PREDICTING FINAL CRICKET SCORES USING LINEAR REGRESSION

PROBLEM STATEMENT

IN CRICKET, PREDICTING THE FINAL SCORE IS IMPORTANT FOR STRATEGY.

THE GOAL IS TO ESTIMATE TOTAL BASED ON:

VENUE

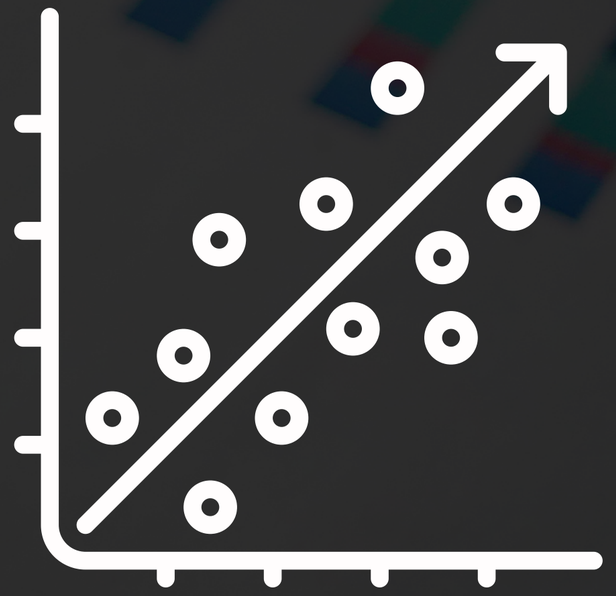
TEAMS PLAYING

TOSS WINNER AND DECISION

TYPE OF PROBLEM & STATISTICAL MODEL TO BE USED

TYPE OF PROBLEM: REGRESSION,

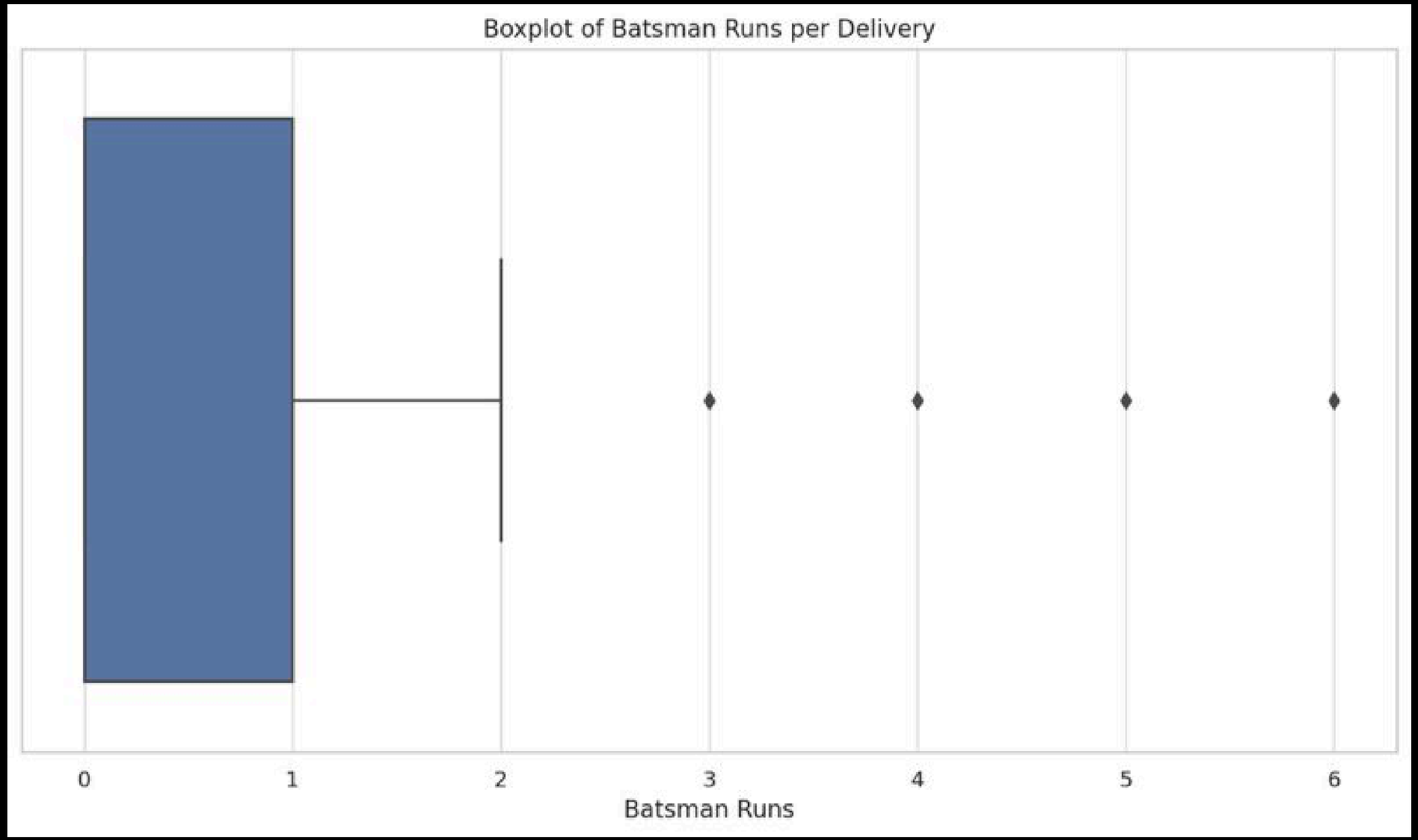
STATISTICAL MODEL: MULTIPLE LINEAR REGRESSION

