CADT បណ្ឌិត្យសភាបច្ចេកវិទ្យាឌីជីថលកម្ពុជា
Cambodia Academy of Digital Technology

CADT IDT វិទ្យាស្ថានបច្ចេកវិទ្យាឌីជីថល
Institute of Digital Technology

SCHOOL OF COMPUTER SCIENCE

DATA VISUALIZATION USING POWER BI

Football Analysis

So Kimlang (IDTB080003)

Sun Sokseka (IDTB080032)

Teang Laksey (IDTB080007)

Songeam Kanha (IDTB080016)

Eav Sarin (IDTB080012)

November, 21st, 2023

(TERM 6, GEN 8)

Lecturer: **CHAN SOPHAL**

Teacher Assistant: **THEAR SOPHAL**

Table of Contents

I. Introduction

This project tends to study football dataset analysis and visualization which contain 7 tables, 199 columns and 199414 rows. The journey will cover the whole process of data analysis from data collection, data cleaning and Exploratory Data Analysis (EDA) and how we do with the dataset to find the insightful information. Moreover, we also discover the source of our dataset and all the methods and tools that we used in the process to get results done. This involves identifying and addressing missing values, inconsistencies, and errors, ensuring the data's integrity and suitability for analysis. With the data cleansed and refined, the exploration phase commences, employing Exploratory Data Analysis (EDA) techniques to illuminate hidden patterns, trends, and relationships within the data.

II. Data Sources

European Soccer dataset is downloaded from kaggle.com in file sqlite. This dataset was collected by web crawling in multiple sources and thorough data collection and processing. The data was original sources from:

- http://football-data.mx-api.enetscores.com/ : scores, lineup, team formation and events
- http://www.football-data.co.uk/ : betting odds. Click here to understand the column naming system for betting odds:
- http://sofifa.com/ : players and teams attributes from EA Sports FIFA games. *FIFA series and all FIFA assets property of EA Sports.*

This dataset include:

- +25,000 matches

- +10,000 players
- 11 European Countries with their lead championship
- Seasons 2008 to 2016
- Players and Teams' attributes* sourced from EA Sports' FIFA video game series, including the weekly updates
- Team line up with squad formation (X, Y coordinates)
- Betting odds from up to 10 providers
- Detailed match events (goal types, possession, corner, cross, fouls, cards etc…) for +10,000 matches

Table	Total Rows	Total Columns
Country	11	2
League	11	3
Match	25979	115
Player	11060	7
Player_Attributes	183978	42
Team	299	5
Team_Attributes	1458	25

III. Tools

Kaggle is a platform for data science and machine learning which is a community and platform that provides datasets, competitions, and a collaborative environment for data scientists, machine learning practitioners, and researchers.

The dataset in this project was downloaded from kaggle in named **European Soccer Database** as file SQLite**.**

SQLite is an open-source, zero-configuration, self-contained, stand-alone, transaction relational database engine designed to be embedded into an application.
We use SQLite in order to convert the file dataset from SQLite to SQL as in POWER BI supports the SQL database server.

SQL Server Management Studio is a free multipurpose integrated tool to access, develop, administer, and manage SQL Server databases, Azure SQL Databases, and Azure Synapse Analytics. SSMS allows you to manage SQL Server using a graphical interface.
We import sql file dataset into SSMS for connecting with the POWER BI.

Microsoft Power BI is used to find insights within an organization's data. Power BI can help connect disparate data sets, transform and clean the data into a data model and create charts or graphs to provide visuals of the data. All of this can be shared with other Power BI users within the organization.
We use the Power BI to visualize and analyze European Soccer dataset by getting the insight to show in the dashboard as graphs, charts, etc…..

IV. Import Dataset into POWER BI

To import dataset into POWER BI, We must connect to a database in SQL server.

1. Convert file from SQLite to SQL

Due, the file dataset has format sqlite so we must convert from sqlite to sql to import into the database. We use the DB Browser (SQLite) tool to convert these files.

2. Import data to SQL server

After we convert the file to sql already, the next step is to import all tables and values into the database in SQL server to connect with Power BI Desktop.

