

Soccer Analysis: Leveraging Data for Comprehensive Innovation

Team:

- Anitha Balachandran
- Aradhya Alva Rathnakar
- Bhavan Kumar Basavaraju
- Mahamaya Panda
- Shashi Kumar Kadari Mallikarjuna



INTRODUCTION

Soccer is a passion shared by millions of fans, players, and managers around the world, with winning being the ultimate goal. However, predicting match outcomes is difficult due to the multitude of factors involved. Traditional methods, such as relying on pundits or psychic predictions, are unreliable and fail to account for the complexity of the game.



THE WHY FACTOR

Our project aims to enhance the analysis and prediction of soccer matches by improving data collection processes. We recognize the growing interest in soccer for placing wagers and winning pride, which makes analysis and prediction crucial. By collecting data from various sources and providing functional information through reports, we can help soccer teams and managers assess performance, make informed decisions, and ultimately achieve success.

PROJECT GOALS

01

To provide accessible and meaningful information to soccer enthusiasts to help team management examine their strengths and shortcomings.

02

To design a robust data model that effectively represents soccer data entities, their attributes, and the relationships between them to enable efficient querying and analysis.

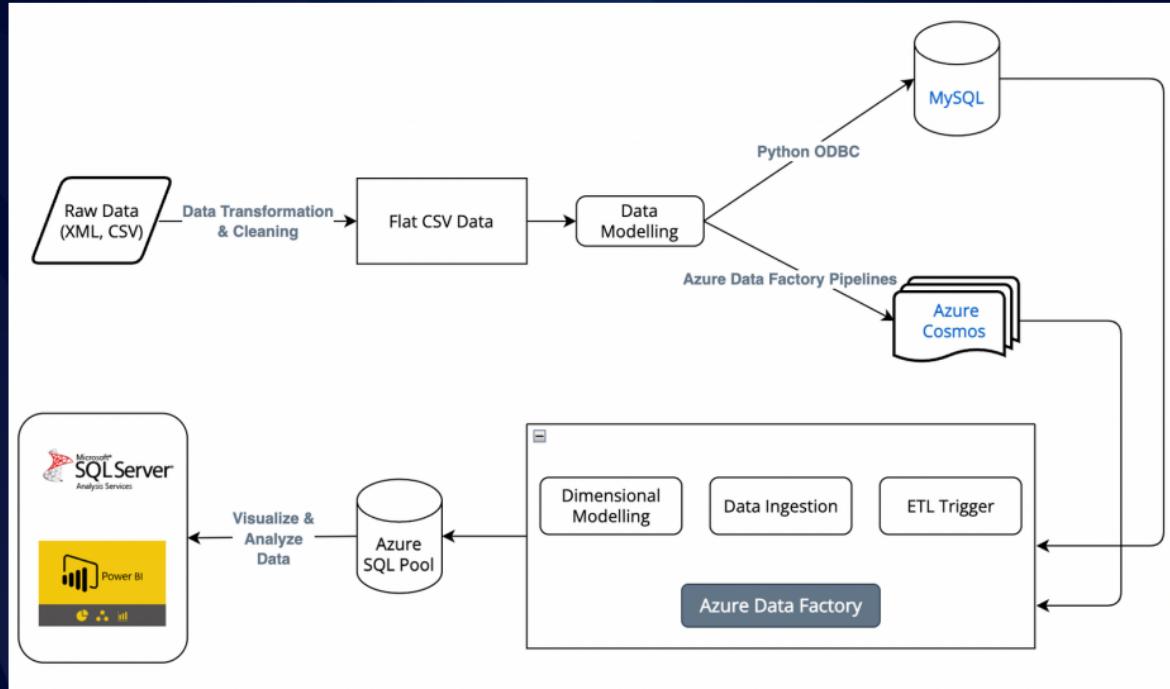
03

To clean and transform soccer data using ETL tools like Azure Data Factory and load it into data warehouse like Azure SQL pool on the cloud for historical analysis and reporting.

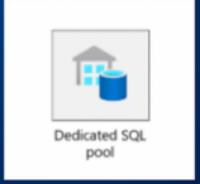
04

To use reporting tool like PowerBI to create visualizations that enhance the user experience in analyzing the information.

