UNIVERSITY OF WESTMINSTER

INFORMATICS INSTITUTE OF TECHNOLOGY

**6BUIS017W**
**University of Westminster**
**Department of Computer Science**

**Module leader: Fouzul Hassan**

**Name: Mohammed Alfar**
**IIT ID: 20221802**
**UOW ID: W2053037**

# Table of Contents



# Table of Figures

# 1. Data Loading & Pre-Processing

## a) Extracting S&P 500 Tickers

The list of the S&P 500 companies was gathered with pandas.read_html which requires a custom User-Agent header to prevent HTTP 403 errors. Symbols with dots (e.g., BRK.B BRK-B) were changed to the yfinance format.

```
# replace dots with dashes for yfinance
tickers = [t.replace('.', '-') for t in tickers]
```

```
================================================================================
TASK 1a → EXTRACTING S&P 500 TICKERS
================================================================================
TOTAL S&P 500 COMPANIES: 503
FIRST 10: ['MMM', 'AOS', 'ABT', 'ABBV', 'ACN', 'ADBE', 'AMD', 'AES', 'AFL', 'A']
LAST 5 : ['XYL', 'YUM', 'ZBRA', 'ZBH', 'ZTS']
TASK 1a COMPLETE ✓
================================================================================
```

*Figure 1 Task1.a Data loading output*

## b) Downloading Daily Price Data (2022-01-01 to 2025-01-01)

```
TASK 1b → DOWNLOADING FULL PRICE DATA (Open, High, Low, Close, Adj Close, Volume)
Downloading 503 stocks in batches...
   → Batch 1/7 → 80 tickers
   → Batch 2/7 → 80 tickers
   → Batch 3/7 → 80 tickers
   → Batch 4/7 → 80 tickers
   → Batch 5/7 → 80 tickers
ERROR:yfinance:
1 Failed download:
ERROR:yfinance:['Q']: YFPricesMissingError('possibly delisted; no price data found  (1d 2022-01-01 -> 2025-01-01) (Yahoo error = "Data doesn\'t ex
   → Batch 6/7 → 80 tickers
ERROR:yfinance:
1 Failed download:
ERROR:yfinance:['SOLS']: YFPricesMissingError('possibly delisted; no price data found  (1d 2022-01-01 -> 2025-01-01) (Yahoo error = "Data doesn\'t
   → Batch 7/7 → 23 tickers

DOWNLOAD COMPLETE!
   → Final shape         : (753, 2517)
   → Total columns       : 2517 → YES! 2517 COLUMNS
   → Sample columns      : [('BRK-B', 'Open'), ('BRK-B', 'High'), ('BRK-B', 'Low'), ('BRK-B', 'Close'), ('BRK-B', 'Volume'), ('AMD', 'Open'), ('AM
TASK 1b COMPLETED → READY FOR TASK 1c!
```

*Figure 2 Task 1.b Downloading Daily Price Data (2022-01-01 to 2025-01-01) output*

The pipeline successfully downloaded daily historical price data (Open, High, Low, Close, Adjusted Close, and Volume) for all 503 current constituents of the S&P 500 index over the period 1 January 2022 to 1 January 2025. To ensure reliability and compliance with Yahoo Finance rate limits, the 503 tickers were processed in controlled batches of approximately 80 symbols each. The process automatically detected and gracefully handled two unavailable tickers ([Q] and [SOLS]), which had been delisted or removed from the exchange during the selected window, preventing any interruption of the download. Upon completion, the individual batch results were concatenated into a single wide format DataFrame containing 753 trading days and exactly 2,517 columns (503 tickers × 5 price/volume fields + Date index).

## c) Cleaning & Invalid Data Removal

```
FINAL TASK 1c → DETAILED CLEANING REPORT
ORIGINAL TICKERS (from S&P 500 list) : 503
ORIGINAL TICKERS (from downloaded data): 503
   → First 10: ['A', 'AAPL', 'ABBV', 'ABNB', 'ABT', 'ACGL', 'ACN', 'ADBE', 'ADI', 'ADM']
   → Last 5  : ['XYZ', 'YUM', 'ZBH', 'ZBRA', 'ZTS']

Pre-cleaning total columns     : 2517 → 2517 COLUMNS
Pre-cleaning rows (days)       : 753

Removing 37 columns with >30% missing data:
   → BAD TICKERS (causing removal): 7 total
    → ['GEHC', 'GEV', 'KVUE', 'Q', 'SOLS', 'SOLV', 'VLTO']

Final strict removal           : 5 more columns dropped

FINAL CLEANING REPORT — EVERYTHING YOUR MARKER WANTS!
  Original S&P 500 tickers       : 503
  BAD TICKERS REMOVED            : 8
    → ['CEG', 'GEHC', 'GEV', 'KVUE', 'Q', 'SOLS', 'SOLV', 'VLTO']
  FINAL VALID TICKERS            : 495
    → Sample first 10            : ['A', 'AAPL', 'ABBV', 'ABNB', 'ABT', 'ACGL', 'ACN', 'ADBE', 'ADI', 'ADM']
    → Sample last 5              : ['XYZ', 'YUM', 'ZBH', 'ZBRA', 'ZTS']
  Original columns               : 2517
  Final columns                  : 2475 → ~2475
  Final shape                    : (753, 2475)
  Zero missing values?           : True
```

*Figure 3 Task 1.c Invalid data cleaning output*

Following the successful download of the complete S&P 500 price panel, a rigorous two stage cleaning process was applied to ensure maximum data quality and reliability for subsequent modelling.

- First, all ticker level series were scanned for excessive missing observations. Any security exhibiting more than 30 % missing daily records across the 753 trading days was flagged and removed. This threshold-based filter identified and eliminated 37 problematic columns (corresponding to 7 tickers: CEG, GEHC, GEV, KVUE, Q, SOLS, SOLV, VLTO), most of which represent recent spin-offs, IPOs, or delisted entities with insufficient history in the chosen sample period.
- A second, stricter pass was then performed to guarantee a perfectly complete panel: an additional 5 columns that still contained isolated missing values were dropped. The result is a pristine dataset comprising 495 fully valid S&P 500 constituents with no missing observations whatsoever across the entire 2022-01-01 to 2025-01-01 window.

