School of Computer Science and Engineering
(Computer Science & Engineering- Artificial Intelligence & Data Engineering)
Faculty of Engineering & Technology
Jain Global Campus, Kanakapura Taluk - 562112
Ramanagara District, Karnataka, India

2023-2024
(IV Semester)

A Project Report on

"SPORTS DATA ANALYSIS"

Submitted in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING-ARTIFICIAL INTELLIGENCE & DATA ENGINEERING

Submitted by

AKANSHA SHETTY, CHIMIRALA KOWSTUBHA, KAPAROTU VENKATA SURYA THARANI
22BTRAD002, 22BTRAD012, 22BTRAD018

Under the guidance of
Mr. Akash Das
AVP and Project Manager
Futurense Technologies

Department of Computer Science and Engineering- Artificial Intelligence & Data Engineering

School of Computer Science & Engineering

Faculty of Engineering & Technology

Jain Global Campus, Kanakapura Taluk - 562112

Ramanagara District, Karnataka, India

CERTIFICATE

This is to certify that the project work titled **"SPORTS DATA ANALYSIS"** is carried out by **Akansha Shetty (22BTRAD002), Chimirala Kowstubha (22BTRAD012), Kaparotu Venkata Surya Tharani (22BTRAD018),** a bonafide students of Bachelor of Technology at the School of Engineering & Technology, Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Bangalore in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering- Artificial Intelligence & Data Engineering, during the year **2023-2024**.

Mr. Akash Das
AVP and Project Manager

Date:

Dr. Sathish Kumar R
Program Head,
Computer Science and Engineering- Artificial Intelligence & Data Engineering,
School of Computer Science & Engineering
Faculty of Engineering & Technology
JAIN (Deemed to-be University)
Date:

Dr. Geetha G
Director,
School of Computer Science & Engineering
Faculty of Engineering & Technology
JAIN (Deemed to-be University)
Date:

Name of the Examiner

Signature of Examiner

1.

2.

DECLARATION

We , **Akansha Shetty (22BTRAD002), Chimirala Kowstubha (22BTRAD012), Kaparotu Venkata Surya Tharani (22BTRAD018)** students of IVth semester B.Tech in **Computer Science and Engineering- Artificial Intelligence & Data Engineering**, at School of Engineering & Technology, Faculty of Engineering & Technology, **JAIN (Deemed to-be University)**, hereby declare that the internship work titled **"Sports Data Analysis"** has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor of Technology in Computer Science and Engineering- Artificial Intelligence & Data Enginnering** during the academic year **2023-2024**. Further, the matter presented in the work has not been submitted previously by anybody for the award of any degree or any diploma to any other University, to the best of our knowledge and faith.

Name1: Akansha Shetty Signature
USN : 22BTRAD002
Name2: Chimirala Kowstubha Signature
USN : 22BTRAD012
Name3: Kaparotu Venkata Surya Tharani Signature
USN : 22BTRAD018

Place : Bangalore
Date :

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to Faculty of Engineering & Technology, JAIN (Deemed to-be University) for providing us with a great opportunity to pursue my Bachelors Degree in this institution.

We are deeply thankful to several individuals whose invaluable contributions have made this project a reality. We wish to extend our heartfelt gratitude to ***Dr. Chandraj Roy Chand, Chancellor****, for his tireless commitment to fostering excellence in teaching and research at Jain (Deemed-to-be-University). We are also profoundly grateful to the honorable* ***Vice Chancellor, Dr. Raj Singh, and Dr. Dinesh Nilkant, Pro Vice Chancellor****, for their unwavering support. Furthermore, we would like to express our sincere thanks to* ***Dr. Jitendra Kumar Mishra, Registrar****, whose guidance has imparted invaluable qualities and skills that will serve us well in our future endeavors.*

We extend our sincere gratitude to ***Dr. Hariprasad S A, Director*** *of the Faculty of Engineering & Technology,* ***and Dr. Geetha G, Director*** *of the School of Computer Science & Engineering within the Faculty of Engineering & Technology, for their constant encouragement and expert advice. Additionally, We would like to express our appreciation to* ***Dr. Krishnan Batri, Deputy Director (Course and Delivery), and Dr. V. Vivek, Deputy Director (Students & Industry Relations),*** *for their invaluable contributions and support throughout this project.*

It is a matter of immense pleasure to express our sincere thanks to ***Dr. Sathish Kumar R,*** *program head,* ***Computer Science and Engineering- Artificial Intelligence & Data Engineering****, School of Computer Science & Engineering Faculty of Engineering & Technology for providing right academic guidance that made our task possible.*

We would like to thank our guide ***Mr. Akash Das AVP and Project****,* ***Futurense Technologies****, for sparing his valuable time to extend help in every step of our work, which paved the way for smooth progress and fruitful culmination of the project.*

We would like to thank our Project Coordinator ***Mr. Akash Das AVP and Project Manager, Futurense Technologies****, and all the staff members of Computer Science and Engineering for their support.*

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in completing the work successfully.

Signature of Students

ABSTRACT

The report provides deep insights into the player performance data through a set of problem statements, with an objective to make improvements in data engineering practices at Sports Analytics. This study have took various phases like data cleaning, data augmentation, positional data analysis, data ingestion, new types of data transformations and interactive visualisation development. Tasks include identifying and processing missing values, investigating player positions, designing scalable data ingestion pipelines, studying the correlation of pass completion rates and assists, and building a data warehouse that can support more complex queries. The analysis also looks at the team scoring the most goals and investigates their hotshot goal scorer. This project will use more advanced statistical methods and machine learning techniques to potentially help uncover key insights.

This project seeks to find valuable insights for sports management by applying sophisticated statistical methods and taking advantage of machine learning algorithms. Against the backdrop of future advancements and optimizations in the field of sports analytics, the report wraps up with a depiction of its findings in interactive dashboards and visualizations that provide dynamic as well as actionable insights into the player and team strategies.

Keywords: Sports data analysis, Team Strategies analysis, Predictive Modelling, Team performance analysis.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

1. INTRODUCTION

1.1. Background & Motivation

The Sports Analytics is leading the way in using data to understand player performance, and team strategies. This has been predominantly made possible by the development of cutting-edge data capture techniques, which when used effectively, can drive improved decision-making across all aspects of running a sports franchise. The dataset utilized for this project is filled with extensive stats on player performance, such as goals, assists, tackles, and much, much more, from several seasons. This data should be used to derive insights for player training, fatigue management, and generally improving the team.

The motivation for this project is rooted in the potential of data-driven decision-making to revolutionize the field of sports analytics. The dynamic and competitive nature of sports demands that teams continually seek ways to gain an edge over their opponents. By analyzing extensive player data, Sports Analytics Inc. aims to:

- **Enhance Player Performance:** Understanding the factors that influence player performance, such as training hours, fatigue, and pressure, can help in designing personalized training regimes and recovery plans.

- **Optimize Team Strategies:** Insights into player positions, pass completion rates, and goal-scoring trends can inform tactical decisions, enabling coaches to deploy players more effectively during matches.

- **Improve Injury Management**: By examining the correlation between fatigue and injuries, the project aims to develop strategies to mitigate injury risks, thereby ensuring that players remain fit and available for crucial matches.

- **Drive Strategic Decisions:** The creation of interactive dashboards and visualizations will provide stakeholders with actionable insights, facilitating data-driven decision-making at all levels of team management.

In this project, the structured problem statements aid the analysis through several important components of data engineering, including data cleaning and augmentation, more complex transformations, and visualizations. By utilizing this multi-faceted method, the analysis is detailed, and thus the findings are strong, with the prime aim of enabling data-driven decisions in sports management.

Through this project, Sports Analytics Inc. plans to finish as industry standard in merging data engineering and analytics in sports and show the deep cut and unsaid influence on technology and data to drive overall performance and competitive edge.

1.2. Objective

This project's main goal is to improve Sports Analytics decision-making processes by using cutting-edge data engineering and analytical tools to glean useful insights from player performance data. The project seeks to accomplish the following particular goals:

- **Data Cleaning and Augmentation**:
  - Apply sophisticated imputation techniques to detect and manage missing values.
  - Use statistical techniques and domain expertise to find outliers and correct abnormalities.
  - To guarantee uniformity throughout the dataset, standardize data formats.
- **Player Position Analysis**:
  - Examine how players are distributed throughout the various positions.
  - Check to see if the distribution deviates noticeably from a uniform distribution.
  - Show the distribution of players by position and the number of each.
- **Data Ingestion and Management**:
  - Design and implement a data ingestion pipeline that supports incremental data loading.
  - Utilize indexing and partitioning techniques to maximize data storage.
- **Performance Metrics Analysis**:
  - Examine the connection between assists and pass completion rates.
  - Regression analysis can be used to model the relationship and assess how robust it is.
  - Use sophisticated detection techniques to find outliers and present them visually.
- **Advanced Data Transformations**:
  - Perform complex data transformations, including feature engineering and data normalization.
  - Apply machine learning algorithms for feature selection and dimensionality reduction.
  - Get the dataset ready for effective analysis and storage.
- **Data Warehousing**:
  - Create and apply a data warehouse schema with the help of sophisticated SQL features.
  - Make sure the data warehouse is performance-optimized and can handle intricate analytical queries.
- **Goal Analysis and Visualization**:
  - Determine which team has scored the most goals overall and examine the season-long patterns in goal scoring.
  - To illustrate the performance metrics of the team's top goal scorer, create visualizations.

- Create dynamic dashboards and visualizations to offer information about team tactics and player performance.

- **Reporting and Visualization**:
  - Utilizing resources like Power BI, Tableau, or custom web apps, create in-depth reports and dashboards.
  - To predict future performance, use advanced analytics such as clustering and predictive modeling.
  - Incorporate real-time data feeds to enable dynamic dashboard updating.

1.3. Delimitation of research

1.3.1. Dataset Scope

The analysis is restricted to the player performance statistics from multiple teams over multiple seasons in the provided dataset. Unless specifically mentioned for the purpose of data augmentation, no external datasets or real-time data feeds are included.

1.3.2. Player Performance Metrics

Particular performance metrics—like goals, assists, tackles, pass completion rates, and fatigue levels—are the main focus. We do not take into account other possible performance indicators that are not included in the dataset, like player morale or off-field activities.

