JGi JAIN DEEMED-TO-BE UNIVERSITY | FACULTY OF ENGINEERING AND TECHNOLOGY

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
(IV Semester)

A Project Report on

“Beyond the Scoreboard: Insights from Football Data Analysis”

Submitted in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Ayushi Tawari, Mutta Datta Sai Vishnu Mohan, Niranjana J

22BTRAD008, 22BTRAD026, 22BTRAD027

Under the guidance of

Mr. Akash Das
Project Practice Head and Mentor
Futurense Technologies

JAIN DEEMED-TO-BE UNIVERSITY | FACULTY OF ENGINEERING AND TECHNOLOGY

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

This is to certify that the project work titled **“Beyond the Scoreboard: Insights from Football Data Analysis”** is carried out by **Ayushi Tawari (22BTRAD008), Mutta Datta Sai Vishnu Mohan (22BTRAD026), Niranjana J (22BTRAD027),** 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 fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering, during the year **2023-2024**.

Mr. Akash Das

Project Practice Head and Mentor

Date:10-06-2024

Dr. Sathish Kumar D,

Program Head,
Computer Science and Engineering,
School of Computer Science &
Engineering(AI & DE)
Faculty of Engineering & Technology
JAIN (Deemed to-be University)

Date: 10-06-2024

Dr. Geetha G

Director,
School of Computer Science &
Engineering
Faculty of Engineering & Technolog
JAIN (Deemed to-be
University)

Date: 10-06-2024

Name of the Examiner

Signature of Examiner

DECLARATION

We, **Ayushi Tawari (22BTRAD008), Mutta Datta Sai Vishnu Mohan (22BTRAD026), Niranjana J (22BTRAD027),** student of CSE (AI & DE) 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 **"Beyond the Scoreboard: Insights from Football 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** 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.

Name: Ayushi Tawari USN : 22BTRAD008	Signature
Name: Mutta Datta Sai Vishnu Mohan USN :22BTRAD026	Signature
Name: Niranjana J USN : 22BTRAD027	Signature

Place : Bangalore
Date : 10-06-2024

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, I take this opportunity to express my sincere gratitude to Faculty of Engineering & Technology, JAIN (Deemed to-be University) for providing me with a great opportunity to pursue my Bachelors 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 honorable* **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 endeavors.*

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 D, Program Head, 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***,* **Project Practice Head and Mentor***,* **AVP and Project Manager at 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 Students

ABSTRACT

This report delves into a comprehensive, data engineering project aimed at harnessing the power of sports data to drive informed decision-making. The project encompasses a wide range of tasks, from meticulous data cleaning and augmentation to the design and implementation of efficient data pipelines. Advanced data transformation techniques are employed to enrich the dataset, followed by the creation of a robust data warehouse optimized for efficient storage and retrieval. The project culminates in the development of interactive dashboards and visualizations, providing a dynamic platform for analyzing player performance, team strategies, and potential areas for improvement. To achieve these objectives, a combination of powerful tools and techniques are utilized, including Python, pandas, SQL, and various data visualization libraries. Advanced methodologies like data imputation, statistical analysis, regression modeling, outlier detection, feature engineering, and machine learning are strategically employed to extract meaningful insights from the data. The project's success is measured by the completeness and accuracy of the resulting dataset, the efficiency of data pipelines, the effectiveness of data transformations, the relevance of the analysis, and the clarity and impact of the visualizations. This report provides a detailed account of the project, including the challenges faced, the methodologies adopted, and the significant outcomes achieved. It offers a comprehensive overview of how data engineering can be leveraged to transform raw data into actionable insights, empowering stakeholders to make data-driven decisions that impact the future of sports.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1. Introduction

1.1 Background and Motivation

Analyzing the football dataset can offer valuable insights into various aspects of the game, spanning from player performance metrics, physical attributes, training data, and even psychological factors.

For Teams and Coaches:

- Improve Player Performance: By analyzing the data, coaches can identify areas where individual players or the entire team can improve.

- Scouting and Recruitment: Analysing data from a wide range of players can help identify talented prospects who might excel in a specific team's system.

- Injury Prevention: Analyzing training data and injury history can help identify players at higher risk of injury and develop preventative measures.

- Tactics and Strategy: By looking at how different formations and tactics affect various performance metrics, coaches can make data-driven decisions about their approach to games.

For Data Analysts and Researchers:

- Understanding Player Performance: The data can be used to develop new models and metrics for evaluating player performance, going beyond traditional statistics like goals and assists.

- The Mental Aspect of the Game: Analyzing psychological factors like "MatchPressure" and "PressurePerformanceImpact" can shed light on the mental aspects of the sport and how they influence decision-making and performance.

1.2 Overall Objective

1. **Player Performance Metrics:** Analyzing player performance metrics involves examining key statistical indicators that reflect a player's contribution to the game.
   These metrics can include:

   - Goals and Assists: Quantifying the offensive impact of players.
   - Pass Completion Rate: Measuring the accuracy and effectiveness of a player's passing.
   - Shots on Target: Indicating a player's shooting accuracy and goal-scoring potential.
   - Tackles Won: Reflecting a player's defensive capabilities and effectiveness in regaining possession.

2. **Physical Attributes:** Physical attributes are crucial in determining a player's suitability for specific roles and their overall potential.
   Key physical attributes include:

   - Height and Weight: Affecting a player's physical presence and endurance.
   - Stamina and Endurance: Crucial for maintaining high performance throughout the match.

3. **Training Data:** Training data provides insights into a player's preparation and readiness for matches.
   Key aspects of training data include:

   - Training Hours: The amount of time dedicated to training sessions.

- Training Intensity: The level of effort and workload during training.

4. **Psychological Factors:** Psychological factors play a significant role in a player's performance and overall well-being.
   Important psychological aspects include:

   - Player Fatigue: Measuring mental and physical exhaustion levels, which can affect decision-making and performance.

   - Match Pressure: Assessing how players cope with high-pressure situations during important games.

   - Injury History and Recovery: Understanding the psychological impact of injuries and the mental resilience required for recovery.

By analyzing these different aspects, the dataset can reveal valuable information about the game and the teams who compete.

1.3 Delimitation of research

- Temporal Scope: The analysis may focus on a specific range of years due to data availability or relevance. The dataset contains information about the players and their participation from 2019 to 2022.

- Scope of Variables: The analysis may focus on specific variables within the dataset, while excluding other variables that are not directly relevant to the research objectives.

- Data Quality: Delimitations may be imposed based on the quality and reliability of the dataset. with a high degree of missing values or inconsistencies that could compromise the validity of results.

- Research Objectives: Delimitations are often defined by the specific research questions or objectives of the study. The analysis may focus on addressing specific research questions while omitting broader or tangential topics.

1.4 Benefits of research

- Enhanced Player Performance: Data analysis empowers coaches to identify a player's strengths and weaknesses with a level of precision that goes beyond simple observation. This allows for targeted training programs that address specific areas for improvement, maximizing a player's potential and propelling the team towards greater success.

- Smarter Scouting and Recruitment: No longer confined to traditional scouting methods, teams can leverage data to unearth hidden gems. By analyzing vast datasets encompassing players from various leagues and age groups, analysts can identify talent that might have flown under the radar, giving smaller clubs a fighting chance against giants with bigger budgets. Data analysis also helps refine transfer strategies by providing a more objective valuation of players based on their performance metrics and potential for growth.

- Data-Driven Tactics and Strategies: Gone are the days of relying solely on intuition. Data analysis empowers coaches to make informed decisions about formations and tactics. By analyzing past performances and how different strategies affected various metrics, coaches can tailor their game plan to exploit an opponent's weaknesses and maximize their own team's strengths. Advanced Player Evaluation: The analysis extends beyond traditional statistics like goals and assists. By incorporating metrics like passing completion rate, distance covered, and tackles won, analysts can create a more holistic picture of a player's contribution to the team. This allows for a fairer and more nuanced evaluation of player performance.

- Unlocking the Mental Game: The data can shed light on the psychological aspects of the sport. By analyzing factors like "MatchPressure" and "PressurePerformanceImpact," researchers can gain valuable insights into how players handle pressure and how it affects their decision-making and performance. This knowledge can be used to develop training programs that help players build mental resilience and perform at their best under pressure.

Chapter 2. Implementation

The dataset contains information about the teams and their players from 2019 to 2022.

2.1 Details of the dataset:

- Rows: 20000
  Columns: 25

Breakdown of the columns:

- Unnamed: This column might be a temporary identifier or player ID.
- Player: This column contains player code.
- Team: This column indicates the team each player belongs to.
- Age: This is the age of each player.
- Height: This represents the height of each player.
- Weight: This indicates the weight of each player.
- Position: This specifies the position a player typically plays on the field (Defender, Midfielder, Forward, Goalkeeper).
- Goals: This is the total number of goals scored by each player.
- Assists: This represents the number of passes that led directly to a goal by a teammate.
- YellowCards: This indicates the number of yellow cards received by each player for minor rule infringements.
- RedCards: This represents the number of red cards received by each player for serious rule infringements or a second yellow card.
- PassCompletionRate: This is the percentage of passes successfully completed by each player.
- DistanceCovered: This indicates the total distance covered by each player during a match.
- Sprints: This represents the number of short, high-intensity bursts performed by each player.
- ShotsOnTarget: This is the number of shots taken by each player that were directed towards the goal.
- TacklesWon: This represents the number of successful tackles made by each player to win possession of the ball.

- CleanSheets: This indicates the number of matches where a goalkeeper did not concede any goals.
- PlayerFatigue: This represents the level of fatigue experienced by each player.
- MatchPressure: This indicates the level of pressure faced by each player during a match.
- InjuryHistory: This might contain information about past injuries or the player's injury risk.
- TrainingHours: This represents the number of hours each player dedicated to training.
- FatigueInjuryCorrelation: This could be a score indicating the relationship between fatigue and injury risk for each player.
- PressurePerformanceImpact: This might be a calculated metric representing the impact of match pressure on a player's performance.
- EffectiveTraining: This could be a calculated score indicating the effectiveness of each player's training regimen.
- Season: This identifies the season in which the data was collected.

Chapter 3. Tasks

3.1 Data Cleaning and Augmentation:

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.

1. Duplicate Removal: We identified and removed any duplicate records present in the dataset. Duplicate records could skew analysis results and inflate counts, leading to inaccurate conclusions. By eliminating duplicates, we ensure that each observation in the dataset is unique. The dataset contained 3372 duplicate records.
   Rows: 16628
   Columns: 25

2. Null Value Imputation: Null values have been imputed using the mean of the respective column.

   Null values:
   - Height: 3864
   - Weight: 3950
   - Goals: 3939
   - Assists: 3935
   - PassCompletionRate: 2410
   - PressurePerformanceImpact: 1711
   - EffectiveTraining: 3228

3. Identify and address inconsistencies: Inconsistences such as the data type of the columns, negative value check were identified and addressed.

4. Outlier Detection: Outliers were identified in numerical columns such as Height, Weight, FatigueInjuryCorrelation, PressurePerformanceImpact, EffectiveTraining, Goals and were addressed IQR method. The values in the column were clipped between a upper and a lower bound.

3.2 Position Analysis:

Problem Statement 2: Analyze 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.

The four positions are : Goalkeeper, Forward, Midfielder, Defender.
The total number of players are 16628, with the highest number in Goalkeeper position i.e. 4398 and the lowest number in Midfielder position i.e. 3962.

Position

- Goalkeeper: 4398
- Forward: 4205
- Defender: 4063
- Midfielder: 3962

A chi-square test was conducted that shows distribution of players across positions is significantly different from a uniform distribution. Expected count for uniform distribution per position is 4157.

