From Points to Paychecks: An Exploratory Analysis of NBA Salaries and Performance (2022–23)

Pacific Boys: Marcelino Bautista, Konstantin Khorobrykh, Tim Platz, Da-Wei Lin

2025-04-17

Contents

1 Introduction **3**
1.1 Research Approach . . . 3

2 Data Source **3**

3 Data Wrangling **3**

4 Operationalization **3**

5 Exploratory Data Analysis **4**
5.1 Salary Distributions and Star Player Comparison . . . 4
5.2 Performance Metric Distributions . . . 4
5.3 Pairwise Scatter Plots . . . 4
5.4 Correlation Matrix . . . 5
5.5 Salary vs. Key Predictors . . . 5
5.6 Preliminary Models . . . 5
5.7 Collinearity Issues . . . 6
5.8 Heteroscedasticity & Robust Standard Error . . . 6
5.9 Model Selection Rationale . . . 6

6 Confirmatory Analysis, Final Model, and Interpretation **6**

7 Model Limitations and Future Directions for Improvement **6**

8 Appendix **8**
8.1 Pairwise Scatter-Plot Matrix . . . 8
8.2 Correlation Matrix . . . 9
8.3 Collinearity Diagnostics: VIF Comparison . . . 9
8.4 Residual Diagnostics for Heteroscedasticity . . . 10
8.5 Breusch–Pagan Test Results . . . 10
8.6 Salary vs. Key Predictors . . . 11
8.7 Raw-Salary Exploratory Models . . . 12
8.8 Log-Salary Exploratory Models . . . 13
8.9 Box-Cox-Salary Exploratory Models . . . 14

8.10 Robust SE for Selected Models 15
8.11 Data Dictionary 16

9 References **16**

1 Introduction

GMs and coaches of NBA teams need to make highstakes decisions when assembling their rosters. With the average NBA player salary topping $6.62 million for the 2022-23 season ([1]), the need for data insights is higher than ever. Our goal is to describe how NBA player salaries relate to performance metrics during the 2022-23 **regular** season. We focus on Points Per Game (PTS) and Minutes Played (MP) as primary predictors, along with other stats like FG, TRB, 3P%, TOV, WS, AST, and GP. This exploratory analysis will derive actionable insights and enable key stakeholders to make data-driven decisions.

1.1 Research Approach

While prior studies use machine learning for salary prediction ([2, 3]), our Ordinary Least Squares (OLS) regression model prioritizes interpretability to identify which performance indicators are most descriptively associated (not causal) with compensation. Because this is a team sport, we acknowledge the presence of dependence among observations and violations of the **IID** (independent and identically distributed) assumption. However, we proceed with the analysis given our focus on descriptive insights. We will revisit these hypotheses in our conclusion to confirm or refine their validity.

2 Data Source

- **Primary Source:** Kaggle NBA Salaries 2022-23
- **All-Star Players:** Basketball Reference All-Star 2023
- **Years of Experience:** Basketball Reference Team Rosters
- **Market Size Data:** NBA Team Market Size Rankings
- **Unit:** Individual NBA players
- **Type:** Cross-sectional snapshot of 2022-23 season

3 Data Wrangling

For data wrangling, we began by obtaining the main dataset from Kaggle and iteratively scraping roster data for each team from Basketball Reference. We then added All-Star selections and merged the dataset with market size information. Duplicate entries, primarily from players traded mid-season, were identified and removed. The cleaned data was split into two parts: 30/70 for exploratory data analysis (EDA) and the remainder for confirmatory analysis. The current analysis focuses on the EDA portion.

The raw NBA dataset contains 140 players and 62 variables for the 2022-23 season. After cleaning, the dataset contains 140 active players with valid salary data. We selected 24 key variables including performance metrics, player characteristics, and team information. We understand that Salary is heavily right-skewed, so we created a log-transformed variable (LogSalary) to normalize the distribution. We also created a Box-Cox transformed variable (BoxCoxSalary) with lambda of 0.1455141 to further improve normality.

