Exploratory Data Analysis (EDA) Memo

Checkpoint 5

Overview

This checkpoint is about seeing what the data is actually saying before you model anything. You'll create clean, readable visuals, summarize patterns, and write short takeaways that connect to your project question. Keep it focused, clear, and useful.

Group Work Option

You may work in groups of up to 3 students, but each student must maintain their own GitHub repository. Collaboration on topic and data collection is encouraged, but each student must submit an individual memo and ensure their repository contains all deliverables.

Assignment Objectives

By completing this checkpoint, you will:

1. Use visuals and summary stats to describe your dataset in a way that informs next steps.
2. Surface trends, comparisons, and relationships that matter for your question. 3. Call out outliers, anomalies, or data gaps you discover.
4. End with actionable next steps for your first analysis in Checkpoint 6.

What to Include in Your Checkpoint

Question Snapshot	A 2-3 sentence refresher on your project question and what "success" would look like.	What are you trying to understand or predict? For whom does it matter? What would be a useful insight?

Data Used (1-2 sentences)	Name the cleaned file from CP4 (row definition, time span). No need to re-explain cleaning.	What does one row represent? What seasons/years/games are included? Any major filters applied?
Univariate Distributions	Show the distribution for 3-4 key variables (histograms or boxplots). Add a one-line takeaway under each figure.	Are they skewed? Any ceilings/floors? Are the typical values where you expected? Any obvious data entry issues?
Relationships (Pairs) Subgroup Comparisons Time or Sequence View (if applicable)	At least 2 simple relationships (scatter or line with a light trendline; or grouped bar/box if categorical). Compare one important split (home vs. away, position groups, opponent tiers). Show a trend over time (season, week, game order) for one metric.	Do the variables move together in a way that fits your sports intuition? Any diminishing returns or non-linear shapes to note? Which group tends to be higher/lower? Is the difference meaningful or just noise at a glance? Are there clear up/down trends, streakiness, or schedule effects (back-to-back)?
Outliers & Anomalies	Identify what stands out and whether to keep, cap, or flag it.	Are the outliers legit performances, or data errors? If you keep them, how will you handle them in analysis?
Missingness & Coverage (brief)	A quick note on any remaining gaps that matter for interpretation.	Do missing values cluster in certain seasons/teams/players? Does it bias a comparison?
Early Takeaways → Next Steps	3-5 bullet takeaways tied to your question, then 2-3 next steps for CP6.	What seems promising? What should be tested first? What comparisons or features should you build next?

Visual Standards (keep it clean)

- Label everything: titles, axes (with units), and categories.
- One clear point per chart; add a short takeaway sentence directly under each figure.
- Avoid 3D and heavy effects; keep colors simple and readable.
- If you show rates/percentages, include n (sample size) nearby.
- If you transform a variable (log, per-minute, per-possession), say so on the axis/title.

Minimum Visuals to Include

- 3-4 univariate visuals (key variables).
- 2 relationship visuals (pairs).
- 1 subgroup comparison.
- 1 time/sequence visual (if your data is considered time-series).

    - If not time series, add one more relationship or subgroup chart.

Submission Details

Document Type	Google Doc (professional memo style) → export to PDF
Submission	Share the Google Doc link on Classroom and upload the PDF to your repo's /reports folder. Include figures in /figures (PNG or PDF).
File Naming	LastName_FirstName_CP5.pdf (example: Doe_Jane_CP5.pdf). Include your name(s), project title, and the checkpoint heading at the top of the memo. EACH STUDENT MUST UPLOAD TO INDIVIDUAL REPO
Format Tips	Number figures (Fig. 1, Fig. 2, …) and reference them in the text. Put a one-line takeaway under each figure.

Deliverables

1. EDA Memo (PDF)
2. Figures saved to your repo's /figures folder (filenames match the figure numbers)

Grading & Rubric (10 points total)