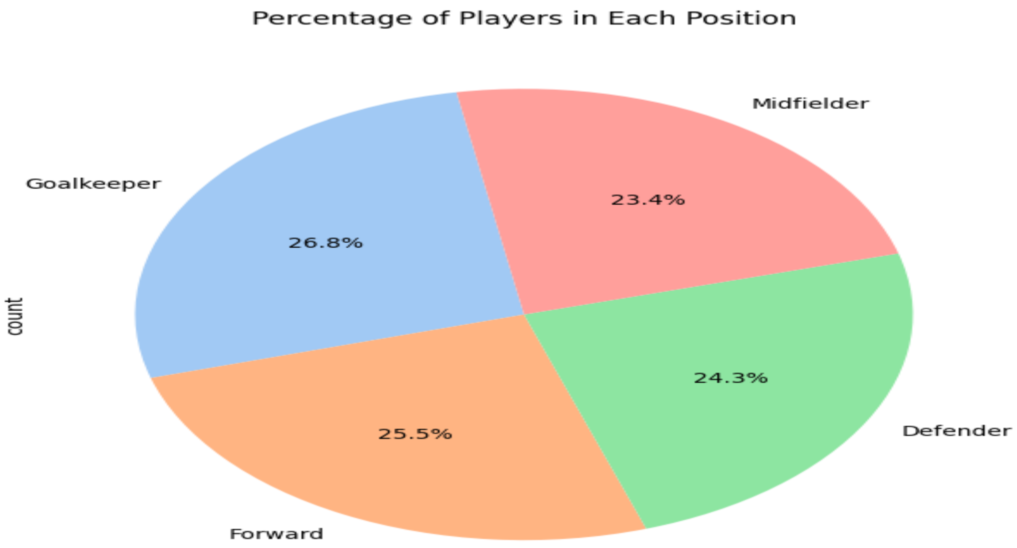


Fig 3.2(a) – Percentage of players in each position

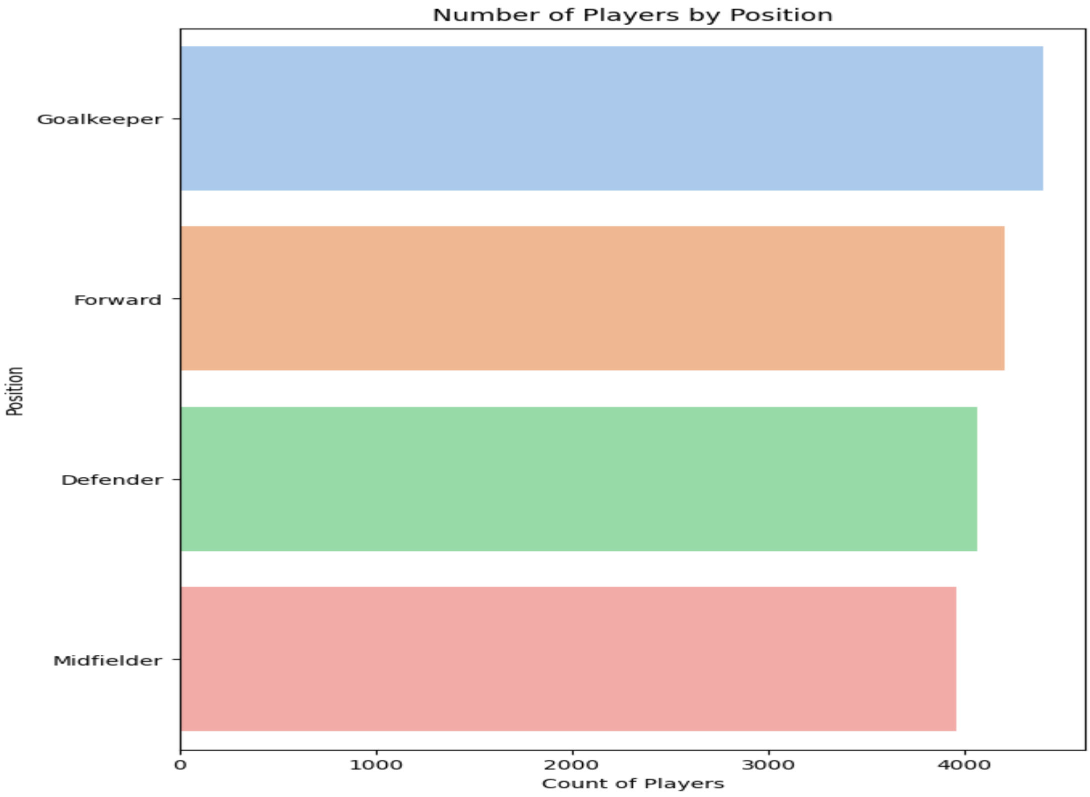


Fig 3.2(b) – Number of players by position

3.3 Data Ingestion Strategies:

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.

It is essential to transfer information from its source to a useable format quickly in the continuously expanding world of data. Here, we provide a programmatic procedure called a data ingestion pipeline that is intended to automate the extraction, transformation, and loading (ETL) of data. This pipeline maximizes query performance, storage usage, and processing efficiency by utilizing strategies including incremental loading, data splitting, and indexing. To further guarantee pipeline dependability and offer insights into its condition, logging and monitoring features are added. In order to facilitate smooth data flow, this article explores the design and implementation aspects of the data intake pipeline using Python, pandas modules, and SQL.

Design Overview:

The process of transferring data points from their original sources into a central place is referred to as data ingestion. Pipelines for ingesting data represent the logic and technology that make this operation possible. They serve as the links between data repositories, such as databases and data lakes, and data sources. The three main components of data ingestion are source that provides the information, the stages of processing that occur in between data sources and destinations, and the destinations the data end up prior to further transformations.

The main components of the data ingestion pipeline process are as follows:

- Data source identification & extraction:
    - ❖ It is the initial step in any data ingestion pipeline. It involves understanding and pinpointing the specific locations and formats where your data resides.
    - ❖ It includes defining the data needs based business requirements and type of data and extracting the required data.
    - ❖ For this project, the sports dataset is provided beforehand hence further sources

identification and extraction processes are not required.
  - ❖ Through this process, data is retrieved from various sources using different techniques and tools.

- Data transformation: Through the process of data transformation, the following steps are taken care of:
  - ❖ Cleaning: It involves removing the duplicates, handling missing values etc.
  - ❖ Formatting: It refers to converting data into a consistent format.
  - ❖ Imputing: It includes estimating and filling in missing values within a dataset.
  - ❖ Outlier analysis: It consists of identifying and dealing with outliers.

- Data validation:
  - ❖ In the data-driven world of today, information integrity must be guaranteed. Data validation serves as a vital precaution, ensuring that data is accurate, consistent, and full before being utilized for analysis or making decisions.
  - ❖ To validate the data, consistency checks are conducted. This ensures that the available data is of the same data type.
  - ❖ After validation, the pre-processed data can be stored in formats that will be suitable for further analysis.

- Data analysis and visualisation
  - ❖ Meaningful insights can be extracted from the data through different analysis techniques involving SQL and visualization approaches using different python libraries such as
    matplotlib, seaborn, plotly etc.

Implementation Details:

1. Define Data Source and Destination:

   - Source: Sports dataset excel sheet including information about different players and their respective 'Team', 'Age', 'Height', 'Weight', 'Position', 'Goals', 'Assists', 'YellowCards', 'RedCards','PassCompletionRate','DistanceCovered', 'Sprints', 'ShotsOnTarget', 'TacklesWon','CleanSheets','PlayerFatigue','MatchPressure','InjuryHistory','TrainingHours', 'FatigueInjuryCorrelation', 'PressurePerformanceImpact', 'EffectiveTraining', 'Season'

- Destination: The ingested data will be stored in a data warehouse or cloud storage platforms like Google Cloud Storage.

2. Choose Data Ingestion Tools and Techniques:

- Tools: Python libraries like pandas and numpy etc. are used for data manipulation. SQL is utilized for database interactions. Cloud platforms offer managed services like Cloud Dataflow (Google Cloud), AWS Glue (Amazon Web Services), and Azure Data Factory (Microsoft Azure) that can simplify the process.

- The different techniques include:
  - ❖ Batch Processing: Suitable for large datasets that are ingested periodically (daily, weekly). Tools like Apache Airflow can be used to schedule batch jobs.
  - ❖ Micro-Batching: Processes data in smaller chunks, offering a near real-time feel compared to batch processing.
  - ❖ Stream Processing: Ideal for continuous data streams requiring real-time ingestion (e.g., sensor data, social media feeds). Apache Kafka is a popular stream processing framework.

In this scenario, since an incremental data loading approach is followed, stream processing is preferred over batch processing since stream processing continuously ingests data as it is generated. This aligns well with the concept of incremental loading, as we can process new data updates as soon as they arrive. Mini-batch processing is also a preferable since it involves data is continuously streamed. Instead of processing each data point individually, the stream is divided into small batches. These micro-batches are then processed at regular intervals, allowing for some efficiency gains while maintaining near real-time updates. Hence it is ideal to choose stream processing with micro-batching.

3. Data Extraction, Transformation, and Loading (ETL):

- Extract: Retrieve data from the source using appropriate methods (e.g., database queries, API calls).
- Transform (Optional): Clean and manipulate the data as needed. This might involve handling missing values, formatting inconsistencies, or deriving new features.
- Load: Transfer the transformed data into your designated storage system.

Through the process of ETL, improved data quality, enhanced analytics, simplified data management, streamlined reporting etc. can be ensured.

4. Design and Develop the Pipeline:

The data processing logic can be programmed using python and SQL according to the requirements.

The steps include:

- Connect to Database: Establish a connection between the Python script and the SQL database using a connector library.
- Data Cleaning and Transformation: Leverage pandas functionalities to clean and transform the data as needed (e.g., handle missing values, remove duplicates, create new features).
- Extract Data: Use SQL queries to retrieve the desired data and insights from the database.
- Load Data into database: Use different libraries to import the retrieved data from the SQL database into a pandas DataFrame or vice versa for further manipulation.
- Analysis and Visualization: Use Python libraries like NumPy and Matplotlib for further analysis and visualization of the processed data.
- Error handling and pipeline scheduling: Implement different error handling approaches during data ingestion. It is also important to schedule the pipeline to run at appropriate timings.

5. Monitor and Maintain:

Monitoring the pipeline's performance to ensure smooth operation. Metrics like processing time, data volume, throughput and error rates. Schedule regular maintenance to address potential issues and adapt to changes in the data source or destination.

Key features of the pipeline:

- Incremental loading: This technique focuses on identifying and extracting only the new or updated data since the last successful load. It minimizes redundant processing and wasted resources by focusing on the most recent changes.
- Partitioning: Partitioning refers to the concept of dividing data into smaller, more manageable subsets based on specific criteria. This technique is employed in various contexts, including data storage, database management, and machine learning.
- Indexing: In the realm of data management, indexing plays a critical role in accelerating data retrieval. It's akin to creating a detailed catalog in a library, allowing you to find specific information quickly and efficiently.
- Logging: Logging is the process of recording events and information within a program or system for debugging, monitoring, and analysis.
- Monitoring: Different metrics like processing time, data volume, throughput, error rates etc. can be used for monitoring.

3.4 Pass Completion Rate vs. Assists:

Problem Statement 4: Analyze 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.

Understanding the relationship between pass completion rate and assists can help in evaluating player performance and making informed decisions for training and strategy. Linear regression provides a foundation for predicting assists based on pass completion rate, which can be valuable for game strategy and player development.

Outliers are data points that significantly differ from the majority of the data. They can occur due to variability in the data or errors in data collection. As observed through isolation forest method there are 1661 rows with outliers.

Statistical Tests: Methods like Z-score, modified Z-score, and IQR (Interquartile Range) can detect outliers in a dataset.

Machine Learning Algorithms: Techniques such as Isolation Forest, DBSCAN, and LOF (Local Outlier Factor) can detect outliers in a dataset.

