Introduction to Pandas

Hariharan Vels

What is Pandas?

- Pandas is a powerful Python library for data analysis and manipulation.
- Provides data structures such as **Series** and **DataFrame**.
- Built on top of NumPy and works well with other libraries like Matplotlib and Scikit-learn.

Key Features

- Fast and efficient DataFrame object.
- Tools for loading data from various sources (CSV, Excel, SQL, etc.).
- Data alignment and handling of missing data.
- Label-based slicing, indexing, and subsetting.
- Merge and join capabilities.

Installation

- Install Pandas using pip:

Command

```
pip install pandas
```

- Import Pandas in Python:

Python Code

```
import pandas as pd
```

Data Structures in Pandas

- **Series**: 1D labeled array.
- **DataFrame**: 2D labeled table.
- **Panel** (deprecated): 3D labeled data structure.

Creating a DataFrame

Example

```
 import pandas as pd
data = {'Name':  ['Alice', 'Bob'], 'Age':  [25, 30]}
df = pd.DataFrame(data)
print(df)
```

Basic Operations

- `df.head()` - View first few rows.
- `df.tail()` - View last few rows.
- `df.info()` - Summary of DataFrame.
- `df.describe()` - Statistical summary.
- `df.shape` - Shape of DataFrame.

Data Selection

- Select a column: `df['column_name']`
- Select multiple columns: `df[['col1', 'col2']]`
- Select rows by index: `df.iloc[0]`
- Select rows by label: `df.loc['index_label']`

Data Cleaning

- Handling missing values:
  - `df.dropna()` - Remove missing values.
  - `df.fillna(value)` - Fill missing values.
- Removing duplicates:
  - `df.drop_duplicates()`

Conclusion

- Pandas is essential for data analysis in Python.
- Provides powerful data structures and functions.
- Works well with other libraries for visualization and ML.