Application of Unsupervised Machine Learning for Quantitative Trading

QF209 - GROUP 7
JEROME WONG JEN HOE
LIM WEI JIE
HTOO MYAT NAING
LIU SIRUI
NGUYEN NGOC QUYNH TRAM (TRACY)

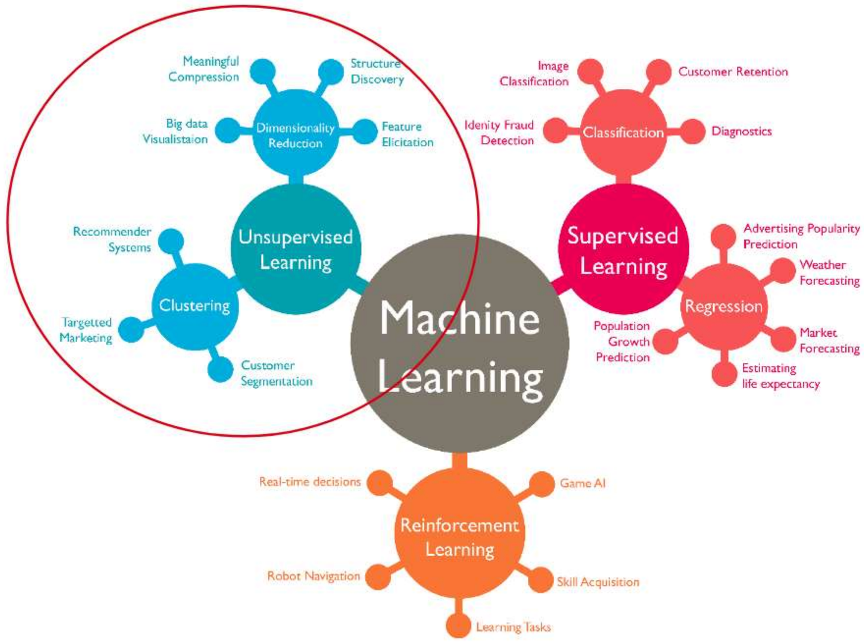
Problem Statement

Problem	Challenge	Solution/Goal
To optimize portfolio management in the financial market through the formation of coherent groups of financial assets.	Identifying the most efficient clustering technique for financial assets.	Develop a trading strategy that enhances risk-adjusted returns and balances risk and reward.

Speaker: Jerome

Overview of Unsupervised Learning

Definition	Objective	Key Techniques
Unsupervised learning is a branch of machine learning that identify patterns and relationships in data without labelled outcomes.	Discover hidden patterns, group data into clusters, or reduce dimensionality.	K-Means, Agglomerative Clustering, Principal Component Analysis (PCA).



Speaker: Jerome

Overview of Clustering in Trading

What	Objective and Advantages	Key Techniques
Clustering is a critical component of trading strategy development.	***Objective***: Group similar financial assets based on historical data. ***Advantage***: *Enhance portfolio diversification*	K-Means, Agglomerative Clustering.

Speaker: Jerome

Overview of Trading Strategies

Components	Objective
K-Means and **Agglomerative Clustering** **Clustering Financial Assets** • Identifies clusters of assets with akin behaviors • Provides foundation for the trading strategy **Capital Allocation Strategy** - Categorizing assets into clusters for efficient capital allocation - Allocate capital to assets with higher expected returns while managing volatility	***Objective***: Achieve maximum risk-adjusted returns while balancing risks effectively

Speaker: Jerome

Success Metrics of Trading Strategies

Key Metrics

(Used to evaluate the strategy's effectiveness and risk-return trade-off)

Annualized returns
- calculates expected returns

Annualized volatility
- measures variation of returns over time

Annualized Sharpe ratio
- quantifies risk-adjusted returns

Maximum drawdown
- maximum loss experienced during trading history

Sortino ratio
- measures risk-adjusted returns, penalize returns falling below target

Calmar ratio
- Compares average annual rate of return to maximum drawdown

Speaker: Jerome

Data Collection and Preprocessing

- Collected a list of S&P 500 companies from Wikipedia
- Downloaded historical data for these companies using Yahoo Finance
- Used the given features (open, high, low, close, adjusted close, and volume) to calculate features such as Garman-Klass Volatility, Relative Strength Index (RSI), Bollinger Bands, Average True Range (ATR), Moving Average Convergence Divergence (MACD), and Dollar Volume

date	ticker	adj close	close	high	low	open	volume
2015-09-29	A	31.588043	33.740002	34.060001	33.240002	33.360001	2252400.0
	AAL	37.361626	39.180000	39.770000	38.790001	39.049999	7478800.0
	AAPL	24.748627	27.264999	28.377501	26.965000	28.207500	293461600.0
	ABBV	37.024632	52.790001	54.189999	51.880001	53.099998	12842800.0
	ABT	33.807266	39.500000	40.150002	39.029999	39.259998	12287500.0
...	...	...	...	...	...	...	...
2023-09-26	YUM	124.010002	124.010002	124.739998	123.449997	124.239998	1500600.0
	ZBH	112.216316	112.459999	117.110001	112.419998	116.769997	3610500.0
	ZBRA	223.960007	223.960007	226.649994	222.580002	225.970001	355400.0
	ZION	33.990002	33.990002	34.700001	33.840000	33.840000	1586100.0
	ZTS	176.869995	176.869995	178.449997	176.270004	176.580002	1463200.0

994088 rows × 6 columns

Speaker: Wei Jie

HMN0

Data Collection and Preprocessing

- Converted the data from daily to monthly data and filtered the top 150 most liquid stocks
- Introduced the Fama-French 5 factors (2x3) model to retrieve financial features (Market Risk Premium, Small Minus Big, High Minus Low, Robust Minus Weak, Conservative Minus Aggressive) for our clustering models, by computing rolling factor betas for each stock using a rolling Ordinary Least Squares (OLS) approach
- Total 18 columns (features) to be used in the clustering models

date	ticker	atr	bb_high	bb_low	bb_mid	garman_klass_vol	macd	rsi	1m_return	2m_return	3m_return	6m_return	9m_return	12m_return	Mkt-RF	SMB	HML	RMW	CMA
2017-10-31	AAL	1.011062	3.994389	3.849110	3.921750	-0.000363	-0.018697	41.051793	-0.014108	0.022981	-0.023860	0.016495	0.007008	0.012702	1.261817	1.306500	0.632023	0.507423	0.557001
	AAPL	-0.906642	3.692324	3.598569	3.645446	-0.000892	-0.039276	69.196794	0.096808	0.015250	0.044955	0.028875	0.038941	0.035228	1.280257	-0.272363	-0.609509	0.663830	0.487736
	ABBV	0.375557	4.307973	4.215227	4.261600	-0.029822	0.473813	55.247815	0.022728	0.098590	0.091379	0.056495	0.047273	0.044026	0.495964	0.376215	-0.016727	0.231814	0.123746
	ABT	-1.040044	3.949284	3.902136	3.925710	-0.004349	0.276133	53.844948	0.021276	0.034308	0.034801	0.038672	0.031320	0.029294	0.831881	-0.213285	-0.539412	0.236860	0.981357
	ACN	-0.986514	4.889487	4.810123	4.849805	-0.003359	0.352343	69.365269	0.064180	0.048455	0.037203	0.028692	0.027398	0.018728	1.197481	-0.157159	-0.330354	0.264165	0.179181
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2023-09-30	VZ	-1.078816	3.563843	3.499366	3.531604	-0.000067	-0.350385	42.222474	-0.056890	-0.016122	-0.033458	-0.021495	-0.014100	-0.006158	0.518738	-0.365186	0.013465	0.313425	0.617454
	WFC	-0.558742	3.798900	3.718132	3.758516	0.000234	-0.282325	40.920274	-0.015500	-0.057917	-0.013554	0.016712	0.000702	0.003255	1.093984	-0.140007	1.309152	-0.750923	-0.409061
	WMT	-0.196379	5.116986	5.081613	5.099300	0.000024	0.399459	54.722508	-0.000676	0.010014	0.012354	0.017574	0.016553	0.020256	0.615232	-0.473891	-0.291126	0.409354	0.729464
	XOM	0.601335	4.793504	4.713293	4.753399	0.000045	1.400623	59.440192	0.046947	0.046139	0.030496	0.012838	0.008747	0.027037	1.168319	0.387311	0.513399	-0.467915	0.799639
	MRNA	-0.529511	4.788149	4.582514	4.685332	0.000146	-0.376899	38.747314	-0.132219	-0.086803	-0.068763	-0.071952	-0.064976	-0.015431	1.301327	-0.186763	-1.183057	1.125993	0.527068

