A Multi-Stage Clustering Algorithm to Re-Evaluate Basketball Positions and Performance Analysis

MSc Research Project
MSCDATOP

Andrew Baumann
Student ID: 20156138

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Andrew Baumann

Student ID: 20156138

Programme: MSCDATOP **Year:** 2022

Module: MSc Research Project

Supervisor: Jorge Basilio

Submission Due Date: Monday 19th September 2022

Project Title: A Multi-Stage Clustering Algorithm to Re-Evaluate Basketball Positions and Performance Analysis

Word Count: 9,548 **Page Count**: 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: A Baumann

Date: 18th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Multi-Stage Clustering Algorithm to Re-Evaluate Basketball Positions and Performance Analysis

Andrew Baumann
20156138
MSCDATOP

Abstract

Availing of computer-based methods of statistical analysis and machine learning has a heavy and budding presence in sport. Data collection and analysis is growing more and more with every passing year, with exciting times lying ahead for data analytics within sport. However, the repetition and staleness are clear to see, and the area is in need of new and innovative approaches. In this report, it is argued that one machine learning methodology remains underutilised, despite displaying opportunity to aid research and the relationship between machine learning and sport.

It is presented in this report that the five traditional playing positions in basketball are outdated for the modern game. The clustering of players based on their performance output and thus their specific skill set proposes a unique approach in fitting an athlete to a position more suitably describing the modern game. New groupings based on a novel multi-level clustering algorithm methodology are proposed in order to better classify and rate players skillset, also allow for ranking of said players and thus assisting in the decision-making and planning for basketball teams, and an overall redefinition of how positions and players are seen in the sport.

Contents

1 Introduction

Data being used to assist in decision-making in all industries is more prevalent now more than ever, and sport is no different. An increasingly analytical and data-driven approach is being taken towards sports, on and off the field. Every aspect, statistical output, and second of a game is dissected and analysed at great length by analysts and spectators alike. As the game of basketball is rapidly advancing in terms of player development and talent levels, a clear shift is evident from the more traditional expectations per position, which have not changed nor seen any development since the sport's invention.

Historically, the five positions on a basketball team all had different roles with different expectations. However, it is clear the lines have become blurred in recent years within these positional groups, with new analysis needed for identifying skillsets and abilities. This is evident in Figure 1, which shows the average statistical output of each position over the most recent ten seasons in the National Basketball Association (NBA) amongst a selection of the more known statistical outputs, such as points, assists, and rebounds. Whilst some differences are still seen, for the most part, they are similar across the five positional groupings far more so than would define different positions in other sports and in the NBA itself in the past.

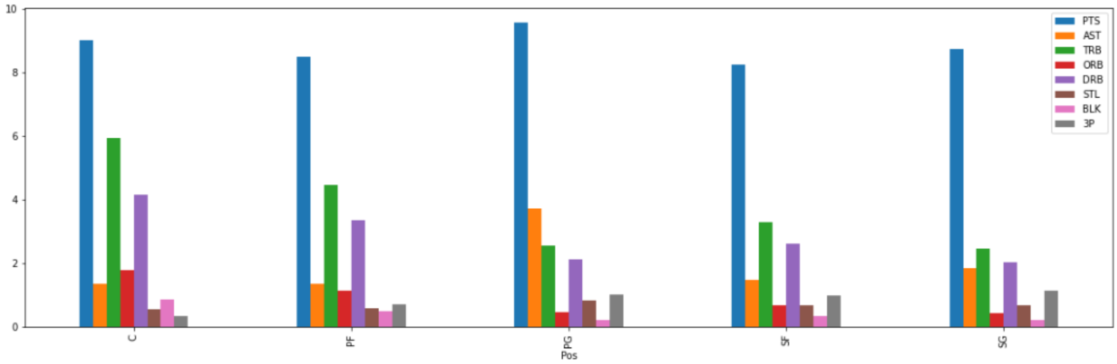


Figure 1: Average output of each position in the NBA over the most recent 10 years of data

In previous decades, the position of Center (C) was seen as a player not often to be the primary points scorer nor have much impact outside of being the largest player on the court and reaching for any loose balls in the air. There's an old saying in the NBA about this; "you can't teach size". Their role was not to shoot from distance nor be a playmaker, but to be a powerful and frankly inconvenient presence on the court. Figure 2 shows how this, amongst other aspects, have changed drastically even in the past ten years. The game itself is evolving and so too must the analytical approaches taken in decision-making and assessing player development within the sport. As such, this speedy evolution of the sport is seeing the need for new and improved methodologies for assessing a player's skill set and talent level, as well as their fit with a specific team and what needs they can fill for this team. Whilst positions have developed in other sports, basketball, which was first played in 1891, has not seen any overhaul of its positions since its creation. That is not to say that a Point Guard in 1891 plays the same role as one in 2022, but it would be difficult to argue against the sport needing a fresh and more modern take of the positions, formations, and roles.

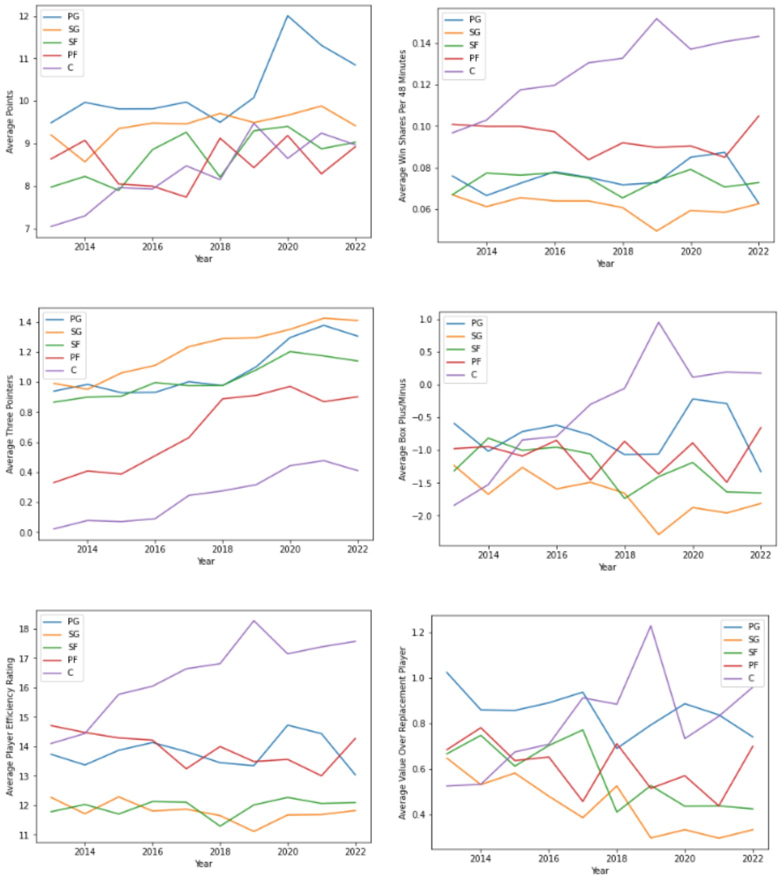


Figure 2: A summary of positional outputs over the most recent 10 years of data

Looking to machine learning methodologies used in sport, there is an oversaturation and repetition for areas centred around classification models such as those seen in result prediction algorithms. However, player groupings using clustering algorithms have yet remained underutilised given the significant contribution it could have on the aforementioned progression of positions within basketball. Machine learning and sport already have a strong relationship, which is growing at a meteoric rate, but there is a need for more innovative and novel approaches in research. As such, whilst the purpose of this research is focussed on re-assessing how positions are viewed within the NBA, there is also a goal within this research to act as an advocate for the exploration of less common and more novel approaches and show the potential contribution they can bring to an environment in danger of becoming stale.

The benefit of data clustering naturally lies in its pattern recognition. Clustering is an unsupervised learning method which generates a group containing data with the most similarity to each other. The potential of different methodologies, such as K-means to avail of feature extraction and segmentation have proved to be advantageous in sporting analytics academia, (Xia et al. 2020), which focussed on such pattern recognition applications. The adaptation of such a methodology in using clusters to assist in player scouting and talent identification is clear. Using methodologies to group and reclassify players will not only assist in redefining basketball positions and players themselves, but I also hope to put forward a case for re-evaluating how clustering methodologies and unsupervised machine learning is viewed in the scope of data analytics within sports as a whole. Thus, highlighting this as an under-explored area of research seeking to spark interest for more innovative approaches to be explored in the sector.

Through research and model development, the role of clustering methodologies, along with additions such as feature selection and dimensionality reduction, is assessed with grouping NBA players on skills and attributes beyond the outdated five positional groupings. Should it be seen that players are grouped correctly by skillset, it will prove invaluable for a team in terms of scouting and player development and changing the perspective on positional value in the sport as a whole. "In other words, player segmentation allows those who are suitable for a particular tactical position in the sport to take advantage of the abilities, services, and program to develop their talent" (Khobdeh, Yamaghani, and Sareshkeh 2021).

Clustering in itself proves to be a novel approach to machine learning exploration within sport, but to increase this innovative approach is the introduction of a multi-stage clustering methodology which leans on Gaussian Naive Bayes (GNB) and Support Vector Machine / Classification (SVM / SVC) algorithms in ensuring validity of level two of this process. Proposed here is that initial clustering shows a concise grouping of players based on particular play style and statistics. The second stage in this framework allows for a second instance of grouping at a more isolated and focussed level, but also act as a means of ranking the players as a sort of quality separation within this insular comparison. This also allows for evaluation metrics to be incorporated with inclusion of supervised algorithms within the proposed methodology.

As mentioned, the intention is to not only introduce innovation into a sector becoming stale, but also to present a case to re-assess both machine learnings place in sporting analysis and player assessment. The objective can be summarised by the below research question.

