Football Data Analytics

Kush Bindal
Data Analytics
San Jose State University
San Jose, California
kush.bindal@sjsu.edu

Shobhita Agrawal
Data Analytics
San Jose State University
San Jose, California
shobhita.agrawal@sjsu.edu

Darpankumar Jiyani
Data Analytics
San Jose State University
San Jose, California
darpankumarpareshbhai.jiyani@sjsu.edu

Dhruv Patel
Data Analytics
San Jose State University
San Jose, California
dhruv.h.patel@sjsu.edu

ABSTRACT

In today's fast-paced world of sports, where the strategic nuances of football continue to evolve, analytics has emerged as a game-changing tool for clubs aiming to gain a competitive edge. Major football clubs like Manchester City and Bayern Munich now rely on dedicated analytics teams year-round, recognizing the pivotal role of data-driven insights in optimizing player selection, squad building, and game strategies. This project harnesses a comprehensive dataset spanning multiple seasons and competitions, featuring over 60,000 games, 400 clubs, and 30,000 players from prominent leagues such as the UEFA Champions League. By leveraging advanced analytics techniques, including player-specific performance analysis based on match conditions, opponent strengths, and tactical scenarios, we aim to reveal hidden insights that revolutionize decision-making both on and off the field. Through detailed exploration of players' dynamics alongside environmental factors and opponent tactics, our analysis provides strategic intelligence that enhances player selection and tactical strategies. Ultimately, our goal is to demonstrate the transformative impact of football data analytics in navigating the ever-evolving landscape of modern sports.

INTRODUCTION

Football, often referred to as the world's most popular sport, captivates millions of fans worldwide with its electrifying matches, passionate rivalries, and moments of sheer brilliance on the field. Behind the spectacle lies a wealth of data waiting to be analyzed and decoded. In recent years, the emergence of football data analytics has revolutionized the way teams, coaches, and analysts perceive the game. By harnessing the power of advanced statistical techniques, machine learning algorithms, and data visualization tools, analysts can delve deeper into the intricacies of player performance, tactical strategies, and match outcomes.

The primary aim of this project is to explore the realm of football data analytics, uncovering valuable insights that can enhance our understanding of the game and provide actionable intelligence for teams, coaches, and enthusiasts alike. By leveraging vast repositories of historical match data, player statistics, and tactical information, we seek to unravel the hidden patterns, trends, and correlations that shape the dynamics of football. At its core, this project seeks to unlock the secrets hidden within the data, revealing patterns, trends, and relationships that offer profound insights into player performance, team dynamics, and match outcomes. By harnessing the power of advanced analytics techniques, we aim to transform raw data into actionable intelligence, empowering teams, coaches, and analysts to make informed decisions and gain a competitive edge on the field.

MOTIVATION

The motivation for football analytics stems from a desire to enhance the sport by advancing beyond conventional manual computations that provide limited insights such as goals scored and assists. Leveraging data from top football leagues over the past decade, our project aims to develop a comprehensive model that predicts current player performance based on historical records from tournaments like the UEFA Champions League. By analyzing individual player dynamics, strategic formations, and match statistics, teams can extract valuable insights to optimize strategies, improve player performance, and gain a competitive advantage. Traditional methods of gaining insights through repetitive SQL queries on extensive datasets are time-consuming and inefficient. Therefore, our initiative seeks to streamline this process by constructing a real-time performance model that harnesses historical data to deliver actionable insights promptly. By empowering coaches and team managers with timely and accurate performance predictions, our project aims to facilitate informed decision-making before matches, leading to enhanced team performance and strategic adjustments. This project represents a significant step towards elevating the standard of football analytics and optimizing performance assessment in the sport.

LITERATURE REVIEW

The research of (Klyuchka, Cherednichenko, Vasylenko, & Yakovleva, 2015) aimed to find the most important factors that are not confidential information and can be easily determined before the start of the football match. It presents that forecasting rules are used to increase the accuracy of predicting the results of football matches by identifying the winning team based on data retrieved from results of previous games championships, adding substantial factors, to understand the influence of results.

According to (Ali, 2011) and raising the bridge to arrive on the study subject, football is a complex sport, requiring the repetition of many disparate actions. For instance, there are

several proofs of concepts that are currently being used such as assessing the physical prowess of players, approaching simple running tests using monitor speed, agility proofs, and repeated sprint performance.

Joining what (Ali, 2011) concludes, (De Silva, et al., 2018) present that performance management of top football players is a complex system involving enhancement of physical performance, skill-based training, tactical training, minimization of injury risk, and psychological support. Managing practice is vital to allowing players to perform at an optimal level throughout a play season's length. (Constantinou & Fenton, 2017) agrees with (Ali, 2011) and (FIFA, 2019) that Football is the most popular sport in the world and it leverages the inspiration of several researchers to use football activities as a real-world application field to test various statistical, probabilistic, and machine-learning techniques.

Central to football analytics is the availability of comprehensive data sources and robust analytical frameworks. In "The Numbers Game" (2013), Chris Anderson and David Sally provided a comprehensive overview of the statistical methods and analytical techniques used in football analysis, highlighting the importance of data-driven decision-making in player recruitment, tactical planning, and performance evaluation. Subsequent works by Simon Kuper and Stefan Szymanski, including "Soccernomics" (2009) and "Moneyball for Soccer" (2016), further underscored the transformative potential of analytics in football, emphasizing the need for clubs to embrace data-driven strategies to gain a competitive edge.

Football analytics encompasses a diverse array of methodologies and techniques, ranging from descriptive statistics to advanced machine learning algorithms. In "Football Hackers" (2017), Christoph Biermann delved into the world of football data science, showcasing innovative approaches to player evaluation, tactical analysis, and predictive modeling. Michael Lewis's "Moneyball" (2003) offered valuable insights into the application of sabermetrics in baseball, inspiring a wave of research in sports analytics and influencing the development of predictive modeling techniques in football.

METHODOLOGY

For the implementation of the whole project, we have followed Agile Methodology using Scrum. The Scrum board is created using Azure Dev-Ops. The link to the Agile Project and the Scrum Board.

Fig. 1. Project Flow

The diagram depicted above illustrates the project approach we've embraced, comprising four main phases: Data Collection & Cleansing, Data Analysis, Result Analysis, and Deployment. Below, we outline the tools utilized in this project.

