National College of Ireland

Project Submission Sheet – 2020/2021

Student Name:	Project Group 07
Student ID:	x21120803, x21153078, x21136921, x21148686
Programme:	MSc Data Analytics
Year:	2022-2023
Module:	Database and Analytics Programming
Lecturer:	Prof. Athanasios Staikopoulos
Submission Due Date:	26/04/2022
Project Title:	Database and Analytics Programming for Sports Analytics: Football
Word Count:	3885

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Project Group H
Date:	08/01/2021

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects** , both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

For Office use only:	
Signature:	
Date:	

Database and Analytics Programming

Vishan Lal
National College of Ireland
Student ID: 21120803

Gaurav Singh
National College of Ireland
Student ID: x21136921

Nishant Bharti
National colleg of Ireland
Student ID: 21148686

Taranjyot Singh
National College of Ireland
Student ID: x21153078

Abstract— **Football is a game of Emotions, Spirit as well as Tactics and Strategy. One of the Important aspects of Team Tactics is Team Formation. A series of Top 5 European football league dataset was considered for the years between 2014-20. Analysis of its dataset was carried to find Various insights by using attributes such as, Games Information, Shots hit by the players, Team Stats and Player appearances. Each football player poses various distinguished characteristics which include their Skill Moves, Pace, shooting, strength, stamina, physics, Passing, Movement, Mentality, and other attributes related to Football. We have compared these attributes as it is a vital aspect for a perfect team selection. Using this analysis an algorithm is implemented to choose the best players for the user-defined formation and the best teams are identified based on the user defined information. The results show that it leads to improved team structure through a systematic analysis of the top 5 European league football data sets. This kind of approach and analytical results can be useful to find the best teams and players attributes**

I. INTRODUCTION

A. Statement of Project

Players are scouted so that they can be trained and groomed early in their careers, with the goal of playing for a longer period of time. An investigation is carried out,
to determine shots distribution and how shots were taken from various spots on the football field in the dataset. One of the most essential components of scout selection is the analysis of yellow and red cards, which indicate a player's discipline. Throughout the Football Season, it is vital for a team to have appropriate information of their opponent's team traits in order to make better actions and decisions. Keeping track of shifting player performance and selecting the player that meets the stated criteria becomes a difficult undertaking. When it comes to choosing a player, numerous traits are considered based on the position of the player. For example, attackers and midfielders are evaluated on their speed, dribbling, crossing, finishing, short passing, accuracy, volley, vision, and reaction. The Defending Marking, Standing Tackle, Sliding Tackle, Aggression, Interceptions, Diving, Handling, and Reflexes of the Defender and Goalkeeper are evaluated. This study tries to improve the various stages of football decision-making.

B. Motivation and Relevance

Soccer has an estimated 3.5 billion fans worldwide, making it the most popular sport in the world. As the number of fans grows, teams strive to keep their supporters happy by winning as many trophies as possible. As data analytics became more sophisticated, its use in sports analytics grew significantly. Sports teams have begun to use Data Analysts to help them make guided and supervised decisions on scouting (buying new players for the club), selecting the best fit team for match day, and selecting the proper playing style and tactics. Scouting used to be done by managers physically analyzing players on the football pitch, but it has now become data-driven with the use of analysis, analytics, and visualizations.

C. Elicitation of Research Question

- Demonstrate the overall preview of the player based on the top 5 leagues
- To study the behaviour of a game in the top 5 league based on various characteristics of the players between the years 2014-2020
- To illustrate the pattern of home and away goals over the period of time from 2014-2020
- To determine whether home conditions are favourable to a team for winning the game
- Demonstrating the impact of the lastAction performed by the player to convert it into the successful goal
- Determining the expectation of goal based on the position from where the shot was taken
- Demonstrating the impact of the player in making the successful shot
- Does home or away determine the outcome of the football match
- Demonstrate whether the shotsOntarget and passes completed within 20 yrds of the goal have an impact on the overall result of the match

- Determine the impact of the yellow and the red cards on the outcome of the game
- Determine the team's performance for the period of 2014-20 for multiple seasons

II. RELATED WORK

The purpose of this study in the research paper [1] is to look at the percentages of shots on target and self-confidence of football players in various leagues. The study included 70 football players from Turkey's Yozgat province who routinely compete in Tier 1, Tier 2, and Regional Amateur Leagues.

The goal of this research paper [2] was to look into goal scoring in European football leagues and see whether elements are linked to projecting Expected Goals (xG). This notion aids in the evaluation of players, particularly strikers, based on the amount of goals they score over the course of a season. As a result, shots from Premier League and Bundesliga games (380 and 306) from the 2012-2013 season were analysed. On the field of play, all of the shots were divided into portions, each with a theoretical goal value.

The authors in research paper [3] provide a new-generation data-driven method for performance measuring and rating of soccer players in the third research article [3]. It enables researchers to investigate the statistical features of soccer performance by giving a score that effectively synthesizes a player's performance level throughout the course of a match or a series of matches. In Python, many data analysis methodologies were used to various properties of data in order to extract the result and apply it to the model. More complex algorithms might be created utilizing various machine learning models to determine a player's location during a match or a percentage of a match.

III. METHODOLOGY

A. *Description of the Datasets*

The chosen datasets consists of vary large data for the top 5 European Leagues which are as follows:

- **Players Appearances:** The purpose to choose this dataset was to know about the players performance on the shots, goals scored, own gaols scored, Assists performed by player, etc. on the various games played by these players between the year 2014-20 in the top 5 European leagues. The players appearances data also was joined with the players data which only consisted of player names against the player ID

- **Games Played:** The purpose to choose this data was to have a better understanding on the individual games played in top 5 leagues of European and to know the facts about the shots taken, home and away goals, probability for home and away win, etc. The Games data was joined with the Leagues data to get information of League names against the league ID.

