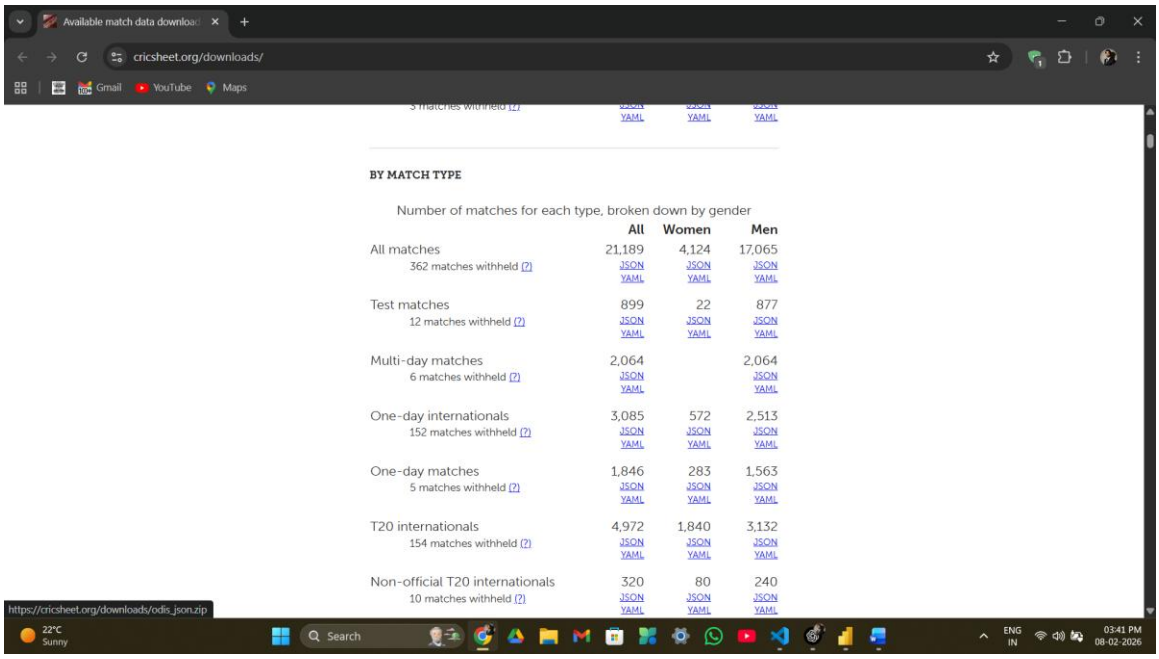
Cricsheet Match Data Analysis Using Python, MySQL & Power BI

