Predicting NBA Team Records Based on Offensive and Defensive Performance

Stat 139 Final Project
John Russell, Max Peng, Praveen Kumar, and Melchior Delloye

Introduction & Motivation

Throughout the history of the National Basketball Association, the regular season has been instrumental in dictating matchups in the playoffs, and often predicting who will win the Larry O'Brien trophy (NBA championship) come June. An intense 82-game regular season slate brings many variables, both expected and unexpected, to the table throughout the course of the year.

Our goal for this project is to analyze different predictors on both the offensive and defensive sides of the floor to create models that estimate the number of wins a given team earns over the course of the season. In doing this, our group hopes that we can determine the most important predictors, both statistically and practically**,** for a team's final win tally before the playoffs begin. In terms of analyzing statistical relationships, we also plan to focus on the effects of purely offensive predictors vs. defensive predictors, and which have more predicting power for how much a team wins.

Hypothesis: We hypothesize that offensive metrics will be more important than defensive metrics, as an increase in 3-point shooting and faster offensive pace have caused games to become higher scoring in recent years. However, we also hypothesize that offensive-only and defensive-only models will each only explain about half of the variation in the response (team performance), while a full model containing both offense and defense will explain nearly all of the variation in the response. Numerically, we expect R^2 values of roughly 0.5 for our offensive-only and defensive-only models (a bit higher than 0.5 for offense and a bit lower for defense), while our full model should have R^2 near 1.

Data & EDA

Dataset Description

Basketball Reference provides detailed statistics and historical data for NBA players, teams, and games across multiple seasons. For this analysis, we collected player and team performance data spanning **8 full NBA seasons**, specifically from the **2014-2015 season to the 2023-2024 season** (2019-20 and 2020-21 seasons were incomplete due to COVID).

Basketball Reference provides detailed historical data and statistics for all NBA players, teams, and games. Because our analysis of team records is done in reference to an 82-game season, we decided to use the 8 most recent *full seasons* for our analysis. By focusing on this period, we ensure that our dataset incorporates robust and comprehensive information across a meaningful timeframe, with consistent sample sizes over time.

Team Statistics

Basketball Reference includes a variety of team-level statistics measuring both offensive and defensive performance, efficiency, and team schedule and records.

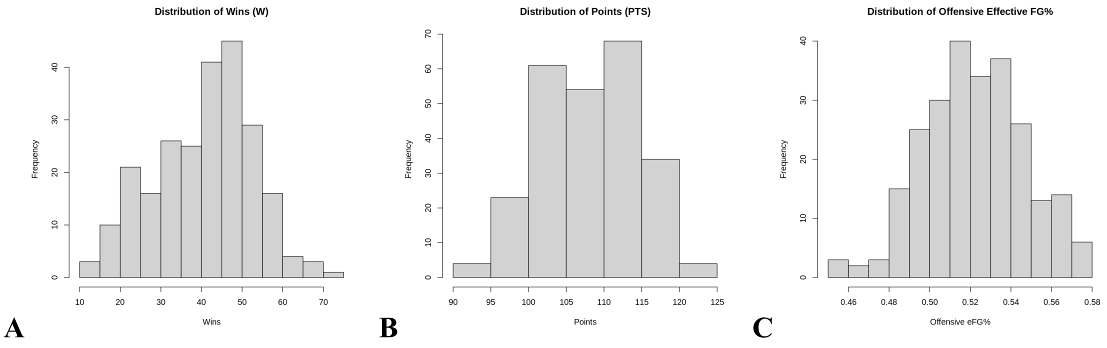
Traditional team statistics include offensive metrics such as points per game (PPG), field goal percentage (FG%), and 3-point percentage (3P%), in addition to defensive metrics such as steals per game (SPG), and blocks per game (BPG). Together, these metrics holistically evaluate a team's performance.

