**Exploratory Analysis on Cricket: Unveiling Insights from ODI Cricket Data**

Vibha Naiknavare

Carlo Lopez Hernandez

University of Notre Dame - Mendoza School of Business

MSBA - Sports Analytics

**Sports Analytics - 60520-01**

**2023 Fall**



In the realm of cricket, One Day International (ODIs) stands as a thrilling fusion, blending the strategic depth of the sport with the intensity of limited overs, Spanning 50 overs for each team, ODI cricket crafts a high-stakes narrative within a single day, delivering encounters that showcase the essence of skill, strategy, and adaptability. With a rich history and global stage, ODI Cricket stands as the centerpiece in international cricket, featuring iconic tournaments and moments that resonate with fans worldwide.

**Motivation:**

The proposed project aims to unravel trends, strategies, and performances in ODI cricket using historical match data. Analyzing these datasets allows us to understand the various factors influencing match outcomes, team strategies, and individual player performances. We want to find interesting things in the numbers, like patterns and connections, to better understand the game of cricket. By digging into these stats, we aim to make the complex world of cricket a bit clearer for everyone

**Problem Framing:**

The project focus aims to explore and uncover insights from the cricket data:

**Exploratory Match Outcome Analysis**: Through statistical exploration and modeling, this project seeks to reveal trends and potential relationships between match outcomes and factors such as toss decisions, venue dynamics, and team compositions. Delving into these variables aims to uncover correlations and patterns that may offer valuable insights into match result forecasts.

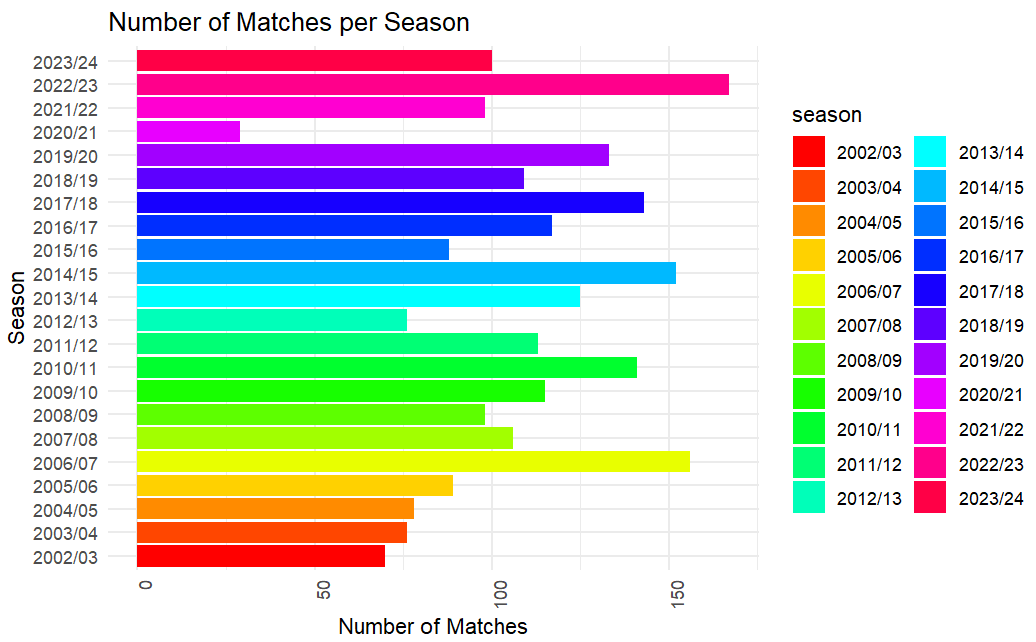
**Data Overview**:

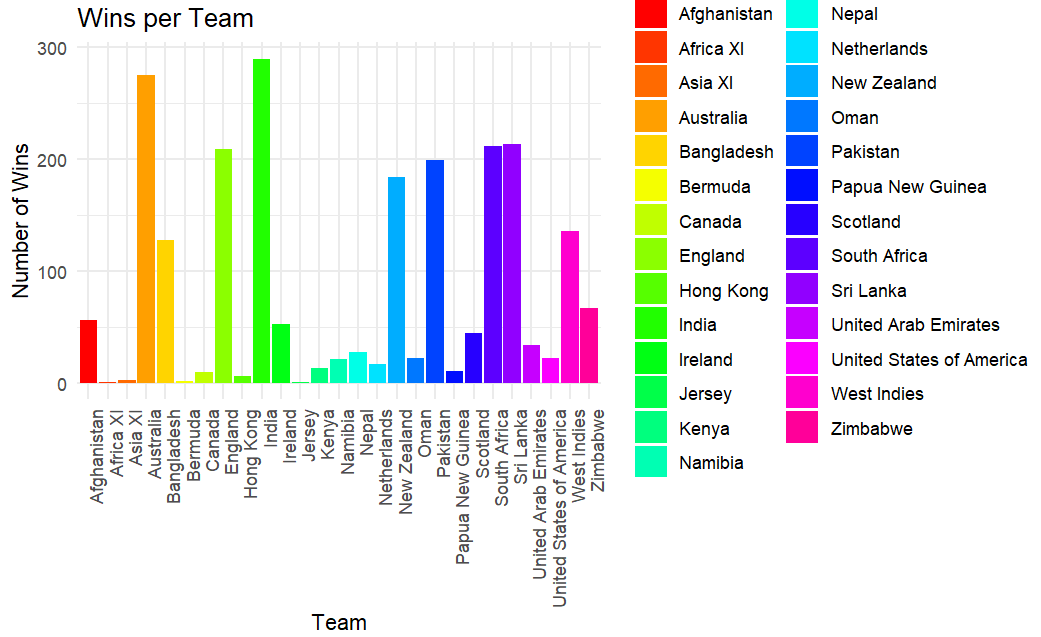
The primary dataset comprises two main sections:

Match Information: Includes details such as season, teams, toss details, match results, DL method, and player awards spanning from 2002 to 2024 (appendix 1).

