College of Professional Studies: Northeastern University

ALY 6140 – Python & Analytics Systems Technology

Instructor: Dr. Harpreet Sharma

Academic Term: Winter 2025

Analysis of NBA Players

Moneyball Approach

Submitted By:

Group 2

Nahimah Suglo
Tawiah Ayeni Micheal
Sheila Kwartemaa Boateng

February 10, 2025

Introduction

In today's world, sports like football, baseball, and basketball have become deeply integrated into entertainment and analytics. Basketball, in particular, dominates the sports landscape in the United States, with the NBA serving as a global hub for elite competition. In the ever-evolving world of professional basketball, data analytics has become a game-changer, reshaping the way teams evaluate talent, make strategic decisions, and build rosters. Much like the "Moneyball" revolution in baseball, the NBA has shifted toward data-driven decision-making, where advanced metrics influence scouting, player development, and in-game strategies.

In this project, we aim to explore key performance metrics that define great players and analyze their career trajectories. By evaluating statistical trends and applying machine learning models, we seek to uncover insights into player efficiency, career progression, and draft potential. Our goal is to bridge raw statistics with strategic decision-making, helping to identify undervalued players, predict long-term success, and provide actionable insights for team management.

Business Questions

1. Identifying Undervalued Talent

Inspired by the Moneyball approach, we aim to identify players who are overlooked but possess strong potential. We focus on players who excel in scoring efficiency, playmaking, and versatility but may be underappreciated in traditional scouting. This analysis can help teams find hidden gems who could thrive in the right system.

2. Predicting Player Efficiency

Using Random Forest, we predict True Shooting Percentage (TS%), a key measure of a player's scoring efficiency that accounts for field goals, three-pointers, and free throws. By leveraging past performance data, we aim to determine which players are most efficient and what factors contribute to their success.

3. Career Projection Based on Age

Player performance naturally evolves over time, but not all players decline at the same rate. Using Gradient Boosting, we analyze how a player's career unfolds based on:

- Age: A primary factor in performance trends.
- Previous Season's Stats: Recent form and consistency.
- Two Seasons Ago Stats: Long-term trajectory and sustainability.

We focus on Points (PTS), Rebounds (REB), and Assists (AST), as these stats often define a player's role and value. This model will help teams anticipate when a player may peak, plateau, or decline.

4. Draft Prediction Using First-Game Metrics

Traditionally, draft projections rely on college or international performance, but in our case, we lacked pre-draft data. Instead, we analyzed players' first professional game metrics to predict draft selection. We examined:

- Scoring efficiency (PTS, FG%)
- Playmaking ability (AST, TO ratio)
- Defensive impact (REB, STL, BLK)
- Overall productivity in limited minutes

  This approach provides insights into how a player's early performance correlates with their draft position and whether teams make the right selections based on early indicators.

Data Overview

The dataset used for this analysis is sourced from Kaggle and includes detailed statistics from various basketball leagues worldwide. The dataset comprises player performance data, including metrics such as games played (GP), minutes played (MIN), field goals made (FGM), field goals attempted (FGA), three-point field goals made (3PM), height, weight, and other relevant statistics. In total the **dataset** contains **53,949 rows** and **34 columns**.

Below is a breakdown of the key columns:

- **League**: Indicates which league the player is associated with (e.g., NBA, CBA, KLS).
- **Season**: The year of the basketball season.
- **Stage**: Whether the data corresponds to the regular season, playoffs, or other stages.
- **Player**: The player's name.
- **Team**: The team the player played for.
- **Performance Stats**: Includes metrics such as **PTS** (Points), **REB** (Rebounds), **AST** (Assists), **STL** (Steals), **BLK** (Blocks), and more.
- **Birth Date**: The player's birth date, which can be used to calculate age.
- **Height & Weight**: Physical attributes of the players.
- **Draft Information**: Includes draft round, draft pick, and draft team.

To ensure a comprehensive understanding of player performance and dynamics, we concentrated on players with connections to the NBA while incorporating their records from other leagues. This approach allowed us to evaluate the influence of international experience on NBA performance and the potential of undrafted players. The dataset was filtered into three main categories:

- NBA Players: All players who have participated in the NBA at any point in their careers. This group forms the core of our analysis, as the NBA's draft system and competitive environment are central to our study.
- NBA Players Who Have Played in Other Leagues: This group includes NBA players who also competed in international leagues before or after their time in the NBA. By including these records, we aim to understand how experiences in different competitive environments impact player development and performance.

- U.S.-Born Players Who Played Exclusively in International Leagues: This category focuses on American-born players who never played in the NBA but pursued professional basketball careers overseas. Including this group helps us analyze undrafted talent and explore how these players might have performed if given an opportunity in the NBA.

Data Preparation and Cleaning

Filtering the Data Our primary focus was to gather records of both **international and local players** who played in the NBA, as well as **U.S.-born players** who may or may not have played in the league. To ensure a structured analysis, we filtered and prepared the data by categorizing players into three main groups:

1. **NBA Players** – All players who have participated in the NBA, regardless of nationality.
2. **NBA Players Who Have Played in Other Leagues** – Players who had careers in the NBA but also competed in international leagues.
3. **U.S.-Born Players Who Played Exclusively in International Leagues** – Players born in the United States who never played in the NBA but pursued professional careers in international leagues.

```
# Filter NBA players
nba_players = df[df["League"] == "NBA"]

# Identify NBA players who have played in other leagues
international_players = df[df["Player"].isin(nba_players
["Player"]) & (df["League"] != "NBA")]

# Identify U.S.-born players in international leagues
us_born_international_players = df[(df["League"] != "NBA")
& (df["nationality"] == "United States")]

# Merge NBA and International Data for the same player
#(keeping all NBA rows)
merged_df = pd.concat([nba_players, international_players, us_born_international_players])

# Reset the index of the merged DataFrame
merged_df.reset_index(drop=True, inplace=True)
```

This reduced the dataset to **21,834 rows** and **34 columns**.

Handling Duplicates in the Data

Duplicate records in a dataset are often indicative of **data quality issues**, which can arise during the data collection, import, or export processes. We identified **1,112 duplicated rows** in the dataset, where the combination of player name, season, and other performance metrics were identical.

Understanding the Duplicate Issue Duplicate rows mean that certain player-season records have been repeated in the dataset, which can distort any analysis by inflating the importance or weight of certain records. In the case of basketball data, this issue could skew performance analysis, as it might make a player appear more consistent or prolific than they actually are.

