Leveling the Court: Modeling Shot Success Across Men's and Women's Professional Basketball

Julia Jeckering, Ananya Manglik, Abigail Schmid, April Wang

Introduction

In professional basketball, shot selection and scoring efficiency are critical components of game strategy. Yet, despite the growing attention to women's sports, there remains a perception that the dynamics of the women's game are fundamentally different, and less exciting, compared to men's basketball. This project tackles the question: **How does shot distance impact scoring probability differently in the NBA and WNBA?**

Understanding these differences can help challenge outdated narratives about women's sports and shed light on how play styles vary between leagues. As fans and analysts push for greater equality between men's and women's sports, having clear, data-driven insights into gameplay patterns is essential.

To explore this question, we analyzed detailed shot-level data from both leagues, focusing on data from the 2024 season. Using a Generalized Additive Mixed Model and Generalized Linear Mixed Model, that account for nonlinear effects and player-specific variability, respectfully, we compared how shot success changes with distance in each league. Preliminary results suggest that while shot success predictably declines with distance in both leagues, the pattern of decline differs: NBA players show a sharper drop-off in scoring probability as distance increases, while WNBA players maintain more stable shooting performance across mid-range and long-distance shots.

Through this project, we aim to provide a nuanced, quantitative view of shot efficiency differences, adding evidence to conversations about competitiveness, entertainment value, and skill in women's professional basketball.

Data

For this project, we analyzed shot selection and efficiency in professional basketball by comparing the NBA and WNBA. We used **play-by-play** data from the 2014 to 2024 seasons, accessed through the `wehoop` and `hoopR` R packages. These packages provide comprehensive and structured access to official game data, including shot attempts, coordinates, and outcomes.

Raw Data

We used the following functions to load the raw data:

- `load_wnba_pbp(seasons = 2014:2024)`
- `load_nba_pbp(seasons = 2014:2024)`

This raw dataset included every shooting play over a ten-year span for both leagues, capturing thousands of unique shot events across regular season and playoffs.

Pre-processing and Cleaning

Since the raw data was highly detailed and varied, we performed several preprocessing steps to make it usable for analysis:

- **Shot Type Simplication**: The NBA dataset contained 205 unique shot types, and the WNBA dataset contained 72 unique shot types. To simplify analysis, we categorized shots into five groups based on text descriptions:
  - Dunk
  - Hookshot
  - Layup
  - Jumpshot
  - 3PT Shot (added based on shot distance thresholds, detailed below)
- **Shot Distance Calculation**: Using raw shot coordinates, we computed shot distance relative to the basket center. X-coordinates were centered by subtracting 25, and Euclidean distance was calculated:

$$\text{shotdistance} = \sqrt{(x-25)^2 + (y)}$$

  For categorizing shots as three-pointers:
  - WNBA 3PT line was set at 22.14583 feet
  - NBA 3PT line was set at 23.75 feet

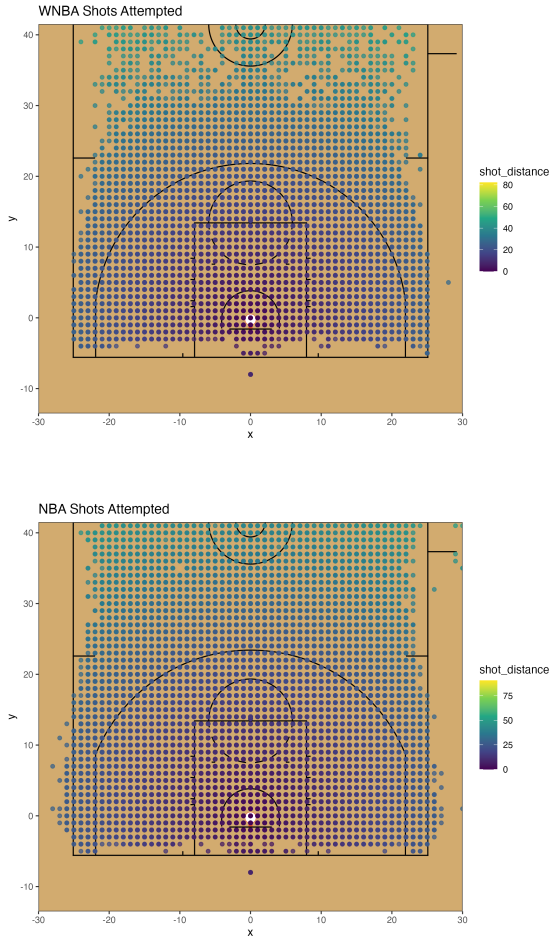
- **League Labels**: Each shot was tagged as either from the NBA or WNBA, and each player ID was prefixed (e.g., "NBA_12345").
- **Filtering Players**:
  - We kept only players who attempted at least 30 shots to avoid unstable statistics.
  - We further filtered out players whose shot distance standard deviation was less than 2 feet, ensuring we only analyzed players with meaningful shot variability.
- **Handling Missing Values**: Shots with missing distance, player ID, or scoring outcome were removed.
- **Relevant Columns for Modeling**: For modeling, we focused on a few key columns from the cleaned dataset:
  - `scoring_play`: Whether the shot was made (1) or missed (0).
  - `shot_distance`: How far the shot was from the basket, in feet.
  - `league`: Whether the shot was taken in the NBA or WNBA.
  - `athlete_id_1`: The player who took the shot, used to capture player-specific effects.