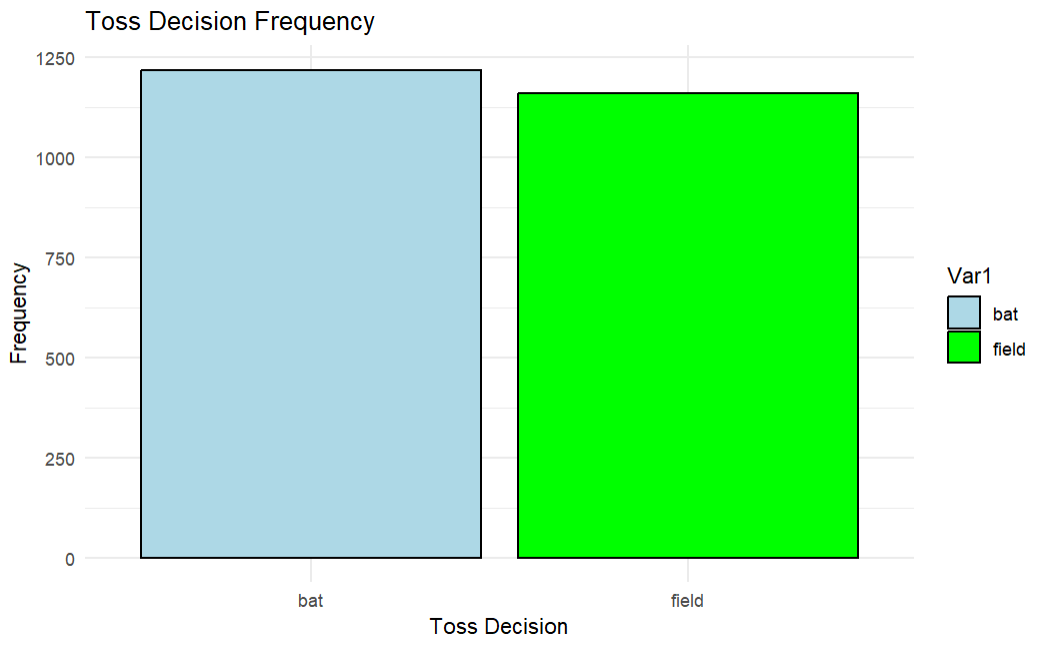
Ball-by-Ball Data: Provides granular insights into each ball, encompassing batting and bowling team details, player involvement, runs scored, extras, and match-specific IDs (appenix2). We will focus on the 2023/24 season for our project.

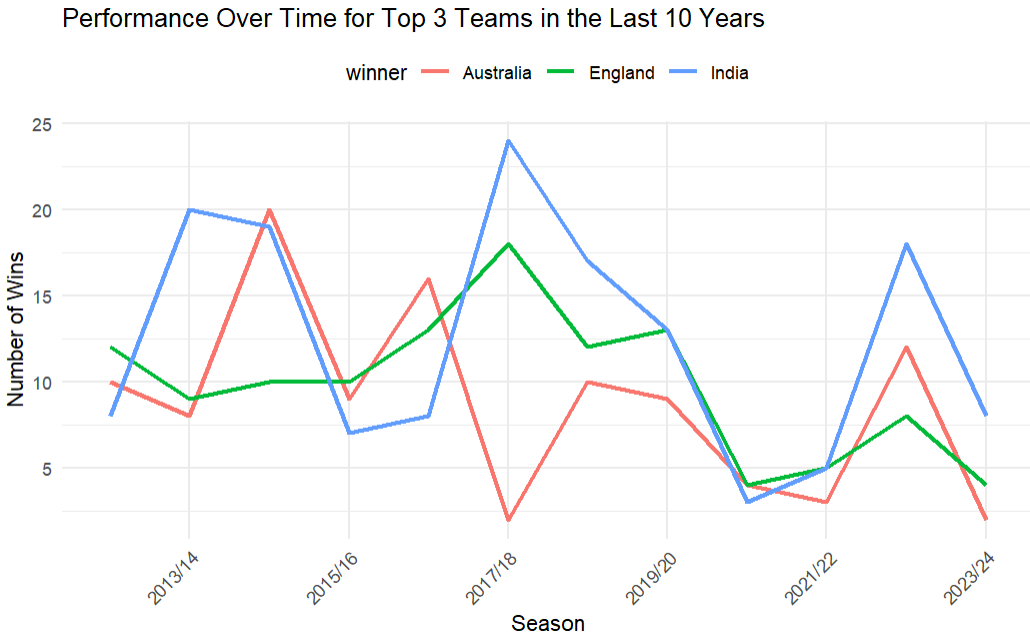
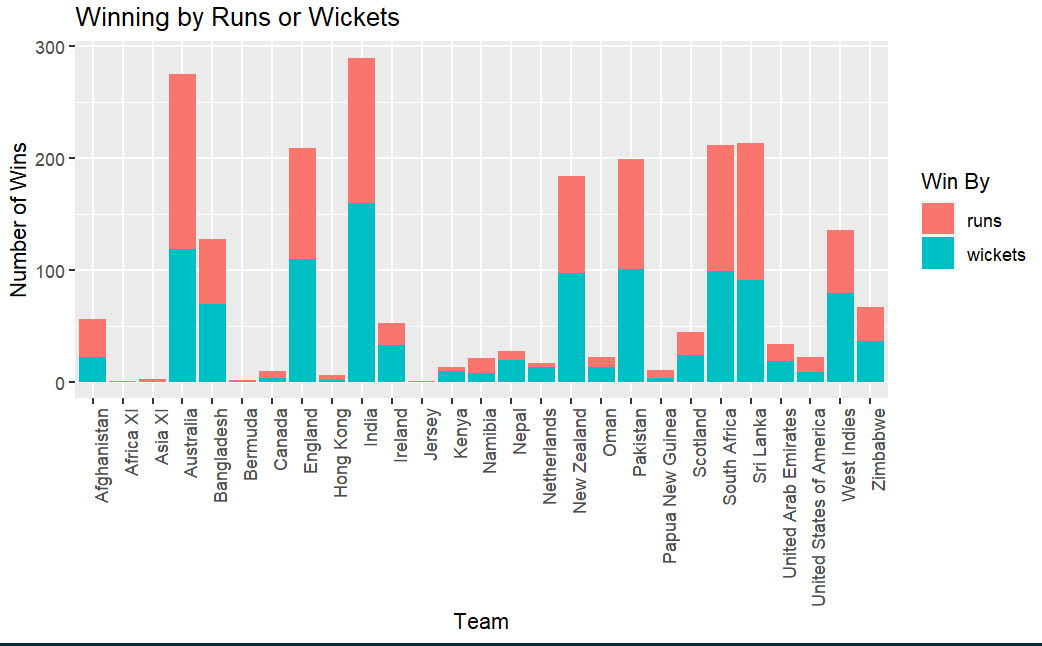
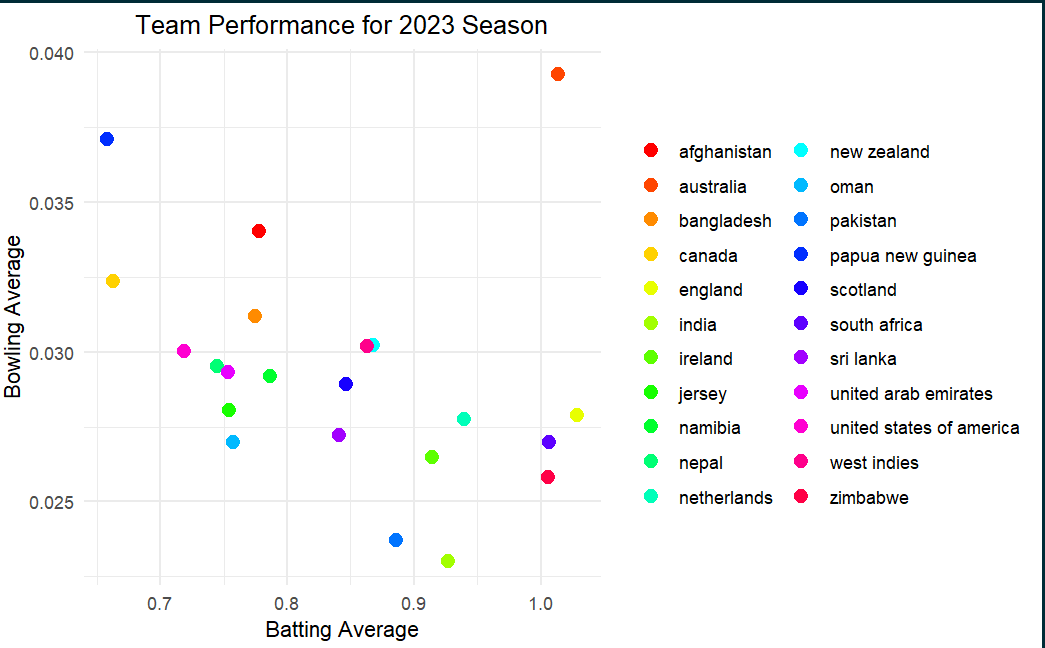
**Summary Visualizations and Findings:**

