The Success of College Basketball Teams can be predicted using several advanced statistics*

A Linear Regression Approach

Thomas D'Onofrio

26 April 2022

Abstract

The March Madness college basketball tournament is one of the most exciting sports events in the world due to the intense games, unpredictable results, and the creation of brackets by millions. In this paper, data displaying each college basketball schools' advanced team statistics was used to create a linear model with Simple Rating System as the response variable, where the explanatory variables are the team statistics that play the largest factor in leading to team success. It was found that true shooting percentage, total rebound percentage, and turnover percentage are the most important team statistics that affect the amount of success a team has, meaning fans can use these statistics to fill out their brackets more successfully. Unfortunately, the data is slightly biased due to inconsistent scorekeeping across teams and arenas, and the racial biases of referees when calling fouls, causing some implications to the findings.

1 Introduction

March is a very exciting time for sports fans in the US and many other countries, as it is when the NCAA Men's Division 1 Basketball Tournament, more commonly known as March Madness, is played. 68 college basketball teams go head to head, playing single-elimination games until there is only one team remaining, making for some of the most intense and unpredictable sports games in the world. As a result of this, March Madness draws in a lot of viewers. 16.9 million people tuned in to the 2021 March Madness finals, making it the 2nd most viewed North American sports event only behind the Superbowl (Baker 2021). A big reason for the massive following is that many people create brackets for the tournament, trying to predict the results. With 9.2 quintillion different bracket possibilities, all the insane upsets that nobody sees coming, and the overall chaotic nature of the tournament, creating a bracket and seeing how well you do is very appealing to both sports fans and regular people, causing all the hype around March Madness (Geiling 2014). People put a lot of work into their brackets, comparing teams to one another and trying to figure out who has the best odds of winning. With this in mind, is there a way people can more successfully predict which teams will have more success than others in the tournament?

In this paper, we attempt to answer this intriguing question by analyzing a variety of datasets from the college basketball section of Sports Reference (Sports Reference 2022), using R (R Core Team 2020). Datasets containing college basketball statistics from 2017 to 2022 were used, where each dataset contains many different advanced team statistics for all NCAA Division 1 schools with a Men's basketball team in a specific season. The data was first graphically analyzed to determine which team statistics have a strong or noticeable relationship with how successful a team is in their games. These statistics were then used to create a linear model that attempts to predict a team's Simple Rating System, the response variable that is being used to measure team success, where only the best and most necessary team statistics were included. As a result of this, it was determined that the following statistics, true shooting percentage, total rebound percentage, and turnover percentage, play the most important factor in leading to team success. Using estimated coefficients, values of these statistics could be used to predict what a teams simple rating system value will be and how

[*] Code and data are available at: https://github.com/TDonofrio62/Predicting-College-Basketball-Success.

successful the team should be. Thus, when filling out their brackets, fans can compare these statistics across teams to determine which teams have a good chance of going far or even winning the tournament, as opposed to just looking at wins or their seeding like many people do.

Unfortunately, there are some implications with these findings. There appears to be several aspects of bias in the datasets being used. For one, there is a lot of human error in the tracking of statistics by scorekeepers, causing certain ambiguously defined statistics in the data to be incorrectly tracked and thus skewed (Williams 2017). This means that data is tracked differently for each team in the league in their respective arenas, made evident by Madison Square Garden's scorekeepers overstating the quality of shots in the NHL (McCurdy 2020). Thus, the data for some teams may be biased, making them look either better or worse relative to the other teams in the league. There is also unfortunate racial biases in basketball, as referees tend to show racism by calling more fouls on players of different races than themselves, leading to these teams scoring less points in games where this racial injustice occurs (Wolfers and Price 2012). With more white referees in the league, teams with more white players may have team statistics skewed to make them seem better than they truly are, where teams with more players of colour will look worse because of the bias. We must be aware that our findings have been affected by these biases and as a result may not be perfectly correct. Nevertheless, a lot can be learned from the datasets which are still quite informative.

The rest of this paper will be as follows: Section 2 goes over the datasets being used in the paper, and analyzes the data with the use of tables and graphs to see which team statistics have a relationship with Simple Rating System. Section 3 includes the creation of a linear regression model that uses only the most important team statistics as explanatory variables to predict Simple Rating System, the response variable that determines team success. Section 4 interprets the coefficients and provides a statistical analysis of the model in order to understand what is being done. Finally, Section 5 discusses what exactly the model and its outcome are telling us, what can be learned from it, and what implications come along with the findings.

2 Data

The data being used to answer our research question in this paper is from Sports Reference, and specifically the college basketball section of the site (Sports Reference 2022). Sports Reference is a fantastic site that contains hundreds of thousands of statistics for all of the most popular American sports, making it a site that many rely on constantly for their sports information. Their college basketball section includes data dating all the way back to the late 1800s, and includes the results from every single game since 2010, giving us a plethora of information to potentially use. This data is very reliable as it is taken directly from the box scores of each game, which have been tracked by both NCAA certified referees and scorekeepers (Sports Reference 2022). Each section of data on their site also gives you an option to export their data, stating that use of their datasets is allowed as long as they are given proper citation, meaning that using their information for research is ethical.

Of the website's abundance of data, we will only be working with team statistics from the 2017-18 season to the most recent 2021-22 season, a 5 season period. The reason for this is that basketball is a constantly changing sport where teams adapt to new trends and game plans that optimize their success. The game is being played differently today than it was even a decade ago, so our results will be able to more accurately predict future results if only more recent data is used. Each of the five datasets contain several advanced team regular season statistics for every NCAA Division 1 school in a given season. These statistics include and average a teams results from each game in the season. With statistics from the 5 datasets being used, which were all combined into one large dataset for use, there are 1762 observations of schools and their results that we can work with. Since only 5 seasons are used, this dataset is still only a sample of all college basketball statistics of all time, as opposed to being true population data. Also, as a result of COVID-19, many games were canceled from 2020 to 2022 due to outbreaks and team sicknesses, so the data still would not be considered population data for the past 5 seasons. Although we are working with sample data, between the amount of seasons used, the number of schools taking part in each season, and the amount of games each team ends up playing, our dataset contains a lot of valuable information. It should give us a large enough sample size to work with that will allow us to accurately make predictions based off of it in our model.

While the data given by Sports Reference is very accurate and comes from a reliable source, there are some issues with the tracking of this data that must be considered as they may cause future implications. Sports data in general is often inaccurate as it is all recorded by scorekeepers who are prone to both human error and having biases. This is especially true for basketball, as there is a lot of ambiguity in the definitions of certain statistics such as assists, blocks, and steals, meaning different scorekeepers interpret them in different ways (Williams 2017). This causes some team scorekeepers to either undercount or overcount certain statistics, and skew the results away from their true values, causing bias in the tracked data. A particularly notable example of scorekeepers tracking data incorrectly due to human error is seen in the National Hockey League. Popular arena Madison Square Garden (MSG), home to the New York Rangers, and its scorekeepers are known to track shot locations far too close towards the net. This causes them to extremely overstate the quality of the shots in games played at MSG, measured with the statistic "Expected Goals" predicting the chance of a shot resulting in a goal, relative to the other arenas in the league (McCurdy 2020). As a result of this, the Rangers team statistics are often extremely biased and skewed to show them having stronger offensive and goaltending performances, and weaker defensive performances (McCurdy 2020). It is likely that situations like this also occur in college basketball, especially considering some NCAA basketball games are played at MSG as well, and would cause our data to be biased. Finally, there is known racial bias in basketball and data may not be tracked ethically. Studies have shown that the referees in the National Basketball Association call fewer fouls against players who share the same race as them, and more fouls against those who are of a different race, ultimately causing players of the same race to score more points in the game being played (Wolfers and Price 2012). This likely goes for NCAA basketball as well and may be even worse since it is a lower and non-professional level of play. Considering there typically are more white referees than other races in basketball, due to the makeup of the US population, this is problematic and unfair towards players of colour. This racial inequality could mean that teams with a larger white makeup will have team statistics skewed positively since they will receive more calls and thus score more points, and those with more players of colour will have negatively skewed team statistics.

Even with these biases, the data should not be affected too substantially, and information can still be learned by analyzing it. This can be done using packages such as Tidyverse (Wickham et al. 2019), Dplyr (Wickham et al. 2022), modelsummary (Arel-Bundock 2022), and more in R (R Core Team 2020), where the data will be looked at graphically and modeled to answer the research question and tell a story about which team statistics play the biggest factor in team success. Before getting to the modeling, we must first take a deeper look into the data that we will be using to do so.

