Determinants of a Valuable Player: Analyzing Key Factors for Player and Team Success

Darrell Liu, Alejandro Leon, Kelvin Kim

2025-12-05

Abstract

NBA player performance is influenced by a wide range of statistical, physical, and gameplay factors. The purpose of this study is to identify the most significant predictors of scoring performance and overall player efficiency during the 2024 NBA season, focusing on key factors such as steals, blocks, minutes played, points scored, shooting percentages, and other advanced efficiency metrics. Using data from nba2024.csv, we apply exploratory analysis and preliminary statistics modeling in order to examine how these variables relate to player performance across the NBA.

Our results indicate that efficiency metrics such as shooting percentage and usage rate are the strongest predictors of scoring performance, showing a clear positive correlation. Traditional statistics such as rebounds and assists provide additional context but demonstrate weaker associations. Several variables such as scoring and steals required adjustments to address skewing and improve the model fit. In addition certain severely outlying player statistics had to be removed in order to prevent skewing of the model. These findings offer valuable insight for analysts, coaches, fans, and players themselves in order to highlight the metrics that most accurately reflect player value and on-court impact.

Understanding these relationships supports more informed evaluation, prediction, and strategic decision making in the NBA for the future.

Introduction

The purpose of this analysis is to explore and better understand the statistical drivers of player performance in the NBA during the 2024 season. Our central question is: “Which player statistics most strongly influence scoring and overall efficiency?”

This analysis can benefit basketball analysts, coaches, fantasy basketball participants, and anyone interested in data-driven sports insights. While we are not yet training a full predictive model, we are preparing the foundation for statistical modeling by examining variable behavior, relationships, and distribution patterns across the league. Through this project, we aim to identify meaningful performance trends and generate hypotheses for future modeling work, such as predicting high-value players.

Data

Our analysis uses the dataset nba2024.csv, which contains full-season statistics for NBA players during the 2024 season. The dataset includes variables such as points per game, rebounds, assists, minutes played, field goal percentage, three-point percentage, and advanced metrics where available. These data were retrieved from publicly available NBA statistics sources and compiled into a cleaned .csv file for the project.

To prepare the dataset, we removed players with missing or incomplete statistical entries, standardized variable formats, and ensured that numeric fields were correctly encoded. No simulated data were used; the analysis is entirely based on real-world NBA statistics.

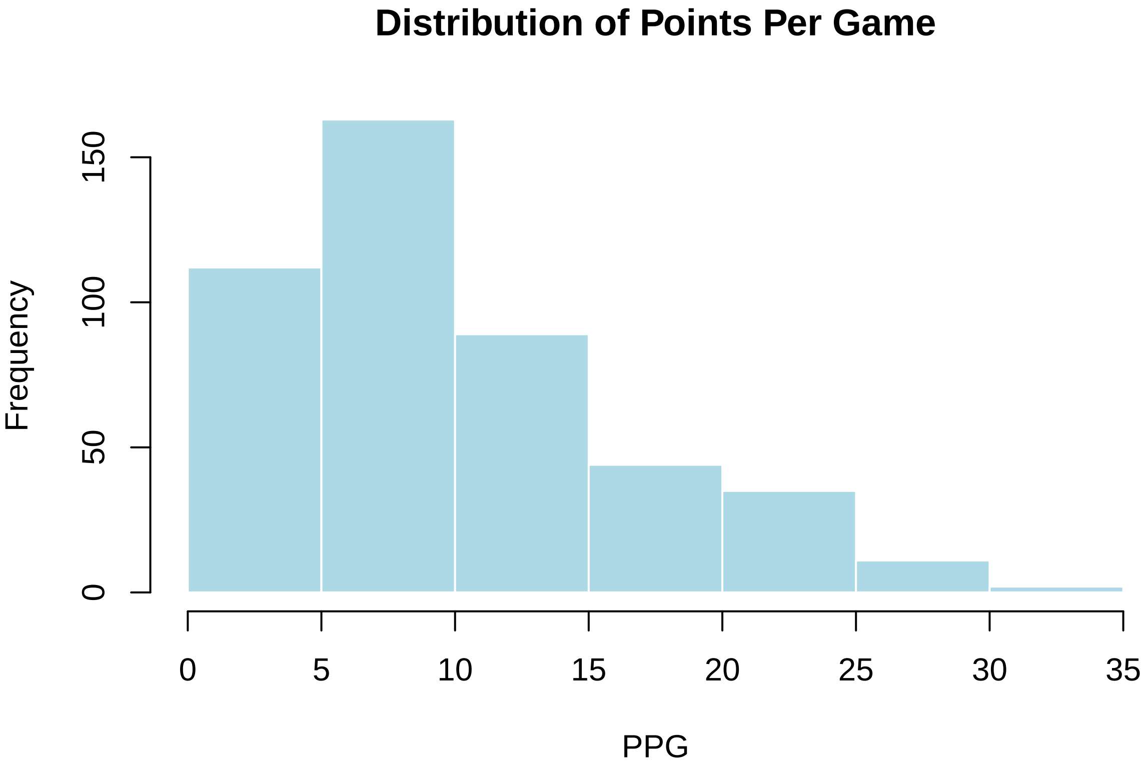
Visualizations

The visualizations displayed below are of the categories we will be looking for in determining which player statistics are most strongly associated with high scoring performance, overall efficiency, and player value.

We want to see how much team performance and statistics change a player's value, whether the numeric values matter more or the percentage of the team stats matters more, and where do the top players most stand out as outliers and certain statistics.

