PRODUCT TRACK - THE AMERICAN EXPRESS CAMPUS CHALLENGE 2024 SOLUTION

Presented by: CHETANYA MAHANA

TABLE OF CONTENT

PARTICIPANT INTRODUCTION

Driven by a passion for data science and coding, I am a pre-final year B.Tech student at Punjab Engineering College, specializing in Data Science Engineering. I thrive on solving real-world problems and am dedicated to continuously expanding my knowledge and skills in development, machine learning, and data analysis. I am proficient in C++, Python, Excel, Data Structures and Algorithms, and Data Science.

PROBLEM STATEMENT

ROUND 1

Objective: Clean and prepare a dataset without anomalies to ensure data integrity for subsequent analysis.

Key Constraints and Tasks:

1. Data Anomalies: The dataset includes various anomalies such as duplicates, missing values (NULLs), outliers, and incorrect data types.
2. Metadata Reference: Understanding of variable definitions and relationships through the Metadata sheet is crucial for identifying anomalies.

Expected Outcomes:

1. Cleaned Dataset: A dataset free of duplicates, with all anomalies addressed according to the outlined correction methods.
2. Accurate Imputations: Missing values are logically and accurately imputed using related data.
3. Consistent Data Types: Data fields are converted to correct and consistent data types.
4. Query Responses: Responses to the queries asked based on the cleaned dataset.

ROUND 2

Objective:

Formulate the best possible cricket team from a given clean dataset, optimizing for performance metrics.

Key Constraints and Tasks:

1. Team Selection:

- Retain a maximum of 4 players from the provided "near best 11."
- Select 11 players with higher cumulative scores than the "near best 11."

2. Team Composition:

- Total of 11 players.
- Minimum 3 batsmen (Cumulative Runs > 100).
- Minimum 3 bowlers (Cumulative Wickets > 10).
- Minimum 2 all-rounders (Cumulative Runs > 10 and at least 1 wicket in minimum 2 matches).
- Minimum 1 wicketkeeper (designated as wicketkeeper in at least 2 matches).

3. Performance Metrics:

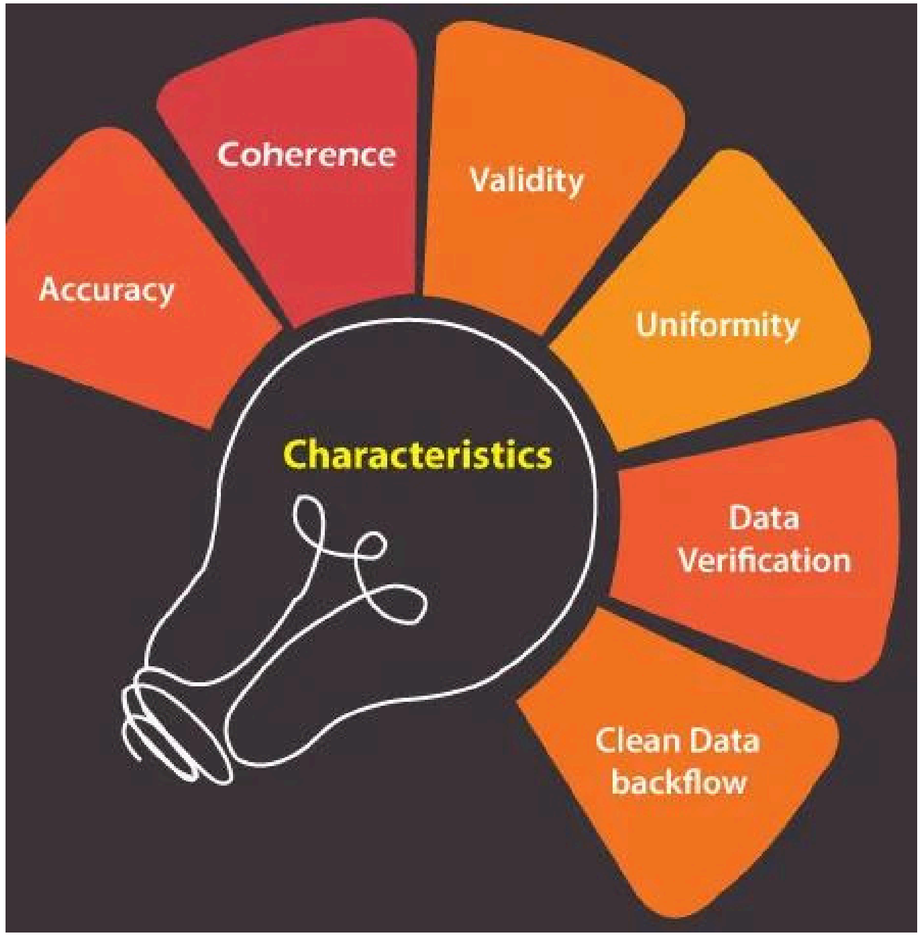
- Using Consistency and Recency of performance to rank players, calculated through methods like weighted average or standard deviation.

