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 state's 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. The next variable was from National Geographic's 2017 survey of happiness level at each state where they 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) for 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 home 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 a random intercept. Minimal information could be collected for the Canadian teams that are part of the major leagues in the United States, thus the 11 data points that were from Canada were removed.

Results

In our dataset after removing the Canadian teams, we had 114 different teams to use to investigate how stadium and state factors influenced home game win percentage. Of those 114 teams, 29 were Baseball (MLB), 29 were Basketball (NBA), 32 were Football (NFL), and 24 were Hockey (NHL). 27 states in the United States, including Washington, DC have at least one major league sports team in our dataset.

A summary of the quantitative variables can be found below:

Table 1:

Statistic	N	Mean	St. Dev.	Min	Max
Happiness	114	3.149	1.271	1	5
Trump.Vote	114	43.320	10.707	4.100	65.300
Year.opened	114	1,994.640	19.787	1,912	2,018
Population	114	15,847,640.000	11,921,447.000	703,608	39,776,830
Capacity	114	39,712.990	23,200.610	15,795	100,000
FranchiseValue	114	1,667,850,877.000	935,646,568.000	300,000,000	4,800,000,000
Attendence	114	86.708	18.128	28.200	109.800
WinPct	114	0.562	0.147	0.000	0.900

When investigating what variables may be influential in our model, we started with potential random effects. The final model contained only one random effect, the random intercept for state. To help explain the need for such an effect, we can see from the map below that the average win percentage for the 2017 season across all four major league teams is highly variable between states. It should be noted that states in white do not have major league teams and thus are not represented within the data. States like Illinois, New Jersey, and New Mexico had very low average win percentatges. On the opposite end of the spectrum, states like Nevada and Lousisiana had very high average win percentages. It is this variability that helps justify the addition of a random effect for state, given there are 50 states and the addition of a categorical variable would not be appropriate to justify such difference. From here, we decided to run a model with all fixed effects interaction terms that appeared significant in the exploratory data analysis. After running a binomial logistic regression with all of the itneractions that appeared significant in EDA, it became clear that the random effect for state was no longer large enough to justify a place in the model. It was also apparent, however, that many of the interaction terms were not significant in this model. To determine what interactions would be influential in the final model,

When determining random effects to use in the model, we opted not to include sport type. From our EDA it became clear that the sport had a large influence on win percentage on its own and whne interacting with other terms in our model. While a random effect might capture some of this effect, given we are interested