Potential Causes for Duplicates:

1. **Data Entry Errors**: Duplicates can arise when data is entered multiple times by mistake, especially during the data collection or processing phase.
2. **System Glitches**: Sometimes, technical issues during the data import or export processes (such as using different tools or systems) can lead to data duplication.

3. **Intentional Duplicates**: While it is less likely in this case, certain situations may involve repeated events, such as a player being tracked multiple times across different datasets or due to specific circumstances (e.g., tracking both regular season and playoff stats separately).

However, in the context of this analysis, **it is unlikely that the duplicates are intentional**, as each player's performance is expected to appear only once per season. Having identical records could distort key analyses, particularly in terms of aggregating statistics.

Hence we removed duplicates values removed by keeping only one record for each player-season combination, ensuring that there are no multiple entries for the same player in a given season.

Handling Data Inconsistency

The **Games Played (GP)** column contained values greater than **82**, which is higher than the typical number of games in an NBA regular season. This was likely due to the inclusion of **playoff games** or possible data entry errors. To explore the data, we decided to cap **GP** values above **82** at **82** for exploratory purposes.

The **draft_round** column showed a maximum value of **7**, which is inconsistent with the current NBA draft format, where there are only **2 rounds**. Hence we cap the draft rounds at 2

Handling Missing Values

Upon inspection, we identified several columns with missing data. The following is a summary of the columns and the number of missing values:

```
# check for missing values
for i in df_no_dup.columns:
    if df_no_dup[i].isnull().sum() > 0:
        print(f"{i}: {df_no_dup[i].isnull().sum()} ")
```

```
Team: 2
birth_year: 102
birth_month: 102
birth_date: 102
height: 1
height_cm: 1
weight: 148
weight_kg: 148
high_school: 1900
draft_round: 12529
draft_pick: 12529
draft_team: 12529
```

These missing values can arise for various reasons, such as incomplete data entry, players playing in multiple leagues, or missing player records in specific seasons. Addressing these missing values appropriately is vital for the success of our analysis.

Treatment of Missing Values **Missing values in 'Team' Values:** The **Team** column had missing values for **Brandon Penn** and **Wesley Myers**, who played in the **Serbian-KLS league (2019-2020)**. After verifying their team affiliation, we confirmed they were with **KK Kolubara** and filled the missing values with **'KK'**.

Missing values 'High School' Values: We replaced the missing **high_school** values with the placeholder **"Unknown"**. But we later drop the column entirely.

Estimating Missing 'birth_year' Values: The **birth_year** column plays an important role in understanding player trajectories and age-related performance trends. We handled missing values in the following way:

- If a player had multiple records, we filled missing birth years using the most common (mode) value within their entries.
- If the mode approach was unsuccessful, we used the **nationality** and **season** information to estimate the **birth_year**. We filled the missing values based on the most frequent **birth_year** for players from the same **nationality** and **season**.
- For players with remaining missing birth years, we used the **minimum season start year** to estimate the **birth_year** based on the player's earliest season.

Handling Missing 'Height' and 'Weight' Values

- **Height** and **Weight** values were imputed based on the **height** (height_cm) of players, using the **mean** height or weight for players with similar physical characteristics.
- If the **weight** was missing, we also converted it to kilograms if needed and imputed the missing values based on the **mean** weight of players with similar height measurements.

Additionally, the **weight** column, which was recorded in pounds, was converted to kilograms for consistency. Any remaining missing **weight_kg** values were filled using the same imputation method, based on the **height_cm** grouping.

Handling Missing Draft Information: For the **draft_round**, **draft_pick**, and **draft_team** columns, we handled missing values based on the following conditions:

- For **U.S.-born players** and players in the **NBA**, missing draft information was filled with **0** for **draft_round** and **draft_pick**, and **"Undrafted"** for **draft_team**. indicating they were not selected in the draft
- For **international players** who did not play in the **NBA**, missing draft information was filled with **0** for **draft_round** and **draft_pick**, and **"International"** for **draft_team**.

This ensured that all players had valid draft-related data, even if they were not drafted or played internationally.

Data Exploratory

In basketball, **player performance metrics** are essential for evaluating and understanding a player's contribution to the team. These metrics provide insight into different aspects of the game, from scoring and shooting efficiency to playmaking, rebounding, and defense. Some players' statistics are above average, while others fall below. In certain cases, the statistics are zero, which could imply a variety of factors. For example, a value of zero might indicate that the player did not participate in a particular category (e.g., no attempts in a specific stat like 3-point field goals or free throws), or it could suggest that the player did not accumulate any meaningful contributions in that stat during the season.

	FG%	3P%	FT%	TS%	AST_TO_Ratio	PTS_per_game	REB_per_game	AST_per_game	STL_per_game	BLK_per_game
count	21276.000000	19744.000000	21176.000000	21276.000000	21248.000000	21278.000000	21278.000000	21278.000000	21278.000000	21278.000000
mean	0.466456	0.318696	0.743086	0.557126	1.385035	11.864022	4.652969	2.288488	0.965484	0.426269
std	0.073312	0.135468	0.119853	0.063261	0.847587	5.287855	2.447668	1.707530	0.521542	0.475375
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.420635	0.279570	0.680000	0.520178	0.800000	8.120303	2.851852	1.058824	0.594203	0.104441
50%	0.458160	0.342857	0.757576	0.557482	1.240000	11.309691	4.142857	1.800000	0.878049	0.267731
75%	0.506329	0.389610	0.822222	0.595951	1.796875	14.736842	5.973684	3.071429	1.250000	0.575000
max	0.909091	1.000000	1.000000	0.900000	15.000000	41.972222	18.666667	13.272727	4.300000	4.333333

Shooting Efficiency

Metrics like **Field Goal Percentage (FG%), 3-Point Percentage (3P%), and Free Throw Percentage (FT%)** are crucial for understanding a player's scoring ability. **True Shooting Percentage (TS%)** offers a more comprehensive view of a player's shooting efficiency by considering all types of scoring attempts, including field goals, three-pointers, and free throws. High shooting efficiency is key for consistent offensive production and overall team success.

- FG% (Field Goal Percentage): The average is 46.6%, which is typical for NBA players. The best shooters (75th percentile) hit over 50.6% of their shots.
- 3P% (3-Point Percentage): The average is 31.9%, with top shooters (75th percentile) hitting 38.9% or more from beyond the arc.
- FT% (Free Throw Percentage): The average is 74.3%, with the best free throw shooters (75th percentile) making over 82.2% of their attempts.
- The average is 55.7%, which is a good efficiency rate. The top 25% of players have a TS% of 59.6% or higher, indicating excellent scoring efficiency.

