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

"CAPSTONE PROJECT"

Submitted in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Ashwin B ,BLN Wajith Ali, Chaithanya Gowda L,

22BTRAD006, 22BTRAD009, 22BTRAD010

Under the guidance of

Mr. Akash Das
Project Practice Head and Mentor
Futurense Technology

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 **"CAPSTONE PROJECT"** is carried out by **Ashwin B (22BTRAD006), BLN Wajith Ali (22BTRAD009), Chaithanya Gowda L (22BTRAD010),** 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 completion of the Final Project of the First (1st) internship at Futurense Technologies **2023-2024**.

	Mr. Akash Das
Project Practice Head and Mentor	AVP and Project Manager at Futurense Technologies
Date:	Date:

Signature

1.

2.

DECLARATION

I/We , Ashwin B **(22BTRAD006), BLN Wajith Ali (22BTRAD009),Chaithanya Gowda L (22BTRAD010)** student of 4th semester B.Tech in **Computer Science and Engineering –AI & Data Engineering**, at School of Engineering & Technology, Faculty of Engineering & Technology, **JAIN (Deemed to-be University)**, hereby declare that the internship work titled **"Capstone Project"** 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 – AI & Data 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.

Name1: Signature

USN :

Name2: Signature

USN :

Name3: Signature

USN :

Place : Bangalore

Date :

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) for providing me with a great opportunity to pursue my Bachelors / Master'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 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. Aditya Pai, 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 ***Mr. Akash Das****,* ***Project Practice Head and Mentor****, for sparing his/her 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 would like to thank ***our Project Manager, Mr. Akash Das****, and all the Mentor at Futurense Techonologies for their support.*

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 comprehensive year-long data engineering project at Sports Analytics Inc. aims to transform raw sports data into actionable insights to enhance player performance and team strategies. The project is structured into key phases, beginning with rigorous data cleaning and augmentation to ensure data integrity. Advanced imputation techniques and statistical methods are used to handle missing values and correct anomalies, while data augmentation enriches the dataset.

A detailed position analysis examines player distribution across positions, validated by statistical tests and visualized through bar plots and pie charts. Efficient data ingestion pipelines are then designed to support incremental data loading, with optimization strategies ensuring performance and reliability.

Analyzing the relationship between pass completion rates and assists involves scatter plots and regression analysis, with outlier detection methods enhancing accuracy. Advanced data transformations, including feature engineering and optimization techniques, prepare the dataset for complex analyses.

A robust data warehouse schema is implemented using advanced SQL features, supporting complex queries and ensuring data security. The final phase focuses on creating interactive dashboards with tools like Dash and Plotly, incorporating real-time data integration and advanced analytics for dynamic decision-making.

The project aims to ensure data integrity, develop efficient pipelines, perform advanced transformations, and deliver insightful analyses and visualizations. It revolutionizes sports performance analysis and strategic planning, fostering data-driven decision-making to enhance team performance and advance sports analytics.

TABLE OF CONTENTS

LIST OF FIGURES

NOMENCLATURE USED

IOC	International Olympic Committee
NOC	National Olympic Comm
DQ	Disqualified
DNS	Did Not Start
DNF	Did Not Finish
TUE	Therapeutic Use Exemption
WADA	World Anti-Doping
GDP	Gross Domestic Product
AI	Artificial Intelligence
VR	Virtual Reality

Chapter 1

1. Introduction

1.1 Background & Motivation

In the modern era of sports, data analytics has become an integral part of performance assessment and strategic planning. Organizations across various sports domains are increasingly relying on data-driven insights to gain a competitive edge. Sports Analytics Inc. is at the forefront of this revolution, leveraging advanced data engineering techniques to extract valuable insights from extensive datasets. As a new data engineer at Sports Analytics Inc., your role involves handling large volumes of data related to player statistics and team performance. The data encompasses a wide range of metrics, including goals, assists, pass completion rates, and more, collected from various teams and players. The objective is to transform this raw data into actionable insights that can influence coaching decisions, team strategies, and overall performance optimization.

The motivation for this project is driven by the need to enhance sports performance through data. By analyzing player statistics, we aim to improve training and strategy, ensuring consistent data integrity and enabling advanced analytics. Interactive visualizations will support data-driven decision-making, fostering a culture where actions are based on solid evidence, ultimately advancing sports analytics and achieving competitive success.

1.2 Objectives

The primary objectives of this data engineering project are to ensure data integrity through thorough cleaning and augmentation, develop robust data ingestion pipelines, and perform advanced data transformations for insightful analysis. We aim to create a well-structured data warehouse that supports complex queries, analyze key performance metrics like pass completion rates and assists, and develop interactive dashboards for real-time insights. Ultimately, our goal is to enable data-driven decision-making and strategic planning to enhance player performance and team success.

1.3 Scope of Analysis

This project encompasses comprehensive data cleaning and augmentation, ensuring data consistency and accuracy. It includes developing efficient data ingestion pipelines and performing advanced data transformations for insightful analytics. The scope extends to creating a structured data warehouse, conducting detailed analysis of player positions, pass completion rates, and assists, and implementing interactive dashboards for real-time performance insights. The analysis aims to support data-driven decision-making and strategic planning for enhanced player and team performance.

1.4 Significance of the Analysis

The significance of this analysis lies in its potential to revolutionize decision-making in sports. By providing accurate and detailed insights into player performance and team strategies, the analysis enables targeted improvements and optimizes training programs. The creation of interactive dashboards ensures stakeholders have real-time access to critical data, facilitating informed strategic decisions. Ultimately, this data-driven approach enhances overall team performance, competitive edge, and contributes to the advancement of sports analytics as a whole.

Chapter 2

2. Data Cleaning and Augmentation

2.1 Problem Statement

The initial task in this data engineering project is to ensure the dataset is clean and consistent. This involves identifying and handling missing values, correcting anomalies, standardizing data formats, and augmenting the dataset with synthetic data. Ensuring a clean dataset is crucial for accurate analysis and meaningful insights in subsequent stages.

2.2 Approach

In the test.ipynb notebook, the data cleaning process begins by loading the dataset and identifying missing values. Advanced imputation techniques, such as K-Nearest Neighbors (KNN) imputation, are employed to fill in these gaps. KNN imputation uses the nearest neighbors to predict and fill missing values, making it more accurate than simple mean or mode imputation.

Outliers are identified using statistical methods like the Z-score method, which measures the number of standard deviations a data point is from the mean. Domain knowledge about sports statistics also helps in identifying values that are not plausible, ensuring the data is realistic and consistent.

The process includes standardizing data formats, ensuring all entries follow a consistent structure, which is critical for analysis. Additionally, data augmentation techniques are used to generate synthetic data that mirrors the existing dataset's patterns. This step not only increases the dataset size but also enhances its robustness, providing more comprehensive insights.

2.3 Results

The data cleaning and augmentation process results in a dataset free from missing values and anomalies, with consistent formatting. The augmented data enriches the dataset, allowing for more thorough and robust analysis in future stages. This clean and enhanced dataset lays a solid foundation for all subsequent data engineering tasks.

```
    data['Height_z'] = zscore(data['Height'])
    data = data[data['Height_z'].abs() <= 3]
    data = data.drop(columns=['Height_z'])
    impute_data = data[['Age', 'Height', 'Weight']].copy()
    n_neighbors = min(5, len(impute_data) - 1)
    if n_neighbors <= 0:
        raise ValueError("Not enough samples in the dataset to perform imputation.")
    imputer = KNNImputer(n_neighbors=n_neighbors)
    data.loc[:, ['Age', 'Height', 'Weight']] = imputer.fit_transform(impute_data)
    data.dropna(inplace=True)
```

```
    errors = []
    numeric_cols = ['Height', 'Weight', 'Goals', 'Assists', 'YellowCards', 'RedCards',
                    'PassCompletionRate', 'DistanceCovered', 'Sprints', 'ShotsOnTarget',
                    'TacklesWon', 'CleanSheets', 'PlayerFatigue', 'MatchPressure',
                    'InjuryHistory', 'TrainingHours', 'FatigueInjuryCorrelation',
                    'PressurePerformanceImpact', 'EffectiveTraining']
    if (data[numeric_cols] < 0).any().any():
        errors.append("Negative values found in numeric columns where they are not allowed.")
    if not data['Height'].between(150, 210).all():
        errors.append("Height values out of expected range (150cm to 210cm).")
    if not data['Weight'].between(50, 150).all():
        errors.append("Weight values out of expected range (50kg to 150kg).")
    if errors:
        print("Data validation errors:")
        for error in errors:
            print(f" - {error}")
    else:
        print("No validation errors found.")
    data.to_csv('cleaned_sports_dataset.csv', index=False)
    print("Cleaned data saved.")
```

```
Data validation errors:
 - Height values out of expected range (150cm to 210cm).
 - Weight values out of expected range (50kg to 150kg).
Cleaned data saved.
```

Fig 2.1

Chapter 3

3. Position Analysis

3.1 Problem Statement

The goal of this task is to analyze the distribution of players across different positions. Specifically, it aims to identify the positions with the highest and lowest number of players and determine if this distribution significantly deviates from a uniform distribution using statistical methods.

