Evaluating NBA On-Court Efficiency

Predicting Per Possession Team Performance

Scott Herman

December, 1st 2016

Predict-456: Sports Performance Analytics

Northwestern University Masters in Science, Predictive Analytics

Introduction

The 2016-2017 NBA season is only 15 games into the 82 game regular season and competition among teams is at an all-time high. In order to find a competitive advantage, front-office executives across the league increasingly rely on advanced team and player performance data to gain an edge over their opponents. This information enables decision makers within each organization to support and validate their day-to-day team and player management strategies.

Thinking in terms of possessions has been claimed as the best way to evaluate the actual strength of a team's on-court performance capability (Glockner, 2016). Any team that is able to earn more possession than their opponent, gains the advantage of creating more opportunities to score than the given opponent. This concept was first introduced in the 1950's by legendary University of North Carolina Basketball Coach, Dean Smith, who emphasized both offensive and defensive rebounding as a crucial way to gain more possessions (Oliver, 2004). This idea has evolved over the years and Dean Oliver has more recently found that the game of basketball is decided by four critical factors measured across a team's offensive and defensive performances. These critical four factors are:

1. **Shooting percentage from the field.**
2. **Getting Offensive Rebounds.**
3. **Committing turnovers.**
4. **Getting to the foul line a lot and making the shots.**

The purpose of this analysis is to measure how well these 'Four Factor' performance statistics translate to a team's output of points scored and allowed. The results from this analysis will help evaluate the variables with the strongest explanatory ability in predicting a team's ability to successfully maximize points scored on offense, while limiting the amount of points given up to their opponent. The performance measures utilized within this analysis have been calculated on a per possession basis in an attempt to simplify the events within the game and promote a more unified level of measurement for team comparison.

Data and Metrics:

The data collected for this analysis consists of regular season NBA team performance records observed by all 30 teams over the course of four regular 82 game seasons spanning from 2012-2013 through the 2015-2016 season. The definitions for each variable analyzed within this report are given below, in Table 1.

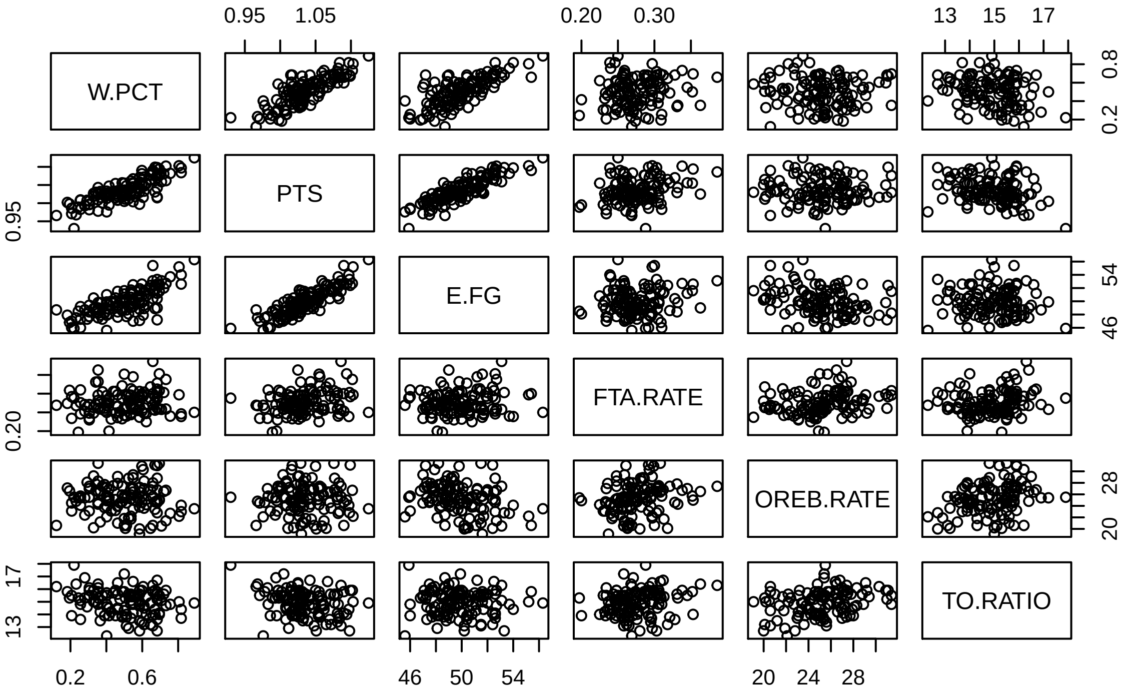
Table 1: Variable Definitions

Variable Name	Data Label	Description
Win Percentage	W.PCT	Team Winning Percentage
Effective FG Rate	E.FG	Team FG Percentage adjusted to make made 3-pointers 1.5 times more valuable than a 2-point shot.
Free Throw Attempt Rate	FTA.RATE	The number of free throws a team shoots in comparison to the number of shots a team shoots from the field (FTA/FGA).
Offensive Rebound Rate	OREB.RATE	The percentage of offensive rebounds a team obtains while on the court.
Turnover Ratio	TO.RATIO	The number of turnovers a team averages per each single possession.
Opponent Effective FG Rate	OPP.E.FG	The Effective FG percentage a team's defense forces their opponent to shoot.
Opponent Free Throw Attempt Rate	OPP.FTA.RATE	The number of free throws an opposing team shoots in comparison to the number of shots that the team shoots.
Opponent Turnover Ratio	OPP.TO.RATIO	The number of turnovers the opposing team averages per each single possession.
Opponent Offensive Rebound Rate	OPP.OREB.RATE	The opponent's percentage of offensive rebounds obtained while on the court.
Team Points Scored	PTS	The number of points a team scores per each single possesion.
Opponent Points Scored	OPP.PTS	The number of points an opposing team scores per each single possession.

