2021-2022 NBA Player Statistics

Mitra Patel* Northeastern Student
Alvin Kannalath† Northeastern Student
Thomas Mcdonagh ‡ Northeastern Student
Nicholas Perrotta§ Northeastern Student

ABSTRACT

The purpose of our project is to create an NBA player visualization tool that will allow coaches and general managers a way to quickly and easily compare the pros and cons of NBA players so they can find the perfect player to fit their roster. One domain task we are looking to implement is a way that allows users to compare player performance across positions. Another task we want to add to our tool is a way to get a sense of player performance across all stat lines.

1 INTRODUCTION

The domain tasks our project will support are vital in analyzing player performance in sports. Our target end-users include coaches, scouts, and general managers who require quick and easy access to player performance data that will allow them to consider new players to trade for or sign in for free agency. The data we will be visualizing includes various statistics related to player performance from the 2021-2022 NBA season, such as points, rebounds, assists, and field goal percentages collected from a website called basketball reference. The first domain task we want to add will allow users to compare player performance across positions, with specific filters for each player's position and statistics that can be chosen from since the effectiveness of some positions is determined by certain stats. This domain task is relevant, as it allows users to compare players on statistics that matter for a specific play style a user could look to find. The second domain task we will implement will allow users to see the entire stat line for a player of interest. This is important for our end users, as it allows them to get a sense of what each player succeeds in and struggles in, so they can get a holistic sense of what their play style is all about, and make a judgment call on whether to pursue them or not.

2 RELATED WORK

2.0.1 Peer Reviewed Paper 1

The first peer review paper analyzes the team statistics for NBA teams that were deemed significant and thus separated winning and losing teams. The paper's research and visualization analyzed regular season data to notice trends between winning and losing teams. The findings of this paper were that more defensive rebounding teams win more games, and more offensive rebounding teams lose more games, as it implies they are missing more shots. There were many visualizations in this peer-reviewed paper. One particular visualization that is relevant to our project was a bar graph comparing the bottom three teams, top three teams, and league-wide average for average defensive rebounds. The visualization clearly and effectively shows the correlation between the winning and losing teams concerning the average defensive rebounds. We can apply a similar visual to our project by looking at individual player stats, such as average points per game. This visual will allows us to compare any given player to the average performances of their position and the league as a whole to determine if they are better, worse, or on par with their colleagues. [2]

2.0.2 Peer Reviewed Paper 2

This paper discusses the recent trend in the NBA of traditional positions no longer being entirely applicable to assembling successful basketball teams. This paper's visualization purpose was to identify to coaches and general managers different players based on their playing styles rather than their position on paper to assemble the most effective teams that can output better winning records. For example, one particular visualization clustered different players into the following categories and displayed how many fell in each cluster via a bar chart: high usage guard, stretch forward, three-point shooting guard, traditional center, versatile role player, floor general, mid-range center, skilled forward, and ball-dominant scorer. Parts of this visualization can be utilized in our project. By creating a scatter plot with the y-axis measuring a player's performance, we can view how players are clustered by the chosen x-axis statistic. A position filter can be implemented to allow the user to look at clusters by position instead of league-wide clusters. [3]

3 USE CASE

To demonstrate the importance of researching statistics for NBA coaches and managers, this visualization tool will present data for all players during the 2021-2022 NBA season. Understanding player performance is vital to constructing the right group of players who can best succeed. Without knowing these facts, it would be difficult for general managers to decide what players they should consider signing on to the team, what players should play together in the lineup rotation, and who should get the most opportunities to score the ball. To make it easier for coaches and managers to understand these statistics, they can use this tool to look up players and filter on certain attributes. They would select from a list of filters that code for certain statistics such as points, assists, rebounds, steals, blocks, turnovers, free throws, etc. Once specific filters are selected, an interactive scatter plot would be visualized to the user where they can see a point for each player. Filtering based on positions would be implemented so that in the case the user only wants to see players who play a certain position, this feature would only show those players on the graph. On the y-axis would be the player efficiency, and on the x-axis would be the stat chosen. This axis layout would allow the user to see how many instances of this particular chosen stat-line a player occurred and how efficient he was in doing so.

This tool could help managers and those who help build the team together. If a team is struggling with rebounds, they could research using this tool and see which players get the most rebounds among their respective positions. They can then make hiring decisions based on these statistics. This website is a quick way of getting all the data that is found on Basketball Reference's website and other statistic tracking websites and consolidating it into one easy mechanism.

