1. Abstract

This study looks into how different factors about basketball players from the National Basketball Association (NBA) league, like minutes played on the court, assists per basketball game, and percent of free throw scores made successfully affect the performance of players. The main factors we will be analyzing are minutes played, field goal percentage, two point percentage, assists per game, and points per game, as well as more outlined in section 4. By studying these factors, we aim to determine the specific characteristics that make basketball players perform better on the court. Prior research has indicated that some of these factors directly reflect a basketball player's performance. The overall goal is to give prospective basketball players, basketball recruiters, and basketball enthusiasts a better understanding of how these features impact the outcome of an athlete on the court, helping them make informed and educated choices when training or recruiting. Our research concluded that free-throw percentage, 2-pointer percentage, 3-pointer percentage, assists per game and total rebounds are predictors that may have an impact on points scored per game. In addition, we found that minutes played and assists have different impacts on points scored when controlling for age and field goal percentage and 2-pointer percentage and 3-pointer percentage have different impacts on field goal percentage when controlling for minutes played and field goal attempts.

2. Introduction

There is a growing emphasis on data-driven decision making in professional basketball and it has transformed how teams evaluate player performance and optimize their team strategies. This study focuses on the Los Angeles Lakers, an NBA franchise with a successful history. By examining key performance metrics of players from the 2003 to 2024 seasons, we aim to uncover the nuanced relationships between different player attributes and how they translate to success on the court. These variables, which include points scored, number of assists, turnovers, minutes played, shooting percentages, and player position, serve as critical lenses through which to analyze the drivers of individual and team performance.

The aim for this research is to move beyond isolated analyses of player statistics and explore how combinations of factors interact to shape outcomes. Building upon work done by prior research, this study integrates traditional statistical techniques with modern analytical frameworks. By employing robust statistical models, we aim to provide actionable insights into player development, team strategy, and resource allocation. As the NBA continues to evolve, such insights are indispensable for fostering strategies that align with immediate performance and long-term success.

3. Background

There are a few variables that we examine that have been used in previous literature such as age, minutes played, and position and their effect on performance metrics such as points scored. The study “Modelling player performance in basketball through mixed models” (Casals & Martinez, 2013) is one such study. In this study, the variable minutes played was highly associated with points and after applying model building procedures, age, player position and minutes played were included in a model where the response variable was points. As a result, we wanted to explore these variables and include them in our models since there was some level of validation.

This study and most others, however, mainly look at variables of a similar nature to age, minutes played, and position and their effect on performance. However, in our study, we also wanted to look at variables that indicate player efficiency and offensive and defensive skills (such as field goal percentage, free-throw percentage, 3-pointer percentage, assists, rebounds, and more) and explore how these skills affect player performance. Shooting percentages relate to player efficiency and offensive skills, while assists and rebounds are an indicator of offensive and defensive skills. As these variables are not metrics commonly studied, we wanted to study them and their effects on performance.

4. Data

The study utilized several datasets from *Basketball Reference,* an open-source website containing extensive professional basketball data (Sports Reference, 2025). Data cleaning was performed on the tables. The rows with discrepancies and missing values were dropped, about 10 observations. This dataset comprises information about all of the current Lakers players, encompassing 157 observations and a total of 32 variables. Following an exploratory analysis, a thorough literature review, and a discussion of the study’s hypotheses, we selected the following variables, including the explanatory and response variables.

Explanatory*:*

RQ1:

- *X1* → Minutes Played (MP): Total number of minutes played on the court.
- *X2* → Assists Per Game (AST): Passing the ball to a teammate that results in that teammate scoring a basket.
- *X3* → Age (Age): Age in years of player.
- *X4* → Field Goal Percentage (FG%): Ratio of shots made to shots attempted, not including free throws.

RQ2:

- *X1* → Free Throw Percentage (FT.): Ratio of free throw shots made to attempted.
- *X2* → 2-Pointer Percentage (X2P.): Ratio of 2-pointer shots made to attempted.
- *X3* → 3-Pointer Percentage (X3P.): Ratio of 3-pointer shots made to attempted.
- *X4* → Turnovers (TOV): Number of times a player loses possession of the ball to the opposing team before attempting a shot

RQ3:

- *X1* → 3-Pointer Percentage (X3P.): Ratio of 3-pointer shots made to attempted.
- *X2* → 2-Pointer Percentage (X2P.): Ratio of 2-pointer shots made to attempted.
- *X3* → Field Goal Attempts (FGA): Any shot taken on the court, excluding free throws

- *X4* → Minutes Played (MP): Total number of minutes played on the court.

RQ4:

- *X1* → Field Goal Percentage (FG%): Ratio of shots made to shots attempted, not including free throws.
- *X2* → Assists Per Game (AST): Passing the ball to a teammate that results in that teammate scoring a basket.
- *X3* → Total Rebounds (TRB): Number of times a player retrieves the ball after a missed field goal or free throw attempt.
- *X4* → Player Position (Pos): Designated role of a player on the court (e.g., point guard, small forward, center forward) that defines their typical responsibilities and play style.

Response:

RQ1:

- Y → Points Per Game (PTS): Average number of points scored per game.

RQ2:

- Y → Points Per Game (PTS): Average number of points scored per game.

RQ3:

- Y → Field Goal Percentage (FG%): Ratio of shots made to shots attempted, excluding free throws.

RQ4:

- Y → Points Per Game (PTS): Average number of points scored per game.

5. Research Questions

1. Research Question (RQ1): *Do minutes played (X_1) and assists per game (X_2) have similar linear impacts on points per game (Y) among Lakers players, while controlling for age (X_3) and field goal percentage (X_4)?*

This question aims to explore whether minutes played, assists per game, age, and field goal percentage exert a significant linear influence on points per game among all Lakers players. The predictor *minutes played* was chosen because we believe as players are on the court more, they have more opportunity to score, which can improve their points per game. *Assists per game* is another predictor, as players with more assists tend to have better ball movement, leading to more points scored. *Points per game* is the predictor variable, as players who score more points are usually better players.

