Shaquille O'Neal NBA Career

By:

Omkar Jadhav – G01441623

Group of 3

(Sabari Mukundth J, Shiwei Kang, Omkar Jadhav)

Project Description:

The use of statistical analysis in sports is crucial for evaluating individual player performance and identifying key factors contributing to success. These tools enable the comparison of teams and assess players' performance in specific situations, upholding the integrity of sports. This dataset focuses on Shaquille O'Neal's points per game, examining the relationships between minutes played, games played, win/loss outcomes, opponent teams, free throws attempted, assists, steals, blocks, personal fouls, and game score. The project aims to predict wins/loss, free throws, game points, and fouls by leveraging other performance indicators, aiding decision-making based on past data and reducing the likelihood of repeating negative outcomes in future games, whether virtual or physical.

Dataset Source and Information:

The dataset comprises detailed regular season statistics for each game throughout Shaquille O'Neal's NBA career. It includes columns for performance metrics like points scored (PTS), rebounds (TRB), assists (AST), steals (STL), blocks (BLK), turnovers (TOV), and personal fouls (PF). Additionally, there are variables indicating the season number, game number, date, age, team, home/away status, opponent team, win/loss outcome, and point difference in each game.

The dataset contains 1207 records with 32 variables; the variables are as follows:

Season	SeasGm	CarrGm	Date
Age	Tm	Home	Opp
Win	teamdiff	GS	Minutes
FG	FGA	FG%	3P
3PA	3P%	FT	FTA
FT%	ORB	DRB	TRB
AST	STL	BLK	TOV
PF	PTS	GmSc	Pls/Mns

Data Source:

The dataset summarizes Shaquille O'Neal's performance in each NBA regular season game. It is selected from https://sports-statistics.com/sports-data/sports-data-sets-for-data-modeling-visualization-predictions-machine-learning/.

Research Question

How are the personal fouls and the offensive rebounds affecting Shaquille O'Neal's points?

Response variable is PTS (Points)

Predictor variables are TRB (Total Rebounds) and PF(Personal Fouls)

Methodology

The methodology for this analysis relies on leveraging the capabilities of the R programming language, along with its specialized packages, to conduct a comprehensive Exploratory Data Analysis (EDA) and to craft informative visualizations. By utilizing R, a powerful statistical computing and graphics platform, we can extract meaningful insights from the dataset encompassing Shaquille O'Neal's performance in NBA regular season games. The process involves a systematic examination of the data's underlying patterns, relationships, and anomalies, facilitated by statistical summaries, correlation analyses, and graphical representations. Key R packages, such as ggplot2 for advanced data visualization and dplyr for efficient data manipulation, will be instrumental in creating insightful charts and graphs that enhance our understanding of Shaquille O'Neal's game-by-game performance trends. The R packages used in this analysis are glmnet, tidyr, caret, tidyverse, ggplot2, rpart, rpart.plot, corrplot. This methodology aims to provide a nuanced and visually intuitive exploration of the dataset, enabling

a deeper comprehension of the factors influencing Shaquille O'Neal's contributions throughout his NBA career.

Exploratory Data Analysis

Relation between Personal Fouls and the Game Scores of Shaq.

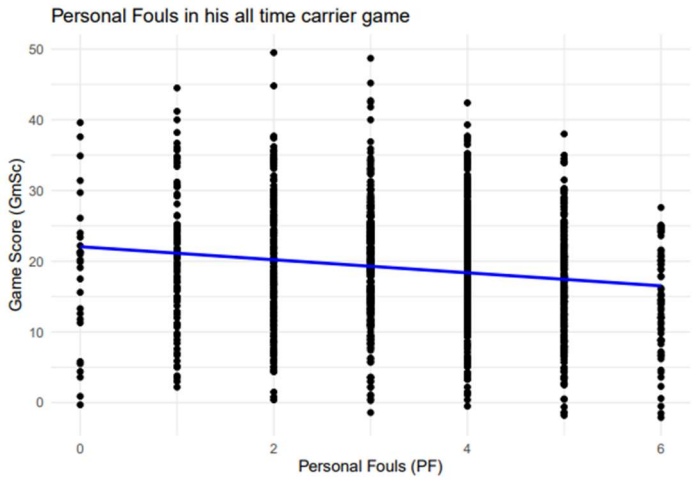

Figure1: Relation between Personal Fouls and Game Scores

The graph shows the player's career game number on the x-axis and personal fouls on the y-axis. The player's highest foul count in a game was 50, with the lowest at 0. A decreasing trend indicates improved discipline over the career, portraying the player as consistently avoiding excessive fouls.

