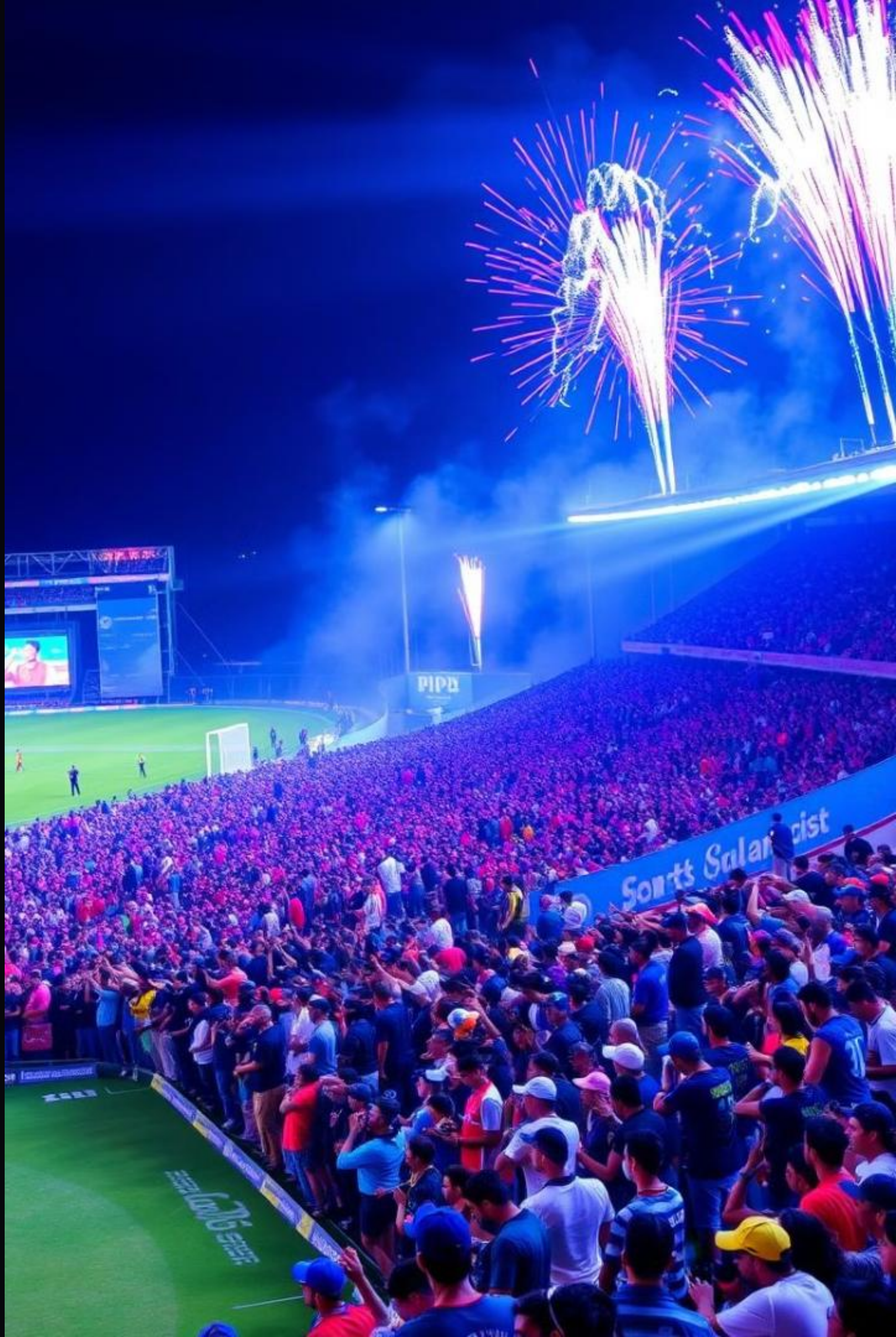
IPL Data Analysis & 2025 Winner Prediction

This presentation outlines our hackathon project. We analyze IPL data from 2008-2024. Our goal is to predict the 2025 IPL winner. We will cover data collection, EDA, and ML models. Join us as we explore cricket data!



Data Collection & Sources

Public APIs

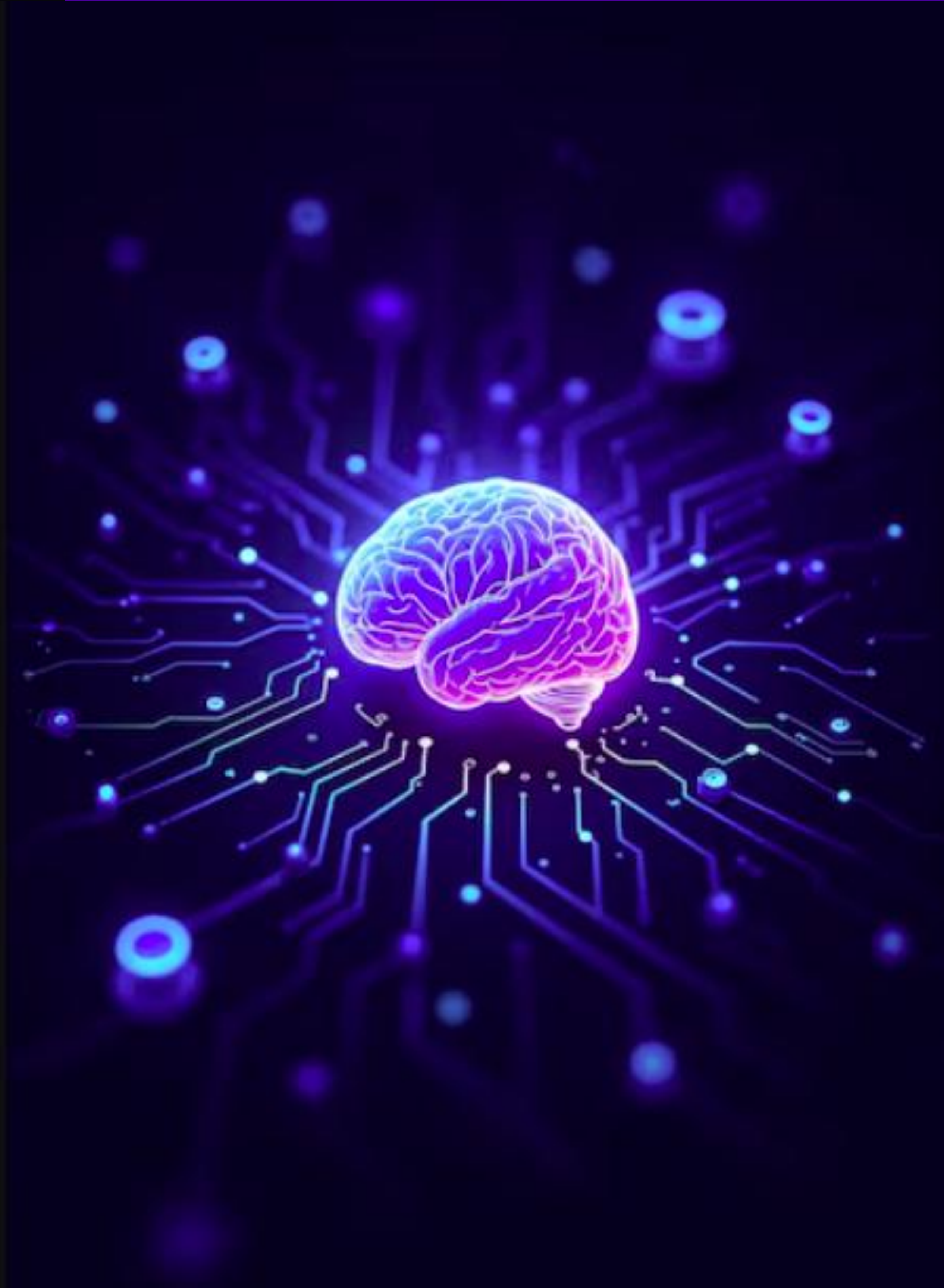
Cricsheet provide key match data. APIs are a direct and efficient way to get organized data.

Kaggle Datasets

Historical IPL datasets are available on Kaggle. These include match results, deliveries, and player statistics. These datasets offer pre-packaged historical information.

Web Scraping

Tools like Beautiful Soup and Scrapy help extract data. We can scrape websites for data not available via APIs. This approach allows access to more diverse sources.



Data Preprocessing & Cleaning

1 **Handling Missing Values**

- Replaced numerical missing values with column mean.
- Replaced categorical missing values with column mode.
- Dropped irrelevant data columns.

2 **Data Type Conversion**

- Standardized dates, scores, margins, and victory types.
- Categorical columns have been intentionally kept categorical so that data analysis becomes meaningful.
- Long names have been abbreviated to shorter ones.

3 **Feature Engineering**

- Ananlyse the data using numpy and pandas methods
- Plotted the meaningful metric using matplotlib, seaborn, Plotly.
- Try to decode hidden patterns from the dataset.

EDA: Win Rates & Team Statistics

Overall Win Rates

- Chennai Super Kings: 58.98%
- Mumbai Indians: 58.21%

These teams have dominated IPL history.

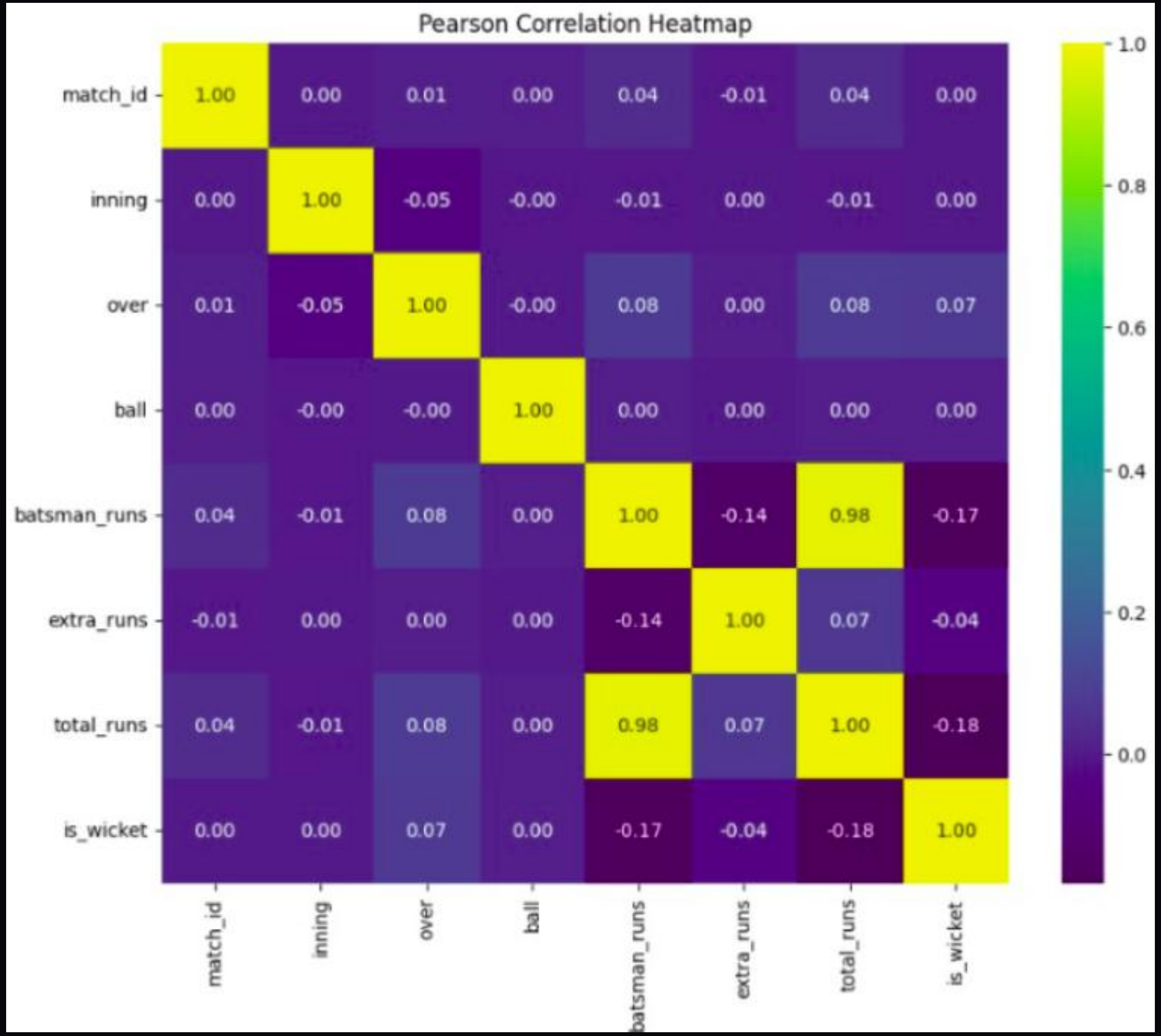
Home vs. Away

Home advantage contributed to 53% of wins.

Impact of Toss

Toss winners won about 52% of matches.