The hypotheses are:

$$H_0: \beta_1 = \beta_2$$

$$H_a: \beta_1 \neq \beta_2$$

The full model to address this question was built as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

The reduced model is represented as:

$$Y = \beta_0 + \beta\ (X_1 + X_2) + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

2. Research Question (RQ2): *Do free throw percentage (X1), two-point percentage (X2), and three-point percentage (X3) have a significant linear impact on points per game (Y), when accounting for turnovers (X4)?*

This question aims to explore whether the percentage of free throws, two pointers, and 3 pointers have a significant linear impact on the number of points scored per game. The percentages of these variables was chosen specifically to test whether efficiency in scoring a certain number of points has an effect on the number of points scored, independent of how many turnovers a player makes. If a certain shooting type has a significant impact on points per game, then improving the efficiency of that shooting type could improve the number of points per game scored, improving player performance.

The hypotheses are:

$$\mathrm{H_0} : \beta_1 = \beta_2 = \beta_3 = 0$$

$$\mathrm{H_a} : \textit{At least one}\ \beta_1,\ \beta_2,\ \beta_3 \neq 0$$

The full model to address this question was built as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

The reduced model is represented as:

$$Y = \beta_0 + \beta_4 X_4 + \epsilon$$

3. Research Question (RQ3): *When accounting for minutes played (MP) and field goal attempts (FGA), does three-point percentage (3P%) or two-point percentage (2P%) have a different effect on field goal percentage (FG%)?*

The question aims to explore whether one shooting efficiency metric (3P% vs. 2P%) matters more in determining overall shooting efficiency (FG%), independent of how much a player shoots (FGA) or plays (MP). The larger coefficient could suggest that improving that shooting type contributes more to FG%. In the real world, we could use this to ask questions such as "Should this player focus on improving 3P% or 2P%?" or evaluate how different players contribute to efficient scoring.

The hypotheses are:

$$H_0 : \beta_1 = \beta_2$$

$$H_A : \beta_1 \neq \beta_2$$

The full model to address the question was built as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

The reduced model is represented as:

$$Y = \beta_0 + \beta\ (X_1 + X_2) + \ \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

4. Research Question (RQ4): *Do assists per game (X_2) and total rebounds (X_3), provide additional explanatory power for points per game (Y), beyond what is explained by field goal %(X_1) and player position (X_4)?*

The question aims to explore whether playmaking (assists) and rebounding contribute to a player's scoring ability, once we already account for shooting efficiency (FG%) and role on the court (position). By checking if assists and rebounds significantly improve the model's ability to predict points per game, we can better understand how different aspects of a player's game drive scoring. In the real world, this could help coaches and analysts identify whether encouraging players to improve in areas like passing or rebounding might also boost their scoring, or if focusing solely on shooting and positional fit is sufficient.

The hypotheses are:

$$H_0\colon\ \beta_2 = \beta_3 = \beta_4 = 0$$

H_A: At least one β_2, β_3, $\beta_4 \neq 0$

The full model to address the question was built as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

The reduced model is represented as:

$$Y = \beta_0 + \beta_1 X_1 + \ \beta_4 X_4$$

6. Methodology

Given the confirmatory observational nature of our study, we performed the following methodologies:

General Linear F Test

To conduct hypothesis testing, the general linear F test was performed. This test defines a full model and a reduced model. A comparison between the error sum of squares (SSE) of these models was executed while considering the degrees of freedom for each model. The general linear F-statistic was used to determine the significance level of the difference between the SSE of both models. The F-statistic is used as a metric to determine whether the observed difference in SSE between the models is substantial enough to support the rejection or acceptance of the null hypothesis.

$$F * \quad = \; (\frac{SSE(R) - SSE(F)}{df_R - df_F}) \; \div \; \frac{SSE(F)}{df_F}$$

Diagnostic Procedures

Various methods were implemented to assess each full model's data and representation. This process addressed limitations within the models and determined their efficacy, whether they performed well, improved using remedial procedures, and inferences made were addressed with caution. The primary objective was to decipher the models' fit, potential limitations and shortcomings, and the steps to improve upon the models.

To test for normality in the errors, a Shapiro-Wilk test was performed for each model, and the QQ-plot was created to visualize the violation of normality. The null hypothesis of the test states that the data follows a normal distribution, and the alternate hypothesis says otherwise. A significant result from this test indicates that the error terms are violating the first assumption of normal distribution.

A statistical test for non-constant variance was performed using Breusch-Pagan (BP) test. The BP test assumes that the error terms are independent and normal, and the variance depends on X. This test was chosen rather than the Brown Forsythe (BF) test because the data does not have to be split into groups, making it a better fit for multilinear regression (MLR).

Multicollinearity was also tested to assess if any of the predictors are linearly dependent. Multicollinearity may change the value of the parameters depending on the order of the variables, yield unstable and unreliable outputs, and may increase parameters' standard deviation, making it hard to interpret the individual effects of each variable. To test for multicollinearity, a scatter plot was plotted and correlation coefficients between each predictors were computed. The Variance Inflation Factor (VIF) was used to further assess.

We also performed the "best" subsets algorithm, as it provides the best subsets according to the specified criterion and identifies several good subsets for each possible number of predictors. Identifying outliers was another pivotal step in the diagnostic process. To detect outliers, we computed studentized residuals, comparing if the absolute value is larger than 2 — if so, it is classified as an outlier. For potential outliers among the independent variables, we evaluated the ith diagonal element of the hat matrix, h_{ii}— if it is larger than twice the as large as the mean leverage, $\overline{h} = p/n$ (number of parameters divided by number of data points). There may be some outliers that are not influential whatsoever, so to test influential points, we employed Cook's distance. Cook's distance measures the influence of the *i*th case on the overall fitted values. These approaches provide a comprehensive analysis on distinguishing outliers and influential points.