Speaker: Wei Jie

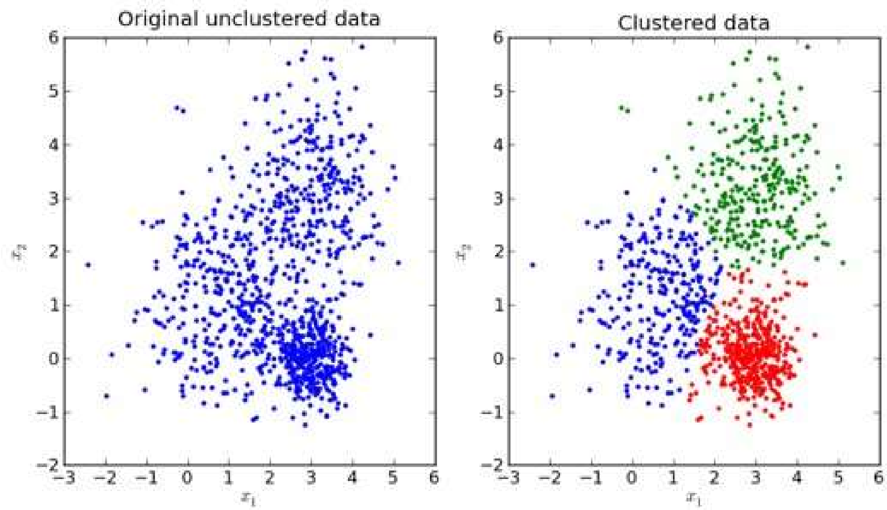
Slide 8

HMN0 Need to roughly explain how we got the FF5 data as well?

Htoo Myat Naing, 2023-11-06T07:11:14.606

Model 1: K-Means Clustering (What)

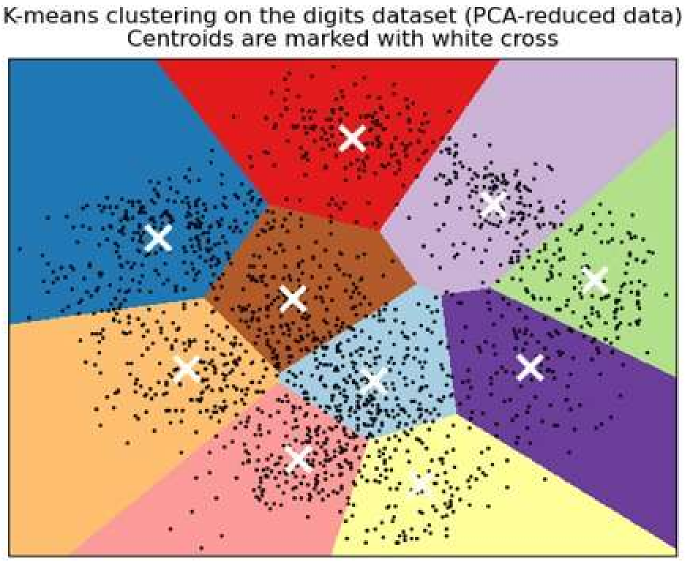
- Unsupervised machine learning algorithm
- Used for segmenting data into distinct groups or clusters based on similarity patterns
- Goal is to minimize the within-cluster variance, which is the sum of squared distances between data points and their respective cluster centroids
- The result is a partitioning of the data into K distinct clusters, where K represents number of clusters defined by the user



Speaker: Wei Jie

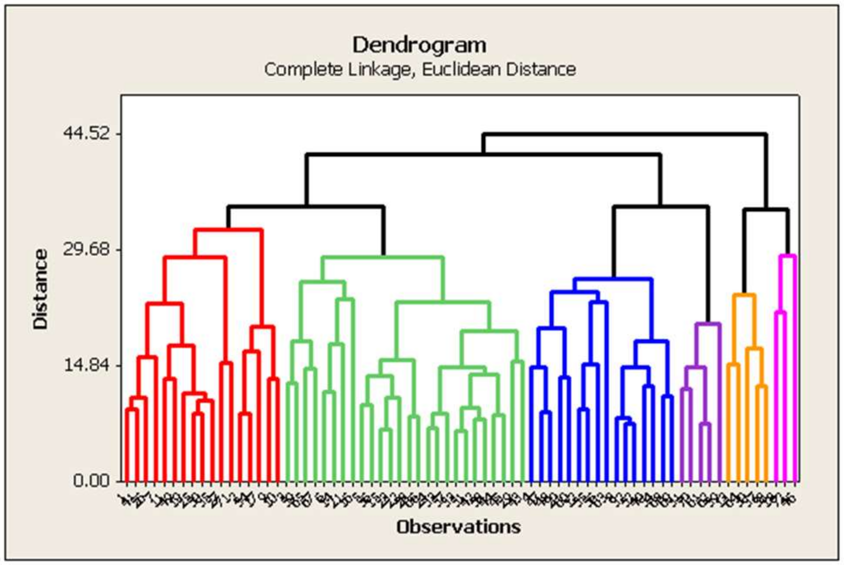
Model 1: K-Means Clustering (Why)

- Used in quantitative finance because of its simplicity and ease of implementation
- Ideal baseline model for understanding financial data patterns
- Computationally efficient and scales well to large and complex datasets, able to quickly yield meaningful insights from them
- A stepping stone for developing more tailored and accurate unsupervised machine learning models suited to specific financial applications



Speaker: Wei Jie

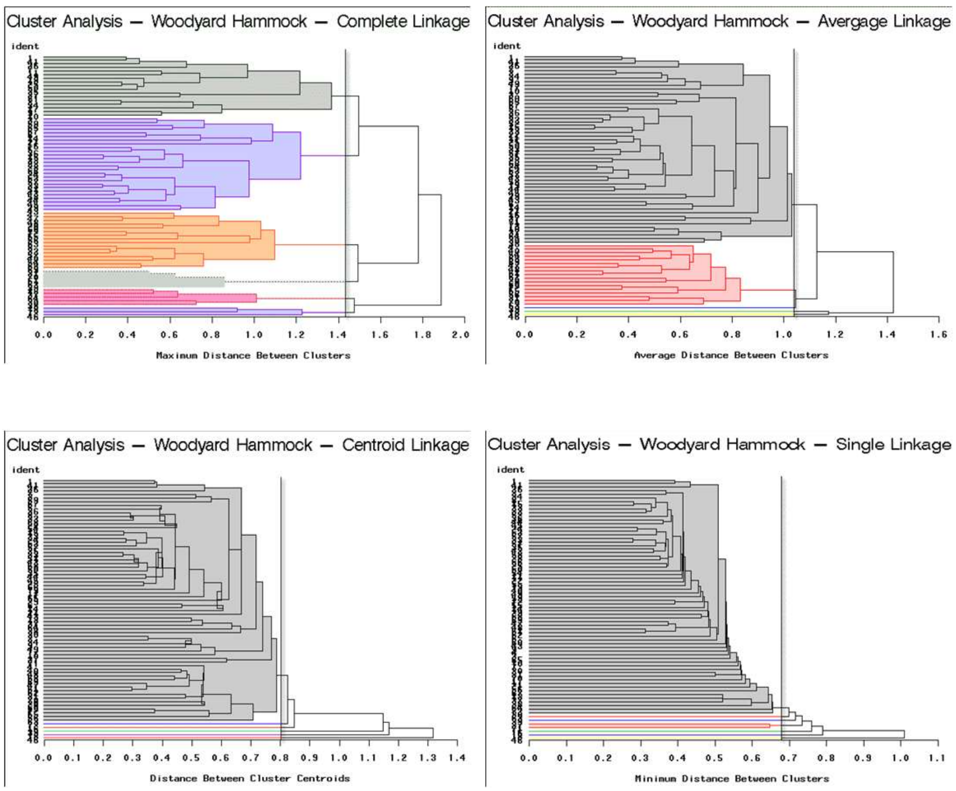
Model 2: Agglomerative Clustering (What)



- A hierarchical clustering method
- Builds a multilevel hierarchy of clusters by starting with individual data points and iteratively merging them into larger clusters ("bottom-up" approach)
- The result can be visualised using a tree-like diagram that records the sequences of merges and shows the arrangement of the clusters produced by the algorithm

