Statistical and Exploratory Data Analysis on Indian Premier League

Erandi Herath
Department of Computer Engineering
University of Sri Jayewardenepura
Colombo, Sri Lanka
erandikghs@gmail.com

Udaya Wijenayake
Department of Computer Engineering
University of Sri Jayewardenepura
Colombo, Sri Lanka
udayaw@sjp.ac.lk

Abstract—**The Indian Premier League (IPL) is a prominent professional cricket league in India that commenced in 2008, featuring eight teams representing different cities. With a massive following, particularly among Indian fans, IPL data analysis holds significant importance. This research delves into a comprehensive statistical and exploratory analysis of the IPL dataset, aimed at extracting meaningful insights for informed decision-making within the league. By critically evaluating limitations and challenges, this study offers strategic recommendations to enhance the overall IPL experience. With a focus on teams, players, matches, and umpires, the analysis encompasses over 20 key investigations. It seeks to unveil performance disparities among teams and players, identify patterns in IPL matches, and assess the league's impact on the selection of Indian cricketers for the national team. Moreover, the research presents valuable findings that could optimize the player selection process. The insights gleaned from this extensive IPL dataset analysis can serve as a valuable tool to boost team dynamics and individual player performance.**

Index Terms—**Indian Premier League, Statistical and Exploratory Data Analysis, Cricket, Optimization**

I. Introduction

Cricket, a widely celebrated sport globally, is played among two teams of eleven players each, culminating in either victory, defeat, or, on rare occasions, a tie where neither team wins. The sport, constantly evolving, is played in three distinct formats: Test matches, One-day Internationals, and Twenty20 (T20) [1].While T20 is a contemporary favorite, ODI and Test Cricket represent the pinnacle of the sport, often contested over five days between two nations [1]. The inception of the T20 format laid the foundation for the Indian Premier League (IPL), a professional cricket league held annually from April to June [2].Spearheaded by the Board of Control for Cricket in India (BCCI), the IPL has transcended into a globally renowned and financially lucrative cricket extravaganza [3]. Commencing with eight teams in 2008, the league expanded to include two more teams for a limited duration [4]. The IPL has transformed the financial landscape for elite cricketers, propelling many to sudden stardom and substantial wealth [3]. Franchise owners, including major corporations, Bollywood luminaries, and media moguls, engage in fierce bidding wars during the IPL player auctions [3].Among the original franchises are the Mumbai Indians (MI), Chennai Super Kings (CSK), Royal Challengers Bangalore (RCB), Deccan Chargers (DCH), Delhi Daredevils (DDV), Kings XI Punjab (KXIP), Kolkata Knight Riders (KKR), and Rajasthan Royals (RR) [5].

The field of data science involves the comprehensive analysis of data to extract actionable insights, subsequently employed for practical applications [6].Consequently, IPL data analysis has emerged as a pivotal instrument for teams seeking a competitive edge. By scrutinizing extensive data on team dynamics, player performances, and match outcomes, teams can make informed decisions regarding player recruitment, strategy development, and real-time tactical adjustments.

This study categorizes the analysis into four primary segments: teams, players, matches, and umpires. Its aim is to gather and analyze data from past matches, extracting meaningful information to facilitate informed decision making. Through statistical and exploratory data analysis, the study seeks to discern performance disparities among teams and players, identify match patterns, and assess the IPL's impact on the selection of Indian cricketers for the national team. Analyzing team performances and weaknesses enables the formulation of strategies to optimize team dynamics, while assessing player performances aids in recognizing top performers and undervalued talents, informing recruitment and contract negotiations. Furthermore, analyzing match data provides invaluable insights into team performances under diverse conditions, unveiling exploitable patterns for future matches. Lastly, analyzing umpire patterns equips teams to anticipate and prepare for specific umpire tendencies. Ultimately, data analysis in the IPL revolutionizes the game by facilitating informed decisions grounded in empirical data, fostering heightened fan enthusiasm and intensified competition.

II. Related Work

Many researchers have found patterns and insights that might help teams improve their performance and make wise decisions by evaluating data on teams' performances, players' performances, and match results.

In the year 2012, P. Bhoyar and P. Agrawal carried out an empirical study on exploratory data analysis of IPL to evaluate IPL games. They used an IPL dataset consisting of 15 seasons from 2008 to 2022. From that study, they analyzed

Identify applicable funding agency here. If none, delete this.

the percentage of teams who won the IPL trophies, the overall winning percentage of each team, and the overall summary of IPL [3].

