(1) Problem

We have limited ability to accurately forecast the outcomes of soccer matches. Paradoxically, the unpredictability of soccer matches makes the business of betting possible. Gamblers try their luck or skill at predicting outcomes, and betting houses set odds that are profitable.

(2) Client

Our client is any betting house or casino that allows sports betting. Betting houses profit only because they set odds that entice clientele but are still favorable to the house. If we can more accurately predict the outcomes of soccer matches, the client will use this information to create odds that are more profitable and less risky for the business.

(3) The data

We will start with a European Soccer Database that was compiled by Hugo Mathien, whose project can be found on Github at <https://github.com/hugomathien/football-data-collection/tree/master/footballData>, and whose compiled database is posted on Kaggle at <https://www.kaggle.com/hugomathien/soccer>. The data was pulled from several sources using crawlers to scrape the data into a .sqlite database. The original sources include:

<http://football-data.mx-api.enetscores.com/>

<http://www.football-data.co.uk/>

<http://sofifa.com/>

We will use this database as a foundation for analysis and starting point for further learning. In addition, we will reverse engineer some of the crawling tools Mathien used, as well as devise some new ones, to:

1. update the data from the original sources and
2. to scrape data from alternative sources.

(4) Approach

We will begin by converting the data into .csv files. This will allow us to apply base R, dplyr, ggplot, and other tools to format, visualize, and pull descriptive statistics. This will provide a high-level view of how wins and losses are distributed and whether there might be any trends.

Once we have a good picture of the data, we will run higher-level analyses on it using the models learned later in this course with the hope of garnering a predictive ability greater than 50%.

(5) Deliverables

Deliverables will be in a Github repository. It will include the original data and the formatted data. It will also include a slide deck to present to the betting house and a report describing deeper analyses and our methodology. We will also include the R code used for the previously mentioned deliverables.

Note: I understand that this data is pulled from Kaggle and is already in good shape. Fortunately, the data is not perfect and contains various null values. In addition to replicating some of what the author had to do to do obtain the data, I plan to make the project more realistic by pulling data from new sources using my own crawlers.