Using PCA and Other Methods to Rank NBA Players

Jackson Bronkema

MAT392: Mathematics Seminar

Dr. Jeremy Case

Fall Semester, 2020

Table of Contents

Abstract

The use of analytics in sports has been changing the ways that players play, coaches coach, and scouts rank and recruit players. Throughout this paper, we will be explaining how we have used a very powerful application of linear algebra – Principal Component Analysis – to develop our own method to rank the careers of the top 50 NBA basketball players of all time. Using Principal Component Analysis (PCA), we were able to take a deeper look into the data we collected. The use of PCA allowed us to observe the covariance between the data we collected and better understand how specific metrics we used to rank players are correlated or uncorrelated to each other. This correlation between specific career accomplishments and statistics helps to determine a player's overall career ranking and explain why some players are ranked higher than others. We will also use PCA in a case analysis, where we will take on the role of an NBA general manager and choose among NBA draft prospect and free agents.

Introduction

There are many different applications of linear algebra to data analysis. In recent years, there has been a surge in the use of applying linear algebra techniques to big data, because it is very applicable to machine learning, dimensionality reduction, and natural language processing (Mahendru, 2019). Throughout this paper, we will expand upon two powerful applications in linear algebra: Principal Component Analysis and the Massey Method. We will use these applications to explore the ways that linear algebra can be used in professional sports to help general managers better understand the true value of players that they are scouting and considering bringing onto their teams.

From a personal perspective, I am personally interested in this topic for a number of different reasons. First, I am very passionate about data science and many of its applications to business and sports analytics. I am currently in the process of applying to Master's in Business Analytics programs at a handful of schools around the United States. To go along with this, I am also very interested in the business side of sports. I have always been interested in the role that owners, presidents, and general managers of sports teams play in creating the best possible team under the salary cap restrictions that exist in professional sports. I am personally interested in conducting my research in basketball because it was my first love as a young boy, and I grew up a die-hard Detroit Pistons fan. I can still remember that feeling when they defeated the Los Angeles Lakers in the NBA Finals in 2004, and I even had an opportunity to hold the Larry O'Brien Championship Trophy. Finally, I felt passionately about using my research to find a statistical way to rank the careers of NBA legends, and to attempt to get an answer to who is the greatest NBA player of all time.

With all this being said, we will first look at two different ways that we can rank the careers of the 50 greatest NBA basketball players of all-time. Linear Algebra will be the staple behind both methods because it is an effective tool that helps us deal with and better understand very large sets of data. Then, in a case analysis, we will take on the role of General Manager of the Minnesota Timberwolves, as we attempt to rebuild the worst team in the NBA through the NBA Draft and NBA Free Agency. In this case analysis we will use PCA and a very powerful application of PCA, called the Statistical Diversity Index (SDI) to draft one college prospect and to sign one high-quality and economically efficient free agent.

Understanding Principal Component Analysis

What is PCA and How does it Work?

Principal Component Analysis is a linear algebra and data analysis technique that was created in 1901 by Karl Pearson. Initially, it was invented as an analogue of the principal axis theorem in mechanics (Semantic Scholar, 2015). It was later expanded upon, developed, and named by Harold Hotelling in the 1930s. Hotelling was an influential economic theorist and mathematical statistician who is also known for creating Hotelling's Law, Hotelling's Rule, and Hotelling's Lemma in economics (Royal Society, 2018). He developed PCA, and it has been widely used in finance, statistics, and computer science applications. PCA can be thought of as a statistical interpretation of taking the singular value decomposition of a mean-centered matrix. What is so useful about PCA is its ability to reduce the dimensionality of very large data sets, while retaining most of the variation of the key variables in the data set. This results in minimal information loss (Manage, Scrariano, 2013). Ultimately, the goal of PCA is to reduce the dimensionality of the data set, trade as little accuracy as possible, and help us better understand our data so that we can make informed decisions based on the variation that we see in our data set.

The Process of PCA

There are four main parts of performing PCA on a set of data. We will briefly explain each of these steps in this section, but we will show the details of how each of these steps can be performed on a set of data in our case analysis.

The first step of Principal Component Analysis is to standardize or normalize your data set. Both terms are used, but they mean the same thing in this case. For sake of familiarity, we will use the term standardize throughout this paper. Within the PCA framework, standardizing our data set means first, arranging our data in a matrix with our observation vectors as columns and the variables we are measuring for each observation vector as our rows. In our research, the observation vectors will be the players we are ranking against each other, and the variables will be the different statistics we will use to rank them. It is important to note that PCA can just as easily be performed with the observation vectors as rows and the variables as the columns. However, we will use the notation laid out in David Lay's: *Linear Algebra and its Applications,* which has the observation vectors as columns and the variables as rows. To standardize our data, we must convert our matrix into mean-deviation form. To do this, we must take the sample mean of each row in our matrix and subtract it from every data value in that row, using the equation:

$$X - \bar{X} = M.$$

Here, X is our data set in matrix form, $\bar{X}$ is the row-wise mean of each variable, and M is our matrix in mean deviation form. Putting our matrix into mean-deviation form helps to normalize our data, so that each of the variables are represented in a zero-mean row.

With our data in mean deviation form, we can then find the sample covariance matrix of the matrix M using the equation:

$$S = \frac{1}{N-1} M M^T.$$

S is our sample covariance matrix and N represents the number of observation vectors we have, or the number of columns in our matrix. We know that any matrix of the form MM^T is positive semidefinite, so S must also be positive semidefinite (Lay, 478). It is vital that we calculate the covariance matrix, because later in the process, this will help us calculate what percent of the variation in our data is represented by each principal component. Using the matrix created from this equation, we will be able to find the variance of each data type and the covariances of respective data types. This will help us in determining how correlated or uncorrelated specific statistics are to each other.

Next, we will find the eigenvalues and eigenvectors of our sample covariance matrix S. You can do this by hand with small data sets, but with large data sets, it is quicker and easier to use a statistical software like MATLAB, Python, or R. Having found the eigenvalues and their corresponding eigenvectors, it is easiest to put them into matrix form. In our research, we put our eigenvalues in a square diagonal matrix with our eigenvalues in descending order along the diagonal of that matrix. We labeled this matrix D. The corresponding eigenvectors are in a different matrix called V. It is important to find both matrices, because they both represent different things within the framework of PCA. Our eigenvectors represent our principal components, and they are used on our data set to determine new variables that will help us rank our observation vectors or players. Ultimately, our principal components are new variables that are constructed as linear combinations of mixtures of the initial variables (Lay, 480). We can use our principal components to see how different players vary and then rank players based on this variation. Our eigenvalues can be used to calculate the amount of variation that each principal component represents within the data set as a whole, which helps us to understand how reliable performing principal component analysis on a set of data really is. I will explain how to calculate our new variables and the variation per principal component in the *Case Analysis*.