Advanced metrics such as offensive rating (ORtg), defensive rating (DRtg) measure overall performance, while metrics such as effective field goal percentage (EFG%) measure efficiency. Complete predictor lists are referenced in the appendix.

Applied Research Question: Based on the directions of the final project, which state that our goal is not just to build the best predictive model, "but instead focus on the interpretations and relationships in the data set", we want to state a more robust applied research question. As we work on this project, we hope to answer an overarching question: "Is offense or defense more important in the NBA?" By building predictive models and analyzing the significance of different statistical variables (many of which are offensively or defensively slanted), we can come to a reasonable conclusion of which side of the ball is more important for winning games in the national basketball association.

EDA

Missingness: Fortunately, the data for each team across each season is complete in our dataset, so missingness was effectively a nonissue. Using R's is.na()function, we found that the only null values present in the dataset were with regards to League Average rows (e.g., League Average does not have a season record nor an arena).



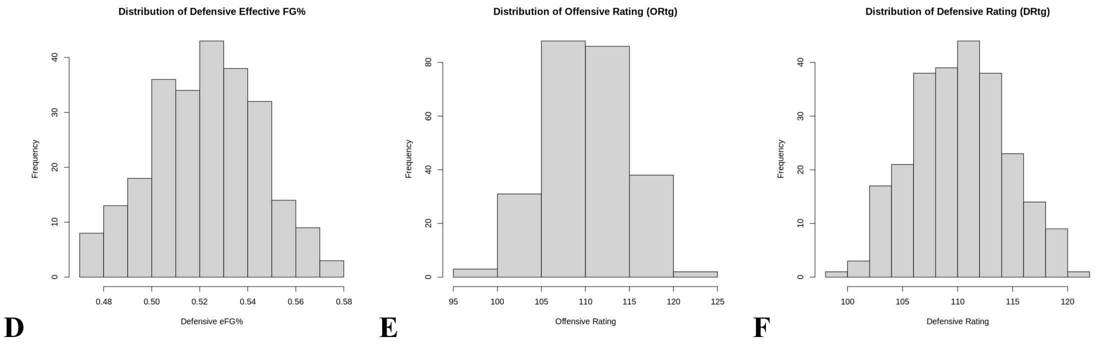


Figure 1: Histograms plotting distributions of Wins (A), Points (B), Offensive EFG% (C), Defensive EFG% (D), Offensive Rating (E), Defensive Rating (F)

Distribution of Variables: Above, we plotted histograms of a handful of our variables. Notice that the center of the Wins histogram is about 40 (which makes sense as 41 out of 82 wins is the mean record), and the distribution is slightly left skewed, meaning that there are a greater number of teams over time with a losing record than a winning record. The distribution of points, effective field goal percentage, offensive rating, and defensive rating are all roughly normal, though offensive effective field goal percentage seem slightly left skewed, which may be related to the same phenomenon as wins.

The box plots below show the relationship between offensive, defensive, and net ratings with the number of wins a team achieved. We binned the ratings into four categories, with ratings in each bin gradually increasing. Higher offensive ratings appeared to be correlated with number of team wins, a positive linear relationship, while lower defensive ratings were correlated with higher team wins, a negative linear relationship.

In terms of statistical significance, due to the overlap between boxes in the offensive and defensive rating plots, we cannot conclude a statistically significant difference between binned groups in these plots. However, when using net rating, we see that the relationship is much stronger and appears statistically significant.

Hence, this suggests that offensive and defensive ratings on their own will not be enough to fully explain the variation in wins. However, including both offensive and defensive predictors should make a model that is statistically viable in predictive power.