In the year 2012, P.K. Dey et al. [7] conducted a cluster analysis of the IPL using a Fuzzy clustering algorithm to classify the batting statistics of IPL T-20 version-3 cricket tournament into several clusters [7]. They used MATLAB to define a membership function and a threshold equation and to measure the distance between the centroid of the clusters and the various points [7]. Finally, they were able to detect N-clusters from the IPL batting statistics dataset.

In the year 2020, P. Banasode et al. [4] designed an application to analyze the data by fetching attributes from the dataset and predicting the future of the match and the players.The prediction was made for various factors, including which player will perform well in tomorrow's game and which side will win the coin toss and the match [4]. Their algorithm achieved over 95% accuracy.

In the year 2019, V. Kanungo and Tulasi B performed data visualization and toss-related analysis of IPL teams and batsmen's performances from the year 2008 to 2018 [1]. They identified hidden parameters, patterns, and attributes that help team owners and selectors to recognize better players [1]. Their paper highlighted player performance, especially batsmen data, and addressed multiple analyses, including the maximum number of man of the match awards, the maximum number of centuries scored by batsmen, top batsmen, and batsmen with the highest strike rate [1]. They used statistics from 696 matches for the experiment, including toss-related analyses such as the count of toss wins, the decision taken by each team after winning the toss, and season-wise and team-wise toss decisions.

In the year 2017, S. Sankaran investigated the relationship between player performance and valuation using K-means cluster analysis [8]. He applied data mining techniques to develop new performance indicators and identified four different groups of bowlers based on their effectiveness. He overlaid cluster data with team ratings in 2013 to confirm the findings and found substantial disparities in the team composition of top and underperforming teams.

The presented works in the literature study focus on specific areas of IPL, whether it is batsman data, toss-related analyses, winner prediction, or cluster analysis of player performance, etc. This study aims to analyze the data of the entire IPL series, categorizing it into four major categories: teams, players, matches, and umpires. The study aims to provide a comprehensive analysis of the IPL considering each category over the years and assess the effectiveness of the IPL for Indian cricketers in getting selected for the national team. Additionally, this analysis will aid in the decision-making process and optimize the player selection process.

III. Methodology

A. Dataset collection

For this work we used three datasets which are "IPL Ball-by-Ball 2008-2020" dataset, which consisted of 193,468 records, "IPL Matches 2008-2020" dataset, which consisted of 816 records and the "Players 2008-2020" dataset which consisted with 2164 records. The datasets are available in Kaggle.

B. Data preprocessing

After data collection next we did data preprocessing. The data preprocessing involved three main steps.

- Initially, we cleaned the data. Data cleaning involved addressing missing values, removing outliers, fixing inconsistent data points, and smoothing noisy data.
- After that we carried out data reduction. This step was employed to minimize the volume of data, consequently reducing the costs associated with data mining and data analysis.
- Lastly data transformation was done. During this stage, the data was converted from one format to another if it was needed. In essence, it involves methods for transforming data into appropriate formats that the computer can learn efficiently from.

C. Data exploration

Following data preprocessing, we undertook an unstructured exploration of the data to reveal initial patterns, characteristics, and points of interest. The approach encompassed the following key steps:

1) Variable Identification: During this phase, our focus was on identifying variables. We initiated the process with a preliminary examination of column names, gaining an initial understanding. Subsequently, we delved deeper into data catalogs, field descriptions, and metadata for a comprehensive review. This systematic exploration aided in understanding the nature of each field and detecting any missing or incomplete data.
2) Outlier Identification: Identifying outliers was a critical task in ensuring data integrity. Outliers, anomalies, or abnormalities within the data could have a substantial impact on the validity of a dataset and skew analysis results. To address this, we employed various methods, including data visualization, numerical techniques, interquartile ranges, and hypothesis testing.
3) Patterns and Relationships Determination: The final step involved exploring patterns and relationships within the data using various visualization techniques. Through comprehensive plotting of the data, we identified and explored patterns, correlations, and trends among variables. This approach ensures a thorough understanding of the dataset, providing the foundation for subsequent detailed analyses and informed decision-making.

