**Project Documentation**

ON

**Football Analysis**

BY

**Rishi Ram**

**&**

**Rejoy Varghese**

MCA-II, SEM-III

2023-25

To

Savitribai Phule Pune University

Pune-411041

In Partial Fulfillment of the Degree of

Master in Computer Application (M.C.A.)

Under The Guidance Of

**Prof. Neha Dhakol**

**Suryadatta Group of Institutes, Pune**

**Suryadatta Institute of Management and Mass Communication (SIMMC)**

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Fields** | **Page No.** |
| **1.** | **Chapter 1 - Abstract** | **3** |
| **2.** | **Chapter 2 - Introduction** | **4** |
|  | 2.1 Objective |  |
|  | 2.2 Need of project |  |
|  | 2.3 Hardware and Software |  |
| **3.** | **Chapter 3 - Literature Review** | **10** |
|  | 3.1 Abstract of 10 Research Papers |  |
|  | 3.2 Limitations |  |
| **4.** | **Chapter 4 - Methods** | **13** |
|  | 4.1 System Design |  |
|  | 4.2 Data Collection |  |
|  | 4.3 Statistical Analysis |  |
|  | 4.4 Implementation |  |
|  | 4.5 Data Visualization |  |
|  | 4.6 Model Development |  |
|  | 4.7 Result |  |
|  | 4.8 Analysis and Testing |  |
| **5.** | **Chapter 5 - Conclusion** | **52** |
| **6.** | **Chapter 6 - Recommendation** | **53** |
| **7.** | **Chapter 7 - References** | **55** |

**Chapter 1: Abstract**

This analysis explores the multifaceted dimensions of football performance, focusing on predictive modeling and data analytics to evaluate player contributions in tournaments. The study aims to develop a robust predictive model leveraging historical data and advanced machine learning techniques to forecast player performance. Key performance metrics, including attacking and defensive statistics, passing accuracy, and consistency, are analyzed to create a comprehensive player assessment framework. Tactical analysis further enhances the understanding of team dynamics, emphasizing formations, attacking strategies, and transition plays. By utilizing advanced statistics such as Expected Goals (xG) and Expected Assists (xA), this research seeks to inform decision-making for player selection and strategic planning within teams. Ultimately, the project highlights the critical role of data analytics in modern football, offering insights for governing bodies and sports organizations to optimize resource allocation and enhance overall team performance. The findings aim to contribute to the growing field of sports analytics, providing a data-driven approach to understanding the complexities of player and team performance in football.

**Keywords:**

1. Football
2. Player Performance
3. Predictive Modelling
4. Data Analytics
5. Machine Learning
6. Football Statistics
7. Tournament Prediction
8. Player Selection
9. Sports Analytics
10. Performance Assessment
11. Historical Data
12. Football Strategy
13. Football Match Analysis
14. Team Performance Analysis

**Chapter 2: Introduction**

**Football**

Football, also known as soccer in some parts of the world, is one of the most popular and widely played sports globally. It is a team sport that involves two teams of eleven players each, competing to score goals by getting the ball into the opposing team's net using any part of the body except the hands and arms, with the goalkeeper being the only player allowed to use their hands within the penalty area.

The game is played on a rectangular field with a goal at each end, and it is governed by a set of standardized rules known as the Laws of the Game. Matches typically last 90 minutes, split into two 45-minute halves, with a brief halftime break. The simplicity of the game, combined with its physical and strategic demands, makes it accessible to players of all ages and skill levels.

Football has a rich history, with roots tracing back to ancient civilizations, but the modern version of the game was formalized in England in the 19th century. Today, it is a sport that unites people across different cultures and nations, with iconic tournaments like the FIFA World Cup and UEFA Champions League capturing the passion of millions of fans worldwide.

The sport emphasizes teamwork, strategy, and individual skill, with players taking on specific roles such as defenders, midfielders, and forwards to achieve their team's objectives. Football is more than just a game; it is a global phenomenon, celebrated for its ability to bring people together and its powerful impact on culture and society.

**Analysis**

1. Tactical Analysis

Tactical analysis focuses on how teams set up and execute their game plan. Key aspects include:

Formations: The shape of the team on the pitch, such as 4-4-2, 4-3-3, or 3-5-2. The formation influences a team’s attack and defense.

Pressing and Defensive Structures: How a team defends, either by sitting deep (low block), pressing high up the pitch, or using a mid-block approach.

Attacking Play: The way a team builds its attack. This can involve counter-attacking football, possession-based play, or using wide areas to cross the ball.

Transitions: How a team moves from defense to attack (and vice versa) during moments of the game.

2. Player Performance Analysis

Analyzing individual players' contributions includes:

Passing Accuracy: The percentage of completed passes, key passes (leading to scoring opportunities), and long passes.

Dribbling Success: The number of successful dribbles compared to failed ones.

Defensive Contributions: Tackles, interceptions, clearances, and blocks made by defensive players.

Goals and Assists: The number of goals scored, assists made, and chances created.

Heatmaps: A visual representation of where a player spends most of their time on the pitch.

3. Statistical Analysis

Modern football uses advanced statistics to give a deeper understanding of player and team performance:

xG (Expected Goals): A metric that evaluates the quality of chances a team or player has had, based on the likelihood of scoring from those positions.

xA (Expected Assists): Measures the likelihood that a pass will result in a goal based on historical data.

Possession Stats: How much of the game a team controls the ball, reflecting their style (e.g., possession-based teams like Barcelona).

Passing Networks: A diagram that shows the flow of passes between players and identifies key passing combinations and player roles.

Defensive Metrics: Include pressures, tackles, and recoveries, offering insight into how hard a team works off the ball.

4. Team Dynamics

Team cohesion and how players work together on the pitch are critical aspects:

Chemistry: How well players combine, especially in key partnerships (e.g., center-backs, midfield duos, or striker-winger combinations).

Leadership and Communication: The influence of key leaders on the pitch who organize the team and make decisions in real time.

Substitutions and Game Management: The impact of changes in tactics or personnel made by the coach during the match.

5. Opposition Analysis

Analyzing the opposition is key to forming a game plan:

Strengths and Weaknesses: Identifying areas where the opposition excels (e.g., strong in aerial duels) and where they are vulnerable (e.g., weak defending against pace).

Set Pieces: Observing how the opposition defends and attacks during set pieces like corners and free-kicks.

Key Players: Assessing which players are likely to have the biggest impact on the game and how to neutralize or exploit them.

**2.1 Objective**

**Prediction**

To create a predictive model that estimates a player's performance in football tournaments.

**Analysis**

To analyze historical data from previous football performance records and determine the best performance metrics for the player.

**Impact Assessment**

Attacking Performance Goals, assists, shots on target, and expected goals (xG).

Defensive Performance Tackles, interceptions, clean sheets, and defensive errors.

Overall Contributions Key passes, successful dribbles, and passing accuracy.

Consistency, adaptability, match-winner Assess the player's consistency across different matches, adaptability to various game situations, and ability to perform under pressure.

**Policy Insights**

To provide insights to football committees, governing bodies, and sports organizations on how to allocate resources and develop strategies to enhance their chances of optimal performance.

**2.2 Need of Project**

Football is a sport rich in data, making it an ideal candidate for analysis through advanced data analytics and machine learning techniques. The primary need for this project stems from the following reasons:

**1. Player Performance Evaluation:**

Accurately assessing player performance is crucial for making informed decisions regarding player selection, transfers, and tactical planning. Traditional methods of evaluating players can be subjective, but using data-driven approaches ensures objectivity and precision in evaluating players' contributions.

**2. Predicting Match Outcomes:**

By utilizing historical data and predictive modeling, this project aims to forecast match results, which can help teams prepare better strategies and allocate resources more effectively.

**3. Improved Tactical Understanding:**

With tactical analysis, teams can better understand their strengths and weaknesses as well as those of their opponents. This allows for more effective game strategies, improving team performance on the field.

**4. Team Management and Player Selection:**

Machine learning algorithms can assist in selecting the best team lineup for any given match based on various performance metrics and conditions. This ensures that teams are making informed decisions about player rotations and team composition.

**5. Advancement in Sports Analytics:**

With the rising importance of sports analytics, the project provides an opportunity to explore cutting-edge technologies in machine learning and data science, contributing to the growing field of sports analysis.

**2.3 Hardware and Software Requirement**

**Hardware Requirement**

* Intel core i3
* 256 GB HDD
* Monitor any
* Input Devices: Keyboard and Mouse
* RAM: 4GB

**Software Requirement**

* Operating System Windows 10 and above
* Coding Language: Python
* Software: Anaconda

**Chapter 3: Literature Review**

**1. Title: Prediction Model of Football World Cup Championship Based on Machine Learning and Mobile Algorithm**

Abstract: Selecting players for a football team should be based on a comparative ranking system across 11 positions, ensuring players are free from injury, fatigue, and mental pressure. The selection is based on performance data from previous matches, analyzed using Python for web scraping and Dartfish for video analysis. Key metrics like goals and shot accuracy for attackers, tackles for defenders, pass accuracy for midfielders, and saves for goalkeepers are used to rank players. An algorithm will recommend the best 11 players, with backup options, ensuring optimal choices for each match. This data-driven approach aims to improve the Indian football team's performance and FIFA ranking by identifying weak areas and enhancing player selection strategies.

**2. Title: Analysis of Football Players' Performance Using Python and Dartfish**

Abstract: The main goal of this research is to develop a predictive model for football matches using machine learning, focusing on analyzing past World Cup data. Python programming is used to implement the K-means and DPC clustering algorithms, which help in forecasting match outcomes. The DPC-K-means algorithm is applied to calculate the accuracy and probability of prediction variables, improving the reliability of the predictions. The study shows that machine learning enhances prediction accuracy by over 55%, demonstrating the effectiveness of mobile algorithm technology in forecasting results in World Cup football matches.

**3. Title: Prediction of Football Players Performance Using Machine Learning and Deep Learning Algorithms**

Abstract: The Player Performance Prediction system addresses the challenge of evaluating football players for transfers, scouting, and strategic planning by analyzing various attributes and skills. It predicts a player's performance value, helping coaches and management at all levels to identify potential talent without bias towards factors like club budgets or league competitiveness. The system uses a data-driven approach to train models that establish relationships between a player's attributes, market value, and performance, considering their position and skills. This enables more objective and effective decision-making in player evaluation and future potential assessment.

**4. Title: The Impact of Football on Cardiovascular Health**

Abstract: Football, a globally popular sport, is known for its physical demands, which make it an effective form of exercise. This study explores the cardiovascular benefits of football training, comparing it with other forms of exercise like running and cycling. The research involved 60 participants aged 20-40, divided into three groups: football training, running, and cycling. After 12 weeks of consistent training, the football group showed significant improvements in cardiovascular health, including increased VO2 max, decreased blood pressure, and lower resting heart rates, suggesting that football offers unique health benefits due to its interval-based play.

**5. Title: The Role of Team Cohesion in Football Performance**

Abstract: This paper examines the role of team cohesion in football and its impact on match performance. Using data from 15 professional teams over a season, the study identified a positive correlation between team cohesion scores and overall team success. The results suggest that well-bonded teams perform better due to improved communication, trust, and collaborative strategies. Implications for team-building activities and managerial decisions in football are discussed.