There are many domain tasks that our visualizations will

*e-mail: Patel.mitr@northeastern.com
†e-mail:Kannalath.a@northeastern.edu
‡e-mail:McDonagh.t@northeastern.edu
§e-mail: Perrotta.ni@northeastern.edu

support. If the user wants to compare a particular player's performance with other players in the same position, that is something that can be done. If a particular stat is more relevant, there can be a filter done to hone into the relevant data only. Another task is gathering a generic sense of how a particular player is doing across all stat lines.

4 Data

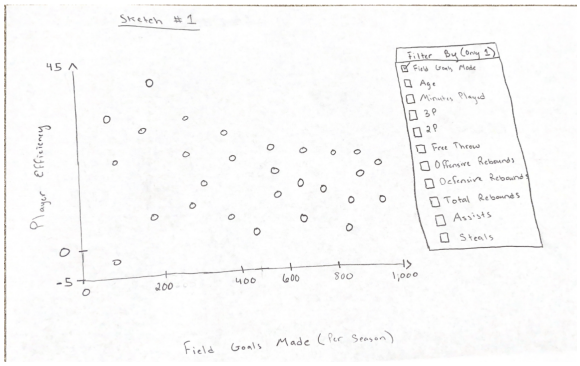
The data collected represents the NBA 2021-2022 season data [1] and comes from Basketball Reference, which is a third-party organization that allows any individual to easily access the raw data for any given player or season. The data is provided SportRadar which is the NBA's official statistics recorder and provider. This organization is responsible for collecting statistics for every game for every player and team, recording it, and then double-checking for any mistakes that were refined after the game had finished. A link to the original database can be found here: Basketball Reference

Although Basketball Reference is not inherently biased or unethical, the data presented may be subject to certain bias or unethical considerations. One potential bias is in the way these statistics are recorded. For example, certain statistics such as assists and rebounds could be subject to the referee and scorekeeper of the game and how they view a particular play. Additionally, certain top players are favored, consciously or subconsciously, based on factors such as media attention or popularity. An ethical consideration that applies to this data is that personal information about every player is publicly available. Although these players provide consent to their performance being tracked and published in their contracts, the widespread access can influence how they are valued in the market as players. Nonetheless, the overall gathering and recording of this data are clean in a manner that does not significantly impact the usage of the data.

The data cleaning we performed started with removing the teams of these players because players can be traded which would complicate the data. It was not relevant as we are strictly interested in the performance of the player rather than their affiliation to a particular team. As a result, for players that played on multiple teams, we removed the multiple rows but replaced them with an accumulation column. Additionally, some of the players were listed as playing multiple positions. This would complicate our filtering process for certain positions, so we did some research on what position those players played mostly and changed it accordingly in the data set. The data itself was relatively clean. There are instances of missing values but these are always percentage calculations that occur when players have not attempted a particular shot. For data additions, we added the following rows that calculate an average per game using the total value divided by the number of games played: points per game, personal fouls per game, turnovers per game, blocks per game, steals per game, assists per game, and total rebounds per game. Additionally, we added a player efficiency (EFF) calculation based on the following formula by Sports reporter and Statistician Martin Manley: (points + rebounds + assists + steals + blocks - missed field goals - missed free throws - turnovers) / total games played.