- **Can a two-tiered clustering algorithm be used to redefine basketball positions and aid in re-evaluating player performance?**

Whilst the suitability of using clustering methodologies as a means of portraying new groups, or positions, here is undoubtedly logical, there are certain areas of potential limitation. Firstly, with clusters, evaluation metrics can be difficult to incorporate as well as assessing any results without levels of subjectivity or an adequate prior knowledge of the area. It may be obvious to any avid basketball fan that Stephen Curry is a player with phenomenal shooting and should be classed as such, for example, but there are little options in grading this objectively in a pure accuracy score sense for model development and submission. This will be a topic of discussion for stage one of the multi-levelled models, but as mentioned, stage two negates some of these subjectivity fears with the GNB and SVC implementations.

This report is separated into six sections, as shown in the table of contents. The following section discusses the research surrounding this field and in itself is broken into subsections to show the progression of machine learning models within sports. Following this, the research methodology discusses the dataset and the justification in forming of the methodology. The design and implementation section is seen after this and dives deeper into the environment, steps, and novel double clustering algorithm proposed here. Finally, each stage is evaluated, before conclusions are discussed.

2 Related Work

The review to follow is divided into five subsections following this introductory paragraph. Firstly, the predominant uses of models and the identification of gaps in this area are explored (2.1), before attention is turned to looking at the novel ideas already researched (2.2). After this focus is turned to the proposed gap identified in using clustering analysis and unsupervised methods for athlete evaluation (2.3) and a discussion on the multi-level proposition (2.4). Finally, conclusions are made based on the literature reviewed (2.5).

2.1 Machine Learning in Sport Background and Identifying Gaps

The decision-making process in the sports world is rapidly evolving. It's difficult to deny that a more analytical approach is being taken to all facets of this, both internally and externally of the sporting organisations themselves. There is a heightened importance on machine learning and model development accentuating this trend and showing the potential of further progression for player, team, and fan analysis. Indeed, even more recent studies, (Horvat and Job, 2020), show the aptitudes of this and the place of machine learning in sport, but mostly focus on high accuracy result predictions, rather than anything at a player evaluation level. Whilst the authors present a review of this field with excitement for its development, it remains in the far too common result outcome prediction field with little deviations seen from analysis on a team outcome level rather than a player assessment level.

That is not to say there are no discussions of player analysis across United States sports. This is seen mostly in simplistic logistic regression models and with one such example (Nandakumar and Jensen, 2019), focussing on analysis within ice hockey and on the scouting and assessment of young players. In this paper, players in the National Hockey League are graded and ranked in terms of numerous expected skills. This paper, whilst maintaining a rather simplistic and not overly promising for future work in itself, shows the beginning stages in machine learning models moving towards player evaluation rather than simply basic result prediction models in the sporting world.

What has been seen on player analysis in the NBA to date has principally been focussed on a player's impact in a predicted win model. As seen, (Nagarajan, Zhao, and Li, 2018), this approach is used in which player performance variables are fed into models for a binary win loss outcome. Whilst simple supervised learning methods, such as linear regression here, are suitable, the tracking of those with more complexity or the use of unsupervised learning becomes less prevalent. Another such example (Huang, Chih-Jen, and Ruby, 2008) looks to deviate from statistical binary responses in assessing player attributes such as points score, as well as the team result. However, this study falls short in the level of player assessment as a whole and intricacy in methodology that this research intends to deliver. Again, the pitfall of exploring an oversaturated area of a team result outcome is evident, whilst an exploration into individual athletes themselves would lead to a more novel and elaborate study. As will be shown to be the aim of this paper.

2.2 Literature Review on More Novel Approaches

The use of cluster analysis to group and tier players and the rarity of such studies makes this an attractive area to explore further. Prior studies, (Terner and Franks, 2021), highlight this necessity to advance beyond simplistic and conventional methods to more novel proposals for both sport analysis as a whole, and for athlete performance analysis within that, with clustering models indicating notable potential to bring new findings to this field. The novelty of such research lies, therefore, in the exploration of methodology which is both more complex and less prevalent in academia centred on machine learning within sport.

It has been said that basketball and the NBA provide a perfect environment for both player and team investigation within the space of deep learning in (Nguyen et al., 2021), due to the enormous amount of data available. This particular paper sees the traditional methods of various forms of regression analysis pitted against deep learning methods, such as deep neural networks, k-nearest neighbours, and so on. Although concentrating largely again on predictive classification, this paper shows the deviation from the exceedingly saturated traditional methods previously discussed. The results in this instance show little to no advancement in the deep learning methods over the traditional regression models, however the failure to include certain additions to the model generation such as standardisation and dimensionality reduction analysis by the authors surely had a negative impact on results here.

The importance of methods such as neural networks, support vector machines, and extreme gradient boosting are discussed (Bai and Xiaomei, 2021), for the feeding of athlete's statistics and metrics are seen in prediction models also. Similarly, the prediction of football player performance using deep learning methods is explored in (Manish, Bhagat, and Pramila, 2021) in which neural network models are used to assess player performance based on their position of play. This final point is noteworthy given the positional significance of players' and their roles in basketball as well as in football. Again, the recency of both papers shows the growing interest in atypical and innovative methodologies for assessing player performance in. The authors argue that the application of systems to predict player performance is a highly sought-after academic area, with tangible results often difficult to achieve, as the case with both.

Additional approaches include the development of efficiency-based models, as evident in (Radovanovic et al., 2013), which intends to judge an athlete's performance level using data envelopment analysis and distance based metric evaluation. The authors’ inclusion of additional metrics, such as a player’s salary, were of note in the paper reviewed. This is particularly pertinent in US sports given that each team operates under a maximum salary allowance, or cap, that cannot be exceeded and thus not be able to simply pool the best players at the highest costs on one team. Looking towards player evaluation more, the main methods seen in prior research make use of regression models and neural networks, whereas clustering is seen as a means of team evaluation rather than on a player level, (Bunker and Thabtah, 2019). Emphasised in this paper is the enlarged use of varying machine learning algorithms in sport as well as the rapidly growing interest for models to aid in planning of strategies, providing further motivation to explore diverse algorithms in this field.

2.3 Literature Review on Clustering Models within Sports

The use of clustering within sports is not totally unexplored. One such paper (Narizuka and Yamazaki, 2019) shows the potential of such approaches as the authors present the idea of identifying similarities of players and their suitability in a specific formation. However, this paper fails in any sort of grading the players' ability and in grouping and classifying players at a style-based level. Both of these shortcomings will prove to be the main objectives in the model generation to follow. To find research which aims to dive into the clustering of players based on their statistical outputs, most only go to a relatively basic level (Sampaio et al., 2015). Here, players are merely grouped into good and poor performers which then leads the authors to uninspiringly conclude that one cluster naturally performs better than the other.

Elementary positional separation analysis is seen in (Duman, Sennaroglu, and Tuzkaya, 2021), where the use of clustering is seen for grouping by playstyle within each of the five positions on a basketball team. Whilst simplistic in nature, the positional grouping shows the potential in application of such methodologies in sorting various athletes into clusters based on variables. This paper attempts to determine the style of play of each player through this positional identification clustering, although more complex methodologies and papers will need to be assessed in order to provide true value to the proposition of this report. I would argue the outdated nature of clustering primarily by position in this manner, given the previous reasons stated for the obsolete nature of the traditional basketball positions and the need for a fresher look at these. Hence the value in the application of this report to follow.

One area for concern in research of this sort is the inevitability of areas of subjectivity or prior knowledge of the sport required in the building and assessing of models. This is particularly evident in clustering, where there is no right or wrong answer in most cases. This can be seen in (Xin, Zhu, and Chipman, 2017), which groups basketball players based on their style of play, including such subjective metrics such as handling of the ball. This perhaps takes the lack of objectivity too far in the manner of giving ratings for a player’s style and for attributes not officially recorded, but rather ones which are judged. My research aims to use statistics and data and alleviate this subjectivity. While this judgement stage holds some level of bias and need for previous knowledge of the sport, I would argue it has been reduced as much as possible and more so than in previous studies, such as this.

"Clustering players based on their abilities, a new perspective and an important opportunity" (Khobdeh, Yamaghani, and Sareshkeh, 2021). Here the authors express the scope of moving into clustering analysis for the identification of talent within basketball, and indeed sport as a whole, primarily due to its ability for pattern recognition amongst the variables. It is put forward that through this and principal component analysis as a form of variable reduction, the grouping of players based on skill set is a feasible approach for talent acquisition and scouting. Significant in this paper was the aim to identify such talent early and thus gain an advantage in said scouting, whilst claiming that the clustering in conjunction with variable reduction and neural network analysis allowed this. The authors also stress that such studies are in their infancy when embracing this sort of methodology.

2.4 Multi-Level Clustering

Implementations of multi-level clustering techniques are rare, but one such example (Inuwa-Dutse, Liptrott, and Korkontzelos, 2021) uses this to identify the strength of relationship and solidarity over groupings made on Twitter. Whilst certainly a differing area of research than the statistical analysis of basketball players in this paper, the identification and use of the microcosms, as the subdivisions of clusters are dubbed by the authors, provides evidence of the potential usefulness in such a theory of double or multi-level clustering. Similarly, further studies (Bouhmala, Viken, and Lønnum, 2016) show this methodology idea as an evolution for “maximizing both the homogeneity within each cluster and the heterogeneity between different clusters”. Highlighting the differences between initial cluster groupings of types of players whilst emphasising skill level differences in the second grouping seems to fit this notion of seeking to improve upon asymptotic convergence from just one cycle.

