Mihir Paranjape

12/20/2022

NBA Regular Season Defense Analysis

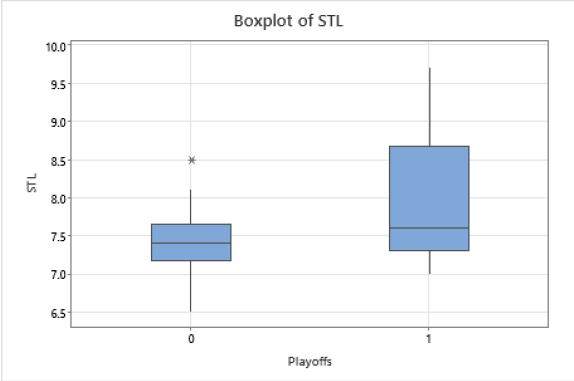
Defense has always been important in the NBA, perhaps even more so than offense. After all, there is the saying: Defense wins championships. But there aren't exactly a litany of defensive statistics that can tell someone what a particular defense needs in order to put their team in the best position to win games. Sure, there is a team's defensive rating, which measures the amount of points per 100 possessions that a team allows. This will obviously indicate how good a defense is, but it does not get into the specifics of defense itself. It does not show what aspects of defense truly move the needle in terms of making the playoffs, for example. There are some box score statistics like steals and blocks that can help show defensive activity, as well as some more advanced stats like defensive rebounding percentage, which estimates the proportion of available defensive rebounds that a team grabs, and opponent free throw attempts per field goal attempted. Today, I will try to see which defensive statistics make the biggest difference for teams, specifically in the context of making the playoffs or not. I will be looking at data from the 2021-2022 NBA regular season, and I will use the four statistics I mentioned as predictor variables and whether the team made the playoffs as a response variable in a logistic regression. Steals and blocks will be measured per 100 possessions rather than per game to control for pace.

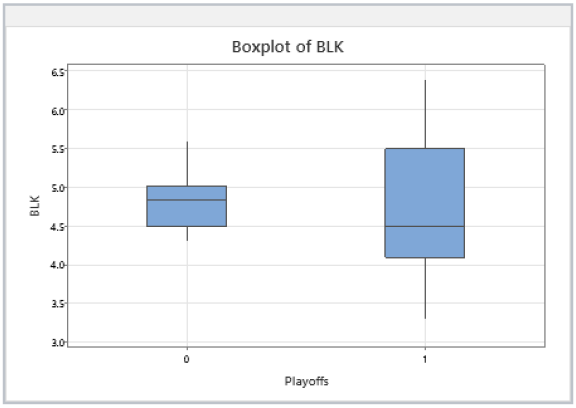
I gathered this data from Basketball Reference, exported it to Excel, added a column that indicated whether the team made the playoffs or not, and pasted the data into Minitab. From here, I was ready to start the logistic regression process. Here is the data I will be using:

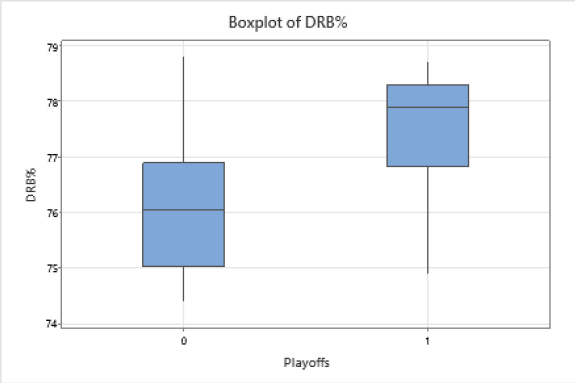
Team	STL	BLK	DRB%	FT/FGA
Atlanta Hawks*	7.3	4.3	76.9	0.177
Boston Celtics*	7.4	6	77.3	0.183
Brooklyn Nets*	7.1	5.5	75.1	0.201
Charlotte Hornets	8.5	4.9	74.8	0.187
Chicago Bulls*	7.2	4.2	78.3	0.199
Cleveland Cavaliers	7.3	4.3	76.5	0.172
Dallas Mavericks*	7	4.1	78	0.185
Denver Nuggets*	7.3	3.8	78.3	0.188
Detroit Pistons	7.8	4.8	75.6	0.226

