**🏏 Cricket Data Analysis – ICC T20 World Cup 2022**

**(Project Sportan)**

**📖 Introduction**

Cricket is not just a sport — it’s an emotion followed by billions of fans across the globe. With the rise of shorter formats like T20, the game has become more dynamic, data-driven, and performance-centric. Teams now rely heavily on **data analysis** to plan strategies, select playing elevens, and optimize match outcomes.

**Objective of the Project**

The **ICC T20 World Cup 2022** concluded with thrilling performances across teams. Inspired by this, we imagine a fun scenario:

**🛸 *Aliens have challenged Earth to a cricket match!  
If they win, the planet is theirs… if we win, they leave us forever.***

**Performance Analysis** – To clean, transform, and analyze ICC T20 World Cup 2022 data using Python, Pandas, and Power BI. This involves evaluating batting, bowling, and team performance through key metrics like Strike Rate, Economy, Batting Average, and Bowling Average.

Using insights from our analysis, we aim to identify the **Best Playing XI** across all countries — Earth’s strongest team to take on the challenge.

## 📂 Dataset Description

### **🔹 Source of Data**

* The raw data was collected from the **ESPN Cricinfo website** in JSON format.
* Since direct web scraping tools were not available, JSON files were downloaded with the help of **Codebasics.io** resources.

### **🔹 Key Tables Created**

After cleaning and transforming the JSON files, the following structured tables were prepared:

1. **Fact\_Batting\_Summary**
   * Columns: match, team Innings, batting Pos, batsman Name, runs, balls, 4s, 6s, SR, out/not out
2. **Fact\_Bowling\_Summary**
   * Columns: match, team Innings, bowler Name, overs, maidens, runs, wickets, 4s, 6s, zeros, no balls
3. **Match\_Summary**
   * Columns: team1, team2, winner, margin, ground, match Date, match\_id, stage
4. **Dim\_Players**
   * Columns: name, team, image, batting Style, bowling Style, playing Role

***Data Cleaning & Transformation***

**Using Pandas**

The raw JSON files required multiple cleaning and transformation steps before analysis. Below were the major actions performed:

1. **Created DataFrames from JSON files** using Python (json + pandas).
2. **Renamed inconsistent columns** for clarity (e.g., scorecard → match\_id).
3. **Derived new columns** from existing data:
   * Extracted *dismissal info* to create an out/not\_out column.
4. **Established match identifiers**:
   * Created a match\_id column to link batting, bowling, and match summary datasets using functions like. iterrows() and .map().
5. **Exported cleaned data to CSV**:
   * Final datasets were saved as .csv files to ensure smooth import into **Power BI** for further modeling and transformation.

📎 *Please find attached preprocessing Jupyter Notebook for reference (GitHub Link), where each step has been implemented and documented.*

**Using Power Query**

After importing the CSVs into **Power BI**, additional cleaning was carried out using **Power Query Editor**:

### dim\_players

* Extracted player names before delimiters (e.g., "Rohit Sharma (c)" → "Rohit Sharma").
* Removed duplicate player records.

### match\_summary

* Created a **Conditional Column Stage** based on match dates (e.g., Group Stage, Semi-Finals, Finals).

### fact\_bowling\_summary

* Created a new column balls = overs \* 6 (to standardize bowling stats).
* Transformed dismissal results into numeric: "out" = 1, "not out" = 0.

### Other Fact Tables

* Applied basic transformations such as datatype corrections, column renaming, and null handling

***🔗 Data Modeling***

The final **data model** was designed as a **Star Schema**:

1. match summary ⟶ fact\_batting\_summary (One-to-Many via match\_id)
2. match summary ⟶ fact\_bowling\_summary (One-to-Many via match\_id)
3. dim\_players ⟶ fact\_batting\_summary (One-to-Many via player\_name)
4. dim\_players ⟶ fact\_bowling\_summary (One-to-Many via player\_name)

This model ensures efficient querying and accurate calculation of batting, bowling, and match KPIs.

