Rene Mejias and Juan Garcia

**Introduction**

Basketball has evolved significantly over the years, with players and teams adopting data-drive strategies to improve their performance on the court. This project seeks to analyze NBA player statistics from the 2022-2023 season to find insights about player efficiency, position-based performance trends, and the relationship between playing time and productivity on the court. With the increasingly important role of analytics in sports teams, understanding such trends can be important for coaches, scouts and fans. The data used in this analysis is a data set on Kaggle, that web scraped the information from a website that tracks statistics of basketball everywhere.

**2. Data Collection**

This project uses three main data sources: Basketball Reference for player statistics and team standings, and HoopsHype for player salary information.

**2.1 Team Standings**  
 Team-level data was gathered from Basketball Reference’s NBA 2023 Standings using the pandas.read\_html() function. Two tables, one for each conference, were extracted and merged into a single DataFrame. We removed division headers and cleaned up team names by stripping asterisks used to indicate playoff qualification. The resulting dataset includes each team's name, total wins, and total losses for the 2022–2023 season.

**2.2 Player Stats**  
 Individual player performance data was collected by downloading a CSV file of the 2022–2023 season from Basketball Reference. This dataset includes a wide range of performance metrics including points, rebounds, assists, blocks, steals, minutes played, field goal attempts and percentages, and free throw efficiency. We removed the Rk (rank) column and ensured column consistency by stripping extra whitespace. We also applied a custom text-cleaning function to fix any encoding issues in player names caused by non-UTF-8 characters.

**2.3 Player Salaries**  
 Player salary data was scraped from HoopsHype’s salary rankings using Selenium. We then used find\_elements to extract both player names and their reported salaries. Salary strings were cleaned by removing dollar signs and commas and converted to integers. Missing or malformed salary values were handled later on.

**2.4 Merging and Cleaning**  
 All datasets were merged using player names as the key. Prior to merging, we applied a custom clean\_text() function to handle character encoding mismatches between datasets. Player names were also stripped of leading/trailing spaces. The salary and stats tables were merged using a left join to retain all player stats, even if salary data was missing.

To handle players without available salary information, we calculated the average salary across all players and used that as the fill value for missing entries. This avoided skewing any downstream salary analysis with zero-filled values. We also dropped duplicate entries for players who had been traded mid-season (e.g., those listed with team "TOT") by keeping only their first occurrence.

The final dataset was saved to CSV and includes cleaned and merged player stats, team wins/losses, salary data, and newly engineered features such as a PerformanceIndex and Efficiency score.

**2.5 Data Sources**

Source 1: <https://www.basketball-reference.com/leagues/NBA_2023_totals.html>

Source 2: <https://hoopshype.com/salaries/players/2022-2023/>

Source 3: <https://www.basketball-reference.com/leagues/NBA_2023_standings.html>

**Final Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| Player | string | Name of the player |
| Pos | string | Player's position |
| Age | int | Age |
| Tm | string | Team abbreviation |
| G | int | Games played |
| GS | int | Games started |
| MP | float | Minutes played |
| FG | float | Field goals made |
| FGA | float | Field goal attempts |
| FG% | float | FG% |
| 3P | float | 3-point field goals made |
| 3PA | float | 3-point attempts |
| 3P% | float | 3-point % |
| 2P | float | 2-point field goals |
| 2PA | float | 2-point attempts |
| 2P% | float | 2-point % |
| eFG% | float | Effective field goal % |
| FT | float | Free throws made |
| FTA | float | Free throw attempts |
| FT% | float | Free throw % |
| ORB | float | Offensive rebounds |
| DRB | float | Defensive rebounds |
| TRB | float | Total rebounds |
| AST | float | Assists |
| STL | float | Steals |
| BLK | float | Blocks |
| TOV | float | Turnovers |
| PF | float | Personal fouls |
| PTS | float | Points scored |
| Salary | float | Player salary USD |
| Team | string | Team abbreviation |
| Wins | float | Team record from standings |
| Losses | float | Team record from standings |
| Performance Index | float | PTS + TRB + AST + STL + BLK |
| Efficiency | float | PerformanceIndex / MP |
| High Salary | int | 1 if above median salary, else 0 |

**3. Data Analysis**

**3.1 Research Question 1: Do different player positions contribute differently to performance?**

Hypothesis: Hybrid players who can play multiple positions (e.g., SF-SG) contribute more statistically across the board than players in a single, traditional role.

