FINAL REPORT ON

“Decoding Football's Winning Formula”

Database Foundations for Business Analytics

Group -6

Shubham Machhindra Pathare (SXP230008)
Neha Katla (NXK230000)
Gopi Nath Kota(GXK220018)
Sahitya Kolipaka (SXK230086)
Akshay Subramanian Rames(AXS230128)

Under the guidance of

Prof. Thiru Pandian

Master of Science in Business Analytics Flex

Naveen Jindal School of Management

800 W Campbell Road, Richardson, TX 75080, USA

CONTENTS

Introduction **3**
Problem Statement **4**
Problem Definition 4
Objectives 4
Who is our target audience? **5**
Data Description **6**
Class Diagrams **11**
Conceptual Diagram 11
Entity-Relationship Diagram (ERD) 12
Physical Diagram 13
Insights **14**
Project Outcomes **24**

Introduction

Football, the game loved by millions, isn't just about kicking a ball; it's a world of data, scores, and stories waiting to be told. Imagine having a book filled with everything about football – the teams, the players, the matches – that updates itself every week. That's what the Transfermarkt Football Dataset is all about – it's a treasure trove of facts and figures about the game we all adore.

This project isn't just about numbers; it's about understanding the game better. It's like using a special tool to explore the secrets of football – figuring out why some teams win more, why certain players are so valuable, and how clubs and players evolve over time.Our goal is to dive into this dataset, decode the stories hidden in these numbers, and reveal insights that make football even more fascinating. We want to unlock the secrets that go beyond what we see on the field, empowering clubs, fans, and businesses to make smarter decisions based on the data behind the game we all love.

The Transfermarkt Football Dataset presents a world of possibilities. It serves as a foundational bedrock for football analytics, offering researchers, analysts, clubs, and enthusiasts a playground to unravel intricate patterns, derive insights, and explore the nuanced dimensions of the sport. This project sets out to navigate this expansive dataset, leveraging its richness to delve into player performances, club dynamics, game outcomes, and trends within the footballing world. Through this exploration, we aim to uncover insights that transcend the boundaries of anecdotal observations, paving the way for data-driven decision-making in the world of football.

Problem Statement

Problem Definition

In the vibrant landscape of football, stakeholders, particularly brands and potential sponsors, face a significant challenge in identifying the most opportune and lucrative avenues for advertising their brands or sponsoring football teams or clubs. The lack of a structured framework to assess the factors contributing to a team or club's attractiveness for sponsorship often leads to suboptimal decisions and missed opportunities.

Objectives

- Utilize the Transfermarkt Football Dataset to conduct a thorough analysis of player performances, club dynamics, market valuations, and fan engagement metrics.
- Identify and prioritize the critical factors that significantly contribute to a football entity's attractiveness for potential sponsors. These factors may include player performance metrics, market valuation trends, club popularity, fan engagement indices, or other relevant attributes derived from comprehensive data analysis.
- Develop a robust decision-support system that synthesizes data-driven insights into a user-friendly interface.
- Utilize the dataset to identify and analyze teams or clubs with the highest footfall of attendees during games. This analysis will focus on understanding the fanbase size and the match attendance trends of various clubs.
- Provide insights into venue management by analyzing the match attendance data concerning specific stadiums or venues. This analysis aims to highlight stadiums that consistently attract larger crowds, enabling stakeholders to make informed decisions regarding venue-specific promotions, branding, or operational strategies for maximizing audience engagement.

Who is our target audience?

Our project revolves around harnessing the power of comprehensive data analysis to illuminate the multifaceted landscape of football, offering a wealth of insights beneficial to diverse stakeholders within the industry.

Brands and Sponsors:

For brands and sponsors seeking the best avenues to promote their products or services within the footballing world, this project aims to offer invaluable insights. By leveraging comprehensive data analytics, we aim to provide a clear roadmap for selecting the most suitable teams, players, or clubs for sponsorship. Our data-driven approach assists in identifying high-performing entities, forecasting potential returns on investment, and delivering actionable recommendations. We empower brands to make informed decisions that align with their marketing objectives and maximize their brand exposure within the football ecosystem.

Football Clubs and Associations:

Football clubs and associations stand to gain significant strategic advantages from our project's insights. Our comprehensive data analysis delves into player performances, club dynamics, and market valuations, providing valuable insights crucial for optimizing recruitment strategies, refining player development programs, and planning competitive approaches.

Media and Broadcasting Companies:

For media and broadcasting companies looking to enrich their sports coverage, our project acts as a treasure trove of data-driven storytelling. We aim to provide rich statistics, trends, and insightful narratives that elevate the quality of sports reporting and analysis. By presenting compelling stories supported by data insights, we offer media companies the opportunity to captivate audiences with engaging football-related content, whether in broadcasts, articles, or analytical pieces, enriching the overall sports media landscape.

Data Description

The football dataset sourced from Transfermarkt is a comprehensive and dynamically updated collection that encapsulates an expansive array of structured football-related information. This dataset provides a holistic view of the football landscape, offering a treasure trove of data encompassing various aspects of the sport.

