Module 2: Data Manipulation

Data Science with Python – BSc Course

Why Data Manipulation?

Real financial data has missing prices on weekends and holidays. Corporate actions split price series overnight. Bloomberg, Reuters, and Yahoo format data differently – merging them requires exact alignment. A single NaN can silently corrupt an entire analysis.

Data quality is not glamorous, but it is where 90% of quantitative work happens. Regulatory compliance depends on complete audit trails. Portfolio analytics require aligned multi-source data.

Clean data is the foundation of reliable financial analysis

Why This Matters

- **Data quality:** 90% of quant time is spent cleaning and preparing data
- **Regulatory compliance:** MiFID II requires complete audit trails with no gaps
- **Portfolio analytics:** Risk metrics require aligned multi-source data
- **Time series analysis:** Financial data is inherently temporal and must be handled correctly

Master data manipulation and you solve real-world finance problems

What You'll Learn

By the end of this module, you will be able to:

- Handle missing data with appropriate imputation strategies
- Perform sorting, ranking, and transformations on DataFrames
- Aggregate data using GroupBy operations (split-apply-combine)
- Merge data from multiple sources with different structures
- Work with NumPy arrays for efficient numerical computing
- Manipulate time series data with proper datetime handling

From messy real-world data to analysis-ready datasets

Lesson Roadmap

Lesson	Topic	Focus
L07	Missing Data and Cleaning	NaN handling, imputation
L08	Basic Operations	Sorting, ranking, apply
L09	GroupBy Operations	Split-apply-combine
L10	Merging and Joining	Concat, merge, join
L11	NumPy Basics	Arrays, vectorization
L12	Time Series Basics	Datetime, resampling

Each lesson: 45 minutes lecture + hands-on exercises

Key Concepts

- **Missing Data Handling** – Detect, drop, fill, or interpolate NaN values
- **DataFrame Operations** – Sort, rank, apply transformations efficiently
- **GroupBy Split-Apply-Combine** – Aggregate data by categories
- **Merging & Joining** – Combine datasets like SQL joins
- **NumPy Arrays** – Fast numerical computation for large datasets
- **Time Series** – Work with dates, times, and resampling

These operations form the backbone of financial data workflows

Finance Application

Scenario: Multi-Source Portfolio Builder
Using the skills from this module, you can:

- Merge stock prices from Bloomberg, Yahoo Finance, and Reuters
- Handle missing dates (weekends, holidays) with forward fill
- Align time series from different data providers
- Calculate portfolio risk metrics across multiple asset classes

This is exactly what risk managers do daily when preparing data for VaR calculations.

Real-world data is messy – this module teaches you how to clean it

Who Uses This?

- **Bloomberg Terminal** – Data cleaning is 60% of the workflow for analysts
- **Central Banks** – ECB merges data from 19 eurozone national central banks
- **Risk Management** – Basel III requires clean, aligned position data
- **Index Providers** – MSCI and FTSE rebalance indices from multiple exchanges

Data manipulation is the unsung hero of finance

What's Next: Module 3 – Statistics & Visualization

Now that you can clean and manipulate data, it's time to analyze it statistically and present your findings visually.
Module 3 covers descriptive statistics, probability distributions, hypothesis testing, and professional data visualization with matplotlib and seaborn.
Prerequisite check: Can you merge two DataFrames, handle NaN values, and resample time series? If yes, you are ready.

Clean data is just the starting point – now we extract insights

Let's Begin!

First up: L07 – Missing Data and Cleaning

Open your laptop and follow along.