2.5 Conclusions and Proposal

In conclusion, innovative methods have the potential to make a sizable contribution to this field. The recent research papers assessed in this review make this clear and evidence has been presented for clustering models to show significant potential in being this innovative method in athlete evaluation. There is also basis for a multi-level clustering solution to emphasise initial groupings while stressing skill level separation within the initial isolated groupings. The expectation of the methodology to follow being to not only prove valuable within basketball and addressing this specific research question, but also in being of tangible benefit to the overall environment of machine learning's relationship with sport analytics.

3 Research Methodology

3.1 Process Overview

The procedure followed in the piece of analysis was the Knowledge Discovery in Databases (KDD) guidelines, meaning the 5 sequences of steps, shown in Figure 3, were taken with iterations back to previous stages when necessary. It is often said with KDD that the primary focus of this framework is to extract information from large amounts of data to arrive to your conclusions (Azevedo, 2019) and that this aligns well with machine learning exploration research projects (Ester and Sander, 2013).

3.2 Architecture and Technologies

Python was used as the primary data processing programming language due its versatility in being able to connect to various data sources, transform data in multiple formats, and analyse and visualise data. There are a number of libraries which enable seamless processing, transformation, mining, model generation, and evaluation, including but not limited to, Pandas, Scikit-learn, and Matplotlib. Due to its better handle of web scraping libraries, the PyCharm IDE was used for this step, whilst the Notebook style of Jupyter was preferred for the data transformation, visualisation, model building, and all else up to this final point.

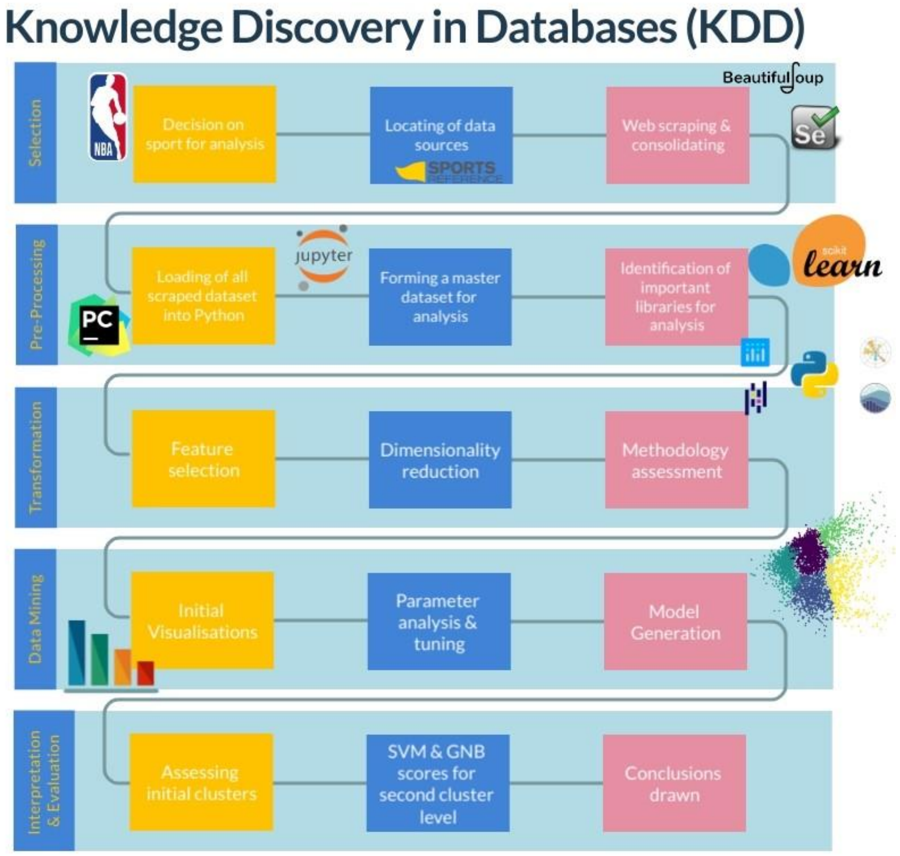

Knowledge Discovery in Databases (KDD)
Selection
Decision on sport for analysis
Locating of data sources
SPORTS REFERENCE
Web scraping & consolidating
BeautifulSoup
Pre-Processing
Loading of all scraped dataset into Python
jupyter
Forming a master dataset for analysis
Identification of important libraries for analysis
scikit learn
Transformation
Feature selection
Dimensionality reduction
Methodology assessment
Data Mining
Initial Visualisations
Parameter analysis & tuning
Model Generation
Interpretation & Evaluation
Assessing initial clusters
SVM & GNB scores for second cluster level
Conclusions drawn

Figure 3: The 5 Process Steps of KDD – with Specific Platforms, Software, and Libraries used

3.3 Datasets

The data that has been selected for this report is centred around basketball and the NBA. Assessing the suitability of basketball and the NBA for such studies, the United States has long since been adopting a more analytical approach to sports when compared to the rest of the world, with large emphasis on measuring every statistic and play from every player in a given game, dating back decades now. Basketball also offers an incredibly wide range of statistical recordings on a player level. Points, assists, rebounds; both defensive and offensive, blocks, steals, and many more. It presents an ideal environment for the analysis to follow and a great representation for sport as a whole.

Data is readily available for the NBA in the public domain and easily accessed from the direct source of the NBA themselves or through trusted stat collection websites such as Basketball-Reference, which uses the official statistic provider of the NBA. Whilst the data is freely listed, the easy download of a file from a trusted source is not, and as such this required web scraping from a trusted and accurate cite. This came from Basketball-Reference in the form of reading the html of the tables and appending to a table within Python before concatenation of the years and download of a master file per report.

For this analysis, a new python script was built to scrape the data using the Selenium and BeautifulSoup libraries, both of which allow for such web scraping and saving of data. My script allows for a simple entry of a start and end year and a dataset is compiled from this scraping and saved locally for analysis. Selenium allows for the iteration through various years as separate chrome instances to view the report on screen, and BeautifulSoup allows for the parsing of these tables before formatting, concatenation, and the eventual download. For this study, a ten-year range was set for the analysis dating from 2013 to 2022 inclusive, given the focus of my research is on recent positional and statistical developments in the NBA.

Three separate sets of data were scraped over the same timeframe. The first is labelled as “Per Game” statistics and shows basic recordings such as points, assists, rebounds and so forth per game for every player in the NBA. The second is more in depth and labelled as “Advanced”. This again is for every player in the league in a given season, but shows metrics such as Player Efficiency Rating, Value Over Replacement Player, and many other advanced calculations widely used in basketball analysis. Lastly the “Shooting” report was also scraped for this timeframe showing the types of shots taken by each athlete as well as their frequency. A key for all of the vast number of columns within each is provided at the end of this report.

Following the scrape, data cleaning and verification was undertaken before this combining of the three scraped datasets into one master file. After the scraping, formatting, and feature selection including the removal of redundant or repetitive variables, a final dataset, of near 50 columns and over 4,500 rows of player data, was used for the initial visualisations and model building analysis. All datasets were scraped from Basketball Reference due not only to its reputation and reliability, but also for consistency purposes in the merging process. Due to this dependability, much of this initial pre-processing was centred around simply getting the dataframe into the most readable format for visualisations and model generation within Python. Additionally, for the sake of dependability within the results, any player with ten or fewer games played in a given season was removed to avoid irregularities or outliers caused by player injuries or other outside variables. More reports were available to be scraped, however, the remaining options showed little value and much repetition to what was already scraped, so it was decided this final master file was more than sufficient.

It is worth highlighting Section 9 of the NBA’s terms of use, which stipulates the terms of using NBA statistics primarily centred around supplying a prominent attribution to the NBA for the use and not using this information for gambling or fantasy sports purposes. The use of NBA statistics in the manner proposed in this report is therefore acceptable given the scope and nature of the research and writing is purely for methodology exploration and academia contribution, rather than betting, fantasy sport, or any solely personal benefits of this sort.

3.4 Objectives

As mentioned, the overall objective here is the examination of not only an underutilised methodology in sports analytics, but also in the proposal of a new and innovative method of addressing player positions, as well as talent scouting and player evaluation within basketball, which has been shown to have significant potential for further development such as this (Khobdeh, Yamaghani, and Sareshkeh, 2021). Whilst moving away from the oversaturation of certain approaches within sporting analytics, as outlined in Section 2.1 of this paper, and simultaneously seeking to address the growing interest, curiosity, and attention within machine learning and sport to propose new points of view.

The goal is the development of a multi-level clustering model will assist in grouping together athletes of similar skill and assist in fitting said athletes to team in need of their style of play, thus redefining outdated positional listings. It is my expectation, given the findings in previous literature, that this will assist in the creation of a novel method of planning and decision-making within the basketball world, and should this be a success, I would see no reason why this currently under-researched methodology could not be broadened to look into other sports too. Furthermore, successful results would lead to the argument for clustering methodologies to be used more in such studies, as at present they remain amongst the lest implemented in such sports studies on data analytics and machine learning.

3.5 Methodology

There are many clustering methods to consider, the first of which being K-Means Clustering, due largely to its simplicity, popularity, and visual capabilities. The K-Means algorithm comprises of the assigning of each of the number of inputs (n) to one of the clusters (k), where k has been outlined ahead of time. Here, the model seeks to minimize the differences within each cluster and maximize the differences between clusters. Whilst the expectation would be to trial multiple methods of clustering to find the most suited in this particular instance, the evaluation of data that holds no exact right or wrong answer in its grouping of the data proves a difficult task, as (De Diego et al., 2019), outline. For what the authors here describe as potentially subjective data, the evaluation would also be subjective in nature.

