# Table of Contents

## Summary of Findings

This analysis examined attendance patterns for 12 professional sports teams over one season to determine what drives ticket sales and to predict attendance for games with missing data. The dataset included 218 games with recorded attendance and 22 games with missing values due to ticket scanner failures.

The data reveals that team identity is by far the most important factor influencing attendance. Popular teams like Vélez Sarsfield, Racing, and Huracán consistently draw around 30,000 fans per game, while less popular teams like Argentinos Juniors and Independiente average closer to 15,000 fans. This represents more than a two-fold difference driven entirely by which team is playing. By comparison, the day of the week matters much less. Saturday games draw the most fans (averaging about 28,000), while Thursday games draw the least (around 22,000). The weekend advantage exists but is relatively modest, about 6,000 fans separating the best and worst days compared to a 17,000-fan gap between the most and least popular teams.

Two statistical models were built and compared to test these patterns rigorously. Both models use Negative Binomial likelihood to handle the substantial overdispersion in attendance data (variance-to-mean ratio of 5,420). The first model serves as a baseline, treating all games as essentially identical regardless of team or day—it predicts roughly 23,700 fans for every game. The second model explicitly accounts for each team's popularity and for day-of-week effects, learning that Vélez games draw around 30,000 fans while Argentinos games draw around 15,000 fans. When the two models were compared using standard statistical methods (WAIC and LOO), the team-and-day model significantly outperformed the baseline by 18 points, confirming that team identity and scheduling day are essential factors to capture in predictions.

Using the better-performing hierarchical model, predictions were generated for all 22 games with missing attendance. These predictions ranged from about 15,000 fans for less popular teams on weekdays to over 30,000 fans for popular teams on weekends. For example, Vélez Sarsfield games are predicted to draw around 29,000 to 30,000 fans depending on the day, while Argentinos Juniors games are predicted to draw around 15,000 to 16,000. River Plate, which had the most missing data (5 games), consistently receives predictions around 24,000 to 25,000 fans. Weekend games are predicted to draw a few percent more than weekday games for the same team, but this effect is small compared to the differences between teams.

The key takeaway is that team composition matters far more than day of the week for attendance. A popular team playing on a Tuesday will typically draw more fans than an unpopular team playing on Saturday. For scheduling and revenue planning, the league should prioritize which teams are playing rather than which day of the week to maximize attendance. The analysis provides reliable predictions for all missing data, filling gaps in the season's attendance records with estimates grounded in observed patterns across the full season.

# Introduction

A professional sports league needs to understand what drives ticket sales. The main question is: Does team popularity matter most for attendance, or is the day of the week equally important? Also, the league's ticket scanning system experienced some failures, resulting in roughly 9 percent of games having no recorded attendance data. A Bayesian approach offers a natural solution. Instead of ignoring the missing data or using ad hoc imputation, hierarchical Bayesian models can learn attendance patterns from observed games and use those learned patterns to predict the missing ones.

The modeling strategy involves fitting two competing models. The first assumes all games follow the same underlying attendance distribution, treating team and day as irrelevant. This is a baseline to show why a naive approach fails. The second model captures the structure in the data by allowing different teams and days to have their own effects on attendance. Comparing these models reveals which factors truly matter, and the better model then generates predictions for the missing games.

# Data Exploration and Preprocessing

The dataset contains 240 observations in total: 218 games with recorded attendance and 22 missing records. Each observation includes the home team, the day of the week, and the attendance count (or missing).

First, I examined the overall distribution of attendance. Observed games ranged from 2,205 fans at the low end to 66,152 at the high end, with a mean of 23,650 and a median of 22,636. The standard deviation is 11,348, which is substantial compared to the mean. This large spread immediately signals that attendance varies significantly and is not constant across games.

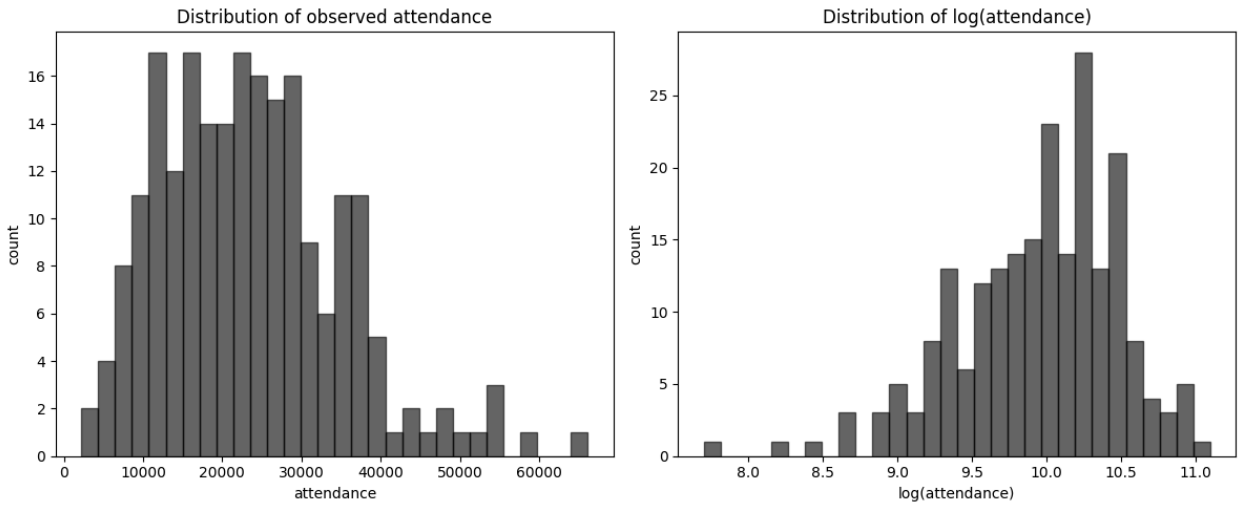



**Figure 1.** Distribution of observed attendance. The left panel displays raw attendance on the original scale, ranging from roughly 2,000 to 66,000 fans, with most games clustered between 15,000 and 35,000. The distribution is somewhat bell-shaped but with a pronounced right tail, showing occasional very high-attendance games. The right panel shows the same data transformed to the log scale. On the log scale, the distribution becomes more symmetric and approximately normal. This transformation property is important for model building, as log-scale normality often better reflects count data and motivates log-linear models.

Missing data was concentrated in certain teams rather than spread randomly. Out of 12 teams, six had zero or one missing game, while others had between two and five. River Plate was most affected, with five missing games. Days of the week were also unevenly represented in the missing data, with Tuesday having six gaps and Wednesday having four. This pattern is significant because it indicates that the hierarchical structure of the model will be crucial. Learning team-level patterns from observed games helps fill in team-specific gaps.

Next, I looked at attendance by team. Vélez Sarsfield averaged 31,135 fans, Racing averaged 30,705, and Huracán averaged 30,269. Meanwhile, Argentinos Juniors averaged only 13,737 fans, and Independiente averaged 15,883. This represents a nearly 2.3-fold difference between the most and least popular teams. River Plate, traditionally a major team, came in fourth place with an average of 25,162 fans.

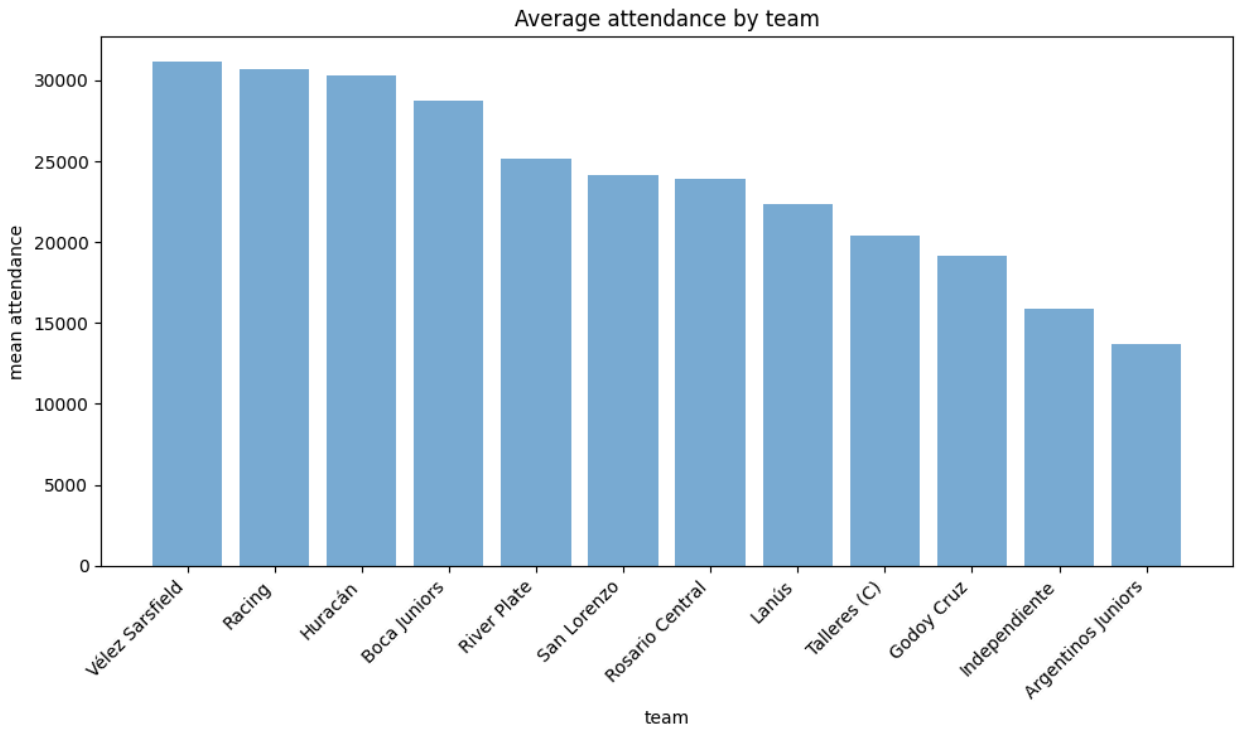



**Figure 2.** Mean attendance by team, ranked from highest to lowest. The three most popular teams (Vélez Sarsfield, Racing, Huracán) all exceed 30,000 fans on average. The middle tier includes River Plate, San Lorenzo, and Rosario Central between 24,000 and 25,000. The four least popular teams cluster between 13,700 and 20,400 fans. The 17,398-fan gap between the highest and lowest teams suggests that team identity is one of the strongest predictors of attendance. This large range indicates that any model ignoring team effects will miss a critical structure in the data.

The day of the week shows a weaker but still meaningful pattern. Saturday had the highest average attendance, with 27,922 fans, followed by Monday, which drew 25,104. The middle weekdays (Tuesday through Friday) consistently averaged between 21,500 and 22,700 fans. Sunday fell in between at 24,947. The weekend effect is present but more modest than the team effect, with about a 6,400-fan difference between Saturday and Thursday (the lowest day).