**📊 DAX Measures Created**

### **🏏 Batting Metrics**

1. **Total Runs**
   * Formula: Total Runs = SUM(fact\_batting\_summary[runs])
   * Meaning: Total runs scored by the batsman.
   * Insight: More runs → better batsman.
2. **Total Innings Batted**
   * Formula: Total Innings Batted = COUNT(fact\_batting\_summary[match\_id])
   * Meaning: Number of innings a batsman got a chance to bat.
   * Insight: More innings → more experience at that position.
3. **Total Innings Dismissed**
   * Formula: COUNTROWS(FILTER(fact\_batting\_summary,fact\_batting\_summary[out]=1))
   * Meaning: Number of innings in which batsman got out.
4. **Batting Average**
   * Formula: Batting Avg = DIVIDE([Total Runs],[Total Innings Dismissed],0)
   * Meaning: Runs scored per dismissal.
   * Insight: Higher average = consistency.
5. **Total Balls Faced**
   * Formula: Total Balls Faced = SUM(fact\_batting\_summary[balls])
   * Meaning: Total number of balls faced.
6. **Strike Rate**
   * Formula: Strike Rate = DIVIDE([Total Runs],[Total Balls Faced],0)\*100
   * Meaning: Runs scored per 100 balls.
   * Insight: Higher SR = more attacking/power-hitter.
7. **Batting Position**
   * Formula: Batting Position = ROUNDUP(AVERAGE(fact\_batting\_summary[batting\_pos]),0)
   * Meaning: Average batting order position.
   * Insight: Shows where the player is usually experienced (e.g., opener, middle order).
8. **Boundary %**
   * Formula: Boundary % = DIVIDE(SUM(fact\_batting\_summary[Boundary runs]),[Total Runs],0)
   * Meaning: % of runs scored from 4s & 6s.
   * Insight: Higher % = power hitter; lower % = anchor.
9. **Avg. Balls Faced**
   * Formula: Avg Balls Faced = DIVIDE([Total Balls Faced],[Total Innings Batted],0)
   * Meaning: Average number of balls faced per innings.

### 🎯 **Bowling Metrics**

1. **Wickets**
   * Formula: Wickets = SUM(fact\_bowling\_summary[wickets])
   * Meaning: Total number of wickets taken.
   * Insight: More wickets → better bowler.
2. **Balls Bowled**
   * Formula: Balls Bowled = SUM(fact\_bowling\_summary[balls])
   * Meaning: Total number of balls bowled.
3. **Runs Conceded**
   * Formula: Runs Conceded = SUM(fact\_bowling\_summary[runs])
   * Meaning: Total runs given away.
   * Insight: Fewer runs → more effective bowler.
4. **Bowling Economy**
   * Formula: Economy = DIVIDE([Runs Conceded],([Balls Bowled]/6),0)
   * Meaning: Runs conceded per over.
   * Insight: Lower economy = bowler is controlling runs.
5. **Bowling Strike Rate**
   * Formula: Bowling SR = DIVIDE([Balls Bowled],[Wickets],0)
   * Meaning: Balls bowled per wicket.
   * Insight: Lower SR = bowler takes wickets more often.
6. **Bowling Average**
   * Formula: Bowling Avg = DIVIDE([Runs Conceded],[Wickets],0)
   * Meaning: Runs conceded per wicket taken.
   * Insight: Lower average = more efficient bowler.

## *🏏 Selection Criteria for Best XI*

To compete with the aliens, we defined a balanced team that can consistently:

* **Score ~180 runs** when batting first.
* **Defend 150 runs** when bowling, ensuring a margin of ~30 runs.

Based on ICC T20 WC 2022 data, the following **role-based criteria** were applied:

### **🔹 Openers (2 Players – Power Hitters)**

* Batting Average > **30**
* Strike Rate > **140**
* Innings Batted > **3**
* Boundaries % > **50%**
* Batting Position < **4**  
  👉 Need for: Aggressive start, quick scoring in powerplay.

### **🔹 Middle Order Anchors (3 Players)**

* Batting Average > **40**
* Strike Rate > **125**
* Innings Batted > **3**
* Avg. Balls Faced > **20**
* Batting Position > **2**  
  👉 Need for: Provide stability, rotate strike, and build partnerships.