At the final stage, we created two datasets:

- For exploratory data analysis (EDA), we used all shot data from 2014–2024 (3,508,432 shots)
- For modeling, we restricted the data to the 2024 season to provide a consistent, contemporary snapshot of shot behavior (336,066 shots)

Exploratory Data Analysis

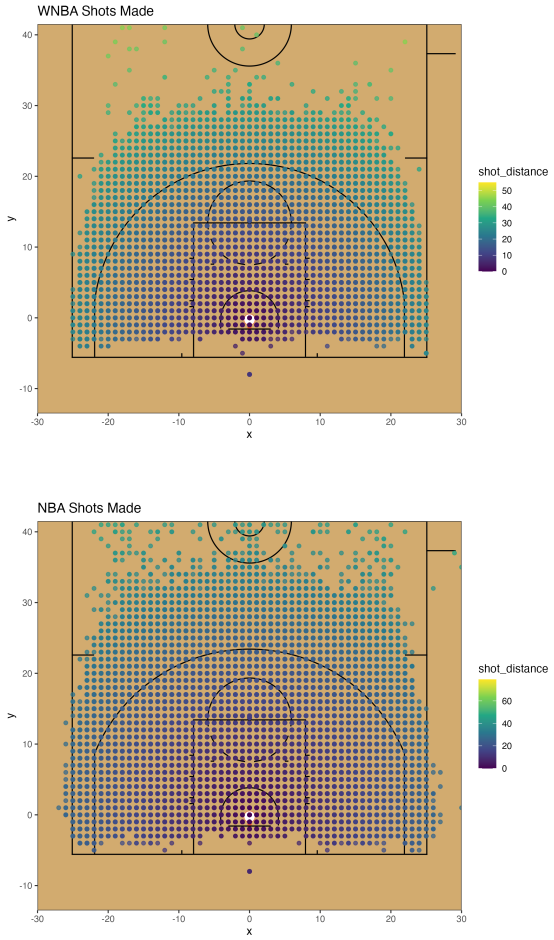
Shots Attempted



- **WNBA**: Shot attempts are more concentrated inside the arc, especially in the paint and floater/mid-range areas
- **NBA**: Shot attempts have dense clusters along the 3-point arc, edges, and corners

The results highlight that NBA players utilize greater floor spacing and long-range opportunities, while WNBA players focus on on high-efficiency zones closer to the basket.