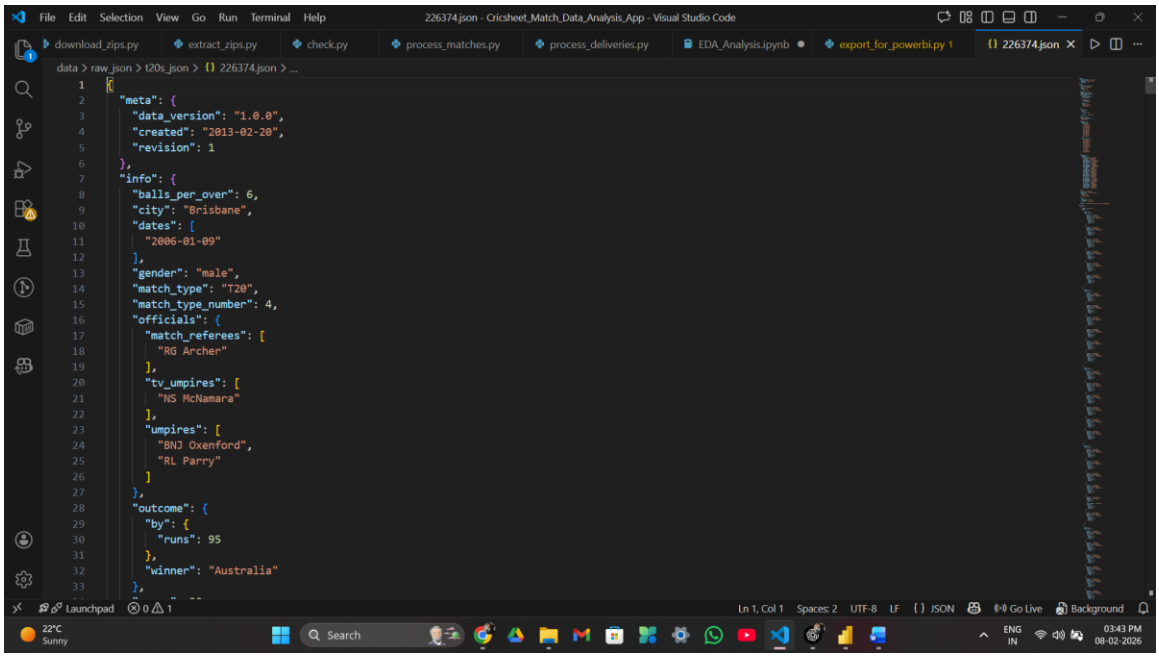
1. Introduction

Cricket is one of the most data-rich sports in the world, generating vast amounts of match statistics, player performances, and ball-by-ball records. With the rise of data analytics, sports organizations and analysts increasingly rely on structured data insights to evaluate player performance, team strategies, and match outcomes.

This project focuses on analyzing historical cricket match data sourced from the Cricsheet dataset. The dataset contains detailed JSON records of matches across multiple formats including Test, ODI, T20, and IPL.

Using modern data analytics tools such as Python, MySQL, and Power BI, this project transforms raw cricket data into meaningful insights through data preprocessing, database modeling, SQL analytics, and interactive dashboards.





2. Problem Statement

Cricket match data is available in semi-structured JSON format, making it difficult to analyze directly. The lack of structured storage and visualization limits the ability to extract insights such as top player performances, team win patterns, and venue trends.

The problem addressed in this project is:

- To collect, process, store, analyze, and visualize cricket match data in a structured format to generate meaningful analytical insights.

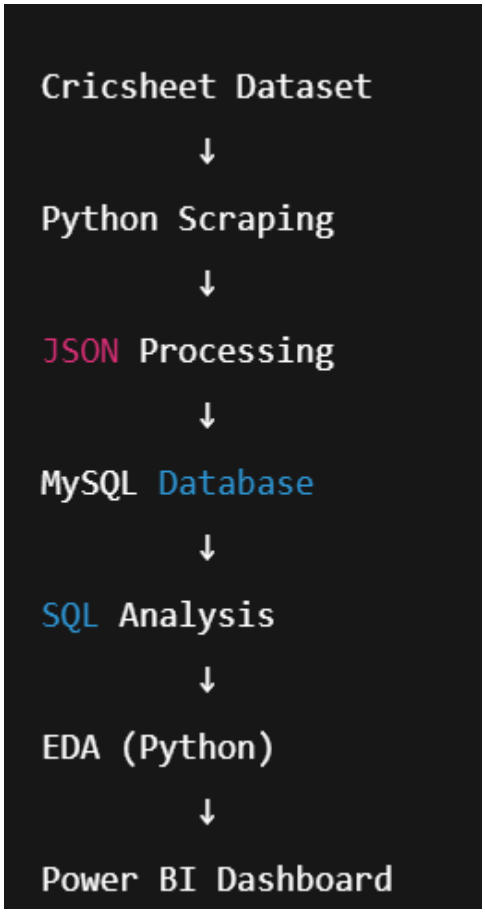
3. Objectives

The key objectives of this project are:

- To scrape cricket match datasets from Cricsheet.
- To preprocess JSON match data using Python.
- To design a relational database for structured storage.
- To perform SQL analytics queries.
- To conduct Exploratory Data Analysis (EDA).
- To build interactive Power BI dashboards.
- To derive insights on teams, players, and match outcomes.