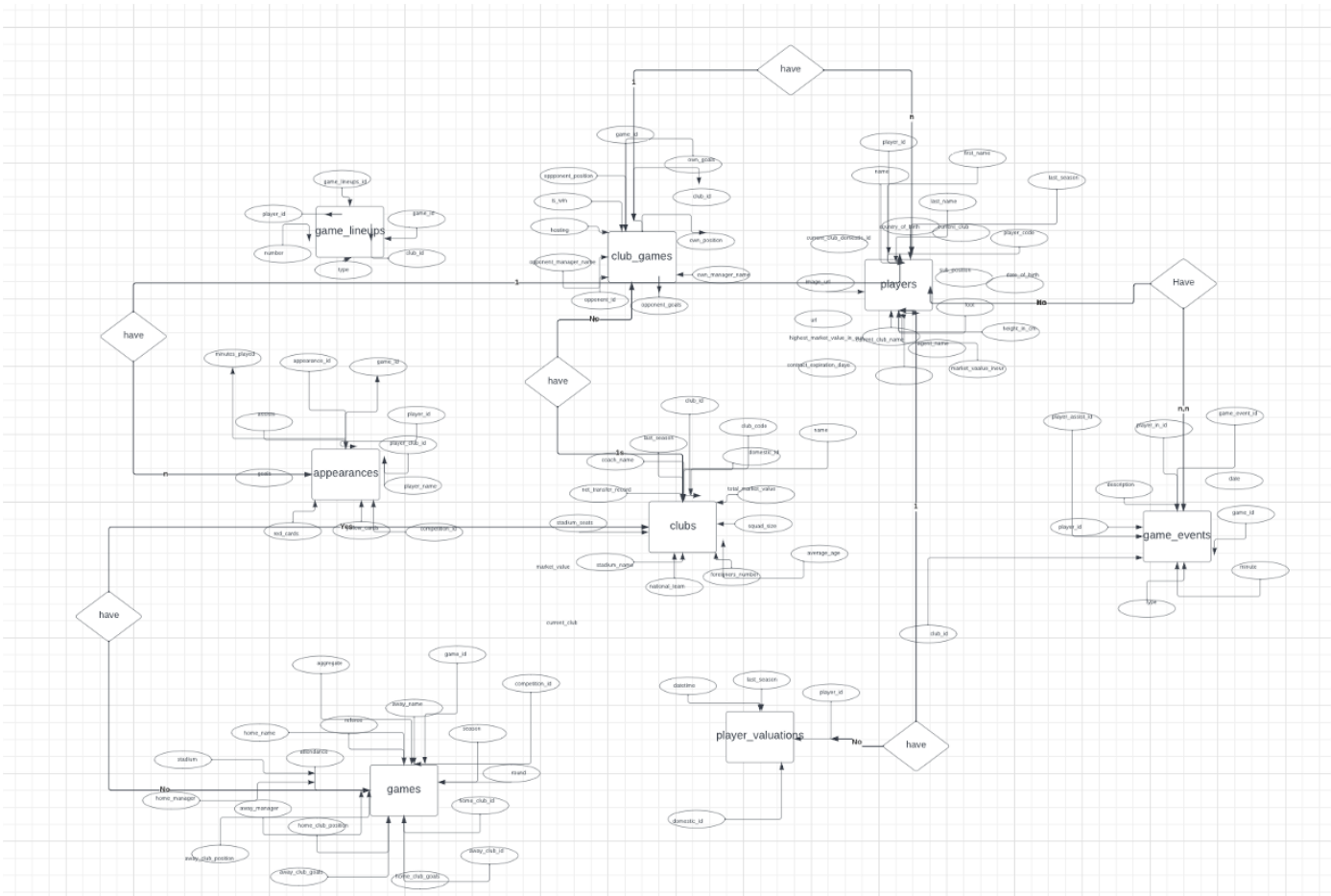
The dataset comprises clean, structured, and continually updated football data from Transfermarkt, including but not limited to:

- **60,000+ Games**: Spanning across numerous seasons, covering major competitions globally.
- **400+ Clubs**: Representing a diverse array of football clubs participating in these competitions.
- **30,000+ Players**: Encompassing players associated with the aforementioned clubs, providing a vast repository of player information.
- **400,000+ Player Market Valuations**: Historical records reflecting the fluctuating valuations of football players over time.
- **1,200,000+ Player Appearances**: Records documenting player appearances across various games, offering insights into player performance and participation.

