

Final Paper

Seth Keim

1/31/2021

Introduction

The never-ending debate: “Who is the greatest basketball player of all time, LeBron James or Michael Jordan?” After LeBron James won his fourth NBA title in 2020, the question is now more controversial than ever. The main issue with this question is that everyone has a different definition of “greatest”. Does greatest mean which player would you choose to be on your team first, or which player was the most successful during their career, or some combination of the two? In the end, “greatest” can be subjective. In this analysis, the impact on the game over time will be analyzed for both Michael Jordan and LeBron James, as well as other NBA players. Impact, in this way, means the amount of influence a player has on the outcome of any particular game.

This metric will still be imperfect. As such, this analysis will not seek to find a definite answer the question of which player is the greatest. However, it will look at the impact of many NBA players over time, in the regular and post seasons. The topic of determining impact in basketball (or any sport, for that matter) is a vital one for coaches, players, and commentators, and an interesting one for fans. For coaches, analyzing players performance can unlock insights that can increase performance of the team. For players, determining their own impact will help them improve that impact.

The question this analysis will focus on is: how much impact did Michael Jordan, LeBron James, and other NBA players have on the games they play, and what are the best ways to quantify and display it? We will look at four main ways to quantify greatness and data visualizations for each. First, looking at per game stats for each season a player plays can show the players quality throughout their career. Some players, with LeBron as an example, change their playstyle as their career progresses. When LeBron was young, he focused on driving to the basket and scoring. However, as he has aged and become less quick, he has focused much more on passing and shooting. Looking at trends over time can reveal these types of things. Second, we will look at career total stats. These are good gages of a players accomplishments and how impressive they are compared to other NBA greats. If a player is a great scorer when they were young, but falls off quick and ends up retiring relatively early, that will show up as fewer career total stats. This seems less impressive as compared to a player which was much more consistent throughout their career and managed to still be competitive at an older age. Third, we look at the impact on their teams, and the success of those teams. While individual stats can be useful in looking at how talented a player is, it also leaves out the impact that player has on the rest of the team. A player can have terrible personal stats and still make the players around them much better, which is hugely beneficial to the team. It can also be the case that a player can have great stats, but they have the ball too much and actually makes the team perform worse. Finally, we look at the rank of a player in a specific stat for each season. This allows comparison of a player against their respective league. It may be that comparing hard numbers for stats between a player in the 1970s and the 2010s is unfair. The three point line did not exist until 1979, so it is likely that players were typically scoring less than there were later on with the 3 point line. The limitations of this would occur in the situation that the players of one era are just worse than another era. A player could be dominant in the worse era, but this would be misleading as the league their playing against is just less competitive than leagues other players have played in.

All of these forms of analysis will contribute to the research question. Likely, the answers will come with nuance. Different players may dominant different aspects of the game. Exploring this nuance will be the main goal of the analysis. As was discussed earlier, there are many different definitions of greatness, and even

impact on the game can be subjective. Analyzing specific parts of the game, like finals, playoffs, and regular season, or comparison against the current league or historically, or defensive or offensive stats can tell the story of the best basketball players to play the game depending on the lens you look at them through.

The data used in this paper is obtained using the NBAstatR R package. This package pulls NBA data from Basketball-Reference, a website that contains NBA throughout its history. It includes player totals, player per-game stats, and team stats.

Methods

Pulling in the data

Create Functions to get data from nbastatr api and return it partially cleaned

To get all the necessary data, we must use the nbastatr package in R. This package contains a multitude of functions that scrape basketball statistics from Basketball Reference, a basketball data site. The package allows calling every basketball player that has ever played in the league, and also has in depth statistics by per game stats or total stats. To get all the necessary data, we must loop through all the players and call the function “players_careers”. This will return the data we need for each specific player. Then, appending each new player stats to one data frame will put all of the data in one place. Since this package hits the website Basketball Reference, it is actually somewhat difficult to call functions a large number of times. In this case, the website recognizes that the same user is getting its data many times and shuts off access to the website. This posed a difficult problem to solve at first. However, by calling the function for 300 players at a time, the website never reached its limit. After 300 players stats were obtained, restarting the R session allowed for the website to reset and repeat the process. There were a total of over 4500 players, so this process had to be repeated 15 times. Each time, the new dataset had to be appended to the existing dataset and saved to a csv so the data was stored and not deleted with the R session was restarted.