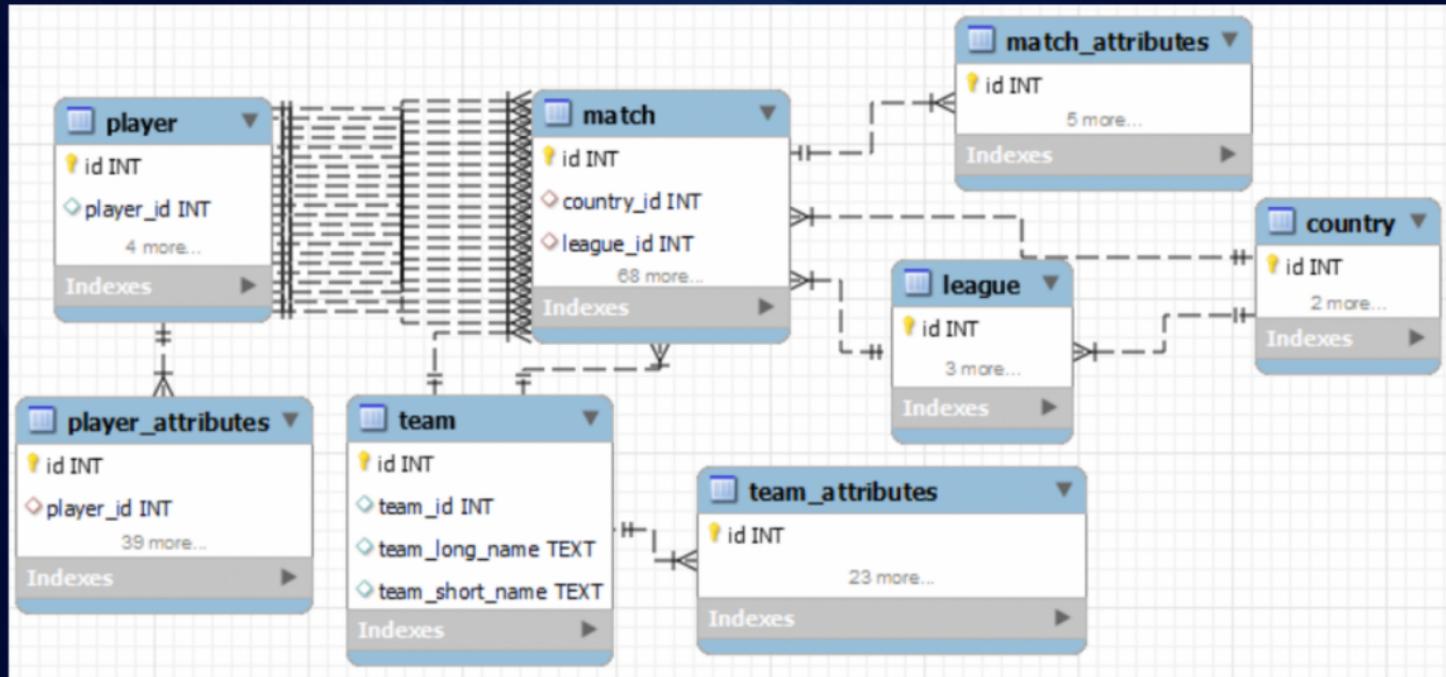
SYSTEM ARCHITECTURE



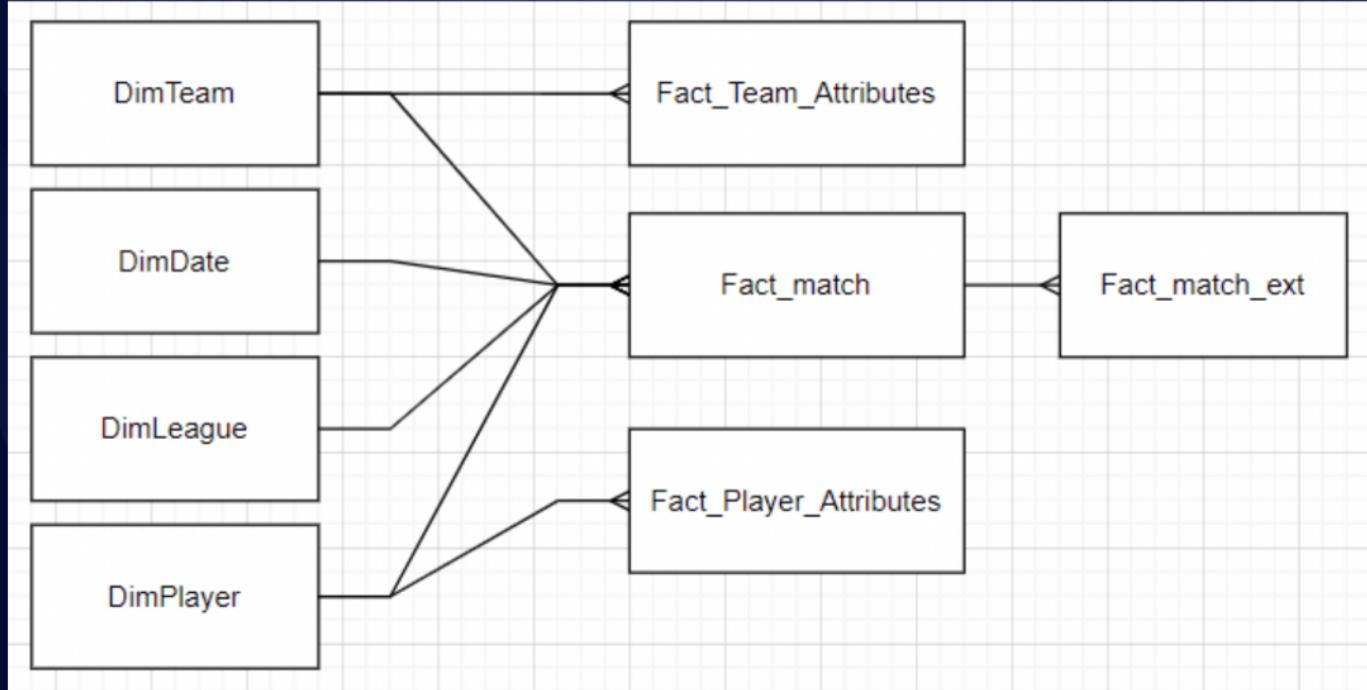
TOOLS USED

Source	OLTP	ETL and scheduling	OLAP/Data Warehouse	Visualize and analyze	Other tools used
 	 			 	     

ER DATA MODELING



DIMENSIONAL MODELING FOR AZURE SQL POOL



TIME FOR.....

LIVE CODE WALKTHROUGH

BEYOND THE FIELD: ANALYZING SOCCER THROUGH DATA

```
21 -- The query provides the maximum chance of goal a team has created throughout its years of gameplay. If there are multiple rows of same team
22 -- then it has created equal chance in different years.
23 SELECT x.team_id,g.team_long_name AS TeamName,x.DateKey,x.Avgchancebypass
24 FROM (
25 SELECT d.DateKey,t.team_id,AVG(t.chanceCreationPassing) AS Avgchancebypass,
26 RANK() OVER (PARTITION BY t.team_id ORDER BY AVG(t.chanceCreationPassing)) AS rank
27 FROM dbo.DimDate d,dbo.Fact_team_attributes t
28 WHERE d.DateKey = t.date_id
29 GROUP BY d.DateKey, t.team_id)x, dbo.DimTeam g
30 where x.rank =1 and x.team_id = g.team_id
31 ORDER BY x.team_id,x.DateKey;
32
33
```

Results Messages

	team_id	team_long_name	date_id	PressureStand	Aggressio...	DefenceSt...
1	1601	Ruch Chorzów	20100222	65.000000	60.000000	50.000000
