Predictive Modeling for Injury Risk and Recovery Time for WNBA Athletes

Submitted in fulfilment of the requirements for
The Dietrich College Senior Honors Thesis

Anahita Hassan

Department of Statistics and Data Science
Dietrich Senior Honors Thesis
Carnegie Mellon University
April 2025

Advised by Prof. Ron Yurko

Abstract

This research develops and evaluates several models for predicting injury risk among WNBA players, using publicly available injury and performance data from the 2024 season. The study addresses gaps in current injury prediction models, which are often based on male athletes and overlook the unique injury risks of female athletes. First, a mixed-effects Gamma regression model is trained on non-injured players to establish a baseline for expected minutes played, using fixed effects (starter status, position) and random effects (team, player) to account for variability across teams and playing styles. The model is applied to injured players, and residuals from games leading up to injury are analyzed to identify deviations that may indicate elevated injury risk. In another section, injury risk is predicted using a binary classification model (logistic regression), which estimates the likelihood of a player sustaining an injury in the time period between the most recently played game and the next future game, using predictors such as prior injuries, minutes played, and total points scored. The research also includes models that predict injury severity, using total missed days due to the injury as a proxy for injury severity. Various models – including Zero-Inflated Negative Binomial (ZINB), Negative Binomial (NB), Multinomial Logistic Regression, and Random Forest – are developed and compared to identify the most effective modeling approach. Ultimately, by modeling injury and performance trends in WNBA athletes, this research aims to support a deeper understanding of injury risk factors, informing future work in sports analytics and athlete health.

Table of Contents

Chapter 1: Exploratory Data Analysis (EDA)

1.1 Introduction

As sports analytics evolves, injury prevention has become a critical focus for safeguarding athletes and optimizing long-term performance. Women's sports, particularly the Women's National Basketball Association (WNBA), are leading this transformation, with the 2024 season marking a 170% increase in viewership, averaging 1.2 million viewers per game across ESPN platforms [1]. This surge, alongside a historic $2.2 billion media rights deal, underscores the need for advanced, data-driven tools to prioritize athlete health and safety [2].

However, existing injury prediction models are limited in several key areas. Current research in injury prevention and modeling has largely focused on male athletes, leading to less accurate predictions and treatment plans for women. Additionally, there is greater focus on acute (in-game) injury prediction models than on the role of long-term performance in injury risk. Furthermore, there are methodological gaps in the field. While current models often rely on classification or regression techniques, mixed-effects models – which account for both fixed and random effects – are underutilized and may offer a more comprehensive approach.

This research seeks to fill this gap by developing injury prevention models for the WNBA, based on the 2024 season stats and injury records. By applying a mixed-effects model, this study aims to identify the key factors that influence injury risk throughout a season – considering player, team, and game level dynamics. This approach will provide a more nuanced understanding of injury risk, considering the unique factors affecting female athletes, ultimately improving injury prevention strategies. This first chapter presents an exploratory data analysis (EDA) of the WNBA injury and performance data, offering insights that can inform these predictive models.

1.1.1 Research Gaps

A critical gap in injury prediction lies in the gender-specific nature of existing models, which are primarily based on male athletes. They are not accurate for female athletes in the WNBA, as women face unique injury risks due to biological differences in biomechanics, balance, and anatomy [3]. Women, for example, tend to have lower centers of gravity than men, which affects their balance and stability, making them more susceptible to lower extremity injuries like ACL tears. Additionally, women often have less muscle mass in certain areas, increasing their vulnerability to muscle strains and overuse injuries. This research aims to develop more accurate and reliable injury prediction models for female athletes.

The WNBA has been grappling with a significant injury problem, which became even more apparent since the 2023 season. *The Next* has reported a sharp increase in injuries led to 176 reported incidents, resulting in 789 missed games and a substantial loss of team performance, measured in win shares [4]. A disproportionate number of these injuries affected ankles, knees, and feet, with backcourt athletes bearing the brunt of the impact. Despite the rising concern, injury occurrence in the WNBA remains vastly understudied. The lack of comprehensive injury data limits the ability to improve rehabilitation processes and develop preventive strategies. Notably, the WNBA was absent from the 2023 Elite Basketball Rehab Conference, a key event for basketball medical providers, where NBA teams presented extensive injury data [5]. This gap highlights the league's insufficient focus on injury tracking. The absence of WNBA injury data in scientific and medical circles not only hinders injury prevention efforts but also complicates the rehabilitation of athletes.

Most existing models focus on in-game injuries, relying on technologies like motion capture, wearable sensors, and videography to detect sudden changes or irregular movements that precede an injury. While these methods are effective for capturing acute risks, there is a lack of research investigating how long-term athlete performance and health deviations could signal a heightened risk of injury. This study aims to address this gap by exploring how gradual changes in performance, health, and injury status leading up to a particular game can serve as early warning signs of injury.

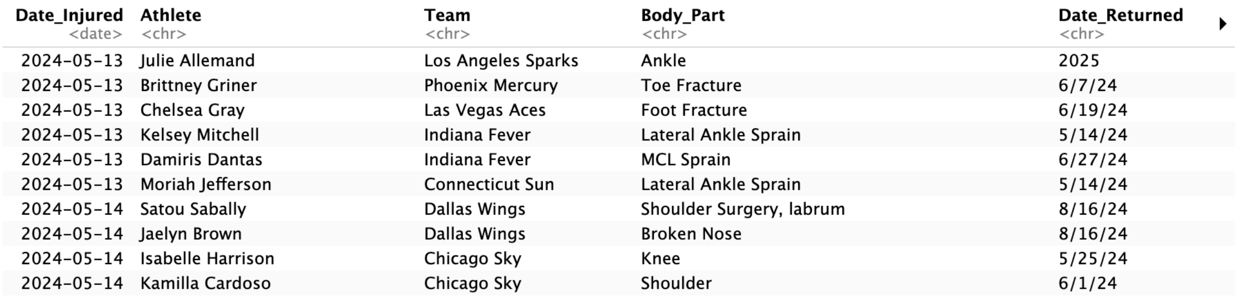
Most current injury prediction models use classification or regression techniques, which focus on predicting discrete outcomes, such as whether an injury will occur. However, this research will use a mixed-effects model, a less common but potentially more effective approach for understanding injury risk. This methodology will allow for a more nuanced analysis that incorporates player, team, and game level factors, providing deeper insights into the complex interactions that contribute to injury risk.

1.2 Dataset Overview

1.2.1 Data Sources

The 2024 WNBA season player statistics were sourced from the *Wehoop* package [6], which provides detailed game level data for all athletes. The dataset includes key performance metrics such as minutes played, points, rebounds, assists, and other player statistics for each game.

The injury data for this study was sourced from *The Next*'s WNBA Injury Tracker [7], which provides a comprehensive and up-to-date log of injuries across the league during the 2023 and 2024 seasons. This publicly available dataset includes details on the body parts affected, games missed, and estimated win shares lost for individual players. It also offers cumulative data, highlighting the most common injuries and their impact on team success. This tracker aims to shed light on injury trends in the WNBA, contributing to better health outcomes and understanding of injury patterns within the league. Here's a preview of the first 10 rows from the dataset, showcasing a subset of the columns:



Date_Injured	Athlete	Team	Body_Part	Date_Returned
<date>	<chr>	<chr>	<chr>	<chr>
2024-05-13	Julie Allemand	Los Angeles Sparks	Ankle	2025
2024-05-13	Brittney Griner	Phoenix Mercury	Toe Fracture	6/7/24
2024-05-13	Chelsea Gray	Las Vegas Aces	Foot Fracture	6/19/24
2024-05-13	Kelsey Mitchell	Indiana Fever	Lateral Ankle Sprain	5/14/24
2024-05-13	Damiris Dantas	Indiana Fever	MCL Sprain	6/27/24
2024-05-13	Moriah Jefferson	Connecticut Sun	Lateral Ankle Sprain	5/14/24
2024-05-14	Satou Sabally	Dallas Wings	Shoulder Surgery, labrum	8/16/24
2024-05-14	Jaelyn Brown	Dallas Wings	Broken Nose	8/16/24
2024-05-14	Isabelle Harrison	Chicago Sky	Knee	5/25/24
2024-05-14	Kamilla Cardoso	Chicago Sky	Shoulder	6/1/24

1.2.2 Research Questions and Approach

This study aims to investigate the relationship between player workload and injury occurrence in the WNBA. The following key questions are addressed:

1. How does playing time impact injury risk?
    - To explore this, a model is trained on non-injured players to establish baseline expected minutes. Applying the model to injured players reveals deviations from expected play, helping identify patterns of overuse that may signal elevated injury risk.

2. Can injury risk be predicted between consecutive games?
    - This question will be addressed using a binary classification model that estimates the likelihood of an injury occurring between the current and next game, based on predictors such as prior injuries, minutes played, and points scored.

3. Can the duration of a player's injury be predicted?
    - Recovery duration can be modeled using cumulative factors leading up to injury – including injury location, minutes played, and prior injuries – to predict days missed due to injury, which is a metric that also serves as a proxy to injury severity.

To address these questions, an Exploratory Data Analysis (EDA) will first be conducted to understand injury distributions and player workload patterns. Then, mixed effects models will be created and evaluated to estimate player usage and injury trends in Chapter 2, addressing the first research question. In Chapter 3, the second and third research questions will be answered by creating a binary classification model to predict injury risk in an upcoming game, along with several other models to predict recovery time.

1.3 Injury Data Patterns and Insights

1.3.1 Injury Data and Type Overview

The available injury data was collected for the 2023 and 2024 seasons. In 2023, there were 175 injuries, and there was an increase to 203 injuries in the 2024 season. Over two years, 96 unique injury descriptions were identified. These descriptions specify the body part affected by the injury and are recorded as string values in the dataset. Examples range from general terms like "Knee" to more specific descriptions such as "Right Knee Hyperextension." To streamline the analysis and identify patterns more effectively, these injuries were manually grouped into 13 broader categories. This grouping approach was essential for simplifying the dataset, reducing ambiguity, and making it easier to observe trends. Injuries were grouped into 13 categories using the `case_when` function in R, which standardizes the variations in terminology into broader groups:

- Face – Includes injuries such as jaw fractures, facial lacerations, and broken noses.
- Head – Covers injuries like concussions and other head-related conditions.
- Abdomen & Back – Encompasses injuries from the lower back to the ribs.
- Shoulder – Ranges from shoulder labrum surgery to shoulder subluxation.
- Arm/Hand – Includes injuries such as broken fingers and hand fractures.
- Hip – Covers conditions like general hip injuries and hip labrum repair recovery.
- Knee – Includes injuries from minor knee issues to major conditions like ACL tears and meniscectomy.
- Ankle – Covers various ankle injuries, including Achilles and lateral ankle sprains.
- Foot – Includes injuries such as toe fractures, plantar fasciitis, and foot fractures requiring surgery.
- General Leg Injuries – Encompasses injuries affecting the leg, calf, and thigh.
- Rest & Mental Health – Covers instances labeled as "Rest" or mental health-related absences.
- Illness – Includes illnesses ranging from general sickness to COVID-related absences.
- Undisclosed – Consists of cases where the injury was not specified.

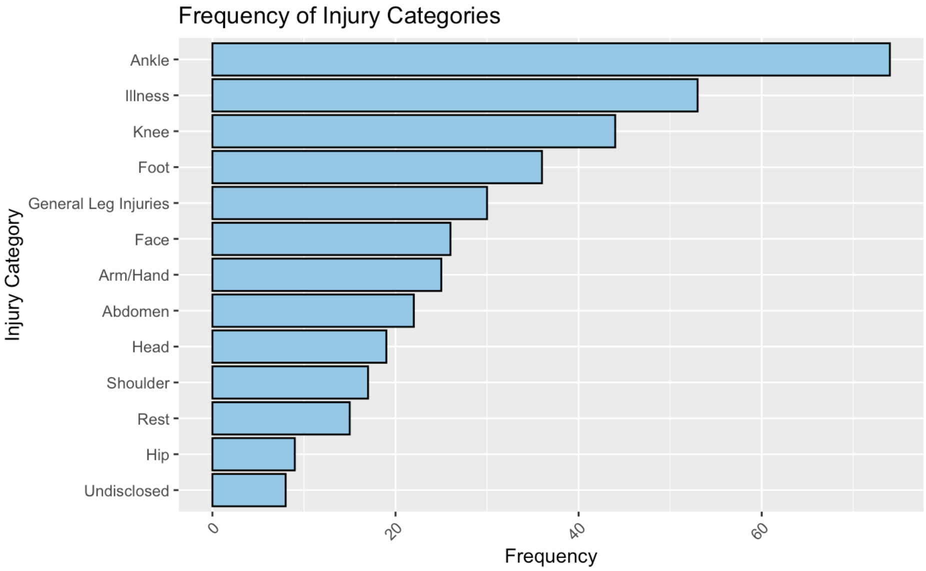


Figure 1.1 **Most common injury categories among reported cases.** Ankle, knee and foot injuries are among the most common, highlighting a predominance of lower extremity cases.

According to Figure 1.1 above, ankle injuries appeared to be most common, followed by knee and foot injuries. This reaffirms the existing research showing women are more susceptible to lower-extremity injuries [8]. Absences due to illness also are extremely common.

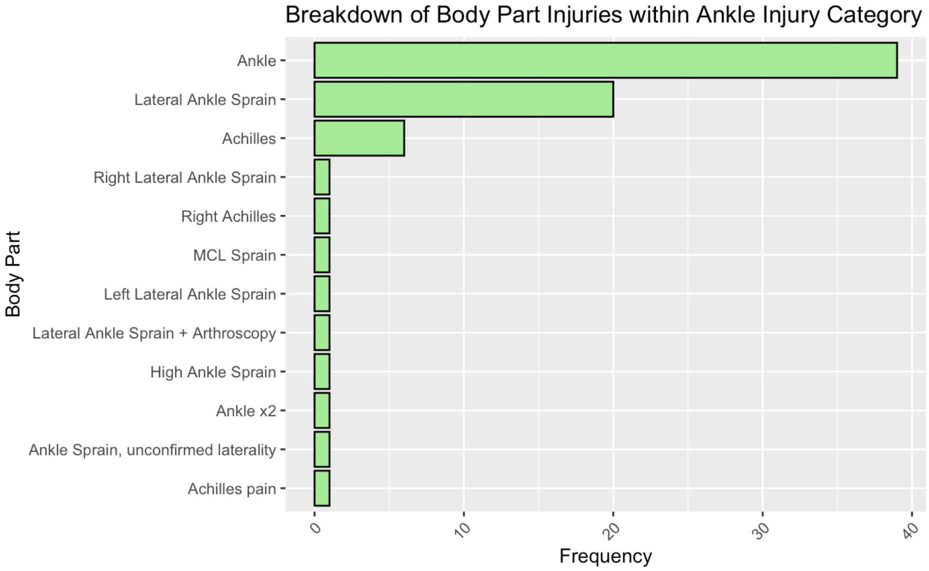


Figure 1.2. **Breakdown of ankle-related injuries.** Vast majority of ankle injuries were just labeled as “ankle” with “lateral ankle sprain” being the next most common.

Shown above, a further breakdown of the "Ankle" category shows that many ankle-related injuries were simply labeled as "Ankle" in the original dataset, with sprains and Achilles injuries being the most common. This highlights the value of using the 13 broad categories that were manually made, like "Ankle," to group vague or overly specific injury descriptions (e.g., "Ankle" or "Lateral Ankle Sprain + Arthroscopy"). This approach simplifies analysis by creating a more organized structure, making it easier to identify patterns and trends across related injuries.

1.3.2 Injury Impact by Player Position

In the player performance data, positions are categorized as center, frontcourt (forwards), and backcourt (guards). However, the injury dataset only uses frontcourt and backcourt. To understand how centers are classified in the injury dataset, we identified all athletes labeled as centers in the performance data and checked their positions in the injury dataset. All of them were listed as frontcourt, indicating that in the injury dataset, "frontcourt" includes both forwards and centers, while "backcourt" refers to guards. According to the injury dataset, across both seasons, there were 63 frontcourt and 83 backcourt players who got injured at least once.

A chi-squared test was conducted to determine if player position (backcourt vs. frontcourt) affects the types of injuries sustained. The p-value from the test was 0.595, well above the 0.05 significance threshold, indicating no significant association between player position and injury type. Thus, position does not appear to influence injury type.

1.3.3 Injury Impact by Recovery Time

Next, we examine how injury category types relate to total days missed due to the injury, which is calculated as the difference between the date of injury and the date of return, with records containing "NA" values or vague dates like "March" or "2025" being excluded.

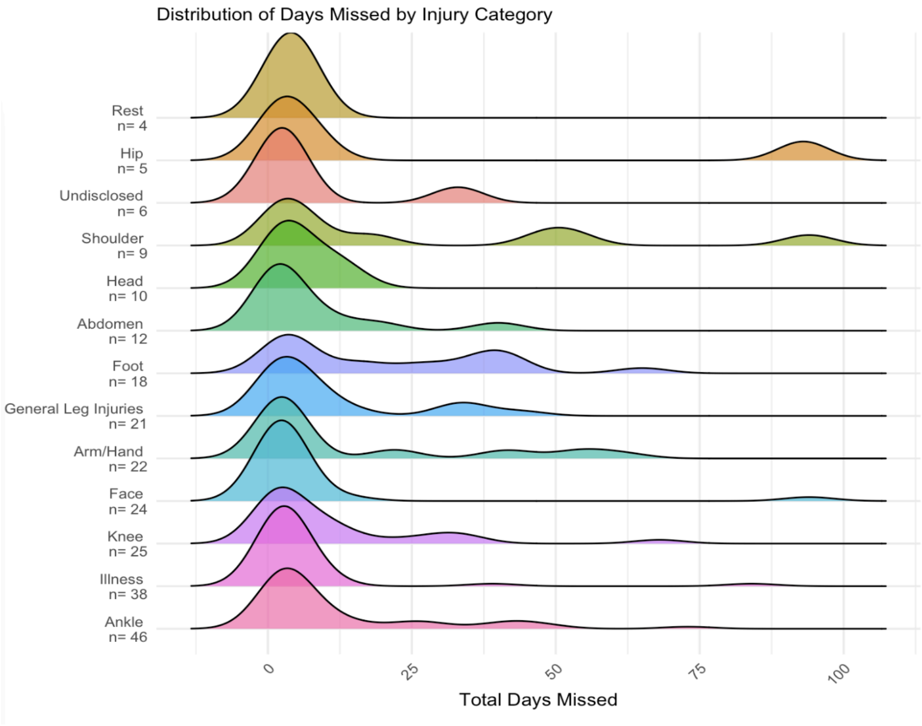


Figure 1.3. **Ridgeline plot of days missed due to injury, by injury category.** Most injuries result in short recovery periods, though some categories show long-tailed distributions.