Create dataset with all current and historical NBA players stats

Once the data scraping is done, we have a dataset with all the players stats per game every season that they played. This ends up being almost 26,000 rows of data, with 29 columns. We can take a look at the dataset here:

```
## [1] "Dataset"

## # A tibble: 6 x 6
##   numberPlayerSeason gp agePlayer pts playerName year
##   <dbl> <dbl> <dbl> <dbl> <chr> <dbl>
## 1 0 43 23 3.1 Alaa Abdelnaby 1991
## 2 1 71 24 6.1 Alaa Abdelnaby 1992
## 3 2 12 25 5.3 Alaa Abdelnaby 1993
## 4 2 63 25 8.2 Alaa Abdelnaby 1993
## 5 2 75 25 7.7 Alaa Abdelnaby 1993
## 6 3 13 26 4.9 Alaa Abdelnaby 1994

## [1] "Column Names"

## [1] "numberPlayerSeason" "slugTeam" "gp"
## [4] "gs" "fgm" "fga"
## [7] "pctFG" "fg3m" "fg3a"
## [10] "pctFG3" "pctFT" "fg2m"
## [13] "fg2a" "pctFG2" "agePlayer"
## [16] "minutes" "ftm" "fta"
## [19] "oreb" "dreb" "treb"
## [22] "ast" "stl" "blk"
## [25] "tov" "pf" "pts"
```

```
## [28] "playerName"      "year"
```

This only includes 5 columns that were selected, so all the column names are printed as well. This data is extensive, and will provide the bulk of the data that we need for the graphs. Even though this leaves out total points, rank of players, and other statistics, most of these can be obtained just by manipulating the data here. For example, totals can be obtained by multiplying points per game by games played for each season, and then summing those values up for each player.

Data manipulation

Per Season Stats LeBron vs Jordan

Next, we need to manipulate the data in order to make the charts that we need. First, creating a dataset that just has statistics for Michael Jordan and LeBron James will allow for over time charts. These charts can be used to compare these two great players. We can take the original dataset with all nba players in it, and filter out by the player's name being either Michael Jordan or LeBron James. This gets us close, but we have to do more. Michael Jordan retired two times midway through his career. In our dataset, this does not show up, he just does not have any rows for those years. However, when graphing that data, it will not show the separations that signify his retired years, instead the chart will just fill up the holes. So, we can add in four more rows to this dataset to represent the four years Michael Jordan was retired before returning to the NBA. For each row, most columns can be NA, except for playerName, year, and agePlayer. With NA in for the stats, it will tell the chart to put a separation in for that year. This will make the chart much more true to the data and give the audience more information.

```
## slugSeason numberPlayerSeason isRookie slugTeam idTeam gp gs fgm fga
## 1 2003-04 0 TRUE CLE 1610612739 79 79 7.9 18.9
## 2 2004-05 1 FALSE CLE 1610612739 80 80 9.9 21.1
## 3 2005-06 2 FALSE CLE 1610612739 79 79 11.1 23.1
## 4 2006-07 3 FALSE CLE 1610612739 78 78 9.9 20.8
## 5 2007-08 4 FALSE CLE 1610612739 75 74 10.6 21.9
## 6 2008-09 5 FALSE CLE 1610612739 81 81 9.7 19.9
## pctFG fg3m fg3a pctFG3 pctFT fg2m fg2a pctFG2 agePlayer minutes ftm
## 1 0.417 0.8 2.7 0.2962963 0.754 7.1 16.2 0.4382716 19 39.5 4.4
## 2 0.472 1.4 3.9 0.3589744 0.750 8.5 17.2 0.4941860 20 42.3 6.0
## 3 0.480 1.6 4.8 0.3333333 0.738 9.5 18.3 0.5191257 21 42.5 7.6
## 4 0.476 1.3 4.0 0.3250000 0.698 8.6 16.8 0.5119048 22 40.9 6.3
## 5 0.484 1.5 4.8 0.3125000 0.712 9.1 17.1 0.5321637 23 40.4 7.3
## 6 0.489 1.6 4.7 0.3404255 0.780 8.1 15.2 0.5328947 24 37.7 7.3
## fta oreb dreb treb ast stl blk tov pf pts playerName year
## 1 5.8 1.3 4.2 5.5 5.9 1.6 0.7 3.5 1.9 20.9 LeBron James 2004
## 2 8.0 1.4 6.0 7.4 7.2 2.2 0.7 3.3 1.8 27.2 LeBron James 2005
## 3 10.3 0.9 6.1 7.0 6.6 1.6 0.8 3.3 2.3 31.4 LeBron James 2006
## 4 9.0 1.1 5.7 6.7 6.0 1.6 0.7 3.2 2.2 27.3 LeBron James 2007
## 5 10.3 1.8 6.1 7.9 7.2 1.8 1.1 3.4 2.2 30.0 LeBron James 2008
## 6 9.4 1.3 6.3 7.6 7.2 1.7 1.1 3.0 1.7 28.4 LeBron James 2009
```

