Extract:

Data were downloaded from footystat.com and fbref.com (urls: <https://footystats.org/download-stats-csv#>, <https://fbref.com/en/comps/20/2109/stats/2018-2019-Bundesliga-Stats>, <https://fbref.com/en/comps/12/1886/stats/2018-2019-La-Liga-Stats>, <https://fbref.com/en/comps/20/2109/stats/2018-2019-Bundesliga-Stats>), and loaded into pandas dataframes. All datasets originated as csv files.

Transform:

Repetitive data were removed with the ‘clean\_df’ function. The ‘leagues’ dataframe was created to summarize data on each league. The ‘english,’ ‘german,’ and ‘spanish’ dataframes contain performance statistics for players of each league. Many columns representing more specialized statistics were removed in order to focus the final presentation.

Load:

Due to the overlap between many of the columns in each dataset, the data were loaded into a relational database (SQL). The lack of a need for scalability also supports this decision. At this stage, the pandas dataframes were simply combined using the ‘to\_sql’ function.