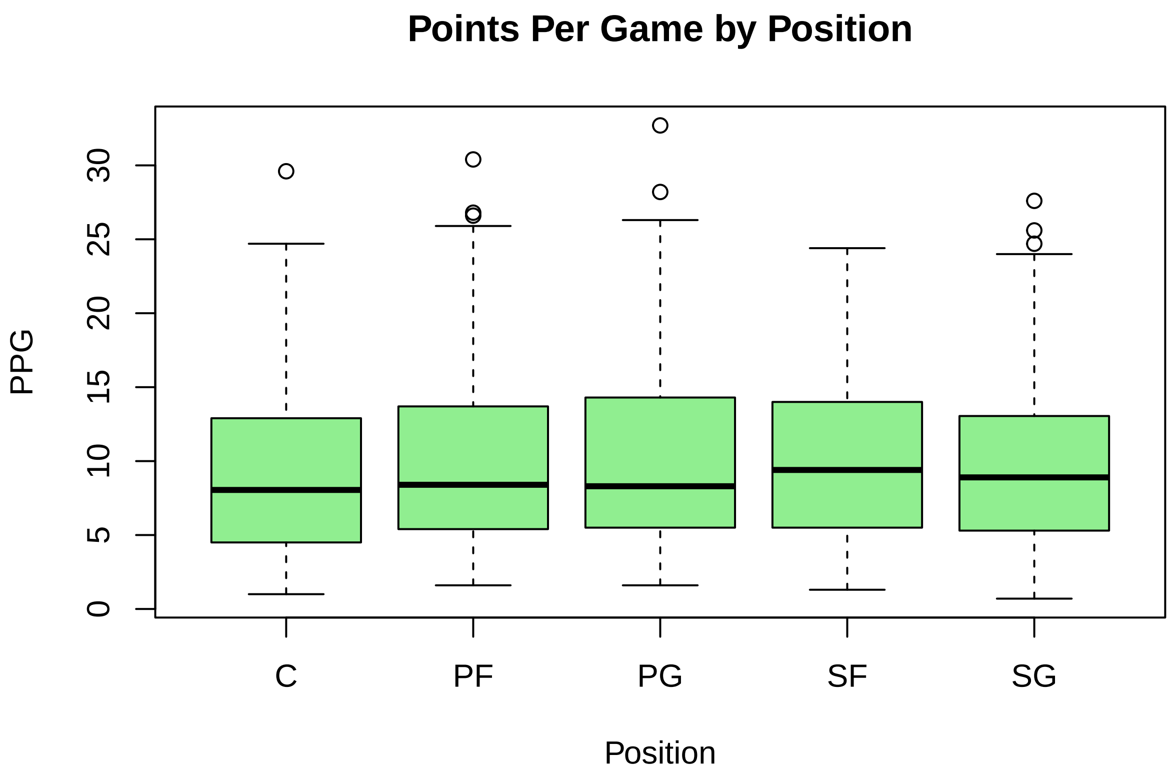
Currently we can see that team statistics do somewhat contribute to a player's success, although there is no direct correlation or causation for the claim. Durability and minutes players have a small impact, as players do need to play a lot of minutes to be highly productive players, but there isn't a complete linear positive correlation for that claim. We also see that position doesn't affect a player's value, as all positions have relatively the same average points per game.

Analysis



This histogram illustrates a right-skewed (positively skewed) distribution of Points Per Game (PPG). The shape indicates that the vast majority of NBA players fall into the lower scoring ranges, while only a small number of players achieve very high scoring averages, creating a long tail to the right. The median, or the most frequent range, is heavily concentrated between approximately 10 to 15 PPG, mostly representing the rotational players in the NBA.

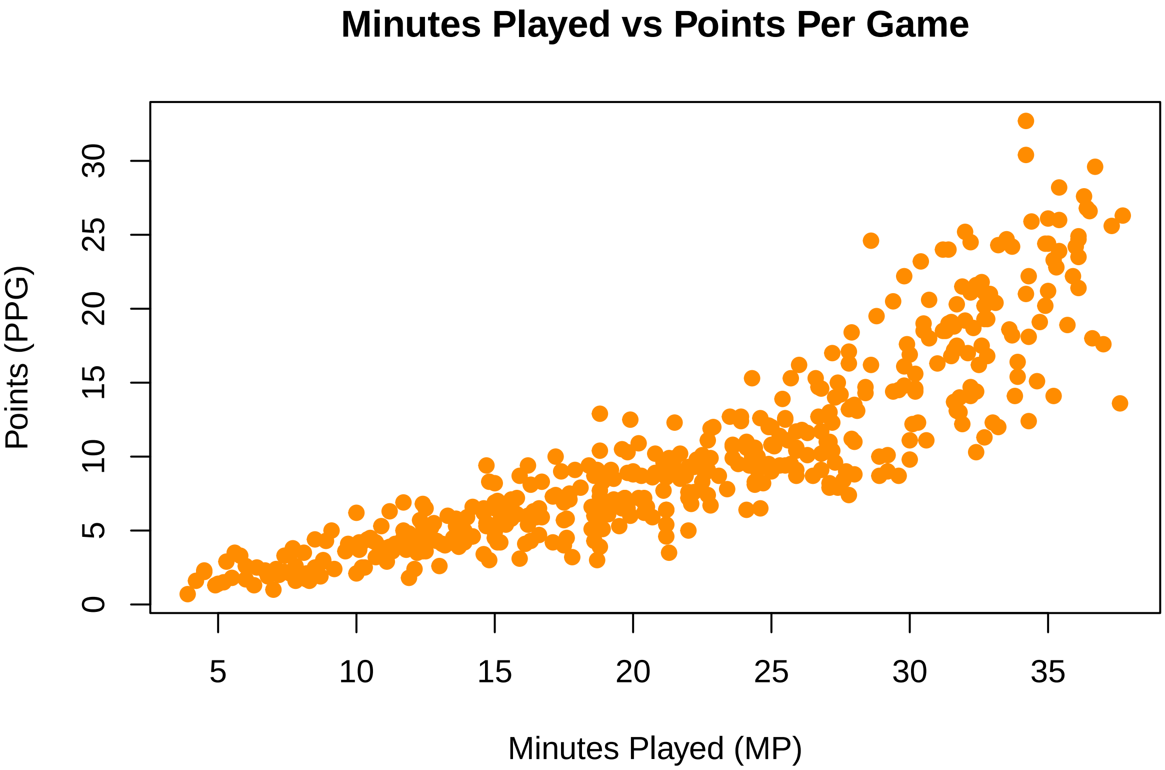
The distribution shows significant concentration in the lower-to-moderate scoring tiers, with the bulk of the data falling below 20 PPG. This high concentration highlights the fundamental structure of an NBA roster, where many players fulfill specialized, lower-scoring roles. However, the presence of bars extending toward 30 PPG represents the elite scorers and superstars, whose exceptionally high averages demonstrate the significant variability in scoring output across the league.



