Big Data
BUSN 41201
Winter 2025
Professor Veronika Rockova

Final Project
NBA Play Styles and Playoff Success: The Impact of Three-Point Shooting and Team Strategies

By Group 1: Hartej Singh, Ryan Bell, Daniel Zhou, Elliot Paschal, and Amanda Huang

15 March 2025

Chicago Booth Honor Code:
"We pledge our honor that we have not violated the Honor Code during this Project."

Section 1 - Introduction

1.0 Background

A pivotal evolution in basketball strategy was the introduction of the three-point line. Initially popularized by the American Basketball Association (ABA) in the late 1960s to incentivize long-range shooting and differentiate itself from the NBA, the three-point line was adopted by the NBA during the 1979-80 season. This innovation transformed offensive strategies, encouraging teams to incorporate perimeter shooting into their playbooks. Initially met with skepticism and used sparingly, the three-point shot has evolved into a central element of modern basketball, with teams averaging over 37 attempts per game in recent seasons. This shift has led to a more dynamic and unpredictable style of play, reflecting the sport's continuous evolution.[1]

Additionally, the NBA season is divided into the regular season and the playoffs. The regular season typically consists of 82 games, where each team competes to secure the best possible record. The performance during this phase determines playoff seeding, with the top eight teams from each conference advancing. The playoffs are a series of elimination rounds culminating in the NBA Finals, where the conference champions compete for the championship title. Playoff basketball is often characterized by heightened intensity and strategic adjustments. Teams analyze opponents meticulously, leading to more deliberate and defensive styles of play. The margin for error diminishes, and the physical and mental demands on players escalate. Historically, successful playoff teams exhibit adaptability, resilience, and the ability to execute under pressure, distinguishing them from regular-season performances.

Given these dynamics, our research aims to explore the evolution of playstyles over the past two decades and their correlation with playoff appearances and success. By analyzing comprehensive datasets encompassing box scores, play-by-play accounts, and game outcomes, we seek to identify patterns and strategies that have contributed to postseason success. This investigation provides insights into how the game's evolution influences team performance during the most critical phases of competition.

To explore these questions, we leverage big data tools to analyze a comprehensive dataset of NBA games from 1996 to 2019. Specifically, we aim to answer the following research questions:

- Are there distinct play styles in the NBA that are indicative of playoff appearances?
    - Can we use individual player statistics per game to predict playoff appearances?
    - Can we generate new features from play-by-play text data to improve our predictive model?
- Are specific play styles that are prevalent for certain teams predictive of greater postseason success (i.e., making the playoffs)?

By addressing these questions, our analysis sheds light on how teams' strategic decisions impact long-term success in the NBA and provides insights into the evolution of modern basketball.

---

[1] https://www.sfchronicle.com/sports/warriors/article/steph-curry-3-pointers-20207343.php

1.1 Data Description[2]

For this study, we sourced three primary datasets from Kaggle: play_data.csv, boxscore.csv, and games.csv, each offering distinct but complementary insights into NBA games between 1996 and 2019.[3] By merging these datasets, we created a comprehensive dataset that captures game-level, player-level, and play-by-play dynamics, allowing for a deep exploration of basketball strategies and their impact on team success.

The play_data.csv dataset contains detailed play-by-play data for over 13 million plays across all games in the selected period. This dataset provides granular insights into team behavior, in-game events, and scoring patterns, including whether a game took place during the regular season or playoffs. The ability to track individual plays allows for an in-depth examination of team tendencies, offensive and defensive strategies, and the evolving role of three-pointers in modern basketball.

The boxscore.csv dataset, with 741,569 observations, focuses on individual player performances per game, offering a summary of scoring, assists, rebounds, steals, turnovers, and other key performance indicators. By analyzing these statistics at a team level, we can evaluate how different play styles—such as three-point reliance versus inside scoring—affect overall team performance and postseason success.

The games.csv dataset provides high-level game summaries for 30,250 NBA games, including final scores, team names, attendance figures, and an indicator for whether the game was part of the regular season or playoffs. This dataset helps establish broader trends, such as changes in scoring over time, home-court advantage, and variations in attendance.

Merging these datasets enabled us to create a holistic view of NBA gameplay, combining micro-level play-by-play insights, player performance statistics, and macro-level game outcomes. For initial feature extraction from the play-by-play data, we extracted playstyle-specific features using regular expression mapping. Given the size of the data set (at its largest 13.6 million rows, 49 variables) and the computational complexity of wrangling the original dataset as well as its subsequent feature extraction, it was necessary to use both the Arrow and BigMemory libraries in conjunction with parallel processing in R. Arrow allowed us to work with the .parquet data type, permitting highly efficient compression to optimize memory as well as temporary storage during complex pattern mapping across multiple text patterns.[4] Then, BigMemory allowed us to implement clustering analyses on massive matrices using limited local memory.[5] After feature extraction, wrangling, and dummying, the final textual analysis data set used for clustering contained 37 rows (one per team) and 18 columns (one per relevant, team-level playstyle feature). These features were then combined with box score data and further analyzed by team and by year from 1996-2019. This integrated approach allows us to investigate how play styles have evolved, whether certain strategies are predictive of playoff

---

[2] Data Dictionary for the final data set is in the Appendix.
[3]https://www.kaggle.com/datasets/patrickhallila1994/nba-data-from-basketball-reference?select=play_data.csv
[4] https://arrow.apache.org/docs/r/
[5] http://www.stat.yale.edu/~mjk56/Research/Prospectus/bigmemoRy-vignette.pdf

appearances, and whether key performance indicators differ between successful and unsuccessful teams.

1.2 Analysis Overview

Our analysis involved multiple steps to uncover insights into NBA play styles and their impact on postseason success. We began with Exploratory Data Analysis (EDA), where we examined historical trends in three-point shooting, differences in play styles across teams, and the relationship between different scoring strategies and game outcomes. Following this, we extracted playstyle-specific features from the play-by-play data, transforming raw text into structured indicators of team tendencies. Next, we conducted clustering analysis to categorize teams based on their playstyles, identifying distinct strategic approaches over time. To understand the relationship between performance metrics and playoff qualification, we ran marginal regressions, testing the significance of various predictors. Finally, we built predictive models, including PCA-based and LASSO logistic regression models, to assess the best indicators of postseason success.

1.3 Key Takeaways

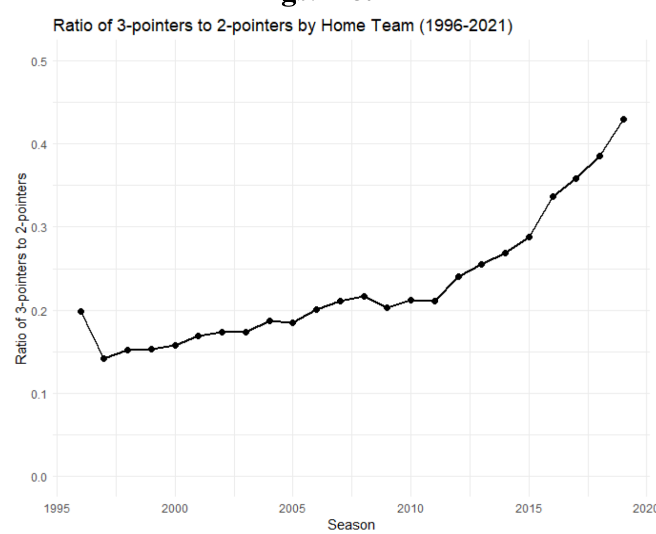
Our findings reveal several key trends in NBA strategy and playoff success. Three-point shooting has grown significantly over the last two decades, becoming a dominant offensive strategy. However, we found that successful teams balance three-point attempts with other strong playmaking abilities, such as assists and efficient shooting percentages (eFG% and TS%). Our regression analysis showed that factors such as assists, three-point makes, turnovers, and shooting efficiency are significant predictors of playoff qualification. Interestingly, our LASSO model indicated that teams with high assist percentage were significantly negatively correlated with postseason success. These findings provide valuable insights into how NBA teams can optimize their strategies to enhance their chances of advancing to the playoffs.

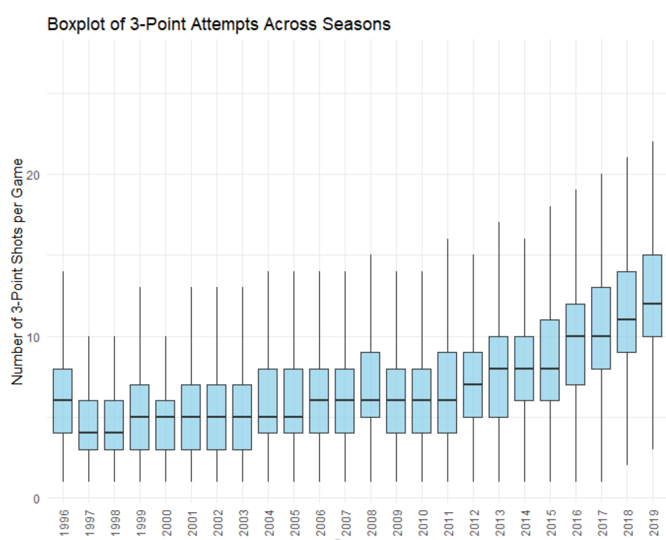
Section 2 - Data Analysis

2.0 Exploratory Data Analysis

Before choosing the appropriate machine learning and regression tools for our analysis, we conducted some exploratory data analysis to analyze trends and patterns in our data. Since the biggest motivation behind our project is to analyze if scoring three-pointers has become a popular strategy over the last few decades, we begin with a time series analysis of the ratio of 3-pointers to 2-pointers scored by the home team between 1996 to 2021 as shown by Fig. 1. We can clearly see an upward trend, implying that teams have been scoring more three-pointers as compared to two-pointers. In fact, the ratio between three-pointers to two-pointers has more than doubled from 0.2 in 1996 to about 0.45 in 2019. The same is also implied by Fig. 2, which shows the marginal distribution of 3-pointers scored per season. We can see an upward trend in the median number of 3-pointers scored per game, with constant variability.

Figs. 1 & 2





Followingly, we were interested to see if scoring three-pointers actually produces better results than scoring two-pointers. Therefore, we analyze the distribution of 2-pointers and 3-pointers scored across wins and losses for the home team in Fig. 3 and Fig. 4, respectively. We can see that wins are associated with scoring more 2-pointers and 3-pointers as compared to losses, and there doesn't seem to be a significant difference between the two-strategies on the final game outcome. This implies that both scoring strategies are equally important, and teams can't rely too heavily on either one. However, it's also important to note that Fig. 3 and Fig. 4 only look at the distribution of 2-pointers and 3-pointers for the home team, implying that there's no difference in game outcome between the two scoring strategies only for the home team.

Fig. 3

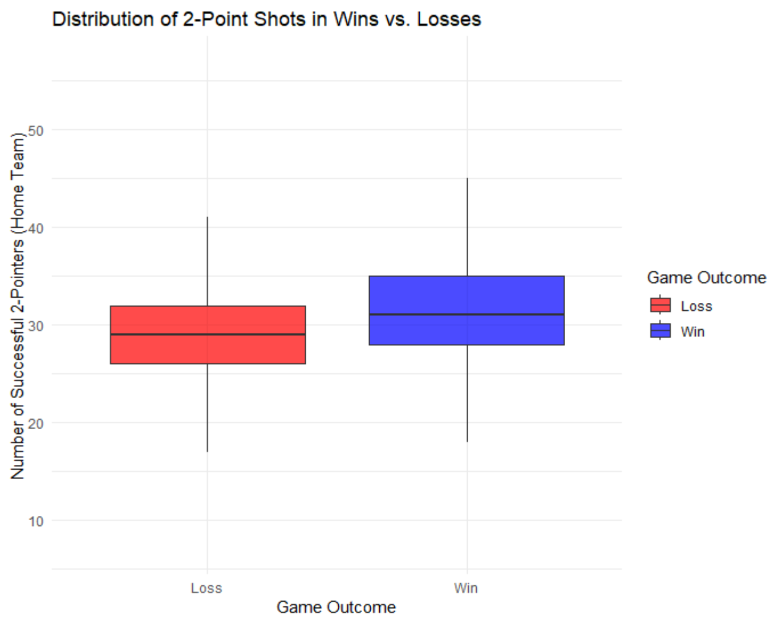
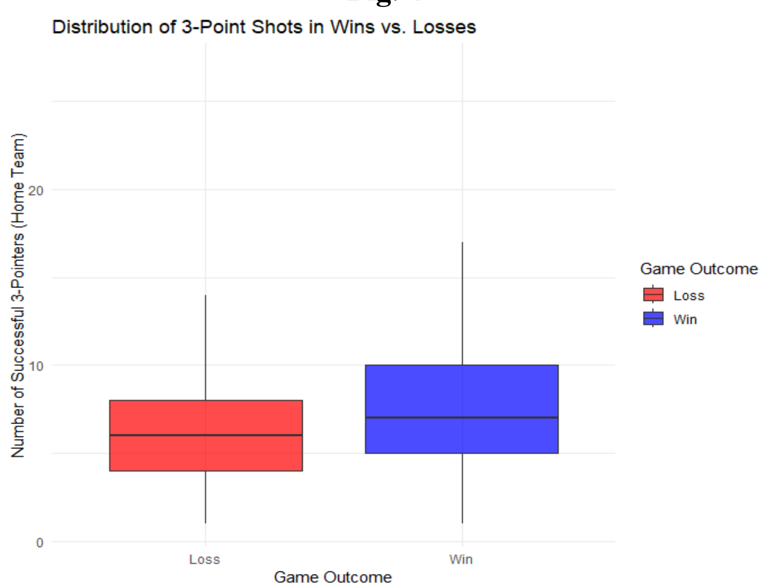


Fig. 4



The above results make sense since if teams use a very aggressive strategy of scoring only 3-pointers, then they are also more likely to lose the ball to the opposition, thereby conceding more points. This may particularly be a problem if a team doesn't have too many three-point scoring players that it can rely on. In such a case, a more conservative approach of scoring two-pointers will yield better results. This is illustrated by Fig. 5 and Fig. 6, where we can notice that there's a positive correlation between scoring two-pointers/three-pointers and the opposition team's final score. However, the correlation seems to be slightly stronger in the case of three-pointers. Nonetheless, we can also notice that the constant variance assumption gets violated in Fig. 5, implying that we might need to log transform the number of points scored by the away team.

