Final Project

Aman Singh

January 21, 2021

Introduction:

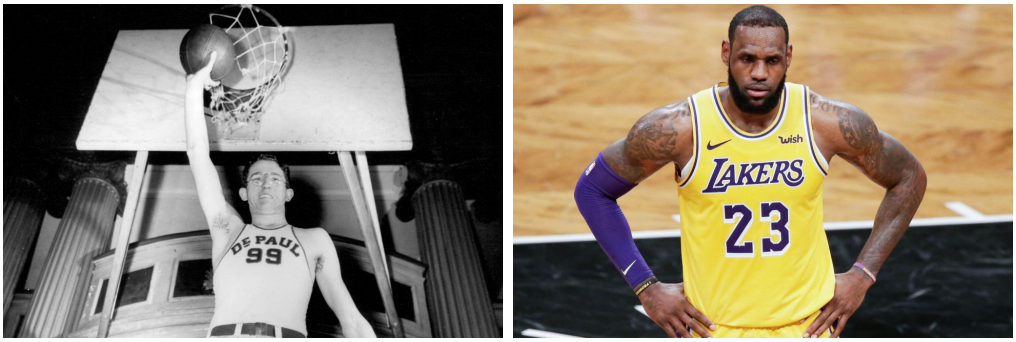

Figure 1: George Mikan pictured on the left and Lebron James pictured on the right

Sports is a huge part of America, starting from the 1800's when Baseball was invented in 1834, and then later other sports such as American Football and then Basketball [1]. Basketball in particular however, really became big in the 1950's when legends such as George Mikan and Bob Shayes put the NBA on notice. Fast forward to the present, and the NBA players right now such as Lebron James and Stephen Curry have been making the league exciting and fun to watch generating massive amounts of revenue and viewership for the NBA [2]. The biggest difference that I see in the picture above is not only the color and the race, but also the fact that it seems that Lebron James has significantly more muscle mass than

George Mikans. George Mikans just looks like a skinny tall NBA player that would really suffer in the NBA today. I was curious to see why Geroge Mikans decided to play back in the NBA in the 1950's because the players back then did not get paid the millions that they are paid currently even after adjusting for inflation. George Mikan's salary in the NBA in the 1950's was just $12,000 and after adjusting in the present days' inflation using the CPI index, only $169,433.41 [3]. That is still low for being one of the best players in the NBA today with Lebron James commanding an outstanding salary of $37,436,858 [4]. This made me start thinking whether the NBA players now play for the love of the game or if they are chasing a paycheck. I decided to compare 2 NBA players one from the past and one from the future, George Mikan and Lebron James. In 1950-1951 season, George Mikan averaged 28.4 points a game while Lebron James, in the 2017-2018 season where he played all 82 games, averaged 27.5 points per game [5, 6]. This is surprising and shows that maybe there are other factors that make an NBA player great such as how many shots they have taken, how many shots actually make, and as such there are many other factors then what the points suggest. There are also a lot of historical changes as well such as you can hold the ball as long as you wanted to in the 1950's because the NBA didn't have the 24 second shot clock violation they do have today which can limit how many points you can shoot [7]. I realized that another factor that could seemingly affect the 2 players could be BMI. However, BMI is largely inflated and does not take into account muscle mass. How BMI is calculated in fact takes into only weight and height with the formula given below:

$$BMI = Weight/Height^2$$

where BMI stands for Body Mass Index, weight is in Kilograms and the height is in Meters. I obtained the dataset that I will be observing from Kaggle which listed all the players in the NBA season of 2014-2015 [8].

Looking at this table that I made, is really just shows that BMI does not really mean much in

Name	Points Per Game	Height	Weight	BMI
George Mikan	28.4	2.08	111.13	25.613
Lebron James	27.5	2.06	113.398	27.4

Table 1: Past NBA stars vs the Present:

the NBA because in terms of BMI, both of these players, who were the best in there respective decades, are considered "overweight". It made me wonder if there are other attributes like I mentioned above if there are other factors involved such as experience, salary, and perhaps even college. This made me come to the colusion that I will use this dataset to analyze:

Is performance in the NBA related to more than the physical attributes of an athlete?

