# Take 5 (to 14)

## RE-DEFINING THE POSITIONS IN THE NBA

Abstract: The aim of this study was to determine whether the current NBA Positional System properly fits player's on-court actions. Machine learning models utilize advanced tracking statistics to predict a player's position. The current NBA positional system has been around since before the creation of the 3-point line, and the game of basketball has evolved significantly, so these positions no longer accurately describe all of a player's on-court actions. New positions were created using K-Means clustering and analyzed for statistical differences to create a more accurate, functional NBA position system. Then a linear model was used to analyze whether age had a statistically significant impact on the new positions, asking the question how much, if at all, age impacts style of play.

Daniel J Brockett
SYRACUSE UNIVERSITY | DR. PAUL

## 1. *Introduction*

In any sport, a position is effectively used as a shorthand to describe what a player's responsibilities are on the playing field or court. In baseball, a pitcher is the person who is responsible for throwing the ball and getting the other team out via strike out or by the ball being put in play. In football, the wide receiver is responsible for playing out wide, subsequently either catching the ball or blocking for a run play. In the sport of basketball, the positions tend to be slightly more complicated as they each have several responsibilities, but they traditionally break down in this way. The point guard is responsible for ball handling, playmaking and distributing, and forcing turnovers by deflecting passes and stealing the ball. The shooting guard is typically the best shooter on the court, someone who excels at scoring and ball-handling, while also being a capable defender. The small forward is often well-rounded with a balanced skill set between the shooting guard and power forward, making up a player who can score relatively well both inside and out, is a good ball-handler, a versatile defender, and capable rebounder. The power forward is usually a taller player that is capable of scoring close to the basket and in the mid-range, while also being an excellent rebounder, a strong interior defender that is a capable shot blocker. The center is usually the tallest player on the court, which means they excel at the game close to the basket as they should be a formidable interior defender and excellent shot blocker in addition to being a great interior scorer.

These positions have been clearly defined throughout the history of basketball, and typically players have flawlessly filled these roles. Wilt Chamberlain was a center who dominated the game inside the paint, to such an extent his records are considered unbreakable. Kobe Bryant was a textbook shooting guard, as he was typically his team's best shooter, scorer, defender, and overall player, but rarely its best playmaker. In the modern NBA, these same traditions are on their way out. Some of the most popular examples of these outdated positions include all of these oversized playmakers, like LeBron James, Nikola Jokic, Ben Simmons, Kevin Durant, and Draymond Green. These players all possess elite playmaking abilities, making them some of the best playmakers in the league, despite the fact that none of them are listed as point guards, with the exception of Ben Simmons, who is a 6'10" point guard. These mold-breaking players could just be considered an exception, but there's a new exception every week in the NBA. So, it makes sense to start to analyze the trends surrounding the hybrid-style players in the NBA.

The NBA has gotten access to lots of new data, which is creating tons of new possibilities for the game to improve. This influx of data if properly analyzed can be used to help make accurate, critical, important findings. This data includes advanced statistics, tracking data, and metrics created using that tracking data. These new performance metrics and tracking data should allow for the creation of new positions. With these new statistics, there should be ways to make connections about players' styles of play. This new information should allow us to compare player profiles and statistics in a brand-new way, which should allow for the creation of new positions. These new positions would be based on the role that the player fills on the court, not some arbitrary, chosen title that they have been assigned.

The best way, at least it seems, to go about classification for the new positions would be some forms of unsupervised machine learning. Unsupervised machine learning allows for the information to provide its own insights, in the efforts to hopefully keep as much of my own personal bias outside of it as possible. These techniques are frequently used with large data sets containing tracking information or metrics created from tracking information. However, these new positions are not going to remain stagnant due to the fact that they are not just simple arbitrary labels. The new positional labels are going to be based on a player's performance on the court, which is going to be impacted by his style of play and physical capabilities. Due to this, players are nearly guaranteed to change position as they age, so evaluating the ways in which players age could provide useful information, not just for the present, but especially for the future as well.

## 2. *Literature Review*

### A. Positional Responsibilities

Traditionally, the positions of basketball have defined what a player is supposed to be doing on the court, as a point guard traditionally is responsible for playmaking and ball-handling, while centers are responsible for rebounding, finishing, and interior defense. In addition to finding differences between on-court play, it has been found that each different position has unique physical characteristics as well.

Utilizing 18 years of data from the NBA combine, it was found that among all 5 positions, height, wingspan, vertical jump height, reach, line agility, and three-quarter sprint test were all significant factors in determining whether or not a player got drafted (Cui 2019). There were some significant positional differences though, indicating that different traits are valued differently at different positions. Leg power, an indicator of explosiveness and vertical jump, was a significant determinant in getting drafted for guards, while agility and speed were significant for power forwards and centers (Cui 2019). This aligns with the findings of several other studies as well. Neural networks, linear, and non-linear models found that different positions could be separated based on certain characteristics. Weight was found to be the biggest determinant of position, while other important factors included the shuttle run, their speed at anaerobic threshold, and their sprint time between 5 and 10 meters (Pion 2018). It's evident that different positions had significantly different traits. Some papers found that even if the classic positions are less than perfect, they still serve a purpose and are valuable given that only 5 players can be on the court at the same time (Duman 2021). To some extent coaches have shown they want a somewhat traditional 5-man alignment on the court, so they would want to know what style of player they should put on the court in replacement of one another to make the most efficient line-ups (Duman 2021).

Unsurprisingly, these findings of positional separation have been found across the world previously. Basketball and the styles of it played are different around the world, so there is some slight variation in the role that different positions play on the court. Data from 471 female FIBA games in the America, Asia, Africa, and Europe Championships indicated that different locations had unique differences between how players performed on the court (Zhai 2021). Centers from Europe made and attempted more 2-point shots and grabbed more offensive and defensive rebounds than forwards from Europe (Zhai 2021). Asian and European guards took and made fewer 2-point shots and grabbed less rebounds on both ends of the court than Asian and European centers (Zhai 2021). Asian and African forwards also took and made more 2-point shots than Asian and African forwards (Zhai 2021). This gives insight into how different regions are teaching/play the game differently, which impacts the role that players fill even though they may be designated the same position. Among professional Serbian basketball players, centers were found to be significantly taller and heavier than both guards and forwards, while forwards were found to be taller and heavier than guards (Ostojic 2006). Vertical jump power was higher in centers than guards, but their VO2 max levels were significantly lower than both guards and forwards (Ostojic 2006). Using actual on-court stats to determine the role they fill on the court as opposed to what is expected of their position could possibly change this outcome. For example, Kevin Durant is 6'10" and 240 pounds while Ben Simmons is 6'11" and 240 pounds, which means the two players theoretically should be doing the same things on the court. Anybody who has watched a game of basketball knows that these two players have wildly different play styles and on-court skills, so classifying them solely by body type would be misleading. Comparing the physical attributes of 45 players across 3 different US National Basketball Teams (U-18, U-20, and the Senior team) it was found that several physical characteristics, including height, body mass, body fat percentage, explosive power, speed, agility, strength, and high-intensity endurance performance, differed based on position and age (Abdelkrim 2010). Centers and power forwards were found to be significantly heavier, taller, and have a significantly higher performance in the 1-rep max for the bench press than the other positions (Abdelkrim 2010). The agility and speed test performance was significantly higher for guards than centers across all three age groups (Abdelkrim 2010). The U-18 players were all significantly younger, shorter, and higher in body fat than the U-20 and senior teams, which indicates how their bodies and physical capabilities are going to change over time, which could potentially lead to their style of play changing as well (Abdelkrim 2010).

### B. Position-less Modern Basketball

As the game of basketball has continued to evolve and change, the traditional 5-position system has become outdated. Players fill multiple roles on the court by possessing skills typically attributed to other positions. Players like Nikola Jokic, Kevin Durant, Ben Simmons, and Russell Westbrook have redefined what is expected of players at certain positions. Some argue that we need new positions to accurately represent play on the court. Some have suggested we do this by re-classifying the existing positions into 5 new positions (Bianchi 2017). Reclassification of what each position is supposed to do would allow players that do the same thing on the court to have the same positional label. There are two issues that arise with the traditional 5-position system, and they are oversimplification and incorrect classification.

Incorrect classification comes from the fact that the game has evolved significantly since the positions were named and invented, and oversimplification comes from the fact that the current positional system is incapable of accounting for the complexities in today’s game. Topological Data Analysis showed the team that 13 clusters would be ideal, resulting in 13 new positions, which would not end up being the best results (Bianchi 2017). Players like LeBron James, Draymond Green, and Russell Westbrook are all located in between different positions regarding their location in the map of statistical profiles (Bianchi 2017). The final recommended number of clusters was 5, which resulted in 5 new positions based on on-court performance called all-around all-star, scoring backcourt, scoring rebounder, paint protector, and role player (Bianchi 2017). These findings suggest a differentiation between specialist style players and well-balanced, all-around players. Utilizing 13 game-related statistics from the NBL and then analyzing players in a professional league using these new characterizations resulted in finding that the percentage of versatile players within the NBL has grown over the observed 9-year period (Rangel 2019). Players were categorized based on the 13 game-related statistics into one of 2 categories: versatile or specialist (Rangel 2019). It does not specify what type of specialist each player is, but indicates whether they play a well-rounded or a specialized game. Specifically, the forward group, made up of small and power forwards, showed the greatest change from being a specialist to a versatile all-around player (Rangel 2019). Perhaps one of the most interesting findings was that the majority of players that make the NBL All-Star Game are actually versatile players (Rangel 2019). This means that versatile players typically make up over half of the 10-20 most talented players in the NBL, despite making up less than half of the league (Rangel 2019).

## C. New NBA Data

As technology has rapidly improved over the past few years, so has the quantity, quality, and styles of data collected. The issues with the traditional positions and or line-up construction in basketball begin with the fact that the positions, including their roles, names, and characteristics, were determined before the three-point line was ever invented (Narayan 2019). Utilizing PCA, K-means clustering, Gaussian mixture models, logistic and linear regressions, and neural networks, 3 significant different categories of shooter were determined by the variables field goals made, missed, and percentage, which takes into account offense only (Narayan 2019). Using all of the variables directly seems to be overwhelming, but limiting attributes to the most important relevant statistic allows for a more accurate analysis. Shooting statistics along with others found crucial by Principal Component Analysis resulted in 10 clusters of "new" positions: ball-handling backup point guards, superstar talents, rebounding-oriented big men, "score-first" point guards, wings that satisfy one half of 3 and D, athletic outside-shooting big men, backup big men, developmental guards and wings, 3 and D wings, well-rounded guards and wings that aren't superstar quality (Narayan 2019). This tracking data has been used differently in different parts of the world, but the findings have continuously produced new information. Tracking data is crucial because it helps determine the difference between simple statistics and build. Ben Simmons and Kevin Durant have similar physical characteristics, but their different play styles should indicate they belong in different positions, but on the flip side Russell Westbrook and Nikola Jokic both averaged nearly a triple-double, but they did it in two completely different fashions.

Studies have been done that utilize both traditional metrics and advanced analytics in order to analyze NBA player performance on the court, in addition to comparing the difference between the two methods of evaluation (Mills 2015). Analytics and predictions based on numbers can be wrong, although it has a higher rate of accuracy than the traditional "gut" feeling (Mills 2015). Due to this, the most accurate method of prediction should lie somewhere between exclusively following numbers and analytics and those "gut" feelings (Mills 2015). In order to do this though there needs to be some way to quantify the qualitative knowledge that scouts, and other members of the basketball community rely upon when making their "gut" decisions (Mills 2015). The quantification of things like intangibles would help make those data-based predictions that much more accurate, if those intangibles are properly evaluated, because it gives more significant, useful information to work with. While it obviously may prove difficult, if not ultimately impossible, being able to quantify traits, like a player's work ethic or drive, would allow for much more accurate predictions of what a player is capable of, which would immensely improve player evaluation and projections.

Player tracking adds so much new information, allowing new, different elements of the game to be analyzed in new ways. Analyzing space-time data from a basketball game played in by a team based out of Pavia, Italy is made possible using on-court tracking location data (Metulini 2017). The players participating in the game wore a microchip on their jerseys that collected their 3-dimensional location (Metulini 2017). This allows for a classification of what a player is doing on the court, how far they run, how fast they move, and several other pieces of key information (Metulini 2017). Using on-court tracking data allows for a much more accurate analysis of a player's time on the court than simple box score data can offer. This data would allow for the most complete classification of a player's role on the court, as the space they occupy and the ways that they occupy it is the most accurate way of representing their style of play. The game of basketball has changed significantly, and the stats of the game have evolved along with it. More basic box score stats like blocks, steals, and three-pointers were added, while new advanced stats like Player Efficiency Rating (PER) and Value Over Replacement Player (VORP) have been created. There's player tracking data that records values for a player's location all over the court. This data obviously is going to contain a wide number of variables, so Principal Component Analysis (PCA) is occasionally used to reduce the dimensionality of the data set without sacrificing a significant portion of the variability (Bruce 2016). PCA allows for the comparison between teams or players in order to find the similarities between them. A metric was created to measure how far away from each other each individual profile for a player or team was from the next called the Statistical Diversity Index (SDI) (Bruce 2016). Being able to compare players and teams based on skill and performance, while eliminating any "star" or "name" value allows for the true analysis of on-court performance (Bruce 2016). This comparison analysis of a player's movements on the court could be useful during trade or free agency negotiations, as it would prove that certain players fill a different role on court in actuality than they may get credit for. Eliminating the nominal value of a player or position allows for just on-court performance to be analyzed, something that is often overlooked or lost in traditional basketball analysis.

