NFL Quarterback Stats: Identifying Key Predictors of Win Percentage

Cooper C, Manas R, Mark B, Elijah R

2025-12-05

Contents

1 Abstract

This report investigates the relationship between various quarterback performance statistics and their associated win percentage in the NFL. Using game log data for starting quarterbacks, we analyze key metrics such as passing yards, rushing yards, passer rating, touchdowns, interceptions, sacks, and fumbles. Exploratory data analysis reveals distinct distributions for different stats and highlights a positive correlation between metrics like Passer Rating and TD Passes with Win Percentage, while Interceptions and Fumbles show a negative correlation. To quantify these relationships, we developed two predictive models: Lasso (Least Absolute Shrinkage and Selection Operator) regression and Stepwise linear regression. Both models identified Passer Rating, TD Passes, Interceptions, and Fumbles as significant predictors of win percentage. The models achieved comparable performance on test data, explaining approximately 37-38% ($R^2 \approx 0.37$) of the variance in win percentage. These findings suggest that while many factors contribute to winning, certain career statistics are significant drivers of success for starting quarterbacks. This analysis provides insights for player evaluation and understanding the components of effective quarterback play.

2 Introduction

The quarterback is often considered the most critical position in American football, directly influencing the outcome of games. Understanding which statistical measures best capture a quarterback's contribution to winning is a central question in sports analytics. This report aims to explore this question using publicly available NFL data.

- **Data Source:** The data originates from the "NFL Statistics" dataset available on Kaggle(https://www.kaggle.com/datasets/kendallgillies/nflstatistics/data), compiled by Kendall Gillies [1]. It contains detailed game logs for various positions, including quarterbacks, spanning multiple seasons.
- **Data Scope:** The initial dataset includes game-by-game statistics. For this analysis, we focused specifically on quarterbacks who were designated as the starter for a given game. The post cleaned dataset (`new_df`) contains 216 unique starting quarterbacks, summarizing their average per-game performance across 9 key statistical features, plus win/loss records.
- **Motivation & Goals:** We were interested in identifying which commonly tracked quarterback statistics have the strongest statistical link to winning games. Our primary goals are:
    1. To explore the distribution of key quarterback performance metrics.
    2. To visualize the relationship between these metrics and career win percentage.
    3. To build and compare statistical models to predict win percentage based on these metric to identify the most influential predictors.

3 Data Cleaning

The data required several cleaning procedures to prepare it for analysis

1. **Loading Data:** The `Game_Logs_Quarterback.csv` file was loaded into R.
2. **Data Type Conversion:** Columns containing statistics were ensured to be in to numeric format.
3. **Zeroed Missing Values:** During numeric conversion, any non-numeric entries became `NA`. These `NA` values, along with pre-existing `NA`s in numeric columns, were changed to 0. This assumes that a missing value often corresponds to zero occurrences of that statistic in a game.
4. **Filtered for Starters:** The dataset was filtered to include only rows where the quarterback had `Games Started` equal to 1, focusing the analysis on primary players for each game.
5. **Aggregation by Players:** The game-level data for starters was aggregated by player (`Name`). We calculated the total games started, total wins, and total losses. For performance statistics such `Passing Yards`, `Passer Rating`, `Ints`, we calculated the *mean* value across all started games to represent the quarterback's average per-game performance.
6. **Chose Specific Feature:** A `Win_Percentage` metric was calculated for each quarterback as `Wins / Games Started`.
7. **Final Check:** Rows with any remaining missing values in the predictor variables were removed to ensure compatibility with modeling algorithms.

4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis helps us understand the characteristics of our data and potential relationships before modeling.

4.1 Distribution of Key Statistics

We examined the distributions of average Passing Yards, Rushing Yards, and Passer Rating per game across all starting quarterbacks in our dataset.

