Lendify

Introduction

As part of the technical interview for potential new Data Engineers at Lendify, we ask applicants to solve a few technical questions beforehand. For each question we ask you to provide a) a short summary of your solution and b) the code that you wrote to solve the problem. The code does not have to be beautiful, but readable. Please note that you will have the opportunity to explain your thinking, etc., at the face-to-face interview.

Send your solutions to Jay Chinnaswamy (jtc@lendify.se) at least one day prior to the actual interview.

Good luck!

Data

Provided is a dataset of all Bundesliga (i.e. the top German football division) matches from the 1993/1994 season to the 2017/2018 season.
The dataset contains 10 attributes:

- Date : the date the match was played
- HomeTeam : the home team
- AwayTeam : the away team
- FTHG : the number of goals the home team scored during the whole match
- FTAG : the number of goals the away team scored during the whole match
- FTR : the full time result, (H)ome, (A)way, or (D)raw
- HTHG : the number of goals the home team scored during the first half
- HTAG : the number of goals the away team scored during the first half
- HTR : the half time result, (H)ome, (A)way, or (D)raw

HTHG, HTAG, and HTR values are missing for the 1993/1994 and 1994/1995 seasons.

Problems

P1: Do some exploratory analysis and provide your most (i.e. 2 - 3) interesting findings.

P2: We wish to know which are the best Bundesliga teams over the entire period. Define a metric, that you think is fair, to rank the teams over time. Compute the top 10 teams over the entire period.

P3: Assuming that you are going build a warehouse come up with ideas what are the dimensions, facts you will build, what could be the structure of those and some challenges or questions you may have in order to build scalable DW

P4. Come up with the ideas how do you take this further (you might have to handle granular data, various formats of data, real time processing etc) to make it more robust and serving all the needs that arise over time for data consumers, Eg. Performance, building real time data pipelines, Handling CI/CD , tools etc