



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

“Comprehensive Performance Analysis of Sports Players”

Submitted in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

M Jeyapathy, Durgam Naga Deeraj Reddy, Eshita Katyal
22BTRAD016, 22BTRAD013, 22BTRAD014

Under the guidance of

Mr. Akash Das

Project Practice Head and Mentor

Department of Computer Science and Engineering
School of Computer Science & Engineering
Faculty of Engineering & Technology
JAIN (Deemed to-be University)



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 “**Comprehensive Performance Analysis of Sports Players**” is carried out by **M Jeyapathy (22BTRAD016), Durgam Naga Deeraj Reddy (22BTRAD013), Eshita Katyal (22BTRAD014)**, a bonafide student(s) of Bachelor / Master 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 / Master 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

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 & Technology
JAIN (Deemed to-be University)

Date: 10/06/2024

Name of the Examiner

Signature of Examiner

1.

2.

DECLARATION

I/ We , **M Jeyapathy (22BTRAD016), Durgam Naga Deeraj Reddy (22BTRAD013), Eshita Katyal (22BTRAD014)** student of 4th semester B.Tech/ M.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 “**Comprehensive Performance Analysis of Sports Players**” has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor /Master 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.

Durgam Naga Deeraj Reddy
22BTRAD013

Signature

Eshita Katyal
22BTRAD014

Signature

M Jeyapathy
22BTRAD016

Signature

Place: Bangalore

Date: 10/06/2024

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. 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 Futureense Technologies** for sparing his valuable time to extend help in every step of my work, which paved the way for smooth progress and fruitful culmination of the project.*

I am also grateful to my family and friends who provided me with every requirement throughout the course. I would like to thank one and all who directly or indirectly helped me in completing the work successfully.

Signature of Student(s)

ABSTRACT

Through this study, we conducted a Comprehensive Performance Analysis of Sports Players to enhance insights into player performance and team strategies via meticulous data management and analysis. The project is divided into four distinct phases: getting familiar with the data, implementing data ingestion strategies, performing advanced data transformations, and ultimately, reporting and visualizing the findings.

In the first phase, we focus on thoroughly cleaning and augmenting the dataset. This includes identifying and addressing missing values, detecting and investigating outliers, and conducting a detailed analysis of player positions to understand their unique roles and contributions. Additionally, we perform preliminary statistical analyses to establish baseline insights.

The second phase involves creating robust data pipelines to ensure seamless data ingestion and processing. Here, we explore relationships between various performance metrics, such as the correlation between pass completion rates and assists, using advanced statistical methods to uncover deeper insights. This phase also includes integrating data from multiple sources to create a comprehensive dataset.

As we progress to the third phase, the emphasis shifts to complex data transformations. We design and implement a data warehouse to support deep analytical queries, enabling efficient data retrieval and analysis. This involves feature engineering to create new variables that enhance the predictive power of our models, normalizing data to ensure consistency, and implementing stringent data security measures to protect sensitive information.

The final phase focuses on creating interactive dashboards and visualizations that facilitate informed decision-making. We utilize advanced analytics and real-time data integration to keep insights current and actionable. These visualizations are designed to be intuitive and impactful, guiding business decisions with clarity and precision. Throughout the project, we measure success by the accuracy and completeness of our data, the efficiency of our data pipelines, the relevance and effectiveness of our data transformations, and the clarity and impact of our visualizations. Ultimately, our goal is to enable strategic decision-making in sports analytics, providing stakeholders with the tools and insights needed to optimize player performance and refine team strategies.

TABLE OF CONTENTS

“Comprehensive Performance Analysis of Sports Players”

1. Chapter 1 - (Data Cleaning and Augmentation)	01
1.1 Duplicates	01
1.2 Null Values	01
1.3 Outliers	02
2. Chapter 2 – (Position Analysis)	03
2.1 Analysis by Positions	03
3. Chapter 3 – (Data Ingestion Strategies)	05
4. Chapter 4 – (Pass Completion Rates vs Assists)	08
4.1 Best Fit line	08
5. Chapter 5 – (Advanced Data Transformations)	10
5.1 Feature Engineering	10
5.2 Normalisation	11
5.3 PCA (Dimensionality Reduction)	11
6. Chapter 6 – (Data Warehousing)	13
7. Chapter 7 – (Team Goal Analysis)	16
7.1 Comparison of Goal Counts	16
7.2 Season by Season Trend	18

8. Chapter 8 – (Reporting & Visualisation)	19
8.1 Sum of Assists and Goals by Player (Line Chart)	19
8.2 Average training hours of each team	20
8.3 Sum of Injury history by team	20
8.4 Player effect Training with its average Training hours	21
8.5 Sum of Assists and Goals by Player (Stacked Column Chart)	22
8.6 Sum of Yellow and Red cards by player and Team	23
Conclusion	24

LIST OF FIGURES

Fig. No.	Description of the figure	Page No.
2.1(a)	Analysis by Positions of players	3
2.1(b)	Observed vs Expected player Distribution by Position	4
3(a)	Managing SQL Server Connection	5
4.1(a)	Line of Best Fit	8
4.1(b)	Pass Completion vs Assists (Outliers Highlighted)	9
5.3(a)	Explained Variance Ratio by PCA components	12
6(b)	Child tables stored in SQL workbench server	14
7.1(a)	Team Goal Comparison	16
7.1(b)	Team Goals by Season	17
7.2(a)	Performance metrics over time for Player A	18
8.1(a)	Sum of Assists and Goals by Player	19
8.2(a)	Average Training Hours of Each Team	20
8.3(a)	Sum of Injury History by Team	21
8.4(a)	Players Effect Training with its average Training Hours	21
8.5(a)	Sum of Assists and Goals by Player	22
8.6(a)	Sum of Red Cards and Yellow Cards by Player and Team	23

CHAPTER 1

1. Data Cleaning and Augmentation

Utilizing advanced imputation techniques to effectively handle missing values and employing statistical methods along with domain knowledge to identify and correct anomalies by detecting outliers. Ensuring data consistency, the project will standardize data formats across the dataset. Additionally, the dataset will be augmented by generating synthetic data through data augmentation techniques and by incorporating additional data collected from public sports databases. All these elements will be integrated into a unified, cohesive dataset to support robust analytical endeavours.

The dataset we received posed significant challenges, as out of 20,000 rows, only 5,000 were unique, with a substantial number of duplicates and outliers. Addressing these issues, particularly deciding whether to drop the outliers or impute them with the mean, proved to be a difficult task for our team.

1.1. DUPLICATES

The dataset initially contained 3,372 duplicates, which were removed during preprocessing. After this step, I was left with 16,629 unique values for my analysis.

```
1 # Check for duplicated rows based on all columns
2 duplicated_rows = df.duplicated()
3
4 # Print the duplicate rows
5 print("Duplicate Rows :\n", df[duplicated_rows])
```

Python

```
[3372 rows x 25 columns]
```

1.2. NULL VALUES

The dataset contained numerous null values causing our team to deliberate between removing these values or imputing them. Removing the null values would significantly reduce the number of rows, complicating further analysis. Ultimately, we decided to impute the missing values with mean values.

However, this led to another issue: fields like goals and cards were filled with float values, which is unrealistic. To address this, we rounded these values to the nearest integers.

1.3. OUTLIERS

Handling outliers presented another significant challenge for our team, as their presence was crucial for visualizing the dataset before and after analysis. Removing outliers would hinder the comparison of the original and processed data. Therefore, instead of eliminating them, we chose to replace the outliers with mean values. This approach allowed us to retain the dataset's overall structure while mitigating the impact of extreme values, ensuring a smoother and more accurate analysis.

Finally, all this was kept in a CSV file named as `cleaned_sports_dataset.csv` and for all further analysis this file was used.