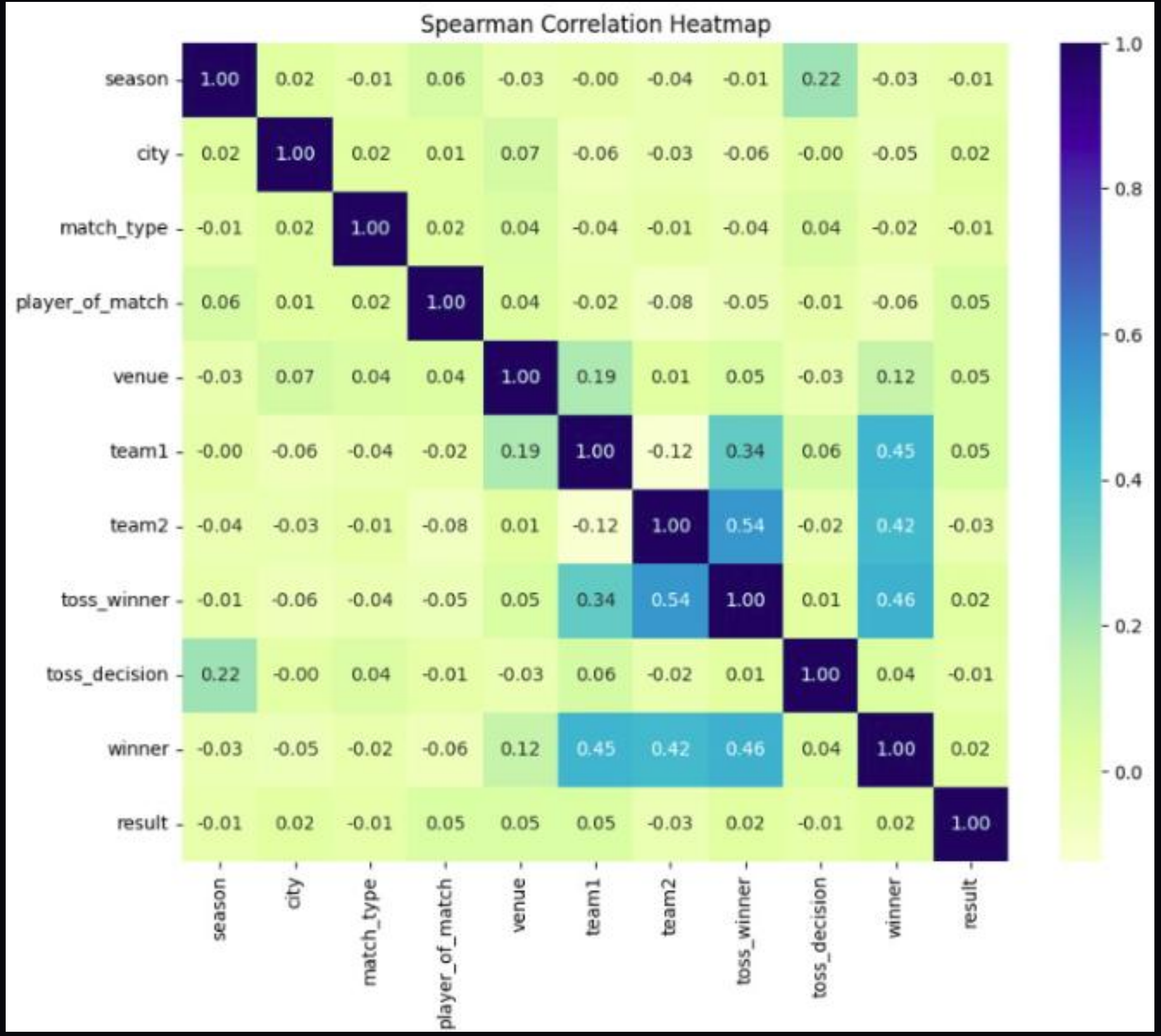
EDA Statistics


Pearson Correlation Heatmap

	match_id	inning	over	ball	batsman_runs	extra_runs	total_runs	is_wicket
match_id	1.00	0.00	0.01	0.00	0.04	-0.01	0.04	0.00
inning	0.00	1.00	-0.05	-0.00	-0.01	0.00	-0.01	0.00
over	0.01	-0.05	1.00	-0.00	0.08	0.00	0.08	0.07
ball	0.00	-0.00	-0.00	1.00	0.00	0.00	0.00	0.00
batsman_runs	0.04	-0.01	0.08	0.00	1.00	-0.14	0.98	-0.17
extra_runs	-0.01	0.00	0.00	0.00	-0.14	1.00	0.07	-0.04
total_runs	0.04	-0.01	0.08	0.00	0.98	0.07	1.00	-0.18
is_wicket	0.00	0.00	0.07	0.00	-0.17	-0.04	-0.18	1.00

Pearson for Deliveries

Using the correlation formula invented by Pearson, we aim to find how the numerical attributes of the datasets are linked to each other.

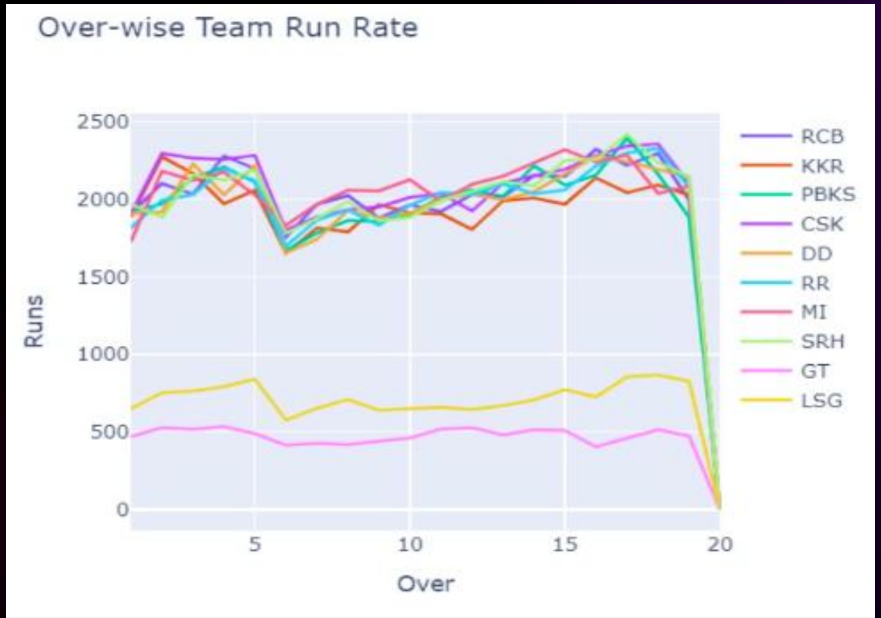

Spearman Correlation Heatmap

	season	city	match_type	player_of_match	venue	team1	team2	toss_winner	toss_decision	winner	result
season	1.00	0.02	-0.01	0.06	-0.03	-0.00	-0.04	-0.01	0.22	-0.03	-0.01
city	0.02	1.00	0.02	0.01	0.07	-0.06	-0.03	-0.06	-0.00	-0.05	0.02
match_type	-0.01	0.02	1.00	0.02	0.04	-0.04	-0.01	-0.04	0.04	-0.02	-0.01
player_of_match	0.06	0.01	0.02	1.00	0.04	-0.02	-0.08	-0.05	-0.01	-0.06	0.05
venue	-0.03	0.07	0.04	0.04	1.00	0.19	0.01	0.05	-0.03	0.12	0.05
team1	-0.00	-0.06	-0.04	-0.02	0.19	1.00	-0.12	0.34	0.06	0.45	0.05
team2	-0.04	-0.03	-0.01	-0.08	0.01	-0.12	1.00	0.54	-0.02	0.42	-0.03
toss_winner	-0.01	-0.06	-0.04	-0.05	0.05	0.34	0.54	1.00	0.01	0.46	0.02
toss_decision	0.22	-0.00	0.04	-0.01	-0.03	0.06	-0.02	0.01	1.00	0.04	-0.01
winner	-0.03	-0.05	-0.02	-0.06	0.12	0.45	0.42	0.46	0.04	1.00	0.02
result	-0.01	0.02	-0.01	0.05	0.05	0.05	-0.03	0.02	-0.01	0.02	1.00

Spearman for Matches

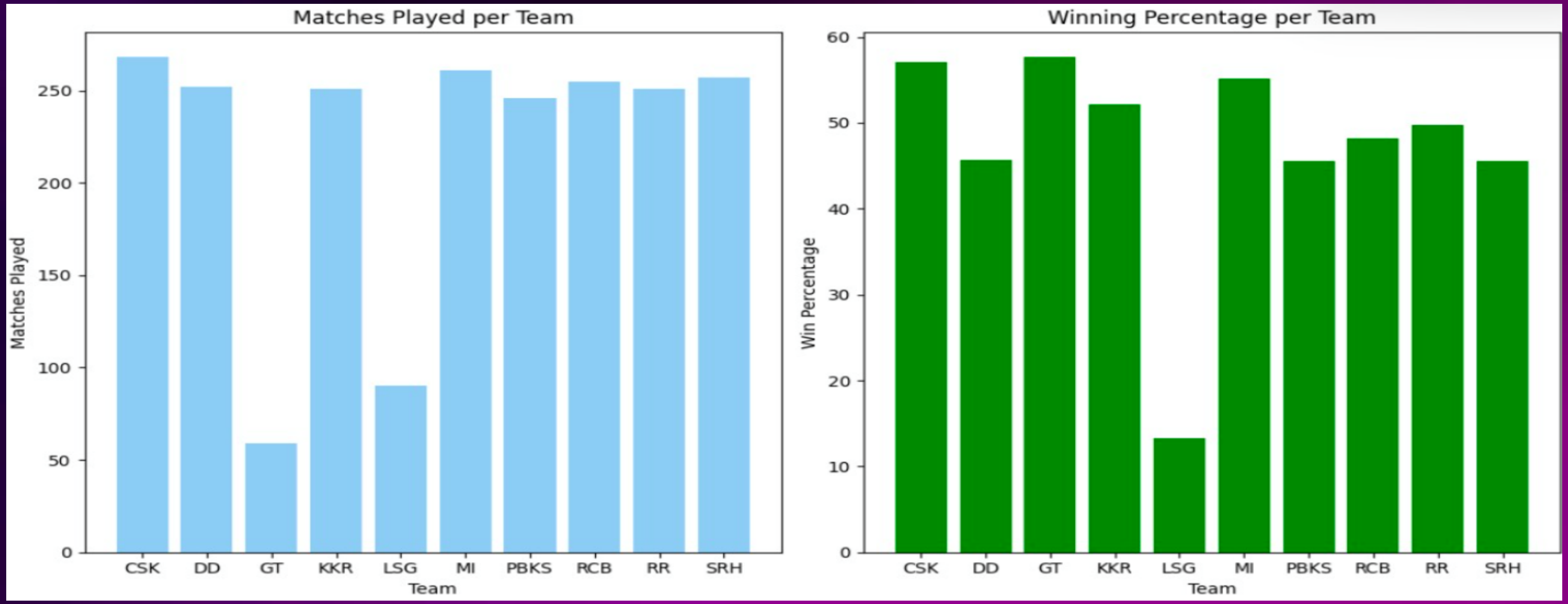
For all the categorical features, we design the Spearman coefficient matrix for all the features and observe all the correlations.

Team Performance



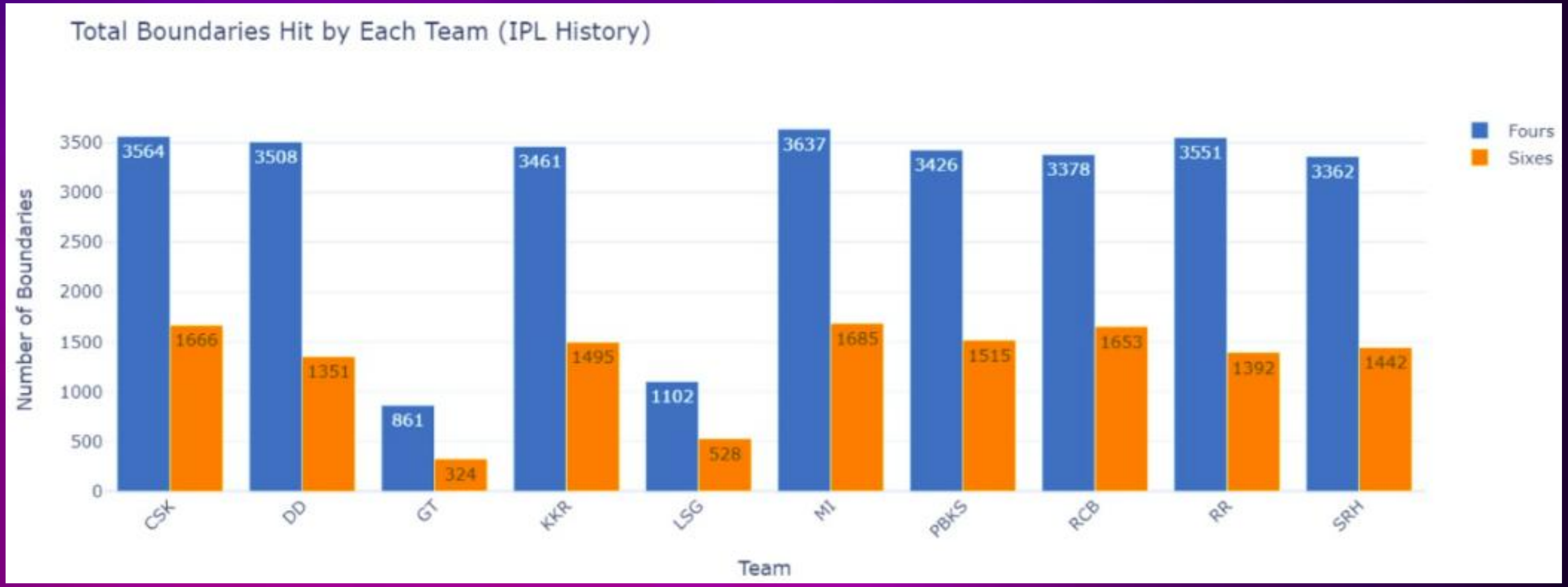
Run per over of each team

Displaying team-wise runs per over across matches to reveal consistency in scoring rates.



Plot Matches Played and Winning Percentages

This graph illustrates the total matches played by each IPL team and their corresponding winning percentages. It helps in understanding the consistency and dominance of different teams over the seasons.