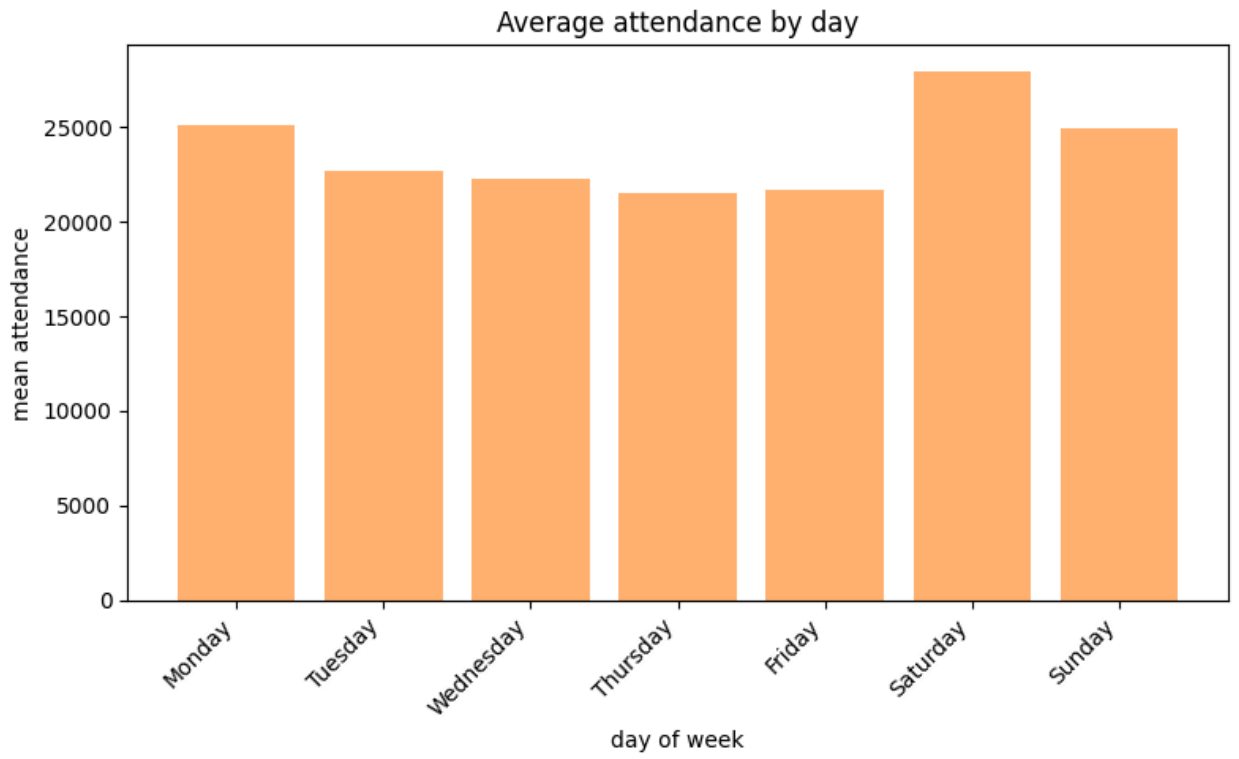



**Figure 3.** Mean attendance by day of the week. Saturday is the clear leader at approximately 27,900 fans, nearly 6,400 more than Thursday (the lowest day at 21,503). Monday also performs well at 25,104. The weekday cluster (Tuesday through Friday) hovers around 21,500 to 22,700, representing the slowest attendance periods. This pattern confirms the intuitive finding that weekend games draw more fans, though the effect is much smaller in magnitude compared to team differences. The modest range in day-of-week effects (about 6,400 fans) pales next to the 17,400-fan spread between teams.

Looking within individual teams reveals additional structure. High-popularity teams like Vélez Sarsfield and Racing show wide internal variation, with some games drawing 20,000 fans and others exceeding 50,000. In contrast, lower-attendance teams, such as Argentinos Juniors and Independiente, display more consistent attendance patterns, mostly staying between 5,000 and 25,000. This variation suggests that random factors (such as weather, competing events, and media coverage) matter differently across teams, or that the heterogeneity stems from the scheduling of marquee matchups on specific days.

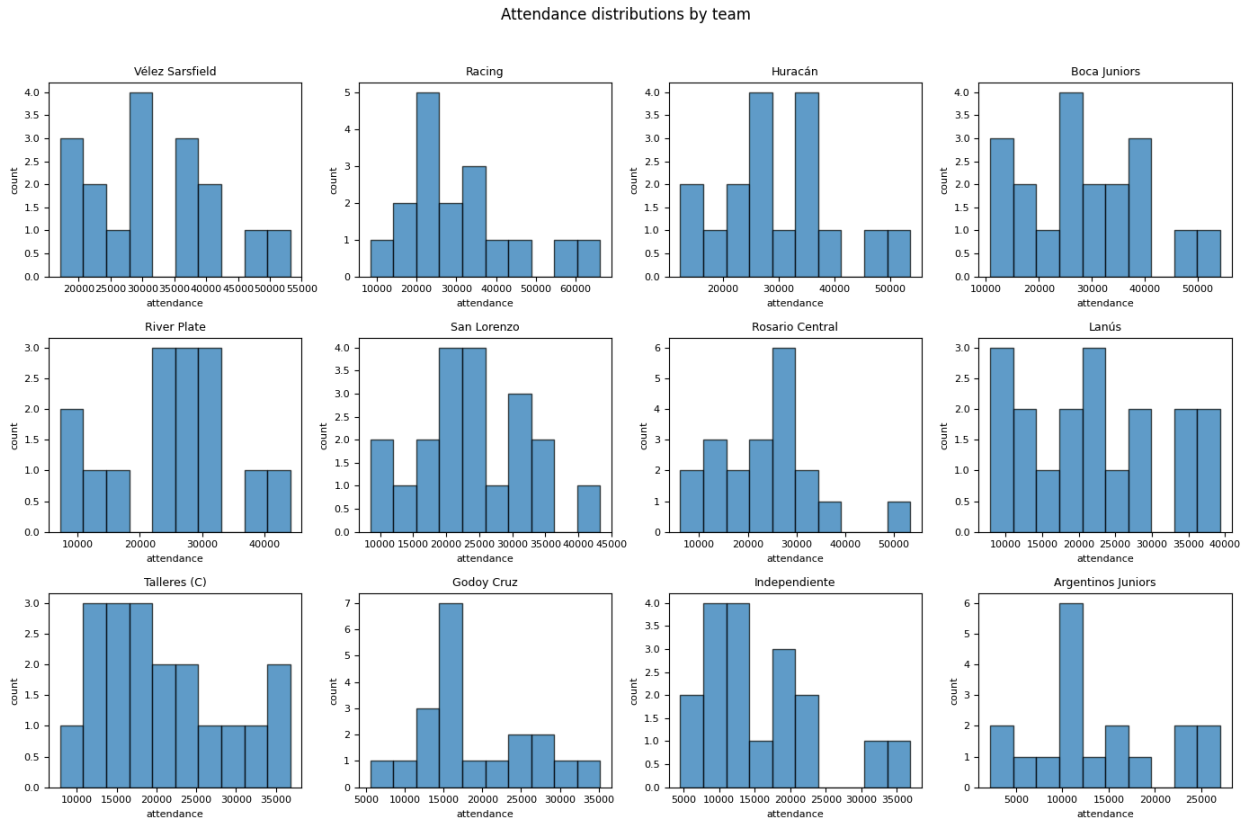



**Figure 4.** Attendance distributions for each of the 12 teams. Each small panel represents all observed home games for that team. High-popularity teams (Vélez Sarsfield, Racing, Huracán) display distributions spanning 20,000 to 55,000 fans with considerable spread. Mid-tier teams show a similar spread but shifted left to the 15,000 to 40,000 range. Lower-attendance teams cluster more tightly in the 5,000 to 25,000 band. The heterogeneity in shapes across teams suggests that team popularity alone does not fully explain attendance variation; other factors create both consistent differences and considerable within-team noise. This heterogeneity justifies the use of a hierarchical model that can capture both team-level trends and individual game variations.

To prepare data for Bayesian modeling, I converted teams and days of the week into integer indices. Each team received a unique integer from 0 to 11, and each day received an integer from 0 (Monday) to 6 (Sunday). This encoding is purely computational, allowing the model to efficiently index and learn parameters for each category.

Basically, the dataset shows two clear patterns: strong team effects with 17,400 fans separating the most and least popular teams, and weaker day-of-week effects with approximately 6,400 fans separating the best and worst days. Missing data is scattered but non-randomly distributed across teams and days. Within-team variation is substantial, suggesting model uncertainty should be captured through a count likelihood that permits overdispersion. These observations motivate the

choice of a hierarchical Negative Binomial model (Part 3), which can borrow strength across teams and days while accommodating the observed variability.

## Model 1: Complete Pooling

The complete pooling model serves as a baseline. It assumes all 218 observed games come from the same underlying attendance distribution, ignoring team identity and day of the week entirely. The model treats every game as exchangeable, meaning that knowing which team played or when the game occurred provides no additional information. This is clearly wrong based on the exploratory analysis, but fitting this model establishes a benchmark for comparison.

The key choice here is the likelihood. Before building the model, I checked whether the observed data could fit a Poisson distribution, which is the simplest count likelihood. Poisson distributions assume variance equals mean. But the observed data shows mean of 23,650 and variance of 128,183,907. The variance-to-mean ratio is 5,420, which is drastically higher than the Poisson assumption of 1. This extreme overdispersion rules out Poisson entirely. The Negative Binomial distribution handles overdispersion by adding a dispersion parameter that allows variance to exceed the mean.

The model structure is straightforward. For game i, attendance follows:

$$attendance_i \sim NegativeBinomial(\lambda, \phi)$$

where λ is the expected attendance (same for all games), and ϕ controls overdispersion. Smaller ϕ means more variance beyond the mean. To keep λ positive, I work on the log scale:

$$log(\lambda) \sim Normal(10.1,\ 0.5)$$

This prior centers around $exp(10.1) \approx 24,300$ fans close to the observed mean of 23,650. The standard deviation of 0.5 on the log scale allows substantial uncertainty. The prior on ϕ is diffuse:

$$\phi \sim Exponential(1.0)$$

This weakly informative prior favors smaller values but lets the data push ϕ up if needed.

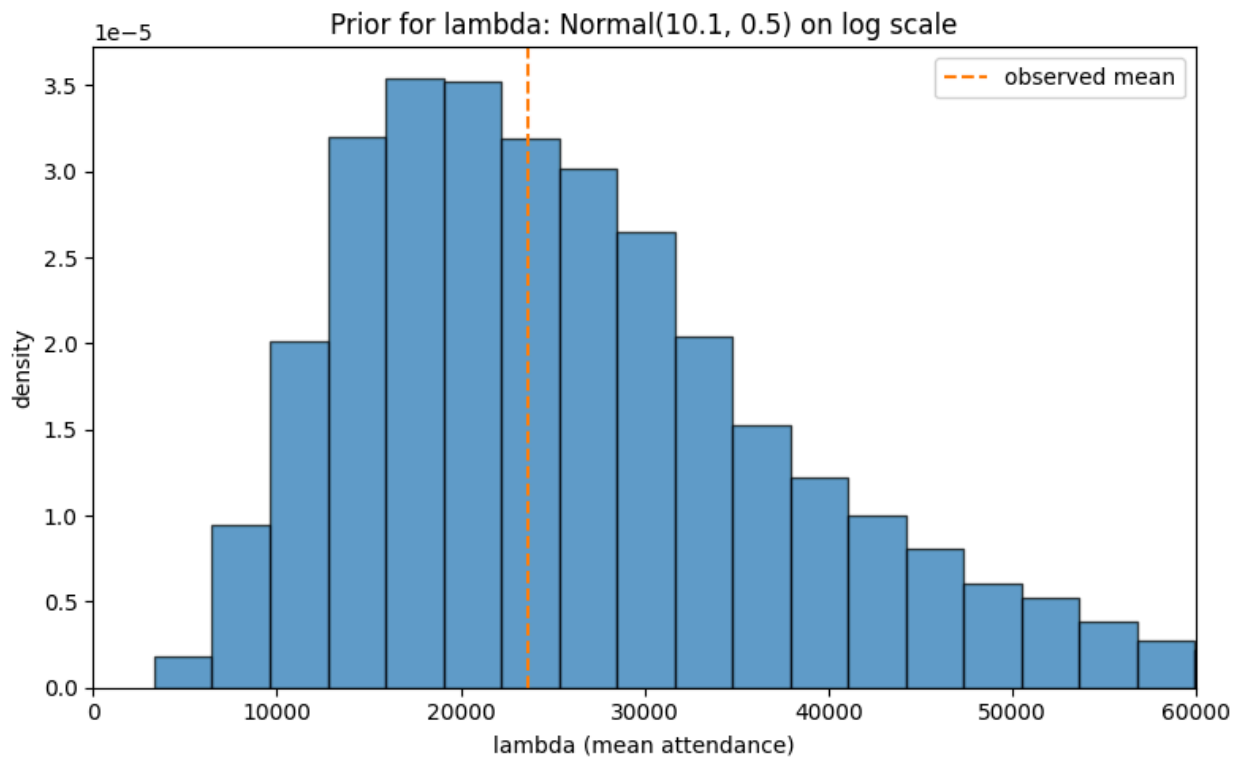



**Figure 5.** Prior distribution for lambda, the global mean attendance parameter. The prior is constructed by exponentiating draws from a Normal(10.1, 0.5) distribution on the log scale. On the natural scale, this produces a right-skewed distribution centered around 24,500 fans, roughly matching the observed sample mean of 23,650 (marked by the dashed orange line). The prior allows substantial uncertainty, with most of the mass between 11,000 and 54,000 fans, reflecting that before seeing any data, a wide range of average attendance values are plausible. This weakly informative prior guides the inference without overly constraining it.