4. Tools & Technologies

Tool	Purpose
Python	Data scraping & preprocessing
Selenium	Automated dataset extraction
Pandas	Data transformation
MySQL	Relational database storage
SQLAlchemy	Python-DB connection
Power BI	Dashboard visualization
VS Code	Development environment



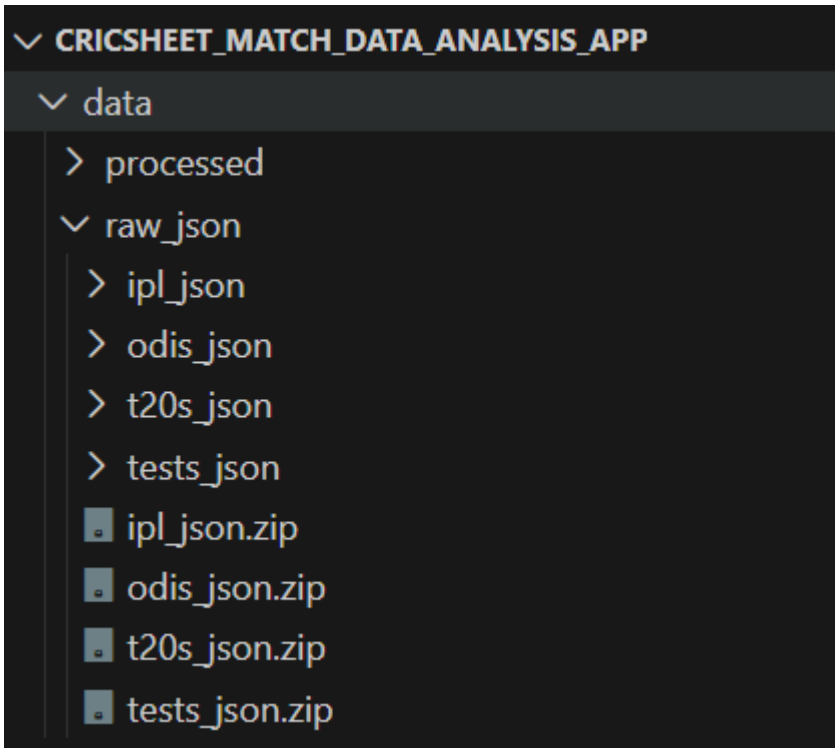
5. Data Collection

The dataset was collected from the Cricsheet website, which provides historical cricket match data in JSON format.

Match formats included:

- Test Matches
- One Day Internationals (ODI)
- T20 Internationals & Domestic T20
- Indian Premier League (IPL)

Bulk JSON archives were downloaded and extracted for processing.



6. Data Preprocessing

Raw JSON files contained nested match information such as:

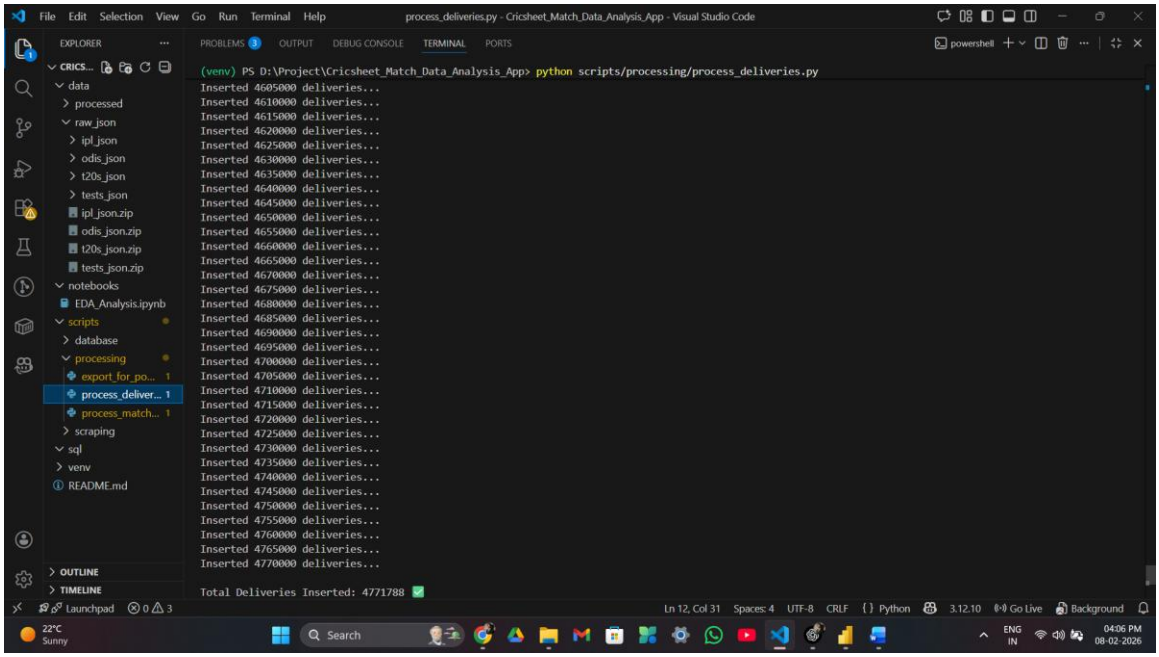
- Teams
- Venue

- Toss details
- Innings
- Ball-by-ball deliveries

Python scripts were developed to:

- Extract match metadata.
- Normalize win margins.
- Handle missing values.
- Skip corrupted files.
- Flatten nested innings structures.

