FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Ebenezer Yeboah	Ghana	ebenezeryeboah46@gmail.com	
Bright Effah	Canada	effahbright.eb@gmail.com	
Shen Yan	China	wella.shen@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	**Ebenezer Yeboah**
Team member 2	**Bright Effah**
Team member 3	**Yan Shen**

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.

1. Skewness

1.1 Definition:

Statistical measure of the asymmetry of a normal distribution is called skewness. According to Ramsey and Schafer (2002), a normal distribution usually exhibits perfect symmetry, showing neither a lean to the left nor a lean to the right. However, deviations from this perfect symmetry suggest that the distribution is skewed to one side, leading to an asymmetry of some type, usually indicated by the distribution's skewness. According to math, a distribution's skewness can be determined as follows (Wooldridge, 2019): $\frac{\sum_{t=1}^{n}(x_t-\bar{x})^3 . p_t}{n \cdot \sigma^3}$

where X = random variable; and Xi= ith random variable; X = mean of distribution; n = number of variables in the distribution; = standard deviation. From the formula, a distribution is called symmetric about zero if it has a symmetric distribution (normal distribution).

1.2 Description:

A distribution of data's asymmetry or lack of symmetry is measured by its skewness. A distribution is said to be skew if the data are not distributed uniformly, which prevents the normal distribution from having a mirror image on both sides. Because the majority of the data is concentrated on the left, a positively skewed distribution has a longer tail on the right side; a negatively skewed distribution has the opposite effect. In order to comprehend the properties of a dataset, skewness is frequently employed in statistical analysis and exploratory data analysis. It offers important insights into the structure of a distribution.

1.3 Demonstration

We use real-world examples of the distribution of daily returns of cryptocurrencies like Dogecoin (DOGE_USD) and Bitcoin (BTC_USD) between 2020 and 2023 from the Yahoo Finance website to show the skewness of a normal distribution. These cryptocurrencies were our choice because of their significant annual volatility. View the detailed statistics of the chosen coins below.

The next table, labeled "Descriptive Statistics for the Daily Percentage Returns," indicates that the skewness of the daily returns for Bitcoin is comparatively close to zero. The daily returns of Dogecoin are significantly positively biased, up to 17, in comparison to Bitcoin. These daily returns display two identical financial instruments with varying skewness.

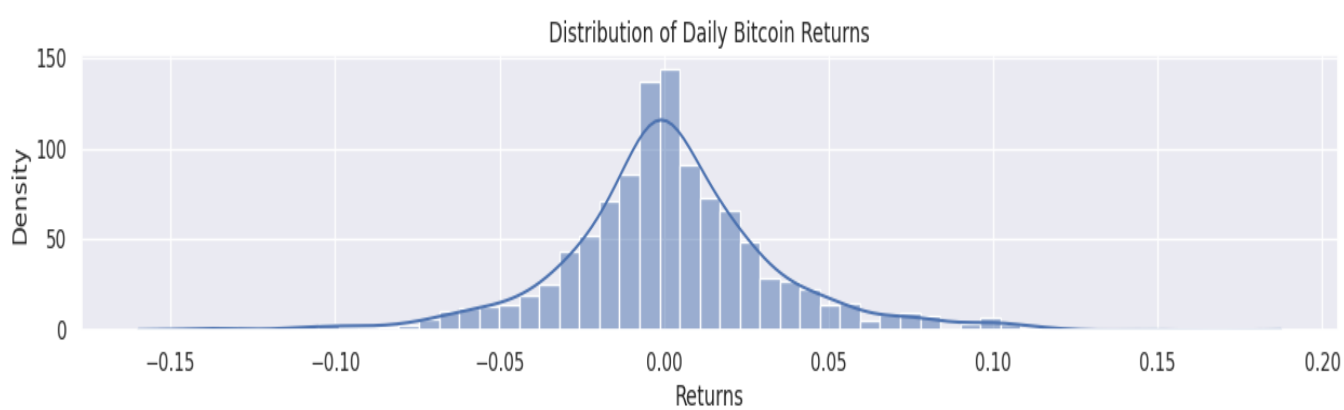
```
import numpy as np
import pandas as pd
from datetime import date, timedelta
import yfinance as yf
```

```
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style="darkgrid")
```