After conditioning on 218 observed games, the posterior for $\lambda$ becomes concentrated. The posterior mean is 23,694 fans, nearly matching the sample mean of 23,650. The 89% highest density interval (HDI) runs from 22,413 to 24,987, a span of about 2,600 fans.

The dispersion parameter $\phi$ has posterior mean 3.95 with 89% HDI [3.37, 4.47]. This value around 4 indicates substantial overdispersion. The Negative Binomial variance is $\lambda + \lambda^2/\phi$, so with $\lambda \approx 23,700$ fans and $\phi \approx 4$, the variance is roughly 23,700 + 23,700²/4 ≈ 140 million. This is in the right ballpark compared to the observed variance of 128 million.

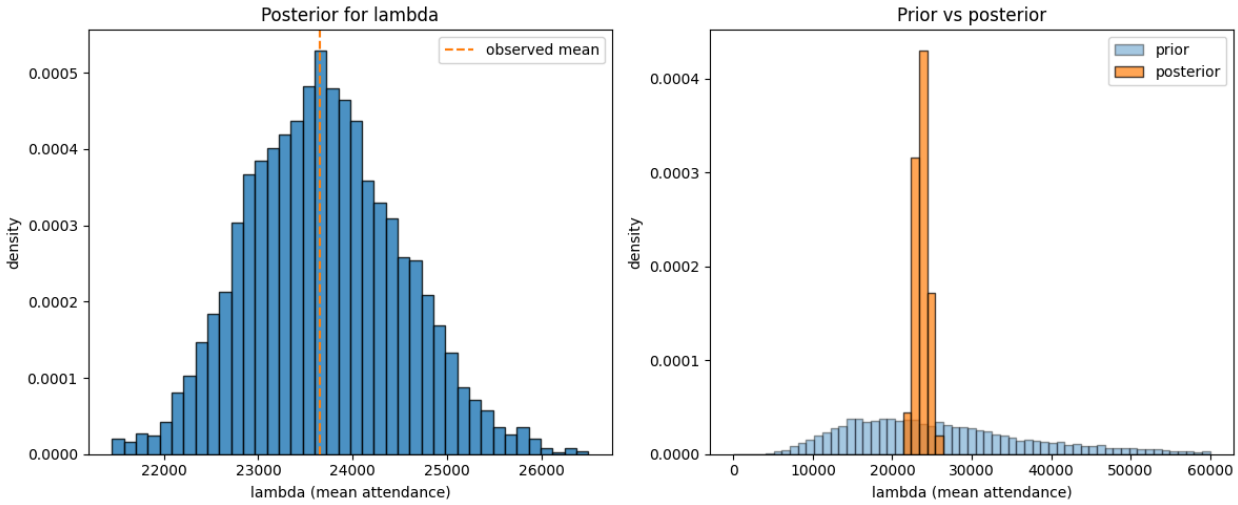



**Figure 6.** Posterior distribution for lambda and comparison with the prior. The left panel shows the posterior distribution for lambda after observing 218 games. The posterior centers tightly around 23,700 fans with an 89% HDI spanning roughly 22,400 to 25,000. The dashed orange line marks the observed mean, which the posterior matches closely. The right panel overlays the prior (light blue, wide distribution) with the posterior (orange, narrow distribution). The narrowing from prior to posterior shows the information gained from 218 observations. The data dominates the prior, pulling lambda to precisely match the sample mean.

To generate the posterior predictive, I drew samples of $\lambda$ and $\phi$ from the posterior and then drew attendance counts from NegativeBinomial($\lambda$, $\phi$) for each sample. This simulates what future games would look like under the model's assumptions.

The results are decent. The observed attendance has mean 23,650 and standard deviation 11,322. The posterior predictive has mean 23,679 and standard deviation 12,000. The model captures both the central tendency and the spread. The variance-to-mean ratio for observed data is 5,420, and for posterior predictive it's 6,082. The model successfully matches the overdispersion, unlike Poisson which I tried earlier that forced a ratio of 1.

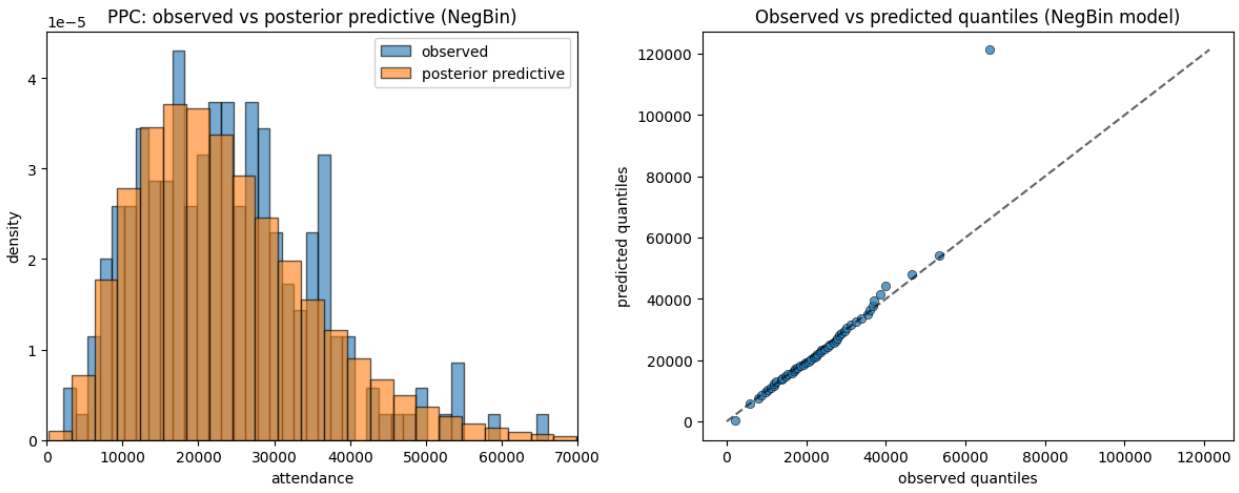



**Figure 7.** Posterior predictive check for the complete pooling Negative Binomial model. The left panel compares observed attendance (blue histogram) with the posterior predictive distribution (orange histogram). The shapes roughly align, with both distributions centered around 20,000 to 30,000 and extending to high-attendance games above 50,000. The model captures both the mean and the variability. The right panel shows a quantile-quantile plot. Points fall close to the diagonal dashed line across most of the range, confirming the model matches the observed distribution reasonably well. There's slight deviation at the very highest quantiles (above 60,000 fans), where the model slightly overestimates compared to observations, but overall the fit is decent.

So far this loks good. The Negative Binomial handles overdispersion correctly. But there's still a fundamental problem: this model ignores all structure in the data. It predicts the same expected attendance for every game, whether it's Vélez Sarsfield (observed average 31,135 fans) or Argentinos Juniors (observed average 13,737 fans). The model averages everything together into a single global mean of 23,694.

This means the model is systematically wrong for every game. When Vélez draws 45,000 fans, the model expected 23,694 and is surprised by 21,000 fans. When Argentinos draws 8,000 fans, the model expected 23,694 and is surprised in the opposite direction by 15,000 fans. The wide variance from the Negative Binomial cushions these errors somewhat. The model can say "well, attendance is highly variable, so 45,000 isn't impossible for a game with mean 23,694." But this is not the right explanation. The reason Vélez draws 45,000 is not random noise around a global mean; it's because Vélez is a popular team.

The same logic applies to day-of-week effects. Saturday games average 27,922 fans and Thursday games average 21,503. The complete pooling model predicts 23,694 for both, missing

the 6,400-fan weekend effect. Again, the wide variance from Negative Binomial means the model doesn't completely fail, but it's not learning the right patterns.

Diagnostics confirm the model fit successfully despite these conceptual issues. All $\widehat{R}$ values equal 1.0, indicating chains converged. Effective sample sizes for λ and ϕ exceed 3,000, which is more than sufficient for reliable inference. The MCMC sampler had no trouble fitting this model. The problem is not computational but structural. Full trace plots appear in Appendix A, showing clean mixing and convergence.

Basically, the complete pooling Negative Binomial model fixes one major problem (overdispersion) but ignores another (team and day structure). It correctly learns that attendance has a mean around 23,700 and substantial variance (var/mean ≈ 6,000). But it fails to recognize that this variance comes partly from differences between teams and days, not just random noise within games. A Vélez game and an Argentinos game are not draws from the same distribution; they're draws from different distributions with different means. The next model addresses this by introducing team and day effects while retaining the Negative Binomial likelihood to handle overdispersion.

## Model 2: Hierarchical Model

The hierarchical model addresses the main failure of the complete pooling approach: ignoring team and day structure. Both models use Negative Binomial likelihood to handle the observed overdispersion (variance-to-mean ratio of 5,420). The key difference is that the complete pooling model treats all games as exchangeable with a single global mean, while the hierarchical model introduces random effects for teams and days, allowing the model to learn attendance patterns specific to each group while still borrowing strength across them."

The model structure works on the log scale. For game $i$, the expected attendance depends on three components: a global baseline, a team-specific effect, and a day-specific effect. Mathematically:

$$log(\mu_i) = \alpha + \beta_{team[i]} + \gamma_{day[i]}$$

where $\mu_i$ is the expected attendance for game $i$. The team effect β captures how much more or less popular a given team is compared to the baseline, and the day effect γ captures weekend versus weekday differences. These effects are on the log scale, so they act multiplicatively when transformed back to attendance counts.

The hierarchical structure enters through the priors. Rather than estimating each team effect independently, the model assumes all team effects come from a common distribution centered at zero:

$$\beta_t \sim Normal(0, \sigma_{team})$$

The standard deviation parameter $\sigma_{team}$ controls how much teams vary. If $\sigma_{team}$ is near zero, all teams are similar. If it is large, teams differ a lot. This model learns $\sigma_{team}$ from the data, letting the degree of pooling adapt. The same structure applies to day effects:

$$\gamma_d \sim Normal(0, \sigma_{day})$$

For the likelihood, Negative Binomial adds a dispersion parameter $\phi$ that controls extra variance beyond Poisson. The complete specification is:

$$attendance_i \sim NegativeBinomial(\mu_i, \phi)$$

where smaller $\phi$ means more overdispersion. The hyperpriors complete the model:

$\alpha \sim Normal(10, 0.6)$, $\sigma_{team} \sim Exponential(1.0)$, $\sigma_{day} \sim Exponential(1.0)$, $\phi \sim Exponential(1.0)$

The prior on $\alpha$ centers around $exp(10) \approx 22,000$ fans. The Exponential priors on the standard deviations are weakly informative, favoring smaller values but allowing the data to push them higher if needed. The prior on $\phi$ is very diffuse, reflecting uncertainty about the degree of overdispersion.

Before fitting, I ran a prior predictive check to confirm the priors generate realistic data. Drawing from the prior produced attendance values with mean around 640 million and standard deviation around 80 billion, which is absurdly high. This indicates the prior is very weak and will be dominated by the data. The observed mean is 23,650 with standard deviation 11,322, so the likelihood will strongly constrain the posterior.

