CAPSTONE PROJECT:FINAL REPORT

CRICKET_WIN_PREDICTION

PGP-DSBA (PGPDSBA.O.SEP23.C)

Name-Kumar Ankit
email-kr.ankit7896@gmail.com

Problem Statement

BCCI has hired an external analytics consulting firm for data analytics. The major objective of this tie up is to extract actionable insights from the historical match data and make strategic changes to make India win. Primary objective is to create Machine Learning models which correctly predicts a win for the Indian Cricket Team. Once a model is developed then you have to extract actionable insights and recommendation.

Also, below are the details of the next 10 matches, India is going to play. You have to predict the result of the matches and if you are getting prediction as a Loss then suggest some changes and re-run your model again until you are getting Win as a prediction. You cannot use the same strategy in the entire series, because opponent will get to know your strategy and they can come with counter strategy. Hence for all the below 5 matches you have to suggest unique strategies to make India win. The suggestions should be in-line with the variables that have been mentioned in the given data set. Do consider the feasibility of the suggestions very carefully as well.

1. 1 Test match with England in England. All the match are day matches. In England, it will be rainy season at the time to match.
2. 2 T20 match with Australia in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.
3. 2 ODI match with Sri Lanka in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.

1. Introduction of the Business Problem

a) Defining Problem Statement

The goal of this report is to analyze cricket match data and identify patterns or factors that can influence the prediction of match outcomes (Win or Loss). The dataset contains various match-related features, such as team composition, match conditions, player performance, and opponent details. This analysis aims to help in understanding the relationship between these variables and the match result, providing insights that can aid in future match planning and strategy.

b) Need of the Study/Project

In the highly competitive world of cricket, understanding and predicting match outcomes can offer a significant advantage. Accurate predictions can aid in strategic planning, optimise team selection, and improve game strategies. This project addresses the need for data-driven decision-making in cricket, offering valuable predictions that can impact team performance and resource allocation.

c) Understanding Business/Social Opportunity

This study presents an opportunity for cricket teams and sponsors to leverage predictive analytics for competitive advantage. By using data-driven insights, teams can enhance their performance, optimise player selection, and ultimately increase their chances of winning. Additionally, accurate predictions can attract more viewers and sponsors, boosting the sport's commercial potential.

2. Data Report

a) Understanding How Data Was Collected

The data used in this study was collected from cricket match records over several seasons. Data collection included various attributes such as player statistics, team composition, and match conditions. The data was gathered at the end of each match and includes historical performance metrics.

b) Visual Inspection of Data

The dataset consists of 2,930 cricket matches, with 23 variables covering match details such as:

- Result: Outcome of the match (Win/Loss)
- Avg_team_Age: Average age of the team
- Match_light_type: Type of lighting used (Day/Night)
- Bowlers_in_team: Number of bowlers in the team
- Wicket_keeper_in_team: Presence of wicketkeeper
- All_rounder_in_team: Number of all-rounders
- First_selection: Whether the team was selected first
- Opponent: Opposing team
- Season: Season of the match
- Audience_number: Number of spectators
- Offshore: Whether the match was played offshore
- Max_run_scored_1over: Maximum runs scored in one over
- Max_wicket_taken_1over: Maximum wickets taken in one over
- Extra_bowls_bowled: Extra bowls bowled
- Min_run_given_1over: Minimum runs given in one over
- Min_run_scored_1over: Minimum runs scored in one over
- Max_run_given_1over: Maximum runs given in one over
- extra_bowls_opponent: Extra bowls bowled by the opponent
- player_highest_run: Highest run scored by a player
- Players_scored_zero: Number of players scoring zero runs
- player_highest_wicket: Highest wickets taken by a player
- Match_format: Format of the match (T20, ODI, Test)

c) Understanding of Attributes

Attributes were reviewed for relevance to the problem. Column names were standardized for clarity, and categorical variables were encoded for analysis.

The data also includes missing values across several columns:

- **Avg_team_Age**: 97 missing values
- **Match_light_type**: 52 missing values
- **Match_format**: 70 missing values
- **Bowlers_in_team**: 82 missing values
- **Audience_number**: 81 missing values

- **Max_run_scored_1over**: 28 missing values

The descriptive statistics of numeric columns revealed several interesting insights:

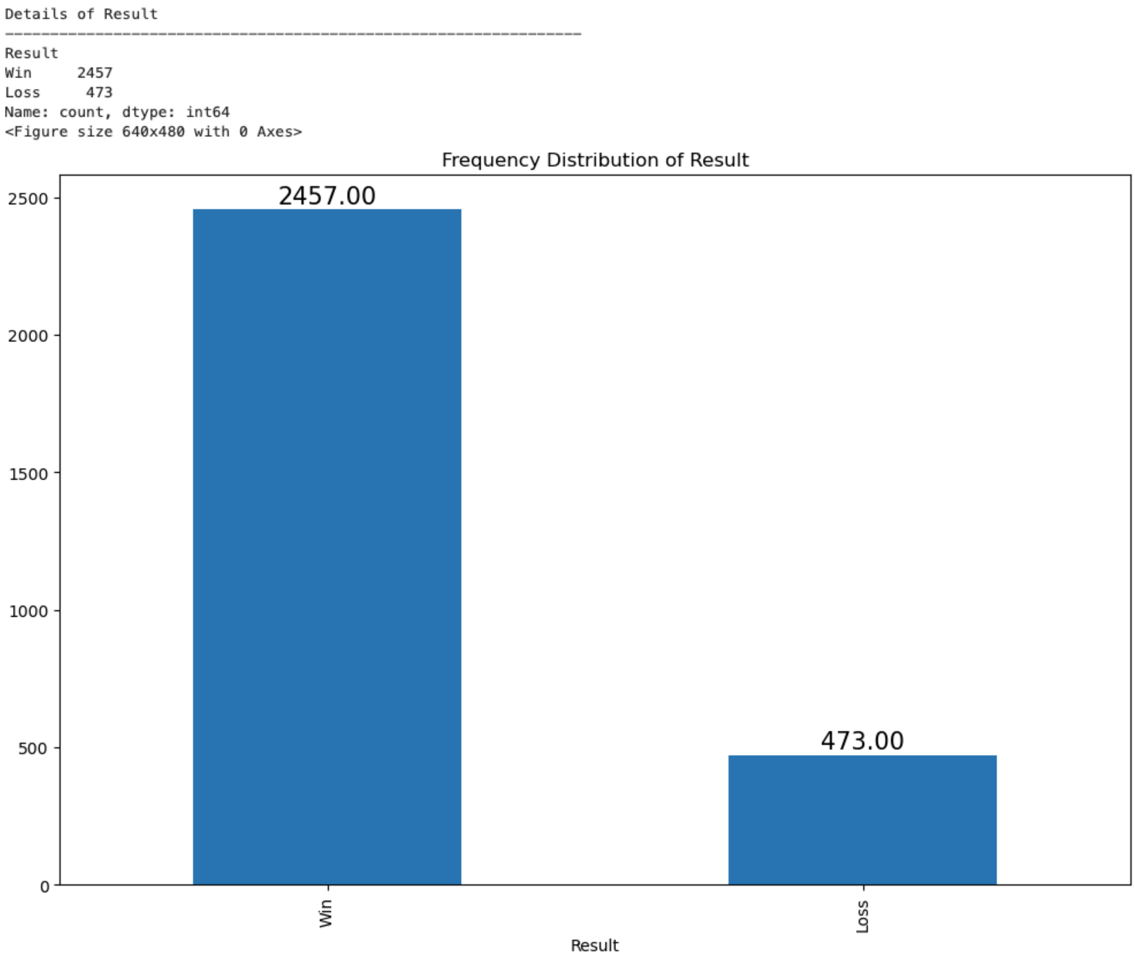
- **Average Team Age**: The average team age is around 29 years, with most teams being between 30 and 70 years old. This suggests that teams have a wide range of player ages, potentially impacting performance.
- **Audience Size**: Matches drew an average of around 46,268 spectators, with some matches drawing over a million spectators.
- **Max Runs Scored in 1 Over**: On average, teams scored 15.2 runs in the most productive over, with a maximum of 25 runs in a single over. This indicates significant variation in scoring ability across matches.

3. Exploratory Data Analysis (EDA)

a) Univariate Analysis

1.Result

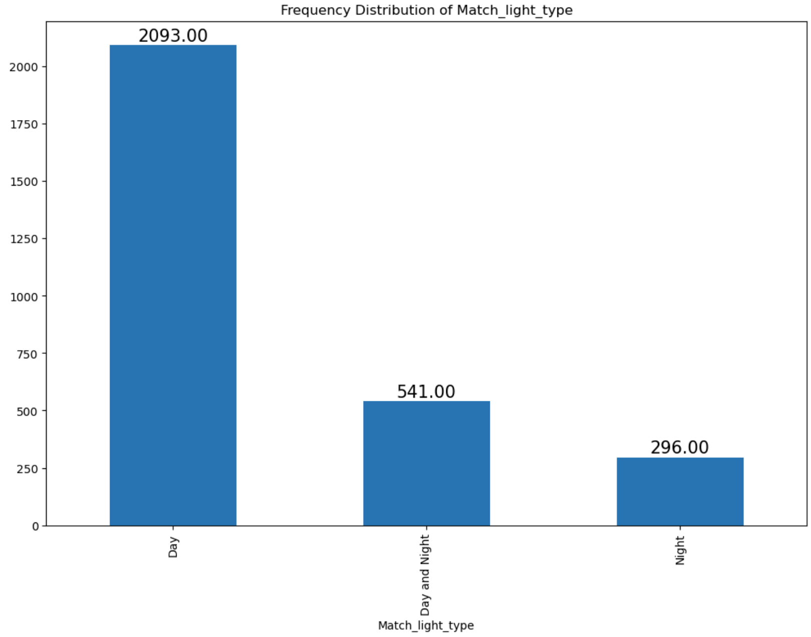
- **Distribution**: The dataset has 2,457 wins and 473 losses, indicating a strong imbalance towards wins.
- **Implication**: The imbalance in the 'Result' variable suggests that the majority of matches result in a win, which could affect predictive modeling efforts.



```
Details of Result
----------------------------------------------------------------
Result
Win     2457
Loss     473
Name: count, dtype: int64
<Figure size 640x480 with 0 Axes>
```

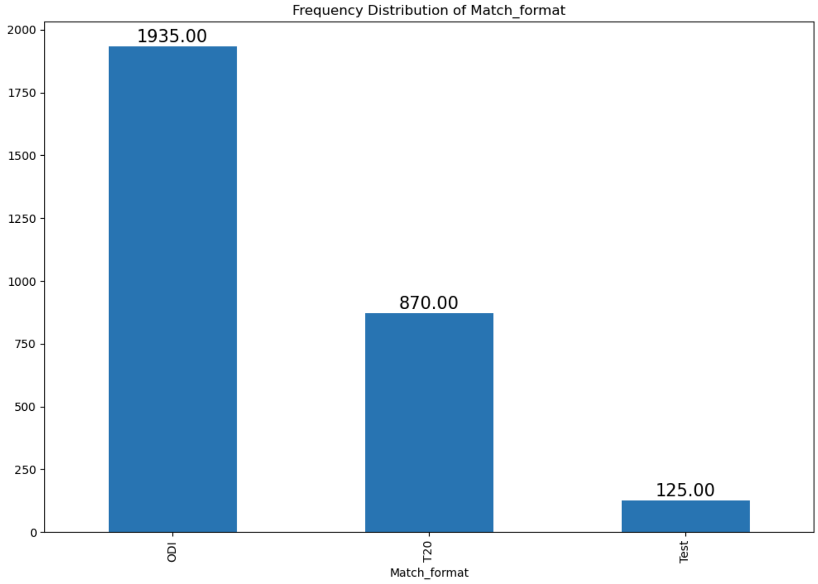
2.Match_light_type

- **Distribution**: Most matches are held during the day (2,093), with fewer matches held during "Day and Night" (541) and "Night" (296).
- **Implication**: This variable might influence match outcomes, especially considering factors like visibility and player performance under different lighting conditions.



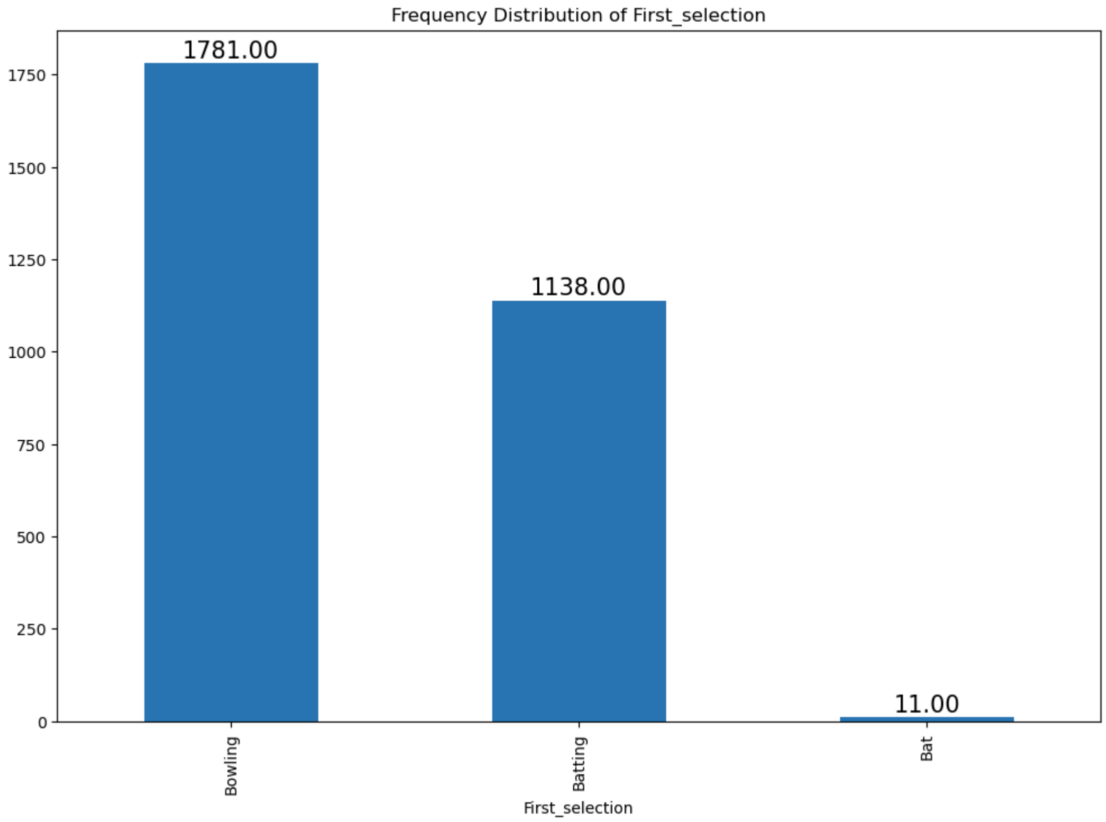
3.Match_format

- **Distribution**: The majority of matches are One Day Internationals (ODI) (1,935), followed by T20s (870), and Tests (125).
- **Implication**: Different formats may require different strategies and team compositions, which could influence match outcomes.



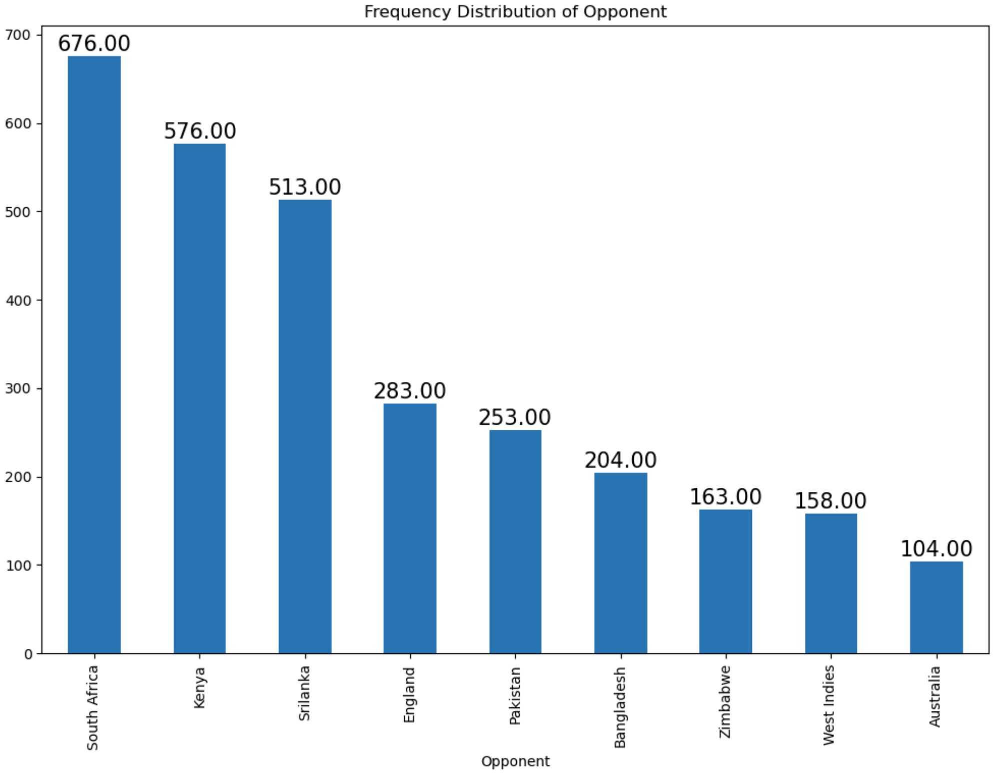
4.First_selection

- **Distribution**: Most teams choose to bowl first (1,781) rather than bat (1,138). A small anomaly exists with 11 matches marked as "Bat".
- **Implication**: The decision to bat or bowl first could be crucial, potentially reflecting the team's confidence or strategy.



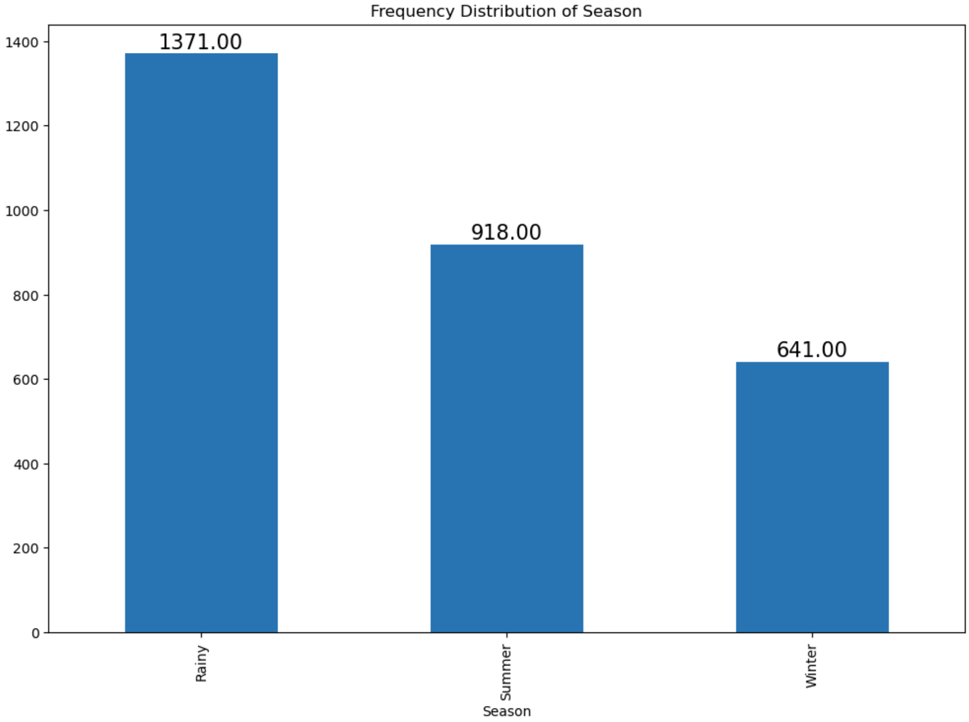
5.Opponent

- **Distribution**: The dataset is dominated by matches against South Africa (676), Kenya (576), and Sri Lanka (513), with fewer matches against other teams like Australia (104).
- **Implication**: The opponent could play a significant role in determining match outcomes, with stronger teams potentially being more challenging to beat.