New advanced metrics allow for a more accurate analysis of a player's performance on the court. There are several new means of evaluating a player's shooting performance, one such case is based on the quality of the shot being taken and how likely that specific player is to make that shot. Shooting percentages are often used to solely determine whether or not a player is a good shooter or not, but that is reductive and doesn't factor in shot quality (Chang 2014). A jump shot from the free throw line could vary greatly in difficulty, as a player could be taking an uncontested jump shot from the free throw line or they could be taking a contested fadeaway. Those two shots vary greatly in difficulty, but they're all counted the same when it comes to determining a player's shooting percentage. New methodologies for two new metrics that could be used to evaluate shot quality, Effective Shot Quality (ESQ) and Effective Field Goal Percentage+ (EFG+) have been created (Chang 2014). EFG+ is EFG minus ESQ and it serves as a measure of how a player is performing compared to the expectation (Chang 2014). Utilizing data that contains player locations on the court, researchers were able to determine how far away a defender was from a player at the time of the shot and whether the shot was taken off the dribble or it was catch and shoot (Chang 2014). In one example, Spencer Hawes has a higher EFG than Kevin Durant, but their EFG+ is nearly even because Durant takes significantly harder shots (Chang 2014). Being able to more accurately classify player skills like shooting should lead to a more accurate analysis of the style of game that a player plays. Someone who is effective at making difficult shots has a very different skill set than someone who is effective at making easy shots. Similarly, other new methods of shooting evaluation have been created. Every shot taken has an opportunity cost, as one player shooting eliminates the opportunity for all other players on the team to take a shot that possession. A new metric for allocative shooting efficiency is created based on a player's shooting percentage and his field goal attempt rate (Sandholtz 2019). This allowed for the analysis of each individual lineup's allocative shooting efficiency, as each lineup was ordered based on the estimated field goal percentage and its real field goal attempt rate to evaluate (Sandholtz 2019). This new method of shooting evaluation allowed for the analysis of each player on an individual level, which meant it was possible to account for individual irregularities, such as players being better shooters from certain sides of the court or players that shoot better from far away (Sandholtz 2019). Analyzing how efficiently a player shoots based on who he shares the court with is another possible means of evaluating his performance on the court in a more accurate and effective manner. Generating stats that can accurately capture different facets of the game without overlapping will allow for the most accurate positional classification.

### D. New positions

Due to this influx of advanced tracking data, many people have tried to use these stats to analyze the game of basketball in a new way. People have found differences amongst the traditional positions, as they have found that there are different styles of each position. For example, there isn't just one point guard, but three different subsections of player type within the point guard position. In one case, CVIs lead to them deciding upon 4 different roles within point guard, 4 within shooting guard, and 4 within small forward positions, 5 roles within the power forward position, and 6 clusters within the center position (Duman 2021). Unsupervised machine learning is one of the most common methods used for clustering because it allows for different styles and quality of play to be grouped together. As basketball has become an increasingly positionless sport, the relevance of the original 5-man position system has faded away. Unsupervised machine learning results indicated that instead of the traditional 5 positions, there should instead be 9 based on the role the player fills on the court (Bosch 2020). The 9 different positions varied based on usage, tendencies, skills, and physical characteristics. This model was created using regular and advanced box score stats from basketball reference over the course of 10 seasons to re-cluster the positions.

Instead of evaluating new positions from the standpoint of positionless basketball, others claim that teams are wasting resources on players who don't play a position that truly fits into their system. Utilizing 3 years of data from basketball reference, Linear Discriminant Analysis (LDA), and Principal Component Analysis (PCA) the dimensions of the data are reduced before being used in the model (Cheng 2018). K-means clustering was then utilized to generate the new clusters of positions and PCA was used to determine the defining features for each position (Cheng 2018). Unsurprisingly, they found that certain roles and types of players contributed significantly less on the court than other types of players. Others have used PCA to reduce the dimensionality of the data instead of LDA, while others have used PCA (Schoch 2018). Although it used slightly different data, the findings were significantly different depending upon the dimensionality reduction used, resulting in either 3 or 8 positions (Schoch 2018). Many people use PCA for dimensionality reduction because it highlights the important variables in separating each position (Cheema 2020). PCA was used to capture 99% of the dimensionality of 189 different variables into just 81 (Cheema 2020). Dimensionality reduction is a critical part of the clustering because it allows for the reduction of hundreds of variables to only tens of variables, which reduces the risk of overfitting and error. Other methods of dimensionality reduction include Pearson Correlation or Variance Inflation Factors (VIF) (Jyad 2020). PCA is commonly used to show what portion of the variation each dimension was accounting for (Jyad 2020). This method resulted in 9 clusters or new positions for the NBA (Jyad 2020).

Other modeling methods include using the elbow and silhouette methods to determine the proper number of clusters (Cheema 2020). This includes Gaussian Mixture Model Clustering, and the optimal number of clusters is determined by the Bayesian information criterion (BIC) method (Cheema 2020). The re-evaluation of positions has occurred across all levels of basketball as well. The positions of college basketball players in the NCAA have been evaluated by their performance based on box score statistics and other performance metrics (Diambra 2018). All of these findings support that players can have the same body type and play two different styles of basketball. This means that their games won't fit together in the same way when it comes time to play. The college findings were similar, despite the quantity, quality, and type of stats being slightly different. In college basketball, topological mapping resulted in the finding of 8 new positions compared to the traditional 5 (Diambra 2018). The game of basketball has changed dramatically over the years, but the positions have mostly remained the same, very few players have a skill set that perfectly matches the traditional position system. Most reclassifications support the idea of 13 positions at the NBA level or at the very least suggest more than the current 5 (Alagappan 2012). The 13 new positions all have different skill sets and significant differences between them, either physically or in their style of play. Thirteen, or any other number greater than 5, positions offer a far wider variety of on-court skill sets that could more accurately represent players than the existing 5-man system does.

### E. Machine Learning in Basketball

Machine learning is frequently used in the analysis of basketball because both unsupervised and supervised techniques have plenty of information to offer. The use of unsupervised techniques is particularly common when it comes to using large sets of tracking data, but supervised techniques can be used as well. Some studies have used unsupervised learning techniques to detect action on the court and analyze player movements on the court (Stephanos 2021). Several different algorithms are used in order to identify the ball handler and then differentiate the unique plays from each other (Stephanos 2021). Being able to analyze plays so deeply and with so much detail and certainty could allow for player performance to be evaluated by certain plays. This would then create the possibility for a player's role to be determined by their performance in different situations, similar to effectively assigning a player a new position. Other uses of unsupervised learning techniques have included specific tracking data captured by SportVu (Sarlis 2020). The data from SportVu is broken down to show how a number of different statistics and performance metrics can be captured through the tracking data (Sarlis 2020). Some of these metrics are revolutionary, which leads to the possibility of new statistics.

Other uses of machine learning are limitless in connection to basketball and sports in general as it can be used to accomplish a number of tasks. Machine learning can be used in basketball to analyze large sets of player data, like the tracking data, help create marketing tactics, and calculating the betting lines for oddsmakers. Machine learning doesn't always have to be used on super complex tracking data, as studies have used simple box score data to calculate what factors impact the game outcome both on an individual and team level (Li 2021). Cluster analysis and a Bayesian model resulted in finding that true shooting percentage, team steals, and committed fouls were the deciding factors in close games, while 2-point field goals made, 3-point field goals made, and defensive rebounds were key factors across all games (Li 2021). These findings indicate that certain statistics matter more in the outcome of a game, leading certain skills to be significantly more valuable in the pursuit of more wins. If certain skills hold more value in contributing to wins then the positions that excel at those skills are going to be significantly more valuable than others.

### F. Aging in the NBA

The saying goes that "Father Time is undefeated" and that certainly holds true in basketball, although some have tried to defy it. As players age, their bodies and physical capabilities are going to significantly change as well. In a sport like basketball, where style of play and effectiveness is dependent upon physical capabilities, one's style of play will likely shift as they age. Numerous previous studies show players do change significantly over time, as do the stats they produce (Vaci 2019). Modeling age-related changes is difficult because instead of numbers holding their typical meaning, they hold a significant, but unique amount of information due to the irregular pattern at which humans age (Vaci 2019). I want to re-classify the positions on the court based on a player’s physical characteristics and his on-court statistics and analyze how a player is going to change positions as they age. If a player’s on-court statistics change as they age, which they should based on previous findings, then the position they fill over time should change as well.

Other uses of aging analysis in the NBA have included evaluating the trajectory of a player’s career in order to help teams decide what to do best in free agency and during trades (An 2016). A machine learning model was used to analyze a player’s career based on his current career statistics, and this model would also be used to predict when a player would exit the league (An 2016). Utilizing a data set containing important per-game statistics on players over the course of 20 years, k-nearest neighbors was the method used to determine a player’s outcome (An 2016). This effectively means that utilizing a player’s nearest neighbors, a prediction could be made on how long a player would survive in the league for. A similar analysis could be done with a player’s position over time, as certain trajectories would become clear, showing that some players are rising to stardom or on their way out of the league.

Aging analysis is frequently done across sports as well. Some studies have modeled how a player’s performance changed over time using Functional Data Analysis (FDA) in order to be able to analyze aging results across different sports (Wakim 2014). Using FDA, functional Principal Component Analysis (fPCA) can be used to show how aging curves differ between power and non-power hitters in the MLB, while the NBA analysis tends to be significantly more complex (Wakim 2014). There are 3 distinct aging patterns among NBA players with each having its own unique scoring patterns and on-court performance characteristics (Wakim 2014). However, it was found that aging is independent of position in the NBA, meaning that positions do not play a role in determining which aging pattern a player fell under (Wakim 2014). The aging analysis done in sports is unique compared to traditional aging analysis because sports careers are short and irregular due to suspensions and injuries. This results in the number of players being sampled in a year being quite irregular, which makes Principal Components Analysis through Conditional Expectation (PACE) useful in this instance (Wakim 2014). PACE was proposed in 2005 to help fit smooth data, so it should help with the abruptness of the ages within the data set (Wakim 2014). This could prove useful in analyzing the positional changes that players experience over time, especially since injuries can significantly impact a player’s physical capabilities, which in turn could impact his play style and position. Other studies have evaluated this very thing by investigating the effects of aging across positions on basketball players' physical and technical game performance. Using data from the 2018-19 NBA season, players were classified into four separate age groups ranging from 19-22, 23-25, 26-29, and 30-42 (Kalén 2020). Variables like distance covered, average speed, minutes played, points scored, and playing efficiency showed that the different age groups did have significant differences in on court performance (Kalén 2020). Players older than 30 covered less distance and were significantly slower than younger players, while centers aged 23-25 scored more points per game than older centers did (Kalén 2020). These findings showed that a player's physical capabilities significantly deteriorated with age as they got slower, covered less distance, and were most efficient around age 26, which is most often associated with an athlete’s physical prime (Kalén 2020). These findings suggest that there should be significant positional differences as players age, if the positions are not based on an arbitrary title, but instead the way a player performs on the court.

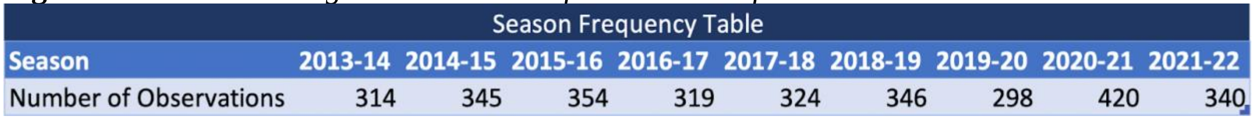
## 3. *Data*

### A. Data Summary

The dataset used in this thesis was largely obtained from the tracking statistics section of NBA.com utilizing a custom scraping function, while a small portion of the information was obtained from Basketball-Reference.com using the nbaStatR package. Due to the width of the net being cast, several overlapping variables were collected, resulting in the use of several different data frames to determine which would provide the most accurate information.

The original data frame contained 3,060 observations of 154 variables. Each observation represents an individual player in an individual season, so 2014-15 LeBron James is a different observation than 2016-17 LeBron James. Each player counts as a different observation for each season because players change and adapt their game between seasons or they can fill different roles on different teams. For example, LeBron James was primarily a Power Forward during his years in Miami, while he's been featured as a Point Guard in his first few years on the Los Angeles Lakers before being listed as a Power Forward yet again. The NBA began collecting tracking data during the 2013-14 season, so the data spans from that year up until the most recently completed NBA season, 2021-22. None of the observations featured any N/A's, as those were filtered out during scraping, so all 3,060 observations could be used in the models. To qualify as a player during a given season, a player must have played upwards of 25% of all possible minutes in a season, meaning in a standard NBA season they must play upwards of 197 minutes. There are approximately equal observations across the seasons, with 2019-20 featuring the fewest players, likely a result of the season interruption and resumption in the bubble. The following 2020-21 season had the highest number of qualified players, likely a result of COVID-19-related absences which forced teams to sign and play more players than they would in a typical season.

**Figure 1:** *A table showing the breakdown of observations from each individual Season.*

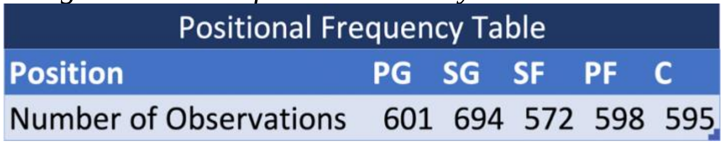


Season Frequency Table

| Season | 2013-14 | 2014-15 | 2015-16 | 2016-17 | 2017-18 | 2018-19 | 2019-20 | 2020-21 | 2021-22 |
|---|---|---|---|---|---|---|---|---|---|
| Number of Observations | 314 | 345 | 354 | 319 | 324 | 346 | 298 | 420 | 340 |

The 154 variables in the dataset were made up of player biographical information, season information, and tracking statistics. The variable of interest is player position. Given that a player's position is intended to describe a player's responsibilities and role on the court, tracking statistics that capture each player's every move should allow for accurate classification of position for each player. Only the tracking statistic variables were used to predict player position, so the season and age variables were removed from the datasets. While many players are often labeled as multi-position players, the first listed position for a player in a given season is the assigned position used. This means the traditional 5-man position system is used: Point Guard, Shooting Guard, Small Forward, Power Forward, and Center.

**Figure 2:** *A table showing the number of observations by listed Position.*



Positional Frequency Table

| Position | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| Number of Observations | 601 | 694 | 572 | 598 | 595 |

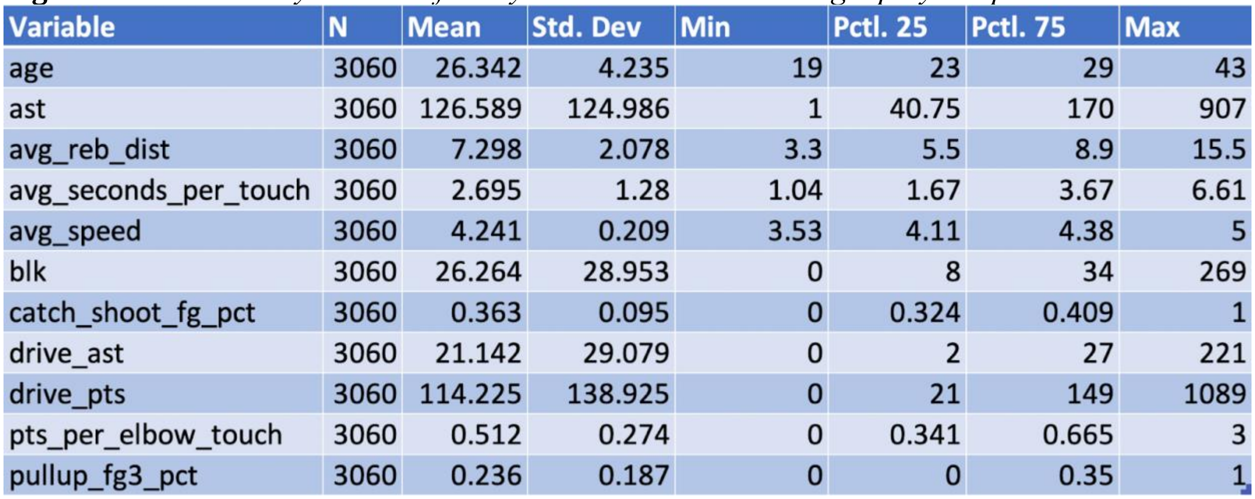
There's a similar number of observations for each position, with roughly 600 observations for each of the 5 positions, a balanced total for just over 3,000 observations. There are the most players classified at shooting guard and the least at small forward, which is likely due to the overlap between the positions and the first listed position being utilized.

