GROUP WORK PROJECT # 3 MScFE 600: FINANCIAL DATA
GROUP NUMBER: 2791

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Marin Yordanov Stoyanov	Bulgaria	azonealerts@gmx.com	
Chia-Cheng Chang	Taiwan	xiong7238@gmail.com	
Azamat Zhaksylykov	USA	a.zhaksylykov@gmail.com	

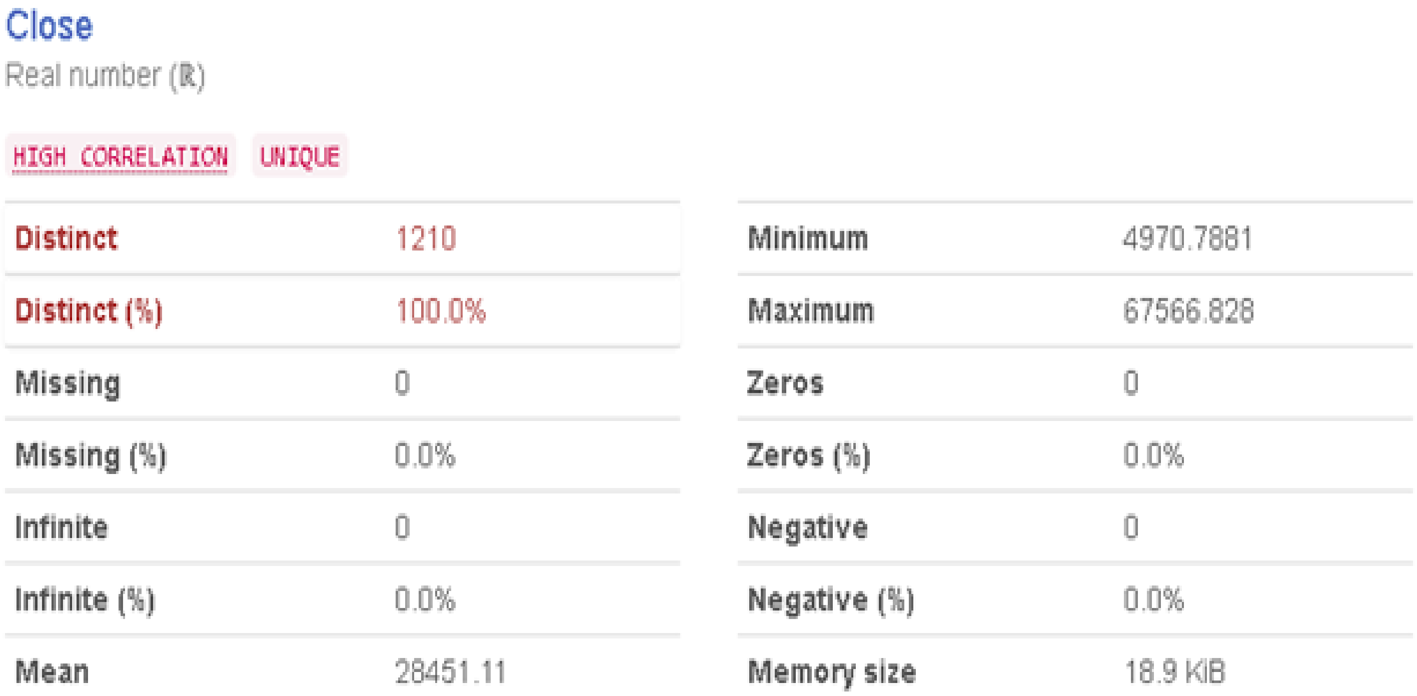
Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).

Team member 1	Marin Yordanov Stoyanov
Team member 2	Chia-Cheng Chang
Team member 3	Azamat Zhaksylykov

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.

Section 1: Statistics

Our group chose to work with Bitcoin data. In this case I will use Bitcoin's Close data for further analysis. Usually with stocks it is the better approach to use the Adjusted Close information because it takes into consideration that stocks do distribute dividends and this affects the instrument price, but in our case with Bitcoin the situation is a bit different. We have taken into consideration that Bitcoin has never had a dividend distribution so the close price and the adjusted close price are all the same and this is why I will use only the close price.

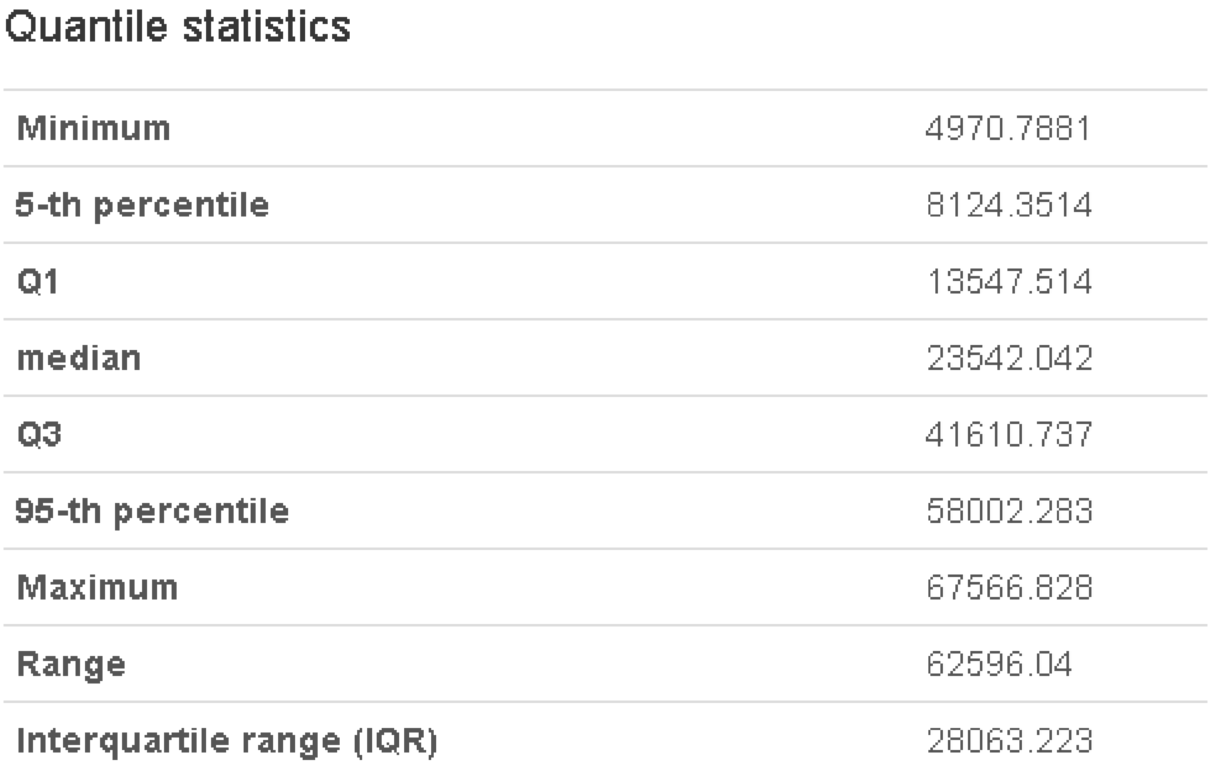


Close
Real number (ℝ)

HIGH CORRELATION UNIQUE

Distinct	1210	Minimum	4970.7881
Distinct (%)	100.0%	Maximum	67566.828
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	28451.11	Memory size	18.9 KiB

This is the statistics information about the Bitcoin's close price and now I will go deeper into the quantile and descriptive statistics:



Quantile statistics

Minimum	4970.7881
5-th percentile	8124.3514
Q1	13547.514
median	23542.042
Q3	41610.737
95-th percentile	58002.283
Maximum	67566.828
Range	62596.04
Interquartile range (IQR)	28063.223

Minimum = min{data}

Maximum = max{data}

Simply the minimal and maximal values of the data.

