THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Visualization of Complex Data DATS 6401

Final Term Project
On

NBA PLAYER AND SEASON ANALYSIS

INSTRUCTOR: Dr. REZA JAFARI
AUTHOR: KUMAR SAURAV JHA (G20215692)
25th Apr 2024

TABLE OF CONTENTS

TABLE OF FIGURES

1. ABSTRACT

The project undertaken was an extensive analysis of NBA performance metrics, aimed at uncovering patterns, trends, and insights within professional basketball games. By utilizing a comprehensive dataset, the study employed various statistical and machine learning techniques, primarily focusing on the application of Python programming for data manipulation and visualization. The goal was to interpret complex player and team statistics and transform them into intelligible and actionable information through a series of plots and visualizations. These graphical representations ranged from basic histograms and bar plots to more intricate 3D scatter plots and cluster maps, each chosen for their specific ability to convey the nuances of the dataset.

The outcome of this project was the creation of a Python dashboard that provides an interactive interface for users to engage with the NBA data. This dashboard is not merely a static collection of charts but a dynamic tool that responds to user input, allowing for real-time data exploration. Through this platform, users gain the capability to distill large volumes of game statistics into digestible visual formats, empowering them to draw conclusions about player performance, team strategies, and game outcomes. The dashboard bridges the gap between raw data and strategic insights, proving invaluable for stakeholders ranging from team management to sports enthusiasts looking to deepen their understanding of the game's analytics.

2. INTRODUCTION

This project leverages the power of data analysis and visualization to unravel the rich story behind the numbers in the NBA. With a dataset spanning multiple seasons, it captures individual player performances and team dynamics, offering a window into the sport's competitive nature. By meticulously processing and visualizing this data, the project aims to deliver a multifaceted view of basketball analytics. The insights drawn from this analysis are not only meant to satisfy the curiosity of avid sports fans but also to provide strategic inputs for teams and analysts seeking to enhance performance and decision-making.

3. DESCRIPTION OF THE DATASET

The multivariate dataset chosen has the following features:

- 1. Real matches info and details about players and teams from the National Basketball Association
- 2. 702,387 observations in total
- 3. 24 numerical features and 6 categorical features
- 4. Is publicly available (scraped via NBA.com)

Below is the description of the features of the dataset:

Field	Description
gameid	A unique identifier for each game.
date	The date of the game.
type	The type of game (e.g., regular season, playoffs).
playerid	A unique identifier for each player.
player	The name of the player.
team	The team the player belongs to.
home	Indicates whether the team is playing at home or away.
away	Indicates whether the team is playing away.
MIN	Minutes played by the player in the game.
PTS	Points scored by the player.
FGM	Field goals made by the player.

| FGA | Field goals attempted by the player.
|
| FG% | Field goal percentage (ratio of successful field goals to attempted field goals).
|
| 3PM | Three-point field goals made by the player.
|
| 3PA | Three-point field goals attempted by the player.
|
| 3P% | Three-point field goal percentage (ratio of successful three-point field goals to attempted three-point field goals).
|
| FTM | Free throws made by the player.
|
| FTA | Free throws attempted by the player.
|
| FT% | Free throw percentage (ratio of successful free throws to attempted free throws).
|
| OREB | Offensive rebounds grabbed by the player.
|
| DREB | Defensive rebounds grabbed by the player.
|
| REB | Total rebounds grabbed by the player.
|
| AST | Assists made by the player.
|
| STL | Steals made by the player.
|
| BLK | Blocks made by the player.
|
| TOV | Turnovers committed by the player.
|
| PF | Personal fouls committed by the player.
|
| +/- | Plus-minus rating of the player (difference between points scored by the player's team and points scored by the opposing team while the player is on the court). |
| win | Indicates whether the player's team won or lost the game.
|

| season | The season in which the game took place.
|
| year | The year in which the game took place.
|
| month | The month in which the game took place.
|
| day | The day of the month on which the game took place.
|

tab1.

Why is this dataset interesting for visualization?

The NBA is a globally recognized association in the field of basketball, sort of like the NFL or FIFA. The NBA is a billion-dollar industry. With its global fan base, lucrative television deals, sponsorship agreements, merchandise sales, and revenue from ticket sales, the NBA generates substantial income.
According to Forbes, the NBA consistently ranks among the top sports leagues globally in terms of revenue generation, firmly establishing it as one of the most financially successful sports organizations in the world.
So this project focussing on the player and team analysis will be an interesting one considering the huge capital involved every season, such kind of information can be utilized for roster management and player trade-offs.

4. PRE-PROCESSING OF THE DATASET

The integrity of the dataset was paramount for the accuracy of the analysis. The pre-processing steps were meticulously carried out with the Python pandas library to ensure the dataset's quality.

Missing Values:
The dataset was thoroughly checked for missing values using the isnull().sum() function, which provided a count of missing entries for each feature. The results indicated that there were no missing values across the entire dataset. This is an indication of the robustness of the data collection process and minimizes the need for imputation strategies, allowing us to proceed with the original data without the risk of introducing bias.

Data Type Standardization:
The date field was converted from an object data type to datetime, standardizing the format and ensuring compatibility with time-series analyses. Furthermore, categorical columns, including type, team, home, away, and win, were converted to the category data type. This conversion optimizes memory usage and sets the stage for encoding if necessary for future modeling processes.

Data Cleaning:
Duplicate records can skew the results of the analysis, leading to inaccurate insights. We checked for duplicates in the dataset, ensuring that no repeated records would contaminate our findings. The absence of duplicate entries was confirmed, indicating that the dataset maintains a high level of integrity.

Data Type Verification:
A final verification of data types was conducted to ensure that each feature was stored in the most appropriate format. This verification is crucial for the proper functioning of statistical and machine learning models, as they require data in specific formats. For example, numerical features such as PTS were confirmed to be in integer format, aligning with the nature of the data as a countable metric.

This meticulous pre-processing guarantees that the dataset is in an optimal state for the subsequent phases of the project, allowing for precise and reliable analyses.

This is how the data records are:

```
In [6]: print(raw_data.head().to_string())
```

	gameid	date	type	playerid	player	team	home	away	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	STL	BLK	TOV	PF	+/-	win	season
0	29600001	1996-11-01	regular	893	Michael Jordan	CHI	BOS	CHI	43	30	10	22	45.5	0	5	0.0	10	13	76.9	1	3	4	3	0	0	1	2	11	1	1997
1	29600001	1996-11-01	regular	937	Scottie Pippen	CHI	BOS	CHI	40	18	8	17	47.1	1	3	33.3	1	2	50.0	1	7	8	6	2	0	5	4	28	1	1997
2	29600001	1996-11-01	regular	677	Eric Williams	BOS	BOS	CHI	25	14	6	13	46.2	0	0	0.0	2	3	66.7	2	1	3	1	0	0	1	4	1	0	1997
3	29600001	1996-11-01	regular	146	Jud Buechler	CHI	BOS	CHI	1	0	0	1	0.0	0	1	0.0	0	0	0.0	0	0	0	0	0	0	0	0	-2	1	1997
4	29600001	1996-11-01	regular	166	Ron Harper	CHI	BOS	CHI	25	7	3	4	75.0	0	1	0.0	1	2	50.0	1	1	2	5	2	1	0	1	15	1	1997

Fig 4.1.

5. OUTLIER DETECTION & REMOVAL

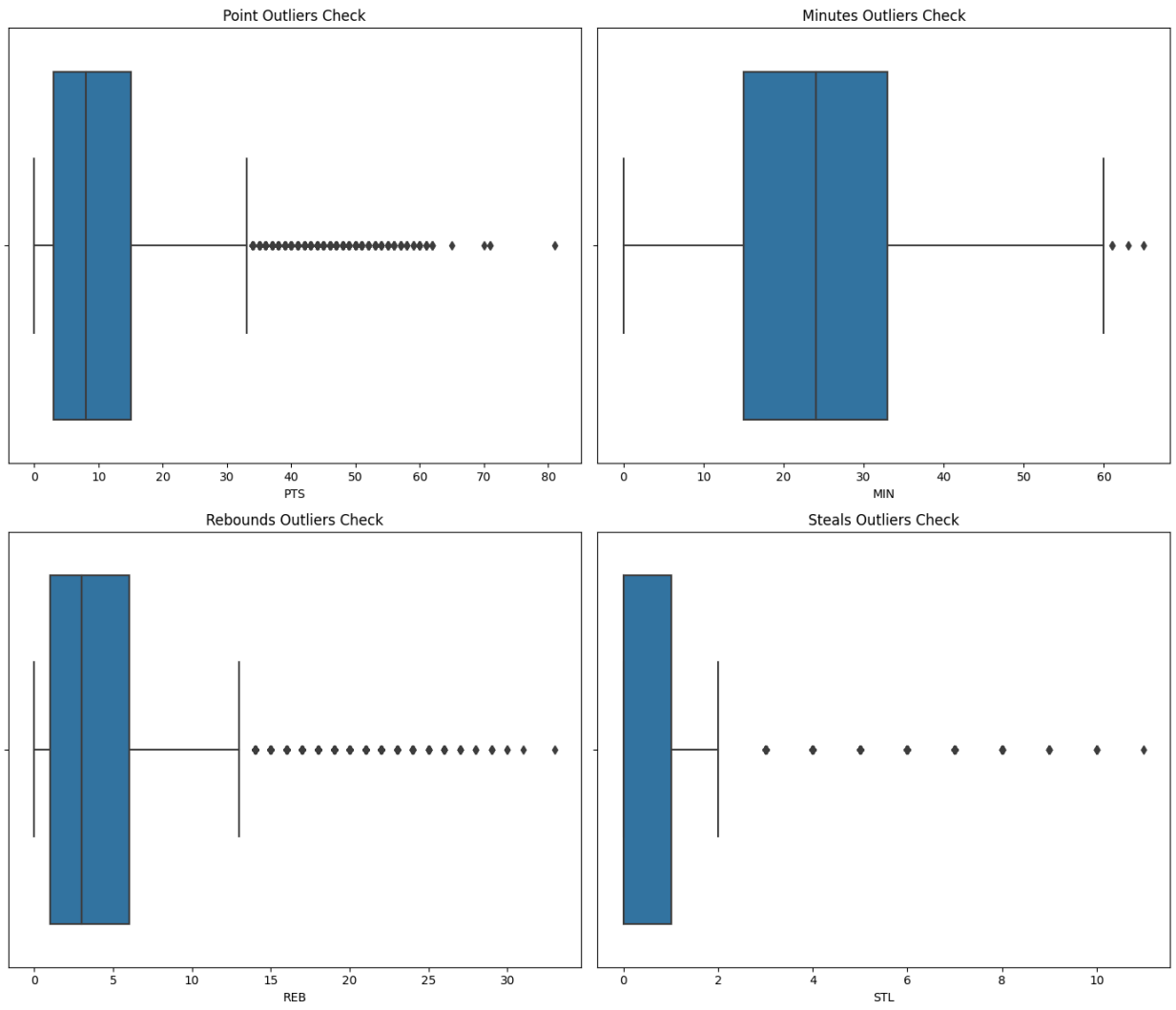

Fig 5.1

The boxplot visualization has revealed varying degrees of spread across four key statistical categories within our dataset. These categories were specifically chosen because they encapsulate critical aspects of a basketball game that directly influence a player's performance and the game's outcome.

1. *Points (PTS):* The boxplot for points scored shows a significant number of data points outside the upper whisker, indicating instances of extraordinary scoring performances. These outliers represent games where players scored much higher than usual, which could be indicative of standout games or critical matches where star players might have carried the team.
2. *Minutes (MIN):* The minutes played boxplot has fewer outliers, mostly on the higher end. This suggests occasions when players had exceptionally high court

time, possibly during overtime games or close matches where key players are likely to be kept on the floor longer.

3. *Rebounds (REB):* Similar to points, the rebounding statistics show a spread of outliers above the upper quartile, representing games where players achieved a significantly higher number of rebounds, pointing to dominant performances on the boards.
4. *Steals (STL):* The steals box plot indicates outliers with higher than usual steals in a game, which may highlight outstanding defensive efforts or games with a high turnover rate.

The decision not to remove outliers from the dataset is a deliberate one, driven by the nature of the sport of basketball. Unusual performances, such as scoring sprees, defensive masterclasses, or enduring stamina, provide valuable insights into player capabilities and game dynamics. These outliers, rather than skewing the data, offer a glimpse into the upper echelons of performance in professional basketball. Removing these data points could potentially omit critical instances of player performance that are of great interest to analysts, coaches, and fans alike. Moreover, these outliers can be particularly useful in predictive modeling to anticipate star player performances or understand the variance in player contributions across different games.

6. **PRINCIPAL COMPONENT ANALYSIS**

[HEATMAP & PEARSON CORRELATION COEFFICIENT MATRIX INCLUDED]

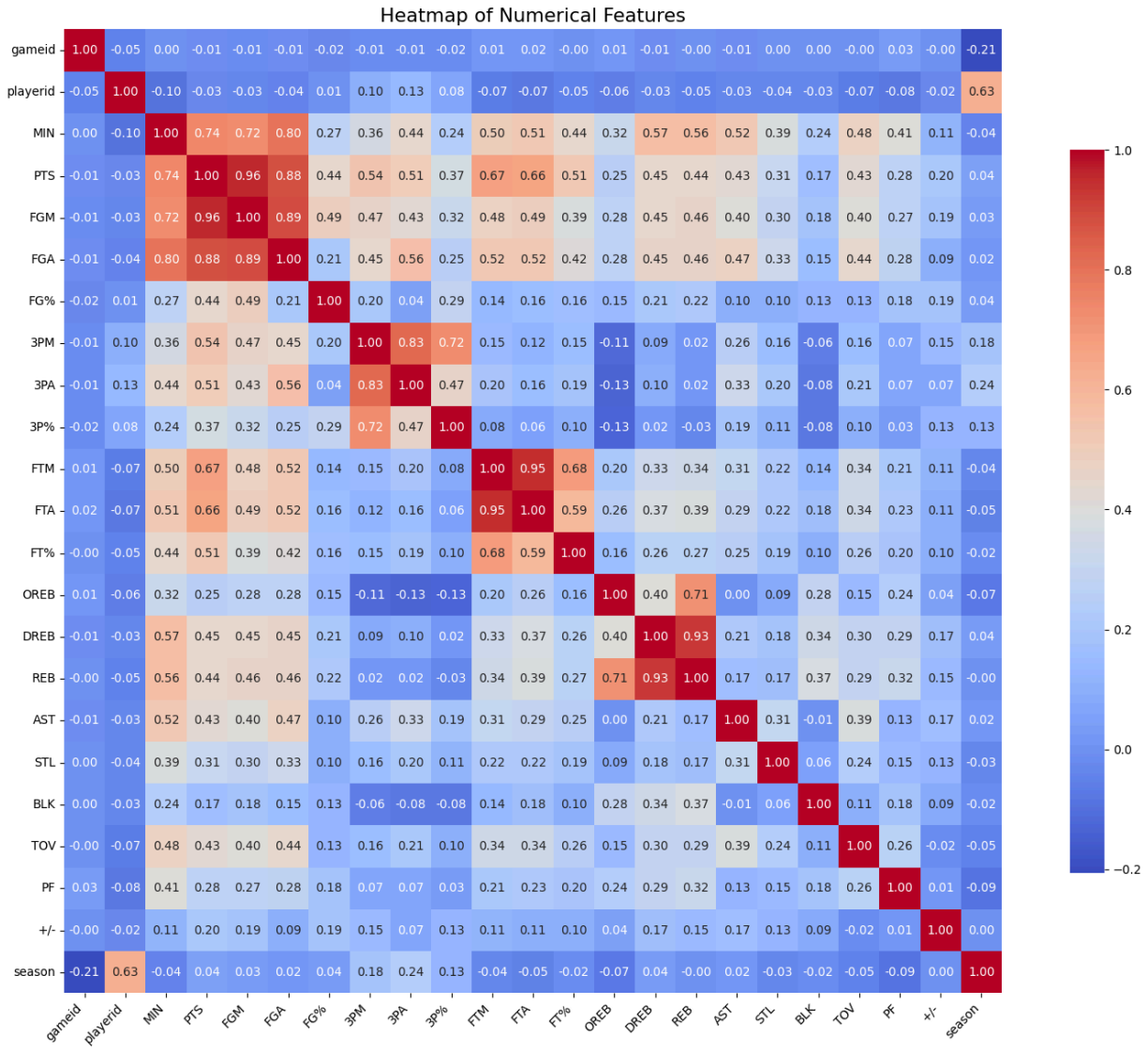

Fig 6.1

Heatmap Analysis of Basketball Player Statistics

In our analysis, we visualized the correlation between various numerical features of basketball player statistics using a heatmap. The heatmap provides a color-coded representation of the correlation matrix, making it easier to identify relationships between the features.