Total Rebounds in his all-time carrier game

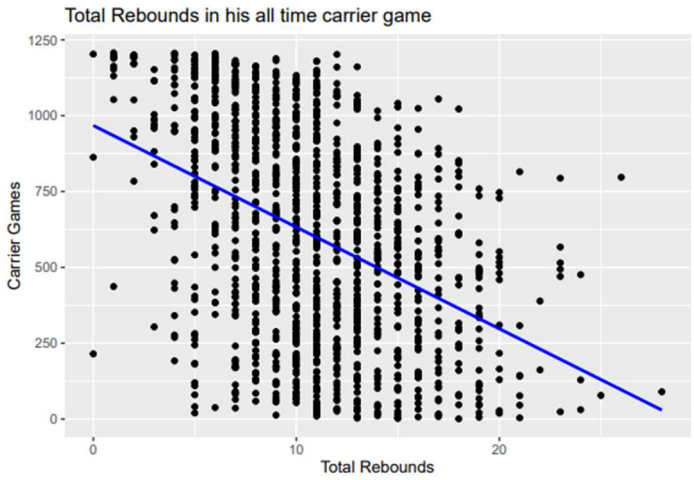

Figure2: Relation between Total Rebounds and Carrier Games

The graph displays a player's career total rebounds, with the x-axis showing rebound count and the y-axis depicting the number of games. The player's peak rebound performance was 20 in one game, with five instances of 15 rebounds and two games recording the lowest count of 0. Overall, the graph indicates a consistently balanced rebounding performance throughout the player's career.

Relation between the Total Rebounds and Personal Fouls

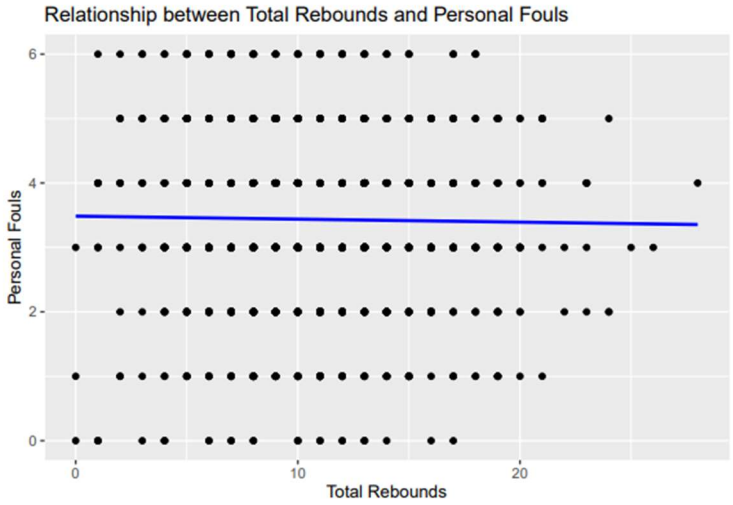

Figure3: Relation between Total Rebounds and Personal Fouls

The graph shows a weak positive correlation between a basketball player's total rebounds and personal fouls, indicating a slight tendency for increased rebounds to be associated with more fouls. However, the correlation is not strong, suggesting variability among players in terms of rebounding and fouling.

Correlation between variables which are used in research questions.

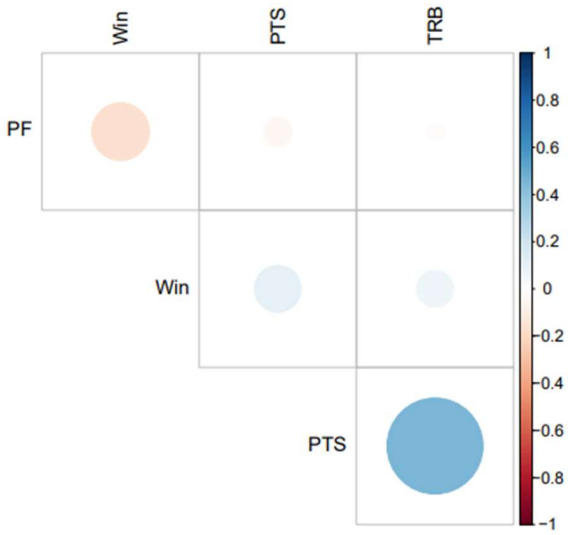

Figure4: Correlation graph!