2	1601	Ruch Chorzów	20110222	46.000000	48.000000	50.000000
3	1601	Ruch Chorzów	20120222	43.000000	44.000000	49.000000
4	1601	Ruch Chorzów	20130920	43.000000	44.000000	49.000000
5	1601	Ruch Chorzów	20140919	43.000000	44.000000	49.000000
6	1601	Ruch Chorzów	20150918	43.000000	44.000000	49.000000
7	1773	Oud-Heverlee Leuven	20120222	43.000000	44.000000	50.000000
8	1773	Oud-Heverlee Leuven	20130920	43.000000	44.000000	50.000000
9	1773	Oud-Heverlee Leuven	20140919	43.000000	44.000000	50.000000
10	1957	Jagiellonia Białystok	20100222	70.000000	70.000000	70.000000
11	1957	Jagiellonia Białystok	20110222	32.000000	56.000000	52.000000
12	1957	Jagiellonia Białystok	20120222	40.000000	50.000000	51.000000
13	1957	Jagiellonia Białystok	20130920	40.000000	50.000000	51.000000
14	1957	Jagiellonia Białystok	20140919	57.000000	56.000000	49.000000
15	1957	Jagiellonia Białystok	20150910	57.000000	56.000000	49.000000
16	2833	S.C. Olhanense	20100222	58.000000	45.000000	60.000000
17	2833	S.C. Olhanense	20110222	50.000000	45.000000	35.000000
18	2833	S.C. Olhanense	20120222	37.000000	24.000000	44.000000
19	2833	S.C. Olhanense	20130920	37.000000	31.000000	44.000000
20	2833	S.C. Olhanense	20140919	37.000000	34.000000	44.000000

