Evaluating Combinations of Play Styles in the NBA

by

Dan Pilsbury

S.B. in Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2022

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by: Dan Pilsbury
Department of Electrical Engineering and Computer Science
May 19, 2023

Certified by: Anette "Peko" Hosoi
Associate Dean of Engineering; Neil and Jane Pappalardo Professor of Mechanical Engineering
Thesis Supervisor

Accepted by: Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Evaluating Combinations of Play Styles in the NBA

by

Dan Pilsbury

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2023, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

This thesis explores the influence of player play styles on offensive performance in the NBA. It offers valuable insight to coaches and managers who seek to understand the significance of play styles and identify the optimal combinations. Through classifying player play styles and analyzing their relationships with team performance, this research reveals that play styles have a tangible impact on performance, even when adjusting for individual skill levels. The findings highlight the importance of three point shooting, the ability to create shot opportunities for teammates, and the benefit of court spacing. Offensive performance is not simply equivalent to the sum of the individual talents, but can be greater or less depending on how the styles of players complement each other. Coaches can use this information to make informed decisions about what players to acquire for optimal offensive performance.

Thesis Supervisor: Anette "Peko" Hosoi
Title: Associate Dean of Engineering; Neil and Jane Pappalardo Professor of Mechanical Engineering

Acknowledgments

I sincerely thank my thesis advisor Peko Hosoi for the invaluable support and guidance, as well as for providing me with the opportunity to work on an incredibly interesting project.

I would also like to express gratitude to my parents. I would not be where I am today without their love and support throughout my life. I am also grateful for my sister Kim, who has been a great role model for me.

My appreciation goes to my friends, and particularly my roommates Peyton and George, for their support. Lastly, I would like to thank Sigma Chi for lasting friendships and support in my academic career.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

"How can I make my team better?" is a question that every National Basketball Association (NBA) owner, general manager, and coach wants to know the answer to. Each NBA team has limited money, time, and players; so making informed decisions regarding resource allocation becomes crucial for optimizing team success. In each season, NBA teams navigate player drafts, trades, and acquisitions to construct a team that provides the best chance at winning games. A simple approach may be to just try to get the 'best' possible players. However, this is not always possible because of limitations in salary cap room and it is not always obvious who the best fit is for a team. Teams want to identify players who can contribute most effectively to their team's success relative to other potential players.

A traditional approach to constructing a basketball lineup has relied on player positions. The five standard positions on a basketball team include the point guard, shooting guard, small forward, power forward, and center [8]. Each position typically corresponds to a specific role, play style, and physical attributes. For instance, centers are usually tall players specializing in rebounding and playing close to the basket, whereas point guards are generally shorter players responsible for facilitating the offense with assists and playing on the perimeter. A conventional approach to constructing a lineup would entail having one player in each position. However, this model may not always produce the most optimal lineup for several reasons.

Firstly, some players do not fit neatly into one of these five defined positions. For example, on nba.com, LeBron James is listed as a forward [8]. However, Lebron often assumes the role of point guard by bringing the ball up and facilitating the offense, while in other instances plays off the ball like a shooting guard, or plays in the post like a power forward. Similarly, certain point guards prioritize facilitating the offense like Chris Paul, while others focus more on scoring like Kyrie Irving. Developing a more accurate model that captures players' unique play styles would offer a higher level of granularity in determining the ideal lineup composition.

Secondly, it is not obvious that an optimal lineup should consist of players from each primary position. The intuition of having one player in each position is to maximize performance across key basketball concepts that contribute to winning, such as rebounding, shooting threes, getting assists, and driving to the basket. With one player in each position, each concept can be addressed by a specialized player, such as a rebounding big man, sharpshooting guard, and a facilitating point guard. It makes sense that a team would not want a lineup of all centers because they would likely not be able to space the court well or defend opponents' guards. However, teams may have lineups with two centers or two shooting guards, so it is useful to determine the benefits and drawbacks of specific player combinations. A recent example of when a team challenged the classical model of basketball lineups is the 2015 Warriors, who won the NBA finals. Their second most common lineup in the playoffs consisted of Stephen Curry, Klay Thompson, Harrison Barnes, Andre Iguodala, and Draymond Green [8]. This lineup did not have a traditional center position, but had Draymond Green fill that role at only 6 feet 6 inches tall. They were able to play a style of 'small ball' where they spaced the floor well, had great shooters, and put a lot of pressure on the defense to guard the perimeter.

It is evident that strictly adhering to the five classical positions is not necessary to perform well. This thesis aims to analyze the performance of various combinations of player styles to identify which styles are most complementary. The findings will contribute to a deeper understanding of how to construct an optimal lineup that maximizes success in the NBA.

1.2 Objectives

This thesis aims to investigate the combinations of play styles in the NBA that perform the best. To achieve this end goal, there are several steps.

The initial step involves identifying a new model for positions by classifying each player according to their play style. The goal is to have a more accurate representation of player positions than the conventional five position framework.

Subsequently, a model will be used to predict the number of points scored on each possession, taking into account the offensive and defensive players on the court. This provides an expected value for each lineup that can be compared to the actual performance.

The third step is to evaluate actual lineups by comparing their performance to the expected performance. This step will provide concrete examples of what players outperform or under-perform relative to what is expected.

The ultimate objective is to group lineups based on their player profiles and evaluate their performance relative to expectations. For example, this evaluation would provide a performance metric for every combination of player types, and insights can be drawn about what combinations exceed expectations and which ones do not.

1.3 Related Work

The concept of creating new player profiles to capture play styles has been explored in previous academic papers and articles [6][7][4]. These works generally employ a similar approach to the one chosen in this thesis. The typical methodology involves selecting specific features that reflect play style and applying clustering algorithms to the resulting data. However, a key difference between this thesis and other related works lies in the choice of features. While this thesis focuses exclusively on offensive features, other studies often consider both offensive and defensive aspects.

One notable piece of related work is a Harvard thesis titled "Evaluating Lineups and Complementary Play Styles in the NBA" [5]. My thesis shares similarities with this research as it classifies players into distinct player profiles. Additionally, it predicts lineup performance based on a regularized adjusted plus/minus (RAPM) model, but evaluates performance based on plus/minus rather than points per possession. In addition, a main focus of the Harvard thesis is improving the RAPM model by adding coefficients according to player profiles. Furthermore, the subsequent analysis in the Harvard thesis diverges significantly from the analysis in this thesis. The Harvard thesis focuses on individual player and lineup performance relative to replacement level players, where this thesis focuses on how combinations of player profiles perform relative to what is expected.

By highlighting these similarities and differences, it becomes evident that while prior research has addressed similar topics, this thesis offers a unique perspective by narrowing the focus to offensive performance evaluation and introducing new methods of evaluating groups of play styles.

Chapter 2

Data

This project utilized a mix of publicly available data as well as private data collected by Second Spectrum, which was provided to me by the San Antonio Spurs. The primary focus of analysis in this thesis was the 2018-2019 NBA season data obtained from these sources. This section is an overview of the data sources used, with a description of the data content. However, it is important to note that the specific rationale and methodology employed in the utilization of this data will be discussed in the subsequent Methods section.

2.1 Second Spectrum Data

Second Spectrum is the official tracking provider for the NBA, and they provide comprehensive game data to all NBA teams. For this project, I had access to Second Spectrum's data for the 2018-2019 season, which was provided to me by the San Antonio Spurs. Second Spectrum processes their collected data and organizes it into tables that capture specific game events. For instance, they have a table storing data for every possession, which includes the players on the court and the corresponding number of points scored. Other tables include detailed information on player dribbles and player shots. In this analysis, the possessions data was used to extract player lineups and points scored per possession. This table contained over 260,000 possessions in the 2018-2019 season. Additionally, the dribbles table was used to obtain the dribbles per possession for each player, containing one row for each player.

To access and explore this data, I used the Google Cloud Platform. This was valuable during the exploration stage of my project, providing convenient access to table schemas and previews. To add, the python package 'bigquery' from google.cloud facilitated data querying. I was able to create a BigQuery client to connect to the Google Cloud Project that contained the possessions and dribbles data. With the appropriate credentials in place, I was able to construct SQL queries to retrieve the desired data efficiently.

	pts_scored	name_o1	name_o2	name_o3	name_o4	name_o5	name_d1	name_d2	name_d3	name_d4	name_d5
0	0	Malik Monk	Marvin Williams	Willy Hernangomez	Devonte' Graham	Kemba Walker	Ricky Rubio	Joe Ingles	Ekpe Udoh	Thabo Sefolosha	Donovan Mitchell
1	3	Malik Monk	Marvin Williams	Willy Hernangomez	Devonte' Graham	Kemba Walker	Ricky Rubio	Joe Ingles	Ekpe Udoh	Thabo Sefolosha	Donovan Mitchell
2	0	Malik Monk	Marvin Williams	Michael Kidd-Gilchrist	Devonte' Graham	Nicolas Batum	Damyean Dotson	Kevin Knox	Tim Hardaway Jr.	Ron Baker	Mitchell Robinson
3	0	Malik Monk	Kemba Walker	Marvin Williams	Devonte' Graham	Bismack Biyombo	Harry Giles III	Justin Jackson	Yogi Ferrell	Bogdan Bogdanovic	Marvin Bagley III
4	0	Malik Monk	Marvin Williams	Devonte' Graham	Nicolas Batum	Bismack Biyombo	Jeremy Lin	DeAndre' Bembry	Alex Len	Vince Carter	Kent Bazemore

Figure 2-1: Example Subset of Second Spectrum Possessions Data

	player_name	dribbles
0	Kevin Huerter	4755
1	Vince Carter	1692
2	Daniel Hamilton	651
3	Alex Len	1055
4	BJ Johnson	51

Figure 2-2: Example Subset of Second Spectrum Dribbles Data

2.2 basketball-reference Data

In constructing player profiles, I relied on player statistics like shooting percentage and points per possession. While these statistics could have been derived from the Second Spectrum data, doing so would have required data processing, which would introduce the potential for calculation errors. To ensure accuracy and streamline the process, I opted to utilize publicly available statistics from basketball-reference.com [2]. This website offers a variety of statistics, and for this project I focused on three main data tables. The first table I used was the per possession stats (Figure 2-3), which includes points, field goal attempts, three point attempts, and rebounds. Additionally, I utilized the shooting stats table (Figure 2-4), which includes the percent of field goal attempts for different regions on the court. Lastly, I utilized the advanced stats table (Figure 2-5), which includes three point attempt rate, and percentage of available rebounds grabbed. I was able to save these tables as csv files on my local machine for use in my analysis.

Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	ORtg	DRtg
1	Álex Abrines	SG	25	OKC	31	2	588	4.4	12.5	.357	3.3	10.1	.323	1.2	2.4	.500	1.0	1.0	.923	0.4	3.4	3.8	1.6	1.3	0.5	1.1	4.2	13.1	103	111
2	Quincy Acy	PF	28	PHO	10	0	123	1.6	7.0	.222	0.8	5.8	.133	0.8	1.2	.667	2.7	3.9	.700	1.2	8.5	9.7	3.1	0.4	1.6	1.6	9.3	6.6	87	116
3	Jaylen Adams	PG	22	ATL	34	1	428	4.1	11.9	.345	2.7	8.0	.338	1.4	3.9	.361	0.8	1.0	.778	1.2	5.3	6.5	7.0	1.5	0.5	3.0	4.9	11.7	99	115
4	Steven Adams	C	25	OKC	80	80	2669	8.4	14.1	.595	0.0	0.0	.000	8.4	14.1	.596	2.6	5.1	.500	6.8	6.5	13.3	2.2	2.0	1.3	2.4	3.6	19.4	120	106

Figure 2-3: Example basketball-reference per 100 possessions data

									% of FGA by Distance						FG% by Distance						% of FG Ast'd	
Rk	Player	Pos	Age	Tm	G	MP	FG%	Dist.	2P	0-3	3-10	10-16	16-3P	3P	2P	0-3	3-10	10-16	16-3P	3P	2P	3P
1	Álex Abrines	SG	25	OKC	31	588	.357	22.3	.191	.102	.025	.019	.045	.809	.500	.500	.500	.333	.571	.323	.800	.951
2	Quincy Acy	PF	28	PHO	10	123	.222	21.7	.167	.111	.000	.056	.000	.833	.667	1.000		.000		.133	1.000	1.000
3	Jaylen Adams	PG	22	ATL	34	428	.345	20.2	.327	.118	.082	.073	.055	.673	.361	.538	.111	.250	.500	.338	.231	.840
4	Steven Adams	C	25	OKC	80	2669	.595	3.5	.998	.645	.294	.058	.000	.002	.596	.667	.458	.511		.000	.642	

Figure 2-4: Example basketball-reference shooting data

Rk	Player	Pos	Age	Tm	G	MP	PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	USG%	OWS	DWS	WS	WS/48
1	Álex Abrines	SG	25	OKC	31	588	6.3	.507	.809	.083	0.9	7.8	4.2	4.3	1.3	0.9	7.9	12.2	0.1	0.6	0.6	.053
2	Quincy Acy	PF	28	PHO	10	123	2.9	.379	.833	.556	2.7	20.1	11.3	8.2	0.4	2.7	15.2	9.2	-0.1	0.0	-0.1	-0.022
3	Jaylen Adams	PG	22	ATL	34	428	7.6	.474	.673	.082	2.6	12.3	7.4	19.8	1.5	1.0	19.7	13.5	-0.1	0.2	0.1	.011
4	Steven Adams	C	25	OKC	80	2669	18.5	.591	.002	.361	14.7	14.8	14.7	6.6	2.0	2.4	12.6	16.4	5.1	4.0	9.1	.163

Figure 2-5: Example basketball-reference advanced data

2.3 Regularized Adjusted Plus Minus Data

Briefly, adjusted plus minus (APM) is a metric that measures the impact of each player on his team's scoring margin after controlling for the strength of each teammate and opponent across each minute played. This metric is enhanced with regularization, which improves the accuracy of the metric when predicting outcomes of future games. More about how this is calculated and how it applies to this thesis is in the Methods section. For now, I will just go over the source and contents of the data. Although the Second Spectrum possessions data could be utilized to derive this metric, I prioritized accuracy and efficiency by choosing to use publicly available metrics. The regularized adjusted plus minus (RAPM) metrics for each player are available at basketball-analytics.gitlab.io [1]. The data provides an offensive, defensive, and total RAPM metric for every player. I was able to access this data through an api endpoint from the website.

Rank	Player	Team	Poss	ORAPM	DRAPM	RAPM
1	Kevin Durant	GSW	12315	5.2647	1.4193	6.684
2	Danny Green	TOR	9830	4.297	2.1749	6.4719
3	Stephen Curry	GSW	10669	4.3916	1.0492	5.4409
4	Paul George	OKC	13310	1.8271	3.5296	5.3567

Figure 2-6: Example RAPM data

Chapter 3

Methods

The goal of this analysis is to evaluate what combinations of play styles outperform or underperform expectations. The first step is creating player profiles that accurately represent play styles. This involves selecting the features that encompass all aspects of possible play styles, and then clustering players based on these features. The next step is predicting the points scored per possession for each lineup. Using these predictions, lineups can be evaluated on their performance relative to the expected points scored. Finally, the same evaluation can be done after grouping lineups by the play styles of the individual players in the lineup.

It is important to note that this analysis focuses only on offensive performance. Basketball is a two way sport, meaning players play both offense and defense. However, when evaluating how well play styles complement each other, I believe focusing on offensive styles is an intuitive first step. In basketball, the offensive players are the ones that dictate the action, whereas the defense acts in reaction to the offense. For instance, the offensive players can position themselves anywhere on the court, and then the defense positions themselves based on where the offensive players are. Therefore, the offense has more control in the style of play. The offense can dictate the pace of the game, types of shots taken, and direction of movement on the court. Yes, the defense can affect the offense if they choose to pressure players on the perimeter versus playing back and crowding the paint. However, in comparison, the offense dictates the style of play much more. Therefore, the player profiles will be determined by offensive stats, and the lineups will be evaluated by the points scored per possession.

3.1 Constructing Player Profiles

3.1.1 Feature Selection

When constructing profiles that represent the style of each player, it is important to consider what features of a player's performance distinguish one player from another. The mainstream statistics used to evaluate players are the points, rebounds, and assists per game. At a high level, these stats encompass the three most important quantitative aspects of basketball. For building accurate profiles, more nuanced and detailed features will be used. When distinguishing offensive player styles, there are features under the categories of scoring, facilitating, rebounding, and ball control. In addition, the features chosen are all rate statistics where the denominator is equal among all players so that things like the number of minutes played do not affect the scale of the metrics.

Table 3.1: Statistics to Represent Offensive Play Style

Metric	Numerator	Denomenator	Aspect of Play Style
PTS/ 100 pos	points	possessions / 100	ability to score
2PA/ 100 pos	2-point attempts	possessions / 100	2-point volume
3PA/ 100 pos	3-point attempts	possessions / 100	3-point volume
% of FGA per region	FGA within region	total FGA	distribution of shot selection
2-point FG%	2-point FGM	3-point FGA	efficiency from 2-point range
3-point FG%	3-point FGM	3-point FGA	efficiency from 3-point range
% of FGM assisted 2P	assisted FGM 2P	FGM 2P	ability to create 2-point shots
% of FGM assisted 3P	assisted FGM 3P	FGM 3P	ability to create 3-point shots
FT%	free throws made	free throw attempts	ability to make free throws
FT/pos	free throw attempts	possessions	ability to get to the foul line
ORB%	offensive rebounds	available offensive rebounds	ability to get offensive rebounds
AST%	assists	total FGM	ability to facilitate
TOV%	turnovers	possessions	ability to control the ball
dribbles/pos	dribbles	possessions	ball handling burden

Since the main objective of the offensive team in basketball is to score, most of these statistics are related to scoring and shooting. The style of scoring can be broken down into the concepts of how often a player shoots, how good they are at shooting, and the types of shots they take. To measure how often a player shoots, field goal attempts are used, broken down into two point attempts and three point attempts. To measure how good players are at shooting, their field goal percentage is used for both two point shots and three point shots. To measure the types of shots players take, the percent of players' made shots assisted is used. This represents how good players are at creating shots for themselves versus relying on teammates to create shots for them. In addition, the types of shots are measured by the distribution of shots by region. This is broken down into these five regions as the percent of total shots within this region.

1. 0-3 feet from the basket. This region represents layups and dunks.

2. 3-10 feet from the basket. This region represents shots close to the basket like floaters or short jumpshots.

3. 10-16 feet from the basket. This region represents mid range jump shots.

4. 16 feet - 3 point line. This region represents longer mid range jump shots.

5. 3 pointers. This region represents three point shots.

To measure a player's ability to facilitate their teammates, their assists as a percent of the team's total made shots is tracked. Another aspect of facilitation as well as ball control is the number of dribbles per possession a player takes. This indicates how much a team relies on a player to handle the ball. The last aspect of facilitation and ball control is how often players turn the ball over. This consists of all turnovers such as bad passes, traveling violations, and offensive fouls. Furthermore, the percent of offensive rebounds a player grabs is used to measure their ability to extend offensive possessions. Lastly, free throws are a significant aspect of offense, so players' abilities to get to the free throw line and to make free throws are included.

3.1.2 Clustering Play Styles

Using the features identified in the previous section, players can be clustered into groups that have similar values for each of the features. This was done with the k-means clustering algorithm. Before the cluster analysis, the data was engineered via normalizing the data and then performing dimensionality reduction.