Fig. 5

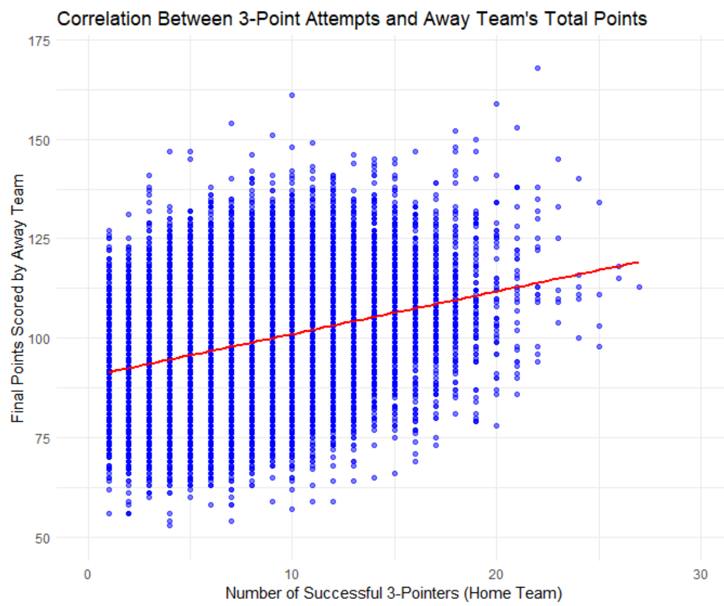
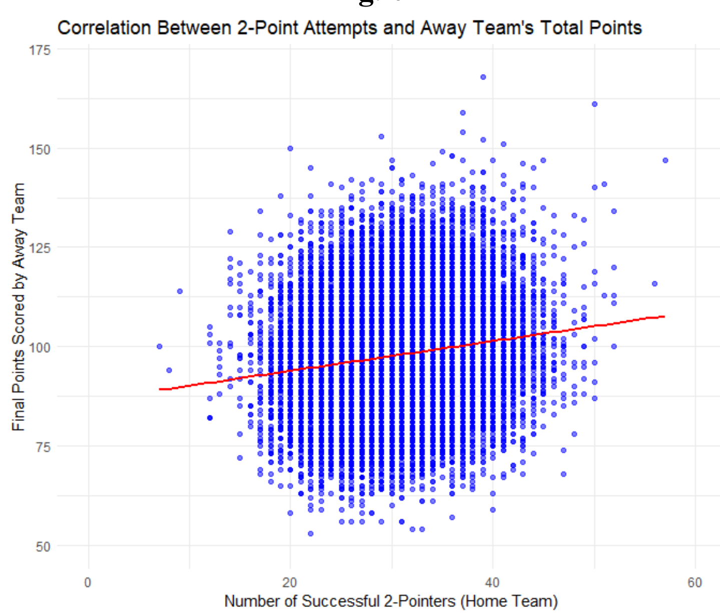


Fig. 6



However, being more defensive, i.e. relying more on two-pointers also has a tradeoff since teams are more likely to score lesser points and is also an indication of a team not having good offensive players. This can be a decisive factor for the game result since the winner is determined by a team's ability to score more points than its opponent. This is illustrated by Fig. 7 and Fig. 8 which show the correlation between winning percentage and average 3-point and 2-point attempts, respectively, irrespective of home team or away team. We can clearly see that the more offensive teams, which attempt more 3-pointers, have a higher winning percentage than the defensive teams, where, in fact, there seems to be a slightly negative correlation between average 2-point attempts and winning percentage. However, we might also need to control for other factors in determining the winning percentage since average 3-point attempts could be endogenous and correlated with other omitted variables such as better players, coaching style, and team's reputation.

Fig. 7

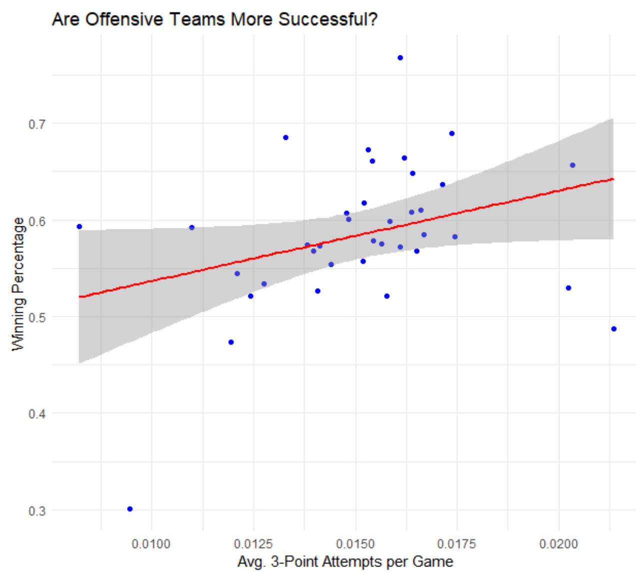
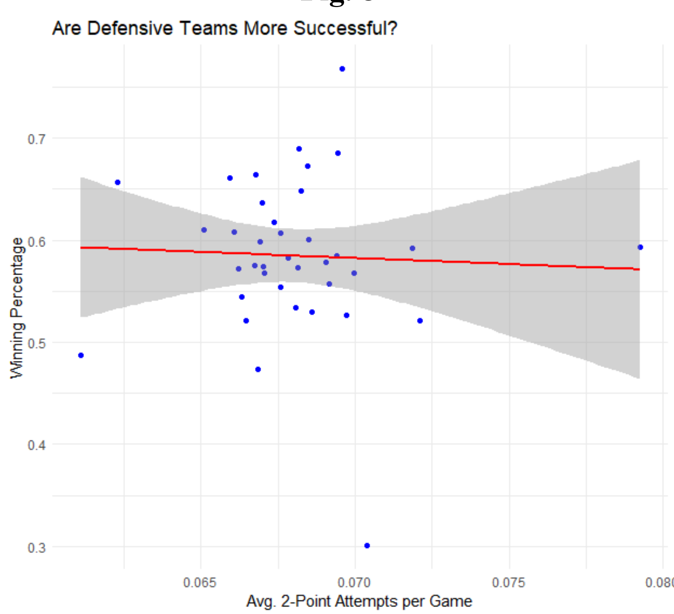
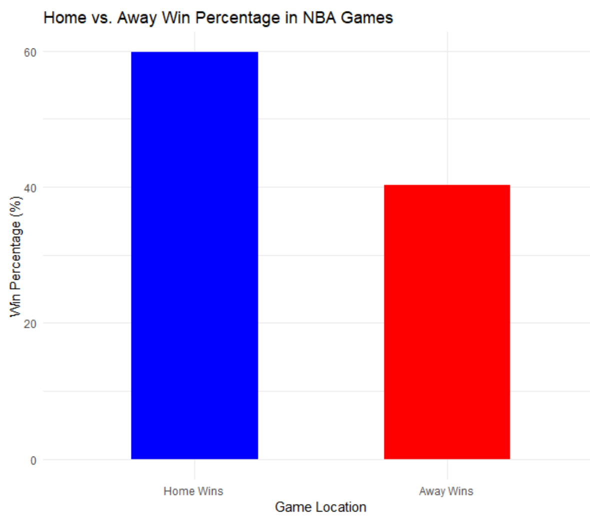


Fig. 8



Now, we can start seeing the reason behind the growing popularity of three-pointers as compared to two-pointers. Given the significance of three-pointers, now, we are interested to see the factors that make it possible for a team to score them. We start by analyzing if a home advantage allows a team to be more aggressive given the moral support from a larger fan base and also greater confidence.

Fig. 9



For this purpose, firstly, we see if a home-advantage actually results in a higher win percentage. As suggested by Fig. 9, there's a clear home advantage, as the win percentage is about 60% for the home team as compared to approximately 40% for the away team.

However, as Fig. 10 and Fig. 11 suggest, the distribution of two-points and three-points is very similar across home and away teams, thereby suggesting that a home-advantage doesn't necessarily improve a team's chances of scoring more three-pointers. This means that there must be other factors, in addition to scoring two-pointers/three-pointers which influence the game outcome.

Fig. 10

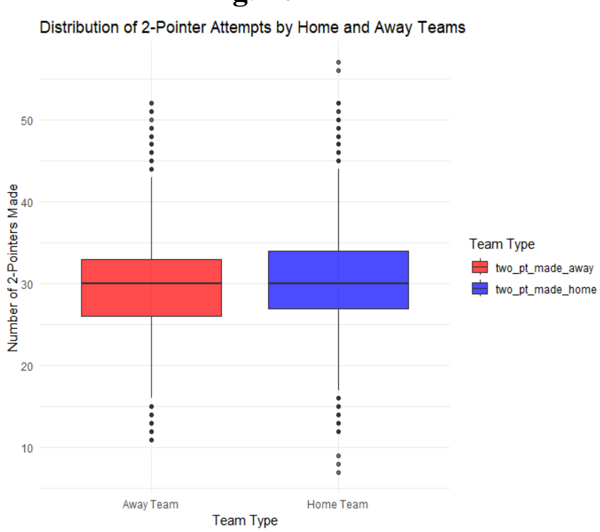
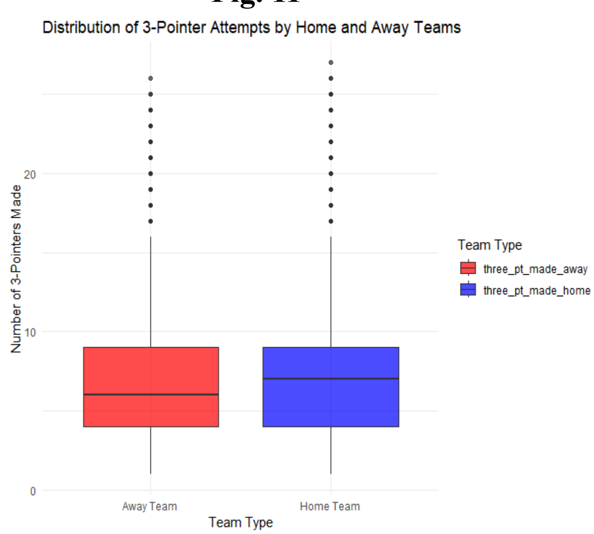
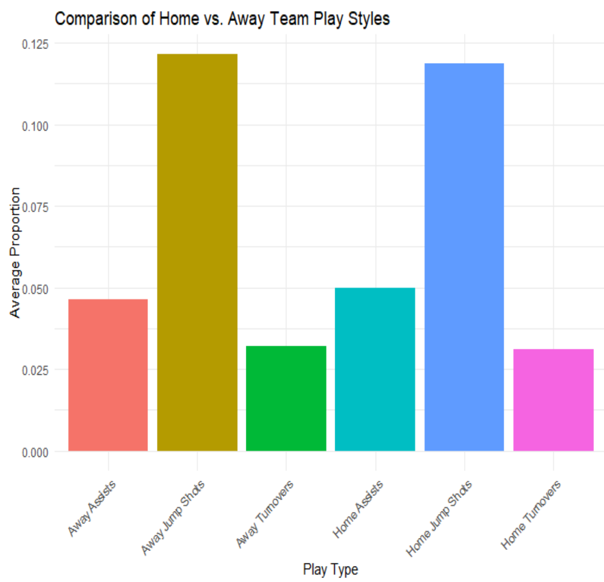


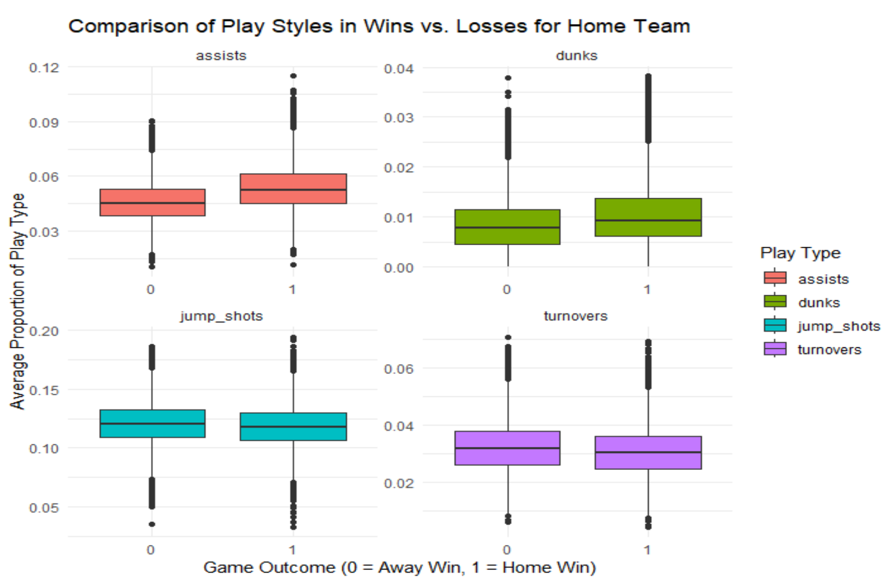
Fig. 11



Followingly, we look for differences in other strategies between the home and away teams. For instance, if three-pointers and two-pointers are similar across home and away teams, then maybe home teams are better at making more assists. As illustrated by Fig. 12, home teams make slightly more assists and also have a lower turnover proportion - although the proportion of jump shots seems to be similar across home and away teams. This result makes sense since home teams showcase a better ball movement (higher assists) and also lose possession less often (lower turnover) than the away team, while jump shots alone don't vary much since the ability to convert jump shots into points matters more in deciding the game outcome. The same is also suggested by Fig. 13 which shows the distribution of assists, dunks, jump shots, and turnovers for the home team across wins and losses. We can clearly see that wins are associated with a greater proportion of assists, dunks, and jump shots, but a lower proportion of turnovers, thereby indicating that these strategies are an important determinant of whether a team wins or loses, implying that relying on three-pointers alone is not an effective strategy.