Remedial Procedures

If a nonlinearity was detected, a transformation of X was performed. If normality or constant variance among errors terms were violated, we considered transformations for Y. Using the Box-cox transformation, we plotted log-likelihood for the optimal lambda, λ, and transformed Y to Y^{λ}.

When non-constant variance was not fixed after transforming Y, Weighted Least Squares regression was implemented to address heteroscedasticity. The weighted matrix, *W*, is used to create a (weighted) data with constant variance, therefore providing valid inference in presence of heteroscedasticity.

For cases with influential outliers, we employed Robust regression. This approach mitigated the influence of these outliers, providing a more accurate parameter estimation of the model.

7. Discussion of Results

The initial analysis made before processing each model was checking the linear relationship of the predictors and response variables. The relationship of each variable is shown below through figures 1 to 4.

In figures 1, 2, and 3, a distinct nonlinear association was observed between the predictor variables and the response variable. To address this, we would perform a transformation of X if there is solely a nonlinear violation. However, from further analysis (covered later), we found further violations, so a transformation of X was not fully executed.

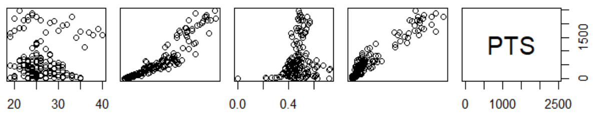


Figure 1. RQ1 Scatterplot matrix of Age, MP, FG%, AST, PTS

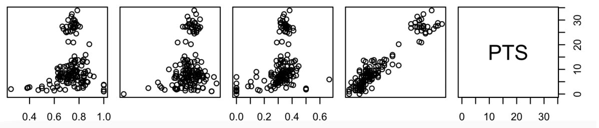


Figure 2. RQ2: Scatterplot Matrix of FT., X2P., X3P., and TOV

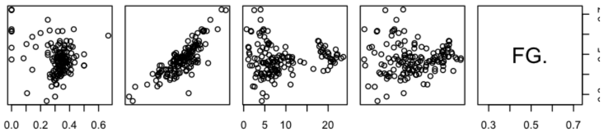


Figure 3. RQ3: Scatterplot Matrix of 3P%, 2P%, FGA, MP, FG%

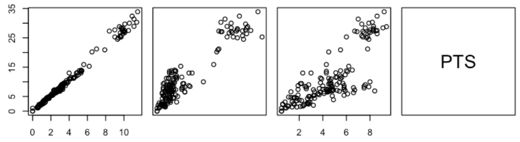
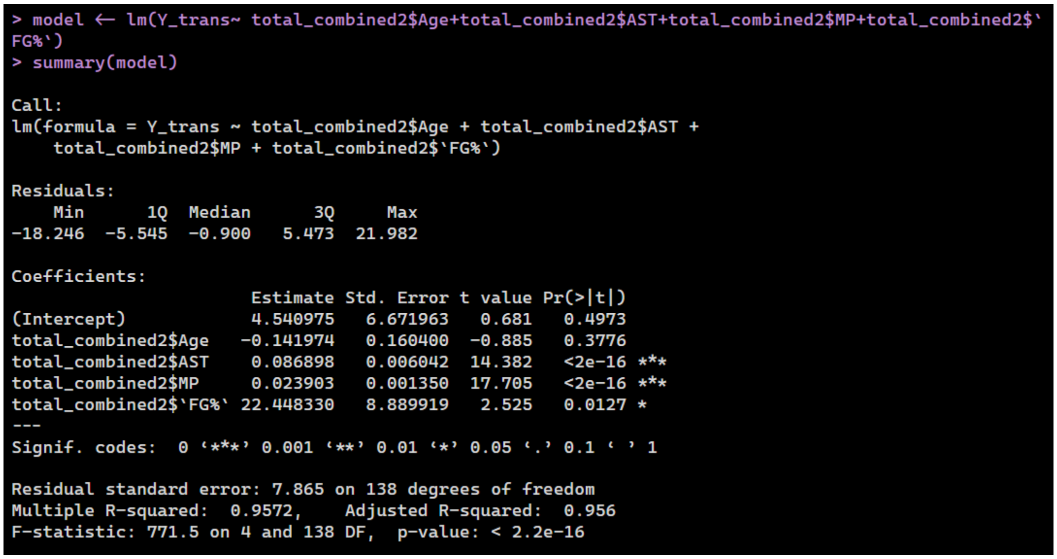


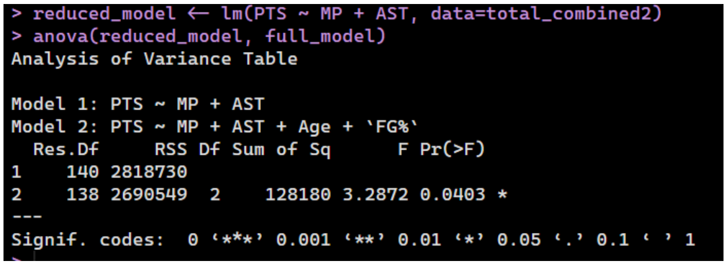
Figure 4. RQ4: Scatterplot Matrix of FG%, Assists, Total Rebounds, and Points per Game