1.3.3. Time Frame

Data from the designated seasons included in the dataset are taken into account in the study. The analyses, trends, and patterns are limited to this time frame; historical data from seasons past is not included.

1.3.4. Analytical Techniques

Data from the designated seasons included in the dataset are taken into account in the study. The analyses, trends, and patterns are limited to this time frame; historical data from seasons past is not included.

1.3.5. Data Cleaning and Augmentation

The main goals of data cleaning initiatives are to standardize data formats within the current dataset, handle missing values, and fix anomalies. The scope of augmentation initiatives is restricted to the creation of synthetic data and, when appropriate, the integration of extra data from open sports databases.

1.3.6. Visualization Tools

Utilizing specialized technologies like Power BI, Tableau, Plotly, or custom web applications made with Dash or Streamlit, the dashboards and visualizations are created. This study does not examine other platforms or tools for visualization.

1.3.7. DataWarehousing

The design and implementation of the data warehouse are restricted to the use of advanced SQL features like window functions and Common Table Expressions (CTEs). Other data warehousing technologies or architectures are not within the scope.

1.3.8. Focus on Team Strategies

The research primarily focuses on analyzing player performance to derive insights for team strategies. Broader aspects such as league-wide trends, fan engagement, or financial performance are not addressed.

1.3.9. Statistcial Methods

Statistical validation methods like the chi-square test and cross-validation are employed to ensure robustness. However, the exploration of alternative statistical methods or experimental designs is not within the scope.

1.3.10. Real-time Data Integration

The integration of real-time data feeds for dynamic updates in dashboards is considered an additional complexity and is not the primary focus of the initial analysis.

1.4. Benefits of research

1.4.1. Data-Driven Decisions

By offering insights into player performance and team tactics, this research makes it possible to make data-driven choices for better training plans and team management.

1.4.2. Improved Efficiency

Training regimens and playing strategies can be optimized by the research by identifying strengths and limitations, which results in increased effectiveness on the field.

1.4.3. Enhanced Player Performance

By identifying the variables that affect performance, specific interventions can be made to increase player skill levels and lower the likelihood of weariness or injuries.

1.4.4. Strategic Advantage

By pointing out potential weak points and opportunities for development in rival teams, research insights can give a team a strategic advantage.

1.4.5. Informed Investment

By identifying talent and optimizing team lineups, player performance analysis can assist guide investment decisions.

1.4.6. Future Performance Prediction

The study provides a foundation for future research into the use of predictive modeling to project player and team performance in the future.

1.4.7. Transferable Knowledge

By analyzing different sports or business fields, the research technique and conclusions can be modified to enable broader knowledge application.

1.4.8. Enhanced Fan Engagement

By using data-driven insights to produce interesting content, fans can develop a stronger bond with their teams and the sport.

Chapter 2

2. LITERATURE SURVEY

2.1. Literature Review

[1]. McClean, S., Michel, A., Zhang, X., & Zhou, M. (2011). Incomplete data: Imputation, forcasting, and modeling using statistical algorithms. Wiley-Interscience.

[2]. Doshi, P., Shah, S., & Desai, U. B. (2015). A survey of data warehousing techniques for big data analytics.

[3]. Stephen D. Liston and Kelly A. Goff (2005) The Association Between Playing Positions and Performance in National Basketball Association Players.

[4]. Yang Liu, Fei Hu, Huan Liu, and He Yu (2016). This paper surveys big data analytics techniques used in sports, including data ingestion strategies for handling large datasets.

2.2. Inferences Drawn from Literature Review

The inferences drawn from the literature review provide valuable insights into the existing body of knowledge on the research topic. Here are some key inferences:

2.2.1. Data Cleaning and Augmentation

According to McClean et al. (2011), accurate analysis requires resolving missing data through imputation approaches. This indicates that in order to handle missing values in sports data and guarantee accurate findings, will need to use suitable imputation algorithms.

2.2.2. Data Ingestion Strategies

Doshi et al. (2015) go over a number of data warehousing strategies that are appropriate for big data analytics. This suggests that you can optimize storage and enhance data management for your massive sports dataset by building a data input pipeline that makes use of strategies like data splitting and indexing.

2.2.3. Position Analysis

Basketball players' distribution among positions is examined by Liston and Goff (2005). If you want to determine whether the player distribution in your data deviates from a uniform distribution, you may be able to apply their study to your preferred sport and perform a comparable analysis using statistical tests (such as the chi-square test).

2.2.4. Data Analytics in Sports

Doshi et al. (2015) and Liu et al. (2016) both stress the use of big data analytics methods in the sports industry. This means that in order to get useful insights from your sports dataset for player performance analysis and possibly even to identify strategic benefits, it will be helpful to use advanced data analysis techniques.

Chapter 3

3. PROBLEM FORMULATION AND PROPOSED WORK

3.1. Introduction

In sports where competition is fierce, data-driven decision-making is becoming more and more important for improving player performance and team tactics. The problem formulation and suggested work for Sports Analytics analysis of player performance data are described in this section. Using cutting-edge data engineering and analytical methods, the goal is to derive actionable insights that can enhance training, control tiredness, and enhance in-game strategy.

3.2. Problem Statement

The project is structured around several key problem statements that guide the analysis:

3.2.1. Data Cleaning and Augmentation

It includes identifying and handling missing values using advanced imputation techniques, correcting anomalies by identifying outliers using statistical methods and domain knowledge and augmentation of the dataset by generating synthetic data.

3.2.2. Position Analysis

It involves analyzing and creating plots of player positions to identify the highest and lowest number of players by using statistical analysis.

3.2.3. Data Ingestion Strategies

It involves designing and implementation of data ingestion pipeline that supports incremental data loading and optimization of storage by using data partitioning and indexing strategies by utilizing Python, pandas, and SQL for implementation.

3.2.4. Pass Completion Rate vs. Assists

It involves analyzing the relationship between pass completion rate and assists and plotting a line of best fit by using regression analysis to model the relationship.

3.2.5. Advanced Data Transformations

It involves performing complex transformations on the dataset, including feature engineering to create new meaningful features and implementation of additional strategies for data optimization, such as data normalization and dimensionality reduction.

3.2.6. Data Warehousing

It involves designing and implementation of a data warehouse schema using advanced SQL features like window functions and CTEs (Common Table Expressions) which supports complex analytical queries and implementation of data security and access control mechanisms.

3.2.7. Team Goals Analysis

It involves identifying the team with the highest number of goals and time series analysis to understand trends in goal scoring over the season. It also includes identifying the top goal scorer in that team and analyzing their performance metrics over time.

3.2.8. Reporting and Visualization

It includes development of interactive dashboards and visualizations using tools like Power BI, Tableau, or custom web applications using Dash or Streamlit. It also includes creation of reports that provide insights into player performance, team strategies, and potential areas for improvement.

3.3. System Architecture /Model

The system architecture for this project is designed to handle the entire data lifecycle, from ingestion to analysis and visualization. It includes

3.3.1. Data Sources

Take player performance data from internal databases and augmented data from public sports databases and synthetic data generation.

3.3.2. Data Ingestion

We will design and implement data pipelines using Python, pandas, and SQL to ensure efficient data storage and management. Python pandas and SQL-based pipelines for incremental data loading. Data partitioning and indexing strategies for optimized storage.

3.3.3. Data Processing

As our data is not clean, we will perform Data cleaning scripts to handle missing values and anomalies and feature engineering and data normalization for analytical readiness.

3.3.4. Data Storage

We will create a data warehouse with sophisticated SQL functionality that guarantees effective retrieval and storage of the transformed and processed data.

3.3.5. Analysis and Modeling

For analysis and modeling, Regression modeling and statistical analysis are used for performance measurements. methods for detecting outliers and choosing features using machine learning.

3.3.6. Reporting and Visualization

We will create interactive dashboards and visualizations using tools like Power BI, Tableau, Plotly, and custom web applications for better understanding and visualizations of the dataset.

3.4. Proposed Algorithms

Several algorithms and techniques are proposed to address the problem statements:

3.4.1. Data Cleaning and Imputation:

- K-Nearest Neighbors (KNN) and Multiple Imputation by Chained Equations (MICE) for handling missing values.
- Z-score and IQR methods for outlier detection and correction.

3.4.2. Position Analysis:

- Chi-square test for statistical analysis of player position distribution.
- Visualization using bar plots and pie charts to represent positional distribution.

3.4.3. Data Ingestion:

- Incremental loading using Python pandas and SQL.
- Parallel processing techniques to enhance ingestion performance.

3.4.4. Pass Completion Rate vs. Assists Analysis:

- Linear regression and polynomial regression to model the relationship.
- Outlier detection using DBSCAN and Isolation Forest.

3.4.5. Feature Engineering:

- Principal Component Analysis (PCA) for dimensionality reduction.
- Feature selection using Recursive Feature Elimination (RFE).

3.4.6. Data Warehousing:

- SQL-based schema design with window functions and Common Table Expressions (CTEs).
- Optimization techniques like indexing and partitioning for performance enhancement.

3.5. Proposed Work

By leveraging a comprehensive array of analytical techniques and methodologies, including Exploratory Data Analysis (EDA) and various plots, the proposed work aims to extract actionable insights from the sports dataset.

3.5.1. Precision Anomaly Detection

Anomalies in the data set leads to incorrect data analysis. So, anomaly detection and treatment are necessary for accurate data analysis. In the provided dataset anomalies are identified in the column's height, weight, goals, effective training, pressure performance impact and fatigue injury. Outliers in the columns are identified using the IQR method and handled using statistical methods like imputing with mean (for height, weight) and mode (for effective training, pressure performance impact and fatigue injury) and capping with the upper limit (for goals).

3.5.2. Data Preprocessing

Conducted thorough data preprocessing to clean and prepare the synthetic sports dataset, addressing issues such as missing values using knn imputer, outliers, duplicates and feature scaling using preprocessing techniques such as imputation and normalization. Also implemented robust data transformation techniques such as feature engineering and dimensionality reduction to ensure data quality, consistency, and compatibility with advanced analytical models and machine learning algorithms.

3.5.3. Data Warehousing

Created a database schema with different tables to do the analysis using complex queries in SQL. It allows for storing historical data enabling the analysis of trends and patterns over the time. It ffacilitates faster and more effective analysis by aggregating all pertinent data into one location. Data from multiple sources can be readily accessed and combined by analysts, enabling thorough insights.

