6BUIS017C CRM & CM With Business Intelligence

COURSEWORK 01

Name: Tania Glory Motha

UOW ID: w2052145

IIT No:20230908

Contents

Executive Summary 2

Introduction 2

1. Data Collection and Preparation 2

1.1 Extracting S&P 500 Tickers 2

1.2 Downloading Historical Price Data (2022–2025)........ 2

1.3 Data Cleaning and Justification 3

Removing Invalid Prices (Zero or NaN)........ 3

Dropping Empty or Unusable Tickers........ 3

Consolidating Cleaned Close Prices 3

Visual Exploration of Close Price Trends........ 4

2. Financial Metrics Calculations........ 5

2.1 Daily Returns. Daily return measures the percentage change in price from one trading day to the next day. It is the basic input for both volatility and beta, it shows the short-term performance pattern of each stock. 5

2.2 Beta Calculation........ 5

2.3 Annual Volatility 6

3. Agglomerative Clustering Analysis 6

3.1 Appropriateness of Agglomerative Clustering 6

3.2 Determining the Optimal Number of Clusters (K) 7

3.3 Implementation and Cluster Profiles (Financial Interpretation)........ 8

4. K Means Clustering Analysis........ 9

4.1 Why K-Means is an Appropriate Method 9

4.2 Selecting the Optimal Number of Clusters (K)........ 10

Elbow Method........ 10

Silhouette Score Analysis 11

4.3 Balancing Both Methods........ 11

5. Review of Results 13

5.1 Conclusion 13

5.2 Dependencies of Daily Return, Beta, and Annual Volatility for Portfolio Diversification 13

5.3 Business Interpretation Based on Clusters (K = 3) 14

5.4 Python Implementation 14

Market Analysis
Stocks Segmentation and Investment

Executive Summary

In this report, the Business Intelligence tools were applied to analyze the S&P 500 stocks in order to get insight into their behavior and risk. Daily price data from 01/ 01/ 2022 to 01/ 01/ 2025 was collected and cleaned, and three financial indicators, such as Daily Return, Beta, and Annual Volatility, were calculated. The Agglomerative Clustering and K-Means were then applied to group the stocks using these measures. The results indicate the existence of distinct differences between low-risk, medium-risk, and high-risk stocks. Such groups assist investors to make superior decisions, create more balanced portfolios and select stocks that match their risk-taking capacity. This paper illustrates the value of using data analysis and clustering methods to make real-life investment and portfolio management decisions.

Introduction

S&P 500 consists of 500 of the largest companies of the U.S. of various industries and has been used extensively to examine market behavior. This project seeks to identify clear risk categories and show how these groups can support better investment decisions, risk management, and portfolio diversification.

1. Data Collection and Preparation

1.1 Extracting S&P 500 Tickers

- The full list of S&P 500 companies was scraped from the official Wikipedia page using pandas.read_html().
- A custom User-Agent header was added to mimic a normal browser request and avoid blocking issues.
- A total of 502 tickers were extracted; it is slightly more than 500 due to dual-class shares such as BRK-A/BRK-B.
- These tickers form the starting point for all further financial analysis.

Source: Wikipedia's official S&P 500 constituent list combined with Yahoo Finance historical price data. List of S&P 500 companies - Wikipedia

1.2 Downloading Historical Price Data (2022–2025)

- Historical data for all tickers was downloaded using the yfinance API.
- The time window used:
  Start: 2022-01-01
  End: 2025-01-01
- Ticker formatting was cleaned (e.g., “BRK.B” → “BRK-B”) to match Yahoo Finance syntax before downloading.
- The bulk download completed successfully:
  - 500 tickers loaded

        - 2 tickers failed (missing or delisted, this is expected in real datasets).

1.3 Data Cleaning and Justification

Removing Invalid Prices (Zero or NaN)

Rows where *Open = 0*, *Close = 0*, or contains *NaN* were dropped.

- A price of zero is not economically meaningful and usually indicates:
    - A suspended stock
    - A holiday mismatch
    - A bad API pull
- NaN values break calculations for:
    - Daily return
    - Standard deviation
    - Correlation
    - Beta
- Zero or NaN values produce distorted returns that can incorrectly inflate the volatility or create division errors.