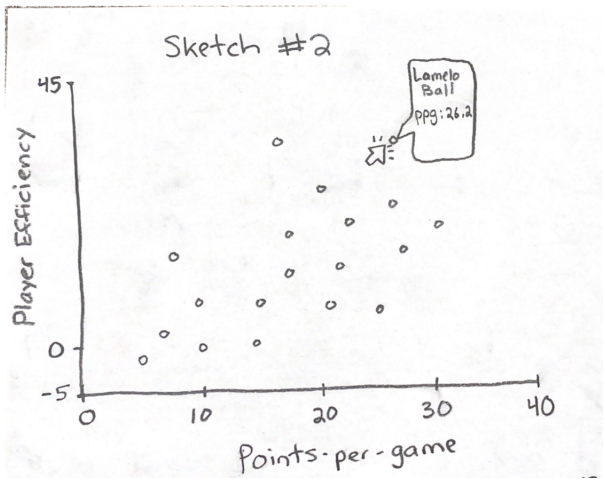
5 Design Process

5.1 Sketch 1



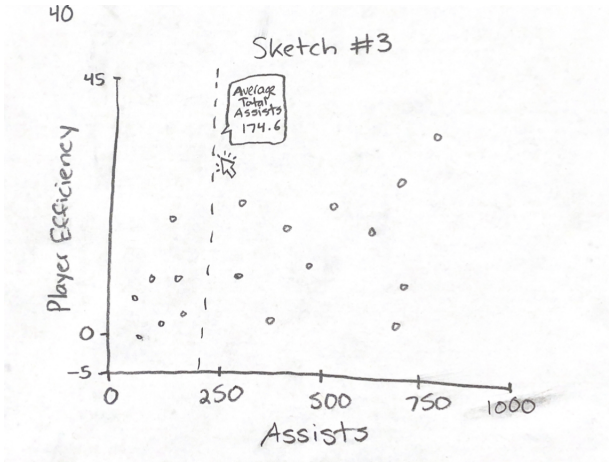
This sketch shows a scatter plot in which each dot represents a different player. On the x-axis the stat chosen, in this case, it is the number of field goals made. On the y-axis, it shows the player's efficiency. We implemented a filter for this sketch where the user can choose any stat of their choosing, and the points will be lined up according to that. For example, if the user changed the stat to steals, a new scatter plot would be shown to accurately represent the data. This influenced the final sketch in that we ended up using a filter where we can filter by any stat along with any position. An addition we added to this final sketch was to link another visualization that allows users to essentially "zoom" into a particular stat and compare a player with the league average for the selected stat. For example, if points are the stat we have selected on the main visualization on the left and the user clicks on a particular player, the right visualization will show a bar representing the selected player's points with another bar representing the average points that a player scores. Each time a new point is clicked on the left visualization, the right visualization updates accordingly based on player and stat.

5.2 Sketch 2



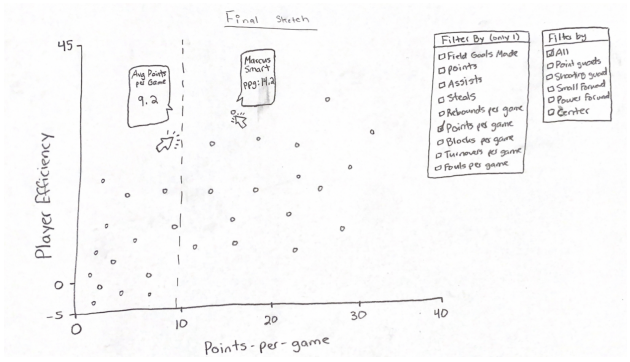
This sketch shows a scatter plot with the x-axis showing the points per game and the y-axis once again showing the player's efficiency. In this sketch, we implemented a tooltip where we could hover over a dot and the player name and stat line would show up. In this example, when we hover over a dot, it shows up as Lamelo Ball with the number of points per game he has. This influenced the final sketch in that we used this implementation of creating a tooltip so that when the user hovers over a point, the player name comes up. This way it creates actually useful information for the user.

5.3 Sketch 3



This sketch once again is presented as a scatter plot with the x-axis this time being assists and the y-axis being the player efficiency. This time, we wanted the ability to compare players to the average. For this, we created a dotted line where if the user were to hover over it, the average assists (or any chosen stat) would pop up. This makes it easy for the user to compare players to the rest of the league. This influenced the final sketch as we wanted to implement this comparison feature to make it easily identifiable where players stood with the rest of the NBA.

5.4 Final Sketch



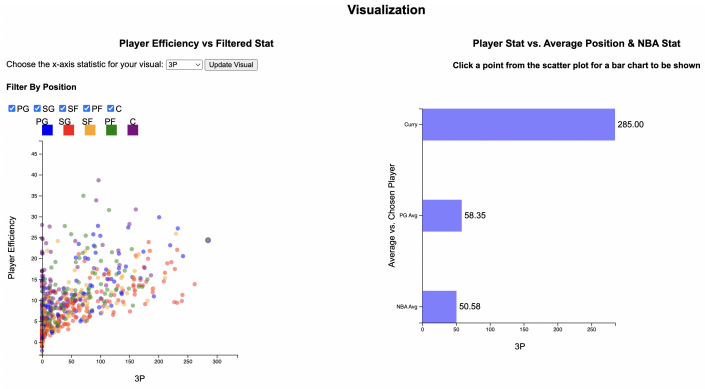
The final sketch depicts one potential option of what the user is able to visualize using our tool. The x-axis and the stat selected by the user to study is points-per-game. The y-axis remains fixed as the player efficiency. We are comparing the points per game of a particular player with the efficiency they are performing with. The visualization is best represented as a scatterplot. The marks in this plot are the points within the scatter plot as well as the vertical line that represents the average of the selected stat. Each point represents a player in the NBA with their associated statistic for the 2021-2022 NBA season data. The channel or the way this information is represented is through the use of position in both the x and y axes. The x-axis represents the stat selected by the user and increases as you move more right. The y-axis represents the player's efficiency and increases as you move up. There is an average points-per-game vertical line drawn in the graph that can be clicked on by the user to display what the average is. If any given point is clicked, the player associated with that point as well as their points per game is displayed in a "tooltip" manner. The tool will implement a filter tool that allows the user to select which stat they want to compare against player efficiency. Additionally, the user can select which positions they want to include data for as an additional filter. All of the features highlighted and displayed within the draft sketches were implemented and brought together in the final sketch.

