EDA Process Documentation

This documentation describes an Exploratory Data Analysis (EDA) process performed on a dataset using Python libraries such as Pandas, Matplotlib, and Seaborn. The dataset used in this analysis is stored in a CSV file called 'cleaned_results.csv'. Below is a step-by-step explanation of the EDA process.

Importing Libraries: The necessary libraries are imported, including pandas, matplotlib.pyplot, and seaborn. Additionally, a warning filter is set to ignore warnings.

Loading the Dataset: The dataset is loaded into a DataFrame named df using the read_csv() function from Pandas.

Examining the Dataset: The contents of the DataFrame df are displayed, showing the first few rows and columns.

Data Type Conversion: The 'date' column in the DataFrame has a data type of 'object'. It is converted to a datetime format using the pd.to_datetime() function.

Histogram for Home Score Values: A histogram is created using the 'home_score' column from the DataFrame. This visualizes the distribution of home scores using the hist() function from Matplotlib.

Statistical Analysis: The mean and median values of the 'home_score' column are calculated using the mean() and median() functions from Pandas, respectively.

Bar Plot for Different Tournaments: A bar plot is generated to display the count of matches played in various tournaments. The 'tournament' column is used for this visualization, and the countplot() function from Seaborn is utilized.

Identifying the Team with the Most Wins: The team with the most wins is determined by comparing the 'home_score' and 'away_score' columns. The home teams with higher scores are extracted using boolean indexing, and their counts are calculated. The team with the maximum count is identified as the team with the most wins.

Top Five Winning Teams: The top five winning teams are determined by calculating the counts of each home team's wins. The value_counts() function is used, and the top five teams are displayed.

Team with the Highest Goal Scored: The goal scores for each home team are calculated by summing the 'home_score' and 'away_score' columns. The total scores are sorted in descending order to identify the team with the highest goal scored.

Best Players of the Era: The dataset is divided into separate DataFrames for each era based on the year. A function named best_team() is defined to display the top five winning teams and the top five teams with the highest scores in each era.

Advantage of Playing at Home: The advantage of playing at home is determined by comparing the number of home wins, away wins, and draws. Percentages are calculated based on the total number of matches.

New Countries over the Years: The number of unique countries participating in matches is calculated for each year. The results are plotted to visualize the change in the number of countries over time.

Matches between Countries Correlation Matrix: A correlation matrix is created to show the correlation between matches played between different countries. The matrix is generated by calculating the correlation coefficient between pairs of countries using the corr() function.

Countries that Have Played Together the Most: The countries that have played together the most are determined by counting the number of matches between each pair of home and away teams.

Countries that Hosted Matches while Not Participating: The countries that hosted matches while not participating as the home team are identified. The number of matches hosted by each country is calculated, and the results are plotted using a bar chart.