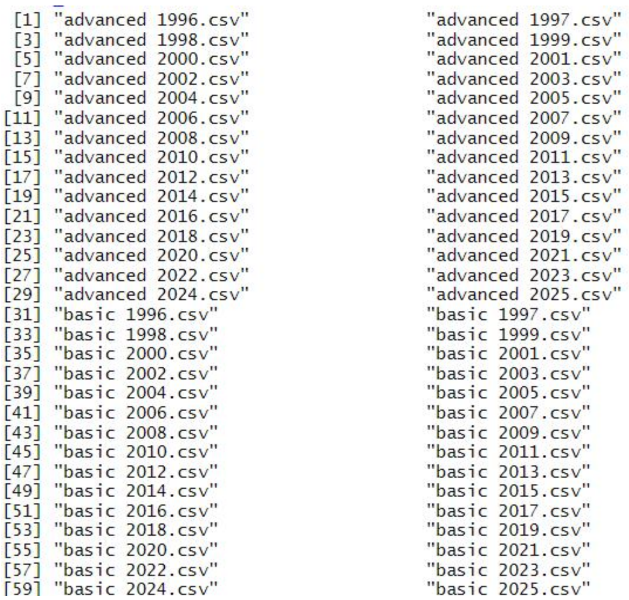
NBA Team Playstyle Evolution

1.Abstract

This study explores how NBA playstyles have changed over the past thirty years. The National Basketball Association has experienced significant evolution over the past three decades. Changes in players' ability, roles, and playing logic had greatly changed the way teams play the game. This study aims to explore how NBA team play styles have changed across the last 30 years using historical data on both basic and advanced NBA statistics. The study was divided into three stages: data cleaning and preprocessing, exploratory data analysis (EDA), and machine learning. After exploratory data analysis, the study applies PCA combined with changepoint detection to segment NBA data from the past three decades into three distinct eras without disrupting temporal continuity. A random forest model is then used to analyze two aspects: whether random forest can successfully classify the data into eras under this context, and whether it can identify the most influential variables for era segmentation through the Gini index. The final result shows that the random forest model successfully classified the data and identified three-point-related variables as the most influential indicators of era differentiation.

2. Data Cleaning and Preprocessing

2.1 Dataset Sources and Structure



```
 [1] "advanced 1996.csv"           "advanced 1997.csv"
 [3] "advanced 1998.csv"           "advanced 1999.csv"
 [5] "advanced 2000.csv"           "advanced 2001.csv"
 [7] "advanced 2002.csv"           "advanced 2003.csv"
 [9] "advanced 2004.csv"           "advanced 2005.csv"
[11] "advanced 2006.csv"           "advanced 2007.csv"
[13] "advanced 2008.csv"           "advanced 2009.csv"
[15] "advanced 2010.csv"           "advanced 2011.csv"
[17] "advanced 2012.csv"           "advanced 2013.csv"
[19] "advanced 2014.csv"           "advanced 2015.csv"
[21] "advanced 2016.csv"           "advanced 2017.csv"
[23] "advanced 2018.csv"           "advanced 2019.csv"
[25] "advanced 2020.csv"           "advanced 2021.csv"
[27] "advanced 2022.csv"           "advanced 2023.csv"
[29] "advanced 2024.csv"           "advanced 2025.csv"
[31] "basic 1996.csv"              "basic 1997.csv"
[33] "basic 1998.csv"              "basic 1999.csv"
[35] "basic 2000.csv"              "basic 2001.csv"
[37] "basic 2002.csv"              "basic 2003.csv"
[39] "basic 2004.csv"              "basic 2005.csv"
[41] "basic 2006.csv"              "basic 2007.csv"
[43] "basic 2008.csv"              "basic 2009.csv"
[45] "basic 2010.csv"              "basic 2011.csv"
[47] "basic 2012.csv"              "basic 2013.csv"
[49] "basic 2014.csv"              "basic 2015.csv"
[51] "basic 2016.csv"              "basic 2017.csv"
[53] "basic 2018.csv"              "basic 2019.csv"
[55] "basic 2020.csv"              "basic 2021.csv"
[57] "basic 2022.csv"              "basic 2023.csv"
[59] "basic 2024.csv"              "basic 2025.csv"
```

Figure 1. List of raw csv files

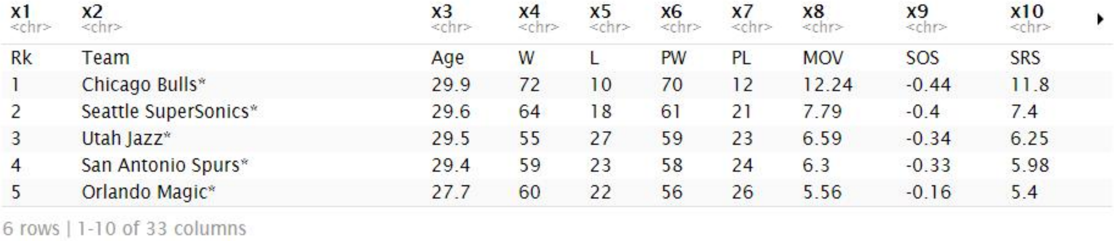
The original data were downloaded from www.basketball-reference.com. There were 60 original files in total, each season's data resulted in two files: one for basic stats and one for advanced stats. Each file contained the statistics for all 30 teams in certain season, along with an additional row for the league average stats for that season. These files were named as shown in Figure 1, season year was contained in each of them.

2.2 File Consolidation and Year Extraction

The first step in preprocessing involved reading all the files and tagging each with the corresponding year. This was achieved by extracting the 4-digit year embedded in the filename and adding it as a new column in each dataset. This ensured that each row of data could be accurately matched to its season during the merging process. This is extremely important because later the data will be merged by years.

