Exploring the Relationship Between NBA Salaries and player performance metrics on NBA players Between 2016 and 2019

STA302H1: Methods of Data Analysis I

University of Toronto

Dec 6th 2024

Contributions

Youyu Fu: *Code, Method, Results, Conclusion, Editing, Editing Demonstration*

Ryan Li: *Introduction, Results, Poster, Editing, Editing Demonstration*

David Lee: *Methods, Results, Conclusion, Poster, Editing*

1. Introduction

This report explores factors influencing NBA player salaries, a topic of interest to fans, economists, and analysts seeking to understand how player attributes affect earnings. Metrics such as points, assists, rebounds, age, position, and popularity are examined for their impact on salaries. Insights from this analysis help NBA teams balance investments in star players with maintaining competitive performance. The study provides actionable information to guide financial decisions.

Our research question is: **Among performance metrics like points per game, assists, rebounds, and factors such as age, position, and Wikipedia views, which predictors most strongly influence an NBA player's yearly salary, positively or negatively?**

We reviewed three peer-reviewed studies. Lyons, Newton Jackson Jr., and Livingston (2018) identified points per game and field goal percentage as the most influential factors, with rebounds and assists playing secondary roles. Sarlis and Tjortjis (2024) found that salaries peak between ages 29-34 and decrease with injuries. They also found that centers earn the most, while guards earn the least. Louivion and Pettersson (2017) emphasized assists and team efficiency, noting that while scoring is critical, playmaking and plus-minus averages also shape salaries.

We used linear regression to address the research question because it quantifies and interprets the relationships between predictors and the response variable through predictor coefficients. Linear regression allows us to evaluate the direction and strength of each predictor's influence on player salaries and identify statistically significant predictors. Additionally, it helps determine which performance metrics NBA teams should prioritize when evaluating player salaries, making it a robust tool for analyzing complex relationships in the data.

2. Method

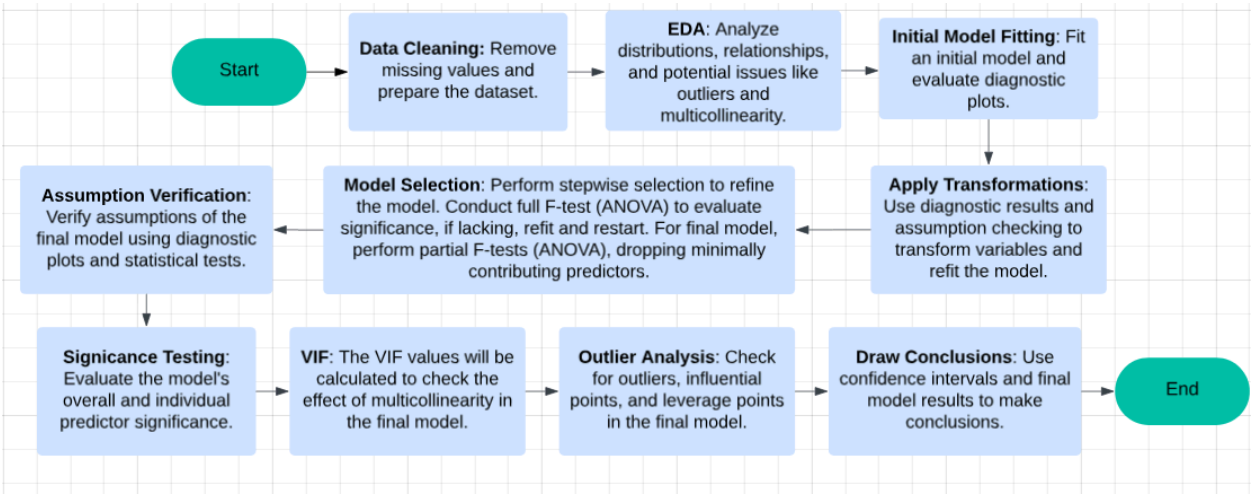
We will begin by cleaning the dataset, removing missing values from the selected variables, and encoding categorical variables for inclusion in the model. During exploratory data analysis (EDA), we will calculate summary statistics, including means, medians, standard deviations, and ranges for numerical variables. To examine the distribution and detect outliers, we will generate boxplots and histograms of the response variable. Scatterplots will visualize relationships between the response and predictors. Multicollinearity will be assessed by fitting models of each predictor against others and evaluating the R^2 score.

