Advanced Programming Final Project Report:

Soccer Data Analysis

Shubham Patil
Divya Shah
Ganesh Makkina

1. Introduction:

Football, also known as soccer globally, fascinates millions of fans worldwide, generating a vast amount of data. Whether it is from players, matches, teams, and tournaments - the statistics are huge. Soccer Data Analytics has revolutionized the way the world perceives the game and the teams strategize, plan, and perform, making it an indispensable tool in this era of sport. Let's explore how this technological marvel is revolutionizing the game.

2. Objectives:

Our objectives for this project were to:

- Analyze the distribution of football teams by country and continent.
- Explore player attributes like height, weight, and nationality.
- Visualize match data, including goals scored over time.
- Evaluate player performance metrics and positions.
- Map the geographical locations of matches and teams.

3. Data Sources:

We sourced data from trusted providers like Figshare and Kaggle, covering all aspects of the game which are teams, players, matches, competitions, attributes, and more.

4. Methodology:

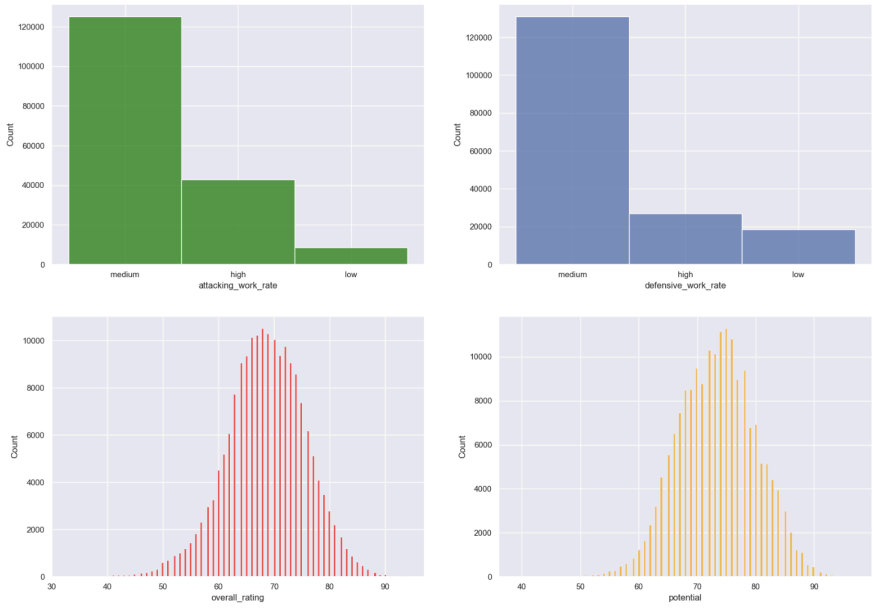
Our methodology involves several key steps to conduct a comprehensive analysis of the soccer data. Firstly, data preprocessing is performed to handle missing values and remove irrelevant columns, ensuring the dataset's cleanliness and relevance. Next, exploratory data analysis techniques are employed to gain insights into the relationships between categorical variables and identify patterns within the data. These preparatory steps lay the groundwork for subsequent analytical techniques, including correlation analysis, visualization, and statistical modeling, aimed at uncovering actionable insights into supply chain dynamics, customer behavior, and operational efficiency. Through this methodology, we aim to derive meaningful conclusions that can drive strategic decision-making and improve the performance of a team by assisting it to identify the areas of improvement.

5. Data Analysis:

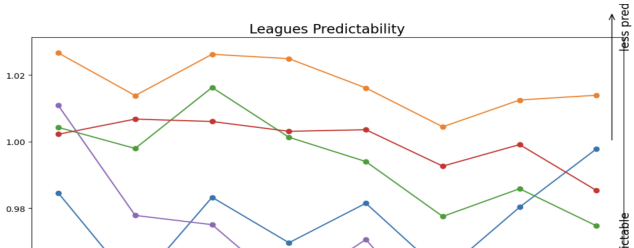
Dataset 1

In this dataset, we delve into a comprehensive soccer repository boasting over 25,000 matches and 10,000 players spanning the premier championships of 11 European countries from 2008 to 2016. Drawing from the esteemed EA Sports' FIFA video game series, we gain access to players' and teams' attributes, continually updated with weekly changes. Detailed match events such as goal types, possession, corners, crosses, fouls, and cards, along with team lineups featuring squad formations and X, Y coordinates, provide a granular view of each encounter. Additionally, betting odds from up to 10 providers add a layer of predictive insight. With the recent inclusion of teams' attributes from FIFA, this dataset offers a robust foundation for advanced data analysis and machine learning applications in the realm of soccer.

