Making 2023 NBA Draft Comparisons using NCAA and NBA Shot Charts

Nicholas Lukowsky

Abstract:

Recent data innovation in basketball over the last ten years has introduced the idea of player-tracking data, which allows for easy creation of shot charts across different levels of basketball. Each year the NBA Draft brings with it comparisons for its prospects with past and present NBA players. This study collects shooting data from the NCAA and NBA 2022-23 seasons and develops a model to make draft comparisons based on similarities in shot charts. A R Shiny application is created that allows users to see the results of this model along with producing shot chart visualization for the NCAA draft prospects and current NBA players.

I. Introduction

Each year the National Basketball Association (NBA) integrates new talent into their league through an amateur draft. Containing prospects from an ever-growing variety of lower leagues and teams, the draft allows NBA teams to select this new pool of talent in a structured and organized fashion. The order of selections is decided beforehand through a lottery process in which the lower quality teams are awarded with earlier picks to promote parity across the league and for the future of the NBA. The draft itself is the beginning of the NBA offseason as teams restructure their rosters over the summer in multiple periods, and will occur on June 22nd, 2023, this year.

For many, the pre-draft process begins years before this date as teams and fans alike become accustomed with the prospects that could be selected in the draft. The largest proportion of these prospects play NCAA basketball before transitioning to the NBA. While increasingly more of the top picks in each draft come from different leagues (along with the potential rule change allowing high school players to enter their names in the draft on the cusp of returning) teams are still the most comfortable selecting college players. For this study, exclusively NCAA basketball players will be considered when looking at potential draft prospects. Furthermore, the time leading up to the draft is very exciting for everyone in the basketball world, as teams get the chance to evaluate the prospective future stars of their organizations and fans just hope their teams make the correct pick. As humans love to do, many try and predict the future during this time, projecting what prospects will be selected where and how they will fare during their career in the NBA. This naturally allows for comparisons to be made between draft prospects and the current and past stars of the NBA, where some comparisons have more merit and validity than others. As the annual NBA Draft has become more of a commercialized event, these comparisons have been blown out of proportion, as amateur players are unfairly projected to have careers that resemble the greatest players the league has seen.

With the rise of the knowledge and use of data and analytics in basketball, teams now have a more comprehensive understanding of all aspects the game. A player’s value has become much easier to quantify and visualize, and with regards to the draft, has helped with the accuracy of pre-draft predictions. One tool that has become more prevalent in usage in the understanding of a player’s game is shot charts. Shot charts are useful in evaluating a player's shooting ability, shot selection, scoring efficiency and how they are effective, with regards to scoring, on the offensive side of the ball. By showing where on the court a player attempts their shots and how effective they are from each location, shot charts provide a visual representation of a player's shooting performance in a game, season or career.

This thesis explores the use of this new-age data of shot charts to assess the prospects in the 2023 NBA Draft. Through the collection of data, the manual creation of NCAA and NBA shot charts, and modeling to find similar players, this study aims to take the bias out of pre-draft predictions. This along with the production of a simple tool that will allow users to interact and visualize who their favorite amateur players could become. Overall, by forgoing the traditional way of making comparisons between college and professional players this research could provide valuable insights for NBA teams and fans alike, as they seek to evaluate the talent in the upcoming draft.

II. Literature Review

The following literature review covers the topics that provide background information for the thesis. Spatial analysis in basketball can be interpreted as research using data such as locations, distances, and angles for basketball players on a court. Most of the literature looks into optimization on the most efficient use of your space on a court on offense. Clustering is a means of grouping data points so that more similarities are found between the data of one group when compared to the other group. With basketball clustering can relate to shot charts, statistics, positions, or anything that uses player-level measures. Visualization is key in expressing data to your audience and can draw interest your work just due to its appearance. Shiny apps are applications with the programming language of R that utilize visualization and allow your audience to have a hands-on, interactive experience with your data. Player-tracking is a general term for newer data in sports that collects data in real time from video of a game. This allows for much more granular data to be collected related to anything that can be identified during a game.

A. Basketball Spatial Analysis

An important part into the spatial understanding of shooting in basketball is how the game has changed and evolved to what it is today. Freitas (2021) looked into the three-point revolution in the NBA over the last 40 or so years, searching for trends in the development of the use of the shot over time. He went on a year-to-year basis comparing the league wide attempts and percentage for three-point shots from the first year with the line in 1979-1980 until the 2018-2019 season. He found no statistical difference between consecutive seasons for the percentage of three-point shots (even including the seasons with the shorter line) and found only one season, 1986-1987, where there was a significant increase in the attempts of three-point shots, which he attributed to the introduction of the three-point contest in the prior season.

The more discrete research into spatial analysis in basketball really originated with B. Reich, Hodges, Carlin, and A. Reich (2006) in which they were able to analyze basketball shot charts through a shot success probability model they created. Sam Cassell's 2003–2004 shot chart data was used, which included 1,270 shots with information on game time, location, angle from the basket, result and binary covariates for each shot. These covariates collected data on star teammates being on the court, characteristics of the game, opponent's defensive stats, etc. The results allowed for many conclusions to be drawn about Sam Cassell's and any other player's shot profile and how it's affected by a group of variables using that analytical technique.

This research was furthered through the introduction of new shot coordinate tracking technology in which Shortridge, Goldsberry, Adams were able to investigate the relative spatial shooting effectiveness in basketball (2014). They visualized shot effectiveness by developing metrics that address the variability of the shot profiles of NBA players. Using shot performance data from the 2011-12 season the authors compared players on these new measures of spatial shooting effectiveness (SSE) and points above league average (POLA). They found that these measures were more clear than other metrics like effective field goal percentage that don't account for spatial factors.

The literature of shot selection spatial analysis was advanced by the idea that a players' own individual shot selection can be influenced by their accuracy (Jiao, Hu and Yan, 2020). Using data on the four premier players of Steph Curry, James Harden, Kevin Durant and LeBron James in the 2017-18 season as well as the top 50 most prevalent shooters, the authors used shot charts to understand what made players effective in different ways. They came to the results that for most a positive association between the shot intensity and shot accuracy existed, meaning players were better at the areas where they took more shots. This provided evidence for the idea of sweet spots or hot zones for players, but also shows how players without that association of intensity and accuracy were harder to guard, ex: Steph Curry.

Spatial analysis in basketball isn't limited to just investigating shooting as Hobbs, Gorman, Morgan, Mooney and Freeston explored the most effective zones with regards to entire offensive side of the ball (2018). This took the research of B. Reich, Hodges, Carlin, and A. Reich (2006) to the next step as it bypassed just investigating shooting with regards to offensive efficiency. The authors also took the current state of literature a step further by applying this form of spatial analysis to the women's game as they used data from the women's basketball tournament at the Rio 2016 Olympic Games. They researched into reoccurring sequences of ball movement on offense and overall, how teams pass the ball across the basketball court effectively. They found that under the basket (right-hand side) and the top of the key 3-point area (right-hand side) were the most effective areas on the court, with the entire right-hand side of the court proving to be slightly more effective than the left. They also found trends in ball movement with many teams, with the ball being dribbled from left to right and then passed into the paint, which suggests the common use of the pick-and-roll.

B. Clustering

Clustering is one of the ways in which you can gain insight to data as it aims at grouping the dataset together into clusters so that data points that show similarities are grouped and data points that severely differ from each other are arranged in different clusters. Benabdellah, Benghabrit, and Bouhaddou (2019) broke down the current state of research into clustering and identified five different categories of clustering algorithms; partitioning, hierarchical, density, grid, and model-based algorithms. They pinpointed the characteristics of each category based on the volume, variety, velocity and value (the 4 Vs of big data) of the data. This created a comprehensive survey on all the types clustering-based algorithms in existence and makes it easier to understand and pick out the correct model for a specific dataset.

