Introduction (Robel)**:** Describe your project scenario. Starting out, what did you hope to accomplish/learn?

Our project centered around examining ESPN NBA player data from the 2001-2002 season to the 2019-2020 season. More specifically, the data captures the top 50 scorers in each of those seasons and includes additional statistical categories like rebounds, assists, etc. From the data, we wanted to identify various historical trends to see how the game of basketball has changed over time within the NBA. We know that the NBA has changed drastically over the past couple decades – there has been more reliance on the 3-point shot, less reliance with the center position, and the top players have evolved. Our objective was used to take the various attributes in the data set to tell a story on what has happened in the NBA the past couple decades. In pursuit of this, part of our goal was to use the data to answer heavily debated topics amongst avid NBA fans. These questions included things like "As players like Steph Curry took over the league, how has the 3-point shot changed over time by position?" and "Should LeBron James have more MVPs?" We also analyzed interesting questions like "How do different players rank in each of the statistical categories?" and "Based on certain stats, which players would we want to select in a Top 5 starting lineup?" From the answers to these questions, we wanted to see how well our analysis matched the reality of the NBA. We wanted to see if the data set that we are using could help to answer heavily debated topics in the NBA and capture the ebbs and flows of the NBA over time.

The Data (Manpreet)**:**

This data set shows the Top 50 NBA players based on Points per Game (PPG) from 2002 to 2020 from ESPN.com. We chose to use this data set because it would allow us to see many statistics of the Top 50 players throughout several seasons such as position, minutes played, points made, three point fields made, rebounds, assists, and several more. This allows us to analyze how trends have potentially evolved over the years, how player rankings can be impacted by different statistics, and much more.

We obtained this data by webscraping ESPN.com with Beautiful Soup. Because it was for several years rather than one year, this involved a for loop to webscrape for the different web

pages for each year. Then, several for loops were used to create appropriate lists of the different players in each season along with their specific statistics.

Experimental Design (Everyone): Describe briefly your process, starting from where you obtained your data all the way to means of obtaining results/output.

Manpreet: In order to understand if there was any correlation between the number of three-point field goals made by position and the number of players in that position, a series was first created by grouping by Season and Position then finding the average three-point field goals. So, this series contained the average three-point field goals for each season and each position. To make it clearer, the positions were separated, so there are now seven series (one for each of the seven positions) which show the average three-point field goals by Season. Then, a similar process was used to find the number of players in each position by creating a dataframe that grouped Season and Position then found the number of players using the unstack function. This was then separated by each specific position, as well, so that eventually plots could be created for each position. Finally, 14 plots were created to show for each position, the average 3-point field goals over the seasons as well as the number of players in each position over the seasons.

Amber: To acquire information on any player that has made the dataset throughout the years, user input was implemented to ask for a player of interest. Based on the player name, a data frame was created with all the columns the player appears in the dataset. If the length of this data frame is less than one, then the code tells the user there is no player by that name. If the length of this data frame is one or greater, that player's first year, first position, last year, and last position in the top 50 list are indexed. The number of times the player makes the top 50, average number of games played, and average number of points scored per game are also calculated from this data frame. All this information is put together in a user friendly print statement to provide the statistics about the indicated player of choice.

In order to acquire a holistic list of the top all around players in the NBA, a ranking system was implemented to compare across the different categories specified in the dataset. The players statistics were grouped by their name, averaging across all the times they appeared in the dataset so they only have one value for each category. The categories that were selected and deemed important statistics when considering the best all-around player were: points per game, field goal

percentage, 3-point field goal percentage, free throw percentage, rebounds per game, assists per game, steals per game, and blocks per game. Within each of these columns, the player's were ranked from 1 to 254 based on their statistics for that category. This means, for example, a player could be ranked number 1 for points per game, but ranked number 254 for free throw percentage. These ranked values were then averaged across all categories per player to give them an all-around average rank score (lower average indicates higher all-around rank). The output displays a list of the top 5 players that were calculated based on this ranking system within these categories to be the top all-around players for this dataset.

Because this code averaged across all the seasons the player appeared in the dataset, additional code was written to list who the top all-around players were for each season. A for loop was created to rank the player's statistics for each individual year of the dataset. The categories that were selected and deemed important statistics when considering the best all-around player were: points per game, field goal percentage, 3-point field goal percentage, free throw percentage, rebounds per game, assists per game, steals per game, and blocks per game. Within each of these columns, the player's value was ranked from 1 to 50 based on their statistic for that category. These ranked values were then averaged across all categories per player to give them an all-around average rank score. The top 5 overall average ranked players for each year were added to a dataframe to display the top 5 players within every season. To inquire about a specific season, user input was implemented where if the user inputs a valid year within the dataset, the code will display the top 5 player's names in order for that season based on the indicated ranking system.

Robel: In NBA circles, an ongoing debate is determining who the top ranked players in the NBA are as well as who is best in a statistical category (points, assists, rebounds, etc.). From our data set, we wanted to see if we could determine an elite NBA starting lineup. There are many attributes that we could examine to determine this Top-5 so we decided to base our starting lineup on points, rebounds, assists, blocks, and Player Efficiency Rating (PER). We decided on using points, rebounds, and assists because these are the three major statistical categories that are calculated for each NBA player. We decided to use blocks to incorporate an attribute that is more reflective of the defensive nature of players. The last element that we used is PER which is a unique statistical category that captures all of a player’s attributes into a numerical value. We determined that these categories would create a starting lineup that would be composed of some of the top NBA players in the last couple decades. These attributes also allow us to develop a starting lineup that allows us to capture various essential skill sets to NBA basketball. We also

developed box plots for each statistical category to visually see the distribution for each statistic in our data set.