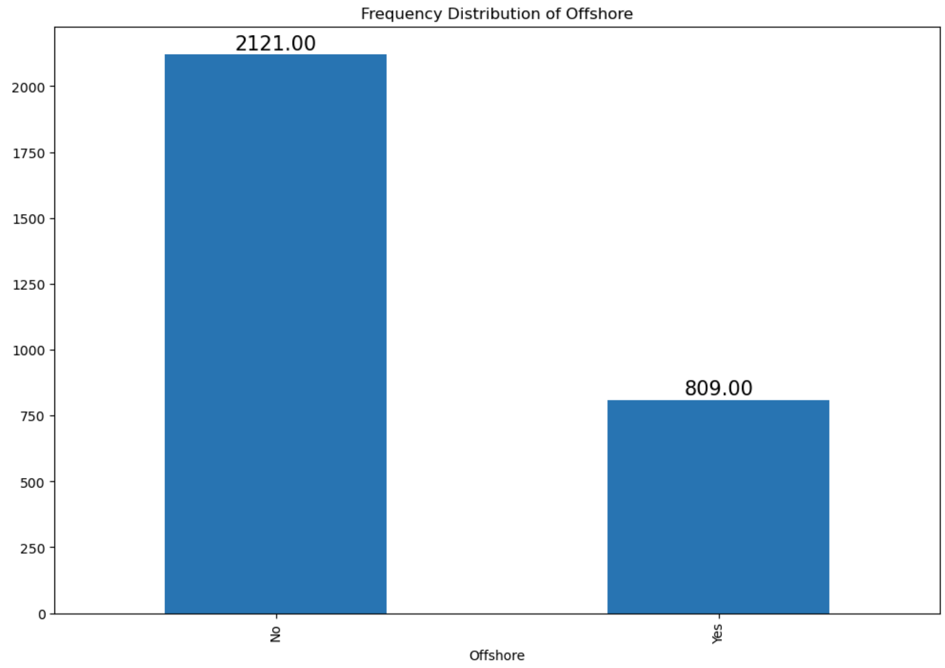
6.Season

- **Distribution**: Matches are fairly evenly distributed across Rainy (1,371), Summer (918), and Winter (641) seasons.
- **Implication**: Weather conditions could affect gameplay, particularly in cricket where rain can lead to match interruptions or affect pitch conditions.



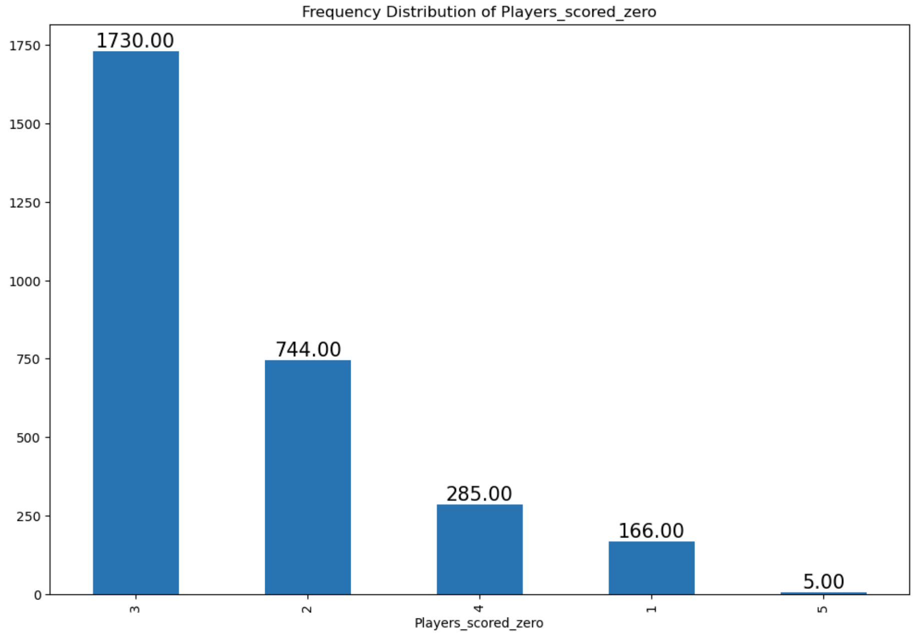
7.Offshore

- **Distribution**: A majority of matches are played onshore (2,121), with 809 matches played offshore.
- **Implication**: Playing offshore could introduce variables such as unfamiliar pitch conditions, which might affect the outcome.



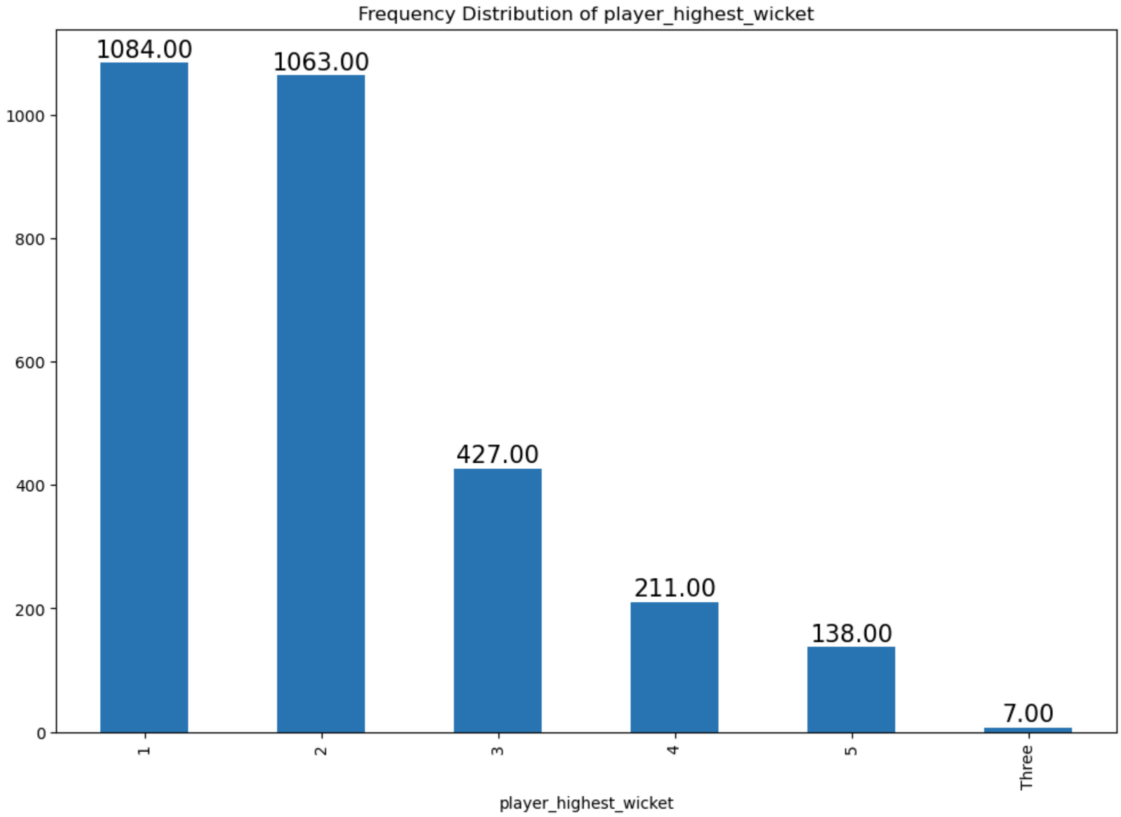
8.Players_scored_zero

- **Distribution**: The number of players scoring zero is most commonly three (1,730 matches), with a smaller number of matches having two (744) or four (285) players scoring zero.
- **Implication**: The number of players scoring zero could reflect the strength or weakness of the batting lineup.



9.player_highest_wicket

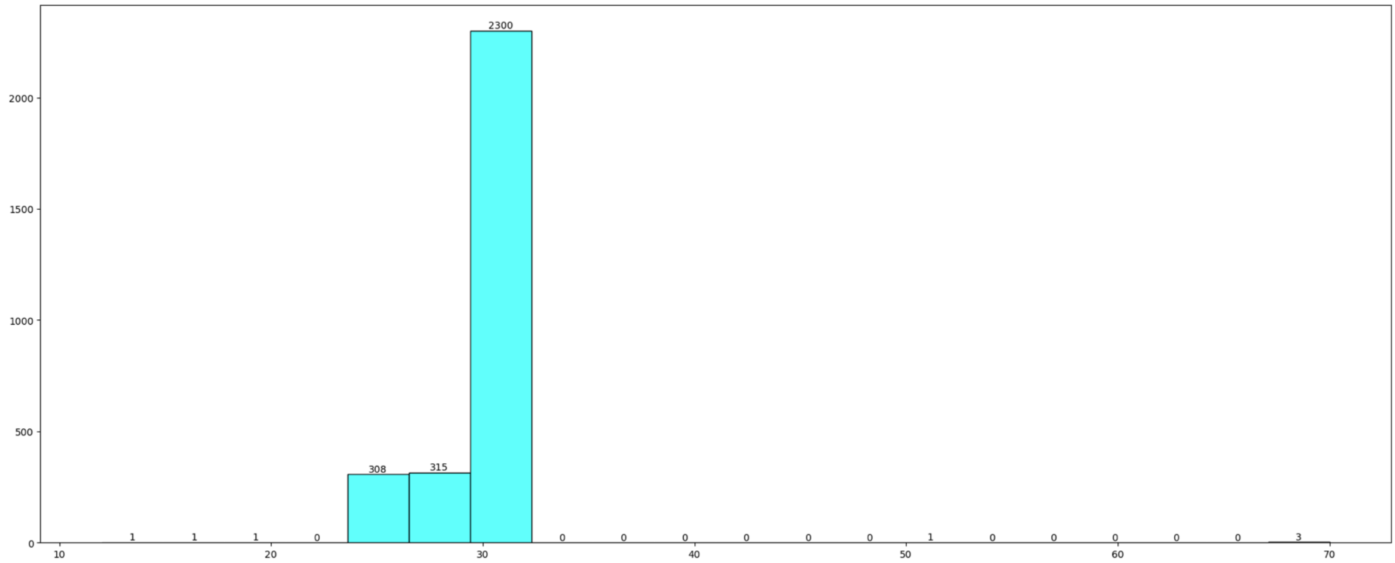
- **Distribution**: Most matches see the highest wicket-taker claiming one (1,084) or two (1,063) wickets, with a decreasing number of matches seeing higher wicket counts.
- **Implication**: The ability to take wickets could be a strong indicator of match success, particularly in limiting the opponent's run rate.

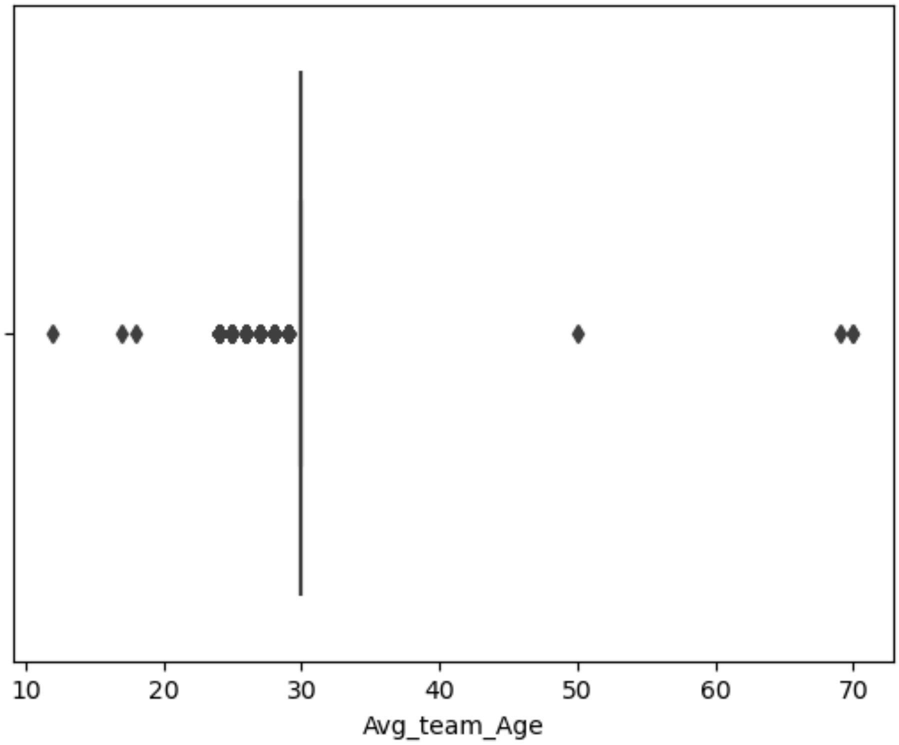


Numerical Variables

1. Avg_team_Age

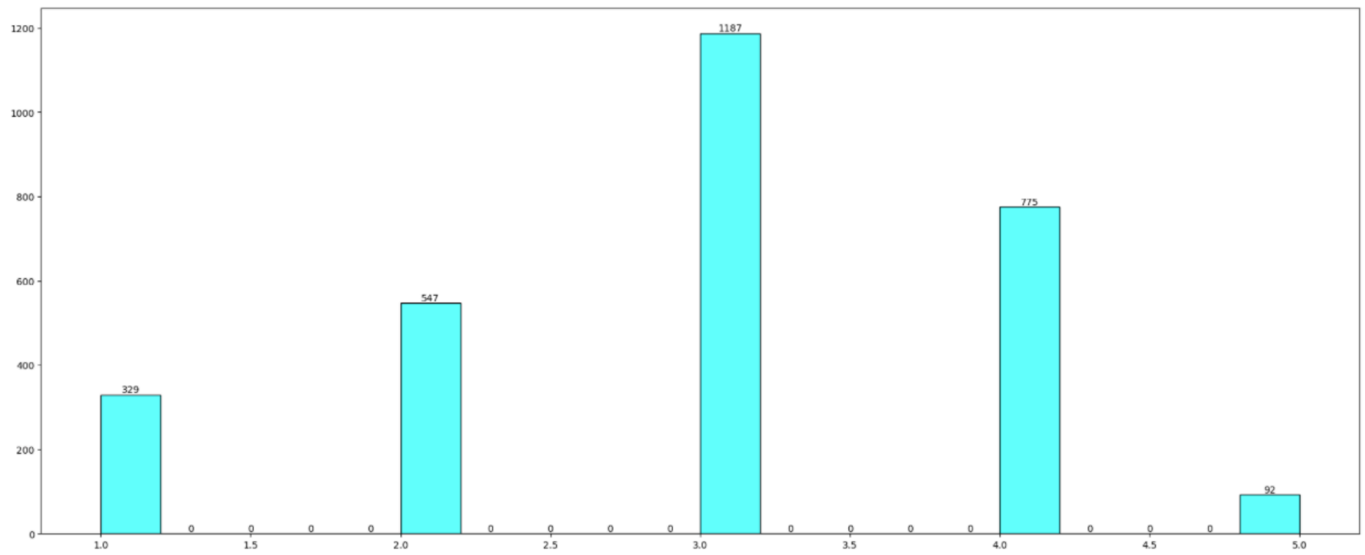
- **Distribution**: The average team age ranges from 12 to 70 years, with a mean of approximately 29.27 years.
- **Implication**: Age could influence player performance, with potential implications for stamina, experience, and injury likelihood.

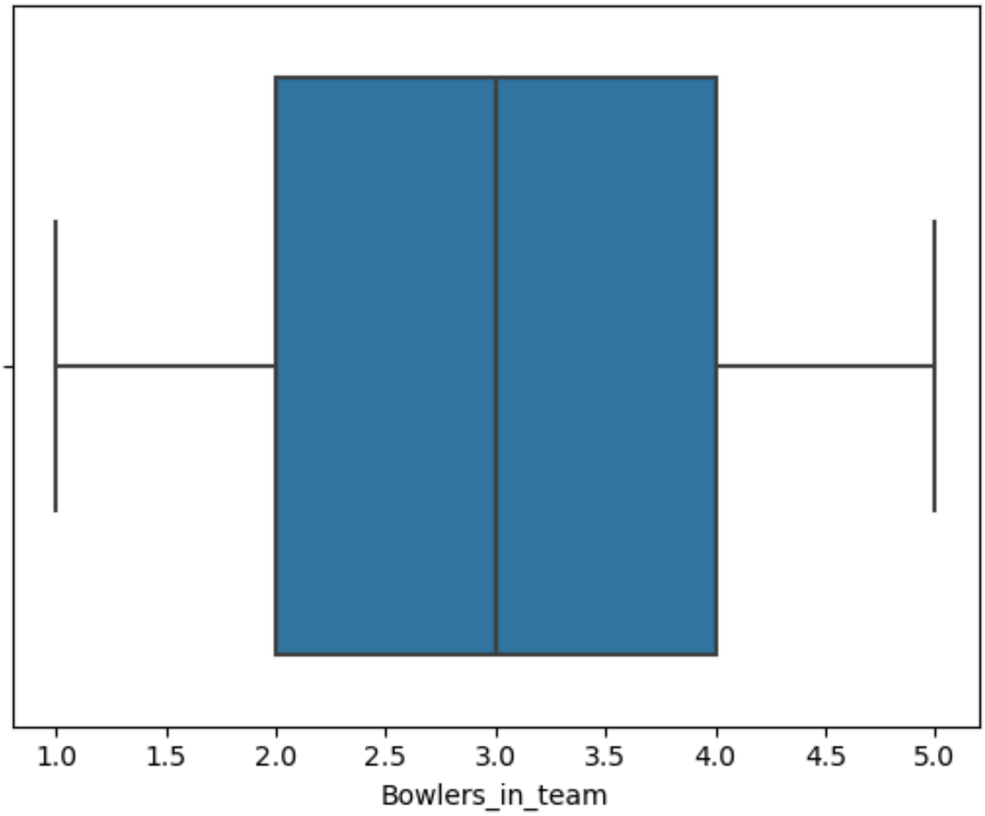




2. Bowlers_in_team

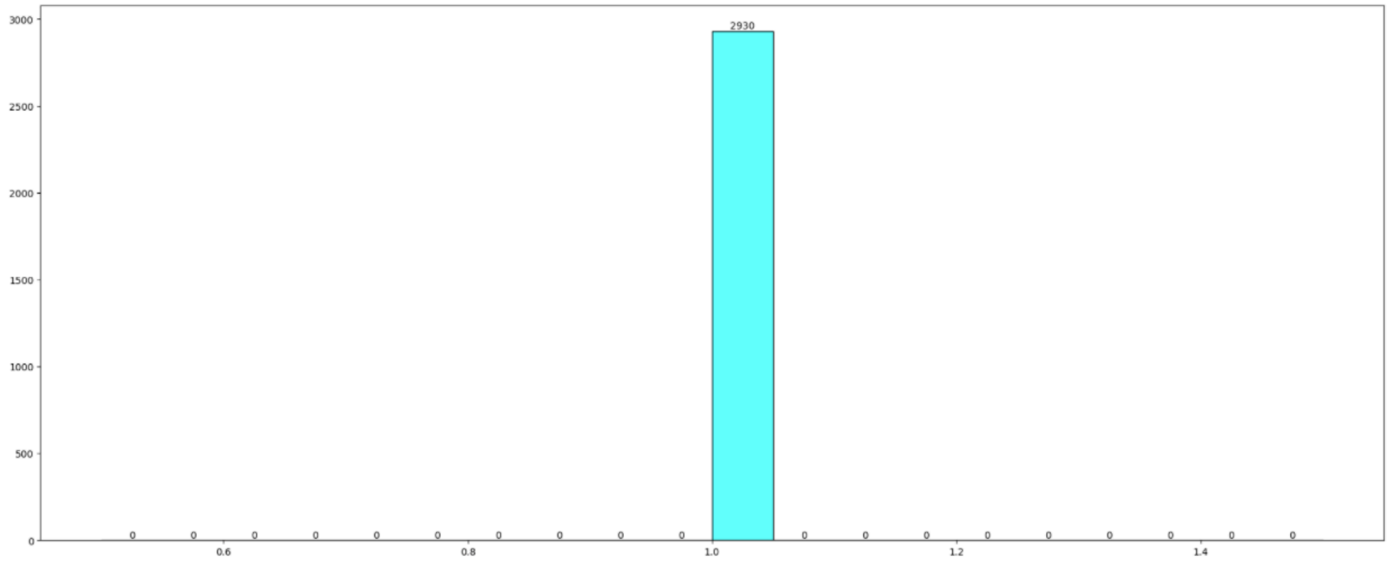
- **Distribution**: Teams generally include around three bowlers, with the number ranging from 1 to 5.
- **Implication**: The number of bowlers might reflect team strategy, particularly in how they balance batting and bowling strength.

