FINANCIAL STOCK ANALYSIS AND CLUSTERING

Analyzing 157 stocks of US based companies operating in the Energy sector

Raza Mehar

Najam Mehdi

Pujan Thapa

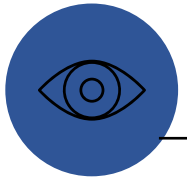
Subject: Hardware and Software for Big Data

Created on: December 24, 2023
Last updated on: January 23, 2024
Version No: 5.0

Università degli Studi di Napoli Federico II

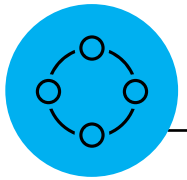
Table of Contents

Data Analysis Workflow



RETRIEVE & EXPLORE

- Crawling data
- Reviewing metadata
- Inspecting initial data samples and data types
- Understanding data dimensions

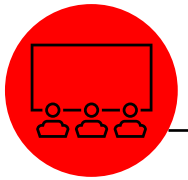


ANALYZE

- Performing descriptive statistics
- Detecting outlier
- Identifying top and bottom 10 stocks
- Displaying time-series visualizations
- Engineering relevant features for technical analysis
- Evaluating stocks and reviewing their returns
- Analyze risks
- Clustering the data based on three different algorithms: Kmeans, Hierarchical and Spectral Clustering

PROCESS

- Flattening the data frame
- Dropping less relevant features
- Handling duplicate and missing values
- Standardizing the data