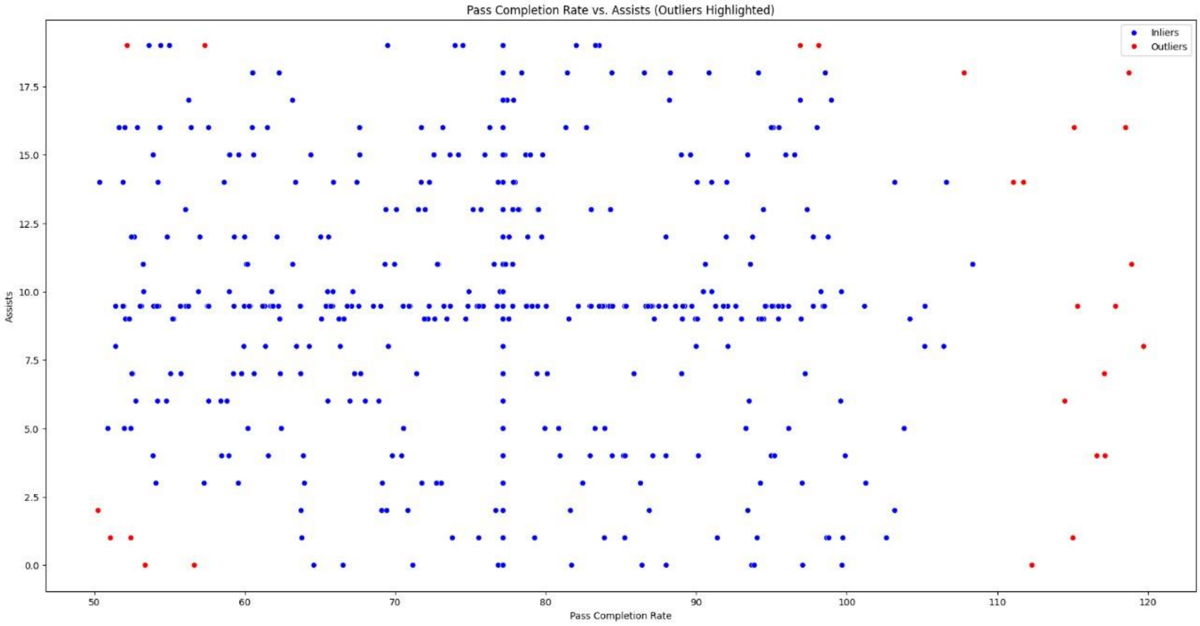


Fig 3.4(a) – Pass Completion Rate vs Assists

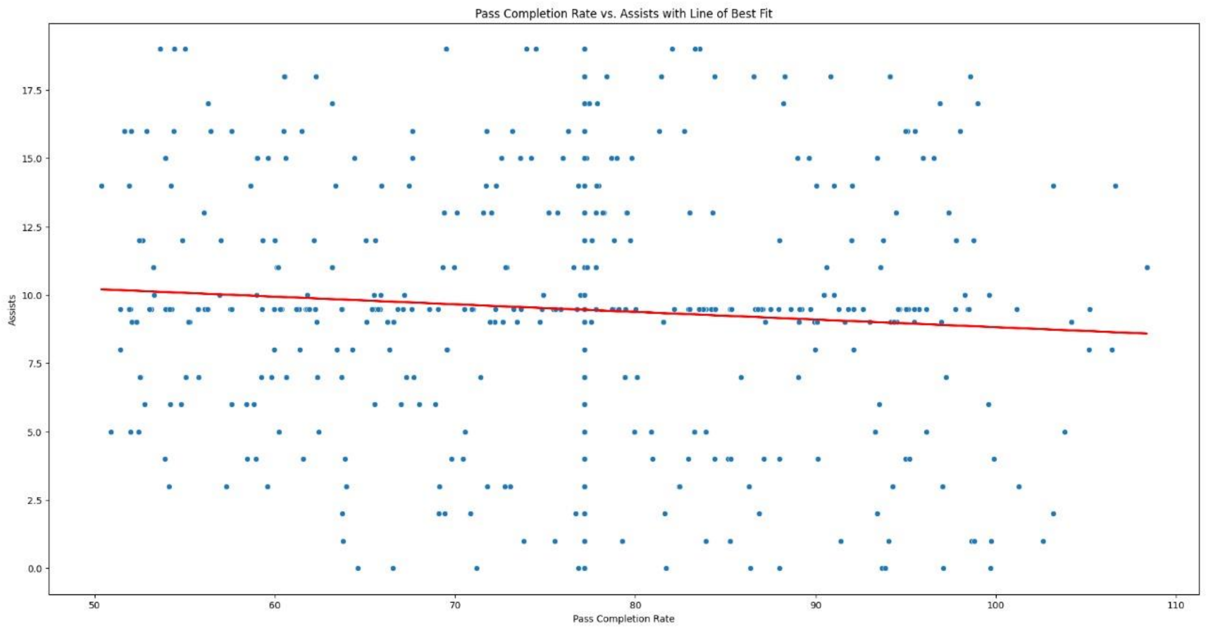


Fig 3.4(a) – Pass Completion Rate vs Assists (Best fit line)

R-squared: 0.006635514692627022

It provides an indication of how well the independent variable explain the variability of the dependent variable. This indicates it is not a perfect (value=1) or imperfect model (value=0)

3.5 Advanced Data Transformations:

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.

- Predicted_Performance_Category

  A new feature named PredictedPerformanceCategory has been introduced to assess player performance based on their training hours and efficiency. This feature categorizes players into different performance tiers by calculating a performance ratio.

  EffectiveTrainingHours=TrainingHours * EffectiveTraining
  Performance Ratio= Goals / EffectiveTrainingHours

AA (Above Average): Players with a high performance ratio (≥ 0.005).

A (Average): Players with a moderate performance ratio (≥ 0.002 but < 0.005).

BA (Below Average): Players with a low performance ratio (< 0.002).

None: Players with no performance ratio, which could be due to missing data or division by zero (where effective training hours are zero).

Coaches and training staff can utilize this feature to identify players who may need additional support or altered training regimes. Players in the "BA" category, for instance, can be given tailored training plans to improve their effectiveness.

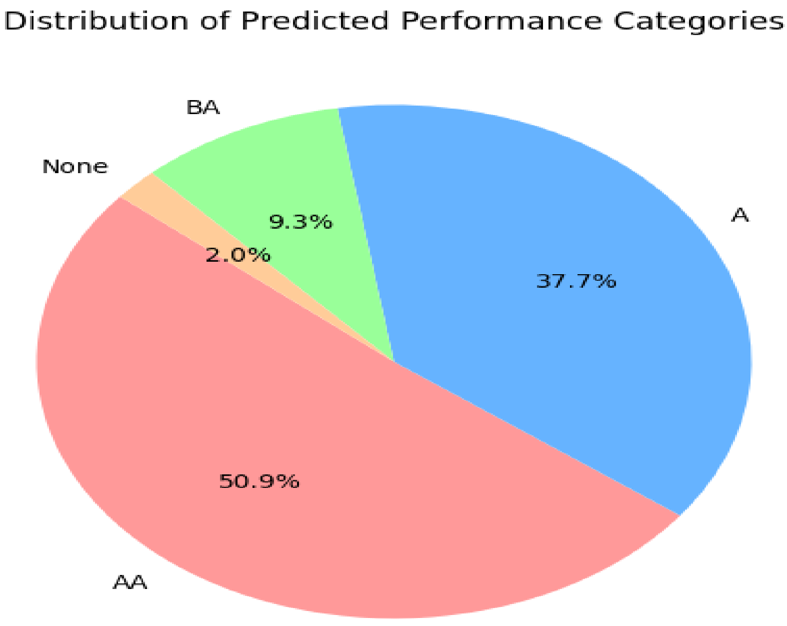


Fig 3.5(a) – Distribution of Predicted Performance Categories

Total number of players is 16628, with most of the players being AboveAverage.

- PredictedInteractionCategory

A new feature, PredictedInteractionCategory, has been introduced to assess whether players are affected by the interaction between fatigue and match pressure. This feature is based on the product of PlayerFatigue and MatchPressure, termed as FatiguePressureInteraction.

FatiguePressureInteraction = PlayerFatigue * MatchPressure

The mean threshold value for this interaction is 24.579950172041613.

Players are categorized as either "Affected" or "Not Affected" based on this threshold:

Affected: Players with a FatiguePressureInteraction >= 24.579950172041613.

Not Affected: Players with a FatiguePressureInteraction < 24.579950172041613.

Identification of players who might be underperforming due to the combined effects of fatigue and match pressure, allowing coaches to manage their workload more effectively.

Provides valuable insights for making strategic decisions about player rotation, rest periods, and match preparation, ensuring optimal performance during games.

Helps in identifying players at risk of injuries due to high fatigue and pressure, allowing preventive measures to be taken to ensure player health and longevity.

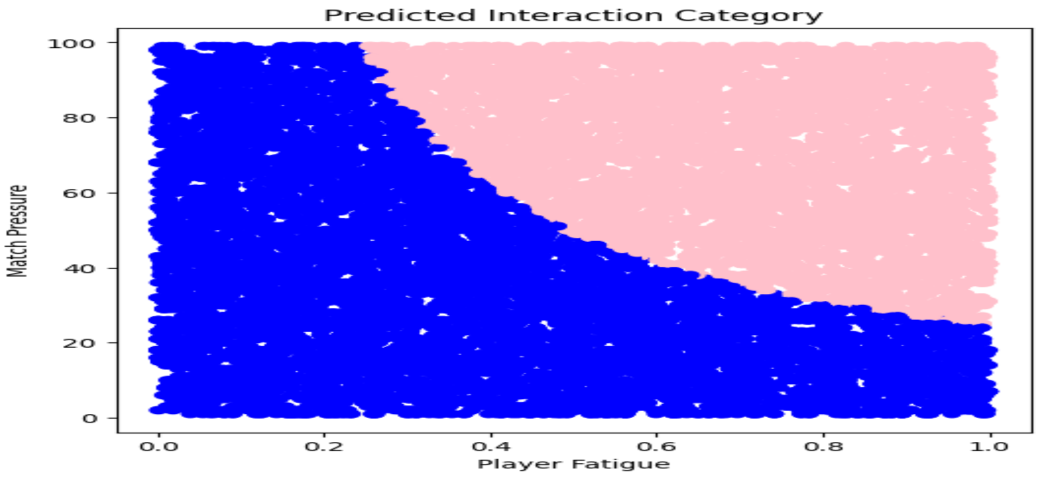


Fig 3.5(b) – Scatter plot of Predicted Interaction Categories

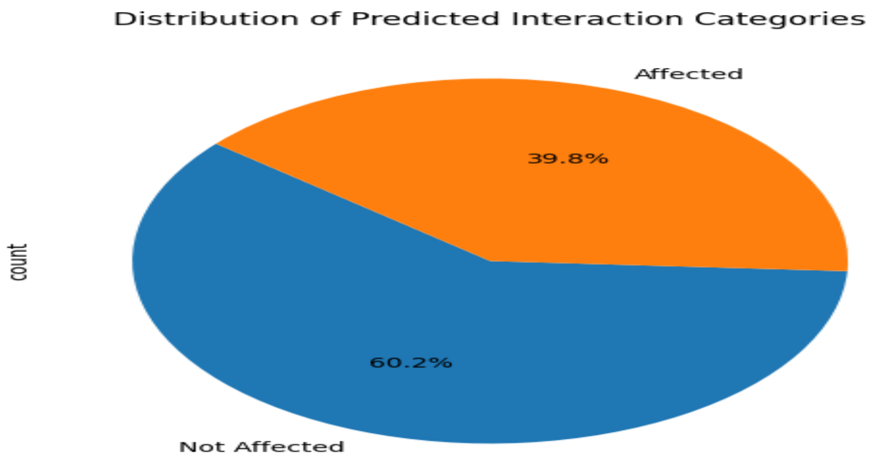


Fig 3.5(c) – Distribution of Predicted Interaction Categories

3.6 Data Warehousing:

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.

Data warehousing is the process of collecting, storing, and managing large volumes of data from different sources into a central repository, called a data warehouse.

Integration: Data warehousing consolidates data from various sources (e.g., databases, spreadsheets, applications) into a single repository, providing a unified view of information.
Consistency: By maintaining a single source of truth, data warehousing ensures data consistency and accuracy across the organization.

Data warehouses are optimized for complex analytical queries, allowing businesses to gain insights from their data quickly and efficiently. Data warehouses store historical data, enabling trend analysis and historical comparisons, which are crucial for strategic planning and forecasting.

