TECHNICAL REPORT:
Player & Match Events Database for 1998-2018 FIFA World Cup

Introduction

Amidst the excitement of this year's world cup, we decided to build an SQL database to capture player performance and to examine if match events (e.g., tournament stage) have any impact on said performance. Users will be able to delve further into statistics about players and teams with the help of this dataset. As the data that we extracted was tabular and structured in nature, a relational SQL database was used to store and query the data tables.

If you would like to find out more about how we went about developing this database, here is the link to the Project's github repository: https://github.com/vn02063007/project-group-2.

Data Sources

We adapted a comprehensive database by Dr Joshua Fjelstul (2022) that summarised the data from 21 World Cups held between 1930-2018. Although it was based primarily on information extracted from Wikipedia, official FIFA match reports were also used as supporting sources.

Given the depth of this dataset, we decided to look at a smaller sample of tournaments. Specifically, we chose to focus on recent tournaments between 1998-2018. These tournaments had the same number of matches and the rules of that era were largely similar. These considerations allowed us to restrict potential confounds to the insights that the current database provides.

There is also an opportunity for users to further reduce the data by focusing on tournaments held between 2006-2018 when the Golden and Silver goals were removed from the rules of football. In this period, two straight 15-minute periods of extra time were played in the event of a drawn match during the knockout stages. If no winner is decided after extra-time, the game will go into a penalty shoot-out.

Data Cleaning & Transformation

The data is pulled from a github repository which contains the World Cup information from 1930 to 2018. For our own purpose, we only focus on the World Cup 1998-2018.

From the github repository, we pulled in total 27 csv files but decided to focus on 7 main csv files that we are interested in (see table below)

Cleaning csv file:

- Load csv file using pd.read function
- Create a new dataframe which only contains related columns. This way we dont overflow with unwanted columns
- Create a new column called “year” in order to identify the year of the event as not every csv file contains this info. Extracted from another column string. For example: In column tournament_id, we got WC-1998, we extracted the last 4 letters as the year and put it in the new column.
- Filter the year > 1998 using new “year” column or “match_date” column
- Show result by calling new dataframe.head

```
new_goals['year']= new_goals['tournament_id'].str.strip().str[-4:]
new_goals = new_goals.loc[(new_goals['year'] > '1997')]
new_goals.head()
```

- Sorted by year or match id if needed

Load cleaned dataframe into local database:

- Connect to local database
- Check for table name
- Use pandas to load csv convert DataFrame into database
- Confirm data has been added by querying the database table

```
# Check that data has been fed into table
pd.read_sql_query('select * from matches', con=engine).head()
```

- Use ‘pd.read_sql_query’ command to join 2 tables to get desire information

```
# Join goals with squad to obtain a table that records player information such as position.

sql_join = r"""SELECT * FROM goals
JOIN squads
ON goals.player_id = squads.player_id"""
pd.read_sql_query(sql_join, con=engine).head()
```

For more information, refer to each jupyter notebook file in the “resources” folder for the table you are interested in

Application

In this section, we discuss some of the potential applications of our database through the discussion of potential questions:

1. Review the total number of goals scored in open play by position and nation.

Firstly, a user could join the ‘squads’ and ‘goals’ tables in this database to create a table that linked a player’s information such as nationality and position to the goals they scored open play during a particular year’s World Cup. Further drill-downs could be used to exclude own-goals and penalties.

2. Examine whether player experience is associated with the amount of goals they have scored.

Next, a user could join the ‘player_appearance’ and ‘goals’ tables in the database to create a dataset that could allow them to count the number of goals scored by a selected player and correlate that with a count of their appearances as either a starter or a substitute. Such a count would provide an estimate of their experience at playing at an elite international level.

