Analysis of Europe's Top Five Football Leagues

IST 652 – Scripting for Data Analysis

GROUP 8:

Chaitanya Mupparaju
Shruti Rao
Gaurav Yadav

Data and Source Description:

The dataset utilized in this project is sourced from Kaggle, specifically from the following link: https://www.kaggle.com/datasets/hugomathien/soccer

Dataset Overview:

The dataset is centered around soccer and encompasses a comprehensive compilation of information related to various aspects of the sport. It includes several tables that collectively provide a multifaceted view of football analytics. The key tables within the dataset are:

Player: Provides general information about individual soccer players, such as names, nationalities, and positions.

Match: Encompasses details about individual soccer matches, including team details, scores, and match-specific statistics.

League: Offers information about different soccer leagues, including names, countries, and potentially other league-specific details.

Country: Provides details about the countries associated with the soccer data, including country names, codes, and possibly additional geographical or demographic information.

Team: Contains information about soccer teams, including team names, home cities, and potentially other team-related attributes.

Data Quality and Usability:

To ensure the dataset's integrity and usability, preprocessing steps, including cleansing and purging of messy or incomplete data, have been undertaken. The dataset is designed to facilitate comprehensive analysis, providing a rich source for those interested in gaining insights into the nuances of soccer through data-driven exploration.

The preprocessing steps mentioned earlier, like cleansing/purging of messy data and aggregation/summarization, indicate a commitment to ensuring data quality and maintaining the integrity of the analysis.

In summary, the dataset seems rich and diverse, providing ample opportunities for in-depth football analytics across different dimensions, including players, teams, matches, and leagues. The variety of tables allows for a multifaceted exploration of football-related phenomena.

Team Contributions and Responsibilities:

Gaurav Yadav:

- Conducted an in-depth analysis of Bundesliga, focusing on Borussia Dortmund's performance decline during specific seasons.
- Explored and compared Manchester United's performance in the Premier League, specifically analyzing the impact of Sir Alex Ferguson's presence on the team.

Chaitanya Mupparaju:

- Led the analysis of Laliga, specifically investigating the predictability of match outcomes within the league.
- Conducted a comprehensive analysis of Ligue 1, exploring Paris Saint-Germain's transformative journey post-QSI takeover.

Shruti Rao:

- Executed a detailed analysis of Serie A, emphasizing the relationship between teams with the most clean sheets and their success in winning the league.

Collaborative Efforts:

- Collaboratively synthesized individual analyses to create a cohesive and comprehensive final report.
- Collaborated on the development of a compelling presentation for the in-class session, integrating findings and insights from each team member's analysis.

Ligue 1: QSI Takeover Resurgence

PSG's Striking Transformation and Dominance in Ligue 1

Methods of Analysis:

Questions to be answered:

1. How did Paris Saint-German's performance metrics, including wins, losses, and goal differentials, evolve in Ligue 1 before and after the Qatari Sports Investments (QSI) takeover?
2. What variations occurred in PSG's average goals per match in Ligue 1 as a result of the transformation before and after the QSI takeover?

Fields Used:

Key fields include metrics related to wins, losses, goals scored, goals conceded, and average goals per match. Utilizing columns such as 'home_team_goal' and 'away_team_goal' for wins and losses, and 'season' and 'date' for categorizing pre and post-QSI periods.

Results Collation:

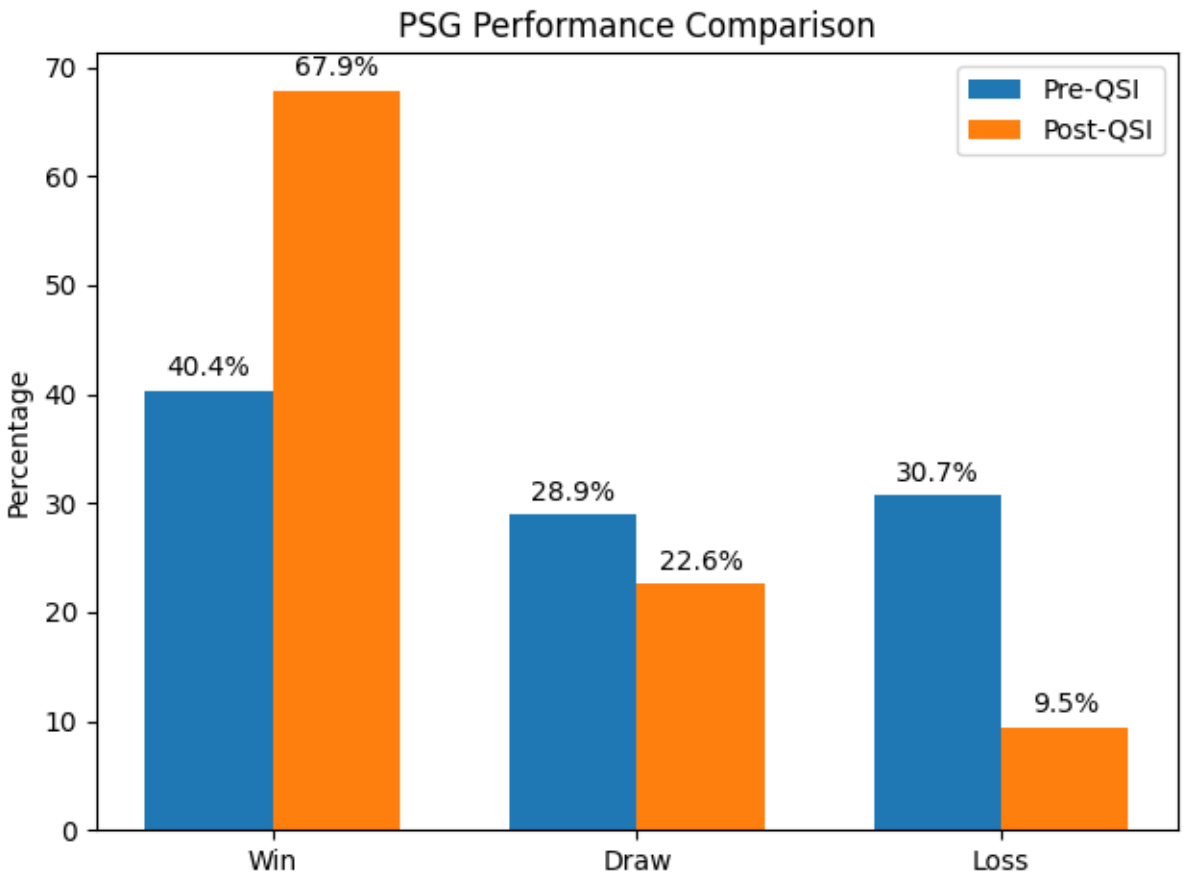
Results are meticulously collated and visualized using bar charts for each crucial metric. These charts provide a concise yet comprehensive representation of PSG's performance evolution over distinct periods.

Overall Python Program Description:

The Python program systematically identifies PSG, filters matches based on the QSI takeover timeline, and employs functions to calculate wins, losses, goals scored, goals conceded, and average goals per match. The generated visual outputs, specifically bar charts, effectively communicate PSG's performance changes over time.

Documentation of Output:

Visual outputs, represented by bar charts, serve as detailed documentation for PSG's performance metrics evolution. Each chart is aptly titled, and data points are labeled, ensuring transparency and accessibility for stakeholders interested in understanding PSG's journey.



Conclusion:

The analysis concludes that PSG underwent a substantial transformation post-QSI, showcasing improved performance metrics. Increased wins, decreased losses, prolific goal-scoring, and enhanced defensive capabilities underscore the profound impact of the QSI takeover on PSG's competitiveness and success in Ligue 1.