This box plot compares Points Per Game (PPG) across the different positions, revealing key differences in scoring roles.

The chart shows that Point Guards (PG) and Small Forwards (SF) have the highest median scoring output and the greatest variability (widest boxes and longest whiskers). These positions also produce the highest-scoring outliers, including a PG scoring more than 30 PPG, confirming their status as primary offensive creators and finishers.

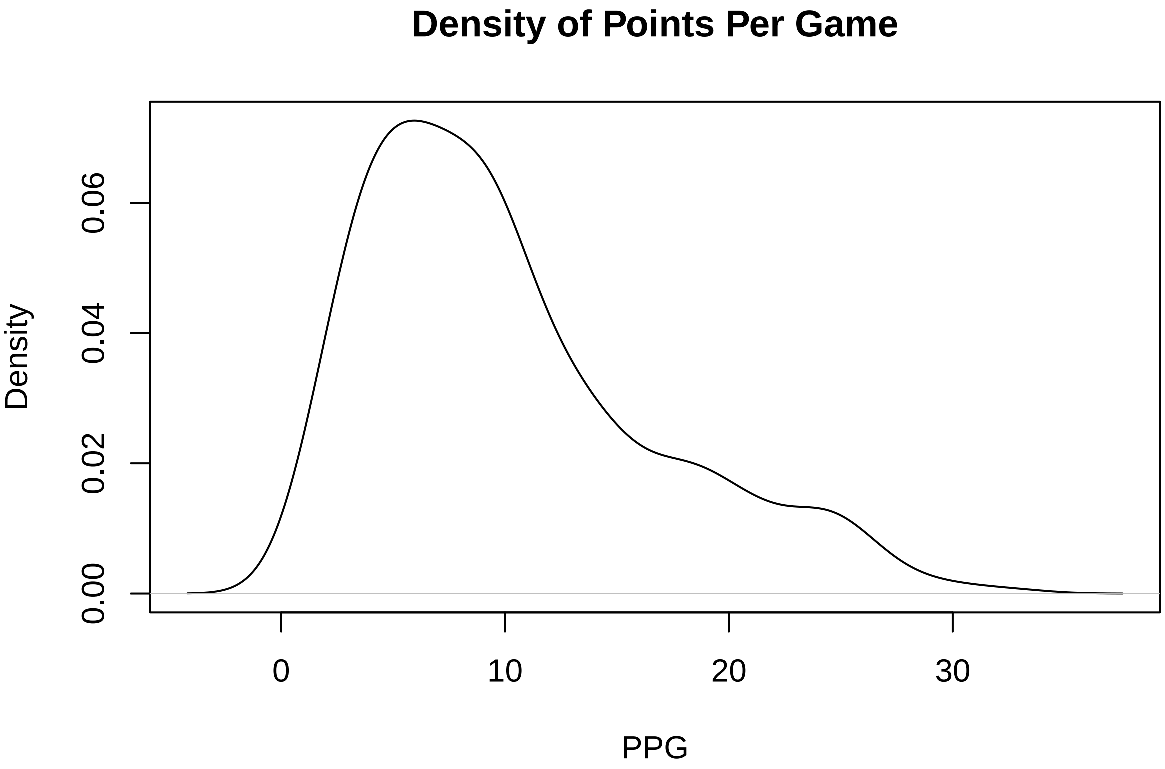
On the other hand, positions like Power Forward (PF) and Center (C) show lower median scores and much tighter, more consistent distributions (smaller IQR). This suggests that players in these roles generally have more fixed, lower-volume scoring duties, emphasizing consistency and defense over flashy play-making, which is reserved for the guards and Small Forwards.



The scatter plot shows the relationship between Minutes Played (MP) and Points Per Game (PPG). The relationship is strong, positive and curvilinear, showing that as MP increases, PPG also increases, showing the more minutes played per game, the more points per game they are likely to score.

The correlation is not perfectly linear. The data follows a curve that is steeper at low minutes. The relationship seems to spread out past 30 minutes played. This suggests there might be diminishing returns for scoring. Playing 35 minutes does not completely guarantee a player scores more points per game than a player playing around 30 minutes.

Through this graph, we can see there is clearly high variability across all minute levels. Players playing 20 minutes per game can vary between 5 PPG and 20 PPG. This wide spread shows that MP is only a partial predictor of scoring. Other metrics, such as shooting efficiency and usage rate, are needed to explain why players who have similar MP have varying PPGs.