2.3 Data Cleaning and Standardization

First, two small datasets are created, advanced data which contains all advanced data and one basic data which contains all basic data. Several steps were taken to standardize and clean these data. The column names were changed to the correct ones. Irrelevant columns like attendance, arena and na columns were removed. Then rows incorrectly containing column names were fixed and removed by filtering out any rows where the 'team' column equaled 'Team'. Asterisks appended to team names were also removed to ensure consistency in merging. Column names were standardized using the clean_names() function, which replaced characters with underscores and converted all names to lowercase. However, some automatically generated column names (e_fg_percent_2 columns) were not clear and were manually changed to comprehensive names we want.



x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
Rk	Team	Age	W	L	PW	PL	MOV	SOS	SRS
1	Chicago Bulls*	29.9	72	10	70	12	12.24	-0.44	11.8
2	Seattle SuperSonics*	29.6	64	18	61	21	7.79	-0.4	7.4
3	Utah Jazz*	29.5	55	27	59	23	6.59	-0.34	6.25
4	San Antonio Spurs*	29.4	59	23	58	24	6.3	-0.33	5.98
5	Orlando Magic*	27.7	60	22	56	26	5.56	-0.16	5.4

6 rows | 1-10 of 33 columns

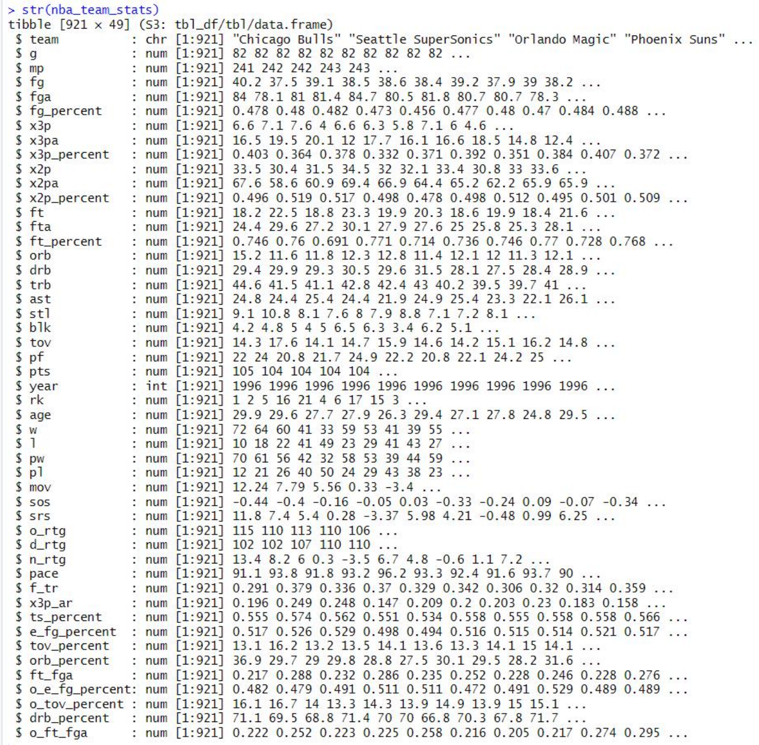
Figure 2. Raw Sample of Advanced Team Statistics

2.4 Dataset Merging

The cleaned basic and advanced datasets were merged into a single dataset using a left join on the columns `team` and `year`. Following the join, duplicated columns (e.g., rk.x and rk.y) were reviewed. `rk.x` was being dropped and `rk.y` was renamed to `rk`.

