Sam Hatamian

NBA Project

Introduction:

The analysis and discussion of various sports are becoming more and more prevalent in our culture due to the large amount of player and team information that is being presented to the public. Discussions that once centered on individual intuition, aptitude, and team chemistry are now being shifted towards discussions of quantifiable metrics. These metrics are starting to give managers, coaches and common fans confidence in making sports related decisions due to the tangible results that are presented by various techniques of statistical analysis.

I have always had a passion for watching basketball but recently, I wanted to have a better understanding of how basketball related statistics actually impacted the game. Simply focusing on number of points scored, or rebounds made in a game were no longer targets of interest for me, as I wanted more of a holistic metric that accounted for game flow, minutes played, points, steals, turnovers etc. In this article, I decided to calculate John Hollinger's PER (Player Efficiency Rating) per player for every game played in the past couple of seasons. I selected this metric because it takes positive contributions that are made in a game and subtracts negative contribution, while adjusting for game flow while also considering total team output from a player's own team and the opposing team as well. The initial goal of my project was to utilize this metric to perhaps then predict future game outcomes. Unfortunately, with time constraints this goal will have to remain in the works as I instead used this metric to explore interesting characteristics about players and their contributions to the team.

Dataset Explanation:

My original data consisted of individual player data of every game played of every season from 2008 until 2013. Player name, team name, points, assists, rebounds, and steals were some of the basic metrics that were included. There were a total of about 130,000 player observations in the data with each observation serving as a game played. Playoff data was also included but I excluded this data when calculating PER since I wanted it to be consistent to PER season averages that are presented on sites such as ESPN.com and Basketball-Reference.com. However, there was a lot of data that I had to add in order to satisfy the PER metric. Since I wanted to calculate the PER per game per player, I had to calculate a lot of aggregate team data for every team game played. Pace of the game was an important factor that was calculated as well. This looks at own team possession and away team possession over a fraction of minutes played. This way a player's success in other key areas can be impacted and almost weighted by own possession and even how fast the ball is exchanging hands between a team. A player who does well in drawn out game with low turnovers might deserve a different PER then a player who does well in a quick, bucket after bucket, possession changing game. I also calculated every team win and loss for every game so I could use this metric as a test for the quality of my PER rating. In the end of course, wins are what matter the most.

Methodology:

All calculations begin with what I am calling unadjusted PER (uPER). The formula is:

```
uPER = (1 / MP) *
     [ 3P
     + (2/3) * AST
     + (2 - factor * (team_AST / team_FG)) * FG
     + (FT *0.5 * (1 + (1 - (team_AST / team_FG)) + (2/3) * (team_AST / team_FG)))
     - VOP * TOV
     - VOP * DRB% * (FGA - FG)
     - VOP * 0.44 * (0.44 + (0.56 * DRB%)) * (FTA - FT)
     + VOP * (1 - DRB%) * (TRB - ORB)
     + VOP * DRB% * ORB
     + VOP * STL
     + VOP * DRB% * BLK
     - PF * ((lg_FT / lg_PF) - 0.44 * (lg_FTA / lg_PF) * VOP) ]
```

Most of the terms in the formula above should be clear, but let me define the less obvious ones:

```
factor = (2 / 3) - (0.5 * (lg_AST / lg_FG)) / (2 * (lg_FG / lg_FT))
VOP    = lg_PTS / (lg_FGA - lg_ORB + lg_TOV + 0.44 * lg_FTA)
DRB%   = (lg_TRB - lg_ORB) / lg_TRB
```

Once we have our uPER we to calculate pace adjustment, which is equal to league pace divided by team pace.

Pace follows the formula 48 * ((Team Poss + Opp Poss) / (2 * (Tm MP / 5)))

Then we get our aPER = (pace adjustment) * uPER

And finally we calculate PER = aPER * (15/ lg_aPER)

