Predicting NBA Scores: An Analysis of Player Performance Metrics

Anton Yang

2024-10-22

Abstract

In this project, we analyze NBA player performance metrics from the 2023 season to predict total points scored using Poisson and quasi-Poisson regression models. The dataset includes 20 predictor variables, such as player position, shooting percentages, rebounds, and turnovers. Exploratory data analysis revealed that factors like player position and turnovers significantly impact scoring. A Poisson model was initially developed, but overdispersion issues led to the use of a quasi-Poisson model. The final model highlights key predictors, including player position and turnovers, while addressing the challenges of overdispersion, providing insights into factors that influence scoring outcomes in the NBA.

Introduction

The National Basketball Association (NBA) is a professional basketball league comprised of 30 teams across North America featuring the best basketball players in the world. NBA was founded on Augst 3, 1949, with the merger of the Basketball Association of America (BAA) and the national Basketball League (NBL). Since then, the NBA has grown into a global sports phenomenon, captivating millions of fans with its fast-paced games, high scoring, and star players. Every games is a competition between team such as team strategies, player performance, defense, nad offensive capabilities come together to influend the final score.

One area of interest for analysts, coaches, and fans is what factors drive the total points scored in a game. Scoring in the NBA is influenced by numerous variables including player efficiency, player's position, shooting accuracy, and defense. Understanding these variables can help in game predictions and performance analysis. Predicting basketball points for each player is one of the most important for basketball analytics. It serves as a crucial performance metric that allows coaches ,analysts, and fans to assess a player's scoring ability and overall offensive contribution to the team. Understanding players' scoring potential aids in strategic decision-making during games, player selection, and talent scouting.

This project is interested in analyzing the factors of player's performance (total points),and the goal is to analyze the NBA dataset and perform predictive analytics on the dataset. To approach this, we perform explanatory data analysis and employ Poisson Regression Model to predict the total points scored. Poisson Regression is particularly good for count data, where the dependent variable represents the number of occurrences of an event in a fixed period. The Poisson distribution is commonly used to model such events.

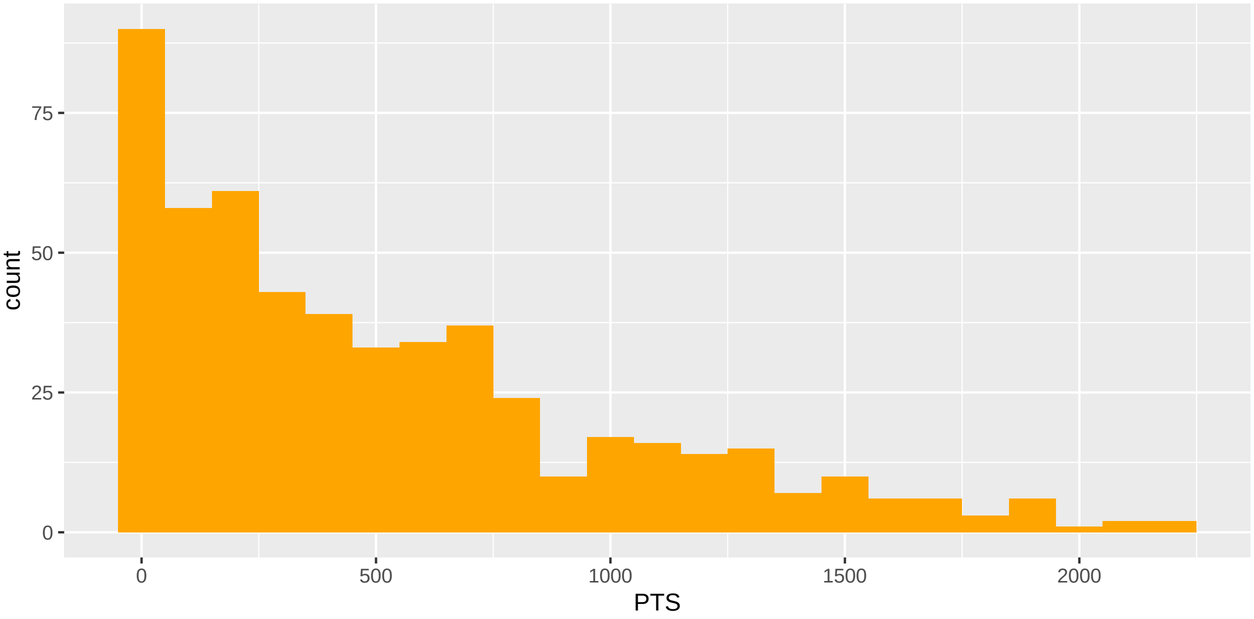
Data

We utilized the NBA Players Stats (2023 Season) dataset from Kaggle, which comprises 20 predictor variables and 1 target variable. This dataset was specifically curated to provide performance metrics for analysis. To enhance clarity and reduce redundancy, several variables from the original dataset were removed. For instance, the original dataset included both free throw attempts and free throw success; these were consolidated into a single variable representing free throw success percentage. Our objective is to leverage this dataset to develop an informed predictive model, facilitating deeper insights into player performance.

Table 1: Explanatory Variable Description

Variables	Description
PName	The name of the basketball player
POS	The player's position in the game
Team	The abbreviation of the team the player is currently playing for this season
Age	The age of the player
GP	The percentage of Wins
WP	The total number of games the player has played in this season
MinP	The average minutes the player has played in this season per game
FGP	The percentage of successful field goals made by the player
X3PP	The percentage of successful 3-point field goals made by the player
FTP	The percentage of successful free throws made by the player
OREBP	The total number of offensive rebounds made by the player
DREBP	The total number of defensive rebounds made by the player
ASTP	The average number of assists made by the player per game
TOVP	The average number of turnovers made by the player per game
STLP	The average number of steals made by the player per game
BLKP	The average number of blocks made by the player per game
PFP	The average number of personal fouls made by the player per game
DD2	The total number of double-doubles made by the player
TD3	The total number of triple-doubles made by the player
FP	The total number of NBA fantasy points made by the player
Score.Difference	The total difference between the player's team scoring and the opponents' scoring while the player is in the game
PTS	The total points made by the player