**6. Title: Injury Patterns and Risk Factors in Professional Football**

Abstract: Football players are prone to various injuries due to the sport's dynamic and physical nature. This paper reviews injury patterns among professional football players, identifying the most common injuries such as hamstring strains, ankle sprains, and ACL tears. Additionally, it highlights key risk factors, including match intensity, training load, and insufficient recovery time. The research concludes with recommendations for injury prevention strategies, focusing on strength conditioning, flexibility training, and proper recovery protocols.

**7. Title: The Psychological Effects of Penalty Shootouts in Football**

Abstract: Penalty shootouts are high-pressure situations in football that can determine the outcome of a match. This study investigates the psychological effects of penalty shootouts on players and how stress levels influence performance. Data were collected through heart rate monitoring, cortisol levels, and self-reported anxiety from 50 professional players. The findings revealed that higher stress levels led to decreased accuracy, but psychological training techniques like visualization and breathing exercises improved performance under pressure.

**8. Title: Analyzing Tactical Evolution in Modern Football**

Abstract: The tactical aspects of football have evolved significantly over the past decades. This paper analyzes the tactical trends from the 1990s to the present day, focusing on formations, pressing strategies, and positional play. Using match footage from major tournaments and leagues, the study tracks the rise of pressing, the decline of traditional 4-4-2 formations, and the increased importance of versatile midfielders. The findings suggest that modern tactics prioritize fluidity, adaptability, and possession-based play.

**9. Title: The Influence of Football on Youth Development**

Abstract: Football is not only a physical activity but also a tool for social and emotional development in young people. This research explores how participating in organized football programs influences youth development in areas such as teamwork, leadership, and discipline. A survey of 200 adolescents involved in football programs revealed improvements in social skills, self-confidence, and academic performance. The study concludes that football can be a significant contributor to the holistic development of young people.

**10. Title: Gender Disparities in Football: A Comparative Study of Men's and Women's Football**

Abstract: This paper explores gender disparities in football, comparing men's and women's professional leagues in terms of wages, media coverage, and sponsorship deals. Data were gathered from top leagues across Europe and the Americas. The results show significant discrepancies in pay and media attention, with women's football receiving only 20% of the coverage given to men's leagues. The study calls for increased investment and promotion of women's football to close this gap and achieve gender equality in the sport.

**Chapter 4: Methods**

**4.1 System Design**

**Table design**

**1.Team Stats**

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| **1. team** | The name of the football club or national team. |
|  |  |
| **2. players\_used** | The total number of different players who have participated in matches for the team during a specified period. |
|  |  |
| **3. avg\_age** | The average age of the players in the team, providing insights into the team's experience and youth. |
|  |  |
| **4. possession** | The percentage of time the team has controlled the ball during matches, reflecting their ability to maintain control. |
|  |  |
| **5. games** | The total number of matches played by the team in a specified competition or season. |
|  |  |
| **6. games\_starts** | The number of matches in which the team has been a starting participant. |
|  |  |
| **7. minutes** | The total number of minutes played by the team in all matches. |
|  |  |
| **8. goals** | The total number of goals scored by the team during the matches. |
|  |  |
| **9. assists** | The total number of assists made by players, contributing to goals scored by teammates. |
|  |  |
| **10. goals\_pens** | The total number of goals scored from penalty kicks. |
|  |  |
| **11. pens\_made** | The total number of penalty kicks successfully converted into goals. |
|  |  |
| **12. pens\_att** | The total number of penalty kicks attempted by the team. |
|  |  |
| **13. cards\_yellow** | The total number of yellow cards issued to players of the team for infractions during matches. |
|  |  |
| **14. cards\_red** | The total number of red cards issued to players of the team, leading to expulsion from the match. |
|  |  |
| **15. gk\_shots\_on\_target\_against** | The total number of shots on target faced by the team's goalkeeper. |
|  |  |
| **16. gk\_saves** | The total number of saves made by the goalkeeper. |
|  |  |
| **17. gk\_save\_pct** | The percentage of shots on target that the goalkeeper successfully saved, indicating their effectiveness. |
|  |  |
| **18. gk\_clean\_sheets** | The total number of matches where the team’s goalkeeper did not concede any goals. |
|  |  |
| **19. shots** | The total number of shots taken by the team during matches, regardless of whether they were on target. |
|  |  |
| **20. shots\_on\_target** | The total number of shots that were on target and would have resulted in a goal if not saved by the goalkeeper. |
|  |  |
| **21. goals\_per\_shot** | The ratio of goals scored to the total number of shots taken, indicating scoring efficiency. |
|  |  |
| **22. shots\_free\_kicks** | The total number of shots taken from free-kick situations. |
|  |  |
| **23. passes\_completed** | The total number of successful passes made by the team during matches. |
|  |  |
| **24. passes** | The total number of passes attempted by the team. |
|  |  |
| **25. assisted\_shots** | The total number of shots taken by teammates that were assisted by a player. |
|  |  |
| **26. passes\_free\_kicks** | The total number of passes made from free-kick situations. |
|  |  |
| **27. through\_balls** | The total number of passes that split the defense and create goal-scoring opportunities for teammates. |
|  |  |
| **28. throw\_ins** | The total number of throw-ins awarded to the team during matches. |
|  |  |
| **29. corner\_kicks** | The total number of corner kicks awarded to the team. |
|  |  |
| **30. passes\_offsides** | The total number of passes that resulted in an offside violation against the team. |
|  |  |
| **31. passes\_blocked** | The total number of passes attempted by the team that were blocked by opponents. |
|  |  |
| **32. tackles** | The total number of tackles made by the team in matches, representing defensive actions. |
|  |  |
| **33. tackles\_won** | The total number of successful tackles made by the team that resulted in regaining possession. |
|  |  |
| **34. blocks** | The total number of shots or passes that were blocked by the team's players. |
|  |  |
| **35. clearances** | The total number of times the ball was cleared from the defensive area by the team. |
|  |  |
| **36. touches** | The total number of times players from the team touched the ball during matches. |
|  |  |
| **37. dribbles\_completed** | The total number of successful dribbles made by the team’s players, indicating skill in maintaining possession. |
|  |  |
| **38. dribbles** | The total number of dribbles attempted by players on the team. |
|  |  |
| **39. passes\_received** | The total number of passes received by players from teammates. |
|  |  |
| **40. games\_complete** | The total number of matches completed by the team. |
|  |  |
| **41. points\_per\_game** | The average number of points earned per game, calculated by dividing total points by the number of games played. |
|  |  |
| **42. fouls** | The total number of fouls committed by the team during matches. |
|  |  |
| **43. fouled** | The total number of times the team’s players were fouled by opponents. |
|  |  |
| **44. offsides** | The total number of times players from the team were caught in an offside position. |
|  |  |
| **45. pens\_won** | The total number of penalty kicks awarded to the team as a result of fouls committed against them. |
|  |  |
| **46. pens\_conceded** | The total number of penalty kicks awarded to opponents as a result of fouls committed by the team. |
|  |  |
| **47. own\_goals** | The total number of goals scored by the team as a result of their own players accidentally putting the ball into their net. |
|  |  |
| **48. ball\_recoveries** | The total number of times the team regained possession of the ball after losing it, reflecting defensive resilience. |

**2.Player Stats**

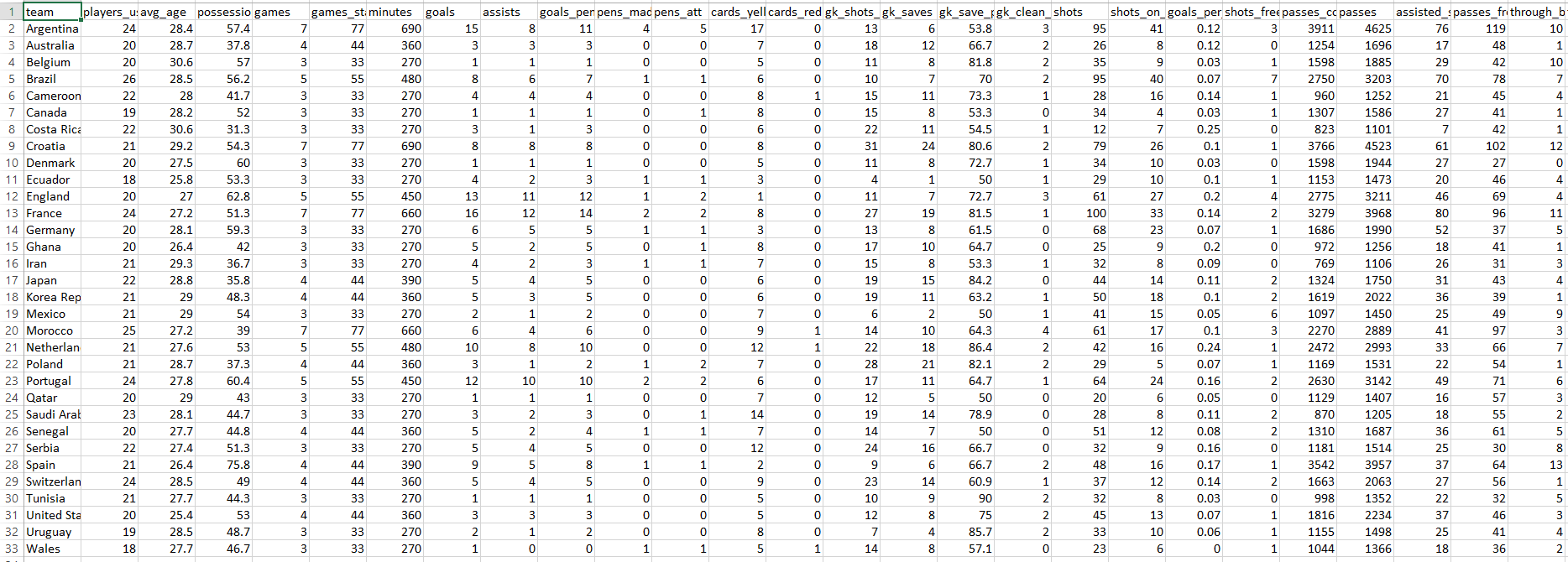
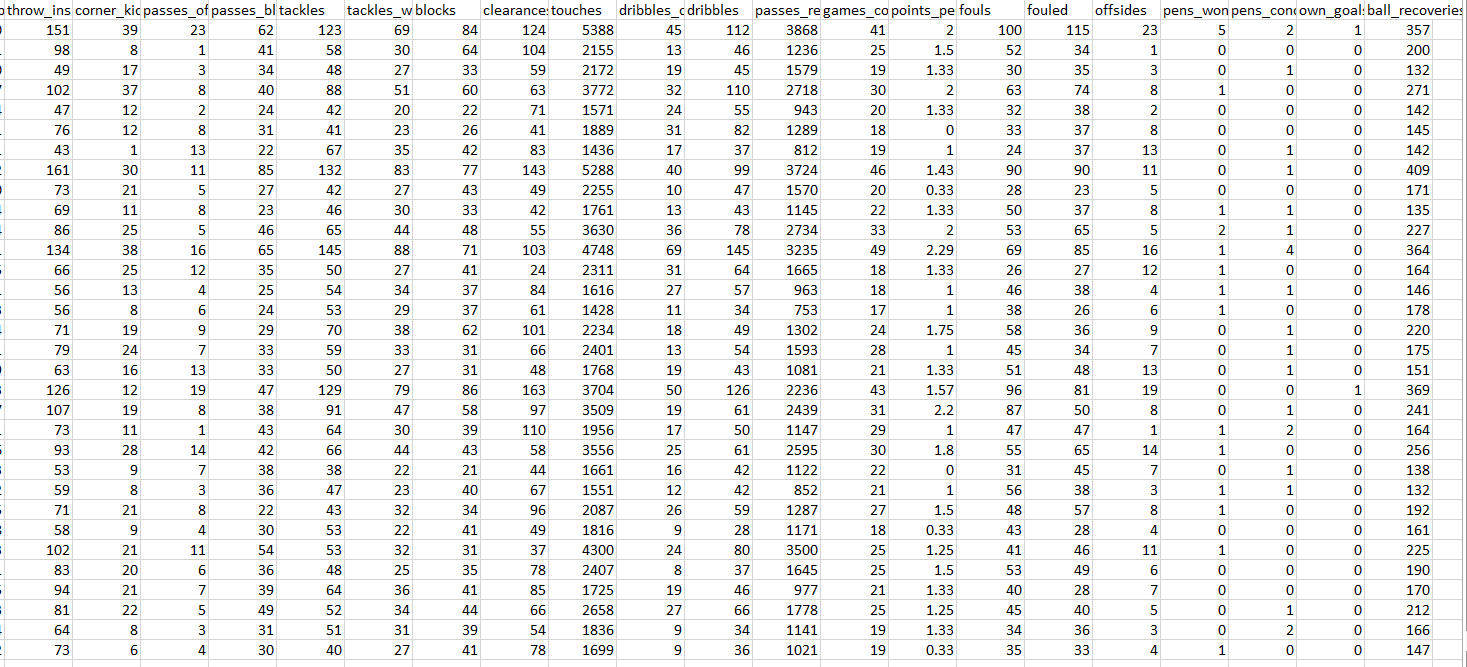
|  |  |
| --- | --- |
| **Column Name** | **Description** |
| **1. player** | The name of the football player being evaluated. |
|  |  |
| **2. position** | The specific position or role the player occupies on the field (e.g., forward, midfielder, defender, goalkeeper). |
|  |  |
| **3. team** | The name of the football club or national team the player represents. |
|  |  |
| **4. age** | The current age of the player, which can indicate experience or youth. |
|  |  |
| **5. club** | The professional football club the player is contracted to, often different from the national team they represent. |
|  |  |
| **6. birth\_year** | The year the player was born, used to calculate the player's age and analyze age trends in performance. |
|  |  |
| **7. games** | The total number of matches the player has participated in during a specified period (e.g., a season or career). |
|  |  |
| **8. games\_starts** | The number of matches in which the player was in the starting lineup. |
|  |  |
| **9. minutes** | The total number of minutes the player has spent on the field during matches. |
|  |  |
| **10. minutes\_90s** | The total number of full 90-minute matches the player has played, providing insight into their endurance and fitness. |
|  |  |
| **11. goals** | The total number of goals scored by the player in matches, an essential measure of their offensive contribution. |
|  |  |
| **12. assists** | The total number of assists made by the player, indicating their ability to create scoring opportunities for teammates. |
|  |  |
| **13. goals\_pens** | The total number of goals scored by the player from penalty kicks. |
|  |  |
| **14. pens\_made** | The total number of penalty kicks successfully converted into goals by the player. |
|  |  |
| **15. pens\_att** | The total number of penalty kicks attempted by the player. |
|  |  |
| **16. cards\_yellow** | The total number of yellow cards received by the player during matches, indicating disciplinary actions. |
|  |  |
| **17. cards\_red** | The total number of red cards received by the player, leading to their expulsion from matches. |