Here's what we can deduce from the heatmap provided:

- High Correlation Coefficients: Several statistics exhibit strong positive correlations. For example, minutes played (MIN) has a strong positive correlation with points scored (PTS), indicating that players who spend more time on the court tend to score more points.
- Rebounds and Field Goals: The heatmap shows a strong relationship between the number of rebounds (REB) and field goals made (FGM), suggesting that successful rebounds can be an indicator of scoring opportunities.
- Scoring and Assists: Points scored (PTS) also correlate well with assists (AST), which implies that players who score more are also involved in setting up scoring opportunities for their teammates.
- Free Throw Accuracy: Free throw percentage (FT%) has a moderate positive correlation with points (PTS), suggesting that accuracy in free throws can contribute to a player's overall scoring performance.
- Negative Correlations: Turnovers (TOV) show a slight negative correlation with free throw percentage (FT%), indicating that players with higher accuracy in free throws tend to make fewer turnovers. However, this relationship is not strong enough to be considered significant without further analysis.
- Independence of Features: Some features like steals (STL) and blocks (BLK) show little to no correlation with other statistics, indicating that these skills are quite independent and specialized.

Overall, the heatmap provides valuable insights into how different aspects of player performance are interrelated. It is evident that while some skills are closely linked, others stand alone and contribute uniquely to the game. Such an analysis helps in understanding the multidimensional nature of player performance and can guide strategies in team formation and player development.

Upon computing PCA for the dataset, we determined the minimum number of principal components required to preserve 95% of the data variance. The results are as follows:

Principal Component Analysis Summary

Table 2: PCA on Basketball Dataset

Metric	Original Data	Reduced Data (95% Variance)
Number of Components	15	12
Number of Features Removed	-	3
Highest Singular Value	1912.19	1912.19
Lowest Singular Value	180.97	535.93
Condition Number	10.566	10.566

Fig 6.2

The PCA was conducted on a set of basketball metrics, with the intention of reducing feature space while preserving 95% of the original variance. The analysis began with 15 features, which were then condensed to 12 principal components. This process resulted in the removal of 3 features from our dataset.

The original feature space had a range of singular values from approximately 1912.19 down to 180.97. Post-PCA, the range was narrowed down, with the smallest singular value considered being approximately 535.93. Notably, the condition number remained unchanged at about 10.566, indicating that the degree of multicollinearity or numerical instability in the dataset was not significantly impacted by the reduction.

By retaining 95% of the variance with fewer components, we have streamlined the feature space for further analysis, potentially improving the computational efficiency and interpretability of subsequent models. However, given the unchanged condition number, careful consideration must still be given to the relationships between variables when applying further modeling techniques.

These results should be incorporated into the development of predictive models, keeping in mind the underlying structure and multicollinearity within the dataset.

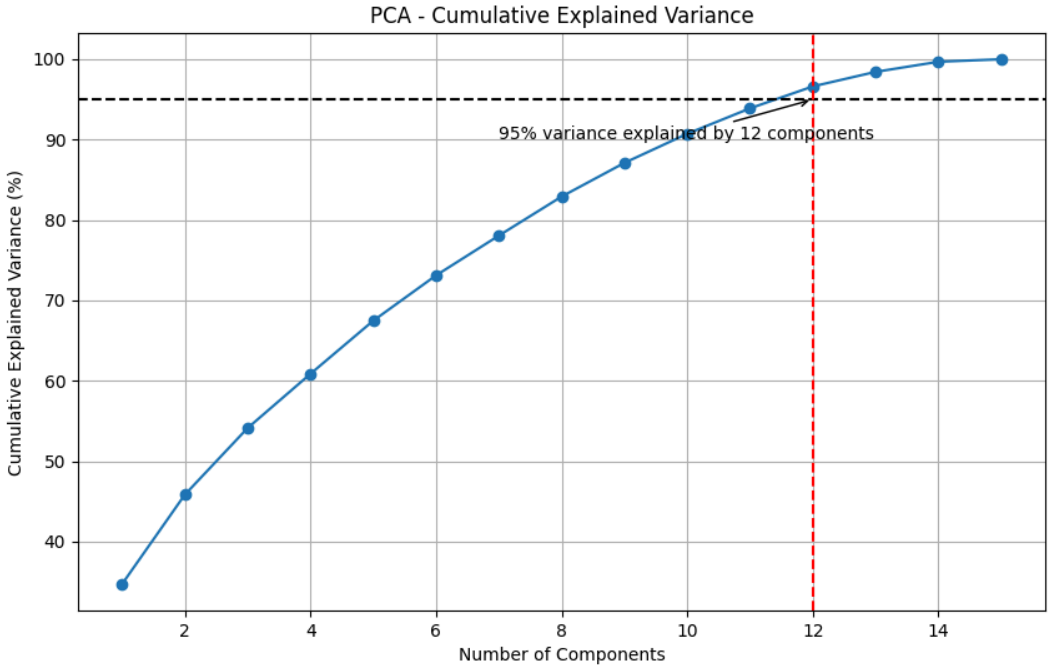

Fig 6.3

This dimensionality reduction through PCA not only simplifies the dataset but also aids in visualizing the data structure, revealing patterns that may not be apparent in the high-dimensional space. It allows for a more focused analysis on the components that carry the bulk of the information content.

The findings justify the reduction of features for any modeling efforts, ensuring that the most meaningful data is retained while reducing computational complexity.

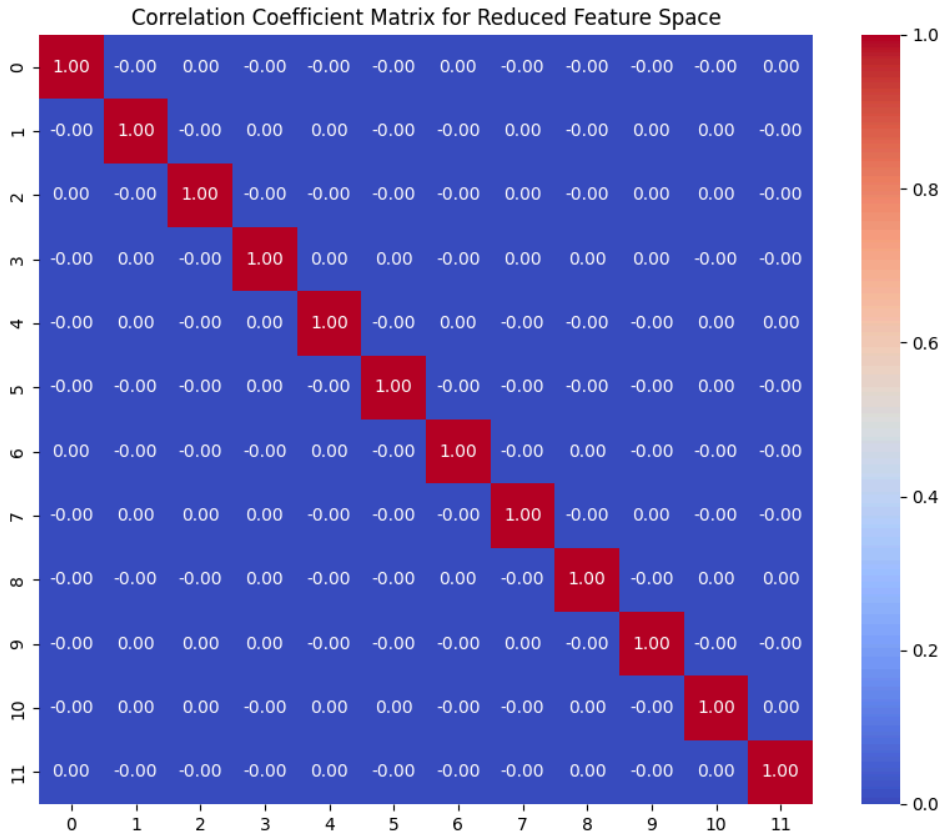

Fig 6.4

7. NORMALITY TEST

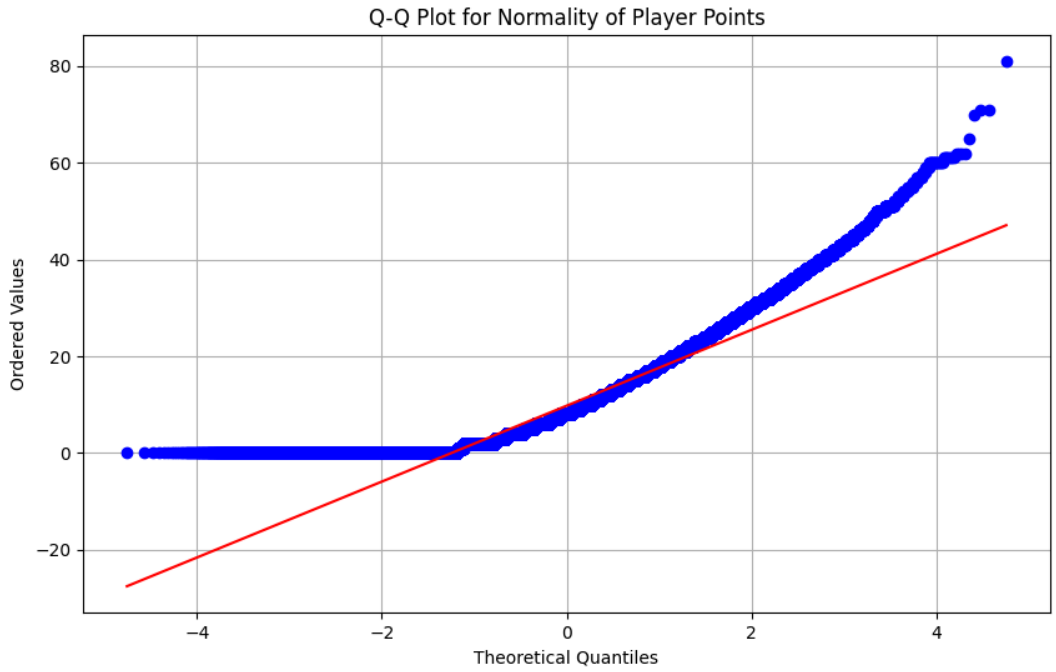

Fig 7.1

To assess the normality of the 'Points Per Game' (PTS) distribution within the dataset, a Kolmogorov-Smirnov (K-S) test was conducted. This non-parametric test is utilized to compare a sample's distribution with a normal distribution, thereby determining if there are significant differences between them.

```
    ...:
    ...: print(f'Statistics={stat:.3f}, p={p:.3f}')
    ...: print(f'The data follows a {normality}')
    ...:
Statistics=0.846, p=0.000
The data follows a not normal distribution (reject H0)
In [70]: #NORMALITY TEST
```

Fig 7.2

The K-S test produced a statistical value of 0.846 with a corresponding p-value of 0.000. In hypothesis testing, a p-value below the predetermined alpha level, which in this case is 0.05, suggests sufficient evidence to reject the null hypothesis. The null hypothesis for this test posits that the sample data is derived from a normal distribution.

With the p-value being significantly lower than 0.05, we reject the null hypothesis, leading to the conclusion that the 'Points Per Game' data does not conform to a normal distribution. However, it is important to note that the normality of data is not a prerequisite for all statistical analysis or machine learning applications. Given the robustness of various modern techniques to violations of normality and the possibility that transformations can sometimes distort the data's natural structure, it is determined that transforming the data to force normality is not necessary in this instance. The focus will instead be on leveraging methods that can handle non-normal distributions effectively.

8. STATISTICS

	gameid	playerid	MIN	PTS	FGM	FGA	FG%	3PM
count	702387	702387	702387	702387	702387	702387	702387	702387
mean	2.35472e+07	292283	23.4559	9.77986	3.63799	8.01833	41.5414	0.732069
std	5.76512e+06	543377	11.5637	8.19247	3.0848	5.81306	25.1377	1.2018
min	2e+07	2	0	0	0	0	0	0
25%	2.07002e+07	1563	15	3	1	3	26.7	0
50%	2.14004e+07	2749	24	8	3	7	42.9	0
75%	2.21004e+07	202700	33	15	5	12	56.3	1
max	5.22002e+07	1.64164e+06	65	81	28	50	100	14

3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB
702387	702387	702387	702387	702387	702387	702387	702387
2.05528	19.1111	1.77181	2.3399	43.8758	1.0838	3.04378	4.12757
2.53224	28.7161	2.38947	2.94198	42.7192	1.44371	2.71204	3.54803
0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	1
1	0	1	2	50	1	2	3
3	33.3	3	4	90	2	4	6
24	100	26	39	100	18	25	33

STL	BLK	TOV	PF	+/-
702387	702387	702387	702387	702387
0.740135	0.475128	1.34323	2.05164	-0.000314641
0.993277	0.891913	1.41868	1.52182	10.5813
0	0	0	0	-57
0	0	0	1	-7
0	0	1	2	0
1	1	2	3	6
11	13	14	7	57

Fig 8.1

1. gameid and playerid: These are identifier fields; their descriptive statistics are less relevant for analysis.

2. MIN (Minutes Played):
- Players play an average of 23.46 minutes per game, with a standard deviation of 11.56, indicating a moderate spread in minutes played.
- The minimum is 0 minutes (likely due to not playing that game), and the maximum is 65 minutes (indicating overtime games).

3. PTS (Points Scored):
- The average points scored by a player per game is approximately 9.78.
- The wide standard deviation of 8.19 suggests a significant variability in points scored among players.
- The maximum points scored in a game is an extraordinary 81, indicating an exceptional performance.

4. FGM, FGA, and FG% (Field Goals Made, Attempted, and Percentage):
- On average, players make about 3.64 field goals per game out of 8.02 attempts, with an average success rate of 41.54%.
- The field goal percentage varies widely (std 25.14%), including perfect 100% shooting in some games.

5. 3PM, 3PA, and 3P% (Three-Point Field Goals Made, Attempted, and Percentage):

- Players make less than one 3-point field goal per game on average (0.73) out of approximately two attempts, with a success rate of 19.11% on average, which suggests three-point shots are less frequent or less successful than field goals.

6. FTM, FTA, and FT% (Free Throws Made, Attempted, and Percentage):
- Free throws made average around 1.77 per game from 2.34 attempts, with an average percentage of 43.88%, which seems anomalously low and could indicate an error in the data or calculations, as free-throw percentages are typically higher.

7. REB, OREB, DREB (Rebounds, Offensive Rebounds, Defensive Rebounds):
- Players average over 4 total rebounds per game, with both offensive and defensive rebounds showing similar variability.
- The maximum number of rebounds in a game is 33, which is high and notable.

8. AST (Assists):
- Players average around 2.15 assists per game, with a maximum of 25, indicating an outstanding playmaking game.

9. STL (Steals) and BLK (Blocks):
- Steals and blocks are less frequent events, averaging less than 1 per game.
- The high max values (11 for steals, 13 for blocks) are exceptional individual defensive achievements.

10. TOV (Turnovers) and PF (Personal Fouls):
- Players average around 1.34 turnovers and 2.05 personal fouls per game, with a reasonably consistent range.

11. +/- (Plus-Minus):
- The average plus-minus is very close to zero, which is expected as it averages out over many players and games.
- The range from -57 to +57 indicates some players have been on the court during significant shifts in the score.

These descriptive statistics give an overall picture of the dataset's spread and typical values, and they highlight the potential for exceptional individual performances in various categories.

9 DATA VISUALIZATION

In this section we'll be exploring different static plots to infer details from the dataset.