### **🔹 Finisher (1 Player – Lower Order Anchor)**

* Batting Average > **25**
* Strike Rate > **130**
* Innings Batted > **3**
* Avg. Balls Faced > **12**
* Batting Position > **4**
* Innings Bowled > **1** (secondary bowling ability)  
  👉 Need for : Close out innings with impactful hitting + contribute with the ball.

### **🔹 All-Rounders (2 Players – Lower Order Batsmen + Bowlers)**

* Batting Average > **15**
* Strike Rate > **140**
* Innings Batted > **2**
* Batting Position > **4**
* Innings Bowled > **2**
* Bowling Economy < **7**
* Bowling Strike Rate < **20**  
  👉 Need for : Flexible balance — contribute with both bat and ball.

### **🔹 Specialist Fast Bowlers (3 Players)**

* Innings Bowled > **4**
* Bowling Economy < **7**
* Bowling Strike Rate < **16**
* Bowling Average < **20**
* Dot Ball % > **40%**  
  👉 Need for : Bowling attack with control (more dot balls) and wicket-taking ability.

# 📊 ***Visualizations and Dashboard***

The Power BI Dashboard was designed with **two major tabs**:

## 🔹 ****Player Analysis Tab****

This section enables detailed analysis of individual players based on their role.  
It contains **five dedicated pages**, accessible through **navigation buttons** at the top:

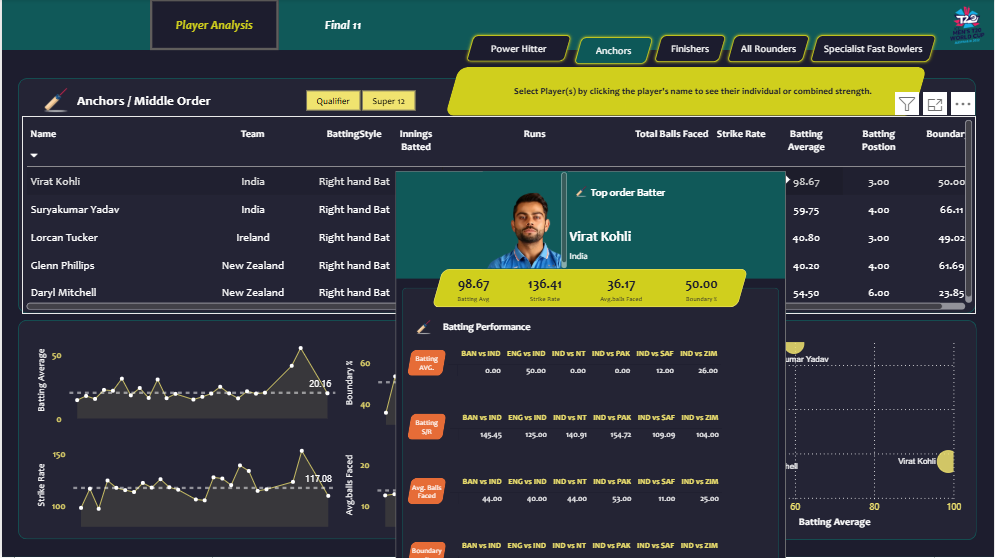
1. **Power Hitters (Openers)**

* Visuals show table including batting average, strike rate, boundary %, innings played. Filters are applied on the visuals using key measures to find the players as per our criteria.
* Line graph showing batting average, strike rate , balls faced , Boundary % of selected players in each match .
* Helps identify the best performing openers.



