RESEARCH METHODS – Data Representation for Data Understanding

COURSEWORK SOLUTIONS

Q1.

Football is widely considered the most popular sport in the world. With a staggering estimated fanbase of over 4 billion as of 2023, football has remained the centre focus for many sports fanatics worldwide (SportyTell Editors, 2023).

Football analytics and statistics have become an integral part of the football game, with clubs investing heavily in devices like trackers, heart rate monitors, and more, in order to capture data which can be used for decision-making processes within teams. These investments in football analytics haven’t just been geared towards securing devices and equipment for data collection, but the growing need for the right personnel who can take this data and draw insights has become paramount, with increasing data job roles within teams’ data groups.

With the growing need for constant analysis and insight generation in the field of football, my dataset obtained from Kaggle, a popular open-source data community which hosts public datasets, is tailored towards “Analysing Teams’ Success in Europe's Top 5 Football Leagues”. The dataset is titled, “Top 5 European Football Leagues Teams Stats”, and covers stats from the English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A, and the French League 1, between the 2010-2011 season to the 2020-2021 season.

The purpose of this analysis is to examine the factors contributing to the success of football teams in Europe's top five leagues over an eleven-year period. By analysing key performance indicators (KPIs) from the various league seasons, we aim to uncover trends and patterns that may have influenced teams' performances and their ability to win league titles.

Q2.

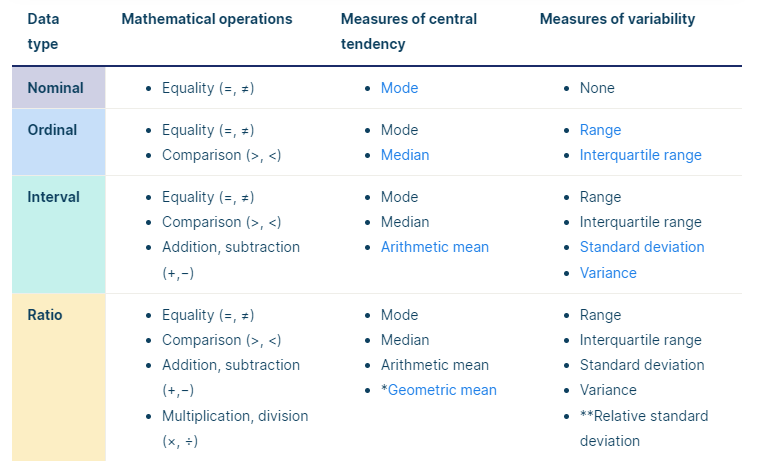
Understanding our data, the variables, and the relationship between each variable, is paramount and should be considered among the first things required by any data professional before data analysis can commence. The dataset we have chosen for this analysis is made up of 1078 rows and 28 columns. The columns in our dataset are regarded as features or variables in data analysis. These variables define new information about the subject of our analysis while the rows indicate a team during a specific season, we have information on which is the subject of analysis.

TABLE 1: Understanding the meaning of the different variables and their respective data types.

|  |  |  |  |
| --- | --- | --- | --- |
| Index | COLUMNS / VARIABLES | MEANING | DATA TYPE |
| 1 | Competition | The name of the competition/league | Categorical - Nominal |
| 2 | Season | The season of football e.g. 2010-2011 | Categorical - Ordinal |
| 3 | Rank | The final position of the club at the end of the season | Categorical - Ordinal |
| 4 | Squad | The name of the club | Categorical - Nominal |
| 5 | Games | Total number of games played by the club | Numeric - Ratio |
| 6 | Wins | Total number of games won by the club | Numeric - Ratio |
| 7 | Draws | Total number of games drawn by the club | Numeric - Ratio |
| 8 | Losses | Total number of games lost by the club | Numeric - Ratio |
| 9 | Goals\_for | Goals scored by the club | Numeric - Ratio |
| 10 | Goals\_against | Goals scored against the club | Numeric - Ratio |
| 11 | Goal\_diff | Goals scored subtracted from goals against the club | Numeric - Interval |
| 12 | Points | Total points accumulated by the club | Numeric - Ratio |
| 13 | Notes | Status of team after season e.g. UEFA Champions League | Categorical - Nominal |
| 14 | Players\_used | Total number of players used throughout the season | Numeric - Ratio |
| 15 | Assists | Total number of assists by the teams’ players | Numeric - Ratio |
| 16 | Pens\_made | Total penalties scored | Numeric - Ratio |
| 17 | Pens\_att | Total penalties given to the club | Numeric - Ratio |
| 18 | Cards\_yellow | Number of yellow cards accumulated by the club | Numeric - Ratio |
| 19 | Cards\_red | Number of red cards accumulated by the club | Numeric - Ratio |
| 20 | Shots\_on\_target\_against | Total shots on target against the club | Numeric - Ratio |
| 21 | Saves | Total number of shots the team goalkeeper catches | Numeric - Ratio |
| 22 | Clean\_sheets | Number of matches without conceding a goal | Numeric - Ratio |
| 23 | Shots\_on\_target | Total shots on target by the club | Numeric - Ratio |
| 24 | Games\_start | Number of games the first eleven play in the league that season e.g. 11 players multiplied by 38 games equals 418 | Numeric - Ratio |
| 25 | Games\_complete | Number of games a team’s first eleven has completed until the final whistle, including any added injury or stoppage time | Numeric - Ratio |
| 26 | Games\_subs | Total number of substitutions used by the club | Numeric - Ratio |
| 27 | Unused\_subs | Total number of unused substitutions by the club | Numeric - Ratio |
| 28 | Points\_per\_match | Average points accumulated by a team per match | Numeric - Ratio |

