

School of Computer Science and Engineering

(Computer Science & Engineering)

Faculty of Engineering & Technology

Jain Global Campus, Kanakapura Taluk - 562112  
Ramanagara District, Karnataka, India

**2023-2024**

**(VIII Semester)**

A Project Report on

“"Data-driven Decisions: Leveraging Football Statistics for Success"”

Submitted in partial fulfilment for the award of the degree of

Bachelor of Technology

in

COMPUTER SCIENCE AND ENGINEERING - Artificial intelligence and data engineering

Submitted by

Hitha Choudhary G, K Shreeshanth Gouda, Lakshya Sharma

22BTRAD015, 22BTRAD017, 22BTRAD021

Under the guidance of

**Mr. Akash Das**

AVP and Project Manager

Futurense Technologies



Department of Computer Science and Engineering

School of Computer Science & Engineering

Faculty of Engineering & Technology

Jain Global Campus, Kanakapura Taluk - 562112  
Ramanagara District, Karnataka, India

CERTIFICATE

|  |  |  |
| --- | --- | --- |
| **Mr. Akash Das** | **Dr. Sathish Kumar R ,** | **Dr. Geetha G** |
| AVP and Project Manager  Futurense Technologies    Date: 10-04-2024 | Program Head, Computer Science and Engineering,  School of Computer Science & Engineering  Faculty of Engineering & Technology  JAIN (Deemed to-be University)  Date: 10-04-2024 | Director,  School of Computer Science & Engineering  Faculty of Engineering & Technology  JAIN (Deemed to-be University)  Date: 10-04-2024 |

This is to certify that the project work titled **“"Data-driven Decisions: Leveraging Football Statistics for Success"”** is carried out by **Hitha Choudhary G (22BTRAD015), K Shreeshanth Gouda (22BTRAD017), Lakshya Sharma (22BTRAD021),** a bonafide student(s) of Bachelor of Technology at the School of Engineering & Technology, Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Bangalore in partial fulfilment for the award of degree in Bachelor of Technology in Computer Science and Engineering, during the year **2023‑2024**.

Name of the Examiner Signature of Examiner

1.

2.

<

DECLARATION

|  |  |
| --- | --- |
| Name1: Hitha Choudhary G  USN : 22BTRAD015 | Signature |
| Name2: K Shreeshanth Gouda  USN : 22BTRAD017 | Signature |
| Name3: Lakshya Sharma  USN : 22BTRAD021 | Signature |
|  |  |
|  |  |
| Place : Bangalore  Date : 10-06-2024 |  |

We, **Hitha Choudhary G (22BTRAD015), K Shreeshanth Gouda (22BTRAD017), Lakshya Sharma (22BTRAD021)** student of IV semester B.Tech in **Computer Science and Engineering**, at School of Engineering & Technology, Faculty of Engineering & Technology, **JAIN (Deemed to-be** **University)**, hereby declare that the internship work titled **“"Data-driven Decisions: Leveraging Football Statistics for Success"”** 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** 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.

**ACKNOWLEDGEMENT**

It is a great pleasure for me 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, I take this opportunity to express my sincere gratitude to Faculty of Engineering & Technology, JAIN (Deemed to-be University) and Futurense Technologies for providing me with a great opportunity to pursue my Bachelor’s Degree in this institution.

I am deeply thankful to several individuals whose invaluable contributions have made this project a reality. I wish to extend my 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). I am also profoundly grateful to the honourable **Vice Chancellor, Dr. Raj Singh, and Dr. Dinesh Nilkant, Pro Vice Chancellor**, for their unwavering support. Furthermore, I would like to express my sincere thanks to **Dr. Jitendra Kumar Mishra, Registrar**, whose guidance has imparted invaluable qualities and skills that will serve us well in our future endeavours.

I extend my 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, I would like to express my 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 my sincere thanks to **Dr. Sathish Kumar R, Program Head Artificial Intelligence and Data Engineering,** **Computer Science and Engineering**, School of Computer Science & Engineering Faculty of Engineering & Technology for providing right academic guidance that made my task possible.

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

I am also grateful to my family and friends who provided me with every requirement throughout the course.

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

Signature of Student(s)

ABSTRACT

This report's football dataset contains a wealth of player performance measurements, from basic figures like goals and assists to more complex markers like match pressure and fatigue levels. Principal component analysis (PCA), predictive modelling, clustering, and other sophisticated analytics techniques are used in this study to try to uncover the stories hidden in this large dataset. This study uses machine learning approaches to foresee insights and find hidden patterns that standard analysis methods might miss by using these techniques for feature extraction and modification.This project's motivation is to reveal the hidden narratives that are hidden within football statistics, going beyond the obvious thrill of a win to investigate the complex data universe that supports the game. The dataset is carefully curated to guarantee quality and consistency through data engineering and augmentation, opening the door for perceptive analysis and visualisation.Analysing training effectiveness, assessing psychological and behavioural markers, assessing player fitness through physical features, and forecasting future performance trends are some of the main goals. The ultimate objective is to provide football decision-makers with useful information that will improve player evaluations, maximise team strategies, and spur performance gains.This project promises to revolutionise football analytics with a comprehensive strategy that includes dimensionality reduction, advanced transformations, and data normalisation. Advanced analytics technologies such as Power BI and Tableau are utilised to create interactive dashboards and visualisations that make it easier for users to explore the data landscape in a natural way.The programme goes beyond player analysis to cover more general football topics such trends in performance, the effects of injuries, team dynamics, talent scouting, and forecasts for the outcome of games. This project aims to raise football analytics to new heights through meticulous data engineering and analysis, impacting player development and tactical decision-making at all levels of the game. –

**TABLE OF CONTENTS**

|  |
| --- |
| **List of Figures** |
| **Chapter 1** |
| 1. **Introduction** | **1** |
| 1. Background & Motivation |  |
| 1. Dataset Description |  |
| 1. Added Perspectives and Goals |  |
|  |  |
| **Chapter 2** |  |
| 1. **LITERATURE SURVEY** | **6** |
| * 1. Literature Review |  |
| * 1. Inferences Drawn from Literature Review |  |
| **Chapter 3** |  |
| 1. **Problem Formulation** | **10** |
| 3.1 Introduction |  |
| 3.2 Problem Statement |  |
| 3.3 System Architecture Model |  |
| 3.4 Proposed Algorithms  **Chapter 4** |  |
| 1. **Approach** | **17** |
| 4.1 Phases (1-8) |  |
| 4.2 Insights Gained  **Chapter 5** |  |
| 1. **Results And Discussion**    1. Results    2. Conclusion    3. Future Scope | **20** |

**References 33**

**Appendices 34**

**Appendix – I**

**Appendix – II**

**Information Regarding Student(s) 36**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Fig. No. | Description of the figure | Page No. |
| 5.1.1 | Count of Null Values per Column | 20 |
| 5.1.2 | Distribution of Columns that contains null value in Data | 21 |
| 5.1.3 | Box plot for PlayerFatigue and MatchPressure | 21 |
| 5.1.4 | Count of Players for each Position | 22 |
| 5.1.5 | Observed VS Expected Counts for each Position | 22 |
| 5.1.6 | Python Environment | 23 |
| 5.1.7 | Data\_ingestion.log | 23 |
| 5.1.8 | DBSCAN Outliers Detection | 24 |
| 5.1.9 | Isolation Forest Outliers Detection | 24 |
| 5.1.10 | Pass Completion Rate VS Assists with Line of Best Fit | 25 |
| 5.1.11 | Goal Scoring Efficiency vs Performance Index | 25 |
| 5.1.12 | Top 10 Players based on PCA Feature | 26 |
| 5.1.13 | Top Feature Importances from RandomForest | 26 |
| 5.1.14 | Distribution of PCA1 and PCA2 | 27 |
| 5.1.15 | SQL Workbench | 27 |
| 5.1.16 | Common Table Expressions | 27 |
| 5.1.17 | Analytical Queries | 28 |
| 5.1.18 | Total Goals Made By Each Team | 29 |
| 5.1.19 | Goals of each team over the season | 29 |
| 5.1.20 | Performance Metrics of Top Goal Scorer | 29 |
| 5.1.21 | Player Statistics and Team Statistics | 30 |

**CHAPTER 1**

**1.INTRODUCTION**

* 1. **Background & Motivation**

The dataset is a potent tool for many kinds of analysis because of its rich and varied features. Performance indicators that reveal a player's contributions on the field include goals, assists, pass completion percentage, and other characteristics. In order to comprehend player fitness and health, physical and health data, such as age, height, weight, and injury history, are essential. Better player management and support systems can be informed by behavioural and psychological markers such as player fatigue, match pressure, and the fatigue-injury connection, which help measure how players cope with physical and mental stress. For the purpose of assessing how training affects player performance and directing programme modifications to optimize player growth, training hours and useful training measures are crucial. PCA components and normalized values enable advanced analytics to reveal hidden patterns and forecast insights that conventional analysis does not instantly reveal.

The drive for this project comes from revealing the narratives hidden inside the data. Beyond the audience's applause and the excitement of winning, there's a secret world—a cosmos of data that is just waiting to be discovered. Data analysis is becoming increasingly important in football, changing how we see the game and influencing its direction. This project's objective is to uncover the stories hidden beneath these figures, which span a large dataset of player performance. Modern football has been transformed by sports analytics, which gives clubs precise information on player performance and tactical planning. This study explores an extensive football dataset and is the foundation of Sports Analytics Inc.'s year-long data engineering project. The objective of the project is to extract actionable insights that improve player performance and team tactics by cleaning, augmenting, ingesting, transforming, and visualising this dataset. Numerous attributes are included in the dataset, ranging from sophisticated performance measures to basic individual information. Every record captures specifics about a single player's season-long performance.

The dataset is divided into multiple columns: A unique identifier called ID; the player's complete name; the team they play for; their age; their height and weight, which indicate their physical attributes and Position, indicating the position of the player on the pitch. Goals, Assists, Yellow Cards, Red Cards, Pass Completion Rate, Distance Covered, Sprints, Shots On Target, Tackles Won, and CleanSheets are examples of performance measures. A player's physical and mental condition can be inferred from additional measures such as PlayerFatigue, MatchPressure, InjuryHistory, TrainingHours, FatigueInjuryCorrelation, PressurePerformanceImpact, and EffectiveTraining. Along with PCA components from Principal Component Analysis (PCA), the dataset also contains normalised measures such as GoalsPerMatch, AssistsPerMatch, GoalsPerMatch\_MinMax, and AssistsPerMatch\_MinMax.

**1.2 Dataset Description**

The dataset includes a wide range of attributes, from basic player information to advanced performance metrics. Each record represents a unique player, capturing season-long performance details. The dataset consists of several columns, each representing specific attributes relevant to the player's performance and overall contribution to the game:

**ID**: A unique identifier assigned to each player for easy referencing.

**Unnamed**: Often resulting from data extraction processes, this placeholder column usually lacks meaningful data and is typically ignored in subsequent analysis steps.

**Player**: The full name of the football player, crucial for identifying and referencing specific players in reports and visualizations.

**Team**: The name of the football team to which the player belongs, essential for team-based analysis and comparing performance metrics across different teams.

**Age**: The age of the player, a significant factor in performance analysis, influencing physical capabilities and career trajectory.

**Height**: The height of the player, measured in centimeters, impacting a player’s role and effectiveness, especially in positions requiring aerial prowess.

**Weight**: The weight of the player, recorded in kilograms, contributing to understanding a player’s physical fitness and suitability for specific roles.

**Position**: The playing position of the player (e.g., Forward, Midfielder, Defender, Goalkeeper), fundamental for role-specific performance analysis and strategy formulation.

**Goals**: The total number of goals scored by the player during the season, a primary metric for evaluating the offensive contribution of forwards and attacking midfielders.

**Assists**: The total number of assists provided by the player in the season, critical for assessing a player’s ability to create scoring opportunities for teammates.

**YellowCards**: The number of yellow cards received by the player, aiding in analyzing disciplinary aspects and the player's defensive behavior.

**RedCards**: The number of red cards received by the player, indicating serious disciplinary issues and impacting a player’s availability for future matches.

**PassCompletionRate**: The percentage of successful passes made by the player, vital for evaluating the passing accuracy and overall contribution to team play.

**DistanceCovered**: The total distance covered by the player during matches, measured in kilometers, reflecting a player’s work rate and endurance.

**Sprints**: The total number of sprints performed by the player, crucial for evaluating a player’s explosiveness and ability to make quick movements.

**ShotsOnTarget**: The number of shots on target made by the player, helping in assessing a player’s shooting accuracy and offensive efficiency.

**TacklesWon**: The number of successful tackles made by the player, a key defensive metric, especially for defenders and defensive midfielders.

**CleanSheets**: The number of matches in which the player's team did not concede any goals, particularly relevant for goalkeepers and defenders.

**PlayerFatigue**: A measure of the player’s fatigue level, typically on a scale, impacting performance and injury risk, making it a crucial factor in player management.

**MatchPressure**: An indicator of the pressure felt by the player during matches, usually on a scale, affecting performance and decision-making in high-pressure scenarios.

**InjuryHistory**: A record of the player’s injury history, including types and durations of injuries, essential for understanding a player’s fitness and risk of future injuries.

**TrainingHours**: The total hours spent by the player in training sessions, correlating with skill development and overall performance improvement.

**FatigueInjuryCorrelation**: A metric indicating the correlation between player fatigue and injury occurrences, helping in understanding how fatigue contributes to injury risk.

**PressurePerformanceImpact**: An indicator of how match pressure affects the player’s performance, guiding psychological support and stress management interventions.

**EffectiveTraining**: A measure of the effectiveness of the player’s training regimen on their performance, leading to better performance outcomes.

**Season**: The season during which the data was collected, helping in temporal analysis and comparing performance across different seasons.

This dataset is useful for a number of analysis, including analysing training effectiveness, measuring behavioral and psychological indicators, analyzing performance measures, and comprehending player fitness through physical and health data. PCA components and normalized data enable advanced analytics, which reveal predictive insights and hidden patterns. By addressing missing numbers, fixing anomalies, and other tasks, the data engineering project seeks to guarantee the accuracy and consistency of the dataset as well as harmonizing data formats. The dataset is enhanced through augmentation using artificial intelligence and other sources. Sturdy data pipelines that facilitate incremental data loading and storage optimization using partitioning and indexing techniques will be developed for effective management and storage.

