Data Warehousing & Business Intelligence Project Report

Topic: T20 International Cricket Data Analytics

Project Members:

Ashish Gole (amgole2@illinois.edu)

Bhagyesha Patil (brpatil2@illinois.edu)

Shravani Bhupal (sbhupal2@illinois.edu)

Sarth Patel (sarthjp2@illinois.edu)

Introduction:

In a world where cricket is not just a sport but a celebrated global phenomenon, our project delves deep into the heart of T20 International cricket, uncovering the analytical trends and performance insights of players and teams.

This project revolves around the analysis and visualization of T20 cricket data, focusing on key metrics like runs, strike rates, wickets, and match outcomes. The primary goal is to provide comprehensive insights into team and player performance trends, enabling cricket enthusiasts and analysts to understand the dynamics of modern T20 cricket.

Using Power BI, we brought this data to life with interactive dashboards that make it easy to explore and draw conclusions. The project aims to give fans, analysts, and decision-makers a clearer view of the game, helping them appreciate the strategy behind every run scored and wicket taken.

Objective:

For cricket enthusiasts, analysts, and fans who live and breathe the excitement of T20 International cricket, understanding the trends and insights behind every match is both fascinating and challenging.

The core problem addressed in this project is the lack of comprehensive and easy-to-understand data analysis regarding player performances, team trends, and match outcomes in T20 Internationals. This includes exploring strike rates, economy rates, top run-scorers, and wicket-takers to identify patterns across matches. By tackling this challenge, we aim to provide cricket lovers, analysts, and decision-makers with a clear and interactive view of the game's ever-evolving dynamics.

Intended Audience:

The primary audience for this project includes:

- Cricket analysts and enthusiasts seeking detailed insights into team performances and player statistics.
- Sports journalists requiring data-driven visuals and trends to enhance their reporting and analyses.
- Team coaches and managers looking to evaluate and benchmark their team's performance against competitors.

Data Source:

The dataset was extracted from the ESPN Cricinfo website using Bright Data as the scraping platform. JavaScript code was utilized to extract and parse the overall match summary from the site to ensure accurate and up-to-date information.

Link of the dataset -

https://www.espncricinfo.com/records/trophy/icc-men-s-t20-world-cup-89

Methodology:

1. Dataset -

The dataset used contains comprehensive information related to cricket matches and player performances for the year **2022**. It includes **219 players**, representing **16 countries**, and provides detailed match summaries, player roles, and performance metrics.

The dataset comprises **5 different tables** with distinct information, structured as follows:

- **`dim_players_no_images.csv`**:

  Contains player attributes without images.
  Columns: `name`, `team`, `battingStyle`, `bowlingStyle`, `playingRole`, `description`.

- **fact_bating_summary.csv**:

  Provides detailed batting performance statistics for each player.
  Columns: `match`, `teamInnings`, `battingPos`, `batsmanName`, `runs`, `balls`, `4s`, `6s`, `SR`, `out/not_out`, `match_id`.

- **fact_bowling_summary.csv**:

Captures bowling performance data for each player.
Columns: `match`, `bowlingTeam`, `bowlerName`, `overs`, `maiden`, `runs`, `wickets`, `economy`, `0s`, `4s`, `6s`, `wides`, `noBalls`, `match_id`.

dim_match_summary.csv:
Summarizes match-level details and outcomes.Columns: `team1`, `team2`, `winner`, `margin`, `ground`, `matchDate`, `match_id`.

- **dim_players.csv**:

  Contains detailed player attributes, including images.
  Columns: `name`, `team`, `image`, `battingStyle`, `bowlingStyle`, `playingRole`, `description`.

2. Data Loading and Integration -

Loading the Data:

The extracted datasets in JSON format (batting, bowling, match results, and player information) were loaded into the Jupyter Notebook using Python libraries such as `pandas` and `json`. Each dataset was stored in a separate DataFrame for processing.

```
import pandas as pd
```

Load JSON files into DataFrames

```
batting_df = pd.read_json("t20_wc_batting_summary. json") bowling_df =
pd.read_json("t20_wc_bowling_summary.json") match_results_df =
pd.read_json("t20_wc_match_results.json") player_info_df =
pd.read_json("t20_wc_player_info.json")
```

Integration of Data:

The individual datasets were linked based on common keys:

- **Batting and Bowling Summaries** were matched on match and teamInnings fields.
- **Match Results** were merged with performance summaries using team1, team2, and matchDate as keys.
- **Player Information** was linked using the batsmanName or bowlerName fields.

Example of merging datasets

```
final_df = pd.merge(batting_df, match_results_df, left_on="match",
right_on="team1", how="inner")
```

This integration allowed us to create a consolidated dataset combining match results, player performances, and key statistics.

