**PROGRAMMING FOR ARTIFICIAL INTELLIGENCE**

**(AI-2001)**

**Spring 2024**

# SEMESTER PROJECT - Exploratory Data Analysis – ODI Batting Figures

**PROJECT BY:**

- **MURSIL HASSAN (22K-2132)**

**SECTION: BAI – 4A**

**INSTRUCTOR: SIR SAMEER FAISAL**

## INTRODUCTION:

Cricket, often dubbed as a game of uncertainties and strategic brilliance, captivates audiences worldwide with its rich history, fierce rivalries, and moments of sheer brilliance on the field. Central to the sport's allure is the performance of batsmen, whose ability to score runs under varying conditions shapes the outcome of matches and defines the narrative of cricketing contests.

In this report, we delve into the fascinating realm of cricket batting performance, analyzing a comprehensive dataset spanning player statistics across different formats and eras of the game. Our exploration aims to uncover patterns, trends, and insights that shed light on the nuances of batting excellence and provide valuable perspectives for players, coaches, and cricket enthusiasts alike.

The dataset under examination, ICC ODI Batting Figures - 1971 to 2019, comprises a trove of batting metrics, including metrics related to player performance such as the number of matches played, innings batted, runs scored, highest score, batting average, strike rate, centuries, half-centuries, and ducks. Additionally, the dataset provides information on the duration of each player's career in terms of the starting and ending year. This dataset offers a comprehensive view of ODI batting performances over several decades.

## OVERVIEW OF THE DATA:

The dataset under analysis contains comprehensive statistics pertaining to the batting performance of international cricket players. It encompasses various metrics that offer insights into the proficiency and impact of players during their careers. The dataset comprises the following columns:

- Player: This column denotes the name and country of the cricket player.
- Span: Indicates the duration of the player's career, usually denoted by the years they were active in international cricket.
- Mat: Represents the total number of matches played by the player.
- Inns: Signifies the number of innings batted by the player.
- NO: Denotes the number of times the player remained not out during their innings.
- Runs: Reflects the total runs scored by the player in their career.
- HS: Stands for the player's highest score in a single innings.
- Ave: Represents the batting average of the player, calculated as the total runs scored divided by the number of times dismissed.
- BF: Indicates the total number of balls faced by the player.

- SR: Denotes the player's strike rate, calculated as the number of runs scored per 100 balls faced.
- 100: Represents the number of centuries (100 runs scored in a single innings) scored by the player.
- 50: Signifies the number of half-centuries (50 runs scored in a single innings) scored by the player.
- 0: Denotes the number of times the player was dismissed without scoring any runs (duck).
- Player_URL: Contains the URL or link to additional information about the player.