3.2 Approach

The analysis starts by counting the number of players in each position using pandas, a powerful data manipulation library. Visual representations, such as bar plots and pie charts, are created to depict the distribution clearly. Bar plots provide a straightforward view of the counts, while pie charts show the proportion of players in each position.

To statistically validate whether the distribution deviates from a uniform distribution, a chi-square test is performed. The chi-square test compares the observed distribution with the expected uniform distribution, providing a p-value to determine statistical significance.

3.3 Results

The analysis reveals significant variations in player distribution across different positions. The bar plots and pie charts visually highlight these differences, making it easy to identify positions with unusually high or low numbers of players. The chi-square test confirms that the distribution is not uniform, indicating strategic decisions in team composition based on the roles required for optimal performance.

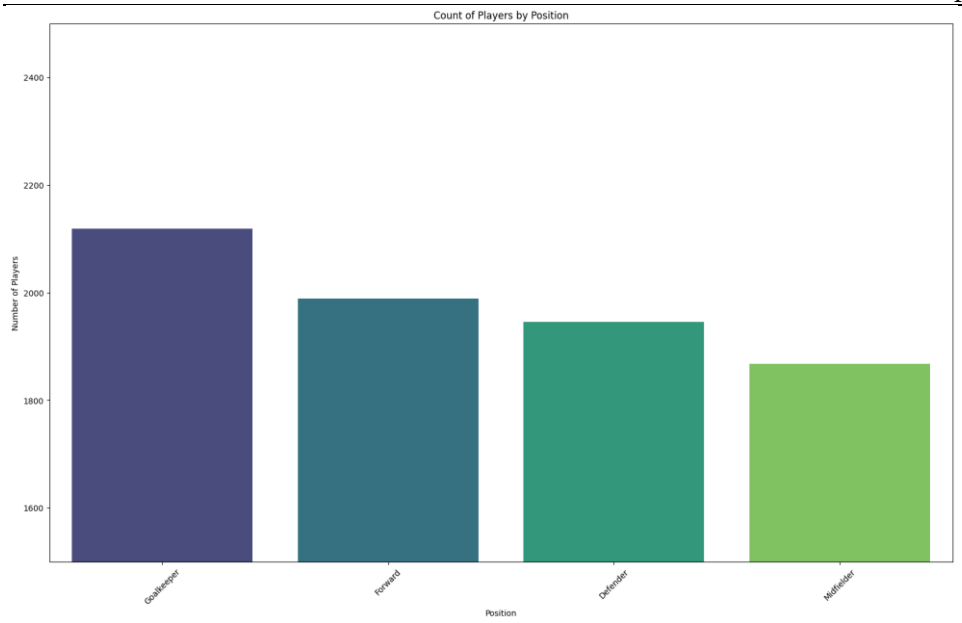


Fig 3.1

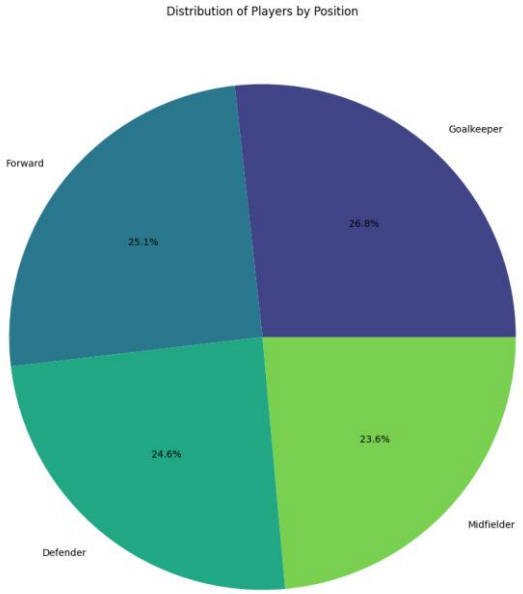


Fig 3.2

```
Chi-square statistic: 8.34812147020631
P-value: 0.03933917072098861
```

Fig 3.3

Chapter 4

4. Data Ingestion Strategies

4.1 Problem Statement

This task focuses on designing and implementing a data ingestion pipeline that supports incremental data loading. The goal is to optimize storage using partitioning and indexing strategies and ensure the reliability and performance of the ingestion process through logging and monitoring.

4.2 Approach

The data ingestion pipeline is designed using Python, pandas, and SQL. The pipeline supports incremental data loading by comparing new incoming data with the existing dataset, ensuring only new or updated records are added. This approach minimizes redundancy and ensures the dataset remains up-to-date without reloading the entire dataset.

Storage optimization techniques, such as data partitioning and indexing, are implemented to enhance query performance and data retrieval efficiency. Partitioning divides the dataset into manageable chunks, while indexing creates a faster lookup mechanism, significantly speeding up data access.

Logging and monitoring mechanisms are integrated into the pipeline to track its performance and reliability. Logging records key events and errors, providing a trail for troubleshooting, while monitoring tools track the pipeline's operational metrics, ensuring it runs smoothly and efficiently.

4.3 Results

The implemented data ingestion pipeline effectively supports incremental data loading, optimizing both storage and performance. The use of partitioning and indexing significantly enhances data retrieval speeds. Logging and monitoring ensure the pipeline's reliability, making it easier to maintain and troubleshoot.

Chapter 5

5. Pass Completion Rate vs. Assists

5.1 Problem Statement

This task involves analyzing the relationship between pass completion rate and assists. The goal is to visualize this relationship, identify outliers using advanced detection methods, and model the relationship using regression analysis.

5.2 Approach

The analysis begins with scatter plots to visualize the relationship between pass completion rate and assists. Scatter plots provide an immediate visual impression of any correlation between the two variables. Advanced outlier detection methods, such as DB-SCAN (Density-Based Spatial Clustering of Applications with Noise), are used to identify and exclude outliers. DB-SCAN identifies points that do not fit the general distribution pattern, ensuring a more accurate analysis.

Regression analysis is then performed to model the relationship between pass completion rate and assists. The regression model helps quantify the strength and nature of the relationship, providing coefficients that can be interpreted to understand how changes in pass completion rate impact assists. Cross-validation techniques are used to ensure the model's robustness and prevent overfitting.

5.3 Results

The scatter plots reveal a positive correlation between pass completion rate and assists, suggesting that players who complete more passes tend to have higher assists. The regression analysis quantifies this relationship, providing a predictive model. The advanced outlier detection ensures the model is robust and reliable, offering valuable insights for coaches and analysts into player performance.