Our dataset is further broadened to 45 variables after exploratory data analysis. The notes column has a lot of inconsistency, hence, to gain insights, it is broken down into 3 columns; UEFA Champions League, UEFA Europa League, and Relegation which are all of categorical nominal data type with values either Yes or No. There is also the issue of Germany having only 18 teams in the league while the remaining leagues we are making comparisons against have 20 teams. This could introduce potential bias in our analysis. Also, in the 2019-2020 season (COVID season), the French league played a maximum of 28 matches in the league that season. These issues could further introduce bias when carrying out analysis. To fix this, we create standardized columns for specific numeric metrics such as adjusted\_wins, adjusted\_draws, adjusted\_losses, adjusted\_goals\_for, adjusted\_goals\_against, adjusted\_goal\_diff, adjusted\_assists, adjusted\_pens\_made, adjusted\_pens\_att, adjusted\_shots\_on\_target\_against, adjusted\_saves, adjusted\_clean\_sheets, adjusted\_shots\_on\_target, adjusted\_points, and adjusted\_points\_per\_match. These new derived variables all follow the same data type as their parent variable from which they were derived.

The variables are categorized into two data types: Categorical and Numerical data type. We have 3 categorical data as nominal, 2 categorical data as ordinal, 1 numeric data as interval, and 22 numeric data as ratio. These data types allow us define the type of descriptive statistics that can be performed. The table below illustrates possible operations that can be carried out on the data types.



SOURCE: [Scribbr](https://www.scribbr.com/statistics/levels-of-measurement/)

TABLE 2: Descriptive statistics on numerical data type – count, mean, standard deviation, minimum value, 25th percentile, 50th percentile or median, 75th percentile, and maximum value.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | VARIABLES | COUNT | MEAN | STD | MIN | 25% | 50% | 75% | MAX | RANGE |
| 1 | Competition | 1078 | - | - | - | - | - | - | - | - |
| 2 | Season | 1078 | - | - | - | - | - | - | - | 11 |
| 3 | Rank | 1078 | - | - | 1 | 5 | 10 | 15 | 20 | 19 |
| 4 | Squad | 1078 | - | - | - | - | - | - | - | - |
| 5 | Games | 1078 | 37.08 | 2 | 27 | 38 | 38 | 38 | 38 | 11 |
| 6 | Wins | 1078 | 13.87 | 5.86 | 2 | 10 | 12 | 17 | 33 | 31 |
| 7 | Draws | 1078 | 9.34 | 2.99 | 2 | 7 | 9 | 11 | 19 | 17 |
| 8 | Losses | 1078 | 13.87 | 5.36 | 0 | 10 | 14 | 18 | 29 | 29 |
| 9 | Goals\_for | 1078 | 50.61 | 16.77 | 20 | 39 | 47 | 59 | 121 | 101 |
| 10 | Goals\_against | 1078 | 50.61 | 12.68 | 17 | 42 | 51 | 58.75 | 94 | 77 |
| 11 | Goal\_diff | 1078 | 0 | 25.63 | -61 | -17 | -5 | 13.75 | 89 | 150 |
| 12 | Points | 1078 | 50.91 | 16.71 | 13 | 40 | 47 | 61 | 102 | 89 |
| 13 | Notes | 450 | - | - | - | - | - | - | - | - |
| 14 | Players\_used | 1078 | 27.87 | 3.28 | 19 | 25 | 28 | 30 | 42 | 23 |
| 15 | Assists | 1078 | 33.52 | 12.73 | 10 | 25 | 31 | 39 | 91 | 81 |
| 16 | Pens\_made | 1078 | 4.31 | 2.5 | 0 | 3 | 4 | 6 | 15 | 15 |
| 17 | Pens\_att | 1078 | 5.55 | 2.9 | 0 | 3 | 5 | 7 | 20 | 20 |
| 18 | Cards\_yellow | 985 | 48.45 | 40.13 | 0 | 0 | 59 | 79 | 150 | 150 |
| 19 | Cards\_red | 985 | 2.73 | 2.7 | 0 | 0 | 2 | 4 | 12 | 12 |
| 20 | Shots\_on\_target\_against | 1078 | 160.69 | 31.49 | 71 | 139 | 162 | 180.75 | 271 | 200 |
| 21 | Saves | 1078 | 112.13 | 22.68 | 45 | 97 | 111 | 126 | 199 | 154 |
| 22 | Clean\_sheets | 1078 | 10.27 | 3.89 | 2 | 8 | 10 | 13 | 24 | 22 |
| 23 | Shots\_on\_target | 1078 | 164.91 | 40.26 | 77 | 137 | 157 | 186 | 318 | 241 |
| 24 | Games\_start | 1078 | 407.82 | 21.96 | 297 | 418 | 418 | 418 | 419 | 122 |
| 25 | Games\_complete | 158 | 278.59 | 31.64 | 217 | 248.25 | 297 | 305 | 339 | 122 |
| 26 | Games\_subs | 1078 | 109.91 | 16.51 | 71 | 101 | 109 | 113 | 189 | 118 |
| 27 | Unused\_subs | 686 | 190.61 | 62.78 | 109 | 153 | 159 | 208 | 345 | 236 |
| 28 | Points\_per\_match | 1078 | 1.37 | 0.44 | 0.42 | 1.08 | 1.26 | 1.63 | 2.68 | 2.26 |

From table 2, competition, squad, and notes, have values allocated to only count as these are categorical nominal data that do not support operations like mean, standard deviation, min, max, median, percentiles, and range. However, we can consider the mode when looking at categorical nominal data.