Table 1: 10 Observations from dataset of NCAA Basketball team statistics

Year	School	Over .500	SRS	Pace	O-rating	3pa Rate	TS %	TRB %	AST %	STL %	BLK %	TOV %	ORB %
2017-18	Austin Peay	TRUE	-3.28	70.1	104.7	0.294	53.2	52.2	49.5	9.4	8.9	16.8	35.7
2017-18	Green Bay	FALSE	-8.51	74.1	102.2	0.401	53.7	48.7	53.4	9.2	7.8	16.2	26.3
2018-19	South Carolina Upstate	FALSE	-15.30	69.4	96.9	0.436	51.5	46.7	56.0	9.2	8.1	17.3	26.2
2018-19	Xavier	TRUE	9.61	66.5	107.0	0.374	55.3	53.4	56.3	8.1	10.6	16.5	32.2
2019-20	Duke	TRUE	22.55	72.9	111.4	0.315	55.9	53.8	52.1	11.2	13.6	15.2	34.8
2019-20	Eastern Michigan	TRUE	-3.12	68.1	95.4	0.378	50.9	48.6	42.9	14.4	11.6	18.3	27.8
2020-21	Central Connecticut State	FALSE	-15.12	72.0	95.2	0.376	51.6	44.7	51.7	9.6	7.1	17.6	23.4
2020-21	Marshall	TRUE	5.70	73.1	108.7	0.410	57.1	48.7	53.2	10.6	13.2	14.6	23.8
2021-22	Eastern Kentucky	FALSE	-6.75	73.0	106.5	0.484	52.5	47.5	54.7	13.7	10.1	13.3	29.7
2021-22	Lipscomb	FALSE	-8.42	71.2	104.0	0.417	56.3	49.8	59.1	6.2	7.0	16.4	23.1

Table 1 allows us to look at the actual raw data as it contains an extract of 10 of the 1762 observations from the dataset being used. Two observations from each of the five seasons were randomly selected to be used. With that being said, each observation shown contains the team statistics for a given school in a given season or year. For example, row one shows the team statistics for the school team of Austin Peay in the 2017-18 season. Of the team statistics, two important variables to notice are SRS and Over .500, as they will be the team statistics used to identify a team's success. SRS, Simple Rating System, is a rating given to a team that takes into account both how much the team has been winning or losing their games, and the quality of opponents their games have been against (Kubatko 2008). This will be our response variable that is used to quantify team success, as it is a better indicator of a team's performance than something like straight-up wins or winning percentage that would typically be used. For example, if Team A beats the first place team by 10, and Team B beats the last place team by 1, although they both have 1 win, Team A's win is clearly more impressive and successful, and Simple Rating System takes this into account. Over .500 is a created boolean variable that is True if a team won more than half their games, or had a winning percentage over .500, and False otherwise. While once again winning percentage is not as good at showing team success as SRS and thus will not be used in the model, it is still a good indicator that will be used in visualizations to show the difference between the better teams that win more and the losing teams.

All the variables to the right of SRS in Table 1 are the numeric team statistics that can potentially be explanatory variables in the model. The team statistics included in the model will only be the most important ones that clearly play a role in making a team better or more successful. There were additional statistics in the original dataset, but they were removed as they did not reasonably seem like they had an impact on how well a team did, so they were removed. The remaining statistics in order are as follows; pace, offensive rating, 3-point attempt rate, true shooting percentage, total rebound percentage, assist percentage, steal percentage, block percentage, turnover percentage, and offensive rebound percentage. Once only the main and important statistics from this list are chosen they will be defined and looked at further, but for now we will focus on analyzing our response variable and its relationship with these team statistics.

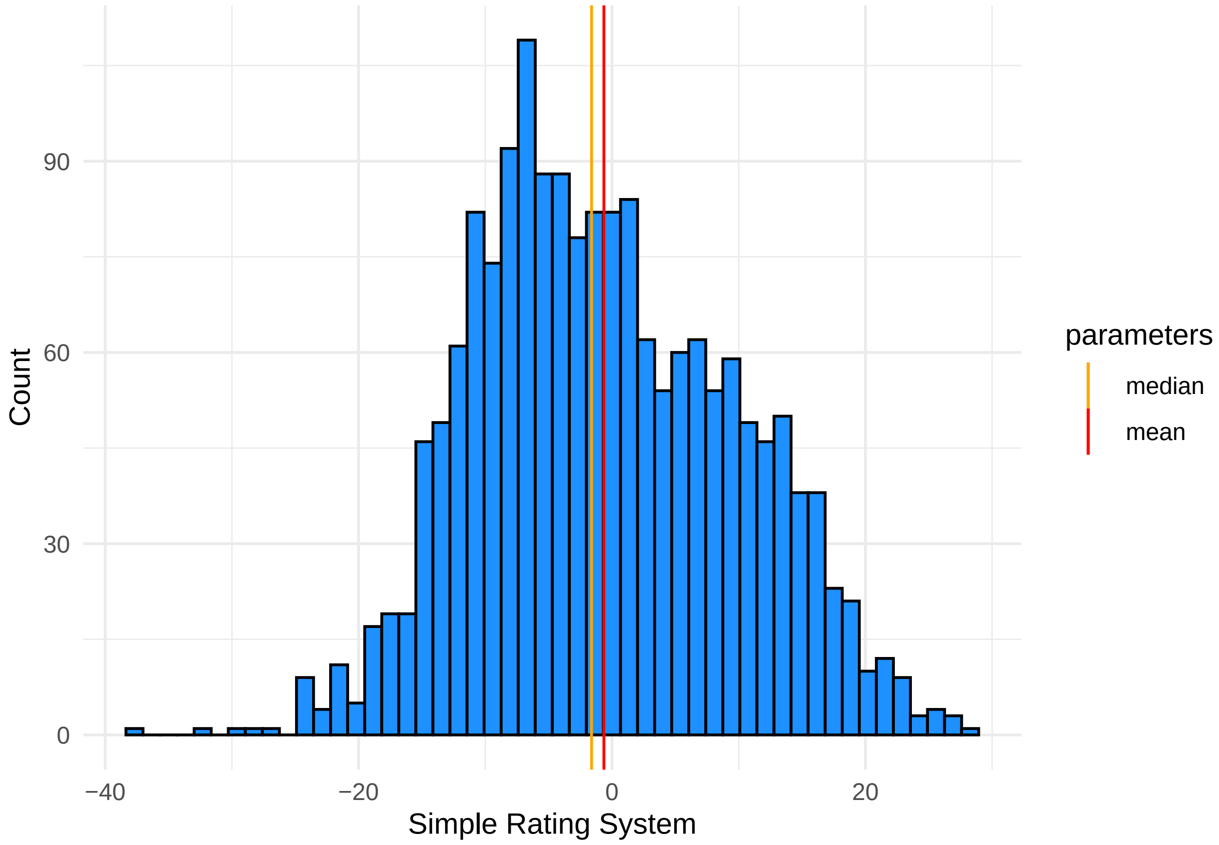

Figure 1: Distribution of Simple Rating System for NCAA Basketball teams

Figure 1 shows the distribution of our response variable simple rating system. It can be seen that the variable is close to resembling a normal distribution based on its shape, and even a standard normal distribution since it is almost centered around 0. The distribution is slightly skewed to the right, evident by the larger mean (red line) than median (orange line), and the mode being lower than both. There are also some noticeable outliers on the negative side of the distribution, which may be dragging the mean down lower than it would be without them. Even with the skew, it seems that there are close to similar amounts of teams with positive and negative SRSs, which makes sense as the better half of the league will often beat the worse half of the league, making the better teams have positive ratings and the poorer teams have negative ratings. Overall, the fact that SRS is almost a normal distribution if not for the slight skew is a good sign that our model will be able to predict the variable successfully. One of the most important model assumptions is the Normality assumption stating the model's errors must be normally distributed. This is more likely to occur when the response is normally distributed, and thus it is likely we will not have to make any transformation on our variable to get more accurate and trustworthy results.

Now, we must compare simple rating system to each of our potential predictor variable team statistics, and Figure 2 shows just this. The figure contains 10 different scatter plots, with each one showing the relationship between simple rating system and a different team statistic. This will allow us to see which team statistics play a role in whether a team is successful and has a higher simple rating system value or not. If SRS increases as a certain team statistic increases, they have a positive relationship and the variable will be considered for the model. The same goes for variables that have a negative relationship with simple rating system, meaning SRS decreases as the variable increases. The plots where the points are randomly scattered all throughout suggest that there is no relationship between simple rating system and a given team statistic. Additionally, the points are coloured in turquoise if the team the point is representing had a winning percentage over .500 in the given season, and a light shade of red if the team had a winning percentage under .500. This gives us an additional piece of insight on whether a variable leads to team success or not. If the blue points representing teams that win more often are more towards a certain side of the plot, it means that a given team statistic causes teams to win more when it is higher or lower. For example, if all the blue dots are on the right of the plot showing the relationship between SRS and statistic x, and all the red dots are on the left, it means more successful teams have higher values of statistic x and worse teams have lower values. Again, this is not as informative as just looking at simple rating system, but it adds a visual that can easily be seen and interpreted which should for the most part back up what the relationship between SRS and another variable is telling us. So, overall, we are looking for variables to include in the model that have a noticeable upwards or downwards trend with SRS, and also for plots where the colours are more separate towards different sides.

