

Analisis Data Menggunakan Python: Memperkenalkan Pandas dan NumPy

Ariya Pannadhitthana Candra¹

¹ Universitas LIA, Jakarta Selatan, Indonesia

lariya@universitaslia.ac.id

ARTICLE INFO

Submit	02-01-2025	Review	09-01-2025
Accepted	11-02-2025	Published	17-03-2025

ABSTRACT

Dalam artikel ini, saya akan membahas pentingnya analisis data di era digital saat ini. Volume data yang terus meningkat, yang diperkirakan akan mencapai 175 zettabyte pada tahun 2025 menurut IDC (Porter & Zingaro, 2024), menuntut adanya alat dan metode yang efisien untuk analisis. Saya percaya bahwa analisis data memainkan peran krusial dalam pengambilan keputusan bisnis dan dapat meningkatkan produktivitas hingga 20% (McKinsey, sebagaimana dikutip dalam Porter & Zingaro, 2024). Selain itu, analisis data memberikan wawasan berharga di berbagai sektor, seperti kesehatan—misalnya dalam memprediksi wabah penyakit—dan pendidikan, yang membantu dalam memahami pola belajar siswa (Sun & Liu, 2024). Fokus artikel ini adalah pada pengenalan analisis data menggunakan Python, dengan memanfaatkan pustaka Pandas dan NumPy. Saya memilih Python karena sifatnya yang ramah pengguna, sintaksis yang jelas, dukungan komunitas yang kuat, serta kemampuannya untuk terintegrasi dengan alat lain seperti Jupyter Notebook (Nur & Cahyani, 2024; Surbakti et al., 2024). Dengan Pandas, saya dapat dengan mudah memanipulasi dan menganalisis data terstruktur melalui struktur data Series dan DataFrame. Sementara itu, NumPy memungkinkan operasi numerik yang efisien pada array multidimensional. Tujuan saya adalah memberikan pemahaman dasar tentang analisis data menggunakan Python, menjelaskan fungsi-fungsi Pandas dan NumPy melalui contoh praktis, termasuk studi kasus analisis dataset penjualan dan data cuaca. Saya bertujuan untuk menunjukkan relevansi keterampilan ini dalam membangun fondasi yang kuat untuk karir di bidang ilmu data. Selain itu, artikel ini juga akan mencakup panduan instalasi dan pengaturan lingkungan kerja Python, serta rekomendasi untuk pembelajaran lebih lanjut.

Kata kunci: *Python, NumPy, Pandas, Analisis data*

1. Introduction

Di era digital saat ini, data telah menjadi salah satu aset terpenting bagi organisasi dan individu. Dengan meningkatnya volume data yang dihasilkan setiap hari, kemampuan untuk menganalisis dan menarik wawasan dari data tersebut menjadi sangat krusial. Menurut laporan dari International Data Corporation (IDC), volume data global diperkirakan mencapai 175 zettabyte pada tahun 2025 (Porter & Zingaro, 2024). Dalam konteks ini, penggunaan alat dan bahasa pemrograman yang efisien untuk analisis data menjadi sangat penting. Python, sebagai salah satu bahasa pemrograman yang paling populer, menyediakan berbagai pustaka yang kuat untuk analisis data, di antaranya Pandas dan NumPy.

Pandas dan NumPy adalah dua pustaka Python yang paling banyak digunakan dalam komunitas data science. Pandas dirancang untuk memudahkan manipulasi dan analisis data, sementara NumPy memberikan dukungan untuk operasi numerik yang efisien. Dengan memanfaatkan kedua pustaka ini, analisis data dapat melakukan tugas-tugas kompleks dengan lebih cepat dan efisien. Oleh karena itu, pemahaman mendalam tentang cara kerja Pandas dan NumPy sangat penting bagi siapa saja yang ingin berkarir di bidang data science. Analisis data memiliki peran yang tidak dapat diabaikan dalam pengambilan keputusan bisnis. Dengan data yang tepat, perusahaan dapat mengidentifikasi tren pasar, memahami perilaku konsumen, dan mengoptimalkan operasi mereka. Menurut survei yang dilakukan oleh

McKinsey, perusahaan yang menggunakan data analitik dalam pengambilan keputusan mereka dapat meningkatkan produktivitas mereka hingga 20% (Porter & Zingaro, 2024). Hal ini menunjukkan bahwa analisis data bukan hanya sebuah pilihan, tetapi suatu keharusan untuk tetap kompetitif di pasar.

