ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«StockLens: Πλατφόρμα Ανάλυσης Χρηματιστηριακής Αγοράς Μέσω Μηχανικής Μάθησης»

Του φοιτητή
Πιρίδης Ιωάννη
Αρ. Μητρώου: 123937

Επιβλέπων
Στέφανος Ουγιάρογλου
Επίκουρος Καθηγητής

Ημερομηνία 30-05-2026

Τίτλος Δ.Ε.: StockLens: Πλατφόρμα Ανάλυσης Χρηματιστηριακής Αγοράς Μέσω Μηχανικής Μάθησης

Κωδικός Δ.Ε. 25206

Ονοματεπώνυμο φοιτητή: Πιρίδης Ιωάννης

Ονοματεπώνυμο εισηγητή Ουγιάρογλου Στέφανος

Ημερομηνία ανάληψης Δ.Ε. 24-03-2025

Ημερομηνία περάτωσης Δ.Ε. 30-05-2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

«Στους γονείς μου»

Πρόλογος

Η παρούσα εργασία εκπονήθηκε στο πλαίσιο των σπουδών μου στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος. Η ιδέα γεννήθηκε από έναν συνδυασμό προσωπικού ενδιαφέροντος για τις χρηματοοικονομικές αγορές και της παρατήρησης ότι, παρά τον τεράστιο όγκο δημόσια διαθέσιμων δεδομένων (SEC EDGAR, APIs χρηματιστηρίου), ελάχιστα εργαλεία τα συνδυάζουν με αλγορίθμους μηχανικής μάθησης σε εύχρηστη μορφή. Τα υπάρχοντα εργαλεία είτε απαιτούν εξειδικευμένες γνώσεις, είτε είναι εμπορικά και κλειστά.

Η ανάπτυξη του StockLens ήταν εκπαιδευτική εμπειρία. Χρειάστηκε να αντιμετωπιστούν πρακτικά προβλήματα στη συλλογή δεδομένων, τον σχεδιασμό της βάσης, την υλοποίηση των αλγορίθμων και το frontend. Ήταν επίσης ευκαιρία να εφαρμοστούν γνώσεις από τις σπουδές σε ένα πραγματικό πρόβλημα και να εξερευνηθούν τεχνολογίες που δεν καλύπτονται εκτενώς στο πρόγραμμα σπουδών, όπως η TimescaleDB και οι αλγόριθμοι μείωσης διαστάσεων.

Η εργασία αυτή ολοκλήρωσε ένα πλήρες σύστημα: από τη συλλογή δεδομένων από δύο επίσημες πηγές, μέχρι την ανάλυση με αλγορίθμους μηχανικής μάθησης και την παρουσίαση αποτελεσμάτων σε διαδραστικό dashboard. Το αποτέλεσμα είναι ένα εργαλείο που μπορεί να χρησιμοποιηθεί άμεσα για εξερευνητική ανάλυση μετοχών, χωρίς να απαιτεί προγραμματιστικές γνώσεις από τον χρήστη.

Περίληψη

Η παρούσα πτυχιακή εργασία αφορά τον σχεδιασμό και την υλοποίηση της εφαρμογής «StockLens», ενός συστήματος ανάλυσης χρηματιστηριακών δεδομένων μέσω τεχνικών μηχανικής μάθησης. Κεντρικό ερώτημα είναι αν το unsupervised clustering βάσει θεμελιωδών οικονομικών μεγεθών μπορεί να αποκαλύψει ομοιότητες μεταξύ εταιρειών χωρίς προκαθορισμένες ετικέτες.

Για τη συλλογή δεδομένων χρησιμοποιούνται δύο πηγές: η Interactive Brokers (ιστορικά OHLCV δεδομένα μέσω TWS API) και η SEC EDGAR (θεμελιώδη χρηματοοικονομικά στοιχεία σε XBRL μορφή). Τα δεδομένα ενοποιούνται βάσει τριμηνιαίων περιόδων αναφοράς και αποθηκεύονται σε PostgreSQL με TimescaleDB. Εφαρμόζεται pipeline κανονικοποίησης με signed log transformation και z-score για ομοιομορφία των δεδομένων μεταξύ 12 features.

Υλοποιούνται τέσσερις αλγόριθμοι clustering: K-Means, Agglomerative Hierarchical, DBSCAN και Gaussian Mixture Models (GMM), καθώς και ο Isolation Forest για εντοπισμό ανωμαλιών. Για οπτικοποίηση εφαρμόζονται PCA, UMAP και t-SNE. Το interactive frontend αναπτύχθηκε με Dash και επιτρέπει παραμετροποίηση και εξερεύνηση αποτελεσμάτων.

Τα πειράματα έδειξαν ότι ο K-Means έβγαλε τα πιο ερμηνεύσιμα αποτελέσματα. Ο αλγόριθμος εντόπισε τις τράπεζες σαν ξεχωριστή ομάδα, μόνο από τη δομή των assets και των cash flows, χωρίς καμία πληροφορία για τον κλάδο. Το σύστημα μπορεί επίσης να εντοπίσει εταιρείες με ασυνήθιστο οικονομικό προφίλ. Τέλος, η ανάλυση σε διαφορετικές χρονικές περιόδους (πριν, κατά και μετά την πανδημία) έδειξε ότι τα οικονομικά προφίλ αλλάζουν με τον χρόνο, με πιο χαρακτηριστική τη μετακίνηση της Nvidia από τις μικρές, ανερχόμενες εταιρείες προς τους κολοσσούς της τεχνολογίας.

«Web Application for the Visualization, Analysis and Clustering of Stock Market Data»

Ioannis Piridis

Abstract

This thesis presents the design and implementation of «StockLens», an interactive web application for stock market analysis using unsupervised machine learning. The central research question is whether clustering on fundamental financial metrics can reveal economically meaningful groupings without any predefined labels.

Data is collected from two sources: Interactive Brokers (historical OHLCV data via the TWS socket API) and SEC EDGAR (quarterly fundamental metrics in XBRL format). The data is merged on quarterly reporting periods, stored in PostgreSQL with TimescaleDB, and normalized through a signed log transformation followed by z-score scaling across 12 features.

Four clustering algorithms are implemented: K-Means, Agglomerative Hierarchical, DBSCAN and Gaussian Mixture Models, alongside Isolation Forest for anomaly detection. Three dimensionality reduction methods (PCA, UMAP, t-SNE) serve visualization. The interactive Dash frontend provides full parameter control and real-time result exploration.

The experiments showed that K-Means gave the most interpretable results. The algorithm identified the banks as a separate group purely from the structure of their assets and cash flows, without any sector information. The system can also flag companies with an unusual financial profile. Finally, the analysis across different time periods (before, during and after the pandemic) showed that the financial profiles of companies change over time, the clearest case being Nvidia moving from the small, rising companies toward the large technology names.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντά μου, Στέφανο Ουγιάρογλου, για την καθοδήγηση και την υποστήριξή του καθ' όλη τη διάρκεια της εργασίας, αλλά και για την υπομονή και την κατανόηση που έδειξε. Η ενθάρρυνσή του ήταν πολύτιμη, ιδιαίτερα στις στιγμές που η κατεύθυνση δεν ήταν ξεκάθαρη, ενώ η προθυμία του να βοηθήσει κάθε φορά που το χρειάστηκα έκανε αυτή την πορεία σημαντικά πιο εύκολη.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες

AIC: Akaike Information Criterion

API: Application Programming Interface

BIC: Bayesian Information Criterion

CSV: Comma-Separated Values

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

EM: Expectation-Maximization

EPS: Earnings Per Share, Κέρδη ανά Μετοχή

GMM: Gaussian Mixture Models

IBKR: Interactive Brokers

KNN: k-Nearest Neighbors

ML: Machine Learning, Μηχανική Μάθηση

OHLCV: Open, High, Low, Close, Volume (τύπος δεδομένων τιμών μετοχών)

PCA: Principal Component Analysis, Ανάλυση Κύριων Συνιστωσών

REST: Representational State Transfer

SEC: Securities and Exchange Commission, Επιτροπή Κεφαλαιαγοράς ΗΠΑ

SQL: Structured Query Language

t-SNE: t-Distributed Stochastic Neighbor Embedding

TWS: Trader Workstation (IBKR trading platform)

UMAP: Uniform Manifold Approximation and Projection

WCSS: Within-Cluster Sum of Squares

XBRL: eXtensible Business Reporting Language

Κεφάλαιο 1ο: Εισαγωγή

1.1 Εισαγωγή

Η παρούσα εργασία ασχολείται με το ερώτημα αν μπορούν αλγόριθμοι μηχανικής μάθησης να ανακαλύψουν διάφορες δομές της χρηματιστηριακής αγοράς αξιοποιώντας οικονομικά στοιχεία των εταιριών, χωρίς καμία προκαθορισμένη κλαδική πληροφορία. Για το λόγο αυτό αναπτύχθηκε το StockLens, μια διαδραστική πλατφόρμα ανάλυσης μετοχών που συνδυάζει δεδομένα από δύο πηγές, εφαρμόζει αλγορίθμους ομαδοποίησης και παρουσιάζει αποτελέσματα μέσω διαδραστικού dashboard.

Στην παρούσα ενότητα παρουσιάζεται το θεωρητικό υπόβαθρο της εργασίας, τα ερευνητικά ερωτήματα και η δομή των επόμενων κεφαλαίων.

1.2 Μετοχές και Χρηματιστήριο

Οι μετοχές είναι ένα από τα πιο βασικά επενδυτικά εργαλεία στις σύγχρονες χρηματοοικονομικές αγορές. Το χρηματιστήριο είναι ο κεντρικός τόπος όπου επενδυτές αγοράζουν και πωλούν μερίδια σε δημόσιες εταιρείες [1]. Η τιμή μιας μετοχής επηρεάζεται από πολλούς παράγοντες όπως τα οικονομικά αποτελέσματα της εταιρείας, τις μακροοικονομικές και γεωπολιτικές συνθήκες, αλλά και την ψυχολογία των επενδυτών [2]. Η κατανόηση αυτών των κινήσεων είναι σημαντική για τεκμηριωμένες επενδυτικές αποφάσεις.

Οι χρηματοοικονομικές αγορές διακρίνονται σε δύο κατηγορίες: την πρωτογενή αγορά (primary market), όπου εταιρείες εκδίδουν νέες μετοχές μέσω IPO, και τη δευτερογενή αγορά (secondary market), όπου διαπραγματεύονται υπάρχουσες μετοχές. Οι κύριοι δείκτες, όπως ο S&P 500 (500 μεγαλύτερες αμερικανικές εταιρείες), ο Dow Jones Industrial Average και ο NASDAQ Composite, χρησιμεύουν ως benchmark για τη μέτρηση της γενικής απόδοσης της αγοράς.

Η fundamental analysis εξετάζει την εσωτερική αξία μιας εταιρείας μέσα από τις οικονομικές της καταστάσεις: balance sheet, income statement, cash flow statement κλπ. Σε αντίθεση με την technical analysis που μελετά patterns τιμών [4], η fundamental analysis ισχυρίζεται ότι η αγοραία τιμή τελικά συγκλίνει προς την πραγματική αξία (intrinsic value) της εταιρείας [1]. Αυτή η λογική είναι η βάση των αλγορίθμων clustering που υλοποιεί η εφαρμογή.

1.3 Ανάλυση δεδομένων μετοχών (Ανάγκη)

Η ποσότητα δεδομένων που παράγεται καθημερινά στις χρηματοπιστωτικές αγορές κάνει την παραδοσιακή, χειροκίνητη ανάλυση πρακτικά αδύνατη. Υπάρχει ανάγκη για αυτοματοποιημένα συστήματα που συλλέγουν, επεξεργάζονται και αναλύουν δεδομένα από διαφορετικές πηγές [3]. Η ανάλυση αυτή επεκτείνεται και σε fundamental data, επιτρέποντας τον εντοπισμό patterns που δεν φαίνονται με μια πρώτη ματιά [4]. Ο αριθμός εισηγμένων εταιρειών στις αμερικανικές αγορές ξεπερνά τις 4.000, με καθεμία να αναφέρει τριμηνιαία και ετήσια οικονομικά στοιχεία. Αυτό παράγει εκατοντάδες χιλιάδες χρηματοοικονομικές εγγραφές ανά έτος, πολύ περισσότερες από ό,τι ένας αναλυτής μπορεί να επεξεργαστεί χειροκίνητα.

Η αδυναμία της χειροκίνητης ανάλυσης δεν είναι μόνο θέμα όγκου, αλλά και συστηματικής σύγκρισης. Ένας αναλυτής που εξετάζει μεμονωμένα 500 εταιρείες δεν μπορεί εύκολα να εντοπίσει ποιες έχουν παρόμοιο οικονομικό προφίλ, ειδικά όταν ανήκουν σε διαφορετικούς κλάδους. Τα εργαλεία ML, όπως το unsupervised clustering, κάνουν αυτή τη σύγκριση συστηματικά και αντικειμενικά.

1.4 Κίνητρο και Συνεισφορά

Το κίνητρο για αυτή την εργασία ξεκινάει από το χάσμα μεταξύ του τεράστιου όγκου ακατέργαστων χρηματοοικονομικών δεδομένων και της δυσκολίας να εξαχθούν χρήσιμα συμπεράσματα από αυτά. Τα κεντρικά ερευνητικά ερωτήματα είναι:

- Μπορούν τα οικονομικά στοιχεία να ομαδοποιήσουν εταιρείες καλύτερα από τις ήδη υπάρχουσες κατηγορίες του κλάδου;
- Ποιος αλγόριθμος clustering (K-Means, Agglomerative, GMM, DBSCAN) αποδίδει καλύτερα για χρηματοοικονομικά fundamentals;
- Ποια μέθοδος μείωσης διαστάσεων (PCA, UMAP, t-SNE) δείχνει καλύτερα τα αποτελέσματα του clustering;
- Μπορεί η παρακολούθηση των διάφορων κλάδων μέσα στο χρόνο να συνεισφέρει σε μελλοντικές επενδυτικές αποφάσεις;

Η συνεισφορά είναι η υλοποίηση του StockLens, ενός ολοκληρωμένου συστήματος που απαντά αυτά τα ερωτήματα μέσω πραγματικών δεδομένων. Το σύστημα ενσωματώνει δεδομένα από Interactive Brokers (OHLCV) και SEC EDGAR (XBRL fundamentals), εφαρμόζει clustering και μείωση διαστάσεων, και παρέχει interactive dashboard για εξερεύνηση αποτελεσμάτων. Το StockLens έχει εξερευνητικό χαρακτήρα: ο αναλυτής μπορεί διαδραστικά να αλλάξει παραμέτρους και να παρατηρήσει την επίδρασή τους στα αποτελέσματα.

1.5 Οργάνωση της εργασίας

Η εργασία χωρίζεται σε οκτώ κεφάλαια. Το Κεφάλαιο 2 περιγράφει τις πηγές δεδομένων (APIs) που χρησιμοποιήθηκαν, τη διαδικασία δημιουργίας του dataset και την αποθήκευσή του. Το Κεφάλαιο 3 εξετάζει τους αλγορίθμους clustering. Το Κεφάλαιο 4 αναλύει τη μείωση διαστάσεων για οπτικοποίηση. Το Κεφάλαιο 5 καλύπτει τεχνολογίες, αρχιτεκτονική και υλοποίηση. Το Κεφάλαιο 6 παρουσιάζει την εφαρμογή. Το Κεφάλαιο 7 παρουσιάζει τα αποτελέσματα τριών πειραμάτων σε πραγματικά δεδομένα. Το Κεφάλαιο 8 συνοψίζει τα συμπεράσματα και τις μελλοντικές επεκτάσεις.

Επίλογος

Στο κεφάλαιο αυτό παρουσιάστηκαν τα βασικά κίνητρα της εργασίας. Στο επόμενο κεφάλαιο περιγράφονται οι πηγές δεδομένων και η διαδικασία δημιουργίας του συνόλου δεδομένων.

Κεφάλαιο 2ο: Δεδομένα μετοχών

2.1 Εισαγωγή

Το κεφάλαιο αυτό περιγράφει πώς χτίστηκε το dataset του StockLens. Χρησιμοποιούνται δύο πηγές: το Interactive Brokers API για ιστορικά OHLCV δεδομένα και η SEC EDGAR για τριμηνιαία fundamentals. Παρουσιάζεται η διαδικασία ενοποίησής τους, η αντιμετώπιση ελλιπών τιμών και η αποθήκευσή τους στη βάση δεδομένων.

2.2 Interactive Brokers API

Η Interactive Brokers (IBKR) επιλέχθηκε ως πηγή για τα price data επειδή παρέχει αξιόπιστα και ιστορικά OHLCV δεδομένα για εκατοντάδες μετοχές. Η σύνδεση γίνεται μέσω του Trader Workstation (TWS) API, ένα socket-based πρωτόκολλο που απαιτεί εκτέλεση του IB Gateway στον τοπικό υπολογιστή.

Η βιβλιοθήκη ibapi παρέχει δύο βασικές κλάσεις:

- EClient: Αποστέλλει requests στον TWS server (reqHistoricalData(), reqContractDetails() κ.α.).
- EWrapper: Λαμβάνει τα αποτελέσματα μέσω callbacks (historicalData(), historicalDataEnd(), error() κ.α.).

Ο custom wrapper που αναπτύχθηκε (backend/ibkr/wrapper.py) κληρονομεί και από τις δύο κλάσεις. Κάθε αίτηση δεδομένων λαμβάνει μοναδικό reqId ώστε ο callback να γνωρίζει σε ποια μετοχή αντιστοιχεί η απάντηση. Όταν ο TWS στείλει historicalDataEnd για ένα reqId, ο wrapper σηματοδοτεί ότι τα δεδομένα αυτής της μετοχής είναι πλήρη.

Ασύγχρονη αρχιτεκτονική: Ο TWS API λειτουργεί asynchronously. Αντί για blocking αναμονή, η εφαρμογή χρησιμοποιεί asyncio.Event() για κάθε reqId. Το κύριο coroutine αναμένει το event, το οποίο «σηκώνεται» από τον EWrapper callback. Έτσι μπορούν να εκτελούνται πολλαπλά requests παράλληλα χωρίς thread blocking.

Rate limiting: Ο IBKR επιτρέπει 50 requests ανά δευτερόλεπτο. Για τη λήψη δεδομένων 127 μετοχών, εισάγεται 10s delay μεταξύ requests για αποφυγή throttling. Κάθε request ανακτά ένα έτος ιστορικών εβδομαδιαίων bars. Τα δεδομένα αποθηκεύονται στον πίνακα stock_prices με upsert βάσει (symbol, timestamp) για αποφυγή διπλότυπων.

2.3 SEC EDGAR API

Η SEC EDGAR (Electronic Data Gathering, Analysis, and Retrieval) είναι η επίσημη βάση δεδομένων της Securities and Exchange Commission των ΗΠΑ. Κάθε δημόσια εταιρεία υποχρεούται να υποβάλλει οικονομικές εκθέσεις τριμηνιαία (Form 10-Q) και ετήσια (Form 10-K). Η EDGAR παρέχει ελεύθερη πρόσβαση μέσω REST API [1].

Η προσέγγιση που χρησιμοποιήθηκε αξιοποιεί το Company Facts API:

Κεφάλαιο 2ο: Δεδομένα μετοχών

- Company Facts: Επιστρέφει όλα τα XBRL-tagged δεδομένα, περιλαμβάνοντας εκατοντάδες financial concepts με ιστορικές τιμές.
- US-GAAP Taxonomy: Τα financial concepts ακολουθούν την US-GAAP ταξινομία (π.χ. «us-gaap/Assets», «us-gaap/NetIncomeLoss»).

Αναπτύχθηκε transformer που απομονώνει 8 βασικά financial concepts:

- us-gaap/Assets: Συνολικά περιουσιακά στοιχεία.
- us-gaap/NetIncomeLoss: Καθαρά κέρδη ή ζημίες.
- us-gaap/EarningsPerShareDiluted: Κέρδη ανά μετοχή (αραιωμένα).
- us-gaap/CashAndCashEquivalentsAtCarryingValue: Μετρητά και ισοδύναμα.
- us-gaap/RetainedEarningsAccumulatedDeficit: Κέρδη εις νέον.
- us-gaap/NetCashProvidedByUsedInOperatingActivities: Ταμειακές ροές λειτουργίας.
- us-gaap/NetCashProvidedByUsedInInvestingActivities: Ταμειακές ροές επενδύσεων.
- us-gaap/NetCashProvidedByUsedInFinancingActivities: Ταμειακές ροές χρηματοδότησης.