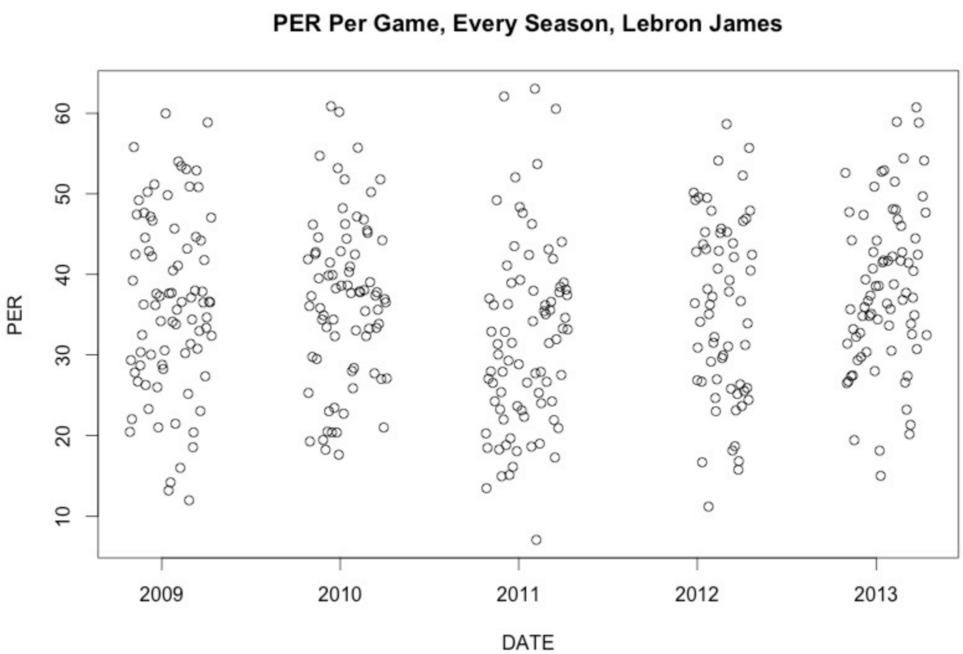
Taken from basketball-reference.com

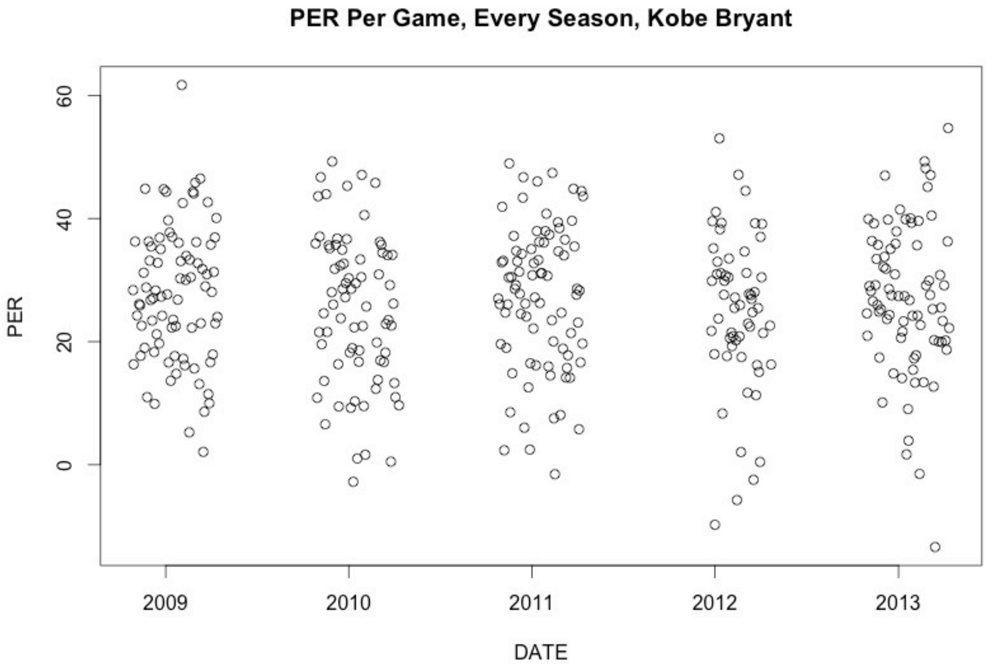
Anything related to the team in the calculations was aggregated within that game. Anything related to the league was aggregated for that season.

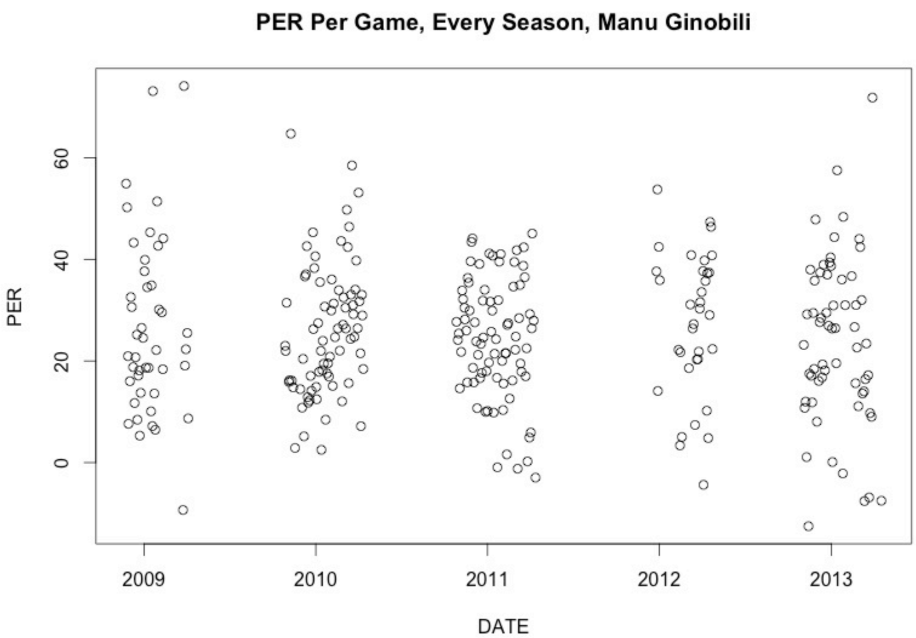
When doing the calculation for a player like Lebron James we find every PER for the player per game during that season and then average this by the total number of games played. We end up getting around 36 for James. His PER as stated online is around 30. I noticed that the better my athletes got the more variation I had between the PER that I calculated and the PER that appeared online. Either way the PER that I calculated was always a little larger. Also the PER that I calculated was done per game and not done for the entire season. A lot of the individual metrics that I end up summing per player are not done by the online season average PER calculations and this is where variation could occur.

Analysis:

I first decided to look at some well known players and look how their PERs were distributed throughout each season.



