

Pandas in Python - Notes

What is Pandas?

Pandas is a Python library used for data analysis, cleaning, and manipulation.

It is commonly used to work with structured data like CSV, Excel, or SQL tables.

It provides two main data structures: Series and DataFrame.

Why Use Pandas?

- Read data from files (CSV, Excel) using `pd.read_csv()`, `pd.read_excel()`
- View and inspect data using `.head()`, `.info()`, `.shape`
- Clean data: remove nulls, fix types, drop duplicates
- Analyze data: filter, group, sort, merge, pivot
- Export data: `.to_csv()`, `.to_excel()`
- Combine with matplotlib/seaborn for visualization

Pandas Data Structures

1. Series - One-dimensional labeled array
2. DataFrame - Two-dimensional labeled table (rows + columns)

Real-World Use Cases

- Data Analysis: Analyze customer or sales data
- Machine Learning: Clean and prepare input datasets
- Finance: Stock price analysis and report generation
- HealthCare: Manage patient records and insights
- E-Commerce: Analyze orders, user behavior

Basic Example

```
import pandas as pd
```

```
df = pd.read_csv('sales.csv')  
print(df.head())  
print(df.groupby('Category')['Sales'].sum())
```

Summary

- Pandas = Python + Data Analysis
- Ideal for real-world structured data tasks
- Fast, powerful, and easy to learn