Another paper, (Page and Quintana, 2015) looked at predicting an NBA athlete's career curve using hierarchical clustering whereby it is argued the algorithm allows for the fit of individual curves while still fitting clusters used to guide the prediction. This type of clustering, however, is seen to struggle with handling larger sets of data, so would not seem to be most suitable here. “Density-based clustering methods do not assume parametric distributions or use variance” (Hahsler et al., 2019), meaning optimal cluster number is selected by the algorithm rather than in entry. Density-Based algorithms, such as DBSCAN, only permit the change of the minimum samples to form a dense region or cluster. This method also marks outliers as an independent density cluster region, for which statistically excellent and poor players would be marked together simply as outliers, thus eliminating it from consideration.

Other clustering methodologies were tested in stage two of the multi-level clustering, however, given the outputs, K-Means was selected as the primary clustering methodology. Not only for these results, but also due to its flexibility and adaptation. This allowed for the models to follow this section to be built from the ground up with additional elements of complexity and novelty added as needed. K-Means boasts the ability to scale to large datasets as well as easier interpretation than most others. Finally, the clusters themselves tend to be tighter, particularly with globular clusters, when compared to hierarchical clustering and can often increase accuracy as number of clusters increase (Kaushik and Mathur, 2014), which is particularly pertinent in this study.

Whilst widely considered the most powerful clustering algorithm, K-Means is not without its problems (Ahmed, Seraj, and Islam, 2020). Most obviously, it required the number of clusters to be defined prior to any analysis. However, the Elbow Method and Shadow Silhouette Coefficient were calculated throughout the analysis to give guidance and thus negate this downfall. On top of this, K-Means shows an inability to handle multiple data types, which is not an issue in such statistical analysis. Lastly, there can be marginal random elements to the initialisation of the centroid fitting. However, it is widely accepted this inconsistency is slight and should not be one to deviate the decision in choice of algorithm.

Using a distance-based method such as K-Means Clustering, it is crucial to implement some manner of feature scaling before the application of any algorithm. Early statistical analysis of the dataframe within this study showed the average points scored per game by a player in the NBA over the past ten years to be 8.37. When compared to the average block rate of 0.39, we can see a clear need for standardisation, which was undertaken here. Given the objective lies in a full analysis of all statistics to better fit player roles, this standardisation and subsequent prevention of variables with larger scales from dominating how clusters are defined proved valuable in the final models.

Principal component analysis (PCA) proved to be a crucial component of this methodology given the vast amount of data and variables to be used in this analysis. Prior research (Bafna and Saini, 2019), state the importance of using dimension reduction to remove features that contribute less to the model, but particularly stress the need to do so for clustering models both for speed and accuracy. In doing so, the authors contend the removal of outliers and noisy data aids their model's predictive power, while lessening any burdens of costs or time. The data involved in this NBA analysis includes a substantial master file, and thus a form of reduction or selection proved to be advantageous if not pivotal.

With the primary pillars of the methodology set in K-Means Clustering, standardisation, and principal component analysis, the forming of the implementation of the multi-levelled clustering can take place. Given the second tier in this multi-level clustering solution is centred on a ranking system and is based upon arranging into two clusters in each of the initial instances, the inclusion of classification allows for a level of confidence in the results of this stage of the model. The SVC algorithm builds on simple logistic styles by creating a hyperplane, or decision boundary line, separating data into classes and finding the best line separator, which acts as the boundary between the two possible output variables. GNB is a simple extension of Naive Bayes that follows Gaussian normal distribution and is used when dealing with continuous data. Given this and the scaling of the data, as has been discussed, GNB and SVC are most suited given the assumed distribution of the data due to this.

3.6 Evaluation Assessment

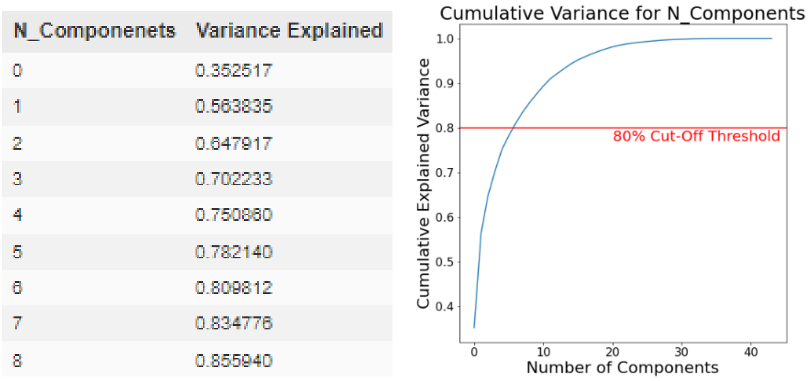
Whilst there are dedicated studies surrounding the more objective means that can be taken to assessing the results of clustering methodologies, interpretation is largely left as a subjective manner (Aljarah et al., 2021). The authors here present the case for reliance on external data labels for generated clusters in many studies. Naturally, it is arduous to argue against external data labels and a somewhat subjective approach in this particular instance. In terms of assessing the initial clusters themselves, much of this comes down to confidence and trust in the models created and the parameters set from testing, as well as some prior knowledge.

Whilst the argument for objectivity has been explored with clustering, the elements of subjectivity and prior industry knowledge, in this instance being basketball, will never be truly removed. However, this subjectivity is alleviated somewhat in the second instance of the clustering and inclusion of the supervised methods as discussed previously. Referring back to the research question and underlying purpose, it can be said that each level of this clustering assists with answering each part of the question. The first level aids in redefining the outdated positions in basketball, the second adds further value to this in supporting the evaluation of player performance, as well as acting as a means of re-enforcing the validity of the first grouping. Thus, showing that a two-tiered clustering algorithm can be used to redefine basketball positions and aid in re-evaluating player performance.

4 Design and Implementation

4.1 Setting of Parameters

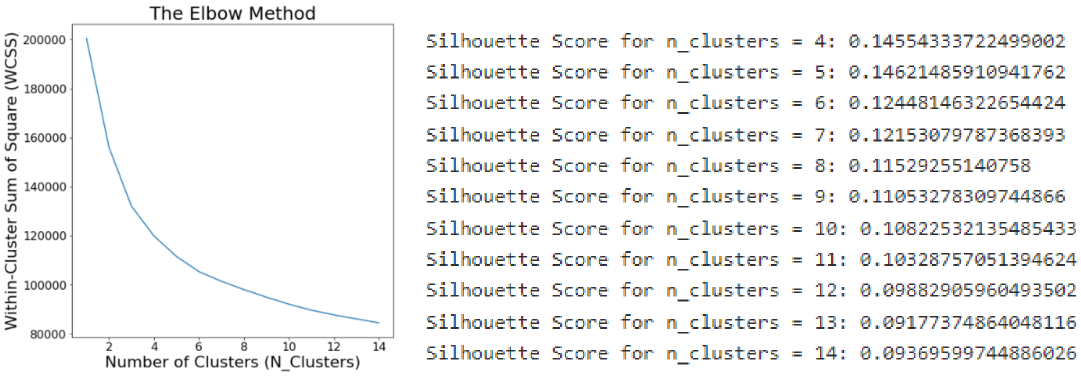
As stated, the primary methodology of this framework centres around K-Means clustering, PCA and standardisation of the variables. Variance explained was calculated and graphed to decide upon the number of components for the models. Given the output provide, seen in Figure 4, a cut-off of 80% was set for variance explained by from the number of components. This left the optimal number of components for the PCA element to this methodology to be set at six. Whilst the variance explained and subsequent cut-off for establishing the number of components is a rather specific topic both to the methodology and particular research question at hand (Gambella, Ghaddar, and Naoum-Sawaya, 2021), a score of 0.8 would generally be seen as more than satisfactory and trials of a cut off at 90% and components size of eleven showed no noticeable improvement or difference over this initial decision.



N_Componenets	Variance Explained
0	0.352517
1	0.563835
2	0.647917
3	0.702233
4	0.750860
5	0.782140
6	0.809812
7	0.834776
8	0.855940

Figure 4: Variance Explained Scores for N_Components

With that in consideration, the optimal number of clusters was in need of being established, given the chosen methodology of K-Means requires a manual input of this. Whilst there is often an argument to be made for an iterative trial-and-error approach in this regard (Ren, et al., 2016), there are calculations that can help give credibility and sanction to this decision. Graphically, the Elbow Method acts as a heuristic in determining the number of clusters in a data set, looking for any obvious kinks or elbows in the graph. Additionally, a silhouette coefficient can be calculated at each cluster number to determine to determine its suitability and overall goodness within the technique. As evident in Figure 5, whilst no clear and obvious elbow is present in the graph, the silhouette score of five clusters proves to be most suitable, with the score dropping with each increase to number of clusters.



```
Silhouette Score for n_clusters = 4: 0.14554333722499002
Silhouette Score for n_clusters = 5: 0.14621485910941762
Silhouette Score for n_clusters = 6: 0.12448146322654424
Silhouette Score for n_clusters = 7: 0.12153079787368393
Silhouette Score for n_clusters = 8: 0.11529255140758
Silhouette Score for n_clusters = 9: 0.11053278309744866
Silhouette Score for n_clusters = 10: 0.10822532135485433
Silhouette Score for n_clusters = 11: 0.10328757051394624
Silhouette Score for n_clusters = 12: 0.09882905960493502
Silhouette Score for n_clusters = 13: 0.09177374864048116
Silhouette Score for n_clusters = 14: 0.09369599744886026
```

Figure 5: Elbow Method and Silhouette Scores for N_Clusters

After this, the clustering models can begin to be generated. However, before showing the outputs and results in the following sections, it is important to again note how this proposed methodology deviates from more traditional clustering proposals. This paper proposes a multi-stage clustering framework, the final results of which reinforced by both Gaussian Naive Bayes (GNB) and Support Vector Classification (SVC) models and outputs. Initial clustering with the parameters supported by previous scores and outputs shows a clear distinction in the grouping of these players, which while difficult to formulate objective measures of evaluation and assessment in supporting the results, prove to be difficult to argue against once the results and five types of players are presented. This element of subjectivity, however, can be reduced with the second stage and the subsequent classification algorithmic evaluations.