Playmaking & Ball Control

The **Assist-to-Turnover Ratio (AST/TO)** is vital for measuring a player's ability to create scoring opportunities while maintaining possession. A higher ratio indicates effective decision-making and ball control, crucial for maintaining offensive flow and reducing costly turnovers.

- The average is 1.39, meaning players generally create more assists than turnovers.
- The top 25% have a ratio of 1.80 or higher, indicating excellent ball handling and decision-making.

Rebounding & Defensive Impact

Rebounds per game (REB) are an important metric for understanding a player's contribution to maintaining possession and limiting the opposing team's opportunities. Additionally, metrics like **steals (STL)** and **blocks (BLK)** highlight a player's defensive impact, particularly in disrupting the opposing team's plays and protecting the rim.

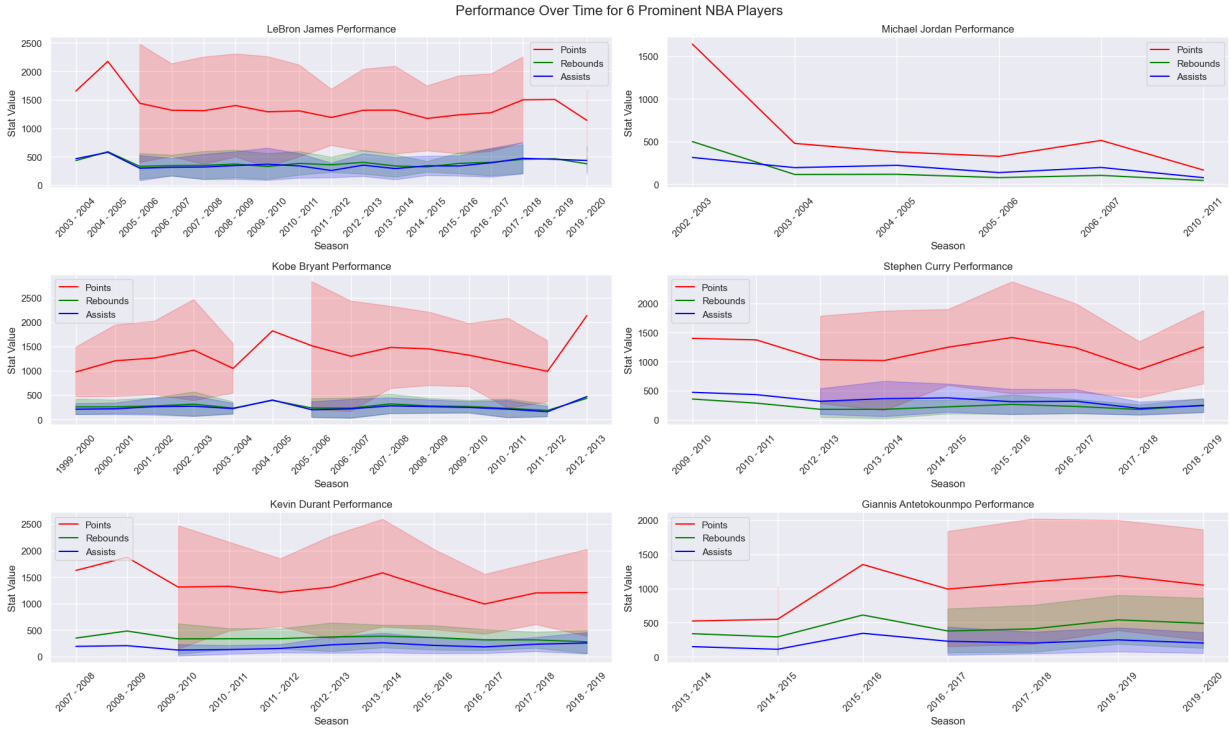
- PTS_per_game: The average is 11.86 points, with the top scorers (75th percentile) averaging 14.74 or more.
- REB_per_game: The average is 4.65 rebounds, with top rebounders grabbing 5.97 or more.
- AST_per_game: The average is 2.29 assists, with the best passers dishing out 3.07 or more.
- STL_per_game and BLK_per_game: These have lower averages (0.97 and 0.43 respectively), which is expected as these are less common events.

Performance Over Time Prominent NBA Players:

This series of line plots visualizes the performance of prominent NBA players (LeBron James, Kobe Bryant, Michael Jordan, Stephen Curry, Kevin Durant, and Giannis Antetokounmpo) over their careers, across key metrics: Points (red), Rebounds (blue), and Assists (green).

Key insights:

- **LeBron James** shows significant improvement in points throughout his career, with notable peaks in the later years (2013-2020). His rebounds and assists also show consistent growth, indicating his all-around contribution to the team.
- **Kobe Bryant's** performance follows a similar trend, with sharp peaks, particularly in points. His assists and rebounds fluctuate over time but don't show as dramatic an increase.



- **Michael Jordan's** performance had a noticeable decline after 2000, especially in rebounds and assists. His points still remain relatively stable during his prime years but decrease significantly towards the end of his career.
- **Stephen Curry** displays impressive growth in points starting from the 2012-2013 season. Rebounds and assists increase as well, though not as drastically.
- **Kevin Durant** and **Giannis Antetokounmpo** show more steady, upward trends in their points, rebounds, and assists, highlighting their growing role as they progress in their careers.