- **Shots Taken:** The primary objective to choose this data was to gather insights on the shots attributes such as, minute of game at which the shot was taken, shot type, shot result, position from where shot was taken(X,Y coordinates), expected goal, etc for the top 5 European Leagues between the year 2014-20. The shots data also was joined with the players data which only consisted of player names against the player ID

- **Teams Statistics**: The purpose to chose this data was to demonstrate the characteristics of the individual team performace for the various games played. Team Attributes for each games such as, shots on target, shots taken, goals scored, fouls committed, yellow and red cards penalty, game result and many more have been taken into consideration to draw better insights for all the teams who have participated in the top 5 European leagues between the years 2014-2020. The team stats data was combined with the teams data to get the team names against the team ID.

B. *Data Gathering*

Kaggle is one of the largest data sources for achieving your data science goals. The four different semi-structured datasets on football are chosen from the repository and are programmatically stored in MongoDB using Python.

C. *Data Pre-Processing*

1. Player Appearance Dataset (Nishant Bharti):

In this dataset, there were null values in those attributes and columns which were not affecting the main variables data, hence the main variables were enough to carry out analysis without the variables containing null values. After further analysis and data pre-processing, outliers were present only in playerid, which won't affect analysis because no numeric major variables and their attributes have outliers.

2. Games Dataset (Taranjyot Singh):

- Deleting the columns PSA, WHH, WHD, WHA, VCH, PSCH,PSCD,PSCA,B365A and B365D as they are irrelevant in out dataset.
- Checked for null values in the data set, but in this dataset there is no null values present in any of the columns.

3. Shots Dataset (Gaurav Singh):

- **Player's Data:** For player's data, the data only contains player ids and names. The columns follow normal distribution although, I added one row which represents the player id as 0 for a Direct Goal.
- **Shots Data:** For shots data, after careful examine deep cleaning and pre-processing is performed on the dataset.

Steps:

- Checked for null values, for assisterID we have 84344 null values.
- Almost 84344 rows for assisterID were null values, because most the goals scored or missed were either penalty kicks, freekicks , set pieces or individual dribbling ability , so for those assisterID is not required so we make it 0. And also , converted the column from float to int type
- In the next step there is a conversion of positionX and positionY into scaled position of football field beause football pitch is 90 and 120 m in length. And also converted xGoals into scaled 100 position.
- Converting the minute column in different labels.
  - Time before half time is First Half in football terms.
  - Time after half time is Second Half in football terms.
  - Time after full time 90 minutes is Extra Time in football terms.
- After full cleaning and pre-processing, the data was added to PostgreSQL.

4. Team Stats Dataset (Vishan Lal):

- **Teams Data:** For teams's data, the data only contains player ids and names. The columns follow normal distribution. The _id column was removed when the data was loaded from mongo db to the python data frame

- **Team Stats Dataset:** For teams Stats data the dataset was cleaned rigorously and pre-processed and the following steps was performed on this dataset while doing the pre-processing and the data cleaning operation.

  Steps:

  - Null value checks was performed on all the columns of this dataset
  - Since the null values were only found in yellow cards column and the number of null values are too less so the nulls are being replaced by the mean value of the number of yellow cards
  - Generated the discipline field for every team based on the number of yellow cards received to the teams
  - Generated the Seasonal team win percenatge value in the team_stats dataframe
  - Generated the shot accuracy for each game for every team in the original dataframe.
  - Created a new data frame (**'team_stats_att'**) which contains the aggregated values of the original team stats datasets in order to have normalization in the dataset and get better insights at the team level.
  - The following aggregated columns were generated :
    - Yellow Cards Sum
    - Red Cards Sum
    - Discipline mean
    - Home and Away team winning percentage
    - Team Winning percentage
    - Total number of games played by every team
    - Shots and Goals Accuracy
    - team_stats_att dataframe.
    - Categorized the team winning, home winning and away winning perentages on the team_stats_att dataframe based on there values in the **team_stats_att** dataframe.
  - After full cleaning and pre-processing, the data was added to PostgreSQL.

D. Technologies Used- Python, MongoDB, PostgreSQL

1. **Python :** Python was chosen because it works on a variety of systems (Windows, Mac, Linux, Raspberry Pi, etc.). It features a fundamental grammar similar to that of English and helps programmers to write fewer lines of code than other programming languages. It runs on an interpreter framework, which means that code may be executed as soon as it is created, making prototyping a breeze.

2. **Mongo DB:** We can't predict all of the queries that will be run to evaluate the dataset while creating a database structure. As a result, we employed MongoDB's ad hoc query, which is a one-time command whose value is determined by a variable. The outcome of an ad hoc query may differ each time it is run, depending on the variables in question.

3. **PostgreSQL:** PostgreSQL is a flexible database that lets you customize data formats, index types, and functional languages, among other things. If you don't like anything about the framework, you may still create a custom plugin that can be tailored to your requirements, for as by installing a new optimizer. In this project, this capability was necessary in order to store the structured dataset in a database.

4. **Process Flow:**

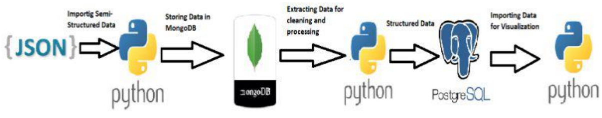
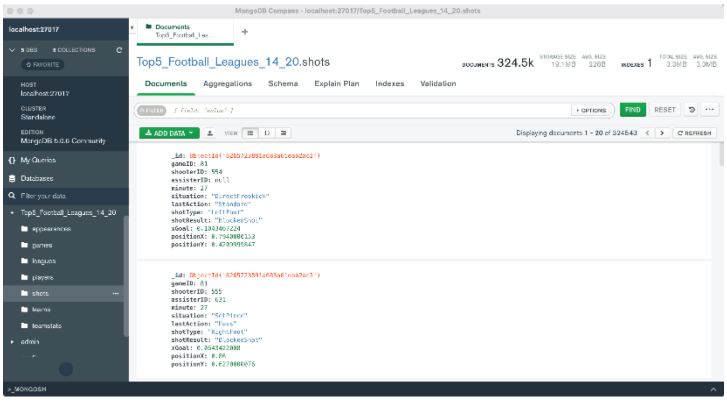

{JSON}
Importig Semi-Structured Data
python
Storing Data in MongoDB
mongoDB
Extracting Data for cleaning and processing
python
Structured Data
PostgreSQL
Importing Data for Visualization
python

Figure 1: Process Flow Diagram

Four JSON dataset of top 5 European leagues between 2014-2020 were taken from Kaggle, the semi-structured JSON was imported into the python environment via a script file (Jupyter Notebook). Now, programmatically, the JSON data in python was stored in the MongoDB database (Top_5_Football_leagues_14_20) with 7 collections storing 7 dataset each for further study and for identifying trends in the data.



