Feature Extraction and Clustering Approach

This presentation explains the process of extracting features from trading data and clustering using PCA and K-means. We cover data merging, feature engineering, dimensionality reduction, and cluster analysis.

VT **by Vishwajit Tekam**

Data Preparation Workflow

File Conversion

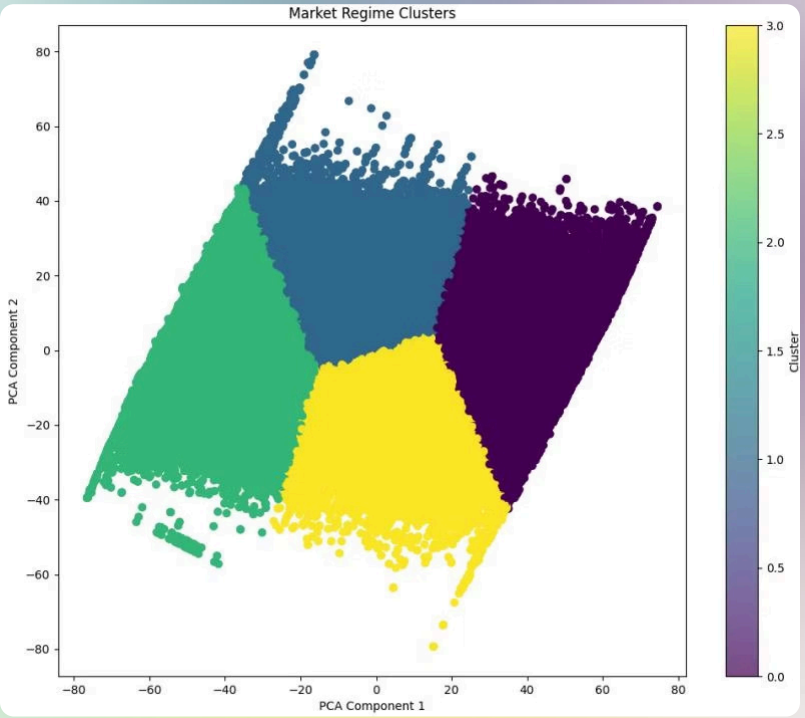
Convert txt files to csv for processing.

Data Merging

Merge aggTrade and depth20 files to create unified datasets.

Feature Extraction

Combine files to extract relevant features for analysis.



Types of Features Extracted

Basic & Liquidity

Mid price, bid-ask spread, and spread in basis points.

Order Book & Depth

Imbalance, microprice, cumulative quantities, and depth slopes.

Price & Volume

Returns, volatility, VWAP, and trade statistics over time windows.

Value	2000	2510	1210	4.00
	1720	3490	3310	4.00
Value	3330	7210	4010	1.00
	1200	1270	4500	4.00
Status	1500	1900	4520	7.00

Normalization and Cleaning

Normalization

All numeric features are z-score scaled.

NaN Handling

NaN values are filled with zero after normalization.

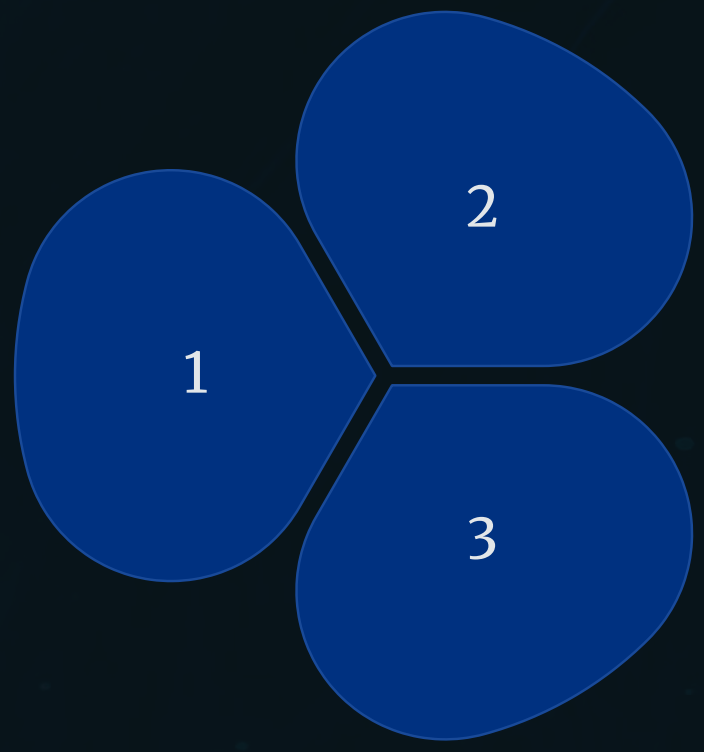
Final Dataset

Over 1.5 million rows with 75 features, 32 used for clustering.

