Chapter 1: Introduction

1.1 Background

Europe's top soccer/football leagues are renowned for their global appeal and competition. Data analysis of these leagues helps identify trends in these leagues and helps understand European soccer's competitive landscape. Since soccer is the most popular sport in the world, soccer fans have traditionally turned to European soccer, and these leagues constantly make improvements to better serve their global fan base. England, Italy, Spain, and Germany leagues have millions of followers. And, the other nations like France, Portugal, Russia, Netherlands and other several countries have been investing heavily to increase their global appeal. In this analysis, the leagues English Premier League from England, Seria A from Italy, La Liga from Spain, Bundesliga from Germany, League One from France, and Russian Premier League from Russia is studied to understand each league and make comparison side by side and analyze the differences between the leagues.

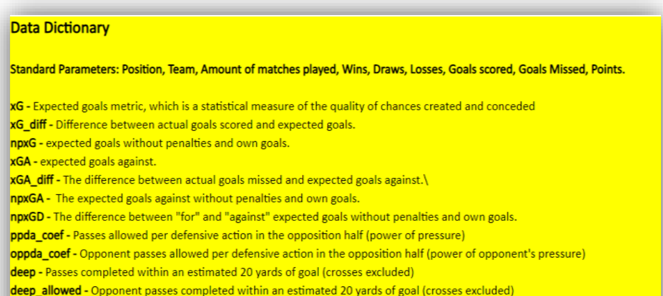
1.2 Importance

This analysis provides valuable insights to various stakeholders involved in soccer, such as leagues, clubs, fans, and businesses related to the sport. Some key importance of the analysis are;

1. It aids in enhancing the overall quality of soccer leagues by **offering strategic guidance and performance evaluations**.
2. This information **benefits businesses that operate in the soccer industry**, including betting platforms, fantasy league games, and other related enterprises.
3. Soccer fans can **deepen their understanding of leagues, players, and competitions** through the analysis, leading to a more enriched viewing experience.
4. Clubs can utilize the findings to make **informed decisions on player selection** and navigate the complex transfer market with more confidence, as through this analysis clubs can **find the trends among the winning teams in their respective league.**

Chapter 2: Dataset Overview

The dataset used in the analysis is sourced from Kaggle. This dataset covers information spanning from 2014 to 2019, offering insights into the performance of various soccer teams during these years. While the dataset contains data for all 20 teams in each league for each year, our analysis will specifically concentrate on the top four teams in each season. This focus allows us to delve into the detailed performance metrics of the most successful teams in the European leagues. The simple metrics to understand football are goals scored, goals conceded and so on. But additional metrics like Expected goals also known as xG and total passes completion on


Data Dictionary

Standard Parameters: Position, Team, Amount of matches played, Wins, Draws, Losses, Goals scored, Goals Missed, Points.

xG - Expected goals metric, which is a statistical measure of the quality of chances created and conceded
xG_diff - Difference between actual goals scored and expected goals.
npxG - expected goals without penalties and own goals.
xGA - expected goals against.
xGA_diff - The difference between actual goals missed and expected goals against.\
npxGA - The expected goals against without penalties and own goals.
npxGD - The difference between "for" and "against" expected goals without penalties and own goals.
ppda_coef - Passes allowed per defensive action in the opposition half (power of pressure)
oppda_coef - Opponent passes allowed per defensive action in the opposition half (power of opponent's pressure)
deep - Passes completed within an estimated 20 yards of goal (crosses excluded)
deep_allowed - Opponent passes completed within an estimated 20 yards of goal (crosses excluded)

Figure 1- Data Dictionary

the opposite half (ppda Coefficient) helps understand the league in much depth. For the analysis, short forms are used, in these long football metric. To better understand the dataset and its contents, refer to the data dictionary presented in the given figure.

Chapter 3: Data Techniques and Libraries

This analysis effectively a range of libraries and techniques for the data analysis. Some of the key libraries and functions used can be listed as;

- **Pandas' library** was used for data manipulation and analysis.
- **Numpy** was used for numerical analysis.
- **Matplotlib and Seaborn library** are used for data visualization.
- The analysis also uses extensive functions like read function for loading data, info for datatypes and null, heads and tails for viewing data, and various other function to replace headers, check duplicates, identify outliers and so on, which is shown in the analysis part below.

Chapter 4. Data Analysis

This analysis covers preparing data for analysis. It includes loading, cleaning, verifying, and exploring the data. Each step are discussed further in the following section.

