Project 2: Discrete and multi-level models

Minerva University

CS146 - Computational Methods for Bayesian Statistics

Prof. Scheffler

November 29, 2025

Project 2: Discrete and multi-level models

Data pre-processing

The dataset contains 240 home games, one row per game, with columns for team, day of week, and scanned attendance. Attendance is observed for 218 games, and 22 games (9.2%) are missing this value.

For modeling in PyMC, I encode team and day as categorical variables with integer indices (team_idx from 0–11 and day_idx from 0–6, ordered Monday–Sunday).

Missingness patterns

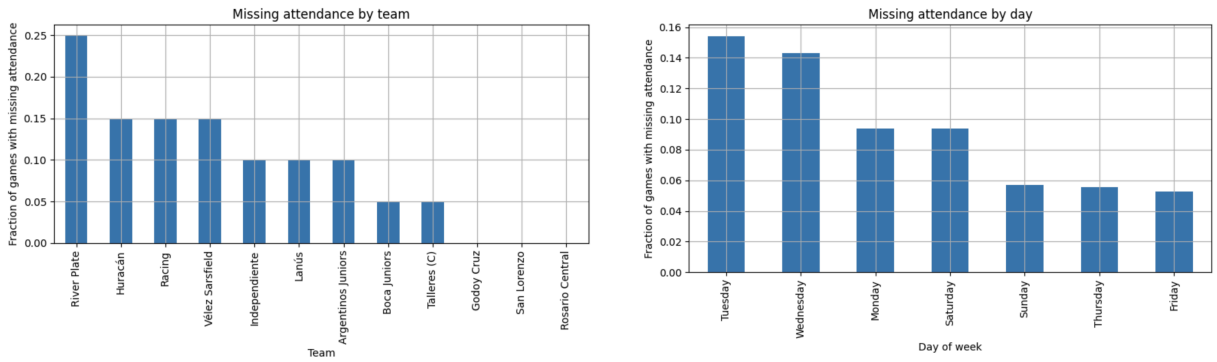

Figure 1. *Missingness patterns by team (left) and day of week (right).*

Missing values are not uniformly distributed. River Plate has missing values for 25% of its games; Huracán, Racing, and Vélez Sarsfield for 15%; and several other clubs around 5–10%, while Godoy Cruz, San Lorenzo, and Rosario Central have none.

By day, Tuesday (15.4%) and Wednesday (14.3%) have the highest missingness, with other days between about 5–9%.

Since missingness is spread across several teams and days rather than concentrated in a single extreme subgroup, I proceed under the assumption that attendance is missing at random conditional on team and day, and I let the model impute these values.

Distribution of observed attendance

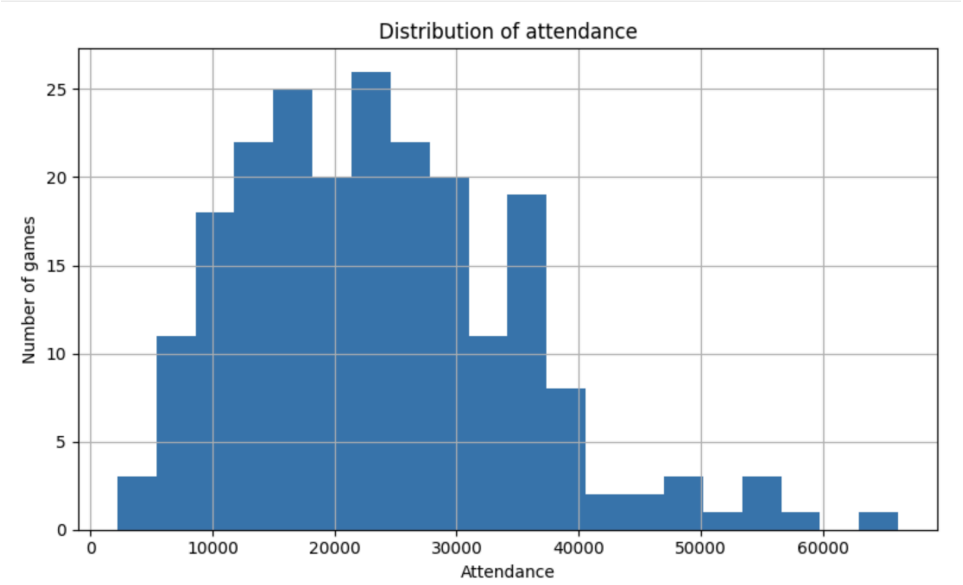

Figure 2. *Distribution of attendance for all observed games.*

Observed attendances range from 2,205 to 66,152, with a mean of about 23,650 and a median of 22,636. The overall histogram is concentrated between roughly 10,000 and 35,000 fans, with a right tail of very well-attended games.