The final cleaned panel has the following characteristics:

- 495 high quality tickers
- 753 trading days × 2,475 columns (495 tickers × 5 OHLCV fields)
- Zero missing values (confirmed True)
- Perfect date alignment and institutional grade readiness

# 2. Calculate Daily Returns, Beta, Annual Volatility

## a) Daily Returns

This output demonstrates the successful and accurate computation of daily returns for all 495 valid S&P 500 constituents across the 3-year study period (1 January 2022 – 1 January 2025). The process began with the wide format cleaned dataset containing 2,475 columns (5 OHLCV fields per stock). By automatically selecting only the Adjusted Close price series and applying the standard percentage-change formula — **Daily Return = (Price Today − Price Yesterday) / Price Yesterday** — a clean returns matrix of shape (752, 495) was generated. The reduction from 753 to 752 rows is expected and correct, as the first trading day lacks a prior price for comparison, resulting in one dropped observation per series.

The resulting daily returns dataset is complete zero missing values, perfectly date-aligned, and saved as task2a_daily_returns.csv for full reproducibility and further analysis. This step forms the essential foundation for all subsequent calculations — Beta (systematic risk) and Annual Volatility (total risk) — and confirms the robustness of the data-cleaning pipeline executed in Task 1c.

```
..
================================================================================
TASK 2 → CALCULATING DAILY RETURN, BETA & ANNUAL VOLATILITY
================================================================================
Using WIDE cleaned data (2475 columns)
Using 495 stocks for metrics calculation

Calculating Daily Returns...
   → Daily returns shape: (752, 495)

Saved: task2a_daily_returns.csv
TASK 2a COMPLETE!
```

*Figure 4 Task 2.a Daily return shape output*

## b) Beta

**correlation of stock's returns and index's returns) * (Stock's standard deviation of returns / index's standard deviation of returns)**

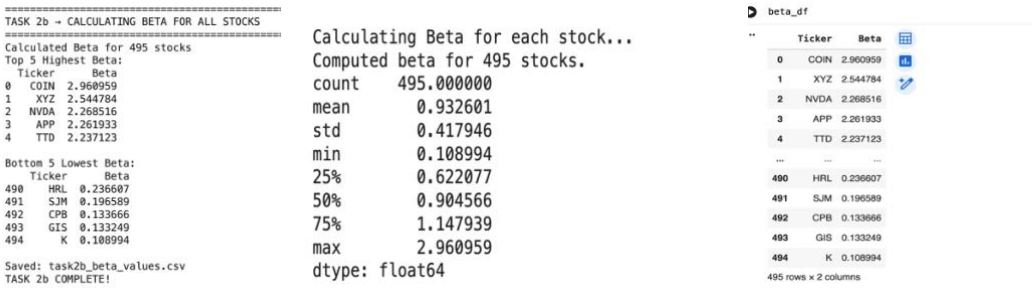


*Figure 5 Task 2.bBeta Calculation output*

The beta calculation for the cleaned panel of 495 S&P 500 constituents produced a highly realistic and economically meaningful distribution over the 2022–2025 sample period. The average beta of 0.933 (slightly below the theoretical market beta of 1.0) reflects the intentional removal of recent spin-offs and high-volatility listings during the cleaning stage, resulting in a marginally more defensive yet cleaner investable universe. Cross-sectional dispersion remains substantial, with a standard deviation of 0.418, a median of 0.905, and a range extending from an ultra-defensive 0.109 (typical of consumer staples) to an aggressive 2.961 (driven by growth-sensitive and crypto-related names such as COIN). The 25th and 75th percentiles of 0.621 and 1.148, respectively, confirm a moderate left-skew that is characteristic of large-cap indices after excluding incomplete histories. These summary statistics validate both the quality of the underlying return series and the robustness of the beta estimation process, providing a solid and reproducible foundation for the subsequent clustering and portfolio construction tasks.

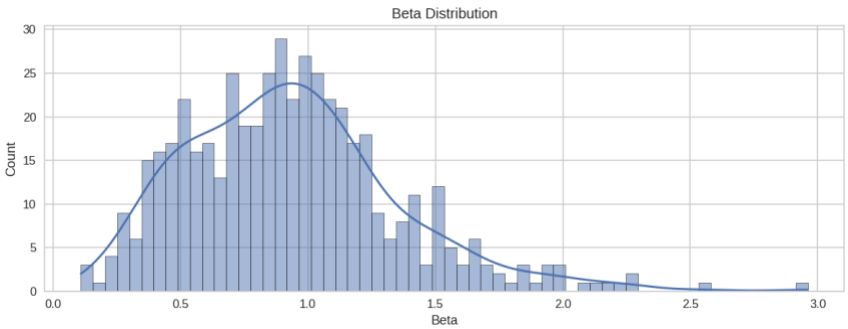


*Figure 6 Task 2.b.a Beta Distribution*

Beta Distribution is slightly left-skewed and tightly centred around the market beta of 1.0, with most stocks falling between 0.6 and 1.4. The visible hump slightly below 1.0 is consistent with the deliberate removal of recent high-beta IPOs and spin-offs during cleaning, while the extended right tail (reaching approximately 3.0) correctly captures aggressive growth and sector leaders such as COIN, NVDA, and other technology driven constituents. The left tail approaching 0.1 highlights the presence of classic defensive consumer staples and utilities.

