Mitchell LaRocque
Max Van Sickle
CSCI 4022 project paper

Who are the best young players in the NBA?

Introduction

In this day and age of the NBA, the conversation of the best players in the league focuses on those who have already accomplished great feats. Lebron James is in the conversation for the greatest player of all time with 4 MVPs and 4 championships (among other accolades), Stephen Curry is definitely the greatest shooter of all time, and Kevin Durant is one of the greatest pure scorers this game has ever seen. But, while superstars like these rightfully take their place in the spotlight, young players in the NBA lurk in the shadows, waiting for a moment to prove their worth and greatness as a star. While the vast majority of the public cares about the best players right now, who among the NBA's young core will prove to be the next face of the league?

Taking a step back, this project was done more out of curiosity than to have any-real world impact. Our conclusions will likely not change any opinions besides our own, although we wish it did. From our point of view, the media focuses heavily on flashy players and players who appeal to a wider audience at the price of moving the spotlight away from more technically sound and possibly better players. Only a die-hard fan could recognize the greatness in players who pour their heart onto the court night-in and night-out with no general acknowledgement. We want to find out if this is true for ourselves, and if we have a bias controlled by the media as well.

For both of us, our largest motivator behind the project was our love for the game. Both of us play recreationally, either with pickup games at the CU Rec Center or on an intramural team. We each have our separate interests that draw us to the NBA and keep us watching and attending games when given the opportunity. With the advent of the league's 75th anniversary, the NBA released a list of the top 76 greatest players to ever lace up. A very nostalgic season has had us reminiscing about the past but also looking towards the future. Especially now, when players our age are entering the league with speculation about their future performance, we chose to seek out which young players show promise to become one of the all-time greats.

Data

We have two main datasets, both of which were sourced from basketball-reference.com. The first dataset contained the 76 players on the NBA 75th anniversary team, and contained the following columns:
Career Years: **FROM, TO**
Total Career Games Played: **G**

Per Game Stats: **MP, PTS, TRB, AST, STL, BLK**
Shooting Stats: **FG%, 3P%, FT%**
Win Share Stats: **WS, WS/48**

The most glaring problem with this dataset is that it is much more limited in statistics when compared to the statistics recorded in the moden NBA, as it does not contain features like Attempts (FGA, 3PA, 2PA, FTA), so stats like eFG% are impossible to compute. It is missing the defensive stat Turnovers (TOV), stats such as STL, BLK, 3P%, and additionally more advanced statistics do not exist for some players. This is because the NBA did not start tracking blocks or steals until the 1973-74 season and the 3pt line was not added until 1979. So, this causes a large problem. This makes it more difficult to compare modern players to some of the players on the 75th anniversary team, as some players lack defensive statistics and thus have less data on which to show their greatness. Players like Bill Russell and Wilt Chamberlain were defensive powerhouses in their time, and reports show that some seasons they each averaged over 8 blocks per game (seen here), and Wilt averaged around 2 steals per game (seen here). So, with less statistics, this makes our basis on which to judge modern players much more limited.

The second dataset contains all NBA players who are under the age of 25, played more than 8 minutes per game, and who played more than 5 games in a year. In order to get a list of the players under 25 we looked at the last three NBA seasons, and filtered it to include weighted averages (by games played) for all stats that appear in the 75th anniversary team data. This dataset included these columns:
Season Stats: **G, GS, TM, AGE**
Per Game Stats: **MP, PTS, ORB, DRB, TRB, AST, STL, BLK, PF**
Shooting Stats: **FG, FGA, FG%,3P, 3PA, 3P%, 2P, 2PA,2P%, FT, FTA, FT%**
Win Share Stats: **WS, WS/48**

Originally, the datasets we downloaded for the players under 25 had much more advanced statistics than appeared in the 75th anniversary team data. These advanced statistics can not be used for comparison to the legends so we excluded them. Win shares and win shares per 48 were not originally in the dataset, so we imported each season's advanced stats and added these columns to our original data. From this, we created a list of career stats from the 75th anniversary team exactly compared with the stats of all players under 25 in the past 3 seasons. We took the last 3 seasons to analyze the top young players in the league because that gives a good metric of where these players will be headed in their career. If we go back too far, we run into problems like players' stats can be skewed by low numbers in their first seasons or players not being in the league yet.