Next, we will fit an initial linear regression model using the predictors identified in the introduction. Diagnostic plots, including residual plots, Q-Q plots, and Cook's distance, will identify potential assumption violations such as non-linearity, heteroscedasticity, or non-normality of errors. If residual plots show a non-zero mean or non-linear patterns in

response-versus-predictor plots, we will identify a linearity violation in the independent error assumption. We will apply transformations, such as logarithmic or square root transformations, to individual predictors. If linearity remains violated, we will transform the response variable. If residual plots show clustering patterns or fanning, we will diagnose a violation of the constant variance assumption and apply transformations like Box-Cox or polynomial transformations to the response variable. Deviations from the 45-degree line in Q-Q plots will indicate a normality violation, prompting transformations of the response variable, such as the Box-Cox transformation. A transformed model will then be refitted using modified variables. Diagnostic plots will be rechecked iteratively until the issues are resolved, and the model achieves a satisfactory fit.

To evaluate the statistical significance of the transformed model, we will conduct a full F-test using ANOVA. If the model lacks significance, we will revise the predictors and refit the model. Partial F-tests will be performed to obtain a final model by removing minimally contributing predictors, improving interpretability and statistical significance. Assumptions of the final model will be reassessed using diagnostic plots and statistical tests. If violations persist, additional transformations will be applied. Individual predictors will be tested for statistical significance with t-tests, and VIF values will measure the effect of multicollinearity. Outliers, leverage points, and influential observations will be examined using scale-location plots, hat value plots, and Cook's distance. Their influence will be assessed, and their removal will be considered if necessary.

Finally, confidence intervals for model coefficients will be computed to evaluate the strength and direction of relationships between predictors and the response variable. These intervals will guide conclusions for the research question.



Simplified Flowchart for Methods

3. Results

3.1 Data cleaning, Summary Statistics, and EDA analysis

Players with missing predictor values were excluded, reducing the dataset from 1408 to 1205 observations. A summary table **(Appendix 1)** was generated for an overview of the data. A Salary boxplot and histogram **(Appendix 1)** revealed several outliers.

Plots of Salary against predictors showed non-linear patterns, particularly in Salary vs. Age and Salary vs. Mean Views. Salary vs. Age displayed a quadratic trend, peaking around 30 years, while Salary vs. Mean Views showed discrepancies between clusters and higher values.

Figure 1: Relationships Between NBA Player Salaries and Key Predictors.

To assess multicollinearity, we regressed Points per Game (PTS) on Age, Assists per Game (AST), Total Rebounds per Game (TRB), Primary Position (Pos1), and Mean Views. A multiple R-squared value of 0.7045 indicated high multicollinearity among these predictors.

3.2 Preliminary Model

We start with the preliminary model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 \mathrm{I}(\mathrm{Pos1}) + \beta_7 x_1 x_2 + \beta_8 x_5 \mathrm{I}(\mathrm{Pos1}) + \varepsilon$$

Where:

y = Salary (player salary)

x_1 = PTS (points per game average)

x_2 = Age (player age)

x_3 = AST (assists per game average)

x_4 = TRB (total rebounds per game average)

x_5 = mean_views (mean daily wikipedia views)

I(Pos1) = dummy variable for Pos1 (categorical, player starting position)

We created interaction terms to explore how NBA teams value specific attributes. The interaction between Age and Points aims to assess whether salary differences exist for players of different ages who average the same number of points, providing insights into how teams value potential. Similarly, the interaction between Position and Mean Views examines whether the effect of popularity, as measured by mean views, varies across player positions.

The Q-Q plot revealed a light-right-tail deviation from the 45-degree line, indicating a violation of the normality assumption for errors. The Residuals vs. Fitted plot showed a fanning pattern as fitted values increased, suggesting a violation of the constant variance assumption **(Appendix 2)**. This was further confirmed by the varying spread of points in the Residuals vs. Age and Residuals vs. Mean Views plots **(Appendix 3)**.

3.3 Transformations

To address these violations, we applied a Box-Cox transformation to the response variable, Salary, selecting a lambda value of 0.3. Age and Mean Views were also transformed. Observing a quadratic relationship between Age and Salary, we squared Age and introduced it as a new variable. To minimize the discrepancy between outliers and other values in the Salary vs. Mean Views relationship, we applied a logarithmic transformation to Mean Views.

The transformed model is given by:

Salary_transformed ~ Age_sq * PTS + AST + TRB + Pos1 * log_mean_views

The revised Q-Q plot aligns much more closely with the 45-degree line, indicating improved normality **(Appendix 2)**. The fanning pattern in the Residuals vs. Fitted plot has diminished **(Appendix 2)**, though a slight decrease in spread remains for fitted values above 400 or below 250. The Residuals vs. Age_sq plot **(Appendix 3)** shows a slight reduction in variance across different values of Age_sq. However, the constant variance assumption remains partially violated, as the spread still decreases for Age_sq values greater than 1000 or less than 600. The Residuals vs. Log Mean Views plot **(Appendix 3)** exhibits improved variance consistency compared to the Residuals vs. Mean Views plot, with less reduction in residual spread as Log Mean Views increase.