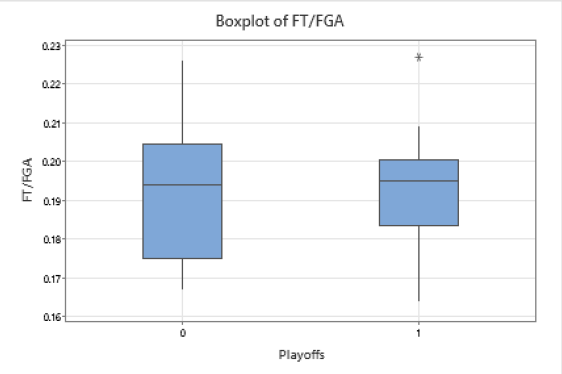
Golden State Warriors*	8.9	4.6	78.7	0.201
Houston Rockets	7.2	4.6	74.4	0.206
Indiana Pacers	7.2	5.6	76.2	0.204
Los Angeles Clippers	7.5	5	74.4	0.167
Los Angeles Lakers	7.5	5.1	75.8	0.192
Memphis Grizzlies*	9.7	6.4	77.8	0.195
Miami Heat*	7.6	3.3	78	0.209
Milwaukee Bucks*	7.6	4	78.6	0.165
Minnesota Timberwolves*	8.7	5.5	74.9	0.227
New Orleans Pelicans*	8.5	4.1	78.2	0.196
New York Knicks	7.2	5	78.8	0.198
Oklahoma City Thunder	7.6	4.7	76.1	0.169
Orlando Magic	6.8	4.5	77.2	0.196
Philadelphia 76ers*	8	5.5	76.8	0.192
Phoenix Suns*	8.6	4.4	77.1	0.195
Portland Trail Blazers	8.1	4.5	76.9	0.22
Sacramento Kings	7.1	4.5	76	0.176
San Antonio Spurs	7.6	4.9	75.1	0.176
Toronto Raptors*	9.3	4.7	75.6	0.199
Utah Jazz*	7.4	5	78.3	0.164
Washington Wizards	6.5	5.1	76.9	0.202

To start off, I took a look at side-by-side boxplots with each of the predictor variables categorized by the response variable. Here are the results:









There seems to be clear separation with defensive rebounding percentage between the groups, but all the other predictors seem to be centered similarly, although variances are quite different(which does not matter in this case with regards to violating assumptions). For blocks, it seems that teams who missed the playoffs have a higher number of blocks per 100 possessions(on average), which definitely surprises me. Maybe this could mean that teams who are better defensively force more outside shots, which inevitably do not get blocked as much.

Initial Regression

I will now do an initial logistic regression with all four predictors included in the model. Here are the results:

Analysis of Variance

Source	DF	Adj Dev	Adj Mean	Likelihood Ratio Chi-Square	P-Value
Regression	4	14.9571	3.7393	14.96	0.005
STL	1	6.8372	6.8372	6.84	0.009
BLK	1	0.1737	0.1737	0.17	0.677
DRB%	1	9.1822	9.1822	9.18	0.002
FT/FGA	1	0.1558	0.1558	0.16	0.693
Error	25	26.4983	1.0599		
Total	29	41.4554			

Looking at the Analysis of Variance(or Deviance) table, it seems that we would strongly reject the null hypothesis that there is no relationship between the predictors and the response variable ($p = 0.005 < 0.05$). From the individual likelihood ratio tests, it seems that steals and defensive rebounding percentage are strongly statistically significant, while blocks and free throw attempts per field goal attempt allowed do not seem to be significant predictors.

Odds Ratios for Continuous Predictors

	Unit of Change	Odds Ratio	95% CI
STL	1.00	7.9002	(1.1413, 54.6845)
BLK	1.00	1.4371	(0.2563, 8.0572)
DRB%	1.00	3.4976	(1.2779, 9.5731)
FT/FGA	0.01	0.8908	(0.4997, 1.5877)

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-111.2	44.9	-2.48	0.013	
STL	2.067	0.987	2.09	0.036	1.48
BLK	0.363	0.880	0.41	0.680	1.32
DRB%	1.252	0.514	2.44	0.015	1.65
FT/FGA	-11.6	29.5	-0.39	0.695	1.15

The intercept has no meaningful interpretation since none of the predictors will have values of zero over the course of a season. The coefficients can be interpreted as follows: An increase of 1 steal per 100 possessions for a team is associated with multiplying the odds ratio of making the playoffs by e^2.067 (7.9002). An increase of 1 block per 100 possessions for a team is associated with multiplying the odds ratio of making the playoffs by e^0.363(1.4371). An increase of 1 percentage point in a team's defensive rebounding percentage is associated with multiplying the odds ratio of making the playoffs by e^1.252(3.4976). An increase of 0.01 in the opponent's free throw to field goals attempted ratio for a team is associated with multiplying the odds ratio of making the playoffs by e^ (0.01*-11.6)= (0.8908). I used 0.01 as the unit of change for this predictor since the values for this predictor range from 0.16 to around 0.22, so it would not make sense if I used 1 as the unit of change. All of these would entail holding all other predictors constant. None of the variance inflation factors for the predictors are anywhere near 10 so there are no signs of multicollinearity.

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	25	26.50	0.381
Pearson	25	34.69	0.094
Hosmer-Lemeshow	8	10.93	0.206

