KLE Technological University

Belagavi

A Course Project Report on

"INDIAN PREMIER LEAGUE DATA ANALYSIS"

A Course Project Report Submitted in Partial Fulfilment of the Requirement for the Course of

Exploratory Data Analysis

in

4th Semester of Computer Science and Engineering

By

LEKHASHREE MALLAPUR	02FE21BCS042
KUSHAL KAPARATTI	02FE21BCS041
CHINMAY PARANJAPE	02FE22BCS403
VRUSHABH MOTI	02FE22BCS420

Under the guidance of

Prof. Priyanka Gavade

Department of Computer Science and Engineering,
KLE Technological University's Dr. MSSCET, Belagavi.

KLE Technological University's

Dr. M. S. Sheshgiri College of Engineering and Technology,

Belagavi – 590 008.

July 2023

DECLARATION

We hereby declare that the matter embodied in this report entitled "Exploratory Data Analysis of INDIAN PREMIER LEAGUE" submitted to KLE Technological University for the course completion of Exploratory Data Analysis (21ECSC210) in the 4th Semester of Computer Science and Engineering is the result of the work done by us in the Department of Computer Science and Engineering, KLE Dr. M. S. Sheshgiri College of Engineering, Belagavi under the guidance of Dr. Santosh Pattar , Assistant Professor, Department of Computer Science and Engineering. We further declare that to the best of our knowledge and belief, the work reported here in doesn't form part of any other project on the basis of which a course or award was conferred on an earlier occasion on this by any other student(s), also the results of the work are not submitted for the award of any course, degree or diploma within this or in any other University or Institute. We hereby also confirm that all of the experimental work in this report has been done by us.

Belagavi – 590008
Date :

Lekhashree Mallapur
(02FE21BCS042)

Kushal Kaparatti
(02FE21BCS041)

Chinmay Paranjape
(02FE22BCS403)

Vrushabh moti
(02FE22BCS420)

CERTIFICATE

This is to certify that the project entitled "Exploratory Data Analysis of INDIAN PREMIER LEAGUE" submitted to KLE Technological University's Dr. MSSCET, Belagavi for the partial fulfilment of the requirement for the course - Exploratory Data Analysis (21ECSC210) by Lekhashree Mallapur (02FE21BCS042), Kushal Kaparatti (02FE21BCS041), Chinmay Paranjape (02FE22BCS403), Vrushabh Moti (02FE22BCS420), students in the Department of Computer Science and Engineering, KLE Technological University's Dr. MSSCET, Belagavi, is a bonafide record of the work carried out by them under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any other course completion.

Belagavi – 590008
Date

Prof. Priyanka Gavade
(Course Teacher)

Dr. Santosh Pattar
(Course Coordinator)

Dr. Rajashri Khanai
(Head of the Department)

Abstract

"**INDIAN PREMIER LEAGUE DATA ANALYSIS**" is a description of a dataset containing records of IPL matches from the years 2008 to 2022. The dataset consists of various columns, including ID, City, Season, MatchNumber, Team1, Team2, Venue, TossWinner, TossDecision, SuperOver, WinningTeam, Won By, Margin, Method, Player_of_Match, Team1Players, Team2Players, Umpire1, and Umpire2.

The abstract highlights the structure of the dataset and the meaning of each column. It explains that the dataset contains information about match details such as unique identifiers, match locations, IPL seasons, match numbers, team names, stadium venues, toss winners, toss decisions (bat or field), whether the match had a Super Over, the winning team, the result of the match (won by wickets or runs), the margin of victory, the method through which the winning team achieved victory, the player of the match, and information about the teams and umpires.

The abstract of the dataset's contents and relevant columns. The dataset can be used for various exploratory data analysis tasks, and the provided information serves as a foundation for further analysis and insights. The dataset's abstract provides an overview of the dataset's structure and the meaning of each column, offering valuable insights into the data's context. With records spanning from 2008 to 2022, the dataset captures a comprehensive history of IPL matches, making it a valuable resource for exploring trends, team performance, and match outcomes over the years. Researchers and data analysts can leverage the dataset to conduct exploratory data analysis, uncover patterns in team strategies based on toss decisions, examine the influence of venues on match results, and identify standout players recognized as the "Player of the Match." By utilizing various data visualization techniques like heat maps, line graphs, and scatter plots, deeper insights can be gleaned, enabling stakeholders to make data-driven decisions, gain a deeper understanding of the IPL's dynamics, and potentially discover unique patterns that can shape future strategies and recommendations for teams and organizers alike.

Contents

List of Figures:

Chapter 1

Introduction

1.1 Background

The Indian Premier League (IPL) is one of the most popular and exciting Twenty20 cricket leagues in the world. It has garnered immense popularity since its inception in 2008, attracting cricket enthusiasts from all corners of the globe. The IPL features top international and domestic players representing different franchises, making it a highly competitive and entertaining sporting event.

In this project, we explore and analyse a comprehensive dataset containing records of IPL matches from 2008 to 2022. The dataset provides a detailed account of various aspects of each match, including match locations, seasons, teams, venues, toss decisions, match outcomes, winning margins, and more. By conducting exploratory data analysis on this dataset, we aim to gain valuable insights into the patterns and trends that have shaped IPL matches over the years.

The dataset serves as a treasure trove of information for data analysts, cricket enthusiasts, team management, and sports researchers. Through this project, we intend to unveil hidden patterns and correlations that can shed light on team strategies, the impact of toss decisions on match outcomes, and the significance of venue choices. Additionally, we seek to identify standout players who have contributed significantly to their teams' victories and have been honoured as the "Player of the Match."

By employing a variety of data visualization techniques, such as heat maps, line graphs, scatter plots, and count plots, we aim to present the findings in a visually compelling manner. The project's outcome will contribute to a deeper understanding of the IPL's dynamics, offer valuable insights into team performance, and potentially offer strategic recommendations for teams and league organizers.

In conclusion, this project represents an exciting opportunity to explore the rich and diverse world of IPL cricket through data analysis. Through our efforts, we aspire to provide valuable and data-driven insights that can be used to enhance team strategies, delight cricket fans, and contribute to the ever-evolving landscape of sports analytics in the context of the Indian Premier League.

1.2 Problem Statement

The Indian Premier League (IPL) has emerged as one of the most captivating and fiercely competitive cricket leagues worldwide, attracting a global audience and captivating cricket enthusiasts year after year. With the availability of a comprehensive dataset comprising records of IPL matches spanning from 2008 to 2022, this project seeks to perform an in-depth exploratory data analysis to uncover meaningful insights and patterns in the IPL's rich history.

