Alex Meci

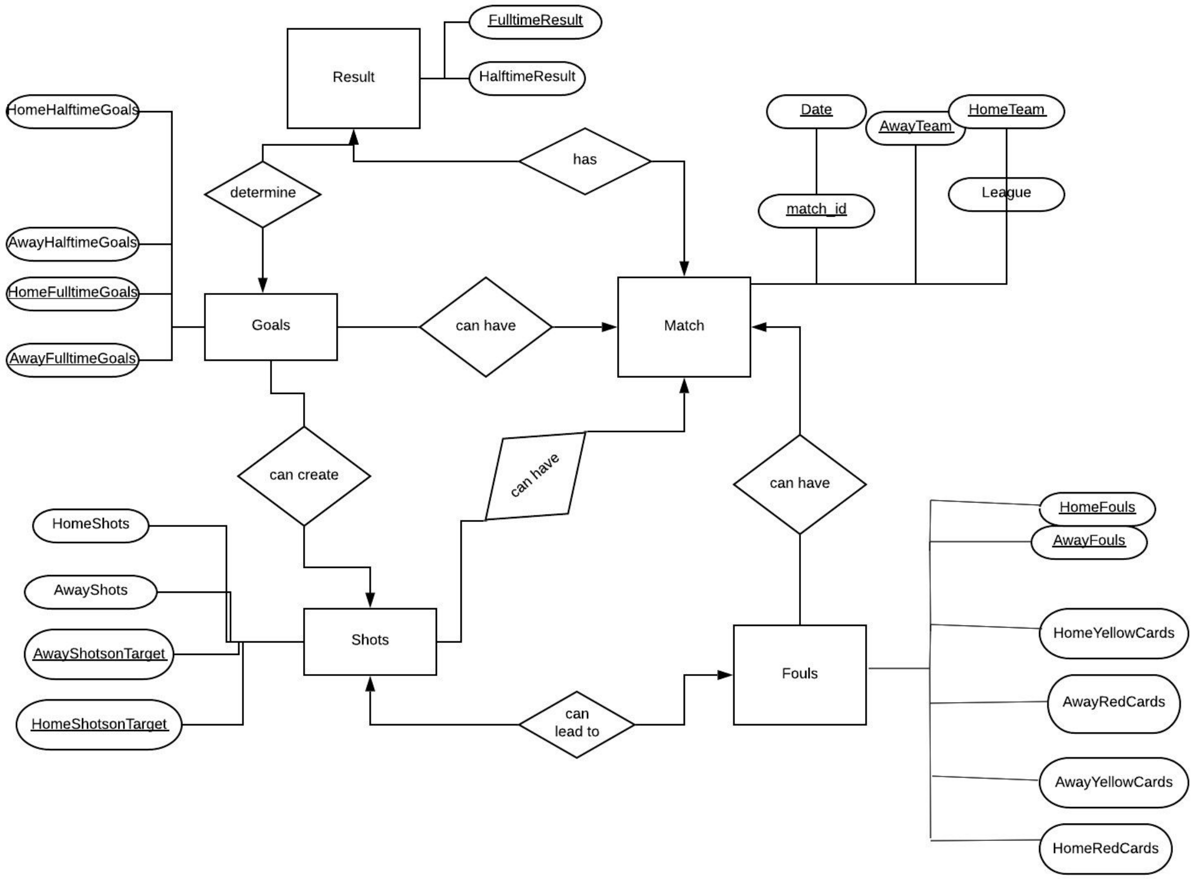
12/10/19

Professor Gregory Schaper

Databases Final Project

When watching soccer games, I am usually only concerned about the final score. However, there are many aspects of the game along with stats that can give more details as to who was truly the better team, and, according to statistics, who should have won. For my project, my purpose was to interpret the data from the 2018-2019 soccer season in the highest Spanish division(La Liga). By developing schemas and queries for this dataset, I think I can come to some reasonable conclusions proved by statistics as to why some teams when it is not expected and why hometeam advantage exists.There are a lot of columns in the tables from which I can extract an interpret different things, such as fouls per half, goals per half, etc. If you need assistance in knowing the labels for some of my tables, here is the key: https://www.football-data.co.uk/notes.txt

Diagram:



Lucidcharts did not allow me to add anymore attributes unless I bought the premium version, but match_id is supposed to be present in each table.

## Schemas:

Match(match_id, Date, AwayTeam, HomeTeam, League)

Fouls(match_id, HomeFouls, AwayFouls, HomeYellowCards, AwayYellowCards, HomeRedCards, AwayRedCards, Date, AwayTeam, HomeTeam)

Match(match_id, Date, AwayTeam, HomeTeam, League)

Shots(HomeShots, AwayShots, match_id, HomeShotsonTarget, AwayShotsonTarget, Date, AwayTeam, HomeTeam)

Match(match_id, Date, AwayTeam, HomeTeam, League, FulltimeResult)

Result(HalftimeResult, FulltimeResult, match_id)

Match(match_id, Date, AwayTeam, HomeTeam, League)

Goals(HomeHalftimeGoals, AwayHalftimeGoals, match_id, HomeFulltimeGoals, AwayFulltimeGoals, Date, AwayTeam, HomeTeam)

Fouls(match_id, HomeFouls, AwayFouls, HomeYellowCards, AwayYellowCards, HomeRedCards, AwayRedCards)

Shots(match_id,HomeShots, AwayShots, HomeShotsonTarget, AwayShotsonTarget, HomeFouls, AwayFouls)

Goals(match_id,HomeHalftimeGoals, AwayHalftimeGoals, HomeFulltimeGoals, AwayFulltimeGoals)

Shots(match_id,HomeShots, AwayShots, HomeShotsonTarget, AwayShotsonTarget, HomeFulltimeGoals, AwayFulltimeGoals)

Result(HalftimeResult, FulltimeResult, match_id)

Goals(HomeHalftimeGoals, AwayHalftimeGoals, HomeFulltimeGoals, AwayFulltimeGoals, FulltimeResult, match_id)

Based on the data, I proposed the following questions:

- Is there a relationship between fouls per game and games won?
  - This depends on both the team and the game, so I did not really find a correlation between who would win the game based off of how many fouls they committed.
- Is there a relationship between yellow cards/red cards and games won/lost
  - If there was a hometeam red card, the away team was more likely to win the game, and vice versa. Some games went either way or to a draw. Yellow cards were not relevant in finding a conclusion.
- Is the amount of shots on target correlated with goals scored?
  - Through running queries where the max number of shots on target was shown first, the team with the most shots on goal almost always won. As the list was more towards the middle where the shots on goal was even, the results of the matches were more inconsistent.
- If a team is leading at half time, are they more likely to win the game?

    - Whoever was leading at halftime was more likely to win at full time, regardless of who was the home team. If half time was a draw the home team was still more likely to win but the results went both ways.
- If it is late in the season, is the home team more likely to win?
    - In the final two months of the season(April to May) I actually found that the home team won most games, as compared to before then where the results were one or the other.
- Does giving more fouls in a game lead to more shots from the other team?
    - When the home team gave more fouls up, the away team had slightly more shots. However, when the away team committed more fouls, the home team had a much bigger difference between shots. I believe this is part of home team advantage.

In terms of normalization, I split my dataset into 5 tables from one, because there were too many dependencies that were unrelated to the primary keys. Instead, for each table, there is the primary key match_id in each table and each table is then in third normal form.

In terms of what grade I believe I deserve, I think that I have put more time into this than any other project so far this year for this class, and I think my division of the dataset along with my queries show that I am showing or at least close to showing what you are looking for. So I think that my grade should be around a B+ to an A.