Team-level patterns

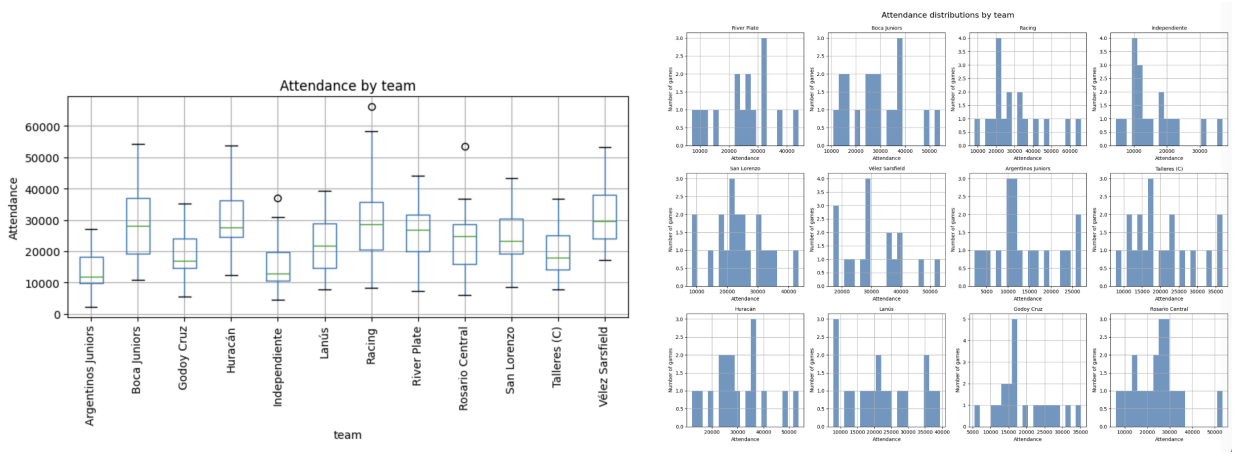

Figure 3. *Boxplots showing the spread of attendance for each team (left) and multiple*

attendance distribution histograms for individual teams (right).

High-attendance clubs such as Racing, Boca Juniors, Vélez Sarsfield, Huracán, and River Plate consistently draw the largest crowds. Their boxplots show high medians (25k–35k), and upper whiskers reach 50k+ (and in Racing's case, even ~65k). As we can see from histograms, the games are clustered around the 30k-40k region.

Mid-tier teams (Lanús, Rosario Central, San Lorenzo, and Godoy Cruz) center around the global median (~22k–27k), with moderate variance. Histograms display more uniformly spread mid-sized crowds.

Lower-attendance teams have noticeably lower attendance, with medians closer to 12k–18k and narrower distributions. Their histograms show thinner right tails and fewer high-attendance outliers

These patterns show that team-level attendance is both large in magnitude and strongly differentiated across clubs.

Day of week patterns

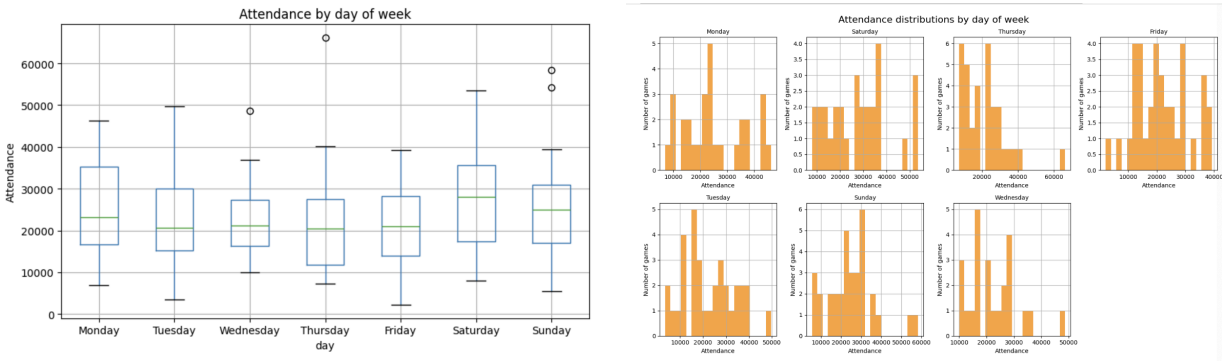

Figure 4. *Boxplots of attendance grouped by day (left) and multiple attendance distribution histograms for each day of the week (right).*

Attendance across days of the week displays much less variation. Medians for Monday

through Sunday fall in a compact band around 21k–27k. Variability increases slightly on weekends (Saturday, Sunday), where big matches are more likely to be scheduled. Midweek games (Wednesday, Thursday) have tighter distributions and slightly lower medians.

Although weekends show higher and more variable attendance, the magnitude of these differences is small relative to the team patterns.

Model 1: Complete pooling

I first fit a complete-pooling model that assumes all games in the league share the same expected attendance. Attendance data consists of non-negative integer counts, so continuous distributions like Normal are not appropriate here. Based on the EDA, the data is highly overdispersed, with an empirical variance ($\approx 128\,million$) much greater than the mean (≈ 23650). Since the Poisson distribution assumes $variance = mean$, it is not suitable either. Instead, I use the Negative Binomial distribution, which is usually a good choice for overdispersed counts. The variance is $var = \mu + \frac{\mu^2}{\alpha}$, so the model can adaptively increase variance above the mean when α is small. Also, the variance-mean relationship of the Negative Binomial allows overdispersion without forcing unrealistic predictions at the left tail (which happens with log-normal).