Top5_Football_Leagues_14_20.shots

DOCUMENTS 324.5k STORAGE SIZE 19.1MB AVG. SIZE 228B INDEXES 1 TOTAL SIZE 3.3MB AVG. SIZE 3.3MB

{ field: 'value' }

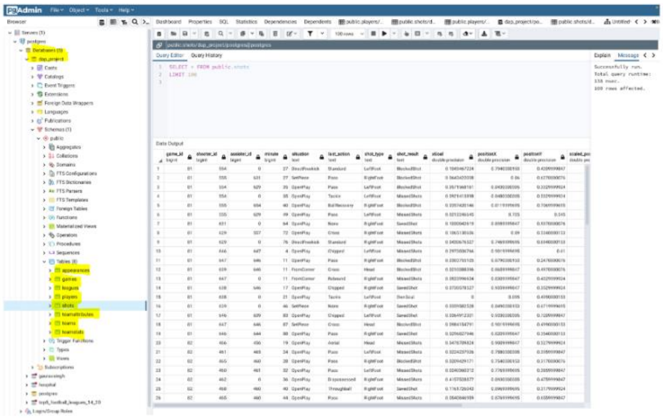
Displaying documents 1 - 20 of 324543

```
_id: ObjectId('6265723891a683a61eab2ac2')
gameID: 81
shooterID: 554
assisterID: null
minute: 27
situation: "DirectFreekick"
lastAction: "Standard"
shotType: "LeftFoot"
shotResult: "BlockedShot"
xGoal: 0.1043467224
positionX: 0.7940000153
positionY: 0.4709999847
```

```
_id: ObjectId('6265723891a683a61eab2ac3')
gameID: 81
shooterID: 555
assisterID: 631
minute: 27
situation: "SetPiece"
lastAction: "Pass"
shotType: "RightFoot"
shotResult: "BlockedShot"
xGoal: 0.0643422088
positionX: 0.86
positionY: 0.6270000076
```

Figure 2: Data Structure In MongoDB

To continue with the study, first, the processing of data must be performed, and a structured database must be created. Dataset analysis has been performed and maintained in a standardized PostgreSQL database. Later, all eight collections were taken from PostgreSQL and visualization was performed on every 4 partial merged data frame individually to get individual insights on Games, Shots, Player Appearances and Teams Stats. Further, all the partial data frame were merged into one data frame with the game ID attribute. Visualization was performed on a single data frame of 71 columns and data trends that shaped a community of chosen players as a team.



public.shots/dap_project/postgres@postgres

```
SELECT * FROM public.shots
LIMIT 100
```

```
Successfully run. Total query runtime: 138 msec.
100 rows affected.
```

	game_id bigint	shooter_id bigint	assister_id bigint	minute bigint	situation text	last_action text	shot_type text	shot_result text	xGoal double precision	positionX double precision	positionY double precision	scaled_po... double pre...
1	81	554	0	27	DirectFreekick	Standard	LeftFoot	BlockedShot	[illegible]	[illegible]	[illegible]	
2	81	555	631	27	SetPiece	Pass	RightFoot	BlockedShot	[illegible]	0.86	[illegible]	
3	81	554	629	35	OpenPlay	Pass	LeftFoot	BlockedShot	[illegible]	[illegible]	[illegible]	
4	81	554	0	35	OpenPlay	Tackle	LeftFoot	MissedShots	[illegible]	[illegible]	[illegible]	
5	81	555	654	40	OpenPlay	BallRecovery	RightFoot	BlockedShot	[illegible]	[illegible]	[illegible]	
6	81	555	629	49	OpenPlay	Pass	LeftFoot	MissedShots	[illegible]	0.725	0.545	
7	81	631	0	64	OpenPlay	None	RightFoot	SavedShot	[illegible]	[illegible]	[illegible]	
8	81	629	327	72	OpenPlay	Cross	RightFoot	MissedShots	[illegible]	0.89	[illegible]	
9	81	629	0	76	DirectFreekick	Standard	RightFoot	MissedShots	[illegible]	[illegible]	[illegible]	
10	81	646	647	4	OpenPlay	Chipped	LeftFoot	MissedShots	[illegible]	[illegible]	0.61	
11	81	647	646	11	OpenPlay	Pass	RightFoot	BlockedShot	[illegible]	[illegible]	[illegible]	
12	81	628	646	11	FromCorner	Cross	Head	BlockedShot	[illegible]	[illegible]	[illegible]	
13	81	647	0	11	FromCorner	Rebound	RightFoot	MissedShots	[illegible]	[illegible]	[illegible]	
14	81	628	646	17	OpenPlay	Chipped	RightFoot	SavedShot	[illegible]	[illegible]	[illegible]	
15	81	628	0	21	OpenPlay	Tackle	LeftFoot	ShotOnPost	0	[illegible]	[illegible]	
16	81	629	0	46	SetPiece	None	RightFoot	SavedShot	[illegible]	[illegible]	[illegible]	
17	81	646	629	83	OpenPlay	Chipped	LeftFoot	SavedShot	[illegible]	[illegible]	[illegible]	
18	81	647	646	87	SetPiece	Cross	Head	BlockedShot	[illegible]	[illegible]	[illegible]	
19	81	646	644	88	OpenPlay	Pass	RightFoot	SavedShot	[illegible]	[illegible]	[illegible]	
20	82	466	456	19	OpenPlay	Aerial	Head	MissedShots	[illegible]	[illegible]	[illegible]	
21	82	461	465	34	OpenPlay	Pass	LeftFoot	MissedShots	[illegible]	[illegible]	[illegible]	
22	82	465	460	39	OpenPlay	Pass	RightFoot	BlockedShot	[illegible]	[illegible]	[illegible]	
23	82	460	461	32	OpenPlay	Pass	LeftFoot	MissedShots	[illegible]	[illegible]	[illegible]	
24	82	462	0	36	OpenPlay	Dispossessed	RightFoot	MissedShots	[illegible]	[illegible]	[illegible]	
25	82	466	460	40	OpenPlay	Throughball	RightFoot	SavedShot	[illegible]	[illegible]	[illegible]	
26	82	465	460	44	OpenPlay	Pass	RightFoot	MissedShots	[illegible]	[illegible]	[illegible]	

