datacamp

COURSE GLOSSARY

Cleaning Data in Python

assert statement: A programming statement that verifies a condition is True during execution and raises an AssertionError if the condition is False, commonly used to validate data-cleaning results

astype(): A pandas method used to convert a Series or DataFrame column to a specified data type (for example int, float, or category) to ensure correct processing and analysis

Blocking (record linkage): A scalability technique that restricts candidate pair generation to records that share the same value on one or more blocking columns (for example state) to avoid comparing all possible pairs

Candidate pairs / Indexer (recordlinkage): The generated set of possible record pairs (often represented by a MultiIndex) produced by an indexing strategy or indexer object, which are then compared and scored for matches

Categorical data / category dtype: Data representing a finite set of discrete values or labels (e.g., marital status) that can be stored efficiently with a category dtype and summarized by counts and unique levels

Data type (dtype): The type of values stored in a variable or column (e.g., integer, float, string, datetime, category), which determines what operations are valid and how values are interpreted

DataFrame: A two-dimensional, tabular data structure provided by pandas that stores labeled rows and columns and is the primary format for data manipulation in Python

flatten: An ndarray method that returns a one-dimensional copy of the array with all elements laid out in a single axis

JSON (JavaScript Object Notation): A lightweight, human-readable data interchange format composed of name-value pairs and arrays that maps naturally to Python dictionaries and lists

fillna(): A pandas method to replace missing values with specified constants or computed statistics (such as mean, median, or mode) as a simple imputation strategy

groupby + agg: A two-step pandas pattern where groupby() groups rows by key columns and agg() computes aggregate statistics (mean, max, sum, etc.) for each group to summarize or combine duplicate records

Minimum edit distance / Levenshtein distance: A measure of how dissimilar two strings are, defined as the smallest number of insertions, deletions, substitutions (and sometimes transpositions) required to transform one string into another

Missingness types (MCAR, MAR, MNAR): Categories describing why data is missing—Missing Completely At Random (MCAR) has no relation to other data, Missing At Random (MAR) is related to observed data, and Missing Not At Random (MNAR) depends on unobserved values

NaN / NA (missing data): A marker used to represent missing or undefined values in datasets, typically appearing as NaN for numeric types and NaT for missing datetimes

pandas .str accessor: A pandas interface for vectorized string operations on Series of text, including methods like strip(), replace(), len(), upper(), lower(), and contains()

pandas: A Python library for data analysis and manipulation that provides DataFrame objects, input/output tools, and many functions for cleaning and transforming data

pandas.concat(): A pandas function used to concatenate or append DataFrames along rows or columns, commonly used to merge non-duplicated rows after record linkage or preprocessing

qcut and cut: pandas functions for binning numeric data into discrete intervals: qcut divides data into quantiles with roughly equal counts, while cut assigns bins using specified numeric breakpoints

Record linkage: The process of identifying and linking records that refer to the same real-world entity across different data sources when unique identifiers are missing or inconsistent

Regular expression (regex): A compact, expressive pattern language used to find, match, or replace complex text patterns (such as non-digit characters) when cleaning or validating strings

String similarity (thefuzz / WRatio): A scoring approach that quantifies how similar two strings are (commonly on a 0–100 scale) using fuzzy matching algorithms such as those implemented by thefuzz (formerly fuzzywuzzy)

to_datetime(): A pandas function that parses strings or other date representations into datetime objects, with options to coerce unparseable values into missing values