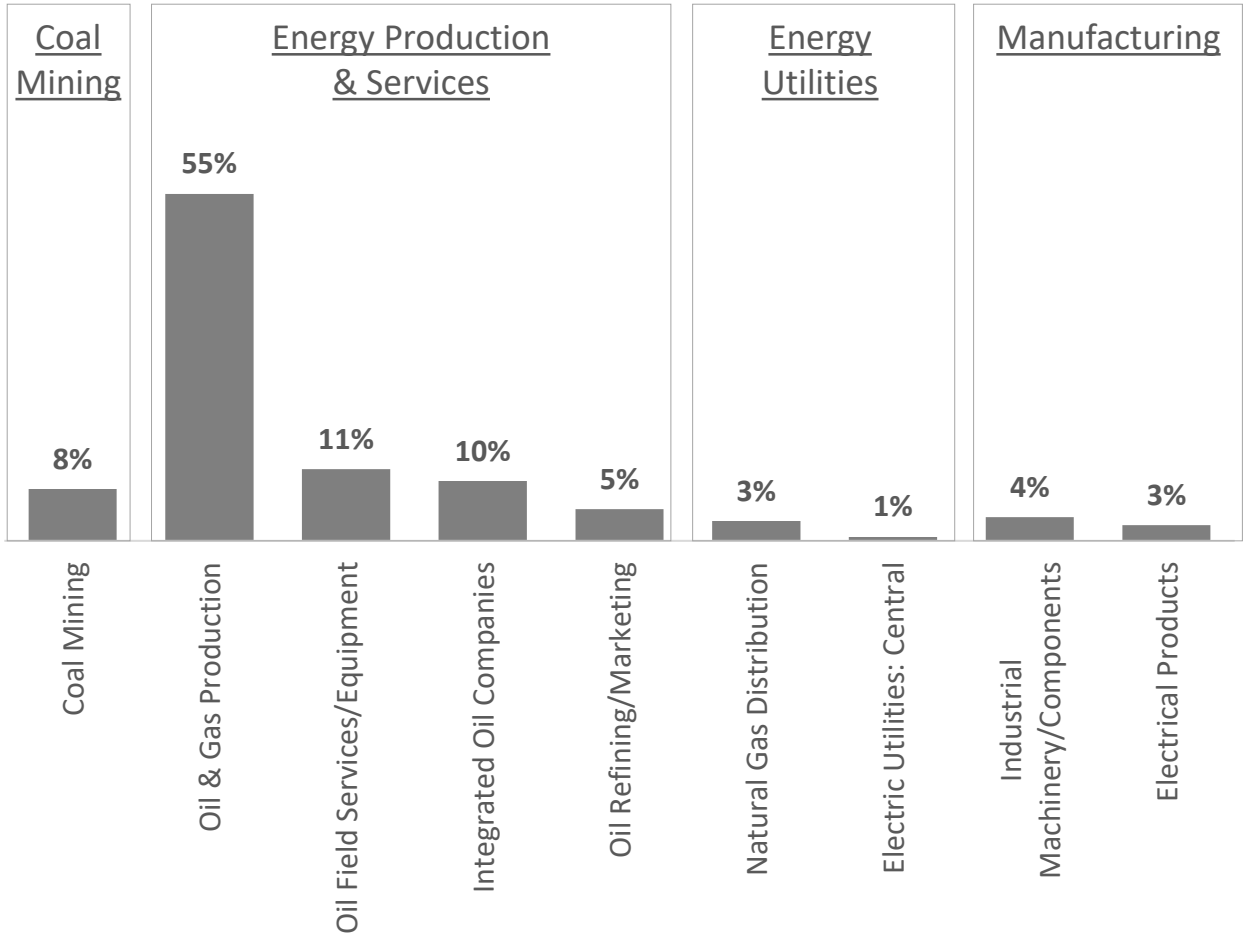


SHARE

- Interpreting clusters
- Storing the processed data in Kakfa

Stock Data Distribution Analysis

Retrieved stock data for all the 157 US energy companies from January to December 2023 from Yahoo! Finance; subsector and industry-wise distribution and names of companies are given below.



Subsector and Industry wise distribution of companies

Company names

Stock Data Review

Below are the features, their data types, and descriptions, along with the dimensions of the data and an output snippet.

Feature	Data Type	Description
Date	Date	Date of the observation in (YYYY-DD-MM) format
Stock	String	Stock symbol of a company
Open	Float	Opening price per stock
High	Float	Highest price per stock in a day
Low	Float	Lowest price per stock in a day
Close	Float	Closing price per stock
Adj Close	Float	Closing price per stock after adjusting for dividends
Volume	Float	Total number of shares traded in a day
Dimension	1853, 8	

Output snippet of the code after features exclusion

```
Head of the dataframe:
         Date Stock   Adj Close        High         Low
0 2023-01-01  ACDC   22.500000   25.440001   19.832001
1 2023-01-01    AE   46.759796   51.500000   37.500000
2 2023-01-01  AMPY    8.660000    9.360000    7.700000
3 2023-01-01   AMR  159.314117  174.554993  130.539993
4 2023-01-01   APA   42.908794   46.980000   41.360001

Tail of the dataframe:
            Date Stock   Adj Close        High         Low
1848 2023-12-01   WTI    3.260000    3.440000    2.860000
1849 2023-12-01  WTTR    7.590000    7.818000    7.065000
1850 2023-12-01   WWD  136.130005  140.729996  132.440002
1851 2023-12-01   XOM   99.980003  104.220001   97.480003
1852 2023-12-01  XPRO   15.920000   17.000000   14.440000
```

