جامعة سيدي محمد بن عبد الله بفاس
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⵎⵎⴷ ⴱⵏ ⵄⴱⴷⵍⵍⴰⵀ ⵏ ⴼⴰⵙ
UNIVERSITÉ SIDI MOHAMED BEN ABDELLAH DE FES

كلية العلوم ظهر المهراز فاس
ⵜⴰⵙⵍⵉⵡⵜ ⵏ ⵜⵓⵙⵙⵏⴰ ⴹⵀⵔ ⵍⵎⵀⵔⴰⵣ ⴼⴰⵙ
FACULTE DES SCIENCES DHAR EL MAHRAZ FES

Computer Science department

BITCOIN PRICES PREDICTION

Prepared by :

- BAKI AMINE
- AZAMI YOUSSEF

جامعة سيدي محمد بن عبد الله بفاس

ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⵎⵎⴰⴷ ⵙⵏ ⵄⴱⴷⵍⵍⴰⵀ ⵏ ⴼⴰⵙ

UNIVERSITÉ SIDI MOHAMED BEN ABDELLAH DE FES

Computer Science department

ACKNOWLEDGEMENT

First of all, we would like to express our gratitude to Mr. Abdelkamel ALJ, our instructor, as well as to all of the other teachers who allowed us to tackle a significant topic (wealth and gender/race disparity) in our project.

TABLE OF CONTENTS

MODELISATION AND SIMULATION

Modeling and simulation (M&S) is the use of a physical or logical representation of a given system to generate data and help determine decisions or make predictions about the system. M&S is widely used in the social and physical sciences, engineering, manufacturing and product development, among many other areas

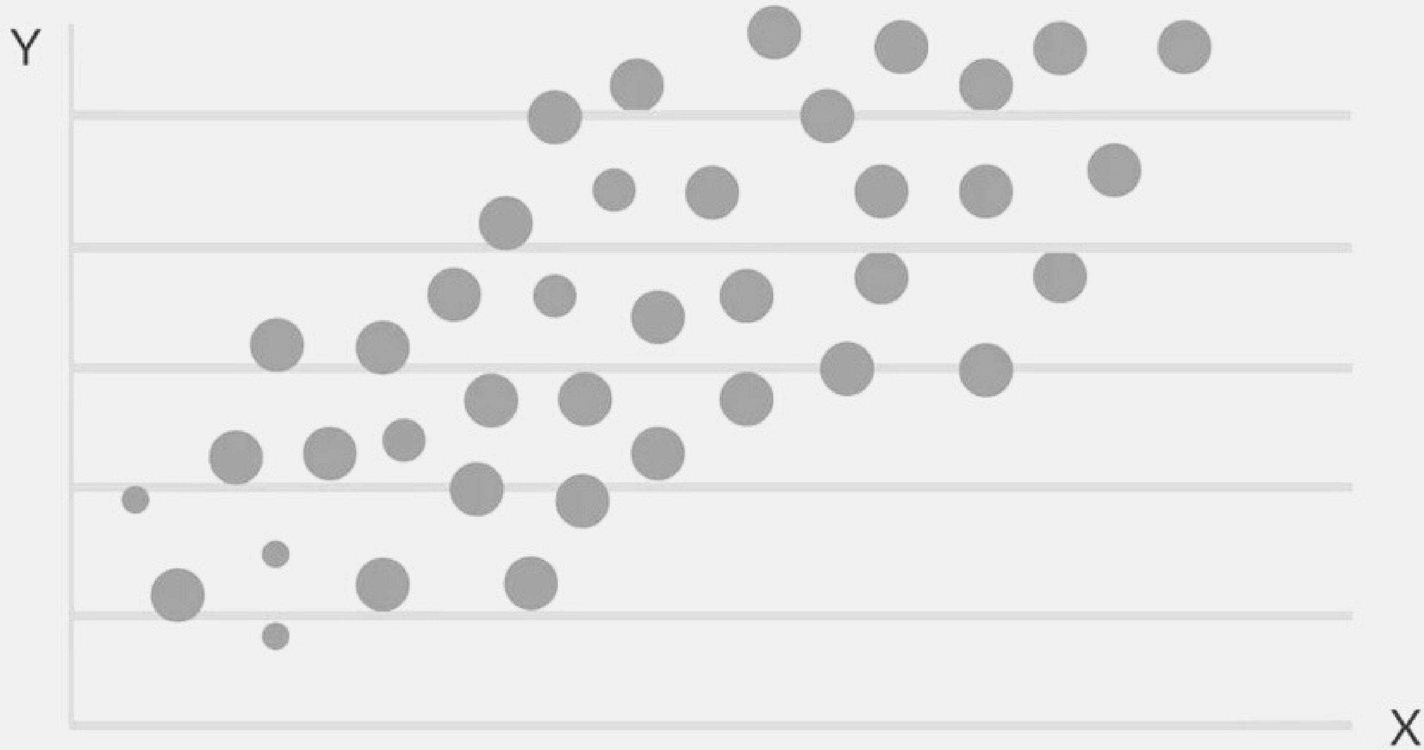
Applications of modeling and simulation include

- Creating models of weather systems, simulating behavior based on available data to generate predictive information for forecasts. A hurricane forecast model, for example, is designed to predict a given storm's track and intensity, as well as related events such as storm surges.
- Creating a program to model a social situation and observing the behavior of individuals in the simulation when the program runs. Social simulations can be used to yield predictive data about how things happen in real-world environments, such as how social norms develop.