Mutate data for Career total stats

Next we need to manipulate the dataset to include total stats. We can do this by multiplying per games stats by games played, grouping by player, and then summing those values. We do this for the 5 main statistics in basketball - Points, Assists, Rebounds, Steals, and Blocks. Finally, in order to allow for 5 different bar charts to show each of these stats, we pivot the dataset and make the stats type a new column.

```
## # A tibble: 6 x 3
## Player stat total
## <chr> <chr> <dbl>
```

```
## 1 Kareem Abdul-Jabbar pts 38387
## 2 Kareem Abdul-Jabbar trb 17440
## 3 Kareem Abdul-Jabbar ast 5660
## 4 Kareem Abdul-Jabbar stl 1160
## 5 Kareem Abdul-Jabbar blk 3189
## 6 Karl Malone pts 36928
```

Mutate data for ranks in each stat by year

Next, we need to manipulate the dataset to include ranks of players in specific stats by year. This is done by grouping by year, and creating a new column for each stat that we want whose value is the sorted order of our original stat. We also need to include the NA data for Michael Jordan that we created earlier. Next, we filter for LeBron and Michael Jordan, since those are the only two players we will be looking at. Next, we create another dataset that takes the averages of the ranks over the years LeBron and Jordan have gotten. Pivoting this dataset allows for multiple charts using the same plot as well.

```
## [1] "Ranks dataset"

## # A tibble: 37 x 9
## # Groups:   year [37]
##   year numberPlayerSea~ gp playerName ptsrank astrank trebrank blkrank
##   <dbl>          <dbl> <dbl> <chr>          <int>   <int>   <int>   <int>
## 1 2004              0    79 LeBron Ja~      15     16    101     87
## 2 2005              1    80 LeBron Ja~       4      8     46     81
## 3 2006              2    79 LeBron Ja~       3     11     43     59
## 4 2007              3    78 LeBron Ja~       6     19     53     68
## 5 2008              4    75 LeBron Ja~       1     11     47     55
## 6 2009              5    81 LeBron Ja~       2     11     44     41
## 7 2010              6    76 LeBron Ja~       2      5     53     59
## 8 2011              7    79 LeBron Ja~       2     16     48    128
## 9 2012              8    62 LeBron Ja~       3     14     32     86
##10 2013              9    76 LeBron Ja~       4     10     32     55
## # ... with 27 more rows, and 1 more variable: stlrank <int>

## [1] "Average Ranks Dataset"

## # A tibble: 10 x 3
##   playerName      stat    avgrank
##   <chr>          <chr>    <dbl>
## 1 LeBron James ptsrank     6.17
## 2 LeBron James astrank     9.89
## 3 LeBron James trebrank    45.3
## 4 LeBron James blkrank   103.
## 5 LeBron James stlrank    31.9
## 6 Michael Jordan ptsrank     4.6
## 7 Michael Jordan astrank    39.2
## 8 Michael Jordan trebrank    71.4
## 9 Michael Jordan blkrank    81.8
##10 Michael Jordan stlrank    11.3
```

That is the final dataset needed to complete the charts we will make. Because of the messiness of obtaining the data, the scraping stage is most time-consuming and difficult. However, once the data is saved to a csv, creating other useful datasets from the original data is relatively easy. Next, we move on to create the charts and see results from the data.

