Regime Detection via Unsupervised Learning from Order Book and Volume Data

Objective

The goal of this project is to segment market behavior into distinct regimes using unsupervised learning,
leveraging high-frequency order book (depth20) and volume (aggTrade) data. The detected regimes are based on three dimensions:

- Trending vs. Mean-Reverting
- Volatile vs. Stable
- Liquid vs. Illiquid

Feature Engineering

We extracted a range of features from order book and trade data to capture short-term market microstructure:

Order Book Features:

- Bid-Ask Spread: ask_price_1 - bid_price_1
- Microprice: weighted price between top bid and ask
- Order Book Imbalance: (bid_qty_1 - ask_qty_1) / (bid_qty_1 + ask_qty_1)
- Cumulative Depth: cum_bid_qty and cum_ask_qty across top 20 levels
- Sloped Depth: quantifying how quickly depth decays from best price

Volatility & Price Action:

- Rolling Mid-Price Returns: log returns over 1s windows
- Short-term Volatility: standard deviation of returns over 10s and 30s

Trade Volume Features:

- Volume Imbalance: difference between buy and sell trade volume
- Cumulative Volume: over 10s and 30s
- VWAP Shift: delta in VWAP across rolling windows
- Trade Wipe Level: average book levels wiped by aggressive trades

Normalization & Dimensionality Reduction

All features were normalized using Z-score normalization to ensure consistent scale.

PCA was applied to reduce dimensionality while preserving variance, easing clustering and visualization.

Clustering & Regime Detection

We tested the following unsupervised clustering methods:

- K-Means: Elbow method used to choose optimal K
- HDBSCAN: Captures non-spherical clusters and noise
- GMM: Probabilistic clustering, allows soft regimes

Silhouette score, Davies-Bouldin index, and visual inspection were used to assess clustering quality.

Best performance observed with HDBSCAN, identifying ~4-6 distinct market regimes.

Regime Labeling & Insights

Each timestamp was assigned a regime label, and aggregate characteristics were computed for each regime:

Regime ID	Behavior	Summary
0	Mean-Reverting, Stable, Illiquid	Tight range, low volume
1	Trending, Volatile, Liquid	Strong moves, high depth, high vol
2	Trending, Stable, Illiquid	Steady direction but low trades
3	Mean-Reverting, Volatile, Liquid	Whipsaw action with heavy volume

Visualization

- t-SNE / UMAP plots revealed well-separated clusters in 2D space.
- Price vs Regime timeline overlays confirmed correlation between regime shifts and market structure.
- Regime transition matrix computed to estimate the probability of switching from one state to another.

Regime Transition Analysis

Observed transitions showed interesting patterns:

- Volatile regimes were often preceded by illiquid + stable states
- Trending + volatile states frequently shifted into mean-reverting regimes

This insight can aid adaptive trading strategies or volatility forecasting models.

Deliverables

- Jupyter Notebook: Data processing, feature engineering, clustering
- Report: Summary of methodology, clustering outcomes, and insights
- Visuals: t-SNE plots, regime heatmaps, price overlays, transition diagrams