Di sisi lain, analisis data juga berperan penting dalam berbagai bidang, termasuk kesehatan, pendidikan, dan pemerintahan. Misalnya, dalam bidang kesehatan, analisis data dapat digunakan untuk memprediksi wabah penyakit dan mengelola sumber daya medis. Dalam pendidikan, data analitik dapat membantu dalam memahami pola belajar siswa dan meningkatkan efektivitas pengajaran (Sun & Liu, 2024). Dengan demikian, kemampuan untuk menganalisis data menjadi keterampilan yang sangat berharga di berbagai sektor. Artikel ini bertujuan untuk memperkenalkan pembaca pada konsep dasar analisis data menggunakan Python, dengan fokus pada dua pustaka utama: Pandas dan NumPy. Melalui artikel ini, diharapkan pembaca dapat memahami pentingnya analisis data, mengenal Python sebagai alat untuk analisis data, serta mempelajari cara menggunakan Pandas dan NumPy untuk melakukan analisis data yang efektif. Selain itu, artikel ini juga akan memberikan contoh praktis dan studi kasus yang relevan untuk memperkuat pemahaman pembaca tentang materi yang dibahas.

2. Research Methods

Penelitian ini menggunakan metode deskriptif dan analitis dalam menganalisis penggunaan Python dan pustaka Pandas serta NumPy dalam analisis data. Data diperoleh melalui studi literatur yang melibatkan buku, artikel, dan sumber daya online yang relevan. Dalam konteks ini, analisis dilakukan untuk mengevaluasi efektivitas kedua pustaka dalam menyelesaikan berbagai tugas analisis data. Selain itu, penelitian ini juga melibatkan studi kasus yang menunjukkan penerapan praktis dari Pandas dan NumPy dalam analisis dataset nyata.

2.1 Apa itu Python?

Python adalah bahasa pemrograman tingkat tinggi yang sangat populer, dikenal karena sintaksnya yang sederhana dan mudah dipahami. Dikenalkan pertama kali oleh Guido van Rossum pada tahun 1991, Python telah berkembang menjadi salah satu bahasa pemrograman yang paling banyak digunakan di dunia, terutama dalam bidang data science, pengembangan web, dan otomatisasi (Klemp, 2024). Salah satu alasan

utama popularitas Python adalah komunitas yang besar dan aktif, yang menyediakan berbagai pustaka dan alat untuk berbagai keperluan.

Python mendukung berbagai paradigma pemrograman, termasuk pemrograman prosedural, berorientasi objek, dan fungsional. Hal ini memungkinkan pengembang untuk memilih pendekatan yang paling sesuai untuk proyek mereka. Selain itu, Python juga memiliki ekosistem pustaka yang kaya, seperti NumPy untuk operasi numerik, Pandas untuk manipulasi data, dan Matplotlib untuk visualisasi data. Dengan demikian, Python menjadi pilihan yang ideal bagi para analis data dan ilmuwan data.

2.2 Mengapa Python Populer di Bidang Data Science?

Ada beberapa faktor yang membuat Python sangat populer di bidang data science. Pertama, kemudahan penggunaan Python memungkinkan pemula untuk dengan cepat belajar dan mulai bekerja dengan data. Sintaks yang bersih dan jelas membuat proses pengkodean menjadi lebih intuitif. Selain itu, Python memiliki berbagai pustaka yang dirancang khusus untuk analisis data, seperti Pandas dan NumPy, yang memungkinkan pengguna untuk melakukan analisis kompleks dengan sedikit kode (Nur & Cahyani, 2024).