CHAPTER 2

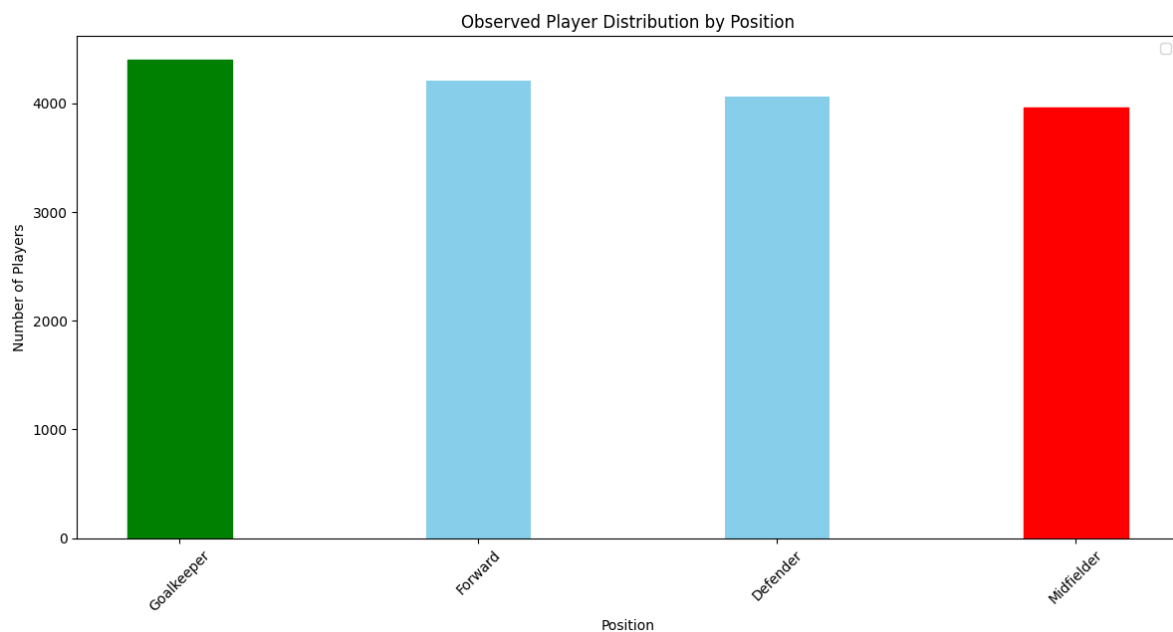
2. Position Analysis

The goal of position analysis in the context is to understand the distribution of players across different positions on the field (like Defender, Forward, Midfielder, Goalkeeper). Analysis is based on:

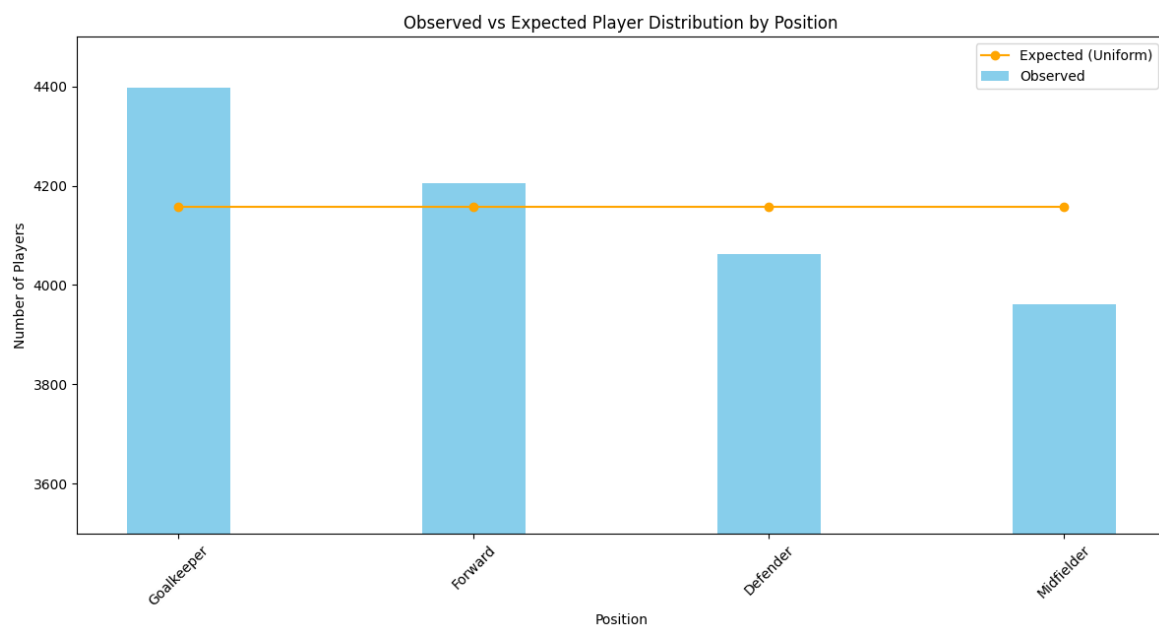
- Which position has the most players?
- Which position has the fewest players?
- Is the distribution of players across positions even or skewed?

Does the distribution of players across positions significantly differ from what we'd expect if the distribution was completely random (uniform)?

2.1. ANALYSIS BY POSITIONS



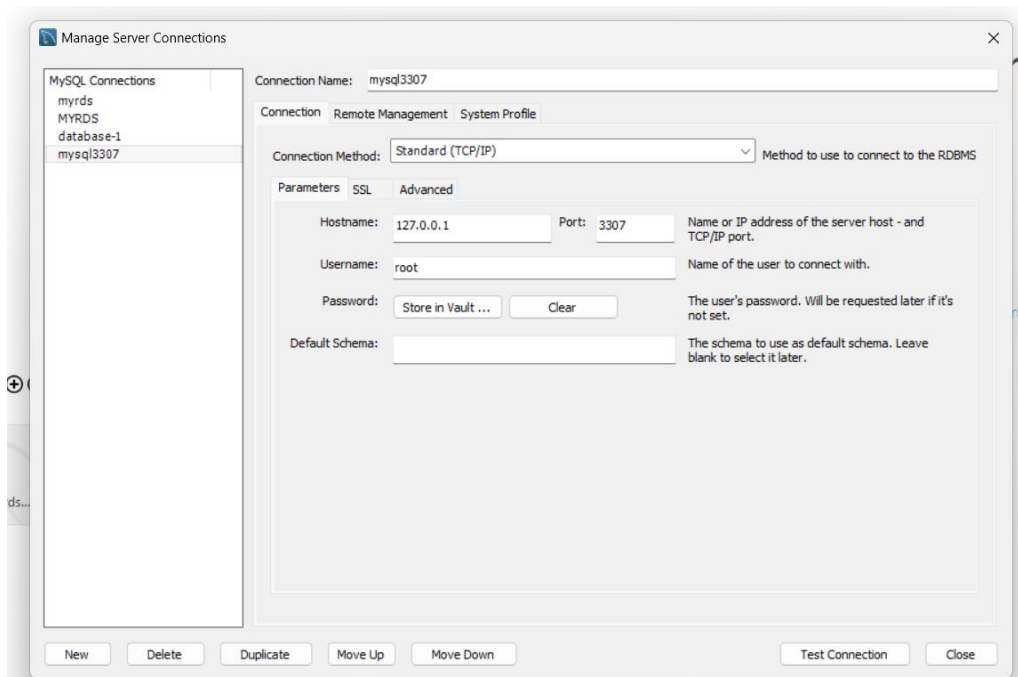
Bar plot that compares the observed frequency of players per position with the expected frequency under a uniform distribution. The x-axis shows different player positions, while the y-axis displays the number of players. Sky blue bars represent the observed player counts, and an orange line with markers represents the uniform expected counts. The y-axis is set between 3,500 and 4,500 to enhance visualization. This plot highlights differences between the actual and expected player distributions, indicating positional imbalances in the dataset.