This analytical research technique can be applied to basketball in a variety of ways as user avyayv (2019) explored the idea of clustering NBA shot charts. They used frequencies of shots at different locations on the court to create a 14-dimensional vector and then, for each player for each season, ran a k-means clustering on that vector. They found the optimal 5 clusters that group all players based on their shot selection. They then did a comparative analysis noticing how the clusters have become more diverse over time with the move towards the 3-&-Key offensive philosophy in the NBA.

The most simple and common way in which clustering is used with regards to basketball analysis is by using individual statistics to group the players. Sampaio, McGarry, Calleja-González, Sáiz, del Alcázar and Balciunas (2015) used clustering to identify key performance indicators in order to differentiate all-star and non all-star players through base-level player-tracking data. Seven different clusters were found, in which clusters 2, 3, 4 contained all of the all-star players. Touches specifically was used as a player-tracking data type that effectively split the clusters into similar groups to roles we see in the NBA today.

Chessa, D'Urso, De Giovanni, Vitale and Gebbia (2022) again looked to grouped professional players based on their performance statistics. However, the authors decided to turn to basketball analyst Dean Oliver and his "Four Factor" statistics to numerically measure the players. They used a weighted complex network analysis based on the statistics for each player and returned four communities of players. After checking the four communities against other general advanced statistics the authors could verify that their results were consistent with the way basketball is played today.

Clustering in basketball can be utilized for other reasons as well such as positioning. Krantz and Shah (2020) looked at clustering in basketball as a way to redefine positions in basketball for optimized roster construction in college basketball to achieve tournament success. They utilized clustering to achieve this by means of player-level college basketball metrics related to physical build, style of play, and shot selection rather than singular advanced performance metrics. They found 9 offensive and defensive clusters and took it a step further by using these defined positions to predict results of NCAA March Madness tournament games over time.

Along with shot charts and positioning, clustering algorithms in basketball can be applied to movement data explore offensive structures in professional basketball (Miller and Bornn, 2017). Similar to how Hobbs, Gorman, Morgan, Mooney and Freeston (2018) looked at entire offensive possessions with regards to spatial analysis in basketball, the authors researched the equivalent for clustering in basketball. Using data with 190,000 possessions of offense from the 2014-15 NBA season where player movement and velocity were tracked, a portfolio of possession maps for every player was created. The authors broke down these possessions into actions based on the movement and velocity data to use a segment clustering algorithm to group these actions into clusters. Then by interpreting these clustered actions as complete possessions the authors used a hierarchal model to group together possessions with similar offensive structure and actions. This allowed for efficient use of an entire database of player-tracking data for game preparation and scouting measures.

C. Data Visualization

Data visualization is a very useful and powerful tool when used correctly and Gomes (2018) focused on data visualization and strategies for presenting engaging and appealing visualizations. Gomes began with some background on the uses and purposes of data visualization, stressing the vast amount of data out there in today's world and the inability for our brains to process and understand it all on our own. She then introduced the most popular types of data visualizations, including line charts, scatter plots, pie charts, etc., and the strengths and weaknesses of each type of charts. Gomes also explained some general rules of thumb when creating visualizations such as understanding the audience, organization, inclusivity, etc. Finally, she showed some example of poor visualizations and explains why they don't achieve what they were intended to or mislead audiences from the true characteristics of the data.

Visualization can be applied to basketball analysis in many ways to understand the dynamics of the game. Hwang (2020) provided a general overview of shooting offense in the current NBA through an analytical viewpoint using many visualizations, including shot charts. He explained modern NBA offenses by first introducing where teams took their shots from, creating a clean shot chart and histogram along with it. This led into shot value, allowing for the understanding of why teams take shots from certain places on the court. Finally, he underwent a simple comparison exercise between the best and worst offensive teams' shot charts of the 2018-19 season, the Warriors and Knicks. He focused on how to visually create the ideas you want others to draw from the shot chart using colors to indicate success and size to indicate volume.

Using extensive data packages these basketball visualization techniques can be applied to the college game as well. Woolf (2018) explored college shot charts with a total of 1,068,844 shots collected and used in visualization. He first made simple observations about intensity, accuracy and efficiency for every shot and came to predicted conclusions, 3-&-Key offensive philosophy, inefficient mid-range shooting, etc. Woolf then separated by shot type; jump shot, layup, dunk, hook shot and tip shot to look for trends amongst these different shot types. Finally, he grouped the shot data by time elapsed in the game to provide insight into how a player's game evolved throughout the game.

Zuccolotto, Sandri and Manisera (2019) visualized shot performance for NBA players with regards to spatial performance indicators using graphs and shot charts to their disposal. The authors used the techniques that are outlined in Hwang (2020) to create shot charts that express the ideas they want readers to understand. They began with regular shot chart analysis with regular data point, density, heat map and hex bin shot charts finding perceived results. Then using the shotchart package they split the court into a specified number of zones providing more context for a player's shot selection. Finally, the authors used the analytical method of classification trees to divide the court into rectangles based on the player's efficiency in different areas.

D. Shiny Applications

An important tool in data analytics to make your research feel more prevalent to the end user is Shiny applications in R. Wrobel (2018) presented background information about Shiny apps along with a basic tutorial in which a R NBA shot chart Shiny app is created. She began with some shiny app basics relating to the transition from one's code to the app's user interface and stresses the importance simplifying the process for the end user. Wrobel then describes the tutorial for NBA shot charts and outlines some of the basic principles that were highlighted in the previous articles. Specifically, she explains the emphasis on creativity and making the app one's own through Shiny's reactive programming capability.

Schneider (2016) outlines a few different approaches for visualizing shot data in a shiny app utilizing an R package he created called, BallR. He presented three different variations of a shot chart you can create for a given player, first being a regular make/miss scatter chart. This used location data to display all shots with a point like a scatterplot would do and color-coded for a make or miss. Next, he created a hexagonal chart which takes the points of the scatter chart and groups them into hexagonal regions across the court. Rather than makes and misses, the color coding is based on efficiency statistics of the shots in that hexagon. Finally, the author discussed heat maps as a shot visualization tool as two-dimensional kernel density estimation was used to create smooth areas where players attempt their shots. This method didn't provide any information on efficiency but brought a unique element that was visually appealing.

Henson (2018) takes this same idea of a shot chart Shiny app from Schneider (2016) but applies it to men's college basketball. He created a Shiny app that presents shot charts to visualize player shooting performance. This shiny app utilized a hexagonal shot chart system to split the court into small hexagonal zones. Efficiency was displayed through the color-coding of the zones by points per shot.

Samangy (2022) brings these concepts together to create a Shiny app dedicated to scouting for the NBA Draft. The author created a hub of information relating to last year's draft with analytically creative tools such as shot charts, radars and percentile curves along with traditional statistics, big boards and other player information. Specifically, this Shiny app used the analytical technique of similarity scores to create NCAA to NBA draft player comparisons through the use of percentile radars for different aspects of a player's game.

E. NBA vs. NCAA Comparison

While the NBA and college basketball are the same game there are many differences in how the game is played at each level that are important to understand before analyzing them together. Nichols (2009) was an early addition to the research as he found the statistics with the highest correlation from college to the NBA for players. Based on their equivalent college statistics the highest correlation were with a player's blocks, assists, rebounds, three-point percentage, and free throw percentage. Free throw attempts, points per game, field goal percentage, and field goal attempts were the statistics with the weakest correlation. Weirdly enough, free throw percentage didn’t have the strongest correlation even though it is the only metric that has the exact same conditions in college basketball and the NBA.