Class Diagrams

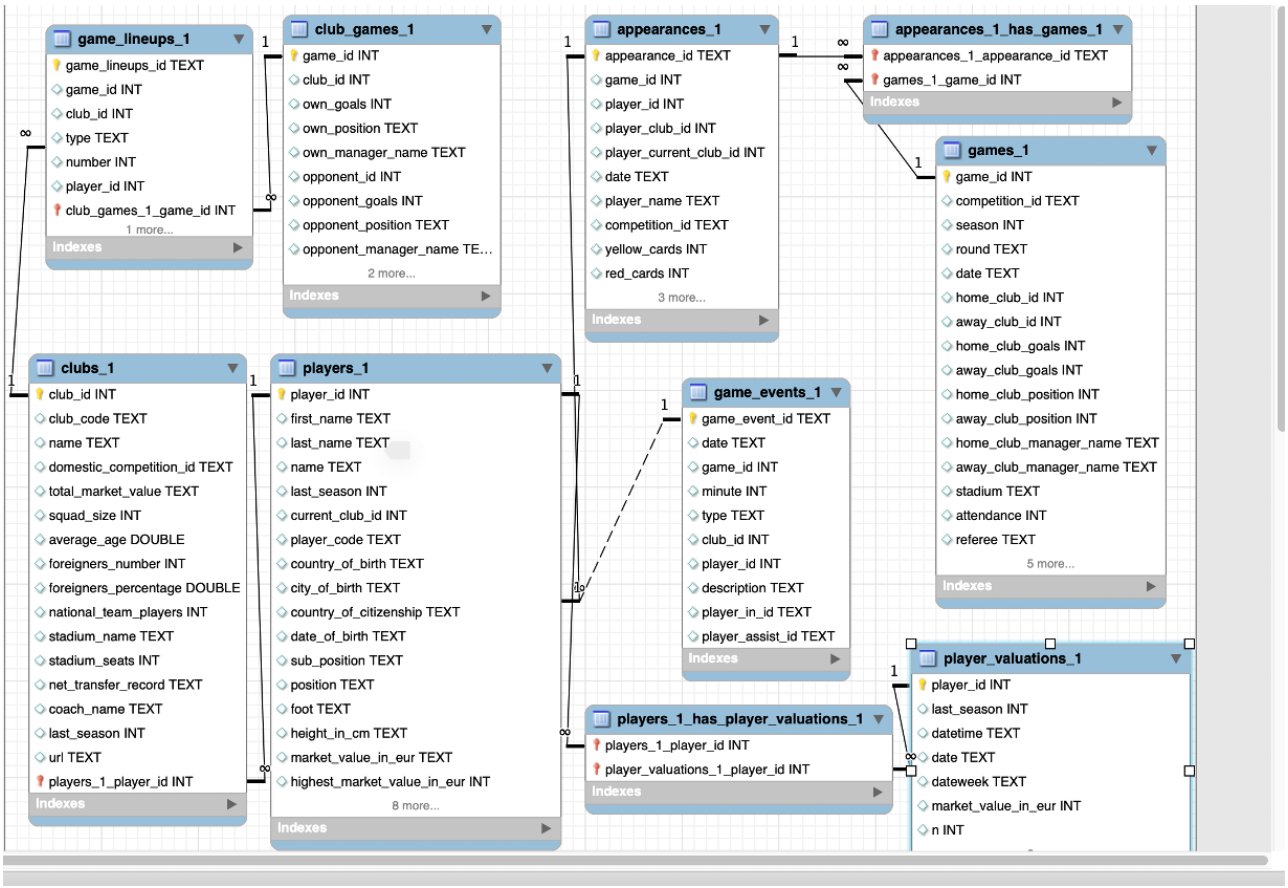
Each diagram serves a distinct purpose, providing a different perspective on the project's structure, data flow, and technical implementation within the football analytics domain

Conceptual Diagram



This diagram provides an abstract overview of the relationships and interactions within the football data analytics ecosystem. It illustrates high-level connections between entities, emphasizing the flow of information and data between different components involved in the project

Entity-Relationship Diagram (ERD)