4 Operationalization

- **Salary:** Player compensation in USD

- **LogSalary:** Natural logarithm of salary (created to address right-skewness)
- **PTS:** Points per game = Total points / games played
- **MP:** Minutes played per game
- **Other metrics:** FG (field goals), TRB (rebounds per game), `3P%` (3-point shooting percentage), TOV (turnovers), WS (win shares), AST (assists per game), GP (games played)
- **TVMS:** TV Market Size - represents the size of the team's market
- **TS%:** True Shooting Percentage - a measure of shooting efficiency

5 Exploratory Data Analysis

5.1 Salary Distributions and Star Player Comparison

The salary distribution is heavily right-skewed (Figure 1). Applying a log transformation makes the distribution more symmetric and closer to a normal shape (Figure 1). Additionally, rookies are clearly at a disadvantage in terms of salary compared to more experienced players (Figure 1).

Figure 1: Salary distributions and boxplots by player category

5.2 Performance Metric Distributions

PTS and FG are both right-skewed and highly correlated, while MP shows a fairly uniform distribution (Figure 2). True Shooting Percentage (TS%), calculated as TS% = PTS / (2 * (FGA + 0.44 * FTA))—which integrates scoring from field goals and free throws—tends to cluster around 0.6 regardless of salary (Figure 2). Notably, there's an outlier with a TS% of 0, and interestingly, an individual with a TS% of 0.8 isn't earning much.

5.3 Pairwise Scatter Plots

Strong correlations among metrics—such as PTS, MP, WS, and FG—suggest these variables tend to move together (see Appendix Figure 8.1). Additionally, histograms reveal significant right skewness,

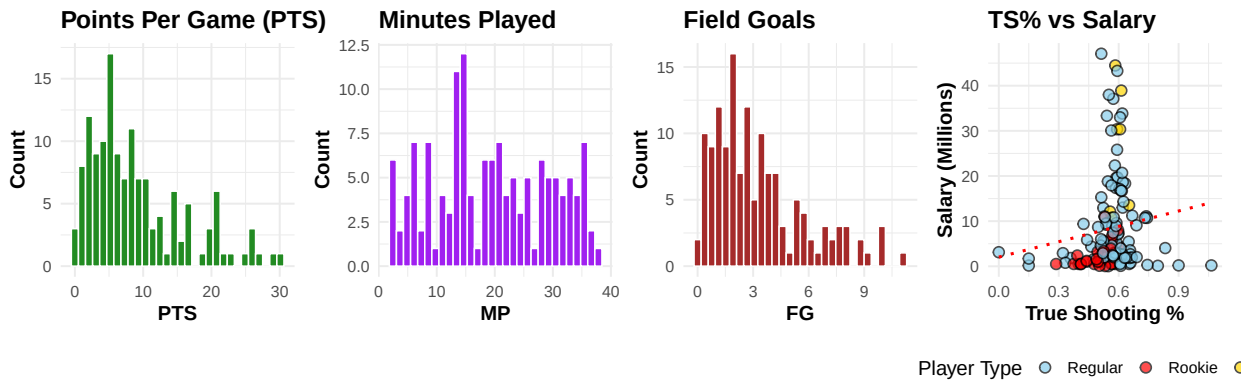


Figure 2: Performance metric distributions

with TVMS displaying an almost bimodal distribution. Therefore, it's crucial to test for variance inflation factors (VIF) before constructing the models to ensure multicollinearity isn't an issue.

5.4 Correlation Matrix

There is a noticeable positive correlation between points (PTS) and other performance metrics such as minutes played (MP), field goals (FG), turnovers (TOV), and win shares (WS) (see Appendix Figure 8.2). This suggests that market-driven factors captured by TVMS are distinct from on-court performance.

5.5 Salary vs. Key Predictors

The analysis indicates that performance metrics such as points per game, minutes played, and win shares are strongly correlated with salary, although there are exceptions—such as a rookie with high minutes but a relatively low salary (Appendix Figure 8.6). Additionally, salaries tend to rise with years of experience before leveling off, highlighting the impact of tenure and potentially age. There are also clear positional differences; for example, centers and shooting guards tend to have lower median salaries compared to power forwards, who generally command higher median earnings. While there is an observation that players on larger market teams may earn more on average, this relationship appears to be complex and could benefit from further evaluation.