3. Data Preprocessing and Cleaning -

Handling Missing and Null Values:

1. **Identification**:
   a. We checked for missing or null values using:

```
print(batting_df.isnull().sum())

print(bowling_df.isnull().sum())
```

2. **Cleaning**:
   a. Null or missing fields (like dismissal types, strike rates, or boundaries) were replaced with appropriate default values. For instance:

```
batting_df['dismissal'].fillna("Not Out", inplace=True)

bowling_df.fillna(0, inplace=True)
```

Duplicate Records:

To avoid redundancy, duplicate rows were identified and removed:

```
batting_df.drop_duplicates(inplace=True)

bowling_df.drop_duplicates(inplace=True)
```

4. Data Transformation -

Derived Metrics:

1. **Batting Data:**
   a. Calculated missing **Strike Rate (SR)**:

Strike Rate = $\left(\frac{Runs\ Scored}{Balls\ Faced}\right) \cdot 100$

```
batting_df['SR'] = (batting_df['runs'].astype(float) /
batting_df['balls'].astype(float)) * 100
batting_df['SR'].fillna(0, inplace=True)
```

b. Derived **Total Boundaries** by summing the number of 4s and 6s:

```
batting_df['Total Boundaries'] = batting_df['4s'] +
batting_df['6s']
```

2. **Bowling Data:**

a. Calculated **Economy Rate** for bowlers:

$$\texttt{Economy Rate} = \left(\frac{Total\ Runs\ Conceded}{Overs\ Bowled}\right)$$

```
bowling_df['economy'] = bowling_df['runs'] /
bowling_df['overs']
```

3. **Match Results:**

a. Converted matchDate to datetime format for consistency:

```
match_results_df['matchDate'] =
pd.to_datetime(match_results_df['matchDate'])
```

4. **Player Roles:**

a. Enriched batting and bowling data with player roles by merging with player_info_df:

```
enriched_df = pd.merge(batting_df, player_info_df,
left_on="batsmanName", right_on="name", how="left")
```

5. Consolidating the Data -

The cleaned and transformed datasets were combined to create a single, structured DataFrame. This consolidated dataset provided:

- Match results (winner, margin, ground, date).
- Batting summaries (runs, balls, SR, boundaries).
- Bowling summaries (overs, wickets, economy rate).
- Player information (playing roles, styles).

```
final_df = pd.merge(batting_df, bowling_df, on="match",
how="inner")

final_df = pd.merge(final_df, match_results_df, left_on="match",
right_on="team1", how="inner")
```

6. Exporting Cleaned Data to CSV -

Once the preprocessing, cleaning, and feature engineering were completed, the consolidated dataset was exported to a CSV file for analysis and visualization purposes:

```
final_df.to_csv("cleaned_t20_data.csv", index=False)

print("Data successfully exported to
cleaned_t20_data.csv")
```

Final Data:

1. **dim_players_no_images.csv**:
   a. name
   b. team
   c. battingStyle
   d. bowlingStyle
   e. playingRole
   f. Description

2. **fact_bating_summary.csv**:

    a. match
    b. teamInnings
    c. battingPos
    d. batsmanName
    e. runs
    f. balls
    g. 4s
    h. 6s
    i. SR
    j. out/not_out
    k. match_id

3. **fact_bowling_summary.csv**:
    a. match
    b. bowlingTeam
    c. bowlerName
    d. overs
    e. maiden
    f. runs
    g. wickets
    h. economy
    i. 0s
    j. 4s
    k. 6s
    l. wides
    m. noBalls
    n. match_id

4. **dim_match_summary.csv**:
    a. team1
    b. team2
    c. winner
    d. margin
    e. ground
    f. matchDate
    g. match_id

5. **dim_players.csv**:
    a. name
    b. team
    c. image
    d. battingStyle
    e. bowlingStyle
    f. PlayingRole
    g. Description

