

NBA Salaries: How Much is that Buzzer Beater Worth?

Eddie and Co.

Introduction

Professional sports salaries, particularly in the NBA, continue to spark discussion because they highlight the intersection of performance, marketability, and team strategy. Understanding how a player’s on-court production translates to salary is not only valuable for analysts and researchers, but also for front offices making contract decisions and fans seeking to understand the economics of the league. Prior research suggests that player compensation is strongly influenced by both measurable performance statistics and more subjective forms of recognition such as awards or team honors (Berri & Schmidt, 2006; Stiroh, 2007; Simmons & Berri, 2011).

Motivated by these findings, our project examines the statistical relationships between NBA player characteristics and their salaries during the 2024-2025 season. We selected three primary variables—points per game, defensive stocks (steals + blocks), and award recognition—because each represents a different dimension of player value. Scoring is one of the most visible and heavily rewarded skills in basketball (Stiroh, 2007). Defensive performance, while often overlooked publicly, provides meaningful team impact and has shown measurable influence on team success (Simmons & Berri, 2011). Awards represent league-wide acknowledgment of excellence and can shape a player’s reputation and contract negotiations (Berri & Schmidt, 2006).

Our general research question is:

What factors best explain variation in NBA player salary during the 2024-2025 season?

To address this, we developed three specific research questions:

Do NBA players who score more points (PTS) earn higher salaries on average?

Is there a meaningful positive relationship between a player’s defensive activity (“stocks”) and their salary?

Do players who have earned at least one major NBA award receive higher salaries compared to players with no awards?

Together, these questions reflect distinct aspects of performance and recognition. By analyzing these variables, we aim to identify which factors are most strongly associated with salary and how our findings compare to patterns documented in the sports-economics literature.

Data Summary

Data Sources

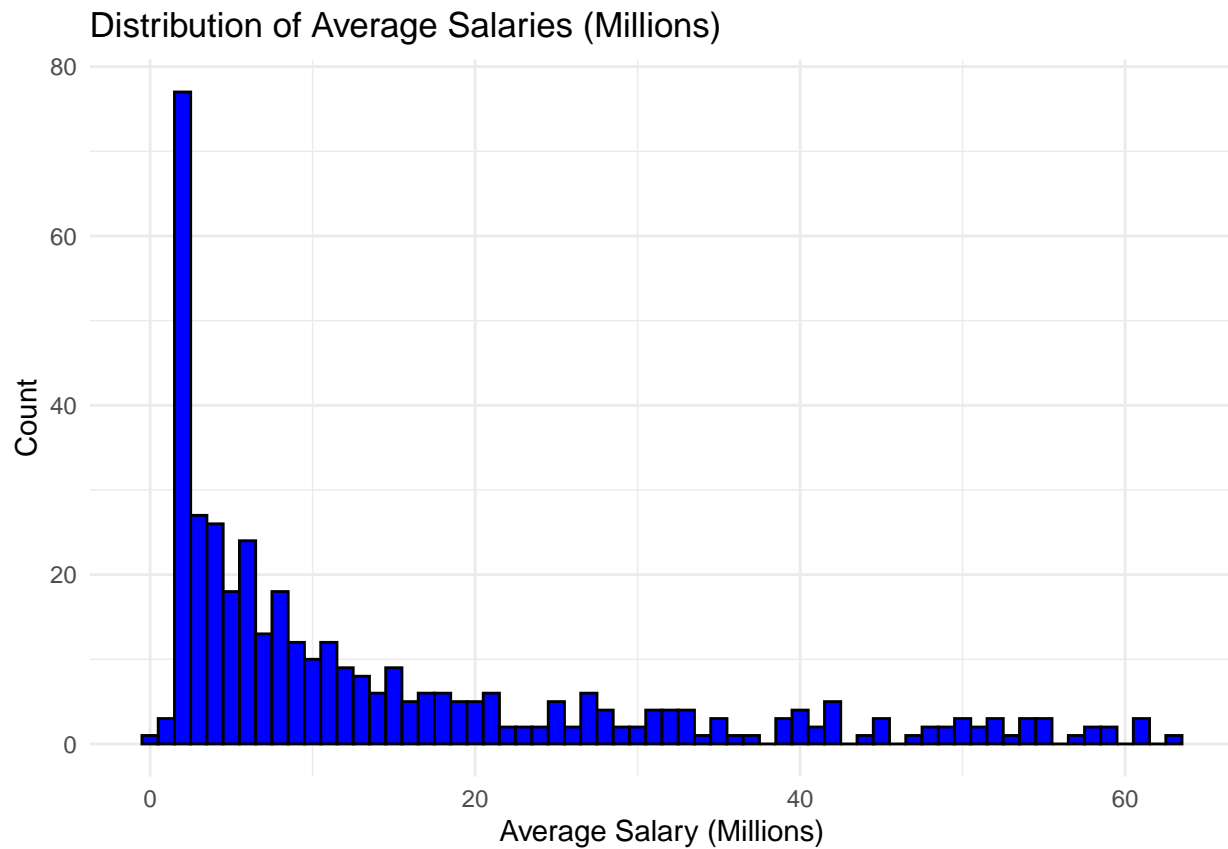
The data in this project were compiled from publicly available and reputable basketball statistics platforms. Performance statistics, biographical information, and award histories were collected from Basketball Reference, a database known for its accuracy, professional documentation, and long-standing use in academic sports analysis. Together, these sources provide a complete picture of each player's performance and compensation.

