Modeling Artefacts

1. Model Choice & Rationale

We employed a **combination of clustering and dimensionality reduction techniques** to uncover patterns in financial data:

- **K-Means Clustering**: Chosen for its scalability and interpretability, allowing us to group companies into peer sets based on financial indicators.

- **Agglomerative Clustering**: Used to explore hierarchical structures in the data, providing insight into nested similarities among companies.

- **DBSCAN**: Applied to capture non-linear clusters and identify outliers, which are especially useful in financial anomaly detection.

- **Spectral Clustering & Birch**: Tested as alternative clustering methods to evaluate performance on high-dimensional feature spaces.

- **Gaussian Mixture Models (GMM)**: Selected to capture probabilistic soft clusters, useful when company behavior overlaps between groups.

- **Dimensionality Reduction (PCA, t-SNE)**: PCA was used to reduce noise and retain variance in fewer dimensions, while t-SNE helped in visualizing high-dimensional structures for diagnostic purposes.

The rationale behind this multi-model approach was to compare clustering techniques under different assumptions (linear separability, density-based grouping, hierarchical structure) and choose the best representation of financial similarity.

---

2. Baselines

As a baseline, we considered:

- **Random Clustering Assignment**: Used to establish a lower-bound benchmark for silhouette score and clustering stability.

- **K-Means (k=2)**: A simple baseline to evaluate the separation of companies into "profitable" vs. "non-profitable" clusters.

This ensured that more sophisticated models were meaningfully outperforming trivial segmentation.

---

3. Configurations & Seeds

- **Random Seeds**: All clustering algorithms were initialized with fixed random seeds (e.g., `random_state=42`) to ensure reproducibility.
- **Data Preprocessing**:
  - Missing values were imputed using **SimpleImputer (median strategy)**.
  - Features were scaled using **RobustScaler** to mitigate the impact of financial outliers.
- **Clustering Configurations**:
  - K-Means: `n_clusters` tuned between **3–10**.
  - Agglomerative: linkage methods (`ward`, `complete`, `average`) compared.
  - DBSCAN: tuned `eps` and `min_samples` using grid search.
  - GMM: number of components tested for **3–10**, covariance types (`full`, `tied`, `diag`, `spherical`).

---

4. Metrics & Diagnostics

To evaluate clustering quality, we employed:

- **Silhouette Score**: Measures cohesion vs. separation of clusters.

- **Davies–Bouldin Index (DBI)**: Captures intra-cluster similarity and inter-cluster difference.
- **Calinski–Harabasz Index (CHI)**: Evaluates variance ratio between clusters.
- **Visualization Diagnostics**:
    - PCA-reduced and t-SNE-reduced 2D plots for intuitive validation of clusters.
    - Heatmaps of feature averages per cluster to interpret financial patterns.

---

5. Ablations & Error Analysis

We conducted ablation studies to analyze the contribution of preprocessing and feature sets:

- **Without Scaling**: Clustering performance deteriorated significantly, especially for DBSCAN and K-Means, confirming the necessity of normalization.
- **Without PCA**: Higher-dimensional clustering showed lower silhouette scores, highlighting PCA's role in denoising.
- **Cluster Stability Across Seeds**: Evaluated by running models with multiple seeds and comparing Adjusted Rand Index (ARI). Results confirmed that DBSCAN was highly sensitive to parameters, while K-Means and GMM were relatively stable.
- **Outlier Analysis**: DBSCAN effectively identified outlier companies that deviated strongly in financial ratios. Manual inspection showed that these often corresponded to companies with unusually high debt or extreme revenue fluctuations.