TABLE 3: Descriptive statistics on categorical nominal data type – count, and mode.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | VARIABLES | COUNT | MODE | OCCURENCE |
| 1 | Competition | 1078 | Premier League, Ligue 1, La Liga, Serie A | 220 |
| 2 | Squad | 1078 | Manchester Utd, Dortmund, Chelsea, Werder Bremen, Hoffenheim, Mainz 05, Bayern Munich, Leverkusen, Nice, M’Gladbach, Montpellier, Saint-Etienne, Bordeaux, Rennes, Paris S-G, Lyon, Wolfsburg, Milan, Lille, Barcelona, Real Sociedad, Atletico Madrid, Athletic Club, Sevilla, Valencia, Real Madrid, Genoa, Inter, Fiorentina, Juventus, Roma, Lazio, Udinese, Napoli, Marseille, Schalke 04, Manchester City, Everton, Liverpool, Tottenham, Arsenal | 11 |
| 3 | Notes | 450 | 🡪UEFA Champions League via league finish | 193 |

Table 3 reveals some insights that we can consider interesting. When looking at the values that constitute the mode for different teams (Squad), the teams which appear as the mode are the teams who avoided relegation across the top five leagues in Europe over the eleven-year period of our analysis. Over the seasons time span from 2010-2011 to 2020-2021, 41 teams successfully avoided relegation and performed at the highest level of football in their respective national leagues.

Q3.

Our data quality assessment on our dataset covers the following; checking for duplicate entries (data redundancy), checking for missing values in columns and understanding what they mean, checking that the data types for each variable are specified correctly, checking data entry issues, checking the source where the data comes from (data reliability), and checking for invalid values or outliers that may indicate errors in data entry or measurement (data validity). These six checks are conducted to analyse the true quality of the data and draw insights on possible next steps.

Checking for duplicate entries (data redundancy) – The dataset has no duplicate records after data entry. This indicates no data redundancy and our dataset passes this check.

Checking for missing values (data completeness) – After successful initial exploratory data analysis, we derive that our dataset has a total of 2126 missing values. These missing values can be further broken down into the following;

TABLE 4: Summary table for missing values with possible solutions made available.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | VARIABLES | MISSING | % MISSING | SOLUTION |
| 1 | Notes | 623 | 58.26 | Create three columns with yes/no answers. The columns to be created are, UEFA champions league, Europa league, and Relegation. This will allow us capture full insights from the notes column |
| 2 | Cards\_yellow | 93 | 8.63 | More information needs to be gotten from other sources, if possible, to fill the gaps in the cards\_yellow and cards\_red column as we found that all five leagues have no data recorded for these columns in the 2017-2018 season, and Ligue 1 (French league) has missing values from 2013-2014 till 2016-2017 |
| 3 | Cards\_red | 93 | 8.63 |
| 4 | Games\_complete | 920 | 85.34 | Ignored in our analysis as it has over 85% of its values missing |
| 5 | Unused\_subs | 392 | 36.36 | The unused\_subs column has missing values across all five leagues between the years 2010-2011 to 2013-2014 which we can skip in our analysis as possible data regarding this may not have been made available |

Table 4 shows a total of five columns out of 28 columns with missing values. It must however be noted that missing values do not necessarily indicate missing information as this is subject to the point of analysis. The column called notes, with 58.26% missing values indicates teams that were neither participating in the UEFA champions league, Europa league, or relegated. All other forms of missing values need to be properly handled before analysis.

Checking data entry issues – In the squad column, Paris Saint-German is specified as Paris S-G which will be considered a data entry issue as all other names include, PSG, Paris SG, or Paris. The notes column has a lot of data entry issues. The source from which the data is collected didn’t do a good job in properly classifying teams that qualified for UEFA Champions League, UEFA Europa League, and teams that were relegated ate the end of the season. To fix this, we made use of an external source – A website called [FlashFootball](https://www.flashfootball.com/) to collect accurate data pertaining to the notes column. All other columns had no data entry issues.

Checking correct data type specification – All columns except the rank column have their correct data type specified. The column rank is given as numeric, however, this should be properly specified as categorical as this attribute or variable explains the position of teams and cannot be considered for mathematical operations.