To prepare the data for analysis, we standardized naming conventions across the two sources and combined the datasets so that each player occupied a single row with salary, performance, and awards information. Awards were originally stored in long strings listing placements, team selections, and voting outcomes, so we simplified these into a three-level qualitative variable representing the general number of accolades received. Age was also simplified into qualitative brackets to match our research questions.

Overall, the combined dataset contains a rich mixture of performance and recognition indicators, allowing us to meaningfully evaluate how both quantitative and qualitative factors relate to NBA salary. While subjective elements such as awards voting and contract timing may influence individual observations, the breadth and reliability of the sources allow us to confidently proceed with statistical analysis.

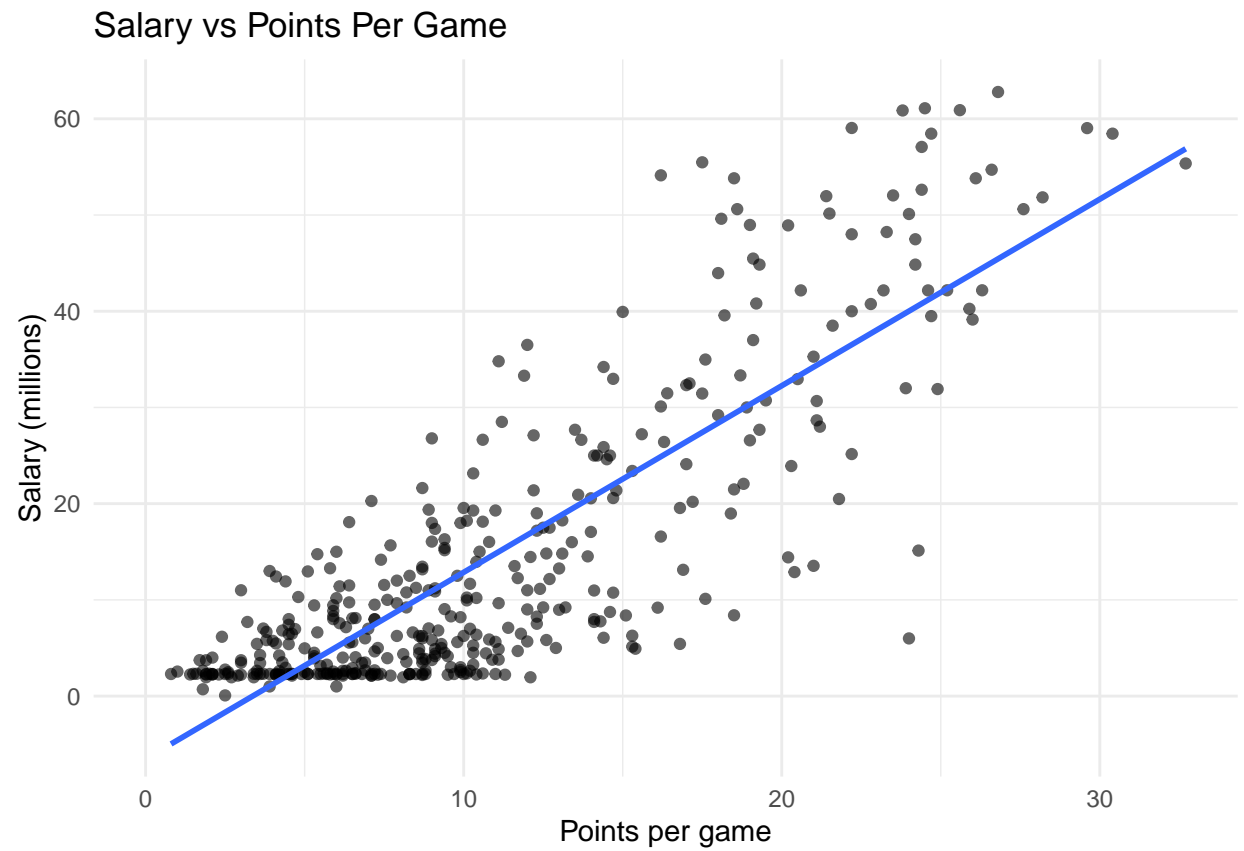
Exploratory Data Analysis

Warning: Removed 172 rows containing non-finite outside the scale range (``stat_bin()``).



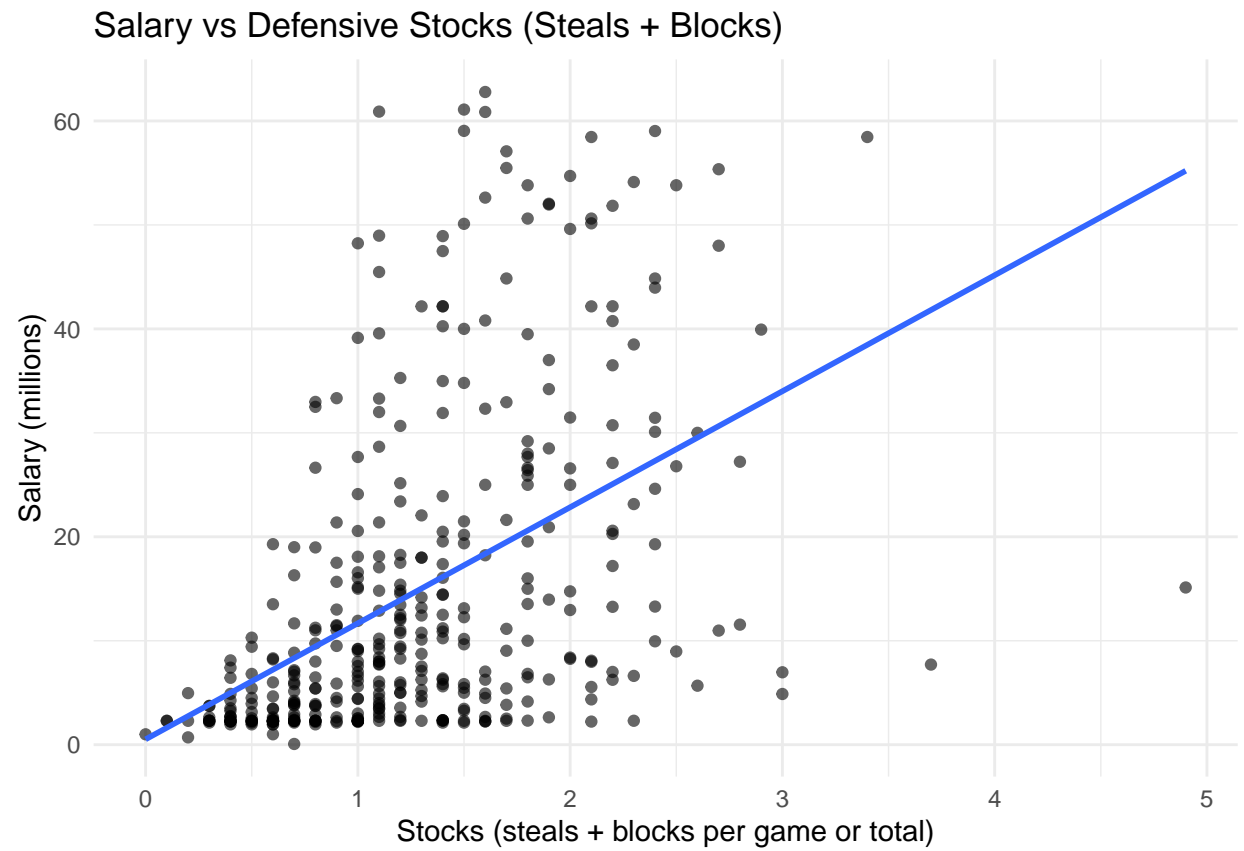
Warning: Removed 172 rows containing non-finite outside the scale range (``stat_smooth()``).