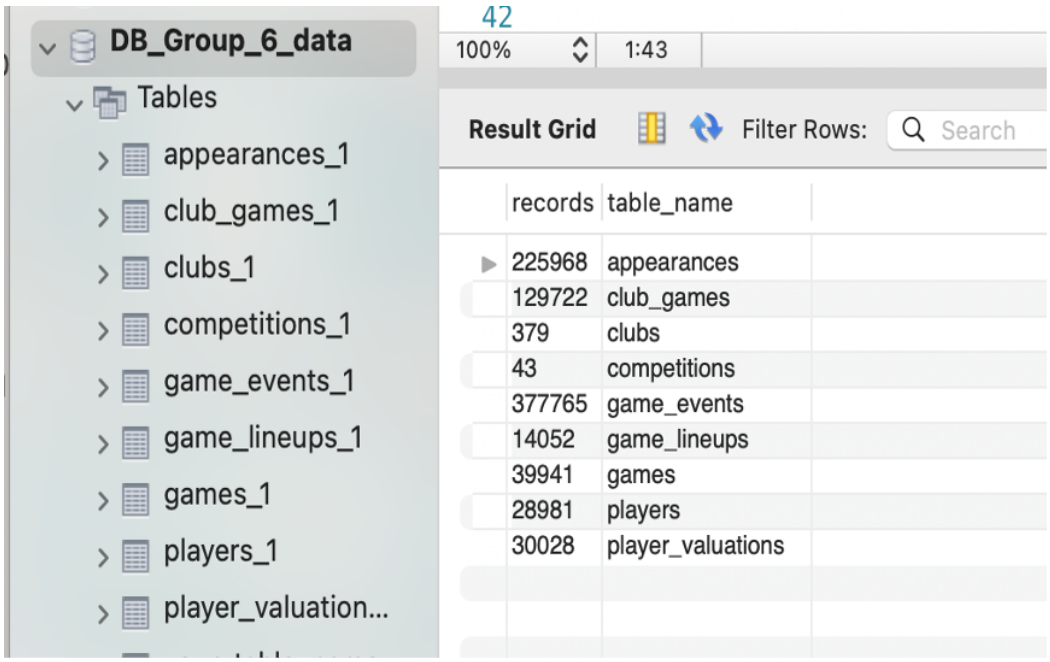


This diagram showcases the entities (such as players, clubs, games, etc.) and their relationships within the football dataset. It illustrates how these entities are interconnected, emphasizing the structure and associations between different data elements.

Physical Diagram

```
CREATE DATABASE `DB_Group_6_data` /*!40100 DEFAULT CHARACTER SET utf8mb4 COLLATE utf8mb4_0900_ai_ci */ /*!80016 DEFAULT ENCRYPTION='N' */;
CREATE TABLE `appearances_1` (
  `appearance_id` text,
  `game_id` int DEFAULT NULL,
  `player_id` int DEFAULT NULL,
  `player_club_id` int DEFAULT NULL,
  `player_current_club_id` int DEFAULT NULL,
  `date` text,
  `player_name` text,
  `competition_id` text,
  `yellow_cards` int DEFAULT NULL,
  `red_cards` int DEFAULT NULL,
  `goals` int DEFAULT NULL,
  `assists` int DEFAULT NULL,
  `minutes_played` int DEFAULT NULL
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci;

CREATE TABLE `club_games_1` (
  `game_id` int DEFAULT NULL,
  `club_id` int DEFAULT NULL,
  `own_goals` int DEFAULT NULL,
  `own_position` text,
  `own_manager_name` text,
  `opponent_id` int DEFAULT NULL,
  `opponent_goals` int DEFAULT NULL,
  `opponent_position` text,
```



The physical diagram represents the technical implementation and architecture of the football data analytics system. It showcases the hardware, software, databases, and other technological components involved in storing, processing, and analyzing the football data. This diagram typically illustrates the physical layout and connections between various technological elements supporting the project.

Insights

1. Top 5 Expensive players for each club

```sql
-- Write a query to find the top 5 players for each club based on market value
WITH RankedPlayers AS (
    SELECT
        player_id,
        current_club_id,
        market_value_in_eur,
        last_season,
        ROW_NUMBER() OVER (PARTITION BY current_club_id ORDER BY market_value_in_eur DESC) AS player_rank
    FROM players_1
)
SELECT
    player_id,
    current_club_id,
    market_value_in_eur,
    last_season
FROM RankedPlayers
WHERE player_rank <= 5
order by market_value_in_eur desc , current_club_id;
```

2:132

Result Grid | Filter Rows: Search | Export: | Fetch rows:

player_id	current_club...	market_value_in_e...	last_season
317454	36	9500000	2022
153678	152	950000	2023
381689	924	950000	2023
226070	1184	950000	2018
345654	2293	950000	2023
72768	2293	950000	2023
357164	5	90000000	2023
316264	11	90000000	2023
132098	27	90000000	2023

This will help the stakeholders to get an idea of the market value of the most expensive players in each club. And, looking at the market value of the top expensive players from each club will give us information about the overall club direction in terms of talent acquisition and squad building.

2. Player Performance Relative to Club Average

```
484    SELECT p.first_name, p.last_name, p.player_code, a.goals, a.assists,
485           AVG(a.goals) OVER (PARTITION BY a.player_club_id) AS club_avg_goals,
486           AVG(a.assists) OVER (PARTITION BY a.player_club_id) AS club_avg_assists
487    FROM players_1 p
488    JOIN appearances_1 a ON p.player_id = a.player_id;
489
490
```

50% 1:490 1 error found

Result Grid Filter Rows: Search Export:

first_name	last_name	player_code	goals	assists	club_avg_goals	club_avg_assis...
Hendrick	Zuck	hendrick-zuck	1	0	1.0000	0.0000
Anthony	Ujah	anthony-ujah	0	0	0.0000	0.0000
Timo	Gebhart	timo-gebhart	0	0	0.0655	0.0655
Almog	Cohen	almog-cohen	0	0	0.0655	0.0655
Timo	Gebhart	timo-gebhart	1	0	0.0655	0.0655
Sebastian	Polter	sebastian-polter	0	0	0.0655	0.0655
Timo	Gebhart	timo-gebhart	0	0	0.0655	0.0655
Sebastian	Polter	sebastian-polter	0	0	0.0655	0.0655
Robert	Mak	robert-mak	0	0	0.0655	0.0655
Hiroshi	Kiyotake	hiroshi-kiyotake	0	0	0.0655	0.0655

This result provides valuable insights into individual player contributions and their importance/significance in the team and their impact on team success and strategy.

