Analysis on Cryptocurrencies data, Financial indexes and Commodities

FINANCIAL DATA SCIENCE
A.A. 2020/2021

Bill Mono

Objective and Dataset

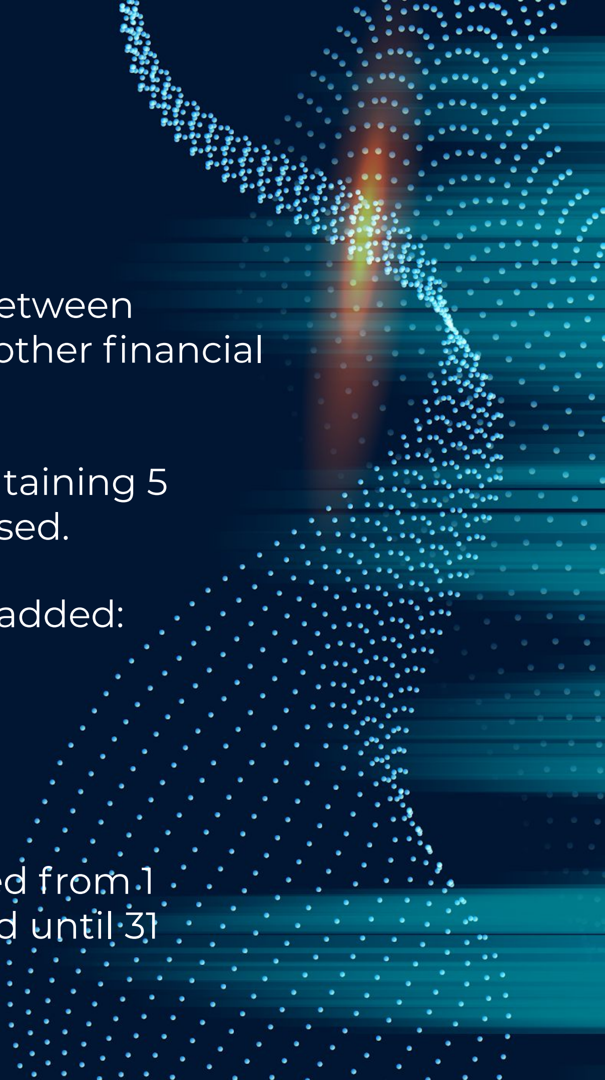
The goal of this analysis is to understand the relationship between cryptocurrencies, and the relationship that they have with other financial instruments.

In order to carry out this analysis, the proposed dataset containing 5 different cryptocurrencies (BTC, ETH, XRP, LTC, XLM) was used.

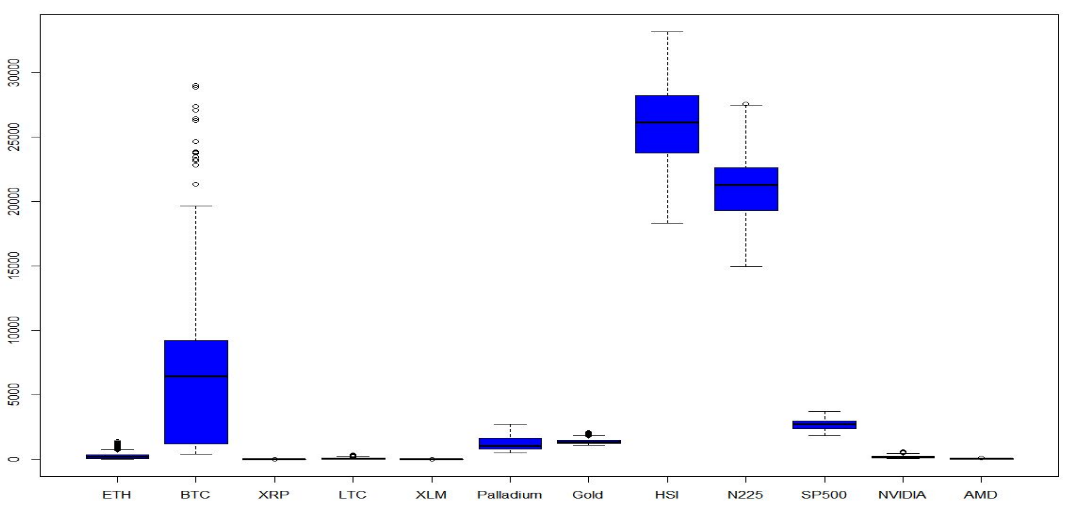
- To the variables listed above, the following have been added:

  - Financial indexes: SP500, N225 and HSI.
  - Commodities: Gold, Palladium.
  - Stock data: NVIDIA and AMD.

