# 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 in order to get a more complete picture of the World Cup events we need to combine them, 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 a new data frame player_teams using the teams in Squads table and players in Players table in the Fjelstul dataset will be **created**.

- 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

The following features will be **created**:

- *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

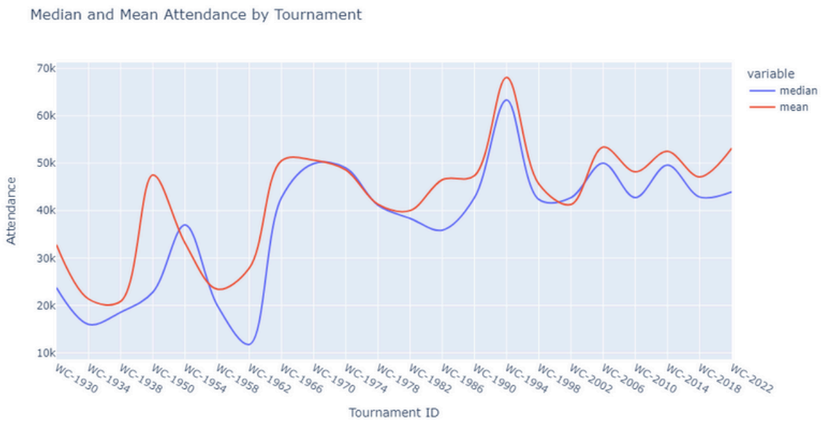
## Tasks

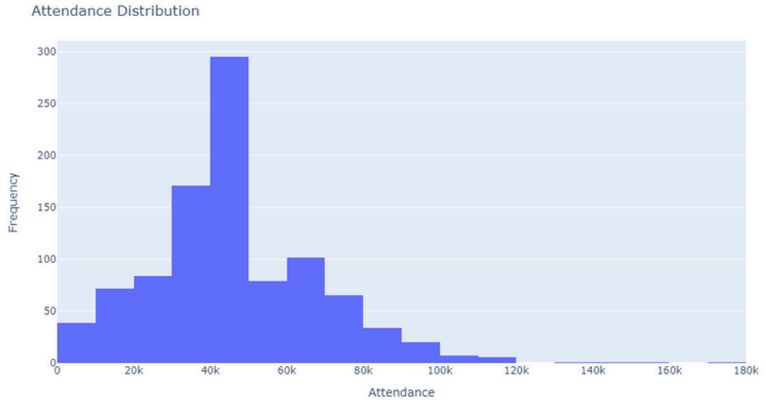
### 3. Exploration and analysis

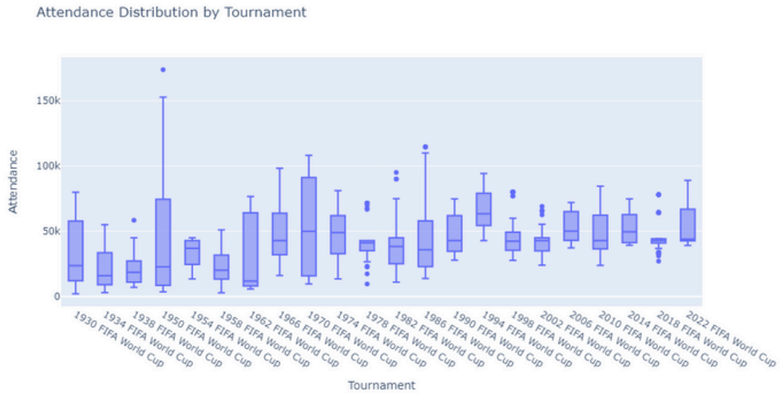
#### a. Attendance case study

I’ve plotted the mean and median of this attribute (Attendance) using a line chart, the histogram distribution of the attribute (Attendance) and the attribute distribution within each round of the tournament using box plot.