D. Statistical and exploratory data analysis

Exploratory Data Analysis (EDA) is a crucial technique for extracting meaningful insights from data by identifying behavioral patterns among variables and forming hypotheses with minimal structures [9]. During this phase, data from previous matches was statistically and exploratorily analyzed to extract valuable information for match result analysis. A critical examination of limitations, challenges, and decisions was conducted, interpreting constraints within the dataset, and decisions were formulated based on the analyses to enhance game quality.

The most essential aspect of data visualization is to present the data visually by using charts and graphs [9]. To present the information visually we used diverse visualization methods such as bar charts, pie charts, tables, boxplots, scatter plots, and line graphs. These analyses were crucial for team selectors, captains, managers, and fans to make informed decisions. The use of NumPy and SciPy libraries facilitated numerical computing tasks, while the Pandas library handled structured data processing. Matplotlib and Seaborn libraries were employed for effective data visualization, creating a variety of informative plots. The entire analytical process was executed in Python, known for its versatility and extensive libraries in data analysis and visualization.

In this work, IPL data analysis was done by categorizing the whole analysis into four major categories which are teams, players, matches and umpires to give a proper analysis of the IPL series considering major aspects of it. There were over 20 analyses done. Following are the analyses under each category.

1) Teams: Number of matches played by each team, Match by match performance of each team over seasons, Number of matches won by each team over seasons, How many times have unique IPL finalists have reached the finals?, Do the teams with highest win counts also tops in the chart of highest win percentage?, Variation of the winner teams' scores in each year.
2) Players: Analysis for IPL debut players who got them a place in the national team, Comparison of top players with highest average runs in IPL, Clustering on batsman data, The top batsman in the IPL and his performance variation throughout the seasons, IPL Bowler wise bowling statistics, Find top 10 high value players over the years, Relationship between player value and Run Above Average (RAA), Most lasting partnerships in each year and they vary.
3) Matches: Probability of match winning with toss winning according to the ground, Number of matches per venue, Does giving more extra runs effects to loss the match?, Season wise IPL matches, IPL venue wise toss decision and venue wise match result information.
4) Umpires: Analysis of umpire performance, Top umpires who took over matches in IPL

IV. Experiments & Results

A. Teams

1) How many times have unique IPL finalists have reached the finals?
In TABLE. I, analyzing IPL finals reveals that CSK holds the record for the highest number of appearances, reaching the finals 8 times, followed by MI with 6 appearances. MI stands out as the most successful, winning 5 finals. Notably, KTK and Pune Warriors (PW) have never reached the IPL finals, while other teams have made finals appearances within the 3 to 1 range. Teams like CSK and MI have consistently demonstrated success by reaching the finals multiple times and clinching the title, while struggling franchises like KTK and PW have yet to make an impact in the IPL finals.
2) Do the teams with highest win counts also tops in the chart of highest win percentage?
In Fig. 1, the top five IPL teams with the highest win counts are MI (120 matches), CSK (106 matches), KKR (99 matches), RCB (91 matches), and KXIP (88 matches). These teams consistently outperform their counterparts. Regarding winning percentage, the top five teams are CSK (59.55%), MI (59.11%), Sunrisers Hyderabad (SUN) (53.23%), KKR (51.56%), and RR (50.31%). This indicates their success in converting matches into wins more frequently than others, with teams having high win counts also boasting high winning percentages, except for outliers like RCB and KXIP.
3) Variation of the winner teams' scores in each year
According to Fig. 2, the given data represents the total runs scored by the winning teams in the IPL over a period of 13 years. The data reveals that the highest total runs scored by a winning team is in 2016 with 208 runs, followed closely by 2011 and 2015 with 205 and 202 runs respectively. The lowest total runs scored by a winning team was in 2017 with just 129 runs, while the other teams scored between 143 and 192 runs. The average score of the winning teams in the IPL is around 180 runs. Further analysis under the team category is added to Appendix A.

B. Players

1) Analysis for IPL debut players who got them a place in the national team
This analysis assesses the impact of the IPL on the selection of Indian cricketers to the national team. Fig. 3 displays the number of matches players played after their IPL debut before being selected for the national team. The positive skew of 1.86 suggests that many players were selected with relatively few IPL matches.