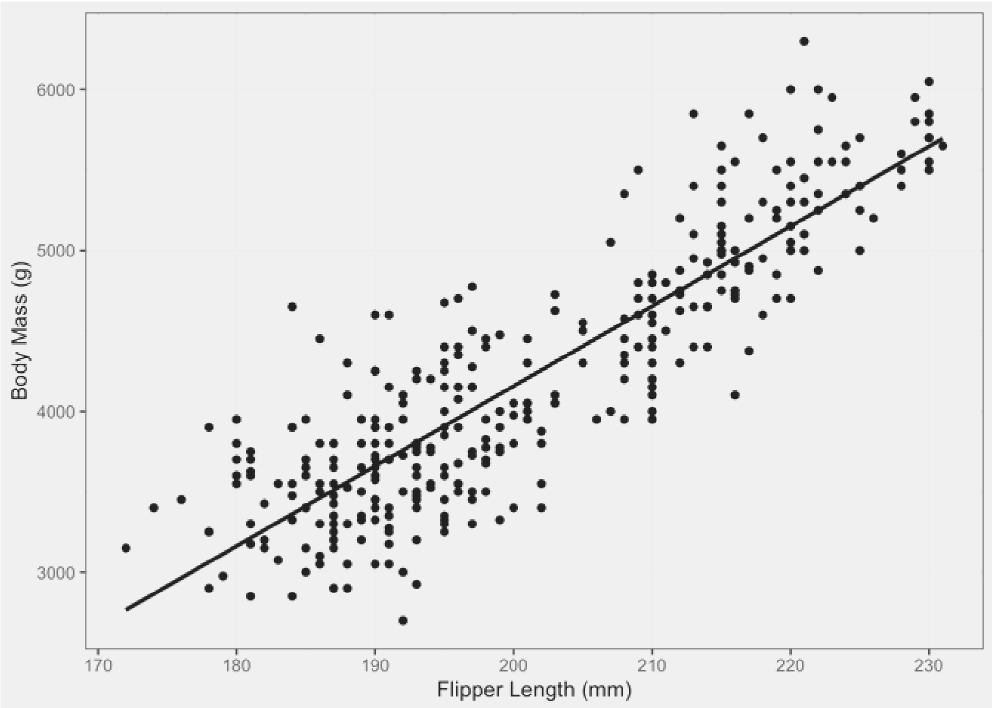
INTRODUCTION

Regression methods are used in different industries to understand which variables impact a given topic of interest.

For instance, Economists can use them to analyze the relationship between consumer spending and Gross Domestic Product (GDP) growth. Public health officials might want to understand the costs of individuals based on their historical information. In both cases, the focus is not on predicting individual scenarios but on getting an overview of the overall relationship.



SIMPLE LINEAR REGRESSION



A simple linear regression aims to model the relationship between the magnitude of a single independent variable X and a dependent variable Y by trying to estimate exactly how much Y will change when X changes by a certain amount.

- The independent variable X, also called the predictor, is the variable used to make the prediction.
- The dependent variable Y, also known as the response, is the one we are trying to predict.

The “linear” aspect of linear regression is that we are trying to predict Y from X using the following “linear” equation.

$$Y = b0 + b1X$$

- b0 is the intercept of the regression line, corresponding to the predicted value when X is null.
- b1 is the slope of the regression line.

MULTIPLE LINEAR REGRESSION

Multiple linear regression model with k variables

$$y^i = \beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \ldots + \beta_k x_k^i + \epsilon^i$$

y^i = dependent variable for the i^{th} observation
x_j^i = j^{th} independent variable for the i^{th} observation
ϵ^i = error term for the i^{th} observation
β_0 = intercept coefficient
β_j = regression coefficient for the j^{th} independent variable

This is the use of linear regression with multiple variables, and the equation is:

Y = b0 + b1X1 + b2X2 + b3X3 + … + bnXn + e

- Y and b0 are the same as in the simple linear regression model.
- b1X1 represents the regression coefficient (b1) on the first independent variable (X1). The same analysis applies to all the remaining regression coefficients and variables.
- e is the model error (residuals), which defines how much variation is introduced in the model when estimating Y.

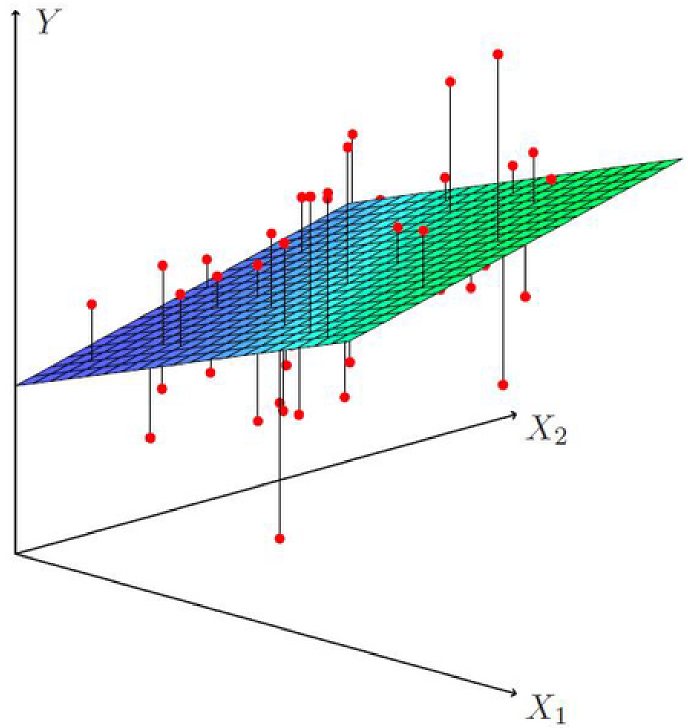
We might not always get a straight line for a multiple regression case. However, we can control the shape of the line by fitting a more appropriate model.

MODEL

Making judgments regarding the nature of the association between the dependent variable and the independent variables is known as inference in multiple regression. Based on the values of the independent variables, we may utilize the regression analysis results to predict the values of the dependent variable.

ESTIMATION

- *Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable*
- *however the least squares method is used to find the best-fitting line for the observed data. The estimated least squares regression equation has the minimum sum of squared errors, or deviations, between the fitted line and the observations.*
- *but befor we go any further we should know establish some assupptions on our model.*



Multiple regression model with two variables

12-1 Multiple Linear Regression Models

12-1.2 Least Squares Estimation of the Parameters

- The least squares function is given by

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

- The least squares estimates must satisfy

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

and

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \ldots, k$$

This is the matrix version of the calculation of the variables using MCO. We usually use it to simplify the work, as if we work with multiple variables with derivation, this will result in very complicated calculations, and this is the result of it .