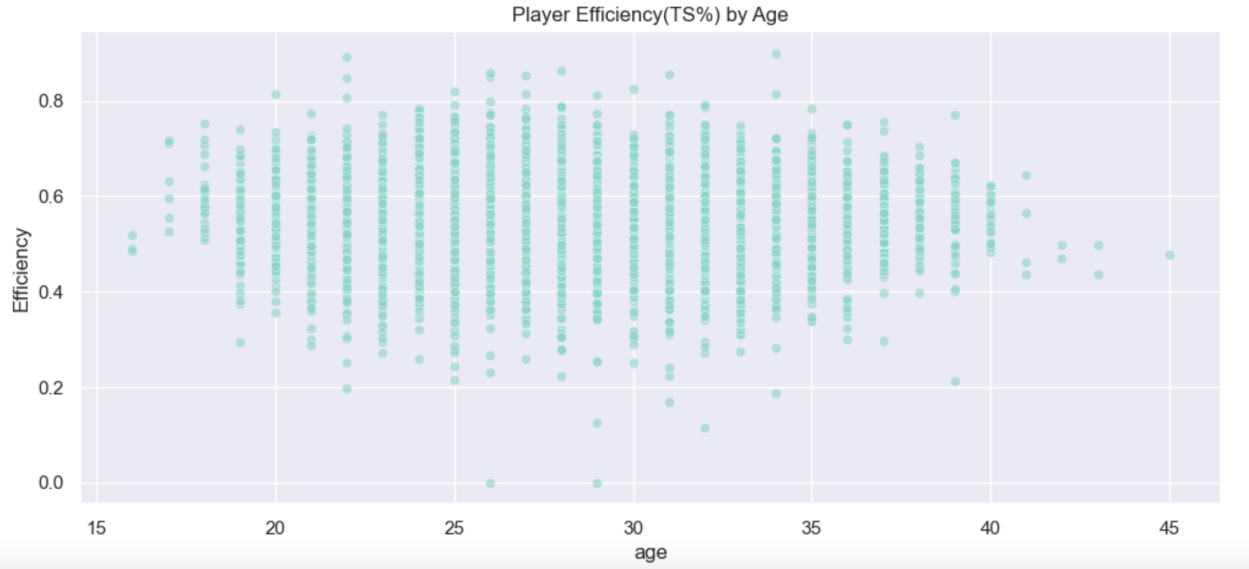
These players demonstrate how performance evolves with age and experience, with scoring typically peaking earlier than other metrics like rebounds or assists. Notably, players who excel in **True Shooting Percentage (TS%)**—a key efficiency metric—continue to dominate across different stages of their careers.

Figure 1: Lebron James, Giannis Antetokounmpo, Stephen Curry, Kevin Durant

Stephen Curry ranks at the top in **TS%** during the regular season, showcasing his elite scoring efficiency. Other standout players in this category include **Kevin Durant, Shaquille O'Neal, LeBron James, and Giannis Antetokounmpo**, all of whom have maintained exceptional efficiency while adapting their games over time. And they excede a composite score above 0.8, they consistently demonstrated efficiency, playmaking, and defense throughout their careers. For a detailed breakdown of these statistics, refer to the **Appendix**.

Players Efficiency By Age

The scatter plot, shows the relationship between a basketball player's age and their True Shooting Percentage (TS%), a measure of scoring efficiency. Players in their early careers, typically between 17 and 23 years old, exhibit greater variability in TS%, likely reflecting skill development and adaptation to the NBA. During their prime years, from approximately 24 to 32, players tend to cluster more densely within a TS% range of roughly 50% to 65%, suggesting a stabilization of performance. As players progress into their later careers, from ages 33 to 45, there appears to be a slight decrease in TS% for some, coupled with increasing variability, potentially due to physical decline or evolving roles within their teams.



Change in Stats

Some NBA players are international talents who transition to the NBA, while others continue their careers in international leagues after leaving the NBA. This shift in playing environments can lead to significant changes in player statistics.

```
Average Change in Stats - International to NBA Transition:
PTS_change    126.153846
REB_change     46.692308
AST_change     34.769231
STL_change      5.923077
BLK_change      4.461538
MIN_change    226.992308
dtype: float64

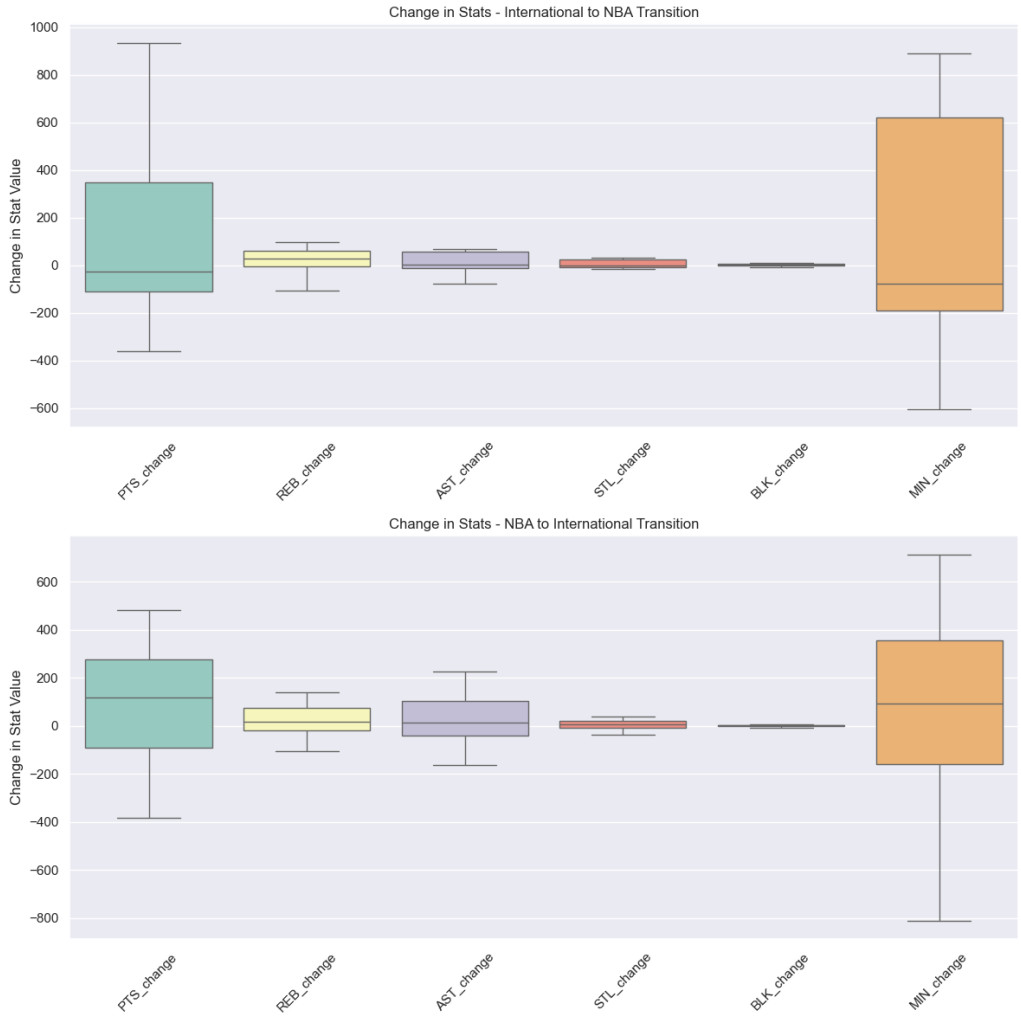
**********************************************************************************************
Average Change in Stats - NBA to International Transition:
PTS_change     89.800
REB_change     26.900
```

```
AST_change     26.550
STL_change      7.250
BLK_change      4.650
MIN_change    113.615
```