2.4 Δημιουργία συνόλου δεδομένων

Η δημιουργία ενός ενιαίου dataset είναι το βασικό βήμα. Τα δύο APIs παρέχουν διαφορετικά δεδομένα: το IBKR δίνει market data (price, volume) ενώ η SEC EDGAR δίνει fundamental data. Η ενοποίησή τους γίνεται βάσει των ημερομηνιών λήξης των οικονομικών εκθέσεων (report end dates). Για κάθε τέτοια ημερομηνία, υπολογίζεται ένα παράθυρο τριμήνου (~90 ημέρες πριν), εντός του οποίου ανακτώνται οι μέσες τιμές (low_avg, high_avg), ο volume_avg και το price_growth. Αυτά συσχετίζονται με τα fundamentals της ίδιας περιόδου.

Η ενοποίηση παρουσιάζει σημαντικές προκλήσεις: οι εκθέσεις της SEC αφορούν ημερολογιακά τρίμηνα, ενώ οι εταιρείες αναφέρουν βάσει οικονομικού έτους που συχνά διαφέρει. Γι αυτό χρησιμοποιείται η ημερομηνία λήξης κάθε έκθεσης (period_of_report) ως σημείο αναφοράς.

Προεπεξεργασία σε δύο βήματα:

- Διαχείριση ελλιπών τιμών: Γραμμική παρεμβολή (linear interpolation) για κενά εντός της χρονοσειράς κάθε μετοχής, με συμπλήρωση τυχόν εναπομεινάντων κενών από τον μέσο όρο του συμβόλου.
- Κανονικοποίηση: Signed log1p transformation ακολουθούμενη από StandardScaler (z-score). Αποτέλεσμα: δεδομένα με mean=0 και std=1.

Η κανονικοποίηση εφαρμόζεται σε όλες τις μετοχές μαζί, οπότε η τιμή 0 αντιστοιχεί στον μέσο όρο όλων των εταιρειών για το feature και +1 σημαίνει μία τυπική απόκλιση πάνω. Με την κοινή κλίμακα μπορούμε να κάνουμε ποσοτικές συγκρίσεις ανεξάρτητα από τις απόλυτες τιμές.

2.5 Αποθήκευση δεδομένων σε βάση δεδομένων

Για τη διαχείριση του όγκου δεδομένων χρησιμοποιείται PostgreSQL με την επέκταση TimescaleDB. Η επιλογή έναντι NoSQL έγινε γιατί η δομή των δεδομένων είναι καλά ορισμένη, τα time-range SQL queries είναι απλά και αποδοτικά, και η SQL επιτρέπει σύνθετα JOINs για την ενοποίηση των δύο πηγών. Η TimescaleDB χρησιμοποιείται επειδή ειδικεύεται στα time-series δεδομένα.

TimescaleDB Hypertables: Ένα hypertable χωρίζεται αυτόματα σε «chunks» βάσει χρόνου. Πλεονεκτήματα:

- Γρήγορα time-range queries: Σαρώνει μόνο τα σχετικά chunks.
- Αυτόματο partitioning χωρίς manual maintenance.
- Compression παλαιότερων chunks για εξοικονόμηση χώρου.

Το schema αποτελείται από έξι πίνακες:

1. stock_prices: Raw OHLCV δεδομένα ανά σύμβολο και timestamp. Hypertable.
2. financial_metrics: Fundamental metrics από SEC EDGAR ανά σύμβολο και report_date. Hypertable.
3. combined_stock_data: Ενοποίηση τιμών και θεμελιωδών μεγεθών με quarterly aggregates.
4. combined_stock_data_enhanced: Εμπλουτισμένη έκδοση του combined με επιπλέον engineered features.
5. normalized_stock_data: Επεξεργασμένα δεδομένα (signed log1p + z-score) του combined.
6. normalized_stock_data_enhanced: Επεξεργασμένα δεδομένα του enhanced, έτοιμα για ML.

Κάποιες εταιρείες αναθεωρούν προηγούμενες εκθέσεις, δημιουργώντας πολλαπλές εγγραφές για την ίδια περίοδο. Εφαρμόζεται deduplication: διατηρείται μόνο η πιο πρόσφατη εγγραφή βάσει filed date για κάθε end date ανά σύμβολο.

Επίλογος

Η ενοποίηση των δύο APIs δεν ήταν τετριμμένη: οι IBKR εγγραφές είναι ημερήσιες, τα SEC fundamentals τριμηνιαία, και οι εταιρείες δεν ακολουθούν το ίδιο οικονομικό ημερολόγιο. Για να λυθεί αυτό το πρόβλημα χρησιμοποιήθηκε η period_of_report ως σημείο αναφοράς.

Κεφάλαιο 3ο: Ανάλυση συστάδων (Cluster analysis)

3.1 Εισαγωγή

Η ανάλυση συστάδων (cluster analysis) είναι ένα από τα βασικά εργαλεία της μη εποπτευόμενης μηχανικής μάθησης (unsupervised learning). Σκοπός της είναι να ομαδοποιήσει δεδομένα έτσι ώστε τα αντικείμενα της ίδιας ομάδας να είναι πιο παρόμοια μεταξύ τους από ό,τι με εκείνα άλλων ομάδων [3]. Η βασική διαφορά από την εποπτευόμενη μάθηση είναι ότι δεν υπάρχουν προκαθορισμένες ετικέτες: ο αλγόριθμος ανακαλύπτει μόνος του τη δομή των δεδομένων.

Στις χρηματοοικονομικές αγορές, το clustering βοηθά να εντοπιστούν μετοχές με παρόμοια θεμελιώδη χαρακτηριστικά, ανεξάρτητα από τον κλάδο στον οποίο ανήκουν. Αντί για ομαδοποίηση βάσει κλάδου, μπορεί κανείς να ομαδοποιήσει εταιρείες βάσει της πραγματικής τους οικονομικής απόδοσης, της κεφαλαιακής δομής και των cash flows [1].

Στο StockLens υλοποιήθηκαν τέσσερις αλγόριθμοι clustering: K-Means, ιεραρχική συσταδοποίηση, DBSCAN και Gaussian Mixture Models (GMM). Επιπλέον, για εντοπισμό ανωμαλιών (outlier detection) χρησιμοποιείται ο αλγόριθμος Isolation Forest. Κάθε αλγόριθμος έχει διαφορετικές παραδοχές και κατάλληλες χρήσεις, οπότε η παρουσία όλων επιτρέπει στον χρήστη να επιλέξει τον πιο κατάλληλο ανάλογα με τα δεδομένα.

3.2 K-Means

Ο αλγόριθμος χωρίζει τα δεδομένα σε k συστάδες επαναλαμβάνοντας δύο βήματα: ανάθεση κάθε σημείου στην πλησιέστερη συστάδα βάσει ευκλείδειας απόστασης, και επανυπολογισμός των centroids ως μέσων τιμών. Σταματά όταν οι αναθέσεις σταθεροποιηθούν [3].

Το κριτήριο βελτιστοποίησης είναι η ελαχιστοποίηση του WCSS (Within-Cluster Sum of Squares): το άθροισμα τετραγωνικών ευκλείδειων αποστάσεων κάθε σημείου από το centroid της συστάδας του. Όσο μικρότερο το WCSS, τόσο πιο «σφιχτές» είναι οι συστάδες.

K-Means++ Αρχικοποίηση

Ένα βασικό πρόβλημα του κλασικού k-means είναι ότι η τυχαία αρχικοποίηση των centroids μπορεί να οδηγήσει σε τοπικά ελάχιστα. Ο αλγόριθμος k-means++ (Arthur & Vassilvitskii, 2007) λύνει αυτό με έξυπνη αρχικοποίηση: το πρώτο centroid επιλέγεται τυχαία, και κάθε επόμενο επιλέγεται με πιθανότητα ανάλογη της τετραγωνικής απόστασης από το πλησιέστερο ήδη επιλεγμένο centroid. Η scikit-learn χρησιμοποιεί k-means++ ως default.

Επιλογή αριθμού συστάδων: Elbow Method και Silhouette Score

Η πιο διαδεδομένη τεχνική για επιλογή k είναι η μέθοδος του αγκώνα (elbow method): υπολογίζεται το WCSS για διαφορετικές τιμές k (2 έως 10) και σχεδιάζεται το γράφημα. Το σημείο όπου ο ρυθμός μείωσης επιβραδύνεται αισθητά σχηματίζει έναν «αγκώνα» που δείχνει τον καταλληλότερο k [3].

Το silhouette score μετρά πόσο καλά κάθε σημείο «ταιριάζει» στη συστάδα του σε σχέση με τις γειτονικές συστάδες. Παίρνει τιμές από -1 (λανθασμένη ανάθεση) έως +1 (ιδανική ανάθεση). Τιμές άνω του 0.5 θεωρούνται αποδεκτές, ενώ για χρηματοοικονομικά δεδομένα, SC > 0.25 θεωρείται αξιόλογο λόγω της εγγενούς ετερογένειας των εταιρειών [3].

3.3 Ιεραρχική συσταδοποίηση

Το βασικό κίνητρο για ιεραρχική προσέγγιση στα fundamentals ήταν το dendrogram: φαίνεται αμέσως αν η JPMorgan και η Goldman Sachs συγχωνεύονται πριν ή μετά από τεχνολογικές εταιρείες χωρίς να χρειαστεί να ορίσει κανείς k εκ των προτέρων. Χρησιμοποιείται η agglomerative προσέγγιση (από κάτω προς τα πάνω): κάθε εγγραφή ξεκινά ως ξεχωριστή συστάδα και οι πλησιέστερες συγχωνεύονται σταδιακά.

Η επιλογή του τρόπου μέτρησης απόστασης μεταξύ συστάδων (linkage criterion) επηρεάζει σημαντικά το αποτέλεσμα:

- Ward linkage: Ελαχιστοποιεί την αύξηση του WCSS σε κάθε συγχώνευση. Τείνει να δημιουργεί συστάδες παρόμοιου μεγέθους και είναι η προεπιλογή στην εφαρμογή.
- Complete linkage: Χρησιμοποιεί τη μέγιστη απόσταση μεταξύ σημείων δύο συστάδων.
- Average linkage: Χρησιμοποιεί τη μέση απόσταση.

Το βασικό πλεονέκτημα είναι ότι δεν απαιτεί εκ των προτέρων τον αριθμό συστάδων. Ο αναλυτής μπορεί να «κόψει» το dendrogram στο επίπεδο που θέλει. Αυτό το κάνει χρήσιμο για εξερευνητική ανάλυση [3].

3.4 Συσταδοποίηση βάσει πυκνότητας (DBSCAN)

Ο DBSCAN συμπεριλήφθηκε κυρίως γιατί ορίζει αυτόματα outliers: εγγραφές σε αραιές περιοχές του 12-διάστατου χώρου λαμβάνουν ετικέτα «noise» χωρίς να χρειαστεί ξεχωριστό βήμα ανίχνευσης ανωμαλιών. Ο αλγόριθμος ομαδοποιεί σημεία υψηλής πυκνότητας ελέγχοντας δύο παραμέτρους [8]:

- epsilon (ε): Η μέγιστη απόσταση εντός της οποίας δύο σημεία θεωρούνται γείτονες.
- minPts: Ο ελάχιστος αριθμός γειτόνων για να θεωρηθεί ένα σημείο «core point».

Κάθε σημείο ταξινομείται σε μία κατηγορία: core point, border point ή noise point. Τα noise points είναι ακριβώς τα outliers που ενδιαφέρουν στη χρηματοοικονομική ανάλυση.

Curse of Dimensionality και DBSCAN

Σε υψηλοδιάστατους χώρους (όπως οι 12 διαστάσεις του dataset), ο DBSCAN αντιμετωπίζει το curse of dimensionality: καθώς αυξάνεται ο αριθμός διαστάσεων, οι αποστάσεις μεταξύ σημείων γίνονται ολοένα πιο ομοιόμορφες. Αποτέλεσμα: δεν υπάρχει «φυσική» τιμή ε που να διαχωρίζει πυκνές από αραιές περιοχές. Στα 12 features × 127 μετοχές, ο DBSCAN δυσκολεύεται να βρει διακριτές συστάδες.

Επιλογή παραμέτρου ε: k-dist graph

Η επιλογή κατάλληλης τιμής ε γίνεται μέσω του k-distance graph: για κάθε σημείο υπολογίζεται η απόσταση από τον k-ο πλησιέστερο γείτονά του. Τα σημεία ταξινομούνται βάσει αυτής της απόστασης

και σχεδιάζεται το γράφημα. Το σημείο «μέγιστης καμπυλότητας» δίνει καλή εκτίμηση για την τιμή ε [3].

3.5 Gaussian Mixture Models (GMM)

Τα GMM αντιμετωπίζουν τις εγγραφές μετοχών ως δείγματα από μίξη κατανομών Gauss. Αντί να αναθέσουν κάθε εγγραφή αποκλειστικά σε έναν cluster, δίνουν πιθανότητες ανά cluster, π.χ. «αυτό το τρίμηνο της AAPL ανήκει 72% στο C1 και 28% στο C3». Αυτή η soft clustering είναι χρήσιμη για εταιρείες που βρίσκονται στα όρια δύο κλάδων, αλλά η ερμηνεία γίνεται πολυπλοκότερη [3].

Οι παράμετροι κάθε κατανομής εκτιμώνται μέσω του αλγορίθμου Expectation-Maximization (EM) [7]. Για την επιλογή k χρησιμοποιούνται κριτήρια BIC (Bayesian Information Criterion) ή AIC (Akaike Information Criterion). Χαμηλότερες τιμές BIC/AIC αντιστοιχούν σε καλύτερο μοντέλο.

Για εντοπισμό ανωμαλιών, η εφαρμογή χρησιμοποιεί τον Isolation Forest. Ο Isolation Forest κατασκευάζει τυχαία δέντρα απόφασης και μετρά πόσες διαχωρίσεις χρειάζονται για να «απομονωθεί» κάθε σημείο. Ανώμαλα σημεία απομονώνονται με λιγότερες διαχωρίσεις γιατί βρίσκονται μακριά από πυκνές περιοχές. Σε αντίθεση με τον DBSCAN, λειτουργεί καλά σε υψηλοδιάστατους χώρους [3].

Επίλογος

Κανένας από τους τέσσερις αλγόριθμους δεν είναι η σωστή επιλογή για κάθε περίπτωση. Εξαρτάται από τα δεδομένα και από το τι θέλει κανείς να βγάλει. Για τα δεδομένα της εφαρμογής, ο K-Means παρήγαγε τα πιο ερμηνεύσιμα αποτελέσματα, το βλέπουμε στα πειράματα του Κεφαλαίου 7. Ο DBSCAN και το GMM παραμένουν χρήσιμα κομμάτια της εφαρμογής, κυρίως για outlier detection και για περιπτώσεις όπου οι συστάδες δεν έχουν σφαιρικό σχήμα. Στο επόμενο κεφάλαιο βλέπουμε πώς οπτικοποιούμε αυτά τα αποτελέσματα.

Κεφάλαιο 4ο: Οπτικοποίηση δεδομένων

4.1 Εισαγωγή

Το κεφάλαιο αυτό εξετάζει τις τρεις μεθόδους μείωσης διαστάσεων που χρησιμοποιεί το StockLens για την οπτικοποίηση των clustering αποτελεσμάτων. Η οπτικοποίηση είναι εργαλείο ανάλυσης. Βοηθά να δούμε αν τα αποτελέσματα του clustering βγάζουν νόημα και να βγάλουμε συμπεράσματα.

Κάθε μετοχή περιγράφεται από 12 features, δηλαδή ζει σε 12-διάστατο χώρο. Χρειάζονται τεχνικές που «συμπυκνώνουν» αυτή την πληροφορία σε 2D χωρίς να χαθεί η δομή των δεδομένων [3]. Υλοποιήθηκαν τρεις μέθοδοι: το γραμμικό PCA (γρήγορο, ερμηνεύσιμο), το μη γραμμικό UMAP (καλύτερη global δομή) και το t-SNE (αποκαλύπτει λεπτές εσωτερικές δομές).

4.2 PCA (Principal Component Analysis)

Το PCA τρέχει πολύ γρήγορα και δίνει άμεση εικόνα αν υπάρχει δομή στα δεδομένα. Κάθε συνιστώσα είναι γραμμικός συνδυασμός των αρχικών features, οπότε μπορεί κανείς να δει ποια features «τραβούν» κάθε άξονα.

Τεχνικά, δουλεύει ως εξής. Πρώτα τυποποιούνται τα δεδομένα (zero mean, unit variance) γιατί αλλιώς features με μεγάλη κλίμακα όπως το assets θα κυριαρχούσαν στο αποτέλεσμα. Στη συνέχεια κατασκευάζεται ο πίνακας συνδιακύμανσης των features και γίνεται eigendecomposition: τα eigenvectors δίνουν τους νέους άξονες και τα eigenvalues δείχνουν πόση διακύμανση εξηγεί ο καθένας. Η πρώτη PC πιάνει τη μεγαλύτερη διακύμανση, η δεύτερη την επόμενη κ.ο.κ. Στην εφαρμογή χρησιμοποιούμε σταθερά 2 components, ώστε το αποτέλεσμα να οπτικοποιείται σε 2D scatter plot.

Το μειονέκτημα είναι η γραμμικότητα: σύνθετες σχέσεις μεταξύ assets και cash flows χάνονται στην προβολή. Γι αυτό χρησιμοποιείται κυρίως ως εξερευνητικό εργαλείο πριν το UMAP [3].

4.3 UMAP (Uniform Manifold Approximation and Projection)

Το UMAP βασίζεται στη θεωρία πολλαπλοτήτων (manifold theory): η βασική ιδέα είναι ότι τα δεδομένα υψηλών διαστάσεων βρίσκονται στην πραγματικότητα κοντά σε μια δομή χαμηλότερης διάστασης. Ο αλγόριθμος δουλεύει σε δύο φάσεις. Πρώτα κατασκευάζει γράφο k-NN στον αρχικό χώρο των επιλεγμένων features, κάθε σημείο συνδέεται με τους k πιο κοντινούς γείτονές του. Μετά βρίσκει μια 2D αναπαράσταση που διατηρεί τη δομή αυτού του γράφου όσο πιο πιστά γίνεται, μέσω βελτιστοποίησης [9].

Σε σύγκριση με το t-SNE, είναι σημαντικά γρηγορότερο και διατηρεί καλύτερα τις αποστάσεις μεταξύ συστάδων και όχι μόνο τις εσωτερικές γειτονιές. Γι αυτό είναι η συνιστώμενη επιλογή για clustering visualization στο StockLens, αν και η εφαρμογή ξεκινά με PCA ως default.

4.4 t-SNE (t-Distributed Stochastic Neighbor Embedding)

Ο αλγόριθμος μετατρέπει αποστάσεις σε πιθανότητες ομοιότητας και βρίσκει 2D αναπαράσταση που ελαχιστοποιεί την KL-απόκλιση μεταξύ αυτών των κατανομών [6], χρησιμοποιώντας κατανομή Student-t για να αποφύγει τη «συρρίκνωση» κοντινών σημείων. Σε αντίθεση με το UMAP, εστιάζει περισσότερο στις τοπικές γειτονιές, κάτι που το κάνει χρήσιμο αν υπάρχει ανάγκη να φανεί η εσωτερική δομή μεγάλων συστάδων.

Στο StockLens η παράμετρος perplexity τίθεται αυτόματα ως min(30, n-1), ώστε σε μικρά subsets να μην προκαλεί σφάλμα. Ο κύριος συμβιβασμός είναι η ταχύτητα: είναι σημαντικά αργότερο από το UMAP, και οι αποστάσεις μεταξύ διαφορετικών συστάδων δεν είναι ερμηνεύσιμες.

Σύγκριση μεθόδων: Το PCA είναι γρήγορο, deterministic και κρατά τις αποστάσεις γραμμικά, οπότε ταιριάζει στον K-Means και βοηθά να καταλάβει κανείς ποια features «τραβούν» τους άξονες. Το UMAP κρατά καλύτερα την δομή και δουλεύει καλύτερα όταν οι συστάδες ορίζονται από πυκνότητα. Το t-SNE δείχνει την εσωτερική δομή μεγάλων συστάδων, αλλά είναι το πιο αργό. Ποια μέθοδος δείχνει καλύτερα τα αποτελέσματα εξαρτάται από τον αλγόριθμο, κάτι που φάνηκε στα πειράματα του Κεφαλαίου 7.