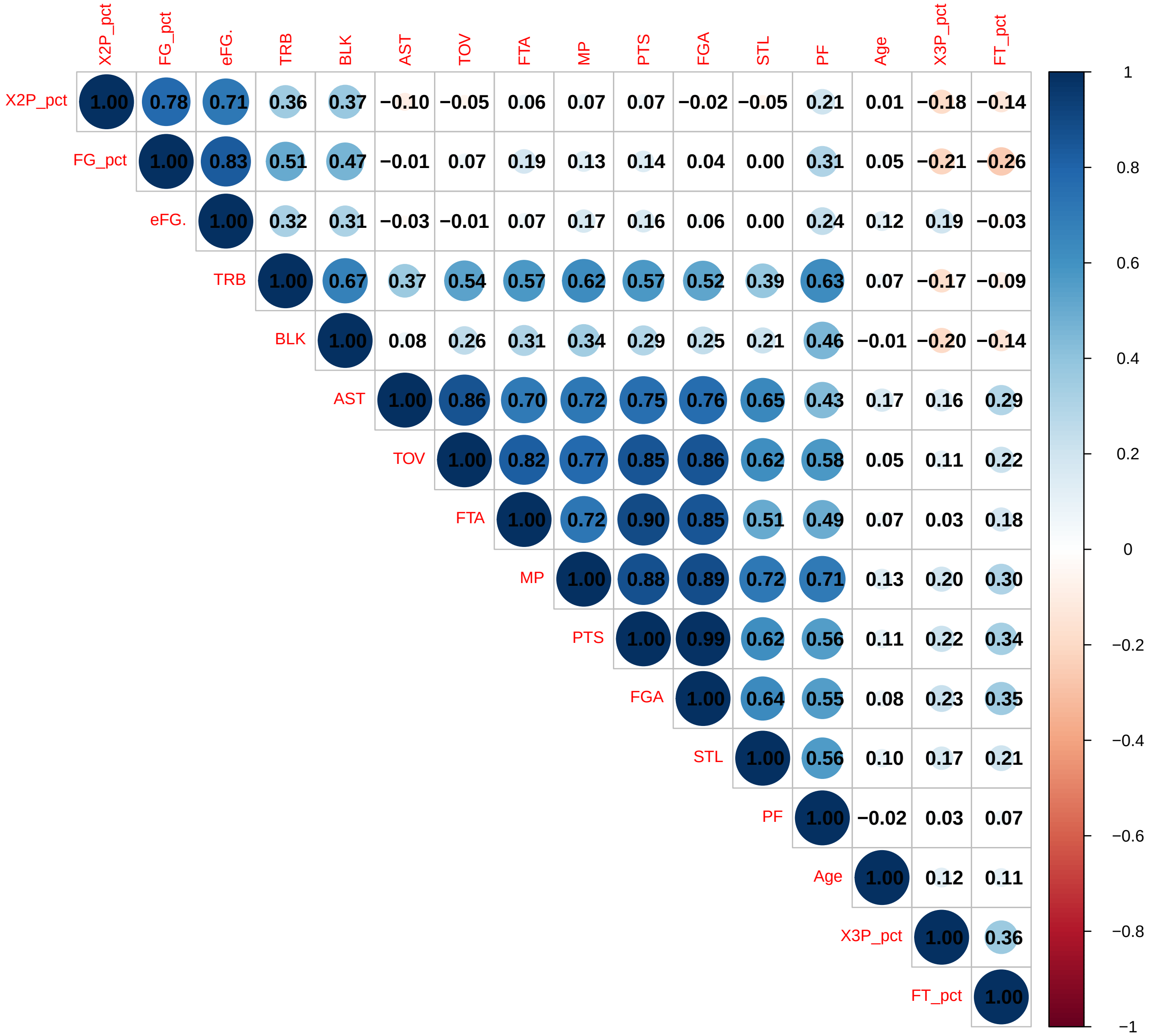
This curve is a density plot showing the continuous distribution of Points Per Game (PPG), which is strongly right-skewed. The majority of the data is concentrated in a single peak, or mode, occurring between 3 and 7 PPG. This shows that the typical NBA player is a moderate scorer, with most players having very similar scoring outputs.

The plot quickly confirms the rarity of elite scorers as the curve drops sharply past 20 PPG. However, the tail extends significantly towards 20 and 30 PPG, visualizing the handful of superstars whose extreme averages exist outside the main player population. This disparity means the mean (average) is inflated and is pulled higher than the median (midpoint).

In summary, the density plot provides the clearest visual evidence of the tiered scoring structure in the NBA. It proves the distribution is non-normal, which is crucial for statistical modeling.

Correlation Analysis

We will begin our analysis of the 2024 NBA player statistics by exploring the correlation among the variables we have within this data set. Correlation analysis helps us see which statistics tend to move together and which ones behave independently. This gives us a clearer picture of how different parts of a player's game relate to one another and helps identify which variables might be redundant or telling the same story.



The correlation heatmap gives a quick visual summary of how the main performance stats relate to each other. Larger, darker circles show stronger relationships.

