Analysis by Simona Valente

GOAL

Analyze the "**European Soccer Database**" to extrapolate useful data and information.

Tools used

- **Data Analysis**: SQL on Google Bigquery
- **Supporting graphs**: Python on Google Colaboratory
- **Presentation**: Google Slides

What you will find in the repository

- **Dataset used**: match.csv, leagues.csv, team.csv, player.csv, match_per_month.csv;
- Relational **schema**;
- A **document** with the ordered **list of queries used**;
- This **presentation**;
- A **python notebook.**

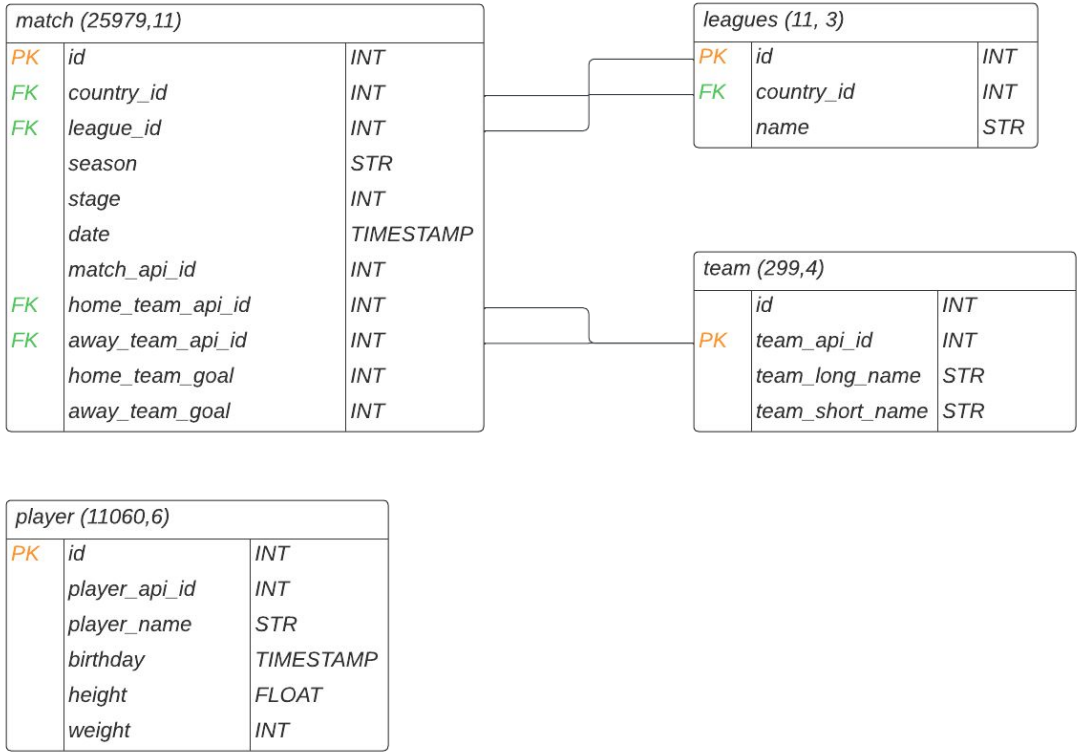
The Database

The analysis was carried out on the **European Soccer database**.

It is composed of 4 datasets:

- match;
- leagues;
- team;
- player.

The relational scheme →



Preliminary phase: let's familiarize ourselves with the database. How many matches and leagues in total are we analyzing?

TOTAL

Row	Tot_match
1	25979

```
SELECT COUNT(id) Tot_match FROM `European_Soccer_Database.match`
```

TOTAL LEAGUES

Row	name
1	Belgium Jupiler League
2	England Premier League
3	France Ligue 1
4	Germany 1. Bundesliga
5	Italy Serie A
6	Netherlands Eredivisie
7	Poland Ekstraklasa
8	Portugal Liga ZON Sagres
9	Scotland Premier League
10	Spain LIGA BBVA
11	Switzerland Super League

```
SELECT DISTINCT(name) FROM `European_Soccer_Database.leagues`
```

What is the reference time range?