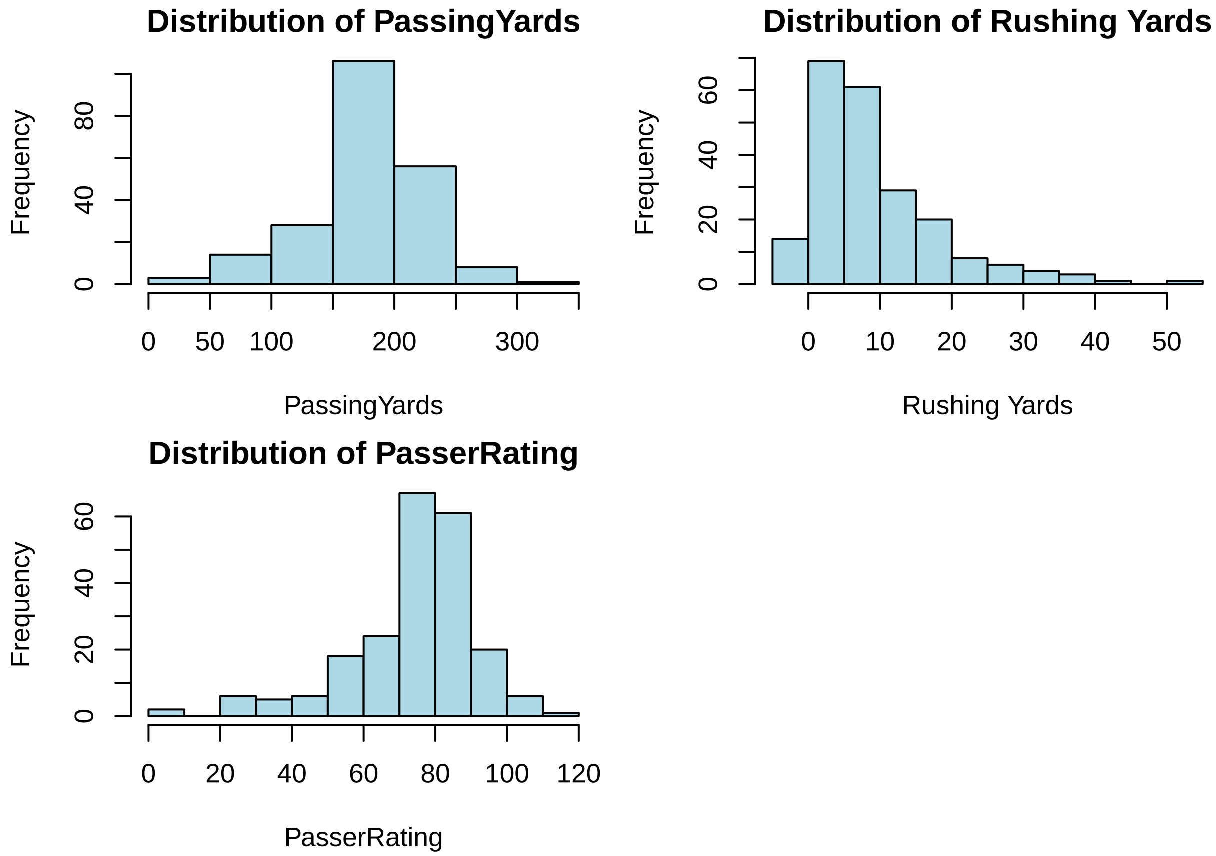

Figure 1: Histograms showing the distribution of average per-game Passing Yards, Rushing Yards, and Passer Rating for starting quarterbacks.

- Average `PassingYards` per game appears roughly bell-shaped, centered around 200 yards, with a slight right skew.
- Average `Rushing Yards` per game is heavily right-skewed, with most quarterbacks averaging very few rushing yards, but a tail of quarterbacks averaging significantly more.
- Average `PasserRating` seems approximately normally distributed, centered around a rating of 80.

4.2 Relationship between Wins and Passing Yards

To explore how performance relates to overall success, we grouped quarterbacks by their total career wins and examined their average passing yards per game.