1. **Number of Matches per Season (appendix 3)**: The chart represents the number of matches per season revealing a fluctuating trend over the year. Notably, there is a significant increase in the number of matches played compared to 18 years ago. The 2020/2021 season season stands out with the lowest match count, attributed to the impact of Covid-19 and cancellations. 
2. **Wins per Team Over All Seasons (appendix 4)**: From insights, predominantly, England, India, Australia, Sri Lanka, South Africa, and Pakistan have the majority of wins, while lesser-known teams like Jersey, Kenya, and Guinea have fewer wins. It's crucial to note that teams with minimal wins are typically not part of major playoffs, as evidenced by only 10 teams participating in the 2023/2024 World Cup.



1. **Toss Decision (appendix 5)**: The data reveals that teams winning the toss opted to bat 1218 times and chose to field 1161 times throughout the season. This indicates a relatively balanced distribution of tess decisions between batting and fielding, showcasing the strategic choices made by teams before matches.



1. **Winning Strategies via Stacked Bar (appendix 6)**: Unveils winning strategies of teams, showcasing preferences in winning by runs or wickets. The presented visualization provides insights into teams' winning strategies. Matches won by runs (when batting first) are depicted by orange bars, while matches won by wickets (when batting second) are represented by blue bars. Notably, Australia and Sri Lanka exhibit a higher number of victories by runs, showcasing proficiency in setting and defending targets. Conversely, West Indies, Ireland, and India demonstrate a substantial count of wins by wickets, emphasizing their strength in chasing down targets. The diversity in winning strategies across teams underscores the impact of factors like team composition, pitch conditions, and opposition strength on match outcomes.
2. **Top Three Teams' Performance Over 10 Years (Line Graph) (appendix 7)**: Tracks the performance trends of the top teams across seasons, revealing consistent and fluctuating performances. The visualization offers insights into the performance trends of the top three cricket teams over a decade. India emerges as a consistently strong performer, securing a substantial number of wins. Noteworthy peaks in India's performance are evident in the 2013/14, 2017/18, and 2022/23 seasons, where they outperformed their counterparts. Certain seasons, such as 2012/13 and 2014/15, witnessed closely contested matches among the top three teams, reflecting a competitive environment and balanced performances. While India's performance remained robust throughout, Australia and England displayed more variability in the number of wins, indicating fluctuations influenced by factors like team composition, strategies, and opposition strength.
3. **Correlation Matrix:** The correlation matrix reveals the relationship between various variables in the match information dataset. The findings demonstrate a weak negative correlation with the DL method (An indicator of whether the Duckworth-Lewis method was applied (1 for applied, 0 for not applied)) with both run-based and wicket-based victories. Moreover, a moderate negative correlation is observed between winning by runs and winning by wickets, suggesting that matches with larger run differences are less likely to be won by a higher number of wickets, and vice versa. Overall, these correlations provide insights into key match-related parameters.
4. **Batting and Bowling Average in 2023 (appendix 8)**: The batting average represents the mean number of runs scored per wicket lost, indicating a team’s ability to score runs. A higher batting average suggests better run-scoring efficiency. In the given findings Australia, England, and Zimbabwe stand out with notably high batting averages, showcasing their ability to score runs effectively. On the other hand, the bowling average signifies the mean number of runs conceded per wicket taken, reflecting a team's effectiveness in limiting opponents' runs. Lower bowling averages indicate stronger bowling performances. Teams such as India and Pakistan exhibit lower bowling averages, highlighting their success in bowling. Together, these averages contribute to a comprehensive understanding of a team's overall performance, balancing offensive and defensive aspects of their gameplay. Teams with high batting averages and low bowling averages are generally considered strong contenders in cricket competitions for the 2023/24 season.
5. **Distribution of Runs in the 2023/24 season:**

