European Soccer Performance Analysis

Dataset Resource: https://www.kaggle.com/datasets/hugomathien/soccer

Project Title: European Soccer Performance Analysis

Project Statement:

Analyze the European Soccer Database to identify key performance indicators and trends in Home-Away Win performance. This project will demonstrate advanced SQL skills, including data extraction, complex querying, and visualization.

Data Description

- +25,000 matches
- +10,000 players
- 11 European Countries with their lead championship
- Seasons 2008 to 2016
- Players and Teams' attributes* sourced from EA Sports' FIFA video game series, including the weekly updates
- Team line up with squad formation (X, Y coordinates)
- Betting odds from up to 10 providers
- Detailed match events (goal types, possession, corner, cross, fouls, cards etc...) for +10,000 matches

Business Questions:

1. **Team Performance Analysis:**
    - Which teams have the highest win rates?
    - How do team performances vary across different leagues and seasons?

2. **Player Performance Analysis:**
    - Who are the top 5 players by rating?
    - How do player performances evolve over their careers?
3. **Match Outcome Analysis:**
    - How does home-field advantage affect match results?

Tools and Skills used

SQL, Tableau, DataGrip, Querying, Data Visualization

Project Steps:

1. **Data Extraction and Preparation:**
    - Connect to the SQLite database containing the soccer data.
2. **Data Cleaning and Transformation:**
    - Clean the data to handle missing or inconsistent entries.

        ▼ Missing Value

        - Country, League, Match, Player, Team, Team_Attributes doesn't have significant missing value for their major features.
        - Player_Attributes has 836 rows of missing data

```sql
select count(*) as missing_value from Team_Attributes
where id is null or date is null;
```

	missing_value
1	0

```sql
select count(*) as missing_value from Player_Attributes
where id is null or player_fifa_api_id is null or player_api_id is null
or date is null or overall_rating is null
or preferred_foot is null;
```

	missing_value
1	836

Due to the reason that the Player_Attributes has total of 183978 rows of data, so we can drop those missing data.

▼ Duplicated Value

- Most Tables don't have duplicated data, except the Team Table.

```sql
select player_name, birthday,count(*)as duplicated_count from Player
group by player_name, birthday
having duplicated_count>1;
```

player_name	birthday	duplicated_count

```sql
select team_long_name,count(*) as duplicated_count from Team
group by team_long_name
having duplicated_count>1;
```

	team_long_name	duplicated_count
1	Polonia Bytom	2
2	Royal Excel Mouscron	2
3	Widzew Łódź	2

Also drop the duplicated data since it's relatively small.

▼ Inconsistent Value

- All tables don't have inconsistent data.

```sql
select count(*)as inconsistent_count from Match
where home_team_goal<0 or away_team_goal<0 or
      home_team_api_id not in (select team_api_id from Team) or
      away_team_api_id not in (select Team.team_api_id from Team);
```

	inconsistent_count
1	0

```sql
select count(*) as inconsistent_count from Player
where birthday>current_date or birthday<'1900-01-01';
```

	inconsistent_count
1	0

Approach to drop the missing data and duplicated data.

```sql
DELETE from Player_Attributes
where id is null or player_fifa_api_id is null or player_api_id is null
or date is null or overall_rating is null
or preferred_foot is null;

with duplicated_data as (
    select *, row_number() over (PARTITION BY team_long_name) as rnum
    from Team
)
DELETE from Team
where team_long_name in (
    select team_long_name from duplicated_data
                          where rnum>1
    );
```

