Phase 1
Uzair Chudhary and Elhan Syed

Domain

For both of us, sports has always been a crucial branch of our lives; whether it was a means of making conversation, constructing new friendships, or getting in a bit of physical activity, sports has played an immense role in our day-to-day lives. Uzair has also coached middle school soccer for multiple years. For that reason, the both of us decided to look into the domain of sports, specifically our favorite sport, soccer (henceforth referred to as football).

Dataset

The dataset we settled on is related to European football, and can be found at https://www.kaggle.com/hugomathien/soccer.

This dataset has 7 tables (Country, League, Match, Player, Player Attributes, Team, Team Attributes), all of which will be used. However, from the table 'Match', 'Player_Attributes', and 'Team_Attributes' not all of the columns will be used. Currently, we are expecting the following columns to be used from the previously stated tables:

'Match': id, country_id, league_id, season, date, homeTeamID, awayTeamID, homeTeamGoal, awayTeamGoal, shoton, shotoff, foulcommit, card, cross, corner, possession.
'Player_Attributes': id, date, overallRating, potential, attackingWorkRate, defensiveWorkRate
'Team_Attributes':id, date, buildUpPlaySpeed, buildUpPlayDribbling, buildUpPlayPassing, chanceCreationPassing, chanceCreationCrossing, chanceCreationShooting, defencePressure, defenceAggression, defenceTeamWidth

In order to really understand and interpret the data, there is certainly some learning that we will have to do. First and foremost, currently the dataset is downloadable in SQLite format. Having never worked with this RDBMS (relational database management system) type before, we might have to learn how to convert the tables into a flat file (i.e. CSV, JSON).

Additionally, in the 'Match' table, the data stored in columns goal, shot on, shot off, foulcommit, card, cross, corner and possession are in XML format. In order to move forward with the analysis, we will either have to find a way to work with the XML data, or clean the data and make it more readable.

To query this dataset, and ultimately make the relationships and connections required for analysis, the both of us will be required to learn more about advanced joins, and how to combine multiple joins without duplication.

Moreover, we will attempt to make the dataset more readable by renaming id columns found within each table (i.e. Country[id] -> Country[countryID]). We will also change all of the columns to follow the camelcase standard (i.e. Player_Attributes[defensive_work_rate] -> Player_Attributes[defensiveWorkRate]).

Further cleaning in this dataset will be ongoing, and as the team sees the lack of use of certain columns (in any specific table), the team will work towards removing those columns in order to reduce and optimize queries and querying times.

Questions

The evolution of modern football, according to the eye test, has placed less emphasis on possession. What is the impact of possession on winning games?

What are the most effective offensive and defensive footballing tactics?

Work rate is highly emphasized by grassroots coaching. Are the most successful teams and players hard working or efficient with their movement?

Schema

Following is the schema of the dataset.

Relations

Country(countryID, countryName)

A tuple in this relation represents a country. The *countryName* is obvious and the *countryID* is a numerical identifier.

League(leagueID, countryID, leagueName)

A tuple in this relation represents a professional football league with a *leagueID* numerical identifier, *countryID* representing which nation the league is based in, and *leagueName* is merely the official name of the league.

Match(matchID, countryID, leagueID, season, date, homeTeamID, awayTeamID, homeTeamGoal, awayTeamGoal, shotOn, shotOff, foulCommit, card, cross, corner, possession)

A tuple in this relation represents a completed football match with a *matchID* numerical identifier, *countryID* referring to where the match was played, *leagueID* referring to the league that the teams played in, *season* is the specific year in the league that the match took place in, *date* is the exact date of the match, *homeTeamID* and *awayTeamID* are the numerical identifiers of the teams involved in the match where home and away are designated by which team hosted the match at their home stadium, *homeTeamGoal* and *awayTeamGoal* tells us how many goals were scored by the respective teams, *shotOn* and *shotOff* refer to the how many shots were on target (reached the area within the goal posts) and how many were not, *foulCommit* details the number of fouls called by the referee, *card* refers to the number of yellow or red cards shown by the ref, *cross* refers to the number of crosses delivered into the box, *corner* refers to how many corners were awarded by the ref, and *possession* lists what percentage of the match each team had the ball.