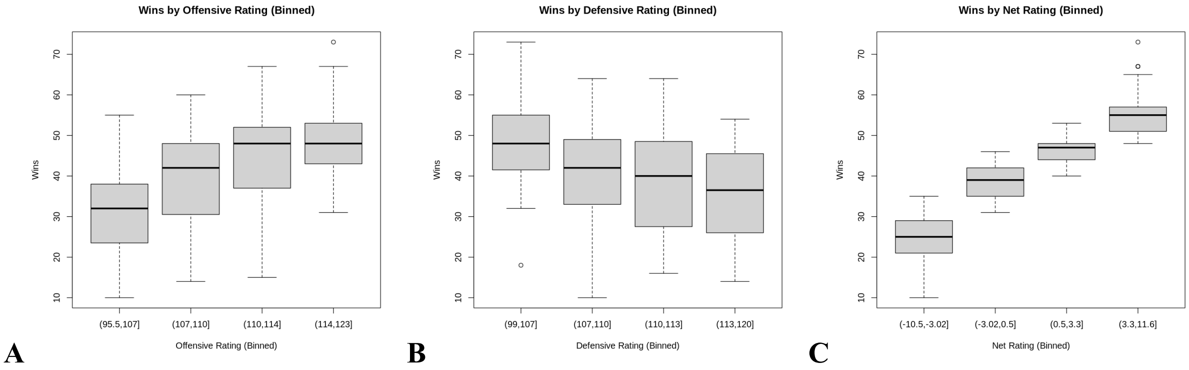


Figure 2: Boxplots of team offensive rating (A), defensive rating (B), and net rating (C), binned into 4 groups per plot.

Naturally, given the linear regression nature of our analysis, we assessed the correlation between predictors that could lead to multicollinearity. The first correlation matrix shows the correlation between offensive predictors. Offensive effective field goal percentage seemed highly correlated to multiple variables such as field goals, offensive rating, and 2 point conversion rate, with correlation coefficients all above 0.89. True shooting accuracy was highly correlated with 2 point conversion, and offensive effective field goal. Both correlation coefficients were above 0.92. The only large negative correlation coefficient was between the number of two point attempts and the number of three point attempts.

The second correlation matrix shows correlation between defensive predictors. The correlation was much lower with only one coefficient being above 0.9. This was Defensive field goal percentage and defensive rating, standing at 0.94. The only other noticeable coefficient was steals and defensive turnover percentage which was 0.82. Ultimately, general correlation was much lower in the defensive predictors than in the offensive predictors. Similarly, correlation was more often negative than in the offensive predictors.