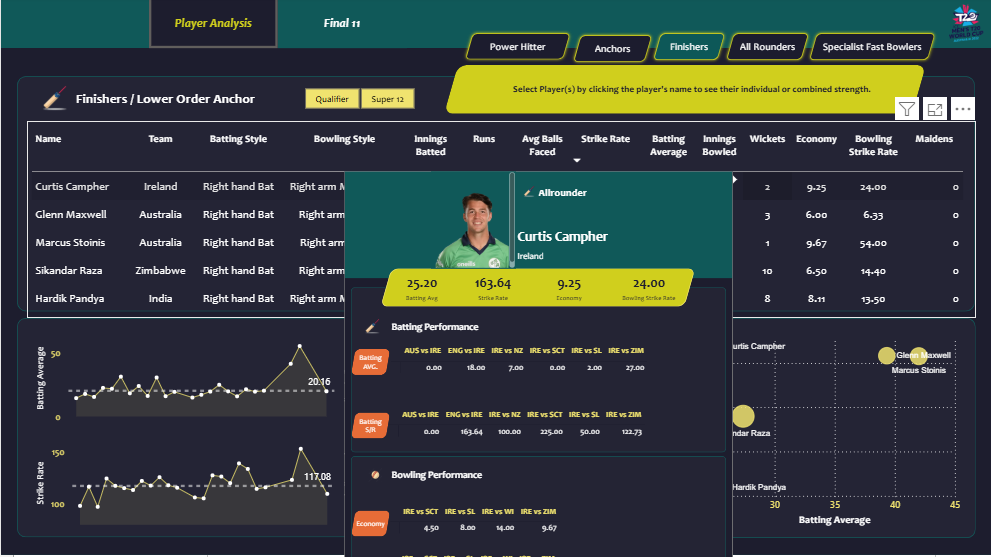
1. **Anchors (Middle Order)**

* Analysis of consistency through batting average, balls faced, and strike rate.
* Tooltips help to see the player role , records in each match and his statistics when we hover over a particular player in the table



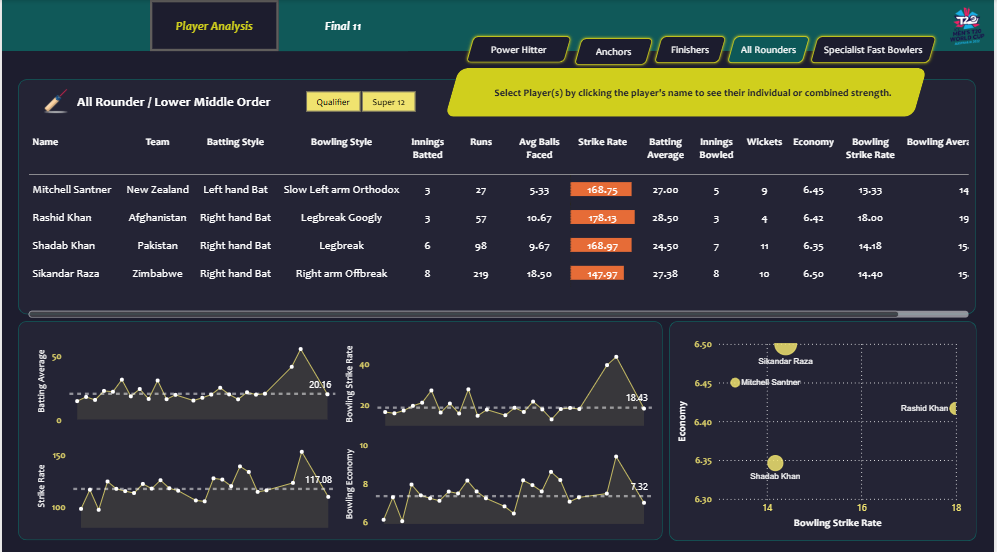
1. **Finishers (Lower Middle Order)**

* Visuals highlight strike rate in death overs, average balls faced, and finishing impact.
* Key for selecting clutch players for the last overs.



1. **All Rounders**

* Combined batting & bowling performance.



1. **Specialist Fast Bowlers**

* Visuals emphasize wickets, bowling average, economy, and dot ball %.



🔹 **Final XI Tab**

This section presents the **team selection outcome**:

* **Final XI Table:**  
  Displays the selected 11 players, role-wise.
* **Team Performance Snapshot:**  
  Shows combined metrics such as: Batting Average , Strike rate , Bowling Average , Economy , Dot Ball %
* **Customizable Slicer (Left Panel):**  
  Users can modify team composition dynamically (e.g., swap players) and instantly view the **recalculated team performance**.