This visualization will encompass many views based on the user selection of the filter options.

The final design visualization evolved in a few different manners as we got further into the implementation of our project for a few key reasons. The first significant change we implemented was to remove an average stat line in our scatter plot to avoid clutter and rather display the same information in a totally separate graph that is easier for the users to break down. We were able to link a separate graph that had 3 bars to compare: the player you clicked and associated stat, the average across that player's position, and the league-wide average for that particular statistic. The second major change we had was to implement a filter onto the scatter plot rather than to create a "zoomed" feature of the filter view as a separate graph. Originally, we were going to allow the user to choose a particular stat or position they were interested in and that graph would be separately displayed on the right side. This was deemed redundant by our group and hence we implemented a different secondary visualization as discussed above.

As for usability testing, we designed an experiment with the following tasks that needed to be performed by our sample user: filter to point guards, change the stat selected to assists, find the center who scored the most points, compare that center's points with the average points scored by centers, find the player efficiency of the player who scored the most three-pointers, and out of all centers and power forwards, find the one who has the most offensive rebounds. The outcomes of the experiment were extremely favorable and we received feedback that all the tasks were self-explanatory. There was one change we implemented which was moving the selection of the x-statistic to be on top and then allowing for the positions to be filtered right under it. This makes more sense logically and allows users to more easily follow along.

6 Final Design



This visualization allows users to see players that exceed or struggle in a certain statistic. This tool is designed in a way that makes it very easy for a user to conduct research on players. It includes a scatter plot that includes all players in the NBA, each represented as a circle. Each position is colored differently, so the position of a player can be easily identifiable. It uses a drop-down menu for choosing stats along with checkboxes to choose the positions you want to look at. It also includes a bar graph on the right-hand side of the page to compare a player with their respective position and NBA average.

To go over the step-by-step process of how the target user would use our visualization, first, the user would click the drop-down menu and choose from a multitude of statistics. Once the desired stat is chosen, the user would click "Update Visual" to form a new scatter plot that correlates with the chosen statistic. From there, the user can check or uncheck any of the position boxes

to see players only from certain positions. From there, the user can hover over any point, and when they do so, a tooltip pops up that shows the name of that player, the stat line for the chosen statistic, and the player's efficiency of that player. If the user clicks the point, a bar graph pops up on the right side of the page with 3 lines. One is the stat line of the clicked player, another is the average stat line of their respective position, and the last one is the average stat line of the entire league. This way, the user can easily tell if a particular player is below, at, or above average compared to their position or league-wide. Once again we implement a tooltip where if the user hovers over the bar lines, another tooltip will show up that says the player name (can also be Position Average or NBA Average), the stat line for the chosen stat (can also be the stat line for the position average or NBA average), and the player efficiency (can also be the position average efficiency or NBA average efficiency). The statistic numbers are also listed to the right of each bar line. Once the user's research is conducted, the user can continue to choose other players or choose other statistics or positions if they desire to do so.

7 Discussion

The visualization tool answers the domain problem we identified at the very start of our project because it provides users with the ability to compare various stats across all players and filter out information that is not relevant to a particular decision. However, there are some features we discovered along the way that would have been interesting and useful. For example, if we are only interested in viewing data from the playoffs, a user should be able to filter out games that are not in that time frame. Similarly, if users want to identify what time period within a particular NBA season a certain player performs best, that is something that our visualization does not currently offer but can be extremely compatible with once implemented.

In terms of future developments, there is one particular area of focus our group would like to prioritize. The current solution offers historical data and in this case, specifically, we are only offering the 2021-2022 NBA season data. However, there would be great value if this tool was able to continuously update with the latest data. For example, once all the games have been completed and the statistics have been logged onto Basketball Reference, we would like our tool to routinely (ideally daily) pull the newest data for users to see. This will provide a real-time application of the tool and be something that coaches can utilize in the game, or during the season to make short-term decisions that will positively impact the long-term future of the team. Another addition we would like to make is to create a shot chart to see what parts of the floor are sweet spots for specific players. This can be done by gathering data of how many feet away their most attempted and made shots are and then visualizing them in a graph that is the basketball court as the base. This helps teams determine where certain players should maximize their shots and also tells defenders where to be on high alert on the court.