International to NBA Transition:

For players who moved from other country to play in the NBA, the result shown that;

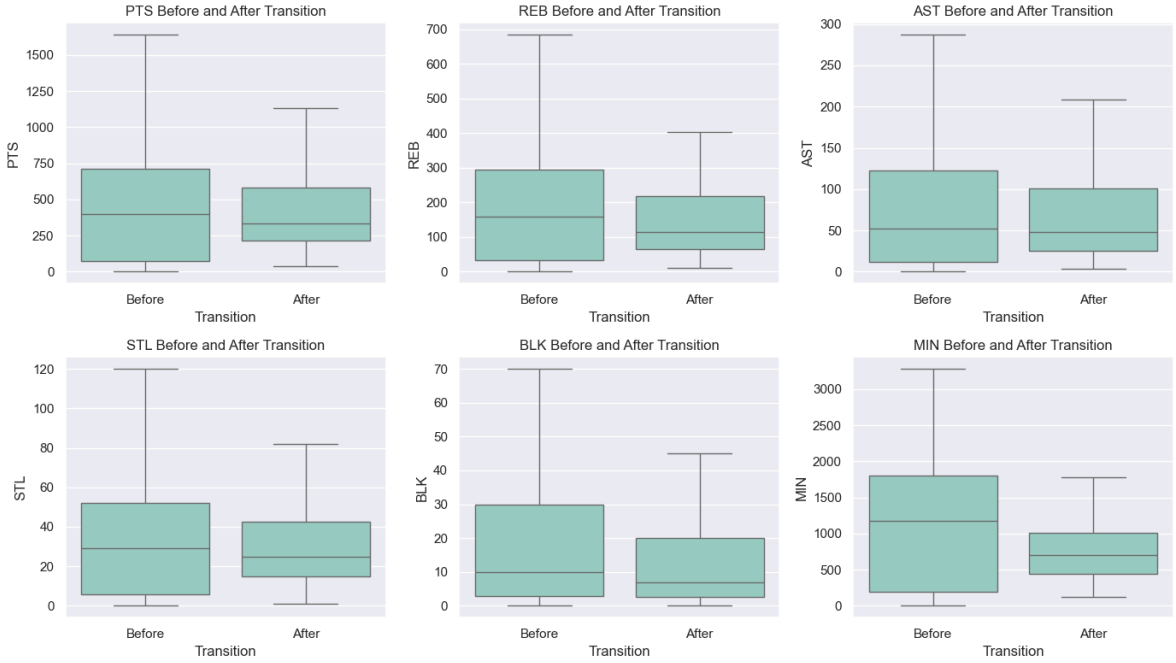
- **PTS Change**Players transitioning from international leagues to the NBA show a large spread in points change, with some players experiencing a significant increase in points.
- **REB Change:** The change in rebounds is much less variable than points, though some players do experience a drop or moderate increase.
- **STL, BLK, and MIN:** There is a noticeable decline in steals and blocks upon entering the NBA, possibly due to the more physical and competitive environment. Minutes played (MIN) also drops in the transition, which could be linked to the role players take on in the NBA.



NBA to International Transition:

- **PTS Change:** Players who move from the NBA to international leagues generally experience a decline in points, although there are outliers who maintain or improve their performance.

- **REBs and AST:** Like points, rebounds and assists also decrease in many cases.
- **STL, BLK, MIN:** These metrics show a slight increase for some players transitioning to international leagues, which could reflect the more prominent role these players take on in their teams abroad.



These findings indicate that international players generally see significant improvements when transitioning to the NBA, especially in scoring, rebounding, and assists. The NBA's pace, coaching, and playing conditions may contribute to these improvements, allowing international players to elevate their overall performance. However, the magnitude of change might vary depending on a player's role, adaptation period, and playing style.

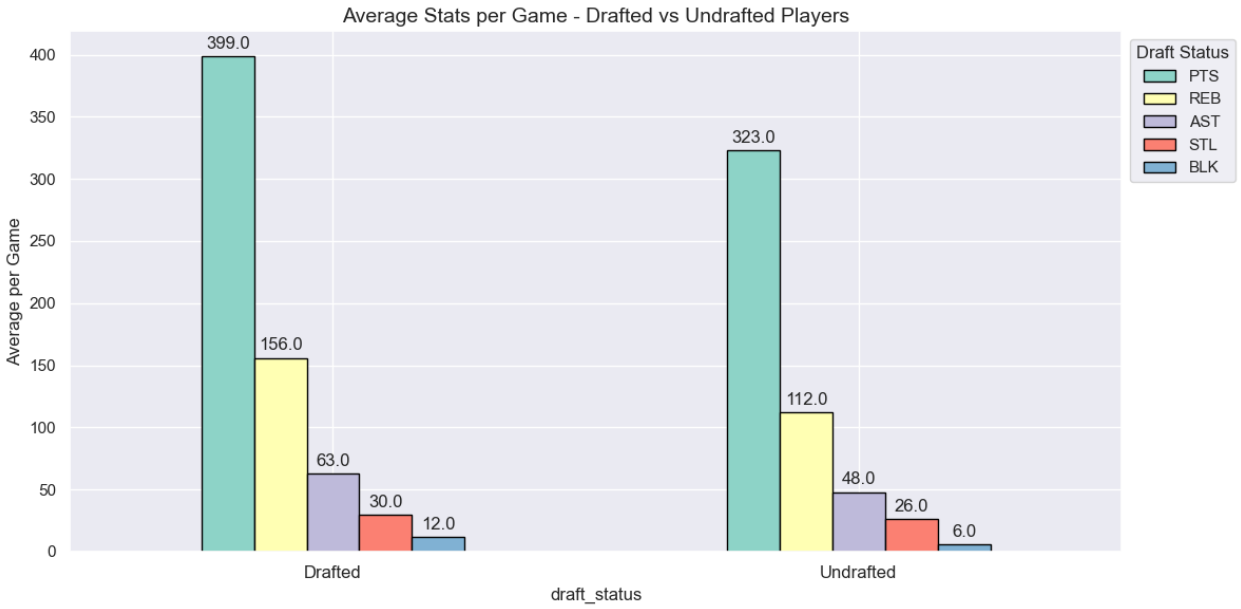
On a broader context: International players have reshaped the NBA with diverse playing styles, emphasizing teamwork, ball movement, and versatility. Stars like Nikola Jokić, Giannis Antetokounmpo, and Luka Dončić have demonstrated that international players can not only adapt but dominate in the NBA. The data supports the hypothesis that international players often improve their performance when transitioning to the NBA https://www.basketballnews.com/stories/an-inside-look-at-how-international-players-are-taking-over-the-nba

