UEFA Euros History and 2024 Prediction
Final Report
Tae Woo (Kevin) Kang

1. Introduction

The UEFA European Football Championship, colloquially known as the Euros, stands as one of the premier international football tournaments, captivating millions of fans worldwide every four years. With a storied history dating back to 1960, the Euros showcase the best of European football talent and provide a platform for nations to compete for continental glory. This project delves into the rich tapestry of the UEFA Euros, exploring historical trends, analyzing team performances, and predicting outcomes for the upcoming Euro 2024 tournament.

2. Motivation and Rationale for Project

The UEFA Euros hold immense significance in the world of football, attracting widespread attention and fervent support from fans across the globe. The tournament serves as a stage for nations to showcase their footballing prowess, fostering national pride and unity. Understanding the historical context of the Euros and unraveling the intricacies of tournament dynamics can offer valuable insights into the evolution of European football and its broader impact on the sporting landscape.

This project seeks to address several key questions and explore various aspects of the UEFA Euros:

- **Historical Analysis:** Investigating the inaugural matches, host cities, and notable team performances throughout the tournament's history provides a comprehensive understanding of its evolution over time.
- **Contemporary Relevance:** Examining the correlation between World Cup success and Euros performance, assessing the tournament's competitiveness relative to other continental championships, and evaluating the influence of hosting nations on tournament outcomes shed light on current trends and dynamics.
- **Predictive Modeling:** Leveraging historical data and advanced analytical techniques to predict match outcomes and forecast the winner of Euro 2024 adds a predictive element to the project, offering stakeholders valuable insights and foresight into the tournament's proceedings.

3. Description of Data Sources

The project draws upon three primary datasets to fulfill its objectives:

1. **FIFA Ranking Data (https://www.transfermarkt.us/statistik/weltrangliste/statistik/stat/plus/0/galerie/0?datum=2024-04-04):** This dataset provides insights into the historical rankings of national teams participating in the UEFA Euros. It serves as a foundational resource for assessing team performance and dynamics over time.
2. **International Football Results Data (https://github.com/martj42/international_results):** This dataset encapsulates the outcomes of international football matches, including those from previous UEFA Euros tournaments. It forms the basis for historical analyses and predictive modeling, allowing for insights into past tournament dynamics and team performances.
3. **2024 Euro Group Team Details(https://en.wikipedia.org/wiki/UEFA_Euro_2024#Group_A) :** This dataset offers information on European nations qualified into this year's competition.

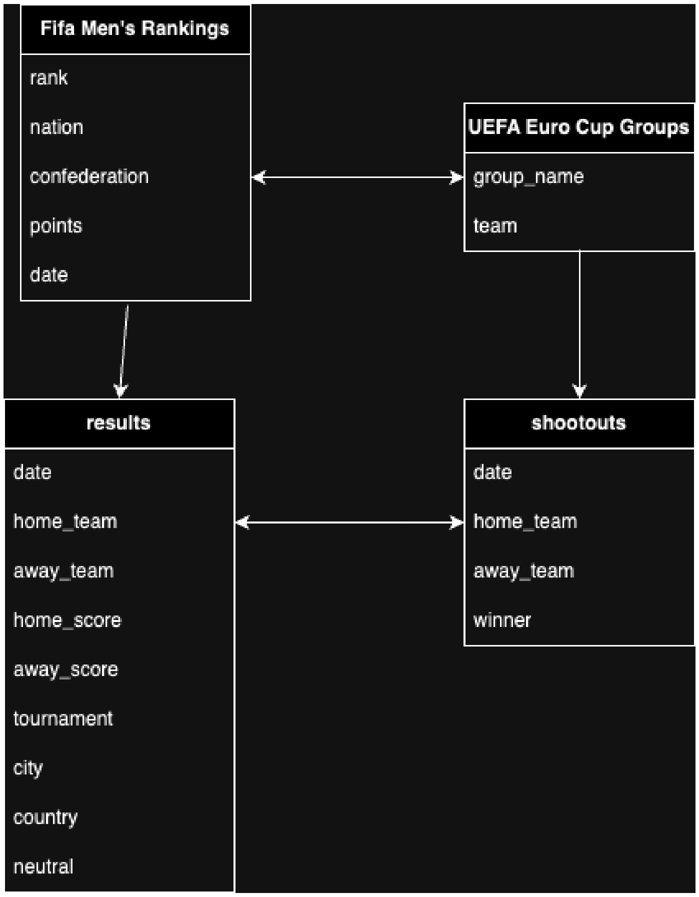
The majority of the data was obtained through web scraping using BeautifulSoup. A significant volume of data was collected, necessitating an exploratory analysis to harness its potential effectively.