Warning: Removed 172 rows containing missing values or values outside the scale range (``geom_point()``).

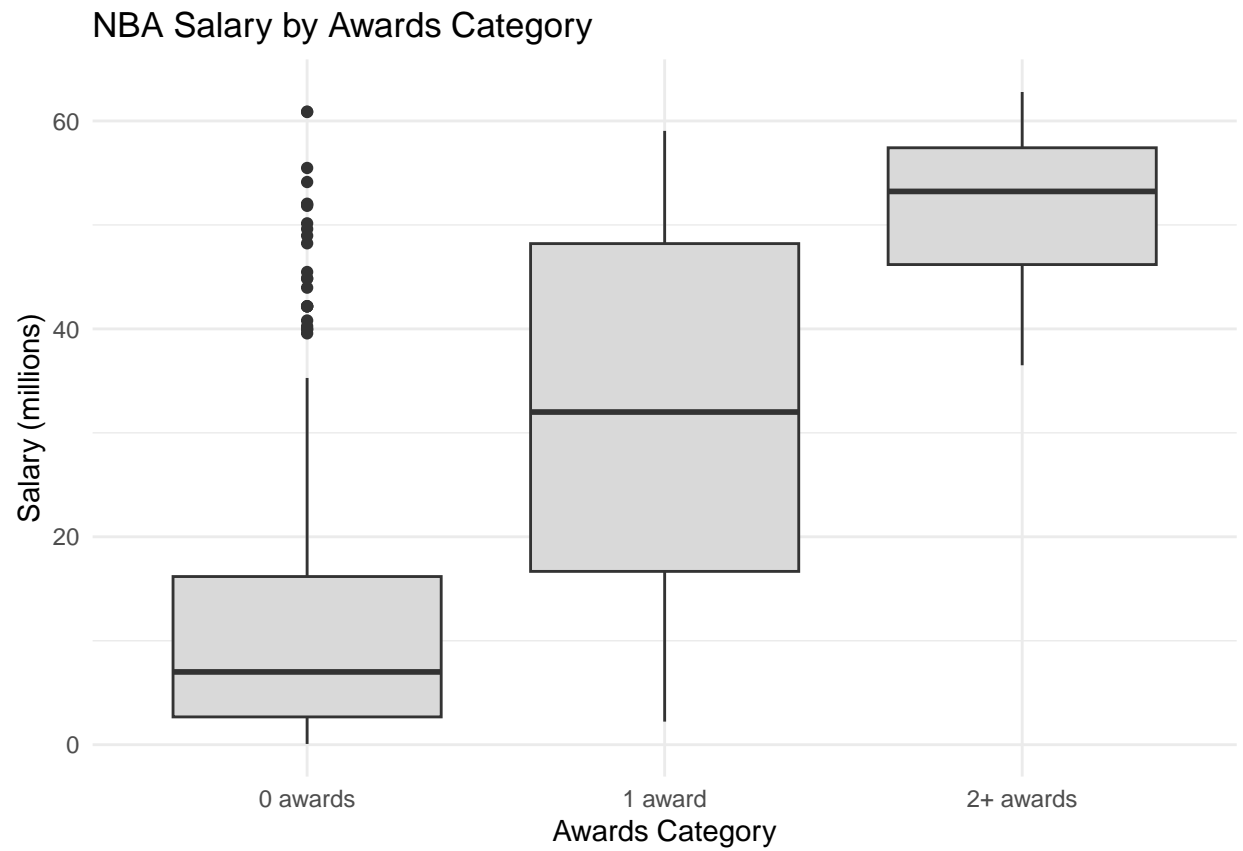


Warning: Removed 172 rows containing non-finite outside the scale range
(``stat_smooth()``).