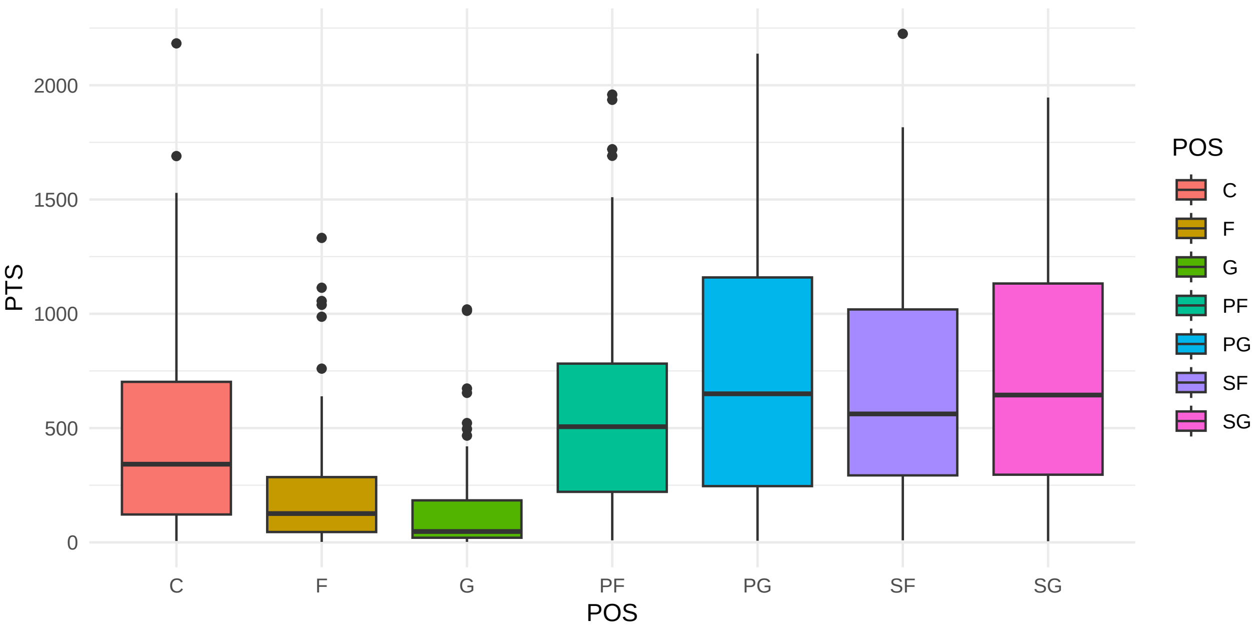
Figure 1: Historgram of Total Points



As illustrated in Figure 1, the distribution of total points is right-skewed, with the majority of players scoring below 1,000 points. Notably, approximately 100 players scored fewer than 100 total points. Given this distribution, we will utilize Poisson Regression to predict total points based on player performance.

To begin, we aim to investigate whether player position has any impact on total points scored. To achieve this, we will employ a boxplot to visually assess the differences in total points across various positions.

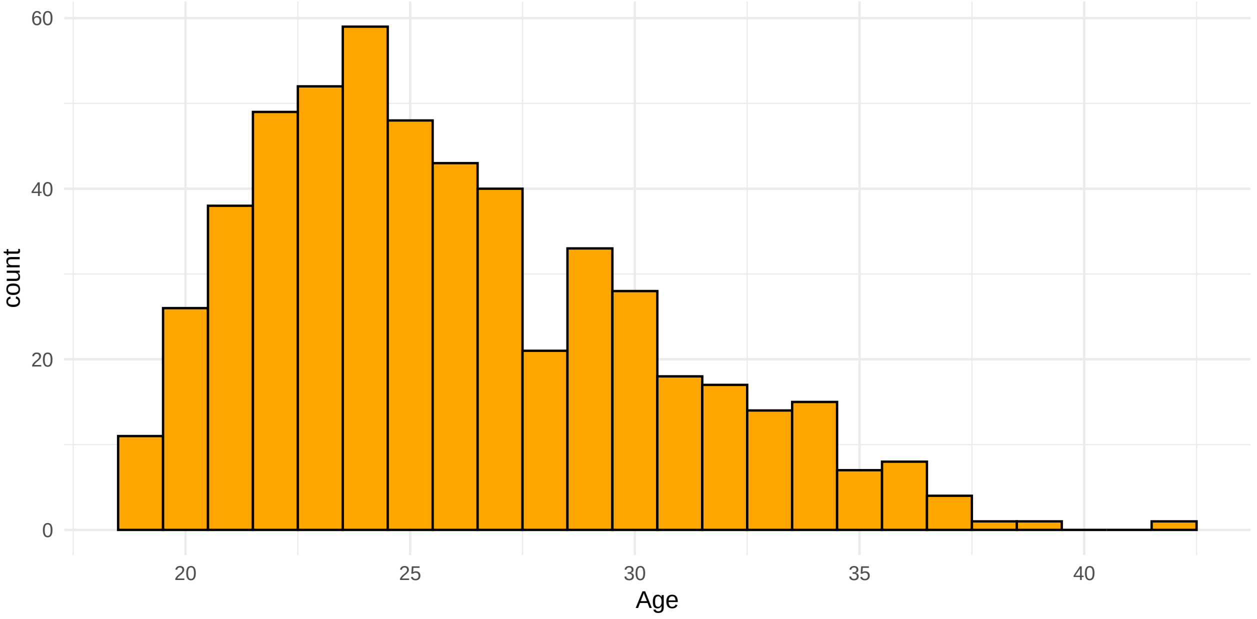
Figure 2: Boxplot of Points Scored by Player Position



As shown in Figure 2, there are significant differences in point scores among the various positions. For instance, the positions of Power Forward (PF), Point Guard (PG), Small Forward (SF), and Shooting Guard (SG) have significantly higher point totals compared to the positions of Guard (G) and Forward (F). This indicates that the variable POS is a significant predictor in the Poisson model.

