Data Cleaning Report NBA Player Seasons

Date:December 8, 2025
Dataset: all_seasons.csv
Environment: Dbeaver (SQLite Engine)

Executive Summary

The raw dataset contained 12.844 rows for NBA player statistics. An initial audit in Dbeaver revealed structural issues (missing primary keys), mixed data types (text in numeric columns), and inconsistent country labeling.
The cleaning process I conducted entirely via SQL queries. 100% of rows were retained, with 'Undrafted' players now properly represented as NULL to ensure accurate statistical aggregation.

Cleaning Methodology

Structural Adjustements

- Issue: The source CSV file contained an unlabelled index column at the start (Unnamed:0 or empty header)
- Action: Renamed this column to row_id to serve as the table's Primary Key.
- SQL logic: ALTER TABLE… RENAME COLUMN

Data Type Standardization

- Issue: The colums draft_year, draft_round and draft_number were imported imported as Text (VARCHAR) because they contained the string 'Undrafted' alongside numbers
- Action;

Executed UPDATE queries to convert 'Undrafted' text values to NULL.
Validated that the remaining data in these columns is strictly numeric.

Results: 2,358 'Undrafted' players are now correctly handled as having no rank, allowing for mathematical operations on the drafted players.

Categorical Cleanup (Country Names)

- Issues: Inconsistent naming conventions caused players from the same nation to be split into multiple groups.
- Action: Standardized variations into single official country names using UPDATE statements.

Bosnia/Bosnia & Herzegovina → Bosnia and Herzegovina
DRC → Republic of the congo
Final schema

Column	Type	Notes
row_id	INT	Unique Primary Key
player_name	TEXT	Cleaned and Trimemd
Country	TEXT	Standardized
draft_year	INT	NULL if Undrafted
Gp,pts,reb	REAL	Game stats