

IE 6700 CRN 19106 Data Management for Analytics Project Report

Group Number: 7

Student Names: Praneith Ranganath and Vikrant Shelke

INTRODUCTION

European football, also known as soccer, is a global phenomenon with billions of fans worldwide. The industry is driven by its players, teams, and leagues, and data is at the forefront of this world-class sport. With the advent of technology, European football has seen a significant increase in data availability, making it essential for organizations to have an analytics solution and data strategy to make sense of it all.

A well-defined European football analytics strategy is critical for teams and individuals to make data-informed decisions. The competition in this industry is intense, and with numerous teams and leagues, organizations need to make sure they are ahead of the game. European football analytics is the process of applying player performance data, team statistics, and fan behavior data to guide decisions about player recruitment, team performance, and fan engagement.

A high-performing European football analytics strategy will help answer some of the following questions:

- Which players impacted the performance of a team?
- What is the impact of player transfers on team performance?
- How are ticket, advertising and merchandise sales contributing to revenue of the team?
- How referees play a crucial role in key games?
- How home teams are benefitted by the support of their fans?

By utilizing a comprehensive analytics strategy, organizations can unlock the value in their data and make informed decisions to enhance the performance of their teams and the overall business impact.

Business Problem Introduction

European football consists of many leagues like LaLiga, English Premier League, Ligue 1 and so on which are played in their respective countries. Each league consists of teams who play matches against each other with each team having their own set of players who are managed by a manager.

The matches consist of two teams, each with 11 players and 5 substitutes. During the match, players score goals against the other team to put their team up front. Then, there are referees who make sure the game is played in a fair way and if there are incidents on the pitch where players are not fair they are shown cards depending on the intensity of the foul/incident they caused. At the end of the game, the team with the most goals wins earning three points but there are instances where both teams walk away with equal goals and earning one point each. At the end of each season the team with the most points win the league.

Problem Statement

Despite the tremendous success of the game, the number of actively available datasets in this region is extremely scarce and mostly contain redundant information without proper modelling. Hence, as a loyal football fan from a young age I have taken this opportunity to model and design a database exclusively for football fans around the world.

CONCEPTUAL DATA MODELLING

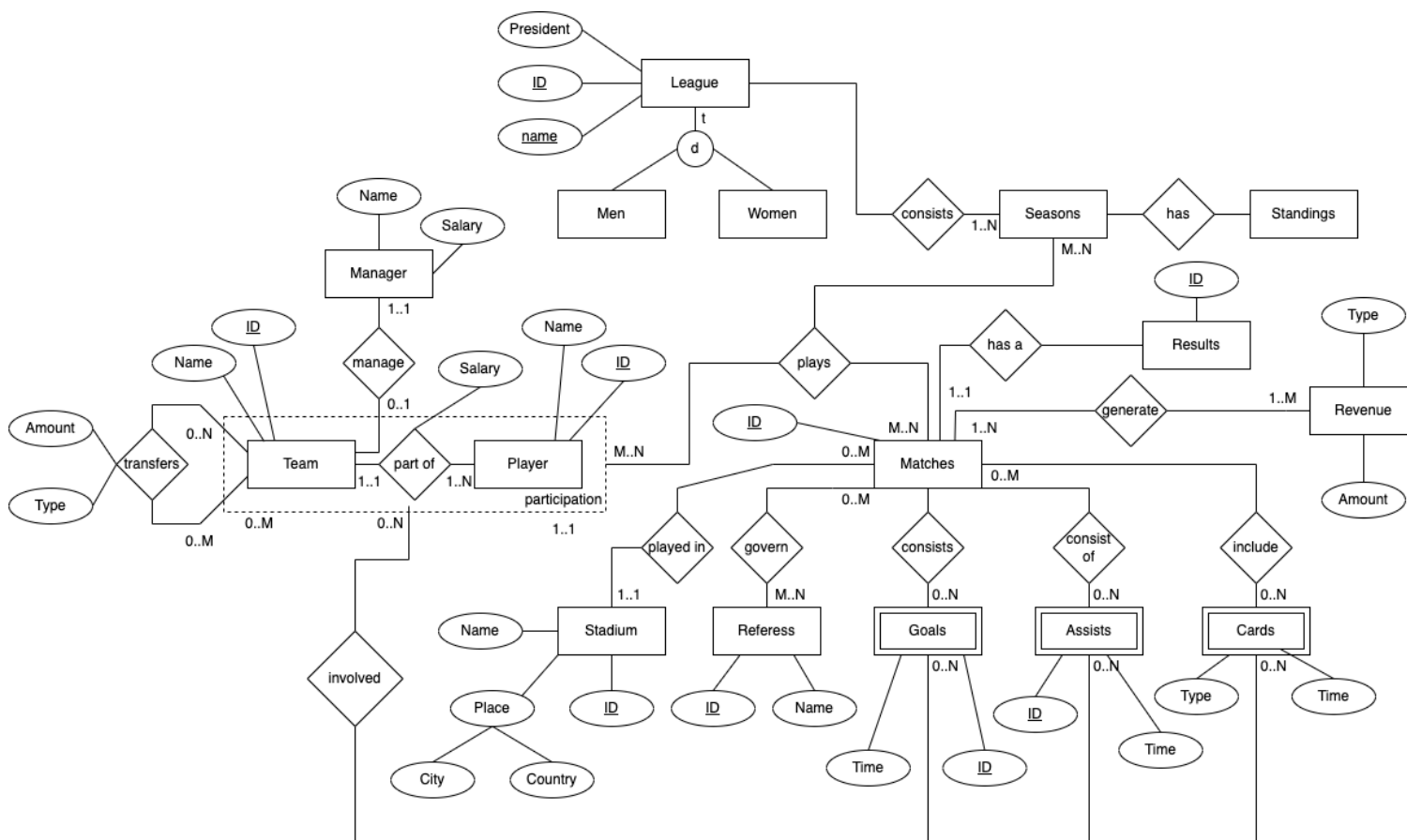
As a first step, the conceptual data model reflecting the game and structure around it is set up using EER as provided below. A detailed explanation of the model is as follows:

- There are leagues and each league has its own unique name as its identity and has a president who governs it and each league has its own set of teams who are identified by a unique name and ID.

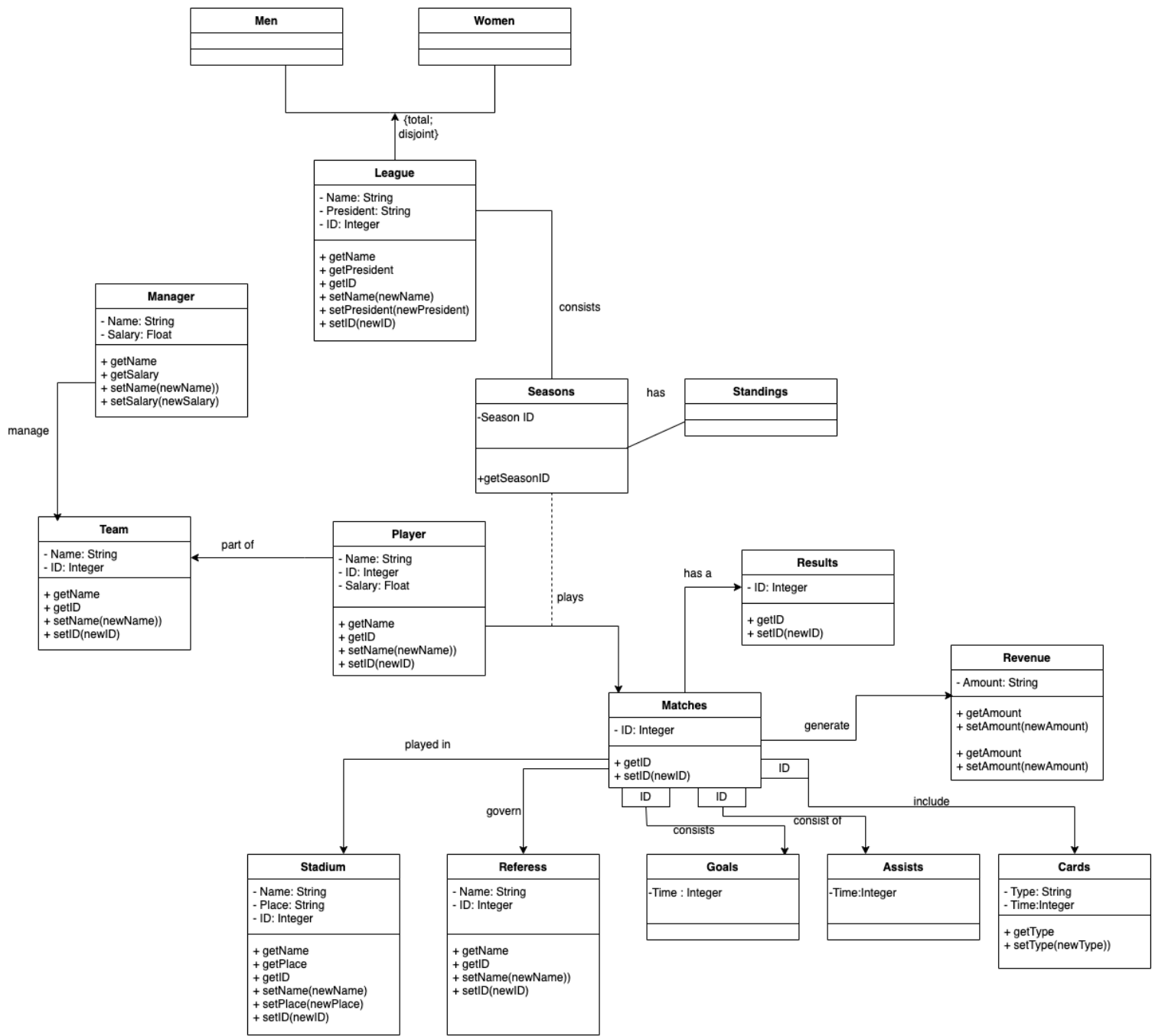
Each league is played for a period of set matches called seasons which is identified by its own unique ID. The teams and players together take part in matches against other teams and players. The matches are played in stadiums identified by their unique ID's and each team has a home stadium associated with them. The matches are governed by referees who have the right to book players for their faults using two types of cards red and yellow. Players score goals assisted by other players in the matches to put their team upfront to win the game.

