ANURAG SARKAR

sarkaranurag2321@gmail.com

9830155739

**Report: Regime Detection via Unsupervised Learning from Order Book and Volume Data**

**Objective**

The primary objective of this project is to segment the financial market into distinct behavioral regimes based on three critical factors:

1. **Trending vs. Mean-Reverting**

2. **Volatile vs. Stable**

3. **Liquid vs. Illiquid**

This segmentation was achieved using unsupervised learning techniques applied to real-time order book and trade volume data.

**Data Overview**

Two datasets were utilized for this analysis:

1. **Order Book Data (depth20):** Contains top 20 levels of bid/ask prices and quantities.

2. **Trade Volume Data (aggTrade):** Includes aggregated trade data such as price, quantity, and trade direction.

**Key Challenges:**

· High dimensionality of order book data.

· Real-time feature extraction and normalization.

· Selection of clustering algorithms suitable for dynamic and noisy financial data.

**Methodology**

**1. Feature Engineering**

Feature engineering was central to extracting meaningful signals from raw data, focusing on liquidity, volatility, and price action characteristics.

**Liquidity & Depth Features:**

· **Bid/Ask Spread:** Spread = AskPriceL1 - BidPriceL1

o Measures market tightness.

· **Order Book Imbalance:**

o ImbalanceL1 = (BidQtyL1 - AskQtyL1) / (BidQtyL1 + AskQtyL1)

o Indicates buying vs. selling pressure at the top level.

· **Microprice:** Weighted average price based on top-level quantities:

o Microprice = (BidPriceL1 \* AskQtyL1 + AskPriceL1 \* BidQtyL1) / (BidQtyL1 + AskQtyL1)

· **Cumulative Depth:** Summation of bid/ask quantities across all levels:

o CumBidQty and CumAskQty.

**Volatility & Price Action Features:**

· **Rolling Mid-price Return:** MidReturn = log(MidPrice\_t / MidPrice\_t-1)

· **Volatility:** Standard deviation of mid-price returns over rolling windows (10s, 30s).

**Volume Features:**

· **Volume Imbalance:** (BuyVolume - SellVolume) / (BuyVolume + SellVolume)

o Highlights directional bias in trading activity.

· **VWAP Shift:** Change in VWAP over time.

**Derived Features:**

· **Sloped Depth:** Linear regression slope of bid/ask quantities across levels to quantify liquidity decay.

· **Trade Wipe Level:** Measures how deep into the order book trades penetrate.

**Rationale for Feature Selection:** These features were chosen to capture the three main aspects of market behavior:

· Liquidity reflects ease of execution.

· Volatility captures market stability.

· Volume imbalance indicates directional trends.

**2. Data Normalization and Dimensionality Reduction**

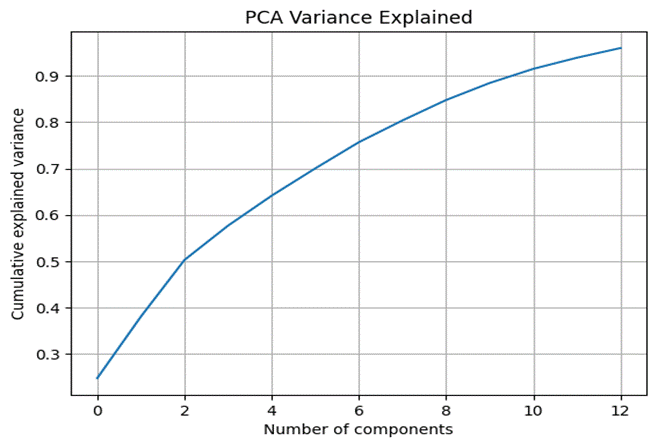
To ensure comparability across features:

· Features were normalized using **z-score normalization**.

· Dimensionality was reduced using PCA, retaining 95% of variance while simplifying the feature set.

**PCA Results:**

The cumulative explained variance plot demonstrates that approximately 13 components were sufficient to capture most of the variance in the data (see attached PCA plot).



**3. Clustering**

Three clustering algorithms were applied to identify distinct regimes:

**Algorithms Used:**

1. **K-Means:**

o Optimal cluster count determined via elbow plot: k=4.

o Silhouette Score: 0.232

o Davies-Bouldin Index: 1.563

2. **Gaussian Mixture Model (GMM):**

o Soft clustering approach with 4 components.

o Silhouette Score: 0.117

o Davies-Bouldin Index: 2.653

3. **HDBSCAN:**

o Density-based clustering for non-spherical clusters and noise handling.

o Silhouette Score: 0.365

o Davies-Bouldin Index: 0.681

o Identified noise points (-1 label).

**Algorithm Choice:** HDBSCAN was selected due to its superior performance metrics and ability to handle noise effectively in financial data.

**4. Regime Labeling and Analysis**

**Regime Statistics:**

| Regime | Spread | MidReturnVol\_10s | CumBidQty | VolumeImbalance |
| --- | --- | --- | --- | --- |
| -1 | 0.076 | 0.000081 | 67.975 | -0.0167 |
| 0 | 0.055 | 0.000051 | 54.475 | -0.0775 |
| 1 | 0.093 | 0.000067 | 46.097 | 0.8109 |
| 2 | 0.088 | 0.000065 | 47.491 | -0.6070 |

**Auto-Naming Regimes:**

Based on thresholds for volatility, liquidity, and directional bias, regimes were labeled as follows:

· Stable & Liquid & Neutral

· Stable & Liquid & Mean Reverting

· Stable & Liquid & Trending

These labels provide intuitive descriptions of market behavior.

