DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

Tittle: IPL Insights: Data-Driven Analysis & Prediction

Submitted by

Harshith Sakhamuri
Registration No - 12326473
Programme and Section – B.Tech. & K23DW
Course Code – INT375

Under the Guidance of

Vikas Mangotra

U.ID – 31488
Faculty of CSE
Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Harshith Sakhamuri bearing Registration No. 12326473 has completed the INT375 project titled “IPL Insights: Data-Driven Analysis & Prediction” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development and study.

Vikas Mangotra

Assistant Professor

Lovely Professional University

Phagwara, Punjab.

Date: 12-04-2025

DECLARATION

I, Harshith Sakhamuri, student of B.Tech. under the CSE discipline at Lovely Professional University, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025

Registration No: 12326473

Signature:

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my mentor, Mr. Vikas Mangotra, for his continued guidance, support, and insightful feedback throughout the project. This project would not have been possible without the faculty and staff of LPU, whose knowledge-sharing helped me build my analytical skills. I am also thankful to my friends and family for their constant encouragement and belief in me. Their motivation kept me focused and helped me complete this project successfully.

Harshith Sakhamuri

K23DW

12326473

TABLE OF CONTENTS

Introduction:

The Indian Premier League (IPL) is a globally celebrated T20 cricket tournament that generates extensive data across multiple seasons. With the massive amount of match and ball-by-ball level data available, there lies a tremendous opportunity to apply data science techniques to derive strategic insights. This project harnesses the power of Python programming and its data science libraries including pandas, seaborn, matplotlib, and scikit-learn to explore the IPL data.

The aim of this project is to clean, process, analyze, visualize, and build predictive models on IPL data ranging from 2008 to 2022. Through this initiative, we uncover hidden performance patterns, understand game strategies such as toss decisions, analyze individual player behaviors, and also develop a predictive model to forecast match outcomes using machine learning.

Source of the Dataset:

The datasets used in this project were sourced from publicly available IPL archives and cleaned versions available on Kaggle(with liscense). The primary files include:

1. “IPL_Matches_2008_2022.csv”– This contains match-level information such as teams, venue, toss, match results, and player of the match.
2. “IPL_Ball_by_Ball_2008_2022.csv” – This file provides granular

delivery-level data like batsman, bowler, runs scored, dismissals, over numbers, and extras.

These datasets serve as the foundation for performing statistical analysis, creating dashboards, and training machine learning models.

Link of datasets:- https://www.kaggle.com/datasets/vora1011/ipl-2008-to-2021-all-match-dataset

Exploratory Data Analysis(EDA) and Cleaning:

Data cleaning and preprocessing is an essential phase in any data science pipeline, especially in sports analytics where accurate alignment between datasets is critical. The cleaning process included the following major steps:

- Missing Value Treatment: Many entries such as batter, bowler, toss winner, and match dates were either missing or malformed. These were fixed using forward fill, logical inference, or removed if deemed unusable.

- Standardization: Teams and venues appeared under different names over seasons. For example, "Delhi Daredevils" was renamed to "Delhi Capitals". All team, player, and venue names were standardized for consistency.

- Column Formatting: Date columns were converted into datetime formats for time-based operations. Numerical fields such as 'runs', 'balls', and 'match ID' were cast to proper types to ensure arithmetic operations worked as expected.

- Duplicates and Logical Validation: Duplicate deliveries and mismatches across ball and match datasets were identified and removed or corrected.

After cleaning, two reliable datasets were generated: `cleaned_match_level.csv` and `cleaned_ball_by_ball.csv`, which served as inputs for the rest of the project.

Objective 1: Cleaning and Standardizing IPL Datasets

This objective focuses on preparing the dataset for accurate analysis. Data cleaning ensures that null values, inconsistent labels, and invalid data entries do not affect the final outcome. The IPL datasets had missing match numbers, dates, team names, and inconsistent player names that were fixed using logic such as forward filling, interpolation, or mapping.

Furthermore, columns like 'innings', 'overs', and 'ballnumber' were recalculated where necessary, and categorical columns were cleaned for whitespace and capitalization issues. Team and player names were standardized, and duplicates were removed to preserve data quality.

