NBA Career Trajectories Analysis

Independent Data Analytics Project

Colton D. DiFranco

May 2025

Introduction and Objective:

This project analyzes NBA player career trajectories using game-by-game data from January 1, 2005, to May 19, 2025. The primary focus is to understand how key performance metrics, such as points, rebounds, assists, and more, evolve over the course of a player's career, measured by season number (i.e., the number of years since a player was drafted). The analysis is segmented by player position (guard, forward, center) to uncover position-specific patterns.

The main objectives of this project are to:

- Determine the average career length for players at each position
- Identify the season in which players tend to peak in various performance metrics
- Provide data-driven insights that can inform strategic decisions, such as:
    - Trading current players
    - Acquiring talent through trades or free agency
    - Drafting based on roster composition and player career stages

These insights aim to help teams optimize player development, roster management, and long-term planning.

Data Description:

The analysis is based on a comprehensive set of NBA data that includes player information, individual game-by-game statistics, game specific information, and team-level statistics. The source of this data is Eoin A. Moore's "NBA Dataset - Box Scores & Stats, 1947–Today" from Kaggle, which is updated daily as new games are played.

For this project, the most critical datasets were:

- **Player Dataset**: Contains individual-level player data, including a unique player ID, first and last name, position, draft year, and other biographical details.
- **Player Game-by-Game Statistics Dataset**: Provides detailed performance metrics for every game a player has played. This includes fields like player ID, game date, game ID, teams involved, minutes played, points, assists, rebounds, steals, blocks, field goal percentage, and more.

Using these datasets, I built a relational database in **MySQL**, organizing the data into structured tables with clearly defined column types, primary keys (to uniquely identify records), and foreign keys (to establish relationships between tables). For example, player ID was used as a foreign key to link game performance data to the player information table.

This structured database allowed for efficient querying, data cleaning, and the integration of relevant information across tables to support the analysis. After creating the database schema, I imported the raw CSV files from Kaggle directly into MySQL, preparing the data for subsequent transformation and exploration.

Data Cleaning:

Once the data was successfully imported into MySQL and the schema was properly defined, the data cleaning process began. Data cleaning is essential to ensure accuracy, consistency, and reliability in any analysis. My process involved several key steps:

1. **Standardizing Categorical Values**
   I began by inspecting string-based columns for inconsistencies in text entries. Using SQL queries to extract unique values from these columns, I identified mismatches in how similar categories were labeled. For example, game types were sometimes listed as both

"West - Conf. Semifinals" and "West Conf. Semifinals." While these refer to the same event, SQL treats them as distinct values. I resolved these discrepancies by standardizing such entries across the dataset. This standardization was applied to multiple columns to ensure uniform data representation.

2. **Filtering by Relevance and Reducing Scope**

   To focus the analysis on the modern era of basketball and improve performance, I removed:

   - All players drafted before 2005
   - All games played before January 1, 2005

   This reduced the dataset size significantly while ensuring that the insights derived are relevant to today's NBA context.

3. **Removing Duplicates**

   Duplicate entries can distort aggregate metrics and trends. I queried for and deleted all duplicate records across key tables to preserve data integrity.

4. **Handling Outliers**

   I reviewed key numerical columns for extreme or implausible values. For example, games where a player scored -22 points or logged 73 minutes (exceeding the 48-minute regulation time) were flagged and removed to prevent skewed results.

5. **Excluding Non-Playing Entries**

   Entries with 0 minutes of playing time for a player were excluded from the analysis. These entries are not useful in performance-based evaluations and would bias averages downward if included.

6. **Managing Null Values**

   I reviewed all null (missing) values and made case-by-case decisions:

   - If a null value appeared in a critical column (e.g., points, assists, rebounds, draft year), the row was removed

    - If a null appeared in a non-essential column not used in the analysis, I retained the row to preserve its useful information

By the end of this cleaning process, the data was consistent, relevant, and ready for robust analysis.