Expected Outcomes:

- A cricket team of 11 players meeting the selection criteria, optimized for performance.

INTERCONNECTION BETWEEN ROUNDS

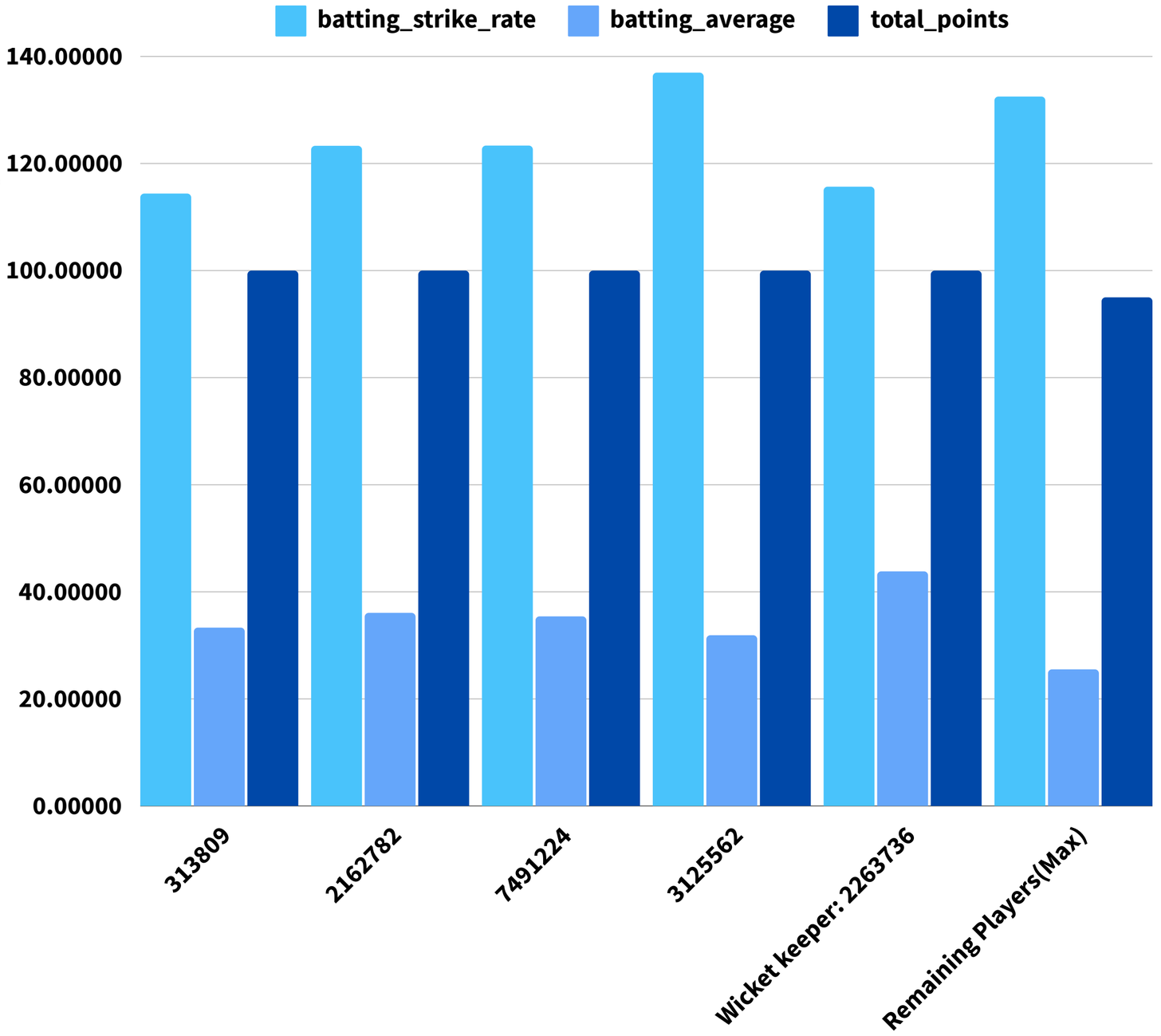
- Input Quality: Round 1's cleaned dataset forms the foundation for Round 2. Ensures data integrity for accurate player selection.
- Performance Metrics: Round 1's data correction enables reliable player metrics. Critical for consistency and recency calculations in Round 2.
- Optimization and Constraints: Round 1's data handling impacts Round 2's optimization feasibility. Cleaner data allows more precise constraint definitions.
- Decision Making: Round 1's documentation provides transparency for Round 2 choices. Enables informed decisions on player attributes and team requirements.
- Continuous Improvement: Round 2 insights can refine Round 1 processes. Creates an iterative loop for ongoing data and selection enhancement.