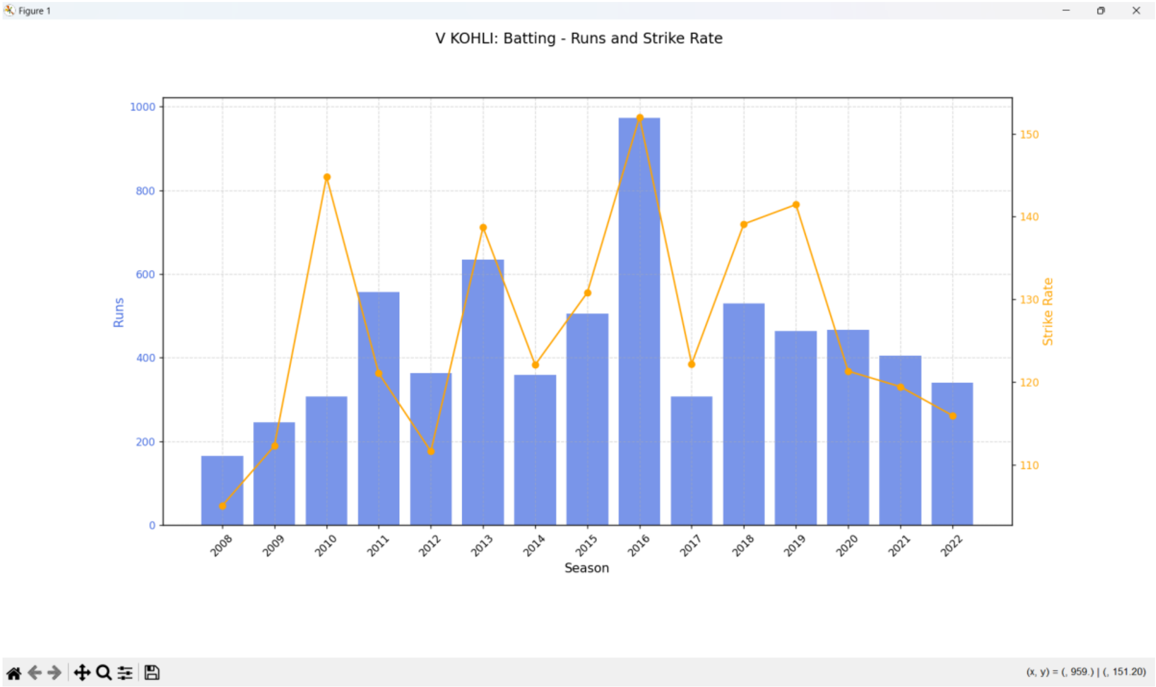
By completing this stage, the datasets became clean, consistent, and ready for further exploration and modeling.

Objective 2: Player Performance Dashboard

This part of the project analyzes individual player performance over the years using visual dashboards and structured summaries. The data was grouped by player and season, allowing us to track a player's form across different seasons.

Statistics such as total runs, balls faced, boundaries hit, strike rate, average, and dismissals were calculated. Similarly, bowling metrics included wickets, economy rate, bowling average, and strike rate. Fielding performance in terms of catches and player-of-the-match awards was also recorded.

All of this information was compiled into a consolidated CSV file. Additionally, graphs and charts provided insights on performance patterns, highlighting consistency or improvement over time. This dashboard is crucial for player comparison and auction strategies.



