Friedrich-Alexander-Universität
Faculty of Sciences

FAU

Clustering

Anuraag Mishra- 23076160
Behnam Babaeian- 23177836
Elaheh Ghorbani- 23208589

AGENDA

1. Introduction
2. Hierarchical clustering
3. Overview of portfolio optimization
4. Coding

Friedrich-Alexander-Universität
Faculty of Sciences

FAU

Introduction

CLUSTERING

Learning methods:

- Supervised
- Semi-supervised
- Unsupervised

Cluster: A collection of data objects within a larger set that are:

- Similar to one another within the same group and,
- dissimilar to the objects outside the group.

Typical applications:

- As a stand-alone tool to get insight into data distribution.
- As a pre-processing step for other algorithms.

CLUSTERING ALGORITHM

Given

- N: number of data points
- M: number of variables (i.e “mass”, “price”, “color”, ...)
- Data X = $\{x_1, \dots, x_N\}$, where $x_n \in R_M$ for all n = 1,...,N
- K: assumed number of clusters

Want

- Assignment: $x_n \rightarrow k_n \in \{1,...,K\}$ for all n = 1,...,N
- Assignment rule: $x \rightarrow k(x) \in \{1,...,K\}$ for all $x \in R_M$
- Reconstruction rule (’representative’): $k \rightarrow m_k \in R_M$

CLUSTERING APPROACHES

- Partitioning approach:
  Construct various partitions and then evaluate them by some criterion. Like: k-means.

- ✓ **Hierarchical approach:**
  Create a hierarchical decomposition of the set of data using some criterion. Like: AGNES.

- Density-based approach:
  Based on connectivity and density functions. Like: DBSCAN.

- Grid-based approach:
  Based on a multiple-level granularity structure. Like: CLIQUE.

- Frequent-pattern-based approach:
  Based on the analysis of frequent patterns. Like: p- Cluster.

And many other approaches.

APPLICATIONS OF CLUSTERING IN FINANCIAL SUBJECTS

- Customer Segmentation
- Credit Risk Assessment
- Fraud Detection
- ✓ **Portfolio Optimization**
- Market Segmentation
- Algorithmic Trading
- Personalized Financial Planning

And many others.

Friedrich-Alexander-Universität
Faculty of Sciences

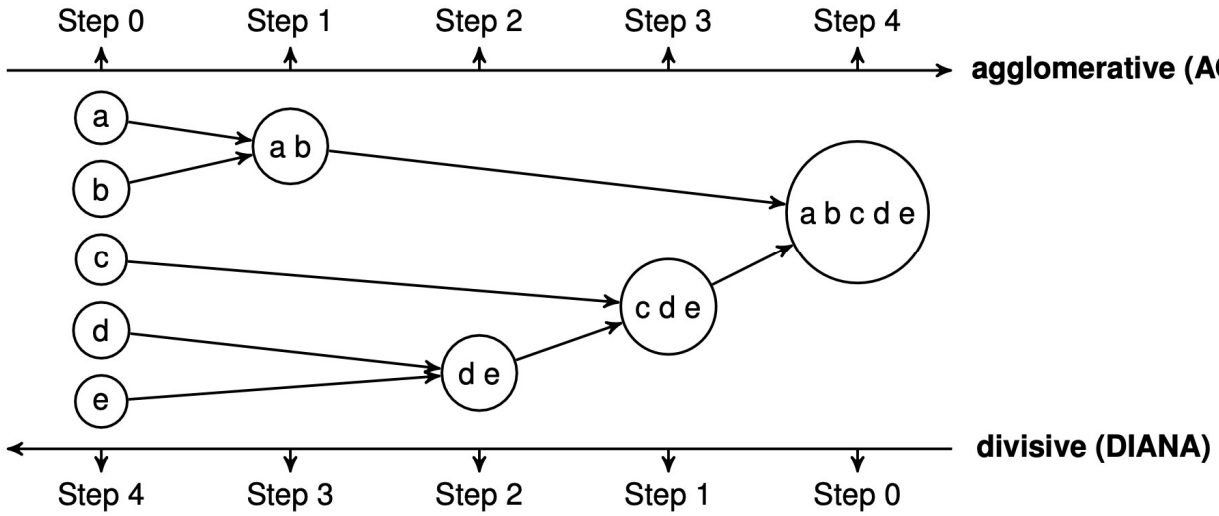
FAU

Hierarchical clustering

Was firstly introduced by Joseph Kruskal in 1964

MOST POPULAR METHODS

The main two methods of hierarchical clustering are AGNES and DIANA:
In this methods, there is no predefined number of clusters, instead there is a termination condition.