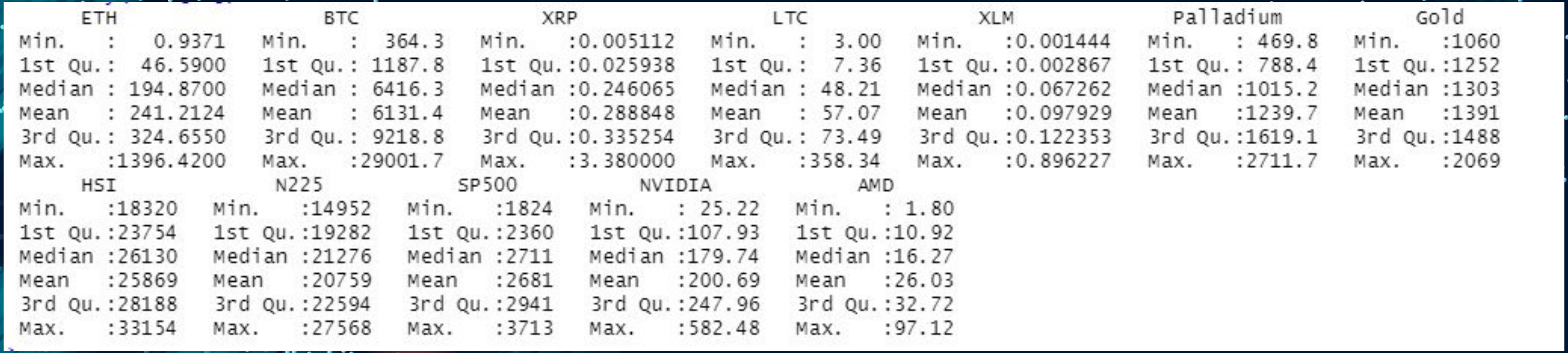
- In addition, the time series of the dataset which started from 1 January 2016 to 30 September 2019 has been extended until 31 December 2020.



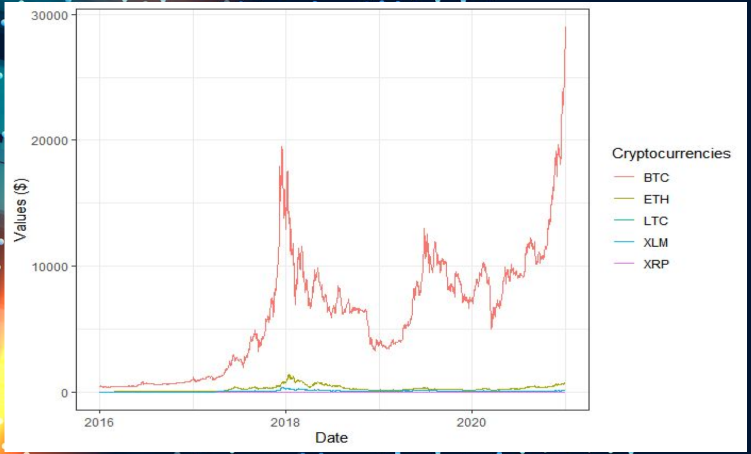
Dataset Overview and Preliminary Analysis



- From the image we can see how there is a difference in the scale of the values of the variables, this makes us understand that the data must be normalized, before using the data for a predictive analysis.