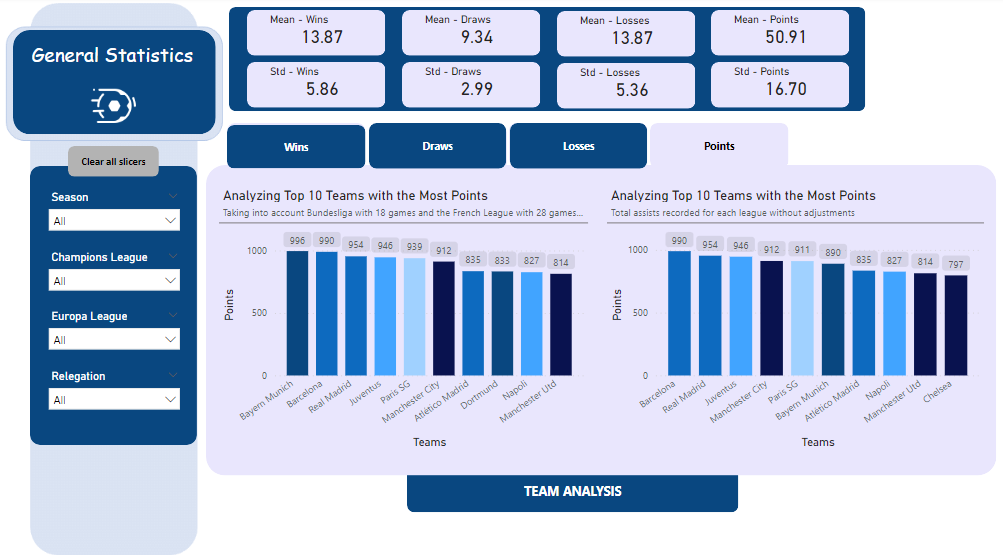
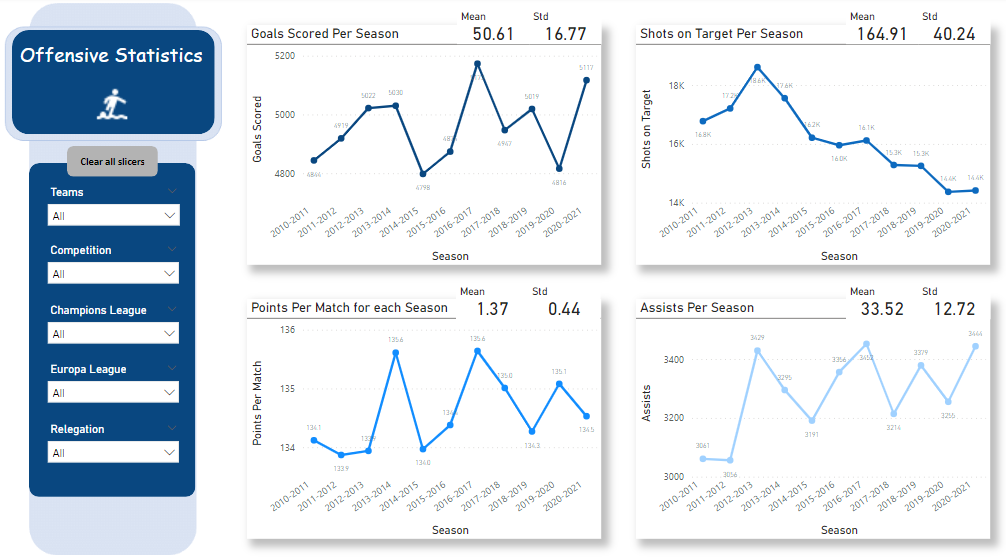
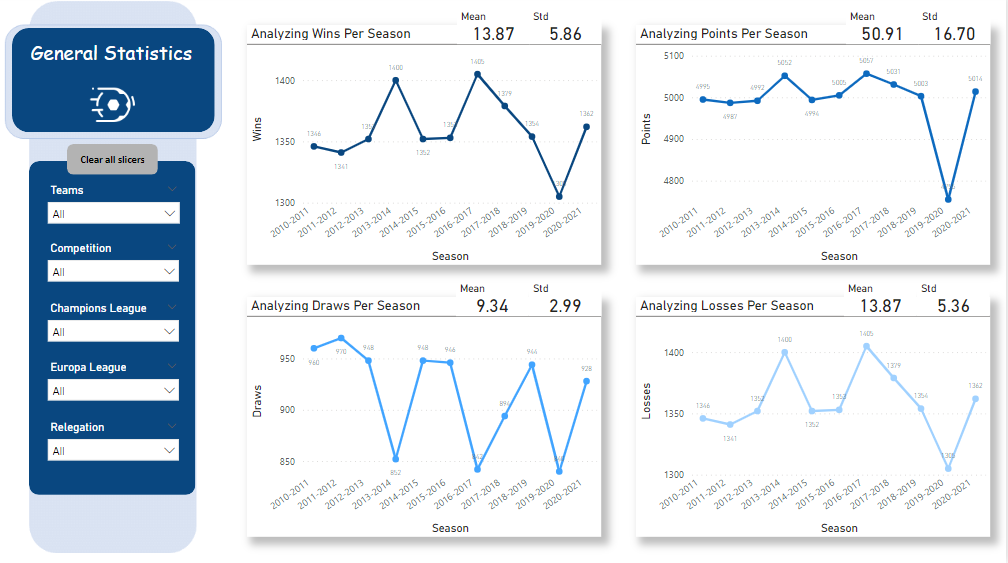
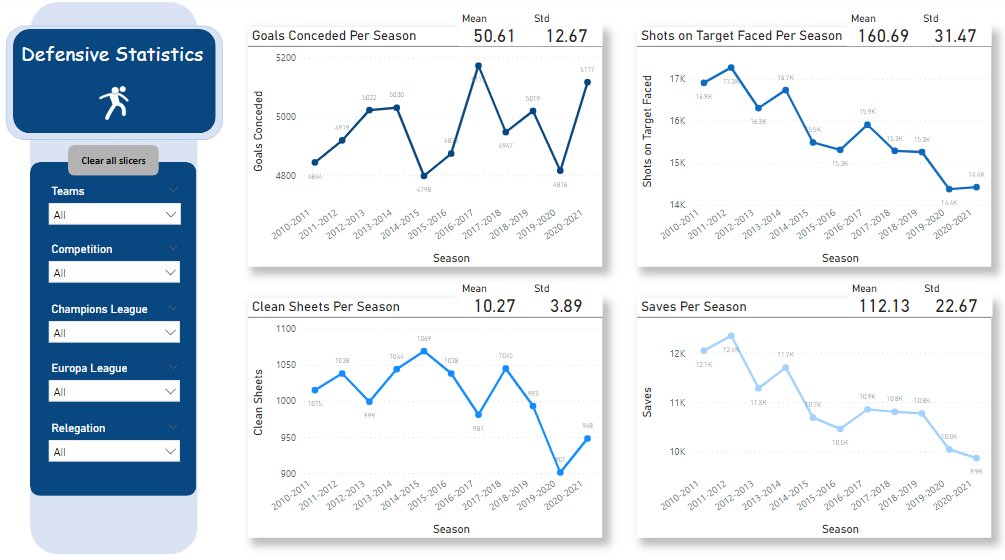
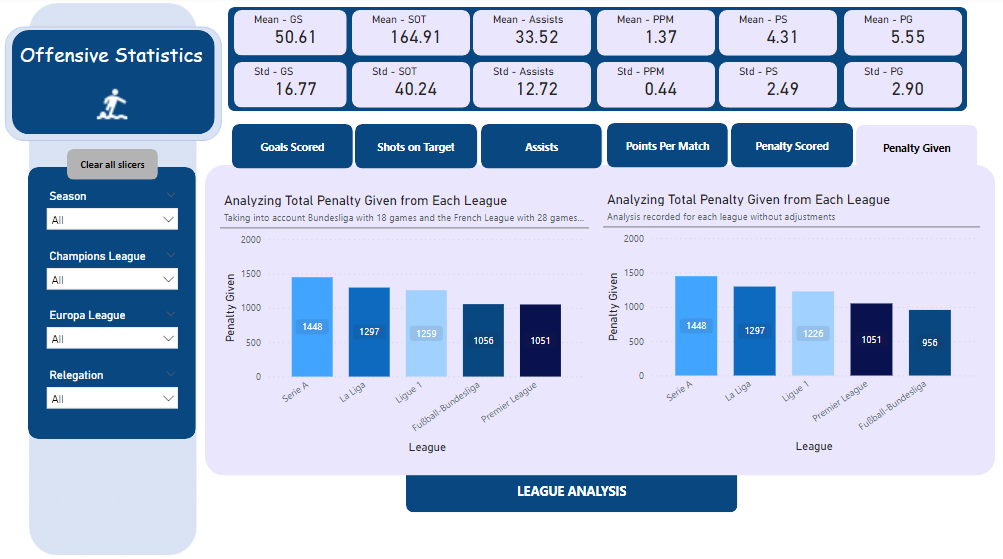
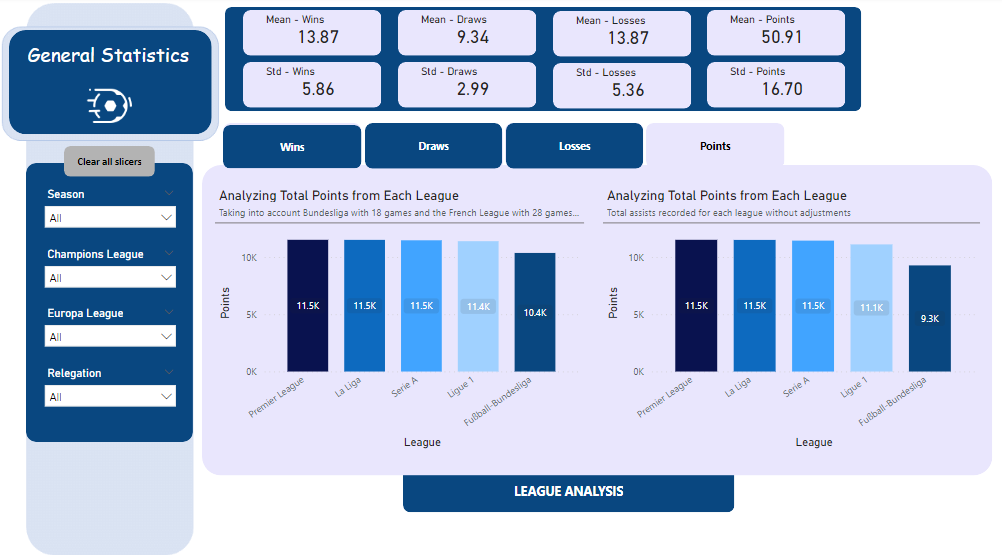
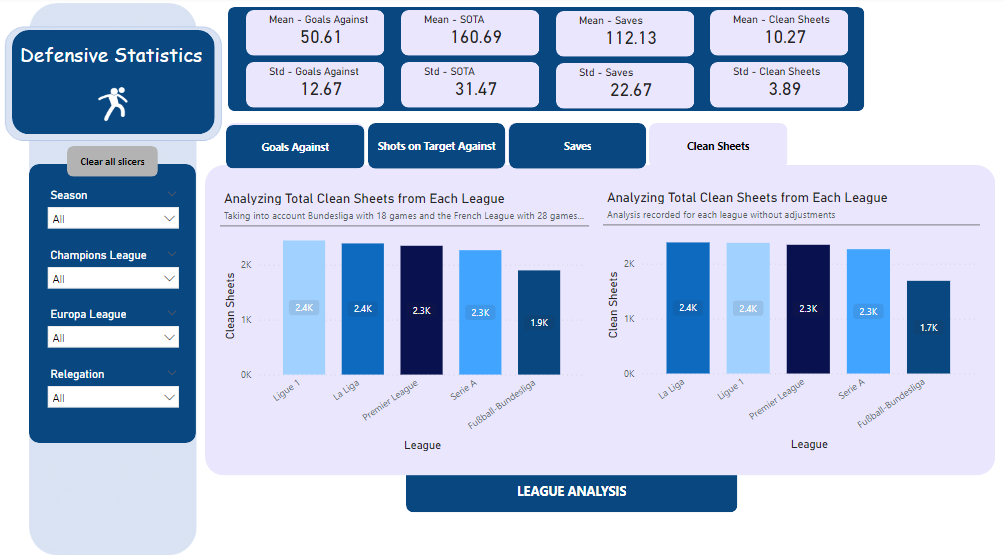
Checking the data source (data reliability) – The data is gotten from a credible source in Kaggle. The dataset itself is properly documented for extracting useful insights. While reliability doesn’t necessarily mean correctness, after exploratory data analysis, we can say a very large portion of the data has little to no issues. The major issues are found in about 4 columns out of 28 columns which can be handled by undergoing proper data preparation and cleaning steps. It’s worth nothing too that the format and structure of the data are well put together.

