CRM & CM With Business Intelligence

6BUIS017W

Coursework 1

Salvador Maya
w1986820

Github: https://github.com/salvadormayaponce-wq/CW1

PART 1: Data Loading and Preparation

Introduction

The coursework requires analysing the companies included in the S&P 500 index. This dataset was specified in the assignment brief and provides a large and diverse sample of firms from different industries. Because the entire analysis depends on the quality and completeness of this data, the first step was to carefully collect, verify, and prepare the dataset before calculating any financial metrics or applying clustering techniques.

1A. Extracting S&P 500 tickers using Pandas

To begin, I extracted the list of S&P 500 constituents directly from the official "List of S&P 500 companies" table on Wikipedia. I used the Pandas read_html() function, which automatically identifies and loads HTML tables into a DataFrame. I selected the table containing the "Symbol" column, which provided all official tickers currently in the index.

This step is important from a data-handling perspective because the S&P 500 list changes over time. Using a live online table ensures that the analysis is based on the most current companies rather than relying on a manually entered or outdated list.

Before applying the full process, I first tested the workflow using only the first 50 tickers. This allowed me to confirm that the extraction and processing pipeline worked correctly and avoided unnecessary computation on the full dataset until the process was validated. After confirming that the test version ran without errors, I continued with the complete set of companies.

	Symbol	Security	GICS Sector	GICS Sub-Industry	Headquarters Location	Date added	CIK	Founded
0	MMM	3M	Industrials	Industrial Conglomerates	Saint Paul, Minnesota	1957-03-04	66740	1902
1	AOS	A. O. Smith	Industrials	Building Products	Milwaukee, Wisconsin	2017-07-26	91142	1916
2	ABT	Abbott Laboratories	Health Care	Health Care Equipment	North Chicago, Illinois	1957-03-04	1800	1888
3	ABBV	AbbVie	Health Care	Biotechnology	North Chicago, Illinois	2012-12-31	1551152	2013 (1888)
4	ACN	Accenture	Information Technology	IT Consulting & Other Services	Dublin, Ireland	2011-07-06	1467373	1989
...	...	...	...	...	...	...	...	...
498	XYL	Xylem Inc.	Industrials	Industrial Machinery & Supplies & Components	White Plains, New York	2011-11-01	1524472	2011
499	YUM	Yum! Brands	Consumer Discretionary	Restaurants	Louisville, Kentucky	1997-10-06	1041061	1997
500	ZBRA	Zebra Technologies	Information Technology	Electronic Equipment & Instruments	Lincolnshire, Illinois	2019-12-23	877212	1969
501	ZBH	Zimmer Biomet	Health Care	Health Care Equipment	Warsaw, Indiana	2001-08-07	1136869	1927
502	ZTS	Zoetis	Health Care	Pharmaceuticals	Parsippany, New Jersey	2013-06-21	1555280	1952

1B. Loading daily stock performance using yfinance and YahooFinancials

The specification requires using both the yfinance and YahooFinancials packages. To meet this requirement, I imported both tools, although the main data extraction was performed using yfinance because it returns price data in a structured and reliable format.

I downloaded daily price information (Open, High, Low, Close, Volume) for each S&P 500 company from 1 January 2022 to 1 January 2025. This period includes different market conditions, which is important because volatility and beta calculations are tied to how a stock behaves during both stable and unstable periods.

From a data-science point of view, collecting daily data over multiple years ensures that:

- financial metrics like volatility and beta are based on enough observations to be reliable
- short-term noise does not dominate the results
- clustering reflects stable patterns rather than random fluctuations

The downloaded prices were stored in a multi-column DataFrame indexed by date and ticker

