Homework Sheet — Stage 7: Outliers + Risk Assumptions

Assignment

Implement reusable functions to detect and handle outliers, run a simple sensitivity analysis, and reflect on assumptions.

Chain: In the lecture, we learned how to detect, remove, or adjust outliers (IQR, Z-score, winsorizing) and to test their impact on simple models. Now, you will adapt these methods to a provided dataset to assess how your choices affect results and to document risks and assumptions.

What You'll Do

1. Open `notebooks/stage07_outliers-risk-assumptions_homework-starter.ipynb`.
2. Load `data/raw/outliers_homework.csv` (provided) or generate the synthetic fallback in the notebook.
3. Implement and document:
    - `detect_outliers_iqr(series)`
    - `detect_outliers_zscore(series, threshold=3.0)`
    - *(Stretch)* `winsorize_series(series, lower=0.05, upper=0.95)`
4. Apply to at least one numeric column; create a boolean outlier flag.
5. Sensitivity analysis (pick one model or summary):
    - Compare summary stats (mean/median/std) **with vs. without outliers**.
    - Fit a simple linear regression and compare coefficients/R^2/MAE **with vs. without outliers**; optionally include winsorized.
6. Reflection (≤1 page in Markdown cell):
    - Which method(s) and thresholds you chose and **why**.
    - Assumptions behind your choices.
    - Observed impacts on results.
    - Risks if assumptions are wrong.

Deliverables (due next class)

- A single Jupyter notebook with:
    - Implemented functions and docstrings
    - Sensitivity comparison (table and 1–2 plots)
    - Reflection write-up

Grading Rubric (100 pts)

- (30) Correct, reusable functions with docstrings and parameterization
- (30) Sensitivity analysis comparing at least two treatments
- (20) Reflection on assumptions/risks tied to results
- (20) Code clarity, organization, and reproducibility (clean cells, seeded randomness)

Stretch Goals

- Implement winsorizing and re-run comparisons
- Visualize residuals and note any change in outlier influence
- Package functions into `src/outliers.py` and import them in your notebook

Example Expectations

- Clear boxplot before/after
- Table of metrics for three variants (all / filtered / winsorized)
- Reflection that ties method choice to the data's distribution and context