**What is the difference in offensive power, throughout the years, for teams in the World Cup?**

Jorge Sedano Fernandez

DSC 530: Statistics for Data Science

June 3, 2023

Statistical/Hypothetical Question:

The question we look to answer in this project is: What is the difference in offensive power, throughout the years, for teams in the World Cup?

Outcome of EDA:

During the exploratory data analysis (EDA) of the FIFA World Cup dataset, several variables were analyzed, including Year, Home Team Goals, Away Team Goals, Half-time Home Goals, and Half-time Away Goals. Various visualizations and statistical measures were used to gain insights into the dataset. I identified some outliers based on extreme values that deviated significantly from most of the data points. For example, in the Home Team Goals variable, there were instances where teams scored a high number of goals in a single match, which were considered outliers. A probability mass function (PMF) was used to compare two scenarios within the data. The PMF could be used to compare the distribution of home team goals before the 90's and after. A cumulative distribution function (CDF) was created for one of the variables, providing insights into the distribution of the data and the probability of observing certain values. The CDF helped address the question of how the variable was distributed and how it related to the overall dataset. Analytical distributions were explored to identify potential fits for the data. The normal distribution was plotted as an example, and the analysis revealed that the data did not perfectly align with a normal distribution. This indicated that the variable's distribution deviated from a typical bell curve shape.

Missed Analysis, Additional Variables, and Incorrect Assumptions:

While the analysis covered several aspects, there are a few areas that could have been further explored:

1. Team-specific variables: The analysis focused on match-level variables, but considering team-specific characteristics, such as historical performance, FIFA rankings, or team composition, could provide additional insights into offensive capabilities.

2. Match context: Factors such as the stage of the tournament, the importance of the match, or the strength of the opponent could influence a team's offensive capabilities. Including these variables in the analysis could offer a more nuanced understanding.

3. Non-linear relationships: While linear relationships were assessed using correlation and regression analysis, exploring non-linear relationships between variables could unveil hidden patterns or associations.

4. Assumptions about data quality: The analysis assumed that the dataset was complete, accurate, and representative of all World Cup matches. However, the presence of missing values, potential data entry errors, or biases in the dataset could impact the results.

Challenges and Areas of Incomplete Understanding:

During the analysis, several challenges were encountered. Handling missing values and outliers required careful consideration to ensure the validity of the results. Deciding on appropriate data preprocessing techniques, such as imputation or exclusion, is crucial to maintaining the integrity of the analysis. Interpreting the results of the regression analysis and understanding the limitations of the model required a solid understanding of statistical concepts. Furthermore, exploring non-linear relationships and advanced techniques like multivariate analysis may require additional expertise. In conclusion, the EDA of the FIFA World Cup dataset provided valuable insights into offensive capabilities. However, there were opportunities to include additional variables, explore non-linear relationships, and address potential assumptions. The analysis highlights the importance of carefully considering data quality, interpreting statistical results, and continually expanding analytical techniques to gain a comprehensive understanding of the dataset.