# World Cup Events




**Conclusion:**

the most common attendance is between 40k - 50k, we can see that the attendance has increased over the time. minimum attendance 1930 was (2000) but minimum attendance 2022 was (39000)

# World Cup Events

## Tasks

### 3. Exploration and analysis

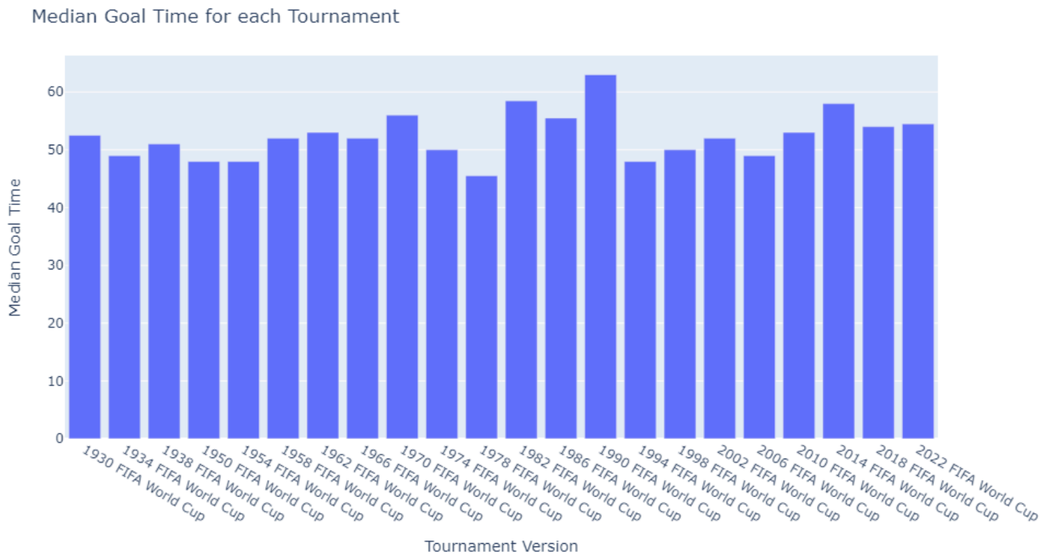
#### b. Goals case study

I’ve done the following:

- Plotted the median goal period for each edition of the tournament using bar chart.
- Plotted histogram of total number of goals per match in the World Cup.
- Calculated the most frequent goal minute and time duration for each edition of the tournament.
- Plotted histogram of total late goals in each edition of the tournament.
- Used bar chart to plot the top 12 goal scorers of all time at the World Cup.
- Used bar chart to plot the top scorer in each edition of the tournament.
- Used bar chart to plot the total number of goals in each edition of the tournament.
- Considered Brazil, Germany and Italy, and used a strip plot for the minute of goal and the short stage name.

# World Cup Events

Median goal period plot

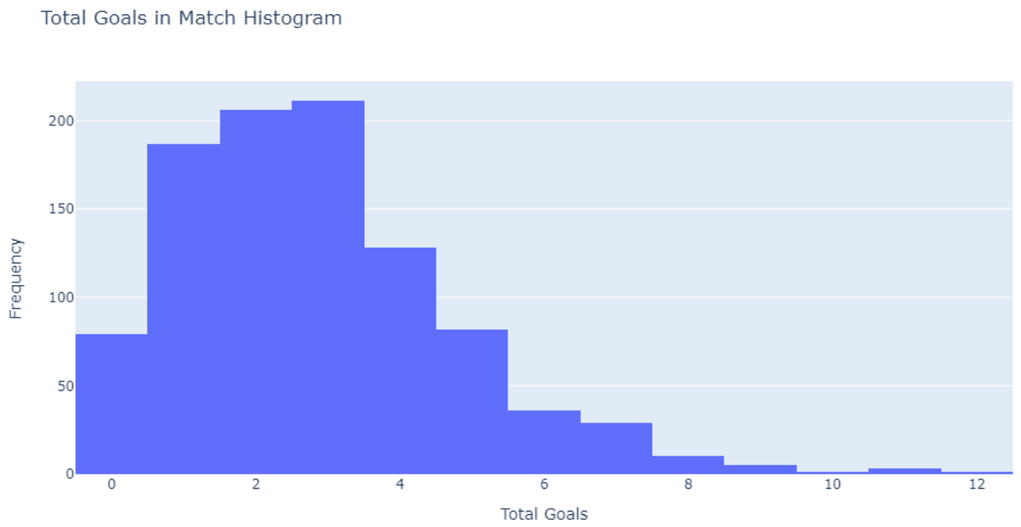



**Conclusion:**

The plot reveals that most tournaments exhibit a median goal time between 40 and 50 minutes, indicating goals are typically scored later in the first half or early in the second half.