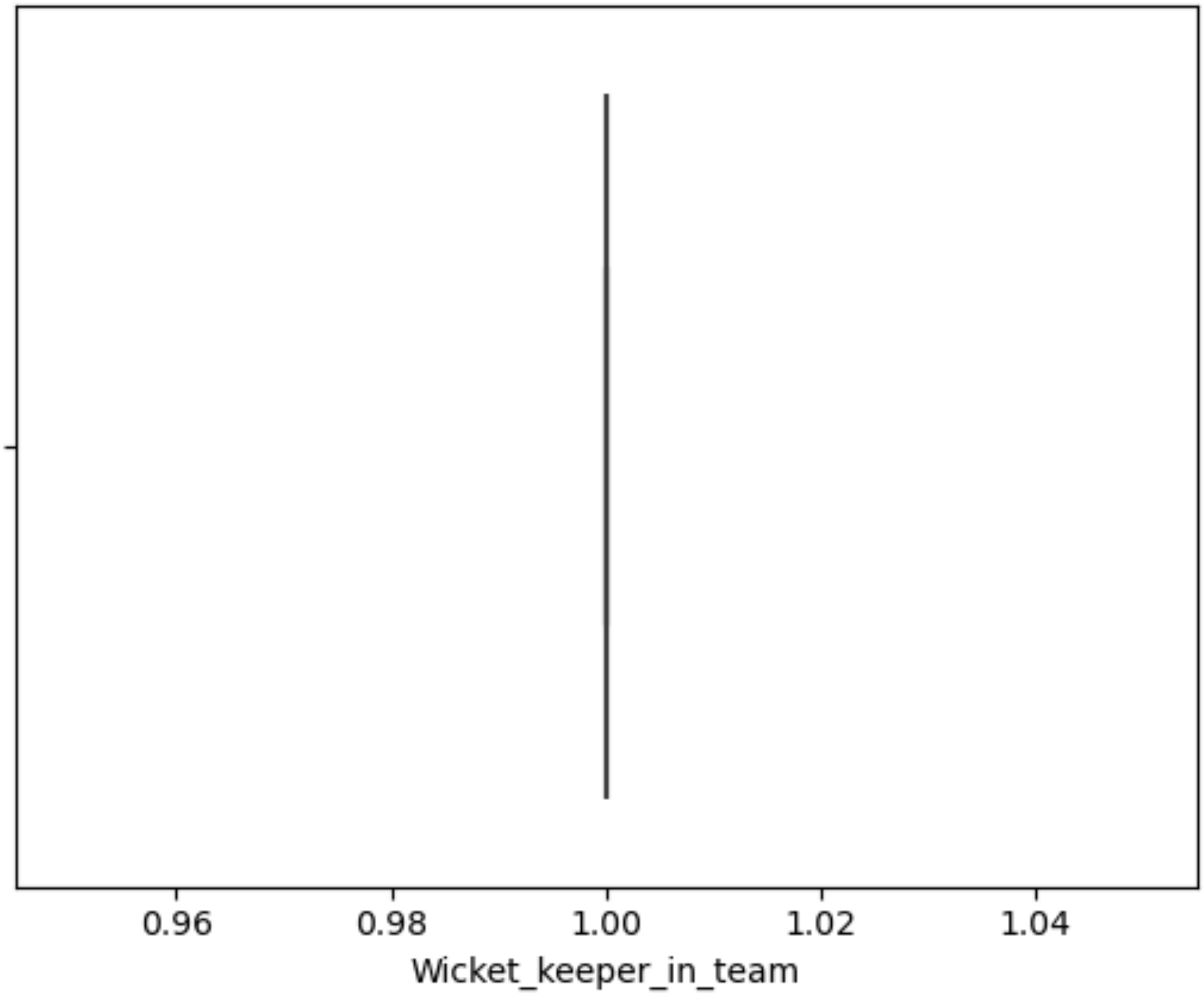




3. Wicket_keeper_in_team

- **Distribution**: All teams have exactly one wicketkeeper.
- **Implication**: As a constant across all teams, this variable may not provide much insight for predictive modeling but is essential for understanding team composition.

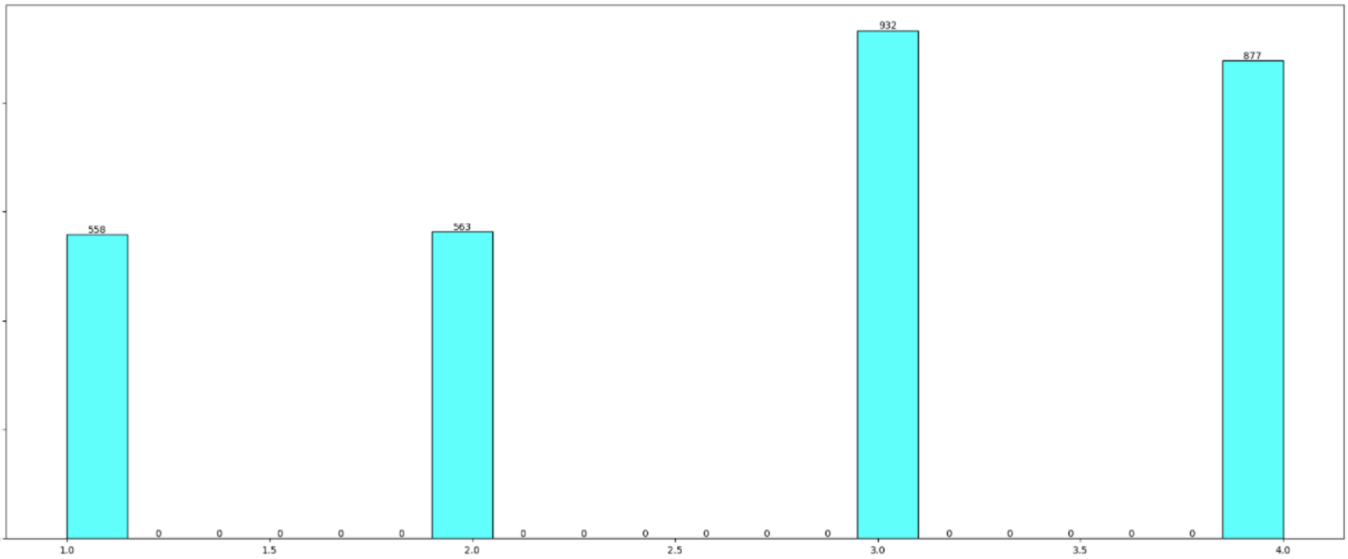


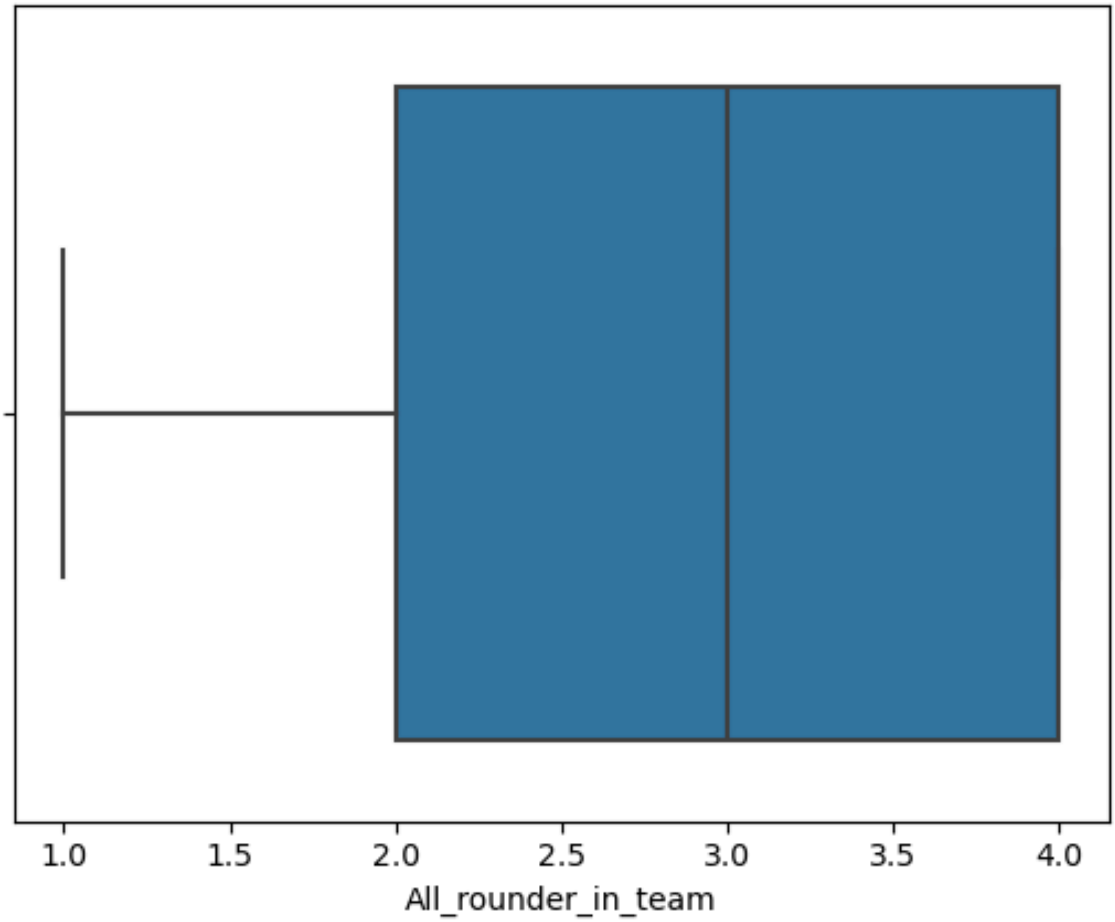


4. **All_rounder_in_team**

   ○ **Distribution**: Teams typically include about three all-rounders, with a range of 1 to 4.

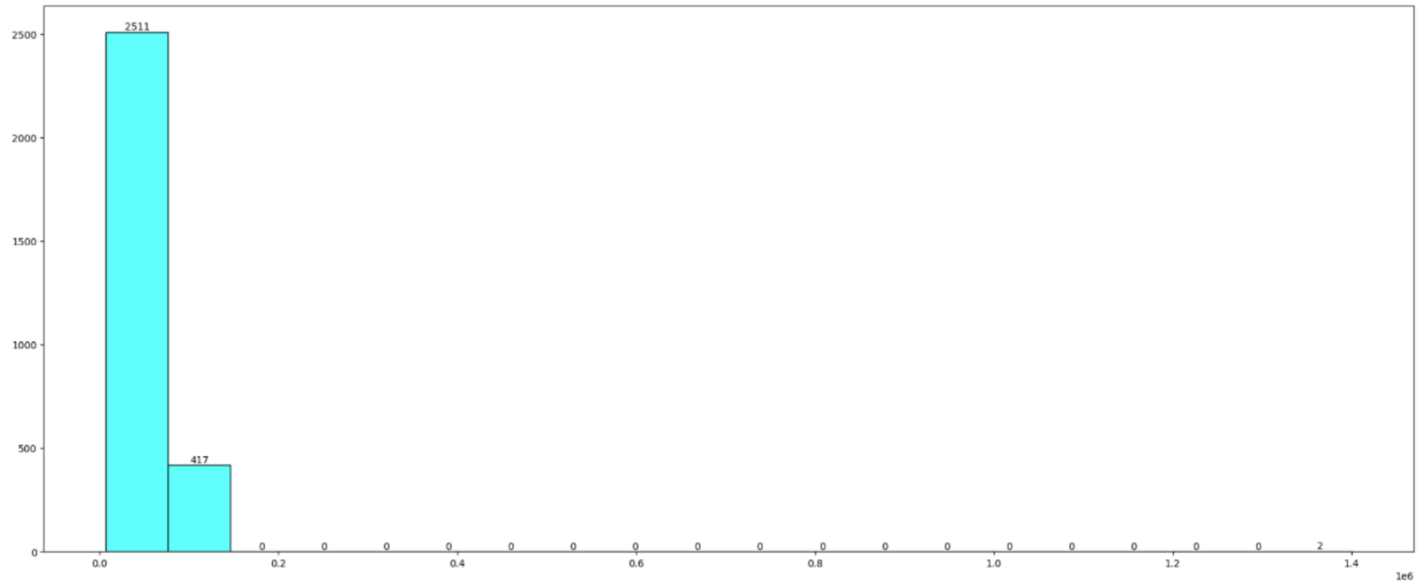
   ○ **Implication**: All-rounders provide flexibility, potentially contributing to both batting and bowling, which could be advantageous.

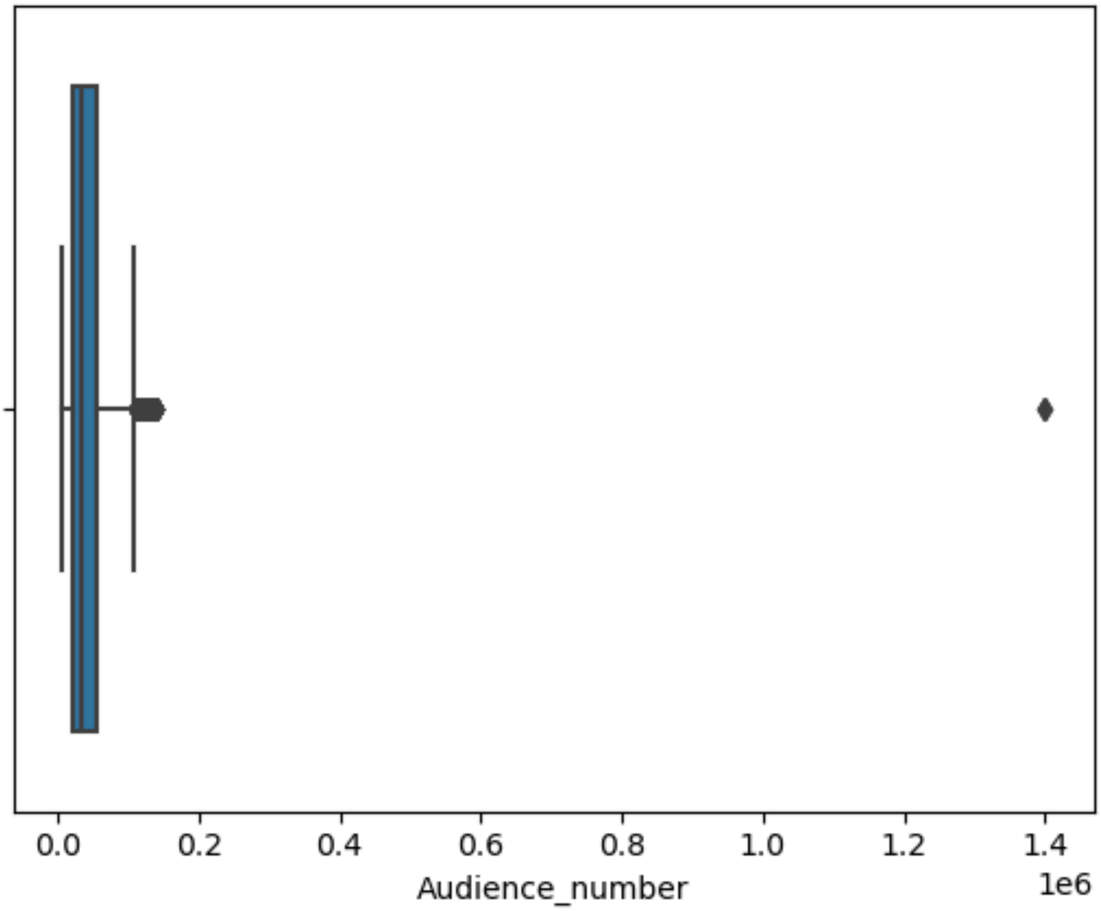




5. **Audience_number**

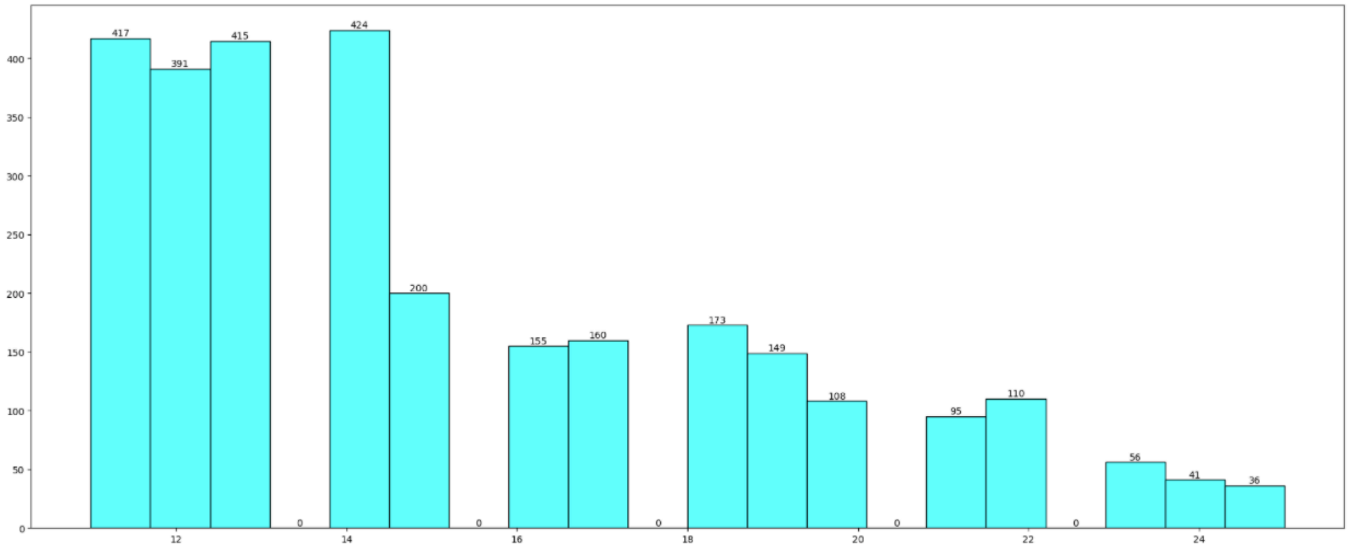
- **Distribution**: Audience numbers vary significantly, with a mean of approximately 45,938.46, but the range is wide (7,063 to 1,399,930).
- **Implication**: Large audiences might correlate with high-profile matches, which could introduce additional pressure or motivation for players.