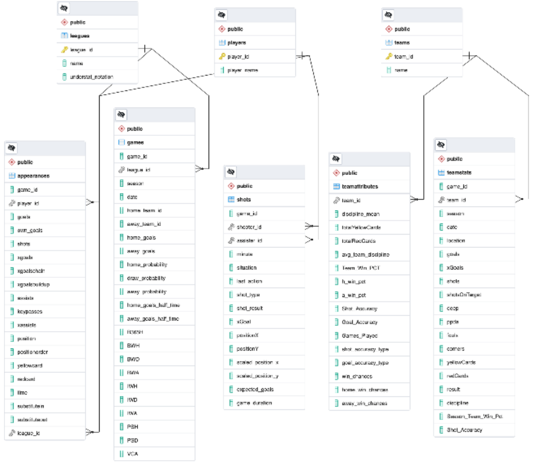


Figure 3: Data Structure and Schema Diagram for all the datasets In PostgreSQL

III. Results

Below are the results obtained from the analysis and visualizations performed.

The visualizations were performed on the following collection of datasets that was taken from PostgreSQL Schema:

1. Player Appearance and players datasets
2. Games and Leagues datasets
3. Shots and Players datasets
4. Team Stats, Team Attributes, Teams Datasets

Data Visualizations For the Player Appearance and Players Datasets (Nishant Bharti)

1. **Player Appearance and players datasets**

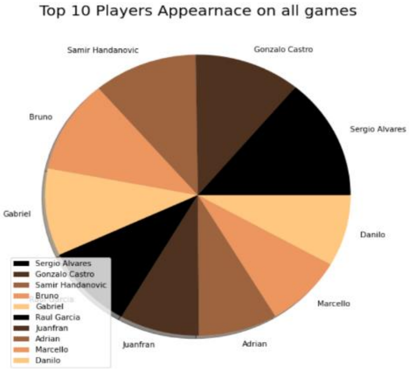


Figure 4 : Top 10 Player Appearances

In above figure, shows the top ten players appeared the most in all five league games.

2. **Goals scored by Player:**

In below scatterplots, first figure is majorly focused on more than three goals scored by players and bar-graph in same figure. And in second plot, the distribution is on basis of assists made by players to score more than three goal. To make plot more discrete and get helpful insight.

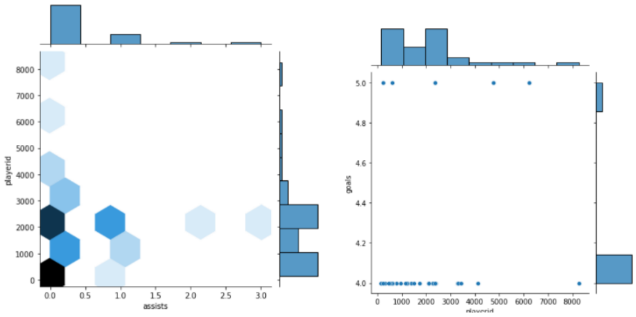


Figure 5: Assists and Goals made by various players

3. **Top Five Forward players to score goals, assists and keypasses and expected goals / expected goals for every possession:**

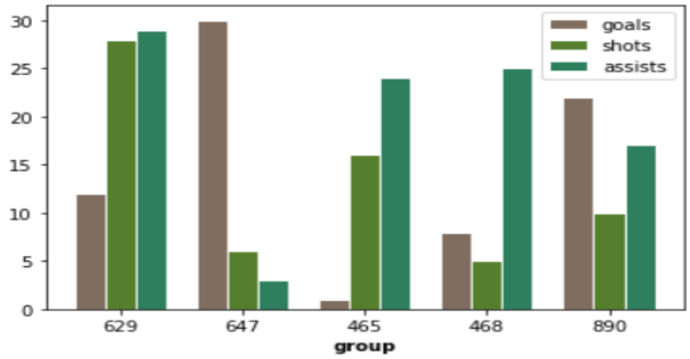


Figure 6: Shots, Assists and Goals Scored by Top 5 players for various players

Above figure, show the expected goals scored by a player for every shot attempted and the consolidated bar graph shows top five player goals scored, assists and shots from *forward* position.

4. **Players Distribution across all major attributes i.e. keypasses, goals scored, expected goals assists:**

The three figures show the overall distribution of player across shots attempted, keypasses made, expected goals (xGoals), expected goals for every posession (xGoalsChain) and expected goals without key passes and shots (xGoalsBuildup). This shows the best player overall in all aspects of the game.