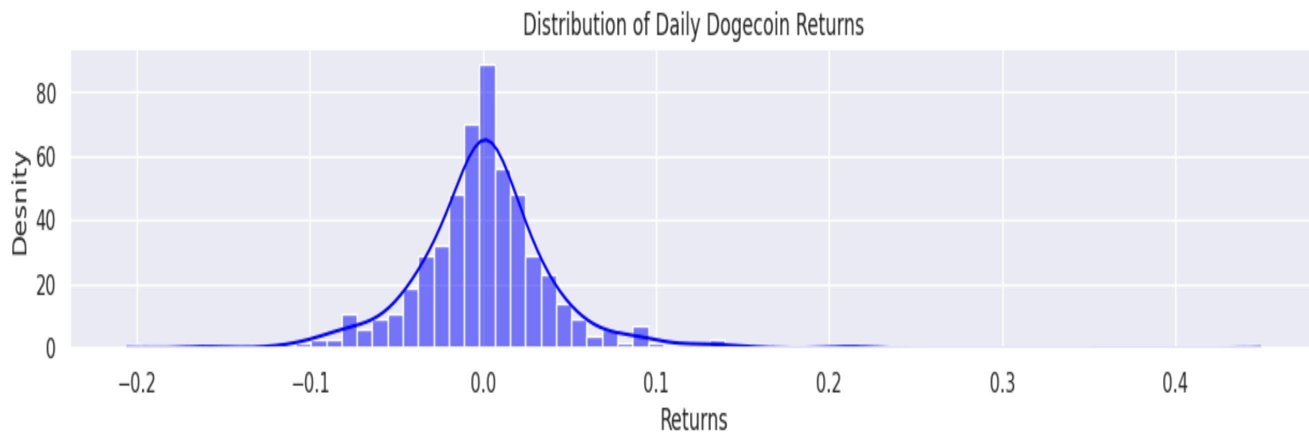
```
# Plot the histogram of Daily Bitcoin Returns
# First histogram
ax1=plt.figure(figsize=(15, 5))
plt.subplot(2, 1, 1)
sns.histplot(returns[:len(returns)//1], x="BTC-USD", kde=True)
plt.title("Distribution of Daily Bitcoin Returns")
plt.xlabel("Returns")
plt.ylabel("Density")

# Second histogram
plt.figure(figsize=(15, 5))
plt.subplot(2, 1, 2)
sns.histplot(returns[len(returns)//2:], x='DOGE-USD', color='blue', kde=True)
plt.title("Distribution of Daily Dogecoin Returns")
plt.xlabel("Returns")
plt.ylabel("Desnity")

plt.show()
```

1.4 Diagram

In the diagrams below, we show the histogram of the returns for these two cryptocurrencies. The diagram below confirms that the daily returns of Bitcoin are relatively normal or symmetric meaning perfectly skewed, although slightly positive.



The diagram below confirms that the daily returns of Dogecoin are positively skewed,



1.5 Diagnosis:

The skewness of a distribution can be determined using the tests listed below:

(i) Statistical descriptions: To do this, use the formula found in *Section 1.1*. A set of data that is balanced would have 0% skewness. Conversely, a positively skewed set of data will have a skewness greater than 0, whilst the opposite is true;

(ii) Examining a frequency plot or histogram visually: Here, the general guideline is to heed the distribution's label, which indicates whether it is favorably skewed to the right or negatively skewed to the left;

(iii) Utilizing Box and QQ Plots: A box plot will show the magnitude of outliers in addition to central tendencies. The median will be lower than the mean in a positively skewed distribution, and the box plot's right whiskers will be longer. Moreover, the opposite is true. Quantiles in a dataset are compared to those of a normal distribution using QQ plots. The data may be skewed if there is a divergence from the straight line.

1.6 Damage:

There are many unmet expectations in a skewed distribution. This is because it results in: (i) skewed estimates as a result of anomalies in the dispersion and central tendency measures. A right-skewed distribution, for instance, usually has a mean that is higher than its median; (ii) the existence of outliers and extreme values, which can result in violations of the normality assumption; (iii) the unreliability of hypothesis testing primarily because of the violations of the normality assumptions; (iv) inaccurate interpretation of the results, since a slight change in one variable can have a major impact on another.

