Cricket Data Quality Analysis Report

Summary

This report presents findings from a data quality analysis of cricket match data contained in several CSV files. The analysis uncovered multiple quality issues that could affect the reliability of any models or analysis built using this data. The most significant issues include:

1. **Legal Ball Counting Errors**: Incorrect increment of legal ball counts for wide balls and no-balls
2. **Run Calculation Inconsistencies**: Discrepancies in total runs between components and recorded totals
3. **Missing Wicket Information**: Incomplete dismissal details affecting player statistics
4. **Incomplete Player Data**: Missing batting and bowling information in player records
5. **Unexplained Innings Endings**: Early innings termination without clear reasons recorded
6. **Inefficient Data Format**: Null values used where zeros would be more appropriate
7. **Missing Over Data and Schema Errors**: Completely missing overs, inconsistent bowler statistics, and structural metadata issues

Methodology

The data quality assessment involved:

- Cross-referencing data across all provided CSV files
- Validating cricket-specific rules and logic
- Checking for internal consistency within each file
- Verifying mathematical correctness of calculated fields
- Analyzing completeness of key fields

Detailed Findings

1. Legal Ball Counting Errors

1.1 Incorrect legal ball counting for extras

In cricket, wide balls and no-balls should not increment the legal ball counter. However, in this dataset, the legal_ball_of_over value incorrectly increments for these deliveries.

Example: In match 1443098, innings 1, over 1:

- Ball 2 is marked as a wide but has legal_ball_of_over = 2, which is wrong. It should not have incremented for this delivery.

2. Run Calculation Inconsistencies

2.1 Inconsistent total_runs_so_far values

The calculated running total of runs (by adding runs_off_bat, runs_wide, no_ball_penalty_runs, runs_bye, runs_leg_bye) differs from the recorded total_runs_so_far values in multiple instances.

Examples:

- Row 682: Calculated total is 84, but recorded as 85
- Row 683: Calculated total is 84, but recorded as 89
- Row 684: Calculated total is 90, but recorded as 95

2.2 Mismatches between datasets

There are discrepancies between the final run totals in ball-by-ball data and the totals recorded in the innings data.

Example: Innings 1449663-2 shows 106 runs in innings.csv but only 68 runs in the final ball-by-ball record.

3. Missing Wicket Information

3.1 Incomplete dismissal details

When wickets are taken, the dataset often lacks complete information about the dismissal. Critical details such as the batter who got out and the bowler who took the wicket are missing in several instances.

This issue is particularly evident in rows 555-610 where wicket counts are recorded but the corresponding player information is missing. Without complete dismissal details, player performance statistics cannot be accurately calculated.

4. Incomplete Player Data

4.1 Missing player attributes

The players.csv reference file is missing important information for player IDs that appear in the ball-by-ball data.

Examples:

- Player ID 1277545 (Abishek Porel) appears in ball_by_ball but data related to batting_hand and bowling_type is missing in players.csv
- Player ID 995037 (Mahidul Islam Ankon) appears in ball_by_ball but data related to batting_hand and bowling_type is missing in players.csv
- Player ID 457280 (Adam Rossington) appears in ball_by_ball but data related to batting_hand and bowling_type is missing in players.csv

This disconnect between the datasets makes it difficult to enrich analysis with player attributes like batting hand or bowling style.

4.2 Missing non-facing batter details

In cricket, two batters are always at the crease. However, in the dataset, approximately 0.63% of records (20 out of 3162) are missing both non_facing_batter_id and non_facing_batter_name, creating an incomplete picture of the match state.

5. Unexplained Innings Endings

5.1 Prematurely ended innings without reason

Some innings appear to end early without the expected conditions:

- Match 1449663- Innings 2: Ended at 14/20 overs without reaching target or being all out (only 6 wickets)
- Match 1475258 - Innings 1: Ended at 12/20 overs without being all out (only 3 wickets)

The dataset should include a dedicated column indicating when a match is suspended or stopped due to external factors such as rain, bad light, flood lights not working, or other reasons. This is particularly relevant for match 1475258, where the first innings ended prematurely.