New Table Creation:

With the data cleaned and organized into a relational database, the next step was to transform and consolidate relevant information into a single, analysis-ready table. While the original database consisted of four interrelated tables, 114 total columns, and over 2 million entries, the necessary insights required aggregating and summarizing this information in a more accessible format.

The goal was to create a new table, titled player_season_stats, that captured key seasonal statistics for each player. This table includes:

- Player ID, first name, and last name
- Draft year and season year
- Season number, calculated as (season year - draft year + 1)
- Position (guard, forward, or center)
- Total games played in that season
- Average per-game statistics, including:
    - Points
    - Assists
    - Rebounds
    - Minutes played
    - Field goal percentage
    - Others if needed

To construct this table, I used SQL queries that:

- Aggregated game-by-game player data into season-level averages

- Grouped by player and season year
- Joined data from multiple tables using JOIN statements, connecting rows based on the player ID key

For example, while player positions and draft years resided in the player information table, individual performance metrics came from the player game-by-game statistics table. I joined these datasets on player ID to combine relevant attributes into a unified table.

As a result, this final table contains one row per player season. For instance, a player active from 2010 through 2018 would have nine entries, one for each season played, containing their average statistics for each season and other relevant information. This structure enabled streamlined querying and analysis for player career trajectories and positional comparisons.

Querying the New Table:

With the consolidated player_season_stats table in place, I was ready to begin extracting insights through SQL queries. A query functions as a question posed to the database allowing for targeted exploration of the data.

The key queries I developed for analysis included:

1. **Player Count by Position and Career Season**
   This query calculated the number of active players in each position (guard, forward, center) for every season since their draft. It provided insight into career longevity and wear patterns across different positions.
2. **Average Statistics by Career Season**
   This query returned league-wide averages for key performance metrics (e.g., points, assists, rebounds) based on how long players had been in the league. It enabled me to assess how player performance evolves over time.
3. **Average Statistics by Position and Career Season**
   A refinement of the second query, this version grouped the statistics by both position and

season number. This allowed me to compare performance trends across positions and determine when players at each position tend to peak in various metrics.

After running these queries, I exported the results to CSV files for further analysis and visualization in Microsoft Excel.

Note: SQL code used for data transformation and querying is included in a separate folder within the GitHub repository associated with this project.

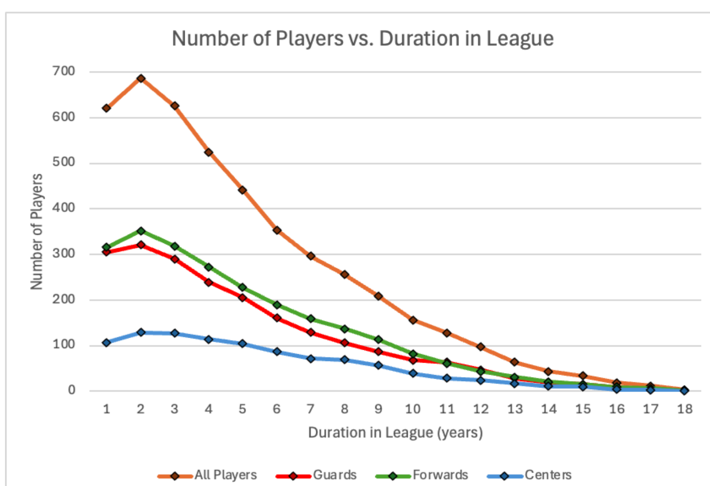
Analysis and Results:

With the results of the three SQL queries imported into Microsoft Excel, I used Excel's functions and plotting capabilities to analyze trends and generate visualizations. To ensure data quality, I excluded players in their 18th season, as only three such players exist in the dataset, and their inclusion could skew the averages.

Player Longevity by Position

The first part of the analysis focused on career longevity and how it differs by player position. I utilized data showing the number of players in each season since their draft year, separated by position. In this structure, if a player had a 6-year career, they appeared six times, once for each season, each time associated with their corresponding year in the league.