Regression in Matrix Form

Assume a model using n observations, k parameters, and $k-1$, X_i (independent) variables.

$$y = Xb + e$$
$$\hat{y} = Xb$$
$$b = (X'X)^{-1}X'y$$

- $y = n * 1$ column vector of observations of the DV, Y
- $\hat{y} = n * 1$ column vector of predicted Y values
- $X = n * k$ matrix of observations of the IVs; first column 1s
- $b = k * 1$ column vector of regression coefficients; first row is A
- $e = n * 1$ column vector of n residual values

ASSUPPTIONS ON THE MODEL

- Multiple linear regression makes all of the same assumptions as simple linear regression:
- ***<u>Homogeneity of variance (homoscedasticity):</u>*** the size of the error in our prediction doesn't change significantly across the values of the independent variable.
- ***<u>Independence of observations:</u>*** the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among variables.
- In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated ($r2 > \sim 0.6$), then only one of them should be used in the regression model.
- ***<u>Normality</u>***: The data follows a normal distribution.
- ***<u>Linearity</u>***: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

INFERENCE

The model can be further generalized to any number of explanatory variables. Note that the slope parameters are termed partial regression coefficients because they measure the change in Y per unit change in the respective X whilst holding all other X variables constant. If the X variables are measured in different units, it may be preferable to use standardized coefficients that are independent of the units the variables are measured in $Y = \beta 0 + \beta 1X1 + \beta 2X2 + \varepsilon$

INFERENCE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio
Regression	k	$SS_{reg} = \sum(\hat{Y}_i - \bar{Y})^2$	$MS_{reg} = SS_{reg} / k$	MS_{reg}/MS_{res}
Residual	n-k-1	$SS_{res} = \sum(Y_i - \hat{Y}_i)^2$	$MS_{res} = SS_{res} / n\text{-}k\text{-}1$	
Total	n-1	$SS_{tot} = \sum(Y_i - \bar{Y})^2$		

- For k = 1 the table above is reduced to simple linear regression
- The F-ratio tests the hypothesis that all coefficients a0 .. an of the independents variables are zero (null hypothesis). The F-ratio is distributed according to an F distribution with k and n-k-1 degrees of freedom. Also, the F value is related to the goodness of fit, r2, through the following equation:
- The residual sum of squares SSres is an estimate of the variability along the regression line. SSres can be used to find the estimated standard errors of the individual regression coefficients ai. The estimated standard error follows a t-distribution with n-k-1 degrees of freedom. The confidence interval for the individual coefficients is given by +/- t(**α**/2, n-k-1)s(ai).
- If two variables xi, and xj, are highly correlated, the regression coefficients are difficult to estimate, and their actual numeric values probably do not reflect real dependencies.

COEFFICIENT OF MULTIPLE DETERMINATION

- *The coefficient of multiple determination, denoted R2�2, in multiple regression is similar to the coefficient of determination in simple linear regression, except in multiple regression there is more than one independent variable. The coefficient of multiple determination is the proportion of variation in the dependent variable that can be explained by the multiple regression model based on the independent variables.*

$$Coefficient\ of\ Deternination \rightarrow \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$Sum\ of\ Squares\ Total \rightarrow \quad SST = \sum (y - \bar{y})^2$$

$$Sum\ of\ Squares\ Regression \rightarrow \quad SSR = \sum (y' - \bar{y}')^2$$

$$Sum\ of\ Squares\ Error \rightarrow \quad SSE = \sum (y - y')^2$$

INTERPRETATION OF THE COEFFICIENT OF DETERMINATION (R^2)

- The most common interpretation of the coefficient of determination is how well the regression model fits the observed data. For example, a coefficient of determination of 60% shows that 60% of the data fit the regression model. Generally, a higher coefficient indicates a better fit for the model.
- However, it is not always the case that a high r-squared is good for the regression model. The quality of the coefficient depends on several factors, including the units of measure of the variables, the nature of the variables employed in the model, and the applied data transformation. Thus, sometimes, a high coefficient can indicate issues with the regression model.

DATASET

The dataset used is the Cryptocurrency Prices Dataset

This dataset contains the historical prices and volume of 4 cryptocurrencies from November 9, 2017 to August 27, 2022.

- BTC - Bitcoin
- BNB - Binance coin
- ETH - Ethereum
- USDT - Tether

PROBLEM

Bitcoin (BTC) is a cryptocurrency, a virtual currency designed to act as money and a form of payment outside the control of any one person, group, or entity, thus removing the need for third-party involvement in financial transactions. It is rewarded to blockchain miners for the work done to verify transactions and can be purchased on several exchanges.

Bitcoin was introduced to the public in 2009 by an anonymous developer or group of developers using the name Satoshi Nakamoto.1

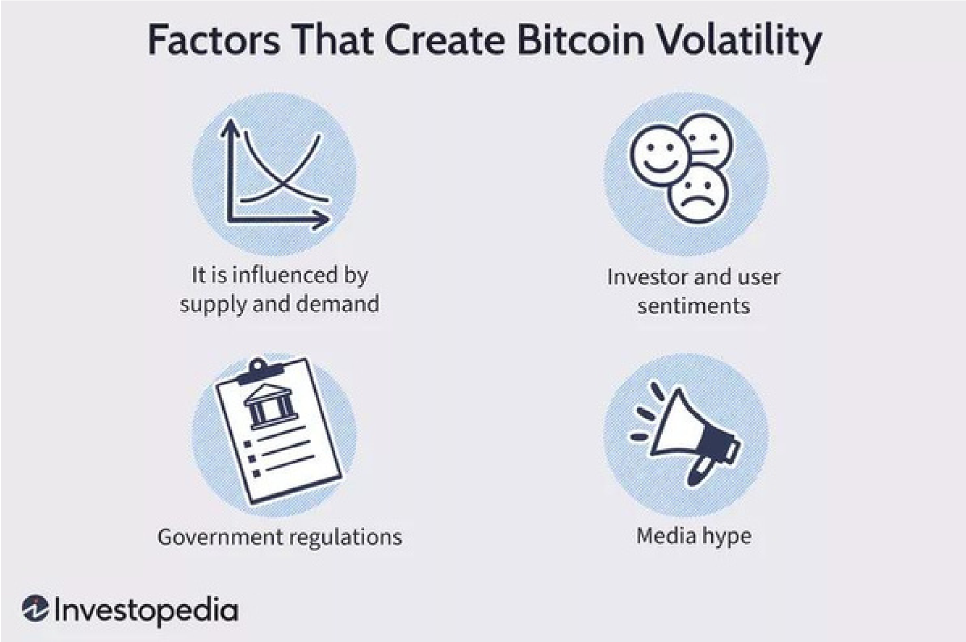
It has since become the most well-known cryptocurrency in the world. Its popularity has inspired the development of many other cryptocurrencies. These competitors either attempt to replace it as a payment system or are used as utility or security tokens in other blockchains and emerging financial technologies.