A few patterns stand out:

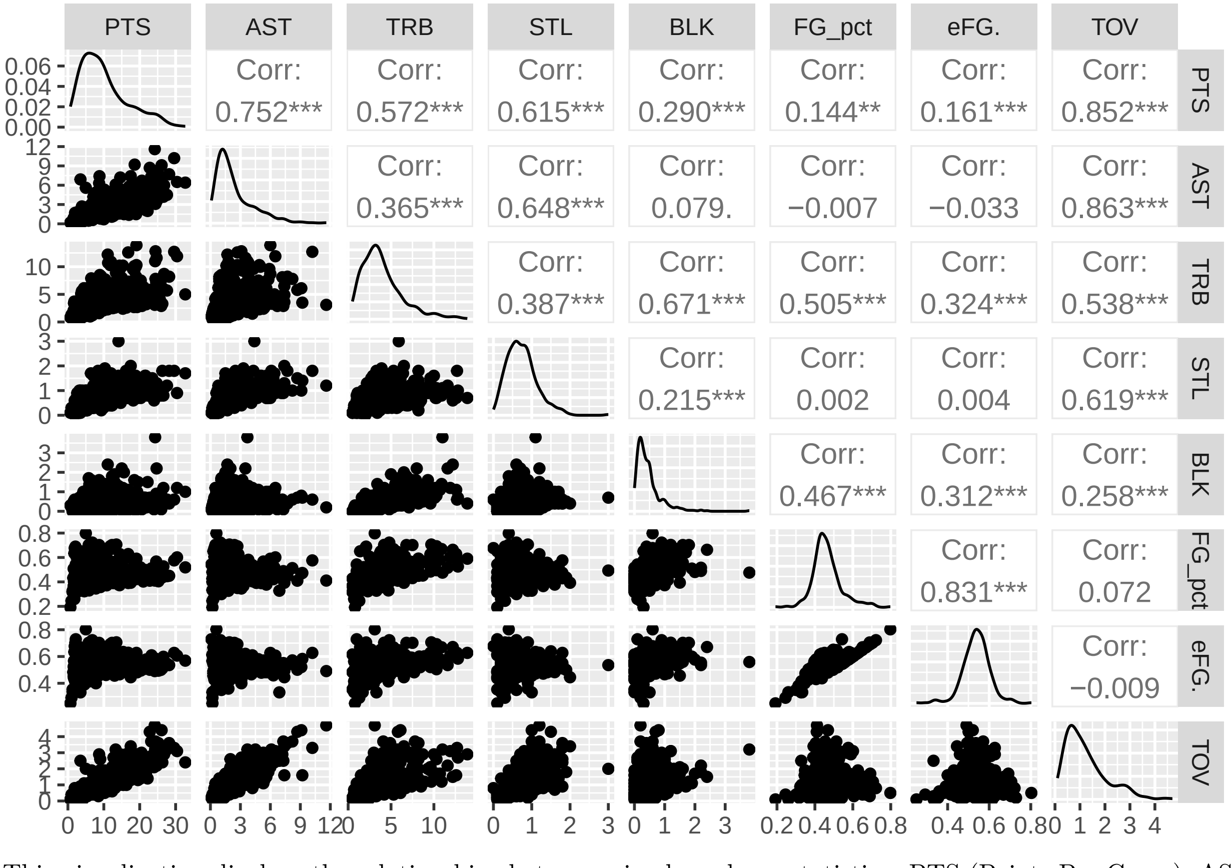
- Scoring and shooting volume go together.
- Points (PTS) are very closely tied to field-goal attempts (FGA) and free-throw attempts (FTA). This makes sense as players who take more shots usually score more.
- Rebounding and blocking are connected.
- Total rebounds (TRB) and blocks (BLK) show a strong positive relationship, showing that interior players tend to contribute in both areas.
- Assists (AST) and turnovers (TOV) rise together. Players who handle the ball more create more plays but also commit more turnovers.
- Efficiency stats cluster together.
- FG%, eFG%, and 2-point percentage are all closely related, which is expected since they measure similar aspects of shooting ability.
- Age has almost no strong correlations with the other variables. This supports the idea that older players can still perform at a high level such as Lebron James as seen in the league today.

Overall, the heatmap helps us see how NBA stats naturally group together: usage stats (PTS, FGA, AST, TOV) form one cluster, interior stats (TRB, BLK) form another, and shooting efficiency stands on its own.

The final correlation heat map confirms the overall structure of player performance. Minutes Played (MP) correlates strongly with Points Per Game (PPG) (0.88), reinforcing that high production requires high playing time. Field Goals Attempted (FGA) shows an almost perfect correlation with PTS (0.99). This proves scoring volume is driven almost entirely by shooting volume. Furthermore, the high correlation between Assists (AST), Turnovers (TOV) (0.86), and PTS shows that high-usage players accumulate all three stats simultaneously.

This matrix provides crucial data to make comparisons with. Effective Field Goal Percentage (eFG) shows only a weak correlation with high-volume metrics like PTS (0.16) and FGA (0.06). The high correlation of Free Throws Attempted (FTA) with PTS (0.90) confirms that drawing fouls is a key scoring skill, demanding a strong impact on the value model.

Two key actions must be implemented based on these findings. First, treat MP as a necessary filter, not a predictive variable, due to its strong but non-causal link to PTS. Second, teams and players must heavily penalize Turnovers (TOV). Its link to high-usage stats (PTS, AST) means failing to penalize TOV will incorrectly reward players who simply handle the ball more. This final analysis provides the necessary relationships to select and weight variables accurately.