Exploratory Data Analysis

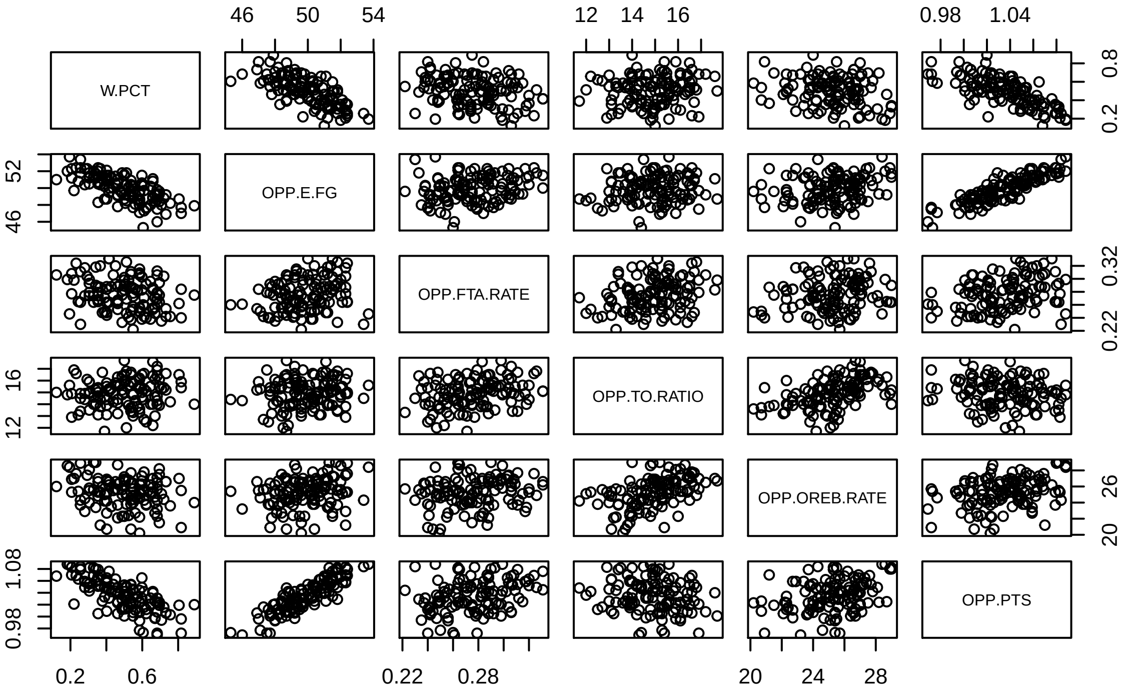
Our data set consists of a total of 120 observations across 12 measurement variables recording each team's offensive and defensive per possession performance statistics. This includes nine variables for prediction along with two response variables to measure against. The response variables are points scored and opponent points allowed. In performing this analysis, linear regression assumptions will be validated to promote reliability statistically accuracy within our results. We begin by visualizing the current relationships that exist bewteen each of the variables contained within our dataset in an attempt to uncover any existing relationships among our predictor set of variables, as well as any variables possessing explanatory power in predicting our response. The scatterplots, below, illustrate the existing correlations among each variable within our data set. The first matrix given, showcases the relationships among the offensive measurment variables, while the second gives the correlations among the opponent performance records.

Figure 1: Scatterplot Matrix of Team Offensive Measures



In examining the above scatterplot matrix, there are a number of strong correlations that immediatley stand out. There appears to be a direct positive linear correlation between W.PCT, PTS, and E.FG, which makes sense intuitively as a team must shoot effectively to score, and further put more points on the board than their opponent in order to attain a wining result.