3.4 Full F-test

We conducted a full F-test using ANOVA to compare the null model, which included only the intercept, with the transformed model. The results indicated that at least one variable in the transformed model was significant, with an F-statistic of 107.78 ($p < 2.2\times10^{-16}$).

3.5 Partial F-tests

For model selection, we used partial F-tests, starting with the transformed model:

Salary_transformed ~ Age_sq * PTS + AST + TRB + Pos1 * log_mean_views

Reduced models were created by systematically dropping one variable at a time. ANOVA was then used to compute the p-value associated with dropping each variable, determining its significance in the model. The results for each dropped variable are summarized below.

Variable to drop	p-value	dropped
Age_sq:PTS	p = 7.515e-06	no
Pos1:log_ mean_views	p = 0.3813	yes
Pos1	p = 0.001582	no
log_mean_views	p = 0.0002622	no
TRB	p = 6.863e-07	no
AST	p = 1.447e-05	no
Age_sq	p < 2.2e-16	no

PTS	p < 2.2e-16	no

Figure 2: Predictor Significance and Retention in the Final Model.

As such, we selected the final model as:

Salary_transformed ~ Age_sq * PTS + AST + TRB + Pos1 + log_mean_views

3.6 Final Model

Based on the summary outputs, our final fitted model is

$$\hat{\text{Salary\_transformed}} = 125.278572 + 0.102745 \cdot \text{Age\_sq} + 1.505158 \cdot \text{PTS} + 9.694878 \cdot \text{AST} + 7.700196 \cdot \text{TRB}$$
$$+ 5.842177 \cdot \log(\text{mean\_views}) + 0.008243 \cdot (\text{Age\_sq} \cdot \text{PTS}) + \text{Pos1Effect}$$

Where:

$$\text{Pos1Effect} = \begin{cases} -12.698336 & \text{if Pos1} = \text{PF (Power Forward)} \\ -40.944888 & \text{if Pos1} = \text{PG (Point Guard)} \\ -9.080352 & \text{if Pos1} = \text{SF (Small Forward)} \\ -13.662659 & \text{if Pos1} = \text{SG (Shooting Guard)} \\ 0 & \text{if Pos1} = \text{C (Center; baseline category)} \end{cases}$$

The model has an adjusted R-squared of 0.551, suggesting that 55.1% of the variability in the response variable is explained by the model, considering the number of predictors used.

We rechecked the linear regression assumptions for the final model using diagnostic plots. These plots showed similar patterns to those of the transformed model, indicating no new violations. As a result, we proceeded with the final model.