Επίλογος

Η επιλογή μεθόδου οπτικοποίησης επηρεάζει αυτό που φαίνεται να συμβαίνει. Το ίδιο clustering με PCA και UMAP δείχνει διαφορετική εικόνα, όχι γιατί αλλάζουν τα αποτελέσματα, αλλά γιατί αλλάζει η προβολή. Γι αυτό το StockLens παρέχει και τις τρεις επιλογές, αφήνοντας τον αναλυτή να αποφασίσει ανάλογα με το τι ψάχνει.

Κεφάλαιο 5ο: Ανάπτυξη του StockLens

5.1 Εισαγωγή

Το κεφάλαιο αυτό παρουσιάζει τον τρόπο με τον οποίο αναπτύχθηκε το StockLens. Αρχικά περιγράφονται οι τεχνολογίες και οι λειτουργικές απαιτήσεις, ενώ στη συνέχεια παρουσιάζονται η αρχιτεκτονική, το backend και το frontend της εφαρμογής. Επίσης παρουσιάζεται επίσης το pipeline κανονικοποίησης που ετοιμάζει τα δεδομένα για τους αλγορίθμους.

5.2 Τεχνολογίες

Η επιλογή τεχνολογιών έγινε βάσει συγκεκριμένων κριτηρίων: απόδοση, οικοσύστημα βιβλιοθηκών ML, υποστήριξη ασύγχρονης επικοινωνίας και ευκολία ανάπτυξης dashboard. Οι κύριες τεχνολογίες είναι:

- Python 3.12: Η κύρια γλώσσα, βιομηχανικό πρότυπο για data analysis και machine learning.
- Dash by Plotly: Framework για interactive web dashboards σε Python. Με αυτό φτιάχνεις πλούσιο UI χωρίς JavaScript, με ενσωμάτωση Plotly για τα γραφήματα.
- PostgreSQL 16 με TimescaleDB: Ειδικεύεται στα time-series δεδομένα μέσω hypertables [10]. Τα range queries σε χρονοσειρές γίνονται σημαντικά πιο γρήγορα.
- Scikit-learn: Υλοποιήσεις k-means, DBSCAN, agglomerative clustering, GMM, Isolation Forest και StandardScaler [5].
- UMAP-learn: Βιβλιοθήκη για τον αλγόριθμο UMAP, βελτιστοποιημένη για μεγάλα datasets.
- Asyncpg και Asyncio: Ασύγχρονος driver για PostgreSQL, χρησιμοποιείται στο pipeline εισαγωγής δεδομένων από IBKR.
- Pandas και NumPy: Επεξεργασία δεδομένων και κανονικοποίηση features.
- IBKR API (ibapi): χρησιμοποιείται για τη λήψη price data από Interactive Brokers. Βασική πηγή δεδομένων.
- SEC EDGAR API (μέσω httpx): για τα financial data (assets, earnings κ.α.). Η δεύτερη βασική πηγή δεδομένων.

5.3 Λειτουργικές απαιτήσεις (User Stories)

- Αναζήτηση raw δεδομένων: Ως αναλυτής, θέλω να επιλέξω ένα ticker και να δω τα ιστορικά δεδομένα του (OHLCV και financial metrics) σε πίνακα, ώστε να καταλάβω τι δεδομένα υπάρχουν.
- Εξέταση normalized δεδομένων: Ως αναλυτής, θέλω να δω τα normalized δεδομένα, ώστε να καταλάβω τι «βλέπουν» οι αλγόριθμοι clustering.
- Εκτέλεση clustering: Ως αναλυτής, θέλω να επιλέξω αλγόριθμο, να ρυθμίσω τις παραμέτρους και να εκτελέσω clustering, ώστε να δω τις ομάδες μετοχών σε scatter plot, να εξάγω συμπεράσματα για να κάνω ανάλυση.
- Οπτικοποίηση και σύγκριση: Ως αναλυτής, θέλω να επιλέξω πολλαπλές μετοχές και metrics και να τις συγκρίνω σε time-series γραφήματα, ώστε να εντοπίσω διαφορές στη χρηματοοικονομική

τους συμπεριφορά.

- Εντοπισμός ανωμαλιών: Ως αναλυτής, θέλω να ενεργοποιήσω Isolation Forest μέσα στο clustering, ώστε να εντοπίσω μετοχές με ασυνήθιστο οικονομικό προφίλ.

5.4 Αρχιτεκτονική της εφαρμογής

Η εφαρμογή ακολουθεί 3-tier αρχιτεκτονική με σαφή διαχωρισμό ευθυνών:

- Data Layer: PostgreSQL/TimescaleDB, αποθήκευση raw και επεξεργασμένων δεδομένων.
- Logic/Service Layer: Python services, μετασχηματισμός δεδομένων, αλγόριθμοι clustering, normalization pipeline. Επικοινωνία με βάση μέσω Repository.py.
- Presentation Layer: Dash frontend, διαδραστική διεπαφή που επικοινωνεί με το Logic Layer μέσω callbacks.

Η ροή δεδομένων: τα APIs (IBKR και SEC EDGAR) τροφοδοτούν τη βάση δεδομένων, από όπου ανακτώνται, επεξεργάζονται και αποθηκεύονται ξανά. Το frontend ανακτά τα επεξεργασμένα δεδομένα, εκτελεί clustering και επιστρέφει τα αποτελέσματα μέσω γραφημάτων.

Βάση δεδομένων: Έξι πίνακες που χρησιμοποιεί το frontend:

- stock_prices: Raw OHLCV δεδομένα ανά σύμβολο και timestamp. Hypertable.
- financial_metrics: Fundamental metrics από SEC EDGAR ανά σύμβολο και report_date. Hypertable.
- combined_stock_data: Ενοποίηση τιμών και θεμελιωδών μεγεθών με quarterly aggregates.
- combined_stock_data_enhanced: Εμπλουτισμένη έκδοση του combined με επιπλέον engineered features.
- normalized_stock_data: Επεξεργασμένα δεδομένα (signed log1p + z-score) του combined.
- normalized_stock_data_enhanced: Επεξεργασμένα δεδομένα του enhanced, έτοιμα για ML.

Όλοι οι πίνακες υλοποιούνται ως hypertables του TimescaleDB χωρισμένοι βάσει των timestamps.

5.5 Υλοποίηση Backend

Repository Pattern

Για να απομονωθεί η λογική πρόσβασης στη βάση από την υπόλοιπη εφαρμογή, χρησιμοποιήθηκε το Repository Pattern. Το module repository.py περιέχει συναρτήσεις που αντιστοιχούν σε συγκεκριμένες λειτουργίες της βάσης δεδομένων. Έτσι ο business κώδικας δεν περιέχει SQL: αν χρειαστεί αλλαγή στο schema, τροποποιείται μόνο ένα αρχείο.

Ασύγχρονη αρχιτεκτονική

Η επικοινωνία με τη βάση γίνεται ασύγχρονα μέσω asyncpg και asyncio. Κατά την εκκίνηση, τα δεδομένα φορτώνονται async στη μνήμη. Αν η φόρτωση ήταν σύγχρονη (blocking), το Dash UI θα

πάγωνε κατά την εκκίνηση. Για τη σύνδεση Dash με asyncio χρησιμοποιείται η βιβλιοθήκη nest_asyncio.

Pipeline κανονικοποίησης

Πριν τα δεδομένα μπουν στους clustering αλγορίθμους, εκτελείται pipeline κανονικοποίησης σε δύο βήματα:

Πρώτο βήμα: signed logarithmic transformation: για κάθε feature x, υπολογίζεται sign(x) * log(1 + |x|). Αυτό συμπιέζει τις ακραίες τιμές χωρίς να χαθεί πληροφορία για αρνητικές τιμές (π.χ. αρνητικά καθαρά κέρδη).

Δεύτερο βήμα: StandardScaler σε mean=0, std=1, ώστε όλα τα features να έχουν ίση βαρύτητα.

Clustering στο Frontend

Ο κώδικας clustering βρίσκεται εντός των Dash callbacks στο frontend/cluster_tab.py. Δύο βασικά callbacks χειρίζονται την ανάλυση: το evaluation callback εκτελεί το k sweep k=2 έως 10 υπολογίζοντας metrics για κάθε k, και το generate callback εκτελεί το τελικό clustering, εφαρμόζει μείωση διαστάσεων και παράγει το scatter plot, το parallel coordinates chart και τον πίνακα σύνοψης συστάδων. Η αρχιτεκτονική αυτή έχει το πλεονέκτημα ότι το clustering τρέχει άμεσα στη μνήμη χωρίς network overhead μεταξύ backend και frontend.

5.6 Υλοποίηση Frontend

Αρχιτεκτονική Dash Callbacks

Το frontend χωρίζεται σε τέσσερις tabs, τα οποία μπορούν να χρησιμοποιηθούν και για ανάλυση:

- Raw Data Search: Αναζήτηση και προβολή raw δεδομένων με Ag-Grid, interactive filtering/sorting.
- Normalized Data: Προβολή του ML-ready dataset (αφού έχει γίνει log transformation και z-score scaling).
- Cluster Analysis: Ο πυρήνας της εφαρμογής. Επιλογή αλγορίθμου, scatter plot αποτελεσμάτων, parallel coordinates διάγραμμα.
- Visualization & Comparison: Time-series γραφήματα για σύγκριση μετοχών ανά metric.

Κάθε tab χρησιμοποιεί Dash callbacks που τρέχουν κάθε φορά που ο χρήστης αλλάζει κάποια επιλογή. Κάθε φορά που τρέχουν, ξαναυπολογίζουν το αποτέλεσμα από την αρχή. Για να μοιράζονται δεδομένα μεταξύ τους, χρησιμοποιείται το dcc.Store, που αποθηκεύει δεδομένα στον browser του χρήστη.

Cluster Analysis Tab

Το cluster tab (frontend/cluster_tab.py) χρησιμοποιεί multiple callbacks:

- Algorithm params callback: Εμφανίζει τις σχετικές παραμέτρους για τον επιλεγμένο αλγόριθμο.
- Evaluation callback: Εκτελεί το k sweep και παράγει διαγνωστικά γραφήματα.
- Generate callback: Εκτελεί clustering, εφαρμόζει dimensionality reduction και παράγει scatter plot.
- Anomaly overlay: Αν είναι ενεργοποιημένο το anomaly toggle, ο Isolation Forest εκτελείται και τα αποτελέσματα ενσωματώνονται στο scatter plot.

Το dcc.Store component αποθηκεύει τα αποτελέσματα clustering client-side, επιτρέποντας σε άλλα callbacks (π.χ. πίνακας μελών, εξαγωγή CSV) να έχουν πρόσβαση στα labels χωρίς επανεκτέλεση.

Dimensionality Reduction

Η μείωση διαστάσεων εκτελείται εντός του Dash callback, αμέσως μετά το clustering και πριν την απεικόνιση:

- PCA: sklearn.decomposition.PCA(n_components=2). Γρήγορο και deterministic.
- UMAP: umap.UMAP(n_components=2, random_state=42). Fixed random_state για αναπαραγωγιμότητα, υπόλοιπες παράμετροι στις default τιμές της βιβλιοθήκης.
- t-SNE: sklearn.manifold.TSNE(n_components=2, random_state=42, perplexity=min(30, n-1)). Το perplexity ρυθμίζεται αυτόματα ανάλογα με τον αριθμό των μετοχών, ώστε να μην προκαλεί σφάλμα σε μικρά subsets. Είναι ο πιο αργός από τους τρεις αλγορίθμους, αλλά για 127 μετοχές ο χρόνος παραμένει αποδεκτός.

Επίλογος

Καλύψαμε όλη την τεχνική πλευρά, από τις τεχνολογίες και τη βάση δεδομένων μέχρι το normalization pipeline και τα Dash callbacks. Το επόμενο κεφάλαιο δείχνει την εφαρμογή σε λειτουργία, με screenshots από κάθε tab.

Κεφάλαιο 6ο: Παρουσίαση του StockLens

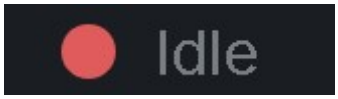
6.1 Εισαγωγή

Το κεφάλαιο αυτό παρουσιάζει τις τέσσερις καρτέλες του StockLens με screenshots από την εφαρμογή σε λειτουργία. Κάθε καρτέλα καλύπτει ένα στάδιο της ανάλυσης, από την εξερεύνηση raw δεδομένων μέχρι την προβολή των αποτελεσμάτων του clustering.

Στην κορυφή της εφαρμογής εμφανίζεται πάντα μια σειρά στατιστικών στοιχείων για τη βάση δεδομένων: ο αριθμός μετοχών, το εύρος ημερομηνιών, ο αριθμός των εγγραφών και ο αριθμός features. Ενημερώνονται αυτόματα κατά την εκκίνηση.

Σχήμα 6.1: Η επικεφαλίδα του StockLens με τα στατιστικά στοιχεία της βάσης δεδομένων (127 μετοχές, 68.461 εγγραφές τιμών, 12 features).

Στην επάνω δεξιά γωνία της εφαρμογής εμφανίζεται μια ένδειξη κατάστασης που δείχνει την τρέχουσα κατάσταση του pipeline δεδομένων. Η ένδειξη αλλάζει αυτόματα ανάλογα με την κατάσταση:



Σχήμα 6.2: Ένδειξη κατάστασης «Idle» (σε αναμονή). Εμφανίζεται όταν δεν εκτελείται το pipeline που κατεβάζει τα δεδομένα.

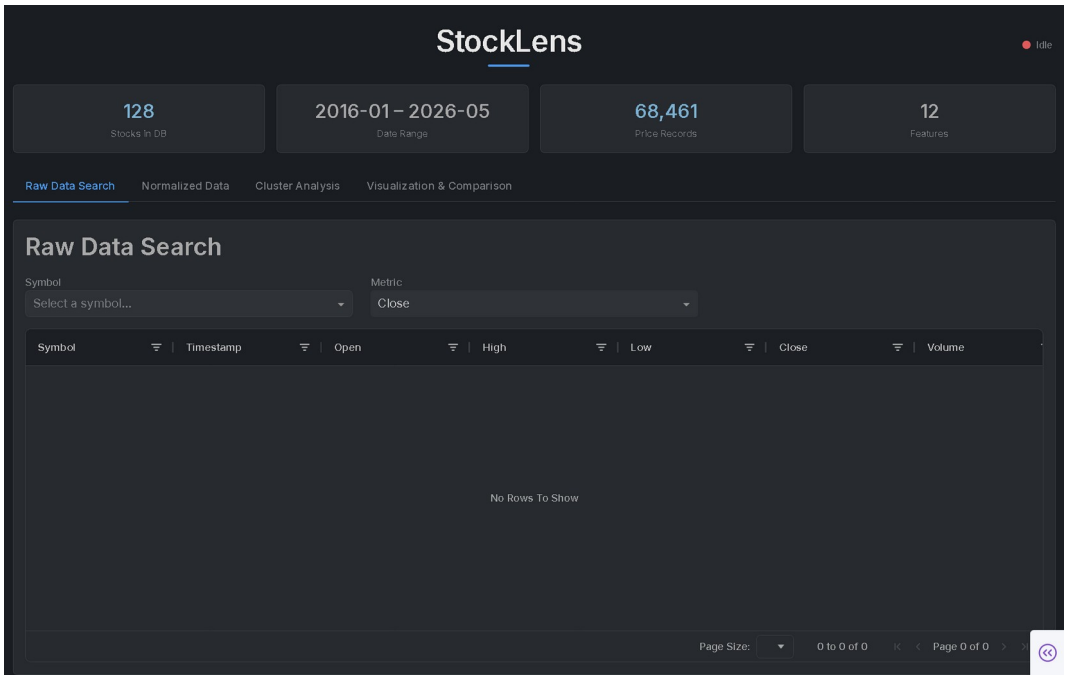
Σχήμα 6.3: Ένδειξη «Pipeline running» κατά τη διάρκεια συλλογής δεδομένων από IBKR ή SEC EDGAR.

Σχήμα 6.4: Ένδειξη «Next run» που δείχνει σε πόση ώρα (ώρες και λεπτά) θα εκτελεστεί ξανά το pipeline.

6.2 Καρτέλα Raw Data Search

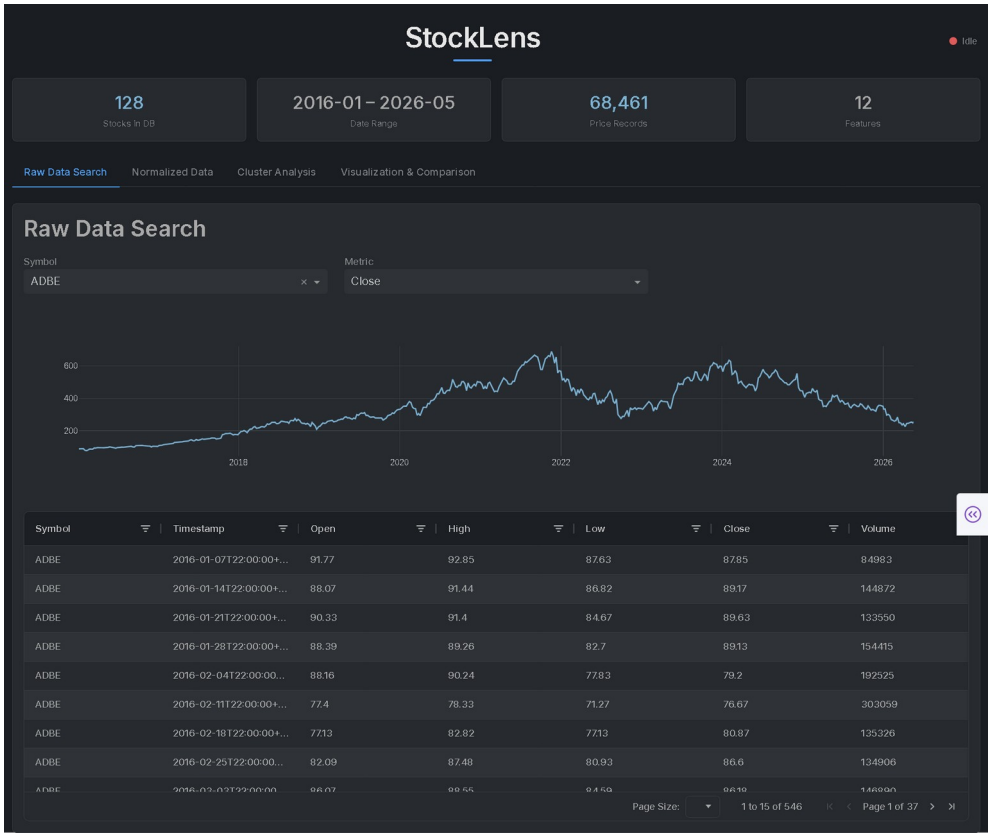
Η πρώτη καρτέλα είναι το σημείο εισόδου στην εφαρμογή. Ο χρήστης επιλέγει το ticker σύμβολο μιας μετοχής από dropdown και ένα feature. Η εφαρμογή ανακτά τα ιστορικά δεδομένα και τα εμφανίζει σε

Ag-Grid πίνακα με ενσωματωμένο time-series γράφημα, αφότου επιλεχθεί το Symbol και το metric.