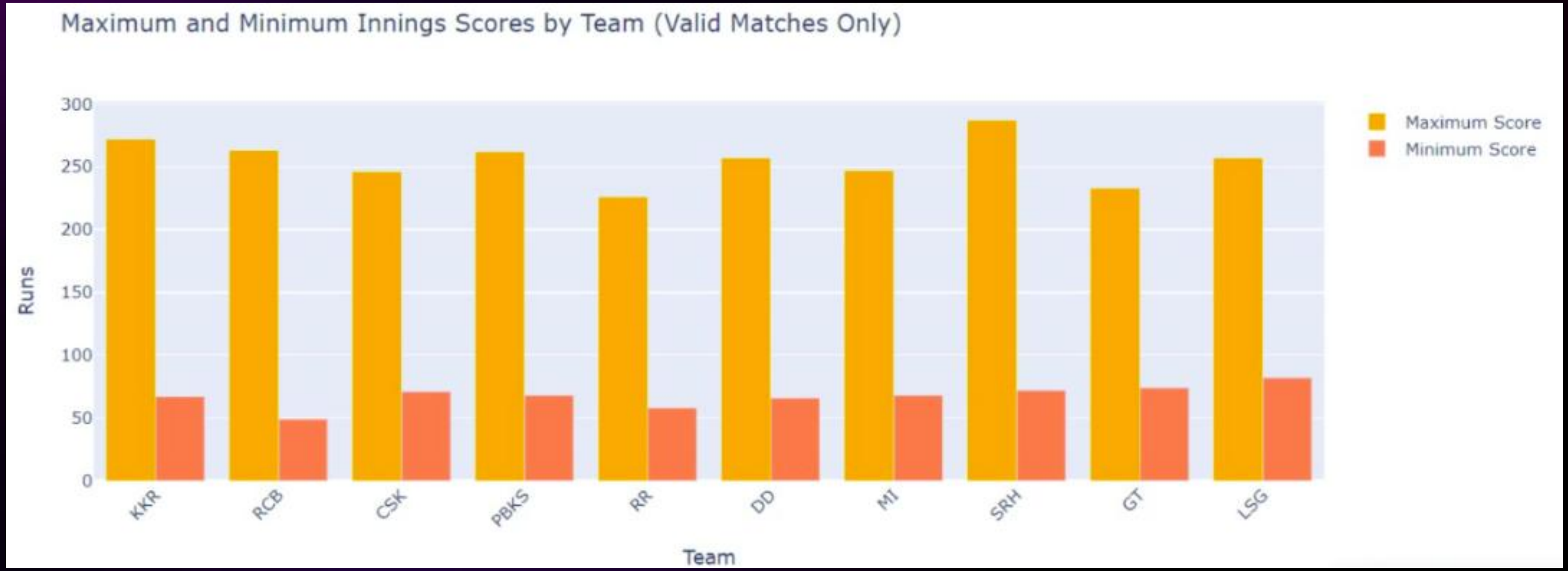


All Boundaries Scored by IPL Teams

This graph depicts the total number of boundaries (fours and sixes) hit by teams across matches. It provides insights into the attacking approach of teams and their ability to score quickly.

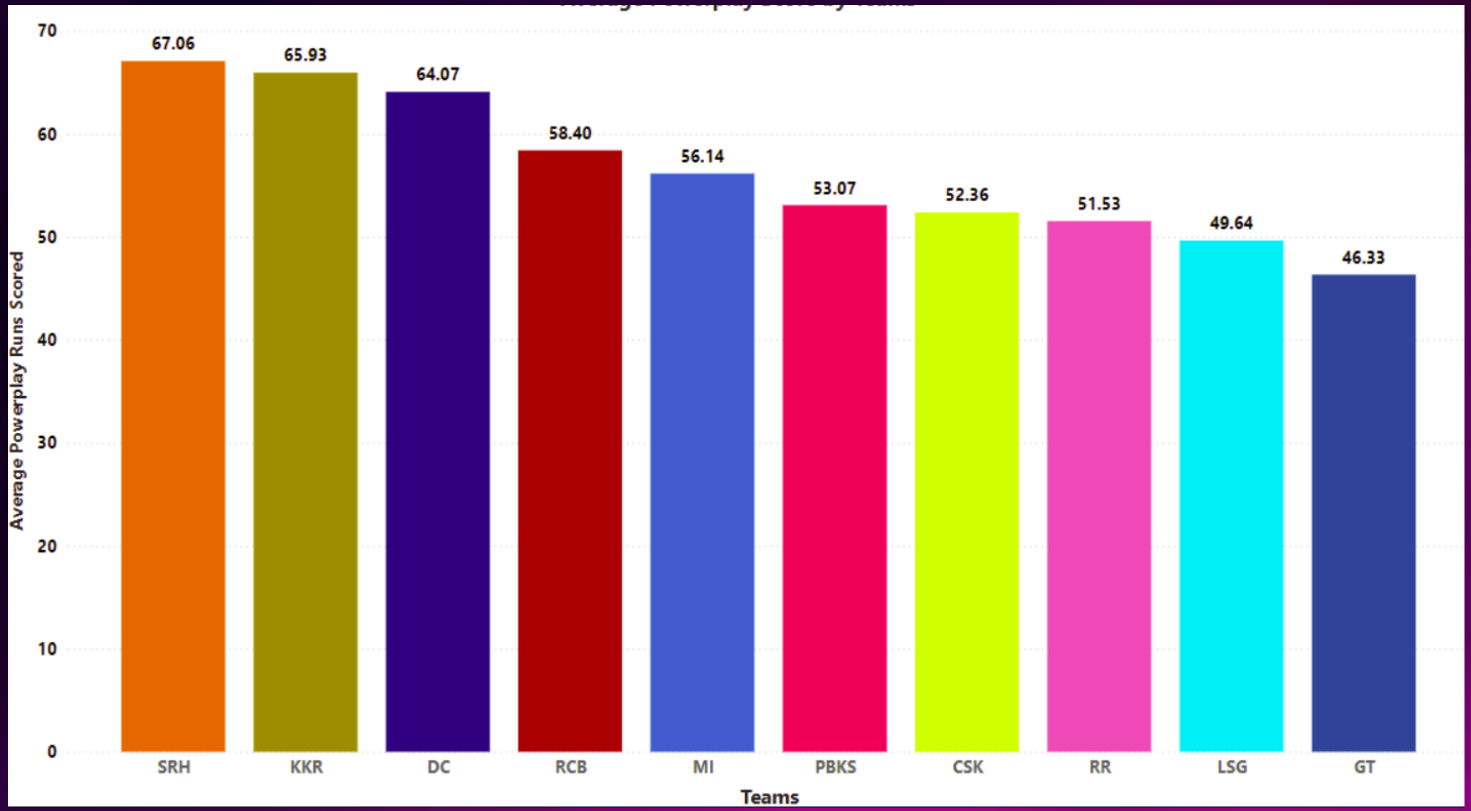
Maximum and Minimum Innings Scores by Team

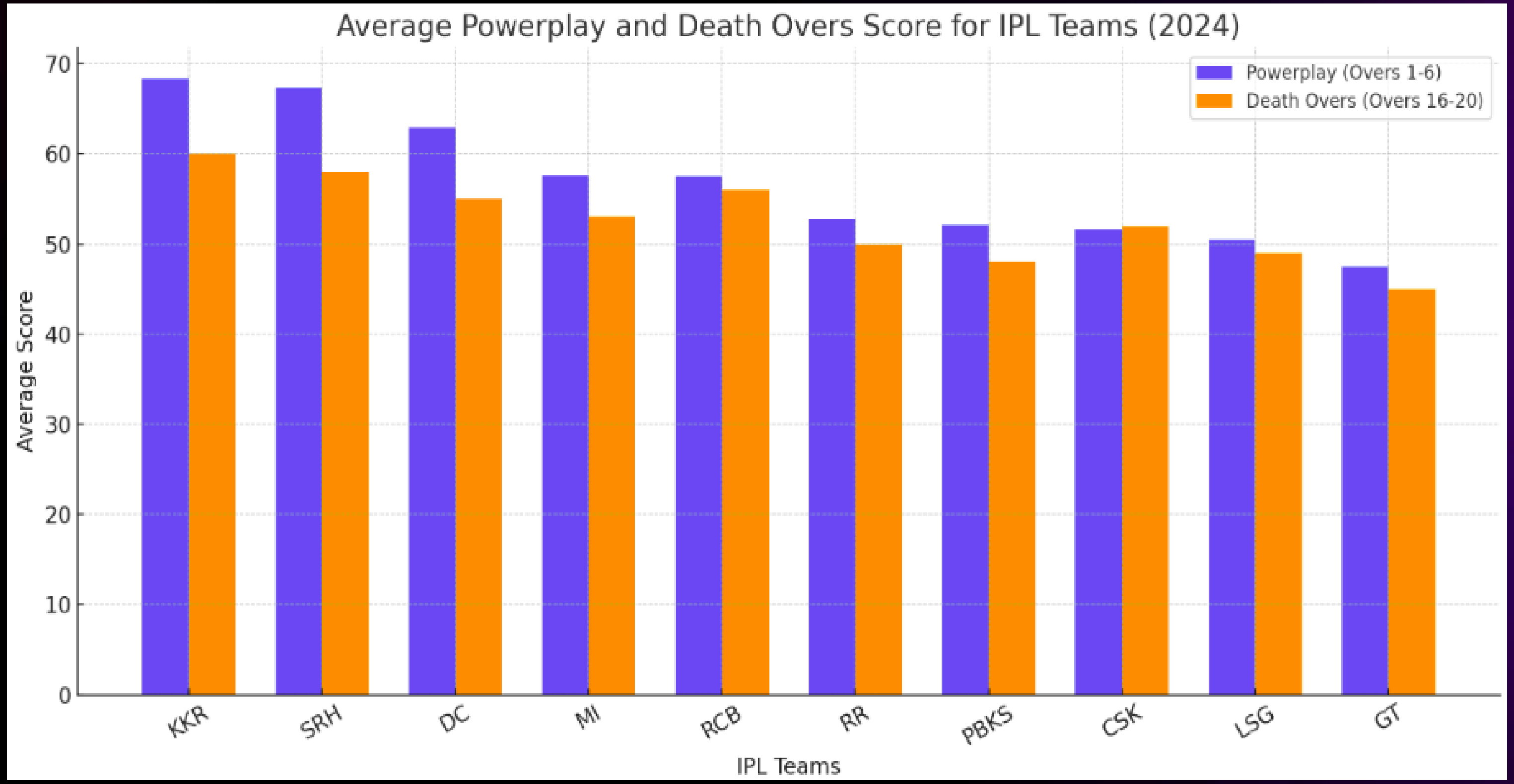
A detailed breakdown of the average runs scored per over by each team. This helps in evaluating the consistency of a team’ s batting performance across different phases of the game.



Average Powerplay Score by Teams

This graph illustrates the average runs scored by teams in the first six overs of IPL 2024. It highlights teams that capitalized on field restrictions to gain early momentum.





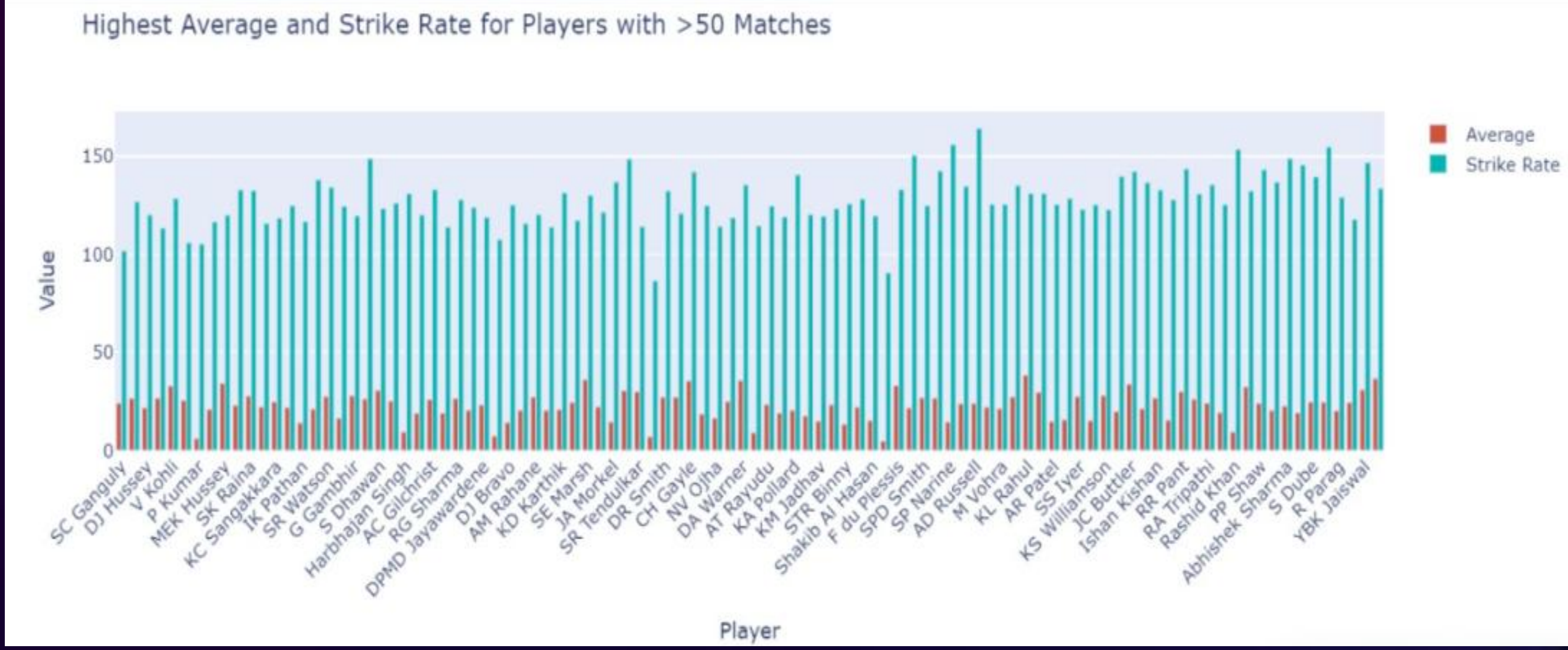
Average Powerplay and Death Overs score for IPL Teams (2024)