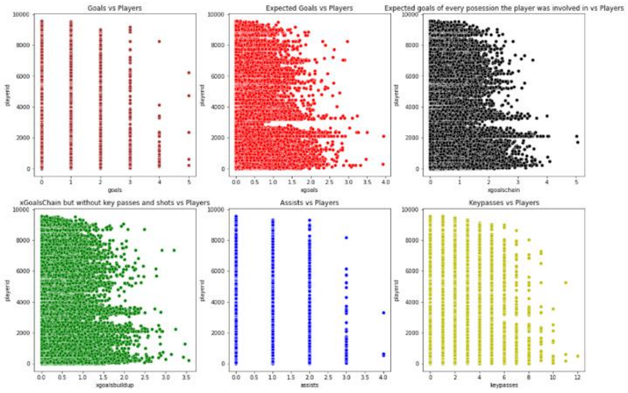


Figure 7: Players Distribution across all major attributes

5. **Goals Score Analysis on the basis of keypasses and postions:**

The below two bar plots is to show maximum number of goals were scored from *forward* Position and in second and third best goal scored by player were forward right wing and forward left wing position.

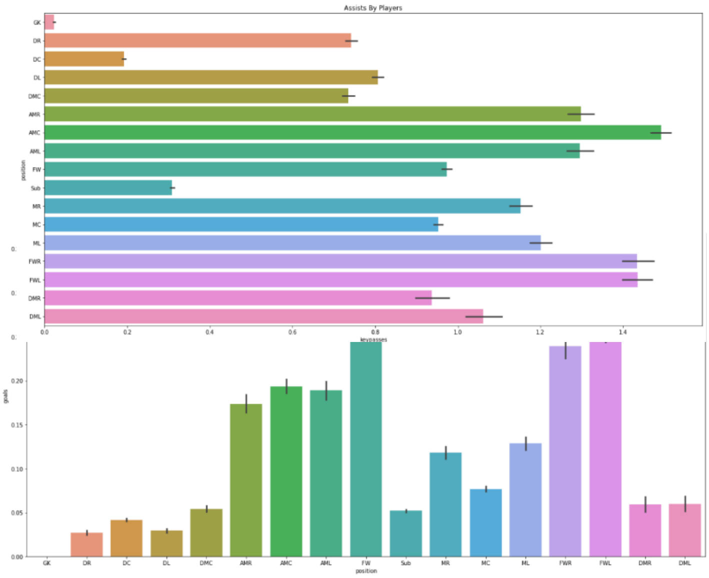
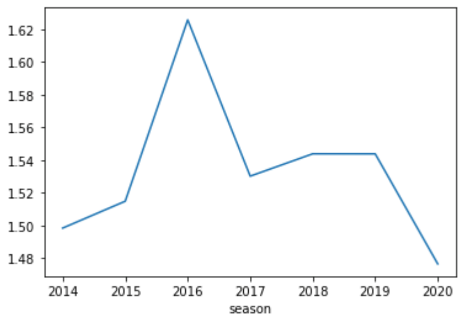


Figure 8 : Goals Score Analysis

Data Visualizations For the Games and Leagues Datasets (Taranjyot Singh):

1. The below chart shows the no of average goals which were made in each season starting from 2014 to 2020. Based on the data, we took the average mean of the home goals for each season.



.

Fig.9- Average home goals per season

2. The below chart shows the bargraph of the away goals which were taken by calculating mean based on the particular league.

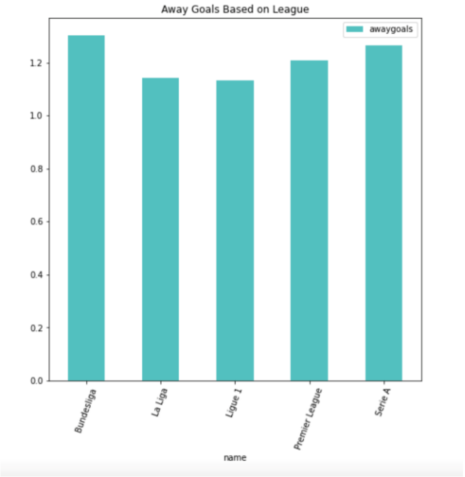


Fig 10: AwayGoals Based On League

3. The below figures shows the plot of both away goals and home goals in each league. The values were calculated by taking the mean of the away and home goals in each league.

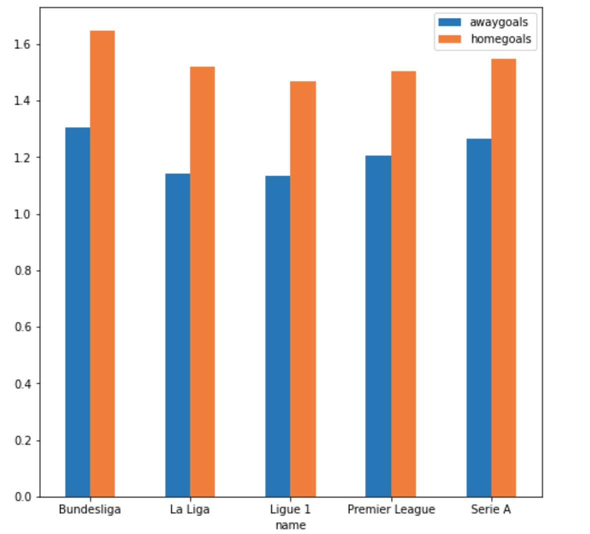


Fig 11: Away and home Goals in each league

4. The below figure shows both away and home goals till the half time. The values were calculated by taking mean or average of the goals scored till the half time.

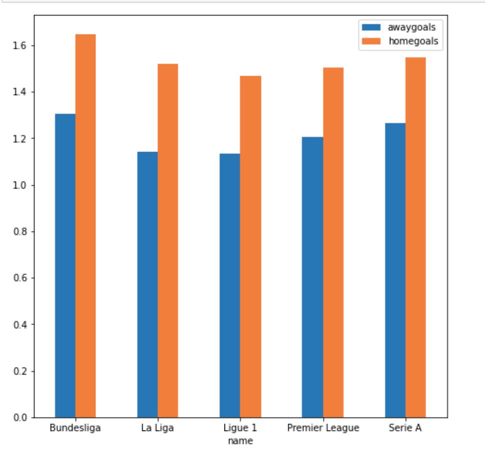


Fig.12. Away and Home Goals Till Half Time

5. The probability was given in the dataset and the pie chart was prepared by taking the average of the home win probability of each league.