Speaker: Wei Jie

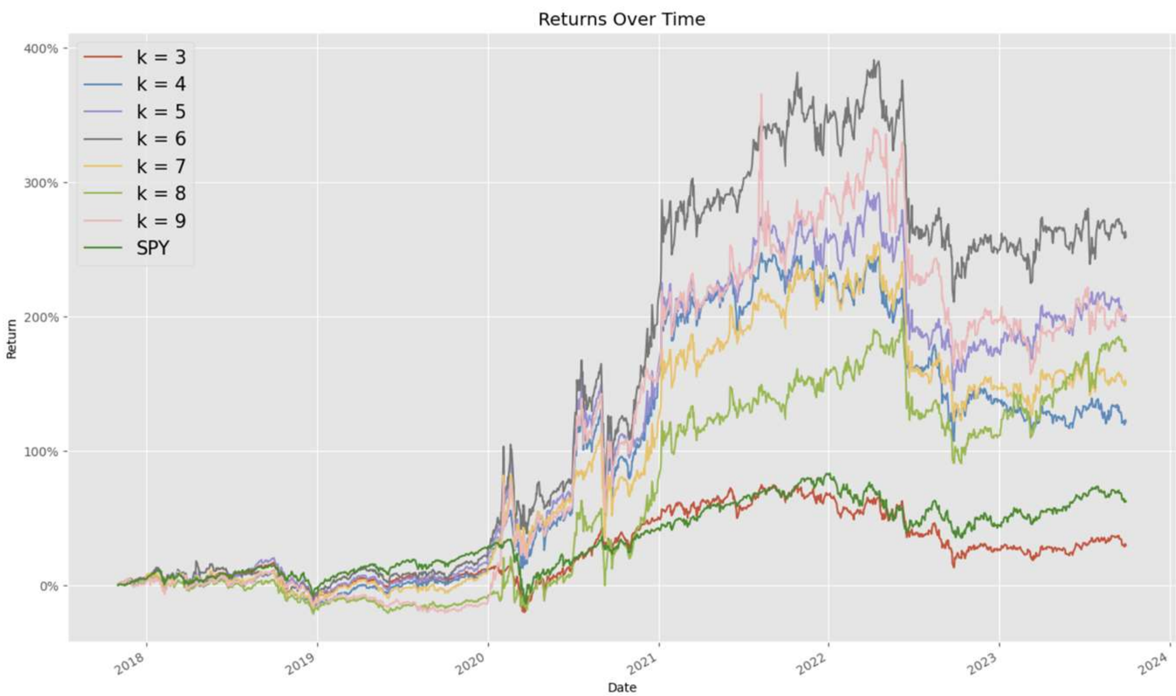
Model 2: Agglomerative Clustering (Why)



- Relationships between different stocks can be complex and not well-defined
- Unlike K-means, which assumes spherical clusters, agglomerative clustering can find clusters of various shapes and sizes
- More robust to noise and outliers as it does not compute a central point (means) that represents a cluster
- Builds clusters based on the proximity of individual data points to one another, and merges clusters based on the chosen linkage criterion
- Clusters formed might be more meaningful and provide more intricate structures and relationships between stocks

Speaker: Wei Jie

Results: K-Means Clustering Algorithm

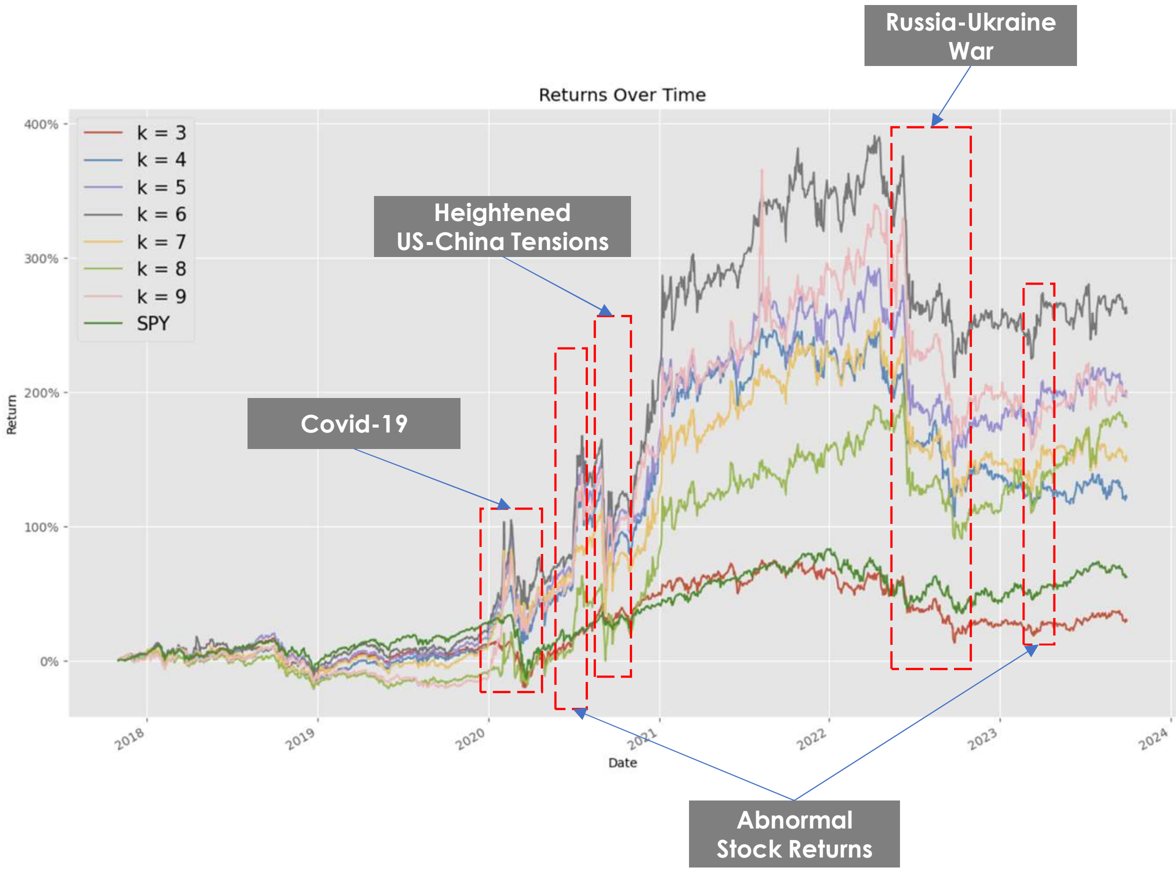


Observation 1:
Portfolio returns are positively correlated to SPY's movement

- As SPY index rises or falls, the portfolio returns tend to follow similarly
- Increase in the portfolio's sensitivity to SPY fluctuations as the value of k rises, particularly to k = 6
- This is likely due to a more selective stock picking process within each cluster as k increases from 3 to 6, leading to larger than proportionate movements in relation to SPY
- As k gets larger from 6 to 9, the portfolio's sensitivity to SPY fluctuations decreases
- This is likely due to chosen cluster to have a highly elevated mean RSI value, resulting in the stocks chosen to likely be largely overvalued, leading to smaller returns compare to the k = 6 portfolio

Speaker: Htoo

Results: K-Means Clustering Algorithm

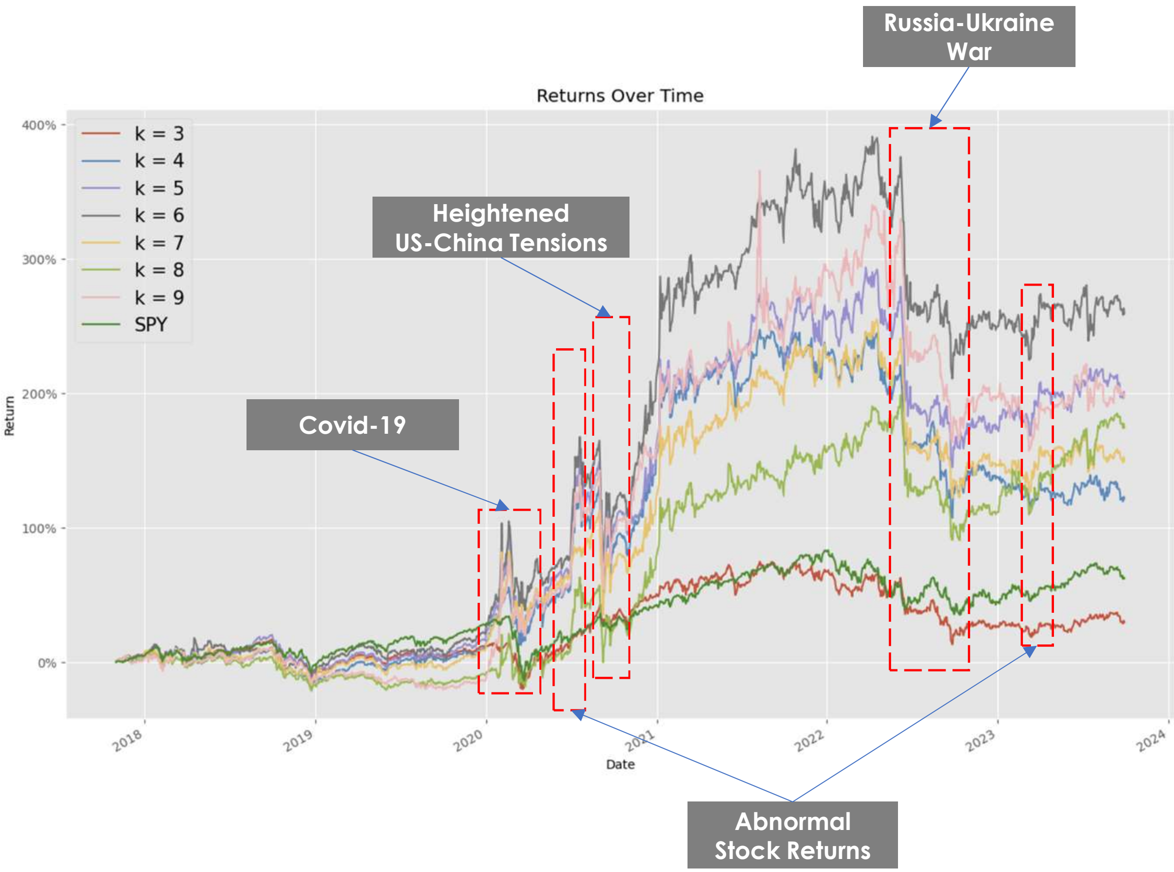


Observation 2:
Market occurrences exert a considerable influence on portfolio returns