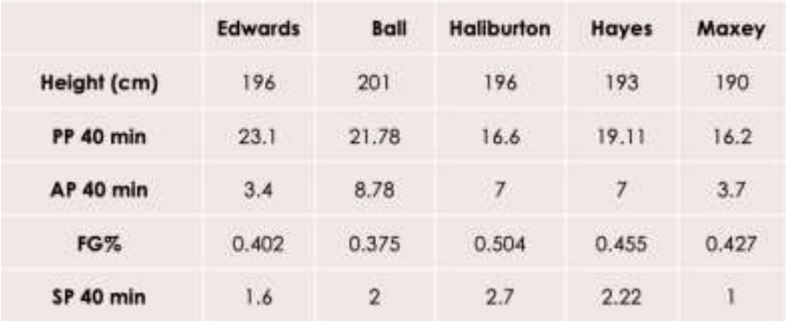
Case Analysis

In this case analysis, we will be taking on the role of the General Manager of the Minnesota Timberwolves, which is an NBA team. The general manager of a professional sports organization supervises the day-to-day operations of a franchise, and they ensure that everything runs as smooth as possible. They are also responsible for collaborating with players, coaches, and the front office to dismiss players from the ream, draft new players, or recruit other professional players through free agency. Ultimately, they oversee handling all team personnel decisions.

We chose the Minnesota Timberwolves for this case analysis, because they are one of the worst teams in the NBA, and they have the first selection the 2020 NBA draft. With the first pick in the draft, they have the power to select any available player in the draft class. In our role as general manager, we will use PCA to rank the top five guard prospects in the draft and using this analysis will we choose which player we want to add to our team for next season. We wanted to pick a guard in the draft because there is a plethora of very talented guards in this draft, and it is much easier to compare players from the same position group using PCA. We will also use PCA and an application of PCA called the Statistical Diversity Index (SDI) to decide which free agent forward we should pursue during the offseason. The Timberwolves have a very weak team, so we want to draft a guard and sign a forward to help improve the roster for next season.

PCA in the NBA Draft

As we mentioned briefly, we will use the first pick in the draft to choose a guard. There are five heavily touted guards in this draft, so we have collected key data on them and will perform PCA on that data set to compare and rank these players. By seeing how these players vary, we can then make an informed decision on which player we want to select in the draft. The players we are considering are Anthony Edwards from University of Georgia, Lamelo Ball from Chino Hills California, Tyrese Haliburton from Iowa State University, Killian Hayes a French International player, and Tyrese Maxey from the University of Kentucky. We will use five main statistics to evaluate these players against one another. These include the players height in centimeters, points per 40 minutes, assists per 40 minutes, steals per 40 minutes, and field goal percentage. We used our statistics per 40 minutes because college basketball games are 40 minutes long and international basketball games are 48 minutes long, so this helped to even the playing field among the prospects. We obtained all our statistics from basketball-reference.com, which is a very reliable source. We believe that these are the five most important statistics to measure how well these players will perform in the NBA. Ideally, we would have used a wider range of statistics to compare these players against each other, but we feel that looking at the five most important statistics will still give us a great gauge at which player we should select. Here is our data set:



	Edwards	Ball	Haliburton	Hayes	Maxey
Height (cm)	196	201	196	193	190
PP 40 min	23.1	21.78	16.6	19.11	16.2
AP 40 min	3.4	8.78	7	7	3.7
FG%	0.402	0.375	0.504	0.455	0.427
SP 40 min	1.6	2	2.7	2.22	1

Now, we put that data into a 5x5 matrix with the players as the columns and the statistics as our rows. We will call this matrix X, and it is shown below:

$$X = \begin{bmatrix} 196 & 201 & 196 & 193 & 190 \\ 23.1 & 21.77 & 16.6 & 19.11 & 16.2 \\ 3.4 & 8.78 & 7 & 7 & 3.7 \\ .402 & .375 & .504 & .455 & .427 \\ 1.6 & 2 & 2.7 & 2.22 & 1 \end{bmatrix}.$$

In order to perform PCA on this data set, we must first standardize our data by converting X to mean-deviation form. To do this we must calculate the sample mean matrix $\bar{X}$. We can obtain $\bar{X}$, by calculating the row-wise mean of X. M represents X in mean-deviation deviation form. Thus:

$$M = X - \bar{X} = \begin{bmatrix} 196 & 201 & 196 & 193 & 190 \\ 23.1 & 21.77 & 16.6 & 19.11 & 16.2 \\ 3.4 & 8.78 & 7 & 7 & 3.7 \\ .402 & .375 & .504 & .455 & .427 \\ 1.6 & 2 & 2.7 & 2.22 & 1 \end{bmatrix} -$$

$$\begin{bmatrix} 195.2 & 195.2 & 195.2 & 195.2 & 195.2 \\ 19.36 & 19.36 & 19.36 & 19.36 & 19.36 \\ 5.98 & 5.98 & 5.98 & 5.98 & 5.98 \\ .433 & .433 & .433 & .433 & .433 \\ 1.9 & 1.9 & 1.9 & 1.9 & 1.9 \end{bmatrix} = \begin{bmatrix} .8 & 5.8 & .8 & -2.2 & -5.2 \\ 3.74 & 2.42 & -2.76 & -.25 & -3.16 \\ -2.58 & 2.8 & 1.02 & 1.02 & -2.28 \\ -.03 & -.06 & .07 & .02 & .01 \\ -.30 & .10 & .80 & .32 & -.90 \end{bmatrix}.$$

Putting our matrix into mean-deviation form helps us to better understand our data. Now, each of the statistics we use are represented in a zero-mean row. Ultimately, we are subtracting the sample mean of a type of data from the actual data value. Looking at our matrix M, we can quickly and easily compare players against each other. A positive value in a player's column means that they are above average compared to the other players, and a negative value in a player's column shows us that they are below average in that specific statistic compared to the rest of the players.

The second part to PCA, is finding the sample covariance matrix of M, which is S. To find the sample covariance matrix of M, we use the equation:

$$S = \frac{1}{N-1} M M^T.$$

Here, N is equal to the number of columns in our matrix, which is also equivalent to the number of players we are studying in our set of data. Thus,

$$S = \frac{1}{4} \cdot \begin{bmatrix} .8 & 5.8 & .8 & -2.2 & -5.2 \\ 3.74 & 2.42 & -2.76 & -.25 & -3.16 \\ -2.58 & 2.8 & 1.02 & 1.02 & -2.28 \\ -.03 & -.06 & .07 & .02 & .01 \\ -.30 & .10 & .80 & .32 & -.90 \end{bmatrix} \cdot \begin{bmatrix} .8 & 3.74 & -2.58 & -.03 & -.30 \\ 5.8 & 2.42 & 2.8 & -.06 & .10 \\ .8 & -2.76 & 1.02 & .07 & .80 \\ -2.2 & -.25 & 1.02 & .02 & .32 \\ -5.2 & -3.16 & -2.28 & -.01 & -.90 \end{bmatrix} = \begin{bmatrix} 16.70 & 7.95 & 6.15 & -.08 & 1.24 \\ 7.95 & 9.37 & .31 & -.11 & -.08 \\ 6.15 & .31 & 5.44 & .01 & 1.06 \\ -.08 & -.11 & .01 & .002 & .02 \\ 1.24 & -.08 & .106 & .02 & .41 \end{bmatrix}.$$

Using the matrix created from this equation, we will be able to find the total variance of dtat set and the covariances of respective data types. This will help us in determining how correlated or uncorrelated our variables are.