Screen Reader Optimized Ln 35, Col 1 Spaces: 4 UTF-8 LF 1,457 rows MSSQL 00:00:03 socceranalysis.database.windows.net : soccerAnalysis

BEYOND THE FIELD: ANALYZING SOCCER THROUGH DATA

Run Cancel Disconnect soccerAnalysis Estimated Plan Enable Actual Plan Parse

```
31
32
33 -- The above code executes to provide us with the report data as per the league and the teams within the league along with the wins, losses and draws over the 9 season period.
34
35 SELECT [dt].team_long_name,
36     COUNT(CASE WHEN match.home_team_goal > match.away_team_goal THEN 1 ELSE NULL END) AS wins,
37     COUNT(CASE WHEN match.home_team_goal < match.away_team_goal THEN 1 ELSE NULL END) AS losses,
38     COUNT(CASE WHEN match.home_team_goal = match.away_team_goal THEN 1 ELSE NULL END) AS draws
39 FROM [dbo].[Fact_match] as MATCH
40 JOIN [dbo].[DimTeam] as dt ON match.home_team_id = dt.team_id OR match.away_team_id = dt.team_id
41 JOIN [dbo].[DimLeague] as dl on dl.league_id=match.league_id
42 WHERE dl.league_name = 'England Premier League'
43 GROUP BY dt.team_long_name;
44
45
46
```

Results Messages

team_long_name	wins	losses	draws
Arsenal	137	94	73
Portsmouth	35	23	18
Hull City	65	46	41
Burnley	39	19	18
Middlesbrough	20	7	11
Cardiff City	18	11	9
Newcastle United	128	73	65
Stoke City	149	69	86
Bournemouth	14	15	9
Birmingham City	36	14	26
Reading	19	9	10
Southampton	68	44	40
Norwich City	74	37	41
Blackpool	15	14	9
Manchester United	151	96	57
Everton	126	78	100
Fulham	109	57	62
Sunderland	127	85	92
Tottenham Hotspur	139	91	74
West Ham United	118	89	88

Screen Reader Optimized Ln 28, Col 1 Spaces: 4 UTF-8 LF 34 rows MSSQL 00:00:03 socceranalysis.database.windows.net : soccerAnalysis

BEYOND THE FIELD: ANALYZING SOCCER THROUGH DATA

Run Cancel ⚙ Disconnect Change Connection soccerAnalysis Estimated Plan Enable Actual Plan Parse

```
1 -- -- The average amount of pressure, aggression and defence shown by each team each year
2
3 SELECT a.team_id,a.team_long_name,b.date_id,AVG(b.defencePressure) AS PressureStand,AVG(b.defenceAggression) as AggressionStand,AVG(b.defenceTeamWidth) as DefenceStand
4 FROM dbo.DimTeam a, dbo.Fact_team_attributes b
5 WHERE a.team_id = b.team_id
6 GROUP BY a.team_id,a.team_long_name,b.date_id
7 ORDER BY a.team_id,b.date_id;
8
9
10
11
12
13
14
15
16
```

Results Messages

	team_id	team_long_name	date_id	PressureStand	AggressionStand	DefenceStand
1	1601	Ruch Chorzów	20100222	65.000000	60.000000	50.000000
2	1601	Ruch Chorzów	20110222	46.000000	48.000000	50.000000
3	1601	Ruch Chorzów	20120222	43.000000	44.000000	49.000000
4	1601	Ruch Chorzów	20130920	43.000000	44.000000	49.000000
5	1601	Ruch Chorzów	20140919	43.000000	44.000000	49.000000
6	1601	Ruch Chorzów	20150910	43.000000	44.000000	49.000000
7	1773	Oud-Heverlee Leuven	20120222	43.000000	44.000000	50.000000
8	1773	Oud-Heverlee Leuven	20130920	43.000000	44.000000	50.000000
9	1773	Oud-Heverlee Leuven	20140919	43.000000	44.000000	50.000000
10	1957	Jagiellonia Białystok	20100222	70.000000	70.000000	70.000000
11	1957	Jagiellonia Białystok	20110222	32.000000	56.000000	52.000000
12	1957	Jagiellonia Białystok	20120222	40.000000	50.000000	51.000000
13	1957	Jagiellonia Białystok	20130920	40.000000	50.000000	51.000000
14	1957	Jagiellonia Białystok	20140919	57.000000	56.000000	49.000000
15	1957	Jagiellonia Białystok	20150910	57.000000	56.000000	49.000000
16	2033	S.C. Olhanense	20100222	50.000000	45.000000	60.000000
17	2033	S.C. Olhanense	20110222	50.000000	45.000000	35.000000
18	2033	S.C. Olhanense	20120222	37.000000	24.000000	44.000000
19	2033	S.C. Olhanense	20130920	37.000000	31.000000	44.000000
20	2033	S.C. Olhanense	20140919	37.000000	24.000000	44.000000