Percentile formula [5]:

$$P_x = \frac{x(n + 1)}{100}$$

P_x = The value at which x percentage of data lie below that value

n = Total number of observations

Median:

The Median is the middle of the data and in order to find its value we have to know if the data is odd or even [2] [3]. So, in these 2 cases we go like this (where n is the count of the data values):

$$\text{if } n \text{ is odd, } \operatorname{median}(x) = x_{(n+1)/2}$$

$$\text{if } n \text{ is even, } \operatorname{median}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$

Quartiles are values that divide a dataset into four equal parts. The formula for calculating quartiles depends on which quartile you want to calculate [1]:

Quartile Formula

$$\text{The Quartile Formula for Q1} = \frac{1}{4}(n + 1)^{\text{th}}\text{term}$$

$$\text{The Quartile Formula for Q3} = \frac{3}{4}(n + 1)^{\text{th}}\text{term}$$

$$\text{The Quartile Formula for Q2} = \text{Q3} - \text{Q1} \text{ (Equivalent to Median)}$$

Range:

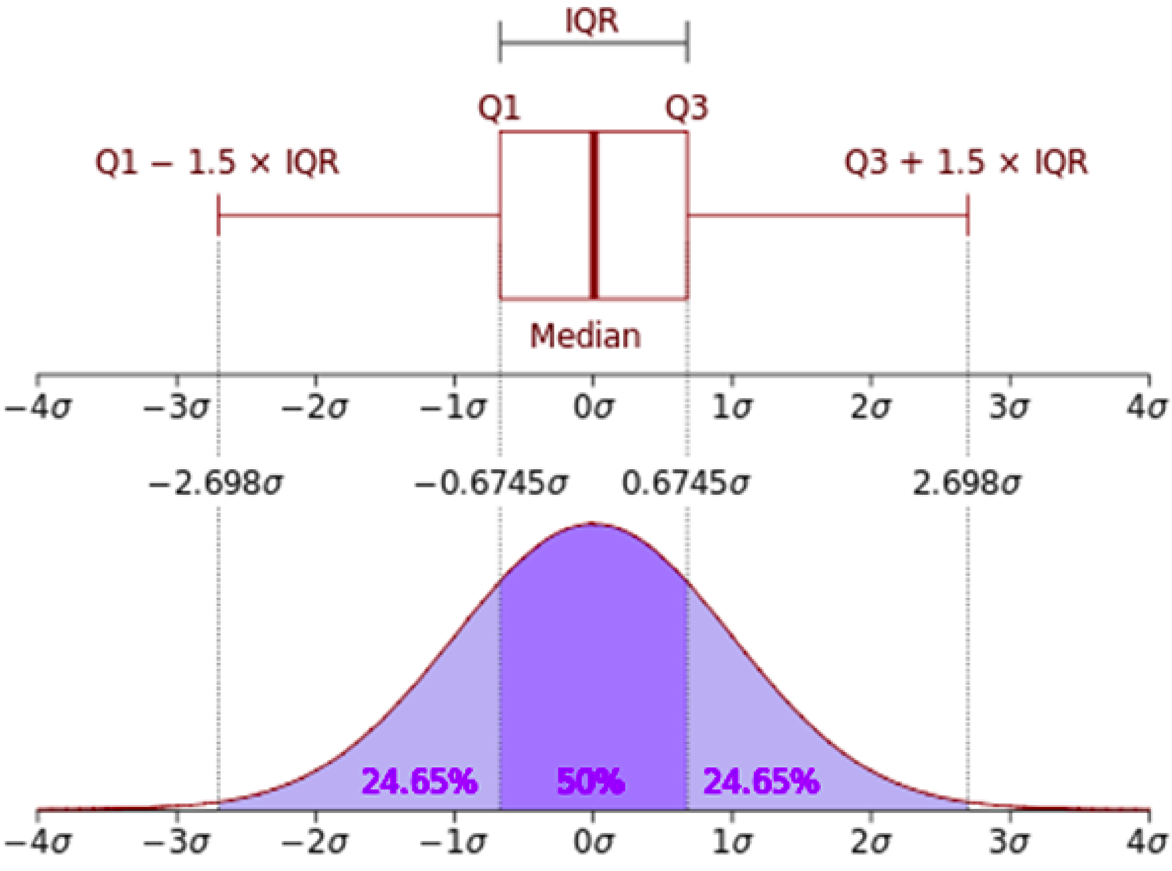
The range of a dataset is the difference between the largest and smallest values in the dataset. So, the range formula looks like this:

Range formula: Range = Max – Min

Interquartile range:

The interquartile range (IQR) is a measure of dispersion that describes the middle 50% of values when ordered from lowest to highest. [6]

The formula for calculating IQR is: IQR = Q3 – Q1

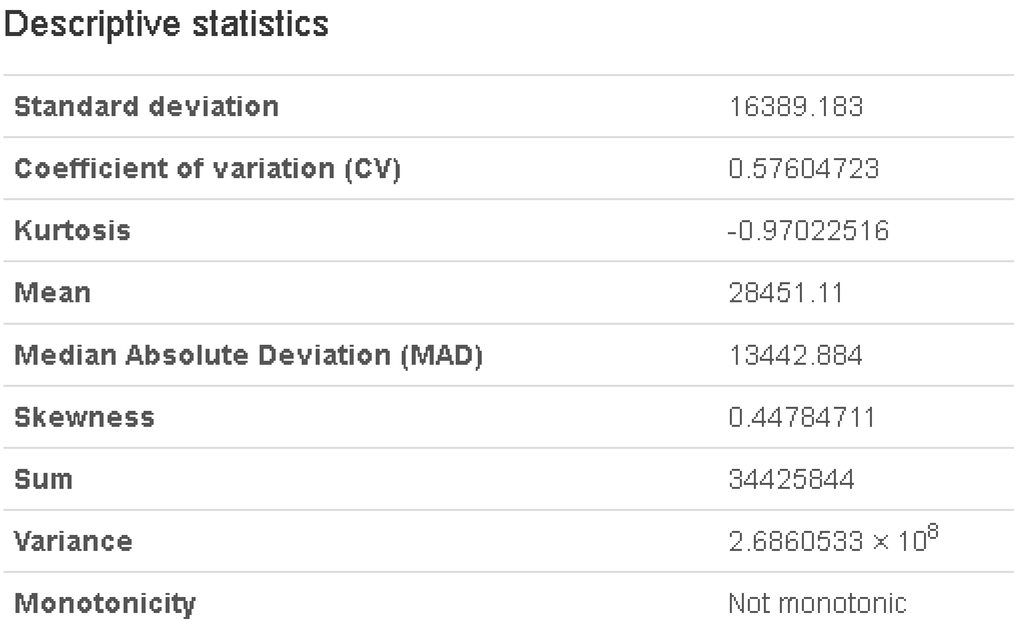


Now is time to look into the descriptive statistics:

Standard deviation formula :

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \mu|^2}{N}}$$

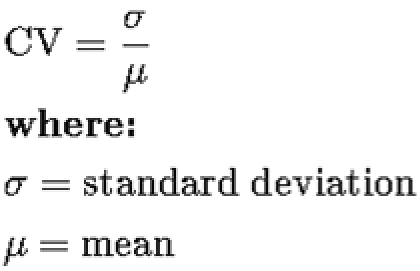
The standard deviation is the square root of the variance. We are using the square root of the variance because the value that comes out of this equation can be presented as distance. The distance can be in both ways: left and right from the mean so it can be negative and positive [7].