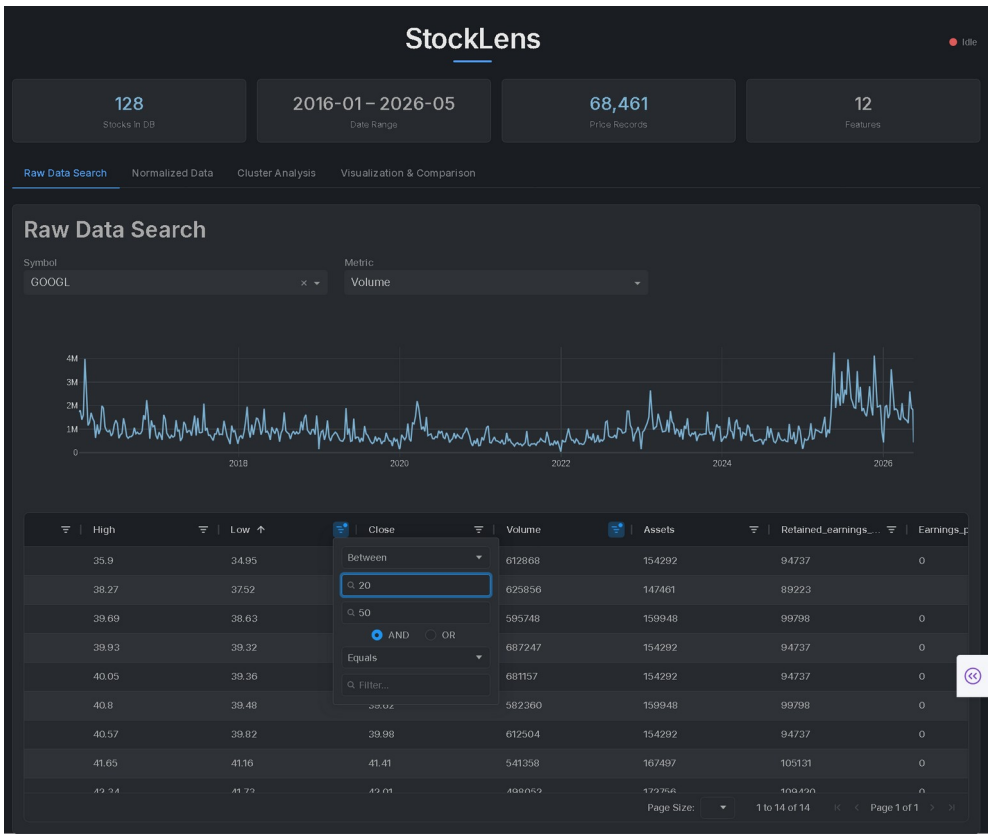


Σχήμα 6.5: Η καρτέλα «Raw Data Search» κατά την εκκίνηση.

Ο Ag-Grid πίνακας παρέχει ενσωματωμένες δυνατότητες φιλτραρίσματος, ταξινόμησης και pagination. Κάθε στήλη διαθέτει ανεξάρτητο φίλτρο με συνθήκες AND/OR για σύνθετες αναζητήσεις.



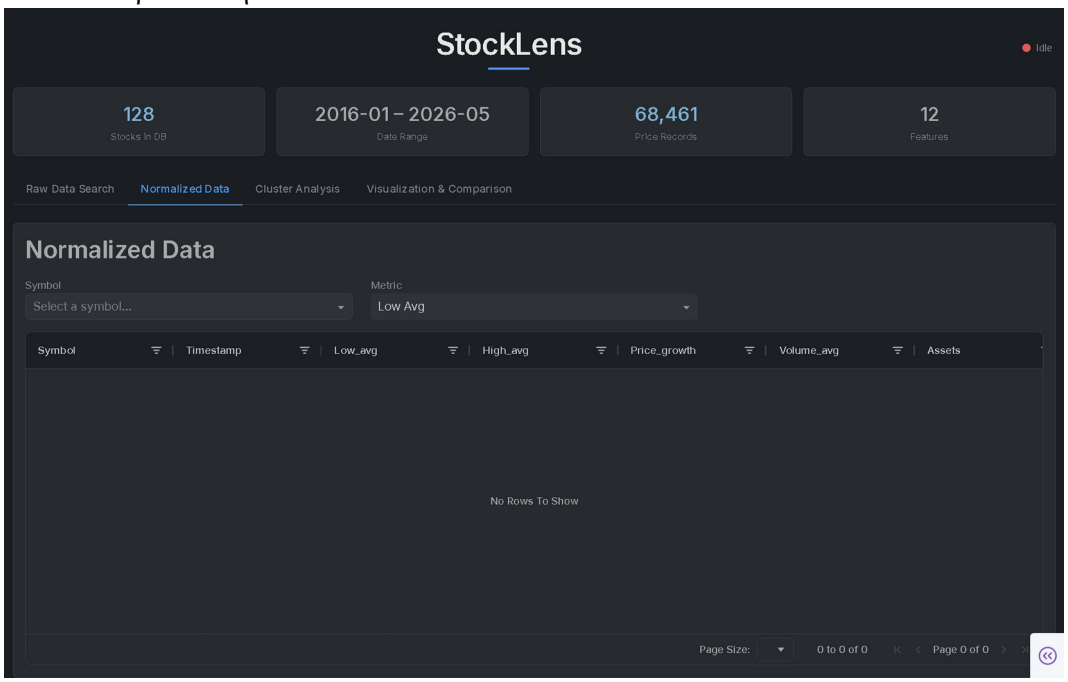
Σχήμα 6.6: Ιστορικά δεδομένα Close price για τη μετοχή ADBE (Adobe), με το time-series γράφημα και τον πίνακα OHLCV.



Σχήμα 6.7: Εφαρμογή φίλτρου σε στήλη του Ag-Grid (GOOGL, metric: Volume). Επιτρέπει συνδυασμό φίλτρων με AND/OR.

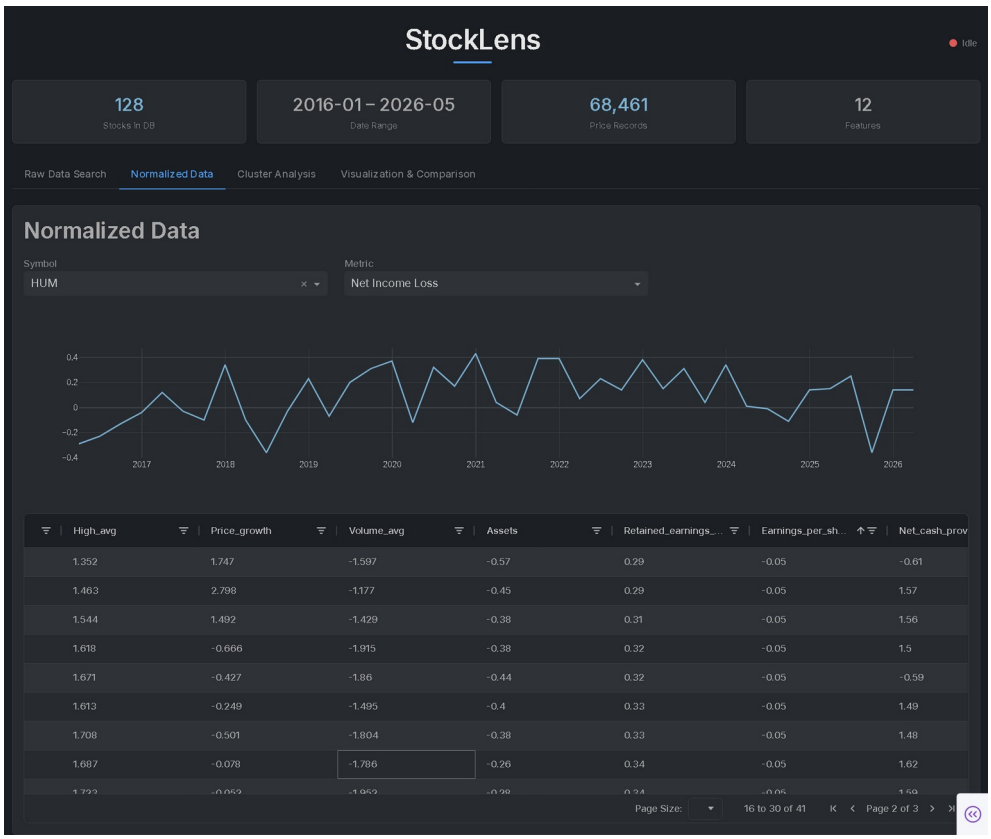
6.3 Καρτέλα Normalized Data

Η δεύτερη καρτέλα δείχνει τα δεδομένα μετά την επεξεργασία: αυτά που «βλέπουν» οι αλγόριθμοι ML. Ο σκοπός είναι να κατανοήσει ο χρήστης τι σημαίνουν οι normalized τιμές και να επαληθεύσει ότι η κανονικοποίηση εκτελέστηκε σωστά.



Σχήμα 6.8: Η καρτέλα «Normalized Data» κατά την εκκίνηση.

Ο πίνακας εμφανίζει normalized τιμές. Για παράδειγμα, η Apple με assets 350 δισ. δολάρια εμφανίζεται με normalized τιμή +2.1, ενώ μια μικρότερη εταιρεία με assets 5 δισ. εμφανίζεται με τιμή -0.8. Αυτές οι τιμές είναι συγκρίσιμες μεταξύ τους.



Σχήμα 6.9: Normalized δεδομένα για τη μετοχή HUM (Humana) με γράφημα Net Income/Loss.

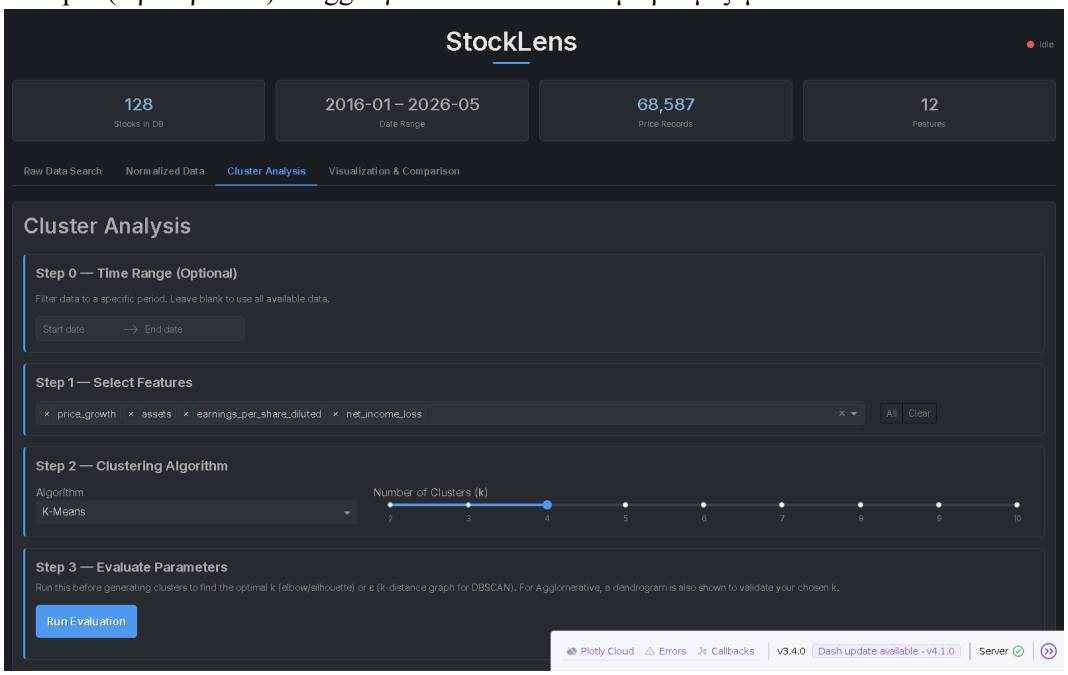
6.4 Καρτέλα Cluster Analysis

Η τρίτη καρτέλα είναι το κεντρικό εργαλείο ανάλυσης. Η ανάλυση οργανώνεται σε διαδοχικά βήματα (Step 0 έως 5) που καθοδηγούν τον χρήστη από την επιλογή παραμέτρων μέχρι την παραγωγή αποτελεσμάτων.

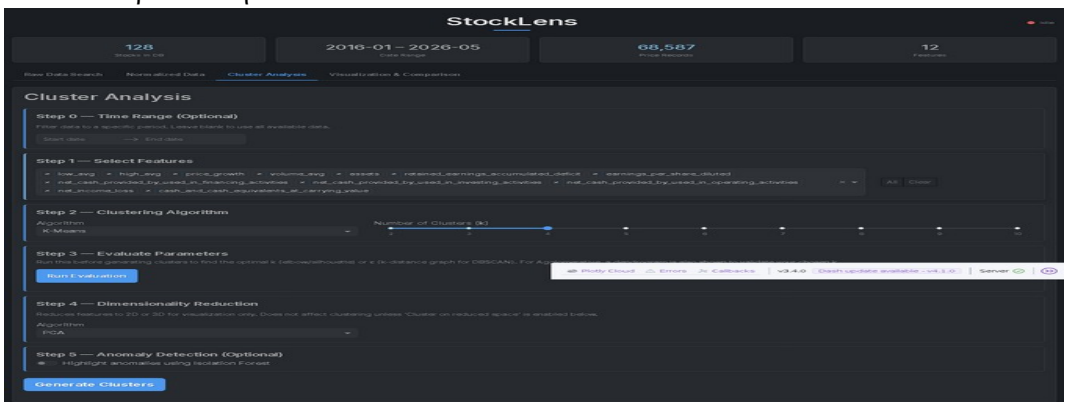
Controls ανάλυσης

Τα βήματα επιτρέπουν πλήρη παραμετροποίηση:

- Step 0 (Προαιρετικό): Date range picker για φιλτράρισμα δεδομένων σε συγκεκριμένη χρονική περίοδο, αν δεν ρυθμιστεί τότε χρησιμοποιούνται όλα τα δεδομένα στα επόμενα βήματα.
- Step 1: Multi-select dropdown με τα 12 διαθέσιμα features. Κουμπιά «All» και «Clear» για γρήγορη επιλογή.
- Step 2: Dropdown με τέσσερις αλγορίθμους (K-Means, DBSCAN, Agglomerative, GMM) και παραμέτρους αυτών.
- Step 3: Κουμπί «Run evaluation» για sweep k=2 έως 10 και εμφάνιση διαγνωστικών γραφημάτων.
- Step 4: Επιλογή μεθόδου οπτικοποίησης (PCA, UMAP, t-SNE). Όλες οι μέθοδοι παράγουν 2D scatter plot.
- Step 5 (Προαιρετικό): Toggle για Isolation Forest με ρυθμιζόμενο contamination rate.



Σχήμα 6.10: Συνολική εικόνα της καρτέλας Cluster Analysis. Φαίνονται τα βήματα Step 0-3.

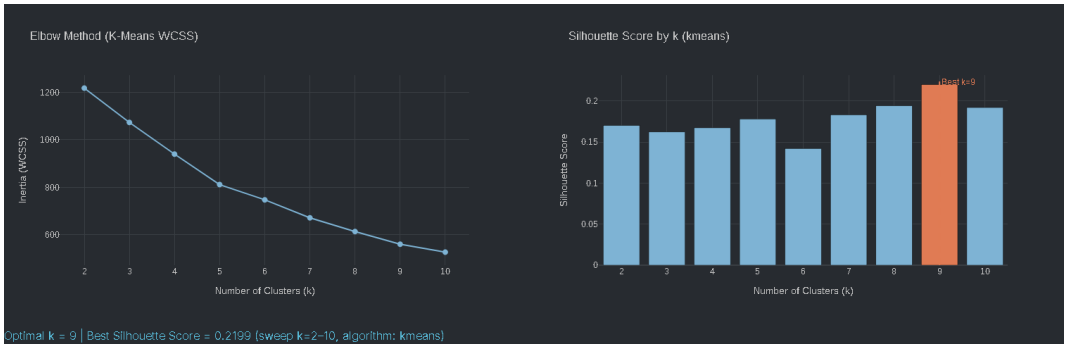


Σχήμα 6.11: Step 1, Επιλογή features. Φαίνονται και τα 12 features επιλεγμένα.

Σχήμα 6.12: Step 0, Date range picker για φιλτράρισμα δεδομένων σε συγκεκριμένη χρονική περίοδο.

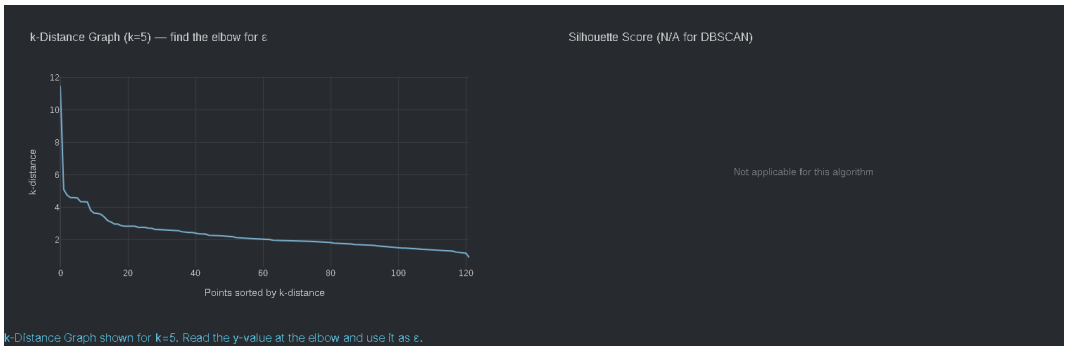
Αξιολόγηση παραμέτρων (Step 3)

Πριν εκτελεστεί το clustering, το Step 3 παρέχει διαγνωστικά γραφήματα προσαρμοσμένα σε κάθε αλγόριθμο. Για τον K-Means εμφανίζεται το elbow plot και το silhouette score ανά k:



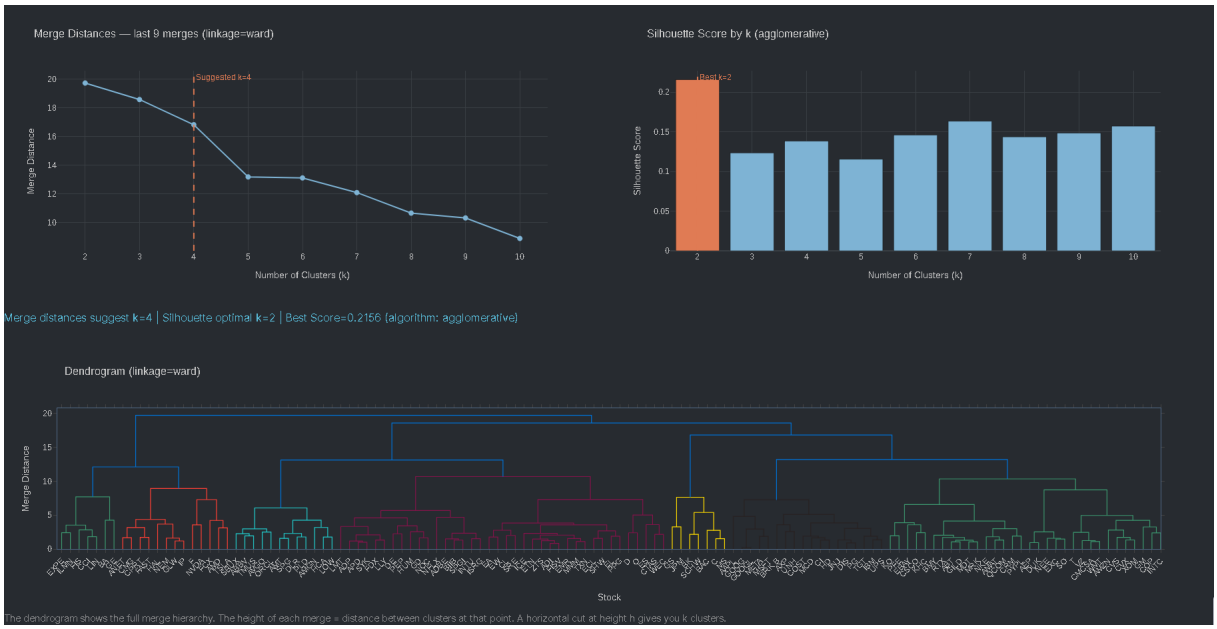
Σχήμα 6.13: Αξιολόγηση K-Means: elbow plot (αριστερά) και Silhouette Score ανά k (δεξιά).

Για τον DBSCAN εμφανίζεται το k-Distance Graph:



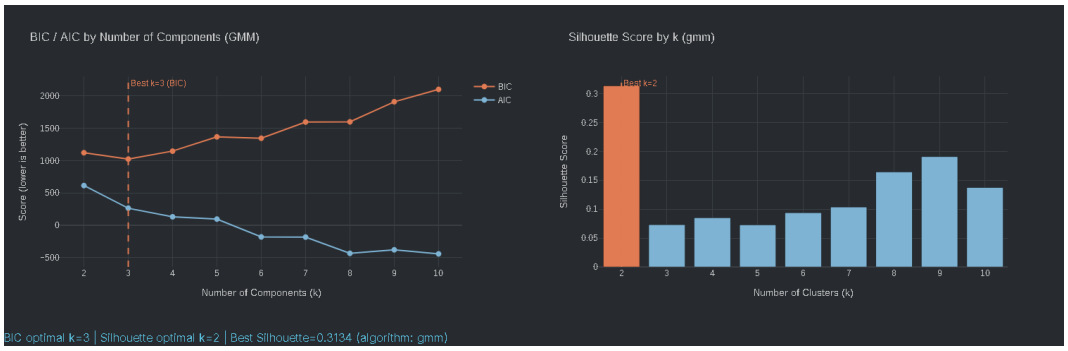
Σχήμα 6.14: Αξιολόγηση DBSCAN: k-Distance Graph για επιλογή ε.