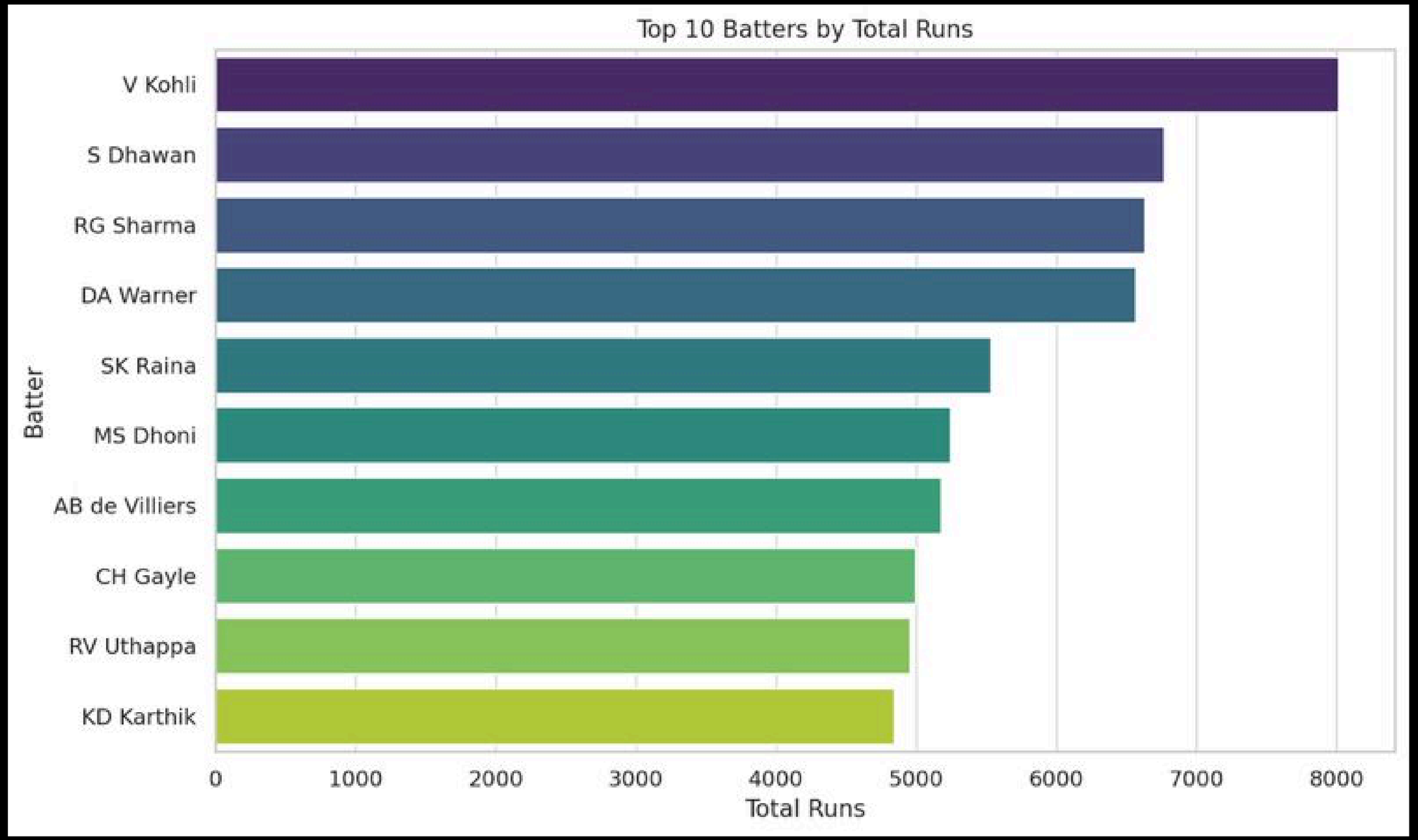


