Data Wrangling in Python

- **Abstract Data Types (ADTs) for Data Wrangling**

  - **NumPy: Numerical Python**
    - NumPy arrays and their attributes
    - Creating NumPy arrays
    - Indexing, slicing, and striding of NumPy arrays
    - Multi-dimensional slicing
    - View vs. Copy of subarrays
  - **Pandas**
    - Pandas Data Model (ADT):
      - 1D Index
      - 1D Series
      - 2D DataFrame
      - 3D Panel
    - Attributes of Pandas ADTs
    - Info Methods for Pandas ADTs
    - Indexing and slicing of Pandas ADTs:
      - Index-based: .iloc and .iat
      - Label-based: .loc and .at
      - Chaining index- and label-based access
    - Pandas CRUD: Create, Read, Update, Delete

- **Data Input and Output (Read and Write)**

  - **NumPy Arrays: I/O helper functions**
  - **Structured Text: CSV and XLSX**
    - Text with Missing value
  - **JSON: JavaScript Object Notation file format**
  - **XML: eXtensible Markup Language file format**
    - JSON vs. XML
    - XML Modules and ElementTree
      - Creating XML
      - Setting and getting attributes
      - Parsing XML
      - Exception handling for XML operations

- Converting XML to Pandas DataFrame
- Converting XML to JSON
- Converting XML to dictionary

o

o **Data Manipulation**

- **Vector operations for faster data manipulation**
- **Subset, Filter vs. Split**
  - Selecting or excluding variables (columns)
  - Filtering or conditional sub-setting
  - Re-encoding categorical variables
  - Adding / deleting rows and columns
  - Splitting for predictive modeling: training, validation, and testing
- **Combining Multiple Datasets**
  - Concatenate and Append
    - Handling duplicate indices
    - Concatenation with different columns: inner and outer joins, join_axes
  - Merge and Join
    - Joins: inner, outer, left, right
    - Joins with duplicate entries: many-to-one or many-to-many
    - Joins by specifying merge keys: on, left_on / right_on, left_index / right_index
    - Mixing index-based and name-based keys
    - Joins with conflicting values in key columns: suffixes
- **Handling Missing Data**
  - NaN, None, Null
  - Finding missing data
  - Dropping missing data
  - Inserting data for missing data
- **Regular Expressions: Introduction**
  - Re module
  - Regex Object
  - Match Objects
  - Groupings
  - Match Flags
  - String Replacements

- **Unit Testing: Introduction**
  - Automated testing framework: unittest
  - Testing a function
  - Creating test cases
  - Failing test and how to respond to it
  - Assertion methods

Not Covered in This Course:

A. Data transformations: Box-Cox, ladder of powers
B. Discretization, normalization, and scaling
C. Gather and spread
D. Handling missing values for predictive modeling and analysis
E. Outliers and influential values
F. Aggregation and Grouping
  a. GroupBy: Split, Apply, Combine
  b. Aggregation, filter, transform, apply
  c. Pivot Tables
G. Dealing with Dates and TimeSeries
H. Hierarchical Indexing: Multi-indexing
  a. Multi-index creation
  b. Indexing and slicing
  c. Rearranging multi-indices:
    - **Sorted, unsorted**
    - **Stacking, unstacking**
    - **Setting, resetting**
  d. Data aggregation on multi-indices