Over 10,000 match records were processed successfully.



7. Database Design

A MySQL relational database was created to store structured data.

Tables Created

1. **Matches Table**

   Stores match-level information:

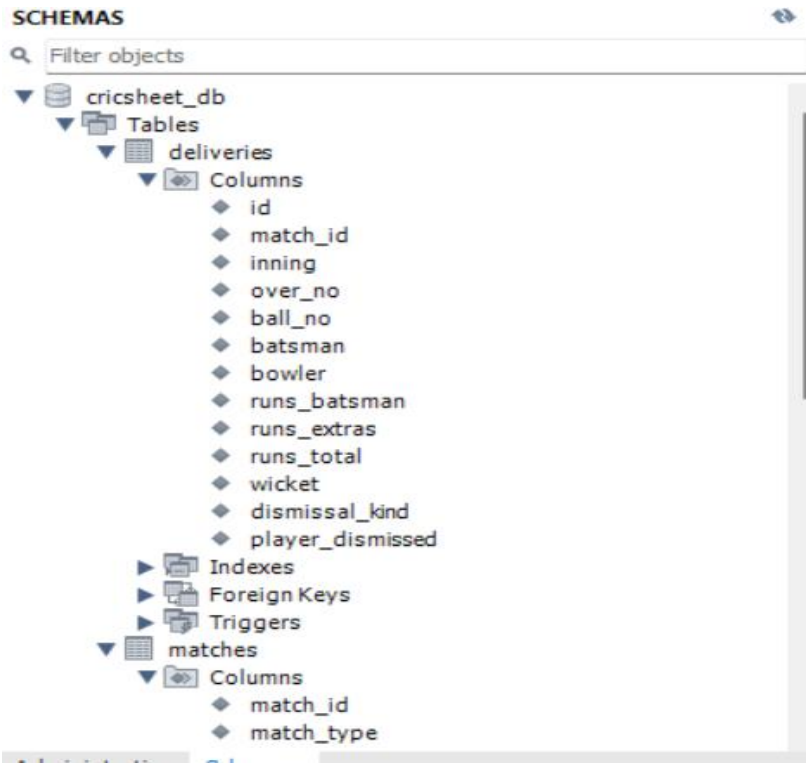
- Match ID
- Teams
- Venue
- Toss decision
- Winner
- Win margin

2. **Deliveries Table**

   Stores ball-by-ball data:

   - Batsman
   - Bowler
   - Runs scored
   - Extras
   - Wickets
   - Over & ball number

Over 4.7 million delivery records were inserted.