Research Question 1



```
> model ← lm(Y_trans~ total_combined2$Age+total_combined2$AST+total_combined2$MP+total_combined2$`
FG%`)
> summary(model)

Call:
lm(formula = Y_trans ~ total_combined2$Age + total_combined2$AST +
    total_combined2$MP + total_combined2$`FG%`)

Residuals:
    Min      1Q  Median      3Q     Max
-18.246  -5.545  -0.900   5.473  21.982

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.540975   6.671963   0.681   0.4973
total_combined2$Age   -0.141974   0.160400  -0.885   0.3776
total_combined2$AST    0.086898   0.006042  14.382   <2e-16 ***
total_combined2$MP     0.023903   0.001350  17.705   <2e-16 ***
total_combined2$`FG%` 22.448330   8.889919   2.525   0.0127 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.865 on 138 degrees of freedom
Multiple R-squared:  0.9572,    Adjusted R-squared:  0.956
F-statistic: 771.5 on 4 and 138 DF,  p-value: < 2.2e-16
```

Figure 5. Model Summary of Final Model



```
> reduced_model ← lm(PTS ~ MP + AST, data=total_combined2)
> anova(reduced_model, full_model)
Analysis of Variance Table

Model 1: PTS ~ MP + AST
Model 2: PTS ~ MP + AST + Age + `FG%`
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    140 2818730
2    138 2690549  2    128180 3.2872 0.0403 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6. ANOVA Summary of Reduced vs. Full Model

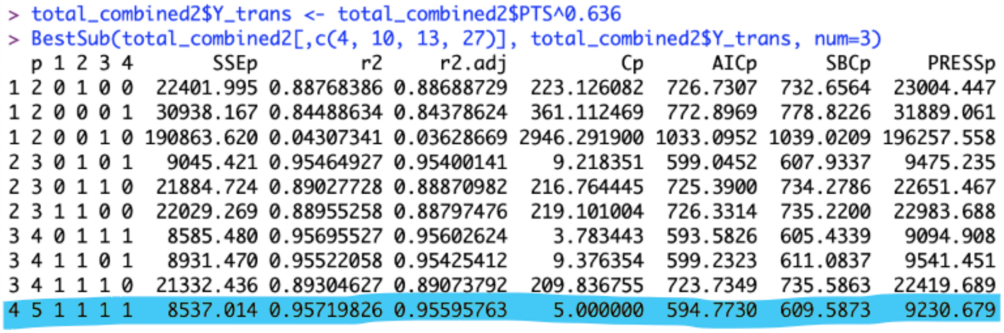
The General Linear Test (GLT) conducted between the Full and Reduced models resulted in an F-statistic of 3.872 with a corresponding p-value of 0.0403, leading to the rejection of the null hypothesis. This rejection indicates that the predictor variables do not have a similar linear impact.

Further diagnostic analyses uncovered pertinent issues. Residual plots were plotted against predictor variables that showed potential heteroscedasticity in the error terms. This is further supported by the Breusch-Pagan (BP) test yielded non-constant variance in the errors, with a p-value of 0.00000553, telling us there is heteroscedasticity. The Shapiro-Wilk test yielded a p-value of 0.0398, indicating the normality assumption is also violated, which is also supported visually through the QQ-plot. Through the Variance Inflation Factor (VIF) analysis, no multicollinearity issues were found as all VIF scores were below the threshold (of 10).

To address the non-normality issue in the response variable, a Y transformation using Box-Cox was implemented. A plot of log-likelihood yielded a peak at $\lambda = 0.636$, transforming the Y to $Y \rightarrow Y^{0.636}$. This resolved the normality problem, as indicated by a successful Shapiro test (p-value = 0.219). Additionally, Weighted Least Squares (WLS) regression was applied to address non-constant variance. The WLS model offered a model that yielded constant variance (homoscedasticity).

Identification of outliers in the response variables (points 65, 79, 81, 84) and the predictor variables (points 79, 81, 83, 84), as determined using studentized residuals and hat matrix values. Cook's Distance analysis didn't show any influential points.

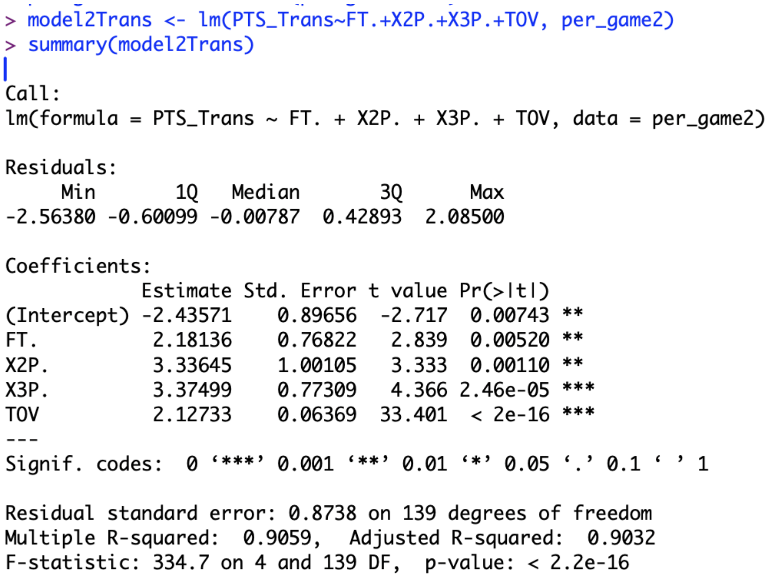
Running the Best subset algorithm, adding all parameters to the model yielded the highest R^2 values, $c_p = p$, and lowest AIC, SBC, PRESS values, indicating this is the best possible subset.