From here, we can use MATLAB to find the eigenvalues and eigenvectors of our sample covariance matrix. D represents a diagonal matrix with the eigenvalues of S in descending order. Putting this matrix in descending order is helpful, because one can more quickly identify which principal component accounts for the most variation in the data set. The eigenvectors of our sample covariance matrix will be in matrix V, and they will be arranged so that they are in the same column as their corresponding eigenvalues. V is also a square matrix. In our case,

$$D = \begin{bmatrix} 23.5 & 0 & 0 & 0 & 0 \\ 0 & 7.12 & 0 & 0 & 0 \\ 0 & 0 & 1.12 & 0 & 0 \\ 0 & 0 & 0 & .19 & 0 \\ 0 & 0 & 0 & 0 & .0001 \end{bmatrix} \text{ and } V = \begin{bmatrix} .829 & .186 & .526 & -.036 & -.003 \\ .473 & -.737 & -.481 & .047 & -.008 \\ .294 & .634 & -.697 & -.159 & -.010 \\ -.005 & .010 & .004 & .077 & -.997 \\ .056 & .144 & -.071 & .982 & .077 \end{bmatrix}.$$

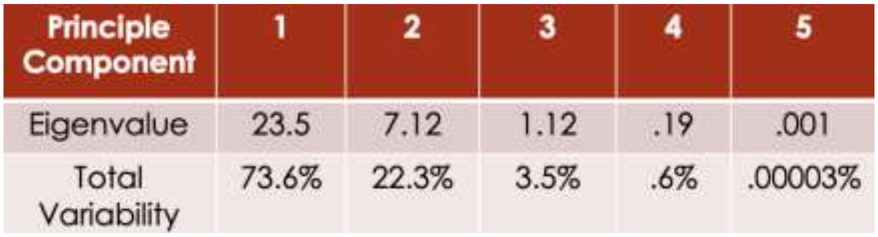
Using this, we can find the total variance of our sample covariance matrix and the percentages of the variance of each of our data types. This will help us to better understand how our model is represented by each of the statistics that we are using in our model. This way, we can better see how our model is generated/created, and we can eliminate any metrics that do not contribute very much to the model. The eigenvectors are the principal components and each of the values in the eigenvectors are called the “loadings” or weights of the principal components. The eigenvalues on the other hand, represent the percentage of variance that each principal component accounts for, and we can easily compare these percentages against one another. To calculate the total variance of the system, we use the equation:

$$Total\,Varaince = \lambda_1 + \lambda_2 + \ldots + \lambda_n = tr(D),$$

Where $tr(D)$ is equivalent to the trace of the matrix D. The trace of a matrix is the sum of its diagonal, for a diagonal matrix. To find the percent of the variance represented by a specific principal component, we use the equation:

$$\%\,Var\;for\,PC_i \;=\; \frac{\lambda_i}{\lambda_1+\lambda_2+\ldots+\lambda_n}.$$

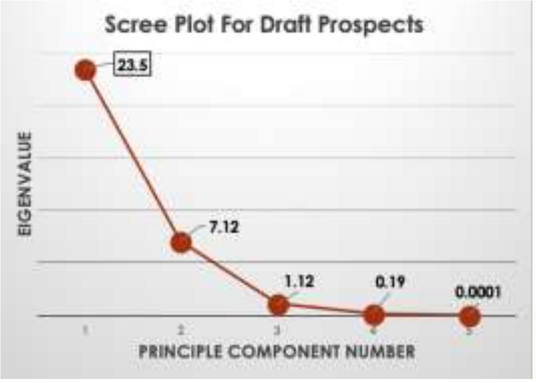
For our data set, here is a table that represents the amount of variability represented by each principal component with its corresponding eigenvalue.



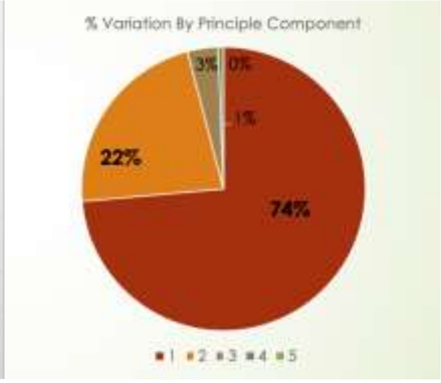
Principle Component	1	2	3	4	5
Eigenvalue	23.5	7.12	1.12	.19	.001
Total Variability	73.6%	22.3%	3.5%	.6%	.00003%

Figure 1

Also, below is a scree plot and pie chart that help us to understand and visualize the differences between each principal component.



Scree Plot 1



Pie Chart 1

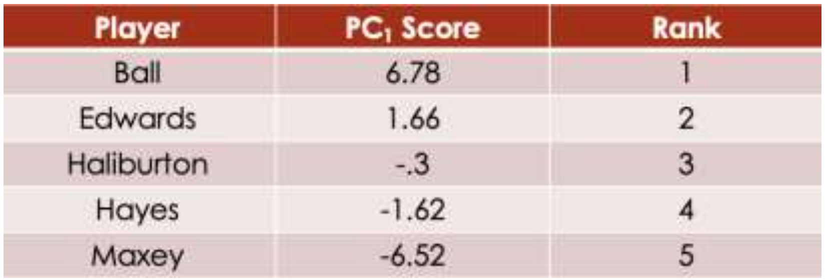
Once we have our principal components, we can then use the largest PC to find variation between our observation vectors or rank them based on their “player rating”. To do this we use the equation:

$$Player\ Rating = M^T \cdot PC_1.$$

For our data,

$$Player\ Rating = \begin{bmatrix} .8 & 3.74 & -2.58 & -.03 & -.3 \\ 5.8 & 2.42 & 2.80 & -.06 & .1 \\ .8 & -2.76 & 1.02 & .07 & .8 \\ -2.2 & -.25 & 1.02 & .02 & .32 \\ -5.2 & -3.16 & -2.28 & -.01 & -.9 \end{bmatrix} \cdot \begin{bmatrix} 8.29 \\ .473 \\ .294 \\ -.005 \\ .056 \end{bmatrix} = \begin{bmatrix} 1.66 \\ 6.78 \\ -.3 \\ -1.62 \\ -6.52 \end{bmatrix}.$$

When we find our player ratings using the equation: $M^T \cdot PC_1$, we are ultimately ranking each player separately using the weights from our first principal component and multiplying that by their statistics in mean deviation form. Using matrix multiplication helps us to calculate all of these scores simultaneously. These ratings can be positive or negative, but the sign is arbitrary, as PCA is applicable in only measuring the variability between players. From this, we can see that Lamelo Ball has the highest rating, since he is the player in row two in the matrix M^T.



