**Article Name** : “Football Analytics: A Comprehensive Examination of Player and Club Performance”

**Course Title: (Microsoft Power BI Engineer)**

**Group code: CAI1\_DAT2\_S1e**

|  |  |  |  |
| --- | --- | --- | --- |
| **Mohamed Khaled Ahmed Ebrahim** | **Islam Mohammed Ahmed Mohammed** | **Bahey Eldeen Mahmoud Bahey Eldeen** | **Student Names** |

**Technical Instructor Name: MARWAN**

**Company name: AAST/MEA**

# Overview

This project focuses on analyzing a football dataset to extract meaningful insights by answering key business questions through SQL queries and Power BI visualizations. The dataset includes various tables such as *players*, *clubs*, *appearances*, *transfers*, and *games*. These tables provide comprehensive data on player statistics, club performance, game outcomes, and transfer activities across multiple seasons.

The primary goal of the project is to derive valuable insights that can aid football managers, analysts, and club officials in making data-driven decisions. Using SQL queries, the data was explored and transformed to answer critical business questions related to player performance, club success, and transfer market trends. Power BI was then used to visualize the results in a dynamic and interactive dashboard, enabling stakeholders to easily interpret the findings.

Key insights include identifying the top-performing players based on goals, assists, and minutes played, club performance trends across various competitions, and patterns in the football transfer market. Additionally, the analysis provided valuable metrics for understanding player valuation trends, the impact of transfers on club success, and performance indicators for managers and clubs.

**Key Metrics**

Some of the critical metrics explored in this report are:

* **Player Performance**:  
  Metrics such as goals scored, assists provided, minutes played, yellow/red cards, and overall contribution to team success were analyzed to highlight top-performing players.
* **Club Standings and Performance**:  
  Win percentage, home vs. away performance, and the relationship between club standings and financial data (e.g., market value, transfer record) were key focus areas for club analysis.
* **Transfers**:  
  Insights into the football transfer market, including the most expensive transfers, net transfer records for clubs, and the relationship between player market value and their transfer fees.
* **Match Outcomes**:  
  Game-by-game analysis, including win/loss ratios, goals scored, goals conceded, and comparisons of club performance in different competitions or against various opponents.
* **Player Valuation**:  
  Trends in player valuations over time, as well as the impact of key performance metrics and transfers on a player’s market value.

# Introduction

**Project Objective**

The primary objective of this project is to conduct a comprehensive analysis of football-related data using SQL Server and Power BI. The goal is to answer important business questions by leveraging data from player statistics, match outcomes, club performance, and transfer activities. Through this analysis, actionable insights can be derived to inform decision-making processes for football managers, analysts, and other stakeholders. The project aims to present these insights through dynamic visualizations in Power BI, allowing for easy interpretation of trends and performance metrics in the football world.

Key objectives include:

* **Player performance analysis**, including metrics such as goals, assists, and minutes played.
* **Club performance** and standings across different competitions.
* **Transfer market trends**, identifying high-value transfers and their impact on clubs.
* **Match outcomes and performance** of clubs in various domestic and international competitions.

**Data Sources**

The data used in this project consists of multiple interconnected tables stored in SQL Server. The main datasets include:

* **Appearances**: Contains data about player appearances in various matches, including information on goals, assists, minutes played, yellow/red cards, and club affiliations.
* **Players**: Provides detailed information about players, including their names, positions, market value, height, and nationality.
* **Clubs**: Contains data about football clubs, including their market value, squad size, stadium information, and last season’s performance.
* **Games**: Holds match-level data such as game results, goals scored, goals conceded, and home vs. away performance.
* **Transfers**: Includes information about player transfers, transfer fees, and the clubs involved in the transfers.
* **Competitions**: Contains details about domestic and international football competitions, including the competition type and participating clubs.
* **Player Valuations**: Records player market values over time and their association with clubs.
* **Club Games**: Captures club-level performance data, including game outcomes, opposition details, and win/loss records.
* **Game Lineups**: Stores information about team formations, starting lineups, and the roles of individual players in matches.

