Лабораторна робота №6

Попередня обробка даних в Pandas

Мета роботи: Ознайомитись з операціями попередньої обробки даних Pandas.

Короткі теоретичні відомості

Функції для читання файлів в Pandas:

- pd.read_csv() – можна зчитуваті дані як з локального csv-файлу, так і за інтернет-посиланням. За замовчанням використовує кому як роздільник.

Основні параметри pd.read_csv():

sep – роздільник

header – заголовок, за замовчуванням header=0, тобто заголовком стає перший рядок.

names – заголовки стовпців

index_col – встановлює стовпець (ці) як індекси

prefix – префікс, що додається до номерів стовпців за відсутності заголовка

dtype – типа даних

skiprows – які рядки пропустити, вказавши кількість або індекс

nrows – кількість рядків, які потрібно прочитати

na_values – які символи розпізнавати як відсутні значення

skip_blank_lines – чи пропускати порожні рядки

comment – вказує, з яких символів починається коментар

converters – словник функцій, які потрібно застосувати до певних стовпців

thousands – роздільник тисяч

decimal – роздільник цілої частини, за замовчанням крапка

- pd.read_table() - за замовчанням використовує табуляцію як роздільник. Має аналогічні до pd.read_csv() параметри.
- Для читання файлу типу json використовується функція pd.read_json

Параметр orient визначає формат рядків JSON. Основні його значення:

'split' : словник {index -> [index], columns -> [columns], data -> [values]}

'records' : список [{column -> value}, ... , {column -> value}]

'index' : словник {index -> {column -> value}}

'columns' : словник {column -> {index -> value}}

'values' : масив значень

- Для читання екселівських файлів використовується функція pd.read_excel
- Для читання html-файлів використовується функція pd.read_html

Попередня підготовка даних не має встановленого переліку операцій; єдина мета полягає в тому, щоб дані після обробки були кориснішими, ніж до неї. На практиці існує три основні задачі, пов'язані з процесом попередньої обробки даних:

• Очищення даних

• Трансформація даних

• Збагачення даних

Основні завдання очищення даних наступні:

- Перейменування
- Сортування та переупорядкування
- Перетворення типів даних
- Обробка повторюваних даних
- Виправлення відсутніх або недійсних даних
- Фільтрація до потрібної підмножини даних

Завдання до лабораторної роботи

Створити програму, яка виконує наступні завдання, використовуючи файл відповідно до варіанту:

1. Читає файл та змінює назви стовпців.
2. Знаходить проблеми з даними та виконує попередню обробку даних для усунення цих проблем.

Оформити звіт. Звіт повинен містити:

— титульний лист;
— код програми;
— результати виконання коду.

Продемонструвати роботу програми та відповісти на питання стосовно теоретичних відомостей та роботи програми.

Варіант 1: Version 1.xlsx
Варіант 2: Version 2.html
Варіант 3: Version 3.json
Варіант 4: Version 4.xlsx
Варіант 5: Version 5.json
Варіант 6: Version 6.html
Варіант 7: Version 7.xlsx
Варіант 8: Version 8.html
Варіант 9: Version 9.json
Варіант 10: Version 10.xlsx
Варіант 11: Version 11.json
Варіант 12: Version 12.html
Варіант 13: Version 13.xlsx
Варіант 14: Version 14.html
Варіант 15: Version 15.json