NBA Salaries: How Much is that Buzzer Beater Worth?

Eddie and Co.

Introduction

Professional sports salaries, especially in the NBA, are a topic of continuous public debate because they sit at the intersection of performance, marketability, and organizational strategy. Understanding what truly drives a player's salary is valuable not only to analysts and researchers, but also to teams making contract decisions and fans trying to understand the economics behind the sport. Prior studies suggest that both quantitative performance metrics and qualitative recognition, such as awards or team roles, significantly influence earnings. Motivated by this, our research explores the statistical relationships between player attributes and their salaries.

Our general research question is: What factors best explain variation in NBA player salary?

From this, we developed three specific research questions, each tied to a chosen predictor and one response variable. The response variable in all three questions is NBA player salary.

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

Quantitative Predictor: Is there a significant relationship between a player's total defensive contributions (steals + blocks, "stocks") and their salary?

Qualitative Predictor: Do players who have won at least one major NBA award (e.g., MVP, All-NBA Team selections) earn higher salaries than those who have not?

Overall, this project is relevant because salary determination is central to team strategy, player valuation, and league-wide competitive balance. By analyzing both quantitative and qualitative predictors, we aim to identify which factors hold the strongest statistical relationship with NBA salary and how these insights compare to findings in the sports economics literature.

Data Summary

Data Sources

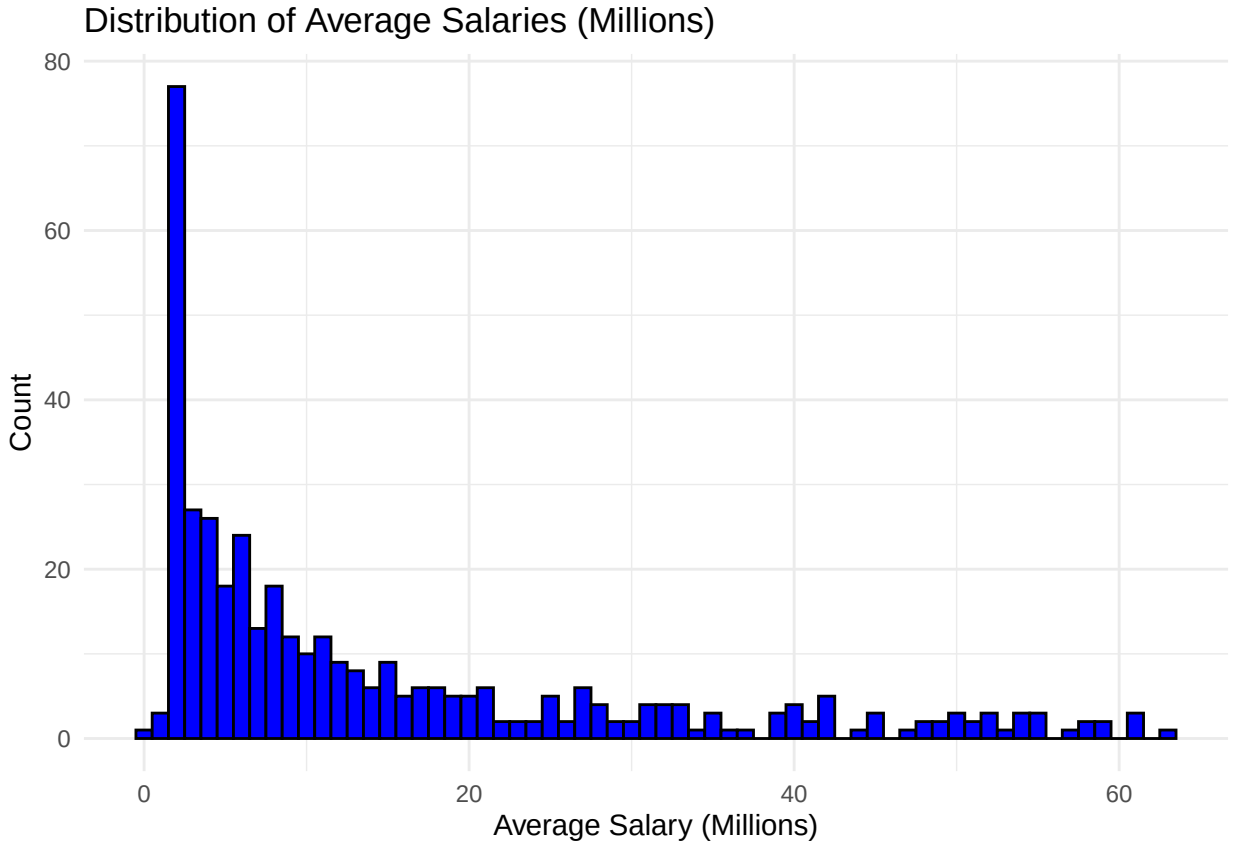
The data in this project were compiled from publicly available and reputable basketball statistics platforms. Performance statistics, biographical information, and award histories were collected from Basketball Reference, a database known for its accuracy, professional documentation, and long-standing 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