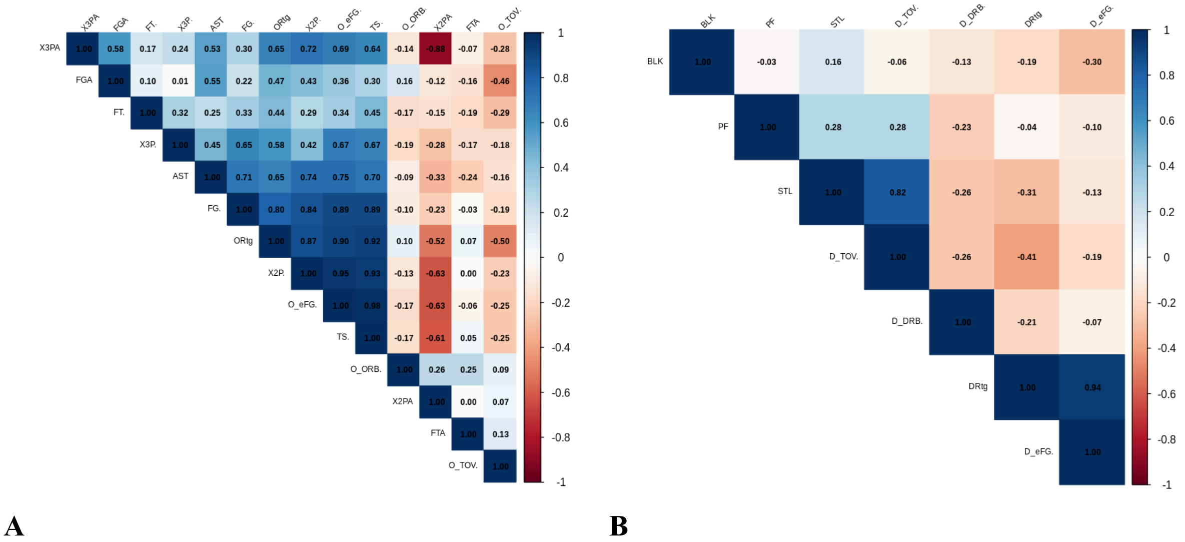


Figure 3: Correlation heatmaps of offensive (A) and defensive (B) predictors

Baseline Model Results

For our baseline model, we wanted to include as few predictors as possible while still holistically measuring offensive and defensive team performance. In order to hypothesize regarding whether offense or defense is more important in predicting team success, we first fitted models using just offensive rating and defensive rating individually.

Table 1: Regression table output of offensive baseline (offensive rating only)

	Variable/Metric	Value	p-value
Coefficients	Intercept	-123.3185	2.76e-14
	ORtg	1.4905	2e-16
Model Fit	Multiple R-squared	0.3293	
	Adjusted R-squared	0.3265	

Table 2: Regression table output of defensive baseline (defensive rating only)

	Variable/Metric	Value	p-value
Coefficients	Intercept	184.3476	2e-16
	DRtg	-1.3004	5.68e-14
Model Fit	Multiple R-squared	0.2116	
	Adjusted R-squared	0.2083	

These simple offensive-only and defensive-only models support our initial hypothesis favoring offense, as the offensive-only model R^2 was notably higher than the defensive model. Still, neither model is particularly effective given the low R^2 values, which makes sense given the overlap of the box plots we noted in EDA. Note that the coefficients match the correlations in the EDA as well.

Next, as a baseline for our full model (including both offensive and defensive), we fit a model with both offensive and defensive rating. Although net rating combines both offensive and defensive rating (as it is calculated as the difference between the two), including both individual predictors in the model allows us to view coefficients and significance for offensive and defensive rating individually.

Table 3: Regression table output of full baseline (offensive rating + defensive rating)

	Variable/Metric	Value	p-value
Coefficients	Intercept	39.92596	4.94e-11
	ORtg	2.44590	2e-16
	DRtg	-2.43641	2e-16
Model Fit	Multiple R-squared	0.9369	
	Adjusted R-squared	0.9363	

We can see that combining offensive and defensive rating significantly improves the model, reaching an R-squared of over 0.93, rather high for just two predictors. Notice that both predictors are statistically significant and almost identical in magnitude, with offensive rating just slightly more important than defensive rating based on coefficient magnitude. From this, we expect that combining offense and defense will significantly outperform using just offense or defense on its own. Additionally, we temper our expectations regarding the difference in importance between offense and defense, as the full model baseline implies that they are roughly equal in importance, with just the slightest edge in favor of offense.

Methods

This section outlines the steps taken to build and evaluate models for predicting team performance.

1) Offense vs. Defense: To assess whether offensive or defensive metrics are more effective at predicting outcomes, we fit separate offensive-only and defensive-only models, containing mutually exclusive sets of statistics generally considered to be offensive and defensive metrics, respectively. We then compare these models on the basis of R^2 (or adjusted R^2), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). These comparisons aim to identify whether offensive or defensive performance better explains team success. Additionally, we hope to gain insight into the most influential features within each model, and to see how these line up with our full model.

To fit the offensive model, the `lm()` function in R was fit to all offensive predictors, as specified in Table A1 of the appendix, with wins as the response. To fit the defensive model, the same was done using all defensive predictors, as specified in Table A2 of the appendix.

2) Full Model: We fit a full linear regression model incorporating both offensive and defensive metrics in order to predict the number of wins using regression analysis. From this, we measure the coefficient magnitude and statistical significance of each predictor to determine its feature importance and relevance in predicting team success. To measure its improvement on the offensive-only and defensive-only models, we again turn to the same metrics of R^2, adjusted R^2, AIC, and BIC, for comparison.