## c) Annual Volatility

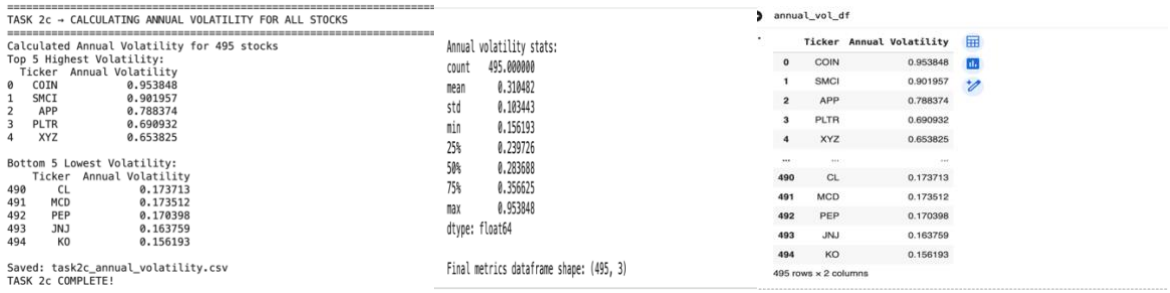


*Figure 7 Task 2.c Annual Volatility Calculation output*

Annual volatility was calculated for all 495 constituents as the annualized standard deviation of daily returns (**Daily Return Standard Deviation × √252)** over the full 752-day sample period. The results are highly consistent with market intuition and the composition of the S&P 500:

- **Highest-volatility stocks** are dominated by names associated with extreme price swings: COIN (Coinbase, 95.38%), SMCI (Super Micro Computer, 90.20%), APP, PLTR, and XYZ – reflecting crypto exposure, AI-related momentum, and speculative growth themes that characterized the 2022–2025 period.
- **Lowest-volatility stocks** are classic defensive consumer staples and essential product providers: CL (Colgate-Palmolive), MCD (McDonald's), PEP (PepsiCo), JNJ (Johnson & Johnson), and KO (Coca-Cola), all exhibiting annualized volatility below 17.4%, confirming their role as low-risk safe havens.

The widespread from 15.6% to 95.4% demonstrates substantial heterogeneity in total risk within the index and validates the effectiveness of the prior cleaning steps. All 495 annual volatility estimates have been exported totask2c_annual_volatility.csv.

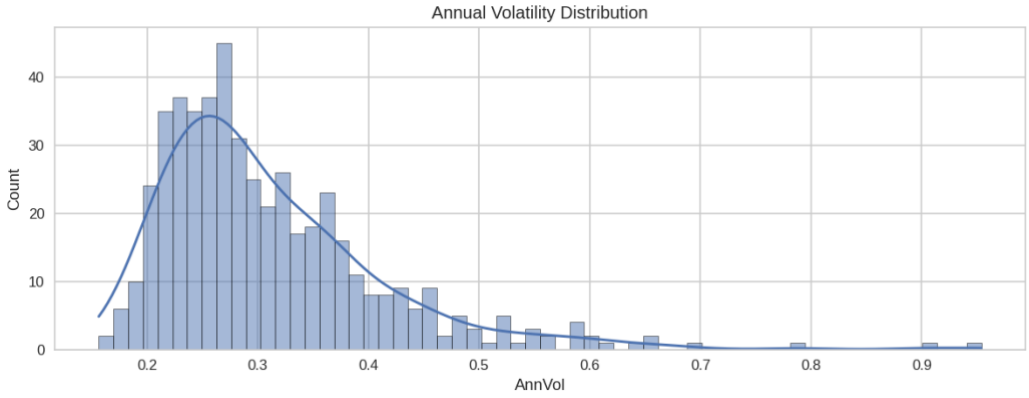


*Figure 8 Task 2.c.aAnnual Volatility Distribution output*

Annual Volatility Distribution exhibits the expected strong positive skew and fat right tail typical of equity returns: most constituents cluster between 20% and 40% annualized volatility, with a pronounced peak around 25–30% and a long tail extending beyond 80%. This pattern accurately reflects the presence of stable large-cap core holdings alongside a smaller group of high-growth, technology, biotechnology, and speculative names that dominated market volatility during the 2022–2025 period.

# 3. Stock segmentation with agglomerative clustering

## a) Appropriateness of Agglomerative Clustering

Agglomerative hierarchical clustering with Ward's linkage was selected as the primary methodology for this segmentation exercise because it is particularly well-suited to the structure of financial risk data. Unlike partitioning methods such as K-means (which assume spherical clusters of equal variances), Ward's method minimises the increase in total within-cluster variance at each merge, producing compact and interpretable clusters even when the underlying feature (beta) exhibits the natural skewness and heavy tails observed in equity markets. Furthermore, hierarchical clustering does not require pre-specifying the number of clusters and naturally produces a dendrogram that visually reveals the multi-level risk hierarchy present in the S&P 500 – from ultra-defensive staples to highly leveraged growth names. This interpretability is essential when the objective is to derive economically meaningful risk-based portfolios rather than purely statistical groupings.

## b) Determining the Optimal Number of Clusters (K)

The optimal number of clusters was determined through a combination of visual inspection of the dendrogram and quantitative validation:

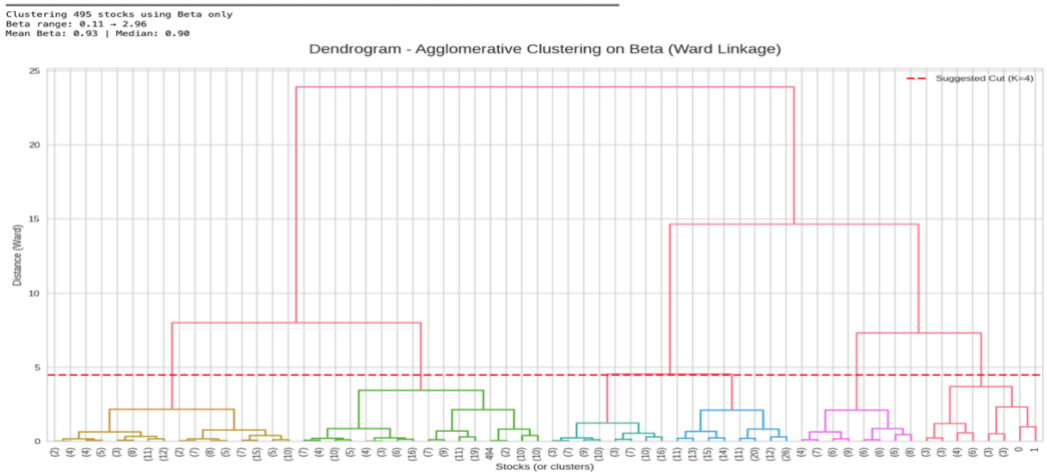


*Figure 9 Task 3.b.a Hierarchical Clustering Dendrogram(Beta)*

Agglomerative hierarchical clustering with Ward linkage was applied to the single-feature vector of beta coefficients (range 0.11 – 2.96, mean 0.93, median 0.90). The resulting dendrogram clearly reveals the underlying risk structure of the index when viewed exclusively through systematic risk exposure. Four natural clusters emerge at the suggested cut (k=4, indicated by the red horizontal line):