These datasets are connected using primary and foreign keys, such as (player\_id), (club\_id), and (game\_id). This structure enables the analysis of player performance in relation to club success, transfer activities, and match outcomes.

# Importance of the Analysis

Football data analysis is critical for clubs, managers, and business professionals in the sports industry. In the highly competitive world of football, data-driven insights are essential for making informed decisions about player signings, team strategies, and financial investments.

* **For Football Clubs and Managers**: This analysis helps identify key performers, assess the impact of transfers, and refine strategies based on past match outcomes. Understanding player and club performance over time can guide decisions on player development, team tactics, and transfer market activities.
* **For Business Professionals**: The financial side of football, including player market value and transfer fees, is closely tied to club success. By analyzing these metrics, business professionals can make better investment decisions and forecast trends in player valuations.

# Data Exploration & Preparation

**Data Cleaning**

During the exploration phase of the football dataset, various data quality issues were identified, including missing values, duplicates, and inconsistencies. Addressing these issues was essential to ensure the accuracy and reliability of the analysis.

**Missing Values:**

* Some tables, such as *appearances* and *transfers*, contained NULL values in critical fields like goals, assists, transfer\_fee, and market\_value\_in\_eur.
* The *games* table also had NULL values for attendance and home\_club\_position in certain matches.

**Duplicates:**

* Duplicate records were identified in the *players* and *game\_lineups* tables. For instance, some players were recorded multiple times for the same game due to inconsistencies in naming conventions or errors in data entry.

**Inconsistencies:**

* Inconsistent data types were discovered, especially with numeric fields stored as strings (e.g., market\_value\_in\_eur and yellow\_cards stored as VARCHAR).
* Date formats were not standardized across tables, with some dates recorded in the YYYY-MM-DD format, while others used different formats or included time information unnecessarily.

**Steps Taken to Clean the Data**

1. **Handling Missing Values:**
   * NULL values in the goals, assists, and market\_value\_in\_eur columns were replaced with zeroes (0) where applicable, as it is reasonable to assume that if a player did not record a goal or assist, the value should be zero.
   * For columns like attendance and home\_club\_position, which are crucial for match analysis, records with missing values were either imputed based on averages or excluded from specific queries.
2. **Removing Duplicates:**
   * Duplicate records were removed by identifying rows with identical values in key fields such as player\_id and game\_id. Duplicates in the *players* table were resolved by merging entries based on unique player identifiers.
3. **Correcting Inconsistent Data Types:**
   * Data types were standardized for numerical fields. For example, market\_value\_in\_eur and transfer\_fee were converted from VARCHAR to INT for better arithmetic operations.
   * Similarly, fields like yellow\_cards, red\_cards, and goals were corrected where they were stored as strings.
4. **Standardizing Dates:**
   * Dates stored in different formats across the dataset were standardized to the YYYY-MM-DD format using SQL’s date conversion functions.

# Data Transformation

Several data transformations were applied to enhance the analytical capabilities of the dataset and to prepare it for visualizations in Power BI.

1. **Date Format Conversion:**
   * The date columns in the *games*, *appearances*, and *transfers* tables were transformed from strings to SQL date types to facilitate time-based analysis, such as analyzing player performance by season or tracking transfer trends over time.
2. **Splitting Player Names:**
   * Player names, which were stored as a single name field in the *players* table, were split into separate first\_name and last\_name fields. This enabled more granular analysis, such as ranking players by last name or sorting them alphabetically.
3. **Merging Data for Club Analysis:**
   * The *club\_games* table was merged with the *clubs* table to enrich the data with additional club details, such as market value and squad size. This allowed for a deeper analysis of club performance in relation to their financial status.
4. **Aggregating Player Valuations:**
   * The *player\_valuations* table was aggregated to analyze the average and maximum market value for players over time. This transformation allowed for trends in player valuations to be tracked across different seasons.
5. **Creating Player Performance Metrics:**
   * Additional metrics such as goal-per-minute ratios and assist-per-minute ratios were calculated by dividing the goals and assists fields by the minutes\_played field in the *appearances* table. This transformation provided a more nuanced view of player performance beyond raw totals.