The figure above reveals that ankle and knee injuries, which are most common, have long right tails, meaning many cases involve short absences, but some result in longer recovery times. Hip injuries have a bimodal distribution, with peaks around 1–2 days and 90 days missed, likely due to surgery and recovery time. Shoulder injuries show a multimodal pattern, with peaks at 1–2 days, 50 days, and 90 days, indicating varying severity. Foot injuries have a wide, flat distribution, showing high variability in recovery time. General leg and arm/hand injuries show similar patterns. Notably, shoulder, hip, and face injuries are the only ones with densities extending into the 75–100 days missed range, suggesting they were likely season-ending. This difference is likely due to treatment, injury severity, and the demands of returning to play.

1.3.4 Seasonal Injury Trends and Player Workload

For the analysis below and throughout the paper, injury data from the 2024 season is the primary focus, as it is the most comprehensive and recent. Additionally, certain dates that could introduce bias into the modeling, such as the 2024 Olympics and the All-Star Game, will be filtered out in later analyses.

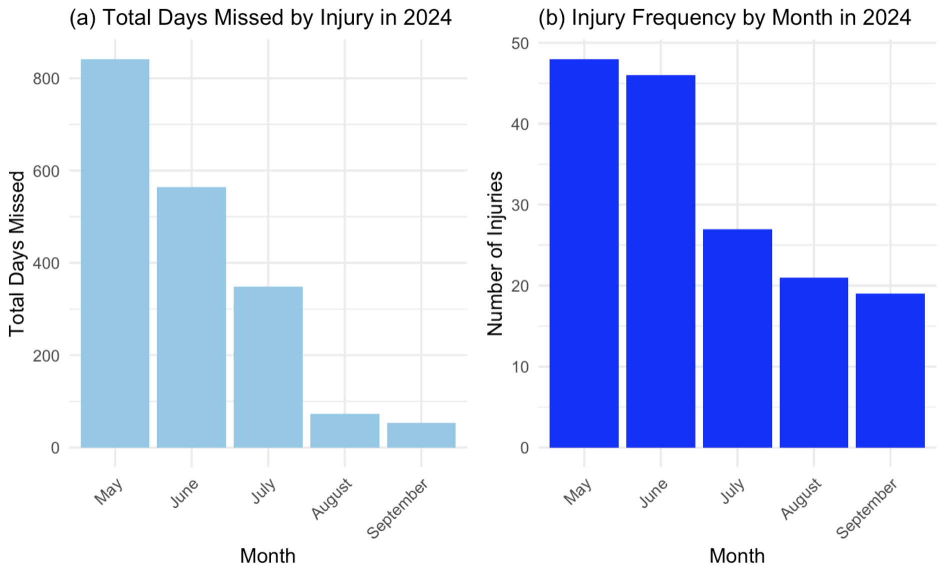


Figure 1.4. **Total days missed due to injury (a) and injury frequency (b) by month in 2024.** May and June accounted for the highest number of days missed and the highest injury counts, suggesting that early-season injuries were more severe or required longer recovery periods.

Figure (a) on the left shows the total number of days missed due to injuries each month in the 2024 season. For instance, if an injury occurred in May resulting in 50 missed days, all 50 days would be counted toward May. This allows us to visualize the impact of injuries on player availability, highlighting periods of greater disruption. May had the highest missed days, while June had significantly fewer. The total missed days decreased over the season, with August and September seeing much lower totals. This may be due to season-ending injuries occurring early in May, with those players no longer contributing to missed days in later months, explaining the drop in missed days in August and September.

Figure (b) shows that the frequency of injuries was highest in May, with June having a similar number of injuries. The other months had about half the number of injuries as May and June. However, since May had significantly more total missed days despite the similar injury frequency, it suggests that the injuries in May were likely more severe, warranting longer recovery times, or players were more frequently rested for longer periods during this month.

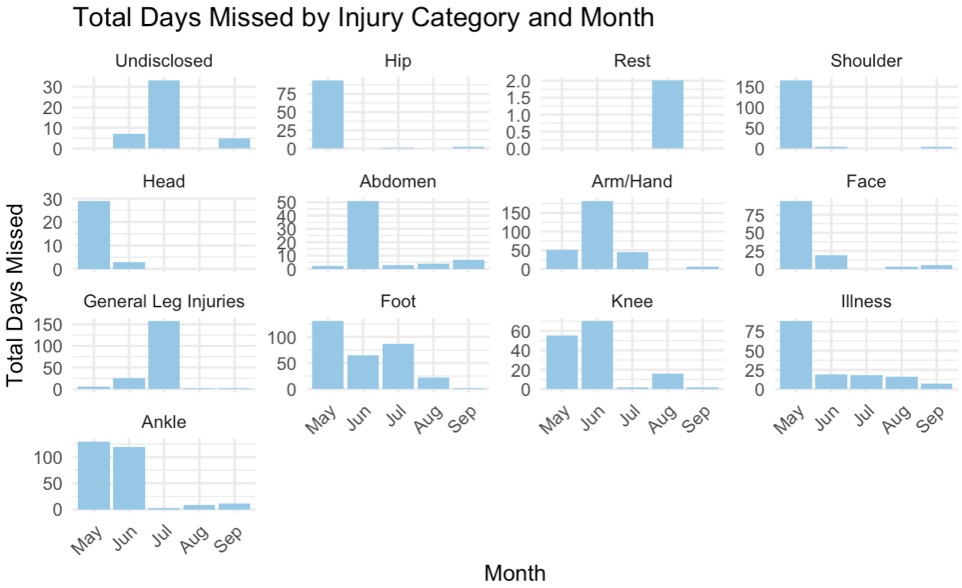


Figure 1.5. **Injury frequency across months in 2024, categorized by injury type.** This highlights the month in which each injury type was most prevalent.

The figure above displays the distribution of injuries across each month, for each injury category. This helps reveal when certain injuries were most common. The work prior to this suggests that most injuries occur in the months of May and June, which is true for many injuries, but there are a few exceptions, such as Abdomen and Arm/Hand injuries peaking in June, and general leg injuries peaking in July.

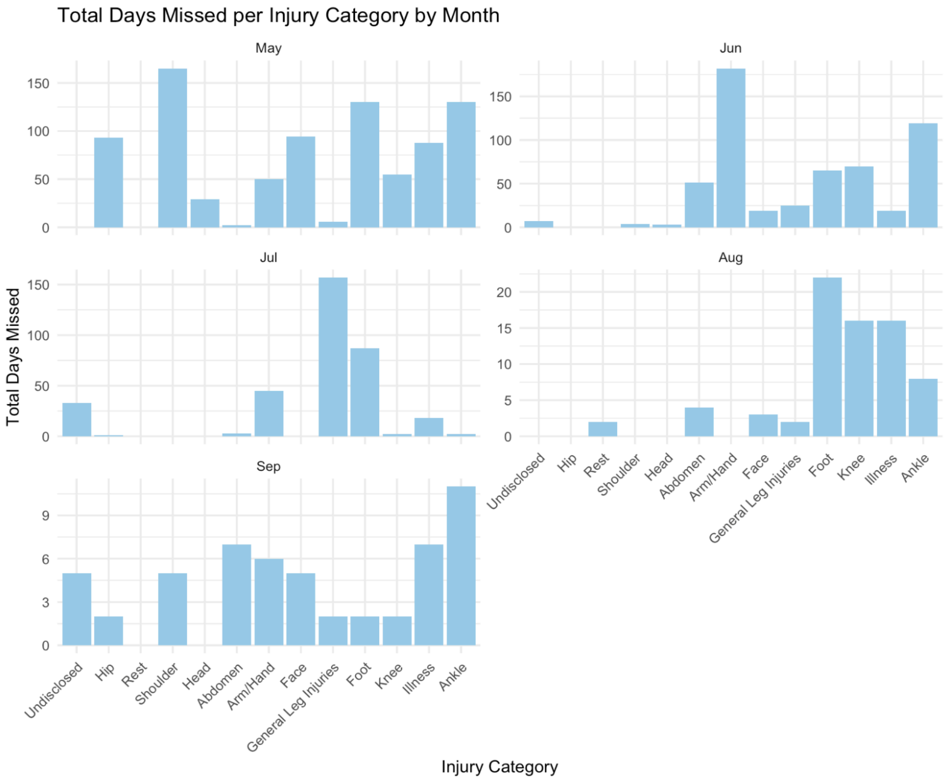


Figure 1.6. **Injury type frequency in 2024, faceted by month.** This highlights the type of injury that was most common in each month.

This next figure displays the same information, but faceted by month. The side-by-side bar chart for each month reveals that the injuries causing the highest number of days missed were shoulder injuries in May, arm/hand injuries in June, and general leg injuries in July. Foot injuries had the highest total in August, while ankle injuries peaked in September, although the magnitude of days missed in these two months was relatively small compared to the earlier months. Overall, injuries appear to be less common in August and September. This could be due to the WNBA's mid-season break or the lower intensity of play as the season nears the playoffs, possibly leading to fewer severe injuries during these months.

Overall, the trends observed in this section show that injuries aren't evenly spread across the season – May and June clearly have the most, and they tend to be more severe early on. Knowing when certain injuries are more likely to happen helps us predict risk better throughout the season. If patterns like this hold in future seasons, we would expect well-designed models to pick up on this seasonality and assign higher risk earlier in the year.

1.3.5 Seasonal Trends Among Injured vs. Non-Injured Players

After analyzing injury data by body part, category, position, and days missed, we now turn to the 2024 season stats. This dataset includes performance metrics for both injured and non-injured players, offering a more complete view of player activity. Using the `wehoop` package, we loaded WNBA player statistics for the 2024 season. Unlike the injury dataset, which has one entry per injury (and excludes uninjured players), the season stats include one record per game for every player, capturing details like minutes played.

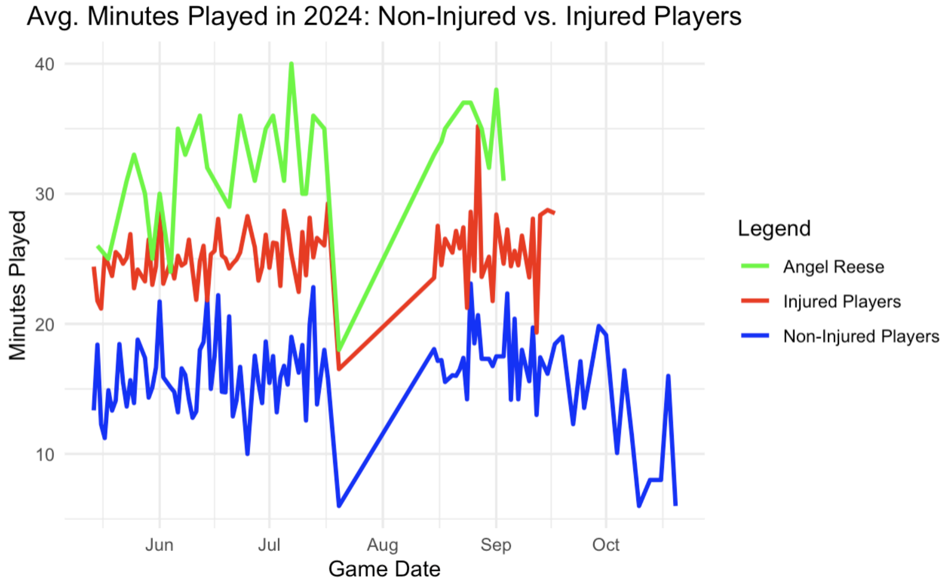


Figure 1.7. **Average minutes played per game in 2024 by injury status, with a separate trend line for Angel Reese.** Injured players averaged more minutes than non-injured players, while Angel Reese consistently played the most.

The figure above compares average minutes played by three groups: non-injured players, injured players before their injury, and Angel Reese as a high-performing baseline. Using merged 2024 injury and season stats data, we calculated minutes leading up to injuries for affected players, and season-long averages for those who remained healthy.

Injured players consistently played more minutes than non-injured ones, and Angel Reese played even more than both groups. This suggests that players logging heavier minutes – especially star players – may be at greater risk of injury. A dip in minutes appears just before August, likely due to the WNBA All-Star Game and the 2024 Olympics. Non-injured players dropped to zero during this rest period, while injured players and Angel Reese still averaged around 15 minutes. The red line (injured players) ends in mid-September, as all injuries in the dataset occurred before then, leaving no post-injury data beyond that point.

This next section explores the differences within the injured and non-injured groups, focusing on patterns like injured players with low minutes and non-injured players with high minutes. Players were split into two groups based on average minutes played: the top 40 players with the highest averages ("high minutes") and the rest ("low minutes"). Each group was then segmented by injury status to highlight how playing time relates to injury trends, as shown in the graph below.

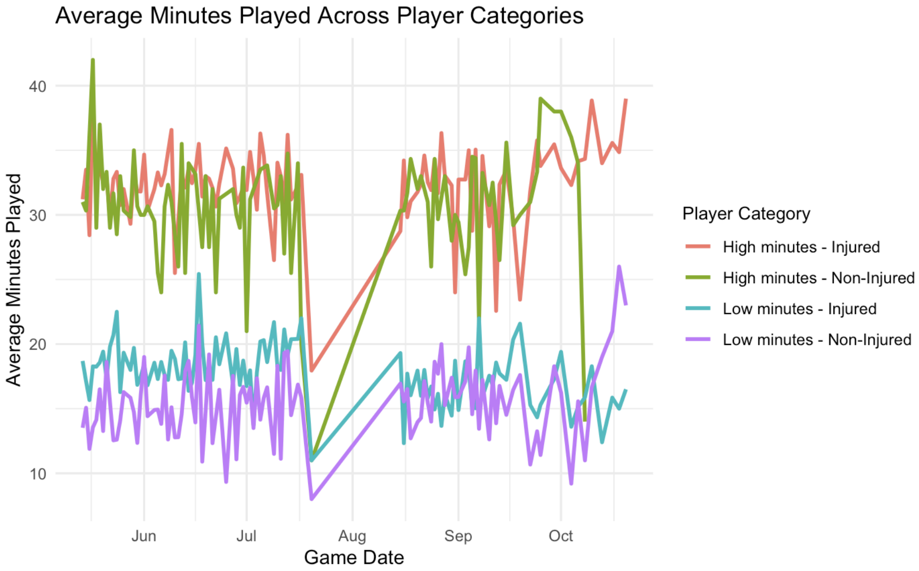


Figure 1.8. **Average minutes played in 2024 across player subgroups by injury status.** Players were split into the top 40 by average minutes played ("high minutes") and the rest ("low minutes"), then further divided by injury status.

The graph above shows that both categories of frequently playing players, regardless of injury status, maintained high average minutes throughout the season, with occasional fluctuations on specific dates. However, around the mid-July WNBA All-Star break, minutes for frequently playing, non-injured players dropped significantly to near zero, falling below those of both categories of less-frequent players. This decline is likely a result of intentional rest during the All-Star break, where key players, who had already logged high minutes and were not suffering from injuries, were given time off to recover and perhaps alleviate fatigue.

Additional patterns observed: within the players who played frequently, the average minutes played were somewhat similar across the season, as the lines were overlapping. The main exception would be around October, when the average minutes played by the non-injured frequently playing athletes dropped. This may reflect rest for star players before the playoffs or fatigue/load management.

Among the less-frequent playing athletes, the injured players consistently had higher average minutes. This is likely because the injured players had fewer games to play, but when they did play, they often played longer to make up for their absence.

1.3.6 Pre- and Post-Olympics Injury Trends

To refine our analysis and prepare the data for modeling, we exclude dates associated with unique game contexts, since these events and the surrounding injury data may introduce bias into predictive models. Specifically, we remove the period spanning the 2024 Paris Olympics, the WNBA All-Star Game, and the playoffs. The final regular-season games before the Olympic break occurred on July 17, 2024, after which the WNBA paused its season to allow players to compete in the Olympics. The WNBA All-Star Game took place on July 20, 2024, and the Olympics officially began on July 25, 2024. The WNBA season resumed on August 11, 2024, following the end of the Olympics. Additionally, the playoffs began on September 22, 2024 and ended on October 20, 2024, so those dates have been removed as well.