A. Data Collection & Cleaning Phase

a) Data Sources:

Utilizing Kaggle as the initial dataset source, we supplemented our data collection by sourcing additional information from various web sources, resulting in a total of nine different tables used for analysis. Kaggle is a popular platform for data science and machine learning enthusiasts, offering a diverse range of datasets, competitions, and collaborative projects. It provides an interactive environment where users can explore data, build models, and share insights with a global community of data scientists.

b) Data Transformation:

Our initial dataset from Kaggle comprised primary fact tables containing match and player performance data. To enrich our dataset, we constructed 9 dimension tables based on unique values extracted from the fact tables. We addressed NULL values by sourcing supplementary data from open sources and ensuring data consistency by correcting misspelled entries. Normalization techniques were applied to optimize data structure, transposing column-oriented data (e.g., seasons table) into row-oriented tables (e.g., seasons_team_history). Establishing primary key-foreign key relationships ensured database integrity, facilitating efficient querying. Additionally, string data was replaced with numerical keys to enhance query performance and efficiency in our football analytics workflows. Through these data management and transformation strategies, we aimed to build a robust and comprehensive dataset for football analytics, enabling sophisticated analysis and insight generation across various dimensions of the sport.

B. Data Analysis Phase

1. **Cloud Tools Work Flow**

- **AWS S3(SIMPLE STORAGE SERVICE)**
  Amazon S3, known as Simple Storage Service allows users to store and access amounts of data from, the internet. It is a cloud-based storage service used for web hosting, content delivery, data lakes, and analytics.

- **SNOWFLAKE**
  For our project needs we rely on Snowflake, a cloud-based data warehousing solution to effectively handle and analyze datasets. Snowflake architecture supports storage and computing capabilities that enable us to execute queries and analytical tasks smoothly. The key benefits of utilizing Snowflake in our project are;

- **Scalability**; Snowflake's flexible structure adjusts to meet varying workloads ensuring performance for data analysis and reporting.

- **Concurrency**; By leveraging Snowflake multiple users can simultaneously query the data warehouse without impacting performance fostering collaborative data analysis efforts.

- **Data Sharing**; With its integrated data sharing features Snowflake facilitates easy and secure sharing of datasets, among teams and organizations promoting collaboration and informed decision making.

- **Native Integration**; Snowflake seamlessly integrates with popular analytics tools and platforms allowing us to make the most of existing workflows and tools for visualizing data and generating reports.

2. **Working with SQL database**

- **MySQL Workbench**
  MySQL Workbench is an all-inclusive graphical tool for database development, administration, and design. Users may develop and run SQL queries, establish database schemas, and view data relationships with its user-friendly interface for administering MySQL databases.

Fig. 2. SQL Tables

3. **Working with No SQL database**

- **Mongo DB**
  MongoDB This document-oriented, open-source database is quite compatible with Binary JSON (BSON). Aggregation Pipelines are what we use for analytics.

- **PyMongo**
  PyMongo is a native Python driver that we have used to link Python to MongoDB and enable us to communicate with the MongoDB database.

C. Result Analysis Phase

- **Query Execution**:
  The queries are performed on different platforms and databases, like MongoDB and Snowflake. The query results are depicted and explained further in the Query Results section.

DEPLOYMENT PHASE

AWS cloud setup

AWS S3

In the AWS (Amazon Web Services) ecosystem, Amazon S3 (Simple Storage Service) is a flexible and extensively used option for creating data lakes. S3 is a data lake storage solution that enables us to store enormous volumes of unstructured and raw data at scale. This covers a wide range of data kinds, including text, pictures, videos, and more. Because of its design, which guarantees scalability,

durability, and accessibility, S3 is a fundamental part of building data lakes that meet the changing demands of analytics and data storage.

A. Creating S3 Bucket

The first step in creating an S3 bucket is to access the S3 service via the AWS Management Console. You can then start the process of making a new bucket from there. Selecting the AWS region in which the data will be kept, giving the bucket a globally unique name, and setting up extra parameters like versioning, logging, and access control are important steps in this process. Your data is centrally stored in the bucket that is formed once these parameters are set.

Fig. 3. S3 Bucket

B. Folder Upload

It is simple to create folders to organize data inside the S3 bucket once it has been created. By selecting the ”Create folder” option, you can add a folder to the bucket and navigate to the desired place. Folders help organize and classify data inside the bucket, which makes data administration more effective. Once the folder was created, we used the ”Upload” feature to pick the folder and upload the CSV file of our project. This stage laid the foundation for further data processing and analysis operations by ensuring that the project data is arranged logically and easily accessible within the S3 data lake. To summarize, the utilization of Amazon S3 as a data lake offered a sturdy framework for the storage and administration of various datasets, and the establishment of buckets and folders expedited the arrangement and retrieval of project-specific data from the S3 environment.

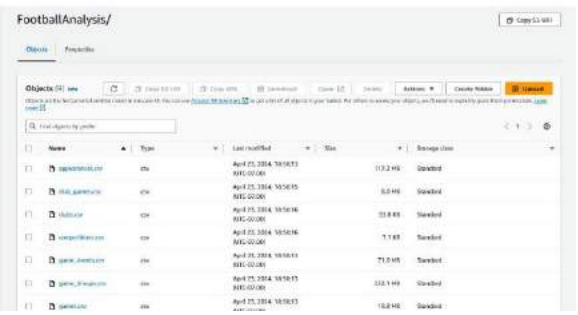


Fig. 4. Folder Upload on Snowflake

C. About S3 Bucket Policy

A bucket policy in AWS S3 is a collection of guidelines established by the bucket owner to manage rights for accessing the objects (files) kept in an S3 bucket. This policy is attached to the S3 bucket and written in JSON (JavaScript Object Notation). A key component of controlling security and access for objects stored in S3 is bucket policies, which let you define who may access your data and what actions they can take.

Fig. 5. S3 Bucket Policy

ETL using SnowFlake

Executing Extract, Transform, and Load (ETL) procedures were part of the procedure. Data was transferred into the data warehouse through the ETL procedure after being extracted from outside sources. After the raw data were transformed, a STAR Schema with one fact table and two dimension tables was produced. The converted data was then put into a consumption zone schema, giving analysts an organized setting in which to run queries and produce useful insights. Within the particular project, Snowflake, a data warehousing solution, was essential. Three schemas were developed for the ” **football DB**” database, which made the converted data easier to retrieve and arrange.

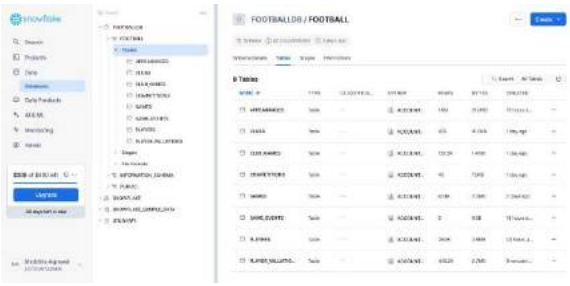


Fig. 6. Football database tables

Using Snowflake as the data warehousing platform gave the project more power because it offered a flexible and scalable environment for data management and querying. By ensuring that unstructured data was transformed into a structured STAR Schema, the ETL procedure enhanced the effectiveness of ensuing analytical queries. Information

retrieval and storage were made even easier with the establishment of the ”footballDB” database, which has separate schemas. This method not only followed industry standards for data warehousing but also set up analysts to gain valuable insights from the carefully selected data in the consumption zone schema. The project’s goals were mostly met because of Snowflake’s ability to manage multi-cluster processing and effective querying, which offered a reliable infrastructure for data transformation, storage, and analysis.