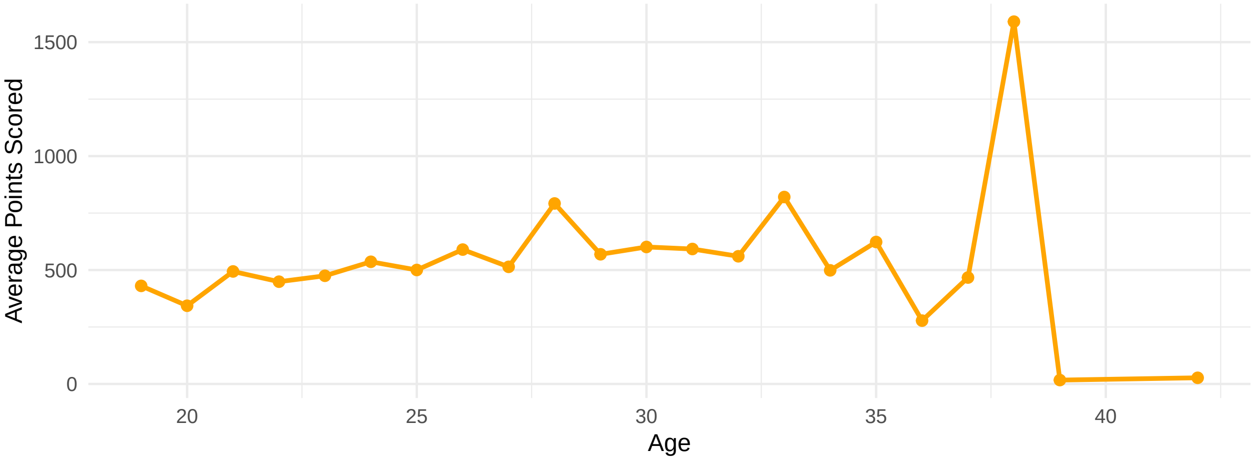
Next, we will examine the distribution of players' ages to determine whether age has an impact on player performance.

Figure 3: Distribution of Player's Age



Based on Figure 3, we can also see that the age is skewed to the right as majority of the player are between the age of 20 and 30.

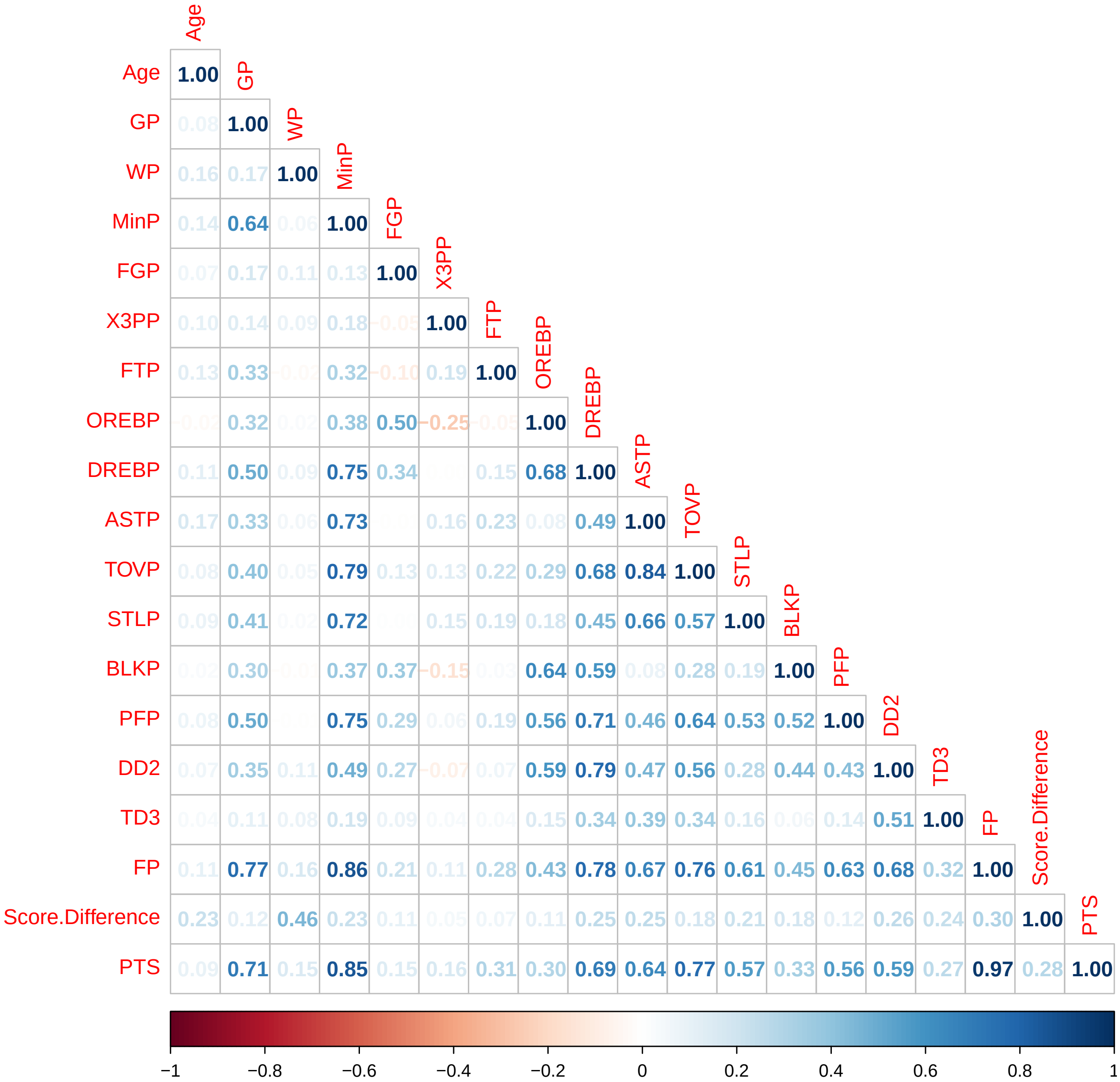
Figure 4: Average Points Scored by Age



As shown in Figure 4, the average points scored by players aged 20 to 30 remain relatively stable. In contrast, players aged 30 and older exhibit more atypical average scores. This variability may be attributed to the limited data available for this age group, which could explain the more pronounced fluctuations in average points as players age.