Looking at the plots, there are four variables that immediately stand out and have a noticeable trend with simple rating system, and should be considered for use in the model. These variables are offensive rating, true shooting percentage, total rebound percentage, and turnover percentage. Offensive rating is simply defined as the amount of points scored per 100 team possessions, so the positive relationship with SRS makes sense as the more points a team scores the more likely they will win games and win them by a larger margin. Next, true shooting percentage is essentially the percentage of all the shots a team takes that they actually make. This was chosen for use over normal shooting percentage as unlike the normal statistic it includes free throws, an important part of the game, and also is adjusted based on where the shot was taken from, whether it was a free throw, 2-pointer, or 3-pointer, through the use of different weightings in the calculation formula (Jacobs 2017). This means the stat truly captures just how strong or weak a team is at overall shooting. Once again, the positive relationship with SRS makes sense as better shooting teams will score more shots and be able to keep up with other good scoring teams, leading them to more success. Futhermore, total rebound percentage, which shows the percent of possible rebounds a team grabs, also seems to have a positive upwards trend with simple rating system. Although on the surface this is less obvious than what was determined for the other variables, it means that teams that grab more rebounds on both ends of the court will have more success. Finally, turnover percentage is defined as the amount of team possessions that end in a turnover. The lower this stat is the better, and thus unlike the other variables, it has a negative relationship with SRS, where teams that turn the ball over more often have lower ratings. This is also reasonable as fewer turnovers should lead to both more scoring opportunities for the team while simultaneously leading to fewer chances for the opposition.

Each of these relationships are further evident by looking at the difference between where the blue dots representing teams above .500 are and where the red points representing teams below .500 are. In the plots for offensive rating and total rebound percentage, the two colours for the most part are almost completely separated, where blue points are more towards the right and red dots are on the left. This means that teams with higher values for these two stats win more games than those with lower values, which agrees with the variable's positive trends with simple rating system. Although the teams with better records are a little more mixed in with the teams with worse records in the plots for true shooting percentage and turnover percentage,

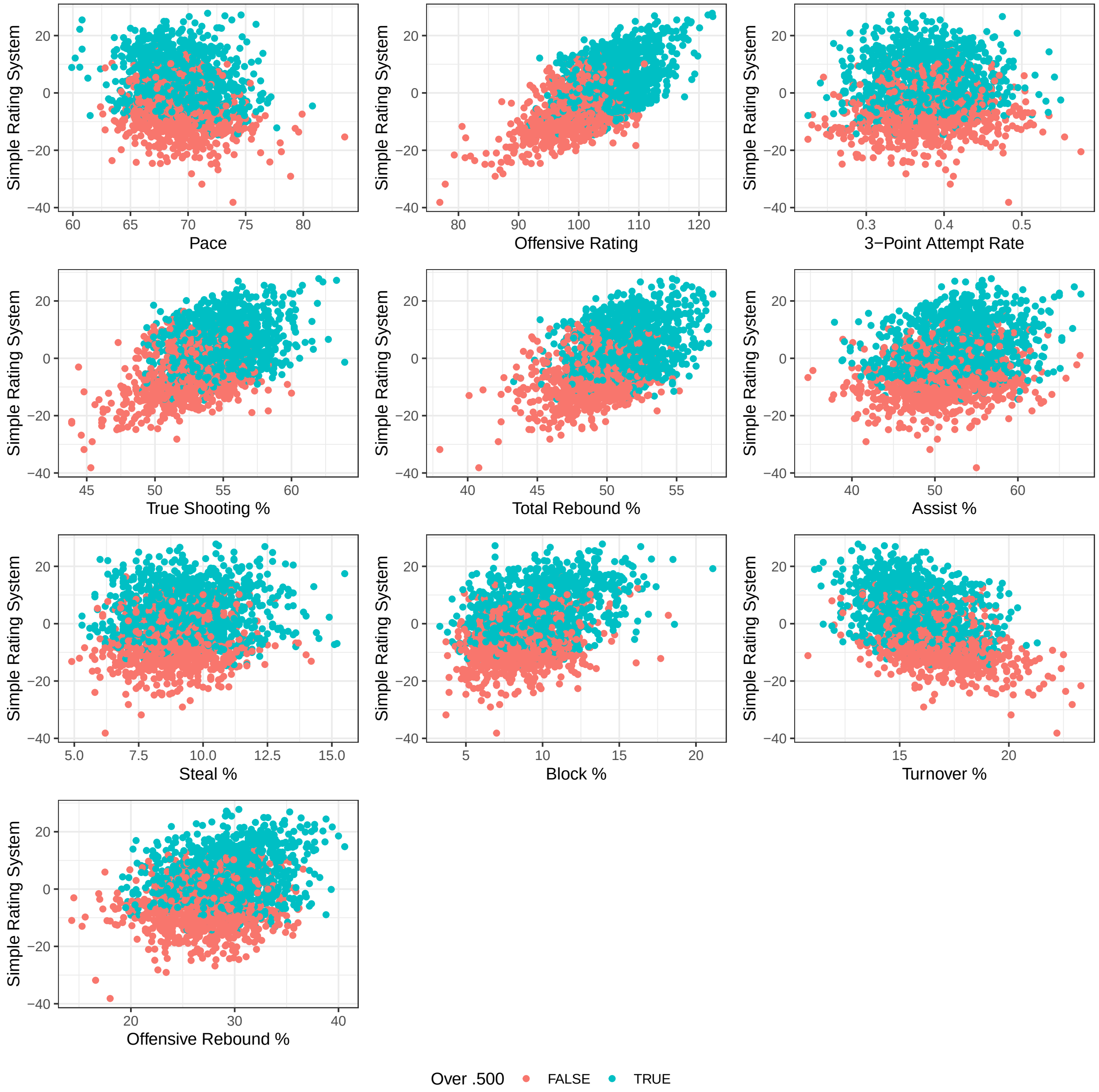

Figure 2: Relationship between Simple Rating System and several adcanced team statistics, coloured by whether a schools winning % was above .500 or not

they still differ greatly towards the far ends of the plots. The right side of true shooting percentages plot with SRS is mostly blue points showing more winning teams, and the left side has more red points showing losing teams, backing up the positive trend of the plot. The same goes for the plot with turnover percentage, although the colours and their sides are flipped since it has a negative relationship with SRS.

The remaining 6 variables from the dataset that were plotted, pace, 3-point attempt rate, assist percentage, steal percentage, block percentage, and offensive rebound percentage, do not seem to have a noticeable trend in their relationships with simple rating system. The points for each plot are randomly scattered throughout, and the structure of the plot does not change or differ as the variables increase or decrease, meaning simple rating system is not affected by any of these team statistics. The two colours showing teams above and below .500 also do not differ greatly in the plots, and are mostly just stacked on top of each other, with the blue dots at the top, which is of course caused by simple rating system as opposed to the variables on the x-axis. This further proves none of these variables have any real and consistent effect on how successful a college basketball team is. The one variable that you can make a case for is offensive rebound percentage as there seems to be a slightly positive trend with SRS in its plot. Despite this, the teams with better records and worse records differ minimally in the plot, telling us that the stat does not affect team success too greatly. Also, offensive rebounding percentage is likely to have a lot of multicollinearity with total rebounding percentage since they have the same definition except offensive rebound percentage only includes offensive rebounds. Due to this, it would be best if we do not include both and just stick with total rebound percentage which had the stronger relationship.

So, none of these 6 variables will be considered for the model, and only offensive rating, true shooting percentage, total rebound percentage, and turnover percentage will potentially be used to predict simple rating system, our response variable, and help us determine what leads to team success. Each of these variables are once again numeric, with offensive rating being a rate per 100 possessions, and the remaining three other variables all being in the form of percentages.

Table 2: Summary statistics of important variables

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
srs	1418	0	−0.6	10.3	−38.2	−1.6	27.8
o_rtg	295	0	102.9	6.1	76.9	103.0	122.3
ts_percent	154	0	53.8	2.8	43.9	53.7	63.9
trb_percent	146	0	50.3	2.6	38.0	50.3	57.6
tov_percent	108	0	16.2	1.8	10.8	16.2	23.3

Table 2 shows several summary statistics of each of these statistics that will be considered for the model, including the response. Unique shows the number of different unique values for each team statistic. Missing shows the percent of values missing in the data for each variable, and each one is 0 since all NA values were removed in the initial cleaning process of the dataset. The noteworthy statistics included in the table are mean, standard deviation, minimum, median, and maximum. We once again see how simple rating system has a mean and median quite close together and is centered very close to one, showing how the distribution is only slightly skewed and is close to being a normal distribution. With a standard deviation of just over 10, there is a decent amount of spread in the distribution but nothing too substantial. The larger max and min may cause some implications in the future but only minimally.

All the potential explanatory team statistics seem to have little to no skew at all as the mean and median of each variable are extremely similar. They are actually the exact same for total rebound percentage, with a mean and median of 50.3, and for turnover percentage, with a mean and median of 16.2. Although they are not identical, the mean and medians for offensive rating and true shooting percentage are nearly the same as well, as they only differ by a tenth for both, as they are 102.9 and 103 respectively for offensive rating, and 53.8 and 53.7 for true shooting percentage. Each variable also seems to have fairly reasonably low standard deviations, meaning no variable has a spread too high that will affect our model. Based on the maximums and minimums, there does not seem to be any crazy high or low values that are serious outliers and would skew our results too drastically either, so overall the data looks quite good.

