Final Paper

Shawn Malik [*1] **Huyen Huynh** [*1] **Wenwan Xu** [*1]

Abstract

The National Basketball Association (NBA) has grown into one of the most valuable sports leagues in the world, with average franchise valuations of roughly $4.66 billion in 2025 and continued growth expected in the coming years. This rapid appreciation raises a natural question about how the labor market for players works in practice: which on-court performance metrics best predict player salaries. In this project, we addressed this question using publicly available season-level statistics and contract data from Basketball-Reference for the 2024–25 season and the corresponding 2025–26 base salaries, aligning performance in one season with pay in the next.

We constructed a cleaned, player-level dataset by consolidating multi-team seasons, merging statistics with contract data using Basketball-Reference identifiers, and transforming salary into a log scale to handle the strong right skew in pay. The predictors included playing-time measures (`G`, `MP`), efficiency and usage metrics (`TS%`, `USG%`), role-based rate statistics (`AST%`, `TRB%`, `STL%`, `BLK%`, `TOV%`), composite impact metrics (`BPM`, `WS`, `VORP`), and demographic variables such as `Age` and position.

We estimated four models: Ordinary Least Squares (OLS), Ridge regression, Lasso regression, and a Random Forest regressor. We evaluated performance using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2.

Across all models, the Lasso regression achieved the best overall predictive accuracy among the linear approaches, with the lowest test RMSE (0.599) and the highest R^2 (0.596), while the Random Forest achieved the lowest MAE but was less accurate for extreme high-salary outliers. Taken together, the models consistently highlighted Usage Percentage (`USG%`), Age (`Age`), Minutes Played (`MP`), and Box Plus/Minus (`BPM`) as the strongest positive predictors of log-salary. Our results suggest that teams primarily compensate players based on how much they play, how central they are to offensive creation, and their overall impact on team performance, rather than narrower rate statistics or positional labels.

[*]Equal contribution [1]University of Virginia, Charlottesville, Virginia, USA.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Data

Sources and Scope

All data, reports, and code for this project are available at:

https://github.com/shawnmalik1/DS3001_F25_Project

Our main research question is: *Which performance metrics best predict NBA player salaries.*

The project combines two season-level player files:

- **Player Performance** (`player_stats.csv`): per-player statistics for the 2024–25 NBA season from Basketball-Reference, including counting metrics (e.g., `G`, `MP`), efficiency and usage rates (e.g., `TS%`, `3PAr`, `FTr`, `USG%`), play-type rate stats (e.g., `AST%`, `TRB%`, `STL%`, `BLK%`, `TOV%`), and all-in metrics (e.g., `PER`, `WS`, `WS/48`, `OBPM`, `DBPM`, `BPM`, `VORP`), plus demographics and role information (`Age`, `Pos`, `Team`).
- **Player Salary** (`player_salaries.csv`): per-player base-salary information from Basketball-Reference for future seasons (e.g., 2025–26, 2026–27) and total guaranteed money.

Both files are sourced from Basketball-Reference, which keeps naming conventions and IDs consistent across statistics and contracts. The modeling target is *next season's* base salary (2025–26), and predictors are taken from the immediately preceding season's on-court performance, aligning the temporal sequence of performance leading to pay.

Unit of Analysis and Coverage

Basketball-Reference reports one row per team stint and, for traded players, sometimes an aggregated multi-team line

(e.g. "2TM" or "3TM"). To keep the unit of analysis consistent, we defined one row per player-season. For players with multiple team stints, we selected the single stint with the most minutes (MP) when a consolidated line was not already provided. This produced a representative stat line for each player and avoided double-counting. In all cases, we used the team associated with that primary stint.

After this filtering, we merged the stats and salary files and retained only players with both a prior-season stat line and a 2025–26 base salary. Some players were dropped because they lacked a next-season salary (for example free agents, two-way or 10-day contracts) or did not log meaningful minutes in the prior season.

Identifiers and Merge Strategy

When available, we used the Basketball-Reference player ID (e.g., lillada01) to create a common key bbref_id and performed an inner join. For the small number of rows missing IDs, we fell back to a name-based merge after light name cleaning, including standardizing suffixes such as "Jr." and "III". The team variable that appears in both datasets (Team in player_stats.csv and Tm in player_salaries.csv) was also used to double-check the merging process after renaming them to match. Using IDs whenever possible minimized mismatches from suffixes, abbreviations, and nicknames.