# World Cup Events

Total number of goals histogram

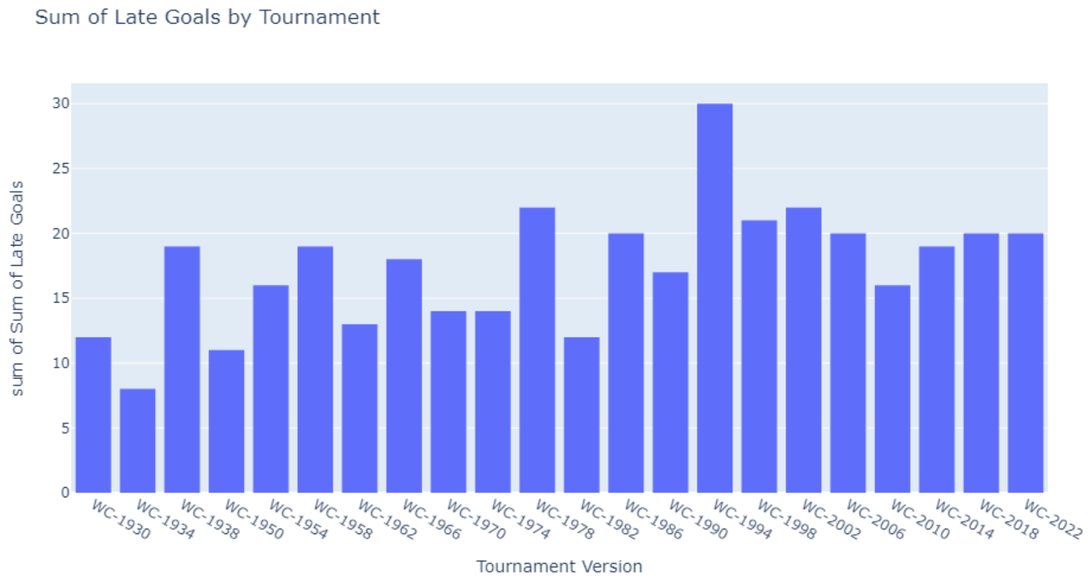



**Conclusion:**

The histogram analysis reveals a peak frequency at matches with 3 total goals, indicating their prevalence in World Cup matches. As total goals per match increase beyond 3, frequency decreases, suggesting that matches with higher goal counts are less common occurrences in World Cup history

