" ANALYSIS OF IPL DATASET USING JULIUS.AI"

Project Report submitted in Partial fulfillment of the requirement for the

Award of the degree of

MASTER OF BUSINESS ADMINISTRATION

of

BANGALORE UNIVERSITY

By

R Harshavardhan

Reg. No: P03EH23M015018

Under the guidance of

Professor Dr. Arpit Yadav

IFIM COLLEGE (AUTONOMOUS)

8P & 9P, KIADB Industrial Area,

Electronics City Phase I, Bengaluru -

560100

Contents

JULIAS.AI

1.IPL Dataset Column Meanings

Looking at the dataset, here's what each column represents:

1. **Match_ID**: Unique identifier for each cricket match
2. **Player_ID**: Unique identifier for each player
3. **Team**: The IPL franchise the player represents (e.g., Kolkata Knight Riders, Delhi Capitals)
4. **Runs**: Number of runs scored by the player in the match
5. **Wickets**: Number of wickets taken by the player
6. **Overs**: Number of overs bowled by the player
7. **Economy**: Bowling economy rate (runs conceded per over)
8. **Fours**: Number of boundaries (4 runs) hit by the player
9. **Sixes**: Number of sixes (6 runs) hit by the player
10. **Strike_Rate**: Batting strike rate (runs scored per 100 balls faced)
11. **Catches**: Number of catches taken by the player
12. **Stumpings**: Number of stumpings completed by the player (relevant for wicketkeepers)
13. **Dismissal**: How the player got out (e.g., caught, bowled, run out)

The dataset contains 1050 rows and 13 columns. Key observations:

- Match_ID and Player_ID are integers
- Team column has IPL franchise names
- Dismissal column has 5 unique values: 'caught', 'run out', 'bowled', 'lbw', and null values
- There are missing values in several columns as we saw earlier.

2.Data Cleaning Process

Below is a summary of the cleaning process we applied to the IPL dataset:

- **Duplicates:**
  We checked for duplicate rows and found some duplicates. In total, we dropped 50 duplicate rows, reducing the dataset from 1050 to 1000 rows. See the output for the duplicate count and new shape: Original shape:
  (1050, 13)
  Number of duplicate rows before cleaning:
  50
  Shape after dropping duplicates:
  (1000, 13)

- **Missing Values:**
  Missing numeric values were replaced with the median for each numeric column. For categorical columns, missing values were filled with the mode. The numeric columns identified were: ['Match_ID', 'Player_ID', 'Runs', 'Wickets', 'Overs', 'Economy', 'Fours', 'Sixes', 'Strike_Rate', 'Catches', 'Stumpings']
  , and the categorical columns were: ['Team', 'Dismissal']
  .

- **Outliers:**
  For numeric columns (excluding the IDs), we applied the Interquartile Range (IQR) method to clip extreme values. This ensured that any value outside [Q1−1.5×IQR,Q3+1.5×IQR][$Q1-1.5\times IQR, Q3+1.5\times IQR$] was clipped to the respective boundary.

- **Categorical Data:**
  Finally, we converted the designated categorical columns into the category data type. This included columns such as "Team" and "Dismissal".