Here is the step by step to can import data in SQL server:

- Create database using command: CREATE DATABASE *databasename*;
- Create each table using command: CREATE TABLE *table_name* (

  column1 datatype,

  column2 datatype,

  column3 datatype,

  ....

  );
- Insert values in each table: INSERT INTO *table_name (column1, column2,column3, ...)*VALUES *(value1, value2, value3, ...);*

3. Connection between SQL server and Power BI

To connect Power BI to database We must:

Get data => Choose SQL server => Input Server name => Connect.

After we connect, we will get all values of the data set and we can use it for cleaning or visualization.

V. Data cleaning

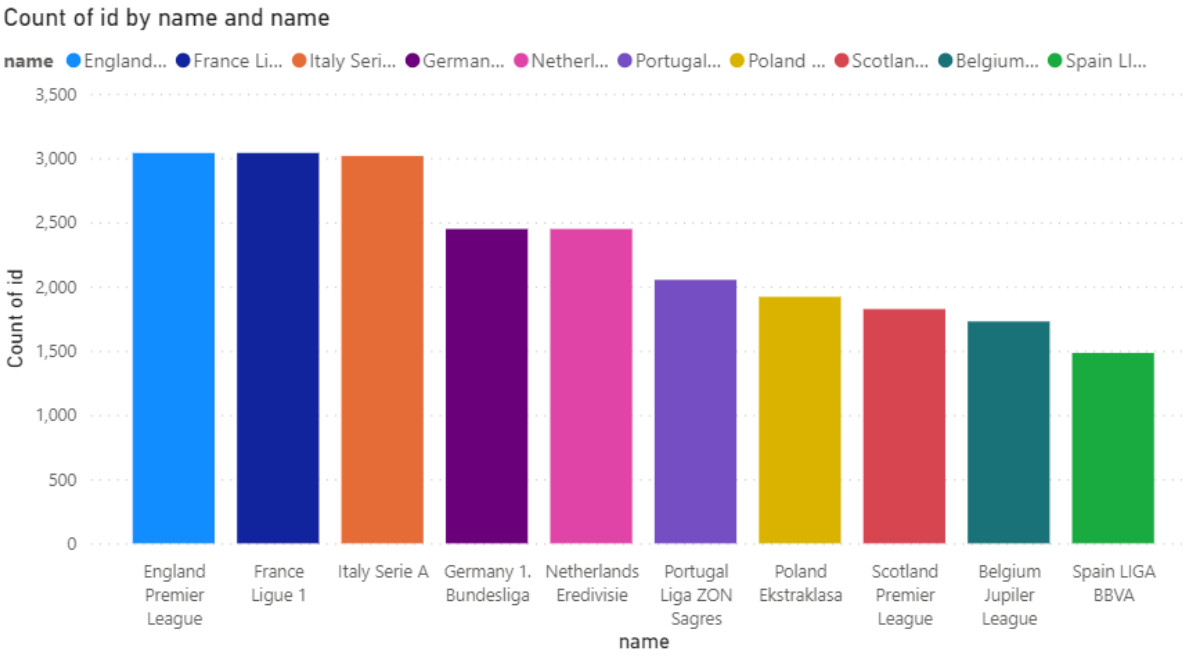
In this process, to get our data clean, we check and remove the duplicated data, data modeling which we need to add the relationship between each table from primary key to the foreign key, change the wrong datatype, fill or remove missing values, and remove unnecessary columns and data modeling. Through these meticulous data cleaning procedures, we ensure that our data is free from errors, inconsistencies, and incompleteness. This refined and reliable data provides a solid foundation for our subsequent analysis, enabling us to extract meaningful insights and uncover the hidden patterns that govern the world of football.

VI. Data visualization and Recommendations

1. Amount of matches in each league

This bar column chart shows the amount of matches that each league contributes. It shows that 3 leagues such as England Premier League, France Ligue 1, Italy Serie A contribute in the most matches. However, Spain LIGA BBVA contributed the least amount of matches.

This graph is useful for us to know which team have most experiences which easily lead to the winner team, since they have many experiences already.