Note: The data was organized by calendar year rather than basketball season, but this distinction has minimal impact on the conclusions drawn from career duration trends.

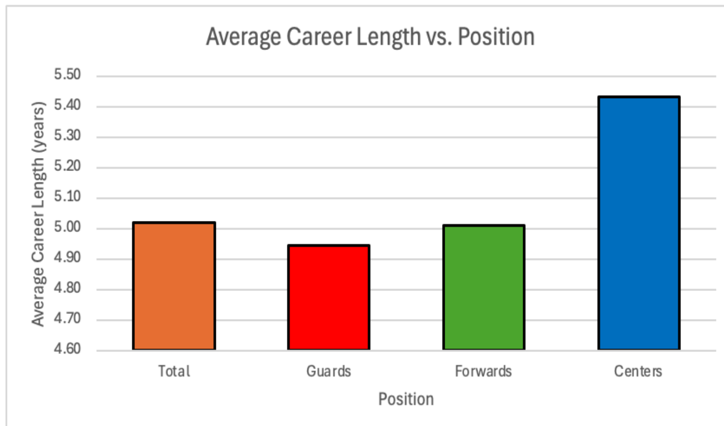


This initial plot revealed a clear trend; the number of players decreases with each successive season in the league.

This is expected, as players retire or leave the league over time. At the start of their careers, there are significantly more guards and forwards than centers. This reflects the structure of a typical basketball lineup. Teams usually field two guards and two forwards, but only one center. Therefore, the league's positional composition naturally contains more guards and forwards.

At the earliest stages (years 1 and 2), there are roughly three times as many guards and forwards as centers. By year 5, this ratio decreases to about 2:1. This narrowing gap may reflect differences in positional career longevity.

To explore this further, I calculated the average career length by position using a weighted average approach: multiplying each season number by the number of players in that season, summing the results, and dividing by the total number of players at that position.



Average Career Length by Position:

The results revealed some key insights:

- **Overall average NBA career length since 2005**: ~5 years
- **Guards**: Slightly below the average
- **Forwards**: Right at average
- **Centers**: 0.49 years longer than guards, and 0.42 years longer than forwards

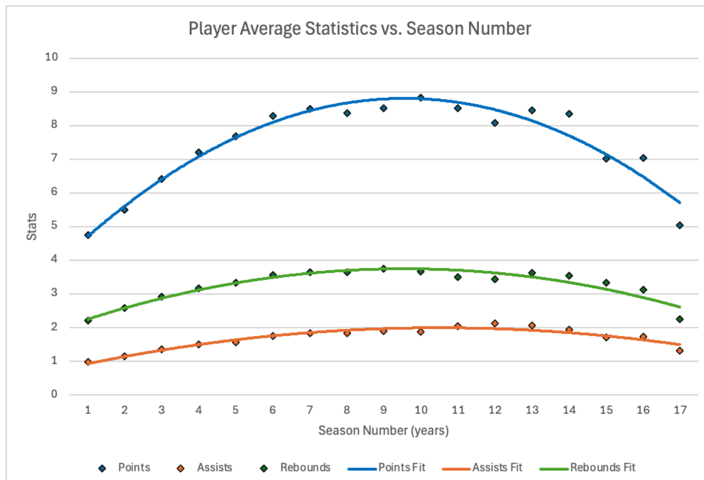
This difference is statistically meaningful and points to potential underlying factors:

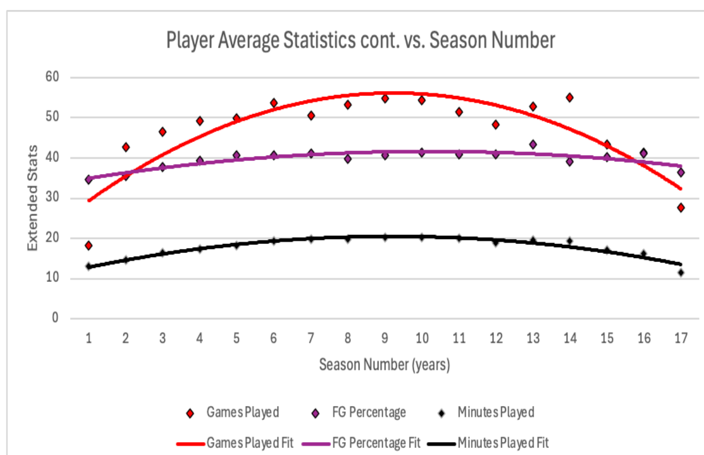
- **Role and physical demands**: Centers rely more on height, strength, and positioning near the basket for rebounding and interior defense. These skills tend to decline less rapidly with age than speed, agility, and quick decision-making, traits more critical for guards and forwards.

- **Injury risk**: Guards and forwards often engage in faster-paced movement, cutting, and driving, which may expose them to greater wear and tear or injury risk over time.
- **Team roster decisions**: Teams may cycle through guards and forwards more often in pursuit of offensive or defensive upgrades, while centers with strong fundamentals may retain value even as they age.

These findings suggest that position plays a significant role in determining a player's longevity in the NBA and may influence front office decisions in areas like drafting, trading, or contract extensions.

<u>Player Statistics by Career Length</u>





This analysis examines how key player performance metrics evolve throughout an NBA player's career. The goal was to determine when, on average, players reach their peak in various statistics such as points, rebounds, assists, games played, minutes per game, and field goal percentage.

Each of these metrics was plotted against the number of years a player had been in the league. These plots typically followed a parabolic shape (starting low, rising to a peak, and then declining) suggesting a second-degree polynomial (quadratic) relationship.

Some metrics, like points per game or games played, showed more dramatic changes over time, while others, such as field goal percentage, appeared relatively stable. However, it's important to account for the scale of each metric. For example:

- Points per game increased from around 5 in a player's rookie season to a peak just under 9, an 80% increase.
- Rebounds and assists showed a more significant relative growth, nearly doubling later in the career, even though the arcs look shallower
- Field goal percentage increased more modestly, from 34.6% to a peak of 43.5%, a 26% rise.

To determine the season in which each metric peaked, I fitted each curve with a quadratic equation of the form:

$$y = ax^2 + bx + c$$

The peak of each parabola occurs at:

$$x_{max} = \frac{-b}{2a}$$

Applying this to each performance metric and then averaging the results (assuming equal weights for all metrics) yielded **peak performance 9.7 seasons into a player's career**. Individually, the peaks were:

- Points: 9.56 years
- Rebounds: 9.54 years
- Assists: 10.46 years
- Games Played per Season: 9.23 years
- Field Goal Percentage: 10.20 years
- Minutes Played per Game: 9.20 years

This may appear surprising when compared to the previous analysis, which found the average career length to be just around 5 years. However, this discrepancy can be logically explained. The players who remain in the league for nine or more seasons tend to be the top-performing athletes, the upper echelon of their draft classes. According to the earlier player count data, fewer than one-third of all NBA draftees make it to year 9. This group likely excludes many role players and includes mostly high-impact starters, All-Stars, and other top talents.

Thus, while the average player retires around year 5, those who survive into their 9th or 10th season tend to be players who not only performed well early in their careers, but continued to develop and contribute meaningfully to their teams.