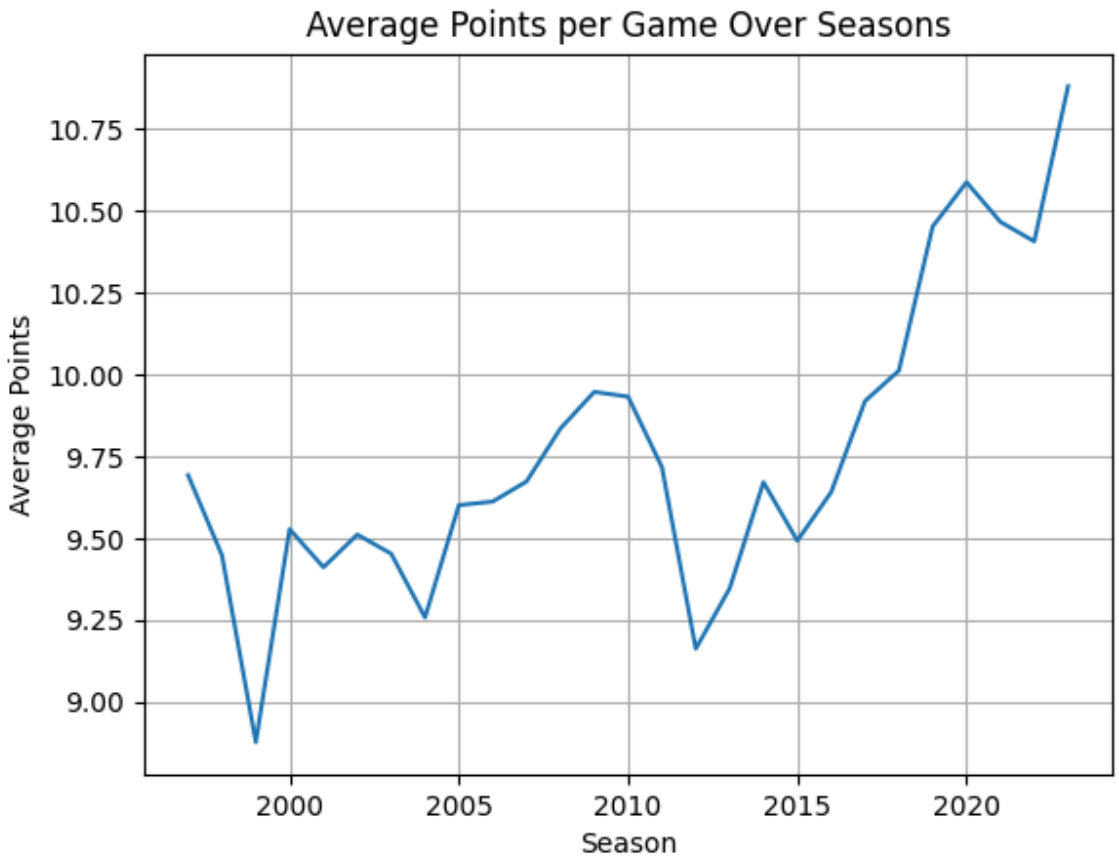

Fig 9.1

Trend Analysis of Average Points per Game

The line graph illustrates a longitudinal trend of the average points scored per game in various basketball seasons from the year 2000 onwards.

Observations:

The early 2000s experienced a minor decline in scoring, possibly reflecting a defensive era in the sport.

Subsequent years display a variable but generally ascending trend in points per game, indicating a shift towards more offensively focused gameplay.
The last segment of the graph shows a marked upsurge, highlighting a recent trend of significantly higher scores. This could be attributable to changes in game dynamics, rule modifications favoring offensive plays, or advancements in player efficiency and strategies.
Implications:
This trend suggests a transformation in the sport's playing style that may influence coaching tactics, player training, and the entertainment value for audiences.

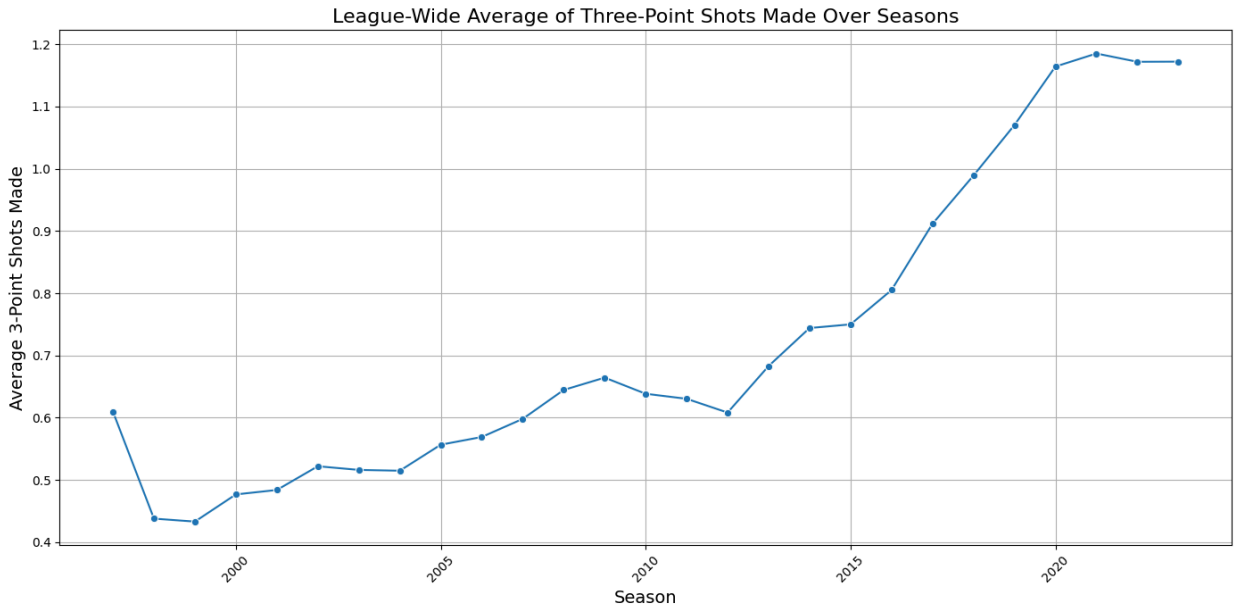

Fig 9.2

Three-Point Shooting Trend Analysis

The graph presents the evolution of the league-wide average of three-point shots made per game across various basketball seasons.

Observations:

There is a noticeable increase in three-point shots made from the early 2010s onwards.

The pronounced rise in the latter part of the graph suggests a strategic shift towards three-point shooting in recent years.
The initial stability followed by a steep ascent implies a change in team compositions, player skill sets, or potentially the influence of analytics favoring three-point shooting.
Implications:
This surge in three-point shot-making reflects a transformative phase in the sport, potentially altering player evaluation, training focuses, and game strategy. It also mirrors broader trends in basketball analytics that emphasize the efficiency of three-point scoring.

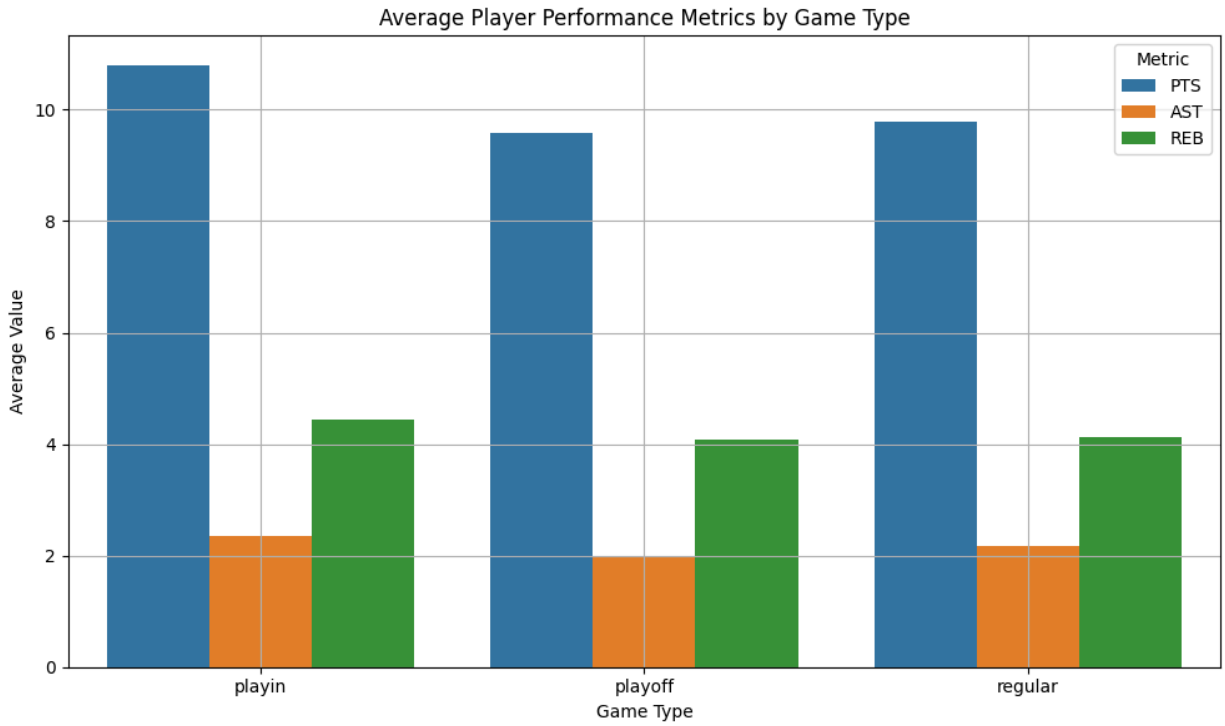

Fig 9.3

Player Performance Across Different Game Types

The bar chart visualizes average player performance metrics—points (PTS), assists (AST), and rebounds (REB)—across different game types (play-in, playoff, and regular season).

Observations:

Points per game (PTS) are consistently the highest metric across all game types, with playoff games showing a significant increase, indicative of higher stakes and possibly a more aggressive offensive strategy.
Assists per game (AST) and rebounds per game (REB) are relatively consistent across the play-in and regular seasons but show a noticeable dip in playoff games.
The drop in assists and rebounds during playoffs may suggest a shift in playing style, where individual scoring takes precedence, or it could reflect tighter defenses where assists become more challenging and rebounding becomes more competitive due to the physical nature of playoff games.
Implications:
The data underscores the importance of scoring in playoff situations and may imply strategic adjustments by teams to focus on point generation. The metrics also hint at the intensity of playoff games affecting the nature and frequency of assists and rebounds.

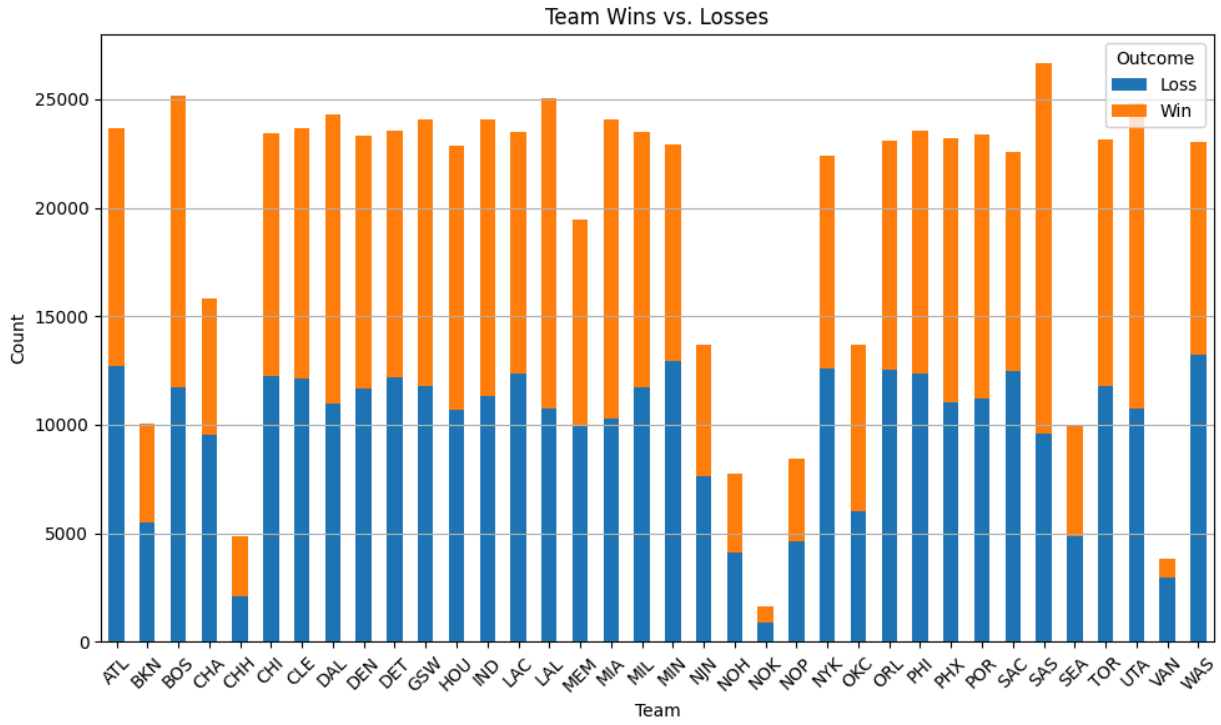

Fig 9.4

Team Wins vs. Losses Analysis
This bar plot provides a clear visual comparison of wins versus losses for NBA teams, displaying the aggregated count of each outcome side by side. The distinct color

coding—orange for wins and blue for losses—allows for immediate interpretation of the data. From a glance, one can discern patterns and anomalies in team performance, such as teams with disproportionately high wins or losses. Moreover, the stacked configuration of the bars showcases the total number of games played by each team, offering a holistic view of their overall activity within the seasons covered by the dataset.

Observations:

Certain teams exhibit a robust track record, as indicated by taller orange segments, suggesting consistent winning performance.
In contrast, teams with predominantly blue segments highlight a history of losses, which could signal areas for strategic improvement.
The overall count variation among teams implies differences in the number of games played, which could be due to several factors including longevity in the league and participation in playoff games.
The data visualization effectively highlights the competitiveness of the teams and may prompt further investigation into the factors contributing to their success or failure.

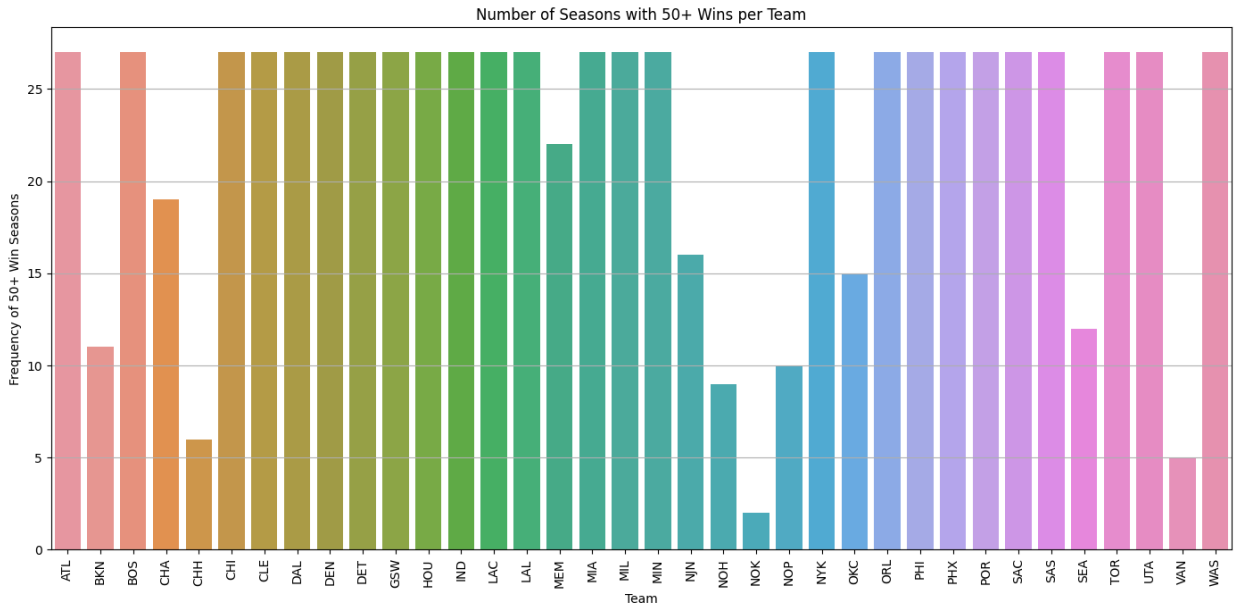

Fig 9.5

Team Performance Analysis: 50+ Win Seasons

The bar plot depicts the frequency of 50+ win seasons for each NBA team, with each bar's height representing the number of times a team has achieved this milestone. The varied color spectrum across the bars adds a visual appeal and aids in distinguishing between teams.