This research is taken a step further an applied to shooting performance across levels in (Schneider, 2018). The author compared shot selection of NBA and NCAA players with regards to shot type, value, closest defenders, etc. Firstly, he examined general shot accuracy between the differing levels of the sport concluding the obvious that NBA players are more accurate shooters than college players. He then dives into analysis relating to predicting the transition to the pros for college players utilizing data on accuracy from NBA 3-point range in college for players.

Miller (2016) examined the topic of NBA Draft player comparisons using player statistics rather than the normal eye test which is used to draw comparisons. He created his own model using the idea of similarity scores, which is an analytical method which produces a numerical score which quantifies the relatedness between two players based on a variety of statistics. He used a group of advanced statistics that are widely accepted in today’s NBA to produce this model and then applied it to a few players from the 2013 NBA Draft drawing more accurate results than others not using the vast array of advanced analytical metrics that we have at our disposal in basketball today.

When comparing and contrasting anything, biases can come into play and Spinella (2021) investigated NBA Draft scouting through a non-statistical lens by looking into how many different types of biases affect one’s judgement of a prospect. He introduced 10 different types of biases and applied basketball to each of them to show how the perception of a player’s ability can be swayed by things that aren’t his play on the court. For example, the bandwagon effect is one of the biases mentioned, as it is very hard for humans to be bold with their ideas as if everyone said a player is good at X it was much easier to agree with that idea rather than go against the grain and disagree.

F. General Player-Tracking Research

The basis for a majority of the developments in the basketball analytics world have been related to introduction of the uniform collection of player-tracking data across basketball. Maheswaran (2015) was a TED Talk by the president of Second Spectrum, whose technology is the official player tracking technology for the NBA and has been since 2017. He introduced the technology and the fine details of how it’s use helped NBA teams through the spatiotemporal pattern recognition abilities of machine learning. He broke down how, through the use of high-tech cameras, Second Spectrum is able to collect data on everything that happened on a basketball court from simple actions like shots and passes to advanced movements such as down screens and pick & roll defensive coverages. He described how a shot can be dissected into two aspects, shot ability and shot quality, to predict shot probability.

Grange (2022) showed how variations of the technology that was discussed in Maheswaran (2015) are being used by current NBA teams. He explained how the Raptors implemented an NBA court length video board alongside the practice courts that provides visuals and instant feedback on a player’s shooting form. The video board was backed by the Noah Basketball technology, which used cameras around the court to track your shooting arc, depth and left/right entry into the basket. This tracked shot chart data that overtime created the perfect shot for your shooting form based on the three characteristics that produced the most makes for you. For the Raptors this was perfect for player development, return from injury protocols, and regular shooting practice. They tracked all your shots you took in a game and practice and see exactly why you were missing shots when you missed them.

Zuccolotto, Manisera and Sandri (2017) looked into shooting and player-tracking data with regards to its effect on high-pressure game situations. Using data from the Italian “Serie A2” league they identified factors that affected the scoring probability of a shot. To investigate the possible multivariate associations between different in-game situations a CART model was used which produced some important insights into game mechanisms due to its tree structure. The authors found the situations that most affected the shot success probability were the end of the shot clock, the second free throw following a miss. Finally, the authors confirmed their study holds across different levels of competition in basketball by reapplying the model to 2016 Olympic basketball data and returning similar results.

The introduction of player-tracking data has also made its way to women’s basketball as the WNBA introduced a new stats website (Nemchock, 2019). The author described how the WNBA website now tracks and collects the same data that has been collected in the NBA, so shot charts and other analysis can be produced in identical fashion to that of NBA.com. He then explained how to analyze the charts specifically with their use of color-coding for FG% vs. league average. The collection of this data for women’s basketball could induce similar analysis that has changed the way NBA teams play the game, developing the WNBA as a sport.

Lichtenstein (2023) introduced the R package, gamezoneR, as it assists in data scraping for college basketball player-tracking data. The author explained how this package is very useful when creating shot charts as well as a variety of other basketball performance visualizations. It provided coordinates for every shot taken by a player or team which allows that to be turned into data points for a future shot chart visualization. Lichtenstein went into the specifics on how the data is scraped from STATS LLC's GameZone and the functions that are built into the package that ease the process of gathering only the data that is useful for you.