Real World Impact:

NBA analytics is very much a real world field. Disregarding NBA teams themselves who constantly use data science strategies to gain a competitive advantage, many companies such as Second Spectrum (https://www.secondspectrum.com/index.html) use cutting edge technologies to get unparalleled machine understanding of sports games. Such establishments could find a much more in-depth answer to our question. Examination of NBA analytics is a very interesting topic evidenced by the large amount of academic papers available. Here are some of the most interesting papers I could find in a miniature literature review: Optimal Endgame Strategy in Basketball (https://calhoun.nps.edu/handle/10945/40337), Scoring and Shooting Abilities of NBA Players (https://www.researchgate.net/publication/46554952_Scoring_and_Shooting_Abilities_of_NBA_Players), Ups and Downs: Team Performance in Best-Of-Seven Playoff Series (https://www.sfu.ca/~tswartz/papers/desperation.pdf). Finally, although the following is not an academic paper, it does attempt to directly answer the question we propose in this project so we deemed it necessary to include here: Top 10 under 25 (https://www.cbssports.com/nba/news/nba-top-10-under-25-trae-young-zion-williamson-looking-up-at-luka-doncic-deaaron-fox-barely-misses-cut/).

Exploratory Results:

After our initial processing of our data we wanted to get a basic understanding of possible 4022 methods and results that we could see in our results. To begin, we noticed that one of the statistics we have in both our datasets were total win shares. Win shares is a really great statistic to demonstrate how much overall success a single player contributes to their team. According to Basketball Reference: win shares is “a metric that estimates the number of wins a player produces for his team throughout the season.” This was the best statistic we had for the greatness of individual players. For our first baseline understanding we printed out the players in the under 25 data frame with the top 5 win shares. The resulting players were: Nikola Jokic, Bam Adebayo, Jarrett Allen, Domantas Sabonis, and Luka Doncic. This exploratory result was already pretty good for estimating some of the best young players in the NBA.

Next, as a secondary exploratory result we took a look at the offensive and defensive volume of individual players compared to the all time greats. Our offensive volume metric was pretty rudimentary being (PTS+AST+REB) while our defensive volume metric was (STL+BLK). We found the outliers from all under 25 players in each of these categories and plotted them against the 75th anniversary team to see how they stacked up. The top outliers for offensive volume were: Karl-Anthony Towns, Trae Young, Nikola Jokic, Devin Booker, Domantas Sabonis, Ben Simmons, and Luka Doncic. These are definitely some of the best young offensive players in the league and they stack up very convincingly against the greats. The top outliers for defensive volume were: Myles Turner, Mitchell Robinson, Elijah Bryant, Ben

Simmons, and Dejounte Murray. These players stack up well defensively against the all time greats, however, with their lack of offensive volume it didn't seem that they would be part of our final results.

Methods:

The first methods we used were mahalanobis distance and i-forest calculations in order to detect outliers for offensive and defensive volume. These methods were used to gain an exploratory result for a baseline understanding. Both of these methods produce expected results for outlier detection. Mahalanobis distance is a multidimensional distance measure of how many standard deviations a player is away from the mean of the distribution of all players. I-forests are a non-parametric probabilistic outlier detection method based on the idea that it is easy to separate anomalous observations from normal observations using less conditions.