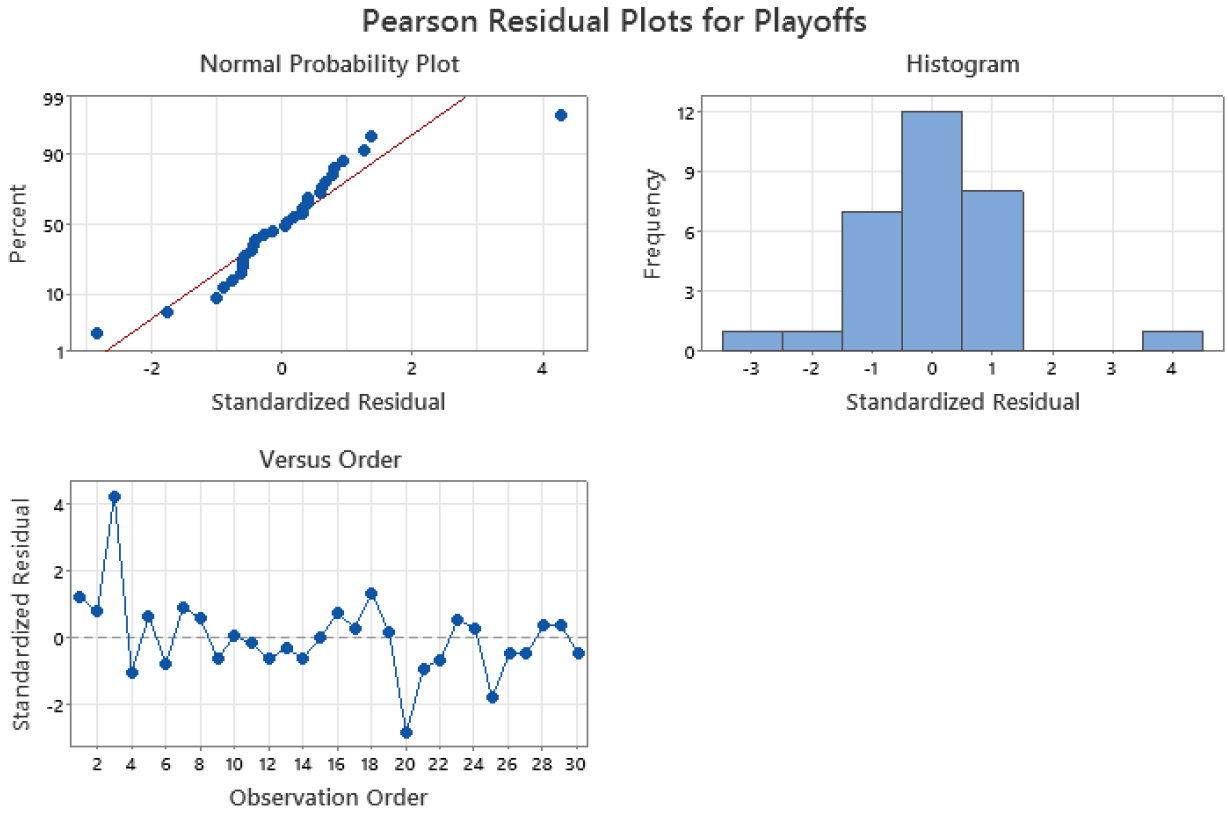
The only relevant goodness-of-fit test for the data is the Hosmer-Lemeshow test since there is no replication in the data. The test suggests that there is no significant evidence of lack-of-fit (p = 0.206 >0.05).

Measures of Association

Pairs	Number	Percent	Summary Measures	Value
Concordant	200	89.3	Somers' D	0.79
Discordant	24	10.7	Goodman-Kruskal Gamma	0.79
Ties	0	0.0	Kendall's Tau-a	0.40
Total	224	100.0		

Association is between the response variable and predicted probabilities

The Somers' D value of 0.79 suggests that there is good separation between the playoff teams and the non-playoff teams with our regression. This measurement is an alternative to the R-squared for ordinary least squares regression.



Because we have a small sample of 30 teams, the residual plots will not be perfectly calibrated, and therefore I will not analyze the normality of the residuals.

Team	SPEARRES	HI	COOK
Atlanta Hawks*	1.24059	0.123902	0.043532
Boston Celtics*	0.80309	0.403727	0.087338
Brooklyn Nets*	4.25526	0.084015	0.332163
Charlotte Hornets	-0.99898	0.291141	0.081977
Chicago	0.67242	0.150245	0.015989

Bulls*			
Cleveland Cavaliers	-0.75275	0.159572	0.021518
Dallas Mavericks*	0.94027	0.156402	0.032782
Denver Nuggets*	0.61335	0.154499	0.013749
Detroit Pistons	-0.60122	0.229839	0.021574
Golden State Warriors*	0.07929	0.023535	0.00003
Houston Rockets	-0.14894	0.052624	0.000246
Indiana Pacers	-0.60474	0.195173	0.017737
Los Angeles Clippers	-0.28397	0.120545	0.002211
Los Angeles Lakers	-0.5891	0.084202	0.006382
Memphis Grizzlies*	0.04231	0.016148	0.000006
Miami Heat*	0.76525	0.349847	0.063023
Milwaukee Bucks*	0.30894	0.122362	0.002661
Minnesota Timberwolves*	1.34652	0.418586	0.261069
New Orleans Pelicans*	0.17751	0.057514	0.000385
New York Knicks	-2.82574	0.175769	0.340556

Oklahoma City Thunder	-0.88796	0.185551	0.035926
Orlando Magic	-0.62473	0.150526	0.013832
Philadelphia 76ers*	0.57644	0.166846	0.013309
Phoenix Suns*	0.3072	0.100549	0.00211
Portland Trail Blazers	-1.77081	0.192019	0.149046
Sacramento Kings	-0.44139	0.112552	0.004942
San Antonio Spurs	-0.45749	0.130161	0.006264
Toronto Raptors*	0.39949	0.230907	0.009583
Utah Jazz*	0.39292	0.178499	0.006709
Washington Wizards	-0.41691	0.182742	0.007773