To determine the appropriate players to place in our Top-5 starting lineup, we had to develop four different loops for each statistical category we observed. These loops found the name of the player, position, season, and statistical value for the player with the best attribute in the observed statistical category. We first had to find the maximum (best) value in each statistical category. From there, we were able to develop an index to find that value's associated player, player position, and season. We did this for each category that we wanted to analyze and put it together to develop our Top-5 starting lineup.

Robel: Another interesting question that we wanted to see was if LeBron deserved to win more than the 4 MVPs he has already been awarded. This question came out of a widely debated topic in the NBA community that LeBron has lost MVPs due to media bias, "politics of the NBA, etc. Many players and fans believe that LeBron has been snubbed of multiple MVP awards because it is widely agreed that he has been the best NBA player in the league for over a decade but only has 4 MVP awards. From this discussion topic, we wanted to compare the statistics of the MVPs of each season in our data set and compare them with Lebron's statistics of that year. We did not include any of the seasons that LeBron had been named the league's MVP.

To perform this analysis, we researched the players that have won the MVP award in each season captured in our data set. We faced no issues with an MVP not being in the top 50 list of the season that they won. We can affirm this because points per game is probably the most important statistic tracked in the NBA and would lead to the conclusion that the MVP would be a top scorer in the league. We placed these MVP players in a list and looped through each player in the MVP list and grabbed their statistics for the year they won the MVP award. Through the use of aggregation and column filtering, we were able to collect the statistics for each player in question for each season. We also created two separate data frames where one stored only the information relevant to that season's MVP and another data frame that stored only the information for LeBron James in the observed season. We selected specific statistical categories to compare between the MVP and LeBron James which included points, assists, rebounds, steals, blocks, and PER. We selected these attributes because we believed that these would be the best attributes in making one-on-one comparisons between the MVP and LeBron. For each comparison, we also

developed a bar chart that visually demonstrated the differences between the MVP and LeBron for each observed statistical category.

Beyond the original specifications (Amber and Robel)**:** Highlight clearly what things you did that went beyond the original specifications. That is, discuss what you implemented that would count toward the extra-credit portion of this project (see section below).

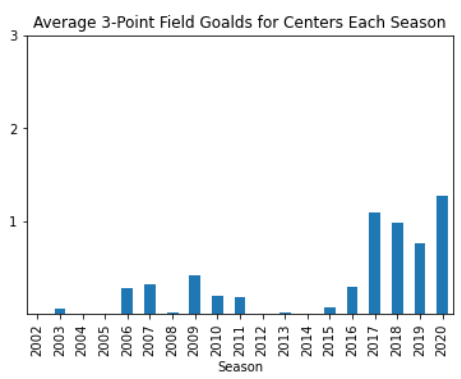
Amber: An additional aspect we added to this project is the use of user input. First, if an individual wanted to know basic information about a specific player, they can input the name of the player and the code will provide information on how many times they were ranked in the top 50, when they first and last made the list, and the average number of games played and points per game if they are on the list. Second, if an individual wanted to know what the top all-around players are for any year of the dataset, they can input the year they want to see and the code will provide a list of the top 5 players based on the specified ranking system along with their city and position.

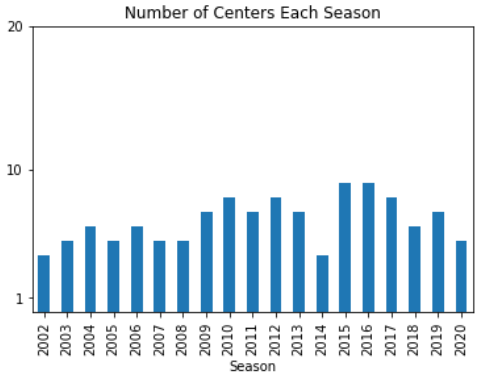
Robel: For this project, we took an extra step with our project to incorporate some web scraping within our Python script. The web scraping portion was based off the earlier web scraping assignment that we had earlier in the semester. We expanded this code to capture data over a longer period of time than we had in our previous assignment. We used the BeautifulSoup library to scrape the data from the ESPN statistics website and output that data into a suitable CSV format. The data on the ESPN website was set up in a way that made it convenient to throw each type of attribute into a list in our Python script and outlay that into a column within the CSV file. We also created an extra column of data to capture the year that each player's statistics were collected for. For each web page that we were collecting data from, we had to create a for loop that went through each web page and scraped all the necessary attributes and then reset the loop so that we could perform the same web scraping on the next page. The web scraping component allowed us to collect all the necessary data we needed to perform our analysis to derive

information and knowledge in support of our project objectives. From the web scraping, we had to code in some formatting to be able to collect the data that we required and insert it into the CSV file.

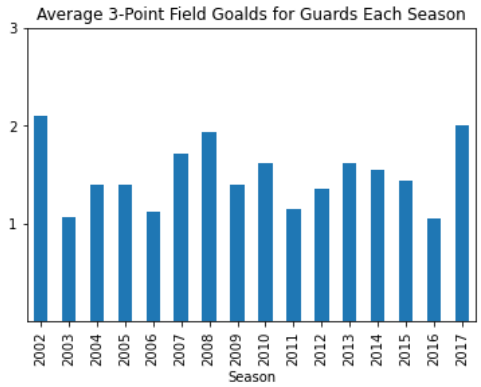
Results (Everyone)**:** Display and discuss the results. Describe what you have learned and mention the relevance/significance of the results you have obtained.

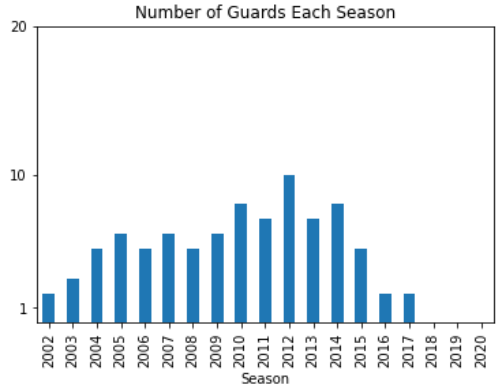
Manpreet: Overall, the average three-point field goals has increased since 2015 for Centers; however, the number of centers from 2015 has actually decreased.