3. **SQL Query Development:z**

- Top teams by win rate

```sql
with match_results as (
    select Match.match_api_id,
           Team.team_long_name as team_name,
           case
               when Match.home_team_api_id=Team.team_api_id then 'Home'
               when Match.away_team_api_id=Team.team_api_id then 'Away'
           end as match_type,
           case
               when Match.home_team_api_id=Team.team_api_id and
                    Match.home_team_goal>Match.away_team_goal then 'win'
               when Match.home_team_api_id=Team.team_api_id and
                    Match.away_team_goal>Match.home_team_goal then 'lose'
               when Match.away_team_api_id=Team.team_api_id and
                    Match.away_team_goal>Match.home_team_goal then 'win'
               when Match.away_team_api_id=Team.team_api_id and
                    Match.home_team_goal>Match.away_team_goal then 'lose'
               else 'draw'
           end as result
    from Match
    join Team on Match.home_team_api_id=Team.team_api_id or Match.away_team_api_id=Team.team_api_id
)
select team_name, count(*) as total_matches,
       sum(case when result='win' then 1 else 0 end) as num_wins,
       round(sum(case when result='win' then 1 else 0 end)*1.0/count(*),3) as win_rate
from match_results
group by team_name
order by win_rate desc;
```

	team_name	total_matches	num_wins	win_rate
1	FC Barcelona	304	234	0.77
2	Real Madrid CF	304	228	0.75
3	SL Benfica	248	185	0.746
4	FC Porto	248	183	0.738
5	Celtic	304	218	0.717

- Team performances vary across different leagues and seasons

```sql
with match_results as (
    select Match.match_api_id,
           Team.team_long_name as team_name,
           League.name as league_name,
           Match.season,
           case
               when Match.home_team_api_id=Team.team_api_id then 'Home'
               when Match.away_team_api_id=Team.team_api_id then 'Away'
           end as match_type,
           case
               when Match.home_team_api_id=Team.team_api_id and
                    Match.home_team_goal>Match.away_team_goal then 'win'
               when Match.home_team_api_id=Team.team_api_id and
                    Match.away_team_goal>Match.home_team_goal then 'lose'
               when Match.away_team_api_id=Team.team_api_id and
                    Match.away_team_goal>Match.home_team_goal then 'win'
               when Match.away_team_api_id=Team.team_api_id and
                    Match.home_team_goal>Match.away_team_goal then 'lose'
               else 'draw'
           end as result
    from Match
    join Team on Match.home_team_api_id=Team.team_api_id or Match.away_team_api_id=Team.team_api_id
    join League on Match.league_id=League.id
)
select team_name,league_name, season,
    ROUND((SUM(CASE WHEN result = 'win' THEN 1 ELSE 0 END) * 1.0 / COUNT(*)), 2) AS win_rate
from match_results
group by team_name, league_name, season
order by league_name, season, win_rate desc;
```

	team_name	league_name	season	win_rate
1	RSC Anderlecht	Belgium Jupiler League	2008/2009	0.71
2	Standard de Liège	Belgium Jupiler League	2008/2009	0.71
3	Club Brugge KV	Belgium Jupiler League	2008/2009	0.53
4	KAA Gent	Belgium Jupiler League	2008/2009	0.5

- Top players by rating

```sql
select Player.player_name, Player_Attributes.overall_rating from Player join Player_Attributes
  on Player.player_api_id=Player_Attributes.player_api_id
group by player_name
order by overall_rating desc
limit 5;
```

	player_name	overall_rating
1	Lionel Messi	94
2	Cristiano Ronaldo	93
3	Neymar	90
4	Manuel Neuer	90
5	Luis Suarez	90

- Top 3 players across each year

```sql
with yearly_rating as (
    select Player.player_name, Player_Attributes.overall_rating,strftime('%Y', Player_Attributes.date) AS year,
        ROW_NUMBER() OVER (PARTITION BY strftime('%Y', Player_Attributes.date)
            ORDER BY Player_Attributes.overall_rating DESC) AS rating_rank
    from  Player_Attributes join Player on Player.player_api_id=Player_Attributes.player_api_id
)
select year, player_name, overall_rating from yearly_rating
where rating_rank<=3
order by year,rating_rank;
```

	year	player_name	overall_rating
1	2007	Gianluigi Buffon	93
2	2007	Wayne Rooney	93
3	2007	Gregory Coupet	92
4	2008	Cristiano Ronaldo	91
5	2008	Iker Casillas	91
6	2008	Gianluigi Buffon	90
7	2009	Iker Casillas	91
8	2009	Cristiano Ronaldo	90