- The winning team is awarded three points and the team with the most points at the end of all matches on the Table win the league for that season.
- Each match generates revenue in three forms Ticket sales, merchandise sales, Advertising and TV rights. While the home team receives the revenue generated from Ticket sales, merchandise sales and Advertising in there home stadiums TV rights revenues are collected by the league and split among the teams based on the Table at the end of the season.
- The manager and the players are paid salaries by their teams and each team can buy/sell players from other teams for a certain amount and loan players from other teams for a certain period.

The EER diagram representing the conceptual model is provided on the next page for your reference.



UML Diagram



MAPPING CONCEPTUAL DATA MODEL TO RELATIONAL MODEL

Refer to the relations below, created to map the conceptual data model to the relational model. While normalizing, the steps have been mentioned if required. The primary key(s) are underlined and the foreign key(s) are italicized. Only if the foreign keys are NULL, it has been called out. Else the foreign key is non-nullable. Similarly, all primary keys are non-nullable. All the below relations have been normalized to at least 3.5 Normal Form,

League(LID, Name, President)

- LID – Primary Key

Men(LID, Name)

- LID - Foreign Key; Referencing League

Women(LID, Name)

- LID - Foreign Key; Referencing League

Seasons(SID, LID, Year)

- SID – Primary Key
- LID - Foreign Key; Referencing League

Player(PID, TID, Name, Salary)

- PID – Primary Key
- TID - Foreign Key; Referencing Team

Plays(PID, TID, MID)

- PID - Foreign Key; Referencing Player
- TID - Foreign Key; Referencing Team
- MID – Foreign Key; Referencing Matches

Team(TID, Name)

- TID – Primary Key

Manager(MID, Name)

- MID – Primary Key

Manager_Manages(MID, TID, Year)

- MID – Foreign Key; Referencing Manager
- TID - Foreign Key; Referencing Team

Transfers(TeamID_To, TeamID_From, Type, Amount)

- TeamID_To / TeamID_From - Foreign Key; Referencing Team

Matches(MID, TeamID_Home, TeamID_Away, SID, StID)

- MID – Primary Key
- TeamID_Home / TeamID_Away - Foreign Key; Referencing Team
- SID – Foreign Key; Referencing Season
- StID - Foreign Key; Referencing Stadium