Correlation is a statistical measure that describes both the strength and direction of the relationship between two quantitative variables. Therefore, we will exclude the categorical variables PName, POS, and Team from our analysis. The correlation coefficient ranges from -1 to +1, where -1 indicates a strong negative correlation, and +1 indicates a strong positive correlation.

Figure 5: Correlation Plot



As shown in Figure 5, most variables display only modest correlations with one another. However, two pairs of variables warrant attention: FP (Fantasy Points) and MinP (Minutes Played), which have a high correlation of 0.86, and ASTP (Assists Points) and TOVP (Turnovers per Game), with a correlation of 0.84. These values exceed the threshold of 0.8, indicating potential collinearity that must be addressed during model building. Furthermore, all variables demonstrate positive correlations with the target variable PTS (Total Points Scored), suggesting that we can expect the model coefficients to be predominantly positive. However, the high correlation of FP and MinP with PTS also raises concerns about potential collinearity, which could impact the reliability of our model's estimates.

Model

For this data, we will be using Poisson model. We have a count responses:

$$Y_i \sim Pois(\mu_i)$$

The predictors $\boldsymbol{x}_i$ and by using a log link function:

$$\log \mu_i = \eta_i = \boldsymbol{x}_i^T \beta \implies \mu_i = e^{\eta_i} = e^{\boldsymbol{x}_i^T \beta}$$

The log-likelihood is

$$\ell(\beta) = \sum_{i=1}^{n} (y_i \boldsymbol{x}_i^T \beta - e^{\boldsymbol{x}_i^T \beta} - \log(y_i!))$$

Differentiating β_j give the Maximum Likelihood Estimate as the solution to:

$$\sum_{i=1}^{n} (y_i - e^{\boldsymbol{x}_i^T \hat{\beta}}) x_{ij} = 0, \forall j$$

For the Poisson model, the exponential family is

$$f(y|\theta, \phi) = \frac{e^{-\mu} \mu^y}{y!} = e^{y \log \mu - \mu - \log(y!)}$$

with $\theta = \log \mu$, which means log link. $\phi = 1$ means that it is a standard Poisson model, and if it's quasi-Poisson, then ϕ will change. Lastly, $a(\phi) = 1, b(\theta) = e^{\theta}, c(y, \phi) = -\log(y!)$. Since the standard Poisson model mean and variance are the same, $\mu = V(\mu) = e^{\theta}$.

Figure 6: Poisson Regression Model Summary

```
##                      Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept)       2.5700e+00  3.9566e-02  64.9567 < 2.2e-16
## POSF              1.0447e-02  1.2135e-02   0.8609  0.389306
## POSG              2.8968e-02  1.6382e-02   1.7683  0.077011
## POSPF             6.7185e-03  9.5132e-03   0.7062  0.480045
## POSPG             2.1082e-01  1.3440e-02  15.6869 < 2.2e-16
## POSSF             1.1144e-01  1.1917e-02   9.3514 < 2.2e-16
## POSSG             1.2161e-01  1.2505e-02   9.7253 < 2.2e-16
## Age              -1.7341e-03  5.7971e-04  -2.9914  0.002777
## GP                1.7352e-02  2.3362e-04  74.2754 < 2.2e-16
## WP                1.8979e-01  2.5627e-02   7.4060 1.302e-13
## MinP              3.7965e-02  7.8124e-04  48.5955 < 2.2e-16
## FGP               8.2657e-03  4.4887e-04  18.4146 < 2.2e-16
## X3PP              2.1147e-03  2.4226e-04   8.7290 < 2.2e-16
## FTP               9.5325e-03  2.5140e-04  37.9183 < 2.2e-16
## OREBP             1.5767e-02  5.6511e-03   2.7901  0.005270
## DREBP             3.2892e-02  3.1279e-03  10.5156 < 2.2e-16
## ASTP             -6.7998e-02  2.8371e-03 -23.9675 < 2.2e-16
## TOVP              2.3905e-01  6.2725e-03  38.1113 < 2.2e-16
## STLP             -7.4490e-02  9.1281e-03  -8.1605 3.335e-16
## BLKP             -1.5150e-02  7.0693e-03  -2.1431  0.032104
## PFP              -7.1027e-02  5.1217e-03 -13.8679 < 2.2e-16
## DD2              -6.4862e-03  4.4470e-04 -14.5855 < 2.2e-16
## TD3               3.6947e-03  1.6949e-03   2.1798  0.029270
## FP                2.8851e-04  9.0409e-06  31.9119 < 2.2e-16
## Score.Difference -1.7478e-04  1.8402e-05  -9.4978 < 2.2e-16
##
## n = 427 p = 25
## Deviance = 11879.85661 Null Deviance = 194913.94303 (Difference = 183034.08642)
```

Before constructing the predictive model, we divided the dataset into a training set and a test set, using an 80-20 split. The training set will be employed for model development, while the test set will serve as a benchmark to evaluate the model's performance. As illustrated in Figure 5, most variables are significant to the model, except for POSF (Forward), POSG (Guard), and POSPF (Power Forward). The model has an AIC (Akaike Information Criterion) of 15093.21, leading us to question whether the model would perform better without the POS variable. According to Figure 6, removing the POS variable increases the significance of STLP (Steals per Game). However, the AIC rises to 15580.73 when the POS variable is excluded, prompting us to retain it in the model.