- In early 2020, there was a synchronized downturn in returns, followed by a recovery, likely reflects market's initial reactions to the Covid-19 pandemic and its subsequent rally as global restriction eases
- In Q3 of 2020, sharp increase in portfolio returns can be attributed to the performance of specific stocks such as Amazon and Microsoft, which significantly exceeded industry expectations in its Q2 earnings
- However, further escalation of the existing US-China tensions contributed to increased volatility and widespread selling during this period, leading to fall in returns thereafter
- In 2022, decline in portfolio returns and heightened volatility is likely due to geopolitical tensions arising from Russia's invasion of Ukraine, which led to uncertainty in the financial markets

Speaker: Htoo

Results: K-Means Clustering Algorithm



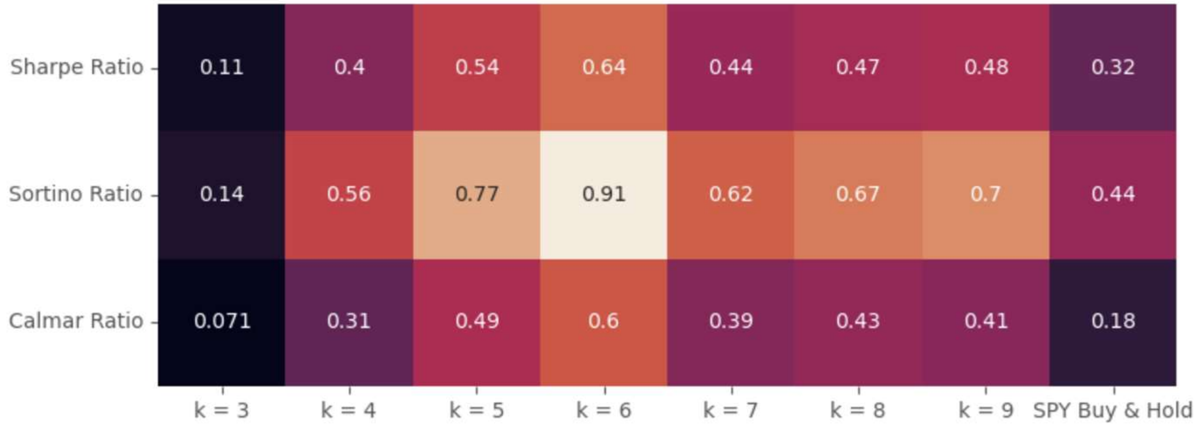
Observation 2:
Market occurrences exert a considerable influence on portfolio returns

- Surge in returns at the start of 2023 attributed to a certain number of socks
- Tesla and Warner Bros Discovery climbing over 60%, and other firms like Catalent, Align Technology and Royal Caribbean seeing ~50% increases
- Despite the diverse industry representation, these stocks were among the lowest performing stocks in 2022
- For portfolios with k >= 5, influence of these stocks are magnified due to the smaller cluster size

Speaker: Htoo

Results: K-Means Clustering Algorithm

	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	SPY Buy & Hold
Annualized Returns	0.045060	0.144383	0.204066	0.241837	0.167783	0.186481	0.202033	0.085671
Annualized Volatility	0.235926	0.313810	0.337998	0.348629	0.335167	0.353780	0.375393	0.204864
Downside Standard Dev	0.175158	0.224091	0.239229	0.243015	0.240216	0.248220	0.258742	0.150440
Max % Drawdown	-0.352543	-0.403325	-0.377897	-0.370571	-0.374851	-0.387648	-0.447837	-0.357459
Sharpe Ratio	0.106219	0.396364	0.544575	0.636313	0.440923	0.470576	0.484914	0.320557
Sortino Ratio	0.143070	0.555057	0.769411	0.912853	0.615209	0.670698	0.703532	0.436525
Calmar Ratio	0.071083	0.308395	0.487079	0.598635	0.394245	0.429464	0.406473	0.183716



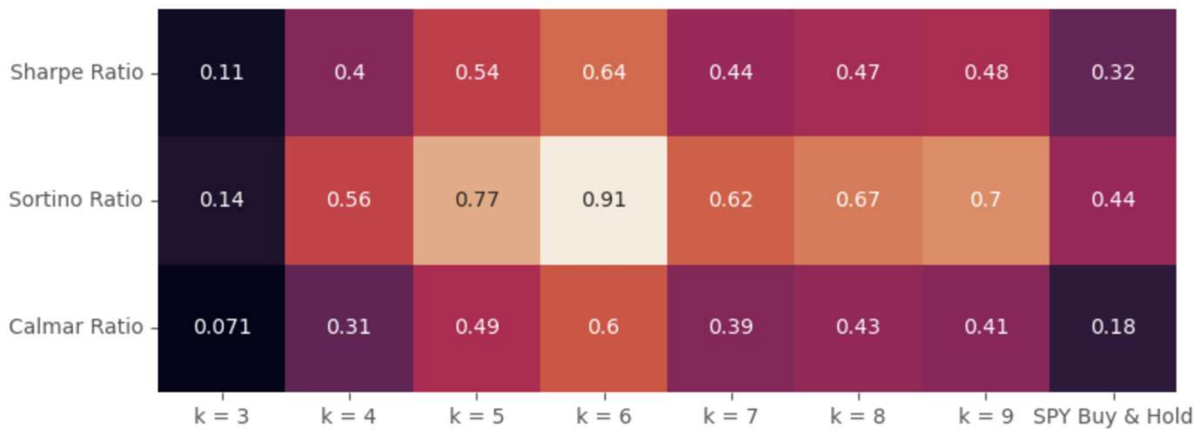
Observation 3:
K Means Clustering algorithm consistently outpaces SPY

- Annualized returns for all values of k > SPY returns
- Uptick in annualized volatility of the portfolio in comparison to SPY, which is likely tied to the portfolio's amplified response to SPY's movements
- As the k value increases, annualized volatility rises, indicating greater variability in returns as size of cluster decreases

Speaker: Htoo

Results: K-Means Clustering Algorithm

	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	SPY Buy & Hold
Annualized Returns	0.045060	0.144383	0.204066	0.241837	0.167783	0.186481	0.202033	0.085671
Annualized Volatility	0.235926	0.313810	0.337998	0.348629	0.335167	0.353780	0.375393	0.204864
Downside Standard Dev	0.175158	0.224091	0.239229	0.243015	0.240216	0.248220	0.258742	0.150440
Max % Drawdown	-0.352543	-0.403325	-0.377897	-0.370571	-0.374851	-0.387648	-0.447837	-0.357459
Sharpe Ratio	0.106219	0.396364	0.544575	0.636313	0.440923	0.470576	0.484914	0.320557
Sortino Ratio	0.143070	0.555057	0.769411	0.912853	0.615209	0.670698	0.703532	0.436525
Calmar Ratio	0.071083	0.308395	0.487079	0.598635	0.394245	0.429464	0.406473	0.183716