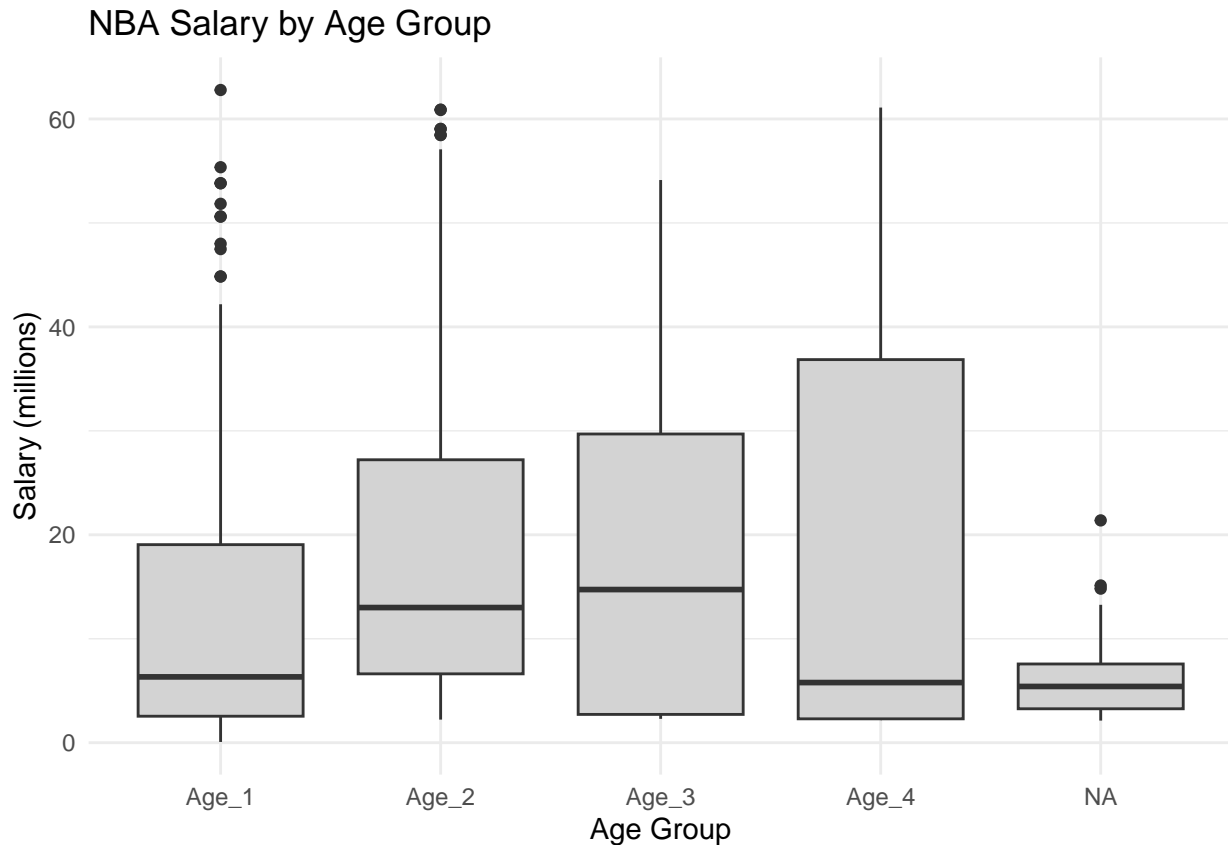
Removed 172 rows containing missing values or values outside the scale range
(``geom_point()``).



Warning: Removed 172 rows containing non-finite outside the scale range
(`stat_boxplot()`).



Warning: Removed 172 rows containing non-finite outside the scale range (`stat_boxplot()`).



EDA Summary

To explore how performance and recognition relate to NBA salaries, we began by examining the distribution of our response variable, average salary in millions of dollars. The histogram shows a heavily right-skewed pattern, with most players earning under \$10 million and a small number of superstars earning between \$40–\$60 million. Numerically, the median salary is approximately \$5.4 million, while the mean is \$9.0 million, reflecting the effect of these extremely high-earnings players on the average. Despite the skewness, salary is a continuous variable with meaningful spread, making it suitable for multiple linear regression. If residual diagnostics later indicate violations of normality assumptions, we may consider transformations such as the logarithm of salary.

We next investigated how scoring performance relates to salary. The scatterplot of salary vs points per game (PTS) displays a strong positive trend, supported by a steep regression line. Players scoring 20 or more points per game typically earn \$35–\$45 million, while those under 5 points often earn below \$5 million. This aligns with economic expectations, as scoring is a highly valued offensive skill in the modern NBA.