Stadium(StID, Name, City, Country)

- StID – Primary Key

Referee(RID, Name)

- RID – Primary Key

Referee_Overseas(MID, RID)

- MID – Foreign Key; Referencing Matches
- RID - Foreign Key; Referencing Referee

Goals(MID, PID, Type, Time)

- MID – Foreign Key; Referencing Matches
- PID - Foreign Key; Referencing Player

Assists(MID, PID, Time)

- MID – Foreign Key; Referencing Matches
- PID - Foreign Key; Referencing Player

Cards(MID, PID, Type, Time)

- MID – Foreign Key; Referencing Matches
- PID - Foreign Key; Referencing Player

Revenue(MID, Type, Amount)

- MID – Foreign Key; Referencing Matches

Results(MID, TeamID_Home, TeamID_Away, HomeTeam_Points, AwayTeam_Points, HomeTeam_Goals, AwayTeam_Goals)

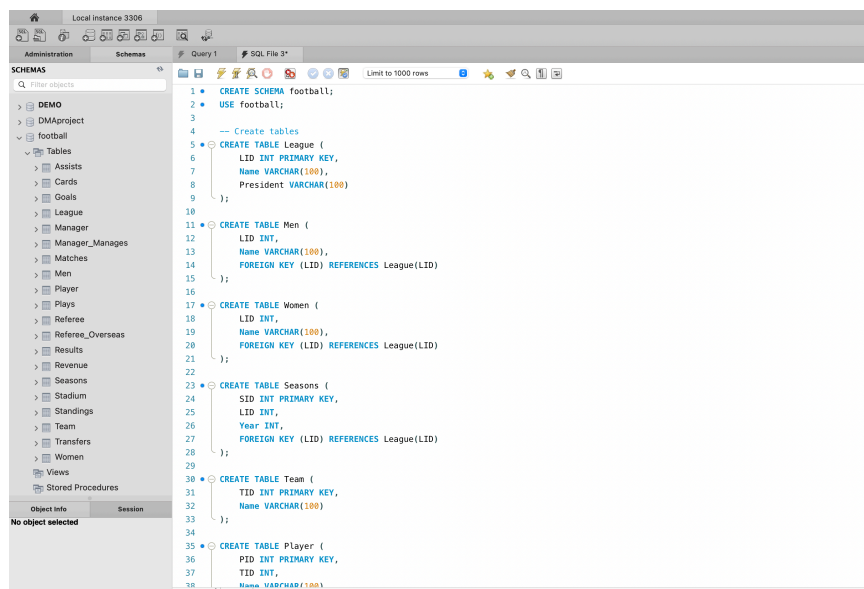
- MID – Foreign Key; Referencing Matches
- TeamID_Home / TeamID_Away - Foreign Key; Referencing Team

Standings(SID, Position, TID, Points)

- SID – Foreign Key; Referencing Season
- TID - Foreign Key; Referencing Team

Implementation of Relation Model via MySQL

The relational model has been implemented in MySQL and MySQL Workbench is used to query the database. Sample data has been populated in all datasets and sample queries have been presented below,






Query 1

Analytical Purpose: Which player has been involved in most goals i.e goals plus assists to understand how a player has impacted a team with both goals and assists

```
SELECT Player.Name as PlayerName,  
       SUM(Goal) + SUM(Assist) AS TotalGoalsInvolved  
FROM Player  
JOIN Goal ON Player.PID = Goal.PID  
GROUP BY Player.PID, Player.Name  
ORDER BY TotalGoalsInvolved DESC  
LIMIT 5;
```

Output:




Result Grid   Filter Rows: <input type="text" value="Search"/> Export: 			
	PlayerName	TotalGoalsInvolved	
▶	William Burke	24	
▶	Dax Park	20	
▶	Titan Webster	18	
▶	Sebastian Mckenzie	17	
▶	James Sharp	15	

Query 2

Analytical purpose: Revenue generated when the team loses versus when the team wins or loses helps us understand how a teams performance on the field is co-related to teams revenue generation

```
SELECT T.Name as HomeTeam , AVG(R.Amount) as AverageRevenue  
FROM Revenue R  
JOIN Matches M ON R.MID = M.MID  
JOIN Teams T ON T.TID=M.HomeTeam  
JOIN Results P on P.MID=M.MID  
WHERE P.HomeTeam_Goals > P.AwayTeam_Goals;
```

