Mini Project 2

IST652 - Scripting for Data Analysis

1. **The data and its source:**

In this project, my original dataset included extensive international football statistics from numerous leagues and events. Kaggle's 'UCL, FC Barcelona(LaLiga), World Cup, and Euros Data' collection shed light on football events and club performances. I concentrated on 2019 Women's World Cup and 2018 Men's FIFA World Cup data to simplify my analysis and study. These data subsets cover match outcomes, scoring, and other important match statistics. Focusing on these datasets allowed me to explore the differences between these two important international football events. This technique provided a deeper understanding of international men's and women's football trends.

a. Dataset Details: The data for this project is derived from two primary sources: 2019_Womens_WorldCup.json: This dataset contains detailed information about the matches played in the 2019 Women's World Cup.
FIFA_WorldCup2018Men.json: This dataset encompasses match data from the 2018 Men's FIFA World Cup.

b. Contents of the Dataset: These datasets include a wealth of information on each match played in these tournaments, such as scores, team details, and other match-related statistics. This data is pivotal for conducting a thorough analysis of patterns and trends in international football tournaments.

2. A description of your data exploration and data cleaning steps

Loading Data

- Data Import: I began by importing necessary Python libraries, including json for handling JSON data and pandas for data manipulation and analysis.
- Custom Function: I defined a function load_json to facilitate the reading of JSON files. This function opens a specified file in read mode and parses the JSON data.
- Dataset Loading: Utilizing the load_json function, I loaded two datasets from the files '2019_Womens_WorldCup.json' and 'FIFA_WorldCup2018Men.json'. These datasets were then converted into Pandas DataFrames for easier handling and analysis.

Data Cleaning and Transformation

- DataFrame Conversion: After loading the data, I converted the JSON data into Pandas DataFrames. This step was essential to leverage the powerful data manipulation features of Pandas.
- Data Flattening: I flattened the DataFrames to transform nested JSON structures into a tabular format. This was achieved by normalizing the data and converting it into a list of dictionaries, making it more suitable for analysis.

- Preview of Data: To get an initial understanding of the data, I displayed the first few rows of each DataFrame. This step was crucial in identifying the structure and key variables within the datasets.

3. **Three clearly stated comparison questions with the unit of analysis, the comparison values and how they are computed.**

Question 1: What is the average number of goals per game in the Women's and Men's World Cups?
In this analysis, I aimed to compare the average goals per game between the 2019 Women's World Cup and the 2018 Men's FIFA World Cup. This was achieved by aggregating the total home and away scores for each match and then calculating their mean. The objective was to gain insight into the offensive characteristics of the tournaments, highlighting potential differences or similarities in the gameplay and strategies between the women's and men's competitions.

Question 2: What is the most common scoreline in the Women's and Men's World Cups?
A key aspect of my study was to determine the most common scoreline in both the Women's and Men's World Cups. This involved grouping matches based on their final scores and identifying the most frequently occurring scoreline. The aim was to reveal the most typical match outcomes, providing an understanding of the competitive nature and level of play across the tournaments.

Question 3: How are match scorelines distributed in the Women's and Men's World Cups?
In this part of the analysis, I focused on the distribution of match scorelines in both tournaments. By categorizing games based on their home and away scores and counting the number of occurrences for each unique scoreline, I was able to examine the variety and frequency of different match outcomes. This analysis was vital for understanding the diversity of scoring patterns, offering a window into the range and unpredictability of results in top-level international football.

Question 4: How do team performances and match distributions compare in the Women's and Men's World Cups?
The final aspect of my project involved a detailed examination of team performances and match distributions in the World Cups. I calculated the total goals scored by each team, analyzed the outcomes of matches for home and away teams, and studied the distribution of matches across different days of the week. This comprehensive approach helped in understanding the scoring abilities of teams, their success rates in various match situations, and the general pattern of match occurrences throughout the tournaments. The analysis provided a nuanced view of the dynamics in both the Women's and Men's World Cups, highlighting the intricacies of team performances.

4. **A description of the program**

i) Data Loading and Normalization:
JSON Parsing: The program begins by importing data from JSON files using the json library. This approach is essential for handling nested data structures typical in JSON format.
DataFrame Creation: After parsing, the data is converted into pandas DataFrames. The use of DataFrames is a strategic choice for efficient data manipulation and analysis.Data Flattening: pd.json_normalize is employed to transform the nested JSON data into a flat, tabular structure, making it more suitable for analysis.

ii) Data Analysis Techniques:
Statistical Computation: The program calculates average goals per match and identifies the most common scorelines. This involves aggregating data and using groupby operations to summarize the data effectively.
Comparative Analysis: The code is structured to compare different aspects of the datasets, such as scorelines, goals, and match outcomes, highlighting differences and similarities between the two tournaments.

iii) Data Transformation and Categorization:
Custom Functions: The program utilizes custom Python functions to categorize matches based on total goals. This demonstrates the flexibility of Python in tailoring data processing to specific analytical needs.
Aggregation: Grouping and aggregating functions are extensively used to create meaningful categorizations and summaries of the data.

iv) Output Generation:
CSV File Creation: The program outputs its findings into CSV files. This is achieved through pandas' to_csv method, indicating a focus on data sharing and reporting.
Structured Data: The output CSV files are structured to provide clear insights, such as score distributions, average goals, and common scorelines, in a format that is accessible and easy to interpret.