## PRE-PROCESSING THE DATA:

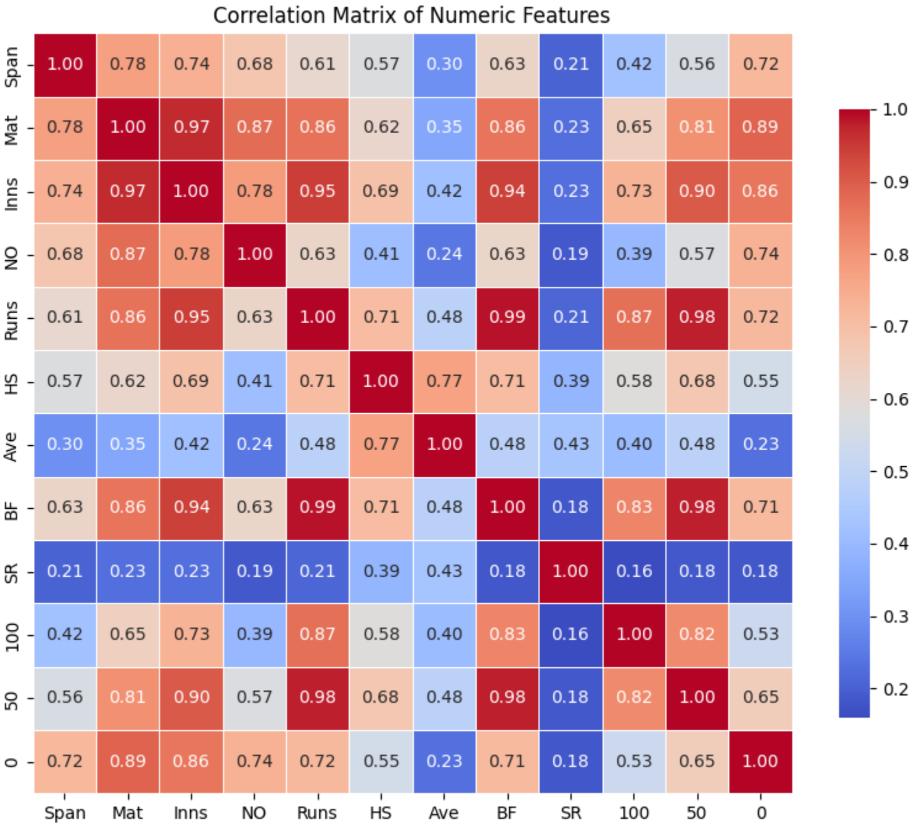
- Loading the Dataset: The initial step involved loading the dataset icc.csv using the Pandas library's read_csv() function. The dataset was encoded using 'iso-8859-1' encoding to handle any special characters.
- Initial Exploration: After loading the dataset, a quick examination of the first 10 rows using the head() function and an overview of column information using the info() function provided insights into the dataset's structure and content.
- Handling Missing Values: The presence of missing values was identified using the isna() function, followed by a calculation of the total missing values for each column using the sum() function. Missing values were handled in several numerical columns by converting them to numeric data types using pd.to_numeric() with errors='coerce'. Additionally, '-' values in the 'Ave' and 'SR' columns were replaced with NaN values.
- Data Imputation: Missing values in the 'Ave' and 'SR' columns were imputed based on other available data. The batting average ('Ave') was imputed as the ratio of 'Runs' to the total number of outs ('Inns' - 'NO'). Similarly, the strike rate ('SR') was imputed as the ratio of 'Runs' to 'BF' (balls faced), multiplied by 100.
- Removing Unnecessary Rows: The dataset contained rows with invalid data towards the end, which were removed using slicing to retain only relevant rows.
- Handling Remaining Missing Values: After removing unnecessary rows, only the 'HS' (highest score) column contained missing values, which could not be imputed. These missing values were retained as NaN since they were not critical for the analysis.
- Feature Engineering: The 'Span' column was split into 'Start_Year' and 'End_Year' columns to calculate the duration of each player's career ('Span'). The original 'Span' column was then replaced with the calculated duration.
- Data Type Conversion: The 'Start_Year' and 'End_Year' columns were converted to integer data type to facilitate calculations.
- Imputing Missing Values with Iterative Imputer: For further data refinement, missing values were imputed using the IterativeImputer from scikit-learn. This approach

leveraged a RandomForestRegressor to estimate missing values iteratively based on other features in the dataset.

- Final Dataset Preparation: The dataset was further cleaned by replacing infinite values and empty strings with NaN values.
- Summary Statistics: Finally, summary statistics were computed using the describe() function to provide an overview of the dataset's numerical attributes.

## QUERIES DEVELOPMENT:

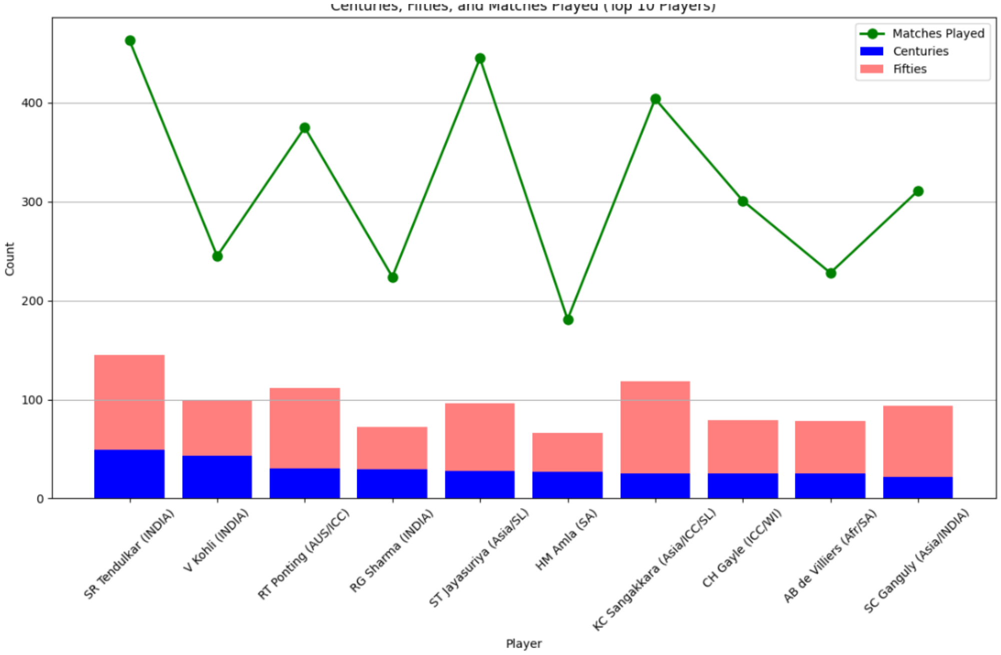
- Correlation Analysis: This query aims to explore the relationships between numerical features in the dataset, such as batting average, strike rate, and other performance metrics. It helps in identifying potential correlations between different aspects of batting performance, providing insights into how certain metrics might influence each other.