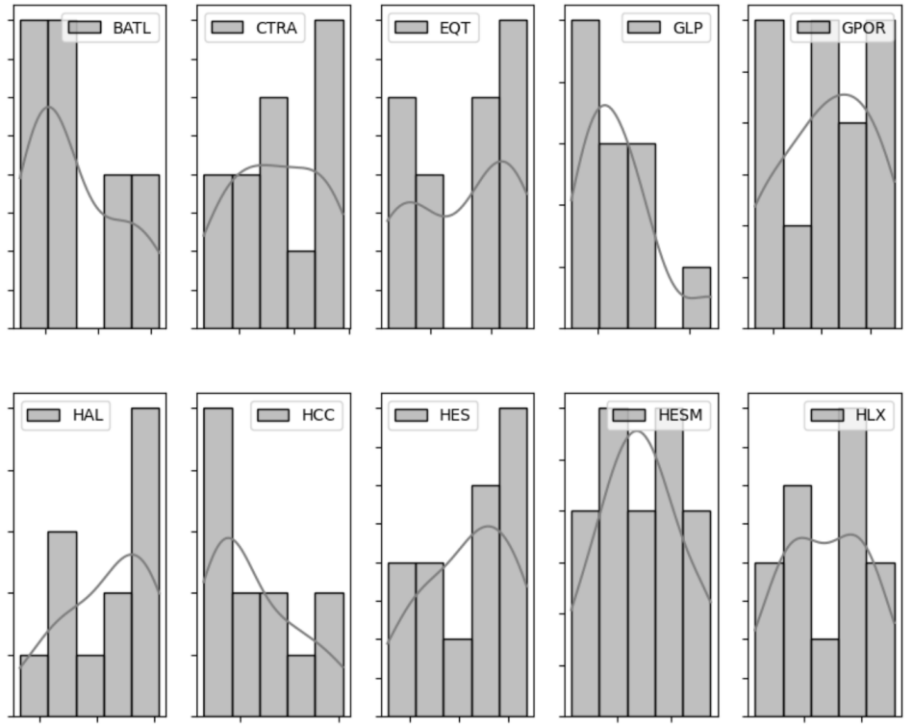
No duplicate or missing values were observed in the data.

NOTE:
Feature Exclusion: *Open*, *Close*, and *Volume* features are irrelevant to the analysis and will be excluded.
Feature Engineering: 6 new features will be created for technical analysis.

Descriptive Statistics

Performing descriptive statistics on adjusted closing price feature on 10 random stocks to observe the data distributions.

Stock	Count	Mean	SD	Min	25%	50%	75%	Max
BATL	12	7.2	1.7	4.9	6.0	6.4	8.6	10.7
CTRA	12	25.4	1.6	23.2	24.9	25.3	27.0	28.5
EQT	12	37.7	4.3	32.1	32.9	37.3	41.1	43.7
GLP	12	33.2	4.0	29.2	31.2	33.2	35.1	37.0
GPOR	12	103.3	23.8	67.9	79.8	99.6	119.1	136.5
HAL	12	35.8	3.9	28.7	33.4	37.9	39.4	40.8
HCC	12	43.0	9.1	32.8	36.7	38.9	44.9	56.0
HES	12	141.7	9.0	126.9	138.6	141.8	149.6	156.6
HESM	12	28.6	1.9	27.5	29.1	29.8	30.4	32.5
HLX	12	8.8	1.4	6.3	7.4	8.2	9.6	11.0



Histograms of 10 random stocks

NOTE:
Upon observing the descriptive statistics, particularly the mean and standard deviation, for each stock, it is evident that the data distributions vary. To facilitate meaningful comparisons and enhance the effectiveness of clustering, data would be standardized, ensuring that all stocks are on a comparable scale.