1.7 Directions:

As stated by Ramsey & Schafer (2002), Moore, McCabe & Craig (2018), and Wooldridge (2019), there are a number of strategies to address some of the damages mentioned above. The use of non-parametric statistical techniques that do not presuppose a particular distribution; the application of robust statistical methods that are less susceptible to outliers and deviations from normalcy; data transformations such as logarithmic, square root, and reciprocal transformations to smooth the distribution of the data; Using classification and grouping to reduce the influence of a few extreme values or outliers in a distribution is known as; using resampling techniques like bootstrapping to assess statistics and produce confidence intervals is known. Using bootstrapping, we are able to produce estimates that are more resilient to data skewness; the selection of tactical statistical methods that are more resilient to data skewness.

2. Kurtosis / Heteroscedasticity

2.1 Definition:

A statistical metric called kurtosis is used to evaluate the form of a distribution's tails. The distribution is compared to a normal distribution to ascertain the weight of the tails. In the context of volatility modeling, this metric aids in identifying the existence of fat tails and extreme price swings (Tsay, 2005).

$$\text{Kurtosis} = \frac{(n-1)\cdot E[(x - \bar{x})^4]}{\sigma^4}$$

When n is the number of data points, μ is the data points' mean, and σ denotes their standard deviation. Heteroscedasticity, on the other hand, describes the uneven or fluctuating volatility levels that are seen in a time history. It suggests that there are times of both high and low volatility because the volatility does not remain constant over time. (Campbell et al., 1997). Heterogeneity $= \frac{\sigma^2}{\mu^2}$ where the mean of the data points is represented by μ and the standard deviation by σ.

2.2 Description:

In the context of volatility modeling, a high kurtosis value suggests that extreme events are more likely than they would be in a normal distribution. This indicates the existence of large price swings and fat tails, both of which must be taken into account for a precise risk assessment.

Contrarily, heteroscedasticity indicates that the volatility is not stable over time. Because volatility fluctuates over time, modeling and estimating risk measurements may be affected, perhaps producing unreliable conclusions.

2.3 Demonstration

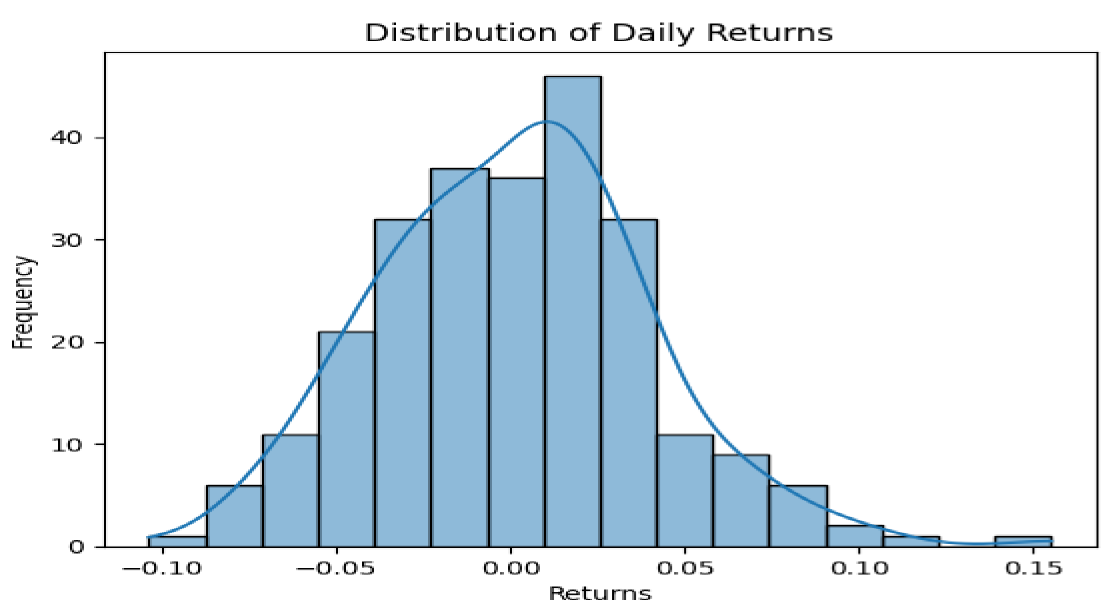
We will employ simulated financial data to illustrate the difficulties associated with kurtosis and heteroscedasticity. We will specifically be working with a dataset that includes a stock's daily returns over a given time frame.