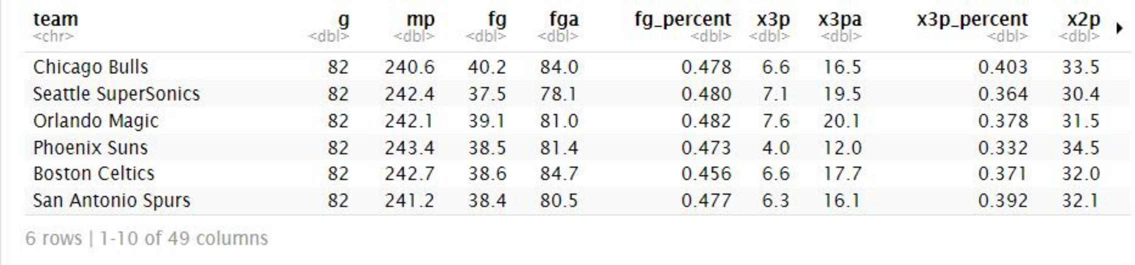
2.5 Final Cleanup and Dataset Export

All columns except for `team` and `year` were then converted to numeric format, previously some of them were stored as character. This final merged dataset was saved as `nba_team_stats_merged_cleaned.csv`. The final dataset comprised 921 rows and 49 variables and served as the input for all subsequent visualizations and modeling procedures.



```
> str(nba_team_stats)
tibble [921 × 49] (S3: tbl_df/tbl/data.frame)
 $ team          : chr [1:921] "Chicago Bulls" "Seattle SuperSonics" "Orlando Magic" "Phoenix Suns" ...
 $ g             : num [1:921] 82 82 82 82 82 82 82 82 82 82 ...
 $ mp            : num [1:921] 241 242 242 243 243 ...
 $ fg            : num [1:921] 40.2 37.5 39.1 38.5 38.6 38.4 39.2 37.9 39 38.2 ...
 $ fga           : num [1:921] 84 78.1 81 81.4 84.7 80.5 81.8 80.7 80.7 78.3 ...
 $ fg_percent    : num [1:921] 0.478 0.48 0.482 0.473 0.456 0.477 0.48 0.47 0.484 0.488 ...
 $ x3p           : num [1:921] 6.6 7.1 7.6 4 6.6 6.3 5.8 7.1 6 4.6 ...
 $ x3pa          : num [1:921] 16.5 19.5 20.1 12 17.7 16.1 16.6 18.5 14.8 12.4 ...
 $ x3p_percent   : num [1:921] 0.403 0.364 0.378 0.332 0.371 0.392 0.351 0.384 0.407 0.372 ...
 $ x2p           : num [1:921] 33.5 30.4 31.5 34.5 32 32.1 33.4 30.8 33 33.6 ...
 $ x2pa          : num [1:921] 67.6 58.6 60.9 69.4 66.9 64.4 65.2 62.2 65.9 65.9 ...
 $ x2p_percent   : num [1:921] 0.496 0.519 0.517 0.498 0.478 0.498 0.512 0.495 0.501 0.509 ...
 $ ft            : num [1:921] 18.2 22.5 18.8 23.3 19.9 20.3 18.6 19.9 18.4 21.6 ...
 $ fta           : num [1:921] 24.4 29.6 27.2 30.1 27.9 27.6 25 25.8 25.3 28.1 ...
 $ ft_percent    : num [1:921] 0.746 0.76 0.691 0.771 0.714 0.736 0.746 0.77 0.728 0.768 ...
 $ orb           : num [1:921] 15.2 11.6 11.8 12.3 12.8 11.4 12.1 12 11.3 12.1 ...
 $ drb           : num [1:921] 29.4 29.9 29.3 30.5 29.6 31.5 28.1 27.5 28.4 28.9 ...
 $ trb           : num [1:921] 44.6 41.5 41.1 42.8 42.4 43 40.2 39.5 39.7 41 ...
 $ ast           : num [1:921] 24.8 24.4 25.4 24.4 21.9 24.9 25.4 23.3 22.1 26.1 ...
 $ stl           : num [1:921] 9.1 10.8 8.1 7.6 8 7.9 8.8 7.1 7.2 8.1 ...
 $ blk           : num [1:921] 4.2 4.8 5 4 5 6.5 6.3 3.4 6.2 5.1 ...
 $ tov           : num [1:921] 14.3 17.6 14.1 14.7 15.9 14.6 14.2 15.1 16.2 14.8 ...
 $ pf            : num [1:921] 22 24 20.8 21.7 24.9 22.2 20.8 22.1 24.2 25 ...
 $ pts           : num [1:921] 105 104 104 104 104 ...
 $ year          : int [1:921] 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 ...
 $ rk            : num [1:921] 1 2 5 16 21 4 6 17 15 3 ...
 $ age           : num [1:921] 29.9 29.6 27.7 27.9 26.3 29.4 27.1 27.8 24.8 29.5 ...
 $ w             : num [1:921] 72 64 60 41 33 59 53 41 39 55 ...
 $ l             : num [1:921] 10 18 22 41 49 23 29 41 43 27 ...
 $ pw            : num [1:921] 70 61 56 42 32 58 53 39 44 59 ...
 $ pl            : num [1:921] 12 21 26 40 50 24 29 43 38 23 ...
 $ mov           : num [1:921] 12.24 7.79 5.56 0.33 -3.4 ...
 $ sos           : num [1:921] -0.44 -0.4 -0.16 -0.05 0.03 -0.33 -0.24 0.09 -0.07 -0.34 ...
 $ srs           : num [1:921] 11.8 7.4 5.4 0.28 -3.37 5.98 4.21 -0.48 0.99 6.25 ...
 $ o_rtg         : num [1:921] 115 110 113 110 106 ...
 $ d_rtg         : num [1:921] 102 102 107 110 110 ...
 $ n_rtg         : num [1:921] 13.4 8.2 6 0.3 -3.5 6.7 4.8 -0.6 1.1 7.2 ...
 $ pace          : num [1:921] 91.1 93.8 91.8 93.2 96.2 93.3 92.4 91.6 93.7 90 ...
 $ f_tr          : num [1:921] 0.291 0.379 0.336 0.37 0.329 0.342 0.306 0.32 0.314 0.359 ...
 $ x3p_ar        : num [1:921] 0.196 0.249 0.248 0.147 0.209 0.2 0.203 0.23 0.183 0.158 ...
 $ ts_percent    : num [1:921] 0.555 0.574 0.562 0.551 0.534 0.558 0.555 0.558 0.558 0.566 ...
 $ e_fg_percent  : num [1:921] 0.517 0.526 0.529 0.498 0.494 0.516 0.515 0.514 0.521 0.517 ...
 $ tov_percent   : num [1:921] 13.1 16.2 13.2 13.5 14.1 13.6 13.3 14.1 15 14.1 ...
 $ orb_percent   : num [1:921] 36.9 29.7 29 29.8 28.8 27.5 30.1 29.5 28.2 31.6 ...
 $ ft_fga        : num [1:921] 0.217 0.288 0.232 0.286 0.235 0.252 0.228 0.246 0.228 0.276 ...
 $ o_e_fg_percent: num [1:921] 0.482 0.479 0.491 0.511 0.511 0.472 0.491 0.529 0.489 0.489 ...
 $ o_tov_percent : num [1:921] 16.1 16.7 14 13.3 14.3 13.9 14.9 13.9 15 15.1 ...
 $ drb_percent   : num [1:921] 71.1 69.5 68.8 71.4 70 70 66.8 70.3 67.8 71.7 ...
 $ o_ft_fga      : num [1:921] 0.222 0.252 0.223 0.225 0.258 0.216 0.205 0.217 0.274 0.295 ...
```

Figure 3. Data Structure of Final nba_team_stats Dataset



team	g	mp	fg	fga	fg_percent	x3p	x3pa	x3p_percent	x2p
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Chicago Bulls	82	240.6	40.2	84.0	0.478	6.6	16.5	0.403	33.5
Seattle SuperSonics	82	242.4	37.5	78.1	0.480	7.1	19.5	0.364	30.4
Orlando Magic	82	242.1	39.1	81.0	0.482	7.6	20.1	0.378	31.5
Phoenix Suns	82	243.4	38.5	81.4	0.473	4.0	12.0	0.332	34.5
Boston Celtics	82	242.7	38.6	84.7	0.456	6.6	17.7	0.371	32.0
San Antonio Spurs	82	241.2	38.4	80.5	0.477	6.3	16.1	0.392	32.1

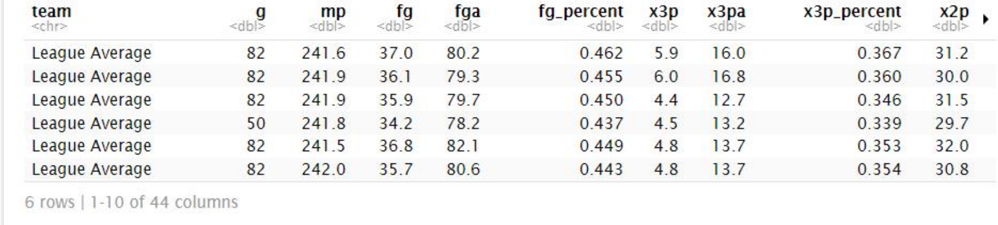
6 rows | 1-10 of 49 columns

Figure 4. Head of Final nba_team_stats Dataset

3. Exploratory Data Analysis (EDA)

3.1 League Average Data Extraction and Preliminary Line Graph Analysis

This part began by extracting rows marked as 'League Average' from the dataset. This subset of data provided a year-by-year snapshot of league-wide trends and was especially useful for identifying macro-level changes in NBA playstyles. Then a line graph was made on each variable to see which several of them are changing the most.



team	g	mp	fg	fga	fg_percent	x3p	x3pa	x3p_percent	x2p
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
League Average	82	241.6	37.0	80.2	0.462	5.9	16.0	0.367	31.2
League Average	82	241.9	36.1	79.3	0.455	6.0	16.8	0.360	30.0
League Average	82	241.9	35.9	79.7	0.450	4.4	12.7	0.346	31.5
League Average	50	241.8	34.2	78.2	0.437	4.5	13.2	0.339	29.7
League Average	82	241.5	36.8	82.1	0.449	4.8	13.7	0.353	32.0
League Average	82	242.0	35.7	80.6	0.443	4.8	13.7	0.354	30.8

6 rows | 1-10 of 44 columns

Figure 5. Head of League Average Dataset

Figure 6. League Average Stat Trends (all variables)

3.2 Changes in Shot Selection

Three-point shooting has seen one of the most dramatic changes. The variables `x3p` (3-point field goals made), `x3pa` (3-point attempts), and `x3p_percent` (3p filed goal percent) were examined over time. Visualization revealed that while both makes and attempts have increased significantly, the shooting percentage has remained relatively flat. This suggests that the increasing three-point shooting was not necessarily due to increased player accuracy but because of its higher expected efficiency.