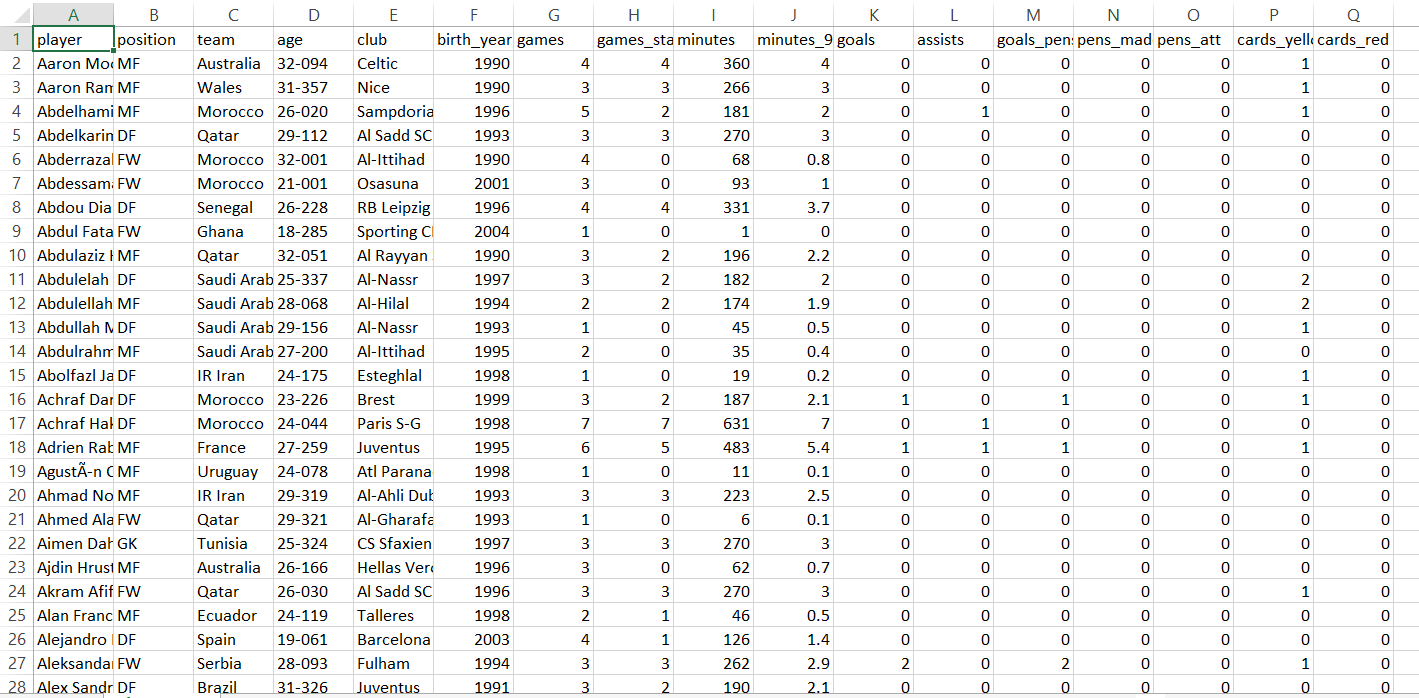
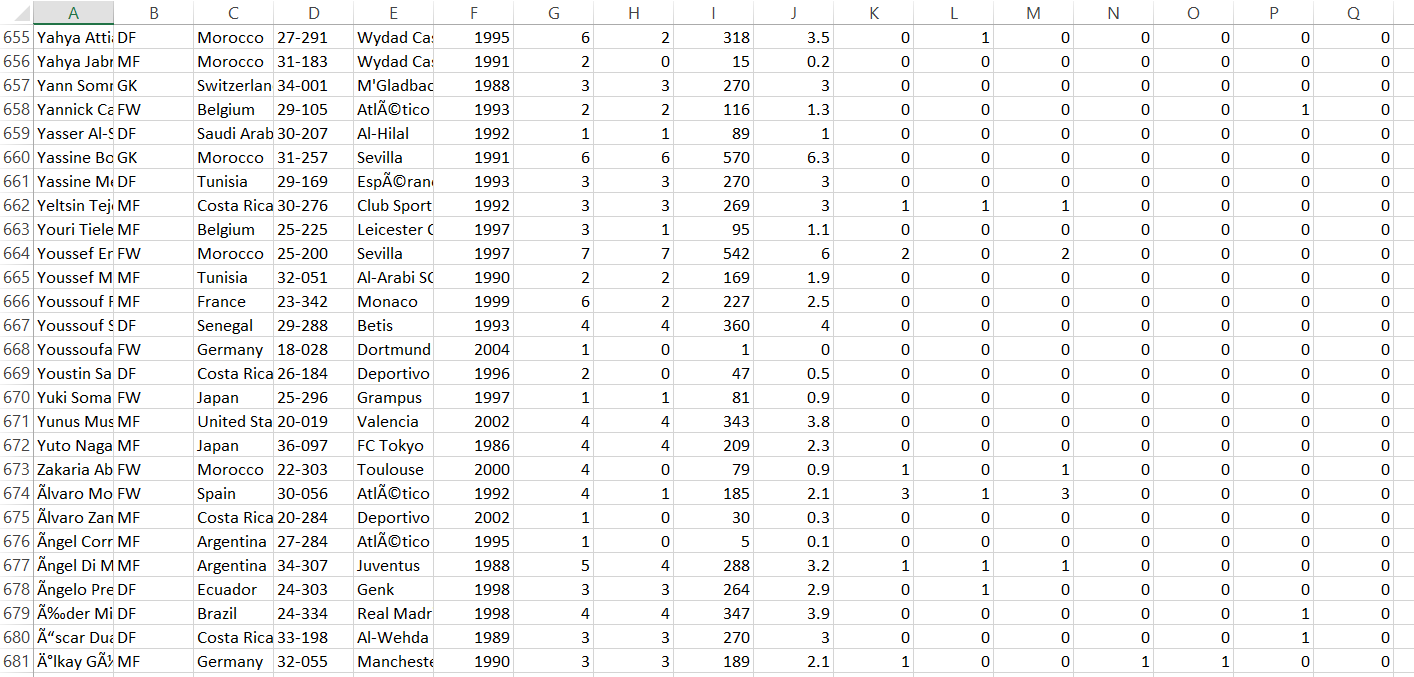
**3.Worldcup**

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| **1. Year** | The year in which the FIFA World Cup tournament took place. |
|  |  |
| **2. Country** | The host country where the FIFA World Cup was held, which may change every tournament. |
|  |  |
| **3. Winner** | The team that won the tournament, becoming the World Cup champion. |
|  |  |
| **4. Runners-Up** | The team that finished second in the tournament, losing in the final match. |
|  |  |
| **5. Third** | The team that secured third place in the tournament, often determined by a playoff match or through group stage performance. |
|  |  |
| **6. Fourth** | The team that finished in fourth place, usually the loser of the third-place playoff match. |
|  |  |
| **7. Goals Scored** | The total number of goals scored throughout the tournament by all teams combined, reflecting the offensive action of the event. |
|  |  |
| **8. Qualified Teams** | The total number of national teams that qualified to participate in the tournament, showcasing the global nature of the competition. |
|  |  |
| **9. Matches Played** | The total number of matches played during the tournament, including group stage and knockout rounds. |
|  |  |
| **10. Attendance** | The total number of spectators who attended the matches, indicating the popularity and draw of the tournament. |

**4.2 Data Collection**

**Kaggle Data Sets**

**1.Team**

**2.Player**

**3.Worldcup**

**4.3 Statistical Analysis**

1. **Player Performance Evaluation**

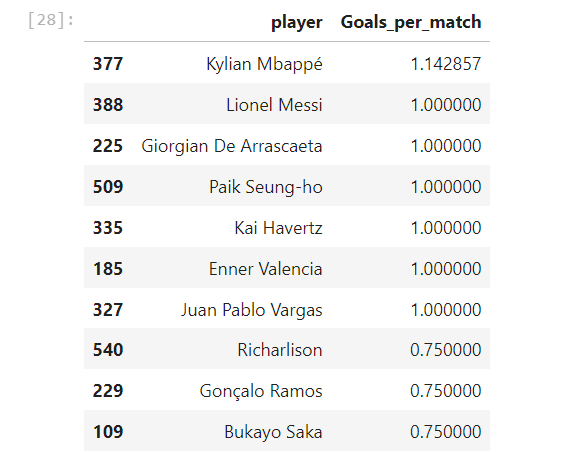
# Calculate goals per match

players['Goals\_per\_match'] = players['goals'] / players['games']