For analysis purposes, we segment the season into two distinct periods:

- Pre-Olympics (Start of season – July 17, 2024)
- Post-Olympics (August 11, 2024 – End of season)

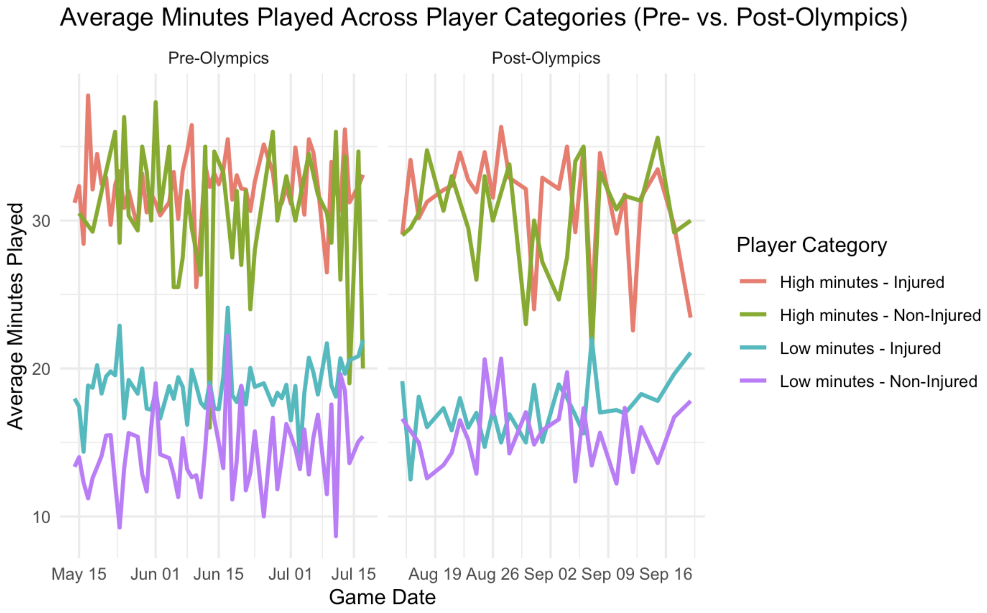


Figure 1.9. **Average minutes played across player subgroups by injury status.** This is excluding data from atypical game contexts (e.g., Olympics, All-Star, playoffs), and the season is divided into pre- and post-Olympics periods for analysis.

Examining pre-Olympic trends, non-injured high-minute players showed greater variation in playing time, with noticeable drops in mid-June and mid-July, likely due to strategic rest. In contrast, injured high-minute players maintained steadier minutes, suggesting heavier reliance until injury. Among low-minute players, injured players consistently logged more minutes, possibly due to increased usage before sustaining injuries.

Post-Olympics, variation patterns shift, with more overlap between injured and non-injured players. Among low-minute players, injured players logged similar minutes early on but saw higher averages nearing the playoffs, suggesting late-season reliance. Meanwhile, injured high-minute players saw reduced minutes late in the season, likely due to load management before the playoffs. These trends align with existing workload management strategies in sports analytics, where preserving key players is prioritized as post-season approaches.

1.4 Key Findings and Implications for Future Modeling

The EDA identified key trends that will shape future injury prediction models, particularly in selecting the most important features for predicting injury risk. Most injuries occurred early in the season, with May and June showing the highest frequency and severity of injuries, particularly in the ankle, knee, and foot categories. Player position did not significantly influence injury types, suggesting that injury risk may be more closely related to performance, workload and other individual factors. Additionally, the analysis highlighted that injured players tended to log higher minutes than non-injured players, supporting the hypothesis that overuse may be a key factor in injury risk. These findings suggest that future models should account for seasonality (month) in injury occurrence, player workload, and specific injury patterns. By incorporating these insights, we can develop more effective models that predict injury risk and recovery time, ultimately contributing to better health outcomes for players.

Chapter 2: Predicting Minutes Played Using Mixed Effects Modeling

2.1 Why Model Minutes Played?

Injury risk is influenced by several factors, with workload being one of the most significant. Although internal indicators like fatigue or soreness are challenging to monitor, minutes played serve as a practical measure of physical demand. Athletes who consistently log high minutes are likely undergoing increased physical stress, which can build up and raise the risk of injury over time. Additionally, minutes played is a controllable metric for coaches, making it valuable to have a model that predicts the optimal number of minutes a player should play in a given game to minimize injury risk.

The aim of this chapter is to build a model that estimates expected playing time in an upcoming game, given the context of the game, the player's role, and team dynamics. The model specifically aims to predict minutes for games that may be leading up to an injury, rather than long-term or future predictions for the entire season. This model will be trained exclusively on data from non-injured players to learn the typical patterns of playing time. Then, the model is applied to injured players to see if their actual playing time starts to deviate from what the model would have expected if they were otherwise healthy.

If such deviations exist, they might serve as signs that a player's usage was abnormal in a way that could be linked to increased injury risk. This approach allows us to explore whether predictive modeling can help identify when a player might be approaching dangerous levels of workload, before an injury actually occurs.

2.2 Justification for Mixed-Effects Regression

The player performance data is inherently hierarchical, as individual athletes play in multiple games across a season, are influenced by team-specific dynamics, and may differ substantially in roles and playing styles. A standard regression model would treat all observations independently and would overlook this nested structure. To address this, we turn to mixed-effects models, which are well suited for this type of data, where games are nested within players, and players are nested within teams.

Mixed-effects models would allow us to estimate *fixed effects*, which are the consistent effects across the population. For instance: impact of player position or starter status on minutes played would be considered consistent since those characteristics are generally associated with predictable differences in playing time, regardless of team or individual. Starters will generally always play more minutes than non-starters, across all teams, and this is why this relationship is treated as a fixed effect.

Random effects, however, capture group-specific deviations from these overall trends. For example, each team might have its own culture, strategies, or coaching decisions that influence playing time patterns. Similarly, players may have individual tendencies or different fitness levels within their team that can also contribute to deviations from the average effect.

Thus, in this model, starter status and position are treated as fixed effects because they represent consistent patterns across the entire population of players. Teams and players are treated as random effects to account for variations that arise from specific team dynamics or individual differences.

2.3 Data Pre-processing

To prepare the data for three-level mixed effects modeling, multiple records for the same player on the same game date were aggregated by summing their minutes. For example, some players had multiple entries for a single game, with one row recording nonzero minutes and others recording zero minutes. The data were consolidated so that each row represents a unique player-game combination. Additionally, cases in which players logged exactly 0 minutes were retained to preserve information about healthy players who did not participate in a given game.

2.4 Exploring the Distribution of Minutes Played

2.4.1 Marginal Distributions

This section examines the distribution of minutes played per game by non-injured WNBA players. The minutes data is sourced from the season stats data, which contains player box data – player performance data from every single game. Injured players were filtered out using injury status information from the WeHoop Injury Tracker dataset. Since the ultimate goal is to make a model that establishes a baseline from healthy athletes, only those who remained uninjured throughout the season are included in this analysis.

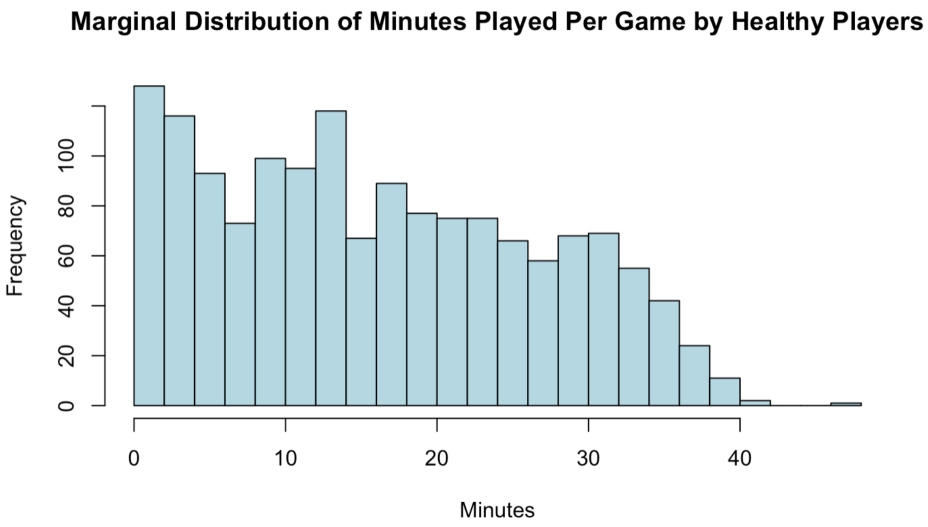


Figure 2.1: **Histogram of minutes played per game (non-injured players).** The distribution is right-skewed, bounded at zero, and most values fall below 40 minutes, which aligns with regular game limits.

The figure above shows the frequency counts for the minutes played per game by all non-injured players, with most values capping at around 40 minutes, as WNBA games typically never exceed 40 minutes (aside from overtime periods). The data is right-skewed, indicating that most players play fewer minutes, with very few logging the full 40-minute game. The distribution is right skewed and bounded at zero – traits that suggest the data may be Gamma distributed.

A Gamma density curve was overlaid on the histogram to assess if the Gamma distribution could be a good fit for modeling minutes played.

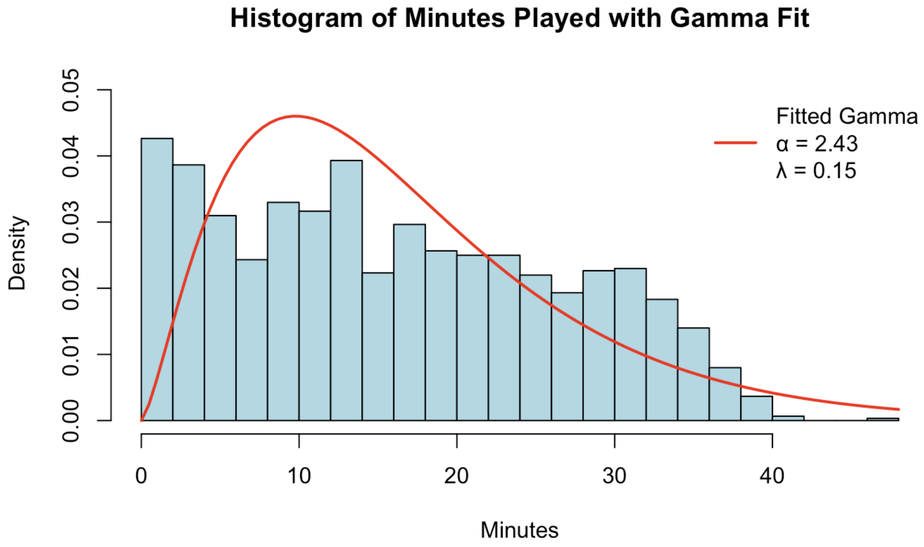


Figure 2.2: **Density plot with fitted Gamma density.** The Gamma curve roughly follows the shape of the histogram. The α (shape) parameter controls skewness, while the λ (rate) parameter controls spread.

The shape and rate parameters were estimated using the method of moments, where the parameters are automatically derived from the original data's mean and variance. This method links the first two moments (mean and variance) of the data to the corresponding moments of the Gamma distribution to estimate the parameters. Specifically, the shape parameter is calculated as the square of the mean divided by the variance, while the rate parameter is the mean divided by the variance. These estimates ensure that the fitted Gamma distribution appropriately reflects the observed data's distribution.

The Gamma density curve aligns reasonably well with the overall shape of the data – capturing the right-skewed structure and the natural lower bound at zero. However, there are some mismatches: in the lower range (0–5 minutes), the Gamma curve sharply increases while the observed data starts at a high density and steadily decreases. Additionally, between 5 to 20 minutes, the observed density consistently falls below the Gamma estimate. And between around 22 to 40 minutes, the observed density consistently lies above the Gamma estimate. These discrepancies suggest that the Gamma distribution may not perfectly capture minute level variation, particularly for cases where athletes play a low number of minutes. Despite these discrepancies, the Gamma distribution is still a reasonable choice given the data's continuous, positive, and right-skewed nature.

2.3.2 Conditional Distributions

While marginal distributions above are useful for understanding the overall shape of a variable like minutes played, they do not account for the influence of other factors. To build a meaningful model, it is necessary to look at conditional distributions – that is, how minutes played varies given specific covariates like starter status, position, or team. This is exactly what a mixed effects model attempts to estimate: the expected distribution of minutes after accounting for both fixed and random effects.

By approximating the conditional distribution given these factors, we can better explain the variation in minutes played. If, even after adjusting for player and team level random effects, the conditional distribution remains skewed – it may suggest that a non-normal model (like a gamma or log-normal) is more appropriate. Ultimately, understanding the conditional distribution helps us choose the right modeling approach for the data.

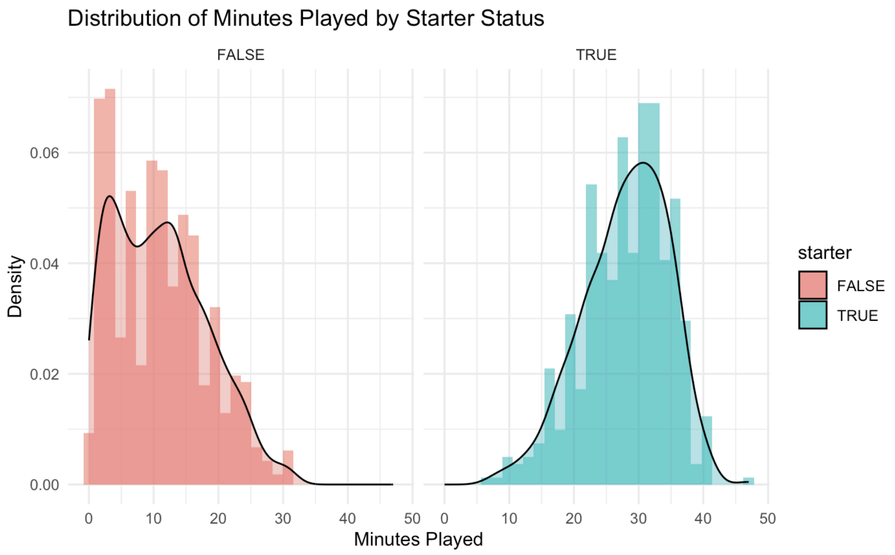


Figure 2.3: **Conditional distribution of minutes played by starter status.**

The figure above shows how the distribution of minutes played differs significantly by starter status. Starters tend to have a more symmetric distribution centered at higher minutes, while non-starters show a pronounced right skew with a concentration of low-minute games. This further indicates that a non-normal distribution – like gamma or log-normal – could perhaps capture the variation across groups.

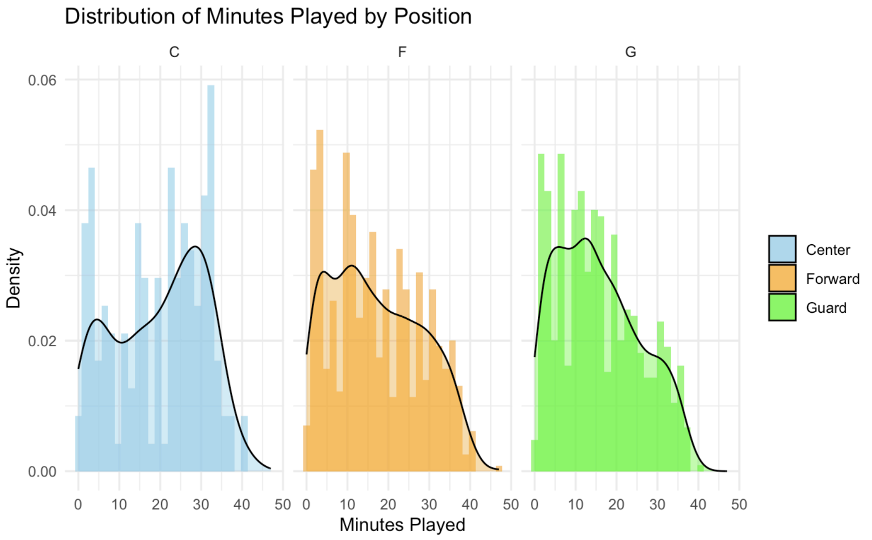


Figure 2.4: **Conditional distribution of minutes played by position.**

The distribution of minutes played also varied when conditioned by player positions. Centers showed a multi-modal and slightly right-skewed distribution, while forwards and guards had stronger right-skewness, with far more observations concentrated at lower minute ranges compared to centers. These patterns suggest non-normality across all groups.

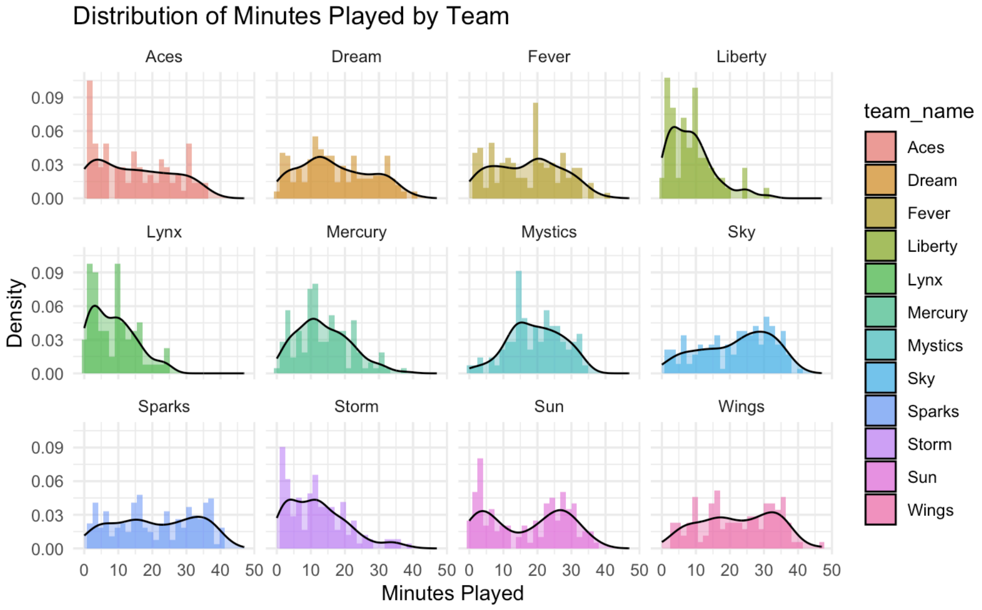


Figure 2.5: **Conditional distribution of minutes played by team.**

Many teams show right-skewed distributions (e.g., Aces, Liberty, Storm), suggesting that normality assumptions may not hold. Similar to the previous figures, this hints that a log-normal or gamma model could be more appropriate than a standard linear model.