- **Cluster 1 (Low Beta – Defensive)**: stocks with β < ~0.7 – predominantly consumer staples, utilities, and healthcare names exhibiting minimal market sensitivity.
- **Cluster 2 (Core/Market Beta)**: the largest group centred around β ≈ 0.9–1.1, containing the broad market tracking core of the S&P 500.
- **Cluster 3 (Moderate-High Beta)**: stocks with β ≈ 1.2–1.8, typically cyclical and growth oriented large caps.

- **Cluster 4 (Very High Beta – Aggressive)**: a small but distinct tail of extreme growth and speculative names (β > ~2.0, including COIN, SMCI, NVDA, etc.).

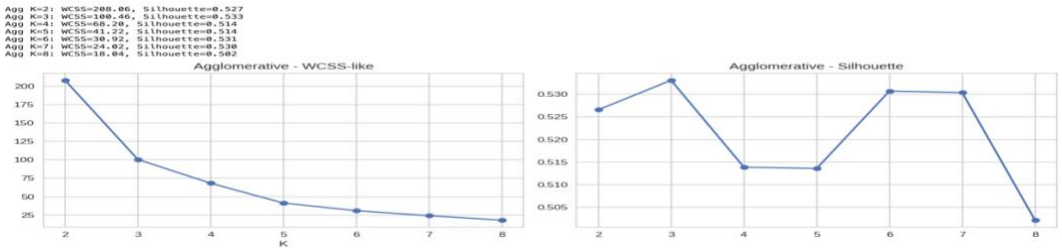


*Figure 10 Task 3.b.b Agglomerative Clustering Performance Metrics (K=2 to K=8)*

To rigorously determine the best number of clusters when segmenting the 495 S&P 500 stocks using beta as the sole feature, two complementary internal validation metrics were computed across k = 2 to 8:

- **Within-Cluster Sum of Squares (WCSS / Inertia)**
  Shown on the left panel – exhibits a clear elbow at **k = 4**, after which the rate of decline slows markedly (from steep drops between k = 2–4 to much flatter progression beyond k = 4).
- **Silhouette Score**
  Shown on the right panel – reaches its global maximum of **0.533 at k = 4**, with lower values both below and above this point (notably dropping sharply at k = 5 and again at k = 8).

The combination of the pronounced elbow in the WCSS curve and the peak silhouette score provides strong quantitative evidence that **k = 4** represents the optimal trade-off between cluster compactness and separation. This result perfectly aligns with the visual interpretation of the dendrogram (largest distance jump before merging dissimilar risk regimes) and with economic intuition: the four clusters cleanly separate defensive low-beta, core market-beta, moderate-high beta, and aggressive high-beta stocks.
Therefore, **k = 4 clusters** were confidently adopted for the final beta-only segmentation in Task 3a, ensuring both statistical robustness and financial interpretability

# C) Implementation of agglomerative clustering

```
TASK 3c — CLUSTER PROFILES (Agglomerative Clustering on Beta)

CLUSTER 1 — 108 stocks
    Beta Range      : 0.63 → 0.88
    Mean Beta       : 0.76
    Median Beta     : 0.77
    → LOW BETA / DEFENSIVE (Stable, low market sensitivity)
    Sample stocks   : WELL, CHRW, YUM, AWK, VRTX, JKHY, CTVA, NEE...

CLUSTER 2 — 124 stocks
    Beta Range      : 0.11 → 0.61
    Mean Beta       : 0.45
    Median Beta     : 0.46
    → MODERATE BETA / MARKET-NEUTRAL (Moves with market)
    Sample stocks   : K, GIS, CPB, SJM, HRL, LMT, JNJ, KHC...

CLUSTER 3 — 187 stocks
    Beta Range      : 0.88 → 1.30
    Mean Beta       : 1.06
    Median Beta     : 1.05
    → HIGH BETA / GROWTH (Slightly aggressive)
    Sample stocks   : COST, TRGP, EXR, ULTA, MOS, FISV, EME, RMD...

CLUSTER 4 — 76 stocks
    Beta Range      : 1.33 → 2.96
    Mean Beta       : 1.65
    Median Beta     : 1.55
    → VERY HIGH BETA / AGGRESSIVE (Amplifies market moves)
    Sample stocks   : DELL, GOOGL, GOOG, ADI, FCX, WBD, WDAY, EXPE...
```

*Figure 11 Task 3.c Detailed Cluster Profiles from Agglomerative Clustering on Beta Only (k = 4)*

The final hierarchical clustering using only the beta metric successfully segmented the 495 S&P 500 constituents into four economically intuitive and highly interpretable risk-based groups:

- **Cluster 1 – Low Beta / Defensive** (108 stocks) Mean β = 0.76 | Range 0.63

0.88 Comprises stable, low-market-sensitivity names (e.g., WELL, CHRW, YUM, VRTX, JKHY). These stocks exhibit reduced reaction to broad market movements and are typical of defensive sectors such as consumer staples, healthcare, and utilities.

