Detecting Market Regimes Using High-Frequency Limit Order Book and Trade Data

I processed the data in 4 pairs of Book Order- Trade Volume data for the four respective files.

1. Feature Engineering

-**Market Microstructure**: Calculated mid-price, spread, and microprice for liquidity, cost insights.
-**Order Book Depth**: Extracts imbalance and depth at multiple levels, capturing market pressure.
-**Volatility**: Computes rolling volatility for price and microprice to track uncertainty.
-**Trade Dynamics**: Includes buy/sell volume, VWAP, & market maker activity for trade flow study
-**Synchronization**: Aligns order book and trade data for consistent feature extraction.

Additional Steps:
Timestamp Cleaning: Converts raw timestamp strings into a consistent datetime format.
Slope Calculation: Uses linear regression to calculate price-quantity slopes for bid and ask sides.
Rolling Window Features: Computes rolling features- total volume, volume imbalance, trade counts.
Normalization: Standardizes features using z-score scaling for model input consistency.

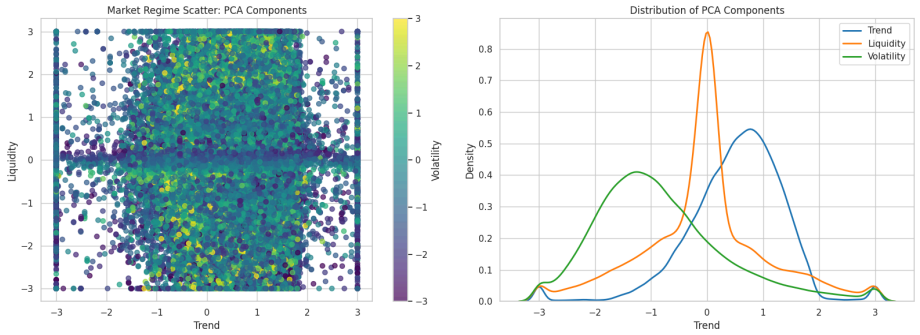
2. Normalization & Clustering

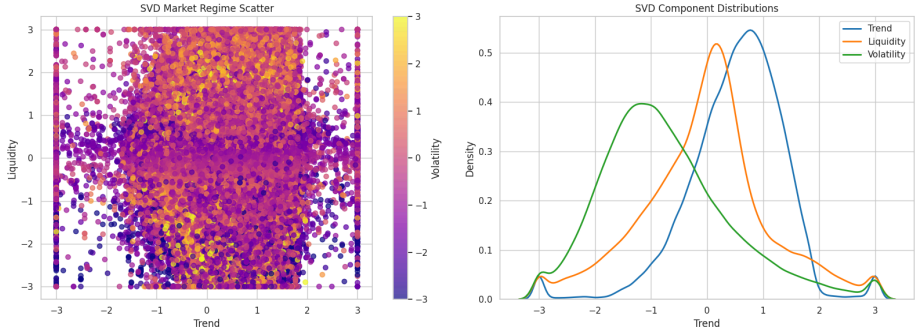
-**PCA**:

Variance: Component 1 captures 99%+ of the data, simplifying to a single dominant feature.
Kurtosis: High in Component 1, indicating rare events.

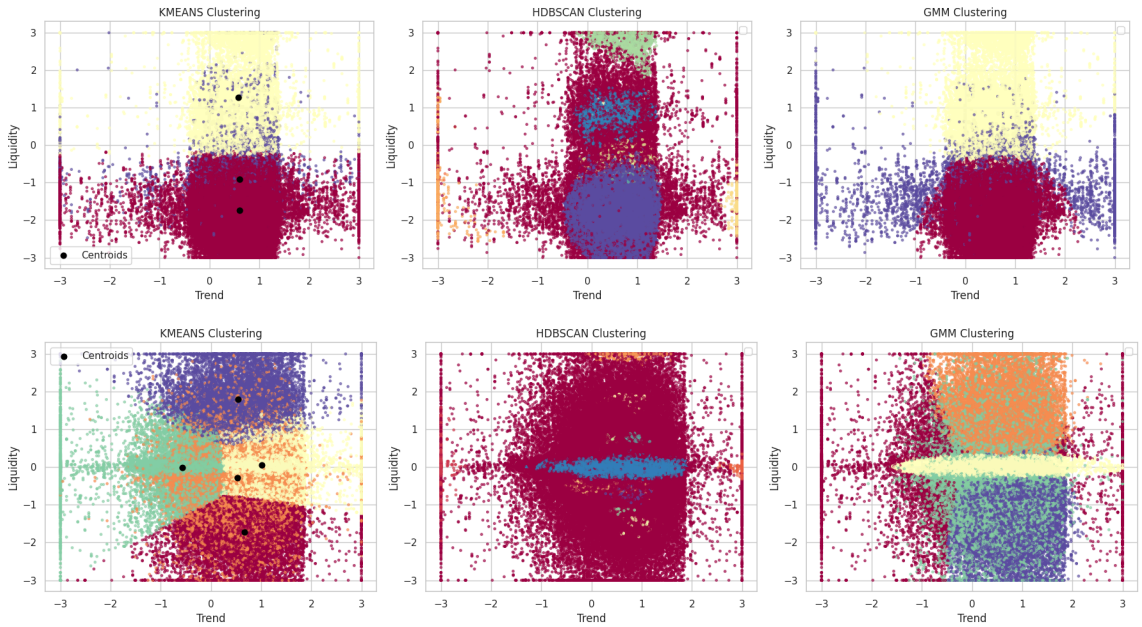
-**SVD**:

Variance: More evenly spread across components, capturing diverse data behaviors.
Kurtosis: High in Component 2, highlighting volatility.





Results from all 4 pairs were almost the same except the first pair. I have included the detailed results with all complete visualizations in the GitHub link.



Clustering Summary:

-KMeans: Silhouette scores (0.278–0.523) suggest moderate clustering, with 3–5 clusters.

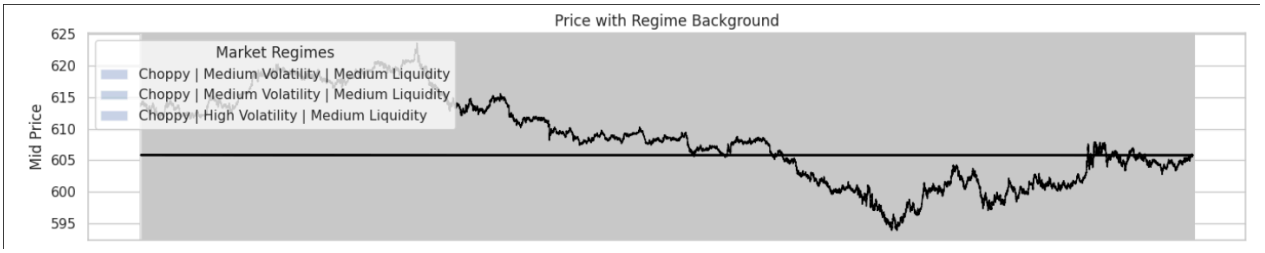
-HDBSCAN: Silhouette scores (–0.069–0.515) indicate good cohesion, with 7–11 clusters.

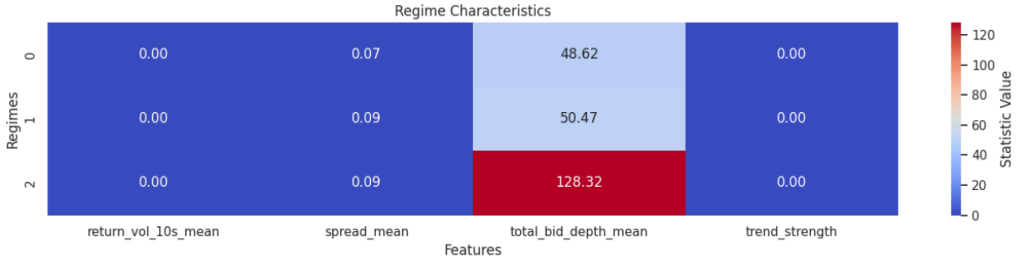
-GMM: Silhouette scores (0.190–0.493) show variable performance, with 3–5 clusters.

Insight: HDBSCAN excels in cohesion and cluster count, while KMeans and GMM perform moderately.

