Forecasting NBA Game Outcomes:

Predicting Spread, Total Points, and Offensive Rebounds Using Statistical Models

Abrah Furbee, Adam Dameron, Danny Zepeda-Martinez, Henoc Codije, Madison Duffy, Vivian Moore

I. Data Information

In order to have an initial dataset to add our outside variables to, we had to first clean our given data. After downloading Nathan Lauga's data from Dr. Mario's website, we decided to not use the Teams dataset. All of the variables in Teams could be found in the other four datasets (Games, Game_details, Players, Ranking). We decided to drop the variables LEAGUE_ID, SEASON_ID, RETURNTOPLAY and CONFERENCE from the Rankings data, as well as renaming STANDINGSDATE to DATE. We deleted these variables because we know we will not need them for further analyses. In the Games dataset, we removed the variables GAME_STATUS_TEXT, TEAM_ID_away and TEAM_ID_home because they were not relevant to the analysis. We also renamed HOME_TEAM_ID to TEAM_ID and GAME_DATE_EST to DATE in order to more easily work with the variables when we built our models. In the Game_details set, we removed the variables COMMENT and TEAM_ABBREVIATION, as well as removing any rows that had missing data. The TEAM_ABBREVIATION was redundant with our other team-level identifying variables, and the COMMENT variable contained notes regarding the game itself which was not relevant to the metrics we aimed to predict. We then narrowed our data selection down to the current 2022-2023 season to ensure we were using the most relevant player and game performance and statistics to best inform our predictions.

When joining together the four datasets, we found it best to create two datasets; one based on the player-level and one based on the game-level. To do this, we inner-merged the Details and Players sets by PLAYER_ID and PLAYER_NAME, and set the rows to all be distinct. We also inner-joined the Ranking and Games datasets by DATE and TEAM_ID. This organization allowed us to best retain all of the information we believed would be useful in

creating our models by representing metrics from two different levels of the sport. During the joining process, we additionally added all of our Home & Away engineered variables, which will be discussed later in this section, to both of the data sets, and created the Matchup_ID variable.

Lastly, we added Left and Right versions of most of the variables in our working dataset. Based on the Matchup_ID, which will be further described in the next paragraph, the "Left" team variables in each game represent the team with the lower valued team ID while the "Right" team variables represent the team with the higher valued team ID. These variables were created and added to the final dataset in order to provide a means to create models in which the two teams themselves were more important than the Home and Away team, specifically. For example, if the Nets and the Bulls score a certain number of total points, we would not expect the location of the game to have an effect on the outcome of this value. In doing this, we hope to explore the possibility that "home court advantage" does not have an effect on the outcome of total points or OREB. The variables that relied on Home and Away team distinction, such as Spread, Predicted Spread, etc., were not converted to the Left & Right representation. Finally, we removed any observations that had no recorded values for the variables. This addition finalized the dataset we utilized going forward when building our models. We did not remove any outliers from our datasets, as they were actually more helpful in our predictions than without.

In order to increase the accuracy of our predictions we created multiple new variables that we added to our final dataset: Matchup_ID, and the TPG, TSG, TOG, PPG, SPG, and OPG variables (see Appendix). Matchup_ID is a four digit value made up of the last two values of the team identification numbers, with the last two digits of the team with the lower value team ID first (the team that becomes the "Left" team in the Left & Right variables) and the last two digits of the team with the higher value team ID after (the team that becomes the "Right" team in the

Left & Right variables). This variable was created to aid in the joining of our data sets to ensure that the final data set could be created accurately and efficiently.

The TPG variables (Left_TPG, Right_TPG, Home_TPG, Away_TPG) are a sum of the total points scored by the team during past games, up to but not including the current game being described. The TSG variables (Left_TSG, Right_TSG, Home_TSG, Away_TSG) and the TOG variables (Left_TOG, Right_TOG, Home_TOG, Away_TOG) similarly describe the sum of the total steals and total offensive rebounds that the team has accomplished in all past games up until the game being described, respectively. These variables provide insight into the performance of the team up to the point of the game for which we are predicting point spread, total points, and offensive rebounds. The past statistics for the season are important in determining future performance, as the teams will likely perform similarly, on average, in a specific game as they have over past games.