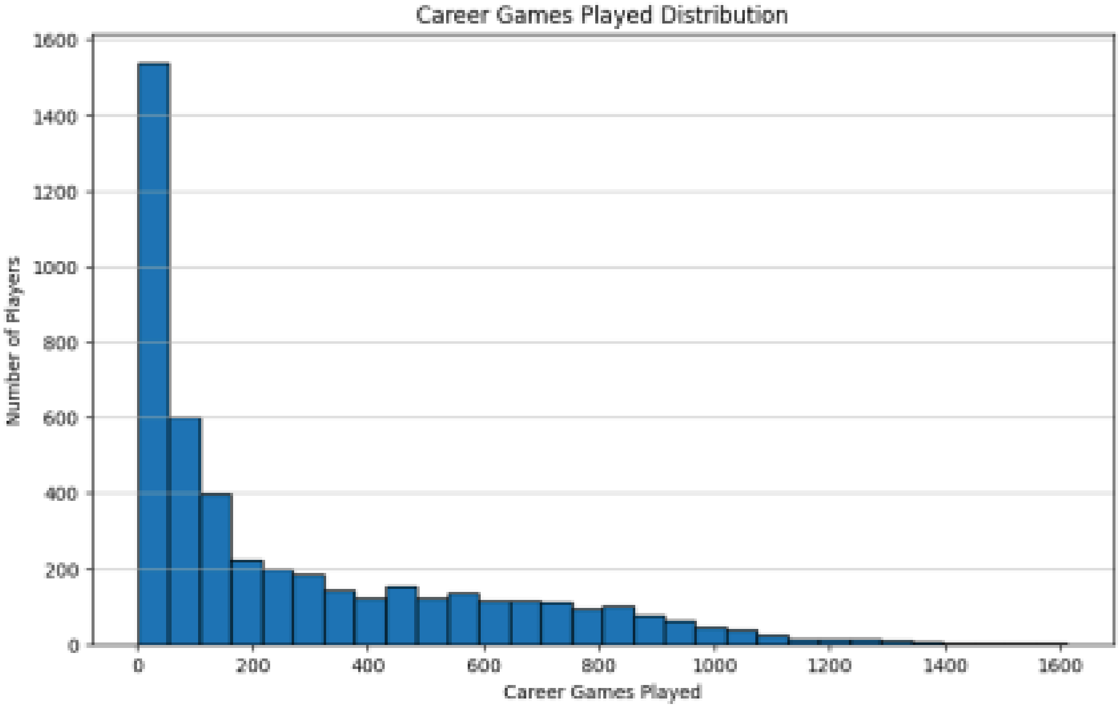
Your Checkpoint 5 will be evaluated on the following criteria:

Relevance to Question	Visuals and comments directly support the project question; no clutter or off-topic charts.	3
Clarity of Visuals	Clean labeling, sensible scales, readable comparisons; one-line takeaways under each chart.	3
Insight & Interpretation	Short, concrete observations (not just "goes up"). Notes on surprising patterns, subgroup differences, or time effects.	2
Outliers & Gaps	You acknowledged outliers/anomalies/missingness and stated how you'll treat them going forward.	1
Professionalism & Reproducibility	Organized memo, correct figure references, on time, files named and stored correctly.	1

Additional Notes

- *Start with more charts than you need, then keep only the ones that tell a story.*
- *If a comparison looks noisy, say that-it’s still useful guidance for CP6.*
- *Keep captions short and specific: “Home teams score ~4.1 more points on average (Fig. 2). Effect is larger in 2023–24 (Fig. 5).”*