Drafted Vs Undrafted

In the NBA, **drafted players** are those selected through the league's annual draft process, typically receiving guaranteed contracts and structured development opportunities. **Undrafted players**, on the other hand, enter the league without a draft selection, often having to prove themselves through training camps, summer leagues, or international play.

Drafted players generally outperform undrafted players across key metrics, including points (PTS), rebounds (REB), assists (AST), steals (STL), and blocks (BLK). Specifically, drafted players record a median of 399 points, 156 rebounds, 63 assists, 30 steals, and 12 blocks per game, whereas undrafted players have a median of 323 points, 112 rebounds, 48 assists, 26 steals, and 6 blocks.

The differences in performance can largely be attributed to opportunities and role expectations. Drafted players typically have a higher chance of securing significant playing time, receiving more structured development, and being positioned for key roles within their teams. Conversely, undrafted players often have to fight for roster spots and may receive fewer minutes, affecting their statistical output.



Undervaluing Players: A Moneyball Approach

Given the metrics above, we wanted to identify players who may be undervalued in the NBA, especially those whose performance metrics suggest they could contribute more if given the opportunity. This aligns with the Moneyball approach, where teams seek players who are overlooked by conventional evaluations but excel in key statistical areas.

Undervalued Scorers

Players like **D.J. Richardson** (0.345 PTS/MIN) and **Mike Myers** (0.324 PTS/MIN) exhibit exceptional scoring efficiency with limited minutes, making them potentially valuable assets if utilized more. Richardson, for example, scored 21.5 points in just 62.2 minutes, suggesting he could have a much larger impact with more court time. While players like **Jimmer Fredette** and **Jeff Withey** show strong scoring, their lower efficiency per minute suggests they were less effective despite potential for greater contributions. From a **Moneyball perspective**, these players would be considered undervalued because their efficiency doesn't match their recognition, and with the right opportunities, they could have significantly impacted teams looking for efficient scoring.

Figure 2: D.J. Richardson (left), **Mike Myers** (Right)

Good Passers

Pass-first players like **John Stockton** and **Chris Paul** are well-known for their elite playmaking abilities, but lesser-known players like **Kirk Hinrich** and **Patty Mills** also demonstrate high assist-to-turnover ratios and solid assists per game. These players may not have the spotlight but play an essential role in ball distribution, offering depth in playmaking, especially for teams in need of efficient facilitators. Their contributions are often overlooked, yet they provide significant value to teams by maintaining ball control and ensuring smooth offensive execution.

Figure 3: John Stockton and Chris Paul

Versatile Players

Players like **Vince Carter**, **Gary Payton**, and **Tim Duncan** were often undervalued for their all-around contributions. Despite their ability to impact both ends of the floor, their multi-faceted skills were sometimes overshadowed by flashier stars. Teams could have benefited greatly from these players' versatility, especially in terms of leadership, defense, and overall contribution.

Figure 4: Gary Payton & Tim Duncan

Three-Point Specialists

With the growing importance of three-point shooting in modern basketball, players like **Marcus Haislip** and **Jimmer Fredette** stand out. Haislip's 65% 3P% and Fredette's 60.3% show their potential as elite shooters, though their careers didn't reflect their abilities due to limited opportunities. Similarly, **Peyton Siva** and **James Nunnally**, despite their strong three-point shooting overseas, weren't fully utilized in the NBA. In today's game, these players could have stretched the floor and created more scoring opportunities, making them valuable contributors for teams looking for reliable shooters.

Modeling

Efficiency Prediction (TS%)

True Shooting Percentage (**TS%**) is a key metric in evaluating a player's scoring efficiency, accounting for field goals, three-pointers, and free throws. As observed above, many of the most prominent players maintained a high **TS%**, highlighting its importance in measuring offensive effectiveness.

To build a predictive model for **efficiency (TS%)**, I selected specific statistics based on their relevance to player performance and their availability in the dataset:

- **PTS, REB, AST, STL, BLK, TOV**: Core box score statistics that contribute directly to a player's overall efficiency.
- **FG% and FT%**: Essential for calculating a player's shooting efficiency.
- **MIN**: Playing time determines opportunities to accumulate stats and influence the game.
- **Age**: Experience and physical prime can impact a player's efficiency.

The data was split into **80% for training and 20% for testing**, ensuring a balanced model evaluation.

Model Evaluation **Random Forest** was applied to predict TS%. The performance of the model was assessed using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and Adjusted R-squared (Adjusted R^2). The following observations were made:

Before parameter tuning:

- $R^2 = 0.7659$, meaning the model explained 77% of the variance in player efficiency (TS%).
- RMSE = 0.03
- MAE = 0.0226

After parameter tuning:

- $R^2 = 0.7671$, a slight improvement but still approximately 77%.
- MAE decreased by 0.001, showing minor improvement.
- RMSE remained the same (0.030).