Fig. 7. Schemas

We set up the ”**Compute WH**” data warehouse in a small size, optimizing its design to improve scalability and performance. A scaling policy that mandates the use of a minimum of one cluster and permits the system to dynamically scale up to a maximum of two clusters depending on workload demands is included in the implementation. This strategy makes sure the data warehouse stays effective and adaptable to changing user needs and query complexity. We optimize query execution times by using a standard scaling approach, taking into account that various queries could need varying amounts of processing power.

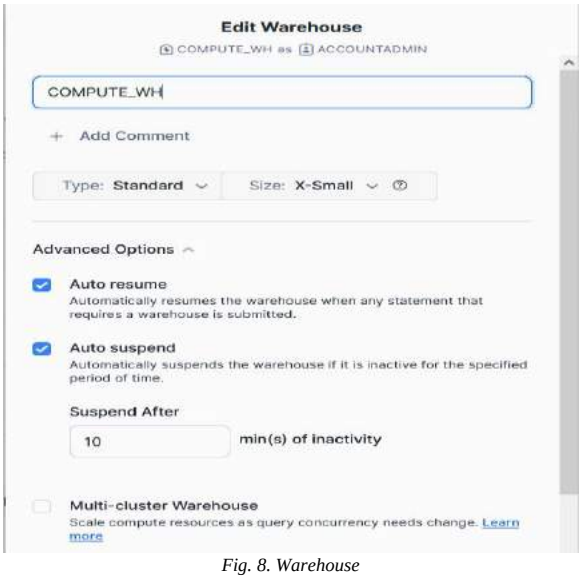


Fig. 8. Warehouse

VISUALIZATION DASHBOARD

Power Bi:

Produced visually appealing graphics for comparing different attributes to derive data-driven conclusions that could benefit overall team performance and management decisions.

MongoDB Atlas:

Designed a visualization dashboard within MongoDB Atlas, utilizing its charts feature to offer a comprehensive overview of the analysis results.

ER DIAGRAM

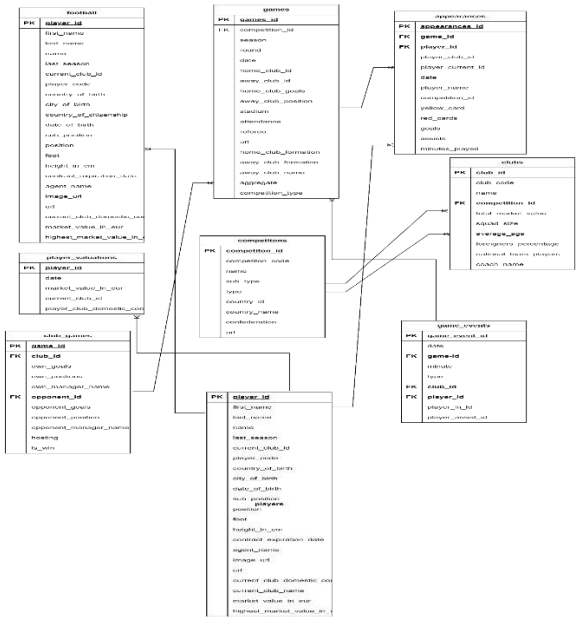


Fig. 9. ER Diagram for the Database Setup

QUERY RESULTS

A. Analyzing with Snowflake

a) Snowflake Query 1:

query retrieves player details along with the total minutes played by each player. It groups the data by player and then sorts the results based on the total minutes played, with the players who played the most minutes appearing at the top.

```
#query retrieves player details along with the total minutes played by each player. It groups the data by

query = """SELECT p.player_id, p.first_name, p.last_name, SUM(a.minutes_played)
AS total_minutes_played FROM appearances a JOIN players p ON a.player_id = p.player_id
GROUP BY p.player_id, p.first_name, p.last_name ORDER BY total_minutes_played DESC
"""

# Execute the query and store results in a DataFrame
cur = conn.cursor()
cur.execute(query)
columns = [desc[0] for desc in cur.description]
rows = cur.fetchall()
df = pd.DataFrame(rows, columns=columns)

df
```

Fig. 10. Snowflake Query 1

	PLAYER_ID	FIRST_NAME	LAST_NAME	TOTAL_MINUTES_PLAYED
0	38253	Robert	Lewandowski	47269
1	28003	Lionel	Messi	44380
2	59377	David	de Gea	44129
3	59561	Dani	Parejo	43830
4	108390	Thibaut	Courtois	43579
...	...	...	...	...
22640	479498	None	Renato Vischi	1
22641	554395	Vasilios	Pavlidis	1
22642	816572	Gilberto	Batista	1
22643	743496	Kévin	Danois	1
22644	373379	Giuseppe	Borello	1

Fig. 11. Query Result for Snowflake Query 1

b) Snowflake Query 2:

This SQL query retrieves aggregated data about each club's performance in matches, including the total number of matches played, total wins, and total losses.

```
query = """
    SELECT c.name AS club_name,
           COUNT(*) AS total_matches_played,
           SUM(CASE WHEN g.home_club_goals > g.away_club_goals THEN 1 ELSE 0 END) AS total_wins,
           SUM(CASE WHEN g.home_club_goals < g.away_club_goals THEN 1 ELSE 0 END) AS total_losses
    FROM games g
    JOIN clubs c ON g.home_club_id = c.club_id OR g.away_club_id = c.club_id
    GROUP BY c.name
    ORDER BY total_wins DESC
"""

# Execute the query and store results in a DataFrame
cur = conn.cursor()
cur.execute(query)
columns = [desc[0] for desc in cur.description]
rows = cur.fetchall()
df = pd.DataFrame(rows, columns=columns)

df
```

Fig. 12. Snowflake Query 2

	CLUB_NAME	TOTAL_MATCHES_PLAYED	TOTAL_WINS	TOTAL_LOSSES
0	Sevilla Fútbol Club S.A.D.	682	329	183
1	Futbol Club Barcelona	676	321	241
2	Real Madrid Club de Fútbol	676	319	245
3	Manchester City Football Club	641	317	228
4	Juventus Football Club	626	315	204
...	...	...	...	...
420	Pendikspor	32	13	10
421	Giresunespor	36	12	15
422	Le Havre Athletic Club	29	11	8
423	Hvidovre Idrætsforening	51	11	34
424	LNZ Cherkasy	31	11	12

Fig. 13. Query Result for Snowflake Query 2

B. Analyzing with MySQL

a) MySQL Query 1:

This SQL query calculates the average market value of players participating in each competition. It joins the player valuation, players, games, and competitions tables to retrieve information about player valuations, player details, game details, and competition details. The query groups the results by competition ID and name, presenting the average market value of players for each competition, and providing insights into the relative value of players competing in different competitions.

```
79 •  SELECT
80        c.competition_id,
81        c.name AS competition_name,
82        AVG(pv.market_value_in_eur) AS average_market_value
83    FROM
84        player_valuations pv
85    JOIN
86        players p ON pv.player_id = p.player_id
87    JOIN
88        games g ON pv.current_club_id = g.home_club_id OR pv.current_club_id = g.away_club_id
89    JOIN
90        competitions c ON g.competition_id = c.competition_id
91    GROUP BY
92        c.competition_id, c.name;
```

Fig. 14. MySQL Query 1

Result Grid | Filter Rows: | Export:

competition_id	competition_name	average_market_value
L1	bundesliga	1948960.8802
FR1	ligue-1	3160566.1298
RU1	premier-liga	2491793.8931
GB1	premier-league	2857391.3043
BE1	jupiler-pro-league	1071323.5294

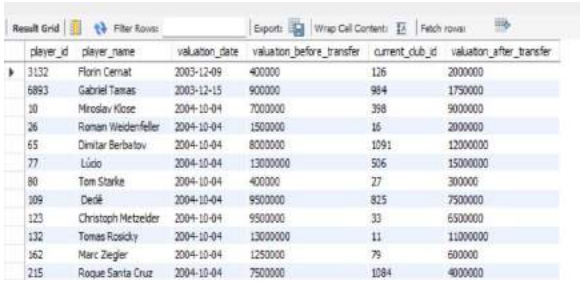
Fig. 15. Query Result for MySQL Query 1

b) MySQL Query 2:

This SQL query retrieves information about player valuations before and after potential transfers. It selects the player ID, player name (concatenated first name and last name), valuation date, valuation before transfer (market value in Euros), current club ID, and valuation after transfer (market value in Euros) for each player. The valuation after the transfer is obtained by selecting the next valuation date after the current valuation date for the same player, ordered by date in ascending order, and limiting the result to 1. This query provides insights into the potential changes in player valuations Over time about potential transfer.

```
64 •  SELECT
65        p.player_id,
66        CONCAT(p.first_name, ' ', p.last_name) AS player_name,
67        pv.date AS valuation_date,
68        pv.market_value_in_eur AS valuation_before_transfer,
69        p.current_club_id AS current_club_id,
70        (SELECT pv2.market_value_in_eur FROM player_valuations pv2
71        WHERE pv2.player_id = pv.player_id AND pv2.date > pv.date ORDER BY pv2.date ASC LIMIT 1)
72        AS valuation_after_transfer
73    FROM
74        player_valuations pv
75    JOIN
76        players p ON pv.player_id = p.player_id;
```

Fig. 16. MySQL Query 2



player_id	player_name	valuation_date	valuation_before_transfer	current_club_id	valuation_after_transfer
3132	Florin Cernat	2003-12-09	400000	126	2000000
6893	Gabriel Tamas	2003-12-15	900000	984	1750000
10	Miroslav Klose	2004-10-04	7000000	398	9000000
26	Roman Weidenfeller	2004-10-04	1500000	16	2000000
65	Dimitar Berbatov	2004-10-04	8000000	1091	12000000
77	Lúcio	2004-10-04	13000000	506	15000000
80	Tom Starke	2004-10-04	400000	27	300000
109	Dedê	2004-10-04	9500000	825	7500000
123	Christoph Metzelder	2004-10-04	9500000	33	6500000
132	Tomas Rosicky	2004-10-04	13000000	11	11000000
162	Marc Ziegler	2004-10-04	1250000	79	600000
215	Roque Santa Cruz	2004-10-04	7500000	1084	4000000

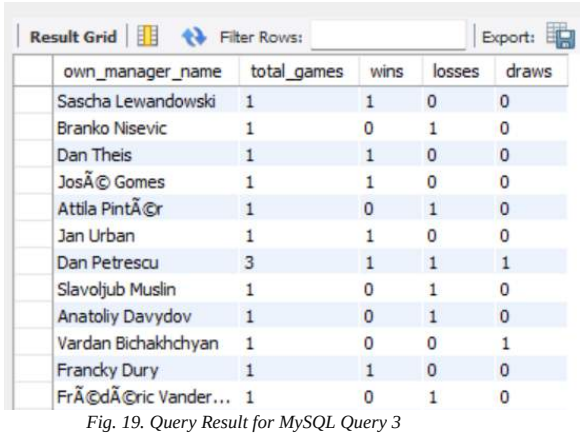
Fig. 17. Query Result for MySQL Query 1

c) MySQL Query 3:

This SQL query provides a summary of the performance of club managers based on the games they have overseen. It counts the total number of games managed by each manager, as well as the number of wins, losses, and draws achieved under their management. This summary allows for an analysis of the managerial performance in terms of game outcomes.

```
SELECT
    cg.own_manager_name,
    COUNT(*) AS total_games,
    SUM(CASE WHEN cg.own_goals > cg.opponent_goals THEN 1 ELSE 0 END) AS wins,
    SUM(CASE WHEN cg.own_goals < cg.opponent_goals THEN 1 ELSE 0 END) AS losses,
    SUM(CASE WHEN cg.own_goals = cg.opponent_goals THEN 1 ELSE 0 END) AS draws
FROM
    club_games cg
GROUP BY
    cg.own_manager_name;
```

Fig. 18. MySQL Query 3



own_manager_name	total_games	wins	losses	draws
Sascha Lewandowski	1	1	0	0
Branko Nisevic	1	0	1	0
Dan Theis	1	1	0	0
JosÃ© Gomes	1	1	0	0
Attila PintÃ©r	1	0	1	0
Jan Urban	1	1	0	0
Dan Petrescu	3	1	1	1
Slavoljub Muslin	1	0	1	0
Anatoliy Davydov	1	0	1	0
Vardan Bichakhchyan	1	0	0	1
Francky Dury	1	1	0	0
FrÃ©dÃ©ric Vander...	1	0	1	0

Fig. 19. Query Result for MySQL Query 3

d) MySQL Query 4:

This SQL query provides a summary of each club's performance in terms of goals scored and conceded, categorized by whether the club played at home or away. It calculates the total goals scored and conceded in home and away games for each club, presenting this information alongside the club names. This summary offers insights into the club's offensive and defensive performance in different match settings.

```
SELECT
    c.name AS club_name,
    SUM(CASE WHEN cg.hosting = 'home' THEN cg.own_goals ELSE 0 END) AS goals_scored_home,
    SUM(CASE WHEN cg.hosting = 'away' THEN cg.own_goals ELSE 0 END) AS goals_scored_away,
    SUM(CASE WHEN cg.hosting = 'home' THEN cg.opponent_goals ELSE 0 END) AS goals_conceded_home,
    SUM(CASE WHEN cg.hosting = 'away' THEN cg.opponent_goals ELSE 0 END) AS goals_conceded_away
FROM
    club_games cg
JOIN
    clubs c ON cg.club_id = c.club_id
GROUP BY
    c.name;
```