Player	PC$_1$ Score	Rank
Ball	6.78	1
Edwards	1.66	2
Haliburton	-.3	3
Hayes	-1.62	4
Maxey	-6.52	5

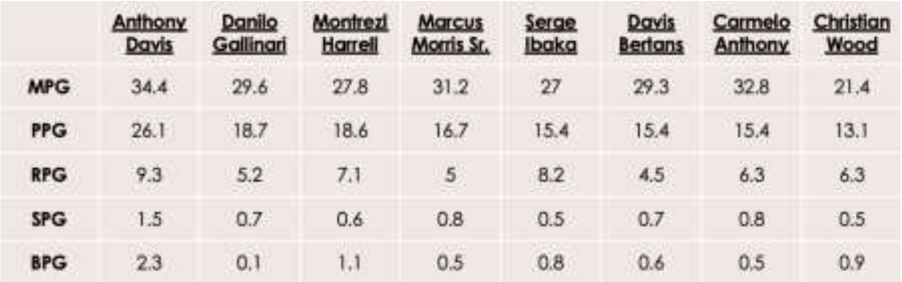
Player Rankings 1

In ranking these five players using PCA, we can now make an informed decision on which player we should select in the draft. Based on these scores from the first principal component, the Minnesota Timberwolves should select Lamelo Ball with the first pick in the draft. It is important to note that this should only make up part of the decision-making process, as the team should interview each player, evaluate them using other statistics or against other players, and think about the play style of each player and think about how that fits with their current personnel.

PCA and SDI in Free Agency

In the second part of this case analysis, we will use PCA and a very powerful application of PCA to decide which free agents we should pursue and try to get to sign with our team. As stated before, we are going to be pursuing one forward free agent to add to our current roster. This decision was made mostly due to team need, and because there are many solid forwards in free agency this season. Also, we already drafted a guard in the NBA draft. This being said, we will be comparing the eight best free agents at that position, and we will use five statistics in our analysis. The players we will be considering in free agency are Anthony Davis, Danilo Gallinari, Montrezl Harrel, Marcus Morris Sr., Serge Ibaka, Davis Bertans, Carmelo Anthony, and Christian Wood. The statistics we will use to evaluate these players will be points per game, assists per game, rebounds per game, steals per game, and blocks per game (Spotrac, 2020). These are all vital statistics in measuring the productivity of forwards in the game of basketball. It is worth mentioning the fact that we could have used more statistics, but in our research, we wanted to focus on these five key indicators of performance for the forward position.

We accumulated our data into an excel sheet here:



	Anthony Davis	Danilo Gallinari	Montrezl Harrell	Marcus Morris Sr.	Serge Ibaka	Davis Bertans	Carmelo Anthony	Christian Wood
MPG	34.4	29.6	27.8	31.2	27	29.3	32.8	21.4
PPG	26.1	18.7	18.6	16.7	15.4	15.4	15.4	13.1
RPG	9.3	5.2	7.1	5	8.2	4.5	6.3	6.3
SPG	1.5	0.7	0.6	0.8	0.5	0.7	0.8	0.5
BPG	2.3	0.1	1.1	0.5	0.8	0.6	0.5	0.9

.

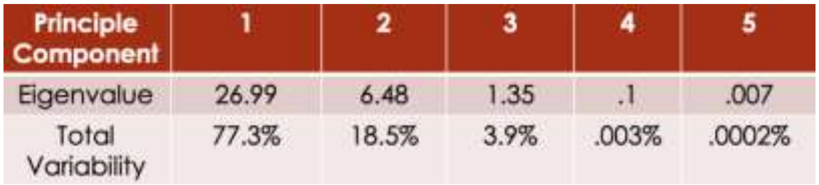
We then converted it into a $5x8$ matrix which we will call X:

$$X = \begin{bmatrix} 34.4 & 29.6 & 27.8 & 31.2 & 27 & 29.3 & 32.8 & 21.4 \\ 26.1 & 18.7 & 18.6 & 16.7 & 15.4 & 15.4 & 15.4 & 13.1 \\ 9.3 & 5.2 & 7.1 & 5 & 8.2 & 4.5 & 6.3 & 6.3 \\ 1.5 & .7 & .6 & .8 & .5 & .7 & .8 & .5 \\ 2.3 & .1 & 1.1 & .5 & .8 & .6 & .5 & .9 \end{bmatrix}.$$

After converting this matrix into mean deviation form and then obtaining the sample covariance matrix, we can find our eigenvalues and eigenvectors of the sample covariance matrix. Once again, D represents a diagonal matrix with the eigenvalues of S in descending order, and V is a square matrix with the corresponding eigenvectors of the sample covariance matrix.

$$D = \begin{bmatrix} 26.99 & 0 & 0 & 0 & 0 \\ 0 & 6.48 & 0 & 0 & 0 \\ 0 & 0 & 1.35 & 0 & 0 \\ 0 & 0 & 0 & .10 & 0 \\ 0 & 0 & 0 & 0 & .007 \end{bmatrix} V = \begin{bmatrix} .69 & -.68 & -.23 & -.04 & .02 \\ .7 & .57 & .41 & -.02 & -.1 \\ .14 & .42 & -.87 & .06 & -.21 \\ .06 & .01 & .001 & .95 & .32 \\ .07 & .17 & -.15 & -.31 & .92 \end{bmatrix}$$

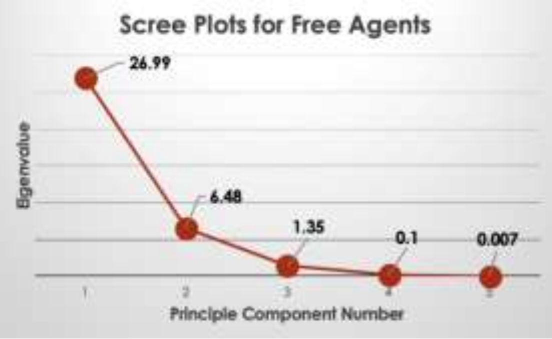
We now have access to our principal components, and we can calculate the percent of variability represented by each PC.



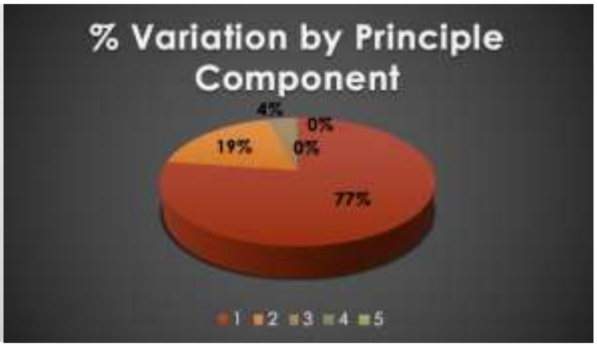
Principle Component	1	2	3	4	5
Eigenvalue	26.99	6.48	1.35	.1	.007
Total Variability	77.3%	18.5%	3.9%	.003%	.0002%

Figure 2

From this chart, we can see that most of the variability in the data can be represented by our first two PCs. Below are a scree plot and a pie chart to represent this.



