FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Quynh Trang Nguyen	Vietnam	nqtrang.2111@gmail.com	
Willytom Tanguy Rigaud	France	willytomrig@gmail.com	
Francis Kwami Dzikpe	USA	Fkdzikpe@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).	
Team member 1	Quynh Trang Nguyen
Team member 2	Willytom Rigaud
Team member 3	Francis Kwami Dzikpe

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed. Note: You may be required to provide proof of your outreach to non-contributing members upon request.

Detailed data and implementation files are available in the zipped folder (GWP2).

Our group choose the Multicollinearity, Non-Stationarity and Regime change challenges. Below are the results of our study on these challenges.

Problem 1: Multicollinearity

1. Definition

Multicollinearity happens when two or more explanatory variables in a regression model are strongly correlated. This condition can be formally measured using the Variance Inflation Factor (VIF), which quantifies how much the variance of an estimated regression coefficient increases because of collinearity among the predictors. The VIF for the j-th variable is defined as:$VIF_j = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient determination of the regression of X_j on all the other explanatory variables [3, 6].

2. Description

Multicollinearity makes it difficult to distinguish the individual effects of correlated predictors on the dependent variable. In finance, macroeconomic indicators such as interest rates, inflation, and stock indices often move together, making the identification and management of multicollinearity crucial for reliable modeling [8].

3. Demonstration

Data Preparation

Monthly data from January 2015 to June 2025 was collected for Apple Inc.'s closing stock price (AAPL), the S&P500 index, Federal Funds Rate, 10-Year U.S. Treasury Yield, Consumer Price Index (CPI), and Brent Oil prices. All data were selected to monthly frequency and merged using the end-of-month values. Missing values were addressed using forward-fill methods.

Model Fitting

The baseline regression model used was:

$$AAPL_{Close_t} = \propto + \beta_1 SP500_{Close_t} + \beta_2 Fed_{Funds_{Rate_t}} + \beta_3 Treasury_{10Y_t} + \beta_4 CPI_t + \beta_5 Brent_{Oil_t} + \in_t$$

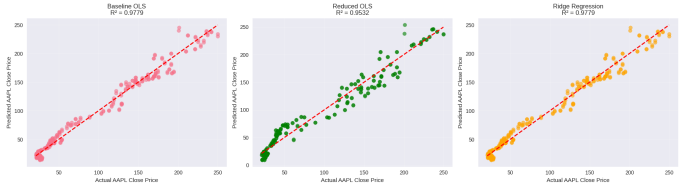
Key outputs:

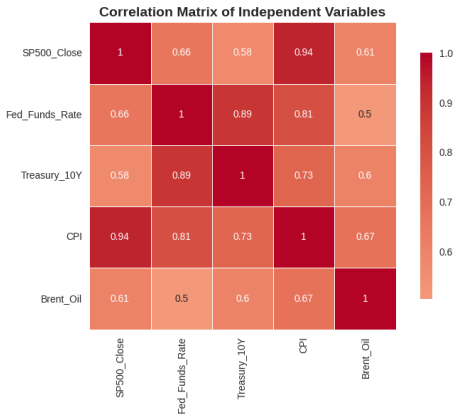
- **Baseline OLS Regression:** R-squared: 0.978; Adjusted R-squared: 0.977; AIC: 1898.33
- **VIF Analysis:** CPI (VIF = 23.3), SP500_Close (VIF = 12.95), Fed_Funds_Rate (VIF = 8.59), Treasury_10Y (VIF = 6.23), Brent_Oil (VIF = 2.32)

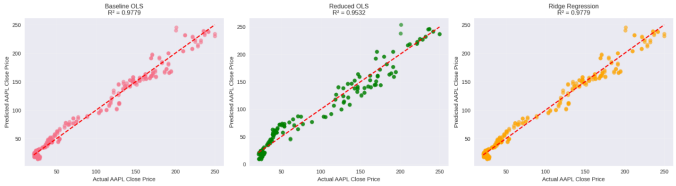
4. Diagram

Exploratory and Diagnostic Plots

- **Correlation Heatmap:** Shows strong correlations among SP500_Close, Fed_Funds_Rate, and CPI, confirming the multicollinearity issue.
- **Residual Plot:** Demonstrates the goodness-of-fit and whether there are any apparent patterns in the residuals.







5. Diagnosis

The high VIF values, especially for CPI and SP500_Close, indicate severe multicollinearity. T This makes coefficient estimates unstable, with some signs being counter-intuitive or standard errors being inflated. These are classic signs of multicollinearity problems [3, 6, 8].