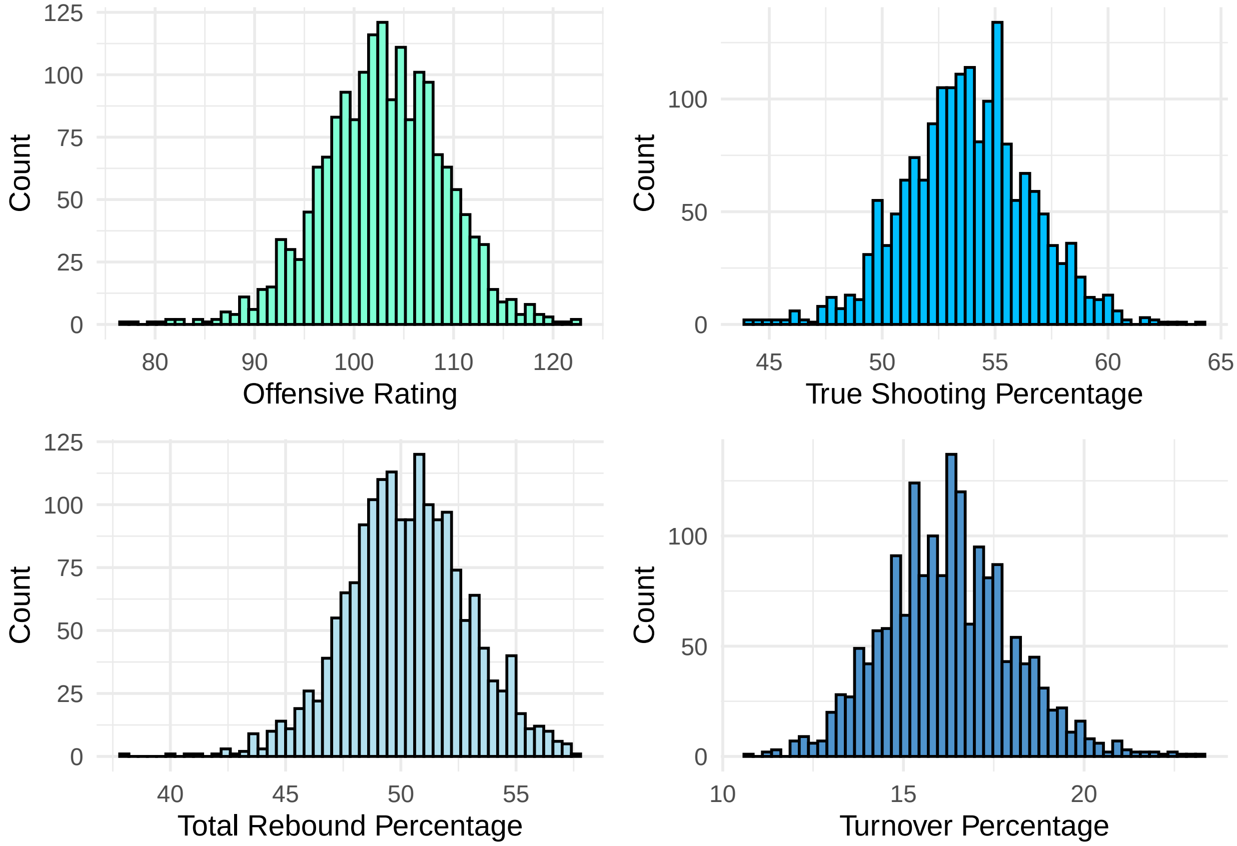

Figure 3: Distributions for important team statistics that affect a college basketball teams success

Figure 3 shows the distributions for each of these important team statistics that may be used as predictor variables, which allows us to more easily and visually see what was learned in the summary table. It can be seen that each variable seems to very closely resemble normal distributions, where they are centered around their mean and are fairly symmetric on both sides of it. Of course, they are not perfectly symmetric as that would be near impossible, but the overall shape of the distributions is still very similar to that of a normal distribution. Offensive rating and total rebound percentage have a slight tail to the left of the distributions skewing them slightly to the left, and true shooting percentage and turnover percentage are the opposite where they are slightly skewed to the right due to their tails. However, the tails are quite small and not too noticeable, so the distributions on not skewed very much, which is evident by the means and medians of each variable being the exact same or nearly the same. Additionally, although there are some points at the end of the tails that may alter the distribution slightly, we can now more clearly see that there does not seem to be any substantially large outliers for any of the variables that would skew results too drastically. All in all, these team statistics should make for great explanatory variables and should not cause any issues in our model.

Now that our dataset and the variables within it have been looked at and analyzed through the use of numeric tables and visualizations, we understand what exactly we are working with and can being to create our model that will help answer our research question.

3 Model

While we do have four strong candidate explanatory variables that may be used in the model, we still must decide what combination of these variables will make for the best possible model. First, it is a good idea to think logically if each of these variables should be included in the model, and the variable in particular we should analyze is offensive rating. Based on the visualizations analyzed in Section 2, it is clear that there is a relationship between offensive rating and a teams simple rating system, but that does not necessarily mean there is direct causality. Of course, if a team scores more points and has a higher offensive rating, they will win more games and have a higher simple rating system value, but a high offensive rating is not what is causing a team to have success. It is instead underlying statistics that cause a team to score more points and thus have a higher simple rating system value. These underlying statistics, such as true shooting percentage, total rebounding percentage, and turnover percentage, affect both offensive rating and simple rating system and thus make it seem like offensive rating and SRS have a causal relationship. For example, teams with a higher true shooting percentage will hit more of their shots causing them to score more points, leading to a higher offensive rating, and thus causing them to win more games by larger amounts, and have a higher SRS value. So, these statistics are the true variables affecting simple rating system, and offensive rating is actually just acting as a mediator for the relationship between them. This relationship can be shown through the use of directed acrylic graphs (DAGs) created with DiagrammeR (Iannone 2022), seen in Figure 4.

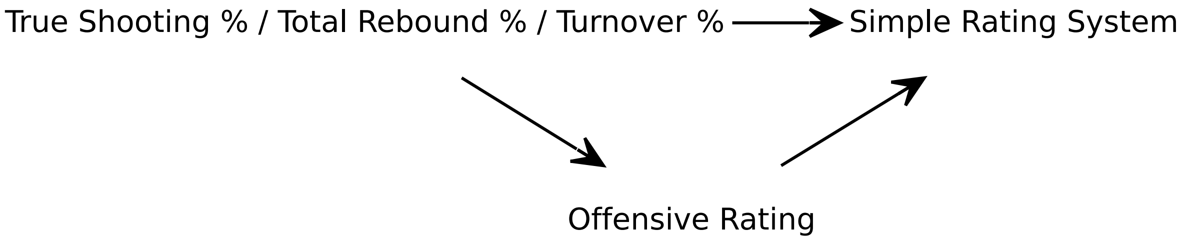

Figure 4: A DAG showing Offensive Rating as a mediator

In the DAG, it can be seen how true shooting percentage, total rebound percentage, and turnover percentage all affect and have causality with simple rating system. It can also be seen that offensive rating affects simple rating system too, but it is also affected by the three variables listed above and is being used as a mediator between these variables and simple rating system. As this mediator, offensive rating is explaining exactly why these three variables affect simple rating system (Bhandari 2021). The three variables affect team success because the three variables cause teams to either score more or less points on offense, giving them either a better or worse offensive rating, and thus causing them to either win or lose more games, which is essentially what the DAG is explaining. As a result of this, offensive rating should not be used in the model as the mediator will affect the strength of the relationship between the true predictor variables and simple rating system.

With the final three predictor variables, there are 7 possible combinations that can be used as predictors for a linear regression model. We can then test these models against one another to determine which model is the best at predicting simple rating system, and the variables used in the model will be the main team statistics that affect how much success a college basketball team has. Before creating the models and testing them, we must first split our data into two groups, a training group and a testing group. This is done to ensure we do not overfit our model, which would make it difficult for our model to be applied to other datasets and accurately make predictions on different data (Alexander 2022). The training data will be used to actually build and fit the model, and thus is also used to get our parameter estimates. As a result of this, the data should be split in a way that the training dataset is larger, allowing it could make more accurate estimates, and thus an 80:20 split has been chosen. The testing data will be used later on to validate our model and see if it can be used to make predictions on different datasets than the one used to build it.

Using our training dataset, the seven different models were created, one with all three final variables included, three with two of the variables included, and three where one variable is used as a single predictor. We can then obtain the values of several different goodness of fit and maximum-likelihood measures that tell us how well each model fits the data. Table 3 displays the estimated coefficients of each potential final model, and the several different goodness of fit measures that were aforementioned, where each column represents a different model. Based on the measurements in the table, it seems as if the largest model that uses true shooting percentage, total rebound percentage, and turnover percentage is the best possible model.