Fig. 20. MySQL Query 4

club_name	goals_scored_home	goals_scored_away	goals_conceded_home	goals_conceded_away
SK Beveren	4	0	1	0
Saint Johnstone Football Club	1	0	1	0
AcadÃ©mica Coimbra	0	0	3	0
Roda JC Kerkrade	1	0	6	0

Fig. 21. Query Result for MySQL Query 1

C. Analyzing with MongoDB

a) MongoDB Query 1:

To identify players who perform best in the final minutes of a match, we analyzed players' goal contributions (goals and assists) specifically in the last 15 minutes of a match. This query filters appearances where the player played at least 75 minutes, calculates their total goal contributions (goals + assists), and then filters appearances where the player played less than or equal to 90 minutes (indicating they played in the final 15 minutes). It then groups the players, sums their total goal contributions, sorts them by total goal contributions in descending order, and finally limits the result to the top 10 players.

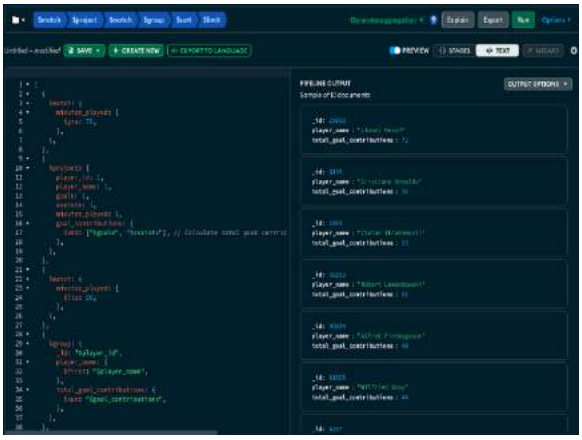


Fig. 22. MongoDB Query 1 & Result

b) MongoDB Query 2:

To find the top 20 players by red cards from the given data in MongoDB, we used the aggregation

framework to group by the player and sum up the number of red cards they have received, then sort the results in descending order by the total number of red cards, and finally limit the results to the top 20 players. This query will give you the top 20 players by the number of red cards they have received.

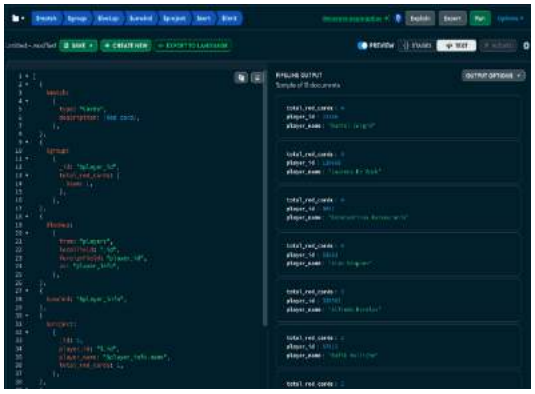


Fig. 23. MongoDB Query 2 & Result

c) MongoDB Query 3:

To find the card distribution (number of cards) versus time intervals from the provided data, we used the "game_events" table to aggregate the number of cards (yellow and red) within specified time intervals

In this query:

1. We first project the minute and type (card type) fields.
2. Then, we use the $group stage to group the events by time intervals. We define time intervals of 15 minutes each (0-15, 16-30, 31-45, 46-60, 61-75, and 76-90) using the $switch operator.
3. Within each time interval, we calculate the total number of cards using the $sum accumulator and the $cond operator to count only the card events.
4. Finally, we sort the results by time intervals.

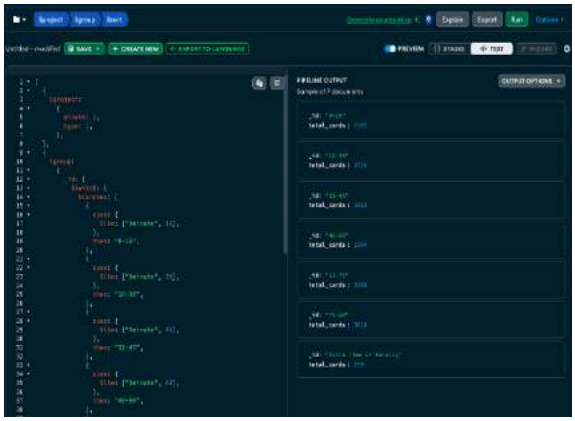


Fig. 24. MongoDB Query 3 & Result

d) DB Connectivity / API calls PyMongo Connection:

A connection has been established between Python and MongoDB using Pymongo for running the queries in Python. The screenshots are provided in the Query results section.

