Analyzing the Influence of Player Performance Metrics on NBA MVP Voting Scores*

Ruiyi Liu

December 24, 2024

This study used linear regression to analyze the impact of NBA player performance metrics on MVP voting scores. Using data from the 2016 - 2019 seasons, key predictors such as scoring, assists, rebounding, and advanced metrics were analyzed. Regression models were trained and validated on separate datasets and diagnostic checks were performed to ensure compliance with linear regression assumptions. The analysis identified key performance metrics that influence MVP results and demonstrated the reliability of the model. These findings provide insights into the evaluation criteria for the NBA MVP award and highlight the role of analytics in sports decision-making.

1 Introduction

In basketball analytics, player performance metrics are critical in assessing individual contributions and overall value to the team.The NBA Most Valuable Player (MVP) award is a prestigious award based on a combination of subjective voting and objective player statistics. The focus of this study is to understand how player-specific performance metrics affect MVP voting results. Using a dataset containing NBA player statistics from the 2016 through 2019 seasons, we develop and evaluate statistical models to identify key predictors of MVP scores.

The purpose of this analysis is twofold: first, to explore the relationship between player metrics (e.g., points, assists, rebounds, and other performance metrics) and MVP voting scores; and second, to validate the stability and generalizability of these models using training and test datasets. By analyzing the statistical significance of these predictors and model assumptions, we aim to gain a deeper understanding of the factors that drive MVP voting decisions.

This report outlines the methods used to clean and preprocess the data, the steps taken to fit the multiple linear regression models, and the validation process to ensure the reliability of

[*] Code and data are available at: https://github.com/zora0131/math_ia.git.

the results. The results of this analysis provide valuable insights into the evaluation criteria for MVP awards and the role of advanced basketball metrics in shaping these decisions.

2 Data Overview

2.1 Measurement

The dataset comprises NBA player statistics from the 2016 to 2019 seasons, encompassing various performance metrics and player information.

Key variables include total points `scored (PTS)`, `assists (AST)`, `rebounds (TRB)`, `minutes played (MP)`, and `field goal percentage (FG%)`.These metrics are standard in basketball analytics, providing insights into a player's scoring ability, playmaking skills, effectiveness in gaining possession, playing time, and shooting efficiency.

The dataset also includes the number of games played (G), which is essential for calculating per-game averages, allowing for standardized comparisons across players. Additionally, the dataset contains the `Score` variable, representing the player's MVP voting results, serving as the dependent variable in our analysis. These measurements are crucial for evaluating player performance and understanding the factors influencing MVP voting outcomes.

Note: The dataset is sourced from Kaggle and includes player statistics from the 2016 to 2019 NBA seasons.

2.2 Data Cleaning

We used the`R`programming language (R Core Team 2022), the `here` package (Müller and Bryan 2023), the `dplyr` package (Wickham et al. 2023), the `ggplot2` package (Wickham 2016), the `knitr` package (Xie 2023), the `kableExtra` package (Zhu 2023), the `car` package (Fox and Weisberg 2023), the `gridExtra` package (Auguie 2023) to clean the data, plot the graphs and tables, fit the models.

Then we select the variables that we think are important in raw data to form cleaned data. Instead of selecting variables such as player names and ids that are not useful for building the model, we chose age, salary, and data about on-field performance as cleaned data.

We split the cleaned dataset into 2 part, train data and test data, each part consists 50% of the cleaned data. We will fit a model using the train data, and do the model validation using the model fitted by the test data.

Below is an overview of the cleaned data, only specific variables are secleted to be shown in the Table 1.

Table 1: The cleaned data overview

	Score	Age	X3P	mean_views	Rk	Salary	Role
2	48.2	32	0.7	11.155738	58	2700000	Back
3	40.0	21	1.0	1713.986339	157	4351320	Front
4	75.5	25	0.2	205.855191	352	2022240	Front
5	42.8	26	1.1	604.341530	10	7680965	Front
6	12.5	30	1.3	1556.382514	203	26540100	Front
7	60.0	32	0.0	603.188525	221	10230179	Front
8	59.5	34	0.5	5.409341	12	1315448	Front
9	85.5	24	0.0	21.213115	464	874636	Front
10	53.0	25	0.6	122.505465	65	9904495	Back
11	27.5	23	1.4	8.311475	1	5994764	Back

The Figure 1 shows the visualization of the cleaned dataset (Selected variables).