These cleaning and transformation steps were critical for ensuring that the dataset was ready for analysis and visualization. By addressing inconsistencies and preparing the data with additional transformations, the dataset became more useful for uncovering insights and presenting them in Power BI visualizations.

# Data Modelling

**Relationships Setup in Power BI**

When importing the football dataset from SQL Server into Power BI, one of the crucial steps was establishing the relationships between the various tables to ensure accurate analysis and visualizations. Power BI’s *Model* view was used to define these relationships, mainly based on foreign key constraints that existed within the SQL Server database.

**Primary Relationships Established:**

1. **Players and Appearances:**
   * **Key Field:** player\_id
   * The *players* table was connected to the *appearances* table using the player\_id as the primary foreign key. This relationship allowed for detailed analysis of player performance metrics such as goals, assists, and minutes played by individual players.

**Challenge:** In some cases, discrepancies arose when linking player\_id across tables due to missing player records or duplicated entries in the *players* table, which required manual data validation.

1. **Clubs and Club Games:**
   * **Key Field:** club\_id
   * The ***clubs*** table was linked to the *club\_games* table via the club\_id. This connection made it possible to analyze how each club performed in different matches and over multiple seasons, adding valuable context to performance metrics like market value, squad size, and foreign players’ presence.

**Challenge:** Some inconsistencies arose in the club\_id fields, where mismatches between the club IDs in the *club\_games* and *clubs* tables led to difficulties in establishing the connection. These issues were resolved through SQL data cleaning before import.

1. **Games and Appearances:**
   * **Key Field:** game\_id
   * The *games* table was linked to *appearances* via game\_id, enabling the analysis of player and team performance at the match level. This relationship made it possible to assess how individual players contributed to specific games and evaluate match outcomes.
2. **Transfers and Players:**
   * **Key Field:** player\_id
   * The *transfers* table was connected to the *players* table using the player\_id. This relationship was essential for analyzing player transfers, movement between clubs, and the financial aspect of football, including transfer fees and market value trends.
3. **Player Valuations and Players:**
   * **Key Field:** player\_id
   * The *player\_valuations* table was linked to the *players* table through the player\_id. This relationship was critical for analyzing player market values over time, providing insights into player performance, career progression, and financial standing within the market.

**Challenges in Setting Up Relationships:**

* **Linking Player Data Across Multiple Tables:**  
  One of the primary challenges was linking the player\_id across various tables. In some cases, the player information was incomplete or duplicated across the *players*, *transfers*, and *appearances* tables. This issue was resolved by carefully cleaning the data in SQL and ensuring that the player\_id field was consistent across all tables before importing it into Power BI.
* **Handling Missing Foreign Key Values:**  
  Certain tables had missing foreign key values, such as NULL club\_id or game\_id, which made it difficult to establish relationships for a few records. For instance, when player transfer data lacked a corresponding from\_club\_id or to\_club\_id, those transfers could not be fully linked to a specific club’s data. These rows were either removed or imputed where possible.

**Key Tables for Analysis**

Some tables in the dataset were more critical than others for answering the business questions posed by the project. Below is an explanation of the essential tables and why they are vital for analysis:

1. **Appearances:**
   * **Why It's Critical:**  
     The *appearances* table provides a detailed account of individual player performances in each game, including goals, assists, yellow cards, red cards, and minutes played. This data is crucial for performance analysis and helps answer business questions like *“Which players are the top performers?”* and *“How do player contributions vary by game and competition?”*

**Use in Analysis:**  
Metrics such as goals per minute, assists per match, and player contribution by game were derived from this table, providing a granular view of player success.

1. **Transfers:**
   * **Why It's Critical:**  
     The *transfers* table is fundamental for tracking player movements between clubs, including the financial details of transfers, such as fees and market values. This table allows us to answer business questions like *“Which clubs have spent the most on transfers?”* or *“How do transfer market values correlate with player performance?”*

**Use in Analysis:**  
Transfer trends, club spending patterns, and player valuations were key metrics derived from this table, offering insights into the financial aspect of football.