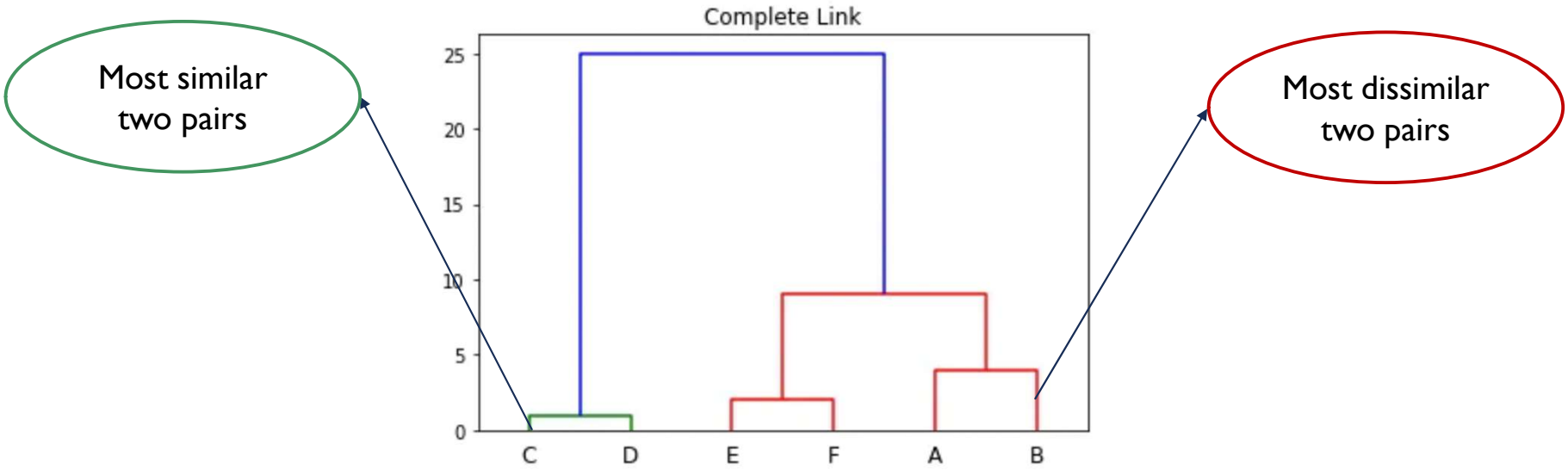
- There are two diagrams to show clustered data in hierarchical clustering:

1. Dendogram
2. Heat map

DENDOGRAM

Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.

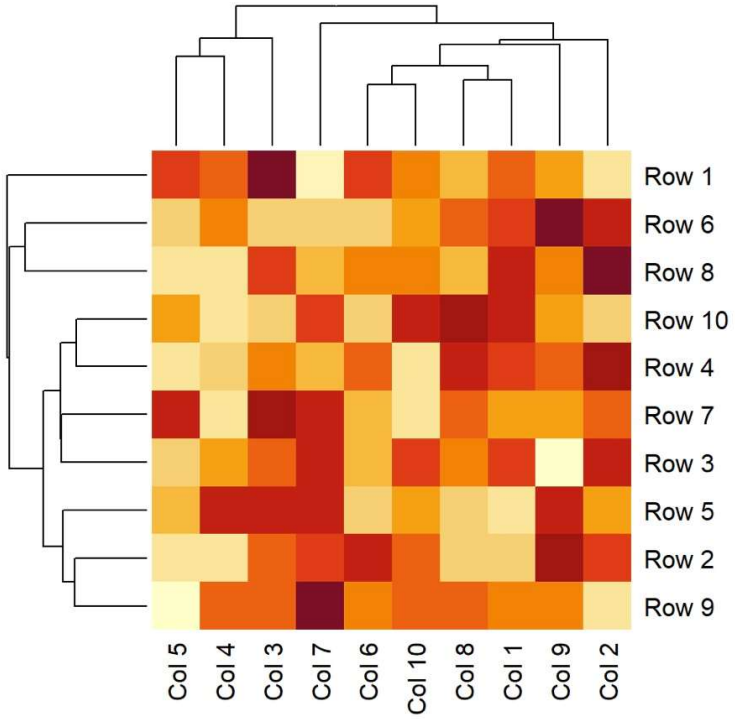
The hight of linkage lines reveals the similarity between previous clusters.



HEAT MAP

Hierarchical clustering is often associated with heat map.
It is a 2-dimensional data visualization technique that represents the magnitude of individual values within a dataset as a color.

Columns represent different samples.
Row represent different attributes.



AGNES (AGGLOMERATIVE NESTING)

Introduced by Kaufman et al.

Merge nodes that have the most similarity (or least dissimilarity).

Go on in a non-descending fashion.

Eventually all nodes belong to the same cluster.

- Before starting the AGNES clustering, you should define the similarity and termination conditions. In this way, the algorithm start to merge data points based on defined similarity and terminate iteration when it reaches to termination condition.

SIMILARITY METHODS

Similarity measures are mathematical techniques used to quantify the degree of similarity between two data points or objects.

- In the context of agglomerative clustering, similarity measures are crucial for determining how close or similar two clusters or data points are to each other (guide the merging process).

Some of similarity measures are:

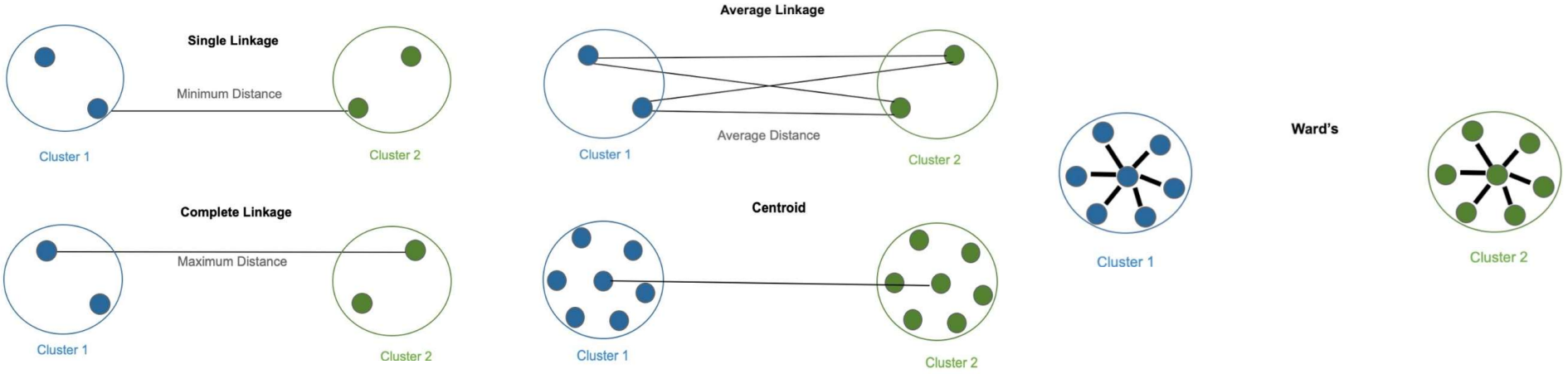
- ✓ **Euclidean distance**: $d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - qi)^2}$
- Manhattan distance: $d(p,q) = |\sum_{i=1}^{n}(p_i - qi)|$
- Cosine similarity: cosine-similarity(p,q) = $\frac{p.q}{\|p\|\|q\|}$
- Jaccard similarity: jaccard-similarity (p,q) = $\frac{|P \cap Q|}{|P \cup Q|}$