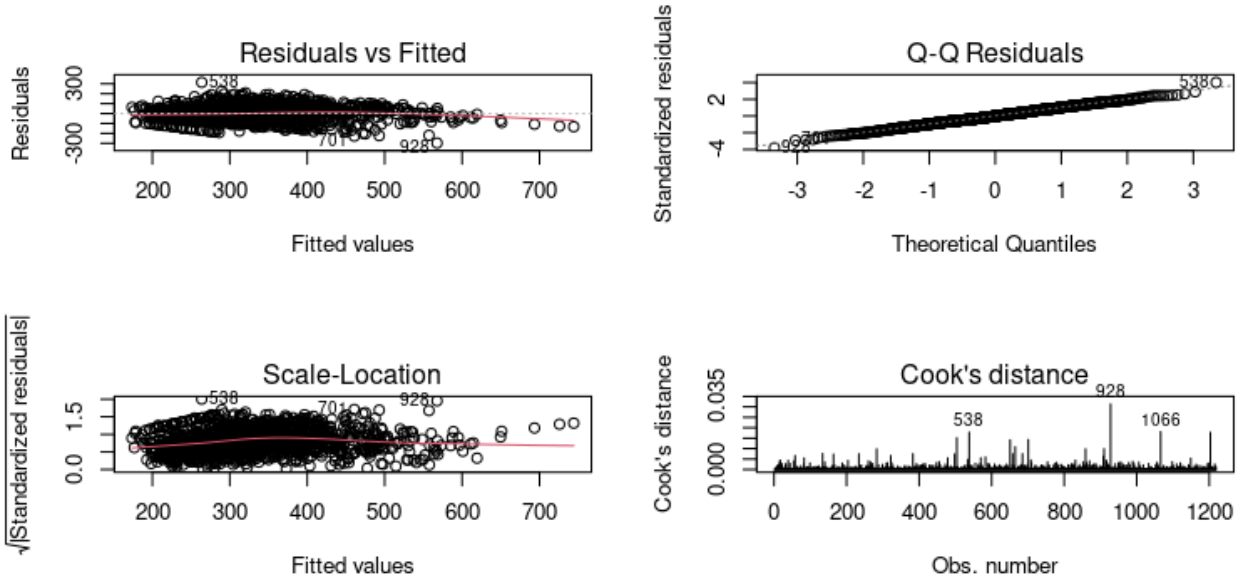


Figure 3: Diagnostic Plots of the Final Model

3.7 t-tests

By reviewing the Pr(>|t|) column in the final model's summary output, we found that most predictors had p-values below 0.05, indicating statistical significance. However, some exceptions included PTS (p = 0.287559), Pos1PF (p = 0.084295), Pos1SF (p = 0.290047), and Pos1SG (p = 0.126463), which had p-values exceeding this threshold. These non-significant results could stem from high multicollinearity among predictors, as previously identified, or the fact that t-tests evaluate the marginal contribution of each predictor without accounting for their joint impact on the model's overall fit.

3.8 VIF

The VIF scores for most predictors are below the standard cutoff of 5, except for PTS (14.41) and the interaction term Age_sq:PTS (15.17). This elevated multicollinearity arises from the structural relationship introduced by the interaction term. Despite this, partial F-tests confirm that Age_sq:PTS is statistically significant, making the high collinearity an acceptable trade-off. However, this collinearity complicates the interpretation of individual predictors, particularly PTS, as their effects cannot be easily disentangled within the model.

3.9: Outliers, Influence points, and High leverage points

We identified observation 538 as an outlier using the standard residuals cutoff method. This was confirmed by the scale-location plot, where it had the largest square-rooted standardized residual. The observation corresponds to Gordon Hayward, often referred to as the

“Most Overpaid Player in NBA History” (Urbina, 2024), which may explain its extreme response value.

The hat values plot revealed multiple high-leverage points, representing observations with extreme predictor values. These points likely correspond to NBA players who significantly deviate from the average.

Using the Cook’s distance plot, we identified observations 538, 928, and 1066 as influential points. These points have a substantial impact on the estimated coefficients and predictions, warranting closer examination.