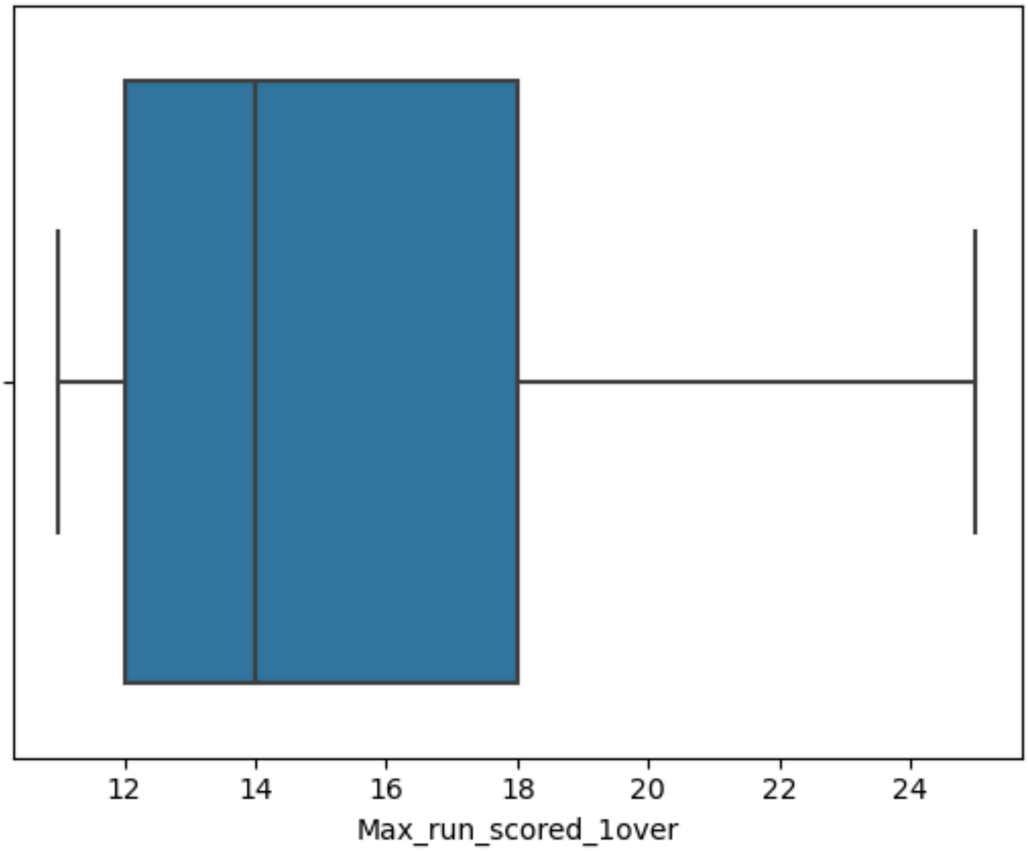




6. Max_run_scored_1over

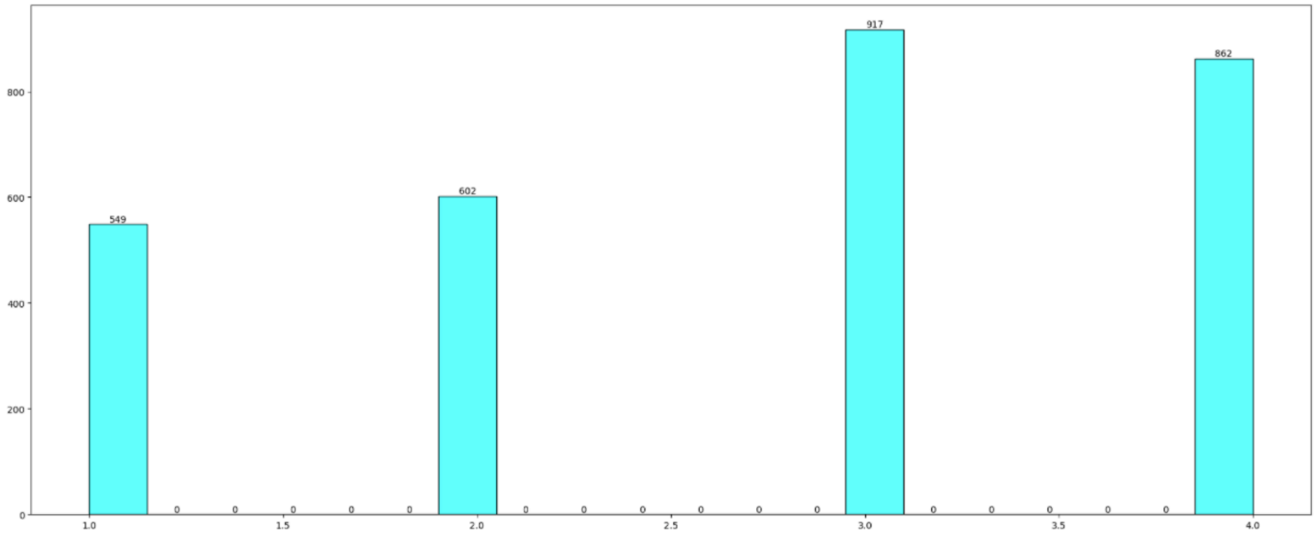
- **Distribution**: The maximum runs scored in an over range from 11 to 25, with a mean of 15.19.
- **Implication**: High-scoring overs could indicate aggressive batting and significantly impact the final score.

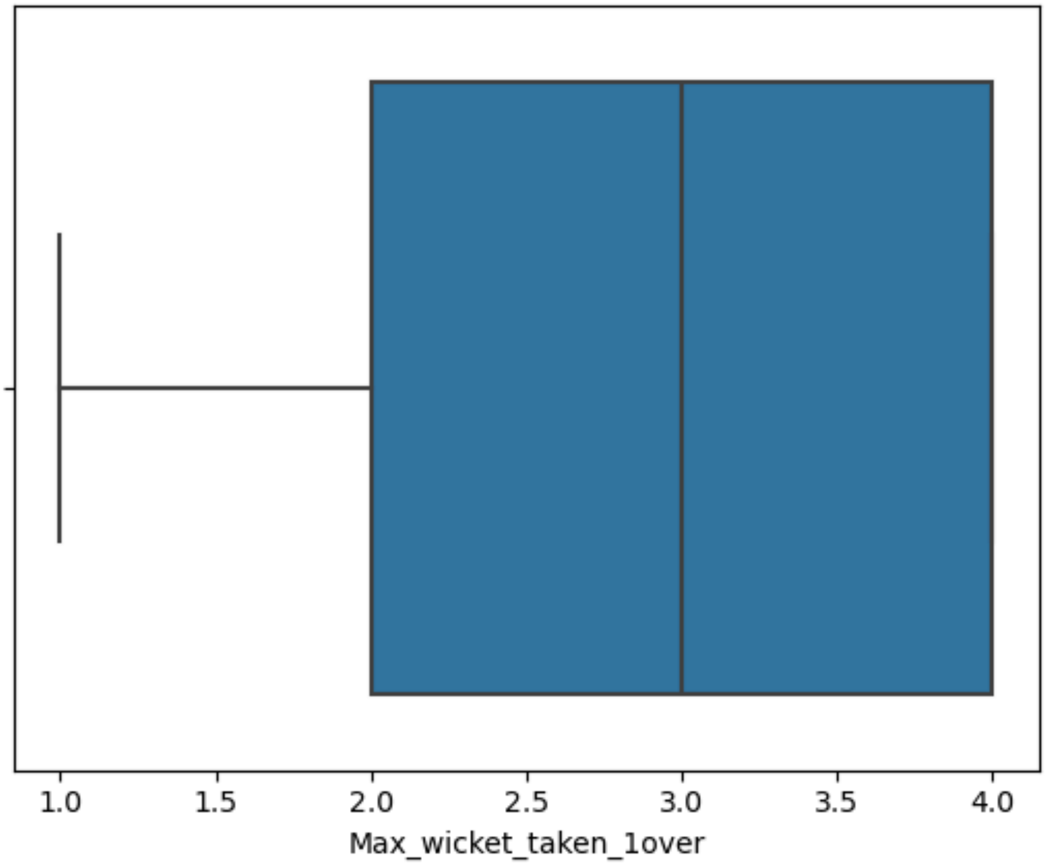




7. Max_wicket_taken_1over

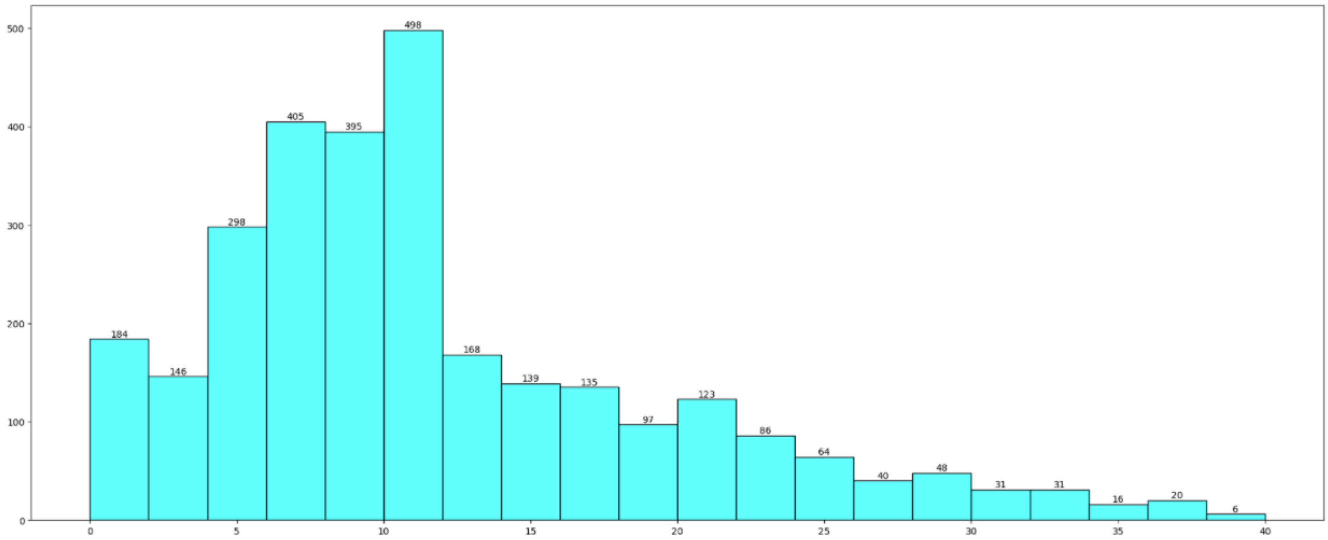
- **Distribution**: The maximum wickets taken in an over range from 1 to 4, with a mean of 2.71.
- **Implication**: This metric could be critical in turning the tide of a match by quickly reducing the opponent's batting strength.

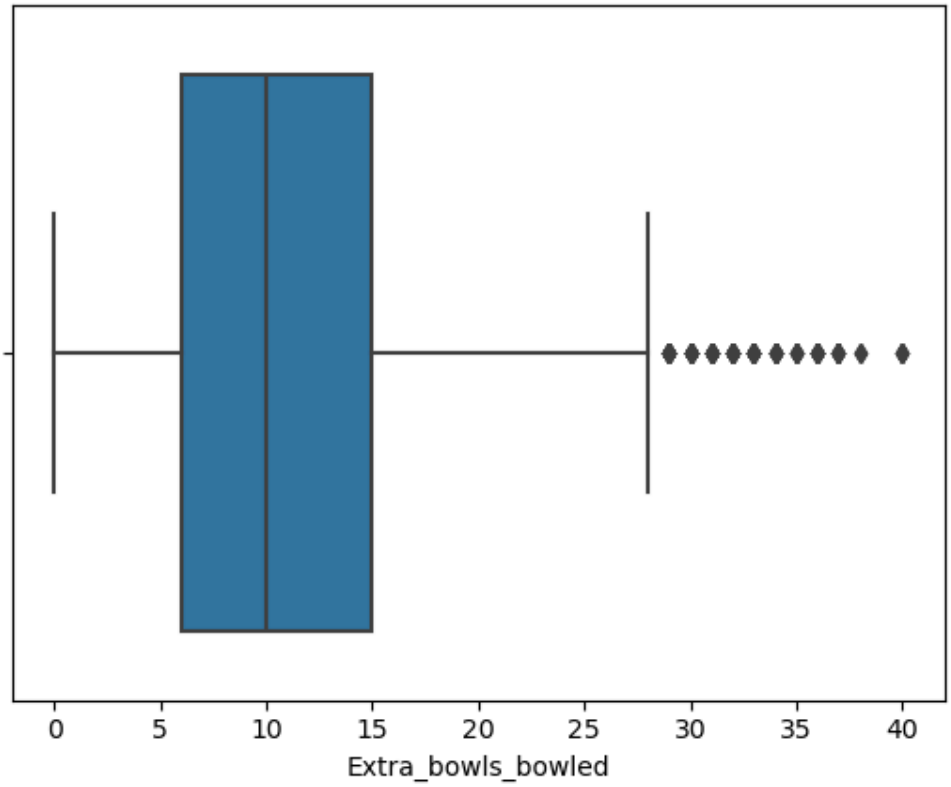




8. Extra_bowls_bowled

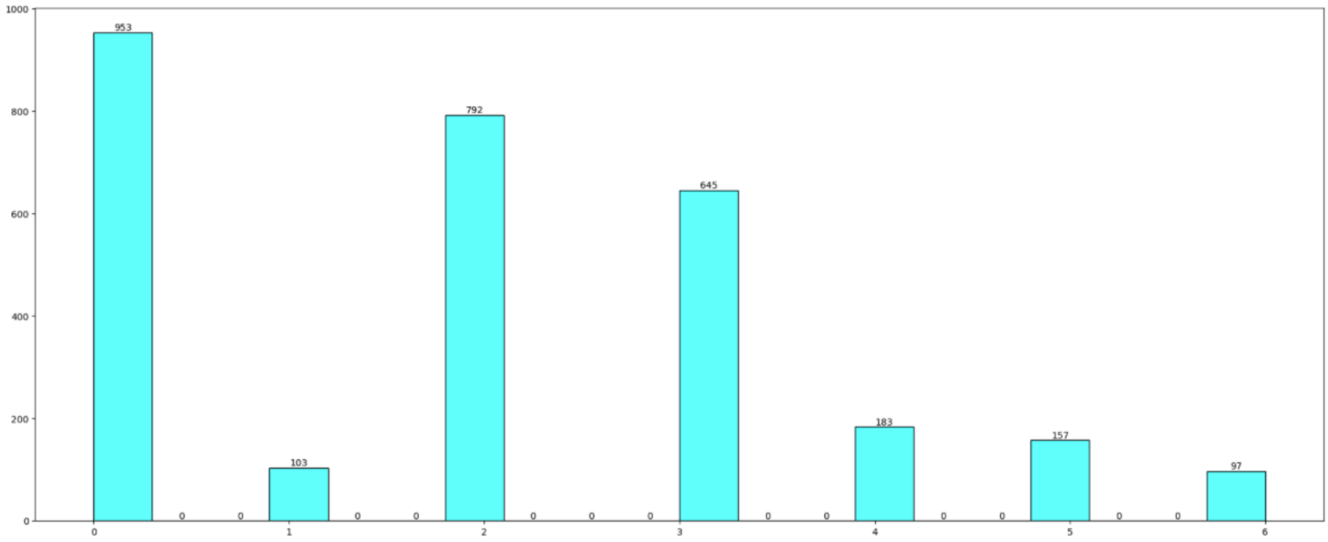
- **Distribution**: Extra bowls bowled range from 0 to 40, with a mean of 11.24.
- **Implication**: Extra bowls could indicate discipline issues within the bowling team, potentially giving the opponent an advantage.

