World Cup Events

Constraints

1. **Programming Language:** Python
2. **Libraries:** Numpy, Scipy, Pandas, Plotly, Scikit-Learn
3. **IDE:** Jupyter Notebook, Colab
4. **Datasets:**
    a. The Fjelstul World Cup Database. (https://github.com/jfjelstul/worldcup)
    b. World Cup Attendance Dataset. (https://drive.google.com/file/d/1-4FNJB6T5LMpSMOtPv3WIa7nOjTFAC3z/view)

World Cup Events

Summary

1. **Topic:** Data Mining
2. **Goal:** Analysis of World Cup Events
3. **Introduction:**

   The FIFA World Cup stands as the pinnacle of international football, orchestrated by the eminent governing body, FIFA. This transcendent tournament, which has seen 22 editions as of the 2022 FIFA World Cup, serves as a spirited battleground for 80 national teams from across the globe.

   As a top-notch football event, the World Cup grabs attention from around the world, drawing fans from all corners. Your task is to dig into the stories of this prestigious competition, finding interesting facts about the tournaments, exciting matches, famous teams, standout players, and the respected stadiums that hold the tales of football history.

World Cup Events

Tasks

1. Data Cleaning and Integration

a. Fill in the gaps

- We have data from different sources and need to combine them in order to get a more complete picture of the World Cup events, So we are interested in merging the attendance dataset and the Fjelstul dataset.
- The final dataset contains all the attributes from the match table as well as The number of crowd attendance from attendance dataset , stadium capacity from the stadium table in Fjelstul dataset.
- To ensure the integrity and completeness of the expanded data set, we handled issues such as null values, column transitions, and duplicate items.

World Cup Events

Tasks

1. Data Cleaning and Integration

b. From rough to polished

- According to FIFA rules, a player is allowed to represent only one national team in official competitions, including the World Cup, however there have been a few instances where a player has played for more than one national team in his career, but not in the same tournament. So we'll **Create** a new data frame player_teams using the teams in Squads table and players in Players table in the Fjelstul dataset.

- The resulting dataset includes player's *first name*, *last name*, *number* and *name of tournaments* (a list) in which he participated from the Players table, along with *team names* (a-list), *team symbols* (also a list), and the *number of teams* that the player represented during his career in the World Cup.

World Cup Events

Tasks

2. Features Engineering

We'll create the following features:

- *total goals in match*, *match for host* (binary feature indicates if the host team is playing) and *used capacity ratio* in the matches table.
- *attendance category* depending on the attendance feature and *relative attendance category* depending on the used capacity ratio feature using Discretization.
- *host country code*, *tournament year*, *full name* (for players in a readable format).
- *winner code* in tournaments table.
- *short stage name* which includes knockout and group stages only.
- *late goal* (binary feature denotes whether a goal was scored late in the match), The goal minute and halftime of the match can be helpful for this feature.

World Cup Events

Tasks

3. Exploration and analysis

a. Attendance case study

We’ll plot the mean and median of this attribute using line chart within the same chart, the histogram distribution of the attribute (Attendance) using an appropriate number of bins and the attribute distribution within each round of the tournament using box plot. then we’ll write our conclusions.

World Cup Events

Tasks

3. Exploration and analysis

b. Goals case study

We'll:

- Plot the mean (or median) goal period for each edition of the tournament using bar chart.
- Plot histogram of total number of goals per match in the World Cup.
- Calculate the most frequent goal minute and time duration for each edition of the tournament.
- Plot histogram of total late goals in each edition of the tournament.
- Use bar chart to plot the top 12 goal scorers of all time at the World Cup.
- Use bar chart to plot the top scorer in each edition of the tournament.
- Use bar chart to plot the total number of goals in each edition of the tournament.
- Consider Brazil, Germany and Italy, and use a strip plot for the minute of goal and the short stage name.

World Cup Events

Tasks

3. Exploration and analysis

c. Matches case study

We'll:

- Calculate match frequencies throughout World Cup history, taking into account that *away_team* and *home_team* are interchangeable World Cup attributes!

- Plot the 10 most frequent matches in the World Cup using bar chart

World Cup Events

Tasks

3. Exploration and analysis

d. Tournament case study

We'll:

- Identify the group of players who represented more than one team and then try to discover the reasons behind this phenomenon.
- Identify whether there is a relationship between the host country and the tournament winner.
- Identify if there is a correlation between the match host and the crowd attendance rate category.
- Identify if there is a correlation between the host country and the category of public attendance