The model is:

$$log\ \mu \sim Normal(log\ (27000),\ 0.3^2)$$

$$\alpha \sim Gamma(2,\ 0.1)$$

$$y_i \sim NegativeBinomial(\mu, \alpha)$$

μ is the shared average attendance on the original scale and α is an overdispersion

parameter.

The Normal prior on the log scale keeps μ strictly positive. I centered it at $log\ (27000)$ after looking up independent statistics on average attendances of football matches in Argentina. According to multiple sources, the league drew around 27,600 spectators per match in 2023. The prior standard deviation of 0.3 on the log scale corresponds to a range of roughly 15k–50k on the original scale, which is tight enough to avoid unrealistic predictions, but wide enough not to dominate the data.

The dispersion parameter (α) is extremely sensitive in Negative Binomial models. When I experimented with Exponential priors, they produced extremely small values, which inflated the variance and generated prior predictive draws with unrealistic attendance in the hundreds of thousands or even millions.

To fix this, I switched to Gamma(2, 0.1), which is a weakly informative prior. It keeps α positive and allows dispersion, but prevents extremely small values that produce huge attendances.

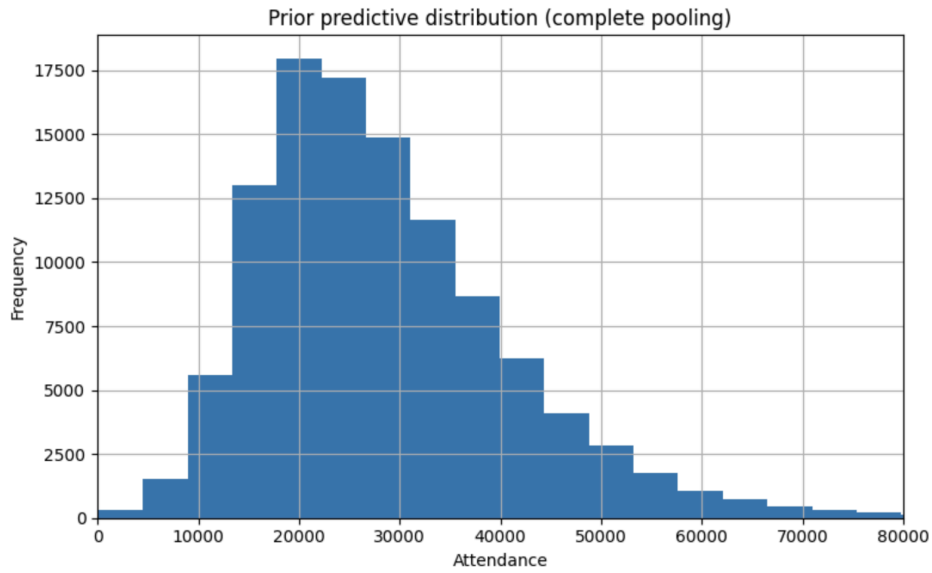

Figure 5. *The distribution of simulated attendance data generated from the priors before seeing the data.*

The prior predictive plot shows a unimodal distribution concentrated around 20–35k, with a right tail extending to around 60–70k. I think this is reasonable because the maximum capacity of the stadiums is usually 80k, and empty stadiums with fewer than 5k attendees are very unlikely.

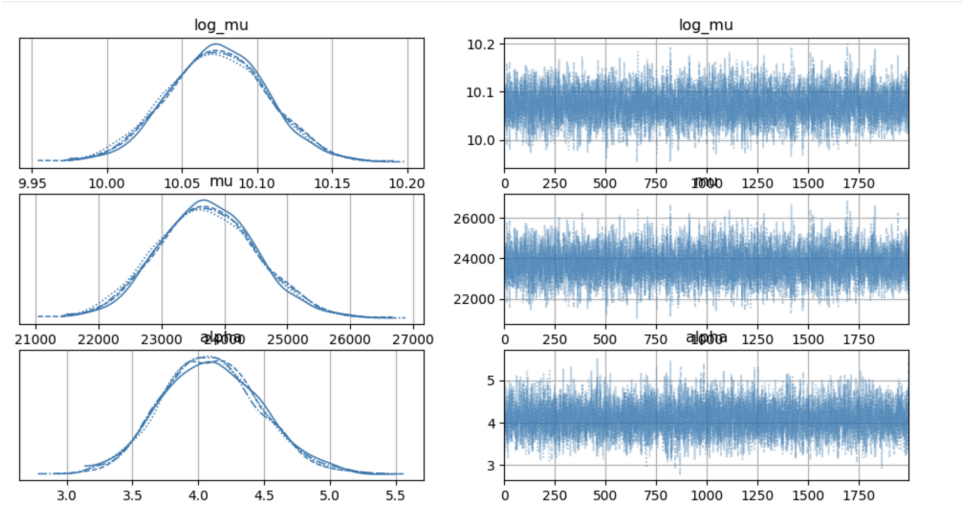

Figure 6. *MCMC trace plots (complete pooling model) for the parameters log_mu and alpha. The fuzzy caterpillar patterns indicate well-mixed chains with no divergences, suggesting the NUTS sampler converged successfully.*

After fitting this model to the 218 games with observed attendance, all chains mixed properly (no divergences, r-hat=1.0, ess are high). A posterior for μ is tightly concentrated around the empirical mean. The posterior for α is centered around ~4-5, meaning strong overdispersion, which is consistent with the heavy right tail observed in EDA.

