STATISTICAL METHODS FOR DATA SCIENCE

STOCK MARKET AND MACROECONOMIC INDICATORS

Student: Iliadis Nikolaos-Anastasios

Personal Number: 199508237311

Contents Table

Step 1: Identify datasets

As a child growing up in Greece, when I was 4 years old a stock market crash happened. Even though I was really young I still have some faded memories of this happening. I remember plenty of people being frustrated, because the prime minister had convinced them to invest in the Stock Market of Athens. Since then, I started being curious about the Stock Market and how it works.

During my bachelor's and first Master's degree in Financial Engineering, I acquired some basic knowledge about stocks. I took courses for portfolio and risk management, but they were too theoretical and I never had the opportunity to get hands-on knowledge and experiment. For this reason, I decided to explore how much some macroeconomic indicators affect the stock prices.

To carry out this I started looking for datasets. Five datasets were used for this project. Three of them are from yahooFinance that were downloaded through Python with the yf.download command. Those are the stock data that were used and they are the NASDAQ100(NDX), NASDAQTech(NDXT), and the volatility index(VIX).

I then looked for macroeconomic indicator's datasets. Since many of the companies in NASDAQ are American, datasets with American indicators were used. The unemployment rates and the inflation rates. Both unemployment and inflation are taking a huge part in our lives, and especially recently that prices are skyrocketing and finding a job is harder than ever.

The datasets used to perform the analysis, were withdrawn from the following websites:

Unemployment rates for America: https://fred.stlouisfed.org/series/UNRATE

Inflation rates for America: https://fred.stlouisfed.org/series/T10YIE

NASDAQ100: https://finance.yahoo.com/quote/%5ENDX/

NASDAQTech: https://finance.yahoo.com/quote/%5ENDXT/

Volatility Index: https://finance.yahoo.com/quote/%5EVIX/ .

The volatility index, commonly known as the VIX, is a real-time market index that represents the market's expectations for volatility over the coming 30 days. It is often referred to as the "Fear Index" because it tends to rise when the stock market experiences significant declines, reflecting increased investor anxiety. [1]

But what is volatility? Volatility is a statistical measure of the dispersion of returns for a given security or market index. It represents how much the price of an asset fluctuates over a certain period of time.In other words is the standard deviation of the close price of a stock.[2]

All of the data were from 2010-01-01 to 2024-12-31.

Multiple datasets were cross-referenced to provide a more comprehensive and insightful analysis of the factors affecting stock market dynamics. By combining NASDAQ-100 (NDX) and NASDAQ

Technology (NDXT) index data, I analyzed broader market trends and sector-specific patterns, offering richer insights than relying on a single dataset. Adding macroeconomic indicators like the Unemployment Rate (UNRATE) and Inflation Expectation (T10YIE) allowed to explore the impact of economic fundamentals on stock market returns.

Step2: Visualize data and compute descriptive statistics

After a long data-cleaning session the boring part was over and it was time to start with the interesting part. However, there were some difficulties, which needs to be further explained. The initial challenge was that the inflation and stock data were calculated daily, whereas the unemployment data were monthly. To address this, all data were converted to a monthly format by calculating the monthly means. This adjustment ensured the data were of the same length, which facilitated merging the datasets and improved the effectiveness of statistical tests and Machine Learning models. The command “resample('M').mean()” was employed for this transformation.

Second issue, was that the data from yahooFinance were two leveled. Hence, the one level had to be removed and rename the second so no columns were lost.

Basic data cleaning was performed, which included removing NaN values, zeros, and unnecessary columns, and setting the date as the index for all datasets. The datasets were then merged by Date. Finally, the returns for each of the Close Prices of the stock data were calculated and added to the merged dataset. Daily returns represent the percentage change in the price of an asset from one day to the next. They are calculated using the formula[3]:

$$DailyReturns = \frac{Ptoday - Pyersteday}{Pyersteday} \cdot 100\%$$

Figure 1:Daily Returns Formula

Where P is the price of the stock.

However, because the Stock data that were used, were not daily but converted to monthly this equation is converted to monthly returns using the values of months.

Then descriptive statistics were calculated with the .describe() command. The results were the following.

2.1 Descriptives

NASDAQ Tech:

NASDAQ TECH	
Count	179
Mean	4301.06
Std	2895.10
Min	1080.59
25%	1780.64
50%	3513.32
75%	6601.05
max	10775.85

Table 1: NASDAQTECH Descriptives

Here we can notice that the difference between the mean and max values is very large. This means that the data will probably need to be standardized for the Machine learning modes.

The Mean is 4301.07. The Standard Deviation, or spread of the closing prices is 2895,1 which indicates a significant variability in the closing prices. This suggests that the index experienced substantial fluctuations over the observed period.

50% of the values are below the mean which indicates a right skewed distribution.

NASDAQ Tech Returns:

NASDAQ TECH RETURNS	
Count	179
Mean	0.013
Std	0.047

Min	-0.0185
25%	-0.0148
50%	0.0184
75%	0.0443
max	0.1242

Table 2: NASDAQTech Returns Descriptives

On average, the monthly return is approximately 1.39%. This suggests a positive overall performance over the period analyzed.

The standard deviation is about 4.75%, indicating the variability of the returns. A higher standard deviation implies more volatility.

The lowest monthly return is approximately -18.57%, showing the worst performance in a single month.

The 25% of the monthly returns are below -1.48%, indicating that a quarter of the time, the returns were negative.

The median return is approximately 1.84%, which is higher than the mean. This suggests that the distribution of returns is slightly skewed to the left (more frequent small positive returns).

The positive mean and median indicate overall growth in the NASDAQ100 Tech. The standard deviation shows that there is considerable volatility.

NASDAQ 100:

NASDAQ 100	
Count	179
Mean	7627.39
Std	5230.94
Min	1784.73
25%	3345.69
50%	5816.11
75%	11867.14

max	21517.41

Table 3: NASDAQ 100 Descriptives

The NASDAQ 100 descriptives are pretty similar to the NASDAQTech. However, it is noticeable that the range of min and max values is much larger, with the max value being 21517.41. Again the median is smaller than the mean and 75% of the data are smaller than 11867.14.

NASDAQ100 Returns:

NASDAQ 100 RETURNS	
Count	179
Mean	0.014
Std	0.039
Min	-0.157
25%	-0.0073
50%	0.019
75%	0.040
max	0.092

Table 4: NASDAQ 100 Returns Descriptives

The mean return is 0.014(1.4%). THis indicates that on average the NASDAQ100 generated positive returns.

