Methods and Statistics Explanation

The main goal of this project is to use logistic regression to predict the outcomes for each first round tournament game. Using the outcome of each game for the 68 tournament teams (2253 games), my final model included offensive rating difference (ORtg) between the two teams, defensive rating difference between the two teams (DRtg), and strength of schedule rating (SOS) provided by sports-reference.com. By doing this, I created a model that used season statistics to predict the outcome of individual games (for the 68 tournament teams). I then applied this model to each first round tournament game. For example, by calculating the difference between Duke and North Dakota State's season offensive rating, defensive rating, and the SOS rating, the model calculates the probability of Duke winning the game.

This project began by downloading and merging each of the 68 tournament teams' regular season game logs into one dataframe of 2253 games. I then merged season overall statistics for each home and opposing team (e.g. every row with Duke as a home team has a variable ORtg and every game with Duke as an away team has a variable ORtg_Opp, both which are 112.9). Using a combination of plots, single-variable models, and tables to find trends between variables and winning, I narrowed my model down to 7 variables: ORtg difference, DRtg difference, SOS, Block difference, Steal difference, Pace difference, and Assist to Turnover ratio. Based on cross-validation and significance levels, I decided to keep the three variables that I used in my final model. Though the model tends to overpredict wins, it also has low collinearity in the predictors, and is able to account for many other variables in just the three that it uses.