Defensive performance, measured through stocks (steals + blocks), also shows a positive

association with salary, although the trend is more moderate. Players averaging two or more stocks tend to cluster around \$20–\$40 million, whereas those producing under one stock per game tend to earn much less. This suggests that defensive impact contributes to player valuation, but not as strongly as scoring.

We also examined how award recognition relates to salary. The boxplot comparing players with “0 awards,” “1 award,” and “2+ awards” shows clear separation between the three groups. Players with no awards have a median salary of about \$8 million, while those with one award have a median closer to \$50 million. Players with two or more awards earn slightly more, with a median of roughly \$55 million. These differences imply that league-wide recognition corresponds with significantly higher pay and may capture aspects of player reputation that go beyond raw statistics.

Finally, to understand salary variation across career stages, we examined a boxplot of salary by age group. Players in the 27–31 age range earn the highest median salaries, reflecting the common peak-performance window in NBA careers. Younger players (22–26) and older veterans (35+) tend to earn less on average, which aligns with typical player development and decline patterns. This demographic plot adds context for interpreting performance and award variables within the broader structure of player career arcs.

Before moving to regression modeling, we considered the possibility of multicollinearity among predictors. While scoring (PTS) may correlate with other offensive variables, our selected predictors—PTS, stocks, Awards, and Age—do not appear strongly correlated with one another based on preliminary inspection. This reduces concerns about inflated standard errors, which we will confirm formally using Variance Inflation Factors (VIFs) in the modeling stage.

Overall, the EDA reveals consistent and meaningful trends:

Salary is right-skewed with a large upper tail of superstar contracts.

Scoring shows a strong positive relationship with salary.

Defensive stocks have a moderate positive association with salary.

Award recognition corresponds with substantial salary increases.

Salary peaks in mid-career age groups, aligning with performance arcs.

These patterns justify inclusion of the selected predictors and provide a clear foundation for developing the multiple linear regression model.

Methods and Analysis

Assessing Multicollinearity

We will begin our analysis stage by looking for multicollinearity between predictors. In our exploratory data analysis, we identified that there may be high multicollinearity between Pts and other predictors related to offensive statistics, like FTA and MP. After examining our correlation heatmap from EDA, we will need to further examine pairwise relationships and VIFs to identify potential hazards in our model.

Our first step will be to acquire numerical information regarding pairwise relationships. That is, identify the correlation between our explanatory variables. We will then analyze multicollinearity through VIF analysis. If individual VIFs are below the threshold of 10, and the average VIF is below 3, then multicollinearity is not a concern and we will move onto model building. However, in the case that multicollinearity is a concern, we will engage in variable screening to identify which predictors may not be necessarily impactful in this model. Using Stepwise Regression, we will identify a set of predictors that may be adequate at predicting Salary. With our new subset, we can reassess multicollinearity and decide which predictors to move forward with.

Building the Model

We will begin building our model by estimating our model parameters, starting with only our quantitative predictors and a quantitative interaction term that we believe may contribute to the model. Our initial model using quantitative predictors is as follows:

$$\begin{aligned}\text{Salary} = & \beta_1 G + \beta_2 MP + \beta_3 PTS + \beta_4 FG_ \% + \beta_5 FTA \\ & + \beta_6 TRB + \beta_7 AST + \beta_8 \text{Stocks} + \beta_9 \text{Value} \\ & + \beta_{10}(PTS \times FTA)\end{aligned}$$

After estimating this Multiple Linear Regression model using the `lm()` function, we will first evaluate its utility using a Global F-Test. Assuming the model is adequate at predicting Salary, we will move forward by testing our most important predictors using individual T-tests. The first T-test we will perform is for our interaction, PTS*FTA. If this interaction is significant at a predetermined alpha (.05), we will keep it in the model and continue with our T-tests. Our other variables of interest, namely PTS and Stocks, will then be tested to ensure they are significant at predicting salary. If any variables are largely insignificant and we believe they may not be a strong determinator of predicted salary, we will remove them from the model and return to a Global F-Test with the updated model.

Following our evaluation of the model with quantitative predictors, we will add in our qualitative predictors, and a Qualitative x Qualitative interaction term we believe may contribute to the model: Awards*Age

After adding our qualitative variables into the model, we will perform a Global F-Test to ensure the model is adequate at predicting Salary. We will then move into the necessary tests to evaluate individual predictors and interactions. We will begin by testing the interaction between Awards and Age with a T-test. If the interaction is significant, we will keep it in the model and continue with our analysis. We will then test our qualitative variable of interest, awards, using a Nested F-Test. If awards are significant at predicting Salary, we will keep it in the model and move forward.