Descriptive statistics

Standard deviation	16389.183
Coefficient of variation (CV)	0.57604723
Kurtosis	-0.97022516
Mean	28451.11
Median Absolute Deviation (MAD)	13442.884
Skewness	0.44784711
Sum	34425844
Variance	2.6860533×10^8
Monotonicity	Not monotonic

Coefficient of variance:

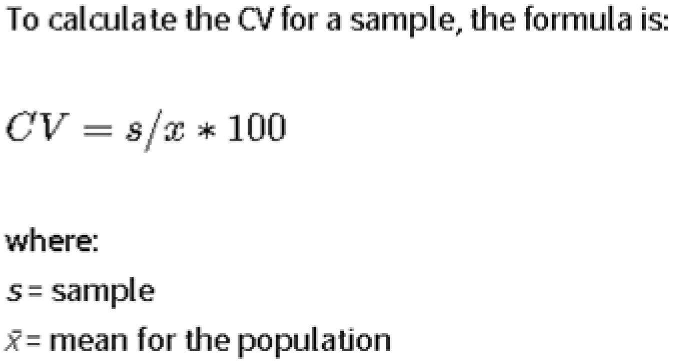


$$\text{CV} = \frac{\sigma}{\mu}$$

where:

σ = standard deviation

μ = mean



To calculate the CV for a sample, the formula is:

$$CV = s/x * 100$$

where:

s = sample

$\bar{x}$ = mean for the population

The coefficient of Variance is a ratio that is calculated by dividing the standard deviation of the dataset (STD) over the mathematical expectation (expected mean). This measure shows the potential expected return of the investment itself and if this is worth the volatility itself. Mainly interested in the potential downside risk that may realize over time. with one word: the expected worst case scenario due to volatility [8].

Kurtosis:

Kurtosis is a measure of how much the data is located and can be found in the tails (left AND right) of the dataset distribution. It can be called also a measure of volatility of volatility or in short, we can call it: vol of vol.

Kurtosis formula is:

$\beta_2 = (E(x)^4 / (E(x)^2)^2) - 3$

It is referred to as the third standardized moment of probability distributions.

Mean:

This is a measure of the arithmetic average and is calculated by summing all the values in the dataset and dividing the sum over the count of the values.

The arithmetic mean is the sum of all the data points divided by the number of data points [9].

$$\text{mean} = \frac{\text{sum of data}}{\text{\# of data points}}$$

$$\text{mean} = \frac{\sum x_i}{n}$$

Median Absolute Deviation (MAD):

This is a measure of how spread the data is and is used when the data is not normally distributed. As a rule of thumb you can say that when data is normally distributed you should use mean and standard deviation as main measures of the spread of the data, and when the data is NOT normally distributed you should use Median Absolute Deviation (MAD) as a measure of spread.

To me Median Absolute Deviation (MAD) looks a lot like some kind of regularization or standardization procedure because: from every value you have to subtract the sample mean, which somehow standardizes the data.

Median Absolute Deviation (MAD) formula [10] is like this:

$MAD = median(|Y_i – median(Y_i|)$

Skewness:

Skewness is a measure of asymmetry of the distribution. The data can be positively or negatively skewed or it can have no skewness. If there is now skew, then:

Zero skew: mean = median

Otherwise, it is either negative or positive, which means that the most amount of data is located left or right of the mean. For skewness calculation we are going to need the mean of the data:

$$\bar{X} = \frac{\sum_i^N X_i}{N}$$

Then we need the standard deviation (STD):

$$\sigma = \sqrt{\frac{\sum_i^N (X_i - \bar{X})^2}{N}}$$

And the final skewness formula [11] looks like this:

$$\text{Skewness} = \frac{\sum_i^N (X_i - \bar{X})^3}{(N-1) * \sigma^3}$$

Sum:

This is the value that comes out of summing all the other values in the data.

Variance:

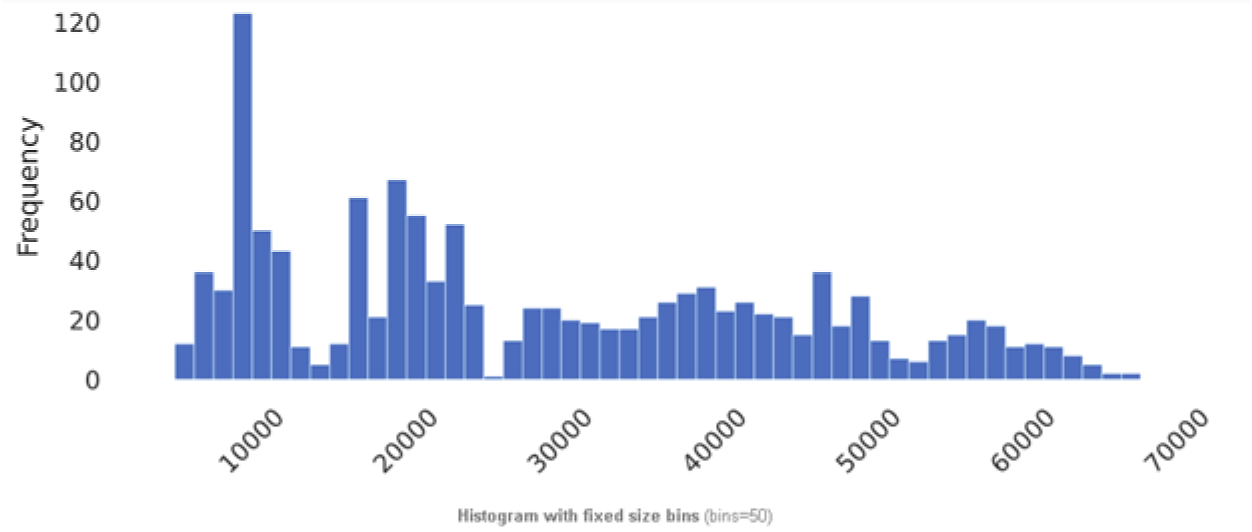
$$\text{Variance} = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{N}$$

This is the variance of data (also called volatility**)**

Monotonicity:

This means whether data is monotonically going down or going up. This simply cannot me True because the returns are measured in percentage growth or decrease and the very next value is independently and identically distributed random variable and it can come up with positive or negative (lets exclude the cases that it comes exactly 0), so there is now way with data to be monotonically rising or falling.

Histogram of the price data distribution:



The histogram shows that the distribution of the data is NOT normal. This can be understood either through the visualization of the data by plotting a histogram of the distribution and then seeing that the shape is not near a bell-shaped curve like it should be.

The other method to prove that this is not a normal distribution is by looking at the descriptive statistics and moreover the kurtosis and skewness parameter's values.

With this data the skewness is: 0.44784 and the kurtosis is -0.97022.

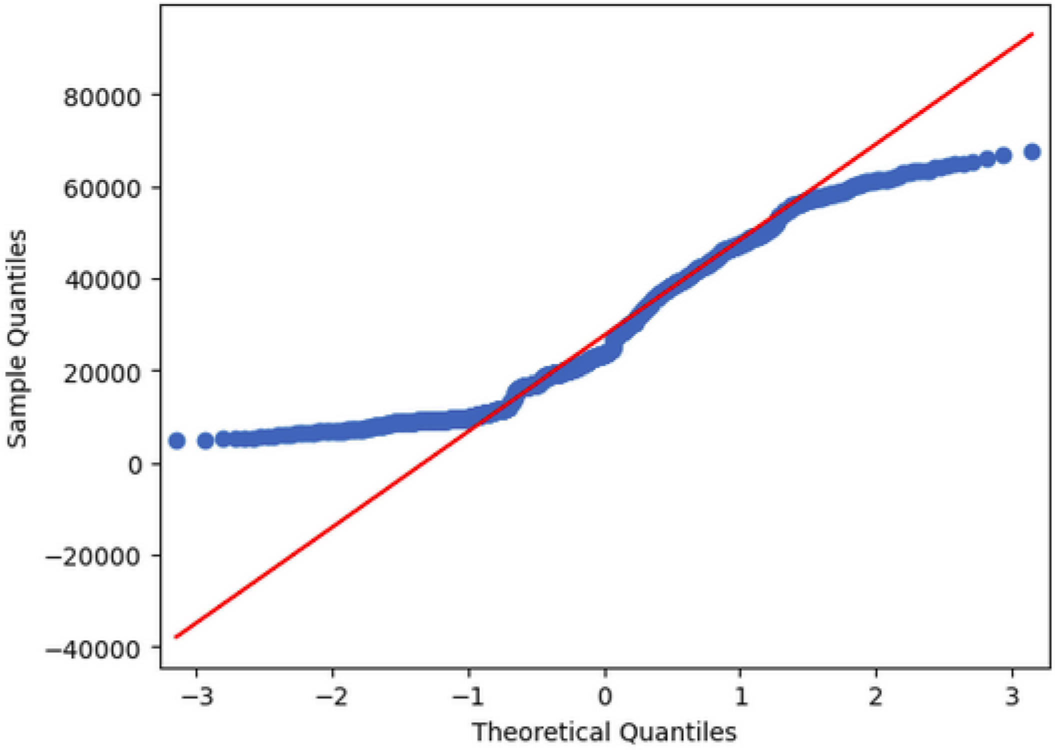
In order to have a normal distribution the skewness should be zero and in our case it is not. Also, the kurtosis should be 3 (as the normal distribution has) but in our case it is not.