Table 3: Comparing goodness of fit measures of several potential models predicting Simple Rating System. Column names represent which variables are included in each model.

	mod_ts_trb_tov	mod_ts_trb	mod_ts_tov	mod_trb_tov	mod_ts	mod_trb	mod_tov
(Intercept)	−98.62	−152.76	−44.00	−58.93	−98.57	−107.96	44.20
	se = 5.42	se = 4.96	se = 5.43	se = 4.53	se = 4.60	se = 4.41	se = 2.26
ts_percent	0.92	1.27	1.44		1.82		
	se = 0.08	se = 0.08	se = 0.08		se = 0.09		
trb_percent	1.63	1.66		1.94		2.13	
	se = 0.08	se = 0.09		se = 0.08		se = 0.09	
tov_percent	−2.04		−2.10	−2.41			−2.77
	se = 0.11		se = 0.13	se = 0.12			se = 0.14
R2	0.512	0.402	0.361	0.461	0.244	0.297	0.221
R2 Adj.	0.511	0.401	0.360	0.461	0.244	0.296	0.220
AIC	9602.5	9886.8	9980.8	9739.6	10 215.0	10 113.4	10 257.6
BIC	9628.8	9907.8	10 001.8	9760.6	10 230.7	10 129.1	10 273.4
RMSE	7.29	8.07	8.34	7.66	9.07	8.75	9.20

The estimated coefficients are not of too much use when comparing the models, and the coefficients of the final model will be analyzed in 4. The standard errors of the estimates are of use to us though, as a smaller standard error means the model has estimated the parameters more precisely. The model with all three variables used has the lowest standard error for each of the slope coefficients, showing it is doing the best job at estimating these parameters. Its intercept standard error is quite high on the contrary, which is expected since it is the largest model with several predictors in use. Next, the R^2 and R^2_{adj} measures also point to the largest model being the best. The measures describe how much of the variance in our response is explained by the model, with R^2_{adj} penalizing larger models since more predictors causes R^2 to increase. Even with the penalty, the model with all three variables still has the highest value at 0.511, which is 0.05 points higher than the next best model, meaning it fits the data much better than the rest. Ideally, that number would be closer to 1, but explaining over half of the variation still makes the model fairly successful. Additionally, the Aikake's Information Criterion (AIC) and Bayesian Information Criteria (BIC) log-likelihood based measures also point towards the most complex model being the best. These measures tell us how well the model fits the data, while also penalizing models with more parameters, with BIC having a stronger penalty than AIC (S. 2010). Even as the largest model and being penalized the most, the model will all three variable once again has the best results, as both its AIC and BIC are much lower than each of the other models' values, and it is not particularly close. While it already fairly evident the full model is the best, looking at the Root Mean Squared Error (RMSE) only backs up this claim further. The RMSE is the variance of all a model's residuals, and shows us how close the models predicted fitted values are to the actual points in our data (Grace-Martin 2013). A lower RMSE means a model has less error and fits the data much better. The largest model once again has the lowest value at 7.29, which is significantly better than the next best model, and further shows that this model fits the data the best and should be used as the final model. The only model that comes relatively close to the full model is the one with total rebound percentage and turnover percentage, as it proved to be the second best model for each of the measurements, with the second highest R^2 and R^2_{adj}, and second lowest AIC, BIC, and RMSE values. While its measurements were fairly good and it was considered as a possibility for the final model, the model with all three variables still was better in each measurement and by a significantly large amount as well, making it the clear and obvious choice to be used as our final model.

So, now that we know which variables will be included in the final and best model, the form of our final model can be seen below.

$$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1}x_1 + \hat{\beta_2}x_2 + \hat{\beta_3}x_3 + \epsilon$$

where:

- $\hat{y_i}$ represents the predicted value of SRS given specific values of the predictor variables
- $\hat{\beta_0}$ is the estimated intercept coefficient
- x_1 represents the predictor variable of True Shooting Percentage
- $\hat{\beta_1}$ is the estimated slope coefficient for True Shooting Percentage
- x_2 represents the predictor variable of Total Rebound Percentage
- $\hat{\beta_2}$ is the estimated slope coefficient for Total Rebound Percentage
- x_3 represents the predictor variable of Turnover Percentage
- $\hat{\beta_3}$ is the estimated slope coefficient for Turnover Percentage
- and ϵ is an error term which should be 0 assuming our model is unbiased

Essentially, the model works by plugging in values of our predictor variables for the x's, and then based these given values and the intercept and slope coefficients, the model will predict what a team's Simple Rating System Value will be. In Section 4, we will analyze the results of the model and its parameters, plug the coefficient parameters into the model, and interpret what exactly the model means.

4 Results

Table 4: Summary of parameter values and statistics for linear model predicting Simple Rating System using various team statistics

	Estimated Coefficient	95% Confidence Interval	P-Value
Intercept	-98.62	(-109.25, -87.98)	1.38e-66
True Shooting %	0.92	(0.77, 1.07)	5.39e-32
Total Rebound %	1.63	(1.47, 1.78)	2.01e-84
Turnover %	-2.04	(-2.27, -1.82)	5.13e-64

Table 4 displays the results of our model's estimated parameter statistics. It displays each of the estimated coefficient values, the 95% confidence interval for each estimated coefficient, and a p-value stating the significance of each variable. The parameters can be interpreted in order for us to better understand the relationship between the three predictor variables used in the model and Simple Rating System, or a college basketballs team success, shown in the model, and this is especially true for the estimated coefficients. The intercept coefficient of -98.62 tells us what a team's SRS will be if all predicted variable values are equal to zero. So, if in a given season a team hits 0% of their shots, grabs 0% of all possible rebounds, and turns the ball over in 0% of their positions, they will have a Simple Rating System value of -98.62. This is a reasonable intercept value as if a team hits no shots they will score no points, and by grabbing no rebounds they will give the other team many chances to score, so the team will lose many games by a large amount, resulting in the predicted very low SRS value. Of course, it is near impossible for a team to have a season where they do not make a shot, do not grab a rebound, and never turn the ball over, but this is just a baseline value that helps us understand how the model works.

Next, the slope coefficient for the true shooting percentage variable is 0.92, meaning that when all other predictor variables are held at a fixed value, if true shooting percentage increases by 1% then their simple rating system will increase by 0.92. Total rebound percentage has a slope coefficient even larger at 1.63. This coefficient value of 1.63 can be interpreted as how much a team's simple rating system increases by when their total rebound percentage goes up 1% and all other variables are fixed. Both of these coefficients make sense as simple rating system will logically increase if a team shoots more efficiently and scores more points or if they grab more rebounds and get more attempts to score that their opposition, which was already determined by looking at the relation graphically earlier. The estimated slope coefficient for turnover percentage is different from the other as it is negative, with a value of -2.04. This lines up with what was seen in the graphical analysis as simple rating system and turnover percentage have a negative relationship, since turning the ball over more will allow the other team to score more and lead to a worse SRS. This negative value tells us that when a team's turnover percentage increases by 1% and all other variables are held fixed, the team's simple rating system value will decrease by 2.04 units.

We can plug all these coefficients into the $\beta's$ in the formula explained in Section 3 to get the completed formula for our final model,

$$\hat{y_i} = -98.62 + 0.92x_1 + 1.63x_2 - 2.04x_3$$

. Doing so makes our model much easier to interpret. In order to estimate what a team's simple rating system value would be, you plug in their true shooting percentage for x_1, their total rebound percentage for x_2, and their turnover percentage for x_3, and the output of $\hat{y_i}$ would be their rating. For example, if a team has a true shooting percentage of 50%, a total rebounding percentage of 40%, and a turnover percentage of 20%, the model estimates that their simple rating system would be $\hat{y_i} = -98.62 + 0.92(50) + 1.63(40) - 2.04(20) = -28.22$. This shows how the model can successfully predict a team's success with the use of advanced team statistics such as our predictor variables that were used, answering the research question.

Table 4 also contains 95% confidence intervals. These intervals give us a range of plausible values for each estimated coefficient, and we are 95% certain that the true population parameter value is within this interval. For example, the true shooting percentage slope coefficient can reasonably be any value between 0.77 and 1.07, and our estimated coefficient value of 0.92 for true shooting percentage falls right in the middle of this range. There is a chance this interval does not capture the true value but it is unlikely, as if we took 100 samples of college basketball data and created models, only 95% of the models and their confidence intervals for the estimated coefficients would capture the true population values. Fortunately, each of the estimated slope coefficients confidence intervals are quite narrow, as the lower and upper bounds only differ by 0.3, 0.31, and 0.45 respectively. The narrowness of the intervals is good as it means that our estimated coefficients are most likely quite accurate, and any coefficients and their confidence intervals that are estimated from other samples would give values very similar to the ones obtained from our sample that was used (Bassett 1999). The intercept coefficient also has a fairly narrow confidence interval as well, from -109.25 to -87.98, even if it may not seem like it relative to the other coefficients. This is because it is not dealing with percentage variables like the slope coefficient and instead has to do with simple rating system which has a wider range of values.

The final value in the summary table is the p-value of each estimated coefficient, which tell us how significant each estimated coefficient is. A low p-value, and one below 0.05 in particular, for a variable's coefficient means that there is evidence against a null hypothesis stating the coefficient should be zero as the variable does not affect the response in the presence of the other variables in the model. Looking at the p-values for our coefficients, they are all extremely small and are very close to 0, meaning the null hypothesis is rejected and thus each of the coefficients are significant and belong in the model. However, this information must be taken with a grain of salt, as p-values are difficult to work with since they assume all model assumptions are satisfied and the data was properly collected and cleaned (Alexander 2022). As a result of this, no decisions were made based on these values, but they do provide us with a piece of information that allows us to further see that the decisions we did make are likely good ones due to their low values.

The relationship between simple rating system and our three explanatory variables can also be visualized despite being a 4D relationship, thanks to the package plot3D (Soetaert 2021).