Initially, the plan involved scraping FIFA ranking data from the official FIFA website (inside.fifa.com/fifa-world-ranking). However, this approach encountered challenges, including a 401 Client Error. As an alternative, we turned to a third-party website, Transfermarkt, which provided basic information on past rankings dating back to 2009. However, this dataset lacked several key data points, such as rank changes and previous points. Consequently, much of the dataset, particularly for parts 2 and 3, focused on dates after 2009, reflecting the limitations of the available rankings dataset. The analysis would have been enriched had the rankings data extended further back in time.

4. Integrated Data Model

The Integrated Data Model aims to integrate and preprocess football match results data with FIFA ranking data. It first loads cleaned match results, shootouts, and ranking datasets using Pandas. The script merges the match results with shootouts data, identifying winning and losing teams. Then, it filters the merged dataset to include only matches from September 2009 onwards. Next, it identifies the closest ranking date for each match date and merges the datasets based on these dates for both home and away teams. After reordering columns and dropping unnecessary ones, it calculates a threshold ranking to define high-ranking teams, based on the top 10% in this example. Finally, it exports the integrated and processed dataset to a CSV file for further analysis.

Note: Further data integration was utilized in the analyze jupyter notebook due to additional needs (see analyze_visualize.ipynb for Part 3)



5. Data Analysis/Visualizations

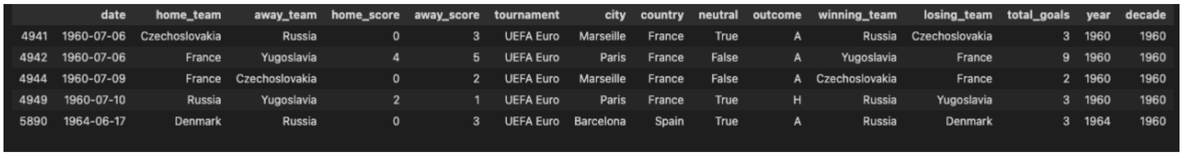
My project is divided into 3 parts: historical analysis, contemporary relevance, and predictive modeling.

Part1: Historical Analysis

We ask three simple questions in our first phase:

1) Which two teams played the first ever game in the UEFA Euros competition?

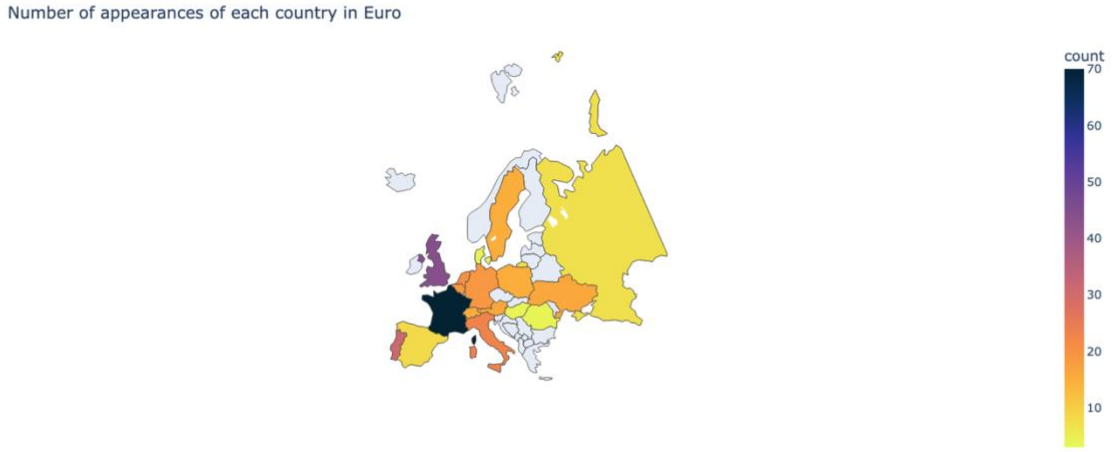
   For this analysis, I used data filtering techniques to extract UEFA European Championship (Euro) match results from the larger dataset of football results. As shown below the first ever game (excluding qualification) was between Czechoslovakia and Russia.