Figure 2: Scatterplot Matrix of Allowed Opponent Measures

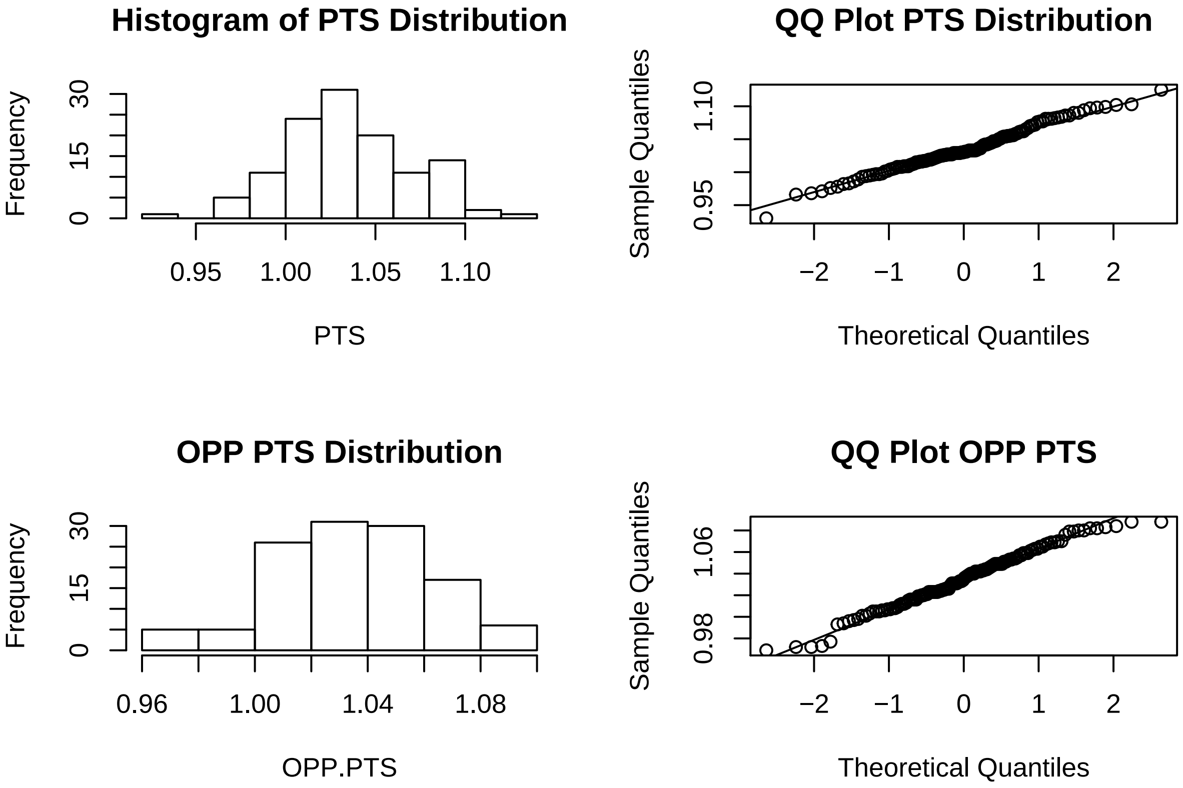


On the other side of the spectrum, it appears that a similar positive relationship exists between OPP.E.FG AND OPP.PTS. OPP.E.FG and OPP.PTS, also seem to have the opposite effect on team W.PCT, as these scatterplots resemble an equally negative correlation. Again, this aligns with the assumption that the more points scored per possesion by the opponent would ultimately limit a team's ability to win.

In analyzing the plotted relationships among the remaining variables within our dataset, there does not seem to be any obvious correlations present. These correlations will be need to be further examined prior to the model development phase in order to further understand and identify where significant correlations occur. Before we move onto testing these correlations, we must first test that the variables within this data set pass the assumptions of ordinal least squares regression.

The first assumption that must be met when utilizing regresison methods is confirmng that each of our response variables possess a normal distribution. These distributions are given by the histograms and QQ Plots in Figure 3, below. Additionally, we have tested this assumption among our set of predictor variables, and it appears that each variable within our data set assumes a normal distribution. For reference, these visuals are given in the Appendix.

Figure 3: Response Variable Distributions



Variable Selection

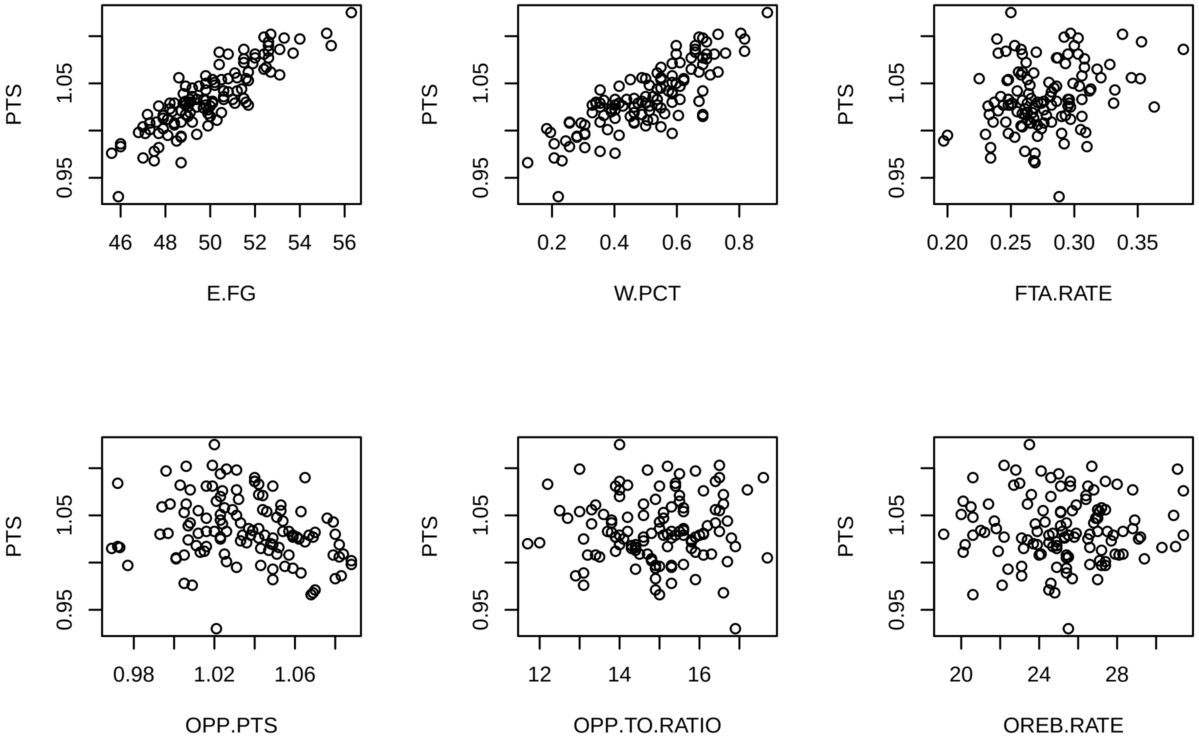
Next, we will utilize Pearson's correlation test to understand the explanatory power our set of variables has on the response outcome. The correlation value can take values between -1 and 1, where -1 indicates a perfect negative correlation and, 0 indicates no correlation, and 1 indicates a perfect positive correlation (Hoffman, 2004). The results showing the correlation between PTS and our set of predictor variables are given below in Table 2.