Figure 5 shows this relationship through the use of a 3D plot, which is technically a 4D plot based on the

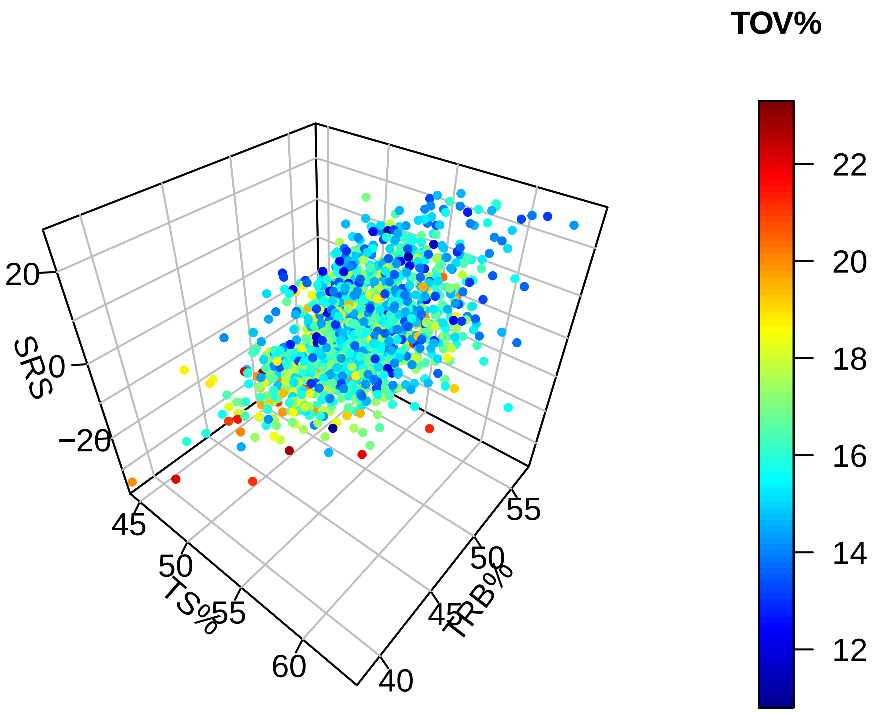

Figure 5: 4D plot showing the relationship between the models response variable, Simple Rating System, and its three predictor variables

use of a colour gradient to show a 4th variable. The vertical z-axis represents our response variable SRS, the x-axis closest to the viewer represents true shooting percentage, and the remaining y-axis on the right represents total rebound percentage. Also, the colour gradient is used to represent turnover percentage where each point is coloured in differently based on its turnover percentage value, giving the plot its 4th dimension.

Although it is difficult, it can be seen that the overall trend of the points moves from the bottom left of the plot toward the top right of the plot. So, as true shooting percentage increases along the x-axis and the points move right, simple rating system increases since the points are also moving upwards. Additionally, as total rebound percentage increases up the y-axis, simple rating system increases as well since the points are moving from the bottom of the z-axis to the top. Finally, the points towards the bottom of the plot where SRS is low are mostly red, orange and yellow. Then as SRS increases, the points gradually become more green and then more blue. This final dimension shows that as turnover percentage decreases and the colours go from red to blue, simple rating system increases. Each of these relationships shown by the 4D plot line up with what has been seen in both the graphical analysis and the analysis of the model's parameter statistics. It should be noted that all three of these relationships are in the presence of the other predictor variables, and when you combine them all together you get the upwards trend and colour change from red to blue seen in the plot, which shows us the overall relationship estimated by the model.

Before going over what we have learned from our model, there are two things that must be done. First, we need to check to see if all of the assumptions that were made with respect to the model were all satisfied, or else there will be some serious limitations to our findings. Then, we need to use our testing dataset to validate our model and see if these results can be used to make predictions on other datasets.

The first assumption is the linearity assumption where all explanatory variables must have a linear relationship with the response, and this was confirmed to be true in Section 2 through the use of scatter plots. The explanatory variables also should not have a strong linear relationship with each other, and this was also already considered and dealt with earlier through the use of DAGs in Section 3. The following three assumptions will be checked using the plots in Figure 6. The next assumption is the independent errors assumption, as errors should be independent of one another and show no correlation. The bottom three plots in Figure 6, which show the relationship between each predictor variable and the model's residuals, prove that this assumption is satisfied. This is because the points are all randomly scattered around a horizontal line at 0 on the y-axis for each one of the three plots, and there are no separated groups of points or noticeable trends that would suggest any correlation in the residuals. The third assumption is the homoscedasticity of errors assumption which simply means the model's residual errors have constant variance. The plot in the top right of the figure shows the residuals have constant variance since the points have an upwards trend along an invisible diagonal line, meaning that variance is mostly similar and constant across all fitted values. Finally, the last assumption states that the errors should be normally distributed. The histogram in the top left of Figure 6 is used to see if this assumption is satisfied, and it should show a normal distribution if it is. The histogram has a distribution very similar to the distribution of our response variable simple rating system, which we previously determined that despite being slightly right-skewed, the distribution is very close to being a normal distribution centered around its mean. This seems to be the case here as well, and while it is not perfect, it is still very close to being so, meaning it should not cause many issues with our model if any at all, and the assumption can be considered satisfied.

Finally, our model must be validated using the testing dataset. One way to do this is by making predictions on the test dataset, using the values of the explanatory variables in the data. We then calculate the RMSE using the errors between the predicted values and actual values of the testing data, and compare it to the RMSE of the original model built with the training data (Arnholt 2021). After using the model to make predictions on the testing data, it was found that the predicted values had an RMSE of 7.74. This is very similar to the original models RMSE based on the training data, which was 7.29. Since these values are very similar to one another, it means that our model did a good job at predicting values using the testing dataset, and thus it should also successfully be able to make predictions on other datasets, like for example future college basketball seasons. This means that the model has successfully been validated using the testing dataset, and we can now analyze what exactly was learned from our model.

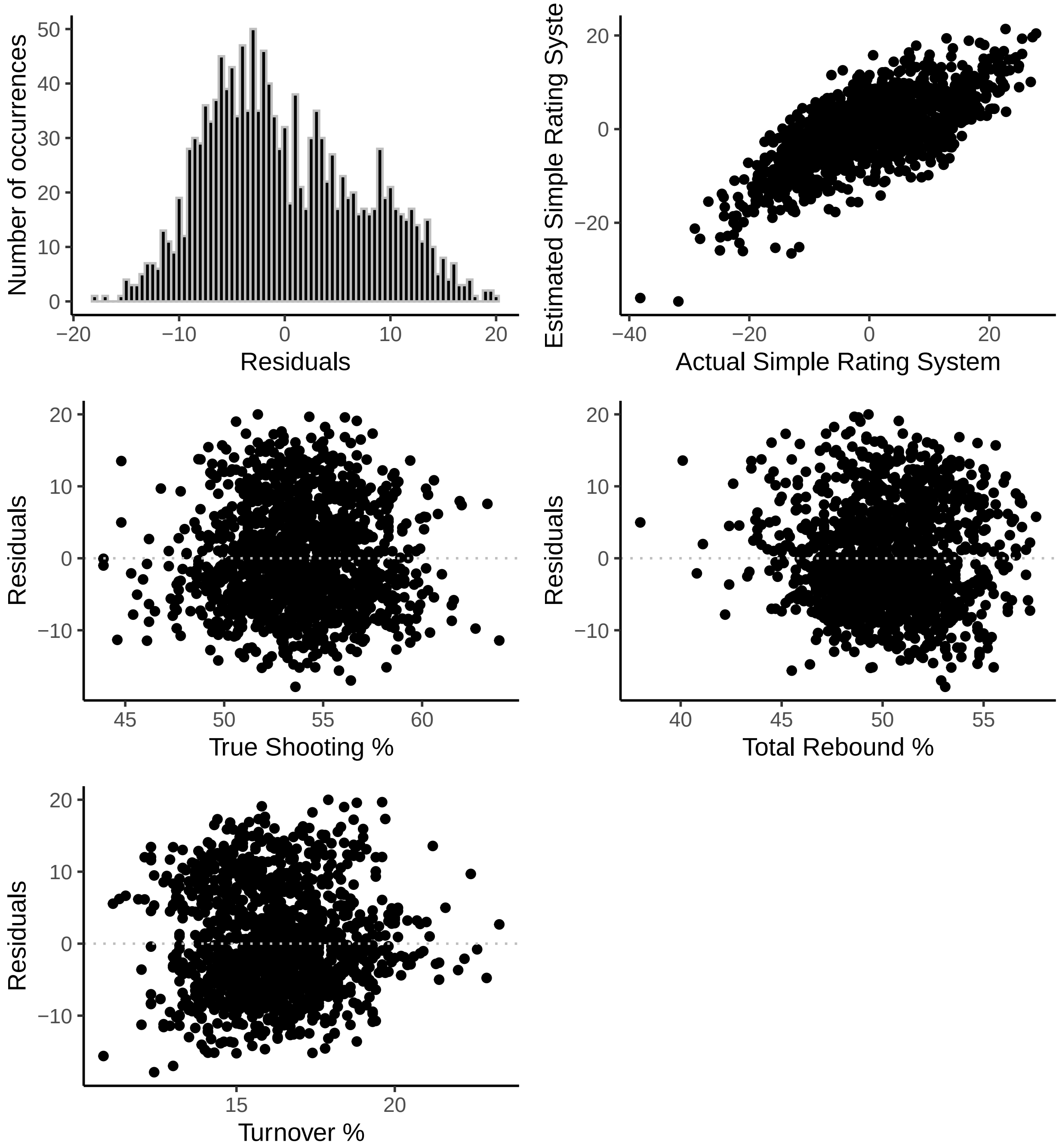

Figure 6: Residuals and Fitted Value plots for linear regression model predicting Simple Rating System using a training dataset on college basketball statistics

5 Discussion

5.1 Simple Rating System and its predictors

In this paper, we used several statistical methods in order to analyze a dataset containing team statistics on NCAA Division 1 men's basketball schools from the past five seasons. This was done to find out which specific statistics played an important role in how successful a team was. Through the use of several visualizations such as Figure 2, it was evident that several advanced statistics had a clear relationship with a team's Simple Rating System (SRS), a value that determines how successful a team is based on the point differential of their games and the quality of their opposition. These advanced statistics were first checked for multicollinearity through the use of DAGs, seen in Figure 4, before being used to create several different linear regression models. Each model had a different combination of explanatory variables used to predict our response, SRS, and the models were tested against each other to determine which one was the best. This was done by comparing several goodness of fits tests across each of the models, and seeing which linear relationship best fit the used data.

