Key NBA Statistics Influencing Minutes Played Per Game: 2023-2024 Regular Season Analysis

Wardah Ali, Navyasri Chinthapatla, Danae McCulloch, Safeen Mridha, Ayda Takehei

DATA 603: **Statistical Modelling with Data**

Mina Aminghafari

FALL 2024

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

1.1 MOTIVATION

1.1.1. Context

Basketball is a sport that has experienced exponential growth worldwide over the last decade. Such growth can be attributed to the increasing popularity of the National Basketball Association (NBA), which is the highest level of professional basketball in the world. This increased traction enhanced the marketability of the sport; with the media and broadcasting-rights deals being an estimated $76 billion, leading players being offered historic contracts that have never been seen before. For example, Jayson Tatum, the star player of the reigning NBA champions, the Boston Celtics, recently received a 5-year contract that will pay him a total of $315 million for that duration to stay with the team, making him the highest paid player in NBA history. This prompted our group to investigate a series of questions. What exactly makes him (and other NBA players) a great player? What can NBA players do on the court that earns the coach's trust to play more than their teammates? What skills can NBA players improve in order to obtain more playing time?

1.1.2 Problem

In basketball, games are 48 total minutes long and each team is allowed to have 5 players on the court at a time. A key assumption that holds true is the fact that not all players are alike, meaning every athlete has different skillsets with lots of variability in terms of what they are skilled at. For this reason, players who are typically more skilled will spend more time on the court as their enhanced skillset gives their team a better chance of winning. The problem we are attempting to address is determining which skillset and counting statistics influence how much playing time a player gets every game. To solve this, we obtained data for the 2023-2024 season from the NBA website that lists the performance statistics for 571 players. From this dataset, we will attempt to build a multiple regression model that takes into consideration the major statistical factors related to basketball, which will help us determine what influences how much time a player gets to play during games.

1.1.3 Challenges

There are a few challenges associated with this project. Although it is reasonable to assume that players who are more skilled and make greater contributions to winning are likely to receive more playing time, the bias of the coach can also play an effect on this. For example, if a coach has a certain preference towards a player, they might be more inclined to give that player more playing time than someone else of equal or even superior skill. Although instances like these are rare, unfortunately no method exists to account for this potential bias; the best we can do is acknowledge this as a potential limitation and move forward with our study with the assumption that playing time partitioning is impartial.

1.2 OBJECTIVES

1.2.1 Overview

The intent of this project is to analyze player statistics from the 2023-2024 NBA season to identify key factors influencing the number of minutes a player spends on the court per game. The playing time is a critical measure of a player's impact and role in a team performance, therefore can greatly influence their career trajectory, salary and overall contribution to their team's success in the season. The identified key factors can then be used as areas of improvement for a player to increase their total minutes played per game.

1.2.2 Goals & Research Questions

The primary research questions of this project are summarized below:

1. What are the most important predictors in response to the Total Minutes played per Game in the 2023 NBA season?
2. What can an NBA player work on to increase their Total Minutes Played per Game?

The primary goal of this project is to perform multiple regression analysis to build a predictive model that identifies the key factors influencing the total minutes a player plays per game. The dependent/response variable in the model is minutes played (MIN), and the independent variables consist of various quantitative performance metrics. Various visuals will be used during the analysis to aid us during this process.

2. METHODOLOGY

2.1 Data

The dataset used in this project was sourced from the official NBA statistics platform website for the 2023-2024 season. It contains a total of 66 columns where 24 columns were used as predictor variables, and one was used as a dependent variable. The explanatory variable is the number of minutes a player plays; MIN. Below are the 24 predictor variables that will be used in the project:

1. MIN: minutes played by a player per game (Response Variable).
2. FGA: total field goals attempted (baskets/shots attempted) by a player per game.
3. FGM: total field goals made (baskets/shots scored) by a player per game.
4. FG_PCT: the average shot-making efficiency of a player per game (calculated by dividing the number of field goals made by the number of field goals attempted)
5. FG3A: 3-pointers (long-range baskets/shots) attempted by a player per game.
6. FG3M: 3-pointer (long-range baskets/shots) made by a player per game.

7. FG3_PCT: the average 3-point shot-making efficiency of a player per game (calculated by dividing the number of 3-point field goals made by the number of 3-point field goals attempted).
8. FTA: free throws (foul shots) attempted by a player per game.
9. FTM: free throws (foul shots) made by a player per game.
10. FT_PCT: the average free throw-making efficiency of a player per game (calculated by dividing the number of free throws (foul shots) made by the number of free throws (foul shots) attempted).
11. OREB: offensive rebounds (grabbing the ball after a missed shot while your team is still on offense) grabbed by a player per game.
12. DREB: offensive rebounds (grabbing the ball after a missed shot while your team is still on offense) grabbed by a player per game.
13. REB: total rebounds (grabbing the ball after a missed shot while your team is still on offense) grabbed by a player per game.
14. PTS: The number of points scored by a player per game.
15. AST: assists (passing the ball to a teammate who subsequently scores) accumulated by a player per game.
16. TOV: turnovers (giving the ball away to the other team while on offense) accumulated by a player per game.
17. PFD: personal fouls drawn (making a player on the other team foul you) by a player per game.
18. PF: personal fouls accumulated (when a player fouls another player on a team) by one player per game.
19. GP: games played by a player throughout the entire season.
20. DD2: the number of double-doubles accumulated by a player throughout the season (having double digits in two counting statistics in a single game; i.e. 10 points and 10 assists; 26 points and 17 rebounds; 13 rebounds and 11 blocks, etc.).
21. PLUS_MINUS: a player's net impact on the game while on the court per game (calculated by adding the amount of points the player of interests' team scores while he is on the court, subtracted by the number of points the opposing team scores while he is on the court).
22. BLK: the number of blocks accumulated by a player per game (a block is when a player swats away/ "blocks" an opposing player's shot).
23. BLKA: the total number of blocks attempted by a player per game (a block is when a player attempts to swat away or "block" an opposing player's shot).
24. TD3: the number of triple-doubles accumulated by a player throughout the season (having double digits in three counting statistics in a single game; i.e. 10 points, 12 assists, 15 rebounds, 36 points, 12 rebounds, 10 steals, 17 assists etc.).
25. STL: the number of steals accumulated by a player per game (a steal is when a player swipes/intercepts the ball from an opposing player)

The remaining 41 variables/columns that were excluded are listed below, along with the rationale for their removal:

Variables to remove:

- PLAYER_ID: for the purposes of regression modelling, the player's ID (a unique numeric code given to ever player) for the corresponding statistic is not relevant
- PLAYER_NAME: like PLAYER_ID, the name of a player is not relevant towards regression modelling. A player's name does not dictate how many minutes he plays every game- his overall skillset does.
- TEAM_ID: Every team has its own unique code. However, the team that a player plays for will not influence how many minutes they play a game. The team a player plays for is not reflective of their skillset.
- TEAM_ABBREVIATION: This column also serves the same purpose as TEAM_ID by giving an alternative method of examining which team a player plays for. Once again, we see that the team a player plays for does not reflect on their skillset.
- AGE: Given that the scope of our project is to analyze which counting statistic affects the minutes an NBA player plays in every game, AGE was removed, as this is not a counting statistic that a player has control over. Therefore, it is not related to the objectives of this project.
- W: This column represents how many wins (W) the player accumulated as a member of their respective team. In basketball, rosters consist of 15 players; out of these 15 players, it is typical for only the top 7-11 players to actually play in the games, meaning that 4-8 players may not even play any given night. Therefore, the number of wins a player accumulates reveals very little about their individual contributions, and therefore, their minutes played.
- L: This column represents how many losses (L) the player accrued as a member of their respective team. Basketball rosters have many players, and it is physically impossible for everyone to play- so it is typical for only the "best" players to get substantial minutes per game, which is determined through individual counting statistics (such as AST, REB, etc.). Furthermore, a player who is skilled (and is therefore given a lot of playing time every game) could be stuck on a bad team that loses a lot. Therefore, the number of losses a player accumulates is irrelevant to the scope of this study.
- W_PCT: This column looks at win percentage (W_PCT). It is determined by dividing the total amount of wins a player accrues by the total number of games they played. Given the fact that the Win column is irrelevant to our study and disregarded, this column (which takes wins into account) should also be removed.
- NBA_FANTASY_PTS: this column calculates the amount of fantasy points that a player accumulates. Fantasy basketball is a game where participants act as general managers of their own fictional team, and they attempt to "draft"/populate their team with real NBA

players. The success of their team is determined by the cumulative "fantasy points" that the players on their team obtain every week. These "fantasy points" are calculated using various metrics and is not a direct counting statistic that these players can obtain during game. Since we're only counting direct counting statistics, this variable is irrelevant withing the scope of our study and was removed.