Figs. 12 & 13





Additionally, we are interested to see if there are any differences in strategies between teams which secure the playoff spots as compared to teams which only play in the regular season. Since the teams which advance to playoffs finish in the top 6 rankings, we expect them to perform better in most parameters. We begin by exploring how the number of 3-point attempts differ between playoff teams and regular teams. Fig. 14 shows that playoff teams have greater number three-pointer attempts on average than regular season teams. This result was expected since playoff teams usually have better quality shooters, which improves their ability to use a more offensive approach. Next, in Fig. 15, we also see if there's a difference in other strategies such as dunks and assists across the regular season games and playoff games. We can notice that the different play styles look mostly similar across playoffs and regular seasons, implying that teams don't usually change their strategies as they advance to the playoffs.

Fig. 14

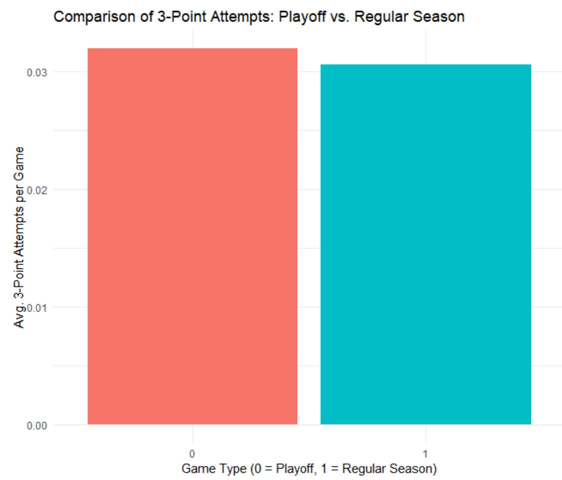
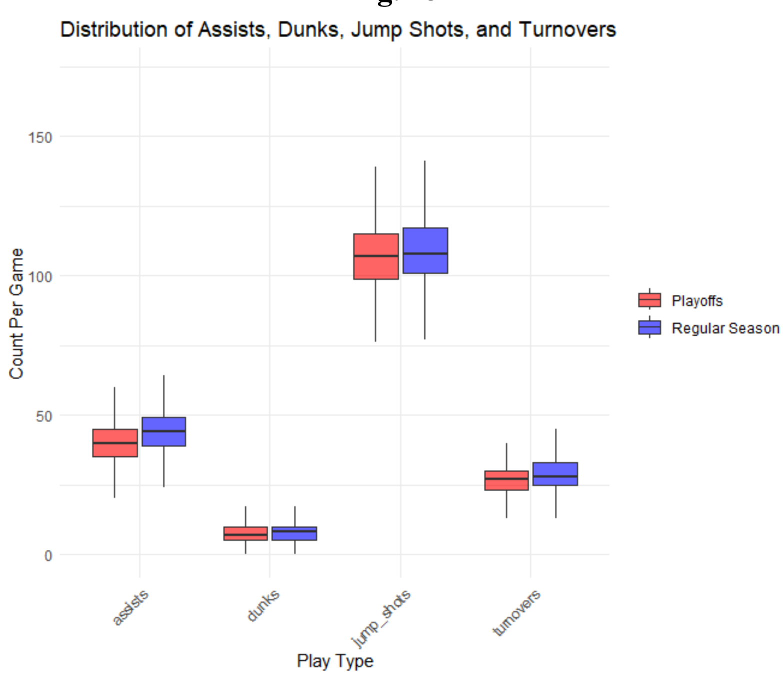
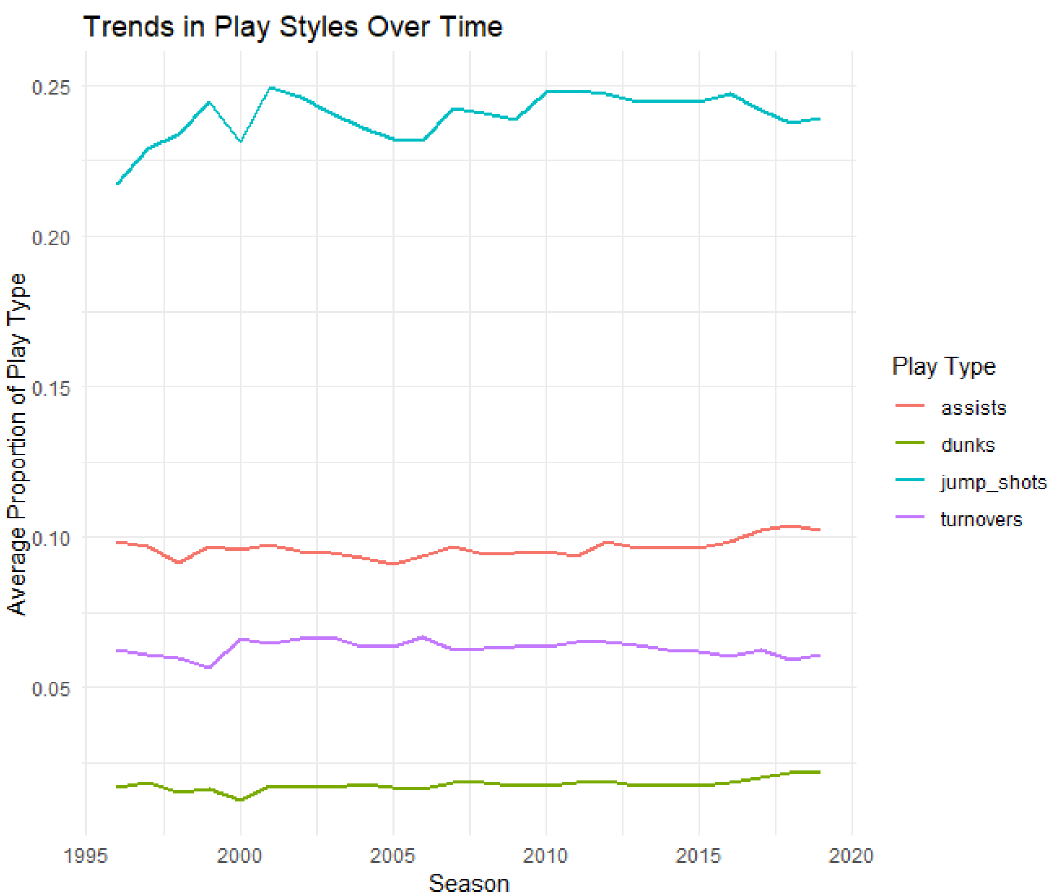


Fig. 15



Lastly, we are also interested to analyse how the play styles have changed over the seasons. We have already clearly seen that teams rely more heavily on three-pointers as suggested by Fig. 1 and Fig. 2. To this end, Fig. 16 shows the change in play styles over the years.

Fig. 16



We can notice that jump shots are the dominant play style across the seasons, and dunks are the least frequent. However, the different play styles (assists, dunks, jump shots, and turnovers) remain constant across the seasons, implying that while three-pointers have witnessed an exponential increase in the last few decades, the same can’t be said for the different play-styles which build up to a two-pointer or three-pointer.

2.1 Initial Clustering Based on Text-Extracted Features

One of the richer elements of our dataset was its play-by-play data, encompassing a string of text describing every play made by both teams for every game from 1996 to 2019. As such, we wanted to mine this information to see what relevant features could be extracted from the text that might enrich our understanding of NBA play styles over time. Originally, the text data frame consisted of two columns and 13.6 million rows. We began by reviewing a randomly selected subset of 100,000 rows of text data to get a feel for which features might be extractable using regular expressions (regex); then, we generated a list of regex patterns that could be mapped to each row of the data, extracting "hits" whenever the pattern matched any part of any string of play-by-play data. In the interest of reducing computational complexity, we combined the individual regex patterns into one string. Then, we created a custom function designed to match any character combination from that string to any corresponding set of characters from the play-by-play data. Finally, we extracted matches to a new column.

```
# Combine patterns into universal pattern:
combined_pattern <- combined_pattern <- paste(patterns, collapse = "|")

## Define custom function to extract pattern matches: ----
extract_matches <- function(txt) {
    hits <- str_extract_all(txt, combined_pattern)[[1]]  # Extract all matches
    if (length(hits) == 0) {
      hits <- NA  # Return NA if no matches are found
    }
    return(hits)
  }
```

Fig. 17

Once text features had been extracted into multiple "feature.hit" columns, we needed to again identify which specific character combination had been detected as a matched pattern in order to create dummy variables on a feature by feature basis. Given the computational complexity of this task, it was necessary to split the data set into four separate chunks of roughly 3.5 million rows each and process individually before rebinding. Even using "chunked" data, it was necessary to use parallel processing across 10-cores to accomplish this task in under 30 minutes per chunk.

```
slice1_dummy <- slice1_tat_matches %>%
  mutate(
    turnover_home_dum = case_when(
      if_any(matches_event_home.1:matches_event_home.4, ~ str_detect(.x, "turnover")) ~ 1,
      TRUE ~ 0
    ),
    turnover_away_dum = case_when(
      if_any(matches_event_away.1:matches_event_away.4, ~ str_detect(.x, "turnover")) ~ 1,
      TRUE ~ 0
    ),
    makes_2_home_dum = case_when(
      if_any(matches_event_home.1:matches_event_home.4, ~ str_detect(.x, "makes 2")) ~ 1,
      TRUE ~ 0
    ),
```

Fig. 18

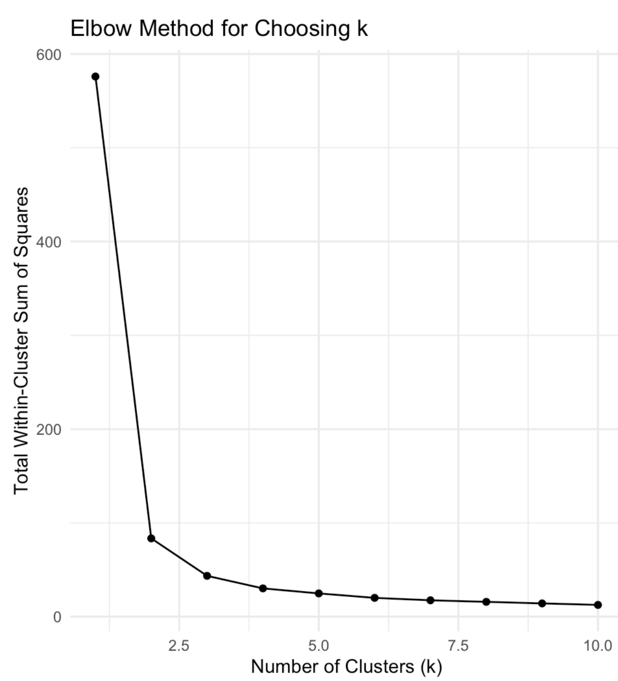
After patterns had been matched across ".hit" columns, it was necessary to pivot the data to take into consideration the fact that all teams played as both home and away during the season. As such, and given that home team plays were recorded in a column separate from away team plays, it was necessary to conduct multiple pivots across parts of the data frame before we would be able to analyze the text on a per-team basis (regardless of whether or not that team was the home or away team in any given game). This was the second most computationally intensive task of all, and required the use of the .parquet data type (used for highly efficient compression and memory optimization) in conjunction with a function that sequentially chunked the data, carried out analysis in a stepwise manner, and deleted all intermediate steps as soon as they were no longer needed. In doing so, we were able to achieve team-by-team features that could ultimately be used for clustering using the BigMemory library.

```r
# Process data in chunks:
for (i in seq(1, total_rows, by = chunk_size)) {
  # Step 1: Slice data into a 1M-row chunk
  chunk <- full_data[i:min(i + chunk_size - 1, total_rows), ]
  # Step 2: Select relevant columns
  chunk <- chunk %>%
    select(game_id, homeTeam, awayTeam, isRegular, ends_with("_dum"))
  # Step 3: Pivot home/away teams into a single column
  chunk <- chunk %>%
    pivot_longer(cols = c(homeTeam, awayTeam), names_to = "homeAway", values_to = "teamName")
  # Step 4: Pivot home/away statistics into long format
  chunk <- chunk %>%
    pivot_longer(cols = ends_with("_dum"), names_to = "stat_type", values_to = "stat_value")
  # Step 5: Match statistics with the correct team
  chunk <- chunk %>%
    mutate(stat_type = case_when(
      str_detect(stat_type, "home") & homeAway == "homeTeam" ~ str_remove(stat_type, "_home"),
      str_detect(stat_type, "away") & homeAway == "awayTeam" ~ str_remove(stat_type, "_away"),
      TRUE ~ NA_character_
    )) %>%
    drop_na(stat_type)  # Remove mismatched cases
  # Step 6: Regroup and summarize
  chunk <- chunk %>%
    group_by(game_id, teamName, isRegular, homeAway, stat_type) %>%
    summarize(total = sum(stat_value, na.rm = TRUE), .groups = "drop")
  # Step 7: Write each processed chunk to a separate Parquet file
  chunk_file <- paste0(output_dir, "chunk_", i, ".parquet")
  write_parquet(chunk, chunk_file)
  # Step 8: Remove chunk from memory
  rm(chunk)
  gc()  # Free memory
}
```