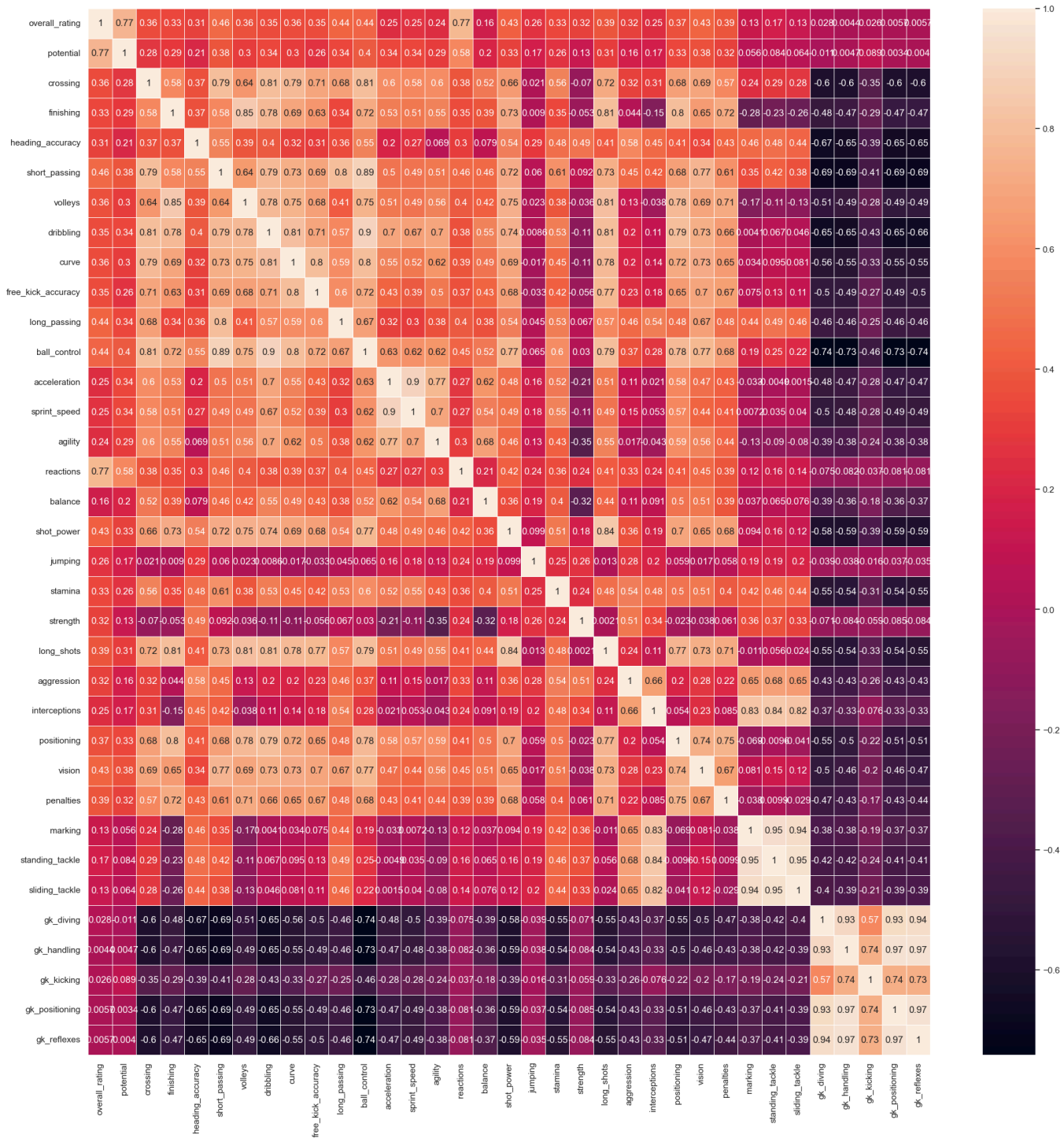
- Attacking Work Rate Distribution: Offers insights into player offensive involvement and their approach towards attacking play.
- Defensive Work Rate Distribution: Provides an understanding of player defensive tendencies and their commitment to defensive duties.
- Overall Rating Distribution: Highlights the distribution of player skill levels, offering an overview of the talent pool.
- Potential Rating Distribution: Reveals the projected development and growth potential of players, indicating future performance trends.



