# Project 2 ETL Data Cleansing and Transforming and Loading Process

## Team:

* Smit Gandhi
* Cheikh Sidi El Moctar

## Phase 0 : defining the project proposal, and data sources.

### Subject:

* European soccer dataset.

### Dataset Source:

* The source of the dataset used in this project is collected from **Kaggle open dataset**, we would love to explore more resources for more datasets, but timeframe doesn’t allow.

### Process

* The dataset is collected as SQLite database from Kaggle, this database contains 7 tables as following:

- Country

- League

- Match

- Player

- Player\_Stats

- Team

- sqlite\_sequence

## Phase 1 : Data Cleaning

In this phase we used pandas to read the data from the database and perform the cleaning in the pandas dataframe.

The cleaning phase consisted in:

* + Formatting the date field to date type
  + Dropping un-needed columns
  + Checking the datatype for each field, to make its in the appropriate datatype

## Phase 2 : Data Transformation

In this phase we carried out some transformation to get the data ready for the analysis, as following:

* + Merging the datasets (Player, Player\_stats)
  + Creating new column “Player\_Age” (date – birthday)
  + Grouping by on the Player\_Age to see the performance vs age.
  + Creating new match table adding names of the teams, countries and leagues
  + Creating two new columns “score\_difference ” and “winner”
  + Grouping by on the winner, to see the most winning team in europe

## Phase 3 : Loading the data into Postgres DB

In this phase we carried out some transformation to get the data ready for the analysis, as following:

* + Creating a database Fifa\_db in postgresql
  + Creating the schema for four tables, players, ages vs performance, match, spain\_league.
  + Using Sqlalchemy and pandas to load the data into postgresql.

## Summary

Given the dataset and the purpose of exploring this dataset, first the data tables were viewed and merged to create a pandas dataframe as required for data analysis. For the ease of analysis a new column ‘age’ was created using the ‘birthday’ column, this allowed for more intuitive comparison such as performance vs. age. Similar process like this would allow us to visualize contribution of players given the skillset, to the overall team performance over the past decade.