2. Amount of total goals in each league

Before we can plot the graph to show the total goals that each league achieve, we first need to calculate the total goals from each league using the DAX as following:

```
TotalGoals =
SUMX(FILTER(Match,
NOT(ISBLANK(Match[home_team_goal]))),
Match[away_team_goal])
```

As we can see in the map, we can compare which team achieved the most total goals for their league. It shows by using the bubble size, the bubble size is big or small according to the total goal that each league found.

The result of the map shows us that some league in Spain, Belgium, and UK have small bubble size which means that they achieve less total goals.

3. Total match result of each team over the seasons

To find the total match result of each team over the seasons, we create a column and use DAX to calculate the amount of wins, losses, and draws:

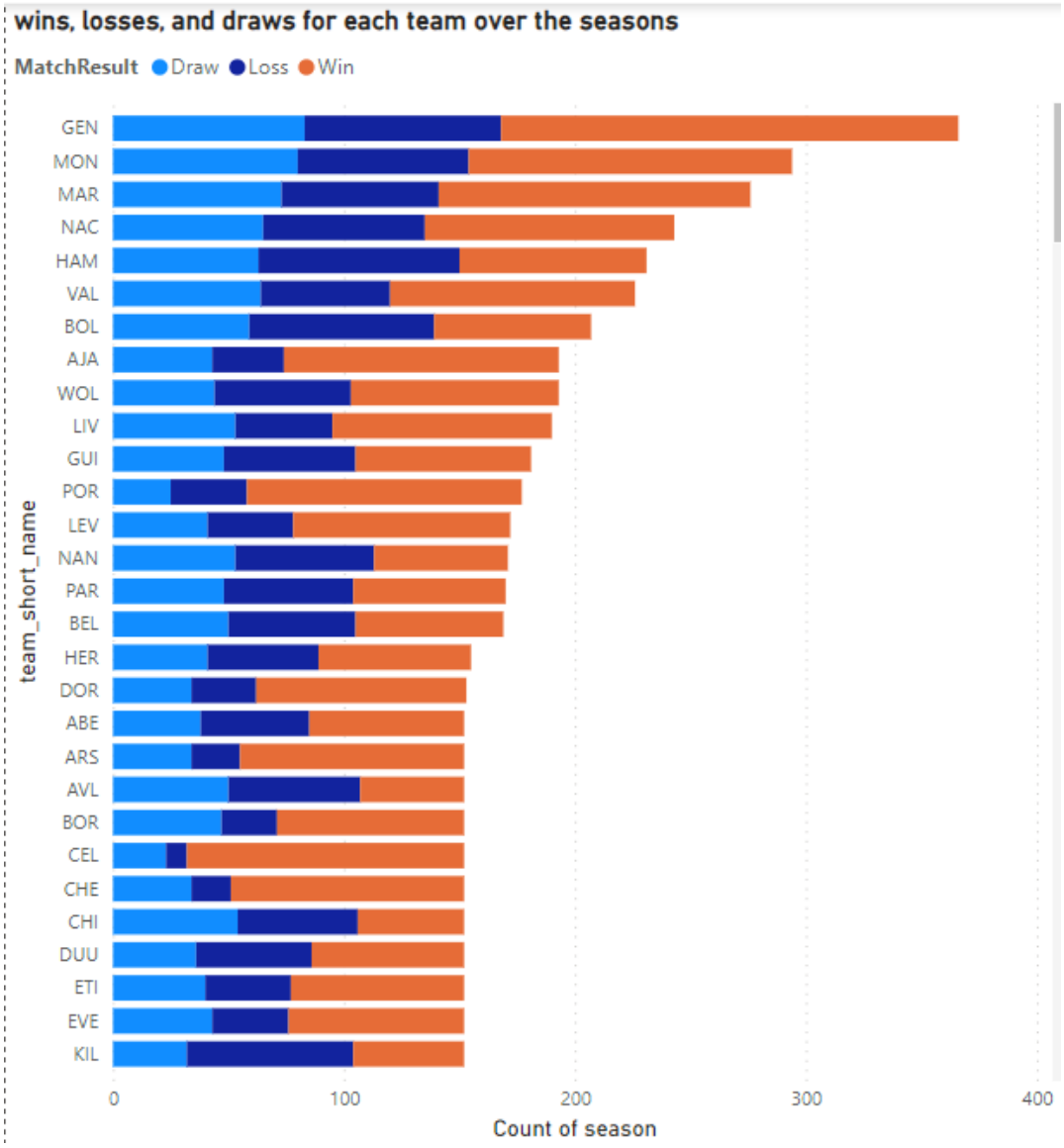
```
MatchResult =

    IF('match'[home_team_goal]                    >
    'match'[away_team_goal], "Win",

    IF('match'[home_team_goal]                    <
    'match'[away_team_goal], "Loss", "Draw"))
```