# World Cup Events

Total late goals histogram

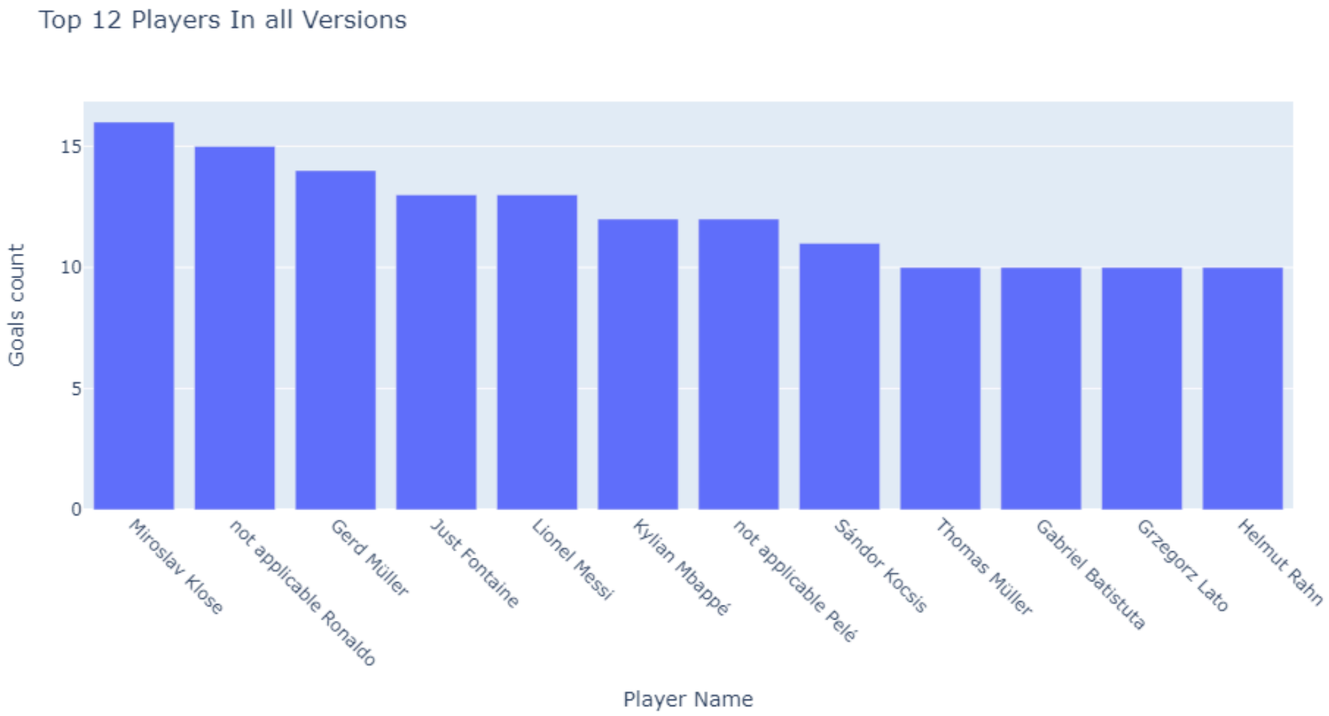



**Conclusion:**

The plot reveals that late goals vary across FIFA World Cup tournaments, with counts ranging from 8 to 30. While fluctuations exist, most tournaments feature double-digit late goals, emphasizing their significance in match outcomes and tournament drama.