The metrics used to represent offensive play style have very different ranges. For example, the points scored per 100 possessions may have a different scale than the three point attempt rate of a player. Therefore, the metrics for each player were standardized and scaled so that each feature had a mean of 0 and standard deviation of 1 for all players. This ensures features that have a different scale do not get weighted differently during the clustering process. In addition to scaling and standardizing the data, dimensionality reduction was performed to make the data more succinct and reduce multicollinearity since many of the features may be correlated with each other. The method of dimensionality reduction was principal component analysis. This is a technique that can be used to linearly reduce the number of dimensions while preserving the maximum amount of information in the data [11].

After performing standardization and principal component analysis, the player data was clustered using k-means clustering. This is an unsupervised learning algorithm that groups data points based on certain similarities [3]. This algorithm was chosen for a few reasons. Firstly, since it is unsupervised, it allowed for the creation of new labels without relying on predetermined player positions like point guard or center. Secondly, the k-means clustering algorithm enables the specification of the number of clusters. This allowed exploration in changing the number of clusters to view the results and then I could pick the number of clusters that make the most sense. I chose 8 clusters, which will be elaborated on in the Results section.

The k-means algorithm works by first specifying the number of clusters k, and then determining k centroids based on randomly sampled data. The algorithm iteratively adjusts the centroid locations to minimize the distance between all data points and the centroids. Finally, each data point is assigned to its nearest centroid. Formally, the algorithm is as follows:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 = \underset{S}{\operatorname{argmin}} \sum |S_i| \operatorname{Var} S_i$$

The set of n player vectors is $(x_1, x_2, ..., x_n)$ where each player vector is the set of features listed above. The k clusters are in sets $S = \{S_1, S_2, ..., S_k\}$. Once all players have an assigned cluster, the clusters can be analyzed to determine their shared qualities.

3.2 Predicting Points Per Possession

When evaluating how well a lineup of players or lineup of play styles performs, we don't want to just look at raw performance. If the evaluation were to just consider raw performance such as the points per possession, the results would be skewed based on the skill of the players. In basketball, the skill of players is a huge factor in performance. For example, even if there is a lineup of players that do not complement each other well, if they are all very skilled, they will likely still perform well. In contrast, there may be lineups that complement each other very well, but if they are not very skilled, their performance may still be below average. To account for the skill of the lineup, the expected performance can be modeled based on the skill of each player in the lineup. With the expected points scored of a lineup, the lineup's performance can be measured relative to what is expected.

Measuring the skill of players can be complicated, so this analysis focuses on a variation of the plus/minus of every player. The raw plus/minus metric for a player is simply the total points scored by that player's team minus the total points scored by the opponent while that player is on the court. However, this metric is highly dependent on what players are on the court. For instance, in the 2020-2021 NBA season, Anthony Edwards had a raw plus/minus of -228 in 76 games with the Minnesota Timberwolves [12]. However, he averaged 19.3 points and 4.7 rebounds per game and finished second in voting for the rookie of the year. This means that he must have had a great year, but he still had a terrible raw plus/minus because his team was not very good. Conversely, in that same season, Royce O'Neale had a raw plus/minus of +471, which was the third best in the league [12]. Royce averaged only 7 points per game, but he was on the Utah Jazz who had the best record in the league that season. Therefore, raw plus/minus does not properly account for an individual player's contribution since it does not factor in the other players on the court.

To mitigate the flaws in raw plus/minus, the metric adjusted plus/minus (APM) was developed. APM estimates a coefficient for each player that represents the impact of the player after controlling for the teammates and opponents on the court [9]. These coefficients are estimated with the following formula [13]

$$y_i = \sum_{o \in P} \beta_{\text{off},o}\, x_{\text{off},io} - \sum_{d \in P} \beta_{\text{def},d}\, x_{\text{def},id} + \beta_{\text{HCA}} x_{\text{HmOff},i} + \beta_{\text{const}}$$

y_i is the points scored on possession i and P is the set of all players. $x_{\text{off},ip}$ is 1 if player p is in the game on offense for possession i and 0 otherwise, $x_{\text{def},id}$ is 1 if player p is in the game on defense for possession i and 0 otherwise. $x_{\text{HmOff},i}$ is 1 if the offensive team for possession i is the home team and 0 otherwise. The coefficients $\beta_{\text{off},o}$ and $\beta_{\text{def},d}$ are the offensive and defensive APM ratings for player p. β_{HCA} is the coefficient to represent home court advantage, and β_{const} is a constant that represents the average points per possession, since the APM coefficients represent the impact relative to the average. The coefficients are estimated by minimizing the ordinary least squares in the model. The expected points scored on a possession can be calculated by plugging in the coefficients of the players on the court to this model. This is called adjusted plus/minus because each player's coefficients are adjusted based on the players on the court at the same time.

APM is a much better representation of a player's contribution to their team's performance, but there are still a few flaws that can be improved. One shortcoming is that there can be high correlation in outcomes for certain players that play together often. For instance, if there are certain lineups that often play together, each of those players will have similar outcomes in this model. In addition, APM can often overfit the data since there are many player coefficients relative to the number of possessions. To fix some of these issues, regularized adjusted plus/minus (RAPM) was developed. This metric introduces regularization to the APM model. Players' coefficients are estimated using the same model for APM, but there is an additional regularization term added to the model. The APM model minimizes the ordinary least squares with the following loss function

$$\sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

where y_i is the points predicted, x_i is the vector of predictors, and β is the vector of coefficients [5]. The RAPM model adds a regularization term, which changes the minimization function to

$$\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \beta^T \beta$$

λ is the regularization coefficient that indicates the weight of the the regularization [5]. The regularization term penalizes coefficients if their magnitude gets too large. The resulting RAPM metrics reduce overfitting and perform better when predicting out of sample outcomes.

The points scored on a possession can be predicted by plugging in the RAPM coefficients to the APM model. The RAPM coefficients represent a player's contribution to their team's performance relative to average players. Therefore, this model predicting points per possession accounts for each individual player's skill. Lineups where each player has a high RAPM coefficient will tend to have higher predicted performance. This model can be used as the expected points per possession of each lineup, and then the lineups can be evaluated relative to what is expected.

3.3 Evaluating Lineups

The possessions data contains the five offensive players, five defensive players, and the corresponding points scored for each possession. With this data, the expected points scored on each possession can be calculated with the RAPM model. Consequently, the total points scored, predicted points scored, and number of possessions can be aggregated for each offensive lineup of five players. Subsequently, the actual points per possession and predicted points per possession are derived. The evaluation metric is simply the difference between the actual points scored per possession and the predicted points scored per possession. The formula is the following

$$P_l = \frac{1}{n_l}(s_l - p_l)$$

where P_l is the performance of lineup l, s_l is the total points scored by lineup l, p_l is the predicted points scored by lineup l, and n_l is the number of offensive possessions played by lineup l. Positive values indicate performing above expectation and negative values denote performing below expectation. Each lineup of five players can be compared based on this metric. To ensure statistical reliability, only lineups that played a minimum of 500 possessions were considered in this evaluation. This threshold prevents exaggerated assessments caused by variance in small sample sizes.

3.4 Evaluating Profile Combinations

This section discusses how combinations of player profiles were evaluated. For clarity, any reference to a lineup will refer to actual players that played together. An example lineup is Stephen Curry, Klay Thompson, Harrison Barnes, Andre Iguodala, and Draymond Green. Any reference to a profile combination will refer to player labels (from the Player Profiles section) that played together. For instance, if each player in the above lineup had labels (1, 2, 3, 4, 5), this combination of labels is the profile combination. There may be multiple lineups that all represent the same profile combination.

The lineup evaluation provided the predicted points scored, actual points scored, and total possession for each lineup. This was used to group each lineup by its profile combination. The end goal is to evaluate these profile combinations. Evaluating these profile combinations is not straightforward because each lineup within a profile combination may have played a different number of possessions. For example, for a given profile combination, it may consist of one lineup that performed very well, but on relatively few possessions played, while another lineup within that profile combination may have performed very poorly on a large number of possessions played. The profile combinations can be evaluated weighting each lineup's performance equal, or they can be evaluated weighting each possession played as equal, which would put more weight on lineups that played more possessions. Therefore, the following two methods of evaluation were considered. The first method treats each lineup's performance within a profile combination as a single data point. The second method treats the performance of each possession for each lineup within a profile combination as a single datapoint.

3.4.1 Method 1: Weighting Each Lineup Equally

The argument for weighting each lineup equally is that the evaluation for a profile combination should not be skewed by lineups that have played a large number of possessions relative to the other lineups within a profile grouping. For example, if there is a lineup that plays very well together and performs very well, they will likely play together a lot. This may cause the evaluation of that lineup's profile combination to converge to the performance of that single lineup. If each lineup is weighted equally, then the number of possessions played will not impact the evaluation of the profile combination. Note that the lineups considered in this method still need to meet a threshold of number of possessions plates in order to reduce variance from small sample sizes.

This method groups every lineup by its profile combination. For a given profile combination, the evaluation metric for each lineup within that combination is considered, which is the difference in points scored per possession and predicted points scored per possession. Then the performance metric of each of the lineups within a profile combination is averaged to get the overall performance metric for that profile combination. The formula is the following

$$P_{c,1} = \frac{1}{|c|} \sum_{l \in c} P_l$$

where $P_{c,1}$ is the performance metric from method 1 for profile combination c, $|c|$ is the number of unique lineups that match profile combination c, P_l is the performance metric for lineup l. This provides a metric for each profile combination to be compared.

3.4.2 Method 2: Weighting Each Possession Equally

The argument for weighting each possession equally is that as the number of possessions a lineup plays increases, the confidence in that lineup's performance metric increases. Therefore, if there is a lineup within a profile combination that has played relatively few possessions, the statistical reliability of that lineup's performance metric is lower and it should be weighted less than a lineup with a larger sample size. The process for calculating the performance metrics of each profile combination in this method has the same initial approach to Method 1. However, once the performance metric of each lineup is calculated, instead of calculating the average, the weighted average is calculated where each lineup is weighted by the number of possessions they played together. The equation is the following

$$P_{c,2} = \frac{1}{t_c} \sum_{l \in c} P_l \, n_l$$

$$t_c = \sum_{l \in c} n_l$$

where $P_{c,2}$ is the performance metric from method 2 for profile combination c, t_c is the total number of possessions for all lineups that match c, P_l is the performance metric for lineup l, and n_l is the number possessions played by lineup l.