Price	Adj Close				Close						...	Volume					
Ticker	BF.B	BRK.B	Q	SOLS	A	AAPL	ABBV	ABNB	ABT	ACGL	...	WY	WYNN	XEL	XOM	XYL	XYZ
Date																	
2022-01-03	NaN	NaN	NaN	NaN	152.320084	178.270309	116.779305	172.679993	128.996109	42.362530	...	3831100	2437800	3501100	24282400	759100	7315700
2022-01-04	NaN	NaN	NaN	NaN	147.170715	176.007751	116.555077	170.800003	125.962341	42.914051	...	3089700	2292300	4197000	38584000	925400	14768500
2022-01-05	NaN	NaN	NaN	NaN	144.649551	171.325989	117.167343	162.250000	125.396393	42.410072	...	3737600	3439900	4166000	34033300	1090200	17546200
2022-01-06	NaN	NaN	NaN	NaN	145.155701	168.466003	116.615440	159.750000	125.377853	42.657307	...	3315200	2583200	2296000	30668500	703400	16244200
2022-01-07	NaN	NaN	NaN	NaN	141.291260	168.632477	116.313622	166.050003	125.767517	42.856995	...	3309900	1720400	2673100	23985400	765000	9426000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2024-12-24	NaN	NaN	NaN	NaN	135.004929	257.037476	173.918823	134.990005	112.625931	92.669998	...	1780100	692800	943900	7807000	379300	2197700
2024-12-26	NaN	NaN	NaN	NaN	134.737091	257.853760	173.145828	135.320007	113.126434	92.930000	...	1736500	1218900	1394900	9652400	575700	2991100

1C. Identifying and removing invalid data

Like many datasets, the downloaded prices contained some missing or inconsistent values. These issues usually come from:

- companies joining or leaving the index
- incomplete trading histories
- temporary data gaps from Yahoo Finance
- stocks with low liquidity

It is essential to remove invalid observations because metrics such as daily returns, beta, and volatility are highly sensitive to missing or infinite values.

I applied a structured cleaning process that included:

- removing rows with missing prices (dropna)
- replacing infinite values with NaN
- excluding stocks with insufficient data for the full 2022–2025 period
- verifying the dataset again after merging returns, beta, and volatility

These steps ensure that each stock included in the analysis has a complete and consistent price history, which is necessary for reliable calculations and clustering.

From a BI/analytics perspective, this process reflects good data-preparation practice: cleaning ensures that later insights are based on trustworthy and reproducible data rather than accidental gaps or errors.

PART 2: Financial Metrics Calculation

This section explains how the financial metrics used later in the clustering models were calculated. The three main measures; daily returns, beta, and annualised volatility, are all standard tools in financial analysis and provide a foundation for understanding stock behaviour and market sensitivity. These metrics were first tested on a smaller sample of 50 stocks to validate the process before applying them to the full dataset.

2A. Daily Returns

Daily returns measure how much a stock's closing price changes from one trading day to the next. I calculated daily returns using the standard financial formula for returns.

This method converts prices into comparable percentage movements, which allows stocks with very different price levels to be analysed in a consistent way.

Daily returns are a necessary first step because both beta and annualised volatility depend on the variation in day-to-day returns. From a scientific standpoint, using returns instead of raw prices avoids misleading results caused by price scale differences and allows later risk computations to be mathematically meaningful.

2B. Beta Calculation

Beta measures how sensitive a stock is to movements in the overall market index, in this case, the S&P 500. A high beta means the stock tends to amplify market movements, while a low beta indicates more stable behaviour.

I calculated beta using the formula specified in the assignment:

Beta = Correlation(stock returns, market returns) × (Standard deviation of stock returns / Standard deviation of market returns)

This formula links two scientific ideas:

- **Correlation**: measures how closely the stock moves with the broader market
- **Volatility ratio**: scales the relationship based on how volatile the stock is compared to the index

To carry this out, I downloaded daily returns for both the S&P 500 index and each stock, ensured the dataset contained no missing or invalid values, and applied the formula to each ticker.

This approach aligns with standard risk analysis methods and demonstrates understanding of how systematic risk is measured in a quantifiable way in finance.