Top and Bottom 10 Stocks

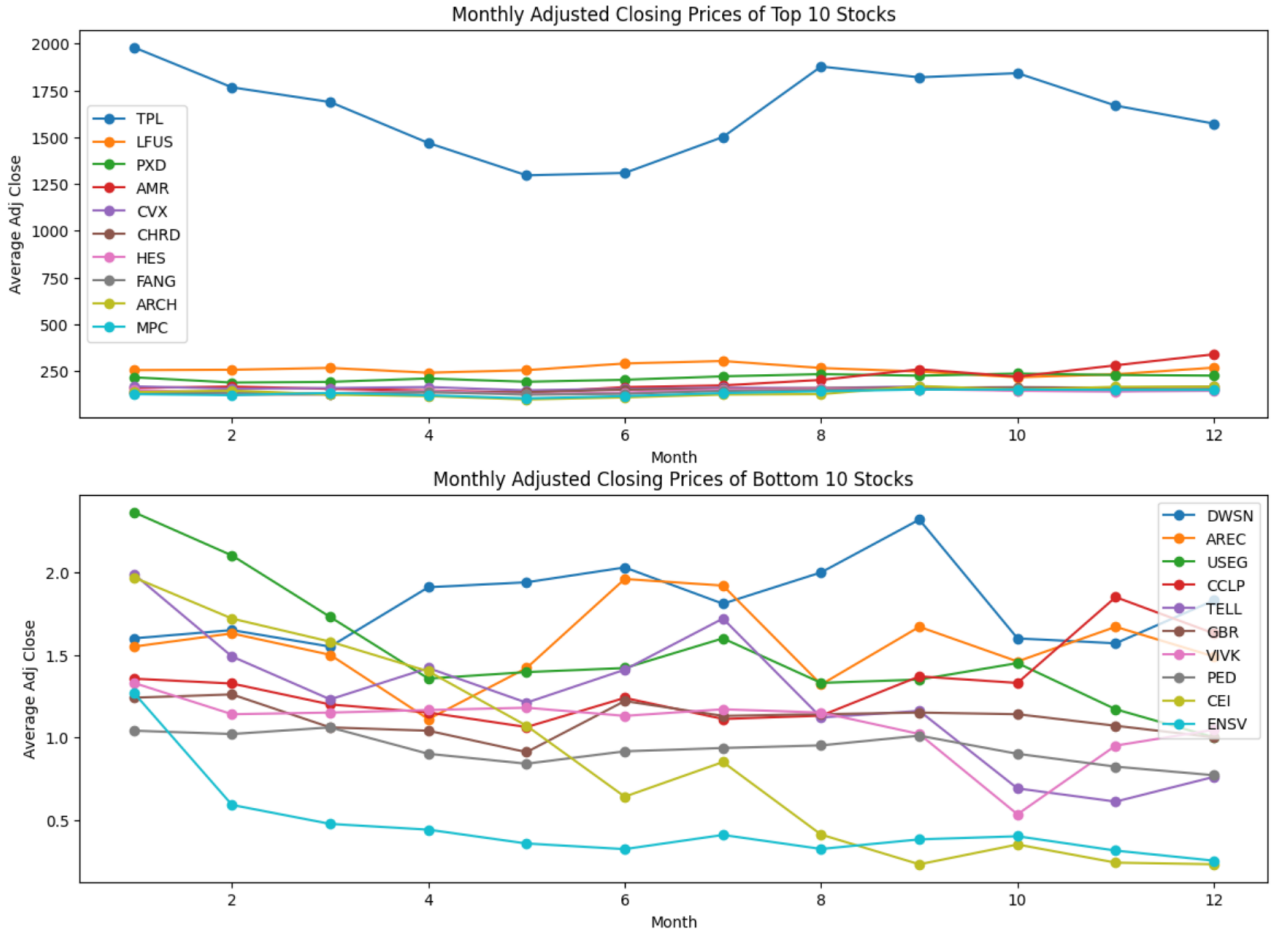
The top and bottom 10 stocks, determined by the mean of the adjusted closing price, are as follows. The time-series visualizations display the monthly adjusted closing prices for both sets of stocks.

Top 10

Stock	Adj Close	Stock	Adj Close
TPL	1649.1	CHRD	146.8
LFUS	257.6	HES	141.7
PXD	213.9	FANG	141.1
AMR	199.5	ARCH	135.5
CVX	155.4	MPC	132.3

Bottom 10

Stock	Adj Close	Stock	Adj Close
DWSN	1.8	GBR	1.1
USEG	1.5	VIVK	1.0
AREC	1.5	PED	0.9
CCLP	1.3	CEI	0.8
TELL	1.2	ENSV	0.4



Time-Series Analysis of top and bottom 10 stocks

NOTE:

TPL seems like an outlier within the group of stocks.

Feature Engineering for Technical Analysis

6 indicators have been created from the existing features that will aid in technical analysis.