7. Summary of Cleaning and Transformation Steps -

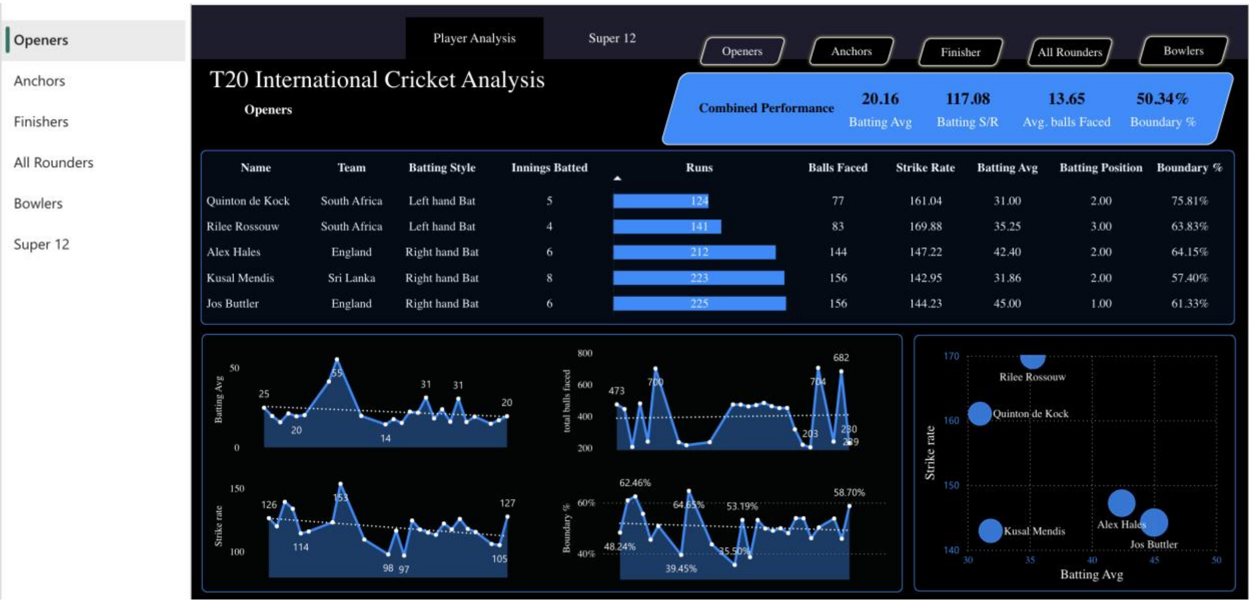
STEPS	DESCRIPTION

Data Loading	Loaded JSON files into DataFrames using pandas.
Handling Missing Value	Replaced null or missing data with appropriate default values.
Duplicate Record	Removed duplicate rows to ensure data consistency.
Derived Metrics	Calculated strike rates, economy rates, and total boundaries.
Data Merging	Combined batting, bowling, and match results with player information.
Data Standardization	Converted `matchDate` fields to `datetime` format for consistency.
Exporting the data	Saved the final cleaned and consolidated data into a CSV file.

8. Visualization -

For the visualization, we used PowerBI Web.
Six distinct interactive dashboards were created using PowerBI.

- **Openers:** The dashboard represents **T20 International Cricket Analysis** for *Openers*, showcasing individual player performance metrics such as runs, strike rate, batting average, and boundary percentage, along with combined performance highlights and visual comparisons.
  - **Player Stats Table**: Runs, strike rate, boundary %, and batting average.
  - **Line Charts**: Batting average, strike rate, total balls faced, and boundary %.
  - **Scatter Plot**: Batting average vs. strike rate.



T20 International Cricket Analysis

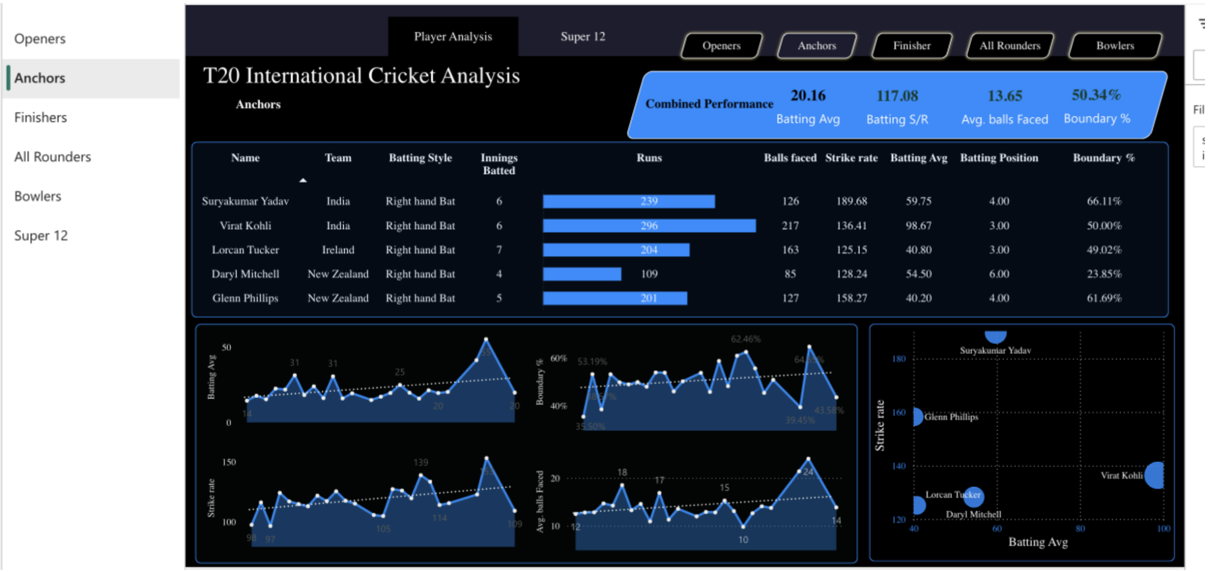
Openers

Combined Performance: 20.16 Batting Avg; 117.08 Batting S/R; 13.65 Avg. balls Faced; 50.34% Boundary %