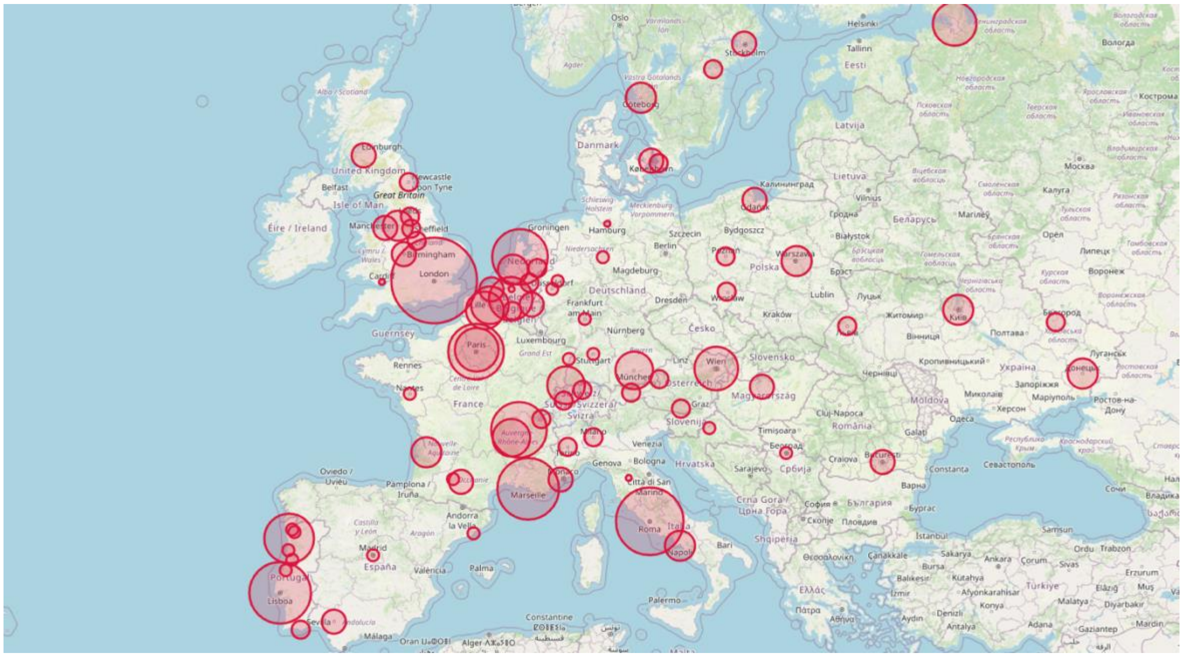


	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral	outcome	winning_team	losing_team	total_goals	year	decade
4941	1960-07-06	Czechoslovakia	Russia	0	3	UEFA Euro	Marseille	France	True	A	Russia	Czechoslovakia	3	1960	1960
4942	1960-07-06	France	Yugoslavia	4	5	UEFA Euro	Paris	France	False	A	Yugoslavia	France	9	1960	1960
4944	1960-07-09	France	Czechoslovakia	0	2	UEFA Euro	Marseille	France	False	A	Czechoslovakia	France	2	1960	1960
4949	1960-07-10	Russia	Yugoslavia	2	1	UEFA Euro	Paris	France	True	H	Russia	Yugoslavia	3	1960	1960
5890	1964-06-17	Denmark	Russia	0	3	UEFA Euro	Barcelona	Spain	True	A	Russia	Denmark	3	1964	1960

2) Which European city/country has hosted the most matches throughout the years?

   For the analysis, I used basic techniques such as data aggregation, manipulation, and visualization. Initially, I calculated the number of appearances of each country in the UEFA European Championship (Euro) tournament by aggregating the data based on the country column. Then, I utilized the Plotly Express library to create a choropleth map representing the number of appearances of each country in Euro. This map provides a visual representation of the participation of different countries in the tournament, with color intensity indicating the frequency of appearances. Additionally, I examined the cities that hosted Euro matches by counting the appearances of each city in the dataset. I used the Folium library to generate an interactive map showing the locations of these cities. The map includes circle markers sized based on the number of matches hosted by each city, providing insights into the distribution of Euro matches across different geographical regions. The figures help understand the geographical spread of Euro matches and the involvement of various countries and cities in hosting the tournament over time.