4.1 Data Loading

Data is mounted and accessed on Google drive for analysis. Before cleaning the data, pandas are installed, which can be seen in the screenshot below;

```
#Mount the raw data to googledrive
#Data from Kaggle https://www.kaggle.com/datasets/slehkyi/extended-football-stats-for-european-leagues-xg
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
[120] #import pandas library for data analysis
#import csv file for analysis
import pandas as pd
import csv
csv_file_path = '/content/drive/MyDrive/43031/raw_data_python/understat.com.csv'
data = pd.read_csv(csv_file_path)
# Display the first few rows of the dataframe
data.head()
```

	Unnamed: 0	Unnamed: 1	position	team	matches	wins	draws	loses	scored	missed	...	xGA	xGA_diff	npxGA	npxGD	ppda_coef	oppda_coef	deep	deep_allowed
0	La_liga	2014	1	Barcelona	38	30	4	4	110	21	...	28.444293	7.444293	24.727907	73.049305	5.683535	16.367593	489	114
1	La_liga	2014	2	Real Madrid	38	30	2	6	118	38	...	42.607198	4.607198	38.890805	47.213090	10.209085	12.929510	351	153
2	La_liga	2014	3	Atletico Madrid	38	23	9	6	67	29	...	29.069107	0.069107	26.839271	25.748737	8.982028	9.237091	197	123
3	La_liga	2014	4	Valencia	38	22	11	5	70	32	...	39.392572	7.392572	33.446477	16.257501	8.709827	7.870225	203	172
4	La_liga	2014	5	Sevilla	38	23	7	8	71	45	...	47.862742	2.862742	41.916529	20.178070	8.276148	9.477805	305	168

5 rows × 24 columns

✓ 0s completed at 12:14 AM

Figure 2- Data Mount and Loading

4.2 Data Cleaning and Verification

4.2.1 Null Values

The function .info is used for null values, and datatype observation in the analysis. There were **no missing values** present in the data. However, *if in case there was one, it would have been filled with the help of statistical measurement or with constant measures or expert advice based on the context of the data, how much and which data are missing.*

```
#   Column        Non-Null Count  Dtype
--- ------        --------------  -----
0   Unnamed: 0    684 non-null    object
1   Unnamed: 1    684 non-null    int64
2   position      684 non-null    int64
3   team          684 non-null    object
4   matches       684 non-null    int64
5   wins          684 non-null    int64
6   draws         684 non-null    int64
7   loses         684 non-null    int64
8   scored        684 non-null    int64
9   missed        684 non-null    int64
10  pts           684 non-null    int64
11  xG            684 non-null    float64
12  xG_diff       684 non-null    float64
13  npxG          684 non-null    float64
14  xGA           684 non-null    float64
15  xGA_diff      684 non-null    float64
16  npxGA         684 non-null    float64
17  npxGD         684 non-null    float64
18  ppda_coef     684 non-null    float64
19  oppda_coef    684 non-null    float64
20  deep          684 non-null    int64
21  deep_allowed  684 non-null    int64
22  xpts          684 non-null    float64
23  xpts diff     684 non-null    float64
```

Figure 3- Null and Datatypes

4.2.2 Data Transformation

- **Change of string** data information to **lowercase** for consistency using *str.**lower*** function
- **Adding headers** for the unnamed headers using ***.rename*** function.
- **Change of data type** of numerical data to date type using *pd.to_datetime* function
- **Duplicate Data** using. duplicated function. But none was found.
- **Filter** data using **numerical operator** only top four teams is shown.
- ***.head*** and ***.tail*** was used to show the **first and last part** of the dataset
- **Outliers**: During, data cleaning, some to identify outliers, box plot was used, which did help in uncovering which gave such outputs, such as

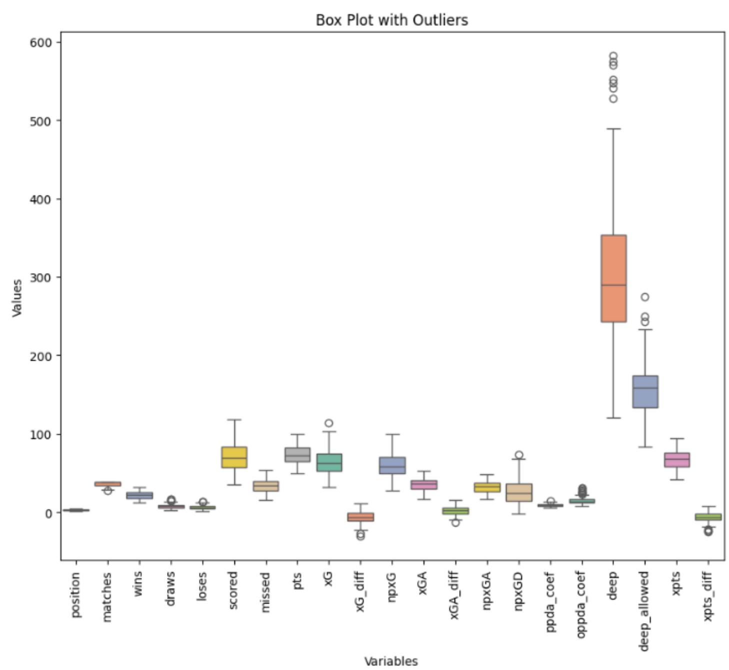


Figure 4- Box Plot for Outlier

In this way, outliers were observed on variables like losses, wins, xG, xGA_diff and a few others. However, handling such outlier depends on the variables are dependent on other factors.

