https://doi.org/10.31891/2307-5732-2025-357-33
УДК 004.942

KRAVCHUK OLHA
Khmelnytsky National University
https://orcid.org/0000-0001-6937-5001
e-mail: kravchukoa2@gmail.com

PROGRAMMING BASICS: PYTHON FOR DATA PROCESSING

The article discusses aspects of efficient data processing. Considerable attention is paid to the problems that arise in data modelling and forecasting. In today's digital world, data processing is becoming an essential skill in many professions, from business analytics to psychology. Today, almost every action leaves a digital footprint - online purchases, page views, answers to questionnaires, health indicators, GPS navigation, etc. This generates huge amounts of data that need to be processed, analysed, and used to make decisions.

A special place in the paper is occupied by a description of the possibilities of software data processing using the Python programming language, which is becoming increasingly popular due to its simplicity, flexibility, open source, convenience of working with data in various formats, as well as many developed packages that facilitate fast and efficient information processing. NumPy, Pandas, which provide data structures and functions that make working with structured data simple and fast, the most popular data visualisation tool Matplotlib and Seaborn, they help to create graphs, charts and other visual representations of data, packages for various computational tasks SciPy, Statsmodels, and a machine learning-oriented package Scikit-learn which provides simple and effective tools for data analysis, it contains a variety of algorithms for classification, regression, and clustering. An example of using Python for data processing tasks is given. This example demonstrates how Python can be used to easily read data, perform basic statistical processing, and build visualisations, which is useful for psychological or sociological research.

The paper contains an analysis of current scientific publications and identifies possible directions for future research.

Keywords: *Python, data processing, modelling, forecasting, programming.*

КРАВЧУК ОЛЬГА
Хмельницький національний університет

ОСНОВИ ПРОГРАМУВАННЯ: PYTHON ДЛЯ ОБРОБКИ ДАНИХ

У статті розглянуто аспекти ефективної обробки даних. Значна увага приділяється проблемам, що виникають при моделюванні і прогнозуванні даних. Особливе місце у роботі займає опис можливостей програмної обробки даних з використанням мови програмування Python, яка набуває все більшої популярності завдяки простоті, гнучкості, відкритому коду, зручності роботи з даними у різних форматах, а також багатьом розробленим пакетам, які сприяють швидкій та ефективній обробці інформації. Розглядаються NumPy, Pandas, які надають структури даних і функції, що дозволяють зробити роботу зі структурованими даними простою і швидкою, найпопулярніший інструмент для візуалізації даних Matplotlib, пакети для різних обчислювальних задач SciPy, Statsmodels, а також пакет, орієнтований на машинне навчання Scikit-learn. Наводиться приклад використання Python щодо задач з оброки даних.

Робота містить аналіз актуальних наукових публікацій та визначає можливі напрями майбутніх досліджень.

Ключові слова*: Python, обробка даних, моделювання, прогнозування, програмування.*

Стаття надійшла до редакції / Received 12.06.2025
Прийнята до друку / Accepted 28.06.2025

Formulation of the problem in general terms and its connection with important scientific or practical tasks

In today's digital world, data processing is becoming an essential skill in many professions, from business analytics to psychology. Today, almost every action leaves a digital footprint - online purchases, page views, answers to questionnaires, health indicators, GPS navigation, etc. This generates huge amounts of data that need to be processed, analysed, and used to make decisions. Data processing is necessary not only for programmers, but also for marketers to analyse the market and customer behaviour; psychologists to analyse questionnaires and research; doctors to study treatment outcomes; educators to assess student progress; and journalists to prepare fact-based infographics.

Knowing how to work with data is not just a bonus, but a key digital literacy. The ability to read data from tables or databases, analyse statistics, visualise results, automate processing, model and forecast makes a specialist more competitive.

The Python programming language has gained great popularity due to its simplicity, flexibility, and powerful data analysis tools. Python is a high-level, versatile programming language that is widely used for scientific computing, data analysis, web application development, machine learning, and task automation.

Thus, Python is an excellent tool for data processing and deepening programming knowledge. Its accessibility, a large number of training materials and libraries make it an ideal choice for both humanities and technical specialists.

Analysis of the latest research and publications

In 2024-2025, Python remains the leading language for data processing, with a focus on performance, scalability, and automation [1]. Here are the key trends and innovative tools.