LINKAGE METHODS

Linkage methods in agglomerative clustering refer to the criteria used to determine how clusters are merged during the clustering process.

- These methods define the distance or similarity between clusters and guide the decision of which clusters to merge at each step.

- There are five different linkage methods.



LINKAGE METHODS

- Single linkage : d(cluster$_1$,cluster$_2$) = min (d(p,q)) for all p ∈ cluster$_1$ and q ∈ cluster$_2$.
- Complete linkage : d(cluster$_1$,cluster$_2$) = max (d(p,q)) for all p ∈ cluster$_1$ and q ∈ cluster$_2$.
- Average linkage : d(cluster$_1$,cluster$_2$) = $\frac{1}{|cluster_1| \times |cluster_2|} \sum_{p \in cluster_1} \sum_{q \in cluster_2} d(p,q)$
- Centroid linkage: d(cluster$_1$,cluster$_2$) = d($\frac{1}{|cluster_1|} \sum_{p \in cluster_1} \bar{p}$, $\frac{1}{|cluster_2|} \sum_{q \in cluster_2} \bar{q}$)
- ✓ **Ward's method** : d(cluster1,cluster2) = $\sqrt{\frac{|cluster_1 \cup cluster_2|}{|cluster_1|}}$ d(cluster$_1$,cluster$_2$)

TERMINATION CONDITION

There are different approaches to terminate iteration of AGNES algorithm.

- Number of Clusters:

AGNES can be terminated when a pre-defined number of clusters is reached.

- Distance Threshold:

The merging process can be terminated when the distance between the nearest neighbors (most similar elements) in the remaining clusters exceeds a certain threshold.

- Dendrogram Cutting:

The dendrogram resulting from the hierarchical clustering process can be cut at a certain height. Clusters below this height are considered the final clusters. This height corresponds to a specific level of dissimilarity or linkage distance.

- Percentage of Variance Explained:

termination may be based on specified percentage of the total variance in the data is captured by the identified clusters.

WEAKNESSES

Major weakness of agglomerative clustering methods:

- Can never undo what was done previously.
- Do not scale well: Time complexity of time complexity of $O(N^3)$ and a memory complexity of $O(N^2)$, where n is the number of objects.

Friedrich-Alexander-Universität
Faculty of Sciences

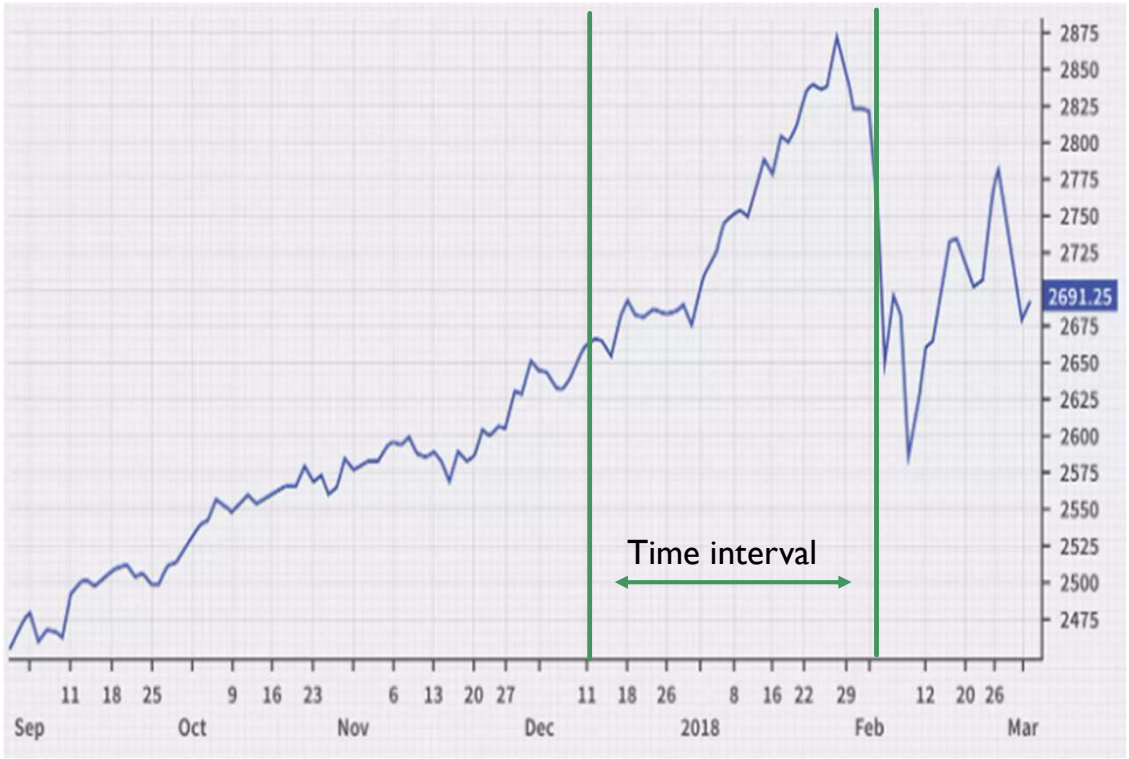
FAU

Overview of Portfolio Optimization

RISK & RETURN OF AN ASSET

$$Return = \frac{P_i - P_{(i-interval)}}{P_{(i-interval)}}$$

$$Risk(std) = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(n-1)}}$$



RISK/RETURN TRADE OFF

PORTFOLIO OPTIMIZATION

Portfolio optimization is a quantitative technique used in finance to construct an investment portfolio that:

- maximizes returns for a given level of risk
- minimizes risk for a targeted level of return

objective

- find the optimal allocation of assets within a portfolio to achieve the best risk-return trade off.

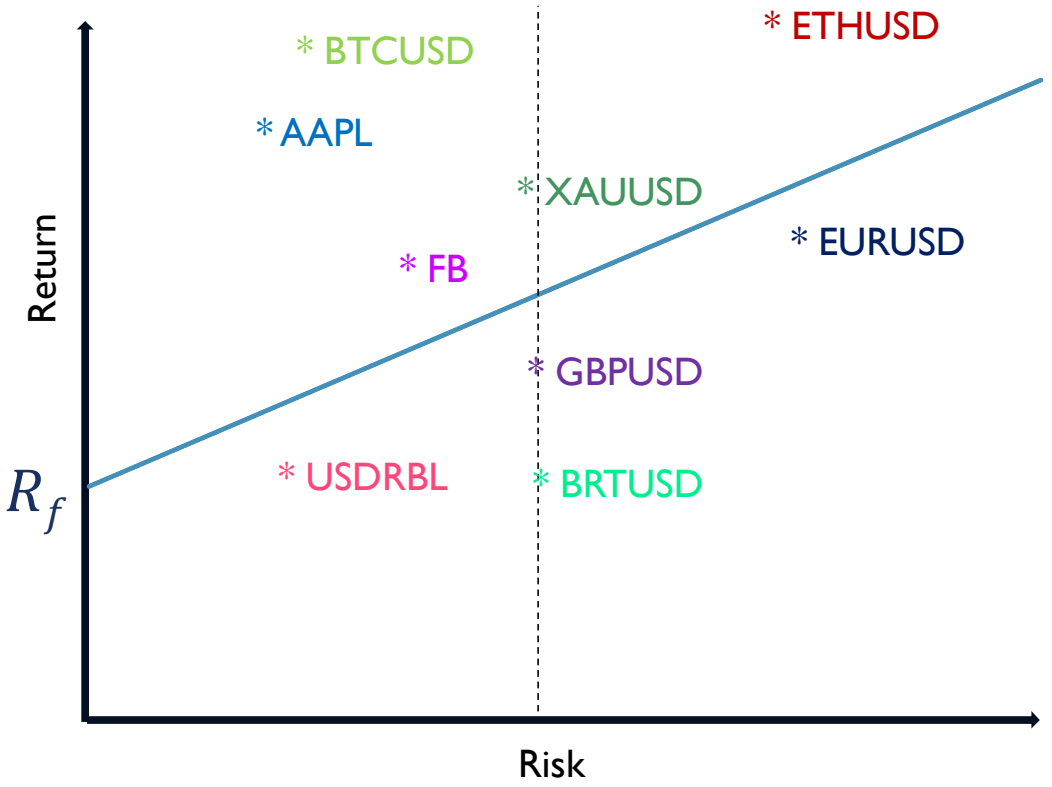
Sharpe ratio

- The Sharpe ratio is a measure of the risk-adjusted performance of an investment or a portfolio. It is widely used in finance to evaluate the return of an investment relative to its risk.

SHARPE RATIO

$$Sharpe\ ratio = \frac{R_e - R_f}{\sigma_P}$$

$$Returns = R_f + \text{sharp ratio} * Risks$$



RISK MANAGEMENT STRATEGIES

✓ **Diversification**
is a strategy that involves investing in a variety of assets to reduce the overall risk of the portfolio.

➢ **Asset Allocation**
is a strategy that involves dividing a portfolio among different asset classes, such as stocks, bonds, and cash.

➢ **Active Management**
is a strategy that involves making investment decisions based on market conditions and individual security analysis.

➢ **Passive Management**
is a strategy that involves investing in a portfolio that tracks a market index, such as the S&P 500

RISK DIVERSIFICATION IN PORTFOLIO MANAGEMENT

Risk diversification is a risk management strategy that involves spreading investments across different assets or types of investments in a portfolio.
Risk diversification in portfolio theory was introduced by Harry Markowitz.