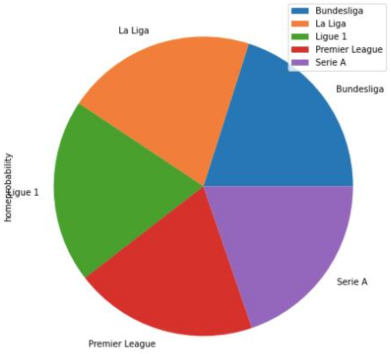


Fig 13: Home Win Probability of Leagues

Data Visualizations For the Shots and Players Datasets (Gaurav Singh):

1. The below chart shows the category of shots distribution for all the players

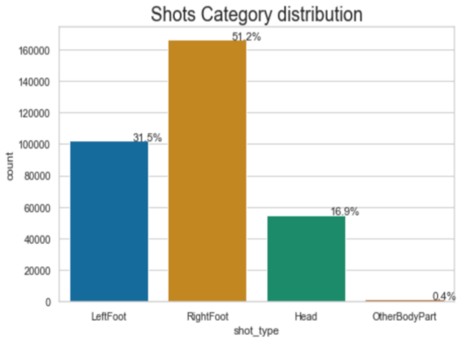


Fig 14: Shots Category Distribution

Conclusion: Most shots were taken by right footed players compared to left footed. 51.2% shots were taken by Right Footed players compared to 31.5% left footed. 17% shots were headers and remaining other body parts

2. The below chart shows the Shot result Distribution for all the players

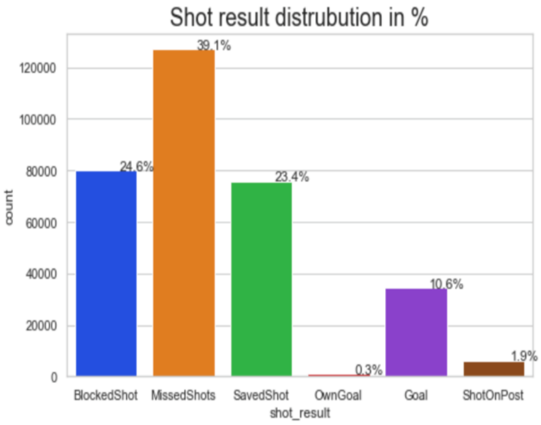


Fig 15: Shot result Distribution

Conclusion: Out of all the shots taken, only 11% shots were converted to Goals (Goals and Own Goals), almost 39.1% of shots were missed shots and 24.6% were blocked shots and 23.4 were blocked shots due to Goalkeeper Brilliance

3. The below chart shows the highest number of goals scored by various players between 2014-20 season

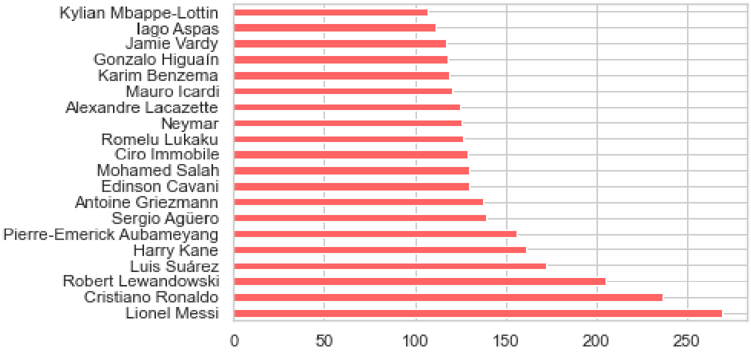


Fig 16: Highest number of goals

Conclusion: As we can see Lionel Messi has the most incredible goal scoring record followed by Cristiano Ronaldo. As both of them are considered the best footballer by their fans we can say both of them have done justice to their fans during 2014 – 2020 period.

Also, Robert Lewandowsky is not far behind and with the rising French star Kylian Mbappe who is about to dominate the goals scoring charts for the next decade.

4. The below chart shows the Overall shots which were converted to goals

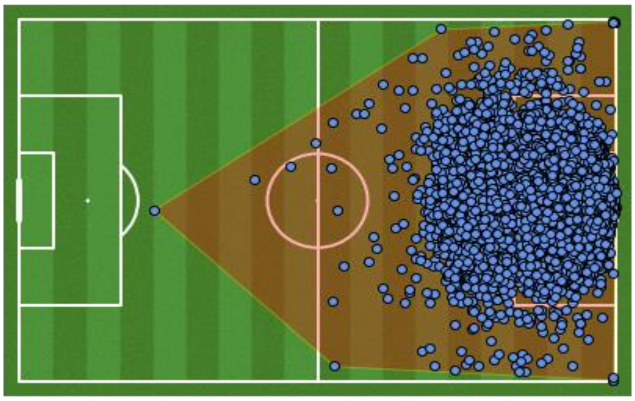


Fig 17: Overall shots which converted to shots

Conclusion : Most shots have been taken in penalty area and very few no of shots are outside of box (penalty area). It’s too difficult to opponent defence in their half.

5. The below Chart shows the shots distribution of Cristiano Ronaldo and Lionel Messi

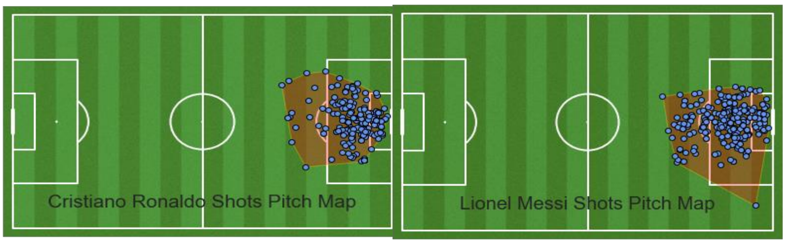
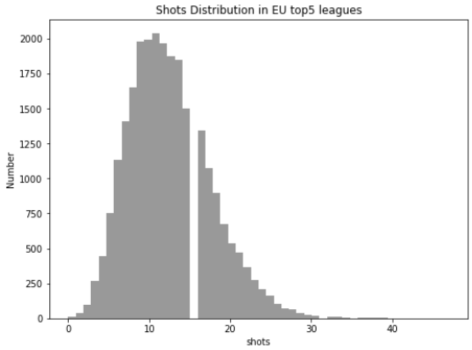


Fig 18: Shots Distribution for Cristiano Ronaldo and Lionel Messi

Data Visualizations For the Teams Stats, Team Attributes and Teams Datasets (Vishan Lal):

1. The below chart shows the distribution of shots and checking the correlation of shots on target with the expected goals for the top 5 leagues in Europe.