The dataset is a potent tool for many kinds of analysis because of its rich and varied features. Performance indicators that reveal a player's contributions on the field include goals, assists, pass completion percentage, and other characteristics. In order to comprehend player fitness and health, physical and health data, such as age, height, weight, and injury history, are essential. Better player management and support systems can be informed by behavioural and psychological markers such as player fatigue, match pressure, and the fatigue-injury connection, which help measure how players cope with physical and mental stress. For the purpose of assessing how training affects player performance and directing programme modifications to optimize player growth, training hours and useful training measures are crucial. PCA components and normalized values enable advanced analytics to reveal hidden patterns and forecast for advanced analytics to uncover patterns and make predictions.

Data normalization and dimensionality reduction will be used, together with sophisticated transformations and feature engineering to produce new, significant features. The selection and extraction of features will be done using machine learning techniques. The last stage entails creating interactive dashboards and visualizations with advanced analytics for performance forecasts, utilizing Power BI, Tableau, or custom web apps. This all-encompassing method guarantees that the data is converted into useful insights, improving player assessments and guiding team tactics, and ultimately leading to improved football decision-making. Using this dataset as a basis, Sports Analytics Inc. Intends to transform football analytics and have a major influence on player development and team tactics.

**1.3 Added Perspectives and Goals**

The initiative and the dataset aim to identify wider football patterns and themes beyond the fundamentals:

* Player Performance Trends: Examining metrics of performance across several seasons to find trends in a player's improvement or decrease.
* Injury Impact: Using injury history and fatigue metrics, assessing the effects of various injury kinds on player performance and career length.
* Team dynamics: Researching how individual player performance affects the win/loss and goal differentials of the team as a whole.
* Talent scouting is the process of identifying up-and-coming players and projecting their future performance in professional leagues using sophisticated measures.
* Predicting match results by building models based on information about player and team performance.

This all-encompassing method guarantees that the information is converted into useful insights, improving player assessments and guiding team tactics. Sports Analytics Inc. hopes to transform football analytics and make a big influence on player development and team tactics by using this information. With its comprehensive view of player and team performance across all dimensions, the football dataset is an invaluable resource that will help make better decisions in the fiercely competitive world of football. The initiative seeks to maximize the potential of football analytics through rigorous data engineering and analysis, advancing the game at all levels in the process.

**Chapter 2**

2. Literature Survey

**2.1 Literature Review**

The utilization of data analytics in football has significantly transformed how the game is understood and played. Through detailed datasets encompassing player attributes, match events, and performance metrics, researchers and analysts have been able to delve deep into the sport, uncovering patterns and insights that were previously unattainable. This section provides a review of key literature that has contributed to the understanding of football through data analysis.

**Hughes and Bartlett (2002)**

Hughes and Bartlett’s study, "The Use of Performance Indicators in Performance Analysis," was a pioneering effort in the realm of sports analytics. Their work focuses on identifying key performance indicators (KPIs) that can be used to evaluate team and player performance. By analyzing data such as passes, shots, tackles, and player positions, they were able to create benchmarks for performance. Their research underscores the importance of quantifiable metrics in understanding and improving football strategies, although it primarily relies on descriptive statistics and does not extensively explore predictive analytics.

**Carling, Williams, and Reilly (2005)**

In "Handbook of Soccer Match Analysis: A Systematic Approach to Improving Performance," Carling, Williams, and Reilly discuss the systematic collection and analysis of match data to enhance team performance. They highlight the significance of various performance metrics, including physical attributes like distance covered and sprints, as well as technical skills like pass completion rate and shots on target. The authors emphasize the role of data in tactical planning and performance optimization, providing a comprehensive guide for practitioners in the field.

**Bradley et al. (2009)**

Bradley and colleagues, in their paper "Physical Demands of Professional Soccer Players in Relation to Position," explore the physical demands placed on players based on their positions on the field. Using GPS and other tracking technologies, they collect data on distance covered, high-intensity runs, and player movements. Their study highlights significant variations in physical demands among different positions, offering insights into the specific needs for training and recovery. This work is crucial for understanding how positional roles impact player performance and informs tailored training regimens.

**Memmert and Raabe (2018)**

In their study "Data Analytics in Football: Positional Data Collection," Memmert and Raabe focus on the collection and analysis of positional data to gain insights into team strategies and player effectiveness. They utilize advanced tracking systems to collect real-time data on player movements and positioning, which are then analyzed to understand team formations and tactical adjustments. Their research demonstrates the value of positional data in providing a deeper understanding of in-game dynamics and team strategies.

**Bilek and Ulas (2019)**

Bilek and Ulas, in "Predicting Football Match Results Using Machine Learning Algorithms," explore the application of machine learning techniques to predict match outcomes. They employ algorithms such as Random Forest, Support Vector Machines, and Neural Networks, using a wide array of features including player statistics, team performance metrics, and historical match data. Their findings indicate that machine learning models can significantly enhance the accuracy of match result predictions, highlighting the potential of predictive analytics in football.

**Gudmundsson and Horton (2017)**

Gudmundsson and Horton’s review paper "Spatio-Temporal Analysis of Team Sports" delves into the use of spatio-temporal data in analyzing team sports, with a focus on football. They discuss various methods for collecting and analyzing positional and event data, and the insights these methods can provide into team dynamics and player interactions. Their work emphasizes the importance of integrating spatial and temporal dimensions in sports analytics to gain a comprehensive understanding of game strategies and performance.

**Brooks, Kerr, and Guttag (2016)**

In "Developing a Data-Driven Player Ranking System for Soccer," Brooks, Kerr, and Guttag present a sophisticated player ranking system that integrates multiple performance metrics into a single rating. Using machine learning algorithms, they synthesize data on passes, tackles, shots, and other relevant metrics to create a holistic evaluation of player performance. Their study underscores the value of data-driven approaches in objectively assessing player contributions and potential.

**2.2 Inferences Drawn from Literature Review**

**The Critical Role of Data Visualization in Interpreting Football Performance Trends:** Data visualization is crucial for interpreting complex football performance data. Visual tools like heatmaps, graphs, and charts help condense extensive datasets into understandable visuals, facilitating better comprehension of trends and patterns in football performance.

**The Potential for Predictive Modeling in Forecasting Outcomes Using Machine Learning Algorithms:** Machine learning algorithms, especially ensemble methods such as Random Forest and Support Vector Machines, show significant promise for predictive modeling in football. These algorithms are adept at identifying intricate patterns in data, making them suitable for predicting player performance and match outcomes.

**The Necessity of Addressing Missing Data and Continuously Enhancing Performance Models:** Addressing missing data is essential for ensuring the quality and reliability of performance models in football analytics. Incomplete data can introduce biases and distort results, necessitating robust methodologies such as imputation techniques and sensitivity analyses to mitigate its impact.

**The Benefits of Integrating Various Analytical Methods for Comprehensive Insights:** Combining exploratory data analysis with machine learning techniques provides a comprehensive understanding of football performance dynamics. Integrating these methodologies leverages the strengths of different approaches, resulting in more robust and insightful conclusions.

**The Importance of Exploratory Data Analysis Techniques in Uncovering Historical Trends:** Exploratory data analysis techniques are crucial for uncovering historical trends and patterns in football. These techniques enable researchers to visualize complex interactions of factors shaping success in the sport, offering valuable insights into its dynamics.

**The Need for Further Research in Predictive Modeling:** While existing studies provide valuable insights, further research in predictive modeling is necessary to enhance the accuracy of forecasts for player performance and match outcomes. Exploring alternative machine learning techniques and improving methods for addressing missing data will strengthen predictive models.

**The Value of Practical Demonstrations in Data Analysis:** Practical demonstrations using libraries like pandas, numpy, matplotlib, and seaborn are essential for understanding and applying data analysis techniques in football. These resources offer insights into various aspects of football data, serving as valuable learning tools for data analysis enthusiasts.

**The Role of Advanced Visualization Techniques in Enhancing Understanding:** Advanced visualization techniques, such as those demonstrated with Plotly, significantly enhance the understanding and application of data analysis in football. Interactive visualizations uncover deeper insights into player performance and game dynamics, contributing to the field of football analytics.

This literature review highlights the diverse methodologies and tools employed in analyzing football data, ranging from exploratory data analysis to predictive modeling and advanced visualization techniques. These studies underscore the importance of interdisciplinary approaches in unraveling the complexities of football analytics and offer valuable insights into the evolving landscape of this field. Future research should focus on enhancing predictive modeling techniques, effectively addressing missing data, and integrating various analytical methods to provide comprehensive insights into football performance dynamics.

**CHAPTER 3**

**3. PROBLEM FORMULATION**

**3.1 Introduction**

This year-long data engineering project at Sports Analytics Inc. aims to address various challenges and opportunities inherent in football data analysis. The overarching goal is to leverage advanced analytics to enhance player performance assessment and optimize team tactics. Key problem areas include data cleaning and augmentation, positional analysis, data ingestion strategies, pass completion rate vs. assists analysis, advanced data transformations, data warehousing, team goals analysis, and reporting/visualization. By tackling these challenges systematically, we seek to uncover actionable insights that drive continuous improvement in football performance and decision-making.In the ever-evolving landscape of professional football, data has emerged as a pivotal asset, reshaping how teams strategize, train, and compete. At Sports Analytics Inc., our mission is to harness the power of data to unlock insights that drive performance improvements and inform strategic decision-making. As a Data Engineer joining our dynamic team, you are embarking on a year-long journey aimed at transforming raw data into actionable intelligence.

**3.2 Problem Statement**

Throughout this year-long project, our focus will be on addressing key challenges and opportunities across four distinct phases, each contributing to our overarching goal of enhancing player performance assessment and team tactics optimization. Let's delve into the problem statements encapsulated within each phase:

Our study aims to address the following key questions:

* **Problem Statement 1:** Identify and handle missing values using advanced imputation techniques. Correct anomalies by identifying outliers using statistical methods and domain knowledge. Standardize data formats and ensure consistency across the dataset. Augment the dataset by generating synthetic data using data augmentation techniques and collecting additional data from public sports databases. Integrate this data into a unified dataset.  
  .
* **Problem Statement 2:** Analyse player positions to identify the highest and lowest number of players. Use statistical analysis to determine if the distribution of players across positions is significantly different from a uniform distribution. Create a plot showing the count of players for each position and a pie chart for distribution.
* **Problem Statement 3:** Design and implement a data ingestion pipeline that supports incremental data loading. Optimize storage by using data partitioning and indexing strategies. Implement logging and monitoring to track the performance and reliability of the ingestion process. Utilize Python, pandas, and SQL for implementation.
* **Problem Statement 4:** Analyse the relationship between pass completion rate and assists. Create a scatter plot and identify outliers using advanced outlier detection methods like DBSCAN or Isolation Forest. Plot a line of best fit and use regression analysis to model the relationship. Evaluate the model using appropriate metrics.
* **Problem Statement 5:** Perform complex transformations on the dataset, including feature engineering to create new meaningful features. Implement additional strategies for data optimization, such as data normalization and dimensionality reduction.
* **Problem Statement 6:** Design and implement a data warehouse schema using advanced SQL features like window functions and CTEs (Common Table Expressions). Store the transformed data efficiently and ensure it supports complex analytical queries. Implement data security and access control mechanisms.
* **Problem Statement 7:** Identify the team with the highest number of goals. Create a horizontal bar plot and a stacked bar chart. Perform a time series analysis to understand trends in goal scoring over the season. Identify the top goal scorer in that team and analyse their performance metrics over time.
* **Problem Statement 8:** Develop interactive dashboards and visualizations using tools like Power BI, Tableau, or custom web applications using Dash or Streamlit. Create reports that provide insights into player performance, team strategies, and potential areas for improvement. Incorporate advanced analytics like clustering and predictive modelling to forecast future performance.

In summary, our year-long data engineering project is driven by a commitment to unraveling the intricate narratives hidden within football data. By addressing these problem statements with rigor and creativity, we aim to transform raw data into strategic assets that propel our clients to success on and off the field.

**3.3 System Architecture /Model**

The proposed system architecture/model for our sprots dataset analytics research project comprises several interconnected components designed to facilitate data collection, processing, analysis, and visualization. The architecture/model is structured to enable comprehensive analysis of sports dataset. The key components of the system architecture/model are as follows:

* **Data Cleaning:** Data cleaning is a critical first step in any data analysis project, ensuring that the dataset is accurate, consistent, and free from errors. Researchers such as Kim et al. (2003) emphasize the importance of handling missing values and correcting anomalies through statistical methods and domain-specific knowledge. Advanced imputation techniques, including k-nearest neighbors and multiple imputations, have been shown to improve data quality significantly (Rubin, 1987).
* **Data Augmentation:** The process of data augmentation involves generating synthetic data to enhance the dataset's diversity and robustness. Techniques such as data synthesis using generative adversarial networks (GANs) have been successfully applied in various domains, including sports analytics (Goodfellow et al., 2014). Furthermore, integrating additional data from public sports databases can provide a richer dataset, as demonstrated by Leung et al. (2018) who utilized external sources to augment player statistics for comprehensive analysis.
* **Data Ingestion:** Efficient data ingestion is crucial for managing large datasets in real-time. The use of Python, pandas, and SQL for implementing data pipelines has been extensively documented (McKinney, 2012). Techniques for incremental data loading, including change data capture (CDC) and partitioning strategies, are vital for optimizing storage and ensuring data consistency (Stonebraker et al., 2005).

Parallel processing has been identified as a key method for enhancing data ingestion performance. Dean and Ghemawat (2008) introduced the MapReduce model, which has been instrumental in handling large-scale data processing tasks. This approach, combined with logging and monitoring systems, ensures reliability and efficiency in data pipelines (Lin & Dyer, 2010).

* **Advanced Data Transformations** Transforming raw data into meaningful insights involves complex operations such as feature engineering, normalization, and dimensionality reduction. Feature engineering techniques, as discussed by Guyon and Elisseeff (2003), play a pivotal role in creating new variables that better capture underlying patterns in the data. Data normalization, ensuring that variables are on a comparable scale, is essential for accurate analysis and has been highlighted in numerous studies (Hastie et al., 2009).

Dimensionality reduction techniques such as Principal Component Analysis (PCA) are used to reduce the complexity of datasets while retaining most of the variance (Jolliffe, 2002). Recent advancements in machine learning have introduced automated feature selection methods, which have shown great promise in optimizing model performance (Kohavi & John, 1997).