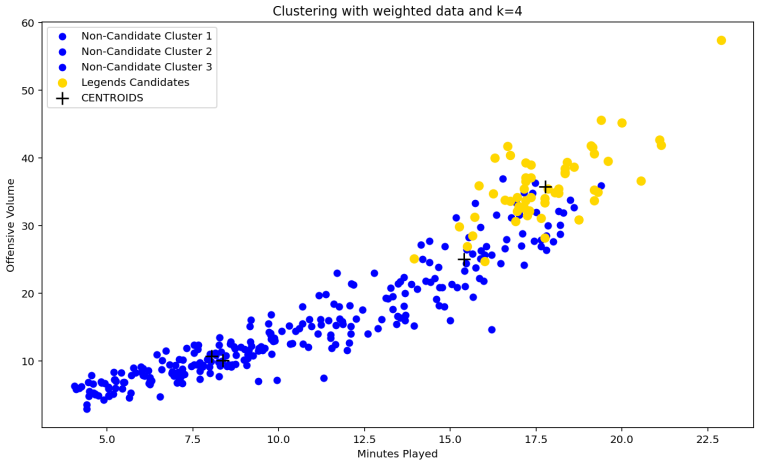
It is difficult to do a complete comparison of the under 25 player data set with the 75th anniversary team dataset because the 75th anniversary team dataset is not complete. There is a lot of missing data for some older players in statistical categories like STL, BLK, and 3P%. The next method we used was a UV decomposition to get an idea of what these missing data could have been. The idea of UV decomposition of a matrix M is to find entries for matrices U and V such that UV = M' and M' is very close to M. First, a couple of arbitrary initialization parameters: the intermediate matrix dimension and the tolerance error for UV convergence were chosen based on the most accurate M' we would get in a short amount of time rather than any kind of computed value. After the UV decomposition we noticed that the resulting missing values for M' were systematically too large. We determined that was because of the following. The UV decomposition has a pretty good understanding of a players' outlierness. This is because the decomposition must produce every players' known stats accurately (because of the low RMSE), therefore for an individual players' unknown stats the decomposition understands that players' relative statistical deviations and thus recognizes outlierness of a players' value in any given statistical column. In order to accommodate for this, we performed a baseline normalization using a known value. Some researchers have found that Wilt Chamberlain averaged around 8.8 blocks and 2.1 steals per game. In order to fix the unknown values we got from UV decomposition, every unknown value for blocks and steals was multiplied by the ratio of Wilt's known stats to his estimated stats. Although the resulting values estimated from UV decomposition aren't necessarily correct they give a good proportional idea of a players' role and outlierness. When comparing players, if we can maintain this sense of proportionality then the comparison will be valid. Finally, the UV decomposition gave us values for 3P% that were not percentages. Again, using the idea of proportionality we found a solution. The 3P% column was tricky because the UV decomposition was giving us values that were >1. In order to accommodate for this we divided the values that were >1 by the total sum of all estimated 3P%.

Again, UV decomposition has a sense of players' role and outlierness. If I divide each estimated value for 3P% by the total estimated sum for 3P%, the proportion of outlierness will be preserved.