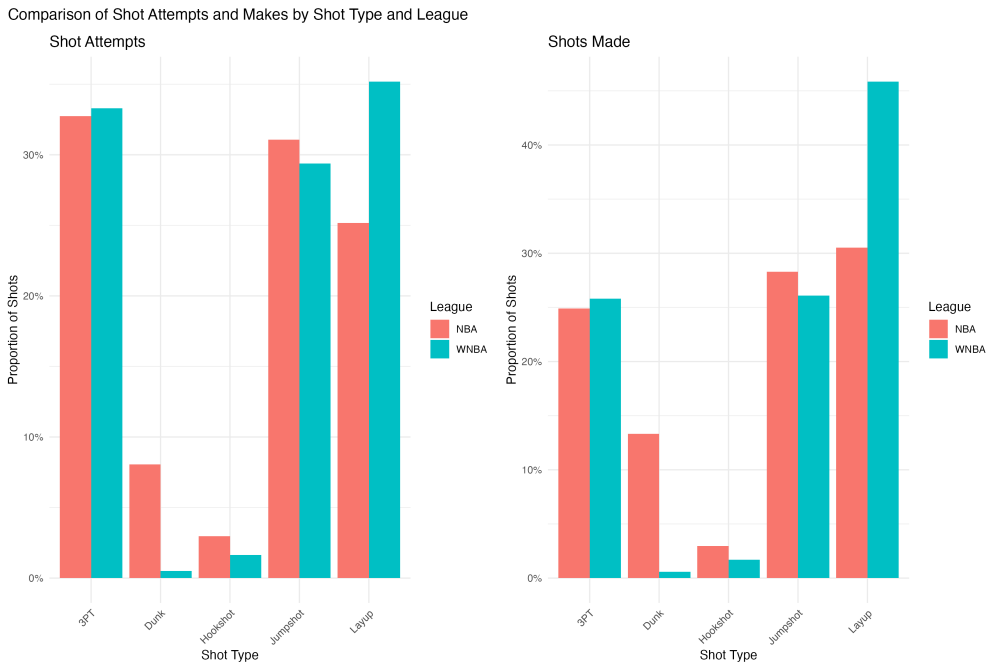
Shots Made



- **WNBA**: Made shots are tightly concentrated within 30 feet, with higher success in paint, floater/midrange areas, and just outside the arc
- **NBA**: Successful shots span a wider range, including deep threes and strong success at the rim

The results highlight that NBA players demonstrate greater long-range shot-making ability, while WNBA players maintain higher efficiency in closer zones.

Distribution of Shots



Shot Attempts:

- 3PT shots are the most frequently attempted shot type in both leagues (~35% of shots), with slightly more 3PT attempts in the WNBA.
- Layups are significantly more common in the WNBA than the NBA (~40% vs. ~25% of shots).
- Jumpshots are more common in the NBA than the WNBA (~32% NBA vs. ~29% WNBA).
- Dunks are almost exclusively an NBA phenomenon (about 8% of attempts in NBA vs. near-zero in WNBA).
- Hookshots make up a very small proportion of attempts in both leagues (~3–4%).

Shots Made:

- Layups dominate the made shots for both leagues, but especially in the WNBA (~45% of all makes).
- 3PT shots have a very similar made proportion in both leagues (~25%), suggesting relatively comparable success rates.
- Jumpshots account for a slightly higher proportion of made shots in the NBA (~28%) than in the WNBA (~25%).
- Dunks are again almost exclusively contributing to made shots in the NBA (~14%).

- Hookshots barely contribute to made shots in either league.

Summary of EDA: Our EDA reveals that distance and shot zone play a major role in shaping shot patterns between the NBA and WNBA. NBA players rely more heavily on volume and floor spacing across the court, while WNBA players focus on efficient, high-percentage zones. These patterns support the need for nonlinear and league-specific modeling approaches. Based on these insights, we proceeded by fitting both a Generalized Additive Model (GAM) and a Generalized Linear Mixed Model (GLMM) to more rigorously quantify and compare shot success trends across the two leagues.

Methods

To study differences in shot success between the NBA and WNBA, we fit two statistical models: a **Generalized Additive Model (GAM)** and a **Generalized Linear Mixed Model (GLMM)**. Both models were chosen because they allow flexible estimation of how shot distance impacts shot success, while accounting for player-specific variability.

Generalized Additive Model (GAM)

We first estimated a GAM of the following form:

$$y_i \sim \text{Bernoulli}(\pi_i) \quad \text{logit}(\pi_i) = \beta_0 + \beta_1 \cdot \text{league}_i + f_{\text{league}_i}(\text{shotdistance}_i)$$

where:

- y_i: indicates whether shot i was made
- π_i: probability shot i was made
- league_i: indicator variable (NBA vs WNBA)
- $f_{\text{league}_i}(\cdot)$: smooth function for shot distance, estimated separately for each league

Assumptions:

- The outcome variable (y_i) follows a Bernoulli distribution.
- Observations are independent conditional on the predictors (shot distance, league).
- Smooth functions $f(\cdot)$ are flexible but penalized to avoid overfitting.

Justification: We chose a GAM because shot success probability is likely nonlinear in distance, especially across different leagues where play style differs. Allowing a separate smooth function by league captures the possibility that distance impacts NBA and WNBA players differently.