So, this proves that the distribution of the data is NOT a normal one.

Let's be more precise and use some other methods to poove that this distribution is not normal.

Q-Q plot:

This is a qqplot of the raw price data:



From this qqplot we can show that the distribution of the data is NOT normal because the blue dots are not located close to the red line which is usual for a normal distribution.

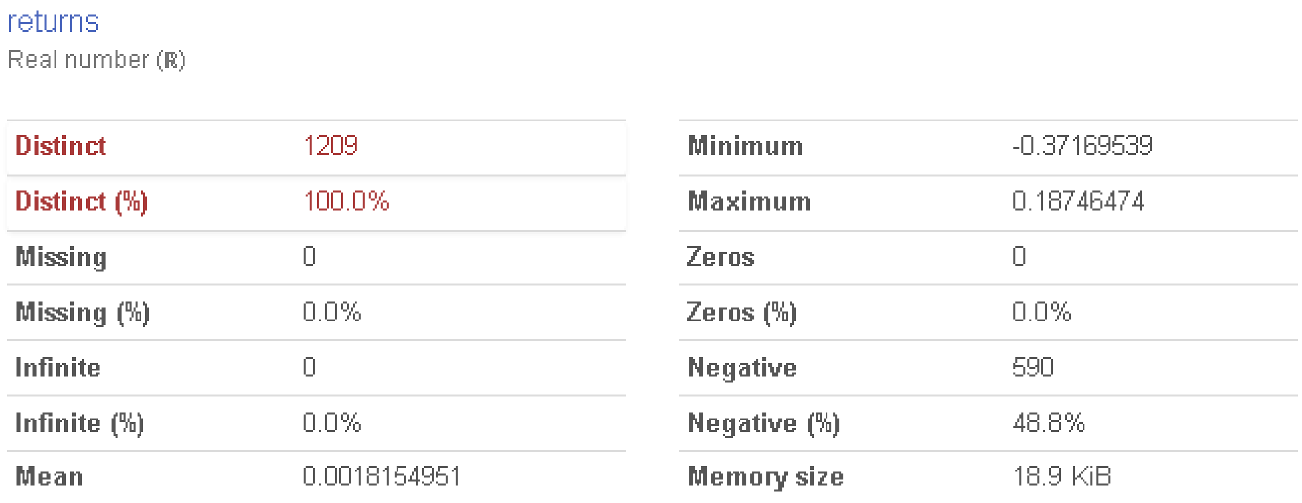
Visualizations are not exactly scientific proof for anything because every human being has his own interpretations and abstract thinking, and it is possible for someone not to be able to see it as it is. In this case I will use the very formal statistical test: Shapiro-Wilk test and it will mathematically (statistically) prove or not whether there is a normal distribution. If the value of the Shapiro test's p-value is less than 0.05. This means that with 95% confidence the data is NOT normally distributed.

The experiment showed a Shapiro-Wilk's p-value = 6.242878596495905e-23

So, this is the mathematical proof that shows the price data is not normally distributed:

Let's do it again but this time I will use the returns as a data source and let's see what will come up.

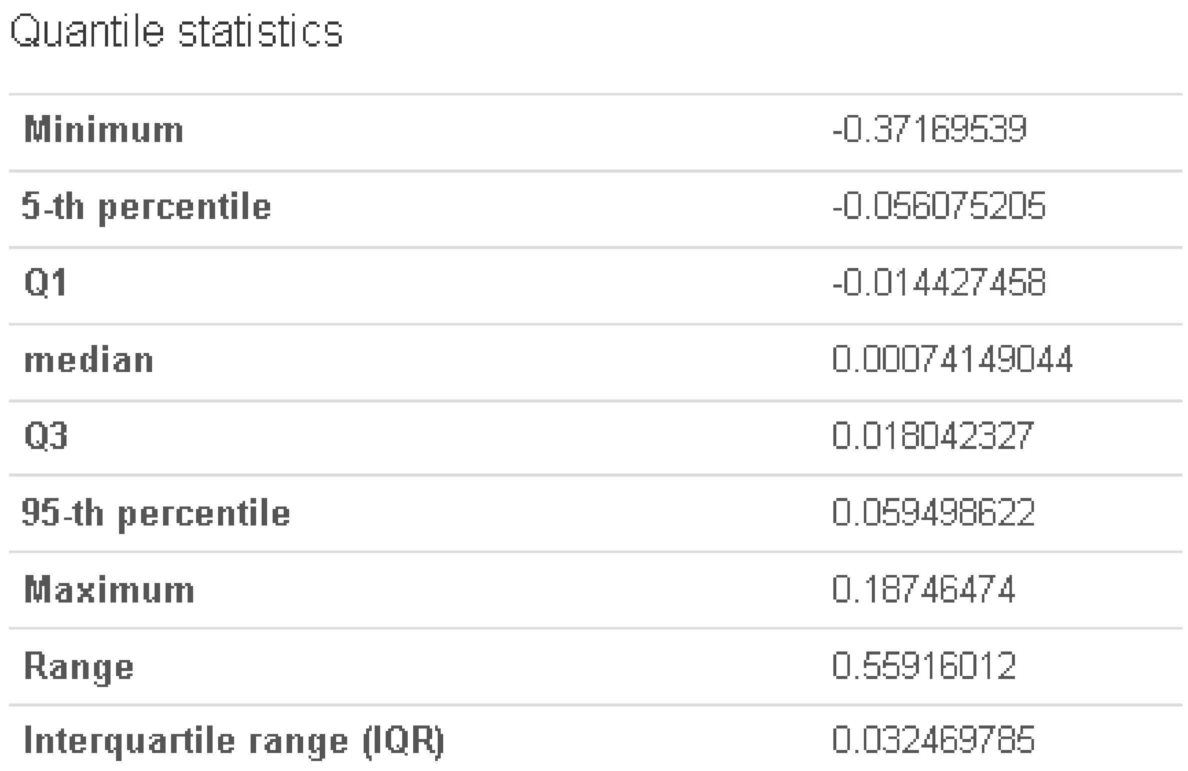
This is the statistics of the returns data:



returns
Real number (ℝ)

Distinct	1209	**Minimum**	-0.37169539
Distinct (%)	100.0%	**Maximum**	0.18746474
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	590
Infinite (%)	0.0%	**Negative (%)**	48.8%
Mean	0.0018154951	**Memory size**	18.9 KiB

