March Madness

A cumulative study of season data to predict tournament results.

Biased Estimators: Alexander Van Roijen, Molly Clark, Rocco Bavuso

May 4, 2018

Table of Contents

Introduction

March Madness is an annual NCAA Division 1 basketball tournament. The tournament occurs in the month of March, thereby earning its name. There are 64 teams in the tournament each year, seeded into four regions that each contain 16 teams. We will be taking a look at how far each team makes it into the tournament, based on their regular season data and prior tournament success.

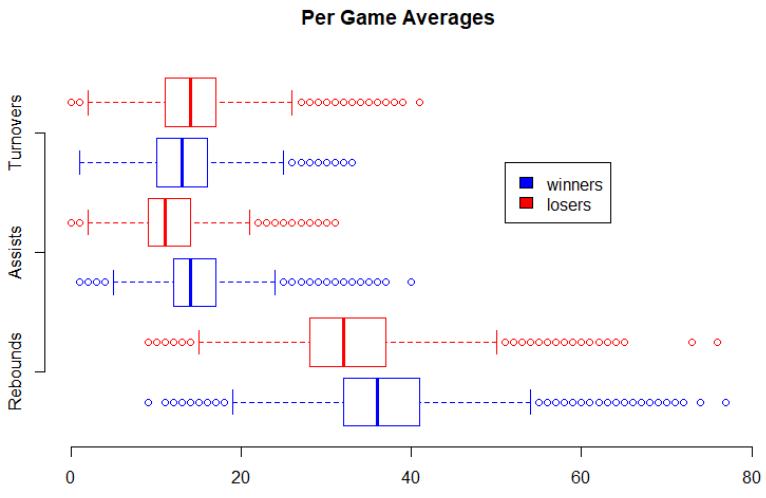
Data Summary

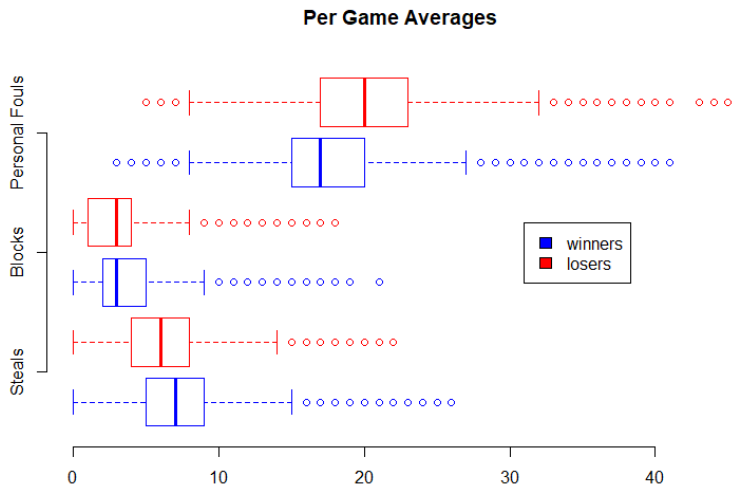
Our data consists of three main components. We have detailed regular season data from 2003 to 2016 for 365 teams, each with a specific numerical code. We have non-team specific game data from the years 1985 to 2016, comprised of winning and losing team scores, and other statistics. Lastly, we have tournament data for each playoff team from the years 2003 to 2016.

Pre-Analysis

Our main goal of the pre-anaylsis was to think about what predictors might be important in our future model. The ideal model would predict a team's success in the NCAA tournament based on regular season statistics and prior tournament success. In order to do so, we compared winning and losing team data for relevant statistics: field goals, assists, rebounds, etc.