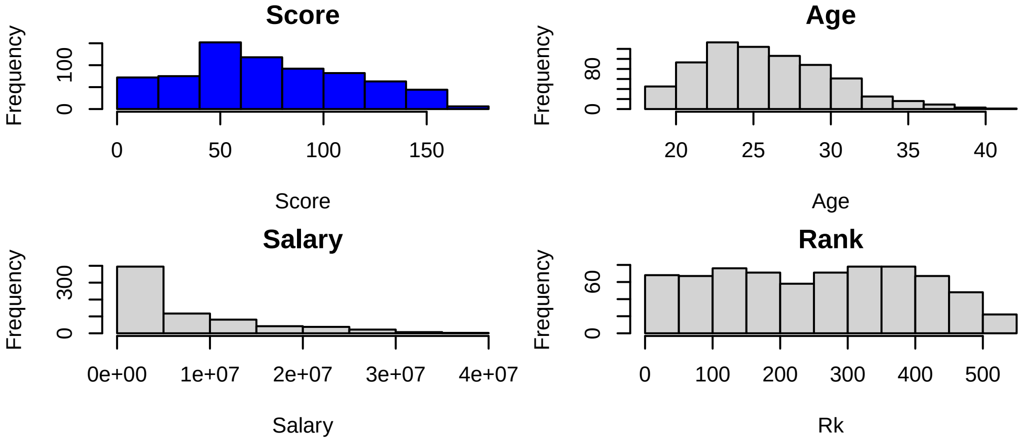

Figure 1: Visualization of the cleaned dataset (Selected variables)

3 Method

We first split the cleaned dataset into 2 parts by 50% : 50%, which are train data and test data. Using the train data to fit the model. We use all the variables in the train data to fit out Model 1, and selected the significant predictors to fit the Model 2, and we randomly select 2 predictors in the Model 2 to fit the Model 3, and apply partial F-test on the Model 2 and Model 3, this is to check if the Model 2 can be simplified. If the model can not be simplified, the Model 2 is our final model. Finally, we use all the predictors in the Model 2 and the test data to fit the Model 4, and then make model vailidation on these 2 models.

4 Model

Linear regression modeling is a statistical method for modeling the relationship between a dependent variable (response) and one or more independent variables (predictors). It assumes that the relationship between the variables is linear and estimates the coefficients of the linear equation that best predicts the response variable based on the predictors. The general form of the model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

where:

- y: the response variable.
- $x_1 \dots x_n$: the predictors.
- β_0: the intercept.
- $\beta_1 \dots \beta_n$: the coefficients of the predictors.
- ϵ: the random error, this should be normally distribute (Montgomery, Peck, and Vining 2012) (James et al. 2013).

And here are 3 assumption on the linear regression model, which are:

- **Linearity**: It is assumed that $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$, indicating a linear relationship between Y_i and x_i.
- **Homoscedasticity**: It is assumed that the variance of Y_i, denoted as $\mathrm{Var}(Y_i)$, does not depend on x_i.
- **Normality**: It is assumed that the residuals follow a normal distribution.

$$\begin{aligned}\widehat{\text{Score}} = {} & 115.8 - 0.3592(\text{Age}) + 15.14(\text{PTS}) \\ & - 1.516(\text{AST}) - 1.503(\text{TRB}) - 0.05607(\text{MP}) \\ & - 19.04(\text{FT}) - 54.36(\text{FG}) + 0.03302(\text{Rk}) \\ & + 7.234(\text{X3P}) + 19.28(\text{X2P}) + 13.82(\text{RoleFront}) \\ & + 5.409 \times 10^{-9}(\text{Salary}) - 3.491(\text{G}) \\ & + 2.436(\text{TOV}) - 0.003031(\text{mean\_views})\end{aligned}$$

From the math we learnt, if the p-value of the variable is smaller than 0.05, then that variable is significant. We first fit the models with all the variables in the train data, from the summary table of the Model 1, we can see that the variables **TRB**, **Rk**, **Role**, **G** and **mean_views** are significant.