****

*🏏* ***Performance Insights and Best XI Justification***

🔹 **Power Hitters**

* Shortlisted 5 players as per criteria: **Rilee Rossouw, Quinton de Kock, Kusal Mendis, Jos Buttler, Alex Hales** → need to pick 2.
* **Jos Buttler** stands out:
* Highest runs, strong batting average & SR.
* Boundary % = 61% (ideal for powerplay).
* Opening experience at position 1 + adds wicketkeeping option.
* For partner, prefer a **left-right combo**. Options: de Kock (low avg, inconsistent) vs Rossouw (high SR, high avg, high boundary %).
* **Rossouw selected** as Buttler’s partner → their combined stats: Avg = 40.67, SR = 153, Boundary % = 62%.
* Projected start: ~55–60 runs in first 6 overs (RR = 10+).
* **Backup option**: Alex Hales (high runs, good avg/SR, experienced at No. 2).

🔹 **Anchors**

* We have 5 players shortlisted as per criteria – **Virat Kohli, Suryakumar Yadav, Glenn Philips, Lorcan Tucker, Daryl Mitchell**, and we need to select **3 players**.
* Looking at the stats, **Virat Kohli** clearly stands out with the highest runs scored. His **batting average is outstanding**, reflecting unmatched consistency. With a strong **strike rate**, solid **boundary %**, and batting position at **No. 3**, he is a natural anchor for the innings. Truly, the “Run Machine” Kohli.
* The second choice is **Suryakumar Yadav**, who brings flair and explosiveness. He has the **second highest runs**, an excellent **average of 59.75**, and an incredible **strike rate of 189.68**. With a **boundary % of 57%** and experience at No. 4, he can both anchor and accelerate depending on the match situation.
* The third anchor is **Glenn Philips**. Apart from being a wicketkeeping option, he provides balance with his **high strike rate** and strong boundary-hitting ability. This makes him an impactful anchor who can stabilize the middle overs and switch gears when required.
* Combined performance (Kohli + Surya + Philips):
  + 🔹 Batting Average – 61.33
  + 🔹 Strike Rate – 156.60
  + 🔹 Boundary % – 58%
* If openers fall early, this trio can collectively add 90–100 runs in 10 overs, ensuring stability and momentum.
* For **backup option**, we considered **Lorcan Tucker**. While his stats are solid, Glenn Philips edges ahead due to his superior strike rate and boundary % contribution.

🔹 **Specialist Fast Bowlers**

* We prioritized selecting **Specialist Fast Bowlers first** since this determines what type of all-rounders and finishers will balance the team later.
* Shortlisted players: **Sam Curran, Anrich Nortje, Shaheen Afridi, Tim Southee**. Requirement: **3 bowlers**.
* The first choice is **Sam Curran**. He tops the list with the **most wickets taken**. His **economy rate (6–7)** is excellent for T20s, his **bowling strike rate of 10.46** means he takes a wicket roughly every 10 balls, and a **bowling average of 11** confirms his efficiency.
* The second pick is **Anrich Nortje**. A genuine right-arm fast bowler, he has taken **11 wickets**, with a superb **economy of 5.37**. His strike rate and average are both impressive, making him a consistent wicket-taker and a bowler who controls runs effectively.
* For the third spot, variety is key. We select **Shaheen Shah Afridi**, a left-arm fast bowler. He provides the angle variation needed against top-order batsmen. His **bowling SR, average, and economy** are all strong, making him an ideal partner to Curran and Nortje.
* **Combined performance (Curran + Nortje + Afridi):  
  🔹 Bowling Average – 11.34  
  🔹 Bowling Strike Rate – 11.20  
  🔹 Economy – 6.08  
  🔹 Dot Ball % – 49.49%**
* If this trio bowls in the **powerplay (first 6 overs)**, they are projected to take around **3 wickets**, dismantling the opposition’s top order and giving us the early advantage