Output:

Result Grid   Filter Rows: <input type="text" value="Search"/> Export: 			
	HomeTeam	AverageRevenue	
▶	Boston	321456	
▶	NewYork	430982	
▶	Washington	458668	
▶	Philadelphia	545677	
▶	Chicago	356751	
▶	California	345213	
▶	Seattle	234570	
▶	Las Vegas	345761	
▶	Los Angeles	467894	
▶	Phoenix	498268	

```

SELECT T.Name as HomeTeam, AVG(R.Amount)

FROM Revenue R

JOIN Matches M ON R.MID = M.MID

JOIN Teams T ON T.TID=M.HomeTeam

JOIN Results P on P.MID=M.MID

WHERE P.HomeTeam_Goals < P.AwayTeam_Goals;

```

Output:

Result Grid		Filter Rows: <input type="text" value="Search"/>	Export:
Home...	AverageRevenue		
▶ Washington	342781		
Seattle	198637		
Phoenix	365209		
Philadelphia	432890		
New York	352137		
Los Angeles	398372		
Las Vegas	279452		
Chicago	278935		
California	281567		
Boston	234575		

Query 3

Analytical Purpose: Win percent of managers helps us to understand how the manager has impacted the teams performance in terms of number of wins with his current team and the previous teams and can help stakeholders determine who is a better fit as a manager to there team

```

SELECT Manager.Name as ManagerName,

COUNT(CASE WHEN Results.HomeTeam_Points > Results.AwayTeam_Points AND Matches.TeamID_Home = Results.TeamID_Home THEN 1 END) +

COUNT(CASE WHEN Results.AwayTeam_Points > Results.HomeTeam_Points AND Matches.TeamID_Away = Results.TeamID_Away THEN 1 END) AS Wins,

COUNT(*) AS TotalMatches,

(COUNT(CASE WHEN Results.HomeTeam_Points > Results.AwayTeam_Points AND Matches.TeamID_Home = Results.TeamID_Home THEN 1 END) +

COUNT(CASE WHEN Results.AwayTeam_Points > Results.HomeTeam_Points AND Matches.TeamID_Away = Results.TeamID_Away THEN 1 END)) / COUNT(*) * 100 AS WinPercentage

FROM Manager

JOIN Manager_Manages ON Manager.MID = Manager_Manages.MID

JOIN Matches ON Manager_Manages.TID = Matches.TeamID_Home OR Manager_Manages.TID = Matches.TeamID_Away

JOIN Results ON Matches.MID = Results.MID

GROUP BY Manager.Name

ORDER BY WinPercentage DESC

LIMIT 5;

```

Output:

Result Grid			Filter Rows:	Search	Export:
	ManagerName	WinPercentage			
▶	Benson Marsh	78.38			
	Braydon Salazar	72.47			
	Nicholas Gibson	70.59			
	Adam Anderson	69.67			
	Theo Wilkinson	64.76			

Query 4

Analytical Purpose: What is the average salary of male and female players this helps us. understand the discrepancy in salary between men and women in the sport