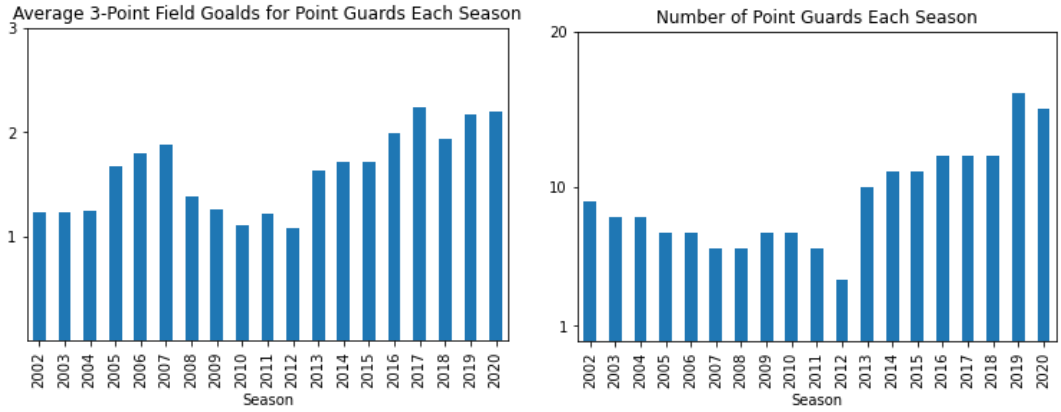


For Guards, there is not a specific trend one can see regarding the 3-point field goals or number of players.

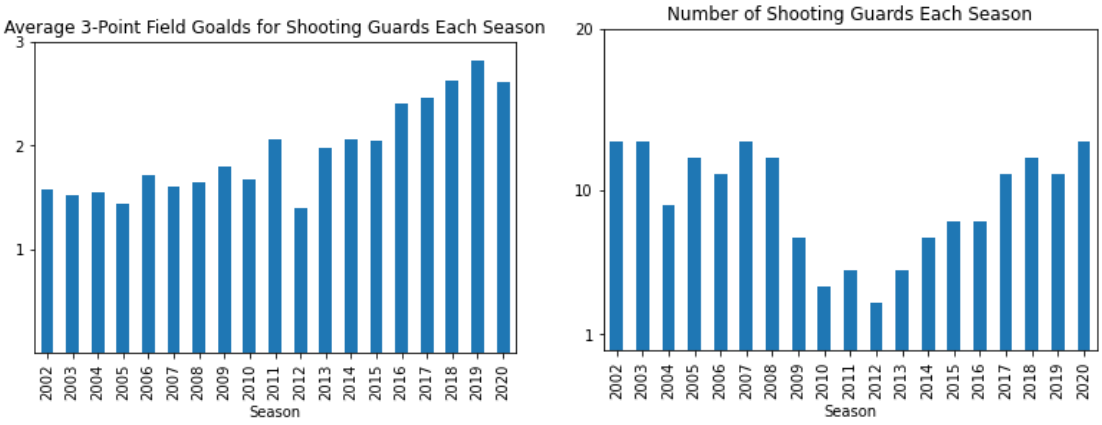




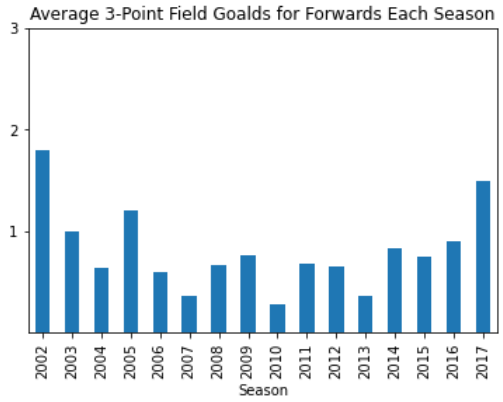
For Point Guards, it seems very clear that the number of three-point field goals has increased steadily from 2012, and the number of Point Guard players has also steadily increased since 2012. So, it does seem that for Point Guards, there may be more of a reliance on 3-point field goals.

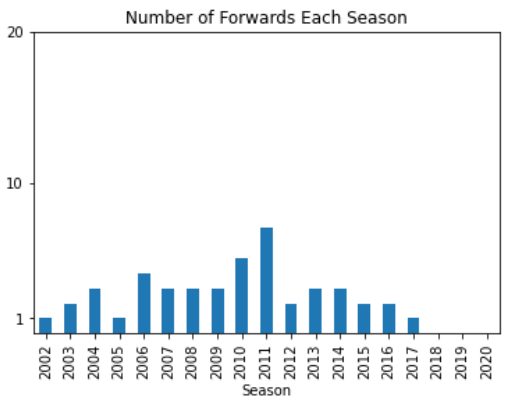


The same steady increase can be seen for Shooting Guards for the three-point field goals as well as the number of Shooting Guards. Therefore, there may also be some kind of correlation for Shooting Guards, as well.