I believe that the argument for method 1 is stronger because it avoids correlation in high performing lineups playing more possessions together. An ideal approach may be a hybrid of these two methods, but this thesis focuses on method 1.

Chapter 4

Results

4.1 Player Profiles

There are two main aspects in analyzing the clusters created from the k-means clustering on players' statistics. Firstly, choosing the number of clusters is important because this analysis depends on having a profile for each player that is an accurate representation of play style. Secondly, quantitative and qualitative analysis on the resulting clusters is necessary to be able to interpret the player profiles.

4.1.1 Selecting Number of Clusters

The number of clusters is equivalent to the number of unique play styles this analysis assigned players. A mix of quantitative and qualitative analysis was used to determine the right number. This analysis wanted players to be grouped into clusters that provide a better representation of play style than the traditional five position framework, so as a general rule, more than five clusters is desirable. One quantitative approach is to examine the sum of squared distances of each player vector to the corresponding cluster centroid. As the number of clusters increases, the sum of distances should decrease, meaning that each player is closer to their assigned cluster centroid. Here is a figure graphing these distances where the x-axis is the number of clusters.

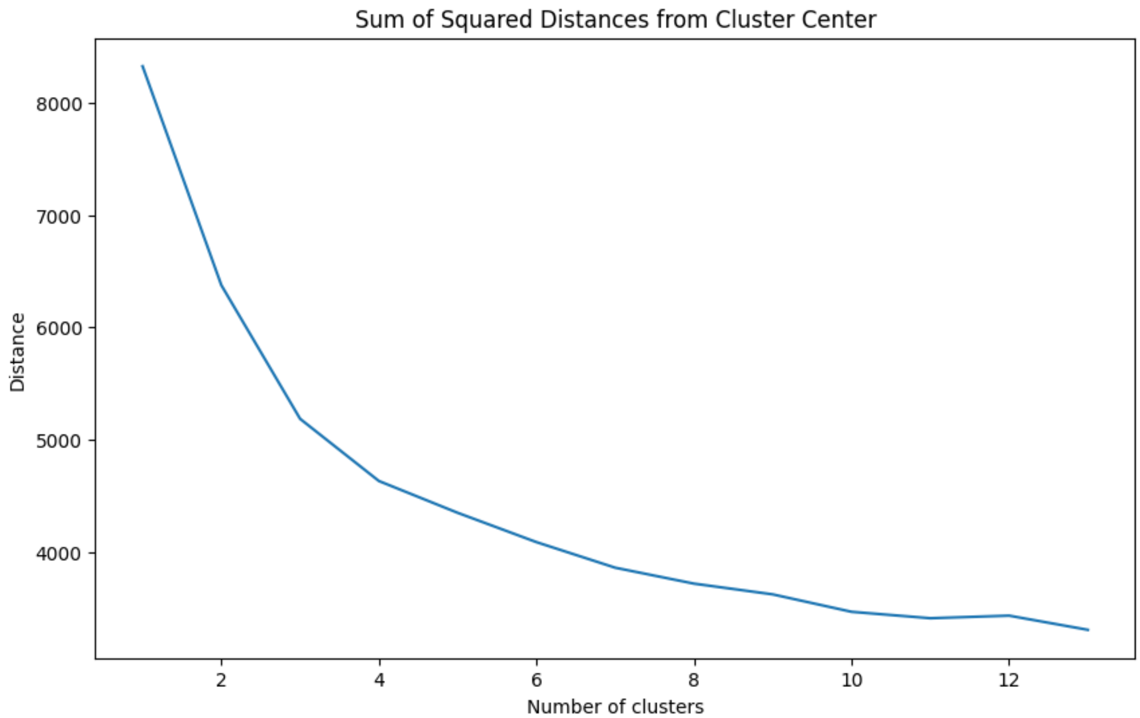

Figure 4-1: Sum of Squared Distances from Cluster Center vs Number of Clusters

A rule of thumb when analyzing these graphs is to look for an elbow point [3]. This is a point where adding one more cluster reduces the total distance significantly, while further additional clusters do not reduce the distance as much. In this graph, there is no clear elbow point. Consequently, each of the results from having 6, 7, 8, and 9 clusters were examined qualitatively. The first desired property of the resulting player profiles is that the analysis of profile combinations has a large enough sample size in terms of players, lineups and possessions. It was subjectively determined that having 9 clusters did not have a large enough sample, while 8 clusters yielded good results. Therefore, to maximize the granularity of each of the player profiles and still maintain a large enough sample size for the following analysis, 8 clusters were used.

4.1.2 Cluster Analysis

Here are the results from k-means with 8 clusters.

Table 4.1: Player Profiles

Cluster	Description	Example Players
1	all-round sorers, facilitators, and playmakers	LeBron James, Kevin Durant
2	shoots mostly 3 pointers with low volume attempts	Pat Connaughton, Kevin Huerter
3	scoring big men with high offensive skill	Giannis Antetokounmpo, Joel Embid
4	facilitators that don't score much	Lonzo Ball, Draymond Green
5	big men with low offensive skill, don't create shots for themselves	Rudy Gobert, Clint Capela
6	secondary scorers, can shoot from midrange and 3	CJ McCollum, Malcolm Brogdon
7	high volume 3 point shooters	Danny Green, Kyle Korver
8	non-scorer, most shots are near the basket	Larry Nance Jr., Kyle Anderson

Table 4.2: Cluster Averages

	cluster							
feature	1	2	3	4	5	6	7	8
points	30.38	16.42	26.08	16.66	18.32	23.04	18.45	16.26
2PA	15.81	6.65	16.6	9.16	12.97	12.59	6.31	9.94
3PA	8.5	7.02	2.26	5.98	0.24	6.82	10	3.55
pct of FGA 0-3 ft	0.25	0.28	0.48	0.24	0.58	0.25	0.13	0.44
pct of FGA 3-10 ft	0.16	0.11	0.27	0.15	0.27	0.16	0.08	0.2
pct of FGA 10-16 ft	0.13	0.05	0.09	0.1	0.08	0.12	0.07	0.05
pct of FGA 16-3P	0.11	0.04	0.06	0.11	0.05	0.12	0.11	0.04
pct of FGA 3P	0.35	0.52	0.11	0.4	0.02	0.35	0.62	0.26
FG pct 2P	0.49	0.54	0.58	0.46	0.57	0.49	0.45	0.51
FG pct 3P	0.35	0.34	0.27	0.32	0	0.35	0.36	0.3
pct FG assisted 2P	0.26	0.67	0.65	0.32	0.69	0.51	0.65	0.62
pct FG assisted 3P	0.61	0.95	0.96	0.75	0	0.87	0.93	0.97
FT%	0.8	0.73	0.72	0.74	0.65	0.78	0.8	0.67
FT rate	0.29	0.2	0.36	0.21	0.43	0.23	0.14	0.32
ORB%	2.55	4.32	10.22	2.28	10.66	3.2	1.9	7.18
AST%	28.82	8.06	12.23	22.24	9.88	12.26	8.47	9.69
TOV%	12.66	11.41	12.4	16.28	14.09	10.3	8.94	13.64
dribbles per pos	2.22	0.29	0.35	1.96	0.27	0.74	0.42	0.37

One note is that dribbles per possession takes into account all possessions offense and defense, which is why it may be lower than expected. Looking at the cluster averages is useful to understand the metrics for each player profile, but the averages as a percentile of all players provides insight to how the cluster average compares to all players. For instance, cluster 1's average points are in the 93rd percentile meaning that the average is greater than or equal to 93% of all players.

Table 4.3: Cluster Averages as a percentile of all players

	cluster							
feature	1	2	3	4	5	6	7	8
points	0.93	0.3	0.84	0.31	0.44	0.72	0.44	0.29
2PA	0.86	0.2	0.89	0.42	0.72	0.7	0.19	0.49
3PA	0.74	0.6	0.17	0.45	0.07	0.58	0.88	0.23
pct of FGA 0-3 ft	0.42	0.53	0.86	0.39	0.92	0.42	0.1	0.83
pct of FGA 3-10 ft	0.58	0.3	0.88	0.53	0.88	0.58	0.15	0.73
pct of FGA 10-16 ft	0.83	0.26	0.55	0.66	0.52	0.79	0.43	0.3
pct of FGA 16-3P	0.73	0.35	0.43	0.71	0.38	0.76	0.71	0.33
pct of FGA 3P	0.44	0.76	0.16	0.54	0.07	0.45	0.91	0.28
FG pct 2P	0.45	0.68	0.84	0.24	0.84	0.4	0.22	0.53
FG pct 3P	0.58	0.56	0.22	0.41	0.09	0.61	0.67	0.3
pct FG assisted 2P	0.1	0.68	0.62	0.18	0.7	0.42	0.64	0.57
pct FG assisted 3P	0.17	0.64	0.66	0.27	0.09	0.41	0.55	0.69
FT%	0.65	0.38	0.36	0.44	0.19	0.58	0.67	0.23
FT rate	0.71	0.39	0.82	0.42	0.91	0.51	0.16	0.75
ORB%	0.39	0.63	0.89	0.31	0.91	0.49	0.24	0.76
AST%	0.94	0.3	0.57	0.84	0.42	0.57	0.33	0.42
TOV%	0.58	0.46	0.57	0.87	0.72	0.34	0.17	0.67
dribbles per pos	0.9	0.35	0.42	0.85	0.33	0.65	0.5	0.45

Cluster 1: all-round playmakers

Cluster 1 represents the typical star guard for a team. These players score a lot of points and take a lot of shots from 2 and 3. In addition, these players often handle the ball, get assists, but also have high turnovers. Furthermore, they are able to create shots for themselves, which is seen with the lowest percent of their two pointers assisted and second lowest for their three pointers. More examples with this profile are Stephen Curry, Devin Booker, Kyrie Irving, and Paul George.