Simple Moving Average

- The average price over a specified time.
- Price above SMA indicates an uptrend (Bullish Signal)
- Price below SMA indicates a downtrend (Bearish Signal)

Exponential Moving Average

- The average price over a specified time calculated by giving more weights to recent prices
- Price above EMA indicates an uptrend (Bullish Signal)
- Price below EMA indicates a downtrend (Bearish Signal)

Relative Strength Index

- The speed and change of price movements
- Ranges from 0 to 1
- RSI value above 70 indicates an overbought stock
- RSI value below 30 indicates oversold stock

Returns

- The percentage change in the price value over a specified time.
- Positive returns indicate a gain
- Negative returns represent a loss

Volatility

- The degree of variation of a trading price series over time
- It is a critical factor in risk management
- Higher volatility implies greater price variability and increased risk

Average True Range

- The measure of volatility
- It is a true range of price movements.
- High ATR value indicates greater volatility
- Low value suggests less volatility

NOTE:
These indicators will help us evaluate stocks, review their returns and losses, and conduct risk analysis.

Stock Evaluation

Bullish stocks in December 2023. Bullish stocks are strong stocks that are expected to rise in value.

Bullish Stocks (SMA)
PAGP, CRT, ARKO, METCB, CLMT, NRP, WTTR, MGEE, MTR, WKC, MPLX, AMR, SUN, GEL, METC, PAA, NGS, DMLP, NS, VTS, EP, TUSK, GPOR, DKL, EGY, FCEL, AROC, WWD, PSX, NSS, CCLP, CAPL, ARCH, THR, POWL, LBRT, SBR, VNOM, CEIX, VIVK, HCC, BE, AREC, GLP, HESM

Bullish Stocks (EMA)
PAGP, CRT, METCB, CLMT, NRP, BTU, MGEE, MTR, WKC, MPLX, AMR, SUN, GEL, METC, PAA, NGS, DMLP, NS, VTS, EP, HPKEW, TUSK, GPOR, DKL, EGY, AROC, WWD, PSX, NSS, CKX, CCLP, PNRG, DK, CAPL, ARCH, THR, METCL, MPC, POWL, LBRT, VNOM, CEIX, VIVK, HCC, BE, AREC, GLP, PARR, HESM

NOTE:
When the adjusted closing price of a stock is above its simple moving average or exponential moving average, it is identified as a bullish stock.

Stock Evaluation

Bearish stocks in December 2023. Bearish stocks are weak stocks that are expected to fall in value.

Bearish Stocks (SMA)	Bearish Stocks (EMA)
ENSV, CVX, LFUS, MGY, NINE, SWN, KLXE, PBT, CRK, HPK, VTLE, PVL, SLDP, MXC, REPX, EQT, AMPY, PTEN, VOC, EOG, DK, MVO, BRN, VLO, HES, HAL, MTDR, COP, WTI, HUSA, BSM, PUMP, AR, NCSM, OII, CRC, BRY, XPRO, CIVI, CNX, VAL, MUR, METCL, CRGY, USEG, DINO, SJT, STR, CEI, CHKEW, RES, AE, XOM, BATL, BTU, TALO, GBR, CHKEL, NC, DVN, BPT, PRT, SM, MARPS, CTRA, ACDC, RRC, SBOW, CHRD, DWSN, NOG, CKX, EPM, TELL, PBF, HNRG, PLUG, NRT, PXD, PED, PARR, ARLP, SD, TELZ, PR, CPE, REI, DO, OXY, FANG, CHK, HPKEW, ROI, HP, PHX, EPSN, HLX, KRP, TPL, RNGR, APA, CVI, PNRG, MRO, TTI, EAF, MPC, VTNR, MMLP, NFG'	ENSV, CVX, LFUS, MGY, ARKO, WTTR, NINE, SWN, KLXE, PBT, CRK, HPK, VTLE, PVL, SLDP, MXC, FCEL, REPX, PTEN, EQT, AMPY, VOC, EOG, MVO, BRN, VLO, HES, HAL, MTDR, COP, WTI, HUSA, BSM, PUMP, AR, NCSM, OII, CRC, BRY, XPRO, CIVI, CNX, VAL, MUR, CRGY, USEG, DINO, SJT, STR, CEI, CHKEW, RES, AE, XOM, BATL, TALO, GBR, CHKEL, NC, DVN, BPT, PRT, SM, MARPS, CTRA, ACDC, RRC, SBOW, CHRD, DWSN, NOG, EPM, TELL, PBF, HNRG, PLUG, NRT, PXD, PED, ARLP, SD, TELZ, PR, CPE, REI, DO, OXY, FANG, CHK, ROI, HP, PHX, EPSN, HLX, KRP, TPL, RNGR, APA, CVI, MRO, TTI, EAF, SBR, VTNR, MMLP, NFG'

NOTE:
When the adjusted closing price of a stock is below its simple moving average or exponential moving average, it is identified as a bearish stock.