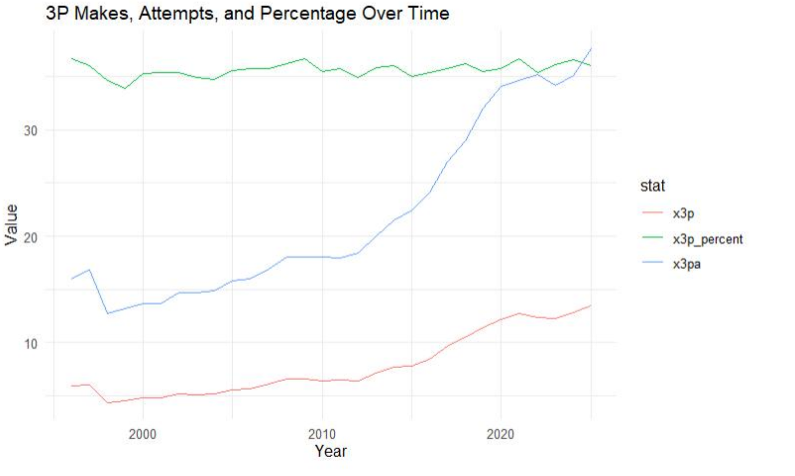


Figure 7. Three-Point Shooting Trends

In parallel with the increase in three-point activity, there has been a decrease in mid-range shots, which was also called inefficient two-point attempts. This conclusion was made because both `x2pa` (2-point attempts) and `x2p` (2-point field goals made) went down while`x2p_percent` (2p filed goal percent) kept going up. This indicated that players were more willing to shoot either highly reliable two-pointers or three-pointers. Effective field goal percentage has risen during the same period, indicating such a shift led us to more efficient basketball.

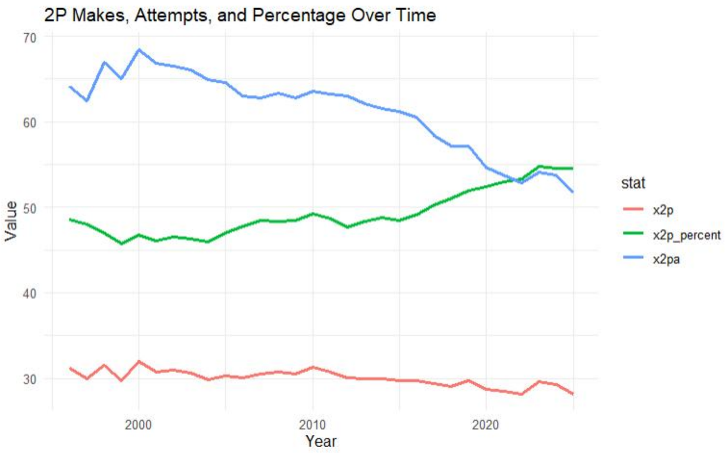


Figure 8. Two-Point Shooting Trends

Another interesting thing worth noticing was that changes in shot selection could also have an influence on rebounding dynamics. When offensive players took more three-point shots, they were positioned farther from the basket, which meant they did not have an advantageous position for offensive rebounds. As a result, the offensive rebounding rate decreased with the rise in three-point attempts, while the defensive rebounding rate increased.

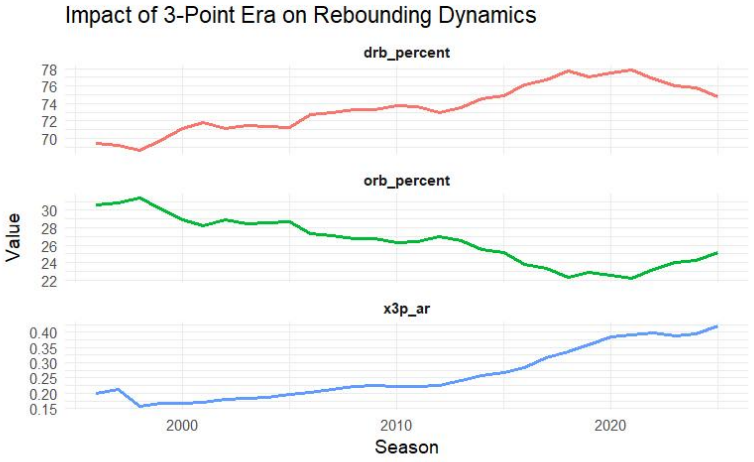


Figure 9. Impact of 3-Point Era on Rebounding Dynamics

3.3 The Fast Pace Era (2013–2020)

Variables including `pace`, `fga`, `trb`, `ast`, and `pts` (pace related) were analyzed to understand the tempo of play. Over the past three decades, these statistics showed a general upward trend. However, the period from 2013 to 2020 appeared the most obvious and accelerated increase.

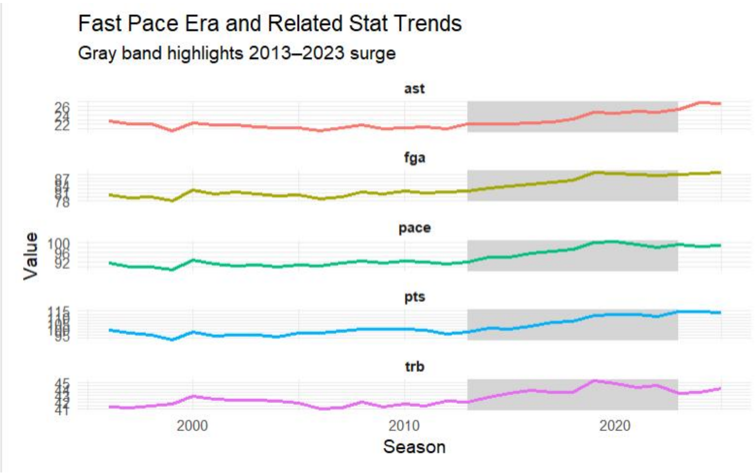


Figure 10. Trends in Key Pace Related Stats (2013–2023 highlighted)

Visualizations supported by regression slope analysis confirmed that this period marked a distinct 'Fast Pace Era' in the NBA. This trend is consistent with the league's shift toward quicker possessions, more shot attempts, and higher overall scoring.