Table 2: Pearson's Correlation to Offensive Points Scored

Rank	Predictor Variable	Correlation to PTS	t statistic	p-value
1	E.FG	0.863613034	15.432	< 2.2e-16
2	W.PCT	0.817723912	18.608	< 2.2e-16
3	FTA.RATE	0.271380094	3.0629	0.002716
4	OPP.TO.RATIO	0.017792351	0.1933	0.8471
5	OREB.RATE	0.007790449	0.084629	0.9327
6	OPP.OREB.RATE	-0.095672970	-1.0041	0.2986
7	OPP.FTA.RATE	-0.109480445	-1.1965	0.2339
8	OPP.PTS	-0.242464643	-2.7149	0.007626
9	TO.RATIO	-0.275075221	-3.108	0.002361
10	OPP.E.FG	-0.298539161	-3.3979	0.0009264

The correlation results displayed in the table above indicate that there are three predictor variables that have a statistically significant and positive relationship to team PTS. These variables are E.FG, W.PCT, and FTA.RATE. In addition, there also appears to be three statistically significant variables that indicate a negative linear relationship to PTS. The scatterplots represented in Figure 4 showcases each of these variables' correlation to PTS.

Figure 3: Scatterplots of Statistically Significant Predictor Variables to Points Scored



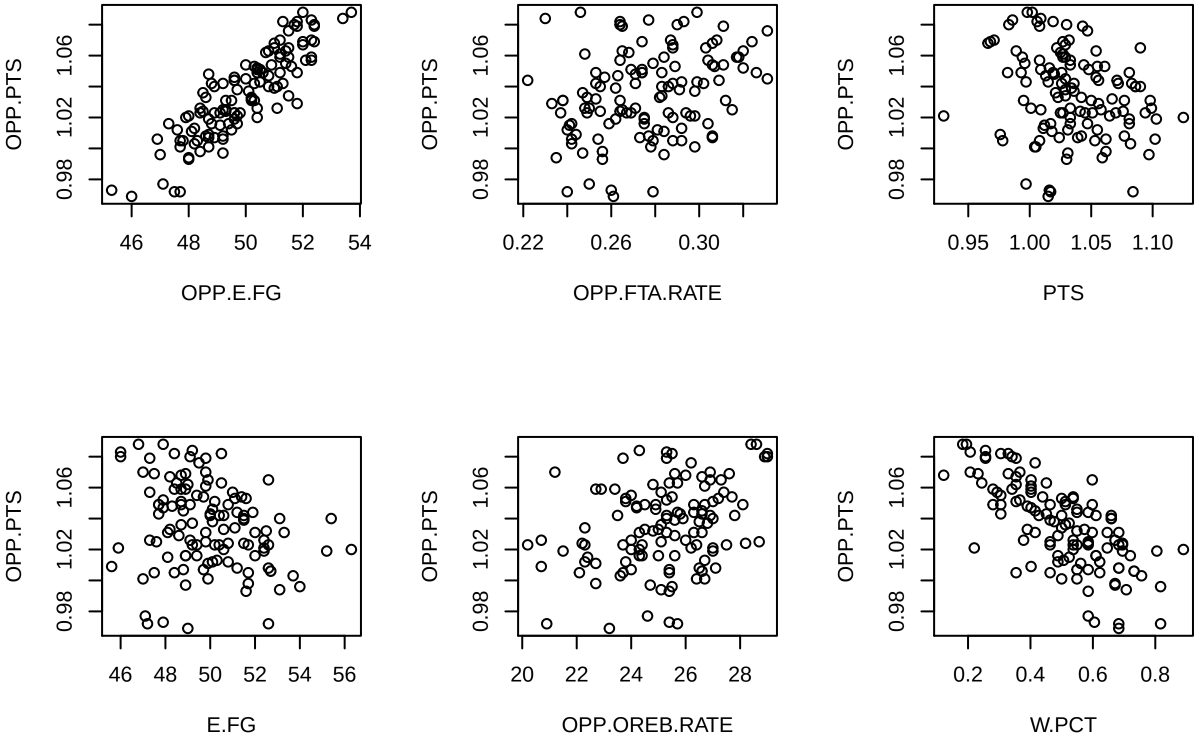
Next, the correlations between variables will be evaluated against Team Opponent Points per Possession. These results are listed in Table 4 and also indicate that there are six total different variables within our data set showing statistically significant correlations to OPP.PTS.

Table 3: Pearson's Correlation to Opponent Points Allowed

Rank	Predictor Variable	Correlation to OPP.PTS	t statistic	p-value
1	OPP.E.FG	0.8749353706	19.627	< 2.2e-16
2	OPP.FTA.RATE	0.3076208388	3.5119	0.0006313
3	FTA.RATE	0.0724149881	0.7887	0.4319
4	TO.RATIO	0.0224393459	0.24382	0.8078
5	OPP.OREB.RATE	0.3254273968	3.7385	0.000287
6	OPP.TO.RATIO	-0.1674007230	-1.8445	0.06762
7	PTS	-0.2424646433	-2.7149	0.007626
8	E.FG	-0.2523294897	-2.8327	0.00543
9	OREB.RATE	-0.0002195301	-0.0023847	0.9981
10	W.PCT	-0.7123284553	-11.025	< 2.2e-16

The statistically significant variable correlations to OPP.PTS are shown visually in the scatterplots given in Figure 5. The two variables indicating the strongest positive correlations of significance are OPP.E.FG and OPP.FTA.RATE, while PTS, E.FG, OPP.OREB.RATE, and W.PCT each display correlations that are negatively related to OPP.PTS.