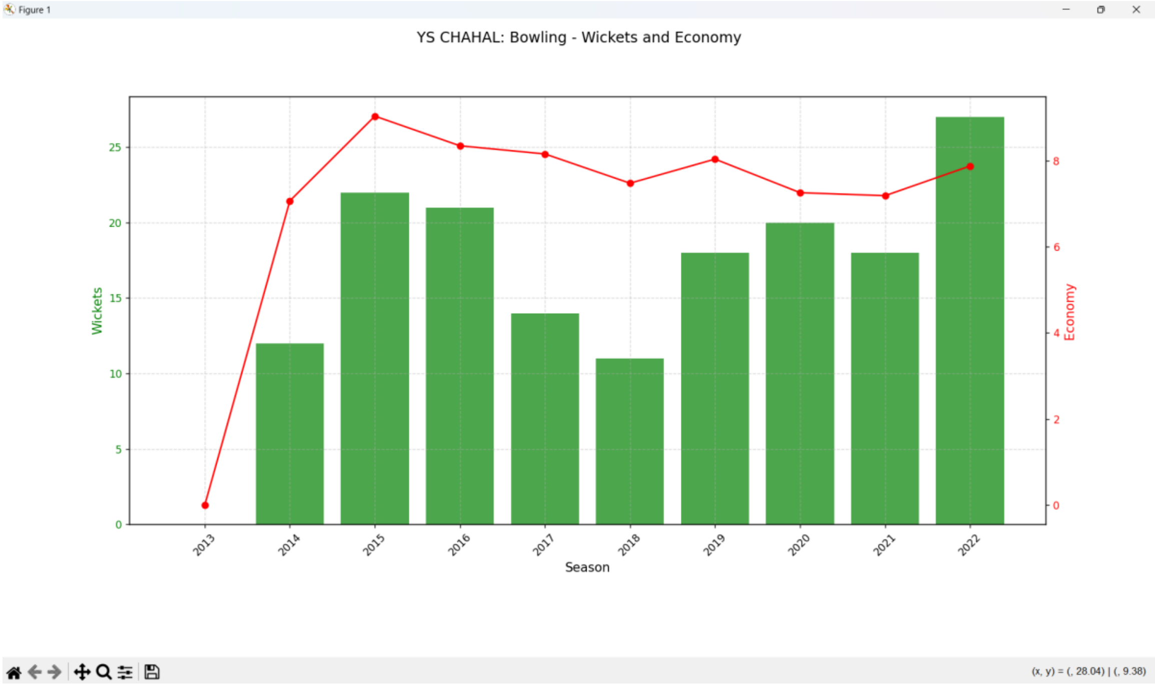


Figure 1
YS CHAHAL: Bowling - Wickets and Economy
Wickets
0
5
10
15
20
25
Economy
0
2
4
6
8
Season
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
(x, y) = (, 28.04) | (, 9.38)

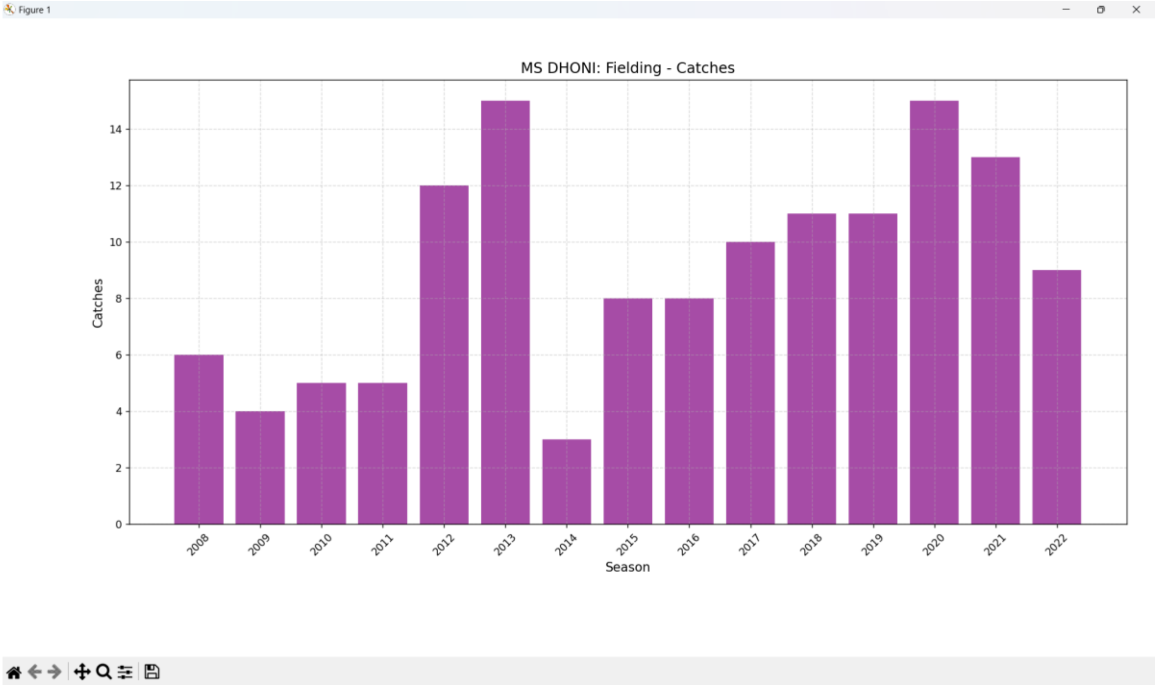


Figure 1
MS DHONI: Fielding - Catches
Catches
0
2
4
6
8
10
12
14
Season
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022