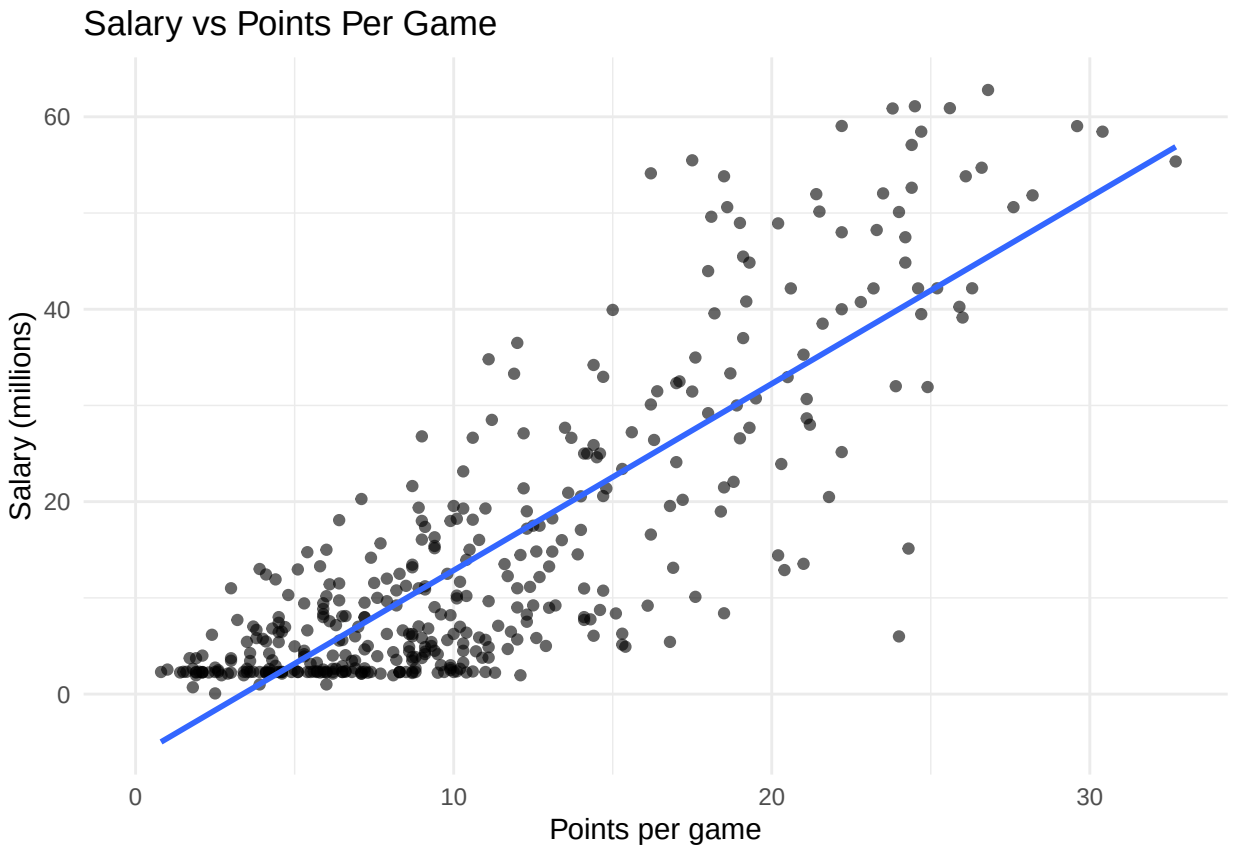
and combined the datasets so that each player occupied a single row with salary, performance, and awards information. Awards were originally stored in long strings listing placements, team selections, and voting outcomes, so we simplified these into a three-level qualitative variable representing the 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