### B. Data Analysis

The variables assists, percentage of team assists, field goal percentage, field goals attempted, field goals made, free throw percentage, free throws made, free throws attempted, passes, percentage of touches that a player passes, personal fouls drawn, percentage of team's personal fouls drawn, points scored, percentage of team's points scored, turnovers, percentage of team's turnovers, touches, and points per touch were all collected for players in regards to Drives, Elbow Touches, Paint Touches, and Post Touches. Standard box score statistics were included as well like general touches, points, assists, steals, rebounds, minutes played, and blocks. Advanced statistics regarding possessions and efficiency were also collected like secondary assists, assist points created, rebounding distance, rebounding contest rates, dribbles per touch, and time per touch. Speed and distance data was also collected as there are variables for player speed and distance covered. Several of the variables also contained shot totals and percentages based on the circumstances of the shot. The variables effective field goal percentage, field goal percentage, 3-point field goal percentage, 3-point field goals attempted, 3-point field goals made, field goals attempted, field goals made, and points scored were all collected for Pull-Up and Catch & Shoot scenarios. Advanced defensive statistics were included as well including the defensive rim field goal percentage, defensive rim field goals faced, and defensive rim field goals allowed.

There were several instances of information overlap, as the player's average speed on offense and defense were separately included along with the player's total game average speed. This happened with rebounding as well, as totals, chances, and contests were all collected for offense, defense, and totals as well.

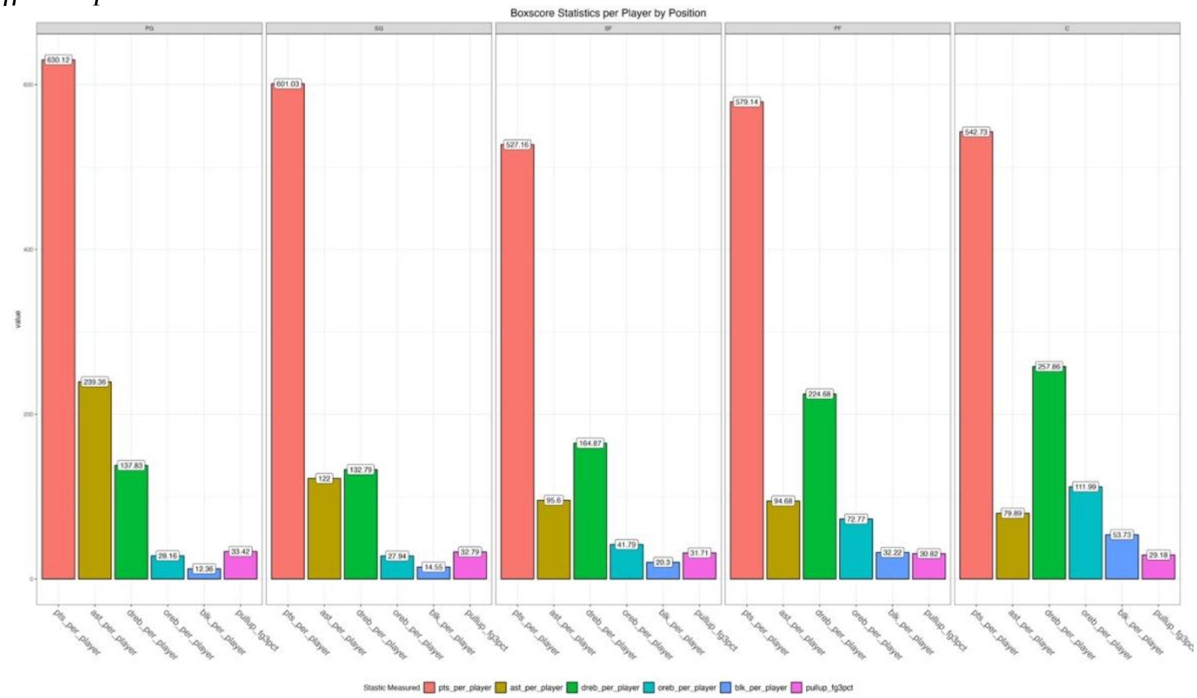
**Figure 3:** *The summary statistics for key variables in determining a player's position.*



| Variable | N | Mean | Std. Dev | Min | Pctl. 25 | Pctl. 75 | Max |
|---|---|---|---|---|---|---|---|
| age | 3060 | 26.342 | 4.235 | 19 | 23 | 29 | 43 |
| ast | 3060 | 126.589 | 124.986 | 1 | 40.75 | 170 | 907 |
| avg_reb_dist | 3060 | 7.298 | 2.078 | 3.3 | 5.5 | 8.9 | 15.5 |
| avg_seconds_per_touch | 3060 | 2.695 | 1.28 | 1.04 | 1.67 | 3.67 | 6.61 |
| avg_speed | 3060 | 4.241 | 0.209 | 3.53 | 4.11 | 4.38 | 5 |
| blk | 3060 | 26.264 | 28.953 | 0 | 8 | 34 | 269 |
| catch_shoot_fg_pct | 3060 | 0.363 | 0.095 | 0 | 0.324 | 0.409 | 1 |
| drive_ast | 3060 | 21.142 | 29.079 | 0 | 2 | 27 | 221 |
| drive_pts | 3060 | 114.225 | 138.925 | 0 | 21 | 149 | 1089 |
| pts_per_elbow_touch | 3060 | 0.512 | 0.274 | 0 | 0.341 | 0.665 | 3 |
| pullup_fg3_pct | 3060 | 0.236 | 0.187 | 0 | 0 | 0.35 | 1 |

The average age for players in the sample is just over 26, which makes sense given that is in the earlier range for a player's prime and younger, more durable players are likely to see minutes during the regular season than older, ring-chasing veterans. It is quite interesting that all of the observations recorded at least 1 assist, likely a result of the minutes floor that was set during data collection. These are all variables that are typically associated with a certain position or group, like point guards are expected to have high assist totals, while a center should have more blocks and more points per elbow touch than other positions, as shown in Figure 4 below. Point Guards and Shooting Guards were the top 2 positions in terms of points scored per player, while Point Guards recorded the most assists per player of the 5 positions by a wide margin. That aligns with the expectations for both of those positions. While they offer an intermediate level of both offensive and defensive rebounding between guards and big men, small forwards surprisingly had little relative offensive production. Many of the league's biggest stars of the 2010's were players primarily considered to be Small Forwards despite their unique offensive skillsets, including players like Kevin Durant, LeBron James, and Kawhi Leonard; however, Small Forwards have the least points scored out of all the positions, while recording similar assist and pull-up 3-point percentage numbers.

**Figure 4:** *A bar graph highlighting the differences between different statistical measures and different positions.*



### C. Data Collection

The variable player age was collected using the nbaStatR package and joined to the information collected from NBA.com. The rest of the information, the remaining 153 variables, comes from Second Spectrum's player tracking system, which stores the data in the tracking section of the statistics page on NBA.com. Second Spectrum is responsible for cameras in the catwalks of every NBA stadium that allows for tracking the movement of the player and ball 25 times per second. The tracking section of NBA.com is broken down into several different pages, including Drives, Defense, Catch & Shoot, Passing, Possessions, Pull Up Shooting, Rebounding, Efficiency, Speed & Distance, Elbow Touches, Post Touches, and Paint Touches. These pages were all scraped and combined into the large data frame mentioned above.

Several of these pages had overlapping variables, so the duplicates were removed to create a new, reduced dataset with 146 variables. This data frame included offensive and defensive splits as well as the averages of the two and shot totals as well as percentages, which meant a lot of redundancy in the model. To accommodate this and test for accuracy and importance four separate data frames were created: one utilizing offensive and defensive splits with shot totals, one with offensive and defensive splits along with shot percentages, one containing averages with shot totals, and one with averages and shot percentages. The variables unrelated to shot totals/percentages and the offense/defense splits, including statistics like assists, seconds per touch, dribbles, and steals, remained the same across all 4 data frames. From this point forward, ***df2*** is the data frame with splits and shot totals, ***df3*** is the data frame with splits and shot percentages, ***df4*** is the data frame with averages and shot totals, and ***df5*** is the data frame with averages and shot percentages.

Any variables that would have caused clear redundancy were also removed, as the percentage of a team's assists that a player has will have some direct relation to the number of assists that player has. It is important that these tracking statistics can accurately predict player positions otherwise it may prove that the NBA and basketball as a whole might need to re-invent the position system it uses.

## *4. Methodology & Results*

There are two different approaches being taken given that there's effectively 2 questions being asked. The first being does the current 5-man position system that the NBA utilizes accurately classify the position of players based on their on-court performance? The second is whether or not age has an impact on the true position (a clustering assignment based on the tracking statistics) of a player. These two questions should work in conjunction to evaluate the current standing of the NBA's position system, as the league has experienced immense change and growth since the original implementation of the system.

### A. Current Position Classification

To answer the question of whether or not the current position system is effective or not, classification models were used to predict the player's position utilizing 78 or 88 variables, depending on the respective data frame used. All 3 of the classification models will be tested with each of the 4 data frames to see which stats are most useful in determining a player's position according to the 5-man position system. All 3 of the classification models utilize both a training and test dataset, so a split was created to use ⅔ of the data in the training set and the remaining ⅓ would be the test set. This meant that the training data had 2,052 observations and the test data had 1,008 observations. All of the numerical variables used were scaled prior to modeling in order to ensure each one carries proper weight.

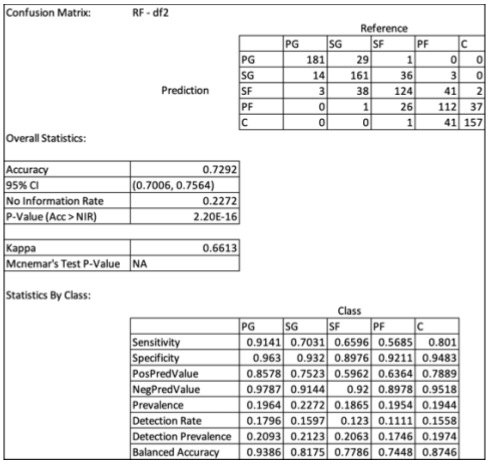
#### a. Random Forest (RF) Classification

Random Forest Classification is a method of classification that utilizes many decision trees to split the data at the points that reveal the most information in determining an object's classification. It uses a large number of individual decision trees as one big ensemble, generating a number of random trees that then split at important values/variables with each tree producing a classification, and the classification that the majority of the trees decide upon being the forest's prediction. Random Forest models tend to work well with high dimensional data, which is what a dataset containing over 70 variables could be considered. The biggest downfall to random forest models is that they struggle to predict data outside the scope of the training data, so any instance that is not accounted for in the initial model will be misclassified. The initial random forest model used all the tracking variables available, in each of the different frames, to predict position. To make each of the initial models more accurate, a search grid and tuning parameters were created and tested with the same original formula. The tuning grid allows for the model to test out its performance with varying numbers of variables sampled as candidates at each split. The more variables tried at a split, the more likely the most accurate split is to be made, but the longer the model takes and the more prone it is to overfitting, which is also a potential issue. At this point ***df2*** and ***df4***, the two datasets containing shot totals in place of percentages, were far outperforming the accuracy of ***df3*** and ***df5*** by roughly 5-10% on both the initial model and the one featuring the search grid.

All 4 models featuring ***df2*** or ***df4*** were then analyzed for variable importance to see which variables were making the biggest impact and if any were causing confusion/adding no information to the model. A third model was then created for each data frame utilizing only the variables that had an importance measure above 8, which gives it a scaled importance measure greater than 0, meaning that it is significantly likely to add to the accuracy of the model. Given that both datasets contain different information, they returned some different variables as providing the most information, but some of the variables were equally high in importance and information added, including average seconds per touch, average speed, blocks, pull-up 3-point field goals made, drive points, and both catch and shoot makes and attempts. The variable importance analysis also highlighted several variables as providing little to no information including free throw makes and attempts from post, elbow, and paint touches.

The models built using just the statistically significant variables from the random forest variable importance analysis were less accurate based solely on the model, while faring slightly worse as well in terms of the accuracy of predictions, indicating that the small amount of value provided by the removed variables may be useful in differentiating between similar groups of players. The p-value of both significant-variable-only models indicate that these results are significant.