Fig. 25. DB Connectivity

DATA VISUALIZATION

A. Matplotlib Visualization

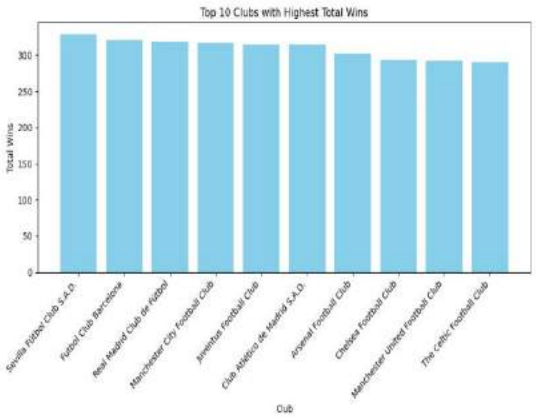


Fig. 26. Matplotlib Visualization

B. Visualization of MongoDB

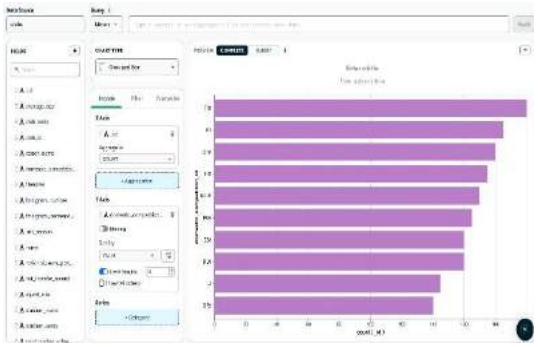


Fig. 27. MongoDB Visualization

C. Visualization Dashboard for Power BI

a. Market Value by Players

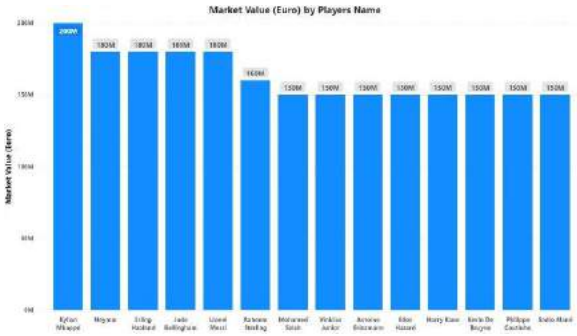


Fig. 28. Market value by player name

b. Number of games by stadium and competition type

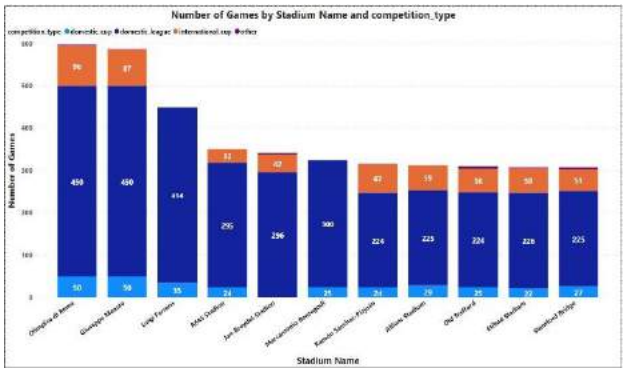


Fig. 29. Visualization of the Number of games by stadium

c) Height Distribution of Players

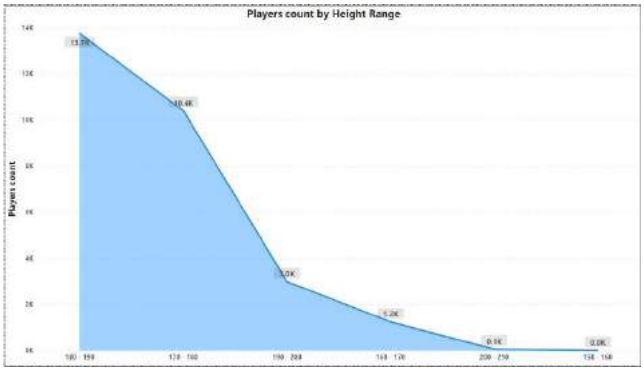


Fig. 30. Height Distribution of Players

d) Top Goal Scorers

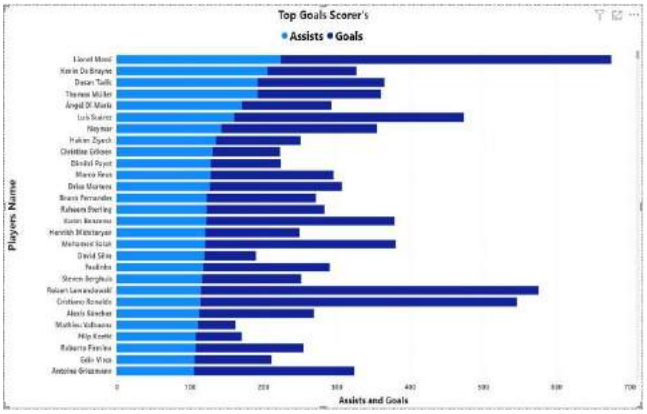
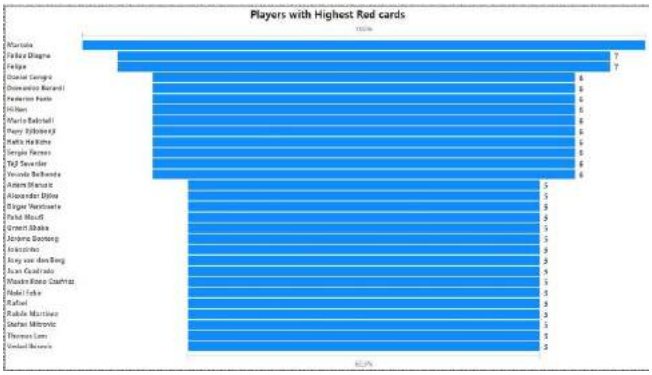


Fig. 31. Top Goal Scorers

e) Players with the highest yellow and red cards



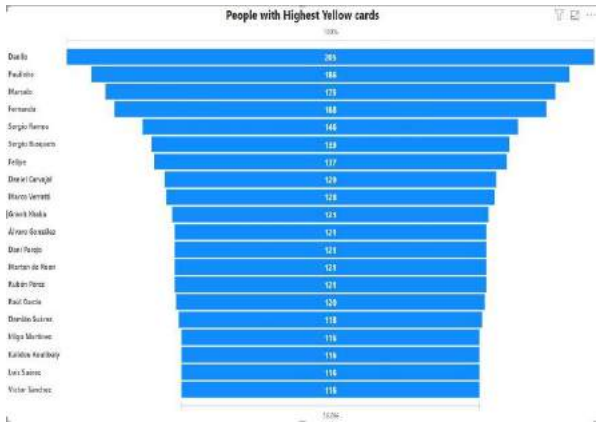


Fig. 32. Highest red and yellow cards

f) Country diversity of players

Fig. 33. Country Diversity of Players

g) Players Profile Dashboard

Fig. 34. Players Profile Dashboard

h) Messi's Profile

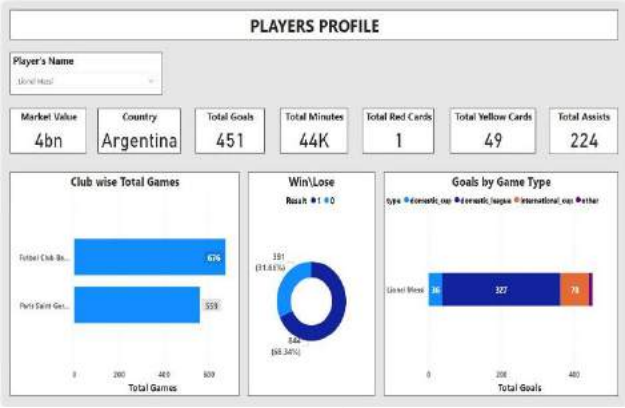


Fig. 35. Messi's Profile Dashboard

KEY FINDINGS

1. **Player Valuation Dynamics:**
   We observed fluctuations in player valuations over time, indicating the potential impact of transfers on market values. This insight sheds light on the valuation trends within the football industry and can inform clubs' transfer strategies.

2. **Managerial Performance Evaluation**:
   By analyzing game outcomes under different managers, we identified trends in managerial effectiveness based on wins, losses, and draws. This analysis provides clubs with valuable insights into the performance of their coaching staff.

3. **Home vs. Away Performance:**
   We examined clubs' performance in terms of goals scored and conceded at home versus away games. This comparison revealed patterns in teams' strengths and weaknesses across different match settings, aiding in strategic planning for upcoming fixtures.

4. **Player Performance Metrics:**
   We aggregated player statistics, including goals scored, assists provided, and disciplinary actions, to assess individual performance across various matches. These metrics offer a comprehensive view of players' contributions to their teams and help identify standout performers.