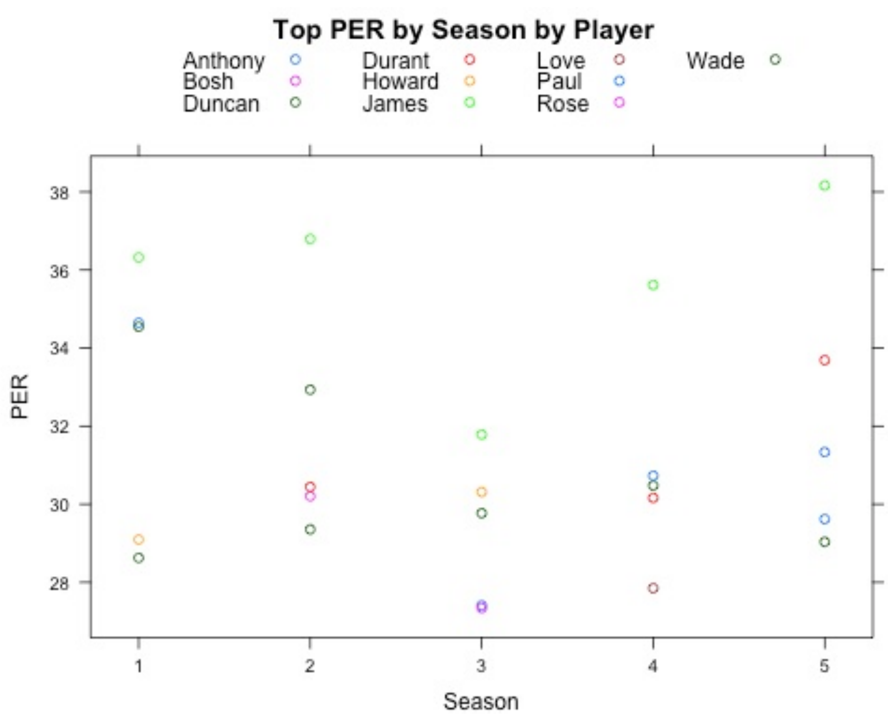


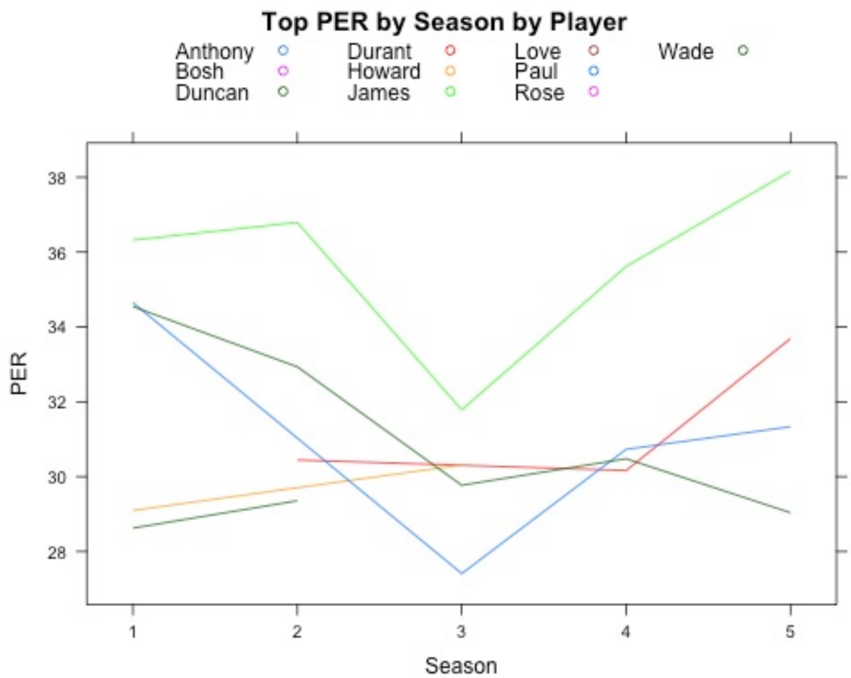


We can see that Kobe has a smaller variance when it comes to PER but Lebron has more games in which he plays better than Kobe in terms of PER. I picked Ginobili because I wanted to look at a player that presented some more variation. We can see that in the 2012-2013 season Ginobili was almost all over the place. More of a tight fit though in 2011. 2012 on the other hand was a shortened season.

I then subsetted my data for players who averaged over 20 minutes a game per season so I could further analyze PER, but more from a sample that didn't necessarily contain many outliers. If a player were to have played a couple of minutes and scored many points and grabbed a rebound, their average PER could technically be off the charts, if not accounted for how many minutes a game they are playing. Usually for seasonal PER, this would be taken care of in the function but because we did our PER based on individual games, we must manually remove these outliers.