- Player’s key performance evolving

```sql
select Player.player_name,Player_Attributes.date,Player_Attributes.overall_rating, Player_Attributes.finishing,
       Player_Attributes.dribbling,Player_Attributes.sprint_speed, Player_Attributes.stamina
from Player join Player_Attributes on Player.player_api_id=Player_Attributes.player_api_id
group by Player.player_name, Player_Attributes.date
order by player_name,Player_Attributes.date;
```

	player_name	date	overall_rating	finishing	dribbl
5	Aaron Appindangoye	2016-02-18 00:00:00	67	44	
6	Aaron Cresswell	2007-02-22 00:00:00	53	48	
7	Aaron Cresswell	2008-08-30 00:00:00	53	48	
8	Aaron Cresswell	2009-02-22 00:00:00	47	48	
9	Aaron Cresswell	2009-08-30 00:00:00	52	40	
10	Aaron Cresswell	2010-02-22 00:00:00	51	39	
11	Aaron Cresswell	2010-08-30 00:00:00	54	40	
12	Aaron Cresswell	2011-08-30 00:00:00	64	53	

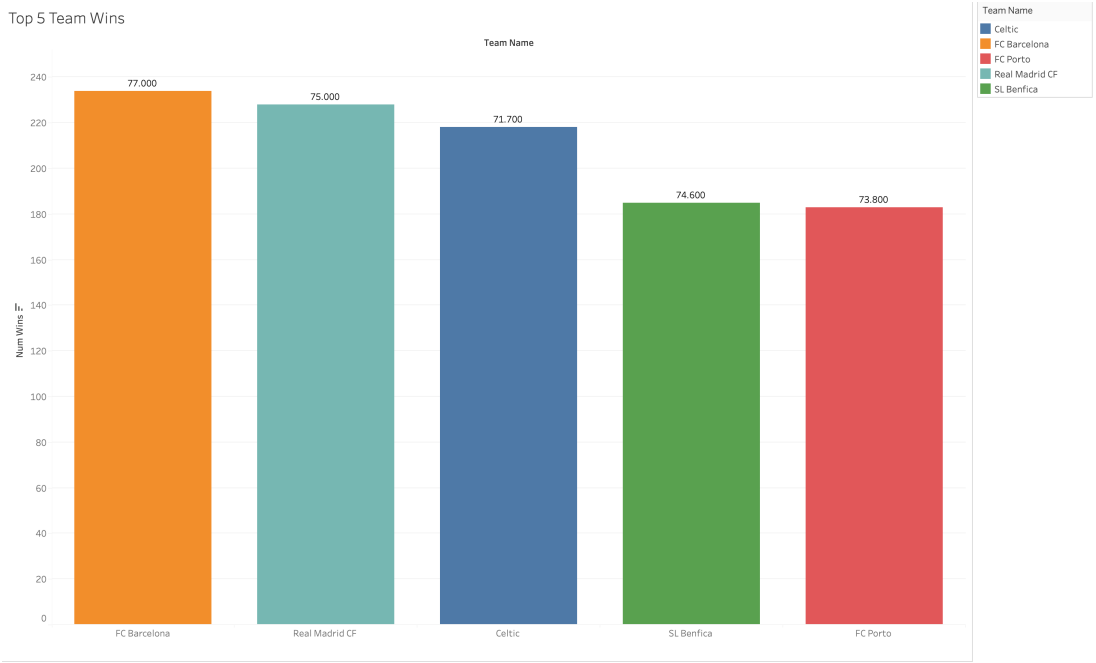
- Home-Away-Wins-Difference

```sql
select home_wins.season,home_wins.total_matches,home_wins.home_wins,away_wins.away_wins,
       round((home_wins.home_wins*100/home_wins.total_matches),2) as home_win_percentage,
       round((away_wins.away_wins*100/away_wins.total_matches),2) as away_win_percentage,
       round(((home_wins.home_wins-away_wins.away_wins)*100/home_wins.total_matches),2) as home_away_wins_difference
from (select season,count(*)as total_matches, SUM(case when home_team_goal>away_team_goal THEN 1 else 0 END)as home_wins
      from Match group by season) as home_wins
join (select season, count(*)as total_matches, SUM(case when away_team_goal>home_team_goal THEN 1 else 0 END)as away_wins
      from Match group by season)as away_wins
 on home_wins.season=away_wins.season
group by away_wins.season;
```