Evaluation Approach: To evaluate the GAM, we compared its AIC (Akaike Information Criterion) value against those from other models (GLM and GLMM). A lower AIC indicates better model fit with appropriate complexity penalization. Additionally, we examined the significance of the smooth terms to assess whether the relationship between shot distance and scoring probability differed meaningfully across leagues. This allowed us to quantify uncertainty and check if nonlinear distance effects were captured appropriately.

Generalized Linear Mixed Model (GLMM)

To account for player-specific effects, we additionally fit a GLMM:

$$\begin{aligned} y_i &\sim \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 \cdot \text{shotdistance}_i + \beta_2 \cdot \text{league}_i + \beta_3 \cdot (\text{shotdistance}_i \times \text{league}_i) \\ &\quad + u_{0,\text{athlete[i]}} + u_{1,\text{athlete[i]}} \cdot \text{shotdistance}_i \end{aligned}$$

where:

- β_0 to β_3: fixed effect coefficients
- $u_{0,\text{athlete[i]}}$: random intercept for player i
- $u_{1,\text{athlete[i]}}$: random slope for shotdistance_i for player i

Assumptions:

- y_i follows a Bernoulli distribution conditional on the random effects.
- Random effects (u_0, u_1) are normally distributed with mean 0 and constant variance across players.
- Observations are conditionally independent given player random effects and fixed predictors.

Justification: Players vary in their individual skill levels and shooting tendencies. Including random intercepts and slopes allows us to capture these player-specific deviations from the population-average relationship between shot distance and scoring probability. This avoids misattributing differences due to player ability to league-level effects.

Evaluation Approach: The GLMM was evaluated through comparison of AIC against the GAM and GLM models to assess model fit. We also examined the significance of fixed effects (league, shot distance, and their interaction) and assessed the variance of random effects to determine the extent of player-specific variability. A small random slope variance indicated that while player-level adjustments were included, overall league patterns dominated, supporting the robustness of league-level conclusions.

To assess the uncertainty in athlete-specific effects, we ran a bootstrap analysis with 100 resamples. Bootstrapping allows us to quantify how stable each player's estimated impact is across different subsets of the data, especially in the presence of sparse or uneven shot distributions.

Results

We compared three models — a Generalized Linear Model (GLM), a Generalized Additive Model (GAM), and a Generalized Linear Mixed Model (GLMM) — to assess how shot distance impacts shot success across the NBA and WNBA. Model performance was evaluated primarily via AIC, while interpretation focused on how distance and league differences influence scoring probability.

Model Comparison

Table 1: Model Summary

Model	AIC	Significant_Terms
GLM	449829.9	Shot Distance, League, Distance × League
GAM	432331.2	Smooths (shot_distance × league)
GLMM	449318.9	Shot Distance, League, Distance × League

Table 2: Model Summary

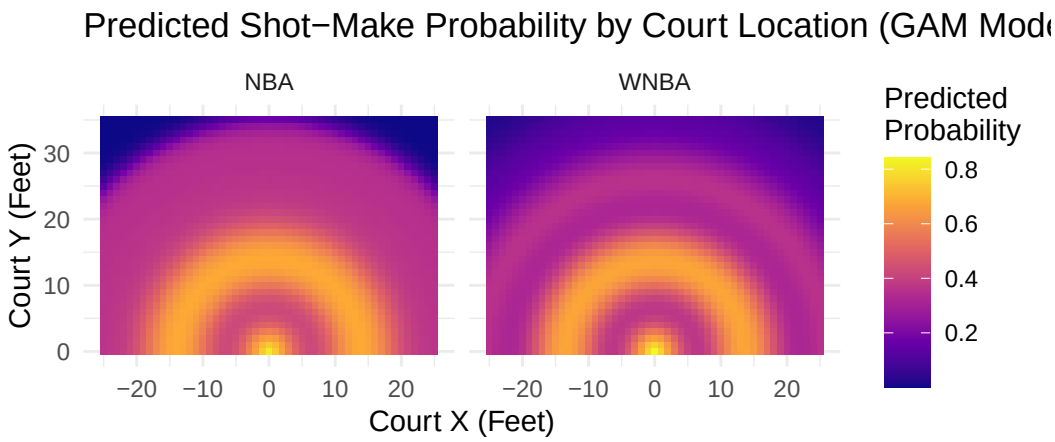
Model	League_Effect	Distance_Effect	Interaction_Effect	Notes
GLM	Yes ($p < 0.001$)	Yes ($p < 0.001$)	Yes ($p = 0.0018$)	Linear; no random effects
GAM	No ($p = 0.297$)	Yes (smooth significant)	Yes (smooths differ)	Nonlinear; flexible smooths
GLMM	Yes ($p < 0.001$)	Yes ($p < 0.001$)	Yes ($p = 0.0046$)	Random slopes (tiny variance)

