## Tuple Attributes:

* ID – manually generated when populated
  + Numeric
* Date – the date the game is played on
  + Follows Date format
  + i.e. 2024-05-201
* Home Team – The team name that hosts the contest
  + Textual
* Home Score – the amount of runs scored by the Home Team
  + Numeric
* Away Team – The team name that isn’t hosting the contest
  + Textual
* Away Score – the amount of runs scored by the Away Team
  + Numeric

## Missing Values:

Out of the 36,270 tuples in “mlbGames.csv”:

* only 185 tuples that have missing values.
* Each of these missing values happens to be just the away team.
  + Missing percentage = 1/5 = 20% (not including ID)

Out of the 30,441 tuples in “mlbGames2.csv”:

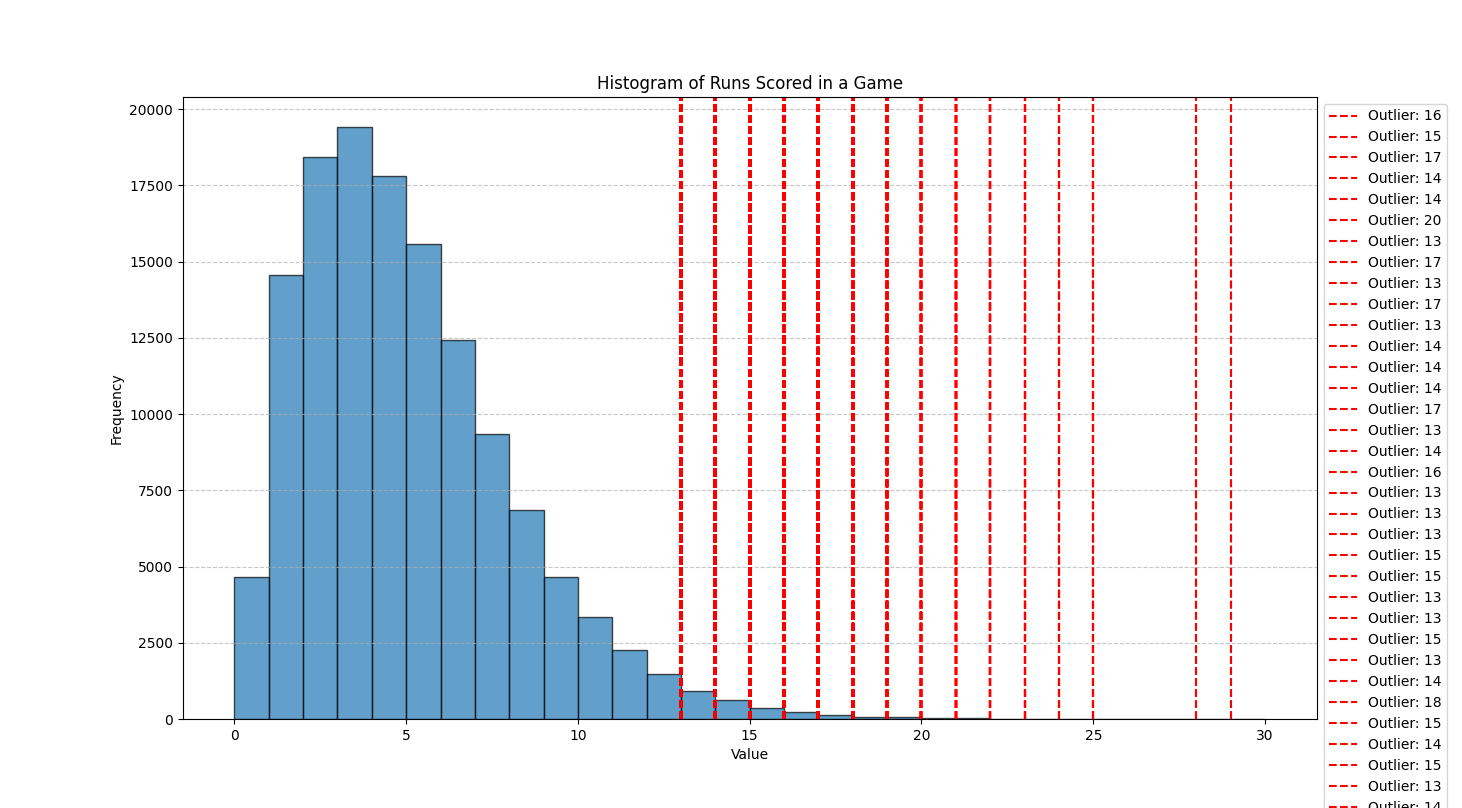
* There are no tuples with missing values.

## Solutions to Missing Values:

Missing Team Names: I’ve noticed a lot of the 185 tuples with missing values is due to how I wrote the scraper. For one of the website, the team names were abbreviated (i.e. Colorado Rockies = COL). So, I created a hashmap that transformed the abbreviation to the team’s full name. The issue lies when the scraper finds COL, which is abbreviated for Colombia during the Baseball World Series Classic. So, the scraper populates COL as Colorado Rockies, then doesn’t know how to populate MEX for Mexico. The fix that makes to most sense it to cross validate the tuples with missing values to the table that has no missing values. The cost of this is fairly small since we would only be cross-validating the 185 tuples, 0.05% of the data set.

## Analysis:

First I took a look at all the runs scored in an MLB game by either team on every game. I found that there were 2,525 outliers in this dataset. The mean was 4.56 runs and the threshold for outliers was 7.67 runs.



Next, I looked at just the Home Team scores. I did this because it is believed that is you host a game on your own field, you have the “advantage” due to familiar field. Less travel, and more fans in the crowd. I found that there were 1,229 outliers in this dataset. The mean was 4.61 runs and the threshold for outliers was 7.57 runs.

A graph of a graph

AI-generated content may be incorrect.

Then, I compared the Home Team scores to the away team scored to see if there is a noticeable difference in the distributions. The average Away Team score was 4.52, 0.09 less than the Home Team score average.

A graph of a number of blue and red lines

AI-generated content may be incorrect.

Notes:

* There are no synonyms in this dataset. (not anymore due to the hashmap mentioned in "Solutions to Missing Values:”)
* Other than the 185 missing attributes, there are no other issues with this data.
* Packages used to create histograms and analyze the data: numpy, matplotlib