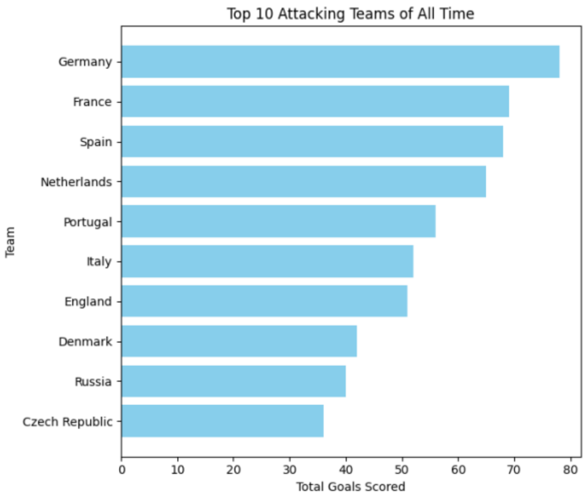


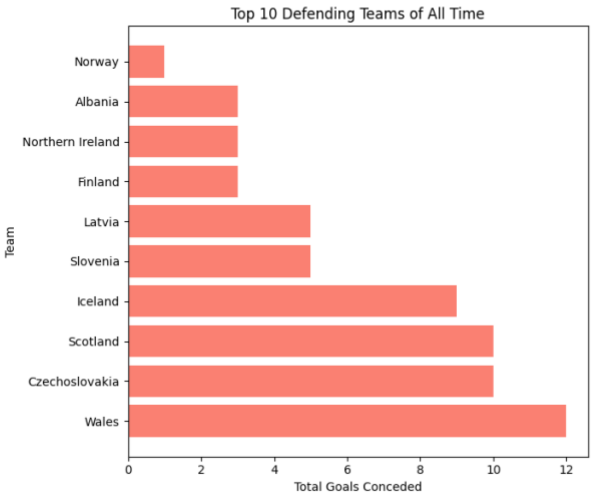


In terms of nations, France stands out as the host of the highest number of matches, totaling 70. However, according to the world map visualization, it is London that emerges as the city with the highest number of hosted matches.

3) Who are the best attacking and defending teams throughout different eras?

The analysis involves calculating the total goals scored and conceded by each team, followed by determining the goal difference. The teams are then sorted based on their attacking prowess and defensive strength. The top 10 attacking and defending teams are identified and visualized using horizontal bar charts. Additionally, the analysis examines the top attacking and defending teams for each year individually, providing insights into the evolving performance over time. Racing bar charts (see Jupyter notebook) are employed to visualize the goals scored and conceded by nations over the years dynamically, showcasing the relative performance of teams in each period. These figures are created using Python libraries such as Matplotlib, Plotly Express, and Bar Chart Race. They offer a comprehensive view of team performance in terms of goal-scoring and defensive capabilities, both historically and across different time periods, facilitating a deeper understanding of the data and trends.