Bitcoin, made publicly available in 2009, began its rise to popularity around 2010 when the price for one token rose from fractions of a dollar to $0.09. Since then, its price has increased by tens of thousands of dollars—sometimes rising or falling thousands of dollars within days.1

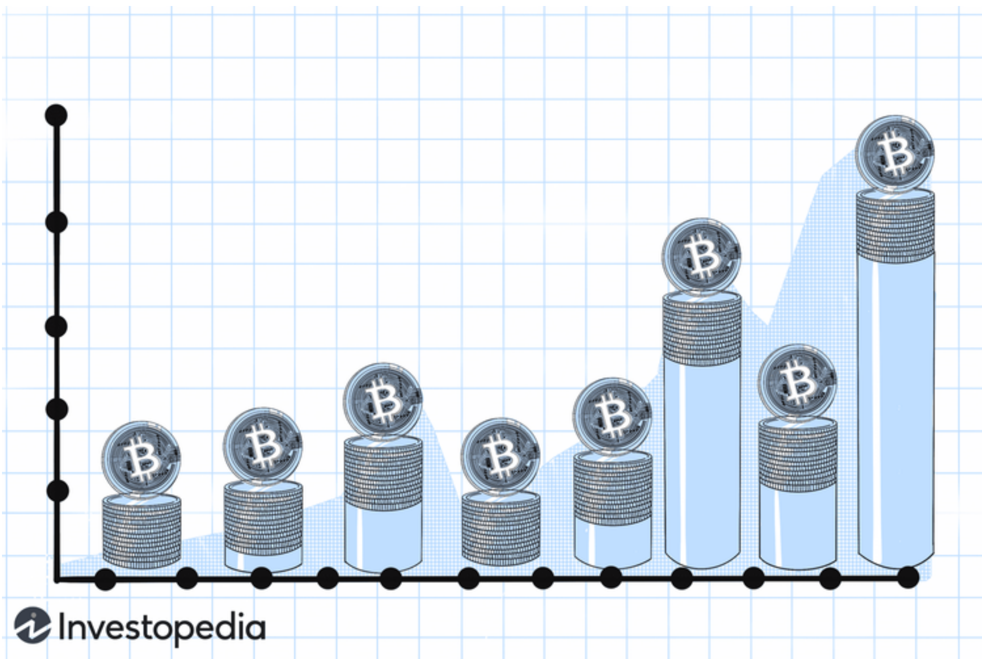
There are several reasons why Bitcoin has such a volatile price history. Understanding the factors that influence its market price can help you decide whether to invest in it, trade it, or continue watching its developments.

KEY TAKEAWAYS

- Like most commodities, assets, investments, or other products, Bitcoin's price depends heavily on supply and demand.
- As an asset adopted quickly by investors and traders, speculation about price movements plays a critical part in Bitcoin's value at any given moment.
- Media outlets, influencers, opinionated industry moguls, and well-known cryptocurrency fans create investor concerns, leading to price fluctuations.



We will just look at the impact of supply and demand. mainly using the price of another crypto currencys, which is the number 1 cause of price variance, we may investigate what the biggest component is. Many sophisticated algorithms use this element as their primary indicator when forecasting future market values or prices.



OVERVIEW

Close..BTC.	Volume..BTC.	Close..ETH.	Volume..ETH.	Close..USDT.	Volume..USDT.	Close..BNB.	Volume..BNB.
67566.83	41125608330	4812.087	19290896267	1.000443	82548510715	654.3150	2828112534
66971.83	42357991721	4735.069	20834172627	1.000202	93002275939	635.1906	2198989754
65992.84	40788955582	4155.992	20338319988	0.999940	70187915900	501.0203	1890415594
65466.84	25122092191	4626.359	12172962219	1.000367	54429579952	650.9181	2101401990
64995.23	48730828378	4636.174	22748160545	1.000097	113809197171	615.2781	3653998344
64949.96	35880633236	4730.384	17933201129	1.000644	83970826408	629.8923	2341652507
64469.53	30474228777	4651.460	14457436261	1.000848	65797472687	650.1041	2106170805
64261.99	40471196346	3877.651	15998757133	1.000117	62387883982	488.1489	1844483840
64155.94	36084893887	4667.115	18316060208	1.000114	82883678293	626.6425	2209163398
63557.87	30558763548	4557.504	16275851299	1.001502	64113625426	633.0486	2146877925
63503.46	69983454362	2299.188	29456642939	0.999275	137132738350	549.5865	9433830832
63326.99	24726754302	4620.555	13541376033	1.001517	55905241687	650.4540	3017167868
63314.01	60954381579	2519.116	32325606817	1.000877	134206127794	542.6321	4608061831
63226.40	37746665647	4584.799	20794448222	1.000815	149878312204	554.4476	2401108112

VARIABLES:

Close (BTC)

Adjacent close price of Bitcoin Coin on that particular date (in USD)

Volume (BTC)

Volume of BTC on that particular date

Close (ETH)

Adjacent close price of Ethereum Coin on that particular date (in USD)

Volume (ETH)

Volume of ETH on that particular date

Close (USDT)

Adjacent close price of Tether Coin on that particular date (in USD)

Volume (USDT)

Volume of USDT on that particular date

Close (BNB)

Adjacent close price of Binance Coin on that particular date (in USD)