CHAPTER 3

3. Data Ingestion Strategies

With a clean and robust dataset at our disposal, it is time to design and implement effective data ingestion strategies using Python, pandas, and SQL to ensure efficient data storage and management. Although I haven't yet built a pipeline model or used Python, we have directly worked in a SQL workbench and extracted data. We will now architect a comprehensive data pipeline that seamlessly ingests, processes, and stores data, leveraging Python for scripting, pandas for data manipulation, and SQL for database interactions. This pipeline will automate the ingestion process, maintain data integrity, and ensure that data is readily accessible for analysis. By focusing on optimizing performance to handle large volumes of data efficiently, we aim to streamline our data operations and facilitate better decision-making through reliable and well-managed data.



Comprehensive Performance Analysis of Sports Players

127.0.0.1
The hostname of the database (Press 'Enter' to confirm or 'Escape' to cancel)

Jeya
The MySQL user to authenticate as (Press 'Enter' to confirm or 'Escape' to cancel)

....
The password of the MySQL user (Press 'Enter' to confirm or 'Escape' to cancel)

Our team has successfully connected SQL to VS Code by entering the username, password, and host ID. After establishing the connection, we created a database that contains the complete dataset with all its rows and columns.

The screenshot shows the VS Code interface with a MySQL connection established. The Explorer panel on the left shows the project structure, including a 'MySQL' section with a '127.0.0.1' connection. The SQL editor in the center contains the following code:

```
use sports_dataset;
select * from sports_data;
```

The Results panel on the right displays a table with 10 rows and 13 columns. The columns are: Sno, Unnamed, Player, Team, Age, Height, Weight, Position, Goals, Assists, Yellow_Cards, Red_Cards, and Pass_Cor. The data is as follows:

Sno	Unnamed	Player	Team	Age	Height	Weight	Position	Goals	Assists	Yellow_Cards	Red_Cards	Pass_Cor
1	0	Player C	Team C	31	164.2382365	64.89955397	Defender	11	2	4	3	81.64491
2	1	Player D	Team C	22	164.4896429	55.6361591	Defender	2	16	8	2	76.28801
3	2	Player A	Team C	27	182.03972267048056	89.32584969	Defender	38	13	8	0	78.24725
4	3	Player C	Team C	29	184.567349	50.95230768	Forward	13	1	3	3	99.74274
5	4	Player C	Team C	27	192.172813	78.83288083	Defender	6	8	1	2	51.43063
6	5	Player D	Team A	28	195.9700123	55.13688117	Forward	6	9	1	4	95.74860
7	6	Player A	Team C	19	184.9501354	78.36580481	Midfielder	13	9	8	3	92.09453
8	7	Player A	Team C	22	160.2252525	52.79525325	Goalkeeper	11	9	8	1	77.16600
9	8	Player C	Team A	24	167.4526711	108.4652082	Goalkeeper	38	1	2	0	94.04731
10	9	Player B	Team A	23	188.4789287911	64.68172958	Goalkeeper	26	3	4	2	82.42806

The Output panel at the bottom shows the execution of the SQL query, indicating that the query was successful and returned 10 rows.

```
1 • use sports_dataset;  
2 • CREATE TABLE PlayerInfo (  
3     Srno INTEGER PRIMARY KEY auto_increment,  
4     Player TEXT,  
5     Team TEXT,  
6     Age INTEGER,  
7     Height INTEGER,  
8     Weight INTEGER,  
9     Position TEXT,  
10    BMI real  
11 );
```

```
• CREATE TABLE PerformanceStats (  
    Srno Integer primary key auto_increment,  
    Player TEXT,  
    Goals INTEGER,  
    Assists INTEGER,  
    YellowCards INTEGER,  
    RedCards INTEGER,  
    PassCompletionRate REAL,  
    DistanceCovered REAL,  
    Sprints INTEGER,  
    ShotsOnTarget INTEGER,  
    TacklesWon INTEGER,  
    CleanSheets INTEGER,  
    GoalContribution INTEGER,  
    Season Integer  
);
```

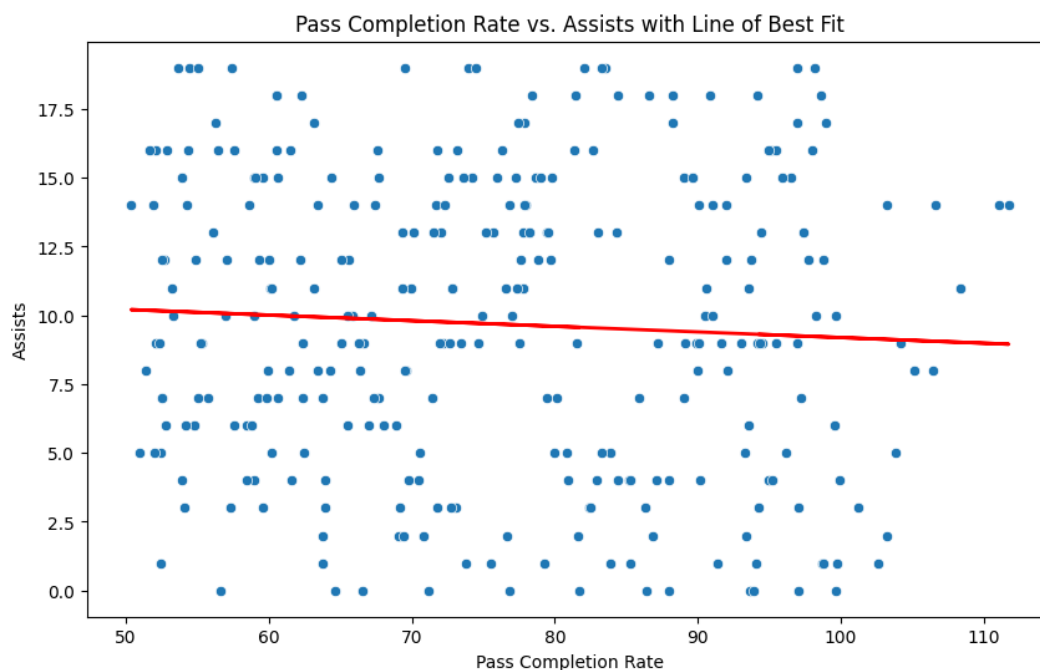
```
• CREATE TABLE HealthFitness (  
    Srno Integer primary key auto_increment,  
    Player TEXT,  
    PlayerFatigue REAL,  
    MatchPressure REAL,  
    InjuryHistory TEXT,  
    TrainingHours INTEGER,  
    FatigueInjuryCorrelation REAL,  
    PressurePerformanceImpact REAL,  
    EffectiveTraining TEXT  
);
```