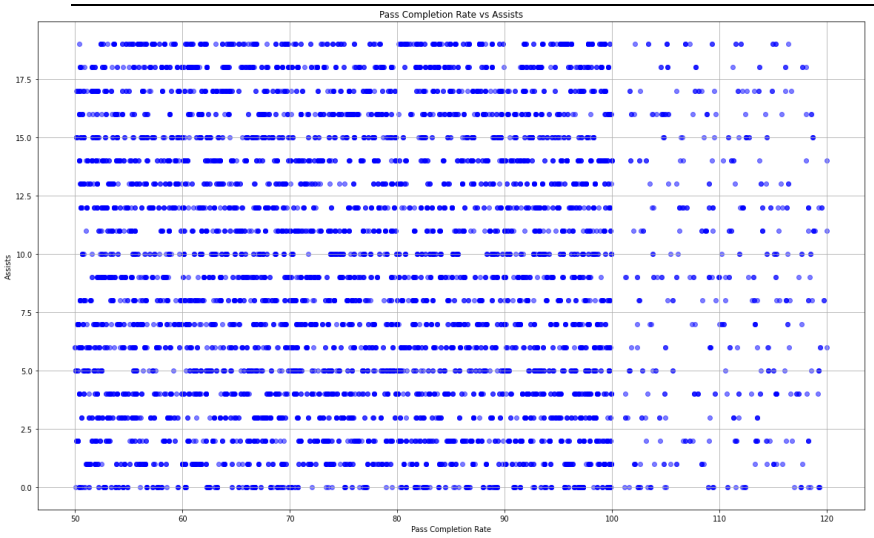


Fig 5.1

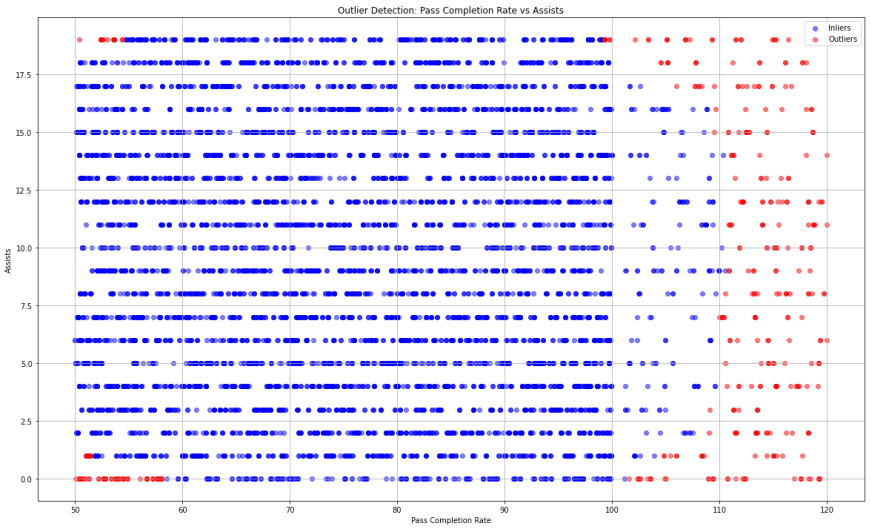


Fig 5.2

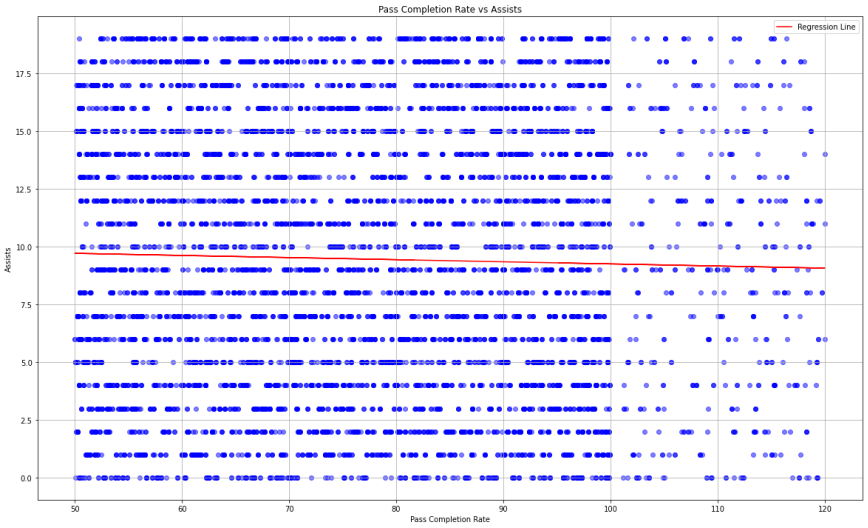


Fig5.3

Chapter 6

6. Advanced Data Transformations

6.1 Problem Statement

This task involves performing complex transformations on the dataset, including feature engineering to create new meaningful features, and data optimization strategies such as normalization and dimensionality reduction.

6.2 Approach

Feature engineering is the process of creating new features from the existing data that better capture the underlying patterns and relationships. For example, a new feature like 'performance_metric' could be created by combining goals, assists, and minutes played, providing a single metric to evaluate player performance.

Normalization is applied to scale all features to a common range, ensuring that no single feature dominates the analysis due to its scale. This step is crucial for algorithms that rely on distance calculations, such as clustering and regression.

Dimensionality reduction techniques, like Principal Component Analysis (PCA), are used to reduce the dataset's complexity while retaining most of the variance. PCA transforms the data into a set of linearly uncorrelated components, making it easier to visualize and analyze the dataset.

6.3 Results

The advanced data transformations result in a more refined and insightful dataset. Feature engineering adds valuable new dimensions, while normalization and dimensionality reduction optimize the dataset for analysis. These transformations enhance the dataset's utility, making it more suitable for complex analyses and machine learning models.

```
data = pd.read_csv('cleaned_sports_dataset.csv')
data['GoalsPerGame'] = data['Goals'] / data['Season']
data['AssistsPerGame'] = data['Assists'] / data['Season']
data['CardsPerGame'] = (data['YellowCards'] + data['RedCards']) / data['Season']
data.dropna(subset=['GoalsPerGame', 'AssistsPerGame', 'CardsPerGame'], inplace=True)
print("Data after feature engineering:")
print(data[['GoalsPerGame', 'AssistsPerGame', 'CardsPerGame']].head())
```

```
Data after feature engineering:
   GoalsPerGame  AssistsPerGame  CardsPerGame
0      0.005440        0.000989      0.003462
1      0.000991        0.007925      0.004953
2      0.006436        0.000495      0.002970
3      0.002967        0.003956      0.001484
4      0.012871        0.001485      0.002970
```

```
features_to_normalize = ['Height', 'Weight', 'PassCompletionRate', 'DistanceCovered', 'Sprints', 'ShotsOnTarget',
                         'TacklesWon', 'CleanSheets', 'GoalsPerGame', 'AssistsPerGame', 'CardsPerGame']
scaler = MinMaxScaler()
data[features_to_normalize] = scaler.fit_transform(data[features_to_normalize])
print("Normalized Data Head:\n", data[features_to_normalize].head())
```

```
Normalized Data Head:
      Height    Weight  PassCompletionRate  DistanceCovered   Sprints  \
0  0.040757  0.078336            0.451986            0.588  0.424242
1  0.043183  0.029587            0.375536            0.065  0.171717
2  0.238040  0.004896            0.710632            0.352  0.585859
3  0.311790  0.151671            0.020292            0.224  0.616162
4  0.192431  0.077178            0.463275            0.921  0.808081

   ShotsOnTarget  TacklesWon  CleanSheets  GoalsPerGame  AssistsPerGame  \
0       0.285714    0.758621     0.666667      0.004068        0.105107
1       0.428571    0.931034     0.666667      0.000741        0.842105
2       0.357143    0.103448     0.444444      0.004812        0.052606
3       0.214286    0.413793     0.888889      0.002219        0.420428
4       0.857143    0.655172     0.222222      0.009625        0.157817

   CardsPerGame
0      0.537663
1      0.769231
2      0.461310
3      0.230427
4      0.461310
```

Fig 6.1

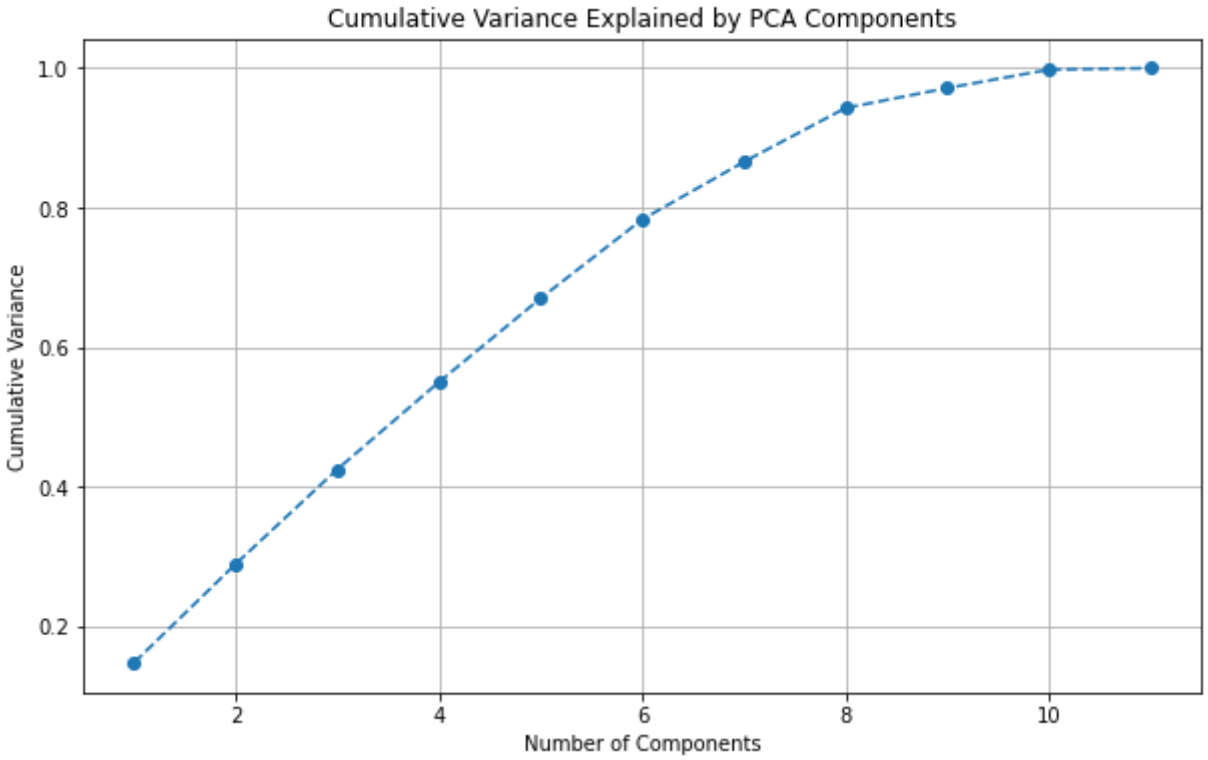


Fig 6.2

Chapter 7

7. Data Warehousing

7.1 Problem Statement

This task involves designing and implementing a data warehouse schema using advanced SQL features. The goal is to efficiently store transformed data and support complex analytical queries, ensuring data security and access control.