1. Define Business Requirements

2. Types of Data Warehouse Schema:
    - Star Schema: Design a star schema with fact and dimension tables. Fact tables store quantitative data (e.g., sales, transactions) and dimension tables store descriptive attributes (e.g., time, product, customer).
    - Snowflake Schema: For normalization, design a snowflake schema, which is a variation of the star schema where dimension tables are normalized into multiple related tables.
    - Fact Tables: Identify the granularity of the fact tables (e.g., daily sales, monthly transactions).
    - Dimension Tables: Identify the dimensions and their attributes (e.g., Date, Product, Customer).

3. Create the Data Warehouse Schema:
    - Use SQL to create the fact and dimension tables.
    - Define primary keys for dimension tables and foreign keys in fact tables to establish relationships.

4. Load Data into the Warehouse:
    - Use SQL scripts or ETL tools to load data into the dimension and fact tables.

5. Implement Advanced SQL Features:
    - Common Table Expressions (CTEs): Use CTEs for simplifying complex queries by breaking them into reusable components.

6. Optimize Data Storage and Query Performance:
    - Indexes: Create indexes on columns frequently used in queries to improve retrieval speed.
    - Partitioning: Partition large tables to improve query performance and manageability.

7. Implement Data Security and Access Control:
    - User Roles and Permissions: Define roles and grant appropriate permissions to ensure secure access to the data warehouse.
    - Data Masking: Apply data masking techniques to protect sensitive data.
    - Encryption: Encrypt sensitive data to ensure its security both at rest and in transit.

8. Test the Data Warehouse:
    - Validate data integrity and accuracy by running test queries.
    - Perform load testing to ensure the warehouse can handle the expected volume of data.

Steps followed:

1. Establishing a connection between python and MySQL.
2. Ingesting data from the dataset into multiple tables of mysql database.

    Tables:
    - Sports_dataset
    - Player_details
    - Player_stats
    - Training
    - Player_performance
3. Implementation of CTEs.
4. Implementing data security and access control mechanisms.

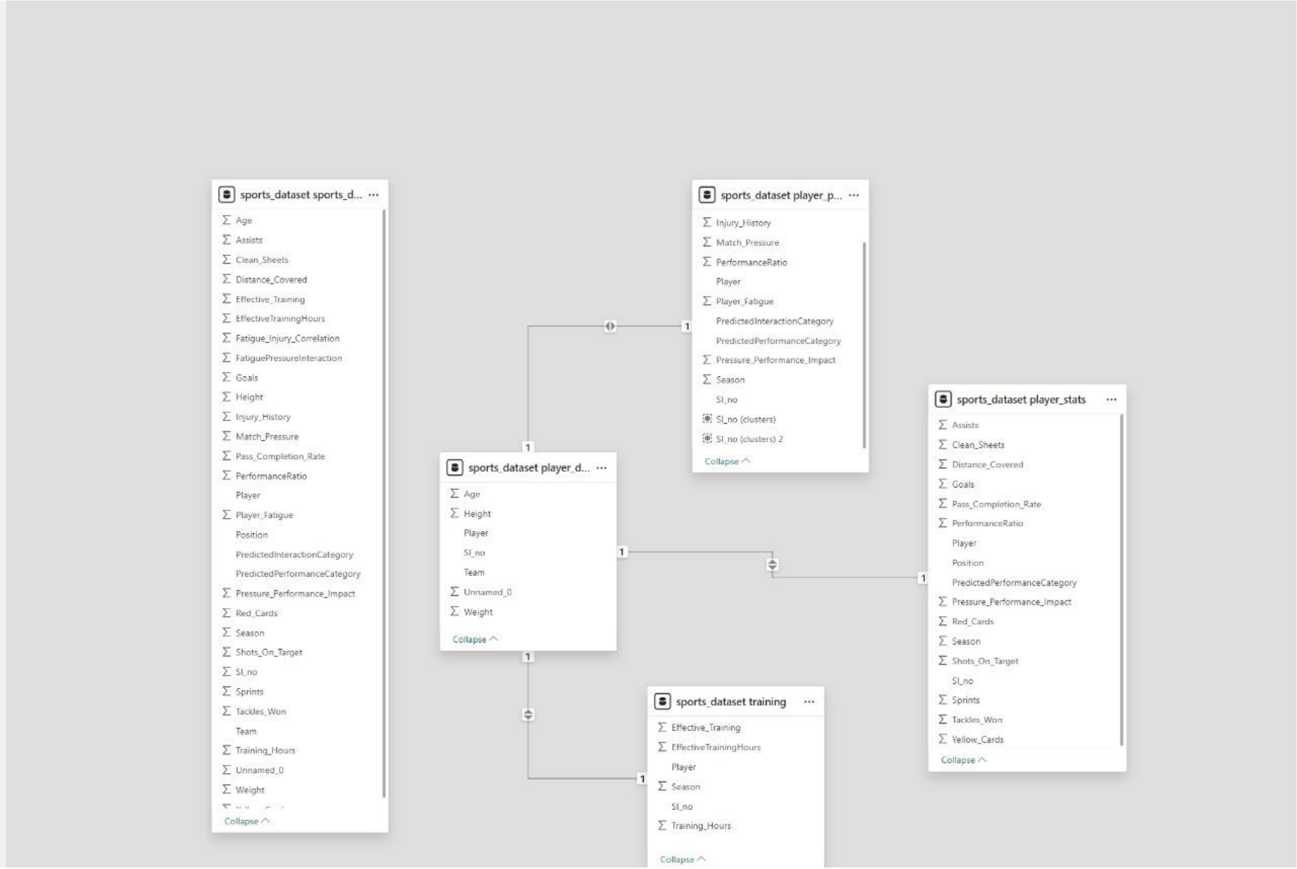


Fig 3.6(a) – Schema Diagram

3.7 Team Goals Analysis:

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 analyze their performance metrics over time.

The team with the highest number of goals is Team B with a total goal score of 108266.

Goals scored by each team:

- Team A: 103243
- Team B: 108266
- Team C: 106814

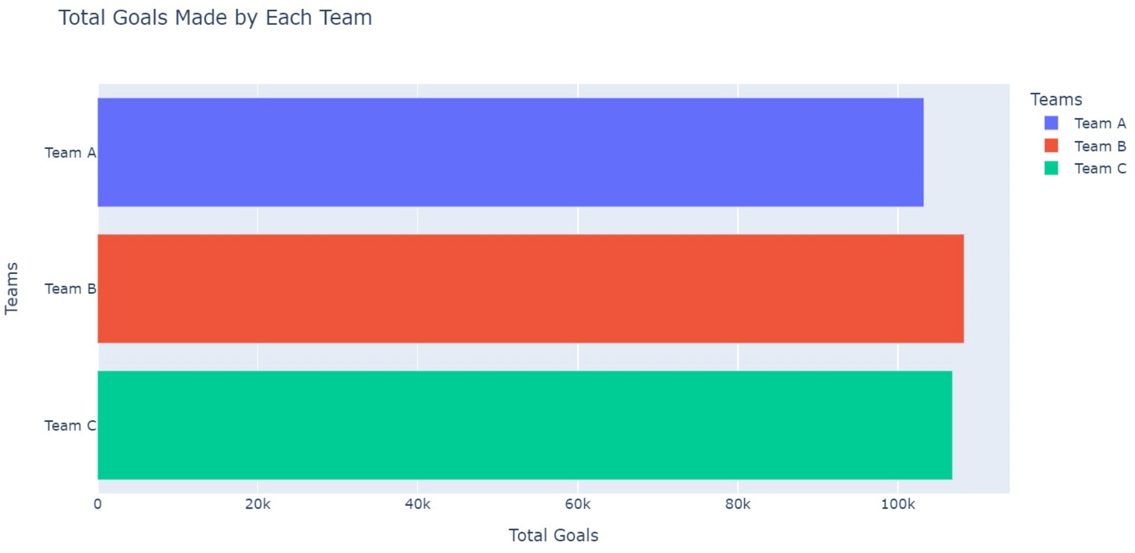
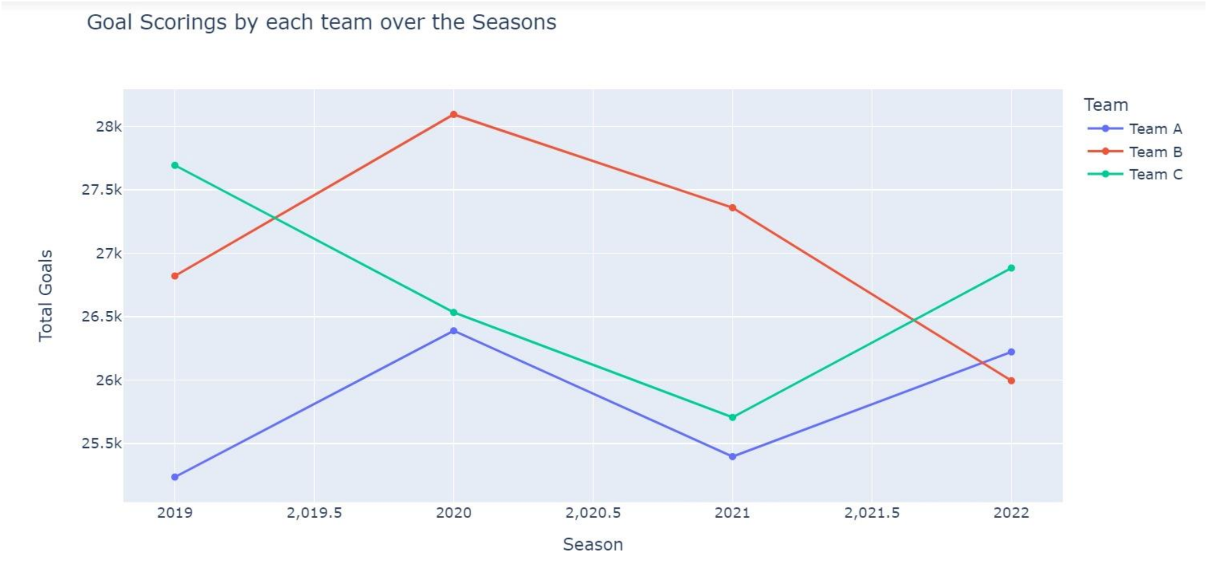