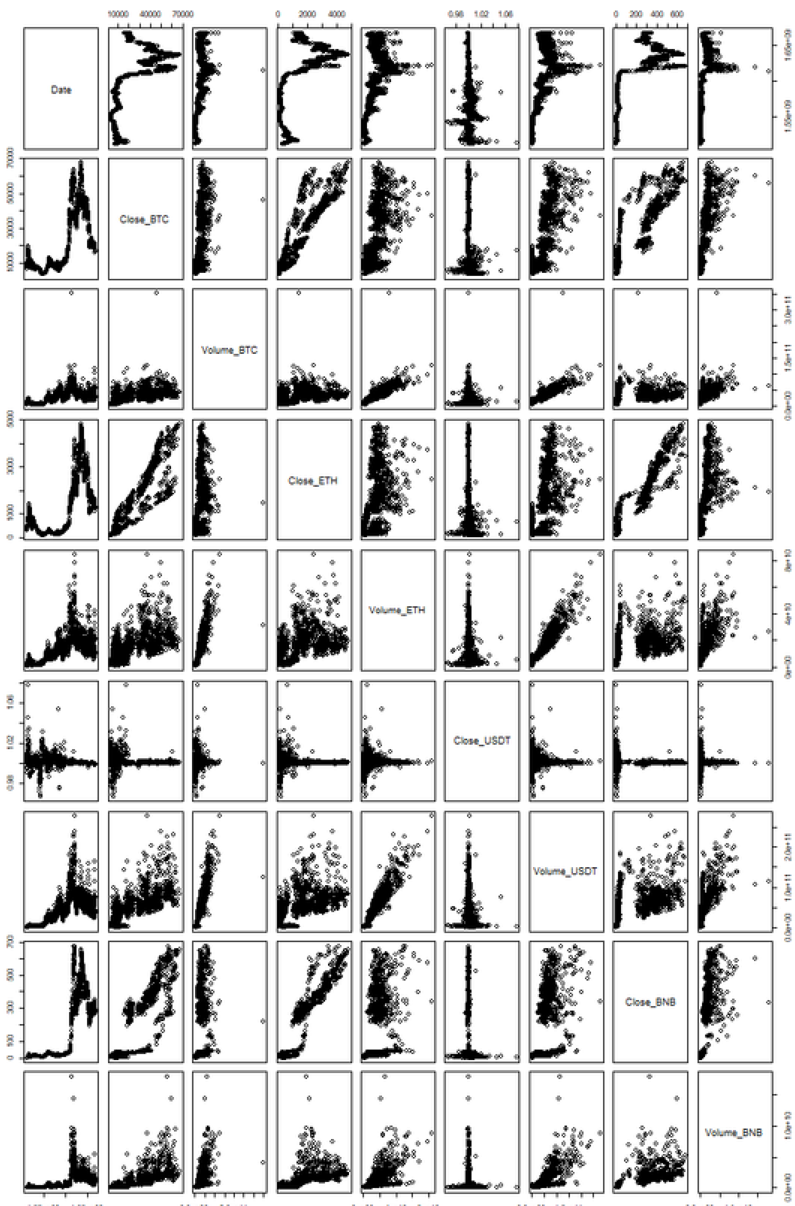
Volume (BNB)sort

Volume of BNB on that particular date

DATA PREPROCESSING

ABERRANT VALUES

After doing correlation analysis, we can see that there are some values in the dataset that are a little bit out of the ordinary. which made us think about aberrant values that could demolish our model and give us very bad results. That's why we introduced the boxplot method, which gave us the opportunity to extract those points.

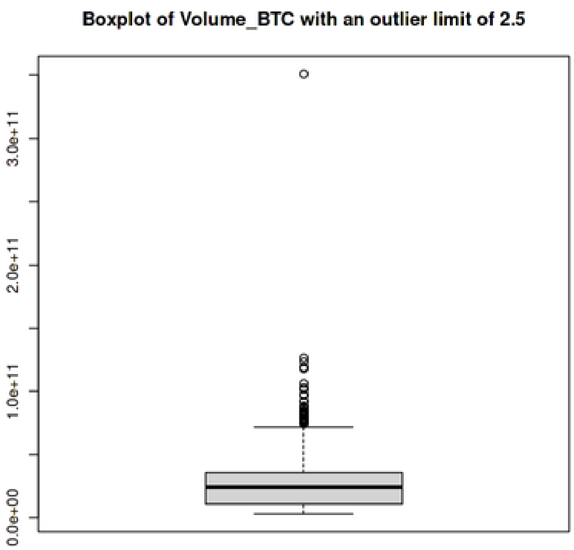


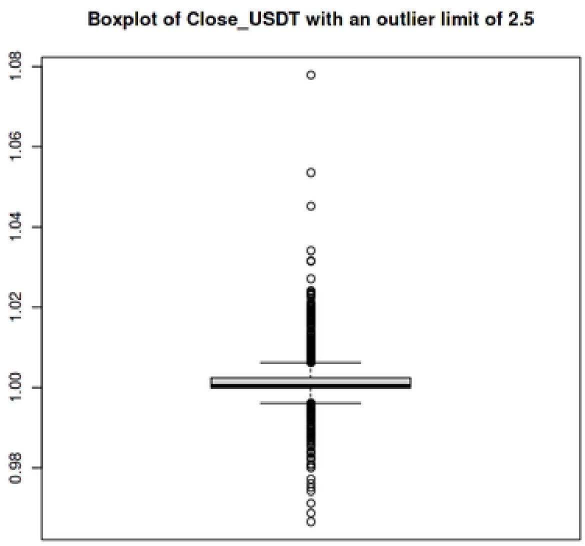
scatter plots of the original data

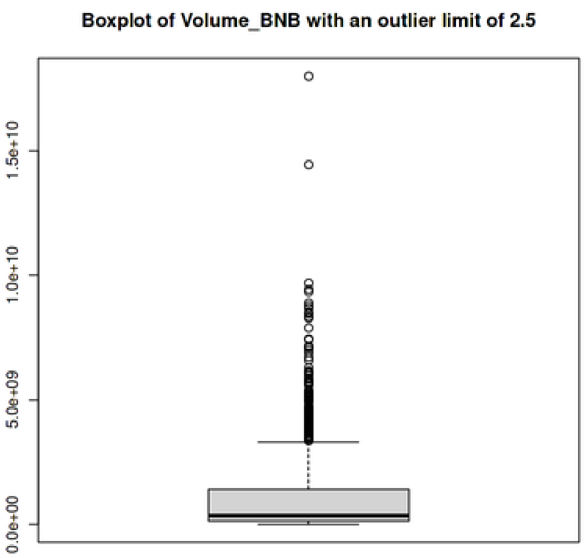
BOXPLOTS

A boxplot is a graph that gives us a good indication of how the values in the data are spread out.Box plots provide some indication of the data's symmetry and skew-n

As you can see this Boxplot give us the exact points that may cause a problem.