4.2 Multi-Stage Clustering Motivation

Proposed withing this framework is the additional and isolated instances of a multi-level clustering inclusion, which not only allows for further grouping of players as a whole, as would be present by a simple increase on number of clusters, but also as a metric of quality separation along with further distinction of groups. The inclusion of the GNB and SVC models act as verification of this separation of class and metrics from our already distinct cluster models. As will be shown in the evaluation section, accuracy, f1, and kappa scores will be used to argue the validity and accuracy in the implementation of this double clustering methodology and the results therein, particularly on a talent division level. With reference to previous experience and to literature (Tan, et al., 2021) a train test split of twenty percent was established throughout given the size of the data is vast but less so when testing on a per year basis. A constant random seed was also established for consistency.

The multi-stage clustering algorithm aims to be a novel proposal to address our research question of re-evaluating the five positions in basketball as well as acting as a means of evaluating player performance and talent level. The inventive solution aims to be a benefit academically for introducing more novelty to machine learning within sports, as well as of practical value to anyone with an interest in basketball or the NBA. The three levels to this can be seen as clustering stage 1, clustering stage 2, and the classification verification. Whilst in themselves simplistic, the consolidation of all, as well as the parameter input optimisation discussed, along with various testing along the way proved to be an intricate task.

4.3 Key Libraries

With that being said, the capabilities of the chosen environment of Python and the strength of the libraries available, primarily BeautifulSoup and Selenium for the scraping and collection of the data and Scikit-learn for model development within machine learning, enable the development of such frameworks, unique or otherwise, for those familiar with them. Function development within Python proved to be pivotal in the trialling of various methods, algorithms, and parameter inputs making this as straightforward and comparable as can be.

4.4 Initial Cluster Visualisations and Investigating

With suitable justification set for the initial parameters fed into the model, the clustering of the standardised and PCA altered data took place. The development of a function in the coding allowed for the passing through of a dataframe, number of clusters, and number of components, which develops the clustering model transforming and fitting the data with PCA and provides the output of the cluster scatterplot graphically, as seen in Figure 6, as well as cluster labels to the dataframe for further analysis. As mentioned, Scikit-learn proved to be the library of the most use for the model generation in itself, while Matplotlib was used in many of the visualisations. More libraries were availed of, but these proved most crucial.

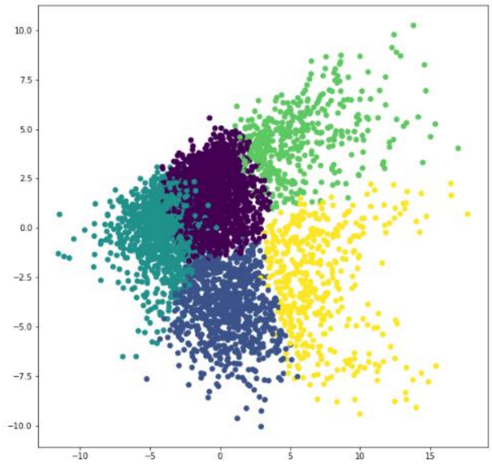


Figure 6: Scatterplot of final data (N_Components as 6 and N_Clusters as 5)

Functions were created for ease of in-depth assessment. Included in these were detailing descriptive statistics of each cluster, a heatmap to the variables showing correlation of a selection of important statistics (Figure 7), the most and least relevant variables for each cluster (Figure 8), visual output of selective important statistics (Figure 12), as well as functions to enable easier searching of specific players, clusters, and years for a more comprehensive analysis of the metrics and members of each of the clusters.

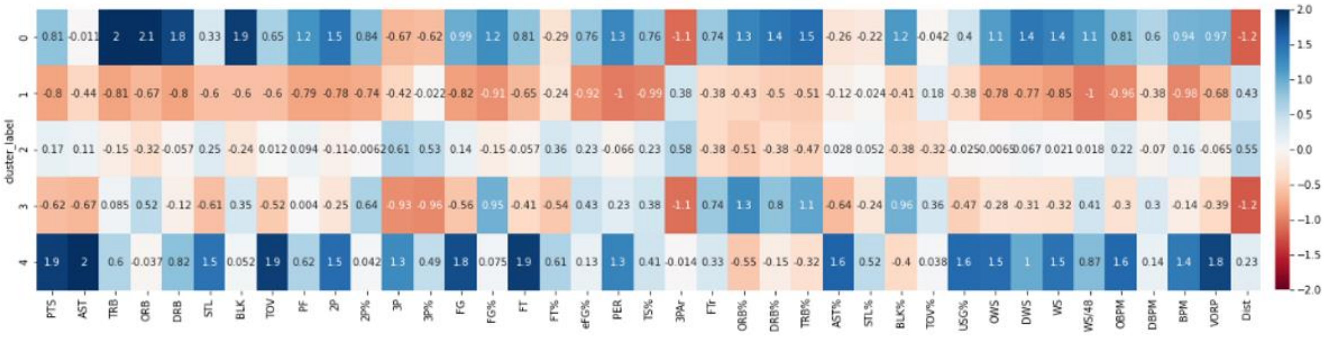
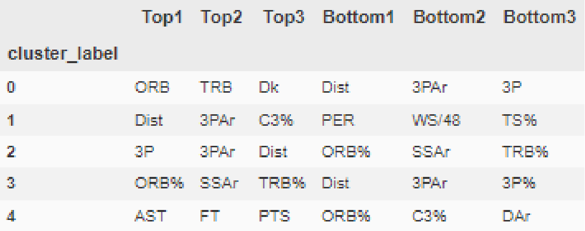


Figure 7: Heatmap of selective important statistics for each cluster



	Top1	Top2	Top3	Bottom1	Bottom2	Bottom3
cluster_label						
0	ORB	TRB	Dk	Dist	3PAr	3P
1	Dist	3PAr	C3%	PER	WS/48	TS%
2	3P	3PAr	Dist	ORB%	SSAr	TRB%
3	ORB%	SSAr	TRB%	Dist	3PAr	3P%
4	AST	FT	PTS	ORB%	C3%	DAr

Figure 8: The 3 most and 3 least relevant metrics for each of the 5 clusters

4.5 First Look at a Second Clustering Instance

A deeper look at the top performing player cluster, discussed in section 6.1, both in terms of the objective statistical analysis and the subjective view of the players included, showed a need for further development of the simplistic clustering model. Implementing the same cluster function but for a further grouping of two for this cluster empathised this, as seen in Figure 9, along with the other previously discussed outputs. A clear division again seen here as the left cluster indicates a more valued player in terms of offensive ability, while the right shows a lower standard but more balanced in terms of assist contribution for example. Whilst the initial clustering and classifying of the players accordingly was to a satisfactory level, this further clustering proved to be of great value for each of the five established groupings.

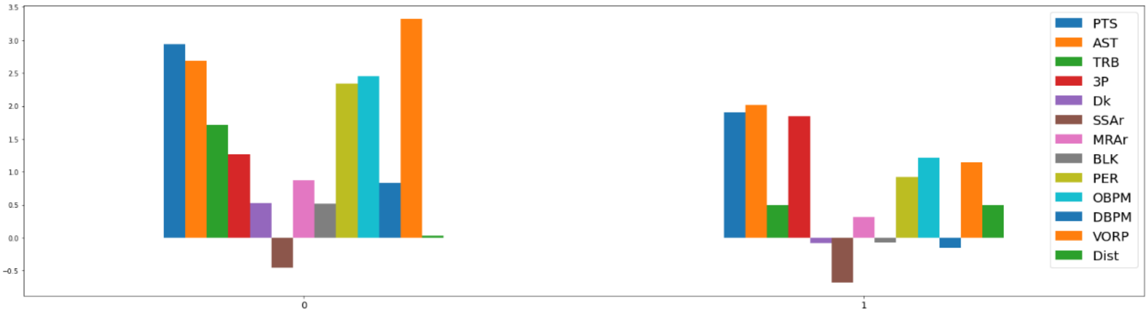
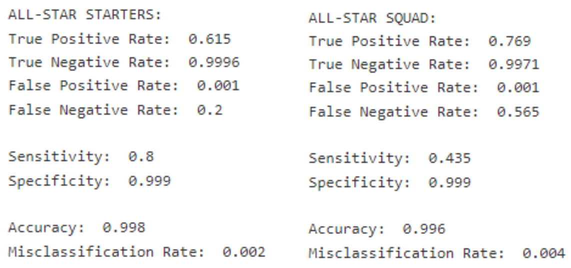


Figure 9: Statistics of top cluster after second instance of clustering

4.6 Trialling Methodologies – Initial Clustering Analysis

Given this cluster showed the top-performing players, it was logical to introduce a metric to verify the proposed methodology, specifically the second clustering and the choice of the algorithm. Confusion matrices and accuracy scores comparing this to the actual top performing players of 2022 (All-Stars) were outputted for K-Means and Hierarchical clustering, as well as other proposed methodology approaches. Overall, K-Means showed the best scores, Figure 10, and thus the methodology proceeded as such. Both academically (Ahmed, Seraj, and Islam, 2020) and in practice, there was no reason to deviated from this methodology, although the testing of other potential algorithms proved to add validation to the choice in the final model in this methodology. This was used as a means of assessing the outputs of the first clustering model here, given much of the first instance is left to trust in the parameters and model itself.



```
ALL-STAR STARTERS:                          ALL-STAR SQUAD:
True Positive Rate:  0.615                  True Positive Rate:  0.769
True Negative Rate:  0.9996                 True Negative Rate:  0.9971
False Positive Rate:  0.001                 False Positive Rate:  0.001
False Negative Rate:  0.2                   False Negative Rate:  0.565

Sensitivity:  0.8                           Sensitivity:  0.435
Specificity:  0.999                         Specificity:  0.999

Accuracy:  0.998                            Accuracy:  0.996
Misclassification Rate:  0.002              Misclassification Rate:  0.004
```

Figure 10: Scores of Top-Level Cluster compared to All-Stars for 2022

4.7 Second Cluster Run

Through looping the separated unique dataframes of each of the five clusters, further groupings were established for each reaching the final output of ten clusters or groups of players. The visual output of each, seen in Figure 11, shows the split from in the second iteration of each cluster. The hypothesised purpose of acting as a second division as well as acting as a means of talent and skill tiering is evident here in this multi-clustering solution. It is worth noting that the simple increase of cluster number from the initial model building did not yield groupings as satisfactory as this approach with the outputs appearing less distinctive, as well as overfit and untidy, both in presentation and in any interpretation thereafter. Additionally, it would be in direct contrast to the parametric testing for the optimal solution. While the final solution does amalgamate to ten clusters, the initial separation to five and the theorised grading thereafter is still under this initial premises and parameters.

