Altitude Effects on Baseball Statistics and Park Factor

Sam Busser
Kevin Nguyen
Armen Arsenian
Blaise Page

Project Description

Since its creation in 1993, Coors field has notoriously been one of the most hitter friendly parks in the major leagues due to the high altitude. The Rockies have tried to combat this by storing baseballs in a humidifier.

There are several parks across the league though that are also considered hitter friendly that are not at a high elevation.We are going to look at teams and players stats throughout the stadiums played in the MLB (pitching, hitting, and fielding) and try to determine how big of an impact altitude has on players stats.

Prior Work

Park factor stat -> PF: ((homeRS + homeRA/(homeG)) / ((roadRS + roadRA) /(roadG)).

This idea of the park factor at Coors Field and many other stadiums across the league has been debated for a long time, thus there is lots of work on this topic already. However, the majority of this work deals with teams at a general level, and does not consider the effects of altitude specifically.

Data sets

We will be using the Retrosheet database. This database contains every single play that has occurred in every game since 1989. In addition to what actually happened on each play, this database also contains how many outs there are, how many runners are on base, how many pitches were thrown in the at bat, where the ball was hit, who was at each fielding position, and much more.

http://www.retrosheet.org/

We all have the dataset downloaded on our respective machines.

Proposed Work

Data Cleaning: We will have to go through our database and get rid of columns and information that we will not need for our project.

Processing: Retrosheet contains every play that occurred, but not individual or team stats, so we will have to calculate these, and compare how the altitude of the stadium affects the statistics.

Integration/Conclusions: We can use probability models like hypothesis testing and correlation tests to see if the altitude plays a big role in the park factor. If this is true, then we could predict how a player would perform on a specific team based on that team's park, and if a team has an advantage because of their park.

List of Tools

MySQLWorkbench to run queries on our database to get the necessary columns and stats.

iPython Notebook to process our data and test if the data is correlated, we will be using libraries such as Pandas, Numpy, and MatPlotLib

Evaluation

Once we gather and process our data from Retrosheet, we will try to see if there is a correlation between altitude and the statistics of each stadium to see if altitude really does play a role on how players perform. If the correlation between altitude and players stats is not very strong, then we can look at other factors as well.