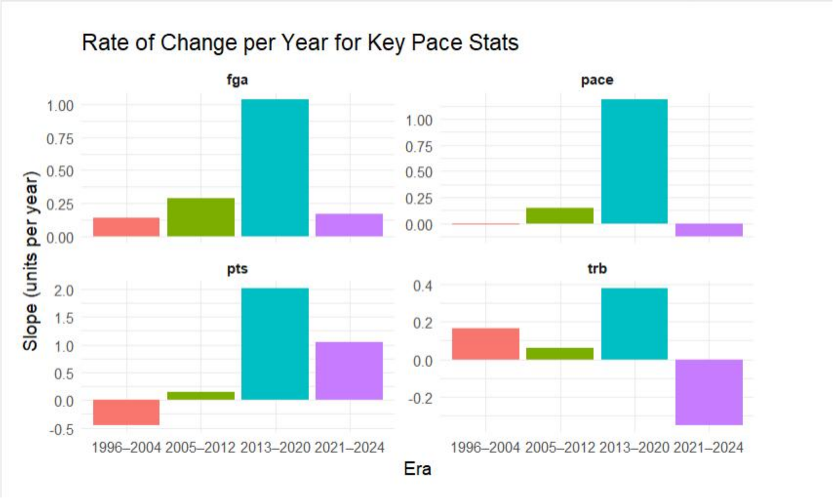


Figure 11. Rate of Change per Year for Key Pace Stats

3.4 Post-2004 Offensive Boost and Rule Influence

Starting from the 2004–2005 season, a sudden and noticeable increase was observed in several key offensive variables. As shown in the visualization, with a red dashed line highlighting the 2004–2005 season, variables such as field goal percentage (fg_percent), free throw attempts (fta), free throw rate (f_tr), and offensive rating (o_rtg) all experienced significant upward shifts.

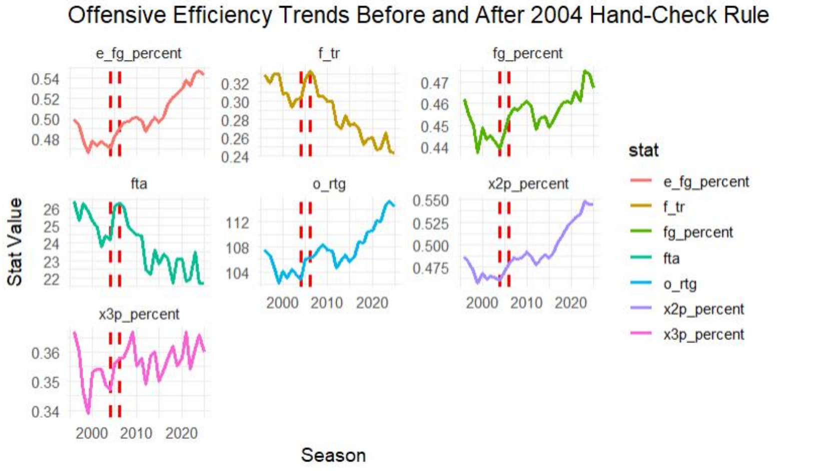


Figure 12. Offensive efficiency trends before and after 2004 hand-check rule change

These variables are all closely associated with a team's offensive capabilities, suggesting a league-wide enhancement in scoring efficiency and offensive ability.

The underlying cause of this abrupt change can be attributed to the introduction of the hand-checking rule in 2004, which limited defenders' ability to apply physical pressure (mainly hand contact) to offensive players. By reducing defensive restrictions, the rule change encouraged offense greatly in the league, leading to immediate improvements in teams' scoring abilities.

4. Data Overview before Modeling

Before proceeding with model analysis, the dataset is first reviewed. This study consists of 921 rows of data and 49 variables (excluding the "team" variable). Based on this dataset, five variables deemed unrelated to playstyle or era segmentation—namely, year, win, loss, Pythagorean win, and Pythagorean loss—are removed, resulting in a total of 44 valid variables.

A specific note is warranted regarding the average-value data: among the 921 rows, 30 correspond to annual averages for each variable over the 30-year period. Since some of these average rows contain missing values (NA) in certain variables, the study recalculates the yearly averages at the beginning of the PCA process to improve data quality. As a result, the PCA is conducted on a dataset of 30 effective observations, which determines the number of principal components (i.e., the matrix rank is 30).

5. Era Segmentation by PCA and Changepoint Detection

5.1 Selection of PCA and Changepoint Detection

Prior to conducting the initial random forest analysis, a critical preparatory step involves determining how each year should be assigned to a specific era—namely, which years belong to Era 1, Era 2, and Era 3. At the outset, this study considered three different ideas for defining these eras:

Idea 1: Divide the timeline into fixed decades, with each 10-year period representing an era: Era 1: 1996–2005, Era 2: 2006–2015, Era 3: 2016–2025.
Idea 2: Use K-means clustering to assign each year to a specific era based on playstyle features.
Idea 3: Apply a combination of PCA and changepoint detection to define three eras in a data-driven manner.

It is clear that the first idea has a significant flaw. Dividing eras solely based on fixed time intervals has no inherent relationship with actual playstyle characteristics. As a result, this form of segmentation is not derived from real changes in playing style but rather from an arbitrarily imposed time structure. Therefore, Idea 1 was quickly ruled out.

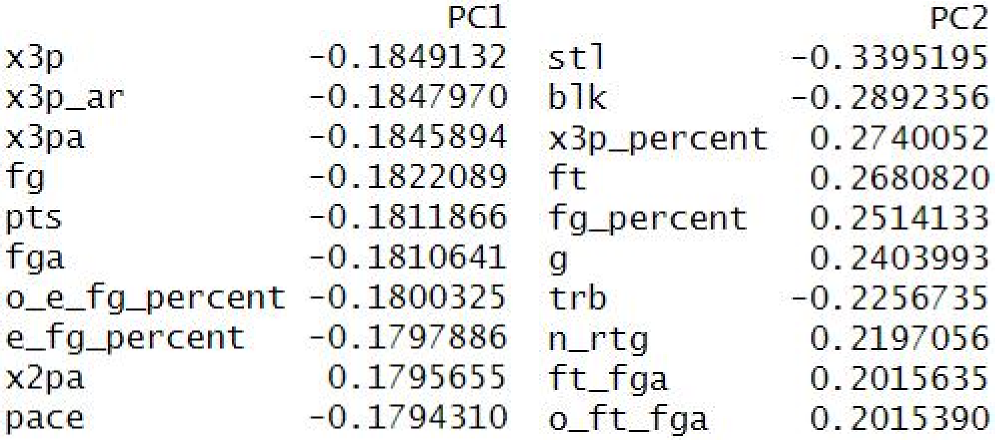
At first glance, the second idea—using K-means clustering to assign years to different eras based on playstyle features—appeared reasonable. Conceptually, it made sense to let playstyle similarity guide the grouping of years. However, during implementation, a critical but easily overlooked issue emerged: the resulting era segmentation was not temporally continuous. For example, K-means might cluster 1996 and 2024 into Era 1, while assigning 2023 to Era 3. This kind of discontinuity contradicts the very research question of this study—how NBA playstyles have evolved over the past thirty years. There is little doubt that 2024 belongs in Era 3; if K-means assigns it to Era 1, the entire segmentation becomes temporally inconsistent, making the research framework untenable.