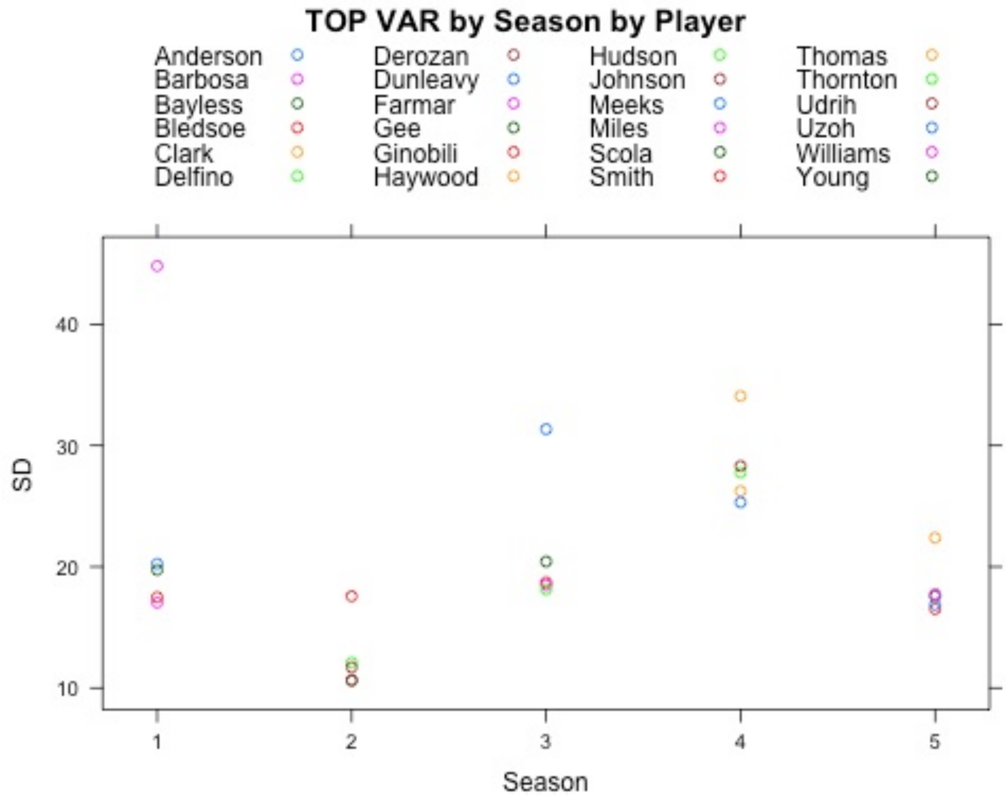
After making sure that we had players who were averaging more than twenty minutes a game I decided to look at who had the highest average PER per season and who had the most variation in PER per season. I did this for the top five in both categories in all five seasons. Below is top PER:





Interestingly enough, Wade and James appear every season. Durant appears almost every season as well. Chris Paul is there for three seasons. There is also a rather large dip in the third season but I think that this has to do with the shortened season.

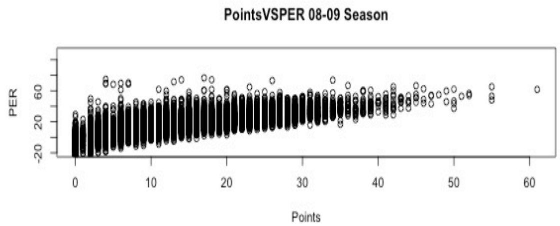
Below we have our variation, measured in standard deviations.

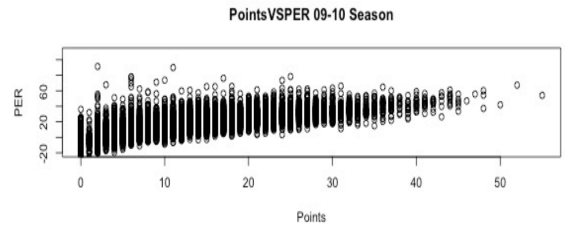


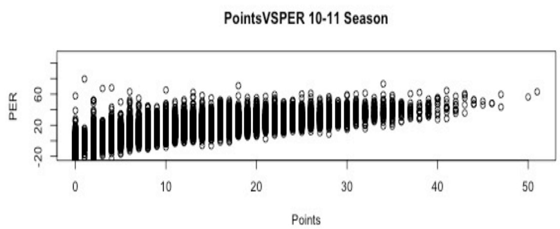
X's hear represent standard deviation. Ginobili seems to show up a few times and Dwight Howard makes our list during his first season of play.

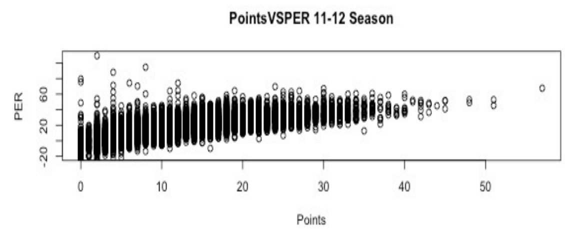
I also wanted to see how an average team per related to winning and losing a game. I made the prediction that if a team were to win, over 90% of the time their average teamPER would be higher than that of the losing team. Going to my data for the five seasons I tested this and found that 76% of the time the with the higher average PER won the game. There were probably a lot of close games that were decided by a few points. Perhaps these games led to the losing team having a higher efficiency rating. I still believe that the 76% leads us to believe that this is a solid predictor.

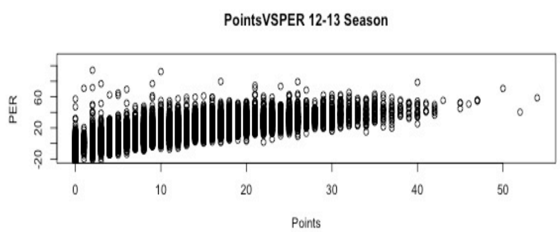
I was also very interested in examining how points affected the PER. The graph below represents number of points scored per game in each season for players averaging over twenty minutes per game with PER being represented on our Y-axis. As the number of points increase it is clear that PER can increase but at the same time there are values of low points in which PER is higher than areas with higher points. This graph only further represents the notion that there are many other factors besides points that go into the calculation of PER.