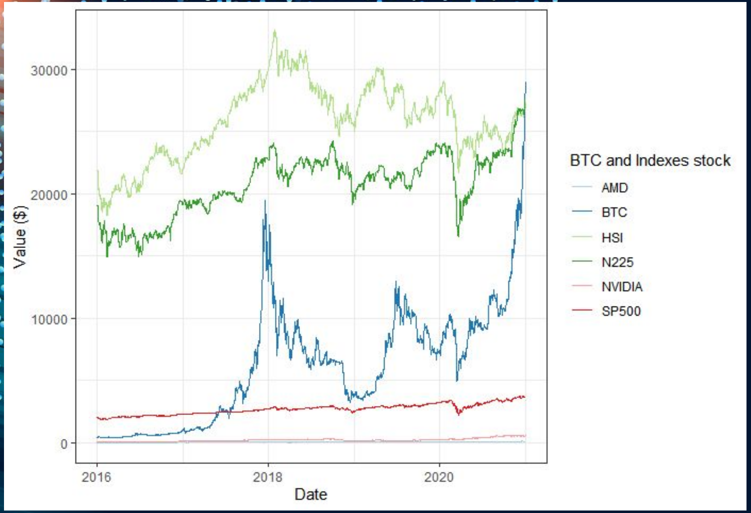


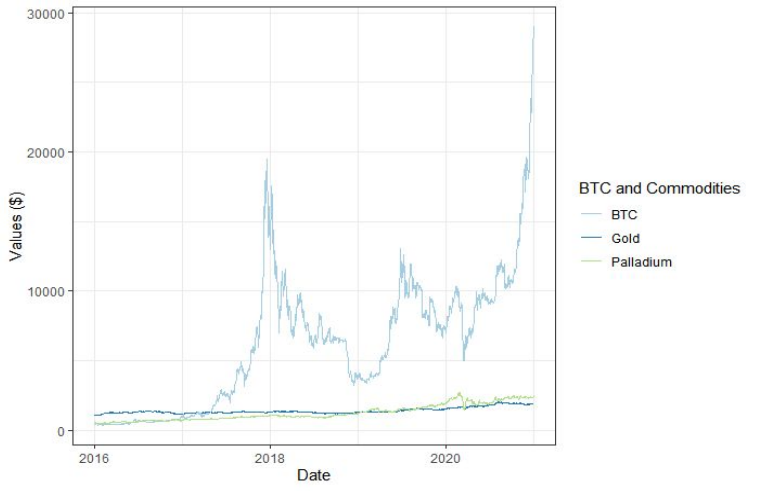
```
      ETH                 BTC                 XRP                  LTC                XLM                  Palladium          Gold     
 Min.   :   0.9371   Min.   :  364.3   Min.   :0.005112   Min.   :  3.00   Min.   :0.001444   Min.   : 469.8   Min.   :1060  
 1st Qu.:  46.5900   1st Qu.: 1187.8   1st Qu.:0.025938   1st Qu.:  7.36   1st Qu.:0.002867   1st Qu.: 788.4   1st Qu.:1252  
 Median : 194.8700   Median : 6416.3   Median :0.246065   Median : 48.21   Median :0.067262   Median :1015.2   Median :1303  
 Mean   : 241.2124   Mean   : 6131.4   Mean   :0.288848   Mean   : 57.07   Mean   :0.097929   Mean   :1239.7   Mean   :1391  
 3rd Qu.: 324.6550   3rd Qu.: 9218.8   3rd Qu.:0.335254   3rd Qu.: 73.49   3rd Qu.:0.122353   3rd Qu.:1619.1   3rd Qu.:1488  
 Max.   :1396.4200   Max.   :29001.7   Max.   :3.380000   Max.   :358.34   Max.   :0.896227   Max.   :2711.7   Max.   :2069  
      HSI               N225              SP500             NVIDIA               AMD       
 Min.   :18320   Min.   :14952   Min.   :1824   Min.   : 25.22   Min.   : 1.80  
 1st Qu.:23754   1st Qu.:19282   1st Qu.:2360   1st Qu.:107.93   1st Qu.:10.92  
 Median :26130   Median :21276   Median :2711   Median :179.74   Median :16.27  
 Mean   :25869   Mean   :20759   Mean   :2681   Mean   :200.69   Mean   :26.03  
 3rd Qu.:28188   3rd Qu.:22594   3rd Qu.:2941   3rd Qu.:247.96   3rd Qu.:32.72  
 Max.   :33154   Max.   :27568   Max.   :3713   Max.   :582.48   Max.   :97.12  
```



