

Merge Troubleshooting Guide

Duplicate Rows After Merge

Symptom: Result has more rows than expected

Cause: Many-to-many relationship (duplicates in key column)

1. Check for duplicates: `df['key'].duplicated().sum()`
2. Remove duplicates before merge: `df.drop_duplicates(subset=['key'])`

Missing Rows After Merge

Symptom: Result has fewer rows than expected

Cause: Key values don't match (spelling, case, whitespace)

1. Check key overlap: `set(df1['key']) & set(df2['key'])`
2. Clean keys: `df['key'].str.strip().str.upper()`
3. Use `how='outer'` to keep all rows

Column Name Conflicts

Symptom: Columns with '_x' and '_y' suffixes

Cause: Both DataFrames have columns with same name

1. Rename before: `df.rename(columns={'col': 'new_col'})`
2. Use suffixes: `merge(df1, df2, suffixes=('_left', '_right'))`

Data Type Mismatch

Symptom: No matches found despite matching values

Cause: Key columns have different dtypes (int vs str)

Pre-Merge Checklist:

1. Check types: `df1['key'].dtype` vs `df2['key'].dtype`
2. Convert: `df['key'] = df['key'].astype(str)`
1. `df1['key'].dtype == df2['key'].dtype` | 2. `df1['key'].duplicated().sum() == 0` | 3. `len(set(df1['key']) & set(df2['key'])) > 0`