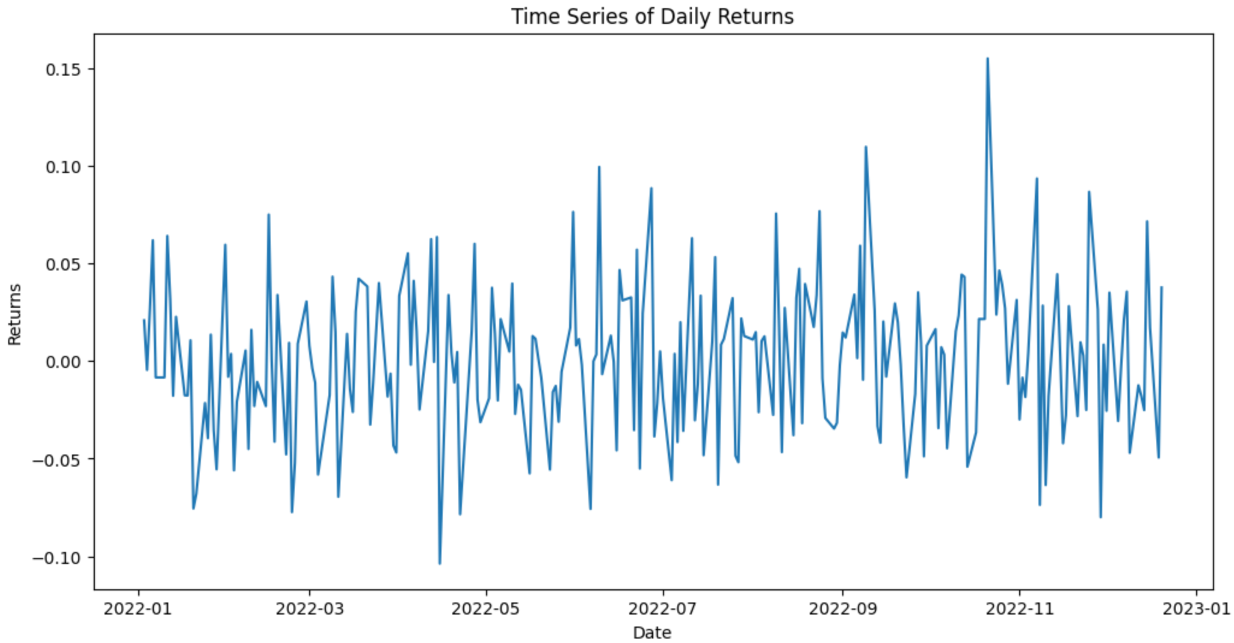
```
import numpy as np
import pandas as pd
import random
import seaborn as sns
import matplotlib.pyplot as plt
# Set the seed value for NumPy random number generator
np.random.seed(42)

# Set the seed value for Python random number generator
random.seed(42)

# Set the parameters
mean_return = 0.001  # Mean return of the stock
volatility = 0.04   # Volatility of the stock

# Set the number of trading days
num_days = 252

# Generate the daily returns
returns = np.random.normal(mean_return, volatility, num_days)
import seaborn as sns
import matplotlib.pyplot as plt

# Plot the histogram of returns
sns.histplot(returns, kde=True)
plt.title("Distribution of Daily Returns")
```

```
plt.xlabel("Returns")
plt.ylabel("Frequency")
plt.show()

# Generate the dates for the time series
dates = pd.date_range(start='2022-01-01', periods=num_days, freq='B')

# Create a DataFrame with dates and returns
df = pd.DataFrame({'Date': dates, 'Returns': returns})

# Plot the time series of returns
plt.figure(figsize=(12, 6))
plt.plot(df['Date'], df['Returns'])
plt.title("Time Series of Daily Returns")
plt.xlabel("Date")
plt.ylabel("Returns")
plt.show()
```

2.4 Diagram

The time series plot of daily returns and the histogram of the distribution of daily returns will be displayed in the diagrams, emphasizing times of high and low volatility. It will give an illustration of the various degrees of volatility.