```
> total_combined2$Y_trans <- total_combined2$PTS^0.636
> BestSub(total_combined2[,c(4, 10, 13, 27)], total_combined2$Y_trans, num=3)
  p 1 2 3 4       SSEp         r2     r2.adj          Cp      AICp      SBCp     PRESSp
1 2 0 1 0 0  22401.995 0.88768386 0.88688729  223.126082  726.7307  732.6564  23004.447
1 2 0 0 0 1  30938.167 0.84488634 0.84378624  361.112469  772.8969  778.8226  31889.061
1 2 0 0 1 0 190863.620 0.04307341 0.03628669 2946.291900 1033.0952 1039.0209 196257.558
2 3 0 1 0 1   9045.421 0.95464927 0.95400141    9.218351  599.0452  607.9337   9475.235
2 3 0 1 1 0  21884.724 0.89027728 0.88870982  216.764445  725.3900  734.2786  22651.467
2 3 1 1 0 0  22029.269 0.88955258 0.88797476  219.101004  726.3314  735.2200  22983.688
3 4 0 1 1 1   8585.480 0.95695527 0.95602624    3.783443  593.5826  605.4339   9094.908
3 4 1 1 0 1   8931.470 0.95522058 0.95425412    9.376354  599.2323  611.0837   9541.451
3 4 1 1 1 0  21332.436 0.89304627 0.89073792  209.836755  723.7349  735.5863  22419.689
4 5 1 1 1 1   8537.014 0.95719826 0.95595763    5.000000  594.7730  609.5873   9230.679
```

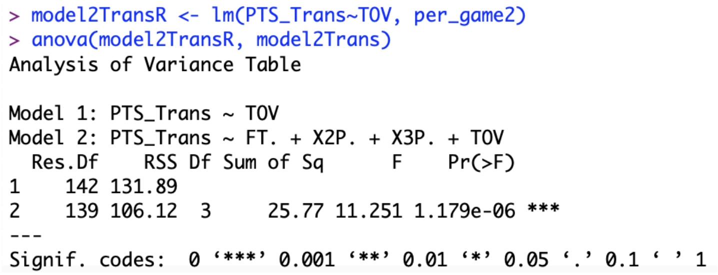
Figure 7. Best Subset Output for RQ1

Research Question 2



```
> model2Trans <- lm(PTS_Trans~FT.+X2P.+X3P.+TOV, per_game2)
> summary(model2Trans)

Call:
lm(formula = PTS_Trans ~ FT. + X2P. + X3P. + TOV, data = per_game2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.56380 -0.60099 -0.00787  0.42893  2.08500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.43571    0.89656  -2.717  0.00743 ** 
FT.          2.18136    0.76822   2.839  0.00520 ** 
X2P.         3.33645    1.00105   3.333  0.00110 ** 
X3P.         3.37499    0.77309   4.366 2.46e-05 ***
TOV          2.12733    0.06369  33.401  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8738 on 139 degrees of freedom
Multiple R-squared:  0.9059,	Adjusted R-squared:  0.9032 
F-statistic: 334.7 on 4 and 139 DF,  p-value: < 2.2e-16
```

Figure 8. Model Summary of Final Model



```
> model2TransR <- lm(PTS_Trans~TOV, per_game2)
> anova(model2TransR, model2Trans)
Analysis of Variance Table

Model 1: PTS_Trans ~ TOV
Model 2: PTS_Trans ~ FT. + X2P. + X3P. + TOV
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1    142 131.89                                  
2    139 106.12  3     25.77 11.251 1.179e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9. ANOVA Summary of Reduced vs Full Model

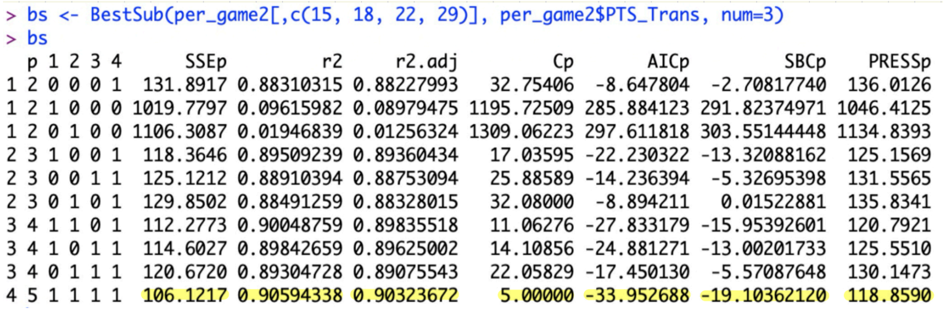
The General Linear Test (GLT) conducted between the Reduced and Full Model yielded a F-statistic of 11.251, with a p-value of 1.179e-06, leading to the rejection of the null hypothesis. This rejection indicates that at least one of the predictors, free-throw percentage (FT.), 2-pointer percentage (X2P.), or 3-pointer percentage (X3P.), has a beta coefficient different from 0, signifying that at least one of predictors has a significant linear impact on points per game, when accounting for turnovers.

In order to validate the assumptions of a linear model, diagnostic procedures were conducted. Residual plots plotted against the predictor variable showed some level of heteroscedasticity, or at the very least, plots not quite consistent with constant variance. Conducting the Breusch-Pagan test yields a p-value of 0.002365, which indicates non-constant variance in the error terms. The Shapiro-Wilks test yielded a p-value of 0.009281. Since this p-value is less than 0.05, it is concluded that the data violates the normality assumption, which is also supported by the QQ plot. Through the Variance Inflation Factor (VIF) analysis, no multicollinearity issues were found as the VIF values for each predictor were below 10, In fact, they were all between 1 and 1.2. Since there is no multicollinearity issue, it is not necessary to consider remedial procedures for multicollinearity like Ridge Regression.