Fig 3.7(a) – Total goals made by each team



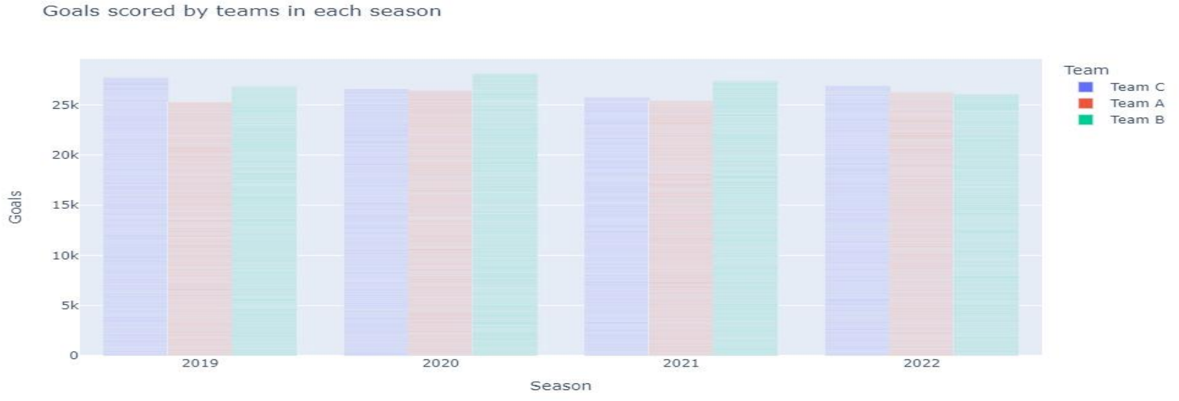


Fig 3.7(b) – Goals scored by teams in all seasons

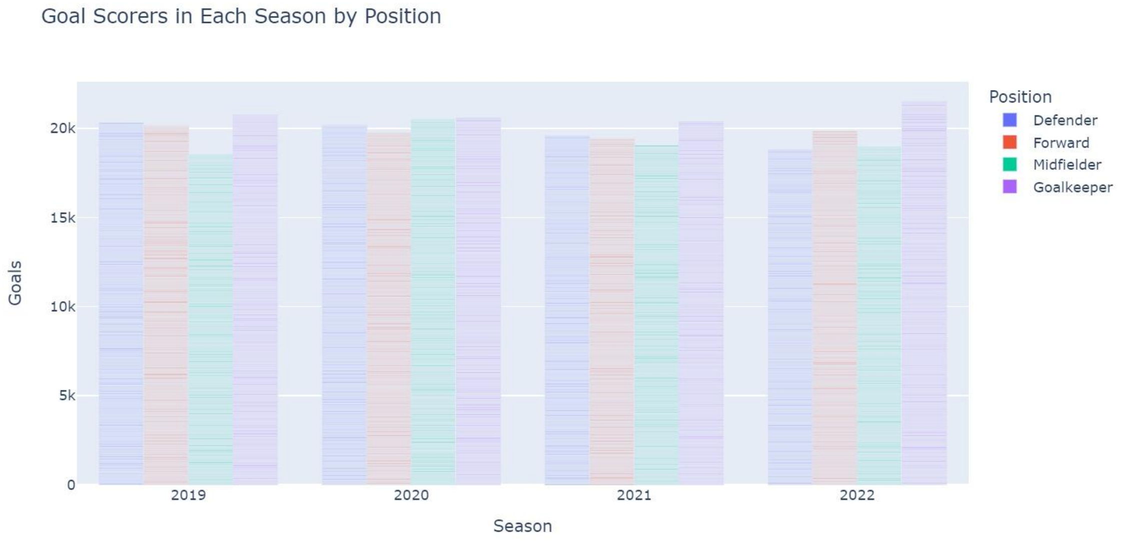


Fig 3.7(c) – Goals scorers in each season by position

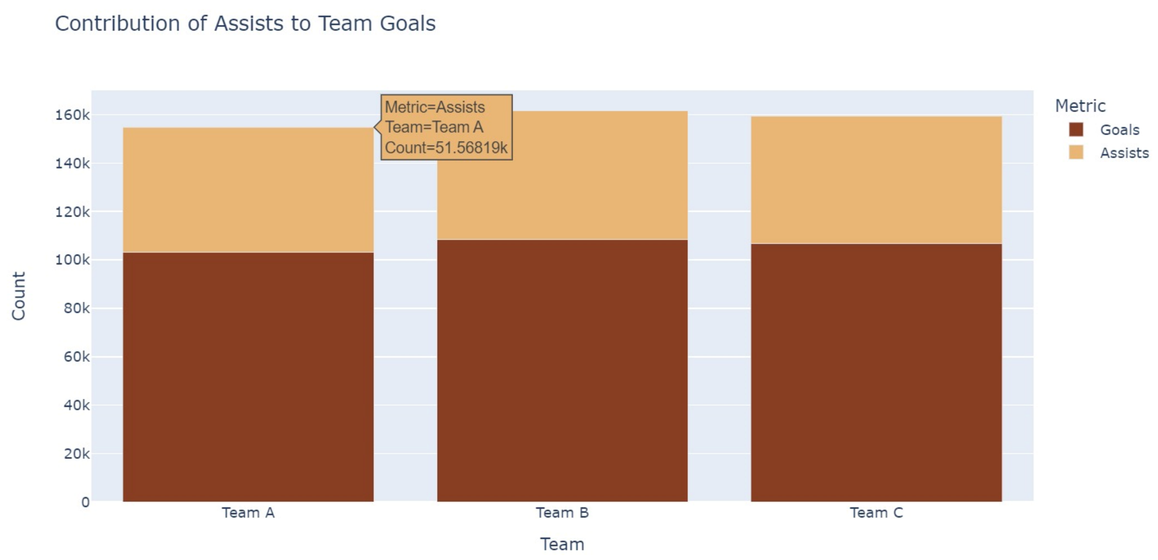
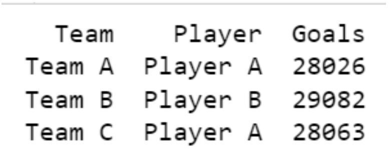


Fig 3.7(d) – Goals and Assists of each team

Each team and their high scorers are as follows:



```
  Team    Player  Goals
Team A  Player A  28026
Team B  Player B  29082
Team C  Player A  28063
```

Team B is team with the highest number of goals and Player B of team B has scored the most number of goals. Analyzing this player's performance statistics provided the following inferences:

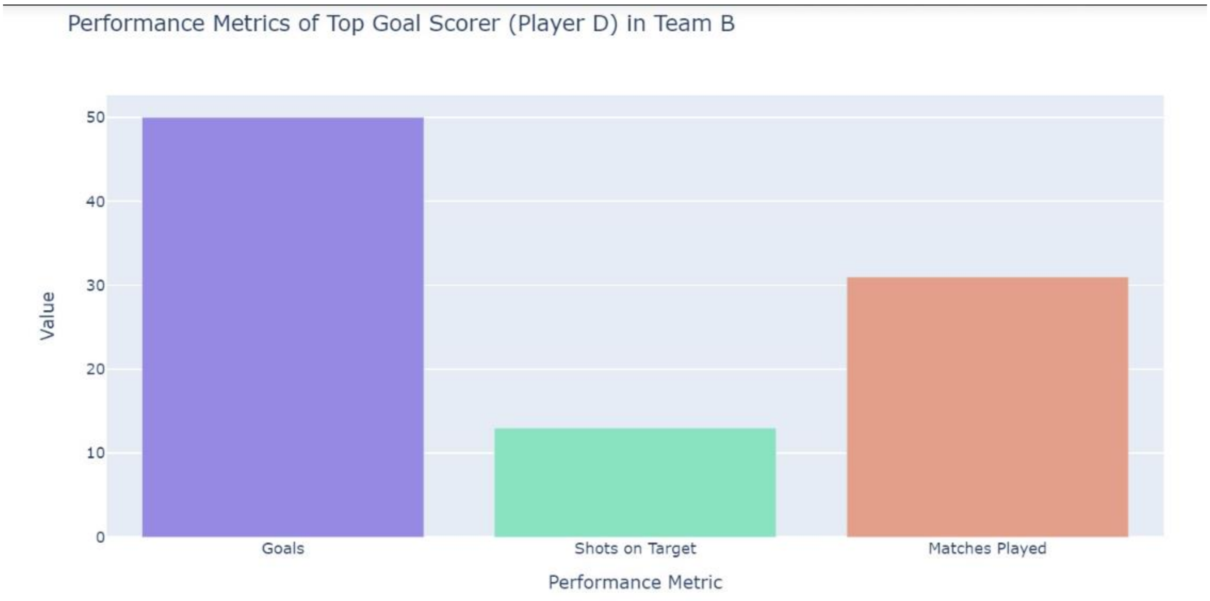


Fig 3.7(e) – Performance metrics of top player of top team

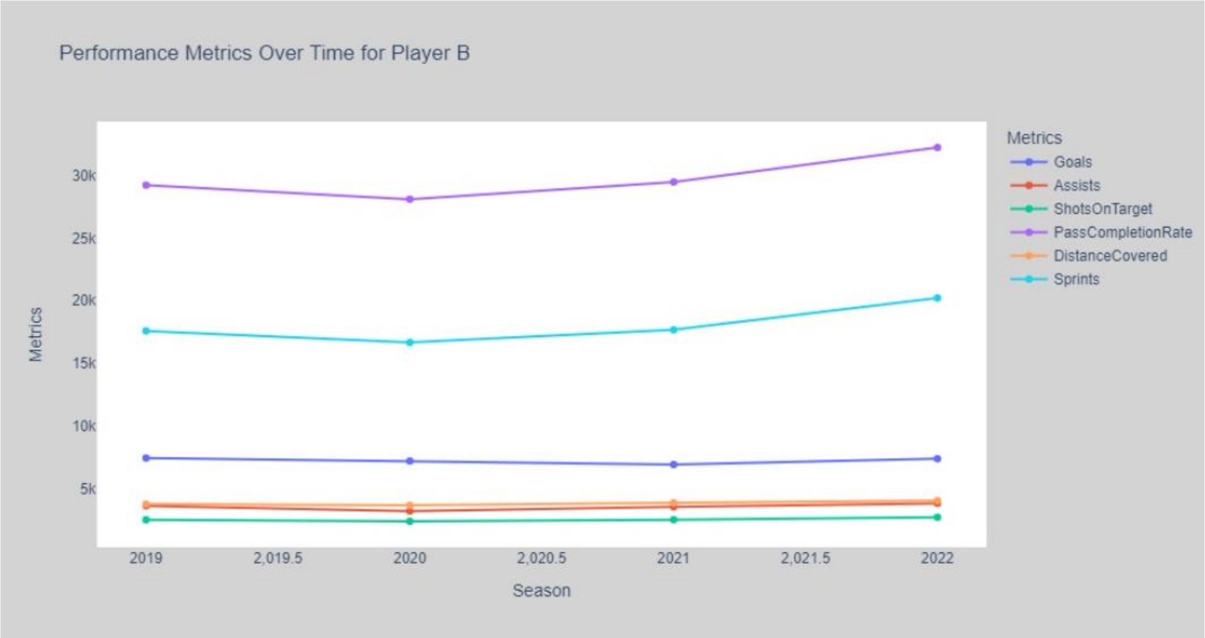


Fig 3.7(f) – Performance metrics of top player of top team- time series plot

3.8 Reporting and Visualization:

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 modeling to forecast future performance.

The plots given below are an comparison between each team and their players.

Plots:

- Performance impact of each player with injuries
- Minimum training hours of each team
- Sum of injury history of each team
- Players effective training hours in total training hours
- Total assists and goals by players
- Sum of yellow and red cards

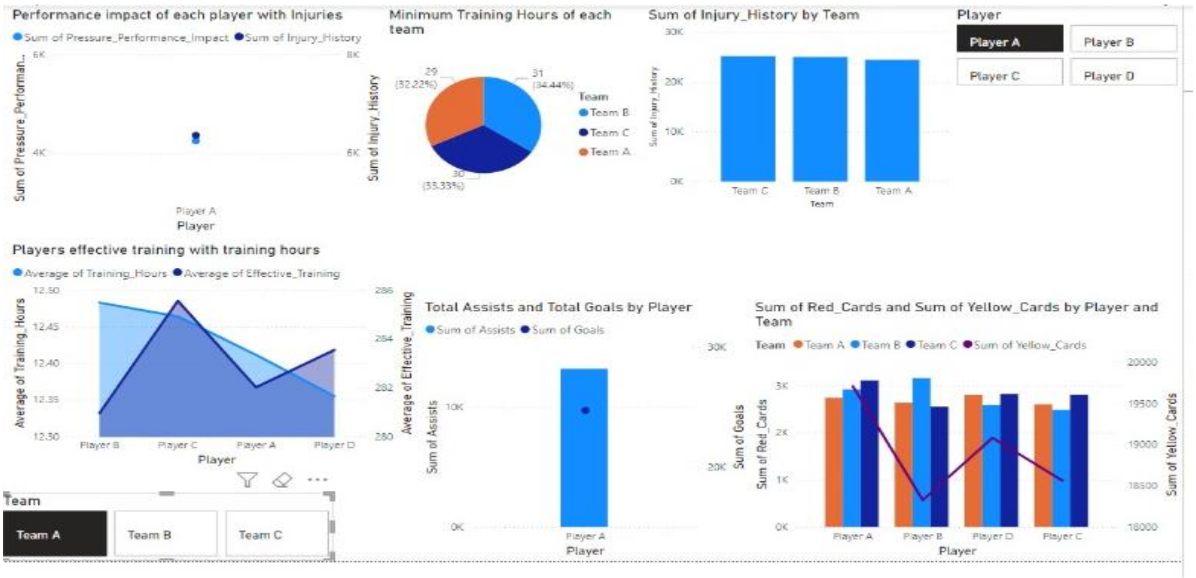

Performance impact of each player with Injuries
● Sum of Pressure_Performance_Impact ● Sum of Injury_History
Sum of Pressure_Performan...
Sum of Injury_History
Player A
Player
Minimum Training Hours of each team
29 (32.22%)
31 (34.44%)
30 (33.33%)
Team
Team B
Team C
Team A
Sum of Injury_History by Team
Team C
Team B
Team A
Team
Player
Player A
Player B
Player C
Player D
Players effective training with training hours
● Average of Training_Hours ● Average of Effective_Training
Average of Training_Hours
Average of Effective_Training
Player B
Player C
Player A
Player D
Player
Team
Team A
Team B
Team C
Total Assists and Total Goals by Player
● Sum of Assists ● Sum of Goals
Sum of Assists
Sum of Goals
Player A
Player
Sum of Red_Cards and Sum of Yellow_Cards by Player and Team
Team ● Team A ● Team B ● Team C ● Sum of Yellow_Cards
Sum of Red_Cards
Sum of Yellow_Cards
Player A
Player B
Player D
Player C
Player