I go to print:

- the **difference** between the *most recent date* and the *most distant one*;
- the two dates at the extremes.

```
SELECT DATE_DIFF(max(date), min(date), day) AS Total_Range,
  FORMAT_TIMESTAMP('%m-%d-%Y', TIMESTAMP(max(date))) AS Most_recent,
  FORMAT_TIMESTAMP('%m-%d-%Y', TIMESTAMP(min(date))) AS Less_recent
FROM `European_Soccer_Database.match`;
```

Row	Total_Range	Most_recent	Less_recent
1	2868	05-25-2016	07-18-2008

It results that our time range is **2868 days**; while the most recent date is **May 25, 2016** and the oldest is **July 18, 2008**.

How many seasons are there in total and how many matches for each season?

SEASON: we have all the data for the seasons ranging from **2008** to **2016**.

Row	Season
1	2008/2009
2	2009/2010
3	2010/2011
4	2011/2012
5	2012/2013
6	2013/2014
7	2014/2015
8	2015/2016

```
SELECT DISTINCT season Season
FROM  `European_Soccer_Database.match`
```

```
SELECT COUNT(id) TotMatch, season Season
FROM `European_Soccer_Database.match`
GROUP BY Season
ORDER BY Season
```

Row	Tot_match	season
1	3326	2008/2009
2	3230	2009/2010
3	3260	2010/2011
4	3220	2011/2012
5	3260	2012/2013
6	3032	2013/2014
7	3325	2014/2015
8	3326	2015/2016

MATCHES FOR SEASON: there are no significant differences, except for the **2013/2014**.

How many matches are there divided by season and league? Do we notice anything out of the ordinary?

```
SELECT DISTINCT (m.season) Season, l.name LeagueName, COUNT(match_api_id) TotMatch
FROM `European_Soccer_Database.match` m
LEFT JOIN `European_Soccer_Database.leagues` l
ON m.league_id = l.id
GROUP BY m.season, l.name
ORDER BY TotMatch DESC
```

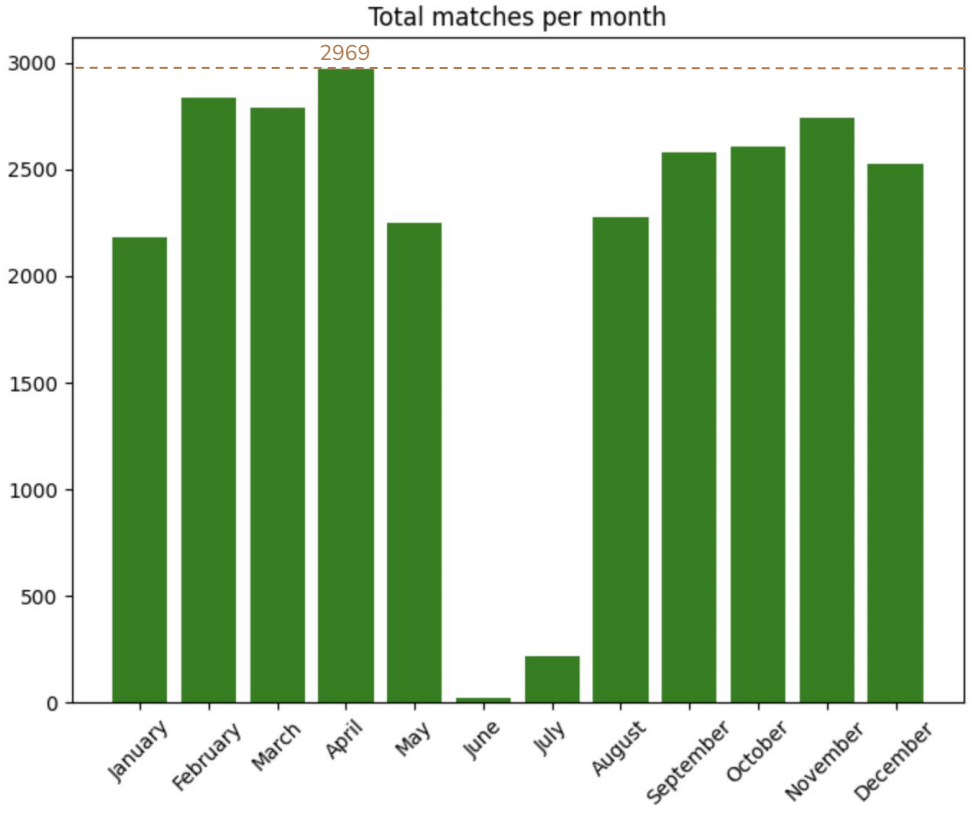
Row	Season	LeagueName	TotMatch
1	2008/2009	England Premier League	380
2	2009/2010	England Premier League	380
3	2010/2011	England Premier League	380
4	2011/2012	England Premier League	380