The PPG variables (Left_PPG, Right_PPG, Home_PPG, Away_PPG) describe the averages of the total points scored by the team in the past up until the game being described (see Figure 1). The SPG variables (Left_SPG, Right_SPG, Home_SPG, Away_SPG) and OPG variables (Left_OPG, Right_OPG, Home_OPG, Away_OPG) describe the average of the total steals and average of the total offensive rebounds completed by the team in past games up to the game being described, respectively (see Figure 1). These variables were made to better measure average performance of the teams over time. By averaging the TPG, TSG, and TOG variables, any outliers in team performance over the past games played up to the current point are smoothed over, which provides us with a metric that better reflects typical team performance and will increase the accuracy of our predictions.

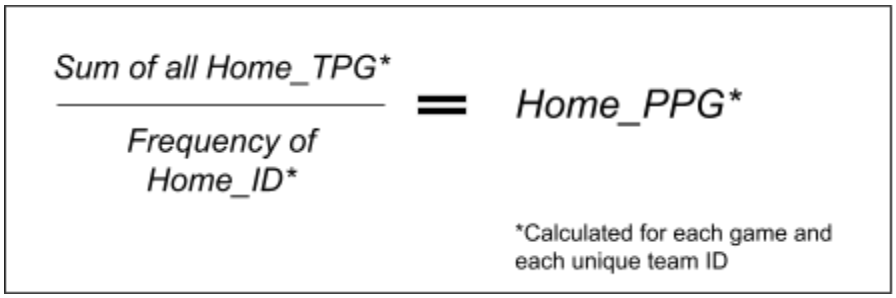

Figure 1: Formula used for all PPG, SPG, and OPG variables.

We chose the outside source of basketball gambling sites because they are attempting to reach the same predicting goals as we are in the project. Sports betting is extremely prevalent and is a growing and innovative industry; so their modeling and predictions are much more in depth and advanced, considering that they are experts in the field. Using their data, which is updated constantly, we hope to get a more accurate prediction.

To supplement the data we used to create our models, we added a predicted spread variable gathered from the Project FiveThirtyEight dataset (Silver). This variable was joined into our final data set by matching the team name to the respective TEAM_ID given in our original datasets. The predicted spread variable describes the point spread predicted using a model created by professionals at FiveThirtyEight. We added this variable in order to better inform our own predictions and compare the values formulated by our models to their predicted spread for the same team match-ups. Additionally, we used the R package nbastatR to gather the Date, Home_ID, Away_ID, Home_pts, Away_pts, Home_oreb, Away_oreb, Home_FGPct, Away_FGPct, Home_3Pct, Away_3Pct, Home_Steals, and Away_Steals variables that we used in our final data set (NBASTATR). These variables feature information regarding performance statistics on a team level, and we included it because it is updated more frequently and will therefore contain values more consistent with current team performances. As a result, our

predictions will likely be more accurate as the models were constructed based on this more recent collection of statistics.

II. Methodology for ***Spread***

Our mission for executing spread started with a handful of plans for the models. Two linear regressions, a K-Nearest Neighbor (or KNN) regression, and a polynomial regression. These would all be tested using either MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) and the best one will be chosen as the final model to run our future predictions on. For the first model, we ran all our initial variables through to see if a linear regression was even worth attempting, as if it came up with a high MAE or RMSE that means that even with additional variables, it would not be a good measure of spread. To do this we performed a forward selection. It ended up giving an overall MAE of 1.612544. This meant that it was a good indicator of spread and we moved on to improving upon it by eliminating any additional variables.

The models after the initial one used the following variables: Home_ID, Away_ID, Spread_Pred, Home_TPG, Away_TPG, Home_TSG, Away_TSG, Home_TOG, Away_TOG, Home_PPG, Away_PPG, Home_SPG, Away_SPG, Home_OPG, Away_OPG, Matchup_ID, Left_ID, Right_ID, Left_TPG, Right_TPG, Left_TSG, Right_TSG, Left_TOG, Right_TOG, Left_PPG, Right_PPG, Left_SPG, Right_SPG, Left_OPG, Right_TSG, Left_TOG, Right_TOG, Left_PPG, Right_PPG, Left_SPG, Right_SPG, Left_OPG, and Right_OPG. Along with this, we split the data into two groups: one that used the Right and Left and all their interactions and the Home and Away and all their interactions. The interactions were done between each of the variables in groups of two, with one for the Home and Away pairs and one for the Left and Right pairs. Ultimately, we found that the interactions had no real impact on the final model chosen.