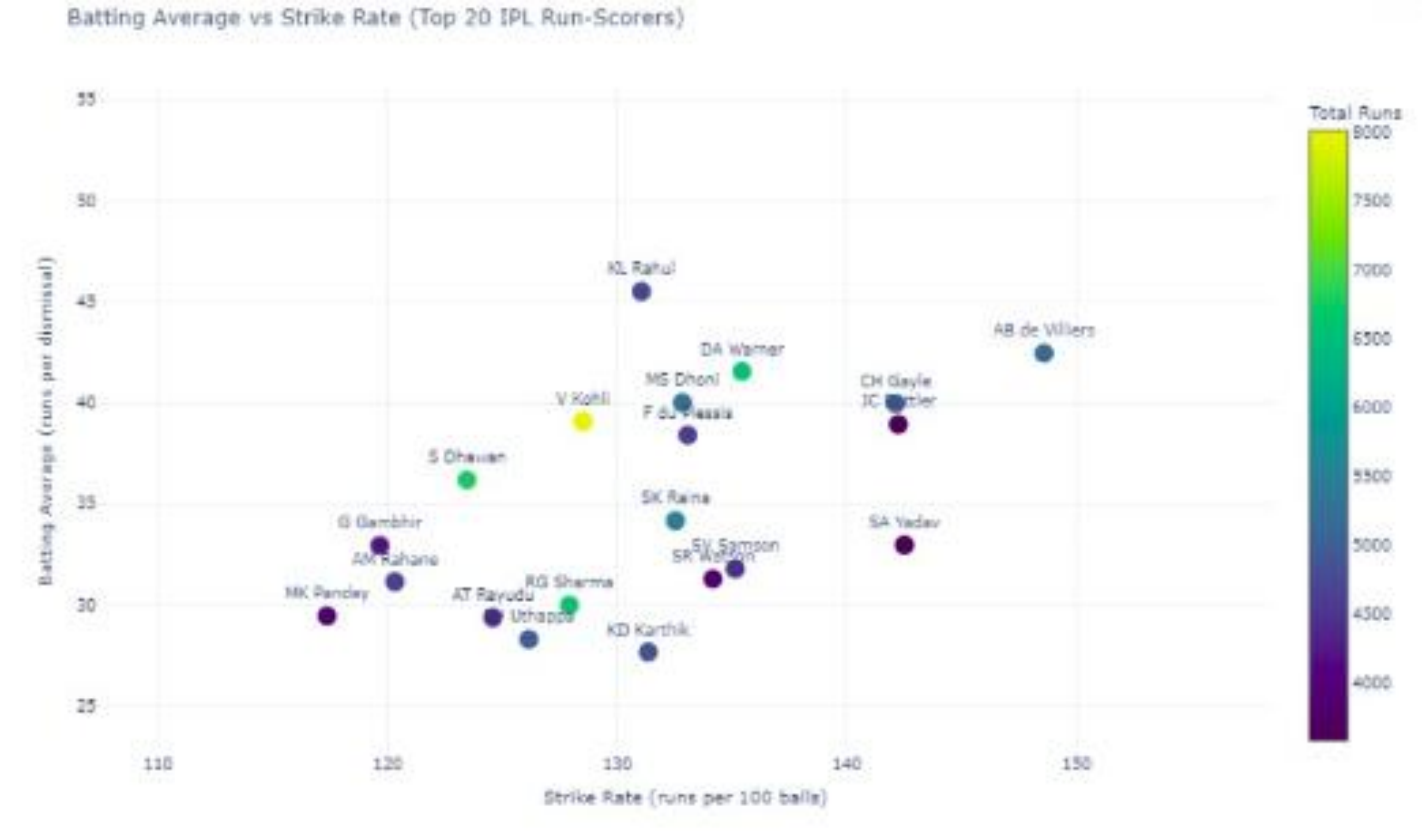
This visualization showcases the average runs scored by teams in the final four overs of IPL 2024. It reflects teams’ ability to accelerate in the closing stages and maximize their total.

Player Performance

Strike Rate Analysis for players with more than 50 Matches

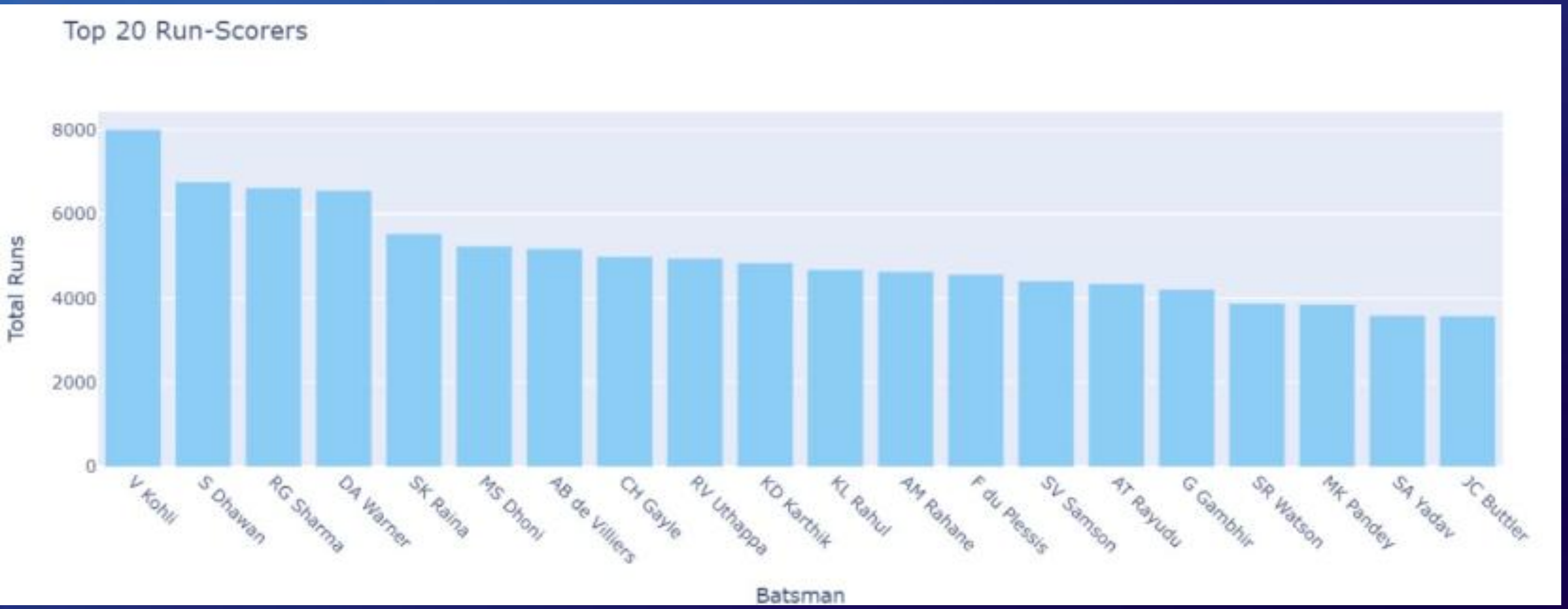
This visualization compares the batting averages and strike rates of the top 20 run-scorers, providing insights into players who balance consistency with aggressive scoring.





Batting Average Vs Batting Strike Rate

This graph groups players based on their batting averages and bowling economy rates, helping identify all-rounders and specialists in batting and bowling.

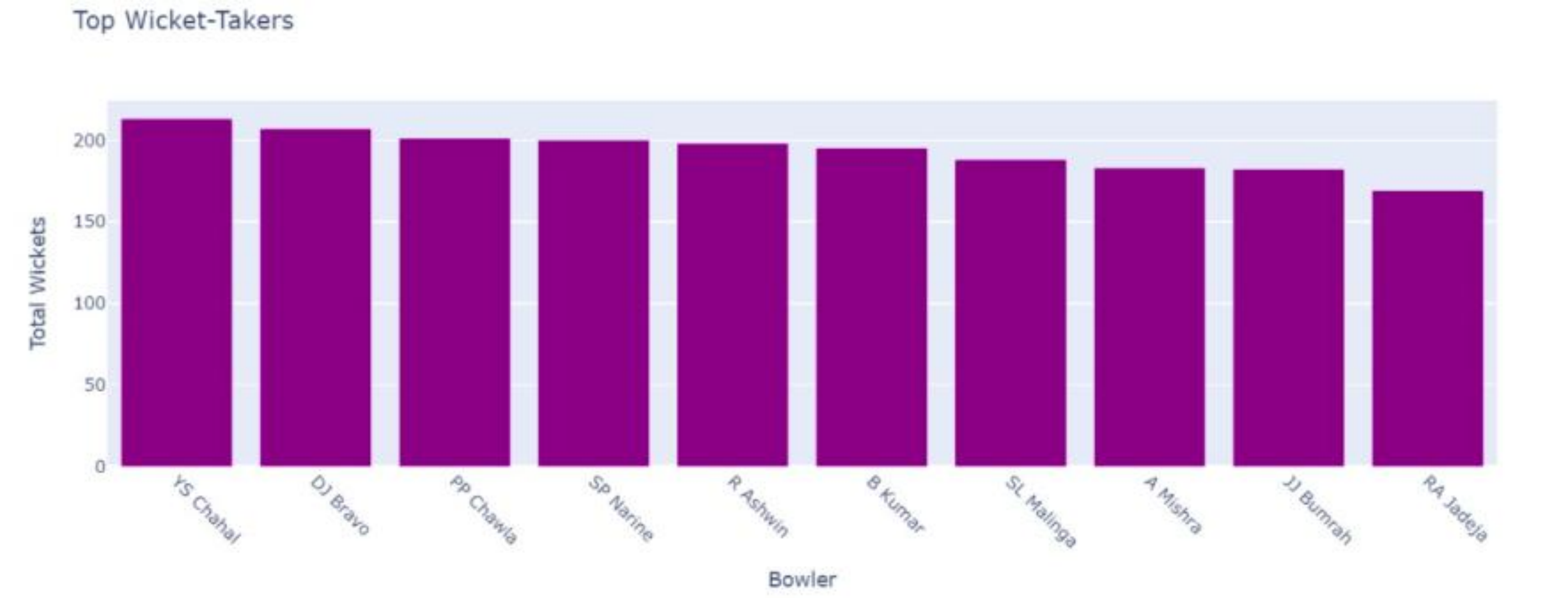


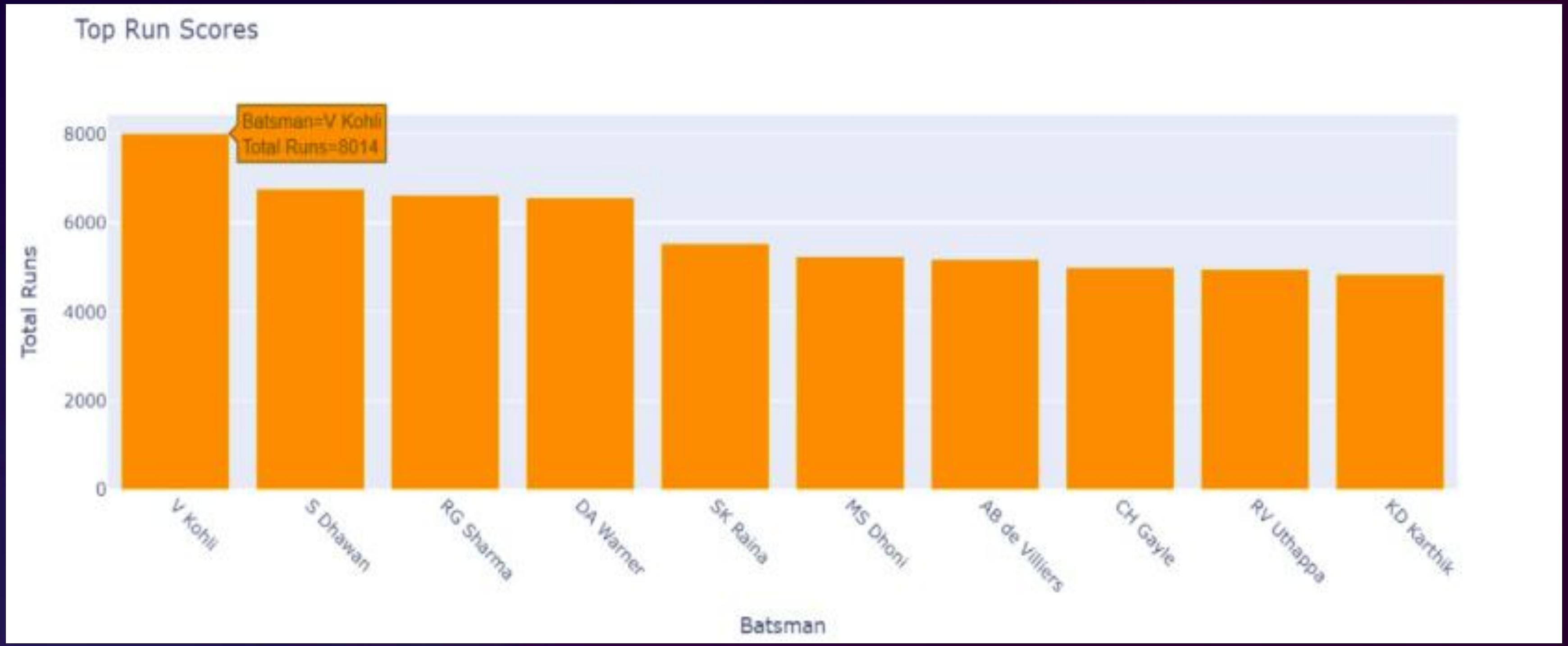
The top 20 run-scorers

A detailed look at the top 20 batsmen with the most career runs in the IPL. This helps identify the most prolific run-makers and their impact across seasons.

Top Wicket Takers

This graph showcases the leading wicket-takers in IPL history, highlighting bowlers who have been the most successful in picking up wickets over multiple seasons.



