

# Daily Summary - Horse Racing Feature Pipeline

Date: 2025-03-28

Summary of Work Completed (March 28, 2025)

## 1. Core Feature Engineering:

- Verified and preserved `class\_num` and `class\_label` across all flattening passes.
- Extracted `GoingStick`, `race\_class`, and contextual race info.

## 2. Statistical Features:

- Parsed and cleaned trainer and jockey win rates and profit figures.
- Normalized stats:
  - \* Converted percent strings (e.g. '13%') to floats.
  - \* Converted profit strings (e.g. '+12.58') to floats.

## 3. ID Handling:

- Ensured `horse\_id`, `trainer\_id`, `jockey\_id` treated as strings.
- Validated horse\_id format to be exactly 7 digits.

## 4. Presence Flags:

- Added `has\_form`, `has\_comment`, etc., using robust logic (not null or empty).

## 5. Final Output:

- All clean features written to CSV with no loss of critical data.
- Verified via forensic inspection of values.

Preparation for Tomorrow:

- Have multiple JSON files in `./data/raw/\*.json`.
- Each should be timestamped with scrape time.
- We'll build a batch flatten script to handle all daily files.

## Keywords (20):

horse\_id\_valid, class\_num, class\_label, GoingStick, presence\_flags, jockey\_last\_14\_win\_pct,

trainer\_last\_14\_profit, flatten\_json, race\_context, form, spotlight, headgear, extract\_features\_v2,  
inject\_country\_and\_flatten, percent\_parser, string\_id\_cast, field\_overwrite\_bug, data\_auditing,  
zscore\_normalization, race\_class\_parser