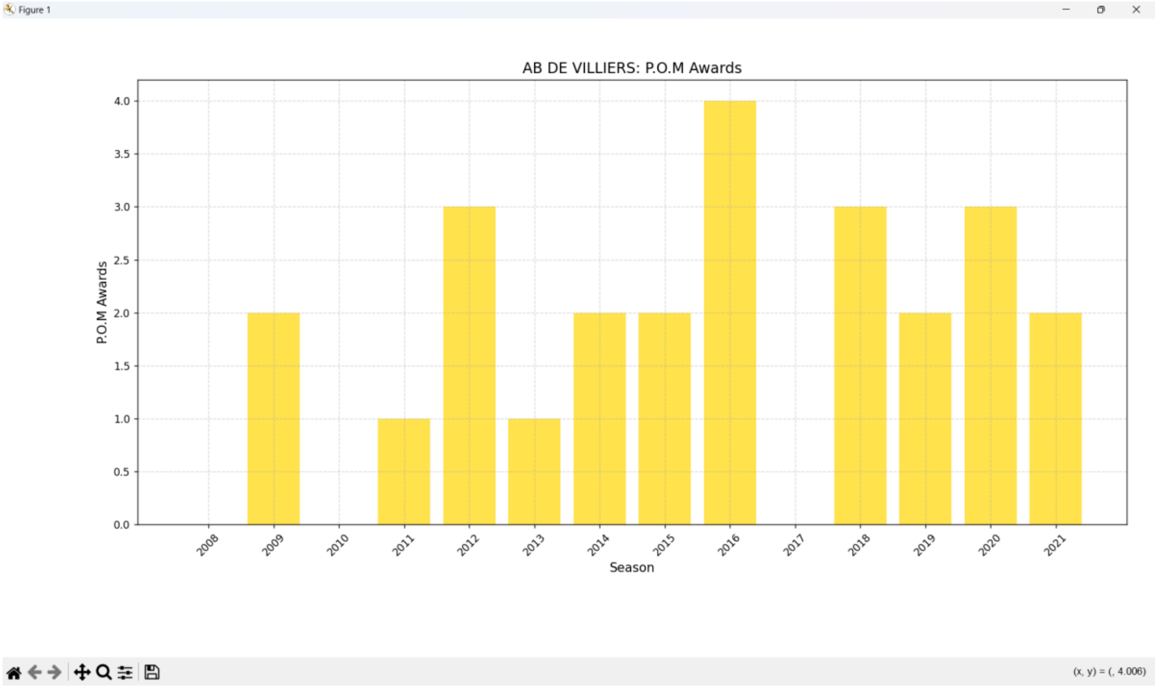


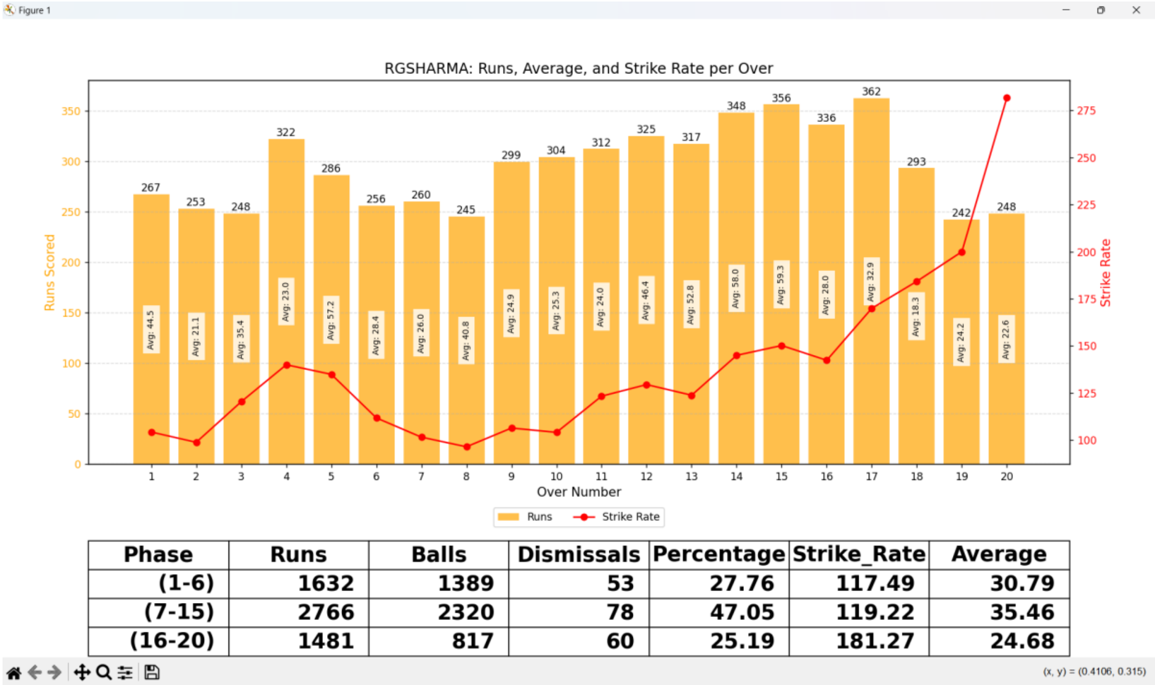
Figure 1
AB DE VILLIERS: P.O.M Awards
P.O.M Awards
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Season
(x, y) = (, 4.006)

Objective 3: Analyzing Player Performance Across Overs

One of the unique insights this project explores is how a player performs across different overs in an innings. The ball-by-ball data enabled us to calculate runs, balls, dismissals, strike rate, and average for each over (1 to 20) of a player’s career.

This was particularly useful in identifying powerplay hitters, middle-over accumulators, and