- WNBA_FANTASY_PTS: this variable is like NBA_FANTASY_PTS but pertains to the WNBA (Women's National Basketball Association) instead. This means that this variable is even further beyond the scope of this study.

The rest of the variables consist of variables that have already been explained and were included in our dataset. The only difference is that the following variables are all measures of RANK. The value, rank, is a discrete variable that will take a counting statistic of interest (i.e. REB) and rank the player with this statistic in comparison to the rest of the players in the NBA. All rank variables were removed for the following reasons:

- Redundancy: the root variable (i.e. the portion of the variable preceding "_RANK") is already included in our study. Therefore, looking at another metric that looks at the same exact counting statistic, while providing no additional insight, is unnecessary. This helps avoid overfitting issues, as no information is lost with the removal of these variables.
- Insignificance and Inaccuracies: ranking NBA players based on specific counting statistics only serves EDA and visualization purposes. Assessing how much playing time they receive based on their "rank" for a statistic compared to the rest of the NBA is not going be useful, as opposed to using the statistic itself. Furthermore, the RANK is not a direct counting statistic- it is not something a player can immediately obtain while playing in game. Therefore, it is beyond the scope of our project.

For these reasons, we have decided to take out the following columns below:

1. GP_RANK
2. W_RANK
3. L_RANK
4. W_PCT_RANK
5. MIN_RANK
6. FGM_RANK
7. FGA_RANK
8. FG_PCT_RANK
9. FG3_PCT_RANK
10. FG3M_RANK
11. FG3A_RANK
12. FTM_RANK
13. FTA_RANK

14. FT_PCT_RANK
15. OREB_RANK
16. DREB_RANK
17. REB_RANK
18. AST_RANK
19. TOV_RANK
20. STL_RANK
21. BLK_RANK
22. BLKA_RANK
23. PF_RANK
24. PFD_RANK
25. PTS_RANK
26. PLUS_MINUS_RANK
27. NBA_FANTASY_PTS_RANK
28. DD2_RANK
29. TD3_RANK
30. WNBA_FANTASY_PTS_RANK

Therefore, the dataset for this project looks like the following:

	GP	MIN	FGM	FGA	FG_PCT	FG3M	FG3A	FTM	FTA	OREB	DREB	REB	AST	TOV	BLK	PF...16	PFD	PTS	PLUS_MINUS	DD2	TD3	STL	BLKA	PF...24	FT_PCT	FG3_PCT
1	42	7.4	1.3	2.9	0.446	0.3	1.2	0.4	0.5	0.3	0.9	1.2	0.5	0.3	0.1	0.5	0.3	3.2	0.4	0	0	0.2	0.2	0.5	0.652	0.26
2	56	11	1.5	3.5	0.423	1.2	3	0.3	0.3	0.2	1	1.1	0.5	0.2	0.1	0.9	0.4	4.5	0.9	0	0	0.2	0.1	0.9	0.895	0.408
3	20	8.5	0.9	3.1	0.29	0.5	2	0.1	0.1	0.1	0.8	0.9	0.3	0.4	0.1	0.3	0.1	2.4	-2.6	0	0	0.1	0.2	0.3	1	0.256
4	73	31.5	5.5	9.8	0.556	0.5	1.9	2.4	3.7	2.4	4.1	6.5	3.5	1.4	0.6	1.9	3.1	13.9	5.8	12	0	0.8	0.8	1.9	0.658	0.29
5	78	16.3	2.4	5.3	0.446	1.1	2.8	0.7	0.8	0.3	1.3	1.6	1.8	0.7	0.1	1.6	0.8	6.6	1.1	0	0	0.5	0.3	1.6	0.921	0.387
6	72	27.7	4.4	8.8	0.496	1.9	4.6	1.5	1.9	0.9	2.9	3.8	1.5	0.9	0.7	3.3	2	12.2	2.2	1	0	0.9	0.7	3.3	0.781	0.419

Table 1: NBA Dataset Head

2.2 Approach

The approach we will use involves performing multiple linear regression to determine the best regression model. This model will help us determine the most influential key factors in predicting the minutes a player plays during a game. The process involves the following steps: building the initial full additive model containing all the variables of interest, exploring first order, interaction and higher order models. Once the best model is selected after completing all the steps, it will then be assessed against the six key assumptions of multiple regression: Linearity, Independence, Normality, Equal Variance, Multicollinearity and Outliers. If any assumptions are violated, appropriate measures will be taken to try and further improve the model such as Box Cox transformations and removal of outliers. After addressing these measures, the final best model will be chosen and used for analysis.

We think this approach will work well because it follows a thorough and iterative process for building and refining the regression model using appropriate statistical methods. With our domain knowledge, we are confident in achieving reliable results.

2.3 Workflow

What steps (workflow task list) are required? Which of these steps is particularly hard? What to do if the hard steps don't work out

Once obtaining our dataset consisting of player statistics from the 2023-2024 season, testing for collinearity will take place first. To do this, a VIF test on the entire dataset will be completed, followed by making a pairwise plot consisting of variables that have critical detection values. By coupling both methods together, any instances of collinearity between variables will be identified; we intend to rely on background domain knowledge to choose the best single variable from these groupings of correlated factors to create a base model. This model will once again be subject to the VIF test to ensure that all instances of collinearity are dealt with. Next, the stepwise approach will be used to create three reduced models: a reduced additive model, an interaction model, and a potential polynomial model as well. In particular, a ggpairs plot will be employed to help identify any potential instances of polynomial relationships within variables. To compare the effectiveness between these models, an F-test will be completed using an ANOVA table. In addition to the ANOVA table, the adjusted R-square will also be used as a baseline to evaluate the models as well. Once the best model is selected, we will complete regression assumption testing. For testing linearity, the residuals will be plotted on a graph to look for patterns (patterns would indicate the assumption fails). If any pattern is present, it would suggest a few things: a potential correlation between variables could be present; there could be interactions between variables; one of the variables must be put to a higher degree. We will assess each circumstance as we go. To test the assumption of independence, we intend on plotting the regression residuals to see if more patterns, markedly clumping occur (which, again, would indicate the assumption does not hold true). A violation of this assumption typically means that the errors are correlated. Next, the assumption of equal variance will be conducted using a Breusch-Pagan test to reveal whether our dataset exhibits homoscedasticity. Next, the assumption of normality will be assessed using a Shapiro-Wilks test; to combat instances where normality is not met, a box-cox transformation will be used on the data to potentially normalize it. Regardless, considering that our sample size is large, the results from the Shapiro-Wilks test will be taken with a grain of salt, as the central limit theorem states that a sample with a large size will naturally move towards a normal distribution. Finally, we will search for the presence of outliers (and leverage points) by calculating Cook’s distance, as well as using a residuals vs leverage plot. Any value of Cook’s distance greater than 0.5 will be classified as an outlier. Once the assumptions are tested, we will train our model and use it to make predictions to assess its effectiveness and accuracy.

2.4 Contributions

Safeen, with his extensive knowledge and passion for basketball, took the lead in determining which variables to remove from the dataset. We all chose to meet up on multiple occasions to work on the R-file to create the model; Danae and Wardah worked on the file interchangeably. While the model was being built, Safeen, Ayda and Navya worked on the report together. The entire group worked on interpreting the results and completing the finishing touches.

3. MAIN RESULTS OF THE ANALYSIS

3.1 First Order Model

3.1.1 Multicollinearity Testing

To construct a first order model, we will begin by testing multicollinearity to determine if a correlation exists between two or more independent variables. If a strong correlation is found between two variables, it indicates the presence of multicollinearity. Multicollinearity reduces the precision of the estimated coefficients thereby weakening the statistical power of the regression model. Multicollinearity will be tested by computing the VIF ("Variance Inflation Factor") for each independent variable to determine which variables should be retained in our model. A VIF value of one indicates that there is no collinearity, a value between 1-5 suggests moderate collinearity but not enough to take corrective measures, and any value above 5-10 indicates critical levels of multicollinearity where the coefficients are poorly estimated.

After conducting the VIF value, based on Table 2, FGM, FGA, FG3M, FG3A, FTM, FTA, OREB, DREB, REB, PFD and PTS has multicollinearity for which we plot a pairwise graph to determine which variables to keep in the model. Based on Figure 1; we decided to keep FG3M, REB and FTM and discard the rest.