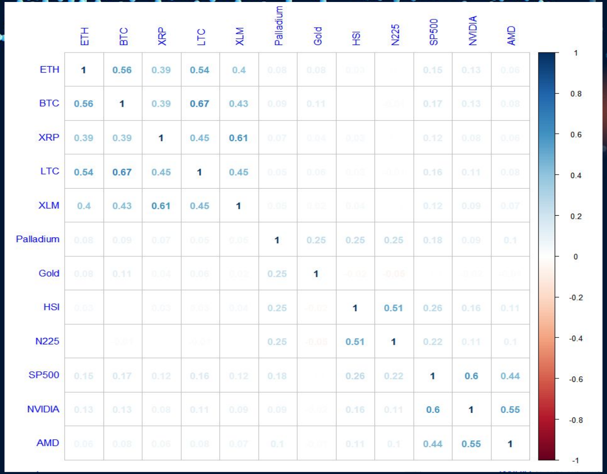
Time Series Plots

- The change in the value of the BTC cryptocurrency over time compared to other cryptocurrencies and other financial instruments.
- We can see how the BTC cryptocurrency is the one with a more unstable trend.

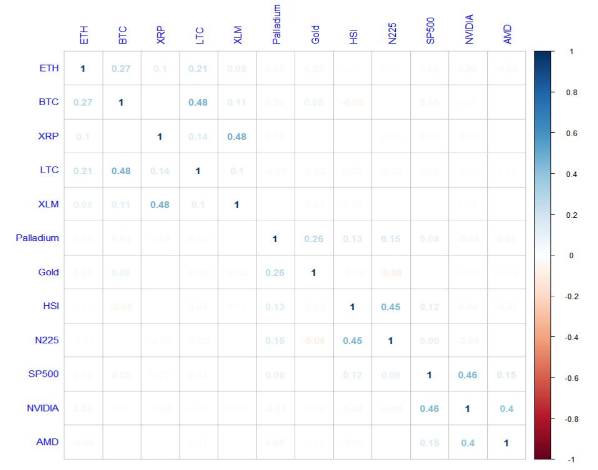




Correlation Analysis



	ETH	BTC	XRP	LTC	XLM	Palladium	Gold	HSI	N225	SP500	NVIDIA	AMD
ETH	1	0.56	0.39	0.54	0.4	0.08	0.08	0.03		0.15	0.13	0.06
BTC	0.56	1	0.39	0.67	0.43	0.09	0.11		-0.01	0.17	0.13	0.08
XRP	0.39	0.39	1	0.45	0.61	0.07	0.04	0.03		0.12	0.08	0.06
LTC	0.54	0.67	0.45	1	0.45	0.05	0.06	0.03	-0.01	0.16	0.11	0.08
XLM	0.4	0.43	0.61	0.45	1	0.05	0.02	0.04		0.12	0.09	0.07
Palladium	0.08	0.09	0.07	0.05	0.06	1	0.25	0.25	0.25	0.18	0.09	0.1
Gold	0.08	0.11	0.04	0.06	0.02	0.25	1	-0.02	-0.05		-0.02	-0.01
HSI	0.03		0.03	0.03	0.04	0.25	-0.02	1	0.51	0.26	0.16	0.11
N225		-0.01		-0.01		0.25	-0.05	0.51	1	0.22	0.11	0.1
SP500	0.15	0.17	0.12	0.16	0.12	0.18		0.26	0.22	1	0.6	0.44
NVIDIA	0.13	0.13	0.08	0.11	0.09	0.09	-0.02	0.16	0.11	0.6	1	0.55
AMD	0.06	0.08	0.06	0.08	0.07	0.1	-0.01	0.11	0.1	0.44	0.55	1