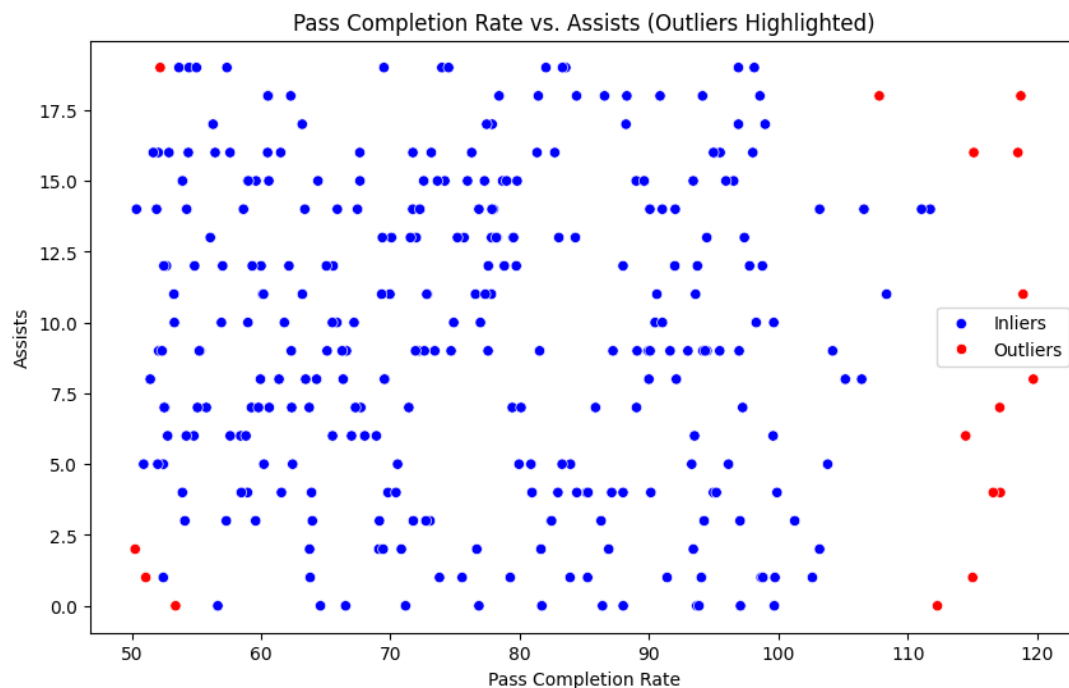
CHAPTER 4

4. Pass Completion Rates vs Assists

We delved into analysing the relationship between pass completion rate and assists. Our initial step involved creating a scatter plot to visually depict this relationship, followed by the implementation of advanced outlier detection methods like DBSCAN or Isolation Forest to identify any outliers within the data. Incorporating a line of best fit onto the scatter plot allowed us to model the relationship between these two variables effectively. Subsequently, we performed regression analysis to quantitatively assess this relationship. Finally, we evaluated the model using appropriate metrics to gauge its accuracy and efficacy in predicting the association between pass completion rate and assists.

4.1. BEST-FIT LINE





The scatter plot is updated to include a line of best fit representing the linear relationship between the two variables. Finally, the performance of the regression model is evaluated using the R-squared metric, which measures the proportion of variance in the dependent variable (Assists) that is predictable from the independent variable (Pass Completion Rate). This analysis provides insight into how well the Pass Completion Rate explains the variation in Assists within the dataset, with a higher R-squared value indicating a better fit of the model to the data.

CHAPTER 5

5. Advanced Data Transformations

```
1 df=pd.read_csv("new_dataset.csv")
2 (variable) df: DataFrame
3 df['GoalCont'] = df['Goals'] + df['Assists']
4 df['BMI'] = df['Weight'] / ((df['Height']/100) ** 2)
5
6 # Drop rows with NaN values generated from the feature engineering step
7 df.dropna(subset=['GoalContribution', 'BMI'], inplace=True)
8
9 print("Data after feature engineering:")
10 print(df[['GoalContribution', 'BMI']].head())
```

Python

We conducted an extensive data transformation process, incorporating advanced techniques such as feature engineering to derive new insightful features from the dataset. Additionally, we optimized the data further by employing strategies like normalization and dimensionality reduction. Through feature engineering, we created additional meaningful features like "GoalContribution" & "BMI", providing deeper insights into player performance. Normalization ensured consistent scaling across all features, preventing dominance by any single metric, while dimensionality reduction techniques such as PCA streamlined analysis and enhanced model interpretability by identifying significant components of variation. These approaches collectively facilitated more effective comparison, reduced overfitting risks, and laid a robust foundation for building predictive models and extracting actionable insights from the sports dataset.

5.1. FEATURE ENGINEERING

In our data enhancement efforts, we introduced two new features to augment the dataset's richness and utility. Firstly, we created the 'GoalContribution' feature by aggregating the 'Goals' and 'Assists' columns, capturing the combined offensive impact of each player. Secondly, we engineered the 'BMI' feature, representing Body Mass Index, calculated as the player's weight divided by the square of their height. Given that height was initially provided in centimeters, we converted it to meters to ensure consistency in BMI calculation. Additionally, to improve dataset organization and facilitate tracking, we appended serial numbers to each entry. These modifications were meticulously integrated into the dataset and subsequently saved as a CSV file, ensuring that the enriched data is readily accessible for further analysis and modeling endeavors.

5.2. NORMALISATION

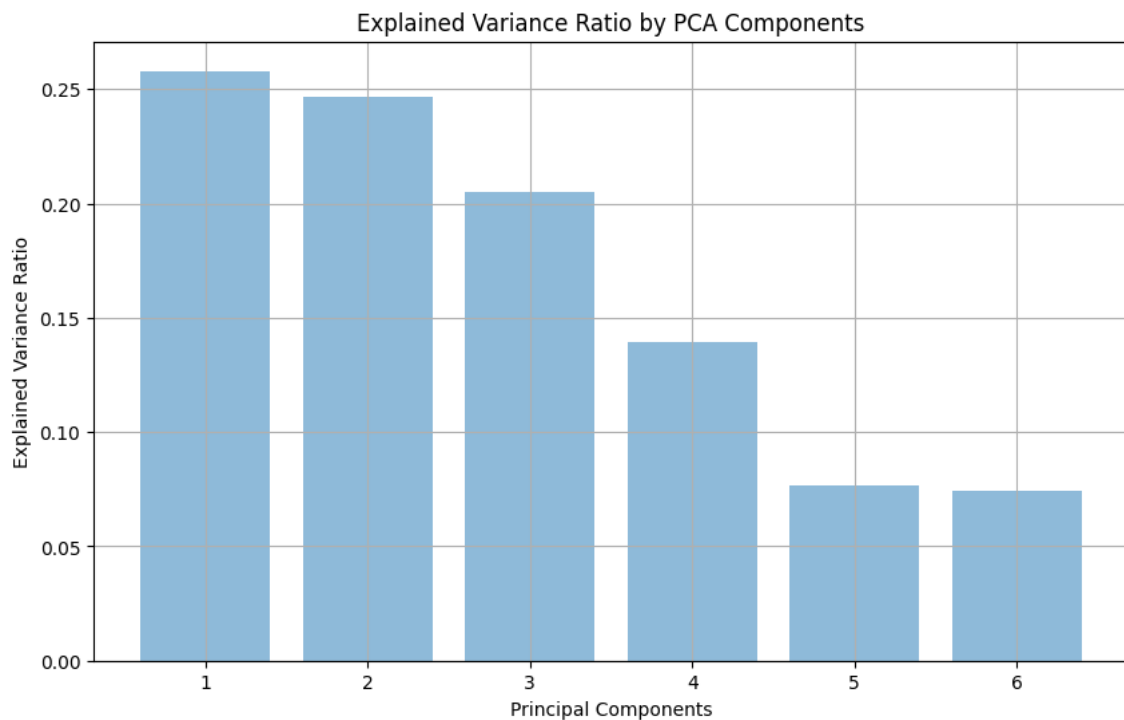
After enriching the dataset with new features and adjustments, we proceeded to enhance its usability by normalizing selected attributes. Utilizing the MinMaxScaler from the sklearn library, we normalized specific features including 'Height', 'Weight', 'PassCompletionRate', 'DistanceCovered', 'GoalContribution', and 'TacklesWon'. Normalization ensures that these features are on a consistent scale, ranging between 0 and 1, which is particularly useful for algorithms sensitive to varying feature magnitudes.

The process involved transforming each feature such that its minimum value becomes 0 and its maximum value becomes 1, preserving the relative relationships between data points. By normalizing these attributes, we facilitate more effective comparisons and analyses while maintaining the integrity of the dataset's information.