To fit the full model, the `lm()` function in R was fit to all predictors listed in Tables A1 and A2 of the appendix.

Feature selection: For both **1)** and **2)**, models were first fit with all appropriate predictors (offensive, defensive, or both), then tuned to include only significant predictors. This was done in order to mitigate multicollinearity between predictors, and to prevent model overfitting due to excessive complexity or too many predictors. Feature selection for each model was done via stepwise regression in both directions, adding or removing predictors until no significant improvement (via AIC or similar criteria) can be achieved by adding or removing additional predictors. This technique was applied to all applicable models in our analysis.

In R, this was done using the built-in `step` function, with `direction = "both".`

Mixed model: Finally, we built a mixed model effect, where we used the teams as random effects; this level of stratification makes intuitive sense, as although there is modest turnover between teams on a year-by-year basis, in general, some teams tend to perform better than others (good management, drafting skill, free agency signings, etc). From this, we can determine the statistically significant predictors and use the aforementioned metrics of AIC, BIC, and more. Note that as marginal and adjusted R^2 are terms used for fixed effects models, we derive the marginal and conditional R^2 (explanation in results section).

Data Preparation & Cleaning

For each season, we obtained two separate datasets from Basketball Reference: The "Per Game Stats" season summary, and the "Advanced Stats" season summary. See the *Dataset Description* section for further details regarding the metrics included in each.

Python vs. R

Both datasets in Basketball Reference are initially in a string format, which mirrors a CSV, but must be converted to a CSV through a Python or R script. Initially, we attempted to use R, however the string parsing capabilities in R are much more limited than those in Python. R failed to parse the strings due to delimiters and newline formatting, so we instead resorted to Python and the Pandas package.

Loading and Merging the Datasets

First, we loaded our datasets into Pandas DataFrames using the `io.StringIO` function. The `pd.read_csv` function was then used to parse and clean the data. We encountered no missingness otherwise.

The two datasets were then merged into a single DataFrame using the `pd.merge` function, on the common column "Team." To prevent redundant columns, we used a left join, and removed the suffix labels automatically appended by Pandas when merging datasets.

This process was repeated for each year's dataset. A column for the season years was added to distinguish team data across time, and then all seasons in our dataset were merged together via a full join.

The final resulting merged dataset was saved as "nba_stats.csv." The final dataset was now well-structured, accurate, and ready for subsequent statistical analysis.

Results

Defensive Model

Table 4: Regression table output of defensive model, pre-feature selection

Category	Variable/Metric	Value	p-value
Coefficients	Intercept	79.7202	0.401281
	DRtg	-1.1926	0.272031
	D_DRB%	1.1865	0.093344
	STL	5.7649	0.000302
	BLK	1.9910	0.056625
	D_eFG%	-1.4792	0.993625
	D_TOV%	-3.9947	0.022477
Model Fit	Multiple R-squared	0.3163	
	Adjusted R-squared	0.2987	
	AIC	1801.677	
	BIC	1829.522	

This defensive model included the variables in the defensive predictors table in the appendix. We intentionally left out certain predictors, such as DRB, because of their potential multicollinearity with other variables measuring similar things, such as DBR%. Overall, this model is weak under the current selection methodology, simply fitting a full model with these defensive predictors.

To address potential multicollinearity seen above, we used stepwise regression, utilizing sequential variable selection in the forward and backwards directions. This process would be iterated until the AIC is minimized and can no longer be improved, returning the optimal model. This returned statistically significant coefficients and results of the model are included below.

The R^2 and the Adjusted R^2 are both relatively low, hovering right around .316 and .301 respectively. The AIC and the BIC are 1799 and 1824 respectively. This demonstrates that a model, based solely off of defensive predictors, likely isn't an amazing set of predictors for the number of wins a team will get in the regular season. As defensive effort and planning have received less emphasis in the past few seasons, it isn't all-too-surprising that defensive statistics aren't entirely accurate predictors of a team's regular season success.