Scree Plot 2



Pie Graph 2

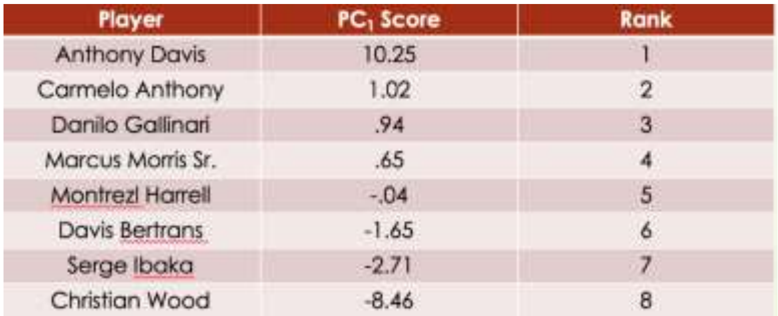
Next, we find each player rating using the equation:

$$Player\ Rating = M^T \cdot PC_1$$

Using our data set,

$$Player\ Rating = M^T \cdot PC_1 = \begin{bmatrix} 5.21 & 8.68 & 2.81 & .74 & 1.45 \\ .41 & 1.28 & -1.29 & -.06 & -.75 \\ -1.39 & 1.18 & .61 & -.16 & .25 \\ 2.01 & -.73 & -1.49 & .04 & -.35 \\ -2.19 & -2.03 & 1.71 & -.26 & -.05 \\ .11 & -2.03 & -1.99 & -.06 & -.25 \\ 3.61 & -2.03 & -.19 & .04 & -.35 \\ -7.79 & -4.33 & -.19 & -.26 & .05 \end{bmatrix} \cdot \begin{bmatrix} .69 \\ .70 \\ .14 \\ .06 \\ .07 \end{bmatrix} = \begin{bmatrix} 10.25 \\ .94 \\ -.04 \\ .65 \\ -2.71 \\ -1.65 \\ 1.02 \\ -8.46 \end{bmatrix}$$

We can then rank each player using the player rating score.



Player	PC₁ Score	Rank
Anthony Davis	10.25	1
Carmelo Anthony	1.02	2
Danilo Gallinari	.94	3
Marcus Morris Sr.	.65	4
Montrezl Harrell	-.04	5
Davis Bertrans	-1.65	6
Serge Ibaka	-2.71	7
Christian Wood	-8.46	8

Player Rankings 2

From here, Anthony Davis is the best available option by a favorable margin. This is what we would have expected based on his statistics in the five areas we used to rank the players. In an ideal world, we would do our best to sign Anthony Davis to the Timberwolves, but that is very unlikely to happen, since he will most likely resign with the Los Angeles Lakers after winning the 2020 NBA Championship. Looking at this data alone, it would be a good idea to pursue either Carmelo Anthony, Danilo Gallinari, or Marcus Morris Sr. in free agency. However, we want to get a better understanding of how these players relate to each other, so we will be using SDI in the next section to help us better understand which free agent we should pursue.

Using SDI to Find Economical Free Agents

The Statistical Diversity Index (SDI) is a very applicable and useful application to principal component analysis. SDI has been used in the past to compare pairwise players against each other to see how similar or different those players are (Bruce, 2016). Then the information garnered from this is used to make decisions on which players to pursue and how much to pay these players in yearly salary.

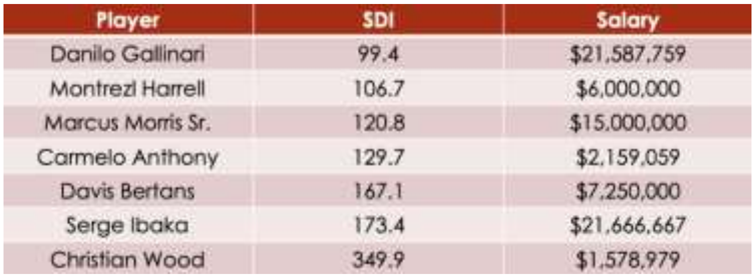
In our case, we will be comparing each of the free agents against Anthony Davis, because he is the ideal candidate for free agency, and it is unlikely that we will be able to get him to sign with our team. For any two players, player i and player j, where $P_{k_{(i)}}$ represents the k^{th} principal component score for player i, the Statistical Diversity Index (SDI) can be calculated as:

$$SDI_{ij} = \sum_{k=1}^{3} \left(P_{k_{(i)}} - P_{k_{(j)}}\right)^2.$$

In our case, Anthony Davis is represented by $P_{k_{(i)}}$ and every player that we will compare against him will be represented by $P_{j_{(i)}}$. We will use the top three principal components, because we can see in *Pie Graph 2* that 99.7% of the variation in our data is represented by the first three principal components. As an example calculation, we found the SDI score of Danilo Gallinari in comparison to Anthony Davis.

$$SDI_{(AD,\ DG)} = \sum_{k=1}^{3} \left(P_{k_{(AD)}} - P_{k_{(DG)}}\right)^2 = (10.25 - .94)^2 + \ (2.81 - (-\,.23))^2 + (-.25 - 1.66)^2 \ = 99.45$$

It is important to note that a lower score is preferable, as that means that a player is more similar to our ideal candidate Anthony Davis. We then found the SDI scores of each of the seven free agents and consolidated those into the table below.



Player	SDI	Salary
Danilo Gallinari	99.4	$21,587,759
Montrezl Harrell	106.7	$6,000,000
Marcus Morris Sr.	120.8	$15,000,000
Carmelo Anthony	129.7	$2,159,059
Davis Bertans	167.1	$7,250,000
Serge Ibaka	173.4	$21,666,667
Christian Wood	349.9	$1,578,979

SDI Scores

What is apparent here is that none of these free agents are very similar to Anthony Davis, as it pertains to both their PC$_1$ and SDI scores, but that is not shocking when we look at their respective statistics. Looking at the data, one's first impulse might be to pursue Gallinari in free agency, but when we take a deeper look, I would suggest that we should pursue Montrezl Harrell instead. Yes, Gallinari has a lower SDI score, but Harrell's is quite similar. When we look at their salaries, Harrell is the much more affordable option. By choosing Montrezl Harrell as the free agent that we want to pursue in free agency, we could possibly save over $15 million dollars in salary cap each season. As the GM, and using both PCA and SDI across the five statistics we chose, I would suggest that we pursue Montrezl Harrell as our number one option in free agency.

From this case analysis, it is apparent that PCA and SDI can be used to help sports management executives make informed decisions on which players to draft and pursue in free agency. Using these techniques, teams can compile the best teams possible, win more games, save millions of dollars against the salary cap, and possibly even win championships.