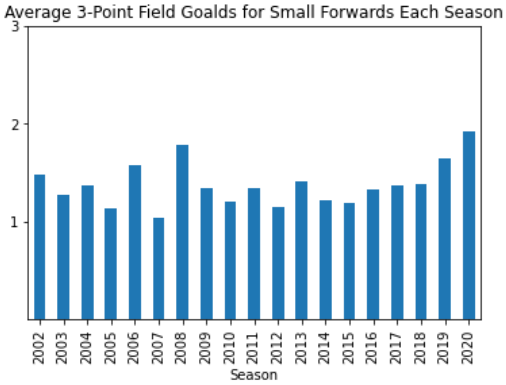


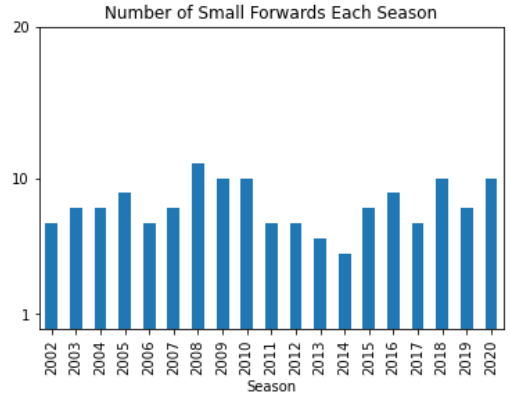
For Forwards, there is not much of a trend for three-point field goals.



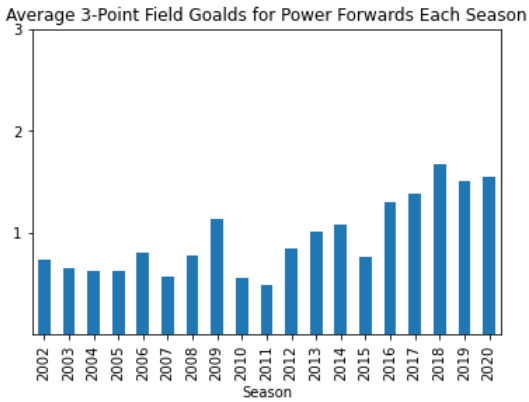


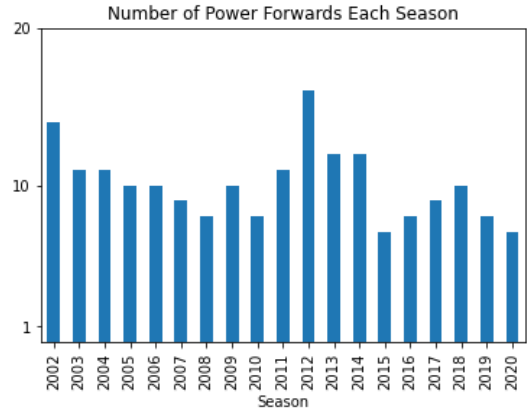
Small Forwards have started to increase the three-point field goals from 2016, but it does not seem that there is much of a correlation to the number of players.





Finally, for Power Forwards, it seems that it follows the same trend as Small Forwards.





Amber: The specified ranking system indicated the top 5 all-around players in the NBA based on their average values through every season they were on this list includes, in order: Kevin Durant, Kawhi Leonard, Stephen Curry, James Harden, and LeBron James. The top all-around players per season can be displayed per user input, but the top all-around players in the 2020 season include in order: Kawhi Leonard, James Harden, Damian Lillard, Anthony Davis, and Nikola Jovic.

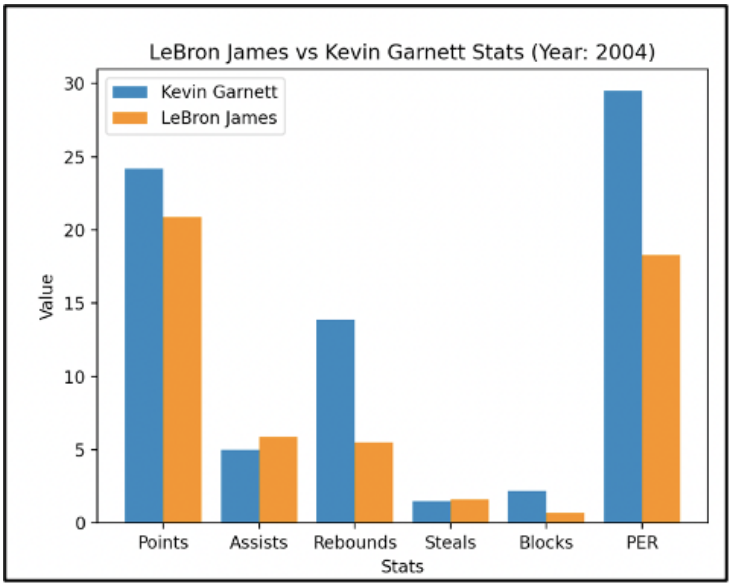
Robel: For the five attributes we focused on for the Top-5, we analyzed the data set to determine which player had the highest value in each category based on the data set's time span. As mentioned previously, we looked at the categories for points per game, rebounds per game, assists per game, blocks per game, and Player Efficiency Rating. We developed visualizations (box plots) to see the distribution between each of the statistical categories. From the visualizations, we can see the range of values for each statistical category as well as the value of the maximum value). These visualizations provide insight into the minimum, median, and maximum values of each category as well. We can see from the visualizations that blocks per game has the tightest range of values and points per game has the widest range of values. Since we are determining the best player in each category, we know that the data point of most interest to us is the maximum value (i.e. the value at the top) of each box plot.

For points, we found that the player that averaged the most points per game in a season was James Harden (a Shooting Guard) by averaging 36.1 points during the 2018-2019 season. For rebounds, we determined that Andre Drummond (a Center) averaged the most rebounds per game in our data set during the 2018-2019 season at 15.1 rebounds per game. For assists, Steve Nash (a Point Guard) averaged the most assists per game (at 11.6 assists per game) during the 2006-2007 season. Tim Duncan (a Center) averaged the most blocks per game during the 2002-2003 season at 2.9 blocks per game. For PER, Giannis Antetokounmpo (a Power Forward) had the highest PER in the data set of 31.94 during the 2019-2020 season. From the attributes that we observed, these five players would be an ideal starting 5 lineup. We believe that this is an ideal starting lineup because four of the five people in our starting lineup have won an MVP (Andre Drummond is the only one who has not won an MVP) and three of those MVPs have won more than once. We also have decent coverage on the different positions and skill sets that these players have which is needed to succeed from an NBA team.