These conditional plots help visualize how minutes played vary across key covariates, but they do not fully isolate individual player contributions. The dataset includes 61 unique players, each with multiple game observations, and player level variation appears to significantly influence minutes played. Figure 1.7 from the previous chapter highlights how individual trends, such as those of standout players like Angel Reese, can differ from broader league trends. While that figure does not adjust for starter status or position, it underscores the importance of modeling individual level variation using random effects.

After reviewing both the marginal and conditional distributions of minutes played, it became evident that a non-normal modeling approach was worth considering. The marginal distribution showed strong right skewness, pointing toward a Gamma or log-normal distribution as a potentially good fit. This skewness persisted across key subgroups in the conditional plots, further supporting the decision to explore alternatives to a standard Gaussian model.

2.5 Linear Mixed-Effects Model

A linear model was initially fitted to establish a baseline for modeling minutes played. This model was fitted using the *lme4* package, which was chosen over *nlme* due to its ability to support both linear and gamma mixed models [9]. As described in Section 2.2, the model includes starter status and player position as fixed effects, and player name and team name are modeled as random effects, ultimately to predict the number of minutes played in a given game. This is considered a three level linear model, as it accounts for variation at the team level, the player level, and the overall game level.

The linear mixed effects model can be described as:

$$y_{ijg} = \underbrace{\beta_1 \cdot (\text{starter status}) + \beta_2 \cdot (\text{position})}_{\text{Fixed effects}} + \underbrace{b_{ig} \cdot (\text{team name}) + b_{jg} \cdot (\text{player name})}_{\text{Random effects}} + \epsilon_g$$

Where:

- y_{ijg} represents the minutes played by player j on team i for a given game g.
- β_1 and β_2 are the coefficients for the fixed effects for starter status and position.
- b_{ig} and b_{jg} are the coefficients for the random effects for team i and player j, for game g.
- ϵ_g represents the residual variation at the individual game level, for given game g.

The fixed effects coefficients β_1 and β_2 capture the average impact of starter status and position across all players, estimated directly from the data. In contrast, the team-level (b_{ig}) and player-level (b_{jg}) random effects account for deviations from these averages and are assumed to follow normal distributions centered at 0. The team-level variance, σ^2_{team}, represents variability in playing time across teams, likewise for players, σ^2_{player}. These random effects reflect that some teams tend to play their players more (or less) on average, and individual players may have their own consistent playing time tendencies. Finally, the residual error term ϵ_g, also assumed to be normally distributed with mean 0, captures additional unexplained variability. The variance of the residual error term, σ^2, captures the unexplained variability in minutes played at the individual game level. In summary:

- $b_{ig} \sim N\left(0,\ \sigma^2_{team}\right)$
- $b_{jg} \sim N\left(0,\ \sigma^2_{player}\right)$
- $\epsilon_g \sim N\left(0,\ \sigma^2\right)$

2.5.1 Linear Model Results and Residual Plots

The model's random effects show substantial variation in playing time across players and teams. The player random effect has a variance of 28.88, meaning there are large differences in minutes played from one player to another, likely because of individual factors like skill or playing style. The team random effect has a variance of 12.74, showing that playing time also varies across teams, possibly due to differences in team strategies or coaching. The residual variance is 32.79, which captures other random variation not explained by the model. Overall, individual player differences have a bigger impact on playing time than team level differences.

The fixed effects analysis reveals that starter status is a strong predictor of minutes played, with starters playing, on average, 12.59 more minutes than non-starters. This effect is highly significant, as indicated by a t-value of 26.81, which is much larger than the typical threshold for significance (around 2). In contrast, athlete position (forward and guard) has minimal impact on playing time. For forwards and guards, the increases in playing time are small (1.47 and 1.84 minutes, respectively), and these effects are not statistically significant, as their t-values are 0.53 and 0.66, respectively, which are too low to indicate meaningful differences. Thus, starter status is the primary factor influencing playing time, while athlete position has little to no effect.

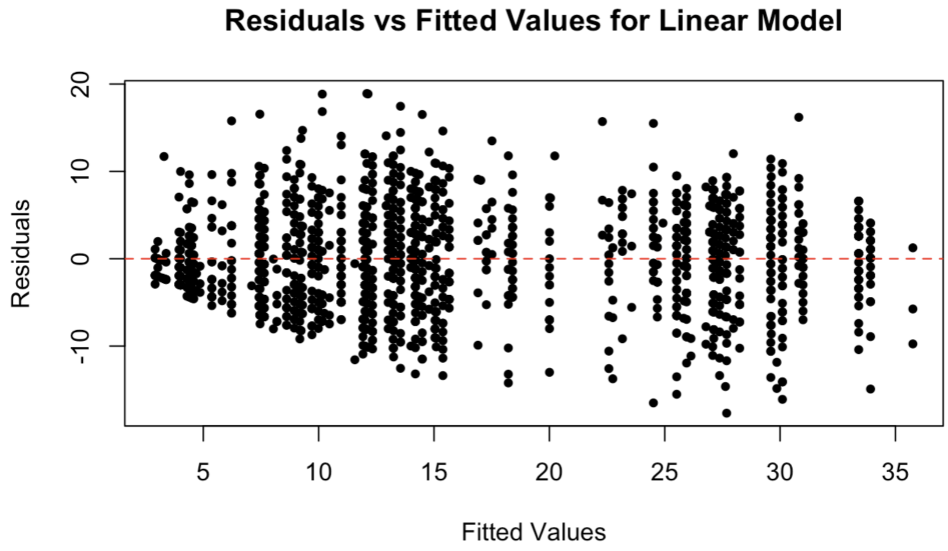


Figure 2.6: **Residuals vs fitted values plot.** The residuals are randomly scattered around zero, suggesting the model's assumptions are reasonable. However, the residual size is large.

The figure shows residuals plotted against the model's predicted values to assess fit and assumptions. Residuals are fairly evenly scattered around zero, indicating no systematic overprediction or underprediction, and supporting model assumptions. However, the residuals have a wide spread, ranging from about -20 to +20 minutes. This suggests that while the model captures overall trends, a significant amount of variation remains unexplained.

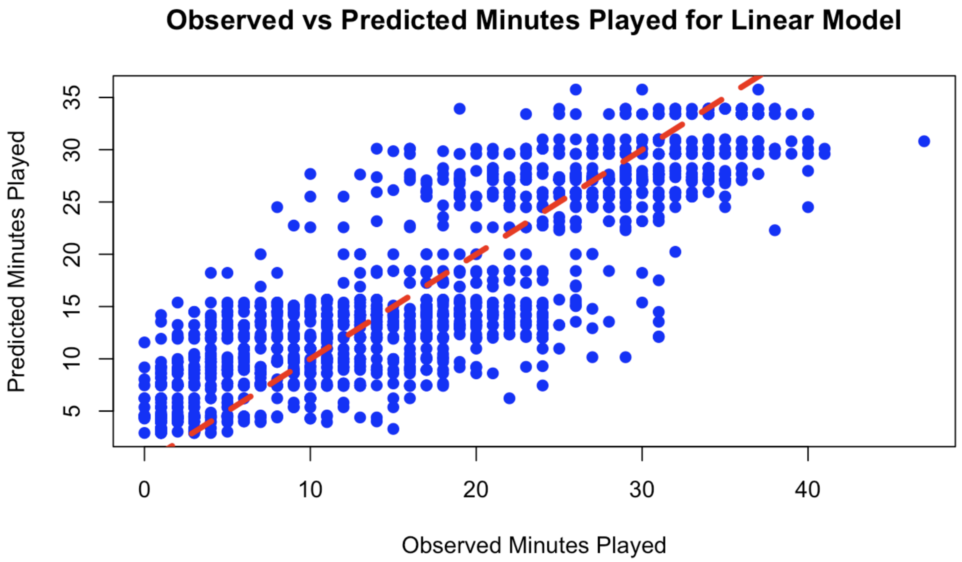


Figure 2.7: **Observed vs predicted minutes plot.** The observed values closely align with the predicted values, indicating general accuracy, but the variation suggests the model may still be influenced by other factors.

The figure above plots the model's predictions against the actual values, allowing us to assess overall accuracy. Most observed values (blue points) lie close to the red dotted line, which represents perfect prediction, indicating generally accurate model performance. However, some points are noticeably scattered, suggesting the model doesn't fully capture all factors influencing playing time.

2.5.2 Random Effects from the Linear Model

In this section, we examine the random effects extracted from the model for both players and teams. These random effects show how much individual group level deviations (from players and teams) differ from the overall average, after accounting for fixed effects.

Among players, Marina Mabrey had the highest deviation at +11.82 minutes, meaning she consistently plays significantly more than average after accounting for starter status and position. In contrast, Caitlin Bickle had the lowest deviation at -8.06 minutes, indicating she plays considerably less. These values reflect how individual players deviate from the model's expected baseline.

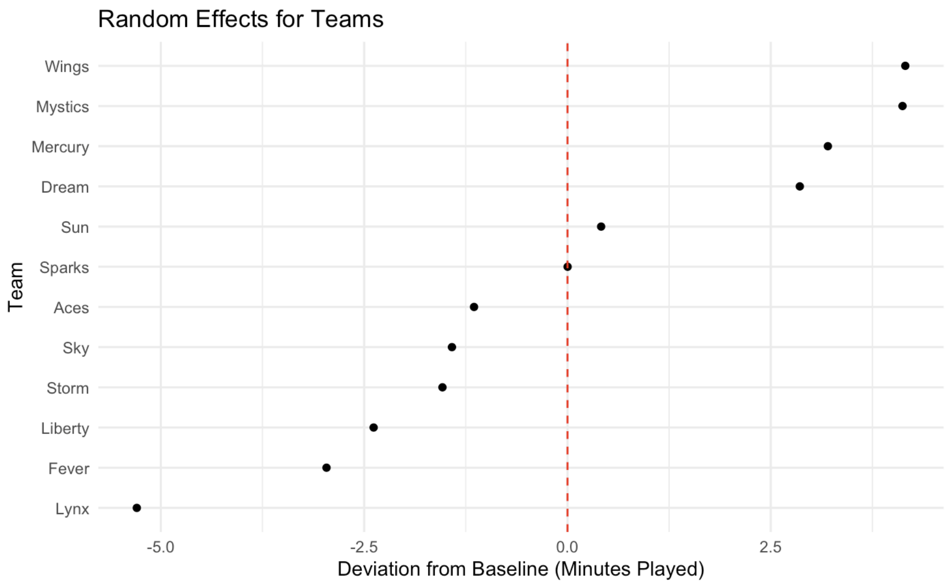


Figure 2.8: **Plot of the random effect intercepts for teams.** This shows the deviations of each team's baseline playing time from the overall average baseline playing time (after accounting for fixed effects).

For teams, the Wings had the highest deviation at 4.15 minutes, meaning their baseline playing time as a team is much higher than the average. The Lynx had the lowest at -5.29 minutes, indicating lower than average playing time. Teams like the Sparks had a deviation of 0.0018, meaning their playing time is almost exactly at the average. Compared to players, teams show a much smaller range of deviations, with players exhibiting more variability in their baseline playing time.

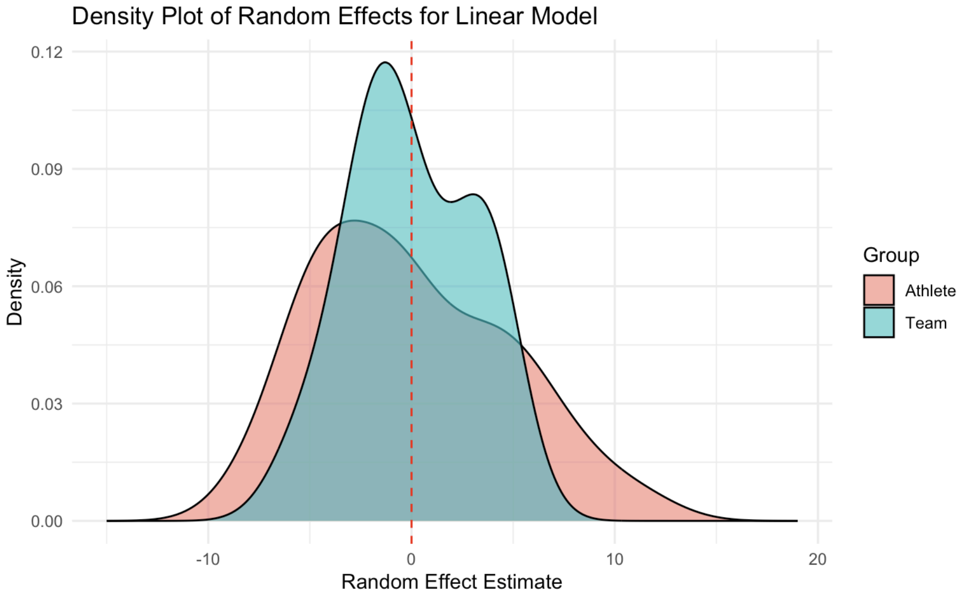


Figure 2.9: **Density plot of random effect estimates for teams and athletes, for the linear model.**

The density plot above shows the distribution of random effect estimates for athletes and teams. Athlete estimates form a unimodal, right-skewed distribution with a wide range (−8.06 to 11.82 minutes), reflecting substantial variation in baseline playing time across players. In contrast, team estimates are multimodal with most values near zero and a narrower range (−5.29 to 4.15 minutes), indicating smaller deviations from the average. These patterns align with earlier findings showing greater variability at the player level than the team level.

In conclusion, the model summary and random effect distributions both show that there is greater variability in baseline playing time at the player level, suggesting that individual player patterns are not fully captured. This points to the need for a more nuanced model to account for these variations. While the linear mixed-effects model serves as a useful baseline, it struggles with the right-skewed nature of the data, as reflected in the wide spread of residuals. To better handle this skew and improve predictions, exploring Gamma or log-normal models is recommended.

2.6 Gamma Mixed-Effects Model

In the Gamma mixed model, the structure from Section 2.4 remains the same, but the response variable (minutes) is assumed to be positive and right-skewed. A log link function is used to ensure that the predictions remain positive by modeling the log of the expected value, which is exponentiated to yield positive values. The key difference in the equation is applying the log link to the response variable, so the model predicts $\log(y_{ijg})$ instead of the raw response y_{ijg}. The model equation is:

$$\log(E[y_{ijg}]) = \underbrace{\beta_1 \cdot (\text{starter status}) + \beta_2 \cdot (\text{position})}_{\text{Fixed effects}} + \underbrace{b_{ig} \cdot (\text{team name}) + b_{jg} \cdot (\text{player name})}_{\text{Random effects}}$$

In contrast to the linear model equation, the error term ϵ_g is dropped because we model the conditional mean directly, and the Gamma distribution itself captures the variance structure of the response variable. Therefore, the residual error term is unnecessary, as the Gamma distribution already defines the variability around the predicted value.

Similar to the linear model, the random effects and the residual error are assumed to follow normal distributions centered at 0. However, for the response variable y_{ijg}, we assume it follows a Gamma distribution with mean μ_{ijg} determined by the fixed and random effects:

- $y_{ijg} \sim \text{Gamma}(\alpha, \beta)$, where α is the shape parameter and β is the rate parameter, as described in Figure 2.2 from Section 2.4, where parameters were automatically estimated using the method of moments.

2.6.1 Gamma Model Results and Residual Plots

The player level random effect had a variance of 0.131, the team random effect had a variance of 0.054, and the residual variance was 0.261 – however, as mentioned above, the residual variance is technically no longer meaningful in the gamma regression. Similar to the linear model, this indicates that individual player differences have a larger influence on playing time than team level variations. The residual variance of 0.261 in the Gamma model is significantly lower compared to the linear model's residual variance of 32.79. While this might initially suggest that the Gamma model better explains the variation in playing time, this difference likely arises because the Gamma distribution inherently accounts for the variance in the response. Therefore, the residual variance in the Gamma model is not directly comparable to that in the linear model, as the Gamma distribution already incorporates this variability.

The fixed effects analysis highlights starter status as a significant predictor of minutes played. Starters, on average, play 0.672 more minutes than non-starters, with a highly significant z-value of 13.274 ($p < 2e-16$), indicating a strong relationship. On the other hand, athlete position shows minimal impact. Forwards and guards only have small increases in playing time (0.145 and 0.270 minutes, respectively), and these effects are not statistically significant, with z-values of 0.422 and 0.774, respectively (both $p > 0.05$). Thus, similar to the results of the linear model, the starter status appears to be more significant in determining playing time, while athlete position has little to no effect.

Since the Gamma model uses a log-link function, its predictions are on the log scale. To create a Residuals vs Fitted Values plot on the original scale of minutes played, we exponentiate the predicted values to convert them back.

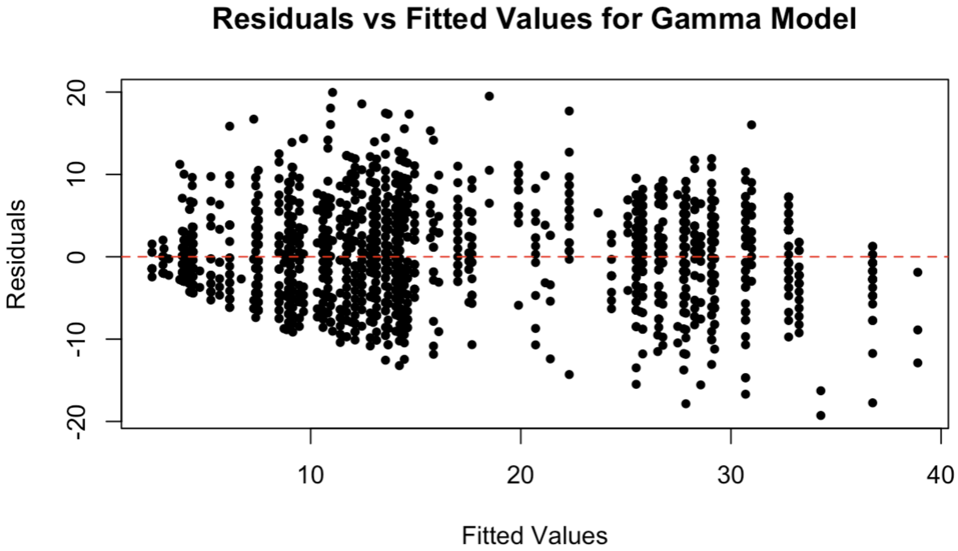


Figure 2.10: **Residuals vs fitted values plot for Gamma model.** The distribution of residuals appears very similar to that of the linear model.