5.6 Preliminary Models

To explore linear relationships between NBA player salaries and various performance metrics, we formulated nine models across three salary transformations. These models assess how different variables influence player compensation during the 2022-23 season.

5.6.1 Model Specifications

We test three model specifications with three different dependent variables:

- **Models 1-3**: Raw Salary
- **Models 4-6**: Log-transformed Salary (LogSalary)
- **Models 7-9**: Box-Cox transformed Salary (BoxCoxSalary)

For each transformation, we test three specifications:

1. **Basic Model**: Only Points Per Game (PTS)
2. **Performance Model**: PTS, Total Rebounds (TRB), and Assists (AST)
3. **Comprehensive Model**: Performance metrics plus Experience (Exp_num) and TV Market Size (TVMS)

5.7 Collinearity Issues

No collinearity concern with the VIF results (<5), see Appendix Table 8.3.

5.8 Heteroscedasticity & Robust Standard Error

From the fitted value v.s. resudul plot, whereas the first three models exhibit heteroscedasticity, the next six (using log and Box-Cox transformations) are homoscedastic. These observations are further supported by the Breusch-Pagan test results. (See Appendix Figure 8.4 and Appendix Table 8.5.)

5.9 Model Selection Rationale

Based on the exploratory models (Appendix Tables 8.7, 8.8, and 8.9), Model 3 (Raw Salary ~ PTS + TRB + AST + Exp_num + TVMS) with additional selected predictive variables provides improvements in R^2 compared to Model 1 and Model 2. While log and Box-Cox transformations (Models 4 - 9) successfully decreased heteroscedasticity, they also decreased R^2 and make interpreting coefficients less direct (Appendix Figure 8.4). Model 3 includes key performance metrics, experience, and market size, all showing significance (especially with robust standard errors). Since we can use robust standard errors to address heteroscedasticity, we can still use the raw salary model (Model 3) for interpretability and significance testing. (Appendix Table 8.10). Therefore, we proceed with Model 3 as final model and apply confirmatory dataset to validate our findings.

6 Confirmatory Analysis, Final Model, and Interpretation

This regression explains about 69% of the variation in NBA salaries, with scoring (PTS), experience (Expnum), and TV market size (TVMS) emerging as the most influential factors. Specifically, each additional point per game is associated with roughly a $942 K boost in salary, each year in the league adds about $1.01 M, and playing in a larger TV market contributes roughly $642 K, all else held equal. Rebounds and assists do not carry significant weight once scoring is in the model, and visual diagnostics show that predictions align well with actual salaries for most players, though superstar contracts at the high end exhibit greater dispersion.

7 Model Limitations and Future Directions for Improvement

As potential next steps, our model could be refined by incorporating longitudinal data to capture temporal trends in player performance. Additionally, addressing omitted variable bias by including contextual factors—such as changes in league policy, evolving salary structures, the influence of agents, and injury history—may further enhance explanatory power. Stratified modeling, such as developing team-specific or region-specific models, could also improve fit and interpretability.

Lastly, exploring nonlinear relationships through the inclusion of quadratic or interaction terms may better capture the complexity inherent in the data.

	Dependent variable:
	Salary
	Final Model
PTS	$942,429.500^{***}$
	(122,593.400)
TRB	225,295.100
	(255,167.000)
AST	19,140.650
	(404,417.300)
Exp_{num}	$1,014,302.000^{***}$
	(118,423.100)
TVMS	$641,778.900^{***}$
	(222,584.200)
Constant	$-7,211,120.000^{***}$
	(848,045.000)
Robust p-val: PTS	0
Robust p-val: TRB	0.378
Robust p-val: AST	0.962
Robust p-val: Exp	0
Robust p-val: TVMS	0.004
Observations	326
R^2	0.689
Adjusted R^2	0.684
Residual Std. Error	6,083,581.000 (df = 320)
F Statistic	141.784^{***} (df = 5; 320)
Note:	*p<0.1; **p<0.05; ***p<0.01