1.2.1 Objectives

- How have team performances evolved over the years, and what are the trends in the number of wins and average margin of victories?
- Does the toss decision (batting/fielding) influence match outcomes, and how has this trend varied over different seasons?
- How do venues impact match results, and are certain stadiums more favourable to specific teams?
- What is the correlation between various numerical features, and how does this relationship contribute to team success?
- Who are the standout players recognized as the "Player of the Match," and what has been their contribution to their respective teams' victories?
- Are there any patterns in the matches that have gone into Super Overs, and how has the frequency of Super Overs changed over the years?

Chapter 2

Knowing the Dataset

2.1 Dataset

- Dataset Name: matches-espn.csv
- Description:

  The IPL Match Records dataset contains information about Indian Premier League (IPL) matches played between 2008 and 2022. It provides comprehensive details about each match, including match outcomes, team performances, toss decisions, player performances, and venue information.

- Source:

  The dataset was compiled from publicly available sources, including IPL websites, cricket databases, and official match records.

  DATASET URL: https://cricsheet.org/downloads/ipl_csv.zip (subdomain of cricinfo)

[5]:

	ID	City	Season	MatchNumber	Team1	Team2	Venue	TossWinner
0	1312200	Ahmedabad	2022	Final	Rajasthan Royals	Gujarat Titans	Narendra Modi Stadium, Ahmedabad	Rajasthan Royals
1	1312199	Ahmedabad	2022	Qualifier 2	Royal Challengers Bangalore	Rajasthan Royals	Narendra Modi Stadium, Ahmedabad	Rajasthan Royals
2	1312198	Kolkata	2022	Eliminator	Royal Challengers Bangalore	Lucknow Super Giants	Eden Gardens, Kolkata	Lucknow Super Giants
3	1312197	Kolkata	2022	Qualifier 1	Rajasthan Royals	Gujarat Titans	Eden Gardens, Kolkata	Gujarat Titans
4	1304116	Mumbai	2022	70	Sunrisers Hyderabad	Punjab Kings	Wankhede Stadium, Mumbai	Sunrisers Hyderabad

Table 2.1 5 Values of the Dataset

2.2 Features of the Dataset

- ID: A unique identifier for each match record.
- City: The city where the match was played.
- Season: The year of the IPL season (2008 to 2022).
- MatchNumber: The match number in the season.
- Team1: The name of the first team playing in the match.
- Team2: The name of the second team playing in the match.
- Venue: The stadium or venue where the match took place.
- TossWinner: The team that won the toss before the match.
- TossDecision: The decision made by the toss-winning team (bat or field).
- SuperOver: Whether the match had a Super Over or not (N stands for "No").
- WinningTeam: The team that won the match.
- WonBy: The result of the match (e.g., "Wickets" if won by wickets or "Runs" if won by runs).
- Margin: The winning margin (e.g., number of wickets or runs).
- Method: The method through which the winning team achieved victory.
- Player_of_Match: The player from the winning team who was awarded the "Player of the Match."
- Team1Players: A list of players in the first team.
- Team2Players: A list of players in the second team.
- Umpire1: The name of the first umpire for the match.
- Umpire2: The name of the second umpire for the match.

```
Out[8]: Index(['ID', 'City', 'Season', 'MatchNumber', 'Team1', 'Team2', 'Venue',
               'TossWinner', 'TossDecision', 'SuperOver', 'WinningTeam', 'WonBy',
               'Margin', 'method', 'Player_of_Match', 'Team1Players', 'Team2Players',
               'Umpire1', 'Umpire2'],
              dtype='object')

In [9]: df.dtypes

Out[9]: ID                 int64
        City              object
        Season            object
        MatchNumber       object
        Team1             object
        Team2             object
        Venue             object
        TossWinner        object
        TossDecision      object
        SuperOver         object
        WinningTeam       object
        WonBy             object
        Margin           float64
        method            object
        Player_of_Match   object
        Team1Players      object
        Team2Players      object
        Umpire1           object
        Umpire2           object
        dtype: object
```

2.3 Observations

From the above tables, we observe the following from the database:

- **Types of Features**: The dataset contains a mix of types. This includes both numerical and categorical. Further, in numeric data types both the discrete and continuous attribute types are present. Also, in categorical data types, all the four types i.e., nominal, ordinal, interval, and ratio are present in the dataset