All-Time NBA Player Rankings

Throughout the remainder of my research, I wanted to use PCA and the Massey Method to compile a ranking of the top 50 NBA basketball players of all time. I specifically chose these two methods because they both use linear algebra techniques, and they both have a proven track record of producing reliable results (Bruce, 2020). The Massey Method in its entirety has never been released to the public, but I am going to apply its main ideas to create my own ranking of the top 50. I will also use PCA to rank these players, and we will compare the results to learn more about how these methods work and how well they can rank players.

Using the Massey Method to Create my Own Ranking

Although powerful and intuitive, principal component analysis is not the only way that linear algebra can be used in data analysis to rank athletes. There are hundreds of different algorithms and metrics that are used to rank and determine the worth of various athletes, but the Massey Method is one of the most powerful and useful methods for ranking teams using a large amount of data. It does extremely well in accounting for the performance of players and teams in the past, and accounts for the variability in that player or teams' performance (Chartier 120). What is particularly interesting about the Massey Method is that it was created by Kenneth Massey during an undergraduate mathematics project, similar to the one that I am currently working on (Chartier, 119). It is a method that was used to rank teams for the BCS and March Madness in collegiate football and basketball, respectively. It is still being used to rank teams in March Madness, but the BCS Standings are no longer used in college football. Another strength of the Massey Method is that it does an excellent job of taking into consideration the overall strength or weakness of an opponent, or in our case, the strengths and weaknesses of particular players in particular statistical areas. In this method, linear algebra and matrix multiplication is used to represent games played and how those results affect the rankings of specific teams.

As we rank all-time great NBA basketball players, we are going to use a similar version to the Massey Method, where we will be weighting each of the statistics, we feel are necessary to represent the worth of a player's career high or low, depending on how important people (or we) deem that stat to be when comparing and ranking the careers of NBA legends. For instance, the number of championships is worth more or has a larger weight than the number of minutes that a player played in their career.

The main piece of the Massey Method that we are going to be using to create our model, is that we will emulate its idea of using a vector representing weights and multiply that vector by a matrix that contains all player statistics. This multiplication will yield player ratings, and then we can rank players based off these ratings.

Looking at my model, there are four different categories that we will use to determine the overall worth of a player's career in comparison to other NBA legends. These categories include *Team Accomplishments, Advanced Metrics, Basic Statistics,* and *Individual Awards.* Each of these categories also have subcategories that make up those 4 categories, as we have listed below. We have also listed the weights of the categories. We have also noted the overall weights for each of the specific sub-categories in parenthesis.

Team Accomplishments 30%:

- Total Championships (11%)
- Total Finals MVPs (8%)
- Total Number of Finals Appearances (6%)
- Total Playoff games played (5%)

Advanced Metrics 20%:

- VORP (Victories above Replacement Player) (4%)
- PER (Player Efficiency Rating) (5%)
- Win Shares (4%)
- Defensive Win Shares (3%)
- Effective Field Goal % (4%)

Statistics 25%:

- Games Played, Total Career Minutes (1.5% each)
- Total Career Points Scored, Career PPG (2.5% each)
- Total Career Rebounds, Career RPG (1.5% each)
- Total Career Assists, Career APG (1.5% each)
- Total Career Blocks, Career BPG (1% each)
- Total Career Steals, Career SPG (1% each)
- FG %, Career FG Made (1.5%)
- 3 Point FG %, 3 Pt. Made in Career (1% each)
- Career FT% (2%)

Awards 25%:

- Number of MVPs (5%)
- Number of DPOYs (3%)
- Number of All-NBA Selections (7%)
- Number of All-Star Selections (4%)
- Number of ROTY (either 1 or 0) (2%)
- Number of All-Defensive Selections (4%)

We have chosen 50 of the top NBA players ever, using Lineup.com's recent article that was released in July 2020, which detailed their list of the 50 greatest NBA players of all time. We have also added one player that we felt could crack the top 50, considering their career accomplishments and statistics.

One thing that was important to us was that the composite scores that we give each of the NBA players for ranking their careers (player rating) was based out of 100 total points. To do that, we will compare the scores of each player against the highest score in that particular category. For example, John Stockton is the all-time leader in assists with 15,806 assists in his career. Michael Jordan has 5,633 assists all time, so we will divide his number by Stockton's total number (since Stockton is the category leader of career assists). This is how we will get the total score that each player earns from each category. In this example,

$$\frac{Total\ Assists_{Player\ being\ Evaluated}}{Total\ Assits_{Category\ Leader}} = \frac{Total\ Assists_{MJ}}{Total\ Assists_{JS}} = \frac{5633}{15806} = .35638.$$

For this specific category, Michael Jordan only gets 35.64% of the points that we assign to the metric of *Total Career Assists*. John Stockton on the other hand would get 100% of the points for the metric *Total Career Assists*. We will use this type of scoring for each of the 32 statistics that we are using in our model. When we use this type of approach for determining the total score for a player's career, this makes it possible that a hypothetical player could have a rating of 100. However, for a player to have a rating of 100, they would need be the all-time leader of all 32 statistics that we are using in our model.

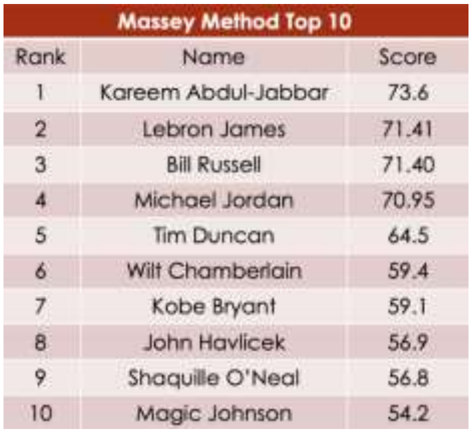
Like the Massey Method, we can use matrix multiplication to ease in ranking these players. Because we have a very large data set with so many rows and columns of data, we will only show part of how we will do this for a few players and a few metrics. We will let row 1 of the $4x4$ matrix below represent Michael Jordan, Row 2 represents LeBron James, Row 3 represents Larry Bird, and Row 4 represents Magic Johnson. The first column of data is their respective number of championships won, the second is PPG, the third is total assists, and the fourth is total minutes played. The $4x1$ matrix or weighting vector represents our scoring weights, with one in the numerator, and the categories' leader score in the denominator. For the first row, we see 11 in the denominator, since Bill Russell won 11 championships (the maximum number for that statistic in league history), so that is the number that each player will be assessed against. For the next row 30.12 is the max number points per game (held by Michael Jordan), so all players will be assessed against that number and so forth and so on. We then execute matrix multiplication to get the rankings of each player:

$$\begin{bmatrix} 6 & 30.12 & 5633 & 41011 \\ 3 & 27.07 & 9346 & 48551 \\ 3 & 24.29 & 5695 & 34443 \\ 5 & 19.54 & 10141 & 33245 \end{bmatrix} \begin{bmatrix} 1/11 \\ 1/30.12 \\ 1/15806 \\ 1/57446 \end{bmatrix} = \begin{bmatrix} 2.62 \\ 2.61 \\ 2.03 \\ 2.31 \end{bmatrix}.$$

