Market Regime Detection unsupervised learning

1. Custom Features and Clustering Strategy

We extracted and engineered a range of features from order book and trade data to characterize market conditions over time. Key features include:

- **Volatility**: Rolling standard deviation of mid-price returns.
- **Spread**: Difference between best ask and bid prices.
- **Order Book Imbalance**: Relative quantity on the bid vs ask side.
- **Trade Intensity**: Frequency of trades within a time window.
- **Price Trend**: Slope of linear regression fitted on mid-price over a rolling window.

These features were combined and standardized using `StandardScaler`. To reduce dimensionality while preserving interpretability, we applied **PCA** and retained components explaining **95% variance**.

We evaluated three clustering models:

- **KMeans**: Silhouette Score = **0.40**
- **HDBSCAN**: Silhouette Score = **0.59** *(best)*
- **Gaussian Mixture Model (GMM)**: Silhouette Score = **0.25**

HDBSCAN was chosen due to its superior performance and ability to identify noise and non-spherical clusters.

2. Clustering Results

We tuned HDBSCAN with a grid search over `min_cluster_size` and `min_samples`, and found the best configuration to be:

- `min_cluster_size = 30`
- `min_samples = 20`

This setup produced **3 major regimes**, with the fourth class (-1) assigned to noise. PCA and UMAP visualizations showed clear separation between dense clusters.

3. Regime Insights

We labeled and analyzed the regimes:

- **Regime 0**: High spread, low volume → "Illiquid & Volatile"
- **Regime 1**: High volatility, intermediate spread → "Volatile & Active"
- **Regime 2**: Low volatility, tight spread, upward trend → "Stable & Trending"

Statistical summary using `groupby('regime').agg(...)` provided insights into:

- Mean volatility, volume, and spread
- Typical price movement directions

A **regime transition matrix** revealed probabilities of switching from one regime to another, helping identify persistent vs. unstable states.

4. Visualizations

We used the following plots to validate and interpret clusters:

- **PCA and UMAP projections** of clustered data
- **Mid-price over time** with regime overlay
- **Volatility and spread charts**, grouped by regime
- **Regime transition heatmap** to study regime persistence and shifts