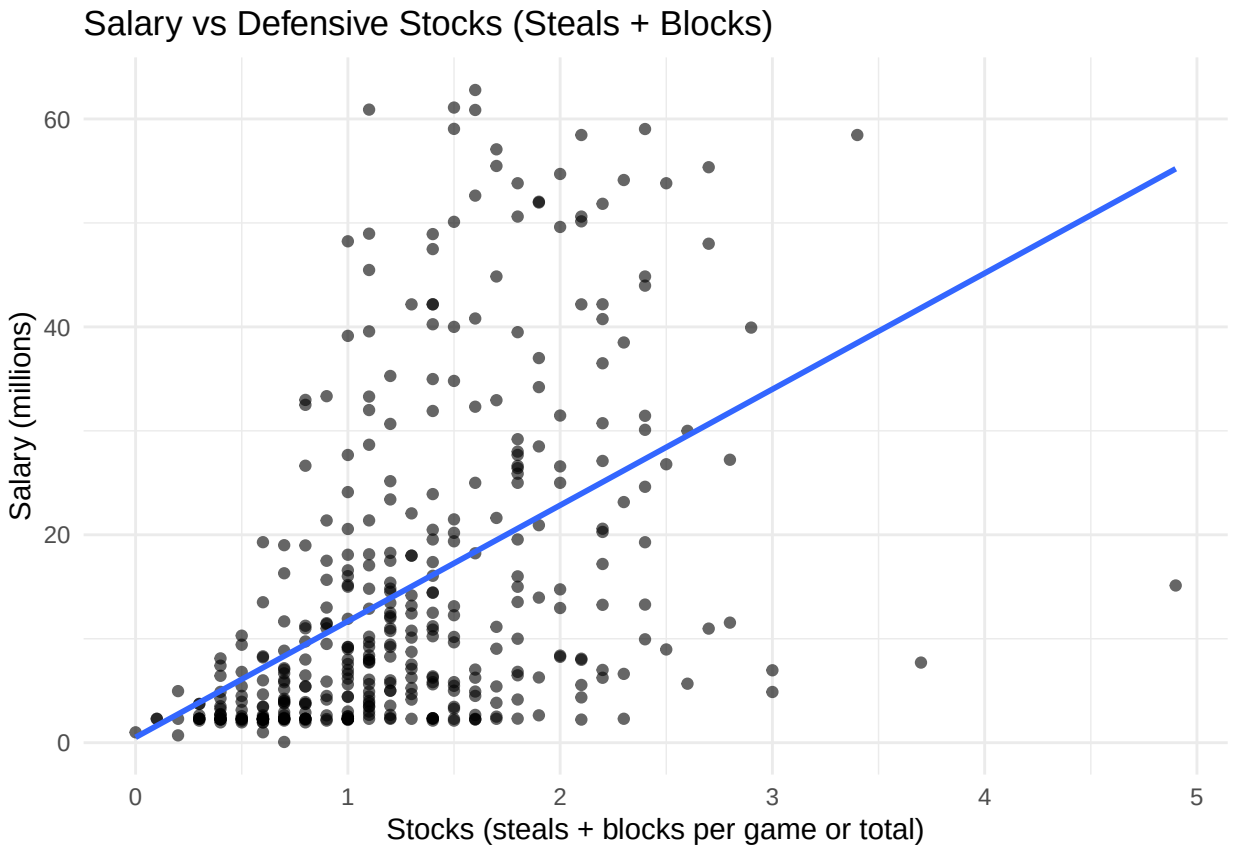
```
Warning: Removed 172 rows containing non-finite outside the scale range
(`stat_bin()`).
```



