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

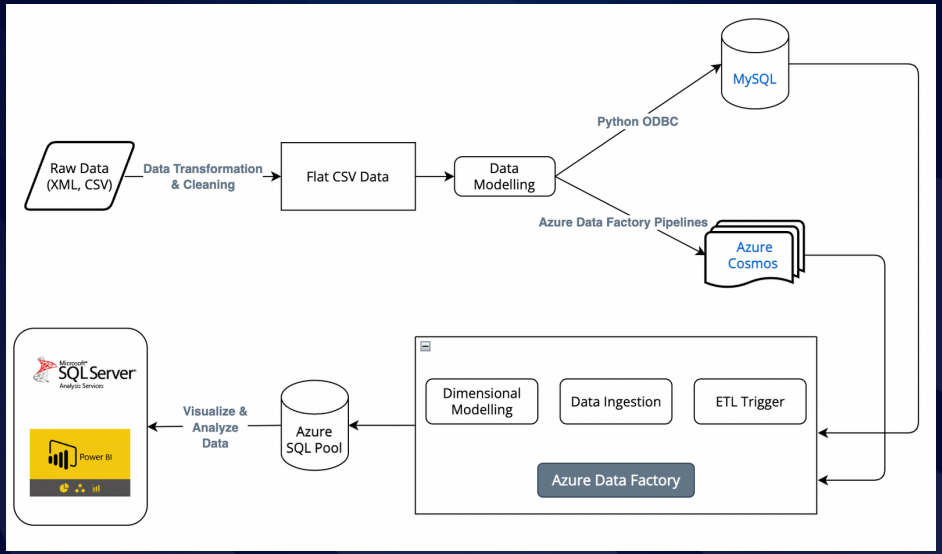
01 To provide accessible and meaningful information to soccer enthusiasts to help team management examine their strengths and shortcomings.

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.