TABLE I
Number of times the teams reached finals

Team	Times Reached Finals	Winning Count
CSK	8	3
RCB	3	0
KKR	2	2
MI	6	5
DC	1	0
RR	1	1
DCH	1	1
KXIP	1	0
SRH	2	1
RPS	1	0
KTK	0	0
PW	0	0

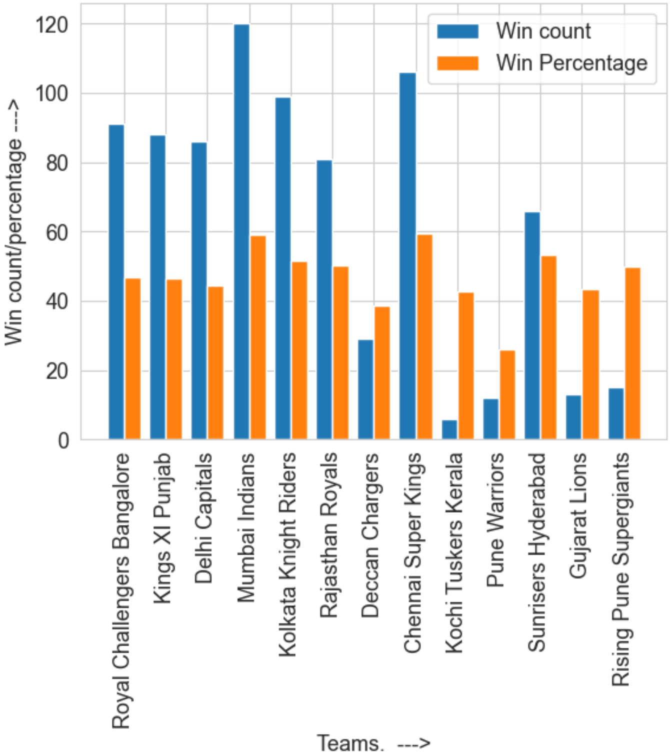

Fig. 1. Win counts and percentages of IPL teams.

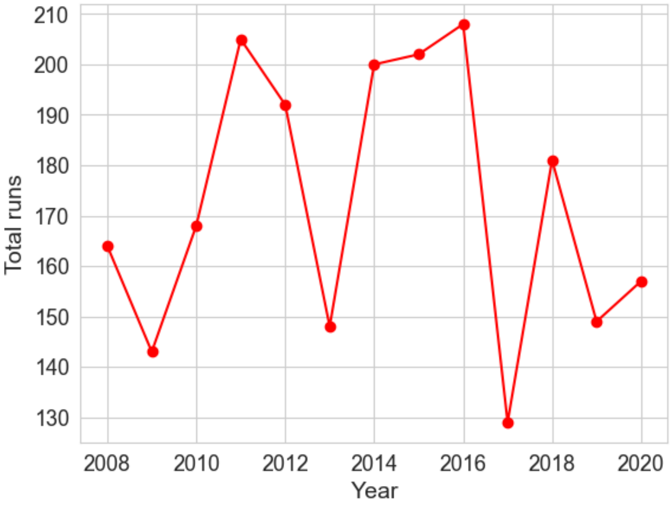

Fig. 2. Variation of the winner teams' scores in each year

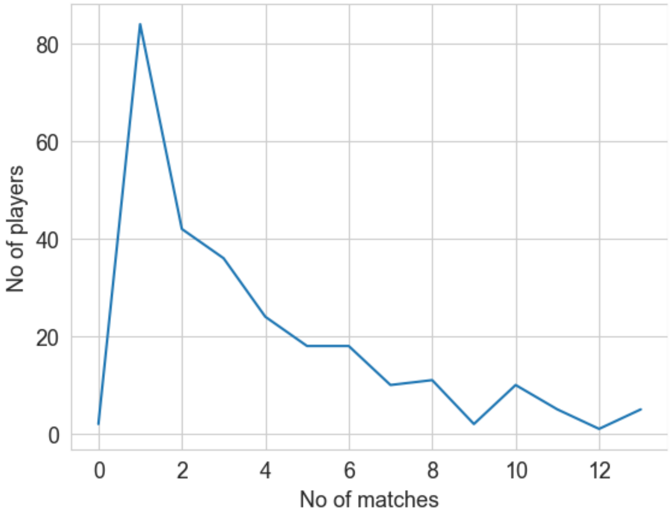

Fig. 3. Number of matches players had to play after their debut to get selected to the national team

While the IPL contributes to player development, other factors like overall performance, skills, and consistency are crucial for national team selection. Nevertheless, data over the years indicates a positive influence of IPL participation, aiding in skill improvement, confidence-building, and overall development, as seen in numerous players successfully transitioning to the national team.