Για τον Agglomerative εμφανίζονται τα merge distances, το silhouette και πλήρες dendrogram:



Σχήμα 6.15: Αξιολόγηση Agglomerative Clustering: merge distances, silhouette και dendrogram.

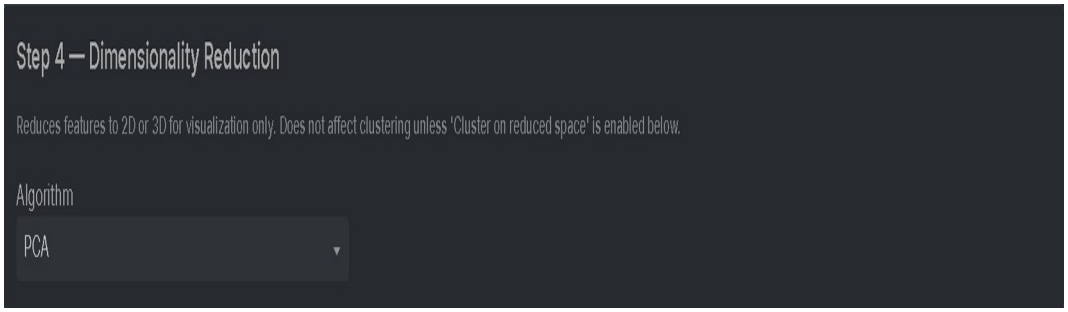
Για τα GMM εμφανίζεται BIC/AIC ανά k:



Σχήμα 6.16: Αξιολόγηση GMM: BIC και AIC ανά k.

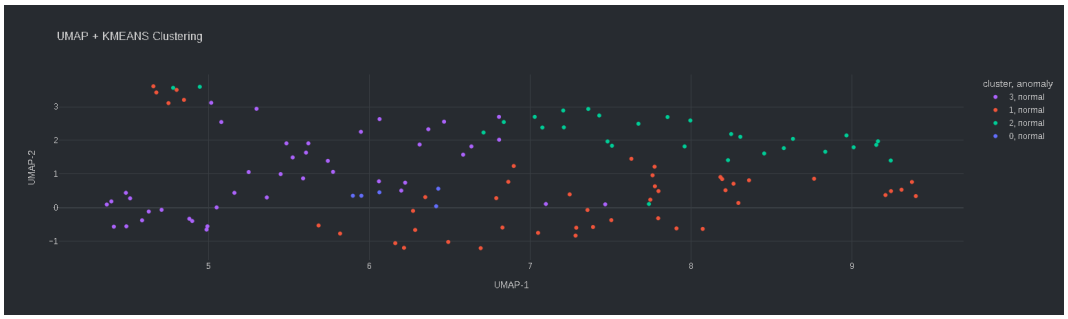
Scatter plot αποτελεσμάτων

Κάθε σημείο αντιστοιχεί σε μία μετοχή. Οι άξονες x και y είναι οι δύο πρώτες διαστάσεις της επιλεγμένης μεθόδου μείωσης. Στο Step 4 διαλέγουμε μεταξύ PCA, UMAP και t-SNE:

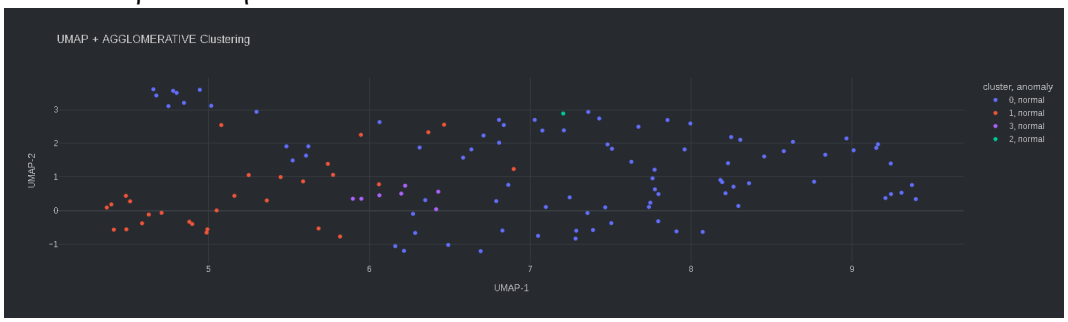


Σχήμα 6.17: Step 4, Επιλογή μεθόδου μείωσης διαστάσεων (PCA, UMAP ή t-SNE) για 2D οπτικοποίηση.

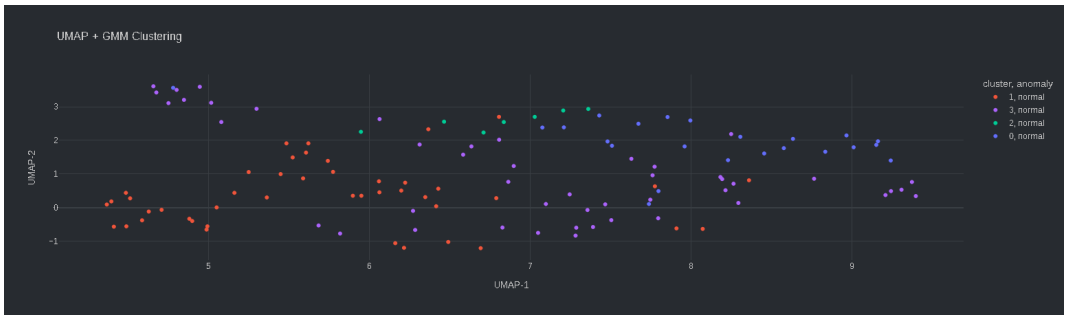
Παρακάτω παρουσιάζονται τα αποτελέσματα και των τεσσάρων αλγορίθμων με UMAP:



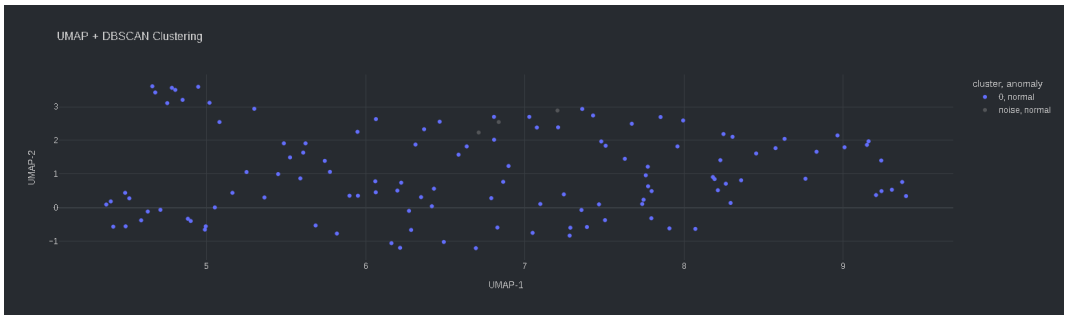
Σχήμα 6.18: K-Means + UMAP, k=4. Ο αλγόριθμος διαχωρίζει τις μετοχές σε ομάδες με διαφορετικά οικονομικά profiles.



Σχήμα 6.19: Agglomerative + UMAP. Παρόμοια διάρθρωση με K-Means.



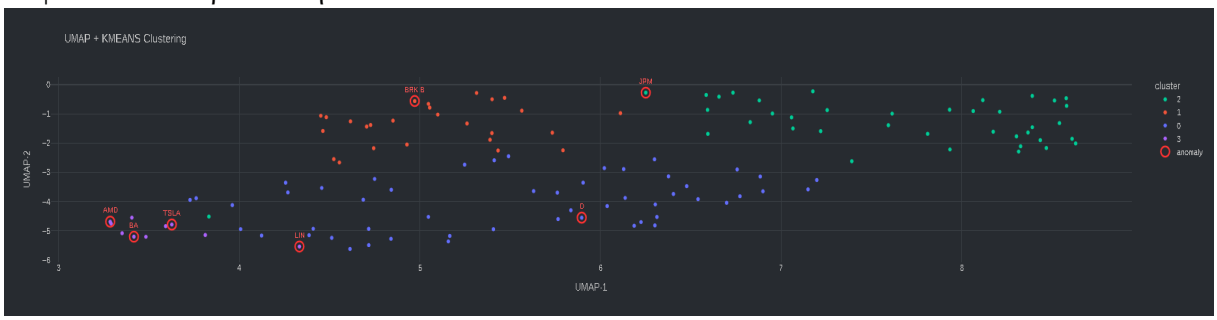
Σχήμα 6.20: GMM + UMAP, probabilistic clustering.



Σχήμα 6.21: DBSCAN + UMAP. Βρίσκει clusters και noise points (γκρι). Τα fundamentals δεν σχηματίζουν πυκνές νησίδες στον 12-διάστατο χώρο.

Εντοπισμός ανωμαλιών (Step 5)

Με την ενεργοποίηση του Step 5, εφαρμόζεται Isolation Forest στα δεδομένα. Τα σημεία κρατούν το χρώμα της συστάδας τους, και όσα βγαίνουν ανώμαλα σημειώνονται με έναν κόκκινο κύκλο γύρω τους και το ticker τους από πάνω. Έτσι φαίνεται με μια ματιά ποιες εταιρείες ξεχωρίζουν ως outliers:



Σχήμα 6.22: K-Means + UMAP με Isolation Forest (contamination=5%). Οι ανωμαλίες σημειώνονται με κόκκινο κύκλο και το ticker τους.

Πίνακες αποτελεσμάτων

Κάτω από το scatter plot εμφανίζονται δύο πίνακες. Ο Cluster Summary δείχνει τον μέσο όρο κάθε feature ανά συστάδα:

Cluster Summary (mean feature values)

cluster	size	low_avg	high_avg	price_growth	volume_avg	assets	retained_earnings_accumulated_deficit	earnings_per_share_diluted	net_cash_provided_by_used_in_fi
0	5	-0.1453	-0.1448	0.0969	0.6436	2.3512	0.6766	-0.05	
1	46	0.6879	0.6838	-0.1463	-0.7376	-0.2652	-0.0216	-0.05	
2	29	-0.4157	-0.409	0.2225	-0.1744	-0.9984	-0.4611	0.1439	
3	42	-0.4291	-0.4319	-0.0918	0.72	0.5837	0.1828	-0.05	

Σχήμα 6.23: Πίνακας Cluster Summary: μέσες τιμές features ανά συστάδα.

Ο Cluster Members εμφανίζει τα ticker symbols κάθε συστάδας:

Cluster Members

Cluster	Tickers
0	BAC, C, GS, JPM, MS
1	ADBE, ADI, ADP, AEP, AMGN, AMT, APD, AXP, BIIB, CI, CL, COST, CRM, DE, DUK, ECL, EMR, ETN, FDX, FISV, GD, HCA, HD, HUM, IBM, ICE, INTU, ISRG, LLY, LMT, LOW, MCD, MMM, NFLX, NOC, PEP, PPG, SHW, SPGI, SYK, TGT, TMUS, TXN, UNH, UPS, ZTS
2	AMD, ANET, BA, CCI, CMG, CPRT, CTAS, D, DLR, EA, EBAY, EL, ES, EW, EXPE, FAST, FIS, GE, GLW, HAL, ILMN, IP, KHC, LIN, NEM, O, PLD, SRE, TSLA
3	AAPL, ABBV, ABT, AMAT, AMZN, AVGO, BK, BMY, CMCSA, COP, CSCO, CVS, CVX, DIS, F, GILD, GM, GOOG, GOOGL, INTC, JNJ, KO, MDT, META, MRK, MSFT, NEE, NKE, NVDA, ORCL, PFE, PG, PYPL, QCOM, RTX, SBUX, SCHW, SO, T, VZ, WMT, XOM

Σχήμα 6.24: Πίνακας Cluster Members: tickers ανά συστάδα.

Parallel Coordinates γράφημα

Κάτω από τους πίνακες εμφανίζεται ένα parallel coordinates γράφημα για σύγκριση feature profiles μεταξύ συστάδων. Κάθε κάθετος άξονας αντιστοιχεί σε ένα feature, κάθε γραμμή σε μία μετοχή και το χρώμα στη συστάδα. Με brush selection φιλτράρονται εταιρείες βάσει εύρους τιμών σε οποιοδήποτε axis:

Σχήμα 6.25: Parallel Coordinates γράφημα, K-Means k=4, UMAP. Κάθε κάθετος άξονας είναι ένα από τα 12 features (normalized).

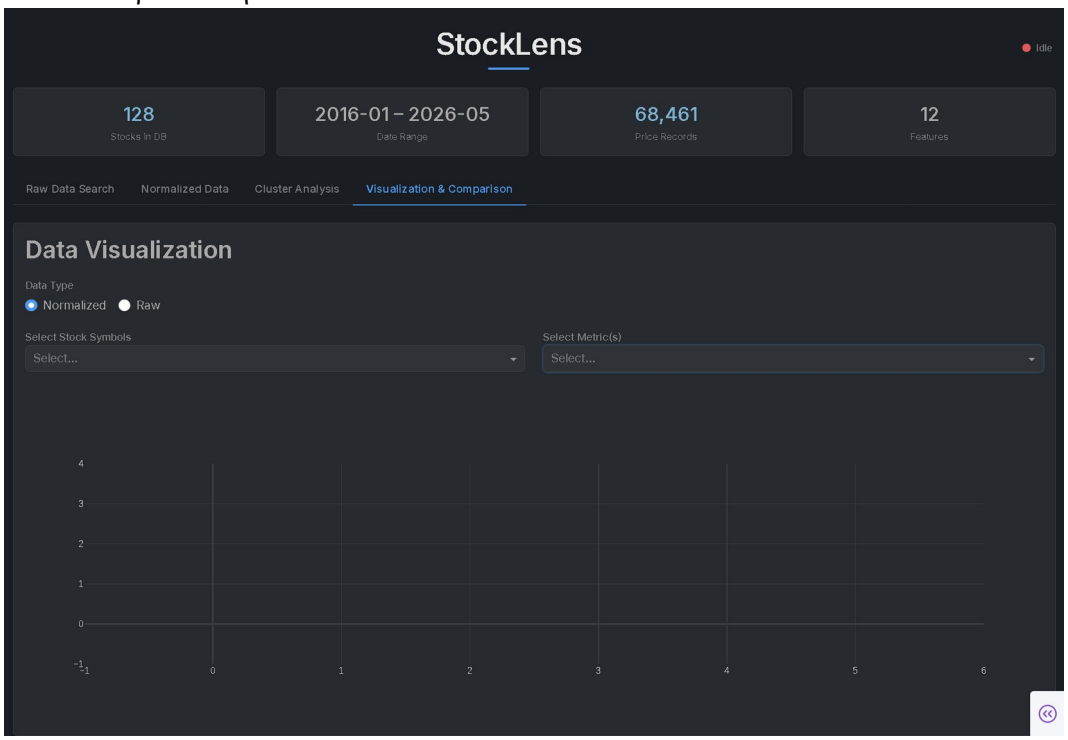
Εξαγωγή αποτελεσμάτων

Στο κάτω μέρος της καρτέλας Cluster Analysis υπάρχει κουμπί εξαγωγής αποτελεσμάτων σε CSV. Το αρχείο περιέχει για κάθε μετοχή το symbol, τον αριθμό συστάδας, την ένδειξη anomaly και τα επιλεγμένα features, επιτρέποντας περαιτέρω ανάλυση εκτός εφαρμογής:

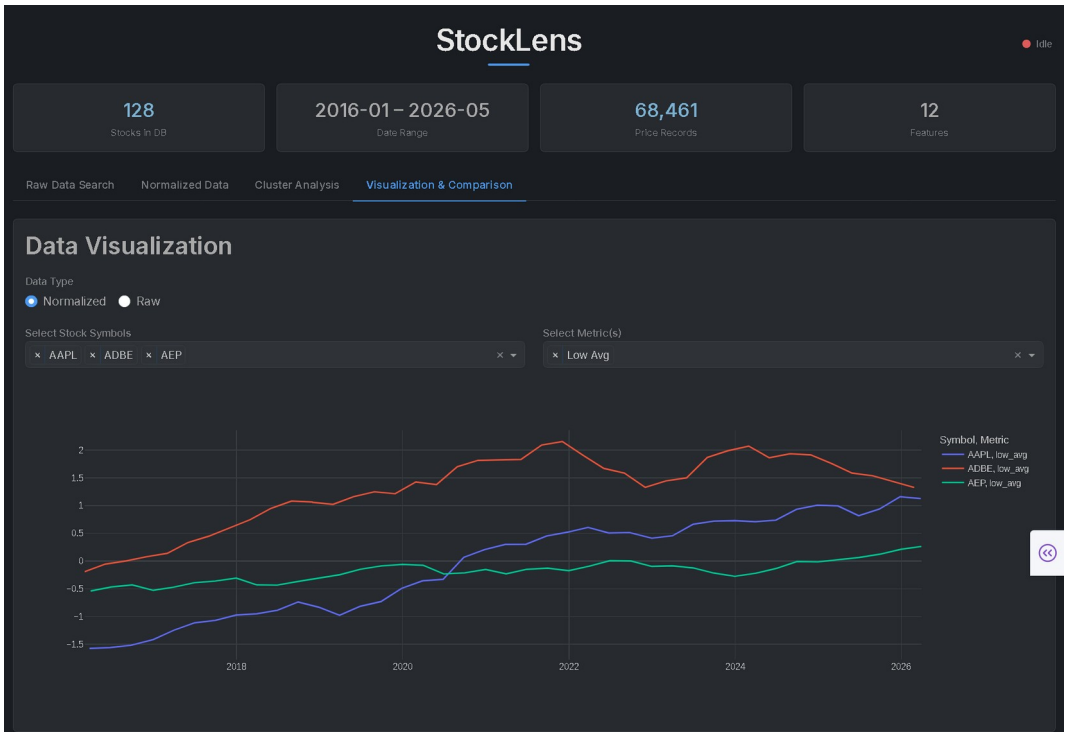
Σχήμα 6.26: Κουμπί εξαγωγής CSV με τα αποτελέσματα clustering.

6.5 Καρτέλα Visualization & Comparison

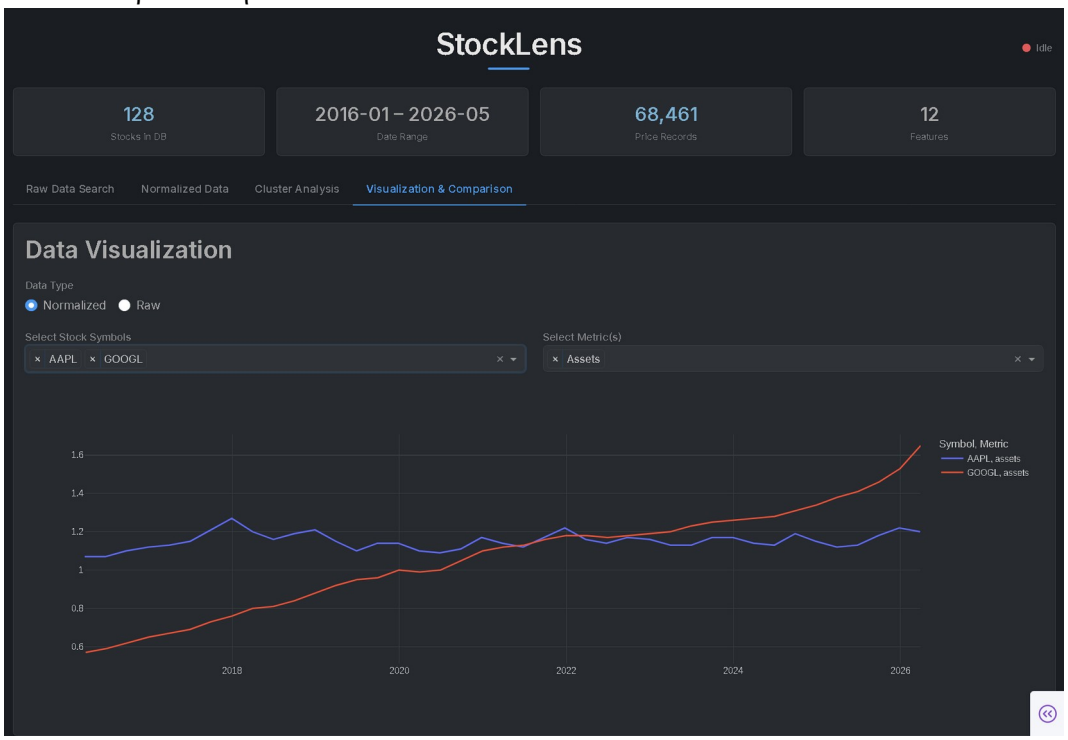
Η τέταρτη καρτέλα επικεντρώνεται στη χρονική εξέλιξη και σύγκριση μεταξύ μετοχών. Ο χρήστης επιλέγει πολλαπλές μετοχές και ένα ή περισσότερα financial metrics. Η εφαρμογή εμφανίζει time-series γράφημα και μας δίνει τη δυνατότητα να συγκρίνουμε τα δεδομένα που έχουμε επιλέξει. Υποστηρίζεται εναλλαγή μεταξύ normalized και raw τιμών.