Correlation Matrix of Numeric Features

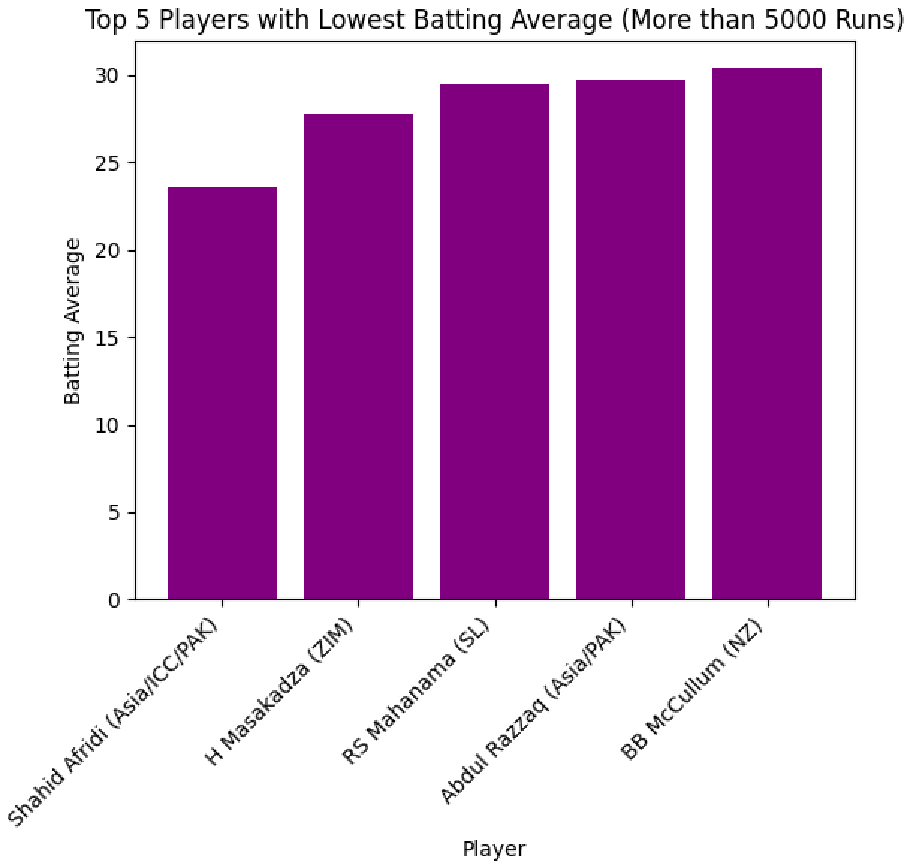
| | Span | Mat | Inns | NO | Runs | HS | Ave | BF | SR | 100 | 50 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Span | 1.00 | 0.78 | 0.74 | 0.68 | 0.61 | 0.57 | 0.30 | 0.63 | 0.21 | 0.42 | 0.56 | 0.72 |
| Mat | 0.78 | 1.00 | 0.97 | 0.87 | 0.86 | 0.62 | 0.35 | 0.86 | 0.23 | 0.65 | 0.81 | 0.89 |
| Inns | 0.74 | 0.97 | 1.00 | 0.78 | 0.95 | 0.69 | 0.42 | 0.94 | 0.23 | 0.73 | 0.90 | 0.86 |
| NO | 0.68 | 0.87 | 0.78 | 1.00 | 0.63 | 0.41 | 0.24 | 0.63 | 0.19 | 0.39 | 0.57 | 0.74 |
| Runs | 0.61 | 0.86 | 0.95 | 0.63 | 1.00 | 0.71 | 0.48 | 0.99 | 0.21 | 0.87 | 0.98 | 0.72 |
| HS | 0.57 | 0.62 | 0.69 | 0.41 | 0.71 | 1.00 | 0.77 | 0.71 | 0.39 | 0.58 | 0.68 | 0.55 |
| Ave | 0.30 | 0.35 | 0.42 | 0.24 | 0.48 | 0.77 | 1.00 | 0.48 | 0.43 | 0.40 | 0.48 | 0.23 |
| BF | 0.63 | 0.86 | 0.94 | 0.63 | 0.99 | 0.71 | 0.48 | 1.00 | 0.18 | 0.83 | 0.98 | 0.71 |
| SR | 0.21 | 0.23 | 0.23 | 0.19 | 0.21 | 0.39 | 0.43 | 0.18 | 1.00 | 0.16 | 0.18 | 0.18 |
| 100 | 0.42 | 0.65 | 0.73 | 0.39 | 0.87 | 0.58 | 0.40 | 0.83 | 0.16 | 1.00 | 0.82 | 0.53 |
| 50 | 0.56 | 0.81 | 0.90 | 0.57 | 0.98 | 0.68 | 0.48 | 0.98 | 0.18 | 0.82 | 1.00 | 0.65 |
| 0 | 0.72 | 0.89 | 0.86 | 0.74 | 0.72 | 0.55 | 0.23 | 0.71 | 0.18 | 0.53 | 0.65 | 1.00 |