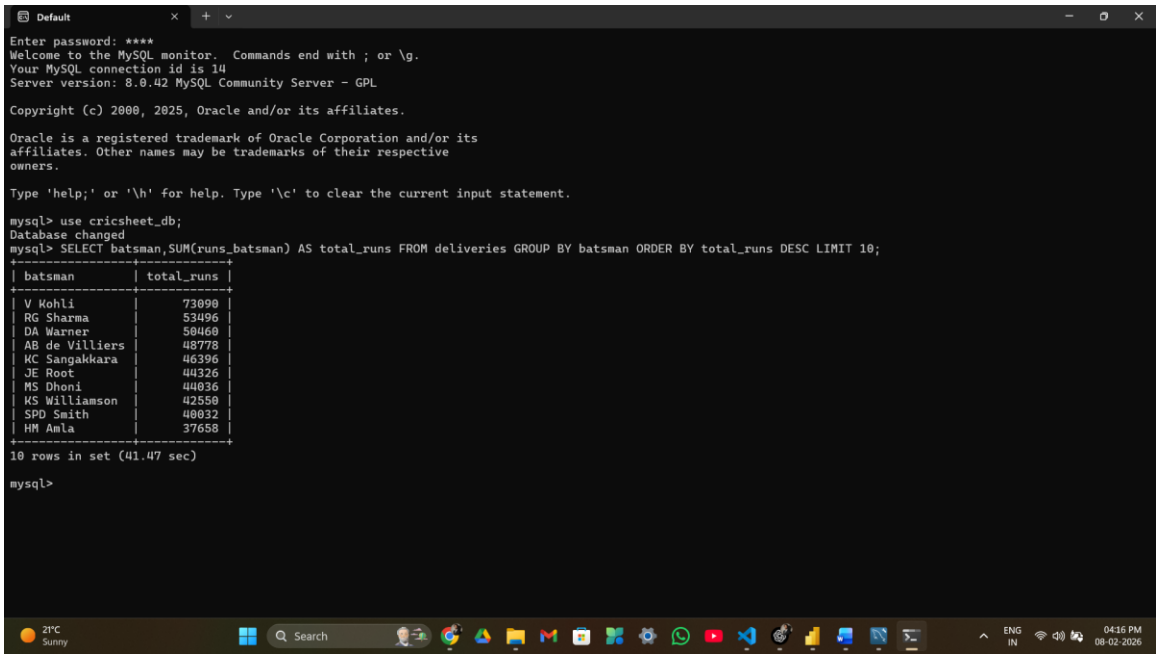


8. SQL Data Analysis

A total of 20 analytical SQL queries were executed to derive insights such as:

- Top run scorers
- Leading wicket takers
- Strike rates
- Economy rates
- Team win counts
- Toss impact
- Venue performance

These queries transformed raw match data into meaningful statistical insights.



9. Exploratory Data Analysis

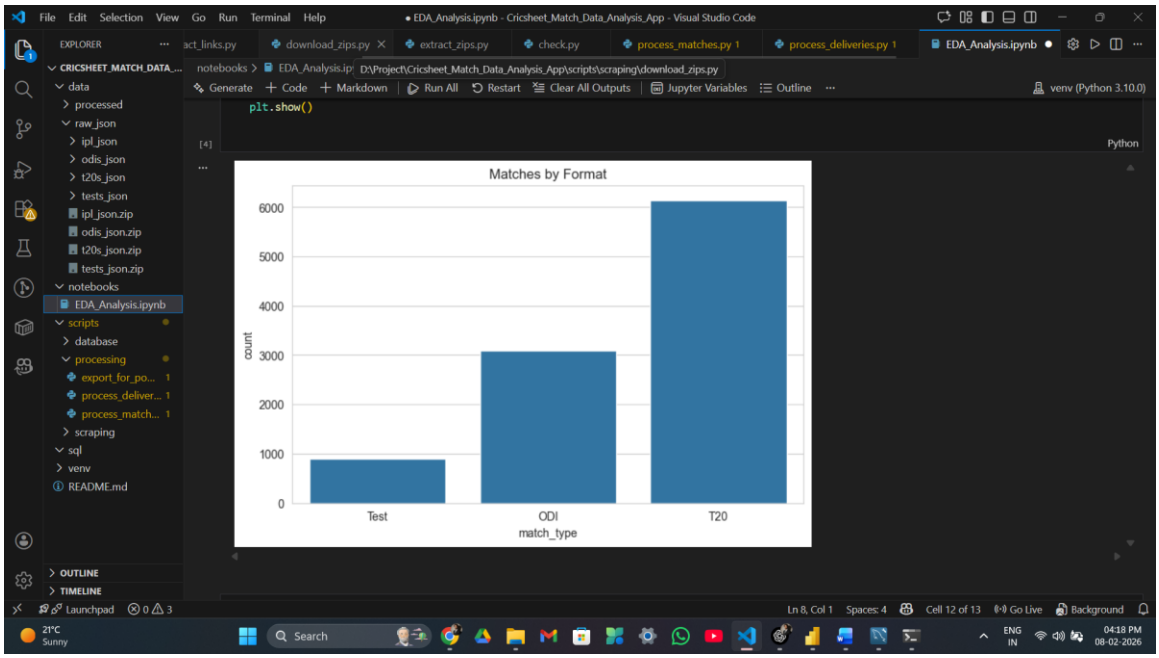
EDA was conducted using Python visualization libraries.