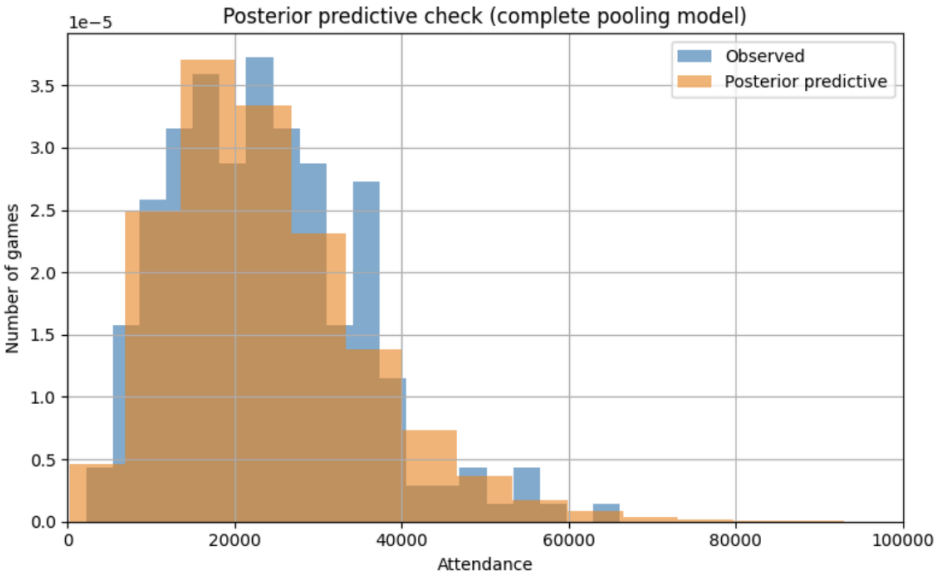

Figure 7. *Posterior-predictive check (complete pooling model). Comparison of observed data (blue) vs. model-generated predictions (orange).*

I then compared the samples from the posterior predictive distribution to the observed attendance distribution. The complete-pooling model captures the overall center, spread, and skewness of the data reasonably well. However, because this model forces all games to share the same μ, it doesn't account for team-level and day-of-week differences, and might overestimate or underestimate attendance for some games.

Model 2: Hierarchical model (partial pooling)

In the second model, I extended the complete-pooling baseline by introducing a hierarchical structure that accounts for systematic variation across teams and days of the week. Instead of assuming that all games share the same mean attendance, this model allows each team and each day to have its own deviation from the global mean. These deviations are themselves modeled as draws from group-level distributions, which leads to partial pooling.

To model this structured heterogeneity while still preventing extreme overfitting, I fit a hierarchical Negative Binomial model.

The hierarchical model assumes that match attendance for team i on day j has a team-specific and day-specific effect added to the overall log-mean:

$$log\_\mu_{ij} = log\_\mu_{global} + u_{team[i]} + v_{day[j]}$$

$u_{team\,[i]} - deviation\ for\ team\ i$

$v_{day\,[j]} - deviation\ for\ day\ of\ week\ j$

$log\_\mu_{global} - global\ average\ attendance$

Global mean and dispersion prior are the same as before:

$$log\_\mu_{global} \sim Normal(log(27000),\ 0.3)$$

$$\alpha \sim Gamma(2,\ 0.1)$$

Initially, I used centered parametrization:

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

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

However, it produced a lot of divergences, even with high target_accept settings. This is common in hierarchical models because when the group-level scale parameters are small or uncertain, the posterior geometry becomes funnel-shaped. The sampler struggles near the narrow part of the funnel, causing divergent transitions and biased estimates.

After confirming that the divergences can't be solved by tuning, I rewrote the model using the non-centered parametrization by introducing standard-normal variables:

$$z_{team[i]} \sim Normal(0, 1)$$

$$u_{team[i]} = z_{team[i]} \cdot \sigma_{team}$$

$$z_{day[j]} \sim Normal(0, 1)$$

$$v_{day[j]} = z_{day[j]} \cdot \sigma_{day}$$

Reparametrizing in this way stretches out the geometry and removes the narrow funnel throat.

The standard deviations are given weakly informative priors:

$$\sigma_{team} \sim HalfNormal(0.2)$$

$$\sigma_{day} \sim HalfNormal(0.2)$$

0.2 on the log scale corresponds to ≈ ±20% variation on the original scale, which is small enough to prevent unrealistic differences and large enough to allow variation.

After changing to the non-centered parametrization, all divergences disappeared.