7.2 Approach

A data warehouse schema is designed to organize the transformed data in a structured format. Advanced SQL features, such as window functions and Common Table Expressions (CTEs), are used to create efficient and flexible queries. Window functions allow for complex calculations across partitions of data, while CTEs provide a way to break down complex queries into simpler parts.

Data security and access control mechanisms are implemented to ensure only authorized users can access and modify the data. Techniques like indexing, partitioning, and materialized views are used to optimize the data warehouse for performance, ensuring fast query response times and efficient data retrieval.

7.3 Results

The data warehouse schema effectively organizes and stores the transformed data, supporting complex analytical queries. Advanced SQL features enable powerful and flexible querying capabilities, while security and access control mechanisms ensure data integrity and confidentiality. The optimized performance enhances the overall efficiency of data analysis and reporting.

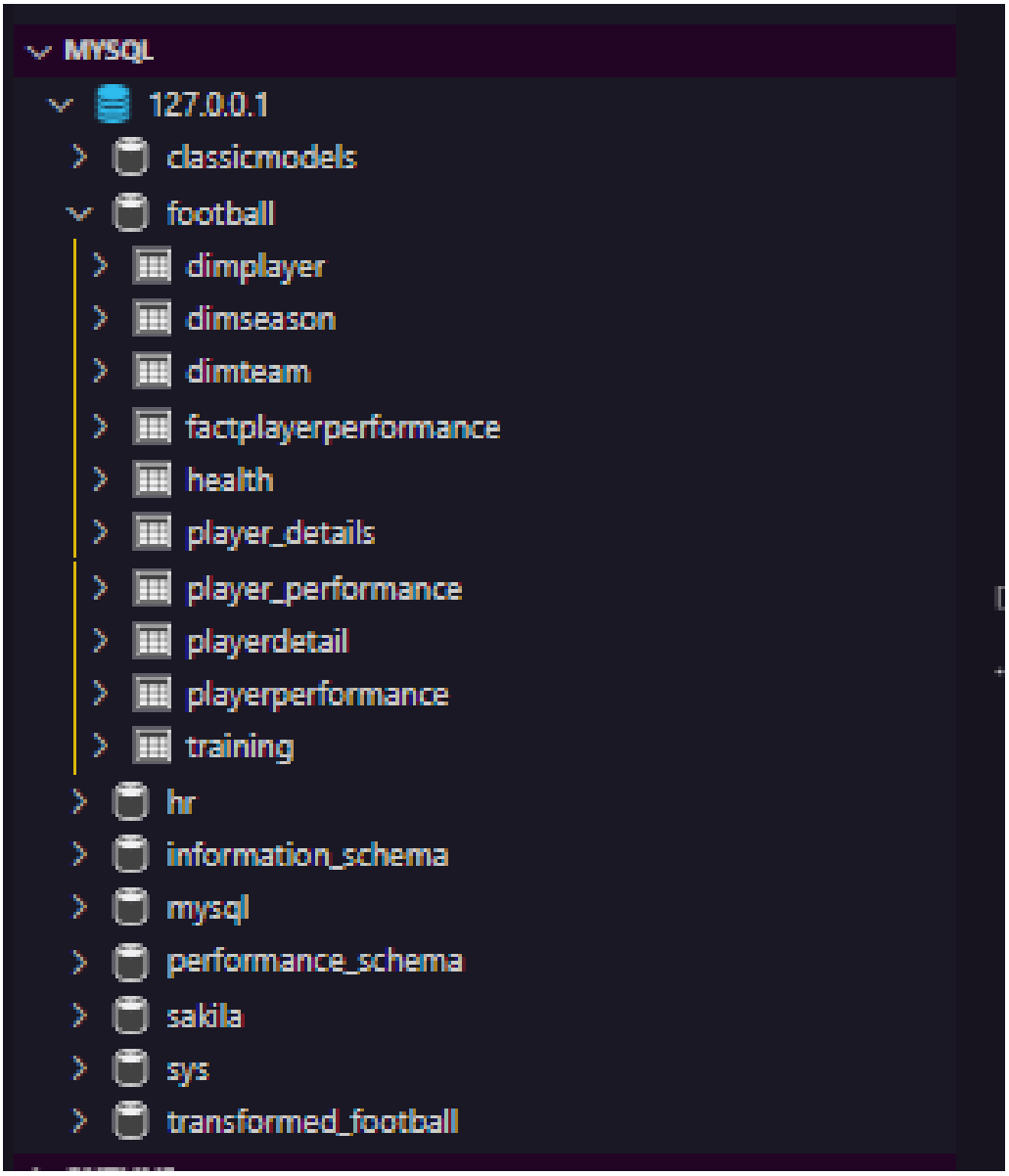

MYSQL
127.0.0.1
classicmodels
football
dimplayer
dimseason
dimteam
factplayerperformance
health
player_details
player_performance
playerdetail
playerperformance
training
hr
information_schema
mysql
performance_schema
sakila
sys
transformed_football

Fig 7.1

Chapter 8

8. Team Goals Analysis

8.1 Problem Statement

This task involves identifying the team with the highest number of goals and performing a time series analysis to understand goal-scoring trends over the season.

8.2 Approach

The analysis starts by grouping the dataset by team and summing the goals to identify the top goal-scoring team. Visualization techniques, such as horizontal bar plots and stacked bar charts, are used to depict team performance visually. These charts provide a clear comparison of goal counts across different teams.

A time series analysis is conducted to understand goal-scoring trends over the season. By resampling the data on a monthly basis, trends and patterns in goal-scoring can be identified. This analysis helps understand the dynamics of team performance over time, providing insights into peak performance periods and potential factors influencing these trends.

8.3 Results

The analysis identifies the team with the highest number of goals and provides a clear visualization of team performance. The time series analysis reveals trends and patterns in goal-scoring over the season, offering valuable insights for team strategy and performance evaluation. These findings can inform coaching decisions and enhance team preparation for future games.