EXPLANATION OF PLAYER SELECTION STRATEGY

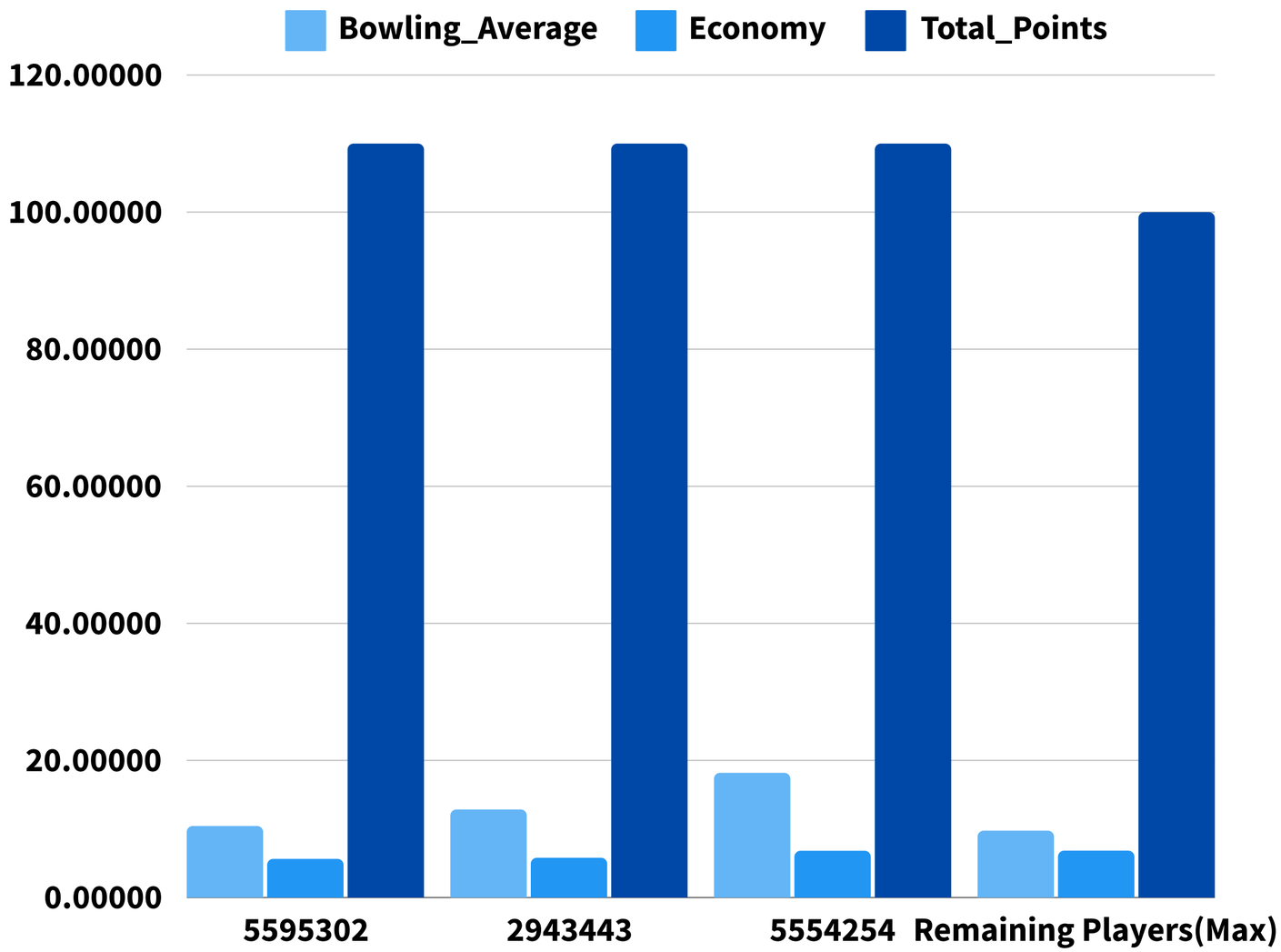
Key Performance Indicators:

- Batsman Selection:
  - Total Points: Sum of points based on Strike Rate, Batting Average, Centuries (100s), and Half-centuries (50s).
  - Consistency Score: Calculated as the Standard Deviation of the Strike Rate.
  - Recency Score: Exponentially Weighted Average (EWA) of the Strike Rate to give more importance to recent performances.
- Wicketkeeper Selection:
  - Total Points: Calculated based on batting points.
  - Designation: Must have been a wicketkeeper in at least 2 matches.



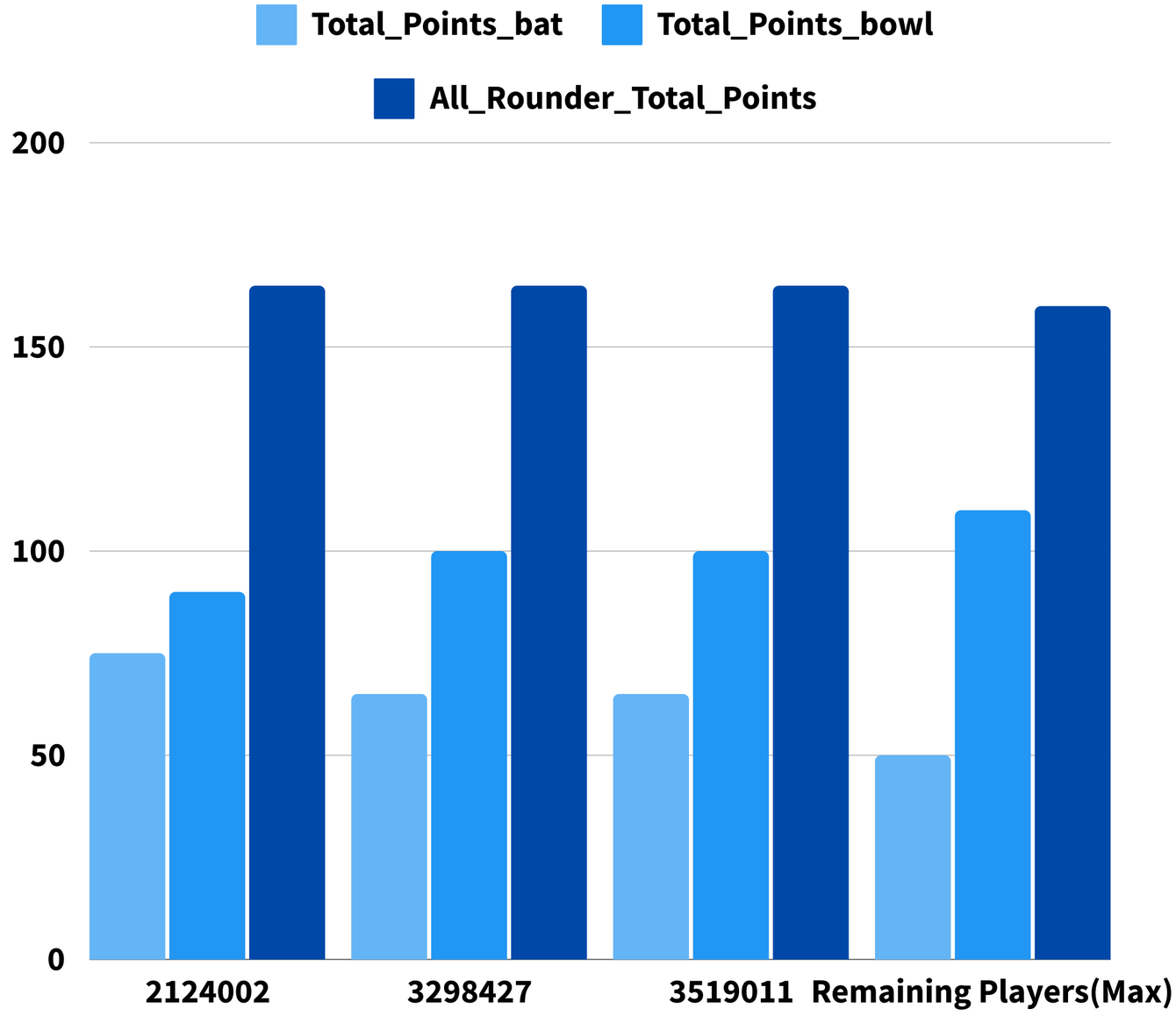
Key Performance Indicators:

- Bowler Selection:
  - Total Points: Sum of points based on Bowling Strike Rate, Economy Rate, Bowling Average, and Four-wicket hauls per innings.
  - Consistency Score: Calculated as the Standard Deviation of the Economy Rate.
  - Recency Score: Exponentially Weighted Average (EWA) of the Economy Rate.



Key Performance Indicators:

- All Rounder Selection:
  - Total Points: Sum of batting and bowling points.



```
batsman['final_score'] = (batsman['Total_Points'] / batsman['Total_Points'].max()) +
(1 - batsman['std_strike_rate_norm']) +
batsman['exp_weighted_avg_sr_norm']
```

```
bowler['final_score'] = (bowler['Total_Points'] / bowler['Total_Points'].max()) +
(1 - bowler['std_economy_norm']) +
bowler['exp_weighted_avg_eco_norm']
```

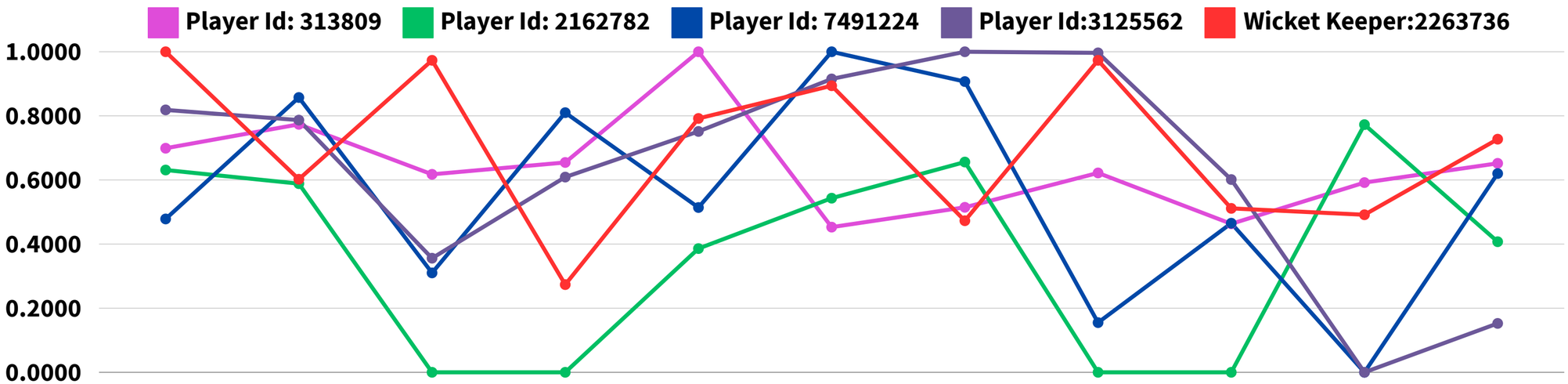
```
allrounders['All_Rounder_Total_Points'] = (allrounders['Total_Points_bat'] +
allrounders['Total_Points_bowl'])
```

OPTIMIZATION/BALANCING TECHNIQUE FOR FINAL TEAM (PLAYING 11) SELECTION

Depth of Statistical Analysis:

- Consistency Calculation: Using Standard Deviation to understand the variability in performance. Lower variability (lower standard deviation) implies higher consistency.
- Recency Calculation: Using Exponentially Weighted Average (EWA) to ensure that more recent performances have a higher impact on the final score, capturing the player's current form.
- Normalization: Scores are normalized to a common scale to ensure fair comparison and aggregation.
- All-Rounder Score: All-rounders provide dual value, so their combined batting and bowling points are considered.