Player(playerID, playerName, birthday, height, weight)

A tuple in this relation describes the biographical data of a professional football player with *playerID* being their numerical identifier, *playerName* being their legal name, *birthday* being their date of birth, and *height* and *weight* referring to their measured height and weight.

PlayerAttributes(playerID, date, rating, potential, attackingWorkRate, defensiveWorkRate)

A tuple in this relation represents a player's footballing abilities with *playerID* being their numerical identifier, *date* being the date the analysis was conducted, *rating* being their overall score as a footballer, *potential* being how high their rating could hypothetically reach and a measure of their natural talent, *attackingWorkRate* being the percentage of effort they put when their team is on the attack, and *defensiveWorkRate* being the percentage of effort they put when their team is on the defence.

Team(teamID, teamName, leagueID, teamAbbreviation)

A tuple in this relation represents a professional football team with the *teamID* being their numerical identifier, *teamName* as the official name of the team, *leagueID* referring to which football league they play in, and *teamAbbreviation* being the team's official name abbreviation.

TeamAttributes(teamID, date, buildUpPlaySpeed, buildUpPlayDribbling, buildUpPlayPassing, chanceCreationPassing, chanceCreationCrossing, chanceCreationShooting, defencePressure, defenceAggression, defenceTeamWidth)

A tuple in this relation represents a team's footballing abilities with the *teamID* being their numerical identifier, *date* being the date the analysis was conducted, *buildUpPlaySpeed* being the speed at which they play when they have possession of the ball, *buildUpPlayDribbling* is how often individual players make runs with the ball, *buildUpPlayPassing* is how often the team passes the ball when in possession, *chanceCreationPassing* is how often the team creates a chance to score with their passing, *chanceCreationCrossing* is how often the team creates a chance to score with their crossing, *chanceCreationShooting* is how often the team creates a chance to score with their shooting, *defencePressure* is how closely they defend when the opposition has possession, defenceAggression is how aggressively they defend when the opposition has possession, and *defenceTeamWidth* is how spread out their defence is.

Integrity Constraints

League[countryID] ⊆ Country[countryID]

Match[countryID] ⊆ Country[countryID]

Match[homeTeamID] ⊆ Team[teamID]

Match[awayTeamID] ⊆ Team[teamID]

Match[leagueID] ⊆ League[leagueID]

PlayerAttributes[playerID] ⊆ Player[playerID]

Team[leagueID] ⊆ League[leagueID]

TeamAttributes[teamID] ⊆ Team[teamID]

$\sigma_{date < 2008 \vee date > 2016} Match = \emptyset$

$\sigma_{date < 2008 \vee date > 2016} PlayerAttribute = \emptyset$

$\sigma_{rating < 0 \vee rating > 100} PlayerAttribute = \emptyset$

$\sigma_{potential < 0 \vee potential > 100} PlayerAttribute = \emptyset$

$\sigma_{defensiveWorkRate < 0 \vee defensiveWorkRate > 100} PlayerAttribute = \emptyset$

$\sigma_{attackingWorkRate < 0 \vee attackingWorkRate > 100} PlayerAttribute = \emptyset$

$\sigma_{buildUpPlaySpeed < 0 \vee buildUpPlaySpeed > 100} TeamAttribute = \emptyset$

$\sigma_{buildUpPlayDribbling < 0 \vee buildUpPlayDribbling > 100} TeamAttribute = \emptyset$

$\sigma_{buildUpPlayPassing < 0 \vee buildUpPlayPassing > 100} TeamAttribute = \emptyset$

$\sigma_{chanceCreationPassing < 0 \vee chanceCreationPassing > 100} TeamAttribute = \emptyset$

$\sigma_{chanceCreationCrossing < 0 \vee chanceCreationCrossing > 100} TeamAttribute = \emptyset$

$\sigma_{chanceCreationShooting < 0 \vee chanceCreationShooting > 100} TeamAttribute = \emptyset$

$\sigma_{defencePressure < 0 \vee defencePressure > 100} TeamAttribute = \emptyset$

$\sigma_{defenceAggression < 0 \vee defenceAggression > 100} TeamAttribute = \emptyset$