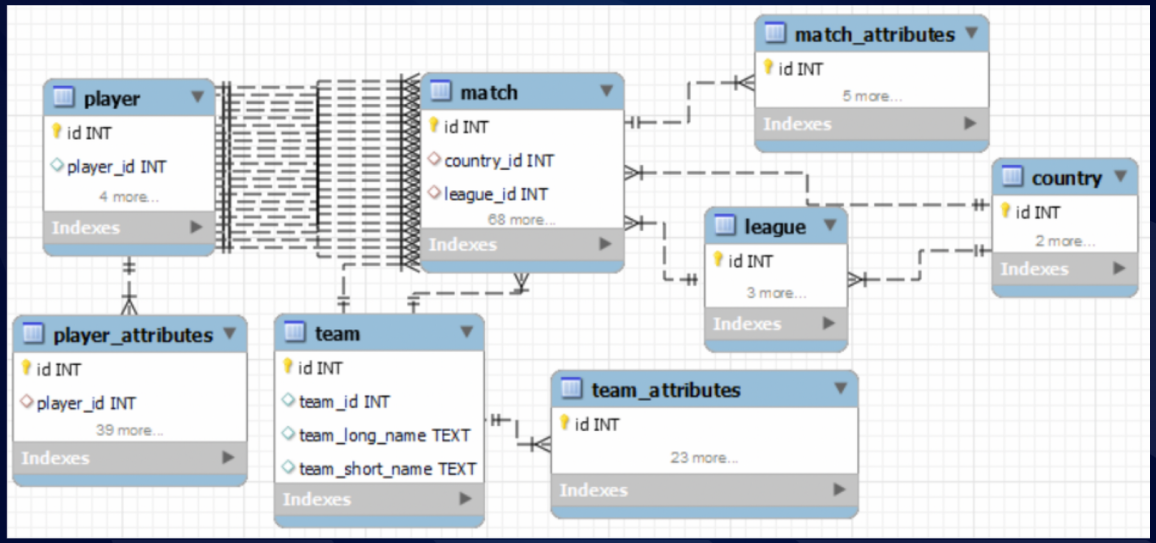
SYSTEM ARCHITECTURE



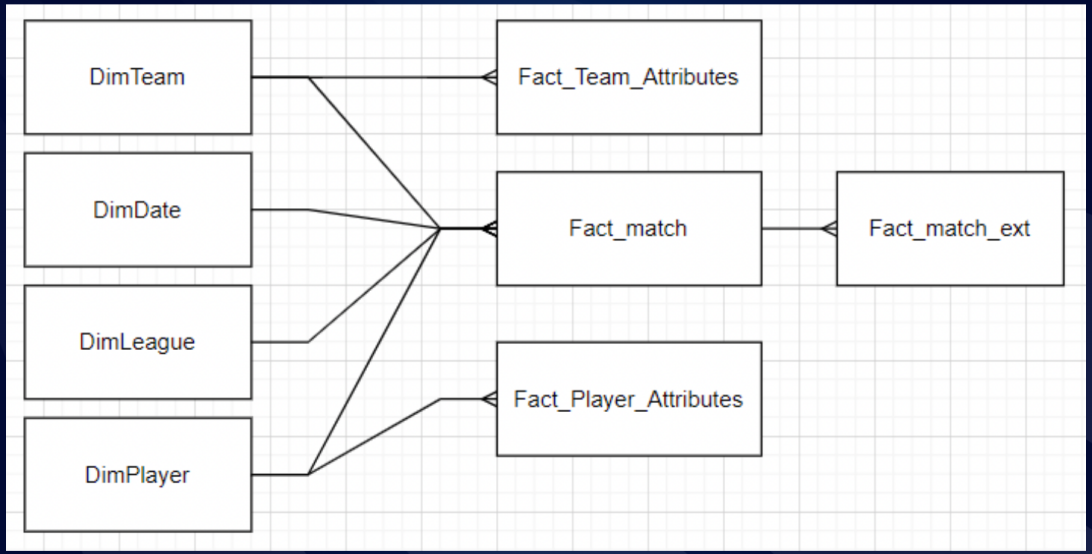
TOOLS USED

Source	OLTP	ETL and scheduling	OLAP/Data Warehouse	Visualize and analyze	Other tools used
CSV XML	MySQL Azure Cosmos DB: API for MongoDB	Azure Data Factory	Dedicated SQL pool	Power BI Microsoft SQL Server Analysis Services	GitHub jupyter Trello The LaTeX Project Prezi zoom

ER DATA MODELING



DIMENSIONAL MODELING FOR AZURE SQL POOL



TIME FOR…..
LIVE CODE WALKTHROUGH

BEYOND THE FIELD: ANALYZING SOCCER THROUGH DATA

```sql
-- 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
-- then it has created equal chance in different years.
SELECT x.team_id,g.team_long_name as TeamName,x.DateKey,x.Avgchancebypass
FROM (
SELECT d.DateKey,t.team_id,AVG(t.chanceCreationPassing) AS Avgchancebypass,
      RANK() OVER (PARTITION BY t.team_id ORDER BY AVG(t.chanceCreationPassing)) AS rank
FROM dbo.DimDate d,dbo.Fact_team_attributes t
WHERE d.DateKey = t.date_id
GROUP BY d.DateKey, t.team_id)x, dbo.DimTeam g
where x.rank =1 and x.team_id = g.team_id
ORDER BY x.team_id,x.DateKey;
```

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	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
1…	1957	Jagiellonia Białystok	20100222	70.000000	70.000000	70.000000
1…	1957	Jagiellonia Białystok	20110222	32.000000	56.000000	52.000000
1…	1957	Jagiellonia Białystok	20120222	40.000000	50.000000	51.000000
1…	1957	Jagiellonia Białystok	20130920	40.000000	50.000000	51.000000
1…	1957	Jagiellonia Białystok	20140919	57.000000	56.000000	49.000000
1…	1957	Jagiellonia Białystok	20150910	57.000000	56.000000	49.000000
1…	2033	S.C. Olhanense	20100222	50.000000	45.000000	60.000000
1…	2033	S.C. Olhanense	20110222	50.000000	45.000000	35.000000
1…	2033	S.C. Olhanense	20120222	37.000000	24.000000	44.000000
1…	2033	S.C. Olhanense	20130920	37.000000	31.000000	44.000000
2…	2033	S.C. Olhanense	20140919	37.000000	24.000000	44.000000