Screen Reader Optimized Ln 20, Col 10 Spaces: 4 UTF-8 LF 1,457 rows MSSQL 00:00:02 socceranalysis.database.windows.net : soccerAnalysis

BEYOND THE FIELD: ANALYZING SOCCER THROUGH DATA

Run Cancel Disconnect Change Connection soccerAnalysis Estimated Plan Enable Actual Plan Parse

```
46
47 -- The code provides us with the reporting data on what types of changes are made to a team in a league with respect to tactics.
48 -- This is also combined with the date on which the change is performed. This helps in detailing the probable tactics that can be implemented by a team.
49
50 select e.league_name,a.team_long_name,b.date_id ,b.buildUpPlaySpeedClass,b.buildUpPlayDribblingClass,b.buildUpPlayPassingClass,b.buildUpPlayPositioningClass,
51 b.chanceCreationCrossingClass,b.chanceCreationPassingClass,b.chanceCreationPositioningClass,b.chanceCreationShootingClass,b.defencePressureClass,b.defenceAggressionClass,
52 b.defenceDefenderLineClass,b.defenceTeamWidthClass from DimTeam a join
53 (select distinct team_id,date_id ,buildUpPlaySpeedClass,buildUpPlayDribblingClass,buildUpPlayPassingClass,buildUpPlayPositioningClass,chanceCreationCrossingClass,
54 chanceCreationPassingClass,chanceCreationPositioningClass,chanceCreationShootingClass,defencePressureClass,defenceAggressionClass,defenceDefenderLineClass,
55 defenceTeamWidthClass from Fact_team_attributes) b
56 on a.team_id=b.team_id join (select distinct d.league_name,c.away_team_id from Fact_match c join DimLeague d on c.league_id=d.league_id) as e
57 on a.team_id=e.away_team_id
58 ORDER BY e.league_name;
59
60
61
```

Results Messages

	league_name	team_long_name	date_id	buildUpPlaySpeedClass	buildUpPlayDribblingClass	buildUpPlayPassingClass	buildUpPlayPositioningClass
1	Belgium Jupiler League	Lierse SK	20110222	Balanced	Little	Mixed	Organised
2	Belgium Jupiler League	Lierse SK	20120222	Fast	Little	Mixed	Organised
3	Belgium Jupiler League	Lierse SK	20130920	Fast	Little	Mixed	Organised
4	Belgium Jupiler League	Lierse SK	20140919	Fast	Normal	Mixed	Organised
5	Belgium Jupiler League	Lierse SK	20150910	Fast	Normal	Mixed	Organised
6	Belgium Jupiler League	KAS Eupen	20110222	Balanced	Little	Mixed	Organised
7	Belgium Jupiler League	KV Mechelen	20100222	Balanced	Little	Mixed	Organised
8	Belgium Jupiler League	KV Mechelen	20110222	Fast	Little	Mixed	Organised
9	Belgium Jupiler League	KV Mechelen	20120222	Balanced	Little	Mixed	Organised
10	Belgium Jupiler League	KV Mechelen	20130920	Balanced	Little	Short	Organised
11	Belgium Jupiler League	KV Mechelen	20140919	Balanced	Normal	Mixed	Organised
12	Belgium Jupiler League	KV Mechelen	20150910	Balanced	Normal	Mixed	Organised
13	Belgium Jupiler League	KSV Cercle Brugge	20100222	Balanced	Little	Mixed	Organised
14	Belgium Jupiler League	KSV Cercle Brugge	20110222	Balanced	Little	Mixed	Free Form
15	Belgium Jupiler League	KSV Cercle Brugge	20120222	Fast	Little	Mixed	Organised
16	Belgium Jupiler League	KSV Cercle Brugge	20130920	Balanced	Little	Mixed	Organised
17	Belgium Jupiler League	KSV Cercle Brugge	20140919	Balanced	Normal	Mixed	Organised
18	Belgium Jupiler League	KSV Cercle Brugge	20150910	Balanced	Normal	Mixed	Organised
19	Belgium Jupiler League	Sporting Charleroi	20100222	Balanced	Little	Mixed	Organised

Screen Reader Optimized | In 4R Col 4 | Snaps: 4 | UTF-8 | F | 1457 rows | MSSQL | 00:00:03 | socceranalysis database windows.net | soccerAnalysis