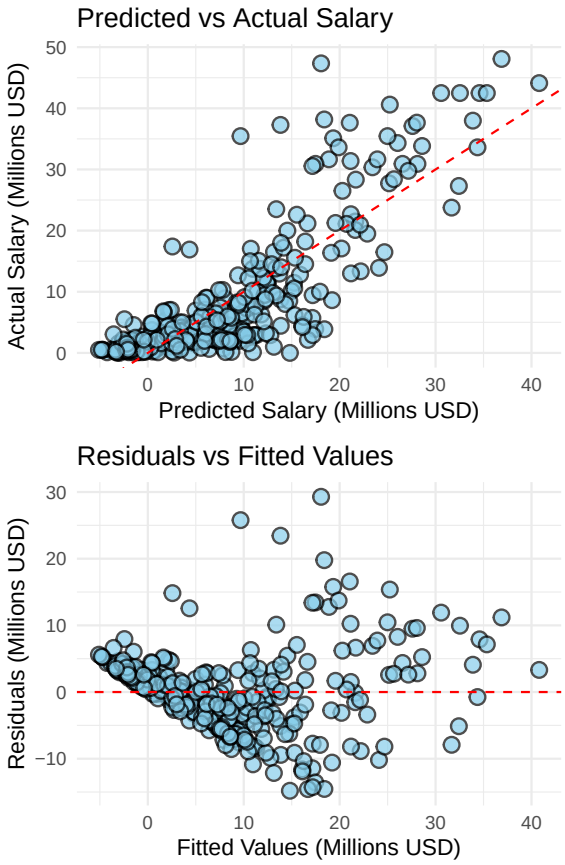
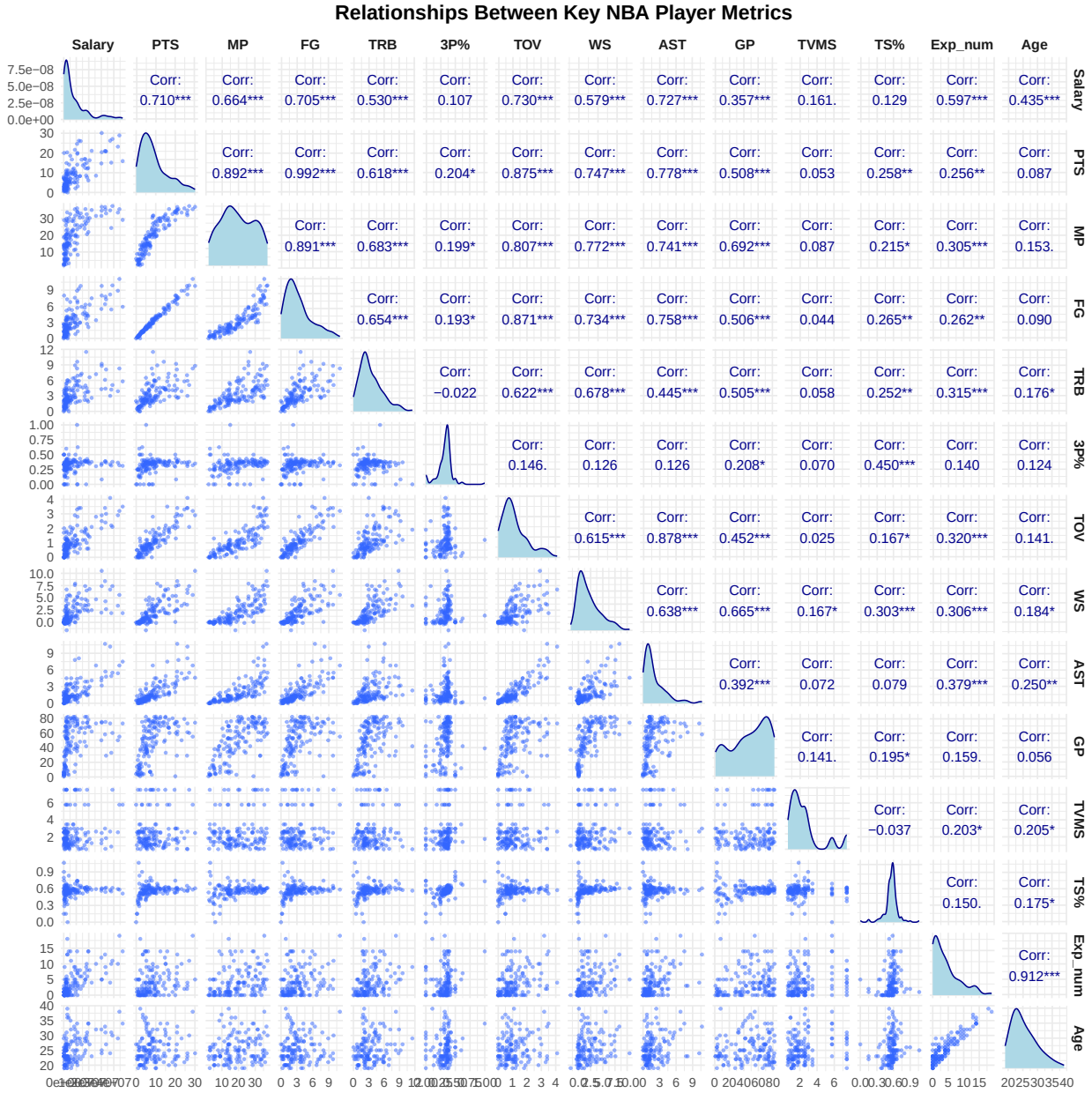