Dimensionality Reduction with PCA

32 Features

Initial set for clustering.

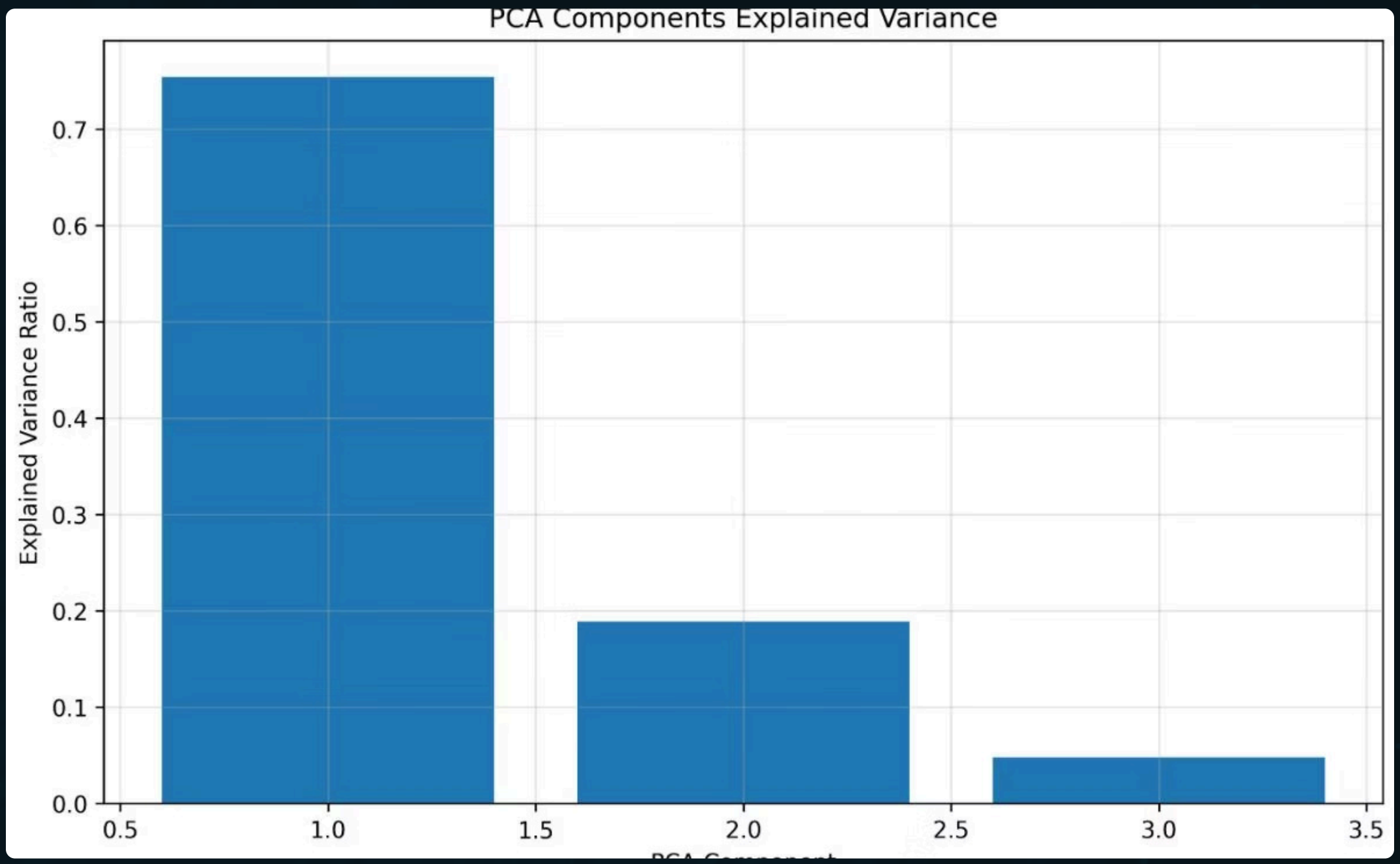


PCA Applied

Reduced to 3 principal components.

Efficient Clustering

Lower dimensions enable faster, clearer clustering.



K-means Clustering Process

1

Feature Selection

Use 32 normalized features.