**Figure 5:** *Confusion matrix and model results for* ***df2*** *for a random forest model using all variables.*



Confusion Matrix: RF - df2

Reference

| Prediction | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| PG | 181 | 29 | 1 | 0 | 0 |
| SG | 14 | 161 | 36 | 3 | 0 |
| SF | 3 | 38 | 124 | 41 | 2 |
| PF | 0 | 1 | 26 | 112 | 37 |
| C | 0 | 0 | 1 | 41 | 157 |

Overall Statistics:

| | |
|---|---|
| Accuracy | 0.7292 |
| 95% CI | (0.7006, 0.7564) |
| No Information Rate | 0.2272 |
| P-Value (Acc > NIR) | 2.20E-16 |

| | |
|---|---|
| Kappa | 0.6613 |
| Mcnemar's Test P-Value | NA |

Statistics By Class:

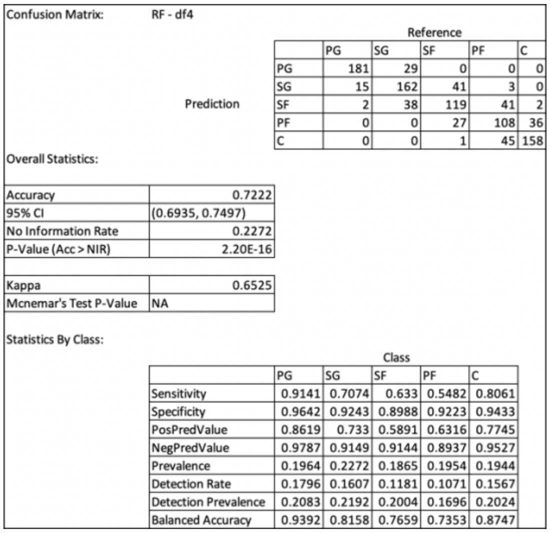
Class

| | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| Sensitivity | 0.9141 | 0.7031 | 0.6596 | 0.5685 | 0.801 |
| Specificity | 0.963 | 0.932 | 0.8976 | 0.9211 | 0.9483 |
| PosPredValue | 0.8578 | 0.7523 | 0.5962 | 0.6364 | 0.7889 |
| NegPredValue | 0.9787 | 0.9144 | 0.92 | 0.8978 | 0.9518 |
| Prevalence | 0.1964 | 0.2272 | 0.1865 | 0.1954 | 0.1944 |
| Detection Rate | 0.1796 | 0.1597 | 0.123 | 0.1111 | 0.1558 |
| Detection Prevalence | 0.2093 | 0.2123 | 0.2063 | 0.1746 | 0.1974 |
| Balanced Accuracy | 0.9386 | 0.8175 | 0.7786 | 0.7448 | 0.8746 |

The confusion matrix, shown in Figure 5, resulting from the model using all variables from ***df2*** shows that small forwards were mis-classified as each of the other 4 positions at least one time, with a majority of the small forward misclassifications coming as a result of being classified as a shooting guard or power forward. The accuracy value of 0.7292 means that 72.92% of the predictions were correctly classified. Considering that a player’s position is supposed to describe their on-court behavior, tracking statistics that capture all of a player’s movements should allow for accurate classification of most players with few outliers. However, here it can be seen that even after model adjustment, tuning, and fitting the highest accuracy that a random forest model can predict a player’s position is approximately 72.92%. Centers and Point Guards were classified as three possible positions, while the others were all classified as at least 4 positions. Point Guards and Centers had the highest positive prediction value, meaning that those two positions were predicted most accurately among all positions. This could be because each of those two positions have very defining traits about them or possibly because many of the players listed at shooting guard, small forward, or power forward are considered to be the same new, modern position group known as “wings”. While point guards and centers were the least mis-classified positions, roughly 85.78% of point guard predictions were correct and only 78.89% of center predictions were correct. While these numbers may not be calling for a redefinition of the current position system, small forward predictions were only correct 59.62% of the time. Predictions for shooting guard and power forward were also only correct less than

76% of the time, indicating that the biggest struggle for correctly predicting player positions comes from wings.

**Figure 6:** *Confusion matrix and model results for* ***df4*** *for a random forest model using all variables.*



Confusion Matrix: RF - df4

Prediction / Reference

| | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| PG | 181 | 29 | 0 | 0 | 0 |
| SG | 15 | 162 | 41 | 3 | 0 |
| SF | 2 | 38 | 119 | 41 | 2 |
| PF | 0 | 0 | 27 | 108 | 36 |
| C | 0 | 0 | 1 | 45 | 158 |

Overall Statistics:

| | |
|---|---|
| Accuracy | 0.7222 |
| 95% CI | (0.6935, 0.7497) |
| No Information Rate | 0.2272 |
| P-Value (Acc > NIR) | 2.20E-16 |

| | |
|---|---|
| Kappa | 0.6525 |
| Mcnemar's Test P-Value | NA |

Statistics By Class:

Class

| | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| Sensitivity | 0.9141 | 0.7074 | 0.633 | 0.5482 | 0.8061 |
| Specificity | 0.9642 | 0.9243 | 0.8988 | 0.9223 | 0.9433 |
| PosPredValue | 0.8619 | 0.733 | 0.5891 | 0.6316 | 0.7745 |
| NegPredValue | 0.9787 | 0.9149 | 0.9144 | 0.8937 | 0.9527 |
| Prevalence | 0.1964 | 0.2272 | 0.1865 | 0.1954 | 0.1944 |
| Detection Rate | 0.1796 | 0.1607 | 0.1181 | 0.1071 | 0.1567 |
| Detection Prevalence | 0.2083 | 0.2192 | 0.2004 | 0.1696 | 0.2024 |
| Balanced Accuracy | 0.9392 | 0.8158 | 0.7659 | 0.7353 | 0.8747 |

The results and confusion matrix for the second model created using ***df4*** are contained in Figure 6 above. The third model created using ***df4*** was slightly less accurate than the model using ***df2***, as the model using all variables from ***df4*** had an accuracy value of 0.7222, which means that 72.22% of all predictions were correctly classified, while the limited model had an accuracy of 0.7087, meaning that only 70.87% of all predictions were correctly classified.

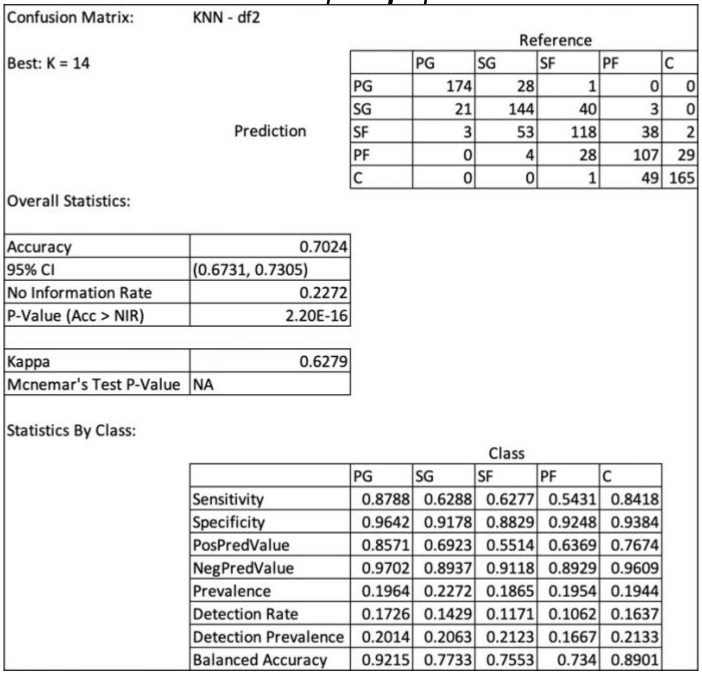
The successful prediction rate for point guards and centers in this model was slightly higher than the ***df2*** models, but it was less accurately able to classify small forwards, as the model correctly predicted 58.91% of its small forward classifications. The no-information rate is exactly the same meaning that no more information is lost/useless in this model than the one using ***df2***. The model using ***df4*** incorrectly mis-classified fewer small forwards as point guards and more small forwards as shooting guards than the model using ***df2***. Again, the wing players are the source of confusion when it comes to classification. This could potentially be happening because the 3 wing positions tend to be more balanced and focused on 2-way play, while the point guard is typically responsible for running the offense and the center responsible for anchoring the defense.

### b. K-Nearest Neighbors (KNN) Classification

K-Nearest Neighbors is a classification technique that predicts an object's classification based on other similar data points in close proximity. It calculates the distance between points and creates K number of classifications before calculating the probability of each observation being each of K classifications and then predicts it as the classification with the highest probability. For each of the 4 data frames, an initial model was created that had position as the variable of interest and the remaining variables were all used as predictors. Then, a second model was created that utilized a search grid and repeated cross-validation in order to help increase the accuracy of the results, while using the same model formula as the original. The search grid contains a number of different K's to try, ranging from 3 to 15 in increments of 1, to see which K produces the most accurate models. The models that were the most accurate at predicting player position were the data frames utilizing shot totals as opposed to percentages, ***df2*** and ***df4***. As a result, only ***df2*** and ***df4*** were used in making a third model that would feature fewer predictor variables. The third model would be created utilizing the same significantly important variables as each of the respective third random forest models. Again, the p-values for each version of the model for both data frames indicate that the results are statistically significant.

The KNN models at each step were less accurate than their respective random forest models using the same formula. This is to be expected because of the high dimensionality of the data and random forest models excelling at that. K-Nearest Neighbors allows for the number of K's to vary, which drastically impacts the accuracy of the predictions. For both ***df2*** and ***df4***, the second model that included all variables provided more accurate predictions than the model utilizing the important random forest variables. The confusion matrix and results for the second model with ***df2*** are pictured below in Figure 7, which shows the model with the highest accuracy had K=14. This means that the most accurate model utilized 14 different groups to be classified into.

**Figure 7:** *Confusion matrix and model results for* ***df2*** *for a KNN model using all variables.*



Confusion Matrix: KNN - df2

Best: K = 14

Reference

| Prediction | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| PG | 174 | 28 | 1 | 0 | 0 |
| SG | 21 | 144 | 40 | 3 | 0 |
| SF | 3 | 53 | 118 | 38 | 2 |
| PF | 0 | 4 | 28 | 107 | 29 |
| C | 0 | 0 | 1 | 49 | 165 |

Overall Statistics:

| | |
|---|---|
| Accuracy | 0.7024 |
| 95% CI | (0.6731, 0.7305) |
| No Information Rate | 0.2272 |
| P-Value (Acc > NIR) | 2.20E-16 |

| | |
|---|---|
| Kappa | 0.6279 |
| Mcnemar's Test P-Value | NA |

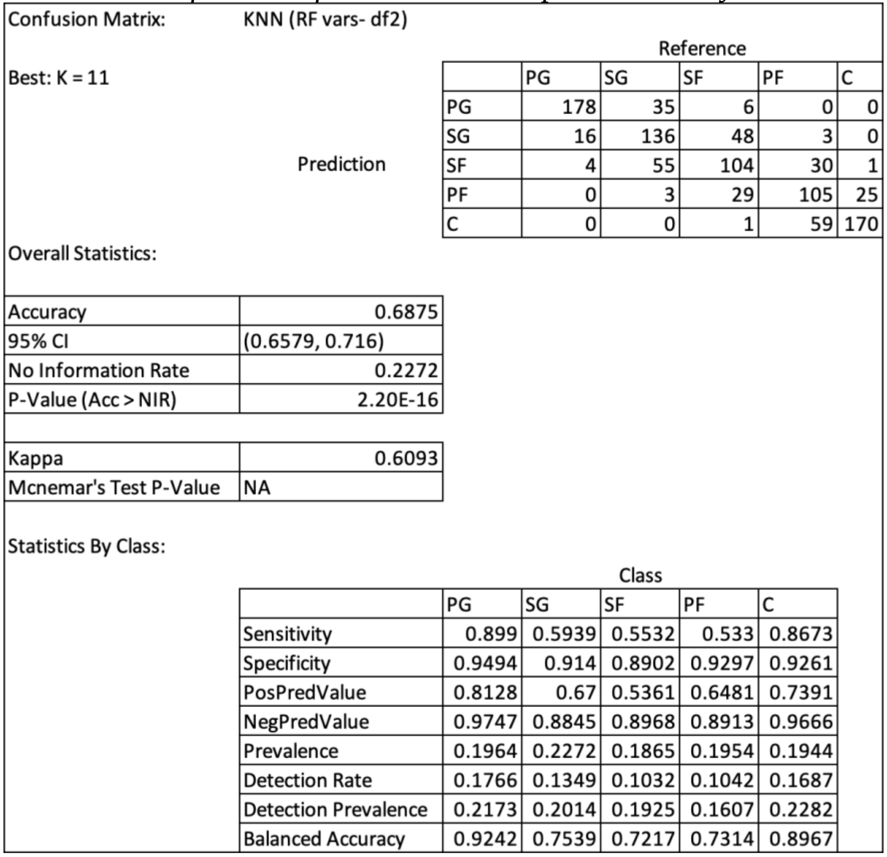
Statistics By Class:

Class

| | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| Sensitivity | 0.8788 | 0.6288 | 0.6277 | 0.5431 | 0.8418 |
| Specificity | 0.9642 | 0.9178 | 0.8829 | 0.9248 | 0.9384 |
| PosPredValue | 0.8571 | 0.6923 | 0.5514 | 0.6369 | 0.7674 |
| NegPredValue | 0.9702 | 0.8937 | 0.9118 | 0.8929 | 0.9609 |
| Prevalence | 0.1964 | 0.2272 | 0.1865 | 0.1954 | 0.1944 |
| Detection Rate | 0.1726 | 0.1429 | 0.1171 | 0.1062 | 0.1637 |
| Detection Prevalence | 0.2014 | 0.2063 | 0.2123 | 0.1667 | 0.2133 |
| Balanced Accuracy | 0.9215 | 0.7733 | 0.7553 | 0.734 | 0.8901 |

The KNN models really struggled with the predicting of small forwards, as they were significantly less accurate than the random forest models, while the models as a whole were less accurate than the random forest models. Figure 7 shows the KNN model using all tracking variables in ***df2*** has an accuracy of 0.7024, which means that only 70.24% of predictions are correctly classified. Again, point guards and centers had the highest accuracy on predictions, while also having the highest detection rates in this model. The second KNN model for ***df2*** also mis-classified small forwards as each of the other 4 positions similar to the random forest model. The KNN model featuring the important random forest variables had an accuracy of 0.6875, meaning that it managed to correctly classify the player's position 68.75% of the time, making it slightly less accurate than the second version. However, the results of this model, pictured below in Figure 8, show that the ideal number of K's is not 14, but 11. These models are showing a lot of the same things as the random forest models: lots of confusion between "close" positions and positions aren't seemingly representative of on-court production. The model using only the random forest variables actually misclassified 6 Small Forwards as Point Guards, a significantly higher amount than any of the other models so far.