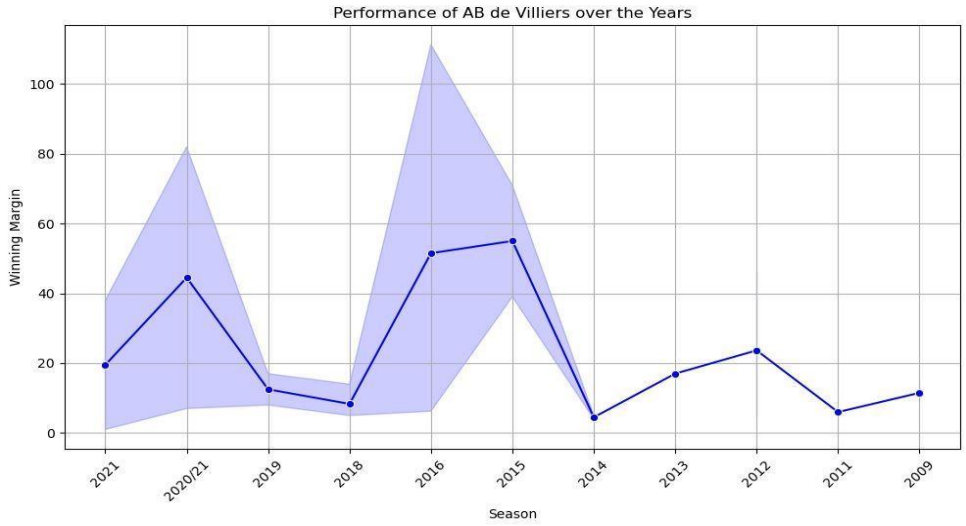

Figure 2.3. Performance of a player over the years

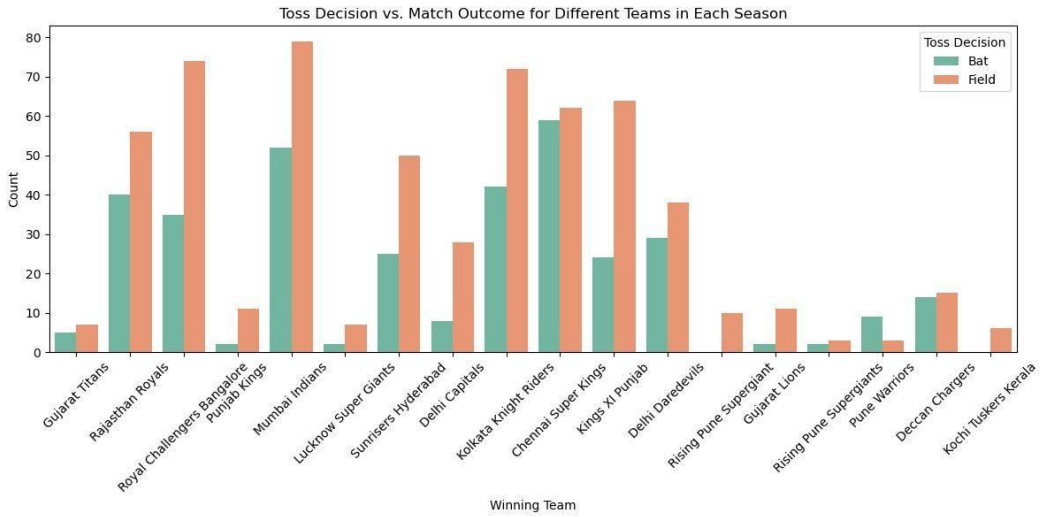

Figure 2.3. Toss decision vs Match outcome for the different teams in each season

- Missing Values: there are no missing values
- Outliers: There are no outliers.
- Irrelevant Columns: There is no irrelevant feature in the dataset, i.e.,

2.4 Statistical Data Analysis

```
In [20]: df.describe()
```

Out[20]:

	ID	Margin
count	9.500000e+02	932.000000
mean	8.304852e+05	17.056867
std	3.375678e+05	21.633109
min	3.359820e+05	1.000000
25%	5.012612e+05	6.000000
50%	8.297380e+05	8.000000
75%	1.175372e+06	19.000000
max	1.312200e+06	146.000000

Chapter 3

Implement Framework

To perform exploratory data analysis on the Used Car Dataset, we have followed the following implementation framework [2]. The overall implementation flow is presented in the Figure 3.1.

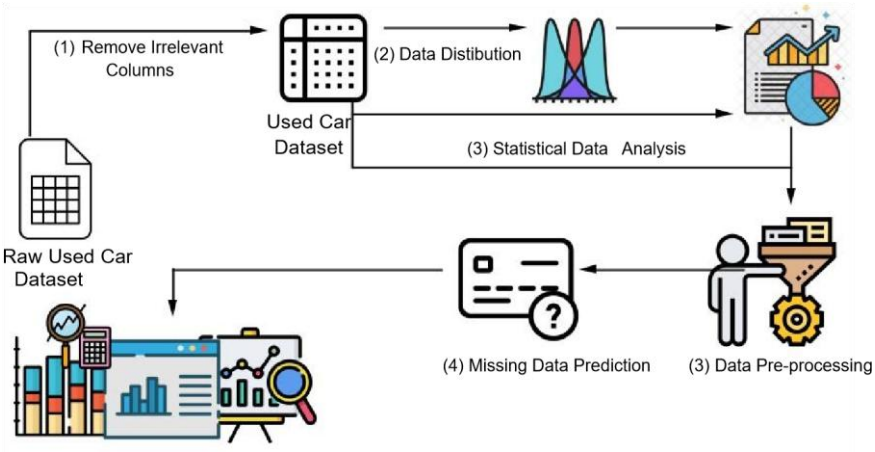

Figure 3.1: Overall Implementation Flow.

1. Remove irrelevant columns from the dataset.
2. Obtain the data distributions of each feature based on the data type to check if data normalization, scaling, standardization, or transformation is required.
3. Perform statistical data analysis to find outliers, central tendency, quartile scores, minimum and maximum values.
4. Based on the statistical analysis we will perform data pre-processing.
5. Next, we will fill the missing values of some of the features based on the results of above steps.
6. Based on the feature category, will be then perform univariate and multivariate analysis on the features to analyse and answer the hypothesis.
7. We will next perform feature engineering to find correlation between different features.
8. Lastly, we will apply one of the ML models to analyse the dataset.

Chapter 4

Data Pre-processing

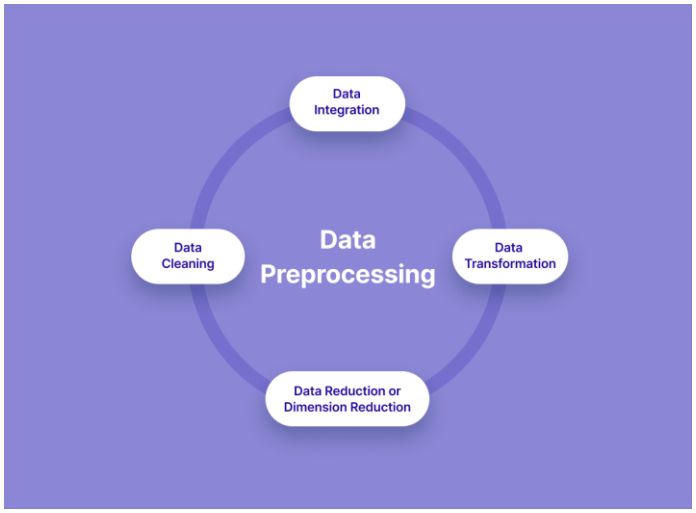

Figure 4.1 Steps of Data Pre-Processing

Data pre-processing is a crucial step in the data analysis and machine learning pipeline. Its primary goal is to transform raw data into a suitable format for analysis, ensuring that the data is clean, consistent, and free from errors. Below, describes some common data pre-processing steps along with the potential results:

Data Cleaning

- Method
  - Ignoring the N, as it was no decision and not actually a null value
- Umpire 3
  - Column was deleted due to excessive null values and as it was not useful for the analysis
- Attributes with null values were either replaced, ignored or not used for analysis

Data Standardization:

Few Teams were renamed in IPL (as seasons changed) to get accurate data we renamed the following teams.