Cluster 2: low volume three point specialists

Players in cluster two do not score a lot of points and most of their shots are from beyond the arc. In addition, they are not great at creating shots for themselves, as most of their shots are assisted. They do not handle the ball often or get many assists, and since they don't drive to the basket often, they also don't shoot many free throws. These players can be thought of as role players expected to make three pointers when open. More examples of players with this profile are Andre Iguodala, Josh Hart, and Mikal Bridges.

Cluster 3: scoring big men

Players in cluster 3 are big men with high offensive skill. These players score a lot of points and shoot a high volume close to the basket, but few three pointers. They are around average for the percent of made field goals assisted, but they still have a relatively high ability to create shots for themselves. Since they are big men and don't handle the ball much, it makes sense that a lot of their makes should be assisted, so being around average for field goals assisted demonstrates a high ability to create shots for themselves. Lastly, these players grab the most offensive rebounds. More examples of players with this profile are Anthony Davis, Nikola Jokic, and Bam Adebayo.

Cluster 4: facilitators

Players in cluster 4 are typically point guards who get a lot of assists, but do not score much. They are about average in field goal attempts, which reflects their lack of scoring since they handle the ball a lot. Their high turnover rate is likely due to the fact that they also have very high dribbles per possession. More examples of these players are Patrick Beverly, Alex Caruso, Kyle Lowry, and Rajon Rondo.

Cluster 5: non-scoring big men

Players in cluster 5 do not score much and shoot almost exclusively 2 pointers. These are the typical big men without much offensive skill, but they can make dunks and shots close to the basket when assisted by teammates. They have the highest percentage of field goals assisted, which reflects their lack of ability to create shots for themselves. In addition, since they shoot mosly close to the basket, they are very efficient and shoot a high field goal percentage. More examples of these players are Steven Adams, DeAndre Jordan, and Tyson Chandler.

Cluster 6: outside shooting secondary scorers

This profile represents good, but not elite scorers. They are the third highest scoring cluster and they can score from anywhere on the court, especially the mid range. These players are typically the second or third scoring option on teams. More examples of these players are Andrew Wiggins, Al Horford, and Kyle Kuzma.

Cluster 7: high volume three point specialists

Players in cluster 7 shoot the most three pointers, and relatively few two pointers. They do not handle the ball much, or create shots for themselves often. Their main role is to make three pointers and not much else, which is seen in their low assist rate, dribble rate, free throw rate, and two point attempt rate. More examples of these players are Jae Crowder, Eric Gordon, Patty Mills, and JJ Redick.

Cluster 8: inside shooting role players

Players in cluster 8 are commonly referred to as role players. They don't excel at any one thing, and most of their shots are within 10 feet. Most of their shots are assisted, indicating that they do not create many shots for themselves. More examples of these players are Daniel Theis, Michael Kidd-Gilchrist, and Bruce Brown.

4.2 Evaluating Lineups

Evaluating lineups does not necessarily provide any special or unique insights, but it is a useful step before evaluating profile combinations. In addition, this will highlight the difference between evaluating raw performance and evaluating performance relative to what is expected. Here are the top and bottom five lineups measured in raw performance as points per possession.

Table 4.4: Lineups Ranked by Points Per Possession

lineup	pts per possession	number of possessions
draymond green, stephen curry, kevon looney, klay thompson, kevin durant	1.2530	739
cj mccollum, evan turner, alfarouq aminu, jusuf nurkic, damian lillard	1.2097	534
thomas bryant, trevor ariza, bradley beal, jeff green, tomas satoransky	1.2011	701
draymond green, stephen curry, andre iguodala, klay thompson, kevin durant	1.1989	734
cj mccollum, alfarouq aminu, jake layman, damian lillard, jusuf nurkic	1.1899	753
⋮	⋮	⋮
donovan mitchell, derrick favors, ricky rubio, joe ingles, rudy gobert	1.0714	1050
cj mccollum, alfarouq aminu, maurice harkless, enes kanter, damian lillard	1.0688	581
derrick white, demar derozan, jakob poeltl, lamarcus aldridge, bryn forbes	1.0507	513
marc gasol, jaren jackson, mike conley, kyle anderson, garrett temple	1.0431	905
dennis smith, wesley matthews, luka doncic, deandre jordan, harrison barnes	0.9907	860

*minimum 1000 possessions

It makes sense that the Warrior's lineup of Kevin Durant, Stephen Curry, Klay Thompson, Draymond Green, and Kevon Looney was the top performing lineup because they had three of the best scorers of all time. These results are unsurprising, but more insights can be drawn from examining the top performing lineups relative to expected performance.

Table 4.5: Lineups Ranked by Points Above Predicted

lineup	pts above predicted	pts per pos	predicted pts per pos
ben simmons, jimmy butler, tobias harris, joel embiid, jj redick	0.0893	1.1888	1.0996
thomas bryant, trevor ariza, bradley beal, jeff green, tomas satoransky	0.0773	1.2011	1.1239
donovan mitchell, jae crowder, ricky rubio, joe ingles, rudy gobert	0.0657	1.1756	1.1099
ben simmons, jimmy butler, joel embiid, jj redick, wilson chandler	0.0576	1.1672	1.1096
blake griffin, bruce brown, andre drummond, reggie bullock, reggie jackson	0.0573	1.1683	1.111
⋮	⋮	⋮	⋮
dj augustin, evan fournier, terrence ross, aaron gordon, nikola vucevic	-0.0407	1.0994	1.1401
draymond green, demarcus cousins, stephen curry, klay thompson, kevin durant	-0.0478	1.1623	1.2101
dennis smith, wesley matthews, luka doncic, deandre jordan, harrison barnes	-0.0625	0.9907	1.0532
cj mccollum, alfarouq aminu, maurice harkless, enes kanter, damian lillard	-0.0639	1.0688	1.1328
derrick white, demar derozan, jakob poeltl, lamarcus aldridge, bryn forbes	-0.0976	1.0507	1.1482

*minimum 500 possessions

As seen in this table, the best performing lineups in terms of points per possession are not necessarily the best performing lineups relative to what is expected. The aforementioned Warriors lineup with Kevin Durant and Steph Curry scored the most points per possession, but since their lineup was full of great players, the expected points per possession was also very high. By comparing lineup performance to what is expected, it adjusts for how skilled a lineup is. This can be beneficial because it provides insight to what players play well together. Coaches can utilize this information of what players complement each other for deciding what lineups to play more often, and what players to build a team around. For instance, if a team has two great point guards and this analysis determines that they perform below expected when on the court together, this may indicate that their play styles don't mesh well. A coach may decide to asjust when they play in the game so they don't play at the same time as often.

Another useful expansion of this analysis is evaluating subsets of lineups. When examining a lineup of five players, the root cause of the performance may not be obvious. To account for this, evaluating subsets of lineups can provide more granular analysis of what players actually complement each other. The evaluation of subsets is done by aggregating all possessions where the given subset was on the court. The total and predicted points scored across all possessions are used to calculate the points above predicted per possession. Here is an example of the best and worst performing subsets of 3 players.

Table 4.6: Lineup Subsets Ranked by Points Above Predicted

lineup	points above predicted	points per pos	predicted points per pos	number of pos
jj redick, wilson chandler, ben simmons	0.073	1.1735	1.1005	1043
jj redick, joel embiid, jimmy butler	0.0705	1.1708	1.1003	1833
bam adebayo, kelly olynyk, josh richardson	0.066	1.1521	1.086	1164
jae crowder, donovan mitchell, joe ingles	0.066	1.1707	1.1048	1716
davis bertans, jakob poeltl, patty mills	0.0658	1.2341	1.1684	1008
⋮	⋮	⋮	⋮	⋮
jaylen brown, kyrie irving, gordon hayward	-0.0696	1.0656	1.1352	1112
justin holiday, wendell carter, zach lavine	-0.0724	0.9759	1.0483	1286
jayson tatum, jaylen brown, al horford	-0.0752	1.0701	1.1453	1312
jaylen brown, kyrie irving, jayson tatum	-0.0752	1.0743	1.1495	1346
rudy gay, patty mills, marco belinelli	-0.0782	1.08	1.1583	1187

*minimum 1000 possessions

One issue is that subsets of players can be highly correlated with lineups that play a large number of possessions together. For instance, two of the top 5 lineup subsets are from the Philadelphia 76ers, so these lineups most likely have some overlap with JJ Redick. Likewise, 3 of the bottom 5 lineups are from the Boston Celtics. It is somewhat surprising that lineups with stars such as Jayson Tatum, Kyrie Irving, and Jaylen Brown performed this poorly relative to expected. One explanation is that this was only Jayson Tatum and Jaylen Brown's second year in the league, so they may not have figured out the proper chemistry yet. In addition, these lineups did not perform poorly in terms of raw performance. But since each of these players have high individual skill, the expectations for these lineups were relatively high. A conclusion for this is that the lineups with these Celtic's stars did not maximize the individual talent.

4.3 Evaluating Player Profile Combinations

The previous lineup evaluation provides insights to how individual players complement each other. The same concept can be applied to evaluating how player profiles complement each other. Instead of analyzing individual lineups, lineups are grouped by the player profile combination that they represent. For example, the lineup of Kevin Durant, Stephen Curry, Klay Thompson, Draymond Green, and Kevon Looney would be represented as the profile combination of (1, 1, 6, 4, 3). It makes sense that Kevin Durant and Stephen Curry have profile 1, which is all round playmaker and scorer; and Draymond Green has label 4, which is facilitating non-scorer. It is somewhat surprising that Klay Thompson is labeled as a secondary scorer that can shoot rather than a three point specialist. This can be attributed to the fact that only 43% of Klay's shots were from three, and he scored a lot more points than the average of all three point specialists with profile 7 or 2.

This player profile evaluation will explore the performance of profile combinations with five players as well as what subsets of profile combinations complement each other.