To investigate this, we created a custom PerformanceIndex for each player by summing five key metrics: points (PTS), total rebounds (TRB), assists (AST), steals (STL), and blocks (BLK). This gave us a broader measure of total contribution across offense and defense. We then grouped players by their listed position and computed the average PerformanceIndex for each.

A bar chart revealed that players in hybrid roles like SF-SG and PF-SF had the highest average performance index, suggesting that position flexibility may lead to higher on-court productivity. In contrast, standard roles like SG and SF scored lower. These results support the hypothesis that hybrid players, due to their diverse responsibilities, tend to contribute more in aggregate.

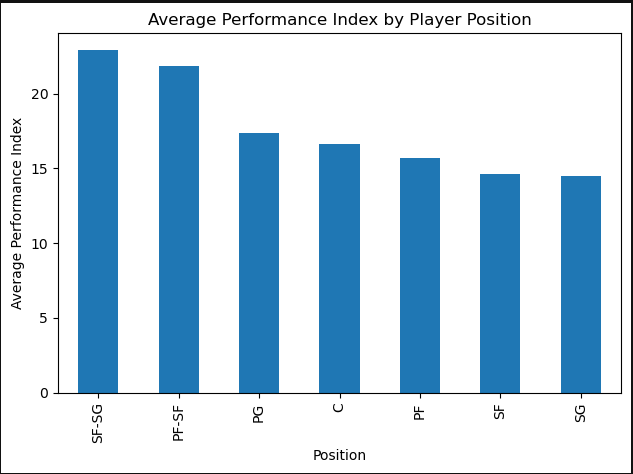


Figure 1. Average Performance Index by Player Position

**3.2 Research Question 2: Is there a relationship between minutes played and player efficiency?**

Hypothesis: Players who play more minutes are more efficient on a per-minute basis.

To test this, we defined a new Efficiency variable as PerformanceIndex divided by Minutes Played (MP). This normalization allowed us to fairly compare starters and bench players.

We created a scatter plot of Minutes Played vs. Efficiency and found a visible upward trend. To quantify the relationship, we calculated a Pearson correlation coefficient of 0.42 with a p-value < 0.0001, indicating a statistically significant moderate positive correlation. In simpler terms, players who earn more playing time tend to be more efficient. This may reflect the fact that coaches reward more efficient, impactful players with more minutes.

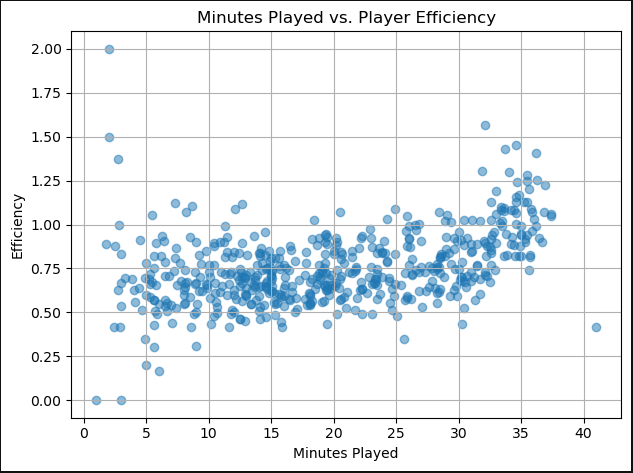


Figure 2. Minutes Played vs. Player Efficiency

**3.3 Research Question 3: Do player salaries align with their on-court performance?**

Hypothesis: Players with higher performance index scores should earn higher salaries.

To explore this, we created visual based the relationship between PerformanceIndex and Salary using a scatter plot. There was a general upward trend, though it was not tightly clustered. A linear regression model was then fit to quantify the relationship, and the model yielded an R² score of 0.42, meaning 42% of the variation in salary can be explained by performance.

While the result supports the hypothesis to some extent, the remaining unexplained variance indicates that salary is influenced by other non-performance factors, such as experience, marketability, contract timing, or previous accolades.