6. Damage:

The model reveals the following issues due to multicollinearity:

- **Inflated standard errors**, making it hard to assess statistical significance.
- **Unstable coefficient estimates**, sensitive to small changes in the data.
- **Complicated interpretation**, as variables compete to explain similar variance..
- **Risk of misleading inference**, potentially causing wrong economic decisions [3, 4, 8].

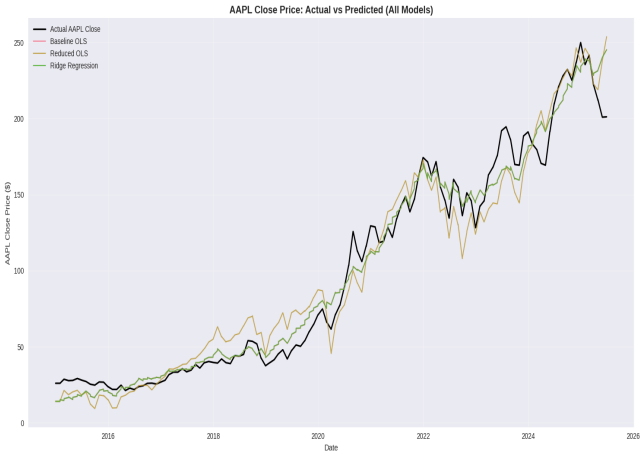
7. Directions: To reduce multicollinearity, we implemented:

- **Removed the variable with the highest VIF (CPI)** and refitted the model: The new model had slightly lower R-squared (0.953), but coefficient estimates became more interpretable and stable.
- **Applied Ridge Regression**, which penalizes large coefficients and effectively handles multicollinearity: Ridge Regression maintained high predictive accuracy (R-squared = 0.978) while reducing coefficient inflation [3, 6].

Future improvements may involve focusing on recent data, applying transformations, or using models like Principal Component Regression to further mitigate collinearity.

8. Deployment

In practice, this approach helps improve investment decisions and risk management for AAPL stock. By handling multicollinearity, the model gives more stable, interpretable results, which are essential for portfolio management and forecasting. Regularization methods like Ridge Regression are particularly helpful when dealing with highly correlated financial indicators [5,6,7,8].



Problem 2: Non - Stationarity

1. Definition

A time-series is considered weakly stationary if its mean, and variance are independent of time, and its autocovariance depends only on the lag between two observations and not the time at which the observation happens. This can be written as
$\forall t, t \in [1, T]$

$$E[y_t] = \mu$$
$$V[y_t] = \sigma^2$$
$$cov(y_t, y_s) = cov(y_{t+h}, y_{s+h}), \forall h$$

2. Description

Lots of financial time series are non-stationary, such as stocks or interest rates. They become interesting if their first difference is stationary however. We call such a time-series I(1), for first order integration. I(1) time-series open the door to many important models for stock analysis. One of the most important trading strategy, statistical arbitrage, is directly based on cointegration. If two I(1) time series can create a linear combination that is I(0), we say that the series are cointegrated. From there on, we can use Vector Error Correction to model the underlying long and short term equilibrium of the two series, and use probabilistic principles to decide when one series is drifting far enough from the equilibrium to make money by betting on it returning to the mean.

3. Demonstration

We used the time series of Visa (V) and Mastercard (MA) stocks. Given that they are stocks, they have days with no observations (week-ends and national holidays). We therefore drop the Not A Number values to get a dataset without missing values. We start by inspecting the plots of the original time-series and the differentiated time series. They indeed look I(1), so we move on to running the Augmented Dickey Fuller test on both for a more thorough diagnostic. The p-value for V and MA original time-series is 0.92 and 0.89, indicating that the series have a unit root and are therefore non-stationary. The first-difference of both series is less than 0.01 so we reject the null of the presence of a unit root : V and MA are both I(1). To see if it makes sense to check cointegration, we run OLS on Visa with Mastercard as the explanatory variable, and check the residuals. If the residuals are I(0), then the series should theoretically be correlated :