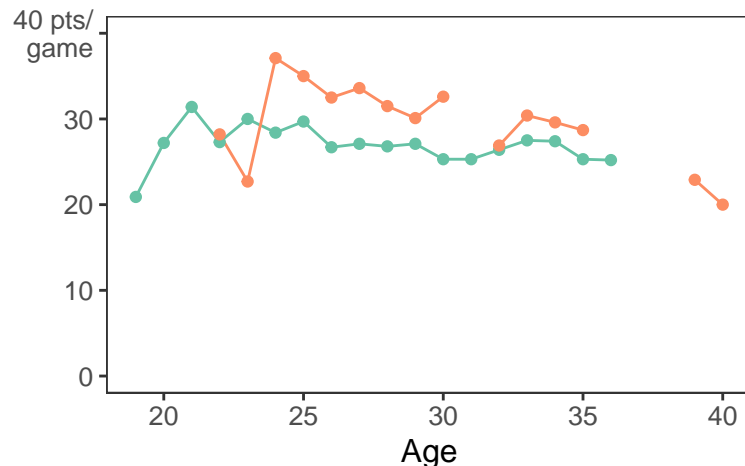
Results

Graph 1: Stats over time

The first chart we create is the per game stats over time chart. This chart allows comparison of players over their career.

Points Per Game Over Time

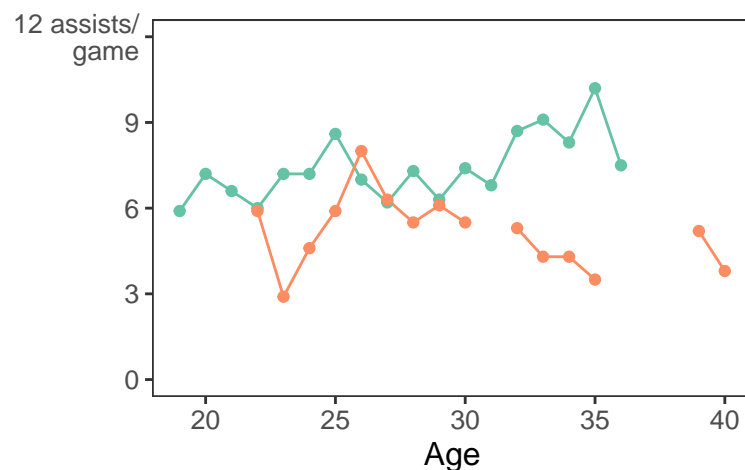
LeBron James vs Michael Jordan



Looking at this chart, the effort from including the NA rows pays off, as Jordan's retired years show up blank. From this chart, it is easy to see that Michael Jordan was a more prolific scorer throughout his career. LeBron was originally known for his scoring ability when he was young, but he has since changed his style of play significantly. So, this chart does not tell the full story. If greatness was measured solely on points scored per game, Jordan would surely be the GOAT. However, there are many more important stats and metrics to compare these two great players.

Assists Per Game Over Time

LeBron James vs Michael Jordan



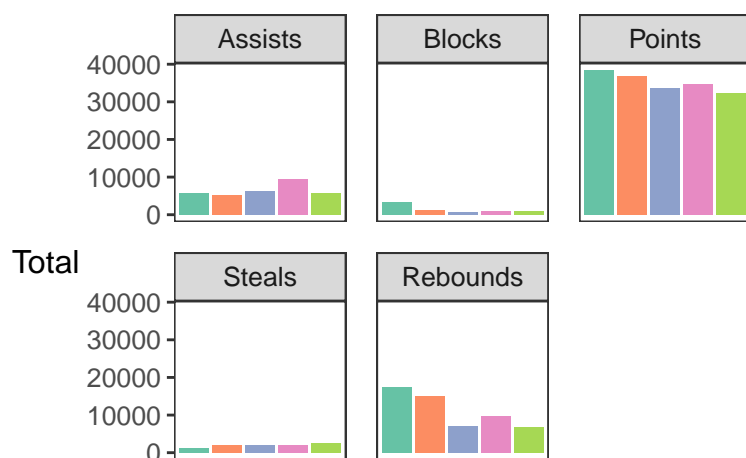
This chart shows the players assists per game over time. Despite Michael Jordan being smaller and a guard, LeBron actually consistently has had more assists per game over his career. In addition, he has continued to rack up more assists per game as he grows older and more experienced. So, while points will be considered by many to be more important, if greatness is measured by versatility, LeBron would probably win. Being a 6'9, 250 lbs small forward that can also dish out passes is a much more versatile player, despite Jordan being a better individual scorer.

