Errors

I identified a total of eleven distinct errors and corrected them across both the BBOps_DQ_26FellowPitch.csv and BBOps_DQ_26FellowContact.csv datasets. These ranged from duplicate identifiers and mislabeled team assignments to physically implausible pitch calculations. Each issue was located, verified, and corrected before combining the datasets.

FellowPitch File

1. Duplicate PitchIds

The dataset contains duplicate entries for PitchIds 30832197 and 30832432. Both appear twice with identical data across all columns, indicating that each pitch could have been recorded twice or that an incorrect duplication occurred. This is an error because each pitch should have its own unique identifier, and duplicates can affect analyses such as pitch counts or performance statistics.

Fix: Remove each duplicate row from dataset.

2. Incorrect PitchId Entry

PitchId 30850000 appears in the dataset, but the correct ID should be 30832365. This conclusion is based on matching timestamps and current count of the at-bat, as well as the absence of the 30832365 PitchId. The error likely occurred due to a manual entry mistake or data substitution, resulting in the incorrect identifier being used.

Fix: PitchId 30850000 was corrected to 30832365.

3. Pitcher Name Typo

The pitcher listed as "Dove, Austin" appears once in the Pitcher column but shares the same PitcherId as "Love, Austin". This indicates that "Dove" is a misspelling. Because a player's name must be consistent and linked to the unique identifier PitchId this is an error.

Fix: Pitcher "Dove, Austin" was corrected to "Love, Austin".

4. Two BatterIds for the Same Player

The batter “Bernal, Leonardo” appears with two different BatterIds, 699024 and 688024. The ID 688024 occurs only once while the ID 699024 occurs in all other instances of the batter, suggesting an error. A player should have a single unique ID, and multiple identifiers for the same individual can cause mismatched or duplicated records.

Fix: PitchId 688024 was corrected to 699024.

5. Batter Name Typo

The batter listed as “Fedko, Kyler” appears once in the Pitcher column but shares the same BatterId with “Fedko, Tyler”. This indicates that “Kyler” is a misspelling. This is an error because consistent naming is essential for tracking player performance accurately.

Fix: Batter “Fedko, Kyler” was corrected to “Fedko, Tyler”.

6. Incorrect BatterSide Label

For player “Rivas, Jeremy” one entry lists him as batting left-handed, while all others indicate that he bats right-handed. This is an error due to him not being a switch hitter, he also could not have changed batting sides mid-at-bat. This inconsistency reflects an error that affects player analysis.

Fix: BatterSide “Left” for Batter “Rivas, Jeremy” was corrected too “Right”.

7. Incorrect Date Entry

The Date column includes two dates: 4/24/2024 and 4/24/2025. The date 4/24/2024 is an error due to only occurring once in the column and the dataset’s source confirming all data is from a game that took place in 2025.

Fix: Date 4/24/2024 was corrected to 4/24/2025.

8. Impossible RelSpeed Value

The `RelSpeed` (Release Speed) column contains a value of `941.433`, which is far higher than the next highest value (96.872 mph). In addition, 941 mph is physically impossible for a baseball pitch, this is clearly an error, likely caused by a misplaced decimal or data-entry mistake. Potential corrections include setting the value to `NULL`, replacing it with the pitcher's average release speed from the same game, or assuming a misplaced decimal and converting it to 94.143. I chose to change the value to `NULL` to avoid introducing potentially incorrect data, especially given that there is already a sufficient sample size of 45 pitches from this pitcher in the game to support accurate analysis.

Fix: `RelSpeed 941.433` was corrected to `NULL`.

9. Incorrect Team Assignment for Catcher

The `CatcherTeam` column lists `"Cardenas, Noah"` as playing for `WIC_SUR` throughout the dataset; however, in 12 rows, he is incorrectly identified as playing for `SPR_CAR`. This discrepancy is clearly an error, as these 12 instances are the only times he is listed with `SPR_CAR`, while all other entries consistently associate him with `WIC_SUR`. Additionally, when he appears as a batter during the game, he is only ever shown as part of `WIC_SUR`, further confirming the mistake. Notably, this error occurs exclusively during the top of the 4th inning, suggesting a data-entry or labeling issue affecting only that portion of the game.

Fix: `CatcherTeam` 12 values of `"SPR_CAR"` during Top of 4th inning was corrected to `"WIC_SUR".`

10. Negative Speed Drop Value

In the `SpeedDrop` column, one record shows a value of `-3.9177`, while all other entries are positive. Since `SpeedDrop` represents the decrease in pitch velocity from release to home plate, it cannot logically be negative because air resistance and gravity naturally slow the ball as it travels toward the plate. The `SpeedDrop` value is calculated as `RelSpeed` minus `ZoneSpeed`, so I verified both values to check for a possible calculation error. However, the formula was applied correctly, indicating that either or both the `RelSpeed` and `ZoneSpeed` values for that record are inaccurate. To prevent using unreliable data, I set the `RelSpeed`, `ZoneSpeed`, and `SpeedDrop` values for that single row to `NULL`.

Fix: `RelSpeed`, `ZoneSpeed`, and `SpeedDrop` row with `-3.9177` was corrected to `NULL.`

FellowContact File

11. Distance Outlier

The `Distance` column contains a value of `752.2577`, which is far greater than the second-largest value of `378.0873`. Hitting a baseball 752 feet is physically impossible, as even the longest recorded home runs do not exceed approximately 600 feet. This indicates a clear data-entry or measurement error.

Fix: `Distance 752.2577` was corrected to `NULL.`