Cole Woods Capstone 1 Mile Stone

- 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.

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.

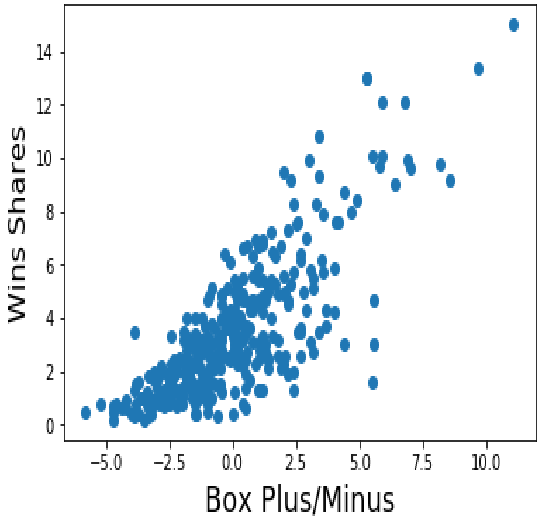
Next I created several scatter plots. It was at this point I noticed that the trends that I noticed when comparing the in game statistics to salary were very similar to the ones I was currently looking at when comparing those statistics to win count.

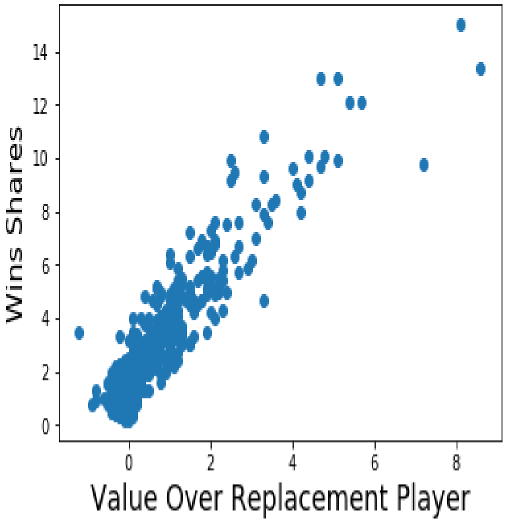
At this point I realized my initial hypothesis was not correct, but I wanted to do some testing, so I ran multiple t tests to try and see if I was just missing something.

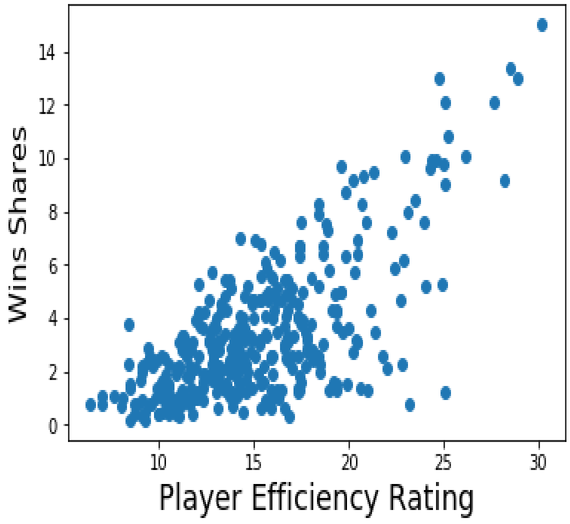
From the multiple tests I gained the following information:

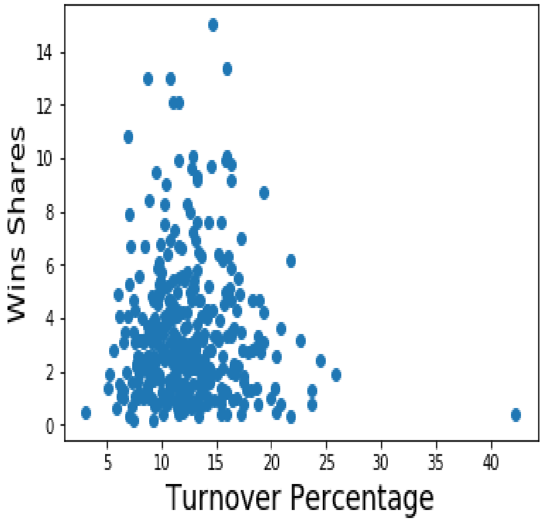
The relationship between holistic statistics and win count is very significant.