SELECT League.Name as LeagueName, MenWomen.MenAvgSalary, MenWomen.WomenAvgSalary

FROM League

JOIN

(SELECT LID, AVG(Salary) as MenAvgSalary

FROM Player

JOIN Men ON Player.TID = Men.TID

GROUP BY LID) MenAvg ON League.LID = MenAvg.LID

JOIN

(SELECT LID, AVG(Salary) as WomenAvgSalary

FROM Player

JOIN Women ON Player.TID = Women.TID

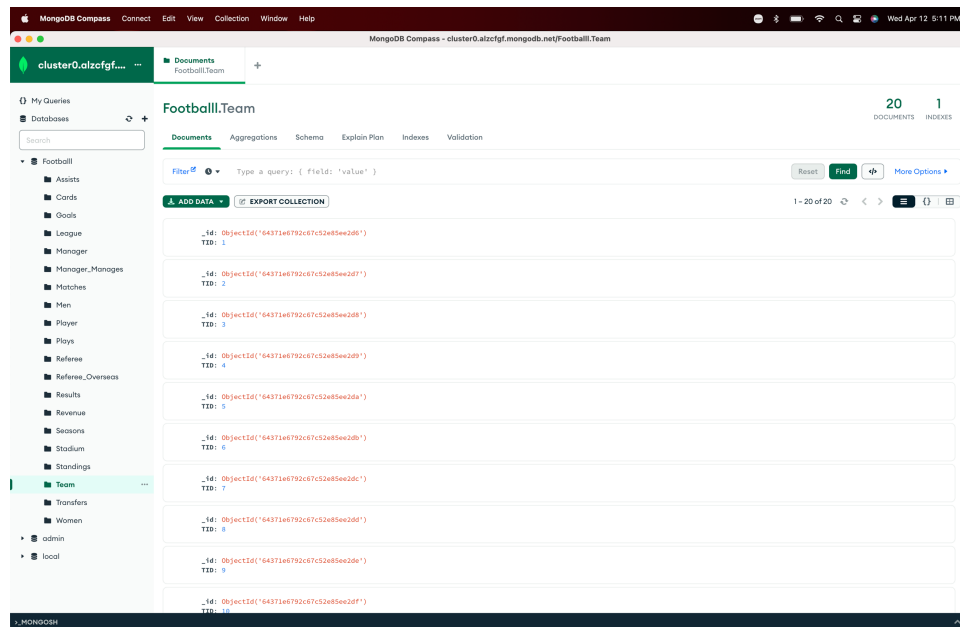
GROUP BY LID) WomenAvg ON League.LID = WomenAvg.LID

Output:

Result Grid				Filter Rows:	Search	Export:
	LeagueName	MenAvgSalary	WomenAvgSalary			
▶	East Coast Cup	173088	68018.8			

Implementation of Relation Model via MongoDB

The relational model has been implemented in MongoDB and MongoDB Compass is used to query the database. Sample queries have been presented below,



Query 1

Analytical Purpose : Find out the goals scored by the teams in a specific season to Understand how many goals each team has scored and how the goal scored difference between the top teams vary

```
Football.Team.aggregate([
  { $lookup: { from: "Matches", localField: "TID", foreignField: { $in: ["TeamID_Home", "TeamID_Away"] }, as: "matches" } },
  { $unwind: "$matches" },
  { $match: { "matches.SID": 2022 } },
  { $lookup: { from: "Goals", localField: "matches.MID", foreignField: "MID", as: "goals" } },
  { $unwind: "$goals" },
  {
    $group: {
      _id: "$Name",
      TotalGoals: {
        $sum: {
          $cond: [
            { $in: ["$goals.Type", ["goal", "penalty"]] }, 1, 0 ] } } },
      { $sort: { TotalGoals: -1 } },
      { $limit: 10 }
    }
  }
])
```

Output:

PIPELINE OUTPUT

Sample of 10 documents

OUTPUT OPTIONS

<div> <div>_id: ObjectId('6437205992c67c52e85ee2f9')</div> <div>TeamName: "Las Vegas"</div> <div>TotalGoals: 30</div> </div>
<div> <div>_id: ObjectId('6437205992c67c52e85ee2fa')</div> <div>TeamName: "Chicago"</div> <div>TotalGoals: 28</div> </div>
<div> <div>_id: ObjectId('6437205992c67c52e85ee2fb')</div> <div>TeamName: "Boston "</div> <div>TotalGoals: 25</div> </div>
<div> <div>_id: ObjectId('6437205992c67c52e85ee2fc')</div> <div>TeamName: "Washington"</div> <div>TotalGoals: 22</div> </div>
<div> <div>_id: ObjectId('6437205992c67c52e85ee2fd')</div> <div>TeamName: "NewYork"</div> <div>TotalGoals: 20</div> </div>
<div> <div>_id: ObjectId('6437205992c67c52e85ee2fe')</div> <div>TeamName: "California"</div> <div>TotalGoals: 17</div> </div>
<div> <div>_id: ObjectId('6437205992c67c52e85ee2ff')</div> <div>TeamName: "Seattle"</div> <div>TotalGoals: 14</div> </div>

