IQIC - Quantitative Tools

#10: Working with Data

What Kind of Data Do You Need?

Price, fundamental, and alternative data

- **Price data:** Open, high, low, close, volume (OHLCV): the foundation of most strategies.
- **Fundamental data:** Earnings, revenue, P/E, book value: used for value and quality signals.
- **Alternative data:** Sentiment, satellite, web traffic: increasingly common but harder to clean.
- **Why types matter:** The data you choose determines what signals you can build.

Data Type	Examples	Common Use
Price	OHLCV, adjusted close	Momentum, mean reversion
Fundamental	EPS, P/E, ROE	Value, quality
Alternative	News sentiment, satellite	Event-driven, alpha

Where to Get Data and How to Organize It

Sources and structure

- **Free sources:** Yahoo Finance, Alpha Vantage, Stooq, FRED: good for beginners.
- **Paid sources:** Bloomberg, Reuters, WRDS: used in professional settings.
- **Structure:** Use a DataFrame with dates as the index and tickers as columns.
- **Consistency:** Align time zones, frequencies, and date ranges across all assets.

Date	AAPL	MSFT	NVDA
2024-01-31	190	400	620
2024-02-29	195	410	650
2024-03-31	200	420	680

Garbage In, Garbage Out

Handling missing values, outliers, and adjustments

- **Missing values:** Forward fill, interpolate, or drop depending on context.
- **Outliers:** Check for data errors, unadjusted splits, or bad ticks.
- **Survivorship bias:** Include delisted stocks to avoid only winners.
- **Corporate actions:** Always use adjusted prices for splits and dividends.

Date	Raw Price	Adjusted Price	Note
2024-01-31	100	100	No split
2024-02-29	50	100	2-for-1 split adjusted
2024-03-31	52	104	Normal

From Prices to Returns

Why we transform data

- **Prices are non-stationary:** they trend and have different scales.
- **Returns are stationary:** easier to compare and model.
- **Simple return:** (Price today / Price yesterday) - 1.
- **Log return:** ln(Price today / Price yesterday): additive over time.

Date	Price	Simple Return	Log Return
Day 1	100		
Day 2	105	5.0%	4.88%
Day 3	103	-1.9%	-1.92%

The Data Pipeline

From raw data to backtest-ready

- **Acquire:** Get data from a reliable source.
- **Clean:** Handle missing values, outliers, and adjustments.
- **Transform:** Compute returns, signals, and other features.
- **Store:** Save as CSV or Parquet for fast access.
- **Iterate:** Repeat as you refine your strategy.

Additional Resources

Where to go next

- **"Python for Data Analysis" by Wes McKinney**

  Free online book: the definitive guide to Pandas and data cleaning.
- **"Financial Data Cleaning & Preprocessing Guide" by PyQuant**

  Article series on handling missing data, outliers, and corporate actions.
- **"Yahoo Finance API Tutorial" by Towards Data Science**

  Step-by-step guide to pulling free financial data into Python.