As a result, this study ultimately adopts Idea 3: using PCA combined with changepoint detection. First, PCA is used to extract the principal components, and PC1 is selected as the most representative component of playstyle variation. Then, changepoint detection is applied to the PC1 scores across years to identify the two most significant change points using the changepoint package in R (note: this paper does not delve into the mathematical formulation of changepoint detection). Based on these two breakpoints, the years 1996–2025 are divided into three eras.

This approach offers two clear advantages. First, it is grounded in PC1, which aggregates rich playstyle information from multiple variables. Second, and more importantly, the resulting era segmentation is continuous over time, aligning naturally with the temporal nature of the research question.

5.2 Implementation of PCA and Changepoint Detection

The study first computes the mean of each variable across all teams within a given year, resulting in a dataset of 30 rows—each representing the average playstyle profile for a single year over the 30-year period. These average values are then standardized to ensure the stability and robustness of the subsequent PCA.



```
                       PC1                       PC2
x3p             -0.1849132  stl           -0.3395195
x3p_ar          -0.1847970  blk           -0.2892356
x3pa            -0.1845894  x3p_percent    0.2740052
fg              -0.1822089  ft             0.2680820
pts             -0.1811866  fg_percent     0.2514133
fga             -0.1810641  g              0.2403993
o_e_fg_percent  -0.1800325  trb           -0.2256735
e_fg_percent    -0.1797886  n_rtg          0.2197056
x2pa             0.1795655  ft_fga         0.2015635
pace            -0.1794310  o_ft_fga       0.2015390
```

Figure 13. Variable Loadings on PC1 and PC2 from PCA Analysis.

The dimensionality of the dataset is ultimately reduced from 30 variables to 2 principal components. The two figures above display the coefficients (loadings) of each variable on PC1 and PC2, respectively. However, since PCA is used in this study solely as a tool for era segmentation rather than for in-depth interpretation, only the scores of the most important principal component, PC1, are utilized. Therefore, PC2 is not further analyzed in this report.

The formula for the PC1 score is as follows:

$PC1_i = \sum w_j \times x_{ij}$

where w_j represents the loading (weight) of the j-th variable in PC1, and x_{ij} is the standardized value of the j-th variable in the i-th year.

Even at this stage, it becomes evident that variables related to three-point shooting are a key driver of changes in NBA playstyle—an insight that will be further reinforced in the Random Forest analysis. It is important to note that PCA is used here primarily for the purpose of defining eras. In other words, the first principal component (PC1) can be interpreted as representing a playstyle characterized by:
more three-point attempts + higher scoring + more field goal attempts + faster pace.

After obtaining the PC1 scores for each year, the changepoint package in R is used to detect the years in which the most significant shifts occur in PC1 values. (The mathematical formulation of changepoint detection is beyond the scope of this report.) The resulting breakpoints are shown in the figure below:

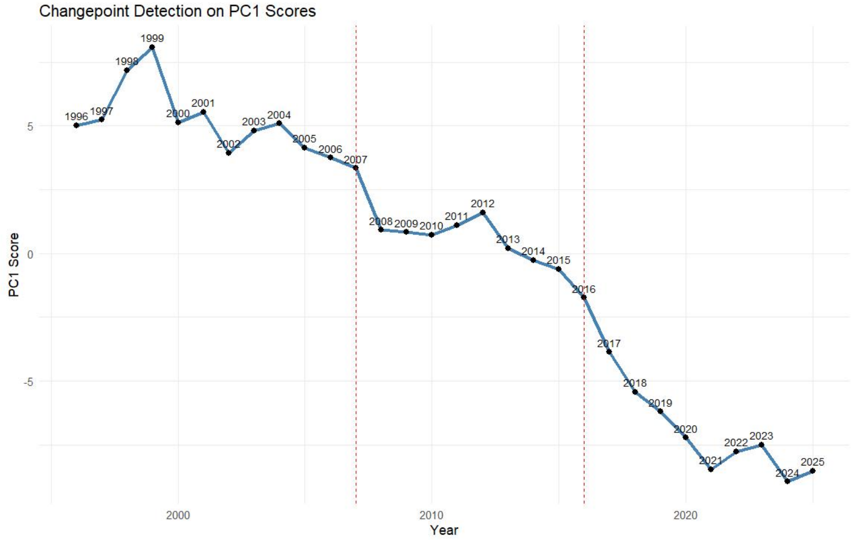


Figure 14. Changes in PC1 Scores Across Years, representing variation in playstyle along the direction of the first principal component.

As shown in the figure, R Studio identifies 2007 and 2016 as the changepoints, indicating that significant shifts in NBA playstyle occurred after these two years. Accordingly, this study defines the era segmentation for the 1996–2025 period as follows:

Era 1: 1996–2007
Era 2: 2008–2016
Era 3: 2017–2025

Notably, this segmentation is consistent with the trends identified in the exploratory data analysis (EDA), suggesting that the changepoint-based division is both statistically and intuitively sound.

For better interpretation of the PCA combined with changepoint detection, this study further presents a figure that overlays the era segmentation with the PCA results.

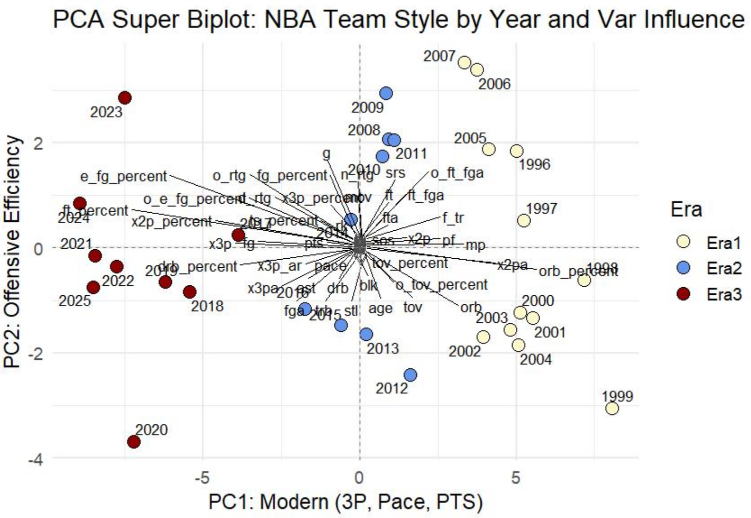


Figure 15. PCA Biplot

As shown in the figure, aided by PCA, the year points corresponding to each era cluster along the PC1 axis. A brief explanation of the position of these points is provided below:

Era 1 year points are concentrated on the right side of the PC1 axis. This is because the original values of modern playstyle indicators during these years were generally below the overall mean. After standardization, these lower-than-average values translated into negative z-scores. Given that most of the PC1 loadings are negative (as shown in Figure X), the product of a negative z-score and a negative loading yields a positive PC1 score, which places Era 1 years on the right side of the graph.