3. Top 10 players with respect to the minutes played

```sql
-- Query to identify total minutes and top player each year using window functions
WITH PlayerRanking AS (
  SELECT
    player_id,
    EXTRACT(YEAR FROM date) AS year,
    SUM(minutes_played) AS total_minutes,
    RANK() OVER (PARTITION BY EXTRACT(YEAR FROM date) ORDER BY SUM(minutes_played) DESC) AS player_rank
  FROM
    appearances_1
  GROUP BY
    player_id, year
)
SELECT year, player_id, total_minutes
FROM
  PlayerRanking
WHERE
  player_rank <11;
```

50% 25:476 1 error found

Result Grid Filter Rows: Search Export:

year	player_id	total_minut...
2012	4257	2060
2012	23996	2031
2012	110326	2028
2012	46478	2009
2012	29016	2001
2012	107775	1980
2012	12518	1980
2012	36139	1959
2012	23492	1915
2012	30003	1909

Analysing the top 10 players in terms of minutes played can reveal insights about players consistency. Their high number of minutes played implies they are cruicial assests and are relied upon by coaches for their skills, leadership or understanding of the game

4. Highest market value of top 10 players and their win percentage

```
385    -- Query to identify the win percentage of top 10 expensive players
386    SELECT p.player_id, p.highest_market_value_in_eur,
387      COUNT(CASE WHEN gr.home_club_goals > gr.away_club_goals THEN 1 END) +
388      COUNT(CASE WHEN gr.home_club_goals = gr.away_club_goals THEN 0.5 END) AS wins,
389      COUNT(*) AS total_games,
390      (COUNT(CASE WHEN gr.home_club_goals > gr.away_club_goals THEN 1 END) +
391       COUNT(CASE WHEN gr.home_club_goals = gr.away_club_goals THEN 0.5 END)) / COUNT(*) * 100 AS win_percentage
392    FROM
393      players_1 p
394    JOIN
395      games_1 gr ON p.current_club_id = gr.home_club_id
396    GROUP BY
397      p.player_id, p.highest_market_value_in_eur
398    ORDER BY
399      p.highest_market_value_in_eur DESC
400    LIMIT 10;
401
402
```