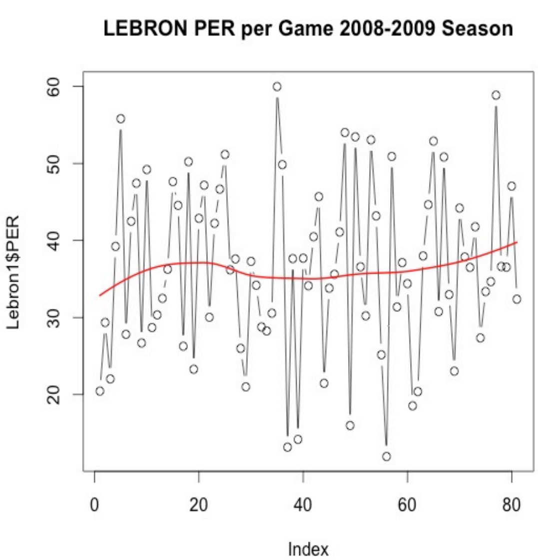


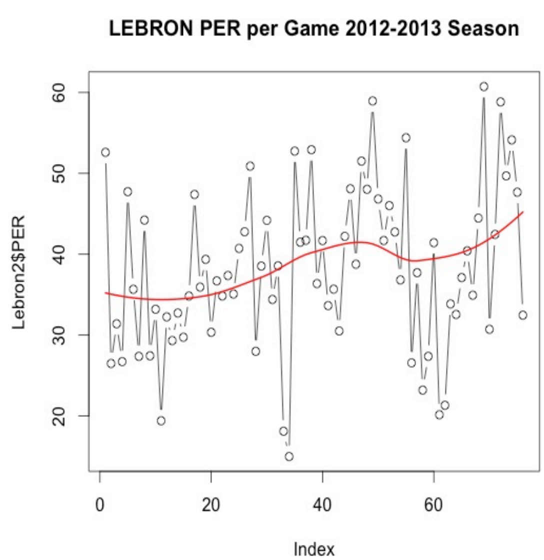


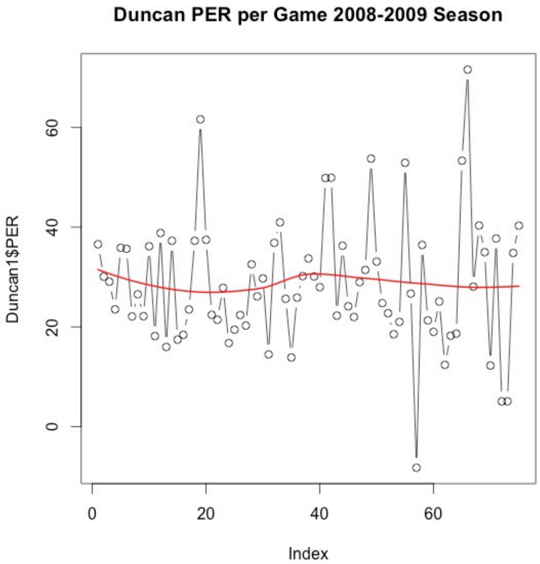


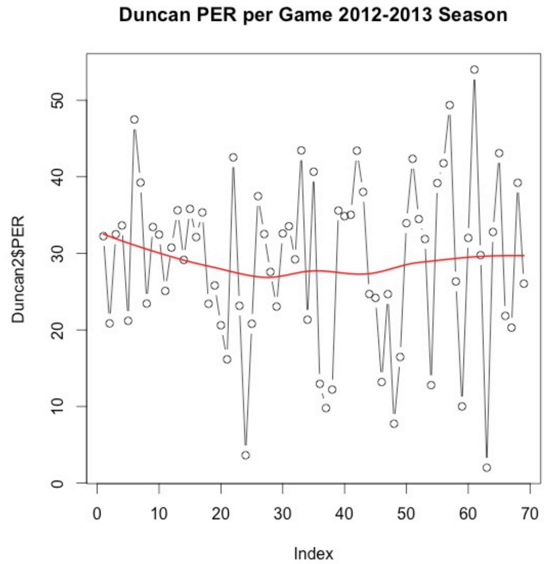


Here I have some graphs of PER of players versus time. I also smoothed the graphs so you could see curves. I first did this for two top PER players.