2.5 Diagnosis

The histogram suggests that the distribution's kurtosis is probably quite similar to a normal distribution. The distribution does not appear to differ significantly from a normal distribution in terms of kurtosis, unless there are exceptionally heavy or light tails present. There appears to be heteroscedasticity in the time series plot, which displays notable volatility variations and fluctuations. The volatility is not constant and varies over different time periods, according to patterns of changing volatility.

2.6 Damage

The potential risk associated with extreme events may be underestimated if the influence of high kurtosis is overlooked. Inaccurate derivatives valuation and insufficient hedging strategies may result from this. Analyzing heteroscedasticity in the same way can lead to bias in volatility estimates and compromise the precision of risk measures. This could lead to decisions about risk management, hedging tactics, and pricing models being jeopardized.

2.7 Directions

Examining models that take fat-tailed distributions into account, like generalized autoregressive conditional heteroscedasticity (GARCH) models, is one way to address kurtosis. More accurate data representation is achieved by GARCH models, which take into account the phenomenon of volatility clustering and capture the time-varying nature of volatility (Tsay, 2005). Heteroscedasticity - robust estimation methods such as weighted least squares (WLS) or generalized least squares (GLS) can be used to handle heteroscedasticity. By accounting for different degrees of volatility, these techniques produce estimates that are more trustworthy (Campbell et al., 1997).

3. Sensitivity to Outliers

3.1 Definition:

According to the National Institute of Standards and Technology, an outlier is "an observation that lies an abnormal distance from other values in a random sample from a population". (NIST)

Commonly, outliers are identified using Tukey Fences (John), where the lower fence is 1.5 times the Interquartile Range (IQR) below the 25th percentile, and the upper fence is 1.5 times the IQR above the 75th percentile.

3.2 Description:

In a dataset, outliers are data points that deviate markedly from the majority, possibly due to measurement errors or representing extreme cases. While there isn't a strict definition, a common approach involves measuring the IQR, the range between the 25th percentile (Q1) and the 75th percentile (Q3), with data points beyond the range (Q1 - 1.5 * IQR, Q3 + 1.5 * IQR) considered outliers.

3.3 Demonstration

Using 1 year BTC-USD return data, we can calculate relevant parameters and find outliers. Here is the summary:

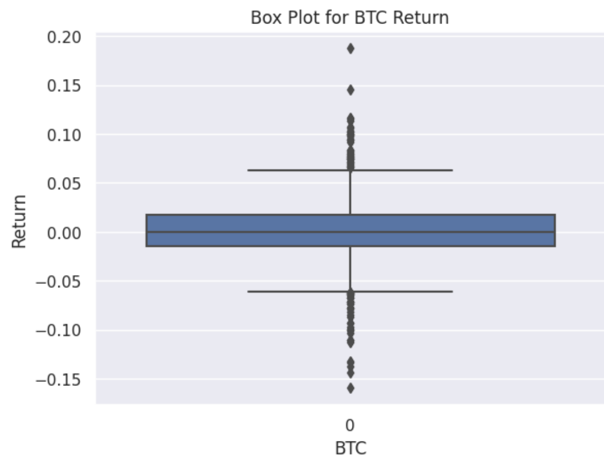
Q1 = -0.014, Q3 = 0.017Interquartile Range (IQR) = 0.031

Lower Fence = -0.061, Upper Fence = 0.064

Number of outliers = 86, Total data points = 1095, Percentage of outliers = 0.079

3.4 Diagram

Boxplot is the most commonly used visualization to identify outliers, first introduced in 1970 by John Tukey. ("Box Plot") The box represents the IQR, the middle line is the median, and whiskers extend to 1.5 times the IQR. Outliers are depicted as rhombus shapes beyond this range.