	ETH	BTC	XRP	LTC	XLM	Palladium	Gold	HSI	N225	SP500	NVIDIA	AMD
ETH	1	0.27	0.1	0.21	0.08	0.02	0.03	-0.01	-0.01	0.02	0.05	-0.04
BTC	0.27	1		0.48	0.11	0.04	0.08	-0.06		0.05	0.01	
XRP	0.1		1	0.14	0.48	0.03			-0.02	0.03	-0.02	
LTC	0.21	0.48	0.14	1	0.1	-0.03	-0.02	0.03	-0.02	0.02	-0.04	0.01
XLM	0.08	0.11	0.48	0.1	1		-0.03	0.02			0.02	
Palladium	0.02	0.04	0.03	-0.03		1	0.26	0.13	0.15	0.08	-0.04	0.05
Gold	0.03	0.08		-0.02	-0.03	0.26	1	-0.03	-0.08		-0.02	-0.02
HSI	-0.01	-0.06		0.03	0.02	0.13	-0.03	1	0.45	0.12	0.04	-0.02
N225	-0.01		-0.02	-0.02		0.15	-0.08	0.45	1	0.09	-0.04	
SP500	0.02	0.05	0.03	0.02		0.08		0.12	0.09	1	0.46	0.15
NVIDIA	0.05	0.01	-0.02	-0.04	0.02	-0.04	-0.02	0.04	-0.04	0.46	1	0.4
AMD	-0.04			0.01		0.05	-0.02	-0.02		0.15	0.4	1

Correlation

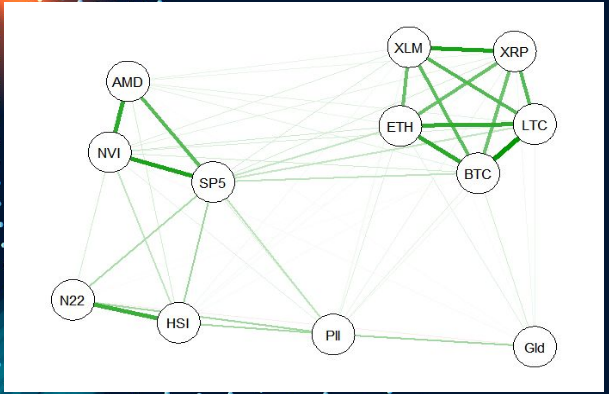
- Most of the variables are correlated with one another, but this effect could be due to fake relationships.
- As was obvious to imagine, there is a strong correlation between the different cryptocurrencies.

Partial correlation

- The partial correlation shows this, and the intensity of some relationships between the variables is different from that obtained in the correlation.

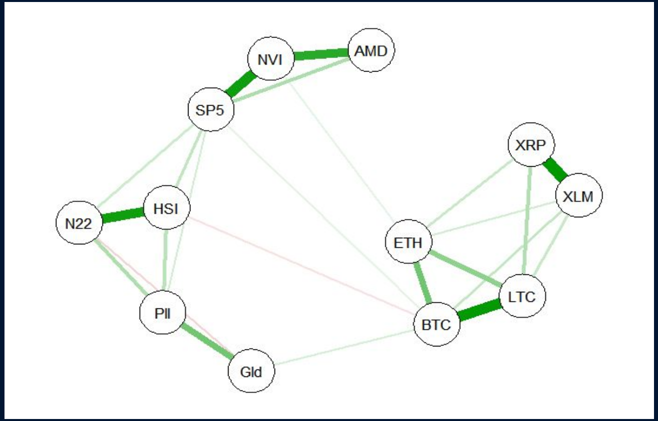
Network Graphs

- The relationship between variables was also examined through network graphs.



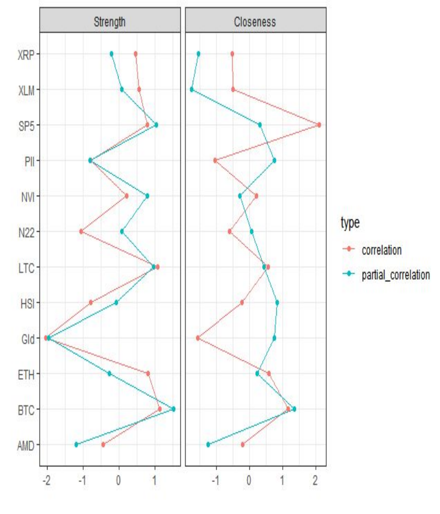
Correlation

In the correlation graph, most of the variables
are clustered and positively-correlated.
It doesn’t provide much useful information in this form.