Key Variables

Base salary 2025–26 (Y) was parsed from currency strings into numeric values in USD. The raw salary distribution was strongly right-skewed, with many players near the league minimum and a long tail of star and supermax contracts.

Candidate predictors (X)

- **Availability and playing time:** G (Games), MP (Minutes Played). Minutes both signal value and help stabilize other statistics.
- **Efficiency and usage:** TS% (True Shooting Percentage), 3PAr (Three-Point Attempt Rate), FTr (Free Throw Attempt Rate), USG% (Usage Percentage).
- **Role and rate indicators:** AST% (Assist Percentage), TRB% (Total Rebound Percentage, with ORB% and DRB% as components), STL% (Steal Percentage), BLK% (Block Percentage), TOV% (Turnover Percentage).
- **Overall impact metrics:** PER (Player Efficiency Rating), WS (Win Shares), WS/48 (Win Shares per 48 Minutes), OBPM (Offensive Box Plus/Minus), DBPM (Defensive Box Plus/Minus), BPM (Box Plus/Minus), VORP (Value over Replacement Player).
- **Demographics and role:** Age, Pos (Position). Team indicators were available for extended models but did not play a central role in the final specification.

Reading, Cleaning, and Preparation

We structured our data preparation in several steps:

1. **Read data from a single source.** We sourced both statistics and contracts from Basketball-Reference (Basketball-Reference, 2025b;a). Reading them as csv files allowed straightforward integration and compatibility with subsequent processing in Python.
2. **Consolidate multi-team seasons.** We retained the consolidated "xTM" line when available; otherwise, we kept the highest-MP stint. This produced exactly one prior-season stat line per player and prevented double-counting.
3. **Parse money fields.** We stripped "$" and commas from salary strings and coerced them to numeric types. We used 2025–26 base salary as the target; guaranteed totals span multiple years and options and were not modeled directly.
4. **Handle missing data.** We dropped rows missing the target salary or essential stats such as MP and Age. We excluded rarely used or mostly-empty fields (for example award flags) from the modeling feature set.
5. **Basic feature set.** We focused on a straightforward feature set consisting of MP, G, TS%, USG%, AST%, TRB%, STL%, BLK%, TOV%, one or two overall impact metrics (primarily BPM and WS), and demographic variables (Age and position dummies).

Challenges

Some challenges we encountered were:

- **Data Consistency Across Multi-Team Seasons:** Basketball-Reference reports separate rows for each team stint and sometimes an aggregated "2TM/3TM" line. Ensuring a consistent unit of analysis—one row per player-season—required resolving conflicts between these representations. For players without a consolidated line, selecting the stint with the most minutes was a practical choice, but it introduced uncertainty about whether that stint fully captured a player's season-long role or performance context.
- **Missing Salary Information and Sample Reduction:** A substantial number of players lacked a 2025–26 base salary (free agents, 10-day contracts, two-way players, and players not retained). This forced us to drop

them, reducing the sample and potentially introducing selection bias, since players without next-season contracts differ systematically (e.g., low-minutes players or veterans near retirement). Our results therefore primarily reflect contracted players rather than the full NBA population.

- **Right-Skewed Salary Distribution:** NBA salaries follow a heavily right-skewed distribution, with most players clustered near the minimum and a long tail of superstar contracts. This skew impacts model stability—especially linear models—and required careful consideration when interpreting residuals, identifying influential points, and evaluating predictive performance. Extremely large contracts can dominate loss functions and inflate error metrics.

- **Merging via IDs and Name Cleaning:** Most rows included a Basketball-Reference ID, but a minority required name-based merging. Name matching introduces risks related to suffixes, abbreviations, and inconsistent spellings (e.g., "Jr.", hyphenated names, multi-last-names). Even light cleaning can create ambiguity when multiple players share similar names or when a player changes how their name is listed across seasons.

2. Methods and Results

Method Overview

We treated the prediction of 2025–26 base salary as a supervised regression problem with log-transformed salary as the continuous target and player-level statistics as predictors. Before fitting models, we performed basic exploratory data analysis to understand the distribution of key variables and their pairwise relationships with salary.