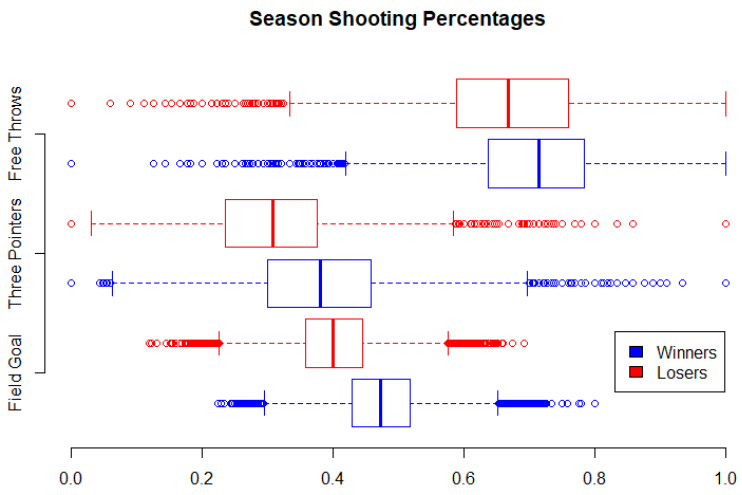
Pre-Analysis

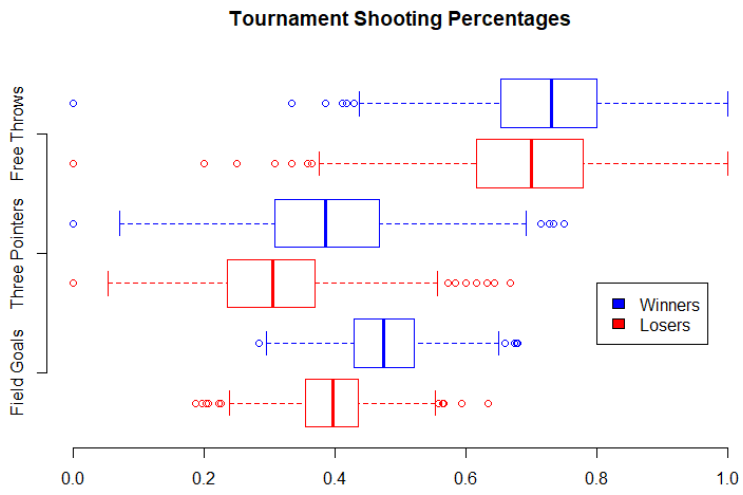




The data shows much of what would be expected, on average. Winning teams hold higher averages in "positive" stats such as assists, rebounds, and steals. On the other hand, losing teams have higher averages in "negative" stats such as turnovers, and personal fouls.