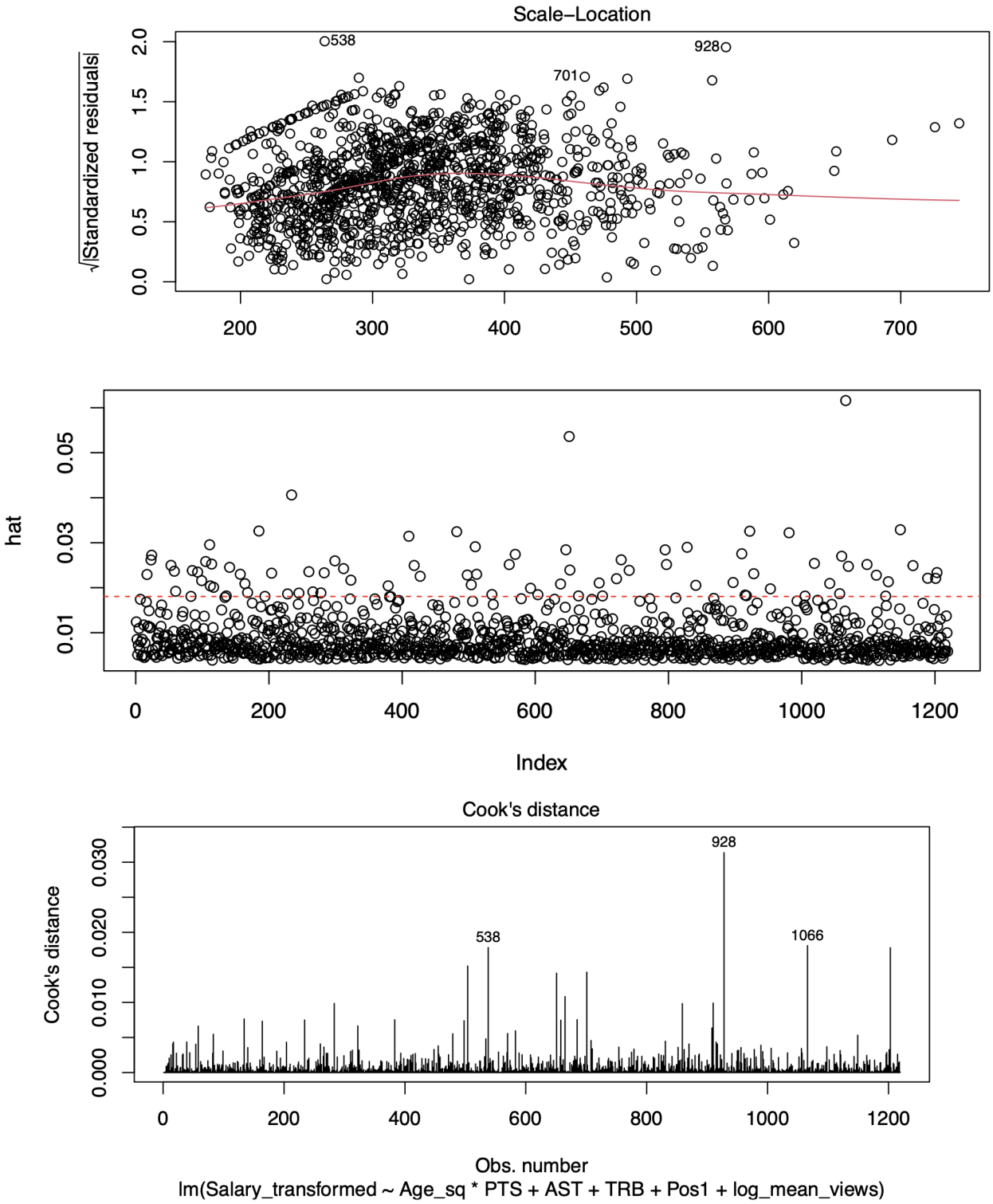


Figure 4: Plot of Outliers, High-leverage points, and Influential points

3.10 Confidence Intervals

Using our finalized model, we constructed 95% confidence intervals for the coefficients of the predictors. Some predictors, such as Pos1PG and Pos1SF, have very wide confidence intervals. Additionally, the intervals for PTS, Pos1PF, Pos1SF, and Pos1SG include 0, indicating that 0 is a plausible value for the population coefficient. This weakens our ability to interpret these variables' contributions to the model, as their effects may not be statistically significant.

Parameter	2.50%	97.50%
(Intercept)	92.2625608	158.294583
Age_sq	0.06861017	0.13687929
PTS	-1.2703332	4.28064918
AST	5.32682973	14.0629263
TRB	4.67371144	10.7266814
Pos1PF	-27.118106	1.7214337
Pos1PG	-62.004101	-19.885675
Pos1SF	-25.911058	7.75035351
Pos1SG	-31.190932	3.8656144
log_mean_views	2.71126309	8.97309121

Age_sq:PTS	0.0046914	0.01179549

Figure 5: Confidence Intervals of Model Coefficients

4. Conclusion and Limitations

4.1 Conclusion

With the 95% confidence intervals from the final model, we conclude that **log_mean_views**, **AST**, and **TRB** have the strongest positive influence on Salary_transformed among the predictors. With 95% confidence, on average, holding all other variables constant:

- A one-unit increase in **log_mean_views** is expected to increase (Salary^0.3 - 1) / 0.3 by 2.71 to 8.97 units.
- A one-unit increase in **AST** is expected to increase it by 5.33 to 14.06 units.
- A one-unit increase in **TRB** is expected to increase it by 4.67 to 10.73 units.

Position also influences salary. Point guards (Pos1PG) generally have the most negative coefficient, indicating lower pay than other positions. In contrast, centers (the baseline category) likely earn the most, as confidence intervals for other positions include negative values, suggesting relatively lower pay.