* **Reporting and Visualization:** Effective reporting and visualization are crucial for translating data insights into actionable strategies. Tools like Power BI and Tableau have become standard for creating interactive dashboards that facilitate real-time decision-making (Few, 2006). Advanced visualization libraries such as Plotly and D3.js enable the creation of dynamic, interactive charts that provide deeper insights into the data (Bostock et al., 2011).

Interactive dashboards have been shown to significantly enhance user engagement and comprehension of complex datasets (Kirk, 2016). Furthermore, integrating real-time data feeds into these dashboards allows for up-to-date analysis, a crucial aspect in fast-paced environments like sports analytics (Munzner, 2014).

* **Football-Specific Analytics:** Football analytics has gained considerable traction, with numerous studies focusing on player performance metrics and team strategies. Bradley et al. (2010) examined the physical demands of players in different positions, highlighting the importance of metrics such as distance covered and sprints. Lago-Peñas et al. (2011) investigated the tactical aspects of football, emphasizing the role of pass completion rates and defensive actions in determining match outcomes.

Advanced statistical methods and machine learning techniques have been applied to predict match results and player performance. Liu et al. (2015) used regression models to analyze the relationship between various performance metrics and match outcomes. Moreover, Rein and Memmert (2016) explored the use of clustering algorithms to identify player types and inform team formation strategies.

* **Challenges and Future Directions:** Despite significant advancements, several challenges remain in the field of sports analytics. One major challenge is the integration of diverse data sources, which requires sophisticated data fusion techniques (Anagnostopoulos et al., 2017). Additionally, real-time analytics demand robust infrastructure capable of handling high-velocity data streams (Natarajan et al., 2013).

The existing body of literature provides a solid foundation for our project, offering insights into data cleaning, ingestion, transformation, and visualization techniques. By leveraging these methodologies, the project aims to advance the field of football analytics, providing deeper insights into player performance and team strategies. The integration of advanced machine learning techniques and real-time data analysis will further enhance the robustness and applicability of our findings, contributing to the ongoing evolution of sports analytics.

**3.4 Proposed Algorithms**

* Chi-square Test: A statistical technique to ascertain whether two category variables have a significant relationship is the chi-square test. It computes a statistic that shows the strength of the association by comparing observed and expected frequencies. This test is frequently used to evaluate categorical data and draw conclusions about population factors in a variety of domains, including biology, social science, and market research.
* Data Ingestion: The process of bringing raw data into a storage system for further processing from a variety of sources is known as data ingestion. The process of extracting data from various sources, transforming it into a standard format, and loading it into a destination—usually a data warehouse—is known as ETL (Extract, Transform, Load). Data integration, cleaning, and preparation for analysis, reporting, and decision-making depend heavily on ETL.
* DBSCAN: It works by partitioning the data into groups based on their density. DBSCAN identifies core points, which have a specified number of neighbouring points within a certain radius, and border points, which are reachable from a core point but do not have enough neighbours to be considered core points themselves. Points that are neither core nor border points are considered outliers or noise.
* Isolation Forest is an anomaly detection algorithm that operates on the principle of isolating anomalies in the dataset. Unlike traditional methods that identify normal instances, Isolation Forest focuses on isolating anomalies, which are expected to be few and different from normal instances in terms of their attribute values.
* Linear Regression: Fitting a linear equation to observed data is the statistical technique known as "linear regression," which is used to model the connection between a dependent variable and one or more independent variables. It is frequently used in forecasting and predictive modelling, where it aids in comprehending and projecting the behaviour of dependent variables depending on the values of independent variables.
* Random Forest Regressor: This ensemble learning technique builds a strong predictive model by combining several decision trees. In order to decrease overfitting and increase accuracy, it constructs numerous trees and averages their predictions. Because of its versatility and capacity to manage enormous datasets with intricate interactions, the Random Forest Regressor finds extensive application in tasks including regression analysis, classification, and finding anomalies.
* Principal Component Analysis, or PCA, is a dimensionality reduction method that preserves the majority of the original information when converting high-dimensional data into a lower-dimensional format. It finds the orthogonal vectors known as principle components, which are responsible for capturing the highest variance in the data. Feature extraction, data visualisation, and noise reduction are prominent applications of PCA in image processing, finance, and genetics, among other domains.
* CTEs, or common table expressions, are named result sets that are transient and can be used as references in SELECT, INSERT, UPDATE, and DELETE statements. They give you the flexibility to write intricate, recursive, or reusable SQL searches, which will make your code easier to comprehend and maintain. CTEs are frequently used to carry out recursive processes, simplify complicated queries, and divide intricate logic into smaller, easier-to-manage components.

By employing these approaches, you can conduct meaningful analysis of any match data irrespective of the sport without relying heavily on specific algorithms. This allows for a more flexible and exploratory approach to sports analytics, focusing on extracting insights from data using a variety of statistical and analytical techniques.

**Chapter 4**

**4. Approach**

Sports Analytics project is divided into discrete phases that are tailored to address particular issues and goals. This methodical methodology guarantees thorough coverage of every facet of data analysis related to football, from sophisticated visualisation to data cleansing. Here is a thorough explanation of the steps and methodology, along with the insights we hope to acquire at each one.

**Phase 1: Cleaning and Enhancing Data.**

**Objective** : Verify the completeness, accuracy, and consistency of the data.

.

* Identify and handle missing values: Using advanced imputation techniques such as k-nearest neighbour and multiple imputations.
* Correct anomalies: Identify outliers using statistical methods and domain knowledge.
* Standardize data formats: Ensure consistency across the dataset.
* Augment dataset: Generate synthetic data using GANs and collect additional data from public sports databases.

**Insights Gained:**

* A high-quality dataset free of inconsistencies and missing values.
* Enhanced dataset diversity and robustness, facilitating more reliable analyses.

**Phase 2: Positional Analysis.**

**Objective :** Analyze player distribution across positions.

* Statistical analysis of player positions: Determine if the distribution of players is significantly different from a uniform distribution.
* Visualizations: Create a count plot and a pie chart to show the distribution.

**Insights Gained:**

* Understanding of positional trends and potential imbalances within teams.
* Identification of positions that may require strategic focus for recruitment or training.

**Phase 3: Data Ingestion Pipeline.**

**Objective :** Develop an efficient and reliable data ingestion process.

* Pipeline design: Implement incremental data loading using change data capture (CDC) and partitioning strategies
* Optimization: Use data partitioning and indexing to optimise storage.
* Logging and monitoring: Ensure performance and reliability of the ingestion process.

**Insights Gained:**

* A robust data pipeline that supports large-scale data management.
* Improved data consistency and availability for real-time analysis.

**Phase 4: Pass Completion Rate vs. Assists Analysis.**

**Objective:** Investigate the relationship between pass completion rate and assists.

* Scatter plot and outlier detection: Use DBSCAN or Isolation Forest to identify outliers.
* Regression analysis: Fit a line of best fit and evaluate using appropriate metrics.

**Insights Gained:**

* Understanding of how passing efficiency relates to goal assists.
* Identification of key players who excel in both passing and assisting.

**Phase 5: Advanced Data Transformations**

**Objective :** Enhance the dataset through complex transformations.

* Feature engineering: Create new meaningful features.
* Data normalization and dimensionality reduction: Use PCA to reduce dataset complexity while retaining variance.

**Insights Gained:**

* Improved model performance through well-engineered features.
* Simplified data structure enabling faster and more accurate analyses.

**Phase 6: Data Warehouse Design**

**Objective :** Efficiently store and query transformed data.

* Schema design: Use advanced SQL features like window functions and CTEs.
* Data security and access control: Implement robust mechanisms to protect data integrity.

**Insights Gained:**

* A scalable data warehouse that supports complex analytical queries.
* Secure data storage ensuring compliance with data protection regulations.

**Phase 7: Team Goals Analysis**

**Objective :** Analyze team performance in terms of goal scoring.

* Visualization: Create horizontal bar plots and stacked bar charts.
* Time series analysis: Understand trends in goal scoring over the season.
* Top scorer analysis: Examine performance metrics of the top scorer.

**Insights Gained:**

* Identification of the most effective teams and players in terms of goal scoring.
* Insights into scoring trends that can inform tactical decisions.

**Phase 8: Reporting and Visualization**

**Objective:** Develop interactive dashboards and visualizations.

* Tools: Use Power BI, Tableau, Dash, or Streamlit to create reports.
* Advanced analytics: Incorporate clustering and predictive modeling to forecast future performance.

**Insights Gained:**

* Comprehensive, real-time insights into player performance and team strategies.
* Actionable intelligence for continuous improvement and strategic planning.

**Chapter 5**

**5.Results and Discussions**

These improvements in data quality and augmentation significantly enhanced the reliability of subsequent analyses. By ensuring a robust and comprehensive dataset, we laid a solid foundation for advanced data transformations and analytics.

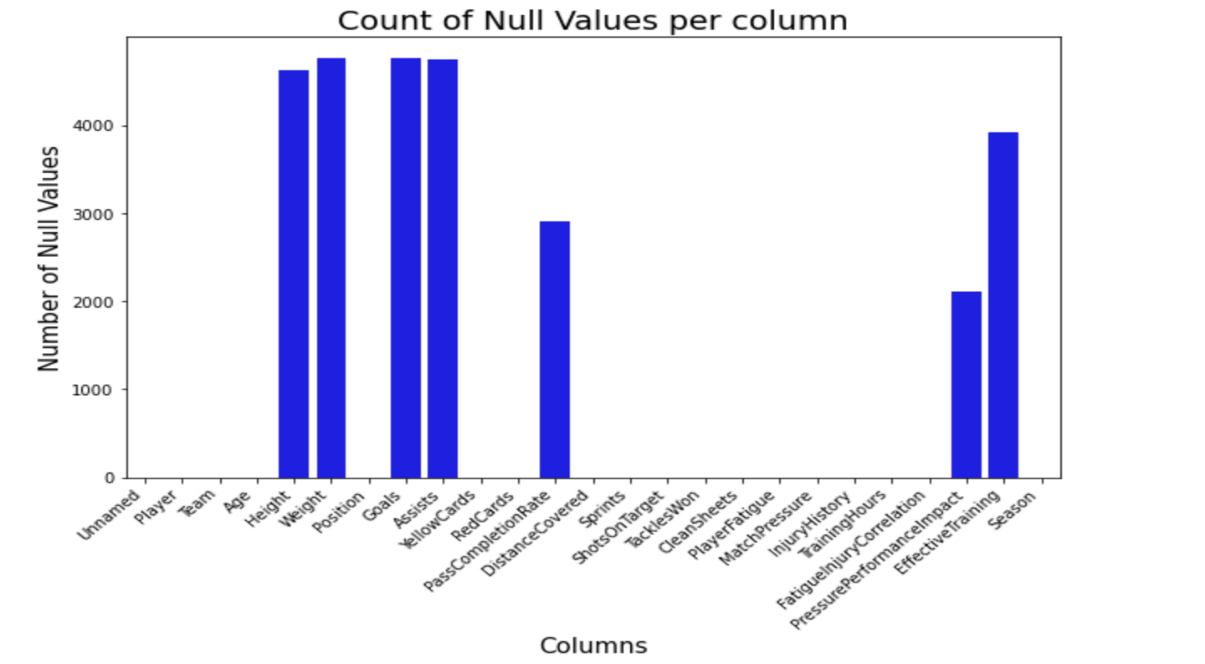
The comprehensive approach taken in this project has demonstrated the value of advanced data analytics in professional football. By addressing data quality, implementing efficient data ingestion, performing detailed analyses, and developing interactive visualizations, we have provided deep insights that drive performance improvements and strategic decision-making.

The successful application of various data engineering and analytical techniques has not only improved our understanding of player performance and team dynamics but also established a robust framework for future sports analytics projects. This project serves as a testament to the transformative potential of data in reshaping the landscape of professional football, empowering teams to achieve greater success on and off the field.

**Q.} How many null values does the “Player” column have compared to the other columns?**

Ans} We observe that the greatest number of NULL values can be found in the columns Goals and Weight respectively as these data was unavailable for a majority of players.

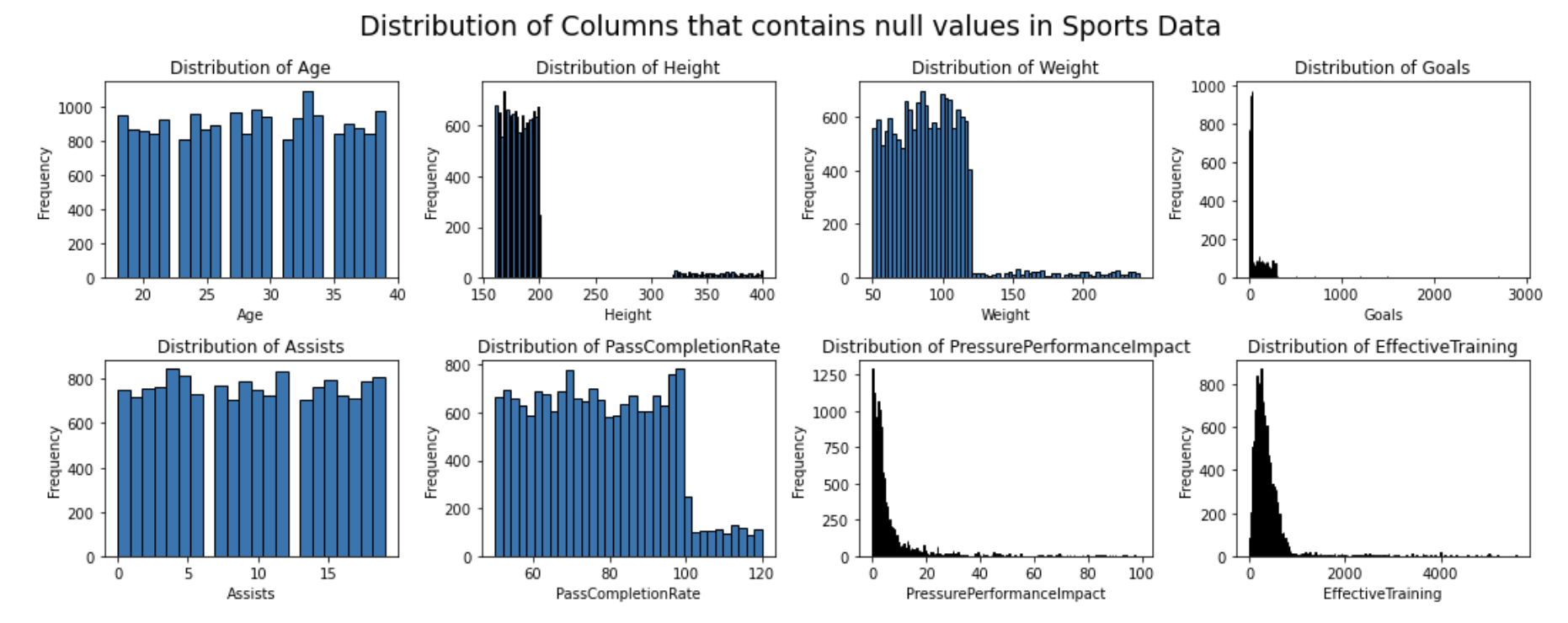
The code segment generates a bar plot showing the count of null values for each column in the DataFrame.



