CHECKPOINT 4 — PREPROCESSING REPORT & DATA DICTIONARY

Nicholas Condos

Nov 16, 2025

1. Pipeline Overview & Target Grain

The objective of this preprocessing phase was to consolidate, standardize, filter, and validate ATP match data from 2008–2024 using Jeff Sackmann's open dataset. The final output of this step is a clean, analysis-ready dataset and a structured, traceable data dictionary.

Target Row Grain:

> One row represents one ATP singles match (winner vs. loser).

Pipeline Steps:

1. Raw Ingestion:
   Pull yearly match files (atp_matches_YYYY.csv) from Sackmann's GitHub repository for 2008–2024.

2. Merge:
   Vertically concatenate all years into a unified table of ATP matches.

3. Initial Filters:
   Remove matches without a valid score and remove rows marked with walkovers/retirements via comment.

4. Reload & Standardize:
   Reload merged file as the raw master input for consistent downstream processing.

5. Clean & Structure:
   Filter singles only, remove qualifying rounds, standardize surface labels, convert date formats.

6. Feature Engineering:
   Create combined efficiency metrics for downstream modeling.

7. Validation:
   Row counts, duplicate checks, missingness analysis, and numeric sanity checks.

8. Export:
   Save a clean, fully processed dataset as:

CHECKPOINT 4 — PREPROCESSING REPORT & DATA DICTIONARY

atp_matches_2008_2024_clean.csv

This pipeline is fully reproducible and documented in the project GitHub repository.

2. ID & Mapping Strategy

Row ID Definition

Each match is uniquely identified using:

unique_key = tourney_id + "_" + match_num

- tourney_id ← identifies the tournament
- match_num ← match-specific index inside a tournament

Duplicate Check:
duplicate unique_key count = 0
→ Confirms stable, unique row identifiers and proper row grain.

Entity Mapping Rules

- winner_id and loser_id are preserved as numeric ATP player identifiers.
- Player names are retained exactly as provided in Sackmann's data.
- Tournament metadata (tourney_name, tourney_level, surface) standardized for consistency.

3. Standardization Rules

The following transformations ensure consistency and interpretability across 17 years of ATP data:

Field	Standardization Rule	Notes
tourney_date	Convert from YYYYMMDD integer to datetime	pd.to_datetime(..., format='%Y%m%d')

CHECKPOINT 4 — PREPROCESSING REPORT & DATA DICTIONARY

surface	Strip whitespace, title-case values (Clay/Grass/Hard)	Fixes inconsistent input formats
Categorical fields	Standardize to known ATP values	e.g., remove doubles (tourney_level=='D')
Numeric statistical fields	Ensure integer/float types	Serve/return stats, aces, BP stats
comment	Filter out all rows where comment is non-null	Ensures only completed matches remain

4. Reshaping & Integration

Reshaping & Filtering Logic

1. Remove Incomplete Matches
   Score must be non-null; comments must be null (excludes WO/RET/ABD).

2. Remove Doubles Matches
   tourney_level != 'D' ensures singles-only.

3. Remove Qualifying & Round-Robin Rounds
   Eliminates data inconsistent with ATP main-draw match performance.

4. Remove Zero-Serve Rows
   Ensures efficiency metrics can be computed safely.

5. Final Column Ordering
   Logical grouping of tournament metadata, player attributes, serve stats, return stats, and engineered features.

Integration Logic

Although the raw dataset consists of a single match table per year, integration occurs via:

- Vertical stacking of years
- Annual tags via match_year

CHECKPOINT 4 — PREPROCESSING REPORT & DATA DICTIONARY

- Tournament-based grouping (tourney_id)
- Player-level consistency (winner_id, loser_id)

5. Feature & Transformation Specification

Below is the required CP4 feature-engineering table:

Feature Name	Definition / Formula	Input Variables	Example (Before → After)	Type	Purpose
w_srv_ret_balance	(w_1stWon + w_2ndWon) / w_svpt	w_1stWon, w_2ndWon, w_svpt	(30 + 12) / 60 → 0.70	Float	Winner's combined serve-return effectiveness
l_srv_ret_balance	(l_1stWon + l_2ndWon) / l_svpt	l_1stWon, l_2ndWon, l_svpt	(24 + 10) / 55 → 0.62	Float	Loser's combined serve-return effectiveness
efficiency_diff	w_srv_ret_balance - l_srv_ret_balance	Two engineered features above	0.70 – 0.62 → 0.08	Float	Winner–loser performance gap metric

Features are intentionally simple, interpretable, and relevant for downstream logistic/regression modeling (Checkpoint 6).

6. Validation Gates (Quality Assurance)

Row Counts

- Initial merged dataset: 48,779 rows
- After completed-match filters: 43,748 rows
- After singles + main-draw filters: 43,366 rows (as shown in your notebook)

CHECKPOINT 4 — PREPROCESSING REPORT & DATA DICTIONARY

- Final cleaned dataset: 43,366 rows

Duplicate Key Check
duplicate unique_key count = 0

Missingness Analysis

(Most fields < 2% missing; no major fields require imputation)

Notebook output confirms:

- No modeling-critical fields exceed typical thresholds.
- Missingness limited to non-essential metadata.

Numeric Range Checks

Using:

df[['w_ace','l_ace','w_df','l_df']].describe()

Results (representative):

- Winner aces: mean ≈ 7.16, max = 113
- Loser aces: mean ≈ 5.28, max = 103
- Winner DF: mean ≈ 2.57, max = 26
- Loser DF: mean ≈ 3.25, max = 26

All fields fall within legitimate ATP match performance ranges.

Checks:

- Winner/loser serve stats always ≥ 0
- No negative/impossible values
- Date parsing successful
- Surface labels standardized

CHECKPOINT 4 — PREPROCESSING REPORT & DATA DICTIONARY

- Serve-point denominators > 0 for all rows

The dataset passes validation.

7. Runbook & Reproducibility

To recreate this dataset:

1. Open Google Colab.
2. Clone or download repo folder structure.
3. Place raw Sackmann CSVs into /data or load from GitHub.
4. Run notebook cells sequentially:
    - Load & merge yearly files
    - Completed-match filtering
    - Main-draw + singles filtering
    - Standardization (surface, date)
    - Feature engineering
    - Validation checks
    - Export cleaned dataset
5. Upload outputs to GitHub:
    - /data/atp_matches_2008_2024_merged.csv
    - /data/atp_matches_2008_2024_clean.csv