Figure 7: Poisson Regression Model without Position Summary

```
##                     Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept)       2.6707e+00  3.5146e-02  75.9880 < 2.2e-16
## Age              -1.9078e-03  5.6946e-04  -3.3503 0.0008074
## GP                1.7663e-02  2.2739e-04  77.6746 < 2.2e-16
## WP                2.1764e-01  2.5479e-02   8.5419 < 2.2e-16
## MinP              3.9511e-02  7.4565e-04  52.9881 < 2.2e-16
## FGP               7.8041e-03  4.2909e-04  18.1876 < 2.2e-16
## X3PP              1.8477e-03  2.3989e-04   7.7022 1.337e-14
## FTP               9.3875e-03  2.4866e-04  37.7525 < 2.2e-16
## OREBP            -1.2837e-03  5.5380e-03  -0.2318 0.8166885
## DREBP             1.7203e-02  2.9843e-03   5.7644 8.194e-09
## ASTP             -5.0109e-02  2.5686e-03 -19.5081 < 2.2e-16
## TOVP              2.3437e-01  6.1191e-03  38.3009 < 2.2e-16
## STLP             -6.3738e-02  8.8561e-03  -7.1971 6.149e-13
## BLKP             -2.7762e-02  6.7461e-03  -4.1152 3.868e-05
## PFP              -8.0892e-02  4.9726e-03 -16.2675 < 2.2e-16
## DD2              -6.7024e-03  4.2615e-04 -15.7278 < 2.2e-16
## TD3               4.6162e-03  1.6297e-03   2.8325 0.0046187
## FP                2.9543e-04  8.9492e-06  33.0114 < 2.2e-16
## Score.Difference -1.8251e-04  1.8309e-05  -9.9685 < 2.2e-16
##
## n = 427 p = 19
## Deviance = 12379.37491 Null Deviance = 194913.94303 (Difference = 182534.56812)
```

According to Figure 7, removing the POS predictor results in a higher deviance for the model, indicating a deterioration in fit. This suggests that including the POS variable is crucial for enhancing the model's performance. Therefore, it is advisable to retain the POS predictor in the final model to ensure optimal predictive accuracy.

Figure 8: Stepwise Poisson Regression Model Summary and Confidence Interval

```
##                     Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept)       2.5700e+00  3.9566e-02  64.9567 < 2.2e-16
## POSF              1.0447e-02  1.2135e-02   0.8609  0.389306
## POSG              2.8968e-02  1.6382e-02   1.7683  0.077011
## POSPF             6.7185e-03  9.5132e-03   0.7062  0.480045
## POSPG             2.1082e-01  1.3440e-02  15.6869 < 2.2e-16
## POSSF             1.1144e-01  1.1917e-02   9.3514 < 2.2e-16
## POSSG             1.2161e-01  1.2505e-02   9.7253 < 2.2e-16
## Age              -1.7341e-03  5.7971e-04  -2.9914  0.002777
## GP                1.7352e-02  2.3362e-04  74.2754 < 2.2e-16
## WP                1.8979e-01  2.5627e-02   7.4060 1.302e-13
## MinP              3.7965e-02  7.8124e-04  48.5955 < 2.2e-16
## FGP               8.2657e-03  4.4887e-04  18.4146 < 2.2e-16
## X3PP              2.1147e-03  2.4226e-04   8.7290 < 2.2e-16
## FTP               9.5325e-03  2.5140e-04  37.9183 < 2.2e-16
## OREBP             1.5767e-02  5.6511e-03   2.7901  0.005270
## DREBP             3.2892e-02  3.1279e-03  10.5156 < 2.2e-16
## ASTP             -6.7998e-02  2.8371e-03 -23.9675 < 2.2e-16
## TOVP              2.3905e-01  6.2725e-03  38.1113 < 2.2e-16
## STLP             -7.4490e-02  9.1281e-03  -8.1605 3.335e-16
## BLKP             -1.5150e-02  7.0693e-03  -2.1431  0.032104
## PFP              -7.1027e-02  5.1217e-03 -13.8679 < 2.2e-16
## DD2              -6.4862e-03  4.4470e-04 -14.5855 < 2.2e-16
## TD3               3.6947e-03  1.6949e-03   2.1798  0.029270
## FP                2.8851e-04  9.0409e-06  31.9119 < 2.2e-16
## Score.Difference -1.7478e-04  1.8402e-05  -9.4978 < 2.2e-16
##
## n = 427 p = 25
## Deviance = 11879.85661 Null Deviance = 194913.94303 (Difference = 183034.08642)
```

```
##                          2.5 %        97.5 %
## (Intercept)       2.4924110391  2.6475050973
## POSF             -0.0133515471  0.0342173445
## POSG             -0.0031796573  0.0610364661
## POSPF            -0.0119237137  0.0253675481
## POSPG             0.1844883986  0.2371703823
## POSSF             0.0880882941  0.1348030048
## POSSG             0.0971117811  0.1461302412
## Age              -0.0028707567 -0.0005983504
## GP                0.0168946218  0.0178103963
## WP                0.1395671585  0.2400230763
## MinP              0.0364335125  0.0394959320
## FGP               0.0073857493  0.0091452745
## X3PP              0.0016390430  0.0025886672
## FTP               0.0090406550  0.0100261114
## OREBP             0.0046864484  0.0268383948
## DREBP             0.0267631743  0.0390244196
## ASTP             -0.0735611015 -0.0624398909
## TOVP              0.2267618946  0.2513498328
## STLP             -0.0923876963 -0.0566061370
## BLKP             -0.0290200193 -0.0013089145
## PFP              -0.0810659487 -0.0609893517
## DD2              -0.0073583464 -0.0056151445
## TD3               0.0003686917  0.0070127965
## FP                0.0002707908  0.0003062305
## Score.Difference -0.0002108486 -0.0001387127
```