# Display top scorers based on goals per match

top\_scorers = players.sort\_values(by='Goals\_per\_match', ascending=False).head(10)

top\_scorers[['player', 'Goals\_per\_match']]

****

1. **Predicting Match Outcomes**

# Since matches have no result we use a proxy for Match\_Result based on goals

# This is a simplified approach to adjust it as needed for our data.

teams['Match\_Result'] = (teams['goals'] > teams['goals'].mean()).astype(int)

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score

# Calculate Pass Accuracy

teams['Pass\_Accuracy'] = (teams['passes\_completed'] / teams['passes']) \* 100

# Prepare data for training

features = teams[['goals', 'possession', 'Pass\_Accuracy']]

labels = teams['Match\_Result']

# Split data

X\_train, X\_test, y\_train, y\_test = train\_test\_split(features, labels, test\_size=0.2, random\_state=42)

# Train the model

model = LogisticRegression()

model.fit(X\_train, y\_train)

# Test the model

predictions = model.predict(X\_test)

print("Accuracy:", accuracy\_score(y\_test, predictions))



1. **Improved Tactical Understanding**

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

# Calculate Pass Accuracy

teams['Pass\_Accuracy'] = (teams['passes\_completed'] / teams['passes']) \* 100

# Apply clustering (excluding Goals\_Conceded if not available)

kmeans = KMeans(n\_clusters=4, random\_state=0).fit(teams[['possession', 'Pass\_Accuracy']])

teams['Play\_Style'] = kmeans.labels\_

# Visualize clusters

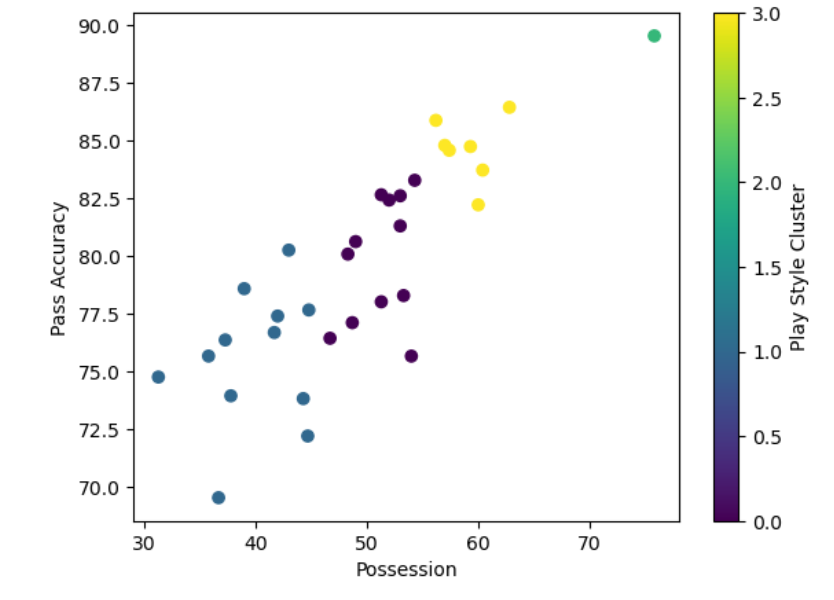
plt.scatter(teams['possession'], teams['Pass\_Accuracy'], c=teams['Play\_Style'], cmap='viridis')

plt.xlabel('Possession')

plt.ylabel('Pass Accuracy')

plt.colorbar(label='Play Style Cluster')

plt.show()

****

1. **Team Management and Player Selection**

# Check players with position as 'Forward'

forwards = players[players['position'] == 'FW']

print("Total forwards:", len(forwards))

# Check players with Goals per Match greater than 0.3

high\_scorers = players[players['Goals\_per\_match'] > 0.3]

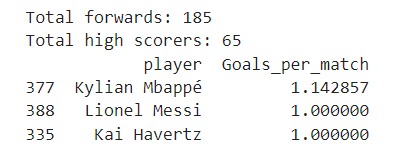
print("Total high scorers:", len(high\_scorers))

# Try a lower threshold for Goals per Match

selected\_players = players[(players['position'] == 'FW') & (players['Goals\_per\_match'] > 0.1)]

selected\_players = selected\_players.sort\_values(by='Goals\_per\_match', ascending=False).head(3)

print(selected\_players[['player', 'Goals\_per\_match']])

****

1. **World Cup Wins Analysis**

# Count the number of World Cups each country has won

world\_cup\_wins = world\_cups['Winner'].value\_counts()

print(world\_cup\_wins)

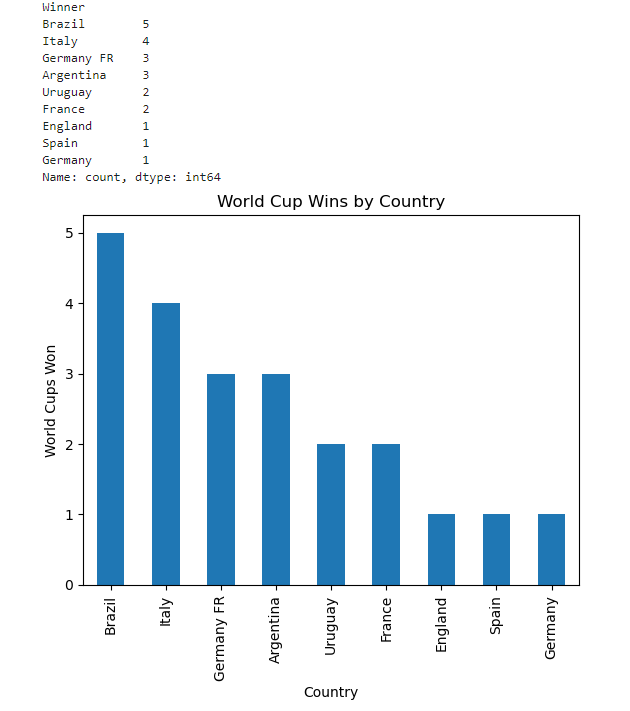
# Plot world cup wins

world\_cup\_wins.plot(kind='bar', title="World Cup Wins by Country")

plt.xlabel("Country")

plt.ylabel("World Cups Won")

plt.show()



**4.4 Implementation**

**Fetch**

**import statements**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

**Load the data**

players = pd.read\_csv("player\_stats.csv")

teams = pd.read\_csv("team\_data.csv")

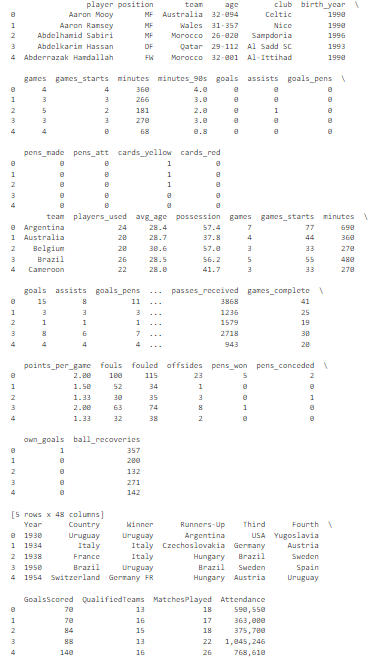
world\_cups = pd.read\_csv("world\_cups.csv")

**Display basic info**

print(players.head())

print(teams.head())

print(world\_cups.head())

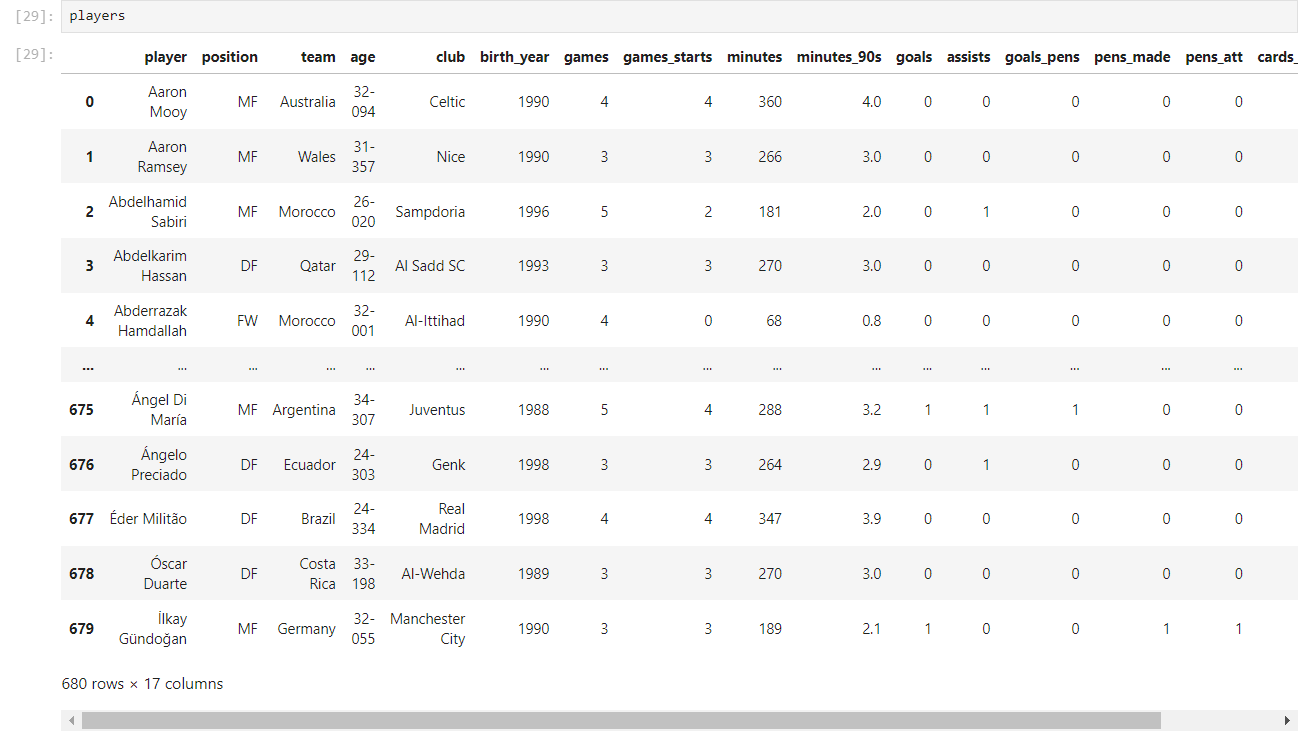


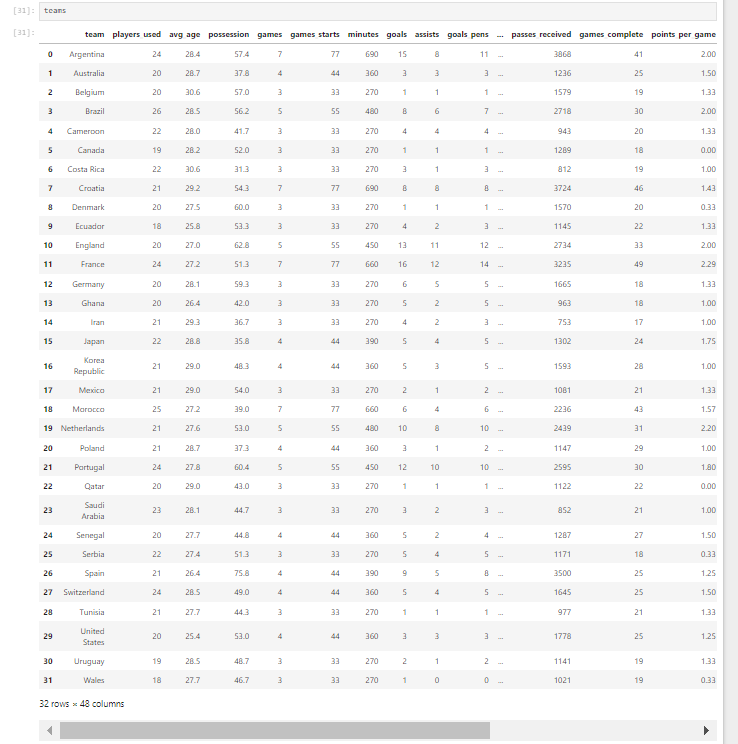
**Check for missing values and handle them**

