BATTING PERFORMANCE ANALYSIS IN IPL (2008-2023) USING PYTHON

PROJECT REPORT

DOMAIN: SPORTS ANALYTICS
TOOLS USED: PYTHON, PANDAS, MATPLOTLIB, SEABORN

SANJAI M
M.Sc Data Science and Analytics Student

CONTENTS

- INTRODUCTION
- OBJECTIVE OF THE PROJECT
- DATASET DESCRIPTION
- TOOLS AND TECHNOLOGIES USED
- DATA CLEANING AND PREPARATION
- EXPLORATORY DATA ANALYSIS (EDA)
- KEY INSIGHTS
- FUTURE SCOPE
- CONCLUSION
- REFERENCES

INTRODUCTION

Cricket is one of the most data-driven sports, where player performance plays a crucial role in team selection and match strategy. Batting performance analysis helps teams and analysts understand a player's consistency, scoring ability, and overall contribution to the game.

This project focuses on analyzing batting performance in the Indian Premier League (IPL) using data analytics techniques and visualizations.

With the rapid growth of limited-overs cricket, especially T20 formats like the IPL, the importance of data-driven decision-making has increased significantly. Traditional evaluation based only on runs scored is no longer sufficient; modern analysis considers multiple performance indicators such as batting average, strike rate, boundary-hitting ability, and match consistency. These metrics provide deeper insights into a batter's playing style and impact on the team.

By applying data analytics techniques to IPL batting statistics, this project aims to objectively evaluate player performance and identify patterns that distinguish consistent batters from aggressive scorers. Visualizations are used to simplify complex data and present insights in an intuitive manner, making the analysis useful for analysts, team management, and cricket enthusiasts.

OBJECTIVE OF THE PROJECT

The main objectives of this project are:

1. To analyze IPL batting performance using historical data in order to understand overall player contributions across multiple seasons.
2. To compare batters based on key performance metrics such as total runs scored, batting average, and strike rate, which are essential indicators of consistency and scoring efficiency.
3. To identify consistent batters who maintain high averages over time, as well as aggressive batters who score at a higher strike rate, thereby highlighting different playing styles.
4. To visualize batting trends and performance patterns using data analytics and graphical techniques, enabling clear interpretation and effective comparison of players.

The main objective of this project is to apply data analytics techniques to cricket batting data and derive meaningful insights that can support player evaluation and decision-making in sports analytics.

DATASET DESCRIPTION

The dataset used in this project consists of IPL batting statistics collected in CSV format. It contains player-wise batting performance data across multiple seasons of the Indian Premier League. Each record in the dataset represents the aggregated batting performance of a player, allowing for comprehensive comparison and analysis.

The dataset includes key attributes such as player name, team, number of matches played, innings batted, total runs scored, batting average, and strike rate. In addition, boundary-related metrics such as the number of fours and sixes, along with milestone statistics like centuries and half-centuries, are also provided. These attributes collectively help in evaluating both the consistency and aggressiveness of batters.

Before analysis, the dataset was examined for completeness and accuracy. Missing values and data type inconsistencies were handled during the data cleaning process to ensure reliable results. The dataset is well-structured and suitable for exploratory data analysis and visualization, making it an appropriate choice for batting performance evaluation in sports analytics.

TOOLS AND TECHNOLOGIES USED

The following tools and technologies were used to carry out the batting performance analysis effectively:

Python was used as the primary programming language for data analysis due to its simplicity, flexibility, and strong support for data analytics libraries. It enabled efficient data manipulation, analysis, and visualization.

Pandas was used for data cleaning, preprocessing, and data manipulation. It helped in handling missing values, converting data types, filtering records, and performing statistical operations on the dataset.

Matplotlib was used to create basic visualizations such as bar charts and scatter plots. It provided control over chart customization, allowing clear representation of batting performance metrics.