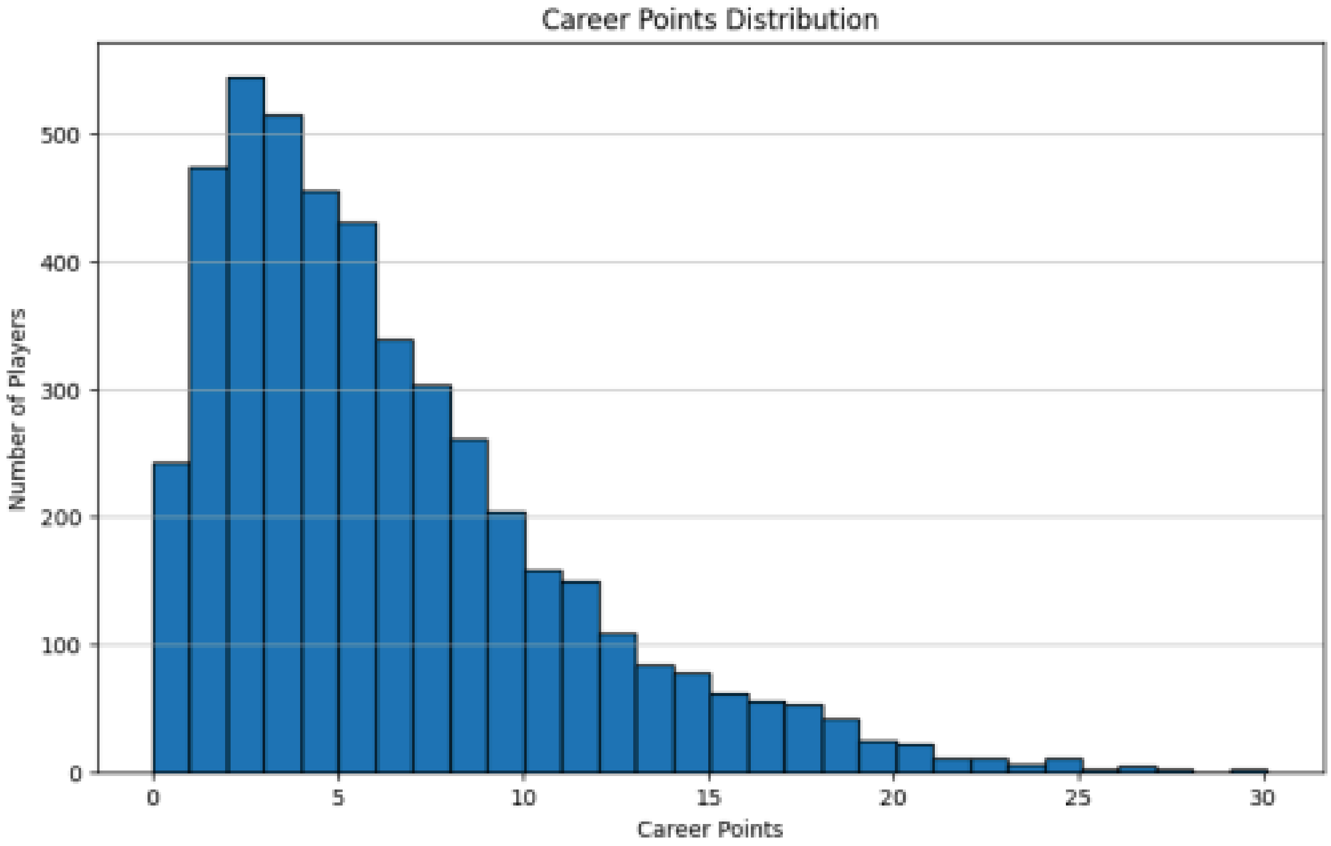
The project asks: Which NBA players in 2017–2018 delivered the most on-court value based on performance and efficiency metrics? Success would entail the identification of players who are outliers for minutes, scoring, efficiency, and impact statistics. This has value to teams (rotation decisions), analysts (feature design), and fans (understanding player contributions). One useful insight would be clear signals on which archetypes-for instance, high-usage scorers, efficient role players-are the ones driving team success.



This is the tallest bar near 0 games, which shows that most of the players had short careers in the NBA, often just a few games or seasons.

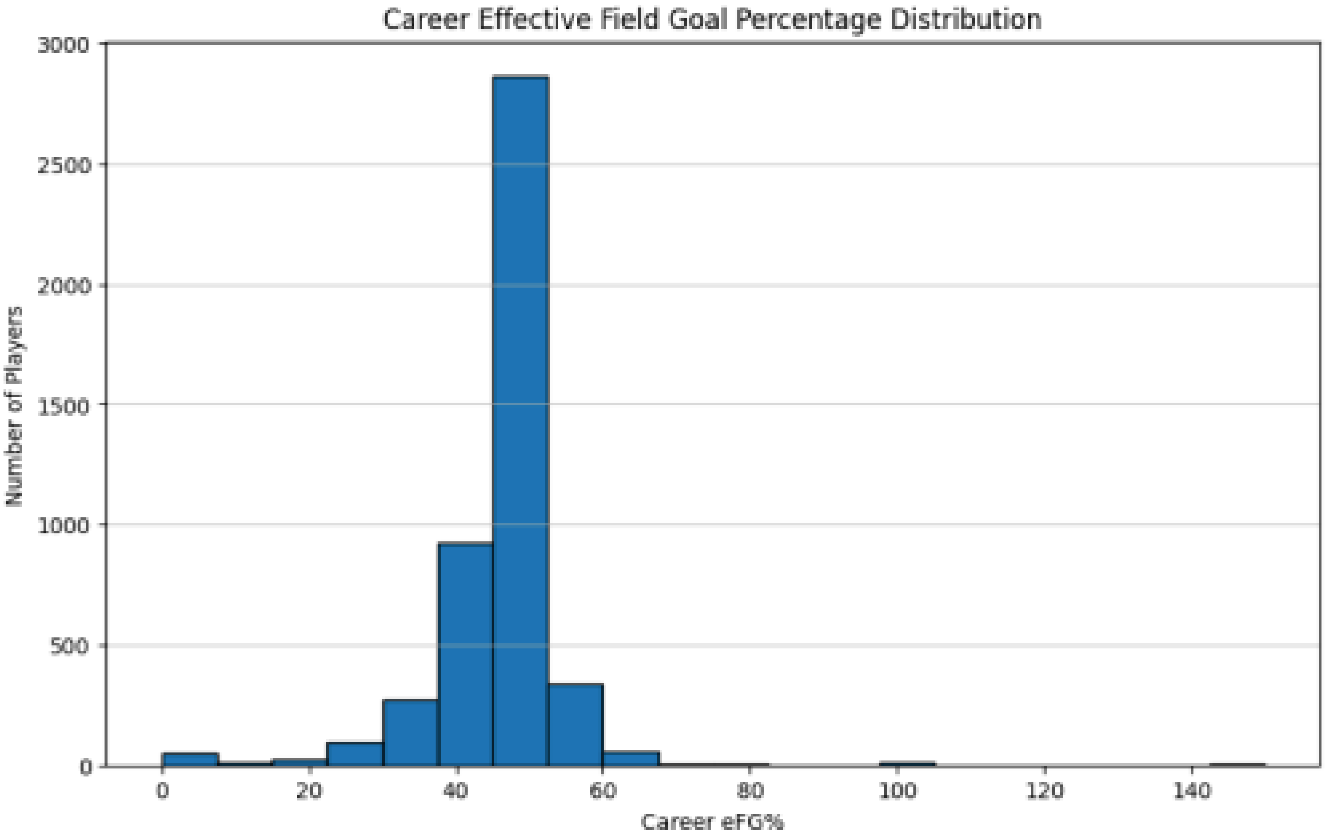
But as the number of games increases, the frequency falls off sharply, showing that longevity is uncommon.