Variables	**VIF**	**Detection**
GP	2.2758	0
FGM	4504.1157	1
FGA	198.6233	1
FG_PCT	2.4223	0
FG3M	209.5221	1
FG3A	83.0241	1
FTM	419.2257	1
FTA	110.8351	1
OREB	235.4159	1
DREB	1423.7317	1
REB	2495.5506	1

AST	5.4586	0
TOV	9.0757	0
BLK	2.3220	0
PF	3.7188	0
PFD	33.4965	1
PTS	8544.5076	1
PLUS_MINUS	1.4922	0
DD2	4.7619	0
TD3	1.8469	0
STL	2.8005	0
BLKA	4.3051	0
FT_PCT	1.6140	0
FG3_PCT	1.6962	0

Table 2: VIF Multicollinearity Diagnostic Results (3.1.1)

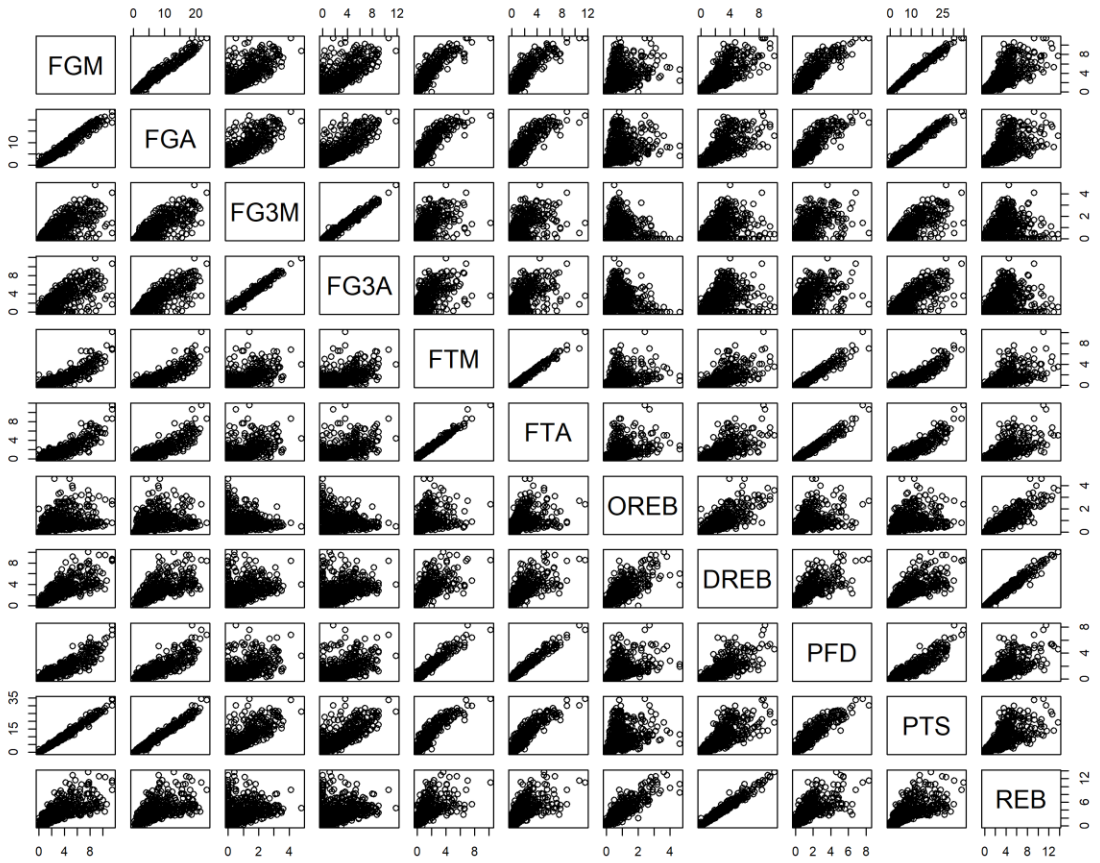

Figure 1: Pairwise plot

The final model after multicollinearity is:

MIN=GP+FG_PCT+AST+TOV+BLK+PF+PLUS_MINUS+DD2+TD3+STL+BLK+BLKA+FT_PCT+FG3_PCT+FG3M+FTM+REB

To confirm the final model above, the multicollinearity was checked again and the VIF values were all zero.

3.1.2 Stepwise & Anova

Stepwise selection is done by iteratively adding and removing variables from the model determining the significance level, which provides p-values for the predictor variables to decide whether a factor is statistically significant or not. This process will help identify that subset of variables contributing the most towards the Minutes played. It further refines the model in selecting the most relevant predictor variables.

Both Forward and Backward selection is applied here using “ols_step_both_p” function which includes both adding (if p-value is less than the given threshold) and removing variables (if p-

value is greater than given threshold). A significance level of 0.05 has been set by which model decides which variables are statistically significant.

Model variables after stepwise selection is:

MIN = GP + PF + FG3M + REB + AST + DD2 + STL + PLUS_MINUS + BLKA + TD3 +

A Full model F-test using Anova is performed to compare the base model and stepwise model. Hypothesis is as follows:

- H_0: The reduced model is a better fit for the data compared to the full model.
- H_a: The full model is a better fit for the data compared to the reduced model.

Using the p-value obtained from the ANOVA test, which is 0.2294, we fail to reject the null hypothesis. Therefore, the stepwise model will be used further as it is considered as reduced model.

The final model after stepwise and Anova is as follows:

$\widehat{MIN}$ = 0.764789 + 0.051002X_{GP} + 1.551311X_{PF} + 3.502251X_{FG3M} + 1.443531X_{REB} + 1.009194X_{AST} - 0.078061X_{DD2} + 3.549834X_{STL} - 0.159623X_{PLUS_MINUS} - 0.247211X_{TD3} + 3.059249X_{BLKA}

Source	**DF**	Sum of Suares (SS)	**Mean Square (MS)**	**F- Statistic**	**P-Value**
Regression	6	52.358	8.7263	1.3584	0.2294
Residual Error	555	3565.4	6.4241		
Total	561	3617.7	6.4487		

Table 3: Analysis of Variance Table for Stepwise and Multicollinearity model (3.2.1)

3.1.3 Interaction

The next approach would be choosing a model that best predicts the factors that influence Minutes Played based on NBA statistics. Interaction model helps in determining if the effect of one predictor variable depends on another predictor variable. The interaction model included all possible two-way interactions between the predictor variables.

Next step was to remove insignificant interactions from the model. Stepwise is the best method to identify significant interactions. The suggested interactive model resulted an Adjusted R-squared of 0.9409 and a p-value of < 2.2e-16 was concluded to be:

MIN = DD2 + REB + FG3M + AST + GP:PF + BLKA + STL + STL*AST + GP + REB*PLUS_MINUS + FG3M*AST + REB*AST + GP*PLUS_MINUS

The final step was to perform the Anova test to ensure the interactive is a better fit compared to the model derived in 3.1.2. The hypothesis remained the same:

- H_0: The reduced model is a better fit for the data compared to the full model.
- H_a: The full model is a better fit for the data compared to the reduced model.

The Anova output a p-value of < 1.888e-12 meaning that the null hypothesis is rejected. Therefore, the final model is:

$$MIN = 0.795761 + 0.031422\ X_{GP} + 1.211096X_{PF} + 4.520723X_{FG3M} + 1.459118X_{REB} + 2.326227\ X_{AST} - 0.024600X_{DD2} + 4.164537\ X_{STL\_} 0.240726X_{PLUS\_MINUS} + 3.269937X_{BLKA} - 0.336148X_{FG3M*AST} - 0.460442X_{AST*STL} - 0.101063X_{REB*AST} + 0.004189X_{PF*GP} - 0.026631X_{REB*PLUS\_MINUS} + 0.004796X_{GP*PLUS\_MINUS}$$

Source	DF	Sum of Squares (SS)	Mean Squares (MS)	F-Statistic	P-Value
Regression	5	391.83	78.366	13.507	1.888e-12
Residual Error	556	3225.9	5.801		
Total	561	3617.7	6.448		

Table 4: Analysis of Variance Table for final Interactive and Stepwise model (3.2.1)

3.1.4 Polynomial

Checking for polynomials is crucial to identify any non-linear relationships between response and predictor variables. It helps in improving our model's fitness and accuracy while ensuring we do not miss any important relationships. From Figure 2, slight curvature is observed between REB and DD2 variables suggesting there might be any non-linear relationship between them. Therefore, a second-degree and third-degree polynomial term was attempted for both DD2 and REB our model. However, adding degrees for both variables did not improve the Adjusted R-square value of 0.9411. As there is no change in the value of the Adjusted R-square, we decided to discard the polynomial model and keep our model simple to avoid complexity and overfitting.