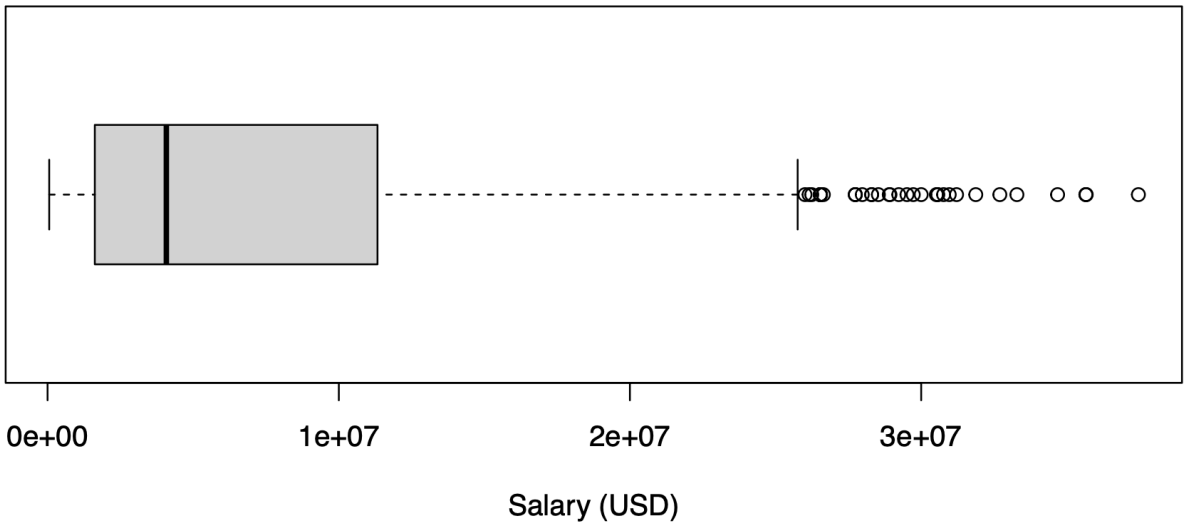
The findings align with the literature in several ways. Assists, identified as a strong influence, are consistent with Louivion & Pettersson (2018). However, our results challenge Lyons, Newton Jackson Jr., & Livingston (2018), who ranked assists and rebounds as secondary factors. Instead, our findings rank assists and rebounds as the top contributors to salary. Our confidence intervals also support Sarlis & Tjortjis (2024), which found centers are paid the most and point guards the least. An outlier is log_mean_views, representing popularity, which emerged as a strong predictor but was not reflected in the literature we reviewed.

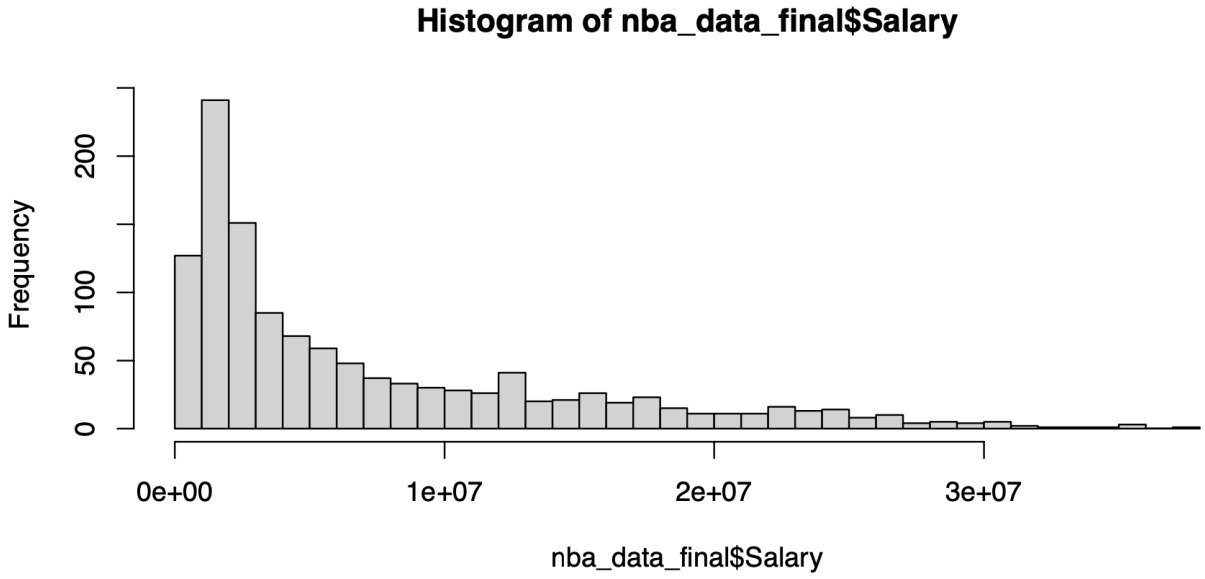
4.2 Limitations

While our analysis offers valuable insights, several limitations remain. Transformations like Box-Cox and logarithmic adjustments, though necessary to meet assumptions, complicate interpretability by altering variable scales. High multicollinearity from structural relationships impairs inference from coefficients, and high-leverage points may skew the model's representation of the dataset.

These issues contributed to wide confidence intervals and a mediocre adjusted R-squaredn reducing the model's precision for salary predictions and limiting its practical utility. Residual plot patterns also suggest unresolved violations of linearity and constant variance assumptions, raising concerns about the model's reliability.