Laliga BBVA: The Predictability of Dominance

Exploration of Football Team Predictability Across Top 5 Leagues:

Methods of Analysis

Questions to be answered:

1. How does the predictability (entropy) of match outcomes vary across the top five European football leagues and individual teams over different seasons?

2. In what ways do the plotted graphs illustrate the predictability of Real Madrid and Barcelona in comparison to other teams across the top five football leagues, providing insights into their consistent performance expectations?

Fields Used:

For league predictability, data from the ‘df_match_top_5’ DataFrame is utilized, focusing on columns 'name' (league name), 'season,' and 'entropy.' For team predictability, data from the ‘df_match’ and ‘df_league’ DataFrames is used, concentrating on columns related to match odds and outcomes.

Results Collation:

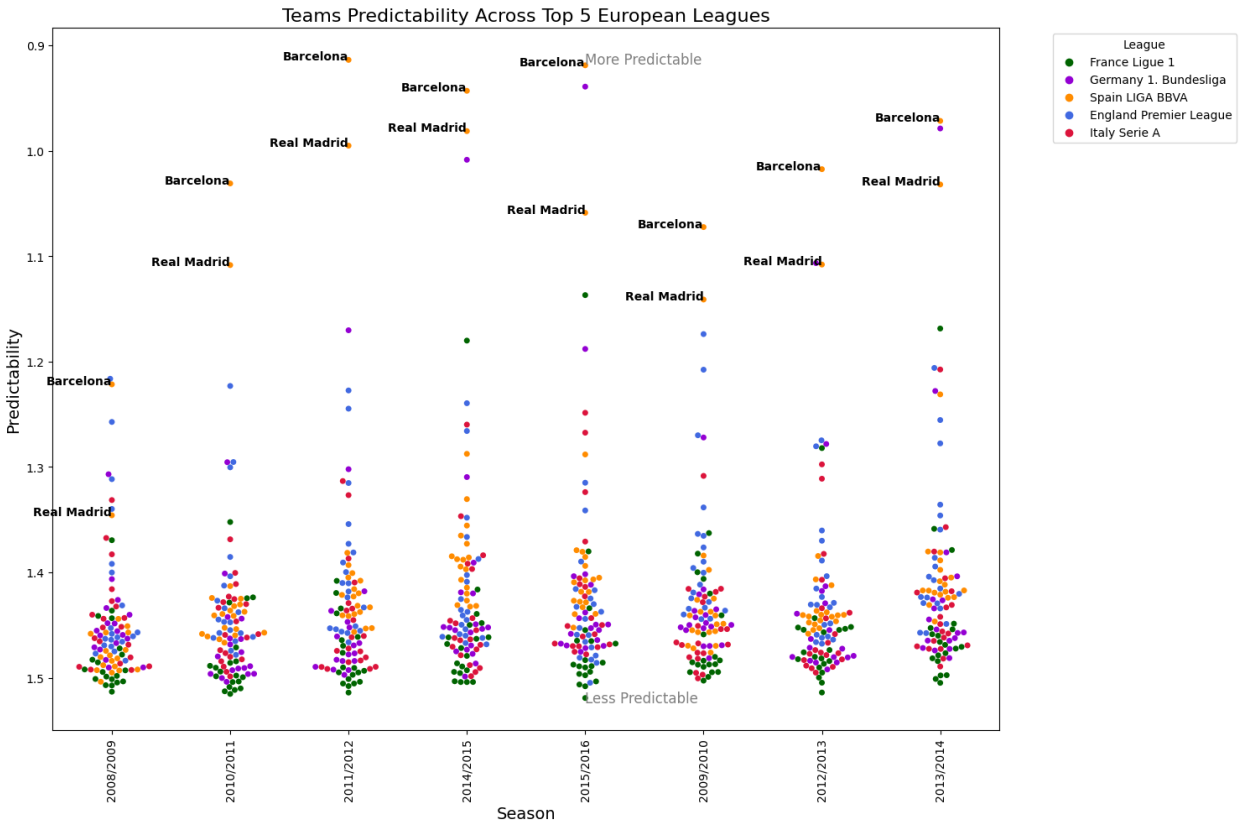
Swarm plot visualizing team predictability across seasons and leagues.