Fig. 19

Then, to carry out our first attempt at clustering, we utilized the BigMemory package in R. BigMemory permits massive matrix computation which we used to generate broad clusters based on the text-extracted features. We created a custom function that would calculate the within-cluster sum of squares and allow us to test multiple values of k as well as multiple random “starting points” within the data frame. To determine the optimal number of clusters, we plotted our within-cluster sum of square results and determined the elbow to be somewhere between two and three clusters. We proceeded with three.

```r
## Determine optimal number of clusters (find elbow): ----

### Define kmeans clustering function using within cluster sum of squares (wcss): ----
wcss <- function(k) {
  set.seed(123)  # Ensure reproducibility

  # Run bigkmeans and handle errors
  kmeans_result <- tryCatch(
    bigkmeans(big_mat_cluster, centers = k, nstart = 5),
    error = function(e) return(NULL)  # Return NULL if clustering fails
  )

  # Check if clustering failed
  if (is.null(kmeans_result)) return(NA_real_)

  # Compute total within-cluster sum of squares manually
  return(sum(kmeans_result$withinss))
}

### Compute WCSS for different values of k: ----
k_values <- 1:10
wcss_values <- map_dbl(k_values, wcss)
```



Figs. 20 & 21

Finally, iterations of both two and three clusters were generated, and we ultimately chose three clusters as the most relevant as it removed teams from our clusters who were not present for every year of data collection (teams that moved to a new city, or were created as expansion teams, for example). We visualized clusters by corresponding play style features, and mapped back to the original data frame such that teams could be assigned to each cluster. At first look, it appears that the same features tend to be significant in defining each style of play, but differ in the extent to which they are present. Cluster two is indicative of a more aggressive style of play, wherein a team is essentially committing more of every kind of play throughout the course of a game (positive offense, negative offense, positive defense, and negative defense). Cluster three appears to indicate less aggressive playing styles, while the teams in Cluster 1, in addition to underperforming, were also the teams who were least present in our data - either they did not exist, their names were changed throughout the course of the two decades included in our data set, or they were newly created since 1996. As such, two distinct styles of play - passive and aggressive - emerged from our initial clusters.

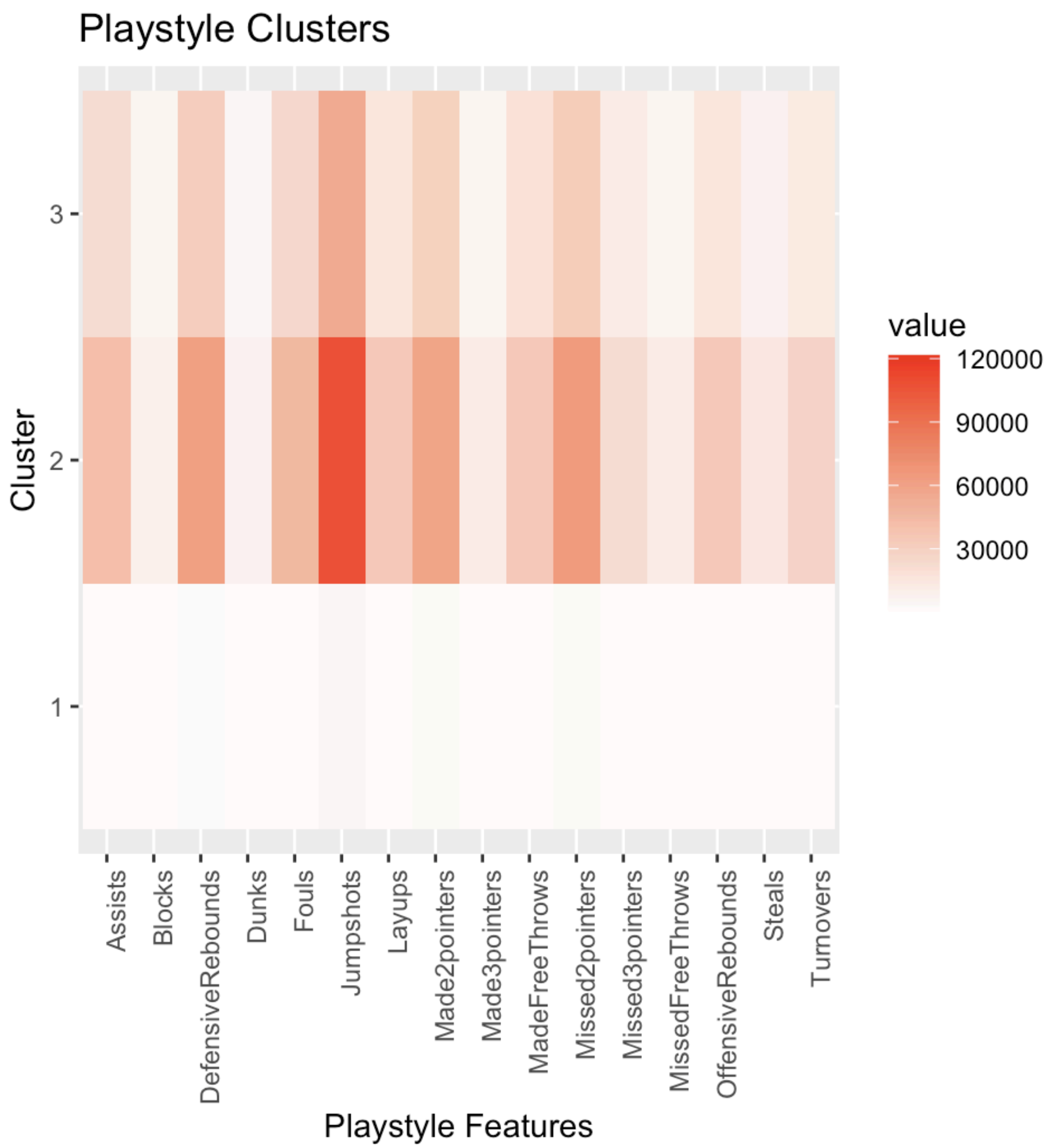


Fig. 22

Finally, in mapping the team names to each play style cluster, it was interesting to note that all winners of all NBA Championships during the analyzed time period came from Cluster 2. Also interesting, losers in the NBA Championship came from a combination of Clusters 1 and 2, seeming to indicate that more aggressive play styles tend to result in more team success at the highest level.

Cluster	Team Names
1	New Orleans/Oklahoma City Hornets, Vancouver Grizzlies, Washington Bullets
2	Atlanta Hawks, Boston Celtics, Chicago Bulls, Cleveland Cavaliers, Dallas Mavericks, Denver Nuggets, Detroit Pistons, Golden State Warriors, Houston Rockets, Indiana Pacers, Los Angeles Clippers, Los Angeles Lakers, Memphis Grizzlies, Miami Heat, Milwaukee Bucks, Minnesota Timberwolves, New York Knicks, Orlando Magic, Philadelphia 76ers, Phoenix Suns, Portland Trail Blazers, Sacramento Kings, San Antonio Spurs, Toronto Raptors, Utah Jazz, Washington Wizards
3	Brooklyn Nets, Charlotte Bobcats, Charlotte Hornets, New Jersey Nets, New Orleans Hornets, New Orleans Pelicans, Oklahoma City Thunder, Seattle Supersonics

Fig. 23

While these results are interesting, we also realized that many of the text features extracted ultimately ended up being features that were already tracked, in conjunction with other variables not present in text data, by the box score data. As such, we determined that it should be possible to extract more distinguishable playstyles using more robust box score data, and decided to use only those text features stylistically relevant that did not originally appear in the box score, namely type of shot taken and shot point value. The results are discussed in more detail in the following section.

2.2 Refined Clustering Based on Player Statistics

Given that our data encompasses 20 years of NBA games, we thought it may be interesting to analyze trends in playing styles over time, especially as new forms of offense have developed.[6] In order to do this, we first look at box score data, which is team totals data for a given game. This data set consisted of about 60,500 rows of data, two rows per game, one for each team, and gave statistics totals, both offensive and defensive for each game. The variables can be found in Fig. 24 below and show both the given variables, as well as derived metrics, starting with Three Point Attempt Rate, which are common variables used for NBA analytics. It's also important to note that although the figure shows PlusMinus (+/-) which is a sum of individual +/- scores, we removed this variable because an aggregate +/- becomes a total point differential at the team level, meaning it's directly correlated with game outcomes rather than distinct playing styles.

Variables Used in Box Score Data

Grouped by Category

Category	Variable	Description
Scoring Profile	PPG	Points Per Game
Scoring Profile	TFGA	Total Field Goal Attempts
Scoring Profile	TFGM	Total Field Goals Made
Scoring Profile	T3PA	Three-Point Attempts
Scoring Profile	T3PM	Three-Pointers Made
Scoring Profile	TFTA	Free Throw Attempts
Scoring Profile	TFTM	Free Throws Made
Scoring Profile	PlusMinus	Total team +/- for the game
Playmaking & Turnovers	APG	Assists Per Game
Playmaking & Turnovers	TTOV	Turnovers Per Game
Rebounding & Defense	TRPG	Total Rebounds Per Game
Rebounding & Defense	TORB	Offensive Rebounds
Rebounding & Defense	TDRB	Defensive Rebounds
Rebounding & Defense	TSPG	Steals Per Game
Rebounding & Defense	TBPG	Blocks Per Game
Rebounding & Defense	TPF	Total Personal Fouls
Plus/Minus	PlusMinus	Team total +/- for the game (point differential)
Derived Shooting Metrics	T3PAr	Three-Point Attempt Rate (3PA / FGA)
Derived Shooting Metrics	TFTAr	Free Throw Attempt Rate (FTA / FGA)
Derived Shooting Metrics	eFG	Effective Field Goal Percentage
Derived Shooting Metrics	TS	True Shooting Percentage
Assist Metrics	AST_PCT	Assist Percentage (APG / (TFGM + APG))
Assist Metrics	AST_TOV	Assist-to-Turnover Ratio
Rebounding Percentages	ORB_PCT	Offensive Rebounding Percentage
Rebounding Percentages	DRB_PCT	Defensive Rebounding Percentage

Fig. 24

[6] https://www.thespax.com/nba/the-rise-of-heliocentrism-in-nba-offenses/

To analyze this data, we then decided to perform a hierarchical cluster, grouping by year in order to see the resulting styles of play. To do this, we ran a for loop that normalizes the data and clusters the data by year with each row representing a certain team's game statistics. This results in season's with different normalized means and variances which means comparing specific statistics across years could be challenging, but for the purpose of within-season clustering, it is sufficient. To cluster the data, we use Ward's minimum variance method which minimizes the total within-cluster variance, aiming to create clusters that are internally cohesive by merging the two clusters that lead to the smallest increase in within-cluster variance. Additionally, we chose a K=3 because it gave the most distinct styles of play - larger Ks resulted in struggles to determine actual playing styles.

To determine playing style, we created labels to determine if a group had a superior statistic compared to other clusters. This gave us a better idea of styles and potentially strategies that teams chose to implement. The labels were determined in Fig. 25 below.

Determination of Playstyles

Criteria for Categorizing Playstyles Based on Statistical Percentiles

Playstyle Category	Statistical Criteria
Elite Scoring & Efficient	PPG percentile ≥ 67% & eFG% percentile ≥ 67%
High-Scoring	PPG percentile ≥ 67%
Three-Point Heavy & Efficient	3PAr percentile ≥ 67% & eFG% percentile ≥ 67%
Three-Point Heavy	3PAr percentile ≥ 67%
Free-Throw Heavy & Inside Attack	FTAr percentile ≥ 67%
Defensive & Strong Rebounding	Defense percentile ≥ 67% & TRPG percentile ≥ 50%
Defensive-Oriented	Defense percentile ≥ 67%
Fast-Paced & Playmaking	PPG percentile ≥ 33% & 3PAr percentile ≥ 33% & AST/TOV percentile ≥ 33%
Slow-Paced & Defensive	PPG percentile ≤ 33% & TRPG percentile ≥ 67%
Elite Rebounding Team	TRPG percentile ≥ 67% & ORB_PCT percentile ≥ 67%
Rebounding-Focused	TRPG percentile ≥ 67%
Balanced Playstyle	No other label assigned

Fig. 25

After averaging the statistics for clusters of each year, we were able to create a table, Fig. 26 shown below, that gave us the labels for each cluster, allowing us to understand what types of playing styles there were for each season. Fig. 26 illustrates which play style was most prevalent among NBA teams over time:

Cluster Playstyles by Season

Summary of Playstyles Assigned to Each Cluster Per Season

Season	Cluster	Playstyle
1996	1	Three-Point Heavy
1996	2	High-Scoring, Defensive & Strong Rebounding, Defensive-Oriented, Elite Rebounding Team, Rebounding-Focused
1996	3	Free-Throw Heavy & Inside Attack, Fast-Paced & Playmaking
1997	1	Free-Throw Heavy & Inside Attack
1997	2	Balanced Playstyle
1997	3	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused
1998	1	Defensive & Strong Rebounding, Defensive-Oriented, Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
1998	2	Free-Throw Heavy & Inside Attack
1998	3	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
1999	1	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
1999	2	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive-Oriented, Fast-Paced & Playmaking
1999	3	Free-Throw Heavy & Inside Attack
2000	1	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2000	2	Free-Throw Heavy & Inside Attack, Fast-Paced & Playmaking
2000	3	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking
2001	1	Free-Throw Heavy & Inside Attack
2001	2	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking
2001	3	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2002	1	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused
2002	2	Free-Throw Heavy & Inside Attack
2002	3	Balanced Playstyle
2003	1	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2003	2	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2003	3	Free-Throw Heavy & Inside Attack, Defensive-Oriented, Fast-Paced & Playmaking
2004	1	Balanced Playstyle
2004	2	High-Scoring, Three-Point Heavy, Fast-Paced & Playmaking, Rebounding-Focused
2004	3	Free-Throw Heavy & Inside Attack, Defensive-Oriented, Fast-Paced & Playmaking
2005	1	Three-Point Heavy, Fast-Paced & Playmaking, Elite Rebounding Team, Rebounding-Focused
2005	2	Elite Scoring & Efficient, High-Scoring, Free-Throw Heavy & Inside Attack, Defensive-Oriented, Fast-Paced & Playmaking
2005	3	Balanced Playstyle
2006	1	Free-Throw Heavy & Inside Attack, Fast-Paced & Playmaking
2006	2	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking
2006	3	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2007	1	Free-Throw Heavy & Inside Attack, Fast-Paced & Playmaking
2007	2	Balanced Playstyle
2007	3	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused

2008	1	Balanced Playstyle
2008	2	High-Scoring, Three-Point Heavy, Free-Throw Heavy & Inside Attack, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused
2008	3	Fast-Paced & Playmaking
2009	1	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2009	2	Free-Throw Heavy & Inside Attack
2009	3	Defensive & Strong Rebounding, Defensive-Oriented, Elite Rebounding Team, Rebounding-Focused
2010	1	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2010	2	Free-Throw Heavy & Inside Attack, Defensive-Oriented
2010	3	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2011	1	Balanced Playstyle
2011	2	Free-Throw Heavy & Inside Attack, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused
2011	3	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2012	1	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking
2012	2	Fast-Paced & Playmaking
2012	3	Free-Throw Heavy & Inside Attack, Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2013	1	Fast-Paced & Playmaking, Rebounding-Focused
2013	2	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Free-Throw Heavy & Inside Attack, Defensive-Oriented, Fast-Paced & Playmaking
2013	3	Balanced Playstyle
2014	1	Fast-Paced & Playmaking, Elite Rebounding Team, Rebounding-Focused
2014	2	Free-Throw Heavy & Inside Attack
2014	3	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive-Oriented, Fast-Paced & Playmaking
2015	1	Balanced Playstyle
2015	2	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2015	3	Free-Throw Heavy & Inside Attack, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused
2016	1	Free-Throw Heavy & Inside Attack
2016	2	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2016	3	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking
2017	1	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking, Rebounding-Focused
2017	2	Balanced Playstyle
2017	3	Free-Throw Heavy & Inside Attack, Defensive-Oriented
2018	1	Fast-Paced & Playmaking, Elite Rebounding Team, Rebounding-Focused
2018	2	Free-Throw Heavy & Inside Attack
2018	3	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive-Oriented, Fast-Paced & Playmaking
2019	1	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2019	2	Defensive & Strong Rebounding, Defensive-Oriented, Elite Rebounding Team, Rebounding-Focused
2019	3	Free-Throw Heavy & Inside Attack

Fig. 26

Again, because the statistics are relative to clusters within a certain year, it is harder to understand how styles of play change over the years; we cannot, for instance, determine how three point shots changed over the years as a playing style. We can, however, find something much more interesting, and that is how the labels are paired with each other throughout the years. Different pairs of labels refer to different playing styles, showing us how different strategies may be allowing teams better performance and co-evolving together. The rise from Three-Point Heavy to Three-Point Heavy & Efficient tells us that there was a motion towards shooting three pointers as a primary strategy of the course of our data. In the early 2000s, Three-Point shooting was often paired with inside scoring and rebounding but shifted to being paired with Fast-Pace play around the 2010s. In the late 2010s, Three-Point Heavy was paired more often with Elite Scoring & Efficient and High-Scoring showing that using three-pointers became a strategy for offensive minded, high octane teams. A similar trend can be seen in Fast-Pace, the speed at which an offense plays. In the earlier years, this was often paired with Inside Attack and Free-Throw Heavy, an era of basketball dominated by big post players who were running the floor and aggressively attacking the basket. But quickly, the Fast-Pace began to be paired with Three-Point Heavy styles of play, often referred to as "run and gun," a style of play that is especially impactful today. As a result of changes such as these, it seems that traditional slow-paced, defensive-oriented teams have diminished, making way for high-tempo offenses that optimize scoring through playmaking, movement, and three-point shooting rather than isolation-heavy, post-dominant schemes.

After finding this, we thought it might be telling to see what style of play championship teams had over the years, hoping to point out similarities between styles that could give us a hint to the optimal style of play for champions. To do this, we merged in a championship column and found out how often they were in a column for a given year.[7]

Surprisingly, we found that championship teams did not hold a certain style of play as consistently throughout a season as we thought they would. For instance, in a few situations, such as the 2016 Cleveland Cavaliers, they split the three styles of play relatively evenly with 36%, 33%, and 31%, respectively. Ultimately though, the broader trend of becoming Three-Point Heavy shows a similarity to the broader trends within the NBA, also correlating with Elite Scoring and Efficient, meaning championship teams have become much more offensive than defensive in contradiction to the old saying that "defense wins championships." However, unlike the general league, there were fewer correlations between different play styles in championship teams suggesting that the best teams tend to dominate with fewer but highly optimized strategies rather than blending multiple styles. This could be extrapolated to the condition that championship teams have a specialization that may be related to specific players, although this was something we could not find with this level of analysis.

The last bit of information we wanted to see was the dendrograms representing the clusters for each year, hopefully getting some insight into the strength of differences between playing styles. These dendrograms were extremely large originally, so we pruned them to K=3 to see where the clusters merged.

---

[7] Refer to Appendix for Fig. 27 - note that the Total Appearances changed significantly throughout the years due to the fact that some seasons were cut short or had a change in seasons played.

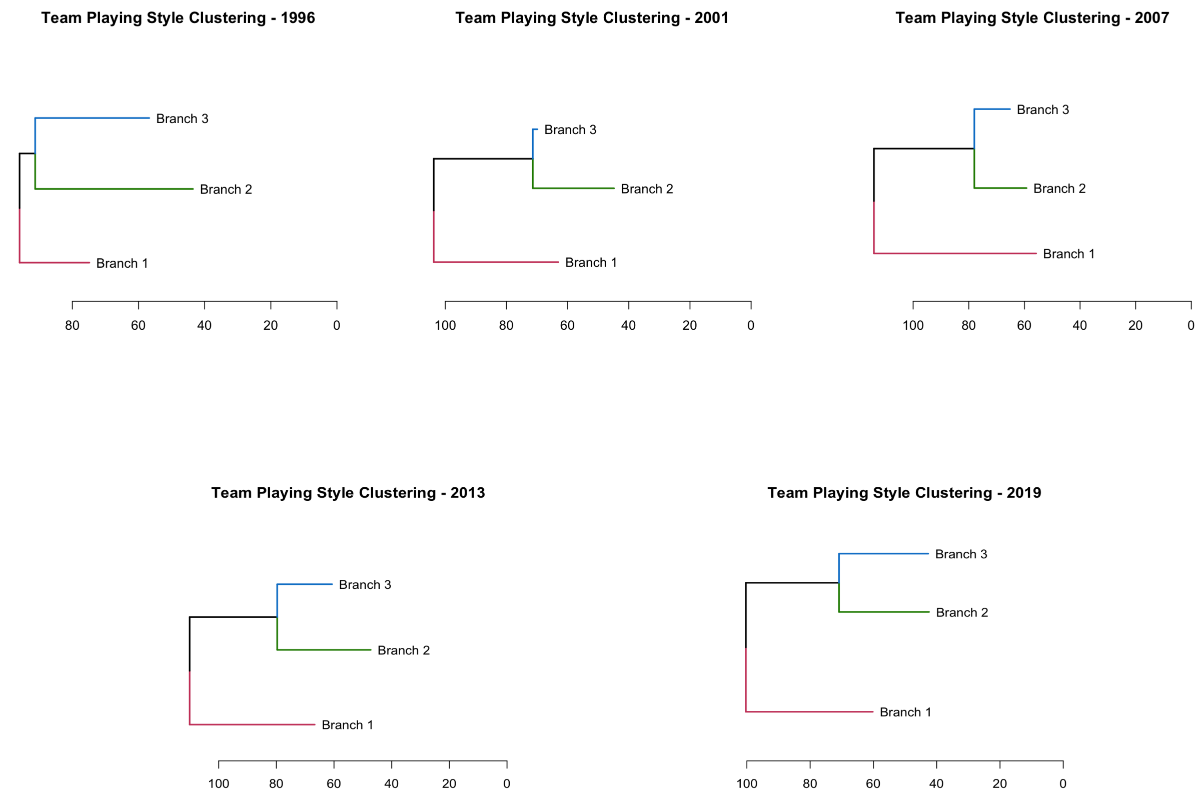


Fig. 28

The dendrograms back up the broader trends in NBA play style evolution, showing how distinct team strategies were in different eras. In 1996 and 2001, teams had clearer play style differences, aligning with an era dominated by slow-paced, defensive, and rebounding-heavy teams. By 2007 and 2013, the clusters merged at larger distances, meaning playstyles became less distinct as teams blended strategies. This shift matches the rise of Three-Point Heavy & Efficient teams, which started as a niche and became the norm, especially among championship teams. The increasing correlation between three-point efficiency and elite scoring suggests that winning teams focused more on efficiency than on varied styles, reinforcing why playstyles looked more similar in later years. By 2013, high-efficiency, three-point-heavy play had taken over, driving the league toward a more unified strategic approach. The dendrograms validate what the data shows—NBA teams used to play in more distinct ways, but modern basketball has pushed them toward a shared formula of efficiency, spacing, and perimeter shooting.

To conclude, our analysis of NBA play styles over time, using hierarchical clustering and dendrograms, reveals a clear shift toward offensive efficiency, fast-paced play, and three-point shooting. In the late 1990s and early 2000s, teams exhibited more distinct playstyles, with some favoring slow-paced, defensive, and rebounding-heavy approaches while others leaned toward high-scoring or fast-paced inside attacks. However, as the league evolved, these distinctions blurred, with styles merging at higher distances in later years. This convergence aligns with the widespread adoption of Three-Point Heavy & Efficient strategies, which were initially paired with inside scoring and rebounding but later became associated with high-tempo, playmaking-oriented offenses.

Championship teams followed a similar trajectory, shifting from varied styles to more specialized, high-efficiency offensive strategies. While early championship teams exhibited diverse playstyles, recent winners have overwhelmingly prioritized elite scoring and efficiency, mirroring broader league trends. Interestingly, championship teams showed fewer correlations between different playstyles, suggesting that dominance often comes from mastering a single strategy rather than blending multiple approaches. This contradicts the traditional notion that defense wins championships, as modern title-winning teams are more offensively driven.

The dendrograms reinforce this trend, showing that NBA play styles were more distinct in earlier eras but have since homogenized. By 2019, teams had largely converged on a shared formula—high-efficiency, three-point-heavy offenses that prioritize spacing, pace, and scoring versatility. This shift reflects the evolution of basketball strategy, where optimization of offensive output has become the defining characteristic of successful teams.

However, winning championships is not the only goal for most teams since advancing to the playoffs is also a major breakthrough. Therefore, in the next section, we now evaluate the different play styles which improve a team’s chances of making it to the playoffs.

2.3 Marginal Regressions of Key Metrics on Playoff Chances

We then proceed with addressing our second set of research questions, i.e. are specific play-styles predictive of post-season success (making it to the playoffs). For this purpose, we grouped the data by team and season to obtain season-level averages for each performance metric, and created a new binary variable, MadePlayoffs, where a team was assigned a 1 if it qualified for the playoffs in that season and 0 otherwise. The resulting dataset allows us to analyze trends over time and train models to predict playoff qualification based on regular-season performance.

For this purpose, we begin with running a series of marginal regressions of our response variable, MadePlayoffs, on all the key metrics for play styles which we derived from the previous sections. Running these independent regressions, we get a set of marginal p-values which have been depicted by Fig. 29 and Fig. 30 below.

