Introduction:

This portion of the assignment focuses on streak data which can help analyze further reasons a team may win more frequently. For example, if a team has a large win streak it could be due to having a much better roster during that time and may not be representative of a future period if it occurred long ago. The main information we are looking for in this section is an average best streak per team, the all-time best win streak by any team, and the median to avoid the best team skewing results. If this streak happened recently, it may be an indication of this team’s likelihood to win the World Cup, although it does not necessarily provide confirmation as streaks can be broken at any time.

About the Data:

The data has been found on Kaggle. Much of the data is not used in this portion of the project and the data which is used has been standardized.

Methods:

For pre-processing, the data is dropped except for dates, home and away teams, and the outcome of the match. Because a draw is not a win of course, these are interpreted by the program as losses on both sides. All home and away teams are merged to a single column named “Teams” and the column for the home team outcome splits to account for both teams and is then merged to one “Result” column. This allows all teams and their data to exist regardless of who they face at any time. All that matters is who they are, when they played, and whether they won or not.

Once the data is ready, it is then analyzed via code which increments per consecutive occurrence of wins or losses and resets the counter when the opposite outcome is achieved. We then go back and negate all losses to a streak of 0 to avoid this impacting the results. Data is sorted to place each team’s streaks in order beginning with the greatest value, and all rows are filtered by distinct team names. This removes all occurrences of a team name after the first, which by our sort, is the highest streak number, effectively giving us only a list of each team and their highest value.

Evaluation:

The data has been evaluated using the max value as well as mean, median, and mode. The data is visualized via a scatter plot and a bar graph showing the density of results and color coded to represent a heatmap. These both show the distribution of best win streaks while the variety in display helps bring clarity to the data.

Storytelling and Conclusion:

Through our data, we can see Spain has the longest win streak by far of 25, greatly going beyond all other teams. It is also shown that the mean is 5.48 and both median and mode are 5. The heatmap helps visualize this proportion and the sharp dropoff as values continue to increase.

Impact:

In our goal to find who is the most likely winner, it may be helpful to see their current or required streak in order to secure a win. For instance, the likelihood of a team overcoming their best win streak since at least 1993 when our data starts during a world cup is understandably slim. If a team for instance has not had more than 3 consecutive wins, expecting them to reach, say, 6 for example, would be an unlikely feat. If any teams have achieved a very high win streak recently and are beginning to show signs of one, it may encourage their chances as well.

Github Repository Link: <https://github.com/maymakhani/3162_Finalproject>

Related Files: WinStreakData.ipynb