The standard deviation is 0.039(3.95%). This illustrates the level of volatility in the index returns .Higher standard deviation suggests more variability in returns. Here the min and max indicate that there were fluctuations, however the period of the data is very extensive.

75% of the returns were below 0.04 and the difference between the median (0.01976) and the mean (0.0146) suggests potential left-skewness, implying a few significant negative returns.

The relatively small standard deviation compared to the extremes min and max suggests occasional outliers.

Inflation Descriptives:

INFLATION	
Count	179
Mean	2.05
Std	0.344
Min	0.986
25%	1.818
50%	2.122
75%	2.303
max	2.884

Table 5: Inflation Descriptives

The average inflation rate per month is 2.05%. The inflation rate varies by 0.34% from the mean. The lowest recorded inflation rate is 0.99% and the median is 2.12%.

75% of the inflation data are below 2.3% and the maximum was 2.88%.

The average inflation rate is relatively stable around 2%, with a narrow range between the minimum and maximum values, indicating moderate inflation variability.

Unemployment Descriptives:

UNEMPLOYMENT	
Count	179
Mean	5.74
Std	2.20
Min	3.4
25%	3.9
50%	5.0
75%	7.5

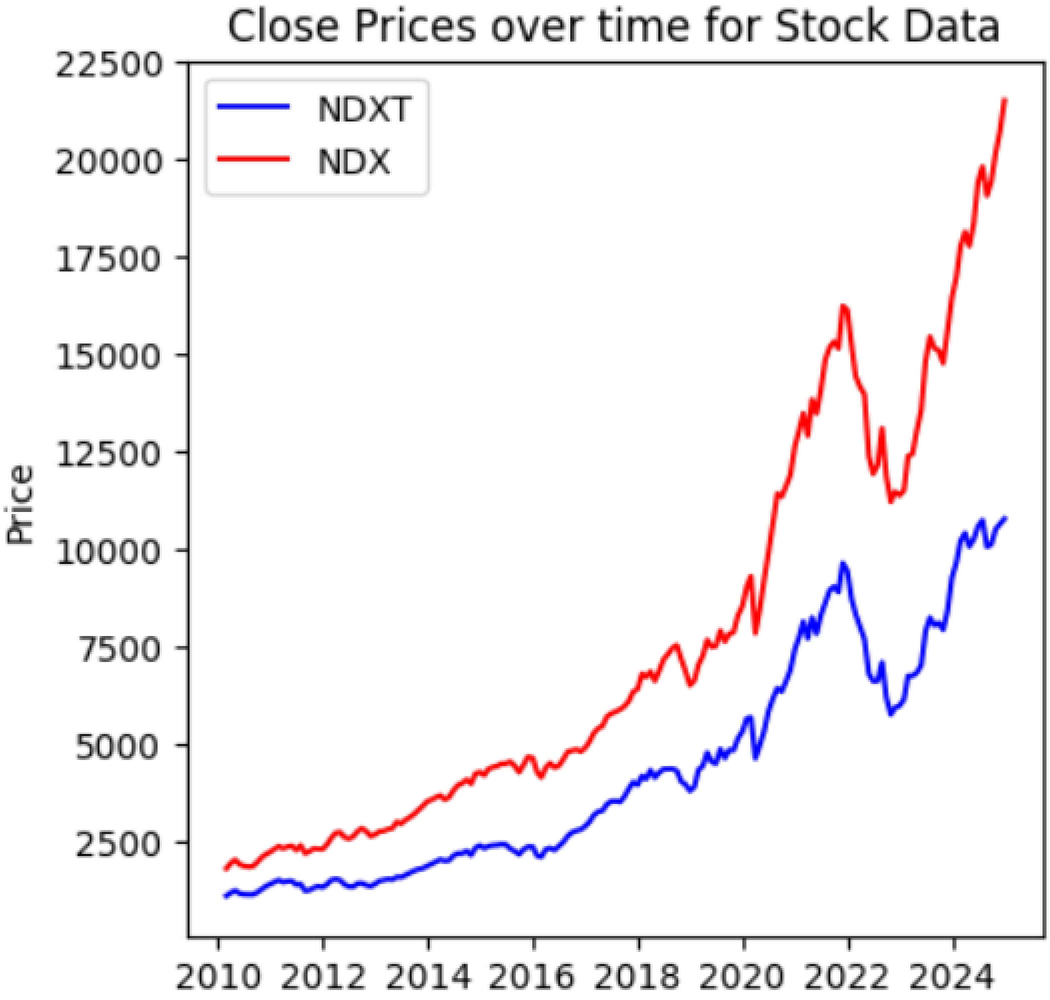
max	14.80

Table 6: Unemployment Decriptives

The average unemployment rate is higher and more variable, with a significant range between the minimum and maximum values, reflecting periods of both low and high unemployment.