Preview of beta values table:

	Symbol	Beta
0	MMM	0.783737
1	AOS	0.924451
2	ABT	0.672316
3	ABBV	0.299098
4	ACN	1.094207

2C. Annualised Volatility

Annualised volatility indicates how much a stock's returns vary over time. I calculated volatility using the standard deviation of daily returns and converted it to an annual measure using √252, which assumes 252 trading days per year:

Annual Volatility = Standard deviation of daily returns × √252

This scaling is commonly used in finance because volatility is time dependent. Without annualization, the values would not be comparable across different time horizons.

The calculation was first applied to the test sample and then repeated for the complete cleaned dataset. Any stocks with incomplete data were excluded to maintain accuracy. This measure captures total risk, meaning it reflects both market-driven movements and company-specific events.

Preview of annual volatility table:

Ticker	
A	0.296259
AAPL	0.270935
ABBV	0.219958
ABNB	0.467875
ABT	0.217839
...	...
XYZ	0.653825
YUM	0.191335
ZBH	0.245157
ZBRA	0.400050
ZTS	0.275016

PART 3: Agglomerative Clustering (Beta)

Agglomerative clustering is a type of hierarchical clustering that groups data by gradually merging observations based on how similar they are. I applied this method using the previously calculated Beta metric, which represents how sensitive each stock is to movements in the overall market. This method helps us see whether companies naturally fall into groups with similar levels of market sensitivity.

The goal of this section is to check if the S&P 500 companies can be separated into different risk categories based on their Beta values alone. To do this, I generated a dendrogram, identified where the biggest separation occurs, and then used that point to decide on the number of clusters.

3A. Why Agglomerative Clustering is Appropriate

Agglomerative clustering is useful in this context because it does not require choosing the number of clusters beforehand. Instead, it builds the clustering structure step by step, starting from individual stocks and merging those that are most similar. This allows me to visually inspect the resulting hierarchy and decide on the number of clusters only after seeing how stocks group together.

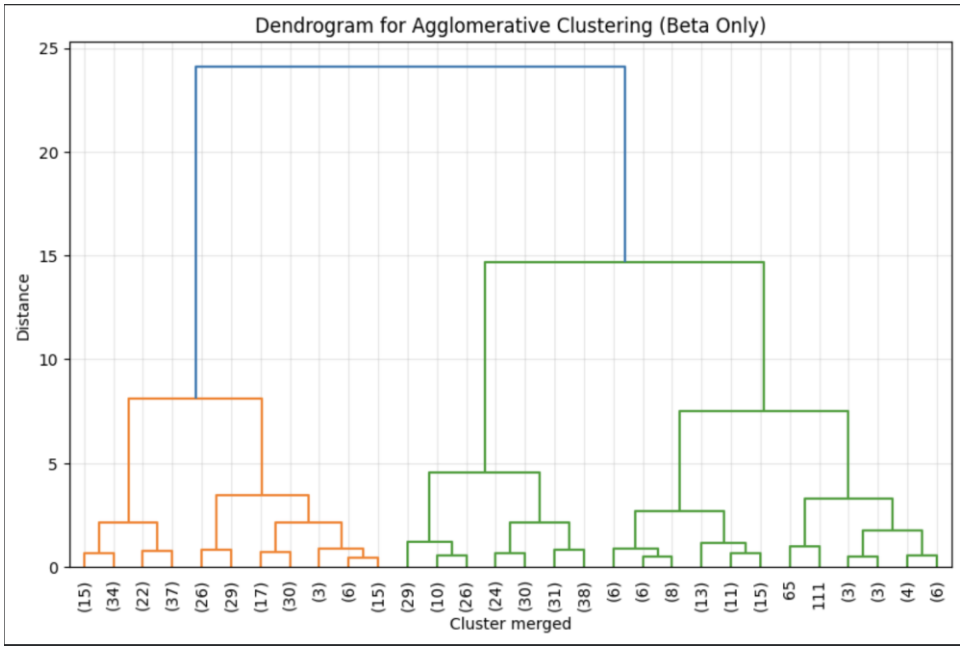
Since we are working with a single metric (Beta), this method offers a clear and intuitive way to observe how firms differ in terms of their exposure to market movements. It is also easy to interpret because the results can be displayed in a dendrogram, which highlights where the most meaningful separations occur.