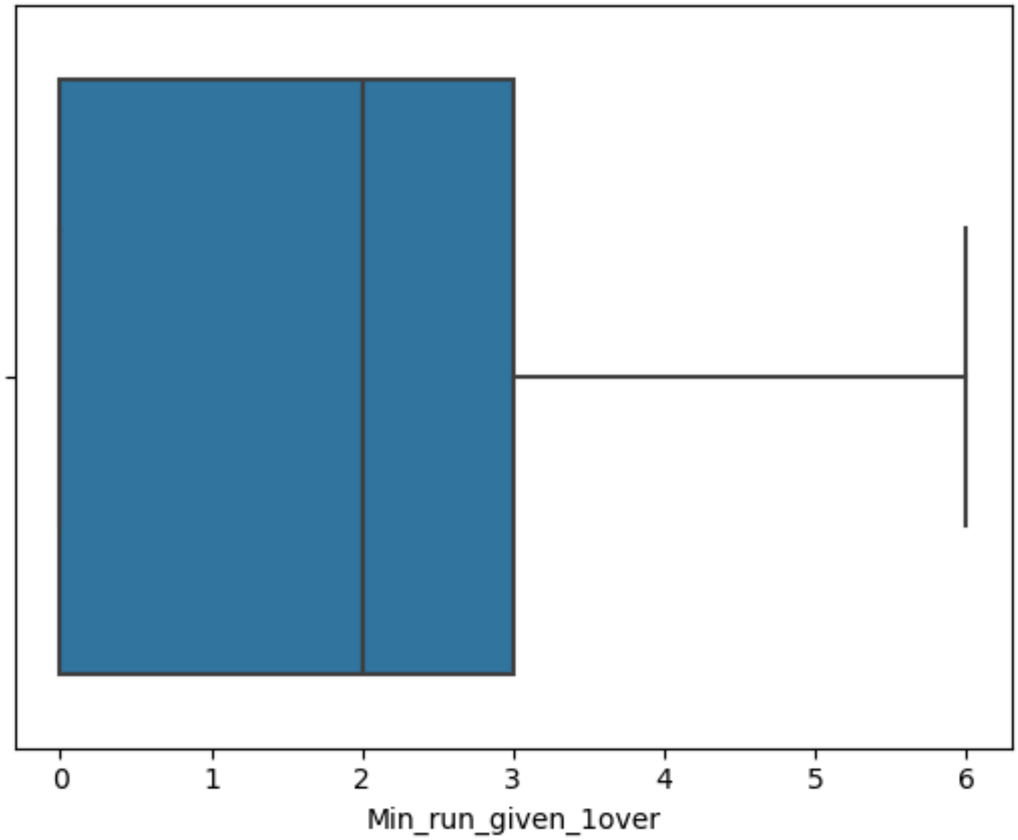




9. Min_run_given_1over

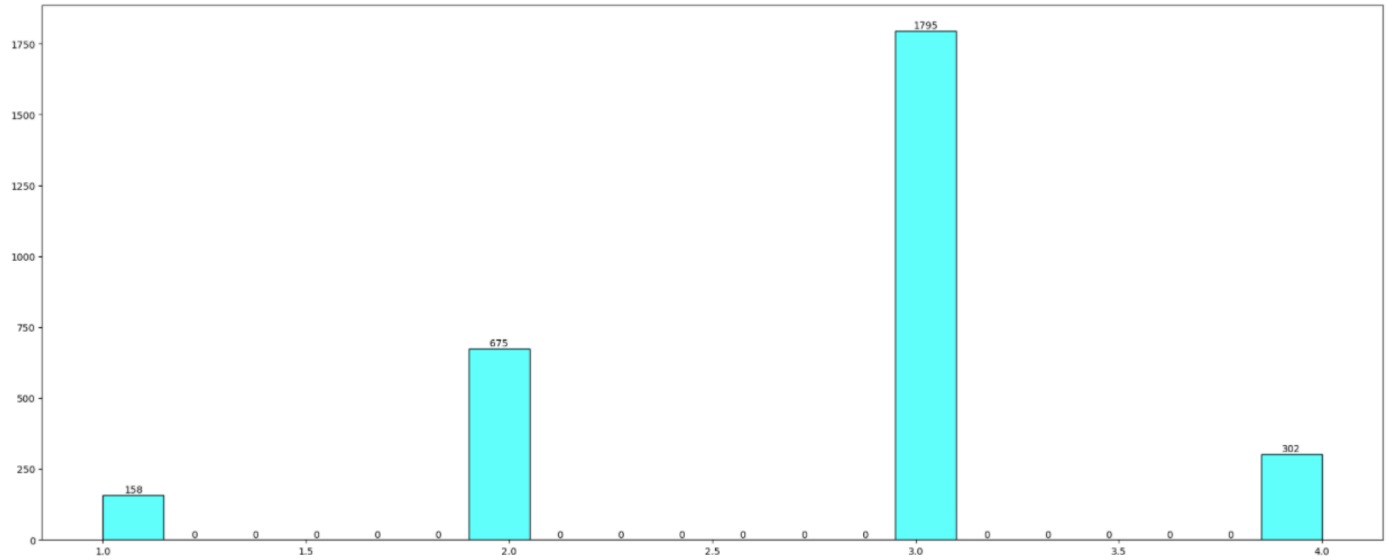
- **Distribution**: The minimum runs given in an over range from 0 to 6, with a mean of 1.95.
- **Implication**: This metric might indicate how well the team is able to contain the opponent's scoring.

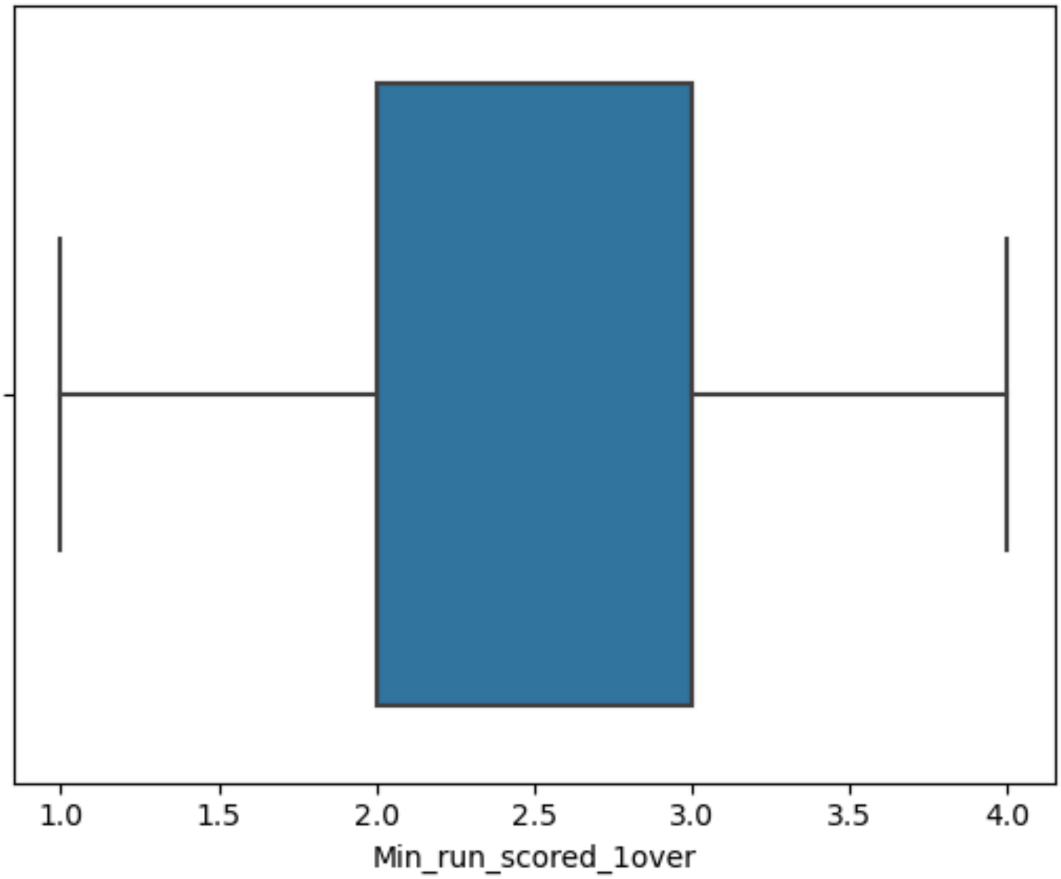




10. Min_run_scored_1over

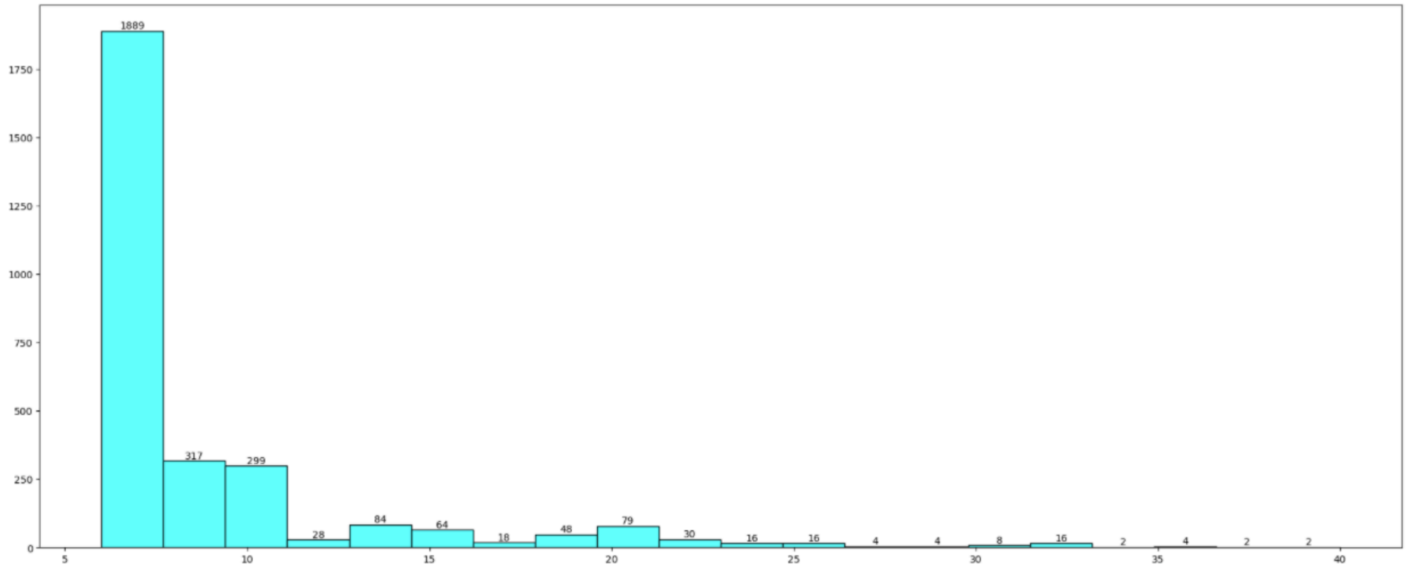
- **Distribution**: The minimum runs scored in an over range from 1 to 4, with a mean of 2.76.
- **Implication**: Consistency in scoring could be crucial for maintaining pressure on the opponent.

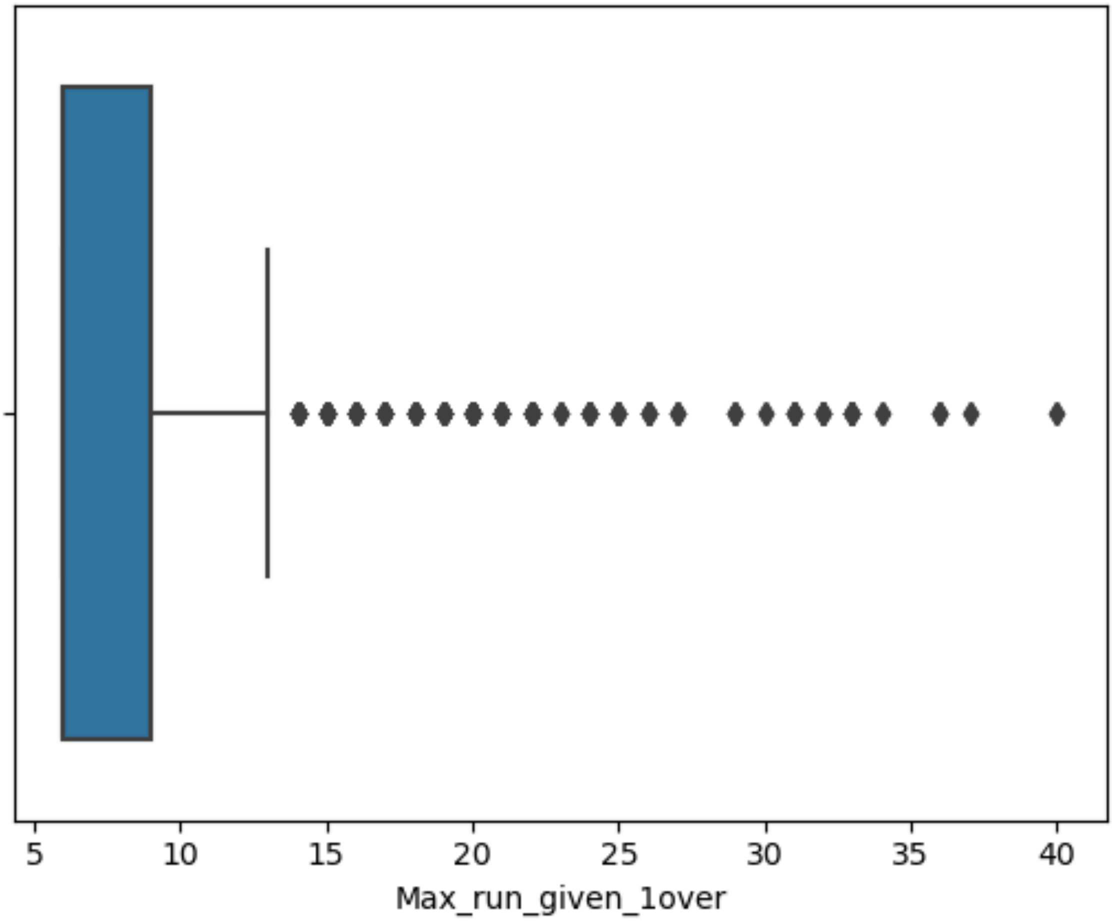




11. Max_run_given_1over

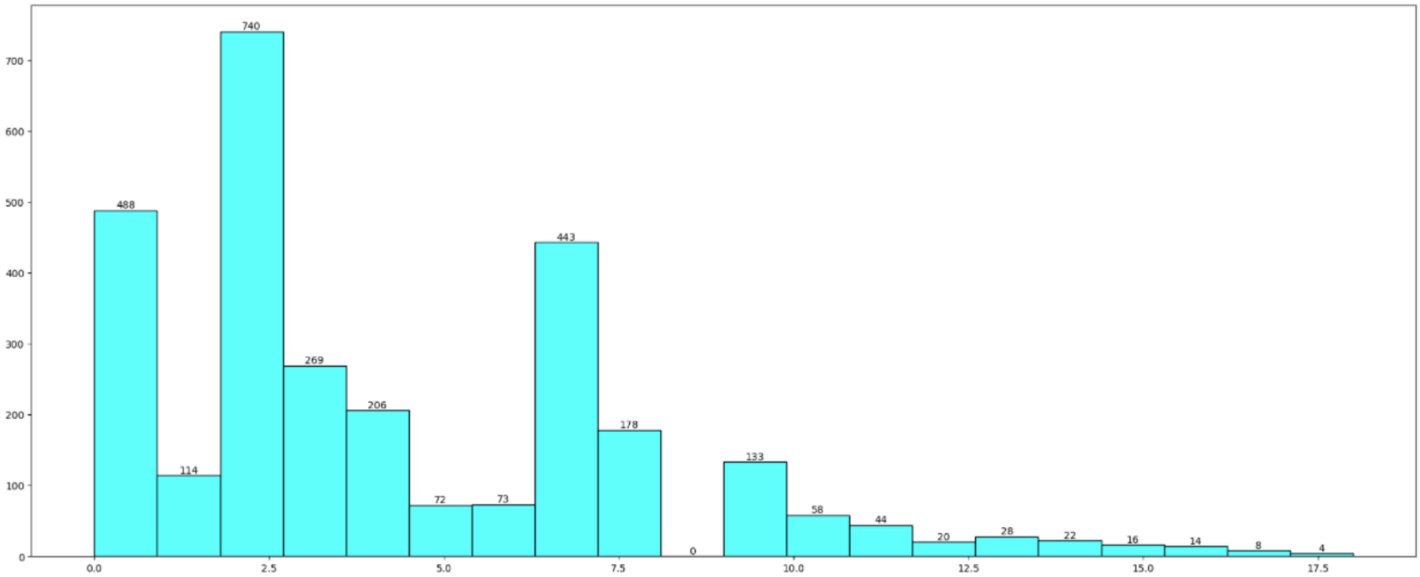
- **Distribution**: The maximum runs given in an over range from 6 to 40, with a mean of 8.64.
- **Implication**: High runs given in an over could indicate weak bowling or a particularly strong opponent batting performance.

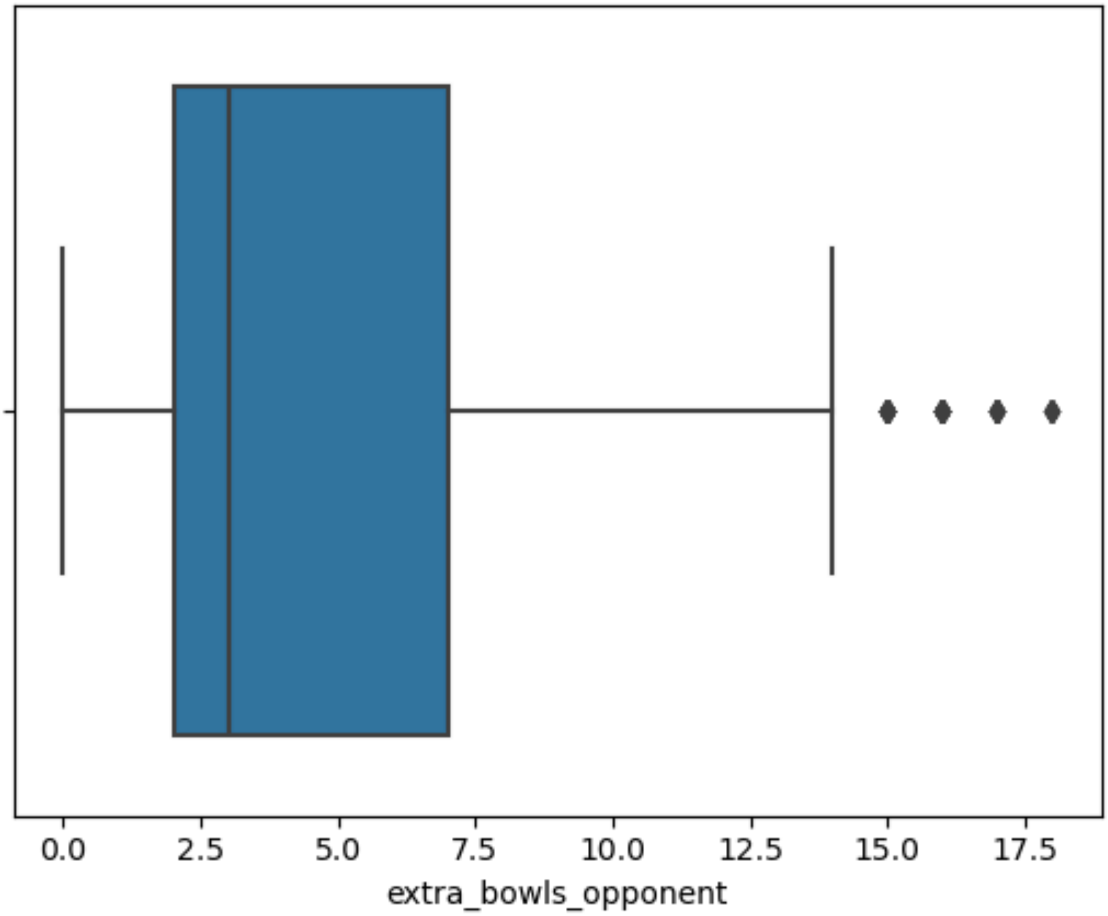




12. extra_bowls_opponent

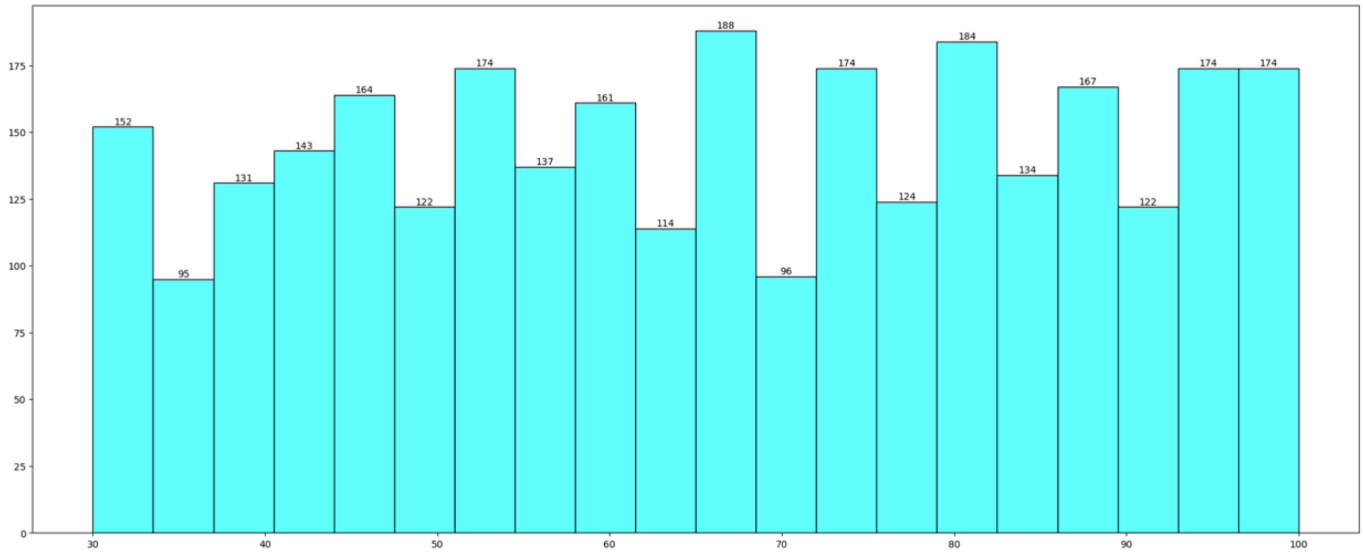
- **Distribution**: The opponent's extra bowls range from 0 to 18, with a mean of 4.23.
- **Implication**: This could provide insight into the opponent's discipline and how it compares to the team's performance.