The variables were chosen because they were either ones that we could have on hand or simply calculate for the teams that were going to play. These were placed into a table with the spread variable and then ran through a stepwise selection linear model. This caused the best output to have an MAE of 5.489101. However, instead of predicting using the variables that we expected, it focused on using the betting data and prediction data from sports blogging sites. To test if this was the best way to go, we reran it with two more model types.

After the linear model, we followed it up with a KNN model to see if we would get substantial results from this new model. It worked better than expected given that spread had a much larger range that is normally used for KNN models, giving us a spread MAE of 10.24442. Being lower than the RMSE of the secondary linear model, it was waived in favor of the aforementioned linear model.

The final model that we checked was a polynomial regression using a degree of 3. When we used a stepwise model selection to check this, we ended up getting a model that also used the betting and prediction data. Because of this we decided to go with the model that used the predictions from the betting and sports blogging data. This had around the same MAE of 5.62863.

Model Type	**MAE**
Linear	5.489101
KNN	10.24442
Poly Regression	5.62863

Table 1: Final Models Mean Absolute Errors

Final Model: lm(formula = Spread ~ 0.24593 + 0.88535*Spread_Pred)

III. Methodology for ***Total***

To model the total points of each game, we explored both fixed effect and mixed effect models. Fixed effect models are models in which it is assumed that each predictor has a set effect on the response; meaning, the variable used to predict will remain constant, and a change in the predictor variable will always affect the response. A mixed effect model is a model that uses these fixed effects, but also takes into account certain effects that have a random effect on the response. In addition, we created mixed and fixed effect models to predict Home Points, Away Points, Left Points, and Right Points individually.

When creating these models, we made sure to only use predictors that we would know before the game began. These included Home_ID, Away_ID, Matchup_ID, and Spread_Pred as well as the season statistics TPG, TSG, TOG, PPG, SPG, OPG, at the values they were before the game began. These variables will be referred to in this section as "Known Variables," as they are known before the game begins. We also split the data into a Home & Away subset, and a Left & Right subset. This ensured that our Home & Away models only included Home & Away variables, and vice versa.

Our initial models predicted Home & Away points using fixed effects. Our first model had Home_PTS as a response, and all the known variables for Home & Away as the predictors. This model had an MAE of 6.084. A similar model was created for Away_PTS, and this model had an MAE of 5.829. At this point, we made our initial predictions for total points by adding up our predictions for Home_PTS and Away_PTS from the respective models. These predictions yielded an MAE of 9.928. We felt this was a good baseline model, as the mean of Total_Points for all games is 229.248, which means this model has an approximate error of 4.33%.

Next, we created a model to predict Total_Points by using all of the Known Variables. However, we would expect this model to have the exact same MAE as the previously fitted models, and this turns out to be true. This is because Total Points = Home_PTS + Away_PTS. As such, adding our sub-models for Home_PTS and Away_PTS achieves the same model as predicting Total Points to begin with. This is still a necessary model to create, as we will achieve different results after implementing some variable selection techniques.

These initial models were effective, but we wanted to determine the significance of the individual variables. To determine this, we performed a "Drop Test." This test is performed on each predictor in the model by excluding that predictor and refitting the model. This new model is then compared to the original model via an ANOVA F-test. A significant P-value in this test ($p < 0.05$) indicates that the variable accounts for a significant amount of variation that cannot be explained by all other predictors.

Performing this Drop Test on our previous models that predicted Home_PTS and Away_PTS separately, the test indicated that both models had only a small number of predictors which were significant. This implies that many of the predictors were likely insignificant, and could lead to an overfitted model. It is for this reason that we elected to perform variable selection. We performed this by iteratively removing variables that decreased the Akaike Information Criterion (AIC) the most. The AIC value for a model is a measure of how effective the model is at predicting, but penalizes the inclusion of a large number of variables. This means that AIC decreases when a predictor is removed that does not have a significant effect on the fit of the model, and increases if a variable is removed that explains a large amount of variability within the model.

The AIC of a model is a good indicator of how well the model will perform for predictions (Chakrabarti, 2011). This is counter intuitive, as one would expect MAE to represent this, but this turns out to not be the case, as a low MAE could be indicative of overfitting. By minimizing AIC, we are less likely to create an overfitted model.