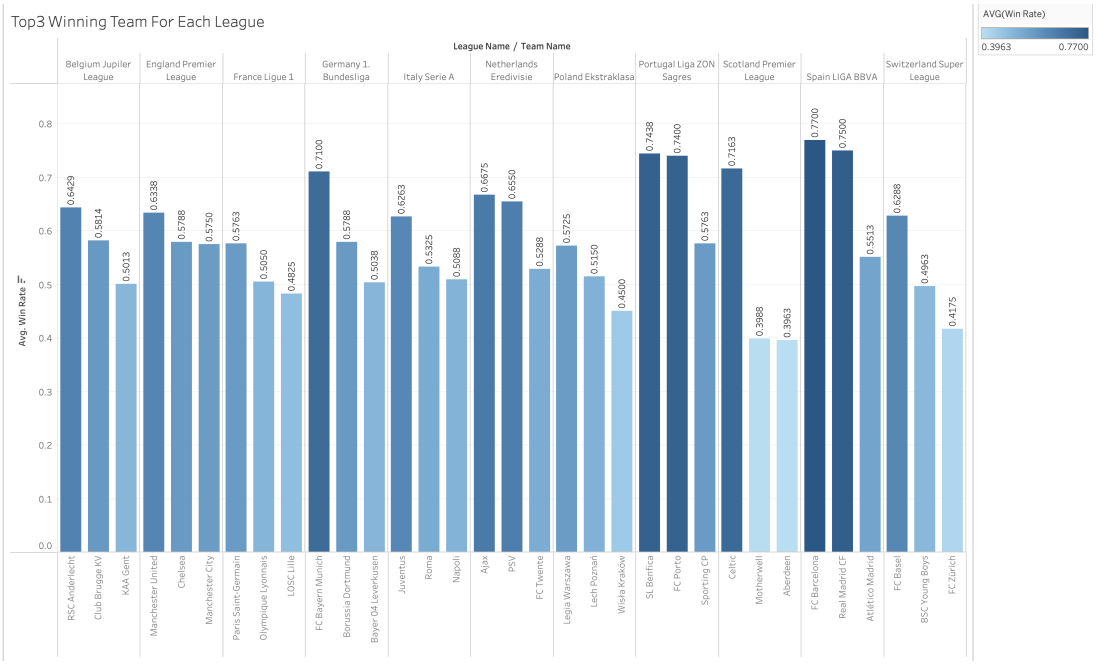
		home_win_percentage	away_win_percentage	home_away_wins_difference
1	929	47	27	19
2	884	47	27	20
3	901	46	27	18
4	904	46	28	18
5	963	44	29	14
6	892	46	29	16
7	981	44	29	15
8	1012	43	30	13