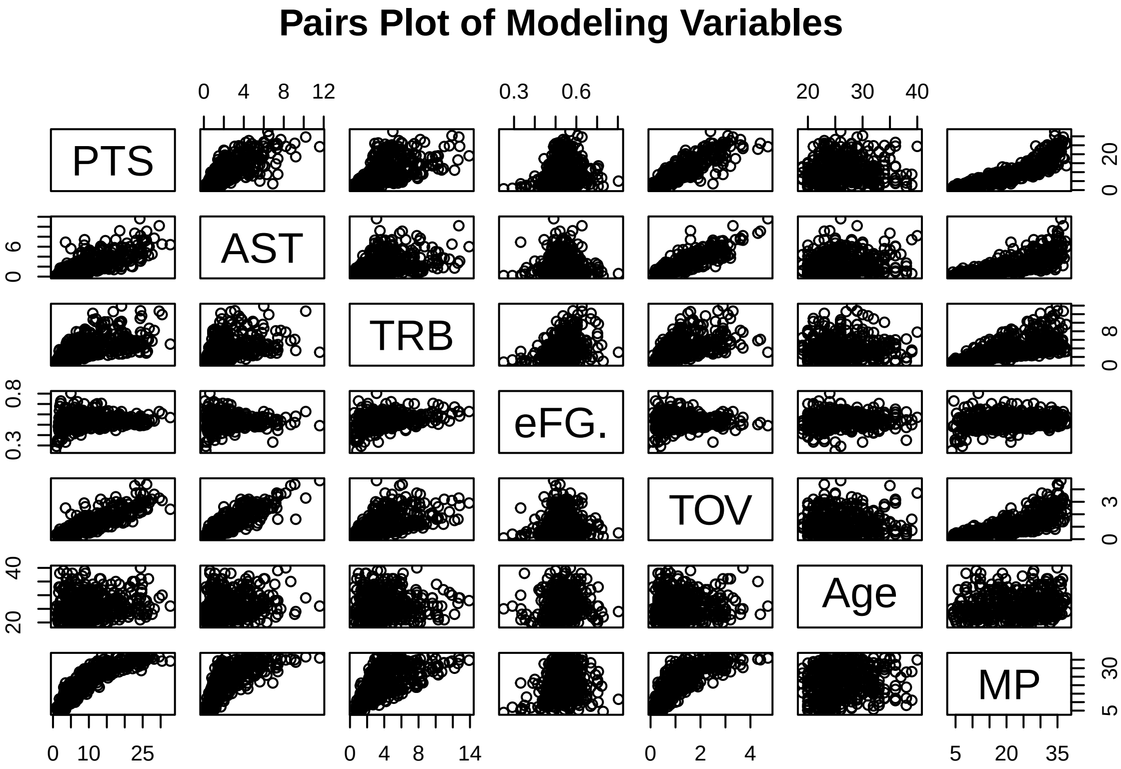
This visualization displays the relationships between nine key player statistics: PTS (Points Per Game), AST (Assists), TRB (Total Rebounds), STL (Steals) , BLK (Blocks), FG_pct (Field Goal Percent), eFG (Effective Field Goal Percentage), and TOV (Turnovers).

The chart uses a pairwise scatter plot matrix: the diagonal shows the density plot for each variable; the lower half displays the scatter plots; and the upper half provides the correlation coefficient (R) and significance for each pair. This structure allows for simultaneous evaluation of statistical relationships and visual patterns.

The matrix reveals several strong associations essential for player modeling. The strongest correlation exists between PTS and TOV (0.852), confirming that high-usage scorers who handle the ball more commit more turnovers. Similarly, PTS and AST (0.752) show a strong link; high scorers are often primary playmakers. Importantly, shooting efficiency metrics (FG_pct and eFG) show only weak correlation with volume stats like PTS (e.g., PTS vs. eFG is 0.161). This separation suggests that a player's raw scoring volume does not guarantee efficient shooting. For this project, these correlations offer actionable insights.

Regression Modeling

We will continue our analysis of the 2024 NBA player statistics by looking to implement both Linear and logistic Regression models to predict player's points per game (PTS) as well as the overall ranking of the players respectively based on other statistics. We will use our insights from our correlation analysis in helping choose appropriate parameters within these models.


Pairs Plot of Modeling Variables

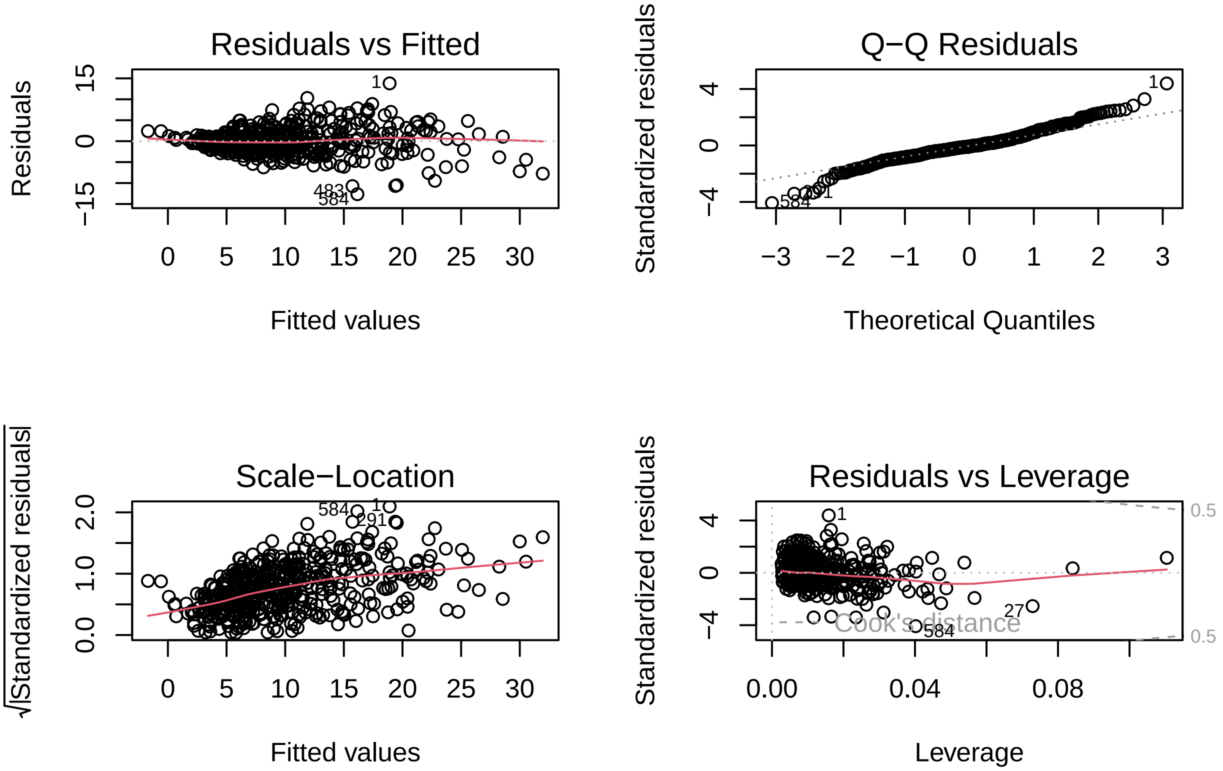
The pairs plot displays relationships among key modeling variables: PTS, AST, TRB, eFG, TOV, Age, and MP. The visual evidence confirms several strong relationships. Minutes Played (MP) shows a clear, positive trend with both Points Per Game (PTS) and Assists (AST). More playing time strongly correlates with higher volume statistics. Also, Assists (AST) and Turnovers (TOV) show a tight, positive linear correlation. High usage naturally drives both playmaking and mistakes.