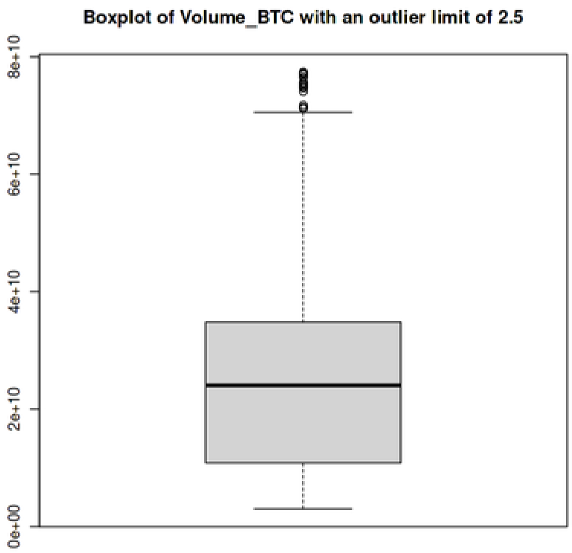


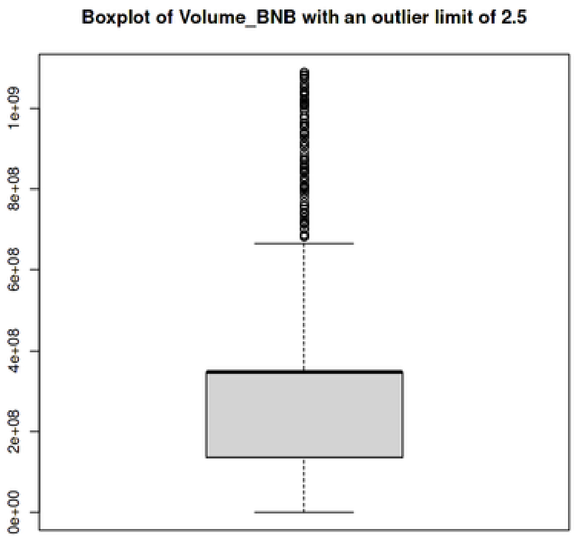


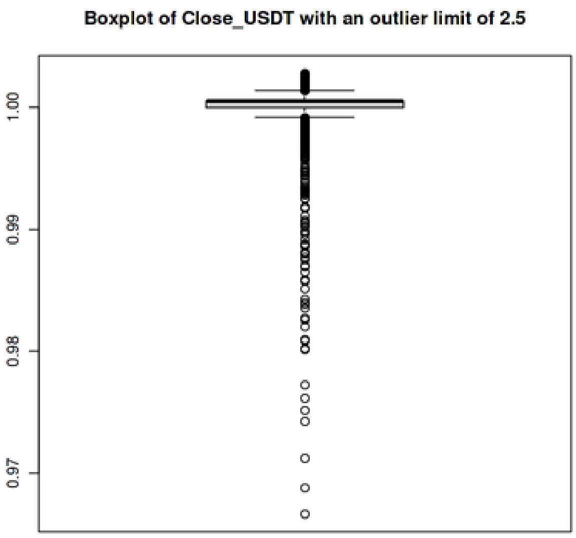


ELIMINATING ABERRANT VALUES

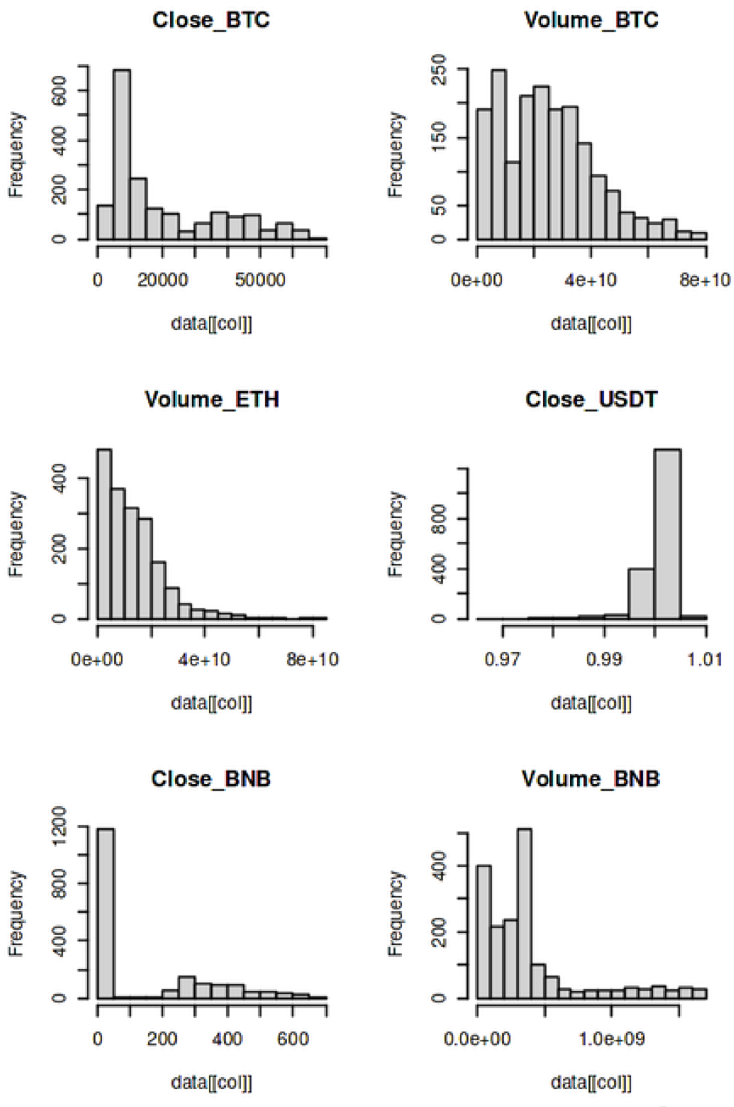
After replacing aberrant values of each independent variable with its the median we obtained the following Boxplots.







