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 "home-field advantage", but just how influential is a team's home field 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 existence 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 primarily 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 heavily divided on their 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 dependent 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 datase

One of the interesting things we found was that stadium capacity varies by sport as seen in Figure 1, which makes sense since some teams play inside like Hockey and Basketball, while Baseball and Football compete outside in large stadiums. Football has the largest stadiums

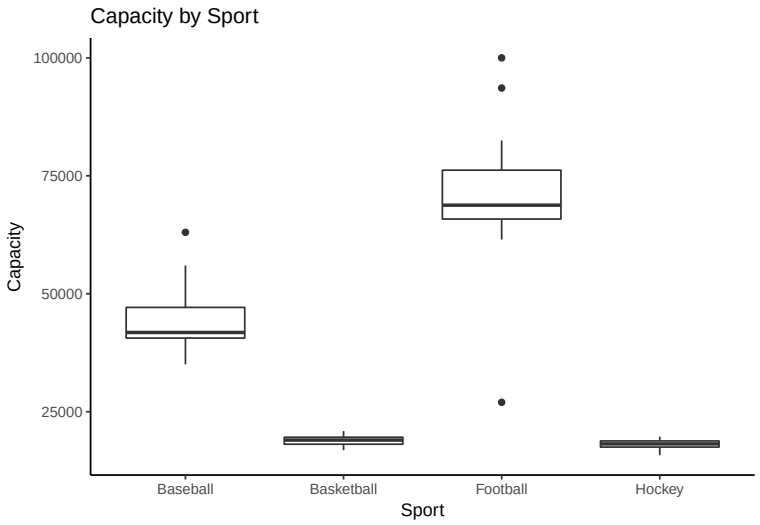

Figure 1: Stadium Capacity by Sport

while Hockey and Basketball have the smallest. In some states, Hockey and Basketball teams actually share the same arena, but since basketball games have floor seats and the ice for hockey takes up more space, the basketball games have more seats. Since capacity differs so much for different sports, we would expect different sports to see different effects on win percentage by capacity. This is analyzed further later in the report.

The variable we are most interested in is home winning percentage. When we plot this by sport in Figure 2, we do not see too much variation. For all sports, the mean is higher than 50% suggesting that home-field advantage may exist. Football and Basketball see the most variation in home winning percentages, with the Football team the Cleveland Browns winning 0% of their home games in the 2017 season. Baseball has the lowest mean as well as variation for home game winning percentage. This could be because teams play so many games at home that they are not as determined to win them, while football teams only have 4 home games in a season, so they really want to please their fans.