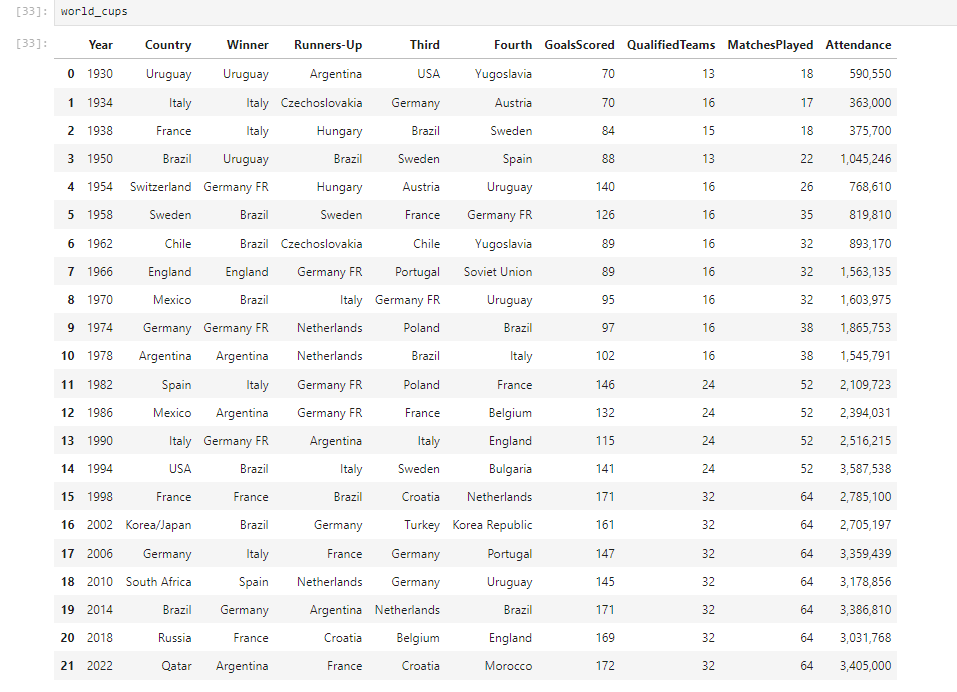
players.fillna(0, inplace=True)

teams.fillna(0, inplace=True)

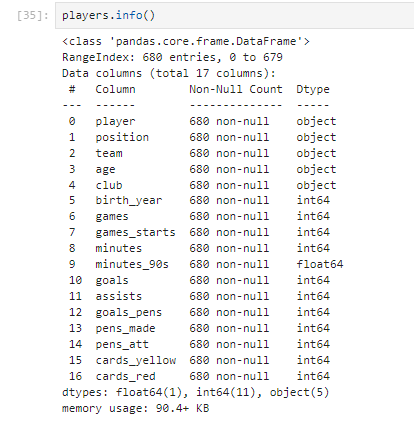
world\_cups.fillna(0, inplace=True)

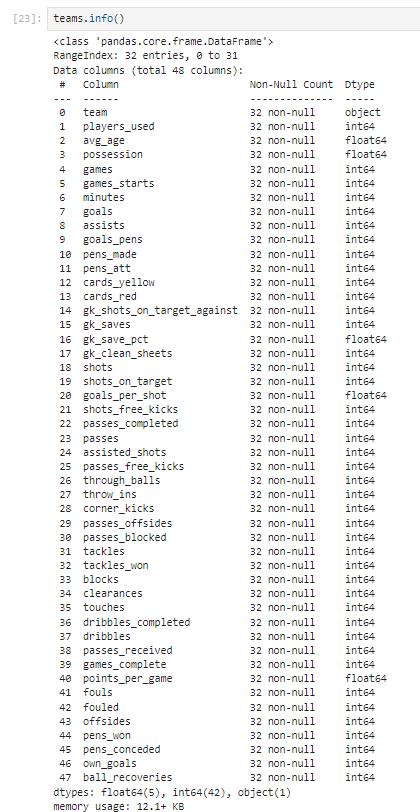
****

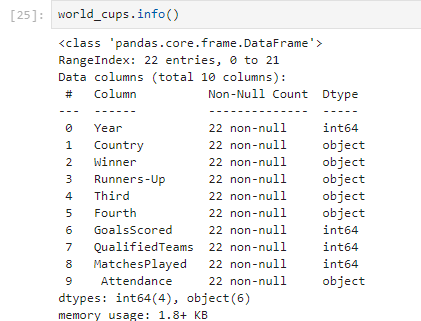
****

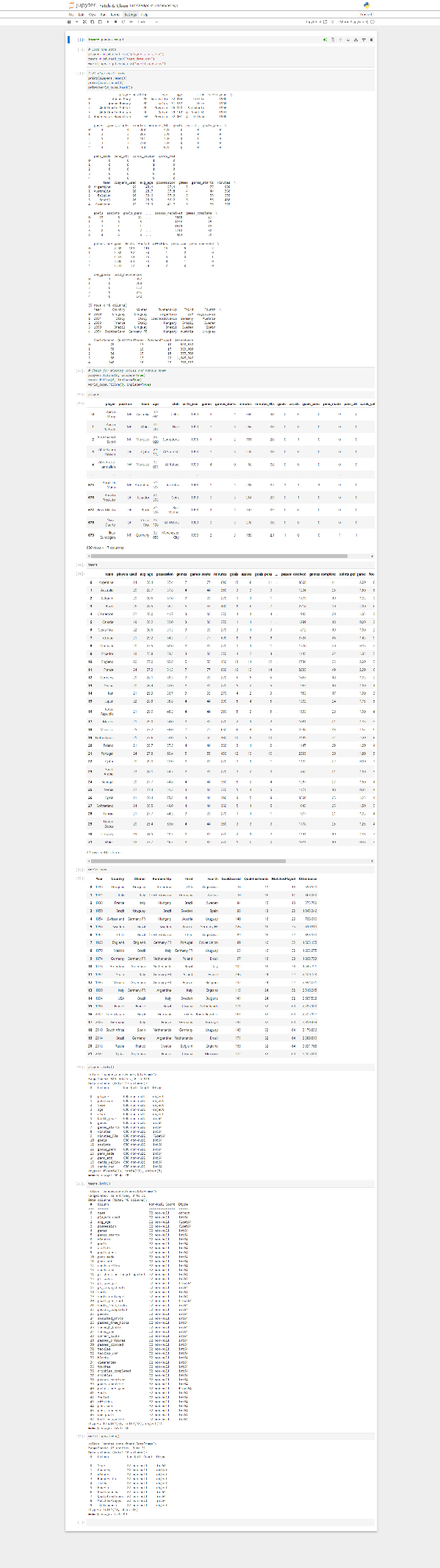
****

**Information of Each Data Sets**

****

****

****

**Fetch & Clean – Full Page Screenshot 🡪4.5 Data Visualization**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

**1. Player Performance Comparison**

# Plot for Top Players by Goals per Match

plt.figure(figsize=(12, 6))

sns.barplot(data=top\_scorers, x='player', y='Goals\_per\_match', color='skyblue')

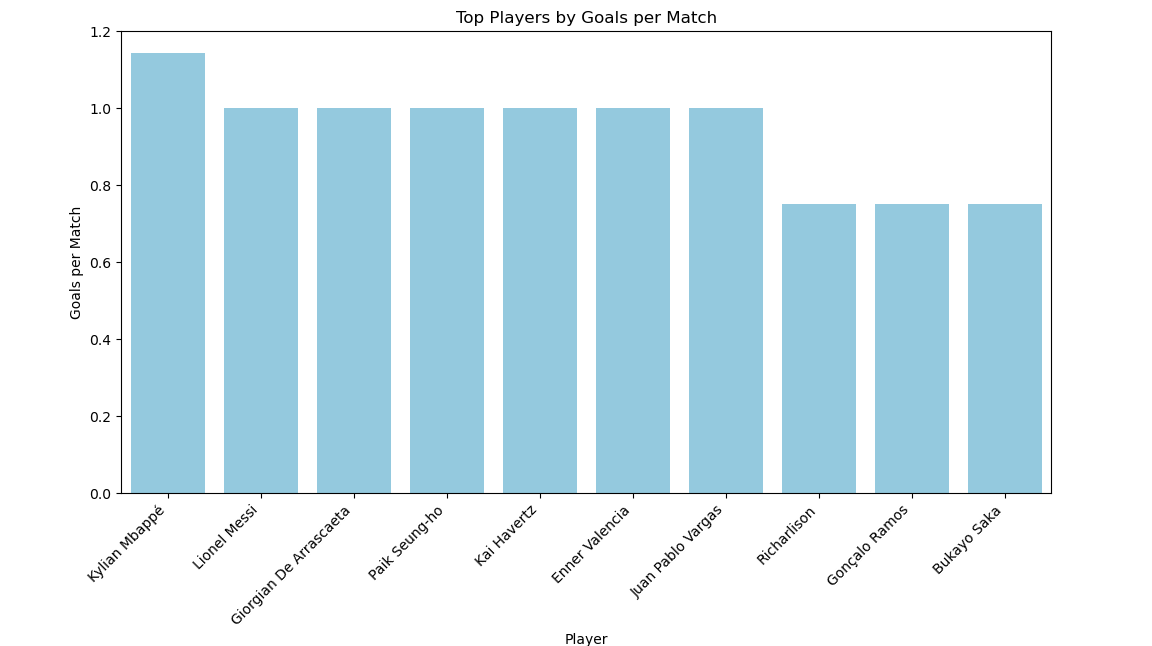
plt.xticks(rotation=45, ha="right")

plt.title("Top Players by Goals per Match")

plt.xlabel("Player")

plt.ylabel("Goals per Match")

plt.show()

****

**2. Team Performance**

top\_teams = teams[['team', 'goals']].sort\_values(by='goals', ascending=False).head(10)

plt.figure(figsize=(12, 6))

sns.barplot(data=top\_teams, x='team', y='goals', palette='viridis')