A particularly interesting comparison is between points per game and field goal percentage over a player's career. While points per game show a substantial rise (80% increase from 5 to 9), field goal percentage improves more modestly (26% increase from 0.346 to 0.435). This divergence suggests that as players gain experience:

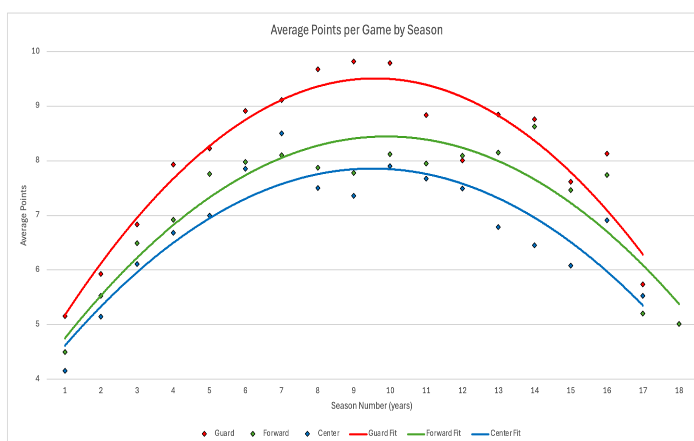
- They take more shots, likely due to increased confidence, role responsibility, or offensive freedom.
- Efficiency increases, but not at the same rate as volume. Players may become more comfortable in offensive schemes or better at selecting shots, but the primary driver of increased scoring seems to be usage, not just accuracy.

These findings highlight the balance between experience, physical decline, and evolving roles throughout a player's NBA journey.

Player Statistics by Career Length by Position

This final analysis builds upon the previous section by examining how player performance trajectories vary across different positions. The goal was to determine whether players at different positions develop differently and reach peak performance at different times in their careers.

As before, each performance metric was plotted against years in the league and modeled using second-degree polynomial (parabolic) fits. These curves generally follow the expected pattern: performance improves early in a player's career, peaks at a certain point, and then declines with age and physical wear.



Scoring Trends by Position

As anticipated, guards score more points on average than forwards, and forwards score more than centers. While all three groups begin their careers with relatively similar scoring averages, the trajectories diverge over time:

- Guards experience the most significant growth in points per game as their careers progress.
- This could suggest that scoring capabilities in guards are more valuable to team offices when compared to other positions or that scoring ability for guards is more heavily influenced by experience compared to other positions.

This distinction may reflect the evolving offensive roles of guards, who often serve as both playmakers and primary scoring options.

Other Metrics and Trends

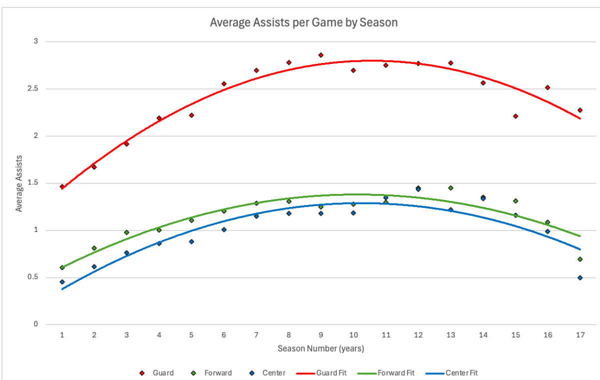
Other key performance metrics were analyzed across positions as well. Although not every trend is discussed in detail here, several notable patterns emerged:

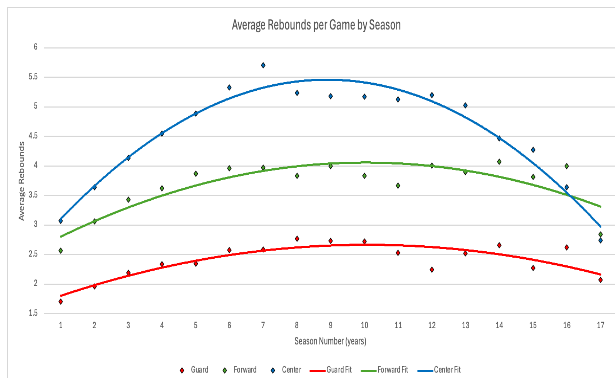
- Assists peak later than other statistics across all positions. This supports the idea that playmaking ability and court vision take longer to develop, possibly due to the complexity of offensive systems and the mental demands of reading defenses effectively.
- Minutes per game vary significantly by position. On average:
    - Guards play the most minutes
    - Centers play the fewest minutes