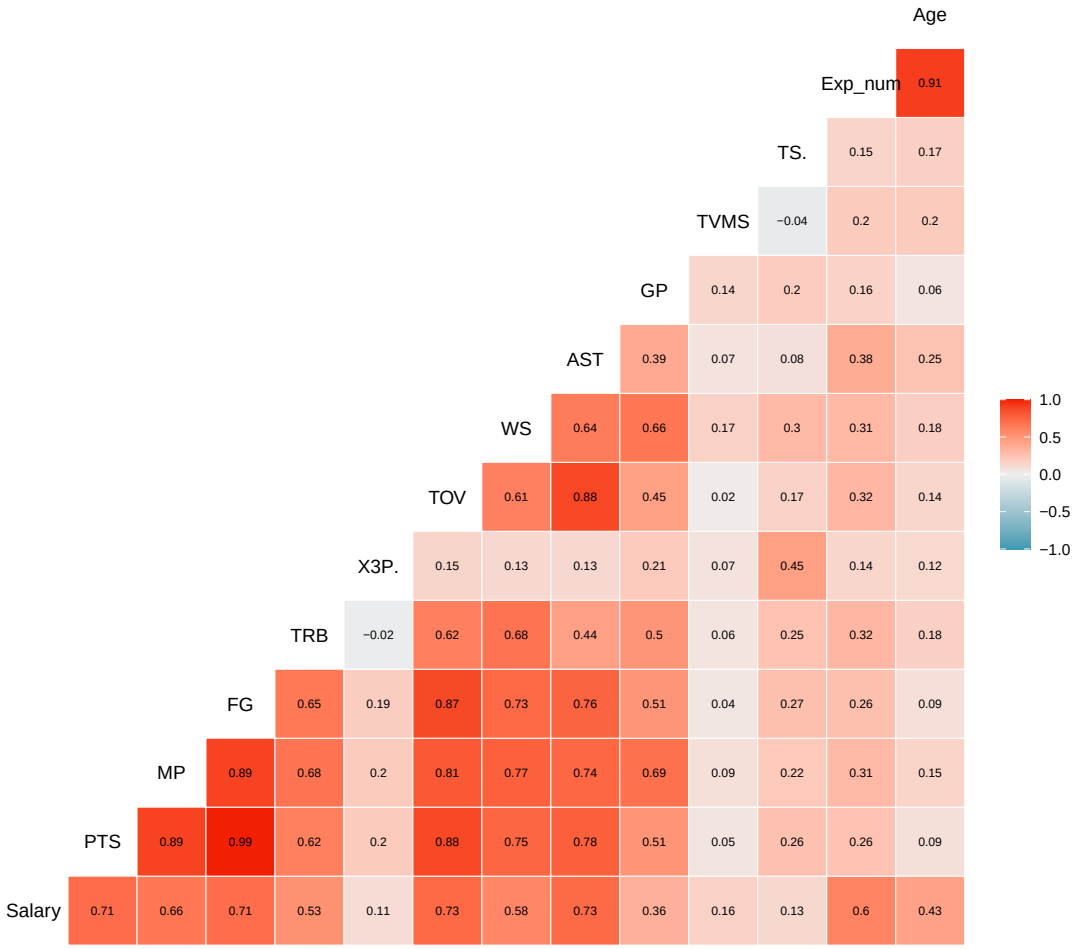
Figure 3: Final Model Results: (a) Coefficients with robust SE; (b) Predicted vs Actual and residual plots.