Stock Evaluation

Overbought and oversold stocks in December 2023. Overbought stocks are those whose prices have increased rapidly, potentially increasing their intrinsic values, while oversold stocks are those whose prices have decreased rapidly, potentially decreasing their intrinsic values.

Overbought Stocks	Oversold Stocks
PAGP, METCB, NRP, MPLX, AMR, SUN, GEL, METC, PAA, NGS, NS, GPOR, DKL, EGY, AROC, WWD, PSX, NSS, CKX, CCLP, CAPL, ARCH, POWL, LBRT, VNOM, CEIX, HCC, THR, HESM	ENSV, CVX, RES, WTI, HUSA, NFG, SD, TELZ, AE, NINE, XOM, BATL, TALO, PBT, CRK, NCSM, CPE, DVN, VTLE, BPT, REI, PRT, OXY, PVL, SLDP, CHK, MARPS, MXC, ROI, FCEL, ACDC, HP, REPX, PTEN, AMPY, XPRO, DWSN, CIVI, EPM, APA, TELL, EAF, VTNR, USEG, PLUG, NRT, PED, STR, CEI

NOTE:
When the relative strength index of a stock is above 70, it is identified as an overbought stock, and when the relative strength index is below 30, it is identified as an oversold stock.

Stock Return Review and Risk Analysis

Profitable, unprofitable, and volatile stocks in December 2023. Positive stocks report positive earnings, while unprofitable stocks report negative earnings. Volatile stocks are those whose prices fluctuate frequently and significantly.

Profitable Stocks (>= 50%)
VIVK

Unprofitable Stocks (< 0%)
ENSV, CVX, MGY, ARKO, NINE, SWN, KLXE, PBT, CRK, HPK, VTLE, MXC, REPX, EQT, AMPY, PTEN, VOC, EOG, MVO, VLO, HES, HAL, MTDR, COP, WTI, HUSA, BSM, PUMP, AR, NCSM, OII, VTS, CRC, XPRO, CIVI, CNX, MUR, CRGY, USEG, DINO, SJT, STR, CEI, CHKEW, RES, AE, XOM, BATL, TALO, GBR, CHKEL, NC, DVN, BPT, PRT, SM, MARPS, CTRA, ACDC, RRC, SBOW, CHRD, NOG, CKX, EPM, TELL, PBF, HNRG, LBRT, PLUG, NRT, PXD, PED, ARLP, SD, TELZ, PR, CPE, REI, OXY, FANG, CHK, HPKEW, ROI, HP, PHX, EPSN, HLX, KRP, TPL, RNGR, APA, CVI, PNRG, MRO, TTI, EAF, MPC, VTNR, MMLP, NFG

Most Volatile Stock (STDV)
TPL

Most Volatile Stock (ATR)
TPL

NOTE:
Volatile stocks are those with the maximum standard deviation and average true range.

Pre-Clustering Preparation

Decided to use KMeans, Agglomerative Hierarchical, and Spectral Clustering algorithms. The reasons are given below. Used the Elbow Curve to find the optimal value for the number of clusters used in these algorithms.