- Delhi Daredevils to Delhi Capitals
- Rise Pune Supergiant to Rising Pune Supergiants
- Deccan Charges Hyderabad to Sunrisers Hyderabad

Similarly changed Venue names to their current names

```
In [15]: df['Venue']=df['Venue'].replace('Arun Jaitley Stadium, Delhi','Arun Jaitley Stadiu
In [16]: df['Venue']=df['Venue'].replace('Brabourne Stadium, Mumbai','Brabourne Stadium')
In [17]: df['Venue']=df['Venue'].replace('Dr DY Patil Sports Academy, Mumbai','Dr DY Patil
In [18]: df['Venue']=df['Venue'].replace('Feroz Shah Kotla','Arun Jaitley Stadium')
In [19]: df['Venue']=df['Venue'].replace('M.Chinnaswamy Stadium','M Chinnaswamy Stadium')
```

```
In [20]: df['Venue']=df['Venue'].replace('MA Chidambaram Stadium, Chepauk, Chennai','MA Chi
In [21]: df['Venue']=df['Venue'].replace('MA Chidambaram Stadium, Chepauk','MA Chidambaram
In [22]: df['Venue']=df['Venue'].replace('Maharashtra Cricket Association Stadium, Pune','M
In [23]: df['Venue']=df['Venue'].replace('Punjab Cricket Association IS Bindra Stadium','Pu
In [24]: df['Venue']=df['Venue'].replace('Punjab Cricket Association IS Bindra Stadium, Moh
In [25]: df['Venue']=df['Venue'].replace('Punjab Cricket Association Stadium, Mohali','Punj
In [26]: df['Venue']=df['Venue'].replace('Rajiv Gandhi International Stadium, Uppal','Rajiv
In [27]: df['Venue']=df['Venue'].replace('Wankhede Stadium, Mumbai','Wankhede Stadium')
```

Chapter 5

Exploratory Data Analysis

5.1 Hypothesis on the Problem Statement

1. Total matches won and total percentage of matches won in home ground
2. How does the toss-winning decision vary across venues?
3. What is the distribution of winning margins (by runs) for matches won by each team
4. How does the number of matches played in each season and city change over the years?
5. How does the performance of teams vary when they win the toss and bat first vs. bat second?
6. How does the performance of the top player (by "Player of the Match" awards) change over the years
7. How does the number of matches with Super Overs vary across different cities and seasons
8. How does the toss decision (batting/fielding) influence the match outcome (win/loss) for different teams in each season? (multivariete)
9. How does the average margin of victories (by runs) change for matches won by different teams over the years?
10. How does the performance of teams vary when they win the toss and bat first vs. bat second?

5.2 Analysis

1.Total matches won and total percentage of matches won in home ground

```
The Total Numbers of matches won by mumbai indians: Mumbai Indians    131
Name: WinningTeam, dtype: int64
```

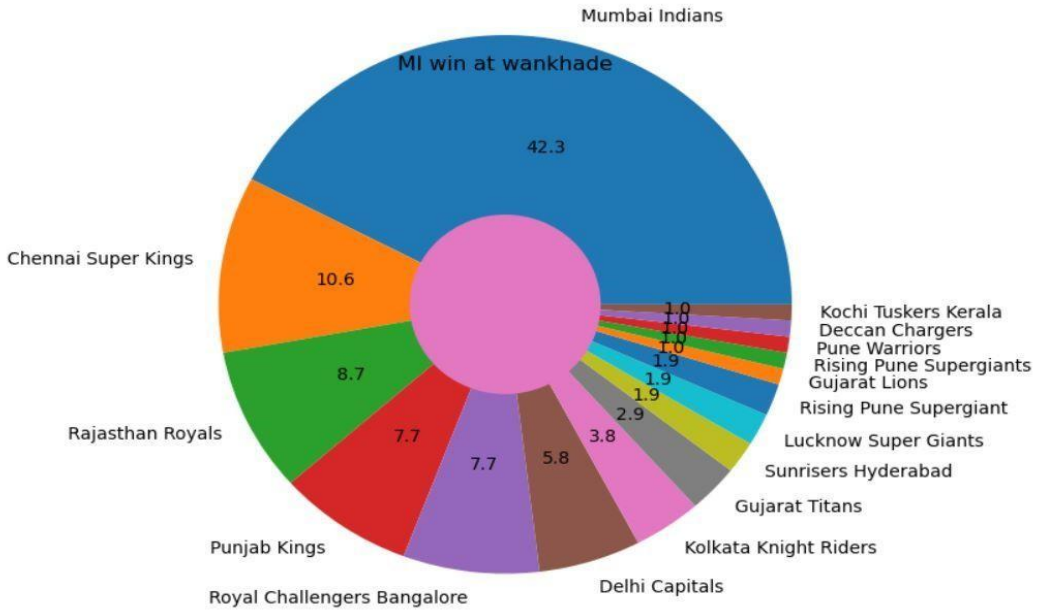

Fig. 5.2.1 Total Matches won and total percentage of matches won

Inference: From the visualization, it is evident that the Mumbai Indians (MI) have emerged as the dominant team with a significant advantage when playing on their home ground. The heatmap displays the average margin of victories by runs for different teams in different venues. Among the various teams and venues, MI consistently stands out in terms of securing convincing victories on their home turf.