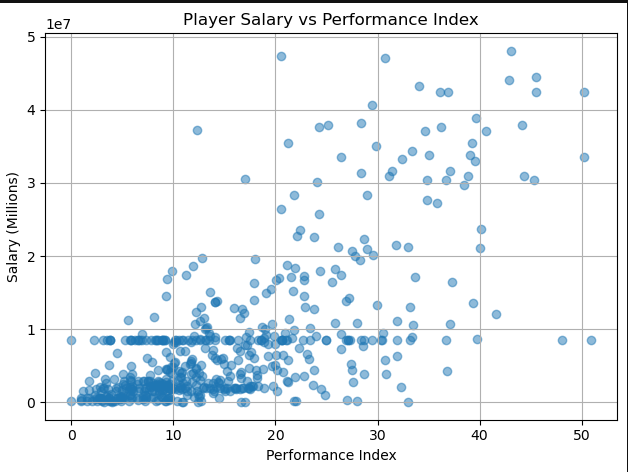


Figure 3. Player Salary vs Performance Index

**3.4 Machine Learning: Predicting Salary and Salary Tiers (Linear Regression)**

We first trained a linear regression model using the five core stats (PTS, TRB, AST, STL, BLK) to predict salary. The model achieved an R² score of 0.42, reinforcing our earlier findings that performance partially, but not fully, predicts compensation.



Figure 4. (Linear Regression) Salary Prediction based on five core stats

In this part of the project, we used a logistic regression model to predict whether an NBA player earns a high or low salary based on their performance stats. We created a new column called HighSalary, which is set to 1 if a player's salary is above the median salary and 0 if it is below. The model used key stats like points (PTS), rebounds (TRB), assists (AST), steals (STL), and blocks (BLK) to make the prediction.

**3.5 Logistic Regression**

We split the data into training and testing sets, using 80% of the data to train the model and 20% to test it. The model was trained for up to 1000 iterations to make sure it fully learned from the data.

The logistic regression model had an accuracy of **70%**, meaning it correctly predicted whether a player had a high or low salary 70% of the time. While the model shows some ability to classify salary levels based on basic stats, the 30% error rate suggests that including more information like playing time, team success, or player experience might improve its accuracy.



Figure 5. (Logistic Regression)

**3.6 Heat Map (Key Metrics)**

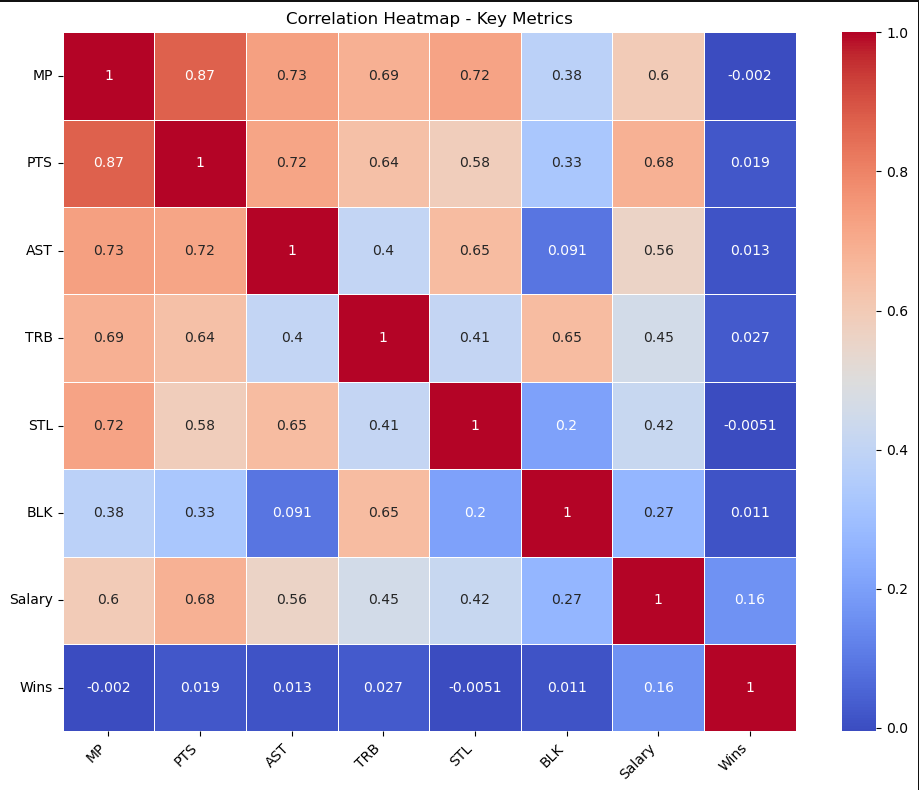


Figure 6. Heat Map based on Key metrics from the game

The heatmap shows how different player stats, salary, and team wins are related. The strongest connection is between Minutes Played (MP) and Points (PTS), with a correlation of 0.87. This means that players who play more minutes tend to score more points. Similarly, there is a strong link between Minutes Played and Assists (AST) (0.73), suggesting that players on the court longer also have more chances to assist their teammates.