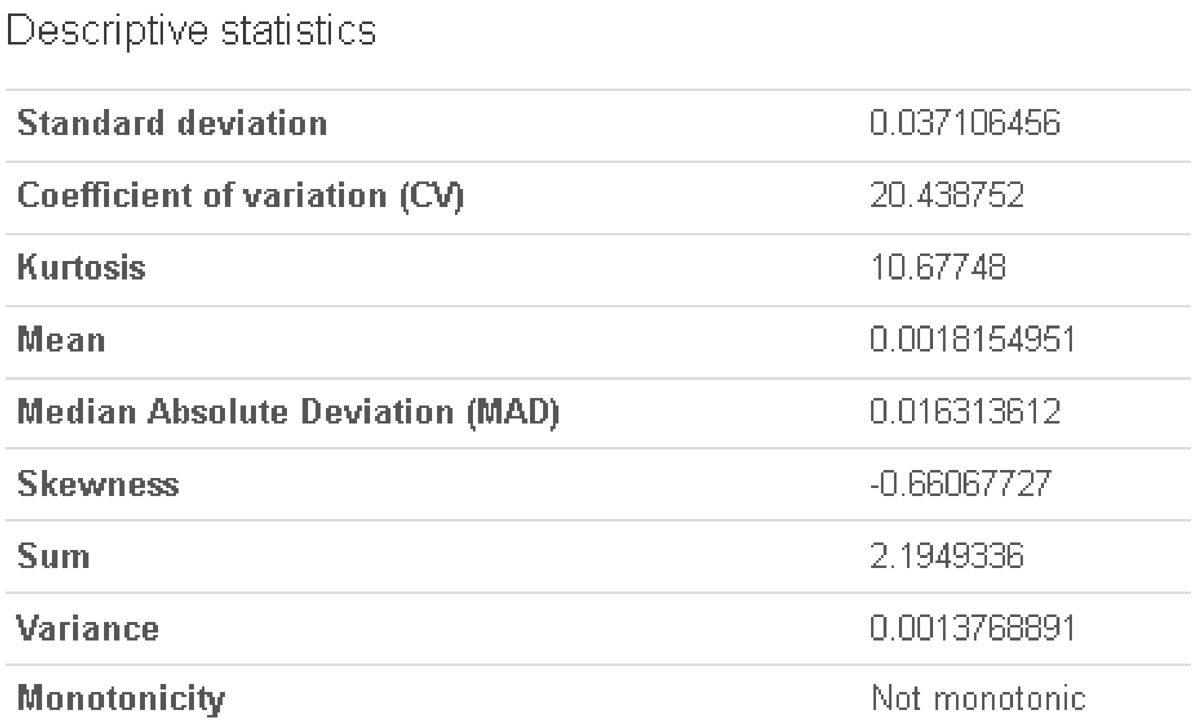
Quantile Statistics:



Quantile statistics

Minimum	-0.37169539
5-th percentile	-0.056075205
Q1	-0.014427458
median	0.00074149044
Q3	0.018042327
95-th percentile	0.059498622
Maximum	0.18746474
Range	0.55916012
Interquartile range (IQR)	0.032469785

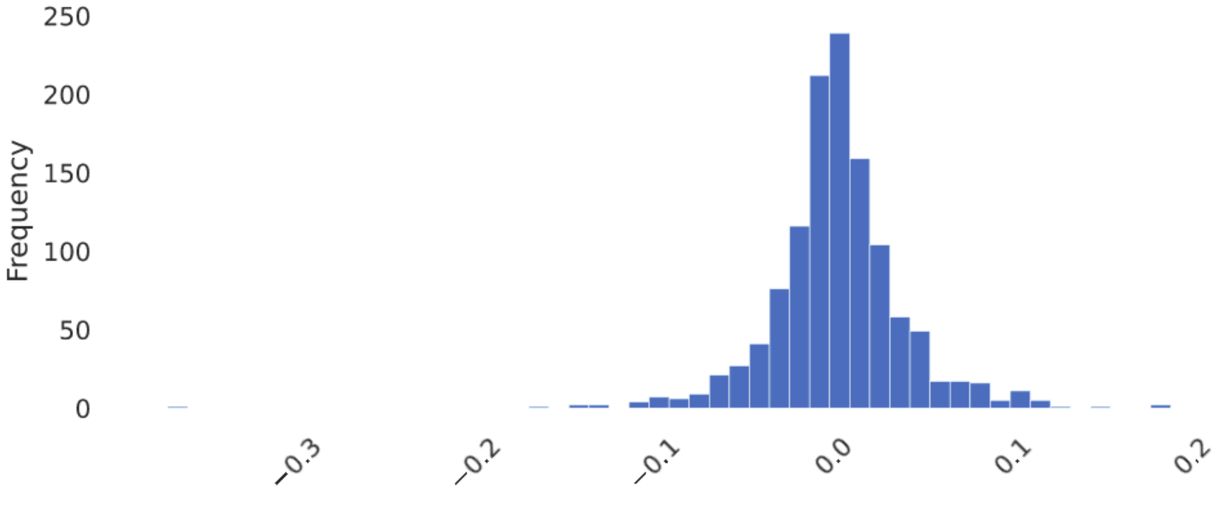
Descriptive Statistics:



Descriptive statistics

Standard deviation	0.037106456
Coefficient of variation (CV)	20.438752
Kurtosis	10.67748
Mean	0.0018154951
Median Absolute Deviation (MAD)	0.016313612
Skewness	-0.66067727
Sum	2.1949336
Variance	0.0013768891
Monotonicity	Not monotonic

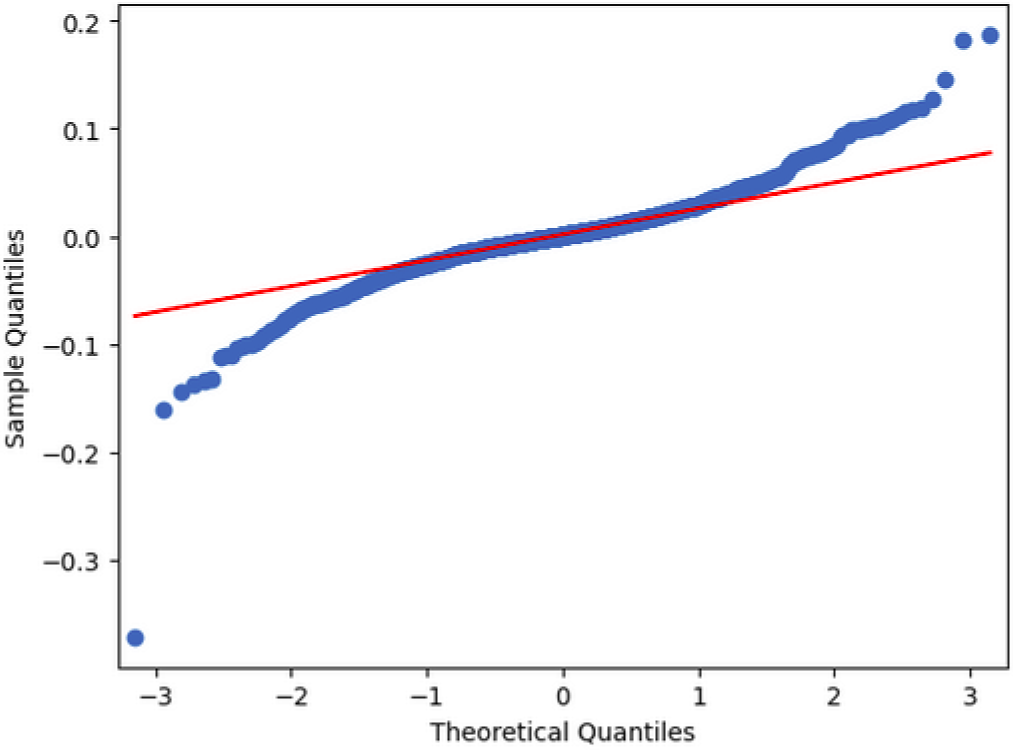
The histogram of the returns data looks like this:



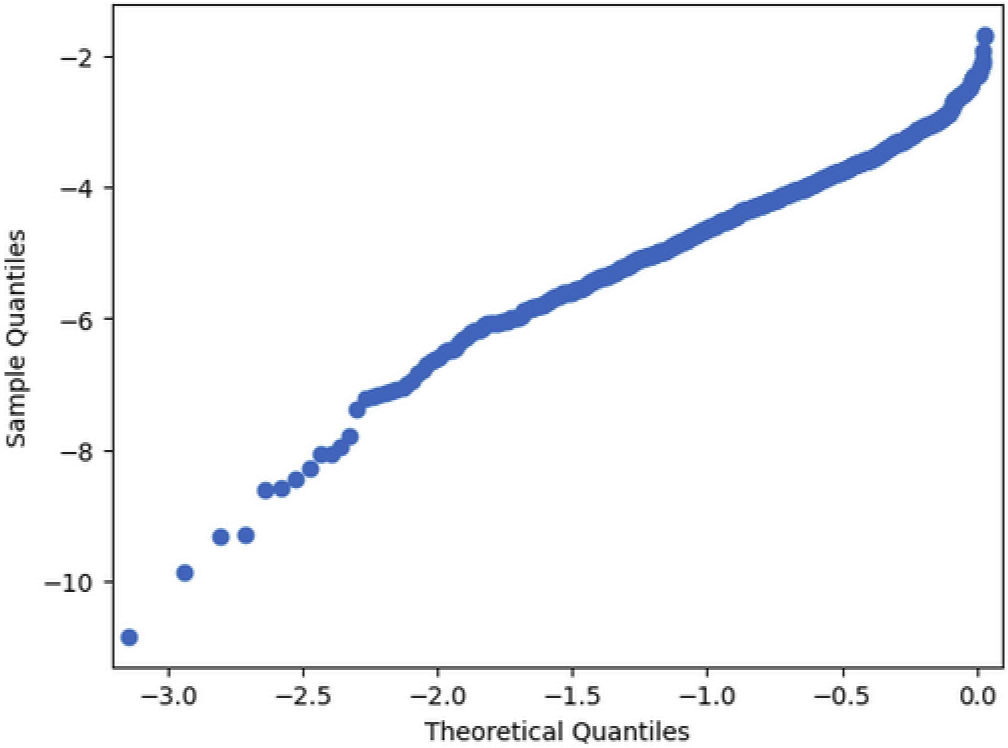
Histogram with fixed size bins (bins=50)

This time it looks a lot like a normal distribution with heavy tails.