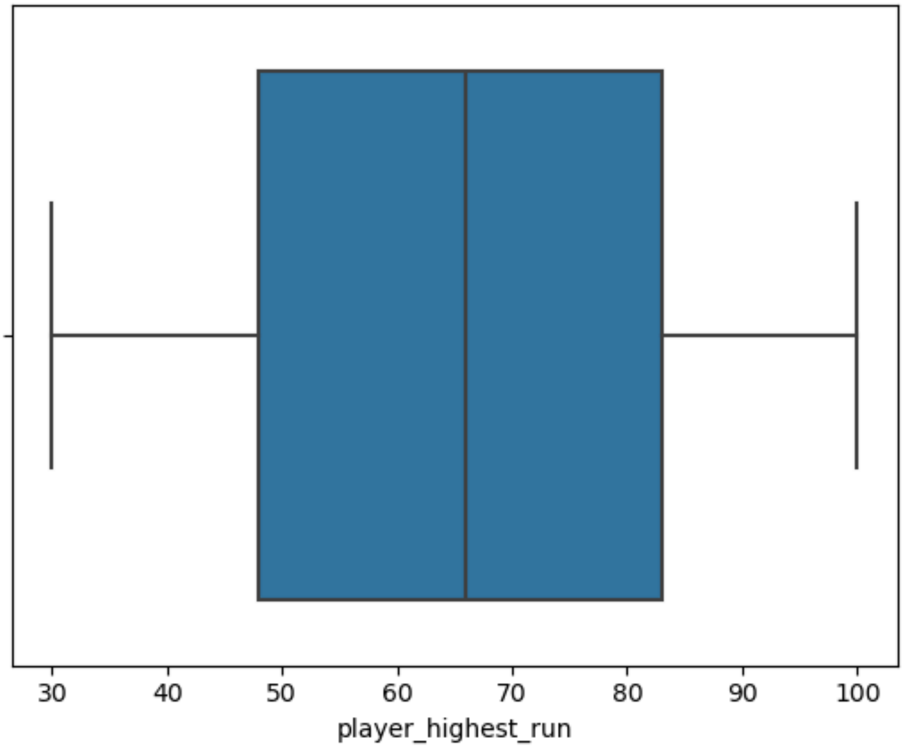




13. player_highest_run

- **Distribution**: The highest runs by a player range from 30 to 100, with a mean of 65.89.
- **Implication**: High individual scores could be key to winning, particularly in matches where one player stands out.





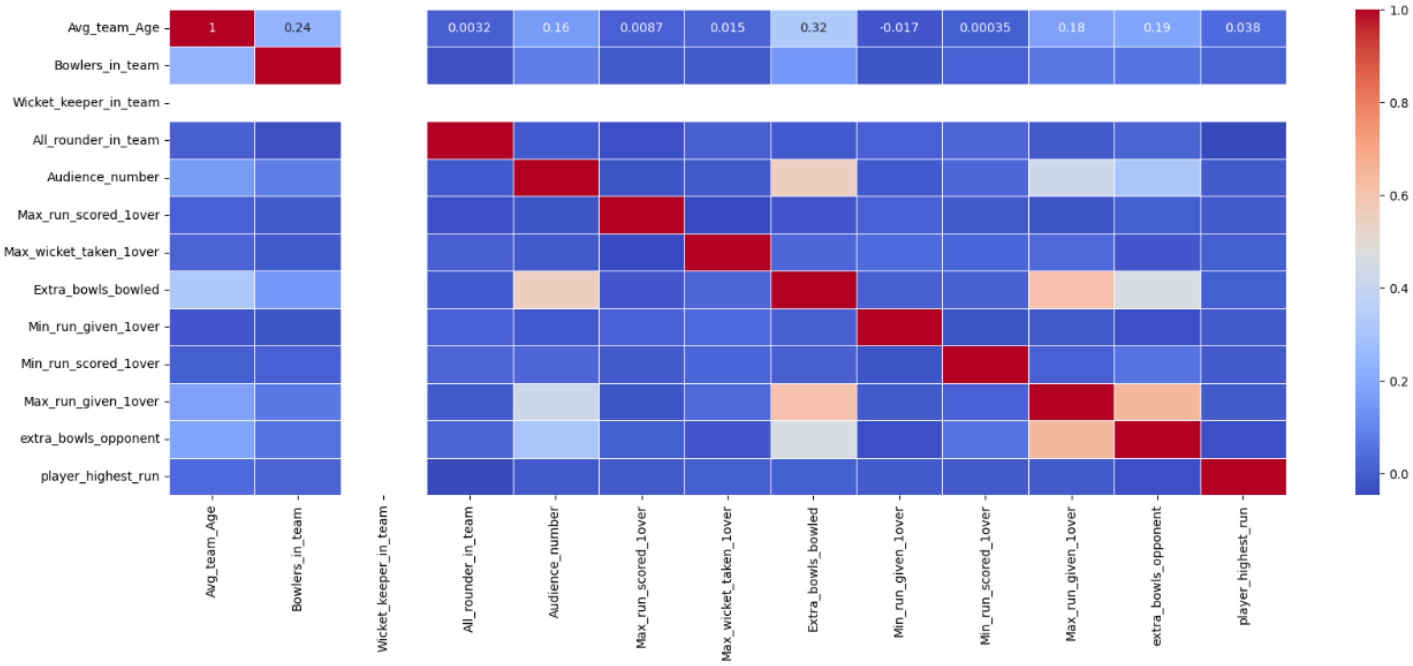
Bivariate Analysis

Correlation Matrix

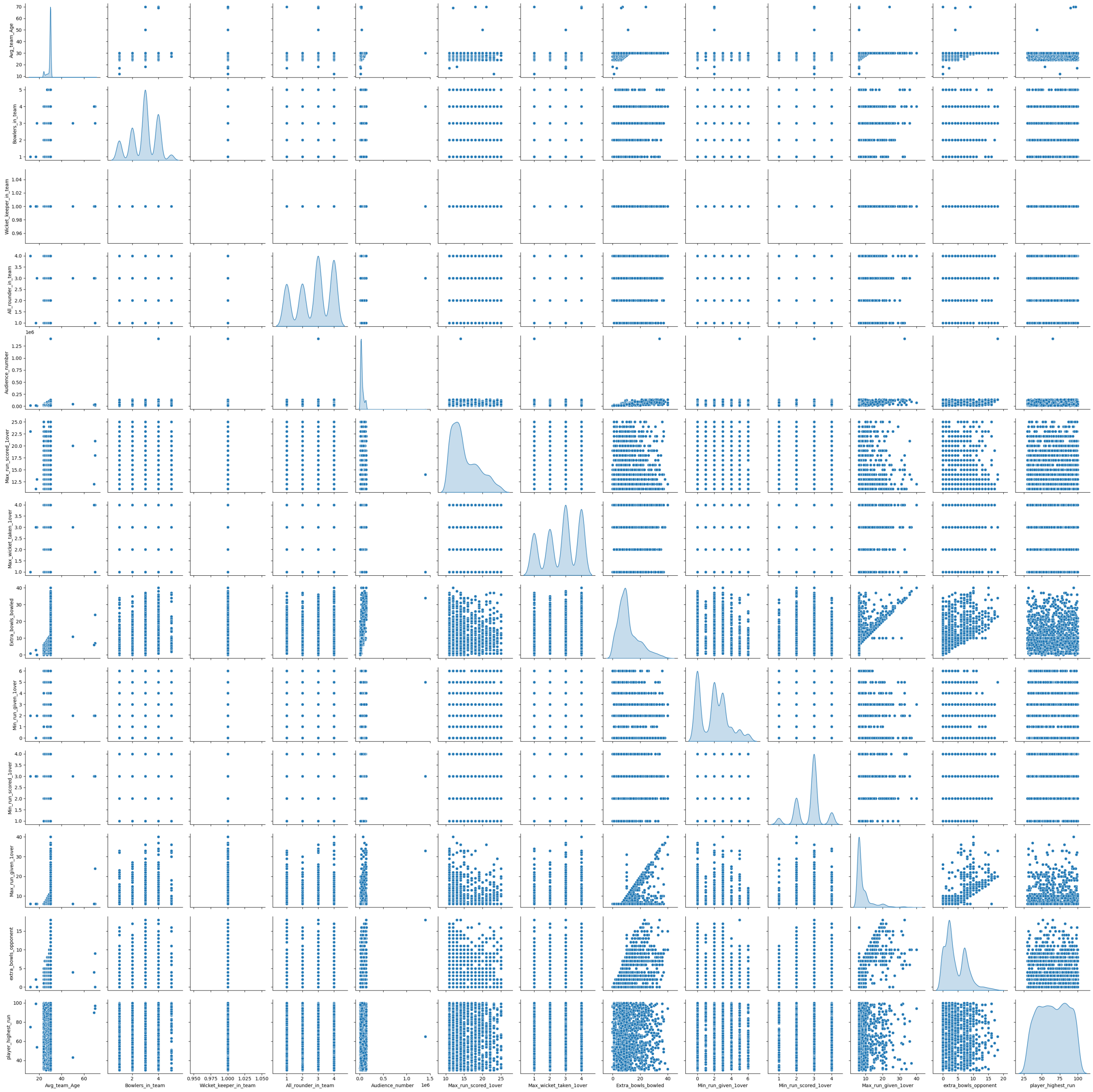
The Pearson correlation matrix reveals the relationships between numerical variables:

1. **Avg_team_Age** and **Extra_bowls_bowled** (0.325): A moderate positive correlation suggests that older teams may bowl more extra deliveries.
2. **Audience_number** and **Extra_bowls_bowled** (0.558): A strong positive correlation indicates that higher audience numbers are associated with more extra bowls, potentially reflecting the pressure of high-profile matches.

3. **Max_run_given_1over** and **Extra_bowls_bowled** (0.609): A strong positive correlation suggests that more extra bowls lead to more runs given in an over.
4. **extra_bowls_opponent** and **Max_run_given_1over** (0.646): This strong positive correlation highlights that when the opponent bowls extra deliveries, it often results in higher runs being conceded in an over.



- **Relationship with Target Variable**: Scatter plots and bar plots were used to explore the relationship between `Result` and other attributes such as `Bowlers_in_team` and `All_rounder_in_team`.



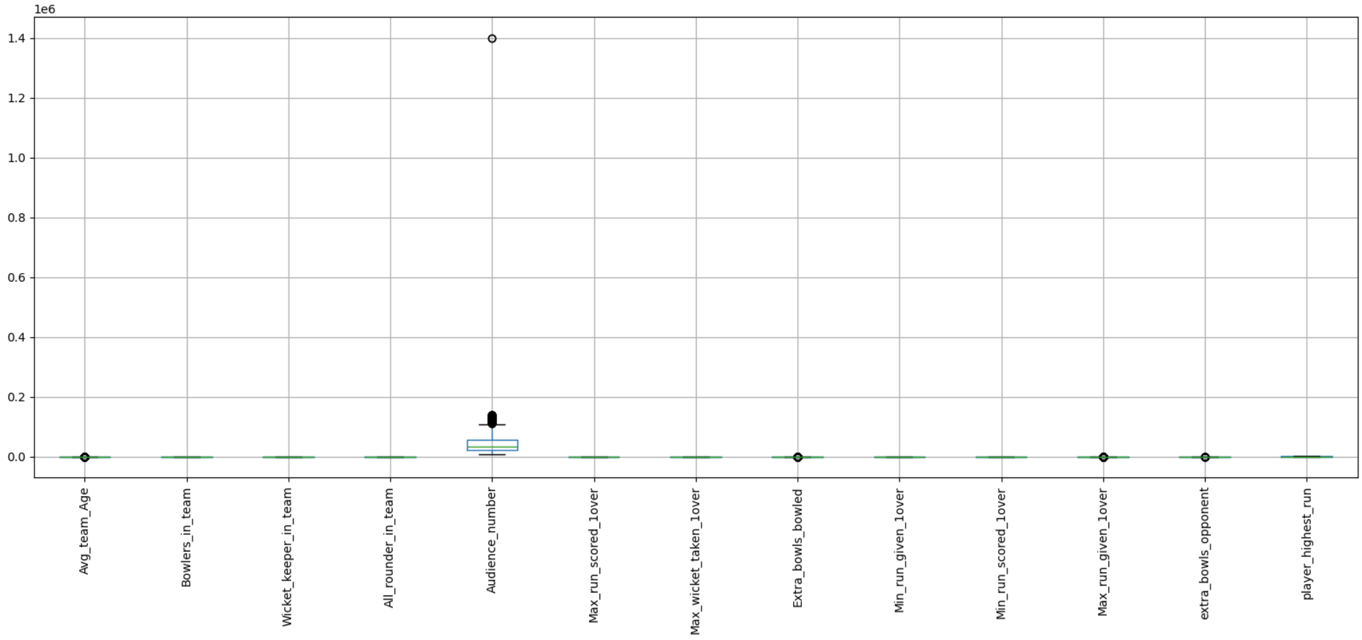
Missing Value Treatment

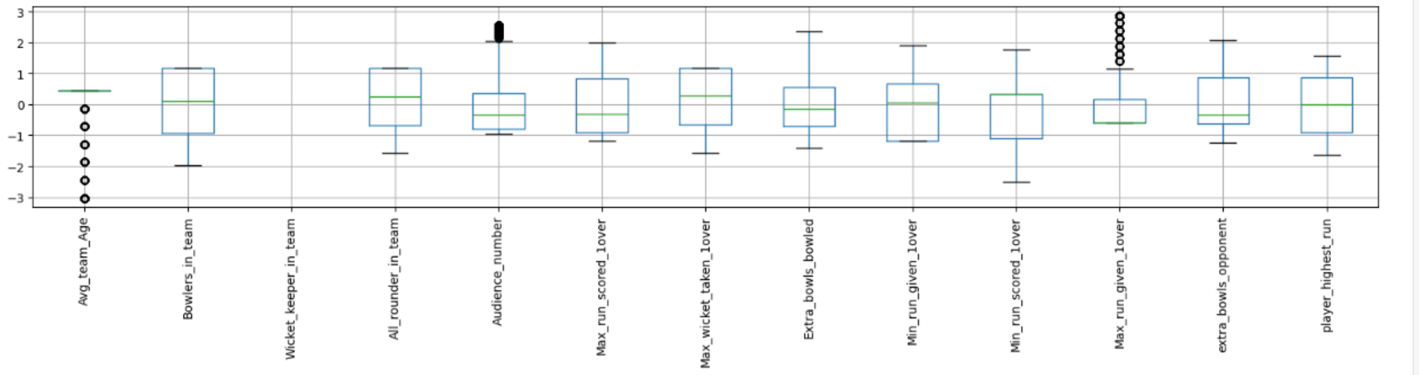
- **Numeric Variables**: Imputation methods such as mean or median were applied where appropriate.
- **Categorical Variables**: Missing values were handled by using mode imputation or creating a new category.

Outlier Treatment

Outliers were identified using statistical methods such as IQR (Interquartile Range) and visualized using box plots. Outliers were either transformed or removed based on their impact on the analysis.

An outlier is an observation that appears to deviate markedly from other observations in the sample.