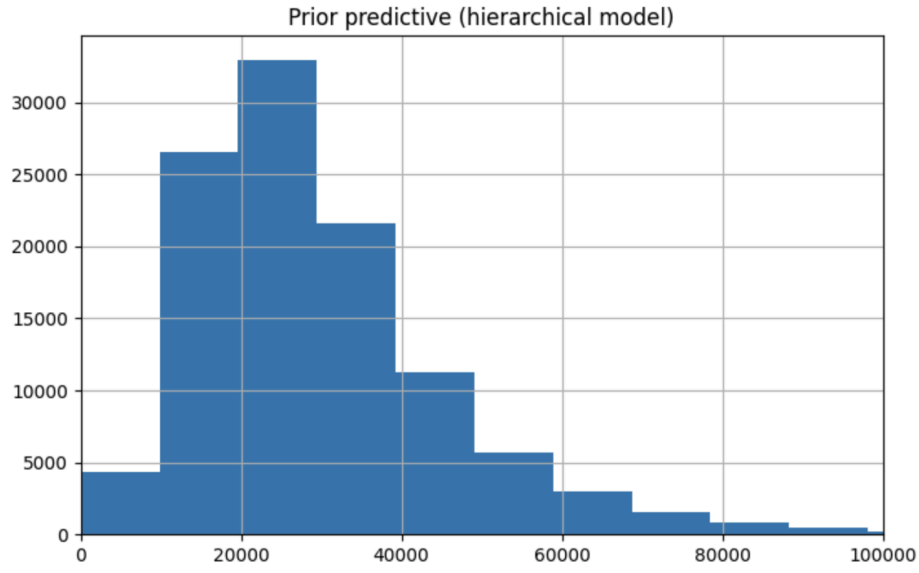

Figure 8. *Prior-predictive distribution of simulated attendance using the hierarchical structure.*

The prior predictive distribution for this model is centered within the expected range

(roughly 10,000–60,000+). It is broader than the complete pooling prior because we allowed variation across teams and days, which naturally widens the set of plausible attendances.

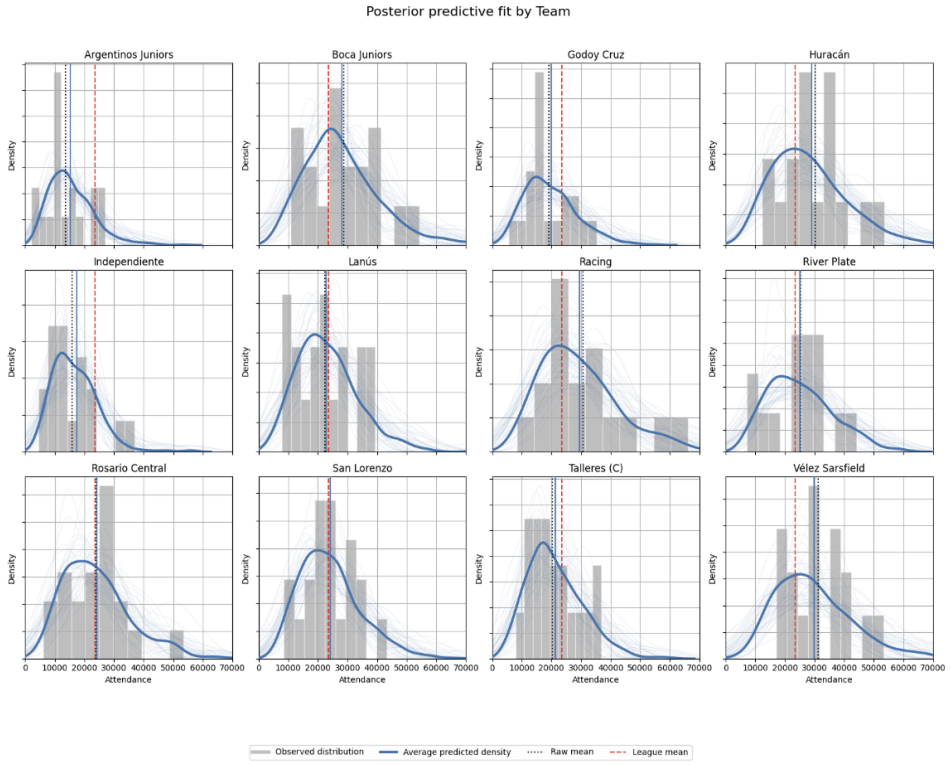

Figure 9. *Posterior-predictive fit by team. Density plots of observed attendance (bars) overlaid with the posterior predictive density (blue line) for each team.*

The posterior predictive checks by team and by day show that the hierarchical model captures the structure of the data well. High-attendance clubs have predicted densities centered around 35k-50k, whereas for low-attendance clubs, it is around 10k-20k.