Graph 2: Career Stats

Next, we will look at career total stats. This should show how Jordan and LeBron measure up to the other greats of the NBA. Some would measure greatness and how many records a player holds, and how impressive their career total stats are compared to other players.

Career Stats of Top 5 Scorers

Kareem Abdul-Jabbar, Kobe Bryant, LeBron James



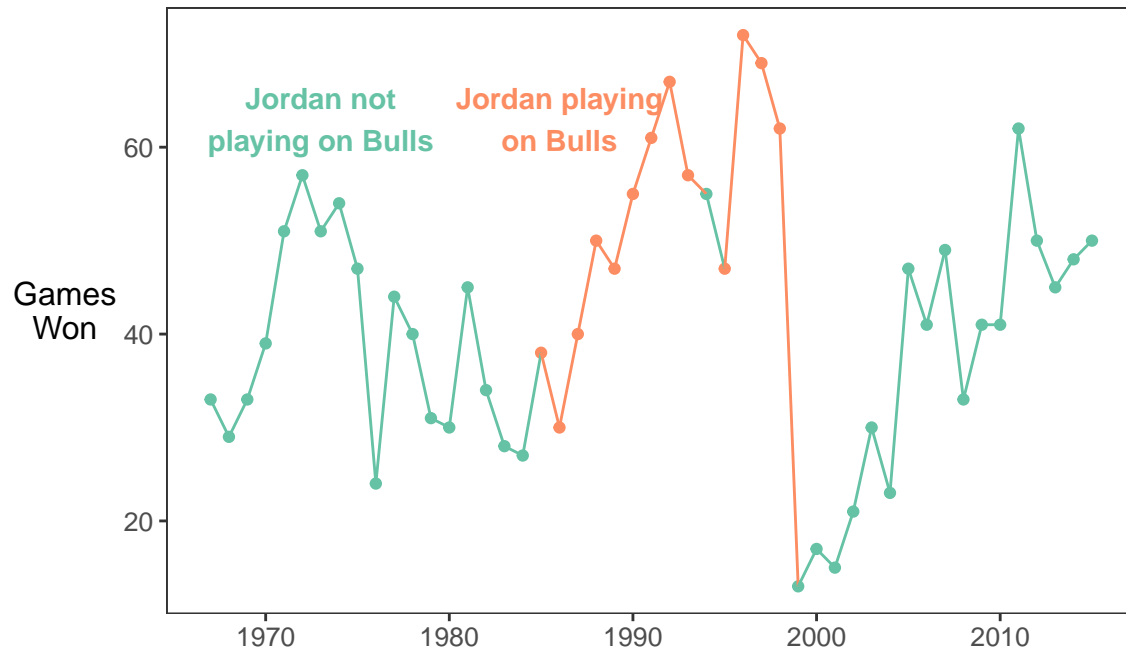
In this chart, the top 5 career total scorers of all time are compared. If greatness was mainly based on career total records, then Kareem Abdul-Jabbar would actually have a great argument for greatest of all time. He has the most career points of any player, and he dominates other categories like blocks and rebounds. Now, this is slightly unfair just comparing those stats with the top scorers, as Kareem was the most traditionally big man out of these players. That being said, Kareem mainly makes his way up these stats because he has played a huge number of minutes and games, and his success was typically with other great players on his team. While in this analysis we only look at regular season, one of the main reasons Kareem is not considered in the same league as LeBron and Jordan is due to his playoff performance. Despite winning 6 NBA Championships, Kareem only won two Finals MVPs, meaning he was not even considered the best player on his team for four of his championships.