4.3.1 Evaluating Size 5 Profile Combinations

Here are the best and worst performing lineups in terms of points scored above predicted per possession. To ensure a large enough sample size, the threshold for minimum number of possessions was set to 500. In addition, a threshold of 3 was used for the minimum number of unique lineups that the profile combination consists of. This threshold was used because if a profile combination only consists of one or two lineups, then the performance will just be a reflection of how well those individual players complement each other. However, the goal of this analysis is to gain more broad insight to how the play styles impact performance.

Table 4.7: Profile Combination Performance

Profile Combination	points above predicted	number of pos	number of unique lineups
(1, 2, 2, 3, 8)	0.1114	621	4
(1, 3, 4, 7, 7)	0.1001	913	8
(1, 2, 5, 7, 7)	0.0847	689	6
(1, 1, 3, 7, 7)	0.082	756	5
(1, 1, 1, 2, 5)	0.0805	1386	5
⋮	⋮	⋮	⋮
(1, 6, 6, 6, 6)	-0.071	959	4
(4, 4, 6, 6, 8)	-0.0903	1081	6
(1, 1, 4, 5, 6)	-0.108	783	5
(4, 4, 5, 6, 7)	-0.1455	968	4
(2, 3, 3, 4, 4)	-0.1671	552	4

Examining these results offers various avenues for insights, as each combination can be individually analyzed. However, delving into the fine details of each combination might lead to overfitting the data and excessive granularity. To avoid this, focusing on common themes among the best and worst profile combinations will provide a general trend without overfitting.

An observation regarding the top-performing profile combinations reveals that four out of the five include multiple three point specialists (labels 2 and 7). This aligns with the current focus of the NBA, where teams increasingly prioritize three point shooting and floor spacing. NBA analytics indicate that shooting more three pointers and fewer midrange shots generally results in greater efficiency [8]. The presence of three point specialists allows for efficient shooting while stretching the defense out and opening space in the middle of the court for other players to drive to the basket. To extend this concept, each of the top five lineups have at least one all round scorer and playmaker (profile 1). The logical connection is that pairing three point specialists with all round playmakers spaces the floor well, which provides the playmakers more room to drive, score, and facilitate from the middle of the court. If the defense crowds the paint to stop the playmaker from driving, there are open shooters; and if the defense is spaced out to the shooters, the playmaker has more room to score. In addition, three point specialists have a high percent of their shots assisted, meaning they are not great at creating shots for themselves. Their shooting efficiency would not be maximized unless their is another player that can create shots for them.

An opposite observation can be made for the worst performing combinations. Three out of the bottom five profile combinations lack a three point specialist, while the remaining two combinations include only one. This aligns with the insights from the best performing combinations. Another observation is that three of the worst combinations include two non-scoring facilitators (label 4). The intuitive reasoning for this could be that teams need players to be able to score. If there are two non-scoring threats on the court, the defense can focus more heavily on stopping the other three players, limiting offensive efficiency.

4.3.2 Evaluating Individual Profiles

The theme found in the top and bottom performing profile combinations was that having more than one of a specific player profile in the lineup may improve or worsen teams' efficiency. This can be examined further for each of the player profiles. The question to be answered is for a given profile label, how do lineups perform when there is one of that profile versus when there is two. In the analysis of all profile combinations, the observations indicated that having more than one three point specialist may be better than having just one. In addition, it indicated that having more than one non-scoring facilitator was detrimental. Here are graphs displaying all instances of exactly one of the given profile label and then all instances of exactly two of the given profile label.

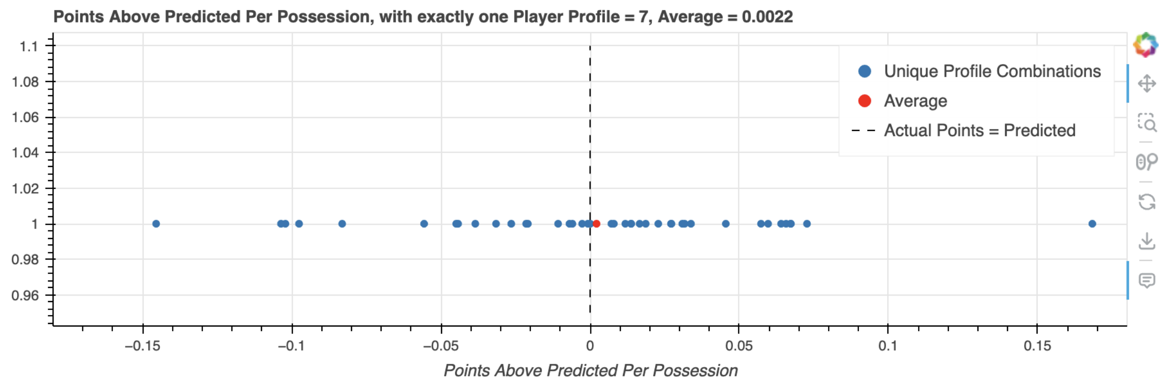

Figure 4-2: Exactly One High Volume Three Point Specialist

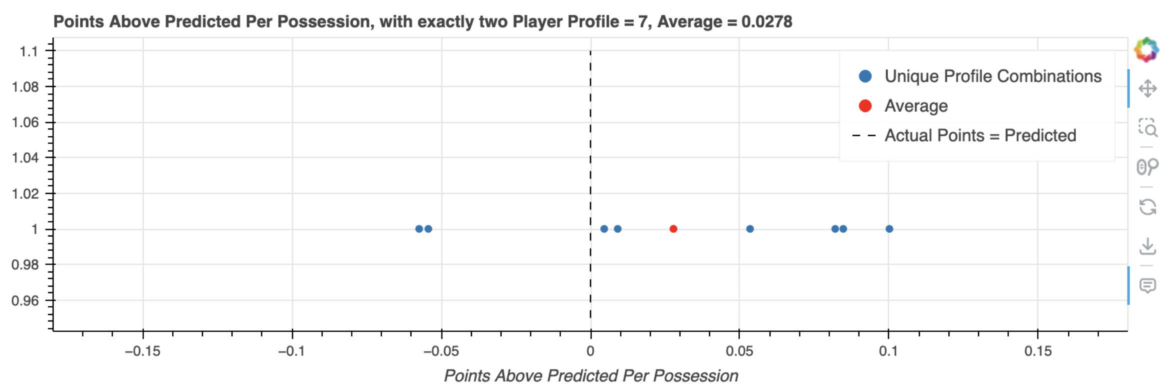

Figure 4-3: Exactly Two High Volume Three Point Specialist

With exactly one high volume three point specialist, the profile combinations perform just as expected with a mean of 0.0022 points above predicted per possession. When there are two high volume three point specialist in the lineup, performance improves drastically to an average of 0.0278 points above predicted per possession. It is important to note that when filtering for exactly two of a given profile type, there is a much smaller sample size of profile combinations that match the filter. There are six profile combinations that perform above expected, while only two that perform below expected.

Here are the same graphs when looking at non-scoring facilitators (label 4).

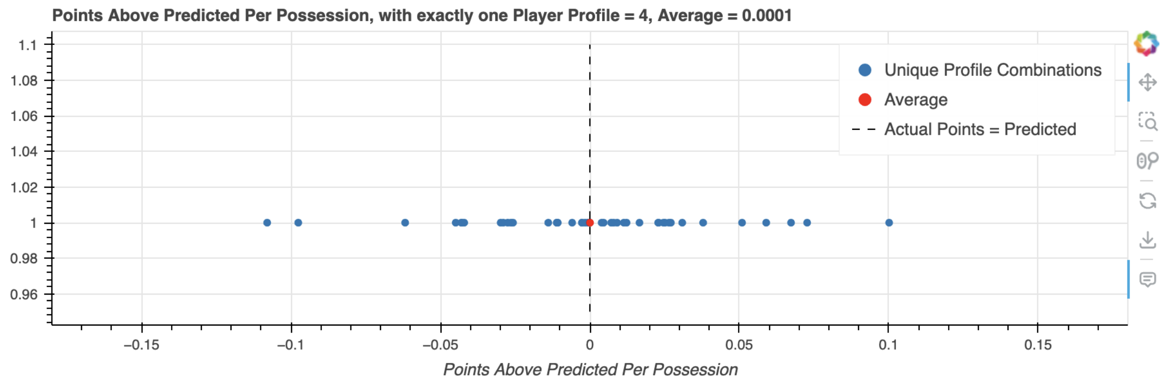

Figure 4-4: Exactly One Non-scoring Facilitator

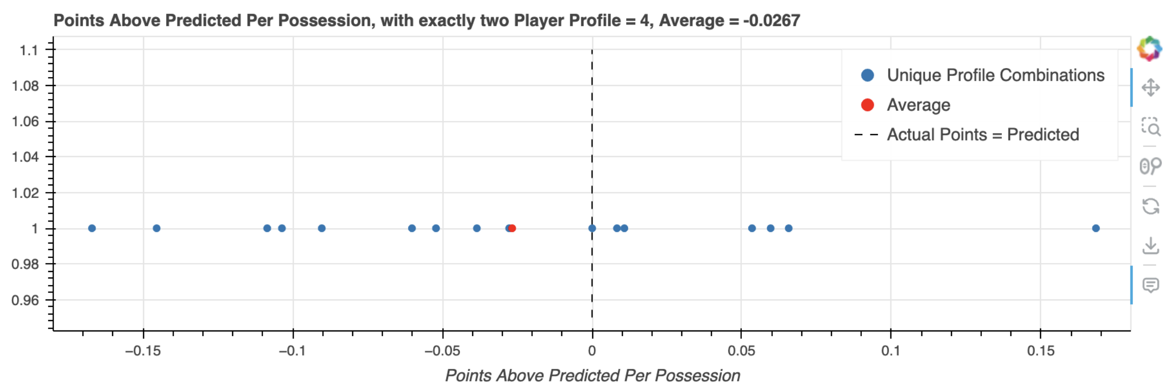

Figure 4-5: Exactly Two Non-scoring Facilitators

This reflects the observations of best and worst profile combinations. Lineups with more than one non-scoring facilitator perform significantly worse than lineups with exactly one non-scoring facilitator.