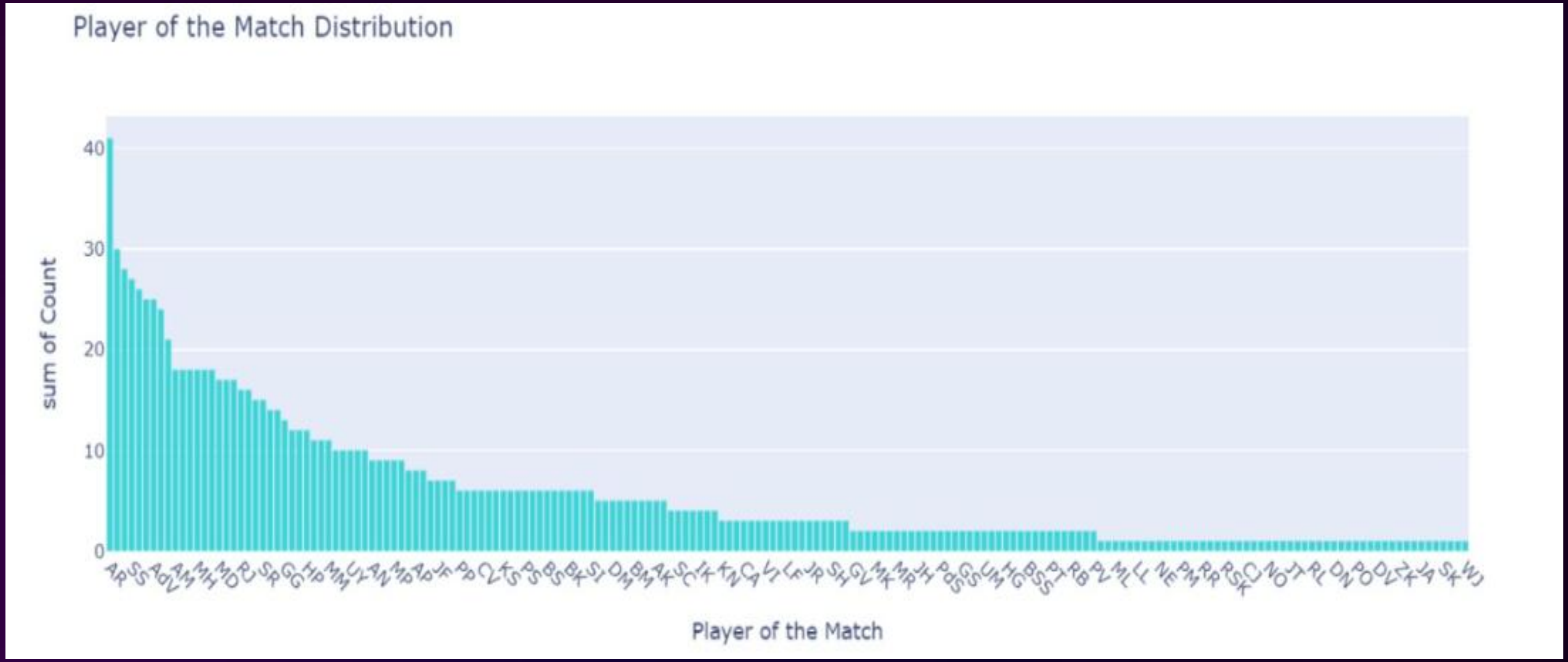


Top Run Scorers

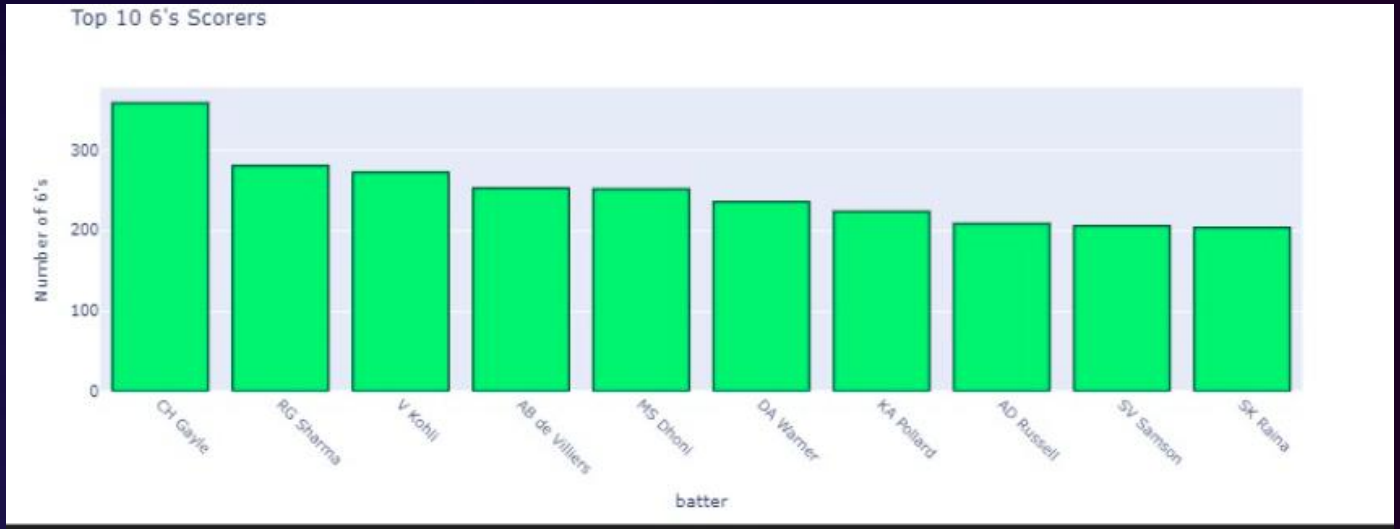
This graph displays the highest run-scorers in IPL history, highlighting players who have consistently performed with the bat and contributed significantly to their teams.

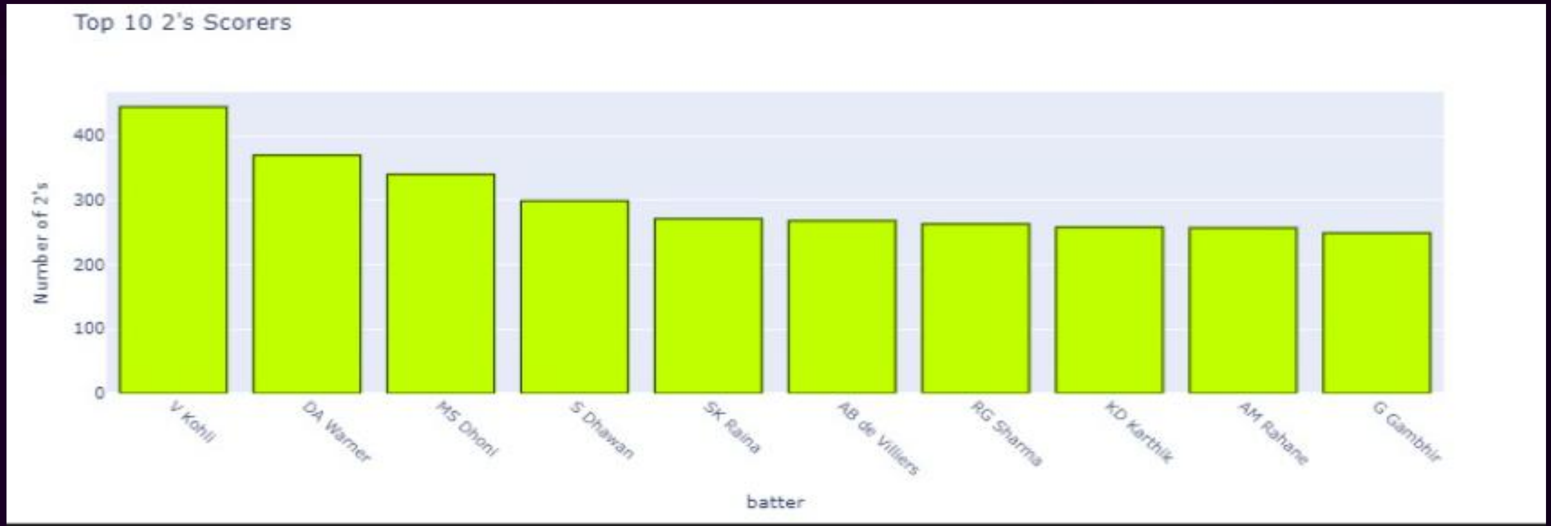
Player of Match over Seasons

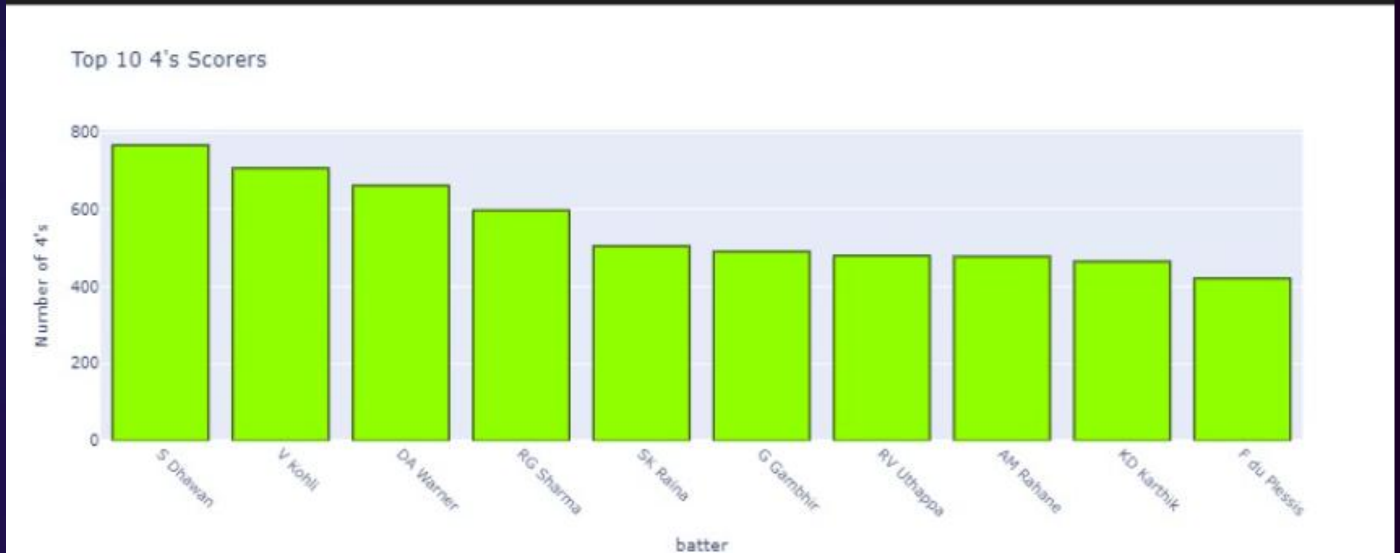
This visualization tracks players with the most ‘Player of the Match’ awards, recognizing those who have delivered match-winning performances consistently.