Figure 4: Scatterplots of Statistically Significant Predictor Variables to Opponent Points Allowed



Model Development and Results

A number of linear regression models were developed for predicting a team's offensive points scored, as well as a separate model predicting opponent points allowed. A number of variable selection procedures were utilized in attempt to identify the best set of variables that yielded results with the highest level of accuracy. In addition to comparing the R-Squared values for each model, the summary results were also evaluated by comparing the goodness-of-fit statistics to understand how well the model fits the actual results observed in our data set.

Ultimately, in this step of the process our goal is to identify the model with the highest level of predictive accuracy. However, interpretability and simplicity were equally important in selecting the 'best' model as this will allow our selected model to be easily replicated by other areas of the organization when attempting to predict future outcomes. We will discuss the results of our selected models below, but have documented the summary statistics for each model developed throughout this process in the Appendix.

Points per Possession Model

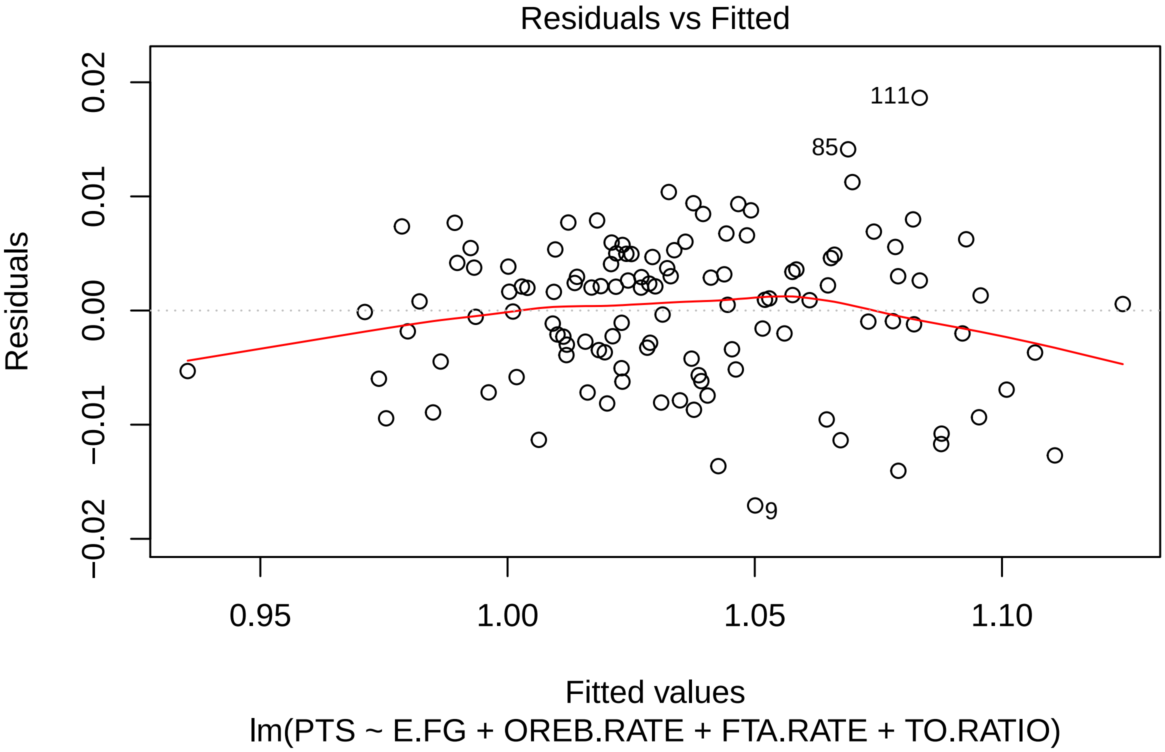
Predicted.PTS $= 0.27339 + E.FG * 0.01589 + OREB.RATE * 0.00485 + FTA.RATE * 0.15654 - TO.RATIO * 0.01319$

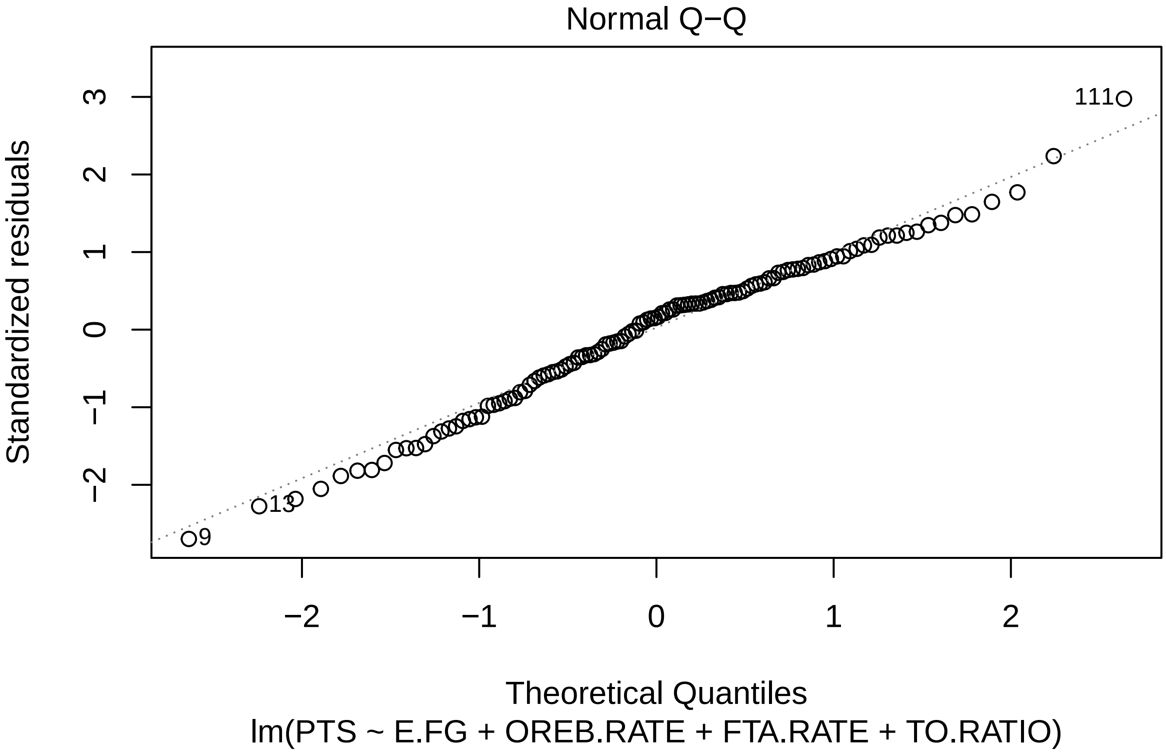
In reviewing the model coefficients for each variable, the formula makes sense logically. The first three variables, E.FG, OREB.RATE, and FTA.RATE, all show a positive impact to the team's points scored, while TO.RATIO indicates a negative relationship. This initially observation does not bring any surprises, however, what is interesting is the weight of the coefficient for FTA.RATE. This suggests that that a team's ability to get to the free-throw line, and shoot well when they get there, holds more importance to offensive scoring production than any of the other three variables. We can further validate this when reviewing the model summary statistics shown in the table below, which indicates that each of our four predictor variables show statistically significant p-values in relation to points scored.