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

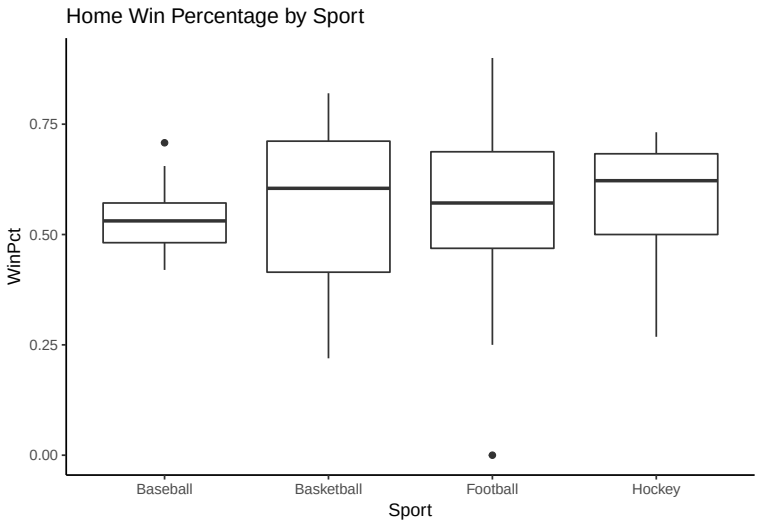

Figure 2: Home Winning Percentage by Sport

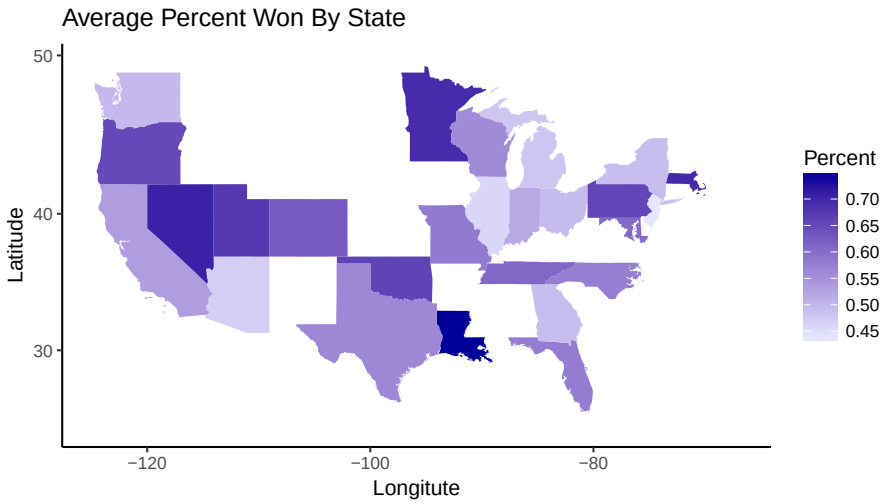

Figure 3: Average Win Percentage by State

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 in Figure 1 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 percentages. On the opposite end of the spectrum, states like Nevada and Louisiana 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.

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

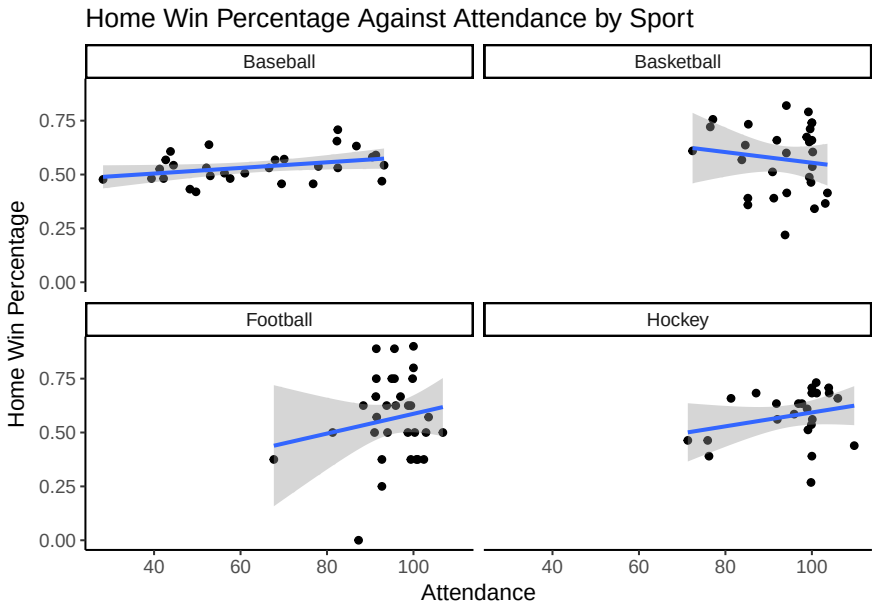

Figure 4: Attendance Against Win Percentage by Sport

its own and when interacting with other terms in our model. While a random effect might capture some of this effect, given we are interested 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 initial 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. To start, we decided to run a model with all fixed effects interaction terms that appeared significant in the exploratory data analysis. This was far too many variables, and many were not significant, so we remade the model starting with the interactions that were most significant in the large model and then added and subtracted terms from there. After the final model fitting, of the fourteen potentially interesting 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 attendance.

Figure 4 displays one significant interaction, that attendance vs home win percentage varies by type of sport. Baseball, Football and Hockey all see a positive relationship, as attendance increases, win percentage increases. Basketball on the other hand sees a slight decrease, but it is hard to tell if this is significant from the data since for most teams, attendance is high.

A similar relationship exists with stadium capacity in Figure 5. However, since stadium capacity varies by sport, it is hard to tell the true effect on win percentage from these graphs. This graph suggests that Football, Baseball and Basketball have a positive relationship with capacity and win percentage while hockey has a negative.