Next, we fit our Model 2 by the significant predictor we selected in the Model 1, the Model 2 is shown below:

$$\begin{aligned}\widehat{\text{Score}} = {} & 103.4738121 - 5.1609752(\text{TRB}) + 0.0316097(\text{Rk}) + 21.7089780(\text{RoleFront}) \\ & - 0.4479400(\text{G}) - 0.0053988(\text{mean\_views})\end{aligned}$$

Then we can check if there is any collinearity between the predictors we selected. The collinearity in statistic refers to the predictors is linear dependent (contributors 2024a). This may leads to the unstable coefficient, difficulty in identifying significant predictors, redundancy among variables and so on.

The collinearity can be checked by compute the VIF (Variance inflation factor) of the predictors in the Model 2. In statistics, the variance inflation factor (VIF) measures how much the variance of a parameter estimate increases when other predictors are included in the model compared to when the model contains only that parameter. The formula for VIF is:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination (contributors 2024b).

We can see that all the VIF values is less than 5, so there is no collinearity between all the predictors.So we maintain our Model 2.

```
   Predictor VIF_Value
1        TRB     1.738
2         Rk     1.007
3       Role     1.269
4          G     1.333
5 mean_views     1.116
```

Next, we need to check if the model can be simplified, we randomly pick 2 variables **G** and **Rk** to fit our Model 3. We apply the partial F-test on Model 2 and Model 3 to see if this model can be simplify, here is the Model 3:

$$\widehat{\text{Score}} = 108.735967 + 0.028256(\text{Rk}) - 0.749699(\text{G})$$

A partial F-test is used to assess whether there is a statistically significant difference between a full regression model and a simpler, nested version of the same model. We will have 2 hypothesis, which are

H_0: The additional predictors in the full model do not provide a significant improvement in the model's fit.

H_A: The additional predictors in the full model provide a significant improvement in the model's fit.

If the p-value in the partial F-test is less than 0.05, we need to reject the null hypothesis H_0. In this case Table 2, the p-value is less than 0.05, so we reject the null hypothesis H_0 and choose our Model 2.

Table 2: The Partial F Test Outcome

Model	Residual DF	RSS	Df	Sum of Squares	F-Statistic	P-Value
Model 2	698	722336	NA	NA	NA	NA
Model 3	701	918333	-3	-195997	63.131	< 2.2e-16

So we finally choose the Model 3 to be our final model.

4.1 Model Quality

Although we successfully selected our current model, our model needs to meet the conditions to be a good model, and these conditions are

$$\text{Condition 1: Linearity}$$
$$\text{Conditon 2: Normality of Residuals,}$$

and we will examine them one by one.

We first check the Condition 1, the Condition 1 can be checked by the following plot Figure 2.

The linearity assumption of the model was assessed using a Residuals vs. Fitted Values plot, where residuals are plotted against the predicted values. The residuals appear randomly scattered around the horizontal red line at $y = 0$, indicating that the relationship between the predictors and the response variable is approximately linear. While slight fanning at higher fitted values suggests potential heteroscedasticity, this does not violate linearity. A few outliers are present and may require further investigation. Overall, the plot confirms that the linearity assumption is satisfied.

Now check for the Condition 2. In the Figure 3, we choose all the numerical variables in train data. By the Figure 3, we can see that there is no linear pattern between these variables, so our Model 2 satisfies the condition 2.