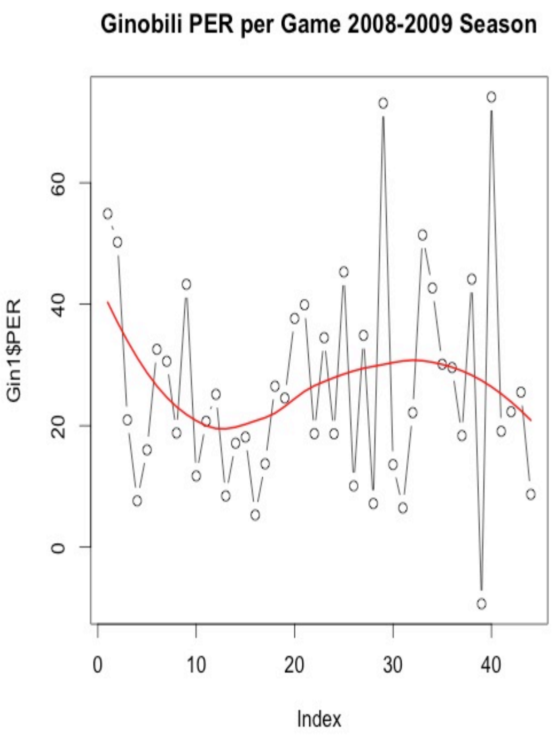


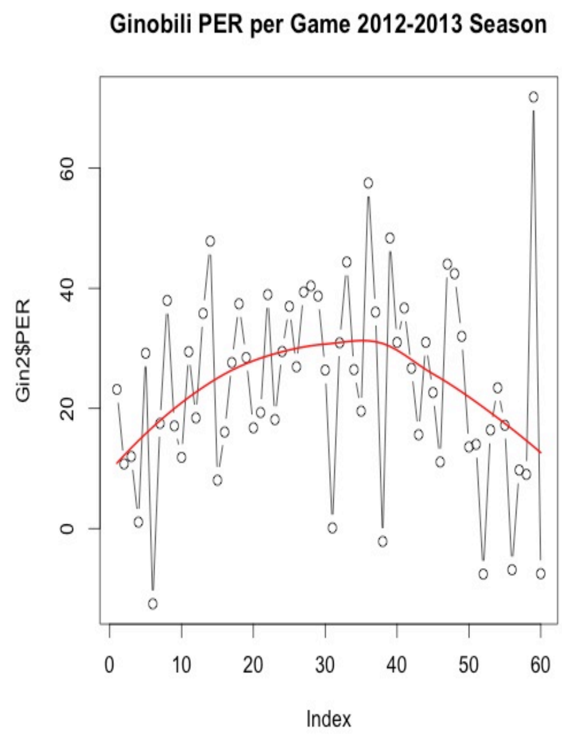






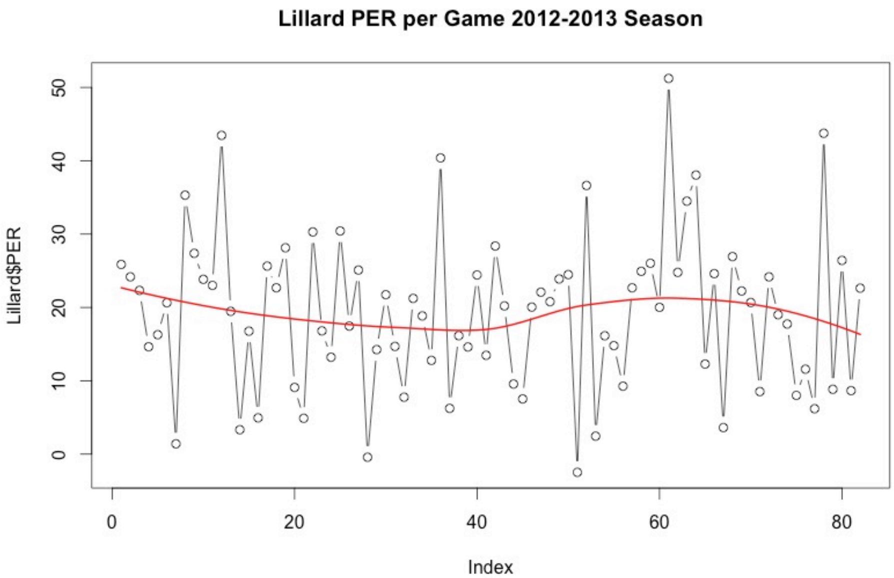
Duncan seems to slightly regress towards the end of seasons while Lebron slightly better. Duncan also shows more variance towards the end of his seasons.