death-over finishers. For instance, some players scored rapidly in overs 16–20 but had lower performance in the powerplay. Others maintained consistent averages in the middle overs, showing resilience under pressure.

This analysis is vital for game planning and player positioning within an innings. Bar charts and over-wise plots visually depicted the player’s trend throughout the 20 overs of an innings.



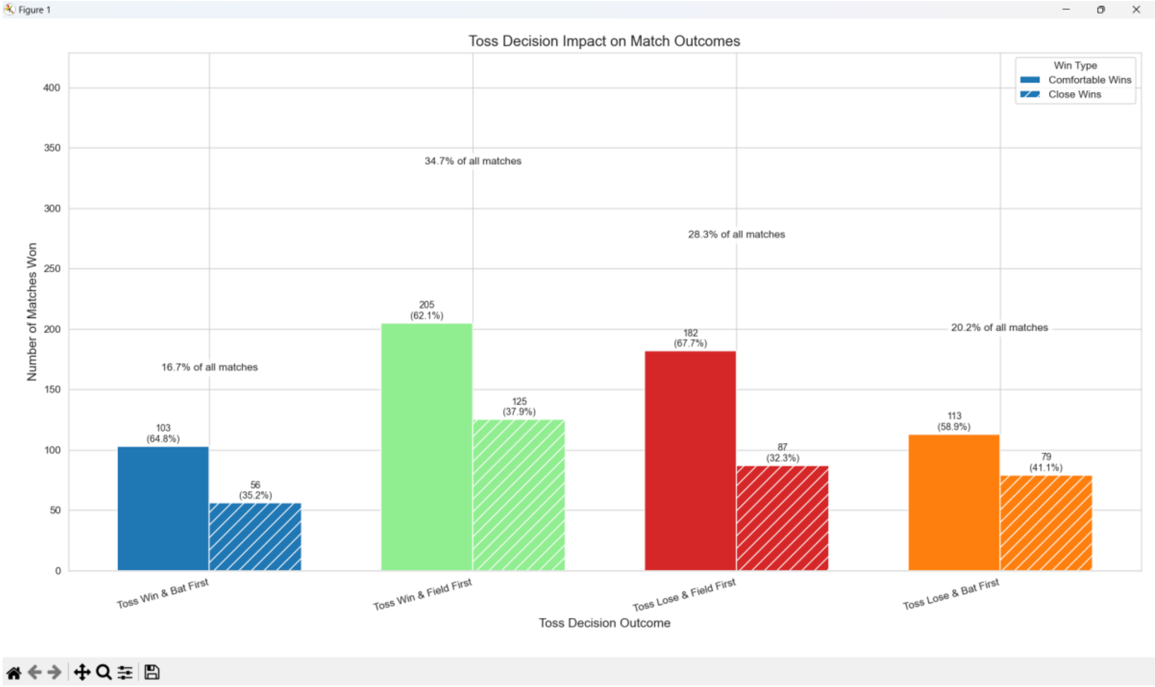
Phase	Runs	Balls	Dismissals	Percentage	Strike_Rate	Average
(1-6)	1632	1389	53	27.76	117.49	30.79
(7-15)	2766	2320	78	47.05	119.22	35.46
(16-20)	1481	817	60	25.19	181.27	24.68