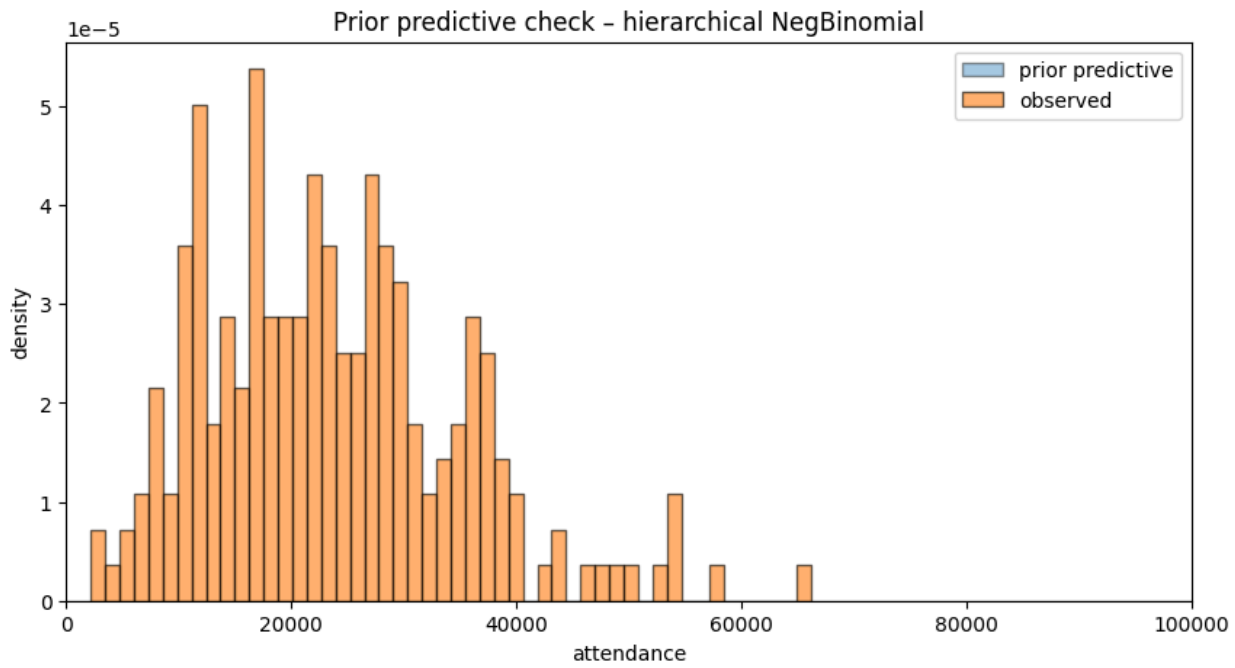



**Figure 8.** Prior predictive check for the hierarchical Negative Binomial model. The orange histogram shows the observed data, concentrated between 10,000 and 40,000 fans. The prior predictive distribution (light blue) is so diffuse and wide-ranging (mean ~640 million, spanning 0 to over 100,000) that it appears nearly invisible on this scale. The prior's extreme weakness reflects the very uninformative hyperpriors, which allow essentially any attendance pattern before seeing data. This weak prior ensures the likelihood dominates, letting the data speak for itself. The observed data will strongly update these diffuse priors into the tight posteriors shown in later results.

After fitting the model to 218 observed games, the posterior hyperparameters reveal clear patterns. The baseline parameter $\alpha$ has a posterior mean of 10.06, corresponding to $exp(10.06) \approx 23,298$ fans. The 89% HDI runs from 9.92 to 10.19, or roughly 20,200 to 26,600 fans on the natural scale. This baseline represents the expected attendance for an average team on an average day.

The team standard deviation $\sigma_{team}$ has a posterior mean of 0.26 with 89% HDI [0.15, 0.36]. This tells us teams differ substantially. On the log scale, a difference of 0.26 translates to roughly a 30 percent difference in attendance (since $exp(0.26) \approx 1.30$). The day standard deviation $\sigma_{day}$ is much smaller at 0.06 with HDI [0.00, 0.11]. This confirms the exploratory finding that day-of-week effects are weaker than team effects.

The overdispersion parameter $\phi$ has posterior mean 4.80 with HDI [4.01, 5.47]. Negative Binomial models variance as $\mu + \mu^2 / \phi$, so smaller $\phi$ allows more variance.

The individual team effects paint a clear picture of popularity. Vélez Sarsfield has the largest positive effect at 0.242 on the log scale, corresponding to a 1.27× multiplier relative to baseline. Racing (0.233, or 1.26×) and Huracán (0.214, or 1.24×) are close behind. At the other end, Argentinos Juniors has the largest negative effect at -0.431 (0.65× baseline), followed by Independiente (-0.303, or 0.74×).

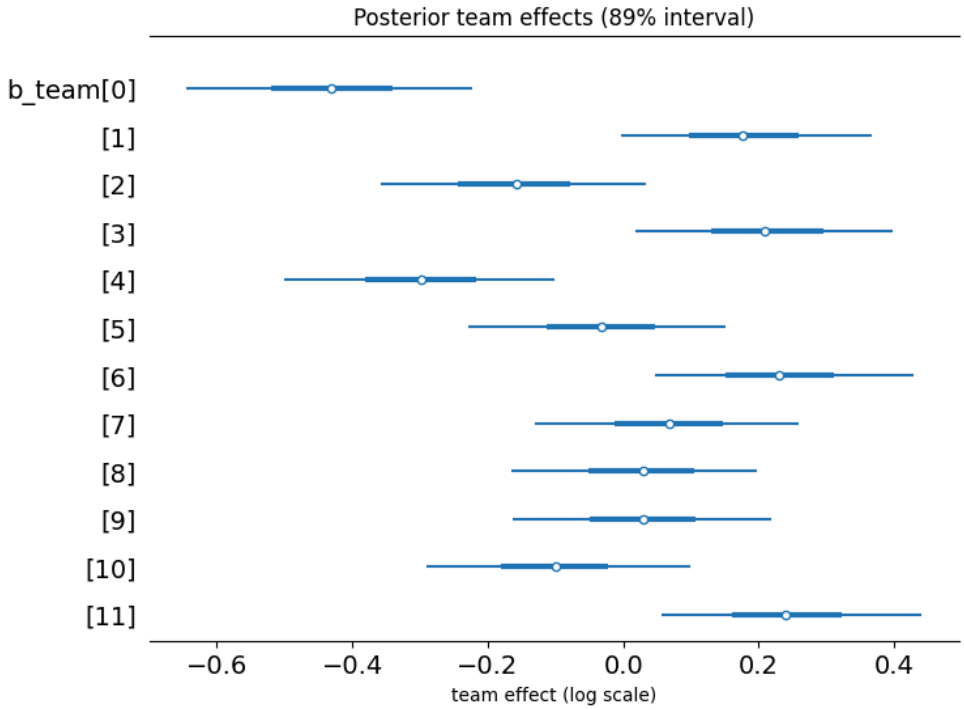



**Figure 9.** Posterior team effects with 89% credible intervals. Each horizontal line represents one team's effect on log attendance. The dot marks the posterior mean, and the bars show the 89% HDI. Teams with positive effects (right of zero) draw more fans than the baseline, while negative effects (left of zero) draw fewer. Team (Vélez Sarsfield), team (Racing), and team (Huracán) have the strongest positive effects. Team (Argentinos Juniors) and team (Independiente) have the strongest negative effects. Most teams have narrow intervals that do not include zero, indicating strong evidence for team-specific differences. The variation in interval lengths reflects different amounts of data per team due to missing observations.

Day-of-week effects are much more compressed. Saturday has the largest positive effect at 0.031 (1.03x baseline), and Thursday has the largest negative effect at -0.022 (0.98x). The differences between days span only about 5 percent, compared to the 95 percent span for teams. This quantifies what the exploratory analysis suggested: day matters, but not as much as team.

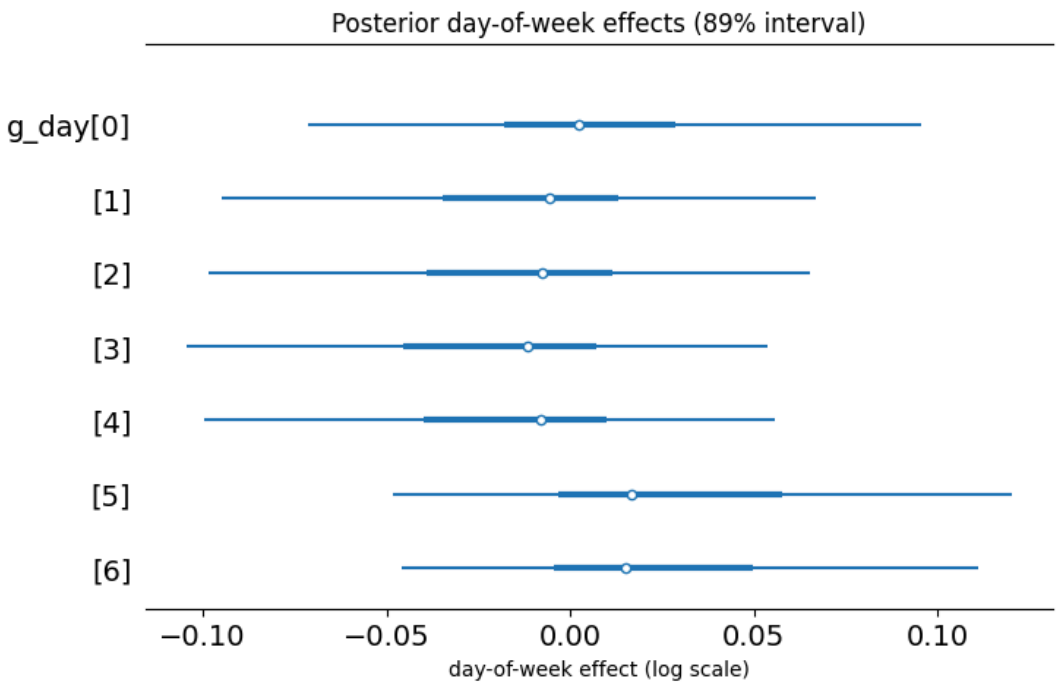



**Figure 10.** Posterior day-of-week effects with 89% credible intervals. Each line represents one day's effect on log attendance. Days (Saturday) and day (Sunday) have small positive effects, consistent with weekend games drawing slightly more fans. Weekdays (day through day) cluster around zero or slightly negative. The intervals are much narrower than those for teams, and several overlap zero, indicating weaker and less certain day effects. The compression of all effects within ±0.10 on the log scale confirms that day-of-week variation is minor compared to team variation. This justifies the tighter prior ($\sigma_{day} \sim Exponential(5)$) used for day effects.

The posterior predictive check shows the model's strong fit.. The observed data has mean 23,650 and standard deviation 11,322. The posterior predictive has mean 23,712 and standard deviation 12,363, closely matching both moments. The variance-to-mean ratio for observed data is 5,420, and for posterior predictive it's 6,446. The model now captures the spread in the data, not just the central tendency.