- Top Performers Analysis: This query identifies the players with the most centuries and half-centuries in their career, highlighting their significant contributions to batting records. It offers insights into the consistency and longevity of players' performances, showcasing their ability to score big runs consistently or contribute crucial half-centuries.

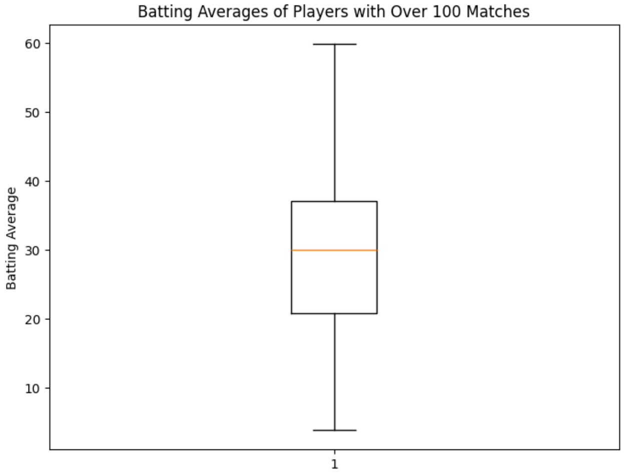
- Century and Fifty Ratios: This query calculates the ratio of centuries and half-centuries scored by players relative to the total number of matches they've played. It helps in identifying players with the highest conversion rates of 100s and 50s per match, indicating their proficiency in converting starts into significant innings.



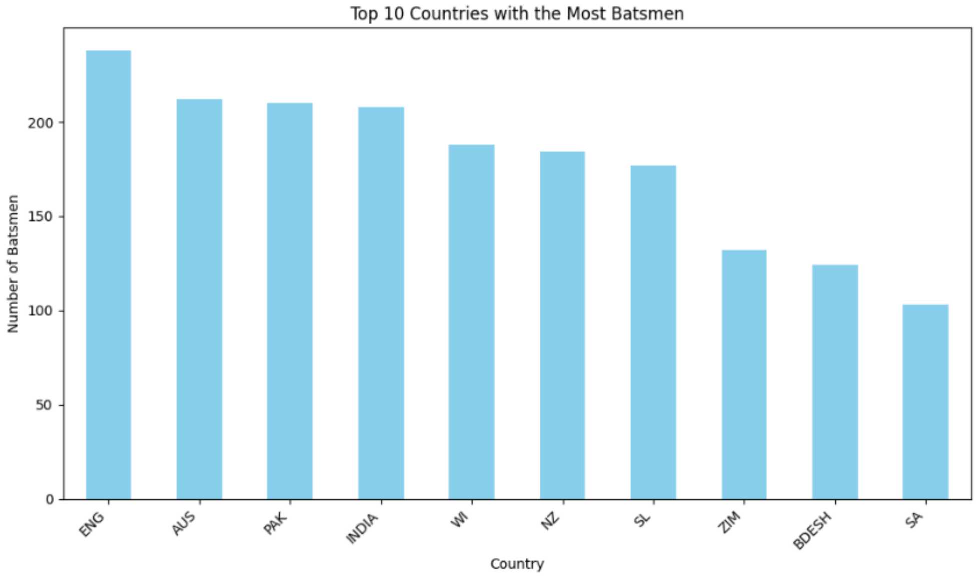
- Player with Longest Span: This query identifies the player with the longest span of international cricket career, showcasing their longevity and endurance in the sport. It provides insights into the players' ability to sustain their performance over an extended period, spanning multiple years or even decades.
- Highest Individual Score: This query identifies the player with the highest individual score (highest runs scored in a single inning) in international cricket. It highlights remarkable individual performances and records set by players in specific matches, showcasing their dominance and skill on the field.
- Most Ducks: This query identifies the player with the most ducks in their international cricket career. It sheds light on the less glamorous aspect of batting performance, highlighting instances where players have struggled or faced challenges in their innings.
- Batting Averages Analysis: This query analyzes the batting averages of players with over 5000 runs in their international cricket career. It identifies players with the highest and lowest batting averages among this group, providing insights into consistency and performance levels among elite batsmen.