Compared to the residual vs. fitted value plot for the linear model in Figure 2.6, both plots have a similar overall residual spread (roughly -20 to +20). There's similar vertical "striping patterns" throughout both plots, where each vertical stripe represents all of the players who had that specific number of minutes played. The minutes played were captured as whole numbers which is why the points appear in lines. Comparing both plots, the residuals appear more densely concentrated in similar regions of fitted values (particularly from 0-15 and from 25-35). However, the greater density of points in the gamma model compared to the linear model could be attributed to the way the predicted values were exponentiated, resulting in continuous (non-discrete) fitted values.

For the Observed vs Predicted Minutes plot, predicted values were exponentiated to return to the original minutes scale for direct comparison with the observed values, which are already on that scale.

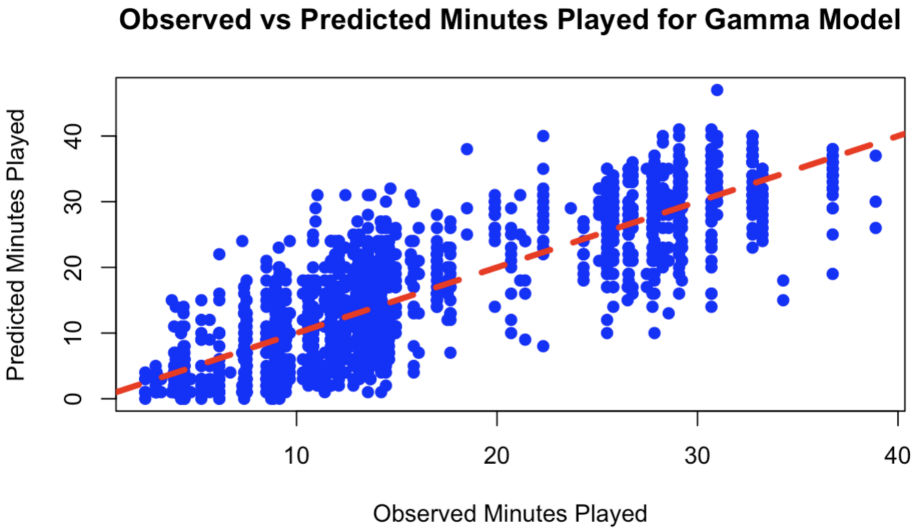


Figure 2.11: **Observed vs predicted minutes plot for Gamma model.** This plot lets us assess how well the model's predictions align with the actual values, evaluating its overall accuracy.

Comparing the observed versus predicted plots reveals distinct differences in how each model handles basketball minutes data. In the gamma model plot, points appear more concentrated, particularly in the 10-15 minute range, due to the log link function that compresses distances between lower values while expanding those between higher values. Despite containing identical data points, this transformation creates different visual distributions across the plots.

In contrast to the gamma model, the linear model (Figure 2.7) appears to overpredict players with few minutes (0-2), as linear model predictions in that region range from 0-15 minutes, while gamma model predictions in that region range more accurately from 0-5. Additionally, the linear model does not produce any predicted values above 35, while the gamma model predicts values up to nearly 50. This suggests that the two models handle the right-skewed nature of the response variable differently. The linear model appears to "average out" extreme values, pulling predictions toward the center of the distribution and struggling to capture the long right tail – resulting in a compressed prediction range. In contrast, the gamma model, which is better suited for right-skewed data, applies an exponential transformation to the linear predictor. This may allow it to more naturally produce higher predicted values, extending further into the right tail.

2.6.2 Random Effects from the Gamma Model

The distribution of the random effects from the Gamma model is quite similar to that of the linear model. The player with the lowest deviation was still Caitlin Bickle, this time with a deviation of -1.18, while Cecilia Zandalasini had the highest deviation at 0.86. This differs from the linear model in that the deviations for the players ranged from -8.06 to 11.82, but for the gamma model it was much more narrow. The random effects for teams also showed similar patterns: the Lynx had the lowest deviation and the Mystics had the highest, consistent with the results from the linear model. The main difference is that in the Gamma model, the team deviations were much smaller, ranging from approximately -0.5 to 0.5, whereas in the linear model, the range was much larger, from about -5 to 5.

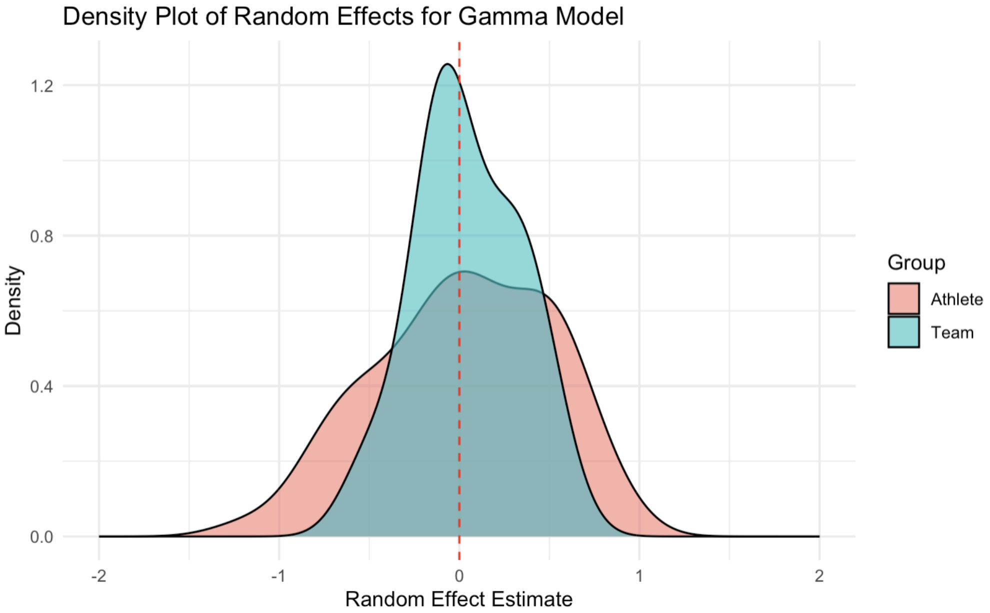


Figure 2.12: **Density plot of random effect estimates for teams and athletes for gamma model.**

Compared to the random effect density plots from the linear model, the Gamma model's curves are more smoothed and the athlete distribution is more tightly centered around 0. The most noticeable difference is the range: in the linear model, random effects ranged widely from approximately -8.06 to 11.82 minutes, while in the Gamma model, the spread is much more condensed, staying within approximately ±2.

The Gamma model provides a more compact estimation of random effects compared to the linear model, capturing the overall variability in playing time more effectively. While the linear model has wider random effect deviations and larger residuals, the Gamma model's residuals are tighter, indicating better overall accuracy. However, the Gamma model shows localized over and underprediction in certain ranges, as seen in Figures 2.10 and 2.11. In general, each model has its strengths: the linear model is more balanced in predictions but less accurate overall, while the Gamma model is more precise but has some systematic biases in specific regions.

2.7 Log-Normal Mixed-Effects Model

The Log-Normal mixed-effects model builds upon the framework for the linear model outlined in Section 2.5, with the primary difference being the assumption of a log-normal distribution for the response variable. A log transformation is applied to the response variable to better handle the right-skewed nature of the data. Specifically, the model predicts the logarithm of the expected minutes played, and the inverse transformation (exponentiation) is used to return to the original scale. The model equation is shown below:

$$\log(y_{ijg}) = \underbrace{\beta_1 \cdot (\text{starter status}) + \beta_2 \cdot (\text{position})}_{\text{Fixed effects}} + \underbrace{b_{ig} \cdot (\text{team name}) + b_{jg} \cdot (\text{player name})}_{\text{Random effects}} + \epsilon_g$$

In the log-normal model, the response variable y_{ijg} is assumed to follow a log-normal distribution – or, in other words, the model assumes that the log of the response (minutes played) follows a normal distribution. Similar to the linear mixed-effects model, the random effects for teams and players are assumed to be normally distributed with mean zero. Additionally, the residual errors after log transformation are assumed to be normally distributed, capturing unobserved variation at the game level.

2.7.1 Log-Normal Model Results and Residual Plots

The player level random effect had a variance of 0.867, the team level random effect had a variance of 0.458, and the residual variance was 0.917. As with the linear and gamma models, individual player differences had a larger influence on playing time than team level variations. The residual variance of 0.917 is substantially lower than the linear model's residual variance of 32.79, but not as low as that of the gamma model’s residual variance of 0.261. This still indicates that the log-normal model accounts for variation in playing time better than the linear model.

The fixed effects analysis identifies starter status as a significant predictor of minutes played. Starters are associated with a 0.782 increase in the log of minutes played, with a highly significant t-value of 9.930, suggesting a strong relationship ($p < 2e-16$). And similar to previous model analysis, athlete position again shows minimal impact. The fixed effects estimates for forwards and guards are small (0.092 and 0.136, respectively) and not statistically significant, with t-values of 0.193 and 0.281 (both $p > 0.05$). These findings mirror those of the linear and Gamma models, reinforcing that starter status plays a more important role in determining playing time than player position.

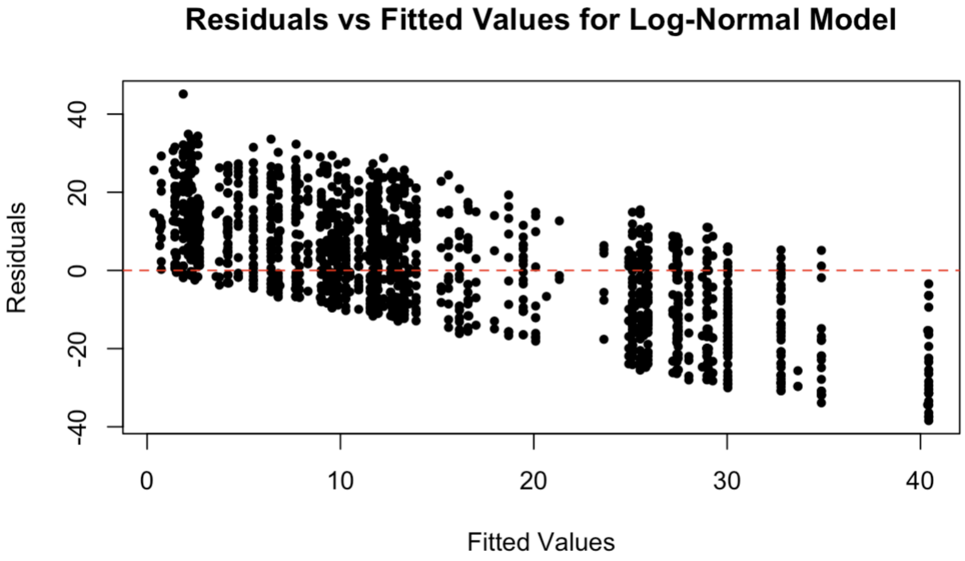


Figure 2.13: **Residuals vs fitted values plot for Log-Normal model.** The distribution of residuals differs greatly from the linear and gamma plots.

In the figure above, the residuals start mostly positive at low fitted values and become mostly negative at high fitted values. This means the log-normal model is underpredicting high values and overpredicting low values. The issue may stem from how the fitted values are exponentiated, which gives the median of the distribution on the original scale rather than the mean. To address this, we attempted to adjust the fitted values by adding the variance of the log scale, which accounts for the difference between the median and the mean in a log-normal distribution. This adjustment is intended to provide more accurate mean predictions on the original scale. However, the graph remained unchanged, likely because the variance adjustment had a minimal impact on the predictions, or the model's fit already adequately represents the underlying distribution, leaving little room for further improvement.

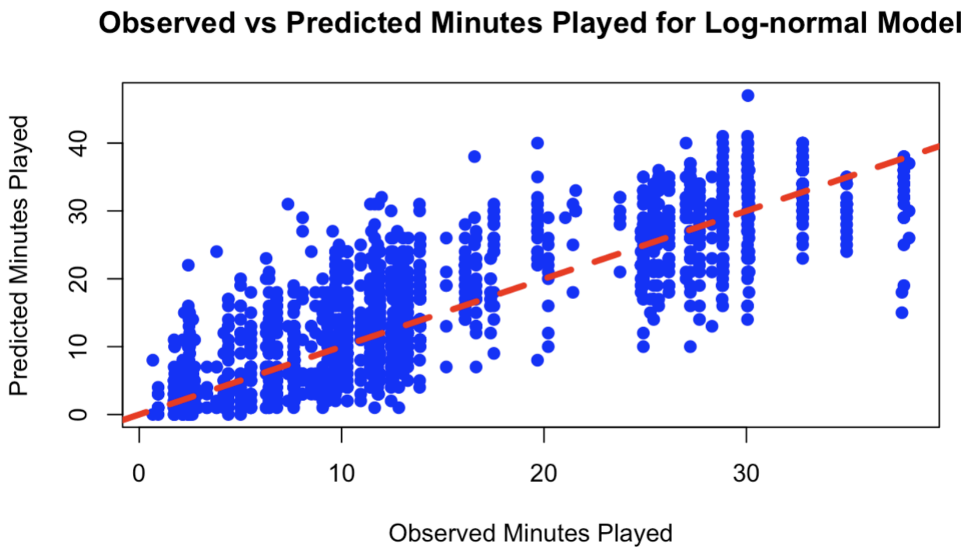


Figure 2.14: **Observed vs Predicted values plot for Log-Normal model.** This shows a strong positive relationship, with a somewhat wide range for predictions.

The observed vs. predicted values plot for the log-normal model closely resembles the one for the gamma model (Figure 2.11) rather than the linear model (Figure 2.7). This is likely because exponentiating the predicted values back onto the original scale had a similar effect as in the gamma model, while the linear model does not involve any transformed scales. Similar to the gamma model, the spread of residuals is narrower at lower minutes and increases toward higher minutes played, suggesting that the model may be underestimating variability at higher values. This pattern aligns with the right-skewed nature of the original response variable, where higher values have more variability and the model may struggle to capture this increased spread effectively.

2.7.2 Random Effects for Log-Normal Model

In the current model, the player with the highest deviation is Cecilia Zandalasini, with a deviation of 1.71, and the player with the lowest deviation is Jakia Brown-Turner, at -3.04. The range of deviations for teams is from -1.5 to 1, with the Lynx at the lower end and the Mercury at the upper end. Compared to the linear model, the player deviations are much narrower, which ranged from -8.06 to 11.82. In contrast to the Gamma model, the team deviations in the current model are slightly wider, ranging from -1.5 to 1, compared to -0.5 to 0.5 in the Gamma model.

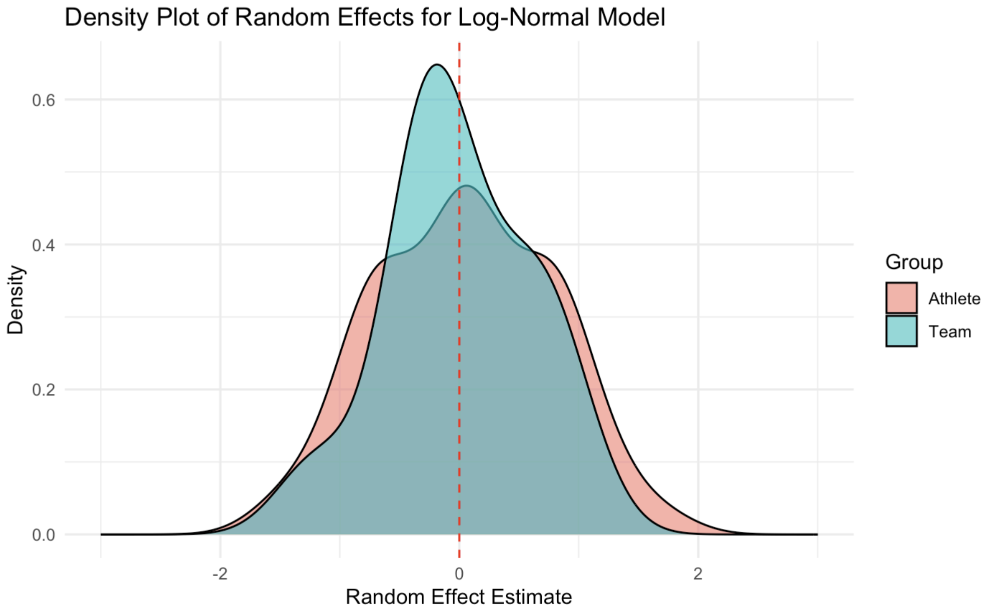


Figure 2.15: **Density plot of random effect estimates for teams and athletes for gamma model.**

Similar to the density plots from the linear and Gamma models, the team random effects in the log-normal model show a sharp peak around 0, indicating that most estimates are clustered near the average. Like in the Gamma model, both the player and team density curves are relatively smooth, with a spread of about ±2.5. This contrasts with the linear model, which had a much wider range of deviations – from approximately -8 to 11. One notable difference in the log-normal model is that the spreads for players and teams are more comparable, whereas in the linear model (Figure 2.9), the team distribution was significantly narrower than the player distribution. Additionally, the random effects distributions in the log-normal model appear more symmetrical and closely resemble a normal distribution, better aligning with the model's assumption of normally distributed random effects than in the previous models.

