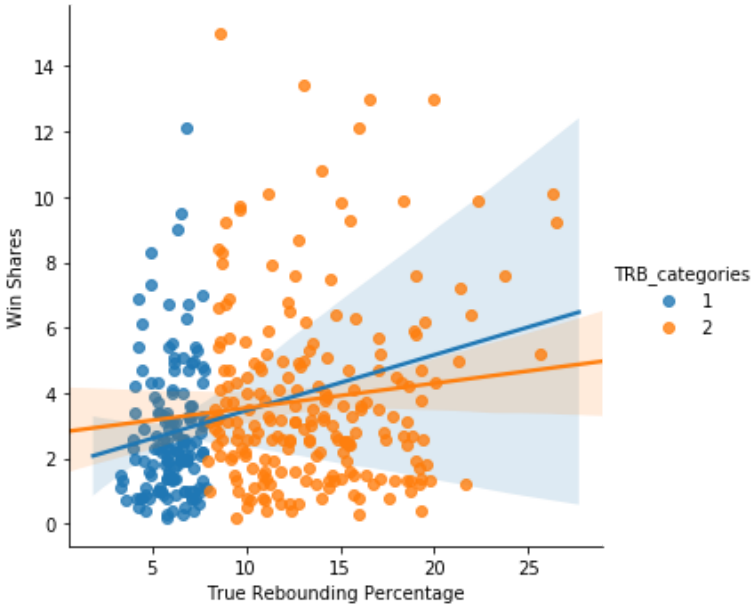


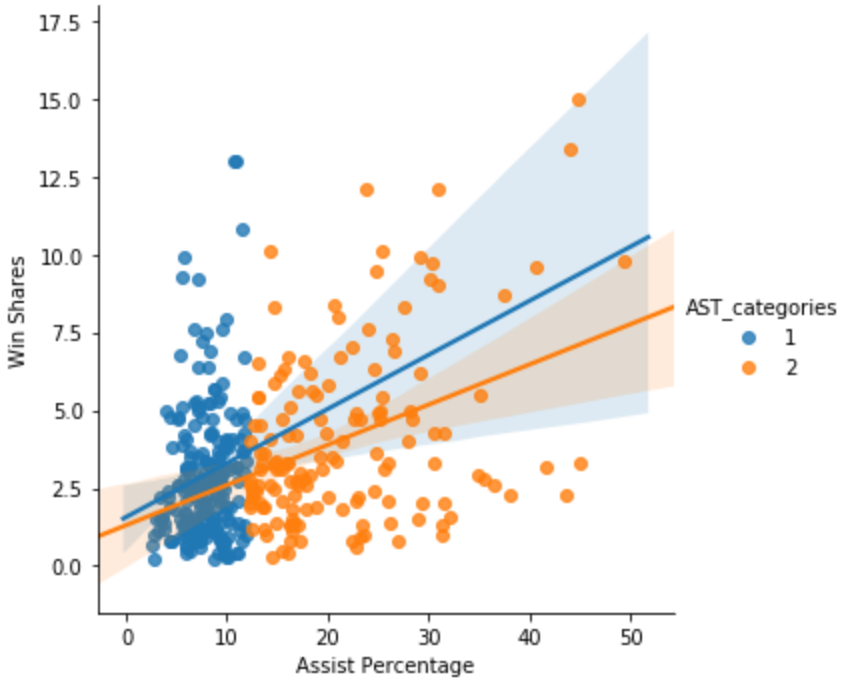


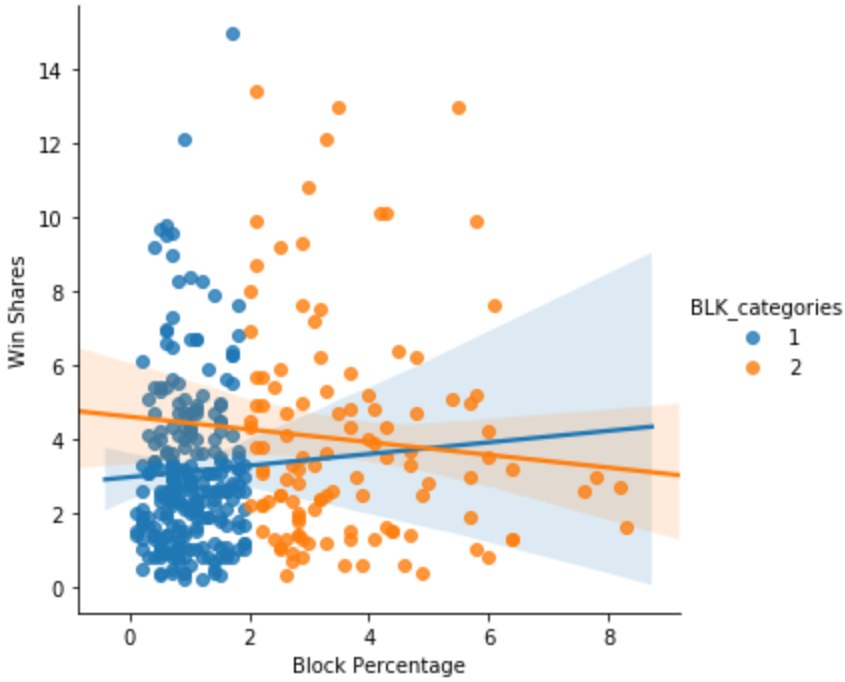

Win Shares
Player Efficiency Rating
0
2
4
6
8
10
12
14
5
10
15
20
25
30

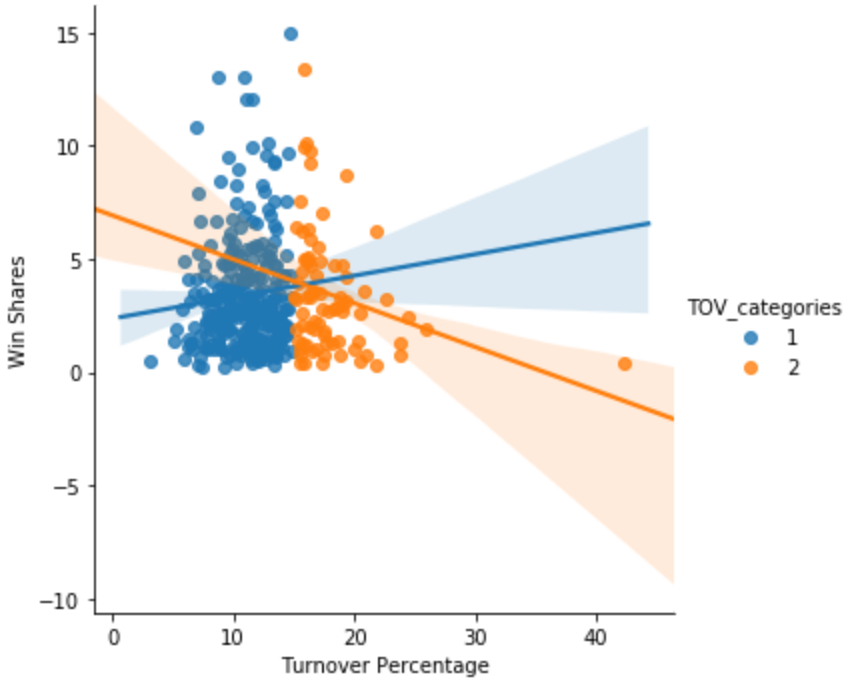

VORP
BPM
PER
Win Shares
Wholistic Statistic
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
−5
0
5
10
15
20
25
30

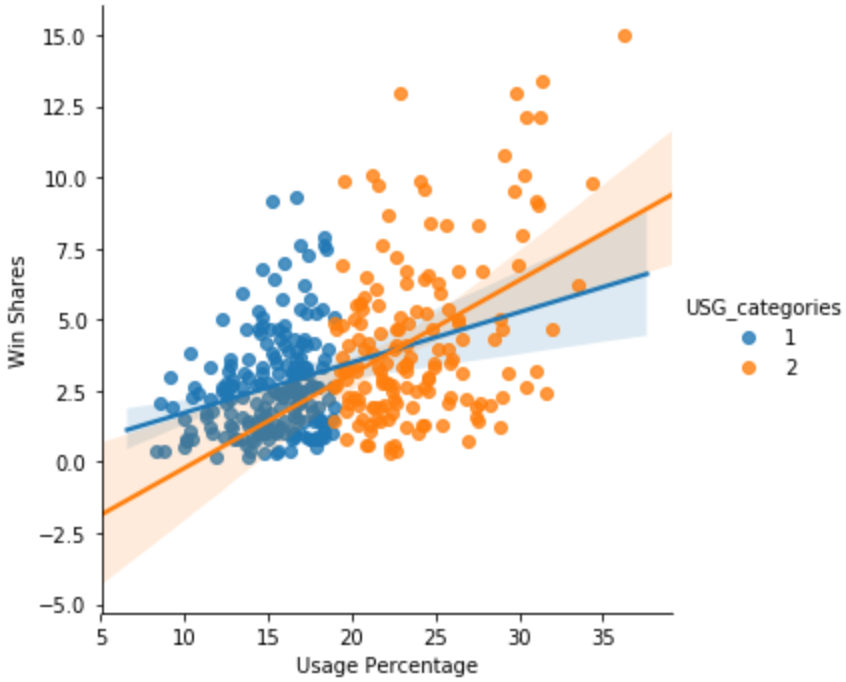

Win Shares
True Rebounding Percentage
TRB_categories
1
2


Win Shares
True Rebounding Percentage
TRB_categories
1
2


17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Win Shares
0
10
20
30
40
50
Assist Percentage
AST_categories
1
2


14
12
10
8
6
4
2
0
Win Shares
0
2
4
6
8
Block Percentage
BLK_categories
1
2


Win Shares
Turnover Percentage
TOV_categories
1
2


Win Shares
Usage Percentage
USG_categories
1
2