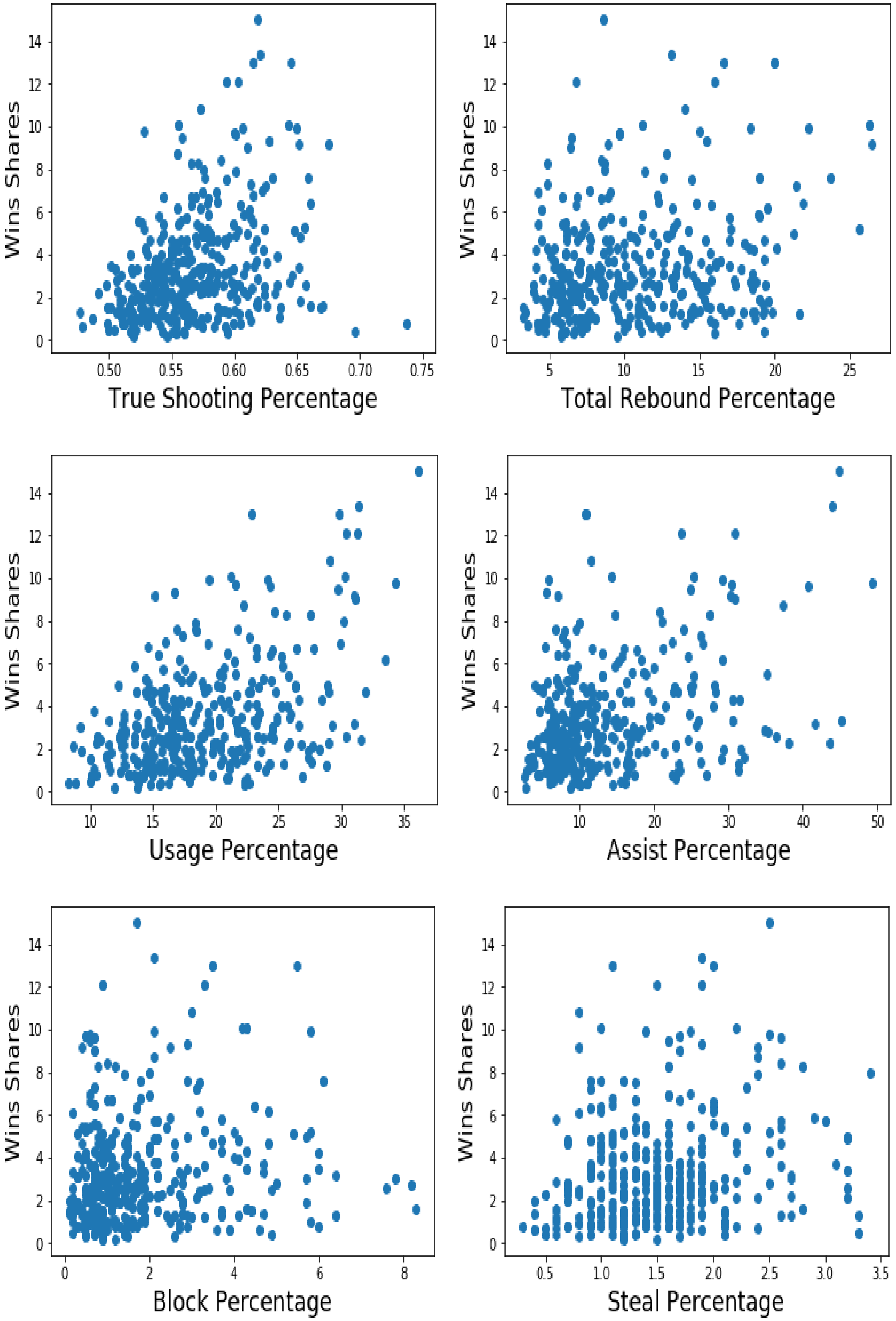
The relationship between individualistic statistics and win count has essentially no significance except for a very, very slight significance with assist percentage.










Wins Shares
0
2
4
6
8
10
12
14
0.50
0.55
0.60
0.65
0.70
0.75
True Shooting Percentage
Wins Shares
0
2
4
6
8
10
12
14
5
10
15
20
25
Total Rebound Percentage
Wins Shares
0
2
4
6
8
10
12
14
10
15
20
25
30
35
Usage Percentage
Wins Shares
0
2
4
6
8
10
12
14
10
20
30
40
50
Assist Percentage
Wins Shares
0
2
4
6
8
10
12
14
0
2
4
6
8
Block Percentage
Wins Shares
0
2
4
6
8
10
12
14
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Steal Percentage