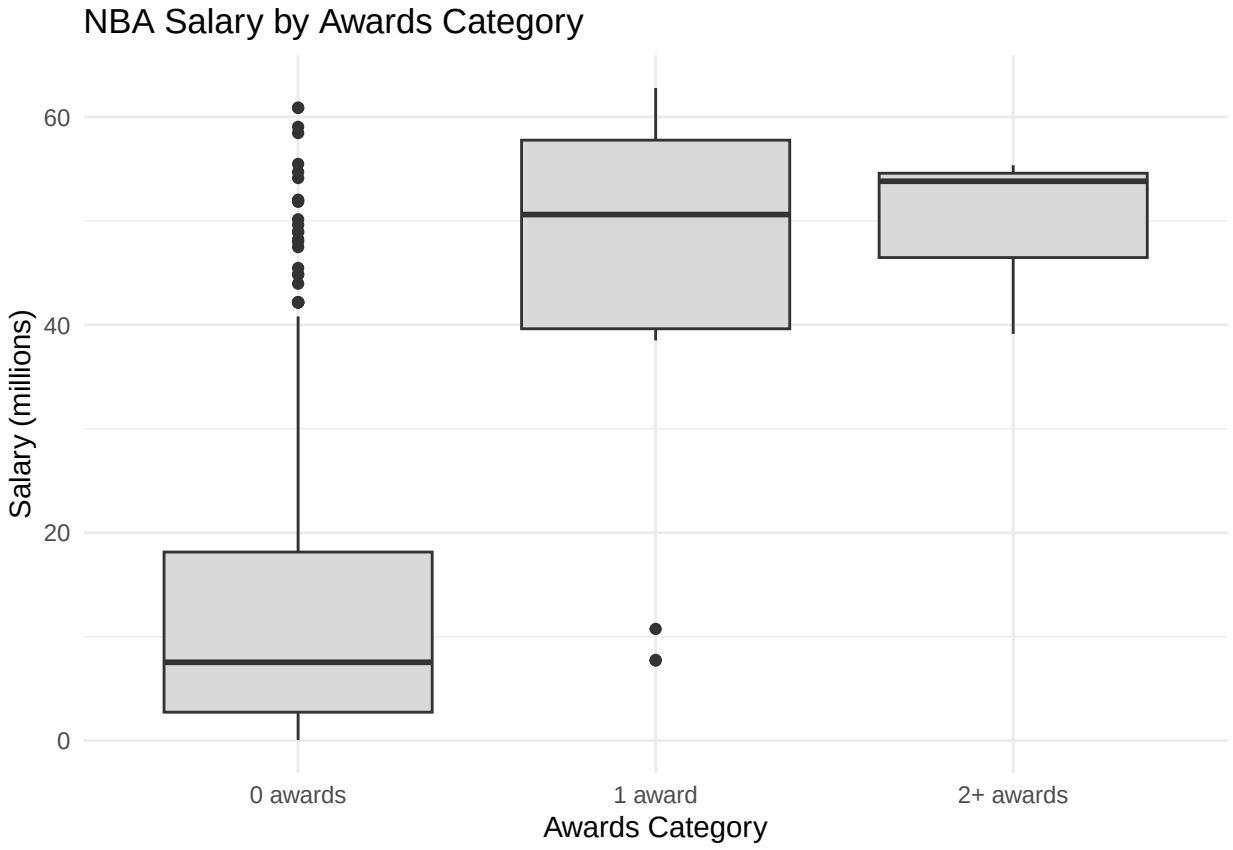

```
Warning: Removed 172 rows containing non-finite outside the scale range
(`stat_smooth()`).

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()`).
```