DATA 6560 - Sports Analytics - Checkpoint 5 Fall 2025



Notice how the tallest bar is close to 2–4 points, indicating that many players have scored only a small number of points in their NBA careers.

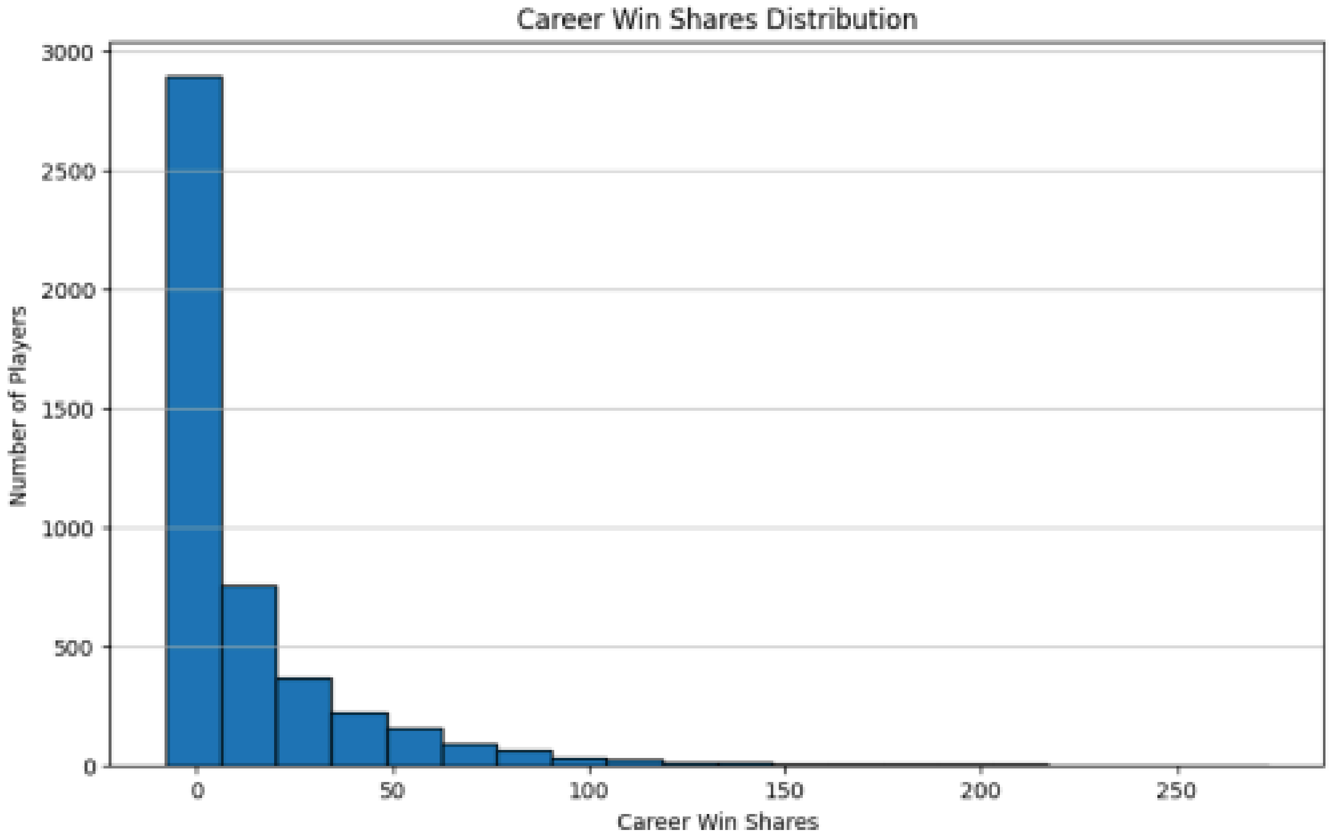
As point totals mount, the number of players decreases rapidly, indicating that high career scoring is rare.



