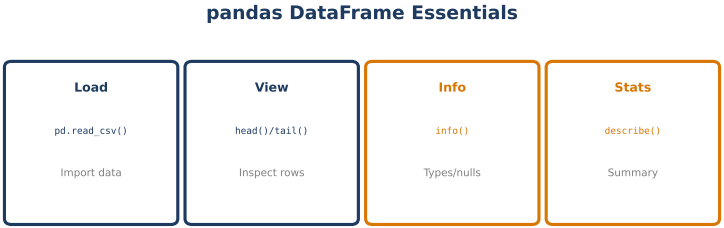
Lesson 05 Summary: DataFrames Introduction

Data Science with Python – Key Concepts

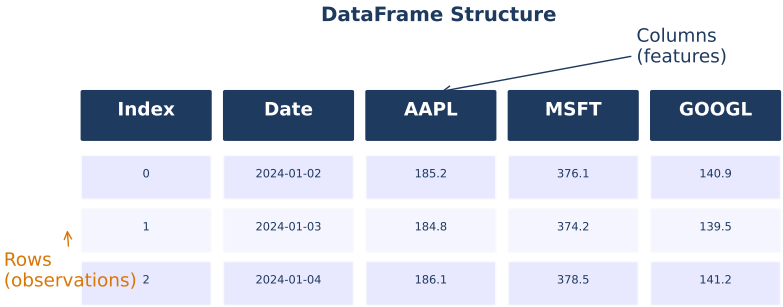
Data Science Program

DataFrames Overview



pandas is THE library for data manipulation in Python

DataFrame Structure



2D labeled data structure with rows and columns

DataFrames are 2D labeled data structures

Series vs DataFrame

Two core pandas data structures:

Series (1D):

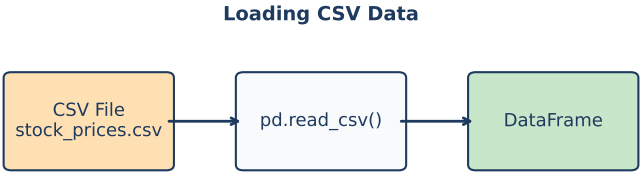
- Single column with index
- Like a labeled array
- Example: `df['Price']` returns a Series

DataFrame (2D):

- Multiple columns sharing an index
- Collection of Series
- Example: `df[['Price', 'Volume']]` returns a DataFrame

Series = 1 column; DataFrame = multiple Series

Loading Data



Common Parameters:

```
filepath: "data/prices.csv"
index_col: "Date"
parse_dates: True
usecols: ["AAPL", "MSFT"]
```

pd.read_csv() handles most CSV formats automatically

Viewing Data

Essential inspection methods:

- **df.head(n):** First n rows (default 5)
- **df.tail(n):** Last n rows (default 5)
- **df.sample(n):** Random n rows
- **df.shape:** (rows, columns) tuple

Always inspect after loading:

```
df = pd.read_csv('data.csv')
print(df.head())  # First 5 rows
print(df.shape)   # (252, 5)
```

Always inspect data after loading!

Data Info

DataFrame Info: df.info()

```
<class pandas.DataFrame>
RangeIndex: 252 entries, 0 to 251
Data columns (5 columns):
  Date    252 non-null datetime64
  AAPL    252 non-null float64
  MSFT    252 non-null float64
  GOOGL   250 non-null float64  (2 missing)
memory usage: 10.0 KB
```

info() reveals data types and missing values

Summary Statistics

df.describe() provides statistical summary:

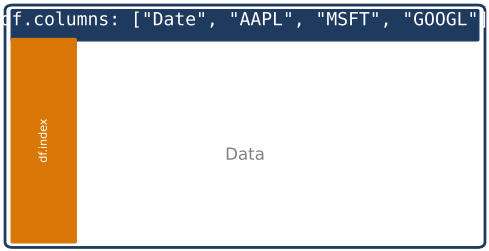
- **count:** Non-null values
- **mean:** Average value
- **std:** Standard deviation
- **min/max:** Range
- **25%/50%/75%:** Quartiles

Finance insight: Compare std (volatility) across assets

---

describe() provides statistical summary of numeric columns

Index and Columns



df.shape: (252, 4) | df.dtypes: column data types

Index labels rows; columns label data fields

Quick Reference

Essential DataFrame Operations:

Operation	Syntax
Import pandas	`import pandas as pd`
Load CSV	`df = pd.read_csv('file.csv')`
View first rows	`df.head(n)`
View last rows	`df.tail(n)`
Data types	`df.info()`
Statistics	`df.describe()`
Shape	`df.shape`
Column names	`df.columns`
Index	`df.index`

Master these for efficient data exploration