Another strong connection is between Points and Assists (0.72), indicating that players who score a lot often also create scoring opportunities for others. There is also a moderate link between Points and Salary (0.68), showing that higher-scoring players generally earn more money.

However, some metrics, like Blocks (BLK), don't have strong connections with other stats, with the highest being 0.38 with Minutes Played. This suggests that blocking shots may be more of a specialized skill that doesn’t heavily depend on playing time.

Interestingly, there is almost no connection between Wins and the other metrics, with correlation values ranging from -0.0051 to 0.16. This means that individual player stats don’t seem to have a direct impact on team wins, which could indicate that other factors like team chemistry or overall strategy are more important.

Overall, the heatmap suggests that while individual stats like points, assists, and playing time are closely related, they don’t seem to tell the whole story when it comes to team success. This points to the need for deeper analysis to better understand what really drives wins.

3.7 Model Accuracy Comparision

Next, we converted the salary variable into a binary feature, HighSalary, indicating whether a player’s salary was above the median. We trained three classification models to predict this label:

Logistic Regression (Accuracy: 70%)

k-Nearest Neighbors (k=5) (Accuracy: 62%)

Decision Tree (Accuracy: 59%)

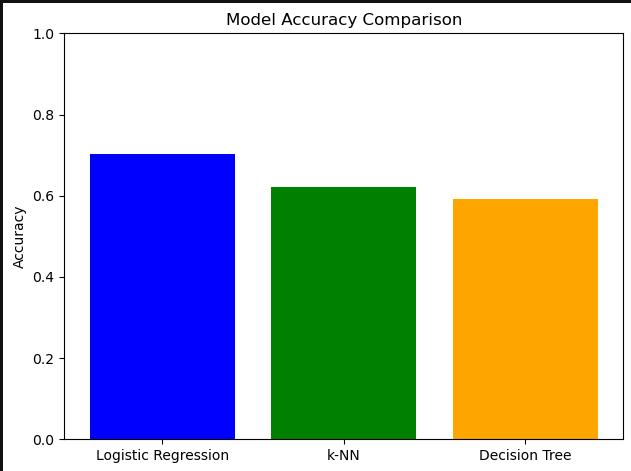


Figure 7. Model Accuracy Comparison

Logistic regression clearly performed the best. The confusion matrix showed that the model correctly identified both low- and high-salary players with reasonable balance. These results suggest that while salary prediction is difficult, it's possible to broadly classify players into salary levels based on their stat lines.

**3.8 Regularized Regression: Ridge and Lasso**

We also tried using Ridge and Lasso regression to see if they could improve how well the model predicts salary. However, both methods performed worse than regular linear regression. Ridge regression had an R² score of 0.24, and Lasso was even lower at 0.22. This suggests that the model is already as good as it can be with the data we have, and adding these methods didn’t help reduce overfitting.

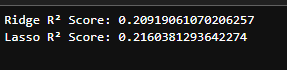


Figure 8. Ridge and Lasso Comparison

**4. Conclusion**

This project looked at NBA player stats, salaries, and team wins to find patterns and insights from the 2022-2023 season. The analysis showed that players who play more minutes tend to score more points and get more assists. Scoring was also closely linked to higher salaries, showing that players who put up big numbers are often paid more.

However, when looking at team wins, the results were less clear. None of the individual player stats had a strong connection to wins, suggesting that other factors, like teamwork or coaching strategies, might be more important for team success.The machine learning models used to predict salary based on performance were somewhat effective, but the accuracy was not very high. This indicates that salary is not just based on stats; it likely includes other factors like market value, reputation, or contract timing.

Overall, the analysis provided some useful insights into how player performance affects salary and how individual stats don’t always translate to team success. Future work could include looking at more seasons of data, exploring advanced metrics, or analyzing how off-court factors, like endorsements or media presence, impact salary.