2.8 Model Comparison and Summary

Below is a model summary table of the results from sections 2.5, 2.6, and 2.7:

Model	Linear	Gamma	Log-Normal
Starter Status Effect	+12.59 minutes, Highly significant	+0.672 minutes, Highly significant	+2.185 minutes, Highly significant
Residuals vs Fitted Values Plot (accesses error	Residuals random but wide spread (-20 to +20), no strong	Similar spread to linear, but denser in specific regions,	Residuals mostly positive at low fitted values, negative at

distribution)	patterns	continuous fitted values	high, underprediction at high values
Observed vs Predicted Plot (accesses accuracy)	Good alignment, some scatter, general accuracy	Similar to Linear, but better captures right-skewed data with extended predictions.	Similar to Gamma, but struggles with variability at higher values.
Random Effects Spread (Player)	Wide (-8.06 to 11.82)	Narrow (-1.18 to 0.86)	Narrow (-3.04 to 1.71)
Random Effects Spread (Team)	Moderate (-5.29 to 4.15)	Narrow (-0.5 to 0.5)	Narrow (-1.5 to 1)
Random Effects Distribution Shape (Player)	Right-skewed, wide spread	Centered at 0, smoother curves	Sharp peak around 0, comparable spread to teams
Random Effects Distribution Shape (Team)	Multimodal	Centered at 0, smoother curves	More symmetrical shape than others

Table 2.16: **Summary of Model Diagnostics and Results for Linear, Gamma, and Log Normal Models.** Includes variance components, starter effect estimates, and key residual and random effect patterns.

Before evaluating model performance, we might expect the Gamma model to perform the best. It has the smallest residual variance (0.26), much lower than both the linear and log-normal models, suggesting it fits the data most tightly. The random effect variances for players and teams are also much smaller, indicating it captures individual and team differences without excessive spread. Additionally, the residual patterns for the Gamma model suggest less bias (although some overprediction at low minutes), and the random effects distributions are centered and smooth, hinting at better model assumptions.

2.8.1 Cross Validation Approach

To compare the predictive performance of the three models, 5-fold cross-validation was conducted separately for each model. However, the cross-validation was performed within the same loop, ensuring that each model was trained and tested on the same data folds. Rather than randomly splitting the data by rows (individual games), folds were created at the athlete level to ensure that all data from a given player appeared exclusively in either the training or testing set. This approach better mimics the real-world scenario, where we want to predict outcomes for players the model hasn't seen before. If the data were split by individual games, the same player could appear in both training and testing sets – making it easier for the model to predict their outcomes based on prior knowledge, and inflating its performance unfairly.

In each fold, models were trained on 80% of the athletes and tested on the remaining 20%, rotating across all athletes over five folds. In each round of cross-validation, 20% of the players are held out as the test set, and the model is trained on the remaining 80%. This process is repeated five times, each time with a different subset of players used for testing, so that every player is eventually part of the test set once. This rotation of the test group was done to average the results to get a more stable and fair measure of how well the model works.

Root mean squared error (RMSE) was used as the performance metric, quantifying how far off the predictions were from the actual minutes played.

2.8.2 Linear Model Demonstrated Superior Performance

The RMSE was calculated for each of the 5 folds, and the mean RMSE for each model was used to compare their predictive performance. The cross-validation results showed that the linear model consistently produced the lowest prediction errors, with a mean RMSE of 8.27. The gamma model followed with an RMSE of 10.26, while the log-normal model had the poorest performance, yielding an RMSE of 18.63. A paired t-test between the two best performing models (linear and gamma) was conducted, and the test revealed a statistically significant difference in RMSE values (p=0.0216), supporting the superiority of the linear model.

This outcome is somewhat surprising given the right-skewed nature of minutes played, which motivated the gamma and log-normal models. However, in this dataset, the added complexity of a log transformation or gamma specification did not translate to better generalization performance. These results suggest that, despite potential skewness, the linear mixed effects model offers a robust and interpretable baseline for modeling minutes played.

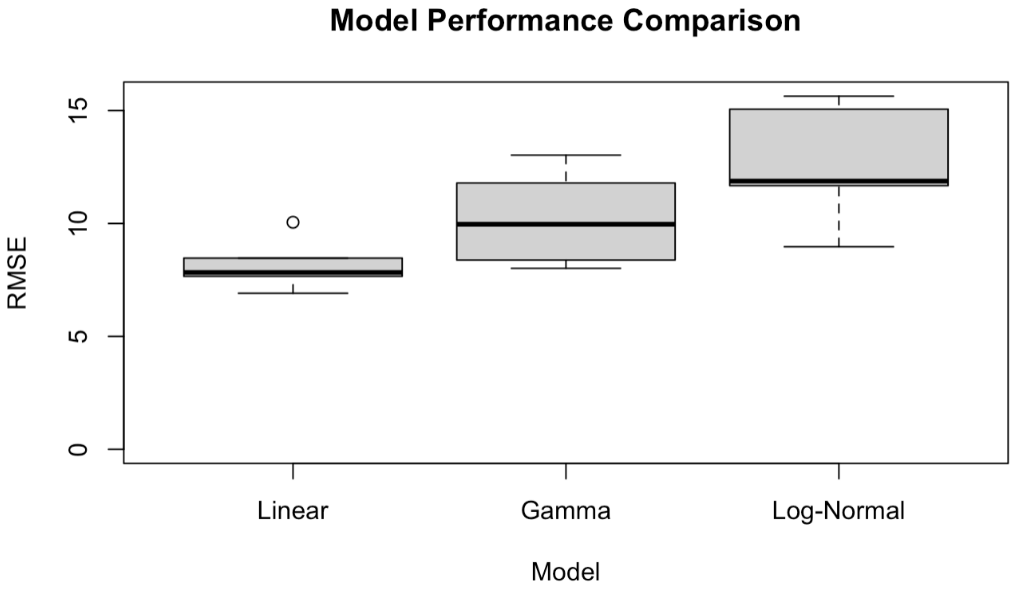


Figure 2.17: **Cross-Validation Performance Comparison of Models.** Boxplots indicate that the linear model had the lowest median RMSE and thus best predictive performance.

The boxplots reveal distinctive RMSE distributions across models. While the gamma model shows a relatively symmetric error distribution with a centered median, both linear and log-normal models display medians positioned toward their lower quartiles, indicating a skew. This suggests both models perform better than their mean RMSEs imply for most athlete subsets, with occasional higher-error folds pulling their averages upward. Despite this similar distribution pattern, the linear model maintains substantially lower and more consistent RMSE values (approximately 7-10) compared to the log-normal model's wider and higher range (approximately 9-15), demonstrating superior generalizability despite sharing a comparable error distribution shape.

The boxplots show some overlap between models, with the log-normal model's best performance (around 9 RMSE) reaching into the linear model's typical range (7-10 RMSE). This means the log-normal model occasionally performs adequately for certain athlete subsets. However, its accuracy varies dramatically across folds, with errors sometimes reaching 15 RMSE. In contrast, the linear model maintains consistent, low errors throughout all cross-validation folds. While the log-normal model may effectively capture relationships for specific player profiles, it lacks the reliable generalization ability of the linear model.

2.9 Model Predictions for Injured Players

After evaluating the three models with cross-validation, the linear mixed-effects model outperformed the others, so we proceed with it to assess players who eventually sustained injuries. It's important to note that the model was trained exclusively on healthy players, serving as a baseline to estimate expected minutes played under normal (non-injured) conditions.

To evaluate how injured players deviated from the model's expectations leading up to their injuries, we applied the trained linear model to generate minute predictions for these players. By comparing actual minutes played by injured players vs. the model's predictions, we calculated residuals – representing deviations from the expected healthy baseline.

The figure below displays the residuals from the linear mixed-effects model for all injured players, focusing on the games leading up to each injury event. For players with multiple injuries, we identified and labeled each injury cycle separately. Each game played by an injured athlete during the 2024 season was matched to all of their injuries and filtered to include only those that occurred on or before the injury date within that cycle. To prevent overlap, we retained only the most immediate upcoming injury associated with each game. If a player was injured on a non-game day, their most recent prior game was treated as the injury-adjacent event. Within each cycle, we calculated the number of games preceding the injury, counting down from a maximum of 40 games – the length of the 2024 WNBA regular season. Using the trained model, we predicted the number of minutes each injured player was expected to play, and calculated residuals as the difference between actual and predicted minutes. Positive residuals indicate overuse relative to healthy player expectations, while negative residuals suggest underuse.

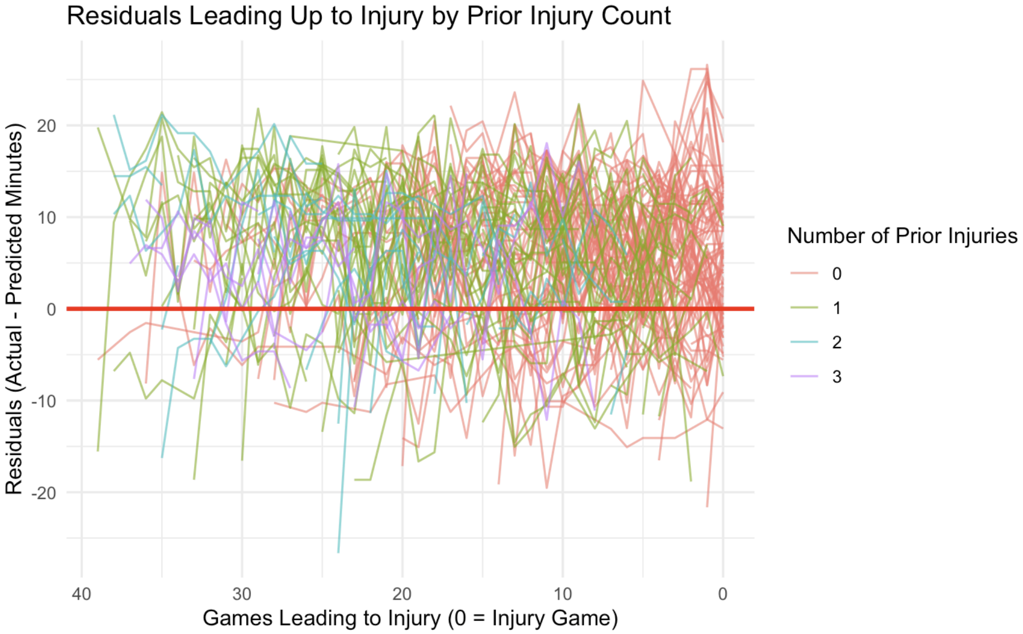


Figure 2.18: **Residuals across games leading to injury events for injured WNBA players, labeled by how many prior injuries they had.** The large presence of positive residuals near injury dates suggest some players may have been overused leading up to their injury.

In the figure above, each line represents an individual injury cycle. The noticeable skew toward positive residuals suggests that many injured players were playing more minutes than the model expected, implying possible overuse. In the final 10–15 games leading up to injury, the residuals become more densely clustered, perhaps reflecting increased game exposure and physical strain. Although there is no strong overall trend, the positive spikes and clustering near the injury event hint that elevated playing time shortly before injury may be a contributing risk factor worth further exploration.

Additionally, there appears to be a pattern among players with few prior injuries. Athletes experiencing first-time injuries have a greater variability in residuals, and many of these athletes played significantly higher than expected minutes. In the five days leading up to their injury, a notable proportion of first-time injured players had the highest positive residuals observed across the dataset, ranging from 20 to 30 minutes. This suggests that these players were playing 20-30 minutes more than expected, had they been healthy, based on the model's predictions.

The original graph is quite cluttered due to all residuals being overlaid in a single plot, so the version below facets the residuals by the athlete's number of prior injuries, for better clarity.

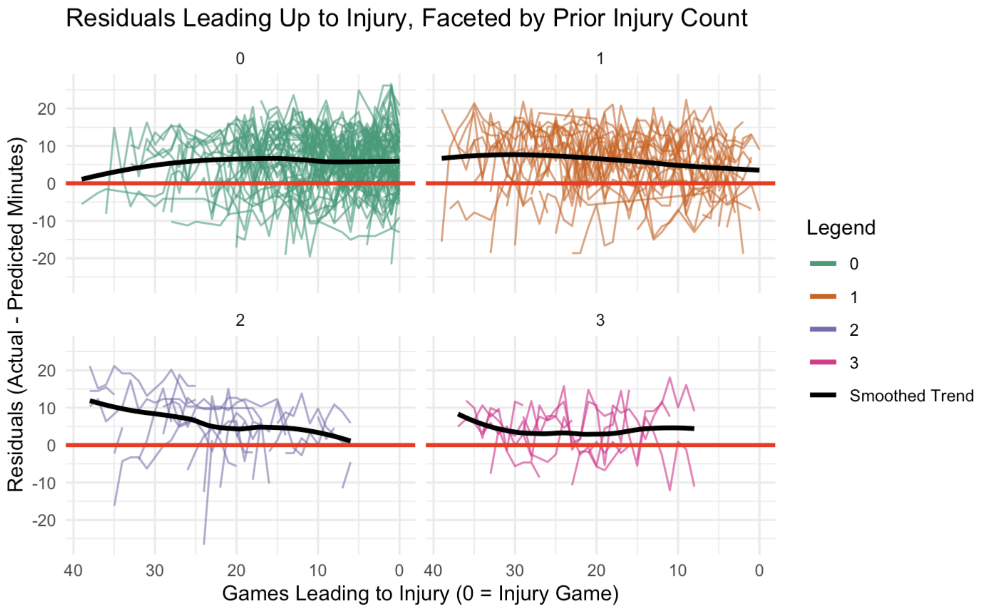


Figure 2.19: **Residuals across games leading to injury events for injured players, faceted by how many prior injuries they had.** Players appear to exceed expected minutes before injury, with this pattern heavily shifting depending on the number of prior injuries.

The figure above builds on Figure 2.18 by overlaying a smoothed regression line in each panel. This trend line is generated using a locally weighted regression (LOESS) fitted to the residuals within each facet, grouped by prior injury count. Each black regression line in each panel highlights the overall pattern of residuals in the games leading up to injury.

Across all groups, residuals are predominantly positive, suggesting that players often played more minutes than expected in the lead-up to an injury, which may reflect patterns of overuse. Some notable differences appear across prior injury counts. Players with no history of injury show a clear trend of increasing positive residuals as the injury date approaches – a pattern not seen in the other groups. This is particularly evident in their panel, where the regression line is the only one that distinctly slopes upward. For players with one prior injury, the residuals appear more stable or even slightly reduced before the day of injury, perhaps indicating more controlled playing time, given that the player and coaches know they've had an injury prior in the season.

2.10 Limitations and Future Work

In figure 2.19, there is a pattern among players with two or three prior injuries, where there is a distinct gap in the residuals during the 5–10 games leading up to the injury. This is unusual since the original dataset retains an entry for each game, even if a player did not participate or played zero minutes, ensuring every player is represented. Thus, the absence of residuals around 5-10 games before the injury suggests that the model might not have made predictions for these games, or it could indicate missing data for those specific games.

Given that this unusual pattern appears only for players with multiple injuries, it's likely related to how injury cycles were categorized, specifically how multiple injuries close in proximity are handled in the code. The code is designed to retain only the closest future injury for each game, which works well in the vast majority of cases. However, when two injuries occur within a short time span, games that should logically lead up to the second injury may instead get grouped under the first. As a result, the second injury cycle appears to have missing data in the lead-up period, creating visible gaps in figures like 2.19. Future work could address this by adding a time-based buffer that skips games occurring shortly after a prior injury, though this may complicate cycle assignments when injuries are closely spaced.

Despite its limitations, the linear mixed-effects model can be used to flag players at risk of injury by comparing actual vs. predicted minutes. Large, sustained positive residuals (like consistently playing 10+ minutes more than expected) may indicate overuse and unhealthy playing patterns. If this pattern continues over several games, it could signal elevated injury risk. Teams could monitor residuals during the season and use a threshold (e.g., +8 minutes for 3+ games) to prompt a change in coaching.

Chapter 3: Predicting Injury Likelihood and Recovery Times

While the previous chapter focused on how minutes played might be linked to injury risk, the main takeaway was that many injuries could be due to overplaying. In this chapter, we move from identifying patterns to actually predicting injury risk, where minutes played is used as one of the key predictors. We do this in two ways:

1. By predicting the likelihood of a player getting injured in the time frame between the current or most recent game and their next game.
2. By predicting recovery time (days missed due to the injury).

The first task frames injury occurrence as a binary classification problem. We use logistic regression to estimate the likelihood of a player getting injured between the most recent game and the future one, using features like prior injury history and average minutes played per game thus far. The goal is to ultimately produce risk scores for players to evaluate their likelihood of getting an injury by their next game.

The second task focuses on injury severity, measured by the number of days missed. Here, we explore and evaluate the performance of multiple modeling approaches: linear regression, zero-inflated binomial regression, standard negative binomial regression, multinomial logistic regression, and random forest.

3.1 Predicting Likelihood of Athlete Injury by the Next Game Day

Before diving into the modeling, it's important to clarify what this research defines as an injury 'by the next game day.' The injury dataset only includes the date of the injury, not the exact time. So, if an injury occurs on a game day, there is no information on whether it happened before, during, or after the game. For this reason, the goal in this section is to predict the likelihood that an athlete will experience an injury between the date of their current game and the date of their next game.