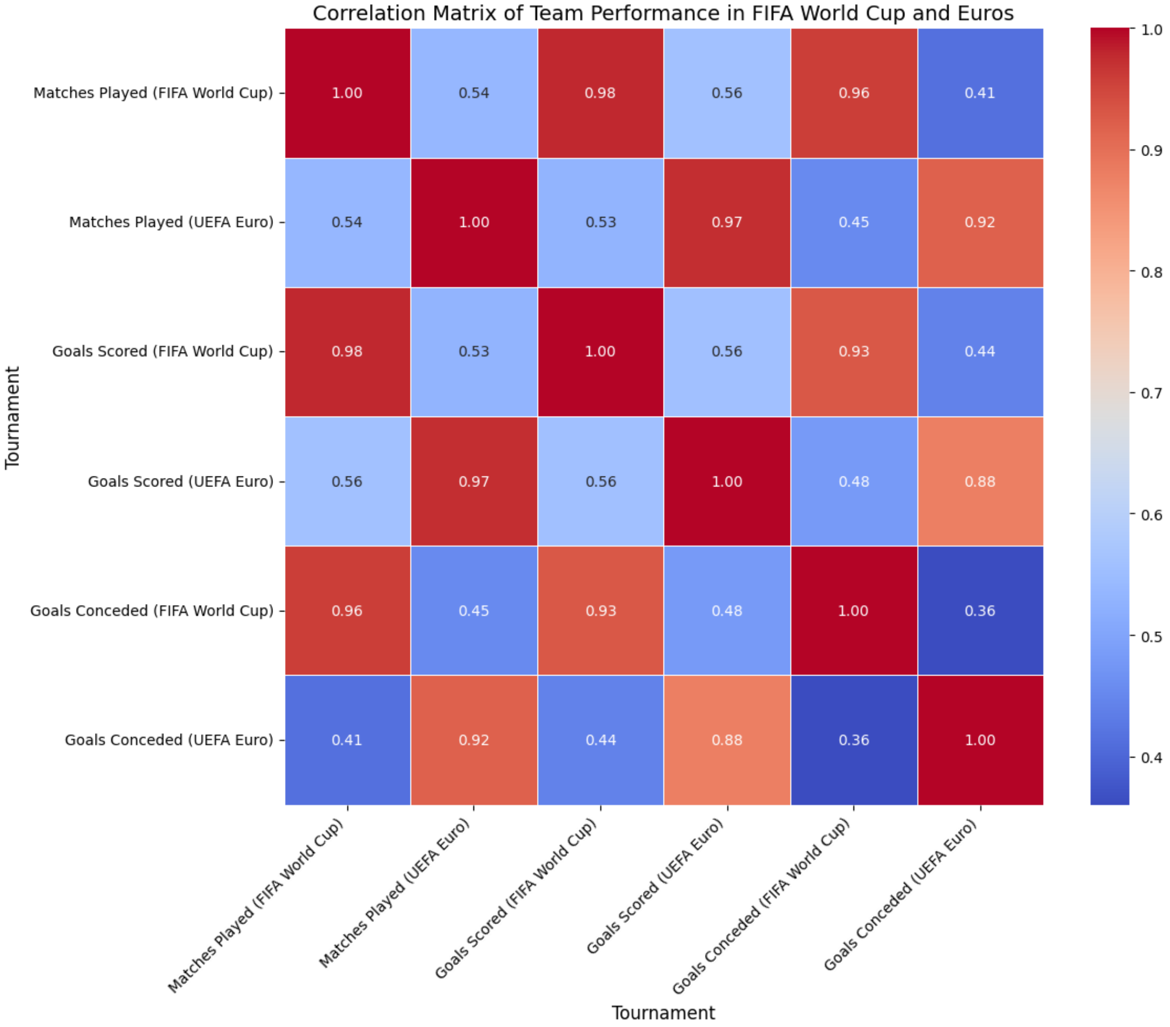




As illustrated by our bar chart, Germany emerges as the most prolific attacking team of all time, leading in terms of total goals scored, while Norway claims the title of the top defensive team, having conceded the fewest goals overall. However, it's essential to consider that these rankings may be somewhat biased by the frequency of appearances of each team.