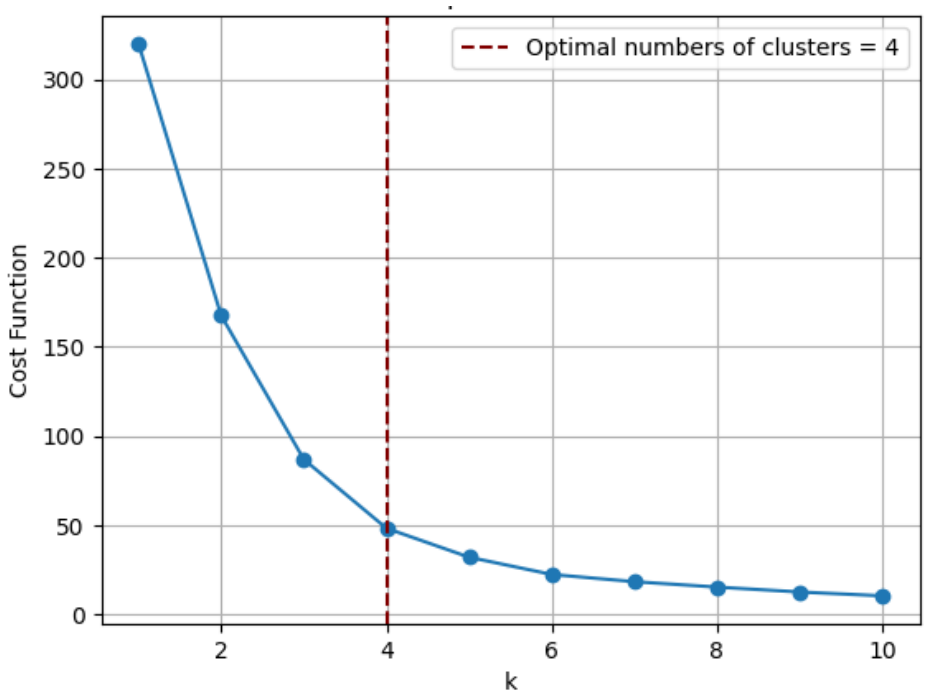
Reasons for choosing different clustering algorithms:

The reason for choosing *KMeans Clustering* was its simplicity of use and computational efficiency.

Subsequently, we aimed to explore potential hierarchical relationships between the stocks, leading us to employ *Agglomerative Hierarchical Clustering*.

Following that, our interest turned to investigating any non-linear relationships among the stocks, prompting the use of *Spectral Clustering*.

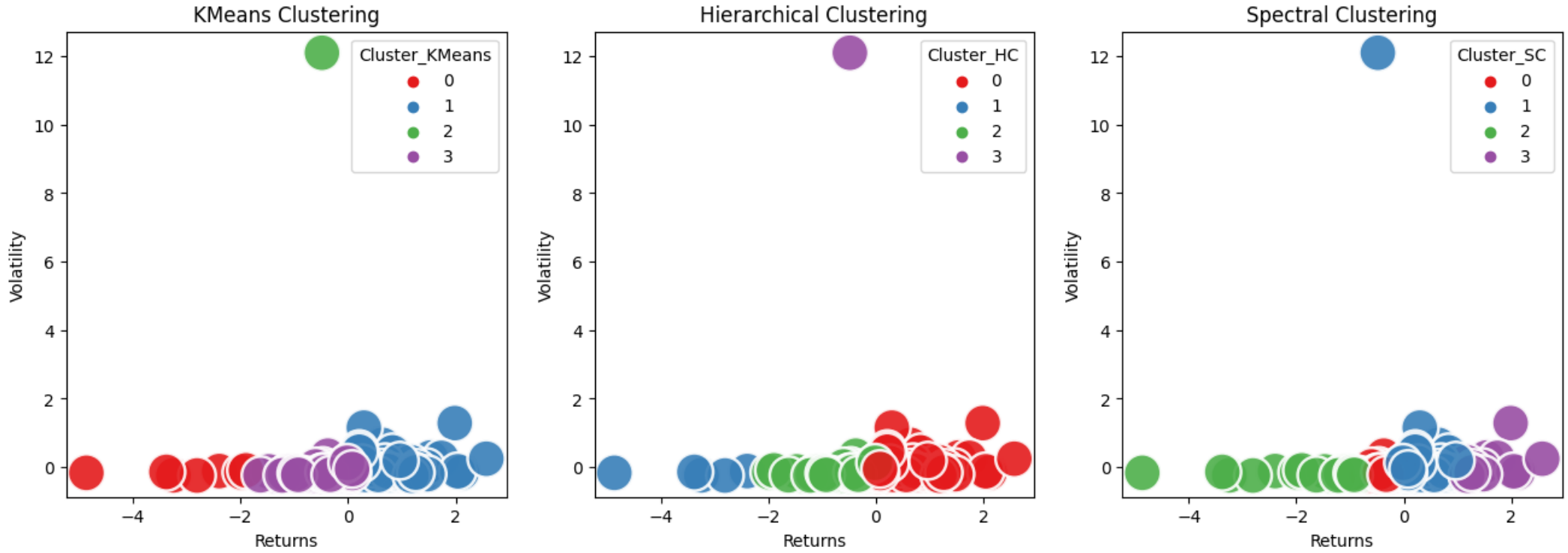
The combination of these algorithms provided insights into the nature of the stocks.