Graph 3: Career Success

Moving on to the next analysis, we look at each player's effect on the teams they played for, and their teams' regular season success. A player's impact on making the team better is an important part of them being great. If a player has great stats, but he ends up taking too many shots to get those stats and makes the team perform worse, that effect would show up in this chart. The data for this only goes up until 2015, which is unfortunate as it leaves 5 seasons off from LeBron.

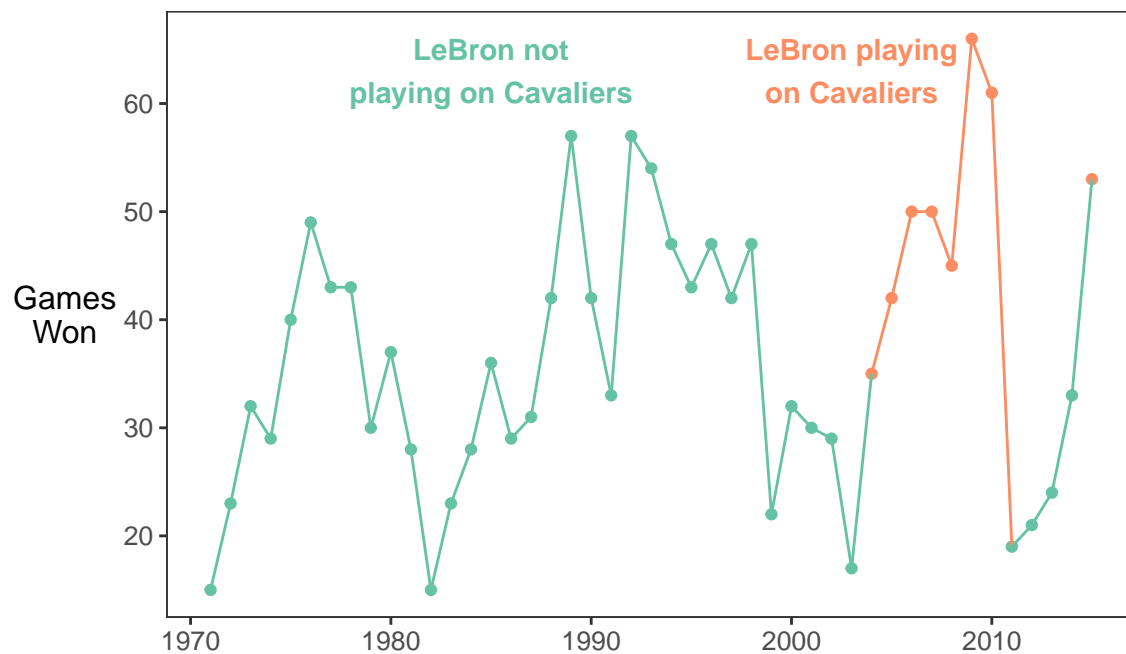
Chicago Bulls Performance with/without Michael Jordan

1967–2015



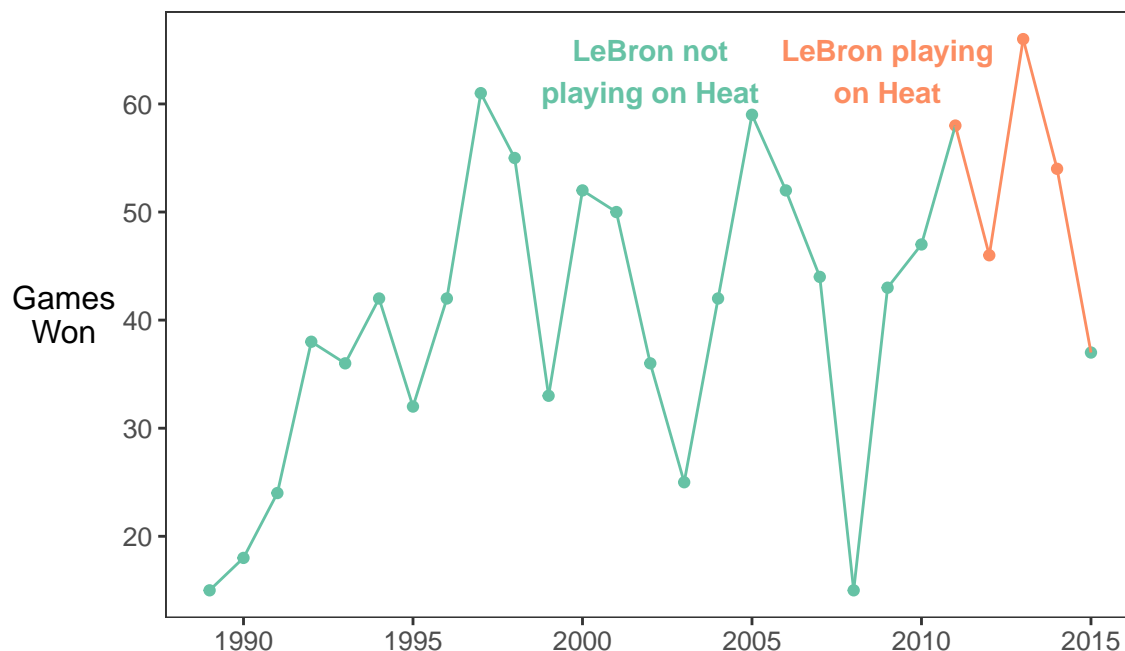
Cleveland Cavaliers Performance with/without LeBron James