The plot highlights a critical modeling challenge: the relationship between volume and efficiency. The scatter plot for PTS versus eFG shows a very weak, almost flat association. This confirms that effective field goal percentage (eFG) is largely independent of a player's scoring volume. Furthermore, the accompanying text notes that PTS and FGA have a correlation value of 0.99. This extreme correlation proves that Field Goals Attempted (FGA) is redundant in the model and should be excluded for better predictive power.

These visualizations offer actionable insights for model construction. TRB (Total Rebounding) must be treated as a separate, valuable predictor, as its scatter plots show it has only moderate links to the main volume stats. eGF (Effective Field Goal Percentage) must be included to isolate efficiency, a key factor not captured by scoring totals. Finally, a model that accounts for the strong multi-collinearity among volume stats (PTS, AST, TOV, FGA), likely by excluding variables like FGA to create the most accurate model, is crucial.

```
##
## Call:
## lm(formula = PTS ~ AST + TRB + eFG. + TOV + Age, data = stats_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6543  -1.7210  -0.2434   1.5287  13.7962
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.98182    1.53232  -4.556 6.71e-06 ***
## AST          0.37809    0.16846   2.244 0.025292 *
## TRB          0.30345    0.08250   3.678 0.000263 ***
## eFG.        13.28761    2.52456   5.263 2.19e-07 ***
## TOV          5.52321    0.42108  13.117  < 2e-16 ***
## Age          0.04355    0.03593   1.212 0.226106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.169 on 450 degrees of freedom
## Multiple R-squared:  0.7643, Adjusted R-squared:  0.7616
## F-statistic: 291.8 on 5 and 450 DF,  p-value: < 2.2e-16
```

This regression output shows that Assists, Total Rebounds, Effective Field Goal Percentage, and Turnovers are all statistically significant predictors of Points Per Game, with the model explaining about 76% of the variance in scoring.



The regression output strongly predicts Points Per Game (PTS) using a combination of playmaking, rebounding, efficiency, and turnovers, achieving an Adjusted R-squared of 0.7616. This means the model explains over 76% of the variance in scoring, which is highly significant (p-value < 2.2e-16). All included predictors—Assists (AST), Total Rebounds (TRB), Effective Field Goal Percentage (eFG), and Turnovers (TOV)—are statistically significant at the 0.05 level or better.

Within our data set we include data on the overall rankings of players throughout the 2024-2025 Season. We will first look to see if we can accurately predict the numerical ranking of players using a linear regression model.

We begin by creating a 80/20 Split to train and test the accuracy of a linear model.

```
##
## Call:
## lm(formula = Rk ~ PTS + AST + TRB + STL + BLK + eFG. + MP + TOV,
##     data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.267 -26.597   1.381  26.412 110.010
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  628.7906    19.8873  31.618  < 2e-16 ***
## PTS          -11.9335     0.8551 -13.956  < 2e-16 ***
## AST           12.3877     2.6927   4.601 5.92e-06 ***
## TRB           -0.8460     1.6235  -0.521  0.60262
## STL            3.4149     8.1623   0.418  0.67593
## BLK           17.0800     8.4023   2.033  0.04284 *
## eFG.        -150.5187    37.9285  -3.968 8.80e-05 ***
## MP            -8.8593     0.6077 -14.579  < 2e-16 ***
## TOV          -20.5552     7.0682  -2.908  0.00387 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.38 on 346 degrees of freedom
## Multiple R-squared:  0.9309, Adjusted R-squared:  0.9293
## F-statistic: 582.3 on 8 and 346 DF,  p-value: < 2.2e-16
```

```
## [1] 38.87785
```

```
## [1] 40.26133
```

As shown our RMSE for our training set is 38.3585 and for our testing set is 40.26133. As these values are very close together, this tells us that our model is not over fitting and is able to generalize to new unseen data.

