**Exploring FIFA World Cup Data: Uncovering Insights**

**Initial Idea and Dataset Source:**

This investigative journey commenced with the aim of answering fundamental questions related to FIFA World Cup matches. The dataset chosen for this project was sourced from Kaggle, specifically the dataset titled "International football results from 1872 to 2023." The focal questions that spurred this analysis were: Do teams have a home advantage? What trends have been observed throughout FIFA World Cup history in terms of goal differences and team parity? Finally, which countries play each other most frequently in the World Cup?

**Exploring the Data:**

The filter tool was employed to extract the FIFA World Cup matches from the entire dataset and to transfer them into a new worksheet.

To determine the goal averages for both home and away teams, the *AVERAGE* function in Excel was employed. Additionally, the *MAX* function was utilized to identify the team that scored the highest number of goals in a FIFA World Cup match.

In the pursuit of understanding the concept of "home advantage," a pivot table was constructed. This analytical tool facilitated the calculation of the average number of goals scored by teams hosting the World Cup. Simultaneously, another Pivot Table was created to generate a summary of the goal-scoring averages for teams competing at neutral venues.

The findings that emerged from this analysis indicated that, on average, teams playing on their home turf scored higher, with an average of 1.83 goals. In contrast, teams participating in FIFA World Cup matches at neutral venues recorded lower average score, with figure standing at 1.53.

**Data Cleaning:**

To ensure the dataset's integrity, a critical data cleaning phase was executed. Among the key issues addressed was the correction of spelling errors in city names, guaranteeing data accuracy. The Spelling tool in the Review section of Excel was used. Additionally, the numerical format of scores was validated using the *ISNUMBER* function, ensuring data consistency and reliability. Below the screenshot of the errors can be found.



**Data Transformation:**

To determine the outcome of each match and whether the home team emerged as the winner, an *IF* function was applied. This function was structured as follows:

*=IF (Home score > Away score, TRUE, FALSE)*

This concise formula evaluated the home score against the away score, producing a logical *TRUE* if the home team won, and *FALSE* if not.

Furthermore, to quantify the goal difference between the competing teams in each match, the *ABS* function came into play. The formula employed was:

*=ABS (Home score - Away score)*

This formula effectively calculated the absolute difference between the home team's score and the away team's score, providing a numerical measure of the goal difference.

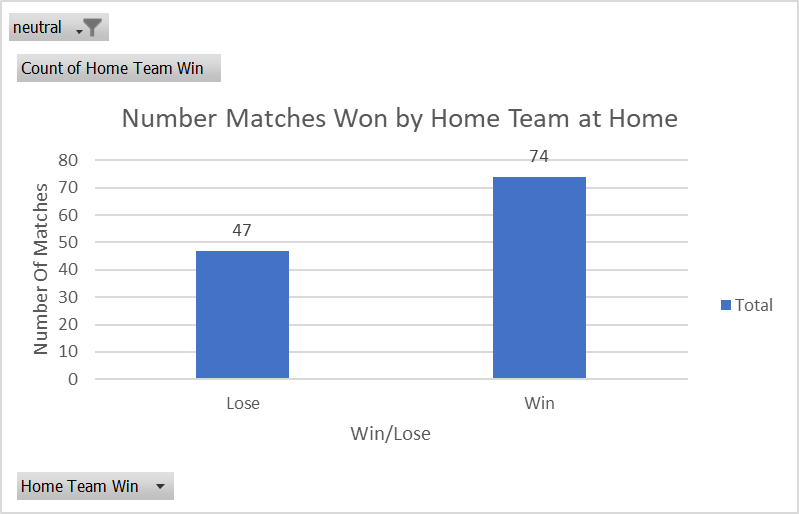
**Data Analysis:**

***Do Teams Have a Home Advantage?***

The pivotal question of whether teams enjoy a home advantage in FIFA World Cup matches was investigated. This analysis focused on the number of wins a team secured when playing on their home turf.

To answer this question, the *COUNTIFS* function in Excel was employed. This function counted the number of matches where the home team not only emerged victorious but also played at home. The result was an understanding of how often home teams won.