This process of removing one variable at a time to decrease AIC was performed via stepwise selection. At the conclusion of the stepwise selection, each variable in the final model will be significant when performing the drop test. This must be true, as any variable that does not explain significant variability in the final model would have been removed to further reduce the AIC. This process was applied to each of our models for Home_PTS, Away_PTS, and Total Points. Figure 2 provides a visual for which variables became less significant, and which ones ended up in the final models.

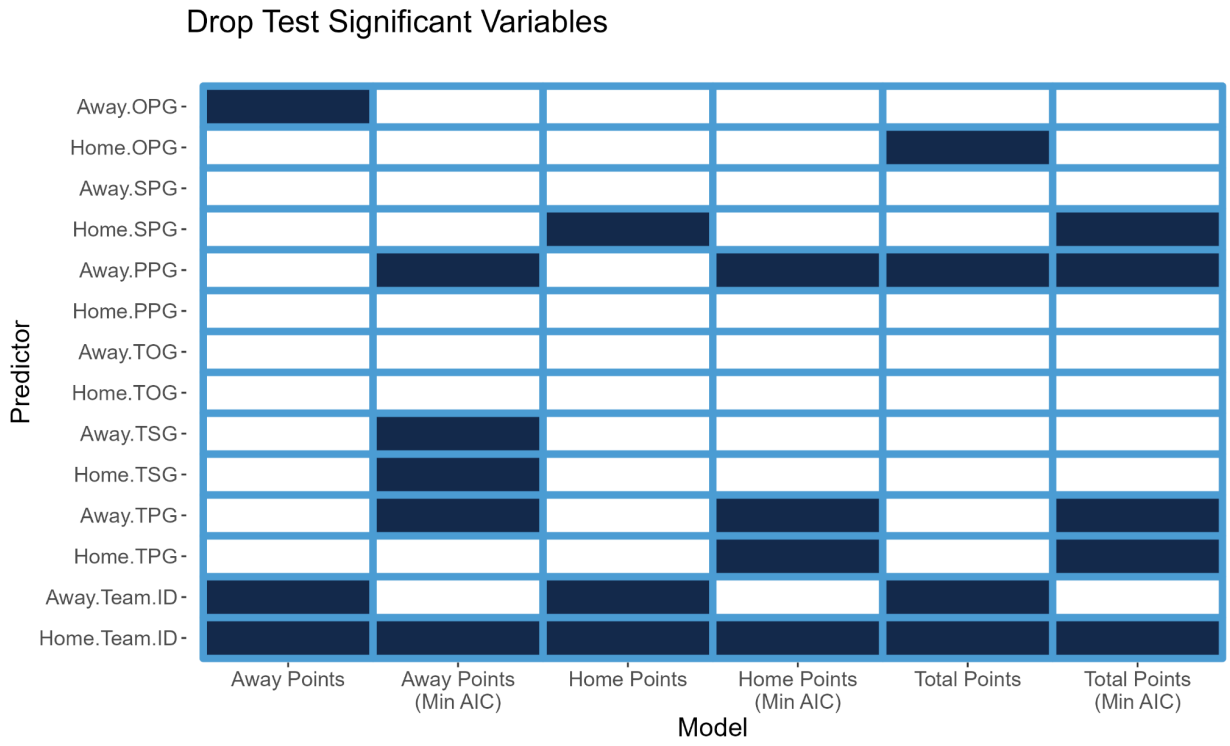

Figure 2: Drop Test of Significant Variables

From Figure 2, one can see that Home_Team had a significant effect in all of our initial models, as well as each model with an optimized AIC. This means that by knowing which team is the one playing from home, one can already explain a significant amount of variability in these models. However, the Home_Team becomes insignificant after creating a model that minimizes AIC. This is interesting because one would expect a similar amount of variability to be explained by both the home and away teams. This turns out to not be the case, and this could be reminiscent of "home court advantage," in which a team tends to perform better when they are the home team. The fact that the home team explains significant variability in both our Home_PTS and Away_PTS models, could imply that home court advantage applies to scoring goals, as well as preventing goals.

Additionally, Away points-per-game (PPG) is significant in the optimized Away_PTS model. This makes sense, as knowing the average points a team scores each game would logically allow one to attempt to predict the amount of points they score. However, one would not expect this to explain significant variability in the amount of home points scored in a game, as is the case in the optimized home points model. Another interesting finding is that the optimal model for predicting Total_Points signifies the Home_SPG (Steals per game) and Away_PPG variables as significant predictors, but does not have Away_SPG or Home_PPG. This is interesting because SPG is a defensive statistic, and PPG is an offensive one. Additionally, TPG is significant for both teams when predicting total points. This is an indicator that the running total of points scored by a team explains significant variability in the total amount of points scored in a game.