I have noticed that in general a plethora of people have wondered how BMI can affect a players performance on the court. In 2005, there was a study done involving Shaq, one of the heaviest players during that time and it showed that despite being one of the heaviest players in the NBA, his body fat percentage was only 13% which is pretty fit for an athlete overall [9]. However, the reason why I decided to not state BMI in my question is because I wanted to include the fact that maybe there are other factors that could be related to performance in the NBA. It has long been stated and there have been players who put in 100% everyday and would get a huge paycheck. However, once they got that paycheck, they wouldn't play as hard anymore and thus could potentially hurt a teams' chance of making the playoffs because they can't sign more NBA players [10]. I will be addressing salary and whether it can hurt a team's chances of making the playoffs. In fact, I believe this question is extremely important in understanding if there are many factors in how a player performs on the court, whether the motivation is money or if they are "fit" in terms of BMI, or if it concerns their experience in the NBA. There are many factors which is why I chose this question to truly find out if the performance in the NBA is related to the physical attributes or more than just physical. I believe this question is very important to people who have a passion of the NBA and not only that but also the General Managers in the NBA who might

find this analysis useful to see what specific attributes they should look for when they go and get a player, whether its considering salary, experience, or even BMI.

Understanding the Dataset:

In this section, I will make an attempt to understand the dataset and understand the representation of each column. When I downloaded this dataset, I noticed that there were 490 rows, with 34 columns depicting a single player's stats as well as other personal information. However, there were a lot of NA's in the dataset. I believe that the website that this data was scraped from had a lot of missing values for a lot of players and thus leading to a lot of NA's. In order to use CV-Lasso, and Bootstrap, we are supposed to not have any NA's in the dataset which I also address later. I also noticed that the author of this dataset wrote collage instead of college, so I decided to change the name to help with later analysis. I believe this dataset is very useful becauase with it not only shows how points they have shot, but also how much shots they have attempted as well which can show a percentage of there efficiency.I also made a table below showing the interpretation of each abbreviated column because a lot of people that don't understand the NBA will not understand the columns that are in this dataset. I also have to understand the fact that there is a huge limitation in the dataset. It only shows relevant statistics in the 2014-2015 year and thus if I wanted to know if BMI and Salary really have a correlation, I would not be able to really analyze that. To be further specific, if I wanted to analyze a player who in the 2014-2015 season had a BMI less than 25 and earned only 1 million per year in salary and than thus compare it to when he gets a huge salary bonus in the offseason, ie 5 million, to see if that affects his game or not, I would not be able to analyze that. It could be that he gets more lazy during the off-season reaping the rewards of his hard work and thus see a shortage both in his performance on the field and his physical body as well. I would not be able to know if the salary or BMI truly affects NBA players performance on the court. However, I can try

my best to use the data that I have and try to answer the question.

Table 2: Interpretation of the columns in the dataset:

Table 3: Part A:

Name of columns	Interpretation
Name	Name
Games.Played	Total Games Played
MIN	Minutes Played
PTS	Points Made
FGM	Field Goals Made
FGA	Field Goals Att.
FG.	Field Goal Perc.
X3PM	3 Points Made
X3PA	3 points Att.
X3P.	3 Point Perc.
FTM	Free Throws Made
FTA	Free Throws Att.
FT.	Free Throw Perc.
OREB	Offensive Rebounds
DREB	Defensive Rebounds
REB	Rebounds Made
AST	Assists Made

Table 4: Part B:

Name of columns	Interpretation
STL	Total Steals
BLK	Total Blocks
TOV	Total Turnovers
PF	Total Personal Fouls
EFF	Efficiency on the court
AST.TOV	Assists/Turnover Perc.
STL.TOV	Steals/Turnover Perc.
Age	Age
BirthPlace	Birth Place
Birthdate	Birth Date
Collage	College
Experience	Experience in the NBA
Height	Height
Pos	Position on Team
Team	Team
Weight	Weight
BMI	Body Mass Index

Cleaning the Data:

When I first imported the dataset, I decided to replace the NA's with meaningful insights. I scoured the web and eventually settled on a website to extract all my insights from [11]. The first NA's that I found were in the Age column and so I used the website to find their respective age during the 2014-2015 season. The next column that I filled in the NA's for was the weight category, which I also filled in using pounds but I realized BMI uses kilograms, which is why I decided to divide by 2.205 to convert into Kilograms. The column I filled in after that was their respective heights which I filled in remembering BMI, in centimeters for easier calculation. I also then calculated the BMI's for the NBA players that did not have their height & weight originally filled in. I then also filled in the position that each player respectively played in the NBA as well. I believe that can lead to some very personal and