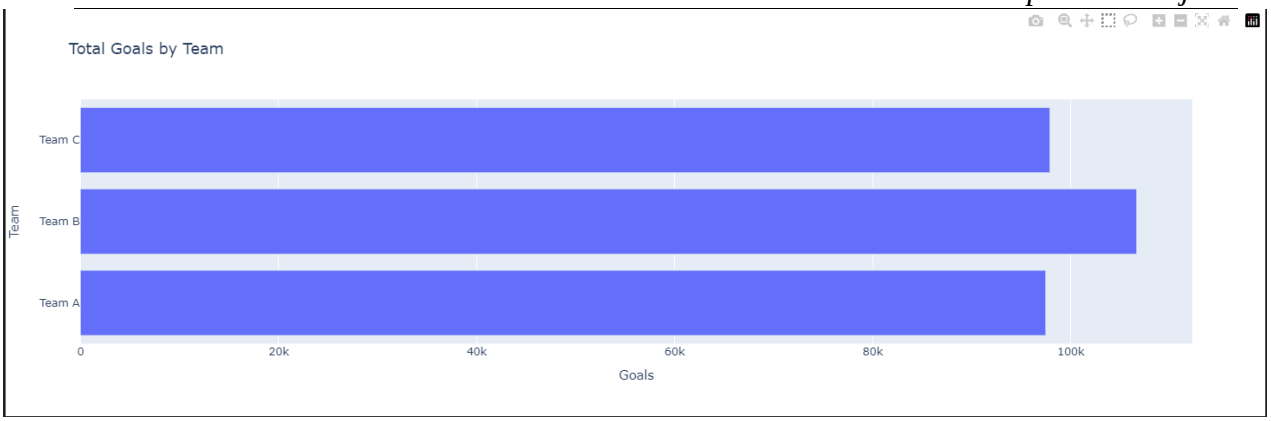


Fig 8.1

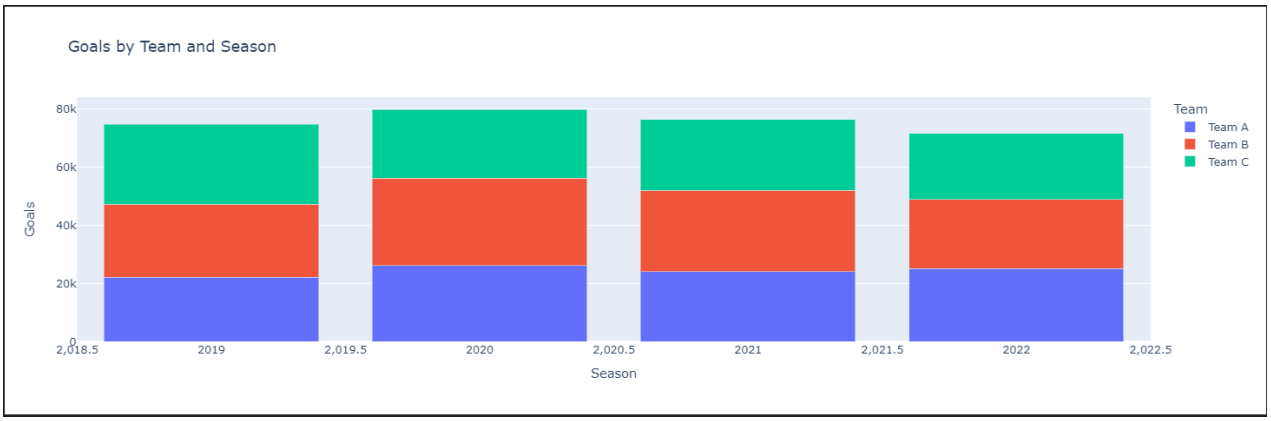


Fig 8.2

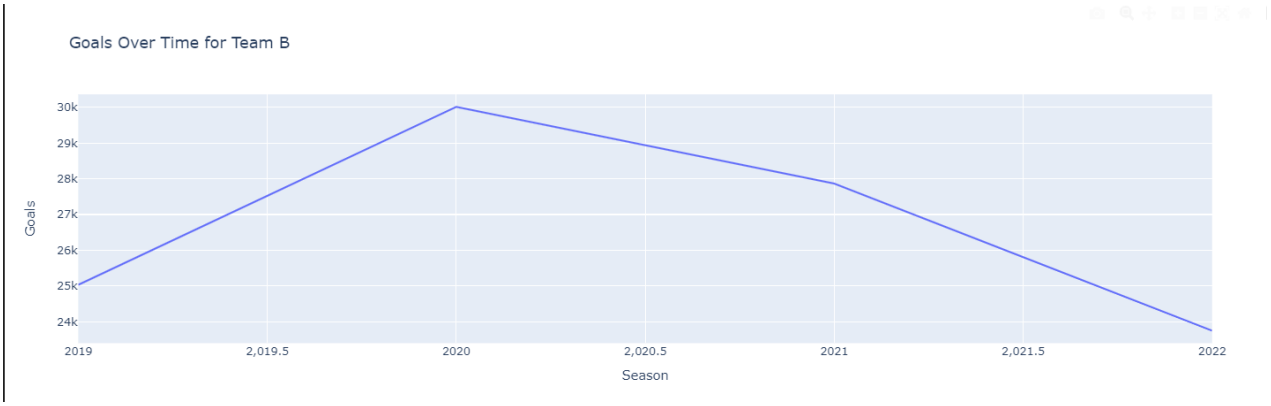


Fig 8.3

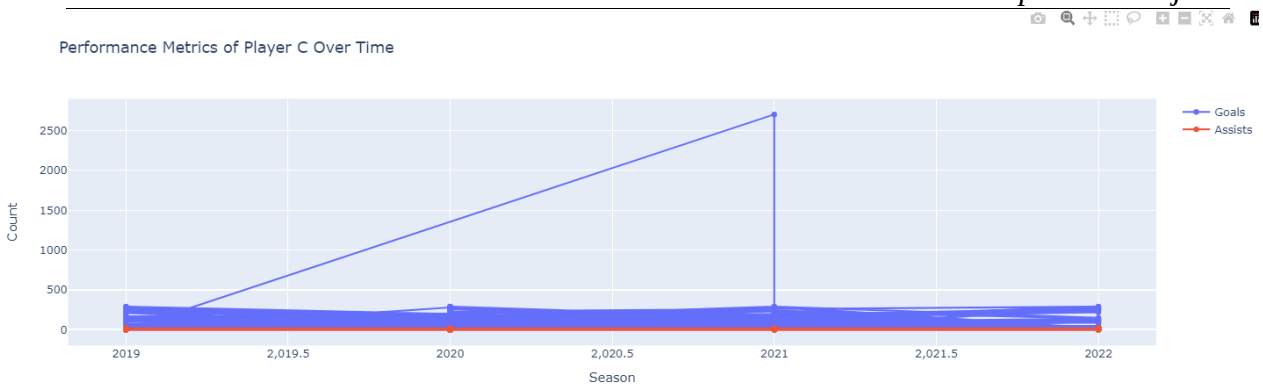


Fig 8.4

Chapter 9

9. Reporting and Visualization

9.1 Problem Statement

The final task is to develop interactive dashboards and visualizations that provide insights into player performance and team strategies. These visualizations should incorporate advanced analytics, such as clustering and predictive modeling, and be capable of integrating real-time data feeds.

9.2 Approach

Interactive dashboards are created using tools like Dash and Plotly, which allow for dynamic and engaging visualizations. Clustering techniques, such as K-means clustering, are used to group players with similar performance metrics, providing insights into player roles and strengths.

Predictive modeling techniques are applied to forecast future performance, leveraging historical data to make informed predictions. Real-time data integration ensures the dashboards are always up-to-date, providing the most current insights for decision-makers.

The dashboards include various visualizations, such as scatter plots to show relationships between performance metrics, bar charts to compare team performance, and line charts to track performance trends over time. These visualizations are designed to be interactive, allowing users to filter and explore the data in detail.

9.3 Results

The interactive dashboards provide a comprehensive view of player and team performance, leveraging advanced analytics to deliver deep insights. The real-time data integration ensures the information is current and relevant, enhancing strategic planning and performance evaluation. These tools empower stakeholders to make data-driven decisions, ultimately improving team performance and outcomes.