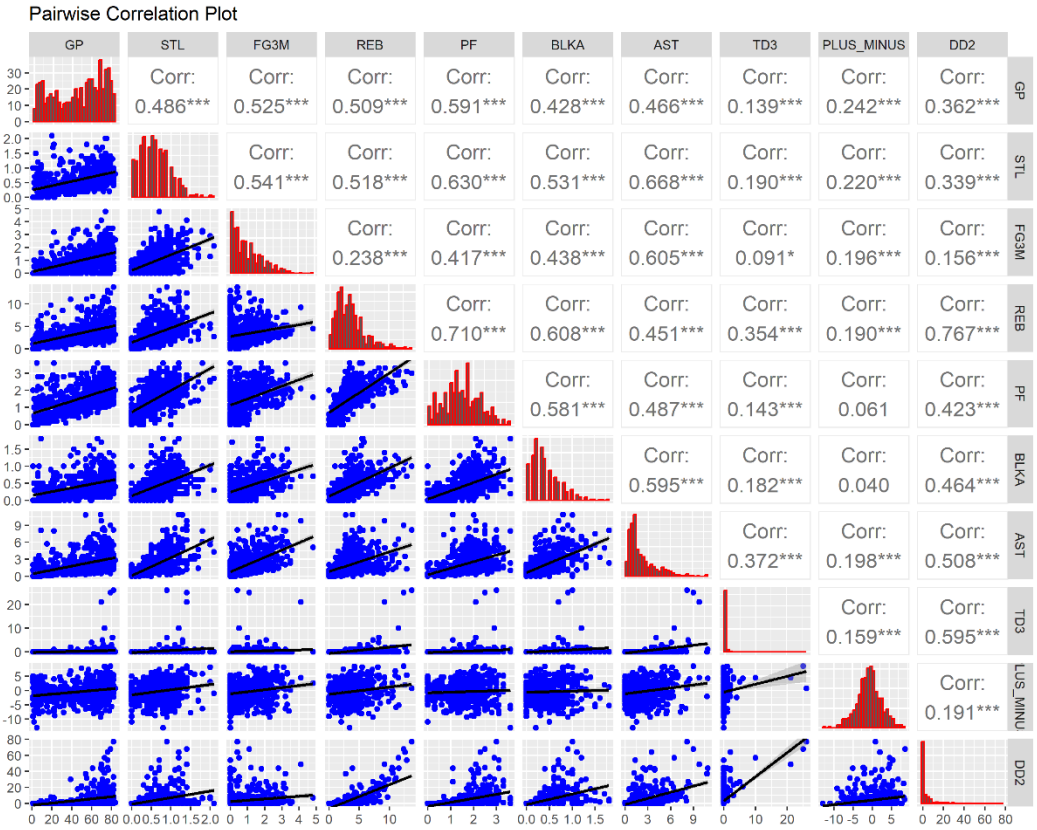

Figure 2: Correlation plot to check for polynomials

3.1.5 Final First Order Model

All Multicollinearity, Stepwise, and Polynomial methods produced a model with removing some unnecessary terms, including interaction terms, and excluded the insignificant interactions. Therefore, Final first-order model is:

$\widehat{MIN}$MIN = 0.795761 + 0.031422 X_{GP} + 1.211096X_{PF} + 4.520723X_{FG3M} + 1.459118X_{REB} + 2.326227 X_{AST} - 0.024600X_{DD2} + 4.164537 X_{STL} – 0.240726X_{PLUS_MINUS} + 3.269937X_{BLKA} - 0.336148$X_{FG3M*AST}$ - 0.460442$X_{AST*STL}$ - 0.101063$X_{REB*AST}$ + 0.004189X_{PF*GP} - 0.026631$X_{REB*PLUS_MINUS}$ + 0.004796$X_{GP*PLUS_MINUS}$

3.2 Multiple Linear Regression Assumptions

3.2.1 Linearity

The linearity assumption in multiple regression states that there is a linear relationship between the independent and dependent variables in a model. This means that the change in the dependent variable is directly proportional to the changes in the independent variables, while holding the other predictor variables constant. To test for linearity, a residual plot was created as shown in Figure 3. The plot shows no discernible pattern and indicates a non-linear relationship; thus, the linearity assumption has been violated.

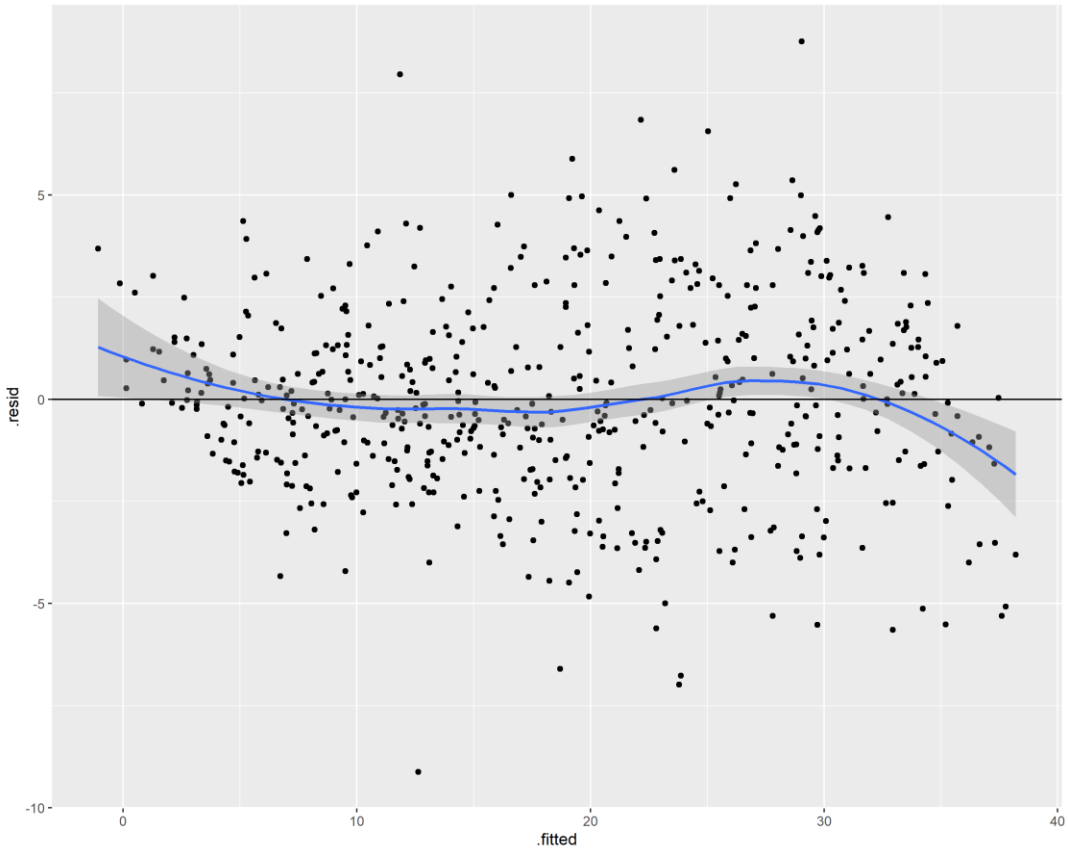

Figure 3: Residual vs fitted plot

3.2.2 Independence

The independence assumption assumes that the residuals (error) for one observation does not influence the error for another observation; error terms are mutually independent and uncorrelated. This assumption is violated when successive errors are correlated, leading to unreliable and inaccurate confidence intervals and p-values. This can be tested by plotting a residual plot and seeing if a pattern, trends, clumping or a correlation can be seen, any of which will suggest a failure of independence. From the residual plot in Figure 4 we do not detect any trends/patterns and see an equal scatter of points; thus, the independence assumption has passed.