Era 2 year points appear around the center of the PC1 axis. As a transitional period between Era 1 and Era 3, the playstyle indicator values in these years tend to be closer to the mean, leading to z-scores near zero after standardization. As a result, the PC1 scores are also near zero, and the corresponding points are located near the origin in the PCA plot.

Era 3 year points cluster on the left side of the PC1 axis. In this era, the values of modern playstyle indicators are above the overall mean, resulting in positive z-scores. When these are multiplied by the negative PC1 loadings, the resulting PC1 scores are negative, hence the leftward positioning.

It is also worth noting that the absolute PC1 scores of Era 3 are greater than those of Era 1. Specifically, while Era 1 scores center around +5, Era 3 scores reach approximately –7.5. This suggests that modern playstyle characteristics are more strongly expressed in Era 3 along the PC1 dimension.

6. Random Forest

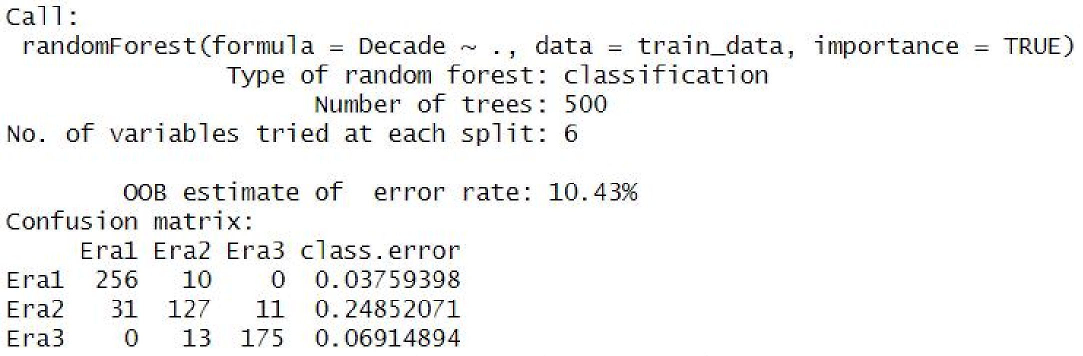
After segmenting the years into eras using PCA and changepoint detection, the study proceeds to analyze the data using a random forest model. Before delving further, the research objectives of the random forest analysis are revisited to clarify the analytical focus.

1. Evaluate predictive performance (quantitative)
2. Identify the variables that best distinguish changes in NBA playstyle across eras (qualitative)

The final model is a random forest consisting of 500 trees, with 6 variables randomly selected at each split—approximately equal to √30, given that the rank of the data is 30, as previously discussed. The dataset is split into 70% for training and 30% for testing, and a fixed random seed is set in R (set.seed(42)) to ensure reproducibility of the results.

6.1 Evaluate Predictive Performance (Quantitative)

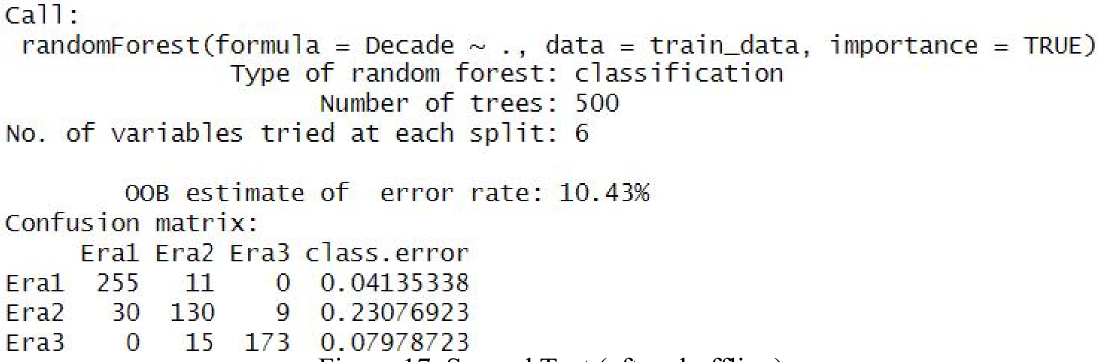
Two separate random forest tests were conducted to evaluate predictive performance. The results of both tests are summarized as follows:



```
Call:
 randomForest(formula = Decade ~ ., data = train_data, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 10.43%
Confusion matrix:
     Era1 Era2 Era3 class.error
Era1  256   10    0  0.03759398
Era2   31  127   11  0.24852071
Era3    0   13  175  0.06914894
```

Figure 16. First Test Result



```
Call:
 randomForest(formula = Decade ~ ., data = train_data, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 10.43%
Confusion matrix:
     Era1 Era2 Era3 class.error
Era1  255   11    0  0.04135338
Era2   30  130    9  0.23076923
Era3    0   15  173  0.07978723
```

Figure 17. Second Test (after shuffling)

As shown in the figures, both tests yielded an Out-of-Bag (OOB) estimate of error rate of 10.43%. The confusion matrices provide detailed classification results: the model performs well in identifying Era 1 and Era 3, with error rates below 5% and 7%, respectively. However, Era 2 exhibits a relatively higher prediction error of around 23–25%, suggesting that its playstyle features are less distinct and more transitional in nature.

These consistent results across both tests support the conclusion that the random forest model employed in this study is both accurate and stable in evaluating predictive performance for NBA era classification.