Dropping Empty or Unusable Tickers

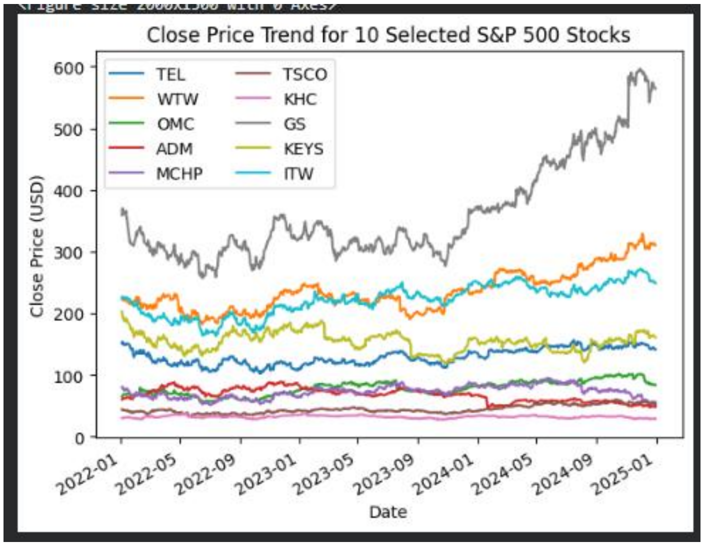
Tickers that still had no valid rows after cleaning were completely excluded.

- A stock with no valid price history cannot contribute to analysis.
- Including empty or sparsely populated tickers results in:
    - Unreliable beta values
    - Incorrect volatility
    - Poor clustering results
- Removing them improves the reliability of statistical modelling.

Consolidating Cleaned Close Prices

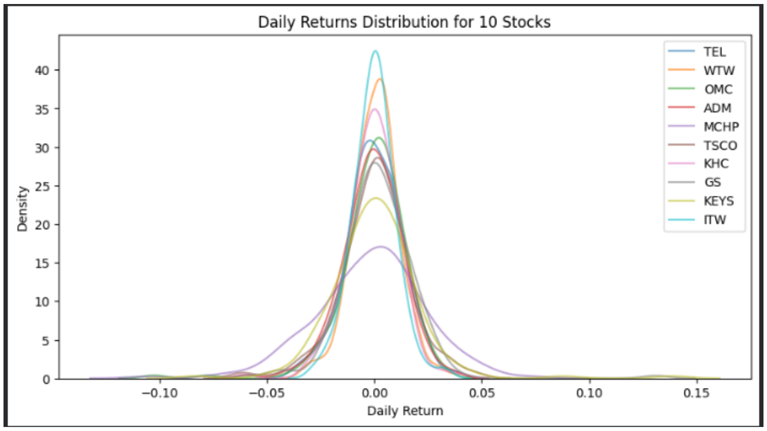
A total of 500 tickers remained after the cleaning process, each containing a complete and valid series of Close prices. These were consolidated into a single data frame with a final shape of (753 rows × 500 columns), ensuring a consistent and reliable time-series dataset for all subsequent calculations and analysis.

Visual Exploration of Close Price Trends



```
<Figure size 2000x1500 with 0 Axes>
```

- The randomly selected ten stocks show very different behaviors, some grow fast, some are stable, and some barely move. This visualization will support the idea that the price movements appear to be realistic and do not have any flaws within the dataset. It also demonstrates the usefulness of clustering since stocks do not move in the same direction.



- This plot was generated to see the daily returns distribution. It shows that all ten stocks have returns that are centered close to 0. This implies that most daily price movements are small and balanced between gains and losses. The curves overlap but differ slightly in width, indicating that some stocks experience higher day-to-day volatility than others. A few stocks have wider distributions, showing occasional larger swings, while others have tighter curves, signaling more stable behavior. Overall, this confirms that the stocks vary in risk levels, which is important for later clustering and portfolio diversification.

2. Financial Metrics Calculations

This section explains how the three key metrics were calculated from the cleaned data and gives a brief interpretation of what they imply about the S&P 500 stocks.