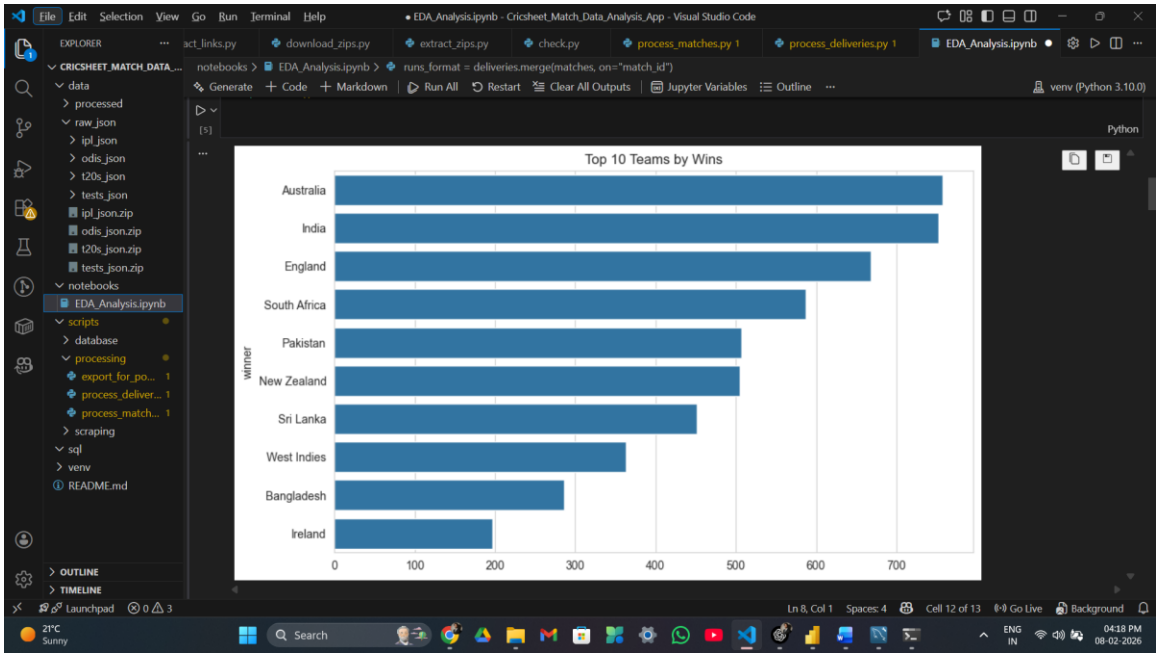
Charts created include:

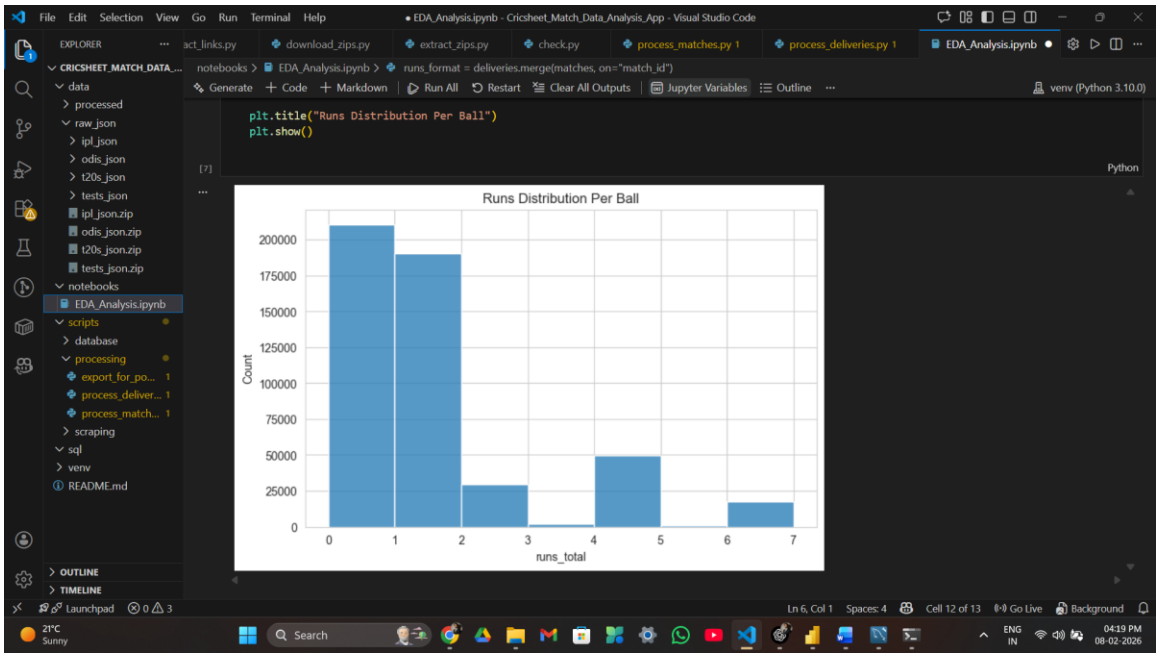
- Matches by format
- Toss decision distribution
- Top batsmen by runs