There seem to be two residuals with high values of 4.255 and -2.825. The leverage value of interest is 2.5((4+1)/30) or 0.41667. There is one observation with a leverage value of 0.41856. The observations with high residuals are the Brooklyn nets and the New York Knicks, respectively. The observation with a high leverage value is the Minnesota Timberwolves. Minnesota had the largest opponent free throw to field goal attempted ratio in the league, alongside the third lowest defensive rebounding percentage and the fourth highest steals per 100 possessions. Brooklyn had a bottom five defensive rebounding percentage, as well as a bottom ten steals per 100 possessions. These two predictors were the strongest in the regression, yet Brooklyn made the playoffs because of the offensive brilliance of Kevin Durant and Kyrie Irving. This explains why they have such a high residual. Meanwhile, New York had the highest defensive rebounding percentage in the league, but failed to make the playoffs, which is why they had a high residual as well. I will now try a regression without these three points to see if anything significantly changes. Here is the message I got:

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

* WARNING * The model could not be fit properly. Maximum likelihood estimates of parameters do not exist due to complete separation of data points. The results are not reliable. Please refer to help for more information about complete separation.

Method

Link function	Logit
Residuals for diagnostics	Pearson
Rows used	27

This means that the removal of these points causes complete separation, and I should look to use a simpler model. I will use the best subsets regression to narrow down the possibilities for models. I am keeping in mind the fact that the likelihood ratio tests indicate that steals and defensive rebounding percentage are the statistically significant predictors. Here are the best subsets results:

Response is Playoffs

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	STL	BLK	DRB%	FT/FGA
1	53.7	51.8	48.5	7.6	0.35335			X	
1	13.5	10.0	3.4	34.1	0.48297	X			
2	64.3	61.3	55.1	2.6	0.31668	X		X	
2	54.3	50.4	45.8	9.2	0.35844			X	X
3	66.4	62.0	54.1	3.2	0.31394	X		X	X
3	64.5	59.9	50.8	4.4	0.32254	X	X	X	
4	66.7	60.6	50.1	5.0	0.31944	X	X	X	X

These results confirm that the model I should be using includes steals and defensive rebounding percentage. The Mallows Cp value is minimized with this two-predictor model.

Best Subsets Regression (3 outliers removed)

Here are the results for this particular model, without the three unusual observations I mentioned:

Analysis of Variance

Source	DF	Adj Dev	Adj Mean	Likelihood Ratio Chi-Square	P-Value
Regression	2	28.011	14.0055	28.01	0.000
STL	1	9.012	9.0118	9.01	0.003
DRB%	1	23.994	23.9943	23.99	0.000
Error	24	9.382	0.3909		
Total	26	37.393			

We strongly reject the null hypothesis that there is no relationship between the predictors and response ($p<0.001$). The individual Likelihood Ratio tests also indicate that both predictors are statistically significant.

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-325	146	-2.24	0.025	
STL	3.45	1.58	2.19	0.029	1.98
DRB%	3.90	1.78	2.19	0.029	1.98

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
STL	31.5585	(1.4291, 696.9125)
DRB%	49.1777	(1.4963, 1616.3290)

The coefficients can be interpreted as follows (holding the other predictor fixed): An increase of 1 steal per 100 possessions for a team is associated with multiplying the odds ratio of making the playoffs by e^3.45 (31.5585). An increase of 1 percentage point in a team's defensive rebounding percentage is associated with multiplying the odds ratio of making the playoffs by e^3.9(49.177). There are once again low variance inflation factors.

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	24	9.38	0.997
Pearson	24	11.90	0.981
Hosmer-Lemeshow	8	5.57	0.695