Seaborn was used for advanced and visually appealing statistical visualizations. It simplified the process of comparing players and identifying trends through high-level plotting functions.

Jupyter Notebook was used as the development environment for writing and executing Python code. It allowed step-by-step analysis with code, output, and visualizations in a single interactive document, making the analysis easy to understand and reproduce.

These tools together provided an efficient and structured workflow for performing exploratory data analysis and deriving meaningful insights from the IPL batting dataset.

DATA CLEANING AND PREPARATION

Data cleaning and preparation is a crucial step in any data analytics project, as the quality of insights depends on the accuracy and consistency of the data. Before performing the analysis, the IPL batting dataset was carefully examined to ensure it was suitable for analytical purposes.

Initially, the dataset was checked for missing values across all columns. Columns containing numerical data such as runs, batting average, and strike rate were verified to identify any null or invalid entries. Where necessary, appropriate data type conversions were performed to ensure numerical values were correctly represented for analysis.

Invalid or inconsistent records were handled to maintain data integrity. Rows with missing or incorrect key performance metrics were removed to avoid biased or misleading results. Additionally, only relevant columns required for batting performance analysis were selected, helping to simplify the dataset and improve analysis efficiency.

Through these data cleaning and preparation steps, the dataset was transformed into a structured and reliable format, making it ready for exploratory data analysis and visualization. This process ensured that the subsequent analysis was accurate, consistent, and meaningful.

3) CLEANING AND TRANSFORMING THE DATA

```
df.isnull().sum()
df['Avg'] = pd.to_numeric(df['Avg'], errors='coerce')
df = df.dropna(subset=['Avg'])
```

This step checks for missing values in the dataset, converts the batting average column into a numeric format, and removes records with invalid or missing average values to ensure clean and reliable data for analysis.

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) was performed to understand the distribution, trends, and relationships within the IPL batting dataset. EDA helps in identifying meaningful patterns and gaining insights before drawing conclusions from the data.

Key batting performance metrics such as total runs, batting average, and strike rate were analyzed to compare players. Bar charts were used to visualize top run scorers, making it easy to identify players with high run contributions. Scatter plots were used to examine the relationship between batting average and strike rate, helping to distinguish between consistent batters and aggressive scorers.

Additionally, comparisons across teams and players highlighted variations in batting styles and performance levels. Visualizations played a crucial role in simplifying complex data and presenting insights in an intuitive and understandable manner.

EDA will provide a strong foundation for interpreting batting performance trends and supported data-driven insights into player evaluation.

TOP 10 IPL RUN SCORERS

The bar chart represents the top 10 highest run scorers in IPL history. It highlights players who have consistently contributed with the bat over multiple seasons. This visualization helps in identifying players with long-term impact rather than short-term performance.


Top 10 IPL Run Scorers
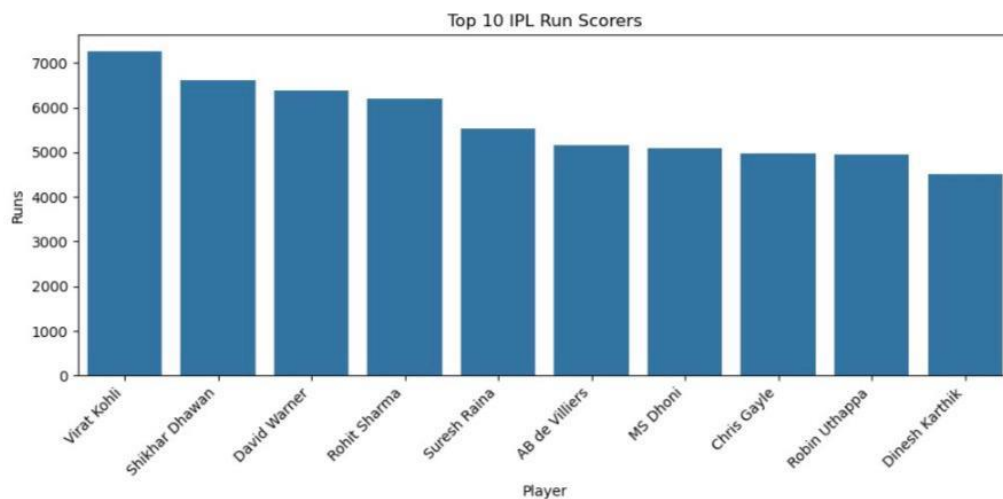
Observation:

Virat Kohli leads the chart as the highest run scorer, followed by Shikhar Dhawan and David Warner. These players have played a large number of matches and maintained consistency across seasons. The presence of experienced batters such as Rohit Sharma, Suresh Raina, and MS Dhoni indicates the importance of longevity and stability in achieving high cumulative run totals in the IPL.

BATTING AVERAGE VS STRIKE RATE

This scatter plot illustrates the relationship between batting average and strike rate for IPL batters. Each point represents an individual player, allowing comparison of consistency and scoring speed.


Batting Average vs Strike Rate
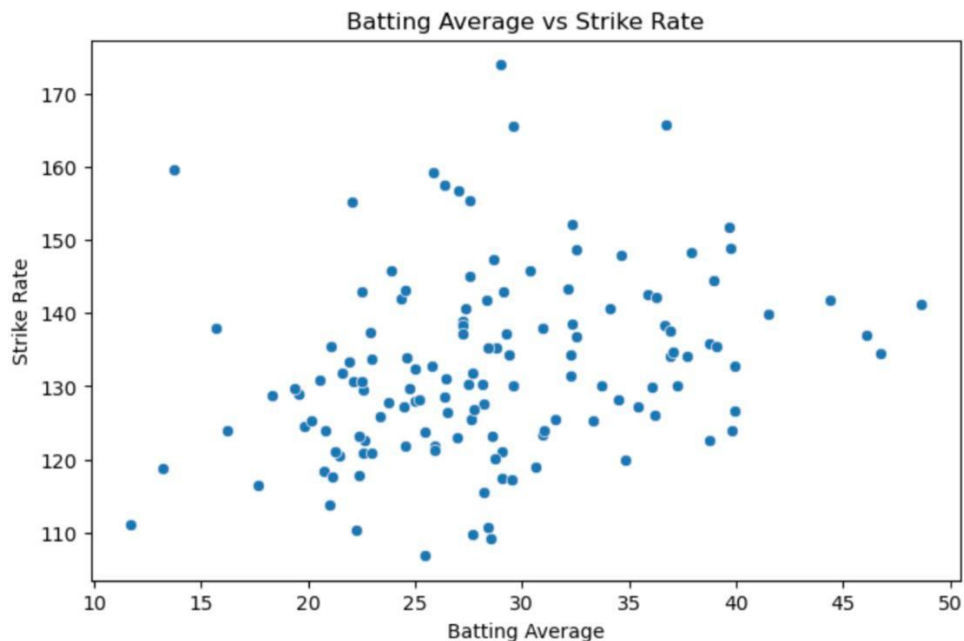
Observation:

The plot shows that players with higher batting averages tend to be more consistent, while players with higher strike rates focus on aggressive scoring. Some players fall into a balanced zone, maintaining both a healthy average and a strong strike rate. This indicates that successful T20 batters often strike a balance between reliability and aggression rather than excelling in only one metric.

BOUNDARY HITTING ANALYSIS (4S + 6S)

The bar chart displays the top 10 boundary hitters based on the combined number of fours and sixes scored. Boundary-hitting ability is a crucial factor in T20 cricket as it directly impacts the run rate and match momentum.