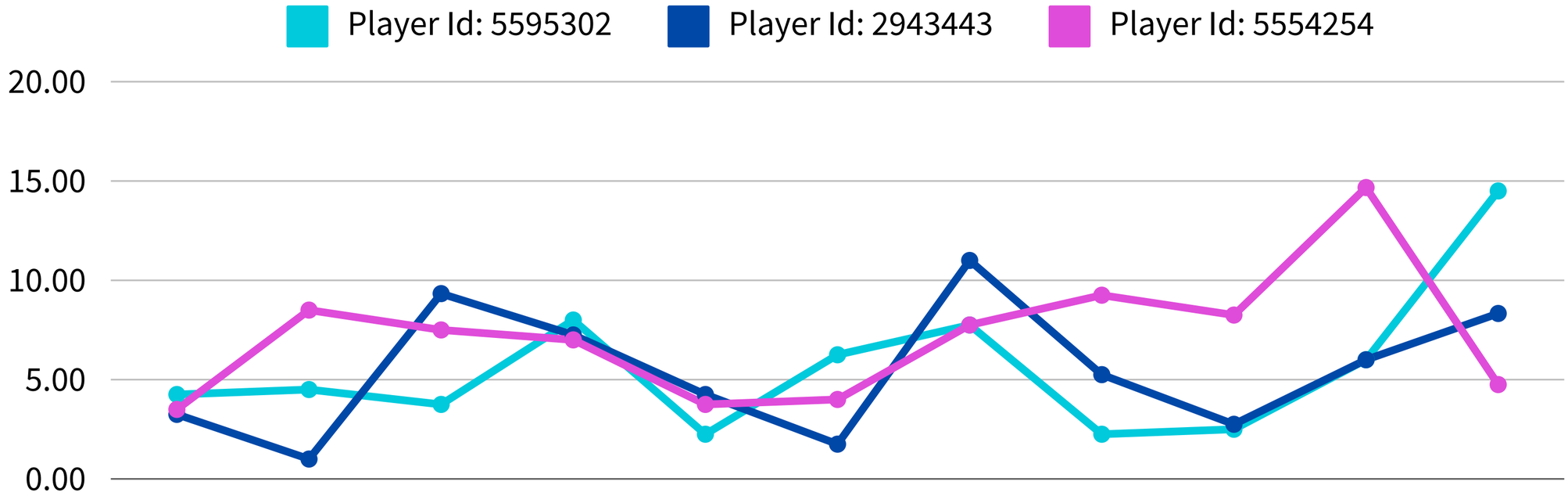
OPTIMIZATION/BALANCING TECHNIQUE FOR BATSMEN AND WICKET KEEPER SELECTION



Hypothesis and Considerations:

1. Hypothesis for Batsmen:
   - Players with higher strike rates and averages are likely to contribute more to the team's total runs.
   - Consistency in scoring runs across matches is crucial, as is recent form (captured through EWA).
2. Hypothesis for Wicket Keeper:
   - Wicketkeepers are primarily evaluated based on their batting performance but must have participated in the role of a wicketkeeper in enough matches.

OPTIMIZATION/BALANCING TECHNIQUE FOR BOWLER SELECTION



Hypothesis and Considerations:

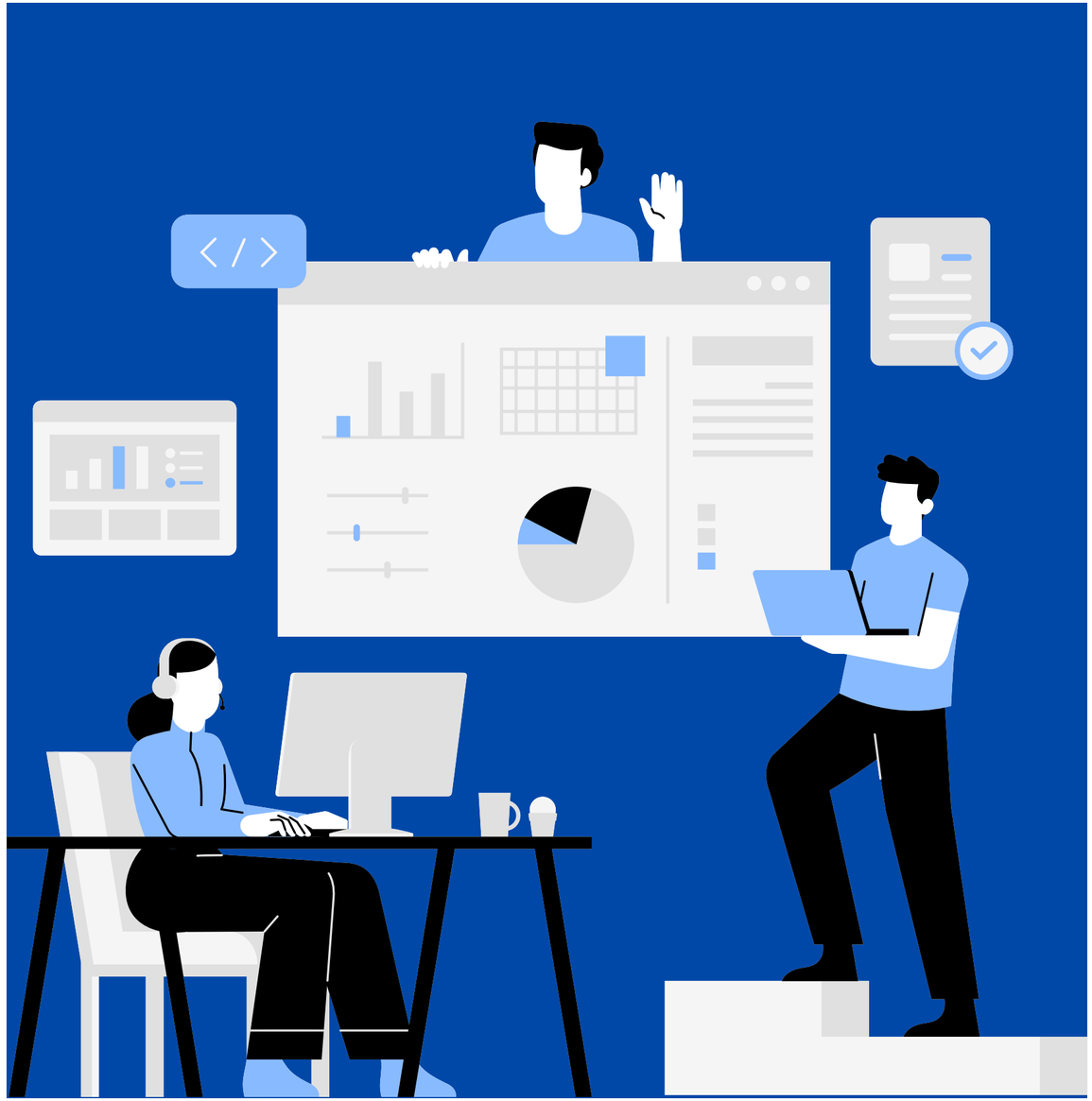
1. Hypothesis for Bowlers:

- Bowlers with lower strike rates and economy rates are more effective in limiting the opposition's scoring.
- Consistency in maintaining low economy rates is important, as is recent form.

SCALABILITY OF APPROACH

1. Automated Data Pipeline
   - Data Ingestion: Regularly import player performance data from APIs.
   - Preprocessing: Clean and normalize data.
2. Dynamic Weight Adjustment
   - Machine Learning: Use historical data to adjust weights for total points, consistency, and recency scores.
   - Optimization: Apply algorithms to find the best weight combinations.
3. Scoring System
   - Calculation: Automate total points, consistency, and recency score calculations.
   - Normalization: Standardize scores for fair comparison.
4. Player Selection Algorithm
   - Sorting: Rank players by final scores.
   - Filtering: Select top players for each role ensuring team balance.

5. Validation
   - Backtesting: Compare model predictions with historical outcomes.
   - Simulation: Test robustness in various scenarios.
6. User Interface
   - Dashboard: Display real-time rankings and insights.
7. Continuous Improvement
   - Feedback Loop: Update model with new data and user feedback.



THANKS FOR WATCHING