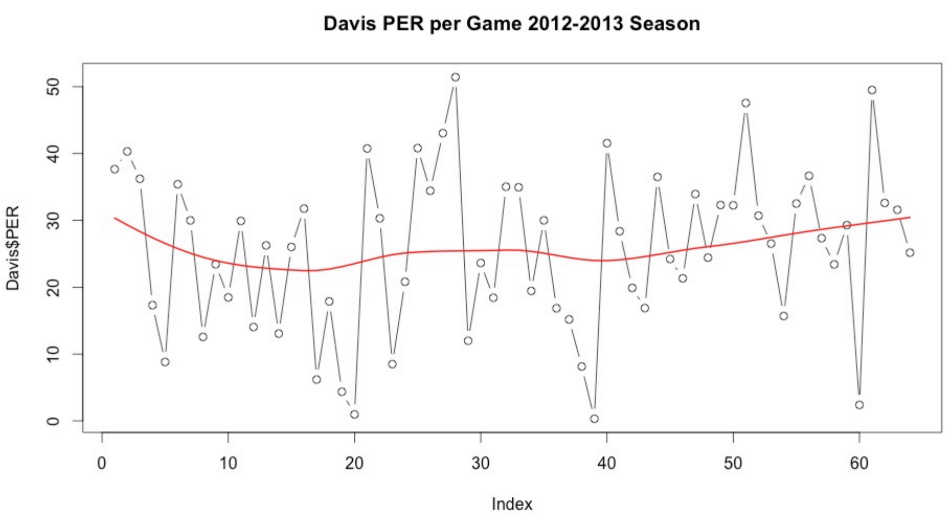




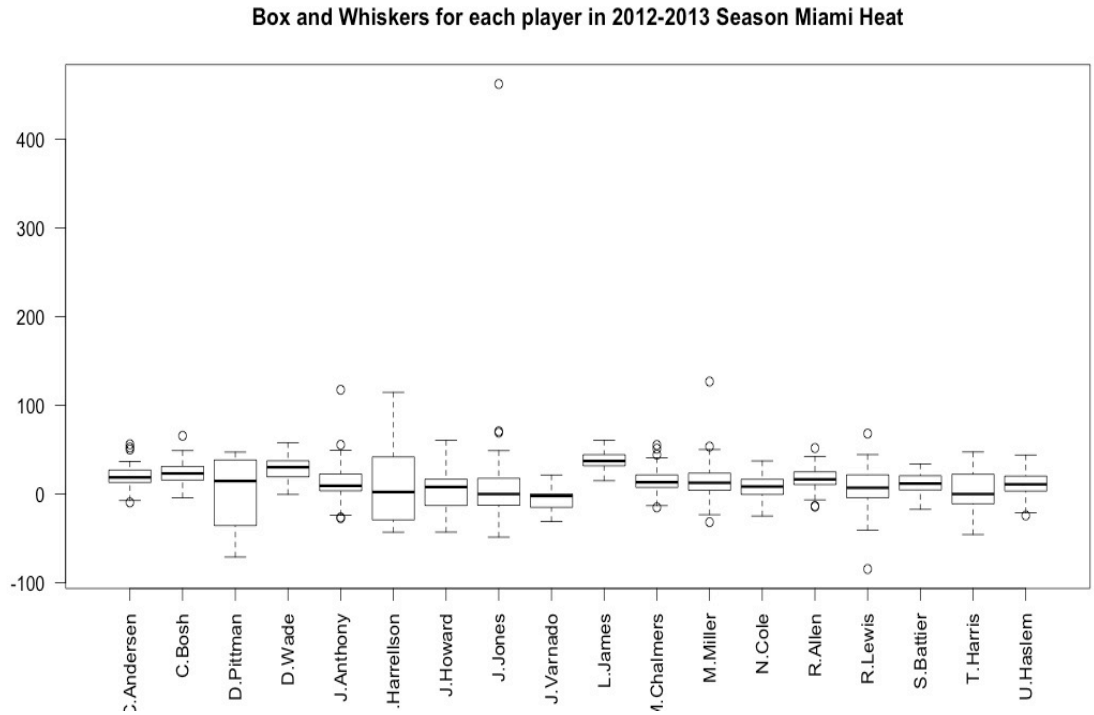
Above I showed the PER per game for Ginobli because he was one of our candidates that had the most variance. He seems to regress at the end of every year but he did really well towards the end of the 2013 season. He also has an extremely high amount of variance.

Below I plotted the 2012-2013 seasons for two famous NBA rookies. The first Damian Lillard won rookie of the year. Anthony Davis was also the number 1 draft pick during that year. They both were pretty stable. I expected Lillard to continually do better throughout the season but he was pretty consistent.



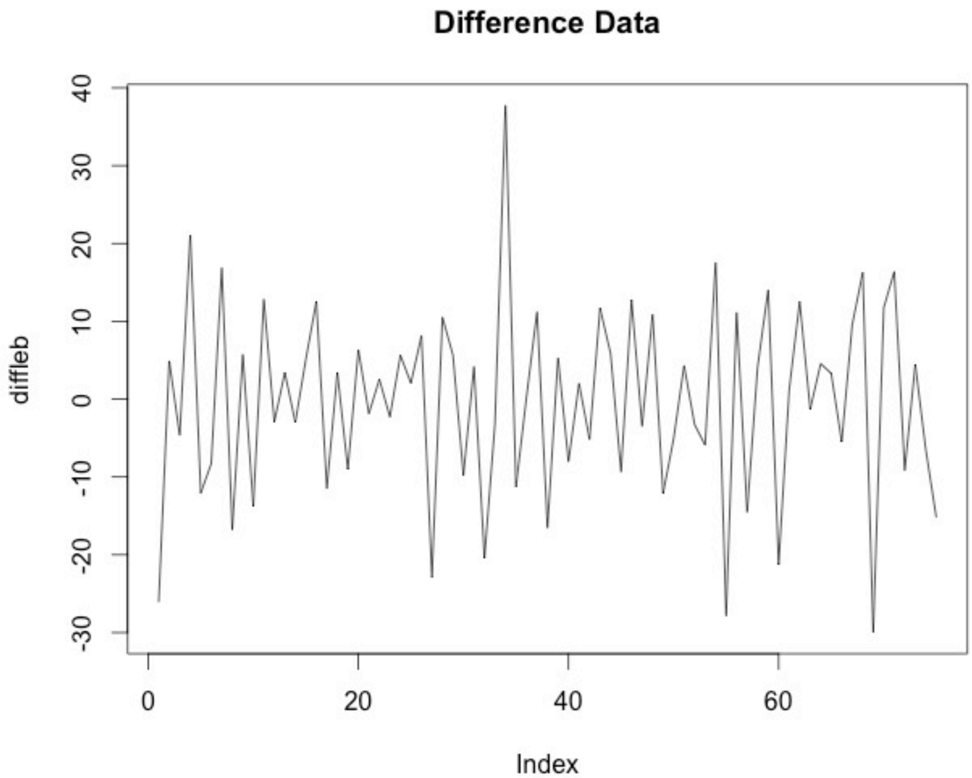


I also wanted to see how efficient each player of my favorite Miami heat team were in the 2012-2013 season so I plotted a box and whiskers plot for each player that represented their PERS for the season. Looking at Lebron's fit we see that it is extremely low in variance and higher than the rest which shows how valuable he is to the team. I would say that Bosh is the second most valuable and the Birdman, Chris Anderson seems to also make a strong impression.