DATA DISTRIBUTION ANALYSIS

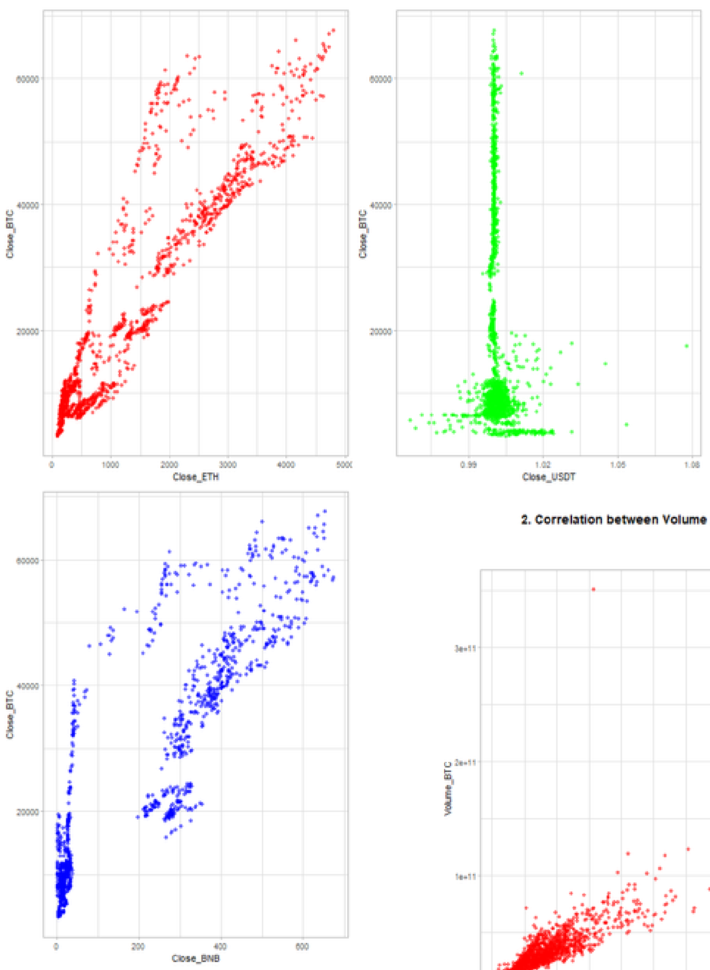


Histogram of columns

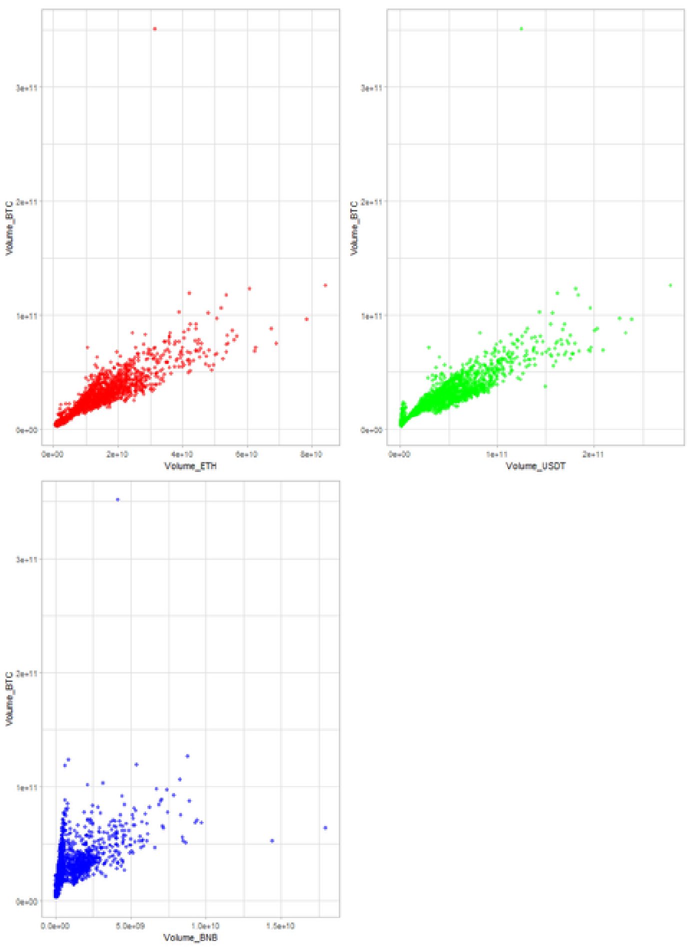
CORRELATION ANALYSIS

As it is represented, there is no direct relation between any of the variables, which leaves us with no shanged or eliminated variables, and that made us think about a stepwise method to eliminate some variables and make our model more meaningful.

1. Correlation between Close (BTC) and Close (ETH) / Close (USDT) / Close(BNB)



2. Correlation between Volume (BTC) and Volume (ETH) / Volume (USDT) / Volume(BNB)



VARIABLE SELECTION

STEPWISE METHODE

Selecting a method allows you to specify how independent variables are entered into the analysis. Using different methods, you can construct various regression models from the same group of variables.

Variable selection procedure in which all variables in a block are introduced in a single operation.
Stepwise . At each step, the program captures the independent variable excluded from the equation with the lowest probability of F, if that probability is sufficiently low. Variables already included in the regression equation are eliminated if their probability of F becomes too high. The process stops when no variable can be introduced or deleted.

```
Residuals:
   Min     1Q Median     3Q    Max
 -8979  -3283   -784   1151  30101

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.133e+05  1.448e+05   1.473    0.144
crypto3.Close..BNB.    2.260e+00  1.407e+01   0.161    0.873
crypto3.Close..USDT. -2.068e+05  1.444e+05  -1.432    0.156
crypto3.Close..ETH.    1.193e+01  2.025e+00   5.891 6.55e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6606 on 90 degrees of freedom
  (1706 observations effacées parce que manquantes)
Multiple R-squared:  0.8656,	Adjusted R-squared:  0.8611
F-statistic: 193.2 on 3 and 90 DF,  p-value: < 2.2e-16
```

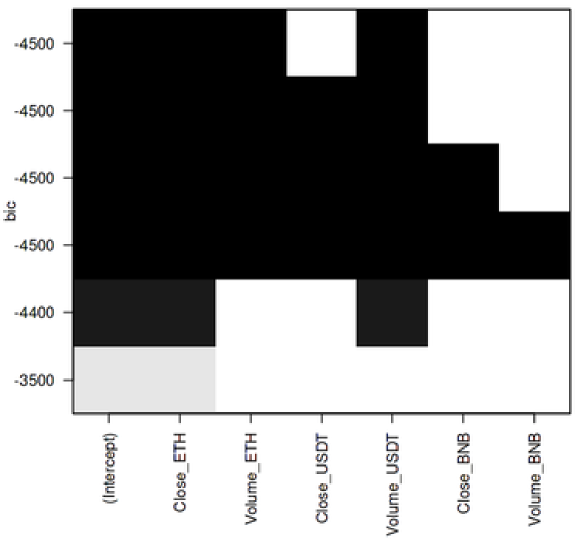
BEFOR USING STEPWISE METHODE

```
Residuals:
       Min         1Q     Median         3Q        Max
-1.320e+10 -2.998e+09 -1.798e+08  2.550e+09  2.734e+10

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.243e+09  1.019e+09   5.148 1.53e-06 ***
crypto3.Volume..BNB.  -4.839e+00  7.799e-01  -6.205 1.64e-08 ***
crypto3.Volume..USDT.  5.031e-01  6.287e-02   8.003 4.03e-12 ***
crypto3.Volume..ETH.   3.430e-01  2.217e-01   1.547    0.125
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.959e+09 on 90 degrees of freedom
  (1706 observations effacées parce que manquantes)
Multiple R-squared:  0.9031,	Adjusted R-squared:  0.8999
F-statistic: 279.6 on 3 and 90 DF,  p-value: < 2.2e-16
```