Checking for invalid values or outliers that may indicate errors in data entry or measurement (data validity) – We conducted exploratory data analysis and no invalid values were found that would raise questions of data validity across any columns.

Note: All data quality checks, data cleaning, and data transformation are done using Python. Visualization is done using Power BI to create dashboards for our analysis.

Q4.   
Data visualization was carried out using [Power BI Desktop](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9). The emphasis is to draw insights on various topics the dataset covers. To handle this, we utilize the creation of dashboards which give us a way a standard to draw insights from variables in our dataset. The creation of dashboards is broken down into 3 broad categories; analyzing teams, analyzing leagues, and seasonal analysis which covers both teams and leagues. Each of these is further broken down into three; general statistics analysis, offensive statistics analysis, and defensive statistics analysis.

Altogether, 9 dashboards are created to aid full getting full insights from our pre-processed dataset. A snapshot the created dashboards can be seen below. While creating the defensive statistics, offensive statistics, and general team statistics, each statistic is compared against an estimated version of that statistic that accounts for the Bundesliga having only 18 teams compared to other leagues with 20 teams and the French league in the 2019-2020 season which saw teams play a maximum of 28 games. Every graph created is followed by its mean and standard deviation. However, it is important to note that for our analysis, only the mean and standard deviation of the actual variables are measured. A breakdown of our major findings in each dashboard can be found below.



When analysing the leagues and teams, a bar chart is used to represent all variables in the general statistics, offensive statistics, and defensive statistics. A bar chart is best employed when making analysis between categories. Given the leagues and teams represent different categories, this justifies our approach of utilizing the bar chart.

Analysis of the various seasons is done using a line chart. The seasons represent activity and events that have happened overtime. Since time is our basis of analysis, then a line chart solves this problem best as it is designed to show trends and provide insights over time thereby justifying our approach. For all seasonal analysis – Season analysis ([General statistics](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)), Season analysis ([Offensive statistics](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)), and Season analysis ([Defensive statistics](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)), a line chart is graphed to draw insights from our data.

In our analysis of the leagues, teams, and various seasons, we make use of other visualisation tools, such as; card – for highlighting the mean and standard deviation, slicers – Helping us view our data from different dimensions, buttons – Not a graph, but used to support our visualization by allowing us switch between pages.

For every general statistics analysis, the variables considered are wins, draws, losses, and points over the eleven-year period of our analysis. While the variables for offensive statistics are goals scored, shots on target, assists, points per match, penalty given, penalty scored. When analysing defensive statistics, we consider, goals conceded, shots on target faced, clean sheets, and saves.

All team analysis – general statistics, offensive statistics, and defensive statistics, taken into account only the top 10 teams for each form of analysis carried out. This allows us define our analysis and streamline our findings.

Our coefficient (formula used) for finding the estimates for all variables used in our analysis, applied to the Bundesliga and the French League 2019-2020 season involves finding the ratio of that variable to the number of actual games played, then multiplying it by 38 games which should be the number of games all teams play at the end of every season.

