Influence of Stadium and State Characteristics on Home Winning Percentage

Clara Livingston and Emily Kaegi

May 25, 2018

Introduction

Typically, when looking at statistics related to sports, researchers tend to focus on the dynamics of the team in question when predicting winning percentages. The goal of our research was to take analysis away from factors related to the team and investigate how factors relating to the location of the team's home stadium affects their home winning percentage. Most people are aware of the phrase "homefield advantage", but just how influential is a team's homefield in boosting their win percentages? SBNation and Sporting News both claim that home field advantage is real across most sports, especially the NBA. What appears harder to determine, however, is the cause of variability in this percentage between different sports and teams.

Instead of focusing on existance of "homefield advantage" we decided to look at what factors relating to a home stadium or arena might be correlated with winning percentage. Specifically, how do characteristics of the arena, franchise, fan base as well as state metrics affect a home team's winning. Are these characteristics consistent across all major league sports? We chose to focus on the NBA, NFL, MLB, and NHL since these professional leagues are the most popular in the United States.

Some of the most interesting potential predictors we wanted to explore were what percentage of a states population voted for Donald Trump in the 2016 presidential election. Does political leaning have any correlation with how well teams play at home? Also, using a metric of states' happiness level, we were curious if happiness is in anyway correlated with winning percentage. Do teams in happier states win more? While we cannot determine if this is casual or even the direction of the relationship (are states happy because their teams win or do players in happy states play better?) it is still interesting to explore.

Methods

Our data is pirmarily in two parts: data related to the state level characteristics and data related to the stadiums of the individual teams. At the state level we chose to focus on three variables related to general state dispositions. The first, population, was gathered from the World Population Review where we collected the approximate population of each state as of the end of 2017. In 2017 National Geographic did a survey of happiness level at each state and ranked the overall well-being of adults within each state on a scale from 1 to 5. Adults (18+) were asked to rate their well-being on a scale form 1-100 based on five primary categories: daily life, physical health, location, finances, and companionship. If a state averaged well in these categories, they would be ranked as a 5 for overall well-being and happiness. More on the study can be found in the National Geographic magazine. Finally, we decided to add a variable for the percentage of registered adults that voted for Donald Trump in the 2016 presidential election as recorded by the New York Times. As the country is heaviliy divided on thier beliefs, we were interested to see if this would affect the home team win percentage of the various teams within a given state.

The second portion of our data was related to the individual major league teams' home stadiums and stadium statistics for the 2017 seasons. We focused our study to baseball (MLB), basketball (MBA), football (NFL), and hockey (NHL). From Wikipedia, we were able to gather the name of the arena or stadium currently being used by active teams, capacity (in thousands), and year opened. It is important to note that across the internet there is some discrepancy in the actual capacity of some of these arenas. Generally, the capacity listed on Wikipedia is fairly accurate for our purposes and will be the variable used in this analysis. In addition

to this basic information, we decided to look at the average attendance percentage (people in attendance divided by capacity) fro the 2017 season for these teams. This data was collected through ESPN's database on sport statistics for the NFL, MLB, NBA, and NHL. Our final explanatory variable of interest was the franchise values of the teams in questions. We gathered this data from Forbes and recorded it in dollar value. Our interest in the variable primarily comes from the likelihood that teams with larger budgets likely have the flexibility to contract better players and spend more on facilities than teams that do not have this luxury.

Our dependant variable, home win percentage, was determined using data from Team Rankings. This site provided the home win-loss record for the 2017 seasons. We then used this information to calculate the win percentage by dividing the wins by total number of games played at home (value between 0 and 1). It is from this data we were able to build our two level model with random intercept.

Results

EDA Paragraphs

To discover the relationship between home game winning percentages and the variables of interest, we used logistic binomial regression. We observed variation in home win percentage by state, but did not want to include a variable for each state, we decided to include state as a random effect in our model.

Insert a couple sentences about modeling process description?

The heirarchical form of the final model is below. Estimates for the β parameters can be found in Table ??. Note that Sport Baseball is the baseline level.