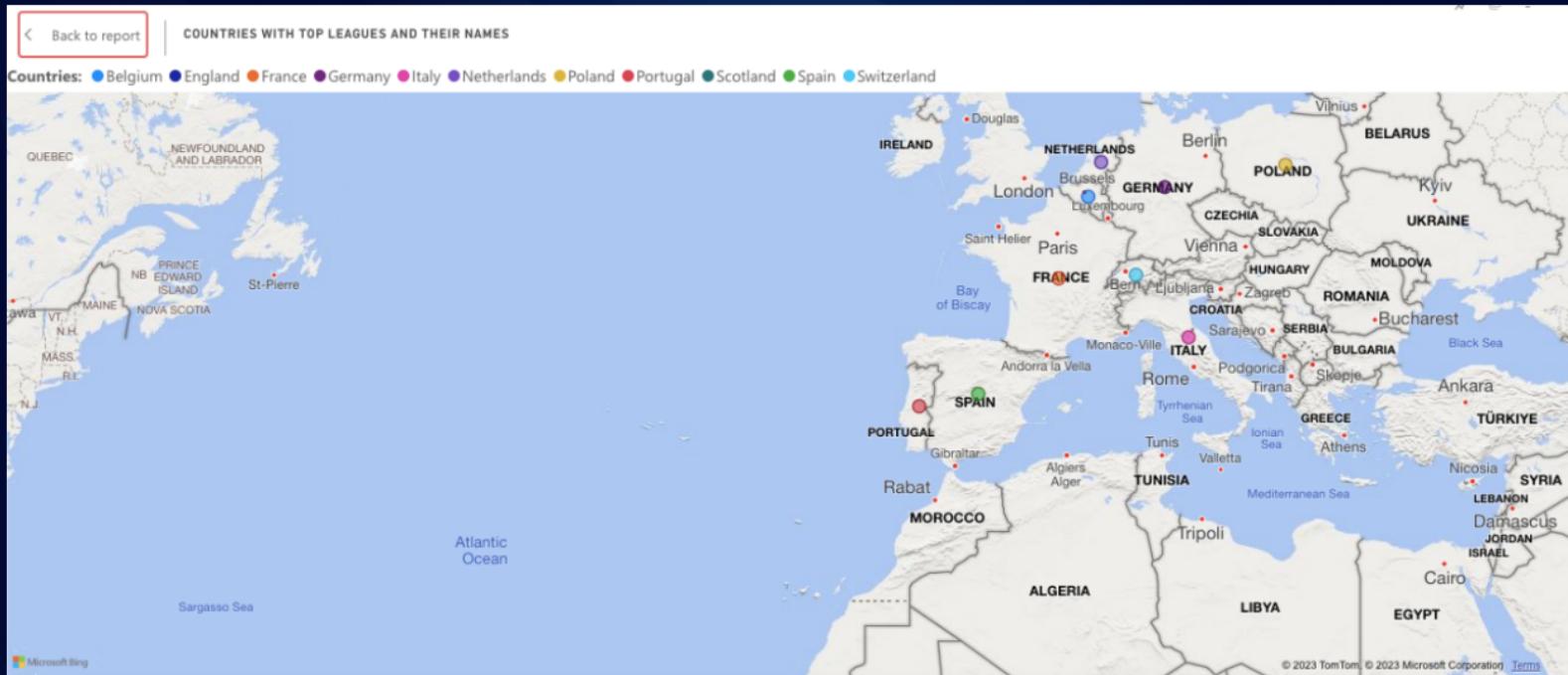
BEYOND THE FIELD: ANALYZING SOCCER THROUGH DATA

```
▶ Run □ Cancel ⚙ Disconnect ⚙ Change Connection soccerAnalysis ▾ ⚙ Estimated Plan ⚙ Enable Actual Plan ✓ Parse  
01  
62  
63  
64 -- Stored procedure to load the above view data into a table for faster retrieval:  
65 CREATE PROC [dbo].[load_players] AS  
66     truncate table dbo.match_players_ext;  
67     insert into dbo.match_players_ext select * from matchPlayers  
68 GO;  
69  
70  
71  
72  
73  
74  
75  
76
```