Out of the 121 matches played by the home team on their own turf, 74 matches culminated in victory for the home team. Conversely, 47 matches ended in defeat for the home team.



**Hypothesis Testing:**

To quantify the significance of these findings, hypothesis testing was conducted. Two hypotheses were formulated:

Null Hypothesis (H0): There is no home advantage; the probability of a home team winning is equal to the probability of an away team winning.

Alternative Hypothesis (H1): There is a home advantage; the probability of a home team winning is greater than the probability of an away team winning.

**Statistical Analysis:**

The observed proportion of team wins, calculated as 74/121, equated to approximately 0.6116. This represented the actual proportion of home team victories in the dataset.

In contrast, the expected proportion under the null hypothesis was 0.5, signifying an equal probability of victory for both home and away teams.

To determine the statistical significance of the observed proportion, a Z statistic was calculated using the formula:

*Z = ((P\_observed - P\_expected) / SQRT ((P\_expected \* (1 - P\_expected)) / N))*

In this equation, N represents the number of matches played (121), P\_observed signifies the observed proportion (0.6116), and P\_expected denotes the expected proportion (0.5).

The calculated Z statistic amounted to approximately 2.4552.

The critical Z value at a significance level of 0.05, denoted as Z value, was obtained using Excel's *NORM.S.INV* function and amounted to approximately 1.6449.

**Conclusion:**

The comparison between the Z statistic and the critical Z value led to a decisive conclusion. Since the Z statistic (2.4552) exceeded the critical Z value (1.6449), the null hypothesis (H0) was firmly rejected. This indicates that there appears to be a home advantage in FIFA World Cup matches.

***What trends have been observed throughout FIFA World Cup history in terms of goal differences and team parity?***

The approach involved transforming the data to focus on the year of each World Cup, enabling a broader view of trends over time. Here's how this analysis unfolded:

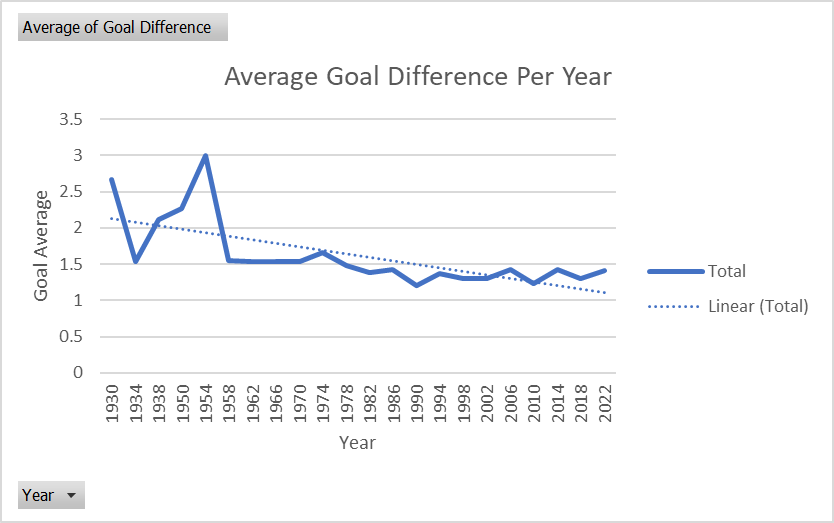
**Data Transformation:**

To facilitate this examination, the date column in the dataset was transformed to display only the year of each World Cup. This transformation was achieved using Excel's *Year* function, which extracted the year from the date.

**Pivot Table and Visualization:**

Subsequently, a pivot table was constructed to calculate the average goal difference for each FIFA World Cup edition. This pivot table organized the data for an understanding of goal differences over time.

A line chart was thoughtfully constructed to visualize the evolving pattern of goal differences throughout the history of the FIFA World Cup. To provide additional insight, a trendline was incorporated into the chart, shedding light on overarching trends.