- Teams Predictability: Illustrates the predictability of different soccer teams across seasons, aiding in understanding the consistency and variability of their performance.



- Sum of Squared Errors (SSE) for Different Cluster Numbers: Displays the variation of SSE values for different numbers of clusters, aiding in determining the optimal number of clusters for KMeans clustering.
- Player Clusters Visualization: Presents a horizontal bar chart for each cluster, showcasing the distribution of players across different clusters. This visualization offers insights into the segmentation of players based on various attributes or features.
- Players' Embedding Visualization: Utilizes TruncatedSVD to reduce the dimensionality of player vectors to two dimensions, facilitating visualization.
- Scatter Plot: Displays player embeddings in a two-dimensional space, where each point represents a player.
- Hierarchical Clustering Dendrogram: Illustrates the hierarchical clustering of countries based on player attributes, providing insights into the similarity and dissimilarity between countries' player profiles.
- Ward's Method: Utilizes Ward's linkage method to measure cluster distances, aiding in identifying meaningful clusters within the data.
- Leaf Labels: Displays country names as labels on the dendrogram leaves, facilitating easy interpretation and comparison.
- Most Successful Teams Analysis: Evaluates the performance of soccer teams across all seasons, comparing total games played, wins, and win percentage.
- Ranking: Ranks teams based on their win percentage, highlighting the most successful teams over the specified period.
- Player Physical Attributes Analysis:
    - Comparative Analysis: Examines the distribution of player height and weight, shedding light on the variability and trends in these physical characteristics.
    - Frequency Distribution: Illustrates the prevalence of different player heights and weights, offering insights into common and rare occurrences within the dataset.
    - Relationship Exploration: Investigates the correlation between player height and weight through a scatter plot, providing a deeper understanding of how these attributes interact.
- Correlation Calculation: Player Attributes Correlation Analysis computes correlations between player skills and performance attributes. It's visualized using a heatmap, highlighting relationships with numerical annotations for clarity.



In our comprehensive analysis of soccer data, we delved into multifaceted aspects encompassing player attributes, team dynamics, and match outcomes. Our exploration commenced with meticulous data preparation and exploration, meticulously scrutinizing datasets containing detailed information on matches, players, and teams. Employing a range of exploratory data analysis techniques, we gleaned insights into the intricate distribution and characteristics inherent in the data, setting the stage for deeper investigations.

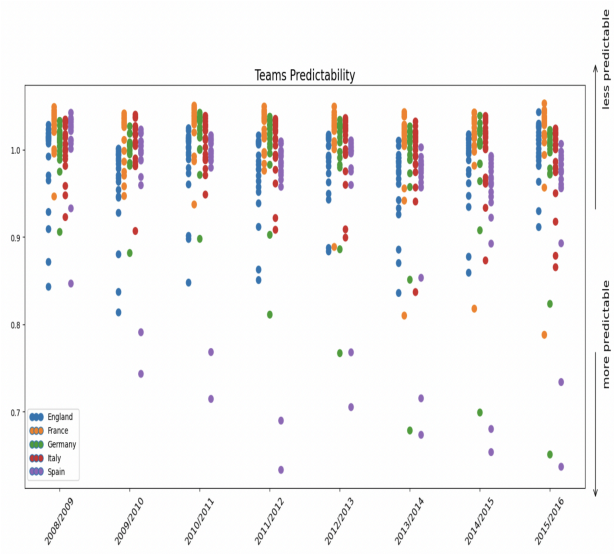
A focal point of our analysis revolved around dissecting player attributes, deciphering the nuanced profiles of soccer athletes. Through meticulous examination of metrics such as height, weight, and preferred foot, coupled with an evaluation of attacking and defensive work rates, we gained invaluable insights into the diverse skillsets and playing styles exhibited by players. Employing an array of visualizations, including box plots, histograms, and count plots, we not only unraveled the distribution patterns but also uncovered subtle interrelationships between different attributes, thereby enhancing our understanding of player dynamics.