Name	Team	Batting Style	Innings Batted	Runs	Balls Faced	Strike Rate	Batting Avg	Batting Position	Boundary %
Quinton de Kock	South Africa	Left hand Bat	5	124	77	161.04	31.00	2.00	75.81%
Rilee Rossouw	South Africa	Left hand Bat	4	141	83	169.88	35.25	3.00	63.83%
Alex Hales	England	Right hand Bat	6	212	144	147.22	42.40	2.00	64.15%
Kusal Mendis	Sri Lanka	Right hand Bat	8	223	156	142.95	31.86	2.00	57.40%
Jos Buttler	England	Right hand Bat	6	225	156	144.23	45.00	1.00	61.33%

- **Anchors:** The dashboard represents **T20 International Cricket Analysis** for *Anchors*, showcasing player performance metrics such as runs, strike rate, batting average, and boundary percentage, along with combined performance highlights and visual trends through charts and comparisons.

  - **Player Stats Table**: Runs, batting average, strike rate, and boundary %.
  - **Line Charts**: Batting average, strike rate, and boundary %.
  - **Scatter Plot**: Batting average vs. strike rate.



T20 International Cricket Analysis

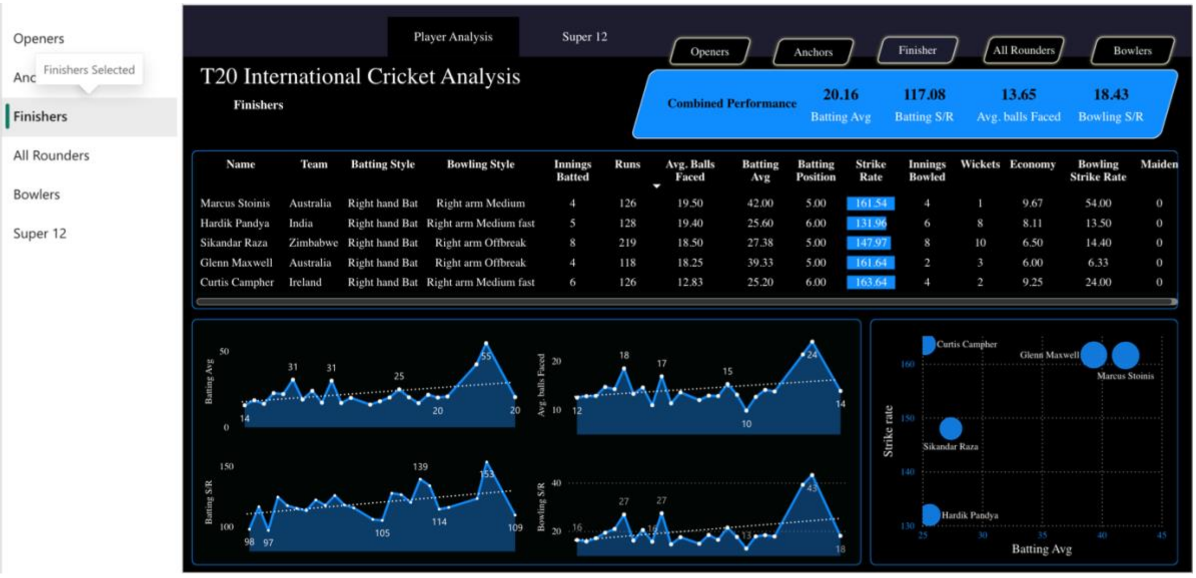
Anchors

Combined Performance 20.16 Batting Avg 117.08 Batting S/R 13.65 Avg. balls Faced 50.34% Boundary %

Name	Team	Batting Style	Innings Batted	Runs	Balls faced	Strike rate	Batting Avg	Batting Position	Boundary %
Suryakumar Yadav	India	Right hand Bat	6	239	126	189.68	59.75	4.00	66.11%
Virat Kohli	India	Right hand Bat	6	296	217	136.41	98.67	3.00	50.00%
Lorcan Tucker	Ireland	Right hand Bat	7	204	163	125.15	40.80	3.00	49.02%
Daryl Mitchell	New Zealand	Right hand Bat	4	109	85	128.24	54.50	6.00	23.85%
Glenn Phillips	New Zealand	Right hand Bat	5	201	127	158.27	40.20	4.00	61.69%

- **Finishers:** The dashboard represents **T20 International Cricket Analysis** for *Finishers*, displaying player performance metrics for both batting and bowling, including runs, strike rate, wickets, economy, and bowling strike rate. It also features combined performance highlights, line charts for trends, and a scatter plot for batting average vs. strike rate.

  - **Player Stats Table**: Runs, strike rate, bowling stats (wickets, economy, and bowling strike rate).
  - **Line Charts**: Batting average, strike rate, average balls faced, and bowling strike rate.
  - **Scatter Plot**: Batting average vs. strike rate.