As a result of these methods, it was determined that there were three team statistics that could be used to predict how successful a team would be. These three variables are a team's true shooting percentage, total rebounding percentage, and turnover percentage. Their inclusion in the model suggests that each of the team statistics plays a large role in how successful a team is since together they all have a relationship with simple rating system. Teams that hit more of their shots and have a higher shooting percentage will score more points, increasing their simple rating system value. Simple rating system will also improve when a team has a higher rebounding percentage, since grabbing more rebounds than the opposition will simultaneously give a team more chances to score will giving the opposition less chances. Turnover percentage has a negative relationship with simple rating system, as teams with less turnovers than others will have possession of the ball more often and for longer, giving them more time and chances to score than their opposition.

While other statistics do also play a role in how good a team is, these statistics in particular play the biggest roles and have the strongest relationship with simple rating system, which is why they were the only variables included in the final model. In the presence of each other, these three statistics and their values will be the main reasons why a team is successful or not. Teams with high true shooting and total rebounding percentages, and low turnover percentages will be the most successful teams in the league, as their simple rating system values will be the highest. On the contrary teams with low true shooting and total rebounding percentages, and high turnover percentages will have low SRS values and struggle to succeed. So, our research question has been answered, and team success can be predicted using advanced team statistics.

5.2 March Madness Brackets, Betting, and Coaching

There are several pieces of information that can be learned and obtained from these findings. To recall, a big reason that this research on college basketball was completed is due to the popularity of the March Madness tournaments, and specifically the creation of brackets by many. The goal was to obtain information that would help people make better predictions when creating their bracket. Since we have learned that true shooting, total rebounding, and turnover percentages are the most important statistics in college basketball, fans can put a large emphasis on these statistics when making decisions in their brackets. Fans could simply substitute a team's values in the current season into the model formula described in 4 and calculate the team's expected simple rating system value. Then, they could choose their bracket based on which teams have the higher value of SRS. Considering the strength and success of the model, doing such a thing would result in a pretty good bracket as the stats should do a very good job at predicting success.

Of course, a fan doing such a thing is not very realistic, as there is little fun in this and many would not have access to the model, but just knowing what was found through the creation of the model would also be very helpful to fans in many other ways. For starters, year after year there are several crazy upsets in the March Madness tournament that shock the world as no one saw them coming (Gleeson 2022). However, using our findings may give fans a one-up over others when it comes to predicting these upsets. If we look at the values of the statistics used as predictors in our model for teams who had poorer records in the season and are thus ranked lower in the tournament brackets, we could get insight on which teams are ranked too low and are more likely to upset better teams. For example, if a fan realizes a low ranked seed has a very good shooting percentage, rebounds the ball often, and does not turn the ball over, based on the model the chances of this team having success in the tournament and upsetting a higher-ranked seed is quite high. This team may have gotten unlucky throughout the season which resulted in their poor record and rank, but based on the underlying important statistics we can realize that they should be a better team, and would be a good underdog pick in the bracket. The same goes for high-ranked teams with poor values of these predictor statistics who got lucky in the season and won a lot of games, as realizing this would allow fans to see picking them to go far is a bad idea. Apart from predicting underdogs, if a fan is having a hard time picking between two teams in a specific match-up and cannot decide which one to pick, they could turn to comparing the two team's true shooting, total rebound, and turnover percentages in order to come to a decision. If the two teams have equal shooting and turnover percentages but one team has a much higher rebounding percentage, the better rebounding team has a higher chance of being successful based on the model, and thus this information would help the undecisive fan come to a conclusion and pick the better team.

For the same reasons the model and its findings are helpful to fans making brackets, it would also be very helpful to people who bet on games. As of recently, sports betting has seen a massive surge in popularity due to the ease of betting online (Hudson 2021). Those who bet on college basketball games can get a huge monetary benefit by using the findings of this paper. By focusing in on teams values for the predictor statistics in the model, bettors can recognize which odds set by oddsmakers in Vegas are most beneficial to them. For example, if one team has very low odds of winning the game but has good underlying statistics and the model predicts they should be a more successful team than the team with high odds of winning, then bettors could place money on the underdog team and have a very high chance of receiving a massive payday. It is very difficult for people to beat the oddsmakers in sports betting, but using the knowledge obtained from this paper can give bettors an advantage that many people do not have.

Finally, although the findings were intended to be for fans, they could also be helpful to the coaches of college basketball teams who can or will play in March Madness. Since we found that good shooting, good rebounding, and limited turnovers lead to team success, coaches can go and recruit players who will help their team obtain better results in these specific categories. They could build their team with players who shoot the ball well, are tall and have advantages rebounding, and are safe with the basketball to improve their simple rating system value, as predicted by the model, and have more success. This could also be done by teaching their players how to improve in these areas both individually and as a team through game plans and practice drills. Using the findings in the paper could give coaches a huge benefit over others, increasing their chances of making their dreams come true and winning March Madness.

5.3 Limitations

While the findings in this paper are mostly trustworthy and extremely helpful, there were several limitations that may have affected our results. As mentioned in 2, there are several biases, both racial and accidental, in the data that was used. As a result of scorekeepers and referees being prone to human error, and the overall ambiguity of many basketball statistics, it may have caused the data for some teams to be skewed. For example, if two scorekeepers in two different arenas have different ideas of what a turnover is, they will track the stat differently. As a result of this, one team who has a scorekeeper who tracks more turnovers relative to scorekeepers of other teams could have a turnover percentage much higher than it should be. This would mean our model predicts they should have a lower SRS and be a worse team even though they do not turn the ball over and their scorekeeper just believes they do. This could occur for any other of the team statistics we looked at, meaning many values in the data are likely biased and skewed, which could have affected our model's estimates and even which variables were included in the model. Also, referees are known to have racial biases and tend to give more calls to players and teams of the same race as them. If a team is predominantly white and has a majority of white referees throughout the season, their team stats will be inflated, and the model will predict that the team is better than they are due to the biases. Due to this, people will use these findings and favour the team that benefited from the biases, which may work until the team gets a referee of a different race who does not help them out, and the team's actual abilities are exposed.

Apart from the biases, there were a few other limitations with the dataset that was used. As everyone is well aware, the COVID-19 pandemic has had a large effect on nearly every aspect of life, and that includes college basketball. Over the past three seasons, hundreds of games had to be canceled because of the virus. 2020 saw the entire March Madness and many conference tournaments go unfinished as the virus first began to spread. It only got worse from here, as no Ivy League schools participated in the following season (Diamond 2021), and in only half of the most recent season, over 70 schools had to shut down their basketball operations (Borzello 2022). Overall, as a result of all these cancellations, many games went unplayed, and thus the sample size of our dataset, and the amount of total games played which are used to calculate values in the data, was much smaller than it would have been in normal seasons. The smaller and affected sample size may have led to our model's coefficient estimations being slightly less accurate than they could have been. Finally, the dataset that was used only contained a certain amount of team statistics. There are many simple and advanced statistics out there that were not accounted for when creating our model, and thus a key predictor of simple rating system could have been overlooked. Despite this, the statistics analyzed were the main and most common advanced statistics, so this should not be too much of a worry.

5.4 Next Steps

Thankfully, there are steps that can be taken in the future to enhance our findings and mitigate the effects of these limitations. To account for the biases in our dataset, we could only use a sample of the data that is more consistent throughout when it comes to scorekeeping and refereeing. For example, only games played on neutral sites could be taken into account to minimize the amount of scorekeeper bias. Also, to remove racial bias in the data, a sample of only games where the refereeing team was diversified and multiple races were equally represented. By doing such a thing, we will ensure that all data is tracked with consistency and that no team is favoured over another.

Furthermore, as more seasons are played in the future, we will obtain more complete data that will give us a larger sample size. This means our dataset would no longer be affected as badly as it was due to the COVID-19 pandemic, assuming our world improves over time in regards to the virus. To further complement this, datasets from various other sources could be searched to find data on advanced statistics that were not included in our dataset. Together, this would give us a dataset that is much larger and contains many more variables, ensuring that our model estimates are as accurate as possible and that all variables that have a relationship with team success are actually included.