Furthermore, our analysis extended to team attributes, where we scrutinized performance metrics to discern strengths, weaknesses, and overarching strategies adopted by soccer teams. Our exploration of match outcomes, entailing an in-depth assessment of total games played, wins, and win percentages across seasons, was complemented by correlation analysis. This endeavor unveiled intricate relationships between numerical player attributes, elucidated through a visually engaging heatmap. Our analysis culminates in a holistic understanding of soccer data, offering actionable insights for stakeholders ranging from team managers to sports analysts, and contributing to the ongoing evolution of soccer analytics.

Dataset 2-

This dataset was in a sqlite format so we used that as it is in our coding in jupyter notebook. Tables named league, players, teams, match and country were used to analyze this dataset.

- League Predictability Analysis: Analyzed the predictability of leagues in England, France, Germany, Italy, and Spain. It helped in assessing the competitive balance and dynamics of each league.
- Team Predictability Analysis: Explored the predictability of individual teams within the mentioned leagues.
- Rating of Japanese Players (2008-2016): Evaluated the performance rating of Japanese players over an eight-year period. Tracking the rating of Japanese players offers insights into the development and progression of talent from Japan in international football.
- Clustering Players by Colors: Utilized clustering techniques to categorize players based on performance metrics. Clustering players by performance allows for the identification of distinct player groups, aiding in talent scouting and team composition.
- Hierarchical Clustering: Used hierarchical clustering to further refine player categorization. It offers a hierarchical structure of player groups, enabling deeper insights into player similarities and differences.
- Database Schema Formation: Developed a database schema with primary and foreign keys for efficient data organization and management. A well-defined database schema enhances data understanding and facilitates seamless data retrieval and manipulation for analysis.
- Correlation Matrix of Player Performance: Created a correlation matrix to explore relationships between different player performance metrics.
- Distribution Plotting of Performance Metrics: Plotted the exact distribution of player performances to determine mean, median, and visualize performance trends.
- Box Plot Analysis of Performance Metrics: Utilized box plots to visualize the distribution and variability of player performance metrics. Box plots offer a clear visual representation of performance distribution, facilitating comparison and identification of performance patterns and anomalies.



Dataset 3-

This dataset was complex and challenging to work with as it contained various datasets in different formats - json and csv. This means that we had to convert all the files to the same format and clean the data for further analysis.

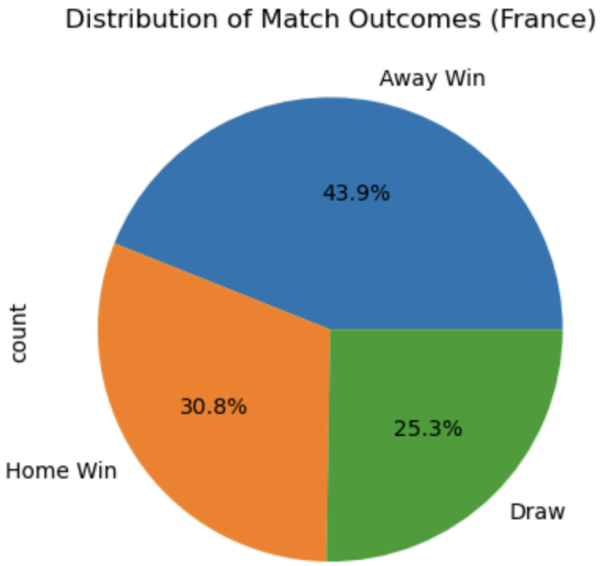
Below is the glimpse of the insights that we gained after performing Exploratory Data Analysis:

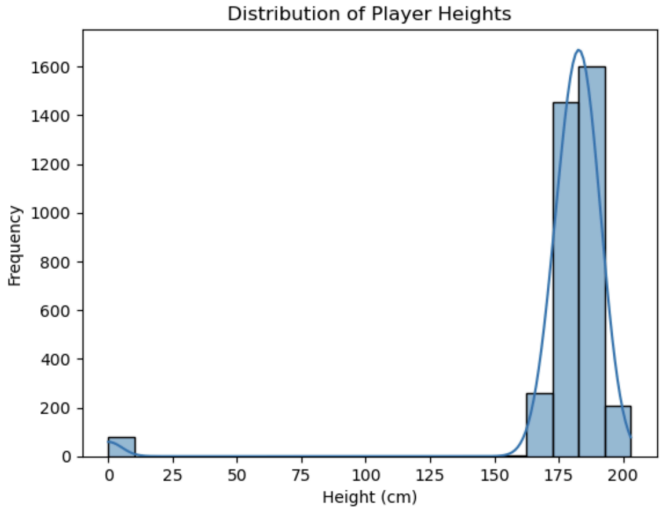
- The teams playing away from home in the tournament in France had slightly higher chances of winning than the teams playing in their home city.
- The scatter plot and the heatmap of Weight vs Height suggests that there is a correlation between the height of a player and the weight of the player. Also, almost all the soccer players were found to be tall and their weight was on the higher side, suggesting fitness is the key to this game.
- Most of the players' height ranged between 180 and 205 cm.