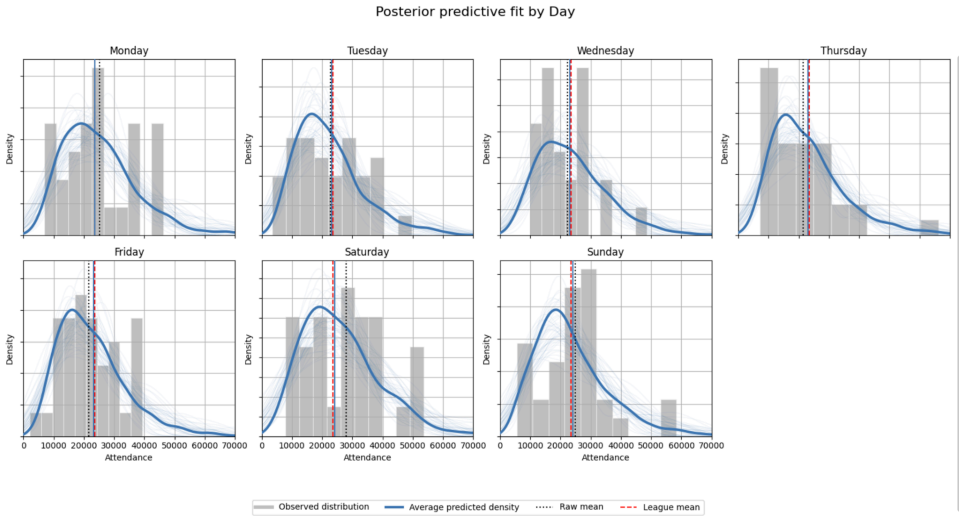

Figure 10. *Posterior-predictive fit by day. Density plots of observed attendance (bars) overlaid with the posterior predictive density (blue line) for each day of the week.*

For days of the week, the posterior predictive curves overlap across Monday through Sunday, and the raw means cluster tightly around the global mean. Day-of-week effects are present but weak relative to team effects.

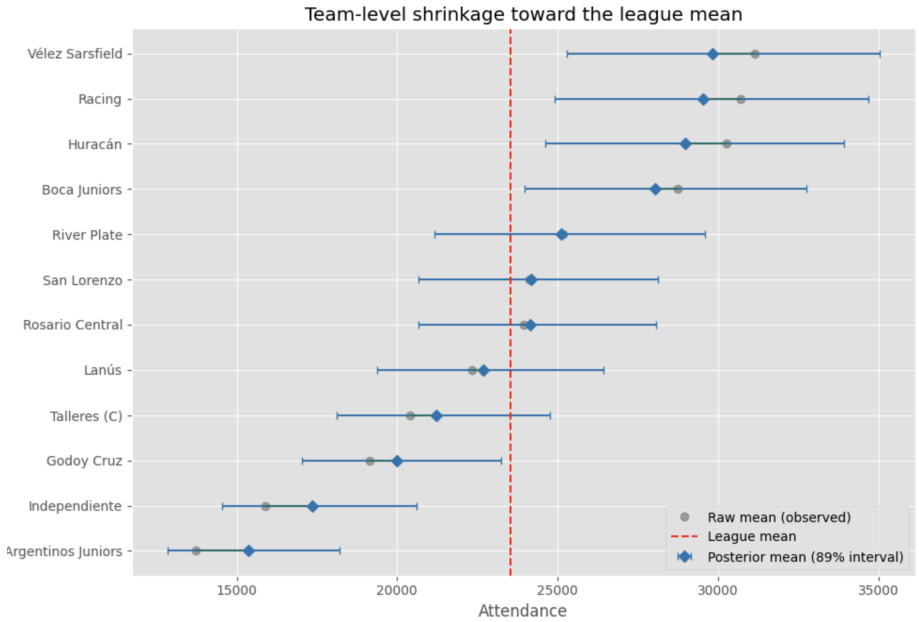

Figure 11. *Forest plot comparing raw observed means (gray dots) with posterior means (blue dots) and the global league mean (red dashed line).*

The forest plot for team-level shrinkage visualizes partial pooling most clearly. For each team, the raw mean (gray dot) is contrasted with the posterior mean, 89% HDI (blue point + interval), and the global mean (red dashed line). Boca Juniors, River Plate, San Lorenzo show very little shrinkage, their raw means and posterior means are nearly identical because the data provide strong evidence about their true underlying averages.

In contrast, teams with extreme means, like Argentinos Juniors and Vélez Sarsfield, show much stronger shrinkage. The model responds by pulling its posterior mean upward, closer to the global mean, possibly meaning we have limited information to justify an extreme deviation.

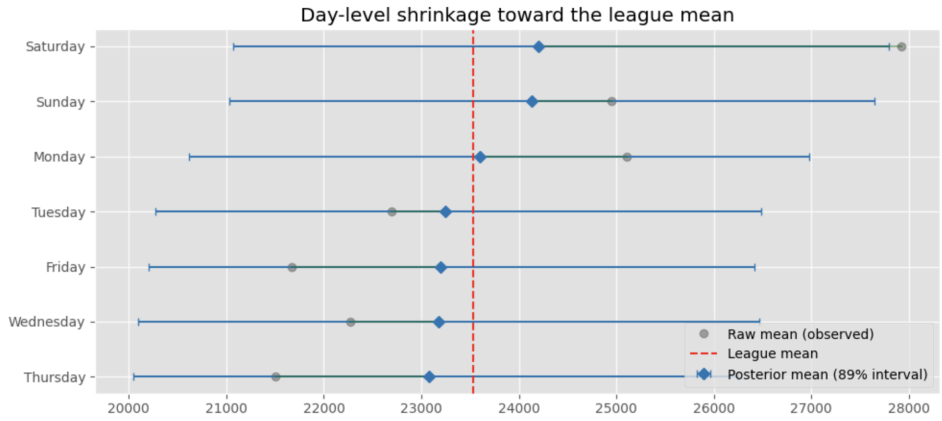

Figure 12. *Forest plot for day-of-week effects. Due to the weak signal in the day-level data, the model applies strong shrinkage.*

Day-of-week effects are weak, and the model expresses this by pulling almost all day-level posterior means tightly toward the global mean. The credible intervals are also wider because the model learns relatively little structure from day-to-day variation.