Table 4: Fitting linear model: PTS ~ E.FG + OREB.RATE + FTA.RATE + TO.RATIO

	Estimate	Std. Error	t value	Pr(>|t|)
E.FG	0.01589	0.0003062	51.89	1.241e-81
OREB.RATE	0.004855	0.0002505	19.38	4.944e-38
FTA.RATE	0.1565	0.01957	8	1.091e-12
TO.RATIO	-0.01319	0.0006112	-21.58	2.986e-42
(Intercept)	0.2734	0.0186	14.7	3.64e-28

Additionally, these results reveal an R-Squared value of 0.9666, which inidcates that this model is able to explain 96.66% of the variation that occurs in predicting points per possession. This initially indicates a model with a very strong level of predictive accuracy, but we will need to also assess how well this model fits the actual results. This is examined in the model diagnostic results presented in the two Figures below.





The first plot given above shows the models residuals vs the fitted values. Although the fitted line doesn't appear to be a perfect fit with the residuals, this is extremely close and indiciative of a strong model fit. Below the Residuals vs Fitted plot, the qqplot of the residuals appear to show a normal distribution, which is also points to a strong model fit.

Opponent Points per Possession Model

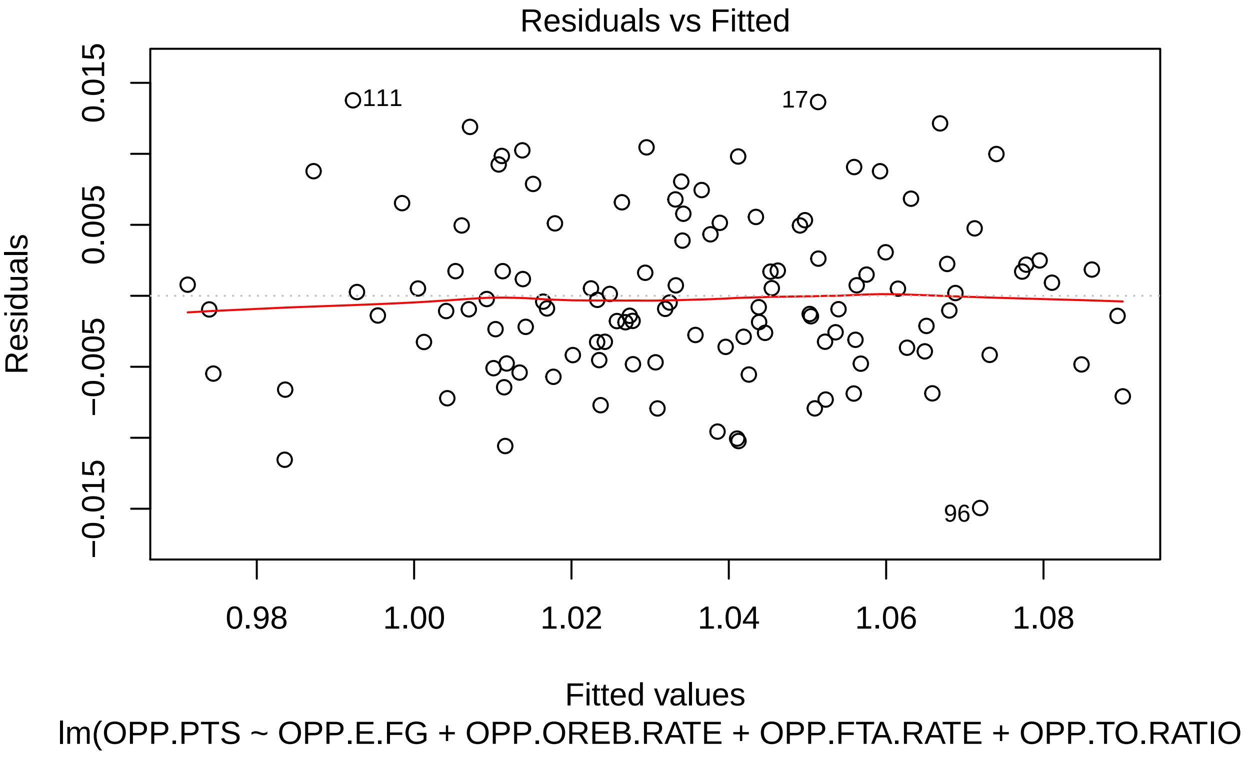
Predicted.OPP.PTS $= 0.32151 + OPP.E.FG * 0.01389 + OPP.OREB.RATE * 0.00475 + OPP.FTA.RATE * 0.20712 - OPP.TO.RATIO * 0.01053$