interesting insights of the data regarding position and points, or position and comparing it to weight respectively even. Upon looking at the Experience column, I realized that the class of the column was a factor which you cannot thus use to graph meaningful insights. Also, a problem in that column was that the rookies in the year of 2014-2015 were listed as "R". I thus changed the class of the column, replaced the "R" with 0 to indicate they have no experience in the NBA, and filled in the NA values as well. I then decided to add 2 more columns, being Salary and Playoffs.made. I decided to make the column Salary because I wanted to see if salary could be a big factor in the NBA in terms of their BMI, or their performance on the court. The salaries that I indexed in are adjusted according to the year 2020's inflation rate [12]. I also decided to make a column called playoffs.made as well to see if depending on multiple factors, if the player was able to play in the playoffs, and I personally believe that it can lead to some meaningful insights. The Playoffs.made column is simply a column indicating that if your team made the playoffs, it would say Yes, and if not, No. The teams that made it to the playoffs in the 2014-2015 season were the **Atlanta Hawks**, the **Milwaukee Bucks** , the **Golden State Warriors**, the **Cleveland Cavaliers** , the **Boston Celtics**, the **Los Angeles Clippers**, the **Toronto Raptors** , and the **Brooklyn Nets** .The Golden State Warriors won that year, playing against the Cleveland Cavaliers in the Finals.

Figure 2: Warriors win first championship after 40 years

Analyzing the Data:

After cleaning the dataset and making sure that there were no NA's, I decided to explore the dataset. Instead of using the traditional graphs in the Base R package, I decided to use Tidyverse and incorporate ggplot2 to create more aesthetically pleasing graphs.

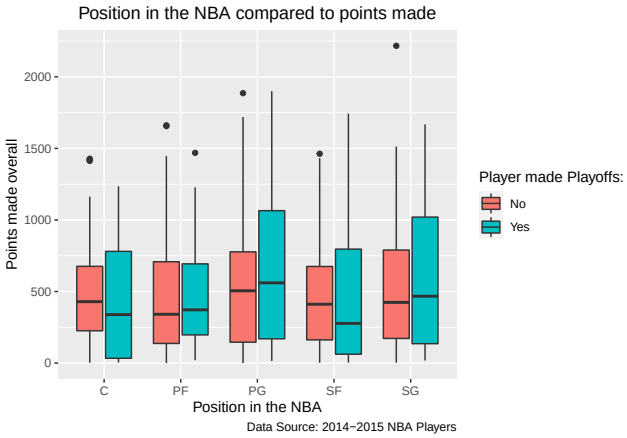

Figure 3: Plot 1

Based on this box-and-whisker plot, this shows that a majority of the players that made the most points overall in the 2014-2015 season were Point Guards regardless if they made the playoffs or not. This makes sense because a Point Guards role in the offense is to run the offense by controlling the ball and is also considered the leader of the team. Thus, it makes sense that the Point Guard on average would shoot more often and thus make more shots on average compared to other positions in the NBA.

A full season in the NBA is 82 games. Noticeably in plot 2, it shows that the majority of the players that played in the 2014-2015 season that played a majority of the games, there BMI, was respectively in between 24-28 which indicates a healthy relation in weight and height. The outlier however, the player with the BMI of 32.72 upon looking at the plot, only seemed to play around 10 games. Using tidyverse, I found that it was Simon Bhullar, who only played 3 games and had a relatively poor performance, and has averaged only 3 minutes and made only 2 points. There could be a possibility that he didn't end up playing after because of an injury or maybe he simply wasn't good enough. Upon searching up his name on the internet, I found an article saying that he had been released by his team, the

Sacramento Kings,that year with only a short stint with the team because he just was not up to par in the NBA. I also noticed that they could have signed him as a marketing technique concluding that he was the first of an Indian descent to play in the NBA. It also turns out that he had weight problems and was just seen as too big and not fit for the NBA [13].It seems like the NBA does care about weight and other aspects as well because after that no team in the NBA took him despite only playing 3 games and not yet proving himself. Despite his young age,no team was willing to take a chance on him, which shows to think that teams in the NBA do care about the weight and try to calculate how much the impact can have on the court. I than wanted to see how many people were motivated by the money and thus let their body gain in weight and thus raise their BMI. I put a salary greater than 5 million in the NBA because I personally believe anything greater than 5 million in the NBA is a lot. I also filtered through their BMI saying that if their BMI is greater than 25, they will be shown in the chart because according to BMI, anything greater than 25 is considered "overweight".

```
## `geom_smooth()` using formula 'y ~ x'
```