Kedua, Python memiliki dukungan kuat untuk integrasi dengan alat dan platform lain, seperti Jupyter Notebook, yang memungkinkan pengguna untuk menulis dan menjalankan kode dalam lingkungan interaktif. Jupyter Notebook menjadi sangat populer di kalangan peneliti dan analis data karena kemampuannya untuk menggabungkan kode, visualisasi, dan dokumentasi dalam satu tempat. Ketiga, Python memiliki komunitas yang besar dan aktif, yang berarti pengguna dapat dengan mudah menemukan dukungan dan sumber daya untuk membantu mereka mengatasi tantangan yang mereka hadapi. Untuk mulai menggunakan Python dalam analisis data, langkah pertama adalah menginstal Python di komputer Anda. Python dapat diunduh dari situs resmi Python (python.org), di mana pengguna dapat memilih versi terbaru yang sesuai dengan sistem operasi mereka. Setelah menginstal Python, pengguna disarankan untuk menginstal Anaconda, sebuah distribusi Python yang mencakup banyak pustaka penting untuk analisis data, termasuk Pandas dan NumPy, serta Jupyter Notebook (Surbakti et al., 2024).

Setelah menginstal Anaconda, pengguna dapat membuka Anaconda Navigator, yang menyediakan antarmuka grafis untuk mengelola lingkungan dan paket

Python. Dari sini, pengguna dapat membuat lingkungan baru dan menginstal pustaka yang diperlukan untuk proyek analisis data mereka. Selain itu, Anaconda juga menyediakan Jupyter Notebook, yang memungkinkan pengguna untuk menulis dan menjalankan kode Python dalam format yang interaktif dan mudah dibaca.

Dengan lingkungan kerja yang sudah siap, pengguna dapat mulai menjelajahi dan menganalisis data menggunakan Python. Penting untuk memahami cara menggunakan pustaka seperti Pandas dan NumPy, karena kedua pustaka ini akan menjadi alat utama dalam proses analisis data.

2.3. Pengenalan Pustaka NumPy

NumPy, singkatan dari Numerical Python, adalah pustaka fundamental untuk komputasi numerik di Python. Dikenalkan pada tahun 2005, NumPy menyediakan dukungan untuk array multidimensi dan berbagai fungsi matematika dan statistik yang efisien (Klemp, 2024). Pustaka ini sangat penting bagi para ilmuwan data dan analis karena memungkinkan mereka untuk melakukan operasi matematis yang kompleks dengan cepat dan efisien. NumPy juga menjadi dasar bagi banyak pustaka lain dalam ekosistem Python, termasuk Pandas dan SciPy.

Salah satu fitur utama NumPy adalah kemampuannya untuk menangani array multidimensi. Berbeda dengan daftar Python biasa, yang hanya dapat menyimpan data satu dimensi, array NumPy dapat menyimpan data dalam bentuk dua dimensi (matriks) atau lebih. Ini sangat berguna dalam analisis data, di mana sering kali kita perlu bekerja dengan data terstruktur dalam bentuk tabel atau matriks.

1. Array Multidimensi

Array multidimensi NumPy, yang dikenal sebagai ndarray, adalah salah satu fitur paling kuat dari pustaka ini. Ndarray memungkinkan pengguna untuk menyimpan dan memanipulasi data dalam bentuk yang lebih kompleks dibandingkan dengan daftar Python biasa. Misalnya, pengguna dapat membuat array dua dimensi untuk mewakili tabel data, di mana setiap baris mewakili satu entitas dan setiap kolom mewakili atribut dari entitas tersebut. Hal ini memudahkan analisis dan manipulasi data (Amelia et al., 2024).

2. Operasi Matematika dan Statistik

NumPy juga menyediakan berbagai fungsi untuk melakukan operasi matematis dan statistik pada array. Pengguna dapat melakukan operasi dasar seperti penjumlahan, pengurangan, perkalian, dan pembagian, serta operasi yang lebih kompleks seperti penghitungan rata-rata, median, dan deviasi standar. Fungsi-fungsi ini

dioptimalkan untuk kecepatan dan efisiensi, sehingga memungkinkan pengguna untuk melakukan analisis data dalam waktu yang lebih singkat (Azzahrah & Sari, 2024).