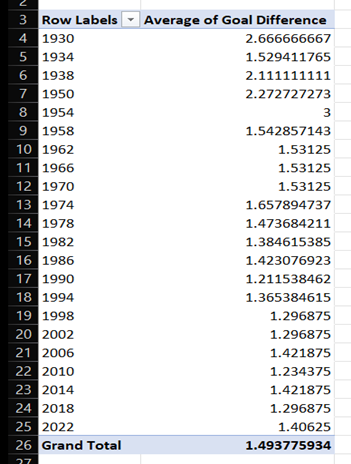
****

It became apparent that the average goal difference has exhibited a decreasing trend since the inception of the FIFA World Cup. The data indicated that the average goal difference fluctuates around an approximate value of 1.3 goals.

**Historical Context:**

This intriguing trend could potentially be attributed to the changing landscape of the tournament. An exploration of the historical context unveiled significant milestones:

* Inaugural FIFA World Cup (1930, Uruguay): Featured 13 teams.
* FIFA World Cup 1934 (Italy): Encompassed 16 teams.
* FIFA World Cup 1938 (France): Maintained the 16-team format.
* FIFA World Cup 1950 (Brazil): Saw a return to 13 participating teams.
* FIFA World Cup 1954 (Switzerland): Reverted to 16 teams.
* FIFA World Cup 1958 (Sweden): Initiated a 16-team standard, continued in subsequent editions.
* From the 1958 FIFA World Cup onward, the 16-team format remained until the 1998 World Cup, when the tournament expanded to include 32 teams.
* Since 1998, the 32-team format has become the standard.

The observed clusters in the pivot table somewhat correspond to the differences in the number of participating teams throughout the years. This suggests that as the FIFA World Cup evolved and expanded, teams may have become more evenly matched, leading to a decreasing trend in average goal differences.



***Which countries play each other most frequently in the World Cup?***

The approach taken involved the creation of a Pivot table to discern frequent matchups. Here's how this analysis was executed:

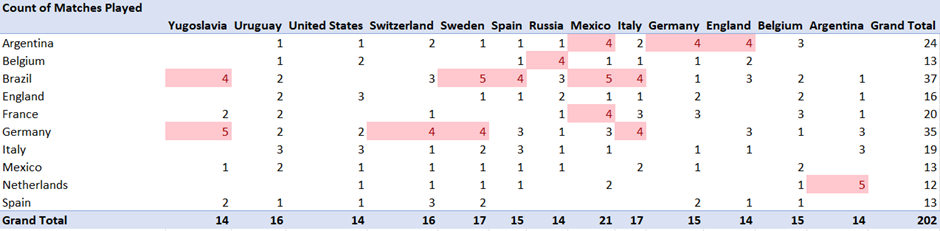
**Pivot Table Setup:**

A Pivot table was configured with "Home Team" in the Row section and "Away Team" in the Column section. This setup allowed for a systematic examination of matches played between specific teams.

The "Home Team" was ingeniously placed in the Values section of the Pivot table. This placement enabled the counting of matches played by each unique pair of teams.

**Conditional Formatting:**

To pinpoint the pairs that engaged in the highest number of matches against each other, conditional formatting was utilized. This formatting technique was applied to highlight the top ten pairs with the most frequent matchups.



As a result of this analysis, the pairs with the most extensive history of facing each other in FIFA World Cup matches were shown. Notably, the Pivot table illuminated the following frequent matchups:

* Germany vs. Yugoslavia
* Brazil vs. Sweden
* Brazil vs. México
* Netherlands vs. Argentina

These pairs have engaged in the greatest number of matches against each other throughout the storied history of the FIFA World Cup.

**Review**

**1. Home Advantage Confirmed:**

Statistical analysis demonstrated the existence of a home advantage in FIFA World Cup matches.

Home teams exhibited a significantly higher likelihood of winning when playing on their own turf.

**2. Goal Difference Trends:**

Analysis of goal differences spanning FIFA World Cup history unveiled a notable trend.

The average goal difference has shown a consistent decrease since the inception of the tournament, stabilizing around an approximate value of 1.3 goals.

This trend may be linked to changes in the number of participating teams over time.

**3. Frequent Matchups Revealed:**

A comprehensive examination using Pivot tables and conditional formatting uncovered the pairs of teams that have faced each other most often in FIFA World Cup history.

Notably, Germany vs. Yugoslavia, Brazil vs. Sweden, Brazil vs. Mexico, and Netherlands vs. Argentina emerged as the pairs with the most frequent matchups.