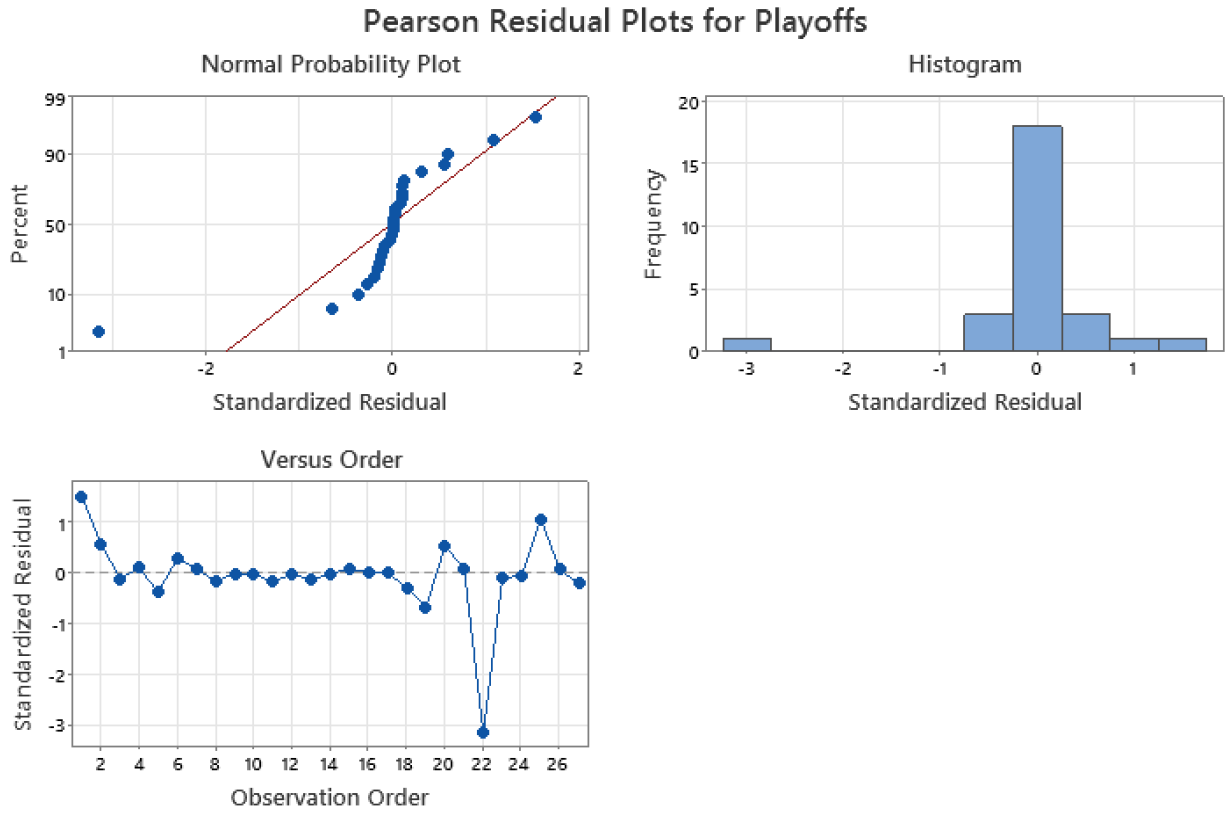
The Hosmer-Lemeshow test suggests that there is no strong evidence of a lack of fit (p is much greater than 0.05).

Measures of Association

Pairs	Number	Percent	Summary Measures	Value
Concordant	178	97.8	Somers' D	0.96
Discordant	4	2.2	Goodman-Kruskal Gamma	0.96
Ties	0	0.0	Kendall's Tau-a	0.50
Total	182	100.0		

Association is between the response variable and predicted probabilities

The Somers' D value of 0.96 indicates that there is extreme separation between making the playoffs and not with the regression.



These plots again do not tell us much about assumptions since there is such a small sample size (27 observations in this model).

Team	SPEARRES	HI	COOK
Atlanta Hawks*	1.53042000	0.20574600	0.20224000
Boston Celtics*	0.59358000	0.21295000	0.03178000
Charlotte Hornets	-0.10130000	0.07656900	0.00028000
Chicago Bulls*	0.10970000	0.06540300	0.00028000
Cleveland Cavaliers	-0.36472000	0.14928200	0.00778000
Dallas Mavericks*	0.30318000	0.21481900	0.00838000
Denver Nuggets*	0.09155000	0.04980800	0.00015000
Detroit Pistons	-0.14415000	0.08151000	0.00061000
Golden State Warriors*	0.00259000	0.00017400	0.00000000
Houston Rockets	-0.00474000	0.00052300	0.00000000
Indiana Pacers	-0.16431000	0.07758800	0.00076000
Los Angeles Clippers	-0.00796000	0.00126600	0.00000000
Los Angeles Lakers	-0.12549000	0.06228500	0.00035000
Memphis Grizzlies*	0.00376000	0.00032400	0.00000000
Miami Heat*	0.09772000	0.04738400	0.00016000
Milwaukee Bucks*	0.02979000	0.00966700	0.00000000
New Orleans Pelicans*	0.01368000	0.00258800	0.00000000
Oklahoma City Thunder	-0.27852000	0.13497100	0.00403000
Orlando Magic	-0.66283000	0.29922400	0.06253000
Philadelphia 76ers*	0.54396000	0.17159900	0.02043000
Phoenix Suns*	0.10018000	0.04317500	0.00015000
Portland Trail Blazers	-3.15185000	0.14379300	0.55612000
Sacramento Kings	-0.09181000	0.04006100	0.00012000

San Antonio Spurs	-0.03719000	0.01349400	0.00001000
Toronto Raptors*	1.08175000	0.74742300	1.15427000
Utah Jazz*	0.07656000	0.03798200	0.00008000
Washington Wizards	-0.19542000	0.11039400	0.00158000

The Portland Trail Blazers are now an outlier (-3.145) and the Toronto Raptors have an extremely high leverage value(0.747) when the value of interest is 2.5(3/27) or 0.2778. The Orlando Magic also cross this threshold with a leverage value of 0.299. Only six teams had a higher number of steals per 100 possessions than the Trail Blazers, and they had the median for defensive rebounding percentage. This regression expected that they would easily make the playoffs but they did not, which explains their high residual. The Raptors had such a high leverage value because they had the second highest steals average and tied for the fifth lowest defensive rebounding percentage. Meanwhile, the Magic had the second lowest steals average and the eleventh highest defensive rebounding percentage, which gave them high enough of a leverage value to be pointed out. I will try another regression with ALL predictors and the removal of these three points. Once again, I got the complete separation error.

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

* WARNING * The model could not be fit properly. Maximum likelihood estimates of parameters do not exist due to complete separation of data points. The results are not reliable. Please refer to help for more information about complete separation.