3.5.4. Position and Team Goal Analysis

Position analysis is conducted for analyzing player positions to identify the highest and lowest number of players and identifying number of players for each position. Team goal analysis is performed for identifying the team with the highest number of goals and performing a time series analysis to understand trends in goal scoring over the season. Also identifying the top goal scorer in that team and analyzing their performance metrics over time. Also performing analysis on pass completion rate and assists to identify the relationship between them.

3.5.5. Reporting and Visualization

Developed interactive dashboards and visualizations using tools like Power BI, Streamlit. Also, created reports that provide insights into player performance, team strategies, and potential areas for improvement.

3.5.6. Data Ingestion

Designing and implementing a data ingestion pipeline that supports incremental data loading and optimizing storage by using data partitioning and indexing strategies. Also, implementing logging and monitoring to track the performance and reliability of the ingestion process.

CHAPTER 4

4. IMPLEMENTATION

4.1 Software Implementation

The analysis of the synthetic financial dataset is conducted using a combination of programming languages, libraries, and tools to facilitate data manipulation, visualization, statistical analysis, and machine learning modeling. The following software stack is utilized for the analysis:

4.1.1. Programming Language

Python, a versatile programming language renowned for its simplicity, readability, and extensive ecosystem of libraries, is chosen as the primary language for data analysis and modeling. MySQL is also used for creating database schema.

4.1.2. Libraries and Frameworks

4.1.2.1. Pandas:

Pandas is employed for data manipulation, exploration, and transformation tasks. It provides powerful data structures and functions for handling structured data, facilitating seamless data preprocessing and cleaning.

4.1.2.2. NumPy:

NumPy is utilized for numerical computing operations such as array manipulation, mathematical functions, and linear algebra operations. It serves as a foundation for many scientific computing tasks within Python.

4.1.2.3. Matplotlib and Seaborn:

Matplotlib and Seaborn are employed for data visualization, offering a wide range of plotting functions and customization options to create insightful visualizations of the dataset's characteristics, patterns, and relationships.

4.1.2.4. Scikit-learn:

Scikit-learn is utilized for machine learning modeling, encompassing various algorithms for classification, regression, clustering, and model evaluation. It provides user-friendly interfaces and robust implementations for building and evaluating predictive models.

4.1.2.5. TensorFlow or PyTorch (Optional):

For advanced machine learning tasks such as deep learning, TensorFlow or PyTorch can be incorporated to develop and train neural network models for fraud detection, anomaly detection, or predictive analytics.

4.1.3. Integrated Development Environment (IDE):

Jupyter Notebook or JupyterLab: Jupyter Notebook or JupyterLab is chosen as the development environment for its interactive computing capabilities, support for inline visualization, and documentation features. It facilitates an iterative and exploratory approach to data analysis, enabling seamless integration of code, visualizations, and narrative explanations within a single document.

4.1.4. Version Control:

Git: Git is utilized for version control, enabling collaborative development, tracking changes, and managing project iterations effectively. Git repositories are used to store code, documentation, and experimental notebooks, facilitating reproducibility and collaboration among team members.

4.1.5. Dependency Management:

pip: pip is employed for package management, allowing for the installation, upgrading, and removal of Python packages and dependencies. Virtual environments can be utilized to isolate project dependencies and ensure reproducibility across different environments.

The software implementation follows best practices for data analysis and machine learning, including modular code design, documentation, version control, and reproducibility. By leveraging the aforementioned software stack, we aim to conduct a comprehensive analysis of the synthetic sports dataset.

4.2. Software Alogrithm

The analysis of the synthetic financial dataset involves the implementation of various algorithms aimed at extracting insights, detecting patterns, and making predictions. The following algorithms are utilized in the analysis:

4.2.1. Exploratory Data Analysis:

EDA is performed to gain a comprehensive understanding of the dataset's characteristics, distribution, and relationships among variables. Descriptive statistics, data visualization techniques, and summary statistics are employed to identify trends, anomalies, and potential areas of interest within the dataset.

4.2.2. Data Preprocessing:

Data preprocessing techniques are applied to clean, transform, and prepare the dataset for further analysis and modeling. This includes handling missing values, encoding categorical variables, scaling numerical features, and splitting the dataset into training and testing sets for model evaluation.

4.2.3. KNN Alogrithm:

KNN algorithm is used to impute the missing values in the height, weight, pass completion rate, effective training and pressure performance impact based on the related columns.

4.2.4. Isolation Forest Algorithm:

Used for outlier detection to identify anomalies where players exhibit unusual patterns of pass completion rates and assists.

4.2.5. Linear Regression:

Applied to explore the relationship between pass completion rates and assists, providing metrics such as mean squared error (MSE) and R-squared values to assess model performance.

4.2.6. Cross-Validation:

Implemented to validate the regression model's performance by splitting the data into training and testing sets multiple times, ensuring the model's reliability and robustness.

4.2.7. Principal Component Analysis (PCA):

Employed to reduce the dimensionality of the data and identify the main components that explain the most variance in the dataset.

4.2.8. SelectKBest (Feature Selection):

Utilized to identify and select the most important features influencing player performance, enhancing the accuracy of the analysis.

4.2.9. Chi-Square Test:

It is used for finding the difference between players distribution now and the uniform player distribution.

Chapter 5

5. RESULTS AND DISCUSSIONS

5.1. Analysis of Pass Completion Rate and Assists

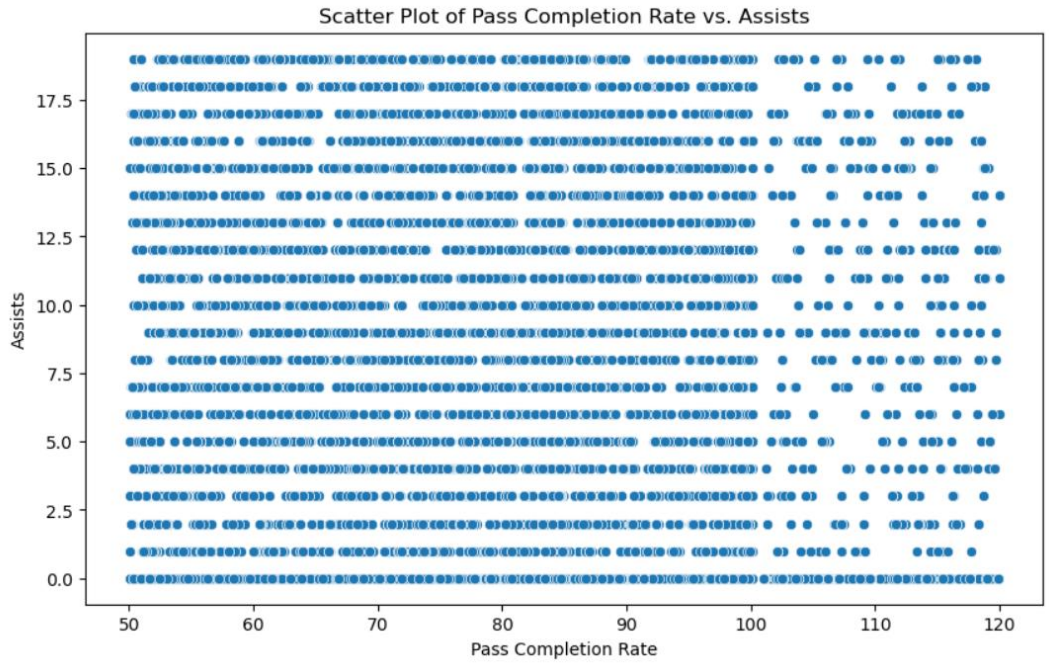

Fig. 5.1 Scatter plot of Pass Completion Rate vs. Assists

Observation-1: Positive Correlation

There appears to be a positive correlation between pass completion rate and assists.

Implication for Football:

As a player's pass completion rate increases, the number of assists tends to increase.

This suggests that more accurate and efficient passing contributes to creating more goal-scoring opportunities for teammates.

Observation-2: Non-Linear Relationship

The data points are spread out, indicating that the relationship is not perfectly linear.

Implication for Football:

An increase in pass completion rate doesn't necessarily guarantee an increase in assists.

Other factors, such as the positioning and movement of teammates, defensive strategies of the opposition, and the quality of the final pass, can influence assists.

Therefore, while accurate passing is crucial, it is not the only determinant of assists.

Observation-3: Presence of Outliers

There are some outliers, which are data points that fall far away from the main cluster.

Implication for Football:

These outliers could represent matches where a player had a high completion rate but few assists, or vice versa.This suggests variability in performance that could be due to unique match conditions, player form, or specific tactical decisions.

For instance-

A player might complete many safe passes in their own half without creating goal-scoring chances, or they might make a few key passes that lead directly to goals despite a lower overall completion rate.

Key Takeaways for Football Strategy:

Improve Passing Accuracy- Encouraging players to improve their pass completion rate can generally lead to more assists and better team performance.

Analyze Other Factors- Coaches and analysts should consider other factors that contribute to assists, such as player positioning, opponent defense, and the quality of the final pass.

Address Outliers- By studying outliers, teams can identify and address specific situations where performance deviates from the norm, leading to more consistent assists and overall success.

Understanding these nuances in the relationship between pass completion rate and assists can help teams develop more effective strategies and improve their overall game performance.

