



Week 13

Introduction to Programming and Numerical Analysis

Ebener, Luca

UNIVERSITY OF COPENHAGEN



Overview

- Lecture topics
- Problem set 4
- Work on problem set 4

Data Merging in Python

Types of Joins

- **Inner Join:** Combines only the rows that have matching values in both datasets.
- **Outer Join:** Combines all rows from both datasets, filling in missing values with NaNs where no match is found.
- **Left Join:** Takes all rows from the left dataset and matching rows from the right dataset.

Implementation in Pandas

- Syntax: `pd.merge(left_df, right_df, how='inner/outer/left', on='key_column')`

Fetching Data with APIs

Tips for Working with APIs

- An API (Application Programming Interface) serves as a communication line between your software (e.g., Python) and another software (e.g., Statistics Denmark's database).

Python Packages for APIs

- Some APIs have associated Python packages (e.g., DstApi for Statistics Denmark or pandas-datareader) that can connect to the API, fetch, and parse data for you.
- Always refer to the documentation provided for the API you're utilizing. It contains crucial information about endpoints, parameters, authentication, and response formats.

Transforming Data: Split-Apply-Combine Approach

Split-Apply-Combine Approach

- The **split-apply-combine** approach involves splitting the dataset into groups, applying a function to each group, and combining the results into a new dataset.

Methods for Transformation

- `.apply`: Applies a function along an axis of the DataFrame.
 - Useful for custom operations on entire rows or columns.
- `.transform`: Performs a transformation for the whole group.
 - Often used for standardizing or normalizing data within groups.
- `.agg`: Allows to calculate more than one statistic.
 - Used for summarizing data, like calculating group means or counts.

Problem Sets

Completion of Previous Problem Sets

- If you didn't complete problem set 3 last time, prioritize completing it first.
- Refer to last week's slides for notes on the problem set and any known bugs.

Problem Set 4: Fetching Data from Statistics Denmark

- In problem set 4, you will be fetching data from Statistics Denmark using DstApi.
- If unsure about which arguments to pass to the `.get_data()` method, consult the documentation.

Working with Pandas

- The tasks are quite syntax-heavy so don't shy away from looking at the solutions.