While both capacity and attendance change their effect on home win percentage with team, we also see an interaction between attendance and capacity. For small capacity stadiums and large stadiums, we see a positive relationship between attendance and home win percentage,

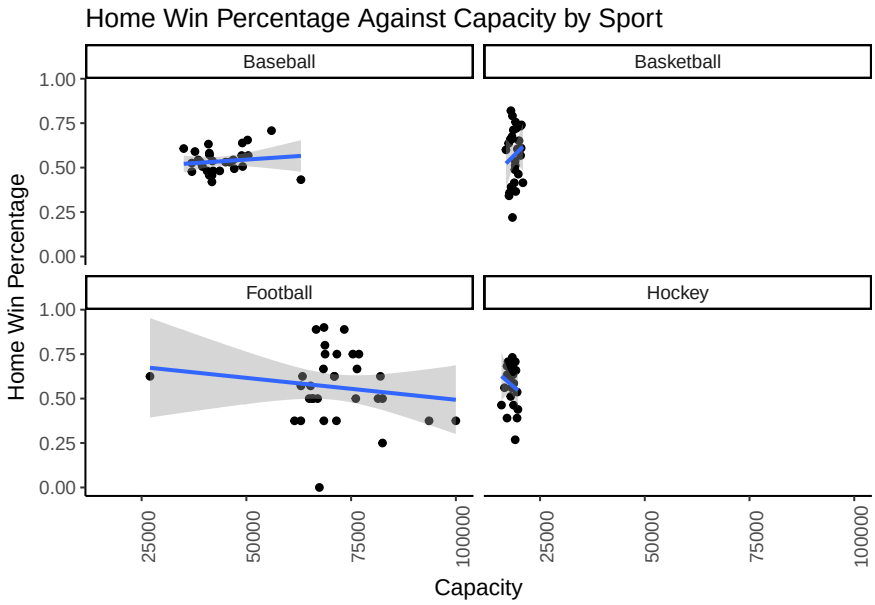

Figure 5: Capacity Against Win Percentage by Sport

while the medium size stadiums don't appear to have a relationship of between attendance and home win percentage. This can be found in Figure 6.

Finally, to examine the interaction at the state level, for the most part, as population increases for each happiness level, home win percentage decreases. However, for happiness level 3, there is a positive relationship between population and home win percentage as seen in Figure 7.