Query 2

Analytical Purpose: Here we try to compare the number of cards handed out by a specific referee and understand which referees is more prone to hand out cards for offences than others thus helping team determine the nature of play

```
Football.Referee.aggregate([
  { $lookup: { from: "Cards", localField: "RID", foreignField: "RID", as: "cards" } },
  { $unwind: "$cards" },
  {
    $group: {
      _id: "$Name",
      CardsHanded: {
        $sum: "$cards.Cards"
      }
    },
    { $sort: { CardsHanded: -1 } },
    { $limit: 5 }
  ]
})
```

Output:

PIPELINE OUTPUT

Sample of 5 documents

OUTPUT OPTIONS

<div> <div>_id: ObjectId('6437257692c67c52e85ee311')</div> <div>RefereeName: ""James Barnhill"</div> <div>CardsHanded: 15</div> </div>
<div> <div>_id: ObjectId('6437260992c67c52e85ee312')</div> <div>RefereeName: "Jack Vest"</div> <div>CardsHanded: 13</div> </div>
<div> <div>_id: ObjectId('6437261792c67c52e85ee313')</div> <div>RefereeName: "George Parker"</div> <div>CardsHanded: 12</div> </div>
<div> <div>_id: ObjectId('6437261a92c67c52e85ee314')</div> <div>RefereeName: "John McDonough"</div> <div>CardsHanded: 12</div> </div>
<div> <div>_id: ObjectId('6437261e92c67c52e85ee315')</div> <div>RefereeName: "Walt Fitzgerald"</div> <div>CardsHanded: 10</div> </div>

Query 3

Analytical Purpose: Top 5 teams which generate the highest revenue helps us understand the financials generated by the club via various income streams such as broadcasting, advertising and so on which can ultimately determine the value possessed in the club to the stakeholders