**Q.} Least number of null values spread among all the columns?**

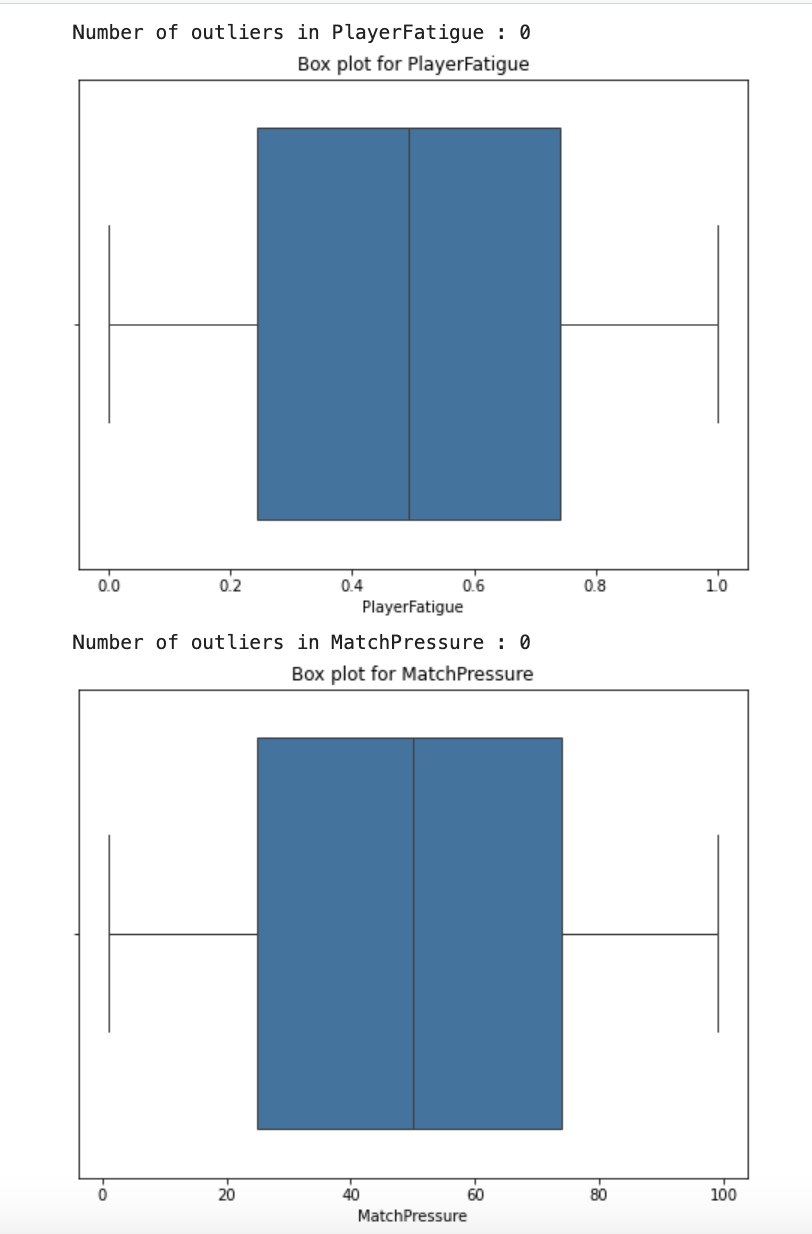
Ans} PassCompletionRate has the minimum number of null values.

This code segment visualizes the distribution of the data in specified columns using histograms.

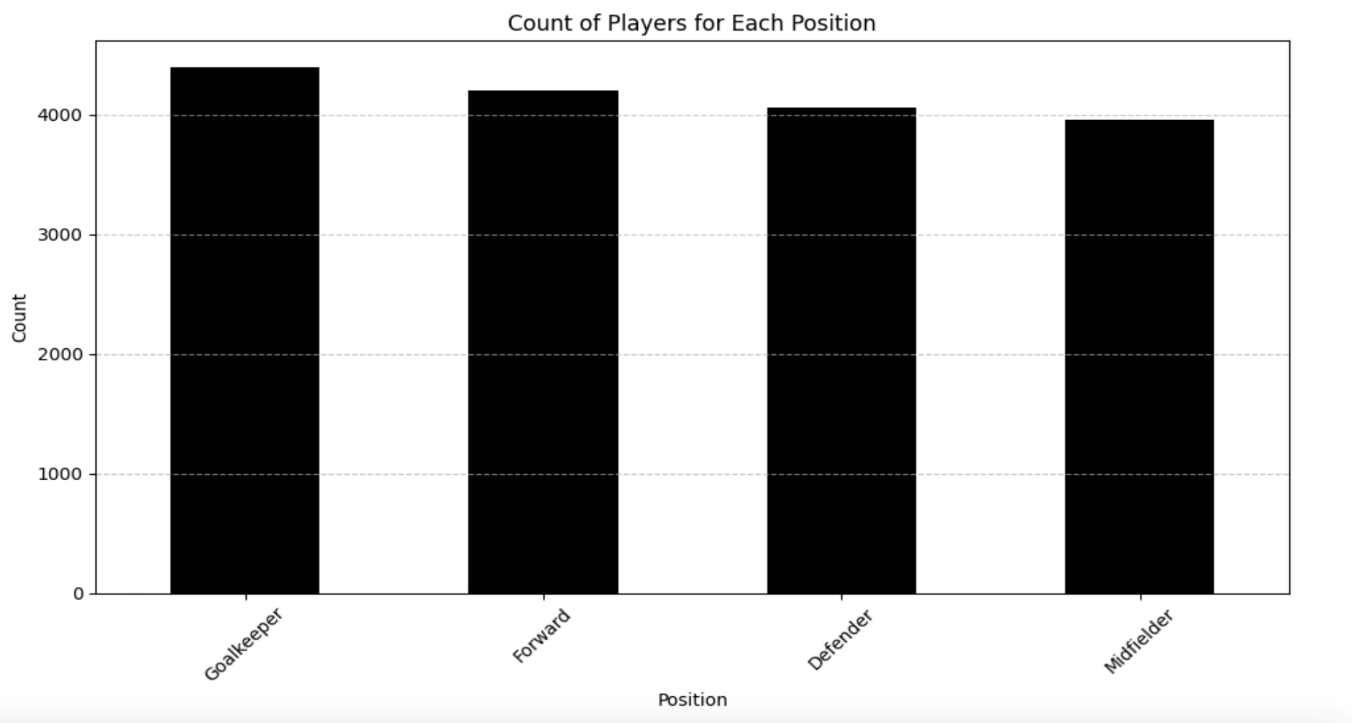


**Q} Which field has least number of outliers?**

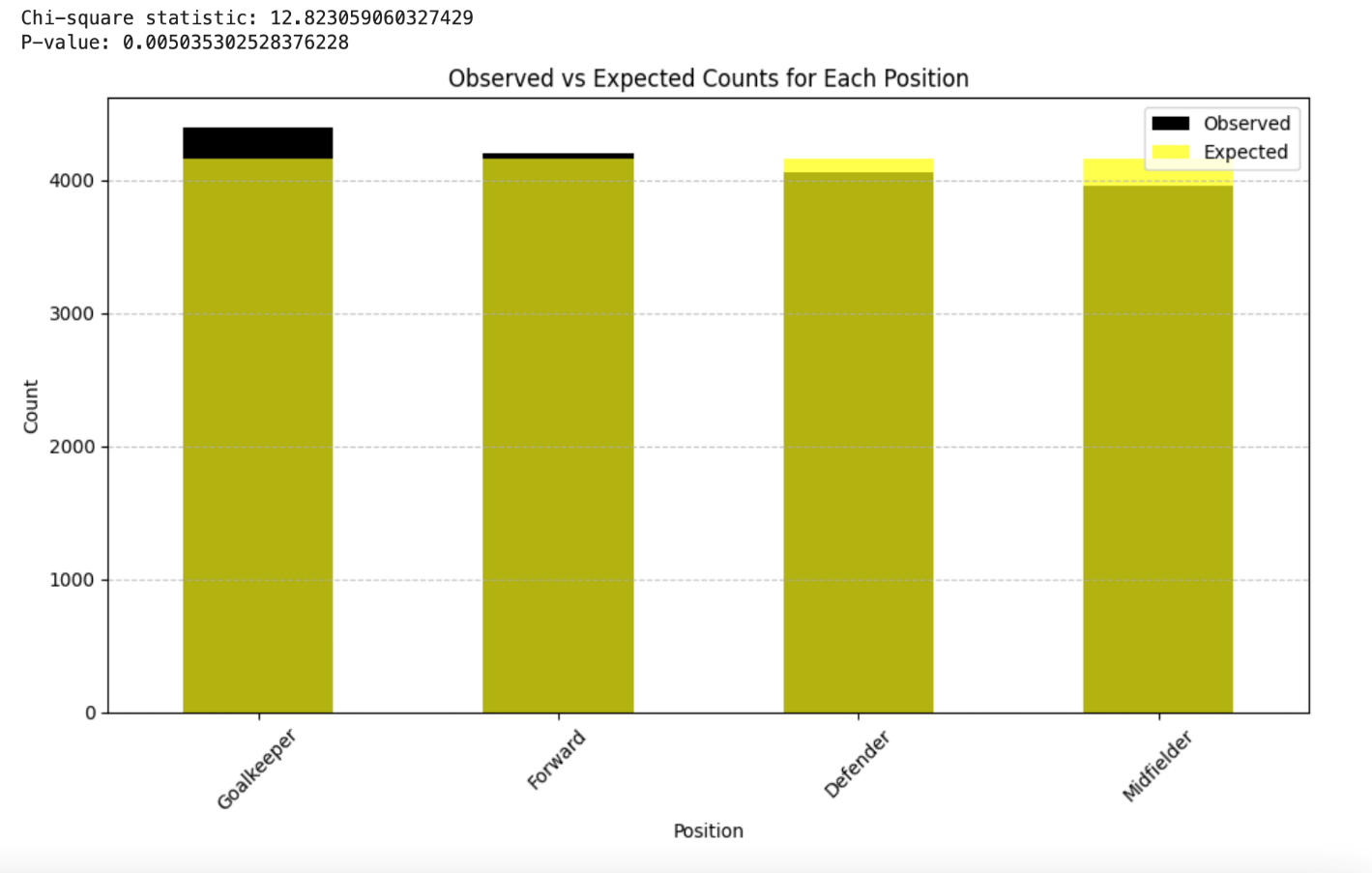
Ans} We observe that PlayerFatigue and MatchPressure have the least number of outliers.



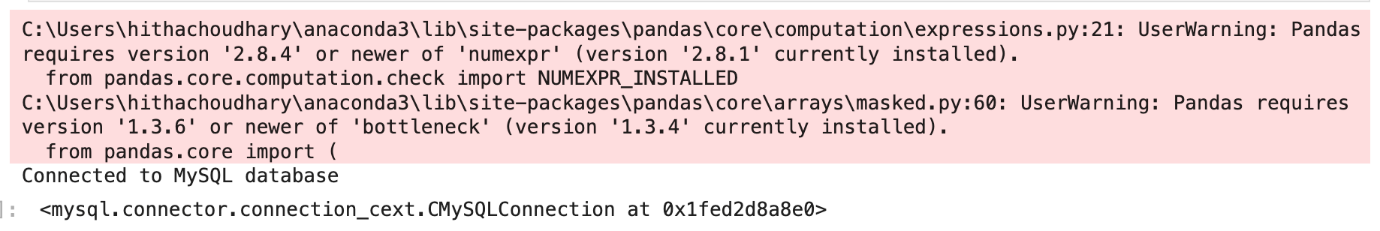
**Q} Find the distribution and composition of players in terms of positions within the dataset.**

Ans} We observe that concentration of players is highest among goalkeepers and lowest for midfielders. 