5.3. PCA (DIMENSIONALITY REDUCTION)

Executed Principal Component Analysis (PCA) on the normalized dataset, aiming to retain five principal components that collectively elucidate a substantial proportion of the data's variance. This strategic reduction in dimensionality enables us to distill essential patterns and structures from the original features while preserving the dataset's integrity. The resulting PCA components, denoted as PC1 through PC5, represent coherent linear combinations of the initial attributes, effectively summarizing the most salient sources of variation. These components, orthogonal to each other, offer a streamlined representation of the dataset, facilitating more efficient analysis and interpretation of the underlying data patterns.

Comprehensive Performance Analysis of Sports Players



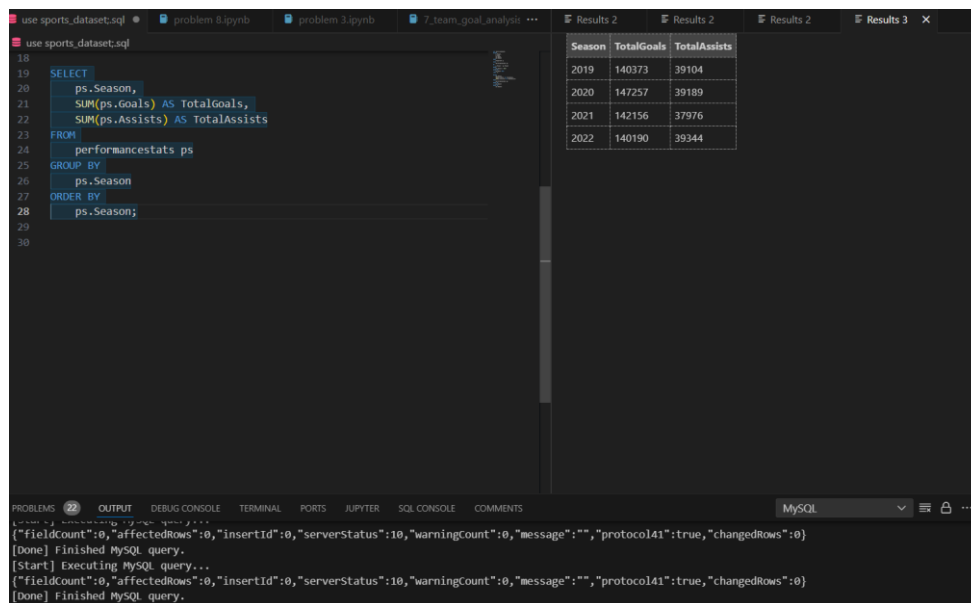
Following this, we computed and presented the explained variance ratio for each principal component, illuminating the extent to which each component captures the dataset's variability independently. Additionally, we calculated the cumulative variance to showcase the cumulative contribution of the components to the dataset's overall variance. To aid in determining the optimal number of components to retain, we visualized the explained variance ratio for each component using a bar plot. This visualization serves as a crucial guide for striking a balance between dimensionality reduction and information preservation, empowering informed decision-making regarding the appropriate number of principal components to retain for subsequent analyses.

CHAPTER 6

6. Data Warehousing

Our team has successfully connected SQL to VS Code by entering the username, password, and host ID. After establishing the connection, we created a database that contains the complete dataset with all its rows and columns. Next, we will design and implement a comprehensive data warehouse schema utilizing advanced SQL features such as window functions. This approach will enable us to store the transformed data efficiently, ensuring that it can support complex analytical queries. Additionally, we will implement robust data security and access control mechanisms to safeguard the integrity and confidentiality of our data.

By focusing on these advanced techniques, we aim to optimize our data storage and management processes, facilitating more insightful and accurate analyses.



The screenshot shows a VS Code editor with a MySQL query in the left pane and its results in the right pane. The query is as follows:

```

SELECT
  ps.Season,
  SUM(ps.Goals) AS TotalGoals,
  SUM(ps.Assists) AS TotalAssists
FROM
  performancestats ps
GROUP BY
  ps.Season
ORDER BY
  ps.Season;

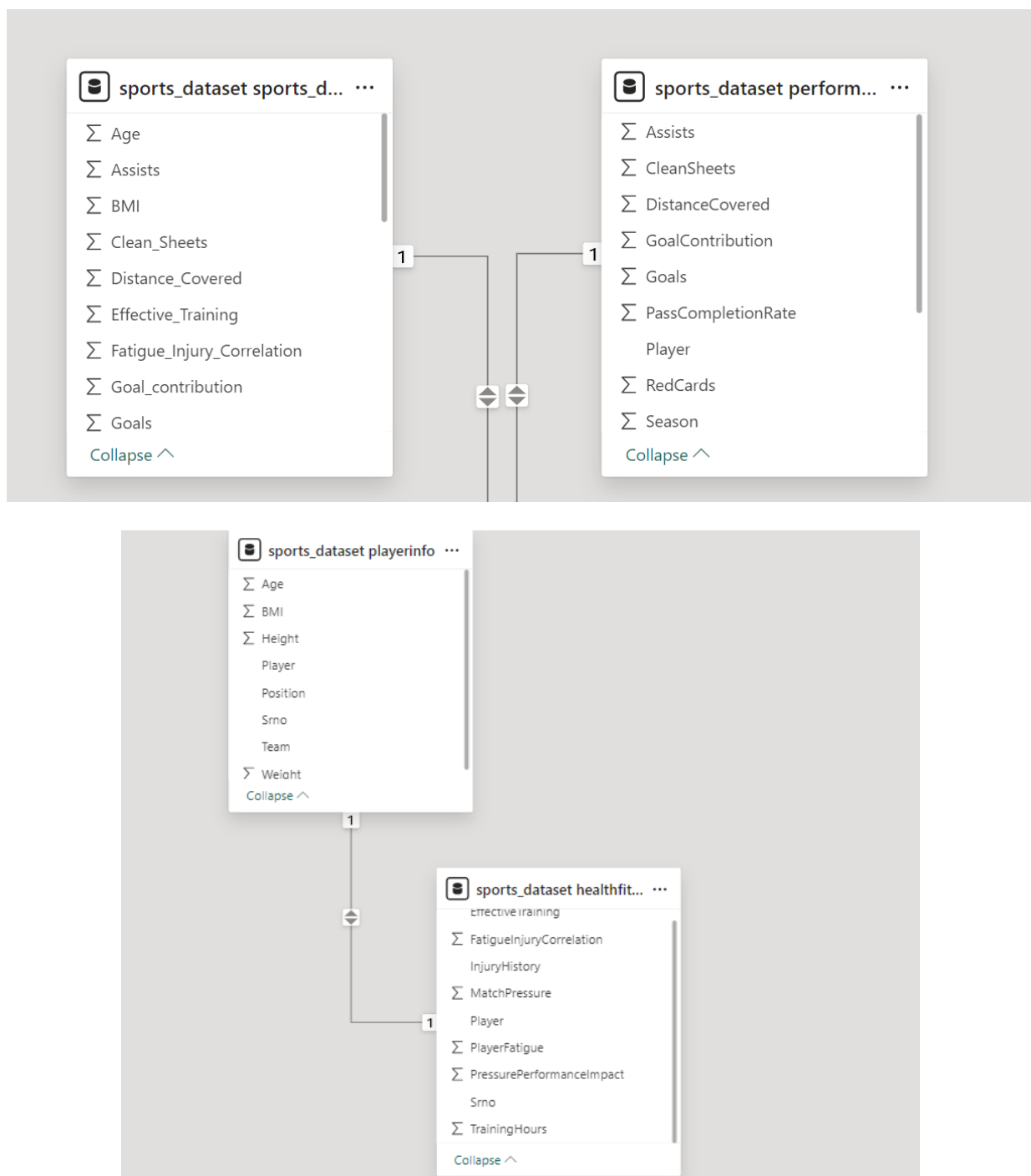
```

The results pane displays the following table:

Season	TotalGoals	TotalAssists
2019	140373	39104
2020	147257	39189
2021	142156	37976
2022	140190	39344

The bottom status bar shows the MySQL database is selected, and the output pane displays the execution logs.