$$y_t = \alpha + \beta x_t + \varepsilon \qquad (1)$$

With $y_t, x_t \sim$ I(1) and $\varepsilon \sim$ I(0), we can rewrite (1) as

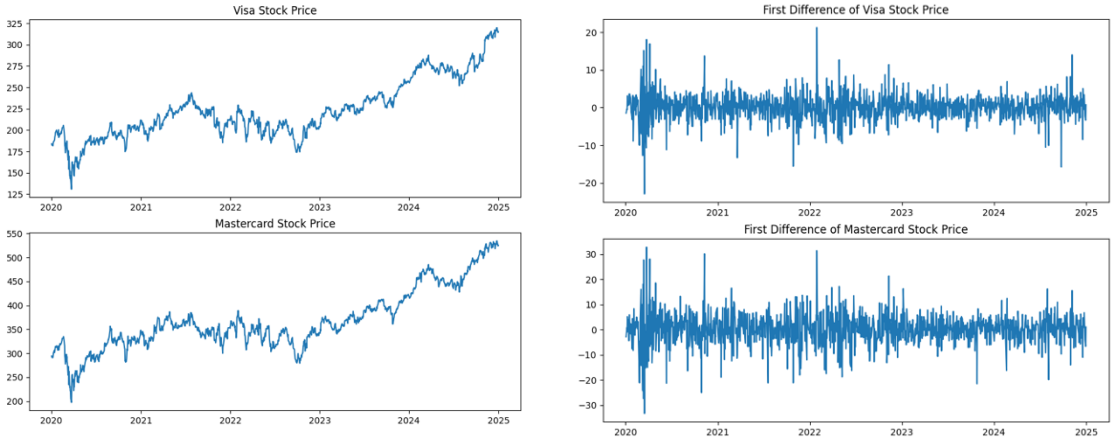
$$\varepsilon = y_t - \alpha - \beta x_t$$

And we therefore have an I(0) time-series created from a linear combination of I(1) variables.

The OLS residuals turn out to be 0 from the ADF test, so we now use a VAR model, and the Johansen cointegration test to tune the hyperparameters of the VEC model :
We therefore start by modeling a Vector Autoregressive Model to find the optimal lag order. The VAR order summary shows that for lag 3, AIC, FPE and HQIC are maximized, but BIC is greatest for a lag of 1. Given that we are trying to model the true relationship, the BIC is a more adequate criterion, and we therefore choose lag 1.
The Johansen cointegration test rejects H0 : rank ≤ 0 at 95% confidence level, but fails to reject rank ≤ 1, for any confidence level. We can thus infer that the rank of our cointegration relationship is 1. We finally move on to modeling the Vector Error Correction model with lag = 1, and cointegration rank = 1. In the end, our VECM gives the following linear combination of V and MA to get an I(0) series :

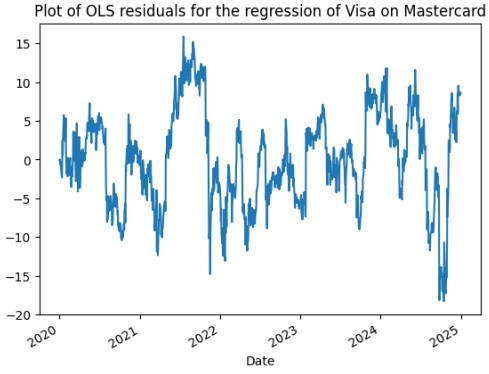
$$S = -22.3724 + Visa - 0.5453 \times Mastercard$$

4. Diagram/Diagnostic



These plots show the original and differentiated time-series of Visa and Mastercard. We see on the left graphs that the series do not look stationary, but that changes on the right graphs.

This plot shows the hunch for cointegrating the series : if the residuals of a OLS regression are stationary, the series are cointegrated. By running OLS and observing the residuals, we can guess whether the cointegration will work.



6. Damage

The OLS regression shows positive autocorrelation, which is also visible in the last graph of the previous question. To get rid of it, we tried modeling the regression as a Generalized Least Squares, while modeling the error term as an AR(1) process. This approach is called GLSAR and is available in the statsmodels linear model package in python. However, while it got rid of the autocorrelation problem, it made the error

not normally distributed anymore. We have decided to keep the original OLS regression because the end results do not change significantly, and they are more interpretable.

7. Directions

The next steps would be to address the violations of the OLS hypotheses by trying to fix the autocorrelation without leading to the violation of any other hypotheses.

8. Deployment

The deployment phase would consist in making the model live by calculating the z-score from the deviations of the securities from their equilibrium, and trading on it when we're above a pre-defined threshold. Basically, we would turn this model into a full statistical arbitrage strategy.