After creating the optimal models, each model saw an improvement in MAE, but the best model was our optimized Total_Points model (the model that did not add together Home and

Away Points models) and had an MAE of 14.08. Meaning that on average, our model makes predictions for total points that are within 14.08 points. This is an error of approximately 6.14%.

This process was repeated for Left & Right statistics; however, with these models we had to further modify our data to use combinations of Left & Right statistics to get statistics for the entire game. For example, to estimate the Total_Points scored in the game, we could not use Left_PPG and Right_PPG, as Left & Right teams are essentially random, and rely on the ID assigned to the team. Thus, even if we achieve a coefficient that estimates the effect Left_PPG has on the total points scored, this value is hard to interpret. However, if we add Left_PPG and Right_PPG as a variable, (called Matchup_PPG), the estimated coefficient value for this is easier to interpret. This was then done for each variable. We summed the per-game statistics, and averaged the "running total before game" stats (see Table 2). After creating these statistics, we developed the same form of Total_Points model as before, with total points as a response, and all the known variables as predictors. This model initially had an MAE of 10.23, and a drop test showed that Matchup_ID, Matchup PPG, and Matchup OPG were significant. We also optimized this model by removing predictors to minimize AIC, and achieved a final model with Matchup TPG, Matchup TSG , Matchup PPG, and Matchup OPG as predictors. This model had an

Matchup Statistics

Predictor	Formulation
Matchup TOG	$0.5 \times$ (Left_TOG + Right_TOG)
Matchup TSG	$0.5 \times$ (Left_TSG + Right_TSG)
Matchup TOG	$0.5 \times$ (Left_TOG + Right_TOG)
Matchup PPG	Left_PPG + Right_PPG
Matchup SPG	Left_SPG + Right_SPG
Matchup OPG	Left_OPG + Right_OPG

Table 2: Matchup Statistics

MAE of 14.84, which is higher than the initial model. This should not be a major issue though, as the lower AIC implies that the model is more robust for prediction purposes. Additionally, this model has an MAE very close to our best Home & Away model. Finally, this model has the lowest AIC of all models so far. It is for these reasons that we do not believe the higher MAE is indicative of any major issues.

In order to truly test which model performs best, we need to split the data into two parts, and train the models on one part, and test the models on the second part. We decided to split the data by day. All games that occurred prior to day 125 were our training set, and all games after (up to day 153) were our test set. Table 3 summarizes the results of these tests.

Train and Test Split MAE

Model	MAE On Test Data
Home and Away Separate	23.518
Total Points (Home and Away Data)	14.331
Total Points (Left & Right Data)	16.903

Table 3: Summary of Test and Train Mean Absolute Error

From our testing, we determined that the best fixed effects model was the model that predicted Total_Points and used Home & Away statistics as predictors. This model had an MAE of 15.206 on the test games.

While this model may seem substantial, it runs into the same issue that many fixed effects models have, which is random variability that cannot ever be predicted. As a result of this, these models are prone to have larger errors as they do not account for this randomness. Mixed effects

models seek to help with this. By accounting for random variation, our model can forgo assigning a coefficient to these variations, and instead will account for the randomness.

The first random effect we explored was the day the game was played. This means that we believed that there was random day-to-day variation in total points scored in a game. We wanted to test this, and achieved so by creating mixed effect models with Day as a random effect. To have a solid baseline, we explored random effects in relation to our previous best model which was the Total_Points model that used Home & Away data.

This first model, which explored Day as a random effect, had an MAE of 13.584, which is a significant improvement over previous models. Additionally, this model showed that the day in the season a game is played, when treated as a random effect, has a standard deviation of 11.702. This means that on a given day, the total points randomly fluctuate by roughly ± 11.072 points. After accounting for this random fluctuation, our model had an MAE of 8.399 points.

This is by far our best model for fit, but we still need to ensure it works on test data. After completing a train and test split, this model had a test data MAE of 16.687. This model performs worse than our previous best Total_Points model which had an MAE of 14.331, so we will use the initial, better performing model, to make predictions.