1970–2015



Miami Heat Performance with/without LeBron James

1989–2015



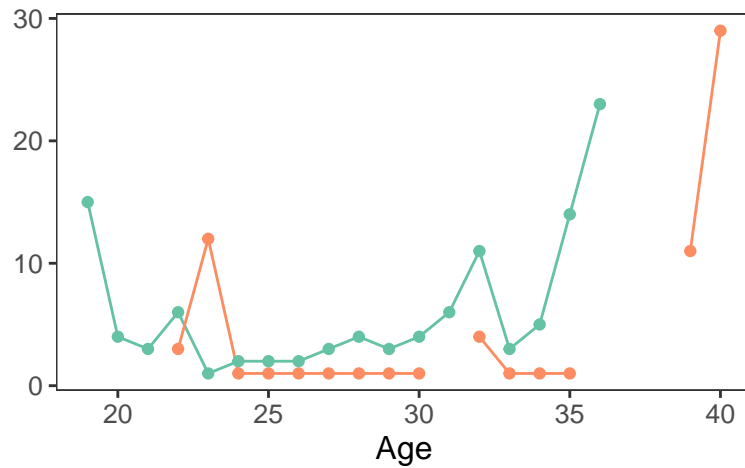
The effect of both Jordan on the Bulls and LeBron on the Cavaliers is significant and clearly shown by these charts. Both teams were subpar, and the signing and emergence of these two stars steadily improved their teams until they were dominant. It can also be seen that when these players left their teams, the dropoff in the teams' performance is huge. The one exception is LeBron and the Heat. While the team does perform better, it is much less of a stark difference. This is partially because the Heat were a decent team before LeBron joined the team, so the improvement is less dramatic. If teams success was a major part of greatness, there would be no clear winner. Both players made their teams much better.

Graph 4: Players' Rank in Stats

The final analysis is players' rank in stats compared to the rest of the players that season. This will provide information as to how dominant a player was in their respective league. The drawback of this analysis is that it leaves out the possibility that the league just became more competitive and talented as time has gone on.