III. Data & Methodology

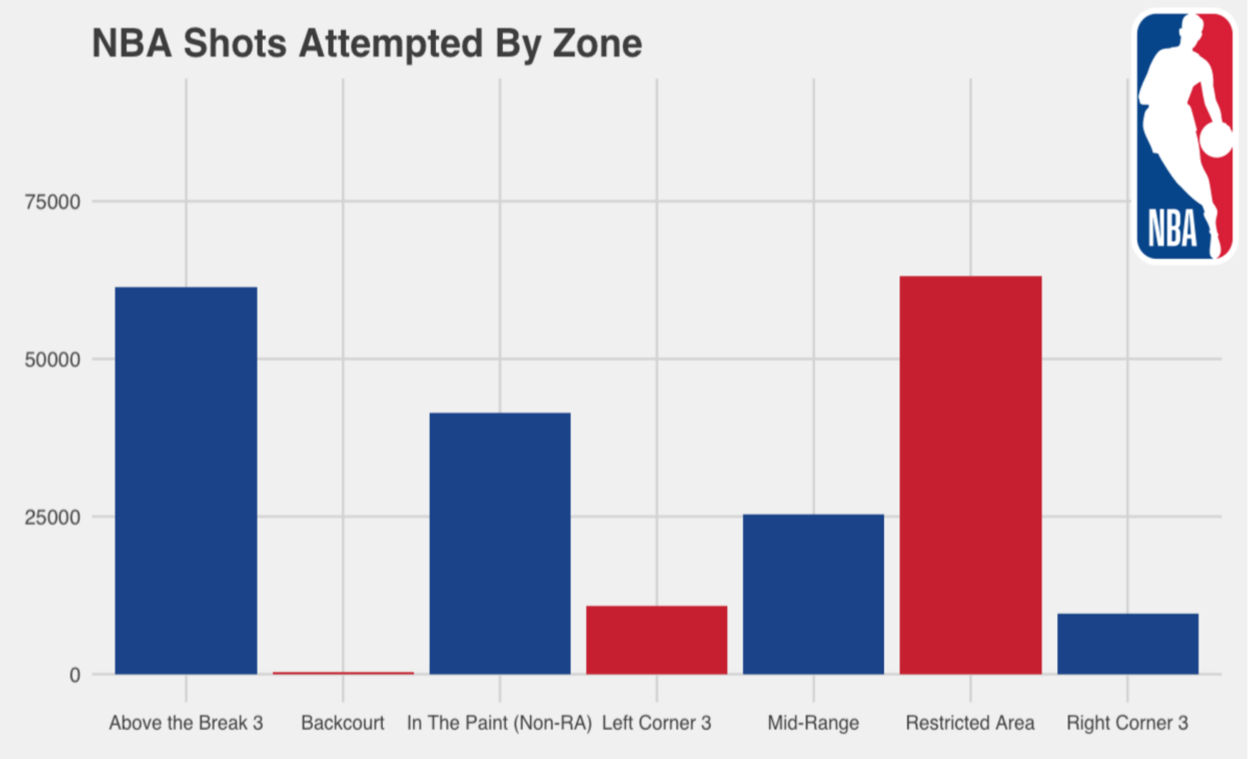
A. Data Collection

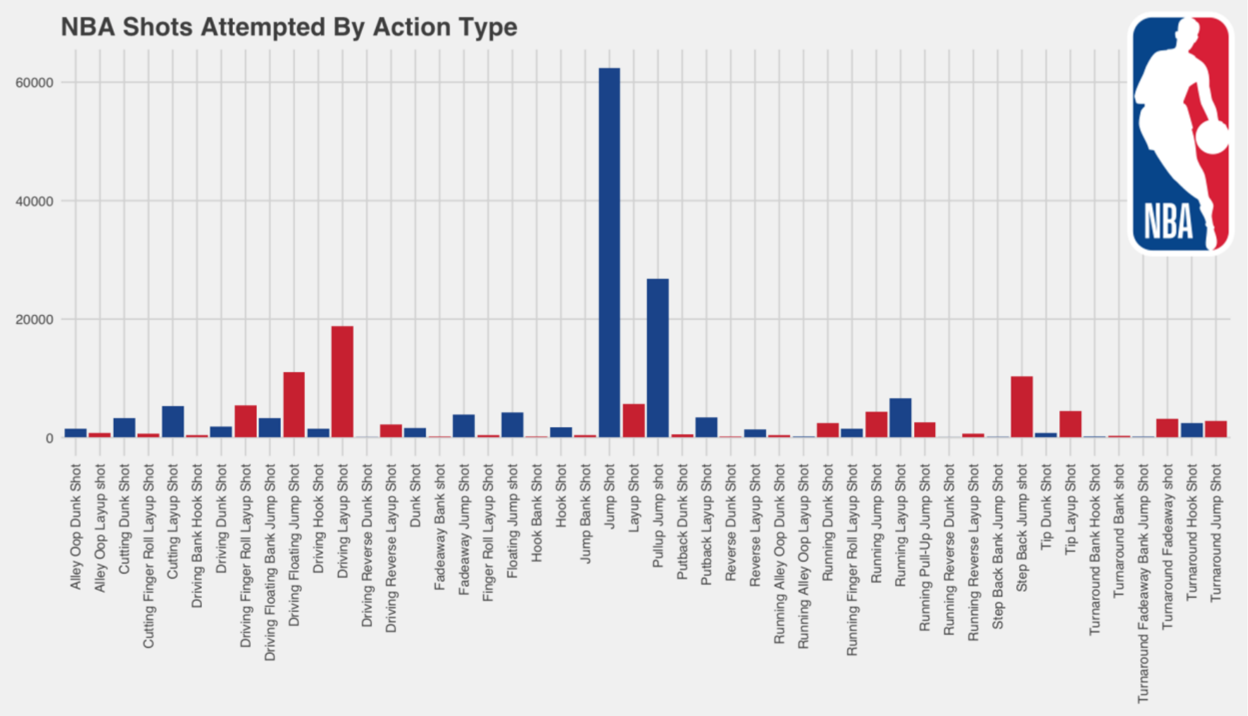
The only data collected for this research was simply play-by-play shooting data for NBA and NCAA players. In order to make comparisons between NBA draft prospects and current NBA players, shooting data was collected for every player to attempt a shot in the 2022-23 NBA regular season using the R package, nbastatR. Developed and maintained by Alex Bresler, nbastatR is a package that acts as an interface for professional basketball data in R. It complies data from web sources such as NBA Stats API, Basketball-Reference, HoopsHype, Basketball Insiders and RealGM into one place for data exploration. Using this package and its teams_shots() function the shooting data for every team was collected form the NBA Stats API one by one and arranged alphabetically by the player’s name. Then, all the dataframes of the team’s shooting data were vertically joined by alphabetical city name to create the entire NBA shots dataset with 217,220 observations of shots. For later analysis purposes all players that recorded less than 100 field goal attempts for the 2022-23 season were removed from the dataset.

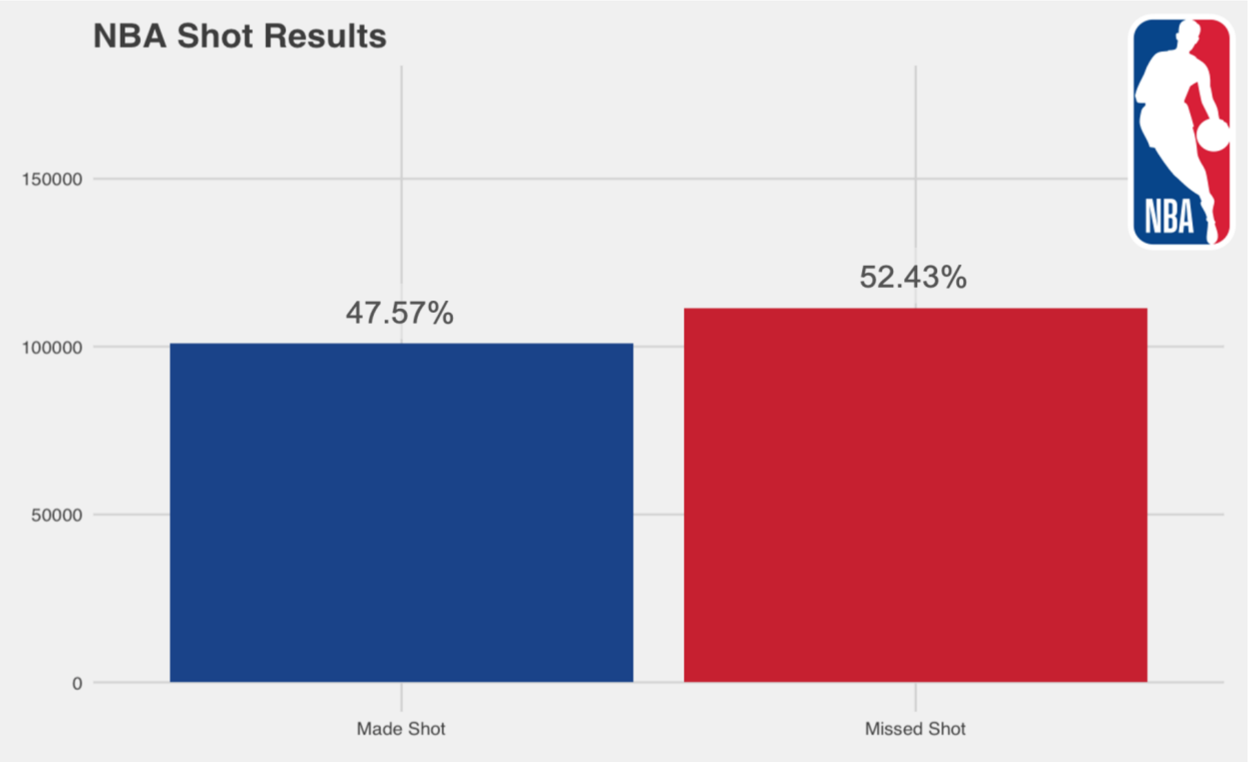
Conversely, for the collection of NCAA shooting data only potential 2023 NBA draft prospects were to be considered. Using a variety of online mock drafts and big boards, along with the Rookie Scale Early Entrant Tracker a pool of prospects to collect shot data for was made. This of course will be updated as dates like the NCAA and NBA Early Entry Withdrawal Deadline in late May & early June draw closer and there becomes a better idea of who will officially keep their name in the 2023 NBA Draft pool. In order to collect this data, another R package, gamezoneR, was used. Created by Jack Lichtenstein, gamezoneR is exclusive to NCAA Men's Basketball as it scrapes play-by-play data with shot locations from STATS LLC's GameZone into a tidy format. As opposed to other sources of NCAA data collection such as ESPN, STATS LLC has a much larger volume of data with almost double the shots charted per season to ESPN. However, although gamezoneR is able to track and collect more data than its counterparts, it still misses a most mid to low-major games, so some less known draft prospects were not included in the NCAA dataset. Also, the package does not track international teams so all non-NCAA prospects including presumed number one overall pick Victor Wembanyama did not have their shots collected. Using the package, play-by-play shooting data was collected for the top 100 NCAA draft prospects for the 2022-23 men's basketball season. Again, for sample size concerns any player with less than 100 observations was removed from the dataset. Finally, the player's shot locations were vertically joined to create the final NCAA dataset and for both datasets the x and y shot coordinates were adjusted to the same scale.