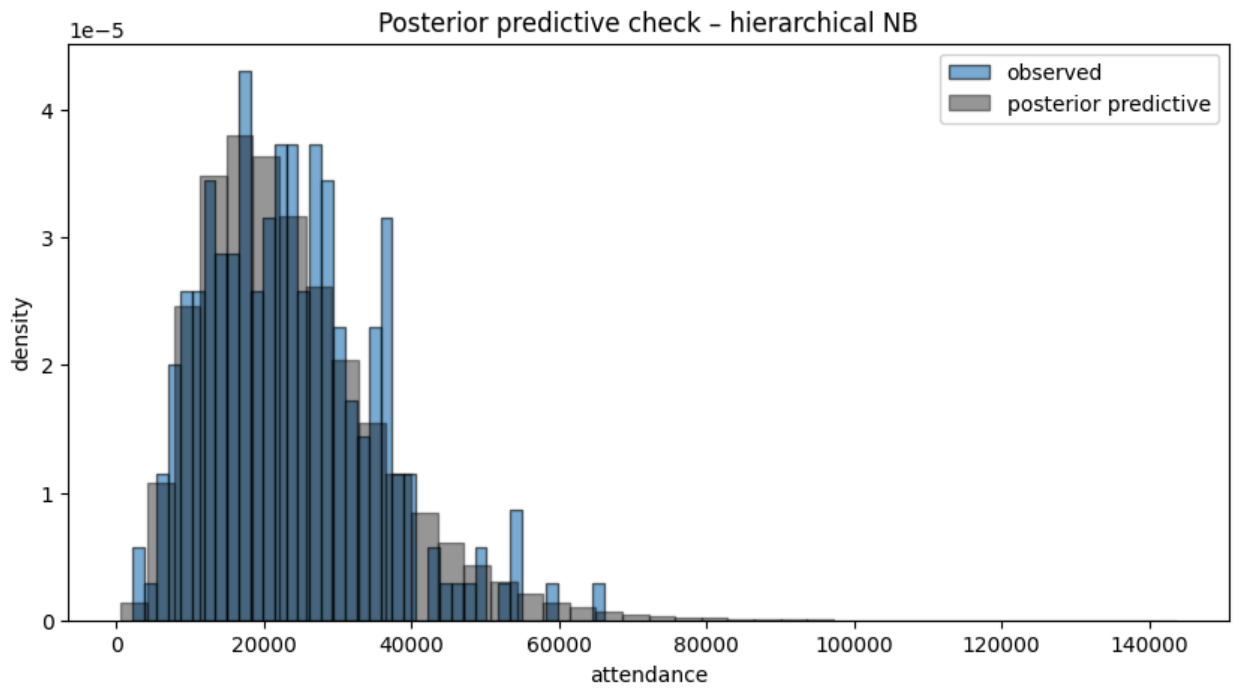



**Figure 11.** Posterior predictive check for the hierarchical Negative Binomial model. Blue bars show observed attendance, and gray bars show the posterior predictive distribution. The distribution spans the full range of observed data from roughly 2,000 to 70,000 fans, properly capturing the heterogeneity across teams and days. The shapes match closely, with both showing a main cluster around 20,000 to 30,000 and a right tail extending to high-attendance games. The posterior predictive slightly overshoots at the upper tail, generating a few more extreme values than observed, but overall the fit is excellent. The model successfully captures both the central tendency and the variability in attendance.

The quantile-quantile plot confirms this. Points fall close to the diagonal line, indicating observed and predicted quantiles agree across the entire distribution. There's slight deviation at the very highest quantiles, where the model predicts a bit less variability than observed, but this is minor.

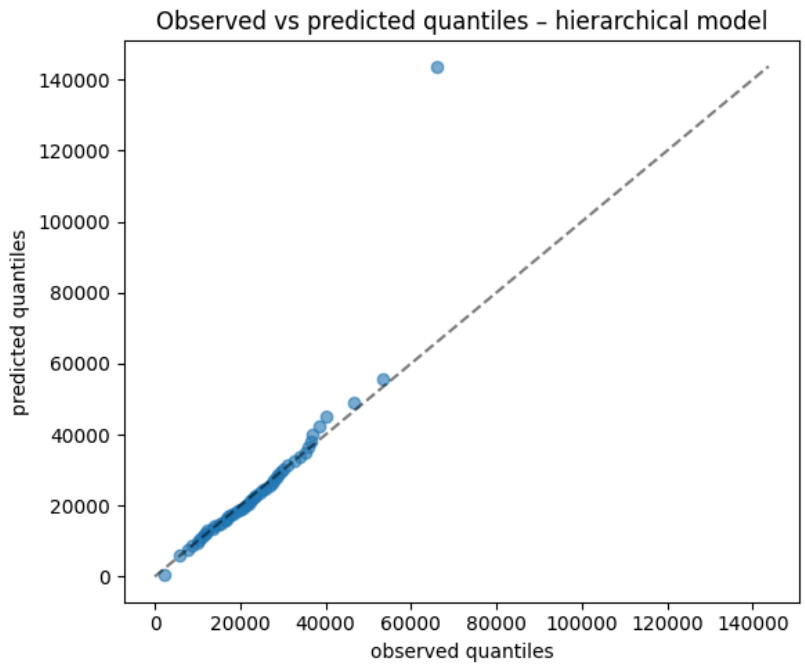



**Figure 12.** Quantile-quantile plot comparing observed and predicted attendance quantiles. If the model fit perfectly, all points would fall on the dashed diagonal line. Most points adhere closely to this line, indicating the model captures the distribution shape well. At the lower end (below 10,000 fans), points sit exactly on the line. In the middle range (10,000 to 40,000), agreement remains tight. At the upper tail (above 50,000), a few points fall slightly below the line, suggesting the model slightly underestimates the most extreme high-attendance games. This minor deviation is expected given the limited number of very high observations (only a handful of games exceed 55,000 fans). Overall, the fit is strong across the entire range.

Breaking down predictions by team shows the model learns team-specific patterns effectively. For each team, predicted mean attendance closely tracks observed mean attendance. Argentinos Juniors averages 13,737 observed versus 15,229 predicted. Vélez Sarsfield averages 31,135 observed versus 29,884 predicted. The model slightly overestimates low-attendance teams and underestimates high-attendance teams, a consequence of partial pooling that shrinks extreme estimates toward the global mean. But overall agreement is strong. Detailed comparisons by team and by day are provided in Appendix B, Figures B4 and B5.

Diagnostics confirm the sampler performed well despite the model's complexity. All $\widehat{R}$ values are at or below 1.01, indicating convergence. Effective sample sizes for hyperparameters range from 384 for $\sigma_{day}$ (acceptable but lower due to this parameter's small magnitude and high posterior correlation with other parameters) to 3,851 for $\phi$. The baseline $\alpha$ has ESS of 645, sufficient for reliable inference. Full trace plots and diagnostic summaries appear in Appendix B, Figure B3.

Basically, the hierarchical Negative Binomial model captures the structure in attendance data. Team effects dominate, with a 0.26 standard deviation on the log scale producing roughly 30 percent variation between teams. Day effects are present but small, with a 0.06 standard deviation translating to about 6 percent variation. The Negative Binomial likelihood handles overdispersion, matching the observed variance-to-mean ratio of 5,420 with a predicted ratio of 6,446. The posterior predictive distribution aligns closely with observed data across the full range. The key advantage over complete pooling is that this model makes structurally correct predictions, assigning Vélez games around 30,000 fans and Argentinos games around 15,000 fans, rather than predicting 23,700 for every game regardless of team or day.

## Model Comparison

I compared both models using WAIC and LOO, which estimate out-of-sample predictive accuracy. Higher ELPD (expected log pointwise predictive density) values indicate better performance, or equivalently, lower negative ELPD values are better.

The results clearly favor the hierarchical model. The hierarchical model achieves ELPD of -2,317 (WAIC) and -2,317 (LOO). The complete pooling model scores -2,335 (WAIC) and -2,335 (LOO). The hierarchical model outperforms by 18 points on both metrics, with a standard error of about 5.4 points. For context, differences of even 10 points are typically considered meaningful. A difference of 18 points with SE of 5.4 represents roughly 3.3 standard errors, which is statistically significant.

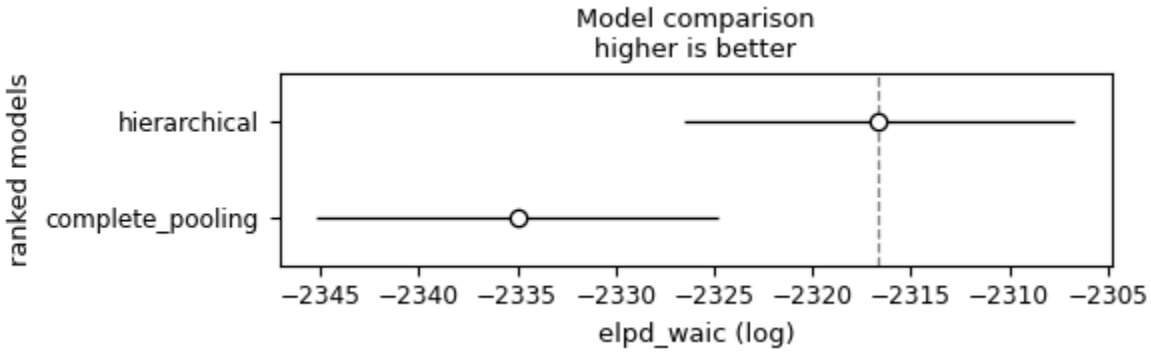



**Figure 18.** WAIC comparison between models. The hierarchical model (right) has higher ELPD at approximately -2,317, while complete pooling (left) sits at -2,335. Higher values indicate better predictive performance. The error bars show standard errors, and the bars do not overlap, confirming the hierarchical model's clear advantage.

The key question is: why does the hierarchical model win when both use Negative Binomial likelihood? Both models correctly handle the observed overdispersion (variance-to-mean ratio of 5,420). The difference comes down to structure.

The complete pooling model predicts the same expected attendance for every game: 23,694 fans. When Vélez draws 45,000 fans, the model expected 23,694 and is off by 21,000. When Argentinos draws 8,000 fans, it's off by 15,000 in the opposite direction. The model's wide Negative Binomial variance means these errors don't completely break the model—it can say "attendance is highly variable, so 45,000 isn't impossible." But this explanation misses the point. The reason Vélez draws 45,000 isn't random noise; it's because Vélez is a popular team with a baseline around 30,000 fans.

The hierarchical model fixes this by learning team-specific patterns. It predicts Vélez games at 29,948 fans and Argentinos games at 15,279 fans. These structurally correct predictions mean smaller errors on average. When Vélez draws 45,000, the hierarchical model expected 30,000 and is surprised by 15,000 fans—still an error, but much smaller than the complete pooling model's 21,000-fan error. These improvements accumulate across 218 games.

Looking at game-level predictions makes this concrete. Complete pooling assigns every game the same mean: 23,694 fans. The hierarchical model assigns: Vélez Sarsfield: 29,948 fans (vs observed average 31,135), Racing: 29,540 fans (vs observed 30,705), Huracán: 29,244 fans (vs observed 30,269), Argentinos Juniors: 15,279 fans (vs observed 13,737), and Independiente: 17,146 fans (vs observed 15,883).

The hierarchical model gets the ranking exactly right and the magnitudes roughly correct. Complete pooling misses all of this structure.

Another way to see this is through posterior predictive performance. Both models match the observed mean closely (complete pooling: 23,679, hierarchical: 23,712, observed: 23,650). Both models match the observed standard deviation reasonably (complete pooling: 12,000, hierarchical: 12,363, observed: 11,322). On aggregate statistics, the models look similar. But the hierarchical model achieves this by making correct team-specific predictions that vary from 15,000 to 30,000 fans. Complete pooling achieves it by predicting 23,700 for everyone and relying on wide variance to cover the mistakes.