8 Appendix

8.1 Pairwise Scatter-Plot Matrix



8.2 Correlation Matrix



8.3 Collinearity Diagnostics: VIF Comparison

Table 1: VIF Comparison Across Models (Excluding Models 1, 4, and 7)

Variable	Model 2	Model 3	Model 5	Model 6	Model 8	Model 9
PTS	3.30	3.41	3.30	3.41	3.30	3.41
TRB	1.63	1.73	1.63	1.73	1.63	1.73
AST	2.54	2.83	2.54	2.83	2.54	2.83
Exp_num	NA	1.29	NA	1.29	NA	1.29
TVMS	NA	1.04	NA	1.04	NA	1.04

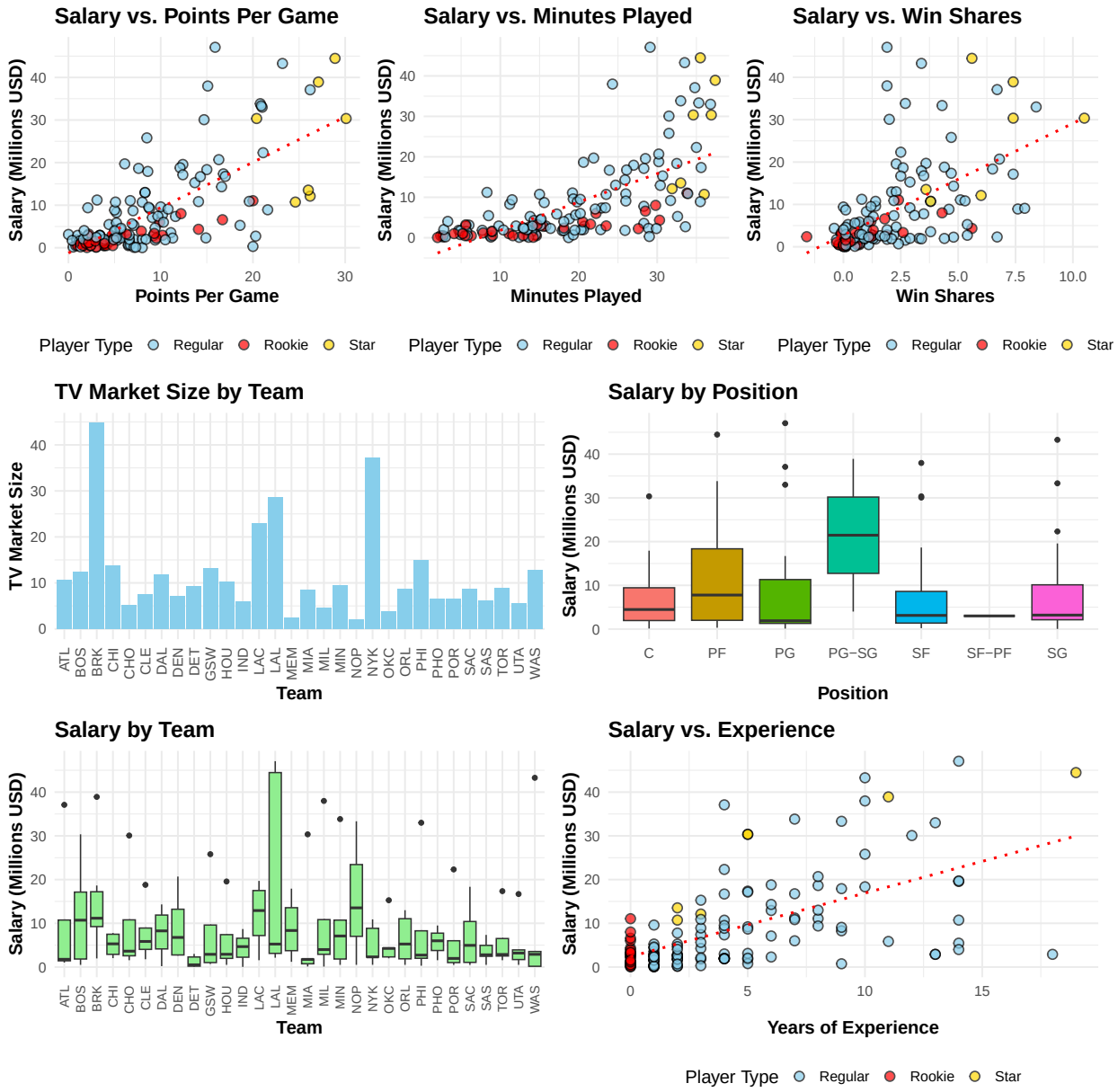
8.4 Residual Diagnostics for Heteroscedasticity