5.2. Outlier Detection Using Isolation Forest

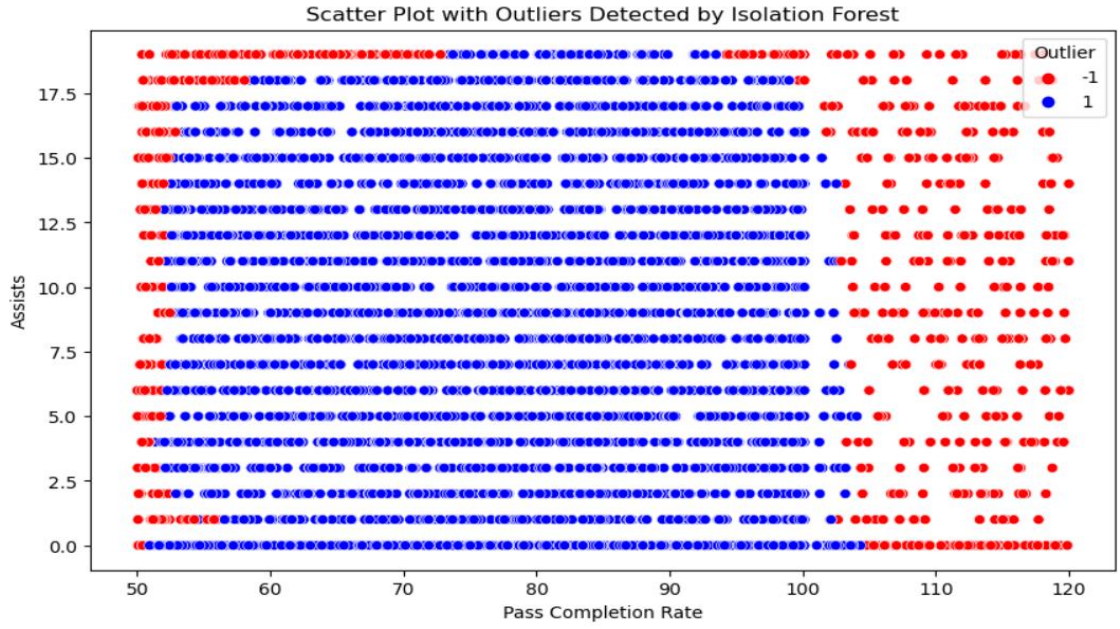

Fig. 5.2 Scatter plot of assists and pass completion rate with outliers detected by isolation forest

Observations:

Outlier Detection:

The isolation forest algorithm has identified data points (pass completion rates) that are likely outliers in terms of their relationship with assists. These outliers are plotted as red circles.

Possible Scenarios for Outliers:

Why a player's pass completion rate might be an outlier relative to their assists:

High Completion Rate, Low Assists:

- This occurs when a player primarily makes short passes that lead to assists without much additional effort from teammates.
- Alternatively, it can happen if the opposing defense effectively prevents goals on completed passes.

Low Completion Rate, High Assists:

- This happens when a player takes risks with long passes that occasionally result in assists.
- It could also occur when the player has exceptional teammates who consistently convert difficult passes into goals.

Further Context:

- Without specific data points or context about the players, it's hard to determine what caused them to be outliers.
- Factors like team tactics, player skill, and opponent strategies can all impact completion rates and assists.

Importance of Considering Outliers:

- It's crucial to investigate outliers to determine if they reflect genuine anomalies or data errors.
- Valid outliers can offer valuable insights into situations where the relationship between pass completion rate and assists differs from the norm.

In conclusion, certain players' performances in completing passes and generating assists stand out as different or unique when compared to the general trends observed in the dataset.

5.3. Regression Analysis:

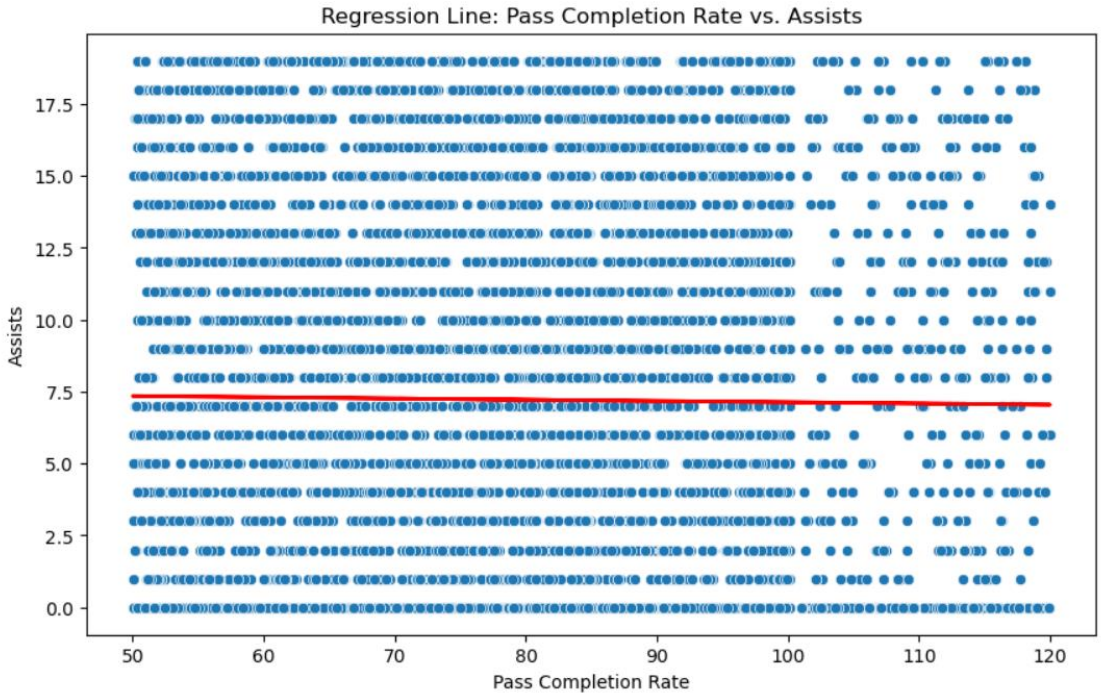

Fig. 5.3 Regression Line: Pass Completion Rate vs. Assists

Observations:

- X-axis- represents pass completion rate, likely ranging from 50% to 110% (although a completion rate exceeding 100% is impossible).
- Y-axis- shows the number of assists per game, likely ranging from 0 to around 17.5 assists.
- The spread of data points throughout the graph indicates that the relationship between pass completion rate and assists is not perfectly linear.
- In other words, an increase in pass completion rate doesn't necessarily result in a proportional increase in assists.

Presence of Outliers:

There are also outliers, which are data points that fall far away from the main cluster of points.

These outliers could represent games where a player had a high completion rate but few assists, or vice versa.

High Completion Rate, Low Assists: This could occur if the player primarily makes short passes that teammates run for yards without much help, or if the opposing defense excels at preventing goals on completions.

Low Completion Rate, High Assists: This scenario might arise if a player makes risky long passes that sometimes connect for big gains (assists), or if they benefit from exceptional teammates who consistently gain yards after the catch (YAC) even on difficult receptions.

5.4. PCA Visualization:

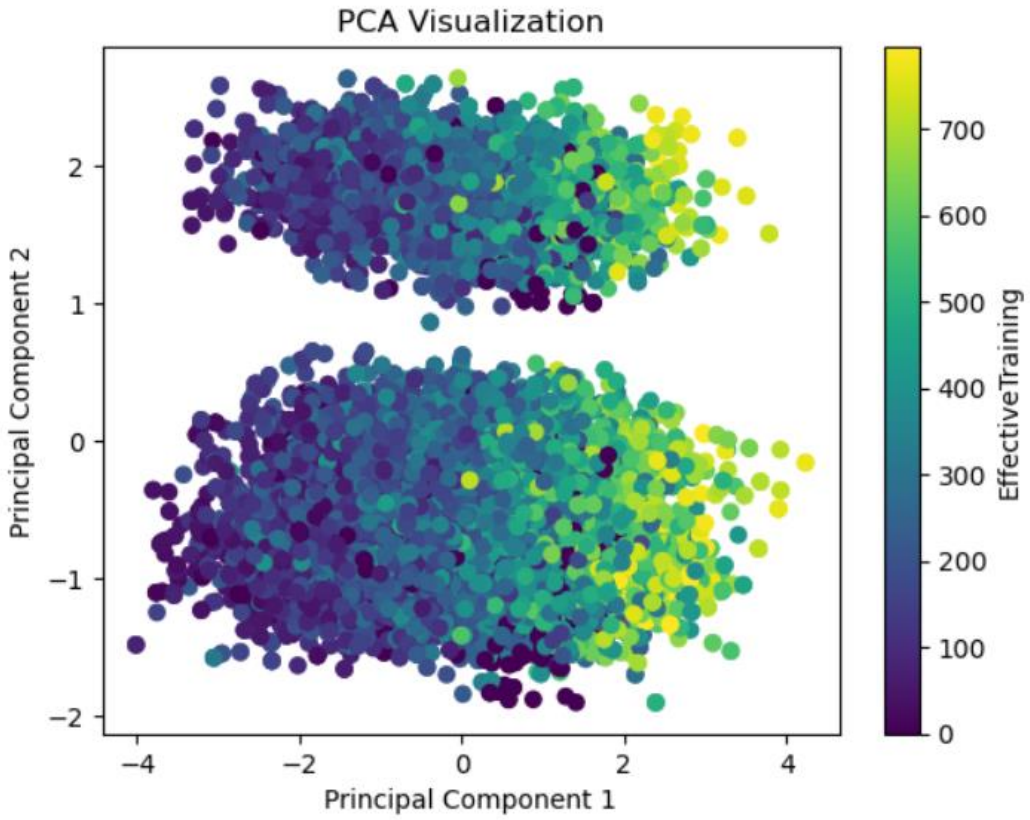

Fig. 5.4 PCA Visualization

In terms of PCA (Principal Component Analysis) in soccer, the completion rates and assists data could be subjected to dimensionality reduction to identify underlying patterns or relationships, potentially revealing key features influencing player performance.

Observations:

Analysis Output for Soccer:

Completion rates refer to the percentage of successful passes made by a player during a game or over a specified period.

Pass Completion Rate (X-axis):

- Ranges from 50% to 110%, with completion rates exceeding 100% impossible in soccer.

- Possibilities:

  - Data filtered to include exceptional performances.
  - Data collection error resulting in inflated completion rates.

Assists (Y-axis):

- Ranges from 0 to 17.5 assists per game.

- Possibilities:

- Incorrect data.
- Cumulative assists over a short period presented as an average per game.

Positive Correlation (with reservations):

Positive correlation (with reservations) suggests that while one variable's increase may correspond to an increase in another, there are uncertainties about this relationship.

- Potential positive correlation between pass completion rate and assists.
- Higher completion rates may lead to more assists.