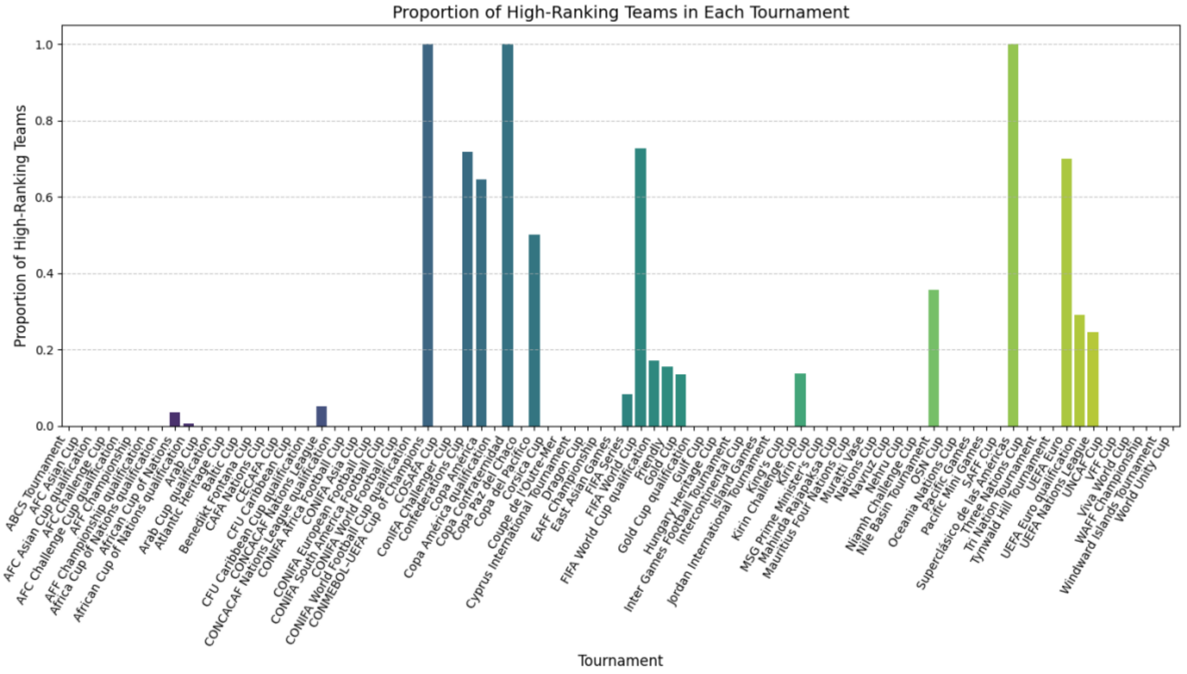
Part 2: Contemporary Relevance

1) Do the teams that are successful in the World Cup also perform well in the Euros and/or vice versa?



For this analysis, I employed various techniques including data filtering, feature engineering, grouping, aggregation, and correlation analysis. Firstly, I filtered the dataset to include matches from both the UEFA European Championship (UEFA Euro) and the FIFA World Cup, excluding qualification matches. Next, I concatenated the filtered datasets to create a comprehensive dataset encompassing matches from both tournaments. Subsequently, I conducted feature engineering to derive a new variable indicating match outcomes (win, lose, or draw). Then, I grouped the data by tournament and home team, aggregating metrics such as total matches played, total goals scored at home, and total goals conceded away. After filtering for relevant columns, I pivoted the table to separate home and away performance for each tournament. Following this, I calculated correlation coefficients to explore the relationship between team performance metrics across tournaments. Lastly, I visualized the correlation matrix using a heatmap, where each cell represents the correlation coefficient between different performance metrics. The heatmap includes custom labels for tournaments and performance metrics, facilitating easy interpretation of the relationships between team performances in the FIFA World Cup and UEFA Euro.

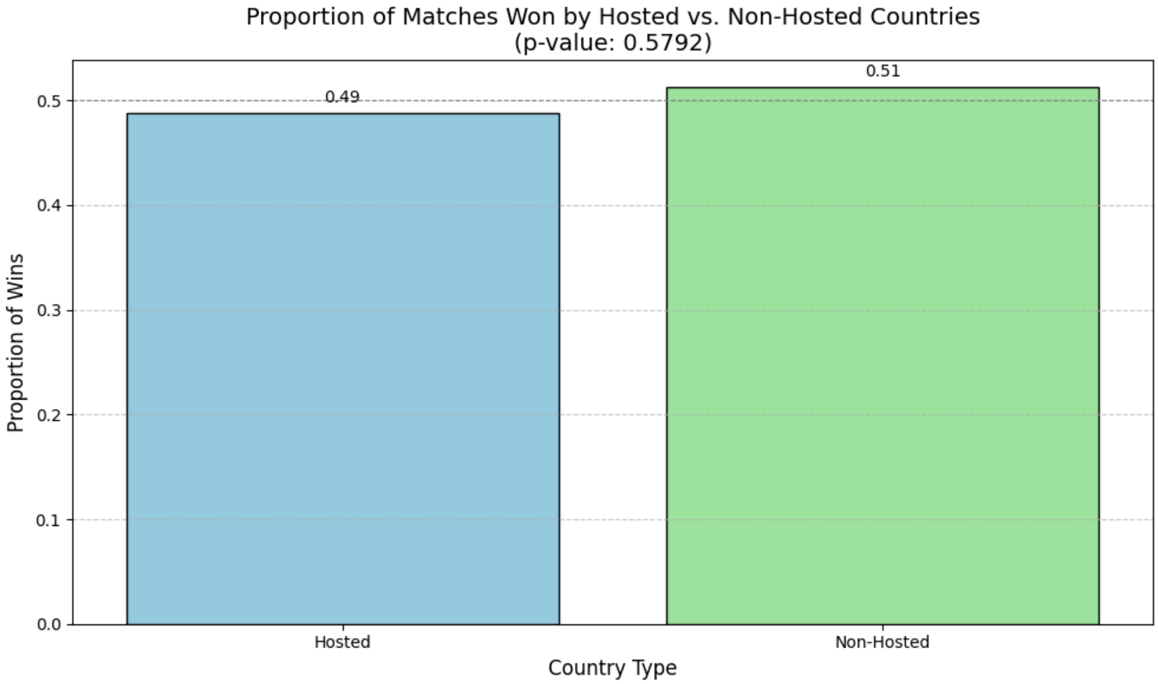
2) Is the Euros competition the hardest continental competition to win regarding the number of teams in high rankings?