We examined histograms for salary (Figure 1) and major predictors. These visualizations helped motivate the log transformation of salary. After examining the scatter plots relating salary to playing time, age, usage, and overall impact metrics (Appendix A), we found that there is a presence of high-salary outliers and low-minute players with noisy performance statistics. Correlation matrices provided a first look at multicollinearity, especially among composite metrics such as `WS`, `BPM`, and `VORP`. We also inspected potential interactions informally, since many basketball statistics are mechanically related.

We began with simple, interpretable linear models and then moved to regularized regression and ensemble methods. This step-by-step approach helped us keep track of how much additional predictive power we gained from extra flexibility, and it kept the connection between model outputs and basketball intuition clear. The analysis assessed both predictive accuracy and how different player attributes contributed to contract value.

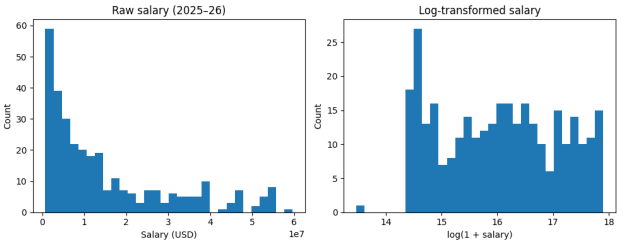

Figure 1. Histogram of Salaries

Models

Our analysis aimed to identify which player performance metrics most strongly predict NBA salaries while balancing model interpretability and predictive accuracy. We adopted a staged modeling approach that starts with transparent regression methods and progresses to a more flexible tree-based ensemble.

Linear Regression. We first fitted a standard Ordinary Least Squares (OLS) regression as a baseline. This model estimated marginal effects: how much log-salary changes, on average, when a metric such as minutes played or win shares increases by one unit, holding other variables fixed. The OLS model provided a clear benchmark and helped diagnose skewness, multicollinearity, and influential observations before adding regularization.

Ridge and Lasso Regression. Because many predictors are highly correlated, we next fit Ridge and Lasso regression models. Ridge regression (L2 penalty) shrinks coefficients toward zero when predictors are correlated, reducing variance but keeping all variables in the model. Lasso regression (L1 penalty) goes further by shrinking some coefficients exactly to zero, which performs variable selection and highlights a smaller set of important predictors. Comparing these regularized models to OLS allowed us to see how much performance we gained from shrinkage and which features remained consistently important.

Random Forest. To allow for nonlinear relationships and higher-order interactions, we also fit a Random Forest regressor. This ensemble method aggregates many decision trees trained on bootstrapped samples of the data and considers random subsets of features at each split. Random Forests can capture complex patterns, such as diminishing returns to minutes or age, without explicitly specifying them. We used the model both as a predictive benchmark and as a way to obtain feature importance scores that complement the linear

coefficient-based interpretation.

Implementation and Preprocessing. All models were implemented in Python using the `scikit-learn` library. Numeric predictors were standardized to zero mean and unit variance, and categorical variables (positions) were encoded with one-hot indicators. The dependent variable, salary, was log-transformed to reduce skewness and stabilize variance, which improved the fit and made residual diagnostics more well-behaved. All models were fit using the same training and test splits and the same preprocessed feature matrix to allow a fair comparison.

Model Validation Plan and Execution

Data Splitting and Cross-Validation. We partitioned the data into an 80% training set and a 20% held-out test set. This split provided a sufficiently large sample for model fitting while preserving enough data to generate an unbiased estimate of generalization performance. Within the training data, we performed 5-fold cross-validation for hyperparameter tuning and model selection, as this fold size balanced computational efficiency with stability in validation estimates.

For Ridge and Lasso regression, we used scikit-learn implementations and tuned the regularization strength (α) via a logarithmic grid search. For Ridge, we evaluated 7 α values from 10^{-3} to 10^{3}, while for Lasso we used a slightly narrower range (10^{-3} to 10^{1}) and allowed up to 10,000 iterations to ensure convergence. For the Random Forest model, we tuned the number of trees (200-400), maximum depth (5-20), and minimum samples per leaf (1-3) to control model complexity and reduce overfitting. Model performance were assessed using RMSE, MAE and R^2 averaged across folds to ensure robustness across accuracy metrics.

This validation strategy balances bias and variance, supports fair comparison among models, and provides a reliable estimate of out-of-sample predictive performance.

