Reading — Stage 7: Outliers + Risk Assumptions

Why Outliers Matter (Especially in Finance)

Outliers can be **errors**, **rare events**, or **early warnings**. In finance, fat-tailed returns, liquidity dry-ups, and flash crashes mean that extreme values may be central to risk understanding rather than noise to delete. Treating them requires **assumptions** about what is “normal.”

Defining Outliers

- **IQR Rule:** Points outside Q1 − 1.5·IQR or Q3 + 1.5·IQR. Robust to skew, quick to compute, widely used with boxplots.
- **Z-score:** Points with |z| > threshold (often 3). Interpretable if data are roughly normal; can over-flag for heavy tails.
- **Winsorizing:** Replace extremes with chosen quantile values. Preserves row count; reduces leverage; but alters data.

Key Point: There is **no single correct** definition. Choose a method, state your **assumptions**, and test **sensitivity**.

Sensitivity Analysis

Ask: *How much do my conclusions change if I change how I treat outliers?*

- Compare summary stats with vs. without outliers.
- Refit a simple model after removal and after winsorizing.
- Report the differences (coefficients, R^2, MAE) in a small table.

Risks & Misconceptions

- **Not all outliers are errors.** A crash day is rare but real.
- **Removing is not always better.** You might delete exactly the signal you seek.
- **Thresholds are choices.** Record them and justify with domain context.

Minimal Engineering Patterns

- Write **reusable functions** (detection, handling).
- Use **boolean masks** for clarity and testability.
- Keep a **sensitivity table** and save it (CSV).
- Document **assumptions** in code and README for reproducibility.

Worked Mini-Example (Conceptual)

1. Fit regression on all data → slope = 1.8.
2. Remove IQR outliers → slope increases to 2.0, MAE drops.
3. Winsorize → slope ~1.9, MAE moderate. Interpretation: Outliers were pulling the fit; removing reduces leverage, winsorizing softens extremes.

What to Write in Your Project Repo

- A `src/` module with your outlier functions and docstrings.
- A notebook that compares **all vs. filtered vs. winsorized**.
- A README section: the thresholds you chose, *why* you chose them, and the observed impact.

Mantra: *Assumptions → Method → Sensitivity → Documentation.*