Cheatsheet

Intro to Python: Pandas for Metadata Transformation and Cleanup (2022)

Read and evaluate

`pd.read_csv()`: Reads a CSV file into a DataFrame.

`df.head()`: Returns the first n rows of the DataFrame.

`df.tail()`: Returns the last n rows of the DataFrame.

`df.columns`: Returns the column labels of the DataFrame.

`df.shape`: Returns a tuple (rows, columns) of the DataFrame.

`df.empty`: Indicates if the DataFrame is empty.

`df['title'].unique()`: Returns unique values of Series.

`df['title'].value_counts()`: Returns Series containing counts of unique value in column (example 'title').

Access rows, columns, and cells

`df['title']` or `df.title`: Select single column with specific name (example 'title').

`df.loc[]`: Access rows & columns by label(s) or a boolean array.

`df.iloc[]`: Purely integer-location based indexing for selection by position.

`df.iat[1, 2]`: Access single value by index.

`df.at[4, 'A']`: Access single value by label.

Clean up

`pd.isna()`: Detects missing values.

`pd.notna()`: Detects non-missing values.

`df.dropna()`: Removes missing values from the DataFrame.

`df.duplicated()`: Returns boolean Series of duplicate rows.

`df.drop_duplicates()`: Removes duplicate rows from DataFrame.

`Series.apply()`: Invoke function on values of Series.

`Series.str.zfill()`: Pads Series with zeros.

`Series.str.strip()`: Strips whitespaces from Series.

Loop through rows

`df.iterrows()`: Loops through DataFrame rows as (index, Series) pairs.

Merge DataFrames

`pd.merge()`: Merge DataFrame or named Series objects with a database-style join.

`how="left"`: Merges on all ids from left DataFrame. Ids not in left DataFrame will not be included.

`how="right"`: Merges on all ids from right DataFrame. Ids not in right DataFrame will not be included.

`how="outer"`: Merges on all ids from both DataFrames.

`how="inner"`: Merges only on ids found in both DataFrames. Ids found in only one DataFrame will not be included.

Reshaping

`df.explode()`: Transforms each element of a list-like to a row, replicating index values.

`df.pivot()`: Reshape data (produce a "pivot" table) based on column values.

`df.pivot_table()`: Create a spreadsheet-style pivot table as a DataFrame.

`lambda`: An anonymous (unnamed) function that applies arguments to various parameters and returns an expression (outcome).

`df.melt()`: Unpivot a DataFrame from wide to long format, optionally leaving identifiers set.

Sort DataFrame

`df.sort_values()`: Sort by the values along either axis.

Create new CSV

`df.to_csv()`: Writes the DataFrame to a CSV file.

Resources

[pandas API reference](#)   [pandas User Guide](#)