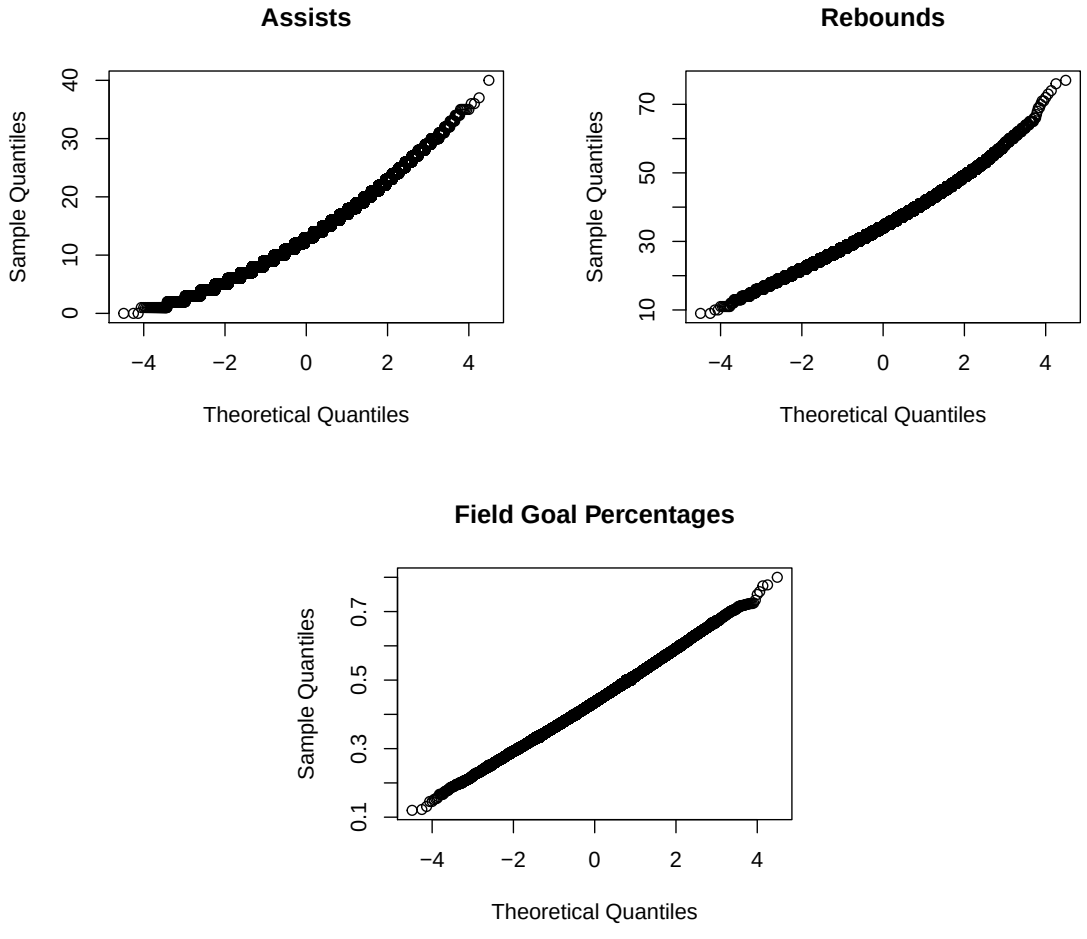
Pre-Analysis





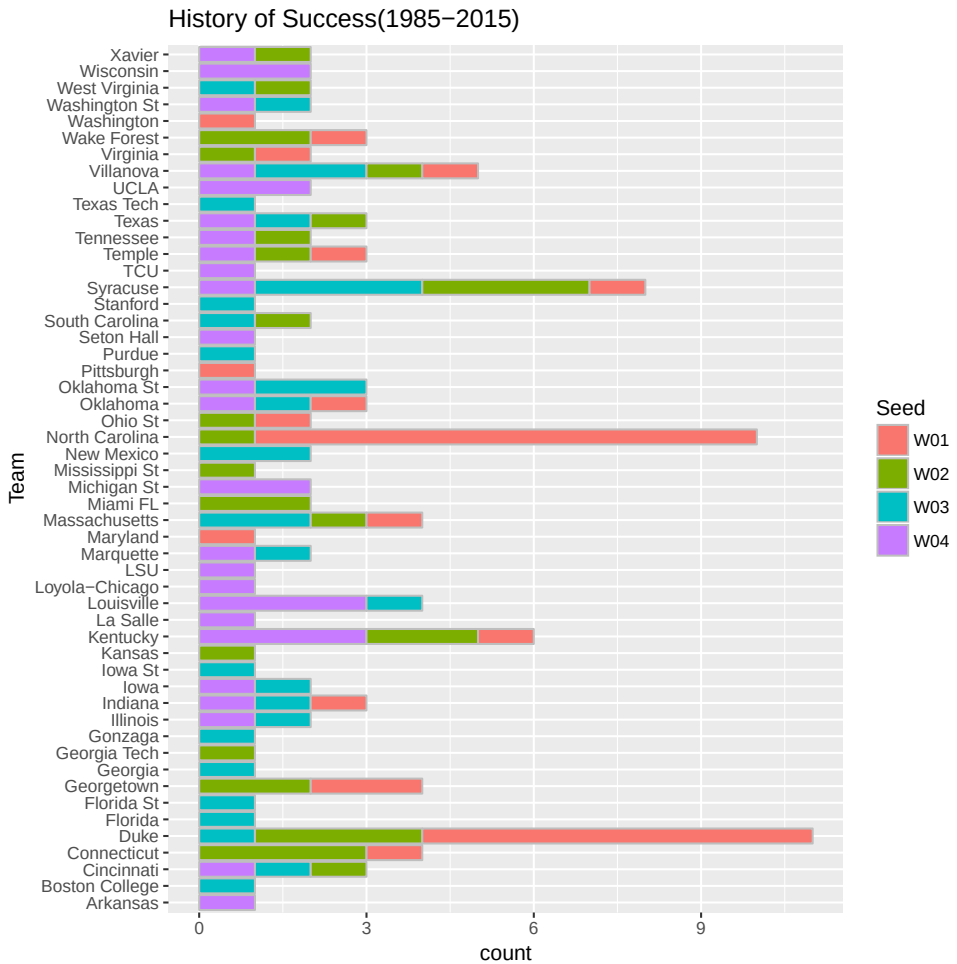
We also analyzed the shooting percentages from different the ranges, looking at season data separate from tournament data. As anticipated, the winning teams averaged higher in all categories, but there are a significant number of outliers in the Season data.

Pre-Analysis



We examined the distribution of parameters of interset. Unsurpisnigly , due to the large sample size, the variables are normally distributed.

Pre-Analysis



Initial Model

$$Y_{daynum} = \beta_0 + \beta_{conf} + \beta_{fgp} + \beta_{tpp} + \beta_{ass} + \beta_{rb} + \beta_{app} + \epsilon$$

We initially looked into this model, where the betas represented the following:

- conf: the conference the team plays in
- fgp: the field goal percentages
- tpp: three point shooting percentage
- ass: assists
- rb: rebounds
- app: prior tournament success

Processing

In processing our data to develop the model, we went through a number of steps:

- Separate out the needed data for each team for each year
- Eliminate unusable predictors, and search out missing values
- Calculate the averages of each team's stats for each year
- Compile the averages into a new csv file that we could use moving forward
- Develop a process for representing how far a team has made it in the tournament

Processing

We quantified our response variable by keeping record of how many games each team plays in the tournament each year, where better teams would be expected to play more games. This did leave us with a lot of missing values(all the teams who didn't make the tournament), so we experimented with models to find how to handle this.

Final Model

We looked into quite a few models before settling on a single one to work with. We analyzed the models for each year and found some variation from year to year caused by:

- Upsets in the tournament
- Great stats in the season but no slot in the tournament (can depend on conference)

Model Analysis

Questions?