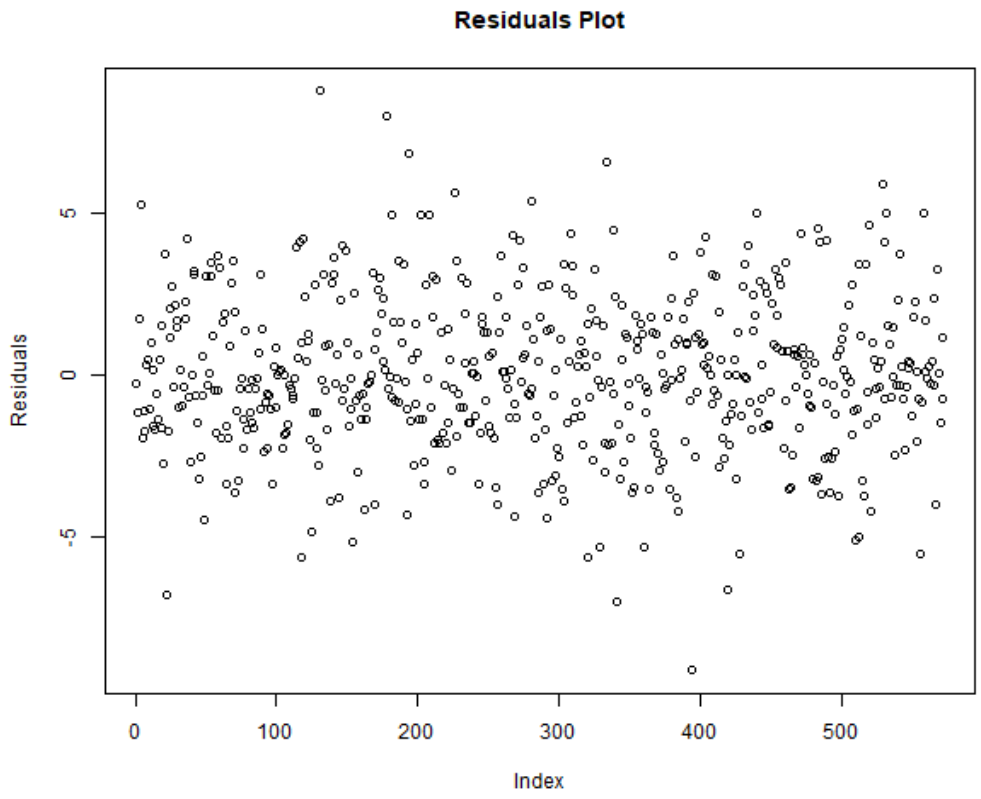

Figure 4: Residual plot

3.2.3 Equal Variance

The next assumptions being tested are equal variance (homoscedasticity) and normality. To evaluate whether the error terms have constant variance, a Residual vs. Fitted plot was created as shown in Figure 5. The plot indicates no clear funneling, as the magnitude of the residuals remains fairly consistent across the range of fitted values. A Scale-Location plot (Figure 6) was also derived to assess whether the residuals are evenly spread across the predictors. Fortunately, both plots appear relatively horizontal, suggesting homoscedasticity visually. However, the Breusch-Pagan test was performed for better detection. The null hypothesis states that the error terms are homoscedastic, while the alternative hypothesis indicates heteroscedasticity. The significance level is stated at 0.05. The test returned a p-value of 8.088e^-6, leading to the rejection of the null hypothesis and confirming the presence of heteroscedasticity. As a result, a Box-Cox transformation is considered as a next step. However, the assumption of normality was checked first since Box-Cox is able to address both non-normality and heteroscedasticity. In simpler terms, the model was tested for normality and equal variance before performing Box-Cox.

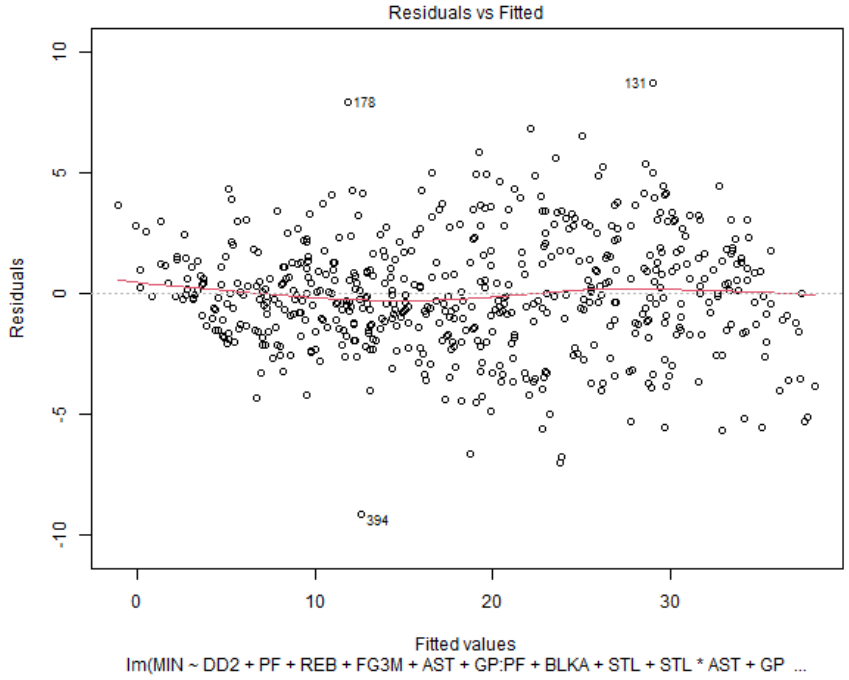

Figure 5: Residual vs. Fitted plot testing for equal variance for NBA data

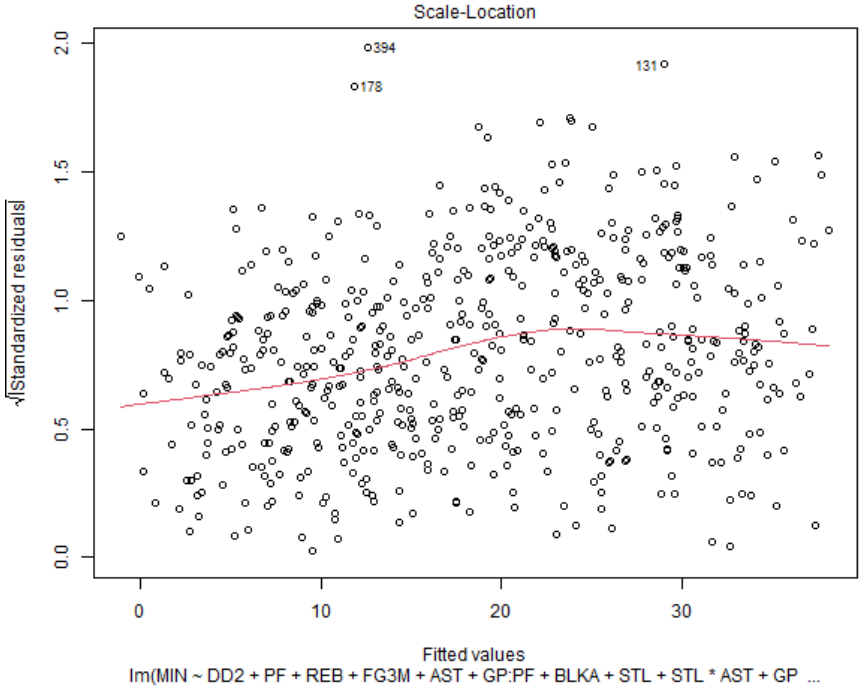

Figure 6: Scale-Location plot testing for equal variance for NBA data

3.2.4 Normality

To test whether the NBA data was normally distributed, the Shapiro-Wilk test and a Q-Q plot were used. The null hypothesis states that the data is normally distributed, while the alternative hypothesis indicates that it is not. For the final interactive model, the test returned a p-value of 0.04761, which is below the 0.05 significance level. This result leads to the rejection of the null hypothesis, suggesting that the data is not normally distributed. As shown in Figure 7, Q-Q plot visually supported this conclusion, as the data points deviated significantly from the trend line at both tails, indicating high kurtosis and non-normality. The histogram performed supported this statement as normal distribution cannot be seen visually (Figure 8). Based on these results, a Box-Cox transformation was performed to address both the failed normality and equal variance assumptions.