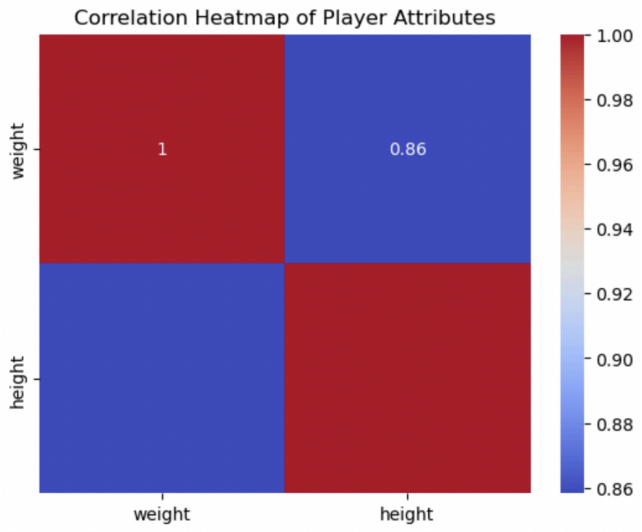
We then analyzed some player statistics with respect to the position of the player. For this, the players.csv, after converting it from json, was used to get this data. It suggested that the Goalkeepers are the tallest and this is because the goalkeepers have to cover a larger portion and reach higher heights in order to stop the opponent from scoring.

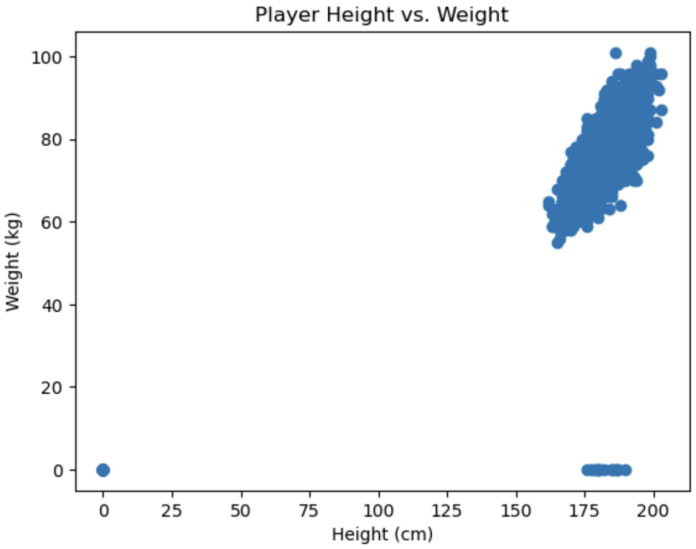
In England and France, the team playing away from home scored more than the one playing in their home and for Germany and Spain, it was the other way around. But for Italy, it was inconsistent- sometimes the home team is scoring the most and sometimes the away team is scoring the most.

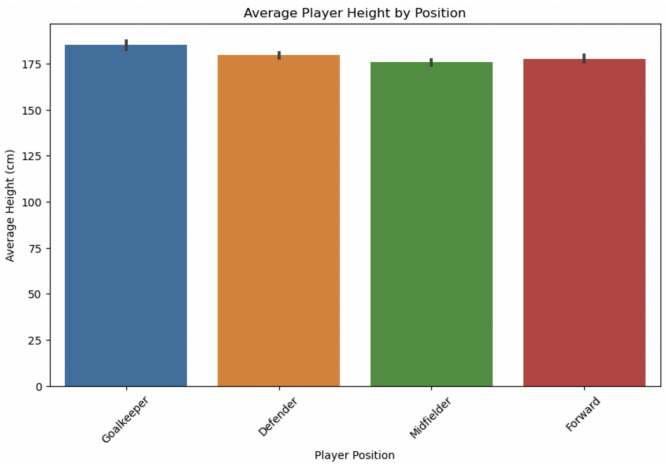
We then used all the datasets with the teams and their respective tournaments info and we plotted the network graph. The network graph depicts the teams who have played against each other. This graph provides information about the teams that frequently played against each other, assisting the teams in strategizing and planning to be on their way to win.

