Each summer 30 Major League Baseball (MLB) teams play 162 regular season games in 27 cities across the United States and Canada. Each team’s success is determined by a multitude of factors including roster, injuries, coaching and weather.

**Data**

Game Statistics for all MLB games from 2010-2015 was collected from a historical baseball statistics website where they record a wide range of metrics from game scores, attendance, to number of stolen bases. Additionally, the weather data for each city was collected from the blank where they measure a precipitation levels, temperature, and other meteorological occurrences.

**Goal**

In this study, we hope to examine how the environment affects player and team performance and the game of baseball as a whole. Then we will attempt to predict game performance using environmental data. Game performance is defined as the measures of offense and measures of defense, and we will examine total number of hits in a game for offense and total number of errors or total number of strikeouts minus number of total number of walks in a game for defense.

Through this study we hope to understand how weather and the environment impact baseball, which could lead to further studies examining similar trends in other activities and sports that do not get as much attention.

**Data Viz**

The heatmap are displayed to find the correlations between variables. The relationship between game statistics themselves and weather data themselves are clear. The outcome also shows that number of hits has a weak correlation with temperature, and the number of pitches has a weak correlation with weather type.

Violin plots and box plots are used for checking the distribution of weather for each game statistics. Temperature effects the number of hits, the number of hits decreases when the temperature approaching 70 degrees of Fahrenheit. The violin plot also shows that the game time (day or night) does not matter the game performances.

K-means

To further understand the data, underlying trends and relationships, we applied k-means clustering. Through this method, we see that there are distinct groupings in the data of game statistics and weather data. Thus, this suggests that there are certain underlying combinations that might lead to these groupings such as certain weather events causing unique game results. Next using the neural networks, we hope that we can use these underlying trends to be able to predict game statistics.

Neural Networks

We created multiple models to predict game statistics such as penalized strikeouts (total number of strikeouts in a game minus total number of walks in a game), total number of hits, and total number of errors.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Penalized Strikeouts | Total Number of Hits | Total Number of Errors |
| RMSE | 0.1063396 | 0.1181028 | 0.1403144 |
| MAE | 0.08397531 | 0.093066 | 0.1098034 |

**Conclusion**

* Weather and the environment appear to play a role in baseball statistics and outcomes.
* Contrary to our initial predictions, number of errors committed in baseball game is not as impacted by weather as pitching or hitting.
* Teams, players, and fans should expect for unusual game results or outcomes when the weather and environment differ from normal conditions.

Each summer 30 Major League Baseball (MLB) teams play 162 regular season games in 27 cities across the United States and Canada.  Each team’s success is determined by a multitude of factors including roster, injuries, coaching, and weather.

Game Statistics for all MLB games from 2010-2015 were collected from a historical baseball statistics website where they record a wide range of metrics from game scores, and attendance, to the number of stolen bases. Additionally, the weather data for each city was collected from National Centers for Environmental Information which measure precipitation levels, temperature, and other meteorological occurrences.

Through this study, we hope to understand how weather and the environment impact baseball, which could lead to further studies examining similar trends in other activities and sports that do not get as much attention.