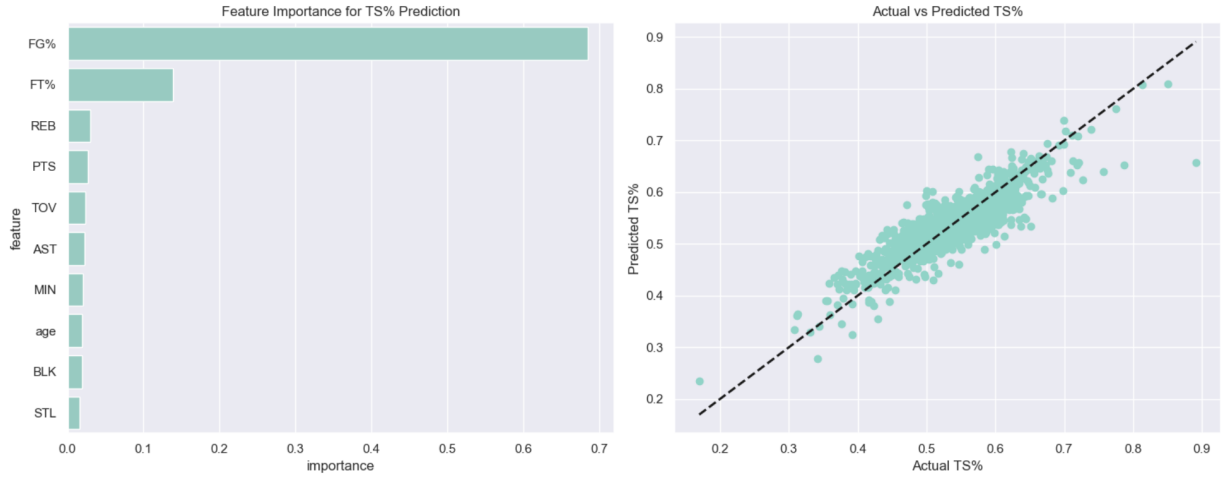
```
TS% Prediction Results:
Best Parameters: {'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}
Root Mean Squared Error: 0.03010722646594936
Mean Absolute Error: 0.022501884921930802
R-squared Score: 0.7671658127820751
Adjusted R-squared Score: 0.7656135848672889

Feature Importance:
  feature  importance
6     FG%    0.684835
7     FT%    0.139206
1     REB    0.030102
0     PTS    0.026531
5     TOV    0.022990
2     AST    0.021842
8     MIN    0.020237
9     age    0.019064
4     BLK    0.018954
3     STL    0.016241
```

Key Feature in TS% Prediction

- Field Goal Percentage (FG%) had the highest influence, contributing 67.3% to the model's predictive power.

In general we can say that the tuned model was best, although the R-square was still around 77%.



Career Trajectory Prediction

As observed, player performance tends to decline with age, though exceptions exist LeBron James, for example, continues to play at a high level even past the typical prime years. To analyze NBA career trajectories, we applied a Multi-output Regression model using Gradient Boosting, predicting three key stats: Points (PTS), Rebounds (REB), Assists (AST)

Model Performance After Tuning (Best Model)

Fundamental Stats	R^2	RMSE	MAE
PTS (Points)	0.4448	404.71	298.53
REB (Rebounds)	0.3999	161.80	118.09
AST (Assists)	0.5478	105.16	70.24

Feature Importance Analysis

- prev2_PTS (two-season lagged points) is the most critical variable in predicting future PTS.
- prev_PTS (last season's points) and prev_REB (last season's rebounds) also contribute significantly.
- Age and assist-related metrics (e.g., prev2_AST) have less influence on projections.
- Tuning the model did not significantly change the ranking of important features.

Key Insights

1. Impact of Performance Metrics
    - Recent seasons' performance (prev2_PTS, prev_PTS, and prev_REB) strongly predicts a player's future trajectory, particularly in scoring.
    - Assists have the highest predictability ($R^2 = 0.547$), while rebounds have the lowest ($R^2 = 0.399$), suggesting different patterns of skill development.

Broader Career Implications In a broader sense Playmaking skills (AST) develop more consistently, while scoring and rebounding are more volatile. Scoring and rebounding trajectories are harder to predict, likely due to external factors like coaching strategies, role changes, and injuries. Moderate predictive power ($R^2 \sim 0.40$ - 0.55) suggests that while general trends exist, individual career paths can vary widely.

While NBA career trajectories can be modeled with moderate success, they remain highly individualized. This highlights the complex nature of player development, emphasizing the need for flexible and personalized training, scouting, and career planning strategies in professional basketball.

Drafted Status Prediction (SVC)

The goal of this analysis is to predict whether a player will be drafted into the NBA based on their performance metrics from their first season. These metrics serve as a proxy for their overall abilities and potential, as we did not have access to their college performance data. The model classifies players as either drafted (Class 1) or undrafted (Class 0) based on their early performance in the league.

Before hyperparameter tuning, the SVM model achieved 79.55% accuracy, with 679 undrafted players correctly classified and 91 drafted players identified. However, 115 drafted players were misclassified, leading to a 44% recall for drafted players and a precision of 52%.

```
SVM Classifier:
Accuracy (Balanced): 0.7954545454545454

Confusion Matrix (Balanced):
[[679  83]
 [115  91]]

Classification Report (Balanced):
              precision    recall  f1-score   support

           0       0.86      0.89      0.87       762
           1       0.52      0.44      0.48       206

    accuracy                           0.80       968
   macro avg       0.69      0.67      0.68       968
weighted avg       0.78      0.80      0.79       968
```

After tuning, the SVM model achieved 83.06% accuracy with optimized parameters (C = 10, gamma = 1, kernel = 'rbf'), balancing true positives (TP) and true negatives (TN) while minimizing false predictions. The confusion matrix shows 669 undrafted players correctly classified and 135 drafted players identified, with 71 drafted players misclassified as undrafted.

The goal was to improve recall for drafted players (66%) while maintaining precision (59%), ensuring a fair balance between identifying true drafted players and avoiding false selections. This trade-off reduces misclassifications while keeping a strong overall performance.

```
Best Parameters from Grid Search:
{'C': 10, 'gamma': 1, 'kernel': 'rbf'}
SVM Classifier:
Accuracy (Balanced): 0.8305785123966942

Confusion Matrix (Balanced):
[[669  93]
 [ 71 135]]

Classification Report (Balanced):
              precision    recall  f1-score   support
```

```
              0       0.90      0.88      0.89       762
              1       0.59      0.66      0.62       206

    accuracy                            0.83       968
   macro avg       0.75      0.77      0.76       968
weighted avg       0.84      0.83      0.83       968
```

Conclusion

In the ever-evolving landscape of professional basketball, where advanced metrics increasingly influence team strategy, our study aimed to bridge the gap between raw statistics and strategic decision-making. Inspired by Moneyball's approach, we sought to unearth hidden talent, project player efficiency, and anticipate career trajectories, enriching traditional scouting with data-driven insights. Our analyses, spanning from True Shooting Percentage prediction via Random Forest to career projection using Gradient Boosting, highlighted both the promise and the complexity of quantifying player performance. Visualizations such as the 'Player Efficiency (TS%) by Age' scatter plot further underscored the dynamic nature of career progression, with youth holding potential, the prime showcasing consistency, and later years demanding adaptability.