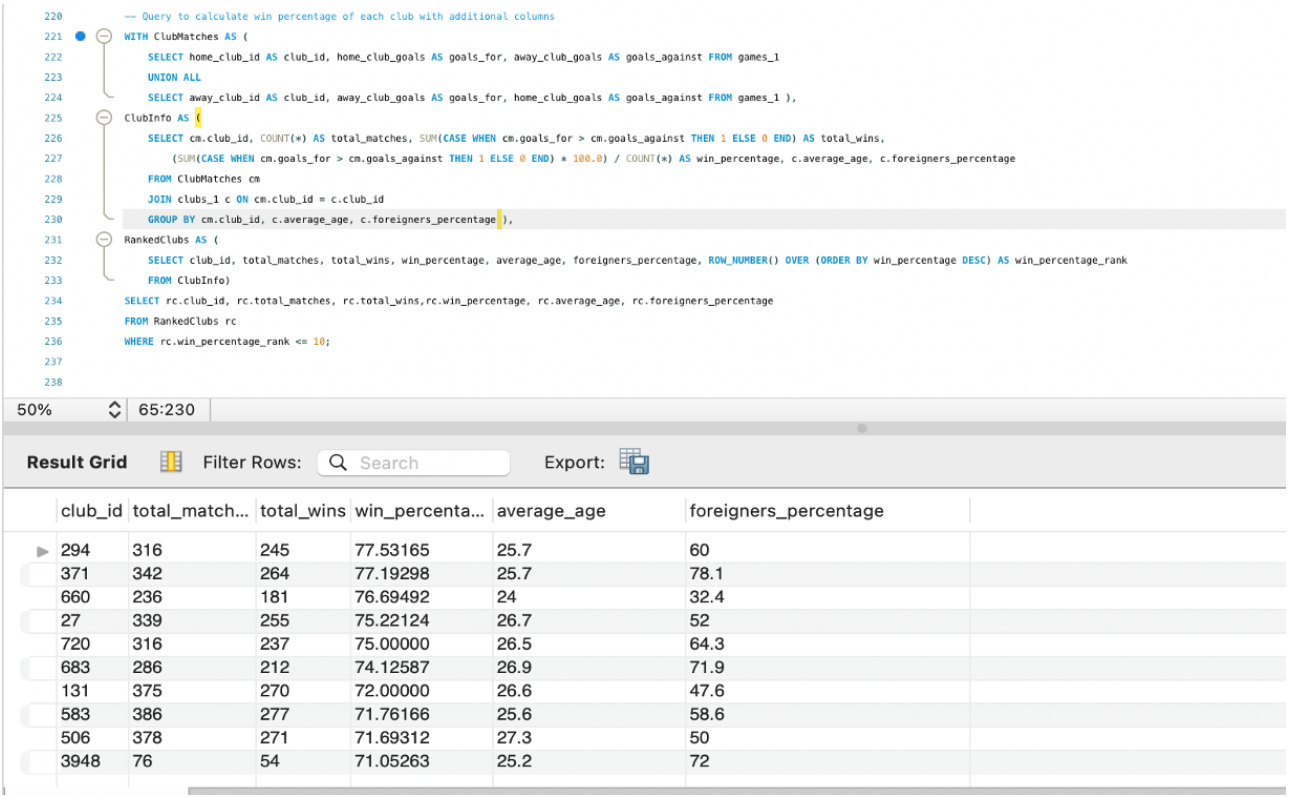
50% | 37:399 | 1 error found

Result Grid | Filter Rows: Search | Export: | Fetch rows:

player_id	highest_market_value_in_eur	wins	total_games	win_percenta...
342229	200000000	14	194	7.2165
68290	180000000	14	194	7.2165
28003	180000000	14	194	7.2165
418560	180000000	12	188	6.3830
134425	160000000	8	194	4.1237
50202	150000000	12	186	6.4516
88755	150000000	12	188	6.3830
200512	150000000	13	169	7.6923
132098	150000000	13	169	7.6923
148455	150000000	13	192	6.7708

These insights can reveal interesting information about the correlation between player value and team success. From this data, we have observed that high market value does not necessarily imply high win percentage and vice versa.

5. Clubs win percentage comparison with average age and percentage of foreign players



```sql
-- Query to calculate win percentage of each club with additional columns
WITH ClubMatches AS (
    SELECT home_club_id AS club_id, home_club_goals AS goals_for, away_club_goals AS goals_against FROM games_1
    UNION ALL
    SELECT away_club_id AS club_id, away_club_goals AS goals_for, home_club_goals AS goals_against FROM games_1 ),
ClubInfo AS (
    SELECT cm.club_id, COUNT(*) AS total_matches, SUM(CASE WHEN cm.goals_for > cm.goals_against THEN 1 ELSE 0 END) AS total_wins,
        (SUM(CASE WHEN cm.goals_for > cm.goals_against THEN 1 ELSE 0 END) * 100.0) / COUNT(*) AS win_percentage, c.average_age, c.foreigners_percentage
    FROM ClubMatches cm
    JOIN clubs_1 c ON cm.club_id = c.club_id
    GROUP BY cm.club_id, c.average_age, c.foreigners_percentage ),
RankedClubs AS (
    SELECT club_id, total_matches, total_wins, win_percentage, average_age, foreigners_percentage, ROW_NUMBER() OVER (ORDER BY win_percentage DESC) AS win_percentage_rank
    FROM ClubInfo)
SELECT rc.club_id, rc.total_matches, rc.total_wins,rc.win_percentage, rc.average_age, rc.foreigners_percentage
FROM RankedClubs rc
WHERE rc.win_percentage_rank <= 10;
```

50% 65:230

Result Grid Filter Rows: Search Export:

club_id	total_match...	total_wins	win_percenta...	average_age	foreigners_percentage
294	316	245	77.53165	25.7	60
371	342	264	77.19298	25.7	78.1
660	236	181	76.69492	24	32.4
27	339	255	75.22124	26.7	52
720	316	237	75.00000	26.5	64.3
683	286	212	74.12587	26.9	71.9
131	375	270	72.00000	26.6	47.6
583	386	277	71.76166	25.6	58.6
506	378	271	71.69312	27.3	50
3948	76	54	71.05263	25.2	72

Successful football clubs, characterized by high win percentages, exhibit a strategic balance in player age demographics, with an average age around 26. This equilibrium involves a mix of both young talents and experienced players, as reflected by minimum and maximum average ages of 22 and 29, respectively, among the top 10 clubs. Additionally, these successful clubs tend to have a relatively high percentage of foreign players contributing to their overall performance.