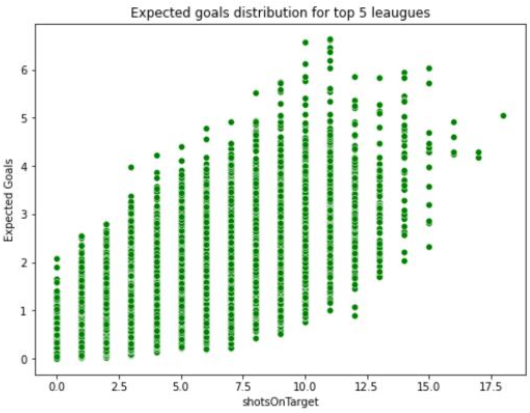


Fig 19: Shots Distribution and Expected goals distribution for the top 5 leagues in Europe

Conclusion: The better the shots on targets are hit the more is the expectancy of the goal to be scored

2. Correlation Scatter Plot to showcase the impact of discipline on the home and away winning percentages. Here, greater the number in discipline, resembles the poor is the discipline of a team.

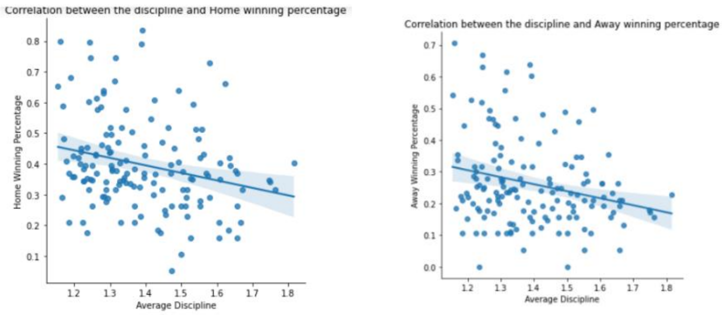


Fig 20: Scatter plots for Home and away Wins with respect to discipline

Conclusion: Discipline has a significant impact in winning the game for both, home and away sides.

3. The below pie chart is showcasing the distribution of the impact of discipline in winning a game for all the teams in the top 5 leagues in Euorpe between the years 2014-20.

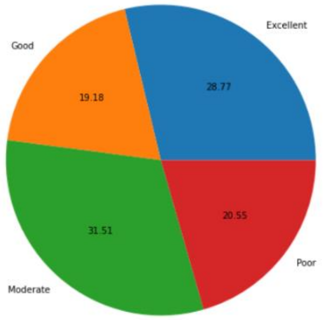


Fig 21: Pie chart distribution for the discipline of all the teams

Conclusion: From the above pie chart it can be inferred that most of the teams have a moderate discipline and there is not much significant difference between the other discipline categories for all the teams participating in the top 5 euro leagues

4. The below chart is to check the impact of number of games played in the top 5 Leagues of Europe with their shot accuracy
   a. For the teams who have excellent shot accuracy vs the Total number of games played

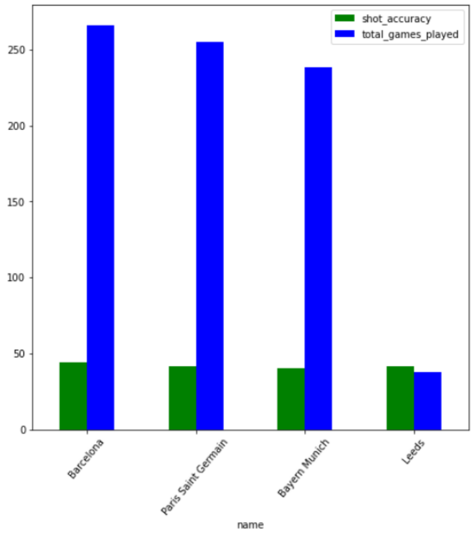


FIG 22: SHOT ACCURACY VS THE TOTAL GAMES PLAYED FOR THE 'EXCELLENT' SHOT CATEGORY GROUP

   b. For the teams who have poorest shot accuracy vs the Total number of games played

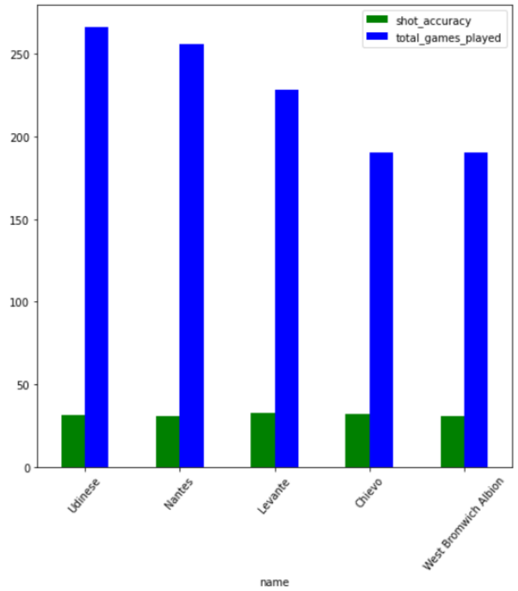


Fig 23: Shot Accuracy vs the total games played for the 'poor' shot category group

Conclusion: It can be inferred from the fig 22 and 23 that number of games played doesn't have a significant impact on the improvement of the shot accuracy as the number of games played for the poor shot category group of teams is the same as excellent shot category group of teams.