B. Understanding the Data

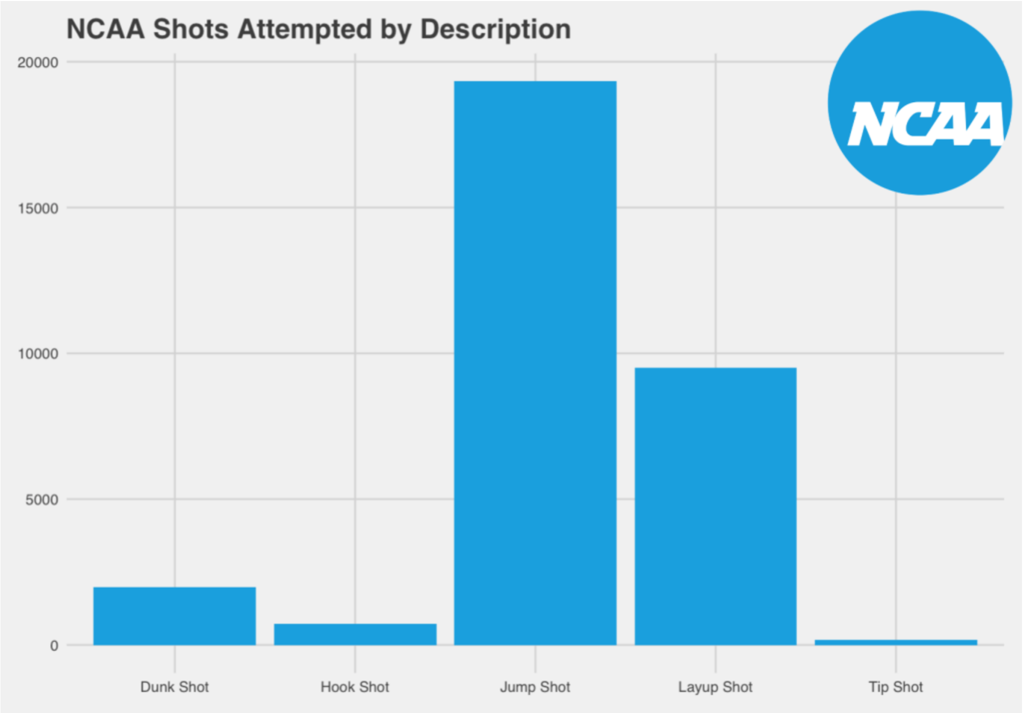
Because the datasets included x and y shot locations along with some shot detail data standard summary statistics didn't apply. Bar graphs were created using the available data to summarize both the NBA and NCAA datasets.

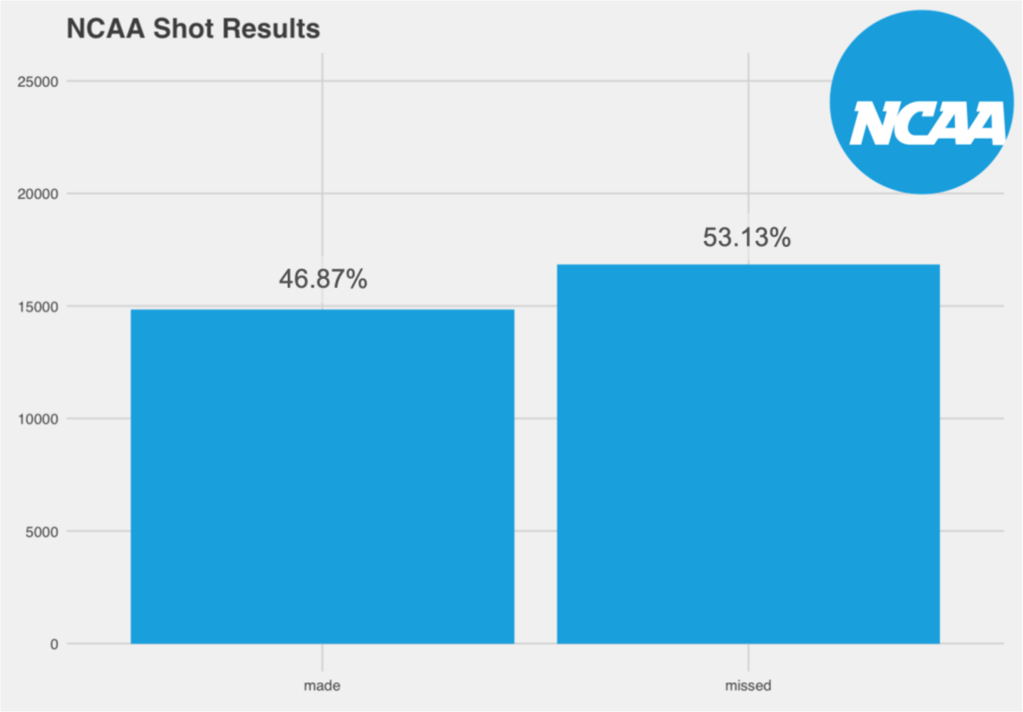






Looking at the NBA summary visualizations, the analytics revolution of the late 2010s is shown in the data with the most common zones being in the paint and behind the three-point line. Along with this the most common shots are all variations of jump shots and layups, and the overall shot success shows a little more shots are missed than made.