Fig. 29

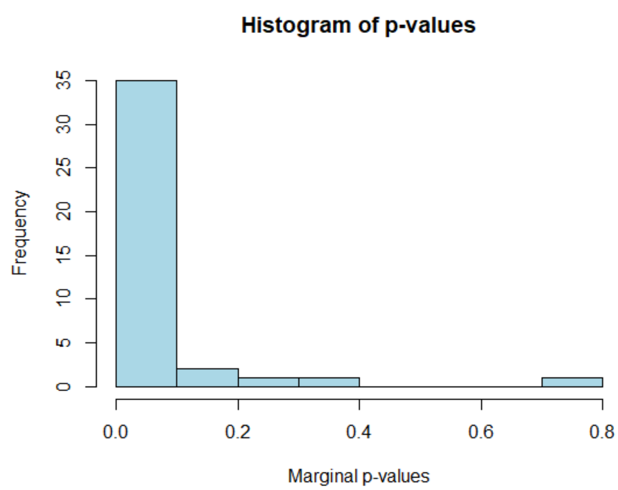
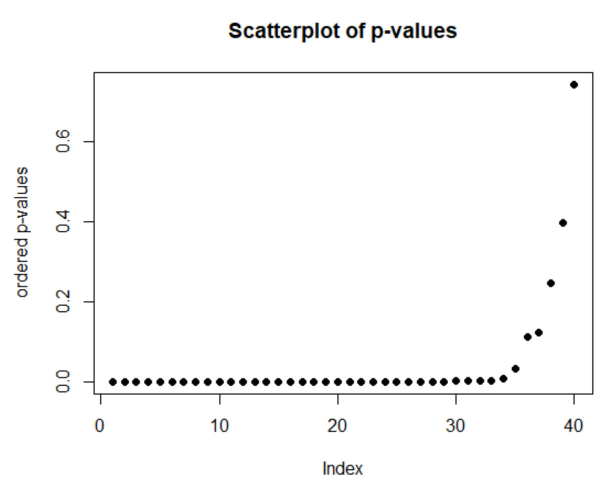


Fig. 30



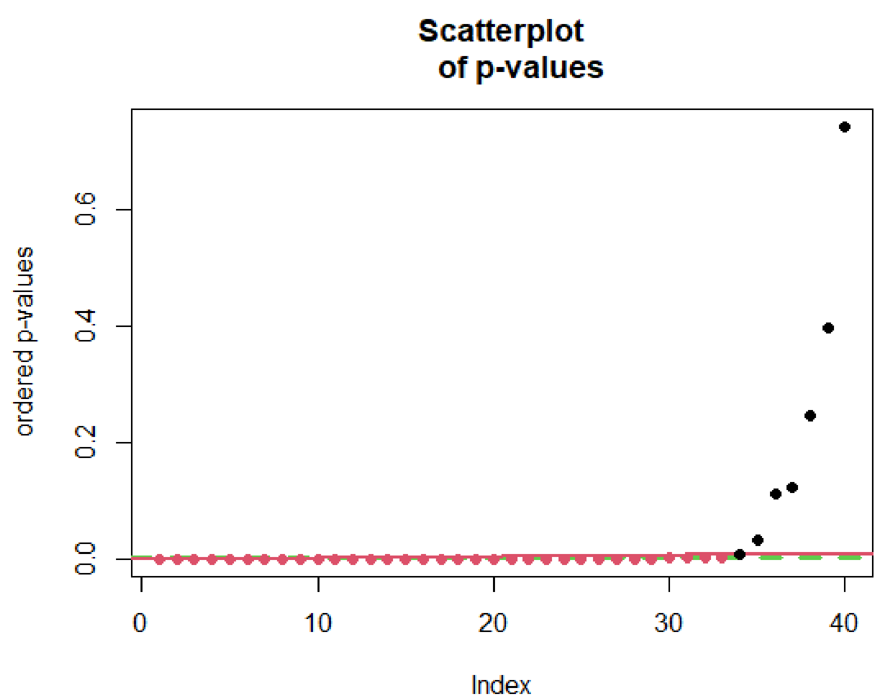
Looking at the scatterplot, we can see that when we plot the ordered p-values, we get a convex curve which means that there are some p-values which are smaller than what we would expect under the null hypothesis that there is no signal at all. The same pattern can also be seen in the histogram for p-values which shows that we don't have a uniform distribution (with a peak near 0), where the uniform distribution would have implied that there is no signal/everything is noise. However, that is clearly not the case here due to the convexity, i.e. some p-values are very close to 0, thereby implying that we probably have some significance.

Followingly, we use the FDR analysis to find the cutoff p-value that we use for our analysis, where q = 1%. FDR analysis makes it possible for us to account for multiplicity when p is large. This means that when p is large, and we conduct p hypothesis tests, then the probability of making a type 1 error / False Discovery keeps on compounding. Therefore, we can't proceed with an alpha = 0.05, and instead the FDR analysis allows us to choose an appropriate level of alpha given a threshold of FDR.

Given an FDR of 1% for this analysis, we find that the corresponding p-value cutoff is 0.0034, where the rejection region (points below the red line) can be depicted by using the plot below.

We find that at this p-value cutoff, 33 of the covariates are significant, and 7 of them are insignificant in affecting the chances of a team advancing to the playoffs. Fig. 32 summarizes the significant predictors from the most significant to the least significant.

Fig. 31



Predictor	Coefficient	P_Value
PlusMinus	0.2065261	0.0000000
TS	41.8920638	0.0000000
eFG	32.3476416	0.0000000
AST_TOV	3.2093593	0.0000000
block_dum	-0.8866369	0.0000000
turnover_dum	-0.5814141	0.0000000
misses_2_dum	-0.1677319	0.0000000
TTOV	-0.5191197	0.0000000
steal_dum	-0.7624004	0.0000000
defensive_rebound_dum	0.2349972	0.0000000
TDRB	0.2211353	0.0000000
APG	0.2212912	0.0000001
assist_dum	0.2200396	0.0000001
ORB_PCT	-12.8650263	0.0000001
DRB_PCT	12.8650263	0.0000001
PPG	0.0616662	0.0000010
TFTM	0.1820847	0.0000101
TFTAr	8.0205624	0.0000447
makes_free_dum	0.1699141	0.0000450
AST_PCT	19.2257990	0.0000452
TORB	-0.2144463	0.0000577
foul_dum	0.1629129	0.0000729
TBPG	0.3735776	0.0000948
TRPG	0.1424128	0.0002021
makes_3_dum	0.1112544	0.0002904
offensive_rebound_dum	-0.1436116	0.0002931
T3PM	0.1092762	0.0003967
TFTA	0.1039077	0.0005302
T3PAr	3.5704512	0.0006199
TSPG	0.2918366	0.0008578
TFGA	-0.0660162	0.0009518
TFGM	0.1074098	0.0018313
TPF	-0.1290964	0.0033829
T3PA	0.0295658	0.0087079
misses_3_dum	0.0375916	0.0320267
layup_dum	-0.0241312	0.1109812
dunk_dum	0.1004576	0.1235179
makes_2_dum	-0.0443937	0.2463063
jumpshot_dum	-0.0167705	0.3965121
misses_free_dum	0.0230835	0.7434479

Fig. 32

In the above table, all the predictors till TPF are significant, given our p-value cutoff. From this analysis, we have been able to find the most important play styles that teams need to incorporate to advance to the playoffs. For instance, assist_dum which represents the total number of assists made in the game is highly significant, as expected, to a team's success with a coefficient of 0.22 which means that an additional assist increases the chances of making it to the playoffs by $(e^{0.22} - 1) * 100 = 24.61\%$. On the other hand, the coefficient for makes_3_dum, i.e. total numbers of 3-pointers scored is 0.11 which means that an additional 3-pointer scored increases chances of making it to the playoffs by $(e^{0.11}-1) *100 = 11.63\%$. This implies that a team should focus more on improving playmaking ability (assists) instead of relying solely on 3-pointers to be able to improve their performance. Interestingly, we also find a negative coefficient of -0.04 for makes_2_dum which implies that scoring an additional 2-pointer actually decreases the chances of advancing to the playoffs. This provides evidence for the fact that teams have been increasingly relying on 3-pointers over the last several years. Simultaneously, as implied by the negative coefficients for poor plays such as turnovers (giving away possession), missing shots, and personal fouls, a team should try to minimize these errors and use a more disciplined approach to improve playoff chances.

2.4 Predicting Post-Season Success

When deciding on the approach of our prediction model, we wanted to make sure the techniques we used aligned with our objective of seeing if any of the metrics are significant in predicting playoff chances. To achieve this, we built a PCA and a LASSO based logistic regression model, comparing the predictive capabilities and interpretability of both techniques.

In our data format from section 2.3, each performance measure represents the team's average of that particular metric for that year. In this case for example, a team's PPG feature is the average points they scored per game in that regular season, and AST_TOV is their average assist-to-turnover ratio.

Thus, our target variable is the binary variable, MadePlayoffs. For each season and team, if the team had at least one postseason game, we label that season as MadePlayoffs = 1 and 0 otherwise. For example, the 1996–97 Chicago Bulls would have MadePlayoffs = 1 because they played postseason games, whereas a team with no playoff games in that season gets MadePlayoffs = 0.

To test the model, we split the data chronologically using an 80/20 training/test split. Thus, all seasons from 1996 up to 2016 were used as the training set, and seasons 2017 to 2021 were held out as the test set, simulating how the model would perform on new, unseen seasons.

Fig. 33

```
# Train-test split
train_data <- filter(model_data, seasonStartYear <= 2016)
test_data  <- filter(model_data, seasonStartYear >= 2017)

x_train <- model.matrix(MadePlayoffs ~ . - teamName - seasonStartYear, data=train_data)[,-1]
y_train <- train_data$MadePlayoffs

x_test <- model.matrix(MadePlayoffs ~ . - teamName - seasonStartYear, data=test_data)[,-1]
y_test <- test_data$MadePlayoffs
```

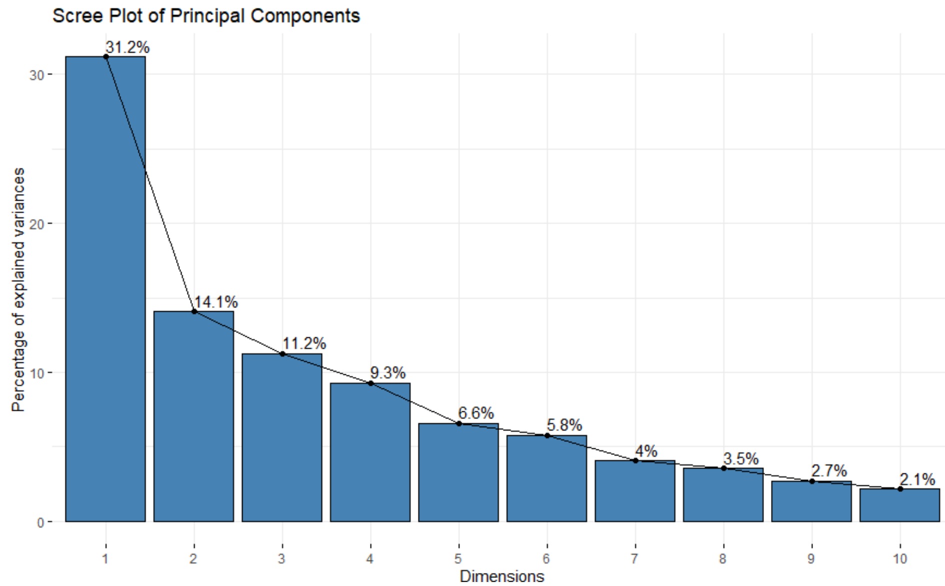


Fig. 34

We first explored a standalone PCA analysis for dimensionality reduction purposes. We applied PCA on the training data features to condense the information into orthogonal components. The above graph shows that the first principal component explains the most amount of variance (31.2%) in the playoff outcome.

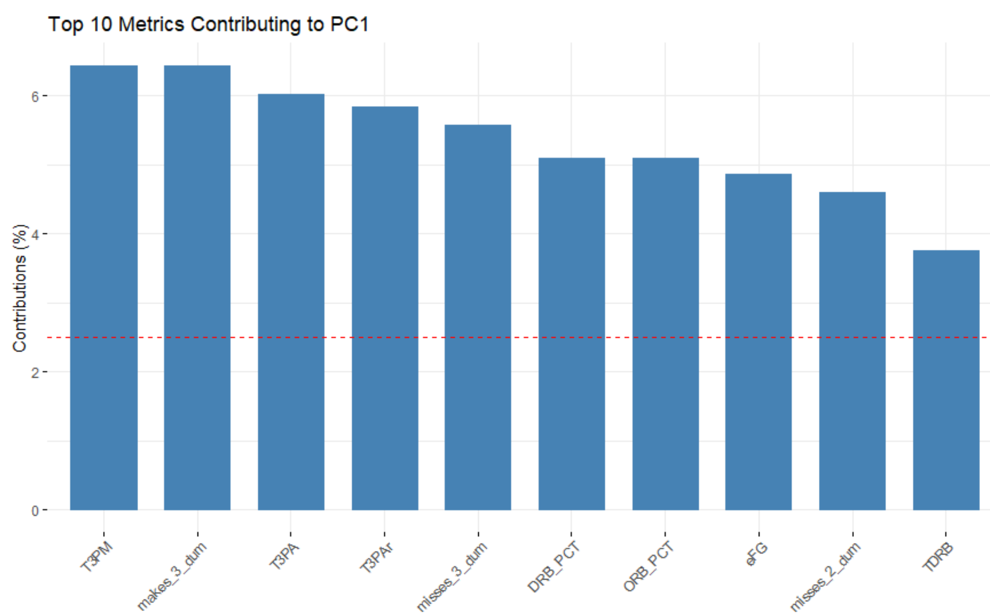


Fig. 35

As shown by Fig. 35 above, PC1 has high loadings on 3-point related stats. This suggests PC1 differentiates teams with a perimeter-oriented efficient offense (high 3PT shooting) compared to those that might rely more on inside play and second-chance points.

We then applied AICC to determine how many factors would be optimal to use for our model. AICC tells us to use 18 principal components to build the model, which can be visually demonstrated in the plot below.