4. **Data Visualization and Reporting:**

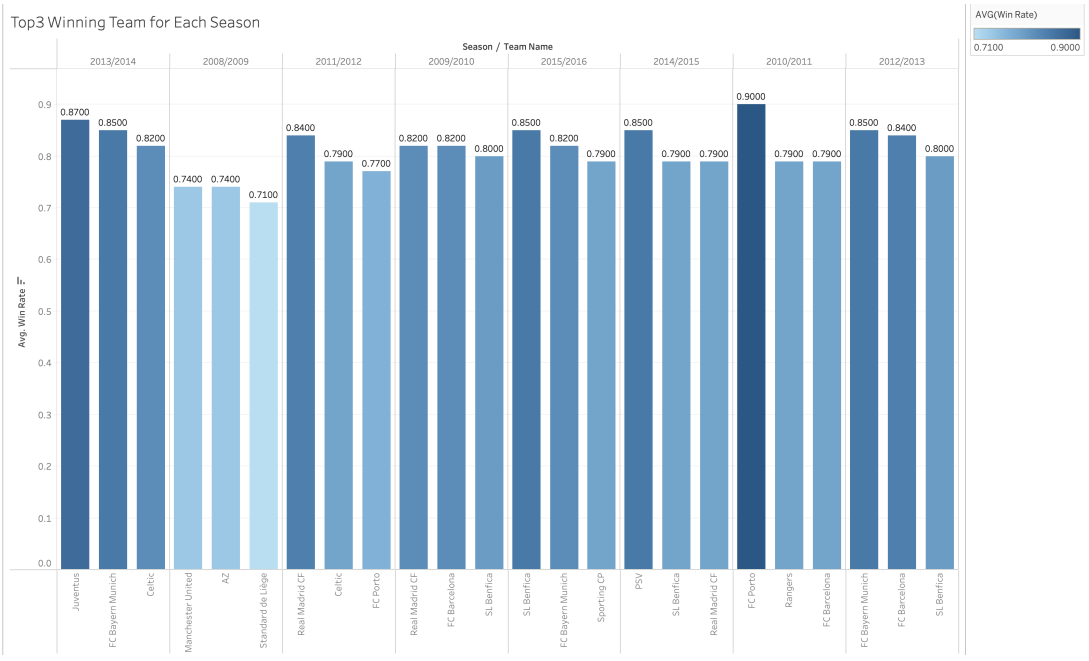
- Top teams by win rate



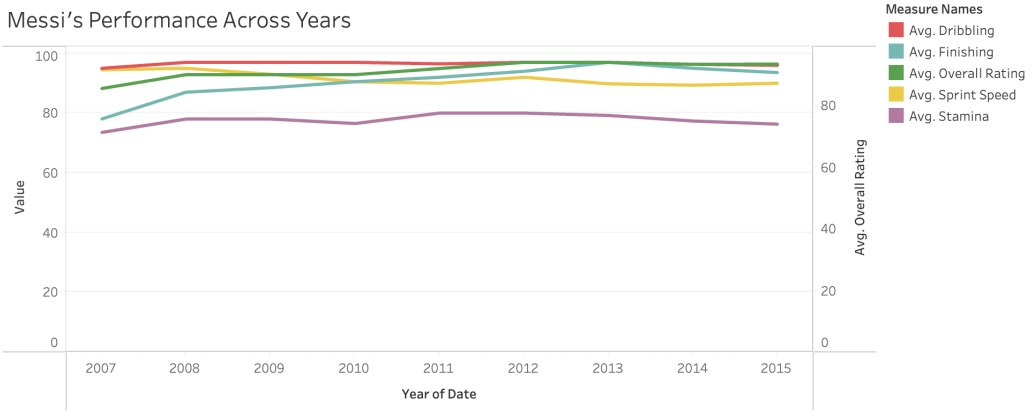
- Team performances vary across leagues.