- **Formatting**: At last, the floats is formatted to **2 decimal places**, which gives the following output at before data exploration

This is the final dataset showcasing the first four rows set after cleaning. Using this dataset further data exploration is conducted.

```
    league  year  position            team  matches  wins  draws  loses  \
0   la_liga  2014         1       barcelona       38    30      4      4
1   la_liga  2014         2     real madrid       38    30      2      6
2   la_liga  2014         3 atletico madrid       38    23      9      6
3   la_liga  2014         4        valencia       38    22     11      5
20  la_liga  2015         1       barcelona       38    29      4      5

    scored  missed  ...    xGA  xGA_diff  npxGA  npxGD  ppda_coef  oppda_coef  \
0      110      21  ...  28.44      7.44  24.73  73.05       5.68       16.37
1      118      38  ...  42.61      4.61  38.89  47.21      10.21       12.93
2       67      29  ...  29.07      0.07  26.84  25.75       8.98        9.24
3       70      32  ...  39.39      7.39  33.45  16.26       8.71        7.87
20     112      29  ...  34.03      5.03  33.29  66.19       6.01       15.06

    deep  deep_allowed   xpts  xpts_diff
0    489           114  94.08       0.08
1    351           153  81.75     -10.25
2    197           123  73.14      -4.86
3    203           172  63.71     -13.29
20   570           163  94.38       3.38

[5 rows x 24 columns]
```

Figure 5- Clean Dataset (First Four Rows)

4.3 Exploratory Analysis

Three major analysis is done in this part, first the sum and average of each numerical values, second the statistical summary of each value, and third the correlation analysis.

- Sum and average of numerical variables using .sum and .mean function
- .describe was used for summary statistics.
- .corr was used for correlation analysis

However, to enhance the exploratory analysis, visualization is used, which provides us with the following output

Sum and Average of Metrics

Metrics	Sum	Average
position	360	2.5
matches	5143	35.72
wins	3158	21.93
draws	1103	7.66
loses	882	6.12
scored	10201	70.84
missed	4793	33.28
pts	10577	73.45
xG	9267.24	64.36
xG_diff	-933.76	-6.48
npxG	8506.17	59.07
xGA	5083.39	35.3
xGA_diff	290.39	2.02
npxGA	4633	32.17
npxGD	3873.26	26.9
ppda_coef	1294.25	8.99
oppda_coef	2082.16	14.46
deep	44216	307.06
deep_allowed	22503	156.27
xpts	9688.27	67.28
xpts_diff	-888.74	-6.17

Figure 6-Sum and Average of Metrics

Summary statistics

	wins	draws	loses	scored	missed	pts	xG	xG_diff	npxG	xGA	xGA_diff	npxGA	npxGD	ppda_coef	oppda_coef	deep	deep_allowed	xpts	xpts_diff
count	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00	144.00
mean	21.93	7.66	6.12	70.84	33.28	73.45	64.36	-6.48	59.07	35.30	2.02	32.17	26.90	8.99	14.46	307.06	156.27	67.28	-6.17
std	4.59	2.93	2.59	18.41	8.42	12.60	16.20	7.70	15.36	7.65	5.50	7.14	16.03	1.65	4.45	97.14	34.82	11.98	6.61
min	12.00	2.00	1.00	35.00	15.00	49.00	31.33	-30.96	27.45	16.84	-12.91	16.08	-2.48	5.68	7.35	121.00	83.00	41.18	-24.72
25%	18.00	6.00	4.00	57.00	27.00	64.75	52.65	-11.02	49.28	29.12	-1.80	26.10	14.68	7.84	11.59	243.00	133.75	58.32	-9.57
50%	22.00	8.00	6.00	69.00	33.50	72.00	62.92	-6.53	57.94	36.44	2.50	32.95	24.35	8.93	13.50	290.00	159.00	68.38	-6.18
75%	25.00	9.00	7.25	83.00	39.00	82.00	74.55	-1.38	69.60	41.04	5.38	36.97	36.33	10.02	16.08	353.25	174.25	75.47	-2.03
max	32.00	16.00	13.00	118.00	54.00	100.00	113.60	10.88	99.48	52.33	15.54	47.85	73.05	14.56	30.47	582.00	275.00	94.38	7.49