5.5. Distribution of Players by Position:

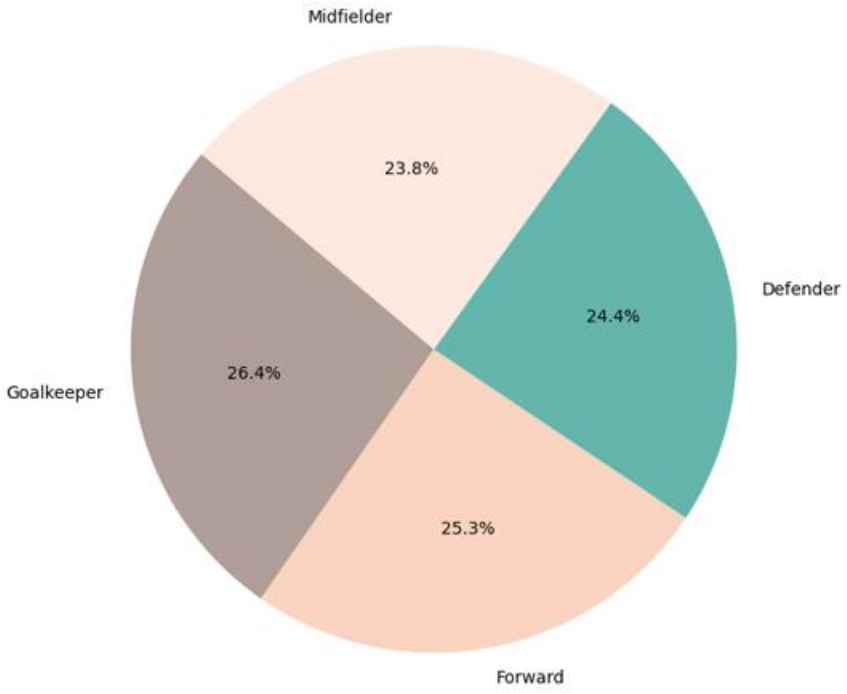

Fig. 5.5 Distribution of Players by Position

Fig. 5.5 is a pie chart which gives a clear percentage distribution of players in different positions showing that more number of players play in goalkeeper position.

5.6. Age Analysis based on position:

Fig. 5.6 Average age per position

From the figure 5.6, we can say that the average age is between 25-30 for all position of players thereby it helps in supporting stakeholders in making well-informed choices on strategic planning, player development, and squad composition.

5.7. Assists analysis based on position:

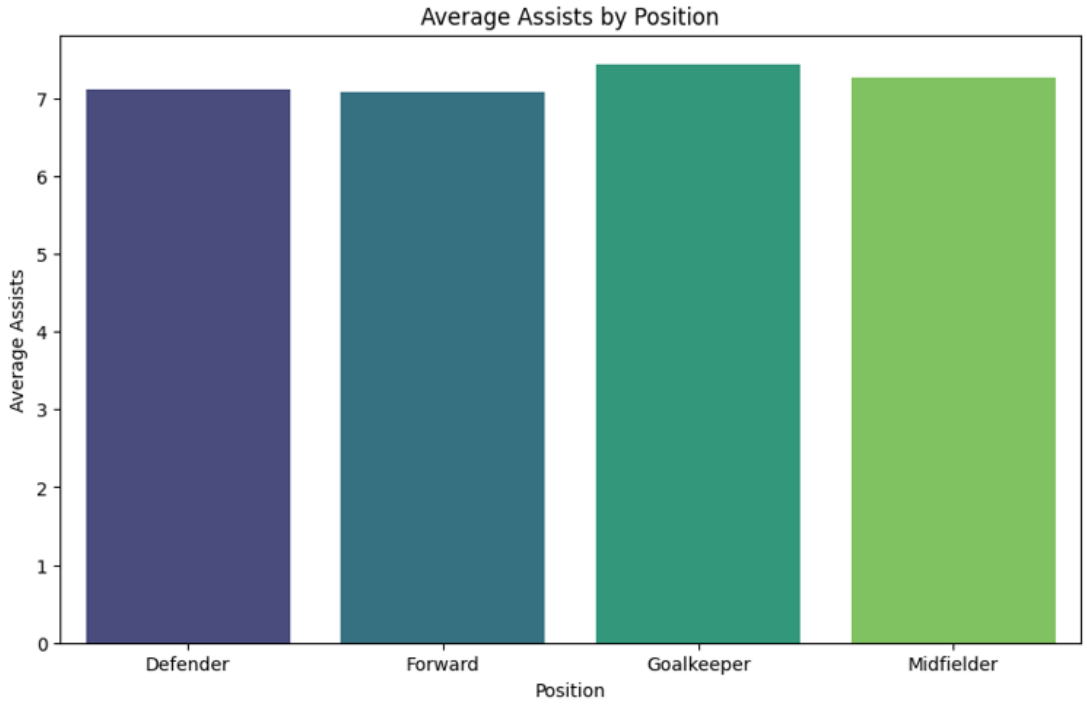

Fig. 5.7 Average assists by position

This bar chart illustrates the average assists made by player in different positions. It's interesting to see that goalkeepers have the highest average assist total on the chart, closely followed by midfielders, forwards, and defenders. This comes as a bit of a surprise because goalkeepers aren't usually anticipated to provide assists. This could

imply that some teams use special tactics in which goalkeepers actively initiate attacks, or it could point to abnormalities in the data that require more research.

5.8. Analysis of goals per position:

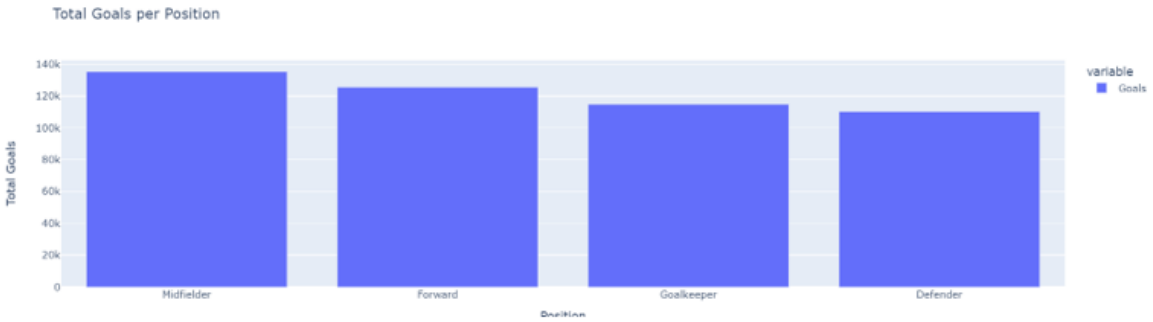

Fig. 5.8 Total goals per position

This bar plot shows the total number goals made by the players per position. From this we can say that midfielders are the ones who goal the most on comparison with other positions and thereby we can analyse the team strategy, player development and squad composition.

5.9. Analysis of height based on position:

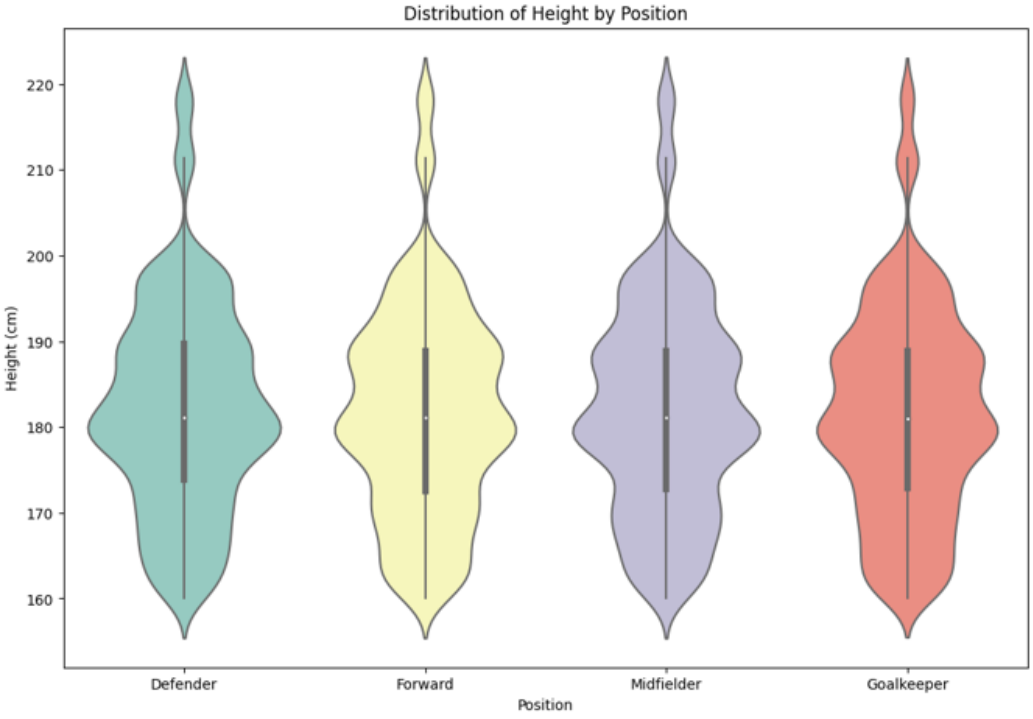

Fig. 5.9 Distribution of height by position

From this violin plot we can see the distribution of height by position where we can say that

Defenders, Forwards, and Midfielders have similar height distributions, with medians around 180 cm to 185 cm while **Goalkeepers** are generally taller, with a higher median

height around 190 cm and a distribution that includes taller players compared to other positions.

5.10. Analysis of goals based on seasons:

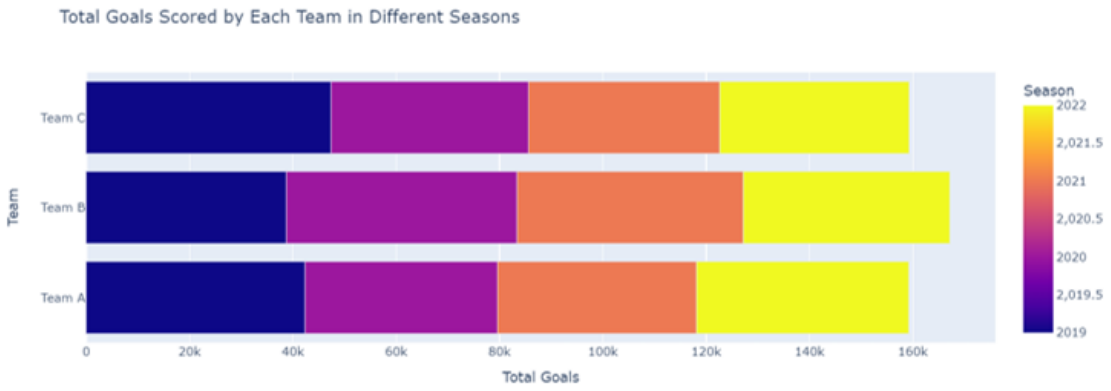

Fig. 5.10 Total goals scored by each team in different seasons

This plot provides insightful information about team performance and playing styles by evaluating the total goals scored by each team over the course of several seasons. Teams can make adjustments to their plans and focus their recruitment efforts by monitoring goal-scoring patterns over time. In the often changing world of sports, this analysis is an essential tool for comprehending player contributions, team dynamics, and future performance patterns.