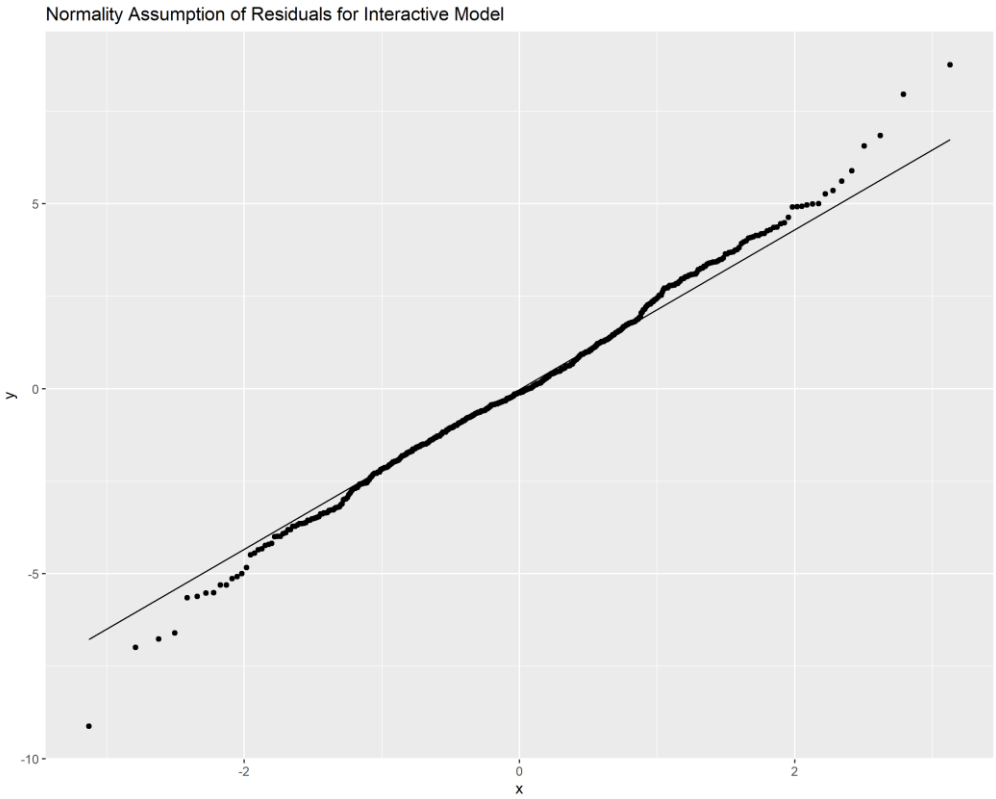

Figure 7: Q-Q plot testing normality for NBA data

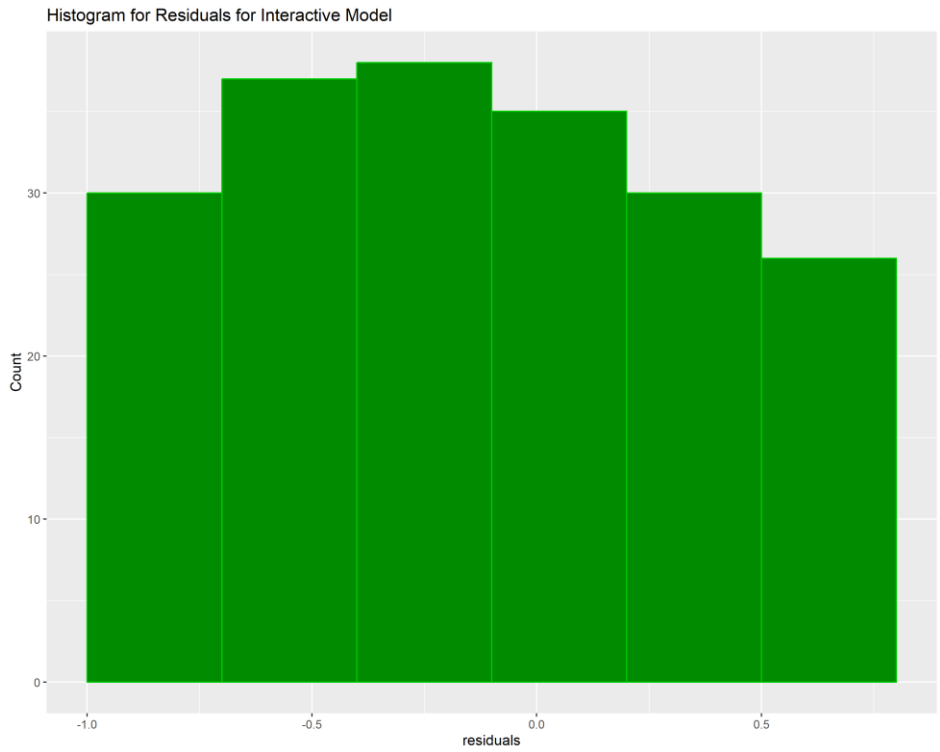

Figure 8: A histogram proving that NBA data is not normally distributed

3.2.5 Box-Cox Transformations

The first step was to derive the best lambda value for the final interactive model as it failed both the normality and equal variance assumptions. This means that a transformation is needed on the response variable, MIN. A Box-Cox plot shown in Figure 9 was derived to find the parameter for the best lambda value between –1.0 and 1.0. An output suggested an estimated of 0.7777778 for the best lambda which falls within the assigned parameter. Results from the Box-Cox Transformations showed that the data was normally distributed as the histogram shown in Figure 10 has improved, however the heteroscedasticity remained unchanged. Interestingly enough, the Q-Q plot shown in Figure 11 shows improvement in the data points as they are found closer to the trend line at both tails. However, the Breusch-Pagan test performed on the Box-Cox model obtained a p-value of 0.0005639 which still rejects the null hypothesis. The Shapiro-Wilk test of the Box-Cox model concluded a p-value of 0.2076 stating that the NBA data is now normally distributed as significance level is stated at 0.05. The Box-Cox transformation may not have resolved the heteroscedasticity because of the nature of the data. Figure 12 shows the trend of equal variances visually, however the Breusch-Pagan test performed suggests otherwise. The NBA sample size is rather large making it harder for the transformation to fully create homoscedasticity.