3B. Selecting the Number of Clusters

To determine how many groups exist in the data, I generated a dendrogram using Ward's linkage method**,** which merges clusters in a way that minimises within-group variance. Before building the dendrogram, I standardised the Beta values to make sure that the clustering was based on consistent scale and not influenced by differences in magnitude across stocks.

The dendrogram clearly showed a large jump in linkage distance near the top of the hierarchy. A common rule when interpreting dendrograms is to cut the tree just below the biggest vertical jump, because that is where the most significant separation between clusters occurs. In this case, the largest jump suggested a natural split into two clusters.

Dendrogram produced in Collab:

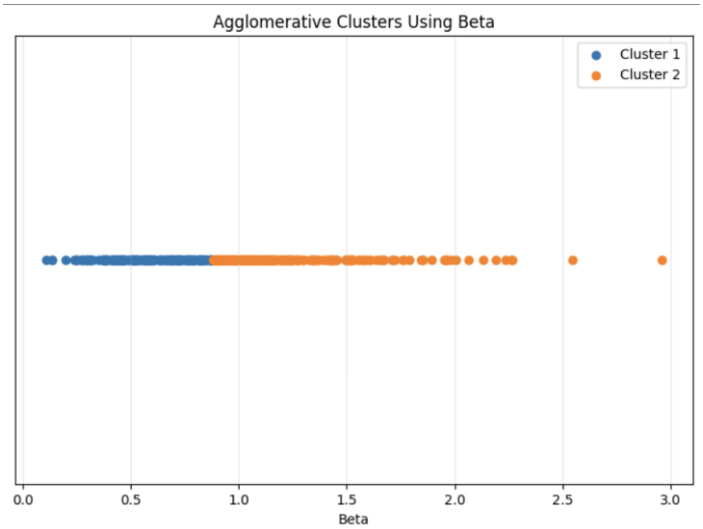


3C. Implementation and Results

After identifying the cut point from the dendrogram, I assigned each stock to a beta-based cluster. This grouping reflects how stocks differ in their sensitivity to market movements. To visualise the results, I created a simple plot placing each stock along a horizontal Beta axis. This approach works well because agglomerative clustering in this section is based on a single metric, so a one-dimensional display is appropriate and easy to read.

The plot confirmed a clear separation:

- One cluster contained stocks with lower beta values, and
- The other cluster contained those with higher beta values.



3D. Interpretation of Agglomerative Clusters

The two clusters represent distinct categories of systematic risk in the S&P 500:

- Lower-Beta Cluster: These companies tend to move less dramatically than the market. They usually experience smaller price fluctuations during broad market changes and are often associated with more stable or defensive sectors. They carry lower systematic risk.
- Higher-Beta Cluster: These companies are sensitive relative to market movements. Their prices tend to rise more during market upturns but also fall more during downturns. This group typically carries higher systematic risk.

This first clustering exercise provides a simple segmentation based solely on market sensitivity. It creates a foundation for the more detailed segmentation in Part 4, where both Beta and annualised volatility are combined using the K-Means method.

PART 4: K-Means Clustering (Beta and Annual Volatility)

In this section, K-Means clustering was used to segment S&P 500 companies based on Beta and Annualised Volatility. These two metrics together provide a more complete picture of a stock's risk profile: beta captures sensitivity to the market, while volatility reflects the total variability of returns. Combining them allows for a more in depth segmentation than using beta alone.

4A. Why K-Means is Appropriate for This Task

K-Means clustering is a suitable method for this task because:

- It works well with continuous numerical variables such as beta and volatility.
- It groups observations by minimising variance in a cluser, which aligns with the goal of identifying stocks with similar risk behaviour.
- It scales effectively to large datasets like the S&P 500, making it efficient.
- It produces clear and interpretable cluster boundaries, valuable for a marketing/strategy target who need simple, visual segmentations.

Before applying the model, I standardised both metrics using a StandardScaler. This is necessary because K-Means relies on Euclidean distance: variables measured on different scales would distort the clustering. Scaling ensures that beta and volatility contribute equally to the model.