2

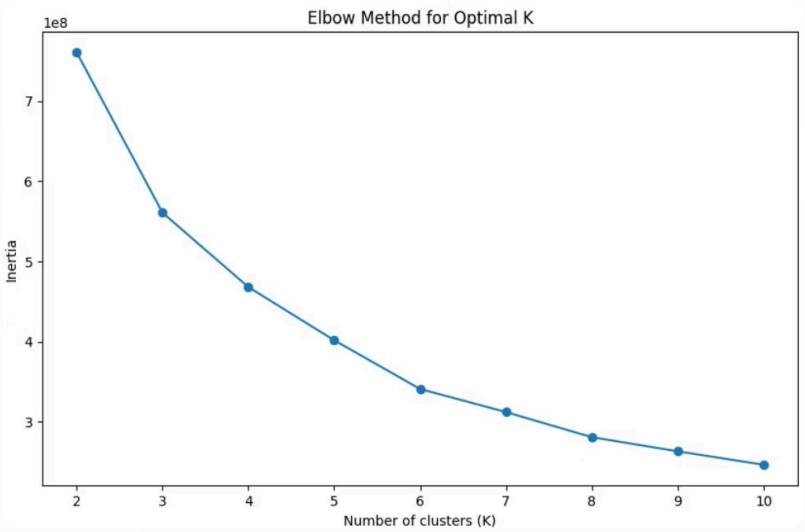
PCA Reduction

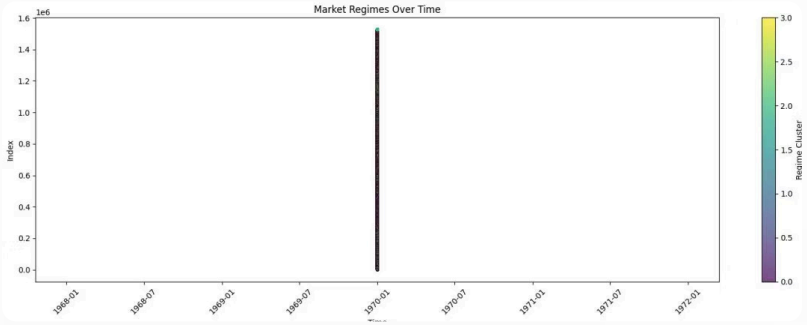
Reduce to 3 dimensions.

3

K-means Clustering

Test different K values for optimal clusters.





Key Cluster Characteristics

Trending vs Mean-Reverting

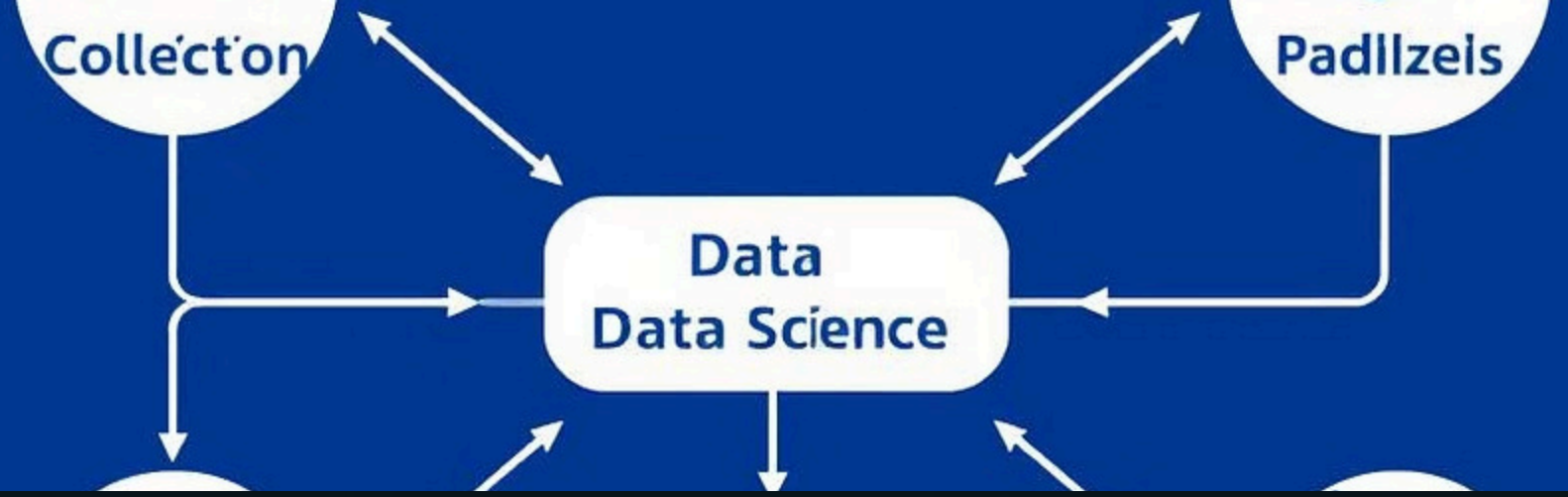
Identifies directional or reverting price behavior.

Volatile vs Stable

Measures price fluctuation intensity.

Liquid vs Illiquid

Assesses market depth and ease of trading.



Summary and Insights

Data Merged

Unified files and extracted features.

PCA Applied

Reduced dimensionality for efficient clustering.

Clusters Analyzed

Identified key market behaviors using three principal features.