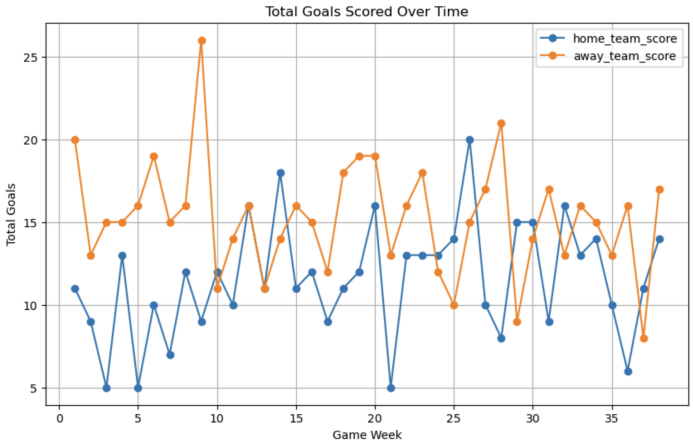

Distribution of Match Outcomes (France)
Away Win
43.9%
count
30.8%
25.3%
Home Win
Draw


Distribution of Player Heights
Frequency
Height (cm)


Correlation Heatmap of Player Attributes
weight
height
1
0.86
weight
height


Player Height vs. Weight
Weight (kg)
Height (cm)

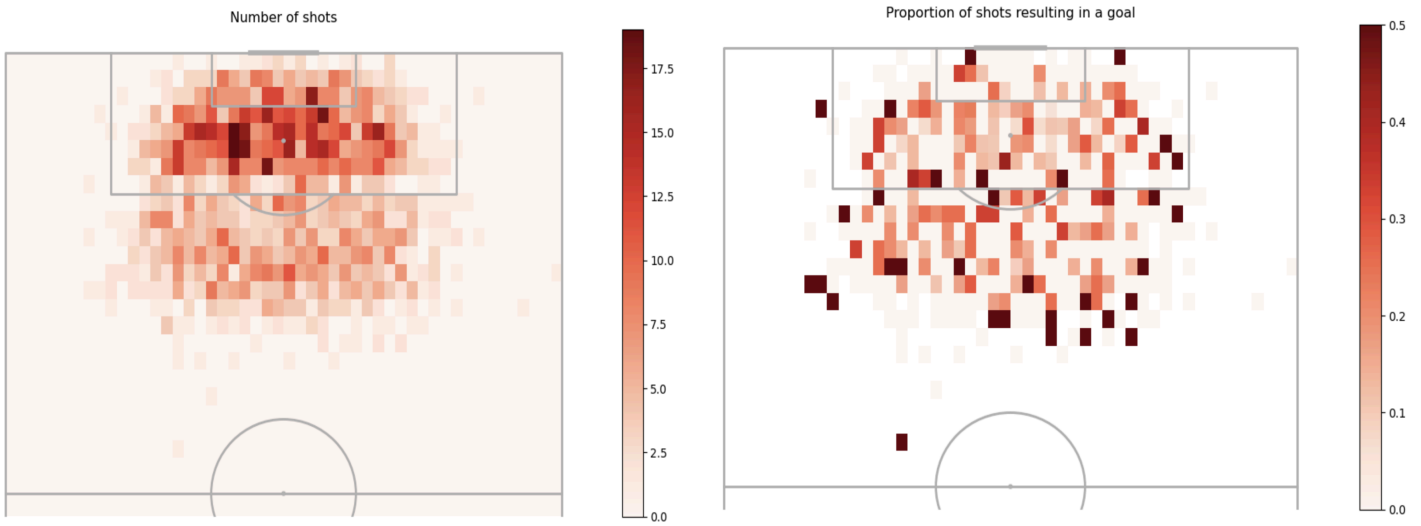

Average Player Height by Position
Average Height (cm)
Goalkeeper
Defender
Midfielder
Forward
Player Position


Total Goals Scored Over Time
home_team_score
away_team_score
Total Goals
Game Week

Dataset 4-

This was the largest dataset in our set and we have used a couple of the CSV files that were inside to perform analysis. Below written are few of the analysis and insights captured from the dataset whose results will help in the predictions.