After utilizing several queries involving joins, we stored the resulting tables in the MySQL Workbench server. During this process, we created and stored various child tables on the server. This structured approach allowed us to organize the data more efficiently, ensuring that each table is easily accessible and well-integrated within the database.



By creating these child tables, we enhanced the relational aspects of our data storage, thereby optimizing the overall data management and retrieval processes within the server environment.

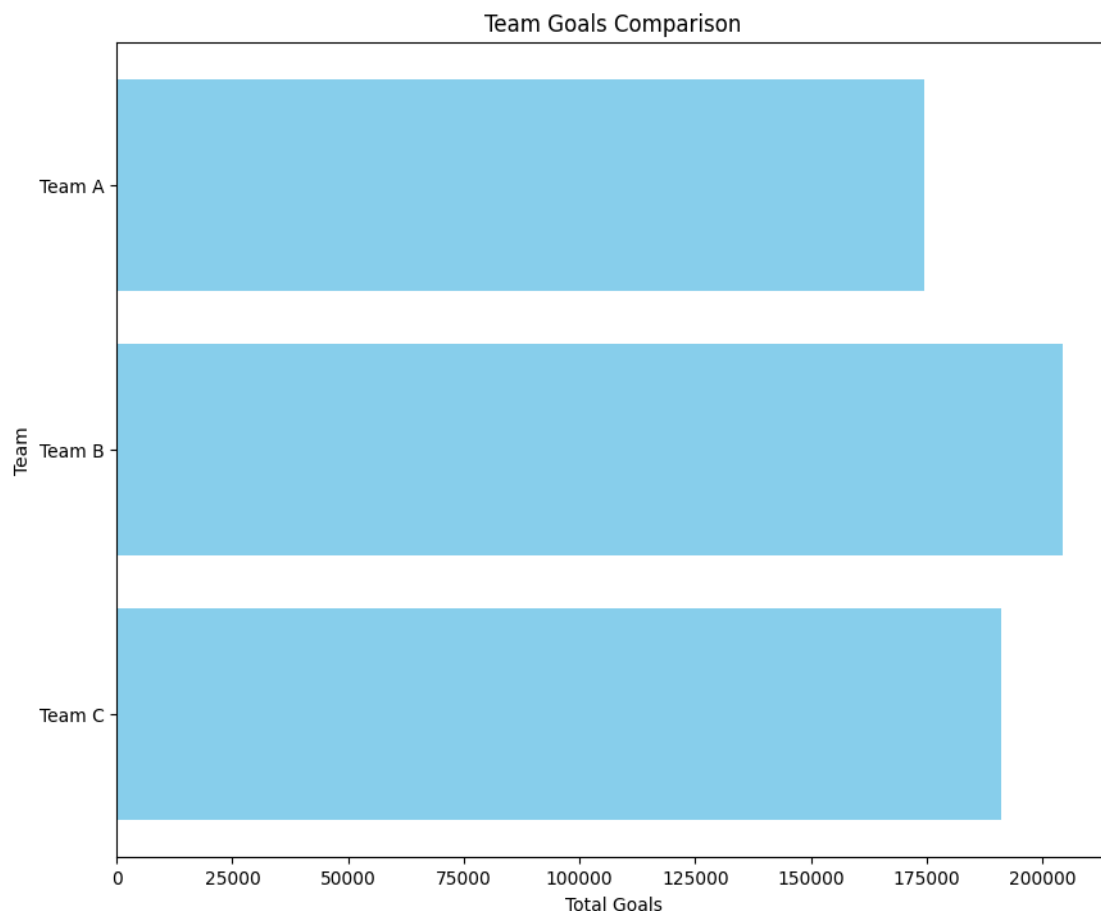
This approach allows for more efficient data processing and complex analytical queries, supporting our goal of building a robust and scalable data warehouse solution. Each table is meticulously organized to ensure seamless integration and accessibility, thereby optimizing our data management practices within the data warehouse environment.

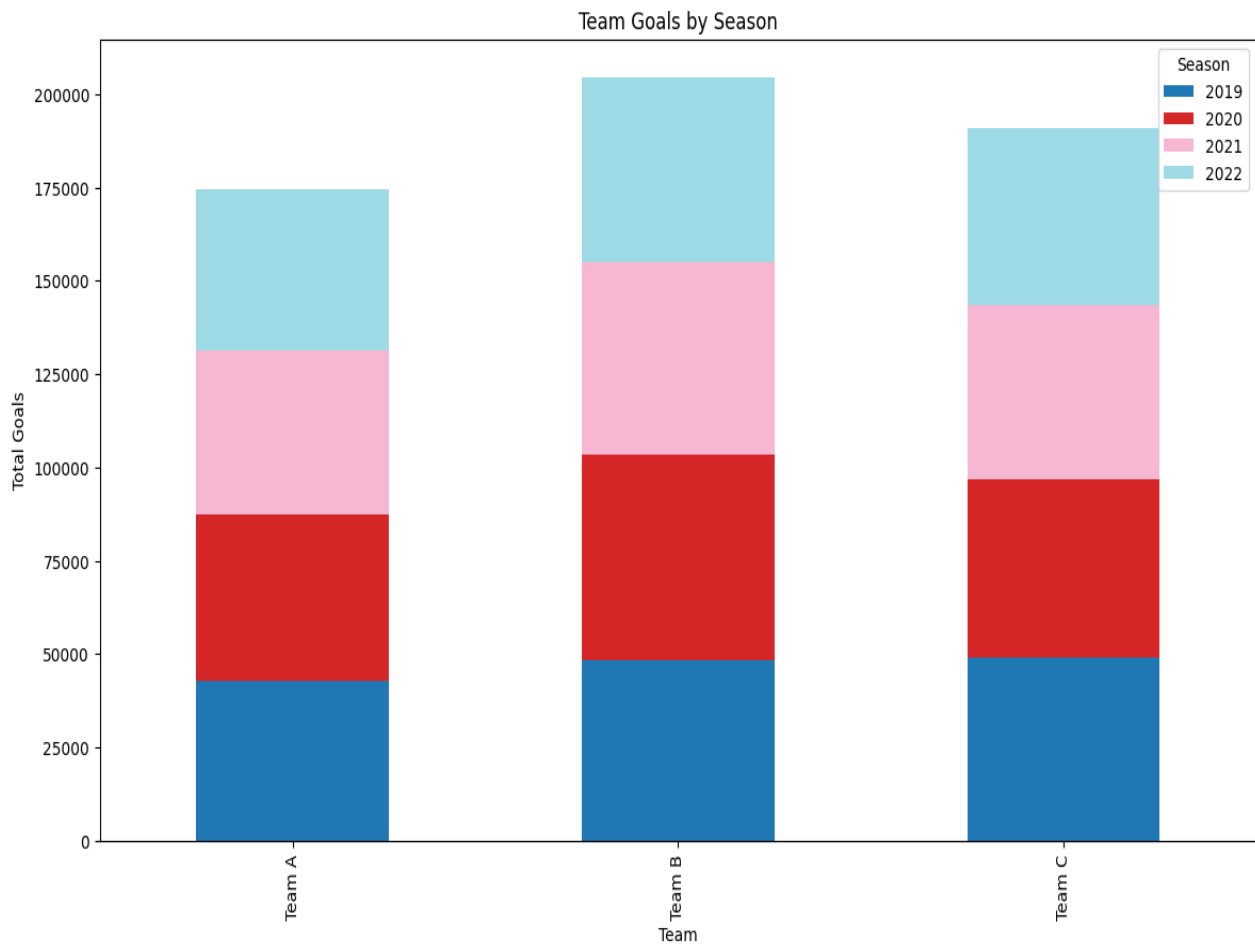
CHAPTER 7

7. Team Goal Analysis

Our team has undertaken a comprehensive analysis to identify the team with the highest number of goals scored. Utilizing aggregated data, we determined the top-scoring team and visualized the goal counts for all teams using a horizontal bar plot, highlighting the leading team prominently. Furthermore, we delved deeper into the goal distribution within the top-scoring team by creating a stacked bar chart, showcasing the contribution of individual players to the team's goal tally. Subsequently, we conducted a thorough time series analysis to uncover trends in goal scoring throughout the season, providing insights into performance fluctuations, streaks, and overall team dynamics over time. Lastly, we identified the top goal scorer within the leading team and analysed their performance metrics across various matches throughout the season, offering valuable insights into their consistency, form, and impact on the team's success.