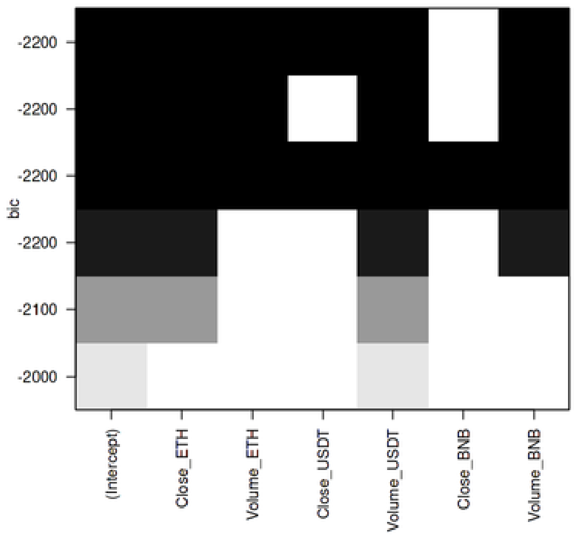
BEST SUBSETS SELECTION

As it is represented, we should conserve
Close_ETH, Volume_ETH, Volume_USTD as variables for Close_BTC

As it is represented, we should conserve
Close_ETH, Volume_ETH, Volume_USTD, Volume_BNB as variables for Volume_BTC



for Close_BTC



for Volume_BTC

TESTING THE MODULES

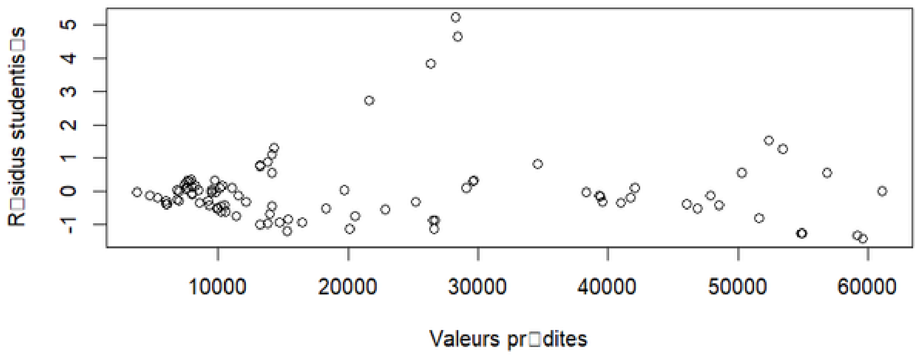
AFTER USING STEPWISE METHODE

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            7.520e+09  8.308e+08   9.051 2.69e-14 ***
crypto3.Volume..BNB.  -2.867e+00  7.217e-01  -3.973 0.000143 ***
crypto3.Volume..USDT.  6.102e-01  2.300e-02  26.526  < 2e-16 ***
crypto3.Close..ETH.   -3.386e+06  5.416e+05  -6.253 1.32e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.041e+09 on 90 degrees of freedom
  (1706 observations effacées parce que manquantes)
Multiple R-squared:  0.9306,	Adjusted R-squared:  0.9283 
F-statistic: 402.6 on 3 and 90 DF,  p-value: < 2.2e-16

> |
```

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            5.154e+03  9.032e+02   5.706 1.46e-07 ***
crypto3.Close..ETH.    1.011e+01  5.115e-01  19.771  < 2e-16 ***
crypto3.Volume..USDT.  3.490e-07  5.590e-08   6.243 1.38e-08 ***
crypto3.Volume..ETH.  -8.605e-07  1.938e-07  -4.441 2.54e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5160 on 90 degrees of freedom
  (1706 observations effacées parce que manquantes)
Multiple R-squared:  0.918,	Adjusted R-squared:  0.9152 
F-statistic: 335.8 on 3 and 90 DF,  p-value: < 2.2e-16

> |
```

As you can see, we have clearly eliminated the write variables. because the R squared value is now closer to one in each of the two models that we have established, and the same thing applies to the F value, which has become larger.

RESIDUAL ANALYSIS



The residuals are obtained by the residuals function, however the residuals obtained are not of the same variance (heteroskedastic). We therefore use studentized residuals residuals, which have the same variance.

The residuals are obtained by the residuals function, however the residuals obtained are not of the same variance (heteroskedastic). We therefore use studentized residuals residuals, which have the same variance.

PREDICTION

```
> colnames(xnew) <- c("crypto3.Volume..BTC.","crypto3.Volume..BNB.",'crypto3.Volume..USD
T.', 'crypto3.Close..ETH.')
> xnew <- as.data.frame(xnew)
> predict(reg.multiple_volume_updated,xnew,interval="pred")
         fit         lwr         upr
1 7282771453 -2863899278 17429442185
> 
```

The residuals are obtained by the residuals function, however the residuals obtained are not of the same variance (heteroskedastic). We therefore use studentized residuals residuals, which have the same variance.

The residuals are obtained by the residuals function, however the residuals obtained are not of the same variance (heteroskedastic). We therefore use studentized residuals residuals, which have the same variance.

CONCLUSION

To conclude we can say that Multiple linear regression is used to evaluate predictors for a continuously distributed outcome variable. The procedure calculates coefficients for each of the independent variables (predictors) that best agree with the observed data in the sample.

Multiple variable regression enables you to:

- Control for confounding: each of the coefficients for the independent variables is adjusted for confounding by all other variables in the model.
- Make predictions: Predicted values from the model can be interpreted either as estimated means (for subjects with a particular profile) or as predictions for individuals.
- Identify relative importance of the independent variables in the model outcome

SOURCE LINKS

here is a summary of all the sources that we have used for our project

N° 01 - **definitions**

https://www.investopedia.com/terms/r/regression.asp
https://corporatefinanceinstitute.com/resources/data-science/coefficient-of-determination/

N° 2- **Matrix calculation**

https://bookdown.org/ripberjt/qrmbook/introduction-to-multiple-regression.html

N° 3- **Dataset**

https://bookdown.org/ripberjt/qrmbook/introduction-to-multiple-regression.html

N° 4- **images**

All the images used are from articles and can be clicked in pdf version