2.4 Contoh Praktis Penggunaan NumPy

1. Membuat Array dan Operasi Dasar

Salah satu cara untuk memulai dengan NumPy adalah dengan membuat array. Pengguna dapat menggunakan fungsi `numpy.array()` untuk membuat array dari daftar Python. Misalnya, untuk membuat array satu dimensi, pengguna dapat menulis:

```
import numpy as np
array_1d = np.array([1, 2, 3, 4, 5])
```

Setelah array dibuat, pengguna dapat melakukan operasi dasar, seperti penjumlahan dan perkalian. Misalnya, untuk menambahkan 5 ke setiap elemen dalam array, pengguna dapat menulis:

```
array_1d_plus_5 = array_1d + 5
```

2. Penggunaan Fungsi-fungsi Statistik

NumPy juga menyediakan fungsi statistik yang berguna untuk analisis data. Misalnya, pengguna dapat menghitung rata-rata dan deviasi standar dari array dengan menggunakan fungsi `numpy.mean()` dan `numpy.std()`. Contoh kode berikut menunjukkan cara menghitung rata-rata dari array:

```
mean_value = np.mean(array_1d)
```

Dengan demikian, NumPy memberikan alat yang kuat untuk melakukan analisis numerik yang kompleks dengan cara yang sederhana dan efisien.

2.5 Pengenalan Pustaka Pandas

Pandas adalah pustaka Python yang dirancang untuk manipulasi dan analisis data. Diperkenalkan pada tahun 2008 oleh Wes McKinney, Pandas menawarkan struktur data yang fleksibel dan alat analisis yang kuat, memungkinkan pengguna untuk bekerja dengan data terstruktur dengan mudah (Lukas Umbu Zogara et al., 2024). Pandas sangat cocok untuk analisis data dari berbagai sumber, termasuk file CSV, Excel, dan database SQL.

Salah satu fitur utama Pandas adalah kemampuannya untuk menangani data yang hilang dan melakukan pemfilteran data dengan cepat. Dengan menggunakan Pandas, pengguna dapat dengan mudah membersihkan, memanipulasi, dan menganalisis data tanpa perlu menulis banyak kode. Karena kemudahan penggunaannya, Pandas telah menjadi salah satu pustaka paling populer di kalangan analis data dan ilmuwan data.

1. Series

Struktur data utama pertama di Pandas adalah Series, yang merupakan array satu dimensi yang dapat menyimpan berbagai tipe data, termasuk angka, string, dan objek Python lainnya. Setiap elemen dalam Series memiliki label yang disebut indeks, yang memungkinkan pengguna untuk mengakses data dengan lebih mudah. Contoh pembuatan Series dapat dilakukan dengan menggunakan fungsi `pandas.Series()`:

```
import pandas as pd
data_series = pd.Series([10, 20, 30, 40])
```

2. DataFrame

DataFrame adalah struktur data utama kedua di Pandas, yang merupakan tabel dua dimensi dengan baris dan kolom. DataFrame sangat cocok untuk menyimpan data terstruktur, seperti data dari file CSV atau Excel. Pengguna dapat membuat DataFrame dari berbagai sumber, termasuk dictionary, list, atau array. Contoh pembuatan DataFrame dari dictionary adalah sebagai berikut:

```
data_dict = {'Nama': ['Alice', 'Bob', 'Charlie'], 'Umur': [25, 30, 35]}
data_frame = pd.DataFrame(data_dict).
```

Salah satu fitur utama Pandas adalah kemampuannya untuk melakukan manipulasi data dengan mudah. Pengguna dapat melakukan operasi seperti penambahan kolom, penghapusan kolom, dan penggabungan DataFrame dengan menggunakan metode yang sederhana. Misalnya, untuk menambahkan kolom baru ke DataFrame, pengguna dapat menulis:

```
data_frame['Kota'] = ['Jakarta', 'Bandung', 'Surabaya'].
```