Final Total_Points Model Coefficients

Variable	Coefficient
(Intercept)	182.506590
Spread_Pred	0.472876
Home_TPG	-0.008719
Away_TPG	0.009412
Away_PPG	0.518116
Home_SPG	-1.634571
Home_ID	Categorical Var.

Table 4: Final Total_Points Model Coefficients

These coefficients show us that by far the biggest quantitative variable effect on the total points scored in a game is the Home steals-per-game. This coefficient has a value of -1.63, meaning that the total points scored each game decreases by 1.63 points, on average, for each successful steal by the home team. This is interesting, because it implies that a successful steal does not necessarily lead to scoring a basket. In fact, it implies that stealing the ball prevents both teams from scoring. This could be due to a steal usually leading to a brief pause in the flow of the game to allow the team that stole the ball to get to their basket and "reset" the gameplay. Additionally, the predicted FiveThirtyEight spreads appear to have an effect, as an increase in the predicted spread correlates with an increase in the total points scored.

Of additional interest is the intercept term of 182.5 points. This is interesting because this value is significantly different from the season average of points per game of roughly 229 points. This implies that significant variation in total points is explained by other variables, and that our model does a good job at predicting total points.

Certain Home_IDs had significant effects on the predicted values. Some teams had coefficients as low as -18, meaning that simply due to that team being the home team, the expected value for Total_Points decreased by 18 points. This means that the home team either is significantly worse at shooting than other teams, or this team has a defensive talent which greatly hinders the ability of the other team to score.

IV. Methodology for ***OREB***

The methodology of the OREB variable will closely mirror that of the Total variable. To create an optimal model for OREB, the first step that was taken was to examine the distribution of the dependent variable, then see what kind of models (normal distribution, poisson, etc.) would work best, and if necessary, apply transformations to improve the distribution of the dependent variable. When considering the figure below, the distribution of the dependent variable seems to more closely resemble that of a normal distribution rather than a Poisson distribution (even though the variable is a count variable).

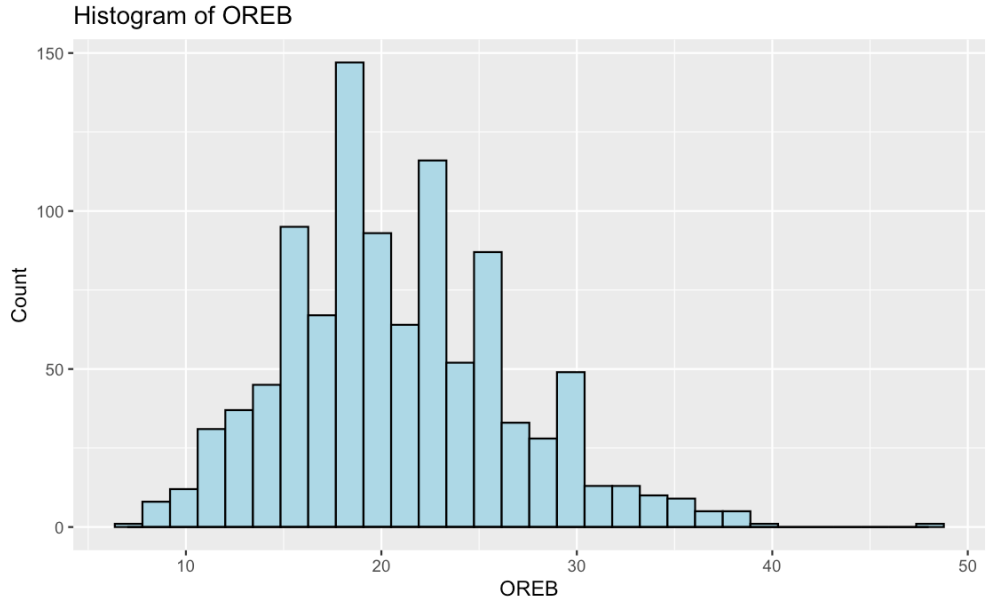

Figure 3: Histogram of OREB

Therefore, we will experiment with models that use the Gaussian distribution as their base. Next, a series of transformations were applied to this variable to see if the OREB variable could more closely approximate a normal distribution. These transformations were taking the natural log of the OREB variable, taking the square root of the OREB variable, and reciprocating the OREB variable. The log transformation didn't change the distribution much (if at all) and the reciprocal made the distribution more right skewed than it initially was. The only transformation that improved the distribution was the square root transformation, which makes sense as a square root transformation corrects for right skewness, as seen in the figure below.