2.2 Visualizations

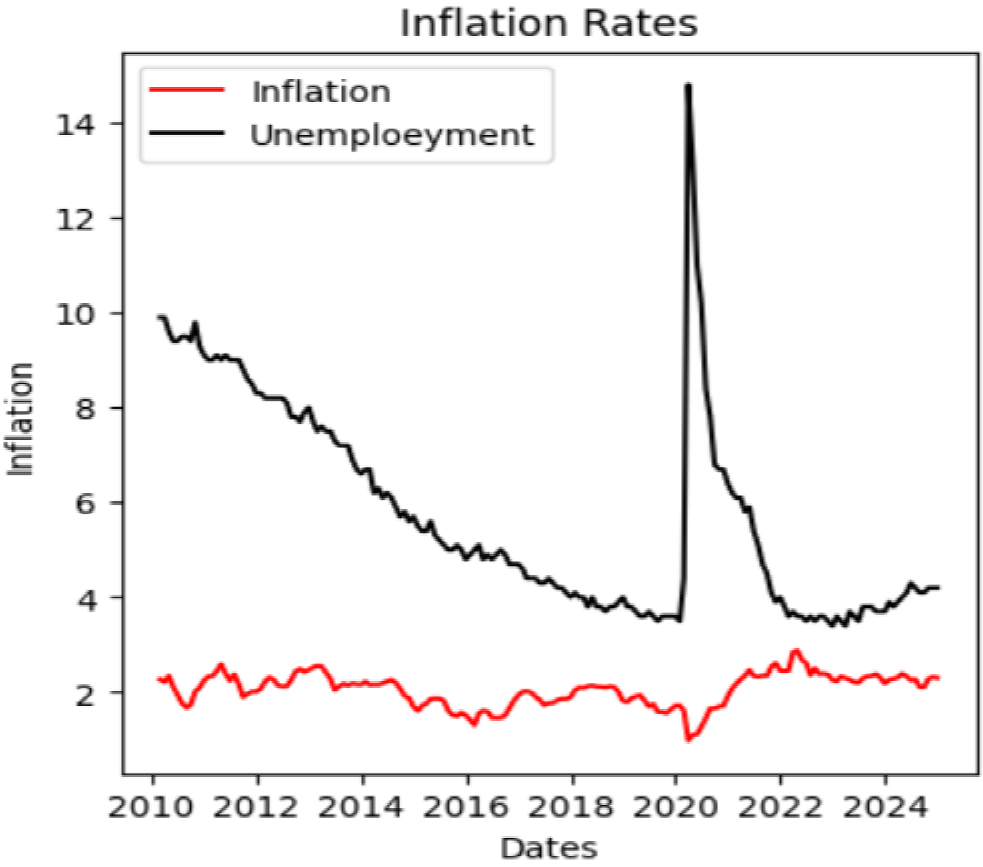
NASDAQ 100 & NASDAQ Tech



Graph 1:NDXT-NDX Close Price Line Plots

This graph depicts the close prices of the NASDAQ100(NDX) and the NASDAQTech(NDXT). Overall, the lines increase as the dates increase with some fluctuations. During 2020 the prices skyrocketed up to 17.5 thousand for NDX and 10 thousand for the NDXT. In 2022 the Close Prices plunged, and then started rapidly increasing again. This graph was selected to depict the overall trend of the Stock Prices over time and their deviations.

Inflation and Unemployment

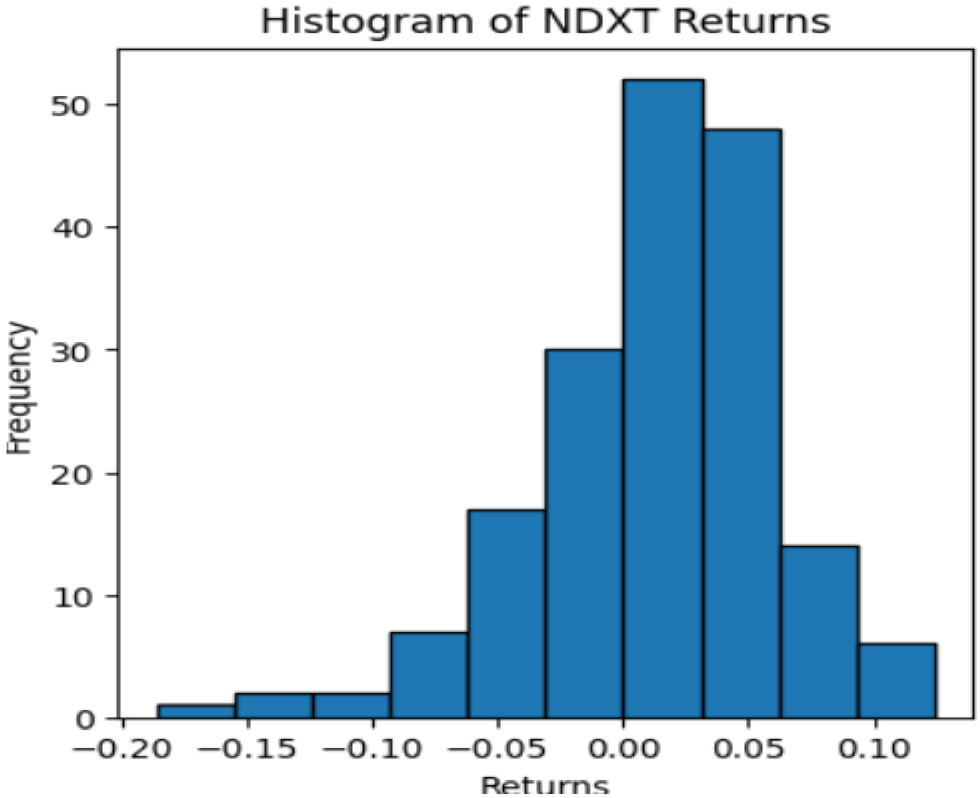


Graph 2: Inflation-Unemployment Rates Line Plots

The graph illustrates the unemployment and the inflation rates over the selected period of 2010 and 2024. The unemployment rate was declining until 2020, when the line sky-rockets from 4%, almost the lowest ever noticed in our sample, to over 14% and then it started declining until it reached 4% again 4%. This can be explained from the COVID-19 pandemic. Many people lost their jobs during the quarantine. Since then, it has been increasing very slowly.

The inflation line depicts that overall, the inflation has increased with some fluctuations. There is a significant decline in 2020, which is easily explained from the COVID-19.

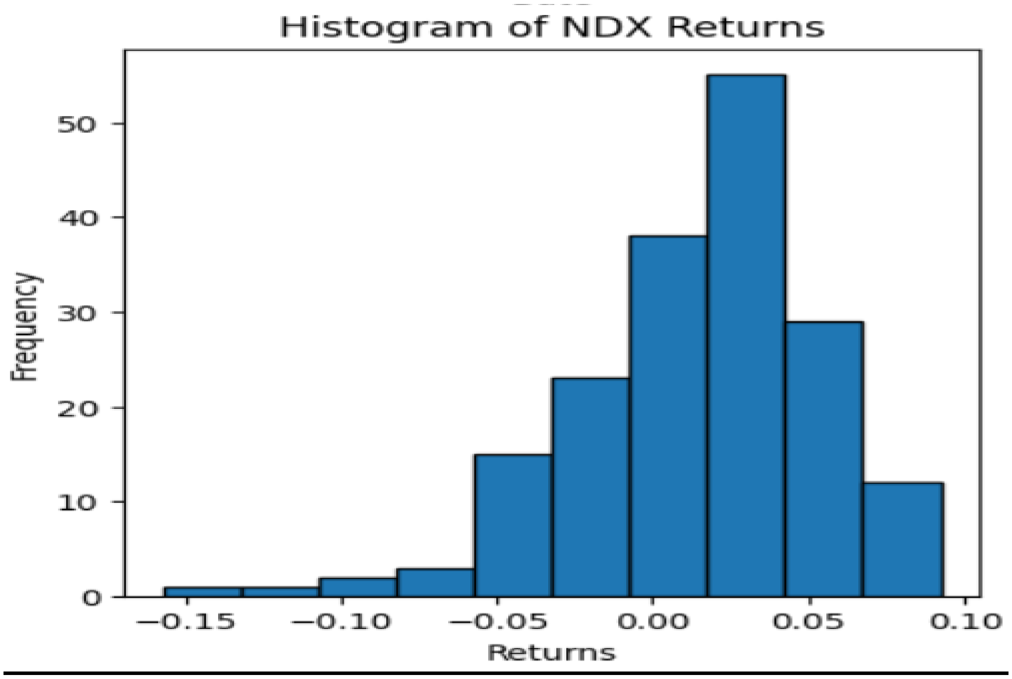
NASDAQTech Returns visual:



Graph 3:Histogram of NDXT Returns

I decided to do a histogram for the returns of stock prices. We can see that the returns in contrast with the close prices follow a Gaussian distribution. From the graph it is confirmed that the returns are left skewed as mentioned in the descriptive analysis. The histogram provides an understanding of the distribution the data follows.

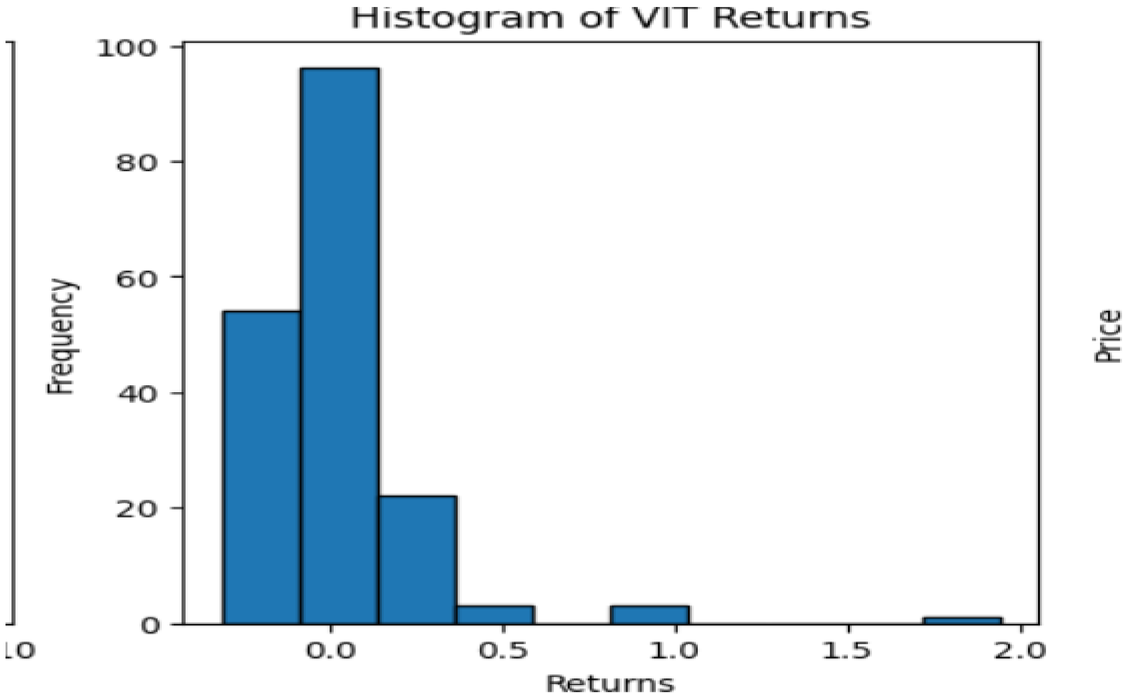
NASDAQ100 Returns visual:



Graph 4:Histogram of NDX Returns

The returns appear to be approximately symmetric and centered around zero, resembling a normal distribution. The returns appear to be approximately symmetric and centered around zero, resembling a Gaussian distribution.A significant number of observations are concentrated around small positive returns (0% to 5%). This indicates that the NASDAQ-100 generally performs with small gains more frequently than large fluctuations.The range of returns spans approximately from -15% to 10%, indicating that the index experienced varying levels of volatility during the time period analyzed.

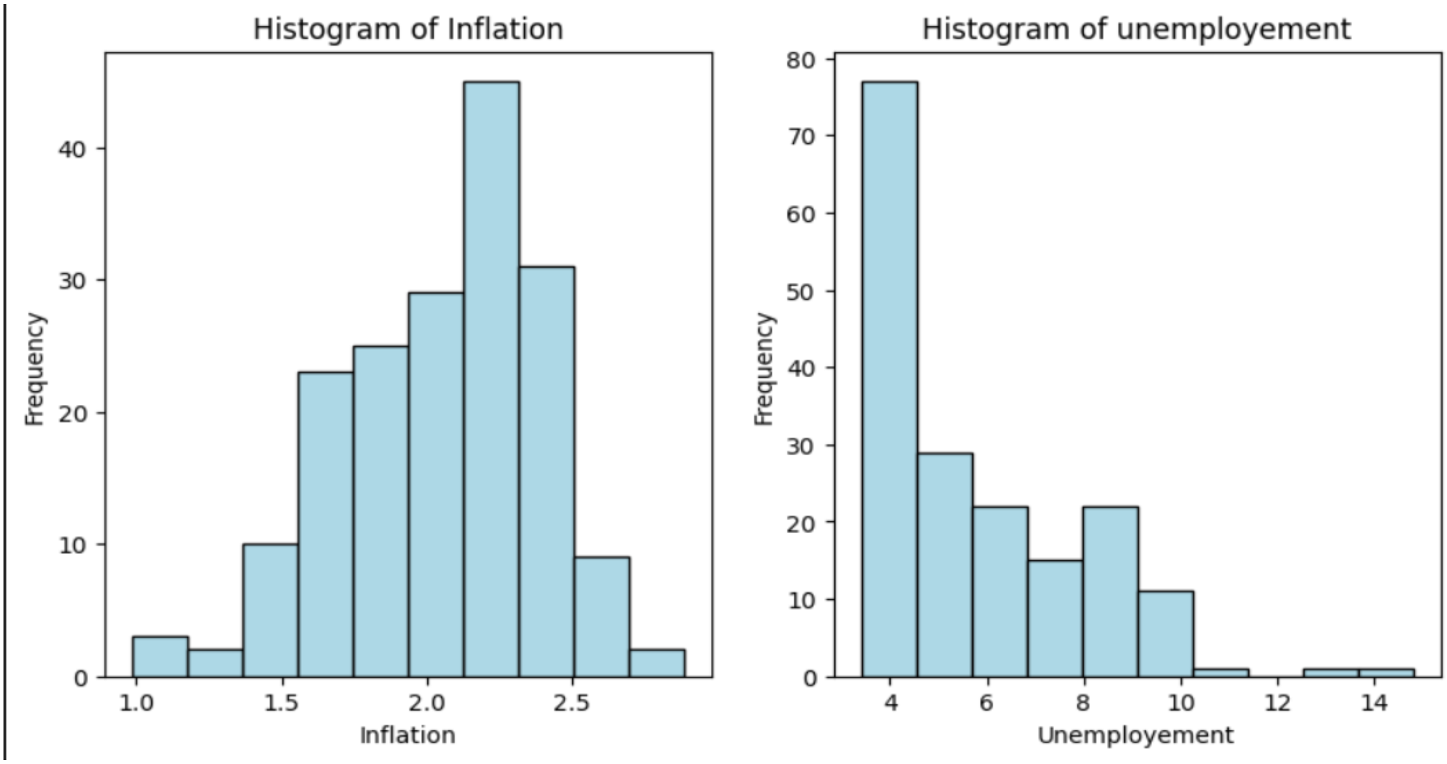
VIT Returns:



Graph 5:Histogram of VIT Returns

The histogram is highly skewed to the right. This suggests that most returns are close to 0, and larger returns are outliers. It suggests that while most of the returns are small or minimal, there are occasional significant positive returns.

Inflation & Unemployment

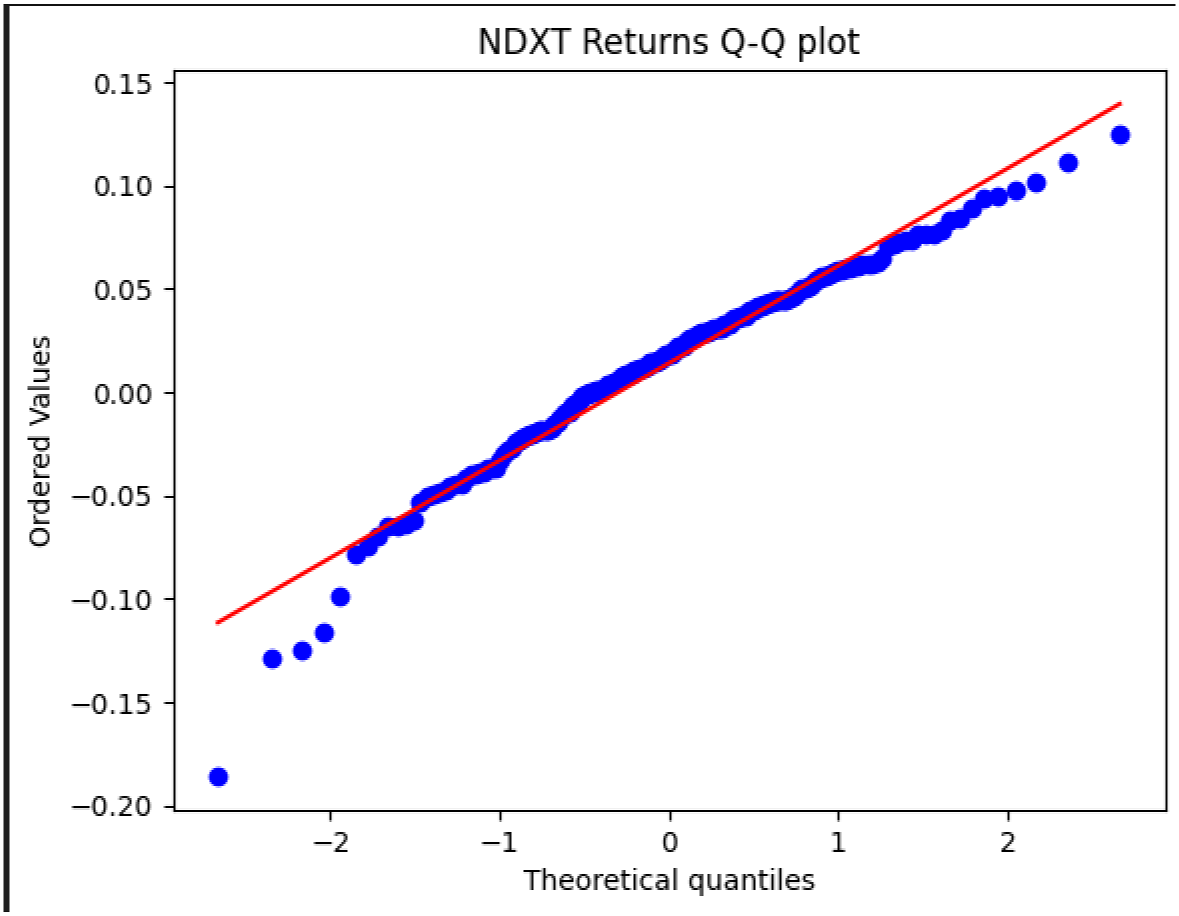


Graph 6: Histograms of Macroeconomic indicators

The histogram of inflation appears to be near-normal centered around 2. Most inflation values are concentrated between 1.5 and 2.5 with the highest frequency occurring near2. The tails of the histogram are small, indicating that extreme inflation is rare. A normal distribution in inflation suggests a stable economy, where inflation rates stay within a reasonable range. The lack of extreme outliers suggests that hyperinflation or deflation events are not a concern.

The histogram of unemployment is right skewed. The majority of unemployment rates are low, concentrated between 4% and 6%. Higher unemployment rates are rare with very few observations in the far right tail of the distribution. The skewed nature indicates that unemployment rates are generally low, but the dataset does contain outliers where unemployment spikes to higher levels up to 14%.