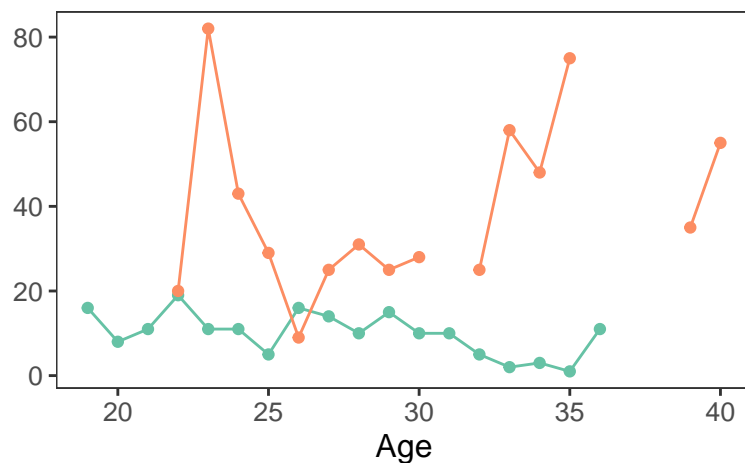
Scoring Rank Over Time

LeBron James vs Michael Jordan



Assist Rank Over Time

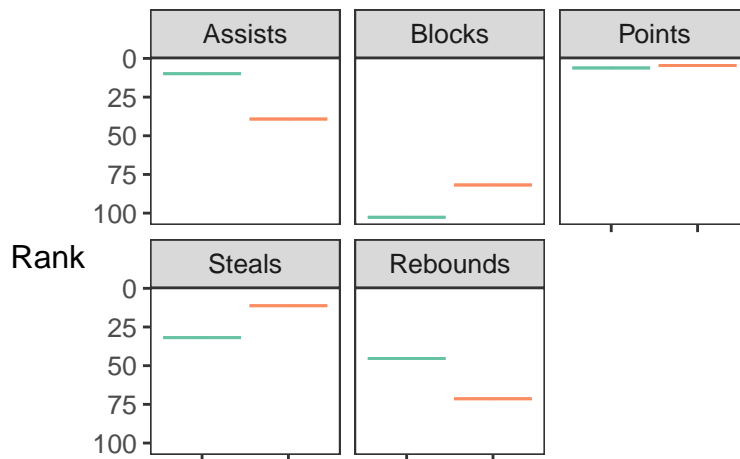
LeBron James vs Michael Jordan



From these charts, we see the same general trend as from the per game stats over time charts. Michael Jordan is a better scorer and LeBron is a better assister. However, this chart illuminates just how dominant Jordan was in scoring. He was ranked number one in scoring for seven years straight. He then retired for one year, was not ranked first for one year, and then he took back first place for three more years. If greatness is mainly a measure for how dominant a player was in their respective league, then Jordan probably wins. He completely dominated during his era, whereas LeBron is still dominant but by smaller margins.

Career Average of Ranks by Stat

LeBron James vs Michael Jordan



The final chart sums up some of the main trends we have been seeing throughout this analysis. It plots average rank for each stat for both LeBron and Jordan. LeBron is more versatile, having a high rank in assists, as well as beating out Jordan in rebounds. The two are close in average points rank. Jordan does have better steal ranks, but steals are more of a niche stat than either assists or rebounds.

Discussion

Conclusion

This paper does not come to a final conclusion about which player is actually greater. This is partially because of the narrow part of the data this project looked at, and it is partially because the answer to this question mainly depends on the definition used for “greatness”. If greatness is mainly about dominating the league that a player played in, Michael Jordan probably wins. However, if greatness includes versatility and consistency over many years, LeBron probably beats out Jordan. To come to more complete conclusion and reveal more of the nuances, more analysis would have to be done.

Limitations

There were many limitations to this analysis. First, this analysis only focused on the regular season. Many of the more important parts of greatness according to some come during the playoffs. NBA Championships, playoff stats, and Finals appearances were all left out of the analysis. This is by design, as this was the initial analysis to be done. In addition, season awards were not analyzed. MVP titles, scoring titles, and all-NBA teams are typically part of the discussion, but these awards were too broad for this narrow focused analysis. That being said, this analysis allows for the discussion to begin, and it outlines many of the general trends that would be seen when looking at the other data. The regular season is an important part of an NBA players’ career, as it makes up the bulk of their time spent on the game.

Future Analyses

The opportunity for future analyses is great. First, looking at the other data that was left out of this analysis, like playoff stats and regular season awards, would be a good start. In addition, looking at more players could be useful. While many solely consider LeBron and Jordan to be the only two players in the GOAT debate, some commentators consider other players to be in the discussion. Finally, broadening the research question could be a next step. The greatest of all time debate is a topical and interesting one, but it is certainly not the only use for NBA stats.

References

Shiny: https://sethrkeim.shinyapps.io/NBA_Stats/?_ga=2.198771496.664884060.1612111709-732976651.1611176855

Kaggle: https://www.kaggle.com/drgilermo/nba-players-stats?select=Seasons_Stats.csv

nbastatR: <https://github.com/abresler/nbastatR>

Basketball-Reference: <https://www.basketball-reference.com/>