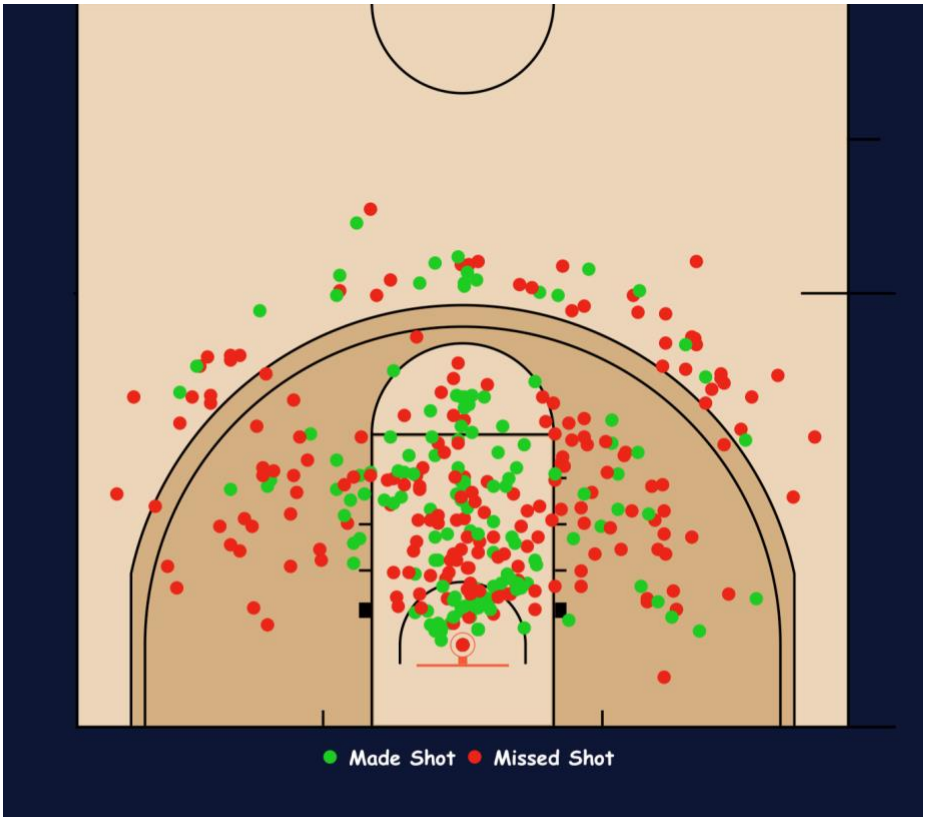


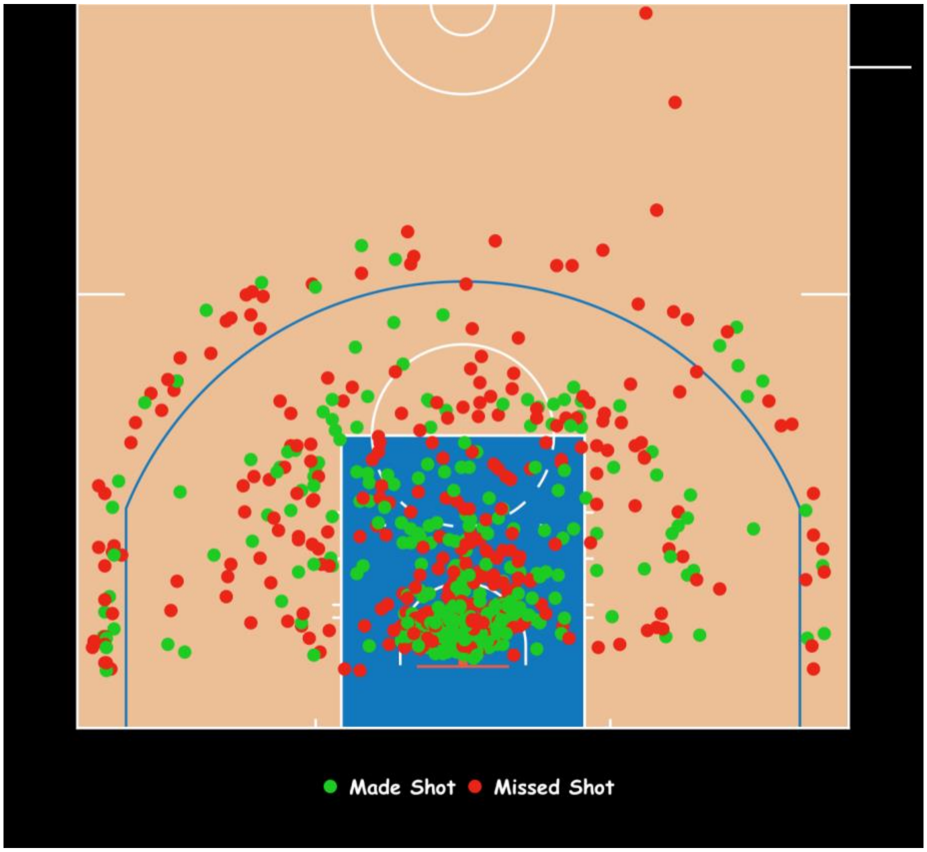


The NCAA summary visualizations show similar trends in the data with jump shots and layups taking the crown. However, you should remember that while the NBA dataset includes all shots attempted in the league for the entire season, the NCAA dataset only includes shot attempts for the selected pool of prospects.

C. Visualizations

As the goal of this research was to make comparisons between different levels of basketball players using their shooting data, the main visualizations that were created were shot charts. Although shot charts don't tell the full story of a player's impact on the offensive side of the ball, they can show how a player is utilized within an offense and when looking for comparisons, can find players that may have a similar influence within a team's offense from a shooting perspective. In order to create shot charts using the available data, another R package was used, this one named sportyR. Written and maintained by Ross Drucker, sportyR is a repository specifically made for visualizing play-by-play tracking data in a variety of sports. The package contains the code necessary to draw scale versions of playing surfaces along with the ability to customize the look of these plots with pre-defined zones that can changed to any color the user would like. As a means to increase the overall design quality of this research, using this package every NBA home core court design was manually created for the display of the NBA shot charts, and NCAA shot chart used the base 2023 Final Four court design for all plots. An example of a potential shot chart comparison between Syracuse guard Judah Mintz and Orlando guard Markelle Fultz using this visualization tool is as follows.





Using these sportyR designed courts along with regular ggplot mechanics these shot charts were produced. To draw conclusions from these outputs you're first drawn to the similarities to where field goal attempts occur for both players. Looking at Mintz, it is easy to see his proficiency as a 3-level scorer as he attempts and makes shots from almost everywhere on the court. This is true as well for Fultz, but it is seen that Fultz is used in the corners much more which is probably a result of playing in an NBA offense where he is off the ball at a higher rate and is asked to spot up in the corner for three-point shots. Another similarity that can be drawn from these charts is both players struggle from the outside of the three-point line, as both players have many occurrences of shots from this area but a majority of them are missed shots. Overall, the visualization tool of shot charts serves as an aid for the later model that can show the user at which places on the court an NCAA player has a similar offensive game to an NBA player.

D. Modeling

The NBA Draft is notorious for comparisons and the live television broadcast of the draft has become commonplace for "experts" to make unfair correlations between the draft selections and NBA stars. This practice almost entirely utilizes the eye test usually minimalizes prospects styles of play to a few standard groups. This is where this research was different, as to make comparisons between NCAA and NBA players, only data was used. By using the x and y coordinates of tracking data for player's shot locations, NBA Draft player comparisons were made. However, it must be remembered that these comparisons were made only on the offensive shooting aspect of player's games and doesn't take into account any defensive or non-shooting impact a player has within a game. To approach this challenge there were a variety of ways to develop a model but what was ultimately settled on was a Kernel density estimation and a Kullback-Leibler divergence model.

Kernel density estimation (a.k.a. KDE) is an application of kernel smoothing to estimate the probability density estimation function of a random variable. It is often used in data analysis to create a smooth curve of the data and for this analysis it was used to estimate the spatial distribution of a player's shot chart two-dimensionally. After the use of this method, we were left with a spatial density distribution that reflected the density and locations of a player's shots but also takes into account the similarity of other, nearby shots. Once these probability density functions were created, the Kullback-Liebler divergence (a.k.a. KLD) was then used to make comparisons. KLD is a way to measure the difference between two probability distributions as the divergence is the difference when one distribution is compared to another. For this research, the divergence measured how wrong you would be if you tried to use one player's shot chart to predict another's, so therefore, the smaller the divergence, the more similar the two shot charts were.

Using the example from the visualization section before we can transform the shot charts of Mintz and Fultz into probability distributions and then measure their divergence to see how good of comparison it is. To achieve this in R the bkde and KLD functions from the KernSmooth and LaplacesDemon packages were used. After using this model for this example the mean sum of the KLD came to a result of 1.053e-01, which is somewhat low meaning that this base comparison by the eye test has some backing in a data sense as well, although there are probably some better comparisons to be made. Overall, this model was used for a selected NCAA player and then iterated over every NBA player in the dataset to find top comparisons for that NCAA player based on the lowest KLD scores.