Elbow Curve showing the most optimal k-value

NOTE:
The optimal value of number of clusters as per the Elbow Curve is 4.

Clusters Visualization



Number of stocks per cluster as per the clustering algorithm

Cluster	KMeans	Hierarchical	Spectral
0	7	89	39
1	68	5	74
2	1	60	25
3	79	1	17

Interpretation

The interpretation of each clustering output from the three algorithms, along with the combined summary, is given below

KMeans Clustering

- **Cluster 0 (red):** Stocks in this cluster are unprofitable and likely to continue losing money in the future.
- **Cluster 1 (blue):** Represents stocks with high profits and lower volatility, considered stable and reliable.
- **Cluster 2 (green):** An outlier containing highly volatile stocks, such as TPL, with potential for medium to higher returns.
- **Cluster 3 (purple):** Includes stocks with medium profits and lower volatility, offering a middle ground between riskier and more stable stocks.

Agglomerative Hierarchical Clustering

- **Cluster 1 (blue):** Stocks are unprofitable and may continue to incur losses.
- **Cluster 0 (red):** Comprises stocks with high profits and lower volatility, indicating stability.
- **Cluster 2 (green):** Includes stocks with medium profits and lower volatility, providing a balanced option.
- **Cluster 3 (purple):** An outlier with TPL, a highly volatile stock offering medium to higher returns.

Spectral Clustering

- **Cluster 0 (red):** Stocks generate medium to somewhat higher profits with lower volatility.
- **Cluster 1 (blue):** Represents stocks with no clear correlation, suggesting a volatile nature.
- **Cluster 2 (green):** Comprises unprofitable stocks that have been losing money.
- **Cluster 3 (purple):** Stocks with less volatility and high returns, indicating strong correlation among them.

SUMMARY:

While each algorithm provides unique insights, there is a common theme across clusters suggesting that many energy sector stocks offer a balance of medium to high returns with relatively low volatility. This stability makes them potentially attractive to investors seeking a manageable level of risk in their portfolios.

NOTE:

The processed data has been assigned to a Kafka topic to be subscribed by other systems.

ADDENDUM

List of Python libraries and modules used

- **time** - Module providing various time-related functions
- **json** - Module for encoding and decoding JSON data
- **pandas** - Data manipulation and analysis library
- **numpy** - Numerical computing library
- **matplotlib.pyplot** - Plotting library for creating visualizations
- **seaborn** - Statistical data visualization library based on Matplotlib
- **talib** - Technical Analysis Library for financial markets
- **yfinance** - Library for downloading historical market data from Yahoo Finance
- **SparkSession from pyspark.sql** - Entry point for using Spark SQL
- **StandardScaler from sklearn.preprocessing** - Standardizes features by removing the mean and scaling to unit variance
- **KafkaProducer** - Python client for Apache Kafka, a distributed streaming platform
- **Kmeans from sklearn.cluster** - K-means clustering algorithm from scikit-learn
- **AgglomerativeClustering from sklearn.cluster** - Agglomerative hierarchical clustering from scikit-learn
- **SpectralClustering from sklearn.cluster** - Spectral clustering algorithm from scikit-learn