Overall Python Program Description:

The Python program employs functions and DataFrame manipulations to calculate entropy, average entropy, and other performance metrics for different entities (leagues, teams) over various seasons. Visualization libraries like matplotlib and seaborn are used for creating informative plots.

Documentation of Output:

Visual outputs include line plots showcasing league predictability trends and swarm plots depicting team predictability. Each chart is titled, and annotations provide insights. Data points are labeled for transparency.



Conclusions:

The combined analysis offers insights into the predictability of match outcomes in top European football leagues and individual teams. Visualizations provide a comprehensive understanding of how predictability evolves over seasons for both leagues and teams, contributing to a holistic view of match outcome trends in European football.

Serie A Defensive Dominance:

<u>Analyzing the Impact of Clean Sheets on League Victories</u>

Methods of Analysis:

<u>Questions to be Answered:</u>

1. Which teams consistently demonstrate the highest defensive prowess in Serie A, as indicated by the most clean sheets across different seasons?
2. Is there a correlation between a team's defensive performance, measured by clean sheets, and its overall success, particularly in terms of winning the Serie A league title?

Fields in the Data Used:

The analysis utilizes fields from the matches DataFrame, including 'season,' 'home_team_api_id,' 'away_team_api_id,' 'home_team_goal,' and 'away_team_goal.' Additionally, information from the teams DataFrame, such as 'team_api_id' and 'team_long_name,' is used to enhance result interpretation.

Results Collation:

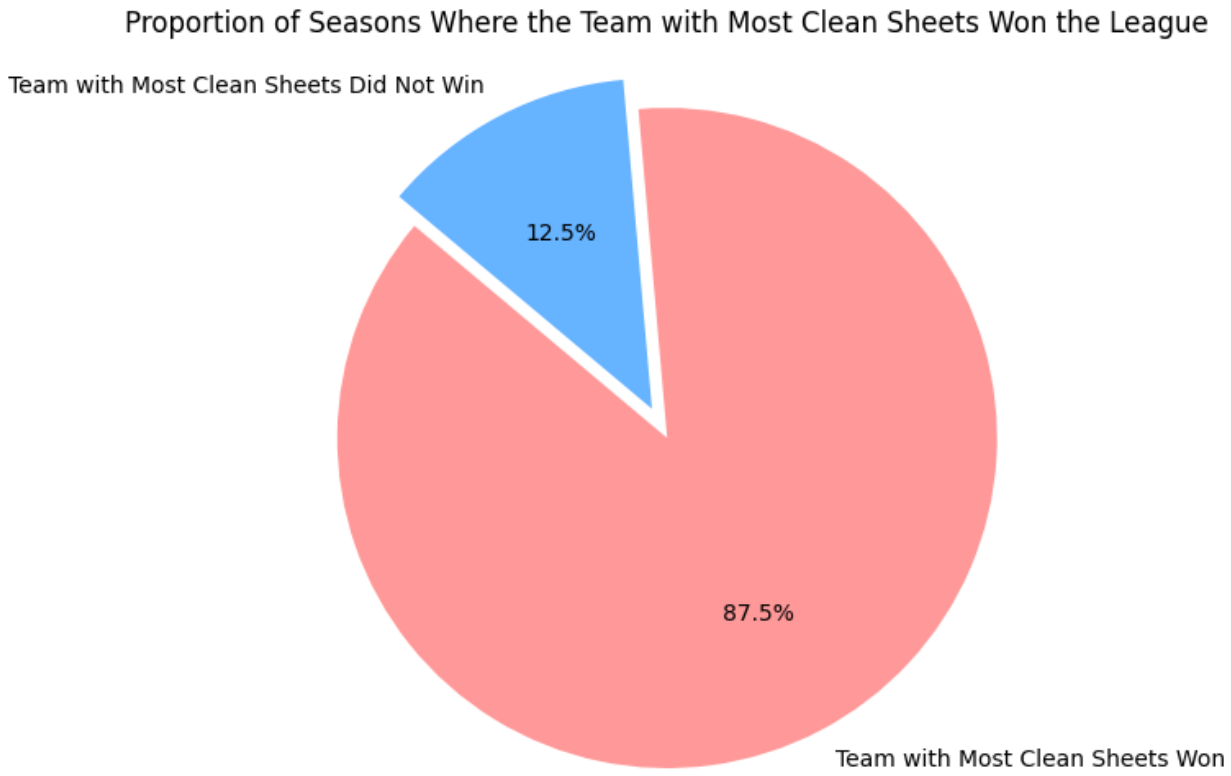
The results are collated through basic calculations on the DataFrame. The number of seasons where the team with the most clean sheets won and where a different team won is determined. These counts are then used to create a pie chart illustrating the proportion of seasons where the team with the most clean sheets won the league.