Before assessing the specifics of each of the subsequent ten groupings and considering that the second iteration was as a means of quality assessment or tiering of the players, a useful integration to the framework is that of a classification prediction to ensure this ranking has taken place within the algorithm's second phase. Here the validity of this claim is ensured with the running of two separate classification models on each of the initial groups enabling such evaluation metrics to be recorded as f1 scores, accuracy, kappa, and roc curves.

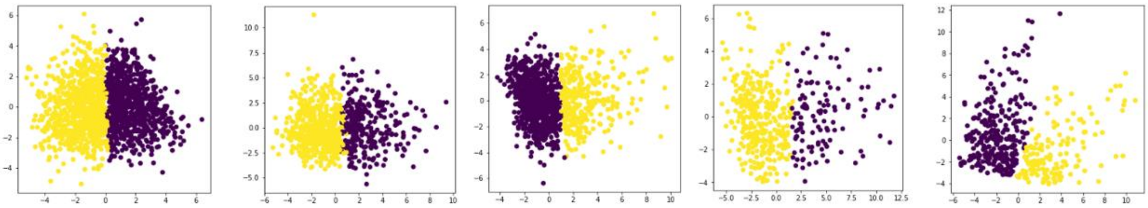


Figure 11: Scatterplots of each of the initial 5 clusters in the second level of clustering

5 Evaluation

Given this is a multi-level clustering algorithm, there are multiple sections of the output to evaluate, taking findings from each. Section 5.1 looks at the first instance of clustering and assesses the output of the initial five groupings created. 5.2 then evaluates the findings of the re-clustering of these initial groupings. Finally, 5.3 explains the classification implementation of the SVC and GNB models, before the overall framework is discussed in 5.4.

5.1 First-Tier Clustering

As discussed, the evaluation of the first stage of clustering is a tricky situation, with the best course to rely on assessing of the breakdown of components and statistics of each of the five clusters here (Aljarah et al., 2021), rather than a specific score measure. Looking to Figure 12, a clear distinction of the five separate types of players from this model is seen. At the beginning of this report, it was shown that the five positions of basketball show a similar statistical output and therefore could be seen as obsolete in the modern NBA. Testing for the optimal parametric inputs conveniently gave an output of five clusters for a direct comparison to be made here. It's hard to deny that the outputs and roles of the new five groupings shows more distinction in the payer style and roles fulfilled than the initial positional breakdown in the introduction. Sport is about many duties and thus a team of these five styles would address all these needs for a team to perform well across all metrics, rather than being wasteful and having five similar players listed as different positions.

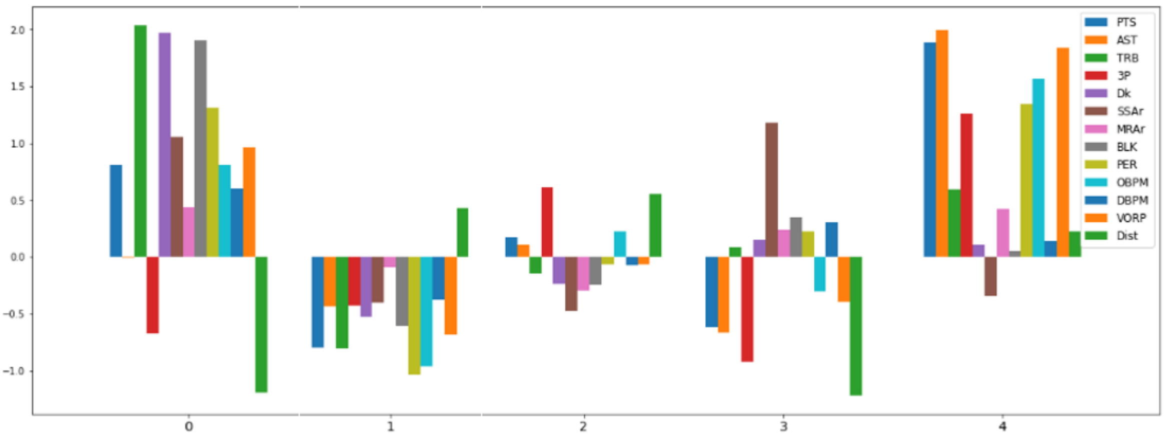


Figure 12: Graphical output of selective important statistics for each cluster

The five new positions in the first stage can be defined as:

- **Big Men**: Tops skills include rebounds and dunking. They have great defensive attributes. Their shooting is close range, and 3-point shooting is far below the norm.
- **Role Players**: These players work hard, and top skills include corner three shooting. They are limited in some areas but are crucial to perform specific roles for the team.
- **Shooters**: One of the most valuable player types in the modern NBA. This cluster excels at all 3-point measurables but will not offer much for rebounds or strength.
- **Close Range and Defence**: In this group, the players rebound well and take short to medium range shots. But lack in 3-point shooting and playmaking abilities.
- **Playmakers**: Points, assists, value over replacements. This bracket are the stars of a team who will score and assist but lack the strength of the big men.

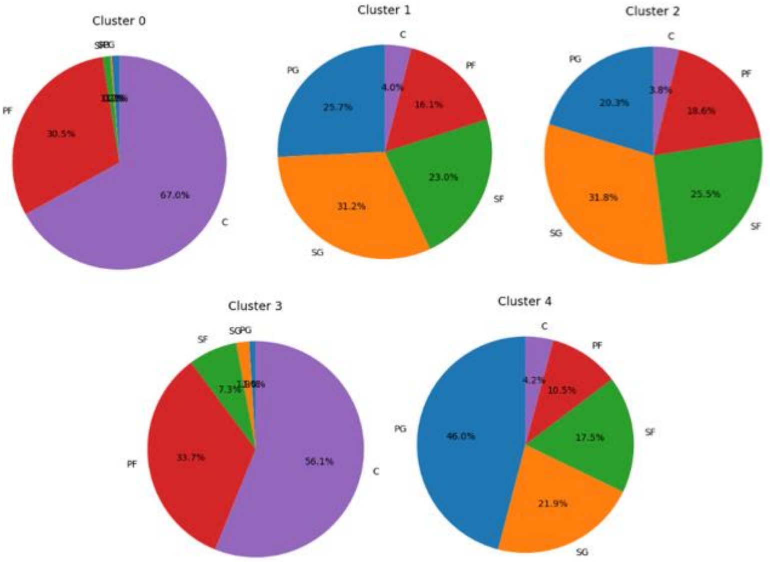


Figure 13: Positional breakdown of each of new 5 clusters

When we look at the positional breakdown of each of our five new definitions for players, we see a clear distinction and breakaway from the five conventional positions in basketball, as show in Figure 13. This further emphasises the breakaway from the traditional positions and skillsets seen in basketball in this proposition for reassessing the players on an NBA team. It is also worth noting the strong dominance of certain positions in these clusters in the figure seen above. As previously stated, the Center position is a growing force in the NBA in terms of its dominance in top performing athletes and its significantly increased statistical output in recent years. This point is accentuated seeing a strong presence in both Cluster 0 and 3.

In itself, a strong argument could be made that these five groupings encapsulate what teams should be looking for in their five players on a basketball court at any one time. It matches a blend on all key attributes and requirements with any shortcoming of one cluster being addressed as a strength in another and covering all ranges of shooting, as well as a blend of size and strength, with agile and hard working. With that being said, and as stipulated previously in Section 4.5, these clusters left too much variability in the players included. Whilst they form an excellent basis for the type of player and how they fit the positional need of a team, the second instance of clustering was introduced to give subcategories to these five and act as distinction of the talent levels therein. This is relevant in the NBA given the salary limits and thus the elimination of simply acquiring a team of best players in all categories.

5.2 Second-Tier Clustering

Each cluster outputted in stage one was taken in isolation with two new groupings created from each, amalgamating in ten total groups of players. Taking each visual output in a group of two (e.g., 0,0 and 0,1 in Figure 14) shows both clear similarities as well as distinct contrasts, underlining that both the skill separation of each of the initial five clusters, and an additional separation of insular ability, skill-set, and overall type of player has taken part.

Visually, it's difficult to argue against the benefit of the second level clustering. There is a clear distinction between each grouping while also maintaining broader similarities in the type of player from level one. The parametric testing of optimisation the output has, therefore, remained true in this execution in a manner that would not have been the case from one instance of larger cluster numbers. As stated, this was also explored for comparative purposes and yielded far less satisfactory and conclusive results.