88	2013/2014	Belgium Jupiler League	12

Belgium Jupiler League only has 12 matches in 2013/2014

How many matches were there for each month of the year?

Row	Tot_match	Month
1	2969	4
2	2834	2
3	2785	3
4	2739	11
5	2608	10
6	2575	9
7	2524	12
8	2276	8
9	2245	5
10	2183	1
11	218	7
12	23	6



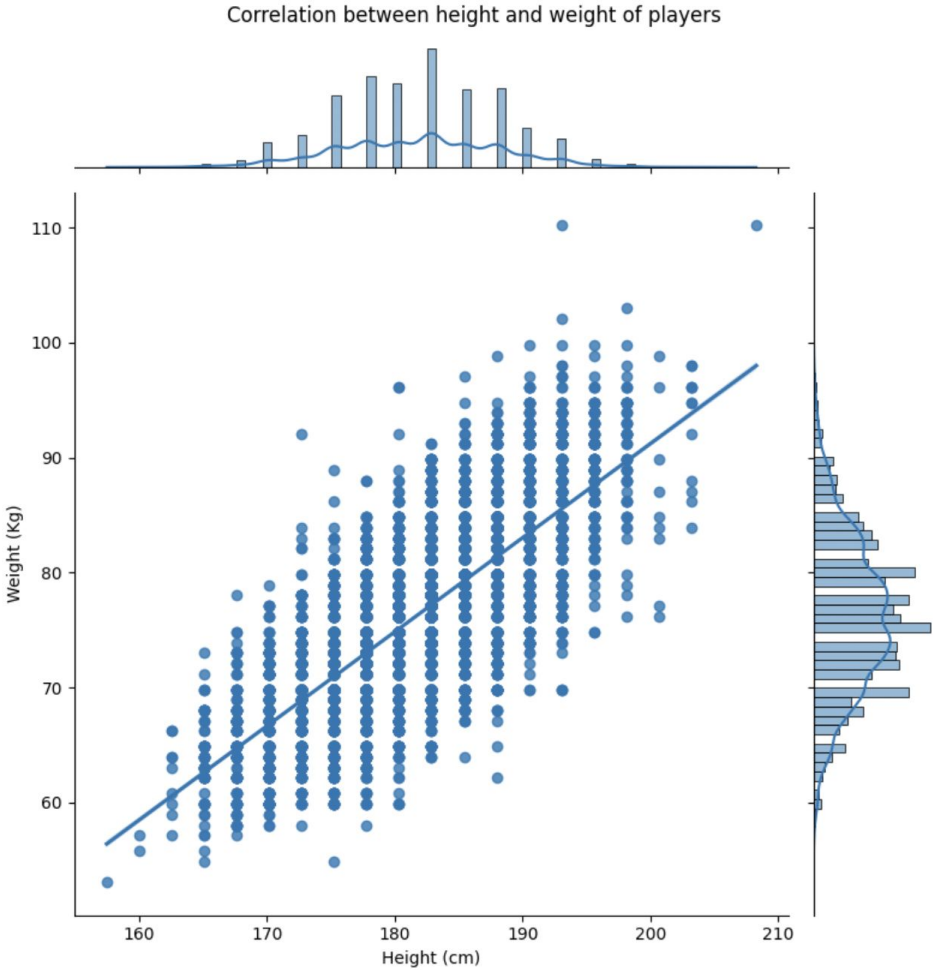
Season and League: statistics on goals scored at home