For this analysis, I utilized statistical testing and visualization techniques. Firstly, I performed a chi-square test of independence to assess the relationship between tournament type (UEFA Euro or FIFA World Cup) and the presence of high-ranking teams. The contingency table was constructed using the crosstab function from pandas, which tabulates the frequency of occurrences for each combination of tournament type and the presence of high-ranking teams. The chi-square test was then applied to determine whether the observed frequencies significantly deviate from the expected frequencies under the null hypothesis of independence.

To visualize the proportion of high-ranking teams in each tournament, I grouped the merged dataset by tournament and the high_ranking column and calculated the mean proportion of high-ranking teams for each tournament. This mean proportion was then plotted using a bar plot created with Seaborn. The x-axis represents the tournament type, while the y-axis indicates the proportion of high-ranking teams. Each bar in the plot corresponds to a tournament, with the height of the bar representing the mean proportion of high-ranking teams. Additionally, I included gridlines to aid interpretation and clarity in the visualization.

3) Does the hosted country have an advantage in terms of advancing further into the tournament?



For this analysis, I employed statistical testing and visualization techniques to compare the performance of hosted matches with non-hosted matches in the UEFA European Football Championship.

Firstly, I filtered the dataset to isolate matches where the home team is the hosted country, creating a subset named "hosted_matches." I then calculated the proportion of matches won by the hosted country and compared it with the proportion of matches won by non-hosted countries. Statistical testing was performed using a chi-square test to assess whether there is a significant difference in performance between the two groups.

To visualize the comparison, I created a bar plot using Seaborn, with "Hosted" and "Non-Hosted" as the groups on the x-axis and the proportion of wins on the y-axis. Each bar represents the proportion of matches won by the respective group, with different colors distinguishing between hosted and non-hosted countries. Additionally, I included data labels above each bar to display the exact proportion of wins. A horizontal dashed line at y=0.5 was added to represent the 50% win rate threshold for reference. The plot title includes the p-value obtained from the statistical test for clarity and interpretation.

Part 3: Predictive Modeling

When analyzing the dataset, I employed various techniques to derive insights and prepare the data for modeling. Here's an overview of the analysis techniques used:

Firstly, I utilized pandas functions to manipulate the DataFrame, including adding new columns based on existing data and filtering relevant information.

Next, I computed statistics for each team based on their past performance in matches. This involved calculating metrics such as mean goals scored, mean goals conceded, mean opponent rank, mean game points, and mean game points per rank for both all past games and the last five games.

I then created a base DataFrame containing essential columns for modeling and computed additional features based on differences between home and away team statistics.

Following that, I split the dataset into features (X) and the target variable (y) and further divided it into training and testing sets using the train_test_split function from scikit-learn.

For modeling, I implemented two classifiers: Gradient Boosting Classifier and Random Forest Classifier. To optimize their performance, I conducted hyperparameter tuning using GridSearchCV to find the best combination of hyperparameters.

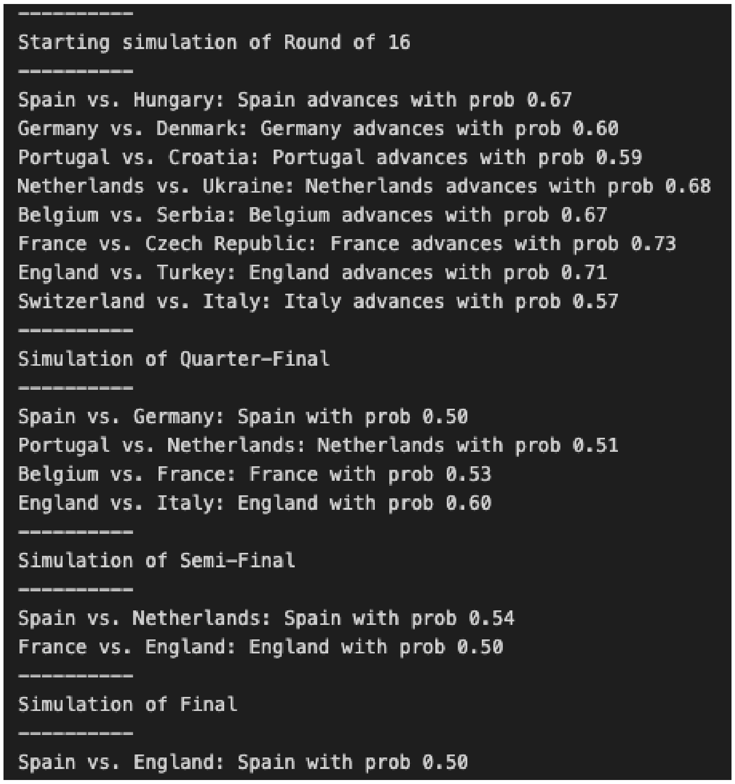
After training the models, I evaluated their performance using the testing set and selected the best-performing model for further analysis.