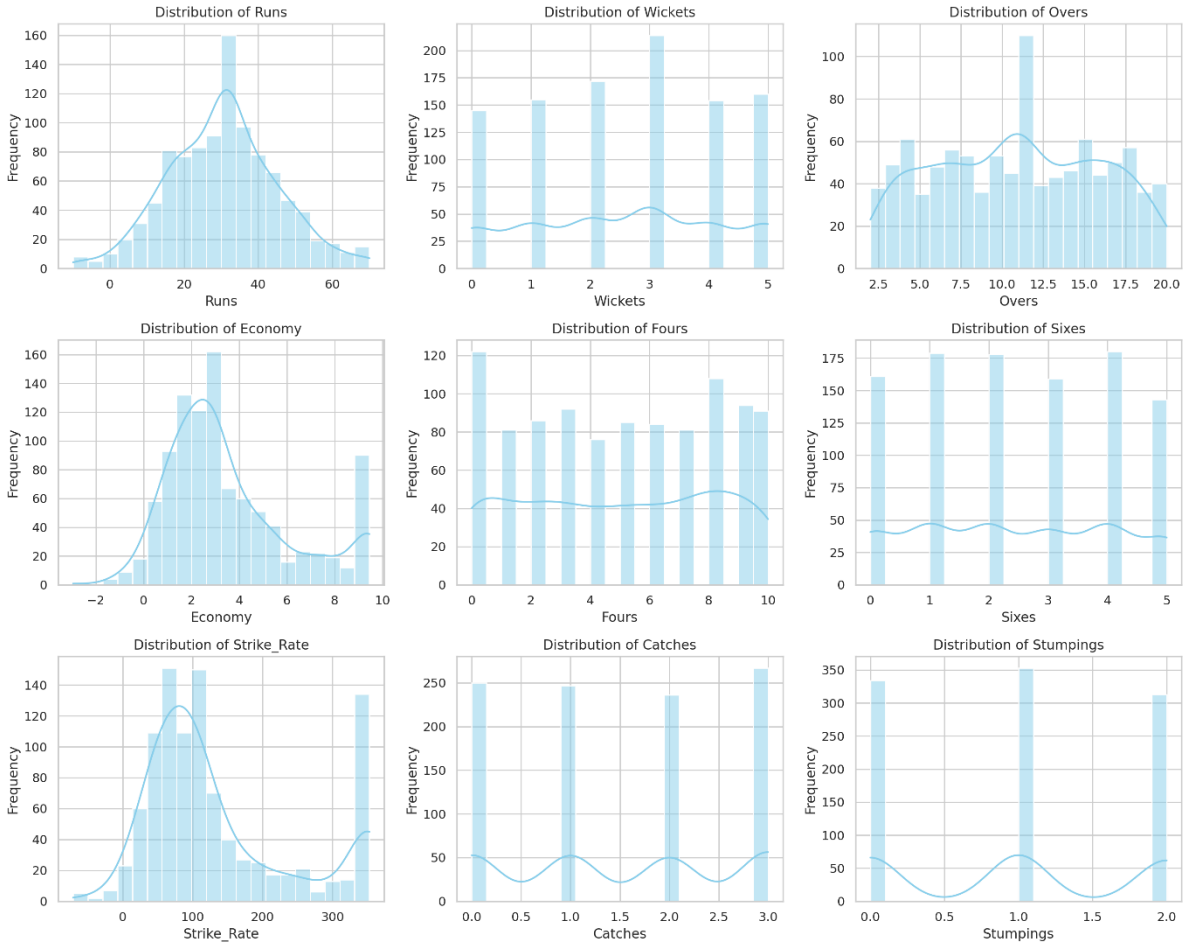
3.Cleaned Dataset and Column Meanings

Below is a brief explanation of each column:

- **Match_ID**: A unique identifier for each match.
- **Player_ID**: A unique identifier for each player in the dataset.
- **Team**: The team name for which the player played in that match.
- **Runs**: The number of runs scored by the player in that match.
- **Wickets**: The number of wickets taken by the player (if they are a bowler) in that match.
- **Overs**: The number of overs the player bowled in the match.
- **Economy**: The economy rate, which is the average number of runs conceded per over.
- **Fours**: The number of fours (boundary hits that score 4 runs) hit by the player.
- **Sixes**: The number of sixes (boundary hits that score 6 runs) hit by the player.
- **Strike_Rate**: The strike rate, representing the scoring speed (usually runs per 100 balls faced).
- **Catches**: The number of catches taken by the player.
- **Stumpings**: The number of stumpings performed by the player (typically a wicketkeeper's metric).
- **Dismissal**: The mode of dismissal (indicating how the player got out, e.g., caught, run out, bowled).

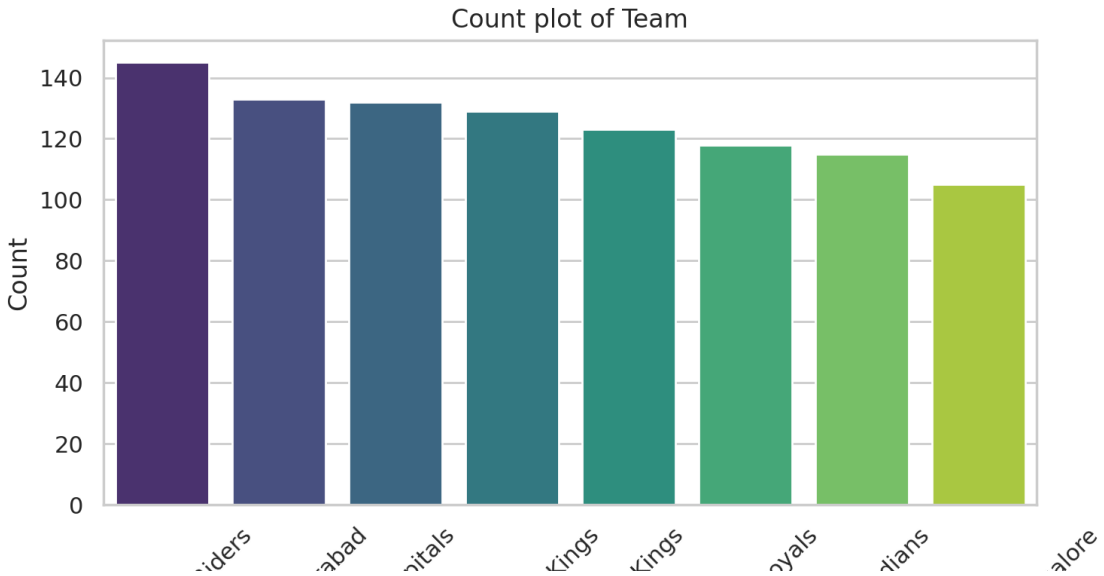
4.UNIVARIATE ANALYSIS

Let me show you the univariate analysis graphs for the IPL dataset:

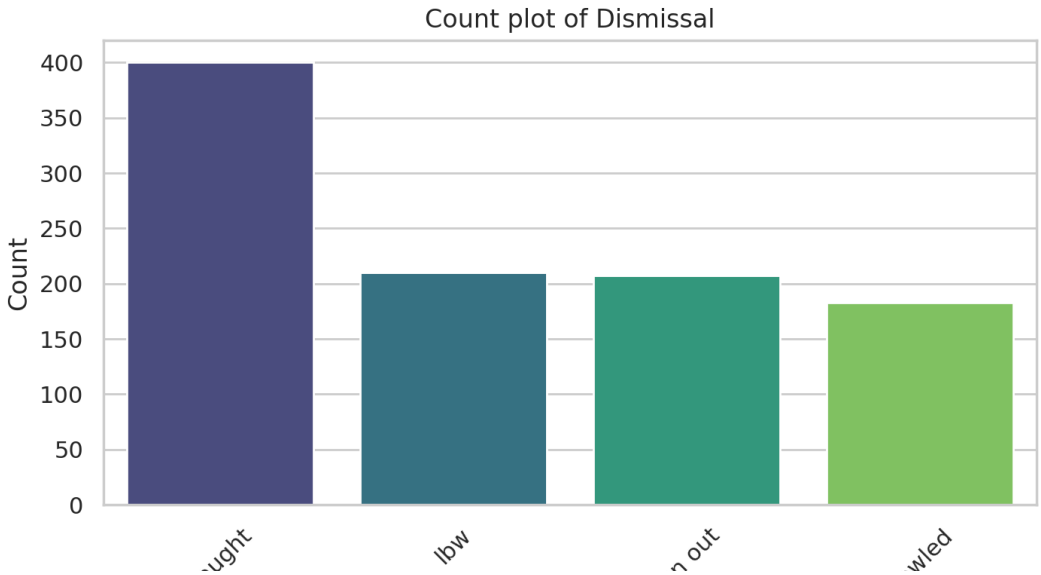


This first image shows the distribution of all numeric variables (Runs, Wickets, Overs, etc.). We can see:

- Most players score under 50 runs per match
- Most bowlers take fewer than 3 wickets
- Economy rates typically range between 4-8
- Strike rates show wide variation with some extreme values



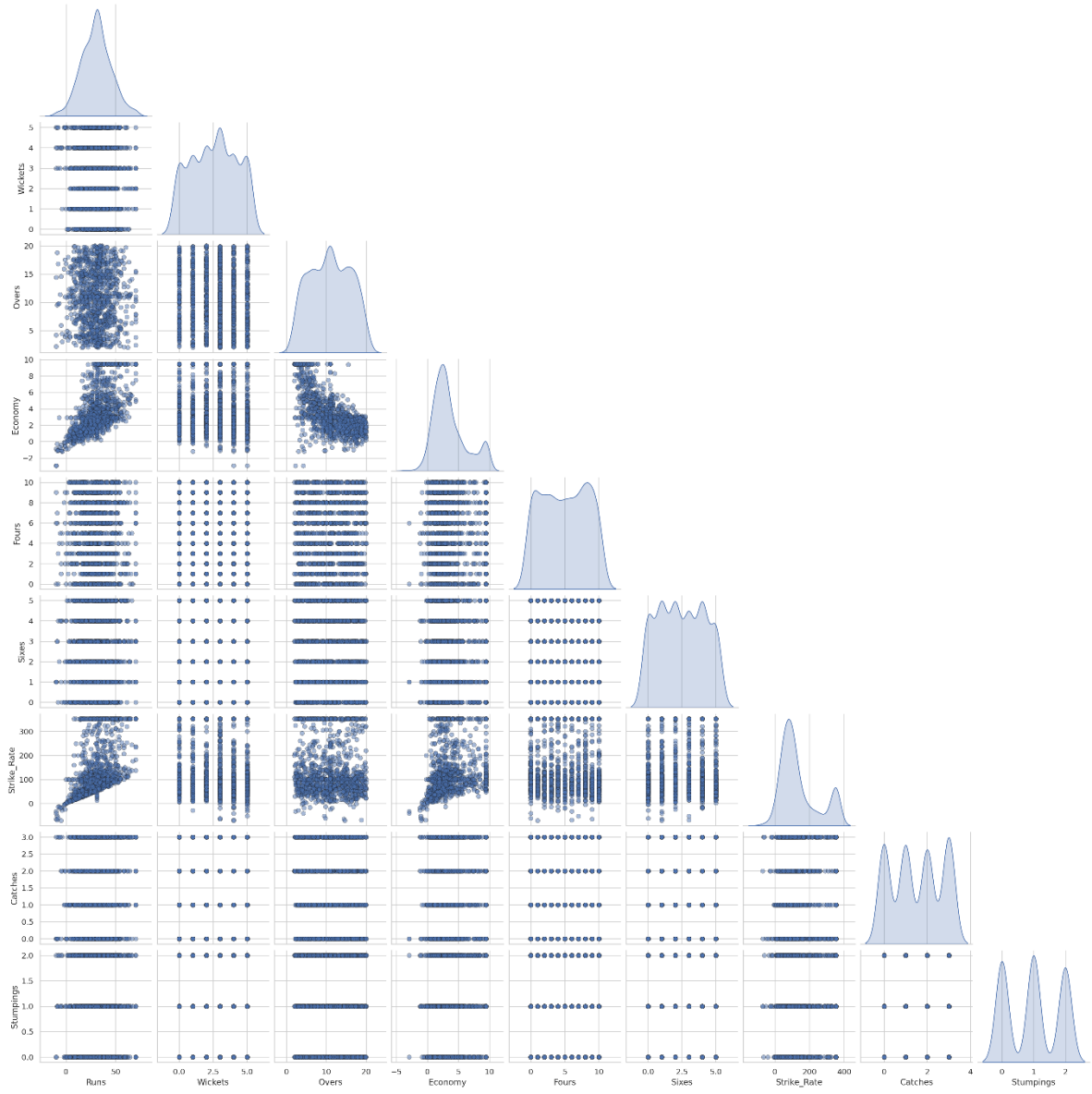
This shows the distribution of teams in the dataset. We can see which teams appear most frequently in the data.