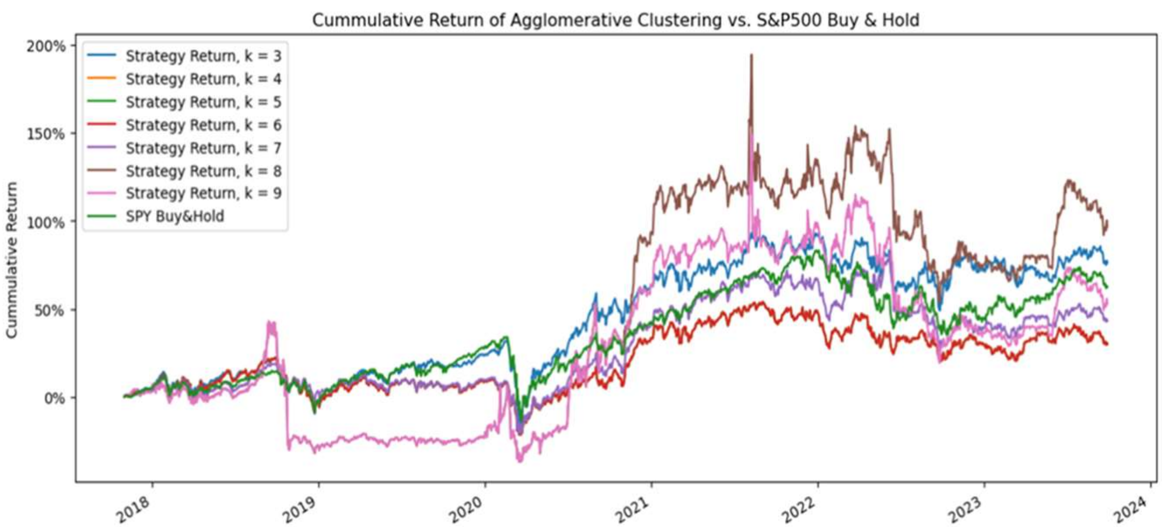


Observation 4:
Optimal trading strategy at k = 6

- Performance peaks at k = 6, suggesting an optimal clustering scenario for the trading strategy
- Sortino Ratios are higher than the Sharpe Ratios where k >= 4, implying that strategy's returns are likely robust against downside risks.
- However, this may also be influenced by a significant chunk of our trading period being in a bull market, which naturally elevates the upside potential and resulting in lower downside volatility
- Similar Sharpe and Calmar Ratios across portfolios indicates uniformity in risk adjusted returns and drawdown risks
- Highlights a consistent performance relative to the risks undertaken and denoting that trading strategy delivers results after accounting for both volatility and potential drawdown

Speaker: Htoo

Results: Agglomerative Clustering Algorithm



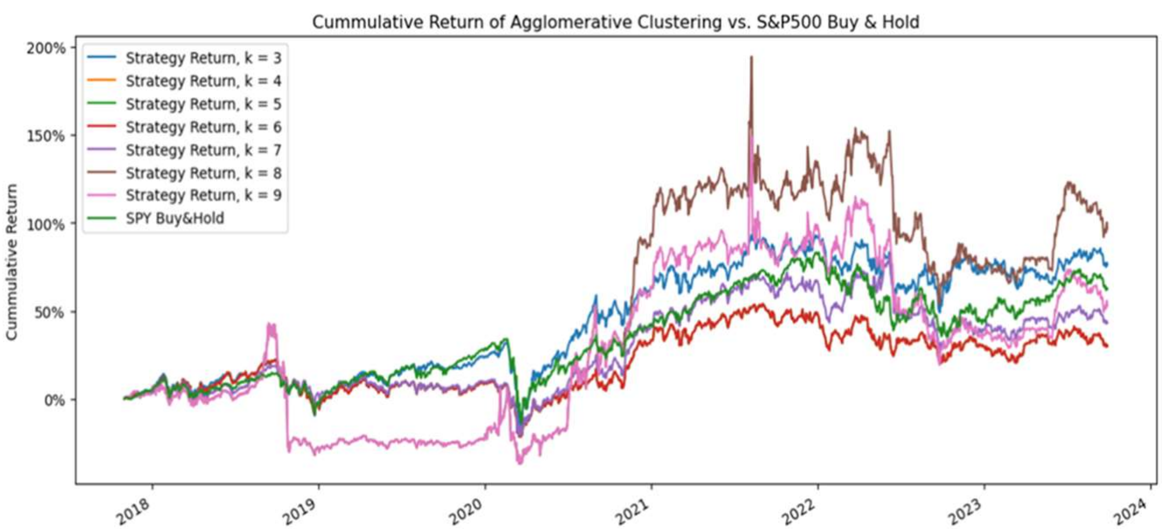
Observation 1:
Portfolio returns are positively correlated to SPY's movement

- As SPY index rises or falls, the portfolio returns tend to follow similarly

** Returns when k=4,5,6 are identical and overlapped in the figure*

Speaker: Sirui

Results: Agglomerative Clustering Algorithm

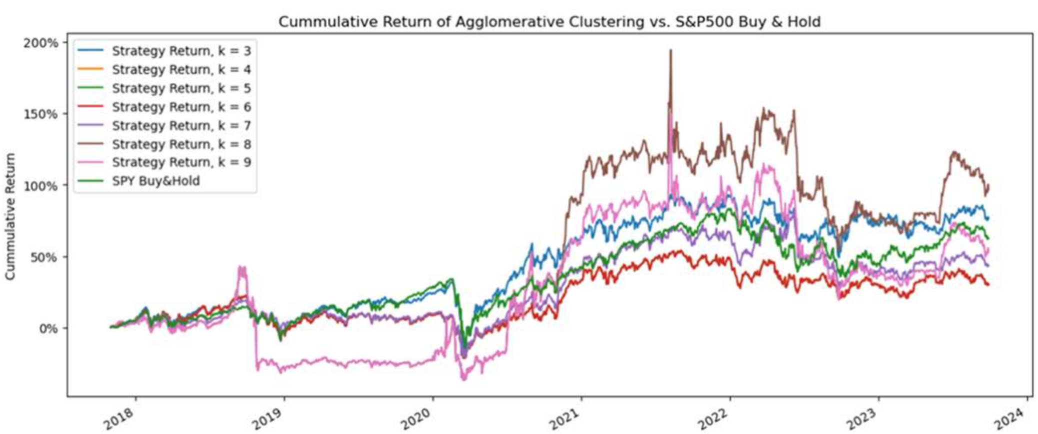


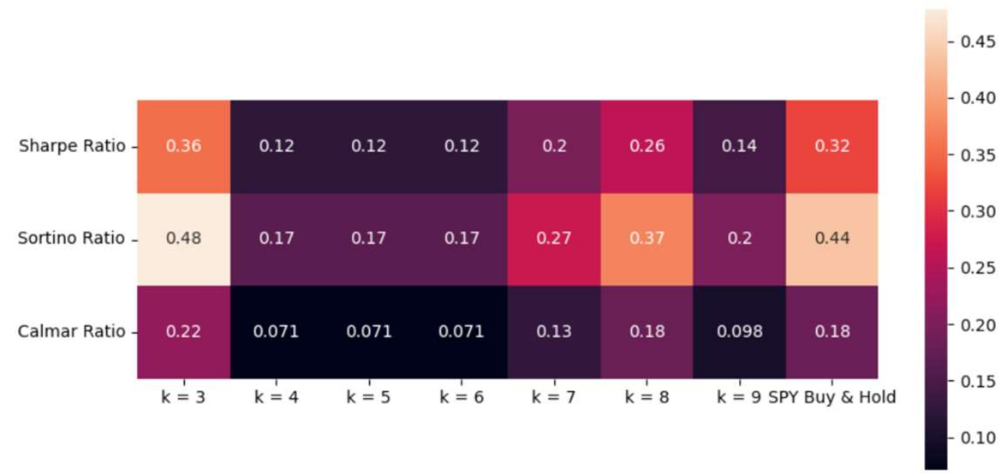
Observation 2:
There is no correlation between k and performance

- Dataset may not have a natural cluster structure. Clustering algorithms assume that the data contains inherent groupings
- As 'k' increases, there could be a risk of overfitting to noise in data

Speaker: Sirui

Results: Agglomerative Clustering Algorithm



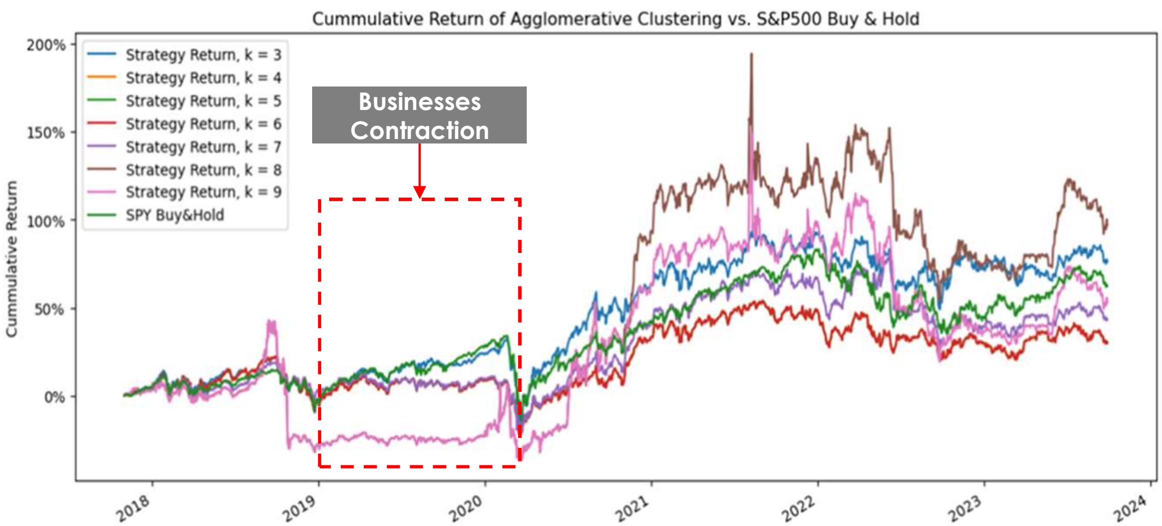


Observation 3:
K=8 outperforms S&P500 Buy and Hold strategy in terms of Returns

- K=8 could represent an optimal level of granularity that captures the most significant divisions in the data
- K=8 may help in reducing noise by focusing on broader movements and trends