1. **Player Valuations:**
   * **Why It's Critical:**  
     The *player\_valuations* table tracks the market value of players over time, making it essential for understanding the financial dynamics of the football market. This data helps answer questions like *“Which players have increased in value over time?”* and *“How does market value fluctuate based on performance?”*

**Use in Analysis:**  
Trends in player market values were visualized, highlighting which players had the highest growth in market value and how market value trends changed across seasons.

1. **Clubs:**
   * **Why It's Critical:**  
     The *clubs* table offers a broad overview of each football club, including squad size, average age, market value, and national team players. This data helps contextualize club performance in relation to financial health and player composition.

**Use in Analysis:**  
Club analysis, including how financial resources (market value) impact performance, was derived from this table. Additionally, comparisons between clubs’ financial metrics and their on-field success were visualized.

1. **Games:**
   * **Why It's Critical:**  
     The *games* table contains essential match data such as results, goals, and positions of the clubs involved. This table was pivotal for answering questions like *“Which clubs perform better in home versus away games?”* and *“What are the trends in match outcomes by season?”*

**Use in Analysis:**  
Match-level analysis, such as home versus away performance and win/loss trends, was drawn from this table, contributing to the understanding of game dynamics.

# DAX Calculations

**Why DAX Was Needed**

DAX (Data Analysis Expressions) is a powerful formula language used in Power BI, Excel, and other Microsoft tools designed for data modeling and analysis. It enables users to create custom calculations, measures, and calculated columns, which enhance the analytical capabilities of the data model.

DAX is essential in the context of this project for several reasons:

* **Dynamic Calculations:** DAX allows for real-time calculations based on filters and user interactions, making the insights in Power BI dashboards dynamic and context-sensitive.
* **Complex Aggregations:** DAX can handle complex calculations and aggregations that go beyond basic sums and averages, facilitating more nuanced analysis of the football data.
* **Performance Optimization:** By defining measures with DAX, calculations are computed on the fly during analysis, which can lead to better performance compared to performing the same calculations in the original dataset.

# Challenges

**Data Challenges**

The initial stages of the project revealed several data-related challenges that impacted the analysis:

1. **Inconsistent Values for Match Dates:**  
   The match dates in the *Games* table were not consistently formatted, which posed difficulties in filtering and aggregating data by date. Some entries used different date formats, making it essential to standardize these formats for accurate time-based analysis.
2. **Inconsistent Player Transfers:**  
   The *Transfers* table sometimes included inconsistent naming for clubs, leading to difficulties in correctly linking player transfers to their respective clubs. Variations in club names (e.g., abbreviations, full names) complicated the analysis of transfer histories and trends.
3. **Missing or NULL Values:**  
   Several key metrics, such as goals scored and player market values, contained NULL entries. These gaps required careful handling to avoid skewing analysis, necessitating a strategy for imputation or exclusion in certain calculations.

**Visualization Challenges**

Creating visualizations in Power BI also presented various challenges:

1. **Formatting Complex Hierarchies:**  
   Visualizing complex hierarchies, such as the relationships between players, clubs, and competitions, required careful design. Ensuring that users could easily navigate from player statistics to club performance and competition outcomes necessitated thoughtful selection of visual types and arrangement.
2. **Handling Large Data Volumes:**  
   The dataset's size posed performance issues when creating visuals, particularly when using multiple filters. This required optimization techniques in Power BI to ensure smooth user interactions, including aggregating data before visualization.
3. **Dynamic Filtering:**  
   Implementing dynamic filters that allowed users to drill down into specific segments (like player positions or club types) while maintaining performance and usability was challenging. Ensuring these filters worked seamlessly across different visuals was essential for effective analysis.

# Findings & Insights

**Player Insights**

1. **Performance Trends:**  
   The analysis revealed notable performance trends among players, including:
   * **Top Goal Scorers:** Players who consistently scored goals emerged as key assets for their clubs, with performance metrics highlighting their contributions to match outcomes.
   * **Most Valuable Players:** Market value analyses indicated that certain players not only excelled on the field but also held significant market value, influencing transfer dynamics.
2. **Club Loyalty vs. Transfer Frequency:**  
   A comparison of player transfers and the duration of their tenure at clubs provided insights into club loyalty. Players with long tenures often contributed more significantly to their clubs, whereas frequent transfers often correlated with lower performance metrics.