3. Explore the number of goals scored at each stage of the tournament.

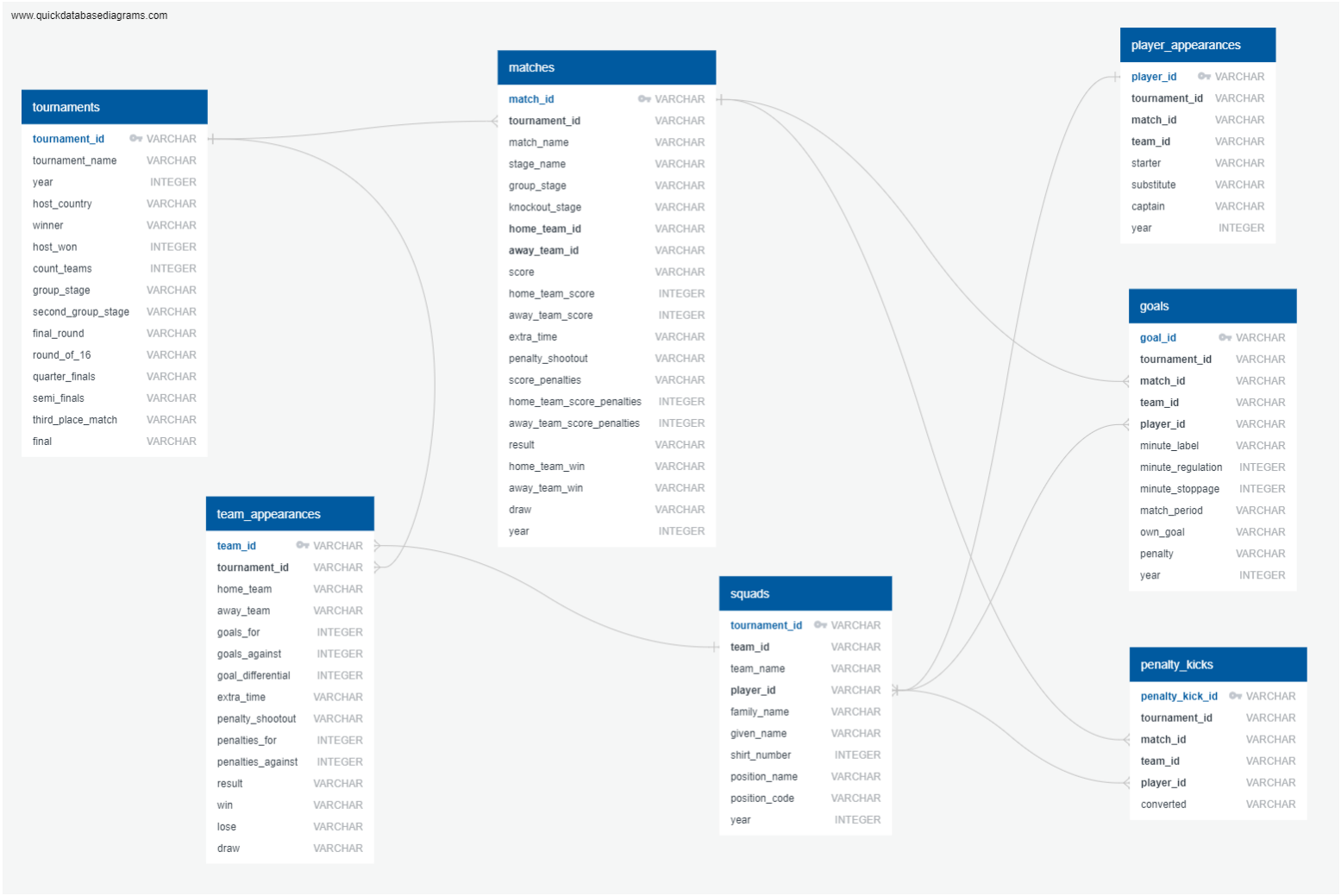
Finally, a join of the tables ‘matches’ and ‘goals’ could provide a summary of how many goals were scored at each stage of the tournament. It is often said that teams get cagey in the business of the World Cup (i.e., in the knockout stages) - a user could explore this notion further with some light wrangling of the data.

7 selected csv files

TABLE	DESCRIPTION
goals	This dataset records all goals. There is one observation per goal. It notes the team that scored the goal, player who scored the goal, the team of the player who scored the goal (to account for own goals), minute of the goal, and whether the goal was scored in the run of play by the opposition, was an own goal, or was a penalty. This dataset does not include converted penalties in a penalty shootout.
matches	This dataset records all World Cup matches between 1998-2018. There is one observation per match per tournament. It includes the home team, the away team, the final score, the score margin for each team, whether the match went to extra time, whether there was a penalty shootout, the number of penalties scored in the shootout (if applicable), the result of the match (home team win, away team win, draw, replayed), and the winner (if applicable).
penalty_kicks	This dataset records all penalty kicks taken during penalty shootouts. There is one observation per penalty kick. This dataset does not include attempted penalty kicks during matches. That information can be found in the 'goals'. It keeps a record of the player who took the kick, and whether the penalty was converted.
player_appearances	This dataset records all player appearances in the matches for the World Cup that occurred between 1998-2018. There is one observation per player per team per match per tournament. It includes players who play in the match, including players who are in the starting eleven and players who come in as substitutes.
squads	This dataset records the composition of each squad. There is one observation per player per team per tournament. It includes the position of each player and the shirt number of each player.
tournaments	This dataset records all the World Cup tournaments held between 1998-2018. There is one observation per tournament. It includes the host of the tournament, the winner of the tournament, the year of the tournament and the number of teams participating in each edition of the World Cup.
team_appearances	This dataset records all team appearances. There is one observation per team per match per tournament. It includes whether the team is the home team or the away team, the number of goals for and against, the goal difference, whether there was a penalty shootout, penalties for and against (if applicable), and whether the team wins, loses, or draws.

Structure of Final Database

The entity relational diagram below provides an overview of the 7 datasets and how they relate to each other.



Opportunities for improvement

1. Introduce a variable that captures match fatigue

Introducing a variable that summarises distance covered by a player in the ‘player_appearances’ table would enable users to examine whether fatigue affects player performance. There is immense potential for examining questions on whether the distance a player covers in a match affects their ability to successfully score penalties should a match remain a draw even after two halves of extra-time.

2. Focus on the dataset is on a single tournament

Given that the current database focuses on just the World Cups, there is an opportunity to introduce data from other international tournaments such as the Copa America, African Nations Cup, Confederations Cups and Asian Cup.

3. Introduce a variable that captures minutes played

Although users could utilise the ‘starter’ and ‘substitute’ variables to operationalise a player’s experience playing at the highest level of international football, capturing the minutes that they were on the pitch for could add further nuances for any analysis that seeks to examine the impact of experience.

References

Fjelstul, Joshua C. "The Fjelstul World Cup Database v.1.0." July 8, 2022. https://www.github.com/jfjelstul/worldcup.