Now it looks a lot better and very close to a normal distributed data, visible in the qqplot of the Bitcoin's return data:



Let's do a logarithm transform over the data and make qqplot again:



This is even better but as mentioned previously we cannot believe in a plot. We have to use mathematical proof for normality, which in our case is the Shapiro Wilk statistical test for normality.

This time the p-value = 1 so the logged returns data of Bitcoin is proved to be normally distributed.

The risks associated with the volatility (especially with Bitcoin volatility) is that very often price spikes can occur in either direction. This event can be explained with the fact that Bitcoin is not a very liquid investment instrument so there is no one to fill the demand/supply of the incoming trades/orders and this can lead to excessive loss. The heavy tails in the distribution histogram suggests that form time to time there could be a rare but very severe event that could affect the returns and a big loss can be realized.

Section 2: Correlations

Given the high trading frequency and liquidity of the 10-year US Treasury bond yield, we make the assumption that any impact from inflation and the actions of the US Federal Reserve have already been incorporated into the yield. Consequently, we will conduct a further analysis to determine how the 10-year US Treasury bond yield affects our trading target, Bitcoin.
By identifying the predictive nature of US Treasury bond yields, we can anticipate periods of high risk and take necessary precautions to avoid them.

Listed below are the resources and tools that we will be utilizing, which consist of mathematical formulas and corresponding definitions.

Return:

$$R_t = \frac{Price_t - Price_{t-1}}{Price_{t-1}}$$

Auto Correlation Function:

$$\frac{\sum_{i=1}^{n}(y_i-\bar{y}_i)(y_{i-h}-\bar{y}_{i-h})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2\ \sum_{i=1}^{n}(y_{i-h}-\bar{y}_{i-h})^2}}$$

Partial Autocorrelation Function:

$$\frac{cov(y_i, y_{i-h}|y_{i-1}, \dots, y_{i-h+1})}{\sqrt{var(y_i|y_{i-1}, \dots, y_{i-h+1}) \cdot var(y_{i-h}|y_{i-1}, \dots, y_{i-h+1})}}$$ [14]

For a time series, the h-th order partial autocorrelation is the partial correlation of yi with yi-h, conditional on yi-1,…, yi-h+1, i.e. [14]

Correlation:

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\ \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

AR(p) model with exogenous variable :

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \gamma_h x_h + \ldots + \epsilon_t$$

Augmented Dickey–Fuller test:

The testing procedure for the ADF test is same testing procedure as the Dickey–Fuller test [16] with the difference being that it is applied to the model

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta y_{t-i} + \epsilon_t$$

Subsequently, the unit root test is performed, assuming the null hypothesis of $\gamma = 0$ [17]

Granger causality:

X Granger-causes Y if the addition of the lagged values of X to the prediction of Y improves the prediction of Y, where the prediction is made using a linear regression model of the form:

$$Y_t = \alpha + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{q} \gamma_i X_{t-i} + \epsilon_t$$

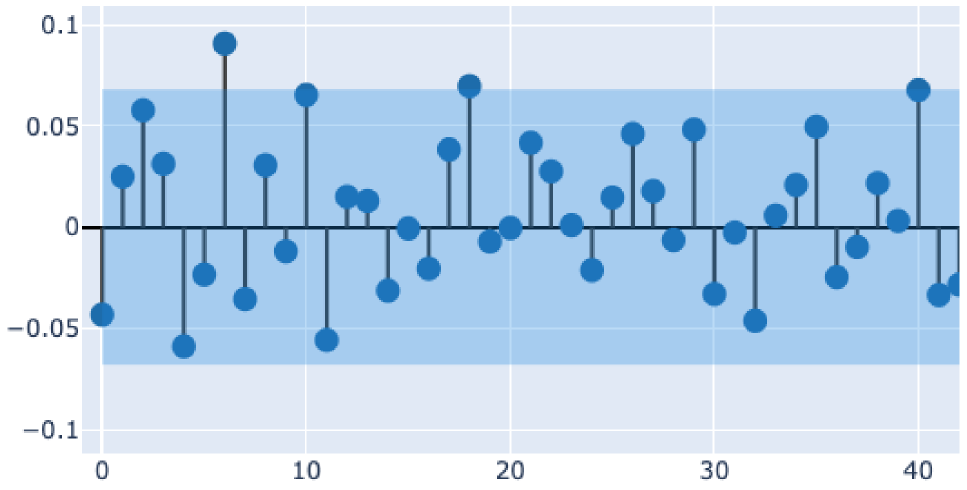
Granger causality is a statistical technique used to determine whether a time series X can predict future values of another time series Y. Specifically, X is said to Granger-cause Y if past values of X, along with past values of Y, are statistically significant predictors of future values of Y. This is typically established using t-tests and F-tests on lagged values of X and Y.[15]

To ensure the stability of the results, it is essential to verify that both time series are stationary before conducting a Granger causality analysis. We will employ the Augmented Dickey-Fuller (ADF) test to confirm stationarity, with the p-value being used as a threshold, where a value exceeding 0.05 indicates non-stationarity.

Upon examination, we discovered that the 10-year US Treasury bond yield has a p-value of 0.94, indicating non-stationarity, whereas the p-value for Bitcoin's return is 1.8e-17, indicating stationarity. To address the non-stationarity of the 10-year US Treasury bond yield, we performed first-order differencing and conducted another ADF test, which resulted in a p-value of 6.9e-13, indicating stationarity as well.

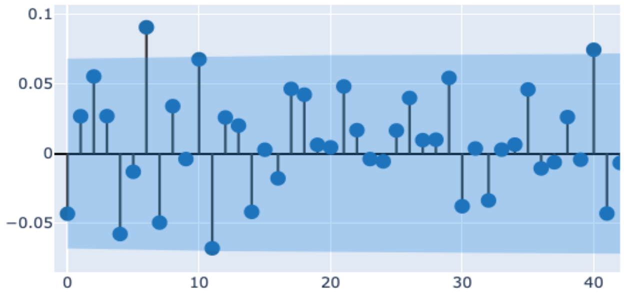
Next, we will examine the ACF and PACF:

Partial Autocorrelation (PACF): Return

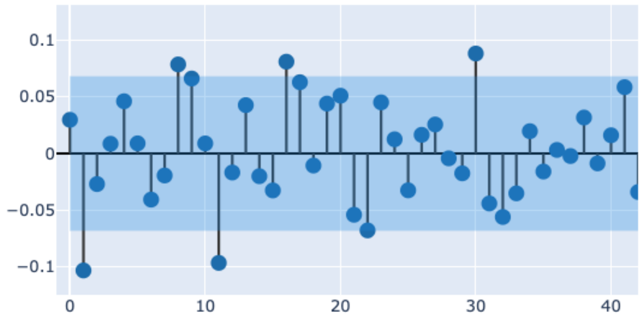


We can observe that there were no outliers in the earlier periods, but a persistent pattern of non-decay and frequent outliers outside of the range appeared in the later periods.