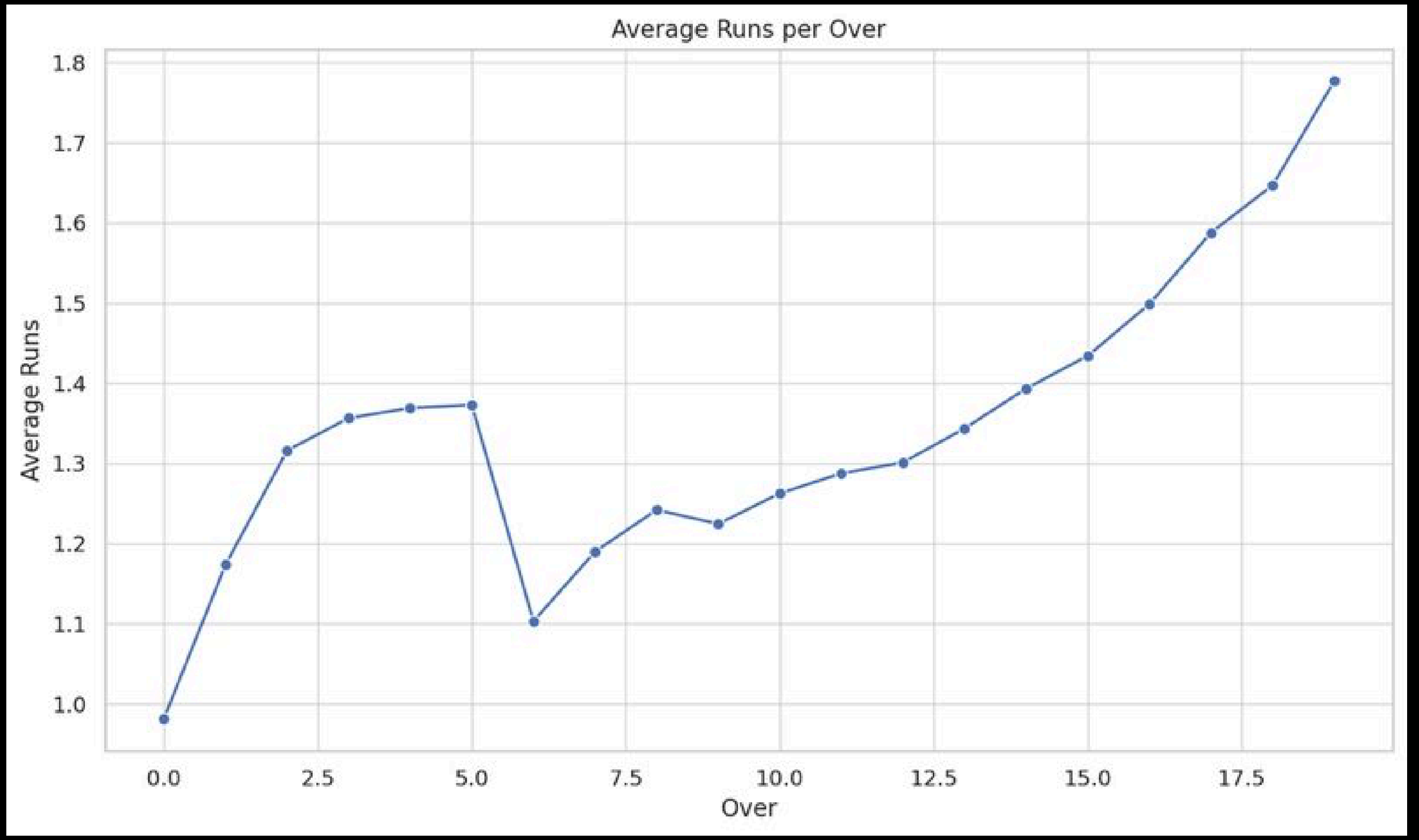
EXPLORATORY DATA ANALYSIS (EDA)