Level 1 (team level):

$$logodds(homewins) = a_i + \beta_0 Capacity_{ij} + \beta_1 Attendence_{ij} + \beta_2 SportBasektball_{ij} + \beta_3 SportFootball_{ij} + \beta_4 SportHockey_{ij} + \beta_5 FranchiseValue_{ij} + \beta_6 YearOpened_{ij} + \beta_7 Capacity_{ij} SportBasektball_{ij} + \beta_8 Capacity_{ij} SportFootball_{ij} + \beta_9 Capacity_{ij} SportHockey_{ij} + + \beta_{10} Attendence_{ij} SportBasektball_{ij} + \beta_{11} Attendence_{ij} SportFootball_{ij} + \beta_{12} Attendence_{ij} SportHockey_{ij} + \beta_{13} Attendence_{ij} Capacity_{ij}$$

Level 2 (state level):

$$a_i = \alpha_0 + \beta_{14} Population_i + \beta_{15} Happiness_i + \beta_{16} Happiness_i Population_i + u_i$$

Where $u_i \sim Norm(0, \sigma)$. Our model predicted $\hat{\sigma} = 0.1434$

Discussion

Appendix

```
Warning: Column `State` joining factors with different levels, coercing to
character vector
```

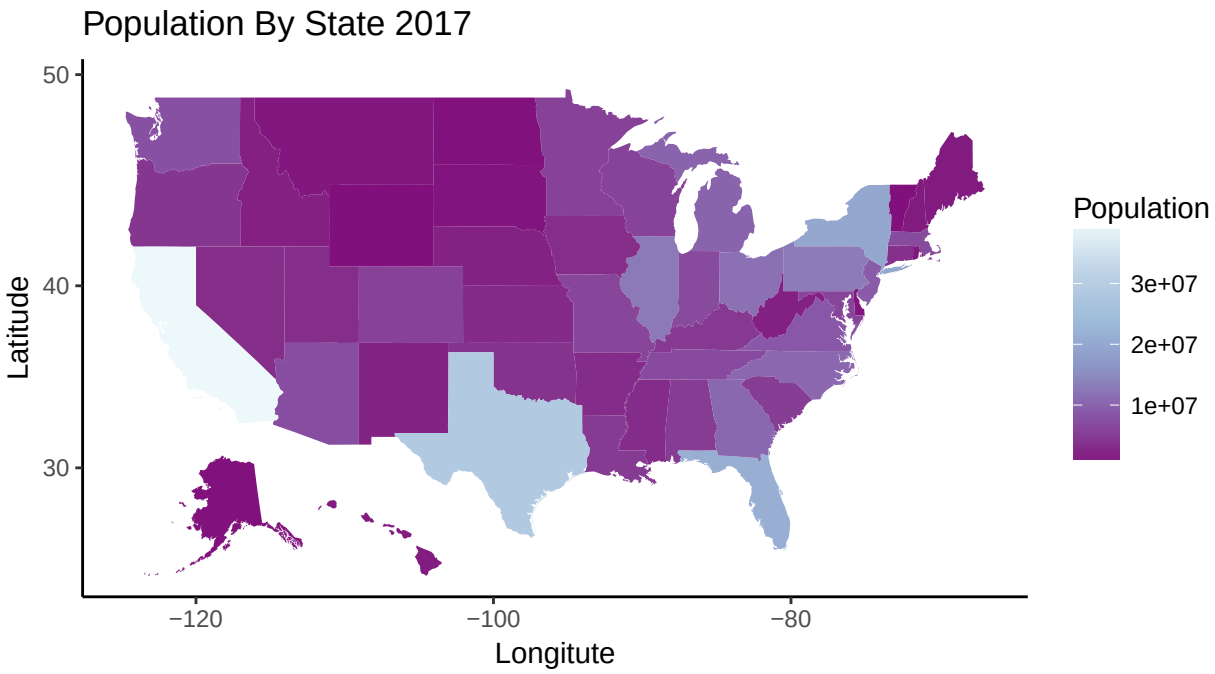

Figure 1: Caption Goes Here

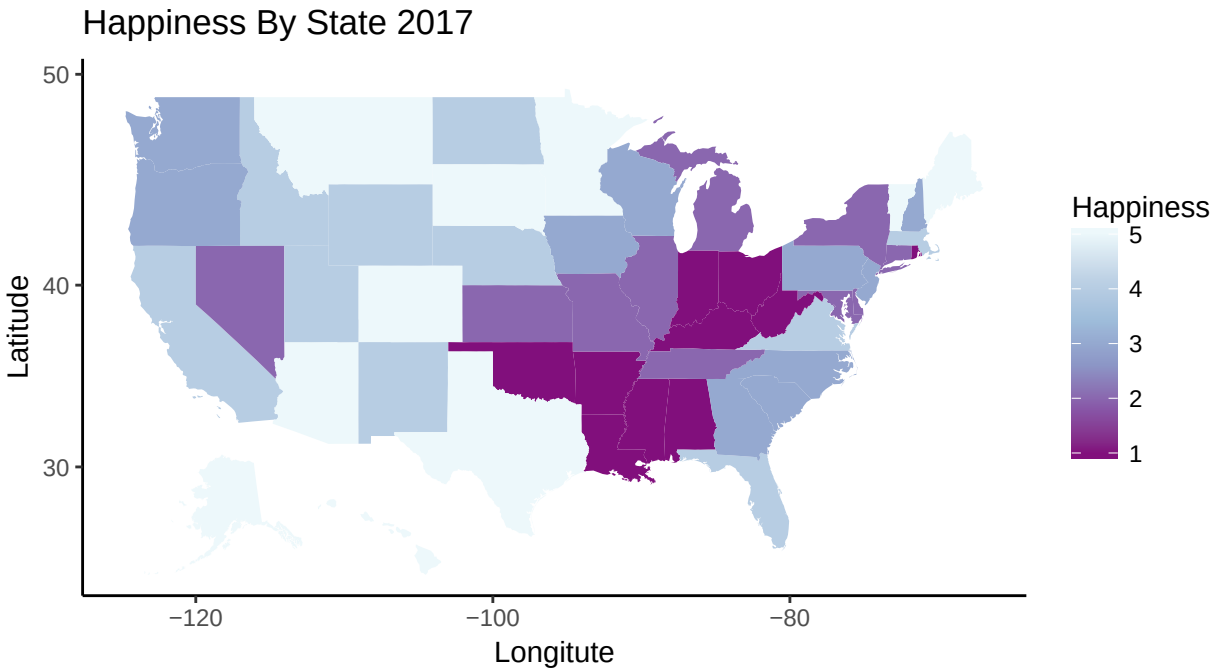

Figure 2: Caption Goes Here

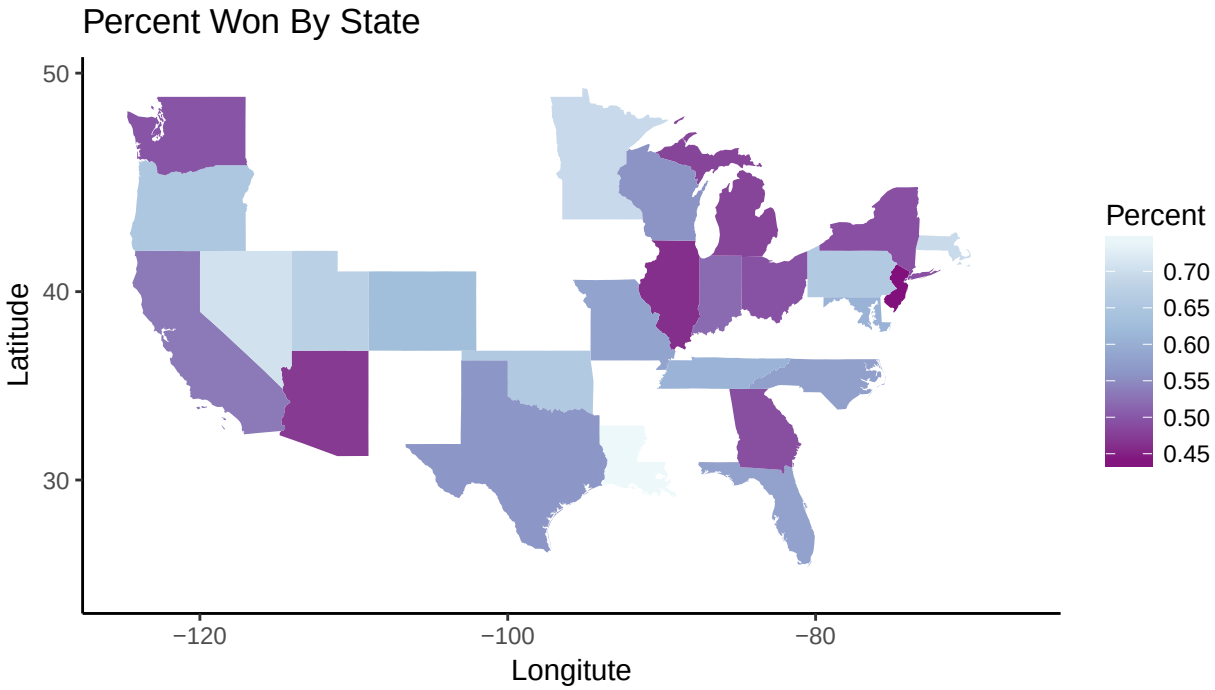

Figure 3: Caption Goes Here