Objective 4: Impact of Toss and Decision on Match Results

This objective analyzed how toss outcomes and decisions influence match results. The dataset was divided into different categories: teams that won the toss and chose to bat, teams that won and chose to field, and their opposites.

The nature of wins was further classified as ‘close’ (margin <15 runs or chased till the 19th over) or ‘comfortable’. Statistical comparisons revealed that fielding first after winning the toss often led to more convincing victories, possibly due to dew conditions or pitch behavior in the second innings.

Bar plots compared these categories, and a chi-square test with a heatmap was used to examine the strength of association between toss decisions and match outcomes. This analysis helps teams decide whether to bat or field after winning the toss under specific conditions.



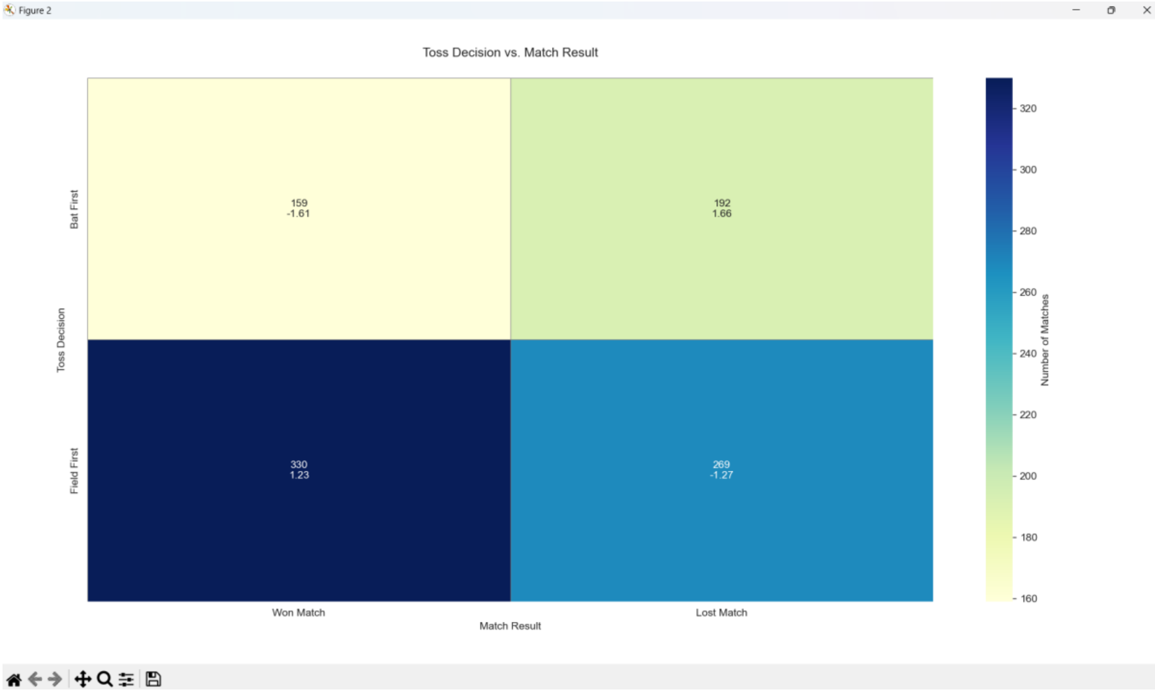


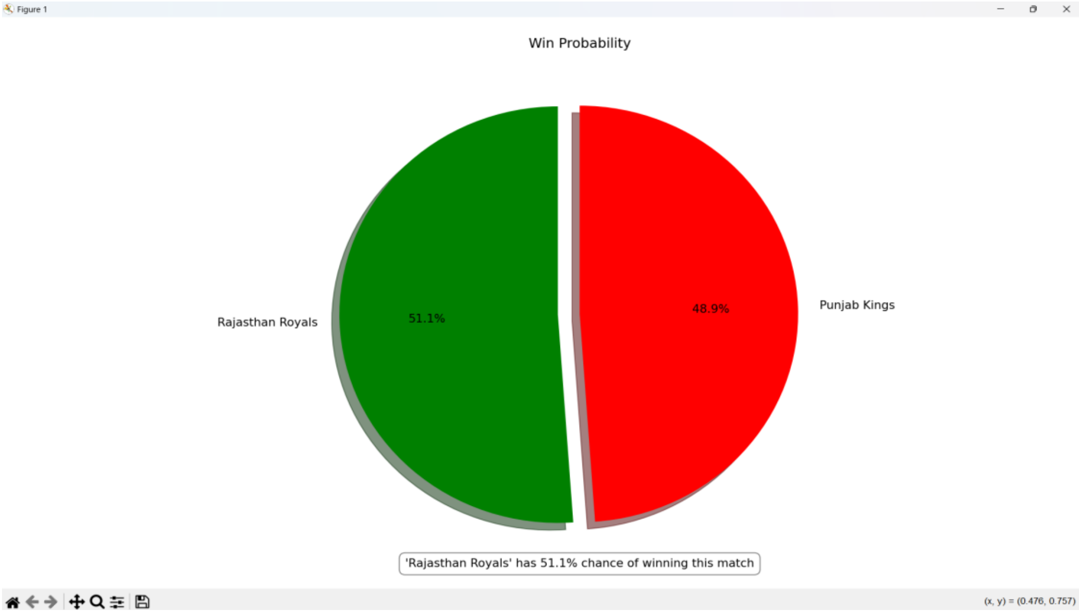
Figure 2
Toss Decision vs. Match Result
Bat First
Field First
Toss Decision
159
-1.61
192
1.66
330
1.23
269
-1.27
Won Match
Lost Match
Match Result
Number of Matches
320
300
280
260
240
220
200
180
160

Objective 5: Match Outcome Prediction Using Machine Learning

The final objective involved building a predictive model using machine learning to forecast match outcomes. A wide range of features were engineered, including:

- Team win percentage and recent form
- Head-to-head win ratios
- Venue impact and home advantage
- Batting and bowling averages for top players
- Innings average scores and economy rates

Using these features, three models were trained: Logistic Regression, Random Forest, and Support Vector Machines. The models were evaluated using accuracy scores and cross-validation.