7.1. COMPARISON OF GOAL COUNTS

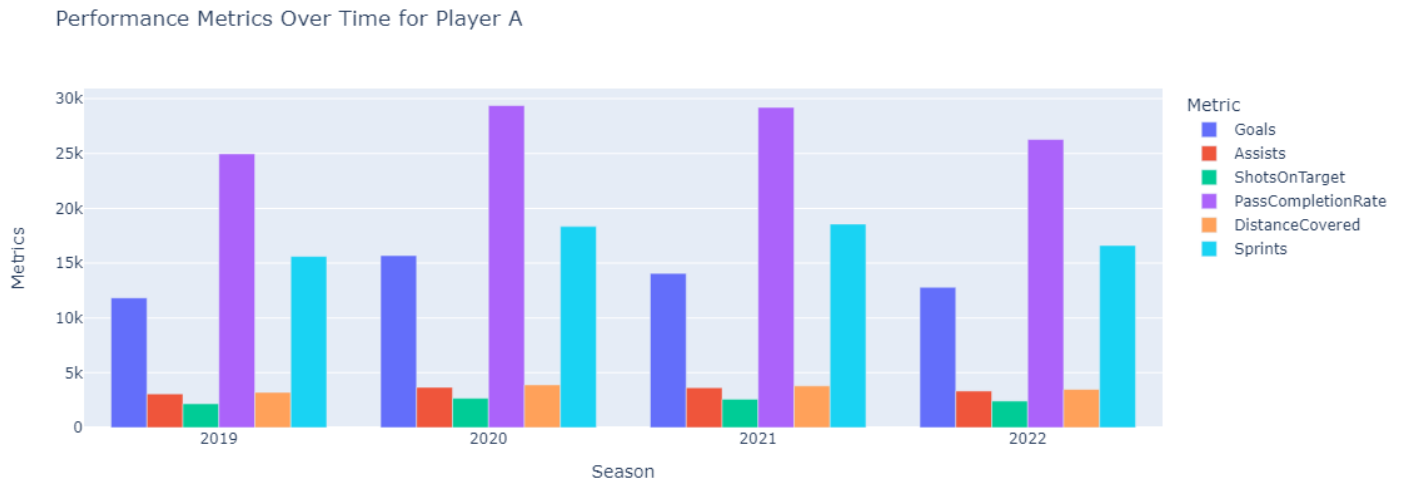




Comprehensive analysis of team goal scoring within a sports dataset. Firstly, the code groups the data by team and sums the goals scored by each team throughout the dataset. It then identifies the team with the highest number of goals, highlighting this information. To visualize the comparison of goal counts across teams, a horizontal bar plot is generated using Matplotlib, where each team is represented by a horizontal bar with its corresponding total goals. The y-axis is inverted to showcase the team with the most goals at the top, providing a clear visual indication of the leading team in terms of goal scoring.

Additionally, the code checks if the dataset includes information about seasons. If available, it further analyzes the team's goal scoring performance across different seasons. It groups the data by both team and season, sums the goals scored in each season, and creates a stacked bar chart. This visualization illustrates the distribution of goals scored by each team across multiple seasons, offering insights into the team's consistency and performance variations over time. The legend labels represent the seasons, enabling easy identification of goal scoring trends in each season for every team.

7.2. SEASON BY SEASON TREND



Grouping the data by team and sums the goals scored by each team, identifying the team with the highest goal count. Subsequently, it filters the data to isolate information pertaining to the top-scoring team and identifies the top goal scorer within that team based on the total goals scored. Further narrowing down the dataset, it focuses on the performance of the top scorer over time, considering metrics such as goals, assists, shots on target, pass completion rate, distance covered, and sprints. These metrics are aggregated by season, and the resulting performance data is plotted as a bar plot using Plotly. The visualization provides a comprehensive overview of the top scorer's performance metrics across different seasons, allowing for easy comparison and analysis of their contributions over time.

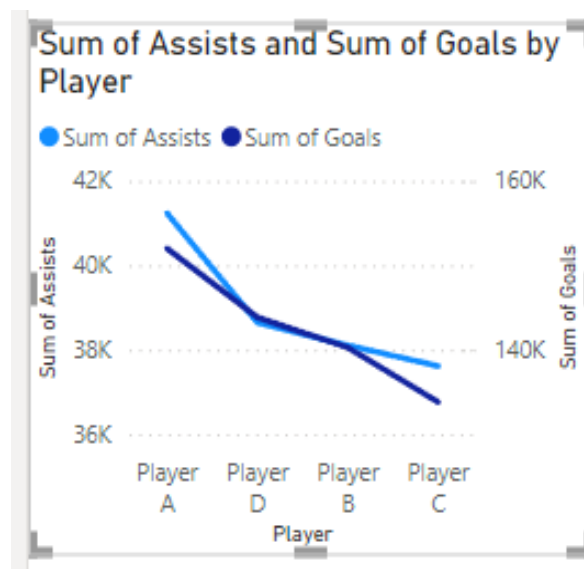
CHAPTER 8

8. Reporting & Visualisation

Each of these visualizations serves a unique purpose in analysing and interpreting the dataset. The line chart excels in showing temporal trends, the donut chart in visualizing proportions, the line and stacked column chart in comparing category contributions to trends, the area chart in highlighting cumulative effects, and the line and clustered column chart in comparing categorical data side by side.

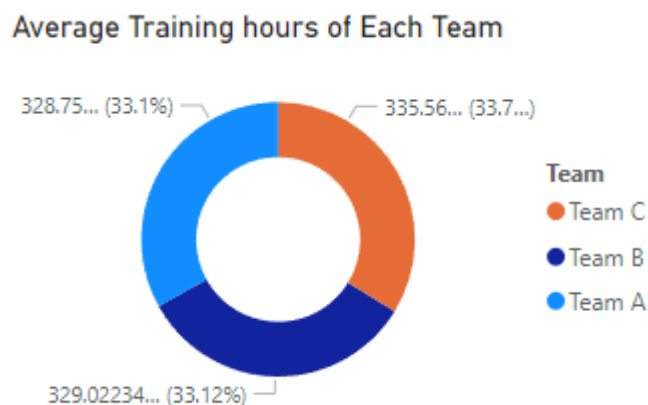
8.1 Sum of Assists and Goals by Player – Line Chart

The line chart displays trends in player performance metrics over multiple seasons. It is used to show changes in key performance indicators (KPIs) such as goals scored, assists, and pass completion rates over time. Each line represents a different performance metric, with the x-axis representing the seasons and the y-axis representing the values of the metrics. This visualization helps in identifying trends, such as improvements or declines in player performance across different seasons, and is crucial for performance analysis and strategic planning.