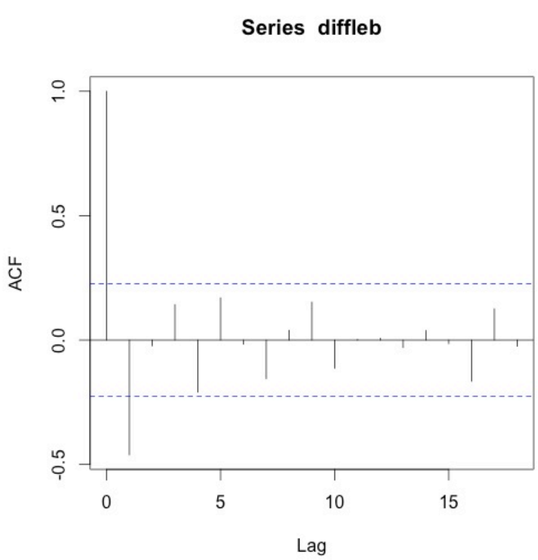


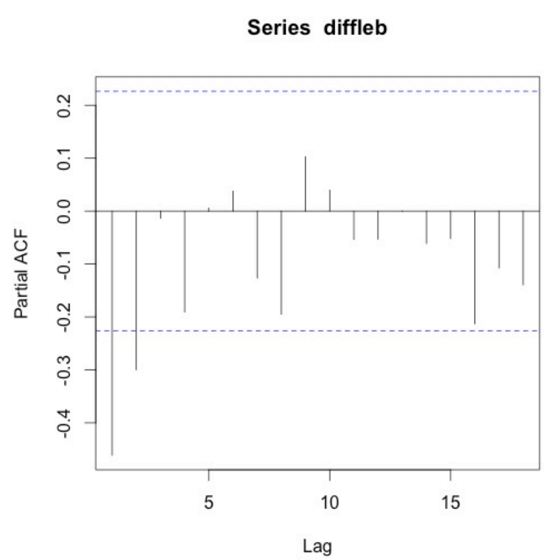
Finally I was interested in examining if there were any autocorrelations for the PER data over time for a certain player. I decided to select Lebron James for the 2012-2013 season. I was curious to see if his performance from previous games affected the performance of his next games.

In order to fit a stationary series, I decided to take the first difference of the Lebron PER data over 82 games in the 2012-2013 season. Here is the plot:

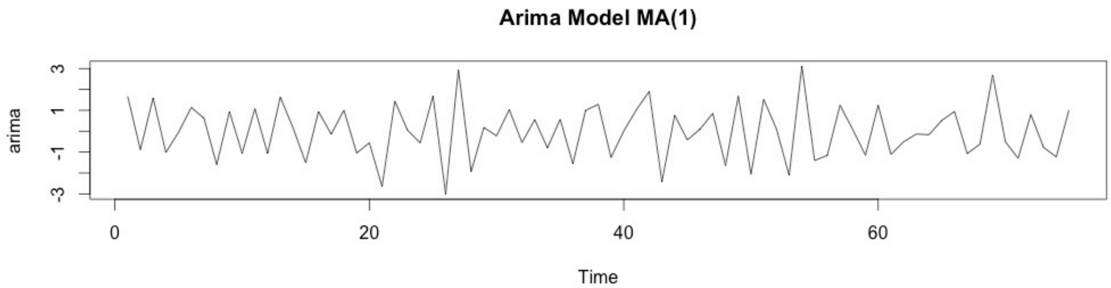


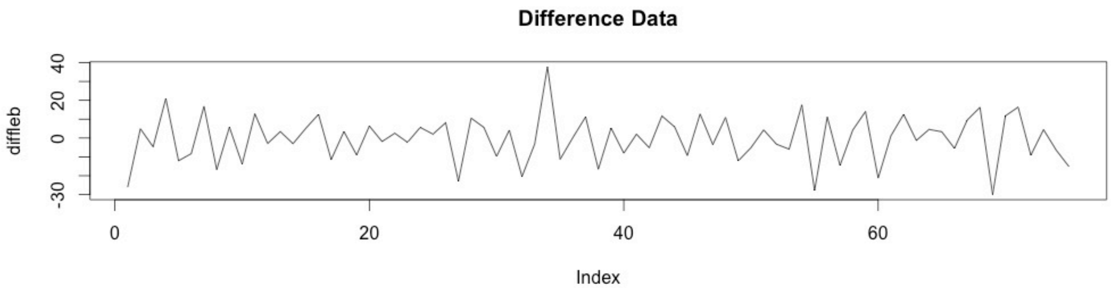
I then took the ACF and PACF plots of the function to see if there were any autocorrelations.





My preliminary guess would be to fit an ARMA(1,1) however using an R package I was able to find that the best fit model was a moving average series of 1. Thus there seems to be some relation at a lag of 1. I fit may MA(1) model below:





Conclusion:

At this point, I am still in a space where I would to continue and explore the data. After performing some of the calculations expressed above, I do believe that the PER metric is a good metric to use in order to analyze NBA data. I also do think that the metric can serve, as a good estimator in order to perhaps predict how well teams will do against one another in the future. It was also interesting to see how the top athletes in the NBA continue to show up at the top of the PER chart season after season. In a league that is already so competitive. This class of individuals seems to be in a group of their own. Call me bias but Lebron James is extremely efficient, not only in maintain low variance in PER but consistently performing at a high rate while continuing to get better. I was curious to see some rookies that perhaps would do very well or fall off. I was also hoping to see some reoccurring faces for players that had high variance. It seems that players can peak very fast, become relevant and lose that momentum quickly as well.

I spent a lot of this project just making sure that my metric would work and was feasible. The meat of my question and exploring how the PER relates to player and team performance still needs to be further explored and analyzed before coming to conclusions that can allow us to perhaps predict games.