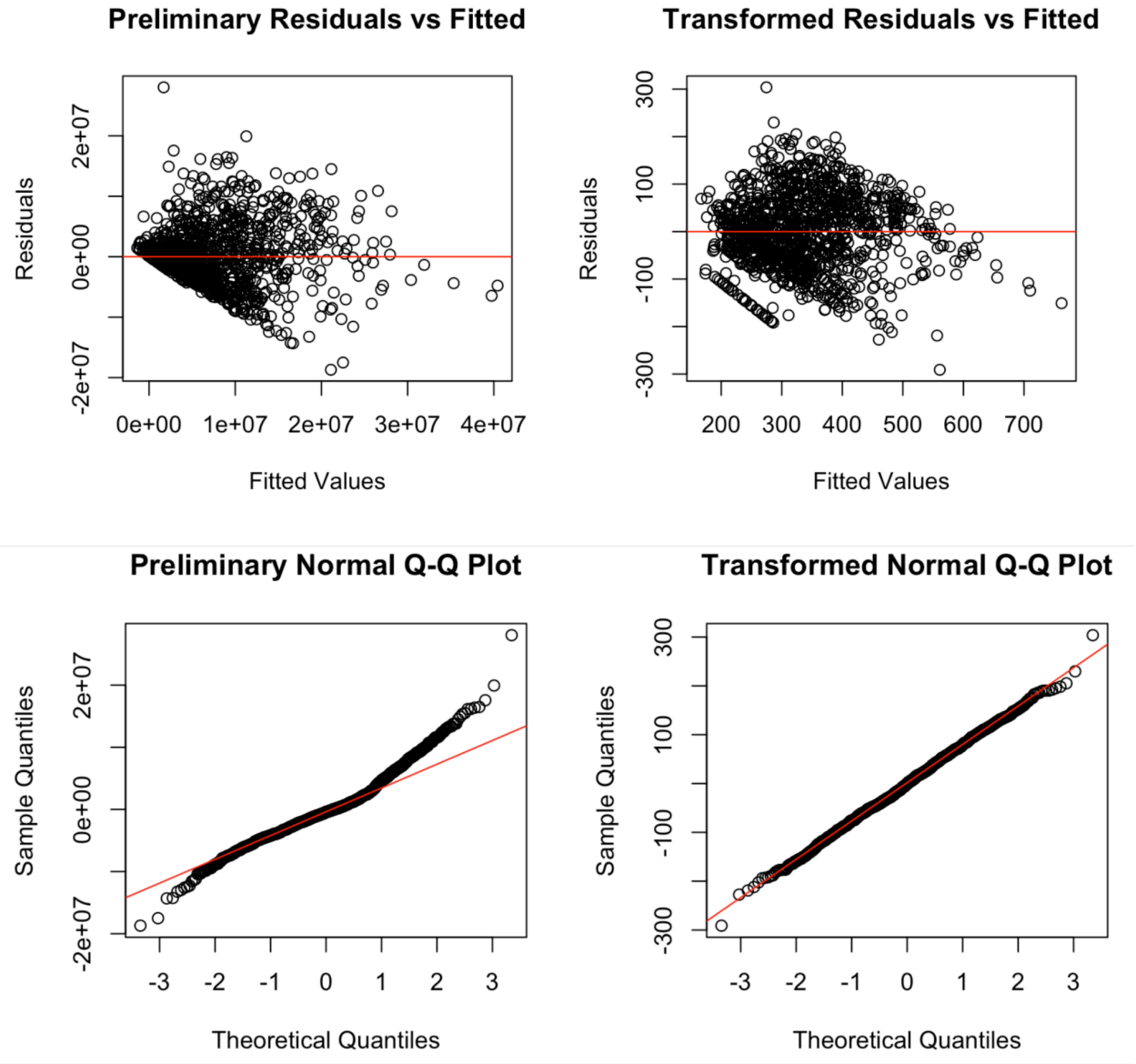
Appendix


0e+00
1e+07
2e+07
3e+07
Salary (USD)