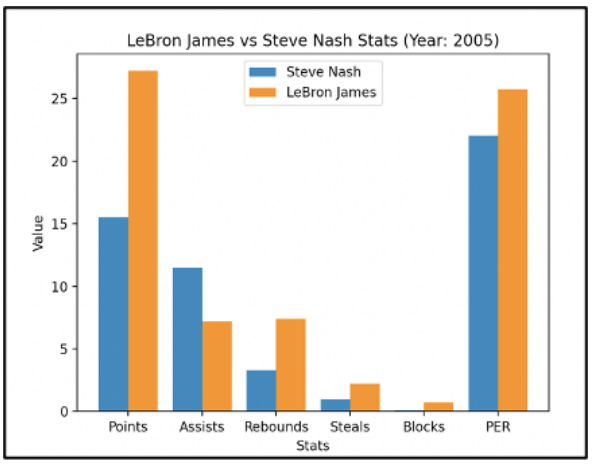
The relevance of this analysis is that it provides insight to avid NBA fans on some of the elite NBA players over the past couple decades. The game of basketball requires an assortment of different skills and abilities and this analysis shows some of the top NBA athletes in terms of offensive and defensive ability. As mentioned, we know that many of the players on this list were former MVPs – the analysis we conducted provides some evidence on why these players won an MVP. They dominated in a specific category and it is probably a safe assumption that this translated them doing well in terms of other statistical categories. From this analysis, we were able to determine all the top scorers who were the best scorer in terms of points per game. In addition, we were able to see how these scorers did in other categories (since the dataset encompassed the top 50 scorers in each season). This information may be useful for an NBA GM that is looking to gain some star talent for their team. They can use this analysis to coincide with the needs of their team and determine the bar of who is the best for these different statistical categories. This information is also useful for avid NBA fans who want to know the players that have been the best at these specific statistical categories over the last couple decades too.

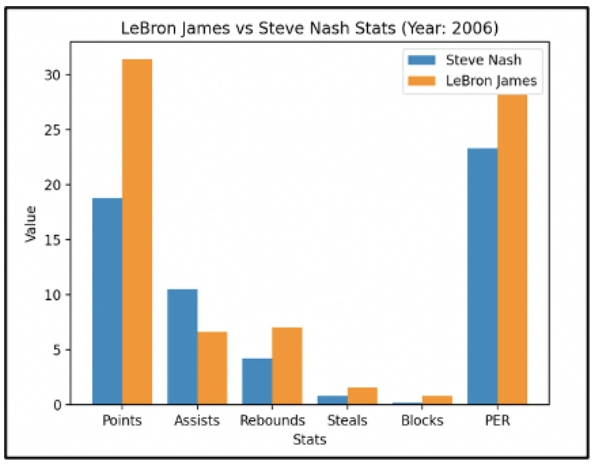
Robel: As we look at the analysis, it is important to note that only a few statistics were observed in these comparisons between the league MVP and LeBron James. There are other factors, outside the scope of this analysis, that also drive the winner of the MVP in each season (i.e. team performance and other statistical categories) From the results, we saw that there were some seasons in which LeBron James could have potentially been snubbed for an MVP award.

In the 2003-2004 season, LeBron James entered the league as a rookie and, during this season, Kevin Garnett won the league MVP. We can see that Kevin dominated a majority of the observed statistical categories and that his PER was much higher than LeBron James' PER that year. We can say that LeBron James, despite his incredible statistics did not get snubbed during his rookie season for the MVP award.

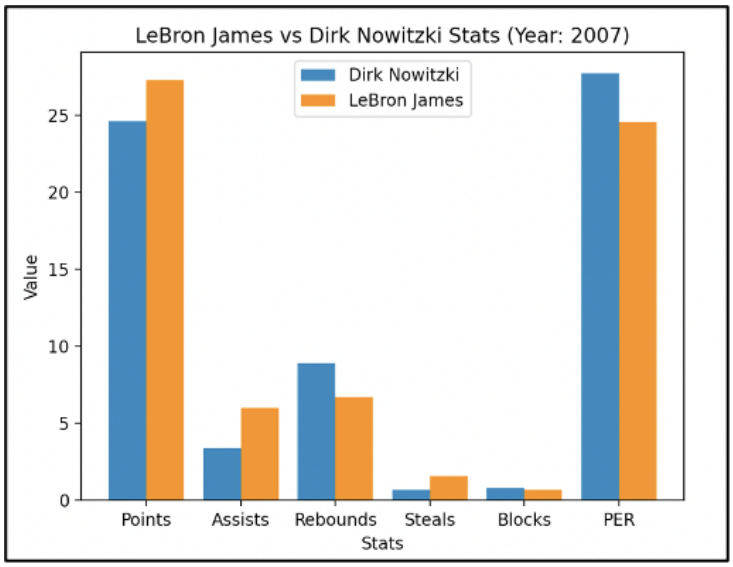


In the 2004-2005 and 2005-2006 season, Steve Nash was awarded the league MVP in back-to-back years. Looking at the statistical categories (points, assists, rebounds, steals, blocks, and PER) in both seasons, Steve Nash only did better in the assist category. In every other category, LeBron James' statistics were better than Steve Nash's statistics. We can see that LeBron's PER was also much higher than Steve Nash's PER in both seasons. This probably indicates that LeBron did better in other statistical categories as well. From the bar plots below, we determined that LeBron James had overall better seasons in 2005 and 2006 and could have won the league MVP in that year.