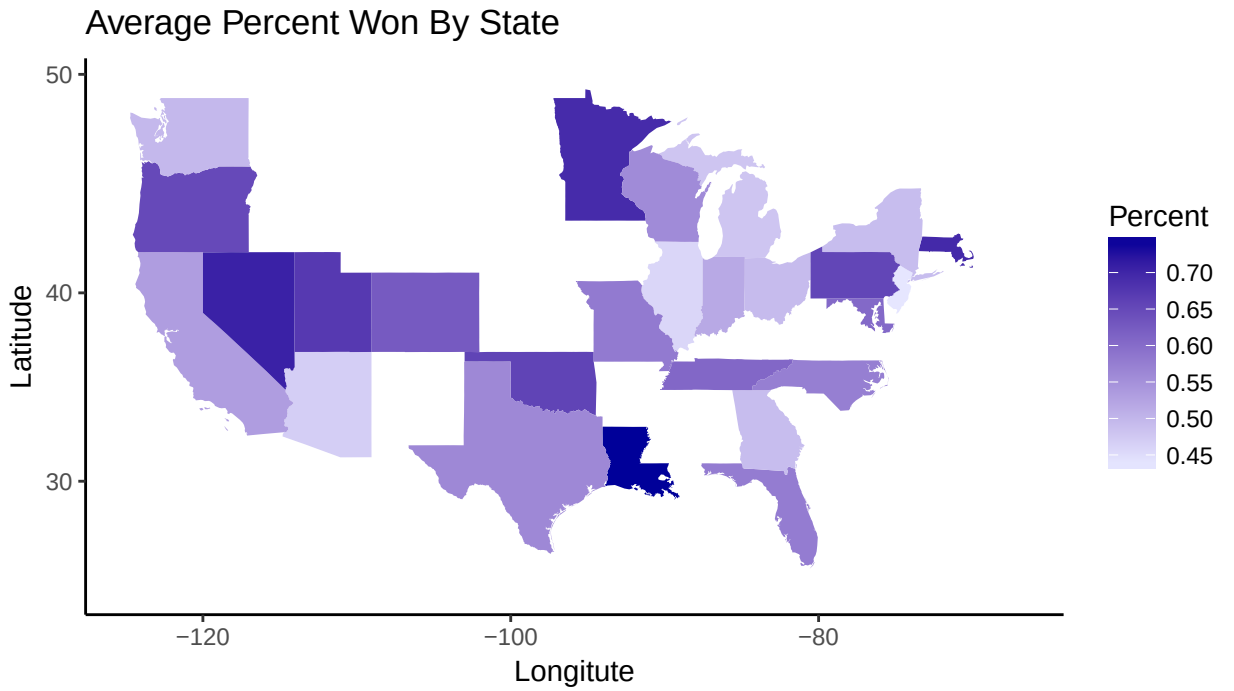

Figure 1: Average Win Percentage by State

in how influential our explanatory variables are, as it relates to win percentage, we believe we could garner more information if sport were a fixed effects variable in the model.

From initally data analysis, we discovered fourteen possible interactions. These included six interactions at the stadium/team level, all three at the state level, and five between levels. Of these fourteen interactions, only four were significant in the final model. These interactions are between sport type and attendance, sport type and capacity, state population and happiness, and capacity and attendence.

EDA of the final model interactions can be found below:

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 $\hat{\sigma} = 0.1434$

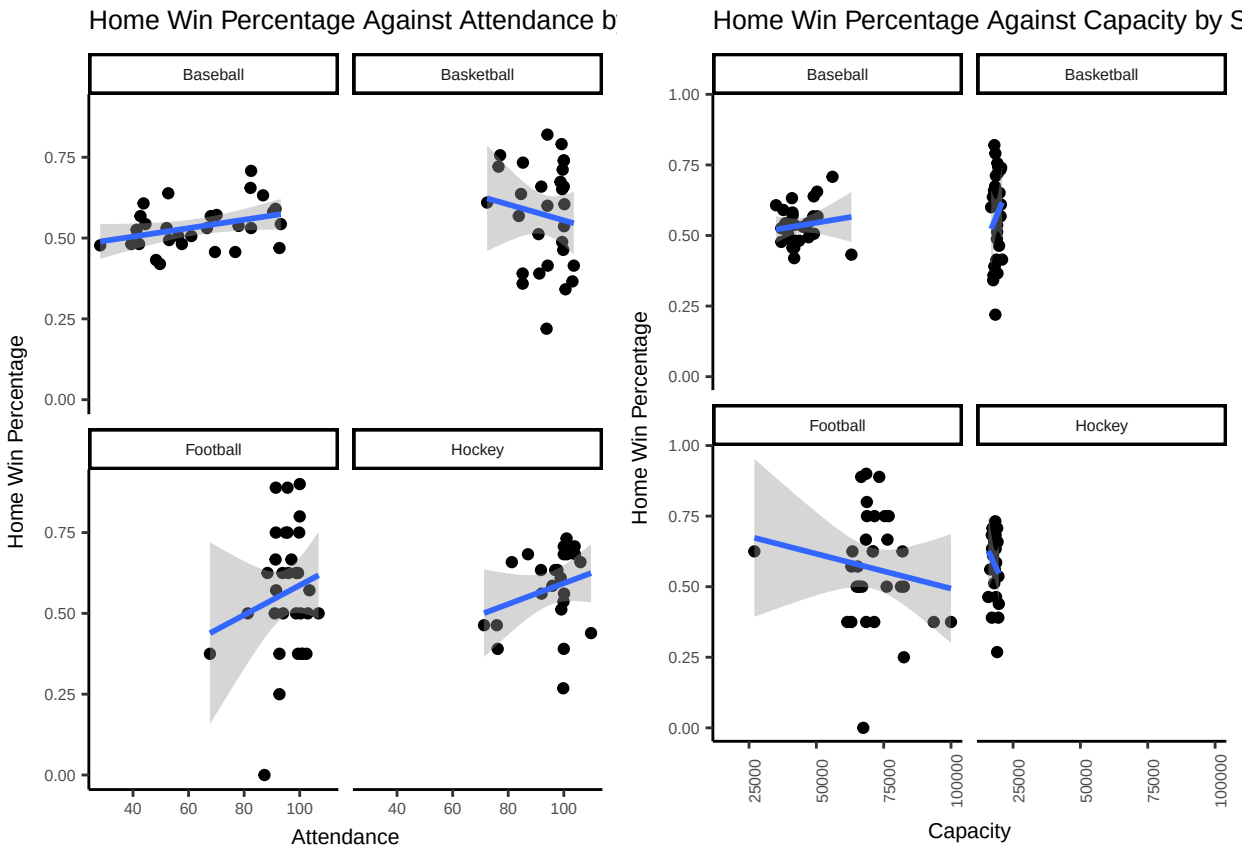

Figure 2: Attendance and Capacity by Sport

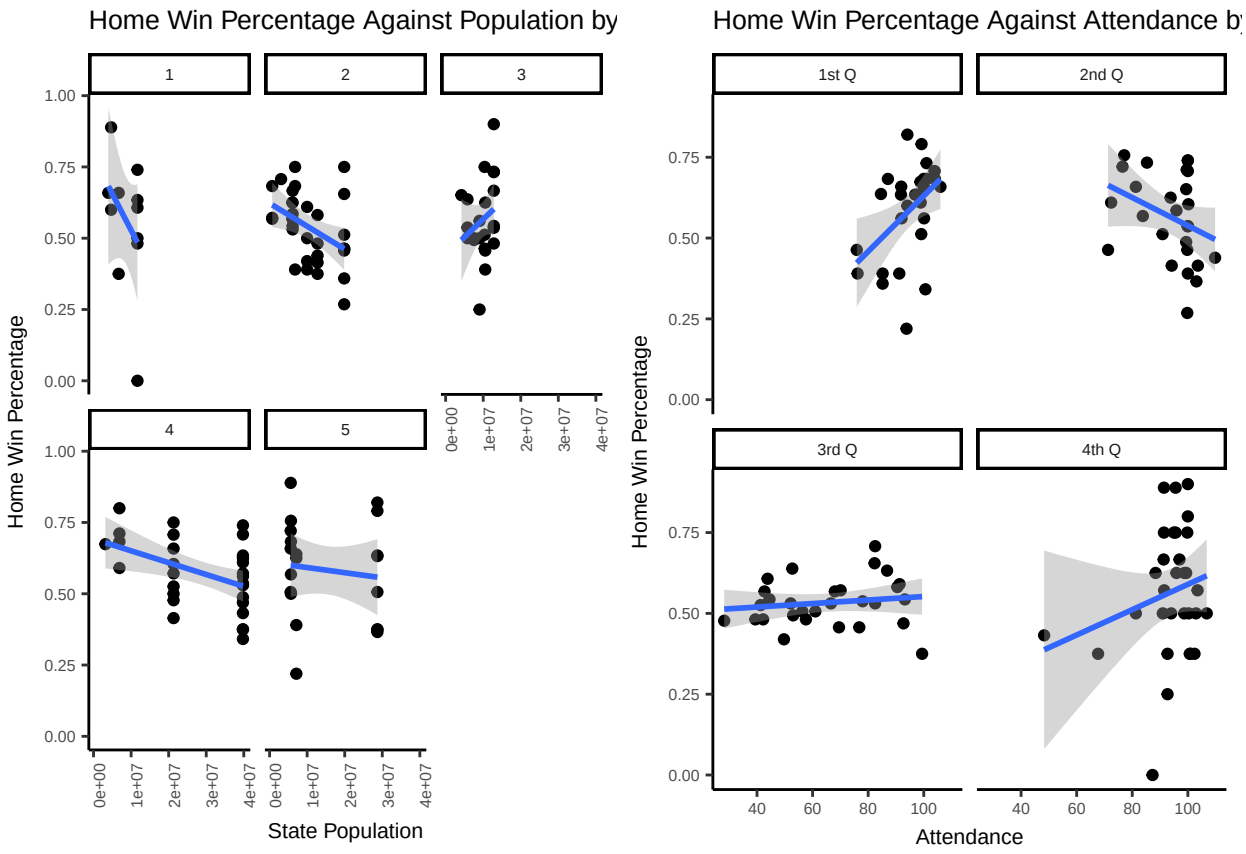

Figure 3: Population by Happiness and Attendance by Capacity Factor

Table 2:

	Dependent variable:
	WinPct
Capacity	0.467 (0.291)
Attendance	0.009 (0.070)
Basketball	0.526^{***} (0.183)
Football	-0.260 (0.365)
Hockey	-0.436^{*} (0.243)
FranchiseValue	0.135^{***} (0.048)
Population	-0.783^{***} (0.250)
Happiness	0.117^{**} (0.047)
YearOpened	-0.067^{*} (0.036)
Attendance*Basketball	-0.245 (0.193)
Attendance*Football	0.150 (0.358)
Attendance*Hockey	0.366 (0.242)
Capacity*Basketball	-0.295 (0.355)
Capacity*Football	-0.266 (0.393)
Capacity*Hockey	-0.846^{**} (0.387)
Population*Happiness	0.175^{***} (0.062)
Capacity*Attendance	0.118 (0.146)
Constant	-0.339 (0.238)
Observations	114
Log Likelihood	-326.122
Akaike Inf. Crit.	690.243
Bayesian Inf. Crit.	742.231
Note:	*p<0.1; **p<0.05; ***p<0.01

Using the model results, the most significant predictors of home winning percentag were stadium capacity, state population, state happiness level, attendence for hockey games, and type of sport. Many of these variables relate to characteristics of the stadium as well as a team's fan base. The only variable that was completely removed from the model was percent of Trump vote meaning that the percent of people in a state who voted for Trump had no effect on the home winning percentage of a team.

First we will examine how team level factors effect home win percentage. The first variable is type of sport. Compared to baseball teams (our baseline), basketball teams are 10.9 times more likely to win at home and football teams are 3 more likely to win at home. Hockey on the other hand sees a decrease in home winning odds of 96% compared to baseball.

For baseball teams, increasing their stadium capacity by 1,039 (the standard deviation of stadium capacity for baseball teams) increased the mean odds of winning at home by 37%. For basketball teams the effect is much stronger. Increasing capacity by basketball arena standard deviation (6,116 seats) increases mean odds of winning at home by 2.8 times. Football teams actually see a decrease of 70.9% in mean home winning odds when they increase their stadium capacity by a standard deviation (11,875 seats). Finally, hockey sees a decrease as well with their home game winning odds decreasing by 99% when they increase their stadium capacity one standard deviation of 998 seats. Increasing stadium size also chances the effect of attendence on odds of home wins since there is an interaction term for these variables. Increasing the baseball stadium capacity by the above amount increases the effect of attendance by 20%. Increasing basketball capacity increases the effect of attendance by 2%, football capacity increases attendence's effect by 38% and hockey 2.7%. Increasing a team's franchise value by $93,5646,568 (the standard deviation of this data) increases odds of winning at home by 14% holding all else constant. Newer stadiums actually decrease the odds of winning. For every 19.8 years newer a stadium is, the mean odds of winning decreases by 8% holding all else constant. Finally, attendance's effect varies by type of sport. Increasing attendence by 18.12 percentage points, decreases the mean odds of home wins by 6% for baseball teams. Increasing attendence by 18.12 percentage points also decreases the odds of home winning for football, but by 60% holding all else constant. For basketball teams, this change increases mean odds by 60% and for hockey, teams odds of winning are 2.3 times higher with higher attendance holding all else constant.

Moving onto state level effects, increasing a state's population by one standard deviation (11,921,447 people) decreases the odds of home wins by 62% but also has an implication on the effect of happiness since there is an interaction term. Increasing happiness level by 1, increases the mean odds of winning at home by 17% holding all else constant. We can interpret the interaction between happiness and population by if we increase populaton by 11,921,447 people, the effect of happiness increases by 24%.

Discussion

for discussion potentially older well established teams have better winning records while newer teams with newer stadiums do not have same legacy for disscussion this could be because these are smaller more intimate venues so attendence matters more

Appendix