- Shots taken: Provides insights into player offensive involvement and shooting behavior.
- Goals scored: Offers a measure of player effectiveness in converting scoring opportunities.
- Scoring probabilities from different field positions: Reveals trends and efficiency of players in various areas of the field.
- Feature importance analysis: Identifies critical variables contributing to player performance outcomes.
- Average VAEP rating: Quantifies the value of player actions by estimating their impact on match outcomes.
- Correlation matrices: Unveils relationships between different performance metrics, aiding in understanding player dynamics.
- K-means clustering: Groups players based on shared characteristics, facilitating segmentation and analysis.
- Silhouette plots: Evaluates the quality and coherence of player clusters generated by K-means clustering.
- Interactive plots: Enables intuitive exploration and visualization of player clusters for deeper understanding.

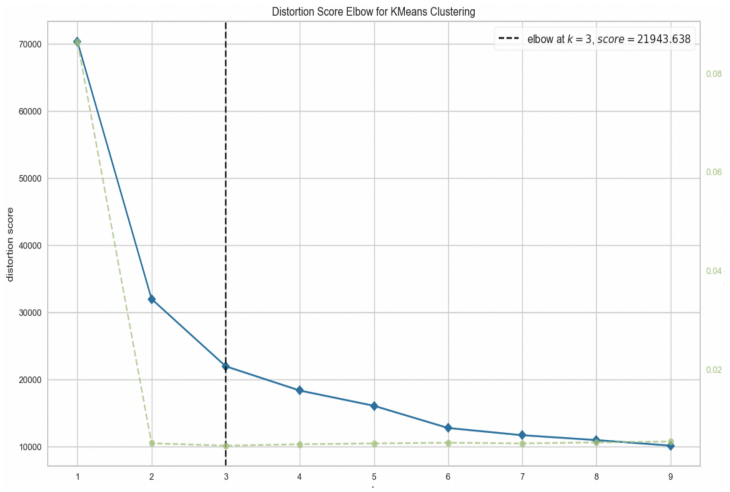


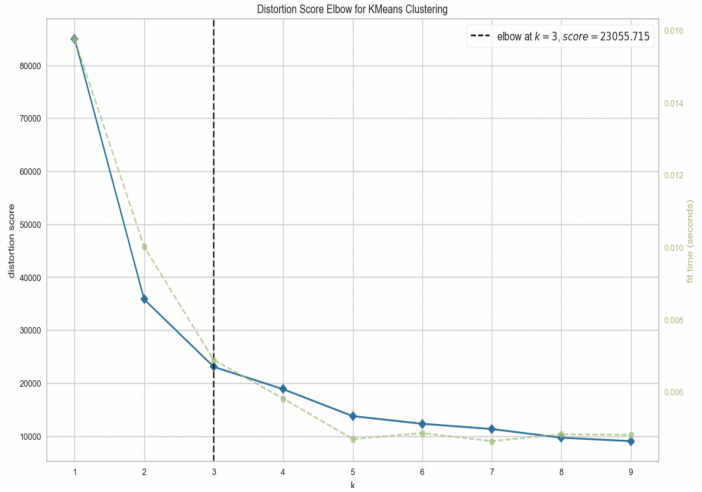
These given plots analyze the total number of shots taken towards the goal and the other photo shows the proportion of the shots taken that actually resulted in a goal. This helps in analyzing the percentage of shots that do end up hitting a goal.

We have then trained the dataset models.csv which shows us the scores and concedes columns. The data is then split into training and testing and it is run on the XGBoost machine learning model where some of the selected features are- start_distance_to_goal-0, End_distance_to_goal-0, Start_distance_to_goal-1, End_distance_to_goal-1, Start_distance_to_goal-2, End_distance_to_goal-2, Start_angle_to_goal-0 and end_angle_to_goal-0. This model then shows us the importance of each feature and which feature is most important and which one is least important. Also the SHAP values have been investigated which tells us the impact on model output.

We have then used the column values in player_games.csv to calculate the VAEP rating of each player. VAEP is Valuing Actions by Estimating Probabilities. It is a metric used in soccer analytics to quantify the value of individual player actions on the field by estimating their impact on the probability of scoring or conceding a goal. It considers different factors such as the location on the field, the context of the action, the involvement of other players, and the outcome of the action. These provide a more granular understanding of player contributions beyond traditional statistics like goals and assists. They help in evaluating a player's overall impact on the game, including their offensive and defensive effectiveness, decision-making ability, and strategic positioning.