Tables 1 and 2 summarize key findings across models:

- Table 1 shows that the GAM achieved the lowest AIC (432,331.2), substantially outperforming both the GLM (449,829.9) and GLMM (449,318.9). This suggests that allowing for nonlinear effects of distance on shot success improves model fit.
- Table 2 highlights that in the GAM, the league fixed effect itself was not statistically significant ($p = 0.297$), while distance and the interaction between distance and league were significant through differing smooths. In contrast, the GLM and GLMM found significant effects for league, distance, and their interaction, but assumed more rigid (linear) relationships.

These comparisons indicate that nonlinear modeling provides a better and more nuanced understanding of shot success differences between leagues, supporting our hypothesis that distance impacts shot probability differently in the NBA and WNBA.

Shot Success Across the Court

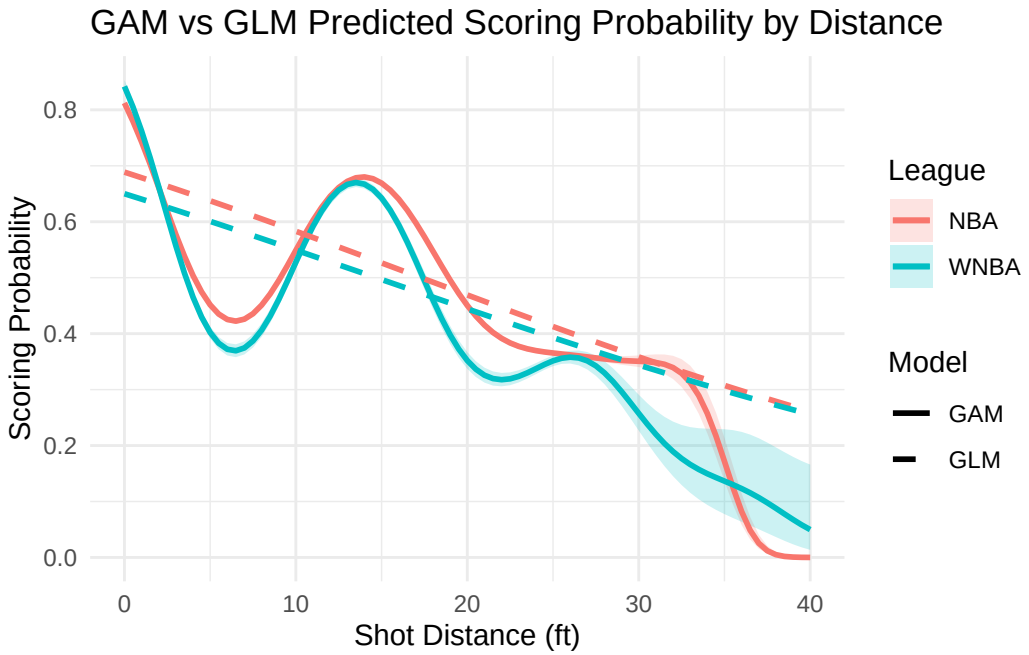


The first figure, "Predicted Shot-Make Probability by Court Location (GAM Model)", visualizes predicted probabilities across court coordinates:

- In both leagues, predicted shot success is highest near the basket and declines with distance.
- NBA players show more intense clustering of high-probability zones near the rim, reflecting greater close-range dominance (e.g., dunking ability).
- WNBA players exhibit a slightly broader spread of moderate-probability regions, suggesting a flatter decay of success with distance.

This spatial pattern supports the idea that shot difficulty rises with distance in both leagues, but NBA players show sharper performance drops beyond close-range zones.