**Club Insights**

1. **Winning Percentages by Season:**  
   The club performance analysis uncovered varying winning percentages across seasons, with specific clubs demonstrating consistent success. Visualizations highlighted trends in performance, indicating which clubs improved or declined over time.
2. **Net Transfer Records and Market Value Impact:**  
   The correlation between net transfer records (transfers in versus transfers out) and the clubs’ market values provided insights into how strategic transfer activities affected overall club valuation. Clubs with positive net transfers often saw a boost in market value.

**Competition Insights**

1. **Performance Analysis Across Competitions:**  
   Clubs' performance varied across different competitions, with some excelling in domestic leagues while struggling in international tournaments. This insight is crucial for understanding clubs' strategic focus.
2. **Home vs. Away Game Results:**  
   The analysis highlighted significant differences in performance based on whether clubs were playing at home or away. Clubs that maintained strong home records often had advantages in league standings and competition outcomes, illustrating the importance of home-field advantage.

# Conclusion

**Summary of Key Findings**

The comprehensive analysis of the football dataset yielded several critical insights that have significant implications for clubs, managers, and stakeholders in the football industry:

1. **Player Performance Trends:**
   * **Top Goal Scorers:** Identifying the top goal scorers not only highlighted key players but also provided insights into potential targets for clubs looking to enhance their attacking capabilities.
   * **Market Value Assessments:** The relationship between player performance and market value revealed players whose contributions were undervalued or overvalued, providing actionable insights for potential transfers.
2. **Club Performance Dynamics:**
   * **Winning Percentages:** Analyzing winning percentages by season allowed for the identification of clubs that consistently perform well, which could inform strategies for future matches and recruitment.
   * **Net Transfer Impact:** The correlation between net transfer activity and market value illustrated how strategic player acquisitions and disposals could influence a club’s financial health and competitive standing.
3. **Competition Insights:**
   * **Performance Across Competitions:** Understanding which clubs excelled in specific competitions provided strategic insights into where clubs should focus their resources and efforts.
   * **Home vs. Away Performance:** The analysis of home versus away performance underscored the importance of creating a strong home advantage, guiding clubs in their preparations and tactics.

**Emphasize Actionable Insights**

1. **Potential Player Transfers:**  
   The analysis highlighted players with exceptional performance metrics who may be suitable targets for clubs seeking to improve their squads. Clubs should consider these insights in their recruitment strategies.
2. **Strategies for Upcoming Matches:**  
   Clubs can leverage the insights on home and away performance to develop tailored strategies for their upcoming matches. Understanding opponent weaknesses and strengths can inform tactical adjustments.
3. **Investment in Player Development:**  
   Identifying players whose performance metrics exceed their market value suggests opportunities for investment in talent development, potentially leading to higher returns through future transfers.

**Future Improvements**

While the analysis provided valuable insights, several areas for improvement could enhance future studies:

1. **Inclusion of Additional Data:**  
   Incorporating more granular data, such as player injuries, training regimens, and psychological assessments, could provide a more holistic view of player performance and club dynamics.
2. **Deeper Analysis of Player Performance:**  
   Conducting deeper analysis using advanced metrics, such as expected goals (xG), assists per 90 minutes, and defensive contributions, would yield a more nuanced understanding of player value and effectiveness.
3. **Longitudinal Studies:**  
   Implementing longitudinal studies to track player performance and club dynamics over multiple seasons would allow for trend analysis and more informed predictions regarding future performance.
4. **Enhanced Visualization Techniques:**  
   Exploring more interactive and advanced visualization techniques in Power BI could provide richer insights and improve user engagement. Incorporating features like drill-down capabilities, predictive analytics, and scenario modeling would enhance the analytical experience.