From a scientific perspective, the choice of K-Means is justified because:

- The algorithm assumes roughly spherical clusters centred around their means.
- Beta and volatility jointly produce patterns that are often compatible with this assumption.

- The technique provides a deterministic method for minimising intra-cluster variance.

K-Means comes with limitations, such as sensitivity to outliers and the assumption of equal variance across clusters, it remains an appropriate choice for market segmentation insights.

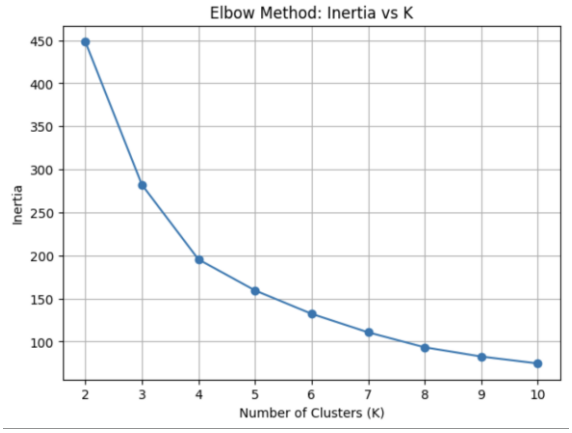
4B. Determining the Optimal Number of Clusters (Elbow Method)

To determine the number of clusters, I applied the Elbow Method, which evaluates how model inertia decreases as K increases. I tested values from K=2 to K=10 and plotted the inertia curve.

The graph showed a clear elbow at K = 4, where the curve shape started to change. This indicates that adding more clusters beyond four adds unnecessary complexity, and unclear clustering.

Choosing K=4 balances interpretability and accuracy, important in a business intelligence context where segmentation must be clear, and easy to communicate to decision-makers.

Elbow Plot:



4C. Implementation in Python and Cluster Assignment

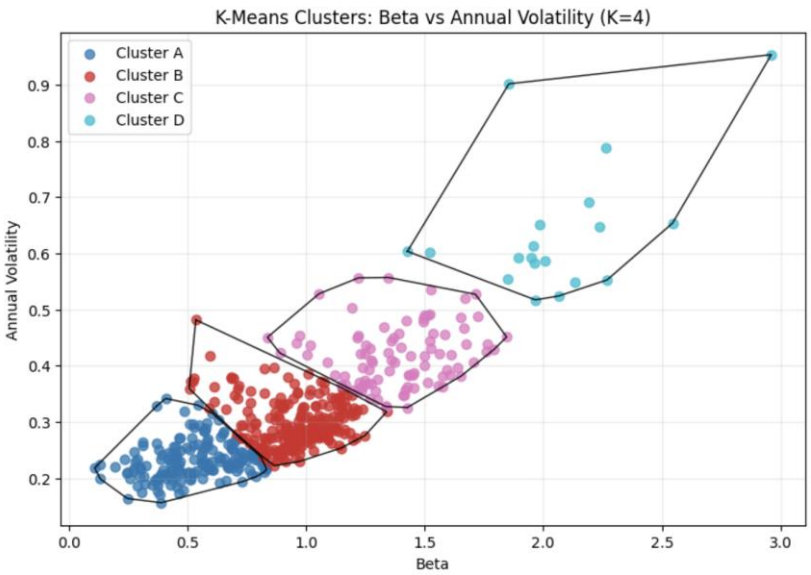
After selecting K=4, I fitted the K-Means model to the scaled dataset. Each stock was assigned a cluster label based on its beta and volatility. I then merged these labels back into the main DataFrame to analyse the characteristics of each segment.

To visualise the segmentation, I created a scatterplot with:

- Beta on the x-axis
- Annualised Volatility on the y-axis
- Colours showing different clusters
- Convex hulls outlining each cluster's boundary

Convex hulls create a clean and intuitive visual boundary without imposing assumptions about shape. This helps marketing and strategy teams understand the structure of the segments without needing statistical knowledge.