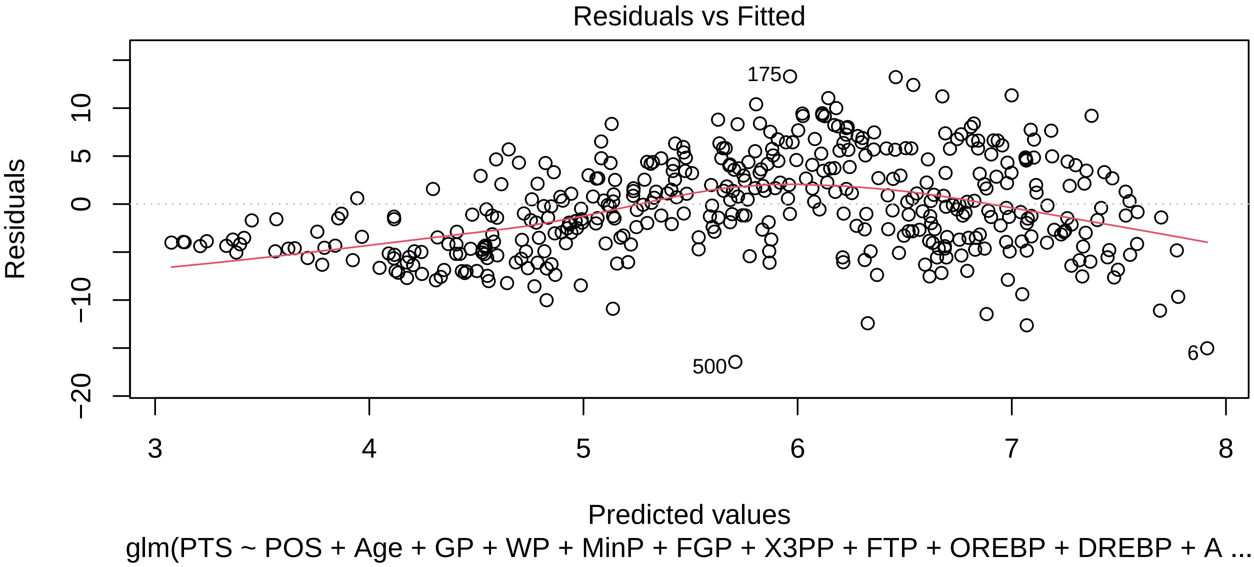
As shown in Figure 8, the stepwise selection process retained all the variables in the model. Consequently, our final model will include all predictors. The intercept is estimated at 2.57, indicating that when the position is at its baseline level and all numeric predictors are set to zero, the mean outcome is 2.57. Among the predictors, POSPG (point guard position) demonstrates the most significant impact on the model. This suggests that players in the point guard position are likely to have a higher mean score. Specifically, when a player occupies the point guard position, the mean increases by a factor of $e^{0.21082}$. Additionally, the predictor TOVP (turnovers per game) also has a substantial influence on the model. An increase in TOVP corresponds to a rise in the mean by a factor of $e^{0.23905}$. These findings highlight the importance of player position and turnovers in predicting performance outcomes.

Next, we want to check the goodness of fit of the model. We will use Goodness of Fit with H_0 : proposed model fits well and H_a : proposed model do not fit well. We will use G-statistics method for the deviance:

$$D = 2\sum_{i=1}^{n}(y_i \log(y_i/\hat{\mu_i}) - (y_i - \hat{\mu_i}))$$

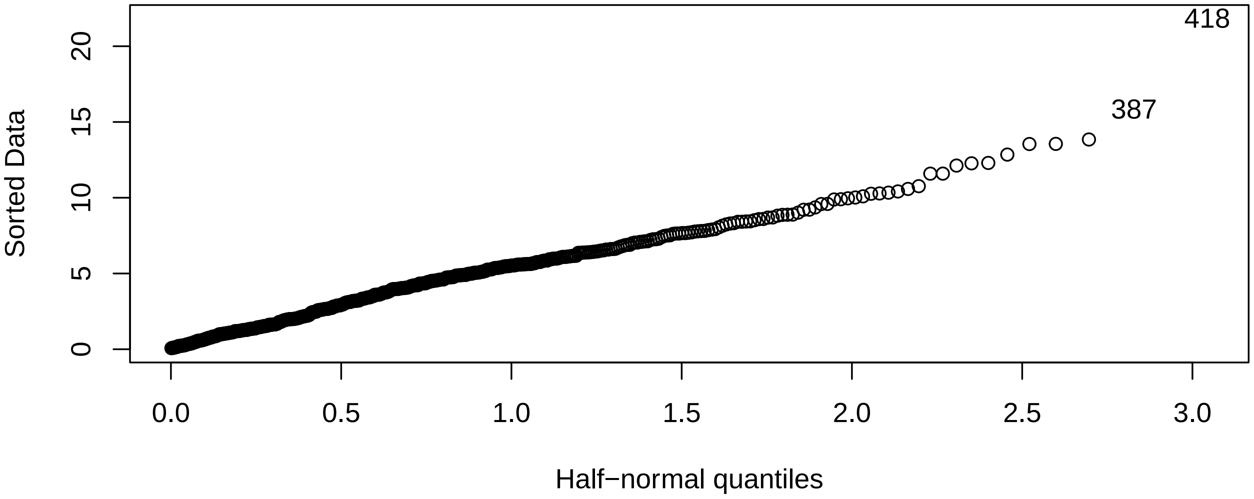
By comparing the deviance aganst χ^2 distribution with degree of freedom, we get the value 0, which means the standard Poisson model doesn't provide a good fit. Now, we want to check the residuals to see ifthe large deviance can be explained by an outlier.