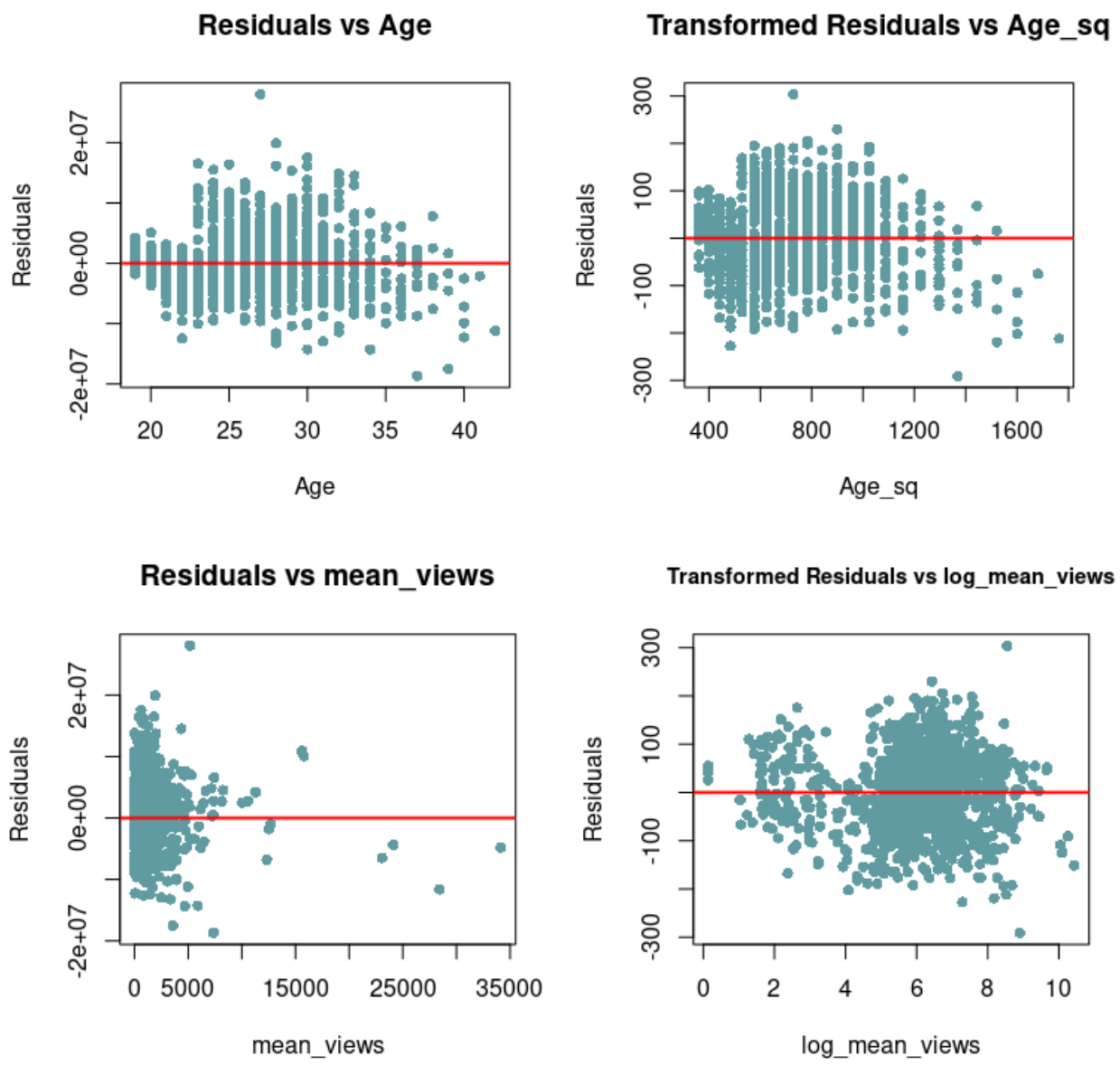

Histogram of nba_data_final$Salary
Frequency
0
50
100
200
0e+00
1e+07
2e+07
3e+07
nba_data_final$Salary

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Salary ($)	543,203	1,245,000	2,345,750	4,672,034	5,980,500	43,006,362
Age	19	22	25	26.1	29	40
PTS	0	4.2	11.5	13.9	19.7	36.1
AST	0	1.2	2.8	3.9	5.4	11.4
TRB	0.6	3.2	5.6	6.7	9.1	15.2
mean_views	2,000	7,800	12,900	14,563	19,400	67,000

Appendix 1: Summary statistics and salary graph



Appendix 2: Comparison of Preliminary and Transformed Models Diagnostic Plots



Appendix 3: Comparison of Residuals vs Predictor Plots for Preliminary and Transformed Models

Bibliography

1. Lyons Jr. R. & Newton Jackson Jr. E. & Livingston, A. (2018, July 13). Determinants of NBA player salaries. The Sport Journal. https://thesportjournal.org/article/determinants-of-nba-player-salaries/
2. Sarlis, V., & Tjortjis, C. (2024, April 21). Sports analytics: Data mining to uncover NBA player position, age, and injury impact on performance and Economics. MDPI. https://www.mdpi.com/2078-2489/15/4/242
3. Louivion, S. & Pettersson, F. (2017). Analysis of performance measures that affect NBA salaries. https://www.diva-portal.org/smash/get/diva2:1114463/FULLTEXT01.pdf
4. Urbina, F. (2024, September 30). *The most overpaid NBA players of all time*. HoopsHype. https://hoopshype.com/lists/the-most-overpaid-nba-players-of-all-time/#:~:text=The%20recently%20retired%20Gordon%20Hayward,to%20fault%20him%20for%20that