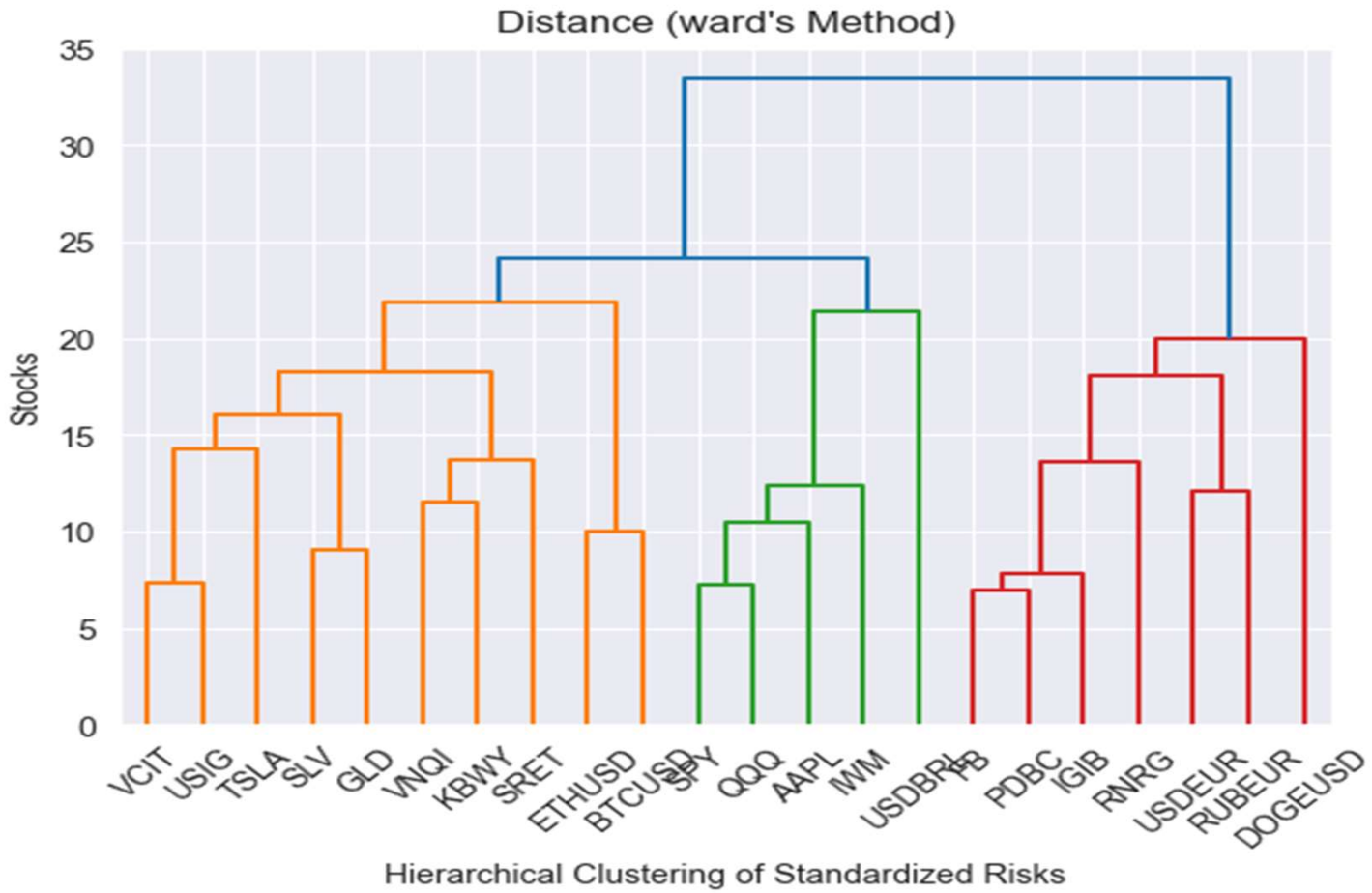
Objectives

- reducing the overall risk of the portfolio by mitigating the impact of poor-performing assets on the entire investment.

Core idea

- Assets may react differently to various economic, market, or geopolitical factors. Thus, in a diversified portfolio, the potential losses from underperforming or volatile investments may be offset by gains in other parts of the portfolio.

- well-diversified portfolio should optimize the trade-off between risk and return.


Distance (ward's Method)
Stocks
35
30
25
20
15
10
5
0
VCIT
USIG
TSLA
SLV
GLD
VNQI
KBWY
SRET
ETHUSD
BTCUSD
SPY
QQQ
AAPL
IWM
USDBRL
PB
PDBC
IGIB
RNRG
USDEUR
RUBEUR
DOGEUSD
Hierarchical Clustering of Standardized Risks

Friedrich-Alexander-Universität
Faculty of Sciences

FAU

Code & Analysis

CODE

- ✓ **Motivation/Challenges**
  - Static data vs Time Series
  - Visualization
  - Pipeline
  - Simple

CODE

Codebase:

- src → finance_ml clustering → ***Hierarchical.py***
- *notebooks* → ***clustering_hierarchical.ipynb***