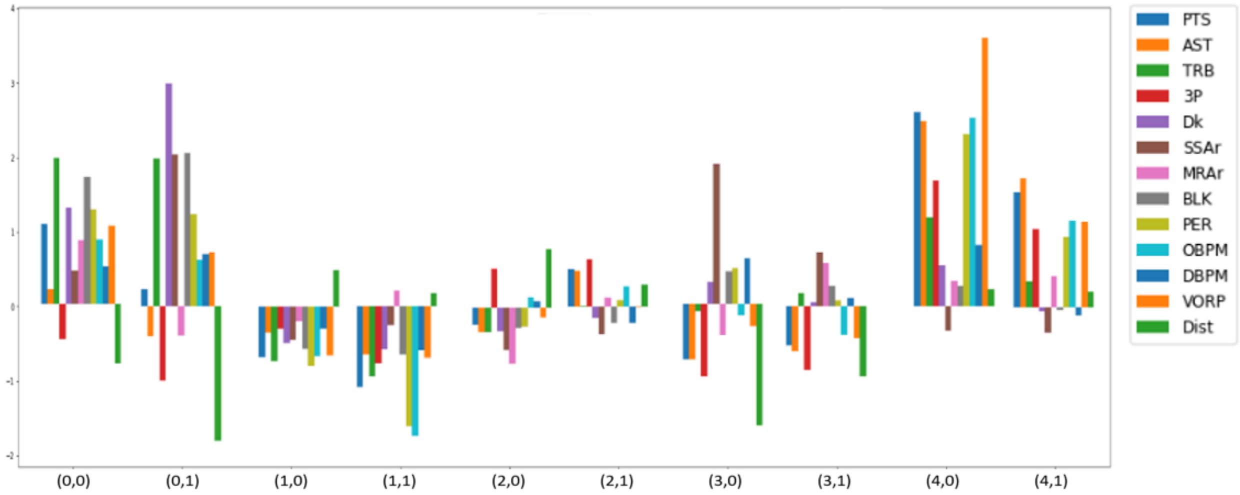


Figure 14: Graphical output after second level of clustering applied

The ten grouping in the output can therefore be described as the following.

Modern Big Men (0,0): The evolution of the largest players, who still dominate close range, but have grown to be the dominant point scorers on their teams from a range of distances, as well as being strong rebounders and all-round talented contributors.
Players: Giannis Antetokounmpo, Anthony Davis, Karl-Anthony Towns, Aaron Gordon

Traditional Big Men (0,1): These players represent the more traditional tall player roles in that they utilise their size and strength in close range scoring, dunks, and defence in particular. They lack the diverse skills of the modern adaptations, but are still pivotal in the game, even if not the primary scorers or playmakers.
Players: Rudy Gobert, Andre Drummond, Jarrett Allen

Corner Players (1,0): Whilst not the most skilled, these players show high usage in 3-point attempts, particularly from the very corner of the court. Deemed the most difficult shot to make in basketball, players who occupy this region are of great benefit to a team, even if only to act as an option for teammates on the court.
Players: Austin Rivers, Avery Bradley, Wesley Matthews

Role Players (1,1): These players would be brought into a team to perform a specific role. Basketball is a high paced game with many substitutions to allow player to catch their breath. Given the salary cap, players who can occupy this replacement role well and on salary friendly contracts are still valuable to teams, even if unglamourous.
Players: Jarrett Culver, Alfonzo McKinnie, Mychal Mulder

Pure 3-Point Shooters (2,0): Players who excel at the 3-point shot are in great demand and command high wages as a result. These players are purely that; 3-point shooters with some minor defensive attributes. While it sounds limited, all of the players in this group are more valuable in the modern NBA now more than ever.
Players: Duncan Robinson, Danny Green, Joe Harris

3-and-D (2,1): Where the pure shooters lack much else, this grouping of 3 and D excels at both 3-point shooting and at defence. They do not tend to go all in on the 3-point attempts as the previous grouping, but they have more attributes in terms of defensive work rate and in assisting other players with their passing.
Players: Klay Thompson, Lonzo Ball, Buddy Hield

Rim to Rim Specialists (3,0): Players who excel in getting from one end of the court to the other. These players shine in close range shooting when in attack, and in defensive box scores on the other end. They are generally taller player with more speed and athleticism than the two Big Men groupings. This combinations of size and speed is their major strength.
Players: DeAndre Jordan, Jaxson Hayes, Gary Payton II

Rebounders (3,1): Similar to the above, however they lack in the athleticism and make amends by shooting from further out. These are strong players who utilise this to rebound and

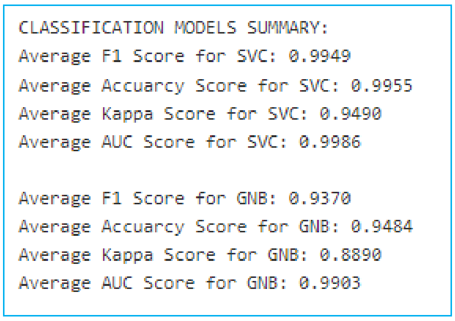
defend well, as well as a good passing range. But their lack of speed means their role is often to hand the ball over to more skilled handlers.
Players: LaMarcus Aldridge, DeMarcus Cousins, Boban Marjanović

Superstars (4,0): The most well-rounded players with the widest range of skills. This cluster was comparable to the All-Star selection each year showing it contains some of the most talented well-known names. Not as strong as Big Men and worse in some shooting attributes to other clusters, but no group contains players with talent levels as comprehensive as here.
Players: LeBron James, Kevin Durant, Joel Embiid, Stephen Curry, Nikola Jokić

Scoring and Rising Stars (4,1): Many similar traits to the Superstars, however this group possess less of the versatile attributes such as defence and rebounding, and instead are more pure scorers and shooters of the ball. Prominent names on this list suggest these players would fit the role as a support player alongside the superstars. The presence of younger rising stars here also shows the scouting and talent identification potential as many can be seen as future Superstars.
Players: Kyrie Irving, Devin Booker, Damian Lillard, Paul George, Darius Garland

5.3 Classification Verification

The inclusion of the classification algorithms in the methodology act as a means of validation in the talent level separation in the second level of clustering. Clustering-Based Classification, as it can be referred to (Krishnaveni and Radha, 2021) has been used in other studies as means of synergising the benefits of both branches of machine learning in an easily interpretable manner. Given there is often no real measure of right or wrong, nor good or bad when it comes to clustering and unsupervised learning, the inclusion of the SVC and GNB allows for accuracy metrics to be used in this second tier, given we have opted for a two-output model per initial five grouping. Meaning there is, in essence, a binary repose. Looking to the average outputs per algorithm across the five distinct models, the scores show to be incredibly high, thus giving us a form of validation to the hypothesised skill set separation of level two to this multi-level clustering solution. Subjectivity and trust in the model and the parameters dictate the output of the first instance of clustering, however, the level left to interpretation can be alleviated in the second level of clustering in this manner.



```
CLASSIFICATION MODELS SUMMARY:
Average F1 Score for SVC: 0.9949
Average Accuarcy Score for SVC: 0.9955
Average Kappa Score for SVC: 0.9490
Average AUC Score for SVC: 0.9986

Average F1 Score for GNB: 0.9370
Average Accuarcy Score for GNB: 0.9484
Average Kappa Score for GNB: 0.8890
Average AUC Score for GNB: 0.9903
```

Figure 15: Output scores of SVC & GNB runs

6 Conclusion and Discussion

The changing nature of positional play in the NBA has been highlighted, with basketball being one of the only major sports never to have had major overhauls to positions, formations, or role expectations of the five teammates on a court at one time. Experiments were conducted throughout, but ultimately the parameters with optimal results from testing elements such as silhouette coefficient and elbow method, as well as testing with varying inputs and variations, showed the two-tiered solution to have the most meaningful results. This was supported by the high SVC and GNB outputs seen in the final stage of the model.

A novel clustering scheme based on player statistical output of near 50 metrics has been identified in reclassifying the outdated positions in basketball to give five new groupings. This was broken down further both on a skill set and type of play level with insular clustering of each of these to give two subcategories of players within each of these groups, totalling ten positions or type of players that better describe the required roles of players in the NBA.

Given the nature of unsupervised learning, there is, however, always an element of subjectivity in the output. As such, the evaluation of the first clustering instance relied largely on trusting the algorithm and parameters, as well as personal interpretation or the need for prior knowledge in basketball. Should there be a viable method to evaluate the initial clustering in this multi-level solution, that would greatly improve both credibility and findings, so any future work would centre heavily on this. The second clustering iteration and inclusion of the classification models intended to alleviate some of these concerns.

Referring back to the research question proposed at the start of this paper, it has been shown that a novel two-tiered clustering algorithm can be used to redefine basketball positions and even aid in re-evaluating player performance within these clusters. It has been shown that through vast amount of data collecting and principal component analysis, a model has been developed based on K-Means to recategorize the five positions in an NBA team. Building upon this, it was also show that insular re-clustering and the use of SVC and GNB metrics, further groupings were made on a primarily talent separation level to a high standard.

“Machine learning is emerging as a powerful, new paradigm for sports analytics, as it provides novel approaches to making sense of the collected data” (Brefeld et al. 2022). Practically, I believe this research could be beneficial in helping to take new approaches in how positions and play styles are seen in basketball, as well as this prove to be invaluable in terms of recruiting, tactics, team building, coaching, scouting, and much more, along with general understanding for fans. But, academically speaking, the goal is ultimately to show the usefulness of such algorithms in the world of machine learning in sport and to spark a discussion on the novel approaches that can be taken. Much was made in this paper about the staleness of something that once showed so much excitement and still continues to garner more interest each day. Novel explorations such as this will, hopefully, continue to aid in this growth and interest and provide further exciting times ahead in this field.