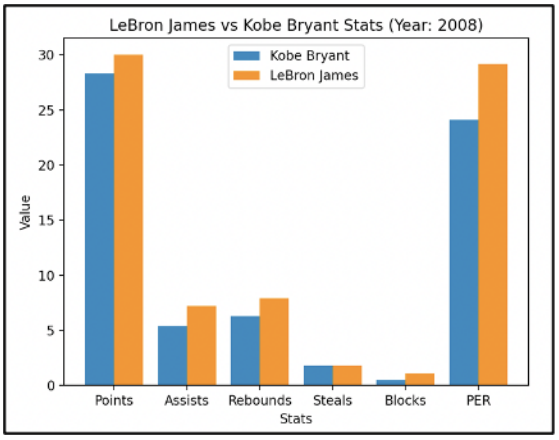


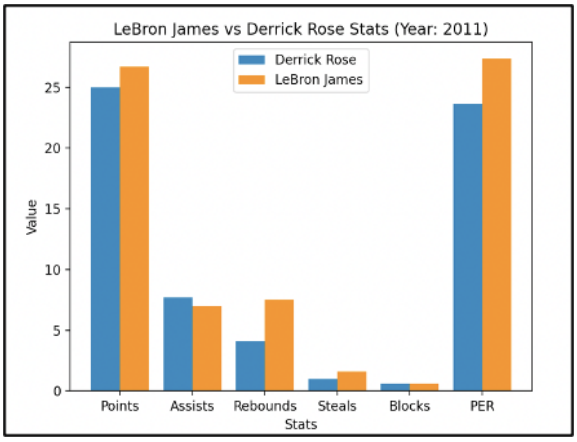
In the 2006-2007 season, Dirk Nowitzki won the NBA MVP. Looking at the statistics between Dirk Nowitzki and LeBron James, we can see that LeBron probably did not get snubbed for the MVP in that year (at least compared to other years). This is further evidenced by Dirk's PER rating. The higher PER rating by Dirk most likely indicates that there were other statistical attributes that influenced him winning the award over LeBron and others. From the bar chart below, we see that Dirk and LeBron had pretty similar statistics in these categories.



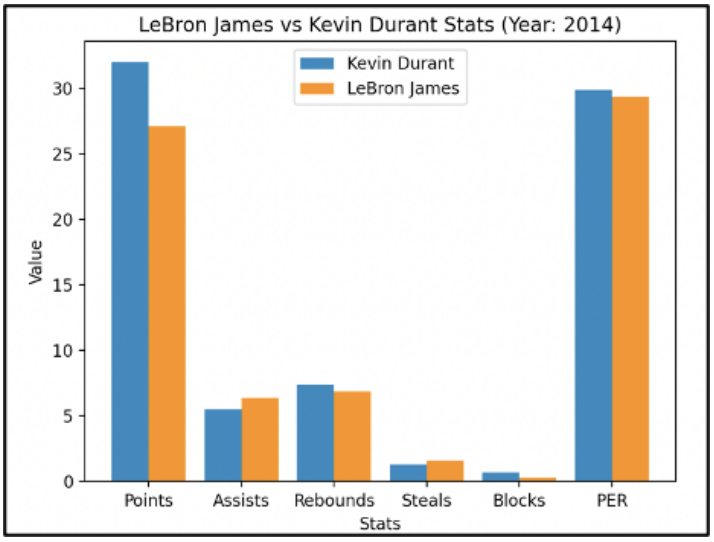
In 2008, Kobe Bryant won the league MVP, and, in 2011, Derrick Rose won the MVP award. Both of these years are potential two seasons in which we could determine that LeBron James was snubbed for an NBA award. The reason is that LeBron, for both seasons, had a stronger showing in a majority of the observed statistical categories. In 2008, the only statistic that LeBron James did not do better than Kobe Bryant in was for steals (where they averaged the same steals per game) and, in 2009, Derrick Rose did better in assists and averaged the same of blocks as James. Otherwise, it would seem that LeBron had a strong case to win the MVP in both of those seasons. From this analysis, we would determine that LeBron James may have been a more viable winner for MVP than Kobe Bryant or Derrick Rose based on these statistics. Again, we also see that LeBron's PER was much higher than the other two athletes which indicates that there are probably other measures that LeBron did better in as well. One note is that LeBron

James won his in 2009, 2010, 2012, and 2013 so we do not provide the statistics for those years in this project.