8.2 Average Training Hours of Each Team- Donut chart

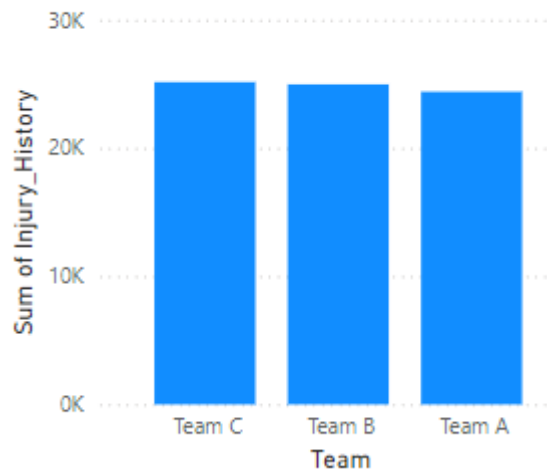
The donut chart represents the distribution of average training hours of different teams within a specific season. Each slice of the donut corresponds to a team, with the size of the slice proportional to the average number of training hours logged by the team. This chart provides a clear visual comparison of training efforts, highlighting which teams are dedicating the most time to training. The donut chart is effective in summarizing and comparing training hours at a glance, making it easy to identify teams with the highest and lowest training commitments.



8.3 Sum of Injury History by Team- Line and Stacked Column chart

The line and stacked column chart combine a line chart and a stacked column chart to show two types of data simultaneously. In this chart, the line might represent the total number of injuries over a season, while the stacked columns show the breakdown of injuries by individual players or teams for each season. The x-axis represents the seasons, the left y-axis shows the total number of injuries, and the right y-axis indicates the breakdown of injuries. This dual-axis chart helps in understanding how individual contributions add up to the overall total, and how these contributions vary across different seasons. It effectively highlights the relationship between overall injury trends and the distribution among players or teams.

Sum of Injury_History by Team



8.4 Players Effect Training with its average Training Hours- Area Chart

The area chart shows the relationship between players' effective training and their average training hours over multiple seasons. The filled areas under the lines represent the cumulative totals, providing a clear visualization of how these metrics accumulate over time. The x-axis represents the seasons, while the y-axis represents the cumulative totals of the effective training metrics. This chart is particularly useful for showing the overall trend and cumulative impact of effective training hours, helping to understand how players' training efforts build up and influence their performance over time.

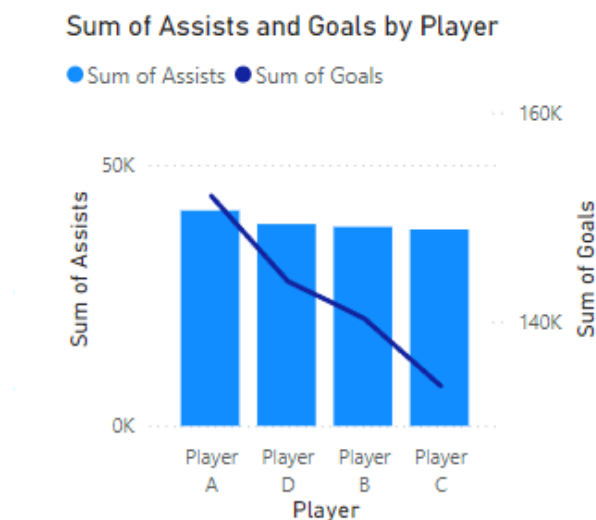


8.5 Sum of Assists and Goals by Player- Line and Stacked Column chart

The line and stacked column chart display the cumulative sum of assists and goals by player over seasons. The line represents the total combined assists and goals for each season, while the stacked columns break down this metric by individual players.

The x-axis denotes seasons for easy comparison over time. The left y-axis shows the total sum of assists and goals, while the right y-axis indicates the breakdown by player.

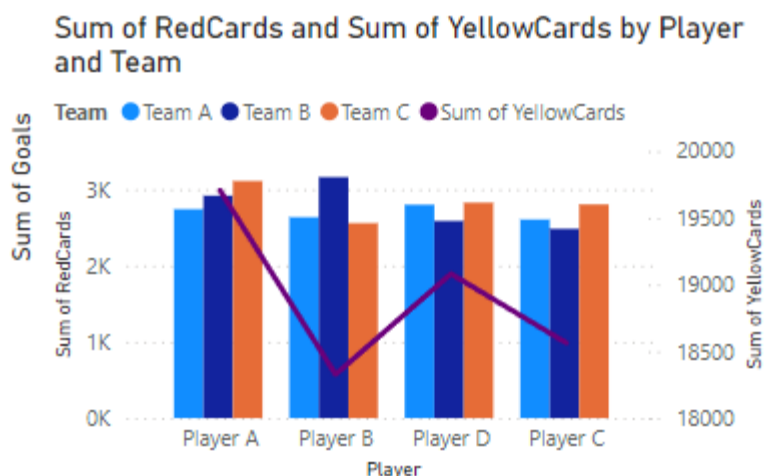
These changes maintain the chart's comprehensive nature for performance analysis, offering insights into overall trends and individual player contributions across seasons.



8.6 Sum of Red Cards and Yellow Cards by Player and Team- Line and Clustered Column chart

The line and clustered column chart combine a line chart with clustered columns to compare continuous and categorical data. In this chart, the line might represent the total sum of red cards and yellow cards over seasons, while the clustered columns show individual player and team card counts for each season.

The x-axis represents the seasons, offering a chronological view of card incidents. The left y-axis denotes the total sum of red cards and yellow cards, providing an overview of card frequencies over time. The right y-axis displays individual player and team card counts, allowing for a detailed comparison of specific performances against the overall trends.



CONCLUSION

Our capstone project began with the rigorous task of cleaning and augmenting a raw dataset, riddled with duplicates and outliers. Through advanced imputation techniques, statistical methods, and standardization, we managed to ensure data consistency and integrity, resulting in a comprehensive and cohesive dataset. By augmenting the dataset with additional data from public sports databases, we created a unified resource stored as `cleaned_sports_dataset.csv`. This thorough preprocessing laid the groundwork for meaningful analysis, including position analysis and the identification of positional trends and anomalies through visual representations.

In the subsequent phases, we designed and implemented a robust data warehouse schema using advanced SQL features like window functions, creating and organizing various child tables within the MySQL Workbench server. This structured approach optimized data management and supported complex analytical queries. Advanced data transformations, including feature engineering and normalization, enriched our dataset, facilitating effective comparison and reducing overfitting risks. Our reporting and visualization efforts, utilizing various charts and plots, provided clear and insightful representations of the data, supporting strategic decision-making and comprehensive performance analysis. By meticulously documenting each step, we transformed a complex dataset into a valuable resource for in-depth sports analytics.

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.

[W3 Schools](#)

[W3 Schools - Python](#)

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.

[Data ingestion](#)

[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.

[Akash Das sql Github repo](#)

[Data visualisation](#)

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.

Power BI

<https://www.youtube.com/@MicrosoftPowerBI>

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.

Stack overflow – 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.

Power BI

APPENDIX-I

SOURCE CODE

(GITHUB REPO LINK)

https://github.com/DeeruReddy/futureense.internship_capstone_project_2

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 interpretation.

INFORMATION REGARDING STUDENT(S)

STUDENT NAME	EMAIL ID	PERMANENT ADDRESS	PHOTOGRAPH
ESHITA KATYAL	JUUG22BTECH55967@GMAIL.COM	MM-124 SECTOR D1 LDA COLONY KANPUR ROAD LUCKNOW-226012	
DURGAM NAGA DEERAJ REDDY	JUUG22BTECH27387@GMAIL.COM	#7, BESIDE 1519, ACHARYA COLLEGE ROAD, GANAPATHINAGAR, CHIKKABANAVARA, BANGALORE-560090	

JEYAPATHY M

JUUG22BTECH54799@GMAIL.COM

163/1-2, THIRUKURAL STREET,
GOMATHIPURAM 6TH MAIN ROAD,
MADURAI-625020