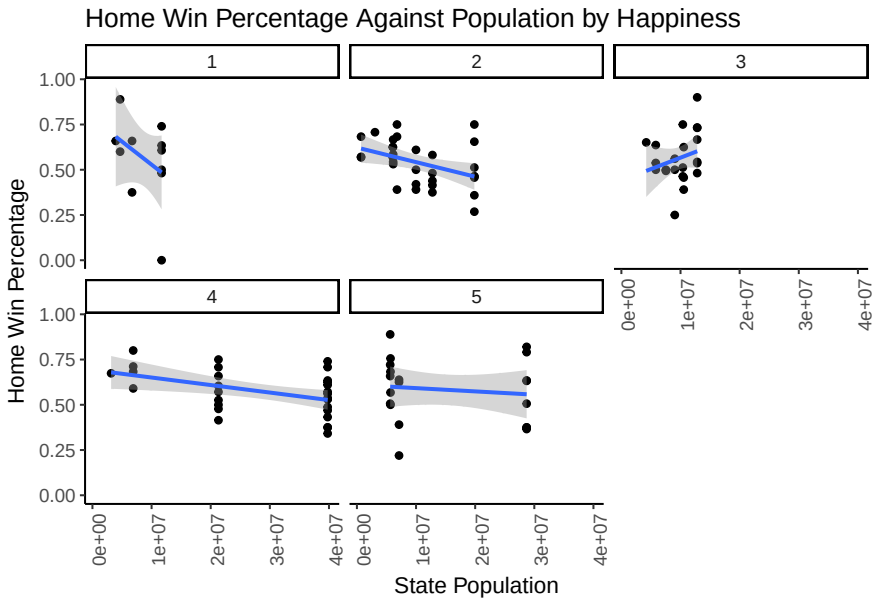

Figure 6: Population Against Win Percentage by Happiness

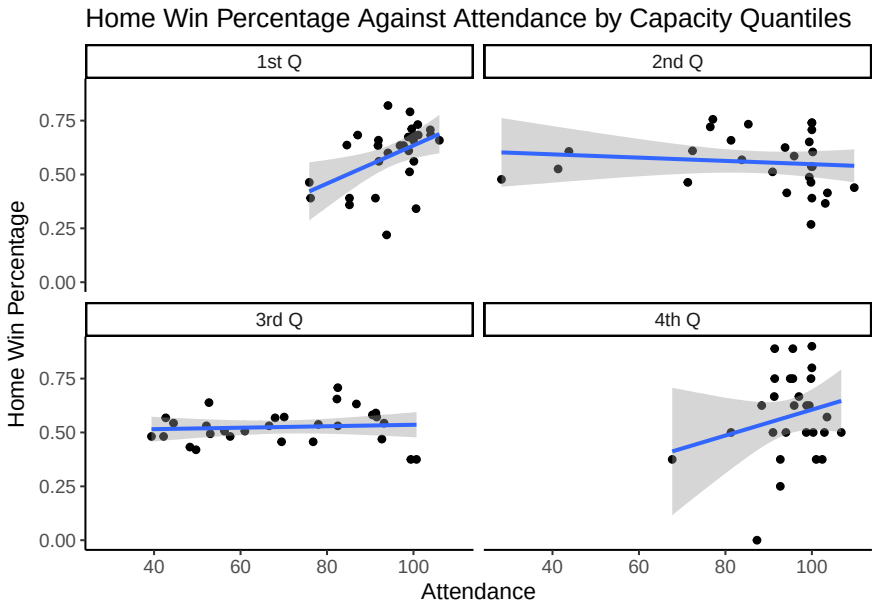

Figure 7: Attendance Against Win Percentage by Capacity Factor

Now that we've examined the interactions, we can move onto the final model. The hierarchical form of the final model is below. Estimates for the β parameters can be found in Table 2. 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 $\hat{\sigma} = 0.1434$

Using the model results, the most significant predictors of home winning percentage were stadium capacity, state population, state happiness level, attendance 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. We are 95% confident that the odds of winning a home game are 0.10 times less to 4.88 times larger for basketball teams than baseball teams. We are 95% confident that the odds of winning a home game are 0.12 to 2.06 times larger for football teams than baseball teams and 6.417 times less to .02 times more for hockey teams than baseball. Football is the only case where 0 is not include in the confidence interval, meaning that the home game win percentage is significantly different for football than baseball. This was something we discussed early in the data analysis for home game win percentage. Football has the least number of home games, so teams might be more focused on winning those.

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

Table 2:

	Dependent variable:
	WinPct
Capacity	1.214^{***} (0.355)
Attendance	-0.062 (0.059)
Basketball	2.389^{*} (1.272)
Football	1.092^{**} (0.496)
Hockey	-3.197^{*} (1.643)
FranchiseValue	0.133^{*} (0.070)
Population	-0.962^{***} (0.252)
Happiness	0.157^{***} (0.047)
YearOpened	-0.085^{**} (0.036)
Attendance*Basketball	0.326 (0.252)
Attendance*Football	-0.855^{*} (0.458)
Attendance*Hockey	0.884^{***} (0.246)
Capacity*Basketball	0.976 (1.471)
Capacity*Football	-1.859^{***} (0.488)
Capacity*Hockey	-5.121^{***} (1.783)
Population*Happiness	0.214^{***} (0.062)
Capacity*Attendance	0.630^{***} (0.195)
Constant	-0.622^{***} (0.210)
Observations	114
Log Likelihood	-321.292
Akaike Inf. Crit.	680.585
Bayesian Inf. Crit.	732.573
Note:	*p<0.1; **p<0.05; ***p<0.01

chances the effect of attendance 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 attendance'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 attendance by 18.12 percentage points, decreases the mean odds of home wins by 6% for baseball teams. Increasing attendance 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 population by 11,921,447 people, the effect of happiness increases by 24%.