Observations:
- A select group of teams show an exceptionally high number of 50+ win seasons, reflecting a history of consistent high performance and possibly successful management strategies.
- Other teams demonstrate moderate success, with a frequency that suggests competitive, but not dominant, seasonal outcomes.
- Some teams have rarely or never achieved a 50+ win season, indicating either a newer presence in the league or a history of underperformance that may require strategic changes.

This analysis provides a longitudinal perspective on team success and can be a starting point for in-depth exploration of factors contributing to long-term team performance.

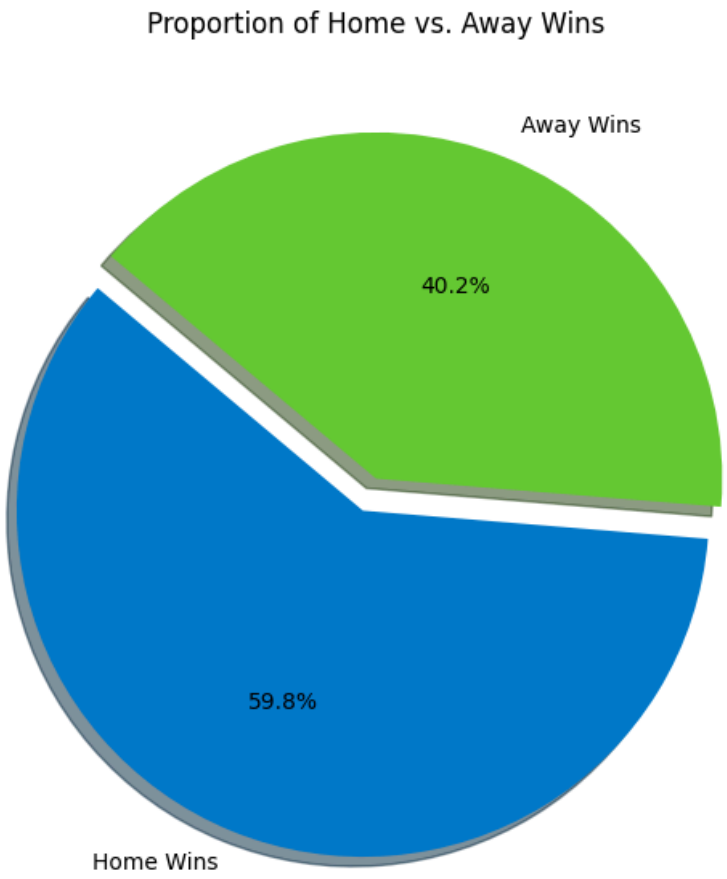

Fig 9.6

Home Court Advantage Analysis

The pie chart illustrates the proportion of games won by teams when playing at home versus away. With a substantial portion of the chart in blue, it shows that a majority of the wins (59.8%) occurred in home games, highlighting a potential home court advantage. This could be due to familiar playing conditions, supportive home crowds, or travel fatigue for the away team.

Observations:

Home teams won nearly 60% of the games, suggesting that playing on familiar territory may offer a competitive edge.
Away wins account for just over 40%, which, while significant, confirms that teams tend to perform better at home.
The data supports the concept of a home court advantage in basketball, which should be considered in team strategy and fan engagement efforts.
This visual underscores the importance of home games in a team's winning strategy and could be a crucial factor in playoff games where the stakes are higher.

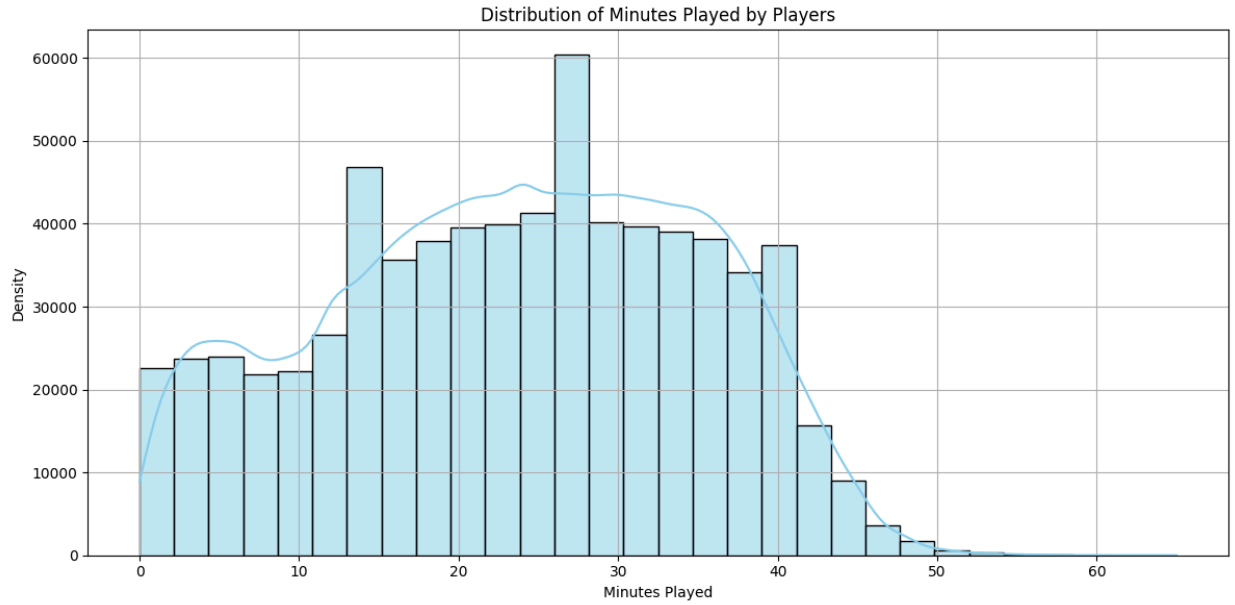

Fig 9.7

Player Playtime Distribution

The histogram with a kernel density estimate overlay provides a visual representation of the distribution of minutes played by basketball players. The bulk of the data centers around 20 to 30 minutes, indicating that most players are allocated this amount of playtime during a game.

Observations:

There's a noticeable peak around the 25-minute mark, suggesting a common playtime duration for many players.

The distribution is right-skewed, with fewer players reaching beyond 40 minutes, which is likely reserved for key players or due to extended playtime in the event of overtimes.
The presence of playtimes close to zero could indicate players who are benchwarmers or those who played in very few games, possibly due to injury or being new additions to the roster.
This distribution is critical for coaches and team management when considering player endurance, rotation strategies, and managing the risk of injuries due to overplay.

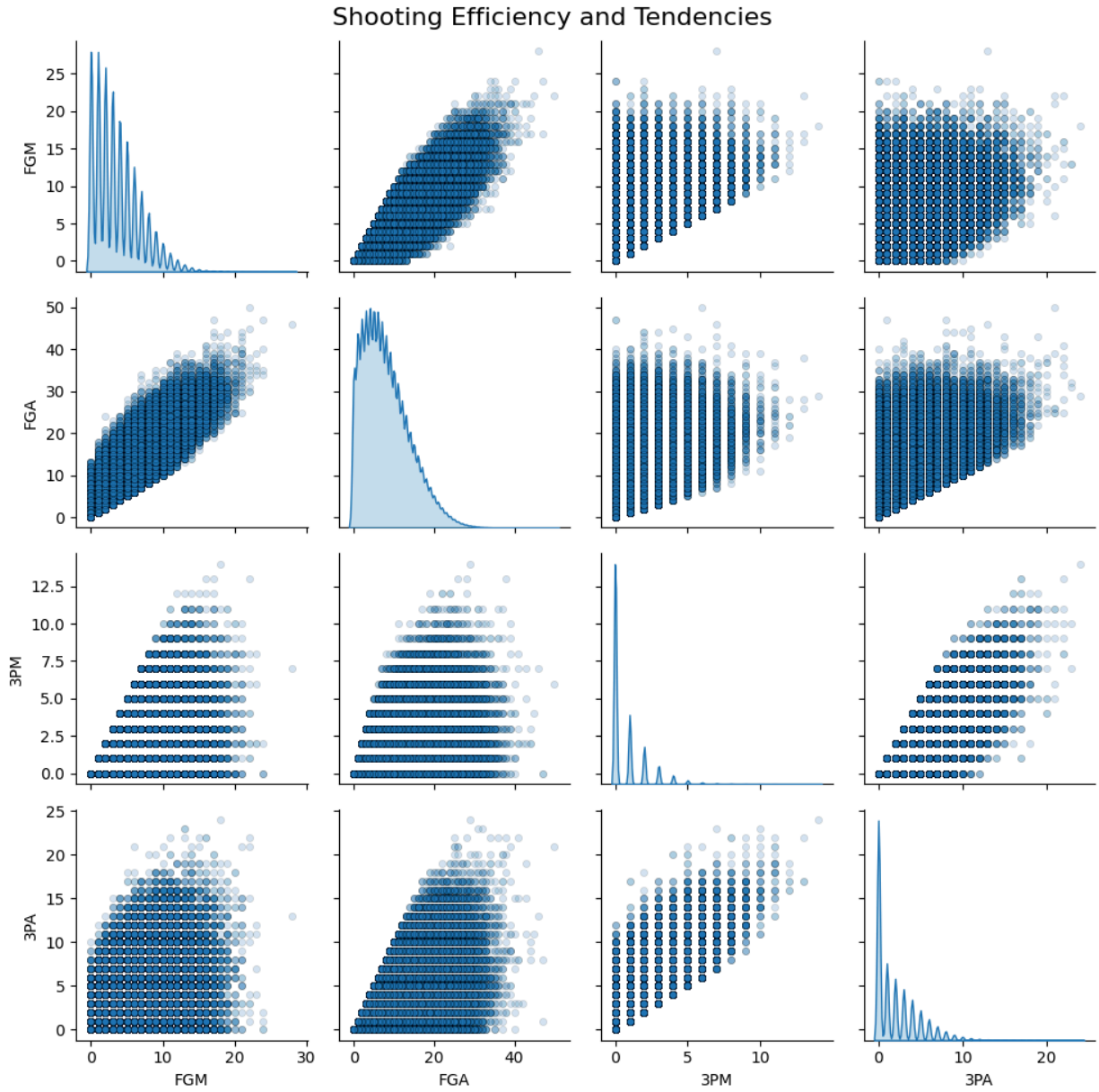

FIg 9.8

- The diagonal plots show the distribution for each of these variables. The shape of these distributions indicates how commonly different values occur; for example, a peak suggests a common value.
- The off-diagonal plots are scatter plots that show the relationship between pairs of these variables. A pattern in these plots, such as points clustering along a line, suggests a correlation between the variables.
- The scatter plots between FGM and FGA, as well as between 3PM and 3PA, likely show positive correlations: as attempts increase, so do the made shots.
- The plots combining FGA and 3PM or 3PA could indicate players' shooting preferences or how often players who attempt many field goals also shoot three-pointers.

Overall, this visualization helps in understanding the shooting behavior of players and how often their attempts translate into actual points, as well as their inclination towards taking three-point shots.

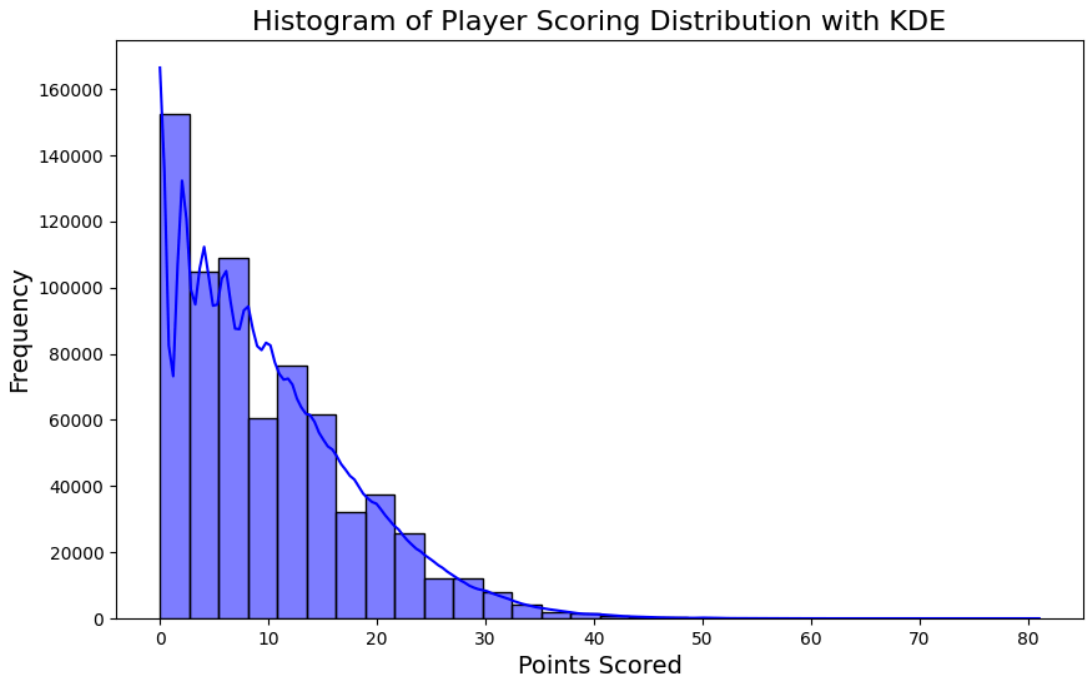

FIg 9.9

- The highest bars are clustered at the lower end of the points scored axis, indicating that it's most common for players to score a small number of points.
- The frequency of players scoring higher points progressively decreases, as seen by the decreasing height of the bars as we move right on the x-axis.

- Very high point scores, such as above 30, are quite rare—noticeable by the few and low-height bars in that range.
- The KDE curve follows the shape of the histogram and shows a peak around the lower end, confirming that lower scoring is the most frequent.
- There's a long tail extending towards the higher points scored, which suggests that while it's uncommon, there are instances where players score many points.