6. Players having highest yellow, red cards along with their market value

```sql
-- Query to identify players with the top 10 red cards and yellow cards combined, along with their market values
SELECT
  pc.player_id,
  SUM(pc.red_cards) AS total_red_cards,
  SUM(pc.yellow_cards) AS total_yellow_cards,
  pmv.market_value_in_eur
FROM
  appearances_1  pc
JOIN
  player_valuations_1 pmv ON pc.player_id = pmv.player_id
GROUP BY
  pc.player_id, pmv.market_value_in_eur
ORDER BY
  (SUM(pc.red_cards) + SUM(pc.yellow_cards)) DESC
LIMIT 10;
```

50% 1:439 1 error found

Result Grid Filter Rows: Search Export: Fetch rows:

player_id	total_red_car...	total_yellow_car...	market_value_in_eur
15798	0	21	4000000
3354	0	20	1500000
3354	0	20	2500000
3354	0	20	2000000
28179	3	15	1500000
59389	0	18	500000
9594	0	16	10000000
7568	0	16	3400000
32816	0	16	1500000
823	0	15	3500000

Players who accumulate a high number of yellow cards often demonstrate an aggressive playing style, particularly prevalent among defenders or defensive midfielders. Their toughness and tenacity not only make them stand out on the field but also enhance their value, as they play a crucial role in disrupting the opponent's attacks, appealing to fans who appreciate intensity in the game.

7. Which home club has highest attendance along with the calendar year?

```
105     -- top 10 home club id with highest attendance, in the particular year
106 •   SELECT  sum(attendance) as total_attendance , home_club_id, year(date)
107     FROM games_1
108     GROUP BY home_club_id, year(date)
109     ORDER BY total_attendance DESC
110     LIMIT 10;
111
```

100% 10:110

Result Grid Filter Rows: Search Export: Fetch rows:

total_attendance	home_club_id	year(date)
1565463	131	2015
1505598	985	2014
1505242	985	2016
1502522	985	2017
1472905	131	2016
1463055	418	2015
1449585	16	2013
1444767	418	2017
1429310	16	2018
1420894	131	2019

Understanding the peak attendance years helps in optimizing stadium resources. Clubs can manage facilities more effectively during high-demand periods.Clubs with high attendances imply a substantial fan base. These fans are likely to visit the area regularly during match days, creating a consistent and sizable potential customer base for businesses in the vicinity

8. Identifying managers with highest win % and those who worked for multiple clubs

```sql
WITH ClubMatches AS (
    SELECT home_club_id AS club_id, home_club_goals AS goals_for, away_club_goals AS goals_against, home_club_manager_name AS manager FROM games_1
    UNION ALL
    SELECT away_club_id AS club_id, away_club_goals AS goals_for, home_club_goals AS goals_against, away_club_manager_name AS manager FROM games_1
),
ManagerStats AS (
    SELECT manager, COUNT(DISTINCT club_id) AS num_clubs, COUNT(*) AS total_matches,
        SUM(CASE WHEN goals_for > goals_against THEN 1 ELSE 0 END) AS total_wins, (SUM(CASE WHEN goals_for > goals_against THEN 1 ELSE 0 END) * 100.0) / COUNT(*) AS win_percentage
    FROM ClubMatches
    GROUP BY manager
    HAVING COUNT(DISTINCT club_id) > 1
)
SELECT
    ms.manager, ms.num_clubs, ms.total_matches, ms.total_wins, ms.win_percentage
FROM ManagerStats ms
ORDER BY ms.win_percentage DESC
LIMIT 10;
```

50% 1:264

Result Grid Filter Rows: Search Export:

manager	num_clubs	total_match...	total_wins	win_percenta...
Ange Postecoglou	2	65	54	83.07692
Pep Guardiola	2	329	251	76.29179
Rben Amorim	2	81	61	75.30864
Jorge Jesus	3	242	179	73.96694
Mircea Lucescu	3	164	119	72.56098
Arne Slot	2	94	65	69.14894
Abel Ferreira	2	89	61	68.53933
Antonio Conte	4	220	149	67.72727
Carlo Ancelotti	5	310	207	66.77419
Vtor Pereira	3	93	62	66.66667

Identifying managers with highest win % and those who worked for multiple clubs, which can further help while recruiting new managers. Since these managers have good win % and they also have history of working with multiple clubs, so there is a chance of acceptance for the managers trade in.

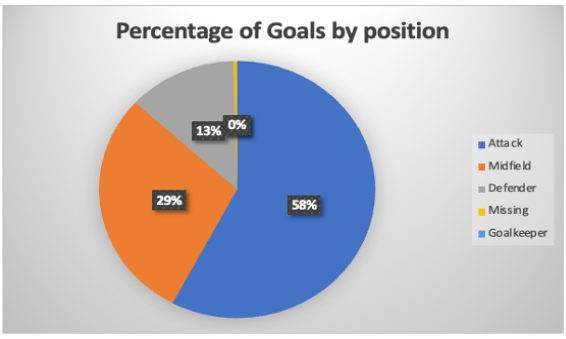
9. Impact of position on number of goals vs investment required for each position