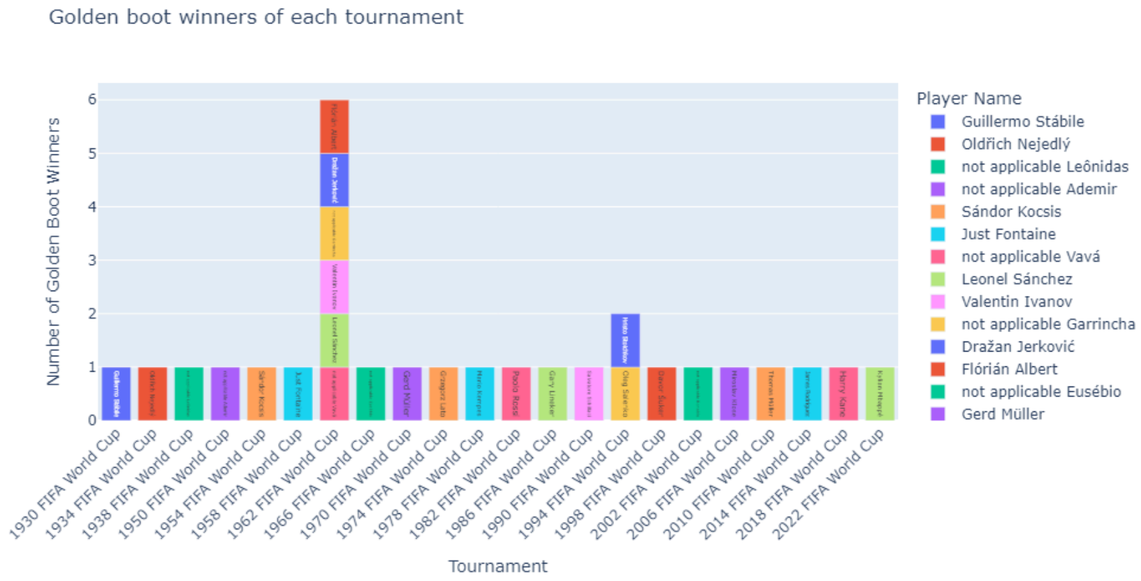
# World Cup Events

Top 12 goal scorers of all time



# World Cup Events

Top scorer in each edition

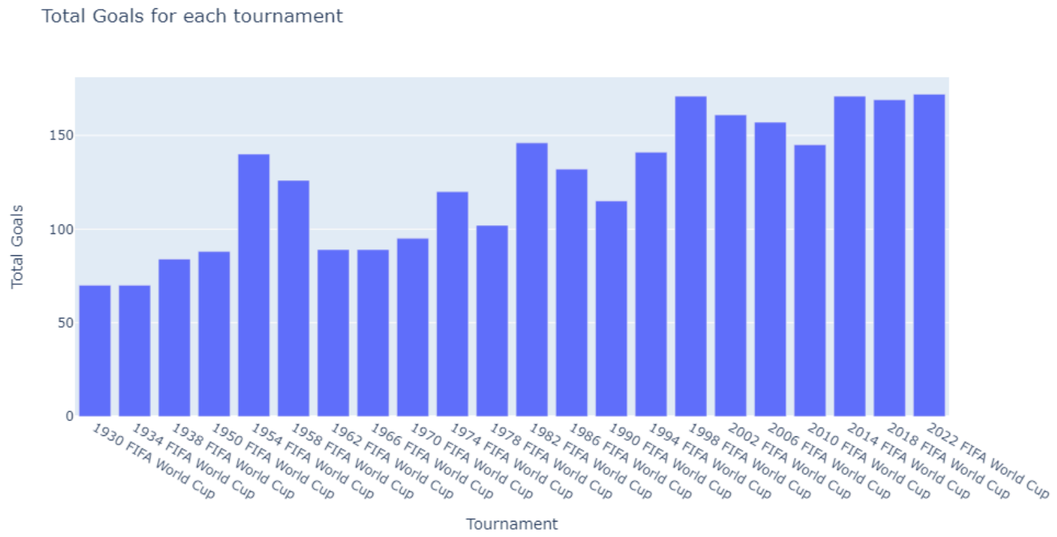



**Conclusion:**

In some of the old versions the golden boot award were shared between more than one player in the same version which is strange phenomena now adays.