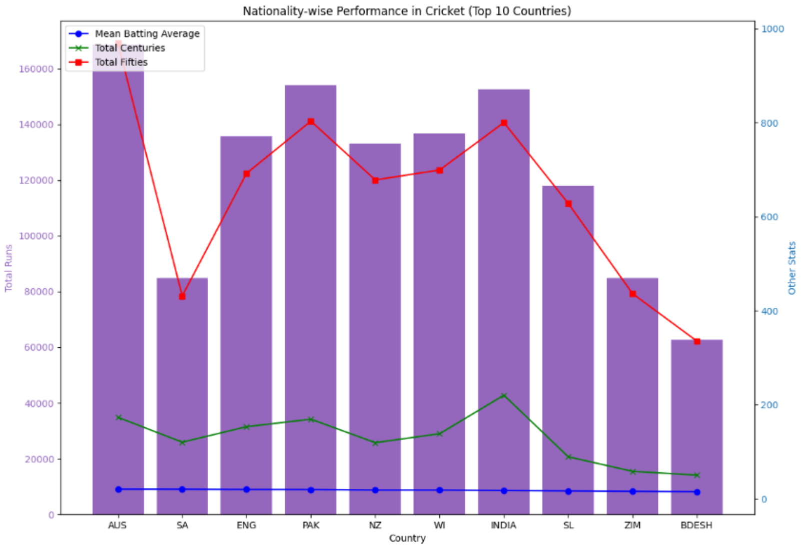
- Not-Out Percentage: This query calculates the percentage of times a player remained not out in their innings, focusing on players with over 100 matches. It identifies players with the highest not-out percentages, showcasing their ability to stay at the crease and contribute to the team's total without getting dismissed.
- Comparison of Batting Averages: This query compares the batting averages of players with over 100 matches, highlighting the player with the highest and lowest batting average among this group. It provides insights into the diversity of batting talent and performance levels across different players.



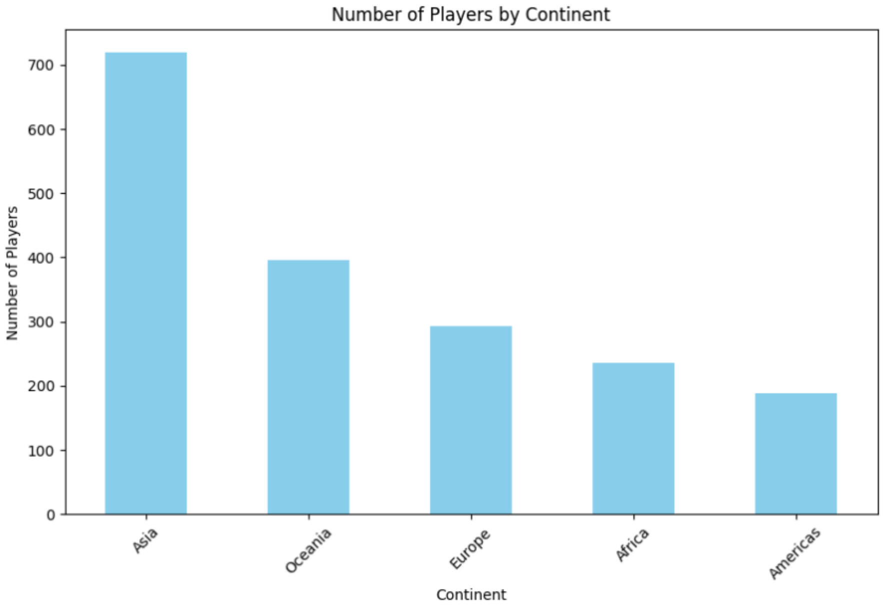
- Comparison of Strike Rates: This query compares the strike rates of players with over 100 matches, highlighting the player with the highest and lowest strike rate among this group. It offers insights into the balance between scoring quickly and preserving wickets among elite batsmen.
- Distribution by Nationality: This query analyzes the distribution of players by nationality, highlighting the countries with the most representation in international cricket. It provides insights into the global diversity of cricket and the dominance of certain cricketing nations in producing top players.