2) Clustering on batsman data
This analysis optimizes player selection in the IPL by employing K-means cluster analysis on boundary percentage and strike rate. The elbow method was utilized to determine the optimal number of clusters, with the squared difference between various k values calculated to assess the within-cluster sum-of-squares (WCSS) [10]. Here, we used (1) to calculate the WCSS (inside-Cluster Sum-of-Squares).

$$WCSS = \sum_{c_k}^{c_n} \left(\sum_{d_i \cdot in \cdot c_i}^{dm} \text{distance}(d_i, c_k)^2 \right) \quad (1)$$

Fig. 4 demonstrates an abrupt change forming an elbow, indicating the optimal number of clusters is 3. Fig. 5 illustrates the K-means cluster analysis, where the second cluster comprises players with high strike rates and boundary percentages, the first cluster with moderate rates, and the zeroth cluster with low rates for both parameters.

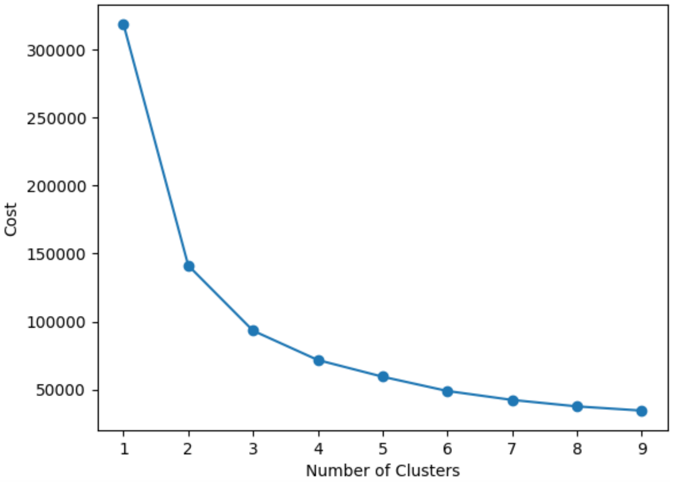

Fig. 4. Elbow curve

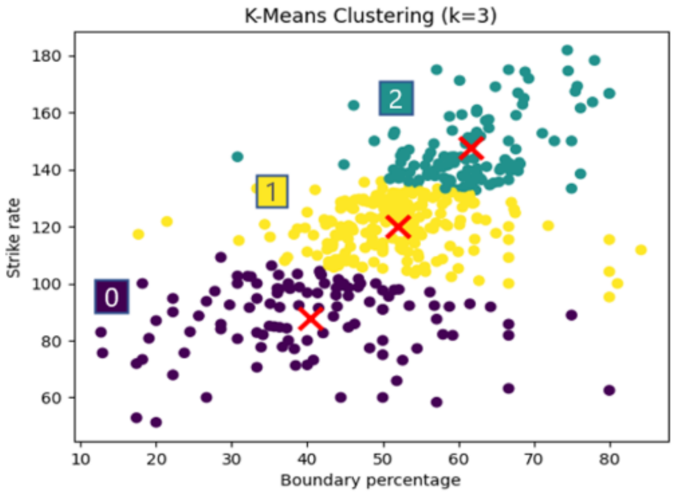

Fig. 5. K-means Cluster Analysis of Boundary Percentage and Strike Rate

3) Relationship between player value and Run Above Average (RAA)
To find the relationship between a player's value and RAA we have used Pearson correlation coefficient. The strength and direction of a linear relationship between two variables is measured by the correlation coefficient [11]. According to the TABLE. II, in the years of 2020, the correlation between RAA and player value is almost 1. For other years also the data values are almost same. Those data is available in Appendix B.
So, we can say that there is a strong positive correlation between RAA and player value. If there is an almost linear relationship between player value and RAA in IPL, it implies that teams are willing to pay a premium for players who have a consistently high RAA.
Further analysis of players is added to Appendix B.