I will try the best subsets once again as well. Here are the results:

Response is Playoffs

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	STL	BLK	DRB%	FT/FGA
1	71.6	70.3	67.8	1.8	0.27757			X	
1	10.5	6.4	0.0	48.5	0.49243	X			
2	74.9	72.5	68.7	1.2	0.26701	X		X	
2	71.7	69.0	65.4	3.6	0.28328		X	X	
3	75.1	71.4	66.4	3.0	0.27238	X		X	X
3	74.9	71.2	64.7	3.2	0.27333	X	X	X	
4	75.2	69.9	62.3	5.0	0.27912	X	X	X	X

I will use the same model with steals and defensive rebound percentage. However, when I do this, I get the same message about a complete separation of data points. This suggests to me that I should put the outliers and leverage points back in, returning to the model with 27 observations (putting the Raptors, Magic, and Trailblazers back in). Before I do this, I will try a logistic regression with all observations but with only the two variables that seem to be the most important: steals and defensive rebounding %. Here are the results:

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	27	26.81	0.474
Pearson	27	34.48	0.152
Hosmer-Lemeshow	8	11.63	0.168

Analysis of Variance

Source	DF	Adj Dev	Adj Mean	Likelihood Ratio Chi-Square	P-Value
Regression	2	14.648	7.3238	14.65	0.001
STL	1	6.878	6.8779	6.88	0.009
DRB%	1	10.057	10.0568	10.06	0.002
Error	27	26.808	0.9929		
Total	29	41.455			

We would still strongly reject the null hypothesis that there is no relationship between the predictors and the response variable($p = 0.001 < 0.05$). Both the individual tests once again indicate statistical significance. The Goodness-of-Fit test here indicates once again that there is no evidence of lack of fit($p = 0.168 > 0.05$).

Measures of Association

Pairs	Number	Percent	Summary Measures	Value
Concordant	197	87.9	Somers' D	0.76
Discordant	27	12.1	Goodman-Kruskal Gamma	0.76
Ties	0	0.0	Kendall's Tau-a	0.39
Total	224	100.0		

Association is between the response variable and predicted probabilities

The Somers' D value indicates that our regression has clearly separated the playoff and non-playoff teams well.

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-104.4	38.7	-2.70	0.007	
STL	1.943	0.916	2.12	0.034	1.30
DRB%	1.169	0.454	2.57	0.010	1.30

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
STL	6.9780	(1.1579, 42.0503)
DRB%	3.2187	(1.3215, 7.8400)

The same interpretations for the coefficients hold, although the coefficients seem to be lower when including all observations. Holding the defensive rebounding percentage fixed, an increase of 1 steal per 100 possessions for a team is associated with multiplying the odds ratio of making the playoffs by e^1.943 (6.978). Holding steals per 100 possessions fixed, an increase of 1 percentage point in a team's defensive rebounding percentage is associated with multiplying the odds ratio of making the playoffs by e^1.169(3.2187). There is no multicollinearity between the predictors.

Team	SPEARRES	HI	COOK
Atlanta Hawks*	1.20601	0.069866	0.036417
Boston Celtics*	0.8664	0.07025	0.018906
Brooklyn Nets*	4.19208	0.068971	0.433949
Charlotte Hornets	-0.922	0.231519	0.085367
Chicago Bulls*	0.60191	0.117341	0.016055
Cleveland Cavaliers	-0.70631	0.071703	0.012845
Dallas Mavericks*	0.88041	0.135937	0.040648
Denver Nuggets*	0.54292	0.10665	0.01173
Detroit Pistons	-0.68447	0.088152	0.015097

Golden State Warriors*	0.08698	0.025825	0.000067
Houston Rockets	-0.18619	0.055372	0.000677
Indiana Pacers	-0.5406	0.081132	0.008601
Los Angeles Clippers	-0.25161	0.073518	0.001675
Los Angeles Lakers	-0.57143	0.077176	0.009103
Memphis Grizzlies*	0.06756	0.022792	0.000035
Miami Heat*	0.47739	0.083931	0.00696
Milwaukee Bucks*	0.33618	0.083966	0.003453
Minnesota Timberwolves*	1.13284	0.280461	0.166738
New Orleans Pelicans*	0.17425	0.052794	0.000564
New York Knicks	-2.52662	0.121176	0.293409
Oklahoma City Thunder	-0.74652	0.067503	0.013447
Orlando Magic	-0.67656	0.131879	0.023179
Philadelphia 76ers*	0.65088	0.078685	0.01206
Phoenix Suns*	0.30739	0.093257	0.003239
Portland Trail Blazers	-1.95623	0.086101	0.120178
Sacramento Kings	-0.4373	0.084756	0.005903
San Antonio Spurs	-0.42107	0.089408	0.005803
Toronto Raptors*	0.40284	0.217444	0.01503
Utah Jazz*	0.49034	0.098181	0.008725
Washington Wizards	-0.42481	0.134254	0.009328