**Figure 8:** *Confusion matrix and model results for **df2** for a KNN model using the variables found to be significant as a result of random forest variable importance analysis.*



Confusion Matrix: KNN (RF vars- df2)

Best: K = 11

| Prediction \ Reference | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| PG | 178 | 35 | 6 | 0 | 0 |
| SG | 16 | 136 | 48 | 3 | 0 |
| SF | 4 | 55 | 104 | 30 | 1 |
| PF | 0 | 3 | 29 | 105 | 25 |
| C | 0 | 0 | 1 | 59 | 170 |

Overall Statistics:

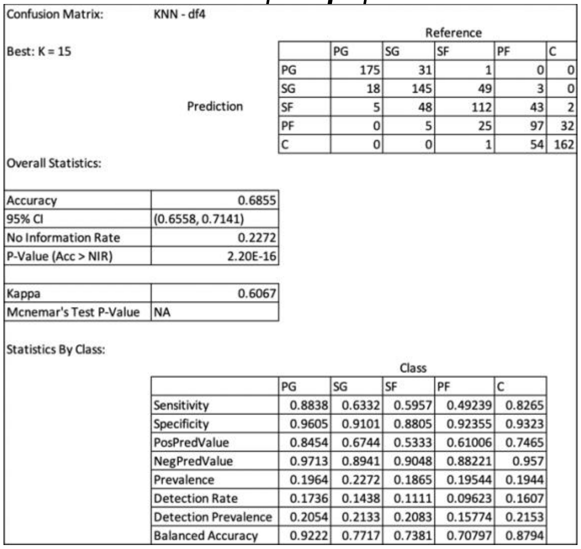
| | |
|---|---|
| Accuracy | 0.6875 |
| 95% CI | (0.6579, 0.716) |
| No Information Rate | 0.2272 |
| P-Value (Acc > NIR) | 2.20E-16 |
| Kappa | 0.6093 |
| Mcnemar's Test P-Value | NA |

Statistics By Class:

| Class | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| Sensitivity | 0.899 | 0.5939 | 0.5532 | 0.533 | 0.8673 |
| Specificity | 0.9494 | 0.914 | 0.8902 | 0.9297 | 0.9261 |
| PosPredValue | 0.8128 | 0.67 | 0.5361 | 0.6481 | 0.7391 |
| NegPredValue | 0.9747 | 0.8845 | 0.8968 | 0.8913 | 0.9666 |
| Prevalence | 0.1964 | 0.2272 | 0.1865 | 0.1954 | 0.1944 |
| Detection Rate | 0.1766 | 0.1349 | 0.1032 | 0.1042 | 0.1687 |
| Detection Prevalence | 0.2173 | 0.2014 | 0.1925 | 0.1607 | 0.2282 |
| Balanced Accuracy | 0.9242 | 0.7539 | 0.7217 | 0.7314 | 0.8967 |

All KNN models using ***df4*** indicated the ideal k=15, which means having 15 positional groups; however, this number continued to rise as the maximum possible K increased in the search grid until the number 20, which is not replicated by any of the other KNN models, including those using ***df3*** and ***df5***. Small forwards present an immense amount of trouble for the models to correctly classify, but shooting guards and power forwards provide plenty of trouble as well for this model.

**Figure 9:** *Confusion matrix and model results for **df4** for a KNN model using all variables.*



Confusion Matrix: KNN - df4

Best: K = 15

| Prediction \ Reference | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| PG | 175 | 31 | 1 | 0 | 0 |
| SG | 18 | 145 | 49 | 3 | 0 |
| SF | 5 | 48 | 112 | 43 | 2 |
| PF | 0 | 5 | 25 | 97 | 32 |
| C | 0 | 0 | 1 | 54 | 162 |

Overall Statistics:

| | |
|---|---|
| Accuracy | 0.6855 |
| 95% CI | (0.6558, 0.7141) |
| No Information Rate | 0.2272 |
| P-Value (Acc > NIR) | 2.20E-16 |
| Kappa | 0.6067 |
| Mcnemar's Test P-Value | NA |

Statistics By Class:

| Class | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| Sensitivity | 0.8838 | 0.6332 | 0.5957 | 0.49239 | 0.8265 |
| Specificity | 0.9605 | 0.9101 | 0.8805 | 0.92355 | 0.9323 |
| PosPredValue | 0.8454 | 0.6744 | 0.5333 | 0.61006 | 0.7465 |
| NegPredValue | 0.9713 | 0.8941 | 0.9048 | 0.88221 | 0.957 |
| Prevalence | 0.1964 | 0.2272 | 0.1865 | 0.19544 | 0.1944 |
| Detection Rate | 0.1736 | 0.1438 | 0.1111 | 0.09623 | 0.1607 |
| Detection Prevalence | 0.2054 | 0.2133 | 0.2083 | 0.15774 | 0.2153 |
| Balanced Accuracy | 0.9222 | 0.7717 | 0.7381 | 0.70797 | 0.8794 |

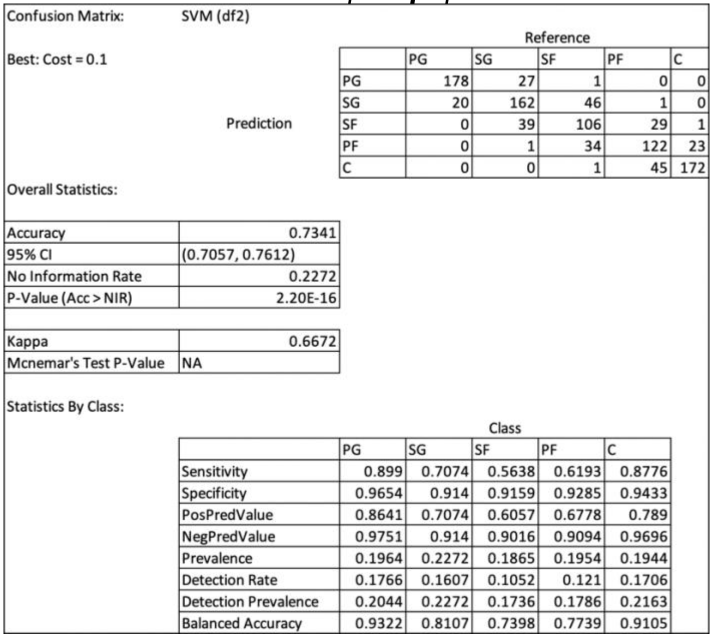
Of all 6 KNN models run, k=5 was never once in the top 5 of most accurate models. At the very least, it appears as if the overlap between small forward and shooting guards and the overlap between small forwards and power forwards warrants an expanded positional breakdown. Every model returned an ideal k >= 10, which re-enforces the idea that the current positional system is inadequate.

### c. Support Vector Machine (SVM) Classification

Support Vector Machine Classification is a method of classification that utilizes a hyperplane to determine the distance between two points before settling equidistant from each and operating as a decision boundary. At the decision boundary, it is decided if something is more like one classification than the other, and the process will continue until each observation is classified into a category based on its location to each decision boundary. SVM attempts to set the decision boundary as far away from each point as possible, leading to it usually being as much in the middle as possible.

The same process as used for Random Forest models and KNN models was utilized with SVMs. A simple model was created for each of the 4 datasets, and then a model utilizing tuning parameters and a search grid was created. The most important tuning parameter for SVMs is the cost, which is how much the model is penalized for misclassification with the smaller costs leaving more room for error and higher costs penalizing the model more heavily for misclassification. A third model was also run using the variables found to be important from the Random Forest variable analysis; however, these results directly matched or were less accurate than the second version of the model, for both ***df2*** and ***df4*** respectively.

**Figure 10:** *Confusion matrix and model results for **df2** for a SVM model using all variables.*



Confusion Matrix: SVM (df2)

Best: Cost = 0.1

Prediction / Reference

| | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| PG | 178 | 27 | 1 | 0 | 0 |
| SG | 20 | 162 | 46 | 1 | 0 |
| SF | 0 | 39 | 106 | 29 | 1 |
| PF | 0 | 1 | 34 | 122 | 23 |
| C | 0 | 0 | 1 | 45 | 172 |

Overall Statistics:

| | |
|---|---|
| Accuracy | 0.7341 |
| 95% CI | (0.7057, 0.7612) |
| No Information Rate | 0.2272 |
| P-Value (Acc > NIR) | 2.20E-16 |
| | |
| Kappa | 0.6672 |
| Mcnemar's Test P-Value | NA |

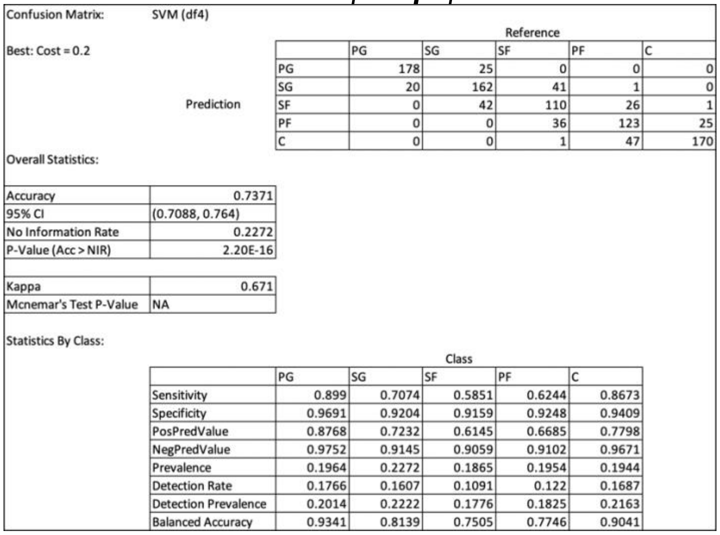
Statistics By Class:

Class

| | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| Sensitivity | 0.899 | 0.7074 | 0.5638 | 0.6193 | 0.8776 |
| Specificity | 0.9654 | 0.914 | 0.9159 | 0.9285 | 0.9433 |
| PosPredValue | 0.8641 | 0.7074 | 0.6057 | 0.6778 | 0.789 |
| NegPredValue | 0.9751 | 0.914 | 0.9016 | 0.9094 | 0.9696 |
| Prevalence | 0.1964 | 0.2272 | 0.1865 | 0.1954 | 0.1944 |
| Detection Rate | 0.1766 | 0.1607 | 0.1052 | 0.121 | 0.1706 |
| Detection Prevalence | 0.2044 | 0.2272 | 0.1736 | 0.1786 | 0.2163 |
| Balanced Accuracy | 0.9322 | 0.8107 | 0.7398 | 0.7739 | 0.9105 |

The SVM models were the first models to not predict any point guards to be small forwards; however, the results for the model using ***df2*** in Figure 10 show that a center was mis-classified as a small forward. The second SVM model using ***df2*** shows that a cost of 0.1 is best, which means not heavily penalizing the model for misclassifications. Yet again, point guards and centers were predicted correctly with the highest frequency with a significant gap between them and the other positions, while small forwards were predicted correctly with the lowest frequency. The accuracy value of 0.7341 means that 73.41% of all classifications were correct. Again, even after using model tuning and accuracy, the SVM models correctly classify the player's position less than 75% of the time with lots of confusion with regards to the wing players. Point guards were correctly classified 86.41% of the time, which is the highest accuracy that has been returned for any of the positions in any of the models, but the positive detection rate for Small Forwards is still only 0.6057, meaning that only 60.57% of Small Forwards are properly classified.

**Figure 11:** *Confusion matrix and model results for* ***df4*** *for a SVM model using all variables.*



Confusion Matrix: SVM (df4)

Best: Cost = 0.2

Reference

| Prediction | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| PG | 178 | 25 | 0 | 0 | 0 |
| SG | 20 | 162 | 41 | 1 | 0 |
| SF | 0 | 42 | 110 | 26 | 1 |
| PF | 0 | 0 | 36 | 123 | 25 |
| C | 0 | 0 | 1 | 47 | 170 |

Overall Statistics:

| | |
|---|---|
| Accuracy | 0.7371 |
| 95% CI | (0.7088, 0.764) |
| No Information Rate | 0.2272 |
| P-Value (Acc > NIR) | 2.20E-16 |

| | |
|---|---|
| Kappa | 0.671 |
| Mcnemar's Test P-Value | NA |

Statistics By Class:

Class

| | PG | SG | SF | PF | C |
|---|---|---|---|---|---|
| Sensitivity | 0.899 | 0.7074 | 0.5851 | 0.6244 | 0.8673 |
| Specificity | 0.9691 | 0.9204 | 0.9159 | 0.9248 | 0.9409 |
| PosPredValue | 0.8768 | 0.7232 | 0.6145 | 0.6685 | 0.7798 |
| NegPredValue | 0.9752 | 0.9145 | 0.9059 | 0.9102 | 0.9671 |
| Prevalence | 0.1964 | 0.2272 | 0.1865 | 0.1954 | 0.1944 |
| Detection Rate | 0.1766 | 0.1607 | 0.1091 | 0.122 | 0.1687 |
| Detection Prevalence | 0.2014 | 0.2222 | 0.1776 | 0.1825 | 0.2163 |
| Balanced Accuracy | 0.9341 | 0.8139 | 0.7505 | 0.7746 | 0.9041 |

The performance between ***df2*** and ***df4*** was even in terms of accuracy when it comes to SVMs, but the model using ***df4*** had a slightly higher cost, meaning the model was penalized more for misclassifications. This model had a ton of overlap in determining the position of shooting guards and small forwards, which fits along with the other models. This model had the highest positive prediction rate for Small Forwards at 0.6145, meaning that at best only 61.45% of Small Forwards could be properly classified based on on-court tracking statistics.

While for the random forest models ***df2***, featuring offense/defense splits and shot totals, fared better than ***df4***, featuring two-way averages and shot totals, in terms of total accuracy. The SVM models for ***df2*** and ***df4*** ended up being the most accurate of the supervised machine learning algorithms, but even with the adjustments, tuning parameters, and variable importance analysis, the algorithm can only predict a player's position with at best roughly 73.71% accuracy. This means that about ¼ of the league's players are classified as the wrong position based on what they actually do on the court. This sparks the questions: how many positions are there? And does age impact what position a player is capable of playing?