Using the Match Result, we can use it to show which team has the amount of win more than loss or draw. This graph shows that team short name that have most experience with a lot of win on the top. It is useful for us to recognize which team

has the most potential to win over which team. It is good for us to bet and predict

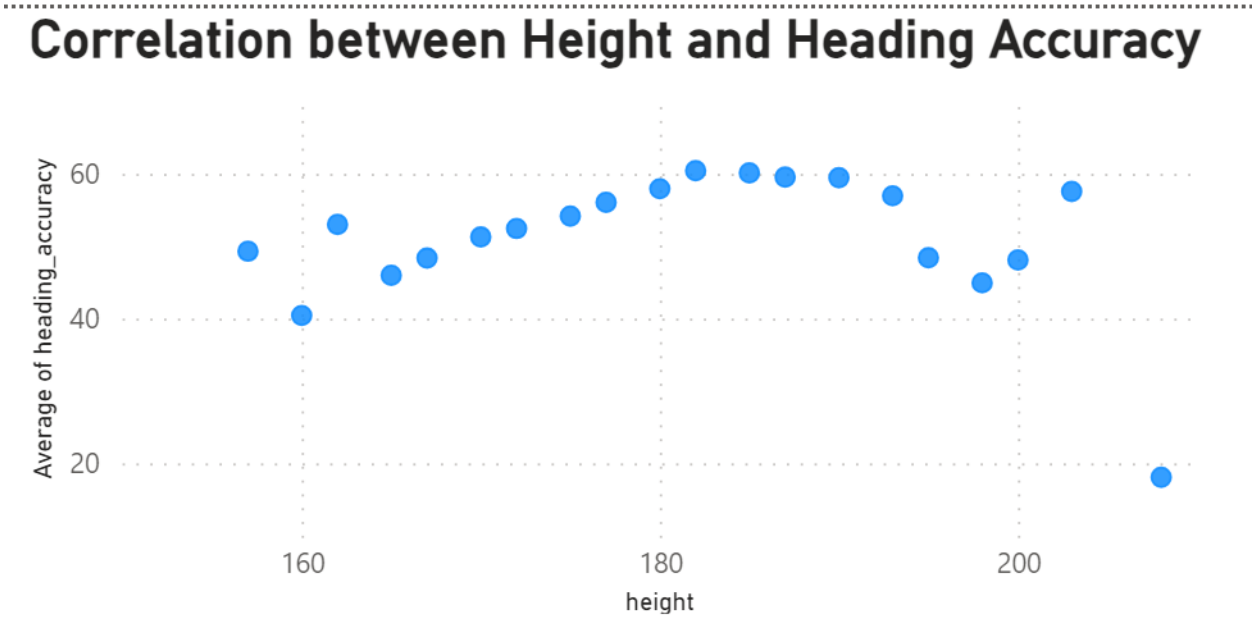


the winner.

4. Scatter Plot

We use a scatter plot to show the correlation between the height and heading accuracy which in this scatter plot shows that the heading accuracy is somehow affected by the height of the player. We can see that players that have a height of around 180 cm to 190 cm have the best heading accuracy.