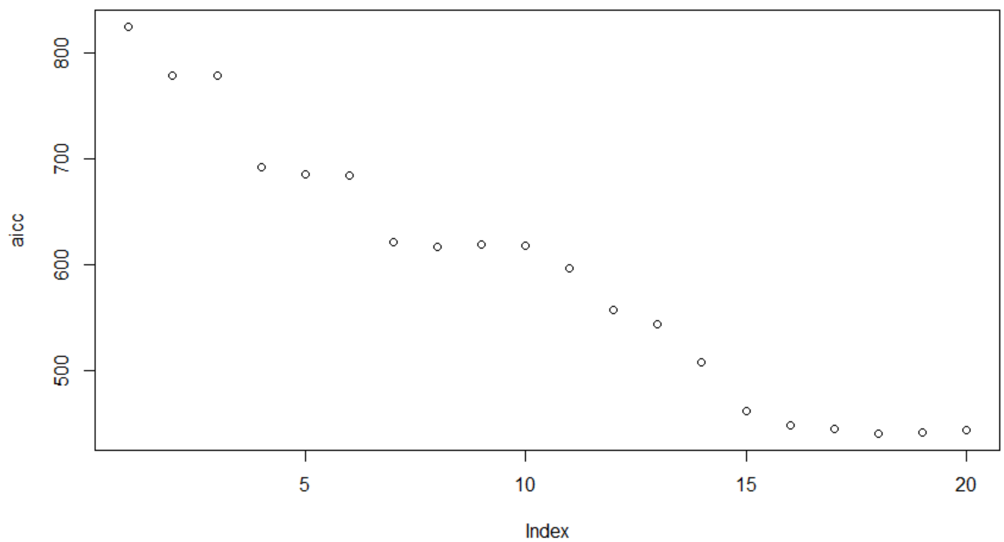


Fig. 36

After constructing the model using our training data and the glm on first K=18 principal components technique, we tested its predictive capabilities on the test data.

```
train_pc <- predict(pca_model, newdata = x_train)
test_pc  <- predict(pca_model, newdata = x_test)
train_pc_df <- as.data.frame(train_pc)
kfits <- lapply(1:32,
                function(K) glm(train_data$MadePlayoffs ~ ., data=train_pc_df[,1:K,drop=FALSE], family='binomial'))
aicc <- sapply(kfits, AICc)
which.min(aicc)
plot(aicc)
summary(MadePlayoffsglm <- glm(train_data$MadePlayoffs ~ ., data=train_pc_df[,1:18]))
pca_preds <- predict(MadePlayoffsglm, newdata = data.frame(test_pc), type="response")
# PCA Evaluation
pca_pred_class <- ifelse(pca_preds > 0.5, 1, 0)
pca_confusion <- confusionMatrix(as.factor(pca_pred_class), as.factor(y_test))
print(pca_confusion)
```

Fig. 37

The standalone PCA logistic regression model achieved a 79% accuracy in predicting out of sample teams to miss or make the playoffs. The results from the confusion matrix below can be interpreted as follows:

- There was only one instance of a false negative where the model incorrectly predicted a team to miss the playoffs. Thus, the false negative rate is 1/24.
- There were 18 instances of the model incorrectly predicting a team to make the playoffs. Thus, the false positive rate is 18/64.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 23  1
         1 18 48

               Accuracy : 0.7889
                 95% CI : (0.6901, 0.8679)
    No Information Rate : 0.5444
    P-Value [Acc > NIR] : 1.206e-06

                  Kappa : 0.5595

 Mcnemar's Test P-Value : 0.0002419
```

Fig. 38

The downside to this PCA approach is interpretability: it’s not always easy to explain to a coach exactly what PC1, PC2, etc. means in practical terms, since each is a mix of many stats. This model is also overly complex given that it uses 18 principal components.

We tried to further refine the model by applying LASSO to the principal components to see which ones will get selected, but did not see much improvement. We do see that the first principal component is selected, but most of the PCs remain selected, which does not improve the issue of complexity and interpretability.

```
PC1      0.028598698
PC2     -0.047492230
PC3      0.001114942
PC4      0.078323817
PC5      0.022066281
PC6     -0.007426301
PC7     -0.092235814
PC8     -0.011609187
PC9      .
PC10    -0.010461653
PC11     0.057787437
PC12    -0.109393843
PC13     0.049339521
PC14    -0.108587268
PC15     0.143177151
PC16     0.076908521
PC17     0.050418234
PC18    -0.035182106
PC19     .
PC20     .
```

Fig. 39

So instead, we decided to ultimately go with a raw LASSO logistic regression model using the original covariates, which will give us an interpretable set of features that are most important. Such simplicity is useful – for instance, a coach/manager could focus on improving a team's three point shooting offense as the primary objective if suggested by the LASSO model, and know that doing so will be the best way for them to improve their chances of making it to the playoffs. The lambda value for the LASSO was chosen using cross validation with nfold=20.

```
# LASSO Model
cv_lasso <- cv.glmnet(as.matrix(x_train), train_data$MadePlayoffs, family="binomial",nfold=20)
best_lambda <- cv_lasso$lambda.min
lasso_model <- glmnet(as.matrix(x_train), train_data$MadePlayoffs, family="binomial", lambda=best_lambda)
# LASSO Predict
lasso_preds <- predict(lasso_model, s=best_lambda, newx=as.matrix(x_test), type="response")
lasso_preds_class <- ifelse(lasso_preds > 0.5, 1, 0)
# LASSO Evaluation
lasso_confusion <- confusionMatrix(as.factor(lasso_preds_class), as.factor(y_test))
print(lasso_confusion)
```

Fig. 40

Overall, the LASSO logistic model was highly predictive and effective. The out of sample results of the raw LASSO are better than the standard PCA model as the accuracy of the model's prediction on the test data was 84%. There were less false positives in the LASSO model (10/55) and slightly more false negatives (4/35). It's worth noting though that some weak-performing teams can sneak into the playoffs due to seeding quirks or from being a part of a weak conference, which may explain the slight increase in false negatives. The in-sample R-Squared for the model was nearly 60%. Out-of-sample R-Squared for the model was also 38.93% meaning the model explained nearly 40% of the variance in the binary outcome of making the playoffs for those seasons.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 31  4
         1 10 45

               Accuracy : 0.8444
                 95% CI : (0.7528, 0.9123)
    No Information Rate : 0.5444
    P-Value [Acc > NIR] : 1.629e-09

                  Kappa : 0.6826

 Mcnemar's Test P-Value : 0.1814
```

Fig. 41

We can also see the exact features that LASSO selected as well as their associated weights:

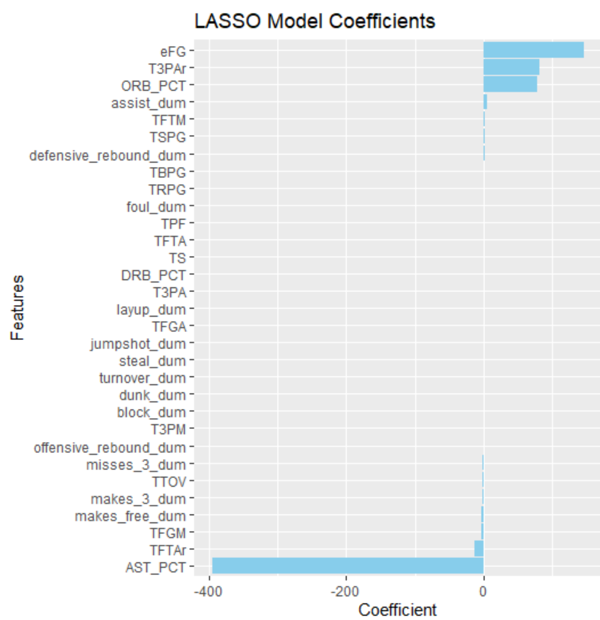


Fig. 42

	feature	coef
1	AST_PCT	-3.945933e+02
2	eFG	1.465254e+02
3	T3PAr	8.212793e+01
4	ORB_PCT	7.934877e+01
5	TFTAr	-1.284606e+01
6	assist_dum	4.612805e+00
7	TFGM	-3.439532e+00
8	TFTM	2.377453e+00
9	makes_free_dum	-2.248761e+00
10	TSPG	1.864653e+00
11	defensive_rebound_dum	1.616365e+00
12	makes_3_dum	-1.218280e+00
13	TTOV	-1.194189e+00
14	misses_3_dum	-8.093133e-01
15	offensive_rebound_dum	-4.020654e-01
16	T3PM	-3.371880e-01
17	block_dum	-3.354984e-01
18	dunk_dum	-3.111871e-01
19	turnover_dum	-3.071264e-01
20	steal_dum	-1.721377e-01
21	TBPG	1.653739e-01
22	TRPG	1.580955e-01
23	foul_dum	1.469159e-01
24	jumpshot_dum	-1.241689e-01
25	TPF	7.980882e-02
26	TFGA	-3.845862e-02
27	layup_dum	-3.531901e-02
28	TFTA	3.163437e-02
29	T3PA	-2.809816e-02
30	TS	1.744808e-02
31	DRB_PCT	-4.682217e-15

Fig. 43

As we can see in Fig. 43, eFG was the variable with the highest positive coefficient, strongly suggesting that teams with a higher eFG% (more efficient shooting) tend to make the playoffs. Thus, high shooting efficiency is critical in successful teams. T3PAr (Three-Point Attempt Rate) was the second-most significant positive predictor, indicating that teams that attempt more three-pointers relative to their total shots are significantly more likely to be playoff-bound. It is clear the NBA highly rewards teams with a playstyle that consists of more perimeter shooting.

AST_PCT has a surprisingly strong negative effect, suggesting that teams overly dependent on assisted field goals (lacking strong isolation scorers or shot-creators) have reduced playoff chances. This suggests teams might benefit from having a hybrid strategy with versatile individual scorers in addition to better playmaking.

Section 3 - Conclusion

The evolution of basketball strategy over the past two decades has been defined by the rise of three-point shooting and increased emphasis on offensive efficiency. Our analysis, spanning over two decades of NBA data, highlights key trends and strategic factors that contribute to postseason success. By integrating play-by-play text analysis, box score statistics, clustering techniques, and machine learning models, we have identified the most important metrics that differentiate successful and unsuccessful teams.

One of the most striking trends in our analysis is the increasing reliance on three-point shooting. Teams have progressively shifted towards perimeter-oriented offenses, as seen in the sharp rise in three-point attempts per game. However, despite the growing emphasis on three-pointers, our results indicate that three-point attempts alone do not guarantee success. Instead, teams that complement their perimeter shooting with strong ball movement, high assist rates, and efficient shot selection tend to have higher winning percentages and a greater likelihood of securing a playoff spot.

Our clustering analysis provided further evidence of this shift, showing that historically, NBA teams exhibited more diverse playstyles, ranging from defensive and rebounding-heavy teams to fast-paced, three-point-oriented offenses. However, in recent years, playstyles have converged, with most successful teams adopting a combination of high three-point volume, efficient shot selection, and strong passing ability. The clustering results also revealed that championship-winning teams disproportionately fall into playstyle clusters associated with offensive efficiency and aggressive three-point shooting.

The marginal regression analysis further solidified these findings by identifying the most significant predictors of playoff success. We found that assists, three-point makes, effective field goal percentage (eFG%), and assist-to-turnover ratio (AST/TOV) were among the strongest indicators of a team's ability to make the playoffs. On the other hand, excessive reliance on two-point field goals was negatively associated with postseason qualification, further supporting the argument that NBA teams must embrace efficient, perimeter-oriented offensive strategies to compete at the highest level.

Our predictive modeling efforts using PCA and LASSO logistic regression provided additional insights. The PCA model confirmed that three-point shooting and efficiency-based metrics were the dominant factors separating successful teams from unsuccessful ones. Meanwhile, the LASSO regression model demonstrated that elite shooting efficiency (eFG%) and effective playmaking (AST/TOV ratio) were the most critical predictors of postseason success. Interestingly, our findings challenge the conventional notion that defensive-oriented teams are more successful, as we observed that modern championship teams are significantly more offensively focused, emphasizing scoring efficiency and spacing rather than traditional defensive schemes.

Taken together, these findings suggest that NBA teams looking to optimize their playoff chances should prioritize efficient three-point shooting, strong ball movement, and disciplined offensive play over traditional, defense-heavy strategies. While defense remains an important factor, the modern NBA rewards teams that can space the floor effectively and generate high-quality scoring opportunities. As the game continues to evolve, teams that fail to adapt to these trends may find themselves struggling to compete against more analytically driven opponents.

In conclusion, the NBA has undergone a strategic transformation over the past two decades, with three-point shooting, offensive efficiency, and playmaking ability emerging as defining characteristics of successful teams. Our research provides a data-driven perspective on these trends, highlighting the key factors that drive playoff success. These insights not only contribute to our understanding of modern basketball strategy but also offer practical implications for NBA teams, coaches, and analysts seeking to optimize team performance in an increasingly competitive league.