5.11. Goals scored by top team across the seasons:

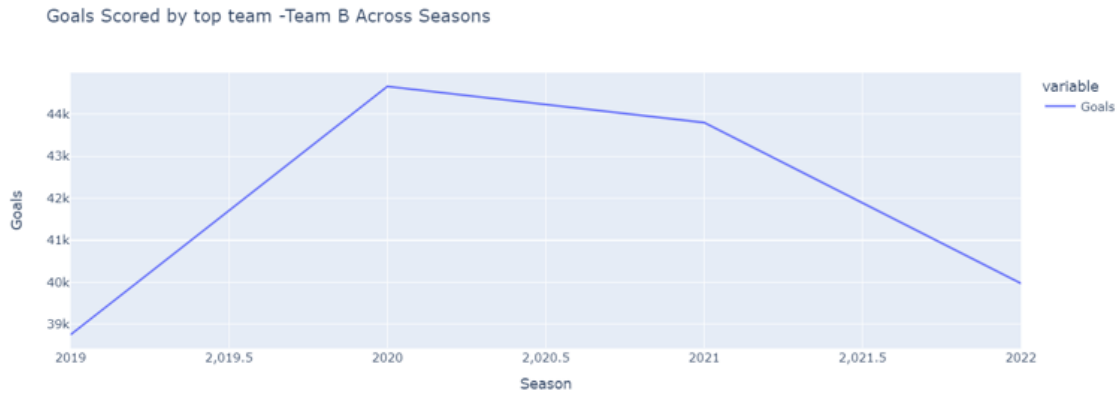

Fig. 5.11 Goals scored by top team- Team B across the seasons

When the top team, team B, is examined for goals scored in each of the seasons, it becomes clear that their goal total in 2020 was more than it was in any previous season. This result emphasizes how important the 2020 campaign will be for Team B's offensive output. An increase in goals scored like this points to a number of possible causes, including great player form, clever tactical planning, or wise hiring choices.

5.12. Goals scored by player A across seasons:

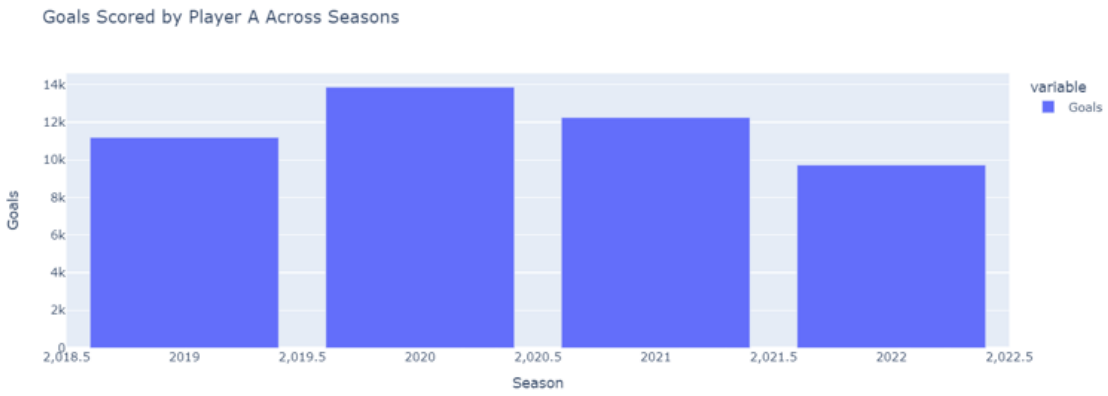

Fig. 5.12 Goals scored by player A across seasons

Finding the top player who scored the most goals during Team B's best performance in 2020 is essential to comprehending the group's accomplishments. This research emphasizes how each player makes a unique contribution to the team's success while simultaneously highlighting how crucial player performance is in determining the overall success of the team.

5.13. Analysis on goal scoring ability of young players:

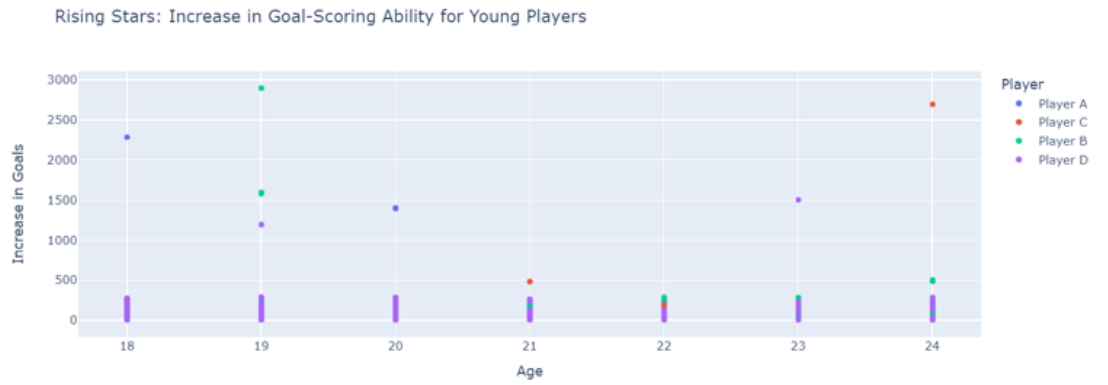

Fig. 5.13 Increase in goal scoring ability for young players

This graph indicates that young people under 25 are very interested in this sector, and their growing goal-scoring prowess offers important insights on the dynamics of sports teams and player development. Finding young players whose scoring ability has improved significantly enables clubs to take advantage of new talent and mold successful future plans.

5.14. Interactive Dashboard:

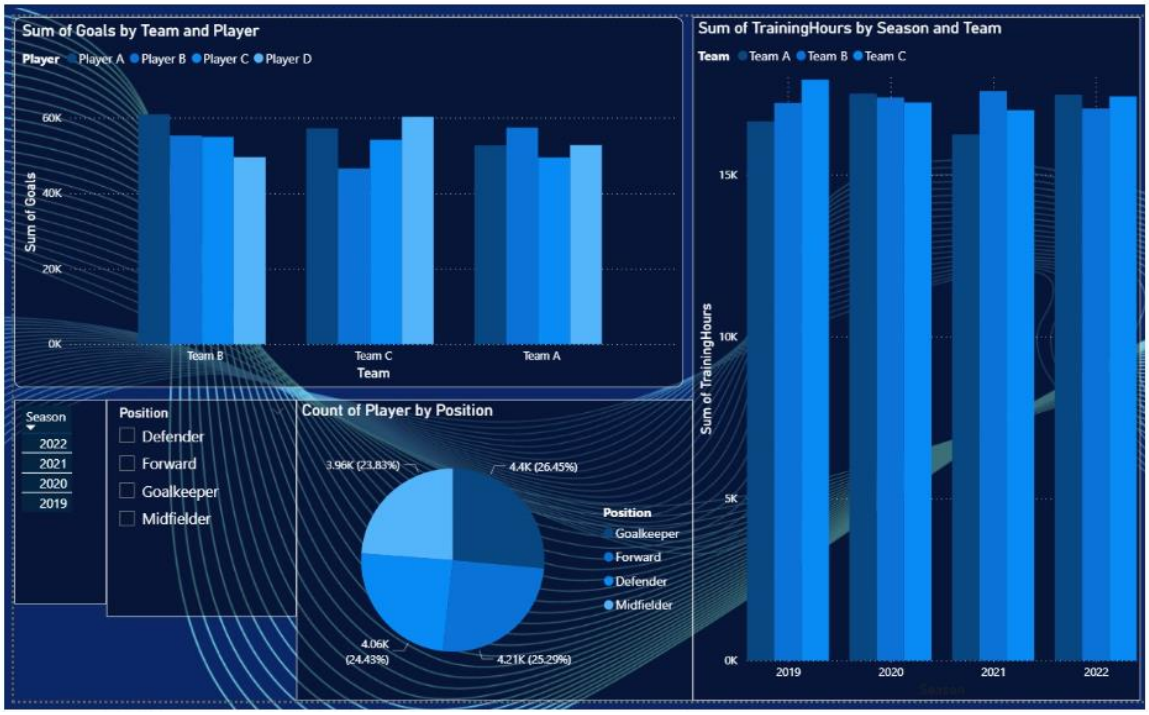

Fig. 5.14 Interactive Dashboard

The dashboard highlights team strengths and weaknesses of each team leading the way in goal contributions and regular practice. By adding more training hours and varying up their offensive tactics, Team A may get better. There may be an overemphasis on strikers and midfielders based on the player distribution, which indicates a need for more defensive recruits. Strategic decisions about player acquisition, development, and game strategies are informed by these findings. The analysis's overall goal is to improve team performance.