T20 International Cricket Analysis

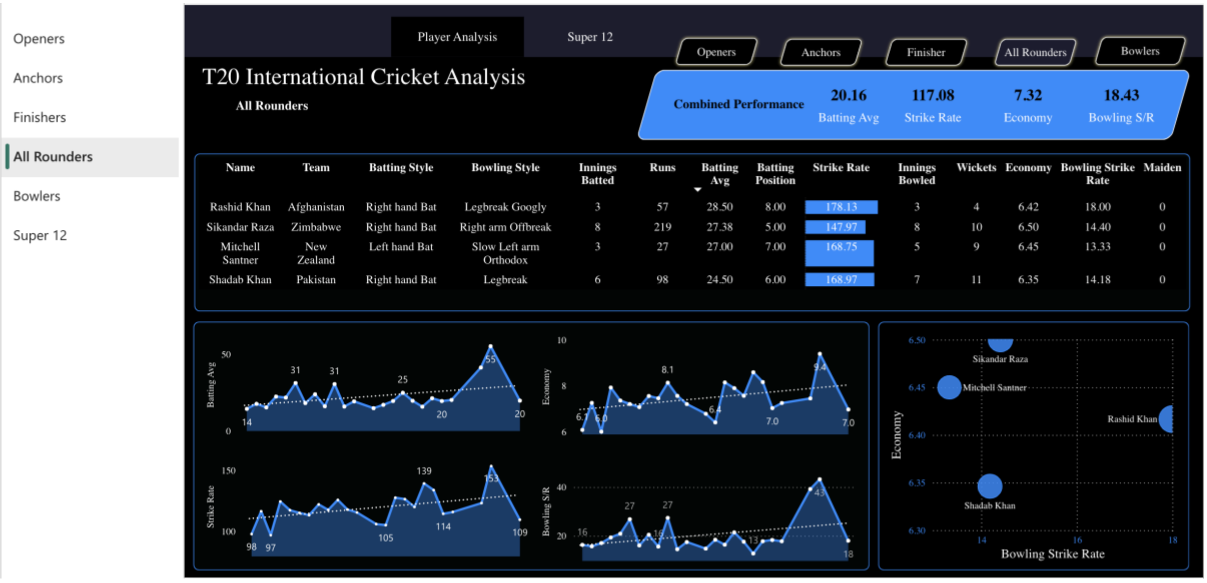
Finishers

Combined Performance: 20.16 Batting Avg | 117.08 Batting S/R | 13.65 Avg. balls Faced | 18.43 Bowling S/R

Name	Team	Batting Style	Bowling Style	Innings Batted	Runs	Avg. Balls Faced	Batting Avg	Batting Position	Strike Rate	Innings Bowled	Wickets	Economy	Bowling Strike Rate	Maiden
Marcus Stoinis	Australia	Right hand Bat	Right arm Medium	4	126	19.50	42.00	5.00	161.54	4	1	9.67	54.00	0
Hardik Pandya	India	Right hand Bat	Right arm Medium fast	5	128	19.40	25.60	6.00	131.96	6	8	8.11	13.50	0
Sikandar Raza	Zimbabwe	Right hand Bat	Right arm Offbreak	8	219	18.50	27.38	5.00	147.97	8	10	6.50	14.40	0
Glenn Maxwell	Australia	Right hand Bat	Right arm Offbreak	4	118	18.25	39.33	5.00	161.64	2	3	6.00	6.33	0
Curtis Campher	Ireland	Right hand Bat	Right arm Medium fast	6	126	12.83	25.20	6.00	163.64	4	2	9.25	24.00	0

- **All Rounders:** The dashboard represents **T20 International Cricket Analysis** for *All Rounders*, displaying batting and bowling performance metrics such as runs, strike rate, wickets, economy, and bowling strike rate. It includes combined performance highlights, trend line charts, and a scatter plot for economy vs bowling strike rate.

  - **Player Stats Table**: Runs, batting average, strike rate, wickets, economy, and bowling strike rate.
  - **Line Charts**: Batting average, strike rate, economy, and bowling strike rate.
  - **Scatter Plot**: Economy vs. bowling strike rate.