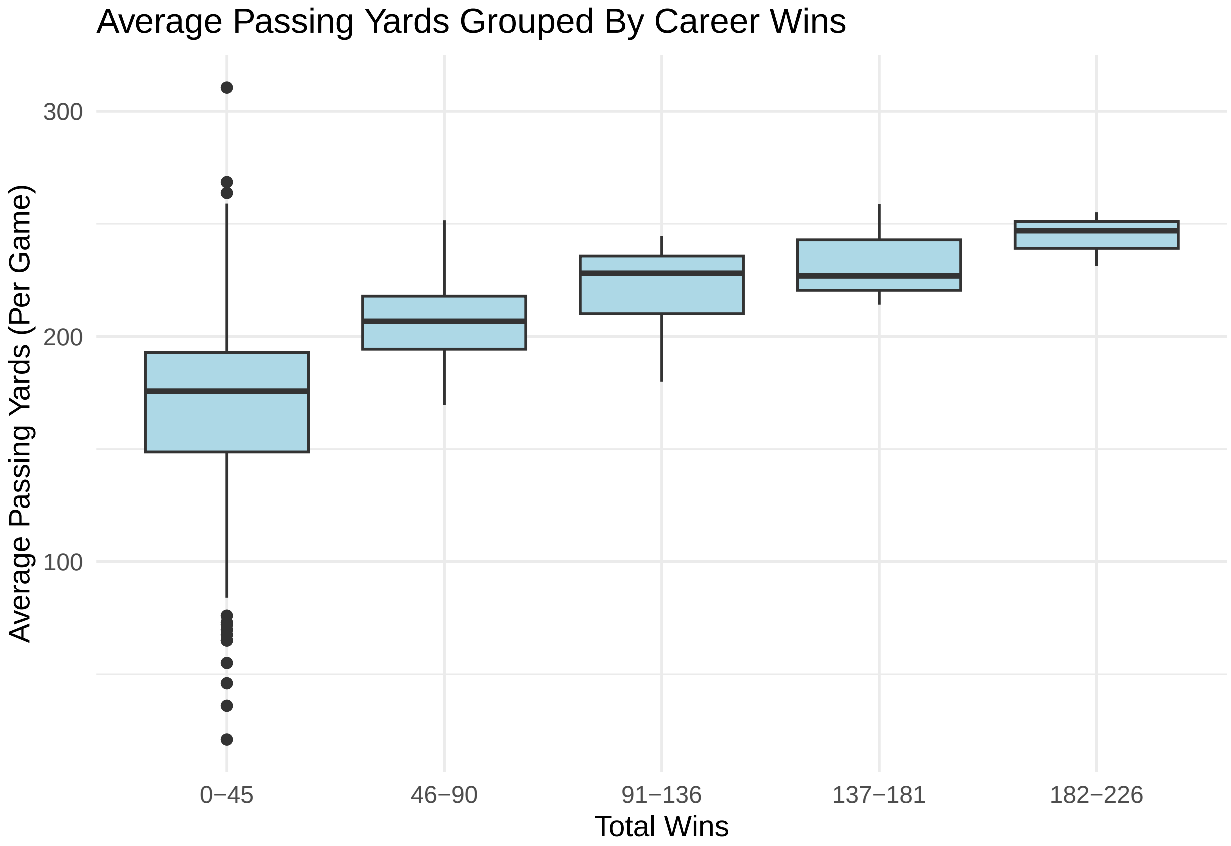

Figure 2: Box plot showing the distribution of average Passing Yards per game, grouped by total career wins.

The box plot suggests a positive trend where quarterbacks in groups with more career wins tend to have higher average passing yards per game. The variability in passing yards also appears to decrease slightly in the higher win groups.

4.3 Correlation with Win Percentage

We calculated the Pearson correlation coefficient between each average per-game statistic and the quarterback's career win percentage to get a preliminary idea of linear relationships.

- **Strongest Positive Correlations:** `PasserRating` shows the highest positive correlation with `Win_Percentage`, followed by `TD Passes` and `PassingYards`.