UNDERSTANDING DATA PATTERNS & PREPROCESSING TECHNIQUES

EXPLORATORY DATA ANALYSIS (EDA)


Boxplot of Batsman Runs per Delivery
0
1
2
3
4
5
6
Batsman Runs


Top 10 Batters by Total Runs
V Kohli
S Dhawan
RG Sharma
DA Warner
SK Raina
MS Dhoni
AB de Villiers
CH Gayle
RV Uthappa
KD Karthik
Batter
0
1000
2000
3000
4000
5000
6000
7000
8000
Total Runs


Average Runs per Over
Average Runs
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Over

BRIEF DESCRIPTION OF THE CHOSEN MODEL

Multiple Linear Regression is used to model the relationship between one continuous dependent variable and two or more independent variables.

USE CASE

Predict the score

LITERATURE REVIEW

(2023). CRICKET SCORE DATA ANALYSIS. INTERNATIONAL RESEARCH JOURNAL OF MODERNIZATION IN ENGINEERING TECHNOLOGY AND SCIENCE. 1. A DATA SCIENCE APPROACH TO PREDICTING THE OUTCOME OF ODI CRICKET MATCHES

BHAT, N., REVANASIDDAPPA, M., & SRINIVAS, S. (2020). DATA ANALYSIS OF CRICKET SCORE PREDICTION. , 465-472.
2. PREDICTING THE OUTCOME OF ODI CRICKET MATCHES: A MACHINE LEARNING APPROACH
A

SU, X., YAN, X., & TSAI, C. (2012). LINEAR REGRESSION. WILEY INTERDISCIPLINARY REVIEWS: COMPUTATIONAL STATISTICS, 4. HTTPS://DOI.ORG/10.1002/WICS.1198.