0 | 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 | Change Connection | 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
1	Arsenal	137	94	73
2	Portsmouth	35	23	18
3	Hull City	65	46	41
4	Burnley	39	19	18
5	Middlesbrough	20	7	11
6	Cardiff City	18	11	9
7	Newcastle United	128	73	65
8	Stoke City	149	69	86
9	Bournemouth	14	15	9
10	Birmingham City	36	14	26
11	Reading	19	9	10
12	Southampton	68	44	40
13	Norwich City	74	37	41
14	Blackpool	15	14	9
15	Manchester United	151	96	57
16	Everton	126	78	100
17	Fulham	109	57	62
18	Sunderland	127	85	92
19	Tottenham Hotspur	139	91	74
20	Aston Villa	118	88	88

0 | 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

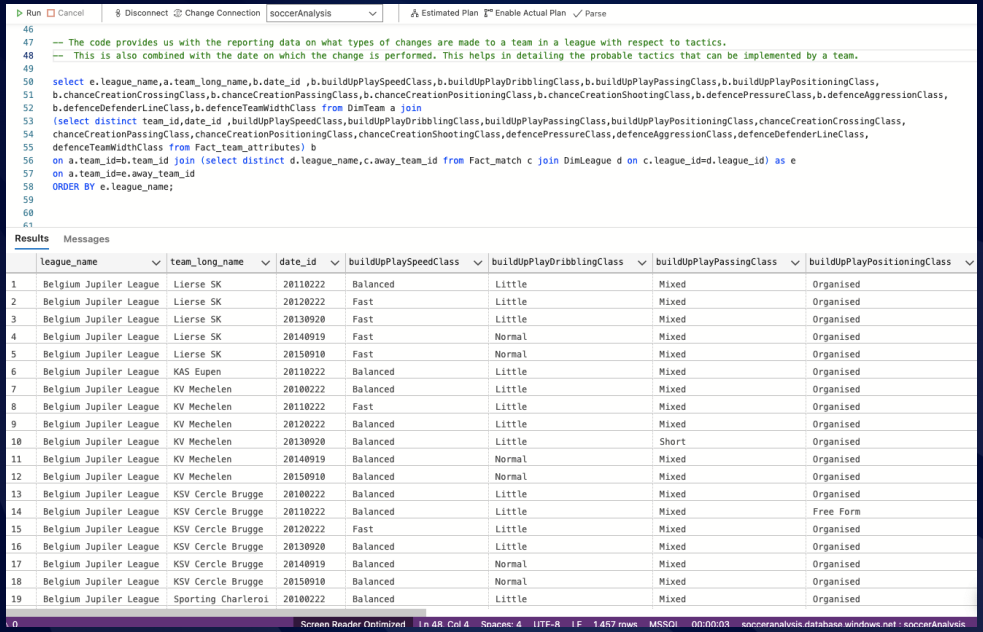
```sql
-- -- The average amount of pressure,aggression and defence shown by each team each year

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
FROM dbo.DimTeam a, dbo.Fact_team_attributes b
WHERE a.team_id = b.team_id
GROUP BY a.team_id,a.team_long_name,b.date_id
ORDER BY a.team_id,b.date_id;
```

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

0 | 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

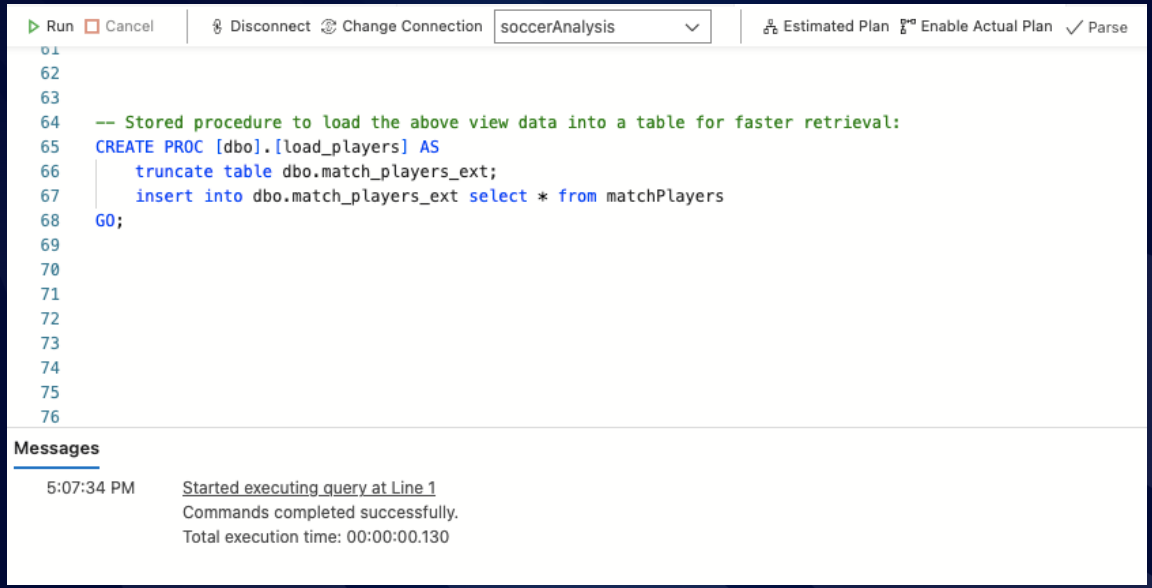



```
46
47  -- The code provides us with the reporting data on what types of changes are made to a team in a league with respect to tactics.
48  --  This is also combined with the date on which the change is performed. This helps in detailing the probable tactics that can be implemented by a team.
49
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
```