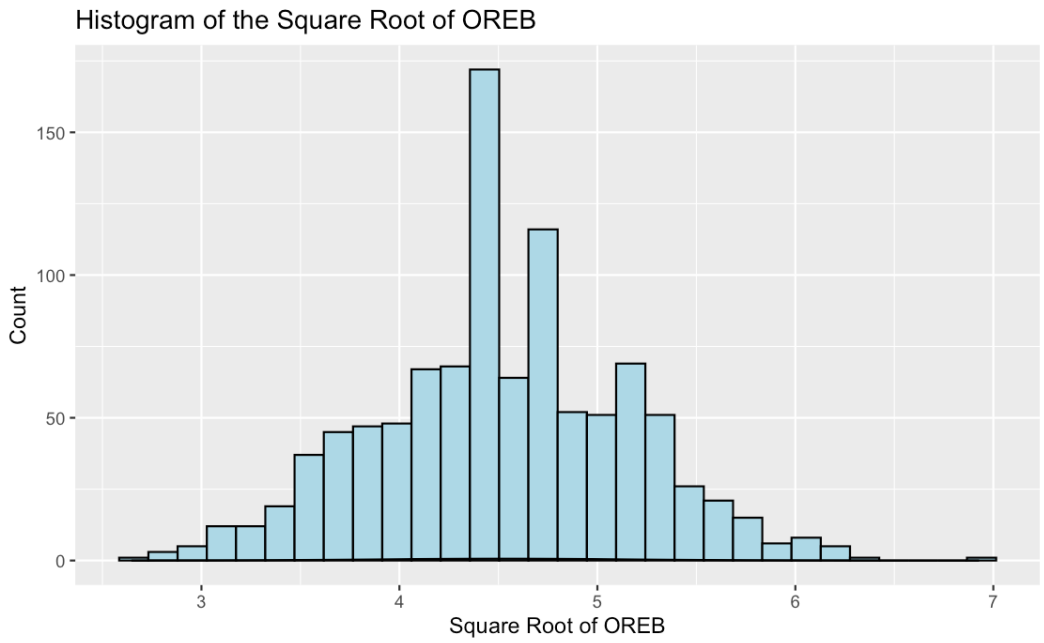

Figure 4: Histogram of the Square Root of OREB

By performing a Shapiro Test, which tests for normality, we found that for a square root transformation, the value was highest at .005, and therefore was closer to normal than without a transformation, which was valued at $3 \cdot 10^{-10}$. Based on the results of this test, the dependent variable for all of the models will be transformed into the square root of OREB (and for MAE, the fitted values will be squared for the calculation).

After determining what kind of distribution the dependent variable was and figuring out what type of models may be best at estimating this variable, the next step was to select all possible independent variables that could be incorporated into the model. Given that we are predicting how many offensive rebounds both teams make in a game, it doesn't make too much sense to incorporate variables that measure this value on a game-by-game basis. This is because we won't know information like the team's free throw percentage for the game or how many points they score in advance. It would be better to use variables that incorporate this information over time, such as the average points that the team scores in a game or the total points they have scored so far in a season (which were written about in the Data Information section of the paper). Including the day variable might also be useful, since teams usually develop better chemistry and improve their sportsmanship as the season goes on.

Another important factor to take into consideration is the concept of home court advantage. Using the variables we have created, we will be able to test for this effect by testing two kinds of different models; one that uses Home & Away variables (where home court advantage is important) and another that uses Left & Right variables (assuming there is no difference where the teams play. The variables for the Home & Away models that will be used are as follows: Home_ID , Away_ID, Day, Home_TPG, Away_TPG, Home_TSG, Away_TSG, Home_TOG, Away_TOG, Home_PPG, Away_PPG, Home_SPG, Away_SPG, Home_OPG, and Away_OPG. The variables for the Left & Right models that will be used are as follows: Matchup_ID, Day, Left_TPG, Right_TPG, Left_TSG, Right_TSG, Left_TOG, Right_TOG, Left_PPG, Right_PPG, Left_SPG, Right_SPG, Left_OPG, Right_SPG. The Home_ID, Away_ID, and Matchup_ID variables were all converted to factor variables (given that there is no inherent hierarchical order to the values of the ID's).