5. **A description of the output files**

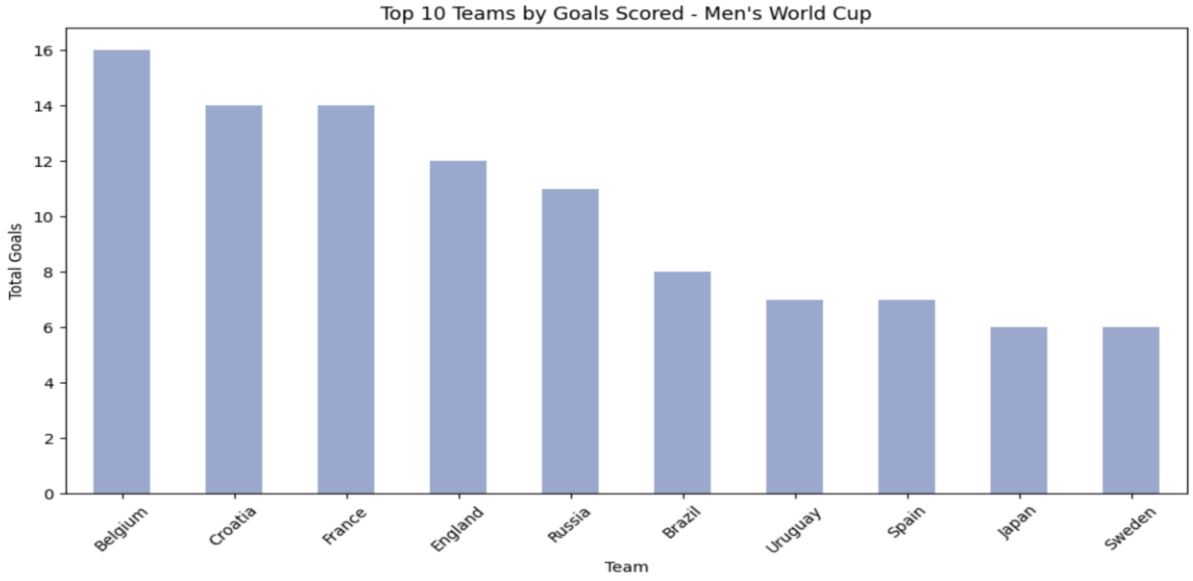
1. Complete Comparative Analysis (Complete_comparitive_analysis.csv)
This file provides a comprehensive comparison between the Women's and Men's World Cup data across multiple dimensions:

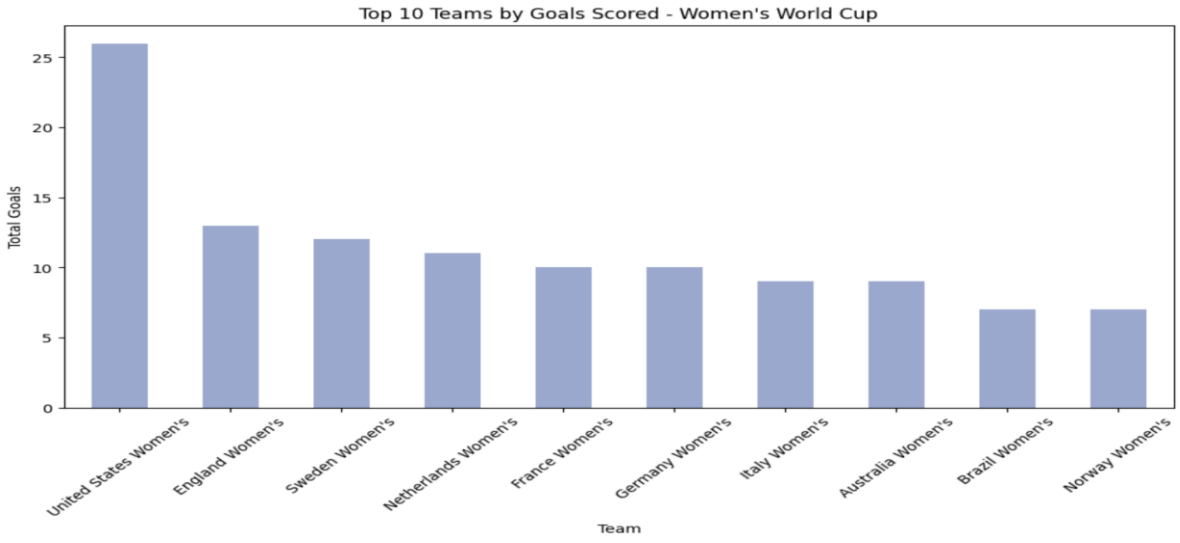
- **Teams and Goals**: It lists the top teams from each gender category along with their total goals scored.
- **Home and Away Performance**: The file details the number of draws, losses, and wins for both home and away teams, allowing for an understanding of how teams perform in different settings.