The two examples above were shown as figures because they are the most insightful and they reflect the observations of the best and worst performing profile combinations. All of the same profile analysis will be summarized in the following table.

Table 4.8: Performance of Individual Profiles

profile	Change from exactly one to two	Average: Exactly one	Average: Exactly two
high volume three point specialist (label 7)	0.02560	0.0022	0.0278
low volume three point specialist (label 2)	0.0196	-0.0016	0.018
all-round playmaker (label 1)	-0.0024	0.0036	0.0012
outside shooting secondary scorer (label 6)	-0.0066	-0.0111	-0.0177
scoring big men (label 3)	-0.0101	0.0029	-0.0072
non-scoring facilitators (label 4)	-0.0268	0.0001	-0.0267
non-scoring big men (label 5)	N/A	-0.0016	N/A
inside shooting role player (label 8)	N/A	-0.0836	N/A

The two profiles of non-scoring big men and inside shooting role player have N/A values because there was not a large enough sample size for the profile combinations that contained exactly two. The threshold for a profile combination to be considered was to have a minimum of 300 offensive possessions and at least three unique line-ups that compose the combination. In addition, at least five data points of profile combinations were needed to calculate an average for a given profile. As mentioned earlier, the results that show having more than one three point specialist improves performance is logical. Having more than one all-round playmaker did not significantly impact performance. When thinking about how two all-round playmakers complement each other, there are two opposing forces. Firstly, there is a beneficial impact where if the defense focuses on stopping one playmaker, the other will have better opportunities to score. Secondly, there is a detrimental force where the playmakers rely on having the ball in their hands, which is seen in their high dribbles per possession and assist rate. However, having two players that rely on having the ball, they may be less effective at sharing this role. Moving on to outside shooting secondary scorers, there is no obvious argument for why having one versus two on the court would make a difference. Lastly, it makes sense that having two scoring big men would not maximize performance. These types of players generally score close to the basket, so having two on the court at the same time may reduce the spacing and crowd the paint, reducing efficiency. In addition, scoring big men are able to create shots for themselves, which is seen in having only the fourth highest percent of two point shots assisted out of all eight profiles. Therefore, if there are two big men that rely on creating shots for themselves, they may be less efficient when having to share the basketball.

4.3.3 Evaluating Subsets of Profile Combinations

The previous analysis examined the impact of having multiple of each individual player profile. There is likely more insights to be gained from looking at combinations across different player profiles. For example, having two three point specialists seems to enhance performance, but what about having a three point specialist paired with a scoring big man? Below is a table of the top performing pairs of player profiles.

Table 4.9: Performance of Profile Pairs

Profile Combination Subset	points above predicted per pos	number of pos	number of unique lineups
(2, 5)	0.0296	20664	140
(3, 5)	0.0264	5187	28
(7, 7)	0.0162	9766	87
(2, 7)	0.015	23951	215
(3, 7)	0.0145	35623	266
⋮	⋮	⋮	⋮
(8, 8)	-0.0243	1413	20
(7, 8)	-0.0276	8226	67
(4, 4)	-0.0325	12600	90
(4, 8)	-0.0349	11619	98
(5, 8)	-0.0422	2508	28

The top performing profile pairs share the aforemention theme that having multiple three point specialists helps maximize performance, as two of these five pairs have two three point specialists. In addition, a logical extension of this is that a three point specialist paired with a big man is beneficial because the three point specialist likely plays on the perimeter and big men play near the basket. If the defense were to focus on defending perimeter shooting, this could create opportunities near the basket, and vice versa. The combination that may be surprising is (3,5), which is a scoring big man and non scoring big man. As seen in the individual profile analysis, having two scoring big men is detrimental to performance. This could be due to more crowding near the basket. In addition, scoring big men often create shots for themselves and rely on having the ball in their hands, which may reduce effectiveness when they have to share the ball. However, these results suggest that having one scoring big man and a non-scoring big man is different. An explanation for this is that a scoring big man often relies on having the ball in their hands, while non-scoring big men typically rely on assists from teammates or offensive rebounds. Therefore, a scoring and non-scoring big men could maximize both the ability for the scoring big man to create shots and the non-scoring big man to score when teammates facilitate shots.

The first concept of high performing profile combinations is having good spacing. As mentioned above, having an inside threat paired with an outside threat seems to maximize each player's efficiency. In addition, from the analysis of size five profile combinations, having a shot creator paired with shot makers seems to maximize performance. The pairs of profiles that perform the worst generally do not have either of these desired properties. Four of these pairs consist of facilitators, non-scoring bigs, or inside shooting role players (profiles 4, 5, and 8). Facilitators are good at creating shots for others, as seen in their high assist percentage. However, they themselves are not good scorers, so for their passing abilities to be capitalized on, they rely on teammates to score. One conclusion is that these profiles may be able to contribute on offense well when paired with scoring threats, but in the absence of a scoring threat, their best skills are not fully utilized.

Chapter 5

Conclusion

The RAPM model quantifies each player's contribution to their team's offensive performance. For a given lineup, the sum of these values corresponds to the team's number of points scored. On average, the team's performance will equal the sum of the parts. This analysis revealed that certain groupings of player profiles can elevate the team's performance beyond the sum of its parts. This validates that play styles do matter for maximizing the individual talents within a lineup.

The understanding that play style influences team performance has long been recognized, as traditional basketball lineups consist of five different positions, each with a different style. This analysis affirmed the belief that the style of individual players has an impact on team success. Additionally, evaluating player profile combinations brought insight to what play styles complement or hinder each other.

5.1 Insights

The evaluation of five-player profile combinations revealed that having multiple three point specialists had a positive impact on performance, while multiple non-scoring facilitators had a negative impact. This theme was reaffirmed when analyzing instances of one versus two players from each profile. Additionally, it was found that having two scoring big men decreased performance compared to having only one. The evaluation of all pairs of player profiles highlighted two key concepts that consistently influenced performance. The first concept is that court spacing improves performance. This is materialized by having three point shooters on the perimeter, who pair well with big men that are efficient in the paint. Secondly, shot creators pair well with efficient shot makers that don't rely on handling the ball. These findings align with the idea that a diverse lineup excelling in various aspects of scoring optimizes individual contributions. Consequently, maximizing offensive performance should consider complementary styles rather than simply assembling the most talented lineup.

This insight has practical implications for coaches and general managers building a team. It suggests that player profiles should be considered when adding new players to the roster. For example, prioritizing three-point threats to complement existing playmakers and skilled big men may be more beneficial than adding another playmaker. While this analysis focused exclusively on offensive performance, it is important to acknowledge that players contribute on both ends of the court. Therefore, this analysis serves as a valuable consideration, but does not serve as a comprehensive assessment of overall performance.

5.2 Future Work

An immediate direction for future work is to expand the analysis to include data from additional years. This would provide a larger and more robust dataset, enhancing the confidence and reliability of the results. In addition, there are a lot of parameters that could be adjusted and explored. For example, the number of clusters or the features selected to create the clusters could be modified. The features used in this thesis were relatively basic and covered multiple aspects of scoring. However, there are still many aspects that could be added such as involvement in pick and rolls, frequency of setting off ball screens, frequency of playing in different locations on the court, or a player's average speed during possessions. Moreover, the methods of evaluation could be altered slightly, as mentioned in the Methods section.

Another valuable extension of this research is to incorporate defense. Coaches and teams would benefit from understanding how play styles interact and influence performance on both ends of the court. This could employ similar concepts utilized in this thesis. For building the player profiles, defensive metrics such as blocks, defensive rebounds, or steals could be used. Furthermore, total plus/minus rather than just points per possession could be used. Another interesting question arises regarding the compatibility between a team's play style and the individual talents of its players. For example, if a team possesses strong three point shooters, it would be logical for them to play a style that shoots a lot of threes. Each team has a unique play style and it would be useful for coaches to know what style is best for their team.

Appendix A

Player Clusters

Table A.1: Cluster 1: all-round scoring playmakers

Kadeem Allen	J.J. Barea	Bradley Beal	Eric Bledsoe	Devin Booker
Trey Burke	Mike Conley	Stephen Curry	DeMar DeRozan	Spencer Dinwiddie
Luka Dončić	Goran Dragić	Kevin Durant	De'Aaron Fox	Paul George
Blake Griffin	James Harden	Jrue Holiday	Kyrie Irving	Reggie Jackson
LeBron James	Zach LaVine	Kawhi Leonard	Caris LeVert	Damian Lillard
Khris Middleton	Donovan Mitchell	Emmanuel Mudiay	Jamal Murray	Shabazz Napier
Victor Oladipo	Tony Parker	Chris Paul	Derrick Rose	D'Angelo Russell
Dennis Schröder	Collin Sexton	Isaiah Thomas	Dwyane Wade	Kemba Walker
John Wall	Russell Westbrook	Lou Williams	Trae Young	

Table A.2: Cluster 2: low volume three point specialists

Quincy Acy	Al-Farouq Aminu	Justin Anderson	Ryan Anderson	OG Anunoby
Trevor Ariza	Keita Bates-Diop	Nicolas Batum	Dragan Bender	Nemanja Bjelica
Jaron Blossomgame	Jonah Bolden	Chris Boucher	Mikal Bridges	Miles Bridges
Ryan Broekhoff	Sterling Brown	Deonte Burton	Bruno Caboclo	DeMarre Carroll
Wilson Chandler	Bonzie Colson	Pat Connaughton	Robert Covington	Torrey Craig
Dante Cunningham	Dewayne Dedmon	Luol Deng	Donte DiVincenzo	Tyler Dorsey
Jared Dudley	James Ennis III	Terrance Ferguson	Dorian Finney-Smith	Jerami Grant
JaMychal Green	Jeff Green	Devin Harris	Joe Harris	Josh Hart
John Henson	Juancho Hernangómez	Solomon Hill	Danuel House Jr.	Kevin Huerter
Andre Iguodala	Ersan İlyasova	Jonathan Isaac	Justin Jackson	Jonas Jerebko
Stanley Johnson	Frank Kaminsky	Maxi Kleber	Rodions Kurucs	Jake Layman
Meyers Leonard	Brook Lopez	Tyler Lydon	Thon Maker	Patrick McCaw
Doug McDermott	Rodney McGruder	Alfonzo McKinnie	Mike Muscala	Abdel Nader
Georges Niang	Royce O'Neale	Semi Ojeleye	Josh Okogie	Kelly Olynyk
Patrick Patterson	Norman Powell	Austin Rivers	Duncan Robinson	Brandon Sampson
Dario Šarić	Thabo Sefolosha	Zhaire Smith	Omari Spellman	Garrett Temple
Lance Thomas	P.J. Tucker	Moritz Wagner	C.J. Williams	Kenrich Williams