Evaluation Metrics. Each model was assessed using three metrics:

- **Root Mean Squared Error (RMSE)** to measure the typical size of squared prediction errors on the log-salary scale.
- **Mean Absolute Error (MAE)** to summarize average absolute prediction error.
- R^2 **(Coefficient of Determination)** to quantify the proportion of variance in log-salary explained by the model.

Looking at all three metrics together helped us understand both accuracy and variance, rather than relying on a single number.

Model Diagnostics. For the linear models, we examined residuals to check key assumptions. Residuals-versus-fitted plots were used to look for systematic patterns, such as increasing variance at higher salaries. Q–Q plots of residuals indicated that the log transformation of salary made the error distribution closer to normal, although some deviations remained in the tails due to superstar contracts. We also verified that residuals did not show obvious dependence patterns across observations, which is consistent with the cross-sectional nature of the data.

Model Comparison and Interpretation. After validation, we compared models based on their test-set performance and interpretability. Scatter plots of predicted versus actual log-salary showed that most models tracked mid-range players well but struggled with extreme high earners. Coefficient estimates from the regularized linear models and feature importance scores from the Random Forest were analyzed together to identify which predictors were consistently influential. Indicators of playing time (`MP`), offensive role (`USG%`), age, and overall impact metrics (`BPM`, `WS`) emerged as the most important factors across model classes.

2.1. Model Performance

Table 1 summarizes out-of-sample predictive accuracy for the four models considered: OLS, Ridge, Lasso, and Random Forest. All models were trained on the same standardized feature set and evaluated on log-transformed salary.

Across the linear models, regularization provided a small but consistent improvement over OLS. The **Lasso model achieved the lowest test RMSE (0.599)** and the highest R^2 (0.596) among the linear models, slightly outperforming Ridge (RMSE 0.600, R^2 0.595) and OLS (RMSE 0.602, R^2 0.591). This pattern reflects the benefit of coefficient shrinkage in reducing variance when predictors are highly correlated.

The Random Forest regressor exhibited a somewhat different tradeoff. It achieved the **lowest MAE (0.481)** among all models, which indicates strong accuracy for typical players, but it had a higher RMSE (0.610) and a slightly lower R^2 (0.581) than Lasso. This suggests that, while the Random Forest did a good job for most players, it struggled more with extreme high-salary outliers than the regularized linear models.

Overall, Lasso provided the best balance of interpretability and predictive performance among the linear models and served as our primary model for coefficient-based interpretation.

Table 1. Test-set performance across models. Lower RMSE/MAE is better; higher R^2 is better.

Model	RMSE	MAE	R^2
OLS	0.602	0.503	0.591
Ridge	0.600	0.506	0.595
Lasso	**0.599**	0.505	**0.596**
Random Forest	0.610	**0.481**	0.581

2.2. Random Forest Feature Importance

To examine nonlinear effects and interactions, we inspected the feature importance scores from the Random Forest model. The most important predictors were:

- **USG%** (importance 0.2499), reflecting offensive role and on-ball usage.
- **Age** (0.2347), capturing nonlinear career-stage salary patterns.
- **MP** (0.1723), representing availability and playing time as central drivers of compensation.
- **BPM** (0.1008), as an overall measure of player impact.
- **WS** (0.0836), summarizing total seasonal contribution.

Secondary predictors such as `AST%`, `TRB%`, `TS%`, `BLK%`, and `STL%` contributed modestly but did not approach the importance levels of usage, age, minutes, and overall impact metrics. Positional indicators (for example `Pos_SG`, `Pos_PG`, `Pos_SF`) had very low importance scores, suggesting that the model captured role differences more effectively through continuous statistical performance metrics than through coarse position labels.

2.3. Linear Model Coefficients: Ridge and Lasso

We examined the Ridge and Lasso coefficients to interpret how individual predictors influence log-salary in a linear framework.

Strong Positive Predictors. Both models identified the same core set of strong positive predictors:

- **MP**: Ridge coefficient 0.556, Lasso coefficient 0.630.
- **Age**: Ridge 0.369, Lasso 0.372.
- **USG%**: Ridge 0.266, Lasso 0.262.
- **BPM**: Ridge 0.197, Lasso 0.198.