The hierarchical model successfully addresses the limitations of the complete‑pooling

model. It represents the heterogeneity in attendance across clubs, regularizes noisy estimates through partial pooling, and retains enough flexibility to adapt to patterns in the data without overfitting.

Model Comparison

To formally evaluate predictive performance, I compared the two models using Pareto-smoothed leave-one-out cross-validation (LOO). LOO evaluates how well a model predicts each observed data point when that point is removed from the dataset. Unlike metrics based solely on in-sample fit, LOO explicitly penalizes models that overfit and rewards those that generalize well to new, unseen games.

	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
hierarchical	0	−2316.595396	12.179417	0.000000	1.0	10.330620	0.00000	False	log
complete_pooling	1	−2335.062519	2.008372	18.467123	0.0	10.465103	5.54858	False	log

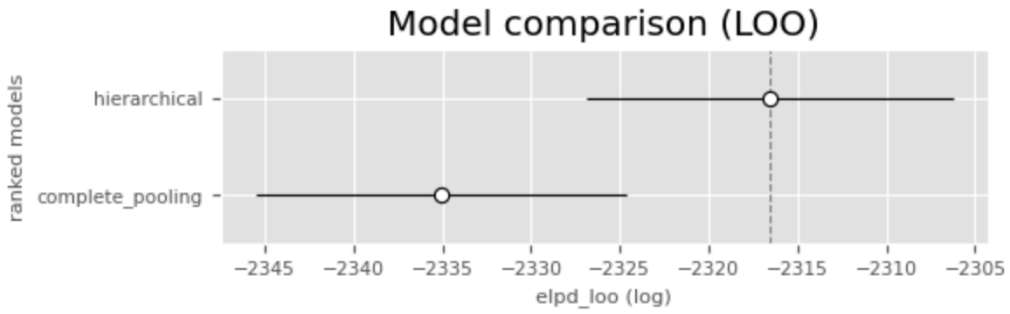

Figure 13. *Model comparison (LOO-CV) of expected log predictive density (elpd) using Leave-One-Out Cross-Validation.*

As we can see from Fig. 13, the results clearly favored the hierarchical model. The hierarchical model achieved an elpd_loo of –2316.6, compared to –2335.1 for the

complete-pooling model. Higher (less negative) elpd_loo values indicate better predictive accuracy, and the difference is about 18.5 units on the log scale.

The LOO comparison also provides model weights that summarize how much support each model receives relative to the others. Using stacking weights, the hierarchical model receives essentially all the model weight (1.0 = 100%), while the complete-pooling model receives 0.

The fact that the hierarchical model performs better despite introducing additional parameters suggests (1) the data contain meaningful structure across teams and days of the week, and (2) partial pooling regularizes these effects enough to avoid overfitting.

Overall, the LOO comparison strongly favors the hierarchical model. It provides more stable parameter estimates and significantly better predictive performance on new games.

Predictions and Interpretation

Based on model comparison, I will use a hierarchical model to generate predictions for missing data. Since 22 of the 240 games in the dataset don't report attendance, recovering these values is essential for understanding overall league patterns and avoiding bias toward teams with more complete data.

	team	day	pred_mean	hdi_lower	hdi_upper
1	River Plate	Thursday	24653	7630	40911
2	River Plate	Tuesday	24795	7384	40788
3	River Plate	Friday	24988	7571	41737
4	River Plate	Monday	25214	7642	41876

5	River Plate	Tuesday	24789	7473	40937
6	Boca Juniors	Wednesday	27603	7913	45716
7	Racing	Tuesday	29140	8944	48350
8	Racing	Sunday	30026	8934	49273
9	Racing	Tuesday	29163	8313	47963
10	Independiente	Tuesday	17101	5252	28110
11	Independiente	Thursday	16973	4623	27978
12	Vélez Sarsfield	Wednesday	29289	8737	48455
13	Vélez Sarsfield	Saturday	30531	8566	49696
14	Vélez Sarsfield	Monday	29862	8271	48472
15	Argentinos Juniors	Tuesday	15179	4821	25632
16	Argentinos Juniors	Saturday	15786	4535	25692
17	Talleres (C)	Saturday	21818	6530	36201
18	Huracán	Wednesday	28611	9113	47636
19	Huracán	Friday	28491	8246	46996
20	Huracán	Wednesday	28674	8386	47400
21	Lanús	Monday	22868	6895	37635
22	Lanús	Sunday	23398	6923	38329

Table 1. *Predicted attendance for missing games. Table listing the posterior mean and 89% High Density Interval (HDI) for the 22 games with missing attendance data.*

Using the posterior draws from the hierarchical model, I generated predictions (Table 1) for each of the missing games by plugging in the team and day-of-week indices for those matches. For each missing game, I computed the posterior mean attendance along with an

associated 89% High Density Interval (HDI). These predictions combine the global mean, the team's estimated effect, and the day effect, and then account for estimation uncertainty through the use of the Negative Binomial likelihood.