The highest concentration of players have a career eFG% between 40% and 60%, which is typical for NBA players.

Very few players exceed 70%, as this usually reflects specialists like dunkers or low-volume shooters.

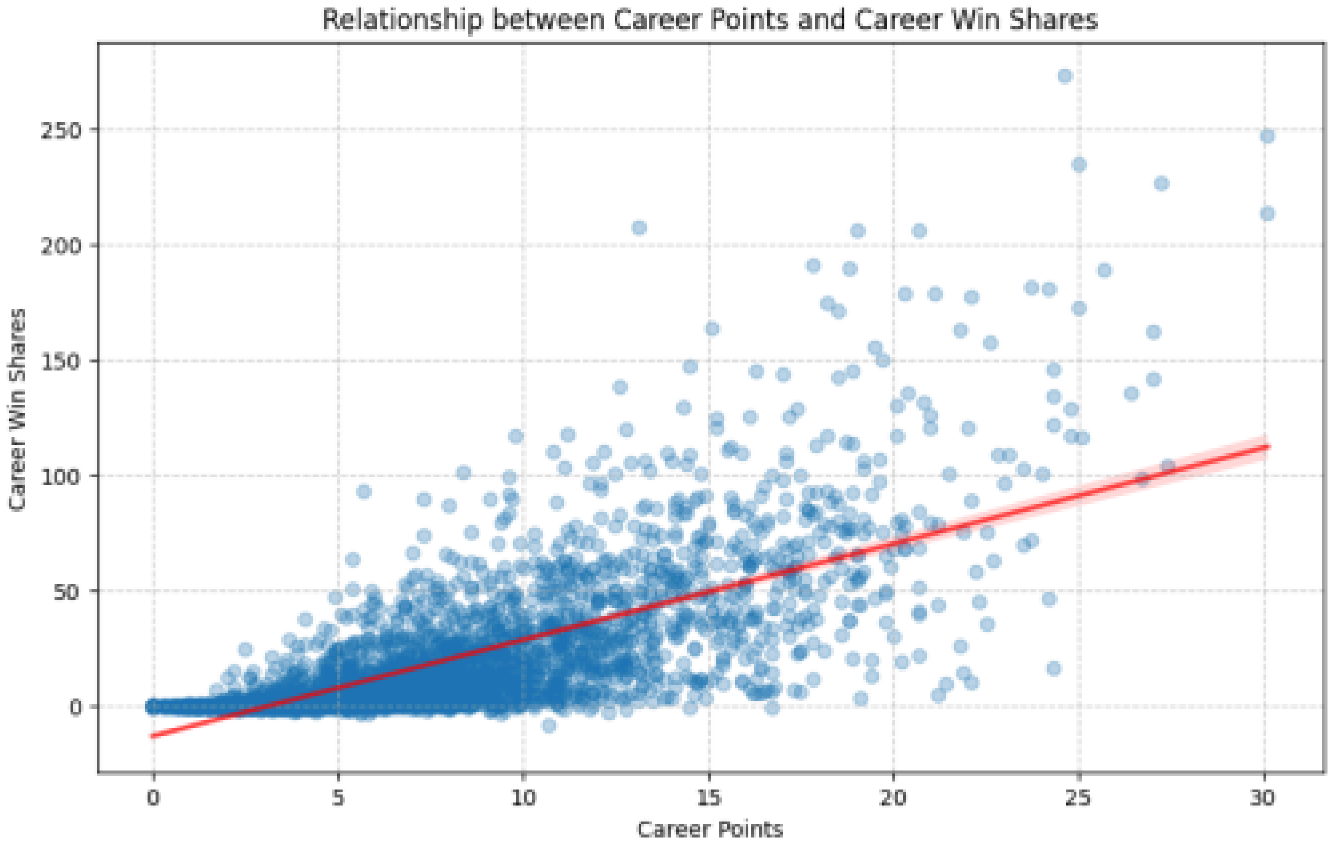
The left tail will fall below 40% and may reflect shorter career players or poor shooting efficiency.

