Home Sweet Home?

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

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

Background

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.

Hypothesis

We expected home field advantage to have the most influence over a team's success in the MLS.

Data

Game data over the period **2003-2022** was collected from five **Kaggle** datasets, one for the games in each league. Each dataset consisted of detailed match data, including information such as team location, attendance, and the match outcome. We collected elevation data using **WolframAlpha** and median household income data from the **U.S. and Canadian censuses** based on each team's location.

Methodology

We first cleaned and joined our five datasets together to form three tables: one for **game data**, one for **team data**, and one for **stadium data**.

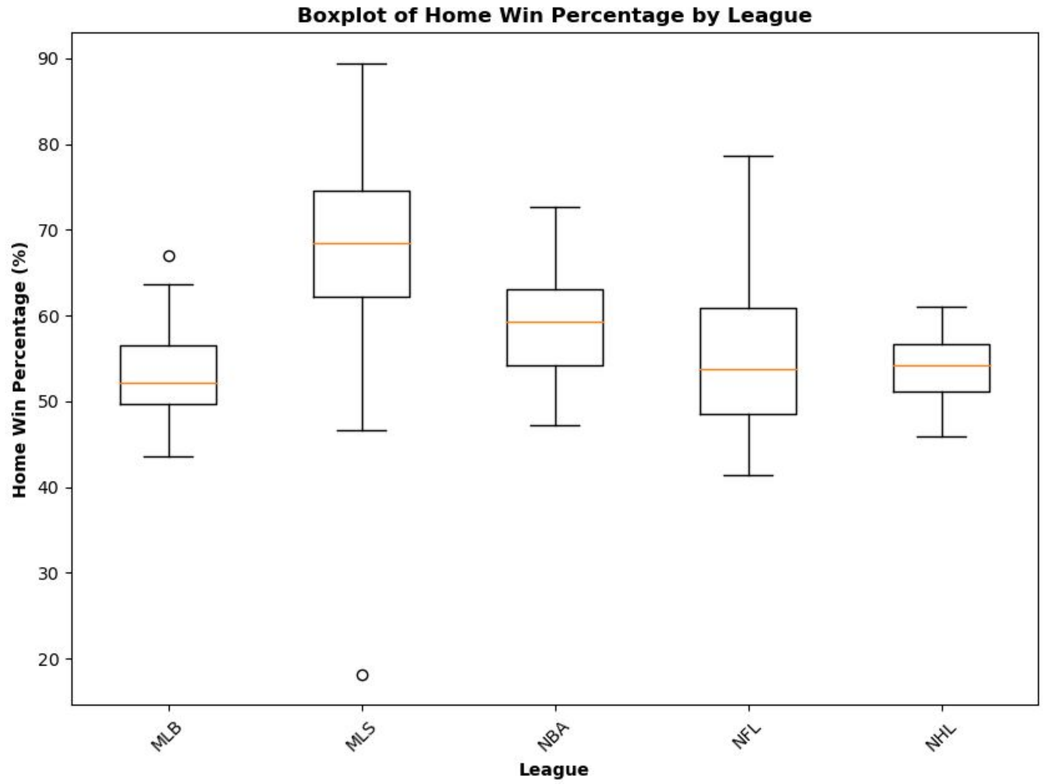
Based on the cities associated with each team, we connected cities to **elevation and income data**.

We first sought to analyze the influence of home field advantage across the five leagues, using average **home win percentage** as our **metric** for success.

Using this metric, we then used **statistical testing** to investigate the relationship between a team's success at home, and elevation and income.

Analysis and Results

- **ANOVA test result:** To investigate our main hypothesis, we used an ANOVA test to compare the difference in win percentage means between each league/sport.
  - We found that *there is a significant difference between home win percentages across the five leagues*. We then generated the boxplot below, confirming our initial hypothesis that playing at home holds the greatest weight in the MLS.



- **Independent t-test result:** We split our teams into low elevation (<= 1000 ft) and high elevation (> 4000 ft) groups.
  - We found that *there is no significant difference in home win percentages between teams playing in high elevation areas and teams playing in low elevation areas*.
- **Correlation test**: We investigated the correlation between home win percentage and game attendance by calculating the Pearson Correlation coefficient.
  - We found *a positive, but very weak correlation between attendance and home win percentage.*

Further Analysis

- **Silhouette score of K-Means clusters:** We generated distinguishable K-Means clusters by considering both the home win percentage and various environmental factors as features in the clustering algorithm.
- **Decision tree:** We made a decision tree to simulate game outcomes between teams with randomly assigned geographical, environmental, and economic characteristics.

ANOVA test	f-statistic	p-value
	11.40	4.27×10^{-8}
Independent t-test	**t-statistic**	**p-value**
	-1.58	0.12
Correlation test	**pearson correlation coefficient**	**p-value**
	0.12	1.87×10^{-44}

Challenges

- We wanted to use GDP to measure a city's wealth, but this by-year data is very difficult to find especially for smaller cities
  - We thus had to settle on household income
- Attendance data was only available for MLB and MLS games–data from the other leagues would have enhance our analysis
- A small sample of teams made it tough to develop refined ML models

Conclusions

Our project provided valuable insights into the dynamics of home field advantage across the five major US leagues. However, it is important to note that the outcome of a game and the presence and magnitude of home field advantage can be influenced by various other factors not considered in our project, such as weather conditions, travel distances, scheduling, and more. Future research could delve deeper into these factors and their specific effects on home field advantage and match outcomes in general. Additionally, our project is by design centered around US leagues and sports, whose relatively timid atmospheres are often duly noted by sports enthusiasts all around the world. An intriguing continuation of our research would be to perform similar testing using game data from a more complete set of countries, sports, and leagues.