3.5 Diagnosis:

A common way to find outliers is to use Tukey's Fences with the following steps:

A. Sort the data points in ascending order.
B. Find Q1 (25th percentile) and Q3 (75th percentile).
C. Calculate interquartile range IQR = Q3 - Q1.
D. Calculate lower fence = Q1 - 1.5* IQR and upper fence = Q3 + 1.5*IQR.
E. Any data point less than lower fence or larger than upper fence is an outlier.

Alternative methods include z-scores or statistical criteria based on residuals.

3.6 Damage

While modeling volatility, outliers can damage the accuracy and reliability of the model and lead to many problems:

1. Biased volatility estimation: The presence of outliers can disproportionately influence the estimation of volatility. Any modeling approach relying on absolute return or squared residuals can be significantly influenced by outliers.
2. Unreliable risk measurement: Volatility is a crucial component in measuring risks. Extreme returns (outliers) can distort volatility measures and damage the credibility of such risk measurements.
3. Unstable model: Outliers can disrupt the normal dynamics assumed by the model, making the estimation inconsistent or failing to converge.

3.7 Directions

To address the problem caused by outliers in volatility modeling, several robust models can be considered:

1. Robust GARCH: Robust GARCH models, such as RGARCH, incorporate robust estimation techniques to reduce the impact of outliers on volatility estimates.(Boudt et al.) These models use robust loss functions, like Huber or Tukey biweight, instead of the standard squared loss function in GARCH models.
2. Stochastic Volatility Models: Stochastic volatility models, such as the Heston model, explicitly model the volatility as a stochastic process, allowing for flexibility in capturing complex volatility patterns. (Krichene) These models can be less sensitive to outliers because they do not assume constant volatility over time.
3. Nonparametric Models: Kernel-based approaches, such as Kernel Density Estimation (KDE) or Local Volatility Models, do not rely on strict parametric assumptions about the distribution of returns. These models can be less sensitive to outliers.

4. Non-stationary

4.1 Definition:

The distribution of these two cryptocurrencies' prices in levels, or their raw form, is non-stationary since it does not follow any pattern, as we demonstrate in Panel A of the diagrams below. But when we look at panel B's first difference, we can see that their trends become mean-reverting and, consequently, stationary. 4.1 Explanation To put it simply, stationarity is a property of time series processes that implies stability due to the time-invariant nature of the series. Because of the constant mean, constant variance, and constant autocovariances for each lag, non-stationarity implies that a time series is time-variant over time (Gujarati, 2018; Brooks, 2019; Wooldridge, 2019). In terms of mathematics, non-stationarity is explained by the following two models: The random walk model with drift (i) and the trend-stationary process (ii) are the two models.

4.2 Description:

When a set of statistical characteristics, such as mean and variance, follow a particular trend over time rather than being randomly distributed, it's referred to as non-stationarity. It is problematic when a time series is non-stationary because it makes it impossible to forecast or precisely estimate statistical measures. For time series analysis and forecasting to be accurate, non-stationarity must be recognized and addressed. Inaccurate predictions and misleading results can result from neglecting to take non-stationarity into account. Using pre-processing methods and selecting models that are suitable for non-stationary data are essential steps when working with time series that exhibit changing statistical properties.

4.3 Demonstration:

We use actual examples of the prices of cryptocurrencies between 2020 and 2023 from the Yahoo Finance website, such as Bitcoin (BTC_USD) and Dogecoin (DOGE_USD), to show the non-stationarity of a time series. The reason cryptocurrencies caught our attention is that they have experienced significant fluctuations over time.