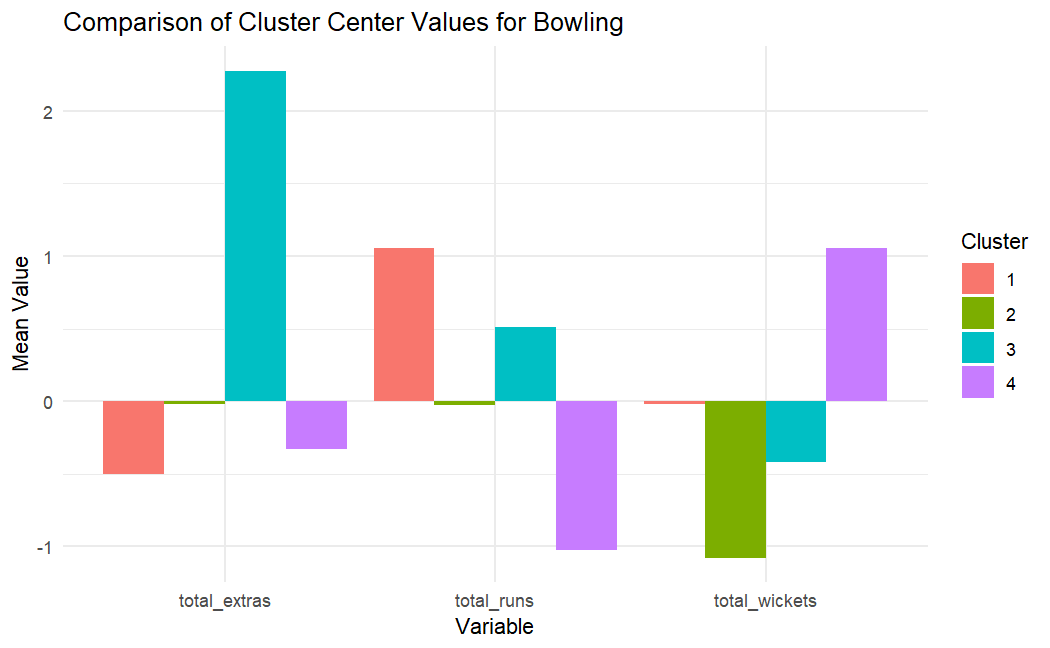
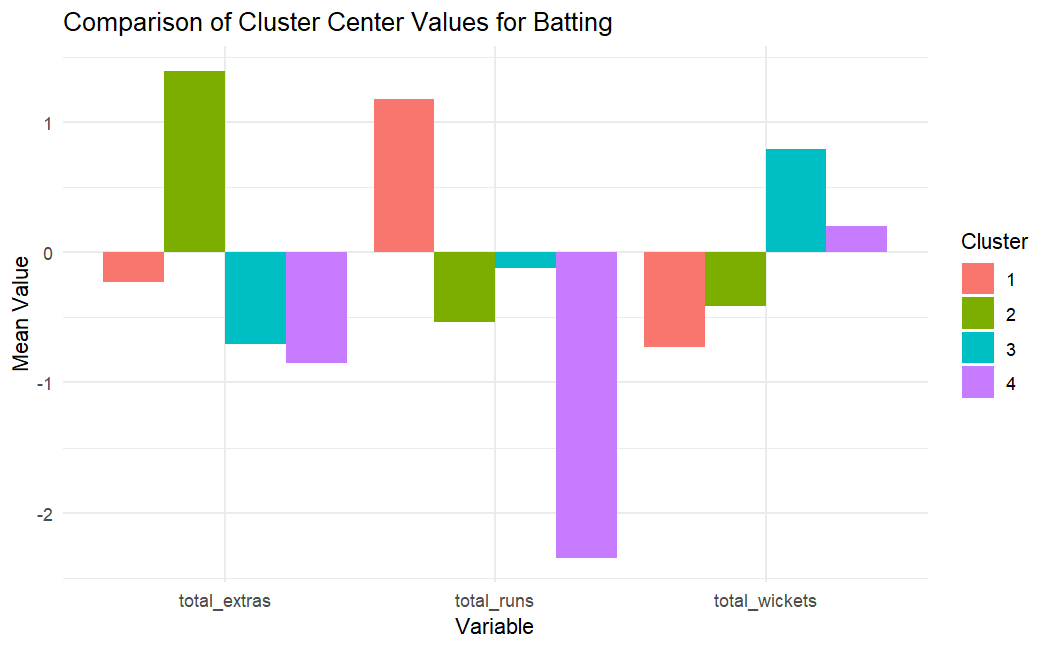
The Distribution of Runs Scored Off the Bat: shows that scoring 0 or 1 run is the most common outcome, with the frequency decreasing as the number of runs increases. This is typical in cricket, where singles are more common than higher-scoring outcomes like boundaries (4 or 6 runs). The Distribution of Extras indicates that extras are relatively rare compared to runs scored off the bat, with most deliveries not resulting in any extras.

**Models and Results:** In our analysis of the recent 2023/24 cricket seasons, we analyzed K-means and Hierarchical clustering to find patterns in team performance. Since teams played different numbers of matches, we focused on a fair sample, taking five matches per team. This approach helps us better understand how teams perform in various situations. The insights gained will be crucial for making informed predictions in future matches, as we unravel the complexities of cricket dynamics.