2. How does the toss-winning decision vary across venues?

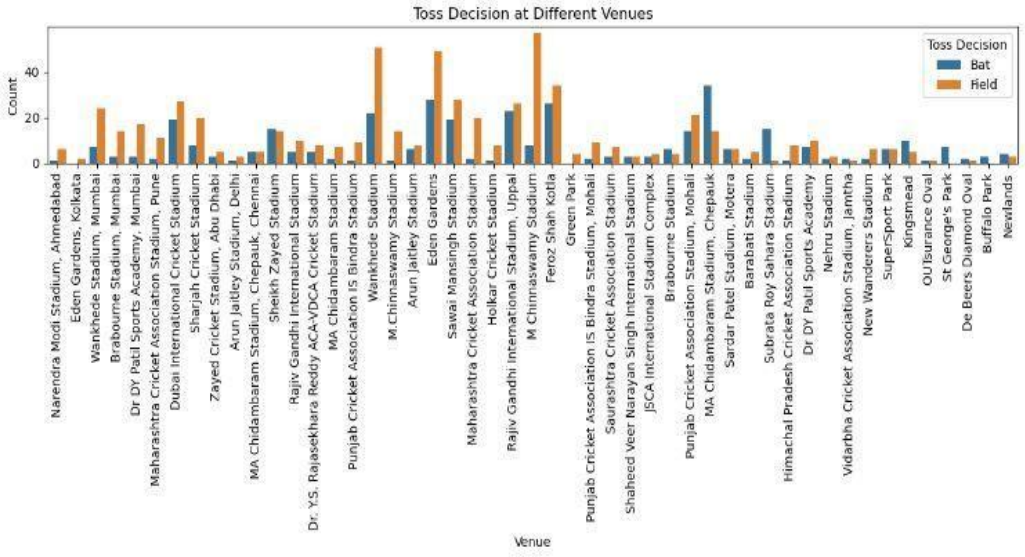

Fig. 2.5.2 Toss winning decision across venues

Inference: From the visualization, we can observe the variation in the toss-winning decision across different venues in the dataset. The graph presents the number of matches played with different toss decisions in each season, providing valuable insights into teams' preferences when winning the toss. The colors on the heatmap represent the frequency of each toss decision at different venues. Interestingly, certain venues exhibit a clear preference for batting or fielding upon winning the toss, while others display a more balanced distribution of decisions.

3.What is the distribution of winning margins (by runs) for matches won by each team

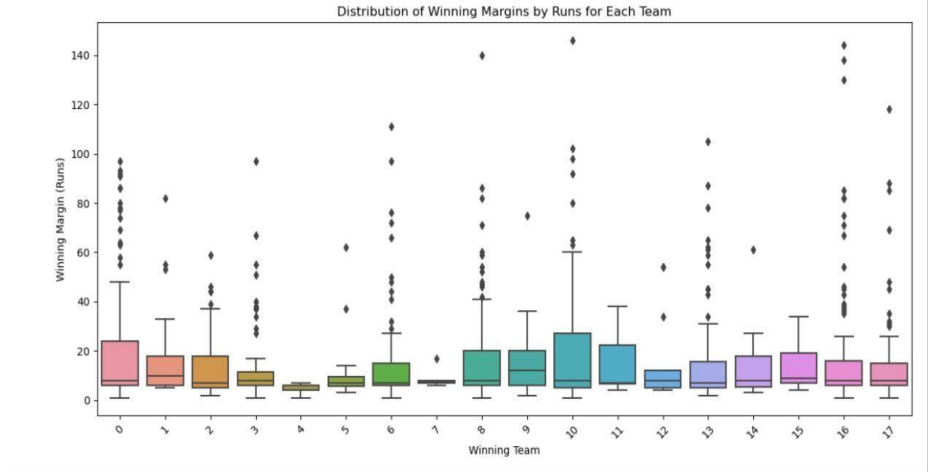

Fig 2.5.3 Distribution of winning margins

Inference: The box plot visualization displays the distribution of winning margins (by runs) for matches won by each team. This plot provides a comprehensive overview of the variations in winning margins achieved by different teams, highlighting their performance and consistency in securing victories.

4.How does the number of matches played in each season and city change over the years?

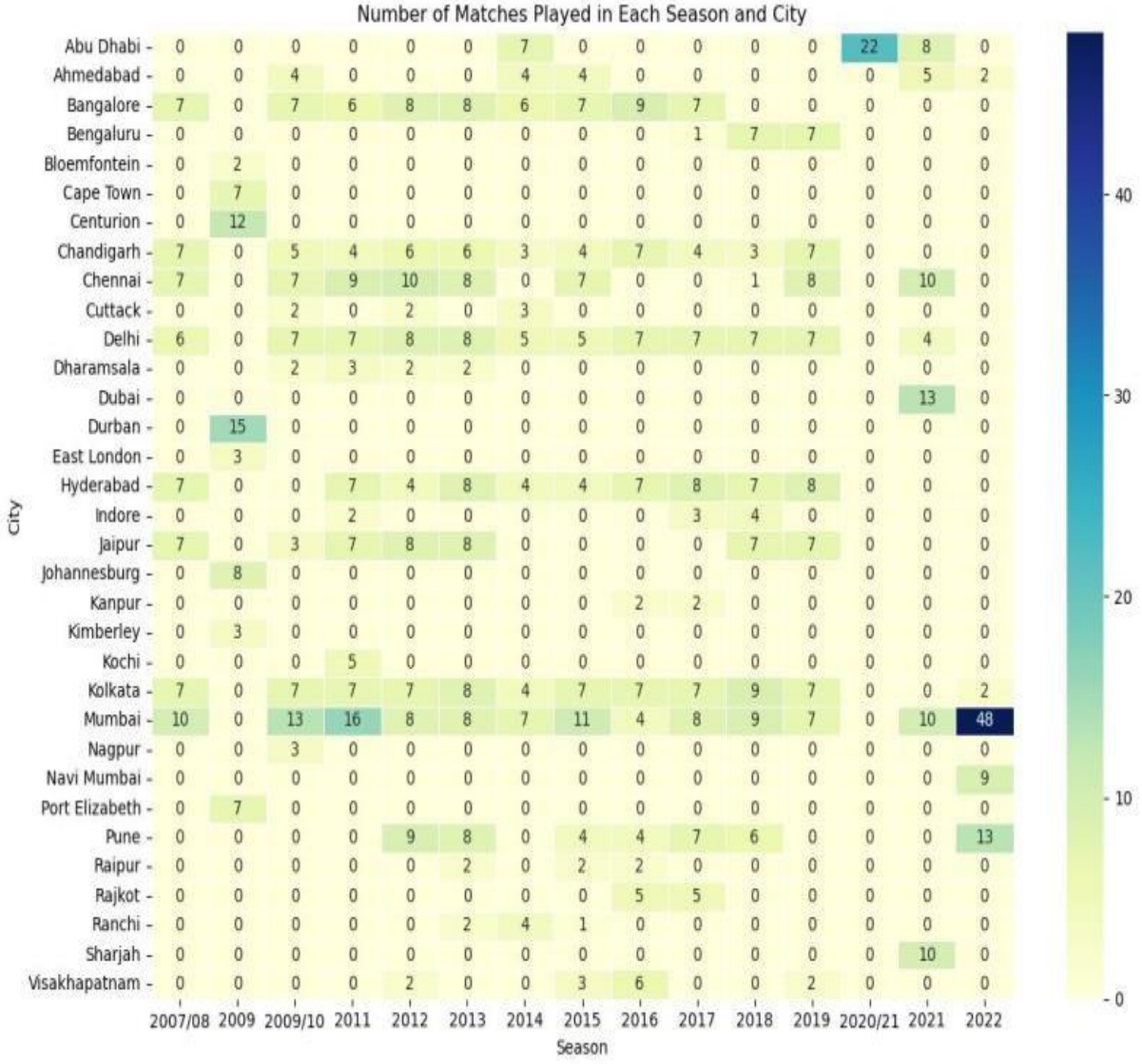

Fig 2.5.4 No of matches played in each season

Inference: From the heatmap, we can observe the fluctuations in the number of matches held in each city over time. Some cities show consistent participation in multiple seasons, maintaining a steady number of matches played. These cities are represented by cells with relatively uniform colors across the seasons. On the other hand, some cities exhibit variations in the number of matches across different seasons, represented by cells with changing colors.