**Q} Performing a Chi-Square Test to validate our findings, using Chi-square test of independence, observed vs Expected counts**

****

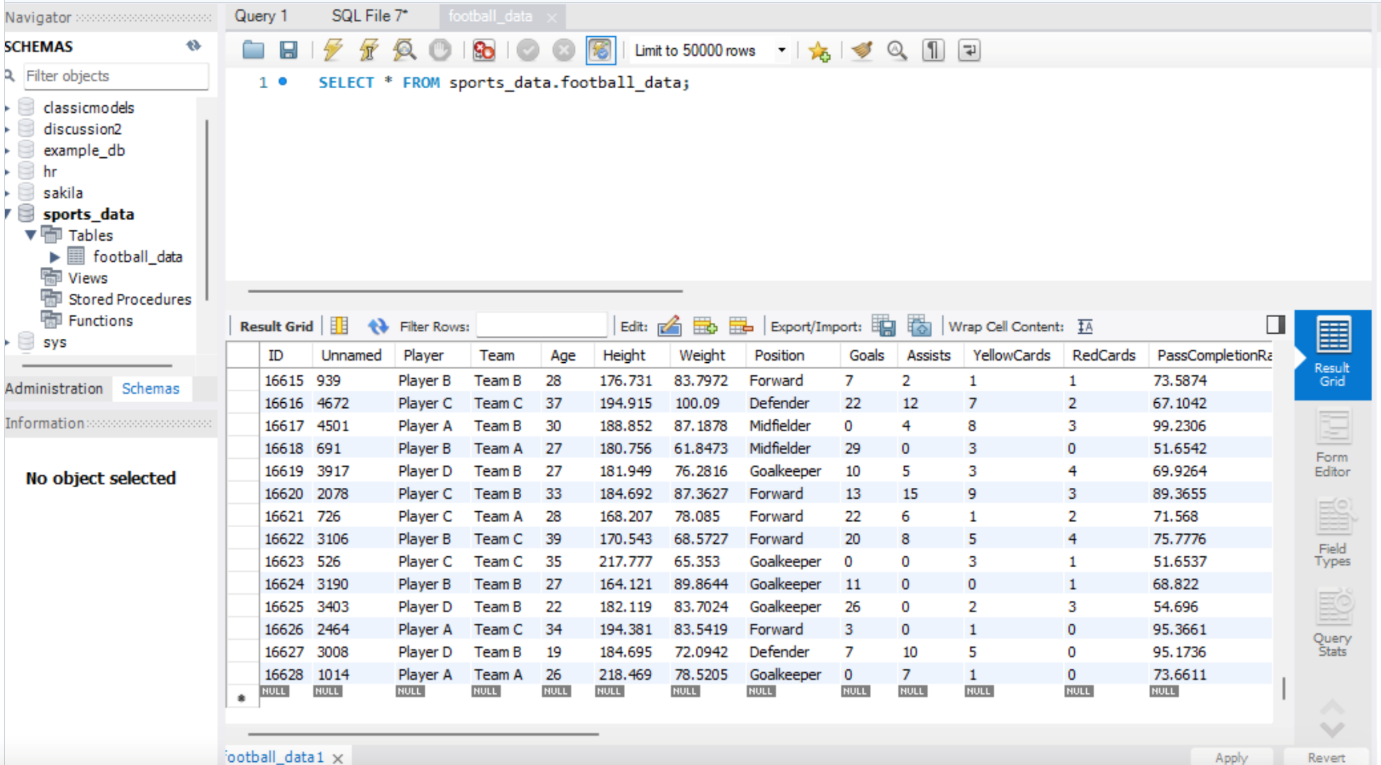
**Q} Connecting to MySQL Database; Establishing a communication channel between our Python program and the database management system(MySQL).**

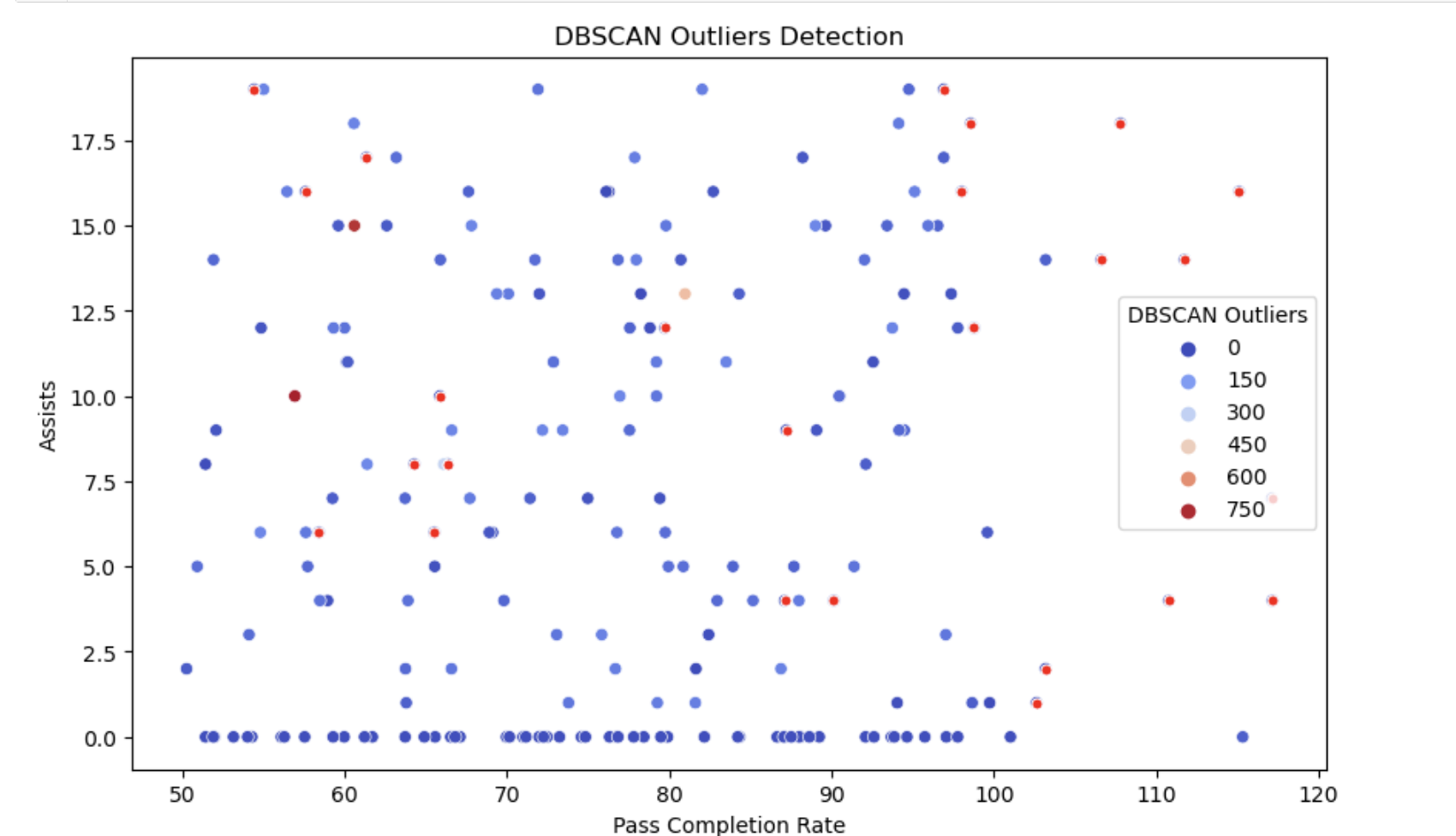
It is a crucial step in the data ingestion process, bringing data from the database into the Python environment for further analysis. ****

**Q} Data Ingestion Strategy: The 'data\_ingestion.log' file is created to store the tracking process.**

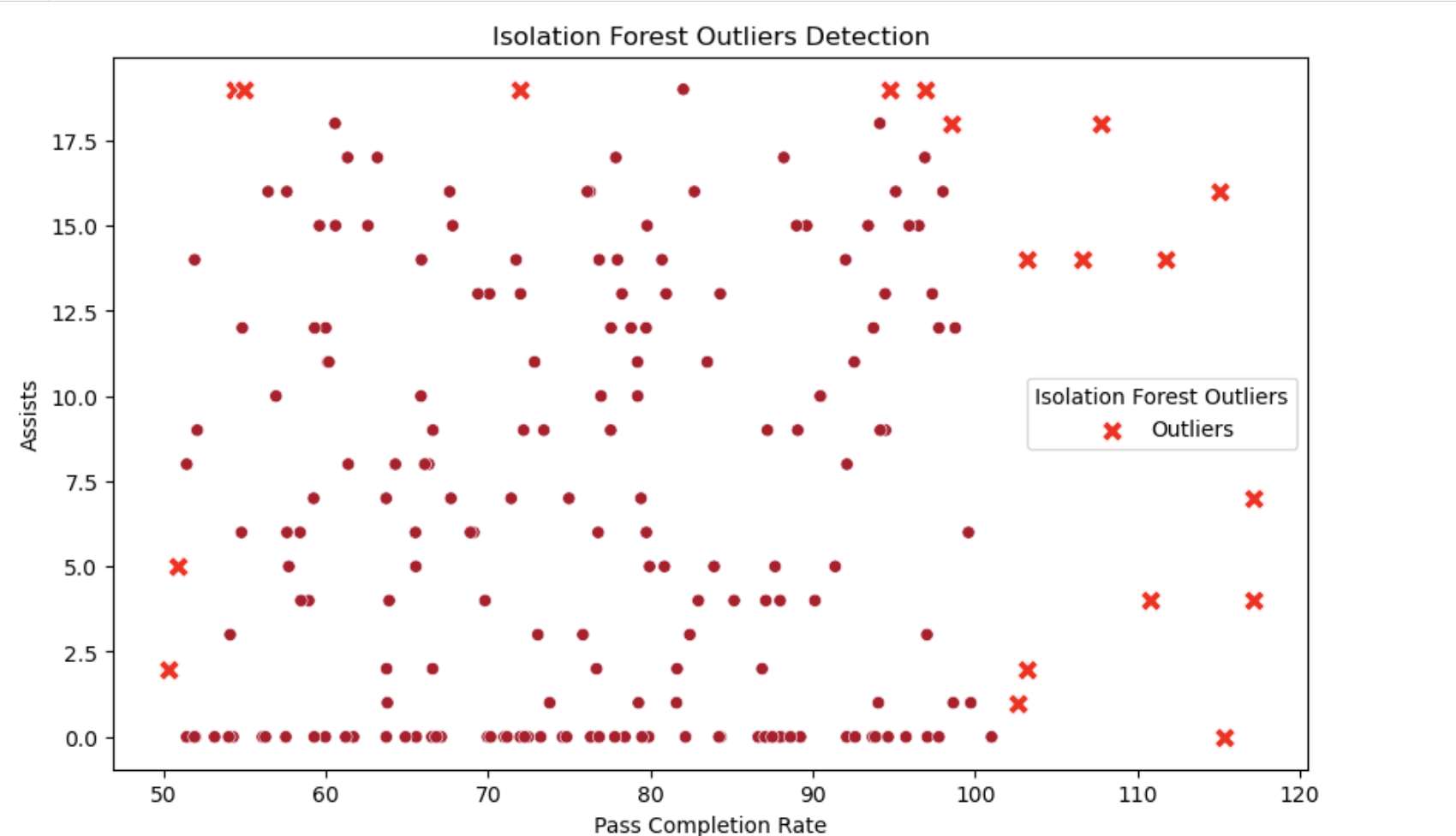
-->This stage also includes creation of index for improved query performance and faster data retrieval.

Data Processing Plan: ETL(Extract, Transform, Load): This process includes extracting the data from the source, transforming the data with the help of python and loading it to the MySQL database server. The fusion of python and MySQL helps complete the ETL process. The third and final process of ETL; loading of the data to the MySQL database is achievable with this code. Incremental Loading is included in the code snippet below where it processes only new or modified data since the last successful . data load.