Outlying Y observations were determined using studentized residuals. When points have an absolute studentized residual value larger than 2, they are considered outliers in the response variable, which was the case with 4 points: 61, 95, 97, 148. Outlying X observations were determined using the hat matrix values. Points with hat matrix values above 2p/n = 0.0694444, such as 95, 97, 104, and 148, are outliers in the predictor variables. Looking at Cook Distance values, the largest value is 0.221602 (Point 148), which is much smaller than the 20th percentile of 0.4674556, indicating point 148 (and other outliers) have minor to no influence. Since there are no major influential points, there is no need to consider Robust Regression.

To address the non-normality issue and the non-constant variance issue, a Box-Cox transformation was applied to the response variable, points per game. A plot of log-likelihood yielded a peak at $\lambda = 0.6969697$, transforming the Y to $Y \rightarrow Y^{0.6969697}$. This resolved the normality issue. Although the QQ plot post-transformation was not perfect, it was better than the original QQ plot. In addition, the Shapiro-Wilks test yielded a p-value of 0.09773, which is not less than 0.05, indicating the data now follows a normal distribution. The Box-Cox transformation also resolved the non-constant variance issue. Performing the Breusch-Pagan test yielded a p-value of 0.1817, so it can be concluded that the error terms have constant variance following the transformation. As the non-constant error variance issue was resolved with the Box-Cox transformation, there was no need to consider Weighted Least Squares (WLS).

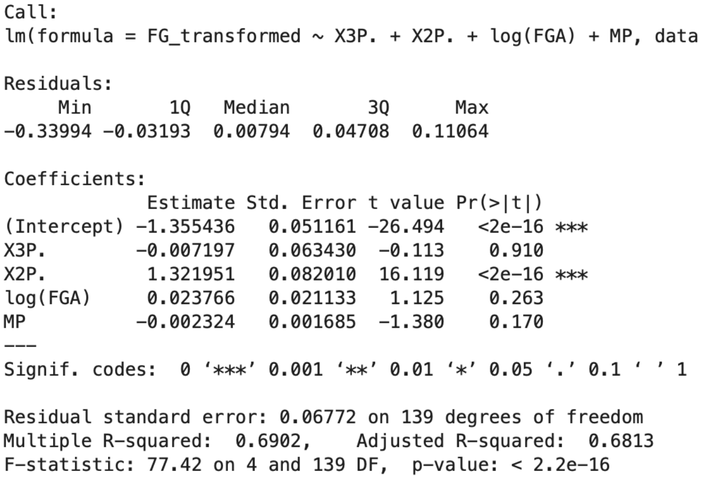
Running the Best subset algorithm, adding all parameters to the model yielded the highest R^2 values, $c_p = p$, and lowest AIC, SBC, PRESS values, indicating this is the best possible subset.



```
> bs <- BestSub(per_game2[,c(15, 18, 22, 29)], per_game2$PTS_Trans, num=3)
> bs
  p 1 2 3 4      SSEp         r2     r2.adj         Cp       AICp         SBCp    PRESSp
1 2 0 0 0 1  131.8917 0.88310315 0.88227993   32.75406  -8.647804  -2.70817740  136.0126
1 2 1 0 0 0 1019.7797 0.09615982 0.08979475 1195.72509 285.884123 291.82374971 1046.4125
1 2 0 1 0 0 1106.3087 0.01946839 0.01256324 1309.06223 297.611818 303.55144448 1134.8393
2 3 1 0 0 1  118.3646 0.89509239 0.89360434   17.03595 -22.230322 -13.32088162  125.1569
2 3 0 0 1 1  125.1212 0.88910394 0.88753094   25.88589 -14.236394  -5.32695398  131.5565
2 3 0 1 0 1  129.8502 0.88491259 0.88328015   32.08000  -8.894211   0.01522881  135.8341
3 4 1 1 0 1  112.2773 0.90048759 0.89835518   11.06276 -27.833179 -15.95392601  120.7921
3 4 1 0 1 1  114.6027 0.89842659 0.89625002   14.10856 -24.881271 -13.00201733  125.5510
3 4 0 1 1 1  120.6720 0.89304728 0.89075543   22.05829 -17.450130  -5.57087648  130.1473
4 5 1 1 1 1  106.1217 0.90594338 0.90323672    5.00000 -33.952688 -19.10362120  118.8590
```

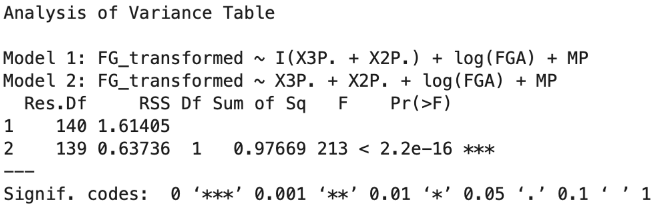
Figure 10. Best Subset Output for RQ2

Research Question 3



```
Call:
lm(formula = FG_transformed ~ X3P. + X2P. + log(FGA) + MP, data

Residuals:
     Min       1Q   Median       3Q      Max
-0.33994 -0.03193  0.00794  0.04708  0.11064

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.355436   0.051161 -26.494   <2e-16 ***
X3P.        -0.007197   0.063430  -0.113    0.910
X2P.         1.321951   0.082010  16.119   <2e-16 ***
log(FGA)     0.023766   0.021133   1.125    0.263
MP          -0.002324   0.001685  -1.380    0.170
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06772 on 139 degrees of freedom
Multiple R-squared:  0.6902,    Adjusted R-squared:  0.6813
F-statistic: 77.42 on 4 and 139 DF,  p-value: < 2.2e-16
```