Speaker: Sirui

Results: Agglomerative Clustering Algorithm



Observation 4: Strategies outperformed by S&P500 from 2019 to COVID-19 stock market crash

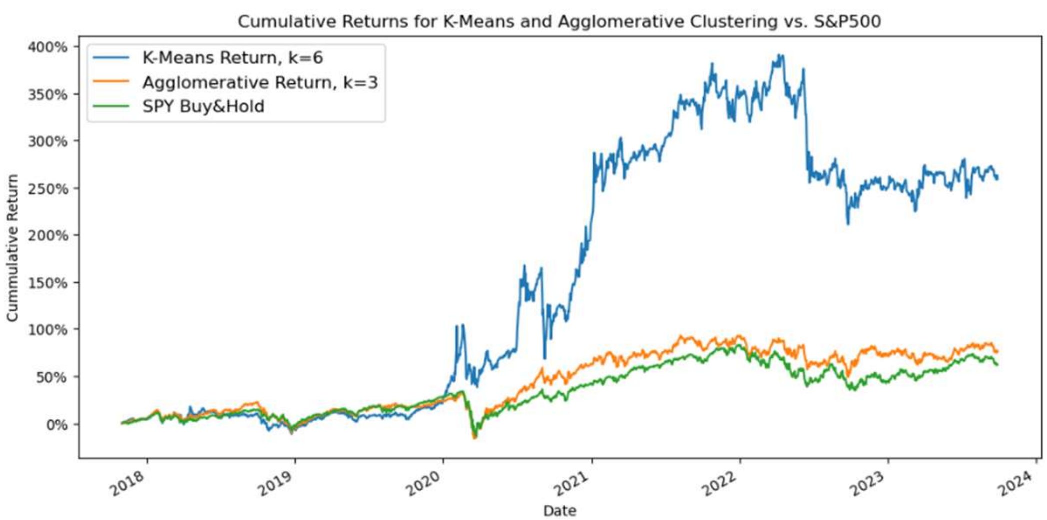
- 2019 ended with a contraction signal in businesses with ISM PMI below 50
- Coincided with former President Trump's aggressive stance against China during that period
- Market participants at that point proceeded with caution against high momentum high value stocks

Speaker: Sirui

Comparing Results of 2 Clustering Methods

** Clusters with highest **Sharpe Ratio** are chosen from each model.*

	K-Means, k=6	Agglomerative, k=3	SPY Buy&Hold
Annualized Returns	0.348629	0.227102	0.204864
Annualized Volatility	0.241837	0.100704	0.085671
Downside Standard Dev	0.243015	0.168825	0.150440
Max % Drawdown	-0.370571	-0.364674	-0.357459
Sharpe Ratio	0.636313	0.355367	0.320558
Sortino Ratio	0.912854	0.478036	0.436526
Calmar Ratio	0.598635	0.221306	0.183716



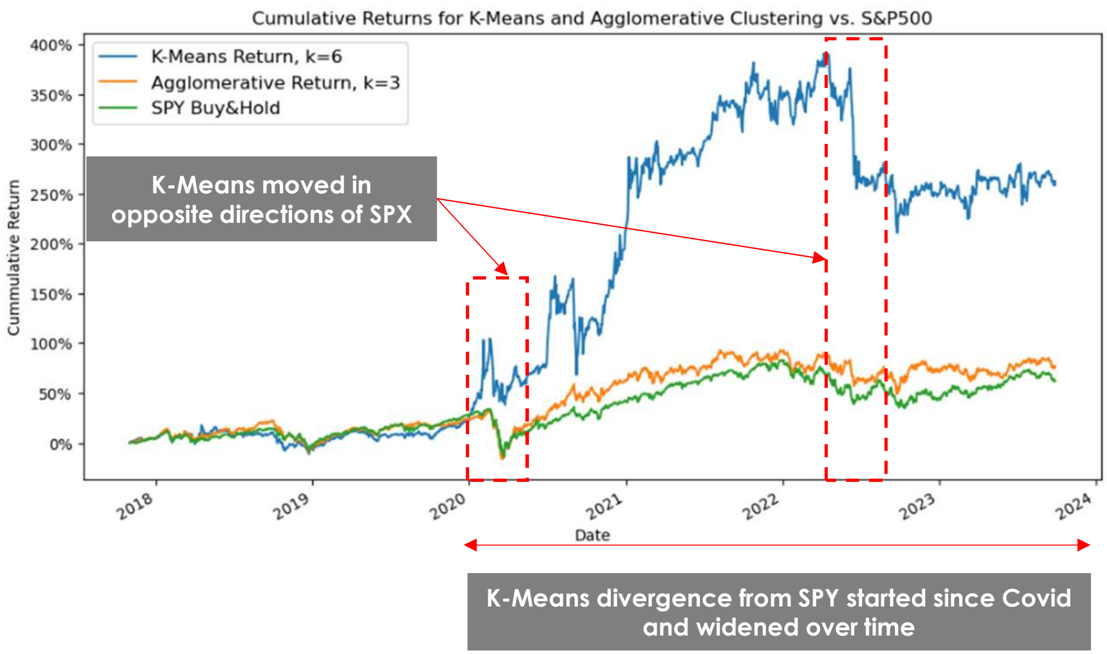
Observation 1:

K-Means Clustering greatly outperforms Agglomerative Clustering and SP500 Buy & Hold Strategy.

- Both strategies outperformed the S&P500 Buy & Hold.
- Although the K-Means Return appears significantly volatile, its Maximum Drawdown (-37%) is almost equal Agglomerative Clustering (-36%), and S&P 500 (-35%).
- **K-Means** (k=6) offers the **higher annualized returns** at a **higher risk**.
- **Agglomerative** (k=3) provides a **balance** between returns and risk while still outperformed the benchmark.

Speaker: Tracy

Comparing Results of 2 Clustering Methods



Observation 2:

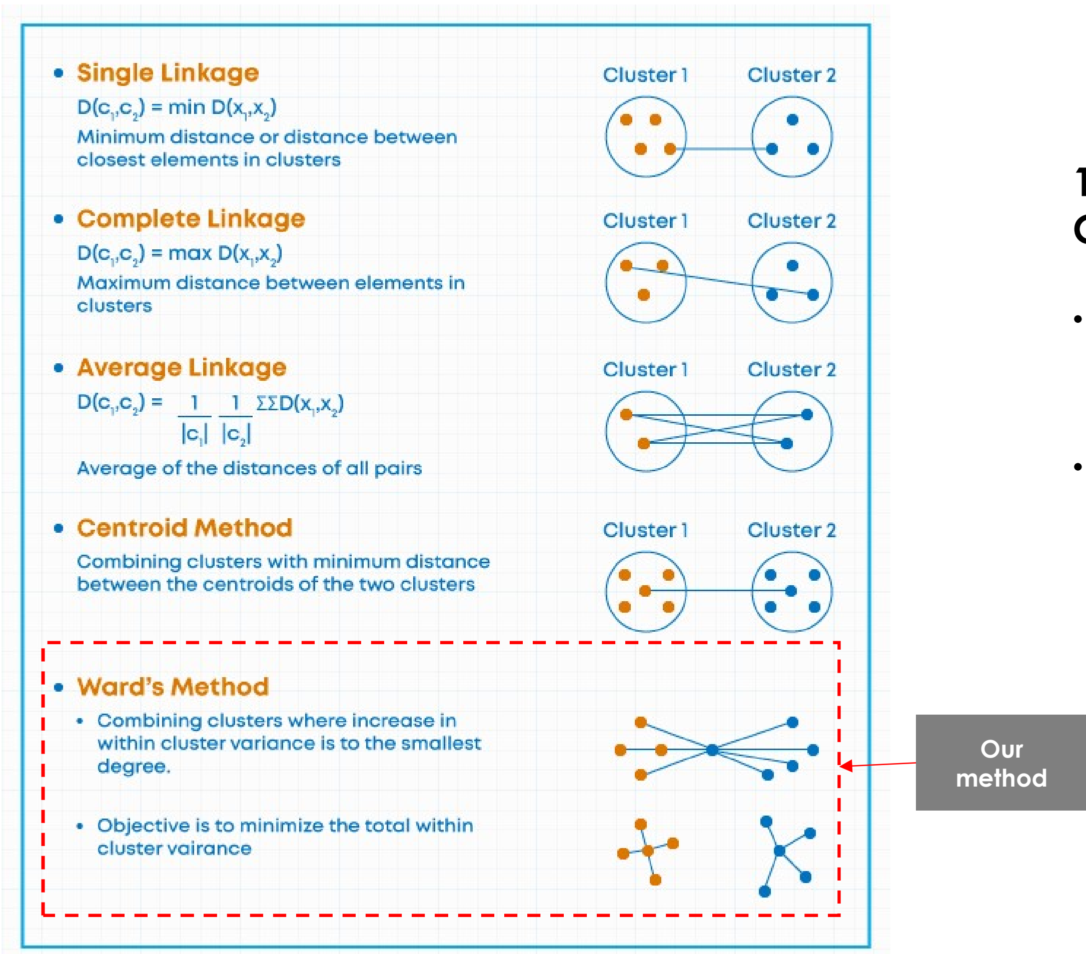
K-Means diverged in volatile markets while Agglomerative mirrored SPY throughout