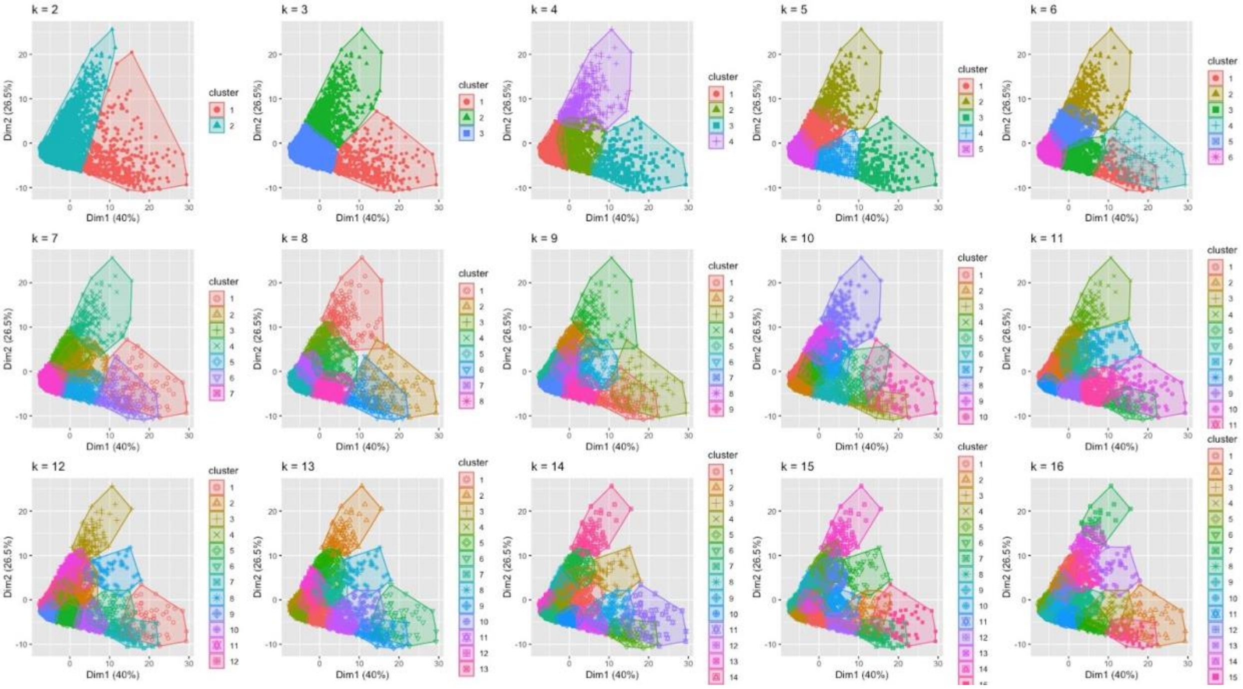
## B. Position-Age Model

In the hypothetical scenario of redefining the NBA's positional system to accurately describe a player's role on the court, determining the number of positions to set is the top priority. K-Nearest Neighbors classification offers up how accurate the model is dependent upon the number of potential classifications, and those results indicated that a number greater than 5 must be the answer. Clustering techniques will help provide the assignments for players at the hypothetical new positions. Then the new positional assignments will be substituted for the old ones. A linear model will then be used to determine what effects age has alongside the tracking statistics on the classification of the new positions. Effectively, new positions will be assigned using clustering techniques and then a linear model will be run to determine the impacts of age on the new, hypothetical positions.

### a. K-Means Clustering

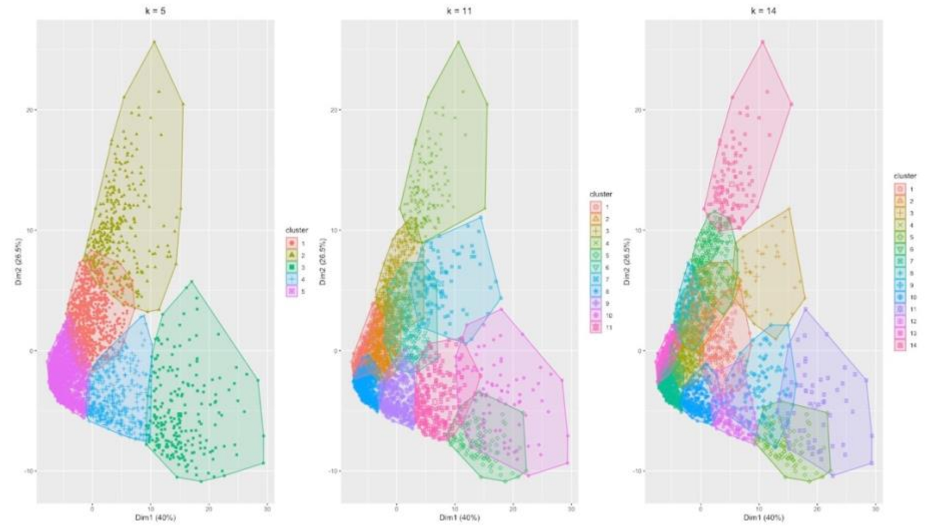
K-Means Clustering is very similar to KNN in principle, but it is an unsupervised machine learning technique, so it identifies the patterns between the numbers/data points on its own without knowing the variable of interest. This is useful for creating new positions as observations with similar variable measurements will automatically be grouped together, so players who have tracking statistics similar to each other will be assigned the same new position.

**Figure 12:** *Graphs of K-Means Clustering results with range in K from 2 to 16.*



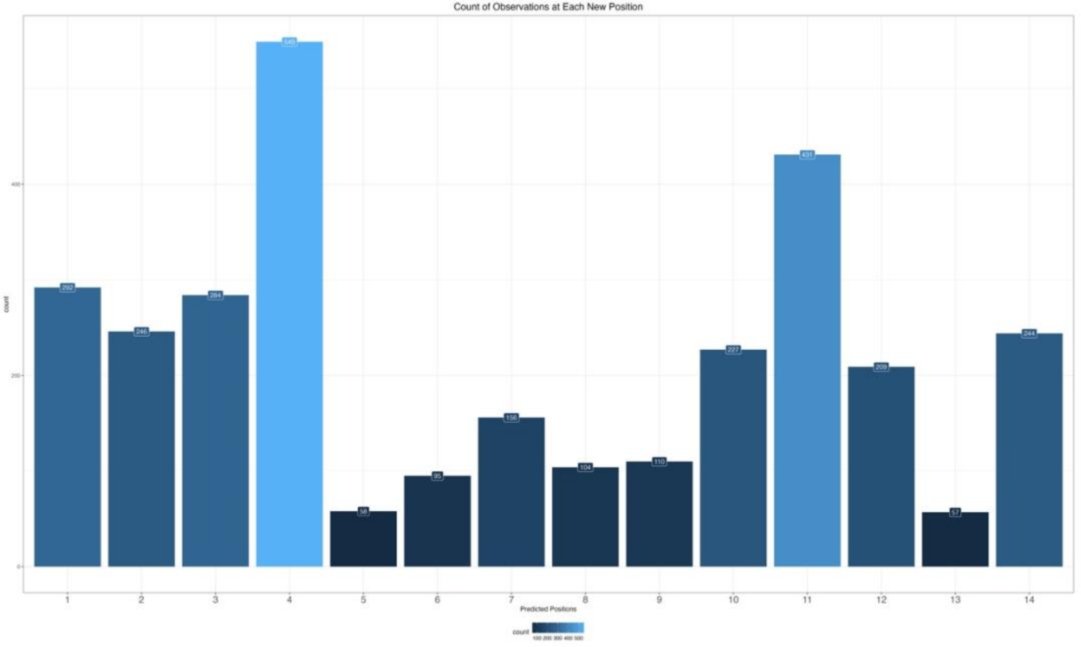
While a smaller K may make the separation between the clusters easier to see, it also produces clusters with a wider variation. K=3 is the highest possible K that has no overlap, but that's a smaller number of positions than currently exists in the NBA. Figure 12 shows the graphs of K-Means clustering results using ***df2***. When using ***df2*** with a KNN model, it indicated that the ideal K should be either 11 or 14 depending on the model used. So Figure 13 below highlights the graphs that have K=5, the current number of positions, in addition to K=11 and K=14, which are the two most likely potential numbers of clusters. The graphs utilize dimensionality reduction to reduce the data to two variables that can account for roughly 66.5% of the variation in the data. It is important to note that the dimensionality reduction is done for the purpose of the graph, but the model doesn't actually need it simplified to just two dimensions in order to run.

**Figure 13:** *Graph of K-Means Clustering results when K=5, K=11, and K=14 respectively.*



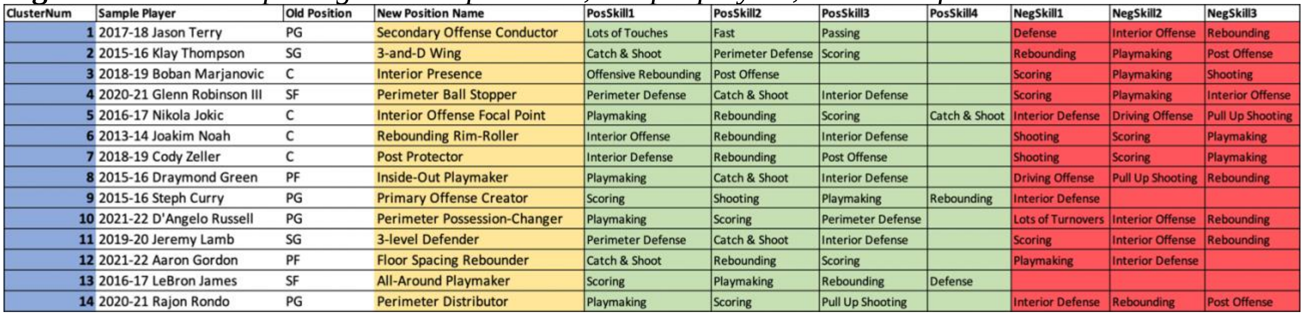
While there is not a lot of overlap between clusters in the graph for K=5, there is also a wide variation within each cluster. While both K=11 and K=14 are comparable in terms of complexity, K=14 offers a cluster split at a very congested area by the coordinates (10, -5), which could prove to be significant in differentiating between similar style positions like small forwards and shooting guards. Considering that both the KNN models were statistically significant, K=14 was slightly more accurate than K=11, and the splits happening at points of great congestion, the ideal number of clusters appears to be 14. This means that in order to most efficiently have positions that describe a player's actual actions on the court, the NBA should switch to a 14-position system. The mean of each variable for each cluster is calculated when they are clustered and using those measurements a name or position title can be assigned to each of the clusters. Statistics that are significant to each cluster can be identified and used to help create names for each of the positions, but for the sake of the linear models the assigned cluster numbers should suffice.

**Figure 14:** *Bar graph showing breakdown of observations across the new positions created as a result of clustering.*



The distribution of players among these new positions is not nearly as even as it was under the 5-man system, but that makes sense given that there is significant differentiation between players that would play the same position. Positions 5 and 13 feature the smallest number of observations by a significant amount, but those two positional groups contain observations like 2020-21 Nikola Jokic, a season in which he won MVP, and 2017-18 LeBron James, respectively. Each of those players played at a level that very few players are capable of achieving, which means that there will be fewer players with similar statistics. While one may argue that other individuals can play a style similar to LeBron, just far less efficiently and effectively, that player is going to impact the game in an entirely different way. Two players can both be good 3-point shooters, and one excels at pull up jumpers and the other excels at catch and shoot jumpers. These two players have different expectations, roles, and responsibilities, so they will move and interact differently on the court, and if positions are supposed to describe a player's actions/roles/responsibilities on the court then these players must be different positions.

**Figure 15:** *Table depicting the new positions, sample players, and skills for each.*



| ClusterNum | Sample Player | Old Position | New Position Name | PosSkill1 | PosSkill2 | PosSkill3 | PosSkill4 | NegSkill1 | NegSkill2 | NegSkill3 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2017-18 Jason Terry | PG | Secondary Offense Conductor | Lots of Touches | Fast | Passing | | Defense | Interior Offense | Rebounding |
| 2 | 2015-16 Klay Thompson | SG | 3-and-D Wing | Catch & Shoot | Perimeter Defense | Scoring | | Rebounding | Playmaking | Post Offense |
| 3 | 2018-19 Boban Marjanovic | C | Interior Presence | Offensive Rebounding | Post Offense | | | Scoring | Playmaking | Shooting |
| 4 | 2020-21 Glenn Robinson III | SF | Perimeter Ball Stopper | Perimeter Defense | Catch & Shoot | Interior Defense | | Scoring | Playmaking | Interior Offense |
| 5 | 2016-17 Nikola Jokic | C | Interior Offense Focal Point | Playmaking | Rebounding | Scoring | Catch & Shoot | Interior Defense | Driving Offense | Pull Up Shooting |
| 6 | 2013-14 Joakim Noah | C | Rebounding Rim-Roller | Interior Offense | Rebounding | Interior Defense | | Shooting | Scoring | Playmaking |
| 7 | 2018-19 Cody Zeller | C | Post Protector | Interior Defense | Rebounding | Post Offense | | Shooting | Scoring | Playmaking |
| 8 | 2015-16 Draymond Green | PF | Inside-Out Playmaker | Playmaking | Catch & Shoot | Interior Defense | | Driving Offense | Pull Up Shooting | Rebounding |
| 9 | 2015-16 Steph Curry | PG | Primary Offense Creator | Scoring | Shooting | Playmaking | Rebounding | Interior Defense | | |
| 10 | 2021-22 D'Angelo Russell | PG | Perimeter Possession-Changer | Playmaking | Scoring | Perimeter Defense | | Lots of Turnovers | Interior Offense | Rebounding |
| 11 | 2019-20 Jeremy Lamb | SG | 3-level Defender | Perimeter Defense | Catch & Shoot | Interior Defense | | Scoring | Interior Offense | Rebounding |
| 12 | 2021-22 Aaron Gordon | PF | Floor Spacing Rebounder | Catch & Shoot | Rebounding | Scoring | | Playmaking | Interior Defense | |
| 13 | 2016-17 LeBron James | SF | All-Around Playmaker | Scoring | Playmaking | Rebounding | Defense | | | |
| 14 | 2020-21 Rajon Rondo | PG | Perimeter Distributor | Playmaking | Scoring | Pull Up Shooting | | Interior Defense | Rebounding | Post Offense |

Each cluster was analyzed to see where the center of the cluster was for each of the statistics and how they compared relative to the other clusters. Figure 15 shows a table depicting the number that each cluster was assigned, the traditional position associated with players in that cluster, a sample player to highlight the prototype for each position, and up to 4 traits/statistics that they are good or bad at to highlight the differences between each of the positions. Figure 15 shows that 8 of the new position groups are created by breaking up the existing categories of point guards and centers, with 4 coming from each. This fits along with the evolution of the league as these two positions have experienced the most change and growth, as Point Guards have improved their ability to score inside and rebound, like Russell Westbrook, while Centers have begun to handle the ball and create offense for the whole team, like Nikola Jokic. The positional names were determined based on the skills that stood out for each player, as well as incorporating in factors like efficiency, opportunities, and minutes played.