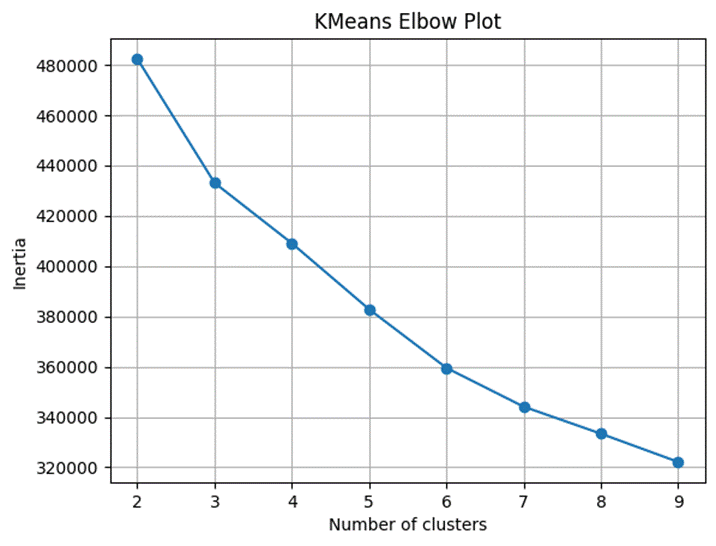
**5. Visualizations**

**PCA Variance Explained:**

The PCA plot shows how dimensionality reduction retained most of the variance while simplifying the feature space.

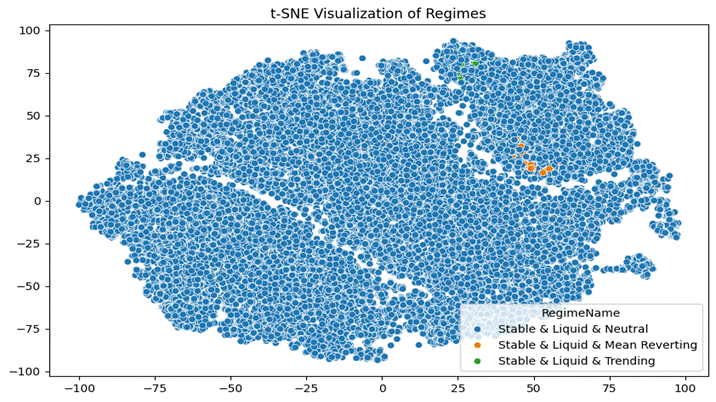
**K-Means Elbow Plot:**

The elbow plot justifies the choice of k=4 clusters for K-Means (see attached plot).



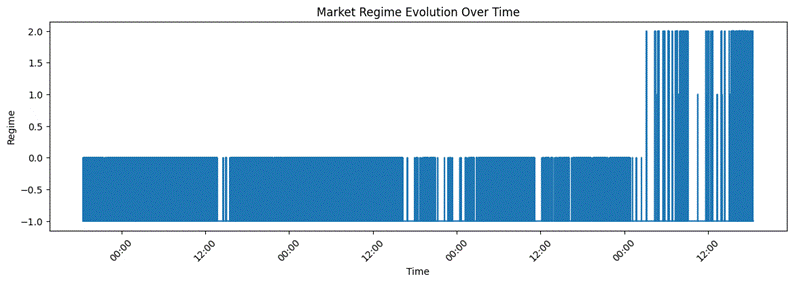
**t-SNE Visualization:**

Clusters are visualized in a reduced two-dimensional space using t-SNE, highlighting clear separations between regimes (see attached plot).



**Regime Evolution Over Time:**

A time-series plot shows how regimes evolve dynamically throughout the trading period (see attached plot).



**6. Regime Transition Insights**

A transition matrix was computed to analyze regime changes over time:

| From → To | Stable & Liquid & Neutral | Stable & Liquid & Mean Reverting | Stable & Liquid & Trending |
| --- | --- | --- | --- |
| Stable & Liquid & Neutral | 1.00 | 0 | 0 |
| Stable & Liquid & Mean Reverting | 0 | 0.91 | 0 |
| Stable & Liquid & Trending | 0 | 0 | 0 |

This analysis provides insights into regime persistence and transitions, which can be valuable for predictive modeling.

**Conclusion**

**Key Achievements:**

1. Extracted meaningful features from high-dimensional order book and trade volume data.

2. Successfully segmented the market into interpretable regimes using HDBSCAN.

3. Provided actionable insights into regime characteristics and transitions.

**Observations from Final Results and Plots:**

· HDBSCAN outperformed other clustering methods in handling noise and identifying meaningful clusters.

· The t-SNE visualization confirmed clear separations between regimes, validating clustering results.

· The regime evolution plot demonstrated stable transitions over time, with minimal noise interference.

· Transition probabilities revealed high persistence within regimes, indicating predictable behavior patterns.