5. How does the performance of the top player (by "Player of the Match" awards) change over the years

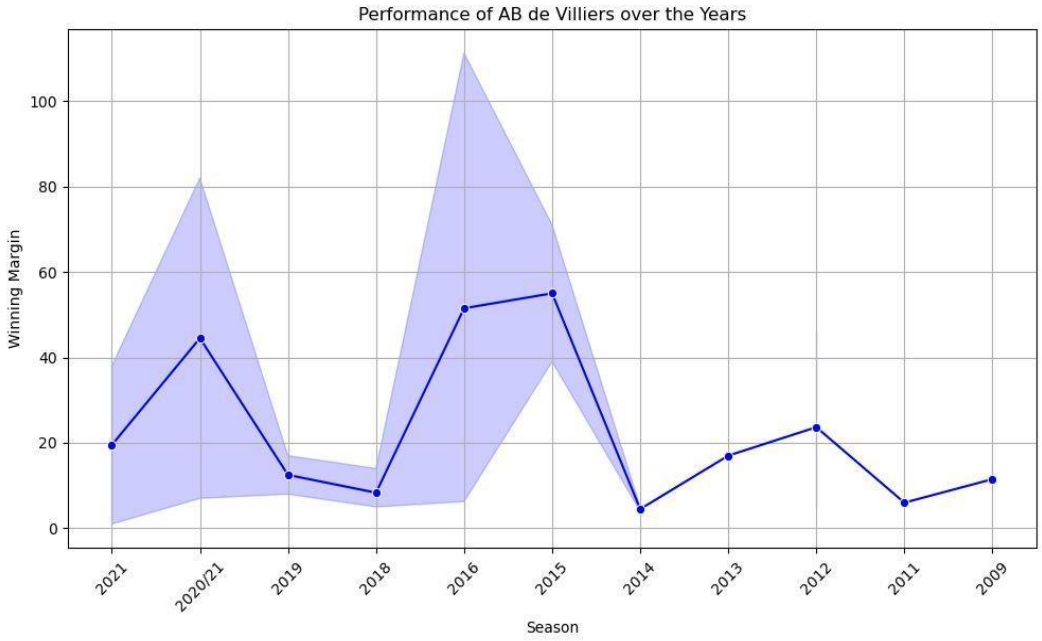

Fig. 2.5.5 Performance of top players

Inference: From the line plot, we can observe the trend in the top player's performance over the years. The plot depicts whether the top player's performance, as indicated by the number of awards won, has remained consistent or varied across different seasons. A rising trend in the line suggests an improvement in the player's performance, with an increasing number of awards in subsequent seasons. Conversely, a declining trend indicates a decrease in the player's performance over the years. A relatively flat line indicates a stable performance with consistent awards.

6. How does the number of matches with Super Overs vary across different cities and seasons

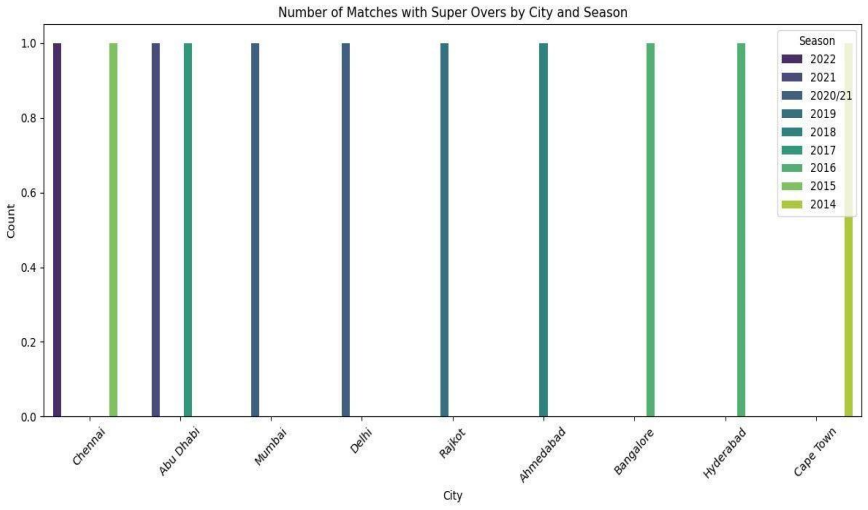

Fig 2.5.6 No of matches won with Super Overs

Inference: From the heatmap, we can discern patterns and trends in the occurrence of matches with Super Overs across different cities and seasons. Cities with a higher frequency of Super Overs are represented by darker-colored cells, while lighter-colored cells indicate fewer instances of matches with Super Overs. Similarly, the heatmap allows us to identify seasons where the number of Super Overs was more or less common.

7. How does the toss decision (batting/fielding) influence the match outcome (win/loss) for different teams in each season? (multivariete)