T20 International Cricket Analysis

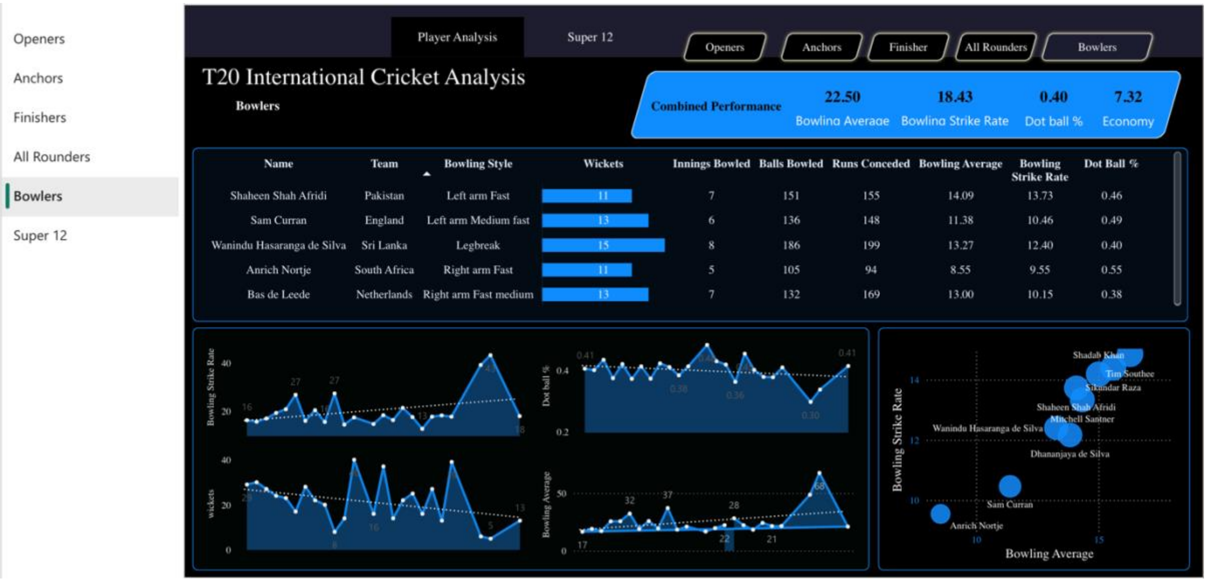
All Rounders

Combined Performance: 20.16 Batting Avg | 117.08 Strike Rate | 7.32 Economy | 18.43 Bowling S/R

Name	Team	Batting Style	Bowling Style	Innings Batted	Runs	Batting Avg	Batting Position	Strike Rate	Innings Bowled	Wickets	Economy	Bowling Strike Rate	Maiden
Rashid Khan	Afghanistan	Right hand Bat	Legbreak Googly	3	57	28.50	8.00	178.13	3	4	6.42	18.00	0
Sikandar Raza	Zimbabwe	Right hand Bat	Right arm Offbreak	8	219	27.38	5.00	147.97	8	10	6.50	14.40	0
Mitchell Santner	New Zealand	Left hand Bat	Slow Left arm Orthodox	3	27	27.00	7.00	168.75	5	9	6.45	13.33	0
Shadab Khan	Pakistan	Right hand Bat	Legbreak	6	98	24.50	6.00	168.97	7	11	6.35	14.18	0

- **Bowlers:** The dashboard represents **T20 International Cricket Analysis** for *Bowlers*, highlighting key metrics such as wickets, bowling average, strike rate, dot ball percentage, and economy. It includes combined performance, trend line charts for bowling statistics, and a scatter plot comparing bowling average and bowling strike rate.

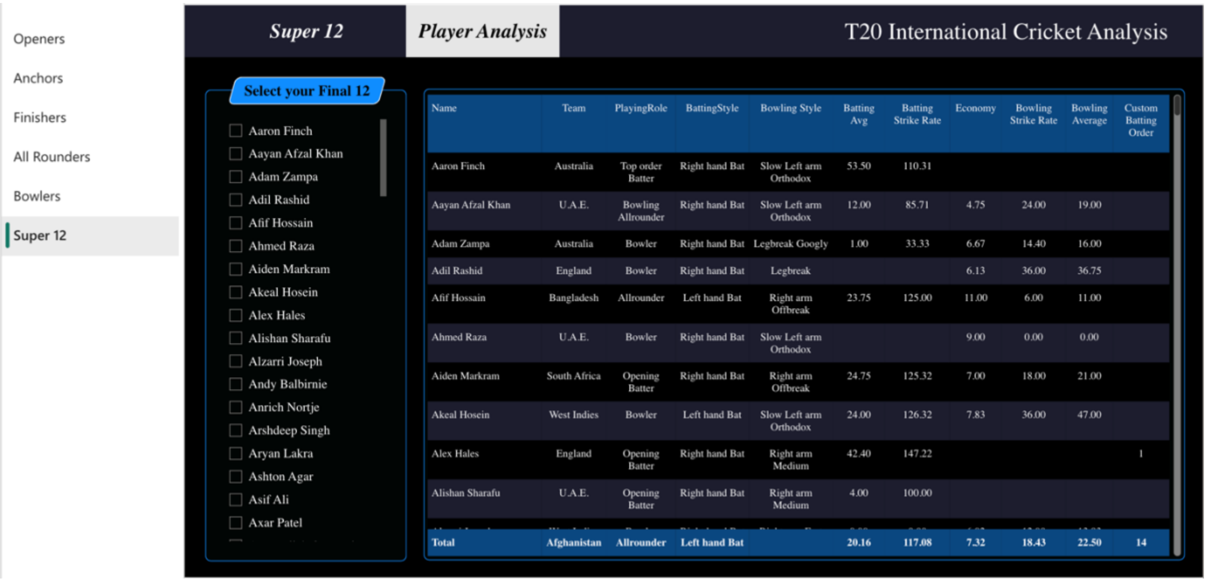
- **Player Stats Table**: Wickets, bowling average, economy, bowling strike rate, and dot ball %.
- **Line Charts**: Bowling strike rate, dot ball %, wickets, and bowling average.
- **Scatter Plot**: Bowling average vs. bowling strike rate.