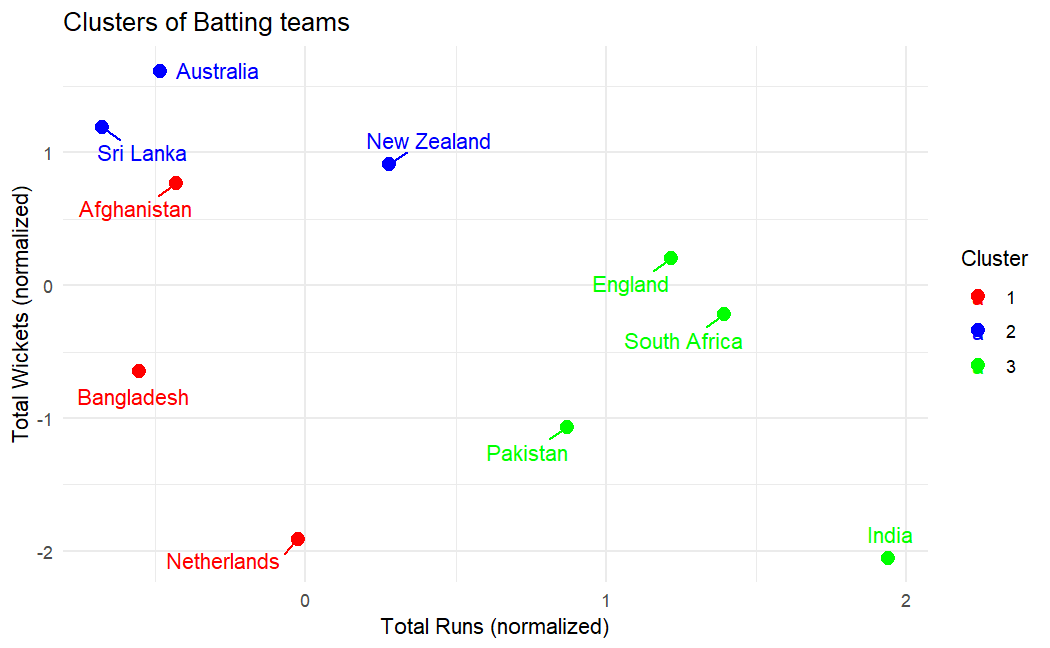
**Initial Clustering:**

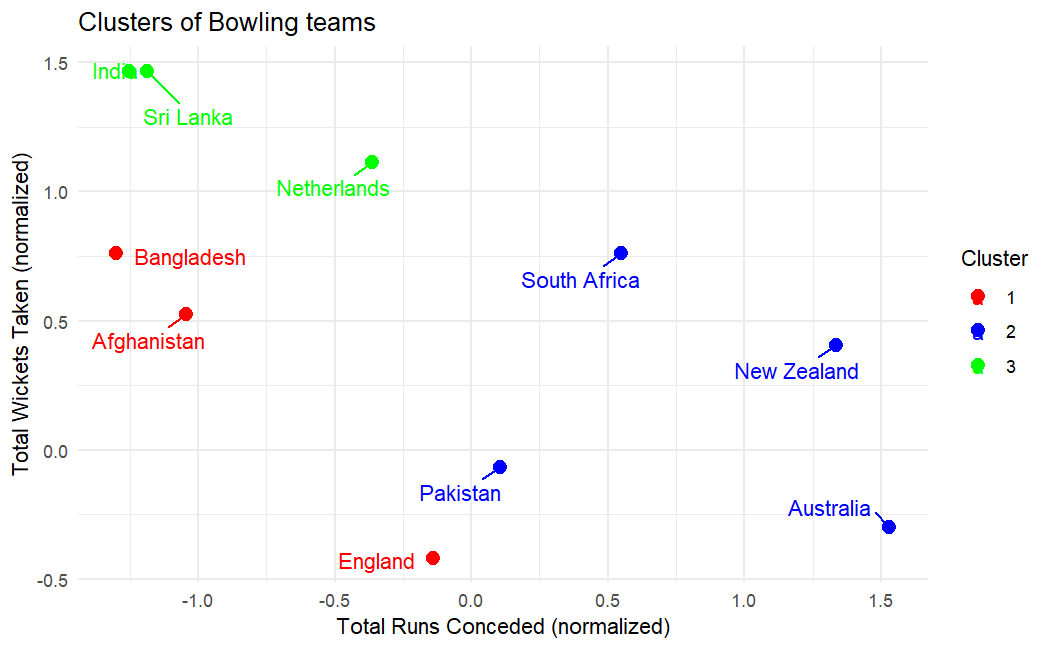
In our initial clustering analysis of batting, four distinctive clusters have been identified based on total runs, total extras, and total wickets. Cluster 1 encompasses teams like England, India, Ireland, Pakistan, and South Africa, indicating a shared batting performance pattern characterized by substantial runs, limited extras, and moderate wickets. Cluster 2 includes Afghanistan, Bangladesh, Netherlands, Oman, and the United Arab Emirates, suggesting a different batting approach with notable variations in runs, extras, and wickets. Cluster 3 consists of Australia, New Zealand, Scotland, Sri Lanka, the United States of America, West Indies, and Zimbabwe, showcasing teams with diverse batting strategies. Notably, Cluster 4 features Nepal, which stands out for its unique batting profile, marked by relatively lower runs, higher extras, and a moderate number of wickets.

In the bowling stats cluster analysis, four distinct groups of cricket teams were identified based on their performance metrics, including total runs, total extras, and total wickets. Cluster 1 comprises well-balanced performers such as Australia, New Zealand, Pakistan, South Africa, and the United Arab Emirates, showcasing consistency across all metrics. Cluster 2 represents teams with varied approaches, including England, Nepal, Oman, the United States of America, and the West Indies. Cluster 3 highlights efficient wicket-taking teams, namely Ireland and Zimbabwe, with a specific emphasis on effective wicket strategies. Lastly, Cluster 4 features aggressive run-scoring teams, such as Afghanistan, Bangladesh, India, the Netherlands, Scotland, and Sri Lanka, indicating a more offensive approach in their bowling strategies. This clustering provides in depth understanding of the distinct batting and bowling styles exhibited by different cricket teams based on their statistical performance.