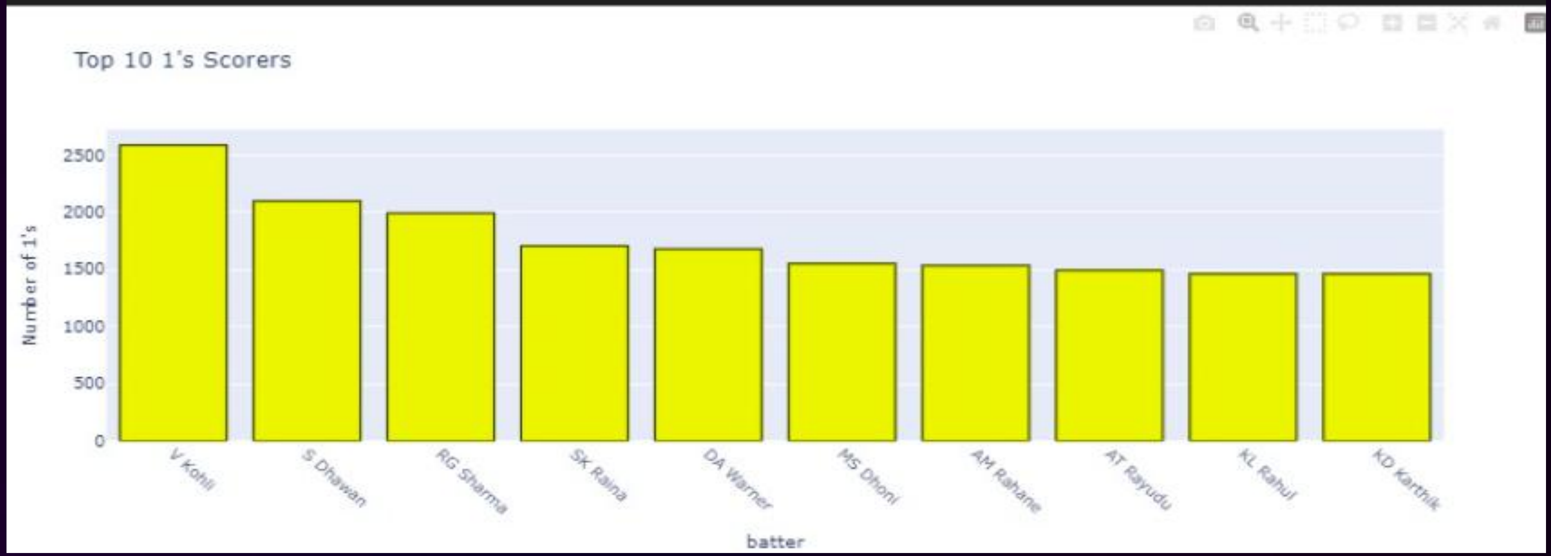


Top 10 Batsmen in each run category

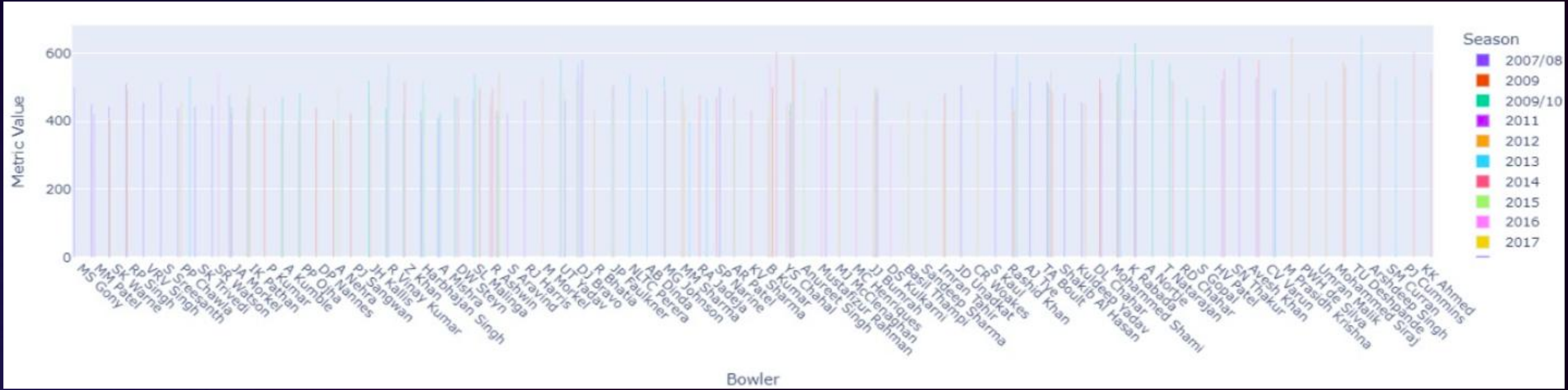




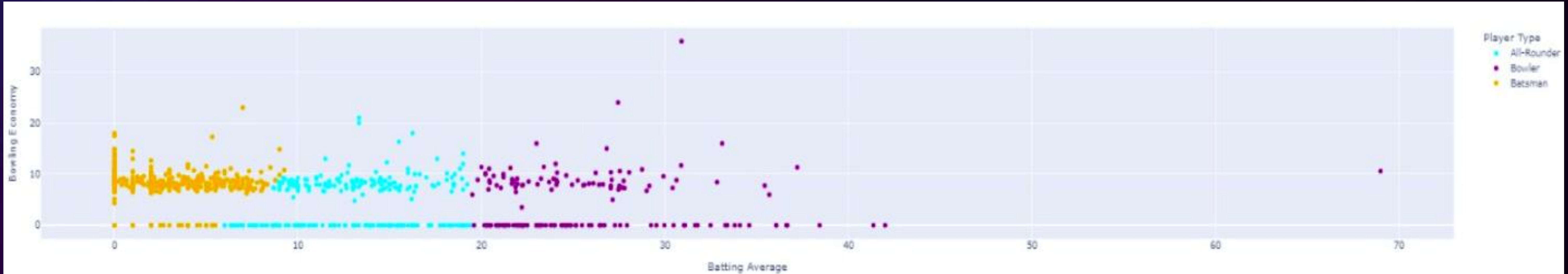




Top 10 Bowlers per season with metrics

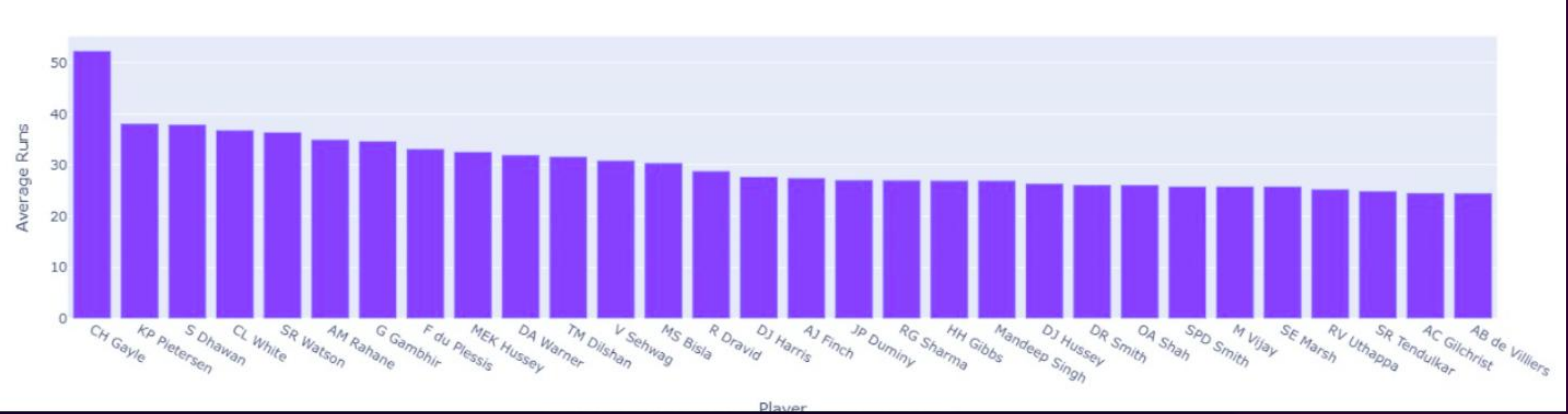


Bowling Economy VS Batting Average

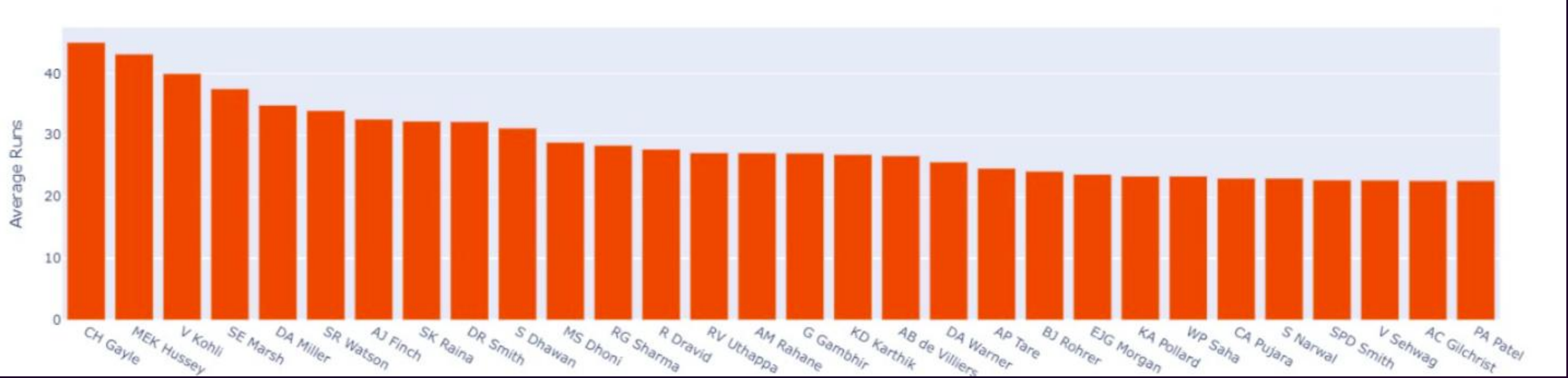


Seasonal Analysis

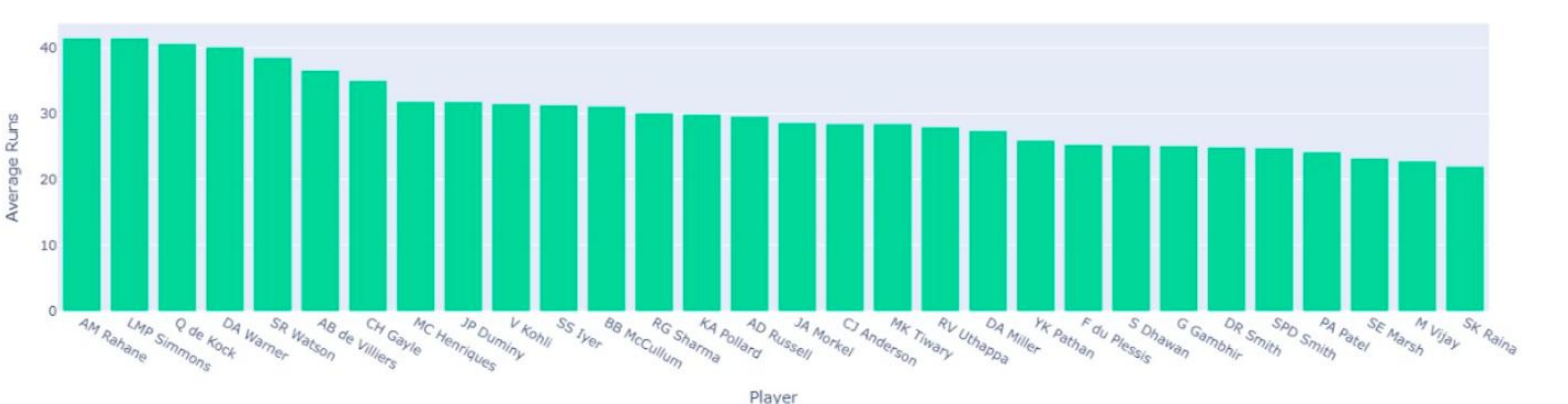
Notable Player Performance in Season 2012