A residual plot is a scatterplot that displays the residuals on the y-axis and the fitted values (predicted values) or another variable on the x-axis. It is used to evaluate the assumptions of a regression model, including linearity, homoscedasticity (constant variance), and independence

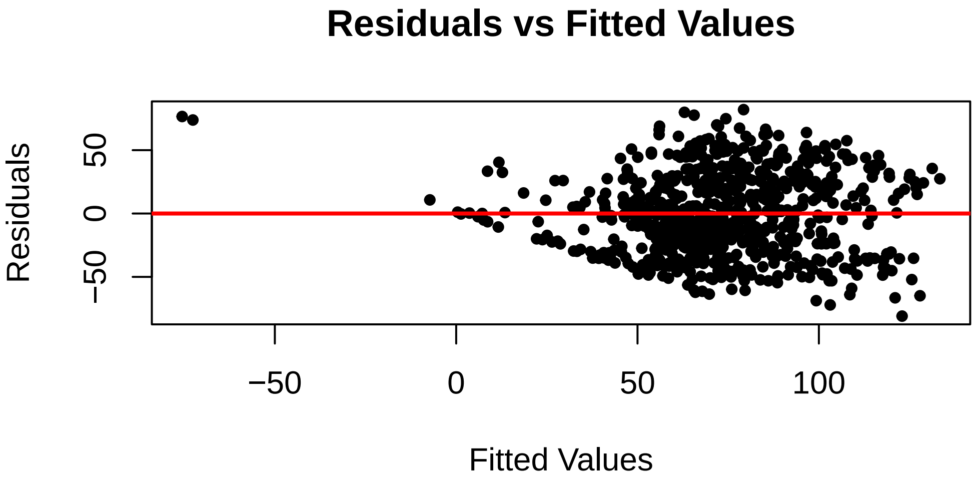

Figure 2: Plot for checking Condition 1 (Model 2)

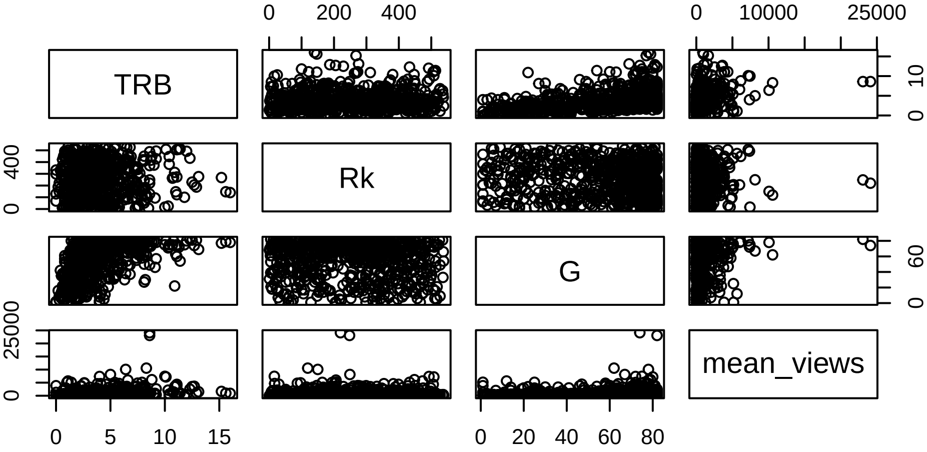

Figure 3: Plot for checking Condition 2 (Model 2)

of residuals. A good residual plot shows no clear patterns, clusters, or trends, indicating that the model's assumptions are likely met (James et al. 2013). The residual plot can check the linearity, homoscedasticity, normality, which are

- **Linearity**: Determines if there is a straight-line relationship between the `Score` and its predictors.
- **Homoscedasticity**: Checks whether the variance of the residuals remains consistent across all levels of the predictors.
- **Outliers**: Identifies any observations of the `Score` that differ substantially from the predicted values of `Score`.

The residual plot Figure 4 revealed no noticeable patterns, clusters, or evidence of heteroscedasticity, confirming that the assumptions of linearity, independence, and constant variance were met.