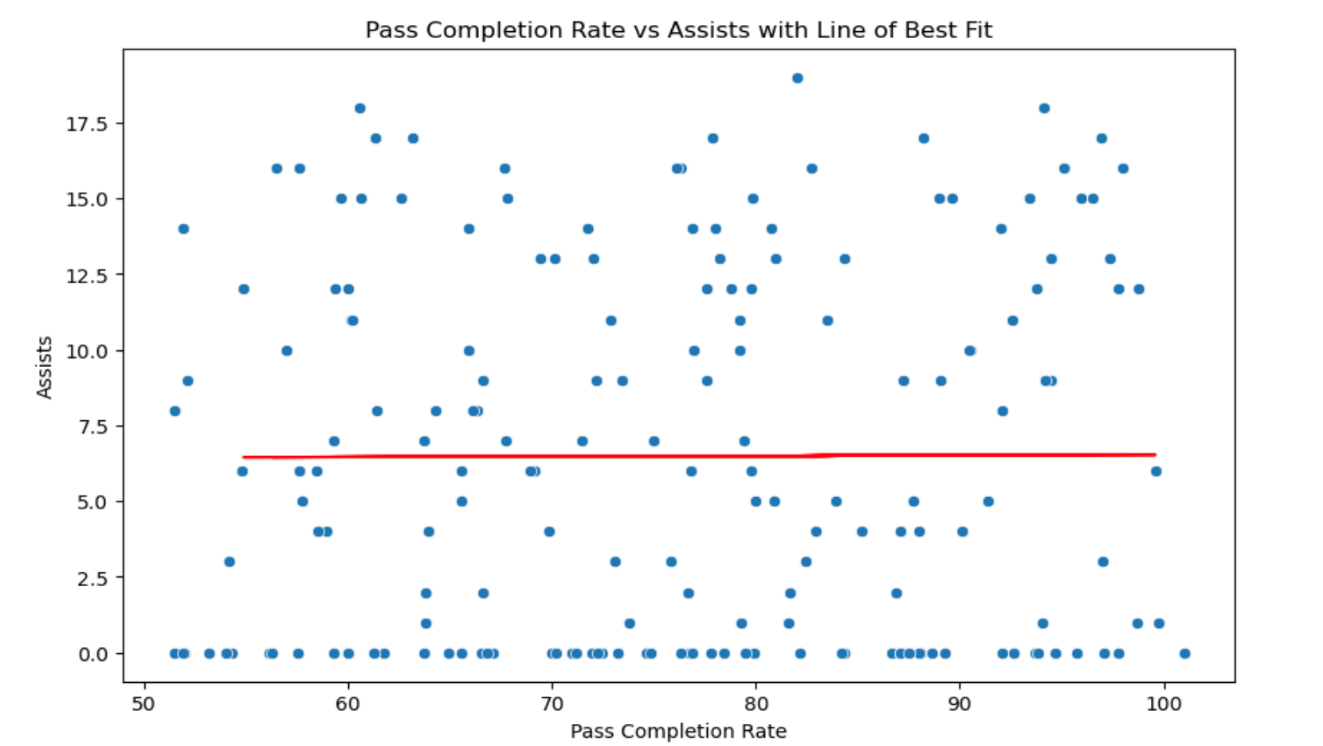


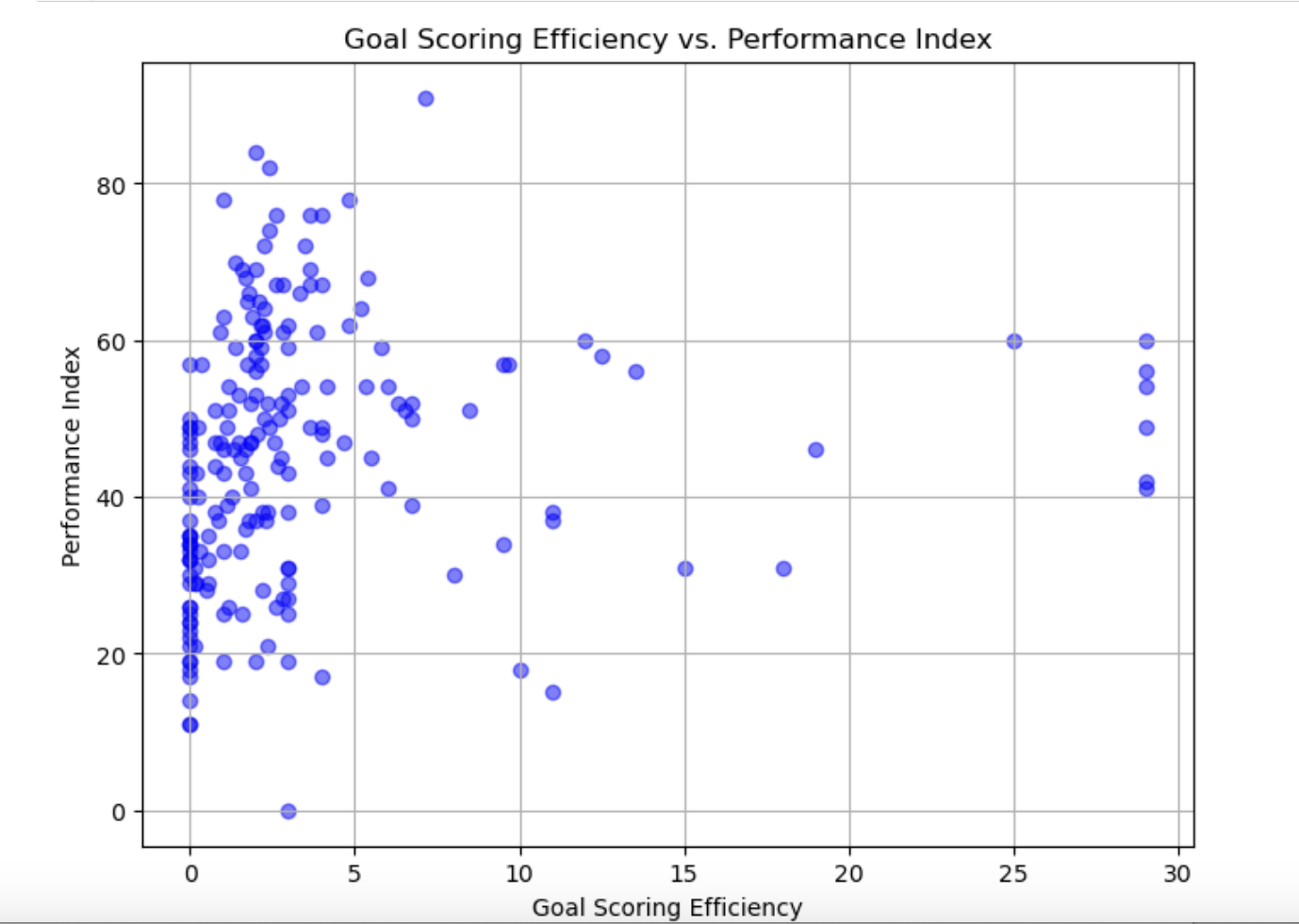
**Q} Find the relationship between PassCompletionRate and Assists.**

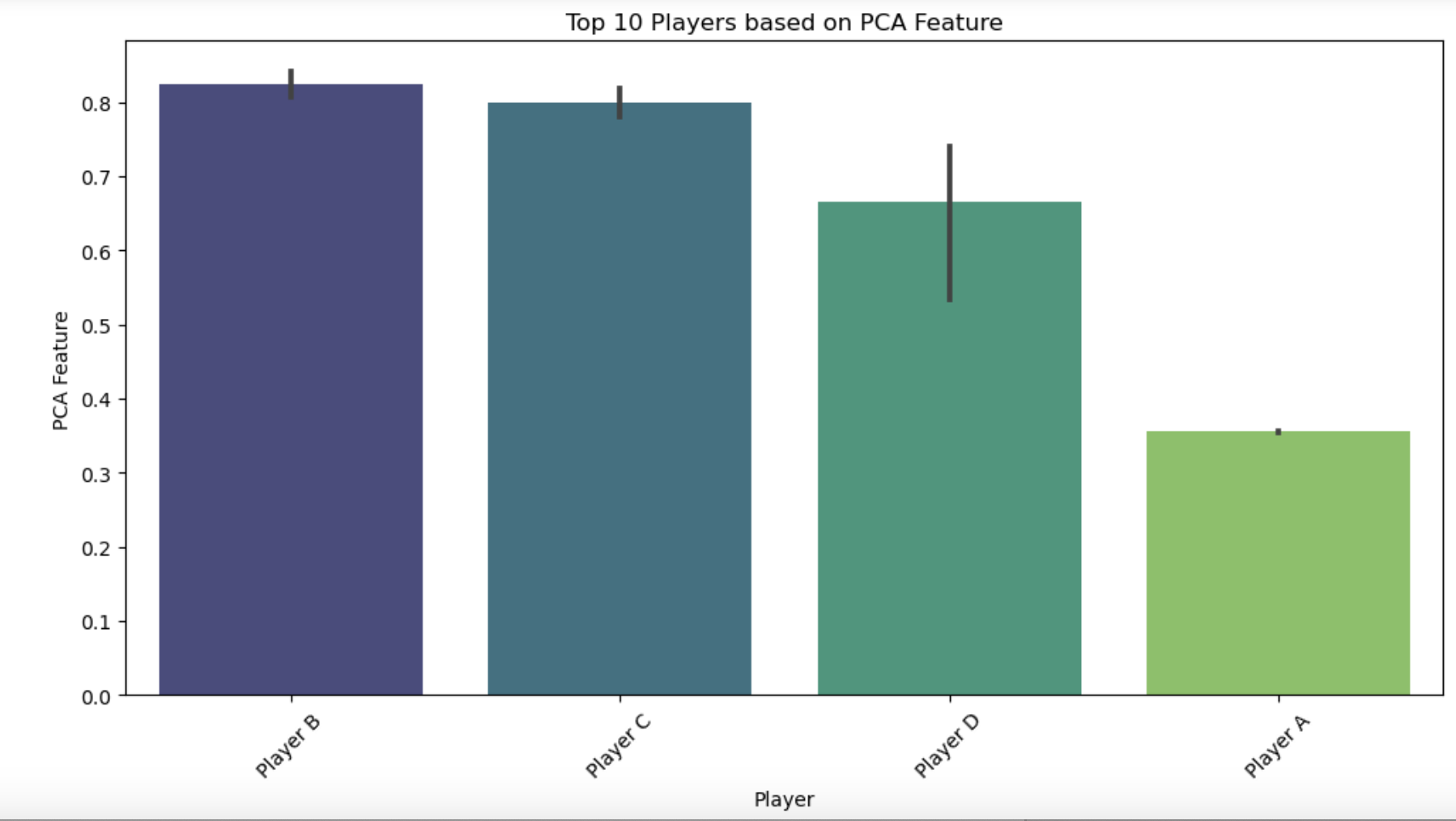
**Using Outlier Detection and Plot using DBSCAN.**

**Q} Find the relationship between PassCompletionRate and Assists using Outlier Detection and plot using Isolation Forest.**

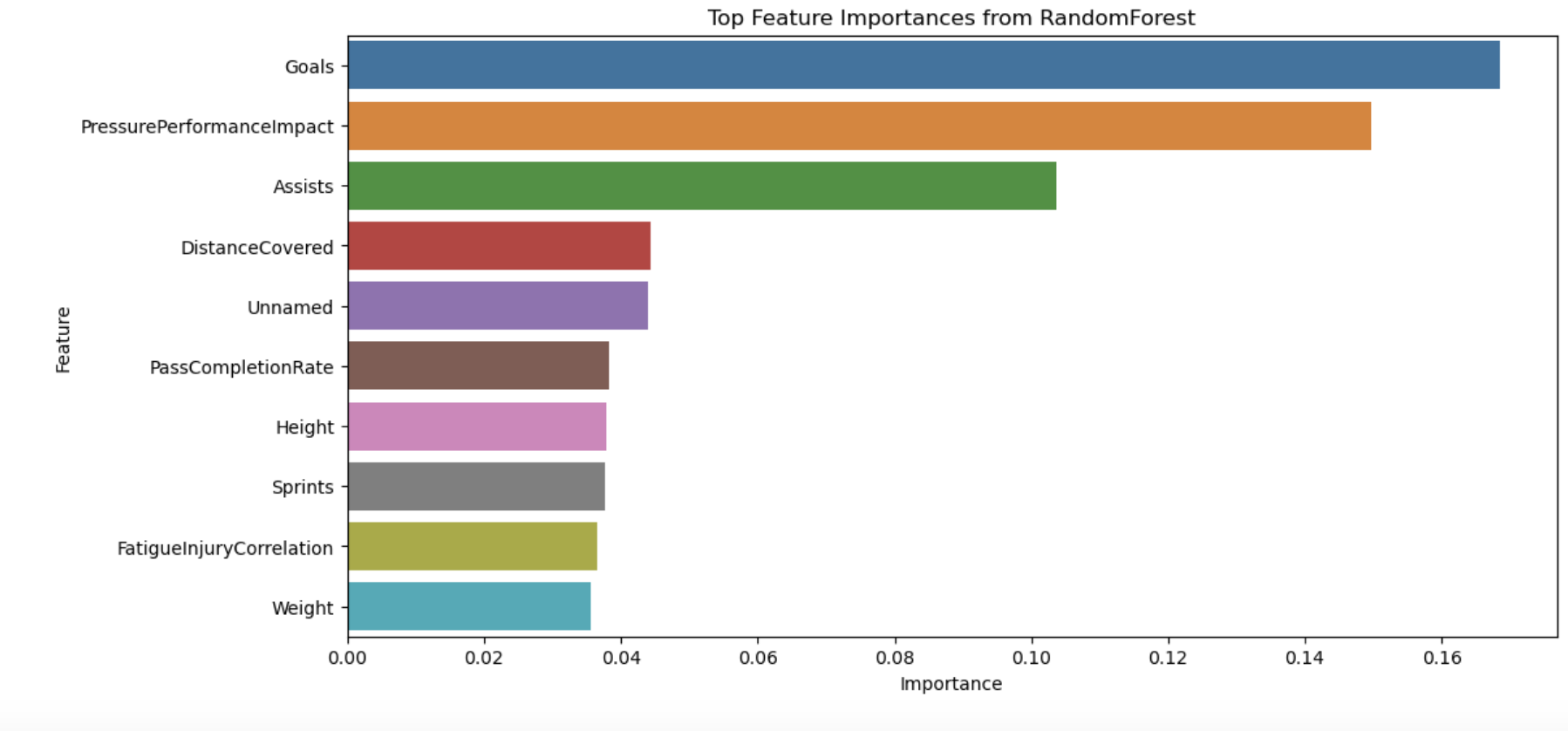
**Q} Regression Analysis model to Create best fit line.**



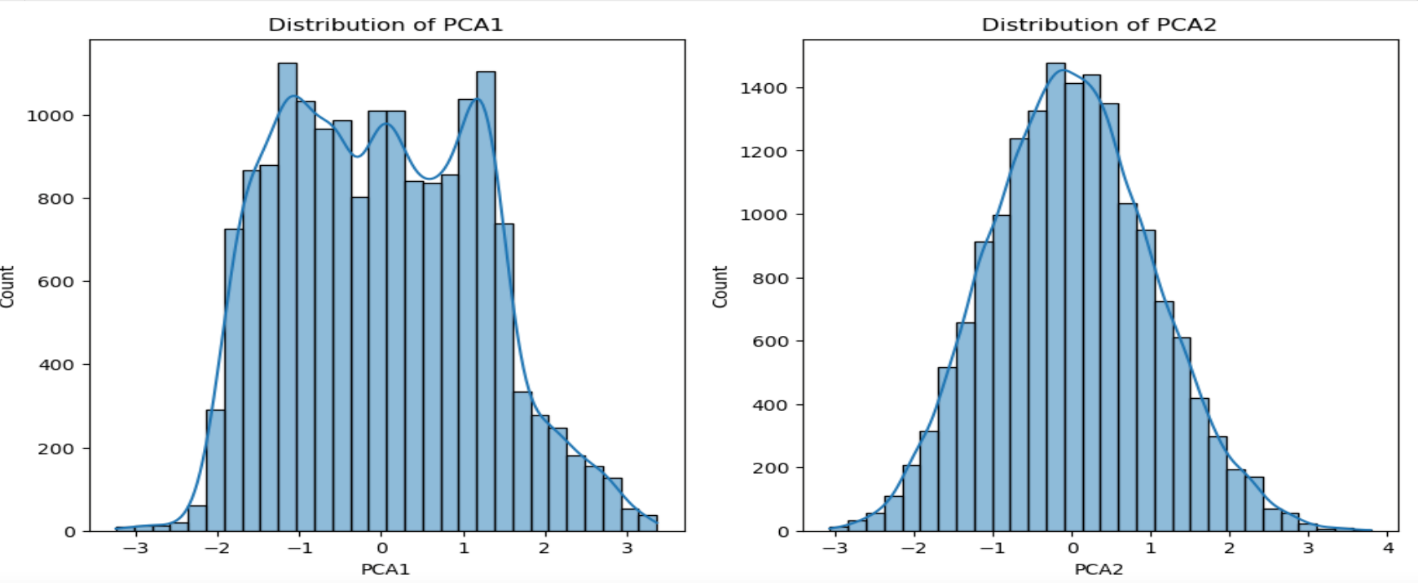
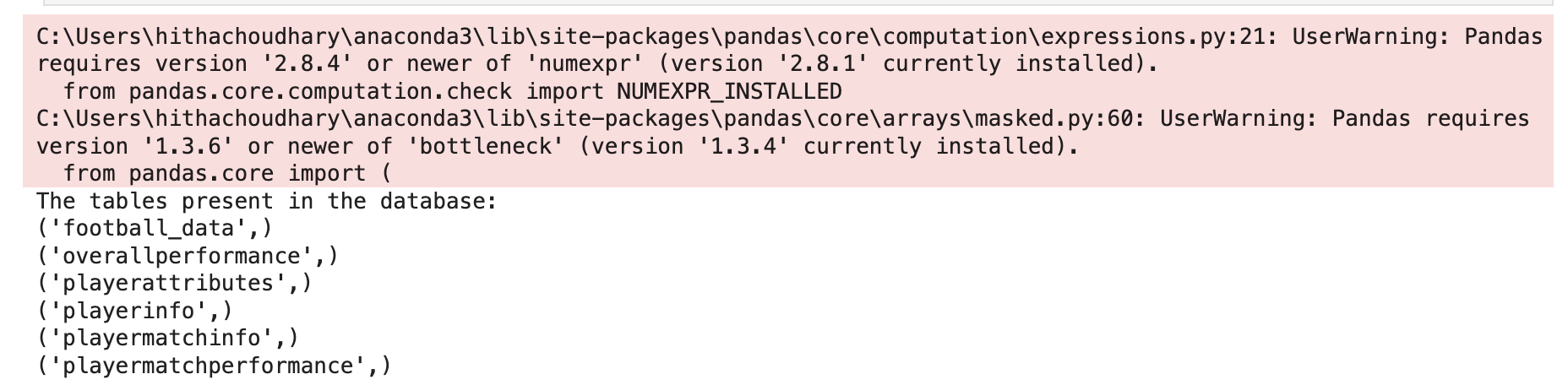
**Q} Creating new features and understand relationship between the created new features, goal scoring efficiency and their overall performance index and understand relationship between the created new feature’s goal scoring efficiency and their overall performance index.**