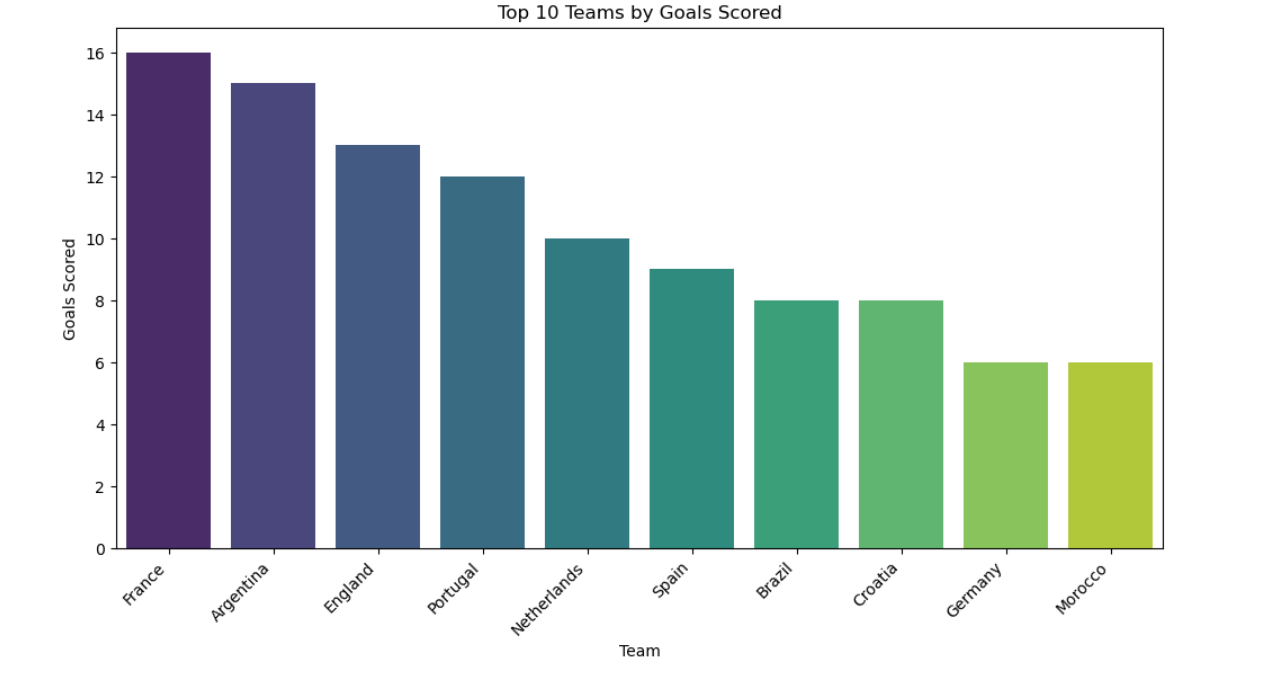
plt.xticks(rotation=45, ha="right")

plt.title('Top 10 Teams by Goals Scored')

plt.xlabel('Team')

plt.ylabel('Goals Scored')

plt.show()



**3. World Cup Wins by Country (Bar Chart)**

plt.figure(figsize=(10, 6))

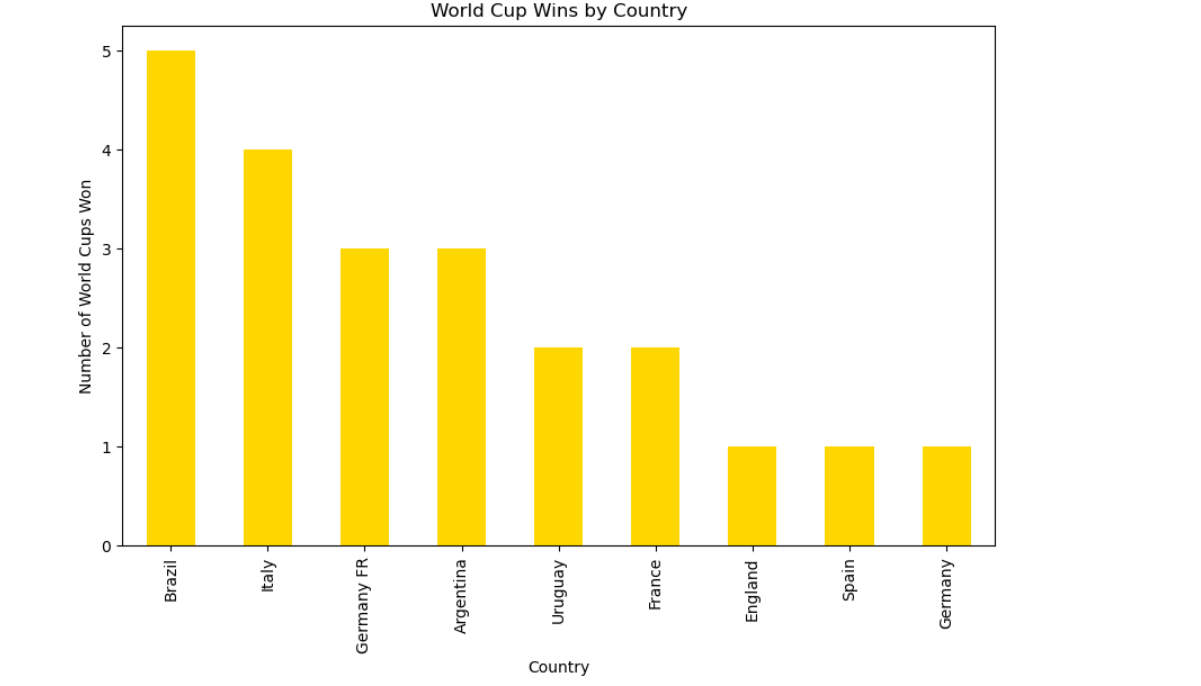
world\_cup\_wins.plot(kind='bar', color='gold')

plt.title("World Cup Wins by Country")

plt.xlabel("Country")

plt.ylabel("Number of World Cups Won")

plt.show()



**4. Team Play Styles (Clustering Result)**

# Calculate Pass Accuracy if not already calculated

teams['Pass\_Accuracy'] = (teams['passes\_completed'] / teams['passes']) \* 100

plt.figure(figsize=(12, 8))

sns.scatterplot(data=teams, x='possession', y='Pass\_Accuracy', hue='Play\_Style', palette='viridis', s=100)

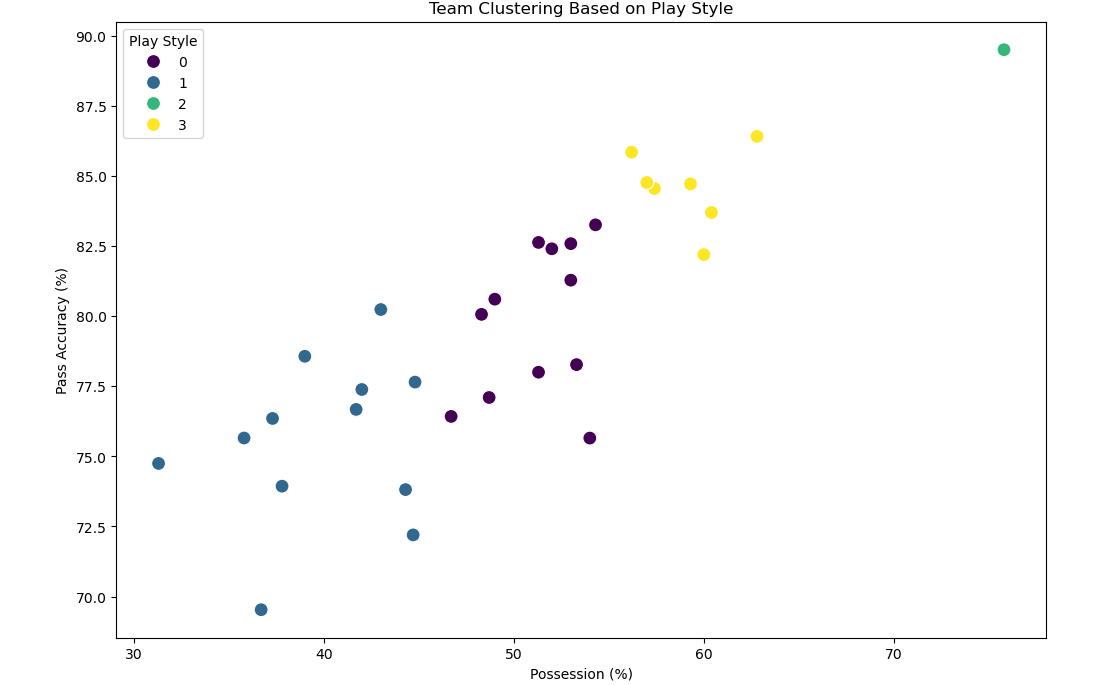
plt.title("Team Clustering Based on Play Style")

plt.xlabel("Possession (%)")

plt.ylabel("Pass Accuracy (%)")

plt.legend(title='Play Style')

plt.show()



**5. Correlation Heatmap for Player Stats**

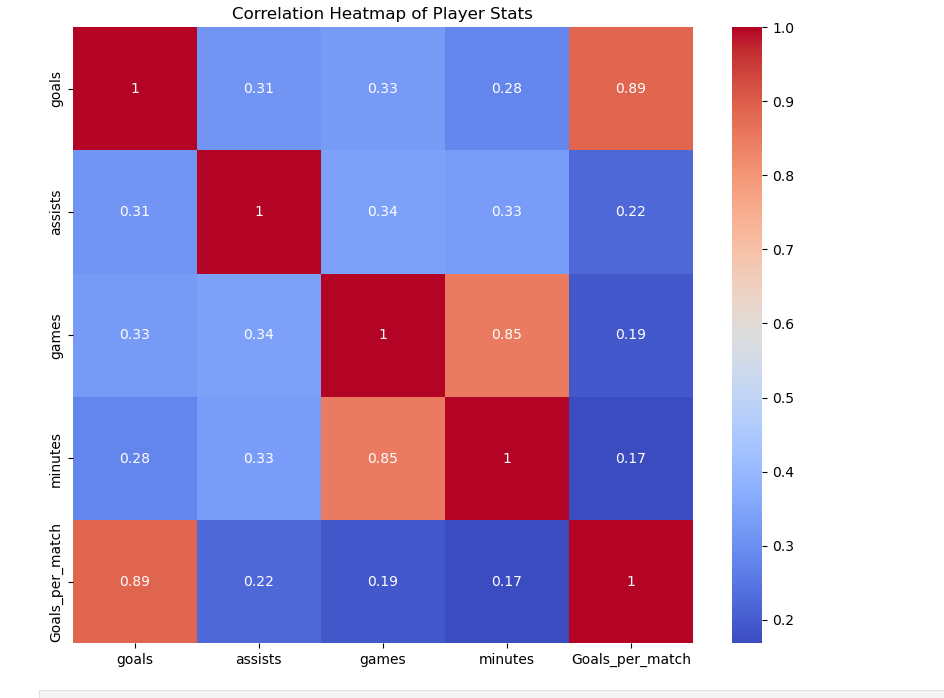
plt.figure(figsize=(10, 8))

correlation\_matrix = players[['goals', 'assists', 'games', 'minutes', 'Goals\_per\_match']].corr()

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm')

plt.title("Correlation Heatmap of Player Stats")

plt.show()



**4.6 Model Development and Evolution**

**Step 1: Predicting Player Performance (Goals Prediction)**

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error, r2\_score

# Selecting features

X = players[['minutes', 'games', 'assists']]

y = players['goals']

# Split data

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Model training

model = LinearRegression()

model.fit(X\_train, y\_train)

# Predictions

y\_pred = model.predict(X\_test)

# Evaluation

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

print(f'Mean Squared Error: {mse:.2f}')

print(f'R² Score: {r2:.2f}')