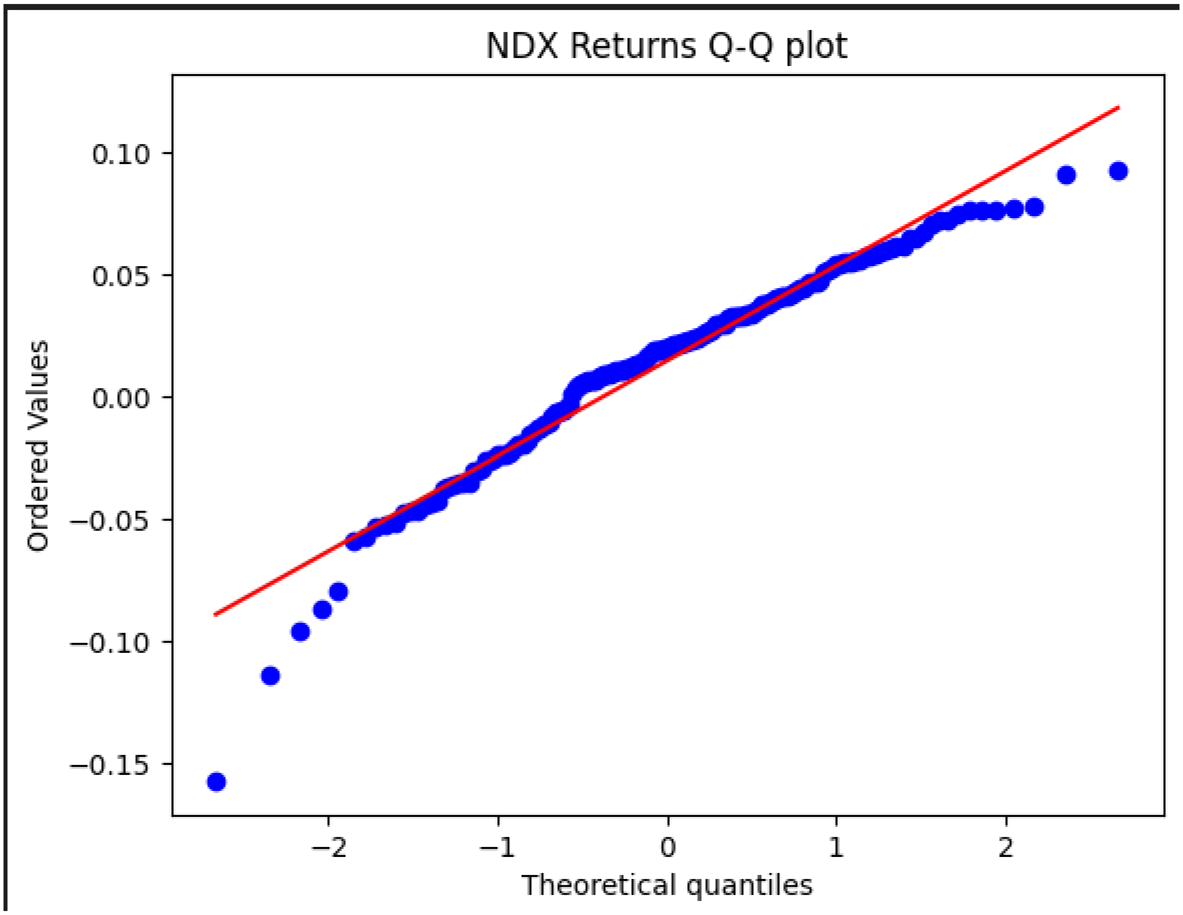
NDXT Returns Q-Q Plot



Graph 7: Q-Q plot of NDXT Returns

The Q-Q plot indicates whether the data follow a specific distribution. In this scenario the normal distribution was tested. The closer to the line the better fit normal distribution is. The NDXT returns follow a normal distribution with deviations in the tails.

NDX Returns Q-Q plot



Graph 8: Q-Q plot of NDX Returns

The NDX returns indicate that the NDX returns also follow a normal distribution with deviations in the tails.

Step 3: Formulate hypotheses

The hypotheses were selected based on my background in financial engineering and an interest in the stock market and trends. The three hypotheses collectively address how macroeconomic indicators influence tech stock indices like NASDAQ. A confidence level of 0.05 was used for all the tests in step 4.

3.1 Hypotheses 1

H1: Is there a significant difference in NASDAQ Tech (NDXT) returns between periods of high and low inflation?

Inflation influences economic conditions, which in turn can affect market sentiment and technology sector performance. Analyzing NDXT returns under different inflation regimes provides insights into the tech market's sensitivity to macroeconomic factors. The objective was to understand if inflation serves as a significant determinant for returns in the tech sector.

3.2 Hypothesis 2

H2: Does market volatility (VIX) significantly impact NASDAQ 100 returns during high-volatility vs low-volatility periods?

Market volatility is a known risk indicator, and its relationship with returns can highlight behavioral trends among investors during uncertain periods. Examining how volatility impacts Returns across different regimes could reveal strategies for portfolio risk management. The objective is to evaluate whether the NASDAQ 100 reacts more to high volatility periods.

3.3 Hypothesis 3

H3: Is the NASDAQ100 more likely to increase after periods of low unemployment than after periods of high unemployment?

Employment levels reflect the health of the economy, influencing investor confidence. Lower unemployment rates may be indicative of stronger economic growth, leading to increased market activity. The objective is to assess whether NASDAQ100 returns tend to be higher following periods of unemployment.

Step 4: Validate or invalidate your hypotheses with statistical analysis

4.1 Hypothesis 1

H1: Is there a significant difference in NASDAQ Tech (NDXT) returns between periods of high and low inflation?

- H_0: The mean NDXT returns are the same during high and low inflation.

$$\mu_{High} = \mu_{low}$$

- H_A: The mean NDXT returns differ during high and low inflation

$$\mu_{High} \neq \mu_{low}$$

For this hypothesis testing, a **two-sample t-test was used**. The reason for that is a two-sample t-test is used to test if two alternative options have different effects by testing if the means differ by a constant.

The Data for this Hypothesis: high inflation NASDAQ TECH returns and low inflation NASDAQ TECH returns.

Random Variable and Assumptions: We assume that the data are independent between them. Additionally, that the data are independent and identically distributed.

According to the Q-Q plot (Graph 7) the data follow a Normal distribution, so they satisfy this requirement.