6.2 Key Variables Distinguishing NBA Playstyle Eras (Qualitative)

In analyzing which variables best distinguish changes in NBA playstyle across eras, this study uses Gini importance as the basis for interpretation. The results are shown in the figure below:

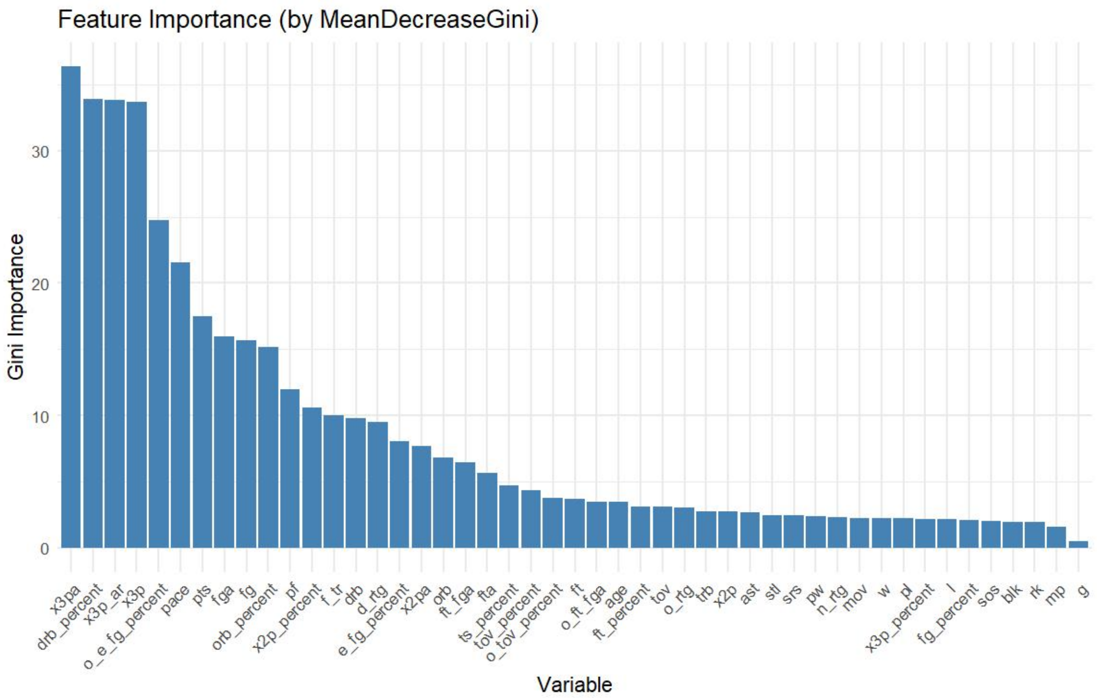


Figure 18. Feature importance histogram ranked by MeanDecreaseGini

Feature	**Description**	**Gini Importance**
x3pa	3-point attempts	37.85
drb_percent	Defensive rebound %	33.13
x3p	3-point makes	32.97
x3p_ar	3-point attempt rate	27.61
o_e_fg_percent	Opponent effective FG%	25.83
pace	Game pace	22.31
orb_percent	Offensive rebound %	16.42
ft_fga	Free throw rate	15.38
ts_percent	True shooting %	13.57
n_rtg	Net rating	13.31

Table 1. Top 10 Gini-Important Variables

This study lists the top 10 variables along with their full descriptions and corresponding Gini importance scores. Notably, three of the top four most important variables are related to three-point shooting, indicating that 3-point frequency and accuracy are highly influential in distinguishing playstyle eras. In other words, the rise of the 3-point era is the most significant feature of Era 3— a key qualitative conclusion drawn from this analysis. To further illustrate this point, the study presents boxplots for the top 5 variables by Gini importance. As shown in the plots, x3p, x3pa, and x3p_ar exhibit significantly larger jumps from Era 2 to Era 3 compared to the other top-ranked variables, drb_percent and o_e_fg_percent. This visual pattern further supports the conclusion that the emergence and dominance of the 3-point era is the defining characteristic of Era 3.

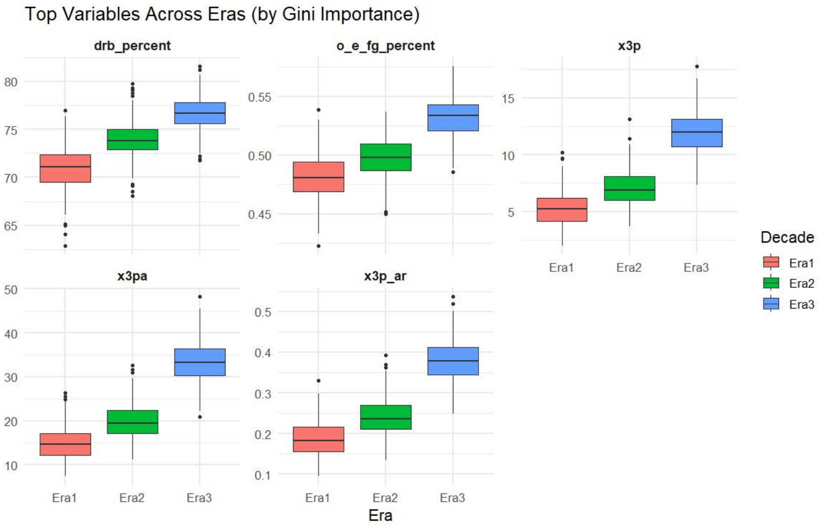


Figure 19. Boxplots of Top 5 Variables Across Eras

In addition to the importance of three-point-related variables, this study also uncovers several other noteworthy findings: drb_percent and orb_percent both rank highly, reflecting evolving strategies in rebounding control. Traditional eras emphasized inside presence, while modern bigs prioritize spacing and mobility. pace, n_rtg, and ts_percent are all in the top 10, suggesting a clear trend of increased possessions and improved offensive efficiency. o_e_fg_percent ranks 5th overall, indicating a strategic shift: Era 1 focused on defensive pressure, while Era 3 traded space for offensive efficiency.

In summary, this can be interpreted as follows:

Era 1 (1996–2007): Focused on rebounding and physicality, with strong defense and a slower pace.
Era 2 (2008–2016): A balanced era between offense and defense — a transitional phase in playstyle evolution.
Era 3 (2017–2025): Defined by 3-point shooting, pace, and efficiency — the rise of the modern basketball system.

7. Conclusion

The study began with comprehensive data cleaning and preprocessing, including file consolidation, variable standardization, and dataset merging to ensure analytical consistency. It then proceeded with exploratory data analysis, which revealed key trends such as the league's increasing reliance on three-point shooting, evolving rebounding dynamics, and notable shifts in pace and offensive efficiency. Building upon these insights, this study proposes an innovative method for segmenting NBA playstyle eras by combining PCA and changepoint detection. Specifically, eras are defined based on significant changes in the PC1 scores, allowing for data-driven segmentation without disrupting the temporal continuity of the timeline. At the same time, this project successfully applied Random Forest to predict era classification based on team statistics (Quantitative) and used Gini importance to identify key variables influencing era shifts, especially 3-point-related features (Qualitative).

Commitment

We would like to acknowledge the contributions of each team member to this project. Zhizhao Wang was responsible for exploratory data analysis (EDA) and data cleaning. Haonan Chen contributed by generating all figures and writing the core data processing code. Yiqi Luo led the modeling and algorithm analysis, including PCA, changepoint detection, and Random Forest implementation. We sincerely thank all members for their dedication and collaborative effort.