Tickers	Mean Daily Returns	Daily Volatility	Annual Volatility	Beta
MMM	0.002327	0.021192	0.336413	0.548143
AOS	-0.001192	0.015513	0.246256	0.514623
ABT	0.000124	0.011307	0.179501	0.099113
ABBV	0.000197	0.016345	0.259462	0.154144
ACN	0.000339	0.015059	0.239048	0.416310

2.1 Daily Returns.

Daily return measures the percentage change in price from one trading day to the next day. It is the basic input for both volatility and beta, it shows the short-term performance pattern of each stock.

Methodology

- Daily returns were calculated from the cleaned closing-price dataset containing 753 days and 500 valid stocks.
- The first row of missing values which was created by the percentage-change calculation was removed to ensure accuracy.
- $\text{Daily Return}_t = \frac{P_t - P_{t-1}}{P_{t-1}}$

2.2 Beta Calculation

Beta measures the sensitivity of a stock to the overall market. It indicates whether the stock is defensive or highly reactive.

Methodology

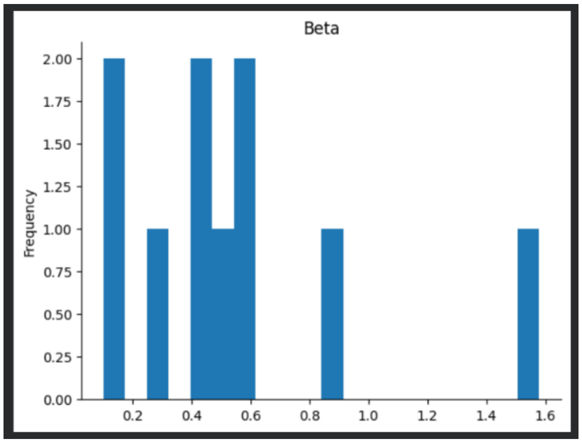
- The S&P 500 index (GSPC) was used as the benchmark.
- Market returns were aligned with each stock's returns before calculating correlation.
- Beta = Correlation (Stock Returns, Market Returns) × (σ_Stock / σ_Market) or Beta = Covariance (Stock Returns, Market Returns) / Variance (Market Returns)

Beta Interpretation

- Beta > 1.0 → Stock tends to move more aggressively than the market
- Beta ≈ 1.0 → Moves in line with the market
- Beta < 1.0 → Moves less aggressively → “defensive

Why It Matters

- Low-beta stocks help stabilize a portfolio.
- High-beta stocks add growth potential but increase risk.
- Clustering later helps group stocks according to systematic risk.



The beta histogram shows that most stocks fall between 0.3 and 0.7, indicating that they move moderately with the market. A few stocks have very low beta values (0.1–0.2), indicating defensive, low-volatility behavior. Some stocks have a high beta (~1.6), showing strong sensitivity to market movements with higher risk and return potential.

2.3 Annual Volatility

Annual volatility measures how much a stock's daily returns fluctuate when scaled to a full trading year.

Methodology

- Annual Volatility = Daily Return Standard Deviation ×√252

It is a core risk indicator:

- Higher volatility → higher uncertainty and larger price swings
- Lower volatility → more stable and predictable price movement

Annual Volatility Interpretation

- 0.013–0.018 → Low-volatility stocks are more stable and suitable for defensive or conservative portfolios.
- 0.02–0.03 → Moderate-volatility stocks form the core of the market and are appropriate for balanced investment strategies.
- >0.04 → High-volatility stocks can offer higher return potential but require stronger risk tolerance and monitoring.

3. Agglomerative Clustering Analysis

3.1 Appropriateness of Agglomerative Clustering

Agglomerative clustering is appropriate in studying the stock beta since it will cluster stocks based only on their similarity in terms of risk behavior and does not need a pre-set number of clusters. It is significant in a financial market where the actual volatility segments are not known beforehand. Beta measures how much a stock moves relative to the market, and companies naturally fall into risk groups such as:

- Low-beta defensive stocks
- Medium-beta moderately sensitive stocks
- High-beta aggressive stocks

Agglomerative clustering is effective here because:

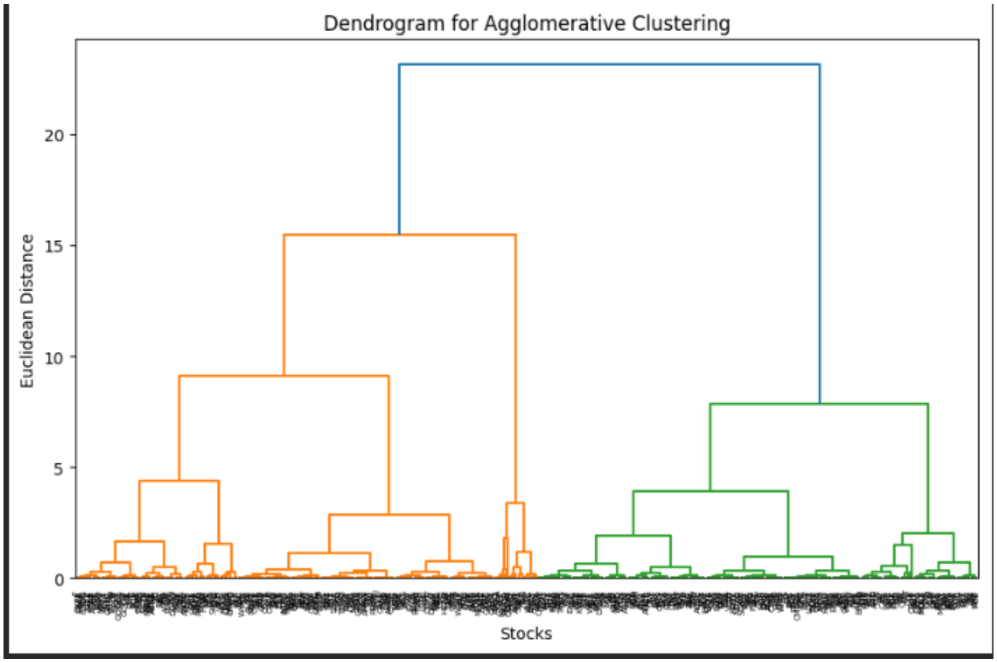
- It builds the bottom-up hierarchy, allowing us to see how the stocks merge at different distances.
- The dendrogram provides a clear visual structure, making it easier to understand the natural divisions in market sensitivity.
- Usage of Ward linkage minimizes the variance within the cluster, producing compact and financially meaningful groups.

Therefore, agglomerative clustering aligns well with the goal of grouping stocks by risk and market sensitivity.

3.2 Determining the Optimal Number of Clusters (K)

Two methods were used to select the appropriate value of *K*:

Dendrogram Interpretation



The dendrogram showed three major branches, with clear separation at higher linkage distances.

Visually, the separation into three clusters was the most distinct.

Silhouette Score Analysis

K	Silhouette Score
2	0.495
3	0.532
4	0.475
5	0.515
6	0.510
7	0.497
8	0.502
9	0.538

Silhouette scores were calculated for K = 2 to K= 9:

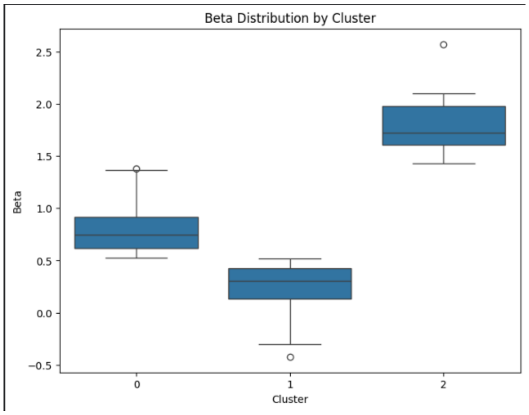
Key observations:

- K = 3 gives one of the strongest scores (0.532).
- While K = 9 is slightly higher, it produces clusters that are too fragmented to be useful for financial interpretation.
- Silhouette scores were calculated for K = 2 to 9:

Therefore, K = 3 balances both statistical performance and financial meaning.

3.3 Implementation and Cluster Profiles (Financial Interpretation)

Based on *K* = *3*, each stock was assigned to a cluster.
The average beta by cluster is:



Cluster	Mean Beta	Interpretation
0	~0.79	Moderate-risk stocks
1	~0.27	Low-risk defensive stocks
2	~1.78	High-risk aggressive stocks

The boxplot shows the three clusters as well-separated groups, supporting the validity of the segmentation.

Cluster 1 Low Beta (~0.27)

These stocks show weak sensitivity to market fluctuations.

