Exercise 3A: Cross-Sectional Standardization

BUSI 722: Data-Driven Finance II

Load `merged.parquet` (produced in Exercise 2C). All features and targets should be standardized cross-sectionally each month.

Submission

Submit a **Jupyter notebook** (`.ipynb`) containing all code, output, and charts. Use markdown cells for any written discussion.

Before building ML models, we need to standardize features cross-sectionally (within each month). For the following, use each month beginning Jan 2023.

1. Using `momentum`, `roe`, and `gp_to_assets`, demonstrate two standardization approaches:
   - **Z-scores:** For each month, subtract the cross-sectional mean and divide by the standard deviation. Report the mean and standard deviation of each variable for 3 sample months to verify they are approximately 0 and 1.
   - **Ranks:** For each month, rank stocks and scale ranks to [0, 1]. Verify that the mean rank is approximately 0.5 for each month.
2. Plot the distribution of raw `roe` vs. z-scored `roe` vs. ranked `roe` for a single month. In a markdown cell, discuss which transformation best handles outliers.
3. Define the full feature set: `momentum`, `lag_month`, `pb`, `roe`, `grossmargin`, `assetturnover`, `leverage`, `asset_growth`, `gp_to_assets`. Each month, convert all features to **percentile ranks** scaled to [0, 1].
4. Each month, convert the target `return` to **percentile ranks** scaled to [0, 1].