K-Means plot:



4D. Cluster Profiles and Interpretation

The clusters can be interpreted as follows:

Cluster A: Low Beta, Low Volatility ("Defensive Stocks")

These companies move less than the market and show stable price behaviour. They are suitable for low risk and defensive investment strategies.

Cluster B: Moderate Beta, Moderate Volatility ("Market-Aligned")

This group exhibits behaviour close to the overall market. These stocks tend to be suitable for balanced or core portfolios.

Cluster C: Higher Beta, Higher Volatility ("Aggressive Growth")

These companies amplify market trends and exhibit more pronounced price swings. They suit investors who are comfortable with higher risk in exchange for potentially higher returns.

Cluster D: Very High Beta, Very High Volatility ("Speculative")

This segment contains the riskiest stocks. Their behaviour is highly sensitive to market movements and internal company events. These companies may offer strong gains during favourable conditions but expose investors to significant downside risk.

PART 5: Review of Results and Business Insights

This section brings together the results from all previous stages of the analysis and evaluates how the calculated financial metrics; daily returns, beta, and annualised volatility, can be used to draw insights or make analysis from the behaviour of S&P 500 companies. It also reflects on the clustering outcomes and how they can be interpreted for a business intelligence and strategy perspective.

5A. Linking Financial Metrics

The three metrics calculated in this project are closely connected and help explain why the clusters formed as they did:

- **Daily Returns** were the foundation for all later calculations. They capture the stock's short-term behaviour and provide the data needed to estimate both volatility and beta.
- **Annualised Volatility** increases when daily returns fluctuate more widely. Stocks with inconsistent or unstable day-to-day movements naturally show higher annualised volatility.
- **Beta** depends on how much a stock's returns move together with the market. A company can have high volatility but still a low beta if its movements do not follow the overall index, and the opposite is also possible.

These relationships explain why volatility and beta do not always move in the same direction, and why the K-Means clusters reflect different combinations of systematic and total risk.

5B. Insights From Agglomerative Clustering

The agglomerative clustering method, applied solely to Beta, showed that the S&P 500 naturally separates into two broad groups:

- **Lower-Beta Stocks**: These firms move less than the market and are typically found in defensive or stable sectors.
- **Higher-Beta Stocks**: These firms amplify market movement and carry higher systematic risk.

This simple two-way classification highlights the first layer of risk segmentation: whether a stock behaves quietly during market movements or reacts strongly to them. This insight is valuable for distinguishing between stable and more aggressive investment opportunities.

5C. Insights From K-Means Clustering (Beta + Volatility)

The K-Means clustering method, which used both Beta and annualised volatility, produced a more detailed segmentation of four distinct clusters. This deeper analysis shows how different stocks combine market sensitivity with total risk:

- Some stocks have low beta and low volatility, making them attractive for defensive strategies.
- Others have high beta and high volatility, aligning with aggressive or growth-focused portfolios.
- The remaining groups fall in intermediate positions, helping identify balanced or moderately risky stocks.

These results show how combining multiple financial metrics provides a clearer and more realistic picture of market behaviour than using beta alone.

5D. Practical Value for Business Intelligence

The clustering results are useful for investors and analysts because they highlight patterns that may not be obvious from looking at individual stocks:

- **Risk-Based Portfolio Construction**: Investors can choose clusters depending on whether they prefer low-risk, balanced, or high-risk strategies.
- **Diversification**: Understanding how stocks behave in relation to the market helps build portfolios that spread risk effectively.
- **Market Sensitivity Analysis**: By identifying companies that react strongly to market conditions, investors can make better decisions during periods of volatility or economic uncertainty.
- **Targeting Growth Opportunities**: High-beta and high-volatility clusters may offer higher potential returns during favourable market conditions.
- **Defensive Positioning**: Low-beta clusters are helpful for protecting portfolios during downturns.

Overall, the clustering methods provide a clear, data-driven way of grouping companies based on their risk characteristics. This can support better decision-making for both short-term trading and long-term investment strategies.