****

**K-means Clustering:**

In our focused analysis on the final playoffs of the ODI World Cup 2023, which included Australia, Netherlands, New Zealand, England, India (host), Bangladesh, Pakistan, Afghanistan, South Africa, and Sri Lanka, we applied K-Means clustering to gain insights into the batting performances of these elite teams. After excluding non-playoff teams, we identified three clusters through the elbow method. Cluster 1 comprised Afghanistan, Bangladesh, and Netherlands, displaying moderate total runs and wickets. Cluster 2 included Australia, New Zealand, and Sri Lanka, showcasing diverse patterns in total runs and wickets. Cluster 3 encompassed England, India, Pakistan, and South Africa, demonstrating a high level of total runs and wickets.



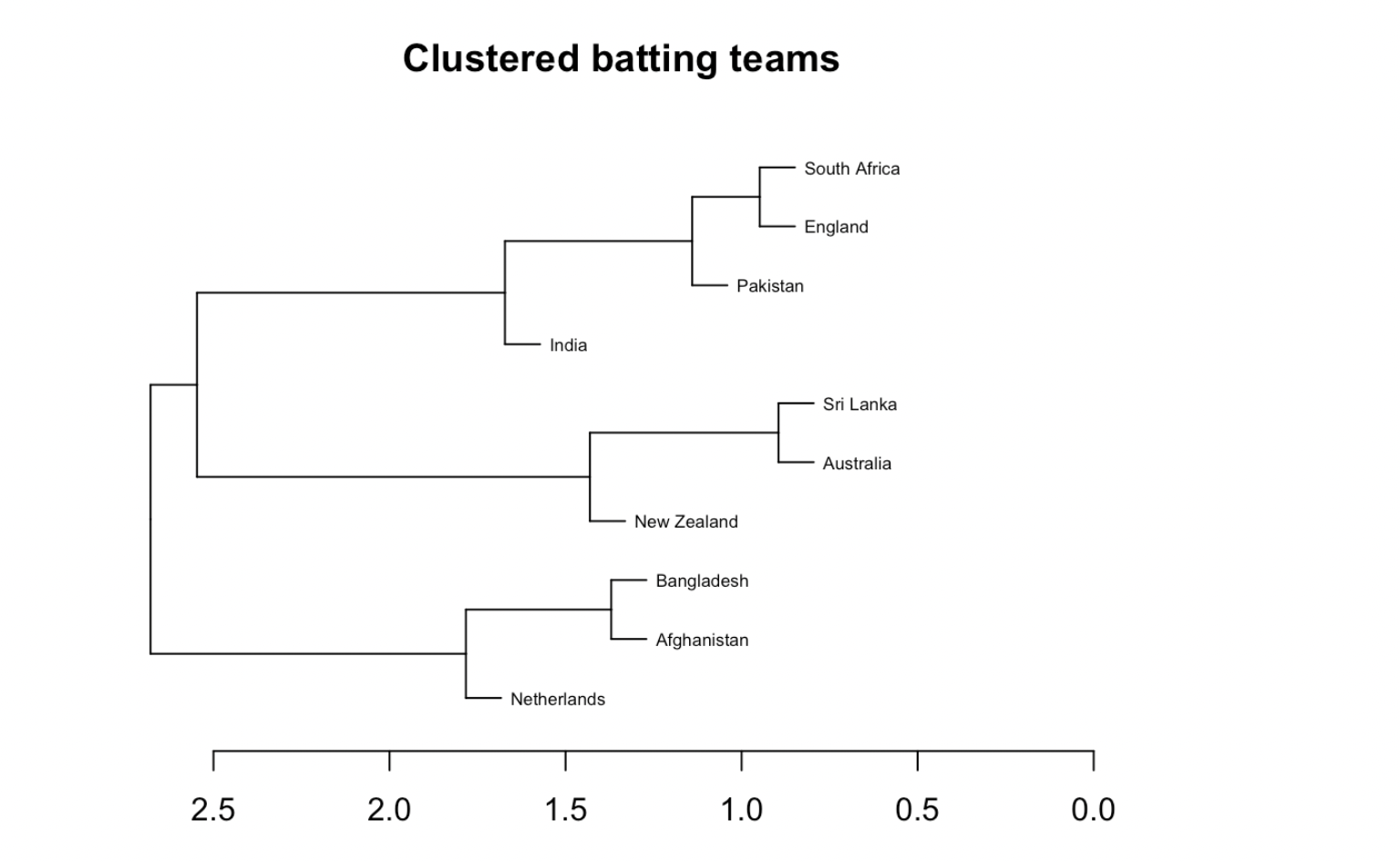
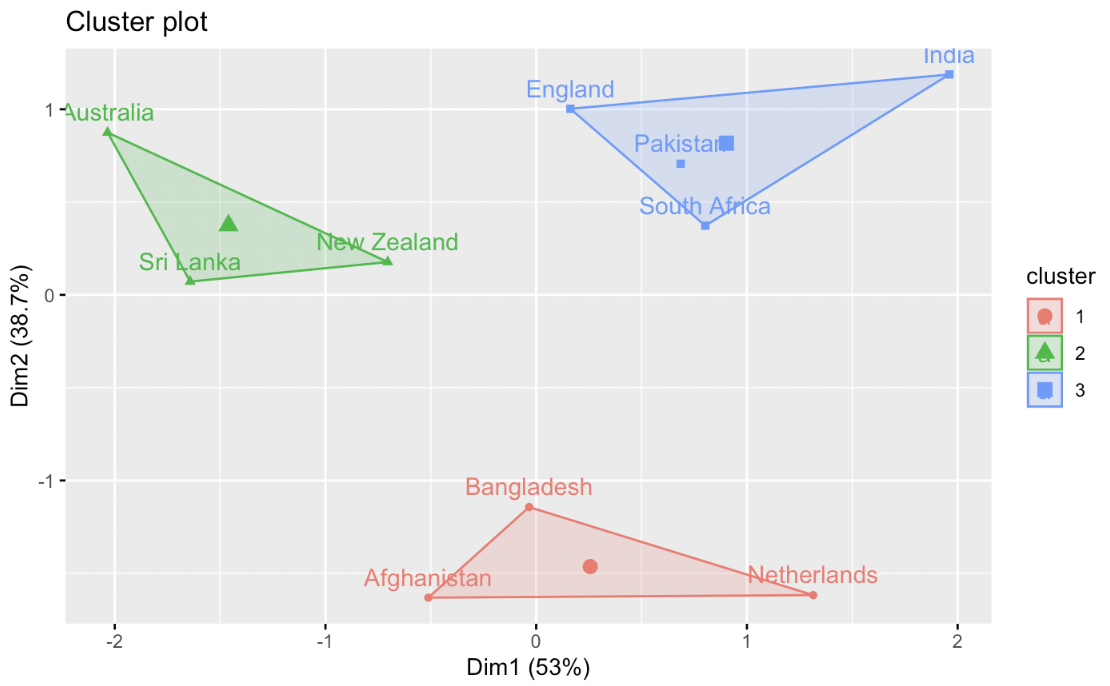
Similarly, we analyzed bowling patterns, cluster 1 comprised Afghanistan, Bangladesh, and England, demonstrating unique patterns in total runs conceded and wickets taken. Cluster 2 included Australia, New Zealand, Pakistan, and South Africa, showcasing diverse characteristics in normalized total runs conceded and wickets taken. Cluster 3 encompassed India, the Netherlands, and Sri Lanka, displaying specific tendencies in the same metrics. The scatter plot visually represents these clusters, offering a comprehensive overview of the distinctive bowling dynamics within the playoff teams.