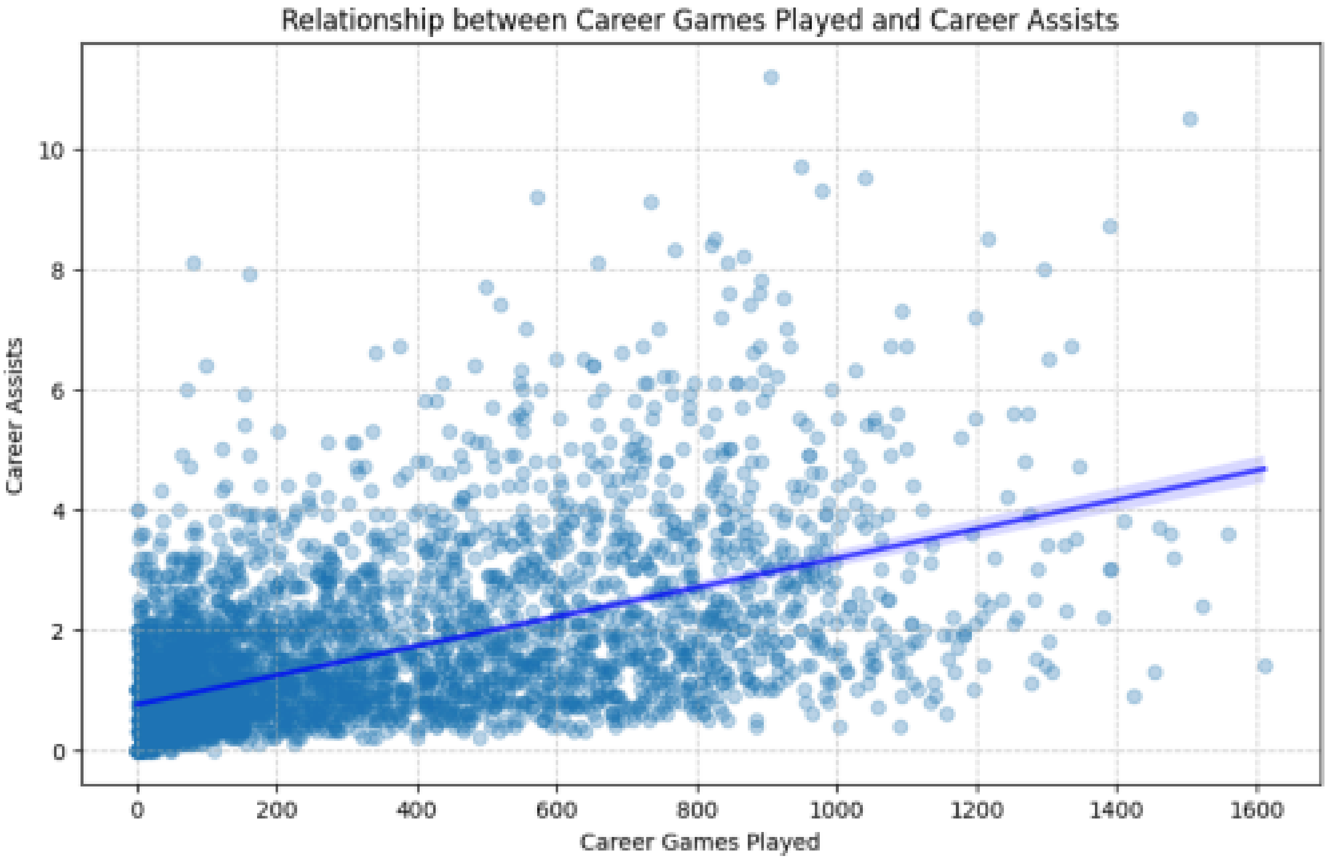


The tallest bar is near zero, meaning that thousands of players contributed minimally to team wins over their careers.

As the win shares increase, the number of players drops off dramatically, showing that sustained impact is rare.

Only a small number of players reach elite win share totals, which reflects both long careers and consistent performance.