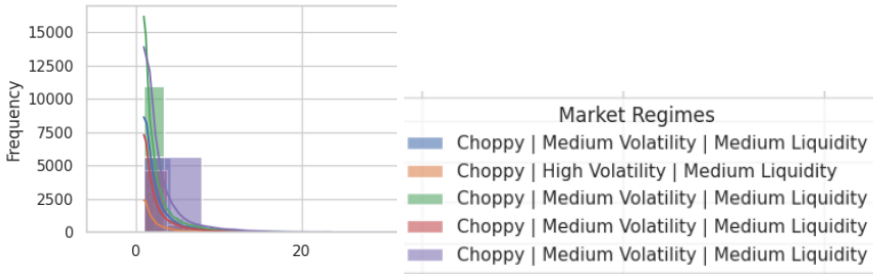
3. Regime Labeling and Analysis & Visualizations

Pair	Volatility	Liquidity	Directionality	Regime Label
1(14th)	Low	High	Mean Reverting	Mean Reverting & Liquid & Stable
2(15th)	Moderate	High	Trending	Trending & Liquid & Volatile
3(16th)	High	Low	Trending	Trending & Illiquid & Volatile
4(17th)	Moderate	Low	Mean Reverting	Mean Reverting & Illiquid & Volatile

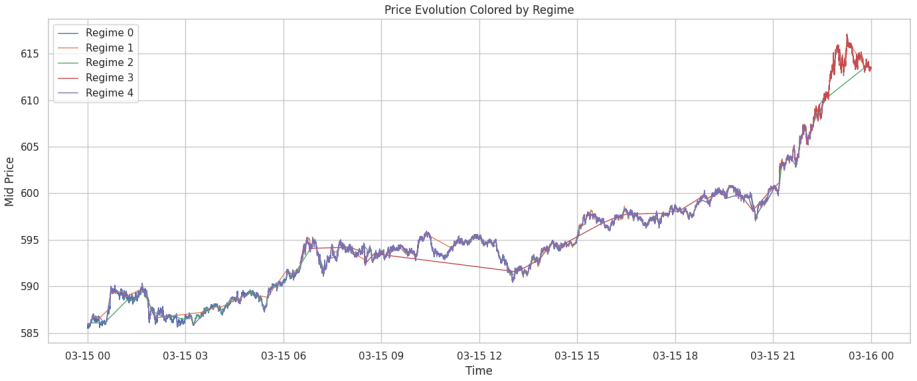


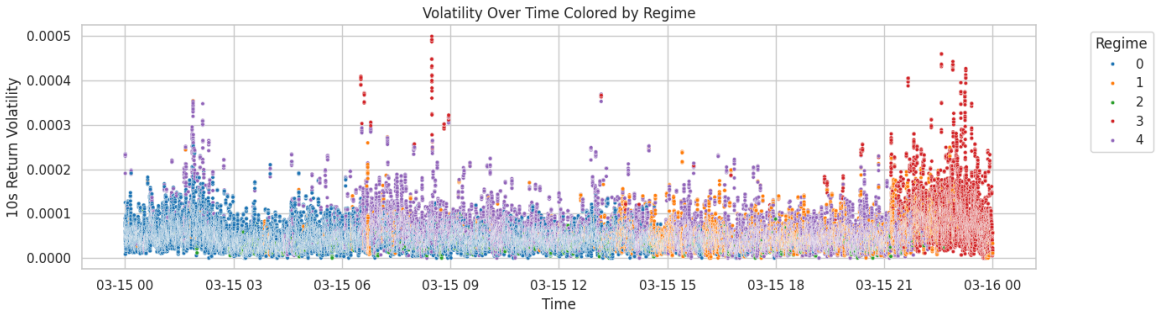


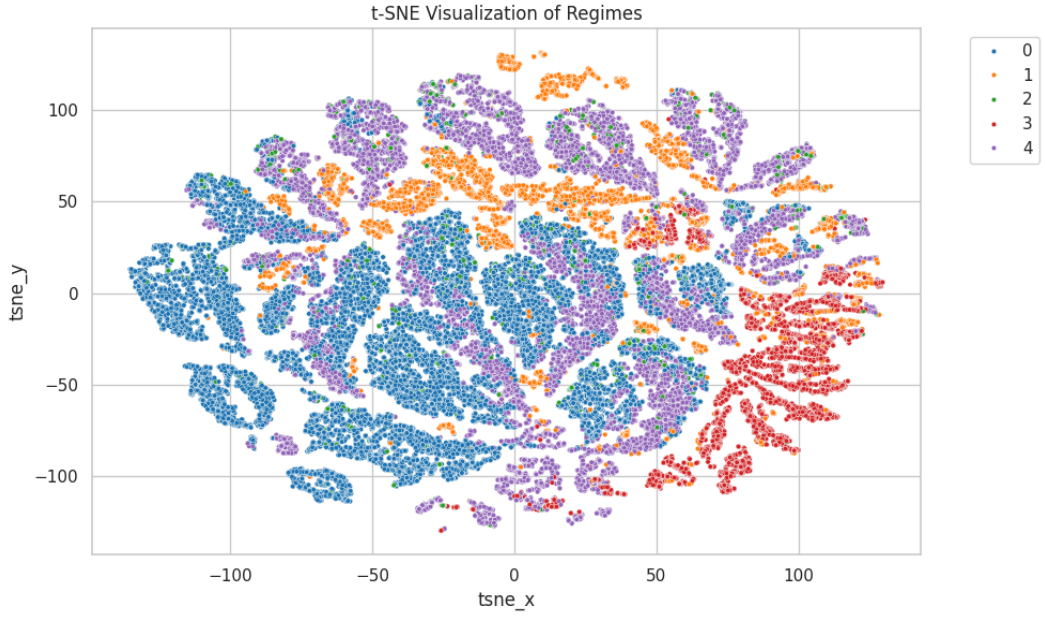
(For pair-3)