2. Penggabungan dan Pengelompokan Data

Pandas juga menyediakan fungsi untuk menggabungkan dan mengelompokkan data. Pengguna dapat

menggabungkan DataFrame menggunakan metode `merge()` dan mengelompokkan data menggunakan metode `groupby()`. Ini memungkinkan pengguna untuk melakukan analisis yang lebih kompleks, seperti menghitung agregat berdasarkan kategori tertentu. Contoh penggunaan `groupby()` adalah sebagai berikut:

```
grouped_data = data_frame.groupby('Kota').mean()
```

1. Membaca dan Menulis Data

Pandas memudahkan pengguna untuk membaca dan menulis data dari berbagai format. Misalnya, untuk membaca file CSV, pengguna dapat menggunakan fungsi `pd.read_csv()`:

```
data_from_csv = pd.read_csv('data.csv')
```

Sebaliknya, untuk menulis DataFrame ke file CSV, pengguna dapat menggunakan metode `to_csv()`:

```
data_frame.to_csv('output.csv', index=False)
```

2. Pemfilteran dan Pengolahan Data

Pandas memungkinkan pengguna untuk melakukan pemfilteran data dengan mudah. Misalnya, untuk memfilter DataFrame berdasarkan kondisi tertentu, pengguna dapat menggunakan sintaks berikut:

```
filtered_data = data_frame[data_frame['Umur'] > 30]
```

Dengan demikian, Pandas memberikan alat yang kuat untuk melakukan analisis data yang kompleks dengan cara yang sederhana dan efisien.

3. Results and Discussions

Integrasi antara NumPy dan Pandas memungkinkan pengguna untuk memanfaatkan kekuatan kedua pustaka dalam analisis data. Pandas dibangun di atas NumPy, sehingga banyak fungsi di Pandas yang menggunakan array NumPy di belakang layar. Ini memungkinkan pengguna untuk melakukan operasi matematis dan statistik yang efisien pada DataFrame dan Series. Misalnya, pengguna dapat melakukan operasi aritmatika antara DataFrame dan array NumPy secara langsung.

Analisis Data Dasar Menggunakan Pandas dan NumPy

1. Deskriptif Statistik

Salah satu langkah awal dalam analisis data adalah melakukan analisis deskriptif untuk memahami

karakteristik data. Dengan menggunakan Pandas, pengguna dapat dengan mudah menghitung statistik deskriptif seperti rata-rata, median, dan deviasi standar. Fungsi `describe()` di Pandas memberikan ringkasan statistik dari DataFrame, termasuk nilai minimum, maksimum, dan kuartil. Misalnya:

```
data_frame.describe()
```

2. Visualisasi Data Dasar

Untuk mendapatkan wawasan yang lebih baik dari data, visualisasi adalah alat yang sangat berguna. Meskipun Pandas sendiri tidak memiliki kemampuan visualisasi yang kuat, ia dapat dengan mudah diintegrasikan dengan pustaka visualisasi seperti Matplotlib. Pengguna dapat membuat grafik sederhana seperti histogram, diagram batang, dan scatter plot untuk menggambarkan data mereka dengan lebih jelas. Contoh penggunaan Matplotlib untuk membuat histogram adalah sebagai berikut:

```
import matplotlib.pyplot as plt
data_frame["Umur"].hist()
plt.show()
```

Sebagai bagian dari studi kasus, pengguna dapat mulai dengan mengimpor dataset sederhana dari file CSV menggunakan Pandas. Misalnya, pengguna dapat membaca dataset penjualan dari file CSV dan menyimpannya dalam DataFrame untuk dianalisis lebih lanjut. Contoh kode untuk membaca dataset penjualan:

```
sales_data = pd.read_csv('sales_data.csv')
```

3.2. Mengolah dan Menganalisis Data

Setelah dataset diimpor, pengguna dapat melakukan berbagai operasi analisis, seperti membersihkan data, memfilter data berdasarkan kondisi tertentu, dan menghitung statistik deskriptif. Dengan menggunakan kombinasi Pandas dan NumPy, pengguna dapat melakukan analisis yang mendalam dan mendapatkan wawasan berharga dari data yang mereka miliki. Misalnya, untuk menghitung total penjualan:

```
total_sales = sales_data["Penjualan"].sum()
```

4. Conclusion

A. Ringkasan Poin-Poin Utama

Dalam makalah ini, telah dibahas pentingnya analisis data dalam era digital dan bagaimana Python, khususnya pustaka Pandas dan NumPy, dapat digunakan untuk melakukan analisis data yang efektif. Pandas dan NumPy menyediakan alat yang kuat untuk manipulasi data, operasi matematis, dan analisis statistik, yang memungkinkan pengguna untuk mendapatkan wawasan berharga dari data yang mereka miliki.