As an indicator for high or low volatility, I used the median. Everything below the median were considered as low inflation and everything over the median was considered as high. I chose the median instead of the mean because I feel that the most common value is a better indicator for separating the sample instead of the mean.

Results

From the two-sample t-test the results where the following:

- T-statistic: -1.10
- P-Value: 0.27

The P-Value is greater than a (0.05). We fail to reject the null hypothesis and there is no statically significant difference in NASDAQTech returns between periods of high and low inflation.

4.2 Hypothesis 2

H2: Does market volatility (VIX) significantly impact NASDAQ 100 returns during high-volatility vs low-volatility periods?

- H_o: There is no significant difference in NASDAQ 100 returns between high-volatility and low-volatility periods as measured by the VIX

$$\mu_{HighVol} = \mu_{LowVol}$$

- H_A: There is a significant difference in NASDAQ 100 returns between high-volatility and low-volatility periods as measured by the VIX

$$\mu_{HighVot} \neq \mu_{LowVot}$$

For this Hypothesis test I used a **paired t-test** because I wanted to test if two alternative options have different effects by testing if the mean of their differences differs from a predefined constant[4].

Data: VIX returns during high volatility and VIX returns during low volatility

Random variable and assumption: We assume that the data are independent and identically distributed.

According to the Graph 8 (Q-Q plot of NDX) the data follow a Gaussian distribution.

In this test, the median was again used to split the data into high and low volatility. An issue arose after converting the data to monthly, resulting in 179 rows, which prevented the test from running. To resolve this, one value was randomly dropped from the sample with more observations (the low volatility returns).

Results

From the paired t-test the results were the following:

- T-Statistic: -3.48
- P-Value: 0.00076

The P-Value is much smaller than 0.05. From this we reject the null-hypothesis.

The t-statistic of -3.4855 suggests a significant difference between the two periods. The negative sign indicates NDX returns are lower during high-volatility periods compared to low-volatility periods.

Given the very low p-value, we reject the null hypothesis (H_0). This means there is significant evidence to suggest that market volatility (VIX) significantly impacts Nasdaq-100 (NDX) returns during high-volatility versus low-volatility periods.

4.3 Hypothesis 3

H3: Is the NASDAQ100 more likely to increase after periods of low unemployment than after periods of high unemployment?

- H0: The probability of the Nasdaq-100 increasing after low unemployment periods is 50%

$$H_0: p=0.5$$

- The probability of the NASDAQ-100 increasing after low unemployment periods is greater than 50%

$$H_A: p>0.5$$

A binomial test was conducted for this hypothesis to determine if the proportion of "success" differs from a predefined constant, which in this case is 0.5. To perform the binomial test, the NDX returns were first converted to binary values (1 for positive and 0 for negative). Subsequently, low unemployment rates were selected by identifying those below the median.

Results

The results of the binomial test were the following:

- Number of Positive Movement (Successes, k):64
- Total Number of Trials (n): 94
- P-Value: 0.00029
- T-Statistic: 0.68
- Alternative Hypothesis: The observed proportion of positive movements is greater than what would be expected under the null hypothesis.

The t-statistic = 0.68 means that 68% of the movements are positive.

The p-value- 0.00029, is very small and much smaller than 0.05. This indicates that the likelihood of observing a proportion of positive movements as extreme as or more extreme than 0.68085, under the assumption of the null hypothesis, is very low.

Since the p-value is so small, we reject the null hypothesis. This suggests there is strong evidence that the proportion of positive movements is significantly greater than what we would expect under the null hypothesis which is 50%.

Hence, we reject the null hypothesis (H_0): The proportion of positive movements is significantly greater than expected.

Step 5: Extend with machine learning predictions

5.1 TASK: Predicting NASDAQ-100 Returns using macroeconomic indicators.

Predicting stock index returns is highly relevant in finance, aiding in portfolio management, risk assessment, and trading strategies. As humans we always try to predict the future and minimize the uncertainty.

To address the task of predicting NASDAQ-100 returns, three regression models were used, Linear Regression, Random Forest, and Support Vector Regressor. Regression is a supervised machine learning technique which is used to predict continuous values[5].

The models were evaluated using Mean Square Error (MSE) and the r2 score.

Mean Square Error (MSE) is a crucial metric for evaluating the performance of predictive models. It measures the average squared difference between the predicted and the actual target values within a dataset. The primary objective of the MSE is to assess the quality of a model's predictions by measuring how closely they align with the ground truth[7]. A low MSE indicates that the model's predictions are closer to the true values, which suggest better performance.

The r2 score is used to measure how well a model fits data, and how well it can predict future outcomes. Simply put, it tells you how much of the variation in your data can be explained by your model. The closer the R-squared value is to one, the better your model fits the data[8].

To carry out the machine learning models, the training data included the standardized close prices of the three stock features, unemployment rates, and inflation. The Close prices were standardized due to their large and varying ranges. The target variable was the NDX returns (NASDAQ 100 returns). The dataset was split into 75% training and 25% validation sets using the sklearn function train-test-split, with a random seed of 101.

To cross validate the models 10-Fold was used.

5.2 Research Questions

RQ1: Can we predict the NDX Returns based on the explanatory variables?

RQ2: Which regression model performs better in predicting NDX Returns as measured by R^2 and Mean Squared Error

RQ3: Which features contribute the most to predicting NDX Returns.

5.3 Models

1. **Linear Regression:**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting[6].

2. **Random Forest:**

Random Forest is a supervised machine learning algorithm that excels in classification, regression, and other tasks by utilizing decision trees. It is particularly effective for managing large and complex datasets, handling high-dimensional feature spaces, and providing insights into feature importance. The algorithm's ability to achieve high predictive accuracy while minimizing overfitting makes it a popular choice in various fields, such as finance, healthcare, and image analysis[9].

3. **Support Vector Regression**

Support vector regression (SVR) is a type of support vector machine (SVM) that is used for regression tasks. It tries to find a function that best predicts the continuous output value for a given input value[10].

5.4 Model Results

Model	MSE	R2	Kfold Mean	Kfold Std
Linear Regression	0.00094	0.25	0.0012	0.0005
Random Forest	0.001	0.17	0.0014	0.0009
Support Vector Regre.	0.0016	-0.28	0.0022	0.0006

Linear Regression has the lowest MSE, indicating that its predictions are closest to the actual values on average. This suggests that the model is quite accurate in its predictions. The R² value of 0.25 means that the model explains 25% of the variance in the data. While this is not very high, it indicates that the model captures some of the underlying patterns but leaves a significant portion unexplained.