Relationship between Career Points and Career Win Shares
Career Win Shares
Career Points
250
200
150
100
50
0
0
5
10
15
20
25
30

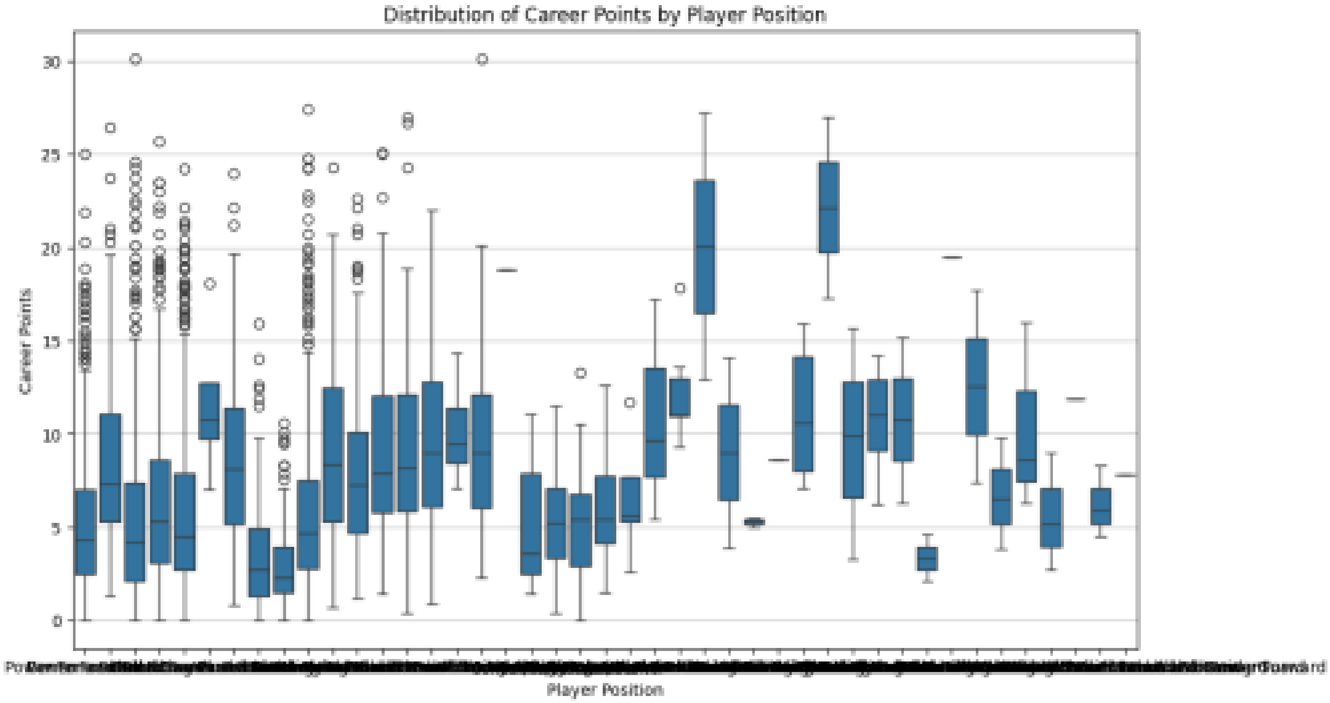


1. Scatter plot depicting the relationship between career_PTS and career_WS:

The scatter plot shows a very strong positive linear relationship between career_PTS and career_WS. This makes a lot of intuitive sense in sports: the more points that one scores, the more value he or she is bringing to the team in terms of wins, and thus the more win shares that player has. A linear trendline would clearly have an upward trajectory; as career points increase, so too do the career win shares. Within the observed range, there are no obvious diminishing returns: higher points mean higher win shares. There is some scatter, so not all high-point scorers have comparably high win shares, and vice-versa, due to other factors such as defensive contribution, efficiency, and overall team success that are not reflected by just points.

2. Scatter between Career Games Played (career_G) and Career Assists (career_AST):

This scatter plot shows a positive linear relationship between career_G and career_AST. This also intuitively makes sense in the context of sports: The more games a player plays, the longer or more consistent his career, the more assists he will have. This positive correlation is reflected in the trendline. Again, similar to win shares and points, the more games one plays, the more opportunities he has to get assists. Although the trend is generally linear, there's a large spread, especially for those with many games, with not all long-tenured players being high-assist players because of different roles, such as dedicated scorers and primary playmakers.



The above box plot describes the distribution of career_PTS across various player positions. Not surprisingly, Guards and Guard-Forwards have the highest median career points. This makes sense according to basketball intuition since the guards tend to be the primary ball-handlers and the main scorers.

The forwards and forward-centers also provide quite a variation in their career points, with a high number of high scorers, but their medians were a bit lower than the pure guards.

Most of the time, Cs tend to have more variability in their career points, meaning that usually, scoring output among the players of this position is variable. Whereas some of them tend to be high scorers, other Cs have specialized roles that might not include high or frequent scoring.