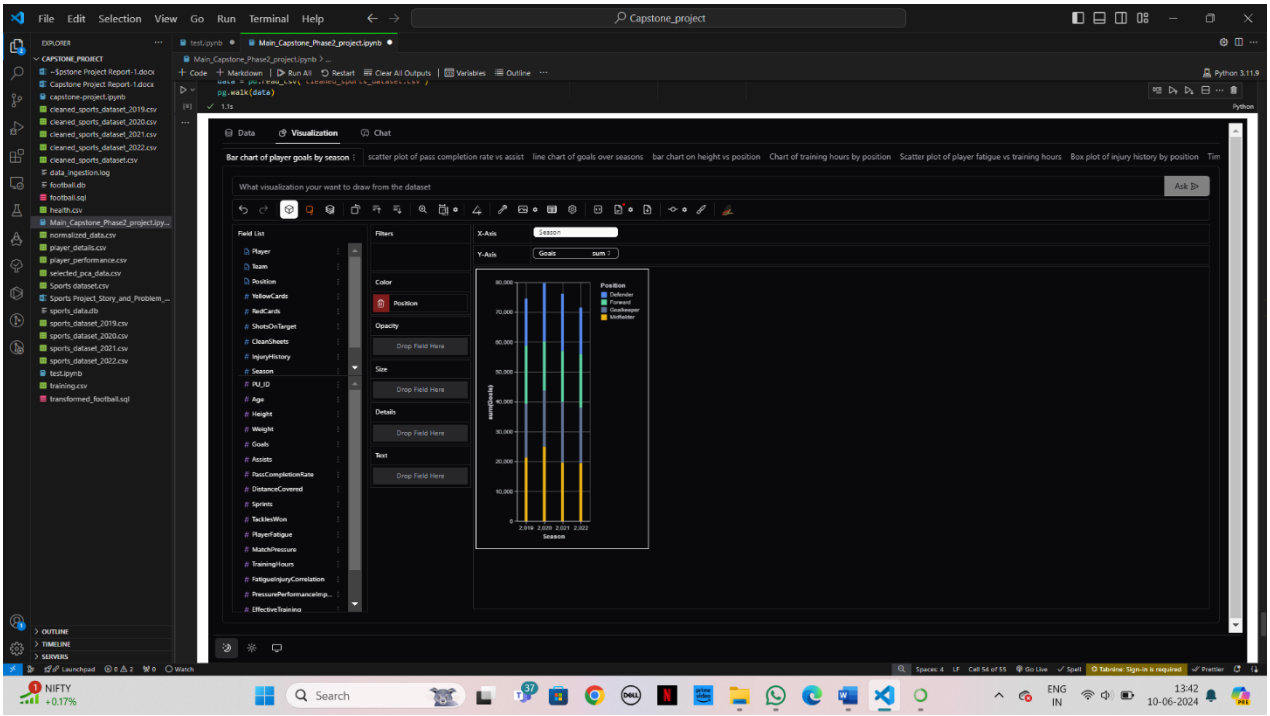


Fig 9.1

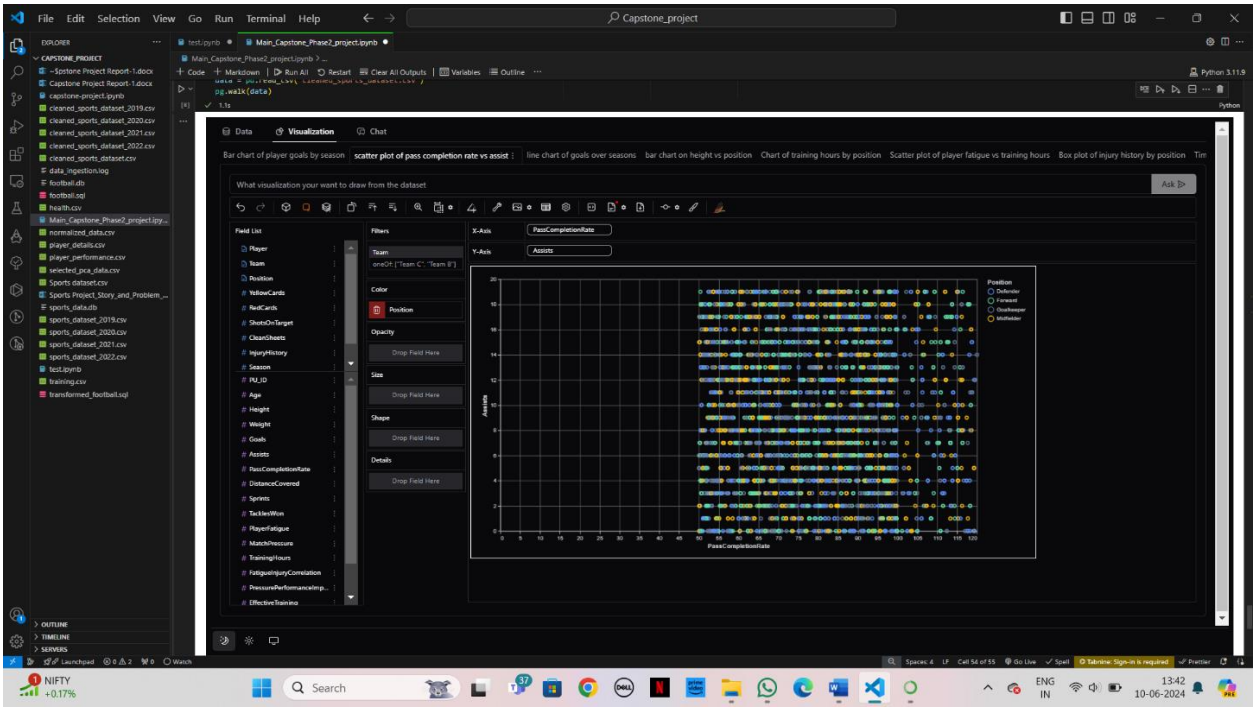


Fig 9.2

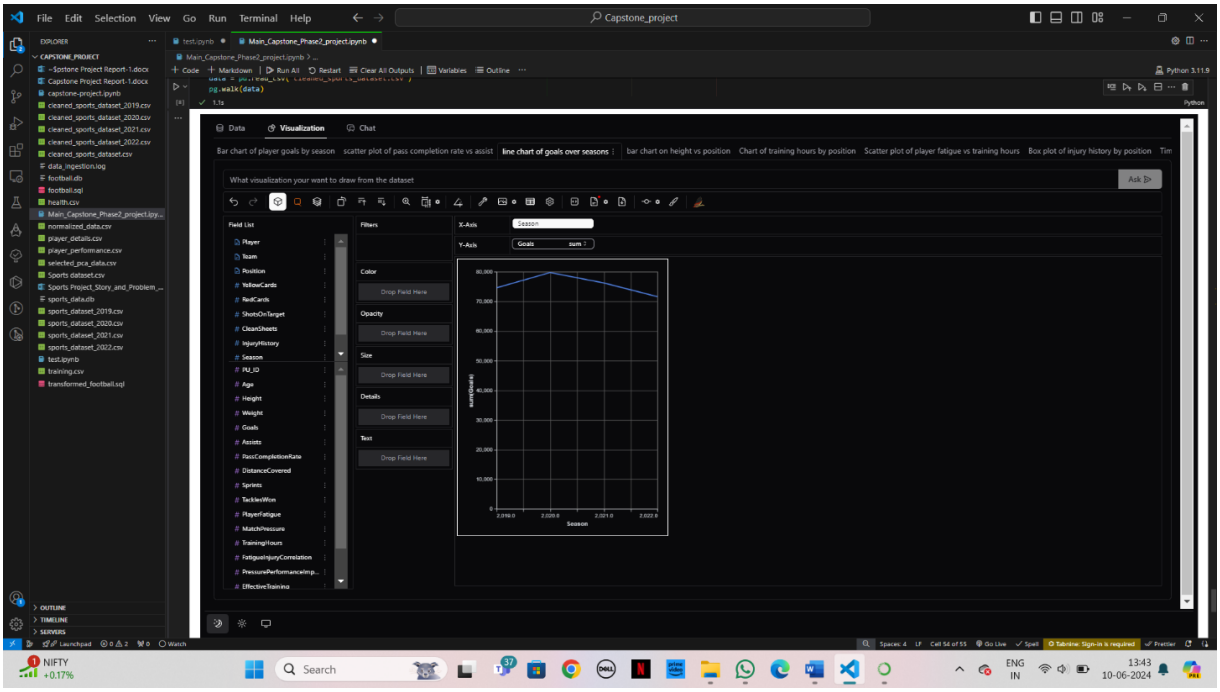


Fig 9.3

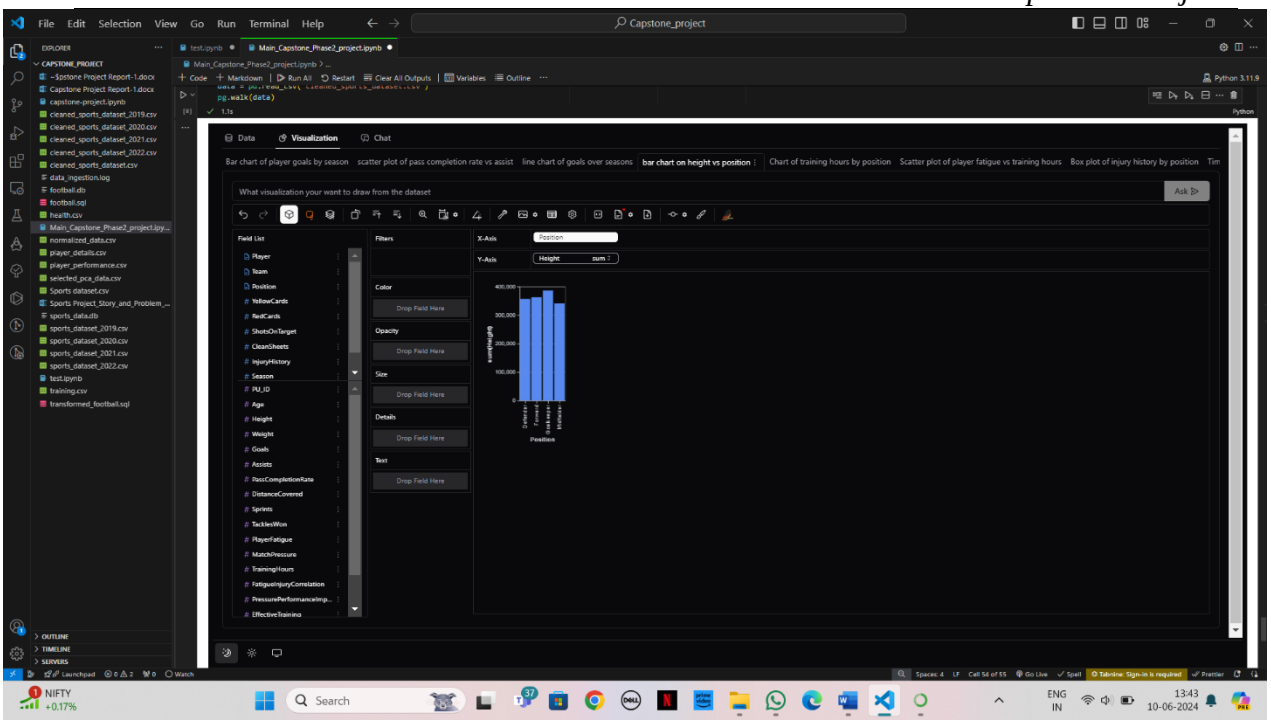


Fig 9.4

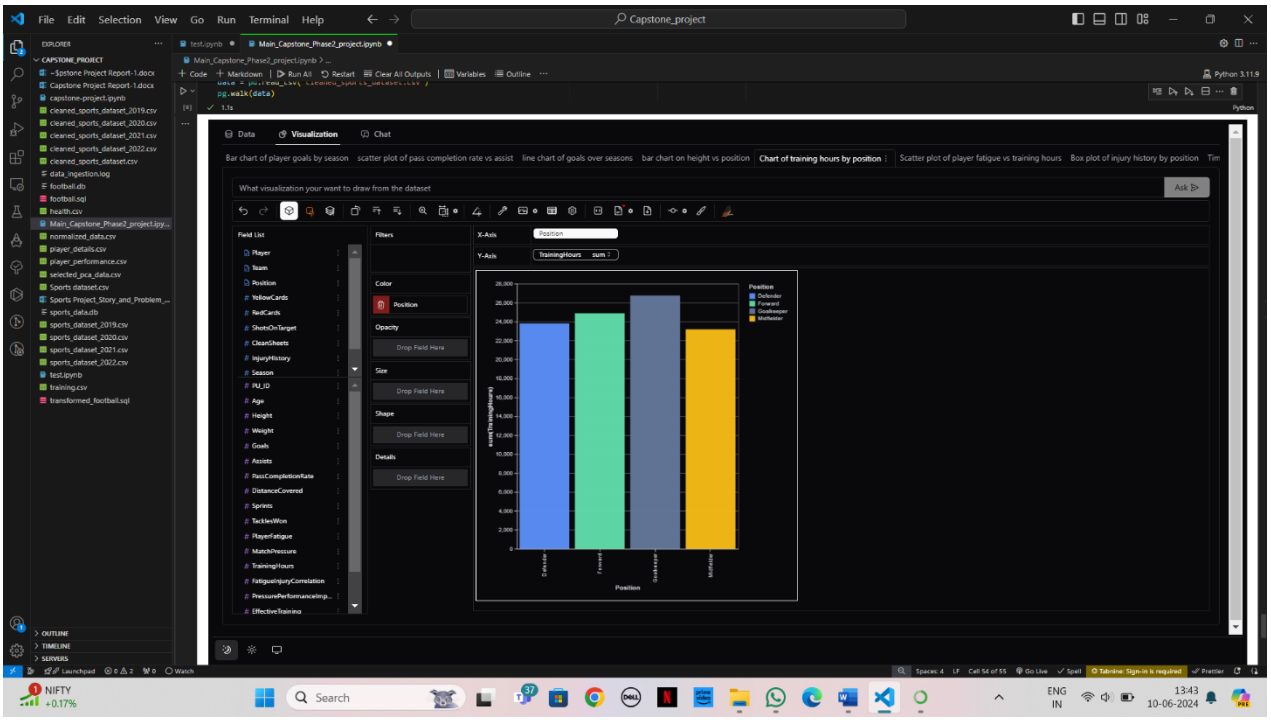


Fig 9.5

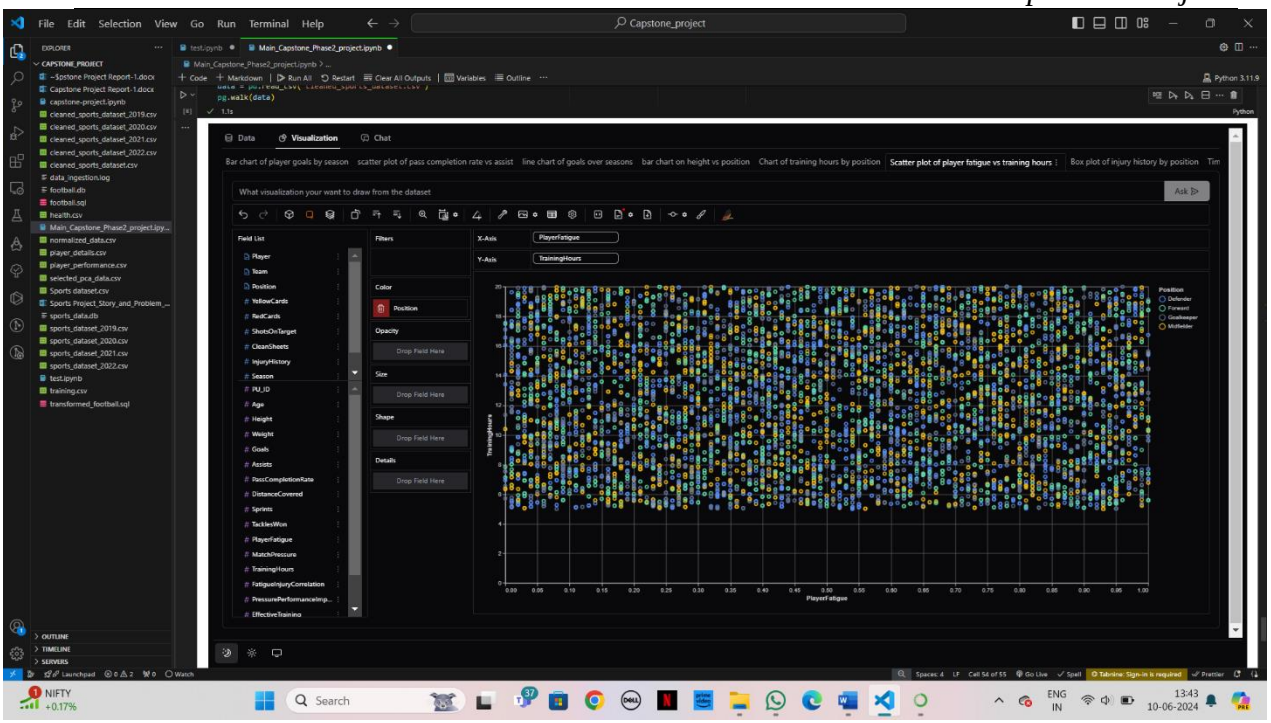


Fig 9.6

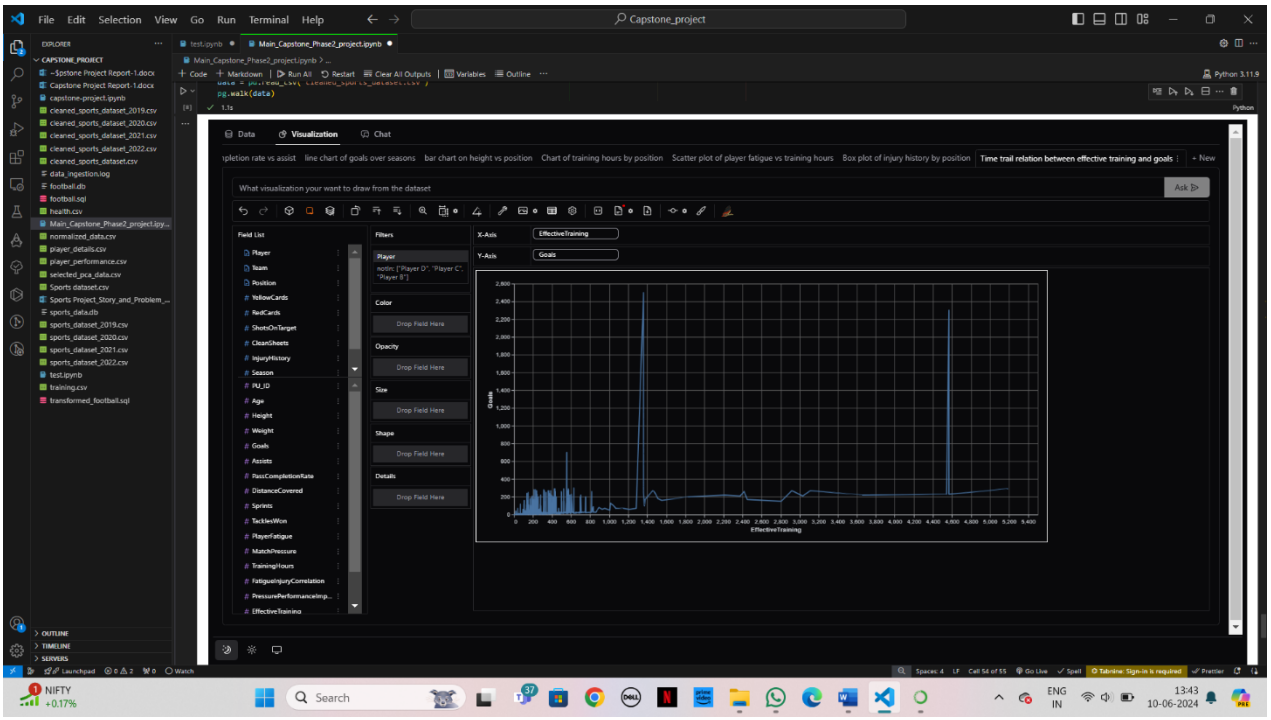


Fig 9.7

REFERENCES

1. "**Data Cleaning: Problems and Current Approaches**"
   Authors: Erhard Rahm, Hong Hai Do
   Link: [Data Cleaning: Problems and Current Approaches]
(https://www.researchgate.net/publication/220602798_Data_Cleaning_Problems_and_Current_Approaches)
2. "**Introduction to Data Mining**"
   Authors: Pang-Ning Tan, Michael Steinbach, Vipin Kumar
   Link: [Introduction to Data Mining]
(https://www-users.cs.umn.edu/~kumar/dmbook/index.php)
3. "**Machine Learning Yearning**"
   Author: Andrew Ng
   Link: [Machine Learning Yearning]
(http://www.mlyearning.org/)
4. "**Principles of Data Wrangling: Practical Techniques for Data Preparation**"
   Authors: Tye Rattenbury, Joseph M. Hellerstein, Jeffrey Heer
   Link: [Principles of Data Wrangling]
(https://www.oreilly.com/library/view/principles-of-data/9781491938923/)
5. "**The Elements of Statistical Learning: Data Mining, Inference, and Prediction**"
   Authors: Trevor Hastie, Robert Tibshirani, Jerome Friedman
   Link: [The Elements of Statistical Learning]
(https://web.stanford.edu/~hastie/ElemStatLearn//)
6. "**Pandas Documentation**"
   Authors: The Pandas Development Team
   Link: [Pandas Documentation]
(https://pandas.pydata.org/pandas-docs/stable/)
7. "**Altair: Declarative Visualization in Python**"
   Authors: Jake VanderPlas, Brian Granger, Jeffrey Heer, Dominik Moritz
   Link: [Altair Documentation]
(https://altair-viz.github.io/)
8. "**Design and Implementation of Data Warehouses**"
   Authors: Matteo Golfarelli, Stefano Rizzi
   Link: [Design and Implementation of Data Warehouses]
(https://link.springer.com/book/10.1007/978-3-540-26641-0)