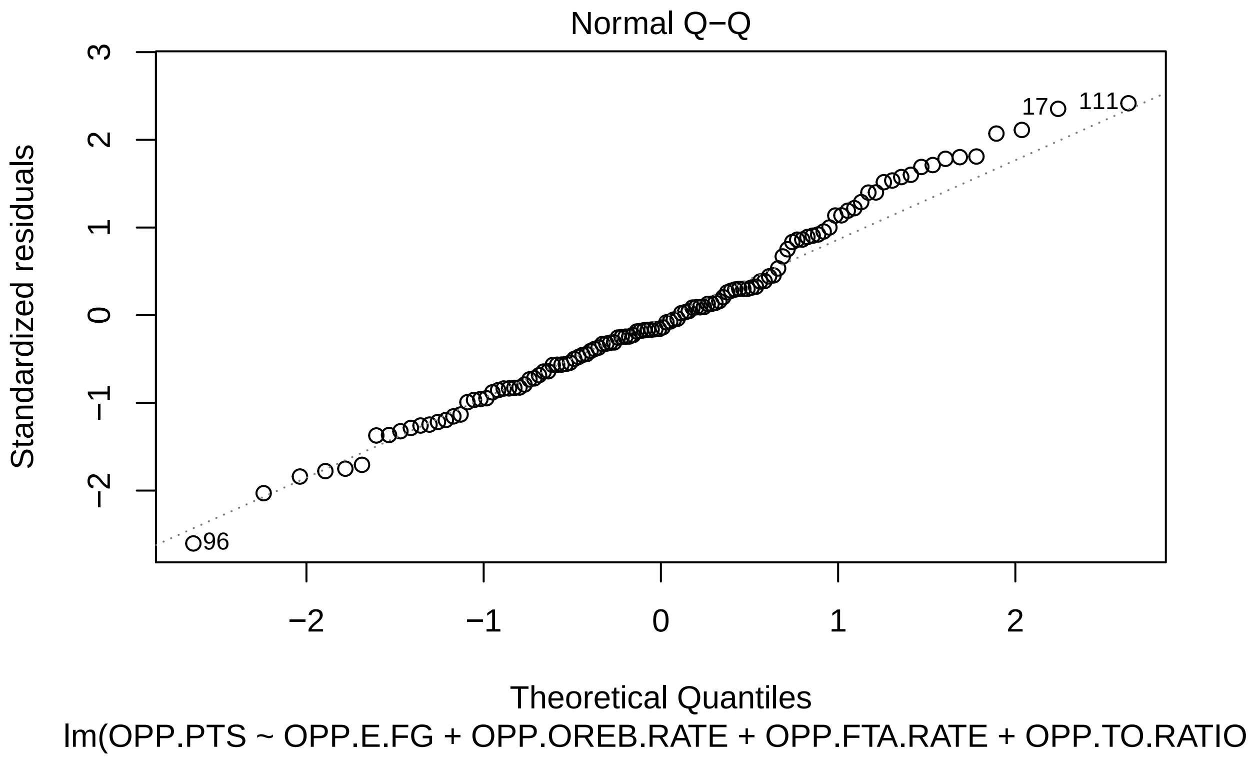
Similar methods were used to develop the opponent points model, as well. The results of this regression model are listed in the summary table below, and also appear to indicate a strong model fit with an R-Squared value of 0.9558.

Table 5: Fitting linear model: OPP.PTS ~ OPP.E.FG + OPP.OREB.RATE + OPP.FTA.RATE + OPP.TO.RATIO

	Estimate	Std. Error	t value	Pr(>|t|)
OPP.E.FG	0.0139	0.0003519	39.49	1.005e-68
OPP.OREB.RATE	0.004758	0.0003278	14.52	9.201e-28
OPP.FTA.RATE	0.2071	0.02288	9.052	4.188e-15
OPP.TO.RATIO	-0.01054	0.0005076	-20.76	1.047e-40
(Intercept)	0.3215	0.01743	18.45	3.738e-36

The opponent points model diagnostics are shown below, which allow us to visual assess the fit of this model.





The residuals for the opponent points regression model also seem to show a normal distribution, which again is a positive sign. The fitted values from this model appear to be a strong fit with the residuals, as well. This allows us to oonclude that the strength of this model is comparable to our selected points scored model.

Implications

The results from this initial exploration allow us to conclude that the four critical factors can be applied to provide an accuate measure of team performance on both the offensive and defensive side of the ball. Additionally, the results from this analysis will enable team decision makers to more accurately assess the strengths and weaknesses possessed by their team, as well as their opponents. The findings from this analysis can further be applied to support teams with developing opponent specific strategies that aim to maximize their winning probability associated wtih any given matchup.