🔹 **All Rounders**

* For all-rounders, we need **flexible players** who can contribute with both **bat and ball**.
* Since we already have a strong fast-bowling lineup, the focus here is on **spin options** to balance the attack.
* Shortlisted players: **Mitchell Santner, Rashid Khan, Shadab Khan, Sikandar Raza**. Requirement: **2 players**.
* First, let’s consider **Sikandar Raza**. He qualifies as both a **Finisher** and an **All-Rounder**, so we will evaluate him separately in the Finishers section.
* **Shadab Khan** emerges as a strong pick. He is a **leg-break bowler** who took **11 wickets** with an excellent **economy of 6.35**. His **bowling strike rate** is also impressive, showing he picks wickets regularly. With the bat, his **batting average and high strike rate** make him a useful lower-order contributor.
* **Mitchell Santner** is the second choice. A left-arm orthodox spinner, he provides variation and balance to the bowling unit. He has taken a solid number of wickets, with good **economy, bowling SR, and bowling average**. Additionally, his **batting strike rate** shows he can add valuable runs when required.
* If Shadab and Santner bowl their full quota of 8 overs, their combined bowling strike rate of **13.8** suggests they could claim around **3 wickets**, giving crucial middle-over breakthroughs.
* **Rashid Khan** is an excellent backup option. If the **pitch favors spin**, he becomes a dangerous weapon with his **googly variations**. However, his **bowling average and strike rate** were on the higher side in this tournament, so he narrowly misses the first-choice list.

🔹 **Finishers**

* We had **5 shortlisted players** who fit the criteria for finishers, but the requirement was to select only **1 player**.
* Looking at runs, **Sikandar Raza** stands out with the **highest number of runs** and a **strike rate of 147**, making him a strong contributor with the bat. Additionally, his **right-arm off-break bowling (economy 6.50)** provides extra bowling support.
* However, his **batting average is on the lower side**, which creates some concern for high-pressure finishing roles. In matches where **batting depth is critical**, **Glenn Maxwell** becomes the better choice with his **higher batting average and superior strike rate**. He can also slot in as an **Impact Player** in case of a top-order collapse.
* If the **pitch is favoring fast bowlers**, then **Hardik Pandya** (right-arm medium fast) is a solid option. He has taken **8 wickets with a Bowling SR of 13.5**, while also maintaining a strong strike rate with the bat – making him a perfect balance between batting firepower and extra seam-bowling option.

**Conclusion:** The **finisher role is condition-based**. Depending on the match scenario:

* **Sikandar Raza** -- for batting + spin support.
* **Glenn Maxwell** -- for extra batting depth.
* **Hardik Pandya** -- for seam-bowling support + finishing power.

### ***🏏 Final Playing XI***

**Openers (Power Hitters)**

* Jos Buttler (wk)
* Rilee Roussouw

**Middle Order Anchors**

* Virat Kohli
* Suryakumar Yadav
* Glenn Philips

**All-Rounders / Finisher**

* Sikandar Raza
* Shadab Khan
* Mitchell Santner

**Specialist Fast Bowlers**

* Sam Curran
* Anrich Nortje
* Shaheen Shah Afridi

### **📊 Combined Team Performance**

* **Batting Average:** 39.60
* **Batting Strike Rate:** 154.54  
  → Expected team total ≈ **185 runs in 20 overs**
* **Bowling Average:** 14.12
* **Bowling Strike Rate:** 13.09
* **Bowling Economy:** 6.47  
  → Opposition expected ≈ **145 runs in 20 overs**
* **Dot Ball %:** 41.15%  
  → ~**48–50 dot balls per innings**

This gives a **30–40 run cushion** between expected score and expected defense → exactly the winning margin you had set at the start.

***Conclusion***

Using ICC T20 World Cup 2022 data, we built a **data-driven framework** to evaluate players and select the best possible XI. Through systematic **data cleaning, modeling, DAX-driven measures, and Power BI dashboards**, we analyzed batting and bowling performances against clearly defined role-based criteria.

The process ensured that selection was based on **evidence, not intuition**, balancing power hitters, anchors, finishers, all-rounders, and specialist bowlers. The final XI demonstrates a strong batting lineup capable of consistently scoring ~180 runs and a bowling attack effective enough to restrict opponents to ~150 runs.

This project highlights how **data analytics can transform sports decision-making**, providing actionable insights for player evaluation and team optimization. The methodology can further be extended to future tournaments, franchise leagues, or even match-specific scenarios (pitch type, opposition analysis, player form).

In short, this case study showcases the power of **analytics in cricket strategy**, where numbers back up intuition and help build a winning combination.