First, modelling with quadratic polynomials. A study by Sipakov et al. 2024 demonstrates the effectiveness of using quadratic models in Python to analyse complex data. The use of the NumPy, Pandas, and scikit-learn libraries allows for the accurate modelling of nonlinear dependencies, providing high accuracy while maintaining simplicity of implementation.

Secondly, optimisation of data processing using databases. PyTond development offers Python integration with the power of SQL databases. This approach allows you to process large amounts of data more efficiently using internal database optimisations, which significantly improves performance compared to traditional Python libraries.

Third, improving Pandas performance using compiler technologies. At PyCon India 2024, we presented FireDucks, a library that optimises Pandas code execution using compilation techniques. This allows to significantly reduce data processing time without changing the existing code.

Fourth, new tools for data analysis. In 2024, new libraries appeared that expand Python's capabilities in data analysis:

1) Dask: supports parallel processing of large data sets that do not fit in memory;
2) VAEX: allows you to work with large data sets using efficient algorithms and memory;
3) PyCaret: automates machine-learning processes, simplifying model building.

Fifth, these are the trends of 2025: AutoML and new frameworks. In 2025, AutoML, an automated machine-learning framework that simplifies model creation, is expected to grow in popularity. In addition, new frameworks, such as Pathway, are emerging that combine streaming and batch data processing, providing flexibility and speed [2].

Formulation of the objectives of the article

In this article, we will consider the aspects of efficient data processing using Python and the problems that arise when modelling and forecasting data.

Presentation of the main material

For data processing, it is advisable to use the technologies that are most appropriate for a particular type of data and the tasks that the researcher is solving.

In some tasks, it is sufficient to use a convenient, understandable and affordable tool such as MS Excel, which has wide capabilities and an analysis package, although somewhat limited, suitable for obtaining results in a first approximation to understand the nature of the data. Spreadsheets cannot be used to run a production model, such as artificial intelligence, but they can be used to analyse the nature of the data, model and predict the outcome. This result can be obtained using the classical approaches of probability theory and mathematical statistics to normalise data, correlation and regression analysis, estimation of predicted point and interval values, and procedures to determine optimal solutions to linear and nonlinear optimisation problems.

To use data processing automation in a programmatic mode, you need to know a particular programming language. However, to understand the analysis used in processing technologies and various application packages such as Statistika, SPSS, etc., it is not necessary to know how to write code. These powerful tools include a variety of analyses, including regression, factor, cluster, neural network, and many others, and also provide a graphical display of the results, if the dimensionality and the task formulation allow. There are several stages in the process of working with data [9].

1). Determining the purpose of the study. This involves preparing a project assignment and evaluating the research objective.

2). Collecting and preparing data, the so-called ‘reconnaissance analysis’. Certain difficulties arise at this stage. The data may be scattered, in different formats, and need to be normalised and brought to homogeneity. Matrices may not be fully filled in, degenerate. It is imperative to choose an algorithm to fill in the blanks. Often, there are significant deviations in the data, i.e. outliers that need to be eliminated, i.e. data cleaning. Otherwise, no modelling methods will lead to an adequate model. Thus, the process of data preparation is very painstaking and routine, almost ‘manual’, requiring an intellectual approach and understanding of the research objective.

3). Data analysis and modelling, i.e. model selection and estimation of its parameters, is called ‘model training’ in machine learning. At this stage, you need to understand how the data are related to each other, estimate data distributions, identify and eliminate outliers, and check for multicollinearity in the system and negative phenomena such as heteroscedasticity and autocorrelation, which require additional variable transformations and special methods. For this purpose, certain statistical methods and simple modelling are used. Questions arise: are the factors and the indicator under study interrelated, is there multicollinearity in the data system, can the number of variables be reduced and thus simplify the model, what form of dependence to choose for modelling, how to reduce the model to a linear form, etc. This stage requires knowledge of the subject area, as well as mathematics, probability theory and mathematical statistics. Only after these studies and data transformations can you use ready-made solutions - software packages. The modelling process itself - ‘model training’ - means building different models on the same set of data randomly selected from the total population. The amount of data can be varied using the parameters set for the chosen method. You can train a data set several times, changing the parameters, and thus achieve the best result. Therefore, building a model is an iterative process and requires the skills of a researcher.