Problem 3: Regime Change

A regime change or regime switching occurs when the underlying data generation process of a time series changes between different states. Such changes may occur due to macroeconomic and financial policy changes, business cycles, global pandemics, wars etc. Structural breaks is mostly used to refer to such abrupt changes in time series. As identified by Ang and Timmermann (2012), regime switching models differ in their means, volatilities, autocorrelations and cross-covariances across regimes hence affecting the dynamic properties of asset prices, forecasting and confidence interval estimations. For these reasons, researchers and analyst pay special attention to regime changes in time series. Different statistical approaches exist to detect regime change. Graphical visualizations is usually a first step. Formal statistical approaches may include the Chow test, CUSUM test, CUSUMQ test and Markov Switching models. Once detected, we can split the sample into pre and post regime change, apply regime switching models or threhold regression models to solve the issue of regime change.

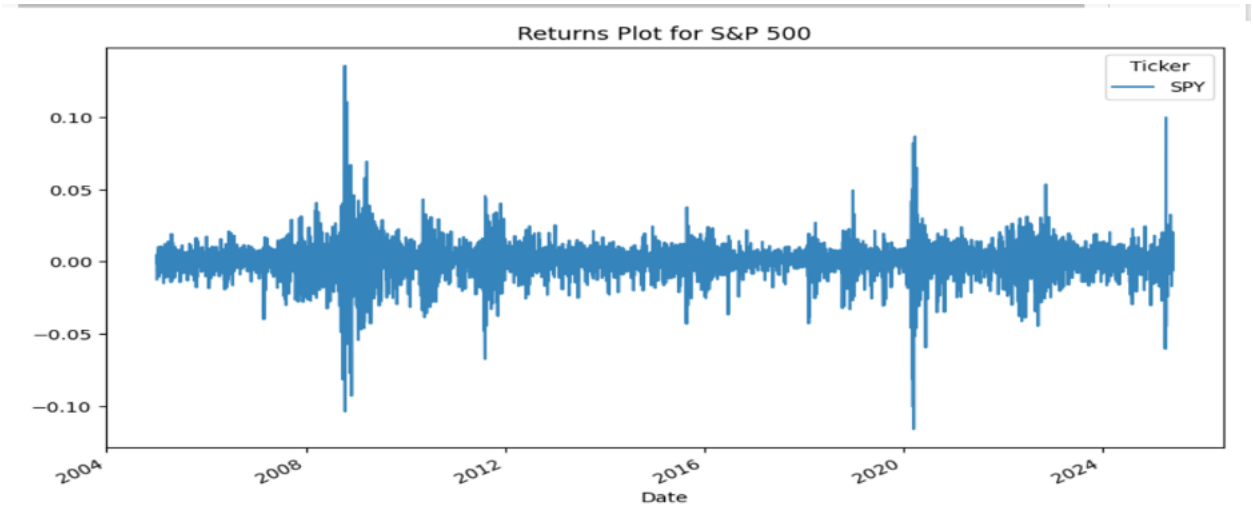
For this exercise, we tested the returns data for several financial time series based our example on the SPY index. We obtain the adjusted close data for the period 01/01/2005 to 06/01/2025 from Yahoo Finance. We clean the data and derive the log returns for the series. We use graps, a Chow test and Markov switching models to detect the presence of regime changes.

For the Markov switching model, if we assume an AR(1) model, we model the data as

$$y_t = c_{st} + \omega\, y_{t-1} + e_t$$

Where S_t is a random variable denoting the different states.

Figure 1 below shows the plot of S&P 500. The figure shows significant spikes in 2009, 2020 and 2025 indicating a possible presence of structural break.



To verify the structural breaks, we test the series using the Chow test with a null hypothesis of no structural breaks. We test structural break on 2008/03/07(financial crises) and 2020/03/01(Covid) Results show the Chow test is significant for the dates confirming the presence of structural breaks or regime change.