**Q} Transformations of features created Bar plot for the top 10 players based on the PCA feature uses normalization and dimensionality reduction**

**.**

**Q} Feature Selection using RandomForestRegressor to find 10 important features.**

**Q} Feature Extraction using PCA. Reduce to 2 components for visualization.**

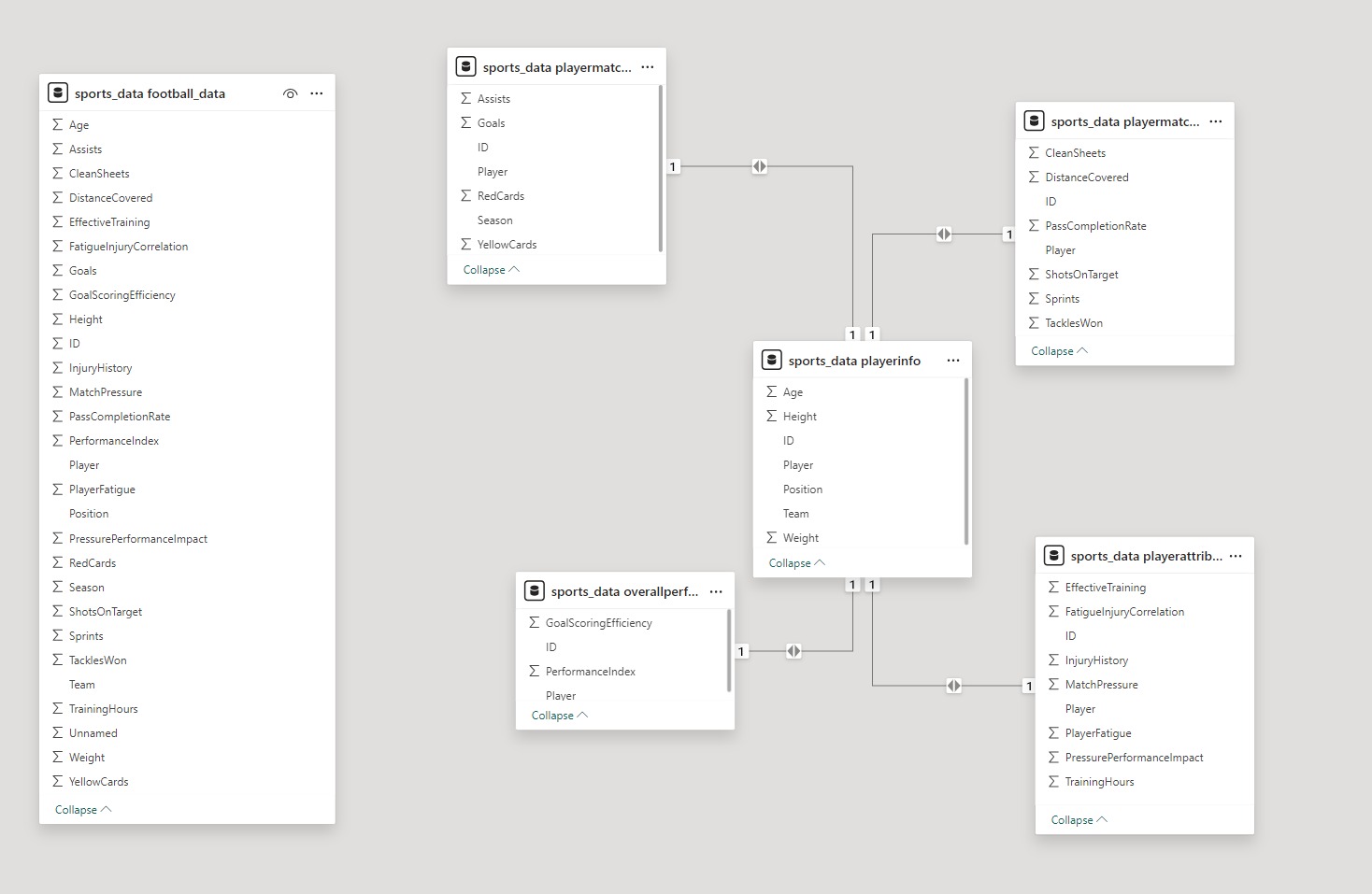
**Q} Creation of tables using MySQL workbench.** 

The tables that are required for the database was created using MySQL Workbench and successfully loaded the data from the csv file to the tables created.

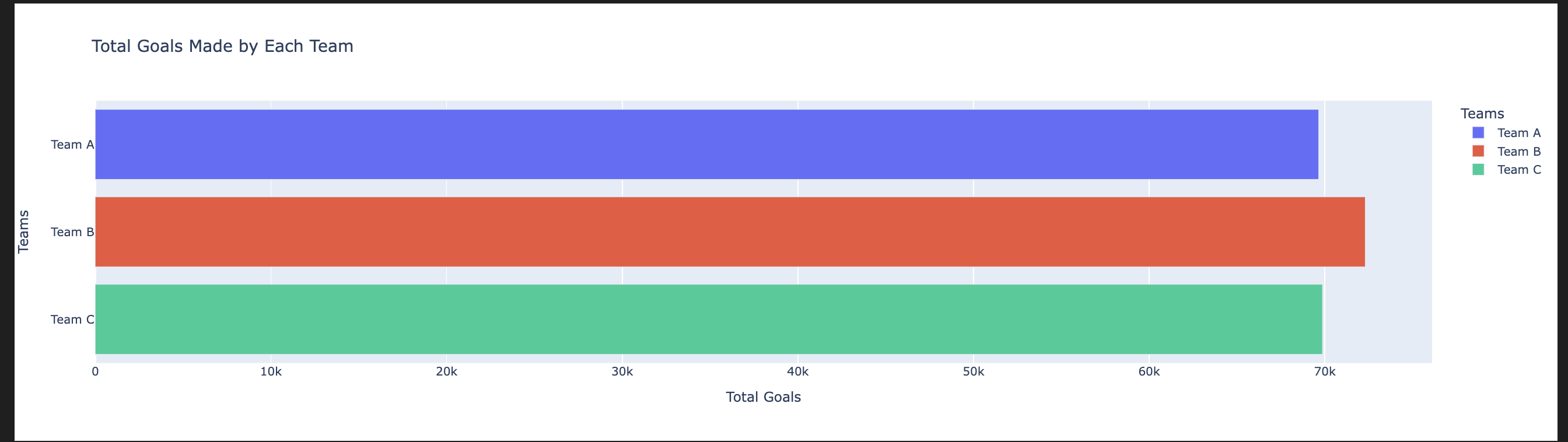
**Q} Advanced SQL features like CTEs-Common Table Expression**

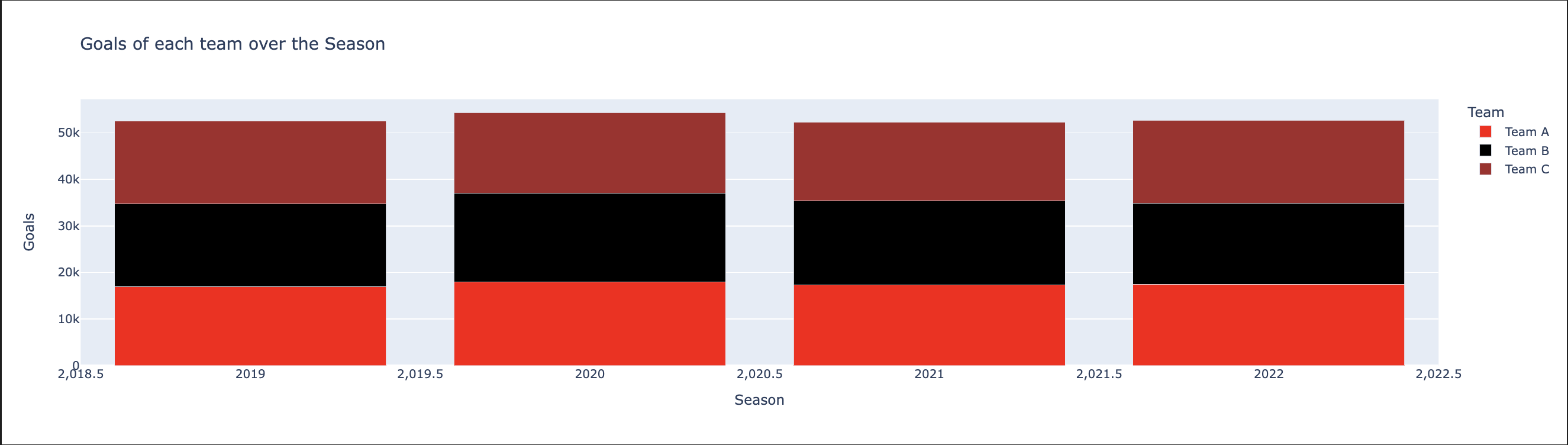
**Q} Complex Analytical Queries based upon:**

* **Average goals per Match**
* **Average Pass Completion Rate**
* **Data Security and Access Control Mechanisms**