We produce a table that shows for each ***Season*** and ***League Name*** the following statistics about the home goals scored: **min**, **average**, **mid-range**, **max** and **sum**.

```
SELECT m.season Season, l.name LeagueName,
MIN(m.home_team_goal) minHTG,
ROUND(AVG(m.home_team_goal),2) avgHTG,
CAST(((MIN(m.home_team_goal) + MAX(m.home_team_goal))/2) AS INT64) midrangeHTG,
MAX(m.home_team_goal) maxHTG,
SUM(m.home_team_goal) sumHTG
FROM `European_Soccer_Database.match` m
LEFT JOIN `European_Soccer_Database.leagues` l
ON m.league_id = l.id
GROUP BY m.season, l.name
ORDER BY sumHTG desc
```

The season with the most home goals was 2009/2010 in England Premiere League

Row	Season	LeagueName	minHTG	avgHTG	midrangeHTG	maxHTG	sumHTG
1	2009/2010	England Premier League	0	1.7	5	9	645
2	2012/2013	Spain LIGA BBVA	0	1.69	3	6	641
3	2011/2012	Spain LIGA BBVA	0	1.68	4	8	638



Let's now analyze the players.

The graph shows the relationship between the height of our athletes and their weight. As we can see, the two variables are directly proportional, with a correlation of 0.77%

Let's create a new '**PlayerBMI**' table in which we insert:

- the **weight** in kg (kg_weight);
- the **height** in meters (m_height);
- the player's body mass index (**BMI**).

We filter the table to only show players with an optimal BMI (18.5 to 24.9).

```
CREATE TABLE `European_Soccer_Database.PlayerBMI` AS SELECT *,
ROUND ((weight / 2.205),2) AS kg_weight,
ROUND ((height / 100),2) AS m_heigth,
ROUND ((weight / 2.205) / power(height / 100,2),2) AS BMI,
FROM `European_Soccer_Database.player`
WHERE (weight/2.205)/power(height/100, 2) between 18.5 and 24.9
```

Row	id	player_api_id	player_name	birthday	height	weight	kg_weight	m_heigth	BMI
1	2868	451335	Edmilson Fernandes	1996-04-15 00:00:00 UTC	190.5	154	69.84	1.91	19.25
2	2901	144989	Efe Ambrose	1988-10-18 00:00:00 UTC	190.5	154	69.84	1.91	19.25

```
SELECT
 (SELECT count(id)
  FROM `European_Soccer_Database.player`) -
 (SELECT count(id)
  FROM `European_Soccer_Database.PlayerBMI`) as PlayerNoBMI
```

Row	PlayerNoBMI
1	863

7,8% discarded

By subtracting the filtered athletes from the total ones, we find that **863 players** do not have an ideal BMI.

Which Team has scored the highest total number of goals during the most recent available season?

```
SELECT h.team_long_name, h.SumOfGoalHome, a.SumOfGoalAway,
h.SumOfGoalHome + a.SumOfGoalAway AS TotalGoal
FROM ( SELECT t.team_long_name, SUM(m.home_team_goal) AS SumOfGoalHome
FROM `European_Soccer_Database.match` m
INNER JOIN `European_Soccer_Database.team` t
ON m.home_team_api_id = t.team_api_id
WHERE m.season = (SELECT MAX(season) FROM `European_Soccer_Database.match`)
GROUP BY t.team_long_name ORDER BY SumOfGoalHome) h
INNER JOIN
(SELECT t.team_long_name, SUM(m.away_team_goal) AS SumOfGoalAway
FROM `European_Soccer_Database.match` m
INNER JOIN  `European_Soccer_Database.team` t
ON m.away_team_api_id = t.team_api_id
WHERE m.season = (SELECT MAX(season) FROM `European_Soccer_Database.match`)
GROUP BY t.team_long_name ORDER BY SumOfGoalAway) a
ON h.team_long_name = a.team_long_name
ORDER BY TotalGoal DESC
LIMIT 1
```