Fully saturated models are a good starting point for any project dealing with linear regression. Based on this notion, we created 4 different types of fully saturated models: one predicting home OREB using the Home & Away set of variables, one predicting away OREB using the Home & Away set of variables, one predicting overall OREB using the Home & Away set of variables, and one predicting overall OREB using the Left & Right set of variables. The results of the regressions that predicted home and away OREB separately were added together to get a resulting Home + Away MAE. The first set of MAE values are listed below:

Model Type	**MAE**
Home + Away	3.945736
Home/Away	3.928675
Left/Right	4.172636

Table 5: Fully Saturated Models Mean Absolute Errors

After doing this, we still felt that our group could do more to create better models (given that the fully saturated models had quite a few variables). A drop1 F-test was performed on each of the fully saturated models, and our suspicion proved to be correct. In all of the models, some coefficients in each of the fully saturated regression models were found to be statistically insignificant under the drop1 F-test. Based on the results of these F-tests, it became clear that some kind of selection technique could prove beneficial to refine the models. Stepwise regression was then performed on all of the fully saturated models. The second set of MAE values for the stepwise regressions are listed below:

Model Type	**MAE**
Home + Away	3.955161
Home/Away	3.935538

Left/Right	4.336732

Table 6: Stepwise Mean Absolute Errors

The selection technique made the MAE values worse. This is most likely a byproduct of the information loss that occurred when certain variables were omitted from the fully saturated regression. There was one final modeling technique we used for the OREB variable, and that was making a mixed effects model (where the Day variable became a random effect). The justification for trying this technique is the same as the justification listed in the section for the total variable. The values for OREB were high some days, and low on others. It is not known why different days can yield different results, indicating that the day variable could potentially be seen as a random effect. This technique was performed on both the full set of independent variables in the fully saturated regressions, and the reduced set of independent variables selected by the stepwise technique. The final set of MAE values are listed below:

Model Type	**MAE**
Home + Away	3.898627
Home/Away	3.931958
Left/Right	4.173443

Table 7: Fully Saturated with Random Effects

Model Type	**MAE**
Home + Away	3.904623
Home/Away	3.93028
Left/Right	4.321122

Table 8: Best Choice Variables (from Stepwise Regression) with Random Effects

After looking at all of the MAE values, the model with the lowest value for OREB was the fully saturated Home + Away mixed effects model. This is a bit surprising, given that fully saturated models are usually not the optimal choice for a regression model. This is probably due to the large information loss that occurs when many of the variables were from the "best model" by selection techniques. While the stepwise techniques did yield lower AIC values and more statistically significant predictors, our group aimed to pick an optimal model based solely on MAE values. Therefore, the best model by MAE results is in fact the fully saturated Home + Away mixed effects model. The formula for calculating the Home OREB is: lmer(sqrt(Home_oreb) ~ Home_ID + Away_ID + Home_TPG + Home_TOG + Away_TOG + Home_PPG + Home_OPG + (1|Day), data = KnownHomeAwayStats). The formula for calculating Away OREB is: lmer(sqrt(Away_oreb)~ Home_ID + Away_ID + Home_TPG + Away_TPG + Home_TSG + Away_TSG + Home_TOG + Away_TOG + Home_PPG + Away_PPG + Home_SPG + Away_SPG + Home_OPG + Away_OPG + (1|Day), data = KnownHomeAwayStats). The predictions from the respective models are squared, and then added together to form a comprehensive prediction for OREB, yielding the final predictions to take the form of predict(Home_oreb, newdata = new)^2 + predict(Away_oreb, newdata = new)^2.

Appendix:

Name of Variable	**Description**	**Type**
TPG	Total points (per) game	Numeric
TSG	Total steals (per) game	Numeric
TOG	Total offensive rebounds	Numeric
PPG	Points per game (Average)	Numeric
SPG	Steals per game (Average)	Numeric
OPG	Offensive rebounds per game (Average)	Numeric

References

"NBASTATR." *RDocumentation*, https://www.rdocumentation.org/packages/nbastatR/versions/0.1.10131.

Siver, Nate. "2022-23 NBA Predictions." *FiveThirtyEight*, 3 Apr. 2023, https://projects.fivethirtyeight.com/2023-nba-predictions/games/?ex_cid=rrpromo.

Chakrabarti, Arijit, and Jayanta K. Ghosh. "AIC, BIC and Recent Advances in Model Selection." *Philosophy of Statistics*, vol. 7, 2011, pp. 583–605., https://doi.org/10.1016/b978-0-444-51862-0.50018-6.