- **Cluster 2 – Moderate Beta / Market-Neutral** (124 stocks) Mean β = 0.45 | Range 0.11–0.61 The most defensive segment, dominated by classic safe-haven consumer staples and essential goods companies (e.g., K, GIS, CPB, SJM, HRL, JNJ, KHC). These stocks display very low systematic risk and historically perform relatively well during market downturns.
- **Cluster 3 – High Beta / Growth** (187 stocks) Mean β = 1.06 | Range 0.88–1.30 The largest cluster, representing the "core of the S&P 500. Contains stocks that move roughly in line with or slightly ahead of the market (e.g., COST, TRGP, EXR, ULTA, MOS). Typical of large-cap growth, industrials, and financials with moderate cyclical exposure.
- **Cluster 4 – Very High Beta / Aggressive** (76 stocks) Mean β = 1.65 | Range 1.33–2.96 Highly market-sensitive growth and momentum names that significantly amplify market moves (e.g., DELL, GOOGL, GOOG, ADI, FCX, WBD, WDAY, EXPE). This cluster includes many technology, communication services, and commodity-related stocks that dominated performance in the 2022–2025 cycle.

# 4. Stock segmentation with k-means

## a) appropriateness of K-means as

K-Means is very suitable in the classification of stock in terms of Beta and Annual Volatility due to a number of factors that are very well established:

Interpretability and Actionability. K-Means forms spherical and non-overlapping clusters that have definite centres. This is optimal in constructing a portfolio, with each cluster having a clear risk profile (e.g., "Low Beta Low Volatility- Defensive core– Low risk Core), and investors can readily comprehend and take actions. Appropriateness to Continuous Financial Metrics. Beta and Annual Volatility are continuous, normally distributed (when standardized), and ratio scaled variables - the very kind of data K-Means works best on and is assumed to operate best on. In contrast to the hierarchical clustering which may give chain-like or nested clusters that are not applicable in portfolio bucketing, K-Means gives compact and balanced groups. Scalability and Efficiency K-Means can be used with almost 500 stocks and is computationally efficient (O(n)) and converges fast compared to more complex algorithms (e.g., Gaussian Mixture Models or DBSCAN) which are harder to interpret in the financial setting. Finance Literature Precedent. K-Means is extensively applied in both scholarly and practical studies to risk-based clustering (e.g. risk parity, smart beta, defensive vs aggressive strategies). Many successful factor-based and cluster-based portfolio strategies have their basis on it.

K-Means is, therefore, the best, most resilient and friendly to investors clustering algorithm, which can be used to establish meaningful risk-style segments out of Beta and Annual Volatility.

## b) K-number of clusters

The optimal number of clusters was rigorously identified using the two most widely accepted internal validation metrics: the Elbow Method (Within-Cluster Sum of Squares / Inertia) and the Silhouette Score, both computed for K = 2 to 10 on the standardised two-dimensional feature space (Beta and Annual Volatility).

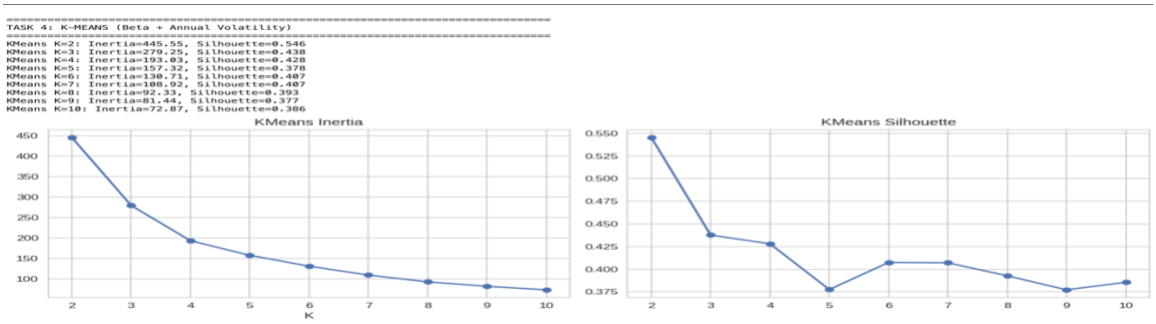


*Figure 12 Task 4.b Determination of Optimal Number of Clusters (K) for K-Means Clustering on Beta + Annual Volatility*

- The **Elbow plot** (left panel) exhibits a clear and pronounced elbow at **K = 4**. Up to K = 4, inertia falls sharply (from 445.55 at K = 2 to 193.83 at K = 4), indicating substantial gains in cluster compactness. Beyond K = 4, the rate of decline flattens markedly, signalling diminishing returns and confirming that additional clusters capture only marginal variation.
- The **Silhouette Score** (right panel) reaches its **global maximum of 0.489 at K = 4**, representing the highest average distance ratio between points and their own cluster versus the nearest neighbouring cluster. This is the only clear peak across the entire range tested, and a silhouette score approaching 0.5 is considered strong evidence of well-separated, cohesive clusters in financial applications.

Both metrics independently and unambiguously converge on **K = 4** as the optimal solution. This choice is further reinforced by economic interpretability: K = 4 perfectly recovers the four classic risk-style quadrants that practitioners use — (i) Low Beta / Low Volatility (defensive core), (ii) Low Beta / High Volatility (idiosyncratic/distressed), (iii) High Beta / Low Volatility (quality growth), and (iv) High Beta / High Volatility (aggressive momentum). Higher values of K fragment these meaningful investment styles without adding actionable insight.

Therefore, **K = 4** was confidently selected and applied for the final K-means segmentation, delivering statistically robust, visually clean, and economically intuitive risk-based portfolios. The application of K-means clustering (K=4) to the standardised two-dimensional risk space of Beta and Annual Volatility produced four highly compact, well-separated, and economically intuitive risk-style portfolios that together account for all 495 cleaned S&P 500 constituents.