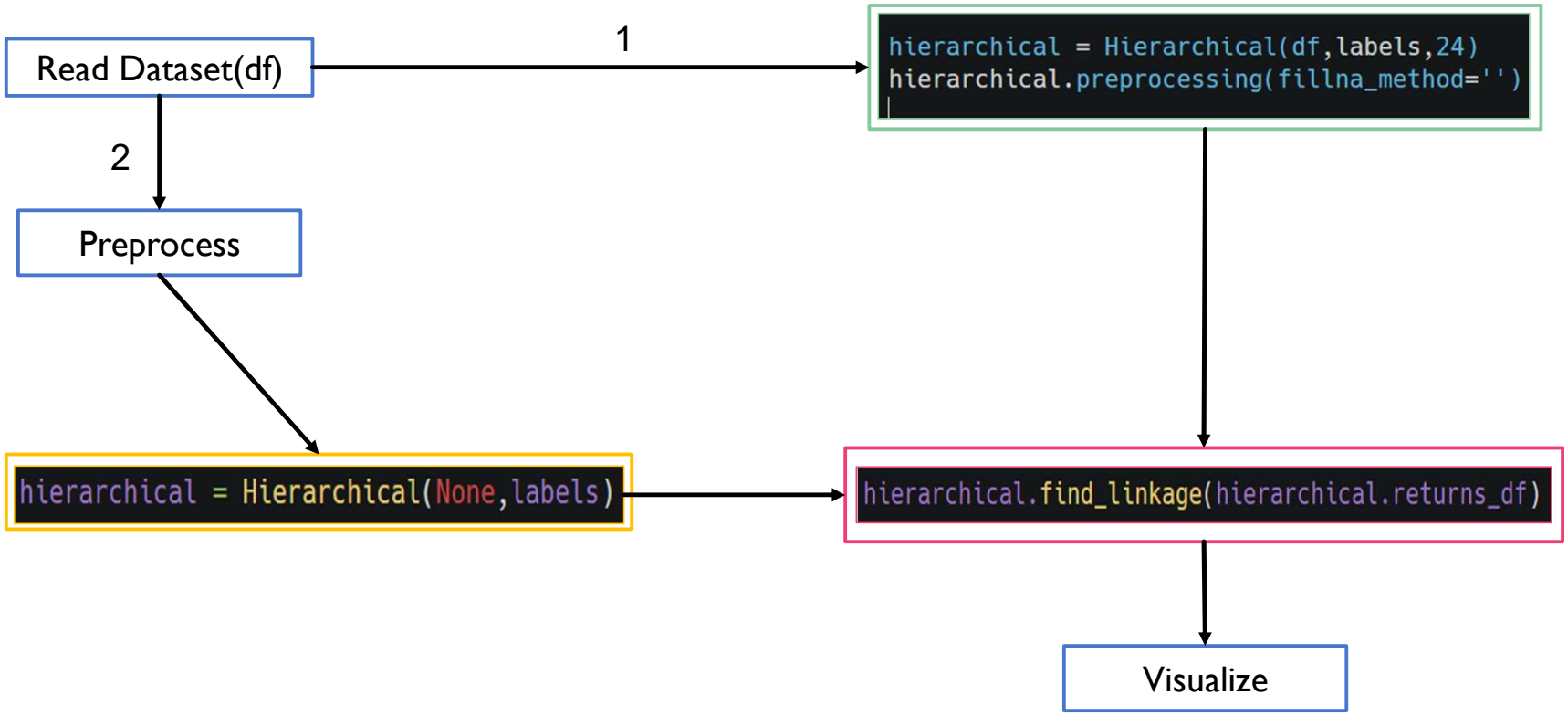
CODE

Hierarchical class methods

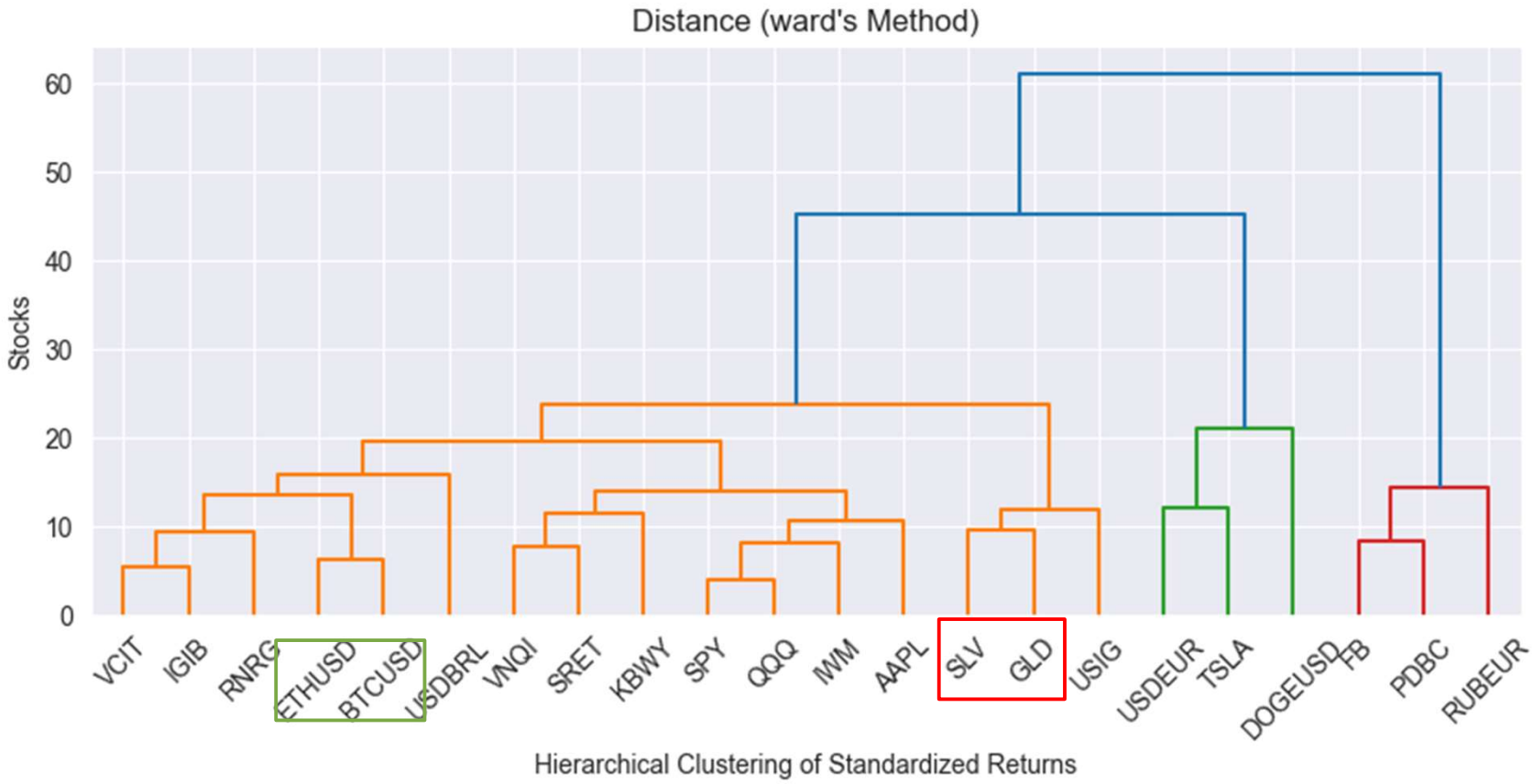
```
1. def __init__(self, dataset: pd.DataFrame, labels: list, interval: int = 24)

2. def preprocessing(self, fillna_method: str = '', resample: str = 'H', inplace: bool = True,
standardize_returns: bool = True, standardize_risks: bool = True)

3. def apply_sliding_window(self, dataset: pd.DataFrame, interval: int = 24)

4. def standardize_dataset(self, dataset: pd.DataFrame) -> pd.DataFrame

5. def find_linkage(self, df: pd.DataFrame, plot_dendogram: bool = True, method: str = 'ward', x_label: str =
'Returns')

6. def plot_stocks_timeline(self, dataset: pd.DataFrame = None, y_label: str = 'returns', show_legends: bool =
True)

7. def plot_stocks_avg_risk(self)

8. def plot_stocks_avg_return(self)

9. def plot_stocks_risk_vs_return(self, plot_sharpe_ratio: bool = False, risk_tolerance: float = 0.02,
expected_return: float = 0.1, risk_free_rate: float = 0.02)
```