6. Inefficient Data Format

6.1 Null values instead of zeros

Several critical fields have a high percentage of null values:

- no_ball_penalty_runs: 99.65% null
- runs_bye: 99.65% null
- runs_leg_bye: 98.86% null
- runs_wide: 96.55% null
- wicket_type: 95.98% null

Instead of leaving these as null values, it would be better to use 0 values in these columns for balls where these extras did not occur, making the data more consistent and easier to process. This would eliminate the need to handle null values separately from zero values, streamlining analysis and reducing the risk of calculation errors.

7. Missing Over Data and Schema Errors

7.1 Inconsistent bowler statistics

The bowler_balls_bowled count doesn't consistently match the actual number of legal deliveries bowled, suggesting an issue with how bowler statistics are tracked.

Examples:

- The most significant discrepancy was found in match 1459548 in the 13th over where bowler Shaheen Shah Afridi's recorded count (1) differs from calculated legal deliveries (19) by 18 balls, but ended with 24 deliveries overall.
- In match 1449663, innings 2, the statistics for bowler Tristan Luus in overs 2 and 8 are completely inconsistent with the actual number of legal balls he delivered, creating critical inaccuracies in bowling performance metrics.

These inconsistencies make it impossible to reliably calculate key bowling statistics such as economy rate and bowling averages, which are fundamental to cricket analysis.

7.2 Gaps in ball sequence data

There are significant gaps in the ball-by-ball sequence data. For instance, in match 1459548, innings 1, over 13 shows only 2 balls registered with the rest missing. Similarly, in match 1459550, innings 2, the entire 19th over is missing, with data for the 18th and 20th overs present but nothing for the 19th over.

These gaps affect all statistical calculations and make it impossible to properly analyze complete match data. Such missing data is particularly problematic for modeling and predicting match outcomes, as it creates an incomplete picture of how innings progress.

7.3 Metadata and schema errors

The dataset contains structural issues in table schemas. For example, in the players.csv file, there are duplicate column names where 'batting_hand' appears twice when one should be 'batting_hand' and the other should be 'bowling_hand'. This schema error makes it difficult to correctly interpret and use the player data, particularly for analyzing bowling statistics and matchups.

Recommendations

Based on the findings, the following recommendations are made to improve data quality:

1. **Correct Legal Ball Counting**: Fix the logic for tracking legal_ball_of_over to ensure it only increments for legitimate deliveries and not for wides or no-balls. Consider starting the counter at 0 instead of 1 to avoid off-by-one errors in calculations.
2. **Validate Run Calculations**: Implement checks to ensure that total_runs_so_far always equals the sum of its components (runs_off_bat, extras, etc.). Reconcile discrepancies between ball-by-ball and innings totals.

3. **Require Complete Wicket Information**: Enforce data validation rules to ensure all wicket records include complete details about the dismissed batter, the bowler, and dismissal type, ensuring accurate player statistics.
4. **Complete Player Reference Data**: Update the players.csv file to include complete batting_hand and bowling_type information for all players, enabling proper analysis of player characteristics.
5. **Add Match Interruption Field**: Implement a dedicated field to record when matches are affected by external factors (rain, bad light, etc.), providing essential context for innings that end prematurely.
6. **Replace Nulls with Zeros**: For extras fields (no_ball_penalty_runs, runs_bye, etc.), use 0 instead of null when these extras didn't occur. This makes data processing more efficient and calculations more straightforward.
7. **Ensure Complete Over Data**: Implement completeness checks to identify and address missing overs and balls. For historical data, flag incomplete sequences so they can be appropriately handled in analysis.
8. **Cross-Validate Data Files**: Implement systematic checks between ball-by-ball data and innings summaries to ensure consistency in totals and statistics.
9. **Document Cricket-Specific Data Rules**: Create clear documentation for how cricket events should be recorded, ensuring consistent application of sport-specific rules in data collection.
10. **Validate Data Schema**: Implement schema validation to identify and correct structural issues like duplicate column names. Ensure consistent and accurate representation of field names and data types across all dataset files.

Conclusion

The cricket match data analyzed contains several significant quality issues that could impact the reliability of any analytics or predictive models built using this data. The problems range from basic data recording errors to missing critical information and structural issues in the dataset.

Addressing these issues is essential before this data can be reliably used for cricket analytics. By implementing the recommended improvements, particularly focusing on consistent legal ball counting, complete wicket information, filling data gaps, and correcting schema errors, the dataset would become substantially more valuable for cricket analysis and modeling.