****

**Step 2: Clustering Players by Play Style**

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

import seaborn as sns

# Selecting features for clustering

features = players[['goals', 'assists', 'games', 'minutes']]

kmeans = KMeans(n\_clusters=3, random\_state=42)

players['Play\_Style'] = kmeans.fit\_predict(features)

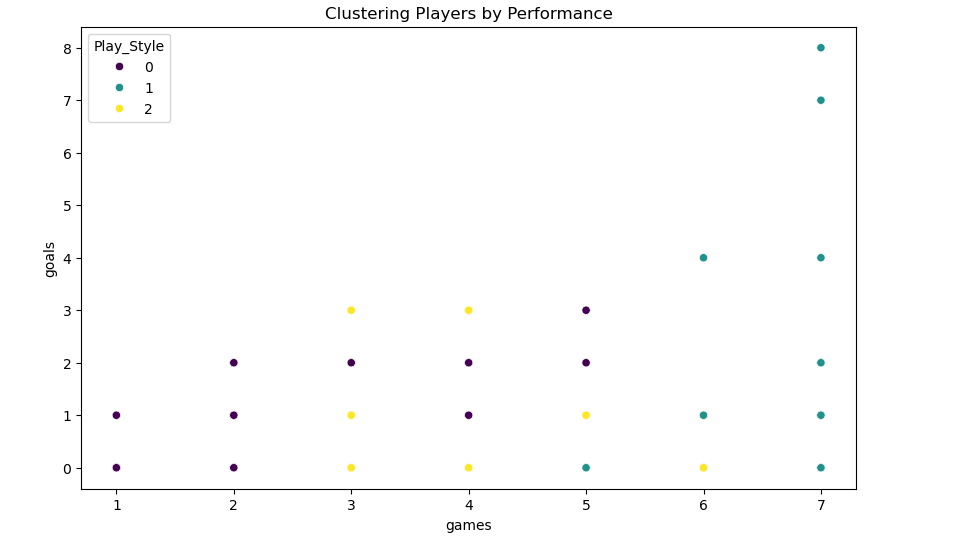
# Visualizing clusters

plt.figure(figsize=(10, 6))

sns.scatterplot(data=players, x='games', y='goals', hue='Play\_Style', palette='viridis')

plt.title("Clustering Players by Performance")

plt.show()

****

**1. KMeans Clustering: Grouping Teams by Play Style**

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

import seaborn as sns

import pandas as pd

# Calculate Pass Accuracy if not already calculated

teams['Pass\_Accuracy'] = (teams['passes\_completed'] / teams['passes']) \* 100

# Select relevant features for clustering

team\_features = teams[['possession', 'Pass\_Accuracy', 'goals']]

# Initialize and fit KMeans

kmeans = KMeans(n\_clusters=4, random\_state=42)

teams['Play\_Style'] = kmeans.fit\_predict(team\_features)

# Plot the clustering results

plt.figure(figsize=(12, 8))

sns.scatterplot(data=teams, x='possession', y='Pass\_Accuracy', hue='Play\_Style', palette='viridis', s=100)

plt.title("KMeans Clustering: Team Play Style")

plt.xlabel("Possession (%)")

plt.ylabel("Pass Accuracy (%)")

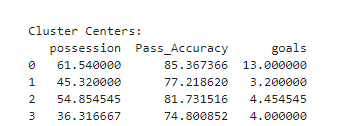
plt.legend(title='Play Style Cluster')

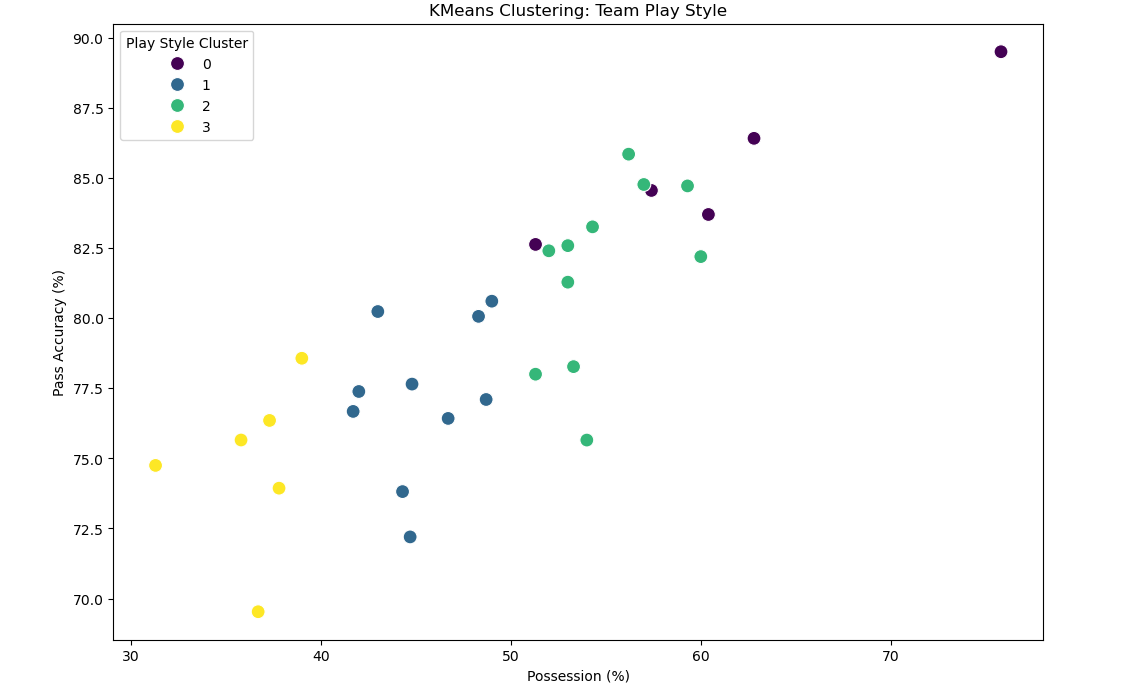
plt.show()

# View cluster centers to analyze the play styles

print("Cluster Centers:")

print(pd.DataFrame(kmeans.cluster\_centers\_, columns=team\_features.columns))

****

****

from sklearn.cluster import KMeans

from sklearn.metrics import silhouette\_score

teams['Pass\_Accuracy'] = (teams['passes\_completed'] / teams['passes']) \* 100

# Perform clustering with KMeans

kmeans = KMeans(n\_clusters=4, random\_state=42)

teams['Play\_Style'] = kmeans.fit\_predict(teams[['possession', 'Pass\_Accuracy', 'goals']])

# Calculate Silhouette Score

silhouette\_avg = silhouette\_score(teams[['possession', 'Pass\_Accuracy', 'goals']], kmeans.labels\_)

print("Silhouette Score for KMeans Clustering:", silhouette\_avg)



**2. K-Nearest Neighbors (KNN): Predicting Match Outcomes**

from sklearn.model\_selection import train\_test\_split

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import accuracy\_score, classification\_report, ConfusionMatrixDisplay

import matplotlib.pyplot as plt

# Calculate Pass Accuracy if not already calculated

teams['Pass\_Accuracy'] = (teams['passes\_completed'] / teams['passes']) \* 100

# Create a basic rule for Match\_Result

teams['Match\_Result'] = teams['goals'].apply(lambda x: 1 if x > 1 else 0)

# Prepare features and labels

features = teams[['goals', 'possession', 'Pass\_Accuracy']]

labels = teams['Match\_Result'] # 1 for Win, 0 for Loss/Draw

# Split data

X\_train, X\_test, y\_train, y\_test = train\_test\_split(features, labels, test\_size=0.2, random\_state=42)

# Initialize and train KNN model

knn = KNeighborsClassifier(n\_neighbors=5) # Using k=5

knn.fit(X\_train, y\_train)

# Make predictions and evaluate

y\_pred = knn.predict(X\_test)

print("KNN Accuracy:", accuracy\_score(y\_test, y\_pred))

print("\nClassification Report:\n", classification\_report(y\_test, y\_pred))

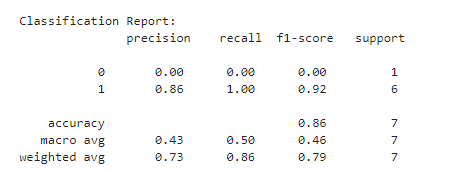
# Plot Confusion Matrix

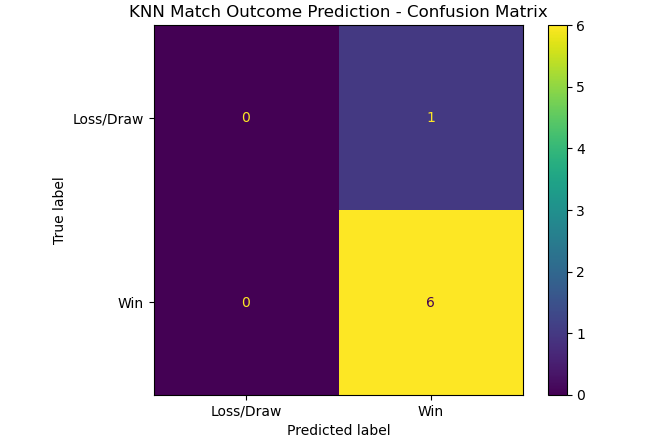
ConfusionMatrixDisplay.from\_predictions(y\_test, y\_pred, display\_labels=['Loss/Draw', 'Win'])

plt.title("KNN Match Outcome Prediction - Confusion Matrix")

plt.show()



****

****

**3. Principal Component Analysis (PCA) for Dimensionality Reduction**

from sklearn.decomposition import PCA

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

team\_features\_scaled = scaler.fit\_transform(team\_features)

# Apply PCA

pca = PCA(n\_components=2)

principal\_components = pca.fit\_transform(team\_features\_scaled)

teams['PCA1'], teams['PCA2'] = principal\_components[:, 0], principal\_components[:, 1]

plt.figure(figsize=(10, 8))

sns.scatterplot(data=teams, x='PCA1', y='PCA2', hue='Play\_Style', palette='Set2', s=100)

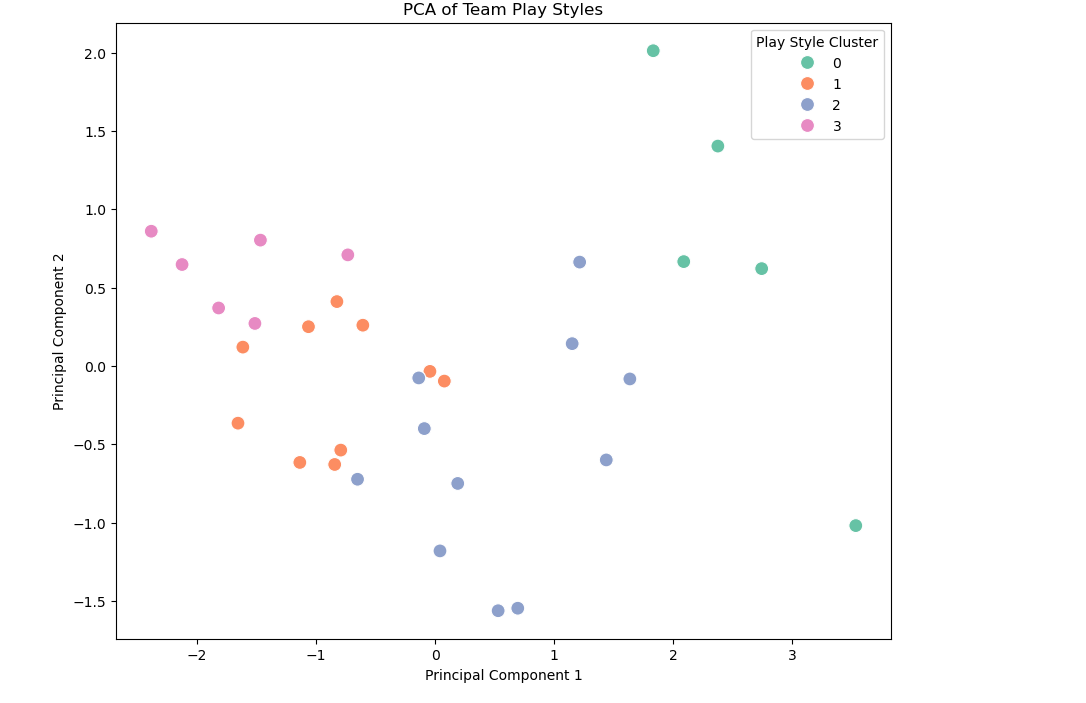
plt.title("PCA of Team Play Styles")

plt.xlabel("Principal Component 1")

plt.ylabel("Principal Component 2")

plt.legend(title='Play Style Cluster')

plt.show()