DATA SET

IPL Complete Dataset (2008-2024)

The latest and complete IPL dataset (Updated till 2024 Season)

DELIVERIES.CSV

MATCHES.CSV

MODEL CHOICE

- Chi-square test for categorical dependencies
- OLS regression for numerical relationships
- Logistic regression for win/loss classification

KEY FEATURES

Explored key features

From matches.csv:

- match_id
- toss_winner
- toss_decision
- winner (target variable for classification)
- result_margin

From deliveries.csv:

- match_id
- batting_team
- bowling_team
- batsman, bowler, runs, etc.
- total_runs

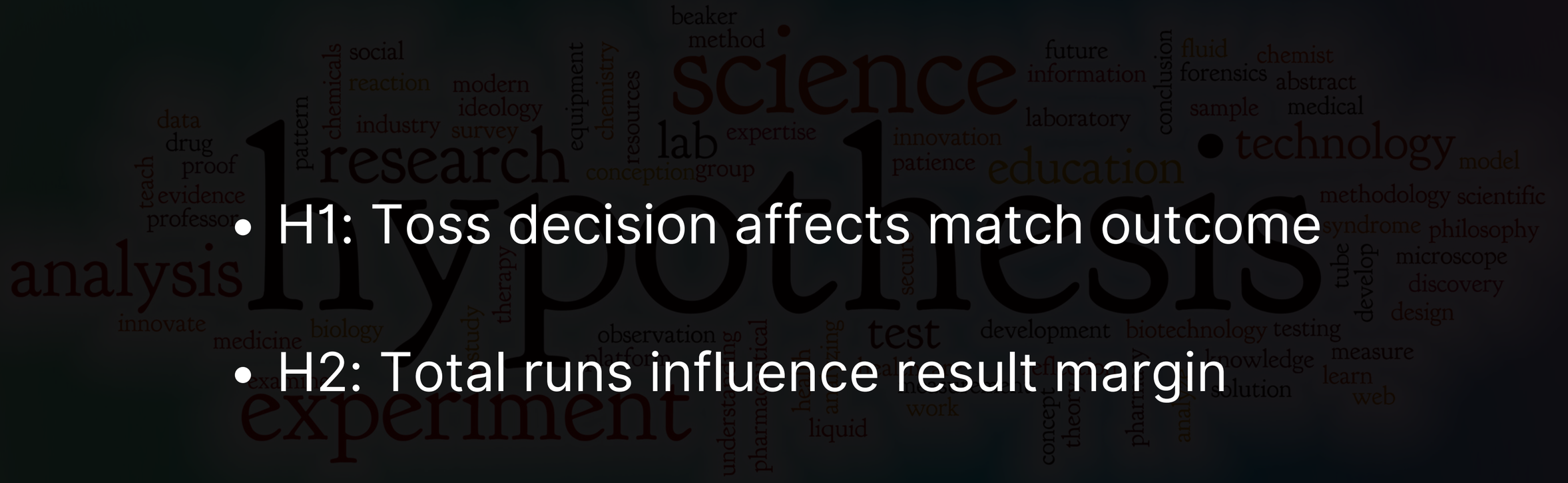
Merged match-level and ball-by-ball datasets.

```
import pandas as pd

# Load data
deliveries = pd.read_csv('deliveries.csv')
matches = pd.read_csv('matches.csv')

# Merge data on match ID
df = deliveries.merge(matches, left_on='match_id', right_on='id')
```

HYPOTHESIS DEVELOPMENT



- H1: Toss decision affects match outcome
- H2: Total runs influence result margin

WEEK 5: STATISTICAL ANALYSIS AND VALIDATION

Descriptive statistics:

- Correlation: total_runs vs result_margin (≈ 0.0047)
- Summary statistics (mean, std, etc.)