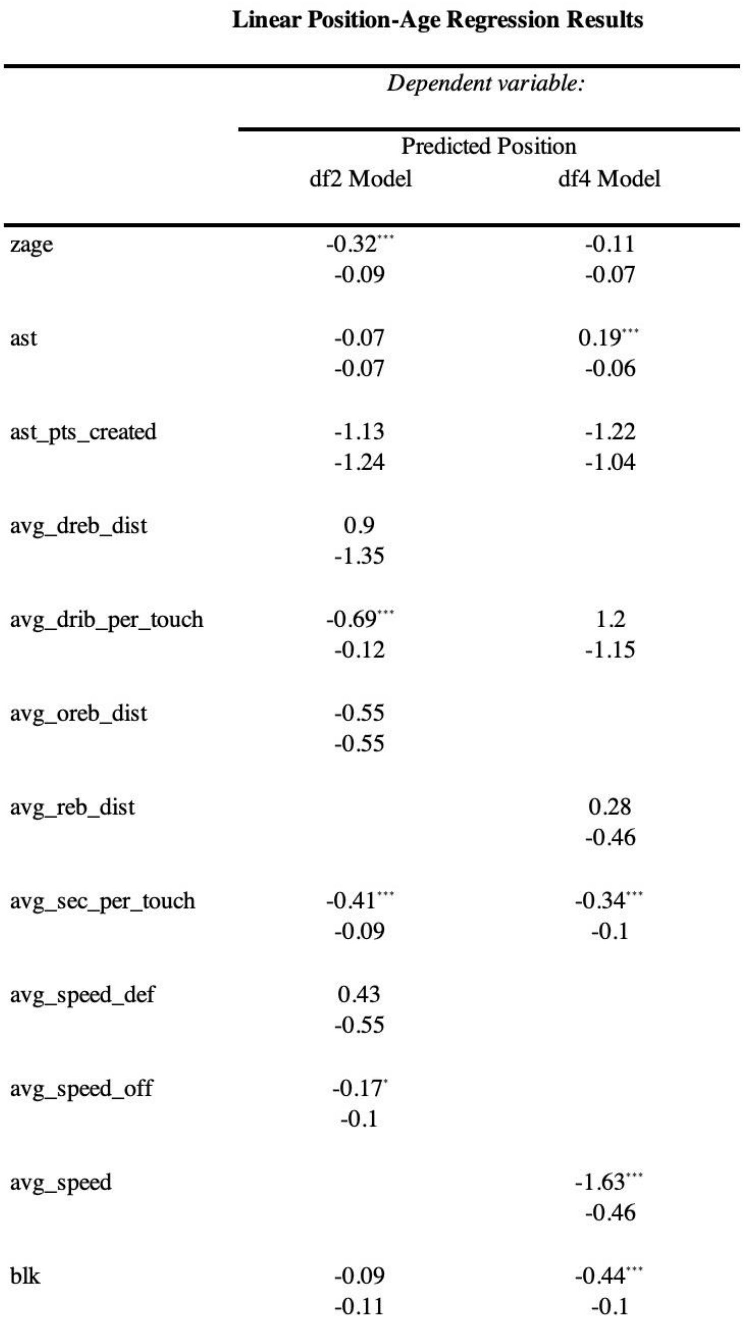
For the most part, players who had multiple seasons as observations had all of their seasons placed into the same clusters with the exceptions being players that experienced injuries that only cost them part of a season, players that changed settings, and players that experienced great changes in skill Players suffering an injury that only cost part of a season is one of the issues resulting from using stat totals, as opposed to rate stats. However, the information provided by the totals allows for a complete evaluation of production, as the amount of time spent on the court and number of opportunities also count as information added. Stat totals provide more information than rate stats in this case because they allow for an analysis of a player's production at scale. For example, take 2 players that average 1 point per minute. If player 1 plays 37 minutes per game and Player 2 plays just 6 minutes per game, the impact they had on the outcome is significantly different, so the scale at which a player performs matters in differentiating their on-court actions and capabilities. While Secondary Offense Conductors and

Perimeter Possession-Changers may be similar, there is a notable difference on the scale at which they perform. Perimeter Possession-Changers produce higher stat totals at similar levels of efficiency to Secondary Offense Conductors, who play fewer minutes and typically do it against an opposing team's bench lineup. Stat totals highlight the differences between players who produce on a large scale and those who produce in fewer minutes.

### b. Linear Model

Now that clustering and classification techniques have been used in order to determine the new number of positions and how they should be assigned based on the tracking information, a linear analysis of whether age impacts a player's style of play can be done. Due to the arbitrary position number given to each cluster, the orientation of the coefficients isn't particularly meaningful from the linear model, but whether or not a variable is significant is. The linear model using ***df2*** returned that age was significant, while the model using ***df4*** indicated that age was not significant.

**Figure 16:** *Linear Model Results.*



**Linear Position-Age Regression Results**

| | *Dependent variable:* | |
|---|---|---|
| | Predicted Position | |
| | df2 Model | df4 Model |
| zage | -0.32$^{***}$ | -0.11 |
| | -0.09 | -0.07 |
| ast | -0.07 | 0.19$^{***}$ |
| | -0.07 | -0.06 |
| ast_pts_created | -1.13 | -1.22 |
| | -1.24 | -1.04 |
| avg_dreb_dist | 0.9 | |
| | -1.35 | |
| avg_drib_per_touch | -0.69$^{***}$ | 1.2 |
| | -0.12 | -1.15 |
| avg_oreb_dist | -0.55 | |
| | -0.55 | |
| avg_reb_dist | | 0.28 |
| | | -0.46 |
| avg_sec_per_touch | -0.41$^{***}$ | -0.34$^{***}$ |
| | -0.09 | -0.1 |
| avg_speed_def | 0.43 | |
| | -0.55 | |
| avg_speed_off | -0.17$^{*}$ | |
| | -0.1 | |
| avg_speed | | -1.63$^{***}$ |
| | | -0.46 |
| blk | -0.09 | -0.44$^{***}$ |
| | -0.11 | -0.1 |

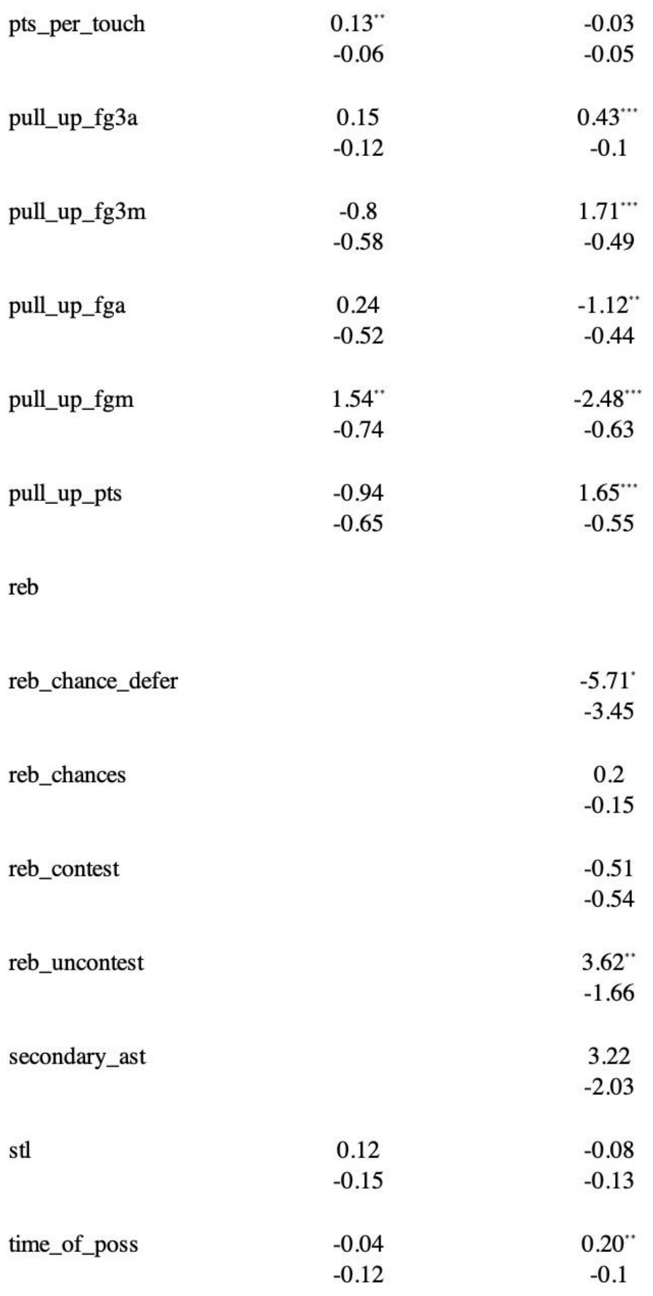
| | | |
|---|---|---|
| dist_miles | | |
| drive_ast | -4.14 | $3.96^{***}$ |
| | -2.83 | -0.91 |
| drive_fga | 0.1 | 0.22 |
| | -0.35 | -0.29 |
| drive_fgm | $-1.70^{**}$ | $1.22^{*}$ |
| | -0.84 | -0.71 |
| drive_fta | 4.27 | -0.47 |
| | -3.2 | -2.71 |
| drive_ftm | 0.19 | -1.08 |
| | -1.84 | -1.56 |
| drive_passes | 0.33 | 0.25 |
| | -1.05 | -0.89 |
| drive_pf | $-1.06^{*}$ | -0.16 |
| | -0.55 | -0.46 |
| drive_pts | 0.4 | 0.96 |
| | -1.82 | -1.54 |
| drive_tov | -5.42 | -0.16 |
| | -3.99 | -3.38 |
| drives | -0.29 | $0.31^{*}$ |
| | -0.22 | -0.19 |
| elbow_touch_ast | $3.53^{**}$ | -1.38 |
| | -1.48 | -1.26 |
| elbow_touch_fga | 0.08 | -0.15 |
| | -0.19 | -0.16 |
| elbow_touch_fgm | -2.05 | -0.49 |
| | -1.33 | -1.13 |

| | | |
|---|---|---|
| dist_miles | | |
| drive_ast | -4.14 | $3.96^{***}$ |
| | -2.83 | -0.91 |
| drive_fga | 0.1 | 0.22 |
| | -0.35 | -0.29 |
| drive_fgm | $-1.70^{**}$ | $1.22^{*}$ |
| | -0.84 | -0.71 |
| drive_fta | 4.27 | -0.47 |
| | -3.2 | -2.71 |
| drive_ftm | 0.19 | -1.08 |
| | -1.84 | -1.56 |
| drive_passes | 0.33 | 0.25 |
| | -1.05 | -0.89 |
| drive_pf | $-1.06^{*}$ | -0.16 |
| | -0.55 | -0.46 |
| drive_pts | 0.4 | 0.96 |
| | -1.82 | -1.54 |
| drive_tov | -5.42 | -0.16 |
| | -3.99 | -3.38 |
| drives | -0.29 | $0.31^{*}$ |
| | -0.22 | -0.19 |
| elbow_touch_ast | $3.53^{**}$ | -1.38 |
| | -1.48 | -1.26 |
| elbow_touch_fga | 0.08 | -0.15 |
| | -0.19 | -0.16 |
| elbow_touch_fgm | -2.05 | -0.49 |
| | -1.33 | -1.13 |

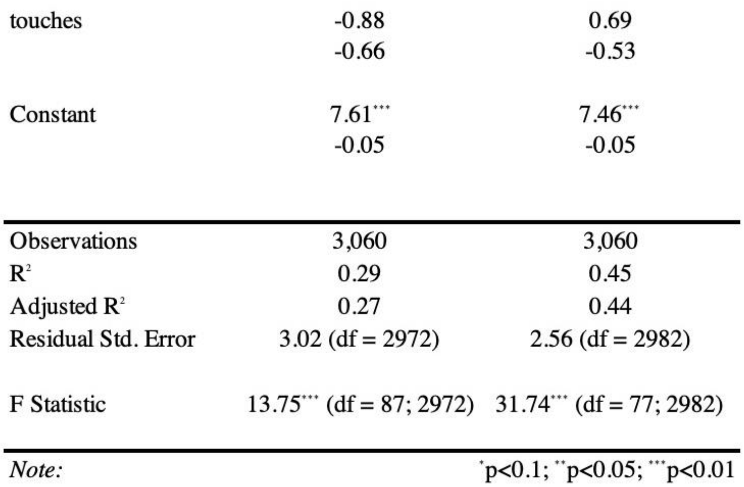
| | | |
|---|---|---|
| elbow_touch_fouls | 1.31 | 2.19 |
| | -2.32 | -1.96 |
| elbow_touch_fta | 0.26 | -0.08 |
| | -0.49 | -0.42 |
| elbow_touch_ftm | 0.36 | -0.31 |
| | -0.63 | -0.54 |
| elbow_touch_passes | -0.39 | 0.56 |
| | -0.58 | -0.49 |
| elbow_touch_pts | -1.47 | -0.88 |
| | -2.08 | -1.76 |
| elbow_touch_tov | -0.71 | -2.98 |
| | -2.67 | -2.26 |
| elbow_touches | -0.11 | -0.12 |
| | -0.26 | -0.22 |
| front_ct_touches | 1.96 | 2.14 |
| | -3.56 | -3.01 |
| ft_ast | -0.12 | -0.99$^{***}$ |
| | -0.42 | -0.34 |
| min | -0.08 | -0.02 |
| | -0.14 | -0.12 |
| oreb | 0.61 | |
| | -1.17 | |
| oreb_chance_defer | 3.61$^{*}$ | |
| | -2.15 | |
| oreb_chances | 0.01 | |
| | -0.1 | |
| oreb_contest | -0.98$^{**}$ | |
| | -0.39 | |

| | | |
|---|---|---|
| oreb_uncontest | -2.61 | |
| | -1.63 | |
| paint_touch_ast | -1.08$^{*}$ | -1.90$^{**}$ |
| | -0.61 | -0.96 |
| paint_touch_fga | 0.11 | 0.15 |
| | -0.16 | -0.13 |
| paint_touch_fgm | 0.51 | -0.34 |
| | -1.92 | -1.62 |
| paint_touch_fouls | 3.86 | -7.25$^{**}$ |
| | -4.19 | -3.54 |
| paint_touch_fta | -0.11 | -0.39 |
| | -0.76 | -0.64 |
| paint_touch_ftm | 0.33 | -1.15 |
| | -0.95 | -0.81 |
| paint_touch_passes | 1.17 | -0.46 |
| | -0.86 | -0.73 |
| paint_touch_pts | -0.73 | -0.16 |
| | -0.81 | -0.68 |
| paint_touch_tov | -6.77 | 7.59$^{*}$ |
| | -4.79 | -4.05 |
| paint_touches | -0.38 | -0.02 |
| | -0.25 | -0.21 |
| passes_made | 2.63 | 0.5 |
| | -3.09 | -2.6 |
| passes_received | 1.84$^{***}$ | -0.42 |
| | -0.44 | -0.36 |
| points | -1.15 | 0.89 |
| | -0.76 | -0.64 |

