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

A Linear Model Approach

Thomas D'Onofrio

31 March 2022

Abstract

First sentence. Second sentence. Third sentence. Fourth sentence.

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 each NCAA Division 1 school 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 many games a team wins. These statistics were then used to create a linear model that attempts to predict winning percentage, 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 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.

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

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

2 Data

Our data is of penguins (Figure 1).

```
## Warning: It is deprecated to specify `guide = FALSE` to remove a guide. Please
## use `guide = "none"` instead.
```

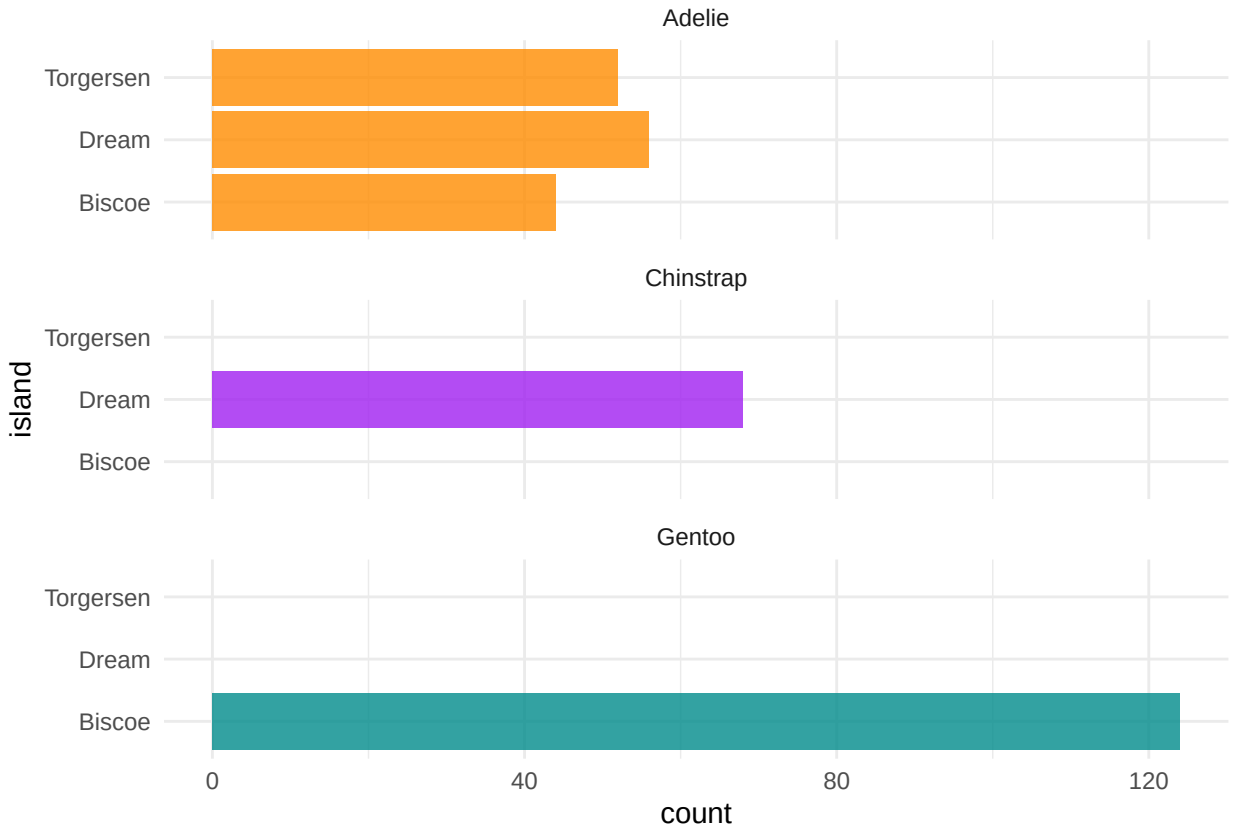

Figure 1: Bills of penguins

Talk more about it.

Also bills and their average (Figure 2). (Notice how you can change the height and width so they don't take the whole page?)

```
## Warning: It is deprecated to specify `guide = FALSE` to remove a guide. Please
## use `guide = "none"` instead.
```

Talk way more about it.

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 (**thereferencecanbewhatever?**) If we were interested in baseball data then (**citeLahman?**) could be useful.

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

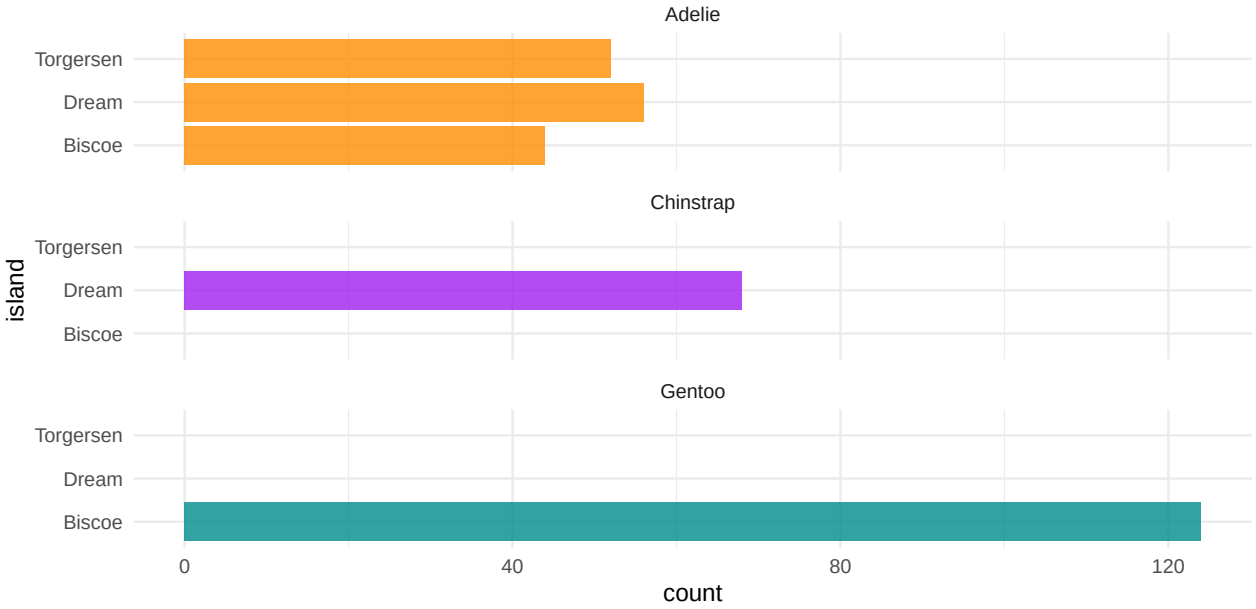

Figure 2: More bills of penguins

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/.

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/.