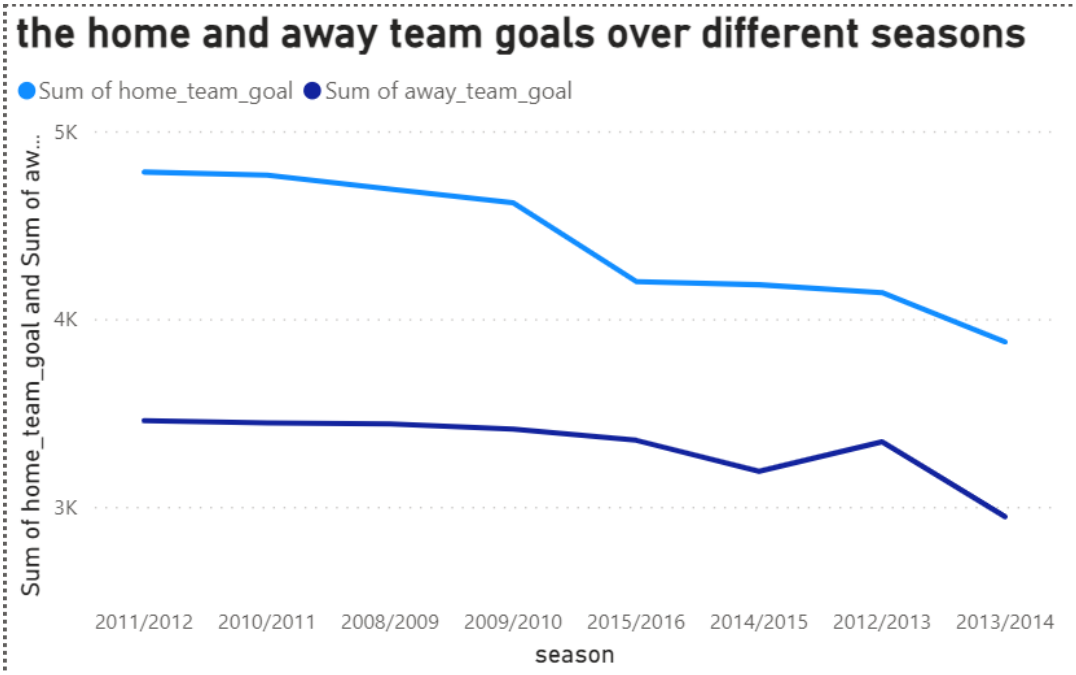
Most player's height between 180 cm to 190 cm is suitable for the player to play with heading.



5. Home team goals vs. away team goals

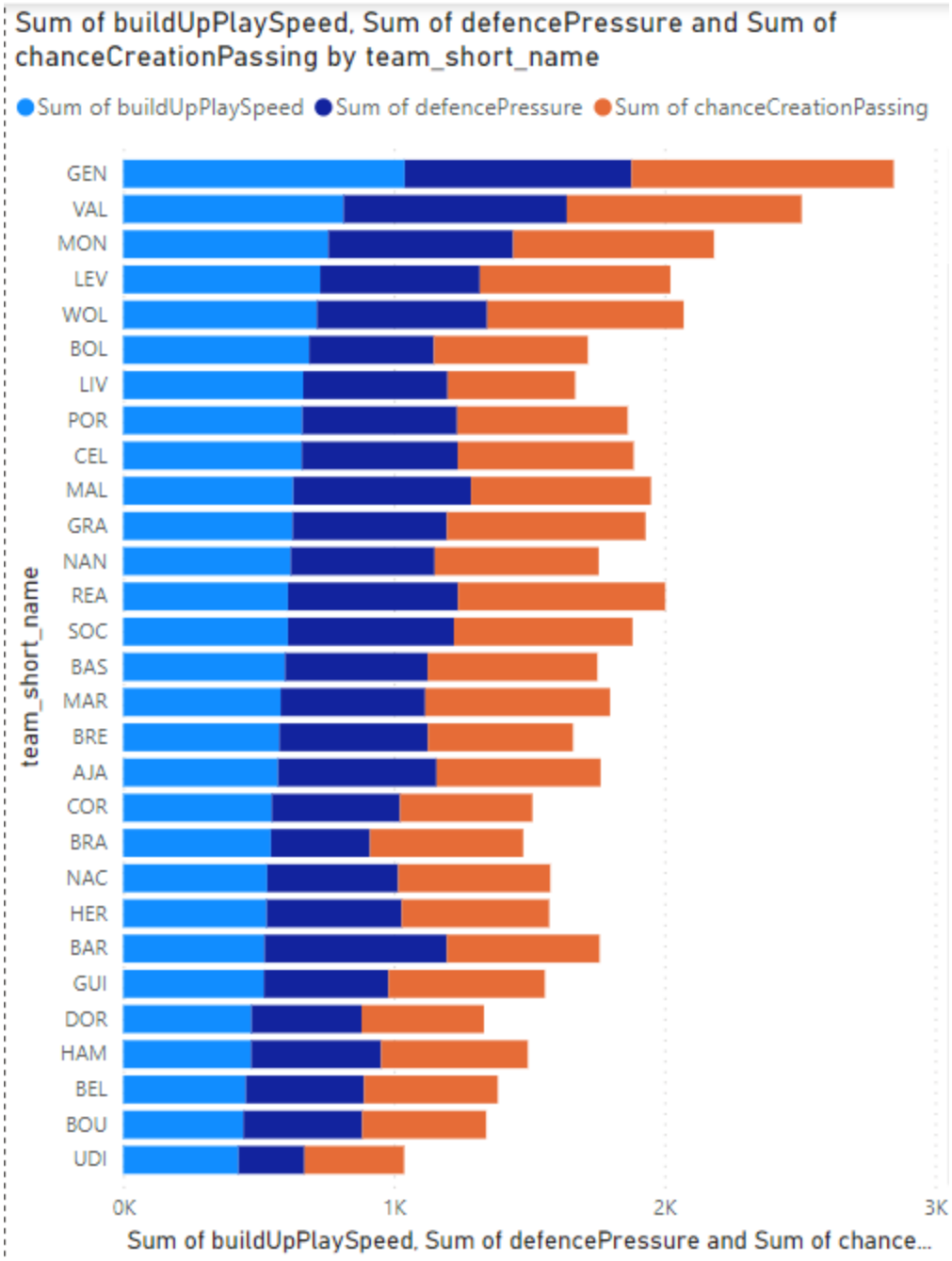
We use line graphs to show the insight comparing whether the home team or the away team achieves more goals. The following graph shows that the home team achieved more goals than the away team in every season.