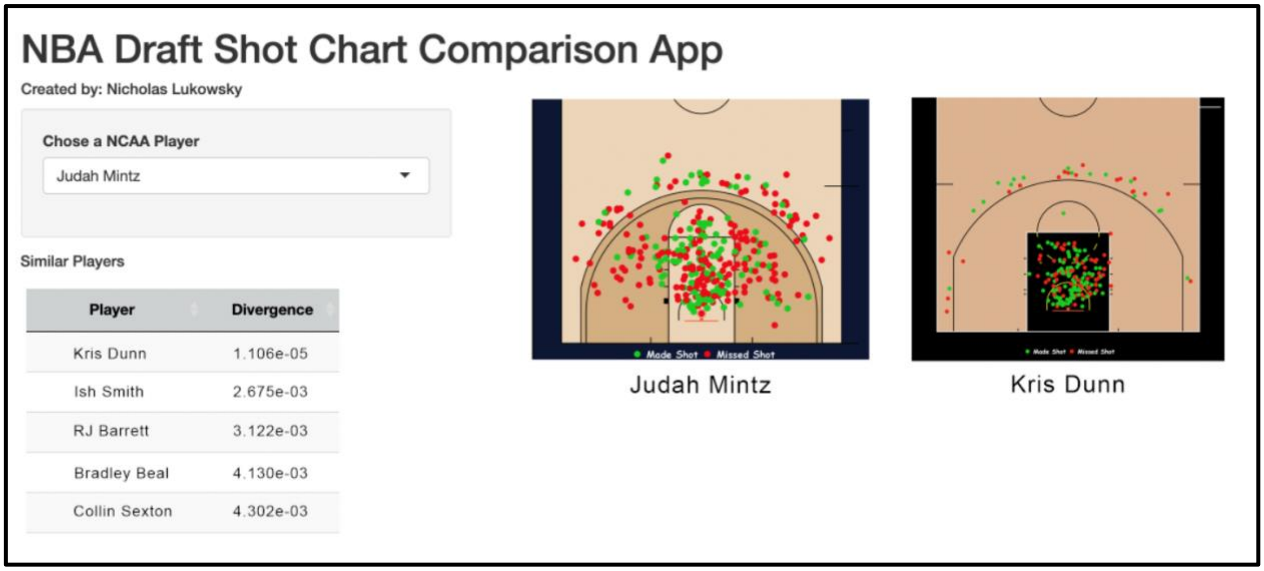
E. Shiny Application

To take this research to the next level, the idea was to make this investigation easier to understand and allow those where interested a chance to manipulate and analyze the data themselves. A R Shiny application was developed to help users compare and visualize the potential offensive output of NBA Draft prospects, enabling them to see what they could become in their career using shot chart data.

R Shiny is a web application framework for R, the programming language that has been used for the entirety of this research for data transformation, visualization, etc. Interactive applications are created through R Shiny that allow users a way to explore data in a simple, user-friendly way. Visualizations can be produced through the click of a button for visitors as developers can specify exactly how the application can be used. The applications can be hosted on a server or run locally on a computer, and they can be customized to meet specific user needs. For this analysis, the purpose is for users to explore comparisons between NCAA and NBA players. The application follows the same process that the Modeling section above outlines using shot charts to make NBA draft comparisons.

To begin, the application opens with a drop-down selection of all the unique NCAA players that were included in the dataset as potential 2023 NBA Draft prospects. The user selects a player to begin the analysis and the R Shiny application then collects the NCAA shot data necessary for this player along with loading in all of the 2022-23 NBA regular season shooting data. Then using the Kernel density estimation approach explained previously the selected NCAA player along with every unique NBA player's shot chart is converted to its corresponding probability density function to find the two-dimensional spatial distribution estimate of that shot chart. Still within the Shiny application, the Kullback-Leibler divergence is found between the NCAA player and every NBA player using a for loop to iterate this model the necessary number of times. A top 5 lowest KLD mean sums table is produced along with the corresponding NBA player to show some NBA players who would be good comparisons to the NCAA prospect with regards to their shot chart. Finally, the NCAA player selected along with the NBA player that produced the lowest KLD mean sum have their shot charts displayed side-by-side to give the user a visual aid to go along with the model results.

Using our example from before with Syracuse guard Judah Mintz, the output of the R Shiny application would be created accordingly:



NBA Draft Shot Chart Comparison App

Created by: Nicholas Lukowsky

Chose a NCAA Player

Judah Mintz

Similar Players

Player	Divergence
Kris Dunn	1.106e-05
Ish Smith	2.675e-03
RJ Barrett	3.122e-03
Bradley Beal	4.130e-03
Collin Sexton	4.302e-03

IV. Discussion

A. Findings & Process

This study attempted to take the data that was available across different levels of basketball and find similarities amongst players transitioning from one league to the next. This data was shot charts, as they are not only a useful data source for understanding a player's impact and performance on the court but are a visually appealing way to take in basketball. This thesis took a different approach as more of an exploratory analysis than a predictive one, which coincides with the creation of a user-friendly investigative tool such as a R Shiny application.

The results produced more variety than what would've been expected as in today's NBA a player's offensive role is more uniform than ever. Due to this, when making comparisons a position restriction was implemented into the shiny application for more accurate comparisons. One thing that this research found that differs from non-data driven draft comparisons is more "star" NCAA prospects being compared to lesser-known NBA players in smaller roles. This could be helpful, as not every draft prospect is going to live up to the expectations given to them and their careers make take them into smaller, yet still effective positions within an NBA team's offense. Teams and fans could use this analysis as a different way to approach projecting the NBA career of a draft prospect, as by finding a current NBA player that is similar to a given NCAA player you can more easily see how successful they might be and how they could fit within the overall offensive dynamic of an NBA team.

B. Further Research & Improvement Areas

Overall, there are a few limitations to this study that must be remembered when applying to the real-life topic of NBA Draft player comparisons. As has been stated before, this analysis only take's a player's offensive shooting repertoire into account on both the NCAA and NBA side of the comparisons. This means that these comparisons aren’t your usual draft day television associations between two players and this study isn’t attempted to use shot chart data as a proxy for overall skill. Furthermore, the NBA Draft doesn’t only include NCAA players as this analysis does. Basketball is becoming a more global sport as time passes on which means more and more NBA talent is being produced internationally. This is more prevalent than ever as in the 2023 NBA Draft Victor Wembanyama is projected to be the first international prospect to be selected first overall since Andrea Bargnani in 2006, and the third ever in the history of the draft. However, future research could dive into these limitations as the availability and volume of data increases. Defensive shot charts could be added to help address similarities in player’s defensive strengths and weaknesses, along with assist frameworks and other new-age basketball data that better quantifies a player’s overall game.

The idea of this research took inspiration from a variety of papers and projects in the overall basketball analytics research community. While the idea of shot charts has been around for a while, the study of (Reich, B. J., Hodges, Carlin, & Reich, A. M., 2006) was one the first main papers to investigate the spatial analysis of basketball shooting data, thus making the study a pioneer for what led to this thesis. The content of player-tracking data that is discussed in (Maheswaran, 2015) is what was officially installed for the NBA starting in the 2013-14 season by SportVU and in the 2017-18 season by Maheswaran’s Second Spectrum. Without this endeavor by the NBA to up their tracking of data in their league this thesis could have never begun. However, (Pickard, 2016) and (Samangy, 2022) were the main inspirations for this thesis as the creation of the R Shiny application here combines ideas from both of these projects.

V. Conclusions

The analysis above tackled the challenge of projecting the career of NBA Draft prospects by finding comparable players already within the league. By using the data of shot charts, NCAA players were examined side-by-side with every player currently in the NBA. A model containing density estimation and divergence found which NBA players had the most similar shooting dispersion through lowest divergence in shot charts. It was found that the comparisons might differ from what expected as the NBA has become more positionless, meaning every player is asked to have a more diverse shooting profile than years before. Ultimately, this analysis took a different route than a lot of other traditional thesis research, but through the creation of a R Shiny application presents an interactive tool for all to explore the idea of analytical NBA Draft comparisons themselves.