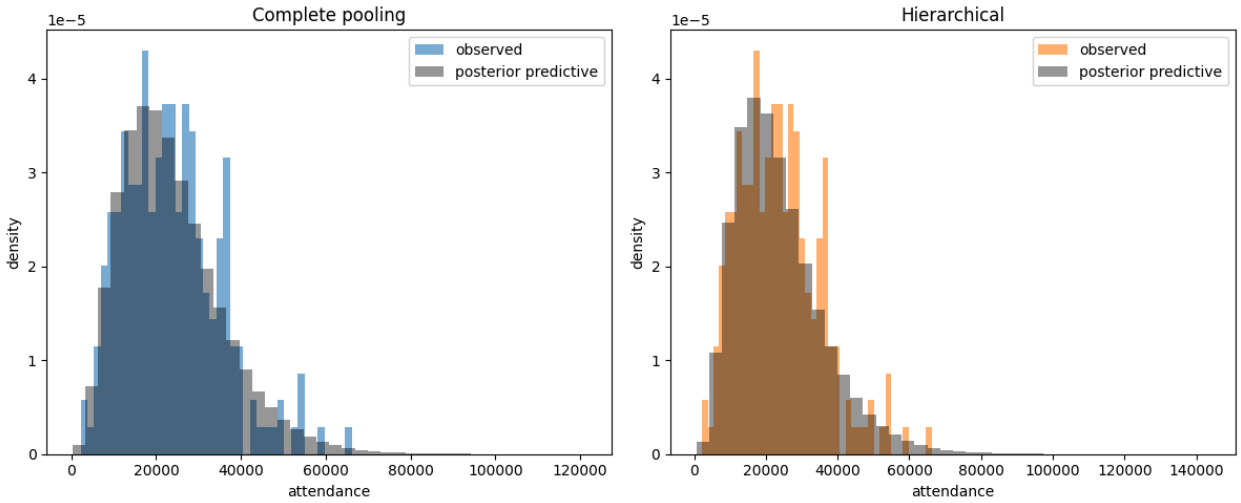



**Figure 19.** Posterior predictive checks. Left: complete pooling spreads attendance around a single global mean of 23,700 fans, capturing the aggregate distribution but missing team structure. Right: hierarchical model matches the same aggregate distribution while making team-specific predictions that align with observed patterns. Both distributions look similar overall, but the hierarchical model's predictions are structurally correct, the right mean for each team, while complete pooling's predictions are statistically wrong for every team.

The win is not dramatic on global fit statistics because both models use Negative Binomial. The Poisson model (that i tried) lost by millions of points due to misspecified variance. Here, both models get the variance right. The hierarchical model wins on structure. It knows Vélez games differ from Argentinos games, and it adjusts predictions accordingly.

My conclusions is that the hierarchical model should be used for all predictions. It outperforms complete pooling by 18 WAIC points (3+ standard errors), captures both team and day effects, and makes structurally sound predictions tailored to each game's context. For the 22 missing games, the hierarchical model is the only reasonable choice. Complete pooling would assign all missing games the same predicted attendance of 23,694 fans, regardless of team or day, which we know from the data is wrong. The hierarchical model accurately predicts Vélez's missing games, which typically occur around 30,000 fans, and Argentinos's missing games, which typically occur around 15,000 fans, reflecting learned patterns from observed data.

## Predictions and Results

Using the hierarchical model, I predicted attendance for 22 missing games. Each prediction includes an 89% credible interval quantifying uncertainty.

Predictions range from 15,061 fans (Argentinos Juniors, Tuesday) to 30,394 (Vélez Sarsfield, Saturday). This spread reflects learned team effects. High-attendance teams (Vélez, Racing, Huracán) average 28,000 to 30,000 fans. Low-attendance teams (Argentinos, Independiente) average 15,000 to 17,000. Mid-tier teams (River Plate, Lanús) fall around 23,000 to 25,000.

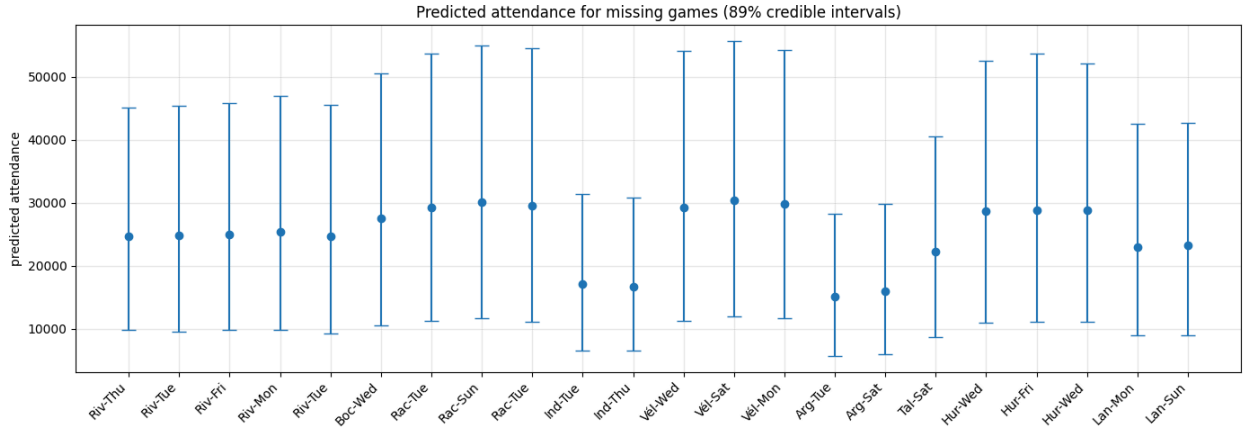



**Figure 20.** Predicted attendance for 22 missing games with 89% credible intervals. High-attendance teams cluster at 28,000 to 30,000 fans with wide intervals (±20,000). Low-attendance teams cluster at 15,000 to 17,000 with narrower intervals (±10,000). Interval width reflects both Negative Binomial dispersion and uncertainty in team/day effects.

Average credible interval width is 36,324 fans, ranging from 22,610 to 43,769. High-attendance teams have wider absolute intervals because their variance is naturally larger, even with similar relative uncertainty.

Predictions strongly correlate with observed team averages ($r = 0.995$). This near-perfect correlation validates that the model correctly learned and applied team-specific patterns.

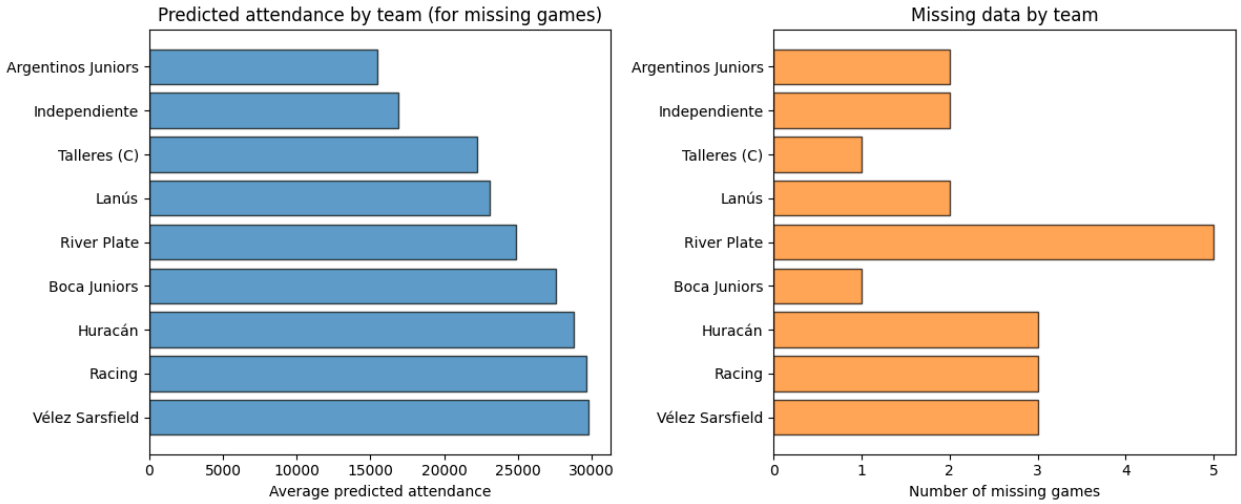



**Figure 21.** Left: Average predicted attendance by team for missing games. Right: Number of missing games per team. River Plate has the most gaps (5 games). Predictions align with team popularity, confirming the model uses learned effects appropriately.

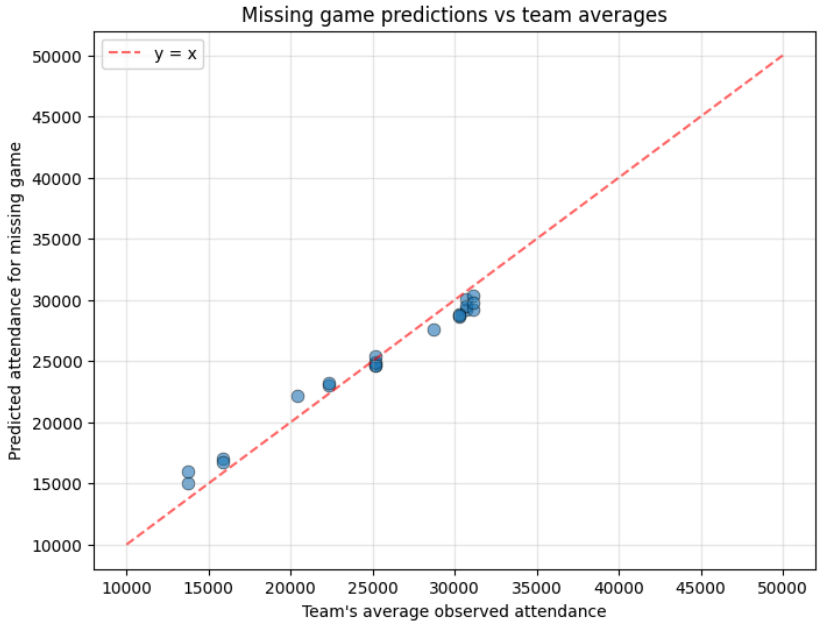



**Figure 22.** Predicted attendance versus team's observed average. The r = 0.995 correlation shows predictions track team patterns. Small deviations reflect day-of-week effects: Saturday/Sunday slightly above average, weekdays slightly below.

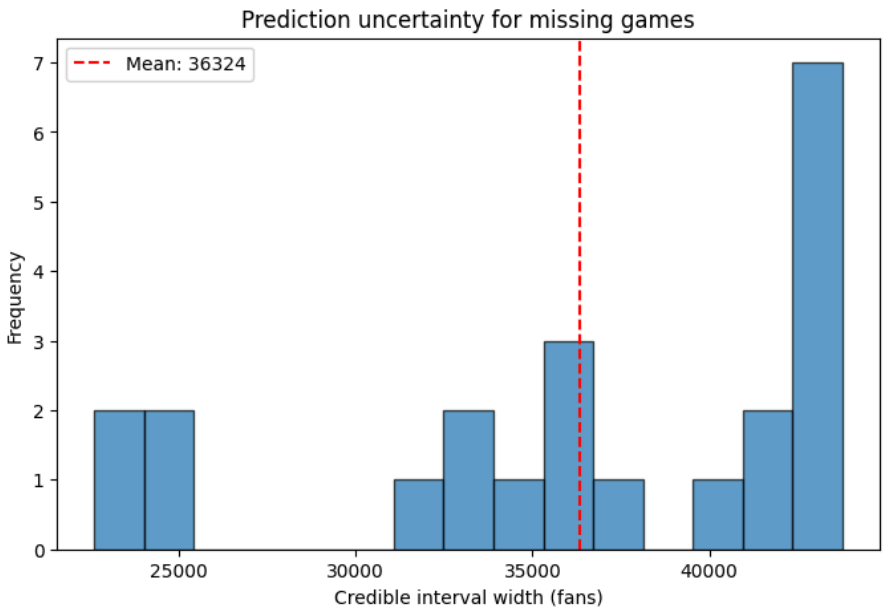



**Figure 23.** Distribution of prediction uncertainty (credible interval width) across the 22 missing games. Most predictions have uncertainty around 35,000 to 38,000 fans, shown by the peak near the mean of 36,324 (red dashed line). A few games have narrower intervals around 23,000 fans (low-attendance teams with tight distributions) and a few have wider intervals exceeding 42,000 (high-attendance teams with more spread). This variability in uncertainty reflects heterogeneity across teams and days, appropriately captured by the hierarchical model.

The patterns make intuitive sense. River Plate's five missing games all predict around 24,000 to 25,000 fans, with slight variation by day (Monday predictions are about 500 fans higher than Thursday). Vélez's three missing games average 29,815 fans. Racing averages 29,606. These align with observed patterns where Vélez, Racing, and Huracán are the top three teams. The model has learned the hierarchy and applies it consistently.