| | | |
|---|---|---|
| post_touch_ast | 0.8 | -1.11$^{**}$ |
| | -0.65 | -0.53 |
| post_touch_fga | 0.06 | 0.01 |
| | -0.24 | -0.2 |
| post_touch_fgm | 1.2 | -0.86 |
| | -1.08 | -0.91 |
| post_touch_fouls | -5.04 | -1.94 |
| | -3.53 | -2.99 |
| post_touch_fta | 1.93 | 0.48 |
| | -1.54 | -1.3 |
| post_touch_ftm | -2.47 | 0.83 |
| | -1.58 | -1.34 |
| post_touch_passes | -1.02 | -1.40$^{*}$ |
| | -1 | -0.85 |
| post_touch_pts | 1.01$^{*}$ | -0.5 |
| | -0.59 | -0.5 |
| post_touch_tov | 6.52 | 2.79 |
| | -4.27 | -3.61 |
| post_touches | 0.18 | 0.04 |
| | -0.24 | -0.2 |
| potential_ast | -3.08$^{*}$ | 1.43 |
| | -1.74 | -1.47 |
| pts_per_elbow_touch | 0.002 | 0.14 |
| | -0.7 | -0.59 |
| pts_per_paint_touch | -0.03 | 0.05 |
| | -0.07 | -0.06 |
| pts_per_post_touch | 0.19$^{**}$ | 0.24$^{***}$ |
| | -0.08 | -0.07 |



| | | |
|---|---|---|
| pts_per_touch | 0.13$^{**}$ | -0.03 |
| | -0.06 | -0.05 |
| pull_up_fg3a | 0.15 | 0.43$^{***}$ |
| | -0.12 | -0.1 |
| pull_up_fg3m | -0.8 | 1.71$^{***}$ |
| | -0.58 | -0.49 |
| pull_up_fga | 0.24 | -1.12$^{**}$ |
| | -0.52 | -0.44 |
| pull_up_fgm | 1.54$^{**}$ | -2.48$^{***}$ |
| | -0.74 | -0.63 |
| pull_up_pts | -0.94 | 1.65$^{***}$ |
| | -0.65 | -0.55 |
| reb | | |
| reb_chance_defer | | -5.71$^{*}$ |
| | | -3.45 |
| reb_chances | | 0.2 |
| | | -0.15 |
| reb_contest | | -0.51 |
| | | -0.54 |
| reb_uncontest | | 3.62$^{**}$ |
| | | -1.66 |
| secondary_ast | | 3.22 |
| | | -2.03 |
| stl | 0.12 | -0.08 |
| | -0.15 | -0.13 |
| time_of_poss | -0.04 | 0.20$^{**}$ |
| | -0.12 | -0.1 |



| | | |
|---|---|---|
| touches | -0.88 | 0.69 |
| | -0.66 | -0.53 |
| | | |
| Constant | $7.61^{***}$ | $7.46^{***}$ |
| | -0.05 | -0.05 |
| | | |
| Observations | 3,060 | 3,060 |
| $R^2$ | 0.29 | 0.45 |
| Adjusted $R^2$ | 0.27 | 0.44 |
| Residual Std. Error | 3.02 (df = 2972) | 2.56 (df = 2982) |
| | | |
| F Statistic | $13.75^{***}$ (df = 87; 2972) | $31.74^{***}$ (df = 77; 2982) |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

While not all of the variables are statistically significant in both models, there are several that stand out for key reasons. The model using ***df4*** was able to account for more of the deviation in the data, as the r-squared value for ***df4*** is 0.45, which means that the data can account for 45% of the variation when it comes to the hypothetical position, while the model using ***df2*** only had an r-squared value of 0.29, meaning that only 29% of the variation in the hypothesized position can be explained by the data. The variables that seemingly had the most impact across both models tended to be related to a player's driving capabilities, post offense, touch numbers, and assists. This makes sense as a lot of the bigger cluster differentiations are going to come at these stats, because offensively gifted players will have more assists and touches than defense-oriented players, while players that are gifted at scoring in the post are likely different than slashers capable of scoring off drives in the paint. In terms of the traditional NBA position system, guards should have lots of assist and touches, while big men should do better in the post. There are going to be important key distinctions though, as not every guard who touches the ball does so in the same way.

Given that the model using ***df4*** had a higher r-squared, age may not be significant on the style of a player's play, but it may be more individually based. This could be backed up by players starting to defy Father Time and find success late in their career as LeBron James is finding success by dominating the NBA at age 38, and players like Kobe Bryant, Tim Duncan, and Dirk Nowitzki all continued to make an impact on the court until they retired at ages 37, 39, and 40, respectively. As training, rehabilitation, and injury prevention methods continue to develop, player's will have greater longevity when it comes to their careers, meaning age has less of an impact on their style of play.

## *5. Conclusion*

The NBA has come a long way since the initial creation of its positional system, and if a player's position is supposed to describe their on-court actions, then the machine learning models above show that it does not do that for roughly ¼ of the league. While it may not be enough to warrant completely overwriting the positional system of the league, there is definitely information to be gained from having more accurately fitting positional groups. Another potential option could be choosing a smaller number of positions like 8 or 11 to make the re-arrangement less significant. Even given the 14 new positions, there are significant differences between each of the groups, regardless of how similar they may seem. Perimeter Distributors and Primary Offense Creators may sound similar and may even do MOST things similar on the court; however, Primary Offense Creators are excellent at shooting off the dribble or on passes, and are far better at scoring, particularly in the paint, while Perimeter Distributors are effective at pull up shooting and tend to be more balanced between their scoring and playmaking. Another two positions that seem similar are Interior Presence and Post Protector, but there are some notable differentiations between the two, including that Post Protectors are better at rebounding and rim defense, while having some semblance of an offensive game in the post. An Interior Presence is just a player who spends his time near the basket without making too significant an impact in any one area except offensive rebounding. These slight differences both in name and responsibility will allow teams to create more efficient lineups, as they could address weaknesses by inserting a player that provides different on-court abilities, despite being the same listed position. It's important to note that despite physical characteristics like height and weight being left out, many of the players in each cluster have similar builds/body types.

As the game of basketball continues to evolve, these new positions will have to be updated, but ideally these should prove to be serviceable until the game experiences a groundbreaking change, like the addition of the 3-point line. In the future, I would like to analyze line-up combinations and their performance utilizing these positions in order to see which positions perform best together, as each contributes different things on the court. Seeing whether lineups featuring certain combinations of players fare better together would prove quite interesting.

The 2016-17 Warriors' Death Lineup featuring Steph Curry, Klay Thompson, Kevin Durant, Andre Iguodala, and Draymond Green featured a perimeter offense creator, a 3-and-D wing, an All-Around Playmaker, a Floor-Spacing Rebounder, and an Inside-Out Playmaker, respectively. This backs up something already widely known about this line-up: it had a deadly combination of offense and shooting that was paired with just the right balance of defense and rebounding. Very few teams historically have a talent level similar to the 2016-17 Warriors, so for most other teams a lineup's performance is far more dependent on fit than pure talent. As the 2022-23 Dallas Mavericks proved, chemistry matters more than talent and finding the right pieces that fit together is difficult using the current positional system. While changing the names of player's positions won't have much of an impact on the individual's style of play, it may allow teams a more accurate way to evaluate their roster and potential line-ups.

## References

Abdelkrim, B. (2010, May). *Positional Role and Competitive-Level Differences in. . . : The Journal of Strength & Conditioning Research*. LWW. Retrieved December 4, 2022, from https://journals.lww.com/nsca-jscr/Fulltext/2010/05000/Positional_Role_and_Competitive_Level_Differences.27.aspx

Alagappan, M. (2012). *From 5 to 13*. Retrieved November 16, 2022, from https://web.math.utk.edu/~fernando/Students/GregClark/pdf/Alagappan-Muthu-EOSMarch2012PPT.pdf

An, M., Liang, E., & Zhang, M. (2016). Predicting the Trajectory of an NBA Player's Career. *Stanford University*. http://cs229.stanford.edu/proj2016/report/AnLiangZhang-PredictingTheTrajectoryOfAnNBAPlayer%27sCareer-report.pdf

Bianchi, F., Facchinetti, T., & Zuccolotto, P. (2017). *Role revolution: towards a new meaning of positions in basketball | Bianchi | Electronic Journal of Applied Statistical Analysis*. Electronic Journal of Applied Statistical Analysis. Retrieved November 16, 2022, from http://siba-ese.unisalento.it/index.php/ejasa/article/view/16464

Bosch, J., & Kalman, S. (2020, March 6). *NBA Lineup Analysis on Clustered Player Tendencies: A new approach to the positions of basketball & modeling lineup efficiency of soft lineup aggregates*. MIT Sloan Sports Analytics Conference. Retrieved November 16, 2022, from https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/5f6a65517f9440891b8e35d0_Kalman_NBA_Line_up_Analysis.pdf

Bruce, S. (2016). A scalable framework for NBA player and team comparisons using player tracking data. *Journal of Sports Analytics*, *2*(2), 107–119. https://doi.org/10.3233/jsa-160022

Chang, Y., Maheswaran, R., Su, J., Kwok, S., Levy, T., Wexler, A., & Squire, K. (2014, February 28). *Quantifying Shot Quality in the NBA - PDF Free Download*. Retrieved December 4, 2022, from https://docplayer.net/52631681-Quantifying-shot-quality-in-the-nba.html

Cheema, A. (2020, June 4). *Using Machine Learning to Classify NBA Players, Part I*. The Spax. https://www.thespax.com/nba/using-machine-learning-to-classify-nba-players-part-i/

Cheng, A. (2018, May 23). *Using Machine Learning to Find the 8 Types of Players in the NBA*. Medium. https://medium.com/fastbreak-data/classifying-the-modern-nba-player-with-machine-learning-539da03bb824

Cui, Y., Liu, F., Bao, D., Liu, H., & Zhang, S. (2019, October 22). *Key Anthropometric and Physical Determinants for Different Playing Positions During National Basketball Association Draft Combine Test*. NCBI. Retrieved November 16, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6820507/

Diambra, N. J. (2018). *Using Topological Clustering to Identify Emerging Positions and Strategies in NCAA Men’s Basketball*. TRACE: Tennessee Research and Creative Exchange. Retrieved November 16, 2022, from https://trace.tennessee.edu/utk_gradthes/5084/

Duman, E., Sennaroglu, B., & Tuzkaya, G. (2021, December 30). *A cluster analysis of basketball players for each of the five traditionally defined positions*. SAGE Journals. Retrieved November 16, 2022, from https://journals.sagepub.com/doi/abs/10.1177/17543371211062064?casa_token=z3ZfL-bn81QAAAAA%3A_OCy2vN_bVk1mZxzPYjDXF4XnMQL7F0U-19atqMWuVNh8IcLGvb24cRhRWyjPCHU5in00F5eRe-7&journalCode=pipa

Jyad, A. (2020, November 16). *https://towardsdatascience.com/redefining-nba-player-classifications-using-clustering-36a348fa54a8*. Redefining NBA Player Classifications Using Clustering. Retrieved November 16, 2022, from https://towardsdatascience.com/redefining-nba-player-classifications-using-clustering-36a348fa54a8

Kalén, A., Pérez-Ferreirós, A., Costa, P. B., & Rey, E. (2020). Effects of age on physical and technical performance in National Basketball Association (NBA) players. *Research in Sports Medicine*, *29*(3), 277–288. https://doi.org/10.1080/15438627.2020.1809411

Li, B. (2021, July 8). *Application of Artificial Intelligence in Basketball Sport | Journal of Education, Health and Sport*. https://apcz.umk.pl/JEHS/article/view/JEHS.2021.11.07.005

Metulini, R. (2017, July 4). *Space-Time Analysis of Movements in Basketball using Sensor Data*. arXiv.org. https://arxiv.org/abs/1707.00883

Mills, J. (2015, August 13). *Decision-Making in the NBA: The Interaction of Advanced Analytics and Traditional Evaluation Methods*. https://scholarsbank.uoregon.edu/xmlui/handle/1794/19127

Narayan, S. (2019, June). *Applications of Machine Learning: Basketball Strategy*. MIT.edu. Retrieved November 16, 2022, from https://dspace.mit.edu/bitstream/handle/1721.1/123043/1127911338-MIT.pdf?sequence=1&isAllowed=y

Ostojic, S., Mazic, S., & Dikic, N. (2006, November). *Profiling in basketball: physical and physiological characteristics of elite players*. NCBI. Retrieved November 16, 2022, from https://pubmed.ncbi.nlm.nih.gov/17149984/

Pion, J., Segers, V., & Bourgois, J. (2018, March 29). *Position-specific performance profiles, using predictive classification models in senior basketball*. SAGE Journals. Retrieved November 16, 2022, from https://journals.sagepub.com/doi/abs/10.1177/1747954118765054?casa_token=XjAo0KBhEPQAAAAA%3A2XIKVFc6LGOo5KFKOuXWMbCuNrv6uMSfbcU5PgRYMcGFevhpN_SfZ3bj6hOKHb_X4-YkDPMU5TD7&journalCode=spoa

Rangel, W., Lamas, L., & Ugrinowitsch, C. (2019, August). *Basketball players' versatility: Assessing the diversity of tactical roles*. ResearchGate. Retrieved November 16, 2022, from https://www.researchgate.net/publication/335010180_Basketball_players&apos;_versatility_Assessing_the_diversity_of_tactical_roles

Sandholtz, N., Mortensen, J., & Bornn, L. (2019, December). *Measuring Spatial Allocative Efficiency in Basketball*. Retrieved December 4, 2022, from https://www.researchgate.net/publication/337904719_Measuring_Spatial_Allocative_Efficiency_in_Basketball

Sarlis, V., & Tjortjis, C. (2020). Sports analytics — Evaluation of basketball players and team performance. *Information Systems*, *93*, 101562. https://doi.org/10.1016/j.is.2020.101562

Schoch, D. (2018, March 4). *Analyzing NBA Player Data II: Clustering Players*. http://blog.schochastics.net/post/analyzing-nba-player-data-ii-clustering/

Stephanos, D., Husari, G., Bennett, B., & Stephanos, E. (2021, April). *Machine learning predictive analytics for player movement prediction in NBA: applications, opportunities, and challenges*. Retrieved December 4, 2022, from https://dl.acm.org/doi/abs/10.1145/3409334.3452064

Vaci, N. (2019, January 25). *Large data and Bayesian modeling—aging curves of NBA players*. SpringerLink. https://link.springer.com/article/10.3758/s13428-018-1183-8?error=cookies_not_supported&code=21377040-6797-4877-bc76-6825a862e141

Wakim, A. (2014, March 28). *Functional Data Analysis of Aging Curves in Sports*. arXiv.org. https://arxiv.org/abs/1403.7548

Zhai, Z., Guo, Y., Zhang, S., Li, Y., & Liu, H. (2021, January 13). *Explaining Positional Differences of Performance Profiles for the Elite Female Basketball Players*. NCBI. Retrieved November 16, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7874149/