Summary of Anomalies and Handling Strategies

1. career_eFG% (0%, 100%, and 150%):

• Observation: There are players with `career_eFG%` values of 0%, 100%, and even 150%. By investigating the `career_G` for these subject players, it becomes clear that nearly all of them have played a very limited number of games, typically 1 to 6 games. The player with 150% eFG% is likely due to having made 3-pointers and having very few attempts. Effective Field Goal Percentage (eFG%) is calculated as `(FG + 0.5 * 3P) / FGA`. If a player only attempts a few shots and these are three-pointers, it's mathematically possible to achieve a very high eFG% if they make a high percentage of them-especially if they've made only 3-point shots and no 2-point shots. For example, if a player attempts 1 FGA and it's a 3-pointer that they make, then their eFG% would be (1 + 0.5 * 1) / 1 = 1.5 = 150%. Similarly, players with 0% eFG% likely attempted shots but made none in their very limited playing time.

Classification: These are not considered data errors, but rather legitimate extreme values, as there are very small sample sizes of career games and attempted shots in this instance. This is a true reflection of their performances, given the limited opportunities.

Handling Strategy: Keep these values as they are since they are statistically correct for the given counts of career game. But in analyses that rely on statistical significance or require bigger samples, these players could be filtered out or may be analyzed separately--for example, set a minimum threshold on `career_G` or `career_FGA`--to prevent their extreme values from disproportionately affecting aggregate statistics or models. One might consider excluding players with fewer than 10 or 20 `career_G` games from certain analyses.

2. `career_PER` (Player Efficiency Rating):

Observation: The minimum `career_PER` is -52.7. PER is a metric intended to summarize a player's per-minute statistical accomplishments while adjusting for pace and opposition. Negative PERs are statistically possible in basketball, reflecting extremely inefficient or damaging play. A player who commits many turnovers, misses many shots, and contributes little else could indeed have a negative PER, especially over limited playing time. The `df.describe()` output shows the 25th percentile at 9.3, a median of 11.7, and a mean of 11.3, so -52.7 is a significant outlier but not necessarily a data error.

Categorization: This is a legitimate extreme value, representing extremely poor performance over a player's career, likely correlated with very limited, unproductive playing time.

Handling Strategy: Leave these values as is. Negative PERs are informative. If analyses may be sensitive to extreme negative values, one might consider analyzing players with very low PERs (e.g., below 0 or some threshold) separately, or include `career_G` as a weighting factor in any aggregate calculations.

3. career_WS (Win Shares)

Observation The minimum career_WS is -7.9. Win Shares is an estimate of the player's contribution to the team's wins. While usually positive, a negative Win Shares value is plausible for players who are highly inefficient and whose presence on the court correlates with their team performing worse. This can happen over short careers or for players who consistently make costly mistakes or have very poor shooting percentages without compensating in other areas.

Categorization: This is a legitimate extreme value, indicating a player who, statistically, detracted from their team's wins.

Handling Strategy: Leave these values as is. Negative Win Shares are by design for this metric to represent contribution. Similar to PER, in any analysis where aggregate metrics are being used, one might want to consider career_G as a weighting factor. The dataset provides information for 4657 unique basketball players that were selected in the NBA draft between 1947 and 2018 without any missing values across the 25 columns. We see that the distributions of career_G, career_PTS, and career_WS are heavily right-skewed, with a high number of players having relatively short or low-impact careers, with far fewer players having long, high-scoring, highly impactful careers. The distribution of career_eFG% is somewhat left-skewed with a peak in the region between 45-50%. The extreme values of 0%, 100%, and 150% for career_eFG% were identified as legitimate statistical outcomes for players with very limited career games - typically 1-6 games, with few shot attempts - rather than data errors. Indeed, there is a strong positive linear relationship between career_PTS and career_WS, which intuitively makes sense given that higher scoring generally contributes to a player's value and wins for their team. Similarly, a positive linear relationship exists between career_G and career_AST, as more games provide more opportunities to accumulate assists. Comparison of career_PTS across player positions indicates that G and GF have the higher median career points. Centers have, on average, a wider range in career points, as they are often utilized for diverse purposes. Outliers, correspondingly considered as excellent scorers, are found across all positions. Outlier values in career_PER range from -52.7, and in career_WS range from -7.9; both are regarded as valid extreme values because they represent extremely inefficient or harmful play and usually coincide with very limited or unproductive playing time. Insights or Next Steps Where necessary (e.g., for some career efficiency metrics for short-career players), some minimum for career_G or career_FGA could be imposed to ensure more statistically reliable comparisons. Further analysis could go into the particular factors driving the negative career_PER or career_WS in players with more significant careers, to better understand nuances in player impact beyond simple cumulative statistics.