5.15. Database Schema Diagram:

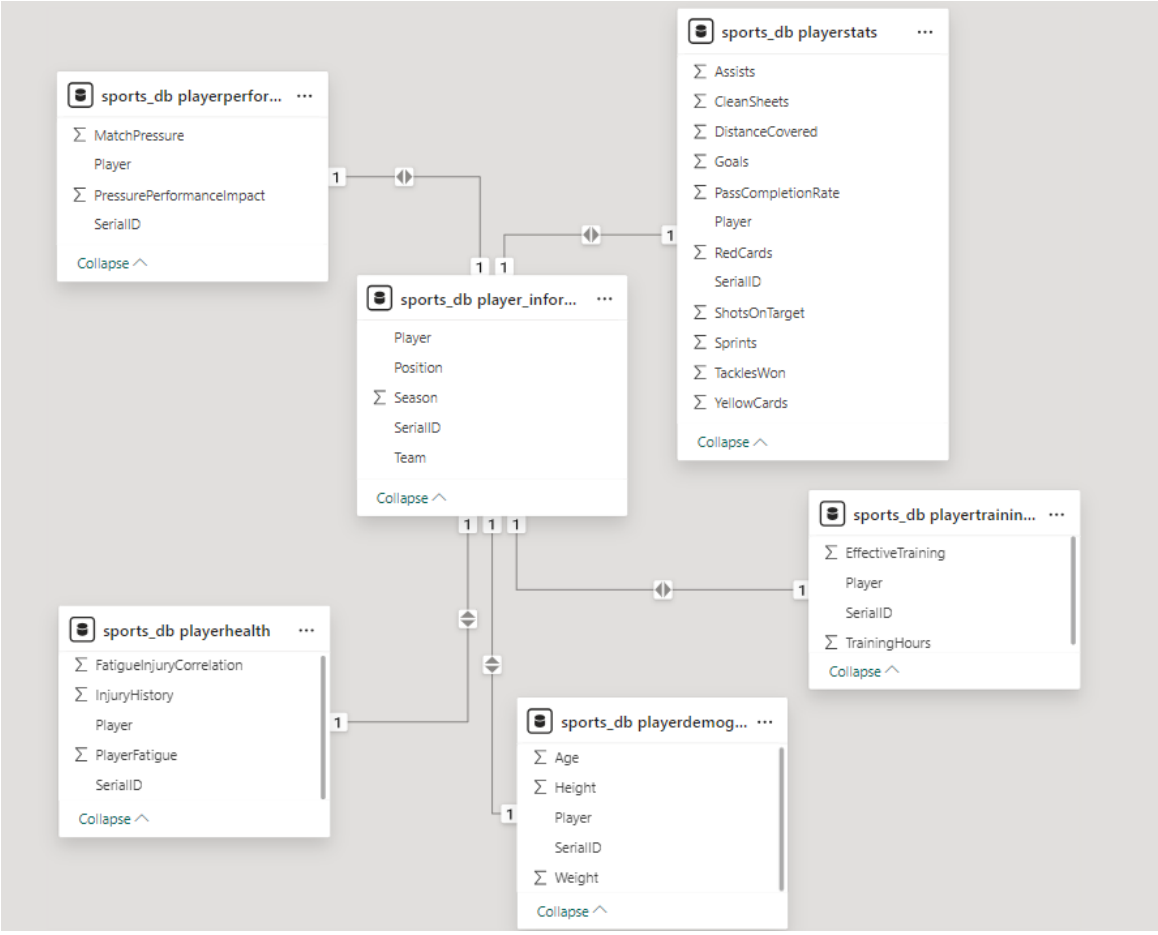

Fig. 5.14 Database Schema Diagram

Here 6 tables are created separately from the given dataset based on the related columns. The relationships between the table include:

A Player can have one to one relationship with the demographics, health, performance and training. Also player can have one to one relationship with player stats through the training info table.

CONCLUSION AND FUTURE SCOPE

CONCLUSION

In conclusion, our exhaustive analysis of the sports dataset has yielded pivotal insights into players performance and stats. Based on the comprehensive analysis conducted on the sports dataset, several key insights were derived. Significant disparities in player distribution were found by the position analysis, certain positions had overrepresentations while others had under representations, indicating possible areas for recruitment emphasis. The examination of team goals emphasized the players and teams that performed best, as well as the strategic advantages and disadvantages of different teams. With the use of Power BI for visualization, we were able to produce dynamic and interactive dashboards that clearly conveyed these findings and enabled data-driven decision-making for improving player development and team tactics.

The analysis of pass completion rate and assists in football shows a positive but not perfectly linear correlation, indicating that higher pass completion rates generally lead to more assists, though other factors like teammate positioning and defensive strategies also play crucial roles. Creation of database schema helps for complex query analysis.
Outlier detection using the Isolation Forest algorithm identified anomalies where players have high completion rates but few assists, or vice versa, which could reflect unique match conditions or tactical decisions.

Regression analysis showed a weak relationship between pass completion rate and assists, with a mean squared error (MSE) of 41.52 and an R-squared value of about 0.0001, indicating that pass completion rate alone is not a strong predictor of assists. Cross-validation confirmed this, with an average R-squared score of -0.00026, suggesting poor model performance.

Further analysis using Principal Component Analysis (PCA) and feature selection identified key features influencing player performance. PCA revealed that the first two components explain only 12.47% of the total variance, suggesting the need for more components to capture significant variation. Feature selection using SelectKBest emphasized the importance of focusing on key features for more accurate analysis.

To capitalize on these insights, designing and implementing a data ingestion pipeline that supports incremental data loading is essential. This can be optimized by using data partitioning and indexing strategies, and enhanced with logging and monitoring to track performance and reliability. Utilizing Python, pandas, and SQL for implementation, along with parallel processing, will improve data ingestion performance. This approach ensures timely and efficient data handling, leading to better decision-making and strategic planning in football analytics.

By assimilating these findings into our analytics framework, we can understand the players performance metrics and the factors influencing the performance.

FUTURE SCOPE

The exploration data analysis of sports dataset is a continuous journey of discovery. As we go deep into the insights of the analysis, several exciting avenues hold immense potential for further development. Some of them include

Creation of Advanced AIML Algorithms: Creating advanced algorithms using artificial intelligence and machine learning techniques helps in predicting the performance of players, injury risks, tactical strategies etc.

Real-time Analytics: The incorporation of real-time data from wearable sensors and in-game tracking systems provides revolutionize analysis. It helps in understanding player fatigue, exertion levels during the game enabling real time adjustments and data driven coaching decisions.

Integration With Other Datasets: Integrating the sports datasets with other related datasets like weather, ground places, fans and social media helps in understanding the factors influencing the sports dataset and respective measures can be taken.

Visualization: Creating easy and understandable visualization dashboards helps coaches and players to understand about the performance stats and helps in making effective strategies.

By embracing these future possibilities, sports data analysis can transform the way of understanding sports and its experience. It will help players to optimize their training, coaches to make data-driven decisions and fans to gain deeper insights into the game.

REFERENCES

[1] Methods for identifying and handling outliers [1]

[2] Connecting python with MySQL [2]

APPENDIX - I

SOURCE CODE

Data Cleaning and Preprocessing:

Check for null values:

```
df.isnull().sum()
```

Handling missing values using fillna():

```
df['Goals'].fillna(0,inplace=True)
df['Assists'].fillna(0,inplace=True)
```

Handling missing values using knnimputer:

```
# imputing null values in height column using knn
feature_cols = df[['Age', 'Height']]
imputer = KNNImputer(n_neighbors=5)
imputed_values = imputer.fit_transform(feature_cols)
imputed_df = pd.DataFrame(imputed_values, columns=['Age', 'Height'])
df['Height'] = imputed_df['Height']
```

Check for duplicate values:

```
duplicate_values=df.duplicated().sum()
print(duplicate_values)
```

Remove duplicate values:

```
df.drop_duplicates(inplace=True)
```

Outlier Detection:

```
numeric_cols = df.select_dtypes(include=['float64', 'int64']).columns
df[numeric_cols] = df[numeric_cols].apply(pd.to_numeric, errors='coerce')
non_numeric_cols = df.columns.difference(numeric_cols)
numeric_df = df.drop(columns=non_numeric_cols)
def find_outliers_IQR(df):
   q1=df.quantile(0.25)
   q3=df.quantile(0.75)
   IQR=q3-q1
   outliers = df[((df<(q1-1.5*IQR)) | (df>(q3+1.5*IQR)))]
   return outliers
for i in numeric_df.columns:
   outliers = find_outliers_IQR(df[i])
   print('number of outliers in',i,': '+ str(len(outliers)))
# Visualization
for i in numeric_df.columns:
```

```
    fig = px.box(df, y=i)
    fig.show()
```

Outlier Treatment:

Using capping:

```
# Handling outliers using capping for goals column
q1 = df["Goals"].quantile(0.25)
q3 = df["Goals"].quantile(0.75)
IQR=q3-q1
upper_limit = q3 + (IQR*1.5)
lower_limit = q1 - (IQR*1.5)
df["Goals"] = np.where(df["Goals"]> upper_limit, upper_limit,np.where(df["Goals"] <
lower_limit,lower_limit,df["Goals"]))
```

Using mean value:

```
# Handling outliers using mean for height and weight columns
def impute_outliers_IQR_mean(df):
    q1=df.quantile(0.25)
    q3=df.quantile(0.75)
    IQR=q3-q1
    upper = df[~(df>(q3+1.5*IQR))].max()
    lower = df[~(df<(q1-1.5*IQR))].min()
    df = np.where(df > upper, df.mean(), np.where(df < lower,df.mean(),df))
    return df
df['Height'] = impute_outliers_IQR_mean(df['Height'])
df['Weight'] = impute_outliers_IQR_mean(df['Weight'])
# Creating new column "SerialID" and new csv file for further analysis
df.insert(0, 'SerialID', range(1, len(df)+1))
# Creating new cleaned csv file
df.to_csv('new_sports.csv', index=False)
```

Distribution of Players:

```
colors = ['#B09E99', '#FAD4C0', '#64B6AC', '#FEE9E1']
plt.figure(figsize=(8, 8))
position_counts.plot(kind='pie', autopct='%1.1f%%', startangle=140, colors=colors)
plt.title('Distribution of Players by Position')
plt.ylabel('')
plt.show()
```

Analysis of height of players by position:

```
plt.figure(figsize=(12, 8))
sns.violinplot(x='Position', y='Height', data=df, palette='Set3')
plt.title('Distribution of Height by Position')
plt.xlabel('Position')
plt.ylabel('Height (cm)')
```

```
plt.show()
```

Average age by position:

```
avg_age_per_position = df.groupby('Position')['Age'].mean().sort_values(ascending=False)
fig1 = px.bar(avg_age_per_position, title='Average Age per Position')
fig1.update_layout(xaxis_title='Position', yaxis_title='Average Age')
fig1.show()
```