### **Appendices**

#### **Appendix A: SQL Queries**

Below is the list of 10 SQL queries utilized to extract, clean, and analyze the football dataset for Power BI visualizations:

1. **Total Goals Scored by Each Player**:

sql

Copy code

SELECT player\_name, SUM(goals) AS total\_goals

FROM appearances

GROUP BY player\_name

ORDER BY total\_goals DESC;

1. **Top Players by Assists**:

sql

Copy code

SELECT player\_name, SUM(CAST(assists AS INT)) AS total\_assists

FROM appearances

GROUP BY player\_name

ORDER BY total\_assists DESC;

1. **Player Transfers Summary**:

sql

Copy code

SELECT player\_name, COUNT(\*) AS transfer\_count

FROM transfers

GROUP BY player\_name

ORDER BY transfer\_count DESC;

1. **Average Market Value of Players by Club**:

sql

Copy code

SELECT p.current\_club\_name, AVG(p.market\_value\_in\_eur) AS avg\_market\_value

FROM players p

GROUP BY p.current\_club\_name;

1. **Top Scorers in a Specific Competition**:

sql

Copy code

SELECT player\_name, SUM(goals) AS competition\_goals

FROM appearances a

JOIN competitions c ON a.competition\_id = c.competition\_id

WHERE c.name = 'Premier League'

GROUP BY player\_name

ORDER BY competition\_goals DESC;

1. **Clubs with the Most Players**:

sql

Copy code

SELECT current\_club\_name, COUNT(\*) AS player\_count

FROM players

GROUP BY current\_club\_name

ORDER BY player\_count DESC;

1. **Average Goals per Game for Each Club**:

sql

Copy code

SELECT club\_id, AVG(home\_club\_goals) AS avg\_goals

FROM games

GROUP BY club\_id;

1. **Player Performance Over Time**:

sql

Copy code

SELECT player\_name, date, SUM(goals) AS goals\_per\_match

FROM appearances

GROUP BY player\_name, date

ORDER BY date;

1. **Transfer Fees Over Time**:

sql

Copy code

SELECT transfer\_date, SUM(CAST(transfer\_fee AS INT)) AS total\_transfer\_fee

FROM transfers

GROUP BY transfer\_date

ORDER BY transfer\_date;

1. **Comparison of Home vs. Away Goals**:

sql

Copy code

SELECT home\_club\_id,

SUM(home\_club\_goals) AS home\_goals,

SUM(away\_club\_goals) AS away\_goals

FROM games

GROUP BY home\_club\_id;

# Appendix B: DAX Formulas

Below is a detailed list of DAX formulas used for creating measures in Power BI:

1. **Performance Index**:

DAX

Copy code

PlayerPerformanceIndex = DIVIDE([TotalGoals] + [TotalAssists], SUM(Appearances[minutes\_played]), 0)

1. **Win Percentage**:

DAX

Copy code

WinPercentage = DIVIDE(SUMX(ClubGames, IF(ClubGames[is\_win] = "Yes", 1, 0)), COUNT(ClubGames[game\_id]), 0)

1. **Average Market Value**:

DAX

Copy code

AvgMarketValue = AVERAGE(Players[market\_value\_in\_eur])

1. **Total Assists**:

DAX

Copy code

TotalAssists = SUMX(Appearances, VALUE(Appearances[assists]))

1. **Net Transfer Impact**:

DAX

Copy code

TotalNetTransfer = SUM(Clubs[net\_transfer\_record])

# Appendix C: Data Sources

1. **Appearances Table**: Data on player appearances, goals, assists, and disciplinary records.
2. **Club Games Table**: Information on matches played by clubs, including outcomes and attendance.
3. **Clubs Table**: Data on clubs, including financial metrics and squad details.
4. **Competitions Table**: Information on various competitions, including league structure and details.
5. **Game Events Table**: Events that occurred during matches, such as substitutions and goals.
6. **Game Lineups Table**: Player lineups for each game, including positions and captaincy.
7. **Games Table**: Match results and statistics, including club positions and manager details.
8. **Player Valuations Table**: Market value changes for players over time.
9. **Players Table**: Comprehensive player information, including demographics and performance data.
10. **Transfers Table**: Details on player transfers, including clubs involved and transfer fees.