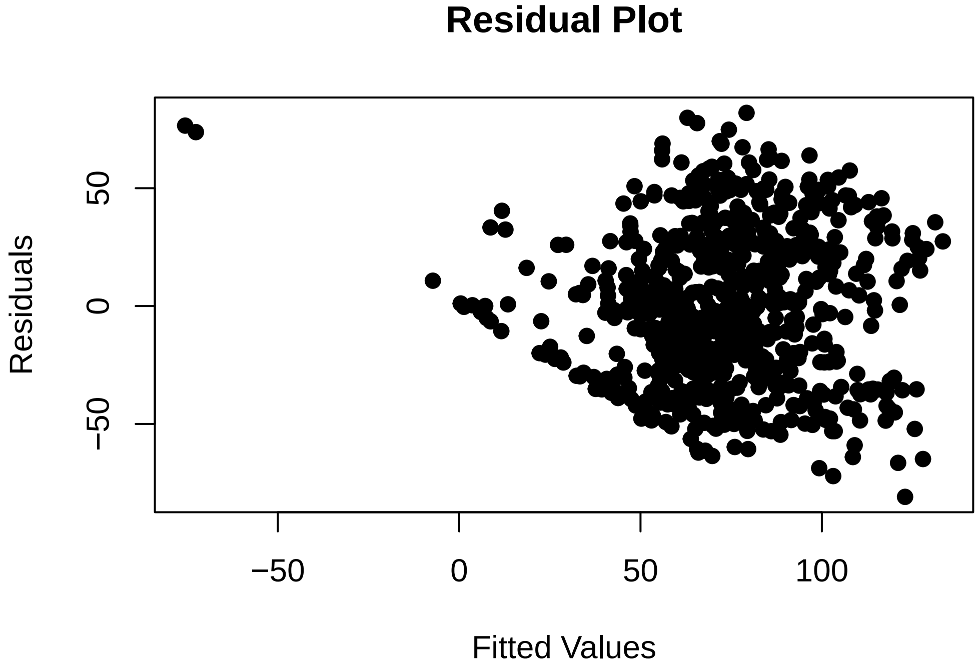

Figure 4: The Residual Plot of the Model 2

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a specified theoretical distribution, typically a normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the data follows the specified distribution, the points in the Q-Q plot will align closely along a 45-degree reference line. Deviations from this line indicate departures from the assumed distribution, such as skewness or heavy tails (Science 2024).

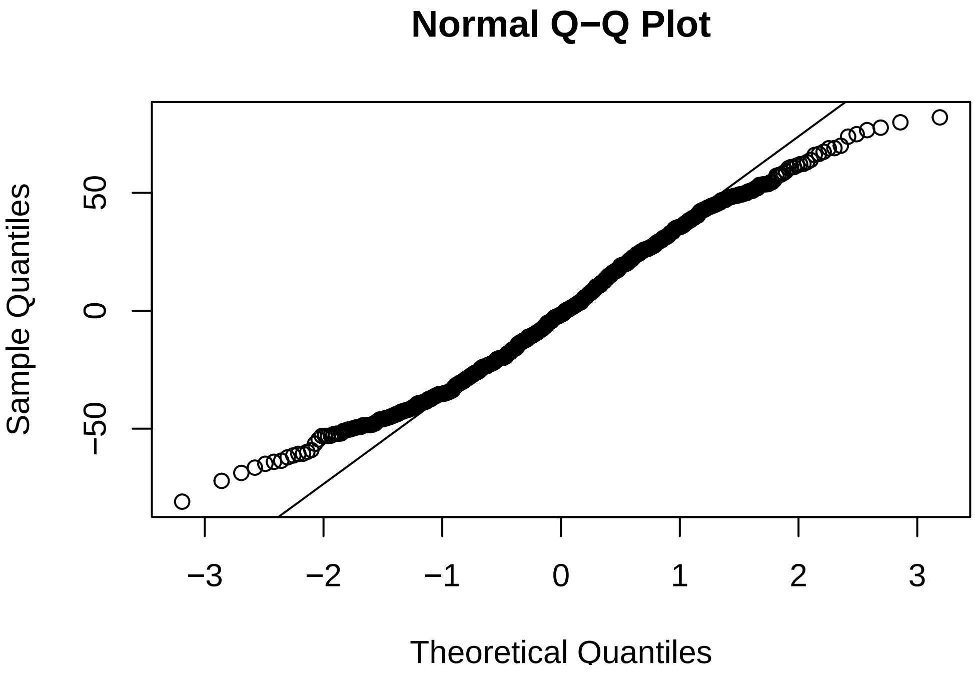

Figure 5: The QQ-Plot of the Model 2

The Q-Q plot Figure 5 demonstrated that the residuals aligned closely with the straight line, suggesting that they were approximately normally distributed.