These variables capture playing time, offensive load, and overall on-court impact, and they are central determinants of player valuation in both the data and basketball intuition.

Weak or Negative Predictors. Some predictors showed weaker or negative associations once the core variables were included:

- **G** had negative coefficients (Ridge -0.219, Lasso -0.257), likely because `MP` already captures availability more precisely and the number of games can vary due to factors like short stints or team context.
- **TS%** had small negative coefficients (Ridge -0.118, Lasso -0.107), probably reflecting collinearity with other offensive metrics rather than a true negative relationship with salary.
- Several predictors, including `TRB%`, `TOV%`, and positional dummies, received coefficients near zero in the Lasso model, indicating limited marginal predictive power after accounting for stronger metrics.

Lasso's sparsity highlighted that a compact set of features (MP, Age, USG%, and BPM) captured most of the linear signal in salary determination.

2.4. Comparison Across Model Classes

Taken together, the models reveal several consistent themes:

- Linear structures explain most of the variance in log-salary, and regularization improves generalization slightly over plain OLS.
- The Random Forest captures some nonlinearities, especially related to age and diminishing returns to minutes and usage, but does not dramatically outperform Lasso.
- The main drivers of salary are robust across model classes: playing time, offensive role, and overall impact metrics consistently appear at the top of coefficient and importance rankings.

These patterns suggest that, for this dataset and prediction horizon, relatively simple regularized linear models are sufficient to capture most of the structure in the salary data.

2.5. Prediction Diagnostics

The residual density plot (Figure 2) and the predicted-versus-actual plot (Figure 3) for the Lasso model reveal two systematic patterns:

1. **Under-prediction of superstars.** Players with the very highest salaries are consistently under-predicted. Their salaries are influenced by max-contract rules, branding value, and market dynamics that are only loosely connected to box score metrics, which limits what any performance-based model can capture.

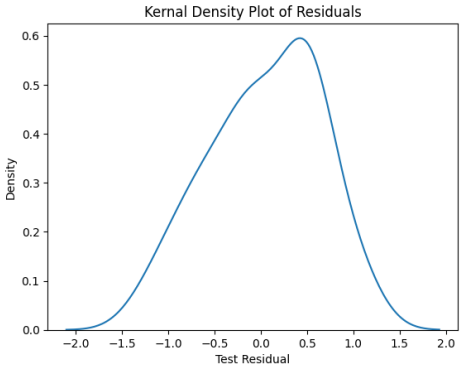

Figure 2. KDE of test residuals for the Lasso model.

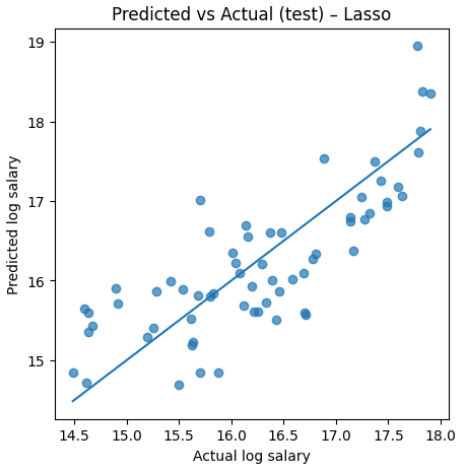

Figure 3. Predicted versus actual log-salary on the test set for the Lasso model.

2. **Noisier predictions for low-minute players.** Bench players and rookies often have salaries shaped by draft position, contract structures, and team development strategies rather than their current-season performance. This leads to higher variance in residuals at the low end of minutes and salary.

For the majority of players in the middle of the salary distribution, predictions are relatively stable and do not show strong systematic bias.

2.6. Summary of Findings

Across all models, a consistent set of predictors explains most of the variation in NBA salaries. The most important factors are:

- **USG%** (offensive role),
- **Age** (career stage),
- **MP** (availability and playing time),
- **BPM** and **WS** (overall impact).

These variables emerge as dominant drivers in both linear and nonlinear frameworks. In contrast, position and narrower rate statistics provide little additional predictive power once these core metrics are accounted for. The results suggest that teams primarily compensate players based on how much they play, how central they are to creating offense, and how much value they generate in aggregate over a season.