4). Checking the adequacy of the model and the significance of the model factors. After obtaining the best result (for example, comparing the sum of squared deviations and choosing the set of parameters that gives the smallest of all), the quality of the model is assessed using statistical criteria. If the quality is unsatisfactory, the model needs to be 'retrained'.

5). Application of the model to unfamiliar data - the so-called 'training set' is selected from the same sample - 'predictive modelling', i.e., a forecast is determined.

The described approach is used for modelling and forecasting tasks in machine learning. Python, for example, has its own Scikit-learn library with various algorithms. Machine learning is currently a very popular and promising technology among data scientists. The machine learning market is growing rapidly. Since 2016, its volume has passed the $1 billion mark, and by 2025, according to forecasts, it may increase to $39.98 billion. 60% of companies in the world are already using machine learning [3].

The tasks that can be solved by machine learning include modelling and forecasting of indicators depending on one or more factors or optimisation tasks. Both traditional methods of econometric analysis are used, including single-factor, multifactor models based on the least squares method, and non-traditional ones, such as decision trees with a large number of set parameters, which provide flexibility in modelling model parameters.

The so-called 'neural networks' are gaining popularity. The modelling uses the concept of risk, the quantitative features of which are calculated in accordance with the numerical characteristics of discrete and continuous random variables.

Over the past decade, Python has become one of the most important programming languages used in data science, machine learning, and general-purpose software development in academia and industry. Improved libraries for Python have made it a serious competitor in solving problems of creating data processing applications.

The Python programming language is a high-level, versatile programming language that is widely used for scientific computing, data analysis, web application development, machine learning, and task automation.

Python is an open-source programming language that is easy to learn thanks to its clear syntax. It has a large community that allows you to quickly find support, examples, and ready-made solutions. In addition, Python supports numerous libraries for working with data:

1) NumPy is a library for working with data arrays. It allows you to perform mathematical operations on large data sets with high performance [4];

2) Pandas is a powerful tool for data processing and analysis. It allows you to work with data tables, perform various filtering, sorting and aggregation operations [5];

3) Matplotlib and Seaborn are data visualisation tools. They help to create graphs, charts and other visual representations of data;

4) Scikit-learn is a machine learning library that provides simple and effective tools for data analysis. It contains various algorithms for classification, regression, and clustering [1,7].

Python supports a variety of data types, including: int - integers; float - fractional numbers; str - strings (text); list - lists; dict - dictionaries. This allows you to conveniently store and process information in the form of arrays, tables, or structured records [7].

The Pandas library allows you to work with data tables (DataFrame), import data from Excel, CSV, databases, as well as perform filtering, grouping, calculating statistics, etc.

Example:

```
import pandas as pd

data = pd.read_csv("results.csv")
print(data.head())
```

Fig.1. Example

The Matplotlib and Seaborn libraries allow you to create graphs, charts, and heat maps to better understand the structure of your data [6].

Matplotlib, один із найпопулярніших варіантів для обробки даних, має різноманітні програми. Його можна використовувати для кореляційного аналізу змінних, для візуалізації довірчих інтервалів моделей і розподілу даних для отримання розуміння, а також для виявлення викидів за допомогою діаграми розсіювання.

Ось деякі з основних функцій Matplotlib для обробки даних: може бути заміною MATLAB, вільний і відкритий джерело, підтримує десятки серверних програм і типів виводу, низьке споживання пам'яті [8].

Example:

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.histplot(data['Стрес'])
plt.title("Розподіл рівня стресу")
plt.show()
```

Fig.2. Example

Scikit-learn включає посилення градієнта, DBSCAN, випадкові ліси в класифікації, регресію, методи кластеризації та опорні векторні машини.

Бібліотека Python часто використовується для таких програм, як кластеризація, класифікація, вибір моделі, регресія та зменшення розмірності.

Ось деякі з основних функцій Scikit-learn для науки про дані:

1) класифікація та моделювання даних;
2) попередня обробка даних;
3) вибір моделі;
4) наскрізні алгоритми машинного навчання.

Here is an example of using Python for data processing tasks.

Task: analyse the level of stress among students.

Goal: read data from a CSV file, calculate the mean, standard deviation, and build a histogram of stress distribution.

What does it do? Visually see how variable the results are. Understand if there are any skews (e.g., stress is higher than normal). Use this data to make decisions (e.g., the need for interventions).

Below is a sample code that demonstrates how you can process the results of a student stress survey:

```
python

import pandas as pd
import matplotlib.pyplot as plt

# Зчитування даних
data = pd.read_csv("students_stress.csv")

# Обчислення статистичних показників
mean_stress = data['Стрес'].mean()
std_stress = data['Стрес'].std()

print(f"Середній рівень стресу: {mean_stress:.2f}")
print(f"Середнє квадратичне відхилення: {std_stress:.2f}")

# Побудова гістограми
plt.hist(data['Стрес'], bins=10, color='skyblue', edgecolor='black')
plt.title("Розподіл рівня стресу серед студентів")
plt.xlabel("Рівень стресу")
plt.ylabel("Кількість студентів")
plt.grid(True)
plt.show()
```

Fig.3. Example

This example demonstrates how Python can be used to easily read data, perform basic statistical processing, and build visualisations, which is useful for psychological or sociological research.

Thus, Python is a great tool for those who want to start analysing data without deep programming knowledge. Its accessibility, a large number of tutorials and libraries make it an ideal choice for both humanities and technical specialists. Python remains the main language for data analysis, thanks to its flexibility and wide range of libraries. Performance optimisation is becoming a key focus, with an emphasis on the use of databases and compiler technologies. New tools are expanding the possibilities of analysing large and complex data sets. AutoML and new frameworks open up new horizons for automation and efficiency in data processing.

Conclusions from this study and prospects for further research in this area

Studies show that Python remains the main tool for data processing, especially in the educational environment, social sciences, business, and research. Modern libraries (Pandas, Dask, PyCaret) allow you to work with both small and very large amounts of data, providing flexibility and scalability. Integration of Python with databases and cloud platforms allows automating data processing and exchange, which is critical in multidisciplinary research.

Prospects for further research include the development of Python tools for the humanities and social sciences: creating specialised libraries for processing questionnaires, text analysis, and behavioural data; using artificial intelligence and AutoML to simplify model building for behavioural prediction, assessment of psycho-emotional state, etc.; studying the effectiveness of new frameworks: investigating the performance of tools on real data sets in various fields (education, psychology, sociology); integrating Python with no-code and low-code platforms, which allows us to attract a wider range of specialists without deep technical training.

References

1. Buki. (n.d.). *Python у науці про дані: Як використовувати Python для аналізу даних та машинного навчання*. Retrieved September 15, 2025, from https://buki.com.ua/blogs/python-u-nauci-pro-dani-iak-vikoristovuvati-python-dlia-analizu-danix-ta-masinnogo-navcannia/

2. Canning, J., Broder, A., & Lafore, R. (2023). *Data structures & algorithms in Python*. Pearson Education, Inc.

3. Chupilko, T., Ulyanovska, Y., Mormul, M., & Lagoda, A. (2021). Python for data processing and modelling of financial and economic indicators. *Information Technology and Computer Innovation (ITCI)*, 51(2), 68–77. https://doi.org/10.32782/ITCI.2021.2.10 *(якщо DOI є – треба уточнити, якщо ні, залишаємо без нього)*

4. Mind the Graph. (2022, July 14). *Python in research*. Mind the Graph Blog. https://mindthegraph.com/blog/uk/python-in-research/

5. Neoversity. (n.d.). *Основи Python для початківців: Чому це важливий інструмент для Data Science ma Analytics*. Retrieved September 15, 2025, from https://neoversity.com.ua/blog/osnovi-python-dlya-pochatkivciv-chomu-ce-vazhliviy-instrument-dlya-data-science-ta-analytics

6. Python Guide. (n.d.). *Python Guide*. Retrieved September 15, 2025, from https://pythonguide.rozh2sch.org.ua/

7. DOU. (2022, November 15). *Чи потрібен Python для Data Science?* Retrieved September 15, 2025, from https://dou.ua/forums/topic/40798/

8. Unite.AI. (2023, February 20). *10 best Python libraries for data science*. https://www.unite.ai/uk/10-best-python-libraries-for-data-science/

9. Kravchuk, O. A., Synyuk, N., & Kravchuk, D. (2025). Review and analysis of data processing information technologies in the course of modern computer science. *Herald of Khmelnytskyi National University. Technical Sciences Series*, (2), 511–514.