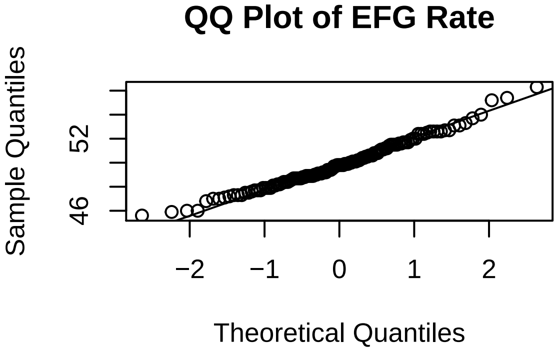
Bibliography

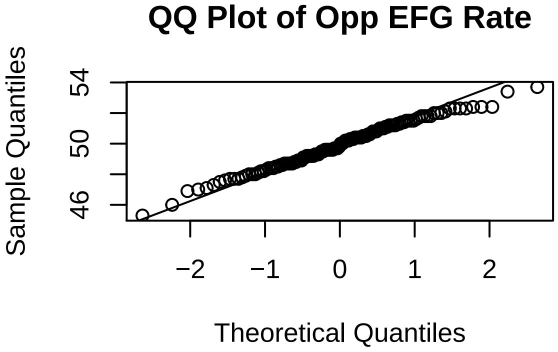
1. Glockner, Andy. *Chasing perfection: a behind-the-scenes look at the high-stakes game of creating an NBA champion.* Boston, MA, Da Capo Press, 2016. Print.

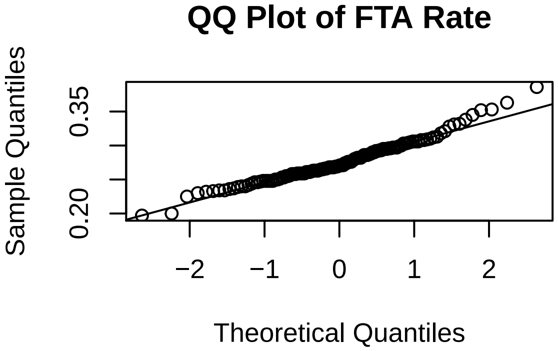
2. Oliver, Dean. *Basketball on Paper: Rules and Tools for Performance Analysis.* Washington, D.C.: Brassey's, 2004. Print.

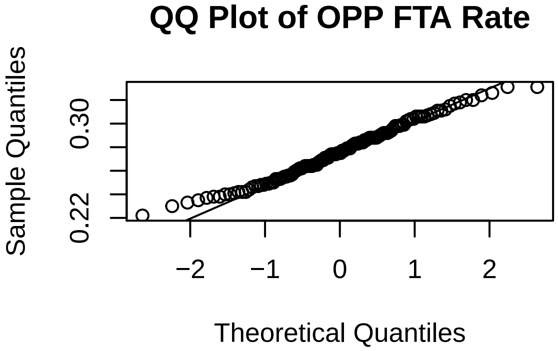
3. "Team per Possession Statistics." *NBA.com/Stats.* N.p,n d. Wed. 01 Dec. 2016.

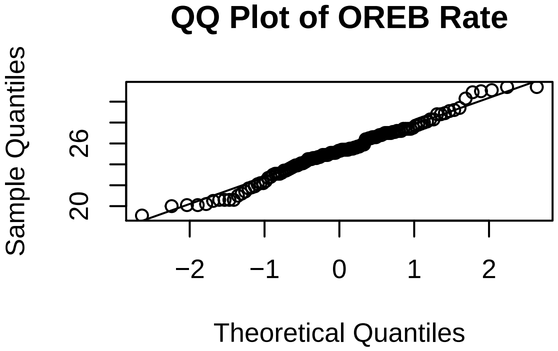
Appendix

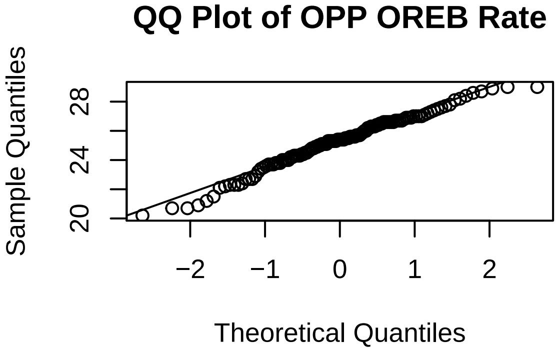

QQ Plot of EFG Rate

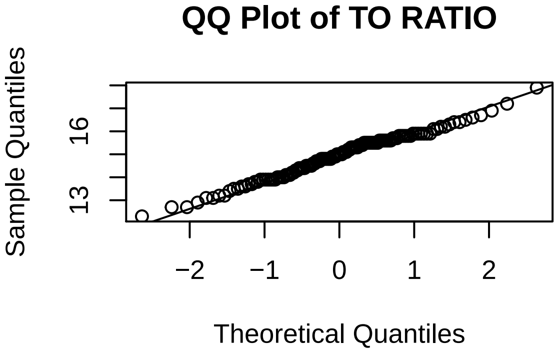

QQ Plot of Opp EFG Rate

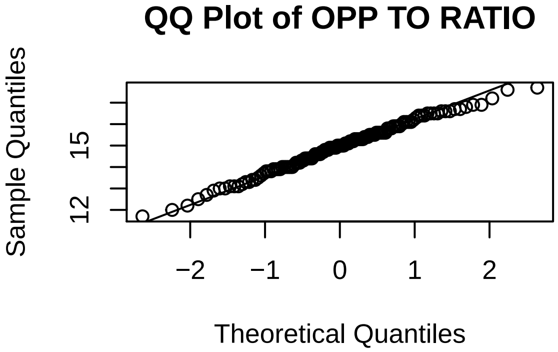

QQ Plot of FTA Rate


QQ Plot of OPP FTA Rate


QQ Plot of OREB Rate


QQ Plot of OPP OREB Rate


QQ Plot of TO RATIO


QQ Plot of OPP TO RATIO