A Appendix

A.1 Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The datasets used and combined in the paper were created by the site Sports Reference. Sports Reference provides data for the purpose of helping fans answer their many questions about sports data and statistics. Their objective is to help fans "grow their appreciation, understanding, and love of the game" by providing them with the informative datasets and resources free of cost (Sports Reference 2022).
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The website and its many datasets were first founded by Sean Forman who was a PhD math student at the University of Iowa (Cannella 2002), and the basketball section in particular was created by Justin Kubatko. There is now a large team of workers who help create, update, and maintain the datasets on the website.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - The websites and datasets were self-funded by the creator.
4. *Any other comments?*
   - No.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - Every instance or observation in the dataset contains team statistics of an NCAA Divison 1 men's basketball school in a given season. Since multiple seasons were used in the dataset, schools may appear up to 5 times in the dataset, with different team statistic values depending on the specific season the observation belongs to.
2. *How many instances are there in total (of each type, if appropriate)?*
   - 1776
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - The dataset contains every instance of a school participating in a college basketball regular season anywhere from 2017-22. Since seasons before 2017 were not used, the dataset is a sample of all team statistics in the history of NCAA basketball. Basketball has experienced many changes over time, and the sport is played much differently than it used to be, so statistics in the sample likely vary from the seasons not taken into account. Since the goal of the model created was predicting future results, using only recent seasons would make our estimate more accurate relative to upcoming seasons as opposed to if the entire population of team data, dating back to the 1800s, was used.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - Each instance consists of a variety of advanced team statistics of a certain school in a season, like for example how many points were scored, shooting rates, percentages, and more. Simpler statistics such as how many games each team won and lost were also included.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - Each instance of team statistics is labeled by two variables. One stating the school of the team,

and the other stating which season is being looked at in the observation.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
    - No, all team statistic values are filled out.
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
    - No, as there are no relationships between instances
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
    - No split was recommended, but the dataset was eventually split into two datasets when creating the linear model, one to train the model and make estimates, and the other to test and validate the model.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
    - No.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The datasets rely on taking data from the box scores, tracked by referees and scorekeepers, of each college basketball game. These will always exist as every game played must have one. The gamesheets are all held on to and archived by the NCAA, and provided to the public, so there are no restrictions.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - No.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - No.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - No.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - Yes, as the public could easily find the roster for each school team in a specific season for every instance of the dataset.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - No.
16. *Any other comments?*
    - No.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the*

data validated/verified? If so, please describe how.

- The data used to compile the team statistics in the dataset was directly observable as the statistics are provided to the public by the NCAA, and then by Sports References in a clear and easily accessible way.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
   - All statistics for each college basketball game are systematically tracked and recorded by referees and scorekeepers. For use in this paper, the data was scraped off of the Sports Reference website by exporting files into excel and reading them into R.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
   - A cluster sampling strategy was used where seasons were grouped into clusters of 5, and the cluster containing the most recent 5 seasons was used in order to use data that was would be the most representative of games being played in the near future.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
   - Thomas D'Onofrio, with no compensation as the project was completed for himself.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
   - The data has been collected by Sports Reference from 2017 to 2022, and the data was simply collected quickly all at once thanks to the work done by the website over time.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
   - No.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - The data was obtained via Sports Reference, who collected the data from the NCAA, who initially tracked and collected the data.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
   - All NCAA teams and players are well aware that the statistics of the games they play are being tracked and distributed. No notice was provided, but it is common knowledge among all athletes at all levels that this occurs.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
   - No.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - Consent was not obtained.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No.
12. *Any other comments?*
    - No.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing,*

tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
    - The data was cleaned where variables were removed, renamed, and also created.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - Yes, the raw datasets exported straight from Sports Reference can be found in the inputs folder at the GitHub page linked in the paper.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R was used, as well as the R packages tidyverse (Wickham et al. 2019) and janitor (Firke 2021).
4. *Any other comments?*
    - No.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Considering the popularity of the Sports Reference website, the datasets used to create the full dataset used in this paper have been used by many. One example of a paper using datasets from Sports Reference can be found at https://arxiv.org/pdf/2007.10550.pdf.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - No.
3. *What (other) tasks could the dataset be used for?*
    - The dataset has little use for tasks outside of analyzing the sport of basketball in many different ways.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - No.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No.
6. *Any other comments?*
    - No.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - The complete dataset used in the paper can be found on GitHub in the inputs folder.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - Apart from being on GitHub, the dataset will not be distributed in any other way.
3. *When will the dataset be distributed?*
    - The dataset is already available on GitHub
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - No. MIT License.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - No.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - No.
7. *Any other comments?*
    - No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Thomas D'Onofrio
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - thomas.donofrio@mail.utoronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - Not as of right now, but in the future instances from future seasons may potentially be added to the dataset to increase the sample size and have updated information. This would be done yearly at the end of every season, which is in April, and the updated datasets will be on GitHub.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - No.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
    - The older versions will not be maintained, as the dataset will just be updated over time. Previous versions of the dataset can be found on GitHub.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
    - Pull Request on Github.
8. *Any other comments?*
    - No.

References

Alexander, Rohan. 2022. *Telling Stories with Data*. https://www.tellingstorieswithdata.com/index.html.

Arel-Bundock, Vincent. 2022. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://vincentarelbundock.github.io/modelsummary/.

Arnholt, Alan T. 2021. *Machine Learning with Caret in R*. https://stat-ata-asu.github.io/MachineLearningToolbox/.

Baker, Alison. 2021. *The Most Watched Sporting Events in the World*. Road Trips: The Ultimate in Sports Travel. https://www.roadtrips.com/blog/the-most-watched-sporting-events-in-the-world/.

Bassett, Mary T. 1999. *Confidence Intervals - Statistics Teaching Tools*. New York State Department of Health. https://www.health.ny.gov/diseases/chronic/confint.htm#:~:text=What%20does%20a%20confidence%20interval,if%20the%20survey%20were%20repeated.

Bhandari, Pritha. 2021. *Mediator Vs Moderator Variables | Differences & Examples*. Scribbr. https://www.scribbr.com/methodology/mediator-vs-moderator/.

Borzello, Jeff. 2022. *How the Latest Covid-19 Spike Will Impact the Rest of the 2021-22 College Basketball Season*. ESPN. https://www.espn.com/mens-college-basketball/story/_/id/33000345/how-latest-covid-19-spike-impact-rest-2021-22-college-basketball-season.

Cannella, Stephen. 2002. *Seamheads, Rejoice Baseball-Reference.com Is the Ultimate Online Statistical Source for Fans with a Sense of History*. Sports Illustrated. https://vault.si.com/vault/2002/12/16/seamheads-rejoice-baseball-referencecom-is-the-ultimate-online-statistical-source-for-fans-with-a-sense-of-history.

Diamond, Dan. 2021. *After a Historic Layoff, Ivy League Basketball Is Back. Did Its Shutdown Go Too Far?* The Washington Post. https://www.washingtonpost.com/sports/2021/11/09/ivy-league-basketball-covid-shutdown-return/.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Geiling, Natasha. 2014. *When Did Filling Out a March Madness Bracket Become Popular?* Smithsonian Magazine. https://www.smithsonianmag.com/history/when-did-filling-out-march-madness-bracket-become-popular-180950162/.

Gleeson, Scott. 2022. *NCAA Men's Tournament Predictions: Four Smart First-Round Upset Picks of March Madness*. USA Today. https://www.usatoday.com/story/sports/ncaab/2022/03/15/ncaa-tournament-predictions-four-smart-first-round-upset-picks/7030388001/.

Grace-Martin, Karen. 2013. *Assessing the Fit of Regression Models*. The Analysis Factor. https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/.

Hudson, Molly. 2021. *More States Embrace Online Sports Gambling as Popularity Soars*. NBC News. https://www.nbcnews.com/politics/politics-news/more-states-embrace-online-sports-gambling-popularity-soars-n1285169.

Iannone, Richard. 2022. *DiagrammeR: Graph/Network Visualization*. https://github.com/rich-iannone/DiagrammeR.

Jacobs, Justin. 2017. *Relationship Between Ts*. Squared 2020. https://squared2020.com/2017/10/10/relationship-between-ts-and-efg/.

Kubatko, Justin. 2008. *The Simple Rating System*. Basketball Reference. https://www.basketball-reference.com/blog/indexba52.html?p=39.

McCurdy, Micah Blake. 2020. *Scorer Bias Adjustments*. Hockey Viz. https://hockeyviz.com/txt/scorerBias.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

S., Prabhat. 2010. *Difference Between Aic and Bic.* Difference Between Similar Terms; Objects. http://www.differencebetween.net/miscellaneous/difference-between-aic-and-bic/#:~:text=AIC%20and%20BIC%20are%20widely,two%20approaches%20of%20model%20selection.

Soetaert, Karline. 2021. *Plot3D: Plotting Multi-Dimensional Data.*

Sports Reference. 2022. *Sports Reference | Sports Stats, Fast, Easy, and up-to-Date.* https://www.sports-reference.com/cbb/seasons/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation.*

Williams, Rob. 2017. *SFU Study Reveals There's Scorekeeper Bias in the Nba.* Daily Hive. https://dailyhive.com/vancouver/sfu-study-nba-scorekeeper-bias.

Wolfers, Justin, and Joseph Price. 2012. *Racial Discrimination Among Nba Referees.* University of Pennsylvania. https://users.nber.org/~jwolfers/papers/NBARace(QJE).pdf.