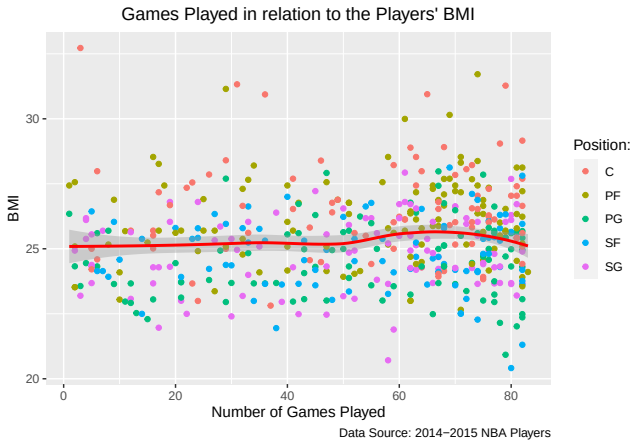

Figure 4: Plot 2

Surprisingly, it was 103/157 people whose BMI was greater than 25 and salary greater than 5,000,000 compared to the salary greater than 5,000,000 thus meaning that 65.6% of NBA

Figure 5: Simon Bhullar pictured above

players in this range have a BMI that is considered "overweight". I then checked their performance on the court to see if money maybe had a factor along with BMI.

It seems like 15.9% (25/157) NBA players' commanded huge salaries and yet only scored less than a 1000 points overall in the season thus averaging approximately only 4.86 points per game. I also calculated the average salary per nba player in this filter, and it thus shows that the average salary for NBA players in this case is 11671929.56 per player. That is extremely high and shows that either a majority of these players are injured, or are commanding a huge salary and are not trying as hard to earn that salary because they are averaging very few points per game with their BMI being considered in the "overweight" & "obese" category. Although it seems like approximately 16% of NBA players are not trying hard enough it begs to differ that maybe if they tried harder in expectations of what the team had when they offered that salary, they quite possibly have led their team to the playoffs.

Upon analyzing the playoffs column, I realized that out of the players that commanded huge salaries and had a BMI over the "fit" category, noticeably only 24% (6/25) players made it to the playoffs. However, despite the fact that I did notice that BMI does have an effect, I noticed that out of all the teams that made it to the playoffs, the Cleveland Cavaliers who had the heaviest team in terms of BMI in the playoffs made it to the Finals. I than realize that just analyzing the data is not enough because it can be misleading and will thus turn to machine learning techniques to figure out if there are factors that affect a performance of

an NBA player.

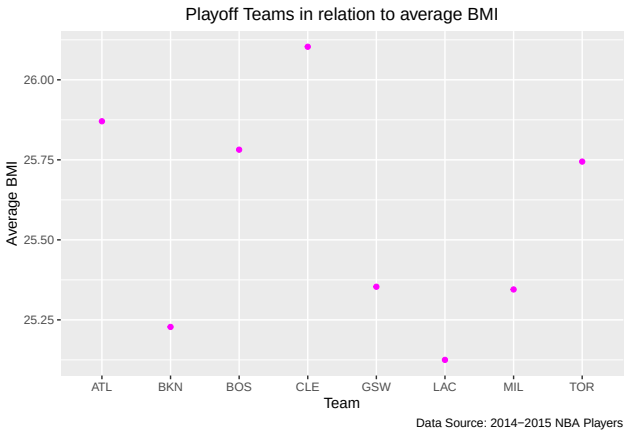

Figure 6: Plot 3

Machine Learning Analysis:

I will start by doing an Cross Validation Lasso to see what variables would be nice with the regressor being BMI. We will do this by estimating the vector of parameters β from the Lasso model:

$$\hat{\beta} = argmin \sum_{i=1}^{n} (yi - x')^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

I am using Cross-Validation Lasso to pick the true λ that can lead to the true prediction error and also picks the best model by penalizing more complicated models. I am also using Cross-Validation Lasso over other best model fits solely because it is one of the best ways to minimize the OOS errors.

This graph shows that there is not that much variability in terms of BMI and thus there is not much we can do to predict it. This graph is basically I suspect that the best models to predict BMI are the complex ones.

I then predicted the in sample deviance and found that for the person I predicted it on was pretty accurate, with it being pretty close with the real BMI. The real BMI of AJ Price was 23.798393 and the predicted BMI was 24.0140392, which is fairly accurate. I will also be

