DATA 6560 - Sports Analytics - Checkpoint 4 Fall 2025 Preprocessing

Report & Data Dictionary Checkpoint 4

Overview.....1
Group Work Option.....1
Assignment Objectives.....1
What to Include in Your Checkpoint.....2
Submission Details.....3
Deliverables.....3
Grading & Rubric (10 points total).....3
Additional Notes.....3

Overview

Checkpoint 3 audited data quality & sources.

Checkpoint 4 is the build step: execute a reproducible preprocessing pipeline that converts raw inputs into one tidy, analysis-ready table, and document the exact rules you applied (inputs → rules → outputs), with evidence (before/after counts, examples, and validation gates).

Group Work Option

You may work in groups of up to 3 students, but each student must maintain their own GitHub repository. Collaboration on topic and data collection is encouraged, but each student must submit an individual memo and ensure their repository contains all deliverables.

Assignment Objectives

By completing this checkpoint, you will:

1. Implement a repeatable preprocessing pipeline (not just a plan).
2. Set your row definition and unique key(s)
    a. What one row represents and which columns make it unique.
3. Standardize units, time, and categorical labels with written rules.
4. Integrate sources with explicit join keys and precedence when conflicts occur. 5. Produce a professional Data Dictionary that mirrors the final cleaned table. 6. Prove readiness with validation gates (counts, ranges, logic checks) and a short runbook.

DATA 6560 - Sports Analytics - Checkpoint 4 Fall 2025 What to Include in Your Checkpoint

Pipeline Overview & Target Grain	State what one row represents (player–game, team–season). List all raw inputs and show a simple flow (inputs → steps → output).	What is a “row” in the cleaned file? What files feed into it? What tools did you use (Excel functions, Power Query, Python/R)?
ID & Mapping Strategy	Specify the smallest set of columns that uniquely identify each row. Include any surrogate IDs and team/player lookup sheets.	Which columns form the unique key? How did you resolve name collisions? Where are the mappings stored?
Standardization Rules	Written rules for units, time zones/dates, and canonical category labels (“LA Clippers” → “LAC”).	Which units did you convert and to what? How did you normalize dates/times and seasons? What is your master list for labels?
Reshaping & Integration	Describe reshapes, filters, and joins (keys, join types, and precedence when sources disagree).	What are your join keys? Inner/left/right? If two sources conflict, which wins and why? What rows were dropped?
Feature & Transformation Spec Validation Gates (QA) Runbook & Reproducibility	A compact table for each derived field: Name → Rule/Formula → Inputs → Example (before→after). Quantitative proof the pipeline worked: row counts before/after, duplicate-key rate, missingness by field, min/max checks, plus 2–3 logic tests. A short “how to re-run” section and a dated ChangeLog. Keep raw intact and export a single clean file.	Which new variables (rates, flags, rolling stats) did you compute and why? Provide one concrete example row. How many rows enter/exit each step? Any duplicate-key violations? Do ranges look realistic? What passed/failed and what you fixed? Where are raw vs. clean files? Exact filenames? Steps to re-run from scratch? What changed since CP3?
Data Dictionary (separate workbook sheet)	One sheet named DataDictionary documenting the final cleaned columns.	Does every column appear with: VariableName, Definition/Unit, Type, Source (original/derived), Notes (rules)? Include Lookup sheet(s).

DATA 6560 - Sports Analytics - Checkpoint 4 Fall 2025 Submission Details

Document Type	Google Doc (Professional Memo Format)
Submission	Share your Google Doc link on Google Classroom and upload a PDF copy to the /reports folder in your GitHub repo.

File Naming	LastName_FirstName_CP4.pdf (example: Doe_Jane_CP4.pdf). Include your name(s), project title, and the checkpoint heading at the top of the memo. EACH STUDENT MUST UPLOAD TO INDIVIDUAL REPO
Format Tips	Use clear section headers and short tables. Keep a consistent naming scheme that matches your Data Dictionary.

Deliverables

1. Preprocessing Report (Google Doc → PDF)
2. Data Dictionary (Excel or Google Sheet link), with optional Lookup sheet(s)

Grading & Rubric (10 points total)

Your Checkpoint 4 will be evaluated on the following criteria:

Pipeline Execution & Evidence	You implemented the pipeline with a clear flow, before/after row counts, and at least one before→after example per key transformation.	3
Row Definition & Unique Key(s)	Row definition is explicit; unique key(s) truly identify rows; mapping strategy (lookups/surrogate IDs) is sound.	2
Standardization & Integration	Potential biases, errors, and data limitations are identified and discussed.	2
Validation Gates (QA)	Memo is well-organized, concise, and professionally written (grammar, style).	2
Professionalism & Reproducibility	Clean structure, concise writing, consistent naming; runbook + changelog make re-running straightforward.	1

Additional Notes

- *Keep raw untouched; export one clean file that matches the Data Dictionary exactly.*
- *Prefer compact spec tables over long paragraphs.*
- *If you used GenAI for formulas/regex/mappings, include the prompt and how you verified outputs.*

Grain: One row = Player–Season (2017–2018) Inputs: Salaries.csv, Stats.csv, Lookup_Teams.csv, Lookup_Players.csv Tools: Excel (clean), Power Query (joins), Python/pandas (final transforms) Output: NBA_2017_2018_PlayerSeason_Clean.csv

ID & Mapping

Unique key: PlayerID + Season

Collisions: Resolved with canonical name + team lookup

Mappings stored: /lookups/Teams.csv, /lookups/Players.csv

Standardization

Salary: USD integer

Percentages: Decimal fractions

Season: Fixed “2017–2018”

Teams: Canonical abbreviations (e.g., “LA Clippers” → “LAC”)

Reshaping & Integration

Join keys: PlayerID + Season

Join type: Stats base (left join salaries)

Conflict rule: Stats win for team; salary wins for pay

Dropped: < 100 minutes played, G-League only

Features
Per-36 stats (PTS, REB, AST)

Efficiency: TS%, WS/48, Usage

Salary_perMinute

ValueFlag_Underpaid (rule-based flag)

???? Verification

Row counts: 540 raw → 418 clean

Duplicates: None

DATA 6560 - Sports Analytics - Checkpoint 4 Fall 2025

Ranges: TS 0.40–0.75, Salary $0.5M–$35M

Logic tests: Passed (minutes >0, team consistency, monotonic salary checks)

Runbook ???? Raw: /data/raw/ Clean: /data/clean/NBA_2017_2018_PlayerSeason_Clean.csv Re-run: Power Query → Python script → QA report ChangeLog: Added PlayerID, minutes filter, ValueFlag_Underpaid Data Dictionary Columns: PlayerID, PlayerName_raw/clean, TeamAbbrev, Salary_USD, Minutes, PTS/REB/AST, per-36 stats, TS, WS/48, Salary_perMinute, flags Lookup sheets: Teams, Players