Table A.3: Cluster 3: scoring big men

Bam Adebayo	LaMarcus Aldridge	Jarrett Allen	Giannis Antetokounmpo	Marvin Bagley III
Michael Beasley	Thomas Bryant	Wendell Carter Jr.	Willie Cauley-Stein	John Collins
DeMarcus Cousins	Anthony Davis	Cheick Diallo	Andre Drummond	Joel Embiid
Kenneth Faried	Derrick Favors	Enes Freedom	Taj Gibson	Montrezl Harrell
Willy Hernangómez	Rondae Hollis-Jefferson	Serge Ibaka	Jaren Jackson Jr.	Nikola Jokić
T.J. Leaf	Jon Leuer	Kevon Looney	Robin Lopez	Boban Marjanović
JaVale McGee	Paul Millsap	Greg Monroe	Jusuf Nurkić	Jahlil Okafor
Mason Plumlee	Dwight Powell	Ivan Rabb	Julius Randle	Domantas Sabonis
Pascal Siakam	Karl-Anthony Towns	Jonas Valančiūnas	Nikola Vučević	Hassan Whiteside

Table A.4: Cluster 4: non-scoring facilitators

Jaylen Adams	Ryan Arcidiacono	D.J. Augustin	Ron Baker	Wade Baldwin
Lonzo Ball	Jerryd Bayless	Patrick Beverley	Isaiah Briscoe	Lorenzo Brown
Jalen Brunson	José Calderón	Isaiah Canaan	Jevon Carter	Michael Carter-Williams
Alex Caruso	Darren Collison	Jamal Crawford	Matthew Dellavedova	Kris Dunn
Jacob Evans	Tyreke Evans	Dante Exum	Raymond Felton	Yogi Ferrell
Tim Frazier	Markelle Fultz	Shai Gilgeous-Alexander	Devonte' Graham	Jerian Grant
Draymond Green	Daniel Hamilton	Andrew Harrison	Aaron Holiday	Joe Ingles
James Johnson	Tyus Jones	Cory Joseph	Brandon Knight	Jeremy Lin
Kyle Lowry	Shelvin Mack	Daryl Macon	Frank Mason III	T.J. McConnell
De'Anthony Melton	Naz Mitrou-Long	Raul Neto	Frank Ntilikina	Elie Okobo
Cameron Payne	Elfrid Payton	Theo Pinson	Chasson Randle	Rajon Rondo
Terry Rozier	Ricky Rubio	Tomáš Satoranský	Jonathon Simmons	Anfernee Simons
Marcus Smart	Dennis Smith Jr.	Ish Smith	Lance Stephenson	Jeff Teague
Miloš Teodosić	Jared Terrell	Evan Turner	Fred VanVleet	Brad Wanamaker

Table A.5: Cluster 5: non-scoring big men

Steven Adams	Deandre Ayton	Jordan Bell	Khem Birch	Bismack Biyombo
Andrew Bogut	Isaac Bonga	Clint Capela	Tyson Chandler	Deyonta Davis
Ed Davis	Drew Eubanks	Cristiano Felício	Harry Giles	Rudy Gobert
Marcin Gortat	Nenê	Richaun Holmes	Dwight Howard	Damian Jones
DeAndre Jordan	Kosta Koufos	Shaun Livingston	Chimezie Metu	Johnathan Motley
Joakim Noah	Nerlens Noel	Zaza Pachulia	Miles Plumlee	Jakob Poeltl
Devin Robinson	Mitchell Robinson	Ben Simmons	Ray Spalding	Tristan Thompson
Ekpe Udoh	Johnathan Williams	Robert Williams	Tyler Zeller	Ante Žižić

Table A.6: Cluster 6: outside shooting secondary scorers

Grayson Allen	Dwayne Bacon	Harrison Barnes	Will Barton	Kent Bazemore
Antonio Blakeney	Bogdan Bogdanović	Bojan Bogdanović	Malcolm Brogdon	Dillon Brooks
MarShon Brooks	Jaylen Brown	Troy Brown Jr.	Alec Burks	Jimmy Butler
Jordan Clarkson	Quinn Cook	Gorgui Dieng	Henry Ellenson	Evan Fournier
Danilo Gallinari	Marc Gasol	Rudy Gay	Aaron Gordon	Tim Hardaway Jr.
Gary Harris	Tobias Harris	Gordon Hayward	Mario Hezonja	Buddy Hield
George Hill	Rodney Hood	Al Horford	Brandon Ingram	Wes Iwundu
Frank Jackson	Josh Jackson	Tyler Johnson	Kevin Knox	Kyle Kuzma
Jeremy Lamb	Courtney Lee	Walt Lemon Jr.	Kevin Love	Trey Lyles
Lauri Markkanen	CJ McCollum	Jordan McRae	E'Twaun Moore	Marcus Morris
Markieff Morris	Monte Morris	Kyle O'Quinn	Cedi Osman	Kelly Oubre Jr.
Jabari Parker	Otto Porter Jr.	Bobby Portis	Taurean Prince	Josh Richardson
Glenn Robinson III	JaKarr Sampson	Wayne Selden	Jayson Tatum	Klay Thompson
Gary Trent Jr.	Allonzo Trier	Myles Turner	Lonnie Walker IV	T.J. Warren

Table A.7: Cluster 7: high volume three point specialists

Álex Abrines	Deng Adel	Carmelo Anthony	Malik Beasley	Marco Belinelli
Dairis Bertāns	Dāvis Bertāns	Avery Bradley	Reggie Bullock	Kentavious Caldwell-Pope
Vince Carter	Gary Clark	Ian Clark	Allen Crabbe	Jae Crowder
Seth Curry	Troy Daniels	Damyean Dotson	Wayne Ellington	Bryn Forbes
Channing Frye	Langston Galloway	Eric Gordon	Treveon Graham	Danny Green
Gerald Green	Justin Holiday	John Jenkins	Wesley Johnson	Luke Kennard
Furkan Korkmaz	Luke Kornet	Kyle Korver	Damion Lee	Timothé Luwawu-Cabarrot
Jarell Martin	Wesley Matthews	Ben McLemore	Jodie Meeks	C.J. Miles
Darius Miller	Patty Mills	Shake Milton	Nikola Mirotić	Malik Monk
Svi Mykhailiuk	Dirk Nowitzki	James Nunnally	Chandler Parsons	J.J. Redick
Cameron Reynolds	Malachi Richardson	Jerome Robinson	Terrence Ross	Mike Scott
Landry Shamet	Iman Shumpert	J.R. Smith	Tony Snell	Nik Stauskas

Table A.8: Cluster 8: inside shooting role players

Rawle Alkins	Kyle Anderson	Mo Bamba	Aron Baynes	DeAndre' Bembry
Corey Brewer	Bruce Brown	Omri Casspi	Marquese Chriss	Zach Collins
Sam Dekker	Hamidou Diallo	Pau Gasol	Maurice Harkless	Shaquille Harrison
Isaiah Hartenstein	Chandler Hutchison	Amir Johnson	Derrick Jones Jr.	Jalen Jones
Jemerrio Jones	Michael Kidd-Gilchrist	Skal Labissière	Alex Len	Ian Mahinmi
Salah Mejri	Larry Nance Jr.	David Nwaba	Quincy Pondexter	Alex Poythress
Edmond Sumner	Caleb Swanigan	Daniel Theis	Sindarius Thornwell	Noah Vonleh

Bibliography

[1] basketball analytics.gitlab.io. Nba rapm. https://basketball-analytics.gitlab.io/rapm-data/season/2018-19/regular-season/, 2019.

[2] basketball-reference. Basketball statistics & history of every team & nba and wnba players. https://www.basketball-reference.com, May 2023.

[3] Imad Dabbura. K-means clustering: Algorithm, applications, evaluation methods, and drawbacks. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a, sep 2018.

[4] Nick Faris. Forget the 5 positions: Inside a new way to classify nba players. https://www.thescore.com/nba/news/1974726, 2020.

[5] Matthew Goldberg. Evaluating lineups and complementary play styles in the nba. Master's thesis, Harvard University, 2017.

[6] Yale Undergraduate Analytics Group. Clustering nba players. https://sports.sites.yale.edu/clustering-nba-players, May 2023.

[7] Ahmed Jyad. Redefining nba player classifications using clustering. https://towardsdatascience.com/redefining-nba-player-classifications-using-clustering-36a348fa54a8, November 2020.

[8] NBA. The official site of the nba for the latest nba scores, stats, and news. https://www.nba.com, May 2023.

[9] NBAstuffer. Nbastuffer adjusted plus-minus (apm). https://www.nbastuffer.com/analytics101/adjusted-plus-minus/, 2023.

[10] OpenAI. Chatgpt. https://chat.openai.com/, may 2023. ChatGPT was used to improve grammar or provide a manner of proofreading of this thesis draft. It was not used to generate text from bulleted lists or from an arangement of phrases.

[11] scikit learn. sklearn.decomposition.pca. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html, 2023.

[12] statmuse. https://www.statmuse.com/nba, 2021.

[13] Aaron Barzilai Steve Ilardi. Adjusted plus-minus ratings: New and imporved for 2007-2008. http://www.82games.com/ilardi2.htm, 2008.