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

A Linear Model Approach

Thomas D'Onofrio

22 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 team winning percentage 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 [insert statistics here once model has been created] are the most important team statistics that affect the amount of wins a team gets the most, 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 fans 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 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 2020 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 teams 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 [Insert statistics here once model is completed], play the most important factor in leading to team success. Thus, when filling out their brackets, fans can compare

[*] Code and data are available at: https://github.com/TDonofrio62/Paper-On-NCAAB-Stats.

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 with different races than them, 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 winning percentage. Section 3 includes the creation of a linear regression model that uses only the most important team statistics as explanatory variables to predict winning percentage, 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 is 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 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 1800's, 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 has 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 giving proper citation, meaning that using their information for research is ethical.

Of the websites 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 score keepers to either under count or over count 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 for 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 less fouls against players who share the same race as them, and more fouls against those who are 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 (**???**), Dplyr (**???**), MASS (**???**), 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 ust 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	0.532	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	0.537	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	0.515	46.7	56.0	9.2	8.1	17.3	26.2
2018-19	Xavier	TRUE	9.61	66.5	107.0	0.374	0.553	53.4	56.3	8.1	10.6	16.5	32.2
2019-20	Duke	TRUE	22.55	72.9	111.4	0.315	0.559	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	0.509	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	0.516	44.7	51.7	9.6	7.1	17.6	23.4
2020-21	Marshall	TRUE	5.70	73.1	108.7	0.410	0.571	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	0.525	47.5	54.7	13.7	10.1	13.3	29.7
2021-22	Lipscomb	FALSE	-8.42	71.2	104.0	0.417	0.563	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 teams 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 teams 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 team statistics that can potentially be explanatory variable 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.

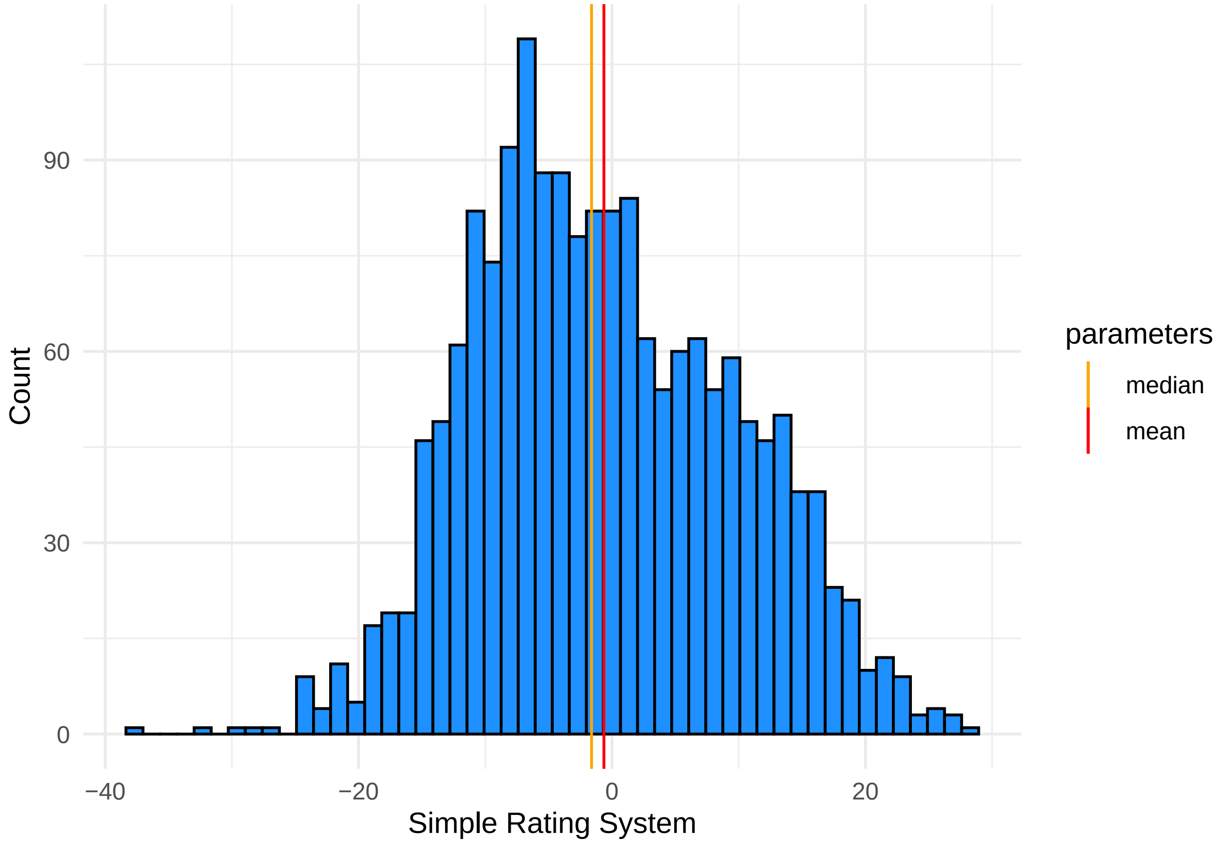

Figure 1: Distribution of Simple Rating System for NCAA Basketball teams

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 postive and negative SRS's, 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 the Normality assumption stating the models 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 another team statistics. 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 on 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 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 that 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 possesstions 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 less turnovers should lead to both more scoring opportunities for the team while simultaneously leading to less 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 variables 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, 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