Predictions by day show modest weekend effects. Saturday games average slightly higher predictions than weekday games for the same team, reflecting the learned day effects (Saturday has +0.031 on log scale, Thursday has -0.022). But these differences are minor compared to team variation. A Vélez Saturday game predicts 30,394 fans versus 29,235 on Wednesday, a difference of 1,159 fans (4 percent). In contrast, Vélez versus Argentinos differs by nearly 15,000 fans (100 percent).

The hierarchical structure enables sensible missing data imputation. For River Plate's five missing games, the model borrows strength from River Plate's 15 observed games, learning that River averages around 25,000 fans. It then adjusts slightly based on each missing game's day of the week. This partial pooling avoids both overfitting (treating each game as completely

independent) and underfitting (ignoring team identity). The predictions are informed by team-specific data but regularized toward reasonable ranges.

So all in all, the hierarchical model produces 22 predictions with credible intervals averaging ±36,000 fans. Predictions range from 15,061 to 30,394 fans, closely tracking team popularity. High-attendance teams get high predictions, low-attendance teams get low predictions, and day-of-week effects cause small adjustments. The strong correlation ($r = 0.995$) between predictions and team averages confirms the model learned meaningful patterns and applies them appropriately. These predictions fill the gaps in the league's dataset with statistically principled estimates grounded in observed attendance structure.

**Word Count:** 3745 (excluding figure descriptions; 5072 with figure descriptions)

# AI Statement

I used AI assistance for code debugging and trimming the report. Early on, I ran into a KeyError when indexing team effects in the ArviZ output. I was not correctly mapping the PyMC parameter names to the actual team indices, so I kept getting misaligned team effect estimates. The AI (ChatGPT) helped me figure out that I needed to explicitly loop through the posterior samples and use the team index from the data instead of just pulling the 12th parameter and assuming it was in the right order. It also caught that I was trying to use coordinate labels when I should've been using raw integer positions, which was causing the plotting functions to fail. Another issue was with missing data imputation. I had NaN values in my indexing that were messing up the posterior predictive draws, and the AI suggested I fill those indices with a missing data indicator and handle them in the likelihood separately, which solved the shape mismatch errors.

Beyond debugging, the AI was really helpful for trimming the report. The initial Model 2 section was way too wordy and stuffed with technical details that made it hard to follow. The AI helped cut it down by removing redundant explanations and consolidating related ideas. Same thing with the Summary of Findings. I had it packed with specific numbers and statistical jargon like WAIC scores and variance ratios that would confuse a non-technical client. The AI suggested stripping those out almost entirely and focusing on the main story, which is team popularity matters way more than day of the week.

# References


Minerva University. (2025a). CS146 Session 6 - [3.2] Diagnostics: When things go wrong [Course session]. https://forum.minerva.edu/app/courses/3708/sections/12797/classes/95539

Minerva University. (2025b). CS146 Session 12 - [7.1] Model comparison 3: Practice [Course session]. https://forum.minerva.edu/app/courses/3708/sections/12797/classes/96885

Minerva University. (2025c). CS146 Session 13 - [7.2] Categorical and count data 1 [Course session]. https://forum.minerva.edu/app/courses/3708/sections/12797/classes/97233

Minerva University. (2025d). CS146 Session 14 - [8.1] Categorical and count data 2 [Course session]. https://forum.minerva.edu/app/courses/3708/sections/12797/classes/97235

Minerva University. (2025e). CS146 Session 16 - [9.1] Hierarchical models 1 [Course session]. https://forum.minerva.edu/app/courses/3708/sections/12797/classes/97308

Minerva University. (2025f). CS146 Session 17 - [9.2] Hierarchical models 2 [Course session]. https://forum.minerva.edu/app/courses/3708/sections/12797/classes/97310

# Appendix

## Appendix A: Complete Pooling Model Diagnostics

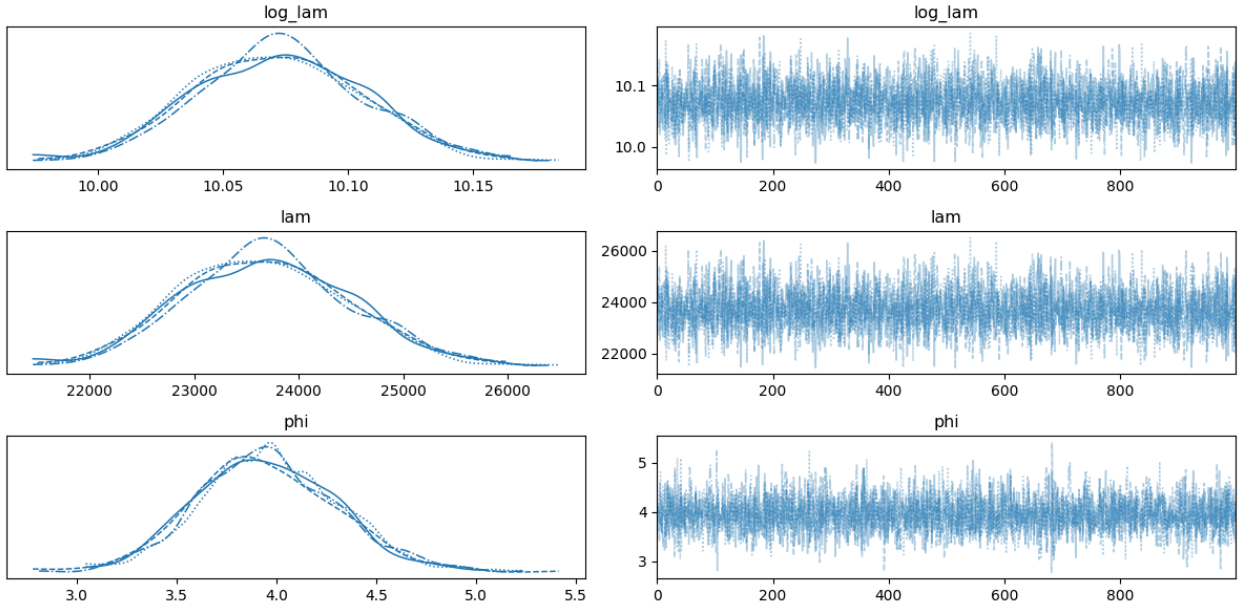



**Figure A1.** Figure A1. Trace plots and marginal posterior densities for the complete pooling Negative Binomial model parameters. The left column shows MCMC chains for log(lambda) (top), lambda (middle), and phi (bottom) across 1,000 post-warmup iterations. All four chains (different colors) mix well and explore the same region of parameter space, indicating convergence. The chains for lambda and phi show healthy variation without trends or sticking, confirming the sampler performed reliably. The right column shows marginal posterior densities pooled across chains. The densities are smooth and unimodal, with all chains contributing equally. Lambda centers tightly around 23,700 fans, while phi centers around 4.0 with slightly more spread. These diagnostics confirm the model converged successfully and posterior uncertainty is well-characterized, even though the model's structural assumptions (ignoring team and day effects) remain problematic.

**Appendix B: Hierarchical Model Technical Details**

B.1 Complete Team and Day Effect Estimates

Table B1 shows the full posterior estimates for all 12 team effects and 7 day effects on the log scale, including posterior means, standard deviations, and 89% HDI bounds.

Table B1. Posterior estimates for team and day effects (log scale).

| **Parameter** | **Mean** | **SD** | **HDI 5.5%** | **HDI 94.5%** | **Multiplier** |
|---|---|---|---|---|---|
| **Team Effects** | | | | | |
| Vélez Sarsfield | 0.242 | 0.089 | 0.103 | 0.387 | 1.27× |
| Racing | 0.233 | 0.089 | 0.092 | 0.375 | 1.26× |
| Huracán | 0.214 | 0.089 | 0.076 | 0.357 | 1.24× |
| Boca Juniors | 0.179 | 0.083 | 0.048 | 0.313 | 1.20× |
| River Plate | 0.068 | 0.094 | -0.080 | 0.219 | 1.07× |
| Rosario Central | 0.029 | 0.083 | -0.101 | 0.162 | 1.03× |
| San Lorenzo | 0.029 | 0.083 | -0.103 | 0.161 | 1.03× |
| Lanús | -0.033 | 0.087 | -0.171 | 0.103 | 0.97× |
| Talleres (C) | -0.101 | 0.083 | -0.234 | 0.031 | 0.90× |

| | | | | | |
|---|---|---|---|---|---|
| Godoy Cruz | -0.161 | 0.083 | -0.293 | -0.029 | 0.85× |
| Independiente | -0.303 | 0.087 | -0.442 | -0.166 | 0.74× |
| Argentinos Juniors | -0.431 | 0.087 | -0.569 | -0.293 | 0.65× |
| Day Effects | | | | | |
| Saturday | 0.031 | 0.048 | -0.044 | 0.106 | 1.03× |
| Sunday | 0.027 | 0.046 | -0.044 | 0.099 | 1.03× |
| Monday | 0.007 | 0.047 | -0.066 | 0.080 | 1.01× |
| Tuesday | -0.012 | 0.044 | -0.080 | 0.056 | 0.99× |
| Wednesday | -0.016 | 0.050 | -0.094 | 0.061 | 0.98× |
| Friday | -0.016 | 0.045 | -0.084 | 0.053 | 0.98× |
| Thursday | -0.022 | 0.045 | -0.091 | 0.047 | 0.98× |

The multiplier column shows $exp(effect)$, representing how attendance scales relative to the baseline. Team effects range from 0.65× (Argentinos Juniors) to 1.27× (Vélez Sarsfield), a span of 0.62 or 95 percent. Day effects range from 0.98× (Thursday) to 1.03× (Saturday), a span of only 0.05 or 5 percent.

B.2 Observed vs Predicted Scatter and Residuals

Figures B1 and B2 provide game-level diagnostic plots showing how well the model predicts individual games.

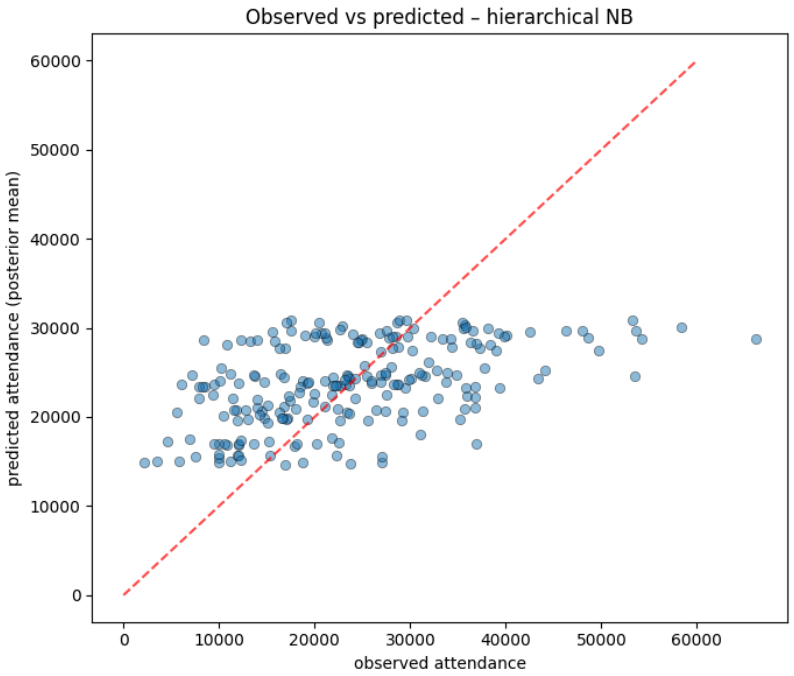



**Figure B1.** Game-level predictions: observed versus predicted attendance. Each of 218 points represents one game. The model achieves reasonable accuracy, with most predictions within ±10,000 fans of the observed value. Outliers (games far from the diagonal) typically correspond to unusually high-attendance matches that exceeded the team's typical pattern due to factors not captured in the model (e.g., playoff implications, rivalry games).