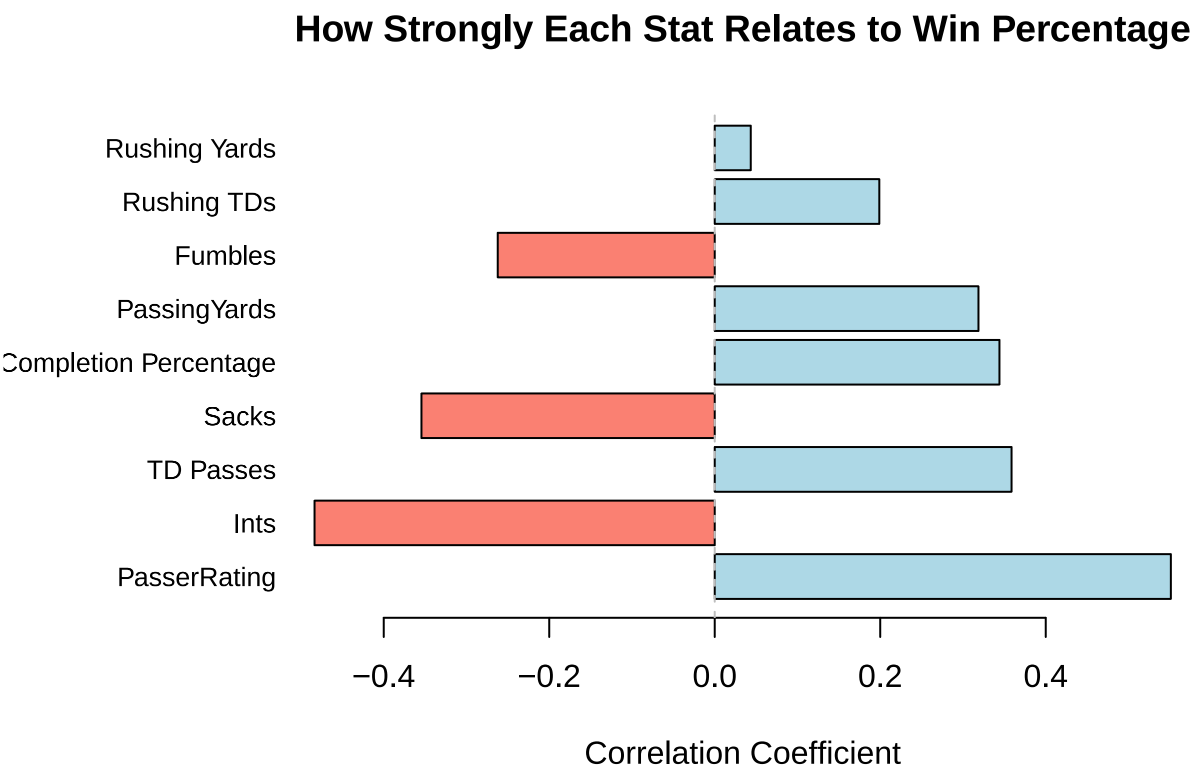

Figure 3: Bar chart showing the Pearson correlation coefficient between various QB stats and Win Percentage.

- **Strongest Negative Correlations:** Ints (Interceptions) and Fumbles have the strongest negative correlations, indicating that turnovers are strongly associated with lower win percentages. Sacks also show a negative correlation.
- **Weaker Correlations:** Rushing Yards and Rushing TDs show weaker positive correlations compared to passing stats. Completion Percentage has a moderate positive correlation.

5 Analysis

Based on the EDA, several statistics appear related to win percentage. We now build regression models to quantify these relationships simultaneously and identify the most important predictors while accounting for potential situatations where two or more independent variables in a model are highly correlated with each other.

5.1 Model Preparation

We split the data into training (70%) and testing (30%) sets to evaluate model performance on unseen data. Win_Percentage is the target variable (y), and the average per-game statistics are the predictors (x).

5.2 Lasso Regression

Lasso regression performs linear regression but adds a penalty (L_1 norm) to the coefficient magnitudes, which can shrink some coefficients exactly to zero, effectively performing feature selection. We use cross-validation to find the optimal penalty strength with lambda.

Table 1: Lasso Model Coefficients (Optimal Lambda)

Feature	Coefficient
(Intercept)	0.2067
Completion Percentage	0.0000
PassingYards	0.0013
Rushing Yards	0.0024
PasserRating	0.0020
TD Passes	0.0076
Rushing TDs	0.0000
Ints	-0.1108
Sacks	-0.0240
Fumbles	-0.1294

The Lasso model selected several predictors, shrinking the coefficients of Completion Percentage and Rushing TDs to zero, suggesting they offer less predictive value when other variables are present. Key predictors include PasserRating, TD Passes, Ints, Fumbles, PassingYards, Rushing Yards, and Sacks.

The Lasso path plot visualizes how coefficients are penalized as lambda increases. Variables that remain non-zero for longer where they are further right are generally considered stronger predictors. PasserRating, Ints, and Fumbles appear particularly influential.