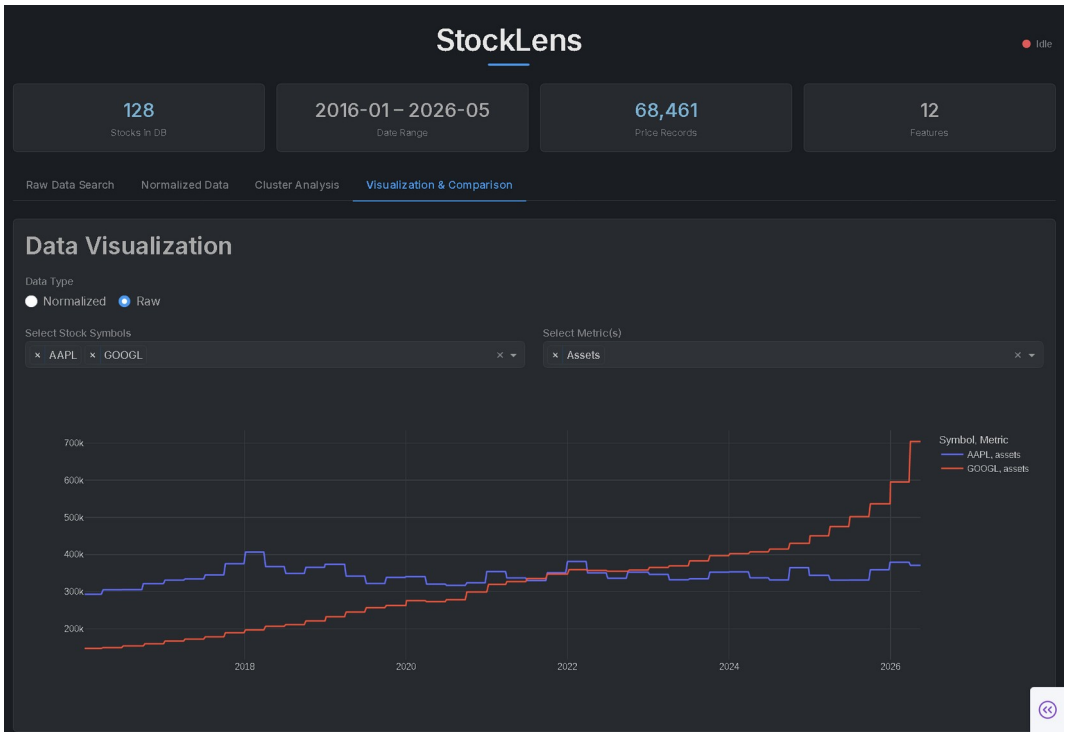
Σχήμα 6.27: Η καρτέλα «Visualization & Comparison» κατά την εκκίνηση.



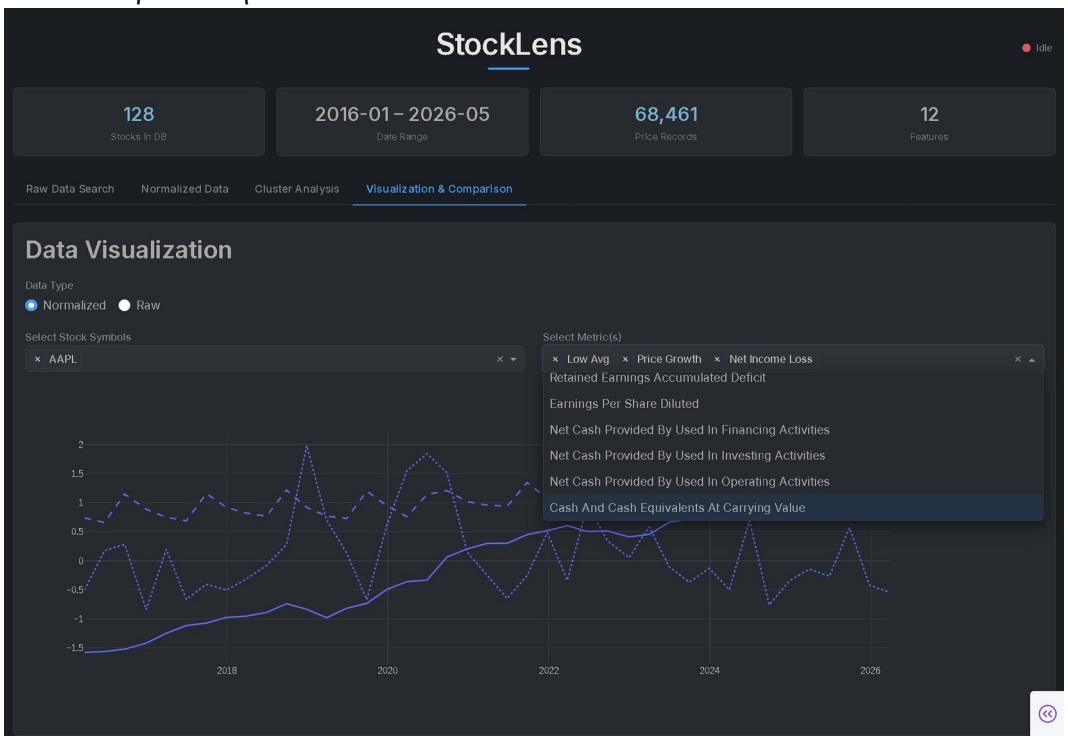
Σχήμα 6.28: Σύγκριση AAPL, ADBE και AEP ως προς normalized Low Avg.



Σχήμα 6.29: Σύγκριση AAPL vs GOOGL ως προς normalized Assets.



Σχήμα 6.30: Σύγκριση AAPL vs GOOGL ως προς raw Assets (εκατ. $).



Σχήμα 6.31: Επιλογή πολλαπλών metrics για AAPL.

Επίλογος

Είδαμε και τις τέσσερις καρτέλες με screenshots, την εφαρμογή σε πλήρη λειτουργία. Στο επόμενο κεφάλαιο χρησιμοποιώ αυτά τα εργαλεία για να τρέξω πραγματικά πειράματα πάνω στα δεδομένα.

Κεφάλαιο 7ο: Cluster analysis μέσω του StockLens

7.1 Εισαγωγή

Το κεφάλαιο αυτό παρουσιάζει τα αποτελέσματα τριών πειραμάτων που εκτελέστηκαν στο πλήρες dataset. Θέλουμε να δούμε αν οι αλγόριθμοι βγάζουν συστάδες που έχουν οικονομικό νόημα, αν εντοπίζουν μετοχές που ξεχωρίζουν, και αν αντικατοπτρίζουν αλλαγές που έγιναν στα fundamentals με την πάροδο του χρόνου.

7.2 Dataset και μεθοδολογία

Τα δεδομένα αφορούν 127 μετοχές για την περίοδο 2016-2026, που ανήκουν σε διαφορετικούς κλάδους:

Τεχνολογία (21):

AAPL, MSFT, NVDA, ORCL, ADBE, CRM, CSCO, INTC, QCOM, TXN, IBM, AMD, AVGO, ADI, AMAT, ANET, GLW, INTU, FIS, FISV, ADP

Υπηρεσίες Επικοινωνίας (10):

GOOG, GOOGL, META, NFLX, DIS, CMCSA, T, TMUS, EA, VZ

Καταναλωτικά (μη βασικά) (13):

AMZN, TSLA, HD, MCD, NKE, TGT, LOW, SBUX, CMG, F, GM, EXPE, EBAY

Καταναλωτικά Βασικά (8):

WMT, PG, PEP, KO, COST, CL, KHC, EL

Χρηματοπιστωτικός (12):

JPM, BAC, BRK B, GS, MS, AXP, PYPL, SCHW, C, BK, SPGI, ICE

Υγεία (21):

JNJ, PFE, LLY, MRK, UNH, ABBV, ABT, BMY, AMGN, GILD, CVS, MDT, ISRG, BIIB, HCA, HUM, ILMN, SYK, ZTS, CI, EW

Ενέργεια (4):

XOM, CVX, HAL, COP

Βιομηχανία (15):

GE, BA, MMM, DE, LMT, RTX, NOC, FDX, UPS, EMR, ETN, GD, FAST, CTAS, CPRT

Πρώτες Ύλες (8):

LIN, SHW, FCX, NEM, PPG, APD, ECL, IP

Ακίνητα (6):

PLD, DLR, AMT, SPG, O, CCI

Υπηρεσίες Κοινής Ωφέλειας (9):

NEE, DUK, SO, AEP, D, EXC, SRE, WEC, ES

Μετά την κανονικοποίηση, το dataset περιέχει 4988 εγγραφές για τις 127 μετοχές.

Τα features για κάθε (μετοχή, τρίμηνο):

- Market features (4): low_avg, high_avg, price_growth, volume_avg.
- Fundamental features (8): assets, retained_earnings, EPS diluted, cash, net income, operating/investing/financing cash flow.

Η μεθοδολογία: για κάθε πείραμα ορίζεται ερευνητικό ερώτημα, εκτελείται η ανάλυση με συγκεκριμένες παραμέτρους, και τα αποτελέσματα έπειτα ερμηνεύονται.

7.3 Πείραμα 1: Επαλήθευση κλαδικής δομής (Sector Validation)

Ερώτημα: μπορεί το σύστημα να ανακτά τα όρια των επιχειρηματικών κλάδων χωρίς να του δοθούν sector labels; Χρησιμοποιήθηκε K-Means με PCA για οπτικοποίηση.

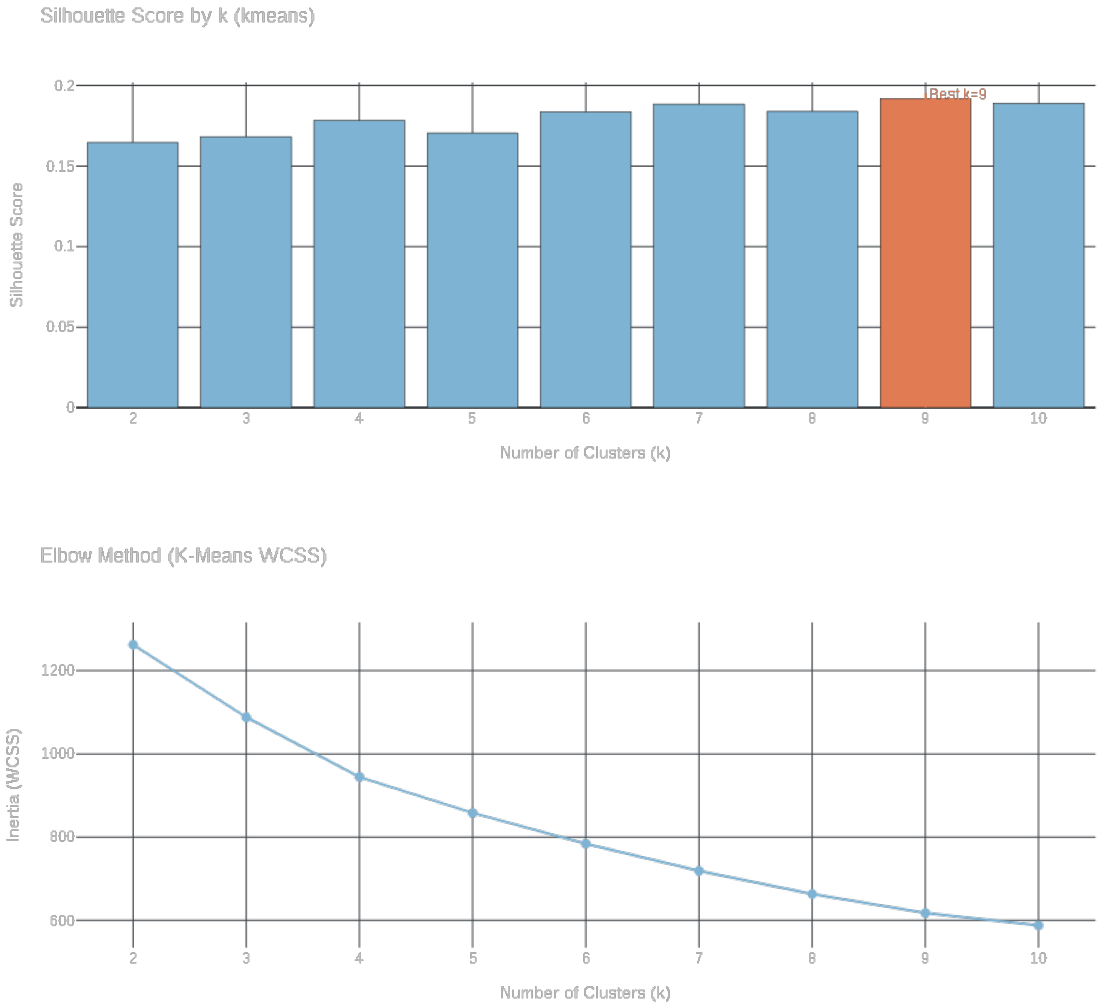
Διαλέξαμε PCA για την οπτικοποίηση γιατί έδειχνε τις ομάδες πιο καθαρά από το UMAP και το t-SNE. Ο K-Means χωρίζει τα δεδομένα με βάση την απόσταση, και το PCA κρατάει αυτές τις αποστάσεις όταν τα φέρνει στο 2D, οπότε οι συστάδες μένουν μαζί και ξεχωρίζουν καλύτερα στο γράφημα.

Σύνολο features	**Features**	**Best K**	**Silhouette**
Market	4	2	0.378
Fundamentals	8	6	0.241
Όλα τα features	12	9	0.192

Πίνακας 7.1: Σύγκριση τριών υποσυνόλων features, εύρεση βέλτιστου K.

Το feature set «Market μόνο» έχει τον καλύτερο silhouette (0.378) επειδή οι τιμές και τα volumes των μετοχών είναι ομαλά και χωρίζονται καθαρά στον 4D χώρο. Αυτό όμως δεν σημαίνει ότι είναι και η καλύτερη επιλογή. Τα βασικότερα χαρακτηριστικά των τραπεζών υπάρχουν μόνο στις fundamental διαστάσεις. Χωρίς αυτές, ο αλγόριθμος δεν μπορεί να ξεχωρίσει τις τράπεζες.

Έτσι, ο χαμηλότερος silhouette με τα 12 features (0.192) δεν είναι χειρότερο αποτέλεσμα. Τα fundamentals προσθέτουν πολυπλοκότητα, αλλά δείχνουν δομή που τα market features δεν βλέπουν. Πολλές φορές ένας χαμηλός silhouette δείχνει ότι οι εταιρείες δεν χωρίζονται σε καθαρές ομάδες, όχι ότι το clustering είναι κακό. Γι' αυτό διαλέξαμε τα 12 features για όλα τα πειράματα.



Σχήμα 7.1: K-Means sweep για k από 2 έως 10 στο πλήρες dataset με τα 12 features. Πάνω ο silhouette score ανά k, κάτω το elbow με το WCSS.

Επιλογή k: Τρέξαμε sweep k=2 έως 10 με silhouette και WCSS (Σχήμα 7.1). Δεν υπάρχει καθαρό βέλτιστο k, κάτι αναμενόμενο για χρηματοοικονομικά fundamentals σε 12 διαστάσεις. Ο silhouette μένει σχεδόν επίπεδος σε όλο το εύρος, με τον υψηλότερο αριθμό στο k=9 (0.192). Η διαφορά όμως από τα μικρότερα k είναι πολύ μικρή, οπότε δεν έχει πρακτική σημασία. Το WCSS κάμπτεται γύρω στο k=4, εκεί που η κάθε επιπλέον συστάδα σταματά να μειώνει αισθητά το σφάλμα. Διαλέξαμε k=4 επειδή δίνει τις πιο ερμηνεύσιμες ομάδες, όχι επειδή είναι το αριθμητικά βέλτιστο. Με k=9 θα είχαμε 9 ομάδες δύσκολες να περιγραφούν, για κέρδος silhouette που ουσιαστικά δεν υπάρχει.

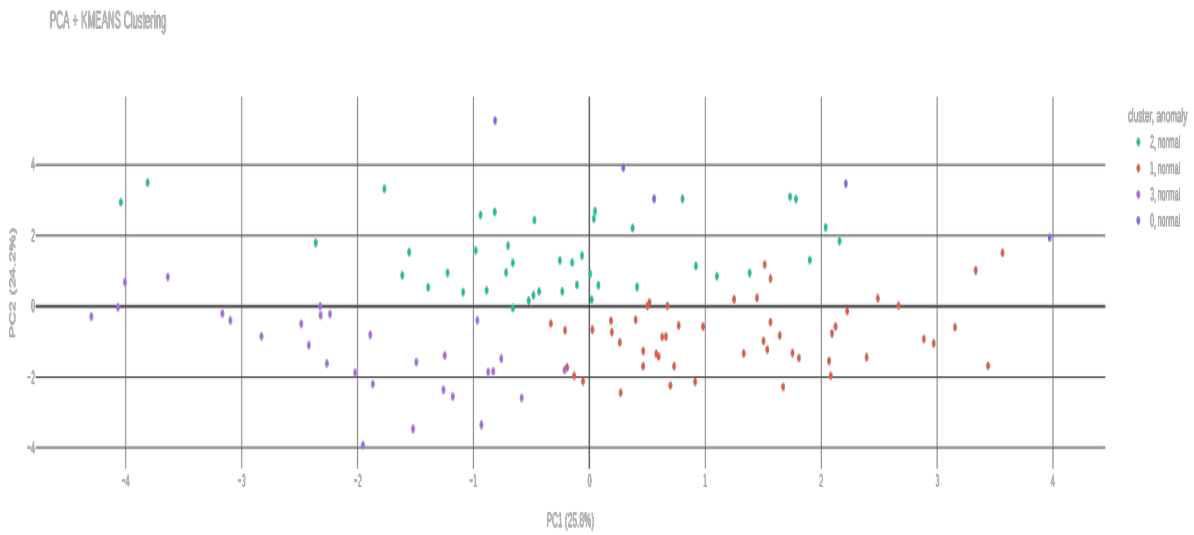
Κεφάλαιο 7ο: Cluster analysis μέσω του StockLens

Συστάδα 0 (5 εταιρείες): οι τράπεζες. BAC, C, GS, JPM, MS. Έχουν τεράστια assets (+2.38σ) και cash (+2.40σ), υψηλό financing cash flow (+1.62σ) και έντονα αρνητικό operating cash flow (-1.93σ). Αυτός ο συνδυασμός είναι η τυπική εικόνα ισολογισμού μιας τράπεζας, και ο αλγόριθμος τις ξεχώρισε μόνος του χωρίς καμία πληροφορία για τον κλάδο.

Συστάδα 1 (51 εταιρείες): υψηλές τιμές μετοχής (low/high avg +0.67σ) και χαμηλό volume (-0.71σ). Είναι οι ώριμες, σταθερές μεγάλες εταιρείες, με θετικό EPS. Εδώ μπαίνουν utilities, healthcare, βιομηχανίες και consumer staples, κλάδοι με σταθερά κέρδη και λιγότερο έντονο trading.