A positive correlation exists between points scored and field goal attempts, indicating that more shots generally result in more points. However, deviations from the trendline and outliers suggest that factors beyond attempts, such as shooting accuracy, contribute to scoring variations in some games.

Creating a linear regression model for the research question.

```
> #creating a linear regression model
> model <- lm(PTS~PF + TRB, data=train_data)
> summary(model)

Call:
lm(formula = PTS ~ PF + TRB, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-22.5462  -5.4061  -0.4959   5.0569  28.1230

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.79998    0.94718  14.569   <2e-16 ***
PF          -0.08193    0.18810  -0.436    0.663
TRB          0.93386    0.06008  15.545   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.05 on 962 degrees of freedom
Multiple R-squared:  0.2008,    Adjusted R-squared:  0.1991
F-statistic: 120.8 on 2 and 962 DF,  p-value: < 2.2e-16
```

Figure5: Summary of linear regression model

The linear regression model predicts points (PTS) based on personal fouls (PF) and total rebounds (TRB). The coefficients indicate that, on average, each additional personal foul is associated with a decrease of 0.08193 points, while each additional rebound is associated with an increase of 0.93386 points. Total rebounds significantly contribute to the model, but personal fouls do not.

The model explains 20.08% of the variability in points. Overall, rebounds have a more substantial impact on points than personal fouls.

```
par(mfrow = c(2, 2))
plot(model)
```

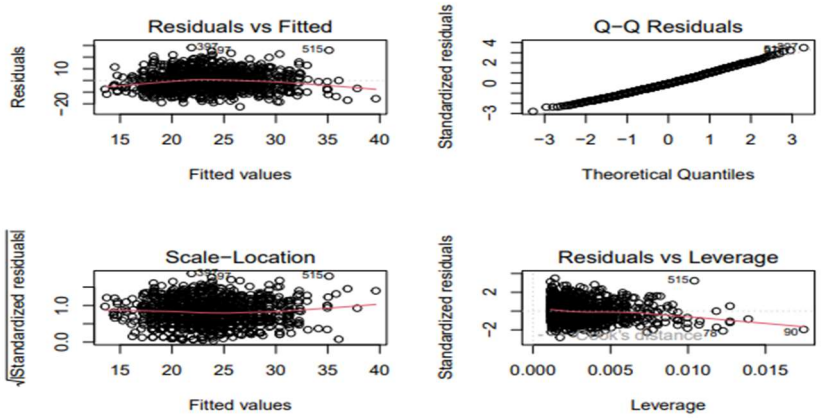

Figure6: Residuals Plot

The Residuals vs Fitted Values plot reveals some clustering around specific fitted values, indicating the linear regression model captures the relationship, though not perfectly. The Q-Q Residuals plot suggests deviations from a perfectly normal distribution of residuals. The Scale-Location plot indicates slightly heavier-tailed residuals than expected in a normal distribution. The Residuals vs Leverage plot identifies a few data points with both high leverage and large residuals, suggesting potential outliers influencing the model. Overall, while the model captures some aspects well, there are deviations and influential outliers to consider.

Logistic Regression to predict the Win/Loss using various variables.

```
> summary(logistic_model)

Call:
glm(formula = Win ~ TRB + PF + PTS + AST + BLK + STL + FT + GS,
    family = "binomial", data = train_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.932600   0.755557   1.234   0.2171
TRB         -0.008710   0.019088  -0.456   0.6482
PF          -0.268214   0.053451  -5.018 5.22e-07 ***
PTS          0.010979   0.010878   1.009   0.3128
AST          0.184265   0.043065   4.279 1.88e-05 ***
BLK          0.071579   0.042934   1.667   0.0955 .
STL          0.051520   0.085329   0.604   0.5460
FT          -0.007493   0.028460  -0.263   0.7923
GS          -0.027004   0.727472  -0.037   0.9704
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1220.5  on 964  degrees of freedom
Residual deviance: 1163.4  on 956  degrees of freedom
AIC: 1181.4

Number of Fisher Scoring iterations: 4
```

Figure7: Summary of logistic model

The logistic regression model predicts the binary outcome variable Win based on predictor variables. Variables used are offensive rebounds, personal fouls, points, assists, blocks, stolens,

full times, and game scores. Key coefficients include a significant negative effect of personal fouls (PF) on the probability of winning, while assists (AST) have a significant positive effect. Other predictors like total rebounds (TRB) and points scored (PTS) show non-significant effects. The model's fit is assessed through deviance values, with a lower AIC indicating a better balance between fit and simplicity.