Among them, the Random Forest model achieved the highest accuracy and was used for prediction. A user-friendly function was created to input two team names and get a predicted winner based on historical trends and match conditions. This shows the practical power of data science in sports forecasting.



'Rajasthan Royals' has 51.1% chance of winning this match

Conclusion:

This project successfully explored and analyzed historical IPL match and ball-by-ball data using a combination of data preprocessing, visualization, statistical insights, and machine learning techniques.

The data cleaning and preprocessing phase ensured the integrity and usability of the dataset by fixing missing values, correcting column formats, standardizing names, and removing duplicates. This robust foundation allowed for accurate and meaningful analysis across various dimensions.

Several key objectives were addressed:

- The cleaning and standardization scripts ensured both match-level and delivery-level datasets were logically consistent.
- A player dashboard was built to visualize seasonal and phase-wise performances.
- Over-wise performance analysis revealed player strengths and weaknesses during each over.
- Toss decision analysis showed a noticeable impact on match outcomes.
- A machine learning model was trained to predict match results using engineered features.

Overall Insights:

- Proper data cleaning significantly enhances the reliability of cricket analytics.
- Player performance varies not just by match but by over and match phase.
- Toss decisions can influence match outcomes, especially in close contests.
- Random Forest model performed well in predicting match outcomes.
- Combining cricket domain knowledge with data science tools leads to powerful insights.