Table 5: Regression table output of defensive model, post-feature selection

	Variable/Metric	Value	p-value
Coefficients	Intercept	39.92596	4.94e-11
	DRtg	-1.2011	1.34e-09
	D_DRB%	1.1820	.005872
	STL	5.7637	.000276
	BLK	1.9932	.047144
	D_TOV%	-4.0041	.001912
Model Fit	Multiple R-squared	0.3163	
	Adjusted R-squared	0.3017	
	AIC	1799.677	
	BIC	1824.042	

Offensive Model

As our baseline model indicated, offensive rating alone is not a sufficient predictor. Below, we analysed a more complex model that included all offensive predictors.

Table 6: Regression table output of offensive model, pre-feature selection

Category	Variable/Metric	Value	p-value
Coefficients	Intercept	143.5316	0.009539
	ORtg	-9.1227	0.000101
	O_ORB%	6.4182	1.41e-06
	AST	0.1651	0.687962
	X3P%	240.0940	0.204211
	X3PA	15.2662	0.179228
	FT%	105.2135	0.392586
	FTA	1.8026	0.225482
	X2P%	-172.9086	0.445337
	X2PA	13.6373	0.227728
	FG%	669.9189	0.420068
	FGA	-15.8771	0.160716
	O_eFG%	641.8394	0.566262
	TS%	620.8780	0.556103
	O_TOV%	-17.0249	4.14e-07
Model Fit	Multiple R-squared	0.5916	
	Adjusted R-squared	0.5662	
	AIC	1693.989	
	BIC	1749.679	

The major issue with the model considering all offensive predictors is that most coefficients have p values above 0.05, indicating that they are not statistically significant. Hence, we can infer that many of these inputs are not significantly contributing to the model. This confirms our findings from EDA where we saw high correlation between offensive predictors. We can conclude that there is high multicollinearity in predictors preventing us from making significant conclusions about the effect of predictors on team wins.

Using a similar stepwise selection approach as above, to account for this high observed multicollinearity, we get the following model and coefficient output:

Table 7: Regression table output of offensive model, post-feature selection

	Variable/Metric	**Value**	**p-value**
Coefficients	Intercept	139.0561	0.007197
	O_rtg	-8.9340	0.000108
	O_ORB%	6.2527	1.67e-06
	X3P%	313.9528	5.33e-09
	X3PA	0.9242	0.004322
	FTA	0.5067	0.133392
	FG%	0.133392	0.051264
	FGA	-1.6289	9.48e-07
	TS%	1425.6738	0.001603
	O_TOV%	-16.5742	5.35e-07
Model Fit	Multiple R-squared	0.5864	
	Adjusted R-squared	0.5702	
	AIC	1687.047	
	BIC	1725.334	

Full Model

As discussed in the method section, we fit a full linear regression model incorporating both defensive and offensive metrics to predict wins.

Table 8: Regression table output of full model, pre-feature selection

Category	Variable/Metric	Value	p-value
Coefficients	Intercept	77.96433	0.1052
	ORtg	0.86761	0.3946
	O_ORB	0.74930	0.1920
	AST	-0.17575	0.2798
	X3P	49.87529	0.4899
	X3PA	9.48766	0.0286
	FT	14.04523	0.7673
	FTA	0.12996	0.8194
	X2P	-39.85156	0.6503
	X2PA	8.90928	0.0389
	FG	279.04870	0.3883
	FGA	-9.08378	0.0356
	O_eFG	-123.05353	0.7759
	TS	174.87086	0.6714
	O_TOV	-2.00834	0.1745
	DRtg	-2.42207	3.53e-07
	D_DRB	-0.27179	0.3291
	BLK	-0.79107	0.0215
	STL	0.90198	0.1040
	D_eFG	-0.91488	0.9909
	D_TOV	-0.68979	0.3357
	PF	-0.09336	0.6960
Model Fit	Multiple R-squared	0.9433	
	Adjusted R-squared	0.9378	
	AIC	1234.105	
	BIC	1314.16	

With this model we do see strong performance, with an adjusted R^2 value of 0.93: this significant increase from the offense-only and defense-only models quells the notion that solely one side of the ball matters in the modern NBA. Similarly to the models above, many of these predictors, however, are not statistically significant. Thus, we attempt stepwise regression once more, giving us the following model and coefficient output.