5. **Late-Game Impact Players:**
   By analyzing goal contributions in the final 15 minutes of matches, we identified players who excel in late-game situations. Understanding the performance of these players during critical moments provides teams with valuable insights for tactical substitutions and strategic decision-making.

TECHNICAL DIFFICULTIES

- One of the primary challenges in the project dealing with data quality issues, such as missing values, inconsistent formatting, and errors in data collection. To address this challenge, you can implement data cleaning and preprocessing techniques, such as imputation for missing values, standardization of data formats, and outlier detection to identify and correct erroneous data points.

- Football data often comes from multiple sources, including match statistics databases, tracking systems, and external APIs. Integrating data from diverse sources can be challenging due to differences in data formats, schemas, and APIs. Consider using data integration tools and techniques, such as **ETL (Extract, Transform, Load)** processes, API wrappers, and data normalization methods, to streamline the integration of disparate data sources into a unified dataset.

- While setting up the AWS environment for our football analytics project, we encountered challenges related to adjusting default data types assigned by the system when creating jobs. Additionally, we faced difficulties in configuring VPC endpoints and specifying file paths during the data-loading process from buckets into crawlers. Ensuring the correct assignment of IAM roles while running ETL (Extract, Transform, Load) jobs was another hurdle we encountered.

TEAM MEMBERS AND THEIR RESPONSIBILITIES

TABLE I. RESPONSIBILITY BREAKDOWN

ID	Members' Names	Responsibilities
017441359	Kush Bindal	Setting up the database in NoSQL (MongoDB (Compass, Atlas, and Mongosh)), Pymongo, Final Data Analysis, and Data Visualization, and setting up data insights.
017552795	Shobhita Agrawal	Understanding Tools, Amazon S3, Intermediate Status Report, Snowflake, Python, Pandas, Matplotlib, Deployment Phase (AWS), Final Report.
017536623	Darpankumar Jiyani	Data cleaning, intermediate status report, Setting up the Database in SQL, Beta Testing, and Preparing Presentation Slides, Final Report
017507100	Dhruv Patel	Data cleaning, data visualization (Power BI), setting up data insights, project proposal, finding the dataset, analysis stage, and evaluation of results.

SIGNIFICANT TO THE REAL WORLD

- The significance of football analytics extends to the real world, driven by the sport's immense global impact and the strategic decisions it entails for teams and stakeholders. With football being a multi-billion-dollar industry, clubs and investors alike benefit from data-driven insights to maximize returns and make informed decisions.

- Club Management Club owners and executives can leverage detailed player and team performance analyses to assess player values and return on investment. By understanding individual player strengths and weaknesses, clubs can make strategic decisions in player recruitment, contract negotiations, and squad management, ultimately optimizing team performance and financial outcomes.

- Coaching Staff Coaches benefit from analytics by strategically deploying players based on their performance metrics and situational suitability. By categorizing players and understanding their strengths under specific match conditions, coaches can maximize player potential and tactical effectiveness, leading to improved team performance on the field.

- Player Development for players and football analytics provide valuable insights into performance metrics, highlighting areas for improvement and skill development. By analyzing records, players can identify weaknesses, enhance their skills, and optimize their performance on the field. This not only enhances individual career prospects but also contributes to increasing their market value and international opportunities.

- Fan Engagement Football analytics enhance fan engagement by providing meaningful insights that go beyond basic statistics. By visualizing complex data and game scenarios, fans can deepen their understanding and appreciation of the sport. This engagement not only fosters a stronger connection between fans and the game but also contributes to the overall growth and commercial success of football as a global phenomenon.
  .

CONCLUSION

The football analytics project has shown the effectiveness of a data-driven approach in comprehending and enhancing football performance. We have discovered factors that have a substantial impact on individual and team performance by studying a range of Football statistics, which has given us invaluable insights into the game of Football. Our analysis has underscored the importance of several critical factors in football, such as goal-scoring rates, defensive efficiency, passing accuracy, and positional play. Additionally, we have identified supplementary variables that influence match results, such as the quality of the opposition, pitch conditions, stadium atmosphere, and team composition. Coaches, players, and teams stand to gain valuable insights from this project, enabling them to enhance their performance and achieve better outcomes on the field. By harnessing the insights gleaned from data and analytics, football teams can make informed decisions and optimize their strategies for success.

KEY LEARNINGS

- Implementing analytics with MySQL, Power BI, Snowflake, MongoDB, Power Query, and AWS

S3 provides insights into the diverse capabilities of these tools for data storage, analysis, and visualization in football analytics projects. Leveraging MySQL and MongoDB for data storage highlighted the flexibility of SQL and NoSQL databases in managing structured and unstructured football data, enabling efficient querying and performance optimization through columnar storage.

- Utilizing Power BI and Power Query facilitated advanced data visualization and transformation, enhancing the presentation and interpretation of football analytics results for stakeholders. Experimenting with AWS S3 showcased the scalability and computational power of cloud platforms, allowing for on-demand resources and efficient execution of complex queries in football analytics workflows.

- Following Agile methodologies ensured effective project planning and team collaboration throughout the research and development phases of the football analytics project, promoting productivity and rapid knowledge sharing among team members. These experiences underscore the importance of utilizing diverse technologies and collaborative methodologies to drive successful football analytics initiatives, enabling teams to navigate challenges effectively and deliver impactful insights to stakeholders.

INNOVATIONS

In our football analytics project, we adopted a detailed bottom-up approach, starting with granular data points and working towards aggregated statistics, akin to the methodology used in cricket analytics. Beginning with the core fact table, which records individual game events such as passes, shots, and tackles, we meticulously analyzed minute factors like player positions, pass accuracy, speed, and field conditions. By delving into these specific data elements, we aimed to uncover nuanced insights that contribute to overall team performance and strategy. This approach allows for a comprehensive analysis of player dynamics, tactical formations, and match scenarios, enabling us to draw meaningful conclusions about individual and team performance. Moreover, our project framework is designed to accommodate women's football data, facilitating comparative analyses between men's and women's football leagues. By embracing this methodological approach and inclusivity in dataset scope, we aim to enrich football analytics and contribute to a deeper understanding of the sport's dynamics across different levels and genders.

FUTURE SCOPE

- Expand Dataset Inclusion: Incorporate women's football data alongside men's football data to enable comparative analysis and gain insights into gender-specific performance trends.

- Explore NOSQL Database Variants: Research and analyze other NOSQL databases such as column-based (e.g., Cassandra), key-value pairs (e.g., Aerospike), and graph databases (e.g., OrientDB and Neo4j) to evaluate their performance in football analytics applications and identify their respective pros and cons through a comparative study.

- Evaluate Player Return on Investment (ROI): Integrate auction data to calculate the ROI for each player, providing valuable insights into player valuation and performance relative to investment.

- Develop a Composite Performance Score: Create a composite performance score derived from multiple metrics to serve as a comprehensive evaluation tool for fantasy football players, facilitating informed player selections and strategy decisions.