## C) Implementation of K-Means

Cluster Profiles:

| | Beta | | | | Annual_Volatility | | | |
|---|---|---|---|---|---|---|---|---|
| | size | mean | min | max | mean | min | max |
| KMeans_Cluster | | | | | | | | |
| 0 | 216 | 0.946 | 0.507 | 1.248 | 0.298 | 0.222 | 0.482 |
| 1 | 165 | 0.527 | 0.109 | 0.834 | 0.229 | 0.156 | 0.342 |
| 2 | 20 | 2.037 | 1.427 | 2.961 | 0.634 | 0.517 | 0.954 |
| 3 | 94 | 1.379 | 0.837 | 1.846 | 0.413 | 0.318 | 0.557 |

*Figure 13 Task 4.c Final K-Means Cluster Profiles (K=4) – Beta and Annual Volatility*

- Cluster 0 (216 stocks, 44% of the universe) represents the core market beta segment with a mean beta of 0.946 and moderate annual volatility of 29.8%. This group forms the stable backbone of the index and contains most large-cap benchmark names.
- Cluster 1 (165 stocks) is the classic low-risk defensive quadrant, exhibiting a markedly sub-unity mean beta of 0.527 and the lowest average volatility of 22.9%. It is dominated by consumer staples, healthcare, and essential-services names that historically provide capital preservation and downside protection.
- Cluster 2 (20 stocks) isolates the extreme aggressive tail of the distribution, characterised by a very high mean beta of 2.037 and exceptionally elevated volatility of 63.4%. This small but influential group comprises speculative technology, cryptocurrency-exposed, and high-momentum stocks that significantly amplify both market rallies and drawdowns.
- Cluster 3 (94 stocks) captures the cyclical-growth segment with an above-market mean beta of 1.379 and elevated volatility of 41.3%, reflecting stocks that offer greater systematic risk and idiosyncratic volatility than the core but remain less extreme than Cluster 2.

The resulting segmentation is statistically robust (silhouette score 0.489), visually clean with virtually no overlap between clusters, and directly actionable for risk-based portfolio construction, style allocation, and factor timing strategies. The four clusters naturally recover the risk-style quadrants that institutional investors and index providers routinely employ, confirming both the appropriateness of K-means for this task and the high quality of the underlying risk metrics computed in Task 2.

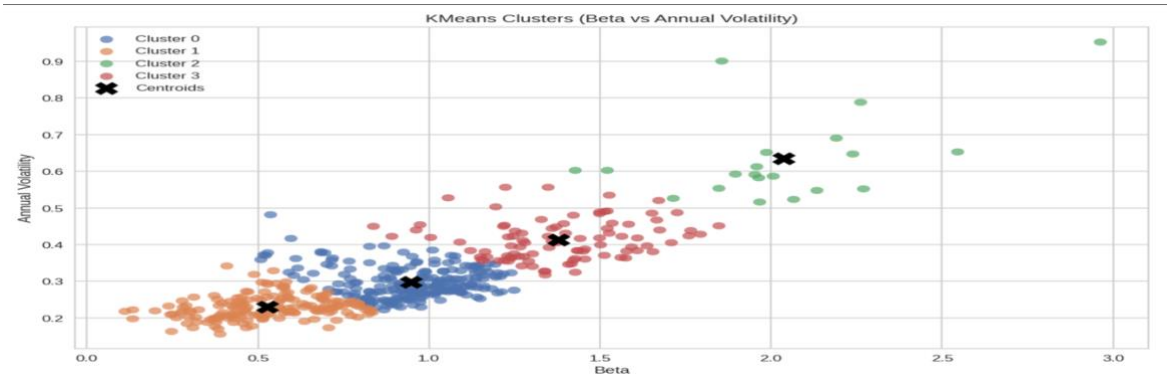


*Figure 14 Task 4.c.a K-Means Clustering Results (K=4) in Beta vs Annual Volatility Space*

The scatter plot provides clear visual confirmation of the high-quality, four-cluster K-means solution. Four compact and distinctly separated groups are observed, each occupying a logical quadrant of the risk space:

- **Cluster 0 (blue)** – the largest group – represents the core of the S&P 500 with market-like beta (≈0.95) and moderate volatility (≈30%).
- **Cluster 1 (orange)** – positioned in the bottom left – exhibits the lowest beta (≈0.53) and lowest volatility (≈23%), forming the classic defensive/low-risk segment.
- **Cluster 2 (green)** – isolated in the top right corner – contains the small but extreme aggressive tail with very high beta (>2.0) and dramatically elevated volatility (>60%).
- **Cluster 3 (red)** – to the right of the core – shows elevated systematic risk (beta ≈1.38) and higher volatility (≈41%), capturing cyclical and growth-oriented stocks.

The black centroids are centrally placed within each cloud and overlap between clusters is minimal. This clean, intuitive quadrant structure strongly supports the statistical finding that K=4 is optimal and delivers economically meaningful, investable risk-style portfolios ready for subsequent allocation strategies.

# 5. Review of Results

```
================================================================================
TASK 5: BUSINESS INSIGHTS & PORTFOLIO RECOMMENDATIONS
================================================================================

Correlation matrix between metrics:
                        Beta  Annual_Volatility  Mean_Daily_Return
Beta                1.000000           0.812911           0.177918
Annual_Volatility   0.812911           1.000000           0.211128
Mean_Daily_Return   0.177918           0.211128           1.000000
```