- Agglomerative Strategy consistently beat SPX by a small margin → ***Recommended*** if investors expect ***bullish trends*** to capture alpha at ***minimal risks***.
- K-Means Strategy can capture ***more alphas in upturns*** but is ***unpredictable in downturns*** → pose higher risks.
  - During **Covid breakout**, K-Means Strategy delivered returns while market went down.
  - But during **Russian-Ukraine war**, while market was only mildly affected, K-Means significantly dipped.

→ K-Means Strategy provides ***benefit of the doubt in market downturns.***

Speaker: Tracy

Recommendations for Improvements (1)



1. Further experiments with Agglomerative Clustering Algorithm:

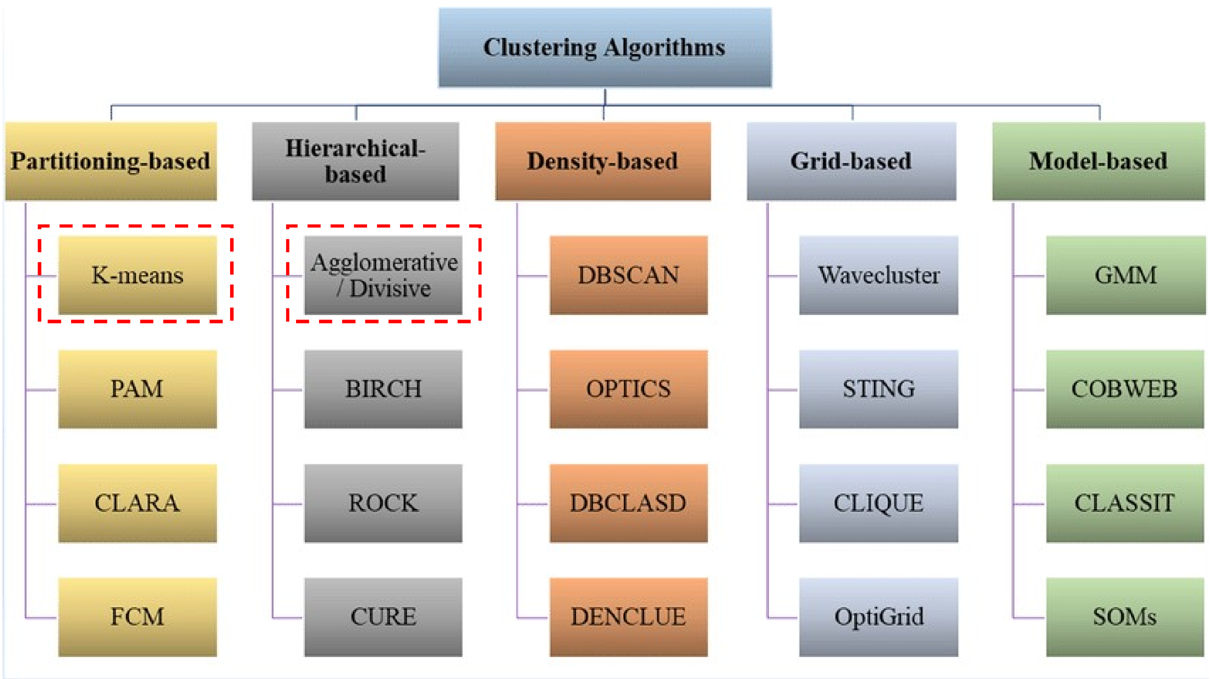
- Irregular patterns observed in performance from k=4 to k=9 suggest further experiments with the setup of **Agglomerative Clustering Algorithm**
- Changing **linkage methods** may provide insights in leads to different performance.
  - Ward's Method is commonly used to reduce noises.
  - However, it is biased towards globular clusters (spherical shapes) → not accurately represent data's structure if clusters have different shapes (elongated or irregular forms).

Speaker: Tracy

Recommendations for Improvements (2)

2. Further experiments with other Clustering Methods:

- Potential Clustering Methods to explore include ***Gaussian Mixture, DBSCAN.***



Speaker: Tracy

Appendix A: Ratio Calculations

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p}$$

$$Treynor\ Ratio = \frac{Portfolio\ Return - Risk\ Free\ Rate}{Portfolio\ Beta}$$

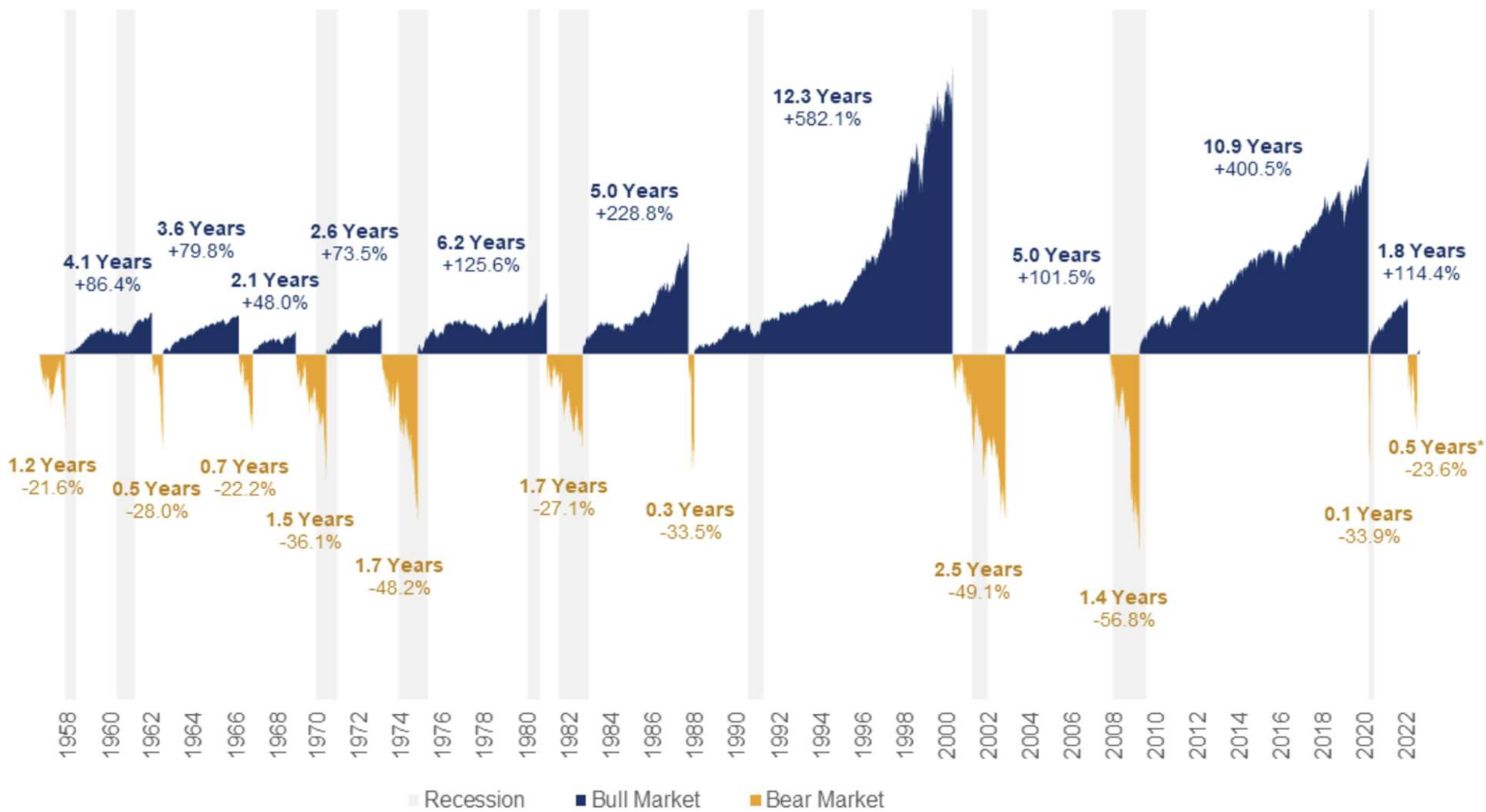
$$\text{Sortino Ratio} = \frac{R_p - r_f}{\sigma_d}$$