B. Pentingnya Penguasaan Pandas dan NumPy bagi Pemula di Data Science

Penguasaan Pandas dan NumPy adalah keterampilan yang sangat berharga bagi siapa saja yang ingin berkarir di bidang data science. Dengan memahami cara menggunakan kedua pustaka ini, pemula dapat membangun fondasi yang kuat untuk analisis data yang lebih kompleks dan mendalam. Hal ini juga akan membuka peluang bagi mereka untuk terlibat dalam proyek-proyek data science yang lebih besar dan lebih menantang.

C. Rekomendasi untuk Pembelajaran Selanjutnya

Bagi pembaca yang tertarik untuk memperdalam pengetahuan mereka tentang analisis data menggunakan Python, disarankan untuk menjelajahi sumber daya online, tutorial, dan buku yang tersedia. Selain itu, praktik langsung dengan proyek nyata akan sangat membantu dalam memperkuat pemahaman dan keterampilan dalam menggunakan Pandas dan NumPy. Mengikuti kursus online atau bergabung dengan komunitas data science juga dapat memberikan wawasan tambahan dan dukungan yang berharga.

Reference

- Porter, L., & Zingaro, D. (2024). *Learn AI-assisted Python Programming*. Simon and Schuster.
- Simon and Schuster. Sun, L., & Liu, J. (2024). Effects of gamified python programming on primary school students' computational thinking skills: A differential analysis of gender. *Journal of Educational Computing Research*, 62(3), 846-874.
- Lukas Uumbu Zogara, Giandari Maulani dkk (2024). *Kecerdasan buatan masa depan manusia dan teknologi*. Igakerta Publisher.
- Nur, A., & Cahyani, P. R. (2024). EVALUASI PENGGUNA JUPYTER NOTEBOOK PADA PYTHON DALAM PEMBELAJARAN DATA SCIENCE (STUDI KASUS: KAPAL TITANIC). *Kohesi: Jurnal Sains dan Teknologi*, 4(10), 21-30.
- Surbakti, N. M., Angelyca, A., Talia, A., Perangin-Angin, C. B., Nainggolan, D. O., Friskauly, N. D., & Tumorang, S. R. B. (2024). Penggunaan Bahasa Pemrograman Python dalam Pembelajaran Kalkulus Fungsi Dua Variabel. *Algoritma: Jurnal*

Matematika, Ilmu pengetahuan Alam, Kebumihan dan Angkasa, 2(3), 98-107.

Klemp, M. (2024). Python. In *Computer Science in Sport: Modeling, Simulation, Data Analysis and Visualization of Sports-Related Data* (pp. 125-131). Berlin, Heidelberg: Springer Berlin Heidelberg.

Azzahrah, A., & Sari, I. P. (2024). Perbandingan Sistem Prediksi Menggunakan Metode Monte Carlo dengan Metode K-NN pada Nilai Peserta Didik Uji Kompetensi Kejuruan. *sudo Jurnal Teknik Informatika*, 3(3), 127-135.

Raharja, A. R., Pramudianto, A., & Muchsam, Y. (2024). Penerapan Algoritma Decision Tree dalam Klasifikasi Data "Framingham" Untuk Menunjukkan Risiko Seseorang Terkena Penyakit Jantung dalam 10 Tahun Mendatang. *Technologia Journal*, 1(1).

Amelia, S. Y., Aryani, D., Warsito, A. B., & Purnomo, Y. (2024). Comparison of the Number of Iterations of Roots of Non-Linear Equations in the Regula Falsi Method, Secant Method, and Bisection Method using Python. *Formosa Journal of Science and Technology*, 3(5), 971-986.



JISED
Journal of Information System
and Education Development