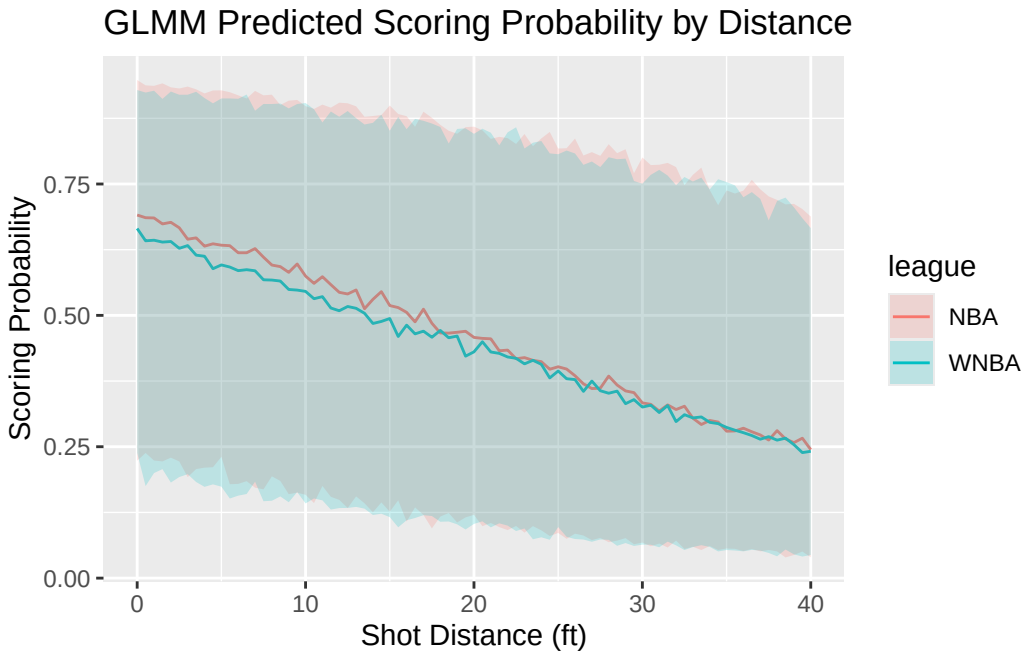
Scoring Probability by Shot Distance



The second figure, "GAM vs GLM Predicted Scoring Probability by Distance," compares model-predicted scoring probabilities:

- Both models show a decline in scoring probability with distance.
- However, the GAM captures subtle nonlinearities — especially for NBA players — that the GLM smooths over with its linear assumption.
- NBA shot success falls off more sharply after ~10 feet compared to the WNBA, consistent with gameplay differences where NBA players often exploit very close shots, while WNBA players maintain more consistent success even at mid-range.

Accounting for Player-Level Variability

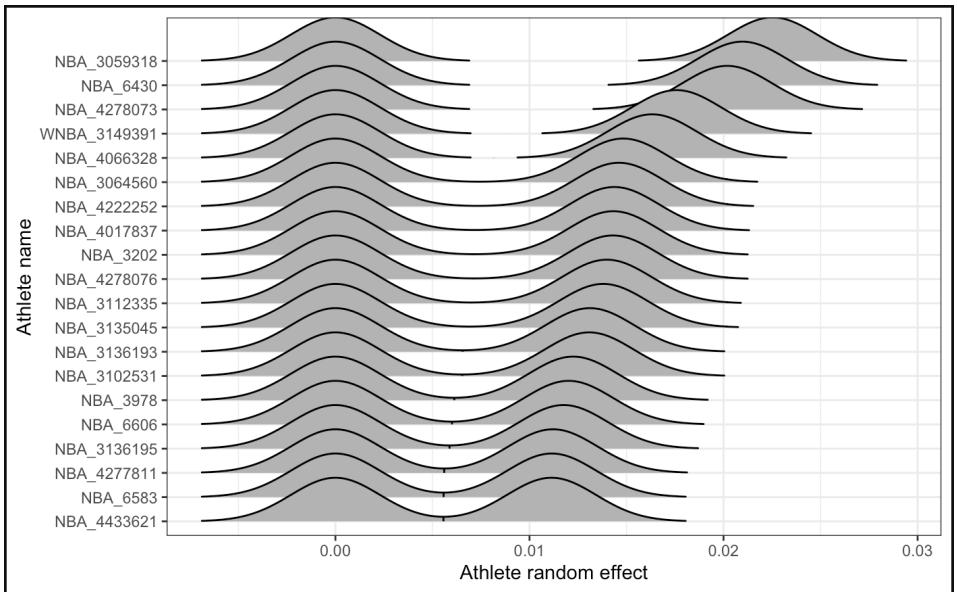


The third figure, "GLMM Predicted Scoring Probability by Distance," shows scoring probability trends accounting for random player effects:

- Even after adjusting for individual players, the overall trend persists: NBA players experience a steeper decline in shot success with distance than WNBA players.
- Random slopes were estimated, but the variance was extremely small, indicating that while individual players vary slightly, the overall league-level patterns dominate.