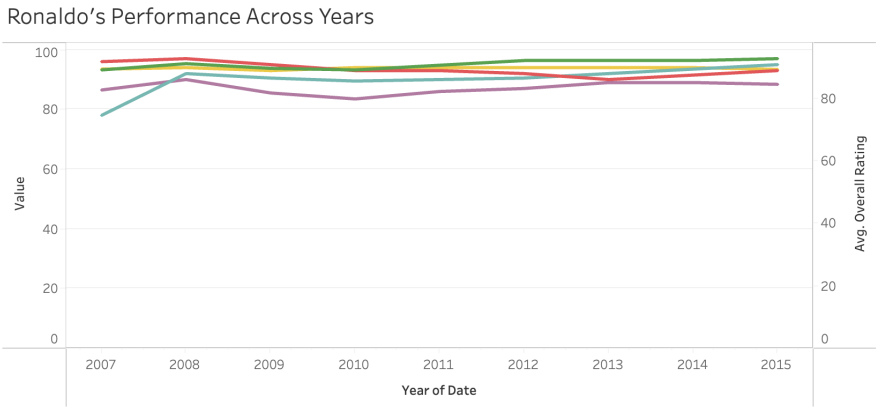


- Team performances vary across seasons

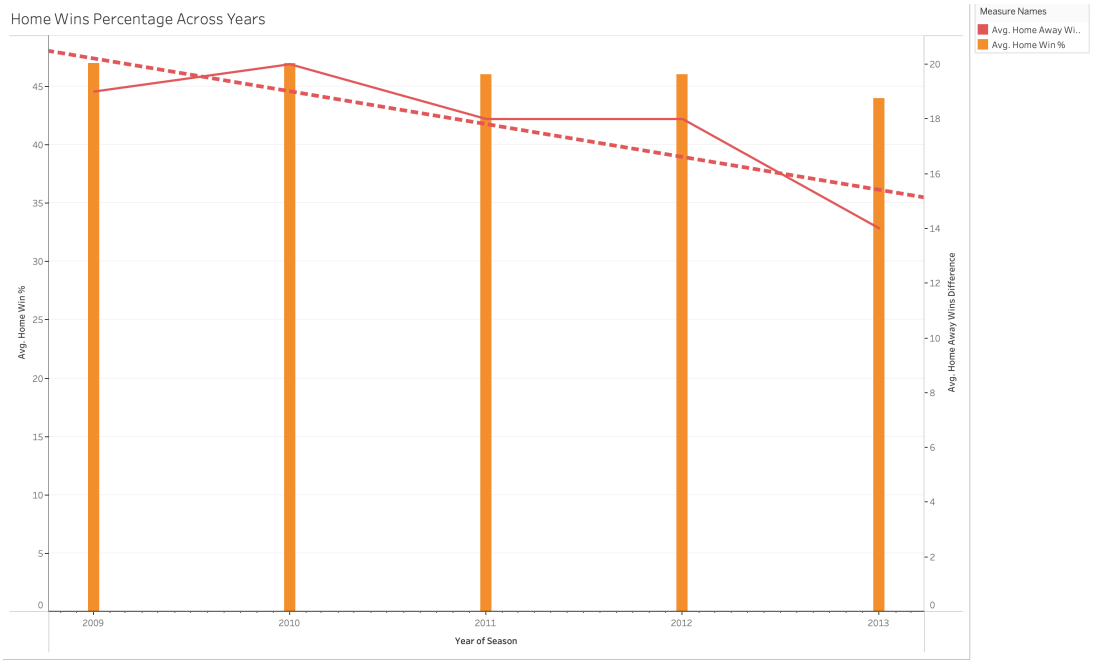


- Player's Performance Across the Year: Messi v.s Ronaldo





- Home Wins Percentage



Insight:

From the analysis on the **European Soccer Database**, we revealed some key insights:

1. Teams with highest win rate are mostly successful across the seasons, eg. FC Barcelona are the highest win rate among all teams, and it has appeared three times in the recent five years.
2. Intensity are varying across different leagues, eg. Portugal LIGA and Spain LIGA both have high intensity since all top 3 team achieve 50% win rate and above with top 2 team all above 70% win rate; whereas the Scotland Premier are less intensity with only one team above 70% win rate and other teams below 40% win rate, such a large win rate gap indicating less intensity in that league.
3. Despite the aging process, both Messi and Ronaldo perform a very consistent elite levels in the major league. And for what fans always arguing about who's the G.O.A.T, Messi has much less physical ability than Ronaldo, but can still competing with him, indicating Messi has some other areas outperform Ronaldo also indicating that Soccer is not just a physical-ability-domain sport.
4. Despite the home team win rate are all above 40%, we can tell by the home-away win-gap is actually narrow down in recent years, suggesting that the home field advantage effect is actually getting smaller. It's a very interest finding that worth to deep research in the future to understand why is the cause of this.

Conclusion

This project using SQL and Tableau to discover some insight and trend from the **European Soccer Database.** We can further use these information and insight to inform strategy for teams and players to enhance their performance and deal with Home-Away game in order to compete in this competitive league.