# World Cup Events

Total number of goals in each edition

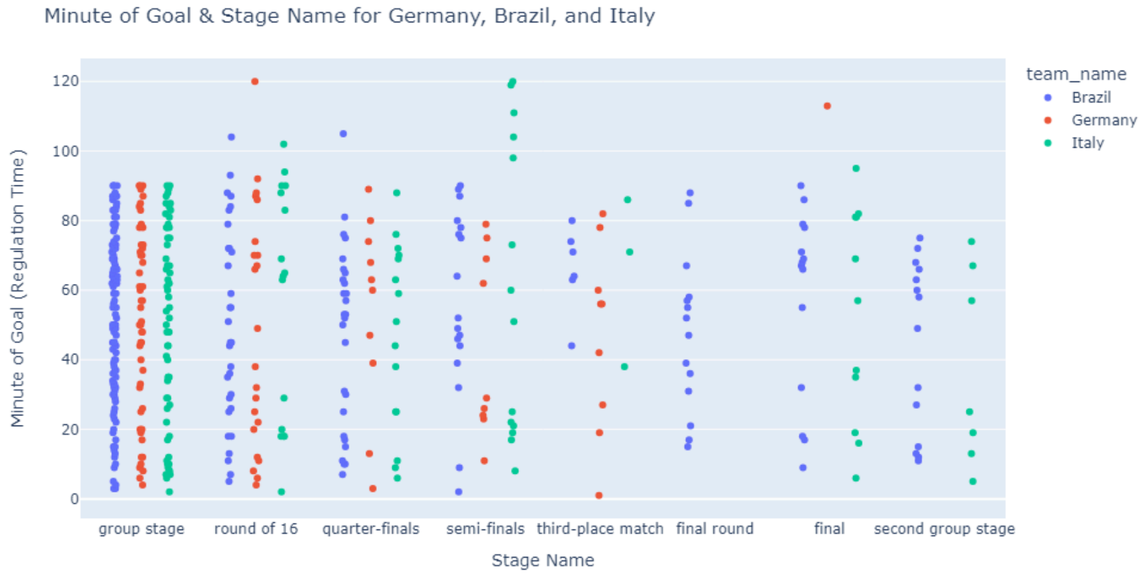



**Conclusion:**

we can see the increase in FIFA world cup goals count in recent versions. This might be a result for many reasons like: Evolution of tactics, Rule changes, Improvement in player skills, Changes in team dynamics, Advancements in sports science and training methods.

# World Cup Events

minute of goal and the short stage name




**Conclusion:**

germany had the most goals scored during various stages of matches across different tournaments, with notable spikes in the group stage and quarter-finals.

brazil had a consistently high number of goals scored across different stages, with peaks in the group stage and final round.

italy had fewer goals compared to Germany and Brazil, with peaks in the group stage and semi-finals. Italy's performance appears to be relatively consistent across different stages.

# World Cup Events

## Tasks

### 3. Exploration and analysis

#### c. Matches case study

I have:

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

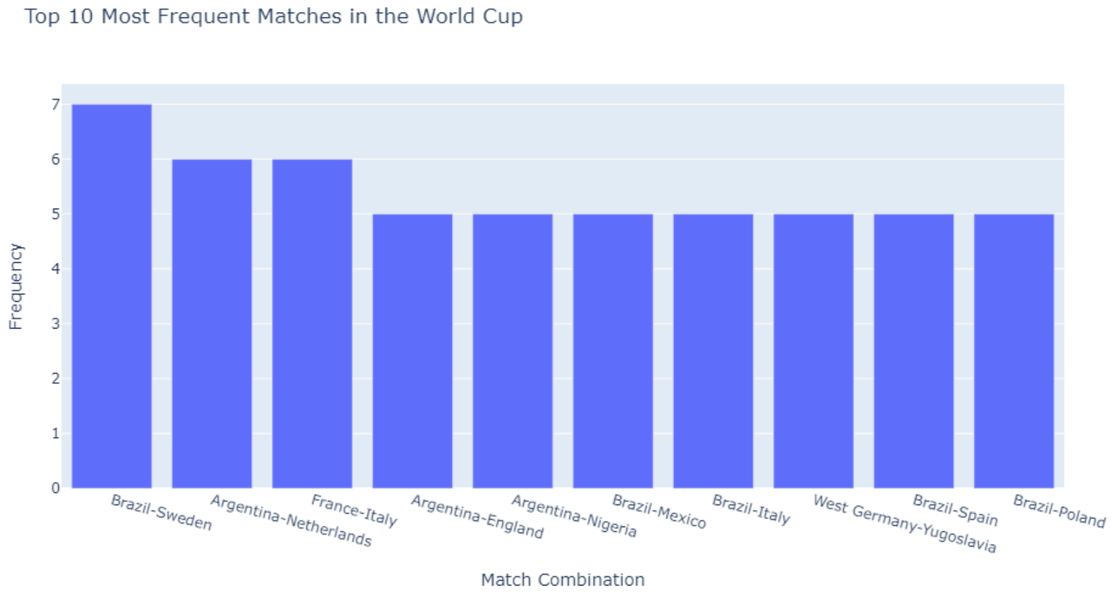
- Plotted the 10 most frequent matches in the World Cup using bar chart

# World Cup Events

## Tasks

### 3. Exploration and analysis

#### c. Matches case study



# World Cup Events

## Tasks

### 3. Exploration and analysis

#### d. Tournament case study

I’ve done the following:

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