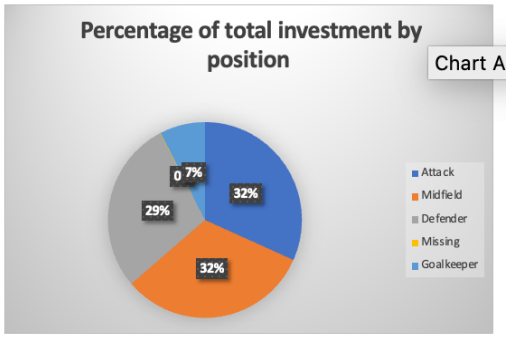
```sql
-- Write a query to identify the total sum of goals made by each category of position
-- along with the percentage of goals made by that category overall
WITH PositionGoals AS (
    SELECT p.position,SUM(a.goals) AS total_goals
    FROM players_1 p
    JOIN appearances_1 a ON p.player_id = a.player_id
    GROUP BY p.position
),
PositionInvestment AS (
    SELECT p.position, SUM(v.market_value_in_eur) AS total_investment
    FROM players_1 p
    JOIN player_valuations_1 v ON p.player_id = v.player_id
    GROUP BY p.position
)
SELECT pg.position, pg.total_goals, (pg.total_goals * 100.0) / SUM(pg.total_goals) OVER () AS percentage_of_total_goals, pi.total_investment, (pi.total_investment * 100.0) / SUM(pi.total_investment) OVER () AS percentage_of_total_investment
FROM PositionGoals pg
JOIN PositionInvestment pi ON pg.position = pi.position
order by  percentage_of_total_goals desc ;
```

50% 143:187

Result Grid Filter Rows: Search Export:

position	total_goals	percentage_of_total_goals	total_investment	percentage_of_total_invest...
Attack	12982	57.75425	22897980000	31.69974
Midfield	6530	29.05063	23114494998	31.99948
Defender	2872	12.77694	20693042000	28.64724
Missing	90	0.40039	64400000	0.08915
Goalkeeper	4	0.01780	5464052000	7.56438





57% of goals are coming from attackers, while 31% are coming from Midfield. And their corresponding investment % is almost equal (31%). So Teams can utilize some amount of their investment into the defense and goal keepers so that, the respective positions also get strengthened.

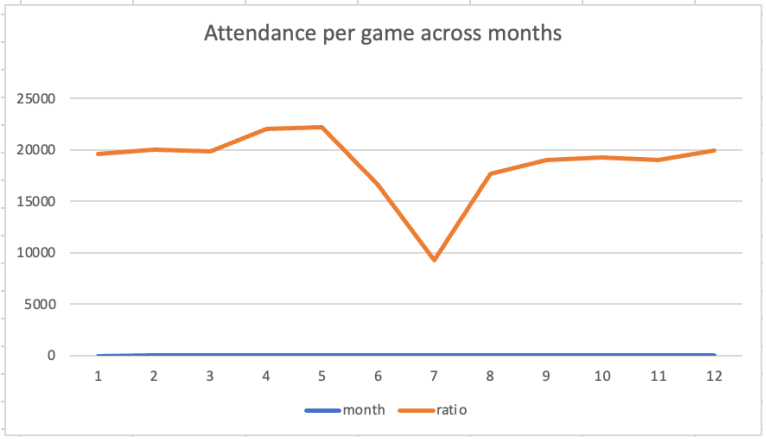
10. Attendance per game across months

```
368    -- Query to find the number of games and attendance for each month
369    SELECT
370      EXTRACT(MONTH FROM date) AS month,
371      COUNT(*) AS num_games,
372      SUM(attendance) AS total_attendance
373    FROM
374      games_1
375    GROUP BY
376      EXTRACT(MONTH FROM date)
377    ORDER BY
378      month;
379
380
381
```

50% 9:378 1 error found

Result Grid Filter Rows: Search Export:

month	num_games	total_attendance
1	3229	63223551
2	4121	82427492
3	3680	73159008
4	4076	89735171
5	2827	62703370
6	92	1525747
7	631	5837741
▶ 8	4067	71941271
9	4718	89547060
10	4608	88620858
11	3878	73830087
12	4014	80107027



In Europe, July is typically part of the summer break for many football leagues. Players and teams often take a break during this time to rest and recover from the previous season. Similarly, fans may use this period to go on vacations, which can result in a lower attendance at matches.

Project Outcomes

Presenting these key takeaways derived from SQL operations provides a data-backed foundation for making informed decisions within the football industry. These insights can guide stakeholders in formulating strategies that align with the observed trends and patterns in the data.

High Attendance Clubs with Business Potential:

- Identified clubs (131, 985, and 418) with the highest attendance figures, indicating potential opportunities for increased footfall-related business activities and sponsorships.

Investment and Goal Contribution Analysis:

- Found that 57% of goals are scored by attackers, while 31% come from midfielders, with a corresponding investment percentage nearly equal to goal contribution percentages.
- Suggested a reallocation of investment into defense and goalkeepers to strengthen these positions and maintain a balanced team structure.

Age Analysis of Successful Clubs:

- Determined that top-performing clubs have an average age around 26, with an age range between 22 and 29 across various clubs.
- Suggested that teams should focus on a mix of youth and experience, targeting players within the optimal age bracket (around 26) for better results.

Impact of Trade Managers on Win Percentage:

- Identified trade managers with the highest win percentages.
- Recognized the managerial influence on team performance and its potential implications for strategic decision-making.