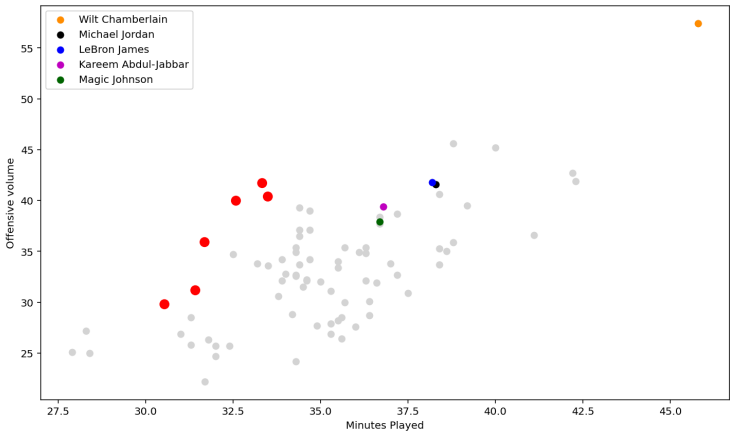
Finally, in order to get an answer to our original question we used k-means clustering of both the 75th anniversary team and the under 25 players dataframe. We performed this k-means clustering in 10 dimensions using all 12 statistical categories of the 75th anniversary team besides GP (games played) and WS (win shares). We didn't use games played in our comparative clustering because it's not a performance measure, it demonstrates longevity and when we are comparing legends to young players we don't want to look at longevity. We ended up dropping win shares due to redundancy with WS/48 and because the stat is a result of time spent in the league, we figured WS/48 was a better comparison for young players. We were then left with 10 columns on which to cluster. There are a total of 249 candidate young players and 76 total players on the 75th Team, for a total of 350 players we're comparing to each other. We began with k=2 clusters, the legends candidates cluster contained 129 people, with 54 being current players under 25 and 75 players being from the 75th team. This is close to what we want, as our goal is to create a cluster that maintains as many 75th team legends as possible while including as few under 25 players as possible. This will give us a small filtered list from the 249 candidates to a small number of players who belong with the legends. 54 current players who belong with the legends is too large of a number so for our next iteration we decided to increase the number of clusters and find a more optimal cluster which contains the best of the best. Additionally, although we are using 10 separate statistical columns for comparison, the weights of each column is not weighted how we like. For example, the range of WS/48 is only .376, while the range of Minutes Played is 37.6. Thus, with our current dataset, Minutes played is 100x more impactful for the resulting cluster because the range is much larger, even though it is not a stat that demonstrates any kind of performance or skill. Furthermore, we know that win shares is a very telling statistical category from our exploratory results so we need it to be weighed much higher. We provided arbitrary weights to each category that we thought would produce the best players from k-means clustering. After deliberation we found the correct initialization and number of clusters to give us a cluster with 6 players under 25 and 57 legends.

Results:

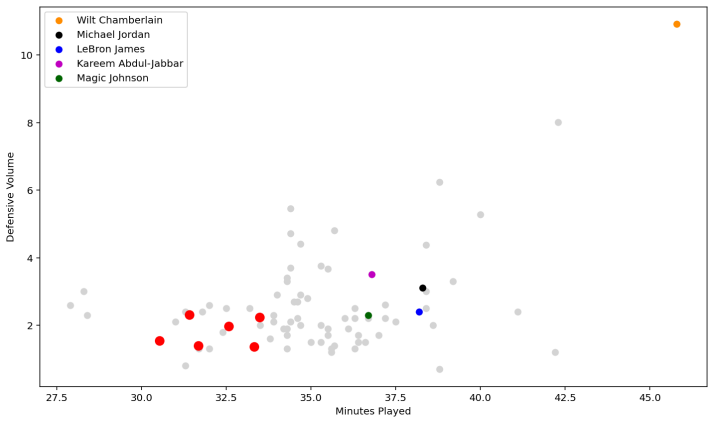
After changing our number of clusters from 2 to 4, and using our adjusted weighted stats, the Legend Candidate Cluster contained 57 total players with 6 of them being under 25, creating our final list. This group had an ideal ratio, as our goal was to find a cluster with as many 75th Team players as possible while finding a small list of players under 25 who belong in this category. Thus, we were left with 6 players under 25, John Collins, Luka Dončić, Nikola Jokić, Kristaps Porziņģis, Karl-Anthony Towns, and Zion Williamson.