**Hierarchical clustering**

Batting Clustering:

For this analysis, we examined teams that participated in the ODI World Cup. We divided the data into two sections to check if there would be distinct clusters for batting and bowling teams. We began our analysis with the batting teams, where we divided them into three clusters. Groupings for batting included:

* Cluster 1: Bangladesh, Afghanistan, Netherlands
* Cluster 2: Australia, Sri Lanka, New Zealand
* Cluster 3: England, Pakistan, South Africa, India

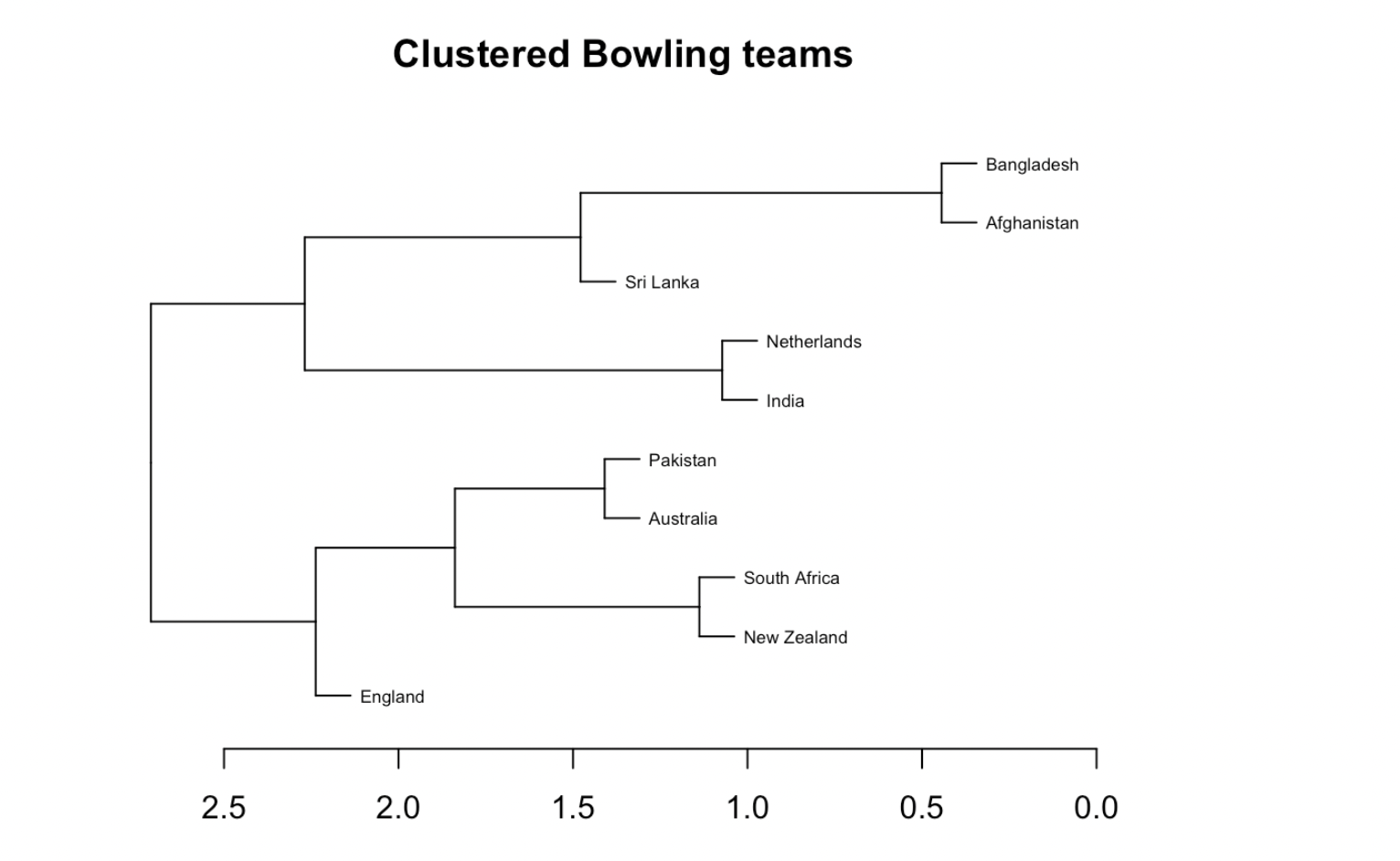
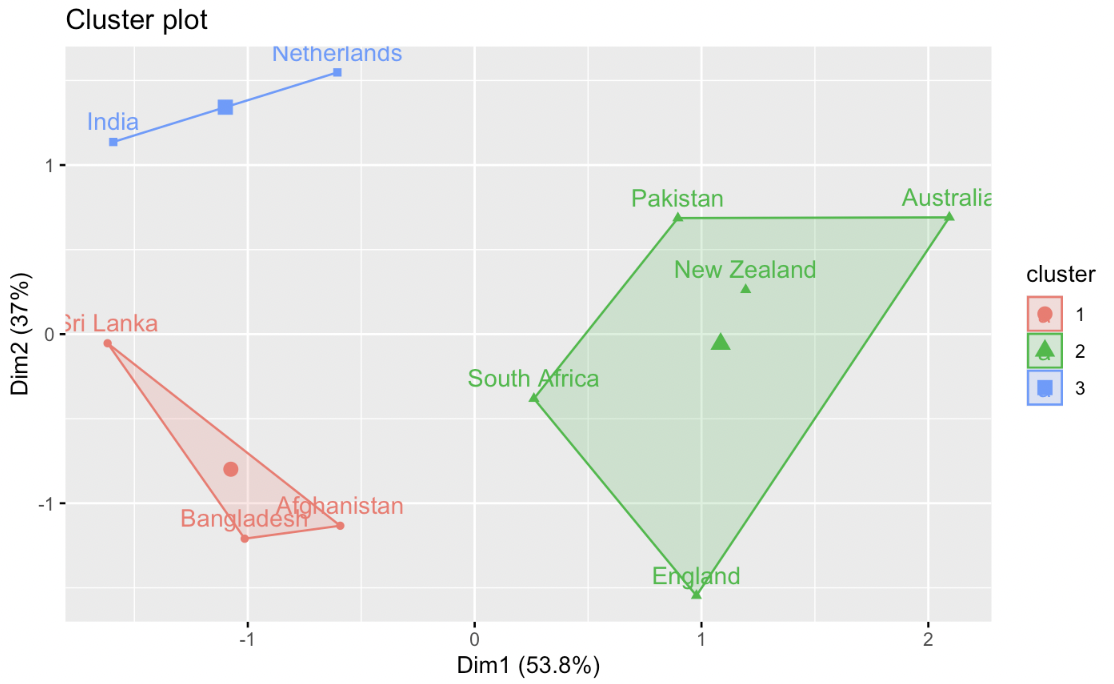
****