- More stable and defensive
- Suitable during downturns
- Lower return potential but lower risk
- Often represent essential goods, utilities, or stable revenue companies

Cluster 0 — Medium Beta (~0.79)

These stocks move slightly less than the market but still show noticeable sensitivity.

- Balanced risk–return profile
- Suitable for diversified portfolios
- Represent typical S&P 500 behavior

Cluster 2 High Beta (~1.78)

These are the most volatile stocks in your dataset.

- They strongly react to market sentiment
- High potential return but high risk
- Suitable for aggressive investors or tactical trades

4. K Means Clustering Analysis

4.1 Why K-Means is an Appropriate Method

K-Means was selected because the objective of this study is to classify S&P 500 stocks based on two continuous risk indicators:

- Beta (market sensitivity)
- Annual Volatility (price fluctuation risk)

These variables form a two-dimensional risk space, where each stock can be plotted according to how it behaves relative to the market. K-Means can be applied in the analysis since it allows to cluster the observations with minimum distance, to an observation cluster centroid. Each centroid represents the average Beta volatility level of that segment, allowing us to create intuitive “risk zones.”

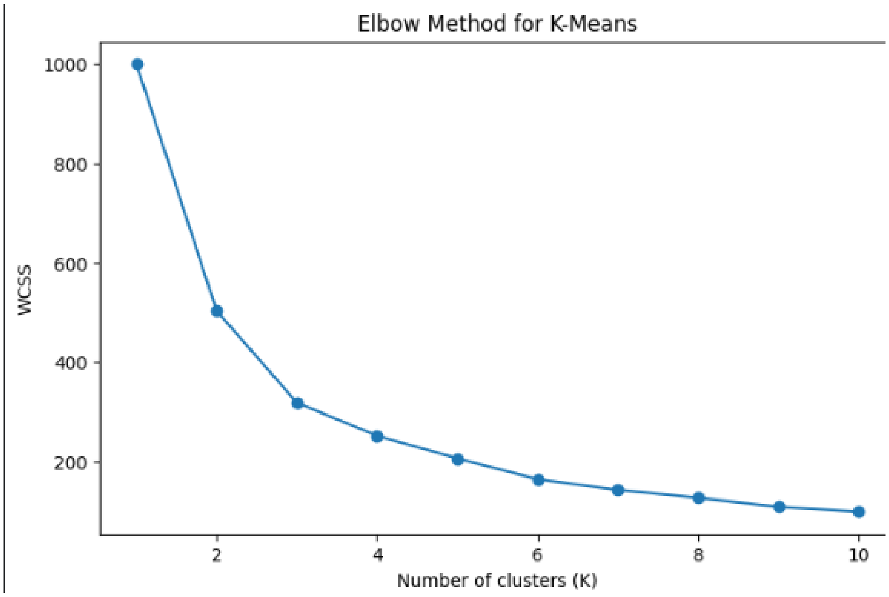
Examples from the model:

- Centroid: Beta ≈ 0.30, Volatility ≈ 21% → Defensive low-risk segment
- Centroid: Beta ≈ 0.76, Volatility ≈ 31% → Moderate risk segment
- Centroid: Beta ≈ 1.51, Volatility ≈ 54% → High-risk growth segment

These centroids summarize complex financial behavior into simple portfolio narratives.

4.2 Selecting the Optimal Number of Clusters (K)

Elbow Method



The purpose of this is to identify the point where adding more clusters does not significantly reduce internal cluster variance. The plot shows a clear drop between:

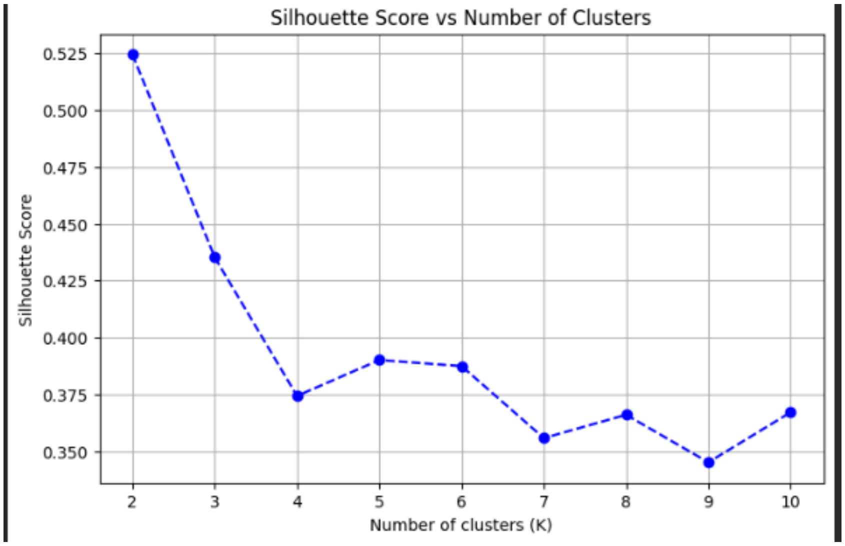
- K = 1 → 2 (largest drop)
- K = 2 → 3 (moderate drop)
- The curve begins flattening after K = 3