- **Match Days and Counts**: It also includes information about the days on which matches were played and the total count of matches, providing insights into the scheduling and frequency of games.

Key Findings:

- The United States Women's team scored the highest total goals (26) in the Women's World Cup.
- In the Men's World Cup, Belgium's home team had the highest number of wins (4).
- The highest number of matches on a single day was 12, occurring on a Saturday in the Men's World Cup.





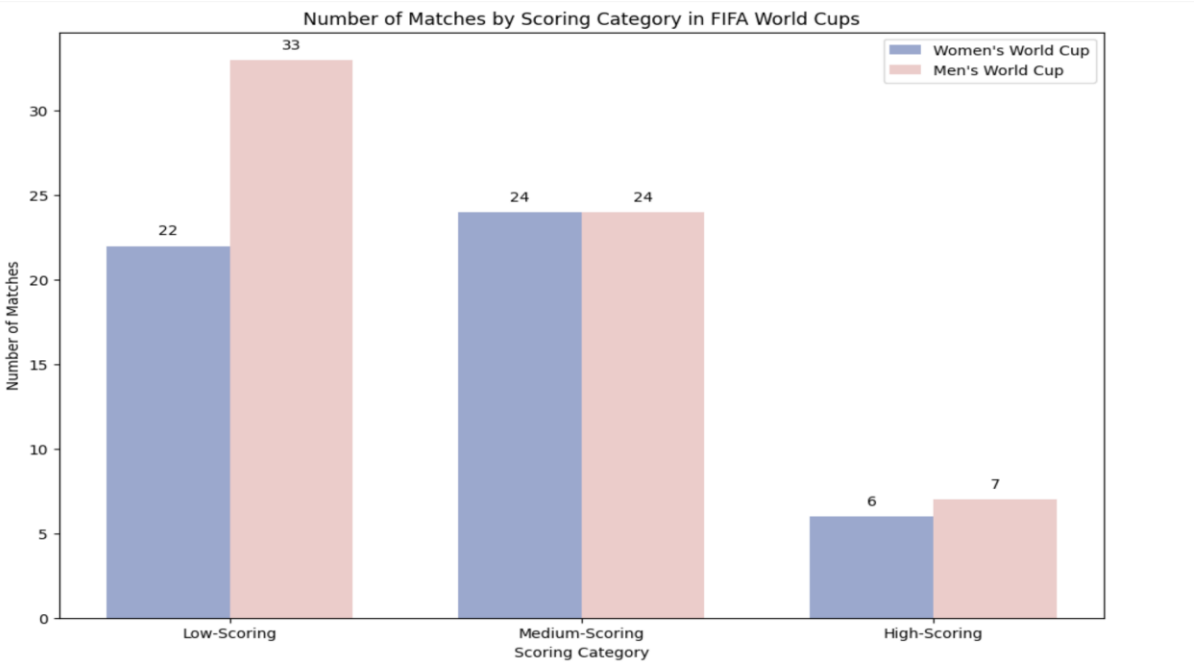
2. Categorized Goals (Categorized_Goals.csv)

This file categorizes matches based on the total number of goals scored (low-scoring, mid-scoring, high-scoring) for both Women's and Men's World Cup:

- **Match Count and Score Distribution**: It shows the distribution of home and away scores for matches in different scoring categories.
- **Gender Comparison**: The file facilitates a comparison between women's and men's matches in terms of scoring patterns.