We observe that South Africa and England have the most similar batting styles, with both nations having a very similar number of runs. Sri Lanka and Australia had the highest number of wickets. The most distinct batting styles belong to India (the highest number of runs with low wickets) and the Netherlands (who had the lowest wickets), with each not having any joins until later on the tree. In batting analysis, teams aim for a high number of runs and extras, paired with a low number of wickets. Therefore, India's playing style would be the most beneficial to mimic.

**Bowling Clustering**:

Our hierarchical clustering analysis was then repeated on each team, but this time focused on bowling. The test grouped the countries in the following way:

* Cluster 1: Sri Lanka, Netherlands
* Cluster 2: Pakistan, New Zealand, South Africa, Australia, England
* Cluster 3: India, Netherlands



In these clusters, we observed that cluster three included countries with the highest number of runs, whereas the first cluster had the lowest. According to the tree diagram, England stood out for having the most distinct playing styles, marked by a very high number of extras and the lowest number of wickets. Unlike batting, in bowling, teams aim for a low number of runs coupled with a high number of wickets. Sri Lanka demonstrates one of the best combinations for these factors, explaining why it is among the more unique countries in the cluster.

**Conclusion and Future Work:**

In conclusion, clustering proved invaluable in uncovering patterns within our dataset. Looking forward, our focus will shift to more targeted exploration, focusing more on the statistical profiles of two specific counties. This approach aims to lay the groundwork for predictive modeling of match outcomes based on deeper insights derived from two teams' statistics. The ultimate goal is to develop accurate forecasts, providing valuable tools for sports analysis and decision-making in Cricket.

**Contribution:**

Carlo Lopez Hernandez: Presentation, Hierarchical Clustering, and Report

Vibha Naiknavare: Data Visualization, Initial, and K-means Clustering, and Report

**Appendix**

Appendix 1: Match information data

1. id A unique identifier for each cricket match.
2. season The season or year in which the match took place.
3. city The city where the match was held.
4. date The date on which the match was played.
5. Team 1 The name of the first cricket team participating in the match.
6. Team 2 The name of the second cricket team participating in the match.
7. toss\_winner The team that won the toss.
8. toss\_decision The decision made by the toss-winning team (bat or field).
9. result The result of the match (e.g., "normal," "tie," "no result").
10. dl\_applied An indicator of whether the Duckworth-Lewis method was applied (1 for applied, 0 for not applied).
11. winner The winning team of the match.
12. win\_by\_runs The margin of victory in runs (0 for wickets, if not applicable).
13. win\_by\_wickets The margin of victory in wickets (0 for runs, if not applicable).
14. player\_of\_match The player awarded the "Man of the Match" title.
15. venue The stadium or venue where the match was played.
16. umpire1 The name of the first on-field umpire.
17. umpire2 The name of the second on-field umpire.
18. umpire3 The name of the third umpire (TV umpire).

Appendix 2: Ball to Ball data:

1. match\_id A unique identifier for each ODI cricket match.
2. season The season in which the match took place
3. start\_date The date on which the match started.
4. venue The stadium or venue where the match was played.
5. innings The innings number (1st innings or 2nd innings).
6. ball A numeric representation of the ball number bowled in the innings.
7. batting\_team The name of the batting team for the current innings.
8. bowling\_team The name of the bowling team for the current innings.
9. striker The batsman who is currently facing the ball.
10. non\_striker The batsman at the non-striker's end.
11. bowler The bowler who is delivering the ball.
12. runs\_off\_bat The number of runs scored off the bat (excluding extras).
13. extras The total number of extra runs (wides, no-balls, byes, leg-byes, penalty) in the current ball.
14. wides The number of wide deliveries bowled in the current ball.
15. noballs The number of no-ball deliveries bowled in the current ball.
16. byes The number of byes scored in the current ball.
17. legbyes The number of leg-byes scored in the current ball.
18. penalty The number of penalty runs awarded in the current ball.
19. wicket\_type The type of wicket taken in the current ball (e.g., caught, bowled, run out).
20. player\_dismissed The player who was dismissed in the current ball.
21. other\_wicket\_type Additional information about the wicket (if any) in the current ball.
22. other\_player\_dismissed Additional player information related to the wicket (if any) in the current ball.
23. cricsheet\_id A unique identifier for the match from Cricsheet.