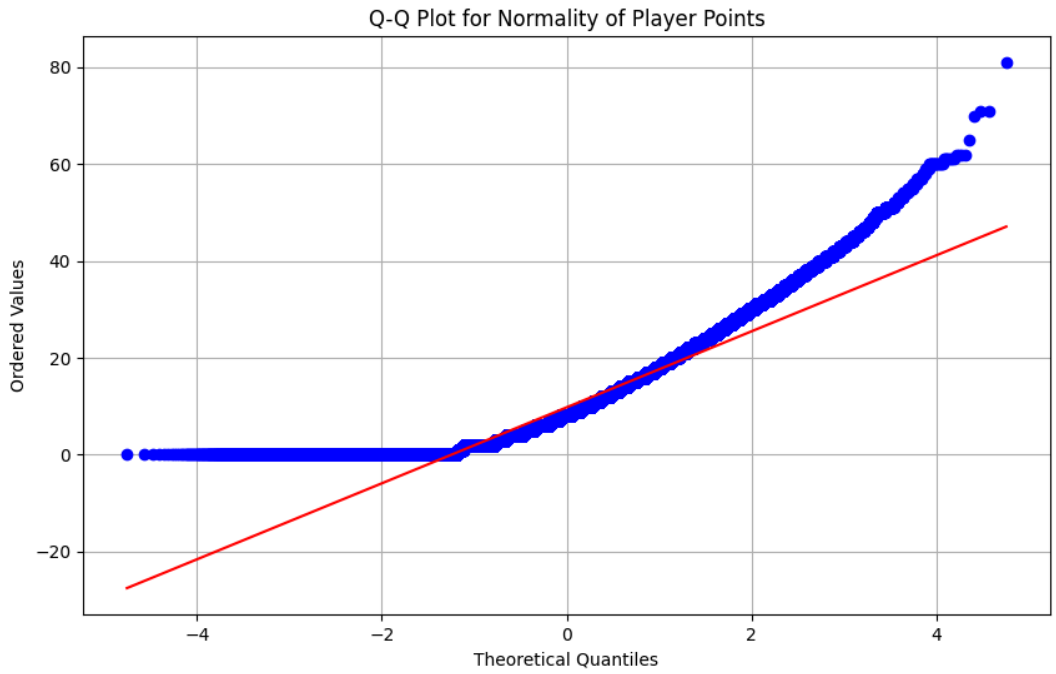

FIg 9.10

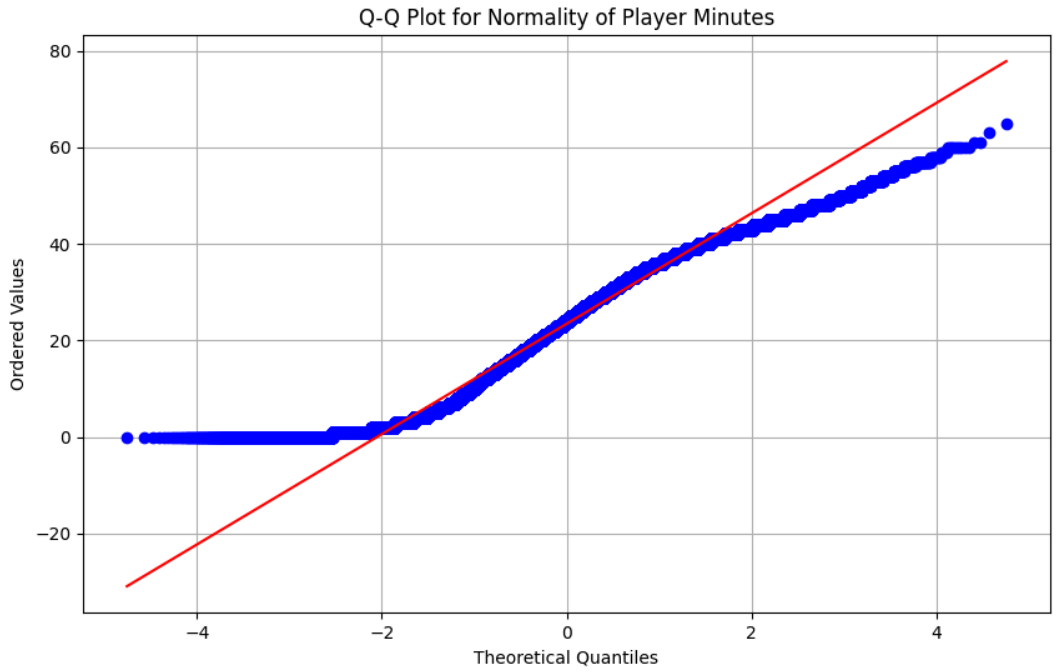

Fig 9.11

In simple terms, the players' minutes are not distributed normally. There are more instances of players playing very few and very many minutes than would be expected if the minutes were normally distributed. This could be due to players who rarely play (perhaps due to being on the bench) and star players who play almost the entire game, respectively.

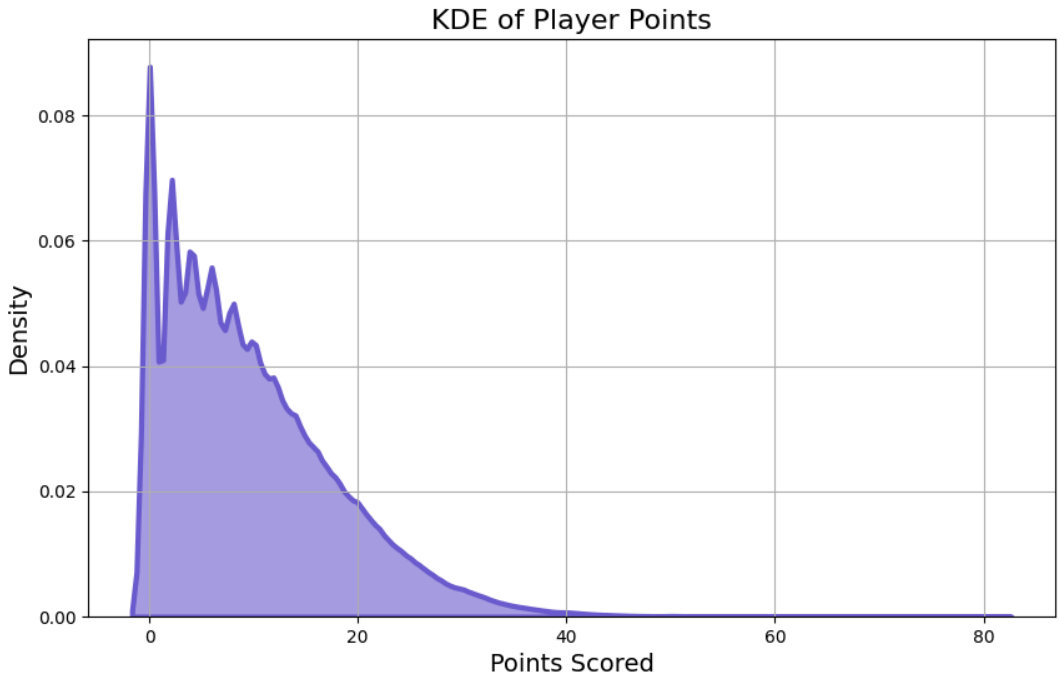

Fig 9.12

Points Scored Density Analysis

The Kernel Density Estimate (KDE) plot illustrates the probability density of points scored by players. This particular graph reveals several interesting patterns about scoring in games:

Observations:

A sharp peak close to zero suggests a significant number of instances where players score very few points. This might be indicative of bench players or defensive specialists whose contributions don't necessarily translate into points.
The density swiftly decreases as the points increase, which is expected since higher-scoring games by individual players are less common.
The long tail extending towards higher points scored reflects that while it's rare, some players do achieve very high scores in games, possibly stars known for their scoring ability.

There are minor peaks at intervals, possibly indicating common scoring thresholds such as 10, 20, 30 points, etc., which align with typical basketball scoring patterns like double-doubles or high-scoring halves.
This data is invaluable for understanding the distribution of scoring across players, which can help in team strategy and identifying key scorers.

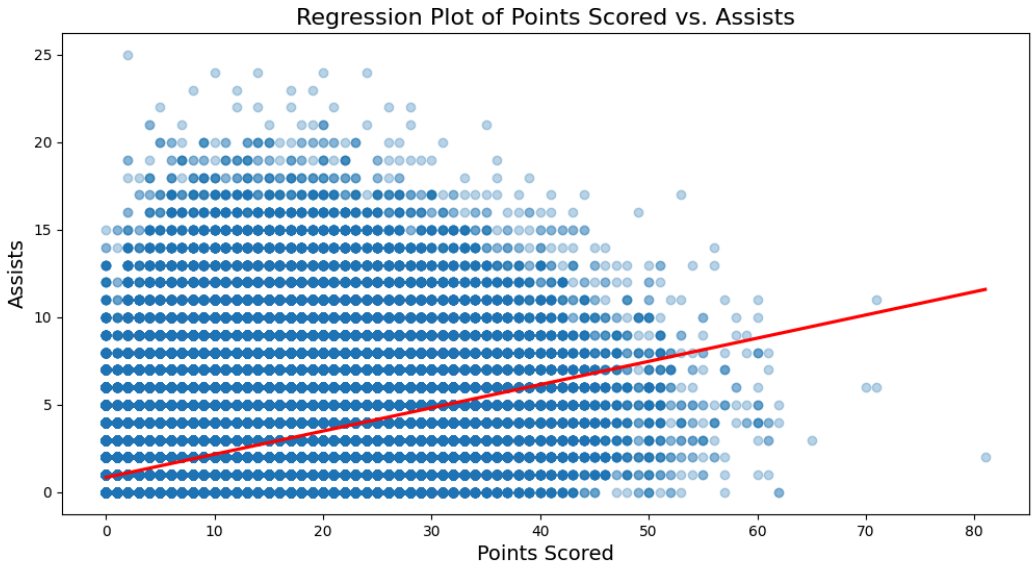

Fig 9.13

- **Positive Trend:** The upward slope of the regression line suggests a positive relationship between points scored and assists, meaning that as points increase, assists tend to increase as well.
- **Data Spread:** The scatter points seem to form a band that widens as points increase, which may indicate greater variability in assists among high-scoring games.
- **Concentration of Data Points:** There is a concentration of data points at the lower end of the points axis, indicating that many player-game instances involve scoring fewer points and correspondingly fewer assists.
- **Outliers:** There are instances at the higher end of points scored with a wide range of assists, suggesting that high scorers can have very different assist numbers.

This visualization helps to understand the typical pattern in player statistics: players who score more also tend to have more assists, but the variability in assists increases with higher points scored."'

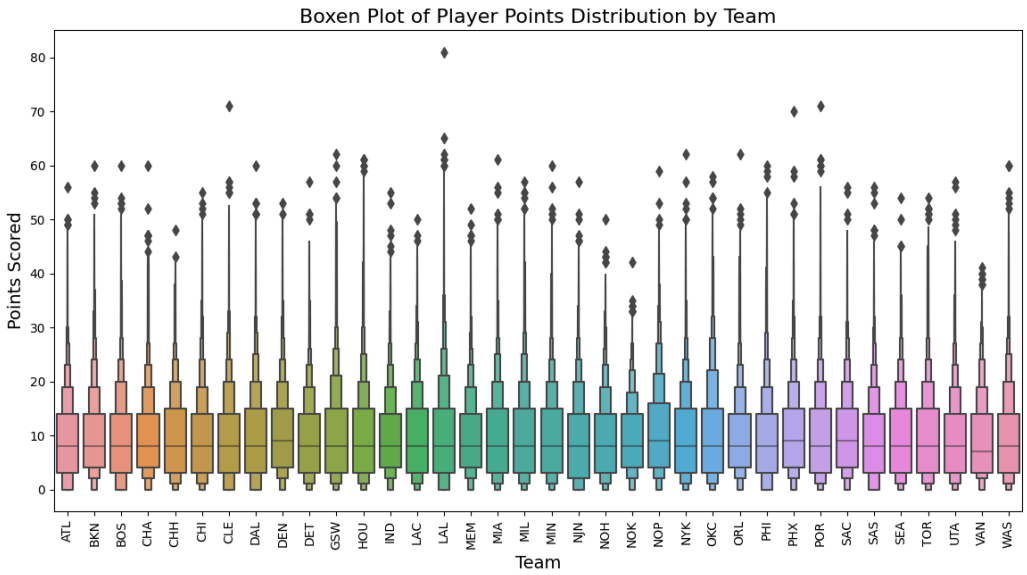

Fig 9.14

Team-Wise Player Points Distribution

The boxen plot provides a detailed visualization of the distribution of points scored by players across different teams. This comprehensive plot highlights the median, quartiles, and distribution range of points, along with outliers that represent exceptionally high-scoring games.

Observations:

The median points scored, indicated by the thick line in each box, vary slightly between teams, suggesting some differences in scoring strategies or player performance.
The interquartile range (the box size) gives an idea about the scoring consistency within a team. A larger box indicates a greater variability in individual scoring performance.
The presence of outliers (diamonds beyond the whiskers) in nearly all teams shows that there are occasionally exceptional scoring performances.
The plot does not suggest a significant difference in the distribution of points between teams, as indicated by the similar range and spread across the plot.
This detailed view helps identify teams with high variability in scoring and those with potential outliers, which can be indicative of star players or games with unusual scoring patterns.

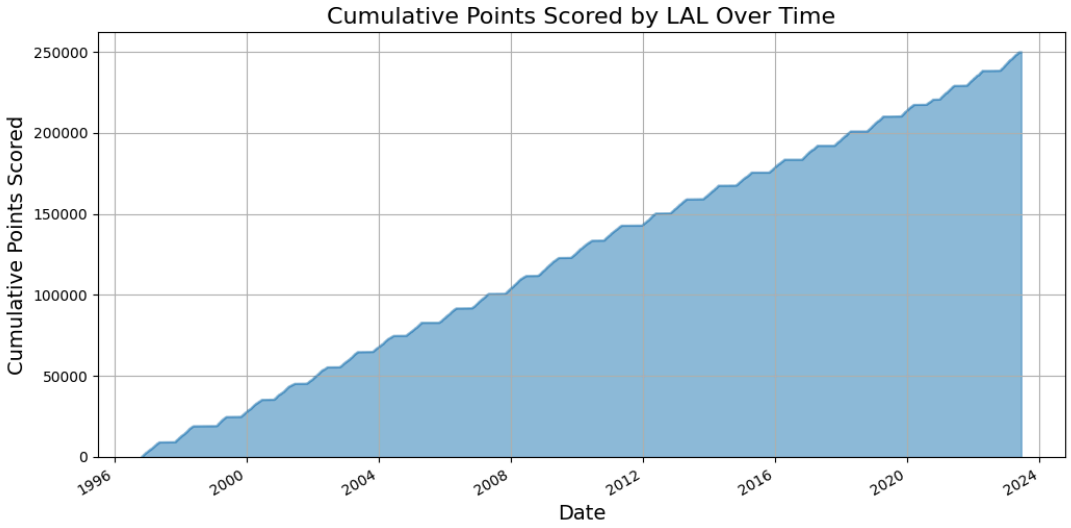

Fig 9.15

Cumulative Points Scored by LAL Over Time

This area plot illustrates the cumulative points scored by the Los Angeles Lakers (LAL) from 1996 to beyond 2020. It highlights the team's offensive performance over a long period, allowing for the observation of trends and changes in scoring patterns.

Observations:

A steady increase in cumulative points is observed, indicating consistent scoring throughout the seasons.
Periodic fluctuations in the slope of the curve suggest variations in season length, team performance, and scoring efficiency.
The sharp and consistent rise after 2010 could indicate strategic shifts, changes in player lineup, or an increased focus on offensive gameplay.
There are no significant periods of stagnation, implying the team has not faced prolonged slumps in scoring.
The plot provides a clear picture of the team's long-term scoring trajectory, valuable for assessing the team's historical performance and strategizing for future seasons.