Top 10 Boundary Hitters (4s + 6s)
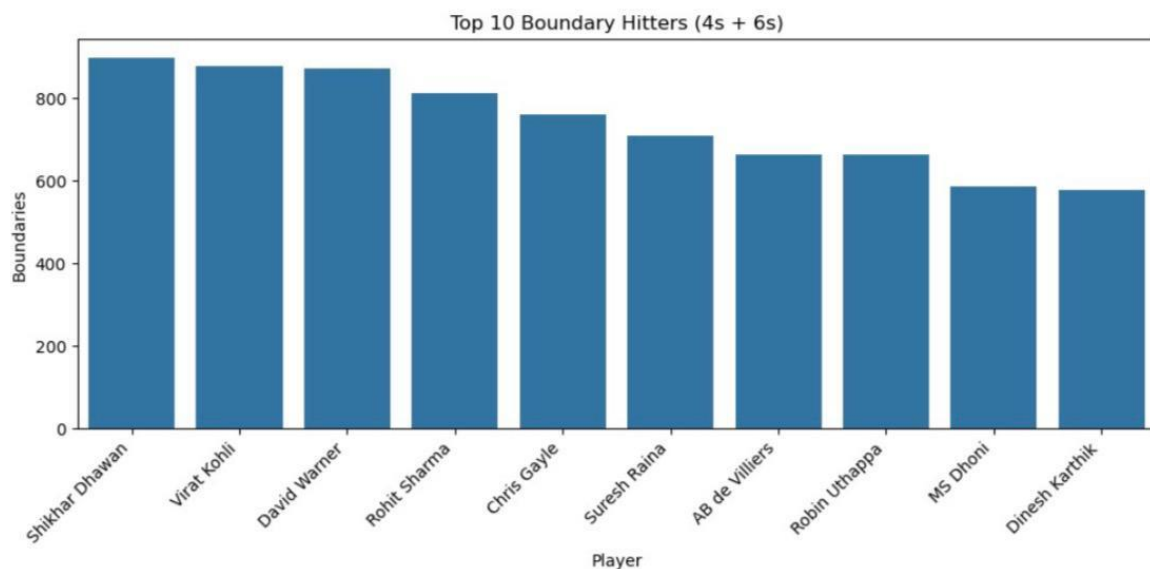
Observation:

Players such as Shikhar Dhawan, Virat Kohli, and David Warner dominate boundary scoring, reflecting their attacking intent and ability to score quickly. Batters with high boundary counts play a significant role in accelerating innings and putting pressure on opposition bowlers. This analysis highlights the importance of power-hitting and shot selection in modern T20 cricket.

TEAM-WISE RUNS ANALYSIS

A bar chart was used to visualize the total runs scored by each IPL team. This analysis helps in understanding the overall batting strength and consistency of teams across seasons.

The chart compares teams based on the cumulative runs scored by their players, highlighting franchises that have maintained strong batting line-ups over the years.
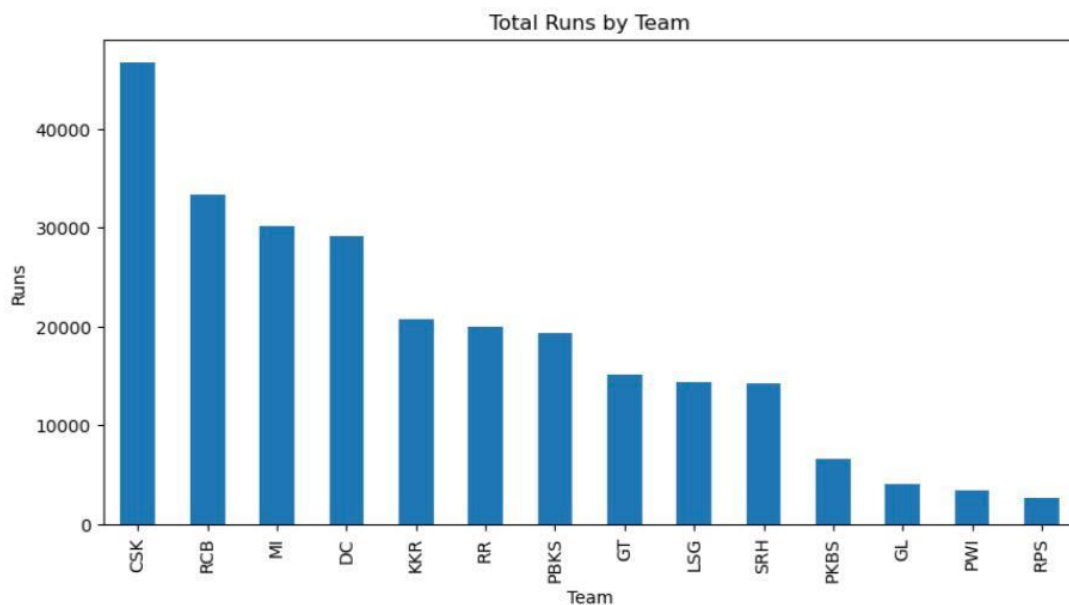
Observation:

Teams such as CSK, RCB, MI, and DC have scored the highest total runs, indicating consistent batting performance and long-term stability in team composition. On the other hand, newer teams and teams with fewer seasons have comparatively lower total runs due to limited match exposure.

KEY INSIGHTS

• Consistent performance across multiple seasons plays a major role in becoming a top run scorer in the IPL. Players with longer careers and regular match participation tend to accumulate higher run totals.

• Batting average and strike rate highlight different aspects of batting performance. While batting average reflects consistency and reliability, strike rate represents aggressive scoring ability. Both metrics are important in evaluating a batter's overall effectiveness.

• Players who maintain a balance between a good batting average and a high strike rate are the most valuable in T20 cricket, as they contribute both stability and quick scoring.

• Boundary-hitting ability has a strong influence on match outcomes. Batters with higher numbers of fours and sixes help maintain high run rates and shift match momentum in favor of their teams.

• Team-wise analysis shows that franchises with stable squads and experienced players tend to score higher total runs across seasons, indicating the importance of team continuity.

• Data visualization techniques such as bar charts and scatter plots make it easier to identify performance trends and compare players effectively.

FUTURE SCOPE

The scope of this project can be extended in several ways to enhance the depth and accuracy of analysis. Future work can include season-wise and match-wise analysis to study player performance trends over time. Additional factors such as opposition strength, match venue, and pitch conditions can be incorporated to provide more contextual insights.

Advanced analytics techniques such as machine learning models can be applied to predict player performance or identify emerging talent. Bowling performance and all-rounder analysis can also be included to build a more comprehensive player evaluation system. Furthermore, interactive dashboards using tools like Power BI or Tableau can be developed to present insights in a more dynamic and user-friendly manner.

Expanding the dataset and integrating real-time data can further improve the relevance and applicability of this project in professional sports analytics.

CONCLUSION

This project successfully analyzed batting performance in the Indian Premier League (IPL) using data analytics techniques and visualizations. By examining key performance metrics such as total runs, batting average, strike rate, and boundary counts, meaningful insights into player consistency and scoring behavior were obtained.

The analysis highlighted the importance of long-term consistency in achieving high run totals, as well as the role of aggressive batting in the T20 format. Visualizations made it easier to compare players, identify trends, and understand the balance between reliability and attacking intent. Team-wise analysis further demonstrated how experienced and stable teams tend to achieve higher cumulative batting performance.

Overall, this project demonstrates the practical application of data analytics in sports and shows how structured data analysis can support performance evaluation and informed decision-making in cricket.

REFERENCES

1.   IPL Official Website – https://www.iplt20.com

2.   ESPN Cricinfo – IPL Statistics and Player Records – https://www.espncricinfo.com

3.   Kaggle – IPL Batting Statistics Dataset – https://www.kaggle.com

4.   Cricbuzz – IPL Player Records – https://www.cricbuzz.com