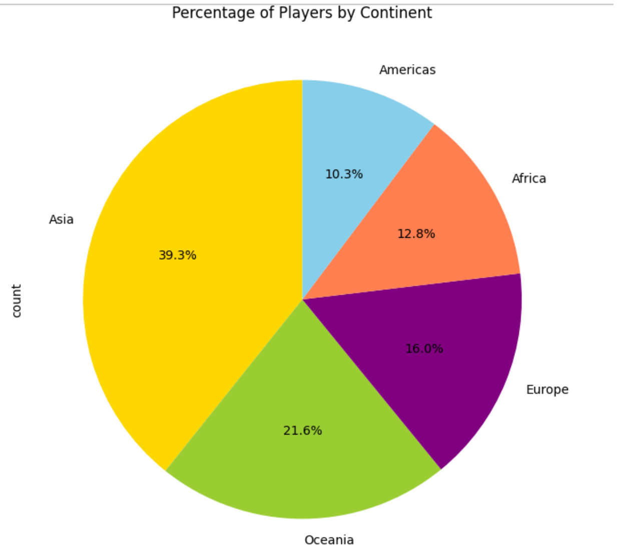


- Performance by Nationality: This query aggregates batting performance statistics by nationality, focusing on the top 10 countries with the most representation. It offers insights into the average batting averages, total runs scored, centuries, and fifties by players from different countries, showcasing national trends in cricketing performance.

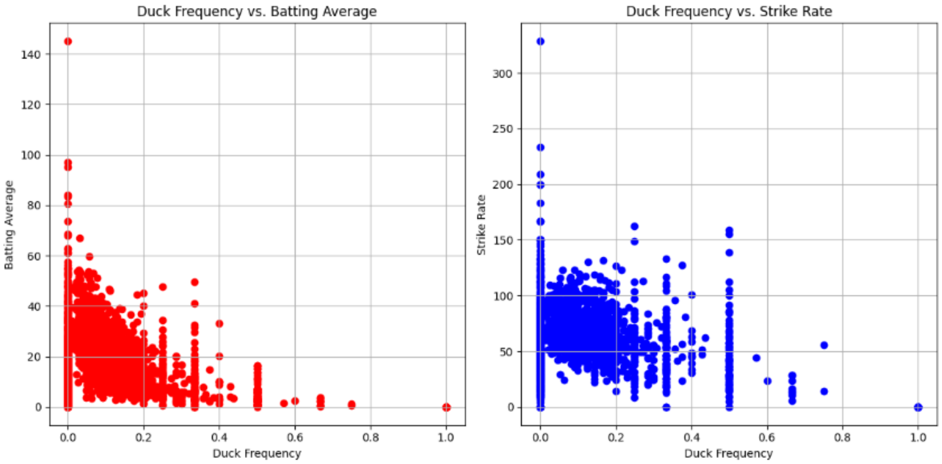


- Player Distribution by Continent: This query categorizes players based on their nationality and maps them to their respective continents. It provides insights into the geographical distribution of cricketing talent, showcasing the prominence of different continents in producing top players.

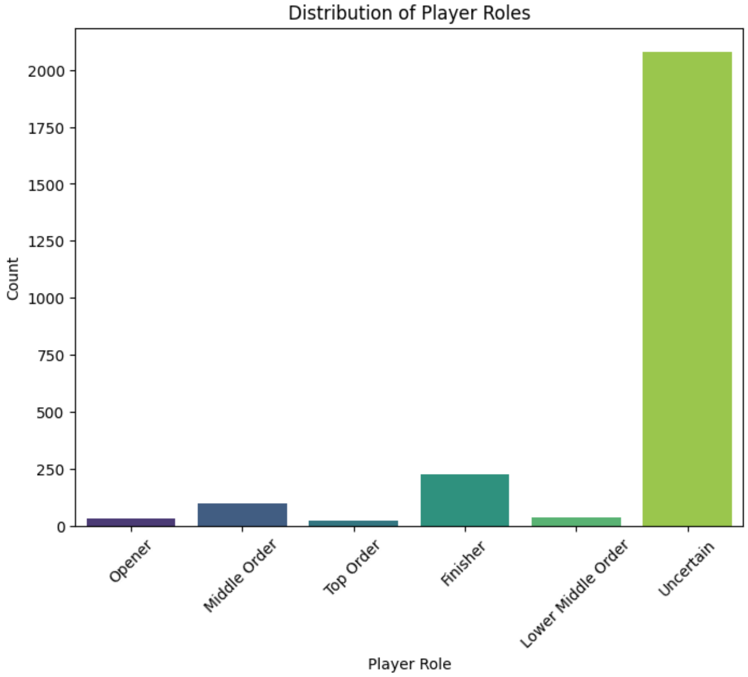




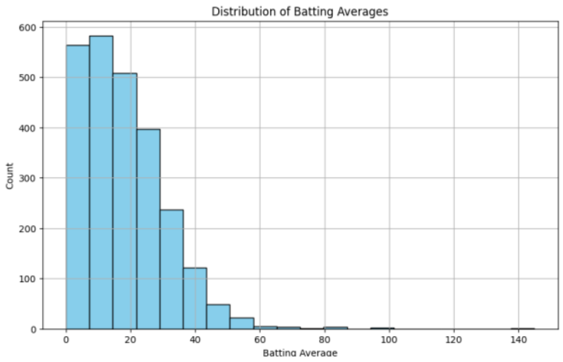
- Consistency Analysis: This query identifies the player with the most consistent batting performance, focusing on the standard deviation of batting averages. It offers insights into the stability and reliability of players' performances over time, highlighting those who consistently contribute to their team's success.
- Correlation with Duck Frequency: This query analyzes the correlation between duck frequency and batting averages/strike rates. It provides insights into the relationship between failures and overall batting performance, highlighting the impact of dismissals on player statistics.