Overall Description of the Python Program:

The Python program focuses on Serie A matches, filtering the DataFrame accordingly. Two key functions, calculate_clean_sheets and identify_league_winners, are employed to gather information on clean sheets and league winners. The program then creates a comparison DataFrame, combining these results and additional team details for clarity. Finally, the DataFrame is printed to provide a concise overview.

Documentation of Output:

The output is a comparison DataFrame featuring columns for 'Season,' 'Team with Most Clean Sheets,' 'League Winner,' and 'Is Same Team.' Each row corresponds to a Serie A season, presenting information on clean sheets and league winners. The DataFrame is appropriately labeled and formatted for ease of interpretation.



Conclusion:

The analysis allows conclusions regarding the defensive performance and overall success of teams in Serie A across different seasons. By identifying teams with the most clean sheets, it provides insights into defensive strengths. The comparison between the team with the most clean sheets and the league winner indicates whether defensive prowess aligns with overall success. The results contribute to understanding the dynamics of Serie A seasons, emphasizing defensive achievements alongside league victories.

Bundesliga Strategic Transition Impact:

Unraveling Dortmund's Performance Decline in Bundesliga Seasons

Methods of Analysis:

Questions to be answered:

1. How have FC Bayern Munich and FC Borussia Dortmund performed in the Bundesliga across different seasons?

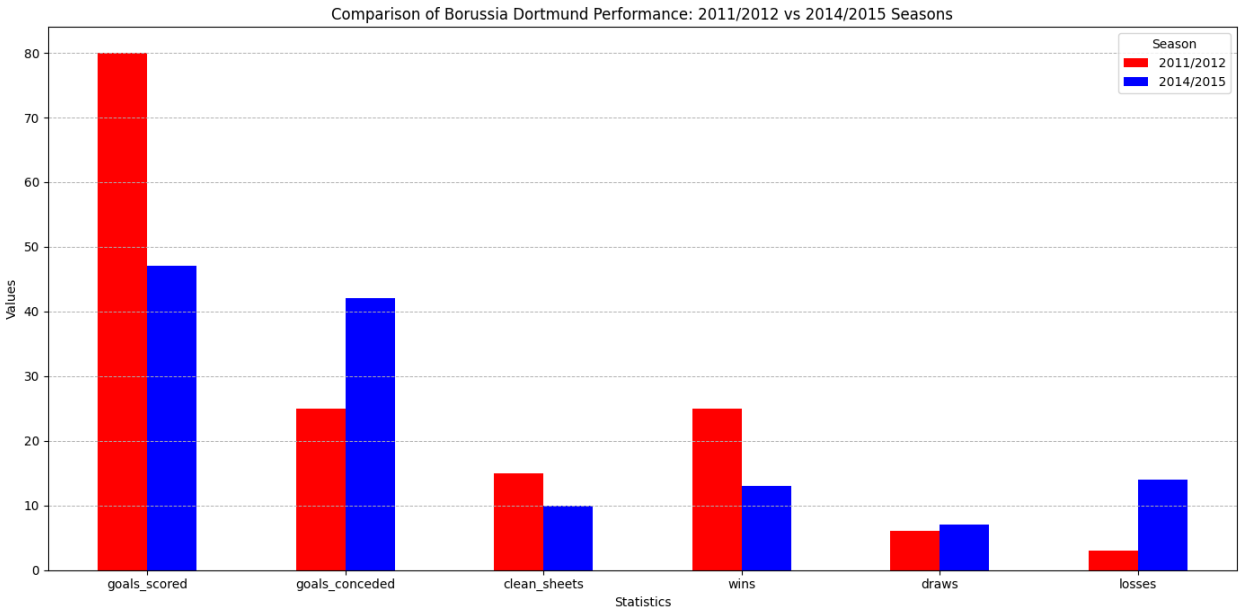
2. What were the key factors contributing to Borussia Dortmund's sharp decline in performance during the 2014/2015 season compared to their best-performing season in 2011/2012?