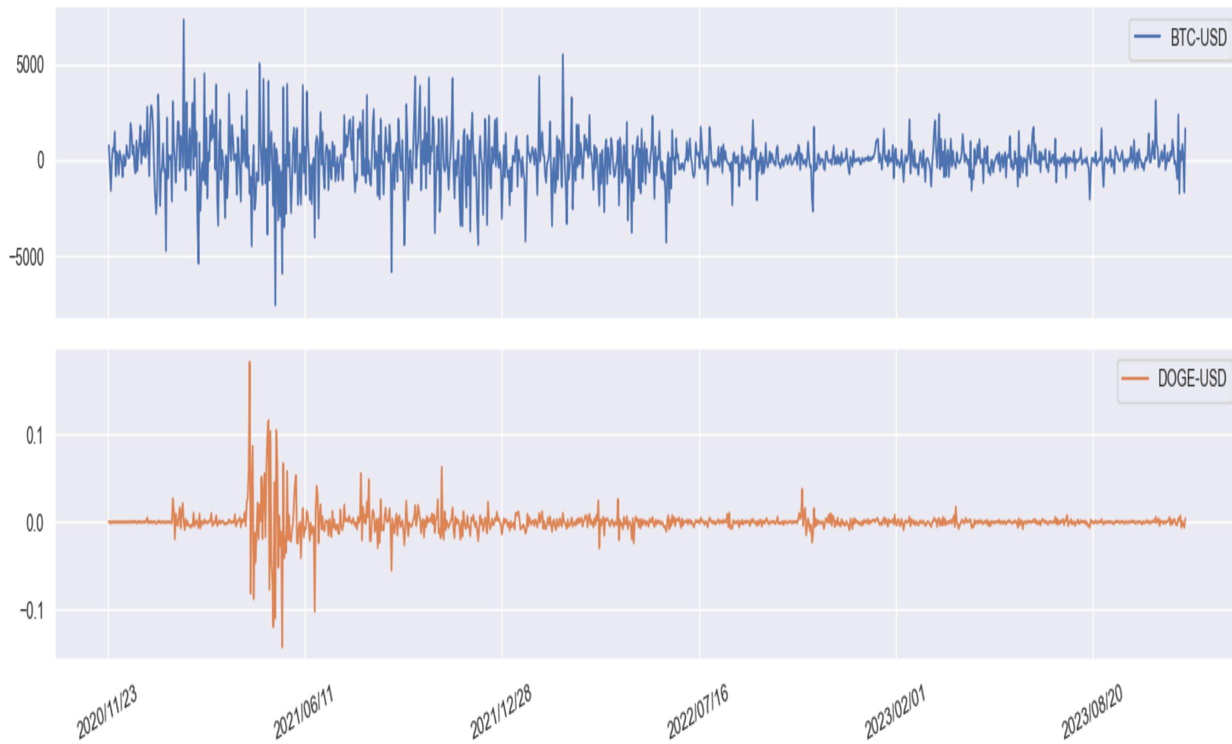
```
# Plotting the daily pricess of bitcoin and dodecoin
port.plot(subplots=True, figsize =(15, 5))
plt.xticks(rotation=20)
plt.tight_layout()
plt.show()
```

```
port_seconddiff=port_firstdiff.diff()
port_seconddiff.plot(subplots=True, figsize=(15, 5))
plt.xticks(rotation=20)
plt.tight_layout()
plt.show()
```