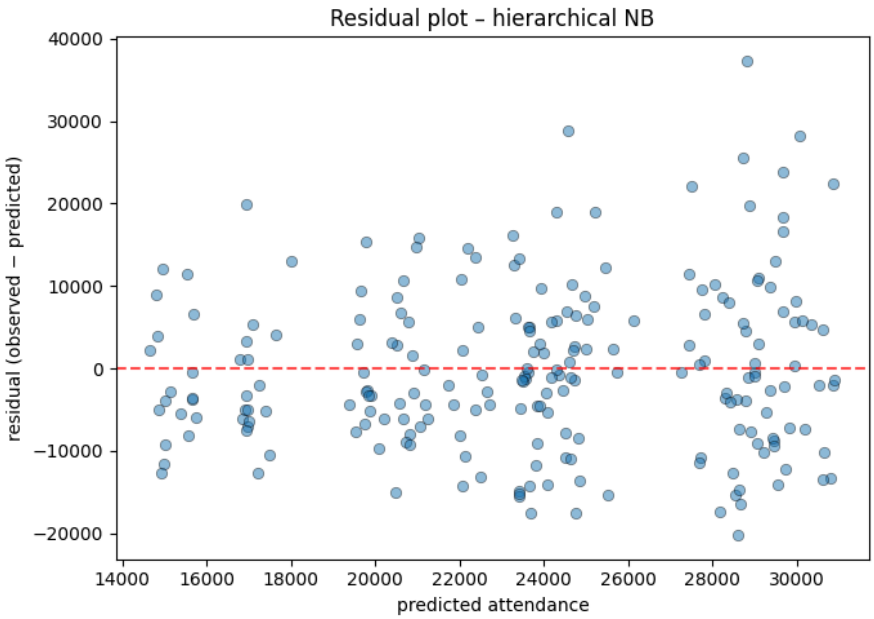
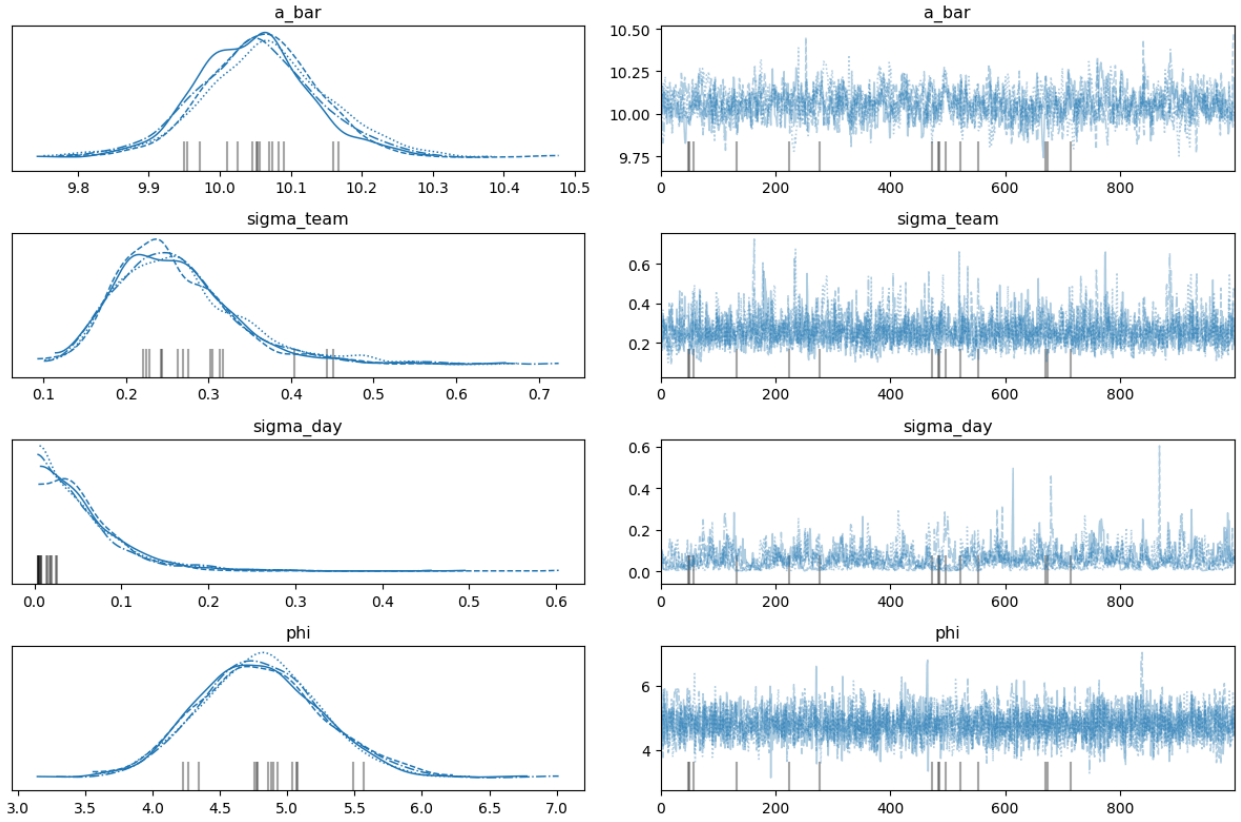



**Figure B2**. Residual plot for the hierarchical model. Residuals scatter around zero with no systematic pattern or funnel shape, confirming the model's assumptions are reasonable. The few

large positive residuals (observed >> predicted) correspond to exceptional games that drew far more fans than the team and day effects alone would predict.

## B.3 Full Trace Plots and Convergence Diagnostics




Full trace plots and posterior densities for hyperparameters (a_bar, sigma_team, sigma_day, phi). All chains converged successfully with $\widehat{R} \leq 1$ and effective sample sizes exceeding 380. The slightly lower ESS for sigma_day (384) reflects this parameter's small magnitude and high posterior correlation with the baseline, but the value is still sufficient for reliable inference.

## B.4 Team and Day Predictions by Group

Figures B4 and B5 show mean predictions aggregated by team and day.

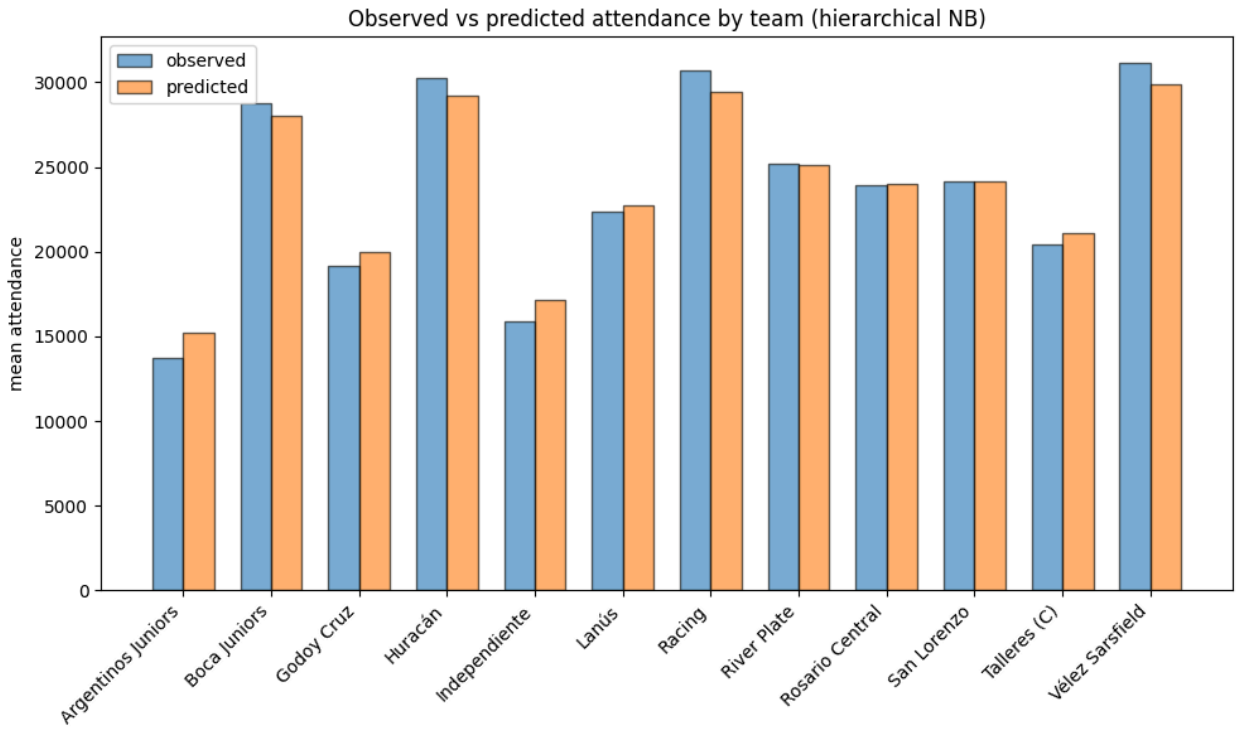



**Figure B4.** Mean attendance by team: observed (blue) versus predicted (orange). The close match validates that the model learned team-specific effects accurately. Slight discrepancies (e.g., Vélez Sarsfield predicted at 29,884 vs observed 31,135) arise from partial pooling, which intentionally shrinks extreme estimates toward the global mean to improve out-of-sample generalization.

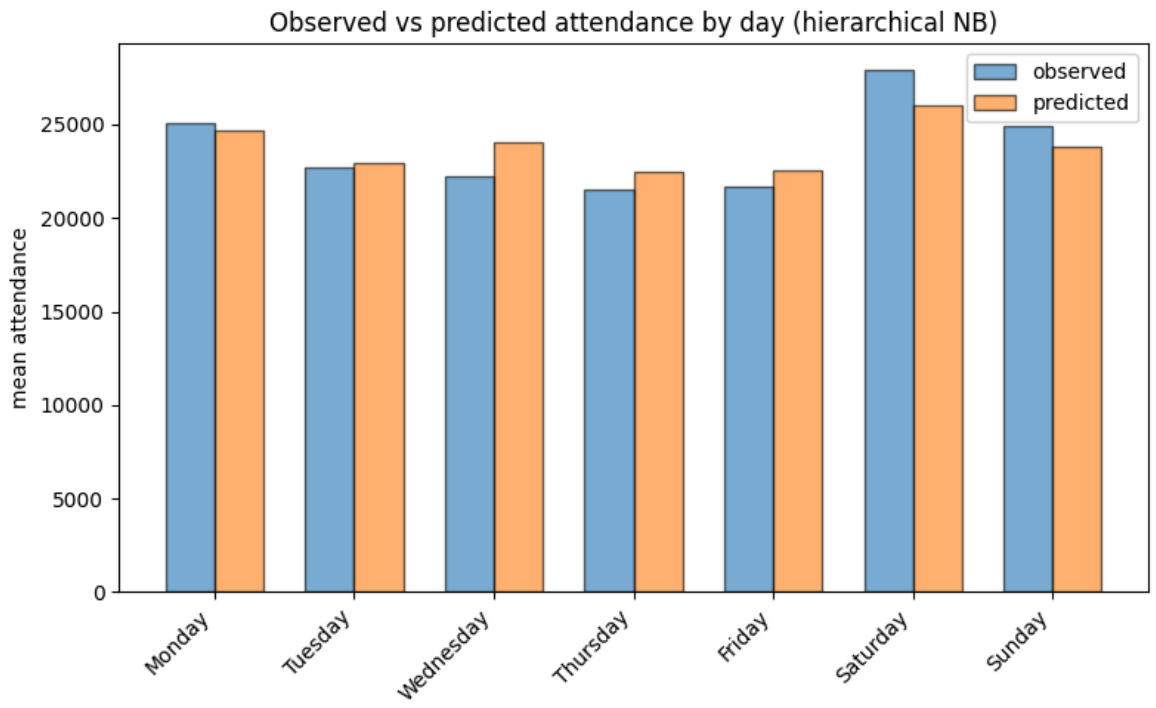



**Figure B5.** Mean attendance by day: observed (blue) versus predicted (orange). Day-level predictions show more variability because day effects are weaker ($\sigma_{day} = 0.06$) and sample sizes per day are smaller. The model appropriately assigns less precision to day distinctions, avoiding overfitting to noise.