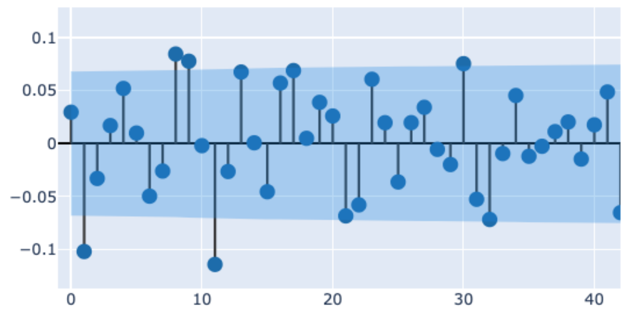
Autocorrelation (ACF): Return



Partial Autocorrelation (PACF): US_10_Year_Treasury_Yield_diff_1



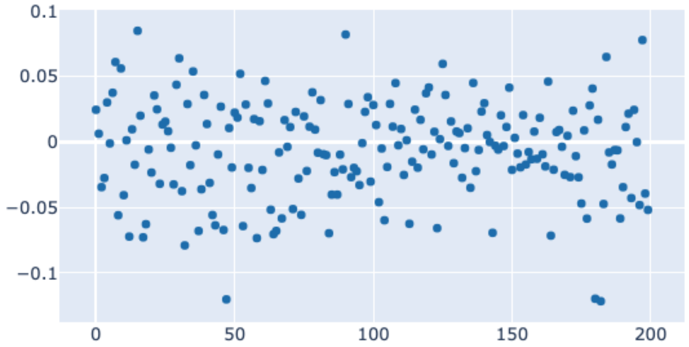
Autocorrelation (ACF): US_10_Year_Treasury_Yield_diff_1



The presence of this phenomenon in both ACF and PACF suggests that when analyzing the impact of the two US Treasury bond yields on Bitcoin, we must consider a large number of lags.

If we observe the correlation between US Treasury bond yields and Bitcoin we can see that there is no obvious relationship. However, we should do granger causality after we consider each lags of themself

Cocorrelation of Return and US_10_Year_Treasury_Yield_diff_1

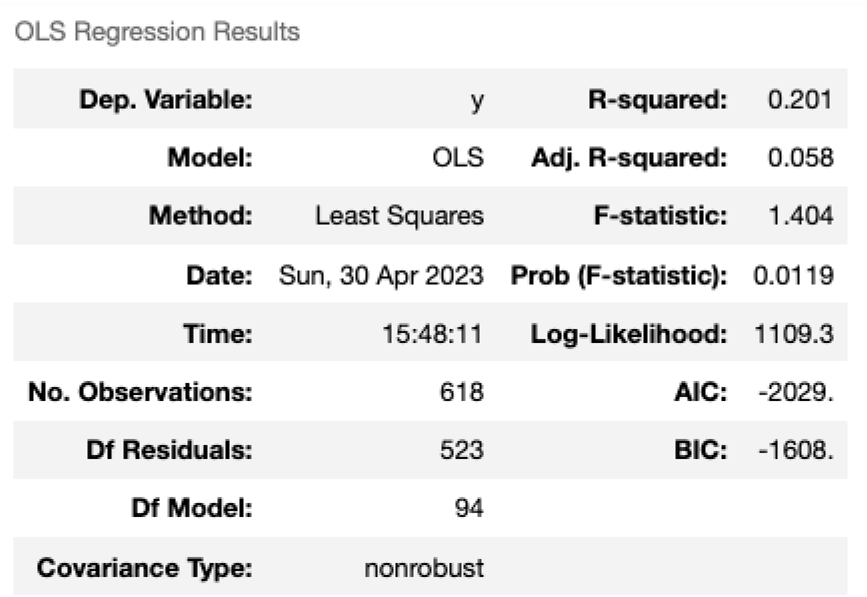


We will examine whether the inclusion of additional lags increases the likelihood of Granger causality through iterative observations. A model that satisfies all Granger causality tests with a sufficient number of lags will be deemed suitable for predicting and mitigating risks.

And in our code, we found that when considering max lags under 47 is the minimum lags to consider to pass all the test.
Here is the p-value of the test under max lags equal to 47 which are all under 0.05:

```
"ssr_ftest": 0.027184323824047157,
"ssr_chi2test": 0.0013994441876435178,
"lrtest": 0.004304714295519853,
"params_ftest": 0.027184323824048208
```

We can now construct our own AR model that incorporates exogenous variables and fit it to the training data. We used 80% of the period as training data(training samples: 665, testing samples: 166)



OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.201
Model:	OLS	**Adj. R-squared:**	0.058
Method:	Least Squares	**F-statistic:**	1.404
Date:	Sun, 30 Apr 2023	**Prob (F-statistic):**	0.0119
Time:	15:48:11	**Log-Likelihood:**	1109.3
No. Observations:	618	**AIC:**	-2029.
Df Residuals:	523	**BIC:**	-1608.
Df Model:	94		
Covariance Type:	nonrobust		

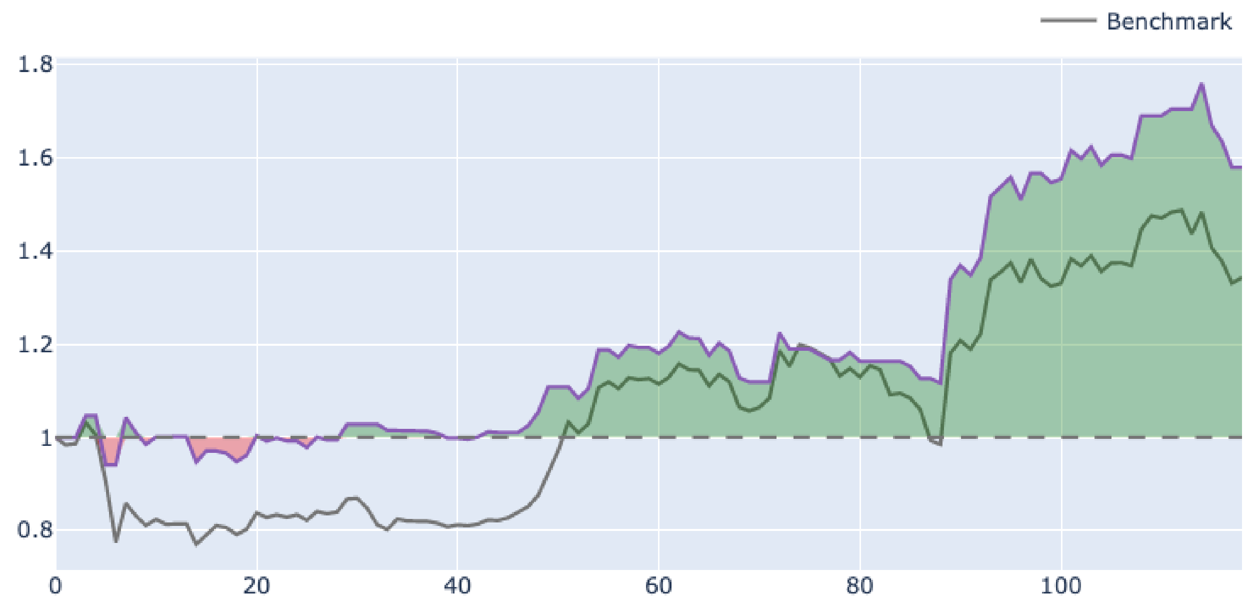
Here is a summary of our model. Its covariance type is "nonrobust" and it is normal, as we use multiple similar lags to construct the model, which can cause multicollinearity. However, interpretability of variables is not necessary for our purpose; we only require predictability.