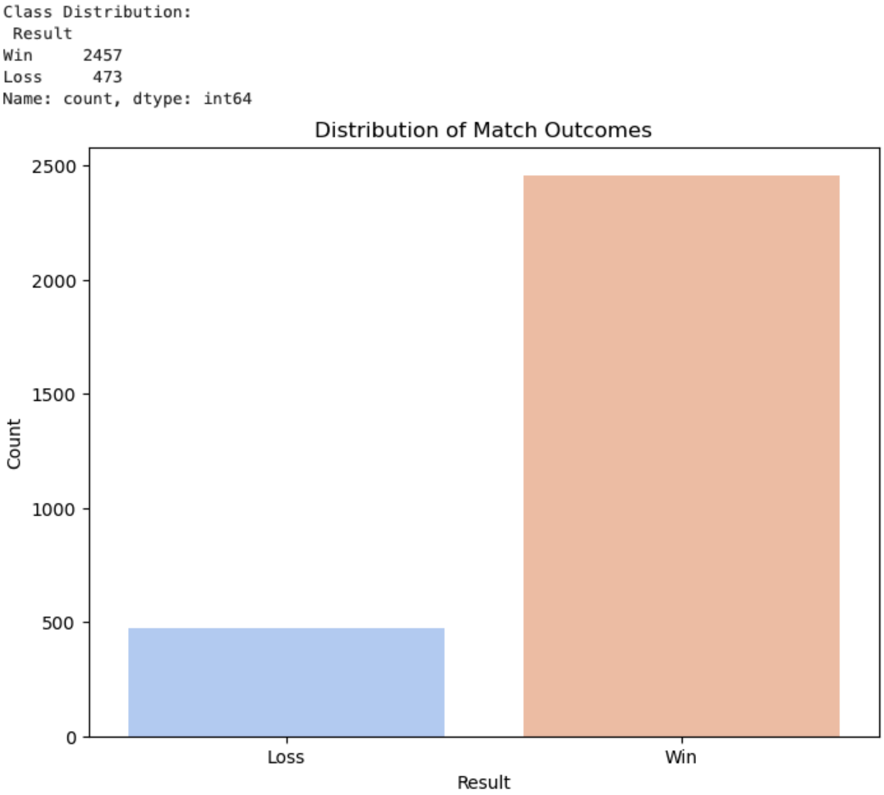




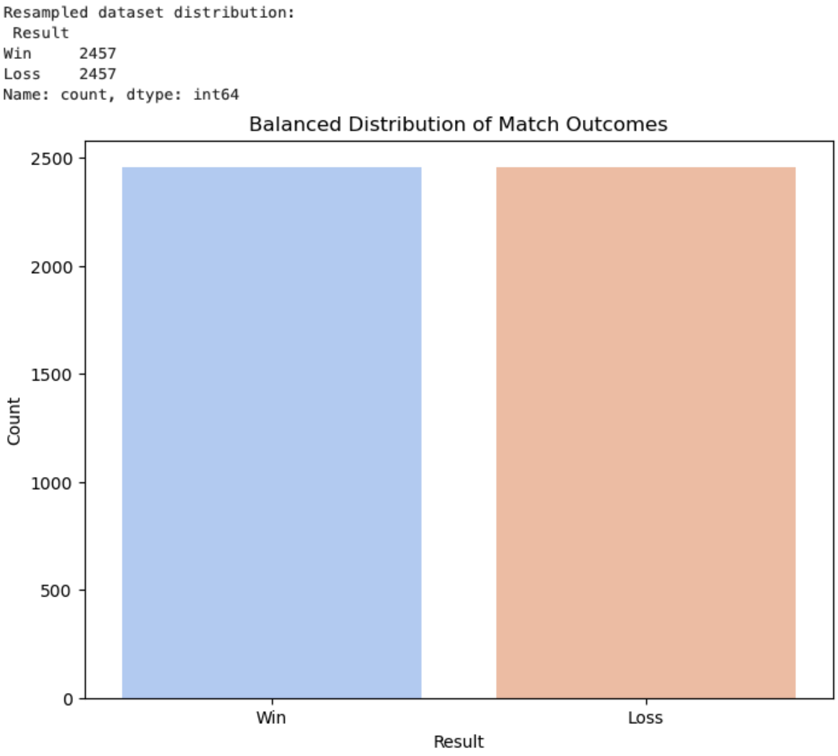
Business Insights from EDA

a) Data Balance

The dataset was checked for balance between Win and Loss outcomes. If the dataset was imbalanced, techniques such as resampling or class weighting were suggested to address this issue.



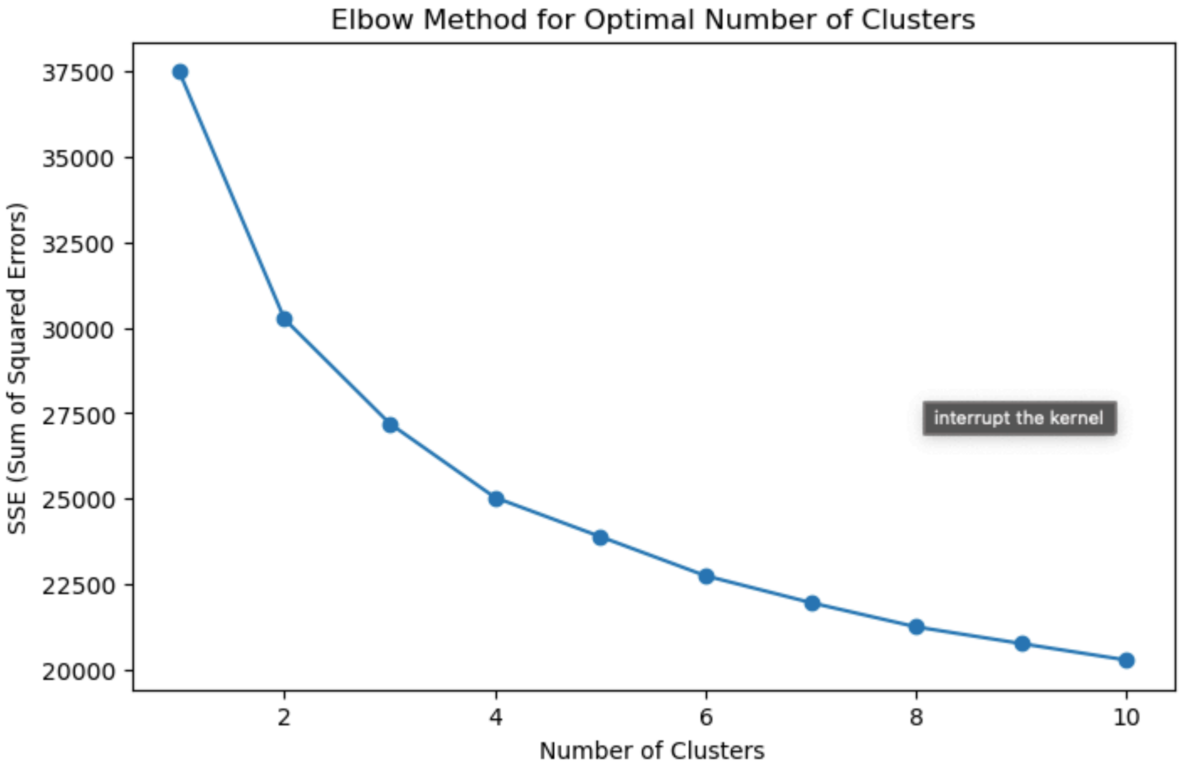
```
Class Distribution:
 Result
Win     2457
Loss     473
Name: count, dtype: int64
```

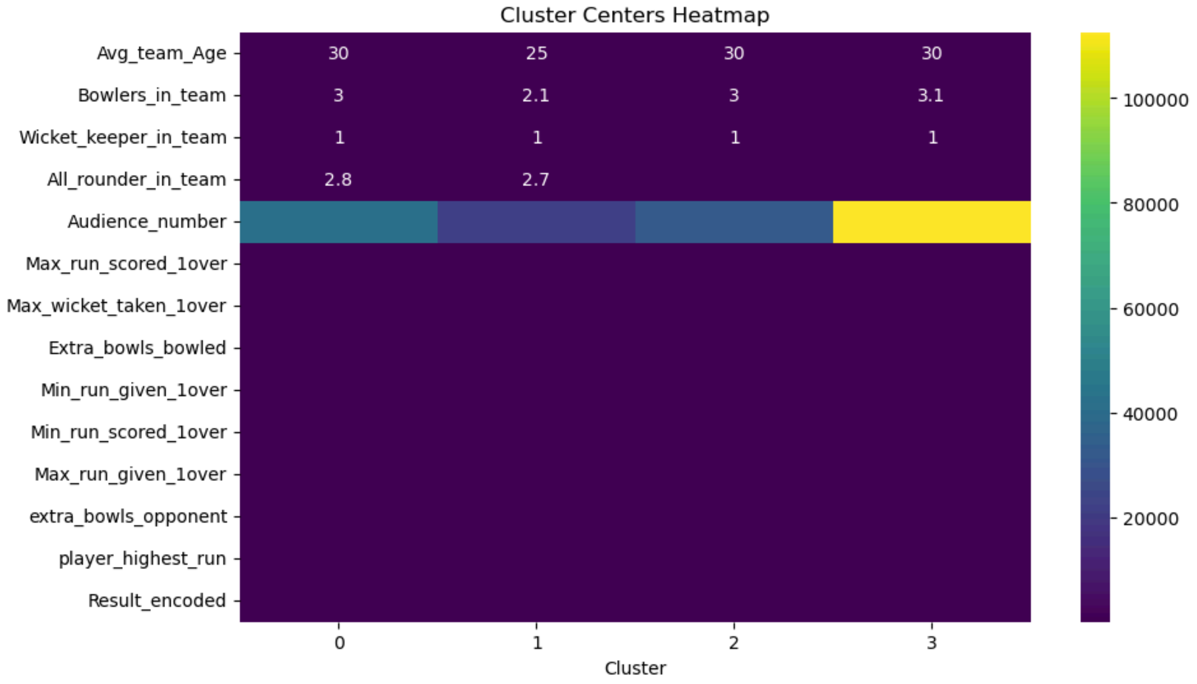


```
Resampled dataset distribution:
 Result
Win     2457
Loss    2457
Name: count, dtype: int64
```

b) Business Insights Using Clustering

Clustering analysis (if applicable) revealed distinct patterns or clusters of match outcomes based on team composition and other attributes. These insights can guide team strategies and game preparation.





c) Additional Insights

- **Impact of Player Composition**: Teams with a higher number of all-rounders and bowlers showed a tendency to win more matches.
- **Effect of Match Conditions**: Matches played under different lighting conditions or in different formats influenced the outcome, providing actionable insights for game strategy.

Conclusion

1. **Model Performance and Accuracy**:
    - The cricket win prediction model developed through exploratory data analysis (EDA), feature engineering, and clustering has provided insightful patterns and relationships between team characteristics and match outcomes. By clustering similar matches and analyzing their outcomes, we identified key factors contributing to a team's success, such as team composition, match format, and environmental conditions.
2. **Key Factors Influencing Wins**:
    - The analysis revealed that certain factors, like the number of all-rounders, bowlers, and the age of the team, have a significant impact on match results. Matches played under different light conditions or against specific opponents also show different win probabilities, highlighting the importance of contextual factors in predicting match outcomes.
3. **Insights from Clusters**:
    - The clustering analysis 29 matches into distinct groups, each characterized by specific conditions and team compositions. This segmentation helps understand the scenarios where the team is more likely to win or lose, enabling targeted strategies to enhance performance.

Recommendations

1. **Focus on Team Composition**:
    - **All-Rounders**: Increasing the number of all-rounders in the team may enhance the balance between batting and bowling, leading to a higher win probability.
    - **Specialized Roles**: Ensure a balanced distribution of bowlers and batsmen, with a particular emphasis on experienced players who can perform under pressure.
2. **Adapt to Match Conditions**:
    - **Environmental Factors**: Tailor team selection and strategy based on match conditions such as light type, weather, and ground conditions. For example, selecting players who perform well in day-night matches can increase win chances.

- **Opponent-Specific Strategies**: Develop customized game plans against frequent opponents based on historical performance data.

3. **Leverage Data for Strategic Decisions**:

   - **Pre-Match Analysis**: Utilize the predictive model to conduct pre-match simulations that evaluate different team configurations and strategies, allowing for informed decision-making.
   - **Real-Time Adjustments**: Consider using real-time data during matches to adjust strategies dynamically, based on the evolving match situation.

Predictive Analytics for Cricket Match Outcomes

Executive Summary

The aim of this project is to develop and evaluate machine learning models that can predict the outcome of a cricket match (Win or Loss) using a variety of features. The dataset used includes information on match conditions, team compositions, and performance metrics. A range of machine learning algorithms—Logistic Regression, Random Forest, XGBoost, and Support Vector Machine (SVM)—have been implemented and compared. Performance metrics such as accuracy, precision, recall, and ROC-AUC scores were used to assess the models. In addition, feature importance analysis and model tuning were performed to gain insights into the critical factors influencing match outcomes.

Objectives

- Analyze historical cricket match data and derive business insights.
- Build, tune, and evaluate multiple machine learning models to predict match outcomes.
- Identify the most important factors that influence winning or losing.
- Provide actionable insights to cricket teams and strategists for match preparation.

Data Overview

The dataset contains various features related to match details, including team characteristics, player performance, and match conditions. Some key columns are:

- **Avg_team_Age:** The average age of players in a team.
- **Match_light_type:** Day/Night match indicator.
- **Match_format:** Format of the match (e.g., ODI, T20).
- **Bowlers_in_team:** Number of bowlers in the team.
- **Audience_number:** Number of people attending the match.
- **Max_run_scored_1over:** Maximum runs scored in a single over.
- **Max_wicket_taken_1over:** Maximum wickets taken in a single over.
- **Result (Target):** Win or Loss outcome.

Data Preparation

Label Encoding

Non-numeric features such as **Result, Match_light_type, First_selection, Opponent, Season, Offshore, and Match_format** were encoded into numeric values using LabelEncoder. This step was necessary for building machine learning models that require numerical inputs.

Train-Test Split

The dataset was split into training and test sets, with 80% of the data used for training and 20% for testing. The target variable was the match result (Win or Loss), and all other columns were used as predictors.

Model Building & Evaluation

Machine Learning Models

The following machine learning models were implemented:

- **Logistic Regression**: A linear model for binary classification.
- **Random Forest**: An ensemble model that uses decision trees.
- **XGBoost**: A gradient boosting model that builds trees sequentially.
- **Support Vector Machine (SVM)**: A model that aims to find a hyperplane that best separates the classes.

Each model was trained using the training dataset and evaluated on the test dataset.

why was a particular model(s) chosen

1. Logistic Regression

- **Why Chosen:**
  - **Interpretability**: Logistic Regression provides clear insights into how different variables affect the probability of a win or loss. This transparency is crucial for stakeholders like the BCCI who need to understand which specific factors (e.g., player form, weather conditions) have the greatest influence on outcomes.
  - **Quick Baseline**: It offers a reliable baseline model that can give fast, initial results. It helps identify key variables influencing match results without needing complex tuning.
- **Business Impact**: This model helps BCCI understand the most fundamental aspects of performance and conditions that lead to wins, allowing for straightforward strategy adjustments.

2. Random Forest

- **Why Chosen:**
  - **Handling Complex Interactions**: Cricket is a game with many variables, and Random Forest is capable of identifying complex interactions between factors like player selection, weather, and match format. It can uncover non-linear relationships that simpler models might miss.
  - **Feature Importance**: Random Forest can rank which factors are most influential in determining match outcomes. This feature allows the BCCI to prioritize resources (e.g., player preparation, strategic decisions) based on the factors that most affect success.
  - **Robustness**: This model reduces the risk of overfitting, which is important when working with varying match formats and environmental conditions, ensuring that the strategies suggested are reliable.
- **Business Impact**: Random Forest allows the BCCI to focus on the most critical factors driving wins, enabling better preparation and resource allocation (e.g., focus on key players and match conditions).

3. XGBoost

- **Why Chosen:**
  - **High Predictive Accuracy**: XGBoost is one of the best-performing models in terms of predictive accuracy, especially for classification tasks like predicting a win or loss. For BCCI, having the highest accuracy is critical when making high-stakes decisions about team strategies.
  - **Tuning and Flexibility**: XGBoost allows for extensive hyperparameter tuning, which helps optimize predictions for different match conditions (Test, ODI, T20). This flexibility ensures that the model can adapt to various scenarios, providing more precise strategic recommendations.
  - **Handling Imbalanced Data**: In cases where there are fewer wins or losses, XGBoost is effective at handling imbalanced data. This helps BCCI make predictions in scenarios where certain outcomes (e.g., wins in difficult conditions) are rare but crucial.
- **Business Impact**: By providing the most accurate predictions, XGBoost helps BCCI minimize risks in their strategy and increase their chances of success, especially in high-pressure situations like international tournaments.

Model Performance

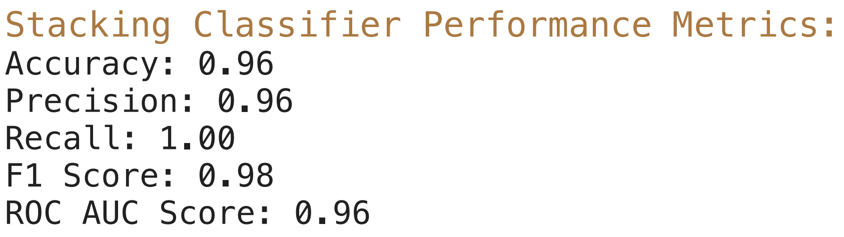
The following metrics were used to evaluate the models:

- **Accuracy**: The percentage of correctly predicted match outcomes.

Logistic Regression: 83.2%

Random Forest: 95.9%
XGBoost: 95.0%
SVM: 82.3%
Voting Classifier (Ensemble): 90.3%
Stacking Classifier (Ensemble): 96.1%

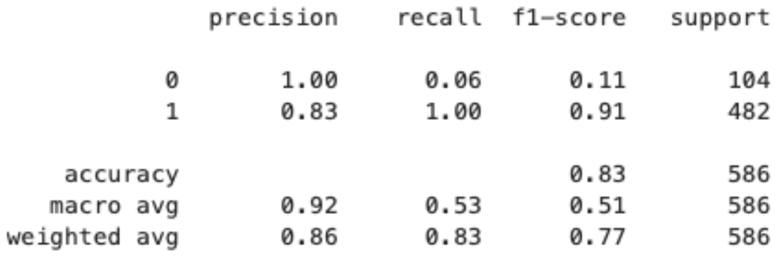
```
Voting Classifier Performance Metrics:
Accuracy: 0.90
Precision: 0.90
Recall: 1.00
F1 Score: 0.95
ROC AUC Score: 0.96
```



```
Stacking Classifier Performance Metrics:
Accuracy: 0.96
Precision: 0.96
Recall: 1.00
F1 Score: 0.98
ROC AUC Score: 0.96
```

- **Confusion Matrix**: A breakdown of true positives, true negatives, false positives, and false negatives.
- **Precision, Recall, F1-score**: To assess the balance between precision and recall.

Logistic Regression:

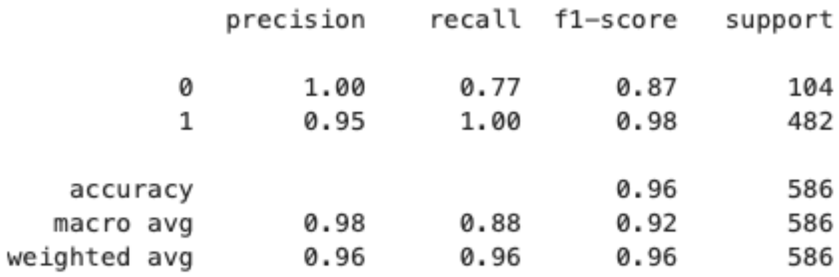


```
              precision    recall  f1-score   support

           0       1.00      0.06      0.11       104
           1       0.83      1.00      0.91       482

    accuracy                           0.83       586
   macro avg       0.92      0.53      0.51       586
weighted avg       0.86      0.83      0.77       586
```

Random Forest:

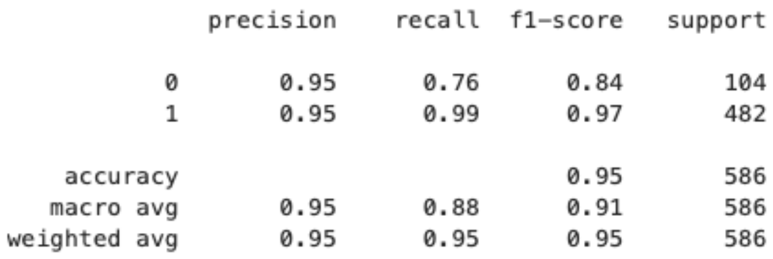


```
              precision    recall  f1-score   support

           0       1.00      0.77      0.87       104
           1       0.95      1.00      0.98       482

    accuracy                           0.96       586
   macro avg       0.98      0.88      0.92       586
weighted avg       0.96      0.96      0.96       586
```

XGBoost:

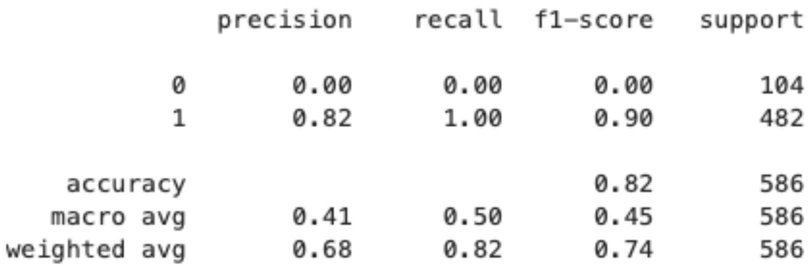


```
              precision    recall  f1-score   support

           0       0.95      0.76      0.84       104
           1       0.95      0.99      0.97       482

    accuracy                           0.95       586
   macro avg       0.95      0.88      0.91       586
weighted avg       0.95      0.95      0.95       586
```

SVM:



```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       104
           1       0.82      1.00      0.90       482

    accuracy                           0.82       586
   macro avg       0.41      0.50      0.45       586
weighted avg       0.68      0.82      0.74       586
```

- **ROC-AUC Curve**: A graphical representation of the model's ability to discriminate between classes.