* [LEAGUE GENERAL STATISTICS](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)

The league general statistics makes comparisons of the various leagues with respect to their wins, draws, losses, and points over the eleven-year period of our analysis. When analysing points without applying any slicers, we see no significant difference between the points accumulated by the different leagues. However, when we apply the slicer – Champions League, which filters to teams which have qualified for the champions league, we find that the Spanish La Liga and English Premier League come in first and second respectively for teams with the most wins and most points over the eleven-year period. ([Full Report and Dashboard](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)).

* [LEAGUE OFFENSIVE STATISTICS](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)

The Serie A and the Premier League are the best leagues when analysing goals scored and shots on target without applying any slicers, however, these margins are not far off from the other leagues. Looking at goals scored and filtering to teams which have qualified for the champion league over the eleven-year period, we find that La Liga and the Premier League perform best respectively but when making these comparisons with the estimated variables, we find that the best performing leagues are the Spanish La Liga and German Bundesliga. These conclusions are the same when analyzing the number of assists from each league. We also find in our analysis that Serie A is the league which concedes the most penalties. ([Full Report and Dashboard](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)).

* [LEAGUE DEFENSIVE STATISTICS](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)

Our analysis of the clean sheets produced in each league reveals Ligue 1 and La Liga have achieved the most clean-sheets over the eleven-year period of our analysis. However, when filtering to teams which have qualified for the champions league within the time frame of our analysis, it is the Premier League and La Liga that have the most clean-sheets. While still analysing the teams who have qualified for the champions league, we observe the goals conceded by these teams in their various leagues; our findings show the La Liga and Premier League have conceded the most goals. When we base this analysis on the estimated variables, it is the Bundesliga and La Liga which have conceded the most goals. ([Full Report and Dashboard](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)).

* [TEAM GENERAL STATISTICS](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9) (Findings from teams who have qualified for champions league)

The top 10 teams with the most wins and most points achieved over the eleven seasons are – Bayern Munich, Barcelona, Real Madrid, Manchester City, Juventus, Paris SG, Dortmund, Atletico Madrid, Chelsea, and Manchester Utd respectively. This analysis is done using the estimated variables to allow us get a glimpse of the true insights from the points gotten by each team. Analysis of teams draws shows the top 10 teams to be – Manchester City, Chelsea, Dortmund, Leverkusen, Real Madrid, Atletico Madrid, Lyon, Manchester Utd, Arsenal, Bayern Munich, and Sevilla respectively.

* [TEAM OFFENSIVE STATISTICS](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9) (Findings from teams who have qualified for champions league)

Barcelona, Bayern Munich, Real Madrid, Manchester City, Paris SG, Dortmund, Juventus, Chelsea, Manchester Utd, and Atletico Madrid respectively are the top 10 teams who have scored the most goals. All findings from our dashboard are based off the estimated variables to allow even and unbiased analysis of the various teams. Analysis of points per match show the top 10 teams to be - Bayern Munich, Barcelona, Real Madrid, Manchester City, Juventus, Paris SG, Dortmund, Atletico Madrid, Chelsea, and Manchester Utd respectively. ([Full Report and Dashboard](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)).

* [TEAM DEFENSIVE STATISTICS](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9) (Findings from teams who have qualified for champions league)

Top 10 teams with the most clean-sheets – Bayern Munich, Manchester City, Paris SG, Barcelona, Atletico Madrid, Juventus, Real Madrid, Chelsea, Dortmund, and Manchester Utd respectively. Across all teams which have qualified for the champions league in the eleven-seasons of our analysis, the top 10 teams with the least goals conceded are – Juventus, Leverkusen, Manchester Utd, Paris SG, Bayern Munich, Chelsea, Barcelona, Manchester City, Real Madrid, and Dortmund respectively.

* [SEASON ANALYSIS](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9) (General Statistics)

The 2019-2020 season (COVID Season) was a season in which the covid-19 virus hit the football world. This is evident in the strong deeps we observe in our graphs. The 2019-2020 season recorded the least wins, draws, losses, and points. These insights are viewed without any filters being applied. The Bundesliga and La Liga recorded its most wins and most points in the 2010-2011 season and its least wins and points in the 2017-2018 season. The La Liga recorded its least wins and points in the 2018-2019 season. The 2019-2020 season in the French league saw them record their least win and least points in any season by a margin. The 2014-2015 saw the Serie A have its least recorded wins and points with the 2016-2017 season recording their best in both departments. The Premier League had its best season in terms of wins and points in the 2018-2019 season, while the 2010-2011 season was their worst in terms of wins and points. ([Full Report and Dashboard](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)).

* [SEASON ANALYSIS](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9) (Offensive Statistics)

Across all leagues, the 2016-2017 recorded the most goals scored, while the 2020-2021 season produced the most assists. The 2012-2013 saw teams produce their highest proficiency when shooting at the post with the highest recorded shots on target.