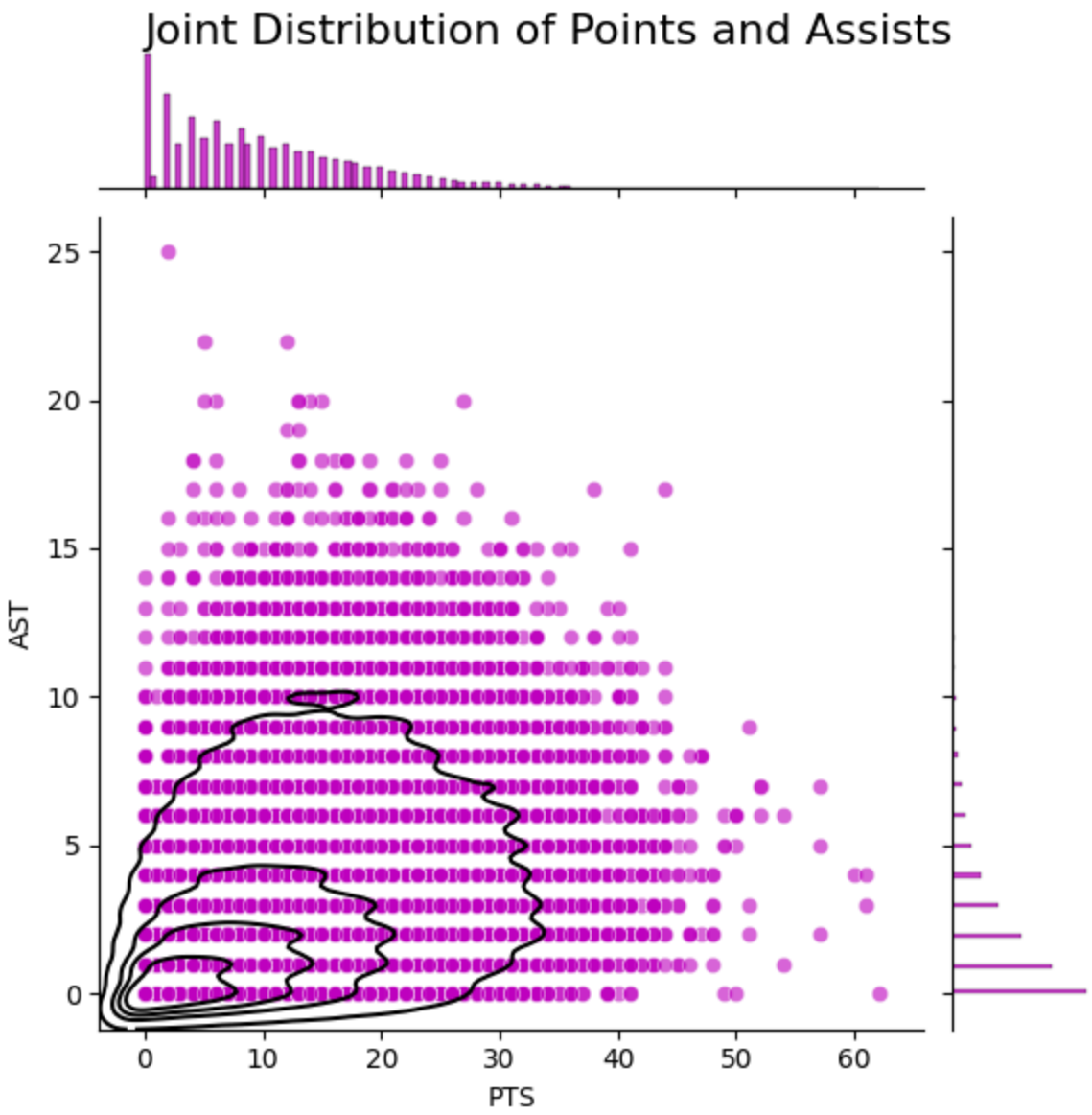

Fig 9.15

- Density: There's a high concentration of data points near the origin, indicating that most players score a low to moderate number of points and assists per game.
- Trend: As the number of points increases, the number of assists also tends to increase, although the relationship is not strictly linear.
- Outliers: There are a few instances of players with high points and assists, suggesting standout performances.

- Marginal Histograms: The histograms on the top and right show the distribution of points and assists, respectively. Most players score 10 points or fewer and make 5 assists or fewer, as indicated by the peaks in both histograms.
- The plot suggests that while scoring and assists are related, there are many players who specialize in one over the other.

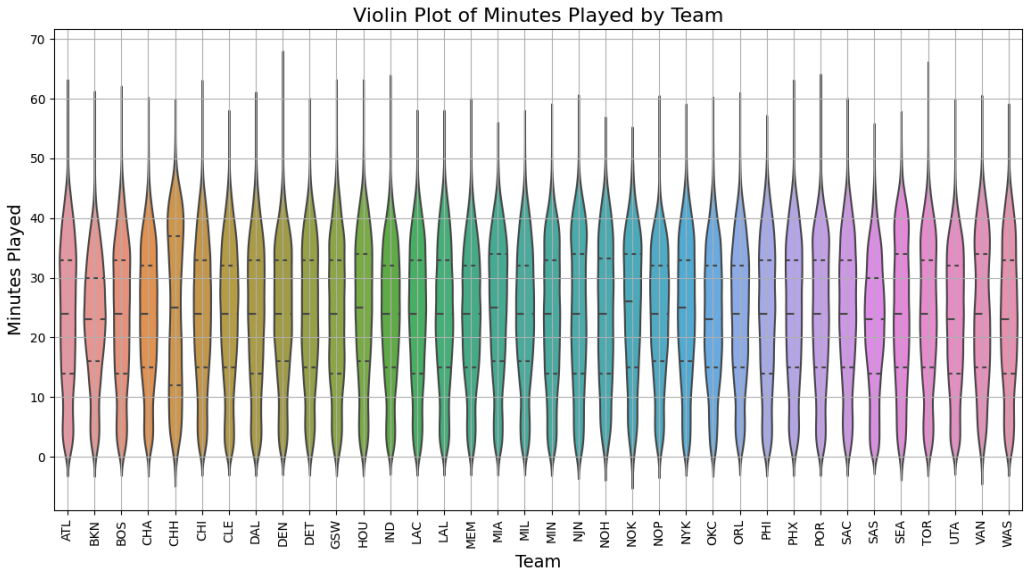

Fig 9.16

Violin Plot of Minutes Played by Team

This violin plot visualizes the distribution of minutes played by players across different NBA teams. It offers insights into the playing time allocated by each team to their players.

Observations:

The width of the violins indicates the frequency of games with certain minutes played, with wider sections showing more common durations.
Internal lines represent the quartiles, giving a sense of the spread and central tendency of the data.
Some teams show a wider range of minutes played, suggesting flexible rotation and varied player utilization.

Other teams have narrower distributions, indicating consistent playing time allocations. This analysis can be useful to identify teams with deep benches versus those relying heavily on their starting lineup, influencing strategies for rest, injury management, and player development.

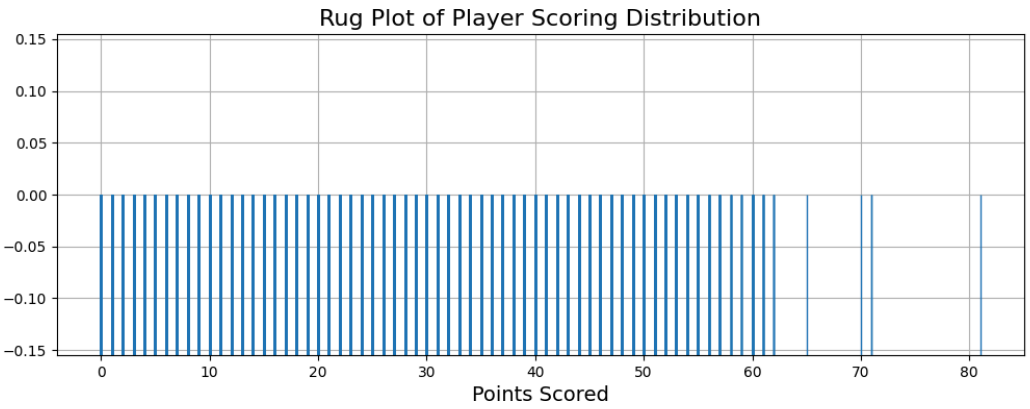

Fig 9.17

The plot shows a set of vertical lines (or "rugs") along the points scored axis, which corresponds to individual instances of player scoring in the dataset. Each rug represents a data point.

Here’s a simple interpretation:

- **Density of Rugs:** The denser clusters of rugs indicate that more scores fall within those point ranges. It looks like there's a higher concentration of scores in the lower point range, which is common in basketball scoring as not all players score high in each game.
- **Spacing Between Rugs:** Areas with wider spaces between rugs indicate fewer scores within that range. The decrease in density as the points increase suggests that high-scoring games are less common.
- **Extremes:** Rugs that appear farther out on the scale, toward higher point values, represent the exceptional high-scoring performances. These are more spread out, indicating such high scoring games are relatively rare.

The rug plot gives a visual sense of where most of the data lies and how it's spread out, which in this case is predominantly toward the lower end of the scoring range.

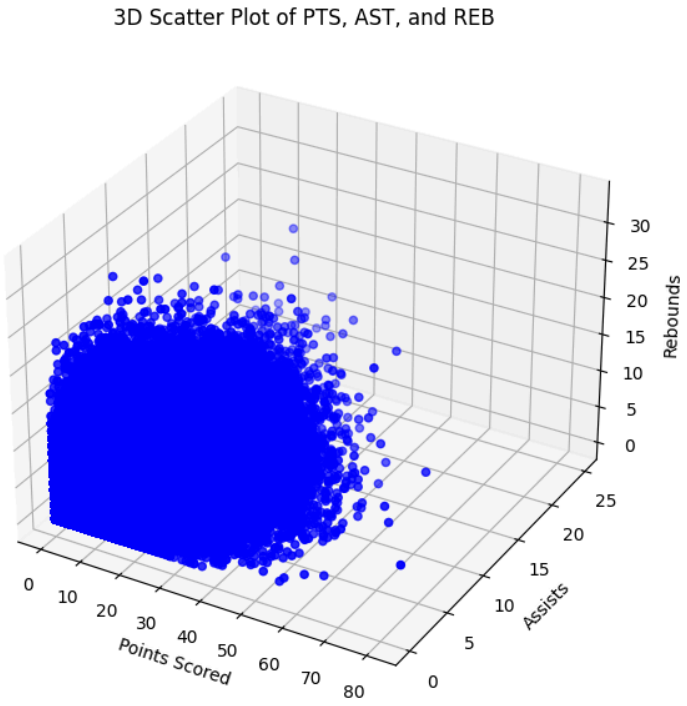

Fig 9.18

- Points scored are displayed along the x-axis, assists on the y-axis, and rebounds on the z-axis, allowing for the observation of patterns and outliers in the data.
- A dense cluster suggests that a majority of the player performances are within a specific range for these three metrics.
- The distribution and density of the points may indicate common performance profiles among players, with rarer occurrences of high values in all three dimensions.
- This plot can be instrumental in identifying players with exceptional all-around performances.

Utilizing this visualization, teams and analysts can pinpoint key contributors and strategize for matchups and player development.

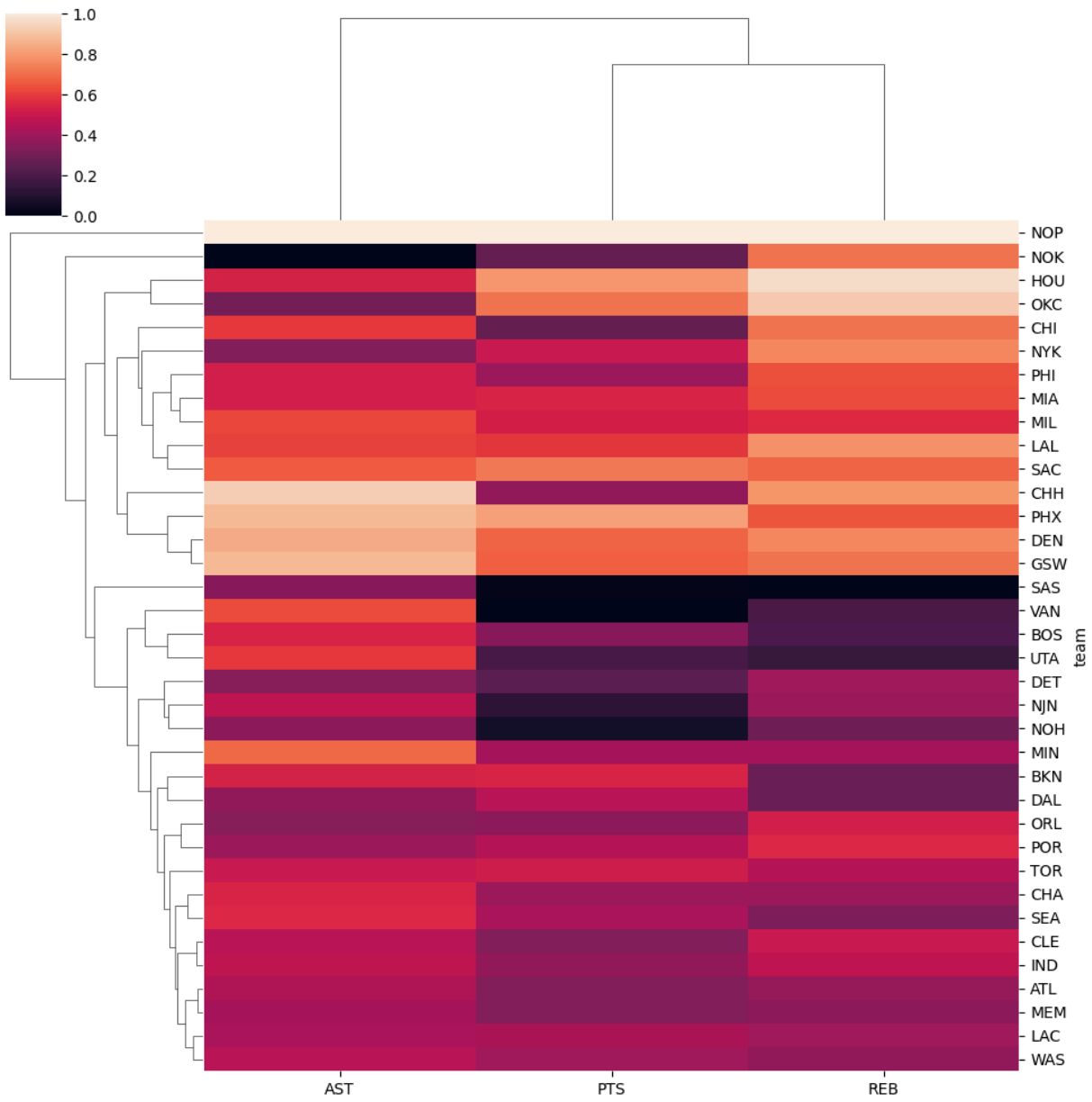

FIg 9.19

Cluster Map of Team Performance Metrics (PTS, AST, REB)

The cluster map synthesizes the average points (PTS), assists (AST), and rebounds (REB) of NBA teams into a single visual framework. This visualization employs hierarchical clustering to reveal natural groupings and performance similarities between teams based on the three selected metrics.

Key Observations:

Teams are clustered based on their performance in points scored, assists, and rebounds, with the color intensity indicating the level of performance after standardization.
Similar color patterns across the three metrics suggest a comparable level of performance amongst grouped teams.
The dendrogram on the left side and top of the heatmap illustrates the hierarchy of similarities, indicating which teams have a similar style or effectiveness in their gameplay.
This analysis could be beneficial for identifying strategic patterns, understanding team dynamics, and comparing team performances in a multi-dimensional metric space.
The cluster map is not just a snapshot of averages; it offers a deeper understanding of how teams are positioned in relation to each other within the league, providing valuable insights for coaching strategies and team building.

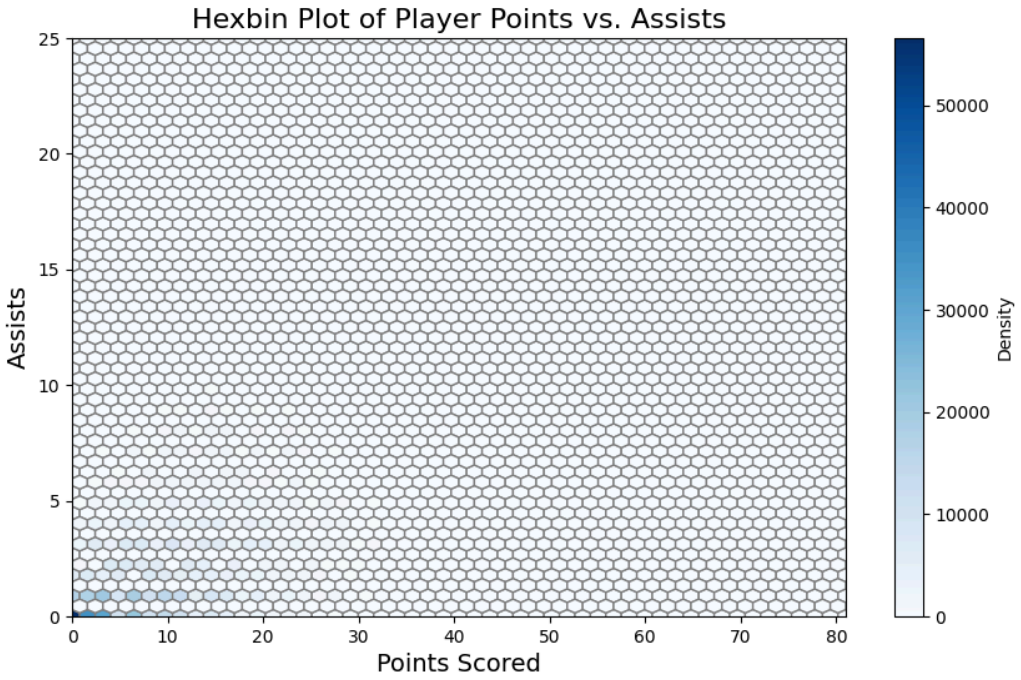

Fig 9.20

- **Clusters of Points:** The vertical lines of dots for each team indicate the number of points scored by players in different games. Where you see a lot of dots clustered around a certain point value, it means many players scored that number of points while playing for that team.