Row	team_long_name	SumOfGoalHome	SumOfGoalAway	TotalGoal
1	FC Barcelona	67	45	112

For each season, which team ranks first in terms of total goals scored?

```
SELECT * FROM
(SELECT h.season, h.team_long_name, h.SumOfGoalHome, a.SumOfGoalAway,
h.SumOfGoalHome + a.SumOfGoalAway AS TotalGoal,
RANK() OVER (PARTITION BY a.season ORDER BY h.SumOfGoalHome +
a.SumOfGoalAway DESC) AS rank_season
FROM (SELECT m.season, t.team_long_name, SUM(m.home_team_goal) AS SumOfGoalHome
FROM `European_Soccer_Database.match` m INNER JOIN
`European_Soccer_Database.team` t ON m.home_team_api_id = t.team_api_id
GROUP BY m.season, t.team_long_name ORDER BY SumOfGoalHome) h
INNER JOIN
(SELECT m.season,t.team_long_name, sum(m.away_team_goal) AS SumOfGoalAway
FROM `European_Soccer_Database.match` m INNER JOIN
`European_Soccer_Database.team` t ON m.away_team_api_id = t.team_api_id
GROUP BY m.season, t.team_long_name ORDER BY SumOfGoalAway) a
ON h.team_long_name = a.team_long_name AND h.season=a.season)
WHERE rank_season = 1
ORDER BY season DESC
```

Row	season	team_long_name	SumOfGoalHome	SumOfGoalAway	TotalGoal	rank_season
1	2015/2016	FC Barcelona	67	45	112	1
2	2014/2015	Real Madrid CF	65	53	118	1
3	2013/2014	Real Madrid CF	63	41	104	1
4	2012/2013	FC Barcelona	63	52	115	1
5	2011/2012	Real Madrid CF	70	51	121	1
6	2010/2011	Real Madrid CF	61	41	102	1
7	2009/2010	Ajax	64	42	106	1
8	2008/2009	FC Barcelona	61	44	105	1

Real Madrid CF is the team that has ranked first for the most seasons.

Create a new table 'TopScorer' containing the top 10 teams in terms of total goals scored

```
CREATE TABLE `European_Soccer_Database.TopScorer` AS
(SELECT h.team_api_id ,h.team_long_name, h.SumOfGoalHome, a.SumOfGoalAway,
h.SumOfGoalHome + a.SumOfGoalAway AS TotalGoal FROM
(SELECT t.team_api_id ,t.team_long_name, SUM(m.home_team_goal) AS SumOfGoalHome
FROM `European_Soccer_Database.match` m INNER JOIN
`European_Soccer_Database.team` t ON m.home_team_api_id = t.team_api_id
where m.season = (select MAX(season) FROM `European_Soccer_Database.match`)
GROUP BY t.team_api_id, t.team_long_name ORDER BY SumOfGoalHome) h INNER JOIN
(SELECT t.team_long_name, SUM(m.away_team_goal) AS SumOfGoalAway
FROM `European_Soccer_Database.match` m INNER JOIN
`European_Soccer_Database.team` t ON m.away_team_api_id = t.team_api_id
WHERE m.season = "2015/2016"
GROUP BY t.team_long_name ORDER BY SumOfGoalAway) a ON h.team_long_name = a.team_long_name
ORDER BY TotalGoal DESC
LIMIT 10)
```

Row	team_api_id	team_long_name	SumOfGoalHome	SumOfGoalAway	TotalGoal
1	8634	FC Barcelona	67	45	112
2	8633	Real Madrid CF	70	40	110
3	9847	Paris Saint-Germain	59	43	102
4	9925	Celtic	55	38	93
5	9931	FC Basel	44	44	88
6	9772	SL Benfica	52	36	88
7	8640	PSV	41	47	88
8	8686	Roma	44	39	83
9	9789	Borussia Dortmund	49	33	82
10	8593	Ajax	49	32	81

Thanks for the attention!

You can find all the queries used within the repository