We will not be testing any Quantitative X Qualitative Interactions.

After we have our final model achieved through testing, we will once again evaluate the utility of the model through a Global F-Test to ensure the model is adequate at predicting Salary.

Assessing the Model (Including Cross Validation)

Once we've finalized our model through variable testing, we will conduct a holistic assessment of our model for a better understanding of its capabilities and limitations. We will evaluate metrics such as R², Adjusted R², Root-MSE to evaluate our model. Separately, we will interpret confidence and prediction intervals and ensure our model performs on a practical level.

In addition to the traditional assessment metrics, we will be incorporating Cross Validation into our final project. Cross validation is a machine learning technique used to assess the performance and accuracy of a model. Performing Cross validation allows the model builder to see whether their model overfits or underfits by training the majority of a data set, and using the remaining data for testing. The trained set is then used to estimate a linear regression model evaluated on metrics such as Root-MSE and R². More specifically, we will use the Validation Set Approach in R, which randomly splits 80% of the dataset into training, and 20% into testing. By performing Cross Validation we will have a strong idea of whether our model is prepared to be used on new data, or if further adjustments must be made.

Checking Model Assumptions

Our last analysis step will be checking the model assumptions through residual analysis. The four assumptions we'll be checking are the Lack of Fit, Constant Variance, Normality, and Independence. Through examination of residual plots, residual vs. fitted plots, QQ plots, and histograms of the residuals we'll be able to tell whether any of these assumptions have been

violated.

We will then identify outliers and influential observations through Cooks Distance, Leverage, Studentized Residuals, and Deleted Studentized Residuals. If there are any influential observations that may be jeopardizing our model, we will consider removing them from the dataset.

Results

Conclusions

Appendix A: Data Dictionary

Variable Name	Abbreviated Name	Description
Games	G	This represents the games played by each respective player.
Minutes Played	MP	This represents the minutes played by each respective player in the overall season.
Points Scored	PTS	This represents the total points scored by each respective player.
Field Goal Percentage	FG_PCT	This represents the percentage of field goals scored by each respective player.
Free Throw Attempts	FTA	This represents the free throw attempts made by each respective player.
Total Rebounds	TRB	This represents the total rebounds made by each respective player.
Total Assists	AST	This represents the total assits by each respective player.
Steals and Blocks Combined	Stocks	This represents the steals and blocks achieved by each respective player.
Franchise Value	Value	This represents the franchise value of the team of each respective player.

Appendix B: Data Rows

	Player	Team	G	MP	FG_pct	FTA	TRB	AST	stocks	PTS
1	Shai Gilgeous-Alexander	OKC	76	34.2	0.519	8.8	5.0	6.4	2.7	32.7
2	Giannis Antetokounmpo	MIL	67	34.2	0.601	10.6	11.9	6.5	2.1	30.4
3	Nikola Jokić	DEN	70	36.7	0.576	6.4	12.7	10.2	2.4	29.6
4	Luka Dončić	2TM	50	35.4	0.450	7.9	8.2	7.7	2.2	28.2
5	Anthony Edwards	MIN	79	36.3	0.447	6.3	5.7	4.5	1.8	27.6
6	Jayson Tatum	BOS	72	36.4	0.452	6.1	8.7	6.0	1.6	26.8
	Value_Billions	awards_1	awards_2plus		avg_salary_millions			Age_22_26		Age_27_31
1	4.35	0		1			55.3591		1	0
2	4.30	0		1			58.4566		0	1
3	4.60	0		1			59.0331		0	1
4	NA	0		0			51.8379		1	0
5	3.60	0		1			50.6117		1	0
6	6.70	0		1			62.7867		1	0
	Age_32_34	Age_35_plus	Pos_PF	Pos_PG	Pos_SF	Pos_SG	Age		Awards	
1	0	0	0	1	0	0	Age_1	2+	awards	
2	0	0	1	0	0	0	Age_2	2+	awards	
3	0	0	0	0	0	0	Age_2	2+	awards	
4	0	0	0	1	0	0	Age_1	0	awards	
5	0	0	0	0	0	1	Age_1	2+	awards	
6	0	0	1	0	0	0	Age_1	2+	awards	

Appendix C: Final Model Output and Plots

Appendix D: References

Background

Data Sources

Additional Help