Using these 4 metrics, we can see that Michael Jordan has the highest score, followed by Lebron James, then Magic Johnson, then Larry Bird. This is with each of these metrics having the same weight of 1, so the highest possible score a player could earn would be 4. If we were to change the weight (which we do in our algorithm), then the scores could change. An example of this is below, where we have given total championships a weight of 4, PPG a weight of 3, total assists a weight of 2, and total minutes a weight of 1. The more important that we believe a specific metric is to ranking the careers of players, the higher the weight will be. To do this within matrix multiplication, we will change the values in the numerators to whatever the weight is, as shown below:

$$\begin{bmatrix} 6 & 30.12 & 5633 & 41011 \\ 3 & 27.07 & 9346 & 48551 \\ 3 & 24.29 & 5695 & 34443 \\ 5 & 19.54 & 10141 & 33245 \end{bmatrix} \begin{bmatrix} 4/11 \\ 3/30.12 \\ 2/15806 \\ 1/57446 \end{bmatrix} = \begin{bmatrix} 6.61 \\ 5.82 \\ 4.81 \\ 5.68 \end{bmatrix}.$$

The highest score one could get would be 10, since $4 + 3 + 2 + 1 = 10$, so it is very difficult to get a "high score." In our model, weights add up to 100, to normalize it and have scores that look clean. That way, scores can more easily be compared to other players' scores. After using the weights, we have listed above, we were able to get rankings for the top 51 careers of NBA basketball players. We have listed the top 10 below, and the full list of ranked players can be found in *Appendix I.*



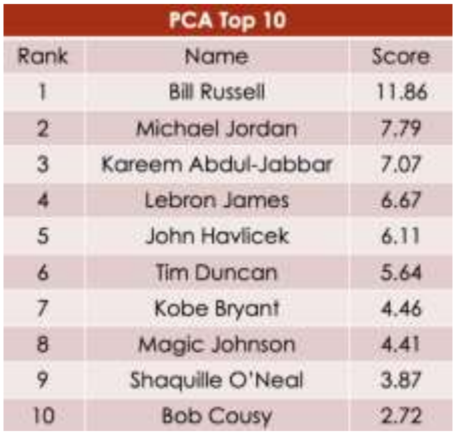
Massey Method Top 10		
Rank	Name	Score
1	Kareem Abdul-Jabbar	73.6
2	Lebron James	71.41
3	Bill Russell	71.40
4	Michael Jordan	70.95
5	Tim Duncan	64.5
6	Wilt Chamberlain	59.4
7	Kobe Bryant	59.1
8	John Havlicek	56.9
9	Shaquille O'Neal	56.8
10	Magic Johnson	54.2

Massey Rankings

Using this model, we were able to rank the worth of the top 51 NBA players of all time. However, there is still a certain level of subjectivity to this model, as we chose our own weights for each of the statistics. A better model would be one that would create those weights for me, based on the variability in my data set. That is where PCA comes into play.

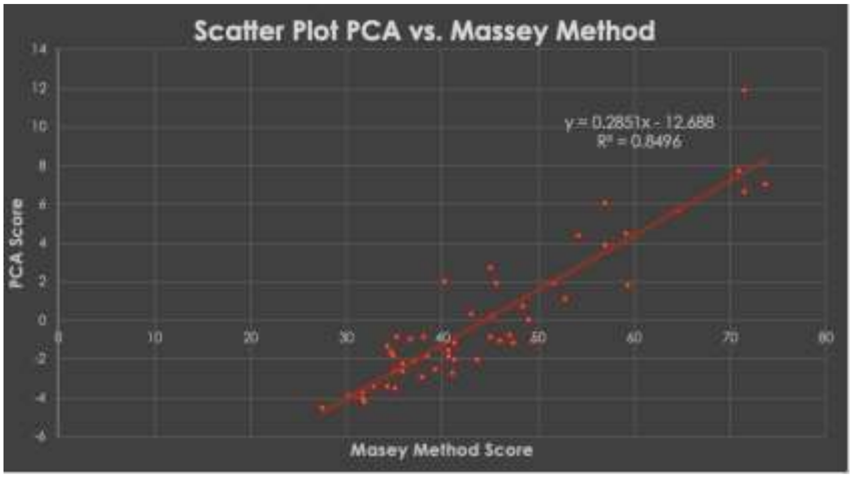
Using PCA to Rank All-Time Greats

Through the process of PCA, we were able to create a different ranking for the top 50 NBA Players of all-time. MATLAB played a vital role in this process, because we were dealing with a much larger data set compared to the data, we used in the case analysis. Having 51 players and 32 different variables or statistics, we will not go into full detail on how we obtained these rankings, but the PCA process is consistent no matter how large that data set. Below is a list of the top 10 players constructed from PCA. A full list of all 51 players with their rankings can be found in *Appendix II.*



PCA Top 10		
Rank	Name	Score
1	Bill Russell	11.86
2	Michael Jordan	7.79
3	Kareem Abdul-Jabbar	7.07
4	Lebron James	6.67
5	John Havlicek	6.11
6	Tim Duncan	5.64
7	Kobe Bryant	4.46
8	Magic Johnson	4.41
9	Shaquille O'Neal	3.87
10	Bob Cousy	2.72

We can clearly see that our results using the Massey Method and PCA to rank the top 10 players are quite different, but it is important to make mention that 9 players appear in both top tens, and that the top four players are the same for each of the rankings (just in different order). This shows that both rankings do well in ranking the careers of NBA players. Next, we wanted to see how similar these rankings were to each other, so we decided to plot them against each other, and see if there was linear correlation between the PCA and Massey Method rankings.



Graphing our Massey Score on the x-axis and our PCA score on the y-axis, we can see that there is a strong correlation between each of these variables. What confirms this is the R^2 value equal to .8496. Upon research, it is clear that the PCA and the Massey Method produce somewhat similar results to each other and very useful methods within sports analytics in ranking athletes.

Areas of Further Research

Due to the timing and the scope of this semester-long research, I was unable to dive deeper into the use of linear algebra techniques in sports analytics. However, there are many different areas of research that I would love to pursue further if I had more time on my hands. Having learned more about PCA, it is a great tool for linear dimensionality reduction. I would love to further my research by studying non-linear dimensionality reduction techniques such as kernel PCA. Kernel PCA is the nonlinear form of PCA, which better exploits the complicated spatial structure of high-dimensional features (Wang, 2012). There are different applications of Kernel PCA in the stock market and to facial recognition in forensic science.

Another area of further research that could be most applicable to this project and to my life as an aspiring data analyst would be to learn how to use many of the technologies necessary for data manipulation and visualization. Throughout my research and during this project, most of the people conducting research in areas like my research used different database tools and coding languages such as R, Python, and SQL. Having a knowledge base of these tools would have saved a lot of time and could have helped me to have performed deeper and more meaningful transformations on the data that I collected. MATLAB worked great for working with my data in matrix form, but it was often very difficult to work with very large data sets in Excel.