8.5 Breusch–Pagan Test Results

Table 2: Breusch-Pagan Test Results for Heteroscedasticity

Model	BP Statistic	P-Value	Conclusion
Model 1	31.826	0.0000	Heteroscedastic
Model 2	53.128	0.0000	Heteroscedastic
Model 3	47.785	0.0000	Heteroscedastic
Model 4	0.274	0.6006	Homoscedastic
Model 5	7.171	0.0667	Homoscedastic
Model 6	6.457	0.2642	Homoscedastic
Model 7	0.284	0.5938	Homoscedastic
Model 8	7.392	0.0604	Homoscedastic
Model 9	6.957	0.2239	Homoscedastic

8.6 Salary vs. Key Predictors



8.7 Raw-Salary Exploratory Models

Table 3: NBA Salary preliminary Models (Raw Salary)

	Dependent variable:		
	Salary		
	Model 1	Model 2	Model 3
PTS	$1,065,213.000^{***}$	$371,271.300^{**}$	$536,860.500^{***}$
	(89,908.840)	(147,836.500)	(127,087.400)
TRB		$816,232.800^{**}$	313,391.500
		(325,873.700)	(283,988.600)
AST		$2,401,759.000^{***}$	$1,444,860.000^{***}$
		(454,548.400)	(405,766.900)
Exp_{num}			$902,796.800^{***}$
			(126,761.900)
TVMS			242,708.300
			(261,655.200)
Constant	$-1,226,388.000$	$-2,545,865.000^{**}$	$-4,754,681.000^{***}$
	(1,024,934.000)	(1,064,685.000)	(1,085,961.000)
Observations	140	140	140
R^2	0.504	0.600	0.718
Adjusted R^2	0.501	0.591	0.707
Residual Std. Error	7,409,796.000 (df = 138)	6,706,668.000 (df = 136)	5,673,890.000 (df = 134)
F Statistic	140.368^{***} (df = 1; 138)	67.932^{***} (df = 3; 136)	68.151^{***} (df = 5; 134)
Note:			*p<0.1; **p<0.05; ***p<0.01

8.8 Log-Salary Exploratory Models

Table 4: NBA Salary preliminary Models (Log Salary)

	Dependent variable:		
	LogSalary		
	Model 4	Model 5	Model 6
PTS	0.133***	0.047*	0.069***
	(0.014)	(0.024)	(0.022)
TRB		0.219***	0.150***
		(0.053)	(0.049)
AST		0.192**	0.061
		(0.074)	(0.070)
Exp_{num}			0.124***
			(0.022)
TVMS			0.039
			(0.045)
Constant	13.911***	13.559***	13.244***
	(0.162)	(0.173)	(0.187)
Observations	140	140	140
R^2	0.387	0.474	0.585
Adjusted R^2	0.383	0.462	0.569
Residual Std. Error	1.170 (df = 138)	1.092 (df = 136)	0.977 (df = 134)
F Statistic	87.166*** (df = 1; 138)	40.775*** (df = 3; 136)	37.716*** (df = 5; 134)
Note:			*p<0.1; **p<0.05; ***p<0.01