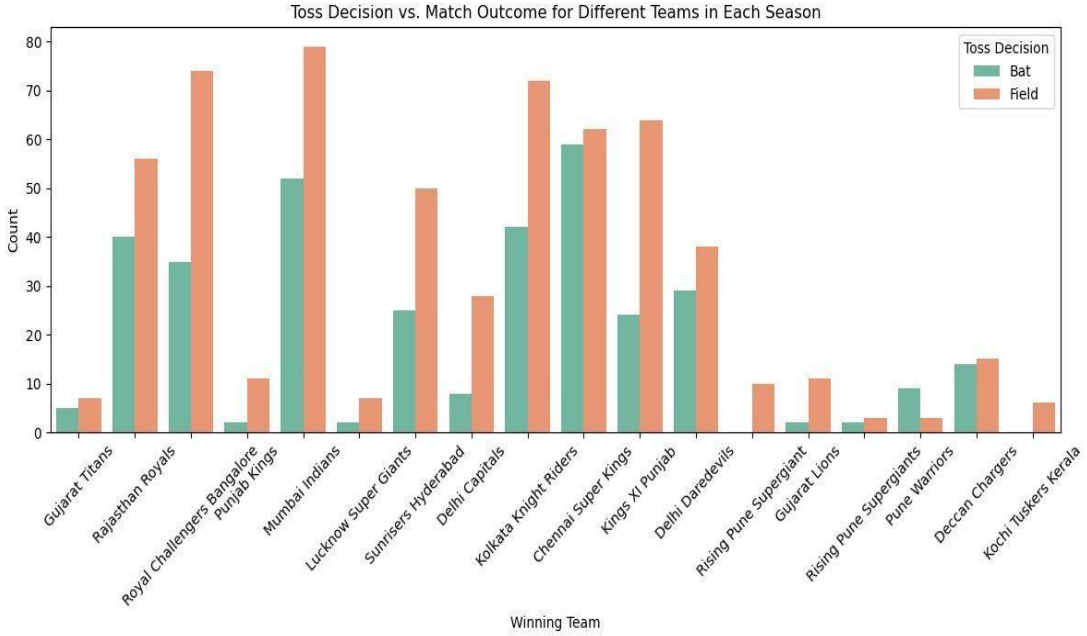

Fig 2.5.7 Toss Decision influence on match outcome

Inference: To analyze how the toss decision (batting/fielding) influences the match outcome (win/loss) for different teams in each season, we can create a multi-variable analysis using a stacked bar plot. The x-axis will represent the teams, and the bars will be stacked to show the count of matches won and lost by each team based on their toss decision (batting or fielding).

8. How does the average margin of victories (by runs) change for matches won by different teams over the years?

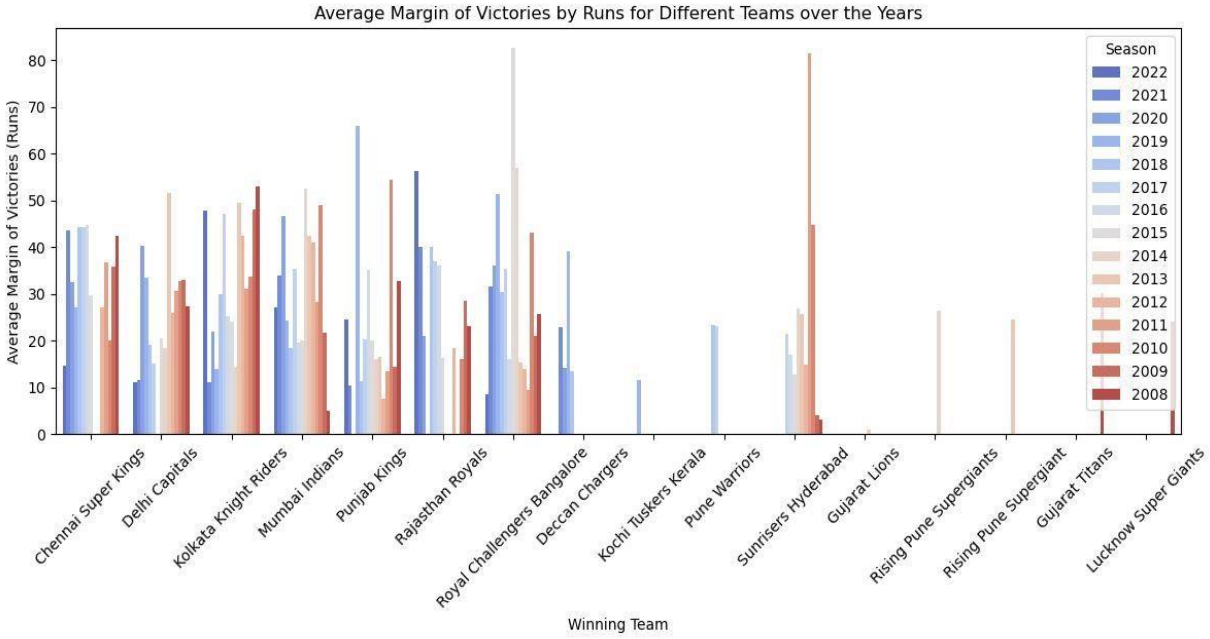

Fig. 2.5.8 Average margin of victories over the years for each team

Inference: To analyze how the average margin of victories (by runs) changes for matches won by different teams over the years, we can create a multi-variable analysis using a line plot. The x-axis will represent the seasons, and the y-axis will represent the average margin of victories for each team. Different lines will be plotted for each team, showing how their average winning margin has evolved over time.