Fig 3.8.1(a) – Team A-Player A

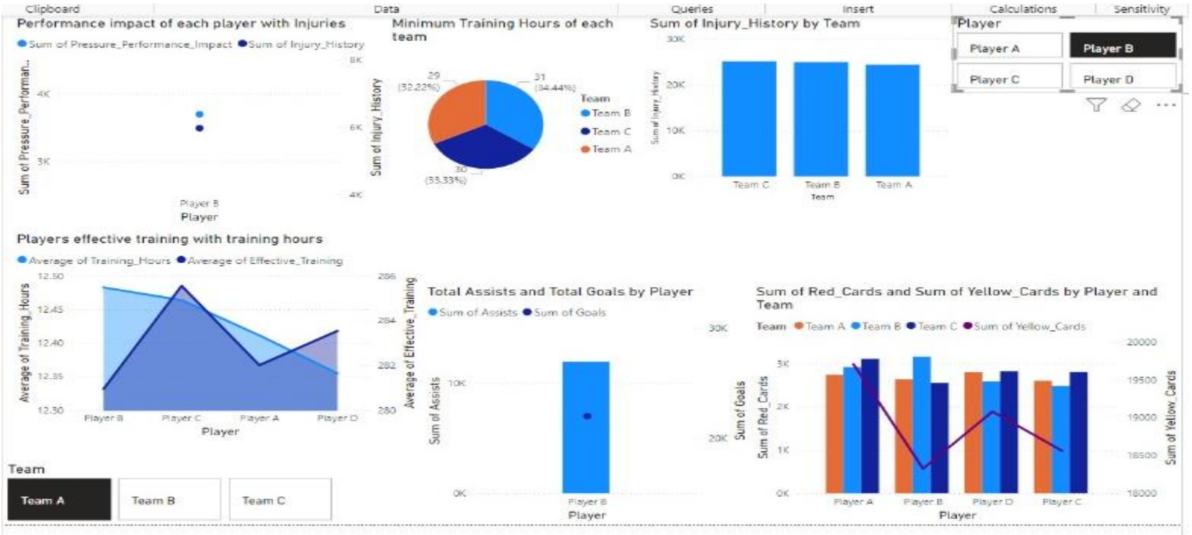

Clipboard
Data
Queries
Insert
Calculations
Sensitivity
Performance impact of each player with Injuries
● Sum of Pressure_Performance_Impact ● Sum of Injury_History
Sum of Pressure_Performan...
Sum of Injury_History
Player B
Player
Minimum Training Hours of each team
29 (32.22%)
31 (34.44%)
30 (33.33%)
Team
Team B
Team C
Team A
Sum of Injury_History by Team
Team C
Team B
Team A
Team
Player
Player A
Player B
Player C
Player D
Players effective training with training hours
● Average of Training_Hours ● Average of Effective_Training
Average of Training_Hours
Average of Effective_Training
Player B
Player C
Player A
Player D
Player
Team
Team A
Team B
Team C
Total Assists and Total Goals by Player
● Sum of Assists ● Sum of Goals
Sum of Assists
Sum of Goals
Player B
Player
Sum of Red_Cards and Sum of Yellow_Cards by Player and Team
Team ● Team A ● Team B ● Team C ● Sum of Yellow_Cards
Sum of Red_Cards
Sum of Yellow_Cards
Player A
Player B
Player D
Player C
Player

Fig 3.8.1(b) – Team A-Player B

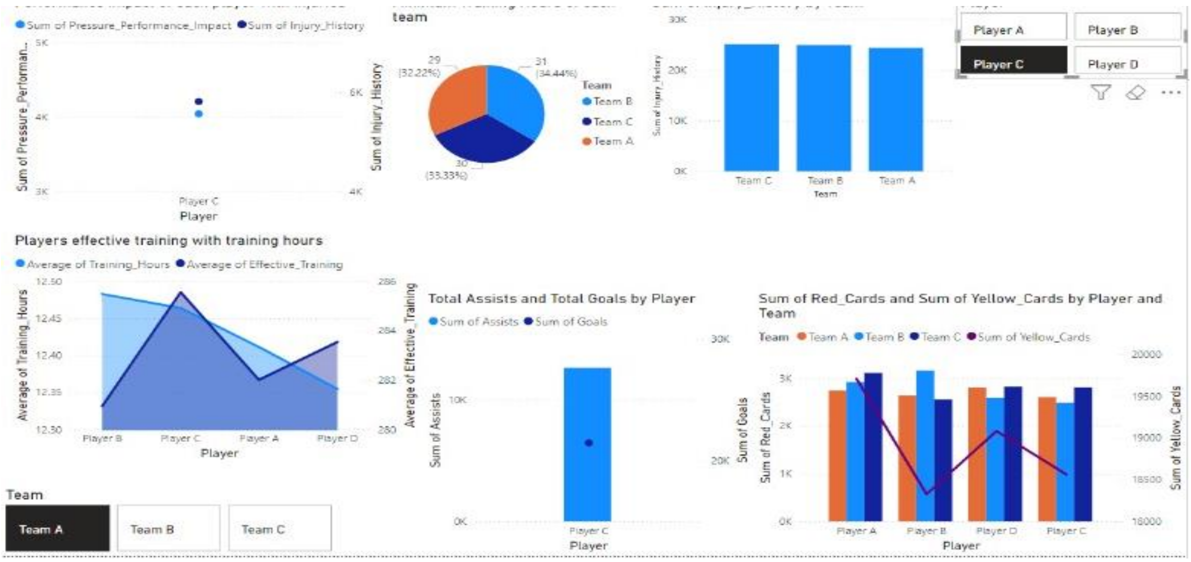
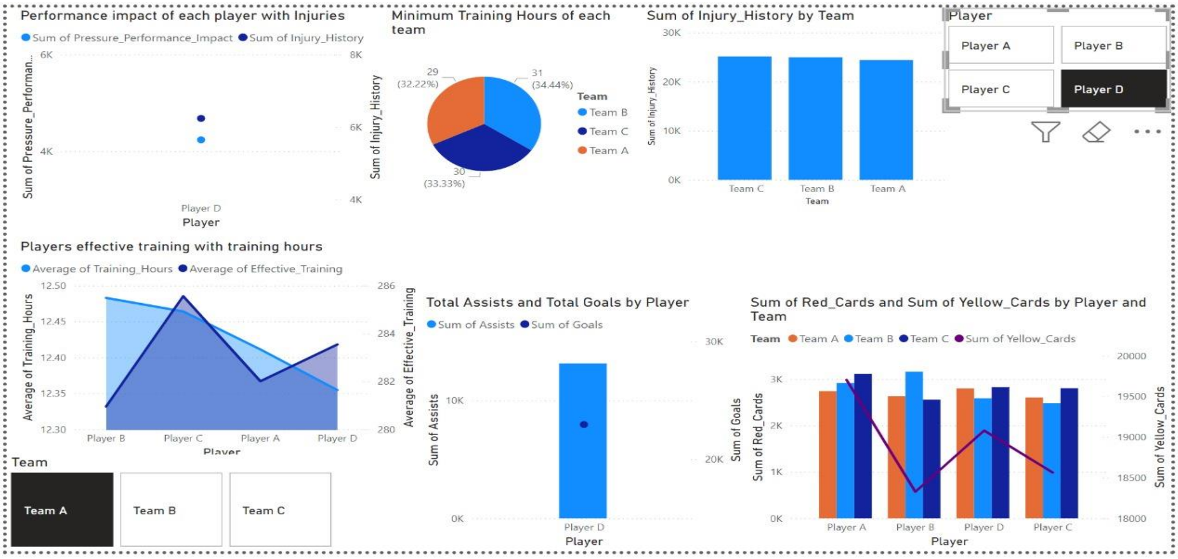


Fig 3.8.1(c) – Team A-Player C



Performance impact of each player with Injuries
● Sum of Pressure_Performance_Impact ● Sum of Injury_History
Sum of Pressure_Performan...
Sum of Injury_History
6K
4K
8K
6K
4K
Player D
Player
Minimum Training Hours of each team
29 (32.22%)
31 (34.44%)
30 (33.33%)
Team
● Team B
● Team C
● Team A
Sum of Injury_History by Team
Sum of Injury_History
30K
20K
10K
0K
Team C
Team B
Team A
Team
Player
Player A
Player B
Player C
Player D
Players effective training with training hours
● Average of Training_Hours ● Average of Effective_Training
Average of Training_Hours
Average of Effective_Training
12.50
12.45
12.40
12.35
12.30
286
284
282
280
Player B
Player C
Player A
Player D
Player
Team
Team A
Team B
Team C
Total Assists and Total Goals by Player
● Sum of Assists ● Sum of Goals
Sum of Assists
10K
0K
30K
20K
Player D
Player
Sum of Red_Cards and Sum of Yellow_Cards by Player and Team
Team ● Team A ● Team B ● Team C ● Sum of Yellow_Cards
Sum of Goals
Sum of Red_Cards
Sum of Yellow_Cards
3K
2K
1K
0K
20000
19500
19000
18500
18000
Player A
Player B
Player D
Player C
Player