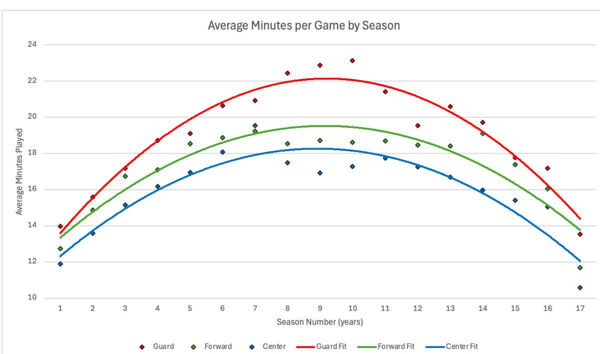
At first, this finding was counterintuitive. One might expect centers to play more minutes, given their typically more stationary style of play, which might result in less fatigue. However, a closer look reveals that guards often carry the offensive load through scoring and facilitating, both of which are critical to winning games. This makes it strategically important for teams to maximize elite guard minutes.

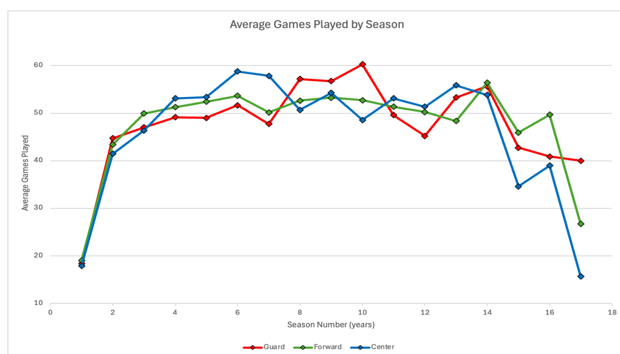
This also reinforces a conclusion from the earlier analysis that centers tend to have longer careers than guards. The higher average minutes played per season for guards likely accelerates wear and tear, contributing to more frequent injuries and earlier retirements. Conversely, lower usage and physical toll for centers may allow them to preserve their bodies and extend their careers. Also, since speed and agility decline heavily with age, it is reasonable to conclude that guards also have shorter careers due to their inability to move as quickly and efficiently when compared to younger counterparts.

Note: Trendlines for games played per season by position were not included, as the data did not follow a consistent or meaningful parabolic trend.









Conclusion and Key Takeaways:

This project set out to analyze NBA player careers and performance trends through structured data analysis using SQL and Excel. By cleaning and consolidating player data into a single comprehensive table, writing targeted queries, and visualizing results, several meaningful insights emerged regarding player longevity, statistical development over time, and positional differences.

Key Takeaways:

- **Average Career Length:** The average NBA career lasts about 5 seasons, with centers having the longest average careers and guards the shortest. This may reflect the physical demands of each position; centers rely more on size and strength, while guards depend on agility and speed, which decline more quickly with age.
- **Career Peaks:** Most key statistics, including points, assists, and rebounds, follow a parabolic trajectory over a player's career. The average peak performance across all metrics occurs around 9.7 years into a career, a point that fewer than one-third of players reach. This highlights how often times only the NBA's most elite athletes and influential players make it that long in their career.
- **Positional Differences in Statistics:**
    - Guards score more, assist more, and play more minutes on average, emphasizing their role in playmaking and scoring.
    - Centers, while less involved in high-tempo actions, appear to benefit from this physically by maintaining longer careers.
    - Assists peak later than other metrics for all positions, suggesting that court vision and playmaking take longer to develop than scoring or rebounding ability.
- **Player Usage Trends:** The higher minutes played by guards may correlate with greater physical wear and tear, contributing to their shorter average careers. In contrast, lower usage and physical exertion may help explain the longer-lasting careers of centers.