Name	Team	Bowling Style	Wickets	Innings Bowled	Balls Bowled	Runs Conceded	Bowling Average	Bowling Strike Rate	Dot Ball %
Shaheen Shah Afridi	Pakistan	Left arm Fast	11	7	151	155	14.09	13.73	0.46
Sam Curran	England	Left arm Medium fast	13	6	136	148	11.38	10.46	0.49
Wanindu Hasaranga de Silva	Sri Lanka	Legbreak	15	8	186	199	13.27	12.40	0.40
Anrich Nortje	South Africa	Right arm Fast	11	5	105	94	8.55	9.55	0.55
Bas de Leede	Netherlands	Right arm Fast medium	13	7	132	169	13.00	10.15	0.38

- **Super 12:** The dashboard represents **T20 International Cricket Analysis** for *Super 12*, allowing users to select their final 12 players. It displays detailed player stats such as batting average, strike rate, economy, bowling strike rate, and bowling average, helping in team selection and performance comparison.

  - **Player Stats Table**: Batting average, strike rate, economy, bowling strike rate, and bowling average.
  - **Selection Panel**: Allows users to select their final 12 players.



Select your Final 12

- ☐ Aaron Finch
- ☐ Aayan Afzal Khan
- ☐ Adam Zampa
- ☐ Adil Rashid
- ☐ Afif Hossain
- ☐ Ahmed Raza
- ☐ Aiden Markram
- ☐ Akeal Hosein
- ☐ Alex Hales
- ☐ Alishan Sharafu
- ☐ Alzarri Joseph
- ☐ Andy Balbirnie
- ☐ Anrich Nortje
- ☐ Arshdeep Singh
- ☐ Aryan Lakra
- ☐ Ashton Agar
- ☐ Asif Ali
- ☐ Axar Patel

Name	Team	PlayingRole	BattingStyle	Bowling Style	Batting Avg	Batting Strike Rate	Economy	Bowling Strike Rate	Bowling Average	Custom Batting Order
Aaron Finch	Australia	Top order Batter	Right hand Bat	Slow Left arm Orthodox	53.50	110.31				
Aayan Afzal Khan	U.A.E.	Bowling Allrounder	Right hand Bat	Slow Left arm Orthodox	12.00	85.71	4.75	24.00	19.00	
Adam Zampa	Australia	Bowler	Right hand Bat	Legbreak Googly	1.00	33.33	6.67	14.40	16.00	
Adil Rashid	England	Bowler	Right hand Bat	Legbreak			6.13	36.00	36.75	
Afif Hossain	Bangladesh	Allrounder	Left hand Bat	Right arm Offbreak	23.75	125.00	11.00	6.00	11.00	
Ahmed Raza	U.A.E.	Bowler	Right hand Bat	Slow Left arm Orthodox			9.00	0.00	0.00	
Aiden Markram	South Africa	Opening Batter	Right hand Bat	Right arm Offbreak	24.75	125.32	7.00	18.00	21.00	
Akeal Hosein	West Indies	Bowler	Left hand Bat	Slow Left arm Orthodox	24.00	126.32	7.83	36.00	47.00	
Alex Hales	England	Opening Batter	Right hand Bat	Right arm Medium	42.40	147.22				1
Alishan Sharafu	U.A.E.	Opening Batter	Right hand Bat	Right arm Medium	4.00	100.00				
Total	Afghanistan	Allrounder	Left hand Bat		20.16	117.08	7.32	18.43	22.50	14

Each report combines **tables**, **line charts**, and **scatter plots** to provide an in-depth analysis of player performances.

Findings and Insights:

The T20 International Cricket Analytics Dashboard provides a comprehensive analysis of player performances across roles—Openers, Anchors, Finishers, All Rounders, and Bowlers. It features key metrics such as batting average, strike rate, economy, and bowling strike rate, supported by tables, line charts, and scatter plots. The dashboard highlights individual strengths, role-specific contributions, and enables data-driven decisions for team selection and performance optimization.

Overall Dashboard Insights:

1. **Role-based Performance:**
   - Openers focus on fast starts with high strike rates.
   - Anchors stabilize innings with better averages and strike rotation.
   - Finishers accelerate at the end with high strike rates.