Chi-square test:

- Toss decision vs match outcome
- p = 0.0000 → strong statistical relationship

```
Chi-square test result: p = 0.0000
                total_runs  result_margin
total_runs        1.000000       0.004743
result_margin     0.004743       1.000000
```

OLS Regression Results

Dep. Variable:	result_margin	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	5.776
Date:	Tue, 24 Jun 2025	Prob (F-statistic):	0.0162
Time:	04:10:11	Log-Likelihood:	-1.1534e+06
No. Observations:	256796	AIC:	2.307e+06
Df Residuals:	256794	BIC:	2.307e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	17.1955	0.055	312.055	0.000	17.088	17.304
total_runs	0.0630	0.026	2.403	0.016	0.012	0.114

Omnibus:	134795.197	Durbin-Watson:	0.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	853481.880
Skew:	2.537	Prob(JB):	0.00
Kurtosis:	10.350	Cond. No.	3.00

- OLS regression: result_margin ~ total_runs
- R-squared = 0.000 ➔ no practical predictive power
- Coefficient p = 0.016 ➔ statistically significant

Total runs alone don’t explain how much a team wins by. The margin depends on many other things — like how many runs the opponent scored, whether the match was close, etc. We should not use total_runs as the only predictor in models for margin.

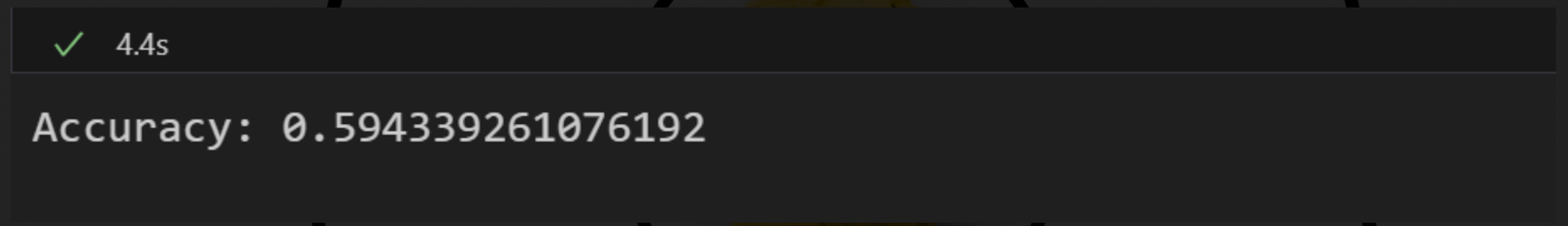
—

WEEK 5: STATISTICAL ANALYSIS AND VALIDATION

```
Chi-square test result: p = 0.0000
                total_runs  result_margin
total_runs        1.000000       0.004743
result_margin     0.004743       1.000000
```