Figure 9: Residuals vs. Fitted Plot



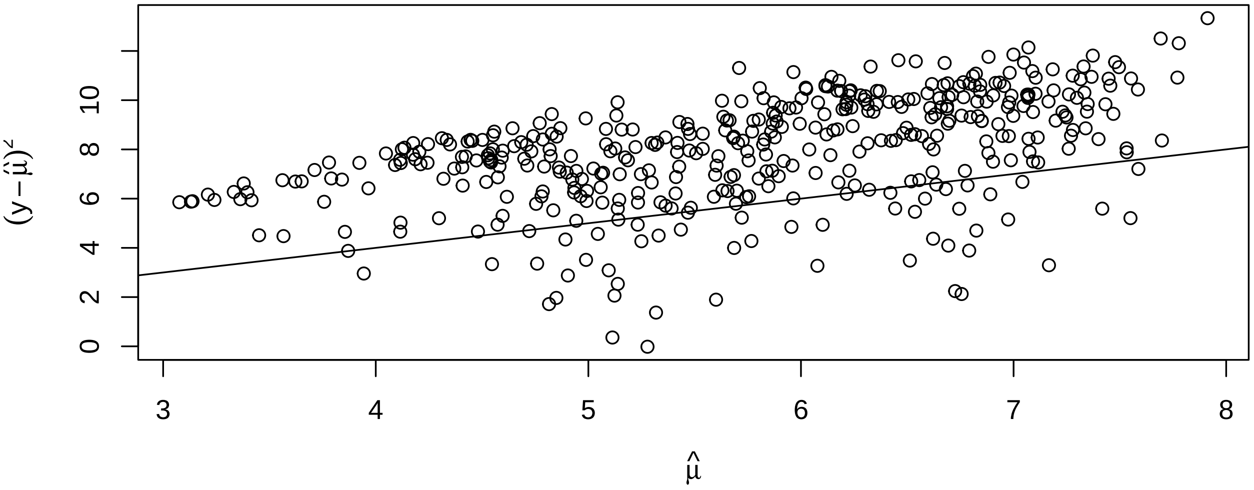
By Figure 9, we can see that the residuals tend to spread as the predicted value gets larger. This suggests that the Poisson model is overdispersed with variance larger than the mean. Before we check the overdiserpersion, we want to check the residuals to see if the large deviance can be explained by an outlier.

Figure 10: Half-Norm Plot



We can see that there are one outlier in Figure 10, but it is not significant enough to suggest that it causes the large deviance.

Figure 11: Mean and Variance Plot



As illustrated in Figure 11, the variance is greater than the mean, indicating that the quasi-Poisson model is more suitable for fitting the data. This model accounts for the dispersion parameter, making it an appropriate choice for our analysis.

Figure 12: quasi-Poisson Model Summary

```
##                     Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)       2.5700e+00  2.0482e-01 12.5479 < 2.2e-16
## POSF              1.0447e-02  6.2820e-02  0.1663 0.8680056
## POSG              2.8968e-02  8.4804e-02  0.3416 0.7328393
## POSPF             6.7185e-03  4.9247e-02  0.1364 0.8915536
## POSPG             2.1082e-01  6.9572e-02  3.0303 0.0026010
## POSSF             1.1144e-01  6.1692e-02  1.8064 0.0715979
## POSSG             1.2161e-01  6.4734e-02  1.8787 0.0610114
## Age              -1.7341e-03  3.0010e-03 -0.5779 0.5636839
## GP                1.7352e-02  1.2094e-03 14.3481 < 2.2e-16
## WP                1.8979e-01  1.3266e-01  1.4306 0.1533116
## MinP              3.7965e-02  4.0443e-03  9.3874 < 2.2e-16
## FGP               8.2657e-03  2.3236e-03  3.5572 0.0004195
## X3PP              2.1147e-03  1.2541e-03  1.6862 0.0925300
## FTP               9.5325e-03  1.3014e-03  7.3248 1.319e-12
## OREBP             1.5767e-02  2.9254e-02  0.5390 0.5902076
## DREBP             3.2892e-02  1.6192e-02  2.0313 0.0428785
## ASTP             -6.7998e-02  1.4687e-02 -4.6299 4.946e-06
## TOVP              2.3905e-01  3.2471e-02  7.3621 1.032e-12
## STLP             -7.4490e-02  4.7253e-02 -1.5764 0.1157195
## BLKP             -1.5150e-02  3.6595e-02 -0.4140 0.6790995
## PFP              -7.1027e-02  2.6513e-02 -2.6789 0.0076895
## DD2              -6.4862e-03  2.3021e-03 -2.8175 0.0050783
## TD3               3.6947e-03  8.7742e-03  0.4211 0.6739171
## FP                2.8851e-04  4.6802e-05  6.1645 1.721e-09
## Score.Difference -1.7478e-04  9.5263e-05 -1.8347 0.0672861
```

```
##
## Dispersion parameter = 26.79811
## n = 427 p = 25
## Deviance = 11879.85661 Null Deviance = 194913.94303 (Difference = 183034.08642)
```

According to Figure 12, quasi-Poisson model's coefficients are different, and many variables are no longer significant.

The quasi-Poisson model has the same exponential family as the standard Poisson model except $\phi \neq 1$. IN this case $\phi = 26.79811$.

The dispersion parameter is estimated using:

$$\hat{\phi} = \frac{\chi^2}{n-p}$$

We used the F-test (using the drop1 function) to identify which variables were not statistically significant. As a result, we decided to remove the predictors BLKP, TD3, OREBP, and Age from our model. This refinement significantly improved the model, resulting in most of the remaining variables being significant or close to significant.

