Мельничук Я.О., магістрант,
Кравченко С.М., старший викладач,
Житомирський державний технологічний університет

АНАЛІЗ ДАНИХ ТА ВІЗУАЛІЗАЦІЯ ЗА ДОПОМОГОЮ МОВИ PYTHON

Python – високорівнева, мультипарадигмова мова програмування загального призначення, орієнтована на підвищення продуктивності розробників та читабельності коду. На відміну від інших інтепретованих мов програмування, Python активно використовується для проведення наукових розрахунків.

В області аналізу та візуалізації даних, Python конкурує з багатьма предмето-орієнтованими мовами програмування та інструментами із відкритим вихідним кодом та із комерційними, а саме R, MATLAB, SAS, Stata та іншими.

Для аналізу та візуалізації даних у Python стануть корисними наступні інструменти:

1. NumPy – основний пакет для виконання наукових розрахунків, надає засоби для ефективної роботи із багатовимірними масивами, операції лінійної алгебри, перетворення Фур'є, генерація випадкових чисел.

2. Matplotlib – бібліотека для створення графіків та інших способів візуалізації двовимірних даних.

3. SciPy – збірка пакетів для вирішення стандартних обчислювальних задач як то: численне інтегрування та вирішення диференціальних рівнянь (scipy.integrate), алгоритми роботи із розрідженими матрицями і рішення розріджених систем лінійних рівнянь.

4. Pandas – надає функції та структури даних для поліпшення роботи із структурованими даними. Пакет надає можливість будувати зведені таблиці, виконувати угрупування, надає доступ до табличних даних, а при наявності matplotlib дає можливість будувати графіки на отриманих наборах даних.

Pandas представляє дві основні структури даних: DataFrame та Series.

Структура Series є об'єктом, схожим на одновимірний масив (список у python), але його відмінною рисою є наявність асоційованих міток, так званих індексів, уздовж кожного елемента із списку. Така особливість перетворює його в асоціативний масив або словник в Python.

У об'єкта Series є атрибути, через які можна отримати список елементів та індексів, а саме values та index. Також до елементів об'єкта Series можна здійснювати доступ за декількома індексами та виконувати масове присвоєння. Series можна легко фільтрувати, використовуючи знаки більше/менше, та одночасно застосовувати математичні операції над об'єктами.

Наступною головною структурою є DataFrame. DataFrame найкраще уявляти собі у вигляді звичайної таблиці і це правильно, адже DataFrame є табличній структурою даних. У будь-якій таблиці завжди присутні рядки і стовпці. Стовпцями в об'єкті DataFrame виступають об'єкти Series, рядки яких є їхніми безпосередніми елементами.

Об'єкт DataFrame має 2 індекси: по рядкам та по стовбцям. Якщо індекс по стовбцям не заданий, то pandas встановить цілочисельний індекс RangeIndex від 0 до N-1, де N кількість рядків в таблиці. Доступ до рядків можливий декількома способами: .loc – використовується для доступу по рядковій мітці, .iloc – використовується для доступу по числовому індексу (починаючи з 0), механізм фільтрації реалізований як і у об'єктів Series. Також реалізований функціонал створення, видалення та перейменування стовбців.

Pandas надає можливість читати та записувати дані у всіх популярних форматах для збереження даних: csv, excel, html, json буфер обміну та інші. Так для ініціалізації DataFrame із csv файлу достатньо викликати метод read_csv.

Групування даних -це операція, яка найбільш використовуються при аналізі даних. В pandas за групування даних відповідає метод groupby, який потрібно викликати на об'єкті DataFrame. Також присутня можливість будувати зведені таблиці, аналогічні зведеним таблицям у MS Excel, для цього потрібно використати метод pivot_table.

Для візуалізації даних pandas використовує бібліотеку matplotlib. З її допомогою можна з легкістю будувати діаграми. Потрібно лише імпортувати модуль matplotlib.pyplot та викликати метод plot() на об'єкті DataFrame.

Як зазначено вище, для мови Python реалізовано багато інструментів для аналізу та візуалізації даних, а завдяки великій підтримці прихильниками відкритого програмного забезпечення, буде створено багато інших інструментів, а існуючі будуть покращуватись. Отже, Python є чудовим інструментом для аналізу та візуалізації даних, а його простота та лаконічний синтаксис дозволить розробникам швидко вирішувати поставлені задачі.