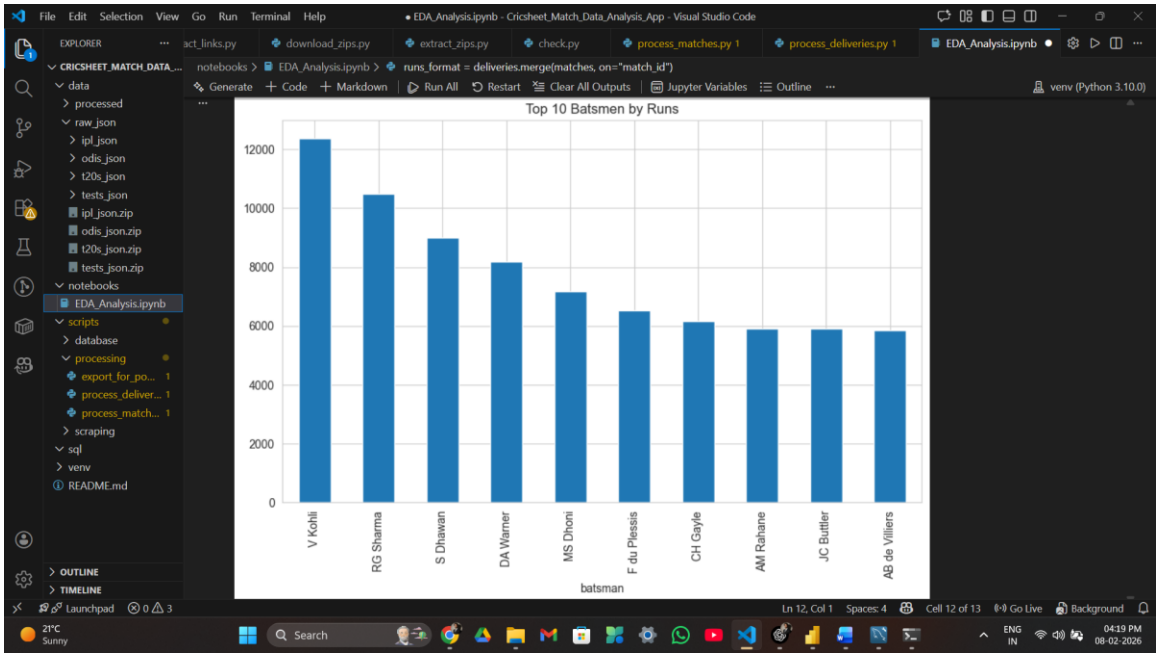
- Leading wicket takers
- Venue analysis
- Win margin types

These visualizations helped identify trends and performance patterns.









10. Power BI Dashboard

An interactive Power BI dashboard was developed with two pages:

Page 1 — Match Overview

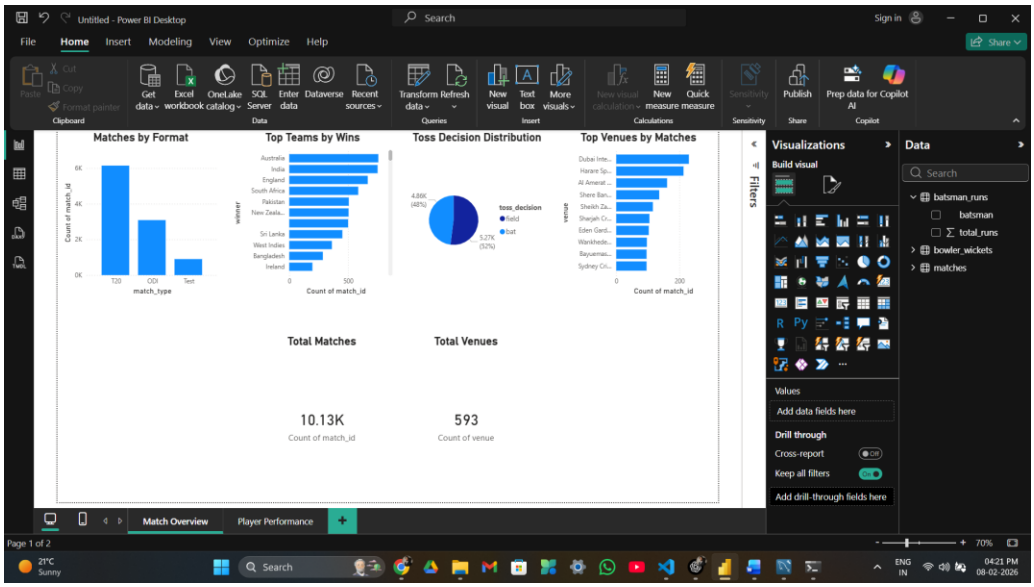
- Matches by format

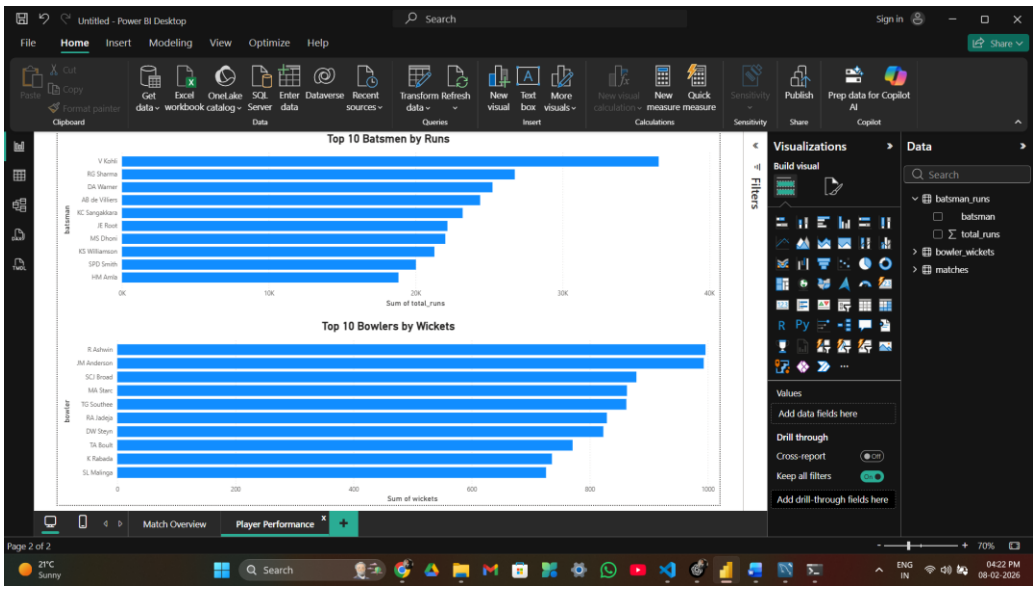
- Toss decision distribution
- Team win leaderboard
- Venue analysis
- KPI cards

Page 2 — Player Performance

- Top 10 batsmen
- Top 10 bowlers

The dashboard allows dynamic exploration of cricket analytics.





11. Key Insights

Some major insights derived include:

- T20 matches dominate modern cricket datasets.
- Fielding first is a common toss decision.
- Certain teams show significantly higher win counts.
- Specific venues host a majority of matches.
- Top players contribute disproportionately to total runs and wickets.

12. Conclusion

This project successfully demonstrates an end-to-end data analytics pipeline, transforming raw cricket JSON datasets into structured insights.

Through scraping, preprocessing, database modeling, SQL analytics, and dashboard visualization, the project highlights the power of data engineering and business intelligence tools in sports analytics.

13. Future Scope

Future enhancements may include:

- Real-time match data integration.
- Player performance prediction models.
- Machine learning-based win forecasting.
- Advanced Power BI drill-through dashboards.