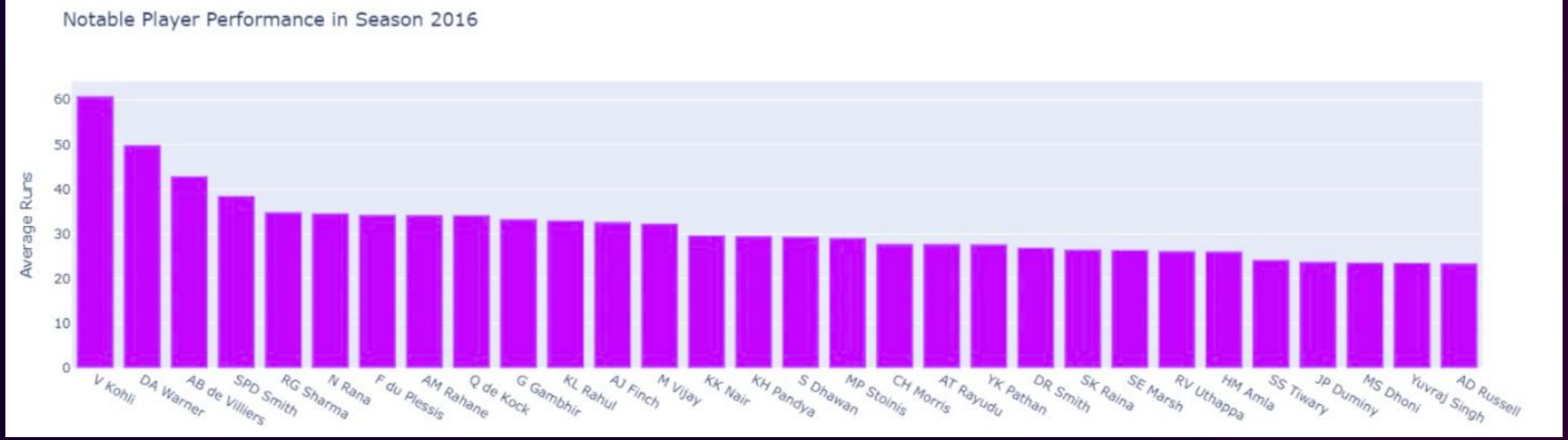


Notable Player Performance in Season 2013

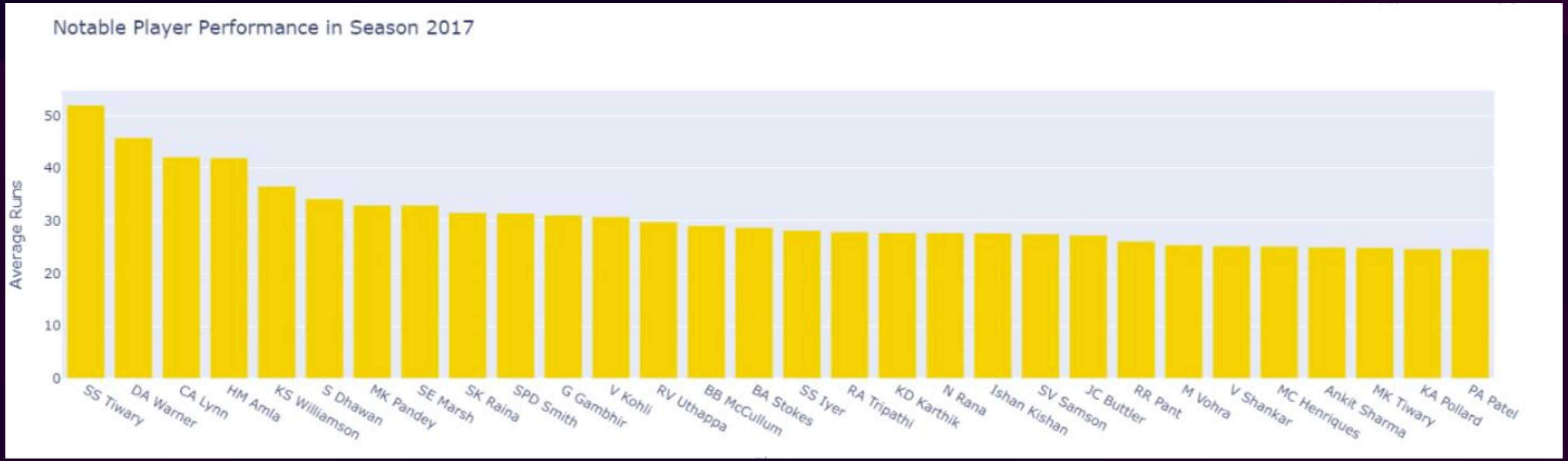


Notable Player Performance in Season 2015

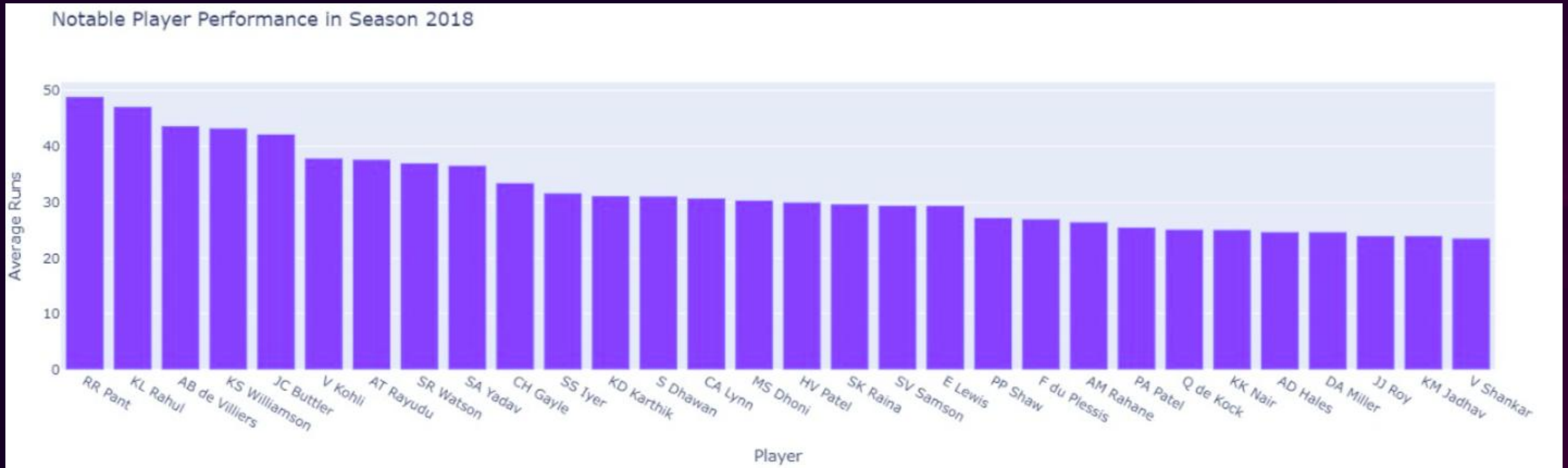




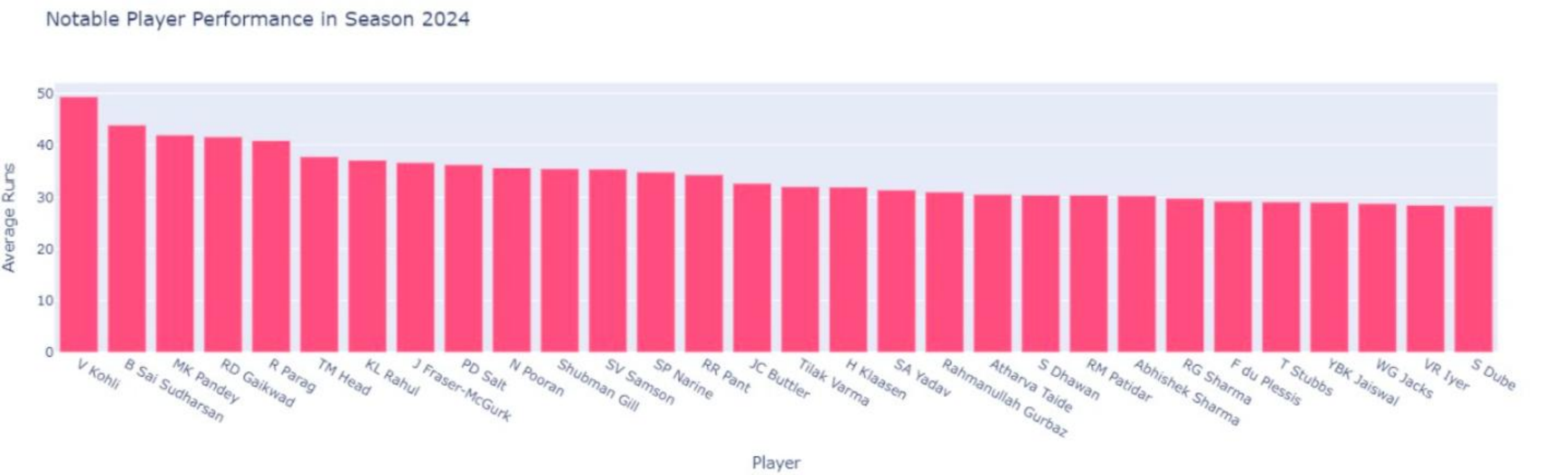
Notable Player Performance in Season 2016

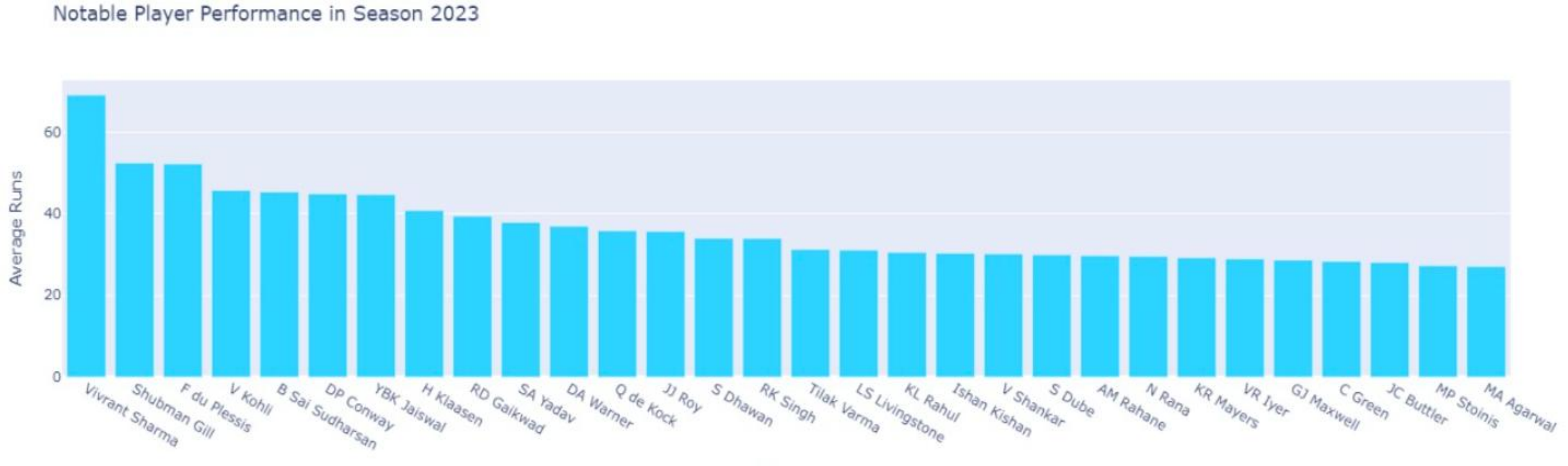


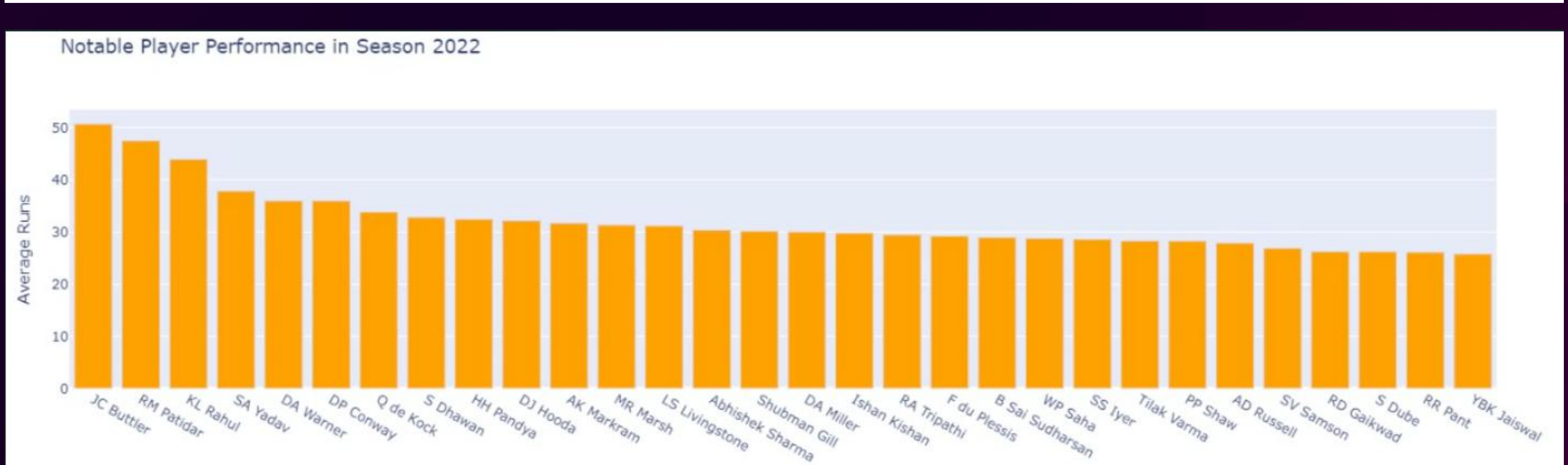
Notable Player Performance in Season 2017