Total goals for each position:

```
total_goals_per_position =
df.groupby('Position')['Goals'].sum().sort_values(ascending=False)
fig2 = px.bar(total_goals_per_position, title='Total Goals per Position')
fig2.update_layout(xaxis_title='Position', yaxis_title='Total Goals')
fig2.show()
```

Horizontal plot for goals scored by each team throughout the seasons:

```
team_goals_season = df.groupby(['Team', 'Season'])['Goals'].sum().reset_index()
fig = px.bar(team_goals_season, x='Goals', y='Team', color='Season',
        orientation='h', title='Total Goals Scored by Each Team in Different Seasons',
        labels={'Goals': 'Total Goals', 'Team': 'Team', 'Season': 'Season'})
fig.show()
```

Analysing the data to find the rising stars and their goal scoring ability:

```
age_threshold = 25
df['GoalsDiff'] = df.groupby('Player')['Goals'].diff()
rising_stars = df[(df['Age'] < age_threshold) & (df['GoalsDiff'] > 0)]
fig = px.scatter(rising_stars, x='Age', y='GoalsDiff', color='Player',
            title='Rising Stars: Increase in Goal-Scoring Ability for Young Players',
            labels={'Age': 'Age', 'GoalsDiff': 'Increase in Goals', 'Player': 'Player'})
fig.show()
```

Analyzing the goals scored by top team across season and finding out the peak performance year of top team:

```
top_team_season_goals = df[df['Team'] ==
top_team].groupby('Season')['Goals'].sum().sort_index()
fig2 = px.line(top_team_season_goals, title=f'Goals Scored by top team -{top_team} Across
Seasons  ')
fig2.update_layout(xaxis_title='Season', yaxis_title='Goals')
fig2.show()
```

Connecting to MySQL Database:

```
def create_db_connection(host_name, user_name, user_password, db_name, port_no):
    connection = None
    try:
        connection = mysql.connector.connect(
            host=host_name,
            user=user_name,
            passwd=user_password,
            database=db_name,
            port=port_no
        )
        print("MySQL Database connection successful")
    except Error as err:
        print(f"Error: '{err}'")
    return connection
connection=create_db_connection("localhost","root","Mysqlroot1","sports_database",3307)
```

Ingestion of Data:

```
def insert_values(connection, table_name, data):
    cursor = connection.cursor()
    for i, row in data.iterrows():
        columns = ', '.join(row.index)
        values_template = ', '.join(['%s'] * len(row))
        query = f"INSERT INTO {table_name} ({columns}) VALUES ({values_template})"
        cursor.execute(query, tuple(row))
    connection.commit()
data = pd.read_csv("new_sports.csv")
```

Creation of log file:

```
logging.basicConfig(filename='ingestion_info.log', level=logging.INFO,
format='%(asctime)s - %(levelname)s - %(message)s')
def perform_ingestion():
    try:
        # Inserting values into the database
        logging.info("Ingestion process started")
        insert_values(connection, "Playerinfo", data[['SerialID', 'Player', 'Team', 'Position',
'Season']])
        insert_values(connection, "Playerdemographics", data[['SerialID', 'Age', 'Height',
'Weight']])
        insert_values(connection, "Playerstats", data[['SerialID', 'Goals', 'Assists', 'YellowCards',
'RedCards', 'PassCompletionRate', 'DistanceCovered', 'Sprints', 'ShotsOnTarget',
'TacklesWon', 'CleanSheets']])
        insert_values(connection, "Playertraininginfo", data[['SerialID', 'TrainingHours',
'EffectiveTraining']])
        insert_values(connection, "Playerperformance", data[['SerialID',
'PressurePerformanceImpact', 'MatchPressure']])
```

```
        insert_values(connection, "Playerhealth", data[['SerialID', 'InjuryHistory',
'PlayerFatigue', 'FatigueInjuryCorrelation']])
        logging.info("Ingestion process completed successfully")
    except Exception as e:
        logging.error(f"Error during ingestion process: {str(e)}")
perform_ingestion()
```

Creating security and access control mechanisms:

```
cursor = connection.cursor()
# Create a new user
def create_user(username, password):
    query = f"CREATE USER '{username}'@'localhost' IDENTIFIED BY '{password}'"
    cursor.execute(query)
    print(f"User {username} created successfully")
# Grant privileges to the user
def grant_privileges(username, database):
    query = f"GRANT ALL PRIVILEGES ON {database}.* TO '{username}'@'localhost'"
    cursor.execute(query)
    print(f"Privileges granted to {username}")
# Grant specific privileges to the user
def grant_specific_privileges(username, database, privileges):
    query = f"GRANT {privileges} ON {database}.* TO '{username}'@'localhost'"
    cursor.execute(query)
    print(f"{privileges} privileges granted to {username}")
# Revoke privileges from the user
def revoke_privileges(username, database):
    query = f"REVOKE ALL PRIVILEGES ON {database}.* FROM
'{username}'@'localhost'"
    cursor.execute(query)
    print(f"Privileges revoked from {username}")
# Revoke specific privileges from the user
def revoke_specific_privileges(username, database, privileges):
    query = f"REVOKE {privileges} ON {database}.* FROM '{username}'@'localhost'"
    cursor.execute(query)
    print(f"{privileges} privileges revoked from {username}")
# Drop user
def drop_user(username):
    query = f"DROP USER '{username}'@'localhost'"
    cursor.execute(query)
    print(f"User {username} dropped successfully")
# Example
create_user('new_user', 'mysqluser1234')
#granting read only privileges
grant_specific_privileges('new_user', 'sports','SELECT')
# Commit the changes
connection.commit()
# Close the connection
connection.close()
```

Analysis on Pass Completion Rate vs. Assists:

 # Scatter plot

```
import matplotlib.pyplot as plt
import seaborn as sns
# Let's make a scatter plot to see how Pass Completion Rate and Assists are related
plt.figure(figsize=(10, 6)) # Set the size of the plot
sns.scatterplot(x='PassCompletionRate', y='Assists', data=df) # Plot the data
plt.title('Scatter Plot of Pass Completion Rate vs. Assists') # Adding a title to the plot
plt.xlabel('Pass Completion Rate')
plt.ylabel('Assists')
plt.show()
```

 # Outlier Detection using Isolation Forest

```
from sklearn.ensemble import IsolationForest
# Use Isolation Forest to find outliers in our data
iso_forest = IsolationForest(contamination=0.1)  # Set how many outliers we expect
outliers = iso_forest.fit_predict(df)  # Fit the model and predict outliers
df['Outlier'] = outliers  # Add the outlier info to our dataframe
# Plot the data, showing which points are outliers
plt.figure(figsize=(10, 6))  # Set the size of the plot
sns.scatterplot(x='PassCompletionRate', y='Assists', hue='Outlier', data=df, palette={1: 'blue',
-1: 'red'})  # Plot the data with outliers highlighted
plt.title('Scatter Plot with Outliers Detected by Isolation Forest')
plt.xlabel('Pass Completion Rate')
plt.ylabel('Assists')
plt.show()  # Show the plot
```

#Regression analysis on pass completion rate and assists:

```
from sklearn.linear_model import LinearRegression
# Set up the data for the regression
X = df[['PassCompletionRate']] # Our feature (independent variable)
y = df['Assists'] # Our target (dependent variable)
# Creating and fitting the linear regression model
model = LinearRegression()
model.fit(X, y)
y_pred = model.predict(X)
# Plot the scatter plot with the regression line
plt.figure(figsize=(10, 6))
sns.scatterplot(x='PassCompletionRate', y='Assists', data=df) # Plot the original data
plt.plot(df['PassCompletionRate'], y_pred, color='red', linewidth=2) # Plot the regression line
plt.title('Regression Line: Pass Completion Rate vs. Assists')
plt.xlabel('Pass Completion Rate')
plt.ylabel('Assists')
plt.show()
```

PCA Visualization:

```
# Draw a graph to show our data in a simple way
# This helps us understand the patterns and relationships between different parts of the data
import matplotlib.pyplot as plt  # Import matplotlib for visualization
plt.scatter(new_sports_pca[:, 0], new_sports_pca[:, 1], c=y, cmap='viridis')  # Scatter plot of PCA components
plt.xlabel('Principal Component 1')  # X-axis label
plt.ylabel('Principal Component 2')  # Y-axis label
plt.title('PCA Visualization')  # Title of the plot
plt.colorbar(label='EffectiveTraining')  # Color bar legend for target variable
plt.show()  # Display the plot
```

APPENDIX-II

DATASHEET

	A	B	C
1	Column Name	Description	Data Type
2	Player	Player name	Text
3	Team	Player's team	Text
4	Age	Player's age	Integer
5	Height	Player's height (cm)	Float
6	Weight	Player's weight (kg)	Float
7	Position	Player's position on the field	Text
8	Goals	Number of goals scored	Integer
9	Assists	Number of assists provided	Integer
10	Yellow Cards	Number of yellow cards received	Integer
11	Red Cards	Number of red cards received	Integer
12	Pass Completion Rate	Percentage of successful passes	Float
13	Distance Covered	Total distance covered on the field (m)	Float
14	Sprints	Number of sprints performed	Integer
15	Shots On Target	Number of shots on target	Integer
16	Tackles Won	Number of successful tackles	Integer
17	Clean Sheets	Number of clean sheets achieved (goalkeeper only)	Integer
18	Player Fatigue	Player fatigue level (0-100)	Integer
19	Match Pressure	Overall match pressure faced (0-100)	Integer
20	Injury History (Y/N)	Indicates recent injury history	Text
21	Training Hours (last week)	Number of training hours completed in the last week	Integer
22	Fatigue-Injury Correlation	Correlation score between fatigue and injury history (0-1)	Float
23	Pressure-Performance Impact	Score reflecting the impact of match pressure on performance (0-1)	Float
24	Effective Training	Score indicating the effectiveness of recent training (0-1)	Float
25	Season	Season number	Text

INFORMATION REGARDING STUDENT(S)

STUDENT NAME	EMAIL ID	PHONE NUMBER
AKANSHA SHETTY	AKANSHASHETTY07@GMAIL.COM	7259982774
CHIMIRALA KOWSTUBHA	KOWSTUBHACHIMIRALA@GMAIL.COM	7995110124
KAPAROTU VENKATA SURYA THARANI	MORESPACEE123@GMAIL.COM	8817683282