Process of Combining Datasets

The process of combining the two TrackMan datasets involved cleaning, validating, and merging data to create a combined record with both Pitch and Contact data. I first identified and corrected errors, then merged the files using Python's Pandas library to produce a single, verified dataset for further analysis.

Steps:

1. Locate and review errors

The first step was identifying potential errors within both original datasets. Using Excel, I examined column value ranges, checked for duplicates using conditional formatting, and used table filters to locate inconsistent or impossible values.

2. Validate identified errors

Next, I verified whether each flagged issue was truly an error. This involved applying general baseball knowledge and arithmetic checks to confirm that the data made sense. I also referenced TrackMan's official glossary to understand the meaning of each field. Though there were other possible errors located I could not confidently validate them as true errors thus they were excluded from my final error report to maintain accuracy.

3. Determine and Implement Corrections

After validation, I decided how to handle each confirmed error. Obvious typos or numeric entry mistakes were corrected, while other uncertain or incomplete values were replaced with `NULL` to preserve data integrity. Once all adjustments were complete, I exported the two cleaned datasets into new CSV files for use in the merging process.

4. Import cleaned data

With the cleaned data ready, I imported both files into a Jupyter Notebook using Pandas. Then performed a quick exploratory check confirming that the import process had succeeded without corruption.

5. Select merge key

Before joining the data, I examined which columns could be good merge keys. Out of twelve shared columns, I selected `PitchId` as the primary merge key because it uniquely identified each pitch in both files and maintained a one-to-one relationship between records. Although I considered joining on multiple columns, `PitchId` was the most reliable and direct choice.

6. Merge the datasets

Since the pitch-level file contained every pitch thrown in the game, while the contact file only included pitches where the batter made contact, I chose to perform a left join. This approach ensured that all pitch records remained intact, with contact information joining where available. After merging, I noticed that eleven columns appeared twice (as `_x` and `_y` variants). To correct this, I removed duplicate columns from the contact dataset and repeated the merge.

7. Post-merge validation

After the merge, I explored the combined dataframe to confirm that row counts matched the original pitch dataset and that no new duplicates or inconsistencies had been introduced.

8. Export

Finally, I exported the completed dataframe to a CSV file named `3002223_CombinedData.csv`. During this step, I replaced all `NaN` values with `NULL` to maintain consistency with the original files. The resulting file represented a fully combined, validated version of the two TrackMan data sources.

Conclusion:

This process accurately merged both TrackMan datasets into a clean, verified file, creating a reliable foundation for continued analysis and exploration.