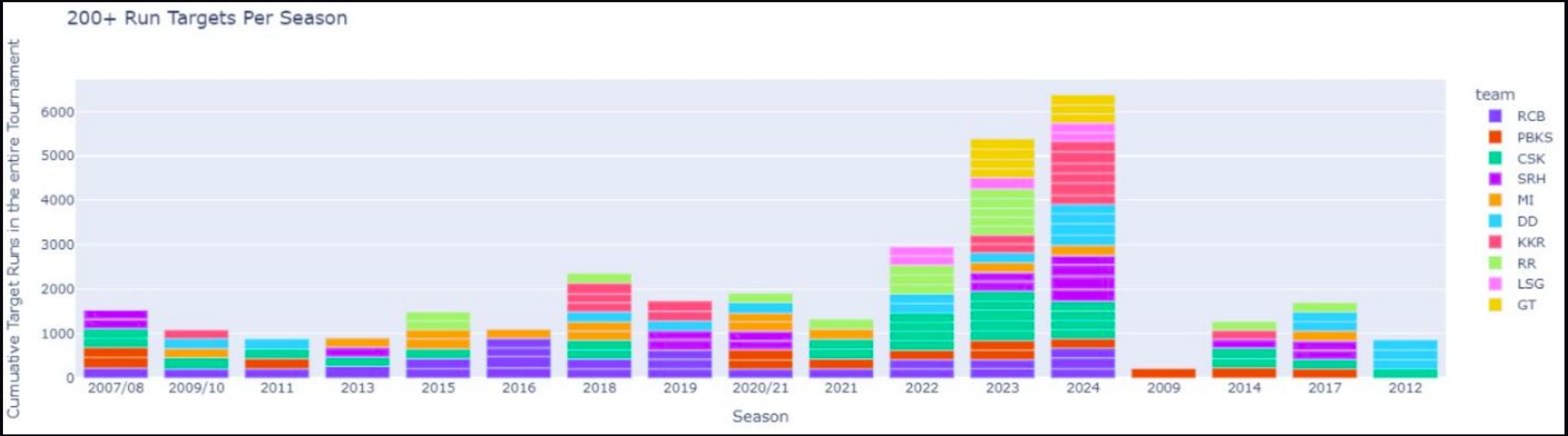


Notable Player Performance in Season 2018

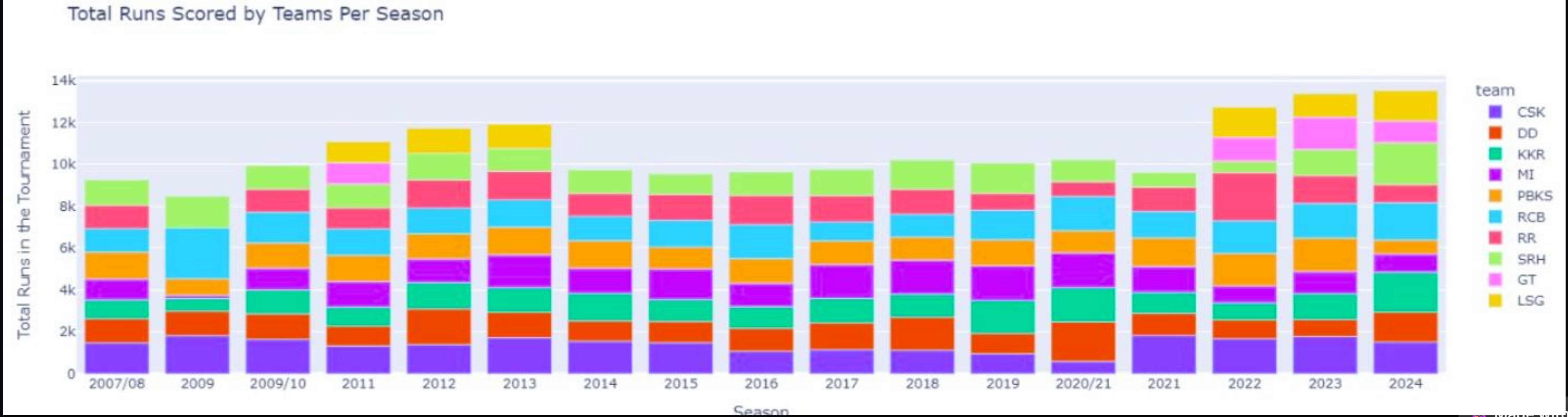

Notable Player Performance in Season 2024


Notable Player Performance in Season 2023

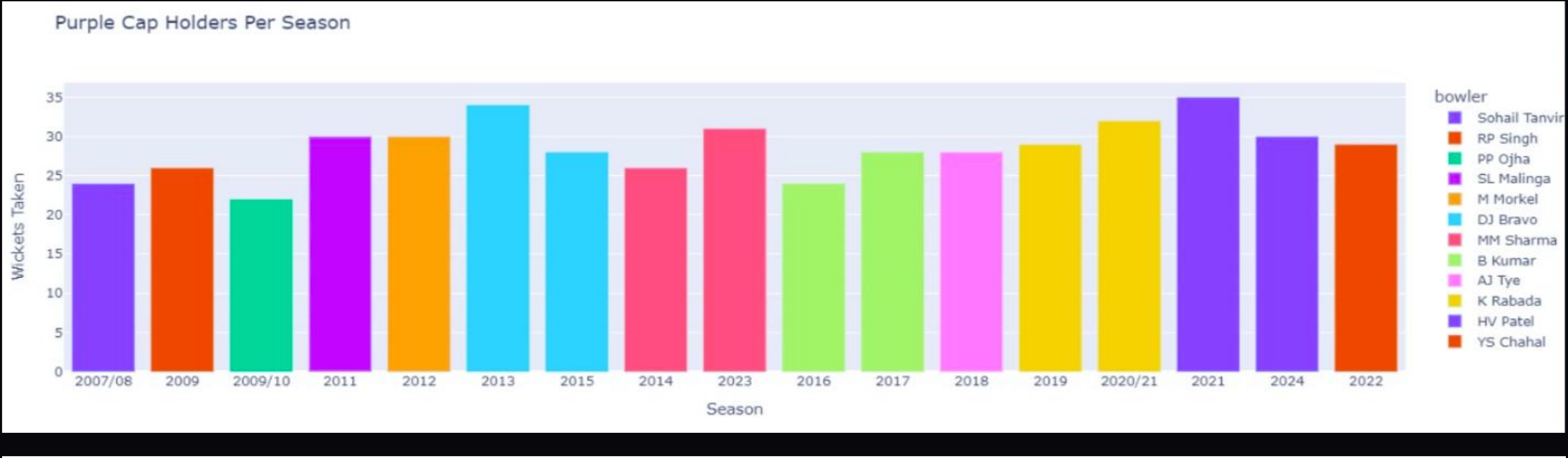

Notable Player Performance in Season 2022



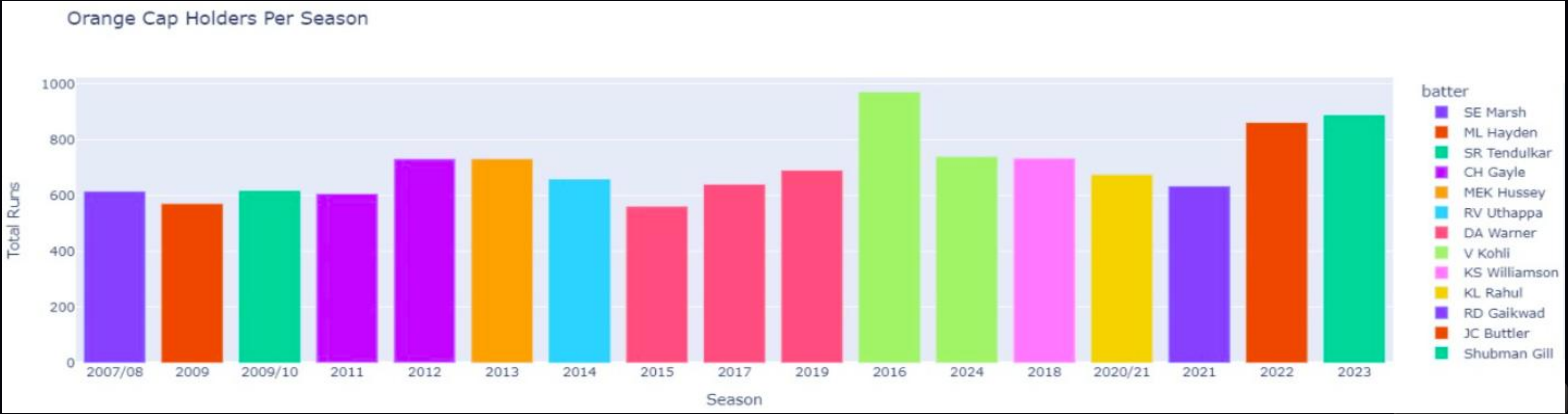
200+ Run Targets Per Season



Total Runs Scored by Teams Per Season



Purple Cap Holders Per Season



Orange Cap Holders Per Season

Machine Learning: Feature Selection

1 Correlation Analysis

2 Feature Importance

3 Key Predictors

- We identify relevant features for prediction.
- More targets runs result in greater winning rates for the score setters.
- Teams like CSK, MI, KKR, RR have higher win percentages than other teams.

Machine Learning: Model Selection & Training

1

Algorithms

We used Random Forests, Decision Tress, XGBClassifier, LightBGM, Categorical Naive Bayes, before finally settling on the MultiLayerPerceptron Classifier.

2

Training Data

80-20 train-test split was done on all the matches in the history of IPL.

3

Validation Data

Optuna automatically generates the validation data for each trial epoch.

Tournament Simulation Using MLP Classifier

1. Generating Match Fixtures

- Extracts unique teams from the dataset.
- Generates all possible matchups using `itertools.combinations()`.
- Initializes a points table with zero points for each team.

2. Synthetic Match Generation

- Samples key match features (umpire, venue, season, toss decision) from historical data.
- Runs **1000 Monte Carlo simulations** per matchup for randomness.
- Generates match conditions: teams, target runs, venue, umpires, match type, etc.

3. Match Outcome Prediction

- Updates the points table dynamically.
- **Mapping Encoded Teams to Original Names**
- Converts encoded team IDs back to original names.
- Merges renamed franchises (e.g., **DD → DC, LSG adjustments**).
- **Season Winner Estimation**
- The team with the highest points is the predicted winner.

4. Mapping Encoded Teams to Original Names

- Converts encoded team IDs back to original names using a mapping dictionary (`label_mapping`).
- Franchise renaming is handled:
  - `"Lucknow Super Giants"` points are merged into `"LSG"`.
  - `"Delhi Daredevils"` (`DD`) is renamed to `"Delhi Capitals"` (`DC`).

Key Takeaways for Tournament Simulation

- **Monte Carlo-style simulations** generate multiple match outcomes for better statistical accuracy.
- The **MLP model generalizes** match conditions and predicts results based on historical patterns.
- **Dynamic points tracking** enables real-time tournament simulation.
- **Post-processing ensures accurate team representation** despite name changes.

This approach can be extended to **multi-stage tournaments**, knockout formats, or **weighted probability simulations** for enhanced realism.

Model Evaluation & Tuning

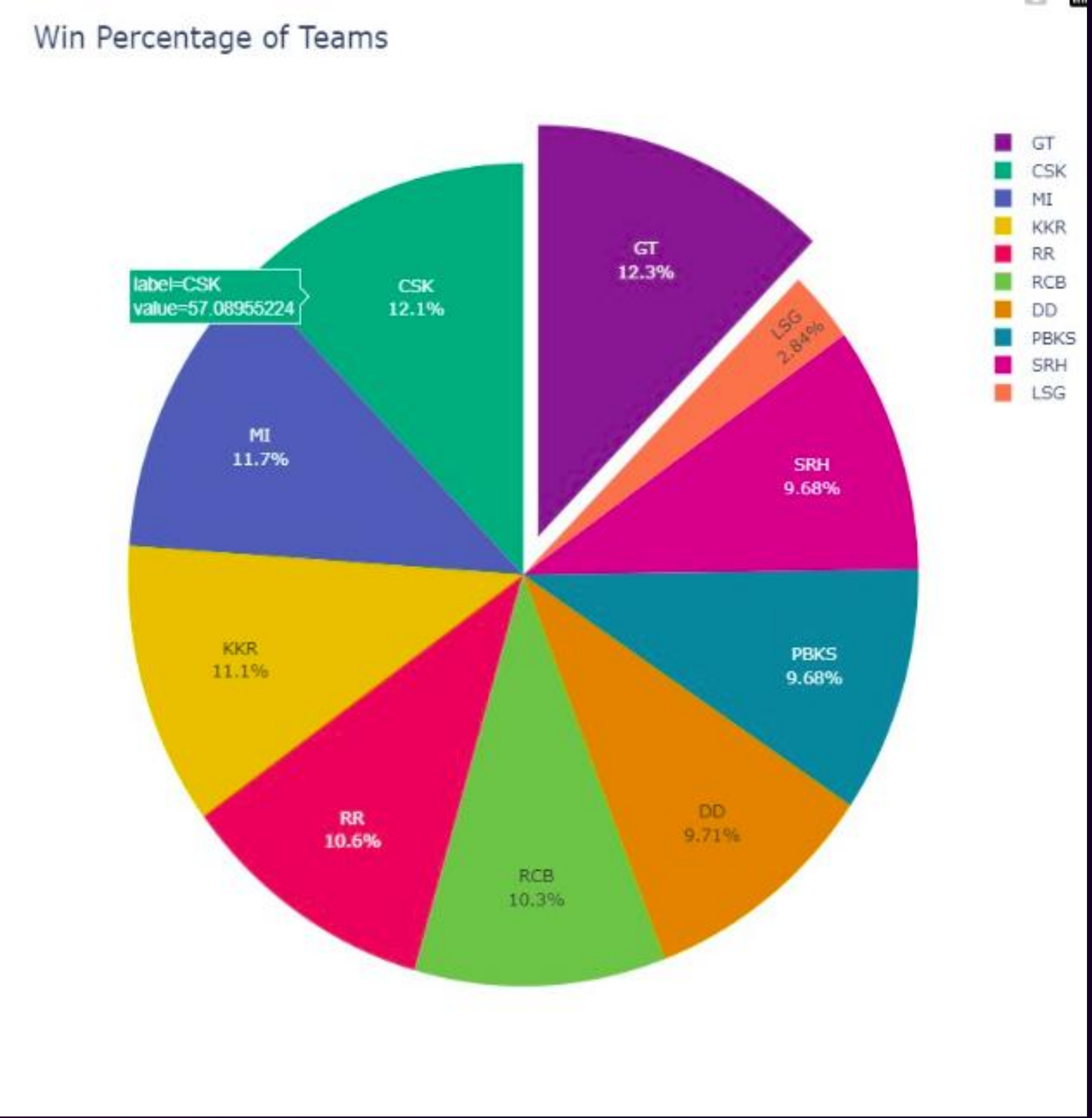
94.82%

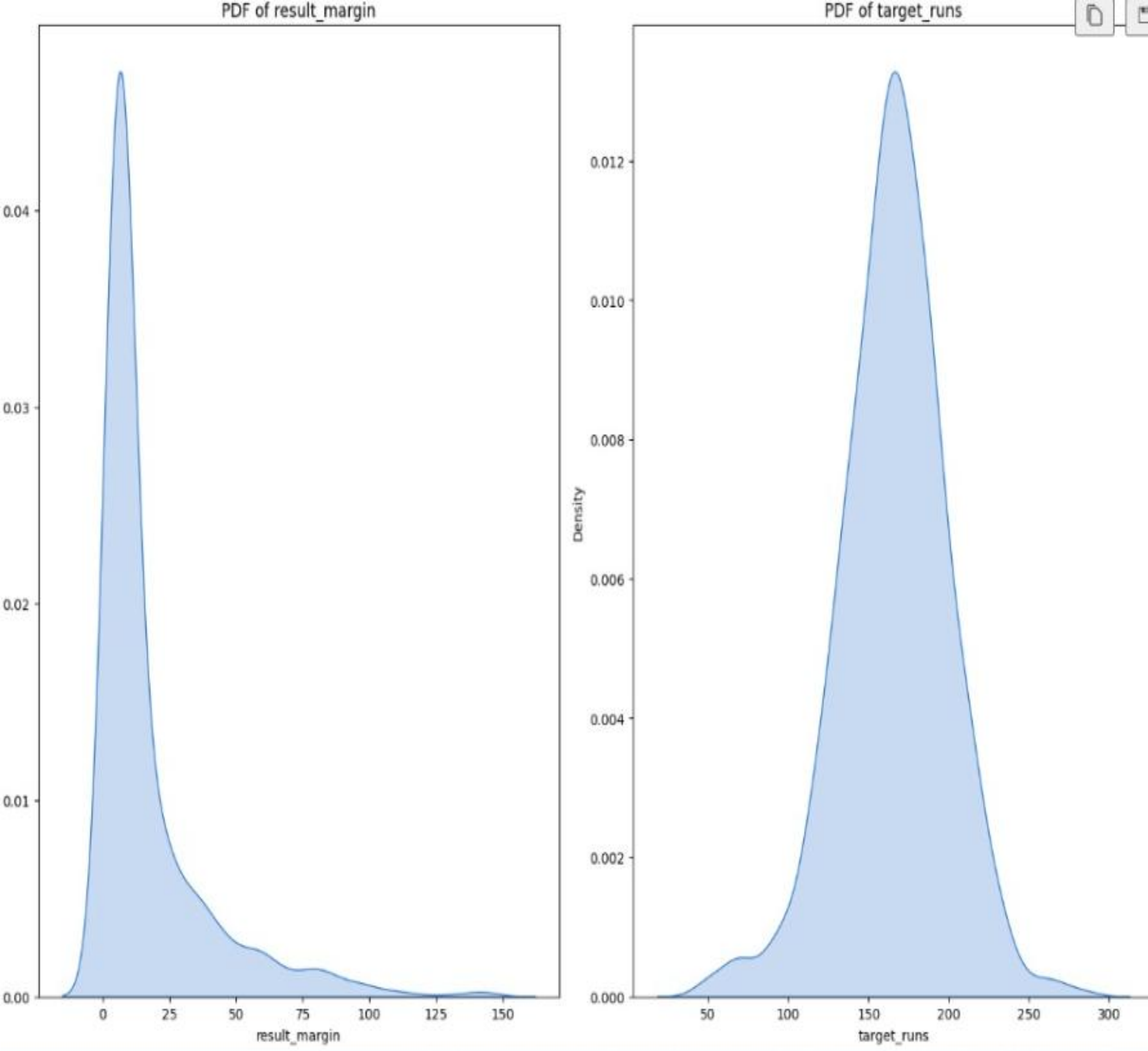
Aggregate tuned MLP accuracy.

97.8%

Highest accuracy achieved during 100 trials of Optuna hyperparameter optimization

We use **Accuracy**, **Precision**, **Recall**, and **F1-Score**. Cross-validation ensures robustness. Hyperparameter tuning optimizes performance.

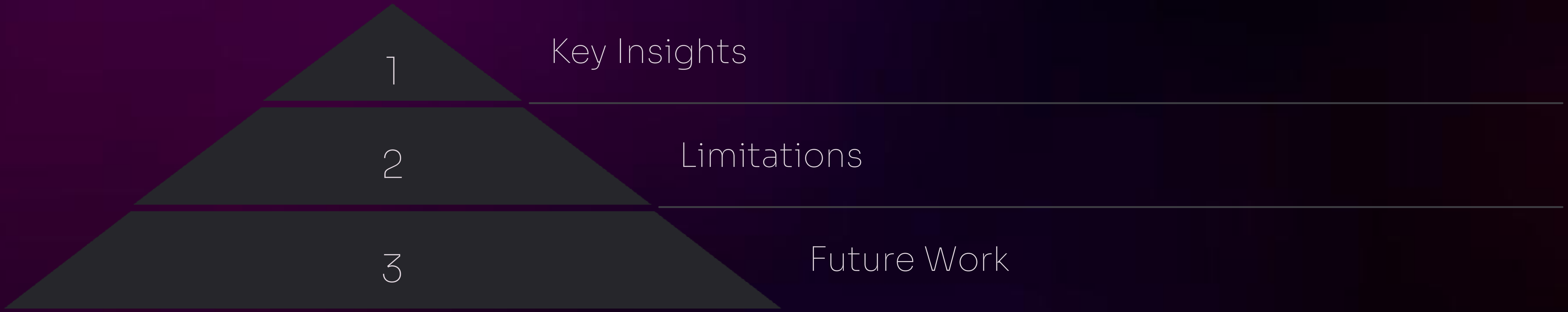




This graph showcases the win percentages of all IPL teams, helping to compare their overall success rates across seasons. It highlights the most dominant teams and those that have struggled.

This visualization shows the probability distribution of target scored and winning margins, whether by runs or wickets. It provides insights into how closely contested IPL matches tend to be.

Conclusion & Future Work



1 Key Insights

2 Limitations

3 Future Work

We will incorporate weather and social media data. We will refine models and build a real-time prediction system. Thanks to data sources, team members, and mentors!

Thank You!!

ISeeData Members: Deepon Halder, Arkapravo Das, Soham Haldar, Aritro Shome, Vedanta Saha