This shows that if we want to bet between the home team and the away team, the home team is better for us to choose. However, it does not actually depend on the home team or away team only, there are many other things to consider when you want to bet on any team and we will tell you in the next insights. It helps us to predict for the next winner as well, according to the insight.



6. Result of each team over the seasons

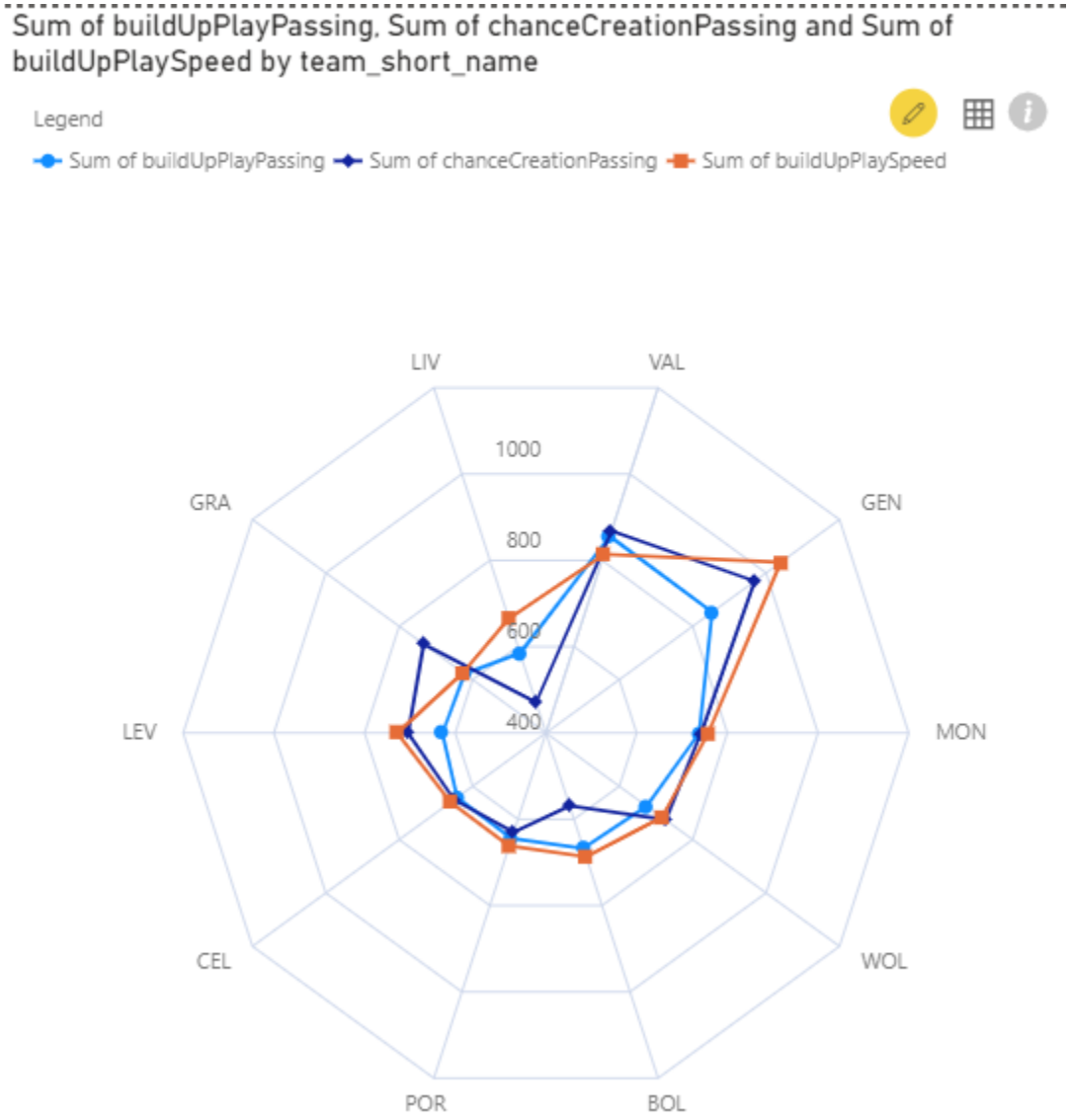
We use the stack bar column to show all teams, whether they win, draw, or lose. We can see that **GEN** is the team that has the most potential since it