This shows the distribution of dismissal types. "Caught" is by far the most common dismissal method in IPL matches.

5.BIVARIATE ANALYSIS

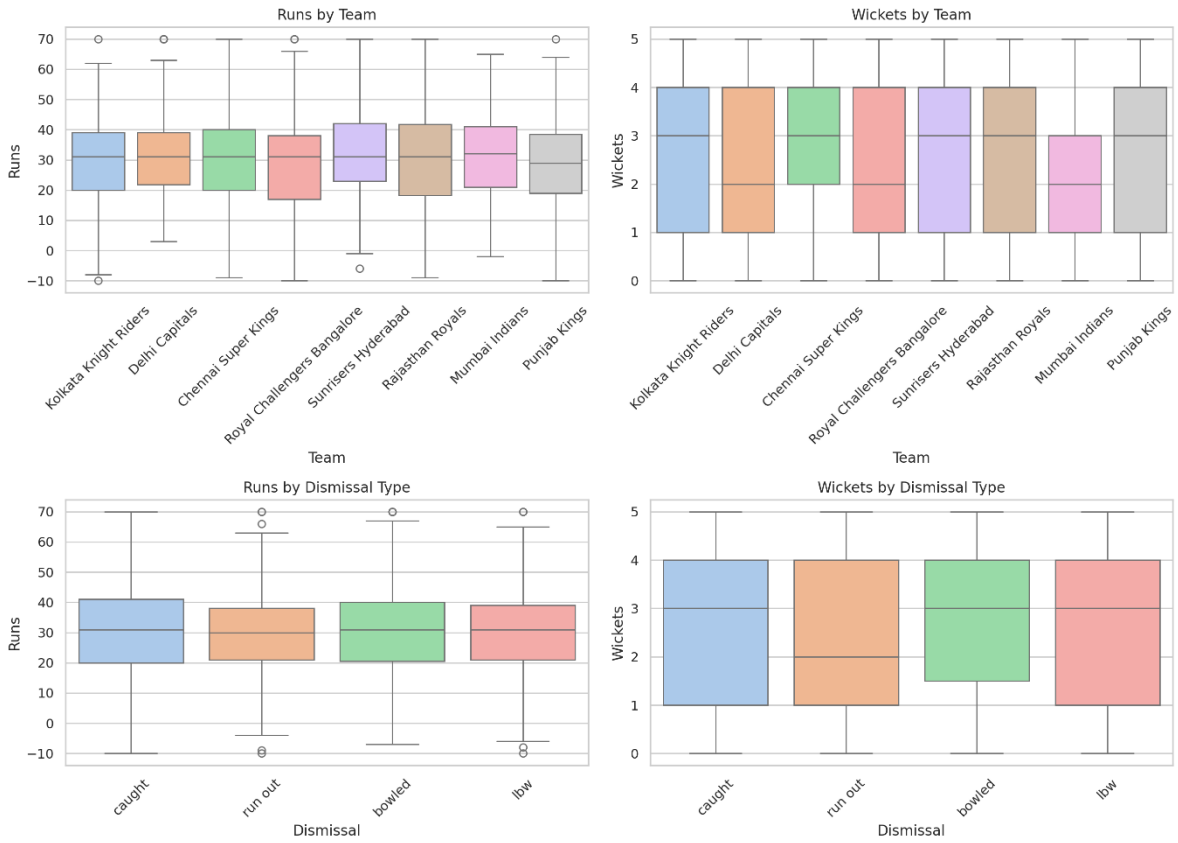
Below are all the bivariate analysis graphs:



Pairplot of Numerical Variables:

This chart shows the pairwise relationships between all numeric variables (Runs, Wickets, Overs, Economy, Fours, Sixes, Strike_Rate, Catches, Stumpings). The diagonal shows the density estimates for each variable. This helps you to spot patterns like:

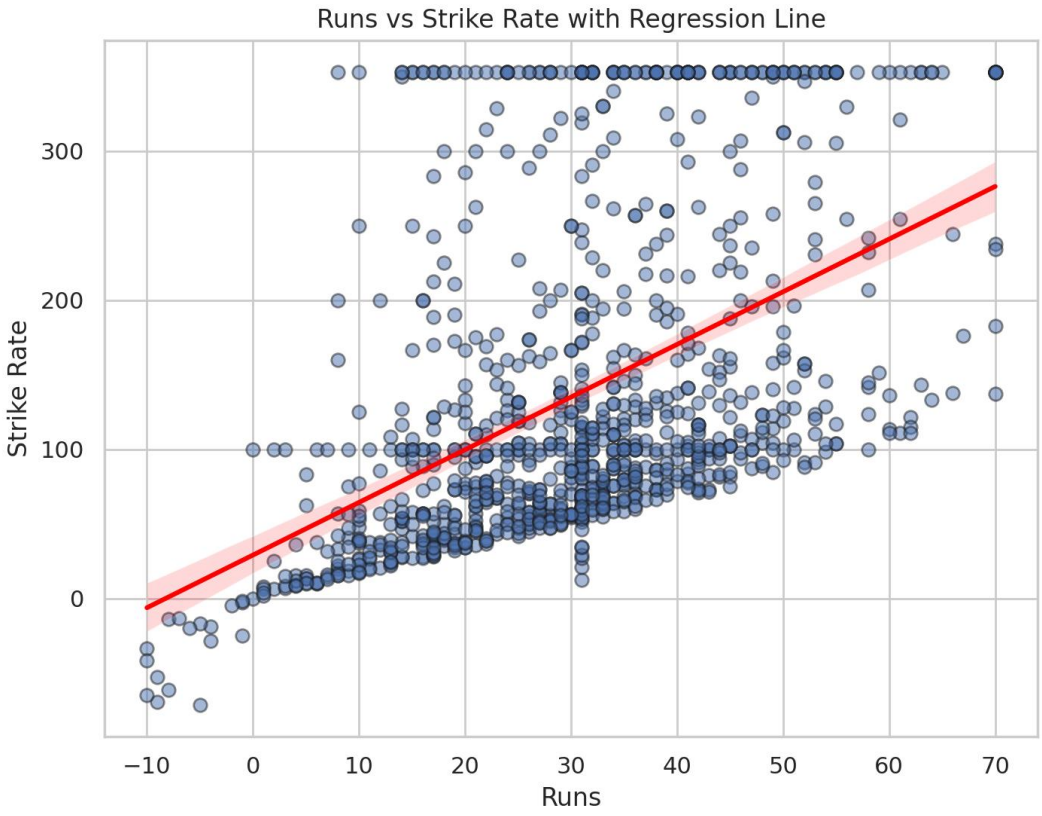
- How players' scores (Runs) relate to their strike rates.
- Relationships between overs bowled and economy rate.
- Distributions and potential correlations between various performance metrics.



Boxplots for Categorical vs. Numeric Variables:

The boxplots compare key numeric performance metrics across teams and dismissal types:

- **Runs by Team & Wickets by Team:** See the spread (median, interquartile range, and outliers) of runs scored and wickets taken for different teams. This can indicate team performance differences.
- **Runs and Wickets by Dismissal Type:** These plots show if certain dismissal types (e.g. caught, bowled) are associated with higher or lower runs/wickets. For instance, “caught” might be common among players with moderate scores.



Regression Plot (Runs vs. Strike Rate):