Fig 3.8.1(d) – Team A-Player D

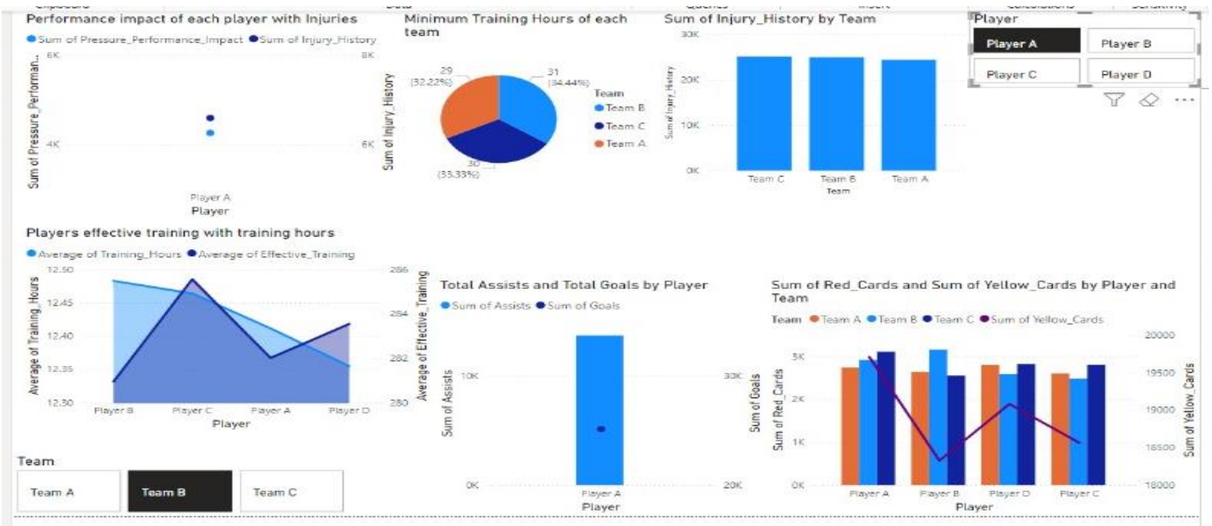


Fig 3.8.2(a) – Team B-Player A

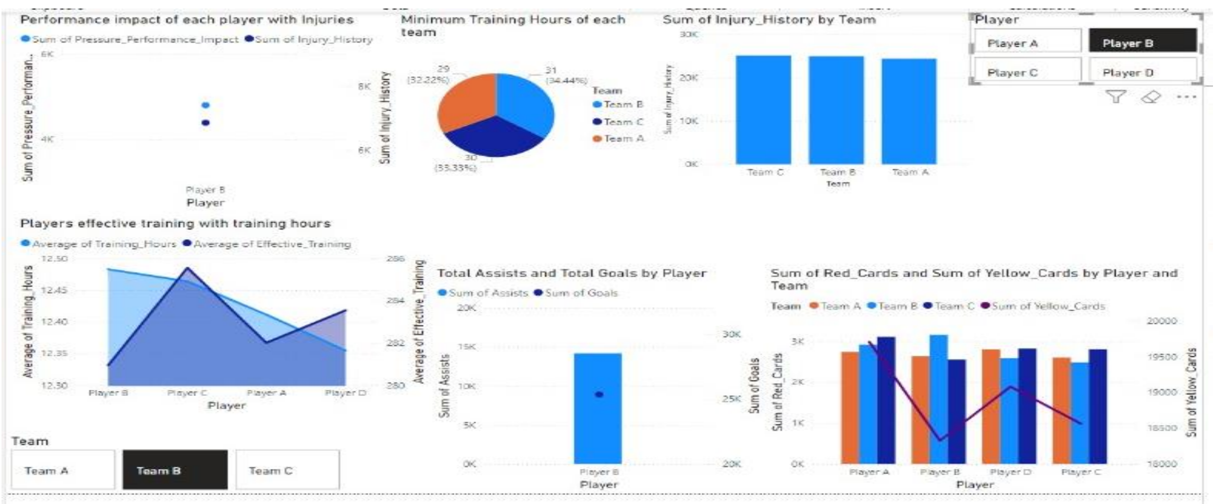


Fig 3.8.2(b) – Team B-Player B

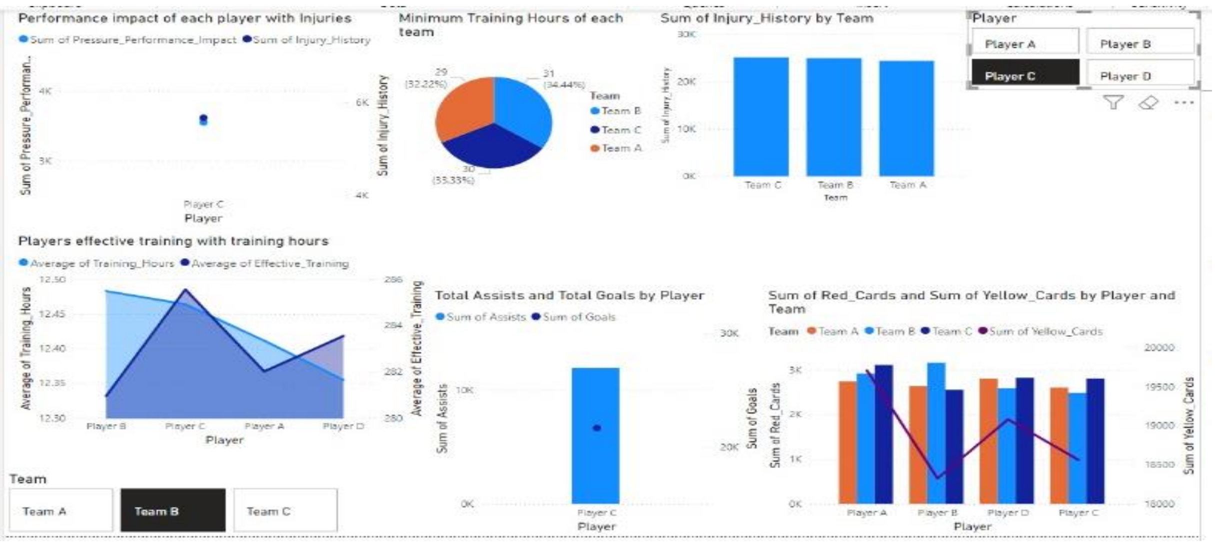


Fig 3.8.2(c) – Team B-Player C

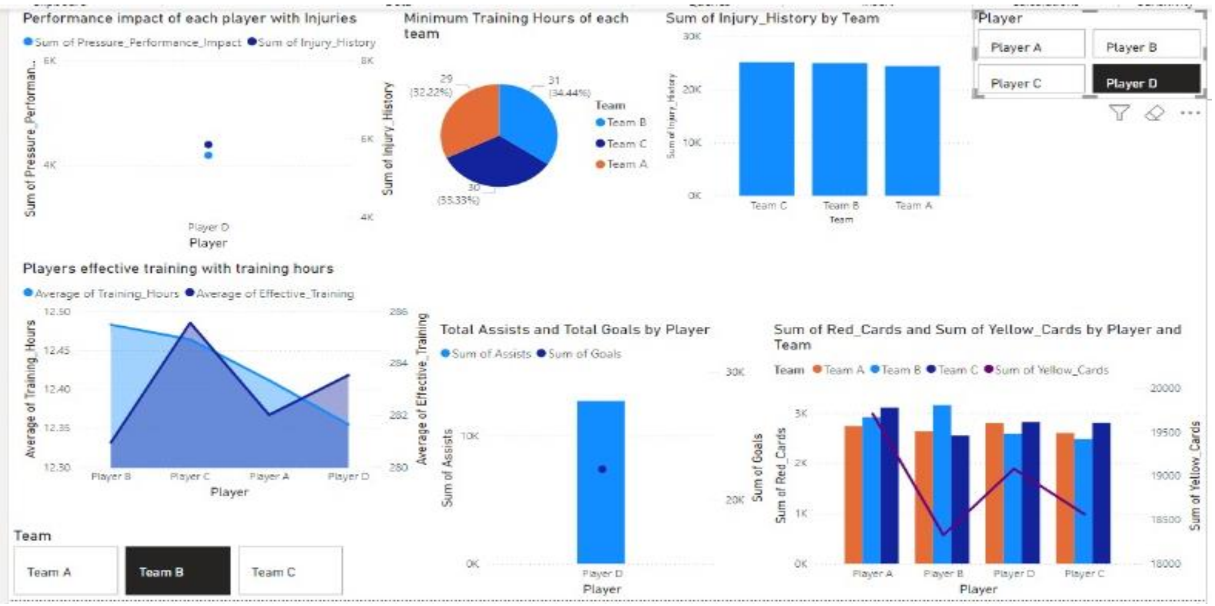


Fig 3.8.2(d) – Team B-Player D

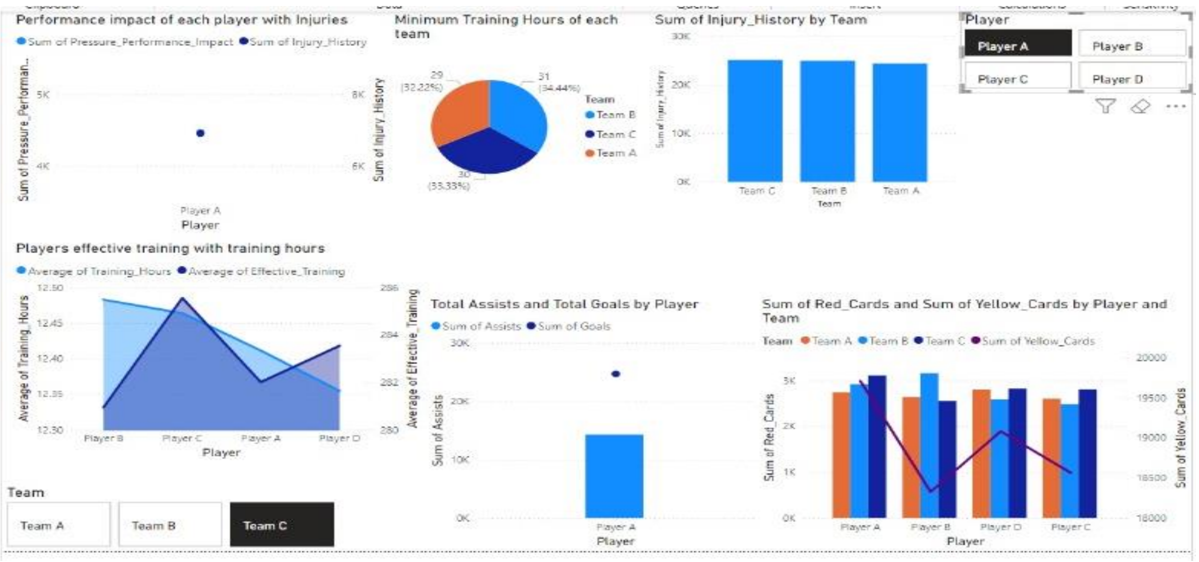

Performance impact of each player with Injuries
● Sum of Pressure_Performance_Impact ● Sum of Injury_History
Sum of Pressure_Performan...
5K
4K
8K
6K
Sum of Injury_History
Player A
Player
Minimum Training Hours of each team
29 (32.22%)
31 (34.44%)
30 (33.33%)
Team
● Team B
● Team C
● Team A
Sum of Injury_History by Team
30K
20K
10K
0K
Sum of Injury_History
Team C
Team B
Team A
Team
Player
Player A
Player B
Player C
Player D
Players effective training with training hours
● Average of Training_Hours ● Average of Effective_Training
12.50
12.45
12.40
12.35
12.30
Average of Training_Hours
286
284
282
280
Average of Effective_Training
Player B
Player C
Player A
Player D
Player
Total Assists and Total Goals by Player
● Sum of Assists ● Sum of Goals
30K
20K
10K
0K
Sum of Assists
Player A
Player
Sum of Red_Cards and Sum of Yellow_Cards by Player and Team
Team ● Team A ● Team B ● Team C ● Sum of Yellow_Cards
3K
2K
1K
0K
Sum of Red_Cards
20000
19500
19000
18500
18000
Sum of Yellow_Cards
Player A
Player B
Player D
Player C
Player
Team
Team A
Team B
Team C