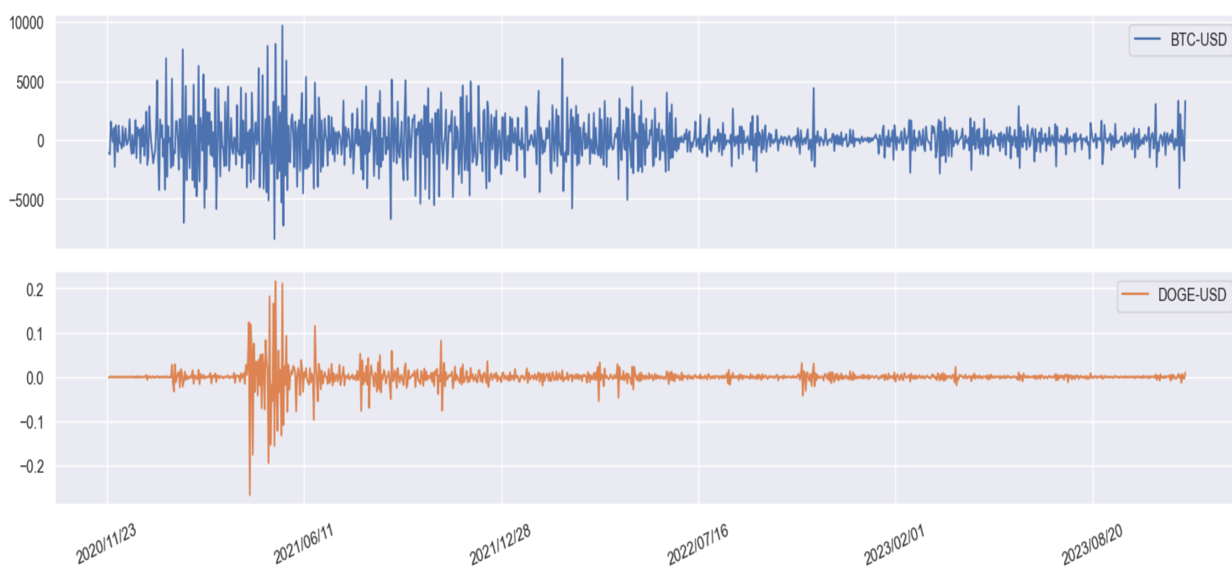
4.4 Diagram

The time series plot of daily returns and the histogram of the distribution of daily returns will be displayed in the diagrams, emphasizing times of high and low volatility. It will give an illustration of the various degrees of volatility.

First Difference of Daily Prices of Bitcoin and Dogecoin between 2020 and 2023



First Difference of Daily Prices of Bitcoin and Dogecoin between 2020 and 2023



4.5 Diagnosis

The histogram suggests that the distribution's kurtosis is probably quite similar to a normal distribution. The distribution does not appear to differ significantly from a normal distribution in terms of kurtosis, unless there are exceptionally heavy or light tails present. There appears to be heteroscedasticity in the time series plot, which displays notable volatility variations and fluctuations. The volatility is not constant and varies over different time periods, according to patterns of changing volatility.

4.6 Damage

The potential risk associated with extreme events may be underestimated if the influence of high kurtosis is overlooked. Inaccurate derivatives valuation and insufficient hedging strategies may result from this. Analyzing heteroscedasticity in the same way can lead to bias in volatility estimates and compromise the precision of risk measures. This could lead to decisions about risk management, hedging tactics, and pricing models being jeopardized.

4.7 Directions

Examining models that take fat-tailed distributions into account, like generalized autoregressive conditional heteroskedasticity (GARCH) models, is one way to address kurtosis. More accurate data representation is achieved by GARCH models, which take into account the phenomenon of volatility clustering and capture the time-varying nature of volatility (Tsay, 2005). Heteroscedasticity-robust estimation methods such as weighted least squares (WLS) or generalized least squares (GLS) can be used to handle heteroscedasticity. By accounting for different degrees of volatility, these techniques produce estimates that are more trustworthy.

Bibliography:

- "7.1.6. What Are Outliers in the Data?" National Institute of Standards and Technology, www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm. Accessed 26 Nov. 2023.
- Boudt, Kris, et al. "Robust Forecasting of Dynamic Conditional Correlation GARCH Models." SSRN Electronic Journal, Elsevier BV, 2010. Crossref, doi:10.2139/ssrn.1721421.
- "Box Plot." Wikipedia, Wikimedia Foundation, 19 Nov. 2023, en.wikipedia.org/wiki/Box_plot. Accessed 26 Nov. 2023.
- John, Tukey. Exploratory Data Analysis. 1st ed., Pearson, 1977.
- Krichene, Noureddine. "Modeling Stochastic Volatility with Application to Stock Returns." SSRN Electronic Journal, Elsevier BV, 2003. Crossref, doi:10.2139/ssrn.879201.
- Ramsey, F.L., & Schafer, D.W. (2002). The statistical sleuth : a course in methods of data analysis.
- Wooldridge, J.M. (2019). Correlated random effects models with unbalanced panels. *Journal of Econometrics*.
- Tsay, R.S. (2005). Analysis of Financial Time Series. Technometrics, 48, 316 - 316.

- ☐ Campbell, K. W. (1997). Empirical near-source attenuation relationships for horizontal and vertical components of peak ground acceleration, peak ground velocity, and pseudo-absolute acceleration response spectra. Seismological research letters, 68(1), 154-179.
- ☐ Brooks, C. (2019). *Introductory Econometrics for Finance* (4th ed.). Cambridge: Cambridge University Press. doi:10.1017/9781108524872