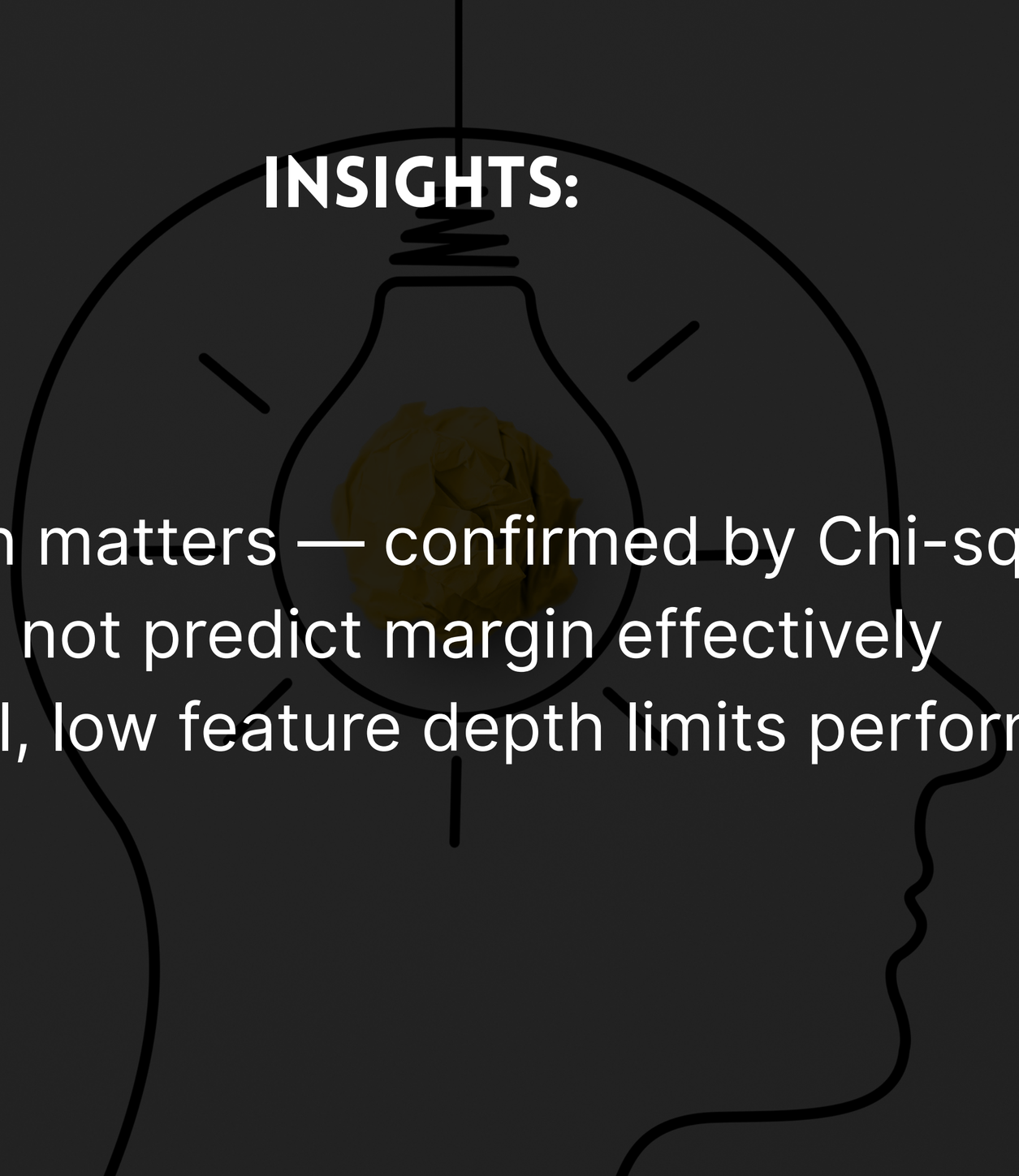
- H1 supported
- H2 statistically significant but not practically useful

WEEK 6: FINAL MODELING AND INSIGHTS



- Built logistic regression model to predict runs scored or conceeded
- Features: Batting team, bowling team, toss decision
- Accuracy: 59.4%

INSIGHTS:



- Toss decision matters — confirmed by Chi-square
- Total runs do not predict margin effectively
- Simple model, low feature depth limits performance

MODEL DEMO

REFERENCE

ASHRAF, S. Z. (2021). VISUAL STORYTELLING WITH BI TOOLS: ENHANCING DATA INTERPRETATION AND COMMUNICATION. INTERNATIONAL JOURNAL FOR MULTIDISCIPLINARY RESEARCH, 3(1). HTTPS://DOI.ORG/10.36948/IJFMR.2021.V03I01.2192

BERNARDINO, J., LAPA, J., & ALMEIDA, A. (2019). COMMERCIAL AND OPEN SOURCE BUSINESS INTELLIGENCE PLATFORMS FOR BIG DATA WAREHOUSING. IN ADVANCES IN DATA MINING AND DATABASE MANAGEMENT BOOK SERIES (PP. 158–181). HTTPS://DOI.ORG/10.4018/978-1-5225-5516-2.CH007

SKIERA, B., REINER, J., & ALBERS, S. (2021). REGRESSION ANALYSIS. IN SPRINGER EBOOKS (PP. 299–327). HTTPS://DOI.ORG/10.1007/978-3-319-57413-4_17

THANK YOU