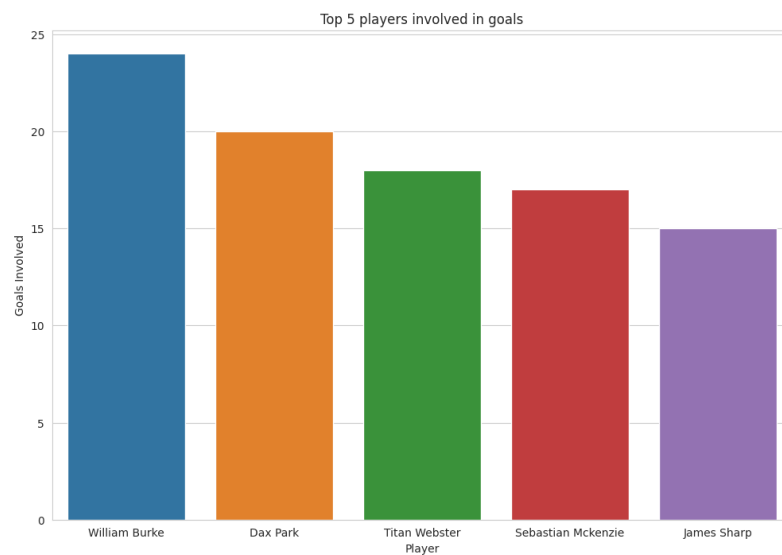
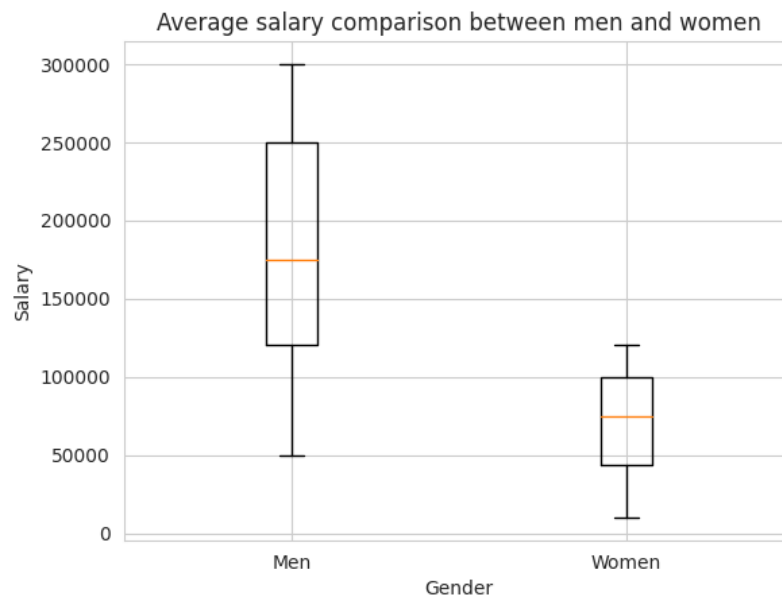
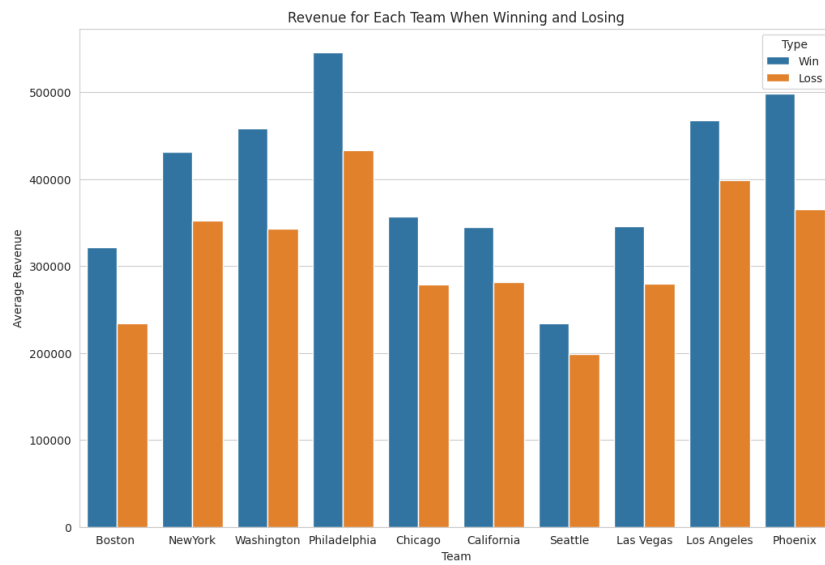
```
Football.Team.aggregate([
  { $lookup: { from: "Matches", localField: "TID", foreignField: { $in: ["TeamID_Home", "TeamID_Away"] }, as: "matches" } },
  { $unwind: "$matches" },
  { $match: { "matches.SID": <season> } },
  { $lookup: { from: "Revenue", localField: "matches.MID", foreignField: "MID", as: "revenue" } },
  { $unwind: "$revenue" },
  {
    $group: {
      _id: "$Name",
      TotalRevenue: {
        $sum: "$revenue.Amount"
      }
    }, { $sort: { TotalRevenue: -1 } },
    { $limit: 5 }
  }
])
```

Output:

PIPELINE OUTPUT	OUTPUT OPTIONS ▾
Sample of 5 documents	
<pre>_id: ObjectId('6437276392c67c52e85ee319') TeamName: "Philadelphia" TotalRevenue: 2728386.4</pre>	
<pre>_id: ObjectId('6437276892c67c52e85ee31a') TeamName: "Phoenix" TotalRevenue: 2491338.65</pre>	
<pre>_id: ObjectId('6437276c92c67c52e85ee31b') TeamName: "Los Angeles" TotalRevenue: 2339471.25</pre>	
<pre>_id: ObjectId('6437276e92c67c52e85ee31c') TeamName: "Washington" TotalRevenue: 2293341.75</pre>	
<pre>_id: ObjectId('6437277192c67c52e85ee31d') TeamName: "NewYork" TotalRevenue: 2154989.45</pre>	

Database access via Python

A connection was set up between the MySQL server and Jupyter notebook in order to access the database via Python and generate insights.



Conclusion and Future Scope

The project has successfully accomplished its goal of creating a scalable database for football. However, the potential of this database extends beyond football and can be expanded to other sports that follow a similar structure of entity relationship pairs. To achieve this, the future scope of the project includes developing Python scripts to extract data from reliable sources and update the database with the latest information.

This project aims to provide an excellent resource for sports enthusiasts worldwide by publishing the datasource for further use. The ability to access up-to-date and accurate data from a reliable source can significantly enhance the experience of following and analyzing sports. Thus, this project has tremendous potential to create a positive impact on the sports industry by providing an accessible and reliable source of information for sports fans worldwide.