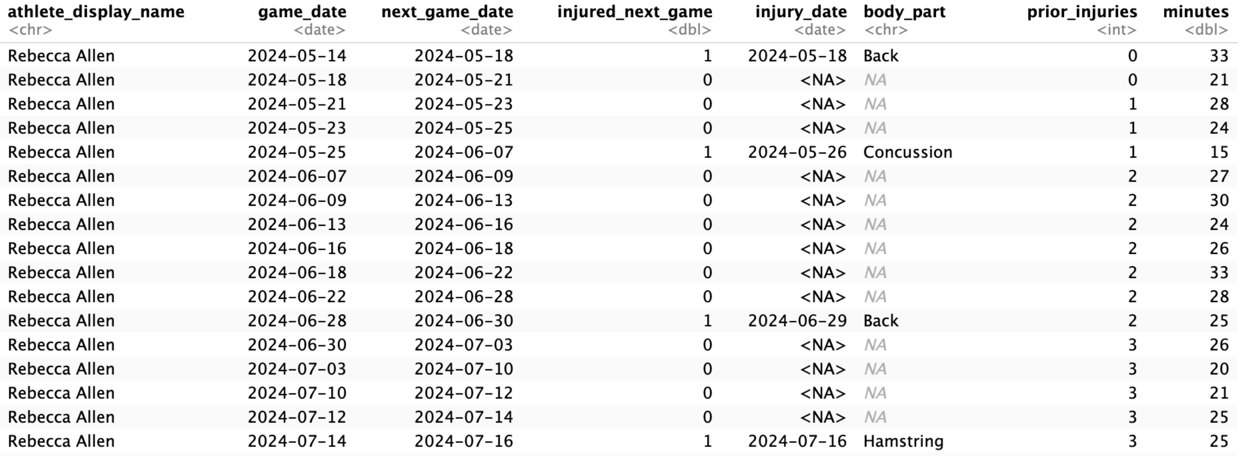
3.1.1 Data Preparation

To address this problem, we aim to build a binary classification model that predicts the probability of an injury occurring between the current game and the next, with the output ranging from 0 to 1, reflecting the model's confidence in the likelihood of an injury during that time frame.

The model is trained on both injured and uninjured players. Since this is a binary classification model, it requires data from both classes (injured and uninjured) to identify patterns for each outcome. If the model were trained only on injured players, for instance, it would be unable to predict outcomes for uninjured players.

The dataset structure mirrors earlier chapters, with each row representing a game played by a specific athlete. To create the binary target variable `injured_next_game`, we identified each athlete's next game and checked whether an injury occurred between the current and next game using the separate injury dataset, where each row represents an injury that occured along with details such as the date it occurred on. A few new features were also engineered:

- `body_part`: lists injury type if an injury occurred after current game date, NA otherwise
- `prior_injuries`: counts number of injuries an athlete had before each game
- `minutes`: playing time in the current game
- `injury_status`: flags whether the athlete was actively injured during the current game



athlete_display_name	game_date	next_game_date	injured_next_game	injury_date	body_part	prior_injuries	minutes
<chr>	<date>	<date>	<dbl>	<date>	<chr>	<int>	<dbl>
Rebecca Allen	2024-05-14	2024-05-18	1	2024-05-18	Back	0	33
Rebecca Allen	2024-05-18	2024-05-21	0	<NA>	NA	0	21
Rebecca Allen	2024-05-21	2024-05-23	0	<NA>	NA	1	28
Rebecca Allen	2024-05-23	2024-05-25	0	<NA>	NA	1	24
Rebecca Allen	2024-05-25	2024-06-07	1	2024-05-26	Concussion	1	15
Rebecca Allen	2024-06-07	2024-06-09	0	<NA>	NA	2	27
Rebecca Allen	2024-06-09	2024-06-13	0	<NA>	NA	2	30
Rebecca Allen	2024-06-13	2024-06-16	0	<NA>	NA	2	24
Rebecca Allen	2024-06-16	2024-06-18	0	<NA>	NA	2	26
Rebecca Allen	2024-06-18	2024-06-22	0	<NA>	NA	2	33
Rebecca Allen	2024-06-22	2024-06-28	0	<NA>	NA	2	28
Rebecca Allen	2024-06-28	2024-06-30	1	2024-06-29	Back	2	25
Rebecca Allen	2024-06-30	2024-07-03	0	<NA>	NA	3	26
Rebecca Allen	2024-07-03	2024-07-10	0	<NA>	NA	3	20
Rebecca Allen	2024-07-10	2024-07-12	0	<NA>	NA	3	21
Rebecca Allen	2024-07-12	2024-07-14	0	<NA>	NA	3	25
Rebecca Allen	2024-07-14	2024-07-16	1	2024-07-16	Hamstring	3	25

Figure 3.1: **Injury timeline for Rebecca Allen.** This data structure helps capture the temporal patterns between playing time and injury history.

The table above presents the engineered features for Rebecca Allen's games during the 2024 season. For example, her first few games occurred on 5/14, 5/18, and 5/21. Because she sustained an injury on 5/18, the `injured_next_game` flag is set to 1 for the 5/14 game – as her next game was on 5/18, and the injury occurred between those dates. While it's unclear whether the injury on 5/18 happened before, during, or after that game, it is still inclusively counted within the 5/14 - 5/18 cycle. Therefore, the `prior_injuries` counter is incremented by 1 after the injury on 5/18 to reflect that the athlete has had a prior injury before the 5/21 game. This logic is applied consistently to all games of all athletes in the dataset.

3.1.2 Model Description

To model whether or not a player will be injured before their next game, we use binary logistic regression. The key predictors we will focus on and use in this model are: `minutes`, `prior_injuries`, and `injury_status`. As previously mentioned, the response variable is `injured_next_game`, an indicator variable that measures whether a player was injured between current game t and their next game t + 1 (1=player was injured, 0=player was not injured).

In the dataset, there are 4389 records of games resulting in a non-injury and only 124 records of games leading to an injury. Because around 97% of the outcomes are non-injuries, a naive model could just predict no injury every time and still achieve 97% accuracy, which is misleading and ineffective. In fact, an initial confusion matrix confirmed this issue, showing that the model failed to correctly identify any injuries.

To address this, the dataset was split into training and test sets for unbiased evaluation, and upsampling was used (from the caret package) to balance out the classes. This technique oversamples the minority class (injuries) so the model couldn't be biased toward the majority class (non-injuries). Predictions were made on the test set using a probability threshold, and a confusion matrix was used to evaluate performance. This process allowed the model to better learn the underlying patterns that contribute to injury risk, making it more useful despite the initial class imbalance.

3.1.2 Model Results

The table below summarizes the key results of the binary logistic regression model:

Predictor	Estimate	Odds ratio	P-value
`minutes`	0.034	1.035	< 2e-16
`prior_injuries`	0.256	1.292	< 2e-16
`injury_status`	-0.767	0.464	0.00026

Figure 3.2: **Binary classification model summary.** All predictors were significant. Minutes played being a significant predictor of injury likelihood aligns with the main takeaways from the previous chapter.

The logistic regression results show that minutes played and prior injuries are both strong positive predictors of injury risk before the next game. Each additional minute increases the odds of injury by 3.5% (odds ratio = 1.035), while each past injury increases it by 29.2% (odds ratio = 1.292). In contrast, if a player was already injured during the current game, their odds of injury in the next game are 53.6% lower (odds ratio = 0.464), likely because they were already playing fewer minutes or recovering from their current injury.

3.1.3 Model Performance

The confusion matrix from this model's predictions on the test set is shown below.

	Reference		
Prediction		0 (actual non-injury)	1 (actual injury)
	0 (model predicted non-injury)	499	8
	1 (model predicted injury)	376	19

Table 3.2: **Regression model confusion matrix.** Compared to the naive model, there are more cases of the model correctly predicting an injury when it occurs.

The model achieved an accuracy of 57.43% – we can see in Table 3.2 this is likely because of the substantial number of non-injury cases that were misclassified as injuries. More notably, the model's **sensitivity** is 70.37%, indicating it correctly identifies actual injury events over 70% of the time – a promising result given the importance of catching potential injuries. However, the **specificity** is lower at 57.03%, meaning the model correctly identifies non-injury cases just over half the time, while falsely flagging around 43% of them as potential injuries.

While these results aren't poor, especially in a context where sensitivity matters more (since it's ideally better to overpredict injuries), the relatively low specificity is still a concern. It suggests the model struggles to clearly distinguish between injury and non-injury cases, possibly due to the class imbalance. Improving specificity – without sacrificing sensitivity – will be key to reducing false alarms and enhancing the model's practical utility.

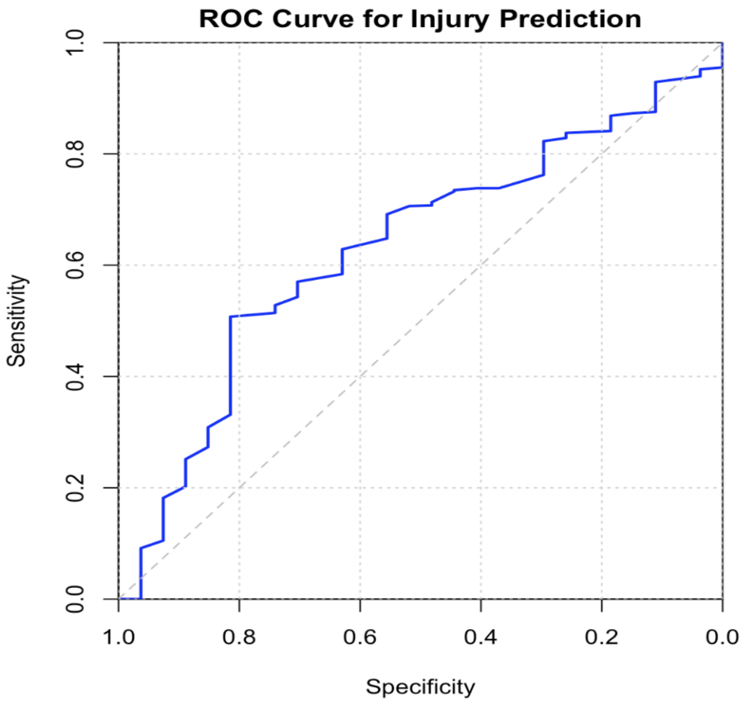


Figure 3.3: **ROC Curve for Injury Prediction Model.** With an AUC of 0.631, the model shows moderate predictive power, outperforming random guessing but leaving room for improvement.

The Receiver Operating Characteristic (ROC) curve above shows how well a model can separate injury and non-injury cases by plotting the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate) across different thresholds. The diagonal line in the figure above represents a model that makes random guesses, held as comparison. Since the ROC curve (in blue) lies above the diagonal line, it indicates that the model performs better than random guessing. However, the curve is not close to the top-left corner of the plot, which is more ideal as it would indicate high sensitivity and specificity.

The Area Under the Curve (AUC) is 0.631, meaning that 63.1% of the time, the model correctly assigns a higher injury probability to a player who actually gets injured compared to one who does not. While this suggests the model captures some predictive signal, there is significant room for improvement – likely restricted by the limited injury data.

3.1.4 Model Optimization

To optimize the model's performance, we can adjust the classification threshold to find the point that maximizes both sensitivity and specificity. In the previous work, a threshold of 0.5 was used, meaning probabilities greater than 0.5 were classified as "1" (injury), and all other probabilities as "0" (no injury). One approach is to maximize the sum of sensitivity and specificity, known as Youden's J statistic, to strike a balanced trade-off between the two [10]. Using the pROC package, we can identify the coordinates on the ROC curve in Figure 3.3 that maximize both sensitivity and specificity, ultimately determining the optimal threshold.

Our analysis revealed that this optimal threshold is 0.4761, slightly lower than the initial 0.5. At this threshold, the model achieves a specificity of 0.8148 (correctly identifying 81.48% of non-injury cases) and a sensitivity of 0.5074 (correctly identifying 50.74% of injury cases). This new threshold optimizes the balance between sensitivity and specificity, which can enhance model performance.

3.1.5 Future Work

To predict injuries with higher accuracy, the model's threshold can be adjusted for classification. Lowering the threshold (as we did earlier, going from 0.5 to 0.47) will increase the model's sensitivity (as it did, going from 70% to 81%), meaning it will correctly identify more injuries. However, this comes at the cost of reduced specificity, leading to a higher number of false positives. For instance, lowering the threshold to 0.25, for instance, will predict more injuries, but it will also misclassify more non-injury games as injuries. In this case, the model might just classify more or even all games as injuries to meet the lower threshold. This trade-off helps reduce the risk of missing actual injuries but increases the likelihood of false alarms.

The above work maximizes the sum of sensitivity and specificity, finding the optimal threshold where both metrics are maximized. But if reducing false alarms is a priority, the threshold can be raised. It can ultimately be adjusted based on the priorities of the coaching staff – whether they prefer to take a more cautious approach and risk false alarms, or focus on reducing unnecessary alarms and risk missing injuries.

Future work could explore more advanced techniques, especially to address the imbalance in the data and improve both the precision of injury predictions.

3.2 Predicting Injury Severity by Modeling Recovery Time

3.2.1 Introduction

In this section, we develop models to predict the severity of injuries, using total days missed as a proxy for severity. The goal is to understand which player or injury-related characteristics – such as game participation, scoring performance, or injury type – are associated with longer recovery periods. We explore several modeling techniques and evaluate their performance to identify the most effective approach for predicting time lost to injury.

3.2.2 Feature Engineering and Dataset Construction

Unlike previous models where each row in the dataset represents a game, for this model the dataset was restructured so that each row corresponds to a unique injury event. New aggregate features were engineered to capture a player's condition leading up to the injury, these include:

- `Total_Days_Missed` (response variable)
- `Total_Games_Missed`
- `Total_Minutes_Played`
- `Avg_Minutes_Played`
- `Total_Points`
- `Prior_Injury_Count`

The distribution of `Total_Days_Missed`, the response variable for the following models in this section is below.

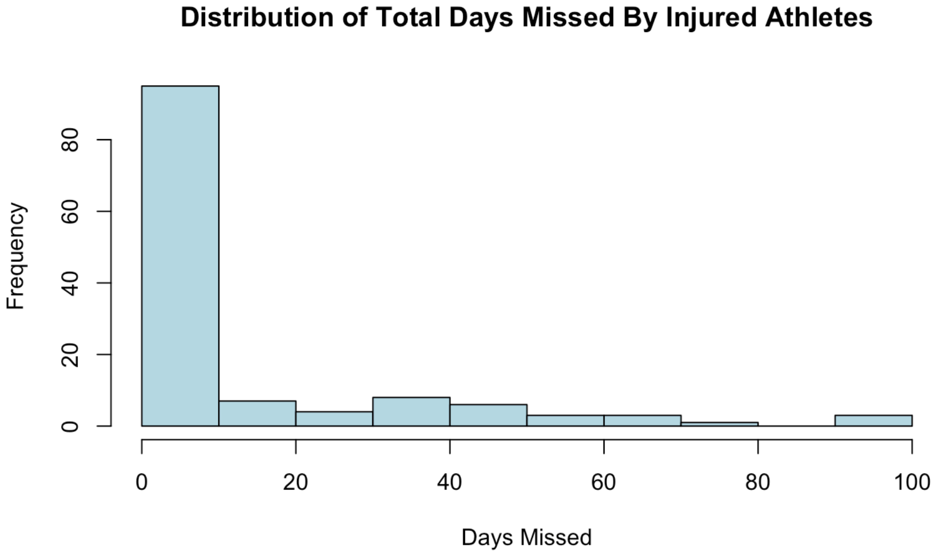


Figure 3.4: **Histogram of total recovery days for injured athletes.** Recovery periods range from 0 to 100, with the distribution being largely skewed towards shorter recovery days.

The figure above shows that the majority of athletes had recovery periods that were roughly 0-10 days long. To further estimate how zero-inflated this was, we found that 9.7% of injured athletes were listed as having 0 days missed (it's likely they had an injury the day of a game, but returned to the game later). Additionally, to be more specific, 57.46% of the injured athletes had a recovery period within 0-5 days, further explaining why the graph appears so right-skewed.

To further investigate the relationship between the new aggregate features, existing variables in the injury dataset, and Total_Days_Missed, their correlations were calculated as shown below:

Feature Name	Correlation with

	Total_Days_Missed
Total_Minutes_Played	-0.249
Total_Games_Played	-0.260
Avg_Minutes_Played	-0.287
Total_Points	-0.219
Prior_Injury_Count	-0.122
Estimated_WS_lost	0.682
Position	-0.145
BMI	-0.163

Table 3.5: **Correlation table for Total_Days_Missed.** Most features aside from estimated win-share loss have very weak correlations with recovery days. (Higher WS loss means that a player getting injured had greater loss for the team.)

As shown in the table above, most features are negatively correlated with recovery days — indicating that as these values increase, the number of days missed tends to decrease. Players who log more minutes, score more points, and have higher win share losses tend to have shorter recovery times, likely because their value to the team encourages quicker returns to play despite their injuries.

Overall, most features were only weakly correlated with recovery duration. However, creating the new features transformed the dataset into a richer representation of injury events, setting the foundation for the predictive modeling in the next section.

3.2.3 Initial Exploration and Linear Regression Models

To explore what factors best predict recovery time after an injury, we tested a series of simple linear regression models focused on different domains. Model 1 tests whether features related to a player's value to their team – such as minutes played, points scored, and starter status – help explain time lost due to injury. Model 2 focuses on characteristics tied more closely to injury severity, including injury type, prior injury history, BMI, and player position. Finally, Model 3 combines both sets of features to see if a holistic model incorporating both team value and injury severity offers a stronger explanation of total days missed. These models help us understand the relative importance of player role versus injury nature in predicting recovery duration. In summary:

- Model 1 (Team value): Total Days Missed ~ Average Minutes Played + Total Points + Starter
- Model 2 (Injury severity and athlete characteristics): Total Days Missed ~ Injury Category + Prior Injury Count + BMI + Position
- Model 3 (Combined): Total Days Missed ~ All predictors above

In Model 1, the only significant predictor of recovery time was average minutes played (coef = -0.67, p = 0.019), suggesting that players who typically play more minutes per game tend to miss fewer days when injured. This could indicate that players who are more valuable to their teams are prioritized for quicker recovery.

In Model 2, Hip injuries proved to be a significant predictor, with a positive relationship to recovery time (coef = 25.16, p = 0.046). This aligns with our earlier exploratory data analysis (Figure 1.3), reinforcing the idea that hip injuries tend to be more severe and require longer recovery periods. BMI also had a surprising negative relationship with total days missed (coef = -2.81, p = 0.015). This suggests that athletes, particularly those with higher BMI (potentially due to greater muscle mass), may have a quicker recovery, possibly due to increased physical resilience.