Fig 3.8.3(a) – Team C-Player A

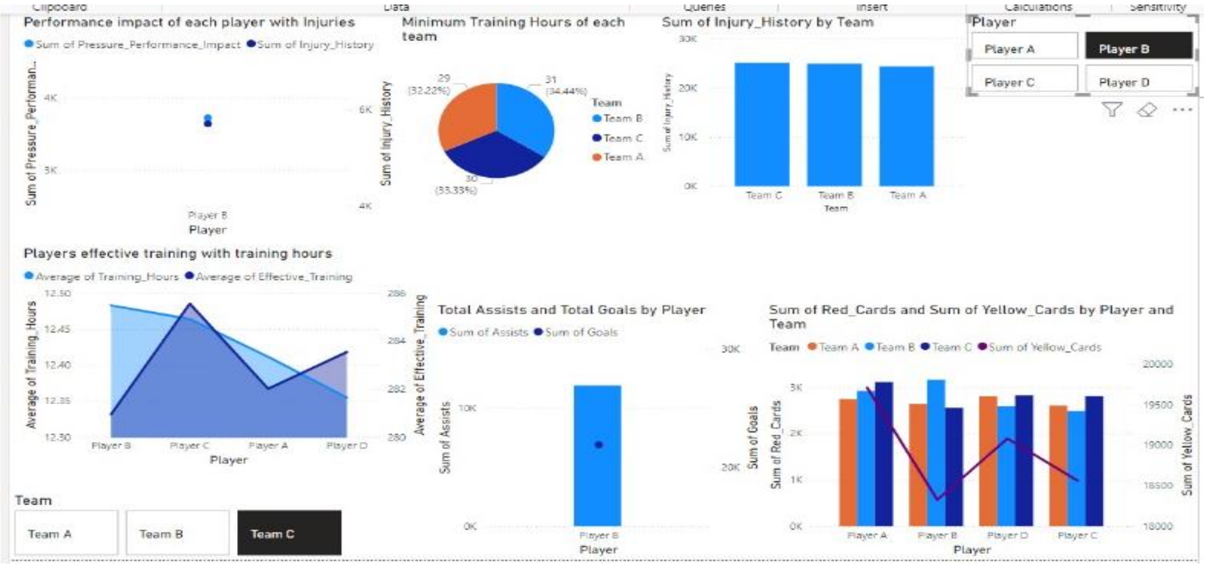

Performance impact of each player with Injuries
● Sum of Pressure_Performance_Impact ● Sum of Injury_History
Sum of Pressure_Performan...
4K
3K
6K
4K
Sum of Injury_History
Player B
Player
Minimum Training Hours of each team
29 (32.22%)
31 (34.44%)
30 (33.33%)
Team
● Team B
● Team C
● Team A
Sum of Injury_History by Team
30K
20K
10K
0K
Sum of Injury_History
Team C
Team B
Team A
Team
Player
Player A
Player B
Player C
Player D
Players effective training with training hours
● Average of Training_Hours ● Average of Effective_Training
12.50
12.45
12.40
12.35
12.30
Average of Training_Hours
286
284
282
280
Average of Effective_Training
Player B
Player C
Player A
Player D
Player
Total Assists and Total Goals by Player
● Sum of Assists ● Sum of Goals
10K
0K
Sum of Assists
30K
20K
Sum of Goals
Player B
Player
Sum of Red_Cards and Sum of Yellow_Cards by Player and Team
Team ● Team A ● Team B ● Team C ● Sum of Yellow_Cards
3K
2K
1K
0K
Sum of Red_Cards
20000
19500
19000
18500
18000
Sum of Yellow_Cards
Player A
Player B
Player D
Player C
Player
Team
Team A
Team B
Team C

Fig 3.8.3(b) – Team C-Player B

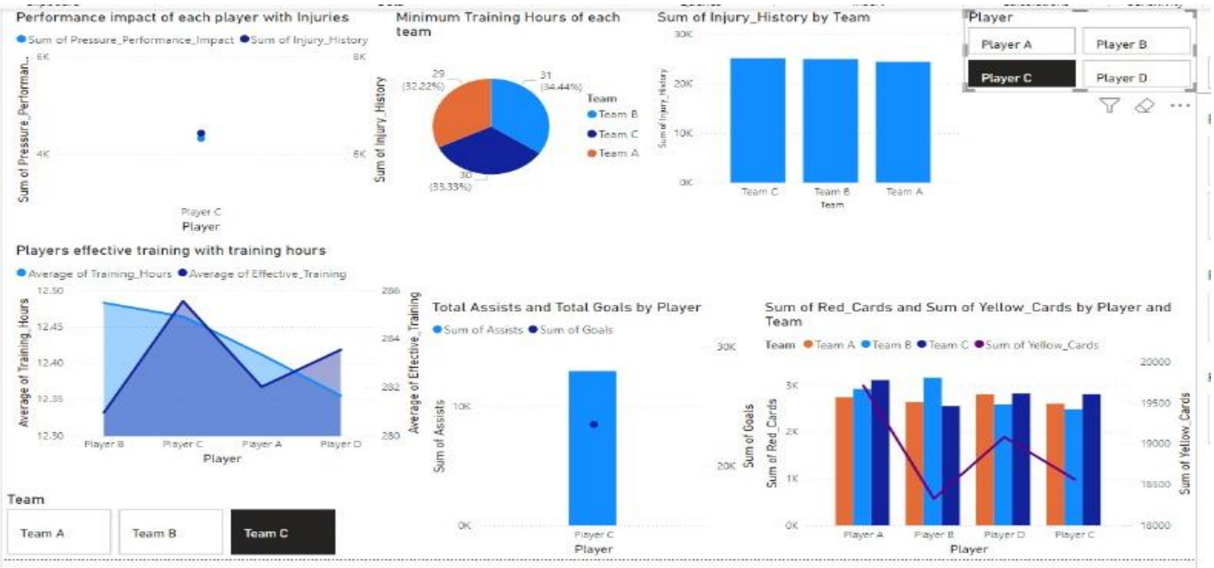


Fig 3.8.3(c) – Team C-Player C

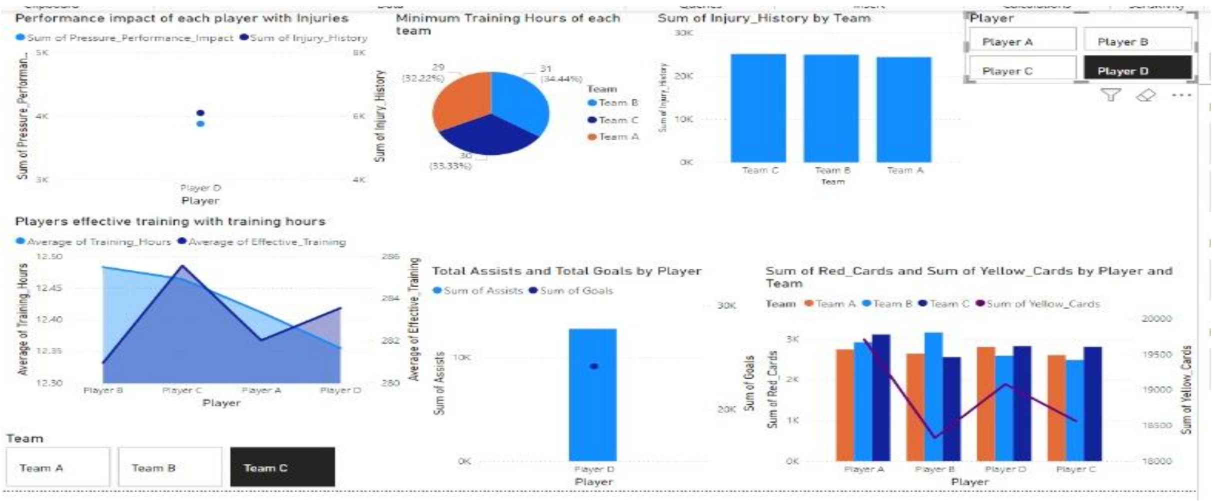


Fig 3.8.3(d) – Team C-Player D

CHAPTER 4. CONCLUSION

This data engineering project has successfully demonstrated the power of data-driven insights in the realm of sports. By tackling a series of complex challenges, including data cleaning, pipeline optimization, and advanced data transformation, we have built a robust and insightful data infrastructure. The resulting interactive dashboards and visualizations provide a powerful tool for analyzing player performance, team strategies, and potential areas for improvement.

The project highlights the importance of leveraging advanced techniques like data imputation, statistical analysis, regression modeling, outlier detection, and machine learning to extract meaningful insights from complex datasets. It also underscores the significance of developing efficient data pipelines and data warehousing strategies to manage and optimize large-scale data.

Moving forward, the insights derived from this project can be further leveraged to enhance decision-making within the sports industry. By continuously refining the data infrastructure and exploring new data sources and analytical techniques, we can unlock even greater value from sports data, driving innovation and improving performance across all aspects of the game.

REFERENCES

1. Data Cleaning and Augmentation:

- Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

- Van der Aalst, W. M. P. (2016). *Process Mining: Data Science in Action*. Springer.

W3schools Data Cleaning

w3schools pandas

2. Data Ingestion Strategies:

- Karau, H., & Warren, R. (2017). *High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark*. O'Reilly Media.

- Makadia, A. (2020). *Practical DataOps: Delivering Agile Data Science at Scale*. O'Reilly Media.

geeksforgeeks Data Ingestion

geeksforgeeks sql indexes

3. Advanced Data Transformations:

- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

github repository sql

github repository (joins)

Data Visualization

4. Data Warehousing:

- Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley.

- Inmon, W. H., O'Neil, B., & Fryman, L. (2010). *Business Metadata: Capturing Enterprise Knowledge*. Morgan Kaufmann.

5. Reporting and Visualization: - Few, S. (2012). *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Analytics Press.

- McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.

uses of power bi

How to use power bi

6. Predictive Analytics and Machine Learning:

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Feature Engineering

7. Interactive Dashboards:

- Munzner, T. (2014). *Visualization Analysis and Design*. CRC Press.

- Jones, K. (2014). *Tableau Your Data!: Fast and Easy Visual Analysis with Tableau Software*. Wiley.

Visualization using power bi

APPENDIX-I

SOURCE CODE

(GITHUB REPO LINK)

GITHUB REPOSITORY LINK

APPENDIX - II

DATASHEETS

Nature of Datasets:

The dataset appears to be a comprehensive collection of statistics for individual players across various teams, likely within a sports context. It includes performance metrics such as goals, assists, yellow and red cards, minutes and matches played, and shots on target. Additionally, it records physical attributes like height, weight, and age, alongside mental and psychological factors such as fatigue levels, pressure impact, concentration, and stress resistance. Other pertinent details might include injury history and training intensity. This dataset is invaluable for analyzing player performance, scouting talent, managing team rosters, and making strategic decisions, offering a holistic view of each player's strengths and areas for improvement. This dataset provides a valuable starting point for exploring player statistics and gaining insights about player performance, physical attributes, and mental aspects within a sports context. By conducting further analysis and addressing limitations, you can extract more valuable knowledge from this dataset.

Observations:

Upon examining the dataset, several observations can be made. There are missing values in the Height, Weight, and Season columns. The dataset comprises mostly numerical data, but some columns such as Player, Team, and Position are categorical. A potential outlier exists in the Goals column, with Player D from Team B scoring 280 goals, which is exceptionally high compared to other players. Additionally, there is noticeable variability across various attributes, including differences in players' heights, weights, and fatigue levels.

Insights:

The dataset reveals several key insights. Team C demonstrates dominance with a high concentration of players achieving substantial numbers of goals and assists. There is notable variety among goalkeepers in terms of goals scored, suggesting different roles or strategies across teams. Defender performance also varies, with some defenders, like Player A from Team C, scoring many goals, while others focus on tackles and maintaining clean sheets, indicating diverse roles within defensive lines. The FatigueInjuryCorrelation column highlights that higher fatigue levels may increase injury risk for some players. The PressurePerformanceImpact column shows that high-pressure situations can negatively affect certain players' performance. Additionally, Player D from Team B stands out with an exceptionally high number of goals, possibly due to a unique playing style, scoring opportunities, or strategic positioning within the team.

Limitations and Further analysis:

For further analysis, addressing missing values using appropriate imputation techniques such as KNN, mean, or median is essential to avoid data loss. Investigating the outlier in the Goals column is also crucial to determine whether it is a data entry error, an exceptionally talented player, or indicative of a different playing style or league. Conducting a correlation analysis between various attributes, like Height and Weight or Training Hours and Performance, could yield additional insights. Creating visualizations such as histograms, scatter plots, and bar charts will help to better understand the relationships and trends within the data. Additionally, building predictive models can estimate future player performance based on the existing data. However, there are limitations to consider: the dataset's relatively small size might not fully represent the broader player population, and the lack of context about the league or specific game scenarios could influence data

Information Regarding Students

Student name	Email id	Permanent Address	Phone Number
Ayushi Tawari	ayushitawari03@gmail.com	yasoram indradhanush apartament td road Ernakulam Kerala	9995951128
Mutta Datta Sai Vishnu Mohan	vishnu.jojo26@gmail.com	dr ngr’s gandhi das apartment,vijayawada, andhra pradesh	9391452963
Niranjana J	niranjanajithendran@gmail.com	Sarangi (H), Thazhvaram road, Kavungal bypass, uphill (p.o), Malappuram, Kerala	8891800705