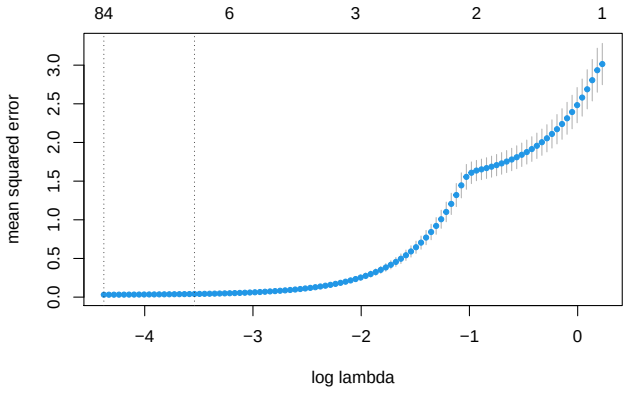

Figure 7: Plot 4

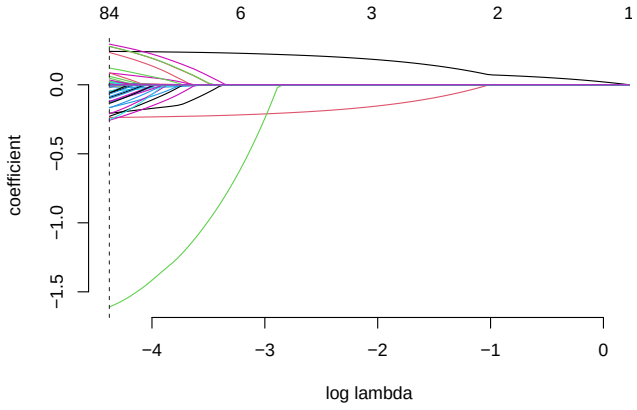

Figure 8: Plot 5

doing now a Nonparametric Bootstrap to essentially see a fairly accurate standard error for the BMI column of my dataset.

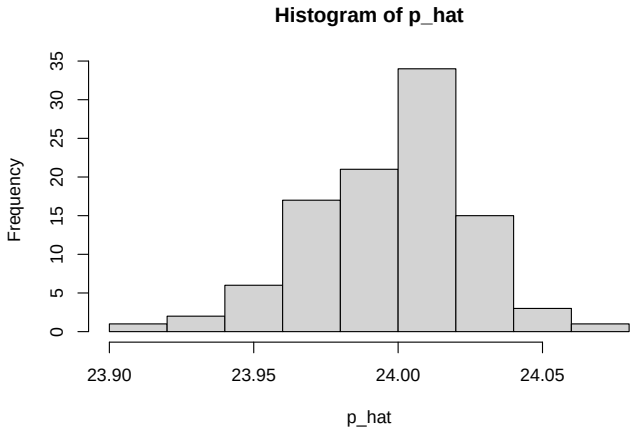

Figure 9: Plot 6

Based off the bootstrap and thus after making an confidence interval, I found that a majority of the players BMI would be between the confidence interval of 25.3066436, 25.4126993. Based off of this histogram, I see that there is a lot of variability and that it is skewed to the left. This has come to my conclusion that we can predict BMI based off on a multitude of factors. However, I have yet to try an out-of sample prediction for BMI. My next step in this project would be to than try to estimate an out-of sample prediction for BMI based off of a lot of variables. I do believe that it is possible to calculate in sample prediction for BMI based of on the performances of an NBA player on the field including his salary as well.However, to answer my questions about whether performance in the NBA is related to more than the physical attributes of an athlete I believe there is more to it because my predictions, while accurate are still off and thus I do believe that there are more factors that we need to even more accurately predict BMI of an NBA player.

References

[1] History Staff: Who Invented Baseball?,

https://www.history.com/news/who-invented-baseball

[2] Richard,

https://www.si.com/media/2016/06/20/nba-finals-game-7-tv-ratings-viewers-cavaliers-wa

[3] Inflation Calculator,

https://www.davemanuel.com/inflation-calculator.php

[4] LeBron James NBA Salary,

https://hoopshype.com/player/lebron-james/salary/

[5] George Mikan Stats,

https://www.basketball-reference.com/players/m/mikange01.html

[6] Lebron James Stats,

https://www.basketball-reference.com/players/j/jamesle01.html

[7] Keely Flanagan: Basketball's Shot Clock: A Brief History,

https://www.wbur.org/onlyagame/2015/04/22/nba-shot-clock-history-basketball

[8] Omri Goldstein: NBA Players Stats - 2014-2015,

https://www.kaggle.com/drgilermo/nba-players-stats-20142015

[9] Associated Press: Athlete Study Exposes Flaw of BMI Obesity Measure,

https://www.foxnews.com/story/athlete-study-exposes-flaw-of-bmi-obesity-measure

[10] Brian Hutchinson: NBA Players Don't Play to Win, They Play for the Money,

https://bleacherreport.com/articles/82851-nba-players-dont-play-to-win-they-play-for-

[11] Basketball Stats and History Statistics, scores, and history for the NBA,

https://www.basketball-reference.com

[12] NBA Players Salaries during the 2014-2015 season,

https://hoopshype.com/salaries/2014-2015/

[13] Ailene Voisin: Opinion: With weight down, Sim Bhullar's game is on the rise,

https://www.sacbee.com/sports/spt-columns-blogs/ailene-voisin/article16700834.html