has the most experience joining the match and winning many times as well. The result here is according to build-up play speed, the sum of defense pressure, and the sum of chance creation passing. By using these attributes, we can see the top team that has the most potential to be the winner. This is one thing that helps us to predict and successfully bet for the winning team.



7. Tactics of each team doing the best at

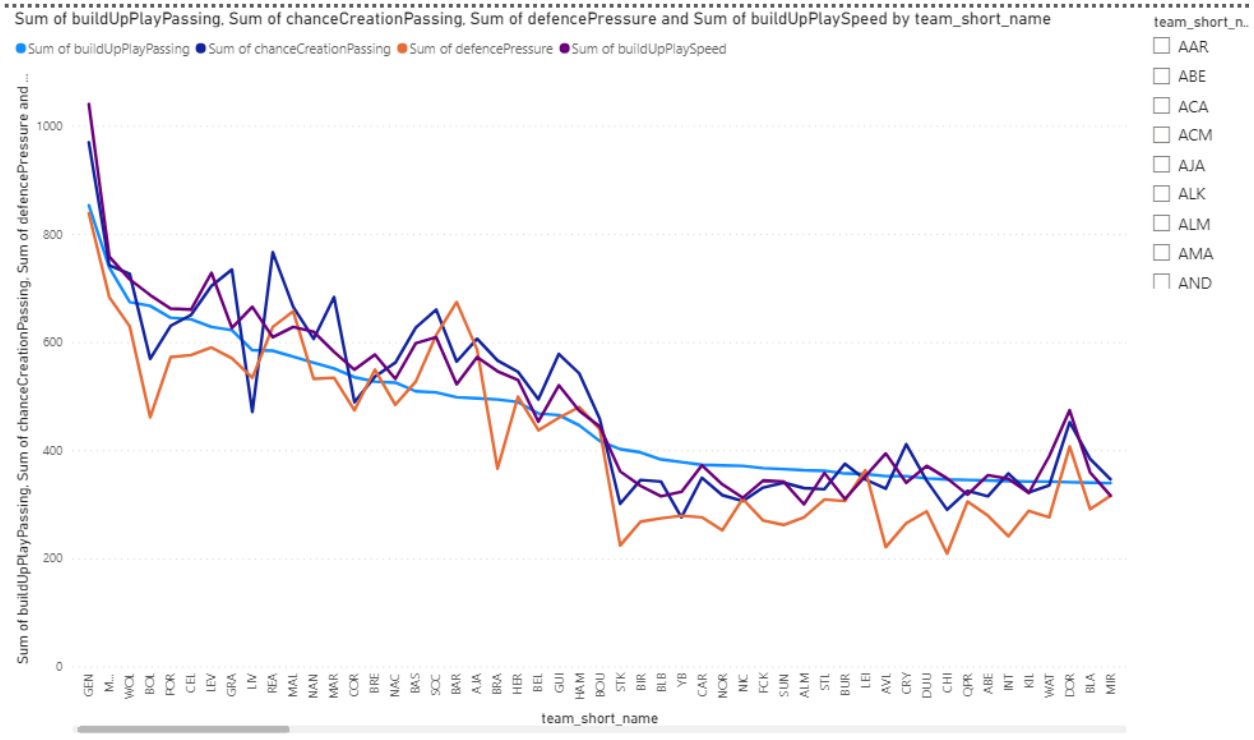
This Radar chart is useful to show the team that has the best tactics of each tactic such as build-up play speed, the sum of defense pressure, and the sum