Key Findings:

- In the Women's World Cup, there were 6 high-scoring matches with an average home score of 4.17 and an away score of 2.5.
- The Men's World Cup had 33 low-scoring matches with close home (0.76) and away (0.73) scores



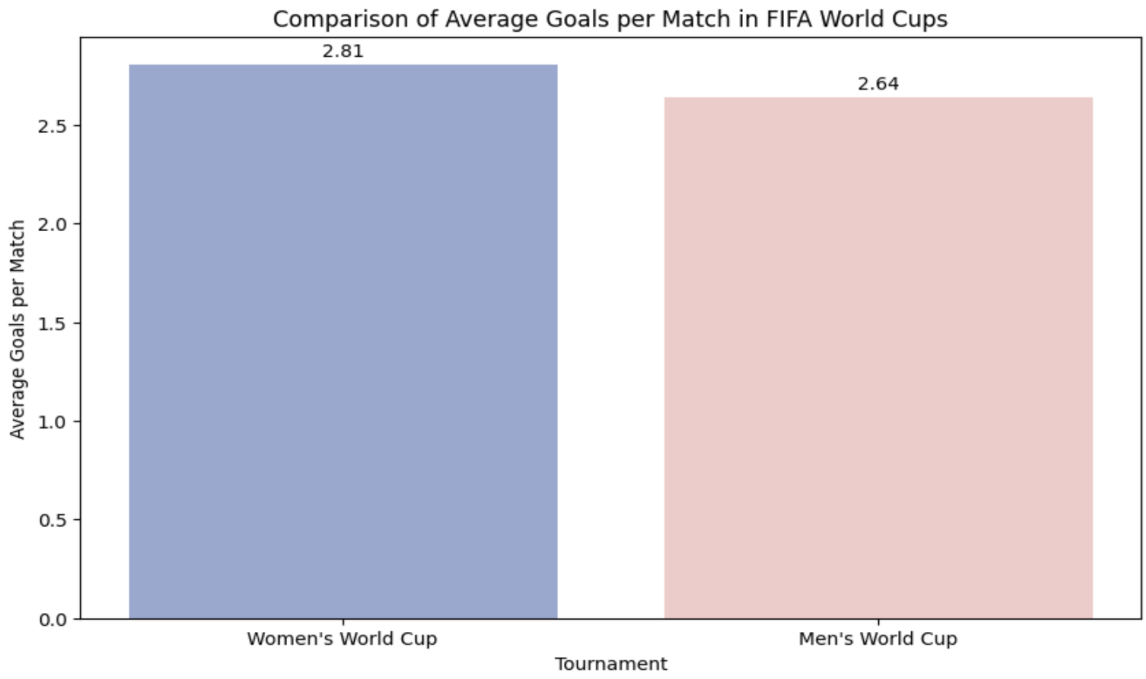
3. Total Goals and Common Scoreline (Total_Goals, Common_Scoreline.csv)

This file summarizes the average goals per match and the most common scoreline in each tournament:

- **Average Goals**: Indicates the average number of goals scored in each match for both tournaments.
- **Most Common Scoreline**: Highlights the scoreline that occurred most frequently.

Key Findings:

- The average goals per match were 2.81 in the Women's World Cup and 2.64 in the Men's World Cup.
- The most common scoreline was (1, 2) in the Women's World Cup and (0, 1) in the Men's World Cup.



4. Scoring Patterns Comparison in Different Stages (Scoring patterns comparison in different stages.csv)

This file compares the scoring patterns in different competition stages of both tournaments:

- **Stage-wise Breakdown**: Provides home and away scores, along with the number of matches for various competition stages.
- **Comparative Analysis**: Enables an understanding of how scoring trends varied across different stages of the tournaments.

Key Findings:

- In the Women's World Cup Final, the home score was 2 with no away goals, while in the Men's Final, the home score was 4 with 2 away goals.

5. Scoreline Distribution (Scoreline_Distrubution.csv)

This file details the distribution of different scorelines in the tournaments:

- **Scoreline Frequency**: Lists the frequency of each specific scoreline (home score and away score).
- **Gender-Based Comparison**: Allows for comparing how often certain scorelines occurred in each gender's tournament.

Key Findings:

- In the Women's World Cup, the (0, 2) scoreline occurred 5 times, while in the Men's World Cup, the most frequent low-scoring match (0, 1) occurred 33 times.

6. Conclusion

The analysis of the 2019 Women's World Cup and the 2018 Men's FIFA World Cup has yielded insightful revelations about the distinct nature of scoring patterns and strategic play in international football. The study highlighted the higher-scoring games in the Women's World Cup, contrasting sharply with the closely contested, low-scoring matches in the Men's World

Cup. This difference not only underlines tactical variations but also emphasizes the unique qualities of each tournament.

A deeper look at total goals and prevalent scorelines further enriched our understanding of the offensive strategies employed by teams in both genders' tournaments. The Women's World Cup was marked by a tendency towards dynamic, attack-oriented play, while the Men's World Cup showcased a more calculated, defense-focused approach.

The stage-wise analysis revealed how teams' strategies evolved through the tournament phases, especially in high-pressure final stages, influencing the game's outcome. These findings offer a nuanced view of the strategic dimensions that define men's and women's football at an international level.

This project, through its data-driven approach, not only quantifies key aspects of football but also narrates the evolving story of this sport on a global stage. The insights gained here are valuable for analysts, coaches, and fans, offering a deeper appreciation of the sport's complexity and its continuous evolution.