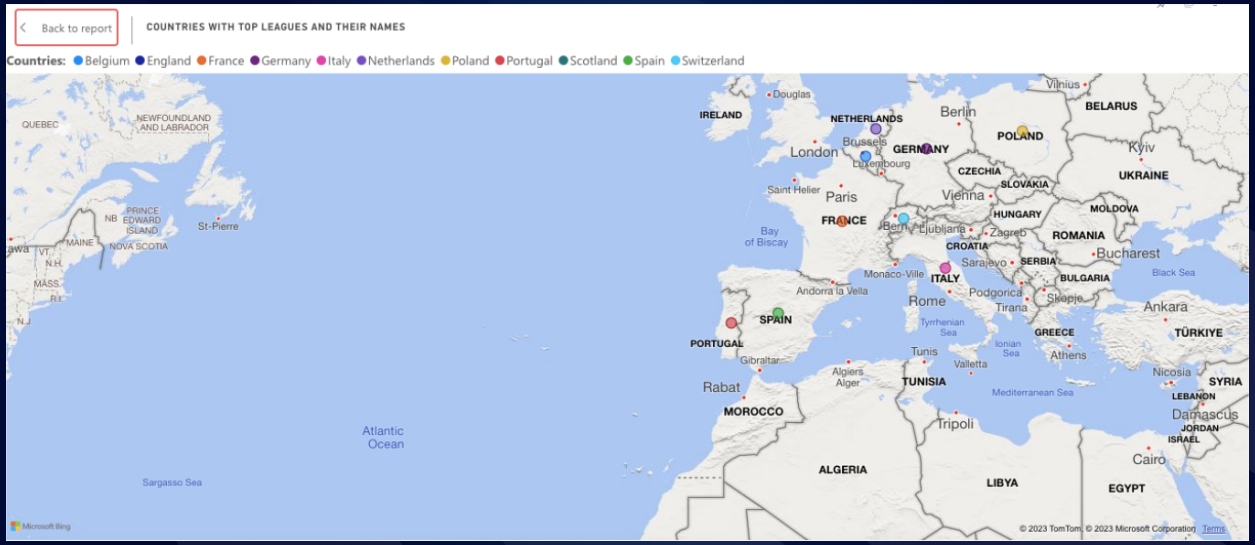
Results Messages

	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

BEYOND THE FIELD: ANALYZING SOCCER THROUGH DATA

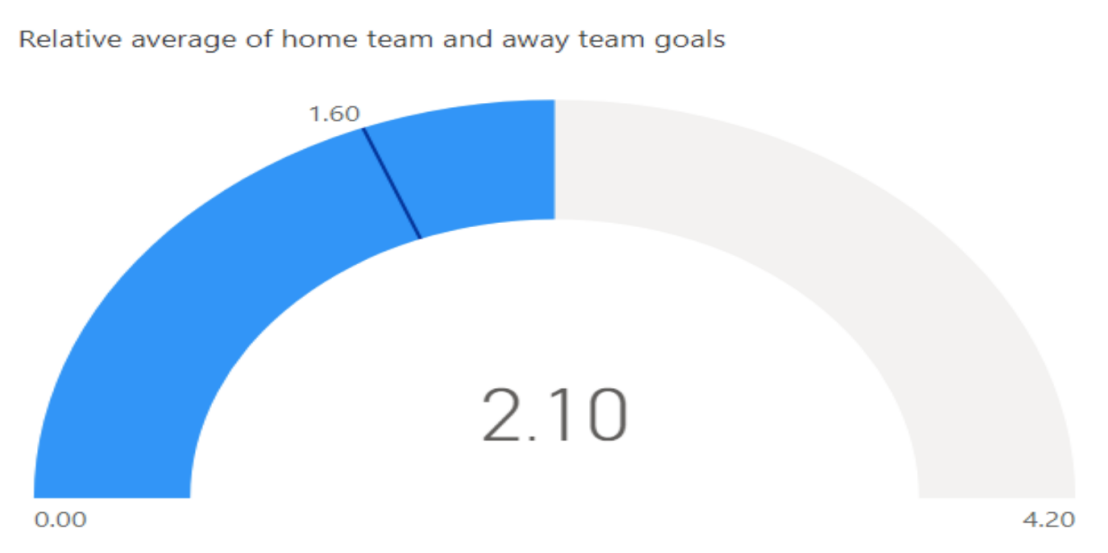


SIMPLIFYING SOCCER DATA: INSIGHTS THROUGH VISUALIZATION

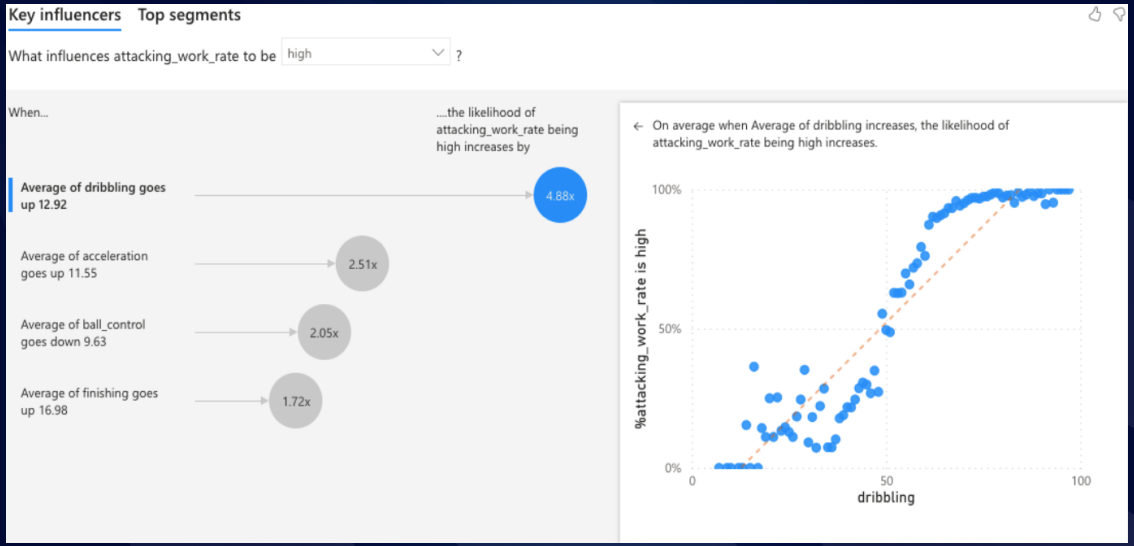


Prezi

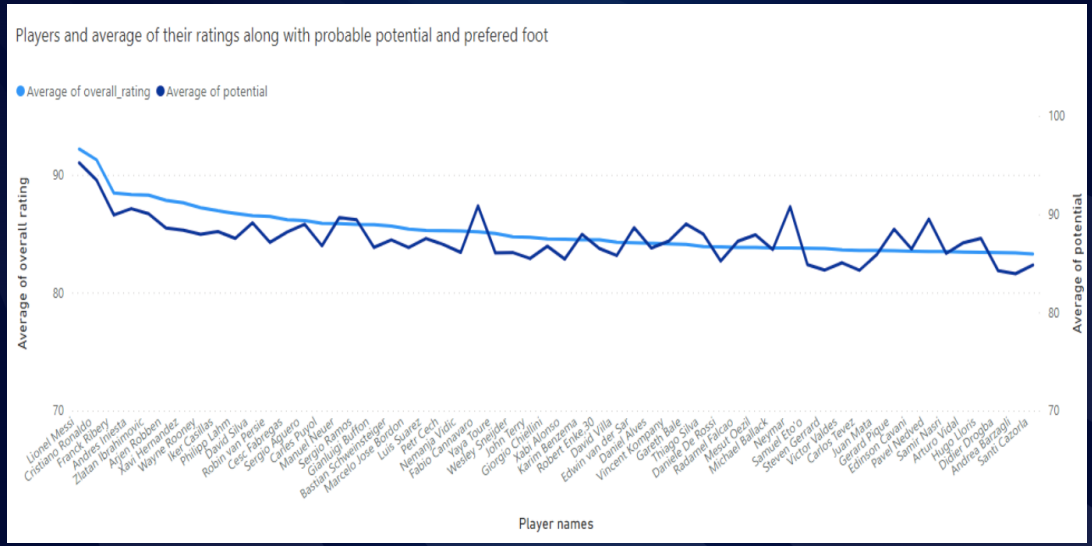
SIMPLIFYING SOCCER DATA: INSIGHTS THROUGH VISUALIZATION



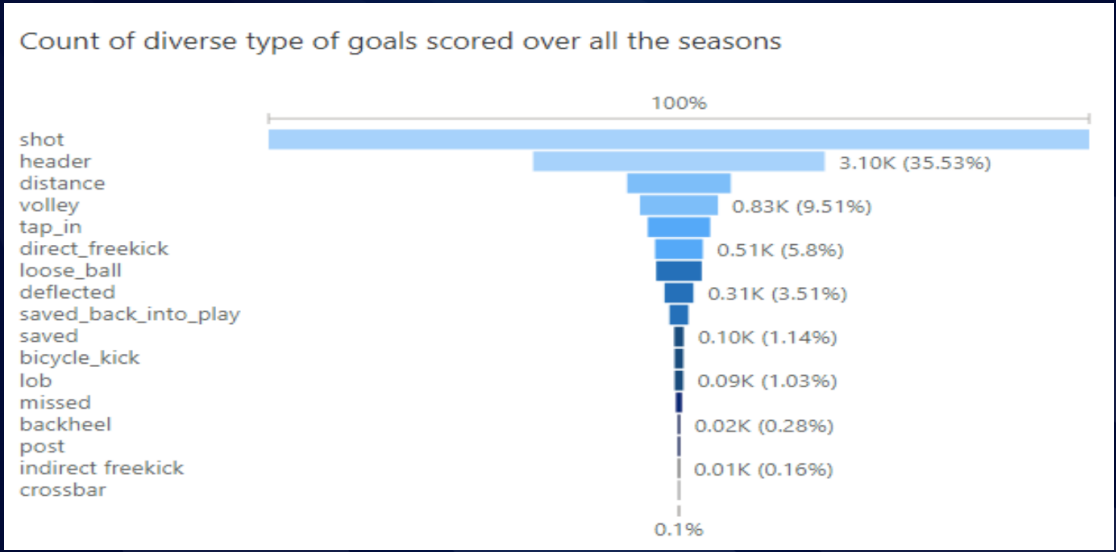
SIMPLIFYING SOCCER DATA: INSIGHTS THROUGH VISUALIZATION



SIMPLIFYING SOCCER DATA: INSIGHTS THROUGH VISUALIZATION



SIMPLIFYING SOCCER DATA: INSIGHTS THROUGH VISUALIZATION



KEY LEARNINGS

1. We got hands-on experience in dealing with the whole data lifecycle.
2. Worked with tools taught in class like MySQL and explored new tools like CosmosDB for Mongo, Azure Data Factory, Azure Dedicated SQL Pools, Power BI, etc., to handle data and perform analysis.
3. We practiced data modeling and dimensional modeling to create a new data project.
4. Agile development methodology can be used to develop data-related projects, and tools such as Trello can help with project management.
5. New presentation and documentation tools like Prezi and Latex were explored and used for this project.
6. Gained experience in pair programming, version control, project management, and teamwork.

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?