Confusion matrix and Accuracy

```
> confusion_matrix <- table(test_data$Win, predicted_classes)
> print(confusion_matrix)
   predicted_classes
    Loss Win
  0   16  56
  1   11 159
> # Calculate accuracy
> accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
> print(paste("Accuracy: ", accuracy))
[1] "Accuracy:  0.723140495867769"
```

Figure8: Confusion matrix and Accuracy

The model correctly predicted 159 Wins and 56 Losses. The model incorrectly predicted 16 Wins as Losses and 11 Losses as Wins. The accuracy of the model is 72.3%, which is calculated as the number of correct predictions divided by the total number of predictions.

Regression Tree for the research question

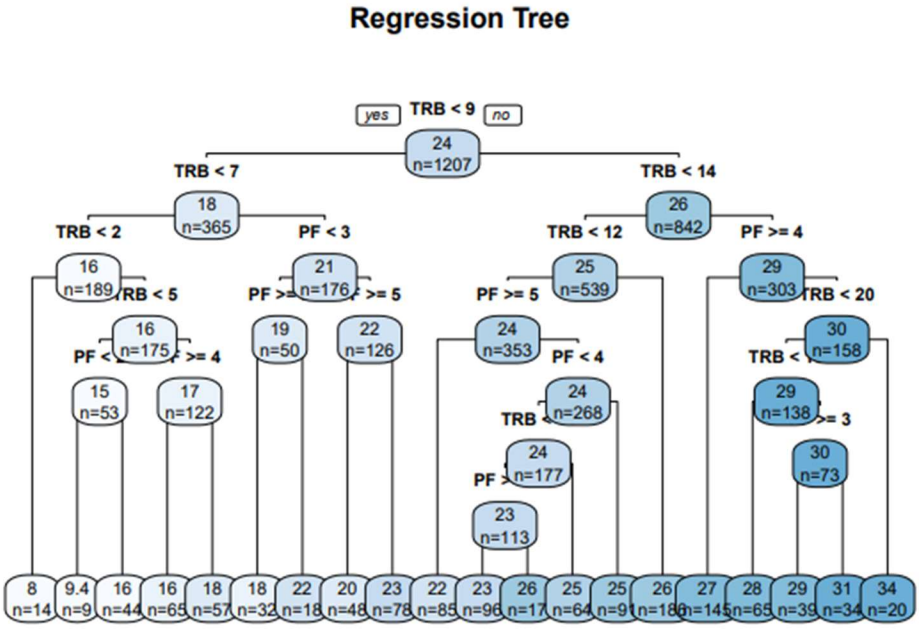

Figure9: Regression Tree

Regression Tree was used to model the relationship between response variable and predictor variables. Here, the response variable is PTS and predictor variables are TRB and PF. The regression tree suggests that there is a positive correlation between personal fouls and total rebounds, but it is not a perfect relationship. Other factors likely play a role in determining a player's rebounding performance.