Finally, I would want to look deeper into and learn more about different methods that are used in ranking players and in sports analytics. One method I did a little bit of research on was the Colley Method, but it did not seem as practical as the Massey Method for the type data manipulation I was trying to perform on my data set. There are plenty of other methods that exist, and we would be very interested in taking my data and creating many different rankings using the same data but different methods for ranking the data. I could then run tests to see which method is "most accurate" by comparing our rankings each other and to available rankings online.

Appendix I

Massey Method Top 51 NBA Players of All-Time

1	Kareem Abdul-Jabbar	73.61449165
2	LeBron James*	71.41854304
3	Bill Russel	71.40267775
4	Michael Jordan	70.95051876
5	Tim Duncan	64.4695744
6	Wilt Chamberlain	59.4026097
7	Kobe Bryant	59.07194509
8	John Havlicek	56.93187836
9	Shaquille O'Neal	56.84688897
10	Magic Johnson	54.20288607
11	Hakeem Olajuwon	52.82970528
12	Larry Bird	51.53949559
13	Karl Malone	49.31077226
14	Julius Erving	48.95804232
15	Jerry West	48.37603245
16	Kevin Garnett	47.32768123
17	David Robinson	46.89612414
18	Oscar Robertson	45.96886209
19	Scottie Pippen	45.54761852
20	Kevin Durant*	45.16674655
21	Dirk Nowitzki	45.04123798
22	Bob Cousy	45.02631143
23	John Stockton	43.53240314
24	Dwyane Wade	43.02056183
25	Jason Kidd	41.15800979
26	Moses Malone	41.11779964
27	Chris Paul*	41.06960692
28	Gary Payton	40.62739669
29	Bob Pettit	40.56802599
30	George Mikan	40.21882855
31	Charles Barkley	39.27281519
32	Rick Barry	38.46346001
33	Stephen Curry*	37.98841363
34	Patrick Ewing	37.77292629
35	Paul Pierce	37.08650589
36	Kawhi Leonard*	36.69431998
37	Ray Allen	35.81098745
38	Clyde Drexler	35.75323798
39	Willis Reed	35.19784873
40	Artis Gilmore	35.13963609
41	Elvin Hayes	34.96524379
42	Walt Frazier	34.78094182
43	Isiah Thomas	34.72495121
44	Allen Iverson	34.31588446
45	Kevin McHale	34.28513824
46	Steve Nash	32.87599686
47	Vince Carter	31.77743131
48	Reggie Miller	31.7365384
49	George Gervin	31.69963765
50	Dominique Wilkins	30.33736787
51	Alex English	27.57227405

***Represents an Active Player**

Appendix II

PCA Top 51 NBA Players of All-Time

1	Bill Russel	11.86
2	Michael Jordan	7.79
3	Kareem Abdul-Jabbar	7.06
4	LeBron James	6.67
5	John Havlicek	6.11
6	Tim Duncan	5.64
7	Kobe Bryant	4.46
8	Magic Johnson	4.41
9	Shaquille O'Neal	3.87
10	Bob Cousy	2.72
11	George Mikan	2.05
12	Larry Bird	1.96
13	Scottie Pippen	1.90
14	Wilt Chamberlain	1.86
15	Hakeem Olajuwon	1.12
16	Jerry West	0.72
17	Dwyane Wade	0.32
18	Kevin Durant	0.29
19	Julius Erving	0.08
20	David Robinson	-0.70
21	Willis Reed	-0.80
22	Dirk Nowitzki	-0.81
23	Stephen Curry	-0.82
24	Kawhi Leonard	-0.91
25	Oscar Robertson	-1.00
26	Karl Malone	-1.06
27	Moses Malone	-1.12
28	Kevin Garnett	-1.13
29	Kevin McHale	-1.36
30	Bob Pettit	-1.52
31	Isiah Thomas	-1.61
32	Rick Barry	-1.81
33	Walt Frazier	-1.82
34	Gary Payton	-1.85
35	John Stockton	-2.06
36	Jason Kidd	-2.06
37	Paul Pierce	-2.16
38	Ray Allen	-2.20
39	Charles Barkley	-2.54
40	Elvin Hayes	-2.59
41	Clyde Drexler	-2.59
42	Chris Paul	-2.75
43	Patrick Ewing	-2.89
44	Allen Iverson	-3.41
45	Steve Nash	-3.43
46	Artis Gilmore	-3.55
47	George Gervin	-3.67
48	Dominique Wilkins	-3.95
49	Reggie Miller	-3.99
50	Vince Carter	-4.21
51	Alex English	-4.52

References

2020 NBA Free Agents Tracker. www.spotrac.com/nba/free-agents/.

Alamar, Ben. *Sports Analytics: a Guide for Coaches, Managers, and Other Decision Makers*. Columbia University Press, 2013.

Bruce, Scott. "A Scalable Framework for NBA Player and Team Comparisons Using Player Tracking Data." *Temple University Department of Statistics*, Jan. 2016, arxiv.org/pdf/1511.04351.pdf.

Chartier, Timothy P. *When Life Is Linear: from Computer Graphics to Bracketology*. The Mathematical Association of America., 2015.

Colley, Wesley N. *Colley's Bias Free College Football Ranking Method:* Colley's Bias Free College Football Ranking Method:

ESPN, ESPN Internet Ventures, www.espn.com/nba/history/leaders.

Khyati MahendruAs a student of B.Tech in Mathematics and Computing. "Applications Of Linear Algebra in Data Science." *Analytics Vidhya*, 23 Apr. 2020, www.analyticsvidhya.com/blog/2019/07/10-applications-linear-algebra-data-science/.

Lay, D. C. (2006). *Linear algebra and it's applications*. Boston: Pearson, Addison-Wesley.

Manage, Amanda B.W., and Stephen M. Scariano. "An Introductory Application of Principal Components to Cricket Data." *Journal of Statistics*, 2013, jse.amstat.org/v21n3/scariano.pdf.

"NBA & ABA Leaders and Records." *Basketball*, www.basketball-reference.com/leaders/.

Reris, Robert, and Paul Brooks. *Principal Component Analysis and Optimization: A Tutorial*. 13 Jan. 2013, pdfs.semanticscholar.org/e0be/f0bd8e07de281230ae5df28daabb4047e8f0.pdf.

"Sports Analytics." *Index Page*, apbr.org/forum/.

Steinberg, L. (2015, August 18). CHANGING THE GAME: The Rise of Sports Analytics. Retrieved October 31, 2020, from https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/?sh=3ee02da24c1f

The Emergence of Sport Analytics. (2016, May 18). Retrieved October 31, 2020, from http://analytics-magazine.org/the-emergence-of-sport-analytics/

"Top 50 NBA Players of All Time in NBA History (Updated List)." *Sports Articles*, 17 July 2020, www.lineups.com/articles/top-50-nba-players-all-time/.

Wang, Quan. “Kernel Principal Component Analysis and Its Applications in Face Recognition and Active Shape Models.” *Rensselaer Polytechnic Institute*, Aug. 2014, arxiv.org/pdf/1207.3538.pdf.