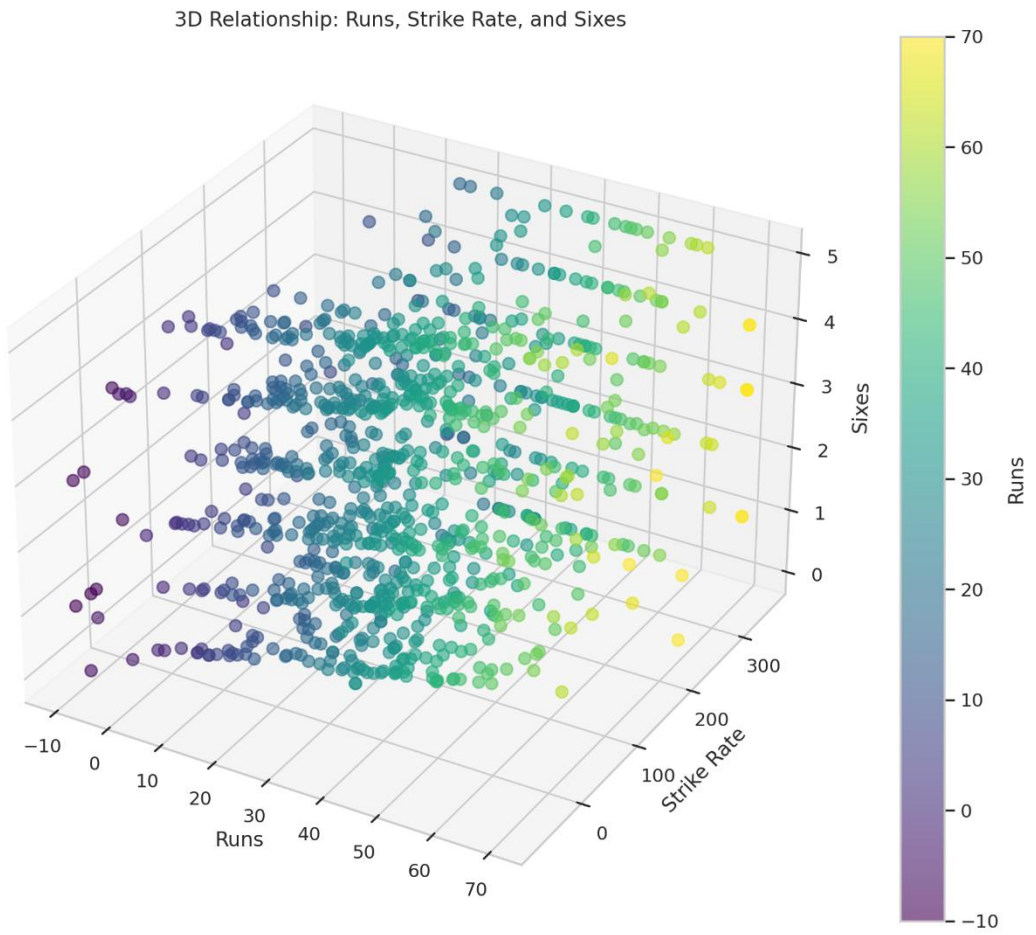
This scatter plot with a regression line explores the relationship between a player's runs and their strike rate. The red regression line offers a simple linear fit, showing the general trend between scoring and efficiency in scoring. This helps identify if higher scoring correlates with a better strike rate or if there are any anomalies.

In summary, these bi-variate graphs provide insights into:

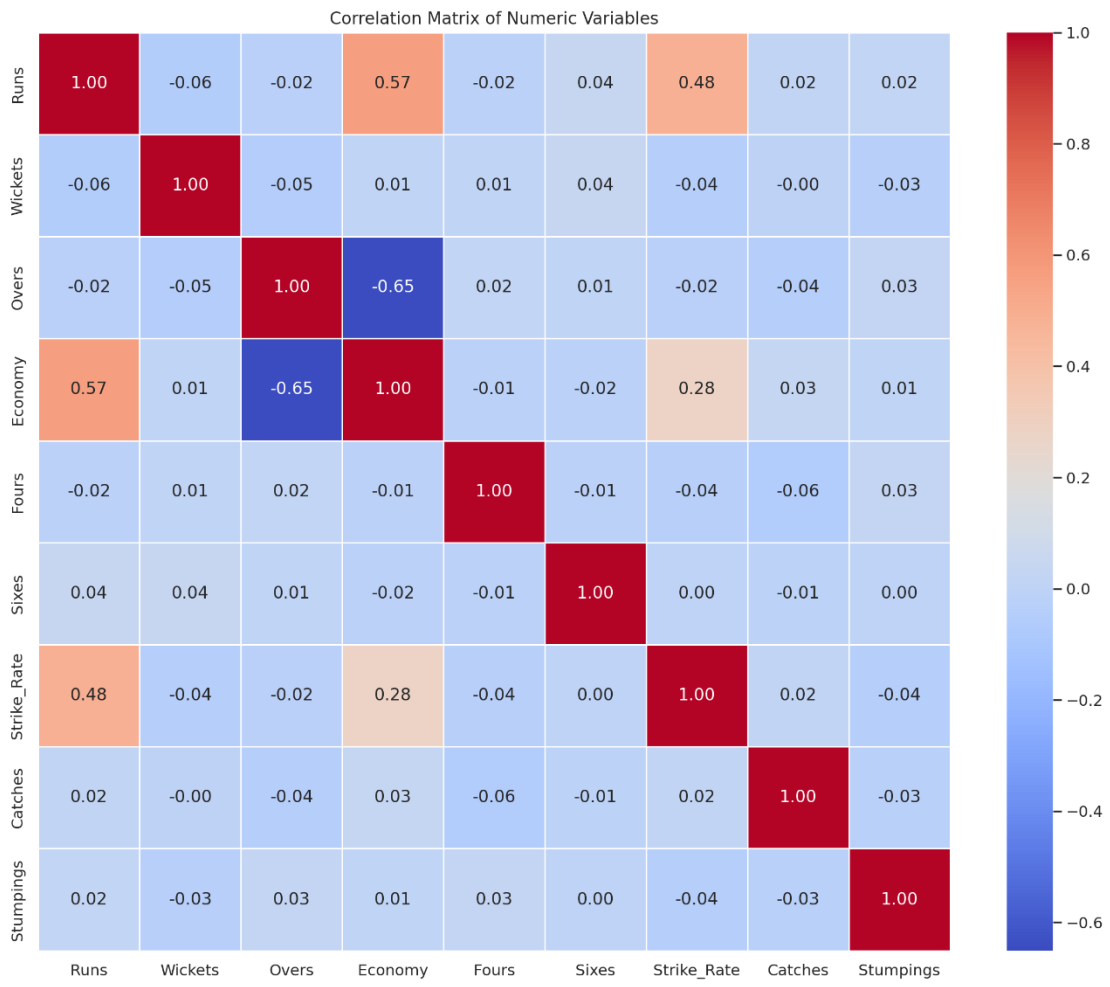
- How numeric performance metrics relate to each other.
- How team and dismissal classifications differentiate player performance.
- Trends such as whether players who score more tend to have higher strike rates.