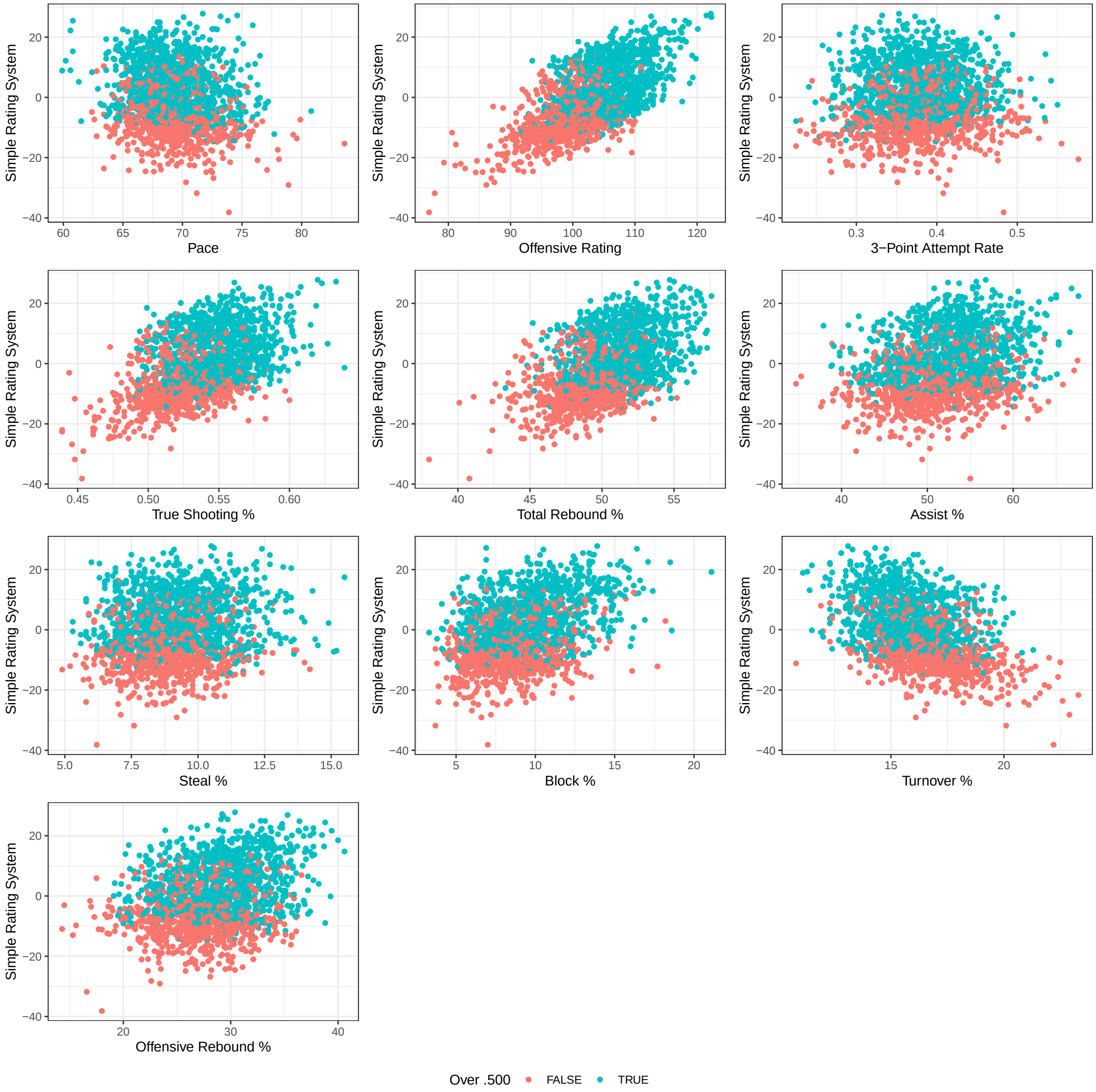

Figure 2: Relationship between Simple Rating System and several adcanced team statistics, coloured by whether a schools winning % was above .500 or not

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 throughtout, 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 varibales have any real and consistent affect 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 slight 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 the 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.

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	0.5	0.0	0.4	0.5	0.6
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

3 Model

$$Pr(\theta|y) = \frac{Pr(y|\theta)Pr(\theta)}{Pr(y)} \tag{1}$$

Equation (1) seems useful, eh?

Here's a dumb example of how to use some references: In paper we run our analysis in `R` (R Core Team 2020). We also use the `tidyverse` which was written by (**???**) If we were interested in baseball data then (**???**) could be useful.

We can use maths by including latex between dollar signs, for instance θ.

4 Results

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional details

References

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/.

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/.

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/.

Sports Reference. 2022. *NCAA Seasons Index.* https://www.sports-reference.com/cbb/seasons/.

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.