C. Matches

1) Number of matches per venue and venu wise toss decision
Fig. 6 highlights that the M. Chinnaswamy Stadium hosted the highest number of IPL matches (80), followed by the Eden Gardens stadium in Kolkata (77). According to Appendix C, teams winning the toss at the MA Chidambaram Stadium in Chepauk predominantly chose to bat first (36 matches), while the Feroz Shah Kotla stadium in Delhi saw a higher inclination for teams to field first (42 matches). Notably, at M. Chinnaswamy Stadium, teams winning the toss often opted to field first, with a 53.8% probability of winning matches in this scenario.

TABLE II
CORRELATION OF THE RAA AND PLAYER VALUE IN 2020

	RAA	Value
RAA	1.000000	0.999997
Value	0.999997	1.000000

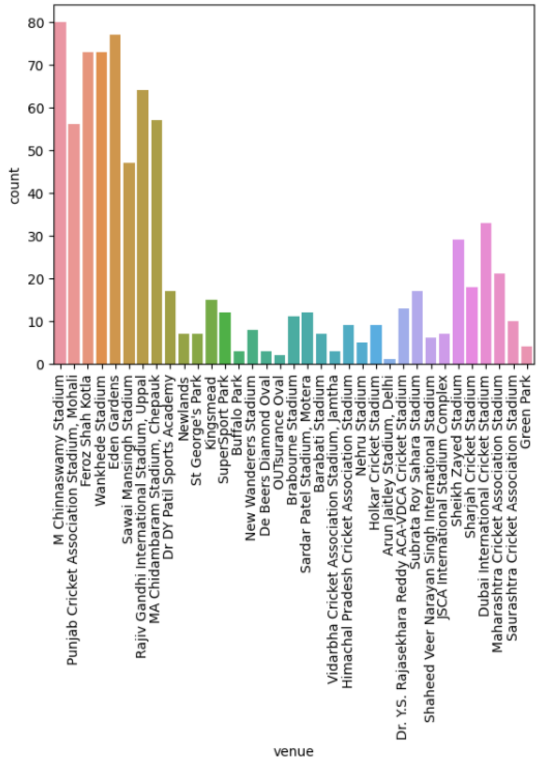

Fig. 6. Number of matches per venue.

2) Does giving more extra runs effects to loss the match?
In Fig. 7, the graphical representation illustrates the number of matches lost by IPL teams due to conceding extra runs, where the X-axis denotes the extra runs given, and the Y-axis indicates the corresponding matches lost. The skewness value of 0.58 suggests a moderately right-skewed distribution, indicating a positive relationship between the number of extra runs given and matches lost. The graph implies that teams conceding fewer extra runs are more likely to win, while those giving more extra runs are prone to losing matches. TABLE. III provides a detailed breakdown of the probability of losing matches based on grouped quartile ranges of extra runs. Notably, in the ranges of 0 to 6, 6 to 13, and 19 to 28, there is approximately a 50% probability of both winning and losing matches. However, the range of 13 to 19 exhibits a higher probability of match loss, reaching 56.59%.
Further analysis of matches is added to appendix C.

TABLE III
PROBABILITY OF LOSING THE MATCH BY GIVING EXTRA RUNS

Extra Runs	Probability of Losing Matches (%)
0 to 6	51.50
6 to 13	47.68
13 to 19	56.59
19 to 28	51.72

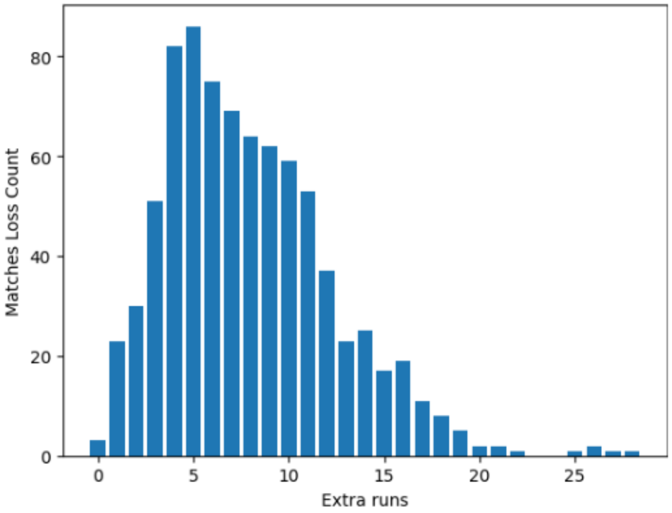

Fig. 7. Matches loss count by giving extra runs.