Messages

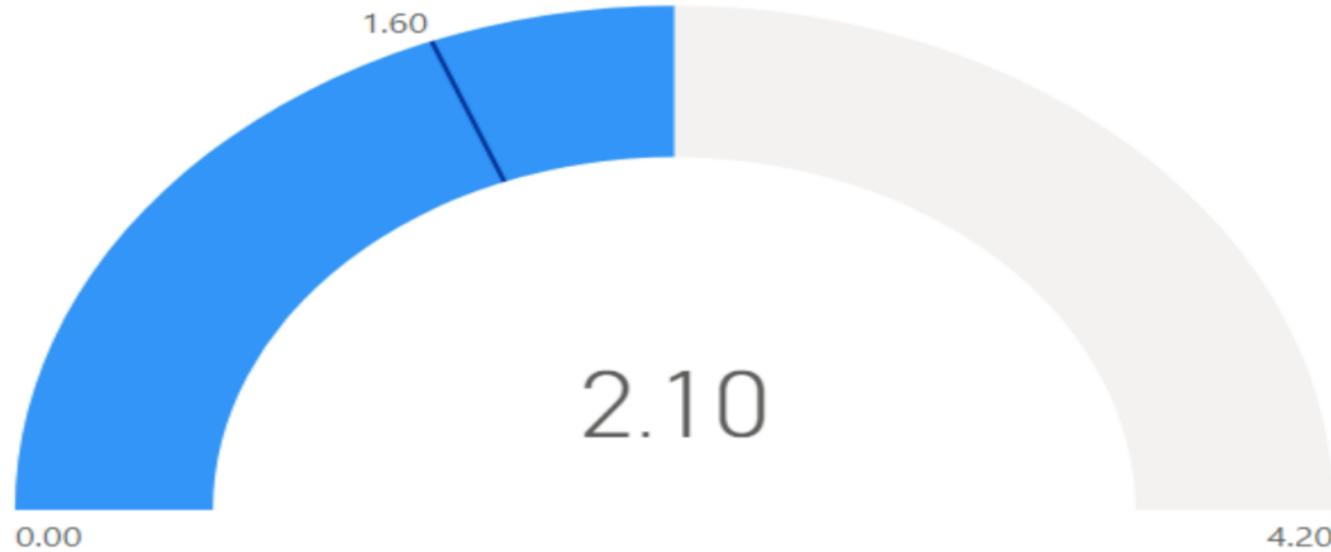
5:07:34 PM Started executing query at Line 1
Commands completed successfully.
Total execution time: 00:00:00.130

SIMPLIFYING SOCCER DATA: INSIGHTS THROUGH VISUALIZATION



SIMPLIFYING SOCCER DATA: INSIGHTS THROUGH VISUALIZATION

Relative average of home team and away team goals



SIMPLIFYING SOCCER DATA: INSIGHTS THROUGH VISUALIZATION

Key influencers Top segments

What influences attacking_work_rate to be ?

When...

Average of dribbling goes up 12.92 → 4.88x

Average of acceleration goes up 11.55 → 2.51x

Average of ball_control goes down 9.63 → 2.05x

Average of finishing goes up 16.98 → 1.72x

...the likelihood of attacking_work_rate being high increases by

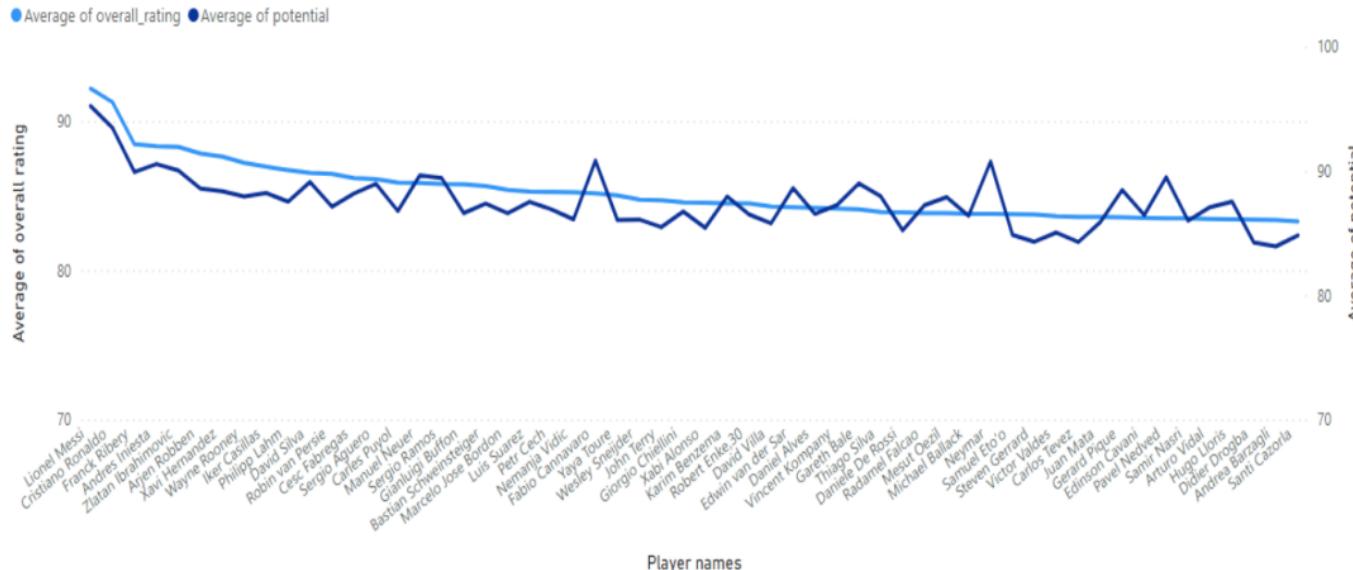
← On average when Average of dribbling increases, the likelihood of attacking_work_rate being high increases.