CODE

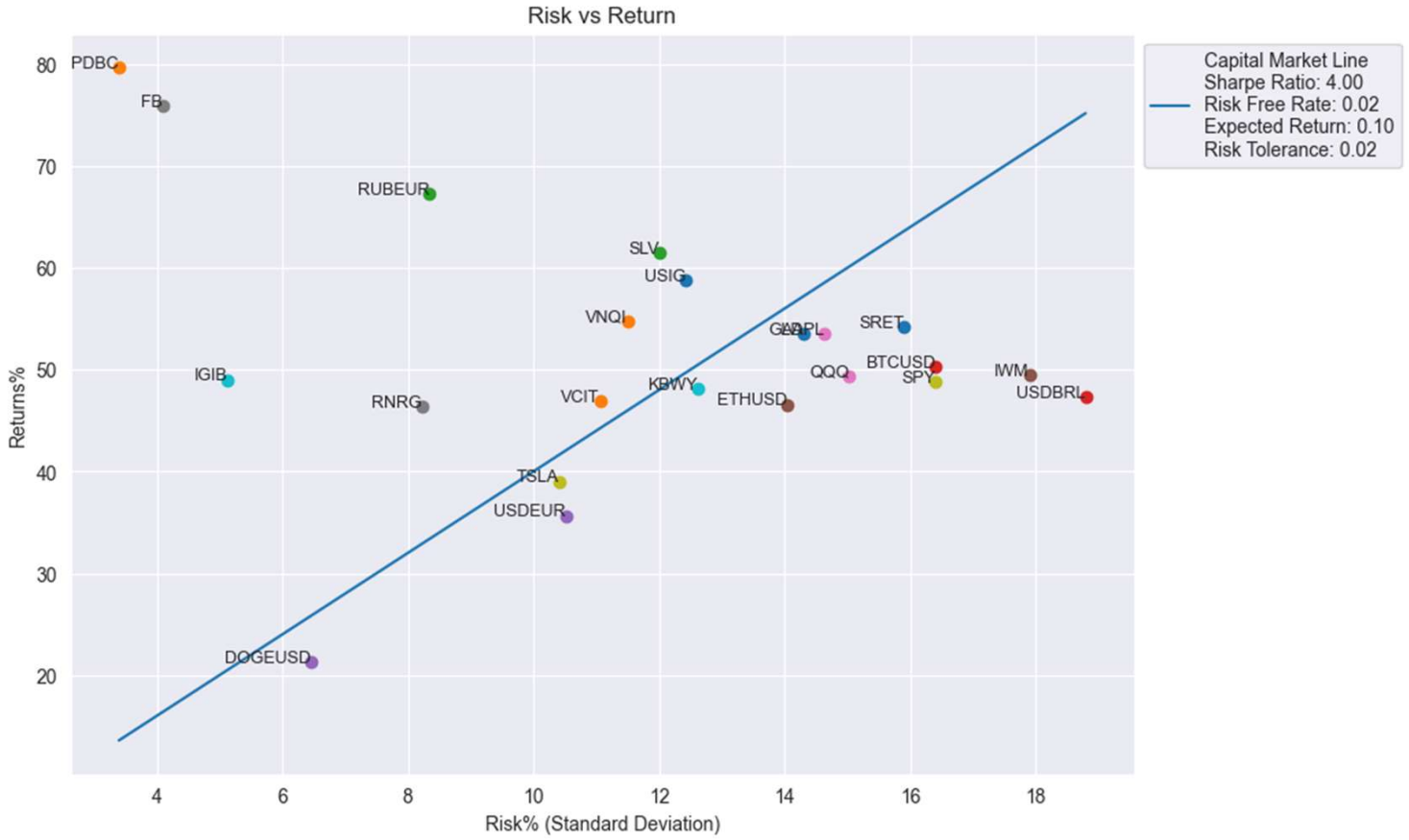
Pipeline



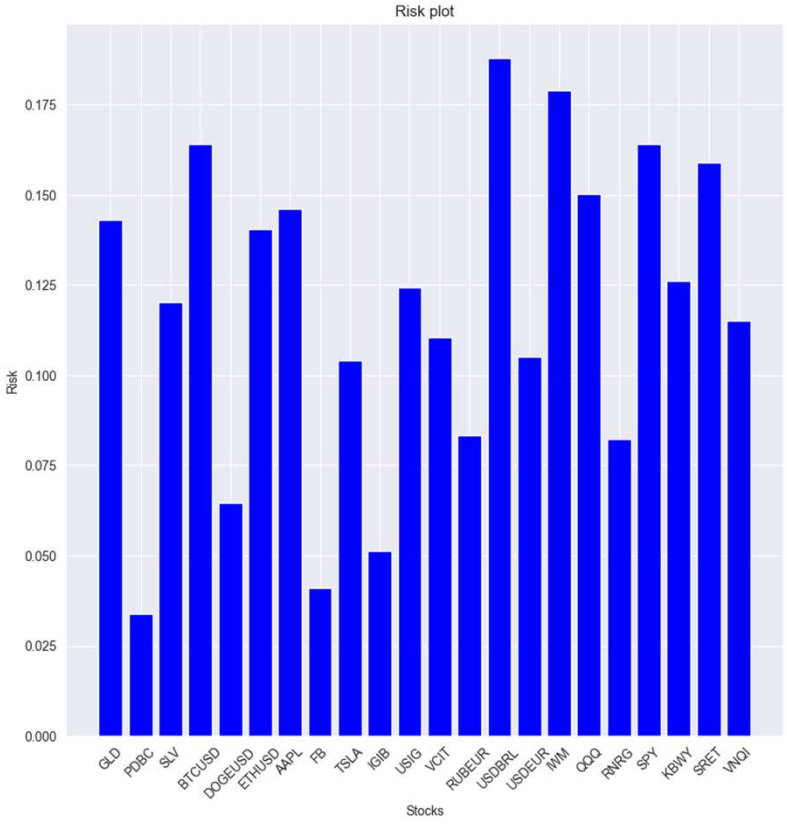
RESULT

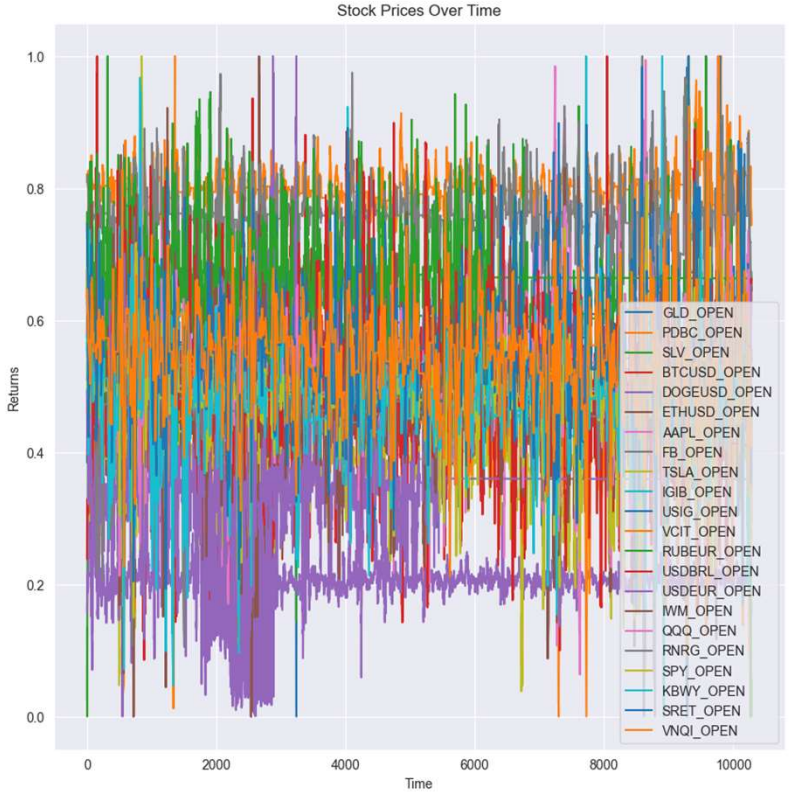

Distance (ward's Method)
Stocks
60
50
40
30
20
10
0
VCIT
IGIB
RNRG
ETHUSD
BTCUSD
USDBRL
VNQI
SRET
KBWY
SPY
QQQ
IWM
AAPL
SLV
GLD
USIG
USDEUR
TSLA
DOGEUSD
FB
PDBC
RUBEUR
Hierarchical Clustering of Standardized Returns

RESULT


Risk vs Return
Returns%
Risk% (Standard Deviation)
Capital Market Line
Sharpe Ratio: 4.00
Risk Free Rate: 0.02
Expected Return: 0.10
Risk Tolerance: 0.02
PDBC
FB
RUBEUR
SLV
USIG
VNQI
SRET
IGIB
BTCUSD
QQQ
SPY
IWM
USDBRL
RNRG
VCIT
ETHUSD
TSLA
USDEUR
DOGEUSD

RESULT


Risk plot
Risk
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
GLD
PDBC
SLV
BTCUSD
DOGEUSD
ETHUSD
AAPL
FB
TSLA
IGIB
USIG
VCIT
RUBEUR
USDBRL
USDEUR
IWM
QQQ
RNRG
SPY
KBWY
SRET
VNQI
Stocks


Stock Prices Over Time
Returns
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Time
GLD_OPEN
PDBC_OPEN
SLV_OPEN
BTCUSD_OPEN
DOGEUSD_OPEN
ETHUSD_OPEN
AAPL_OPEN
FB_OPEN
TSLA_OPEN
IGIB_OPEN
USIG_OPEN
VCIT_OPEN
RUBEUR_OPEN
USDBRL_OPEN
USDEUR_OPEN
IWM_OPEN
QQQ_OPEN
RNRG_OPEN
SPY_OPEN
KBWY_OPEN
SRET_OPEN
VNQI_OPEN

Questions

THANK YOU FOR YOUR ATTENTION