D. *Umpires*

1) Analysis of umpire performance
The umpires' details throughout the years have been attached to appendix D. According to the data it could be seen that most umpires have played the role as both umpire 1 and umpire 2 meanwhile some of them didn't perform as umpire 2.
2) Top umpires who took over matches in IPL
TABLE. IV presents the top umpires in IPL matches, with S Ravi officiating the highest number at 121 matches. HDPK Dharmasena follows with 94 matches, AK Chaudhary with 87 matches, C Shamshuddin with 82 matches, and M Erasmus with 65 matches. These umpires boast significant experience, each officiating more than 65 matches.

TABLE IV
TOP UMPIRES WHO TOOK OVER MATCHES IN IPL

Umpire	Matches
S Ravi	121
HDPK Dharmasena	94
AK Chaudhary	87
C Shamshuddin	82
M Erasmus	65

V. Conclusion

This study provides a comprehensive overview of the IPL tournament, covering teams, players, matches, and umpires. Our analysis reveals key trends and patterns, highlighting the performance disparities among teams and players. Notably, teams like MI and CSK consistently perform well, while others face challenges in maintaining stability.

In terms of player performance, variations are observed among individuals, with players like Virat Kohli and Suresh Raina consistently demonstrating excellence. The relationship between IPL performance and national team selection is complex, suggesting that strong IPL performance may enhance a player's chances of selection. The study carries significant implications for teams, players, and umpires seeking to enhance IPL performance. Insights can guide improvements, aiding teams and players in identifying areas for development and formulating strategies. Additionally, umpires can use this information to maintain high standards throughout the tournament. In conclusion, our analysis contributes valuable insights to the comprehensive understanding of the IPL.

VI. Acknowledgement

This research was supported by the Science and Technology Human Resource Development Project, Ministry of Education, Sri Lanka, funded by the Asian Development Bank (Grant No. STHRD/CRG/R1/SJ/06).

References

[1] V. Kanungo and T. B, "Data visualization and toss related analysis of ipl teams and batsmen performances," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, p. 4423, 2019.

[2] P. Kansal, P. Kumar, H. Arya, and A. Methaila, "Player valuation in indian premier league auction using data mining technique," in *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, 2014.

[3] P. Bhoyar and P. Agrawal, "Exploratory data analysis of indian premier league: An empirical study," *IJFANS International Journal of Food and Nutritional Sciences*, vol. 11, no. 3, pp. 4125–4130, 2022.

[4] P. Banasode, M. Patil, and S. Verma, "Analysis and predicting results of ipl t20 matches," *IOP Conference Series: Materials Science and Engineering*, vol. 1065, no. 1, p. 012040, 2021.

[5] A. Longmore, "Indian premier league," https://www.britannica.com/topic/Indian-Premier-League, Mar 2023, encyclopædia Britannica, 20-Mar-2023.

[6] S. G, A. Swaminathan, J. B. J, S. R, and L. Nelson, "Ipl data analysis and visualization for team selection and profit strategy," in *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, 2023.

[7] P. K. Dey, G. Chakraborty, P. Ruj, and S. Sarkar, "A data mining approach on cluster analysis of ipl," *International Journal of Machine Learning and Computing*, pp. 351–354, 2012.

[8] S. Sankaran, "Comparing pay versus performance of ipl bowlers: An application of cluster analysis," *International Journal of Performance Analysis in Sport*, vol. 14, no. 1, pp. 174–187, 2014.

[9] P. Abelairas-Etxebarria and I. Astorkiza, "From exploratory data analysis to exploratory spatial data analysis," *Mathematics and Statistics*, vol. 8, no. 2, pp. 82–86, 2020.

[10] M. Cui, "Introduction to the k-means clustering algorithm based on the elbow method," *Accounting, Auditing and Finance*, vol. 1, no. 1, pp. 5–8, 2020.

[11] A. T. A. E. B. Erhardt. Passion driven statistics, 13.2 the correlation coefficient. https://statacumen.com/teach/S4R/PDS_book/the-correlationcoefficient.html. [Online].

Appendix A

Analysis of teams

Appendix B

Analysis of players

Appendix C

Analysis of matches

Appendix D

Analysis of umpires