Discussion

The goal of this research was to determine if factors related to a sports team's home stadium and state have any effect on the home winning percentages. What we found, not surprisingly, is that the sport itself and interactions related to the stadiums have a significant influence on home winning percentages. At the base sport level, basketball and football were more likely to win at home than baseball, while hockey was less likely to win at home compared to baseball. This could be due to the spectator environment that each sport is known for, although from the exploratory data analysis, the win percentages for all four sports appear pretty equal, and this difference is small. The interaction term that sport has with capacity of the stadium gives a more interesting interpretation. We see that as capacity increases for hockey and football, win percentage decreases. As capacity increases for basketball and baseball, win percentage increases. Although speculation, we believe this could be due to the fan bases these sports attract. Typically, football and hockey are known for being rowdier and hostile compared to baseball and basketball. Increasing the capacity for these fans to interact with their teams might facilitate environments that helps the team perform (baseball and basketball) or hinder the team's performance (football and hockey).

Similar trends occur when sport type is related to attendance. We see that as attendance increases, basketball and hockey see increased win percentages compared to baseball, football has the opposite relationship. As hockey has opposite effects related to capacity and attendance, perhaps NHL teams play better at home when the arena is full, and not necessarily perform worse because the area is large. If a hockey team has increased capacity,

we predict they perform worse, but if they have large attendance, they perform better. Thus, if the increased capacity causes the arena to be partially empty, this would hinder the home performance of the team. The other three major leagues have a similar relationship between capacity and attendance. Basketball and baseball perform better in large arenas or stadiums with high attendance while football performs best in smaller stadiums with less attendance, compared to baseball.

The final effect that was particularly interesting was at the state level. The model concluded that increases in state population alone hurts the home win record for the sports teams within said state. However, as there is an interaction between population and happiness, as happiness levels and population increase, so do winning home records. This is further amplified by happiness increasing win percentages overall. This means that if a state is particularly happy, increases in population can benefit the home teams. Alternatively, if states rank lower on the happiness scale, population increases could damper or reverse this effect. Thus, the influence of population increases on home winning percentage is also highly dependent on the state's happiness.

While we originally had percent of Trump vote as a predictor variable in our model, it was removed since it was not significant in predicting home game winning percentage. While this makes sense, it would be strange if states that liked Trump won more, it would have been an interesting finding. One of the reasons we decided to this variable in the original model is because of the debate over the national anthem protests at football games. Trump claims that support of the NFL declined after his comments criticizing the players who protested the anthem. Many people also believe that football fans tend to be more right leaning. However, our analysis suggests that voting for Trump had no effect on home game wins (sorry, Trump).

These findings are interesting because it shows that sport franchises might be able to make improvements to their stadium or fan base to increase home game winning percentage that has nothing to do with players. This could give team owners more control over their team's win record and more ownership on how well their team performs. As sports business becomes a larger and larger field, sports business consultants (like Kraft Analytics Group) could harness this information to advise teams to build certain size stadiums or figure out the right level of attendance at their games. One may notice that some of the attendance values in our data were over 100%. That is because some teams, like the Minnesota Twins offer standing room seating, while teams like the New England Patriots only offer seat tickets so they can never be at over 100% attendance. Having standing room tickets is an interesting concept for teams as they look to make games more accessible to younger, less wealthy fans. Maybe these standing room seats actually can help the atmosphere of the game and thus increase home game wins. This could be an area of further study.

There are many potential directions to go from this analysis. Since teams run renovations to stadiums all the time, a natural experiment could be created to see if our predictions about capacity effects are accurate. One could compare home game winning percentages for years before renovations with fewer seats, to years after with more seats and see if there is a difference. We also only used one year of data in this analysis: the 2017 seasons. It would be interesting to run this analysis with many years to data and include team as a random effect to see if the significant predictors we observed are significant over time. Another possible

direction is to look at how away teams perform in other teams' stadiums. Maybe the positive factors we observed for home teams, negatively affect away teams. Adding in more variables about a team's fan base could also be an interesting addition to the model.

Some potential problems with our analysis is that there are many variables that could contribute to home game winning percentage so there is no way to include them all. For this reason, we do not expect our model to be accurate at predicting home game winning percentage. However, it still can illuminate interesting correlation between variables and home winning percentages as we have found. Another problem with our analysis is since we are doing binomial regression for random effects, we could not run model diagnostics such as leverage or cooks distance to remove influential points. As the field of statistics develops and these methods are discovered, we could do a better job making sure our model is not affected by outliers.

While there are potential problems with this analysis, we did find many interesting effects on home winning percentage including the fact that these effects depend on what sport you are looking at. This shows that not all sports are the same, so there is no magic formula for running a successful team or building a successful stadium.