- Optimize Query Performance: Enhance query performance by leveraging join operations instead of subqueries, improving the efficiency and speed of data retrieval for advanced football analytics and reporting.

- These objectives aim to enhance the depth and breadth of our football analytics project, enabling comprehensive analysis, data-driven decision-making, and performance evaluation across various dimensions of the sport.

USE OF GRAMMARLY

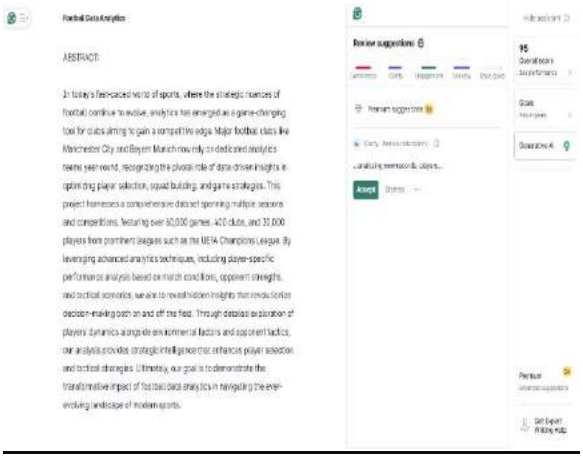


Fig. 36. Generated Grammarly Report

Version Control

We have used GitHub where we have upload all our documents and reports. Below We have attached the link:

https://github.com/labcad11/FootballDataAnalytics

REFERENCES

1. Carling, C., Williams, A. M., & Reilly, T. (2005). Handbook of soccer match analysis: A systematic approach to improving performance. Routledge.

2. Lago-Peñas, C., Lago-Ballesteros, J., Dellal, A., & Gómez, M. (2010). Game-related statistics discriminated against winning, drawing, and losing teams from the Spanish soccer league. Journal of Sports Science & Medicine, 9(2), 288-293.

3. Liu, H., Hopkins, W. G., Gómez, M. A., & Molinuevo, S. J. (2013). Inter-operator reliability of live football match statistics from OPTA Sports data. International Journal of Performance Analysis in Sport, 13(3), 803-821.

4. Kuper, S., & Szymanski, S. (2009). Soccernomics: Why England Loses, Why Germany and Brazil Win, and Why the US, Japan, Australia, Turkey--and Even Iraq--Are Destined to Become the Kings of the World's Most Popular Sport. Nation Books.

5. Anderson, C., & Sally, D. (2013). The Numbers Game: Why Everything You Know About Football Is Wrong. Penguin UK.

6. Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions Speak Louder than Goals: Valuing Player Actions in Soccer. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 7240-7247).
7. Liu, H., Gomez, M. A., Gonçalves, B., Sampaio, J., & Sáiz, S. (2016). Technical performance and match-to-match variation in elite football teams. Journal of Sports Sciences, 34(6), 509-518.

APPENDIX

CRITERIA	PTS	EXPLANATION
Presentation Skills Includes Time Management	5Pts	
Code Walkthrough	3Pts	The whole code and SQL scripts are uploaded on GitHub. The link to the GitHub document is uploaded on Canvas.
Discussion/ Q&A	4Pts	
Demo	5Pts	
Version Control Use of Git / GitHub or equivalent; must be publicly accessible	3Pts	The link to our document is the following: GitHub Document
Significance to the real world	5Pts	Most football analytical projects do not go beyond conventional statistical models. This is why this project tries to establish insights based on player-to-player match-ups which are much more useful in real-world scenarios.
Lessons learned Included in the report and presentation? How substantial and unique are they?	5Pts	The key learnings are included in both the report and presentation.
Innovation	5Pts	The innovation part is mentioned in the report as a separate section.
Teamwork	5Pts	The complete Team Responsibility breakdown is provided in the report in tabular form
Technical difficulty	4Pts	Mentioned in the report as a separate section.
Practiced pair programming. See https://en.wikipedia.org/wiki/Pair_programmingLinks to an external site.	2Pts	The whole team has worked on the project both physically and virtually using Zoom meetings. The MOM for all the virtual meetings is provided on Canvas.
Practiced agile / scrum (1-week sprints). Submit evidence on Canvas - meeting minutes, other artifacts	3Pts	Working on the project is done in sprints, and the tasks and issues are created using Azure DevOps and the Scrum Boards are created for the project. The link to the The project in Azure DevOps is mentioned in the Methodology Section.
Used Grammarly / other tools for language? Grammarly free version is sufficient; you can use other tools as well. Submit a report screenshot on Canvas.	2Pts	Grammarly Report has been generated for the Project Report and has been uploaded on canvas
Slides	5Pts	The presentation slides are submitted on canvas
Report Format, completeness, language, plagiarism, whether Turnitin could process it (no unnecessary screenshots), etc.	7Pts	The report is created in IEEE format with the help of Google Docs and is provided as well on Canvas.
Used unique tools E.g.: LaTeX for writing a report (submit .tex that is not generated from another format such as .docx; generating from .lyx and similar LaTeX editor outputs is fine. Also, check out https://www.overleaf.com/LinksLinksLinksLinks to an external site.	5Pts	The report is created in IEEE format with the help of Google Docs and the presentation slides are very Interactive and visually pleasing and are made using the Prezi tool.
Performed substantial analysis using database techniques The project must include an analytics component	3Pts	Aggregation Pipelines used for analytics in MongoDB Compass. We have used

		Windows functions for ranking and aggregation in MYSQL. Visualizations are created using Power BI and MongoDB Atlas.
Used a new database or data warehouse tool not covered in the HW or class	3Pts	Amazon S3 as a data lake, Power query is used to perform ETL, for warehouse Snowflake is used.
Used appropriate data modeling techniques	5Pts	For the data modeling, we have used a combination of MySQL reverse engineering(ER diagram -the link of which is mentioned in the report) and Microsoft PowerPoint features to create diagrams for the report.
Used ETL tool	1Pts	We have used Power Query as the ETL tool for the project.
Demonstrated how Analytics support business decisions	3Pts	An interactive Dashboard in Power BI has been created which gives the analysis results, we can observe the player and Team performance analysis across different seasons which can aid in the decision-making of which player to consider for playing in different match conditions.
Used RDBMS	1Pts	For the Aggregation and Ranking, we have used the Windows function in MYSQL, for populating the database for derived data we have created different Stored Procedures, Common table expressions, and sub-queries.
Used Data warehouse The idea is to exercise as many topics from the course as possible	1Pts	Snowflake is our cloud data warehouse to run SQL queries and run additional analysis.
Includes DB Connectivity / API calls Possibly using Python	1Pts	A connection has been established between Python and MongoDB using
Used NOSQL	1Pts	Pymongo for running the queries in Python. The screenshots are provided in the query result selection.
Total Points	85Pts	