5. The Below set of charts explains the total number of goals that have been scored for each discipline category over the years 2014-2020. The second chart explains the conversion rate of corners to the goals for all the teams.

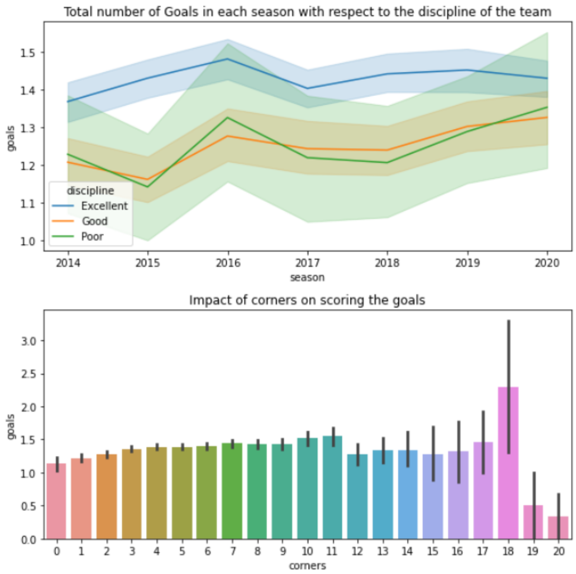


FIG 24: GOALS PATTERN OVER THE YEARS AND IMPACT OF CORNERS ON THE GOALS

Conclusion: The better the discipline of the team the more are the chances to hit a goal.

6. The below chart shows the Comparative analysis of the performance of top 10 home winning teams vs with their performance in the away side

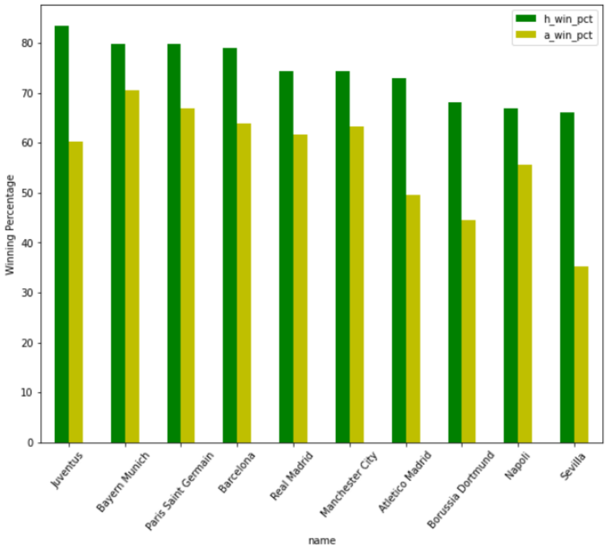


Fig 25: Analysis of top 10 home winning team and their comparison with the away performance

Conclusion: The home winning rates are always higher than away winning rates and the top 10 home winning team does not resemble the consistent 'away' winning rates as well. From the above chart it can be seen that Juventius is better than Bayern Munich at home but the trend reverses in the away matches for both the teams.

7. The below K-Means Clustered plot showcases the relation between shot and goal accuracy.

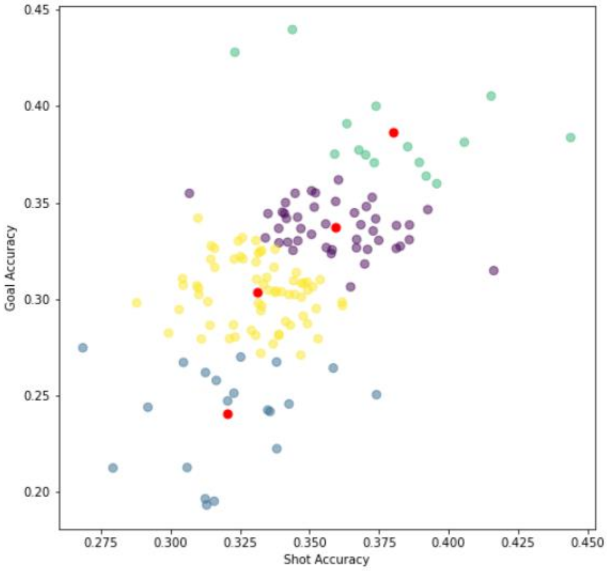


FIG: 26: K-MEANS CLUSTERRED PLOT FOR THE SHOT VS THE GOAL ACCURACY

Conclusion: There is a very positive correlation seen between the Shot accuracy and the goal accuracy and the similar set of points have been clustered to have abetter understanding of the data points.

IV. CONCLUSION

This study presented a data science strategy for selecting players and forming a dream squad from different countries and clubs based on criteria such as Skill Moves, Pace, Shooting, Strength, Stamina, Physics, Passing, Movement, Mentality, and other football-related characteristics. Such methodologies and empirical data may aid in the formation of a dream team from a group of selected players.

V. FUTURE SCOPE

- One of the future scope of this project could be to do the budget analysis for various teams and select the optimal team which suits the wages budget for a team
- Provide an Optimal analysis on the role distribution of Attacker vs the Midfielder vs the Defender
- Demonstrate an analysis of the Player's wages based on the players performance and the age group of that player.

REFERENCES

[1] Çoban, Oktay; Baykan, Erol; Gürkan, Oguz; Yildirim, Mehmet, “The Analysis of Football Players' Percentages of Shot on Target and Levels of Self-Confidence in Different Leagues” Sep 2020.

[2] ALEX RATHKE, “An examination of expected goals and shot efficiency in soccer” , 6th ISPAS International Workshop, 22-23 March 2016. International Society of Performance Analysis of Sport. Carlow, Ireland.2016.

[3] Pappalardo, Luca & Cintia, Paolo & Ferragina, Paolo &Massucco, Emanuele & Pedreschi, Dino & Giannotti, Fosca. (2019). PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach. ACM Transactions on Intelligent Systems and Technology. 10. 1-27. 10.1145/3343172.

[4] https://pandas.pydata.org/

[5] https://www.psycopg.org/docs/usage.html