8.9 Box-Cox-Salary Exploratory Models

Table 5: NBA Salary preliminary Models (Box-Cox Salary)

	Dependent variable:		
	BoxCoxSalary		
	Model 7	Model 8	Model 9
PTS	1.245***	0.461**	0.669***
	(0.118)	(0.198)	(0.174)
TRB		1.828***	1.197***
		(0.436)	(0.389)
AST		1.913***	0.712
		(0.608)	(0.556)
Exp_{num}			1.131***
			(0.174)
TVMS			0.323
			(0.358)
Constant	45.203***	42.265***	39.453***
	(1.343)	(1.423)	(1.487)
Observations	140	140	140
R^2	0.448	0.536	0.656
Adjusted R^2	0.444	0.525	0.644
Residual Std. Error	9.706 (df = 138)	8.965 (df = 136)	7.768 (df = 134)
F Statistic	111.830*** (df = 1; 138)	52.273*** (df = 3; 136)	51.199*** (df = 5; 134)
Note:			*p<0.1; **p<0.05; ***p<0.01

8.10 Robust SE for Selected Models

Table 6: NBA Salary Models with Robust Standard Errors

	Dependent variable:		
	Salary Model 3	LogSalary Model 6	BoxCoxSalary Model 9
PTS	$536,860.500^{***}$	0.069^{***}	0.669^{***}
	(165,364.900)	(0.023)	(0.200)
TRB	313,391.500	0.150^{***}	1.197^{***}
	(320,133.600)	(0.039)	(0.334)
AST	$1,444,860.000^{**}$	0.061	0.712
	(597,919.800)	(0.063)	(0.558)
Exp_{num}	$902,796.800^{***}$	0.124^{***}	1.131^{***}
	(183,344.700)	(0.021)	(0.188)
TVMS	242,708.300	0.039	0.323
	(297,389.300)	(0.037)	(0.324)
Constant	$-4,754,681.000^{***}$	13.244^{***}	39.453^{***}
	(997,226.400)	(0.219)	(1.626)
Robust p-val: PTS	0.001	0.003	0.001
Robust p-val: TRB	0.329	0	0
Robust p-val: AST	0.017	0.339	0.204
Robust p-val: Exp_{num}	0	0	0
Robust p-val: TVMS	0.416	0.29	0.321
Observations	140	140	140
R^2	0.718	0.585	0.656
Adjusted R^2	0.707	0.569	0.644
Residual Std. Error (df = 134)	5,673,890.000	0.977	7.768
F Statistic (df = 5; 134)	68.151^{***}	37.716^{***}	51.199^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

8.11 Data Dictionary

Variable	Description
Salary	Player salary in USD
LogSalary	Natural logarithm of salary
PTS	Points per game
MP	Minutes played per game
FG	Field goals made per game
TRB	Total rebounds per game
3P%	3-point shooting percentage
TOV	Turnovers per game
WS	Win shares
GP	Games played
AST	Assists per game
TVMS	TV Market Size
TS%	True Shooting Percentage
star_player	All-Star status (1=Yes, 0=No)
Position	Player position
Age	Player age
Exp_num	Years of NBA experience

9 References

1. Southwest Journal. (n.d.). *NBA Players' Salaries: How Much Do NBA Players Make?* Retrieved from https://www.southwestjournal.com/sport/nba/nba-players-salaries/?utm_source=chatgpt.com

2. Papadaki, I., & Tsagris, M. (2020). *Are NBA Players' Salaries in Accordance with Their Performance on Court?* In *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies* (pp. 405–428). https://doi.org/10.1007/978-3-030-85254-2_25

3. Sigler, K.J., & Sackley, W.H. (2000). *NBA Players: Are They Paid for Performance? Managerial Finance, 26*(7), 46–51. https://doi.org/10.1108/03074350010766783

4. Bautista, M., Khorobrykh, K., Platz, T., & Lin, DW. (2024). NBA Salary and Performance: Exploratory Data Analysis [R Markdown file]. GitHub repository. Retrieved April 17, 2025, from https://github.com/mids-w203/lab-2-pacific-boys-1/blob/main/notebooks/analysis_lab2_short/lab2_report_short.Rmd