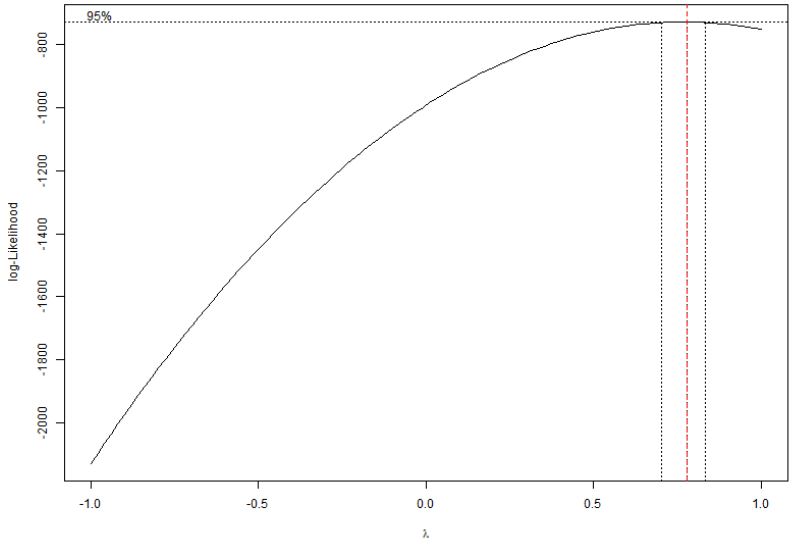

Figure 9: Box-Cox Plot estimating the best lambda value for NBA data

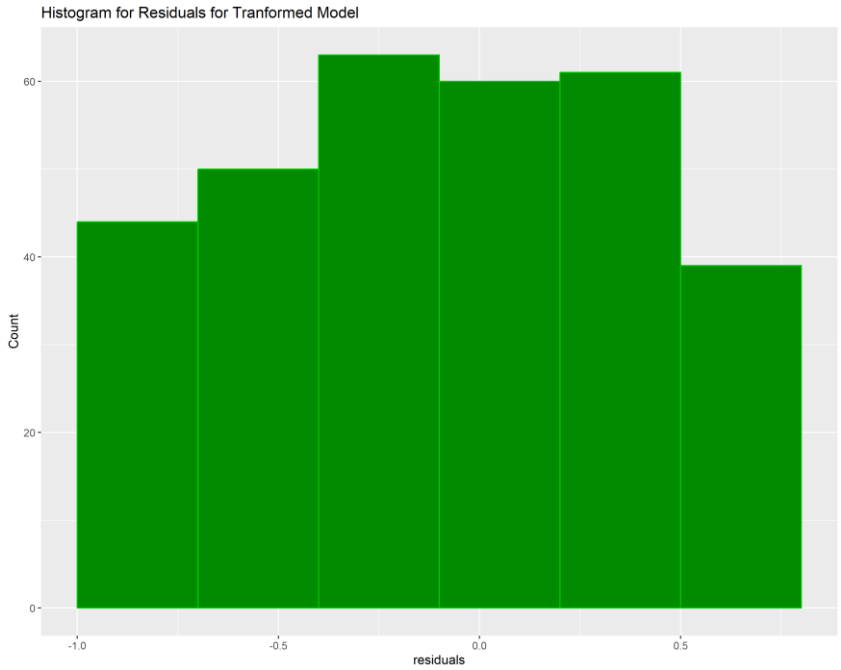

Figure 10: Histogram representing normality for the Box-Cox transformed NBA data

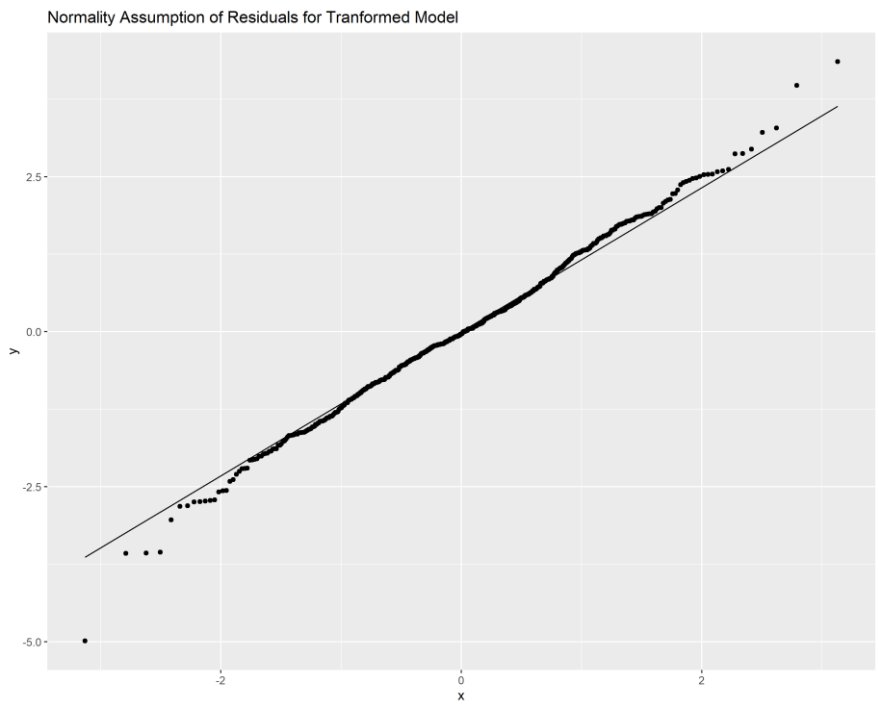

Figure 11: Q-Q Plot testing normality for the Box-Cox transformed NBA data

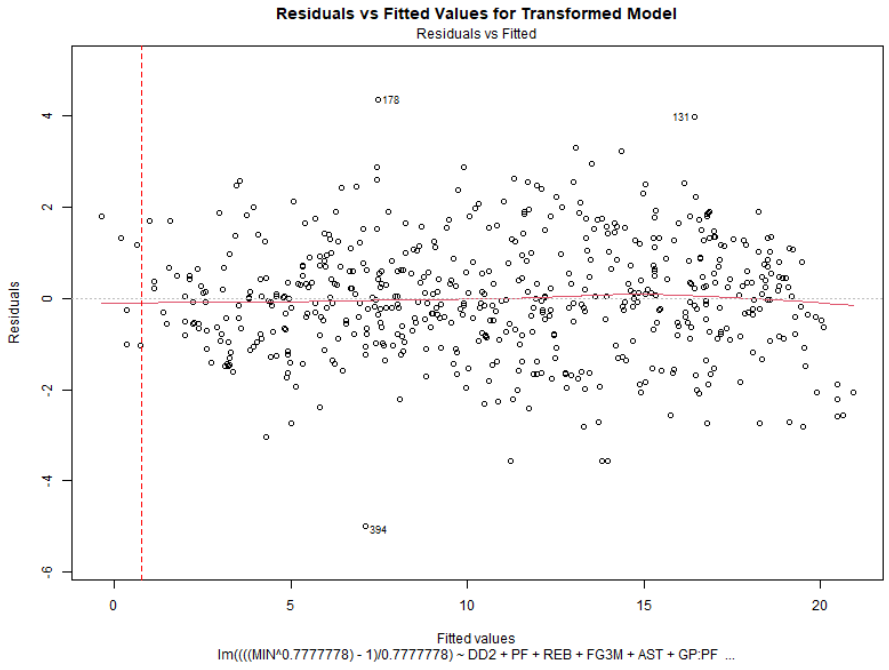

Figure 12: A Residual vs. Fitted plot of the Box-Cox transformed NBA data

3.3 Outliers

To identify the outliers in the model, a residual vs leverage graph was plotted. Most of the points lie close to zero indicating the model is a good fit as seen in Figure 13 but three points i.e. 394, 360 and 529 have large residuals or higher leverage comparatively, therefore these points might be influential. To identify whether to keep these points in the model or not, the cook's distance was plotted, as shown in Figure 14. Point 529 has the Cook's distance of 0.05 approximately, point 360 has the cook's distance of 0.07 approximately and point 394 has the cook's distance of 0.09 approximately. Since all these points have the Cook's distance below 0.5, we decided to keep them in the model.