Figure 13: Final quasi-Poisson Model Summary and the Confidence Interval

```
##                     Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)       2.5160e+00  1.9157e-01 13.1336 < 2.2e-16
## POSF              2.3798e-02  5.9632e-02  0.3991 0.6900434
## POSG              4.2326e-02  8.0703e-02  0.5245 0.6002388
## POSPF             1.4991e-02  4.6949e-02  0.3193 0.7496640
## POSPG             2.2042e-01  6.5514e-02  3.3645 0.0008397
## POSSF             1.2205e-01  5.6651e-02  2.1545 0.0317890
## POSSG             1.3401e-01  5.9165e-02  2.2650 0.0240395
## GP                1.7518e-02  1.1729e-03 14.9355 < 2.2e-16
## WP                1.7911e-01  1.3118e-01  1.3653 0.1729065
## MinP              3.7740e-02  3.8490e-03  9.8051 < 2.2e-16
## FGP               8.7257e-03  2.1226e-03  4.1109 4.771e-05
## X3PP              1.9825e-03  1.2024e-03  1.6487 0.0999840
## FTP               9.3470e-03  1.2716e-03  7.3508 1.095e-12
## DREBP             3.2112e-02  1.5440e-02  2.0798 0.0381745
## ASTP             -6.9319e-02  1.3149e-02 -5.2720 2.198e-07
## TOVP              2.4272e-01  3.1356e-02  7.7405 7.948e-14
## STLP             -6.8189e-02  4.5975e-02 -1.4832 0.1388013
## PFP              -6.7551e-02  2.4942e-02 -2.7083 0.0070491
## DD2              -5.5039e-03  2.0095e-03 -2.7390 0.0064341
## FP                2.8175e-04  4.3889e-05  6.4196 3.819e-10
## Score.Difference -1.8002e-04  9.4124e-05 -1.9126 0.0564979
##
## Dispersion parameter = 26.63487
## n = 427 p = 21
## Deviance = 11907.93099 Null Deviance = 194913.94303 (Difference = 183006.01204)
```

```
##                          2.5 %       97.5 %
## (Intercept)       2.1377485774  2.888657e+00
## POSF             -0.0935316822  1.402578e-01
## POSG             -0.1170877135  1.993538e-01
## POSPF            -0.0769086189  1.071435e-01
## POSPG             0.0921894039  3.490094e-01
## POSSF             0.0111324414  2.332097e-01
## POSSG             0.0182244195  2.501559e-01
## GP                0.0152263259  1.982427e-02
## WP               -0.0778930095  4.363342e-01
## MinP              0.0301908540  4.527890e-02
## FGP               0.0045639911  1.288420e-02
## X3PP             -0.0003951689  4.318125e-03
## FTP               0.0068757102  1.186007e-02
## DREBP             0.0019319989  6.245682e-02
## ASTP             -0.0951515073 -4.360936e-02
## TOVP              0.1812769023  3.041945e-01
## STLP             -0.1585038269  2.171640e-02
## PFP              -0.1164809132 -1.870754e-02
## DD2              -0.0094570906 -1.579897e-03
## FP                0.0001956866  3.677297e-04
## Score.Difference -0.0003644959  4.461462e-06
```

Thus, we have arrived at our final model, as illustrated in Figure 13. The dispersion parameter has been adjusted to $\phi = 26.63487$, reflecting the model's ability to account for overdispersion. Next, we performed a diagnostic check on the final quasi-Poisson model, revealing that its diagnostic results are relatively similar to those of the standard Poisson model. Despite this similarity, the quasi-Poisson model is better equipped to handle the increased variance in the data. As shown in the summary, the intercept has less impact compared to the standard Poisson model, with a coefficient of 2.5160. This indicates that if the position (POS) is at the baseline level and all numeric variables are set to zero, the mean outcome is 2.5160. Furthermore, POSPG (point guard) remains the most influential position affecting total scores, while TOVP (turnovers per game) continues to have the highest coefficient among the variables.

Next, we want to perform predictions using both the standard Poisson model and the quasi-Poisson model to evaluate which one provides better predictive performance. By comparing the models, we can determine which yields more accurate predictions and assess the overall efficiency of each approach. This analysis will allow us to identify the most suitable model for predicting total scores, especially considering the potential presence of overdispersion in the data.

Table 2: Models Prediction Result

Models	Deviance
Standard Poisson Model	3100.003
quasi-Poisson Model	3047.401

The results indicate that the standard Poisson model has a deviance of 3100.003, while the quasi-Poisson model achieves a lower deviance of 3047.401. This reduction in deviance suggests that the quasi-Poisson model provides a better fit to the data, making it more effective at predicting the player's total score. By accounting for overdispersion, the quasi-Poisson model captures the variability in the data more accurately, leading to improved prediction efficiency.

We use the Poisson Deviance formula to access the prediction power:

$$D = \sum 2y_i \log\left(\frac{y_i}{\hat{\mu_i}}\right) + 2(y_i - \hat{\mu_i})$$

The best model for predicting total scores is the quasi-Poisson model, as it accounts for the overdispersion in the data. Based on the diagnostic plot, it's evident that the variance in total scores increases as the scores themselves increase, highlighting the need for a model that can handle this variability effectively. The quasi-Poisson model adjusts for this by allowing the variance to be greater than the mean, offering a more robust fit for predicting total scores in the presence of overdispersion.

Conclusion

In conclusion, this study enhances our understanding of the performance metrics needed to predict a player's total score. It enables us to visualize various aspects of player performance, such as age, position, and the correlations among different variables. Notably, we identified that the total number of fantasy points has a strong correlation with total scores, along with average playing time.

By applying Poisson regression to our dataset, we discovered that the quasi-Poisson model outperforms the standard Poisson model in terms of predictive accuracy. This finding underscores the effectiveness of the quasi-Poisson model for our analysis, as it provides superior prediction results.

However, one area for improvement in this study is the method of selecting variables for the quasi-Poisson model. We observed that the quasi-Poisson model indicated more non-significant variables than the standard Poisson model. Given that the quasi-Poisson model lacks AIC or BIC metrics, we are unable to perform stepwise selection. Another area for improvement is to improve the predictive ability of the model for the player with extremely low points and extremely high points. The model do not perform well to the player with very high tota scores and very low total scores. Additionally, exploring alternative regression techniques, such as Random Forest and Support Vector Machines, could yield better prediction results.

The Poisson model is well-suited for count datasets, making it applicable to any scenario where the response variable consists of whole numbers. This model can be utilized to predict the number of insurance claims, NFL total scores, and total wins.

Ultimately, this study provides valuable insights into how performance metrics influence a player's total score, highlighting the potential of the Poisson model in predictive analytics within sports and beyond.

References

Mirzaie, A. H. (2023). NBA players stats (2023 season) [Data set]. Kaggle. https://www.kaggle.com/datasets/amirhosseinmirzaie/nba-players-stats2023-season