Recommendation

1. **Embrace Multifaceted Evaluation:** Rather than relying solely on traditional scouting or advanced metrics,teams should synergize these approaches to gain a more comprehensive understanding of player potential, performance, and fit within the team.
2. **Invest in Talent Development and Coaching**: Acknowledge the uniqueness of each player's career trajectory by tailoring development plans to address specific areas for improvement,considering age and role evolution.
3. **Strategically Leverage Analytics in Scouting:** Utilize advanced metrics like True Shooting Percentage to identify undervalued players, but consider the potential and the specific team context in which the player will be utilized. This is especially true with age and projected output, for planning beyond one season.
4. **Cautious Use of First-Game Metrics**: The first impression and performance should not be weighted significantly due to its volatility.

Ultimately, as basketball continues to evolve, so too must the strategies for talent acquisition and team building. By embracing a holistic approach that marries data analytics with strategic decision-making, teams can optimize their performance and unlock unparalleled success. Through continuous adaptation and a commitment to innovation, basketball organizations can harness the full potential of their players and teams.

Reference

Alperendes. (2021, March 3). *NBA Analysis - How Basketball has changed?* Kaggle. https://www.kaggle.com/code/alperendes/nba-analysis-how-basketball-has-changed

Lewis, M. (2003). *Moneyball: The art of winning an unfair game.* W. W. Norton & Company.

Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, M. (2009). *Basketball on paper: Rules and tools for analytical decision-making.* Potomac Books.

Skansi, S. (2017). *Basketball data science: With applications in R.* CRC Press.

Fewell, B. M., Armatas, V., Neville, J., McGuire, R.T., Williams, M. R., & Davids, K. (2012). Modeling representative task constraints in Australian football. *Journal of Science and Medicine in Sport, 15*(2),

153-159.

Appendix

```
Top 10 Efficient Scorers:
                Player Team       TS%  PTS_per_game    MIN  Efficiency
8012     Primoz Brezec  KRY  0.855696     20.833333  183.8    0.113348
8855   Jimmer Fredette  PAN  0.824176     13.666667  342.3    0.039926
16474  D.J. Richardson  SPI  0.820611     21.500000   62.2    0.345659
20199      Jeff Withey  IRO  0.811924     14.888889  246.7    0.060352
16442      Seth Tuttle  LIM  0.782584     18.333333  178.5    0.102708
15525      Tony Gugino  RIL  0.774246     12.666667  182.6    0.069368
3006         Raja Bell  PHX  0.772025     13.600000  214.8    0.063315
16992   James Nunnally  FEN  0.771899     11.866667  635.1    0.018685
17516    Cory Bradford  BOS  0.770021     15.000000   92.4    0.162338
17544       Mike Myers  OOS  0.767045     13.500000   41.7    0.323741
```

```
Top 10 Good Passers:
                  Player Team  AST_per_game  AST_TO_Ratio
3708        Kirk Hinrich  CHI      4.000000     10.000000
7186         Patty Mills  SAS      3.571429      8.333333
7582          Joe Ingles  UTA      4.714286      8.250000
17534      Maarty Leunen  AVE      5.125000      8.200000
611        John Stockton  UTA     11.400000      8.142857
5874       Zaza Pachulia  DAL      3.200000      8.000000
2021         Earl Watson  MEM      3.750000      7.500000
5811          Chris Paul  LAC      7.250000      7.250000
19931  Cliff Clinkscales  HAL     10.125000      7.147059
13113     Matthew Pettit  MIN      3.500000      7.000000
```

```
Top 10 Versatile Players:
             Player Team  PTS_per_game  REB_per_game  AST_per_game  \
1     Vince Carter  TOR     25.695122      5.804878      3.926829
2      Karl Malone  UTA     25.548780      9.500000      3.707317
4      Gary Payton  SEA     24.170732      6.451220      8.926829
6       Grant Hill  DET     25.756757      6.621622      5.202703
7    Kevin Garnett  MIN     22.925926     11.802469      4.950617
8   Michael Finley  DAL     22.621951      6.317073      5.341463
9     Chris Webber  SAC     24.453333     10.493333      4.600000
10       Ray Allen  MIL     22.060976      4.378049      3.756098
12      Tim Duncan  SAS     23.189189     12.405405      3.162162
13  Glenn Robinson  MIL     20.901235      5.987654      2.382716

    STL_per_game
1       1.341463
2       0.963415
4       1.865854
6       1.391892
7       1.481481
8       1.329268
9       1.600000
10      1.341463
12      0.891892
```

```
13      0.962963
```

```
Top 10 Three-Point Specialists:
                Player Team       3P%  3PA
12656   Marcus Haislip  DON  0.650000  160
8855   Jimmer Fredette  PAN  0.602564   78
8853       Paul Zipser  BAY  0.579710   69
8783       Paul Zipser  BAY  0.579710   69
16992   James Nunnally  FEN  0.571429  119
15328   Chasson Randle  NYM  0.548077  104
20438      Peyton Siva  BER  0.541667   72
19714      Peyton Siva  BER  0.541667   72
6021         Pau Gasol  SAS  0.538462  104
16387      Jon Diebler  GAL  0.538462   91
```

- ```
  Top 15 Players - Regular Season TS%:
  Player
  Stephen Curry            0.623333
  Kevin Durant             0.612656
  James Harden             0.609455
  Kawhi Leonard            0.594411
  Shaquille O'Neal         0.589213
  LeBron James             0.587365
  Chris Paul               0.585591
  Giannis Antetokounmpo    0.584268
  Dirk Nowitzki            0.580947
  Kobe Bryant              0.555138
  Dwyane Wade              0.551497
  Tim Duncan               0.548705
  Allen Iverson            0.523202
  Michael Jordan           0.491147
  Name: TS%, dtype: float64

  Top 15 Players - Playoffs TS%:
  Player
  Kawhi Leonard            0.611560
  Stephen Curry            0.605011
  Kevin Durant             0.594372
  LeBron James             0.585926
  Chris Paul               0.583343
  James Harden             0.581835
  Dirk Nowitzki            0.569571
  Giannis Antetokounmpo    0.557766
  Shaquille O'Neal         0.552266
  Kobe Bryant              0.544648
  Tim Duncan               0.543229
  Dwyane Wade              0.541905
  Allen Iverson            0.496732
  ```
-
  ```