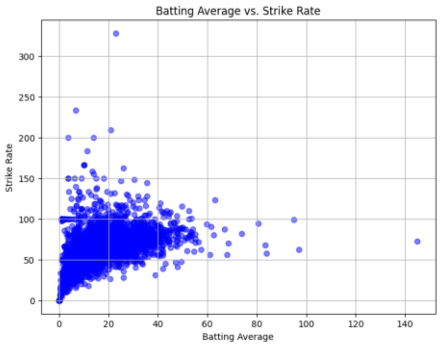


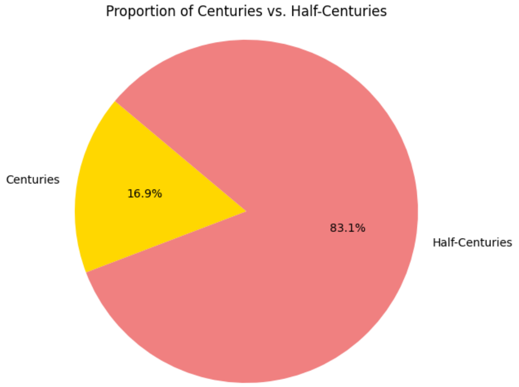
- Player Role Identification: This query categorizes players into different batting roles (e.g., opener, top order, middle order, finisher) based on their batting averages and strike rates. It helps in identifying the best players for specific batting positions, providing insights into team composition and strategy.

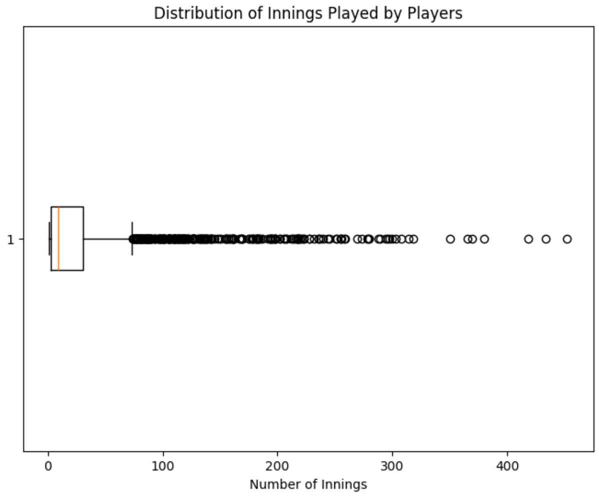


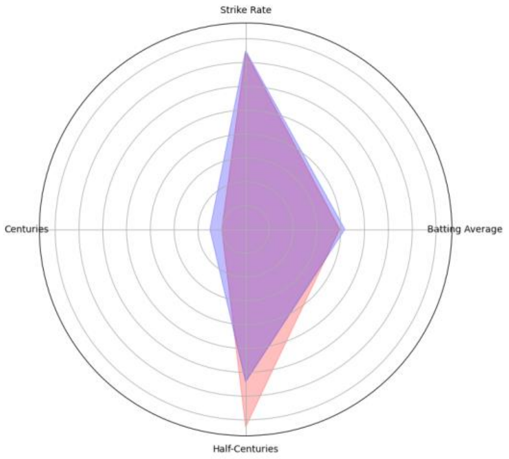
# FURTHER VISUALIZATIONS:

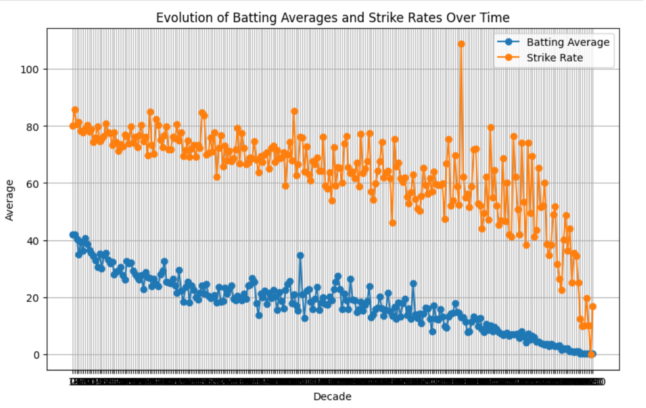












# RESULTS INTERPRETATION:

The analysis of cricket batting performance yielded several key findings and notable trends that provide valuable insights into the dynamics of the sport.

