Football Analysis:

Data Dictionary:

Column Name	Data Type	Description
appearance_id	String	Unique identifier for a player's appearance in a match
game_id	Integer	Unique ID of the match/game
player_id_x	Integer	Unique ID of the player
date_x	Date	Date of the match
player_name_x	String	Name of the player
competition_id_x	String	ID of the competition/tournament
yellow_cards	Integer	Number of yellow cards received by the player in the match
red_cards	Integer	Number of red cards received by the player in the match
Goals	Integer	Number of goals scored by the player in the match
Assists	Integer	Number of assists made by the player
home_club_position	Float	League position of the home club during the season
home_club_manager_name	String	Name of the manager of the home club
away_club_manager_name	String	Name of the manager of the away club
Stadium	String	Name of the stadium where the match is played
attendance	Float	Number of spectators in the match
Referee	String	Name of the referee officiating the match
home_club_name	String	Name of the home team
away_club_name	String	Name of the away team
aggregate	String/Time	Aggregate score for two-legged matches

Business objective questions

Performance analysis:

Business questions:

Which players have the highest goal contribution per 90 minutes?
How do yellow/red cards correlate with performance?

What is the probability distribution of goals per match?

Business insights:

Top Performers:

Christian Cappis leads in goal contribution per 90 minutes, followed by Christian Pulisic and Kenny Saief.
These players are the most efficient in creating and scoring goals.

Impact of Cards:

Correlation between yellow/red cards and goal contribution per 90 minutes = 0.0124
Indicates that receiving more cards does not significantly affect player performance.

Scoring Patterns:

Most matches have 2–3 goals (~22% each), typical for football.
Low-scoring matches (0–1 goal) occur ~20% of the time.
High-scoring matches (>5 goals) are rare (<5%).
.5 goal bins reflect average goals per team, not total match goals.

Player Profile

Business questions:

Do minutes played and competition type influence the number of goals/assists?
Which players perform better in home vs away matches?
Who are the top scorers and assist providers across competitions?

Business insights:

Players tend to score and assist more as the match progresses, with contributions peaking in the 61–90 and 91–120 minute periods. Domestic competitions show the highest contributions, while international and other competitions have less consistent patterns. Overall, both minutes played and competition type influence player output.

Players' offensive efficiency is similar in home and away matches, indicating that they maintain performance regardless of match location.
• Aron Johannsson is the top scorer with 128 goals. • Christian Pulisic contributes both goals and assists, showing versatility. • Coaches can rely on top performers for strategic offensive plays .

- What factors influence stadium attendance (team popularity, day of week)?

- Main factor: Home team popularity — the more popular the team, the more people come.
- Secondary factor: Day of week — weekend matches (Friday/Saturday/Sunday) have slightly higher median attendance and more high-outlier matches than midweek.
- Interaction: Popular teams on weekends = highest attendance.

Team comparison

Business Questions:

- Which teams have the highest average stadium attendance?
- Which teams score the most goals per season?

Business insights:

- This business questions solve by tableau.We done separate visualization for home club and away club teams for attendance stadium. From this visualization we know that home club team i.e., Broussia Dortmund have 87,69,538 avg attendance compared to other teams.
- Season wise 2012 to 2020 Broussia Dortmund have a highest goal.

Attendance & Stadium Analysis

Business Questions:

- What factors influence stadium attendance (team popularity, day of week)?
- Can we classify stadiums into high/medium/low attendance using KNN?
- Which stadiums host the highest scoring games

Business Insights:

- Attendance increases with the home team's popularity, regardless of the day of the week. However, weekend matches (especially Saturday) tend to have higher or more variable attendance than weekday matches, indicating that both team popularity and scheduling affect crowd size.
- Top 10 stadiums by average goals per game show that Hermann-Neuberger-Stadion, MDCC-Arena, Hacker Wiehenstadion, Erzgebirgsstadion, Jahnstadion Regensburg, and Vodafone Park lead with 8.5 goals per game..
- We applied KNN to classify stadiums into High, Medium, and Low attendance based on match features.
- The model achieved 88% overall accuracy, performing best on High and Low attendance stadiums.
- Medium attendance stadiums are sometimes misclassified, likely due to overlapping features.
- This analysis helps identify stadiums likely to attract large or small crowds, useful for planning, ticketing, and staffing

Refree Analysis

Business Questions:

- Which referees issue the most yellow/red cards?
- Do referees favor home or away teams in fouls or cards?
- Is there a significant difference between referees' decisions across leagues?

Business insights:

Top card issuers: Felix Zwayer (16Y /1R), Guido Winkmann (13Y /0R), Felix Brych (12Y /0R)

Red cards are rare, mostly 0–1 per referee.

We analyzed fouls and cards given by referees to home and away teams.
The data shows nearly identical counts for both sides, indicating no systematic favoritism.

Any differences are minor and within normal match variation.

Domestic leagues → most referees, highest card counts → high intensity.

International cups → fewer cards → possibly less aggressive or stricter enforcement.

Substitutions Analysis

Business Questions:

At what time are most substitutions made?

Which players are substituted most frequently?

Are some teams more aggressive in substitutions than others?

Business Insights:

Most substitutions occur between the 64th and 79th minute, with the next highest in the 80th–95th minute, and very few in the first 30 minutes or after the 96th minute.

Fabian Johnson and Christian Pulisic are substituted most frequently, around 59 times each.

Substitution patterns indicate strategic timing in later stages of matches, highlighting key players who are rotated regularly.

In type_x column filter only substitution. then using pivot . Separately doing pivot for home and away club teams. comparing both team,finally away club teams are more aggressive.

Event analysis:

Business Questions:

Which event types occur most frequently per match?

Are more goals scored after substitutions?

Which minutes of the match have the most goals?

Business Insights:

Goals event types occur most frequently .

76-90 minutes have the highest goals.

Before substitutions players scored more goals.

Competition Analysis:

Which competitions have the highest average goals per match?

Are international competitions more competitive (closer scorelines) than domestic leagues?

Do competitions have seasonal trends (more goals or fouls in

first/second half of the season)?

Business Insights:

Other category competitions have the highest avg goals per match.

Then International cup competitions are more competitive than domestic league.

Yes, every two years all competitiors have the highest goals

Player Attributes & Demographics

Business Questions:

Can we cluster players by performance and demographic attributes using K-Means?

Do younger players outperform older players in certain metrics?

Business Insights:

We successfully segmented the players into three distinct groups based on age and goal contribution per 90 minutes.

The analysis reveals a large cluster of low-performing players across many ages, a smaller cluster of moderate performers, and a tiny cluster of very high performers (outliers).

This clustering can be used to target different contract, training or scouting strategies for each group instead of treating all players alike.

This suggests that, in this sample, older players are making more concentrated contributions per match, while younger players are contributing more in raw counts but less per 90 minutes.

Contract Management:

Business Questions:

What is the average contract length for top-performing players?

Which clubs tend to give the longest contracts for top-performing players?

Do players with more yellow/red cards receive shorter contracts?

Business Insights:

We trained a model to predict contract length from age and other factors.

The model achieved 100% accuracy, indicating a very clear pattern.

Younger players are much more likely to receive medium/long contracts, while older players get short-term deals.

Clubs like Aarhus Gymnastik Forening and TSG 1899 Hoffenheim top the list for awarding the most long contracts to high-performing players.

Overall, a handful of clubs dominate long-term deals, showing a stronger commitment to retaining top talent compared to others

Yellow cards: Players in different contract categories show a significant difference in yellow cards ($p < 0.001$), suggesting disciplinary record affects contract length.

Red cards: No significant difference across contract categories ($p > 0.05$), so red cards don't seem to influence contract length.