8 Conclusion

In this project, the goal was to develop a visualization tool that can aid coaches and general managers in quickly comparing NBA players' performance across the NBA given their position and statistic that is currently being looked at. Both the scatter plot and the linked bar graph obtain these goals. Although the project proved to be a lot of work, we met regularly and worked independently to produce a well-made project. Mitra contributed to the demo video and overall webpage design, while Alvin was helpful with the linking of the two visual encodings and the group's presentation. Tom and Nick helped with the scatter plot's implementations and filters. Additionally, Nick spearheaded the data abstraction, while Tom put a lot of time and effort into the report write-up. Overall, from start to finish, the group worked cohesively together for our DS 4200 Final Project.

9 Appendex

9.0.1 Data Abstraction

The player row represents the name of each player. The position represents the position that each player plays with the following abbreviations spelled out: C - Center, PF - Power Forward, SF - Small Forward, SG - Shooting Guard, and G - Guard. Age represents the age of a player. G represents the games played during the season. GS represents the number of games started during the season. MP represents minutes played during the season. FG represents field goals made. FGA represents field goals attempted. FG% represents what percentage of shots a player has made. 3P represents the number of 3-point shots made by a player. 3PA represents 3-pointers attempted. 3P% represents what percentage of 3-pointers a player has made. 2P represents the number of 2-pointers a player made. 2PA represents the number of 2-point shots attempted. 3P% represents what percentage of 2-point shots a player has made. eFG% represents effective field goal percentage, adjusting for the fact that 3 points are worth more than 2 points. FT represents the number of free throws a player has made. FTA represents the number of free throws a player attempts. FT% represents the percentage of free throws a player has made. ORB represents the number of offensive rebounds. DRB represents the number of defensive rebounds. TRB represents the total rebounds for a player. AST represents assists for a given player. ASTPG represents the assists per game. STL represents the steals a player has done. STLPG is the steals per game. BLK is the number of blocks a player has. BLKPG is the average blocks per game. TOV represents turnovers. TOVPG represents the average turnover per game. PF represents personal fouls. PFPG represents average personal fouls per game. PTS is the total number of points scored during that season. PPG represents the average points scored per game. EFF represents the efficiency of a player (calculation explained in the data portion). Player and position are categorical attribute types. Eff row is quantitative (diverging) because a player can be below 0. The remaining are considered quantitative (sequential) attribute types.

9.0.2 Task Abstraction

In order to compare a particular player's performance with other players of the same position, there are a few information-specific tasks our visualization will have to fulfill for this domain task. The high-level task that this visualization would have to perform is a consume-discover task, as the user is using the information that the visualization is displaying to gain new knowledge and insight. The mid-level task that this visualization would have to perform is a search browse task, as the user might not know exactly what player they want to look at, but know certain statistics they would want to see. The low-level task that this visualization would perform is a query - identify as the user is trying to use the tool to identify specific players that will fit their roster. In regards to targets for this abstraction, we want to do an all data - outliers approach, as end users are looking to find players that stand out from the crowd. For our second domain task, we are looking to visualize the all-around statistics of a specific player. The high-level task abstraction for this visual will be a consume-discover approach, as the end user wants to look at specific players to discover how they are doing across the board and gain new insight into what the player's strengths and weaknesses are at a high level. The mid-level task that we want to incorporate into this visual is a search lookup, as the end-user has identified a player of interest and wants to get a more detailed view of their all-around game. The low-level task abstraction that

we want to incorporate into our visualization is a query-summarize, as the players' all-around statistics will be summarized within one visualization. For the target of this visualization, using an all-data-features approach will allow the user to see all important statistics (or features) for a specific player of the user's choice.

REFERENCES

[1] *2021-22 NBA Player Stats: Totals*. Philadelphia, Pennsylvania, 2021.

[2] J. Hewko, R. Sullivan, S. Reige, and M. El-Hajj. Data mining in the nba: An applied approach. pp. 0426–0432, 2019. doi: 10.1109/UEMCON47517.2019.8993074

[3] S. Kalman and J. Bosch. Nba lineup analysis on clustered player tendencies: A new approach to the positions of basketball modeling lineup efficiency of soft lineup aggregates. pp. 1–19, 2023. doi: 1548738