APPENDIX-

SOURCE CODE

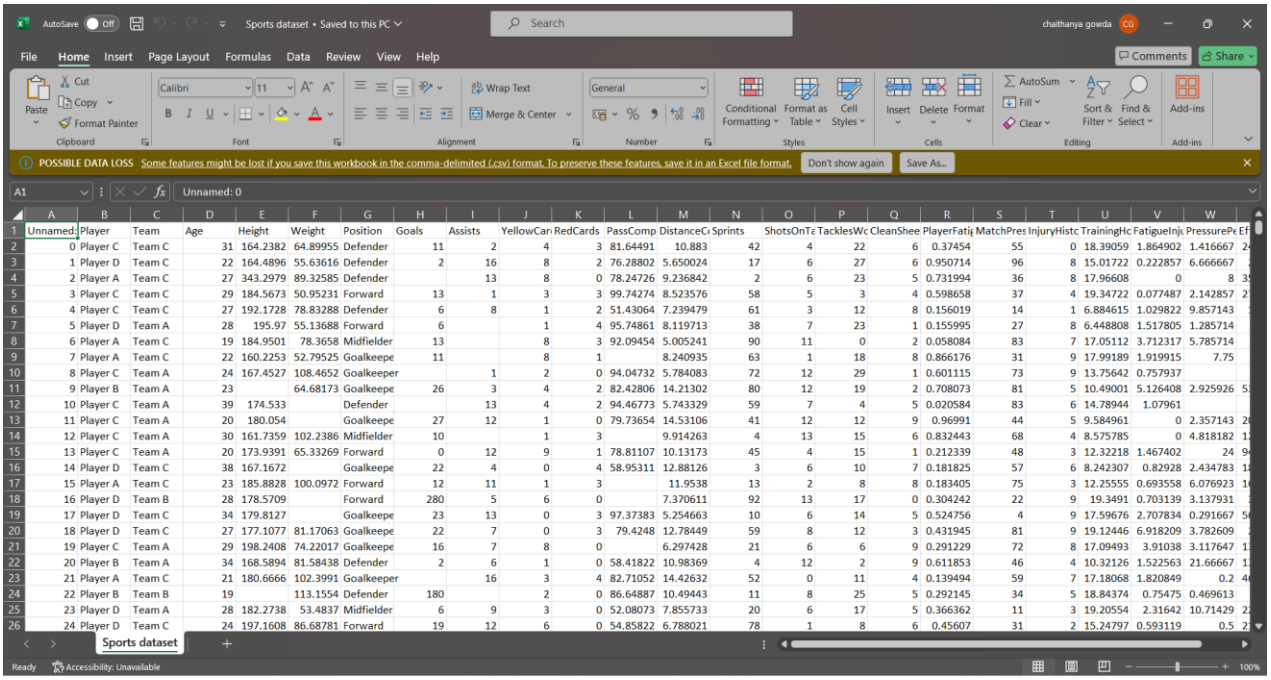
Ashwin B: https://github.com/ashwinbalaji05

BLN Wajith Ali: https://github.com/BLNWajith

Chaithanya Gowda L: https://github.com/chaithanyagow

APPENDIX -

DATASHEETS



Unnamed: 0	Player	Team	Age	Height	Weight	Position	Goals	Assists	YellowCar	RedCards	PassComp	DistanceCi	Sprints	ShotsOnTa	TacklesWc	CleanShee	PlayerFati	MatchPres	InjuryHistc	TrainingHc	FatigueInj	PressurePe	Ef
0	Player C	Team C	31	164.2382	64.89955	Defender	11	2	4	3	81.64491	10.883	42	4	22	6	0.37454	55	0	18.39059	1.864902	1.416667	24
1	Player D	Team C	22	164.4896	55.63616	Defender	2	16	8	2	76.28802	5.650024	17	6	27	6	0.950714	96	8	15.01722	0.222857	6.666667	1
2	Player A	Team C	27	343.2979	89.32585	Defender		13	8	0	78.24726	9.236842	2	6	23	5	0.731994	36	8	17.96608	0	8	3
3	Player C	Team C	29	184.5673	50.95231	Forward	13	1	3	3	99.74274	8.523576	58	5	3	4	0.598658	37	4	19.34722	0.077487	2.142857	2
4	Player C	Team C	27	192.1728	78.83288	Defender	6	8	1	2	51.43064	7.239479	61	3	12	8	0.156019	14	1	6.884615	1.029822	9.857143	
5	Player D	Team A	28	195.97	55.13688	Forward	6		1	4	95.74861	8.119713	38	7	23	1	0.155995	27	8	6.448808	1.517805	1.285714	
6	Player A	Team C	19	184.9501	78.3658	Midfielder	13		8	3	92.09454	5.005241	90	11	0	2	0.058084	83	7	17.05112	3.712317	5.785714	
7	Player A	Team C	22	160.2253	52.79525	Goalkeepe	11		8	1		8.240935	63	1	18	8	0.866176	31	9	17.99189	1.919915	7.75	
8	Player C	Team A	24	167.4527	108.4652	Goalkeeper		1	2	0	94.04732	5.784083	72	12	29	1	0.601115	73	9	13.75642	0.757937		
9	Player B	Team A	23		64.68173	Goalkeepe	26	3	4	2	82.42806	14.21302	80	12	19	2	0.708073	81	5	10.49001	5.126408	2.925926	5
10	Player C	Team A	39	174.533		Defender		13	4	2	94.46773	5.743329	59	7	4	5	0.020584	83	6	14.78944	1.07961		
11	Player C	Team A	20	180.054		Goalkeepe	27	12	1	0	79.73654	14.53106	41	12	12	9	0.96991	44	5	9.584961	0	2.357143	2
12	Player C	Team A	30	161.7359	102.2386	Midfielder	10		1	3		9.914263	4	13	15	6	0.832443	68	4	8.575785	0	4.818182	1
13	Player C	Team A	20	173.9391	65.33269	Forward	0	12	9	1	78.81107	10.13173	45	4	15	1	0.212339	48	3	12.32218	1.467402	24	9
14	Player D	Team C	38	167.1672		Goalkeepe	22	4	0	4	58.95311	12.88126	3	6	10	7	0.181825	57	6	8.242307	0.82928	2.434783	1
15	Player A	Team C	23	185.8828	100.0972	Forward	12	11	1	3		11.9538	13	2	8	8	0.183405	75	3	12.25555	0.693558	6.076923	1
16	Player D	Team B	28	178.5709		Forward	280	5	6	0		7.370611	92	13	17	0	0.304242	22	9	19.3491	0.703139	3.137931	
17	Player D	Team C	34	179.8127		Goalkeepe	23	13	0	3	97.37383	5.254663	10	6	14	5	0.524756	4	9	17.59676	2.707834	0.291667	5
18	Player D	Team C	27	177.1077	81.17063	Goalkeepe	22	7	0	3	79.4248	12.78449	59	8	12	3	0.431945	81	9	19.12446	6.918209	3.782609	
19	Player C	Team A	29	198.2408	74.22017	Goalkeepe	16	7	8	0		6.297428	21	6	6	9	0.291229	72	8	17.09493	3.91038	3.117647	1
20	Player B	Team A	34	168.5894	81.58438	Defender	2	6	1	0	58.41822	10.98369	4	12	2	9	0.611853	46	4	10.32126	1.522563	21.66667	1
21	Player A	Team C	21	180.6666	102.3991	Goalkeeper		16	3	4	82.71052	14.42632	52	0	11	4	0.139494	59	7	17.18068	1.820849	0.2	4
22	Player B	Team B	19		113.1554	Defender	180		2	0	86.64887	10.49443	11	8	25	5	0.292145	34	5	18.84374	0.75475	0.469613	
23	Player D	Team A	28	182.2738	53.4837	Midfielder	6	9	3	0	52.08073	7.855733	20	6	17	5	0.366362	11	3	19.20554	2.31642	10.71429	2
24	Player D	Team C	24	197.1608	86.68781	Forward	19	12	6	0	54.85822	6.788021	78	1	8	6	0.45607	31	2	15.24797	0.593119	0.5	2

INFORMATION REGARDING STUDENT(S)

STUDENT NAME	EMAIL ID	PHONE NUMBER
ASHWIN B	22btrad006@jainuniversity.ac.in	9945691004
B.L.N. WAJITH ALI	22btrad009@jainuniversity.ac.in	6366183482
Chaithanya Gowda L	22btrad010@jainuniversity	9380082604