Home Sweet Home? A Look into Home Field Advantage Among Major U.S Sports

Nicolas Perez, Ty Pham-Swann, Zaul Tavangar, John Manning

Goal

For sports fans, it's tacit knowledge that home teams have a major advantage over their opponents. But how does the importance of home field advantage compare across America's five major sports leagues, the NFL, NBA, NHL, MLB, and MLS? We sought to investigate whether certain sports have a more significant home field advantage. We also looked to examine whether economic and environmental factors, such as elevation or income, influenced a team's home-field success.

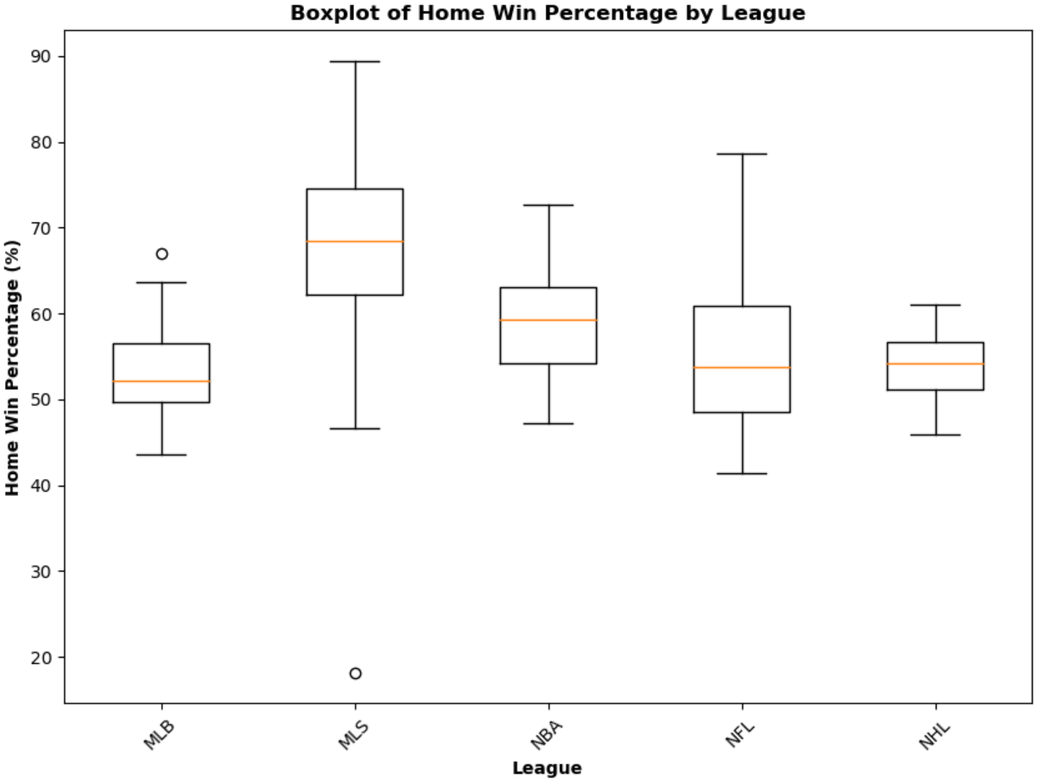
Data

We collected team, stadium, and game data for the period 2003-2022 from various sources, primarily from Kaggle as well as from official league sources. We gathered one Kaggle dataset for each of the five major U.S. sports leagues–NFL, NBA, MLB, MLS, and NHL. Each dataset consisted of detailed match data over the era described above, including information such as team location, attendance, capacity, league/sport, and most importantly, home team matchday wins. Some of the datasets represented the result of a match as a boolean, others as a string, and others as an integer, thus a major component to our cleaning process was to produce a common "home_wins" column across the five leagues, representing whether or not the home team won the match. Finally, to enhance our amalgamated dataset, we incorporated elevation data using WolframAlpha and median household income data from the U.S and Canadian censuses based on each team's location.

Findings

Claim #1: There is a significant difference in the mean home win percentages between the five major U.S leagues–we expect the MLS to have the highest mean home win percentage.

Support for Claim #1: We calculated the home win percentage for each team, and then performed an ANOVA (Analysis of Variance) test to compare the difference in means between each league/sport. Our null hypothesis was the assumption that there was no significant difference in the mean home win percentages between different sports. Our test yielded an f-statistic of 11.40 and a p-value of 4.27 x 10^{-8}. The f-statistic suggests that there is a significant difference in the home win percentages across the five leagues, and our small p-value indicates this result is significant. Thus we can successfully reject the null hypothesis. The boxplot below in Figure 1 visualizes the variance in home win percentages across the five leagues and confirms the second part of our hypothesis, that the MLS has the highest mean home win



percentage.