3. Conclusion

This project aimed to identify which on-court performance metrics most strongly predict NBA player salaries by analyzing publicly available Basketball-Reference data from the 2024–25 season alongside 2025–26 base salary figures. After cleaning, reconciling, and merging player statistics with contract information, and ensuring that performance precedes salary chronologically, we constructed a player-level dataset suitable for a supervised regression task.

We evaluated four models within a unified pipeline: OLS, Ridge, Lasso, and Random Forest. Predictors were standardized, salaries were log-transformed, and model performance was assessed via cross-validation using an 80/20 train–test split. Metrics included RMSE, MAE, and R^2. The regularized linear models provided modest improvements over OLS, with Lasso offering the strongest balance of lower RMSE and higher R². Random Forest achieved the lowest MAE but exhibited a higher RMSE and slightly reduced R^2. These results suggest that regularization effectively mitigates issues arising from correlated predictors, and that highly nonlinear approaches are not essential given the feature set used.

Across all modeling approaches, a core group of variables consistently stood out. Usage Percentage, Age, Minutes Played, and Box Plus/Minus were robust positive predictors of log-salary in both linear and tree-based models. Win Shares also contributed meaningfully, though to a lesser extent. In contrast, positional indicators and several more granular rate statistics added limited incremental value once the primary variables were accounted for. From a basketball standpoint, these findings align with the notion that teams compensate players who log significant playing time, assume major offensive responsibilities, and generate strong overall impact throughout the season.

At the same time, our analysis has clear limitations. First, we modeled only a single season and a single-year salary outcome. The sample selection is biased as only players have recorded statistics and signed contrasts specifically for the year of 2024-25 and 2025-26. True player valua-

tion depends on multi-year performance, injuries, contract timing, and league-wide cap dynamics, none of which are fully captured in one year of box score data. Second, our models rely entirely on performance statistics and a few demographics. They do not include market factors such as team market size, international appeal, or off-court revenue potential, even though these factors are known to influence high-end contracts. Third, our treatment of contract structures is simplified. We focused on base salary for a particular season, but many contracts contain options, incentives, and non-linear escalation clauses that complicate the mapping from performance to pay. Fourth, our model could perform poorly on the extreme cases. Certain types of players such as rookies or other roles who only have small amount of playing minutes, causing smaller sample playing size for them, that could not fully capture their performance with their actual salary.

These limitations also point to natural directions for future work. A richer project could extend the dataset to multiple seasons, allowing models to incorporate trends in player performance over time and to separate long-term value from single-season spikes. Including team and market-level variables, such as franchise revenue, playoff appearances, and cap space, could improve predictions, especially for stars whose contracts reflect both on-court impact and broader business considerations. More advanced modeling techniques, such as gradient boosting, generalized additive models, or partial pooling models that borrow strength across positions or roles, might capture subtle nonlinearities and interactions more effectively. Finally, future work could focus on uncertainty quantification, for example by constructing prediction intervals for salaries and studying how uncertainty varies across different types of players. Also, comparing model residuals across different positions or salary tiers might help with the issue of extreme cases, so we could separate them deeper into categories.

Despite these caveats, our results already provide a useful and interpretable summary of how NBA salaries relate to on-court performance. For most players, especially those in the middle of the salary distribution, relatively simple models based on minutes, usage, age, and overall impact metrics can explain a large share of salary variation. Their salaries track statistical contributions closely, suggesting a relatively rational market. For superstars and players in unusual contract situations, salaries are less tightly connected to standard statistics. This pattern highlights both the power and the limitations of using machine learning on publicly available sports data to study labor markets in professional sports. These patterns shed light on how efficiently the NBA labor market prices production. This could be possibly explained by the branding, scarcity, and negotiation dynamics appear to decouple compensation from measured performance at the high end, highlighting the dual athletic and entertainment nature of NBA player value. We hope our results could provide insights on front-office decision making, contract negotiations, or salary projections, while curious fans could also use it to explore more about the basketball industry through a different lens.

References

Basketball-Reference. 2024–25 nba player stats: Advanced, 2025a. URL https://www.basketball-reference.com/leagues/NBA_2025_advanced.html. Accessed: 2025-09-26.

Basketball-Reference. 2025–26 nba player contracts, 2025b. URL https://www.basketball-reference.com/contracts/players.html. Accessed: 2025-09-26.

A. Exploratory Data Analysis: Predictor vs. Log-transformed salaries scatter plots