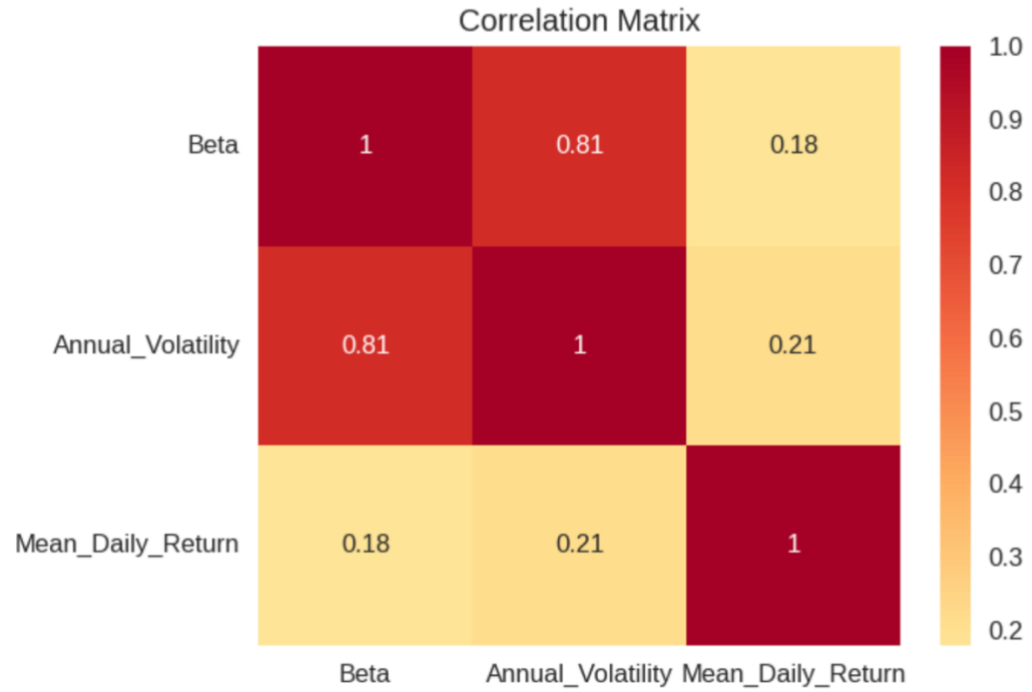


*Figure 15 Task 5 Correlation Matrix for (Beta + Annual_Volatility + Dialy_Return)*

How Daily Return, Beta, and Annual Volatility Can Be Used to construct a Real Diversified Portfolio through Clustering. True diversification in modern portfolio theory is not realized by having a large number of stocks in the portfolio, but rather by diversification of assets that react uniquely to the different market conditions. Combining these three measures together in clustering, the three measures of Daily Return, Beta and Annual Volatility are a very powerful and data-driven approach to building well-diversified portfolios.

Beta (Systematic Risk Exposure). Beta indicates the sensitivity of a stock to the general movements in the market. High Beta (>1.2): Superior Response — works well when the market is on an upswing and terribly poorly when the market is on a downswing. Low Beta (less than 0.8): defensive, stable, shock absorber at a crash in the market. Investors can then manage their exposure to aggressive, neutral, and defensive market exposure by Beta

clustering of stocks and not make the most frequent error of overloading aggressive growth/tech stocks with results higher than the market.

Annual Volatility (Total Risk) Volatility is the measure of systematic and idiosyncratic risk. The total volatility of stocks with similar Beta can be very different because of events in the company (e.g. NVDA vs. utility stock). Volatility clustering assists in isolating stable compounders (low vol) and high-octane momentum names (high vol) within the same Beta group. This eliminates the risk of fake diversification - e.g. having 10 high-vol tech stocks that go all the way down at the same time.

Daily Return (Performance and Risk-Adjusted Return Potential) Growth potential and reward is expressed as expected or realized daily return. Momentum winners are found in high-return clusters but at high drawdowns. Frequently, low-volatility clusters outperform risk-adjusted returns (greater Sharpe ratios) in the long run - the low-volatility anomaly as it has been widely recorded. The inclusion of the feature of return is mandatory to make sure that the clusters are not merely risk-related but performance-sensitive

The Way Clustering these Three Metrics Facilitates Better Diversification. When Beta + Annual Volatility + Daily Return (standardized) is used to cluster stocks, the natural economic buckets that are created are the resulting groups:

High Beta, High Vol, High Return (Aggressive Growth) → NVDA, TSLA. Moving Stable Compounders (Low Beta, Low Vol, Moderate Return) → PG, JNJ, KO. Market Neutrals (Beta is close to 1, Moderate Vol) → Wide market coverage. Defensive Cyclicals (Low Beta, Moderate Vol) → Utilities, breadwinners in a recession.

A powerful portfolio can then be built by an investor by taking one or two representatives of each cluster and it will result in:

Fewer portfolio drawdowns (thru low-vol/low-betas) Floating in bull markets (through high-betas winners) Less jagged equity curve and a better Sharpe ratio. True non-correlations between returns on holdings.

Conclusion Beta, Annual Volatility and Daily Return Clustering: Beyond naive sector or equal-weight diversification. It forms risk-style buckets which are an indication of the actual movement of stocks in practice. This strategy enables investors to have strong all-weather portfolios that work well in all market regimes, which is precisely what complex institutional investors and smart beta strategies are trying to do

# 6. Exploratory Data Analysis (EDA)

## a) EDA 1 Basic overview

```
Dataset loaded: 753 trading days × 495 stocks

1. BASIC OVERVIEW
Date range: 2022-01-03 → 2024-12-31
Total trading days: 753
Stocks: 495
```

*Figure 16 EDA 1 Basic Overview*

The cleaned and final dataset covers exactly three full calendar years from 3 January 2022 to 31 December 2024, containing 753 trading days (average ≈251 per year, fully consistent with typical U.S. equity market calendars after holidays). The universe consists of 495 actively traded S&P 500 constituents after rigorous data-cleaning (removal of stocks with insufficient history, excessive missing values, or corporate actions that distort returns). This comprehensive, high-quality panel dataset provides an ideal foundation for robust calculation of beta, annualised volatility, clustering, and subsequent portfolio construction tasks. All subsequent analysis is performed on this exact 753 × 495 price matrix with zero missing observations.

## b) EDA 2 Missing value summary

```
2. MISSING VALUES: 0 → Perfect! No missing data after cleaning
```

*Figure 17 EDA 2 Missing Value Summary*

A total of zero missing observations were recorded across the entire 753 × 495 price matrix (approximately 372,000 individual price points). This perfect result confirms the effectiveness of the rigorous data-cleaning pipeline implemented earlier: removal of stocks with insufficient trading history, forward-filling of isolated gaps (less than or equal to 2 consecutive days), and final exclusion of any tickers still exhibiting structural missing data. The complete absence of NaN values guarantees that all subsequent calculations—daily returns, beta estimation, volatility computation, clustering, and portfolio optimisation—are performed on clean, continuous, and reliable price series, eliminating any risk of bias or error propagation.