We can see the statistic of training return is as below:

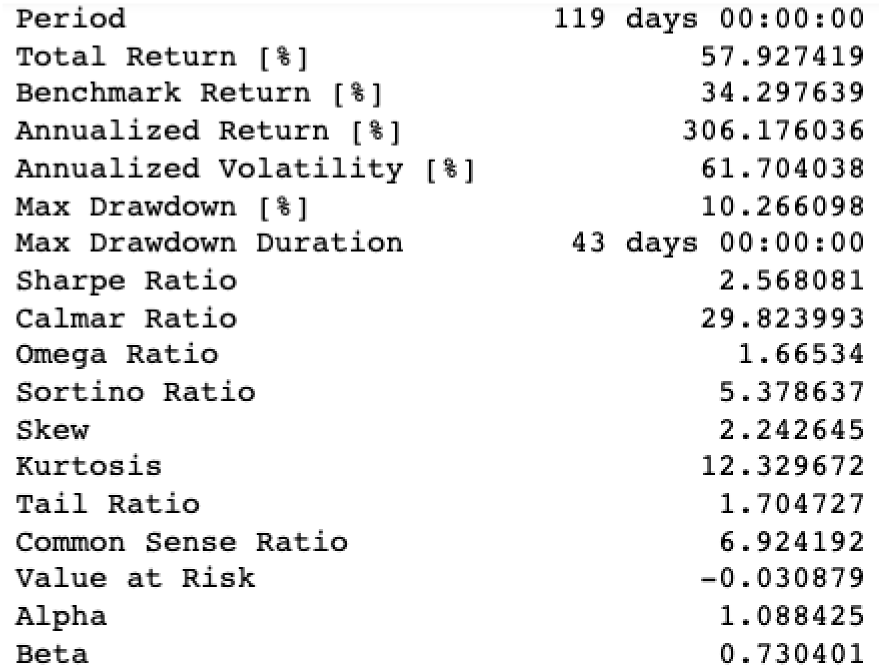
```
count    618.000000
mean       0.003379
std        0.020208
min       -0.077180
25%       -0.008667
50%        0.002950
75%        0.014720
max        0.096807
```

We can simply set the “risky” threshold as the 25% quantile of training return which is -0.008667. Our strategy is to retrieve the bitcoin as cash whenever we predict the return is lower than -0.00866.

Here is a comparison between the buy and hold strategy and our risk-avoidance strategy:



The cummulative return curve of our model is significantly better than the buy and hold strategy.



```
Period                        119 days 00:00:00
Total Return [%]                      57.927419
Benchmark Return [%]                  34.297639
Annualized Return [%]                306.176036
Annualized Volatility [%]             61.704038
Max Drawdown [%]                      10.266098
Max Drawdown Duration          43 days 00:00:00
Sharpe Ratio                           2.568081
Calmar Ratio                          29.823993
Omega Ratio                             1.66534
Sortino Ratio                          5.378637
Skew                                   2.242645
Kurtosis                              12.329672
Tail Ratio                             1.704727
Common Sense Ratio                     6.924192
Value at Risk                         -0.030879
Alpha                                  1.088425
Beta                                   0.730401
```

The total return is 57.9% which is 1.69 times of benchmark 34.3%.

Section 3:

The paper "The New Risk Management: The Good, the Bad, and the Ugly " written by Philip H. Dybvig and David A. Marshall is discussed below. The article was published in Economic Research Federal Reserve Bank of St. Louis in 1997. The authors of the paper discuss changes happening in risk management at the end of last century. They claim that new risk management techniques could be "bad", even "ugly" if those are not used properly.

The paper starts with discussing option-pricing tools that are used in new risk management. Authors claim that Black-Scholes model is mainly used for hedging strategies, such as insurance, by risk managers and it helps the market to be liquid. They show examples of copper manufacturer's hedging strategies using naive hedging using underlying assets and using the option to hedge for three economic conditions (well, good, and bad). They claim that naive approach increases the risk exposure while dynamic hedging using options reduces it. Our complex and evolving markets need more diversified "new risk management" because traditional and naive risk management do not work any more. The main ways of reducing volatility of the asset price are risk control, risk transfer and risk financing.

Then authors discuss benefits of new risk management techniques. At that time new risk management techniques were state of art methods to evaluate risk exposure of the firm. Because of the evolution of markets, naive risk hedging techniques did not work any more and new risk management methods developed using the following principles. Paying more attention to risk control by actively monitoring it and managing it, if possible transfer risk, and understanding all risks, not only market risk. New risk management used such tools as Value at Risk (VAR).

After that authors described the "bad" and the "ugly" consequences of new risk management techniques.

New hedging methods are affecting the accounting principles. For example, hedging using long-term options are not recorded to the accounting books until they are not realized gains or losses. These could lead to missguide investors or analysts who are only looking at reports of the firm. Therefore, companies should work to improve accounting standards which includes the all risk exposures of a firm. Authors claim that companies should also work to improve internal control and policies of the firm. Warnings of authors were justified with the accounting fraud examples of Enron company which happened in 2001, 4 years after the article was published.

Authors also discuss moral hazard and potential conflict of interest problems produced by new risk management methods. Usually if you have insurance for your car accident, then you try to take more risk compared to a case of no insurance. And this is a case for moral hazard. Same situation happens when a firm manager hedges their risk, they are eager to take more risk and it is a potentially dangerous situation for the company's future. Because models used in the hedging are not always 100 percent correct and blindly believing in the models could be disastrous as it was in the 2007-2008 crisis with the world and Long Term Management Management Company collapse in 1998. As Warren Buffet said “options are financial weapons of mass destruction ” , overusing options also have huge negative consequences. Some people claim that the 2007-2008 financial crisis was primarily caused by derivatives in the housing market. Therefore, misuse of new risk management techniques leads to the “ugly” consequences, such as the 2007-2008 crisis, as the authors of this paper warned.

I believe authors are missing the role of correlation in risk management, because of globalization companies are now more connected than before. Therefore, the assumption of independence and zero correlation is not valid anymore. Including correlation into risk management techniques could save as from the 2007-2008 crisis.

REFERENCES:

1. Harsch Katara and Dheeraj Vaidya, Quartile Formula, WallStreetMojo, Quartile Formula | How to Calculate Quartile in Statistics | Example (wallstreetmojo.com)

2. Pritha Bhandari, How to find the Median, Scribbr, 02/19/2023, How to Find the Median | Definition, Examples & Calculator (scribbr.com)

3. Median - Formula, Meaning, Example | How to Find Median? (cuemath.com)

4. Quantile: Definition and How to Find Them in Easy Steps - Statistics How To

5. Avijeet Biswal, SimpleLearn, 04/03/2023, Percentile in Statistics: Overview & How to Calculate | Simplilearn

6. Jim Frost, Statistics by Jim, Interquartile Range (IQR): How to Find and Use It - Statistics By Jim

7. Khan Academy, Calculating standard deviation step by step, Standard deviation: calculating step by step (article) | Khan Academy

8. Adam Hayes, Somer Anderson, Amanda Bellucco-Chatham, Co-efficient of Variation Meaning and How to Use It, Investopedia 09/16/2022, Co-efficient of Variation Meaning and How to Use It (investopedia.com)

9. Khan Academy; Mean, median and mode; Mean, median, and mode review (article) | Khan Academy

10. Statistics How To, Median Absolute Deviation, Median Absolute Deviation - Statistics How To

11. Ashish Kumar Srivastav, Dheeraj Vaidya, Skewness Formula, WallStreetMojo, Skewness Formula | How to Calculate Skewness? (with Examples) (wallstreetmojo.com)

12. Diksha Keni, Dheeraj Vaidya, Population Variance Formula, WallStreetMojo, Population Variance Formula | Step by Step Calculation | Examples (wallstreetmojo.com)

13. Dybvig, Philip H., and William J. Marshall. “The New Risk Management: The Good, the Bad, and the Ugly.” *Economic Research Federal Reserve Bank of St. Louis*, vol. 79, no. 6, 1997, pp.9-21,https://research.stlouisfed.org/publications/review/1997/11/01/the-new-risk-management-the-good-the-bad-and-the-ugly. Accessed 30 April 2023.

14. Partial Auto Correlation Function Formula

15. Granger_causality

16. Dickey–Fuller test

17. Augmented Dickey–Fuller test

18. Leslie, J. R., Stephens, M. A., and Fotopoulos, S. (1986). Asymptotic distribution of the Shapiro–Wilk W for testing for normality. Ann.Statist. 14, 1497–1506