of chance creation passing. As we can see in the graph, the team that the best at building up play speed is **GEN.**



8. Line graph of each team's tactics

This line graph contains build-up play speed, the sum of defense pressure, and the sum of chance creation passing for each team. This can be used to click on each team’s name to see which skill they are good at most.



VII. Challenges and Limitation

Throughout this project, our journey into the realm of football data analysis and visualization was something else, and we faced several hurdles, each posing unique obstacles to our progress. Here are some challenges and obstacles that we have faced through the whole process:

1. Data collection challenge

For the data collection part, it has been so tough for us to collect a good data set with the lecturer's requirement. After that, we have decided on choosing this final dataset from Kaggle with the type of data is SQLite and then we converted it into SQL standard dataset and lastly inserted it into power bi. In this process, the hardest part for us is converting file SQLite to SQL, and the way we need to

connect databases into Power BI, and it is pretty complicated since we usually only import file csv or excel into Power BI.

2. **Data cleaning complexities**

The data cleaning process, though essential, presented its own set of challenges. Identifying and addressing missing values, inconsistencies, and errors demanded careful attention and sophisticated techniques. We employed a variety of data cleaning methods, meticulously examining each data point to ensure its integrity. Moreover, since our dataset is super large and complicated, data cleaning became one of those struggle parts in the whole process.

3. **Data visualization constraint**

Transforming abstract data into compelling visualizations was not without its challenges. Effectively conveying complex relationships and patterns through visual representations required careful consideration of design principles and audience perception. We experimented with various visualization techniques, striving to create clear, informative, and engaging visualizations. Also, in each analysis, we need to think a lot about its insight into whether the outcome visualization could be useful to predict.

4. **Computational limitation**

As mentioned, our dataset is a large dataset which is the obstacle that we faced the most in this process in computer error or super slow running machine. This leads us to not do our project smoothly and also time constraints are definitely included.

5. **Resource constraint**

We are also lacking in resources of the dataset since the dataset has so many columns that are not useful, errors, and null.

6. **Time constraint**

On the other hand, time constraints are also affecting us so much on finishing this project.

VIII. Conclusion

Last but not least, throughout this project, we have learned a lot and improved our analysis skills including data cleaning, EDA, data visualization, group work, and so on. Also, we have done the project completely as the lecturer has provided and learned a lot from that.

IX. References

Link to original dataset
https://www.kaggle.com/datasets/hugomathien/soccer/data

Link to Power BI Service:
https://app.powerbi.com/links/dRzFBSYTs6?ctid=1e9461ec-5362-4329-ae46-61fa3e91c6d2&pbi_source=linkShare&bookmarkGuid=77880357-0d89-4588-bcc1-2587a6c99fdb