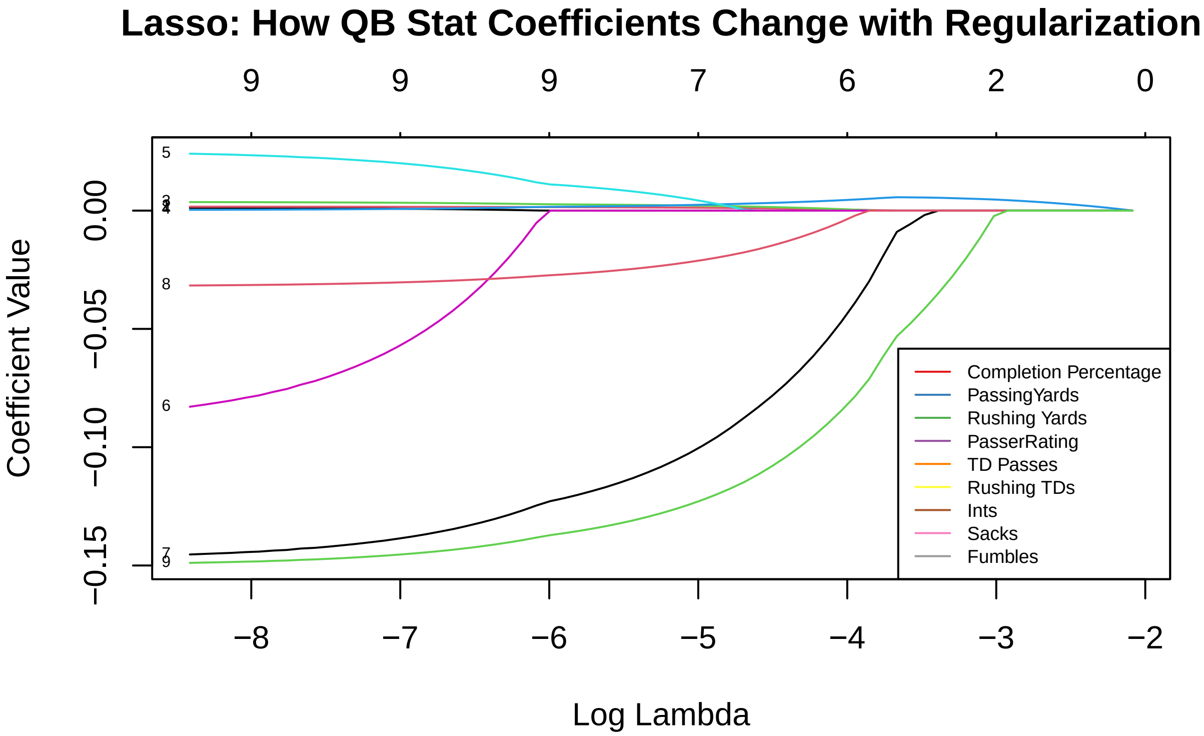

Figure 4: Lasso coefficient paths as a function of the log of the penalty parameter (lambda). Labels indicate when variables enter the model.

5.3 Stepwise Regression

Stepwise regression automatically selects variables for a linear model by iteratively adding or removing predictors. We use a bidirectional approach, considering both adding and removing variables.

Table 2: Stepwise Regression Final Model Summary

Feature	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	0.3228	0.0677	4.7673	0.0000
PassingYards	0.0020	0.0003	6.6202	0.0000
Rushing.Yards	0.0030	0.0016	1.7982	0.0742
Ints	-0.1583	0.0281	-5.6289	0.0000
Sacks	-0.0323	0.0195	-1.6613	0.0988
Fumbles	-0.1610	0.0408	-3.9506	0.0001

The final Stepwise model selected `PassingYards`, `Rushing Yards`, `Ints`, `Sacks`, and `Fumbles` as significant predictors. Looking at it closley we see it excluded `PasserRating` and `TD Passes`, likely due to their high correlation with other variables like `PassingYards` and `Ints`. The coefficients align with expectations where it is positive for yards and negative for turnovers.

6 Results

6.1 Model Performance Comparison

We compare the performance of the Lasso and Stepwise models using the R^2 (coefficient of determination) value on both the training and testing datasets. R^2 represents the proportion of the variance in the dependent variable (Win Percentage) that is predictable from the independent variables (QB stats).

Table 3: Model R-squared Comparison

Model	Data Set	R-squared
Lasso	Train	0.4263
Lasso	Test	0.3667
Stepwise	Train	0.4275
Stepwise	Test	0.3812

Both models show similar performance. The R^2 values on the training data are slightly higher than on the test data, which is expected. The test R^2 values of approximately 0.37 for Lasso and 0.38 for Stepwise indicate that the selected quarterback statistics explain about 37-38% of the variability in win percentage in this dataset on unseen data. While this shows a significant relationship, it also implies that over 60% of the variance is due to other factors not included in the model.