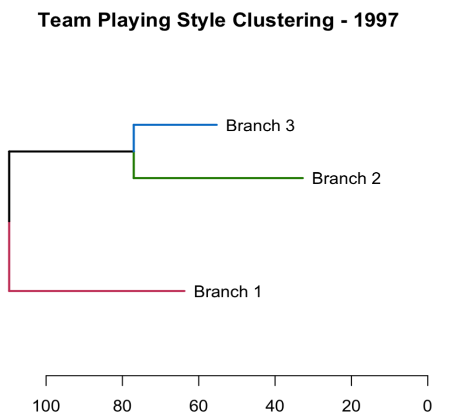
Appendix

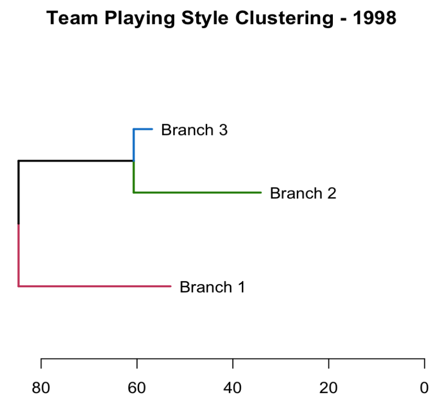
Data Dictionary

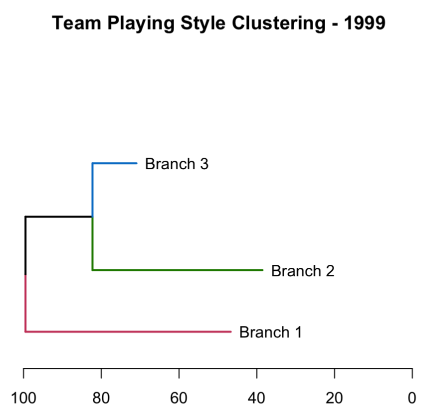
- **game_id**: Unique identifier for each NBA game.
- **teamName**: Name of the team.
- **seasonStartYear**: The starting year of the NBA season (e.g., 1996 for the 1996-97 season).
- **MadePlayoffs**: Binary indicator (1 = Playoffs, 0 = Didn't make Playoffs).
- **homeAway**: Indicator of whether the team was playing at home (homeTeam) or away (awayTeam).
- **assist_dum**: Total number of assists.
- **block_dum**: Total number of blocks.
- **defensive_rebound_dum**: Total number of defensive rebounds.
- **dunk_dum**: Total number of dunks.
- **foul_dum**: Total number of personal fouls.
- **jumpshot_dum**: Total number of jump shots.
- **layup_dum**: Total number of layups made.
- **makes_2_dum**: Total number of **made** two-point field goals.
- **makes_3_dum**: Total number of **made** three-point field goals.
- **makes_free_throw_dum**: Total number of **made** free throws.
- **misses_2_dum**: Total number of **missed** two-point field goals.
- **misses_3_dum**: Total number of **missed** three-point field goals.
- **misses_free_throw_dum**: Total number of **missed** free throws.
- **offensive_rebound_dum**: Total number of offensive rebounds.
- **steal_dum**: Total number of steals.
- **turnover_dum**: Total number of turnovers.
- **PPG (Points Per Game)**: Total points scored by the team in that game.
- **TFGA (Total Field Goals Attempted)**: Total number of shots attempted (both 2-pointers and 3-pointers).
- **TFGM (Total Field Goals Made)**: Total number of successful field goals.
- **T3PA (Three-Point Attempts)**: Total number of three-point shots attempted.
- **T3PM (Three-Point Makes)**: Total number of successful three-point shots.
- **TFTA (Total Free Throws Attempted)**: Number of free throw attempts.
- **TFTM (Total Free Throws Made)**: Number of free throws successfully made.
- **APG (Assists Per Game)**: Total assists made by the team in that game.
- **TTOV (Total Turnovers)**: Number of turnovers committed.
- **TRPG (Total Rebounds Per Game)**: Total rebounds (offensive + defensive).
- **TORB (Total Offensive Rebounds)**: Total number of offensive rebounds.
- **TDRB (Total Defensive Rebounds)**: Total number of defensive rebounds.

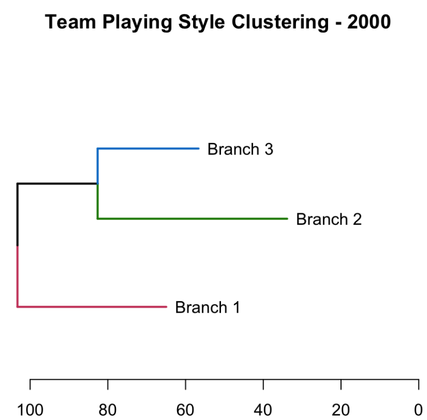
- **TSPG (Total Steals Per Game)**: Number of steals recorded.
- **TBPG (Total Blocks Per Game)**: Number of blocks recorded.
- **TPF (Total Personal Fouls)**: Number of personal fouls committed.
- **PlusMinus**: The **point differential** when the team was on the court. A positive value means the team outscored their opponent while they were playing, while a negative value means they were outscored.
- **T3PAr (Three-Point Attempt Rate)**: The ratio of three-point attempts to total field goal attempts. This measures how much a team relies on three-point shooting.
- **TFTAr (Free Throw Attempt Rate)**: The ratio of free throw attempts to total field goal attempts. This reflects a team's tendency to draw fouls.
- **eFG% (Effective Field Goal Percentage)**: Added value of three-pointers; adjusts field goal percentage to account for the extra point value of three-pointers.
- **TS% (True Shooting Percentage)**: A comprehensive shooting efficiency metric, accounting for three-pointers and free throws. It measures a player's efficiency at scoring.
- **AST_PCT (Assist Percentage)**: The percentage of a team's made field goals that were assisted.
- **AST_TOV (Assist to Turnover Ratio)**: The number of assists per turnover, reflecting playmaking efficiency.
- **ORB_PCT (Offensive Rebound Percentage)**: The percentage of available offensive rebounds a team secures.
- **DRB_PCT (Defensive Rebound Percentage)**: The percentage of available defensive rebounds a team secures.

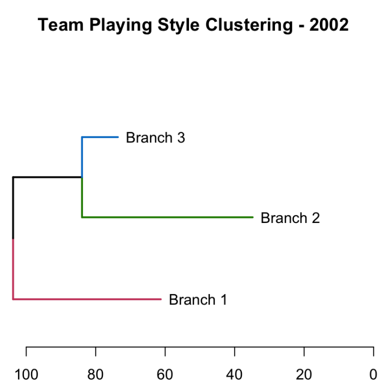
Team Playing Style Clusters

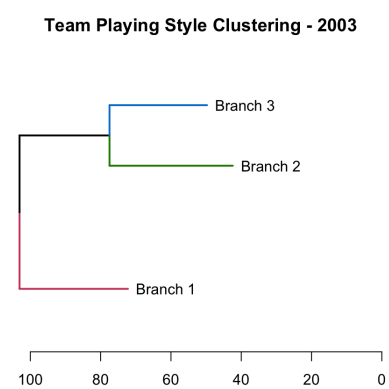












NBA Champions Cluster Breakdown

Cluster Assignments of Championship Teams (1996-2019)

Season	Champion Team	Cluster	Appearances in Cluster	Total Appearances	Percentage Time in Cluster	Playstyle
1996	Chicago Bulls	2	58	101	57.4	High-Scoring, Defensive & Strong Rebounding, Defensive-Oriented, Elite Rebounding Team, Rebounding-Focused
1996	Chicago Bulls	1	31	101	30.7	Three-Point Heavy
1996	Chicago Bulls	3	12	101	11.9	Free-Throw Heavy & Inside Attack, Fast-Paced & Playmaking
1997	Chicago Bulls	2	65	103	63.1	Balanced Playstyle
1997	Chicago Bulls	3	27	103	26.2	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused
1997	Chicago Bulls	1	11	103	10.7	Free-Throw Heavy & Inside Attack
1998	Chicago Bulls	1	25	50	50.0	Defensive & Strong Rebounding, Defensive-Oriented, Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
1998	Chicago Bulls	2	18	50	36.0	Free-Throw Heavy & Inside Attack
1998	Chicago Bulls	3	7	50	14.0	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
1999	San Antonio Spurs	1	44	86	51.2	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
1999	San Antonio Spurs	3	29	86	33.7	Free-Throw Heavy & Inside Attack
1999	San Antonio Spurs	2	13	86	15.1	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive-Oriented, Fast-Paced & Playmaking
2000	Los Angeles Lakers	2	47	98	48.0	Free-Throw Heavy & Inside Attack, Fast-Paced & Playmaking
2000	Los Angeles Lakers	1	31	98	31.6	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2000	Los Angeles Lakers	3	20	98	20.4	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking
2001	Los Angeles Lakers	2	55	101	54.5	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking
2001	Los Angeles Lakers	1	38	101	37.6	Free-Throw Heavy & Inside Attack
2001	Los Angeles Lakers	3	8	101	7.9	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2002	Los Angeles Lakers	1	46	94	48.9	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused
2002	Los Angeles Lakers	3	35	94	37.2	Balanced Playstyle
2002	Los Angeles Lakers	2	13	94	13.8	Free-Throw Heavy & Inside Attack
2003	San Antonio Spurs	1	60	92	65.2	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2003	San Antonio Spurs	2	18	92	19.6	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2003	San Antonio Spurs	3	14	92	15.2	Free-Throw Heavy & Inside Attack, Defensive-Oriented, Fast-Paced & Playmaking

2004	Detroit Pistons	1	81	107	75.7	Balanced Playstyle
2004	Detroit Pistons	3	17	107	15.9	Free-Throw Heavy & Inside Attack, Defensive-Oriented, Fast-Paced & Playmaking
2004	Detroit Pistons	2	9	107	8.4	High-Scoring, Three-Point Heavy, Fast-Paced & Playmaking, Rebounding-Focused
2005	San Antonio Spurs	3	36	95	37.9	Balanced Playstyle
2005	San Antonio Spurs	2	30	95	31.6	Elite Scoring & Efficient, High-Scoring, Free-Throw Heavy & Inside Attack, Defensive-Oriented, Fast-Paced & Playmaking
2005	San Antonio Spurs	1	29	95	30.5	Three-Point Heavy, Fast-Paced & Playmaking, Elite Rebounding Team, Rebounding-Focused
2006	Miami Heat	1	43	86	50.0	Free-Throw Heavy & Inside Attack, Fast-Paced & Playmaking
2006	Miami Heat	3	23	86	26.7	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2006	Miami Heat	2	20	86	23.3	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking
2007	San Antonio Spurs	3	36	99	36.4	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused
2007	San Antonio Spurs	1	32	99	32.3	Free-Throw Heavy & Inside Attack, Fast-Paced & Playmaking
2007	San Antonio Spurs	2	31	99	31.3	Balanced Playstyle
2008	Boston Celtics	3	42	96	43.8	Fast-Paced & Playmaking
2008	Boston Celtics	1	37	96	38.5	Balanced Playstyle
2008	Boston Celtics	2	17	96	17.7	High-Scoring, Three-Point Heavy, Free-Throw Heavy & Inside Attack, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused
2009	Los Angeles Lakers	1	51	105	48.6	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2009	Los Angeles Lakers	3	35	105	33.3	Defensive & Strong Rebounding, Defensive-Oriented, Elite Rebounding Team, Rebounding-Focused
2009	Los Angeles Lakers	2	19	105	18.1	Free-Throw Heavy & Inside Attack
2010	Los Angeles Lakers	1	47	92	51.1	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2010	Los Angeles Lakers	2	30	92	32.6	Free-Throw Heavy & Inside Attack, Defensive-Oriented
2010	Los Angeles Lakers	3	15	92	16.3	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2011	Dallas Mavericks	2	29	70	41.4	Free-Throw Heavy & Inside Attack, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused
2011	Dallas Mavericks	3	22	70	31.4	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2011	Dallas Mavericks	1	19	70	27.1	Balanced Playstyle
2012	Miami Heat	2	51	105	48.6	Fast-Paced & Playmaking
2012	Miami Heat	1	40	105	38.1	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking
2012	Miami Heat	3	14	105	13.3	Free-Throw Heavy & Inside Attack, Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused

2013	Miami Heat	2	72	102	70.6	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Free-Throw Heavy & Inside Attack, Defensive-Oriented, Fast-Paced & Playmaking
2013	Miami Heat	1	22	102	21.6	Fast-Paced & Playmaking, Rebounding-Focused
2013	Miami Heat	3	8	102	7.8	Balanced Playstyle
2014	San Antonio Spurs	3	51	89	57.3	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive-Oriented, Fast-Paced & Playmaking
2014	San Antonio Spurs	2	24	89	27.0	Free-Throw Heavy & Inside Attack
2014	San Antonio Spurs	1	14	89	15.7	Fast-Paced & Playmaking, Elite Rebounding Team, Rebounding-Focused
2015	Golden State Warriors	2	81	106	76.4	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2015	Golden State Warriors	3	14	106	13.2	Free-Throw Heavy & Inside Attack, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking, Rebounding-Focused
2015	Golden State Warriors	1	11	106	10.4	Balanced Playstyle
2016	Cleveland Cavaliers	3	36	100	36.0	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive & Strong Rebounding, Defensive-Oriented, Fast-Paced & Playmaking
2016	Cleveland Cavaliers	1	33	100	33.0	Free-Throw Heavy & Inside Attack
2016	Cleveland Cavaliers	2	31	100	31.0	Slow-Paced & Defensive, Elite Rebounding Team, Rebounding-Focused
2017	Golden State Warriors	1	63	103	61.2	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking, Rebounding-Focused
2017	Golden State Warriors	2	23	103	22.3	Balanced Playstyle
2017	Golden State Warriors	3	17	103	16.5	Free-Throw Heavy & Inside Attack, Defensive-Oriented
2018	Golden State Warriors	3	72	104	69.2	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Defensive-Oriented, Fast-Paced & Playmaking
2018	Golden State Warriors	1	16	104	15.4	Fast-Paced & Playmaking, Elite Rebounding Team, Rebounding-Focused
2018	Golden State Warriors	2	16	104	15.4	Free-Throw Heavy & Inside Attack
2019	Toronto Raptors	1	48	83	57.8	Elite Scoring & Efficient, High-Scoring, Three-Point Heavy & Efficient, Three-Point Heavy, Fast-Paced & Playmaking
2019	Toronto Raptors	3	18	83	21.7	Free-Throw Heavy & Inside Attack
2019	Toronto Raptors	2	17	83	20.5	Defensive & Strong Rebounding, Defensive-Oriented, Elite Rebounding Team, Rebounding-Focused

Fig. 27