****

**Q} Identifying the teams with the highest number of goals**

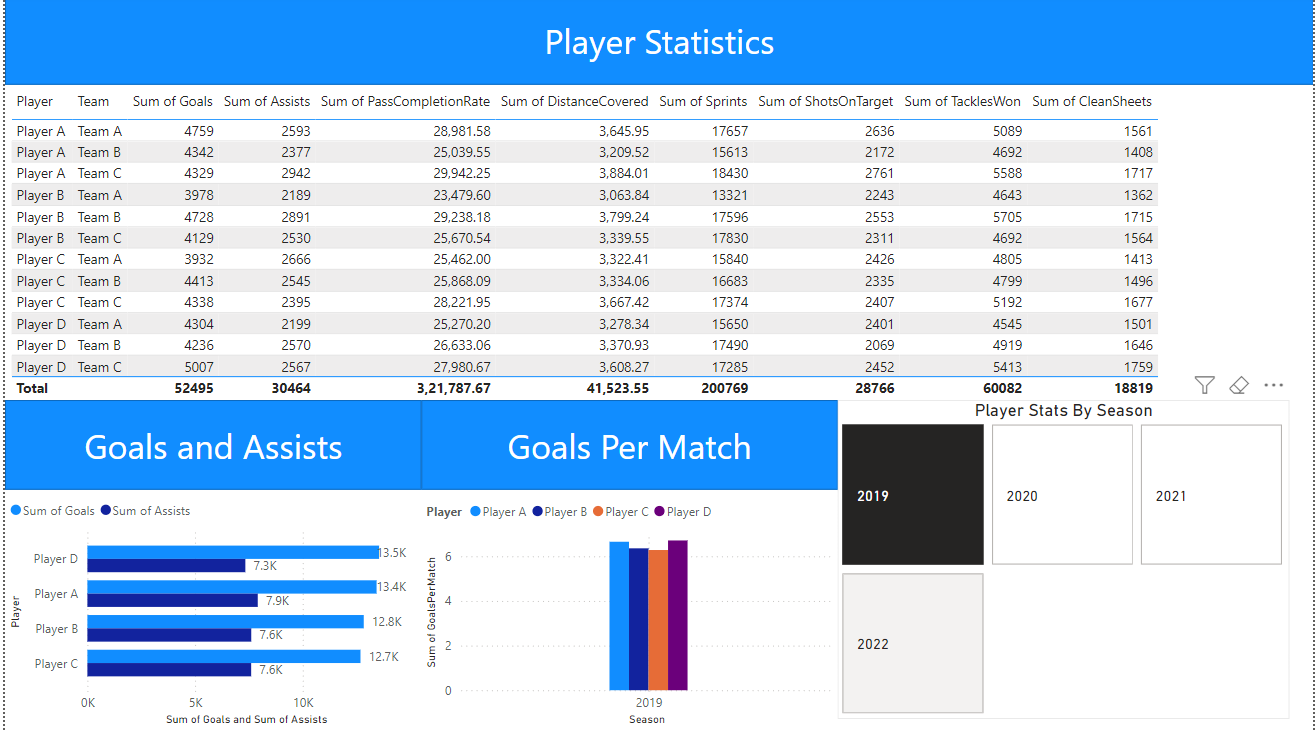
****

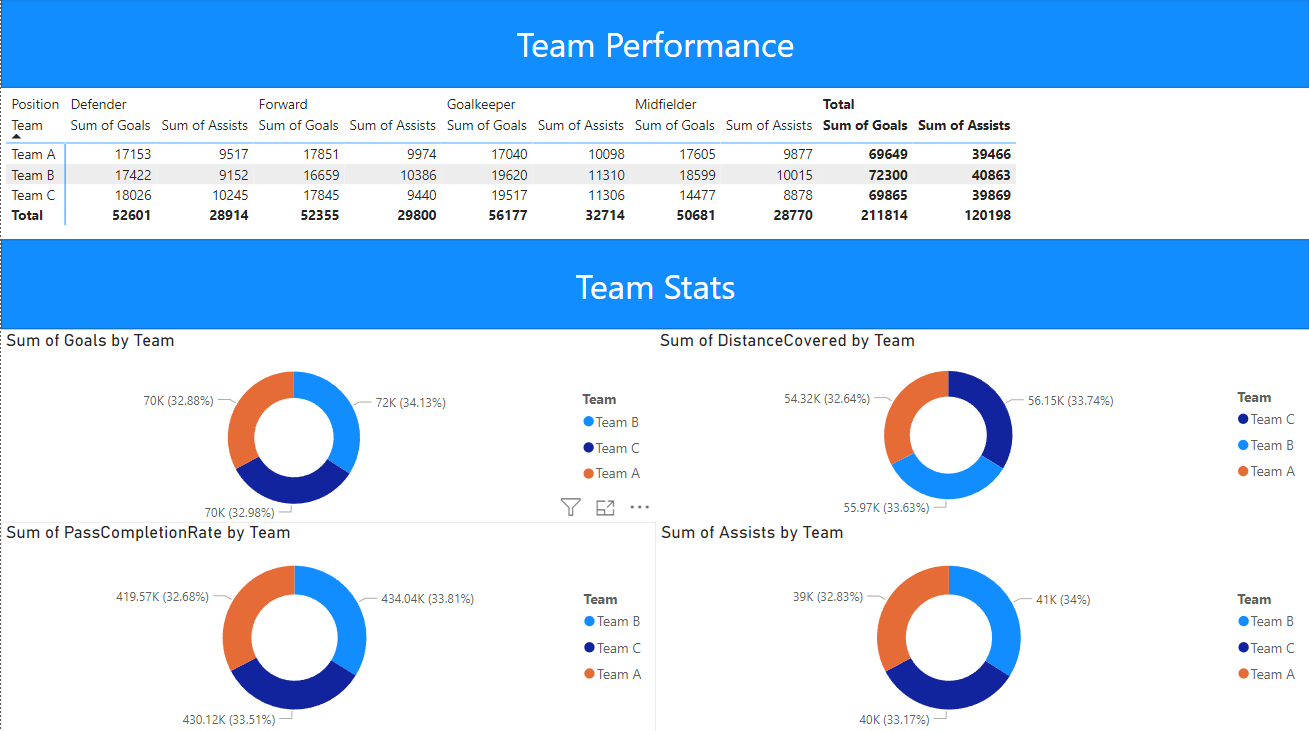
**Q} To visualize the goals made by each team over the season.**

**Q} Analyzing the performance metrics of the overall top scorer .**



**Q} Find Player Statistics filtered by each season year.**



**Q} Analyze team performance for each season.** 

**Conclusion**

The year-long data engineering project at Sports Analytics Inc. has been meticulously designed to tackle the multifaceted challenges inherent in football data analysis. Through a systematic approach, encompassing data cleaning, augmentation, advanced analytics, and visualization, the project aims to transform raw data into actionable insights, driving enhancements in player performance assessment and team tactics optimization.

**Key Achievements:**

1. **Data Quality Improvement:** By employing advanced imputation techniques and anomaly detection methods, we have ensured the accuracy, consistency, and completeness of our dataset. This foundational step is crucial for reliable downstream analysis.
2. **Enhanced Data Robustness:** Through data augmentation, including the use of GANs and integration of external data sources, we have significantly enriched our dataset, providing a more comprehensive basis for analysis.
3. **Efficient Data Ingestion:** The development of a robust data ingestion pipeline, optimized with partitioning and indexing strategies, ensures efficient and reliable data management, facilitating real-time analysis.
4. **Insightful Positional Analysis:** Statistical analysis and visualization of player positions have highlighted key trends and imbalances, offering strategic insights for team management and player recruitment.
5. **Pass Completion and Assists Correlation:** Advanced outlier detection and regression analysis have elucidated the relationship between pass completion rates and assists, identifying key players who excel in these areas.
6. **Complex Data Transformations:** Feature engineering, normalization, and dimensionality reduction have enhanced the dataset, enabling more accurate and insightful analyses.
7. **Scalable Data Warehousing:** The design and implementation of an advanced data warehouse schema ensure efficient storage and querying of transformed data, supporting complex analytical queries and secure data management.
8. **Team Goals Analysis:** Comprehensive analysis of team goals, coupled with time series analysis, has provided deep insights into scoring trends and top performers, informing tactical decisions and strategic planning.
9. **Interactive Reporting and Visualization:** The development of interactive dashboards and visualizations using state-of-the-art tools has facilitated real-time decision-making and provided actionable intelligence to stakeholders.

**Strategic Impact:**

This project underscores the transformative potential of advanced data analytics in professional football. By systematically addressing key problem areas, we have uncovered actionable insights that drive continuous improvement in player performance and team strategies. The integration of advanced machine learning techniques and real-time data analysis has further enhanced the robustness and applicability of our findings, contributing to the ongoing evolution of sports analytics.

**Future Directions:**

While significant progress has been made, the dynamic nature of sports analytics presents ongoing challenges and opportunities. Future efforts should focus on integrating more diverse data sources, enhancing real-time analytical capabilities, and continuously refining machine learning models to keep pace with the evolving landscape of professional football. Additionally, expanding the scope of analysis to include psychological and physiological data could offer even deeper insights into player performance and team dynamics.

In conclusion, this year-long project at Sports Analytics Inc. has laid a strong foundation for leveraging data to enhance football performance and strategy. By turning raw data into strategic assets, we have empowered our clients to achieve greater success on and off the field, positioning Sports Analytics Inc. at the forefront of innovation in sports analytics.

**REFERENCES**

**1.https://www.researchgate.net/publication/283522005\_anomaly\_and\_event\_detection\_for\_unsupervised\_athlete\_performance\_data2.https://www.astera.com/type/blog/build-a-data-pipeline/**

**3.https://community.sap.com/t5/technology-blogs-by-sap/outlier-detection-by-clustering-using-python-machine-learning-client-for/ba-p/13469349.**

**4.https://machinelearningmastery.com/feature-selection-machine-learning-python/**

**5.https://dev.mysql.com/doc/connector-python/en/connector-python-example-connecting.html**

**6.https://dev.mysql.com/doc/connector-python/en/connector-python-example-connecting.html**

**7.https://www.stratascratch.com/blog/chi-square-test-in-python-a-technical-guide/**

**8.https://www.kaggle.com/code/parulpandey/a-guide-to-handling-missing-values-in-python**

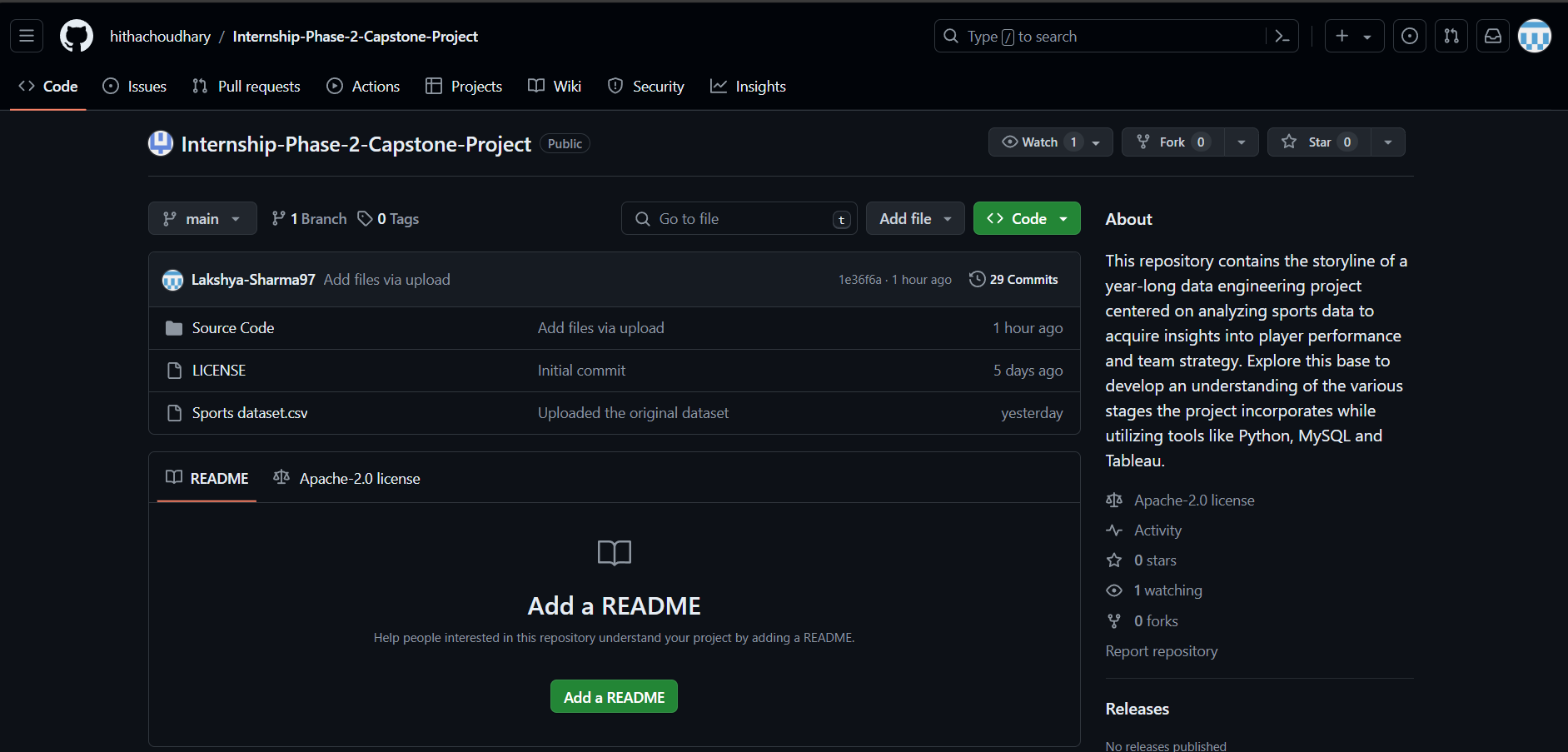
**9.https://www.analyticsvidhya.com/blog/2021/07/anomaly-detection-using-isolation-forest-a-complete-gu**

**10.https://www.geeksforgeeks.org/principal-component-analysis-pca/**

**Appendix - I**

**Source Code**

Github Repository:



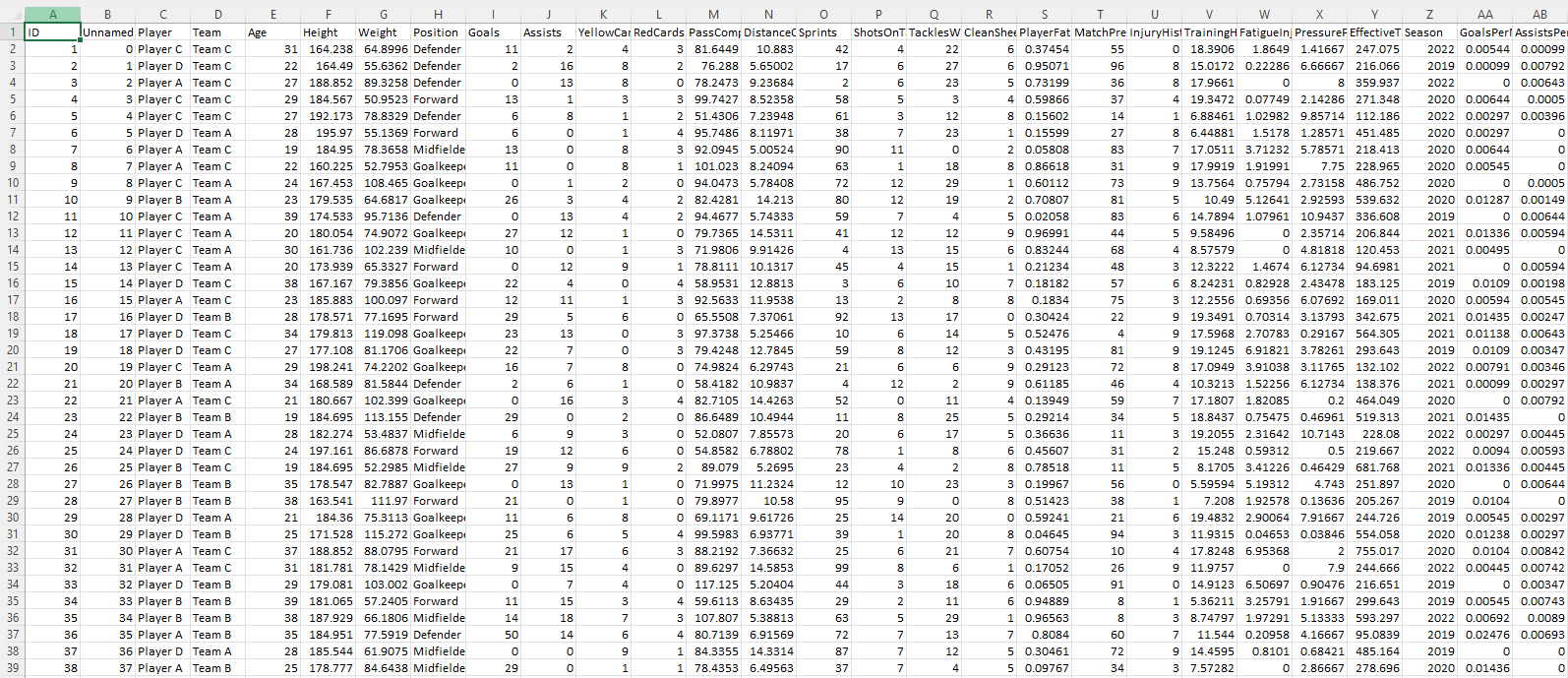
Github Link: <HTTPS://GITHUB.COM/HITHACHOUDHARY/INTERNSHIP-PHASE-2-CAPSTONE-PROJECT>

**Appendix-II**

**Datasheets**

“Sports Dataset” dataset consists of one csv file:

sports\_dataset.csv:



this dataset consists of a total of 20000 rows and 28 columns

**Information Regarding Student(s)**

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Email Id** | **Phone Number** |
| Hitha Choudhary G | hithachoudhary5559@gmail.com | 9741629859 |
| K Shreeshanth Gouda | kannarishreeshanthgouda@gmail.com | 8095496999 |
| Lakshya Sharma | lakshyasharma9797@gmail.com | 8894453852 |