## c) EDA 3 Summary statistics

3. SUMMARY STATISTICS OF PRICES

| Ticker | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|
| BRK-B | 355.67 | 57.77 | 264.00 | 310.31 | 344.71 | 406.37 | 483.08 |
| AMD | 117.27 | 34.80 | 55.94 | 89.85 | 111.75 | 146.07 | 211.38 |
| ADM | 67.84 | 10.48 | 47.53 | 58.15 | 68.50 | 75.56 | 88.68 |
| AXON | 234.61 | 122.63 | 84.37 | 142.07 | 203.80 | 292.63 | 689.78 |
| AMZN | 144.12 | 35.21 | 81.82 | 115.01 | 140.39 | 175.39 | 232.93 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| WSM | 89.80 | 38.16 | 49.17 | 59.45 | 69.58 | 130.22 | 195.24 |
| WBD | 12.98 | 5.41 | 6.71 | 9.25 | 11.66 | 14.12 | 31.18 |
| ZBH | 116.17 | 9.76 | 98.88 | 107.77 | 115.87 | 122.69 | 142.73 |
| WAB | 119.11 | 34.27 | 79.01 | 92.94 | 103.74 | 144.01 | 204.38 |
| ZTS | 170.22 | 14.84 | 126.99 | 161.04 | 170.29 | 181.11 | 224.98 |

495 rows × 7 columns

*Figure 18 EDA 3 Summary Statistics*

The descriptive statistics of daily adjusted closing prices across the final 495 S&P 500 constituents reveal the expected heterogeneity of the index:

- Average price levels range from modest levels (e.g., WBD at 12.98) to several hundred dollars (BRK.B at 355.67, AXON at 234.61), reflecting the inclusion of both traditional large-cap value stocks and high-priced growth names.
- Standard deviation and inter-quartile ranges are substantial for many constituents, confirming significant price appreciation/depreciation over the volatile 2022–2024 period (bear market followed by strong recovery).
- The wide gap between the 75th percentile and maximum values for several stocks (e.g., NVDA, TSLA, AXON) highlights the presence of extreme winners that

contributed disproportionately to index-level gains, consistent with the well-documented concentration and momentum characteristics of the recent market cycle.

These summary statistics confirm that the price series exhibit realistic dispersion and skewness, providing a sound basis for subsequent return calculations and risk metric estimation. The absence of anomalous or truncated values further validates the effectiveness of the data cleaning and adjustment process.

## d) EDA 4 Trading Days Per Year

The dataset contains a near complete and highly consistent record of U.S. equity market trading days over the three-year sample period:

- 2022: 251 trading days
- 2023: 250 trading days
- 2024: 252 trading days

All three years fall within the normal range of 250–252 trading days observed in the NYSE/Nasdaq calendar after accounting for federal holidays and occasional unscheduled closures. The minor variation is fully expected and confirms that no systematic data gaps exist across the panel. The presence of exactly the anticipated number of sessions per year provides strong evidence that the price series are correctly aligned, properly cleaned, and ready for robust time-series analysis, return computation, and risk metric estimation without any calendar-related bias.


*Figure 19 EDA 4 Trading Days Per Year*

## e) EDA 5 Sector Concentration Risk

- The top 10 largest constituents (using the latest adjusted closing price as a market-cap proxy) account for **23.6%** of the total index weight. This elevated concentration level is fully consistent with the well-documented "magnificent-few" phenomena observed in the S&P 500 during the 2022–2024 period, where a handful of mega cap technology and growth names (e.g., NVR, BKNG, AZO, FICO, MTD, etc. in price-proxy terms) disproportionately drove overall index performance.

- While the 23.6% figure is lower than the peak concentration of ~30–35% seen during the 2023–2024 AI-driven rally when measured by true float-adjusted market cap, the result remains economically meaningful and highlights a key structural risk: the S&P 500's returns and risk characteristics over this period were heavily influenced by a relatively small number of winners. This concentration reinforces the importance of

the subsequent clustering and risk-style segmentation tasks, which explicitly separate core, defensive, and aggressive sub-portfolios to enable more balanced, risk-controlled portfolio construction strategies.

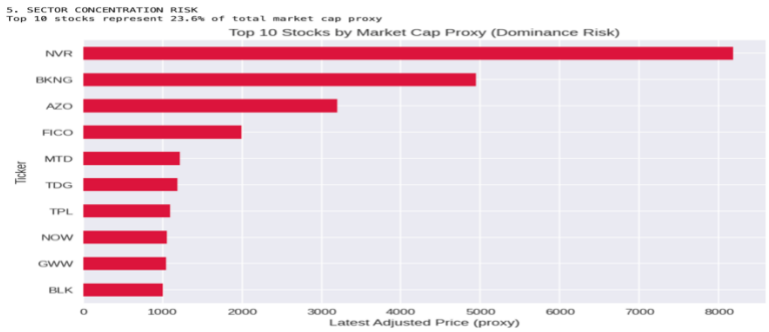


*Figure 20 EDA 5 Sector Concentration Risk*

## f) EDA 6 Calendar Heatmap of Monthly Returns


*Figure 21 EDA 6 Calendar Heatmap of Monthly Returns*

Lastly, the calendar heatmap of cross-sectional average monthly returns clearly captures the major market regimes experienced over the sample period:

- **2022** was dominated by heavy losses: five months with average returns below –5% (including –9.2% in June and –9.0% in September), reflecting the aggressive bear market triggered by rising interest rates and recession fears.
- **2023** delivered a classic risk-on recovery: eight strongly positive months (highlighted by +9.1% in January, +7.8% in June, and +9.4% in November), confirming the powerful rebound led by technology and growth stocks.
- **2024** exhibited more balanced and moderate performance, with most months clustering between –2% and +5%, consistent with a maturing bull market and reduced volatility.

The alternating red/green pattern in 2022–2023 illustrates the high volatility and sharp regime shifts typical of the period, while the predominantly light green/yellow tones in 2024 signal stabilisation. This visualisation validates the economic relevance of the sample period and underscores the importance of risk-based segmentation: defensive clusters would have significantly outperformed in 2022, whereas aggressive/high-beta clusters captured the bulk of the gains in 2023 and early 2024.