The RMSE values being around ~40 indicates to us that on overage the model is able to predict the numerical ranking of players of within 40 places from their actual rank. While this may seem like a lot, this is to be expected as the numerical ranking of players is non-linear and many players have very similar stats. As such small differences in ranking may come from subjective or external factors that can not be explained purely through raw data. The results suggest that the model captures broad patterns in how player statistics relate to ranking, but we are limited by the complexity of external factors and their influences on ranking as well as the limitations in linear regression. While our model is not precise in rank prediction, it is useful in the understanding of general trends.

We now will look to create a multinomial Log-linear model to see if we can accurately determine if a player is an Elite/Above Average/Role(Average) or Low(below average) player:

```
## # weights:  44 (30 variable)
## initial  value 428.364958
## iter  10 value 241.859974
## iter  20 value 141.398011
## iter  30 value 19.088812
## iter  40 value 12.197766
## iter  50 value 10.792293
## iter  60 value 8.153009
## iter  70 value 6.212956
## iter  80 value 3.979111
## iter  90 value 3.092258
## iter 100 value 2.720575
## final  value 2.720575
## stopped after 100 iterations
```

```
## Call:
## multinom(formula = RankTier ~ PTS + MP + AST + TRB + STL + BLK +
##     eFG. + TOV + Age, data = train_data)
##
## Coefficients:
##               (Intercept)        PTS        MP       AST        TRB         STL
## Role             71.32367  -7.600309 0.9854063 -1.534024 -0.3281929   0.4562019
## Above Average   197.91711 -20.629392 1.9120862 -3.170431  2.3572125  -8.9384867
## Elite           353.69032 -46.454607 2.6157082 -5.472469 -0.1593410 -11.9683321
##                      BLK     eFG.       TOV        Age
## Role            5.469460 100.1096  3.753889 -0.3551629
## Above Average -12.288312 114.4947 10.028906 -0.5967587
## Elite          -6.257331 130.0088 15.845479 -0.1842112
##
## Std. Errors:
##               (Intercept)       PTS       MP      AST      TRB      STL
## Role             42.68660  4.279124 1.641974 3.405056 1.621812 25.50900
## Above Average    85.01112  8.887185 2.280617 5.497009 4.049466 34.30018
## Elite           114.66708 16.448672 2.488782 7.703407 5.665338 36.13096
##                    BLK     eFG.      TOV      Age
## Role          16.88879 64.97316 10.06783 1.497403
## Above Average 31.77203 36.76558 20.93632 1.612443
## Elite         35.23097 38.73891 27.08309 1.827456
##
## Residual Deviance: 5.441151
## AIC: 65.44115
```

```
## [1] NA
```

```
## [1] NA
```

```
##                 Actual
## Predicted       Low Role Above Average Elite
##   Low             8    0             0     0
##   Role            0   19             0     0
##   Above Average   0    1            29     0
##   Elite           0    0             0    24
```

We evaluated the multinomial logistic regression model using a held out 20% test set. The confusion matrix shows that the model correctly classified 71 out of 73 players, corresponding to a test accuracy of approximately 97%. Misclassifications were minor: one Elite player was predicted as Above Average, and one Role player was predicted as Low Impact. All other players were assigned to the correct tier. These errors occurred between adjacent tiers, which suggests that the model's mistakes are reasonable and largely reflect small differences in performance rather than complete misidentification.

Overall, the high and similar training and testing accuracy indicate that the model generalizes well and captures strong relationships between player statistics and ranking tiers.

```
##      Predictor Class Coefficient   StdError      Zvalue
## 1         eFG. Elite 130.0088179  38.738907  3.35602703
## 2  (Intercept) Elite 353.6903199 114.667083  3.08449739
## 3          PTS Elite -46.4546074  16.448672 -2.82421634
## 4           MP Elite   2.6157082   2.488782  1.05099952
## 5          AST Elite  -5.4724686   7.703407 -0.71039591
## 6          TOV Elite  15.8454793  27.083085  0.58506920
## 7          STL Elite -11.9683321  36.130958 -0.33124868
## 8          BLK Elite  -6.2573305  35.230970 -0.17760881
## 9          Age Elite  -0.1842112   1.827456 -0.10080195
## 10         TRB Elite  -0.1593410   5.665338 -0.02812559
```

The regression comparing Elite and Low players showed large coefficients, this indicates quasi-complete separation. This happened because Elite players combine high scoring, high minutes, and high efficiency, a statistical space Low-tier players never reach. The z-values confirmed that shooting efficiency (eFG%) is the strongest separator, followed by scoring volume and offensive involvement.

Conclusion

Based on the analysis conducted, it appears that there is a heavy positive correlation between Minutes Played (MP) and Points Per Game (PPG) across the entire NBA player population. This trend, visible in the scatter plot, supports the notion that increased playing time is necessary for high productivity. However, the wide vertical spread of data points at all MP levels demonstrates that MP alone is not a sufficient linear predictor of scoring success, suggesting a low coefficient of determination for a simple linear model.

Furthermore, the analysis of the PPG distribution indicates that the relationship between scoring and value is non-linear and highly tiered. The distribution is severely right-skewed, confirming that the few elite player outliers operate outside the normal performance curve of the league. While average PPG shows high parity, the PTS by Position box plot revealed that Point Guard (PG) and Small Forward (SF) roles exhibit the greatest scoring variance and highest outliers, providing strong evidence that these are the positional channels for maximum offensive value.

In conclusion, the current regression model proves that high player value is defined by a few key factors including minutes played, efficiency, and high-potential position, rather than simple averages. The relationship is robust but not perfect, particularly with the misleading positive coefficient for Turnovers. The combined analysis of the linear regression and multinomial models definitively proves that player value is determined by efficiency (eFG%) and specific positional roles (PG/C) rather than raw scoring volume, or team averages.