References

Ahmed, M., Seraj, R., and Islam, S.M.S., (2020). “The k-means algorithm: A comprehensive survey and performance evaluation.” Electronics, 9(8), p.1295.

Aljarah, I., Habib, M., Nujoom, R., Faris, H. and Mirjalili, S., (2021). “A comprehensive review of evaluation and fitness measures for evolutionary data clustering.” Evolutionary Data Clustering: Algorithms and Applications, pp.23-71.

Azevedo, A., (2019). “Data mining and knowledge discovery in databases”. In Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics (pp. 502-514). IGI Global.

Bafna, P. B., and Jatinderkumar, R. S., (2019). “Identification of significant challenges in the sports domain using clustering and feature selection techniques”. In: 2019 9th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-19). IEEE, pp. 1–5.

Bai, Z. and Xiaomei, B., (2021). “Sports Big Data: Management, Analysis, Applications, and Challenges”. In: Complexity 2021.

Bouhmala, N., Viken, A., and Lønnum, J. B., (2016). "A multilevel K-Means algorithm for the clustering problem,". *IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA),* 2016, pp. 115-121, doi: 10.1109/ICCCBDA.2016.7529544.

Brefeld, U., Davis, J., Lames, M. and Little, J.J., (2022). “Machine Learning in Sports (Dagstuhl Seminar 21411)”. In Dagstuhl Reports (Vol. 11, No. 9). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Bunker, R. P. and Thabtah, F., (2019). “A machine learning framework for sport result prediction”. In: Applied computing and informatics 15.1, pp. 27–33.

De Diego, I. M. et al., (2019). “Subjective data arrangement using clustering techniques for training expert systems”. In: Expert Systems with Applications 115, pp. 1–15.

Duman, A. E., Sennaroglu, B., and Tuzkaya, G., (2021). “A cluster analysis of basketball players for each of the five traditionally defined positions”. In: Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology, p. 17543371211062064.

Ester, M. and Sander, J., (2013). “Knowledge discovery in databases: Techniken und Anwendungen. Springer-Verlag”.

Ferreira, P., (2018). “Chosen International Clubs Approach to Technical Scouting”. B.S. thesis. Høgskolen i Molde-Vitenskapelig høgskole i logistikk.

Gambella, C., Ghaddar, B. and Naoum-Sawaya, J., (2021). “Optimization problems for machine learning: A survey”. European Journal of Operational Research, 290(3), pp.807-828.

Hahsler, D., Piekenbrock, M. and Doran, D., (2019) “Fast density-based Clustering with R”, Journal of Statistical Software, (91), p.1.

Horvat, T and Job, J., (2020). “The use of machine learning in sport outcome prediction: A review”. In: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10.5, e1380.

Huang, T., Chih-Jen, L., and Ruby, C. W., (2008). “Ranking Individuals by Group Comparisons.” In: Journal of Machine Learning Research 9.10.

Inuwa-Dutse, I., Liptrott, M. and Korkontzelos, I., (2021). “A multilevel clustering technique for community detection”. Neurocomputing, 441, pp.64-78.

Kaushik, M. and Mathur, B., (2014). “Comparative study of K-means and hierarchical clustering techniques.” International Journal of Software & Hardware Research in Engineering, 2(6), pp.93-98.

Khobdeh, S. B., Yamaghani, M. R., and Sareshkeh, S. K., (2021). “Clustering of basketball players using self-organizing map neural networks”. In: Journal of Applied Research on Industrial Engineering 8.4, pp. 412–428.

Krishnaveni, N. and Radha, V., (2021). “Performance evaluation of clustering-based classification algorithms for detection of online spam reviews”. In Data Intelligence and Cognitive Informatics (pp. 255-266). Springer, Singapore.

Manish, S, Bhagat, V., and Pramila, R. M., (2021). “Prediction of Football Players Performance using Machine Learning and Deep Learning Algorithms”. In: 2021 2nd International Conference for Emerging Technology (INCET). IEEE, pp. 1–5.

Nagarajan, R., Zhao, Y., and Li, L., (2018). “Effective NBA player signing strategies based on salary cap and statistics analysis”. In: 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA). IEEE, pp. 138–143.

Nandakumar, N. and Jensen, S. T., (2019). “Historical perspectives and current directions in hockey analytics”. In: Annual review of statistics and its application 6, pp. 19–36.

Narizuka, T. and Yamazaki, Y., (2019). “Clustering algorithm for formations in football games”. In: Scientific reports 9.1, pp. 1–8.

Nguyen, N. H. et al., (2021). “The application of machine learning and deep learning in sport: predicting NBA players’ performance and popularity”. In: Journal of Information and Telecommunication, pp. 1–19.

Page, G. L. and Quintana, F. A., (2015). “Predictions based on the clustering of heterogeneous functions via shape and subject-specific covariates”. In: Bayesian Analysis 10.2, pp. 379–410.

Radovanovic, S. et al., (2013). “A novel approach in evaluating efficiency of basketball players”. In: Management 67, pp. 37–45.

Ren, M., Liu, P., Wang, Z. and Yi, J., (2016). “A self-adaptive fuzzy c-means algorithm for determining the optimal number of clusters”. Computational intelligence and neuroscience, 2016.

Sampaio, J. et al., (2015). “Exploring game performance in the National Basketball Association using player tracking data”. In: PloS one 10.7, e0132894.

Tan, J., et al., (2021). “A critical look at the current train/test split in machine learning”. arXiv preprint arXiv:2106.04525.

Terner, Z. and Franks, A., (2021). “Modeling player and team performance in basketball”. In: Annual Review of Statistics and Its Application 8, pp. 1–23.

Xia, K. et al., (2020). “Racquet sports recognition using a hybrid clustering model learned from integrated wearable sensor”. In: Sensors 20.6, p. 1638.

Xin, L., Zhu, M., and Chipman, H., (2017). “A continuous-time stochastic block model for basketball networks”. In: The Annals of Applied Statistics 11.2, pp. 553–597.

Key for Column Names of Scraped Data

Dataset 1 – Player Stats Per Game

Player
Pos - Position
Age
Tm - Team
G - Games
GS - Games Started
MP - Minutes Played
FG - Field Goals
FGA - Field Goal Attempts
FG% - Field Goal Percentage
3P - 3-Point Field Goals
3PA - 3-Point Field Goal Attempts
3P% - 3-Point Field Goal Percentage
2P - 2-Point Field Goals
2PA - 2-Point Field Goal Attempts
2P% - 2-Point Field Goal Percentage
eFG% - Effective Field Goal Percentage; the formula is (FG + 0.5 * 3P) / FGA.
FT - Free Throws
FTA - Free Throw Attempts
FT% - Free Throw Percentage
ORB - Offensive Rebounds
DRB - Defensive Rebounds
TRB - Total Rebounds
AST - Assists
STL - Steals
BLK - Blocks
TOV - Turnovers
PF - Personal Fouls
PTS - Points

Dataset 2 – Advanced Statistics

Player
Pos - Position
Age
Tm - Team
G - Games
MP - Minutes Played
PER - Player Efficiency Rating; a measure of per-minute production
TS% - True Shooting Percentage; a formula taking various other shooting metrics into account
3PAr -3-Point Attempt Rate; percentage of FG attempt from 3-point range
FTr -Free Throw Attempt Rate; number of FT attempts per FG attempt
ORB% - Offensive Rebound Percentage; percentage of offensive rebounds available a player had
DRB% - Defensive Rebound Percentage; percentage of defensive rebounds available a player had
TRB% - Total Rebound Percentage; percentage of total rebounds available a player had
AST% - Assist Percentage; percentage of assists available a player had
STL% - Steal Percentage; percentage of steals available a player had
BLK% - Block Percentage; percentage of blocks available a player had
TOV% - Turnover Percentage; percentage of turnovers available a player had
USG% - Usage Percentage; percentage of team plays used by player while on the floor
OWS - Offensive Win Shares; estimate of number of wins contributed by player due to offense
DWS - Defensive Win Shares; estimate of number of wins contributed by player due to defence
WS - Win Shares; estimate of number of wins contributed by player
WS/48 - Win Shares Per 48 Minutes
OBPM - Offensive Box Plus/Minus; the offensive points per 100 possessions a player contributed above the league average
DBPM - Defensive Box Plus/Minus; the defensive points per 100 possessions a player contributed above the league average
BPM - Box Plus/Minus
VORP - Value Over Replacement Player; estimate per 100 team possessions

Dataset 3 – Shooting Statistics

Player
Pos - Position
Age
Tm - Team
G - Games
MP - Minutes Played
FG% - Field Goal Percentage
Dist. - average distance (ft.) of Field Goal Attempts

% Of FGA by Distance - distance of field goals attempted percentage
  2P - 2-point range
  0-3 - feet
  3-10 - feet
  10-16 - feet
  16-3P - 16 feet to 3-point mark
  3P - 3-point range
SSAR – short shooting attempt rate – % Of FGA by Distance 0-3 feet
MRAR – mid-range attempt rate – % Of FGA by Distance 3-16 feet

FG% by Distance - distance of field goals made percentage
  2P - 2-point range
  0-3 - feet
  3-10 - feet
  10-16 - feet
  16-3P - 16 feet to 3-point mark
  3P - 3-point range

% Of FG Ast'd -percent of team's field goals assisted
  2P - 2-point range
  3P - 3-point range

Dunks
  %FGA - number of field goal attempts that are dunk attempts
  # - number of made dunk attempts

Corner 3s - a 3-point shot from the very corner of the rectangular court
  %3PA - number of 3-point attempts that are corner 3 attempts
  3P% - corner 3 shooting percent

Heaves - shots from beyond half-court
  Att. - number of shots from beyond half court attempted
  # - number of shots from beyond half court made