VI. References

Avyayv. (2019, December 24). *Clustering NBA shot charts (part 2)*. AnalyzeBall. Retrieved from https://analyzeball.com/2019/12/24/clustering-nba-shot-charts-part-2/

Benabdellah, A. C., Benghabrit, A., & Bouhaddou, I. (2019). A survey of clustering algorithms for an industrial context. *Procedia Computer Science*, *148*, 291–302. https://doi.org/10.1016/j.procs.2019.01.022

Bresler, A. (2023). nbastatR: R's interface to NBA data. R package version 0.1.152, https://github.com/abresler/nbastatR

Chepkevich, J. (2023, April 7). *2023 NBA Draft Early Entrant Tracker*. Rookie Scale. Retrieved from https://www.rookiescale.com/early-entrant-tracker/

Chessa, A., D'Urso, P., De Giovanni, L., Vitale, V., & Gebbia, A. (2022). Complex networks for community detection of basketball players. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04647-x

Drucker, R. (2022). sportyR: Plot Scaled 'ggplot' Representations of Sports Playing Surfaces. R package version 2.1.0, https://CRAN.R-project.org/package=sportyR

Freitas, L. (2020). Shot Distribution in the NBA: Did we see when 3-point shots became popular? *German Journal of Exercise and Sport Research*, *51*(2), 237–240. https://doi.org/10.1007/s12662-020-00690-7

Gomes, M. M. (2018, June 14). *Data Visualization – best practices and foundations*. Toptal Design Blog. Retrieved from https://www.toptal.com/designers/data-visualization/data-visualization-best-practices

Grange, M. (2022, October 5). *Raptors investing in new people, technology with hopes of improving shooting*. Sportsnet.ca. Retrieved from https://www.sportsnet.ca/nba/article/raptors-investing-in-new-people-technology-with-hopes-of-improving-shooting/

Henson, T. (2018). NCAA Shot Chart App. Retrieved from https://tedhenson.shinyapps.io/NCAA_Hex_Chart/

Hobbs, W., Gorman, A. D., Morgan, S., Mooney, M., & Freeston, J. (2018). Measuring spatial scoring effectiveness in women's basketball at the 2016 Olympic Games. *International Journal of Performance Analysis in Sport*, *18*(6), 1037–1049. https://doi.org/10.1080/24748668.2018.1550892

Hwang, J. P. (2020, January 20). *Visualising basketball shots in 2020 - the (big) fundamentals.* Visual/Noise. Retrieved from https://www.visualnoise.io/visualising-basketball-shots-the-basics/

Jiao, J., Hu, G., & Yan, J. (2020). A bayesian marked spatial point processes model for Basketball Shot Chart. *Journal of Quantitative Analysis in Sports*, *17*(2), 77–90. https://doi.org/10.1515/jqas-2019-0106

Krantz, C., & Shah, K. (2020). *Defining Offensive and Defensive Positions in College Basketball to Build Optimal Rosters for Maximum Tournament Success*. Syracuse University. Retrieved from https://www.stat.cmu.edu/cmsac/poster2020/posters/Shah-CollegePosition.pdf

Lichtenstein, J. (2023). gamezoneR: A Scraping Interface for STATS LLC GameZone College Basketball Data. https://jacklich10.github.io/gamezoneR https://www.github.com/JackLich10/gamezoneR

Maheswaran, R. (2015). *The Math Behind Basketball's Wildest Moves*. *TED Talks, YouTube.* Retrieved from https://www.youtube.com/watch?v=66ko_cWSHBU.

Miller, A. C., & Bornn, L. (2017). *Possession Sketches: Mapping NBA Strategies*. MIT Sloan Sports Analytics Conference. Retrieved from http://www.lukebornn.com/papers/miller_ssac_2017.pdf

Miller, J. (2016, December 29). *Identifying Player Comps using Similarity Scores & Play Index Tools*. Roundball Reasons. Retrieved from https://www.roundballreasons.com/identifying-player-comps-using-similarity-scores-play-index-tools/

Nemchock, E. (2019, August 7). *Navigating wnba.com's new stats page: Shot distances and charts*. Swish Appeal. Retrieved from https://www.swishappeal.com/wnba/2019/8/7/20755305/wnba-navigating-new-stats-page-shot-distance-location-charts-minnesota-lynx-aja-wilson

Nichols, J. (2009). *How do NCAA statistics translate to the NBA?* Basketball Statistics. Retrieved from https://basketball-statistics.com/howdoncaastatisticstranslatetothenba.html

O'Connor, K. (2023, April 4). *The Ringer's 2023 NBA Draft Guide*. The Ringer's 2023 NBA Draft Guide. Retrieved from https://nbadraft.theringer.com/

Pickard, C. (2016, October 25). *Nylon Calculus: Finding and quantifying similar shooters in the NBA*. FanSided. Retrieved from https://fansided.com/2016/10/25/finding-quantifying-similar-shooters/

Pickard, C. (2016). NBA Shot Chart Finder. Retrieved from https://ctpickard-3.shinyapps.io/NBA_Shot_Chart_Finder/

Reich, B. J., Hodges, J. S., Carlin, B. P., & Reich, A. M. (2006). A spatial analysis of basketball shot chart data. *The American Statistician*, *60*(1), 3–12. https://doi.org/10.1198/000313006x90305

Samangy, D. (2022). 2022 NBA Draft App. Retrieved from https://dsamangy.shinyapps.io/2022_NBA_Draft_App/

Sampaio, J., McGarry, T., Calleja-González, J., Jiménez Sáiz, S., Schelling i del Alcázar, X., & Balciunas, M. (2015). Exploring game performance in the National Basketball Association using player tracking data. *PLOS ONE*, *10*(7). https://doi.org/10.1371/journal.pone.0132894

Schneider, T. (2016, March 8). *Ballr: Interactive NBA Shot Charts with R and shiny*. toddwschneider.com. Retrieved from https://toddwschneider.com/posts/ballr-interactive-nba-shot-charts-with-r-and-shiny/

Schneider, T. (2018, April 2). *Assessing shooting performance in NBA and NCAA Basketball: R-bloggers*. R-BLOGGERS. Retrieved from https://www.r-bloggers.com/2018/04/assessing-shooting-performance-in-nba-and-ncaa-basketball/

Shortridge, A., Goldsberry, K., & Adams, M. (2014). Creating space to shoot: Quantifying spatial relative field goal efficiency in basketball. *Journal of Quantitative Analysis in Sports*, *10*(3). https://doi.org/10.1515/jqas-2013-0094

Spinella, A. (2021, December 9). *Fighting biases: Avoiding the trap (pt. 1)*. The Box and One. Retrieved from https://theboxandone.substack.com/p/fighting-biases-avoiding-the-trap

Statisticat, LLC. (2021). LaplacesDemon: Complete Environment for Bayesian Inference. Bayesian-Inference.com. R package version 16.1.6. https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software

Wand, M. (2021). KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995). R package version 2.23-20, https://CRAN.R-project.org/package=KernSmooth

Woolf, M. (2018, March 19). *Visualizing one million NCAA basketball shots*. Max Woolf's Blog. Retrieved from https://minimaxir.com/2018/03/basketball-shots/

Wrobel, J. (2018, May 8). *Learning shiny with NBA Data*. juliawrobel.com. Retrieved from http://juliawrobel.com/tutorials/shiny_tutorial_nba.html

Zuccolotto, P., Manisera, M., & Sandri, M. (2017). Big Data Analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions. *International Journal of Sports Science & Coaching*, *13*(4), 569–589. https://doi.org/10.1177/1747954117737492

Zuccolotto, P., Sandri, M., & Manisera, M. (2019). Spatial Performance Indicators and Graphs in Basketball. *Social Indicators Research*, *156*(2-3), 725–738. https://doi.org/10.1007/s11205-019-02237-2