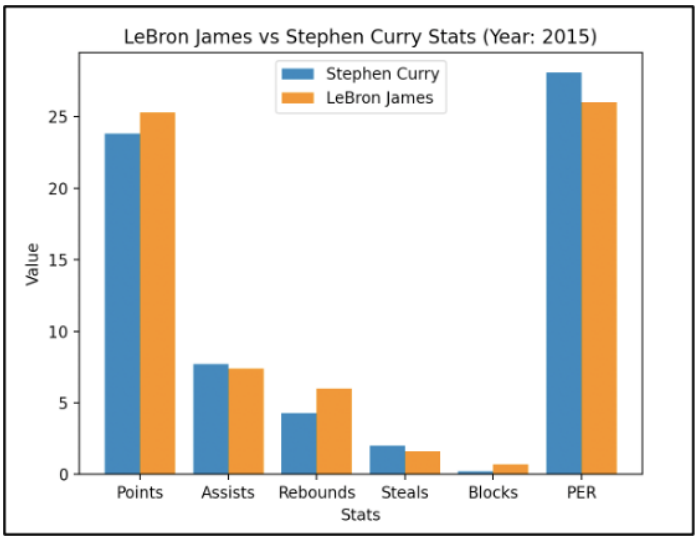




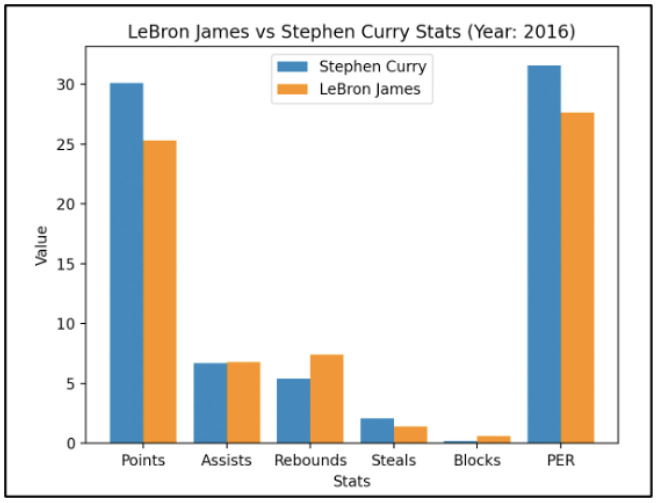
In 2014, Kevin Durant won the MVP award and, from looking at our analysis, it does not seem that LeBron was snubbed in that year. For most categories, Kevin Durant and LeBron James seemed to be pretty close in their statistics for the 2013-2014 league though Kevin Durant did seem to have the edge in 4 of the 6 categories (with a big lead in points per game). We would determine that the 2013-2014 season was not a season in which LeBron was snubbed.



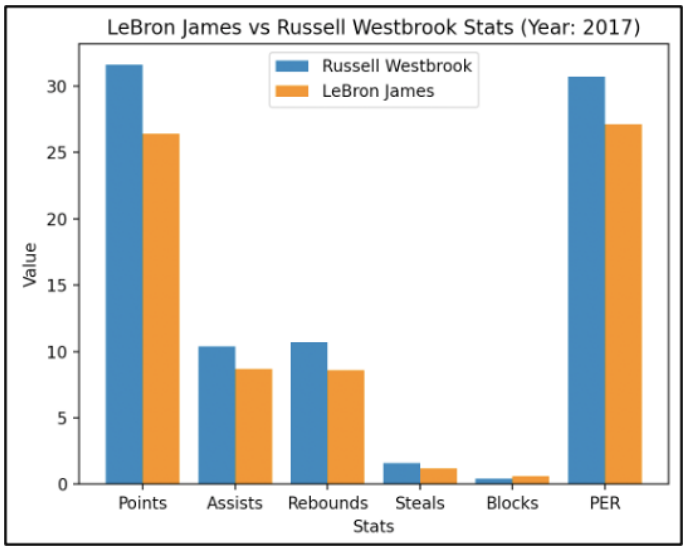
In the 2014-2015 season, Stephen Curry won his first league MVP award. From our analysis, we see that Stephen Curry and LeBron James had similar statistics in the categories under observation. We see that Stephen's PER was a bit higher than LeBron's which might indicate to us that there are other statistical categories we are not taking into account that might help explain the leverage Stephen had over LeBron for MVP that season. We also know, from our own knowledge of the NBA, that this was the time where Stephen took over the culture of the NBA by increasing the reliance on the 3-point shot.



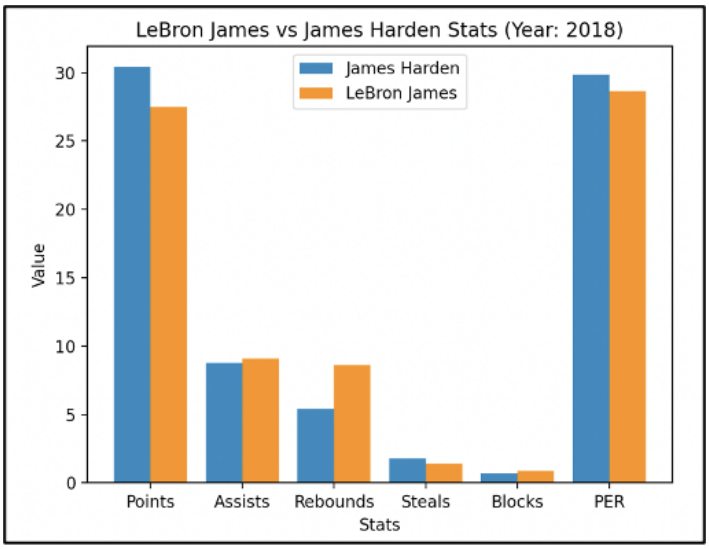
In the 2015-2016 season, Stephen Curry statistics were even better than when he won the MVP in the previous season while LeBron's James more or less stayed the same. It is also important to note that this was the season that Stephen Curry was the unanimous MVP, becoming the first MVP to accomplish this milestone. We determined then that this was not a year where LeBron was snubbed for an MVP award.



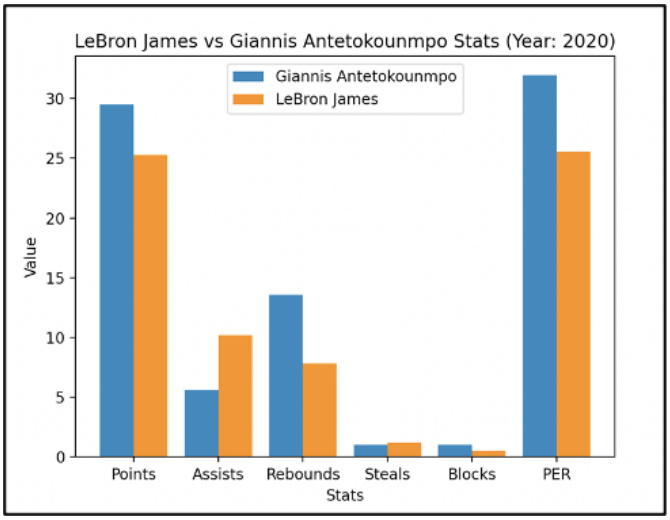
In the 2016-2017 season, Russell Westbrook took home the hardware for league MVP. This season is probably the clearest picture to demonstrate that LeBron did not deserve the MVP in that year. The only statistic that LeBron did better than Russell (out of the observed categories) was blocks (and only by about 0.2 blocks per game more). We can say with confidence that LeBron did not get snubbed for an MVP award in this season. We can further cross-check this because Russell Westbrook became only the second player in league history to average a triple-double (in points, assists, and rebounds) which is an extremely difficult thing to do a couple times a season, let alone for the entirety of a season.