**4. Predicting Player Performance using Linear Regression**

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error, r2\_score

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

# Calculate Goals per Match if not already calculated

players['Goals\_per\_match'] = players['goals'] / players['games']

# Define features and target (without Tackles)

player\_features = players[['minutes', 'assists']] # Adjusting for available columns

target = players['Goals\_per\_match']

# Split data

X\_train, X\_test, y\_train, y\_test = train\_test\_split(player\_features, target, test\_size=0.2, random\_state=42)

# Train Linear Regression model

regressor = LinearRegression()

regressor.fit(X\_train, y\_train)

# Predict and evaluate

y\_pred = regressor.predict(X\_test)

print("Mean Squared Error:", mean\_squared\_error(y\_test, y\_pred))

print("R^2 Score:", r2\_score(y\_test, y\_pred))

# Plot Actual vs Predicted values

plt.figure(figsize=(10, 6))

plt.scatter(y\_test, y\_pred, alpha=0.7, color="blue")

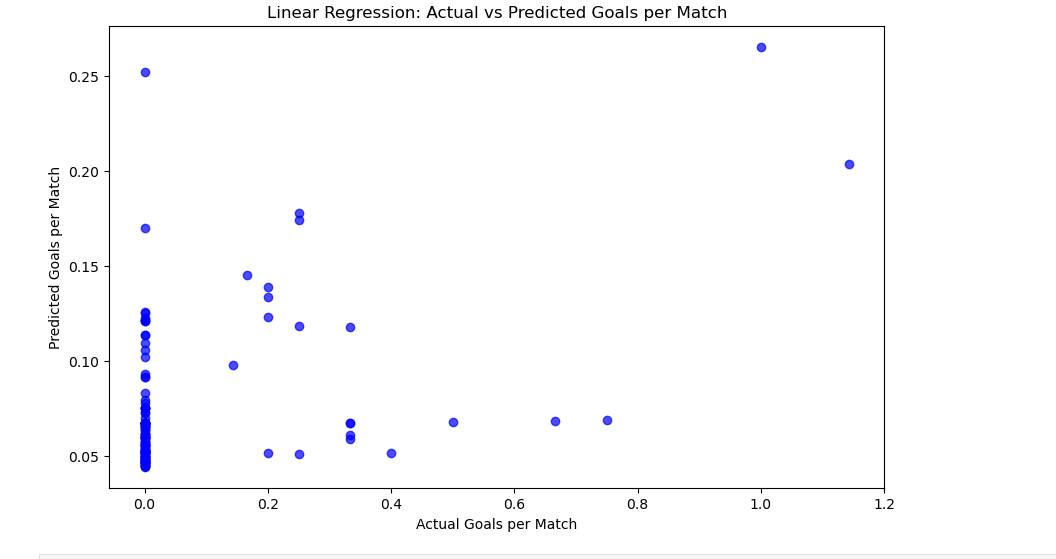
plt.xlabel("Actual Goals per Match")

plt.ylabel("Predicted Goals per Match")

plt.title("Linear Regression: Actual vs Predicted Goals per Match")

plt.show()

****

****

**4.7 Result**

**Predicting FIFA 2026 Tournament Winner**

We can use Monte Carlo Technique for the future prediction of the tournament.

Below are the technique which is used for predicting the winner.

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score, classification\_report

teams['Match\_Result'] = (teams['goals'] > teams['goals'].mean()).astype(int)

teams['Pass\_Accuracy'] = (teams['passes\_completed'] / teams['passes']) \* 100

# Feature set (e.g., possession, goals scored, goals conceded, player stats)

features = teams[['possession', 'goals', 'Pass\_Accuracy']] # Update with relevant columns

labels = teams['Match\_Result'] # Win (1), Loss (0)

# Train-test split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(features, labels, test\_size=0.2, random\_state=42)

# Initialize Random Forest model for classification

rf\_classifier = RandomForestClassifier(random\_state=42)

rf\_classifier.fit(X\_train, y\_train)

# Predict match outcomes (win/loss) for the test set

y\_pred = rf\_classifier.predict(X\_test)

# Evaluate model performance

accuracy = accuracy\_score(y\_test, y\_pred)

print(f"Accuracy of the match outcome prediction model: {accuracy:.2f}")

# Future prediction: Use the model to predict 2026 tournament matches

# Example: Predict the winner of match between Team\_A and Team\_B

match\_data = [[55, 2, 1, 80]] # Example: [Possession, Goals\_Scored, Goals\_Conceded, Pass\_Accuracy]

predicted\_outcome = rf\_classifier.predict(match\_data)

print(f"Predicted Match Outcome: {'Win' if predicted\_outcome == 1 else 'Loss'}")



**4.8 Analysis and Testing**

|  |  |
| --- | --- |
| **Model Name** | **Accuracy** |
| **KMeans Clustering** | **Silhouette Score: 0.31054393887977333** |
| **K-Nearest Neighbours (KNN)** | **0.8571428571428571** |
| **Principal Component Analysis (PCA)** | **Graph** |
| **Linear Regression** | **Mean Squared Error: 0.02655452704756836**  **R^2 Score: 0.14786525217641766** |

**Chapter 5: Conclusion**

This AI-driven approach to football analysis provides objective insights that empower data-informed decisions across team management, match predictions, and player performance evaluations. By combining clustering, classification, and regression techniques, teams can better understand their strengths, optimize their lineups, and improve pre-match strategies.

Future Directions:

* Explore advanced models such as ensemble learning or neural networks for richer predictions.
* Continuously validate and update models with new data to keep analyses current.
* Integrate external factors like weather, crowd influence, and season phases to improve contextual accuracy.

By implementing these recommendations, the analytical models will become more precise, making them valuable tools for teams, coaches, and sports analysts.

**Chapter 6: Recommendation**

**1. Enhance Data Quality and Feature Selection**

* Data Consistency: Use high-quality, consistent data sources, especially for team and player statistics. Inconsistent data can affect model performance, particularly for KMeans and Linear Regression.
* Feature Engineering: Introduce new features, such as opponent strength, home vs. away performance, and player fatigue to capture contextual factors impacting outcomes.
* Incorporate Advanced Metrics: Integrate advanced football metrics like expected goals (xG), expected assists (xA), and progressive passes to better capture play styles and individual contributions.

**2. Experiment with Advanced Models and Tuning**

* Explore Ensemble Models: Techniques like Random Forests and Gradient Boosting often outperform simpler models by combining multiple decision trees, enhancing predictive accuracy for KNN and regression tasks.
* Hyperparameter Tuning: Use methods like GridSearchCV or RandomizedSearchCV to fine-tune KMeans clusters and KNN neighbors. This ensures the models are optimized for the specific structure of your dataset.
* Non-Linear Models for Player Performance: If Linear Regression shows poor fit (low R² score), try Polynomial Regression or Neural Networks to capture potential non-linear relationships, especially for player statistics.

**3. Model Validation and Robust Testing**

* Cross-Validation: Implement k-fold cross-validation across all models to confirm stability and avoid overfitting on a single data split.
* Residual and Error Analysis: Regularly assess residuals in regression models to check for patterns or biases. Large residuals may indicate missing features or areas where model performance could be improved.
* Update Models with Seasonal Data: Football dynamics change seasonally; models should be re-trained periodically with the latest data to adapt to new strategies, player transfers, and form changes.

**4. Focus on Interpretability and Visualization**

* Feature Importance: For models like Random Forest, extract and visualize feature importances to understand which factors most influence outcomes. This can provide actionable insights for coaches.
* Cluster Visualization with PCA or t-SNE: For KMeans, regularly visualize clusters with PCA or t-SNE to verify that team styles remain relevant. This also helps analysts understand shifts in team dynamics over time.
* Performance Metrics Dashboard: Create an interactive dashboard (using tools like Tableau or Power BI) that shows model predictions, performance metrics, and trend analysis for real-time insights during the season.

**5. Integrate Real-Time and Contextual Factors**

* Injury and Player Availability: Incorporate real-time data on player injuries, availability, and rotation. These affect match outcomes and player performance predictions significantly.
* Environmental Variables: Factors like weather, stadium conditions, and crowd size can impact game play and should be integrated for more accurate match predictions.
* Match Context: Consider factors like match importance (league vs. friendly), which can affect player motivation and team lineup strategies, especially for predictive models.

**6. Expand Model Applications in Game Strategy**

* Pre-Game and In-Game Decision Support: Use predictions to help coaches in real-time game strategy adjustments. For instance, if a model indicates a high likelihood of conceding goals, coaches can adjust defensive tactics.
* Player Rotation Optimization: By predicting performance drop-off or injury likelihood, optimize player rotation and substitution plans to keep the team fit and competitive across the season.
* Opponent Analysis: Use clustering and predictive models to identify and counter specific tactics of upcoming opponents based on historical data and identified play styles.

**Chapter 7: References**

**Books & Research Papers**

1. "Python for Data Analysis" by Wes McKinney
   * This book offers practical guidance on data manipulation, analysis, and visualization using Python’s pandas library, which is crucial for any data-driven sports analysis project.
2. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron
   * Covers machine learning algorithms with practical Python examples, including KMeans, KNN, and regression techniques. This is ideal for implementing predictive models in football analysis.
3. "Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers" by Benjamin C. Alamar
   * Focuses on how data science and analytics are used in sports, including concepts applicable to team performance analysis, player selection, and game strategy.
4. "Practical Time Series Analysis" by Aileen Nielsen
   * Provides methods for analyzing and forecasting time series data, which is useful for predicting match outcomes and analyzing player or team trends over time.
5. "Machine Learning Yearning" by Andrew Ng (available for free online)
   * A concise guide on how to set up machine learning projects and make effective improvements, which can help structure and refine football analysis models.
6. "Introduction to Sports Biomechanics: Analysing Human Movement Patterns" by Roger Bartlett
   * Although focused on biomechanics, this book offers insights into measuring and interpreting player movement, which can enhance player performance analysis.

**Online Resources**

1. Prediction Model of Football World Cup Championship – 13th September 2021.
2. Analysis Of Football Players’ Performance Using Python And Dartfish V.V. Prasanth1 , G.Nallavan2 – 05th May 2022.
3. Prediction Of Football Players Performance Using Machine Learning Algorithms - March 2022.
4. Kaggle FIFA World Cup 2022 Datasets.