Συστάδα 2 (42 εταιρείες): χαμηλότερη τιμή μετοχής αλλά πολύ υψηλό volume (+0.78σ) και μεγάλα assets (+0.57σ). Οι μεγάλες εταιρείες με έντονο trading. Εδώ βρίσκονται οι τεχνολογικοί κολοσσοί όπως AAPL, MSFT, NVDA, META, GOOGL, AMZN, μαζί με ενεργειακές (XOM, CVX) και μεγάλες φαρμακευτικές. Είναι ενδιαφέρον ότι μετοχές από τελείως διαφορετικούς κλάδους μπαίνουν μαζί, επειδή μοιράζονται το ίδιο χαρακτηριστικό, τον μεγάλο όγκο συναλλαγών.

Συστάδα 3 (29 εταιρείες): οι μικρότερες εταιρείες. Αρνητικές σχεδόν σε όλα, assets -0.89σ, net income -0.70σ, EPS -0.63σ, cash -0.65σ. Έχουν όμως το υψηλότερο price growth (+0.29σ). Εδώ μέσα βρίσκονται και η HAL και η TSLA.



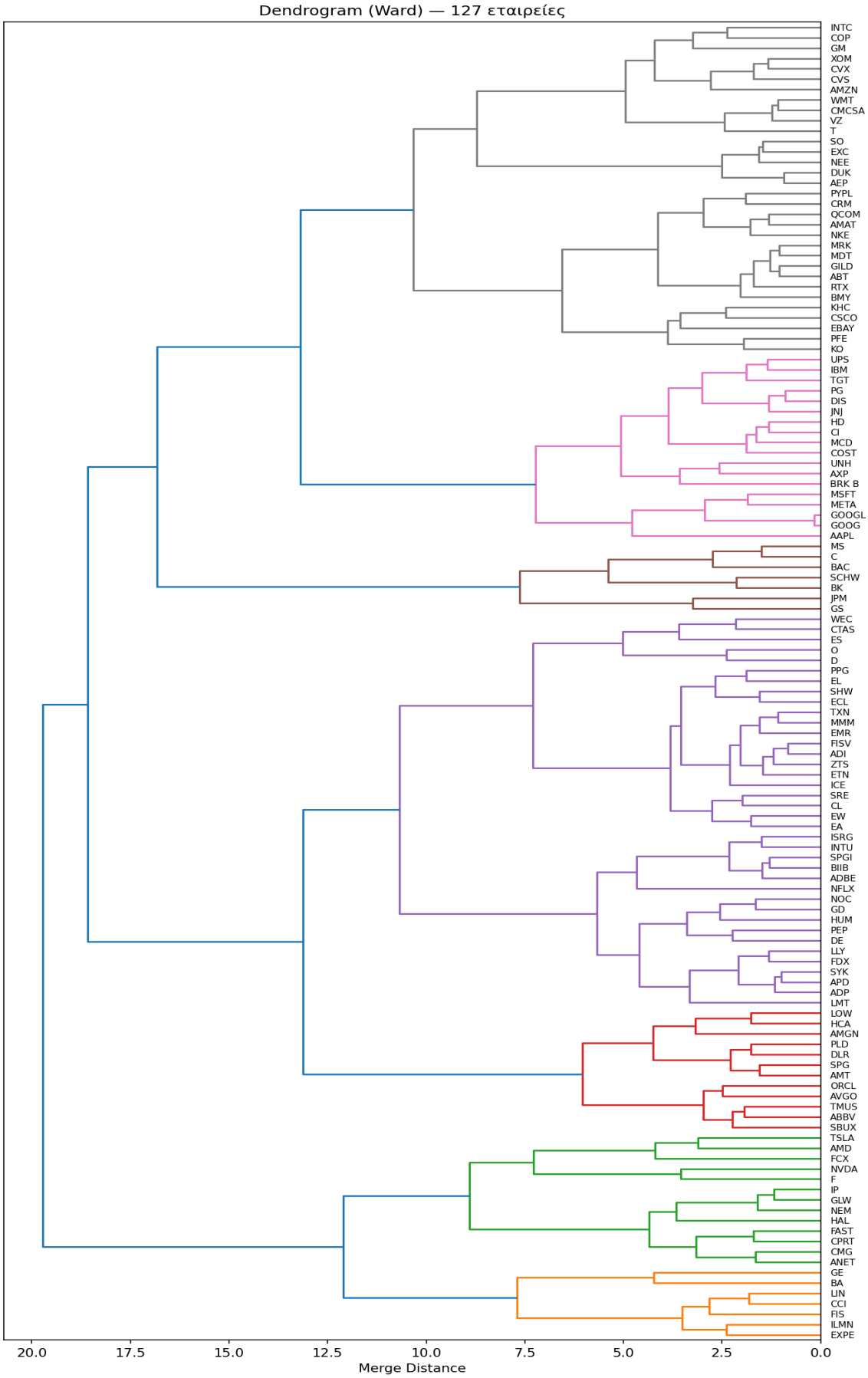
Σχήμα 7.2: K-Means k=4 με PCA, scatter plot των 127 εταιρειών, χρωματισμένο ανά συστάδα. Οι δύο πρώτοι άξονες εξηγούν PC1: 25.8% και PC2: 24.2% της διακύμανσης.

Κύριο εύρημα: ο αλγόριθμος εντόπισε τις τράπεζες χωρίς καμία πληροφορία για τον κλάδο, μόνο από τη δομή των assets και των cash flows τους. Ενδιαφέρουσα είναι και η Συστάδα 2, που βάζει μαζί tech κολοσσούς, ενεργειακές και φαρμακευτικές. Είναι εταιρείες από τελείως διαφορετικούς κλάδους, που όμως μοιράζονται μεγάλο όγκο συναλλαγών και μεγάλα assets. Αυτή η ομαδοποίηση δεν υπάρχει σε καμία παραδοσιακή κλαδική ταξινόμηση.

Αλγόριθμος	Παράμετροι	Silhouette	Παρατήρηση
K-means	k=4	0.178	Καθαρές, ερμηνεύσιμες ομάδες, απομονώνει τις τράπεζες
GMM	k=4	0.178	Παρόμοιο με k-means
Agglomerative	k=4, Ward	0.138	Λίγο χειρότερος διαχωρισμός, αλλά χρήσιμο το dendrogram
DBSCAN	ε=3.0, minPts=5	-	1 ομάδα και 12 noise, δεν βρίσκει δομή

Πίνακας 7.2: Σύγκριση αλγορίθμων clustering στο πλήρες dataset, όλοι με k=4.

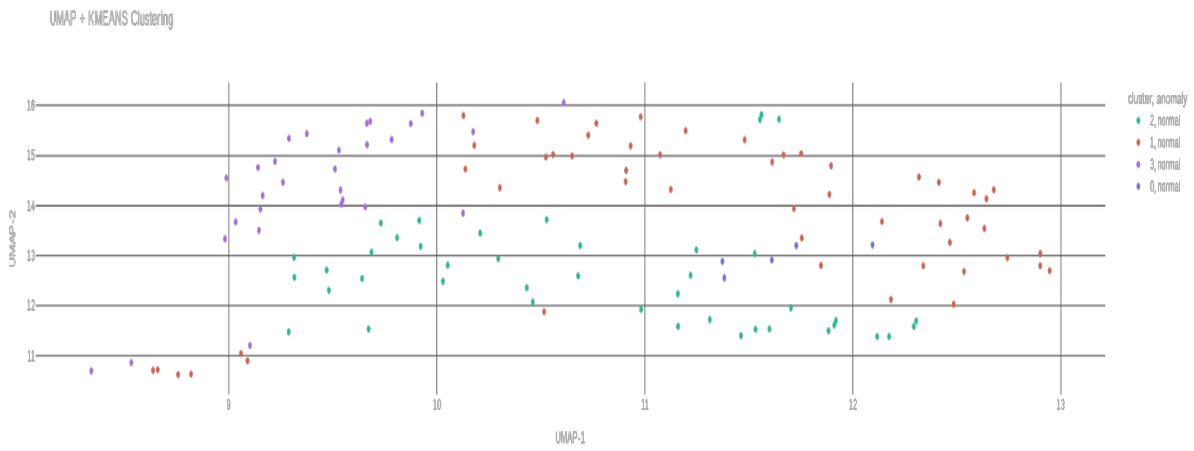
Και οι τρεις πρώτοι αλγόριθμοι τρέχουν με k=4, για να είναι δίκαιη η σύγκριση. Ο K-Means και ο GMM βγάζουν σχεδόν τον ίδιο silhouette (0.178) και τις ίδιες τέσσερις ομάδες, οπότε στην πράξη δεν υπάρχει διαφορά. Ο Agglomerative με Ward πέφτει λίγο πιο χαμηλά (0.138), δηλαδή ο διαχωρισμός του είναι λιγότερο καθαρός, αλλά το dendrogram του βοηθά να δεις την ιεραρχία ανάμεσα στις εταιρείες. Ο DBSCAN δεν ταιριάζει σε αυτά τα δεδομένα. Με ε από το k-distance knee, γύρω στο 3.0, βάζει σχεδόν τα πάντα σε μία ομάδα και αφήνει 12 εταιρείες σαν noise. Αν μικρύνουμε το ε για να βγουν περισσότερες ομάδες, το noise εκτοξεύεται, στο ε=2.0 φτάνει το 44%. Αυτό δεν σημαίνει ότι οι εταιρείες είναι ανώμαλες, απλά ότι τα fundamentals απλώνονται ομοιόμορφα στον 12-διάστατο χώρο και ο DBSCAN, που ψάχνει πυκνές περιοχές, δεν έχει πού να πιαστεί.


Dendrogram (Ward) — 127 εταιρείες
INTC
COP
GM
XOM
CVX
CVS
AMZN
WMT
CMCSA
VZ
T
SO
EXC
NEE
DUK
AEP
PYPL
CRM
QCOM
AMAT
NKE
MRK
MDT
GILD
ABT
RTX
BMY
KHC
CSCO
EBAY
PFE
KO
UPS
IBM
TGT
PG
DIS
JNJ
HD
CI
MCD
COST
UNH
AXP
BRK B
MSFT
META
GOOGL
GOOG
AAPL
MS
C
BAC
SCHW
BK
JPM
GS
WEC
CTAS
ES
O
D
PPG
EL
SHW
ECL
TXN
MMM
EMR
FISV
ADI
ZTS
ETN
ICE
SRE
CL
EW
EA
ISRG
INTU
SPGI
BIIB
ADBE
NFLX
NOC
GD
HUM
PEP
DE
LLY
FDX
SYK
APD
ADP
LMT
LOW
HCA
AMGN
PLD
DLR
SPG
AMT
ORCL
AVGO
TMUS
ABBV
SBUX
TSLA
AMD
FCX
NVDA
F
IP
GLW
NEM
HAL
FAST
CPRT
CMG
ANET
GE
BA
LIN
CCI
FIS
ILMN
EXPE
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Merge Distance

Σχήμα 7.3: Dendrogram του Agglomerative (Ward) για τις 127 εταιρείες. Οι εταιρείες είναι πολλές για να διαβαστούν μία μία, αλλά φαίνεται η ιεραρχία: το δέντρο σπάει σε λίγες μεγάλες ομάδες, και οι τράπεζες ενώνονται μεταξύ τους σε ένα κλαδί.

Μάλιστα, αν κόψουμε το δέντρο σε 4 ομάδες, και οι 7 τράπεζες (BAC, BK, C, GS, JPM, MS, SCHW) πέφτουν μαζί σε ένα κλαδί, ακόμα πιο καθαρά απ' ότι στον K-Means.

Σύγκριση μεθόδων μείωσης διαστάσεων: δοκιμάσαμε και τις τρεις μεθόδους (PCA, UMAP, t-SNE) για να δούμε ποια δείχνει καλύτερα τα clusters του K-Means. Το PCA έδειξε τις ομάδες πιο καθαρά, με τις συστάδες να μένουν μαζί και να ξεχωρίζουν πάνω στους δύο πρώτους άξονες. Γι' αυτό το χρησιμοποιήσαμε στο Σχήμα 7.2. Το UMAP, αντίθετα, σκόρπισε τις ίδιες συστάδες και ανακάτεψε τα χρώματα στο γράφημα. Αυτό συμβαίνει γιατί το UMAP αλλάζει τις αποστάσεις με μη γραμμικό τρόπο, οπότε ταιριάζει καλύτερα σε αλγορίθμους που ψάχνουν πυκνότητα, όχι στον K-Means. Το t-SNE ήταν το πιο αργό από τα τρία και ούτε αυτό βοήθησε να ξεχωρίσουν οι ομάδες. Έτσι, για αυτό το πείραμα, το PCA ήταν η καλύτερη επιλογή για οπτικοποίηση.



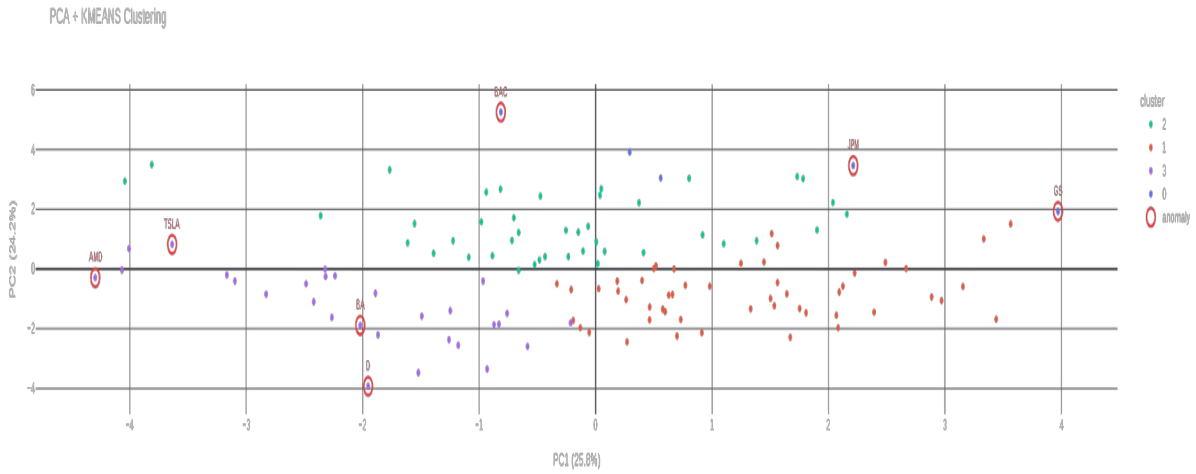
Σχήμα 7.4: Το ίδιο αποτέλεσμα (K-Means k=4) με UMAP. Οι ίδιες συστάδες εμφανίζονται σκορπισμένες, σε αντίθεση με το PCA του Σχήματος 7.2.

Πίσω στο αρχικό ερώτημα: το σύστημα ανακτά μόνο εν μέρει τους κλάδους χωρίς labels. Τις τράπεζες τις βρίσκει καθαρά, και αυτό γιατί ο ισολογισμός τους είναι τόσο διαφορετικός, με τεράστια assets και αρνητικό operating cash flow, που ξεχωρίζει αμέσως. Οι υπόλοιποι κλάδοι όμως δεν μπαίνουν σε δικές τους ομάδες. Οι τεχνολογικές για παράδειγμα σκορπίζονται, άλλες πάνε με τις mega-cap στη Συστάδα 2 και άλλες με τις ώριμες στη Συστάδα 1. Το ίδιο γίνεται και με την υγεία ή την ενέργεια. Αυτό που βλέπει στην πραγματικότητα ο αλγόριθμος δεν είναι ο κλάδος, αλλά το μέγεθος και το προφίλ συναλλαγών της κάθε εταιρείας. Δύο εταιρείες από τελείως διαφορετικούς κλάδους μπορεί να καταλήξουν μαζί, απλά επειδή έχουν παρόμοια έσοδα και assets.

7.4 Πείραμα 2: Εντοπισμός ανωμαλιών

Ερώτημα: ποιες εταιρείες έχουν ασυνήθιστο οικονομικό προφίλ; Χρησιμοποιήσαμε τον Isolation Forest (Step 5) με contamination 5% στον πλήρη 12-διάστατο χώρο, και μετά διασταυρώνουμε τα ευρήματα με τα noise points του DBSCAN.

Isolation Forest (contamination 5%): με 127 εταιρείες, το 5% βγάζει 7 ανωμαλίες. Χωρίζονται σε τρεις ομάδες. Πρώτον, τράπεζες (BAC, GS, JPM), που ξεχωρίζουν λόγω του ακραίου ισολογισμού τους, με τεράστια assets και cash και πολύ αρνητικό operating cash flow, που στην JPM φτάνει τα -5σ. Δεύτερον, εταιρείες υψηλής ανάπτυξης (AMD, TSLA), με πολύ μεγάλο price growth και υψηλό volume, και EPS που κάνει έντονες διακυμάνσεις. Τρίτον, δύο ειδικές περιπτώσεις, η Boeing (BA) που βγαίνει λόγω μεγάλων ζημιών (net income -4.2σ, operating cash flow -2.8σ), πιθανότατα από τα προβλήματα του 737 MAX και την πανδημία, και η Dominion (D) που ξεχωρίζει με ασυνήθιστα χαμηλά assets, cash και net income σε σχέση με τις υπόλοιπες.



Σχήμα 7.5: K-Means k=4 με PCA και Isolation Forest (contamination 5%). Οι 7 ανωμαλίες είναι σημειωμένες με κόκκινο κύκλο και το όνομά τους. Φαίνονται όλες στις άκρες του cloud, κυρίως τράπεζες (BAC, GS, JPM) και εταιρείες υψηλής ανάπτυξης (AMD, TSLA).

Διασταύρωση με DBSCAN: όπως είδαμε στο 7.3, ο DBSCAN δεν φτιάχνει κανονικές ομάδες σε αυτά τα δεδομένα, αλλά αφήνει 12 εταιρείες σαν noise. Αν δούμε ποιες βγαίνουν ανώμαλες και από τις δύο μεθόδους, μένουν έξι: AMD, BA, BAC, GS, JPM, TSLA. Αυτές έχουν πραγματικά ακραίο προφίλ, αφού τις πιάνουν δύο τελείως διαφορετικοί αλγόριθμοι.

Τα υπόλοιπα noise points του DBSCAN, όπως AAPL, NVDA, GE, είναι μάλλον αποτέλεσμα της δυσκολίας του σε υψηλές διαστάσεις, όχι πραγματικές ανωμαλίες. Γι' αυτό ο Isolation Forest είναι πιο αξιόπιστος εδώ [3].

Stability test K-Means k=4: τρέξαμε τον K-Means 10 φορές με διαφορετικό random_state. Οι 5 τράπεζες (BAC, C, GS, JPM, MS) έπεσαν στην ίδια συστάδα και τις 10 φορές.

Αυτό δείχνει ότι η τραπεζική υπογραφή στα δεδομένα είναι ισχυρή και δεν αλλάζει από την τυχαιότητα της αρχικοποίησης.

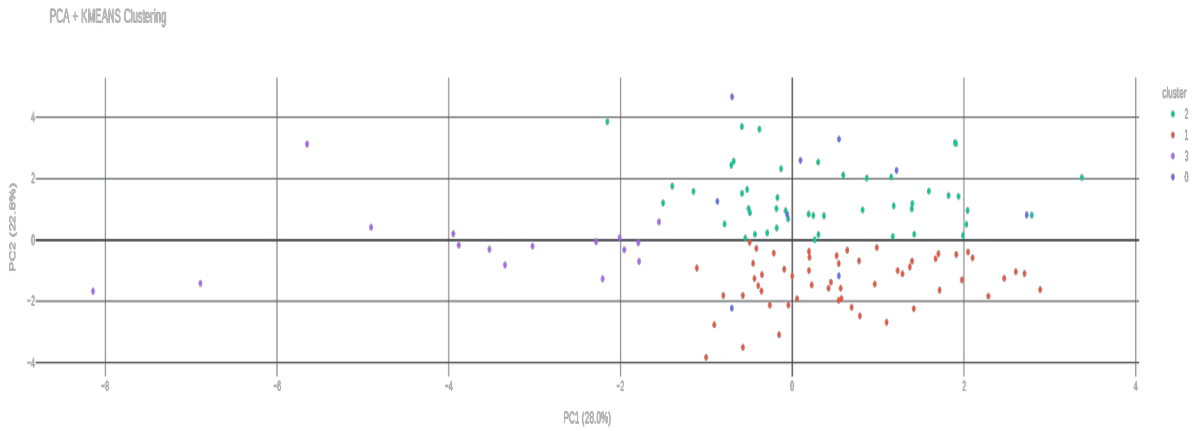
Πίσω στο ερώτημα: οι εταιρείες με το πιο ασυνήθιστο προφίλ είναι αυτές οι έξι. Δεν είναι «κακές» εταιρείες, απλά ξεχωρίζουν για συγκεκριμένους λόγους, οι τράπεζες λόγω ισολογισμού, η AMD και η TSLA λόγω ανάπτυξης, η Boeing λόγω ζημιών. Στο γράφημα φαίνονται όλες στις άκρες του cloud, μακριά από το κέντρο.