Table 9: Regression table output of full model, post-feature selection

	Variable	Value	P-value
Coefficients	(Intercept)	104.69031	5.27E-07
	O_ORB%	1.23513	2.00E-16
	X3P%	36.8175	0.04474
	X3PA	8.4271	0.04097
	X2PA	8.35459	0.04288
	FGA	-8.39887	0.04189
	TS%	414.77634	2.00E-16
	O_TOV%	-3.37506	2.00E-16
	DRtg	-2.40597	2.00E-16
	D_DRB%	-0.28631	0.03526
	BLK	-0.80858	0.00861
	STL	0.86203	0.1022
	D_TOV%	-0.69777	0.08679
Model Fit	Multiple R-squared	0.9424	
	Adjusted R-squared	0.9394	
	AIC	1219.82	
	BIC	1268.548	

As expected, while our multiple R^2 decreases slightly, truncating the number of predictors we use leads to the adjusted R^2 increasing slightly.

Mixed Model

Looking at the results of the mixed model (trained to output p-values as well), we see that the random effects are not demonstrably large at first glance, with a variance of 0.6 and standard deviation of 0.78. In the table below the three highest and lowest random effect teams are listed.

Table 10: Mixed model random effects by team

Team	Value
Utah Jazz	-0.921
Minnesota Timberwolves	-0.678
New Orlean Pelicans	-0.587
Milwaukee Bucks	0.571
Portland Trail Blazers	0.785
Memphis Grizzlies	1.026

In terms of significant predictors, the model indicates that there are five, with the three statistically significant offensive predictors corresponding to the three different types of shooting attempts around the court (save free throws), alongside defensive rating and blocks.

Furthermore, the marginal and conditional R^2 values are shown below. The conditional R^2 value being around 0.01 higher indicates that the random effects do have some explanatory power in the proportion of variance in wins.

Table 11: Regression table output of mixed model

	Variable/Metric	**Value**	**p-value**
Coefficients	X3PA	9.795	0.0226
	X2PA	9.222	0.0309
	FGA	-9.387	0.0283
	Drtg	-2.534	7.8e-08
	Blocks	-0.764	0.0271
Model Fit	Marginal R-squared	0.937	
	Conditional R-squared	0.941	
	AIC	1187.009	
	BIC	1270.545	

Conclusion & Discussion

After exploring multiple models throughout the course of this project, let us now revisit our applied research question, initial hypothesis, and bridge them together with our final conclusions. To rehash, in the beginning of the paper we stated how we wanted to determine the most important predictors for the amount of wins per team, and also gauge the relationship between offensive predictors and defensive predictors when it comes to winning.

Our initial hypothesis was that offensive metrics would be more important than defensive metrics, in large part from the evolution of the modern NBA in the 3-point era. Looking at our models, that does somewhat seem to hold true. Our offense-only model achieved an R^2 value of 0.59, nearly twice that of our defense-only model (which had an R^2 value of 0.31). This lends credence to the notion that offensive metrics may explain a greater amount of variance than defensive metrics overall.