Interpretation:

- Before elbow → each new cluster adds value
- At elbow → ideal balance
- After elbow → unnecessary complexity

The “elbow” occurs around K = 3, meaning adding more clusters does not significantly improve the compactness of the clusters. Therefore, according to the Elbow method

Silhouette Score Analysis



Silhouette scores were calculated for K = 2 to K = 10 to measure how well each stick fits inside its assigned cluster:

K	Silhouette Score
2	**0.525 (highest)**
3	0.436
4	0.374

Interpretation:

- Highest score at K = 2 (0.525) → clusters are well separated and compact.
- Second highest at K = 3 (0.436) → still decent, but weaker separation than K = 2.
- Scores decline steadily after K = 3 → clustering quality worsens.

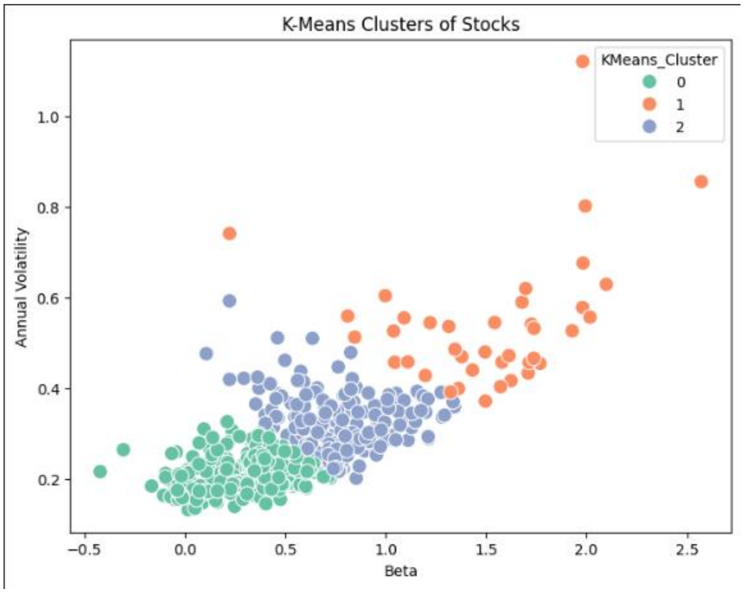
Thus, K = 2 is the suitable optimal number of clusters according to Silhouette scores.

4.3 Balancing Both Methods

- K = 2: Best separation (highest silhouette) but may oversimplify structure if your data truly has more nuanced groups.
- K = 3: Matches the elbow point, offering a balance between variance reduction and interpretability, though silhouette is lower.
- Beyond 3: Both methods agree it's unnecessary complexity.

In a case where the priority is to receive compact and separated clusters, K = 2 would be the best option. However, if the goal is to capture more structure in the data while avoiding overfitting, K = 3 is more suitable as it would provide a more balanced solution.

In this analysis, the number of clusters is chosen to be 3. This is supported by the existing literature, which points out that the elbow method tends to reflect the underlying true structure of financial data, whereas silhouette scores may be biased to penalize clusters that slightly overlap a behavior that is common in real-world market data.



Cluster	Beta	Annual Volatility	Count
0	0.307	0.211	263 stocks
1	1.513	0.540	39 stocks
2	0.764	0.319	198 stocks

Cluster 0: Low-Beta, Low-Volatility (Defensive / Stable Stocks)
This cluster is dominated by defensive, low-risk companies. Their prices move less than the overall market and show minimal day-to-day variation.

