Thomas Hucorne Théo Rochette Clément Lévêque Quentin Ching Ewad Jamin

Project Ando

I. Data Choice

Our team chose to work with **football data** due to widespread interest in the sport. We sourced 10 interconnected datasets from Kaggle, linked by keys such as `id_player`, `club_id`, and `game_id`, and containing **integer**, **date**, and **string** data (e.g., player names, club names).

For our analysis, we have several target variable options: we could predict **game** outcomes or assess a **player's future success**. In the latter case, using `market_value_in_eur` as the target variable would reflect a player's demand in the market.

II. Description Datas

As depicted before there are 10 files, we will try to describe the data that we can find in each file and how to link one file to another.

The columns represent the attributes of each file and the rows represent the number of observations.

Firstly we have the **appearances** file which contains **13 columns** and **1 618 667 rows**.
This structure allows for a comprehensive analysis of **player performance** across games, including **discipline** (yellow and red cards), **scoring contributions** (goals and assists), and **playing time**.

Then **club_games** files which contain **11 columns** and **141 544 rows**.
This dataset is focused on **match outcomes**, **team performance**, and **managerial information** for each game.

Clubs files which contain **17 columns** and **439 rows.**
This dataset is primarily focused on the **financial, demographic,** and **operational characteristics** of the clubs.

Competitions file which contains **11 columns** and **44 rows**.
This dataset would be useful for exploring the structure of football competitions globally, identifying **major leagues**, and understanding which governing bodies are associated with each **competition**. It could also help analyze the **hierarchy** between **domestic** and **international leagues**.

Game events file which contains **10 columns** and **982 888 rows.**
This data is valuable for analyzing i**n-game events**, tracking **player contributions** like goals and assists, and understanding the flow and **impact of substitutions** and other **key actions** during matches.

Game lineups file which contains **10 columns** and **2 191 911 rows.**
This dataset provides insights into **team formations**, **player roles,** and **leadership** on the field, which is valuable for analyzing game strategies, player consistency in roles, and captaincy patterns.

Games file which contains **23 columns** and **70 772 rows.**
This dataset provides detailed information on **individual football matches,** including **scores**, **teams** and their **formations**, **attendance**, **stadium**, and specific details like **referee** and **managers**. It enables precise **match-by-match analysis** to identify trends or compare team performances.

Player valuations file which contains **5 columns** and **483404 rows.**
This data will be useful for tracking how **players' market values** change over time, possibly reflecting their performance, age, and transfers between clubs and leagues.

Players file which contains **23 columns** and **32 273 rows.**
This dataset would be valuable for understanding **player demographics, career progression**, **physical attributes**, and **market valuations**, allowing for comparisons across various clubs and leagues.

Transfers file which contains **10 columns** and **73 760 rows.**
This dataset will be useful for **tracking player movement** across clubs, **analyzing trends** in transfer fees, and **comparing market values** during different seasons or for specific players.

III. Data preprocessing

To ensure the data is both manageable and insightful, we will conduct careful preprocessing steps. First, we'll format each dataset to **maintain consistency** in data types and units, ensuring that dates, numerical values, and categorical variables are **standardized**. Then, we'll merge the datasets by keys such as `id_player`, `club_id`, and `game_id` to create a comprehensive, unified dataset. We'll select only the most **relevant features** to avoid redundancy and focus on **impactful data**, emphasizing variables that contribute to **predicting player market value** and game outcomes. Finally, to address the high dimensionality of the dataset, we'll employ PCA to reduce complexity, preserving only those features that contribute significantly to the overall variance and predictive power of our model.