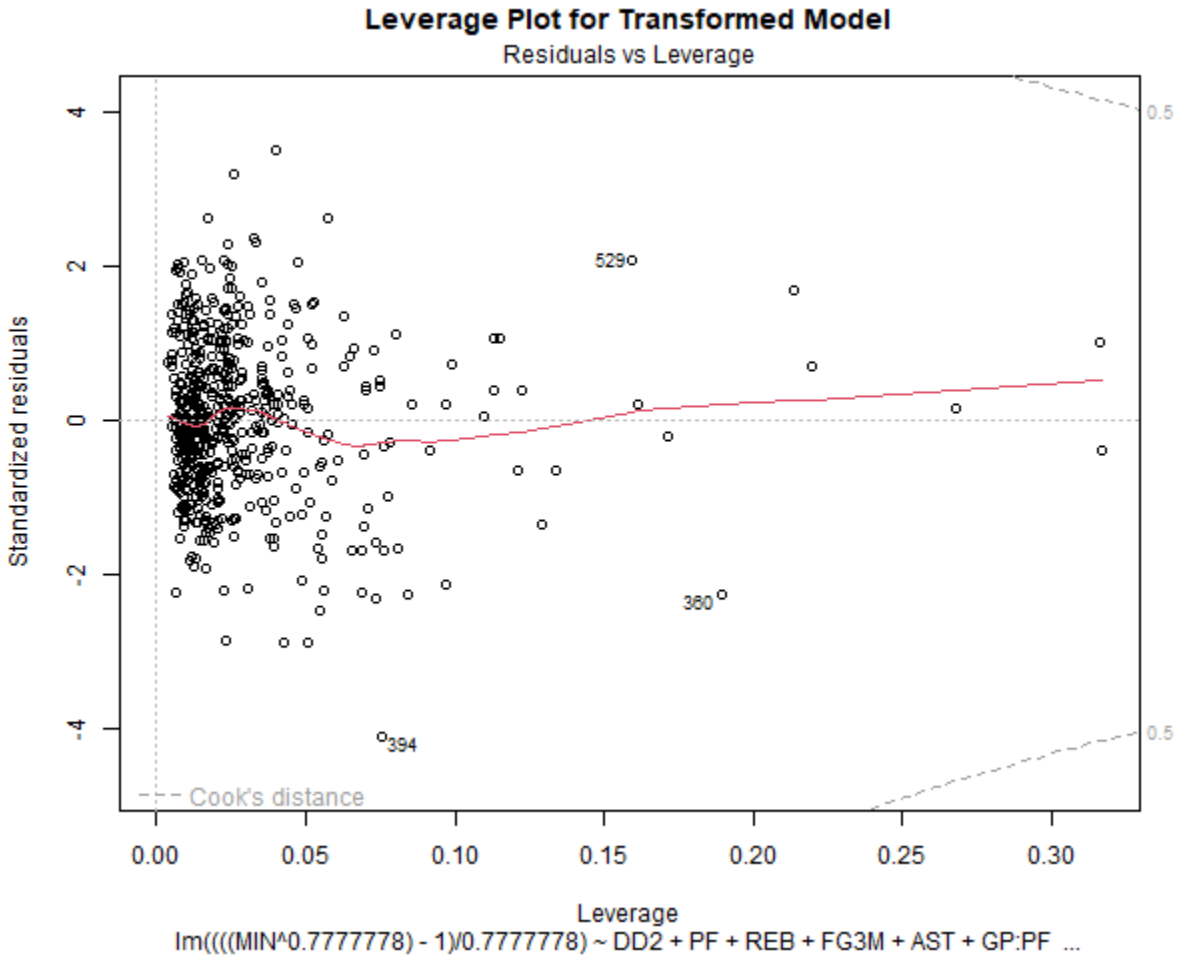

Figure 13: Residual vs Leverage Plot of the Box-Cox transformed NBA data

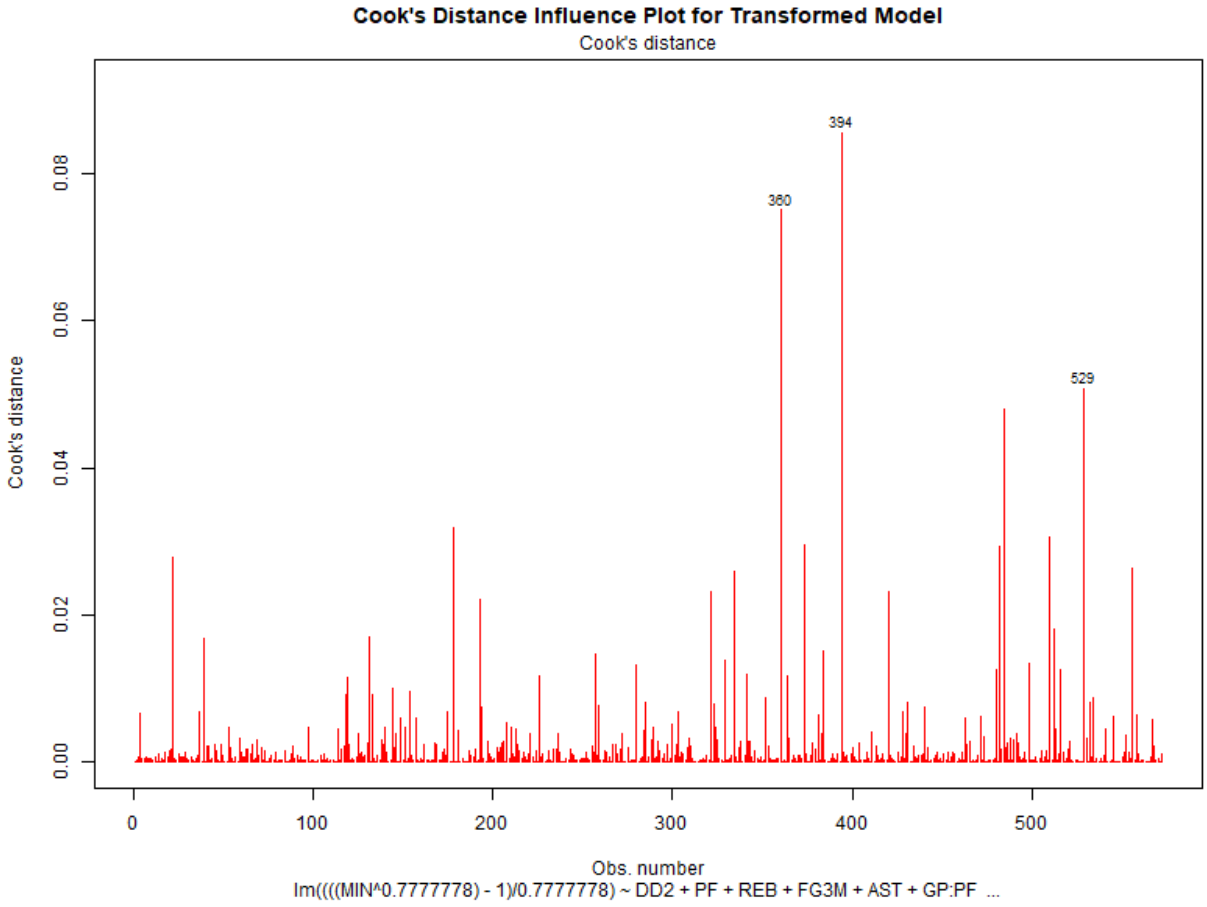

Figure 14: Cook's Distance Plot of the Box-Cox transformed NBA data

3.4 Predictions & Final Model

The NBA data was divided into training and testing, with eighty percent of the data being used for training, having 459 rows, and twenty percent for testing, having 113 rows. Firstly, the Box-Cox transformed model was used for training and testing. This model has a training accuracy of 13.76% and a testing accuracy of 4.68%, with a mean absolute error of 86.19. These results clearly suggest that the model is underfitted and is not suitable for prediction of Minutes played of the player when given other statistics.

Therefore, a final interactive model was used so check if this performs better than the other one. The interactive model has a training accuracy of 94.52% and the testing accuracy of 92.71%, with mean squared error of 6.59. This model is good for predictions as both training and testing accuracy are close to one another. However, this model did not pass the normality and equal variance assumptions.

Models	Mean Absolute Error	Training Accuracy	Testing Accuracy
Transformed Model	86.19	13.76%	4.68%
Interactive Model	6.59	94.52%	92.71%

Table 5: Training and Testing Accuracies for Transformed and Interaction Model (3.4)

4. CONCLUSION AND DISCUSSION

4.1 Approach

The final model identified some important positive influences of the minutes played per player in the NBA. These key factors include Games Played (GP), Personal Fouls (PF...16), Three-Point Field Goals Made (FG3M), Rebounds (REB), Assists (AST), Steals (STL), Attempted Blocks (BLKA)and interaction term, GP * PLUS_MINUS. Our final model indicates that players with the higher values in these key factors tend to have increased playing time. This concludes that players who excel in these aspects are valuable to their teams and see more playing time. Therefore, a player wanting more playing time can focus on their improvement on these variables.

Key factors that include Double Doubles (DD2), Plus-Minus (PLUS_MINUS) and interaction terms, PF * GP, AST* STL, REB * PLUS_MINUS, FG3M * AST, and REB * AST show a negative impact on playing time. These variables and interactions suggest that even though a player excels in some areas, there can be a limit to extra play time. It could lead to reducing their playing time even if they are performing well.

The approach that we took during this project was promising, as each step was rigorous and thorough. Our methodology was particularly effective at addressing and preventing the issues of overfitting, as the preliminary use of the VIF test on all the variables in our dataset worked to identify factors that could exhibit collinear behavior. Furthermore, pairing this test with the use of a pairwise plot consisting of variables with critical values further helped us determine which exact variables had collinearity between each other; we were then able to use domain knowledge to select which correlated variable to keep and which one to drop. A final VIF test was then ran on this model to ensure that the selected variables no longer exhibited correlation. Overall, this approach was very effective, as the issues of collinearity and overfitting between variables were minimized before we began searching for our ideal regression model.

Despite our model following a thorough and rigorous approach, the accuracy of our model can be questioned, given the fact that our selected model failed the assumptions of linearity, equal variance and normality assumptions. However, after applying the Box Cox transformation and performing the Shapiro-Wilk test on the transformed model, the results indicated that the normality assumption was successfully addressed, thereby enhancing the validity of the final

model. The Box Cox transformation was unable to address the equal variance assumption possibly due to the large sample size of the dataset. In conclusion, while the model provides valuable insights into the relationships between the minutes played per game and the key predictor variables, some limitations remain.

4.2 Future Work

Directions we can take in the future regarding this topic can be to investigate how player and team performance can have an impact on the overall revenue and value of the team itself. It goes without a doubt that the presence of higher skilled players will help increase the odds that a team will win their game on a nightly basis. Therefore, the use of logistical regressions could potentially be employed to see which individual/team statistics have the greatest influence on winning. This would be interesting, as some teams in the NBA are worth billions of dollars more than other teams; typically, these higher valued teams have greater team success and appear as more competitive contenders for the NBA title every year. Therefore, identifying which statistic has the highest impact on winning can help teams prioritize obtaining players that fulfill these needs, while also favoring coaching strategies that emphasize these facets.

5. REFERENCES

1. Lipman, D. (2023). *Multiple Regression Models: Part 1* [Lecture notes]. University of Calgary. Based on content created by T. Ngamkham, Fall 2024.
2. Lipman, D. (2023). *Multiple Regression Models: Part 2* [Lecture notes]. University of Calgary. Based on content created by T. Ngamkham, Fall 2024.
3. Lipman, D. (2023). *Multiple Regression Models: Part 3* [Lecture notes]. University of Calgary. Based on content created by T. Ngamkham, Fall 2024.
4. Lipman, D. (2023). *Multiple Regression Models: Part 4* [Lecture notes]. University of Calgary. Based on content created by T. Ngamkham, Fall 2024.
5. National Basketball Association. (2023, July 15). *Celtics' Jayson Tatum agrees to 5-year supermax extension*. NBA.com. Retrieved November 30, 2024, from https://www.nba.com/news/celtics-jayson-tatum-supermax-extension
6. National Basketball Association. (n.d.). *NBA stats*. NBA.com Retrieved November 29, 2024, from https://www.nba.com/stats
7. RunRepeat. (n.d.). *NBA team values analysis*. RunRepeat.com. Retrieved November 30, 2024, from https://runrepeat.com/nba-team-values-analysis