Now, let's delve into the figures created and their significance:

The figures generated include matchup simulations for different stages of the tournament, such as the Round of 16, Quarter-Final, Semi-Final, and Final. These simulations provide insights into the probabilities of each team advancing to the next round based on their past performance and calculated features.

For each matchup simulation, the probabilities of the home team and the away team winning are presented, allowing for an understanding of the predicted outcomes and potential advancements in the tournament.

Overall, the analysis techniques utilized enable a comprehensive understanding of team performance, inform decision-making in predicting match outcomes, and offer insights into potential tournament results based on historical data and statistical features.



```
----------
Starting simulation of Round of 16
----------
Spain vs. Hungary: Spain advances with prob 0.67
Germany vs. Denmark: Germany advances with prob 0.60
Portugal vs. Croatia: Portugal advances with prob 0.59
Netherlands vs. Ukraine: Netherlands advances with prob 0.68
Belgium vs. Serbia: Belgium advances with prob 0.67
France vs. Czech Republic: France advances with prob 0.73
England vs. Turkey: England advances with prob 0.71
Switzerland vs. Italy: Italy advances with prob 0.57
----------
Simulation of Quarter-Final
----------
Spain vs. Germany: Spain with prob 0.50
Portugal vs. Netherlands: Netherlands with prob 0.51
Belgium vs. France: France with prob 0.53
England vs. Italy: England with prob 0.60
----------
Simulation of Semi-Final
----------
Spain vs. Netherlands: Spain with prob 0.54
France vs. England: England with prob 0.50
----------
Simulation of Final
----------
Spain vs. England: Spain with prob 0.50
```

6. Conclusions

I'll discuss only the findings from Parts 2 and 3, as I've already covered the results of Part 1. Our correlation analysis revealed that success in the World Cup doesn't guarantee success in the Euros. Interestingly, the number of matches played correlates moderately at 0.54 with tournament progress, indicating a team's advancement. Notably, Goals Scored appears to have a greater impact on matches played than Goals Conceded in both competitions.

Regarding tournament difficulty, the Euros ranks among the top three toughest contests, though not the most challenging. Excluding certain competitions like CONMEBOL, Superclasico, and Copa Confraternidad, which aren't considered international, the FIFA World Cup emerges as the most demanding, followed by the Confederations Cup and then the UEFA Euro Cup. The chi-square statistic of 2186.72 highlights a significant disparity between observed and expected frequencies, with a p-value of 0.00 indicating a highly significant association between tournament type and the presence of high-ranking teams.

Regarding performance differences between host and non-host nations, our analysis yields a p-value of 0.57, indicating no significant distinction. Surprisingly, non-hosted teams hold a slight advantage in win proportion at 0.51 compared to hosted nations at 0.49. We employed sensitivity analysis to confirm these findings.

Part 3 provided an enjoyable exercise, revealing Spain as the projected winner of the final with a probability of approximately 0.50.

7. Future Work

With more time at my disposal, I would have expanded my analysis to include other tournaments besides Euro 2024. Specifically, I would have explored relationships between various soccer actions and events, such as possession and tactics, using a Graphical Markov Network. This approach involves defining nodes and edges to represent different events of interest, such as possession, goals, and tactics formations. By constructing a Markov network, we can depict the probabilistic relationships between these actions and events over time.

To achieve this, we would utilize inference algorithms like Gibbs (Markov Chain Monte Carlo) to analyze the Markov network. This process would enable us to uncover patterns, dependencies, and transitions between different actions and events, providing valuable insights into the dynamics of soccer matches.

Furthermore, I would have refined and integrated Part 3 of the analysis into the existing data integration file to enhance clarity and organization.