Figure 11. Linear Model Output



```
Analysis of Variance Table

Model 1: FG_transformed ~ I(X3P. + X2P.) + log(FGA) + MP
Model 2: FG_transformed ~ X3P. + X2P. + log(FGA) + MP
  Res.Df     RSS Df Sum of Sq   F    Pr(>F)
1    140 1.61405
2    139 0.63736  1   0.97669 213 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12. Anova Output of Hypothesis Test

The General Linear Test (GLT) conducted between the reduced and full models yielded an F statistic of 220.86, leading to the rejection of the null hypothesis. This rejection signifies that the effects of three-point percentage (3P%) and two-point percentage (2P%) on field goal percentage (FG%) are significantly different, even after accounting for minutes played (MP) and field goal attempts (FGA).

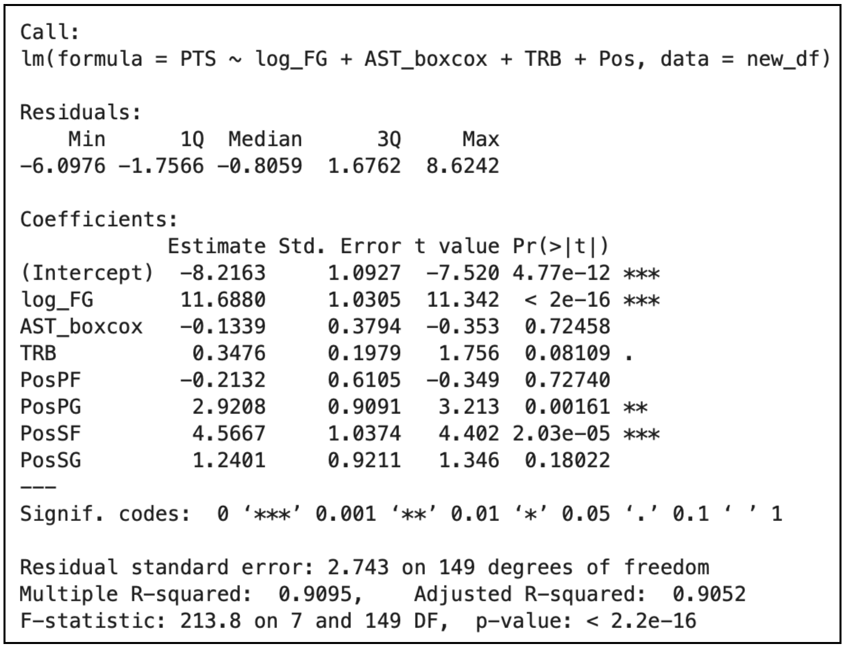
Diagnostic procedures were conducted to validate the assumptions of the linear model. Histograms of the predictor variables indicated that field goal attempts (FGA) were right-skewed; thus, a log transformation was applied prior to model fitting. This transformation successfully addressed the skewness and also improved the distribution of residuals and minimized the influence of outliers.

Variance Inflation Factor (VIF) analysis revealed no severe multicollinearity, with all VIF values well below the threshold of 10 (X3P% = 1.28, X2P% = 1.26, log(FGA) = 7.95, and MP = 7.96). Examination of Cook's Distance showed no influential points (largest Cook's D = 0.18), and outlier analysis via standardized residuals identified three rows (85, 95, and 138), but none necessitated removal or remedial action.

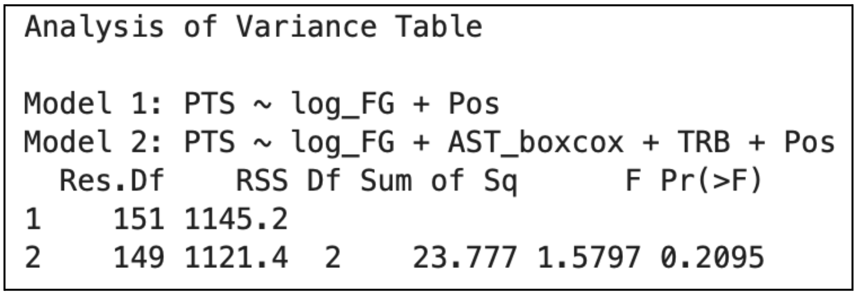
The Breush-Pagan (BP) test initially indicated the presence of non-constant variance with a p-value of 0.02844, and the Shapiro-Wilk test confirmed deviations from normality in the residuals with a p-value of 0.0001186. Following these findings, a Box-Cox transformation was applied to the response variable, yielding a lambda value of approximately 0.384. Post-transformation, the BP test p-value increased to 0.07825, suggesting that heteroscedasticity was effectively corrected. However, the Shapiro-Wilk test remained significant with a p-value of 1.528e-06, indicating a continued non-normality of residuals.

Despite the non-normality issue, no further transformations or remedial measures were required, as other diagnostic checks showed. The model outcomes align with expectations, and show the difference in impact of two-point and three-point shooting percentages on overall field goal success, after controlling for volume of shots taken and playing time.

Research Question 4



```
Call:
lm(formula = PTS ~ log_FG + AST_boxcox + TRB + Pos, data = new_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0976 -1.7566 -0.8059  1.6762  8.6242 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -8.2163     1.0927  -7.520 4.77e-12 ***
log_FG       11.6880     1.0305  11.342  < 2e-16 ***
AST_boxcox   -0.1339     0.3794  -0.353  0.72458    
TRB           0.3476     0.1979   1.756  0.08109 .  
PosPF        -0.2132     0.6105  -0.349  0.72740    
PosPG         2.9208     0.9091   3.213  0.00161 ** 
PosSF         4.5667     1.0374   4.402 2.03e-05 ***
PosSG         1.2401     0.9211   1.346  0.18022    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.743 on 149 degrees of freedom
Multiple R-squared:  0.9095,	Adjusted R-squared:  0.9052 
F-statistic: 213.8 on 7 and 149 DF,  p-value: < 2.2e-16
```