This analysis provides a data-driven foundation for understanding the trajectory of NBA careers. It also opens the door for further exploration, such as comparing individual players, incorporating more advanced statistics, or evaluating how these trends have shifted over time with changes in playing style and league structure.

Analysis Limitations

While this project offers valuable insights into NBA player career trends and performance by position, there are several limitations to keep in mind:

1. Averaging Across Players

The analysis uses averages of player statistics per season, which smooths out individual differences. This approach does not account for outliers, such as exceptionally high-performing players or those with unusual career arcs (e.g. late bloomers or injury-prone players). Also, this analysis averages based on games, not on minutes. Some results could potentially be misleading as some game statistics reflect 30 minutes of playing time while others reflect 5.

2. Career Length vs. Peak Performance

Although the study found that player performance tends to peak around year 9.7, most players don't make it that far. This means the peak reflects a subset of players who are likely already above average in ability, health, or opportunity. The analysis does not differentiate between these elite players and others, which could bias interpretations of "average" career trajectories. To get a different analysis, we could split the players into subgroups of players (short careers, medium careers, and long careers) and examine how the arcs change specifically for these groups to not allow it to be overpowered by elite, extended career players.

3. Position Classification

Player positions were treated as static across seasons, yet in reality, many players switch or evolve roles over time (e.g., forward to guard). A more dynamic position tracking method might yield different insights.

4. No Control for External Variables

The analysis does not account for important external factors that could influence career progression, such as:

- Injuries and health

- Coaching strategies
- Team quality and roster depth
- Rule changes over time

5. Visualization and Modeling Simplicity

Polynomial fits (parabolas) were used to model trends, assuming a smooth rise and fall over time. While useful for capturing general arcs, these models may oversimplify the nuanced and non-linear nature of player development.

Future Work Possibilities

This project uncovered a number of valuable trends related to NBA player performance, longevity, and career progression. However, there are several directions in which this analysis could be expanded to provide deeper or more nuanced insights:

1. Segmenting Players by Career Length

As noted in the limitations section, players with long careers disproportionately influence average statistics in later seasons. A natural extension would be to categorize players by career length, for example, into short (1–3 seasons), medium (4–8 seasons), and long (9+ seasons) career groups and analyze each group's trajectory independently. This would reveal how performance evolves for typical players, rather than being skewed by the small percentage who have exceptional longevity.

2. Calculating Stats per Minute Rather Than per Game

This analysis has calculated all of the average stats of player performance based on games that they played. In reality, all players play a different amount of time in every game. It would be a great analysis to see how all of these metrics change on a per playing minute basis. This would more effectively communicate player efficiency but fail to show player usage.

3. Accounting for Position Changes

Future work could attempt to track positional changes over time, as many players transition roles throughout their careers. Using game-by-game or season-level position data would enable a more dynamic and realistic view of how role affects performance and career trajectory.

4. Incorporating Advanced Metrics

The current analysis focuses on traditional box score stats. Including advanced basketball metrics such as PER (Player Efficiency Rating), Win Shares, or BPM (Box Plus/Minus) could provide a more refined understanding of a player's impact beyond raw averages.

5. Exploring Injury and Health Data

Since career longevity is often impacted by injuries, integrating injury history or games missed due to health could explain variance in career arcs. This could also help analyze whether certain positions or play styles are more injury-prone.

6. Contextualizing with Team and League Trends

Performance and usage patterns may be influenced by team dynamics, coaching philosophies, or league-wide trends (e.g., the rise in 3-point shooting). Factoring in contextual variables could help explain shifts in statistics or playing time over time.

7. Machine Learning Applications

For a more predictive approach, future analyses could employ machine learning models to forecast career trajectories based on early-career performance, physical attributes, or draft position. Clustering players by similar developmental arcs could also identify patterns not visible through average-based analysis.

By expanding in these directions, this project could evolve from descriptive analytics into more targeted and predictive tools for understanding player development and informing decisions in basketball scouting, coaching, or roster management.