Fields in the data Used:
'season,' 'home_team_api_id,' 'away_team_api_id,' 'home_team_goal,' 'away_team_goal,' 'team_long_name.'

Results Collation:
Analyzing FC Bayern Munich and FC Borussia Dortmund's Bundesliga performances, line charts showcase their points accumulation trends over multiple seasons. The comprehensive data aggregation process involved calculating points for each team, offering a concise yet insightful depiction of their dynamic trajectories in the league.

Overall Python Program Description:
The Python program reads Bundesliga datasets into Pandas DataFrames, preprocesses the data, and applies specific functions for each analysis. Functions include assigning points based on match results, calculating total goals, and creating visualizations using Matplotlib. The program generates visual outputs, including line charts and bar charts each appropriately labeled for clarity.

Documentation of Output:
The program produces visual outputs showcasing trends in team performance and a comparison of statistics for the 2014/15 and 2011/12 seasons. Each visualization is labeled and presented in a clear format, making it accessible and informative.



Conclusion:

The comprehensive analysis focused on calculating total points for FC Bayern Munich and FC Borussia Dortmund in each Bundesliga season, providing a clear trajectory of their performance over time. A notable observation was Dortmund's sharp decline in total points during the 2014/2015 season, prompting a deeper investigation into the underlying factors. To contextualize this decline, a detailed comparison was made between Dortmund's best-performing season in 2011/2012 and the challenging 2014/2015 season. Key statistics such as total goals scored, goals conceded, clean sheets, wins, losses, and draws were scrutinized. The decline in Dortmund's performance during the 2014/2015 season aligns with significant changes, notably the departure of their manager Jurgen Klopp.

English Premier League Dynamics:

Deciphering Manchester United's Post-Sir Alex Ferguson Decline

Methods of Analysis:

Questions to be Answered:

1. How have Premier League teams, including Manchester United, performed over different seasons?
2. What are the differences in performance between the era with Sir Alex Ferguson and the post-Ferguson era?

Fields Used in the Data:

In conducting both analyses, we utilized a set of common fields: Season, Team API ID, Points, Goals Scored, and Goals Conceded. The Team API ID was used selectively to isolate Manchester United's data for targeted analysis.

