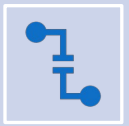
DAY 4

Instructor: Balu Mohandas Menon

Christian B. Wiberg
Philip Jess Teining

COMBINING DATASETS: MERGE AND JOIN

One essential feature offered by Pandas is its high-performance.



in-memory join and merge operations.

The main interface for this is the pd.merge function.

CATEGORIES OF JOINS

- The pd.merge() function implements a number of types of joins:
- *one-to-one*,
- *many-to-one*, and
- *many-to-many* joins
- All three types of joins are accessed via an identical call to the pd.merge()
- interface; the type of join performed depends on the form of the input data.

SPECIFICATION OF THE MERGE KEY

The on keyword

The left_on and right_on keywords

The left_index and right_index keywords

HANDLING MISSING DATA

The difference between data found in many tutorials and data in the real world is that real-world data is rarely clean and homogeneous

many interesting datasets will have some amount of data missing.

To make matters even more complicated, different data sources may indicate missing data in different ways.

discuss how Pandas chooses to represent it, and demonstrate some built-in Pandas tools for handling missing data in Python

TRADE-OFFS IN MISSING DATA CONVENTIONS

revolve around one of two strategies:

using a mask that globally indicates missing values,

or choosing a sentinel value that indicates a missing entry.

OPERATING ON NULL VALUES

isnull(): Generate a boolean mask indicating missing values

notnull(): Opposite of isnull()

dropna(): Return a filtered version of the data

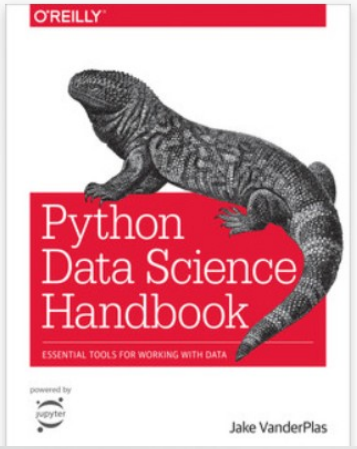
fillna(): Return a copy of the data with missing values filled or imputed

FILLING NULL VALUES

provides the fillna() method

```
# forward-fill
data.fillna(method='ffill')
# back-fill
data.fillna(method='bfill')
```

REFERENCE



Python Data Science Handbook

by Jake VanderPlas
Released November 2016
Publisher(s): O'Reilly Media, Inc.
ISBN: 9781491912058

Read it now on the O'Reilly learning platform with a 10-day free trial.

O'Reilly members get unlimited access to live online training experiences, plus books, videos, and digital content from O'Reilly and nearly 200 trusted publishing partners.

- This notebook contains an excerpt from the Python Data Science Handbook by Jake VanderPlas; the content is available on GitHub.
- The text is released under the CC-BY-NC-ND license, and code is released under the MIT license. If you find this content useful, please consider supporting the work by buying the book!