Here's a paradoxical result. It is important to note that the single predictor of defensive rating was one of our most statistically significant predictors; it was statistically significant in the defense only model, the full model (both before and after stepwise selection), and the mixed model, often having a p-value thousands of times smaller than the other statistically significant predictors. After some discussion, we interpret this surprising finding as the possibility that while the cumulative predictive power of all the defensive metrics is on the lower side, the individual predictive power of "Defensive Rating" is quite high because of the lack of defensive metrics. For example, when examining the offensive predictors in the appendix, note how shots are split between free throws, three pointers, and two pointers. Such detail is not widely available for defense, and so therefore a catch-all metric like "Defensive Rating" holds such proportionally high statistical significance.

Other important predictors for a team's win (with a focus on the full and mixed models) included X3PA and True shooting. Similar to defensive rating, true shooting holds high statistical significance because it encompasses both three-point accuracy, two-point accuracy, and free throw accuracy. The statistical significance of a predictor like X3PA also follows our initial hypothesis, which is that post-2016, the league has increasingly evolved into a three point shooting contest. Because three is 50% more valuable than two, it makes sense that that amount of three pointers attempted would be important.

Going back to our models' explanatory power, we initially hypothesized that our full models would reach an R^2 value of nearly 1. While our full model did well, at ~0.94, even incorporating the teams as random effects in the mixed model did not improve that accuracy by much. This means that beyond offensive, defense, and each individual team's competency, there is still variance in how many wins a team achieves.

In conclusion, our analysis does support the notion that offensive metrics are slightly more impactful than defensive metrics, but we highlight the importance of both when it comes to predicting team success.

Limitations & Future Steps

While our models did effectively answer our research questions and fulfill the purposes of our analysis, there remain a few limitations we are aware of.

Dataset limitations: Due to the COVID-19 pandemic, the 2019-2020 and 2020-2021 seasons were shortened. As a result of their smaller sample sizes, we opted to omit these seasons from our analysis. However, this does create a time discontinuity which could affect our results, particularly if we decided to look at trends over time in further analysis. Finding a way to either adjust for sample size, such as simulating, bootstrapping or imputing to expand these seasons into full samples of 82 games, would be ideal.

Predictors imbalance: Additionally, while our hypothesis that offensive modeling would outperform defensive modeling was supported, there is a notable limitation to mention: we had several more offensive predictors than defensive predictors, giving offense an inherent edge in explanatory power. Ideally, we would have a more equal number of metrics, somewhat symmetrical, for both models.

Multicollinearity: Per our EDA and modeling, it is evident that multicollinearity among predictors is present. In order to address this, we used stepwise regression to maximize significance of predictors and to limit multicollinearity. However, this comes at the tradeoff of being unable to include all predictors originally intended. Ideally, we would test all predictors for significance in various scenarios, rather than simply exclude some altogether in our final models. However, this would require various models, and exceeds the scope of this paper.

Future Steps: In addition to addressing the limitations mentioned above, our future steps would likely include addressing trends over time, to see if offense has increased in importance or defense decreased, as we hypothesize given trends in the NBA over time. Additionally, if we wanted to adopt the goal of tuning the best predictive model, there are steps we could take for that, too. Factoring in variables which we lack, such as non-offensive or non-defensive variables like strength of schedule and injuries, may give our full model greater explanatory power.

Appendix

Table A1: List of offensive predictors

Predictor	**Interpretation**
ORtg	Offensive Rating
ORB%	Offensive Rebounds Percentage
AST	Assists
3P%	3 Point Percentage
3PA	3 Point Attempts
FT%	Free Throw Percentages
FTA	Free Throw Attempts
2P%	2 Point Percentages
2PA	2 Point Attempts
O_eFG	Offensive Effective Field Goals
FG	Field Goals
FGA	Field Goal Attempts
TS%	True Shooting Percentage
O_TOV%	Offensive Turnover Percentage

Table A2: List of defensive predictors

Predictor	Interpretation
DRtg	Defensive Rating
DRB%	Defensive Rebound Percentage
BLK	Blocks
STL	Steals
D_ eFG%	Defensive Effective Field Goals
D_ TOV%	Defensive Turnover Percentage