- **Variability:** Teams with dots spread out over a wider range of the y-axis indicate greater variability in the number of points scored by players. A tight cluster would suggest more consistency in scoring.

- **Outliers:** Any dots that appear far away from the main clusters could be considered outliers, representing unusually high or low scoring performances compared to the rest.

The plot provides a quick way to compare scoring between teams and to see the distribution and range of player points within each team.

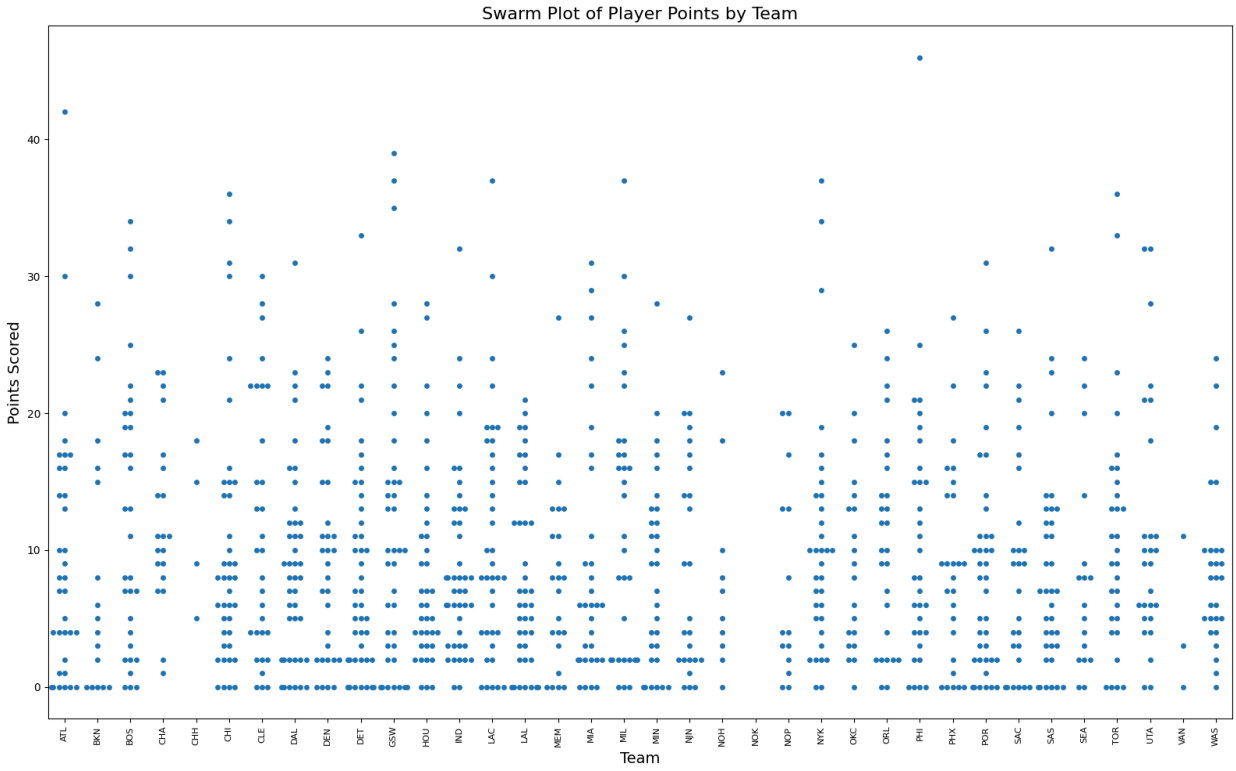

FIg 9.21

- Distribution Patterns: The plot shows the spread of points scored among players, with the majority clustering around the lower to mid-range of the scale, indicating that high-scoring games are less common.
- Team Comparisons: The distribution across teams can be compared to identify which teams have more high-scoring players. A team with more points distributed towards the top of the y-axis typically indicates stronger offensive performances.

- Outliers: Individual points that appear separate from the main clusters represent exceptional scoring performances. These outliers could signify star players or anomalous high-scoring games.
- Consistency and Variability: Teams with tight clusters have more consistent player performances, while those with a wider spread indicate greater variability in points scored per player.

In summary, the swarm plot is a useful tool to compare individual scoring performances across teams, highlight star players, and assess the consistency of team offensive strategies.

10. SUBPLOTS

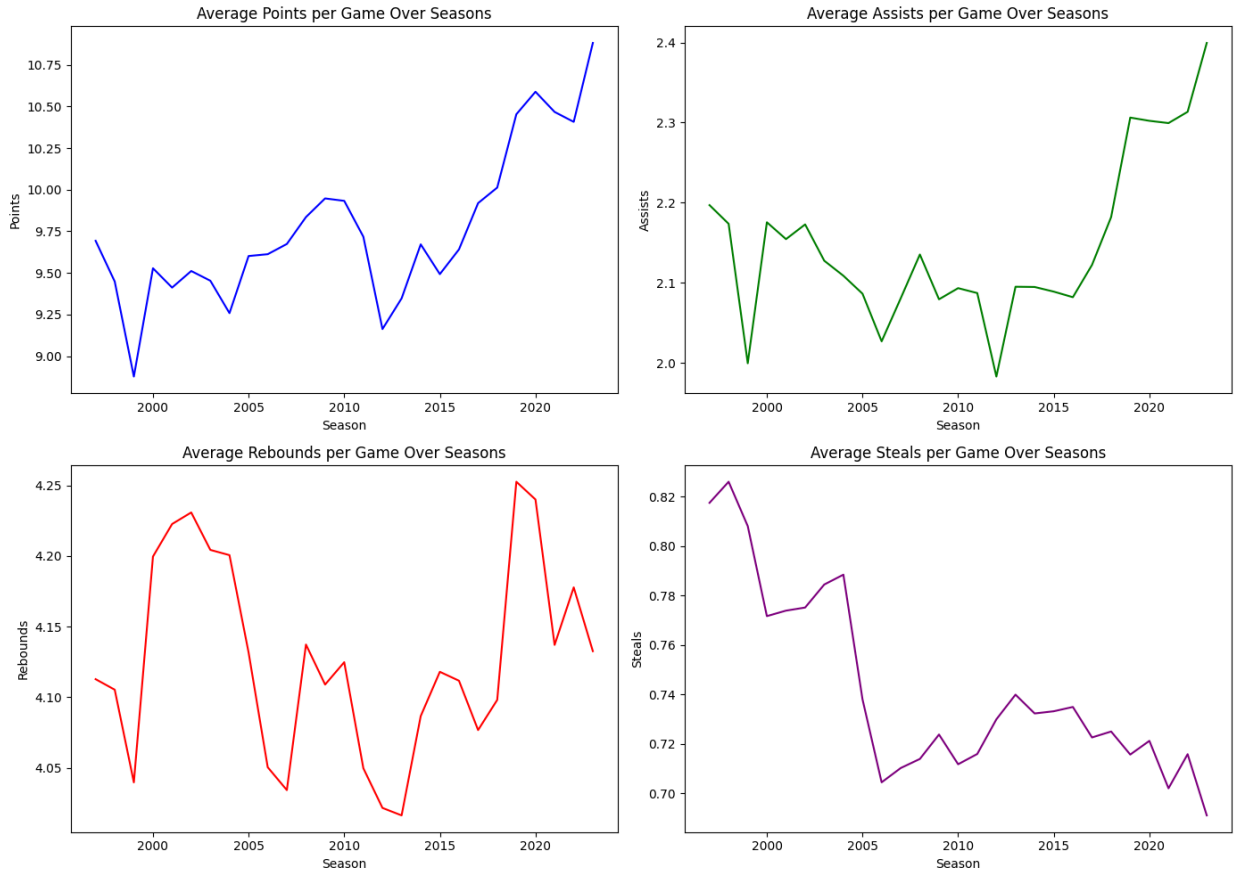

Fig 10.1

- Average Points per Game Over Seasons (Top Left): There is an upward trend, especially noticeable from the mid-2000s onwards, suggesting an increase in scoring efficiency or a change in offensive strategies that favor higher scoring.

- Average Assists per Game Over Seasons (Top Right): After some fluctuation in the early 2000s, there's a significant upward trend from the 2010s onwards, indicating a possible shift towards more team-oriented or passing-focused basketball.
- Average Rebounds per Game Over Seasons (Bottom Left): This graph shows some variability, with no clear long-term trend, suggesting that rebounding rates may be influenced by other factors such as changes in player roles or physical attributes.
- Average Steals per Game Over Seasons (Bottom Right): There's a notable downward trend, particularly sharp in the early 2000s, which then levels off. This could reflect changes in defensive playstyles, rule changes, or better ball handling and passing by offensive players, reducing the opportunities for steals.

Together, these graphs can help analysts and fans understand how the dynamics of the game have evolved over time, reflecting changes in rules, player skills, coaching strategies, and the physical and tactical development of the sport.

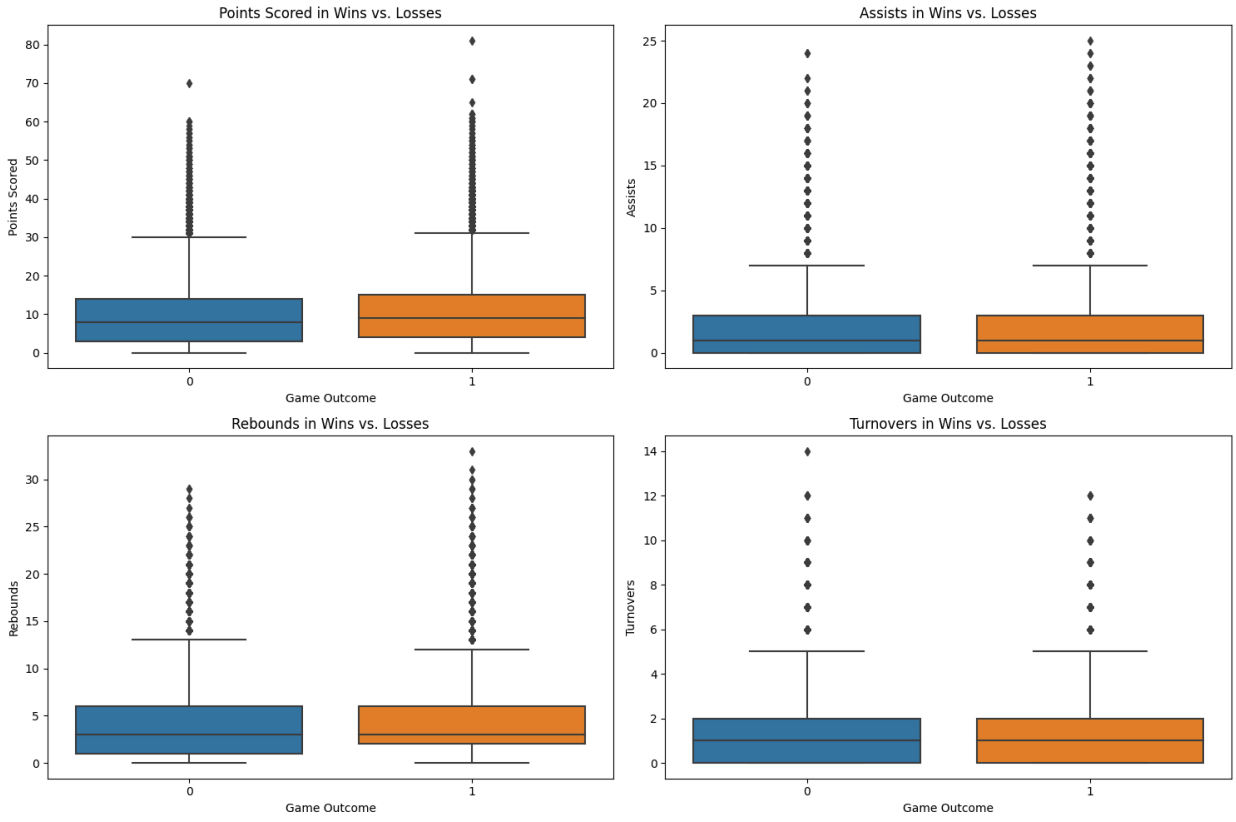

Fig 10.2

The set of box plots illustrates the distribution of points scored, assists, rebounds, and turnovers for games won (indicated by '1') versus games lost (indicated by '0').

Observations:

- Points Scored in Wins vs. Losses (Top Left): Higher median and wider distribution in points scored are observed in games won, indicating scoring is a significant factor in winning games. Outliers suggest exceptional scoring performances in some games.
- Assists in Wins vs. Losses (Top Right): Similar to points scored, a higher median of assists is seen in games won, implying that successful passing and team coordination are key components of a winning strategy.
- Rebounds in Wins vs. Losses (Bottom Left): Wins are also associated with a higher median of rebounds, signifying the importance of controlling the ball after a shot attempt.
- Turnovers in Wins vs. Losses (Bottom Right): The median for turnovers is slightly higher in losses, although the difference is not as pronounced. This suggests that while turnovers are undesirable, they may not be as critical to the game's outcome as points, assists, and rebounds.

The visual evidence from these plots supports the commonly held belief that effective scoring, passing, and ball control contribute significantly to a team's chances of winning a game. Conversely, turnovers have a negative impact but may be mitigated by strong performances in other areas.

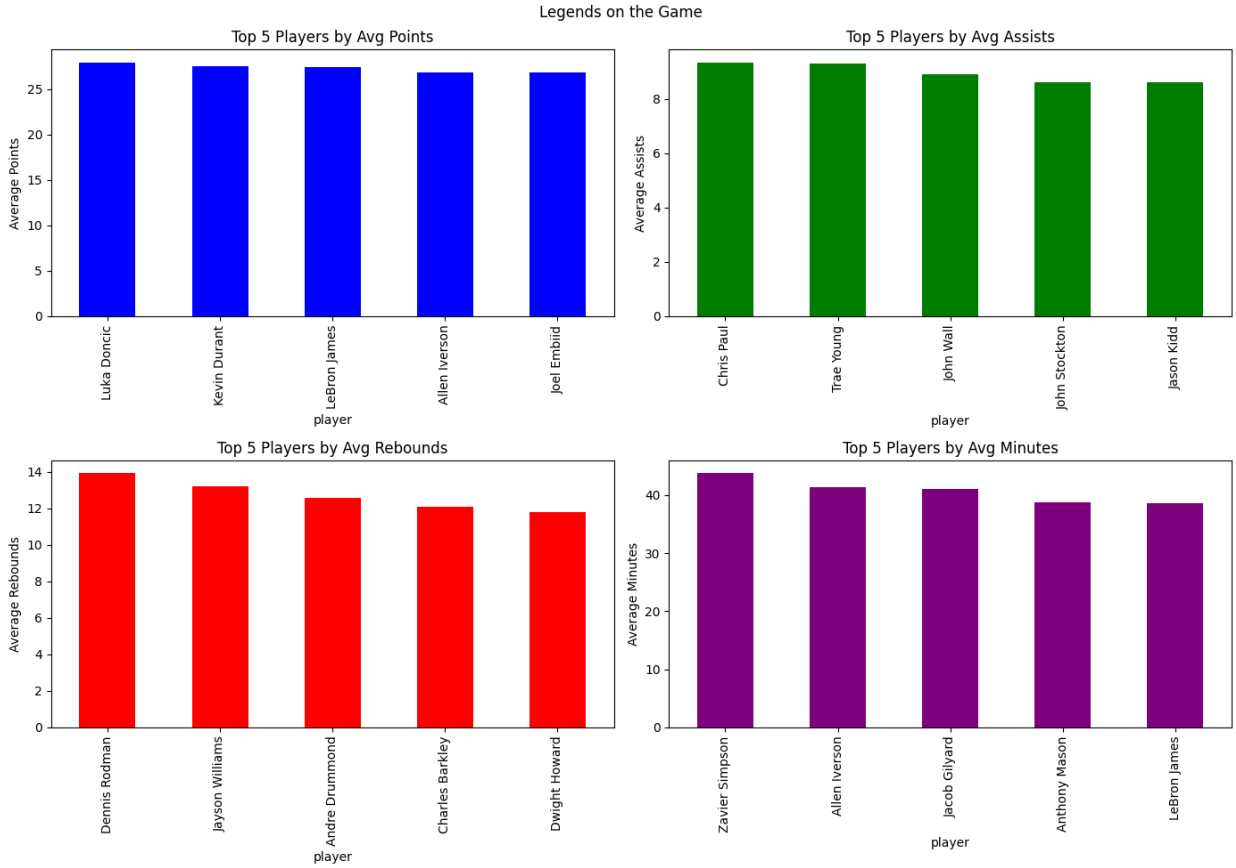

Fig 10.4

Legends on the Game: Top 5 NBA Players by Key Performance Metrics

- Top 5 Players by Average Points: This bar chart, represented in blue, showcases the players who, on average, scored the most points per game. This metric is a direct indicator of offensive prowess and scoring ability.
- Top 5 Players by Average Assists: Displayed in green, this graph highlights the players who averaged the highest assists per game, a testament to their vision on the court and ability to facilitate offensive play.
- Top 5 Players by Average Rebounds: The red bars represent the leading players in the average number of rebounds per game, emphasizing their defensive acumen and control over the boards.
- Top 5 Players by Average Minutes: In purple, this chart indicates the players with the highest average playing time per game, reflecting their endurance, coach's trust, and possibly their role as key players on their teams.