This could be due to the simplicity of the model or the presence of non-linear relationships in the data that Linear Regression cannot capture.

Random Forest has a slightly higher MSE than Linear Regression, meaning its predictions are a bit less accurate on average. However, the difference is not substantial.The R^2 value of 0.17 indicates that the model explains 17% of the variance in the data, which is lower than Linear Regression.

SVR has the highest MSE, indicating its predictions are the least accurate on average. This suggests that the model is not performing well in terms of prediction. The negative R^2 value (-0.28) suggests that the model does not perform well.

The best performing model is the Linear Regression which has the lowest MSE and the highest R2. Even though all the models have MSE very close to zero, the R2 of the models are not close to 1. This indicates that the models are most likely facing underfitting. This can be solved by creating new features or transforming existing ones.

5.4 Statistical Tests

Paired t-tests were performed to compare the predictions of the three models. A 10-fold cross-validation with negative mean squared error as the scoring metric was used. The mean of the 10-fold cross-validation for each model was stored and converted back to positive. Additionally, the standard deviation was calculated to understand how much the MSE varies across the different folds.

5.5 Statistical Test Results

Paired-Test	t-statistic	P-Value
Linear vs RandomForest	-1.3037	0.2247
Linear Vs SVR	-4.4190	0.0017*
RandomForest vs SVR	-2.64	0.0265*

Table 7: Paired Tests for Machine Learning Models

*Statistical Significance

Linear Regression vs Random Forest

H_0: There is no significant difference between the predictions of the Linear Regression model and the Random Forest model

The p-values is greater than 0.05, indicating that there is no statistically significant difference between the predictions of the Linear Regression and Random Forest models. This suggest that, on average the predictions from these two models are not vary different from each other and we fail to reject the H0.

Linear Regression vs SVR

H_0: There is no significant difference between the predictions of the Linear Regression model and the Support Vector Regression (SVR) model

The p-value is less than 0.05, indicating a statistically significant difference between the predictions of the Linear Regression and SVR models. This means that the predictions from these two models are significantly different from each other. Hence, we reject the H0.

Random Forest vs SVR

H_0: There is no significant difference between the predictions of the Random Forest model and the Support Vector Regression (SVR) model.

The p-value is less than 0.05, indicating a statistically significant difference between the predictions of the Random Forest and SVR models. This means that the predictions from these two models are significantly different from each other. Hence, we reject the null hypothesis.

Step 6: Final conclusion and potential pitfalls

6.1 Conclusion

This project explored how macroeconomic indicators such as inflation, unemployment, and market volatility affect stock returns, with a specific focus on the NASDAQ-100 and NASDAQ Technology. The analysis revealed several critical insights:

1. Macroeconomic Indicatiors and Stock Returns:
   a. The statistical analysis demonstrated that market volatility significantly impacts NASDAQ-100 returns, particularly during high-volatility periods. This aligns with the understanding that uncertainty in the market often results in lower investor confidence and reduced returns.
   b. Inflation and unemployment showed less direct and significant impacts on stock returns. While the binomial test for unemployment demonstrated a strong correlation with positive returns, inflation's effect on NASDAQ Tech was statistically insignificant.
2. Machine Learning models
   a. Linear Regression outperformed Random Forest and Support Vector Regression in predicting NASDAQ-100 returns, as evidenced by its lower MSE and higher R^2. However, the relatively low R^2 values for all models suggest that the features used may not fully capture the complexity of stock market behavior.
   b. The standardization of the Close Prices helped improve model performance.
3. Integration
   a. Cross-referencing datasets provided a comprehensive perspective, combining market indices with macroeconomic indicators.

Potential Pitfalls

1. Dataset Limitations
   a. The reliance on historical data limits the ability to account for structural changed in the economy or stock market. For instance, COVID-19 pandemic introduced unique dynamics that may not align with historical trends.
   b. The use of monthly aggregated data, while necessary, have diluted details and volatility observations in daily data.
2. Machine Learning
   a. The relatively low R^2 scores highlight potential underfitting, suggesting that additional features or non-linear relationships could improve predictive performance.
3. Assumptions in Hypothesis Testing
   a. The assumptions of normality and independence in the data may not hold entirely, which could affect the validity of the t-tests and other statistical analyses.

b. The median-based threshold for splitting high and low regimes might oversimplify the analysis.

Bibliography

1. CBOE Volatility Index (VIX): What Does It Measure in Investing? Investopedia. Accessed January 18, 2025. https://www.investopedia.com/terms/v/vix.asp
2. Volatility: Meaning in Finance and How It Works With Stocks. Accessed January 18, 2025. https://www.investopedia.com/terms/v/volatility.asp
3. Intraday Return: What it Means, How it Works. Investopedia. Accessed January 18, 2025. https://www.investopedia.com/terms/i/intraday-return.asp
4. Yinan Yu. Lecture 12: Hypothesis testing part II-Statistical Methods for Data Science. Lecture presented at: DIT863 Statistical Methods for Data Science; December 19, 2024; Gothenburg, SE
5. Regression in Machine Learning: Definition and Examples | Built In. Accessed January 18, 2025. https://builtin.com/data-science/regression-machine-learning
6. Python | Linear Regression using sklearn - GeeksforGeeks. Accessed January 18, 2025. https://www.geeksforgeeks.org/python-linear-regression-using-sklearn/
7. Mean Square Error (MSE) | Machine Learning Glossary | Encord | Encord. Accessed January 18, 2025. https://encord.com/glossary/mean-square-error-mse/
8. R-Squared: Coefficient of Determination in Machine Learning. Accessed January 18, 2025. https://arize.com/blog-course/r-squared-understanding-the-coefficient-of-determination/
9. Random Forest Classifier using Scikit-learn. GeeksforGeeks. 00:44:15+00:00. Accessed January 18, 2025. https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/
10. Support Vector Regression (SVR) using Linear and Non-Linear Kernels in Scikit Learn. GeeksforGeeks. 00:23:19+00:00. Accessed January 18, 2025. https://www.geeksforgeeks.org/support-vector-regression-svr-using-linear-and-non-linear-kernels-in-scikit-learn/