TABLE 5: Summary table of insights from seasonal offensive analysis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Index | LEAGUE | MOST GOALS | MOST SHOTS ON TARGET | MOST POINTS PER MATCH | LEAST GOALS SCORED | LEAST SHOTS ON TARGET | LEAST POINTS PER MATCH |
| 1 | Bundesliga | 2019-2020 | 2013-2014 | 2010-2011 | 2014-2015 | 2017-2018 | 2014-15 and 2017-18 |
| 2 | La Liga | 2016-2017 | 2011-2012 | 2010-2011 | 2020-2021 | 2020-2021 | 2018-2019 |
| 3 | Premier League | 2018-2019 | 2012-2013 | 2018-2019 | 2014-2015 | 2017-2018 | 2010-2011 |
| 4 | Serie A | 2020-2021 | 2013-2014 | 2016-2017 | 2010-2011 | 2020-2021 | 2014-2015 |
| 5 | Ligue 1 | 2020-2021 | 2012-2013 | 2014-2015 | 2019-2020 | 2019-2020 | 2010-2011 |

From table 5, we see that La Liga didn’t perform well offensively in the 2020-2021 season, recording their least goals scored and least shots on target in any season. It is also worth nothing the 2020-2021 season for Serie A, as they recorded the most goals scored in any season as well as recording the least shots on target that same season showing just how potent they were on the attack that season. ([Full Report and Dashboard](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)).

* [SEASON ANALYSIS](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9) (Defensive Statistics)

The 2014-2015 season is the season with the least goals conceded across all leagues while the 2016-2017 saw teams concede the most goals.

TABLE 6: Summary table of insights from seasonal defensive analysis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Index | LEAGUE | MOST GOALS CONCEDED | MOST SHOTS ON TARGET FACED | MOST CLEAN SHEETS | LEAST GOALS CONCEDED | LEAST SHOTS ON TARGET FACED | LEAST CLEAN SHEETS |
| 1 | Bundesliga | 2019-2020 | 2013-2014 | 2014-2015 | 2014-2015 | 2017-2018 | 2019-20 and 2020-21 |
| 2 | La Liga | 2016-2017 | 2011-2012 | 2013-2014 | 2019-2020 | 2020-2021 | 2016-2017 |
| 3 | Premier League | 2018-2019 | 2010-2011 | 2013-2014 | 2014-2015 | 2017-2018 | 2010-2011 |
| 4 | Serie A | 2020-2021 | 2013-2014 | 2011-2012 | 2010-2011 | 2015-2016 | 2019-2020 |
| 5 | Ligue 1 | 2020-2021 | 2011-2012 | 2016-2017 | 2019-2020 | 2019-2020 | 2019-2020 |

From table 6, we learn that Serie A and Ligue 1 conceded their most goals in the 2020-21 season, while for Bundesliga, this was in 2019-20 season, for La Liga, 2016-17 season, and the Premier League, 2018-19 season. The 2013-14 season is the season La Liga teams and Premier League teams kept the most clean-sheets. The 2019-20 season (Covid season) is the season with the least clean-sheet across three different leagues – Bundesliga, Serie A, and Ligue 1. ([Full Report and Dashboard](https://app.powerbi.com/view?r=eyJrIjoiYzA3NzU3NGUtM2ZiNC00YzIyLTg5MTYtY2M3ZDc4YTkzNGRjIiwidCI6IjY2NjYxMWFjLTE1NjktNDhjYy1iYjg5LWY2MjZkY2JmMjkxMSJ9)).

Q5.

Our analysis of the top 5 leagues in Europe, provided insights from – The English Premier League, La Liga, Bundesliga, Ligue 1, and Serie A. Our dataset had to be properly cleaned and data quality checks had to be conducted to allow for even analysis of the various teams and leagues without introducing bias in our analysis.   
We learn from our analysis, the impact of covid-19 on football across the major leagues in Europe as these created problems for various teams and leagues at large and this is evident in that season’s statistics especially the French league. The 2019-2020 season (Covid-19 season) allowed a maximum of 28 matches out of a possible 38 matches due to health concerns around the world. Only two teams were relegated from the French league that season due to these concerns. Asides, the French league, other leagues like the Bundesliga and Serie A suffered their worst defensive record in the 2019-2020 season, recording the least clean sheet. The 2019-2020 season saw the Bundesliga concede the most goals within the eleven-seasons of our analysis.   
Analysis of teams who have qualified for the champions league over the eleven seasons of our analysis show that Bayern Munich, Manchester City, Paris SG, Barcelona, Atletico Madrid, Juventus, Real Madrid, Chelsea, Dortmund, and Manchester Utd, are the top 10 teams with the best offensive ability. With Bayern Munich, Man City, Barcelona, Real Madrid being the more dominant forces across Europe while Juventus, Leverkusen, Manchester Utd, and Paris SG, can boast conceding the least goals.  
The decision to create a coefficient that adjusts for the Bundesliga with 18 teams while the remaining leagues have 20 teams and to account for the 28 matches played in the 2019-2020 season in the French League allowed analysis of our dataset in a robust manner, reducing bias and introducing new insights that guide the conclusions we were able to derive.  
While our analysis covers leagues, teams, and seasonal analysis across the top 5 leagues in Europe, further work can be done to improve this by introducing more variables that capture true offensive and defensive performances of teams e.g. xG, xGA, and xGD. Also, focus should be placed on analysing the impact of covid-19 on the various leagues to learn more from the 2019-2020 season across Europe.