$$Jensen's\ \alpha = Rp - Rf - \beta p(Rm - Rf)$$

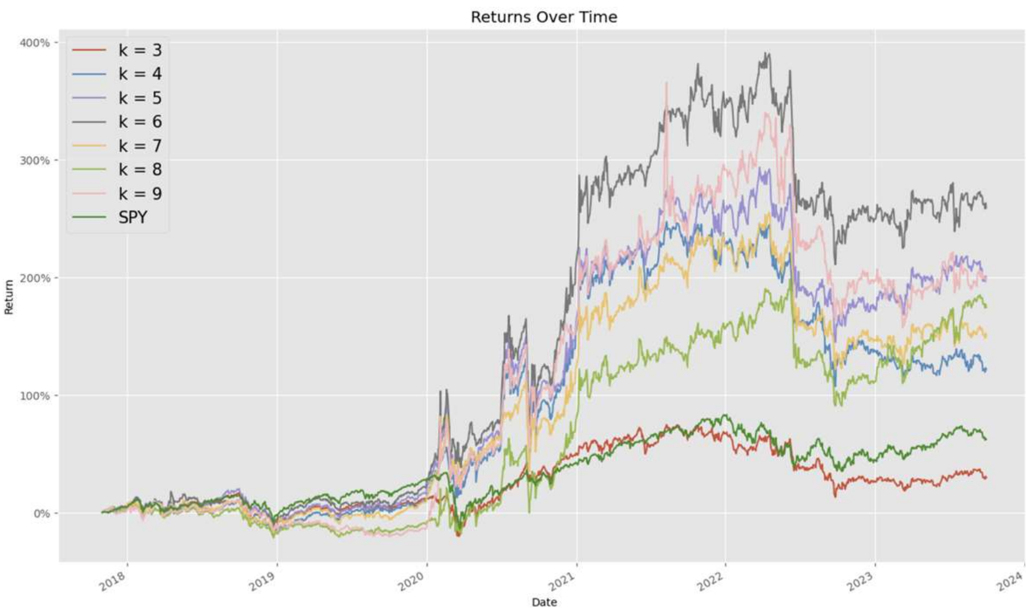
$$Calmar\ Ratio = \frac{r_p - r_T}{D_{Max}}$$

$$\beta_p = \frac{Cov(r_p, r_b)}{Var(r_b)}$$

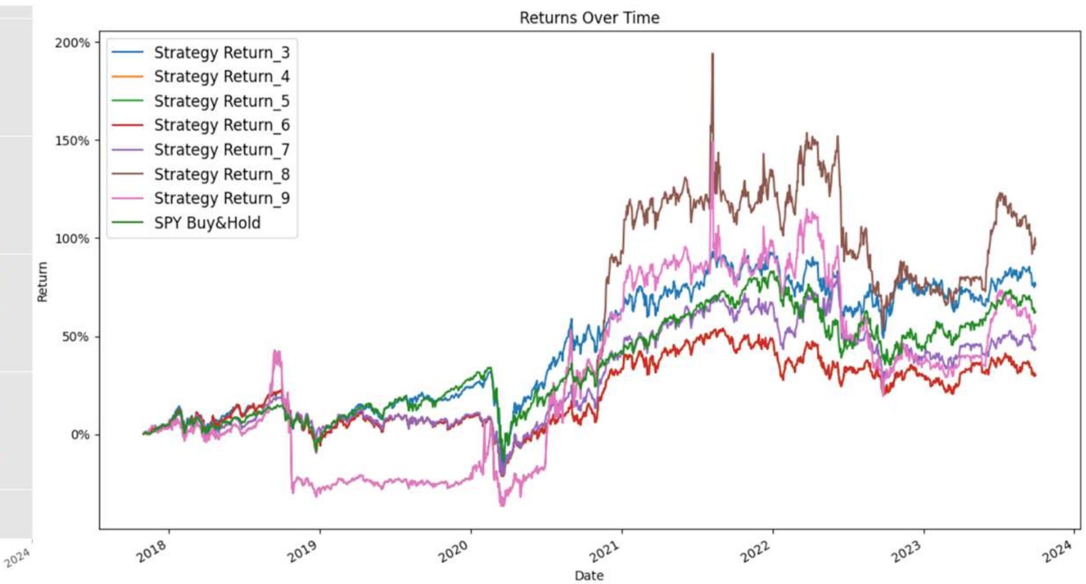
Appendix B: History of U.S equity of bull & bear markets by RBC



Appendix C: Side by Side comparison of K-Means vs Agglomerative



K Means



Agglomerative

Appendix C: Side by Side comparison of K-Means vs Agglomerative

K Means

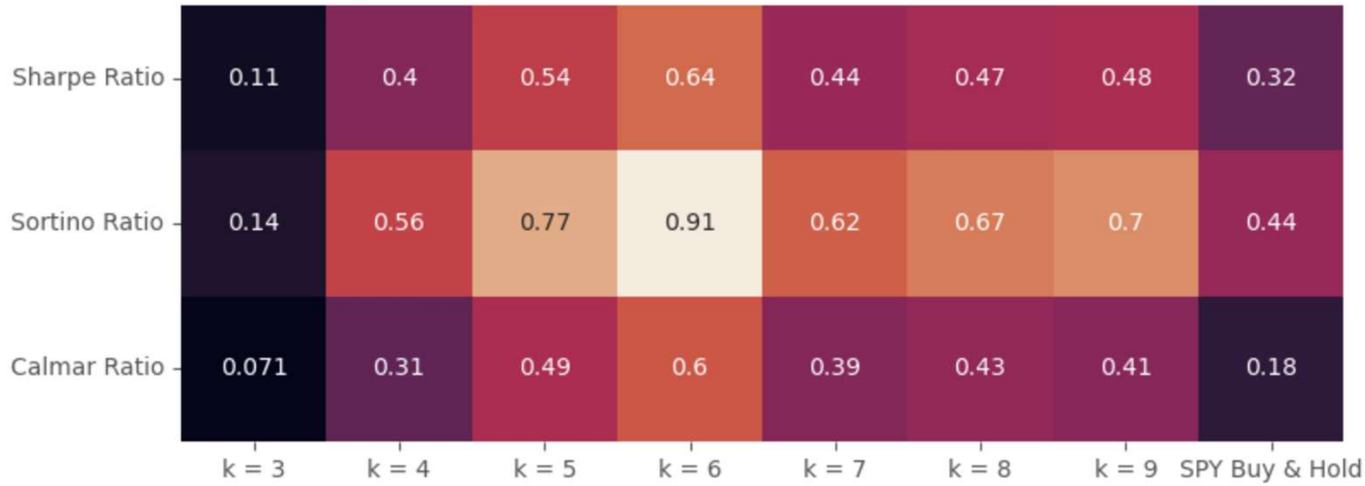
	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	SPY Buy & Hold
Annualized Returns	0.045060	0.144383	0.204066	0.241837	0.167783	0.186481	0.202033	0.085671
Annualized Volatility	0.235926	0.313810	0.337998	0.348629	0.335167	0.353780	0.375393	0.204864
Downside Standard Dev	0.175158	0.224091	0.239229	0.243015	0.240216	0.248220	0.258742	0.150440
Max % Drawdown	-0.352543	-0.403325	-0.377897	-0.370571	-0.374851	-0.387648	-0.447837	-0.357459
Sharpe Ratio	0.106219	0.396364	0.544575	0.636313	0.440923	0.470576	0.484914	0.320557
Sortino Ratio	0.143070	0.555057	0.769411	0.912853	0.615209	0.670698	0.703532	0.436525
Calmar Ratio	0.071083	0.308395	0.487079	0.598635	0.394245	0.429464	0.406473	0.183716

Agglomerative

	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	SPY Buy & Hold
Annualized Returns	0.100705	0.045498	0.045498	0.045498	0.062772	0.121418	0.074674	0.085671
Annualized Volatility	0.227101	0.207211	0.207211	0.207211	0.215308	0.386604	0.387010	0.204864
Downside Standard Dev	0.168825	0.151859	0.151859	0.151859	0.156604	0.270653	0.272404	0.150440
Max % Drawdown	-0.364673	-0.357938	-0.357938	-0.357938	-0.333815	-0.557110	-0.557110	-0.357459
Sharpe Ratio	0.355368	0.123054	0.123054	0.123054	0.198656	0.262330	0.141273	0.320557
Sortino Ratio	0.478039	0.167906	0.167906	0.167906	0.273124	0.374717	0.200709	0.436526
Calmar Ratio	0.221307	0.071236	0.071236	0.071236	0.128132	0.182043	0.098139	0.183716

Appendix C: Side by Side comparison of K-Means vs Agglomerative

K Means



Agglomerative