EDA Summary

To understand how performance and recognition relate to NBA player salaries, we began by examining the distribution of our response variable, average salary (in millions of dollars). The histogram shows that salaries are heavily right-skewed, with the majority of players earning under $10 million per year. Numerically, the median salary in our sample is approximately $5.4 million, while the mean salary is higher at $9.0 million, reflecting a small group of extremely high-earning players who pull the average upward. This skewed structure is expected in the NBA, where superstars sign "max contracts" much larger than the salaries of role players. Importantly, salary is a continuous quantitative variable, making it appropriate as a response variable for multiple linear regression. Although the data are skewed, the linear model can still be interpreted meaningfully; however, if model assumptions are violated later, we may consider transformations such as logging salary.

Relationship Between Scoring and Salary A scatterplot of salary versus points per game (PTS) shows a clear positive trend: players who score more tend to earn more. The fitted regression line rises steeply, suggesting scoring is strongly associated with higher pay. Numerically, high scorers (20+ PPG) earn an average of $35–$45 million, while low scorers (under 5 PPG)

typically earn below $5 million. This relationship aligns with economic expectations. Scoring is perhaps the most visible and rewarded skill in basketball, and our EDA indicates it is likely to be a significant predictor in the regression model.

Relationship Between Defensive Stocks and Salary We also explored the connection between salary and defensive activity, measured by "stocks" (steals + blocks). The scatterplot shows a weaker but still positive trend. On average, players who average 2 or more stocks tend to earn $20–$40 million, while those with fewer than 1 stock per game cluster near lower salaries. The numerical spread suggests that although defense contributes to player value, it may not influence salaries as strongly as scoring, which is consistent with league-wide patterns where offensive stars typically command the highest contracts.

Salary Differences Based on Awards To evaluate whether league recognition affects earnings, we created a categorical variable Awards with three levels: "0 awards," "1 award," and "2+ awards." The boxplot shows clear separation among these groups. Players with 0 awards have a median salary of about $8 million, while players with 1 award have a median salary of roughly $50 million. Those with 2+ awards earn even more on average, with a median close to $55 million. These numerical patterns strongly suggest that award recognition, representing public and league-wide acknowledgment of excellence, is associated with substantially higher pay.

The broad salary range within each group also highlights the importance of including additional variables, since recognition alone does not explain all variation. Still, the clear upward shift in salary distributions across the award categories implies that awards are likely to be an important qualitative predictor.

Suitability of Variables and Consideration of Multicollinearity Before finalizing our regression model, we assessed potential multicollinearity among predictors. Scoring (PTS) and some other offensive statistics (not included in our model) can be highly correlated, which may inflate standard errors if included simultaneously. However, in our chosen model, stocks and awards do not appear strongly correlated with scoring, meaning multicollinearity should be minimal. Variance Inflation Factor (VIF) checks later in the analysis will confirm this formally.

All selected predictors: PTS, stocks, and Awards, show meaningful variation and substantive relationships with salary. The EDA supports the idea that each variable ties directly to one of our research questions and is suitable for inclusion in a 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:

$$Salary = Beta_1(G) + Beta_2(MP) + Beta_3(PTS) + Beta_4(FG_pct) + Beta_5(FTA) + Beta_6(TRB) + Beta$$
$$Beta_{10}(PTS * FTA)$$

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

Appendix C: Final Model Output and Plots

Appendix D: References

Background

Data Sources

Additional Help