This minimal random slope variance reassures us that the differences we observe are not driven solely by a few outlier players, but reflect broader structural differences between leagues.

Uncertainty in Athlete-Level Effects from Bootstrap Resampling



To further assess the variability in athlete-specific effects, we ran a bootstrap analysis with 100 simulations. In each round, we randomly resampled game data and refit our model, which accounts for shot distance and league, while allowing each athlete to have their own influence on scoring. We sampled within each league and game_id to avoid impossible datasets that will not show the variability we are interested in. We focused on the top 20 athletes with the highest median effects and looked at how their estimates varied across simulations.

The ridge plot below shows this uncertainty, athletes with narrow curves had more consistent results, while those with wider curves had estimates that changed more from one simulation to the next. This means that even if an athlete has a high average effect, we should be cautious if their estimate isn't stable. For all the athletes, the distribution of random effects appeared bimodal, indicating that the estimated effect shifted significantly across bootstrap samples. This could reflect instability in the estimation process — often due to sparse or inconsistent data for that athlete. As a result, we interpret these cases with caution, since the model does not consistently converge on a single estimate for their effect. Overall, this helps us avoid over-interpreting results and reminds us that some player effects are more certain than others.

Uncertainty Estimates

We accounted for uncertainty in several ways across our models. For the GLM and GLMM, we assessed statistical significance using p-values for fixed effects, focusing on whether distance, league, and their interaction were meaningfully associated with shot success. In the GAM,

we evaluated uncertainty through the significance of the smooth terms, which allowed us to capture and test for nonlinear effects across leagues. To address variability across individual players, the GLMM incorporated random intercepts and random slopes for shot distance, though the random slope variance was very small, suggesting that player-specific deviations were minimal relative to league-wide trends. Together, these approaches ensured that our interpretations of league and distance effects reflect real underlying patterns rather than noise or sampling error.

Discussion

Conclusions

Our analysis highlights important differences in how shot distance affects scoring probability between the NBA and WNBA. Across all models, we consistently found that scoring probability declines with distance in both leagues. However, NBA players experience a sharper drop-off in scoring success at longer distances compared to WNBA players, who maintain more consistent shooting performance across mid- and long-range attempts. This suggests that while NBA players may dominate in close-range opportunities (such as dunks), WNBA players display a steadier efficiency even as shot distance increases.

The findings provide evidence against the perception that women's professional basketball is less skilled or less exciting. Tnstead, it reveals different, but equally sophisticated, gameplay strategies between leagues. By modeling both nonlinear effects (through the GAM) and player-level variability (through the GLMM), we were able to robustly support these conclusions while addressing potential confounding factors like individual player skill.

Limitations

While our project provides valuable insights, there are several limitations worth noting. First, although our GAM and GLMM captured important aspects of the data, we recognize that a Generalized Additive Mixed Model (GAMM) would have been a more appropriate modeling choice. A GAMM would have allowed us to simultaneously model nonlinear effects of shot distance while accounting for random player-specific variation. However, due to computational limitations, specifically, model fitting times exceeding four hours even on powerful personal machines, we were unable to reliably fit a GAMM for the full dataset.

Additionally, while we accounted for player variability using random intercepts and slopes in the GLMM, the random slope variance was extremely small, limiting the potential benefits of mixed modeling. This could reflect true low variability among players, but it may also point to a need for models better suited to capture subtle player-level differences or interactions with game context (such as shot clock or defensive pressure). Finally, by restricting modeling to the 2024 season for comparability, we sacrificed the ability to study trends over time or examine season-to-season shifts in shot behavior.

Future Work

Future work could address several of these limitations and build on our findings. First, using more powerful computing resources would allow us to fit a full GAMM, giving a more complete picture of how nonlinear effects and player-specific variability interact. Second, expanding the models to include additional game context variables (such as shot clock timing, defender

proximity, or score differential) could yield deeper insights into shot selection strategies in late-game or high-pressure situations.

Another promising direction would be to model shot types (e.g., layups, floaters, three-pointers) directly as a multinomial outcome, rather than focusing solely on made/missed shots. This could uncover whether league differences also manifest in how types of shots are selected at different distances. Finally, longitudinal analysis across multiple seasons could help assess whether the observed differences are stable over time or shifting as the WNBA and NBA evolve.