- Correlation Analysis: The correlation matrix revealed significant correlations between batting averages, strike rates, and other performance metrics. It showed that batting average and strike rate are positively correlated, indicating that players with higher batting averages tend to have higher strike rates as well.
- Top Performers Analysis: SR Tendulkar emerged as the top performer in terms of scoring centuries and half-centuries, showcasing his consistency and ability to make substantial contributions to his team's total runs.
- Century and Fifty Ratios: The analysis of century and fifty ratios highlighted players' abilities to convert starts into significant scores. Top players like SR Tendulkar, V Kohli, RT Ponting, RG Sharma, ST Jayasuriya, HM Amla, KC Sangakkara, CH Gayle, AB de Villiers, SC Ganguly and SR Tendulkar, KC Sangakkara, JH Kallis, Inzamam-ul-Haq, R Dravid, RT Ponting, DPMD Jayawarde, MS Dhoni, SC Ganguly, ST Jayasuriya demonstrated impressive ratios of centuries and half-centuries relative to the number of matches they've played.
- Player with Longest Span: SR Tendulkar emerged as the player with the longest international cricket career, underscoring the importance of longevity and endurance in sustaining performance over time.
- Highest Individual Score: RG Sharma achieved the highest individual score in international cricket, showcasing remarkable individual performances and setting records in specific matches.
- Most Ducks: ST Jayasuriya had the unfortunate distinction of having the most ducks in their international cricket career, highlighting the challenges and occasional setbacks faced by players in their batting innings.
- Batting Averages Analysis: Analysis of batting averages among players with over 5000 runs revealed both the highest and lowest batting averages, offering insights into consistency and performance levels among elite batsmen.
- Not-Out Percentage: Players like EJ Chatfield, MS Dhoni demonstrated high not-out percentages, indicating their ability to stay at the crease and contribute to their team's total without getting dismissed.
- Comparison of Batting Averages and Strike Rates: The comparison of batting averages and strike rates among players with over 100 matches highlighted both the highest and lowest averages and strike rates, showcasing the diversity of batting talent and performance levels across different players.
- Distribution by Nationality: The distribution of players by nationality revealed the dominance of certain cricketing nations in producing top players, with England having the most representation in international cricket.
- Player Distribution by Continent: Categorization of players by continent highlighted the geographical distribution of cricketing talent, showcasing the prominence of different continents in producing top players, with Asia being the most represented continent.

- Consistency Analysis: B Lee emerged as the player with the most consistent batting performance, demonstrating stability and reliability in their performances over time.
- Correlation with Duck Frequency: The correlation analysis between duck frequency and batting averages/strike rates revealed insights into the relationship between failures and overall batting performance, highlighting the impact of dismissals on player statistics.
- Player Role Identification: Identification of players into different batting roles based on their batting averages and strike rates provided insights into the best players for specific batting positions, contributing to team composition and strategy.

## CONCLUSIONS AND RECOMMENDATIONS:

Based on the analysis conducted, several conclusions and actionable recommendations can be drawn to enhance player and team performance in cricket:

- Selection Strategies: Teams can use insights from top performers and consistent players to inform their selection strategies, focusing on players with proven track records of scoring big runs and maintaining consistency in their performances.
- Training Focus Areas: Coaches and players can identify areas for improvement based on batting averages, strike rates, and other performance metrics. Emphasizing training in areas such as converting starts into big scores, improving consistency, and minimizing dismissals can lead to enhanced batting performance.
- Team Tactics: Teams can tailor their tactics and batting lineups based on player roles and performance trends identified in the analysis. Strategic decisions such as player positioning in the batting order, partnerships, and match situations can be optimized to maximize run-scoring opportunities and team success.
- Further Research: Areas for further research or analysis include exploring the impact of external factors such as pitch conditions, match formats, and opposition strengths on batting performance. Additionally, longitudinal studies tracking player development and performance trajectories over time can provide deeper insights into the evolution of batting excellence in cricket.

## Conclusion:

In conclusion, the analysis of cricket batting performance using data-driven approaches has provided valuable insights into the intricacies of the sport. Understanding the nuances of batting performance, identifying top performers, and leveraging data insights can play a pivotal role in enhancing player and team performance in cricket. By harnessing the power of data analytics and strategic decision-making, cricket teams can optimize their resources, maximize player potential, and strive for excellence on the field, contributing to the rich tapestry of cricketing history and legacy.