7.5 Πείραμα 3: Χρονική ανάλυση (Pre/During/Post-COVID)

Ερώτημα: πώς αλλάζει η δομή της αγοράς σε διαφορετικά οικονομικά περιβάλλοντα; Εκτελέστηκε K-Means k=4 ξεχωριστά σε τρεις περιόδους χρησιμοποιώντας το date range picker (Step 0) της εφαρμογής.

Pre-COVID (2017-2019)

127 εταιρείες, silhouette 0.178. Τέσσερις ομάδες: C0 (9), C1 (55), C2 (47), C3 (16). Η C0 είναι πάλι οι τράπεζες, και οι 7 (BAC, BK, C, GS, JPM, MS, SCHW), μαζί με δύο εταιρείες που εκείνη την περίοδο είχαν παρόμοιο προφίλ έντονου δανεισμού, τη NFLX και τη DE. Η C2 έχει τις μεγάλες κερδοφόρες εταιρείες, AAPL, MSFT, META, AMZN και άλλες. Η πιο ενδιαφέρουσα είναι η C3, μια μικρή ομάδα 16 εταιρειών με υψηλό price growth αλλά χαμηλά κέρδη. Εκεί μέσα βρίσκονται η NVDA και η TSLA, μαζί με AMD, HAL και άλλες μικρότερες. Δηλαδή πριν την πανδημία, η NVDA και η TSLA ήταν ακόμα στην ομάδα των μικρών, ανερχόμενων εταιρειών, όχι μαζί με τους κολοσσούς.

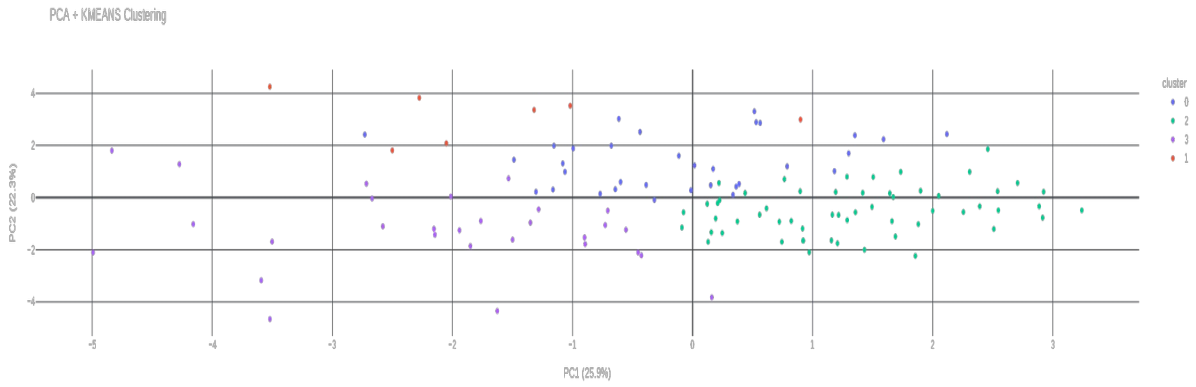


Σχήμα 7.6: K-Means k=4 με PCA, περίοδος Pre-COVID (2017-2019). Η NVDA και η TSLA βρίσκονται στη μικρή ομάδα των ανερχόμενων εταιρειών.

COVID (2020-2021)

126 εταιρείες (μία δεν είχε πλήρη δεδομένα για την περίοδο), silhouette 0.17, το χαμηλότερο από τις τρεις περιόδους. Τέσσερις ομάδες: C0 (33), C1 (7), C2 (57), C3 (29). Οι τράπεζες (C1) παραμένουν απομονωμένες, και μαζί τους μπαίνει η GM, που έχει δικό της χρηματοδοτικό βραχίονα. Η C0 είναι η ομάδα των μεγάλων κερδοφόρων, AAPL, MSFT, META, AMZN και άλλες, με υψηλό operating cash flow και volume. Η NVDA και η TSLA είναι ακόμα στη C3, την ομάδα των μικρότερων με χαμηλά

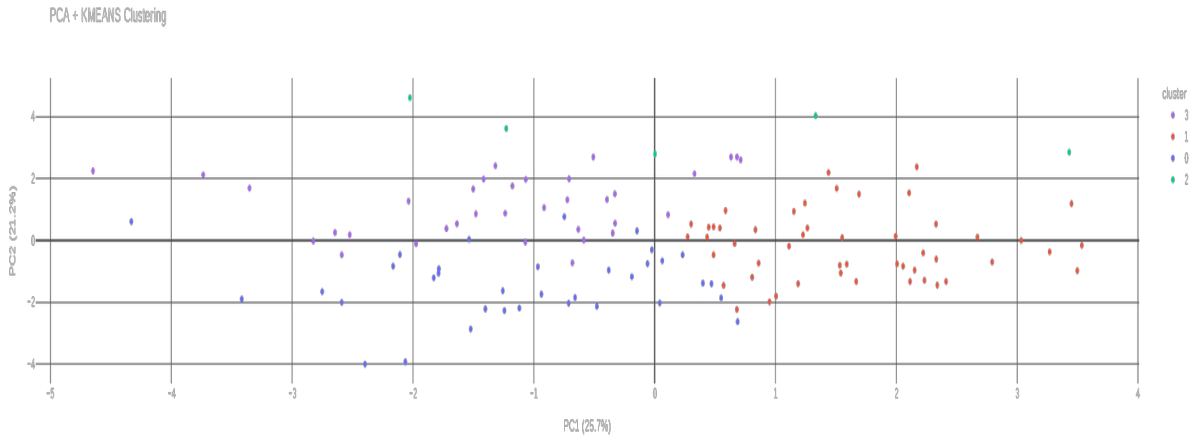
κέρδη και υψηλό price growth. Δηλαδή ούτε στην πανδημία δεν έχουν περάσει στους κολοσσούς. Ο χαμηλός silhouette δείχνει ότι η αναταραχή της περιόδου ανακάτεψε λίγο τη δομή.



Σχήμα 7.7: K-Means k=4 με PCA, περίοδος COVID (2020-2021). Η NVDA και η TSLA παραμένουν στην ομάδα των μικρότερων εταιρειών.

Post-COVID (2022-2024)

125 εταιρείες, silhouette 0.172. Τέσσερις ομάδες: C0 (35), C1 (50), C2 (5), C3 (35). Οι τράπεζες (C2) μένουν πάλι μόνες τους, και οι 5 (BAC, C, GS, JPM, MS). Το πιο σημαντικό όμως είναι η C3, μια ομάδα με υψηλό volume όπου βρίσκονται μαζί η NVDA, η TSLA, η AAPL, η AMZN και η GOOGL. Εδώ είναι η αλλαγή: η NVDA και η TSLA, που πριν ήταν στην ομάδα των μικρών, τώρα κάθονται δίπλα στους κολοσσούς. Η άνοδος της τεχνητής νοημοσύνης το 2023-2024 και η εκτόξευση της Nvidia φαίνονται και στα fundamentals, όχι μόνο στην τιμή της μετοχής.



Σχήμα 7.8: K-Means k=4 με PCA, περίοδος Post-COVID (2022-2024). Η NVDA και η TSLA έχουν περάσει στην ομάδα με υψηλό volume, μαζί με AAPL, AMZN και GOOGL.

Εταιρεία	Pre-COVID(2017-19)	COVID(2020-21)	Post-COVID(2022-24)
NVDA	μικρές / ανερχόμενες	μικρές / ανερχόμενες	κολοσσοί(υψηλό volume)
TSLA	μικρές / ανερχόμενες	μικρές / ανερχόμενες	κολοσσοί(υψηλό volume)
AAPL	μεγάλες κερδοφόρες	μεγάλες κερδοφόρες	κολοσσοί(υψηλό volume)
AMZN	μεγάλες κερδοφόρες	μεγάλες κερδοφόρες	κολοσσοί(υψηλό volume)
MSFT	μεγάλες κερδοφόρες	μεγάλες κερδοφόρες	Ώριμες κερδοφόρες
JPM	τράπεζες	τράπεζες	τράπεζες

Πίνακας 7.3: Κίνηση εταιρειών ανάμεσα στις ομάδες στις τρεις περιόδους (K-Means k=4, PCA, 12 features). Οι ομάδες περιγράφονται με βάση τον χαρακτήρα τους, γιατί τα νούμερα των clusters αλλάζουν σε κάθε run.

Από τον πίνακα βγαίνουν τρία πράγματα. Πρώτον, η μετανάστευση της NVDA και της TSLA. Και στις δύο πρώτες περιόδους ήταν στην ομάδα των μικρών, ανερχόμενων εταιρειών.

Post-COVID πέρασαν στους κολοσσούς, δίπλα στην AAPL και την AMZN. Η άνοδος της AI το 2023-2024 άλλαξε τα fundamentals τους, όχι μόνο την τιμή της μετοχής. Δεύτερον, οι τράπεζες (JPM, BAC, GS, MS, C) έμειναν στη δική τους ομάδα και στις τρεις περιόδους, η πιο σταθερή υπογραφή στα δεδομένα. Τρίτον, μετά την πανδημία η μεγάλη ομάδα των tech σπάει στα δύο, η AAPL και η AMZN πάνε με τις εταιρείες υψηλού volume, ενώ η MSFT και η META μένουν στις ώριμες κερδοφόρες.

Περιορισμοί

Τα αποτελέσματα έχουν κάποιους περιορισμούς που καλό είναι να τους κρατάμε στο μυαλό μας.

Ο silhouette είναι χαμηλός σε όλες τις περιόδους, γύρω στο 0.17 με 0.18. Αυτό δεν σημαίνει ότι το clustering απέτυχε. Δείχνει ότι οι εταιρείες δεν χωρίζονται σε καθαρές, ξεχωριστές ομάδες, αλλά απλώνονται σε ένα συνεχές φάσμα. Η μία περνάει ομαλά στην άλλη.

Υπάρχει και survivorship bias. Το dataset έχει 127 εταιρείες που είναι σήμερα στον S&P 500. Όσες βγήκαν από τον δείκτη μέσα στην περίοδο δεν μπαίνουν στην ανάλυση, οπότε η εικόνα είναι λίγο πιο αισιόδοξη από την πραγματικότητα.

Ένας ακόμα περιορισμός είναι ο τρόπος που φτιάχνουμε τα σημεία. Παίρνουμε τον μέσο όρο κάθε εταιρείας μέσα σε όλη την περίοδο και τον κάνουμε ένα σημείο. Έτσι χάνονται οι αλλαγές που γίνονται μέσα στην ίδια περίοδο. Αν μια εταιρεία άλλαξε πολύ μέσα στο 2022-2024, εμείς βλέπουμε μόνο τον μέσο όρο της.

Τέλος, οι περίοδοι δεν είναι καθαρές. Μεγάλα οικονομικά γεγονότα, όπως η αύξηση των επιτοκίων το 2022, χτυπούν πολλές εταιρείες ταυτόχρονα. Οπότε όταν μια εταιρεία αλλάζει ομάδα, δεν μπορούμε να το αποδώσουμε μόνο στην πανδημία.

Πίσω στο ερώτημα: η δομή της αγοράς όντως αλλάζει ανάλογα με την περίοδο. Κάποια πράγματα μένουν σταθερά, όπως οι τράπεζες, αλλά άλλα μετακινούνται καθαρά, όπως η NVDA και η TSLA που ανέβηκαν κατηγορία. Τα fundamentals δηλαδή δεν είναι στατικά, ακολουθούν τις πραγματικές αλλαγές στις εταιρείες και στην οικονομία.

Επίλογος

Από τα τρία πειράματα κρατάμε τα εξής. Το σύστημα δεν ανακτά όλους τους κλάδους, αλλά βρίσκει πολύ καθαρά τις τράπεζες, μόνο από τα fundamentals και χωρίς labels. Στον εντοπισμό ανωμαλιών, ο Isolation Forest ξεχώρισε εταιρείες με πραγματικά ακραίο προφίλ, και η διασταύρωση με τον DBSCAN έδειξε ποιες από αυτές είναι οι πιο αξιόπιστες. Η χρονική ανάλυση έδειξε ότι τα fundamentals αλλάζουν με τον χρόνο, με πιο χαρακτηριστική τη μετανάστευση της NVDA από τις μικρές εταιρείες στους κολοσσούς.

Το πιο απρόσμενο εύρημα ήταν αυτή ακριβώς η μετανάστευση. Στην αρχή περιμέναμε η ομάδα της τεχνολογίας να είναι σταθερή διαχρονικά. Το ότι η Nvidia άλλαξε ομάδα ανάμεσα στο 2019 και το 2024 δείχνει ότι τα fundamentals πιάνουν πραγματικές αλλαγές, όχι απλώς θόρυβο. Το επόμενο κεφάλαιο συνοψίζει τα συμπεράσματα.

Κεφάλαιο 8ο: Συμπεράσματα και μελλοντικές επεκτάσεις

8.1 Εισαγωγή

Το κεφάλαιο αυτό συνοψίζει τα κύρια ευρήματα της εργασίας και αξιολογεί κατά πόσο επιτεύχθηκαν οι στόχοι που τέθηκαν στο Κεφάλαιο 1. Παρουσιάζονται επίσης κατευθύνσεις για μελλοντική επέκταση του StockLens.

8.2 Συμπεράσματα

Ο σχεδιασμός και η ανάπτυξη του StockLens ολοκληρώθηκαν με επιτυχία. Τα συμπεράσματα οργανώνονται σε τρεις κατηγορίες.

Τεχνολογικά

Ο συνδυασμός Python, Dash και PostgreSQL/TimescaleDB λειτουργεί καλά για real-time data analysis. Η TimescaleDB ταίριαξε πολύ καλά με τα time-series δεδομένα: τα queries range-by-date εκτελούνται ταχύτερα λόγω automatic chunking. Η ασύγχρονη αρχιτεκτονική με asyncpg και nest_asyncio διατηρεί το UI αποκριτικό κατά την εκτέλεση βαρέων queries. Το Repository Pattern απλοποίησε τη συντήρηση. Το pipeline κανονικοποίησης (signed log + StandardScaler) δούλεψε καλά για χρηματοοικονομικά δεδομένα που καλύπτουν πολλές τάξεις μεγέθους.

Αλγοριθμικά

Η σύγκριση των αλγορίθμων στα πραγματικά δεδομένα έβγαλε σαφή ευρήματα. Ο K-Means με k=4 (silhouette 0.178) έδωσε τα πιο ερμηνεύσιμα αποτελέσματα, με πιο ισχυρό εύρημα την ομάδα των τραπεζών (5 μετοχές) που ξεχωρίζει μόνη της. Ο GMM με k=4 έβγαλε σχεδόν τον ίδιο silhouette (0.178), δηλαδή παρόμοιο αποτέλεσμα. Ο Agglomerative με Ward(0.138) είχε λίγο χειρότερο διαχωρισμό, αλλά το dendrogram του είναι χρήσιμο για την ιεραρχία και μάλιστα ομαδοποιεί καθαρά και τις 7 τράπεζες. Ο DBSCAN δεν ταίριαξε σε αυτά τα δεδομένα: με ε γύρω στο 3 βάζει σχεδόν τα πάντα σε μία ομάδα και αφήνει λίγα noise points, λόγω του curse of dimensionality. Ως προς τη μείωση διαστάσεων, για τα clusters του K-Means το PCA έδωσε την πιο καθαρή εικόνα, ενώ το UMAP ταιριάζει καλύτερα σε αλγορίθμους που ψάχνουν πυκνότητα.

Χρηματοοικονομικά

Τα αποτελέσματα επιβεβαίωσαν ότι τα θεμελιώδη μεγέθη εμπεριέχουν αρκετή πληροφορία για οικονομικά ερμηνεύσιμες ομαδοποιήσεις:

- Κλαδική αναγνώριση χωρίς labels: ο K-Means εντόπισε την ομάδα των τραπεζών μόνο από τη δομή των assets και των cash flows, με χαρακτηριστικό το έντονα αρνητικό operating cash flow [1].
- Υπέρβαση κλαδικών ορίων: εταιρείες από διαφορετικούς κλάδους που πέφτουν στην ίδια ομάδα έχουν παρόμοιο οικονομικό προφίλ. Αυτό έχει πρακτική αξία για τη διαφοροποίηση χαρτοφυλακίου [2].
- Χρονική εξέλιξη: η μετακίνηση της NVDA προς τους τεχνολογικούς κολοσσούς μετά την πανδημία αντικατοπτρίζει πραγματική αλλαγή στα fundamentals, και δείχνει τη χρησιμότητα της χρονικής

8.3 Μελλοντικές επεκτάσεις

Οι σημαντικότερες κατευθύνσεις επέκτασης:

Εμπλουτισμός δεδομένων

Η πιο προφανής κατεύθυνση είναι ο εμπλουτισμός των features με μακροοικονομικούς δείκτες (επιτόκια Fed, VIX, GDP growth) και η διεύρυνση περισσότερων από 127 μετοχών. Η ενσωμάτωση sentiment score από news feeds θα ήταν ενδιαφέρουσα, αν και το sentiment μπορεί να είναι ένα θορυβώδες feature.

Αλγοριθμικές επεκτάσεις

- Deep Learning: Αντικατάσταση PCA/UMAP με Autoencoder νευρωνικό δίκτυο για μη γραμμικές αναπαραστάσεις.
- Temporal Clustering: Χρήση LSTM/Transformer για μοντελοποίηση χρονικής αλληλουχίας. Ο αλγόριθμος θα μαθαίνει «trajectories» αντί για στατικά snapshots.
- Ensemble Clustering: Συνδυασμός αποτελεσμάτων πολλαπλών αλγορίθμων για πιο robust ομαδοποίηση.
- SHAP values: Εξήγηση ποια features οδήγησαν κάθε μετοχή στη συγκεκριμένη συστάδα.

Επεκτάσεις πλατφόρμας

- Real-time Alerts: Ειδοποιήσεις όταν μια μετοχή αλλάζει συστάδα ή εντοπίζεται ως anomaly.
- Backtesting Module: Αξιολόγηση επενδυτικών στρατηγικών βάσει cluster membership ιστορικά.
- API Endpoints: Εξαγωγή clustering αποτελεσμάτων μέσω REST API για ενσωμάτωση σε άλλα συστήματα.

Επίλογος

Η εφαρμογή StockLens υλοποίησε ένα πλήρες σύστημα ανάλυσης χρηματιστηριακών δεδομένων: από τη συλλογή από δύο πηγές, μέχρι την ανάλυση με αλγορίθμους ML και την παρουσίαση σε διαδραστικό dashboard. Το κεντρικό εύρημα είναι ότι τα fundamentals εμπεριέχουν δομή που αντικατοπτρίζει κλαδικά και λειτουργικά χαρακτηριστικά εταιρειών, κάτι που φαίνεται μόνο όταν εξεταστούν όλες μαζί. Αυτό αποτελεί σημείο εκκίνησης για περαιτέρω έρευνα στην κατεύθυνση temporal clustering και ενσωμάτωσης alternative data.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] B. Graham and D. Dodd, Security Analysis. New York, NY: McGraw-Hill, 1934.

[2] B. G. Malkiel, A Random Walk Down Wall Street. New York, NY: W. W. Norton & Company, 2019.

[3] M. López de Prado, Advances in Financial Machine Learning. Hoboken, NJ: Wiley, 2018.

[4] J. J. Murphy, Technical Analysis of the Financial Markets. New York, NY: New York Institute of Finance, 1999.

[5] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[6] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579–2605, 2008.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Series B, vol. 39, no. 1, pp. 1–38, 1977.

[8] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf.

Knowledge Discovery and Data Mining (KDD-96), Portland, OR, 1996, pp. 226–231.

[9] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018. [Online]. Available: https://arxiv.org/abs/1802.03426

[10] Timescale Inc., "Time-series data management with TimescaleDB," 2024. [Online]. Available: https://docs.timescale.com