```
Analysis of Variance Table

Model 1: PTS ~ log_FG + Pos
Model 2: PTS ~ log_FG + AST_boxcox + TRB + Pos
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    151 1145.2                           
2    149 1121.4  2    23.777 1.5797 0.2095
```

The General Linear Test (GLT) conducted between the Reduced and Full models yielded an F statistic of 6.13, leading to the rejection of the null hypothesis. This rejection indicates that not all beta coefficients are equal to zero, suggesting that at least one of the added predictors has a significant linear relationship with points per game. In this case, assists and rebounds provided additional explanatory power beyond field goal percentage and player position, highlighting their meaningful contributions to a player's scoring ability.

Diagnostic analyses conducted to validate the assumptions on the model revealed further issues. The model was tested for heteroscedasticity with the Breusch-Pagan test, which yielded a p value of 0.00000000002152, and non-normality with the Shapiro-Wilk test which yielded a p value of 0.00005156. Both these p values are well below the chosen alpha of 0.05, indicating both normality and constant variance assumption violations. The existence of non-normality is further supported by the QQ-plot. Through the Variance Inflation Factor (VIF) analysis, no multicollinearity issues were found as all VIF scores were below the threshold (of 10).

To address the issue of heteroscedasticity, a Y transformation using Box-Cox was implemented. A plot of log-likelihood yielded a peak at $\lambda = 0.212$, transforming the Y to $Y \rightarrow Y^{0.212}$. This solved the issue of heteroscedasticity with the Breusch-Pagan test now yielding a p value of 0.2188.

This is higher than our chosen alpha of 0.05, meaning there is no longer an issue of non-constant variance.

To address non-normality, several transformations were attempted, including X transformations (log transform on X_1 (Field Goal Percentage) and Box-Cox transformation X_2 (Assists Per Game). However, these techniques failed to yield a lower Shapiro-Wilk p value, and simultaneously increased the Breusch-Pagan test p value, bringing back heteroscedasticity. As such a final model with only a Box-Cox transformation on Y was chosen, keeping the non-normality issue.

An examination of Cook's Distance showed no influential points with the largest Cook's Distance being at 0.17, and outlier analysis via standardized residuals identified two rows, 52 and 75, but neither necessitated removal or remedial action.

Despite the non-normality issue, no further transformations or remedial measures were required, as other diagnostic checks showed. The model outcomes align with expectations, and show the additional impact of assists per game and total rebounds, beyond what is explained by field goal percentage and player position.

K-Fold Cross Validation of the Models

Model	**RMSE**	**R-squared**
Model 1 (RQ1)	7.802	0.957
Model 2 (RQ2)	0.9018	0.9081848
Model 3 (RQ3)	0.0662	0.6821203
Model 4 (RQ4)	0.5818	0.8771848

Table 1. K-fold Cross Validation Output for Each Research Question

We performed K-fold cross validation to test the performance of our models on new, unseen data. Table 1 summarizes the best performing models per question, detailing the Root Mean Squared Error (RMSE) and R-squared values. An optimal RMSE value is a smaller value, which is reflected by the table, suggesting good predictive accuracy. The R squared values are high, close to 1.0, indicating the model's effective ability to explain variability in the data.

However, based on the way the data was collected, the model's reliability in accurately predicting outcomes beyond observed data raises concerns. This is because initial observations revealed non-normality and non-constant variance in most of the models. These statistical violations could impact the model's overall performance in unobserved scenarios, prompting us to consider further research and adjustments to ensure robustness and reliability.

8. Conclusions and Limitations

Conclusion RQ1:
Because the p-value=0.00369 < 0.05, we reject the null hypothesis and conclude that the effects of minutes played and assists per game are significantly different, after accounting for age and field goal percentage. The analysis concludes that MP and AST have different impacts on PTS when controlling for Age and FG%.

Conclusion RQ2:
Because the p-value=1.179e-06 < 0.05, we reject the null hypothesis and conclude that free-throw percentage, 2-pointer percentage, or 3-pointer percentage has a significant linear impact on points per game, after accounting for turnovers. The analysis concludes that shooting efficiency does affect performance, in this case, measured as points per game. This makes sense as shooting percentages are related to scoring points as making more shots means earning more points. This question is limited in two ways. First, we now know that shooting efficiency does affect performance, but this question does not research whether the efficiency of shooting a certain number of points affects performance more than the other. In addition, the final model used a Box-Cox transformation, so we are comparing the predictors' effect on points per game raised to the 0.6969697 power and not points per game directly.

Conclusion RQ3:
The ANOVA output shows an F-statistic of 213 with a corresponding p-value of less than 2.2×10^{-16}. Since the p-value is much smaller than 0.05, we reject the null hypothesis. This provides strong evidence that the effects of three-point percentage and two-point percentage on field goal percentage are significantly different, even after adjusting for minutes played and field goal attempts. The analysis concludes that 3P% and 2P% have different impacts on FG% when controlling for MP and FGA.

Conclusion RQ4:
Because the p-value=2.2e-16 < 0.05, we reject the null hypothesis and conclude that assists per game and total rebounds provide additional explanatory power for points per game beyond what is explained by field goal % and player position. The analysis concludes that AST and TRB have significant impacts on PTS when controlling for FG and Pos. This makes sense since assists are negatively correlated with points, as a player making an assist by definition means that that player did not score that point, and rebounds are positively associated with points as players who secure more rebounds often create additional scoring opportunities for themselves and their team.

9. References

Basketball-Reference.com. (n.d.). *2024-25 Los Angeles Lakers roster and stats*. Sports Reference LLC. Retrieved April 28, 2025, from https://www.basketball-reference.com/teams/LAL/2025.html

Casals, Martí & Martinez, Jose. (2013). *Modelling player performance in basketball through mixed models. International Journal of Performance Analysis in Sports.* 13. 64-82. 10.1080/24748668.2013.11868632.