2. **All-Round Versatility:**
   - Players like *Sikandar Raza* and *Shadab Khan* highlight the importance of all-rounders in maintaining team balance.

3. **Bowling Dominance:**
   - Bowlers' strike rates, dot ball percentages, and economy rates are key indicators of success in restricting runs and taking wickets.

4. **Top Performers:**
   - *Virat Kohli*, *Marcus Stoinis*, *Sam Curran*, and *Wanindu Hasaranga* emerge as standout players in their respective roles.

5. **Combined Performance:**
   - Metrics like batting average, strike rate, bowling average, and economy provide a holistic view of player contributions.

Key Takeaways:

The dashboard delivers a comprehensive analysis of player performances, enabling data-driven decisions for team selection, strategy optimization, and identifying strengths and weaknesses in various roles (batting, bowling, and all-round contributions).

Challenges and Solutions:

During the process of data extraction, transformation, and cleaning, several challenges can arise that need careful attention. One major challenge is dealing with **dynamic content** on websites like ESPN Cricinfo, where data is loaded via JavaScript. This requires advanced tools such as **Bright Data** or Selenium to extract the data reliably while adhering to website scraping policies.

Another issue involves handling **missing or inconsistent values** within datasets. For instance, fields like runs, balls, or strike rates may be null, requiring careful imputation techniques such as replacing missing values with zeros or calculating derived metrics like strike rates and economy rates using logical formulas. Additionally, **duplicate records** can inflate the dataset size and skew analysis results, which necessitates identifying and removing them using deduplication techniques.

Data integration poses another layer of complexity, as combining multiple datasets (e.g., batting, bowling, and match results) may lead to mismatches in keys such as player names, match identifiers, or team names. Standardizing these keys and ensuring alignment is critical for accurate merging. Transforming data into useful insights introduces challenges such as computing derived metrics (e.g., total boundaries, player strike rates, or bowling economy) while managing computational efficiency for large datasets. Performance concerns like memory consumption or slow processing can arise when working with JSON files containingthousands of records, and these can be mitigated using optimized tools like **Pandas**, chunk processing, or libraries like **Dask**.

Furthermore, **outliers** and extreme values can distort analytical results, requiring proper validation and statistical checks to maintain data accuracy. By systematically addressing these challenges—ensuring clean, consistent, and reliable data through preprocessing, deduplication, validation, and integration—the extracted data can be transformed into a structured format suitable for insightful analysis and visualization.

Feedback:

Throughout the project, we consulted two seniors from the MSIM program who had previously completed the IS525 course and had significant expertise in data visualization. The feedback they provided included:

- Feedback: Graph placement for improved visualization.
- Action: We repositioned the graphs from the top to the bottom to enhance visual clarity and effectiveness.

Future Scope:

Following are the features that can be considered as a future addition to this project:

- **Integration of Real-Time Data:** Real-time match data can be integrated into the existing Power BI dashboard to provide live updates on player statistics, team performances, and match outcomes. This feature would allow fans, analysts, and team management to monitor games as they happen, enabling on-the-spot analysis and decision-making. Real-time data can also improve the relevance and accuracy of insights during live matches.
- **Enhanced Player Performance Analysis:** Deeper analytics can be introduced to analyze player performances across multiple dimensions, such as strike rates under pressure, performances in different weather conditions, or consistency across venues and opponents. By incorporating metrics like player fitness, workload, and recent form, teams

and analysts can gain comprehensive insights to make informed decisions for team selection and strategy.

- **Incorporation of Sentiment Analysis:** By integrating social media platforms like Twitter and Instagram, sentiment analysis can capture fan reactions and opinions about players, teams, and matches. Using natural language processing (NLP) techniques, the project can analyze trends such as fan enthusiasm, criticism, or appreciation. This feature would provide stakeholders with a comprehensive view of the emotional and social aspects of the game.

Conclusion:

This project provided a deep dive into the exciting world of T20 International cricket, uncovering valuable insights into player and team performances for the year **2022**. With data covering **219 players** from **16 countries**, we meticulously processed and analyzed information across **five tables**, ensuring it was clean, consistent, and ready for meaningful analysis.

Through this work, we highlighted key aspects of the game—runs, strike rates, wickets, economy rates, and more—shedding light on how players excel in their roles, whether as Openers, Anchors, Finishers, All-Rounders, or Bowlers. By using interactive **Power BI dashboards**, we brought this data to life, allowing cricket fans, analysts, and decision-makers to explore performance trends and identify standout players with ease.

The project emphasizes how data can help us understand the strategy behind every run scored and wicket taken, showcasing the dynamic nature of modern cricket. Looking ahead, features like **real-time match data** and **sentiment analysis** can take this analysis even further, enabling live insights and connecting the numbers to the emotions of fans worldwide. This work not only celebrates the sport but also opens the door for smarter, data-driven decisions in cricket.