%attacking_work_rate is high

dribbling

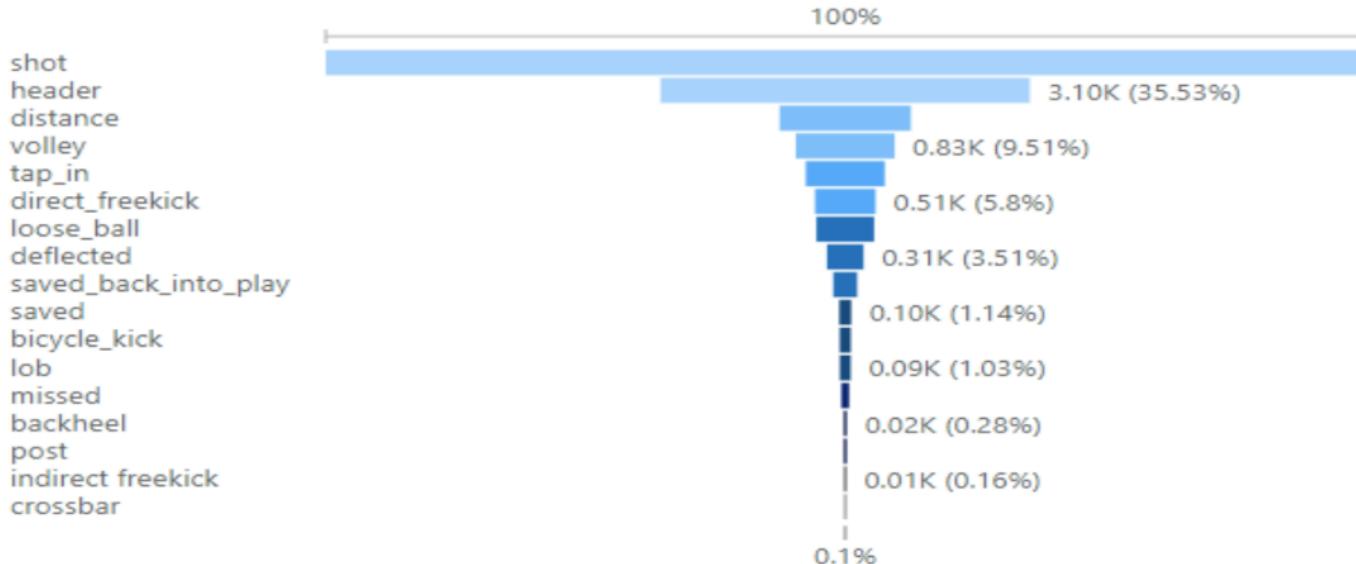
SIMPLIFYING SOCCER DATA: INSIGHTS THROUGH VISUALIZATION

Players and average of their ratings along with probable potential and preferred foot



SIMPLIFYING SOCCER DATA: INSIGHTS THROUGH VISUALIZATION

Count of diverse type of goals scored over all the seasons



CONCLUSION

The project aimed to analyze the soccer team's performance through data-driven technologies. Various data loading techniques were used such as Bulk insert and Upsert using Azure Data Factory. This allowed the team to use advanced analytics techniques to identify patterns that informed team strategy and improved performance. Hence the project demonstrated the effectiveness of data-driven technologies in improving soccer team performance.



FUTURE ENHANCEMENTS

-  Incorporating more real-time data to improve the accuracy and relevance of our data analysis and reporting.
-  Implementing AI/machine learning algorithms to gain deeper insights from the data and identify patterns that can help in predicting game outcomes.
-  Including social media data and sentiment analysis to gain a better understanding of fan engagement and its impact on team performance, and to create more personalized experiences for users.

THANK YOU!

Q/A?

