

Data Quality Audit Report

Dataset Overview

Dataset Name: file1

Source Path: data\_guardian\test\_data\movies.csv

Loaded At: 2025-05-08 13:43:19

Dimensions: 9999 rows, 9 columns

Quality Scores

Completeness Score:	81.26 / 100
Uniqueness Score:	95.69 / 100
Consistency Score:	100.00 / 100
Validity Score:	90.00 / 100
Overall Quality Score:	90.36 / 100

Detected Issues Summary

Total Issues Found: 17

MissingValue (6 occurrences)

Column	Description	Severity	Affected (%)	Suggestion
--------	-------------	----------	--------------	------------

YEAR	644 standard missing values (NaN) found in column 'YEAR'.	Low	6.44	Consider imputation (mean, median, mode, model-based) or row deletion if appropriate.
GENRE	80 standard missing values (NaN) found in column 'GENRE'.	Low	0.80	Consider imputation (mean, median, mode, model-based) or row deletion if appropriate.
RATING	1820 standard missing values (NaN) found in column 'RATING'.	Medium	18.20	Consider imputation (mean, median, mode, model-based) or row deletion if appropriate.
VOTES	1820 standard missing values (NaN) found in column 'VOTES'.	Medium	18.20	Consider imputation (mean, median, mode, model-based) or row deletion if appropriate.
RunTime	2958 standard missing values (NaN) found in column 'RunTime'.	Medium	29.58	Consider imputation (mean, median, mode, model-based) or row deletion if appropriate.

Gross	9539 standard missing values (NaN) found in column 'Gross'.	High	95.40	Consider imputation (mean, median, mode, model-based) or row deletion if appropriate.
-------	---	------	-------	---

DuplicatedRows (1 occurrences)

Column	Description	Severity	Affected (%)	Suggestion
N/A (Dataset Level)	431 fully duplicated row(s) detected in the dataset.	Medium	4.31	Review and remove duplicated rows to ensure data integrity and prevent skewed analysis.

LeadingTrailingWhitespace (4 occurrences)

Column	Description	Severity	Affected (%)	Suggestion
MOVIES	3526 value(s) with leading or trailing whitespace found in column 'MOVIES'.	Low	35.26	Trim whitespace from values in this column (e.g., using <code>.str.strip()</code> ).
GENRE	9919 value(s) with leading or trailing whitespace found in column 'GENRE'.	Low	100.00	Trim whitespace from values in this column (e.g., using <code>.str.strip()</code> ).

ONE-LINE	9999 value(s) with leading or trailing whitespace found in column 'ONE-LINE'.	Low	100.00	Trim whitespace from values in this column (e.g., using <code>.str.strip()</code> ).
STARS	9999 value(s) with leading or trailing whitespace found in column 'STARS'.	Low	100.00	Trim whitespace from values in this column (e.g., using <code>.str.strip()</code> ).

MixedCaseValues (2 occurrences)

Column	Description	Severity	Affected (%)	Suggestion
MOVIES	Column 'MOVIES' contains values that differ only by case (e.g., 'Apple' vs 'apple'). Original unique count: 6817, Lowercase unique count: 6816.	Medium	100.00	Standardize casing (e.g., convert all to lowercase or title case) to ensure consistency.
ONE-LINE	Column 'ONE-LINE' contains values that differ only by case (e.g., 'Apple' vs 'apple'). Original unique count: 8688, Lowercase unique count: 8686.	Medium	100.00	Standardize casing (e.g., convert all to lowercase or title case) to ensure consistency.

MixedDataTypeInObjectColumn (2 occurrences)

Column	Description	Severity	Affected (%)	Suggestion
MOVIES	Column 'MOVIES' (object type) contains mixed data: ~15 numeric-like values and ~9984 non-numeric string values.	Medium	99.85	Investigate non-numeric values in 'MOVIES'. Standardize or clean them if the column should be numeric, or confirm mixed type is intentional.
VOTES	Column 'VOTES' (object type) contains mixed data: ~4429 numeric-like values and ~3750 non-numeric string values.	Medium	45.85	Investigate non-numeric values in 'VOTES'. Standardize or clean them if the column should be numeric, or confirm mixed type is intentional.

NumericalOutlier (2 occurrences)

Column	Description	Severity	Affected (%)	Suggestion
RATING	165 numerical outlier(s) detected in column 'RATING' using IQR method (multiplier 1.5). Examples: [3.3, 3.7, 2.7]....	Medium	1.65	Investigate outliers in 'RATING'. They may be errors or genuinely extreme values. Consider transformation, capping, or removal if they are errors.

RunTime	105 numerical outlier(s) detected in column 'RunTime' using IQR method (multiplier 1.5). Examples: [395.0, 209.0, 201.0]...	Medium	1.05	Investigate outliers in 'RunTime'. They may be errors or genuinely extreme values. Consider transformation, capping, or removal if they are errors.
---------	---	--------	------	---