Further, we applied a two-state Markov switching model for a high volatility and low volatility states. Results are in the table below. We observe that, the variance of error term (sigma2) for regime 0 is 0.0004 yielding a standard deviation/volatility of 0.02067 while regime 1 has a variance of 0.00045 and standard deviation of 0.0067 implying regime 1 is more stable than region 0. The regime transitions also shows that the probability of staying in regime 0 is 0.9556 while a transition to regime 1 is 0.044. the probability of staying in region 1 is 0.0158 while the transition from 1 to 0 has a probability of 0.9841. put together we find a significant presence of structural breaks. We also find the expected during of region 0 to be about 22 days as against 63 days for region 0. That is even though there is a high persistence and stability of region 0, the system tends to linger in region 1 more. The smooth probabilities in low and high volatility alongside the structural breaks are stated in figure

Markov Switching Model Results

Dep. Variable:	SPY	No. Observations:	5134
Model:	MarkovRegression	Log Likelihood	16601.946
Date:	Mon, 30 Jun 2025	AIC	−33191.891
Time:	01:33:27	BIC	−33152.629
Sample:	0	HQIC	−33178.149
	− 5134		
Covariance Type:	approx		

Regime 0 parameters

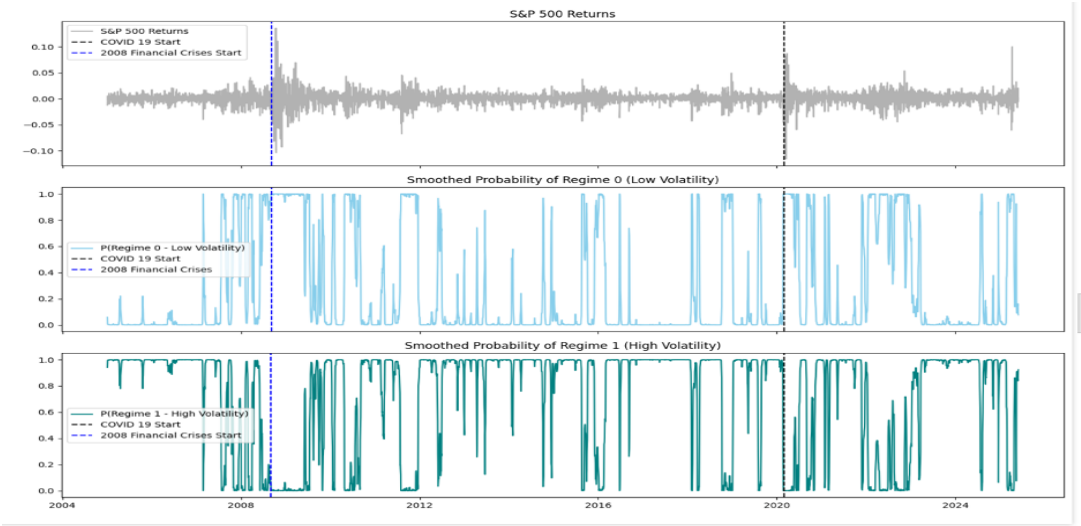
	coef	std err	z	P>|z|	[0.025	0.975]
const	−0.0013	0.001	−2.236	0.025	−0.002	−0.000
sigma2	0.0004	2.21e−05	19.381	0.000	0.000	0.000

Regime 1 parameters

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0010	0.000	8.409	0.000	0.001	0.001
sigma2	4.509e−05	1.74e−06	25.890	0.000	4.17e−05	4.85e−05

Regime transition parameters

	coef	std err	z	P>|z|	[0.025	0.975]
p[0−>0]	0.9556	0.008	116.477	0.000	0.940	0.972
p[1−>0]	0.0158	0.003	5.276	0.000	0.010	0.022


S&P 500 Returns
S&P 500 Returns
COVID 19 Start
2008 Financial Crises Start
Smoothed Probability of Regime 0 (Low Volatility)
P(Regime 0 - Low Volatility)
COVID 19 Start
2008 Financial Crises
Smoothed Probability of Regime 1 (High Volatility)
P(Regime 1 - High Volatility)
COVID 19 Start
2008 Financial Crises Start
2004
2008
2012
2016
2020
2024

References

1. Ang, A., & Timmermann, A. (2012). Regime changes and financial markets. Annu. Rev. Financ. Econ., 4(1), 313-337
2. Tan, Z., & Wu, Y. (2025). On Regime Switching Models. Mathematics, 13(7), 1128
3. James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.
4. Kutner, Michael H., et al. *Applied Linear Statistical Models*. 5th ed., McGraw-Hill, 2005.
5. Montgomery, Douglas C., et al. *Introduction to Linear Regression Analysis*. 5th ed., Wiley, 2012.
6. Wooldridge, Jeffrey M. *Introductory Econometrics: A Modern Approach*. 7th ed., Cengage, 2020.
7. "Federal Reserve Economic Data (FRED)." Federal Reserve Bank of St. Louis, https://fred.stlouisfed.org/.
8. "Yahoo Finance." https://finance.yahoo.com/.