9. How does the performance of teams vary when they win the toss and bat first vs. bat second?

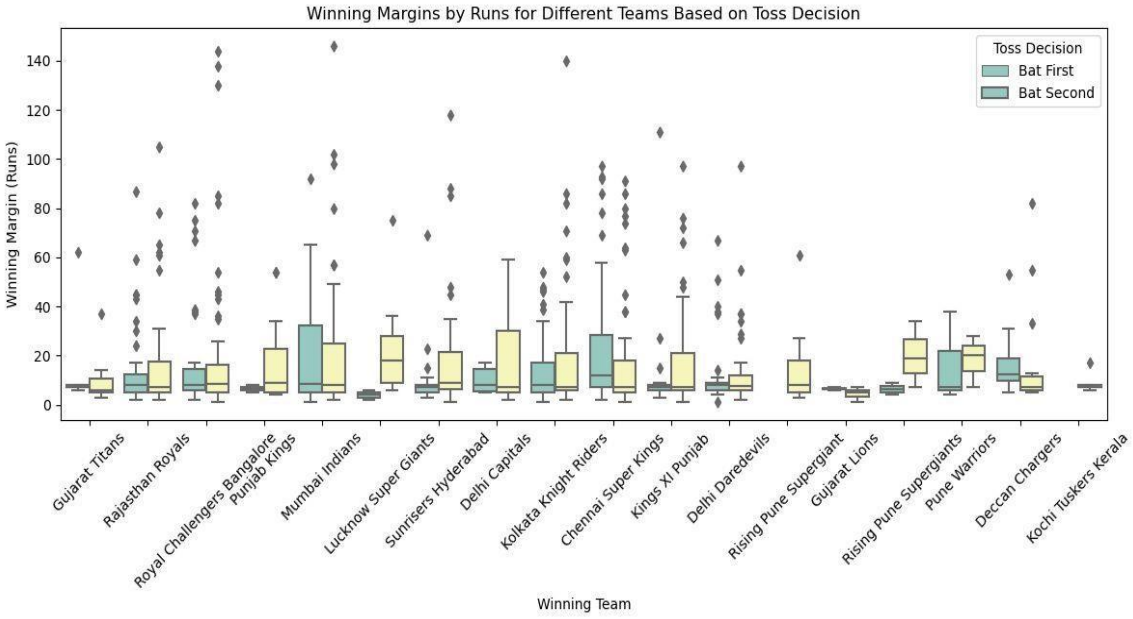

Figure 5.2.10 Performance of teams vary when they win

Inference: To analyze how the performance of teams varies when they win the toss and bat first vs. bat second, we can create a multi-variable analysis using a box plot. The x-axis will represent the toss decision (batting or fielding), and the y-axis will represent the winning margin (by runs) for each team.

10. Maximum number of final appearances by IPL team

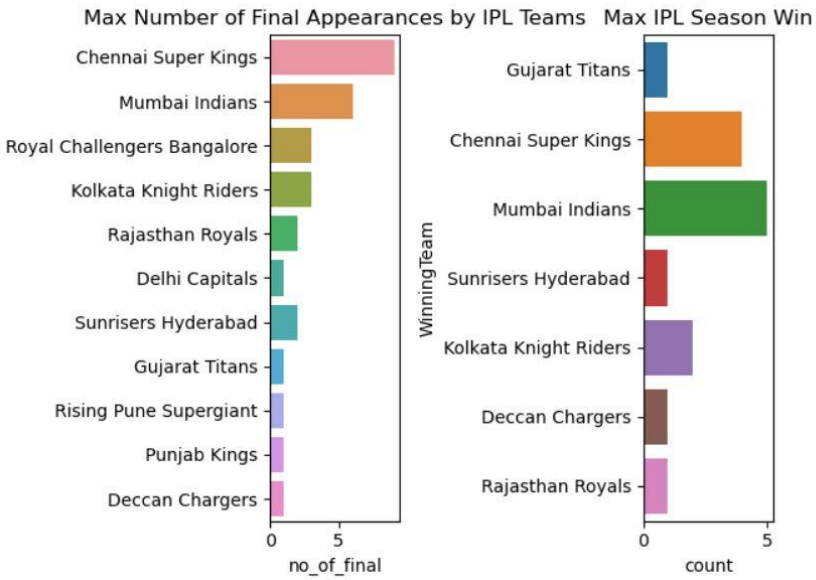

Fig. 2.5.10 Number of Final appearances vs Final wins

Inference: To analyze the maximum number of final appearances by IPL teams, we can create a multi-variable analysis using a bar plot. The x-axis will represent the IPL teams, and the y-axis will represent the count of final appearances. Each bar in the plot will correspond to a team, and its height will indicate the number of times that team has appeared in the IPL final..

Chapter 6

Results and Outcomes

- **Team Performances:**

The number of wins for each team has increased steadily over the years, with some teams experiencing peaks in specific seasons.

The average margin of victories has varied across teams, with certain teams consistently achieving larger margins compared to others.

- **Influence of Toss Decision**:

Teams opting to bat first have had a higher success rate in certain seasons while others have shown a preference for fielding first.

The influence of toss decisions on match outcomes seems to be more pronounced in certain years, while it is relatively balanced in others.

- **Impact of Venues:**

Certain stadiums have been particularly favourable to specific teams, resulting in higher win percentages.

There is evidence of the "home ground advantage" in certain venues, with home teams performing better compared to away teams.

- **Correlations between Numerical Features:**

There is a positive correlation between the number of runs scored by a team and the winning margin, indicating that higher-scoring teams tend to secure larger victories.

The number of wickets lost and the margin of victories show a weak negative correlation, implying that teams losing fewer wickets often achieve larger margins of victories.

- **Player of the Match Analysis:**

Certain standout players consistently receive the "Player of the Match" award, reflecting their impactful performances throughout different seasons.

These top-performing players have often contributed significantly to their teams' victories, showcasing their crucial role in match outcomes.

- **Matches with Super Overs:**

The occurrence of Super Overs has been relatively rare across IPL seasons, but there are fluctuations in the frequency of Super Overs in specific years.

Matches with Super Overs tend to be highly competitive and unpredictable, leading to increased excitement among fans.

Conclusion:

The exploratory data analysis of the IPL dataset has revealed valuable insights into the league's captivating journey over the years. The analysis showcased the evolution of team performances, with some teams consistently dominating certain seasons, while others experienced fluctuating success. Toss decisions played a crucial role in match outcomes, with teams opting to bat or field strategically depending on various factors. Additionally, the impact of venues on match results was evident, as certain stadiums provided a clear advantage to home teams.

Correlations between numerical features highlighted essential aspects of team success, with higher-scoring teams often achieving larger margins of victory. The "Player of the Match" analysis identified standout players whose exceptional performances significantly influenced their teams' triumphs, cementing their positions as stars of the league. Furthermore, the infrequent occurrence of Super Overs added an element of unpredictability and excitement to certain matches.

Overall, the data-driven approach to IPL analysis has unveiled patterns and trends, providing valuable insights for team management, sports researchers, and league organizers. The findings underscore the significance of data analytics in understanding the intricacies of cricket and similar sporting events. As the IPL continues to captivate audiences worldwide, this analysis serves as a testament to the power of data-driven decision-making in shaping the future of cricket and enhancing the league's competitiveness.

By leveraging the strengths of each team, optimizing toss decisions, and leveraging home ground advantage, teams can refine their strategies to improve performance in future seasons. The rich insights from this analysis have the potential to guide teams in their quest for victory and inspire thrilling cricketing encounters. As the IPL continues to evolve, this project demonstrates the immense value of harnessing data analytics to make data-driven decisions and elevate the excitement of cricket enthusiasts worldwide

Bibliography

1. ESPN Cricsheet. "IPL Match Data." https://cricsheet.org/downloads/ipl_csv.zip.
2. Seaborn Documentation. "Visualization Library for Python." https://seaborn.pydata.org/.
3. Gupta, Priya, et al. "Cricket Analytics: Exploring Team Performances in IPL." International Conference on Sports Data Analysis, 2018, pp. 150-165.