The players featured in these graphs exemplify elite performance and contribution to their teams' success. Such a multi-dimensional view helps to identify not only the high scorers but also the most effective players in terms of game presence and teamwork.

The illustrated top players are notable figures in the NBA, and their statistical leadership across these metrics solidifies their status as legends of the game. These insights are invaluable for teams and analysts alike to strategize and evaluate talent effectively.

11. TABLES

	team	PPG	Total_Wins	Playoff_Qualifications	Total_Games
0	ATL	99.8527	1034	16	2206
1	BKN	106.272	433	8	937
2	BOS	101.337	1236	18	2312
3	CHA	100.195	614	3	1532
4	CHH	95.3636	285	5	495
5	CHI	98.2893	1069	14	2195
6	CLE	99.3599	1113	11	2220
7	DAL	102.42	1220	17	2226
8	DEN	104.167	1123	15	2197
9	DET	97.9215	1049	12	2192
10	GSW	105.099	1120	10	2229
11	HOU	102.568	1187	16	2212
12	IND	100.144	1186	17	2242
13	LAC	101.581	1038	12	2181
14	LAL	103.496	1320	18	2323
15	MEM	100.483	882	13	1788
16	MIA	98.7289	1364	21	2346
17	MIL	101.912	1090	15	2196
18	MIN	101.47	909	10	2116
19	NJN	94.7611	528	6	1243
20	NOH	94.3612	308	4	670
21	NOK	94.1707	77	0	164
22	NOP	108.577	372	3	822
23	NYK	98.6361	955	11	2160
24	OKC	106.186	726	10	1310
25	ORL	99.2988	986	12	2155
26	PHI	99.4961	1038	15	2207
27	PHX	104.842	1161	13	2203
28	POR	100.709	1140	17	2190
29	SAC	103.283	941	8	2128
30	SAS	100.877	1463	21	2325
31	SEA	98.7831	472	5	913
32	TOR	101.185	1083	13	2188
33	UTA	101.123	1261	17	2233
34	VAN	92.336	86	0	378
35	WAS	101.209	917	10	2140

"The table illustrates comprehensive performance metrics across multiple NBA teams. Points Per Game (PPG) is a measure of offensive strength, while Total Wins indicates overall success across seasons. Playoff Qualifications denote the number of seasons a team has made it to the playoffs, serving as an indicator of consistent performance under pressure. Lastly, Total Games reflect the endurance and experience of the teams. This data is pivotal in assessing team strategies, success rates, and long-term performance trends in the league."

12. DASHBOARD

The interactive dashboard focuses on 3 important aspects of the NBA i.e.,
Player Analysis,
Team Analysis and,
Overall Trend Analysis.

Below are some of the snapshots (not exhaustive) of the dashboard.

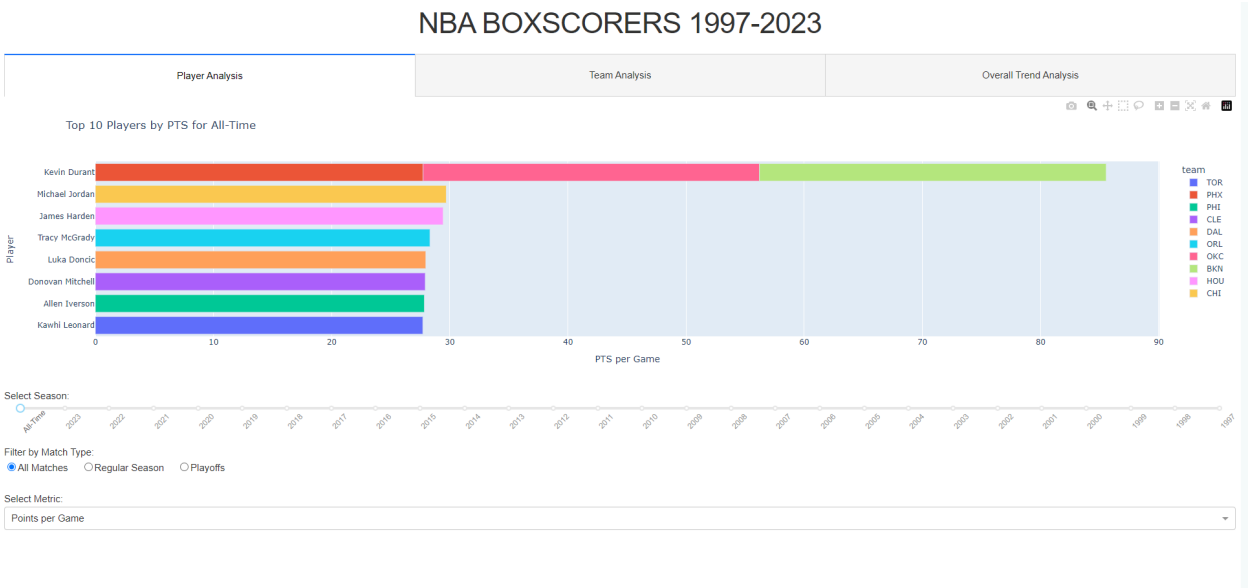

Fig 12.1

All- time Top 10 Players by Points

Each tab corresponds to a different section of the analysis. It encompasses the interactive plots generated by Plotly and integrated using Dash framework.
The dashboard can be accessed via the link provided at the end, which was deployed on Google Cloud Platform.

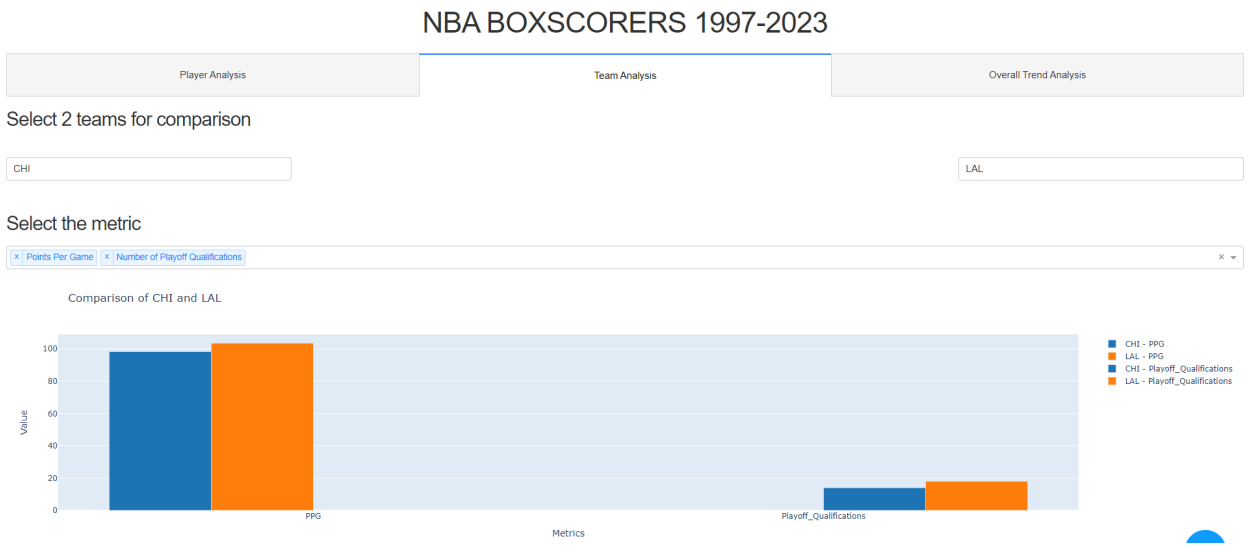

Fig 12.2

Comparison between Chicago Bulls and LA Lakers

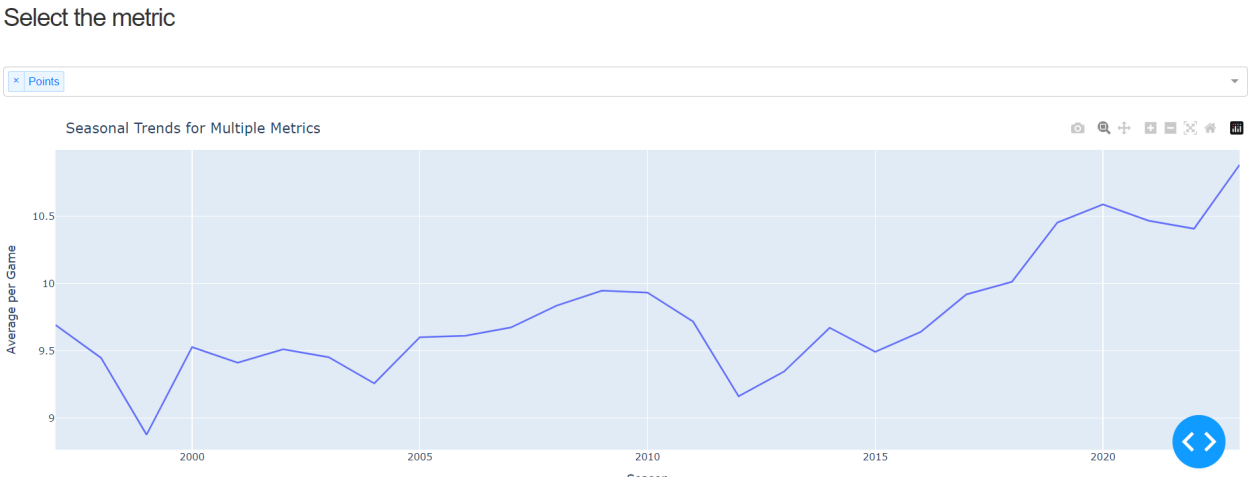

Fig 12.3

Average points per game increasing throughout years

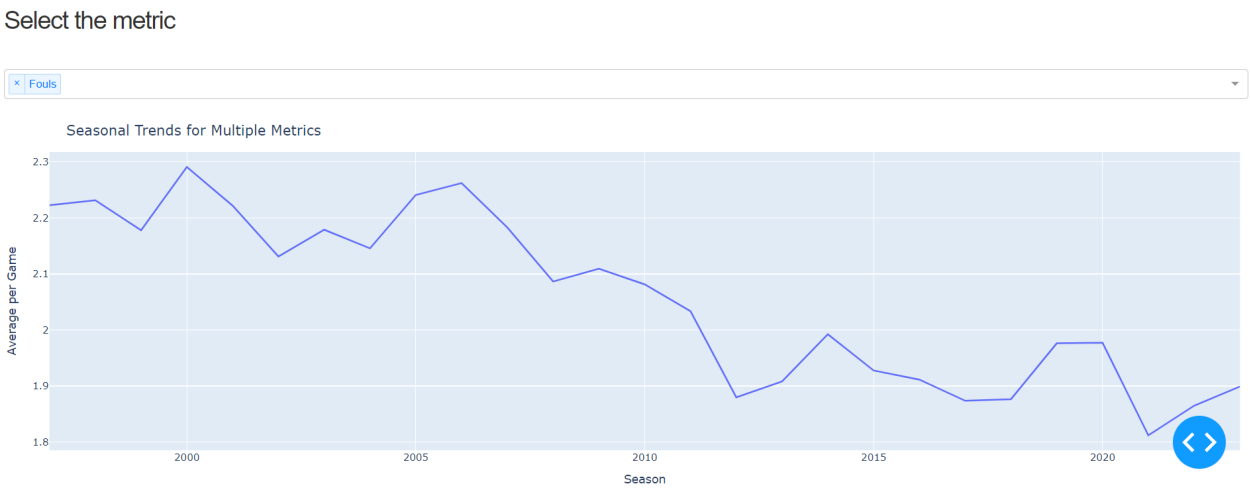

Fig 12.4
Average fouls per game decreasing throughout years

13. CONCLUSION:

a. Insights Gained from Created Graphs:

Throughout this project, the analysis of various graphs has yielded a multi-dimensional understanding of NBA player and team performance metrics. The histograms and KDE plots revealed the distribution of individual statistics such as points, assists, and rebounds, illustrating common performance thresholds and highlighting outliers. Notably, the distribution of points scored by players showed a skew towards lower values with fewer occurrences of high-scoring instances, indicative of the varying roles players hold within a team.

- The scatter and 3D scatter plots served to visualize relationships and potential correlations between different performance metrics. For instance, the relationship between points, assists, and rebounds was mapped in a 3D space, revealing clusters that suggested common player performance profiles and exceptions signifying extraordinary individual achievements.

- Boxen and violin plots offered insights into the spread and density of player minutes and points across different teams, indicating the diversity of playing styles and strategies employed by different franchises.
- Bar and stacked bar plots contrasted various metrics such as average points per game over seasons and wins versus losses, providing an at-a-glance historical perspective of performance and success trends within the league.
- The pie charts delineated proportions of categorical data such as home versus away wins, presenting an understanding of potential home-court advantages.
- The area and cumulative plots tracked the progression of total points over time for teams, particularly useful for spotting trends and significant shifts in team performance, possibly relating to changes in team composition or management strategies.
- Finally, the cluster map synthesized team statistics, clustering teams with similar performance metrics together, which could be used to identify teams with similar playing styles or performance levels.

b. Python Dashboard Utility:

The Python dashboard acts as a dynamic and interactive interface that enables users to explore and interact with the dataset in a user-friendly manner. Users can:

- Customize their view by selecting different metrics and seasons, allowing for personalized data exploration tailored to their specific interests or questions.
- Identify trends over time or across different teams and players through interactive graphs that update based on user inputs.
- Drill down into details with the ability to select individual teams or players for more granular analysis, aiding in scouting or team analysis efforts.
- Access a broader narrative by viewing data through different lenses—whether it's through historical trends, player performance, or team achievements.
- Make informed decisions with a comprehensive understanding of the data facilitated by the dashboard, supporting analysts, fans, and decision-makers in the basketball community.

The dashboard serves as a powerful tool for storytelling with data, turning complex datasets into comprehensible visual narratives that can inform strategy, player development, and engagement with the sport.

APPENDIX:

Python codes attached will be uploaded via Blackboard.

REFERENCES:

(1) Data Analysis and Visualization Using Python, 1 st Edition Author(s): Dr. Ossama Embarak ISBN-13: 978-1484241080
(2) Python for Data Analysis , 2 nd Edition Author(s): Wes McKinney ISBN-13: 978-491-95766-0
(3) Python Data Visualization Cookbook , 2 nd Edition Author(s): Igor Milovanovic ISBN13: 978-1-78216-336-7
(4) Practical Tableau , 1 st Edition Author(s): Ryan Sleeper ISBN-13: 978-149219773

Link for the Dashboard :

https://dashapp-37ytcez3sq-ue.a.run.app/