1. We would like to create a model that can choose the best players for the particular needs of a football club. We would achieve this by looking at the current players that a club has and determining where they are lacking quality. The model would then suggest players that can solve that lack of quality. We have several datasets that could be useful to us:

   Football Data from Transfermarkt (kaggle.com);

   https://www.kaggle.com/datasets/rishikeshkanabar/premier-league-player-statistics-updated-daily;

   https://www.kaggle.com/datasets/hugomathien/soccer;

   https://www.kaggle.com/code/desalegngeb/english-premier-league-players-statistics;

   We chose these datasets because they have the relevant statistics that can help us achieve the completion of our model, such as player performance statistics, club statistics, player prices on the market, among many other such data, and some of them have nice preprocessing.

2. The datasets contain extensive statistics about each individual player with different criteria depending on their positions.
   The data will be filtered during the processing thanks to the filtered search provided in the input by the club that is in search of an ideal player / player profile.

   (a) As for preprocessing, the databases already takes care of some of it:

   - Separating objects and columns that might not be divided by appearances.
   - Making sure players are labeled with their respective positions.
   - Adding an extra label to players who have made at least 38 appearances, which is the equivalent of a whole Premier League season.

   We can use Python scripts to merge the datasets and organize the data to make comprehensive player profiles and store them in vector form, to be fed to the model. The player profile vectors can have elements such as (position, goals per 90 minutes, assists per 90 minutes, ratio of passes completed, club strength

relative to domestic league.,...) and many other statistics that we may add in the future that would improve the model.

(b) Depending on a particular club's needs (position, role) and their current budget and team stats, the trained model will compare all concerned players' stats with what the club needs (averages of certain stats) and can afford. With the amount of knowledge we currently possess, we might consider a linear regression model with multiple intricate statistics being involved. Our goal is to provide precise results, but it could take an overwhelmingly long amount of time to compute because of the sheer amount of data that is available to process. We will have to learn to manage this and solve this issue.

Also, a decision tree could be better at assessing a unified decision and making the process gradually simpler thanks to an ensuing stats-driven process of elimination. Another algorithm that could be useful is a KNN algorithm especially when sorting by position.

(c) Since we are mainly considering linear regression, we would use Mean squared error as our evaluation metric. As for the accepted results, those depend on individual teams and their requirements or goals, as well as the ability of the concerned players to fit those requirements, note that the model is purely statistical and does not account for off-pitch situations or criteria such as discipline, mental state of players or player personal affair which could drastically influence their performance on the pitch but cannot be measured.

3. We would make a simple webapp that takes in a team or club as an input via text and other optional inputs such as player position, budget, age range of players, etc. also via text. These are subject to change as we develop the model. The output would be the suggested players that are a good fit for the team's requirements along with their pictures, names and price on the current market.

   To note: the less parameters the input includes, naturally the more suggestions the model will provide, as there are more possible players that fit the input requirements.