The same story with residuals and leverage values from the first model appear here, with the Knicks, Nets, and Timberwolves, being classified as unusual. The leverage value of interest here is 2.5(3/30) = 0.25, and the Timberwolves have a leverage value of 0.28. The Knicks and Nets have residuals of -2.52 and 4.19 respectively. Given this regression and the other regression without these three observations, I can choose either model to continue with since both have an appropriate goodness of fit, and I cannot further remove outliers without running into the problem of complete separation. Also, I realized early on from multiple best subsets as well as the individual likelihood ratio tests that steals and defensive rebound

percentage are the most significant factors towards making the playoffs. I will choose the regression model without the three observations to continue with because it had a much stronger Hosmer-Lemeshow result (0.695 as p-value compared to 0.168). As a reminder, here are the results once again:

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-325	146	-2.24	0.025	
STL	3.45	1.58	2.19	0.029	1.98
DRB%	3.90	1.78	2.19	0.029	1.98

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
STL	31.5585	(1.4291, 696.9125)
DRB%	49.1777	(1.4963, 1616.3290)

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	24	9.38	0.997
Pearson	24	11.90	0.981
Hosmer-Lemeshow	8	5.57	0.695

Measures of Association

Pairs	Number	Percent	Summary Measures	Value
Concordant	178	97.8	Somers' D	0.96
Discordant	4	2.2	Goodman-Kruskal Gamma	0.96
Ties	0	0.0	Kendall's Tau-a	0.50
Total	182	100.0		

Association is between the response variable and predicted probabilities

Analysis of Variance

Source	DF	Adj Dev	Adj Mean	Likelihood Ratio Chi-Square	P-Value
Regression	2	28.011	14.0055	28.01	0.000
STL	1	9.012	9.0118	9.01	0.003
DRB%	1	23.994	23.9943	23.99	0.000
Error	24	9.382	0.3909		
Total	26	37.393			

Team	SPEARRES	HI	COOK
Atlanta Hawks*	1.53042000	0.20574600	0.20224000
Boston Celtics*	0.59358000	0.21295000	0.03178000
Charlotte Hornets	-0.10130000	0.07656900	0.00028000
Chicago Bulls*	0.10970000	0.06540300	0.00028000
Cleveland Cavaliers	-0.36472000	0.14928200	0.00778000
Dallas Mavericks*	0.30318000	0.21481900	0.00838000
Denver Nuggets*	0.09155000	0.04980800	0.00015000
Detroit Pistons	-0.14415000	0.08151000	0.00061000

Golden State Warriors*	0.00259000	0.00017400	0.00000000
Houston Rockets	-0.00474000	0.00052300	0.00000000
Indiana Pacers	-0.16431000	0.07758800	0.00076000
Los Angeles Clippers	-0.00796000	0.00126600	0.00000000
Los Angeles Lakers	-0.12549000	0.06228500	0.00035000
Memphis Grizzlies*	0.00376000	0.00032400	0.00000000
Miami Heat*	0.09772000	0.04738400	0.00016000
Milwaukee Bucks*	0.02979000	0.00966700	0.00000000
New Orleans Pelicans*	0.01368000	0.00258800	0.00000000
Oklahoma City Thunder	-0.27852000	0.13497100	0.00403000
Orlando Magic	-0.66283000	0.29922400	0.06253000
Philadelphia 76ers*	0.54396000	0.17159900	0.02043000
Phoenix Suns*	0.10018000	0.04317500	0.00015000
Portland Trail Blazers	-3.15185000	0.14379300	0.55612000
Sacramento Kings	-0.09181000	0.04006100	0.00012000
San Antonio Spurs	-0.03719000	0.01349400	0.00001000
Toronto Raptors*	1.08175000	0.74742300	1.15427000
Utah Jazz*	0.07656000	0.03798200	0.00008000
Washington Wizards	-0.19542000	0.11039400	0.00158000

This model seems to be strongest due to the goodness-of-fit tests and the removal of outliers to the point where it does not have perfect separation between the groups in the response variable. In addition, these two predictors are significant in the regression, as told by the best subsets models and the individual likelihood ratio tests. Now, I will see how well the model can classify, putting the unusual observations back in for prediction purposes. For Minnesota (the leverage point), the predicted Fits value was 0.027, which means they were predicted to not make the playoffs yet they did. The two outliers will be misclassified, so the Nets will be predicted to miss the playoffs while they actually made it, and the Knicks will be predicted to make the playoffs while they actually missed it. Here are the results for the classification matrix:

Rows: Playoffs Columns: Predict

	0	1	All
0	12	2	14
	40.00	6.67	46.67
1	3	13	16
	10.00	43.33	53.33
All	15	15	30
	50.00	50.00	100.00

Cell Contents
Count

The rows are the actual results while the columns indicate the predicted results. Adding up the upper left and bottom right probabilities, it seems this model correctly classified 83.33% of teams in making the playoffs or not. The C_{max} in this dataset is 16/30 = 53.33%, the larger group being teams that made the playoffs. The C_{pro} here would be 62.5% [1.25(0.5*0.4667+0.5*0.5333)]. The correct classification percentage is higher than both C_{pro} and C_{max} which shows the strength of this model. Another way we can evaluate the model's classification performance is by validating the model on new data. I could choose a different NBA season to look at and see if it predicted a playoff berth correctly. I will try this with the 2018-2019 NBA season because that would be the closest full season without external factors like COVID-19 influencing the NBA. To do this, I had to take all the same steps to import the data as before. Then, I predicted using columns of values for steals and defensive rebounding percentage for all thirty teams. Here is all the data:

Team	STL	DREB%	2018-19 Playoffs	PFITS	2018-19 Predict
Atlanta Hawks	7.9	76.4	0	0.37816	0
Boston Celtics*	8.6	77	1	0.98602	1
Brooklyn Nets*	6.4	76.4	1	0.00342	0
Charlotte Hornets	7.2	77.1	0	0.45343	0
Chicago Bulls	7.3	77.3	0	0.71858	1
Cleveland Cavaliers	6.7	77	0	0.09093	0
Dallas Mavericks	6.5	77.5	0	0.2602	0

Denver Nuggets*	7.9	78	1	0.99678	1
Detroit Pistons*	7.1	78.7	1	0.99667	1
Golden State Warriors*	7.5	77.1	1	0.7003	1
Houston Rockets*	8.7	74.4	1	0.00396	0
Indiana Pacers*	8.8	76.2	1	0.86179	1
Los Angeles Clippers*	6.7	76	1	0.00203	0
Los Angeles Lakers	7.3	76.4	0	0.0712	0
Memphis Grizzlies	8.5	77.6	0	0.99807	1
Miami Heat	7.8	77.6	0	0.97879	1
Milwaukee Bucks*	7.2	80.3	1	1	1
Minnesota Timberwolves	8.3	74.9	0	0.00697	0
New Orleans Pelicans	7.2	76.8	0	0.20498	0
New York Knicks	6.8	76.1	0	0.00422	0
Oklahoma City Thunder*	9	78.2	1	0.99997	1
Orlando Magic*	6.7	79.7	1	0.99973	1
Philadelphia 76ers*	7.2	78.6	1	0.99652	1
Phoenix Suns	8.8	72.5	0	0	0
Portland Trail Blazers*	6.7	77.9	1	0.76916	1
Sacramento Kings	8	75.5	0	0.02513	0
San Antonio Spurs*	6.2	79.4	1	0.99514	1
Toronto Raptors*	8.2	77.1	1	0.96321	1
Utah Jazz*	8	80.3	1	1	1
Washington Wizards	8.1	74.1	0	0.00016	0

Here is the classification matrix for the 2018-2019 NBA season:

Rows: 2018-19 Playoffs Columns: 2018-19 Predict

	0	1	All
0	11	3	14
	36.67	10.00	46.67
1	3	13	16
	10.00	43.33	53.33
All	14	16	30
	46.67	53.33	100.00

For this data, the model correctly classified 80% of teams in making the playoffs or not. C_{max} is still 53.33% in this case and C_{pro} in this case is (0.4667*0.4667+0.5333*0.5333) or 50.222%. The C_{pro} in this case would not be multiplied by 1.25 because we are validating on new data. Even if the 1.25 was multiplied by the current C_{pro}, this number would be 62.777%. This is still below the correct classification proportion of the model. I believe this truly shows the success of the model in predicting whether a team makes the playoffs or not.

Some key takeaways I have from this whole process would be that rebounding seems to be very important for teams to win games in the regular season. An increase of 1 percentage point for rebounding percentage is associated with multiplying the odds ratio by 49 according to our strongest model, and about 3.218 for our model including outliers initially. I think this makes sense because rebounding can be quite the momentum shifter for teams when it is a close game. Limiting a team to one shot per possession puts the pressure on them to be effective for that one shot, without the comfort of another possession in a row. Of course, rebounding on the defensive side can also lead to fastbreak opportunities for a team, which means that the net change between a team grabbing the defensive rebound versus them letting the offense have another possession can be monumental in situations. I was surprised to see that limiting free throws per field goal attempt was not a significant factor in the model. Free throws are the most efficient offense for any team, so I thought it would be intuitive that limiting free throws per possession or field goal attempt would have a similar effect as grabbing the defensive rebound every time. This would come from the offense getting frustrated about a lack of fouls calls against a particular team, and the fact that repeated trips to the free throw line could be demoralizing for a defense. However, this does not seem to show itself in the model. I was also surprised to see that blocks did not really have any impact on making the playoffs. In fact, the initial box plots showed that non-playoff teams had higher blocks per 100 possessions on average, which caught me off guard. This could speak to the fact that the truly good

defensive teams do not even let the offense get in positions where they have to block them (in the paint mostly). As for steals, I'm guessing that steals often lead to fastbreak points on the other end, which is why they are a good predictor for regular season success. Steals also imply that a defense is putting pressure on ball-handlers and jumping the passing lanes, which leads to an offense that might be more cautious and less aggressive.

While this analysis taught me a lot about how some particular defensive statistics can be important in determining whether a team makes the playoffs or not, I still think there needs to be more data that paints a picture for how a team is good defensively, rather than simply stating the fact that they are good defensively (defensive rating does this, for example). Perhaps something regarding how well they navigate screens or how well they close out to shooters and rotate when there is a lapse in the defense. Even if this data does exist, I assume it will not be widely available, as it may be expensive to track these sorts of niche occurrences, which I believe are still vital to determining the strength of a defense. Nonetheless, this analysis still confirms how important defense is, as the measures I chose did not directly relate to clear defensive measures like defensive rating, yet they still had predictive success for whether a team made the playoffs or not.