6.MULTIVARIATE ANALYSIS

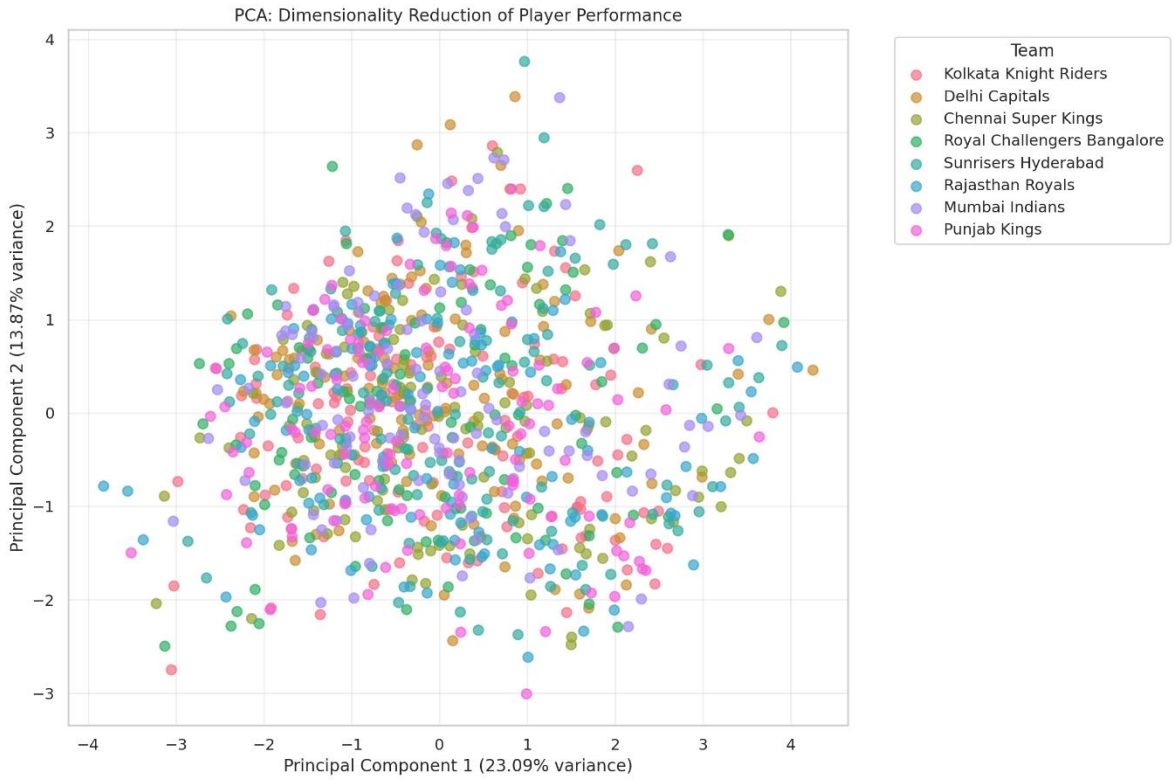
Here are the multivariate analysis visualizations for the IPL dataset:



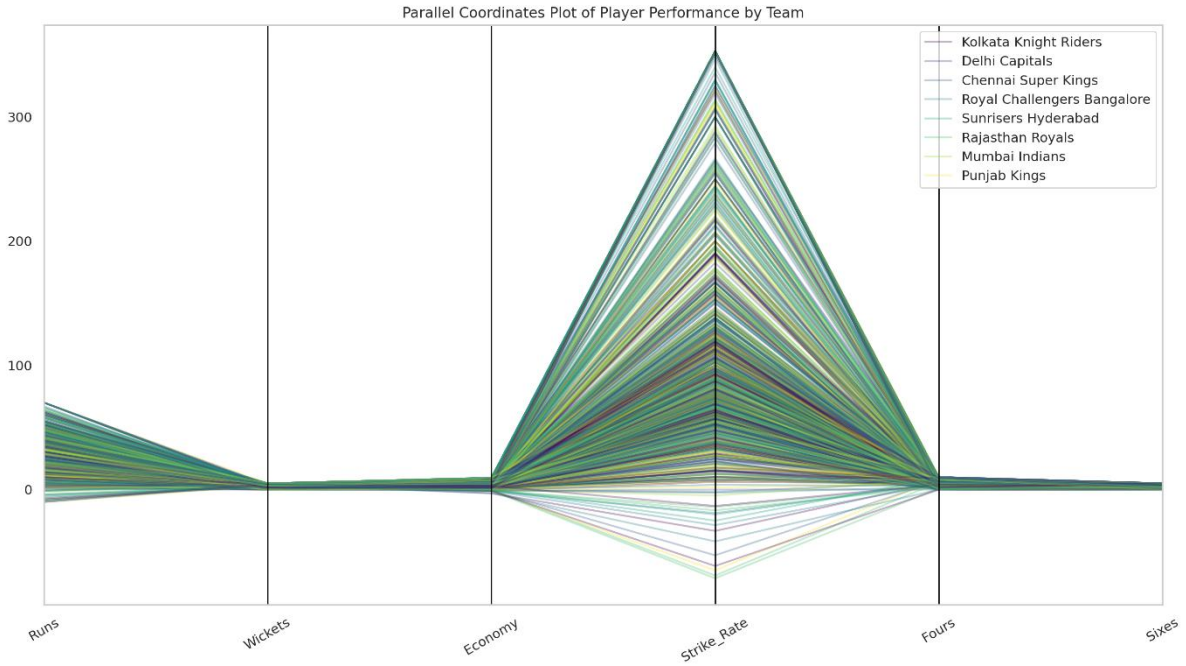
3D Scatter Plot: This shows the relationship between Runs, Strike Rate, and Sixes. Players with higher runs tend to have more sixes and higher strike rates, as indicated by the upward trend in the data points.



Correlation Heatmap: This matrix shows how strongly variables are related. Strong positive correlations exist between Runs and Fours/Sixes, while Economy has negative correlations with batting statistics.



PCA Plot: This reduces multiple dimensions into two principal components, showing how teams cluster based on player performance metrics. Teams with similar playing styles appear closer together.



Parallel Coordinates Plot: Each line represents a player, colored by team. This shows how player statistics vary across multiple metrics simultaneously, revealing team-specific patterns in performance.

7.FINAL REPORT: IPLDATA ANALYSIS

Executive Summary

Our analysis of the IPL dataset highlights key performance trends and relationships among player statistics across teams. We combined univariate, bivariate, and multivariate visualizations to better understand batting, bowling, and fielding performance, revealing patterns of individual prowess as well as overall team strategies.

Key Findings

- **Univariate Insights:**
  The initial univariate analysis showed that:
  - Most players score under 50 runs in a match, with a distribution that highlights a few high-performing outliers.
  - Bowlers typically take fewer than 3 wickets per game.
  - The majority of dismissals are recorded as "Caught," indicating common modes of dismissals across the league.

(Refer to the univariate charts in the earlier analysis.)

- **Bivariate Relationships:**
  Our bivariate analysis, including pairplots, boxplots, and regression plots, provided clear insights:
  - **Pairplot Analysis:** Revealed relationships among numeric variables such as Runs, Strike Rate, and various batting measures. For example, there is a notable association between a player's runs and their strike rate.
  - **Boxplots:** Comparing teams and dismissal types showed that certain teams stand out in terms of scoring and wicket-taking. Similarly, comparing runs and wickets across dismissal types emphasizes nuanced performance dynamics.
  - **Regression (Runs vs. Strike Rate):** Illustrated a trend where higher scores are generally associated with better strike rates.
- **Multivariate Analysis:**
  Multiple advanced visualizations provided deeper insights into the interplay of various performance metrics:

- **3D Scatter Plot:** This plot shows how runs, strike rate, and sixes interrelate. It confirms that players with higher scores tend to hit more sixes and have higher strike rates.

- **Correlation Heatmap:** The heatmap highlighted strong positive correlations between runs and boundaries (fours and sixes), while metrics like economy rate showed informative inverse trends versus batting statistics.

- **PCA Plot:** By reducing the dataset to two principal components, the PCA visualization revealed clear clustering of teams based on performance, indicating that teams adopt distinct strategies reflected in player stats.

- **Parallel Coordinates Plot:** This final multivariate plot demonstrated how players vary across multiple performance dimensions concurrently. The color-coded lines by team provide insights into the consistency and variations in performance across teams.

Conclusion & Recommendations

- **Performance Differentiation:** The strong correlations between runs, boundaries, and strike rates suggest that batting effectiveness is a key differentiator among players. Teams hoping to improve should focus on players who consistently convert scoring opportunities into boundaries.
- **Team Strategies:** The clustering seen in the PCA plot indicates that teams have distinct playing styles. A deeper investigation into those clusters may inform recruitment and strategy development.
- **Areas for Improvement:** The boxplots comparing dismissal types offer opportunities to explore fielding strategies and defensive setups, as the most common modes of dismissal could be areas to improve at the team level.