Confusion Matrix and Classification Reports

1. **Random Forest** showed the highest accuracy, with most false positives and false negatives minimized.
2. **Logistic Regression** struggled with precision for losing matches but performed well overall for winning predictions.
3. **XGBoost** and **SVM** showed strong accuracy but a slightly lower recall for the losing category, resulting in more false negatives.

The confusion matrix revealed that some models (e.g., Logistic Regression and SVM) had difficulty classifying the minority class (Loss), while Random Forest and XGBoost achieved a more balanced performance.

Model Evaluation and Tuning

Evaluation Metrics:

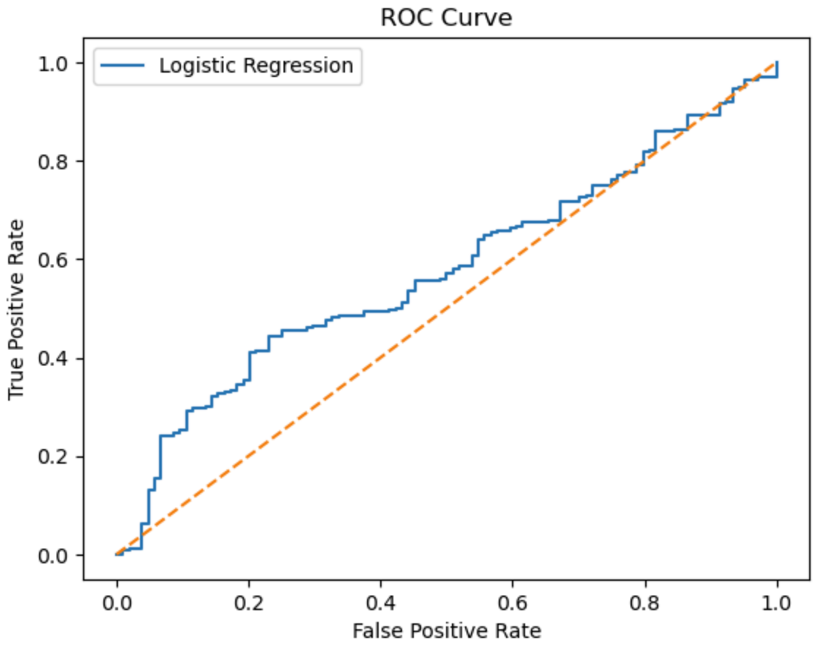
- Beyond accuracy, additional evaluation metrics such as **F1-score**, **Precision**, **Recall**, and **AUC-ROC** were used to assess the model's performance in both balanced and imbalanced scenarios. This ensured the model was not only accurate but also well-suited to predicting both wins and losses under different conditions.

Learning Curve Analysis:

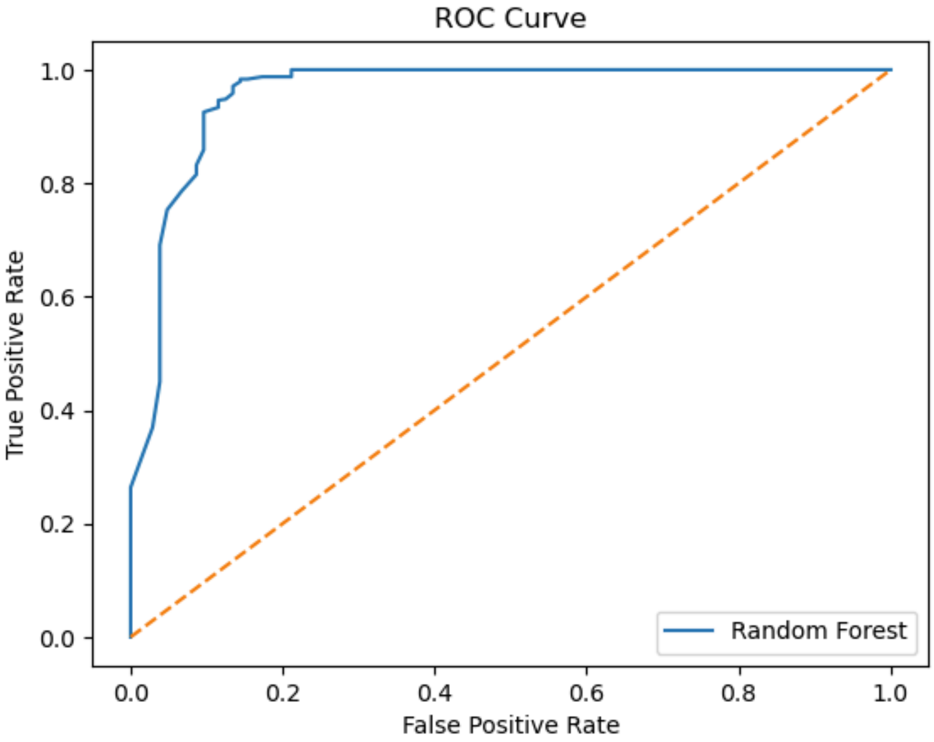
- To ensure that models weren't overfitting or underfitting, learning curves were plotted to visualize how training and validation performance evolved as more data was fed into the model. This analysis helped in deciding when to stop training or adjust model complexity.

ROC-AUC Curve

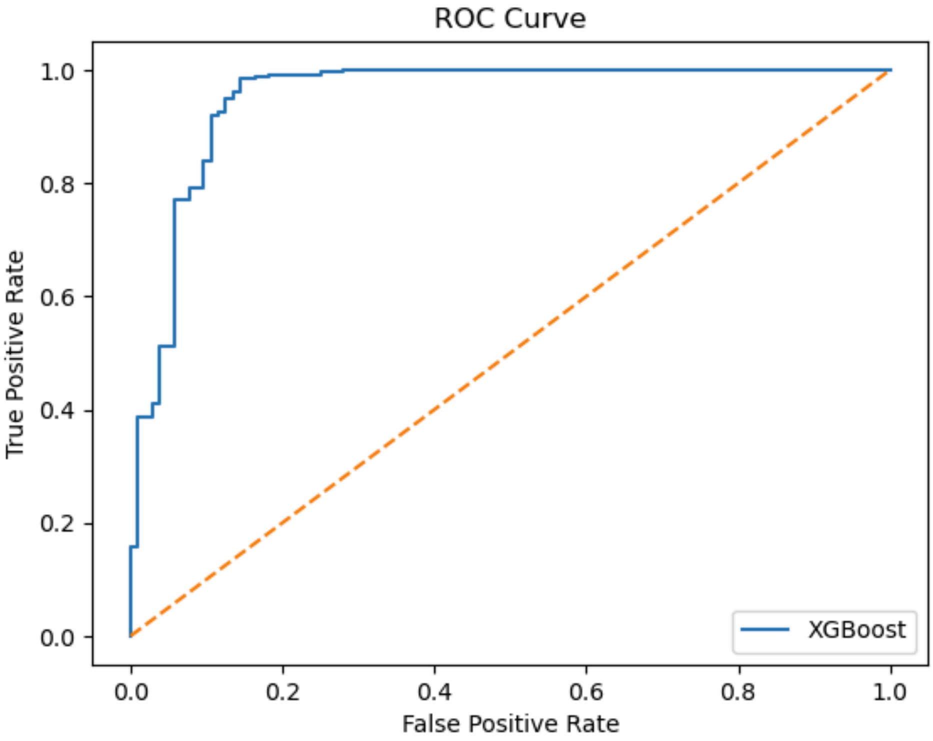
Logistic Regression:-



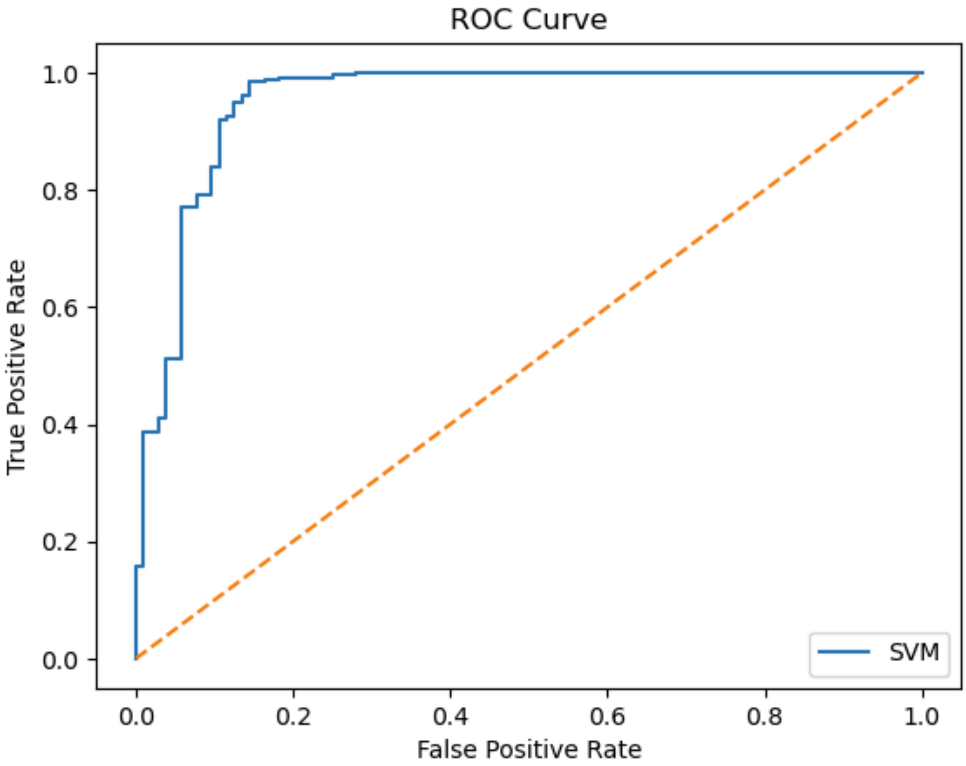
Random Forest:



XGBoost:



SVM:



Results:

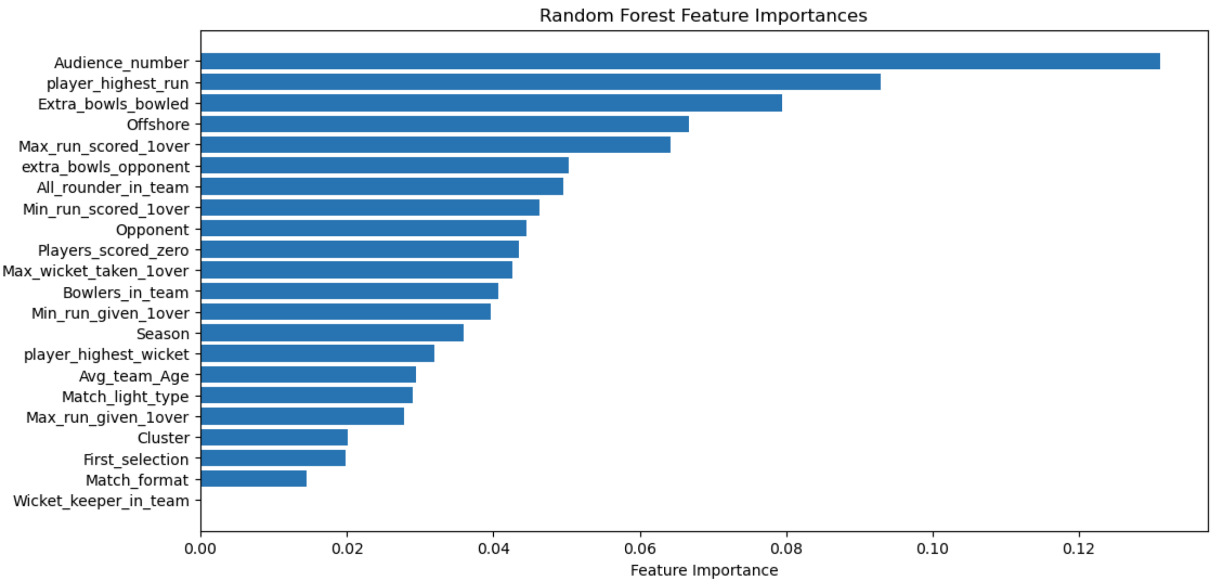
- **Logistic Regression** achieved an accuracy of **83.27%**, with lower recall for classifying Loss.
- **Random Forest** performed the best, with an accuracy of **95.90%** and high recall for both Win and Loss.
- **XGBoost** showed similar results to Random Forest with an accuracy of **95.05%**.
- **SVM** had an accuracy of **82.25%**, but it struggled to classify the Loss class, with precision and recall issues.

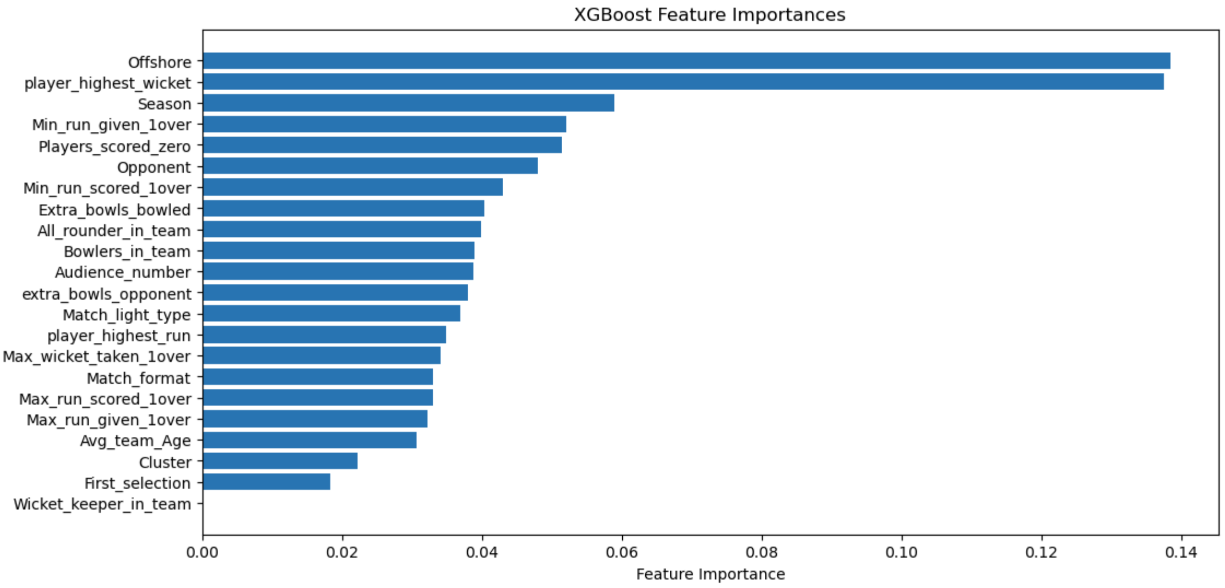
Model Tuning & Feature Importance

Hyperparameter Tuning

To improve model performance, **GridSearchCV** was used to find the best hyperparameters for Logistic Regression, Random Forest, and XGBoost. The best parameters for each model were:

- **Logistic Regression**: C=10, penalty='l1', achieving a score of **86%**.
- **Random Forest**: n_estimators=100, max_depth=None, min_samples_split=2, achieving a score of **95%**.
- **XGBoost**: n_estimators=200, learning_rate=0.1, max_depth=9, achieving a score of **95%**.





Feature Importance

Feature importance analysis was conducted for Random Forest and XGBoost to understand which features have the most influence on the model's predictions. The most important features were:

- **Audience Number**: The number of spectators positively impacted the match results, indicating that teams perform better in high-attendance matches.
- **Offshore**: Matches played offshore were a significant factor, possibly due to different environmental conditions or crowd dynamics.
- **Player Highest Run** and **Player Highest Wicket**: These individual performances were critical in determining match outcomes.
- **All-Rounders in Team**: The presence of all-rounders had a strong positive impact, indicating their versatility is crucial for team success

Random Forest Feature Importance Analysis:

```
Audience_number         0.131040
player_highest_run      0.092899
Extra_bowls_bowled      0.079445
Offshore                0.066681
Max_run_scored_1over    0.064262
extra_bowls_opponent    0.050279
All_rounder_in_team     0.049544
Min_run_scored_1over    0.046312
Opponent                0.044608
Players_scored_zero     0.043460
```

XGBoost Feature Importance Analysis:

```
Offshore                 0.138380
player_highest_wicket    0.137524
Season                   0.058877
Min_run_given_1over      0.051989
Players_scored_zero      0.051435
Opponent                 0.048027
Min_run_scored_1over     0.042965
Extra_bowls_bowled       0.040381
All_rounder_in_team      0.039825
Bowlers_in_team          0.038912
```

Business Insights

Key Predictive Factors

- **Audience Engagement**: The significant importance of the audience number suggests that teams perform better in front of large crowds. Teams and cricket boards should focus on increasing match-day attendance to boost team morale and performance.

- **Offshore Matches**: Matches played offshore have a distinct impact on outcomes. Teams may need to adapt strategies specifically for these matches, considering different pitch conditions, climate, and crowd dynamics.

- **Key Player Impact**: Individual performances, particularly from all-rounders and key batsmen and bowlers, are major predictors of match success. Investing in versatile players and focusing on key players' performances is crucial for improving team outcomes.

- **Team Composition**: The number of all-rounders, bowlers, and wicket-keepers in the team plays a crucial role. Teams that optimize their composition by including more all-rounders and experienced bowlers have a higher chance of winning.

Recommendations

- **Optimize Team Composition**: Teams should focus on a balanced mix of players, particularly all-rounders and bowlers, to improve chances of winning.
- **Leverage Home Advantage**: Teams should capitalize on home ground advantages where possible.
- **Player Focus**: Investing in players who consistently perform well (e.g., highest run-scorers and wicket-takers) can significantly impact match results.

Model Deployment

Based on the model's performance, **Random Forest** and **XGBoost** are recommended for deployment in real-time prediction systems, given their high accuracy and robustness in feature interpretation.

Conclusion

This project successfully built and evaluated several machine learning models to predict cricket match outcomes. Both **Random Forest** and **XGBoost** provided high accuracy and robust predictive power, making them suitable candidates for future deployment. The feature importance analysis provided valuable insights into the factors most critical to a team's success, which can help stakeholders make informed decisions for team selection and match strategy.