We have analyzed the playerrank.csv file to firstly find out all the different roles and positions that are available for all the players. Then we moved one to define the name in a better way doing data preprocessing and then we made a correlation matrix of all the different player performance metrics. Hopkins hypotheses is used to assess the spatial clustering tendency of a dataset. It helped us in identifying spatial clustering of player actions, detecting defensive formations, assessing offensive strategies, analyzing player positioning and examining spatial trends in goals conceded. We used this to perform K-means clustering and Silhouette plot of it. After that we made a few adjustments and then plotted these same graphs to understand the differences in it. The adjustments made to the attributes were to involve filtering players based on minutes played (>90), aggregating duel and passing actions, normalizing aggregated values per 90 minutes and selecting a subset of relevant attributes for analysis.





6. Key Findings:

We found that most teams are based in Italy (14.8%) and Spain (14.8%), followed by England (14.1%) and France (14.1%). This distribution suggests that these countries have a strong soccer culture and infrastructure. This implies that soccer is a big deal for Europeans and its competitive nature of the game across Europe.

Moreover, we found that there are 770 Away Wins and 680 Home wins. This suggests that the team playing away from their home stadium has a slightly higher chance of winning. Also, the number of draws (480) suggests that the matches in football are often neck and neck, implying that the teams often end up having equal scores.

Coming to player statistics, we found that most of the players' height fall between 180 and 190. Number of players having height between the category of 190-200 is the minimum. The heatmap also provides evidence that the height and weight of the player are highly correlated.

We also found that Goalkeepers are the tallest followed by defenders, then midfielders, and lastly forwards. This might be due to the fact that goalkeepers are required cover more area and reach higher heights to stop the opponent from scoring.

6. Conclusion:

In this project, we delved into player performance and match predictions, uncovering insights into playing styles and key performance indicators. Through interactive graphs and predictive models, we provided tools for exploration and decision-making. These analyses offer valuable insights into predicting player performance and match outcomes, underscoring the value of data-driven approaches in soccer analytics.

7. Future Work:

Opportunities for future work include:

- Further exploration of player performance metrics and team strategies.

- Development of machine learning models for outcome prediction.

- Collaboration with football clubs and organizations for richer datasets.

8. References:

1. James, P. (2010). Soccer Analytics: Successful Coaching through Match Analysis.

2. Memmert, D., & Raabe, D. (2018). Data analytics in football: Positional data collection, modelling and analysis. Routledge.

3. Pappalardo, L., Cintia, P., Rossi, A., et al. (2019). A Public Data Set of Spatio-Temporal Match Events in Soccer Competitions. Scientific Data, 6, 236.

4. Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. SpringerPlus, 5(1), 1410.

5. Sarmento, H., Marcelino, R., Anguera, M. T., et al. (2014). Match analysis in football: a systematic review. Journal of Sports Sciences, 32(20), 1831-1843.This systematic review covers various methods and technologies used in soccer match analysis, offering a comprehensive overview of the field's evolution.

6. Lucey, P., Bialkowski, A., Monfort, M., et al. (2014). Quality vs Quantity: Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data.

Chatgpt Prompts:

1. How can I preprocess soccer match data using Python? I need to handle missing values and remove irrelevant columns.

2. I need to perform KMeans clustering on soccer player attributes. Can you show me how to determine the optimal number of clusters using the elbow method in Python?

3. How can I use hierarchical clustering to group soccer teams based on their performance metrics? I'm particularly interested in using Ward's method.

4. Can you help me write a SQL query to select all players with an attacking work rate above average from a soccer database?

5. What are some effective Python libraries for visualizing geographic data related to soccer teams and matches?

6. Can you provide a code snippet for generating a heatmap of correlations between different player attributes in a soccer dataset using seaborn?

7. What is the best way to visualize the distribution of player heights and weights in my soccer dataset using matplotlib?

8. I need to compare player performance across different teams. What statistical methods should I consider for this analysis?

9. How do I write a function in Python to calculate the win percentage of soccer teams from match data?

10. Can you assist me in writing the methodology section of my report on soccer data analytics, emphasizing the statistical techniques used?