In the 2017-2018 season, James Harden won the MVP award in that year. We see from the bar plot below that, in most statistical categories, James and LeBron had similar statistics. We would determine that LeBron did not get snubbed for an MVP award in this season for a couple reasons. As we saw in our "Top-5 Starting Lineup" analysis, James Harden averaged the most points per game in a season in the past couple decades (and probably ever) and we have already identified points per game as probably the most important statistic when comparing players. Also, we see that James Harden's PER outperforms Lebron James' PER. This leads us to believe that there may have been other factors that improved James Harden's chances to win MVP.



For the 2018-2019 season, LeBron James did not play enough games (70%) in the regular season to make the list so we do not show any analysis for this season. In the 2019-2020 season, Giannis Antetokounmpo won his second MVP award (he was also the winner in the 2018-2019 season). We see from the visualization below that we can determine that LeBron did not appear to be snubbed for the award. Giannis performed better in most of the observed statistical categories and his PER was much larger than Lebron's We should also point out that Giannis was also in our Top-5 player analysis for having the highest PER in our data set. From this, we determine that LeBron was not snubbed for an MVP during that season.



It is important that the statistical categories under observation are essential, but not encompassing, of all the factors that go into the MVP decision. Our analysis was based off each player's statistics and cross-checked from our own personal understanding of the NBA. With additional data and more research, we could provide an even more in-depth analysis on how each MVP won over LeBron James. It is also important to note that for many of these seasons that there were other players under consideration for MVP that did not win, and that LeBron ended up in the top tier of voting for many of the seasons he lost.

Testing (Everyone)**:** Describe what testing you did. Describe the unit tests that you wrote. Show a sample run of 1 or 2 of your tests (screen captures or copy-and-paste is fine).

Robel: For our Top-5 starting lineup analysis, we wanted to confirm that we were capturing the maximum values for each category of interest. We used these maximum values as a way to set the index in the data set so that we could grab other attributes related to the maximum value (i.e. player name and position). If we didn't find the correct max value, then the rest of our analysis would be incorrect based on those observed statistical categories. For our testing, we used the

“unittest” library to test this out and applied a glass-box/white-box testing so that we could use our code to create our tests. We called our main project code and imported the unittest.TestCase method to do our unit testing. We used assertEqual() methods to compare the player we expected with the output of our code. We were able to find the maximum values for our categories of interest and compare those with the output that our code was getting. We incorporated assertEqual() methods to compare the maximum values with the code output.

Continuing with testing the Top-5 starting lineup script, we ran similar unit tests to determine that the appropriate players were put into the starting lineup. We used the “unittest” library for our unit testing and used multiple assertEqual() statements to compare our code’s output with the correct result. We developed five separate unit tests to find the five players for our starting lineup and confirm that our code was collecting the appropriate members. From the analysis, we wanted the players with the highest points per game, rebounds per game, assists per game, blocks per game, and PER (if there was a tie, go with the player that appeared first in the data set). From the data set, we were able to determine the players that should be in the list – James Harden (points), Andre Drummond (rebounds), Steve Nash (assists), Tim Duncan (blocks), and Giannis Antetokounmpo (PER).

For our comparison of LeBron James vs the MVPs of season’s in which he did not win, we employed a similar testing strategy that we did for our Top-5 starting lineup unit testing. We wanted to confirm that we were picking up the appropriate players to compare with LeBron in each year. We researched online which players won the MVPs and collected them into a list. Based on the order of the MVPs, we knew where they would be positioned in the list. Our code collected each MVP in our script into a list and then we compared that list with the player we expected to be at a certain index. We used assertEqual() methods to make these comparisons and test that feature of our code out. From there, we were able to confirm that we were comparing the right players to LeBron in the appropriate seasons.

Amber: To unit test the ranking system, first we wanted to see if the code was sorting the data set correctly. This was performed by grouping the data by name to get the mean values of each statistical category and then being able to sort based on each category. The category “Points Per Game” was used as an example to unit test. The expected top player for mean points per game throughout all the seasons was Allen Iverson. Once the list was grouped and sorted, it returned Allen Iverson as the first name and we can assert that these are equal and the code is sorting properly.

Next, we wanted to unit test the ranking system itself. This was performed by grouping the data by name to get the mean values of each statistical category. These values were then sorted from highest to lowest and given a ranked value from 1 to 254 (there are 254 unique players in the dataset). Therefore, the player with the best statistic for that category should be ranked 1. The category “Field Goal Percentage” was used as an example to unit test. The expected top player for mean field goal percentage was Clint Capella. Once the list was grouped, sorted, and ranked, it returned Clint Capella as the first name and we can assert that these are equal and the ranking system is ranking the categories properly.

Conclusions (Amber)**:** Summarize your findings, explain how these results could be used by others (if applicable), and describe ways you could improve your program. You could describe ways you might like to expand the functionality of your program if given more time.
Amber: These results could be used by others who want to look at historical trends within the NBA. It could be important for coaches or sports analysts to understand quantitatively how the NBA has changed over the years to maximize their player and team efficiency. To improve the program, more unit testing could be performed on the data to ensure adequate responses. There are also infinite additional queries that could be performed on this dataset to analyze how the NBA has changed over time. A way to expand on our program could be to predict performance of players in future seasons. We could predict which players might be on the list of highest ranked players or how teams may distribute the player's positions.