Finally, in Model 3, which combined both team value and injury severity predictors, only average minutes played (coef = -0.749, p = 0.013) and BMI (coef = -2.85, p = 0.012) remained significant. Their significance in the combined model was similar to their individual effects in the previous models, indicating that these factors play a consistent role in predicting recovery time across different contexts.

The AIC values for models 1, 2, and 3 are 1190.745, 1166.854, and 1162.515, respectively. Model 3, which incorporates both team value and injury severity predictors, achieved the lowest AIC, suggesting it provides the best fit to the data. This indicates that combining these two types of predictors yields a more accurate model for predicting total days missed due to injury. Based on this analysis, all models moving forward will utilize predictors from model 3.

Additionally, model 2, which focused solely on injury severity predictors, outperformed model 1, which included only team value predictors. This result implies that injury-related characteristics are stronger predictors of recovery time than team-related variables like minutes played and points scored.

3.2.4 Advanced Statistical Models

Recalling Figure 3.4 from the EDA in the previous section, we saw that the distribution of total days missed was right-skewed and somewhat zero-inflated. This suggests that a simple linear regression model may not adequately capture the complexities of the data, as it could fail to account for the non-normal distribution. As a result, we move on to more advance statistical models:

- Zero-inflated negative binomial model: to account for the excess zeros in the data
- Negative binomial model: to handle the overdispersion, since the variance of the distribution (436.77) is significantly greater than the mean (12.79)
- Multinomial logistic regression: to model recovery periods in categories (e.g., “short” vs “long” recoveries)

Zero-Inflated Negative Binomial (ZINB)

The results of the zero-inflated negative binomial (ZINB) model show a significant relationship between several predictors and total days missed. The negative binomial count model reveals that average minutes played (coef = -0.0496, p = 0.015) and BMI (coef = -0.2849, p = 0.0007) are significant predictors, suggesting that athletes who play more minutes and those with higher BMI tend to miss fewer days due to injury. Notably, Total_Points, Starter, and most injury categories did not significantly predict recovery time. The zero-inflation model had an intercept with an exceedingly large standard error, indicating no meaningful relationship between the zero-inflated part of the model and the predictors. The AIC for the model is 889.0634, and the log-likelihood is -426.5.

Negative Binomial (NB)

The ZINB model produced nearly identical coefficient estimates and significance levels for the count portion of the model, with average minutes played and BMI remaining significant. It is not surprising that these models produced similar results. If the data had excess zeros, we would have expected the ZINB model to outperform the NB, but since only 9% of athletes had 0 days of recovery, the data only had very few excess zeros, and thus explaining why the ZINB model behaved so similarly to the NB model.

Furthermore, the zero-inflation component was not statistically meaningful — the intercept had a very large standard error and a p-value of 0.98, suggesting that modeling excess zeros with a separate process did not improve model performance. The ZINB model's AIC was slightly higher (889.06), and the log-likelihood (-426.5) was nearly indistinguishable from that of the NB model.

However, both models have significantly lower AIC values compared to the linear regression model (AIC=1162.51), suggesting they provide a substantially better fit. This is likely due to their ability to account for overdispersion in the number of days missed.

Multinomial Logistic Regression

While the ZINB and NB above models offered valuable insights into the continuous count of days missed, a different approach may help achieve a better model fit. We categorized total days missed into recovery periods:

- 0 days = No Recovery
- 1–2 days = Short Recovery
- 3–14 days = Medium Recovery
- 15–94 days = Long Recovery

This categorical framing allows us to model recovery outcomes as discrete events using multinomial logistic regression. While it's not a requirement for each category to have an equal number of observations, the categories were designed to be relatively balanced. The distribution is shown below:

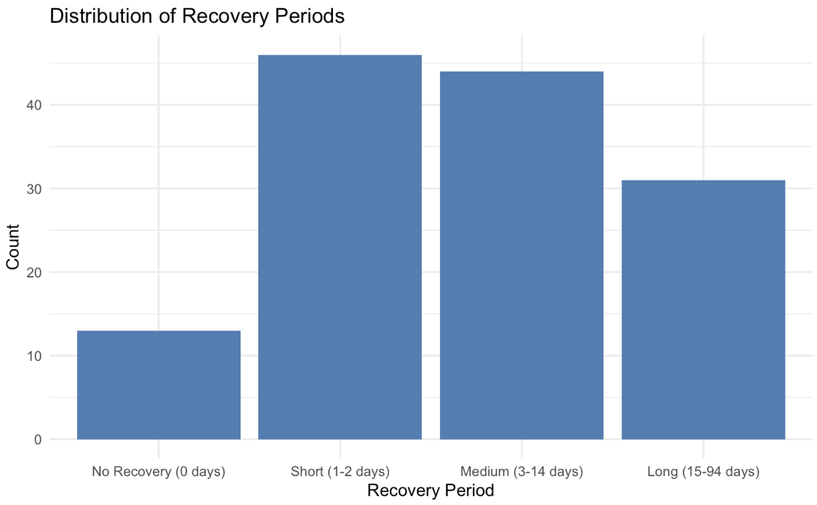


Figure 3.6: **Histogram showing the frequency of athletes across the different recovery period categories.** The time intervals were manually defined to ensure the classes were relatively balanced.

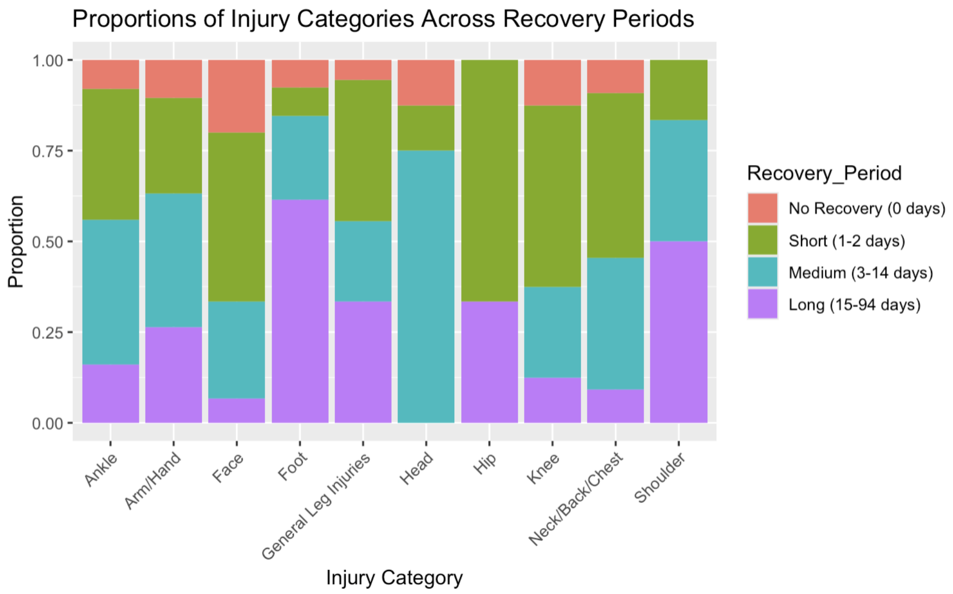


Figure 3.7: **Stacked bar chart depicting proportion of recovery periods based on injury type.** Face injuries had the highest proportion of no recovery, hip injuries had the highest proportion of short recovery, head injuries had the highest proportion of medium recovery, and foot injuries had the highest proportion of long recovery periods.

The figure above offers deeper insight into patterns observed in earlier analyses. For example, while previous models identified hip injuries as significant predictors of total days missed, this graph reveals that hip injury recovery tends to be either very short or very long — the latter often associated with surgeries. However, it's important to note that the recovery period categories were manually defined and somewhat arbitrary; adjusting these boundaries could significantly alter how the distribution appears and is interpreted.

For the model output, we selected "Short Recovery" (1-2 days) as the baseline category because it represents a more typical recovery period that is often observed in injury data. Unlike "No Recovery" (0 days), which may involve cases of minor injuries or no injury at all, "Short Recovery" reflects injuries that do require some recovery time, making it a more meaningful category for comparison.

Below is a summary of the only predictors that were significant:

Predictor	Recovery Category	Estimate	p-value
Average minutes played	Long	-0.104	0.0406
Foot injury	Long	2.655	0.0409

Head injury	Long	-14.311	< 2e-16
Hip injury	No recovery	-14.036	< 2e-16
Hip injury	Medium	-14.692	< 2e-16
Shoulder	No recovery	-13.670	< 2e-16

Table 3.8: **Summary of significant predictors.** Most predictors for recovery period were various injury types.

Aside from the various injury types, average minutes played was also a significant predictor. Since "Short Recovery" is the baseline, the negative estimate of -0.104 for "Average minutes played" in the "Long Recovery" category suggests that for each additional minute played, the likelihood of a player experiencing a long recovery (15-94 days) decreases. In other words, players who play more minutes are less likely to experience a long recovery compared to a short recovery. This is consistent with our previous analysis, where we observed that injured players who play more minutes tend to be key, high-value players for the team, and their recovery periods are often expedited to get them back on the field sooner.

In comparison to the baseline "Short Recovery," significant predictors for "Long Recovery" include fewer minutes played and foot injuries, while head injuries are strongly associated with long recovery periods. For "No Recovery," hip injuries and shoulder injuries are strongly negative, indicating that these injuries are less likely to require no recovery time. This is all consistent with our EDA from Figure 3.7.

Evaluating model performance, the multinomial model had an AIC of 374.86, which is far lower than the AIC of all previous models.This suggests it offers a much better fit by more effectively capturing the categorical variation in recovery periods, compared to models predicting continuous days missed.

3.2.5 Random Forest

Moving forward, we could use a random forest model to predict recovery time due to its ability to capture complex, non-linear relationships that the previous models could not. This approach may offer improved predictive accuracy.

To ensure an apples-to-apples comparison with the multinomial regression model, the Random Forest will also predict recovery period, rather than continuous total missed days due to injury. Since the random forest model cannot be directly compared using AIC, we will evaluate its performance based on AUC metrics.

To create the random forest model, we utilized the caret and randomForest packages. The data was split into training and testing sets using a 80-20% split. The model uses bootstrapping with 25 resampling repetitions to improve model stability and robustness. The mtry tuning parameter, which controls the number of variables randomly sampled at each split, was optimized across different values, with the final model selected based on the best accuracy (mtry=8). This means the final random forest model achieved its highest classification accuracy when considering 8 predictors at each split. This resampling procedure was done to ensure that the model generalizes well to new data by repeatedly sampling and fitting the model.

3.2.6 Model Comparison

The results from the Random Forest model indicate an overall accuracy of 48%. In comparison, the accuracy for the multinomial model was 40%. Greater insight can be made by looking at the AUC values, which tells us how well a model can distinguish between a given class and all other classes.

AUC values	**Multinomial Logistic Regression**	**Random Forest**
No Recovery (0 days)	0.9347	0.8695
Short (1-2 days)	0.3472	0.7048
Medium (3-14 days)	0.5000	0.6985
Long (15-94 days)	0.8859	0.7543

Table 3.9: **Summary of AUC values.** The Random Forest model is better at distinguishing Short and Medium recovery periods, whereas the Multinomial model performs better in identifying No Recovery and Long recovery periods.

The AUC measures how well a model distinguishes between classes. For a given category, it reflects the model's ability to differentiate whether a player belongs to that specific category versus all others. For example, the AUC for the "No Recovery" class was 0.8695, meaning the model had an 87% chance of correctly distinguishing between a randomly selected player with no recovery and one with a different recovery outcome (short, medium, or long). Since an AUC of 0.5 indicates no better than random guessing, and 1.0 indicates perfect classification, an AUC of 0.87 suggests the model has strong discriminatory power for identifying the "No Recovery" class.

Comparing the AUC values, the random forest model demonstrated stronger discrimination between short and medium recovery categories, likely due to its ability to capture non-linearities and interactions. In contrast, the multinomial model showed better performance in distinguishing the long and no recovery classes, potentially because these patterns are more linearly separable or rarer. It's also possible that the multinomial model may be overfitting to these less common classes, while the Random Forest model distributes learning more evenly across all categories.

3.2.7 Summary of Findings

The key findings from Section 3.2 are summarized below:

- **Most recoveries are very short**: Over half (57.46%) of injured athletes recovered within 0–5 days, and 9% returned with no missed days.
- **Injury severity predictors were stronger predictors of recovery time than player value predictors.** The model predicting recovery time solely based on injury severity related predictors outperformed the model which only included team value related predictors. Ultimately, a model with combined predictors performed the best.
- **Negative binomial models outperform linear ones**: Because of overdispersion and many short/no recovery cases, advanced models like the negative binomial provided a much better fit than standard linear regression.
- **Multinomial regression captures risk tiers well**: This model grouped recovery into 4 buckets and found hip injuries and minutes played to be strong predictors of severe (14+ day) recovery, which aligned with EDA.
- **The random forest and multinomial regression models had tradeoffs:** Random Forest excels at distinguishing Short/Medium recoveries from others, while Multinomial better identifies No/Long recoveries.

Chapter 4: Reflection

As we reflect through the journey of this study, the most common result that emerges across these three chapters is the critical role of minutes played in injury risk. This was revealed in Chapter 1, where we also found critical insights into injury factors, establishing the foundation for subsequent modeling. Lower extremity injuries (ankle, knee, foot) were most prevalent, reflecting known biomechanical risks in female athletes, with injuries and missed games rising sharply in recent seasons. While most injuries caused short absences, shoulder/hip injuries often led to prolonged recovery, and early-season injuries (May/June) proved more severe, resulting in longer recovery times and significant team disruption. Most critically, injured players averaged more minutes than non-injured players – this made sense as the best players generally have greater playtime, increasing injury exposure and reinforcing the link between high minutes played and injury risk.

Given that minutes played is a critical factor contributing towards injury risk, Chapter 2 moved onto building a model to predict the optimal number of minutes a player should play in a given game, to minimize injury risk. A three-level mixed effects linear model-trained on non-injured players to establish healthy play habits-predicted minutes played using starter status (which was found to be highly influential) and position as fixed effects, and team/player as random effects. The linear model (assuming a Gaussian distribution) outperformed Gamma/Log-Normal alternatives via RMSE. When comparing the distributions of the random effects, it was found that there was greater variation in minutes played on the athlete compared to the team-level. When the model was applied to injured players, the residuals showed that the model was severely underestimating the injured players' actual minutes, indicating those players were consistently overplayed. Residuals grew larger approaching injury dates, confirming excessive minutes likely contributed to injuries.

In the first section of Chapter 3, a binary classification model was created to determine the likelihood of an injury between the date of a player’s current game and the date of their next game. It was found that minutes played in the current game, number of prior injuries, and whether or not the player was already injured were all significant predictors. This model had an accuracy of 57% but sensitivity of 70%, meaning the model correctly identifies actual injuries over 70% of the time, which is promising given the importance of catching potential injuries over having false alarms. The AUC of the model revealed that 63% of the time, the model will correctly assign a higher injury probability to a player who actually got injured compared to one who was not.

In the second section of Chapter 3, models predicting total days missed due to the injury were compared. A multinomial logistic regression categorized recovery periods (0 days = No Recovery; 1–2 = Short; 3–14 = Medium; 15–94 = Long), while a random forest model also predicted these categories. The random forest excelled at identifying Short/Medium recoveries, whereas the multinomial model better detected No/Long recoveries.

Ultimately, these findings mean injury risk is both predictable to an extent and modifiable through workload management, particularly by optimizing minutes played and monitoring high-exposure periods. The models developed here can be used to inform real-time coaching decisions, such as rotating players during high-risk periods (e.g., early season) or adjusting minutes for injury-prone athletes, while prioritizing early intervention for those flagged by predictive algorithms. These findings provide a valuable resource for coaches, trainers, and medical staff aiming to implement targeted injury prevention and management strategies in women's basketball, balancing performance demands with athlete health.

References

1. ESPN Press Room. (2024, October). *The 2024 WNBA season delivers record viewership across ESPN platforms*. https://espnpressroom.com/us/press-releases/2024/10/the-2024-wnba-season-delivers-record-viewership-across-espn-platforms/
2. WNBA. (2024, April 24). *WNBA announces historic media rights deal with Disney, Prime Video and NBC Sports*. https://www.wnba.com/news/media-rights-deal-disney-prime-nbc
3. Trasolini, N. T., Lu, W. H., Schultz, B., & Pandya, N. K. (2022). Injury epidemiology in women's professional basketball: Has the WNBA met the NBA standard? *Orthopaedic Journal of Sports Medicine, 10*(11), 23259671221136847. https://doi.org/10.1177/23259671221136847
4. Gordon, A. (2024, April 15). *The WNBA's injury problem*. The Next. https://www.thenexthoops.com/features/injury-problem-statistics/
5. Gordon, A. (2023, July 14). *WNBA missing from Elite Basketball Rehab Conference*. The Next. https://www.thenexthoops.com/wnba/wnba-missing-from-elite-basketball-rehab-conference-sports-injury-data/
6. Gilani S, Hutchinson G (2024). _wehoop: Access Women's Basketball Play by Play Data_. R package version 2.1.0, <https://CRAN.R-project.org/package=wehoop>.
7. Seehafer, L. (2023, July 5). *WNBA Injury Tracker: Who gets hurt, how often, and why it matters*. The Next. https://www.thenexthoops.com/wnba/wnba-injury-tracker-who-gets-hurt-how-often-and-why-it-matters/
8. Evans, J. (2024, October 28). *Injury in the Women's National Basketball Association (WNBA) from 2015 to 2019*. KT Insights. https://articles.kangatech.com/wnbainjuries&#8203;:contentReference[oaicite:0]{index=0}
9. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). *Fitting linear mixed-effects models using lme4* (Version 1.1-37) [Computer software manual]. CRAN. https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf
10. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2023). *pROC: Display and analyze ROC curves* (Version 1.18.5) [Computer software manual]. CRAN. https://cran.r-project.org/web/packages/pROC/pROC.pdf