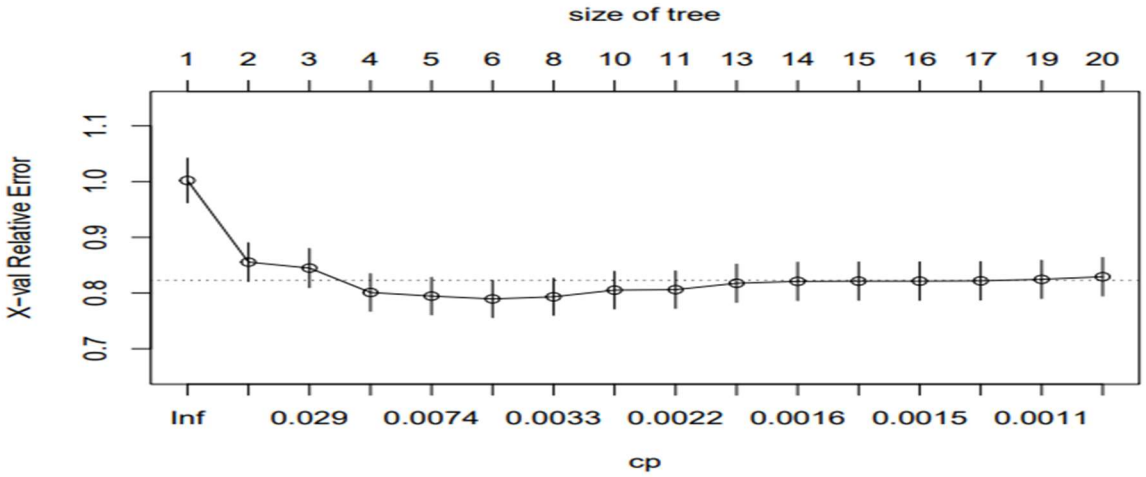

Figure10: Complexity Parameter plot

The complexity parameter plot illustrates the relationship between a decision tree's size and cross-validated relative error. As the complexity parameter (cp) decreases, indicating larger trees, there is an initial substantial drop in error. However, beyond a certain point, further reduction in cp offers minimal error reduction. The plot suggests that an optimal balance occurs, stabilizing relative error around 0.8. The secondary x-axis indicates the increasing size of the tree as cp decreases. Overall, the plot helps identify the optimal cp value for an effective decision tree model.

Summary

In this comprehensive analysis of Shaquille O'Neal's NBA career, statistical tools and R programming are harnessed to conduct Exploratory Data Analysis (EDA), craft informative visualizations, and build predictive models. The research question centers on the influence of personal fouls and offensive rebounds on Shaq's points per game. Exploring trends, it is observed that Shaq's personal fouls show a decreasing trend, indicating improved discipline, while total rebounds exhibit consistent performance. A weak positive correlation between rebounds and fouls suggests a nuanced relationship among players. Correlation analysis identifies significant associations, and a linear regression model predicts points, emphasizing the substantial impact of rebounds. Logistic regression models the probability of wins/losses, highlighting the negative effect of personal fouls and the positive impact of assists. The regression tree explores the

relationship between points, total rebounds, and personal fouls, acknowledging additional contributing factors. The complexity parameter plot aids decision tree model optimization. In summary, this project offers a robust statistical exploration of Shaquille O'Neal's performance, unraveling nuanced interactions between personal fouls, rebounds, and game outcomes. The findings provide valuable insights for informed decision-making in team strategies and player development.

References

1) Glmnet, Friedman J, Tibshirani R, Hastie T (2010). “Regularization Paths for Generalized Linear Models via Coordinate Descent.” _Journal of Statistical Software_, *33*(1), 1-22. doi:10.18637/jss.v033.i01 <https://doi.org/10.18637/jss.v033.i01>.

2) Tidyr, Wickham H, Vaughan D, Girlich M (2023). _tidyr: Tidy Messy Data_. R package ver sion 1.3.0, <https://CRAN.R-project.org/package=tidyr>.

3) Caret, Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

4) Tidyverse, Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemu nd G, Hayes A, Henry L,Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Oo ms J, Robinson D, Seidel DP,Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.”_Journal of Open Source Software_, *4*(43), 1686. d oi:10.21105/joss.01686<https://doi.org/10.21105/joss.01686>.

5) Ggplot2, H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016

6) Rpart, Therneau T, Atkinson B (2023). _rpart: Recursive Partitioning and Regression Tree s_. R package version 4.1.21, <https://CRAN.R-project.org/package=rpart>.

7) Rpart.plot Milborrow S (2022). _rpart.plot: Plot 'rpart' Models: An Enhanced Version of ' plot.rpart'_. R package version 3.1.1, <https://CRAN.R-project.org/package=rpart.plot>.

8) Corrplot, Taiyun Wei and Viliam Simko (2021). R package 'corrplot': Visualization of a Co rrelation Matrix (Version 0.92). Available from https://github.com/taiyun/corrplot

9) https://sports-statistics.com/sports-data/sports-data-sets-for-data-modeling-visualizati on-predictions-machine-learning/.