Collation of Results:

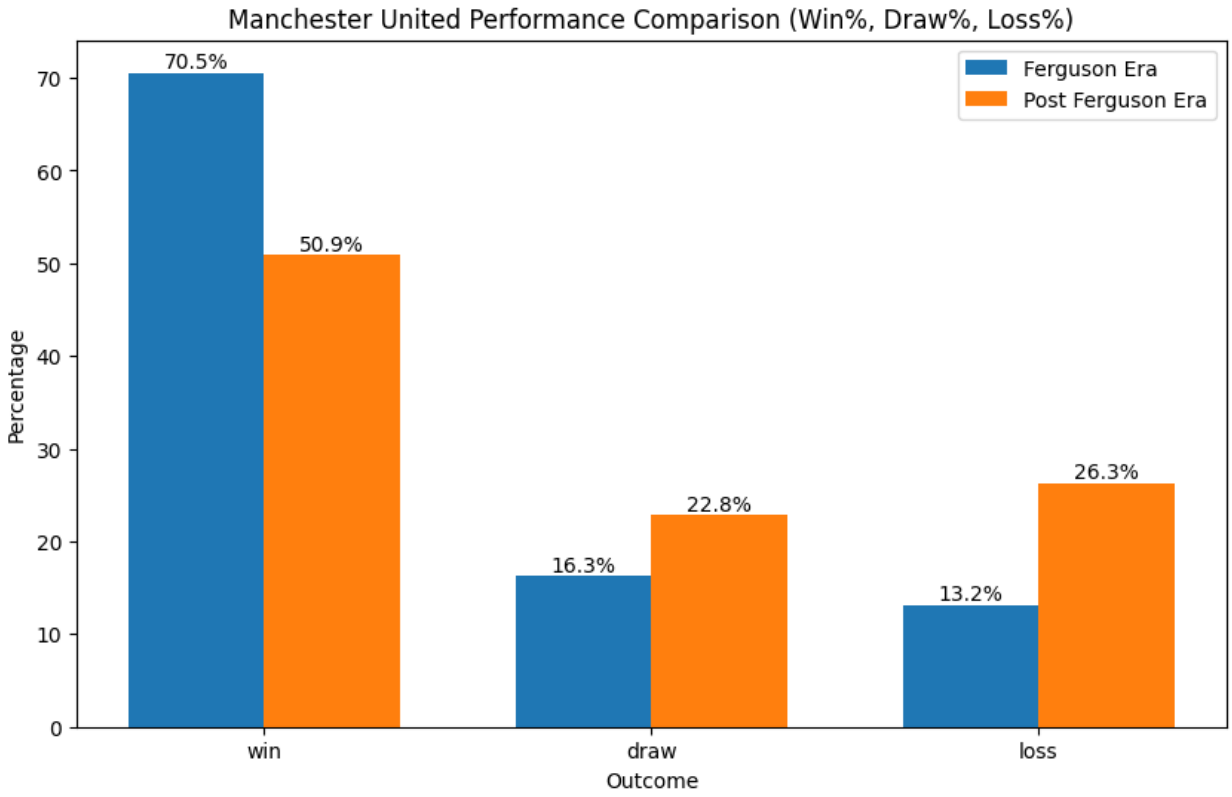
For the comparative analysis of Manchester United's stats with and without Sir Alex Ferguson, relevant metrics such as total points, goals scored, and goals conceded were calculated for each season. Plots were created to visualize the trends, emphasizing the differences in performance metrics between the two eras.

Overall Description of Python Program:

The Python program employs the Pandas library for data manipulation and analysis. It is structured into distinct sections, each focusing on specific analyses. Functions are utilized for modularity and readability. The code reads datasets, preprocesses data, performs analyses, and outputs meaningful visualizations.

Documentation of Output:

The output encompasses Pandas DataFrames and visualizations, providing a comprehensive view of team and Manchester United-specific performance. The Manchester United analysis includes plots displaying trends in total points, goals scored, and goals conceded, differentiating between the Sir Alex Ferguson and post-Ferguson eras.



Conclusion:

The analyses reveal a notable decline in Manchester United's performance after Sir Alex Ferguson's departure. Specific metrics such as goals scored and total points highlight this decline. Intriguingly, the worst season under Sir Alex Ferguson (68 points) still outperforms the best season without him (64 points). This stark contrast emphasizes the profound impact of managerial transition on a football powerhouse, showcasing the enduring legacy of Sir Alex Ferguson at Manchester United.