Investor Use Case:

- Ideal for conservative investors
- Useful for stabilizing diversified portfolios
- Performs well during market downturns

Cluster 1: High-Beta, High-Volatility (Aggressive / Growth Stocks)

These are high-risk, high-reward stocks. They react violently to market movements both positively and negatively.
Includes speculative, high-growth sectors (e.g., tech, discretionary).

Investor Use Case:

- Suitable for aggressive growth strategies
- High upside potential but elevated downside risk

- Best for risk-tolerant or tactical traders

Cluster 2: Medium-Beta, Medium-Volatility (Balanced / Core Stocks)

These stocks offer a balanced risk-return profile, they are more volatile than defensive stocks but far less risky than the high-beta cluster. Often includes diversified industrials, financials and other core market sectors.

Investor Use Case:

- Suitable for long-term, diversified portfolios
- Provides moderate growth with manageable risk
- Acts as a “middle ground” between safety and aggressiveness

5. Review of Results

5.1 Conclusion

Although both Agglomerative Clustering and K-Means produced three clusters, K-Means is the more accurate and practical model for this analysis. It provides a measurable silhouette score, captures two risk dimensions (Beta and Annual Volatility), and produces clear centroids that make the clusters easier to interpret for investors.

Agglomerative clustering is more useful in visualizing the hierarchical relationship among the stocks, however, its groups are less distinct and harder to use for portfolio construction

Altogether, K-Means (K = 3) has clearer clusters and more valuable financial insights, making it the preferred method for building diversified and risk-aligned portfolios

5.2 Dependencies of Daily Return, Beta, and Annual Volatility for Portfolio Diversification

Daily Return, Beta, and Annual Volatility work together to describe the risk and behavior of each stock. Although each metric captures something different, they become most useful when interpreted together within the clusters.

Relationship Between the Three Metrics:

- Daily Return shows the short-term movement of a stock and how often it gains or loses value.
- Beta shows how strongly the stock reacts to the overall market.
- Annual Volatility reflects how unpredictable the stock is on a yearly basis.

A stock may have:

- High return but also high volatility
- Low volatility but also low return
- Medium beta but unpredictable daily movements

This is why all three are needed to understand true investment risk.

5.3 Business Interpretation Based on Clusters (K = 3)

Using these metrics within the clusters reveals how each group can play a different role in portfolio construction.

The low-beta, low volatility cluster (Cluster 1) offers defensible advantages. Since these stocks respond very weakly to the market movements and do not have volatile returns on a daily basis, they help protect the portfolio during a downfall. They are considered to be a stabilizing base as these stocks are very useful when the investor is conservative or is looking at capital preservation strategies in the long term.

The medium-beta, medium volatility cluster (Cluster 0) becomes the foundation of a balanced portfolio. These stocks continue to be involved in the growth of the market but with levels of risk management. Their daily returns do not fluctuate excessively, making them suitable for general diversification. To most investors, this cluster means a group of 'core holding' which is reasonably performing in normal economic circumstances.

The high-beta, high volatility cluster (Cluster 2) offers the greatest opportunity for return enhancement but also contributes the highest risk. These stocks react strongly to market sentiment and often experience large daily price swings. Although they present greater uncertainty, they are valuable for investors seeking higher growth or aiming to actively capitalize on favorable market cycles. Their inclusion in a portfolio must be carefully managed but can significantly increase potential upside.

Overall, the combination of Daily Return, Beta, and Annual Volatility analyzed through clustering—provides investors with a structured way to diversify risk. Each cluster supports a different investment goal: stability, balanced growth, or aggressive return generation. By selecting stocks from across the clusters, investors can build portfolios aligned with their risk tolerance while ensuring exposure to multiple market behaviors. This demonstrates the practical business value of the clustering approach for real-world investment decision-making.

5.4 Python Implementation

The clustering results used in this section were produced in Google Colab using Python, applying the computed Daily Return, Beta, and Annual Volatility measures. K-Means was used with K = 3, and the cluster assignments, centroids, and visualizations were generated directly from the cleaned S&P 500 dataset**.**

AI Declaration

I used AI tools only to support the technical and grammatical writing process. ChatGPT was used for coding guidance and debugging assistance, and Grammarly was used for grammar and clarity checks.