Figure 7 Summary Statistics

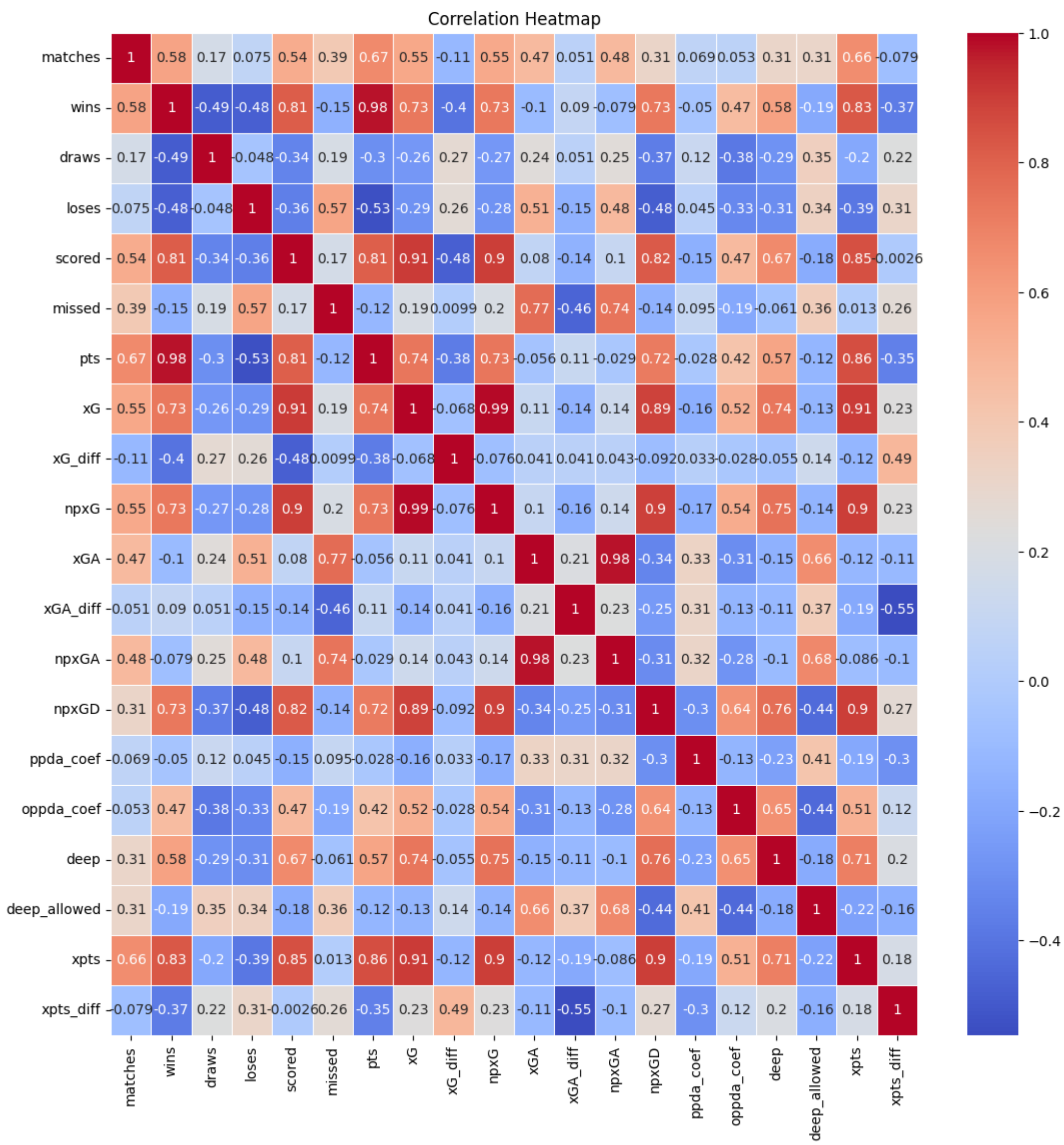


Figure 8- Correlation Heatmap

Chapter 5. Conclusion

During the analysis, important steps such as data cleaning, verification, and exploration were conducted. Here are some key findings from the analysis process:

- No duplicate entries or missing values were identified in the dataset.
- Outliers were detected and visually represented using a box plot.
- A table showcasing summary statistics, correlation matrix, and the sum and average of metrics was created during the exploration phase.

These findings provide a solid foundation for further analysis of the data.