4.2 Model Validation

We use the test data and the same predictors in the Model 2 to fit our Model 4, the Table 3 shows the coefficient compare table of the Model 2 and the Model 4. we can see that there is slightly difference between the coefficients of the predictors, and all the predictors in the Model 4 are significant. This means our model can work well even if we fit it on the different datasets, means our coefficients are stable, here is the Model 4:

$$\begin{aligned}\widehat{\text{Score}} = {} & 99.6246334 - 5.8121747(\text{TRB}) + 0.0355633(\text{Rk}) + 25.6532125(\text{RoleFront}) \\ & - 0.4656068(\text{G}) - 0.0034080(\text{mean\_views})\end{aligned}$$

Table 3: Summary of the coefficients and metrics for the training and testing models

	Variable	Model 2 Estimate (Train)	Model 4 Estimate (Test)
(Intercept)	(Intercept)	103.4738121	99.6246334
TRB	TRB	-5.1609752	-5.8121747

	Variable	Model 2 Estimate (Train)	Model 4 Estimate (Test)
Rk	Rk	0.0316097	0.0355633
RoleFront	RoleFront	21.7089780	25.6532125
G	G	-0.4479400	-0.4656068
mean_views	mean_views	-0.0053988	-0.0034080

5 Conclusion

This paper uses linear regression modeling to explore the relationship between NBA player performance indicators and MVP voting scores. Key predictors, such as total rebounds, rankings, roles, games, and average viewpoints, were identified as significant factors influencing MVP voting results. The study utilized a robust methodology that included data cleaning, training and test dataset segmentation, and model validation. Both residual and diagnostic plots confirmed that the model satisfied the assumptions of linear regression, including linearity, residual normality, and independence.

The stability and generalizability of the final model is demonstrated by consistent results on both the training and test datasets. Significant predictors in the training model remained valid in the test model, enhancing the reliability of the model. In addition, the partial f-test validates the necessity of including key predictors and rejects any simplification of the final model.

While these results provide valuable insights into the factors that influence MVP voting, there are some limitations. The dataset only covers the 2016-2019 seasons and may not fully capture changes in player evaluation criteria over time. Additionally, the subjective nature of MVP voting introduces an element of variability that cannot be fully explained by statistical models. Future research could expand the dataset to include more seasons and explore nonlinear relationships or machine learning techniques to enhance predictive power.

Overall, this study highlights the importance of advanced analytics in understanding sports decision-making and contributes to the broader field of basketball performance assessment.

Reference

Auguie, Baptiste. 2023. "gridExtra: Miscellaneous Functions for "Grid" Graphics." https://CRAN.R-project.org/package=gridExtra.

contributors, Wikipedia. 2024a. "Multicollinearity." https://en.wikipedia.org/wiki/Multicollinearity.

———. 2024b. "Variance Inflation Factor." https://en.wikipedia.org/wiki/Variance_inflation_factor.

Fox, John, and Sanford Weisberg. 2023. "Car: Companion to Applied Regression." https://CRAN.R-project.org/package=car.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in r*. New York, NY: Springer. https://www.statlearning.com/.

Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. 2012. *Introduction to Linear Regression Analysis*. 5th ed. Hoboken, NJ: Wiley. https://www.wiley.com/en-us/Introduction+to+Linear+Regression+Analysis%2C+5th+Edition-p-9780470542811.

Müller, Kirill, and Jennifer Bryan. 2023. "Here: A Simpler Way to Find Your Files." https://CRAN.R-project.org/package=here.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Science, Penn State Eberly College of. 2024. "Normal Probability Plots (q-q Plots)." https://online.stat.psu.edu/stat501/lesson/7/7.2.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. "Dplyr: A Grammar of Data Manipulation." https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2023. "Knitr: A General-Purpose Package for Dynamic Report Generation in r." https://CRAN.R-project.org/package=knitr.

Zhu, Hao. 2023. "kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax." https://CRAN.R-project.org/package=kableExtra.