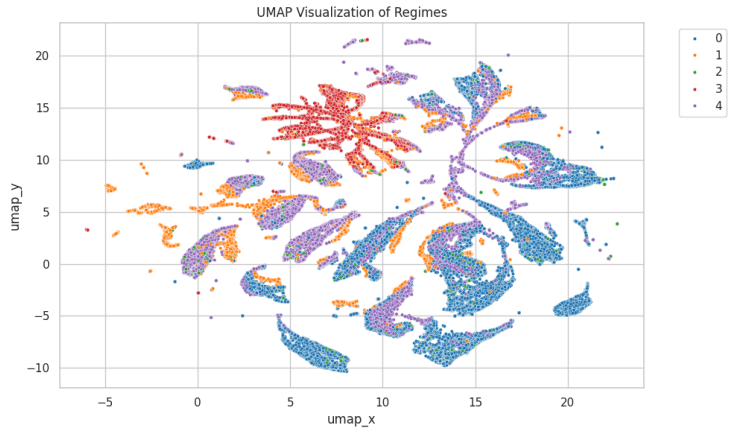


Regime Duration Distribution

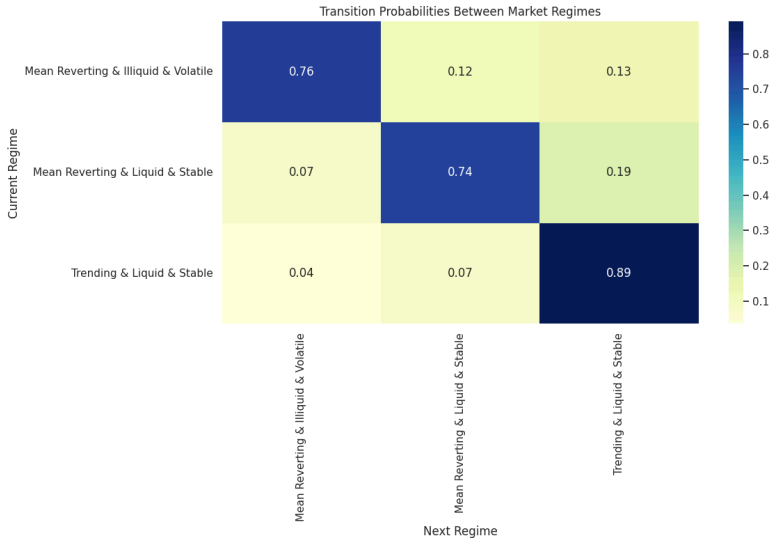








4. Regime Change Insights



Although the results varied a bit among the 4 pairs(provided in GitHub link), for the above result, it could be summarised in the following way:

1.Transition Matrix:

-High Self-Transition Probabilities:

Regime 0 → 0: 76%

Regime 1 → 1: 74%

Regime 2 → 2: 89%

-Interpretation: Regimes exhibit stability, with Regime 2 (Trending & Liquid & Stable) being the most persistent. This persistence is useful for predictive modeling and suggests regimes change infrequently.

2.Feature Distributions by Regime:

Regime 0: High volatility, low liquidity.

Regime 1: Stable, moderate liquidity.

Regime 2: High liquidity, trending price action.

-Interpretation: Feature distributions across regimes verify economic intuition, highlighting distinct characteristics for each market state.