6.2 Key Predictors Summary

Comparing the variables selected by both models and their coefficients

- **Consistently Important (Negative Impact):** `Ints` and `Fumbles` were identified by both models as having a significant negative impact on win percentage. This strongly suggests that ball security is paramount. `Sacks` also showed a negative impact in both models.

- **Consistently Important (Positive Impact):** PassingYards was significant and positive in both models. So making sure the quarterback has consistently connecting throws is important rather then focusing on touchdown throws or any other maybe more risky plays.
- **Model Differences:**
  - Lasso retained PasserRating and TD Passes (positive) and Rushing Yards (positive), while shrinking Completion Percentage and Rushing TDs to zero.
  - Stepwise selected Rushing Yards (positive) but excluded PasserRating and TD Passes, likely due it conflicting with the other independent variables.

Overall, the analysis points towards passer efficiency, touchdown generation, and avoiding turnovers/sacks as the most statistically important indicators of higher win percentages among starting quarterbacks in this dataset.

7 Discussion and Conclusion

This analysis aimed to identify which quarterback statistics best predict win percentage using game log data for NFL starters. Through EDA and regression modeling, we found statistically significant relationships between several performance metrics and winning. The data consistently showed a strong positive correlation and was a key predictor in the Lasso model. Interceptions and Fumbles demonstrated strong negative correlations and were significant negative predictors in both models. TD Passes (Lasso) and Passing Yards (both models) were positively associated with winning. Both Lasso and Stepwise models explained a similar, moderate amount of variance ($R^2 \approx 0.37 - 0.38$ on test data), indicating that while QB stats are important, they are only part of the complex equation of winning NFL games.

Limitations

- **Omitted Variables:** The models do not account for many other factors influencing game outcomes, such as defensive performance, special teams play, opponent quality, coaching strategies, or game situation (such as the weather).
- **NA Imputation:** Imputing missing stats with 0 is a simplification.
- **Correlation vs. Causation:** While we identified statistical associations, these models do not prove causation. Good stats might lead to wins, or winning teams might enable quarterbacks to achieve better stats.
- **Data Granularity:** Using per-game averages over a career removes any variation and doesn't capture situational performance.

Future Work

- Incorporate opponent adjustments such as strength of game scheduling or defensive rankings.
- Include team-level variables such as defensive performances and offensive line performances.
- Explore non-linear relationships between variables.
- Analyze data on a play-by-play level for more granular insights.

In conclusion, this analysis confirms the statistical importance of efficient, high-scoring, and secure quarterback play for achieving wins in the NFL. Coaches in the NFL will likely want to focus on making sure their Quarterback is strongly defended while ensuring they have good passing efficency to improve the win percentage of the team overall. But while individual stats don't tell the whole story, metrics related to passing efficiency, scoring TDs/gaining Yards, and avoiding mistakes such as Interceptions/Fumbles/Sacks are demonstrably linked to success.

8 Appendix: R Code

R code used with packages such as `tidyverse`, `readr`, `dplyr`, `glmnet`, `ggplot2`, and `knitr`. Used AI to get help for certain abstract problems such as graph and model inspiration (Google Gemini 2.5, 2025 and ChatGPT Model 4o, 2025)

9 Acknowledgements

1. Mark Brown - Power-point & Basic Graph
2. Cooper Crow - Data Wrangling & Basic Graph & Report Writing
3. Manas Reddy - Modeling & Basic Graph & Report Writing
4. Elijah Robledo - Power-Point & Report Writing

10 Bibliography

[1] Gillies, Kendall. "NFL Statistics". Kaggle, 2023. Accessed May 12, 2025. URL: https://www.kaggle.com/datasets/kendallgillies/nflstatistics/data

[2] RDocumentation for kable. Accessed May 12, 2025. URL: https://www.rdocumentation.org/packages/knitr/versions/1.48/topics/kable

[3] Geekforgeeks Article for Stepwise Regression in R, 2024. Accessed May 12, 2025. URL: https://www.geeksforgeeks.org/stepwise-regression-in-r/

[4] Statology Article for Lasso Regression in R, Zach Bobbitt 2020. Accessed May 12, 2025. URL: https://www.statology.org/lasso-regression-in-r/

[5] R-bloggers Article for the Pearson Correlation in R, 2021. Accessed May 12, 2025. URL: https://www.r-bloggers.com/2021/10/pearson-correlation-in-r/

[6] Tidyverse Documentation for ggplot2. Accessed May 12, 2025. URL: https://ggplot2.tidyverse.org/reference/