Claim #2: There is a significant difference in home field win percentages between teams playing in higher elevation areas and teams playing in lower elevation areas–teams that play in higher areas of elevation have a higher home win percentage.

Support for Claim #2: To investigate this hypothesis, we split our set of teams into two groups. One low elevation group (teams that play at elevations less than 1000 ft) and one high elevation group (teams that play at elevations greater than 4000 ft). Note that some teams that fall into the 1000-4000 ft are left out, but we deemed this sacrifice necessary in order to create a significant split between the low and high elevation groups. After splitting the data, we conducted an independent t-test to explore our hypothesis. Our null hypothesis was that there is no significant difference in home field win percentages between teams playing in higher elevation areas and teams playing in lower elevation areas. Our t-test yielded a t-statistic of -1.58 and a p-value of 0.12. Although the t-statistics suggests there is a significant difference, given p-value of 0.12, greater than our chosen significance level of 0.05, we fail to reject the null hypothesis.

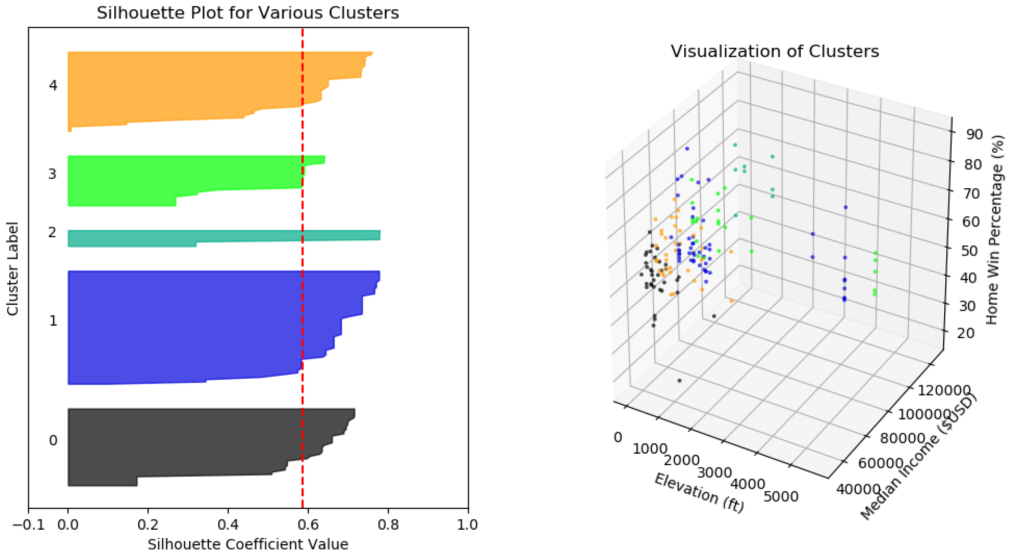
Claim #3: Attendance is positively correlated with home win percentage.

Support for Claim #3: We wanted to investigate the correlation between these two variables, and deemed calculating the Pearson Correlation Coefficient appropriate to measure the strength and the direction of the relationship between attendance and home win percentage. Note that we disregarded all the games for which we lacked attendance data, i.e. all MLS, NFL, and NHL games. Nevertheless, we judged the 14,180 games that remained from the MLS and MLB largely sufficient to investigate this correlation. Our test yielded a correlation coefficient of 0.12, and a p-value of 1.87 x 10^{-44}. Our coefficient suggests a positive, but quite weak correlation between attendance and home win percentage, and our miniscule p-value suggests that this result is statistically significant.

Claim #4: Clustering major U.S sports teams based on home win percentage and environmental factors leads to high silhouette scores, and thus clearly distinguishable categorical groups.

Support for Claim #4: To investigate this hypothesis, we tried to cluster the U.S major sports teams based on home win percentage and environmental factors (such as elevation and median regional income) using KMeans clustering. We iterated through different KMeans models using different k values and picked the optimal k value that would produce the maximal silhouette score. By clustering the major sports teams into 5 different groups, we found that the silhouette score was 0.59. Since the silhouette score was between 0 and 1, this indicated that the major sports teams could be categorized into clearly distinguishable groups given home win percentage and environmental features.

Silhouette Analysis of Major U.S Sports Teams based on Environmental Factors


Silhouette Plot for Various Clusters
Cluster Label
Silhouette Coefficient Value
Visualization of Clusters
Home Win Percentage (%)
Elevation (ft)
Median Income ($USD)