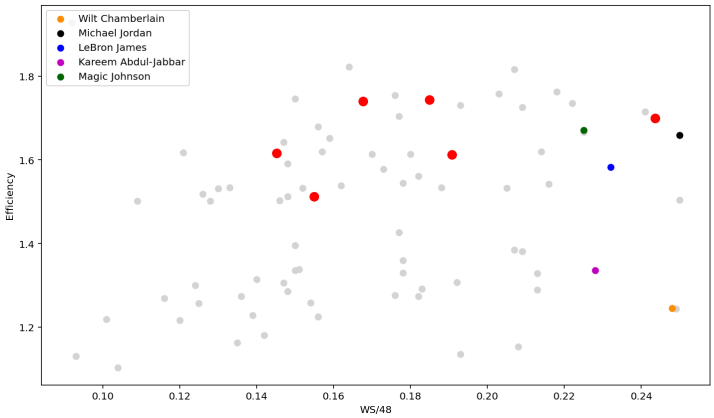
Final Clusters Visualized on Offensive Volume vs MP



Top young players vs legends in offensive volume vs minutes played



Top young players vs legends in defensive volume vs minutes played



Top young players vs legends in efficiency vs win shares per 48

A k value of 4 was much more appropriate, as this way we can get an upper cluster of just the best players our of both sets. The remaining 3 clusters have much less importance, and while we could likely rank these remaining clusters, they have no value for us.

Some of the names on this list made a lot of sense to us, as people like Luka Dončić and Nikola Jokić are clearly all-star talent and are easily some of the best young players in the league. But other names, like John Collins and Kristaps Porziņģis, were very surprising because neither of us would consider either of these players to be apart of the “top 10” players under 25. But, the results of these two players were important for us, to check our own biases of who we thought belonged on this list of 6. We easily would have picked other players before these two, but it is nice our list contains players we wouldn’t have thought of.

One question that came up for us after seeing these results is why only 51 players of the 75th team were in this cluster, and if the remaining 25 players were considered "worse" than the

6 young players in the cluster. This actually makes sense to us, as these 6 young players easily could be better than some names on the 75th list because the skill ceiling of modern players is continuously increasing. Additionally, because these young players are not far enough into their career, they were not even considered for the list in the first place. They have much more to prove, despite their greatness already.

We had the Final Ranking of the 6 young players by PER:

	Player	PER
0	Nikola Joki	31.3
1	Zion Williamson	27.1
2	Luka Don	25.3
3	Karl-Anthony Towns	23.1
4	Kristaps Porzi	21.3
5	John Collins	20.6

To answer the question of the BEST young player, we used PER (Player Efficiency Rating) to measure this stat, as this shows a rating of a player's productivity. This is an excellent advanced stat, as it's per-minute and is pace-adjusted, meaning it normalizes lots of factors between players to create a very fair statistic. Because this stat is extremely powerful, we thought it was a great way to rank players.

Nikola Jokić is the best young player under 25, and the most likely to be considered the "GOAT".

Conclusion:

There are many unanswered questions after obtaining our results. What happened to Devin Booker and Jayson Tatum? Our clusters are very volatile and depend heavily on initialization, so we stopped at k=4 in order to maintain consistent results. How can we get consistent results that are not heavily dependent on initial conditions, would these results be better than our current results? We tried to maintain the most amount of legends in our legend candidate cluster. What if we simply asked which player under 25 is most likely to come from the legends set rather than which player under 25 is the best? Would these results be the same? A lot of these questions could be answered using a neural network or a linear model. We used some arbitrary numerical values for weighting the k-means clustering algorithm and UV decomposition. Using a linear model to compute these arbitrary numerical values rather than personal discretion would give us the optimal unbiased values to run our model better. A neural network could be a better way of answering our initial question using our datasets. Neural networks could decompose given statistics for each player into abstract concepts to place players into categories more effectively than our k-means model. However well a different approach could do, our model still produced very satisfactory results. We knew that some player's like

Nikola Jokić and Luka Dončić would definitely be among the legends so our model did a good job as they were in our top 3 young players. The other players in our results make a lot of sense as well, it was good to see players like Kristaps Porziņģis and John Collins who are very good young players appear in our results. Although we might have expected Devin Booker or Jayson Tatum, our model might have picked up on something that we couldn't. Our results unquestionably gave us some of the best young players in the NBA in an unbiased fashion, giving us a lot to think about in terms of our opinions about young NBA athletes.