$\sigma_{defenceTeamWidth < 0 \vee defenceTeamWidth > 100} TeamAttribute = \emptyset$

Data Dictionaries

Country

Attribute	Description	Type	Required	Default
countryID	The ID of the country	INT	Yes	
countryName	The name of a country	TEXT	Yes	

League

Attribute	Description	Type	Required	Default
leagueID	The ID of the league	INT	Yes	
countryID	The ID of the country the league is in	INT	Yes	
leagueName	The name of the league	TEXT	Yes	

Match

Attribute	Description	Type	Required	Default
matchID	The ID of the match	INT	Yes	
countryID	The ID of the country where match took place	INT	Yes	
leagueID	The ID of the league in which match took place	INT	Yes	
season	The season of the league the match took place during	TEXT	Yes	
date	The date of the match	TIMESTAMP	Yes	

homeTeamID	The home team’s teamID	INT	Yes	
awayTeamID	The away team’s teamID	INT	Yes	
homeTeamGoal	Number of goals home team scored	INT	Yes	
awayTeamGoal	Number of goals away team scored	INT	Yes	
shotOn	Number of shots on target	INT	No	
shotOff	Number of shots off target	INT	No	
foulCommit	Number of fouls	INT	No	
card	Number of cards shown	INT	No	
cross	Number of crosses	INT	No	
corner	Number of corners	INT	No	
possession	Amount of possession for each team	INT	No	

Player

Attribute	**Description**	**Type**	**Required**	**Default**
playerID	The ID of the player	INT	Yes	
playerName	The name of the player	TEXT	Yes	
birthday	The player’s date of birth	TIMESTAMP	Yes	
height	The player’s height in cm	INT	Yes	
weight	The player’s weight in lbs	INT	Yes	

PlayerAttributes

Attribute	**Description**	**Type**	**Required**	**Default**
playerID	The ID of the player	INT	Yes	
date	The date the analysis was conducted	TIMESTAMP	Yes	
rating	The player’s overall rating	INT	Yes	
potential	The player’s potential rating	INT	Yes	
attackingWorkRate	The player’s work rate percentage on attack	INT	Yes	
defensiveWorkRate	The player’s work rate percentage on defence	INT	Yes	

Team

Attribute	**Description**	**Type**	**Required**	**Default**
teamID	The ID of the team	INT	Yes	
teamName	The team’s name	TEXT	Yes	
leagueID	The ID of the league the team plays in	INT	Yes	
teamAbbreviation	The team’s abbreviated name	TEXT	Yes	

TeamAttributes

Attribute	Description	Type	Required	Default
teamID	The ID of the team	INT	Yes	
date	The date the analysis was done	TIMESTAMP	Yes	
buildUpPlaySpeed	The speed with which they play on attack	INT	Yes	
buildUpPlayDribbling	The percentage of attacks that are built by individuals dribbling the ball	INT	No	
buildUpPlayPassing	The percentage of attacks that are built by team passing	INT	Yes	
chanceCreationPassing	The percentage of chances created by passing	INT	Yes	
chanceCreationCrossing	The percentage of chances created by crossing	INT	Yes	
chanceCreationShooting	The percentage of chances created by shooting	INT	Yes	
defencePressure	How closely the team defends the opponent	INT	Yes	
defenceAggression	How aggressively the team defends the opponent	INT	Yes	
defenceTeamWidth	How spread out or wide the team defends	INT	Yes	

Justification of Design

The team has directly translated the structure of the dataset having only changed the column names (no structural changes). This action was thoroughly contemplated and was agreed upon due to the dataset already being very well structured. However, if there are issues with extracting XML data effectively and efficiently, it may be necessary to modify our relations to accommodate for the differences.

Currently, each of the 7 tables have a primary key, and contain a column that allows for the building of a relationship with another table (i.e. Country[countryID] and Match[countryID]). Thus, each table is joinable with at least one other table, ultimately allowing the team to assemble and relate the data to answer the questions (found in the 'Questions' section of this paper) at hand.

In addition, columns in the dataset have a clear type. Whether numerical, text, or date, the data does not contain overlaps of types. This is essential in understanding our data, and will allow us to move forward with our analysis without adding another step in our data cleaning process.