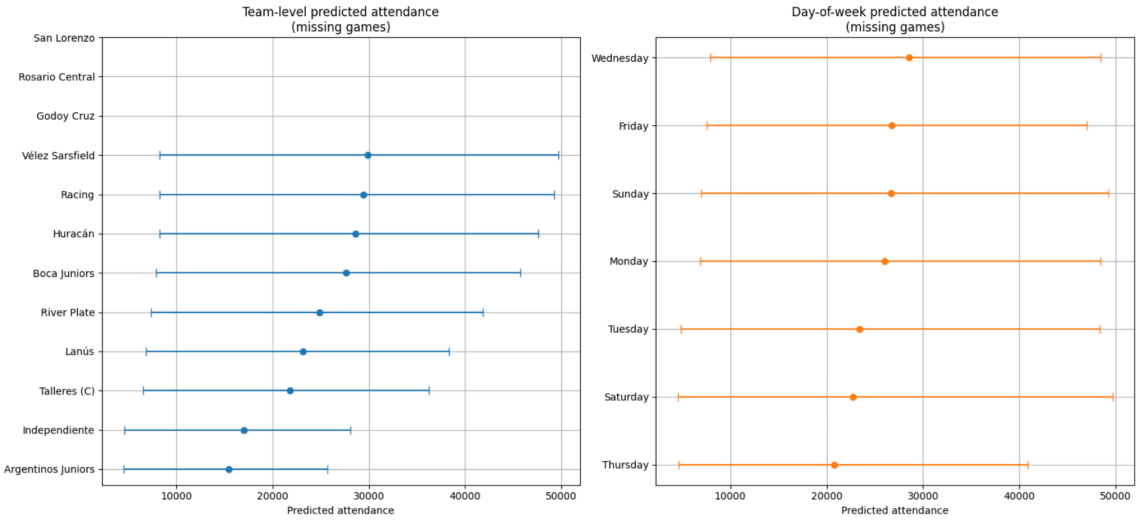

Figure 14. *Summary of predicted attendance for missing games grouped by team (left) and day (right). The team plot shows distinct stratification, whereas the day plot shows overlapping intervals, confirming that team identity is the primary driver of attendance counts.*

To interpret these predictions more systematically, I aggregated them at the team and day-of-week levels and visualized the results (Fig.14). Now, we can not only see which teams or days tend to have higher predicted attendance, but also which factor contributes more to uncertainty.

From the left panel, we can see that teams such as Vélez Sarsfield, Racing, Boca Juniors, and Huracán consistently show the high predicted attendance (≈29,000–31,000). Argentinos Juniors, Independiente, and Talleres (C) fall on the lower end, with predictions as low as 15,000–22,000. These differences align closely with the group-level effects estimated in the model and reflect meaningful underlying heterogeneity between clubs. Also some teams have

wider HDIs, while other teams have narrower uncertainty bands, which represent the team-level variation effects estimated earlier in the model.

Unlike teams, the day effects show considerably less variation. All predicted day-of-week means lie within a narrow range around 25,000–30,000 attendees. The HDIs for different days overlap almost completely.

In conclusion, team explains far more variation in predicted attendance than day of the week. The imputed values are plausible, coherent with the observed data, and appropriately uncertain. Without partial pooling, teams with missing data would be either overconfidently assigned their raw means or severely overfitted based on too little information. The hierarchical model instead provides balanced, data-driven predictions that respect both within-team information and league-wide trends.

Summary of findings (for non-technical audience)

The whole point of this project was to figure out what actually drives match attendance in the Argentine Primera División. In simple terms, why do some matches pull 50,000 fans while others barely hit 15,000?

After analyzing the data and fitting a model that accounts for differences between teams and days of the week, I came to the conclusion that the attendance depends almost entirely on the team. Big clubs like Boca, River, Racing, San Lorenzo consistently draw massive crowds. Smaller clubs don’t come close, and this gap is way larger than anything related to scheduling.

The day of the week does matter a little, but not enough to change the story, the range is tiny compared to the difference between clubs. If Boca plays on a Tuesday, they still fill the

stadium. If Argentinos Juniors plays on a Sunday, they still won't suddenly pull 40,000 people. The "team effect" basically dominates everything else.

Since about 10% of the games in the dataset were missing attendance numbers, I used the model to fill them in. The model does this by combining what we know about the team, the day of the week, and what typical league-wide attendance looks like. For example, missing Boca or Racing games naturally get higher predicted values, while missing games for smaller clubs land much lower. And instead of giving one exact number, the model gives a reasonable range.

So Argentine match attendance is driven mainly by club size and fanbase, not by the day of the week. The hierarchical model captures that extremely well and gives realistic predictions for the missing games.