Partial correlation

The partial correlation graph shows that there are some negative relationships. It’s also
possible to identify some variables that are less correlated with the rest.

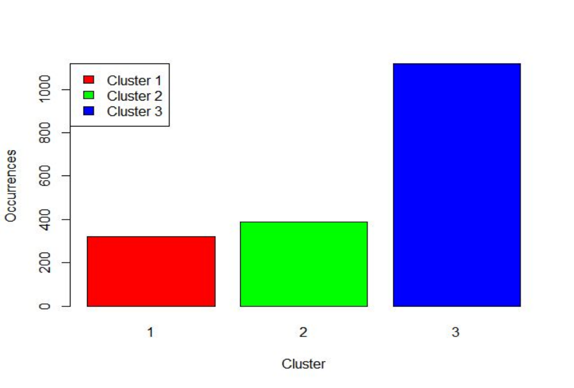


Centrality

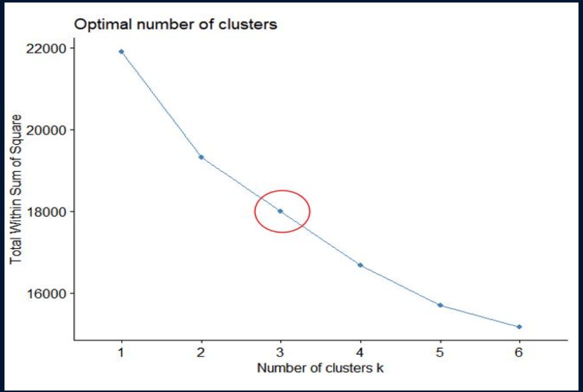
we can see how there is a variability in both correlation and partial correlation, from the chart we can also see how BTC has a strong centrality, so it is easier to predict than the other variables.

Clustering: K-means

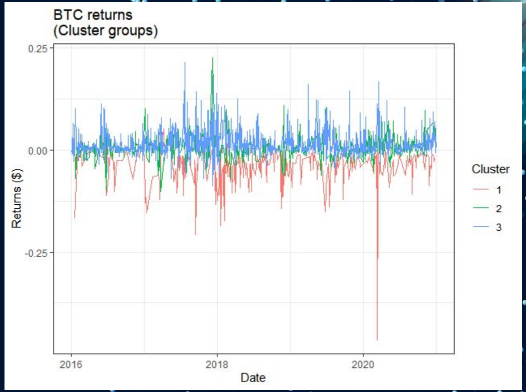
K-means clustering was performed on the data provided to find the groups which may provide additional information for analysis.



We can see how for cluster number 3 there are more observations than for the other 2.



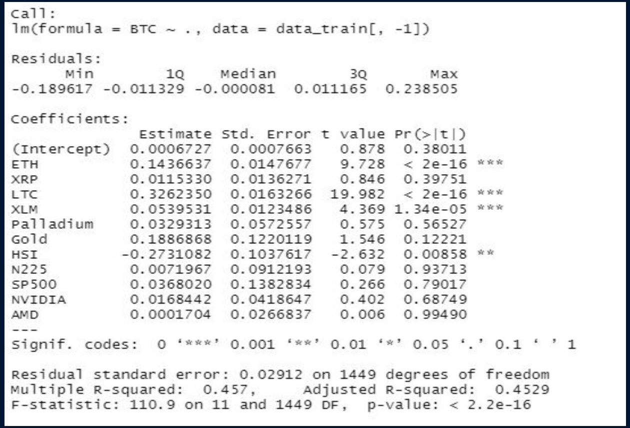
Through the elbow method, 3 was found to be the optimal number of clusters.



The image shows the price trend, intended as returns with respect to the 3 clusters. We can see how for cluster 1 there is a negative trend for almost the whole time period considered.

Linear Regression

- For the first model, It was used the linear regression. Seeing the BTC variable as response variable, and all others as predictors.
- The data division adopted for model training was 80-20 training-test split.
- The metric used to estimate the goodness-of-fit of the models is the R squared.



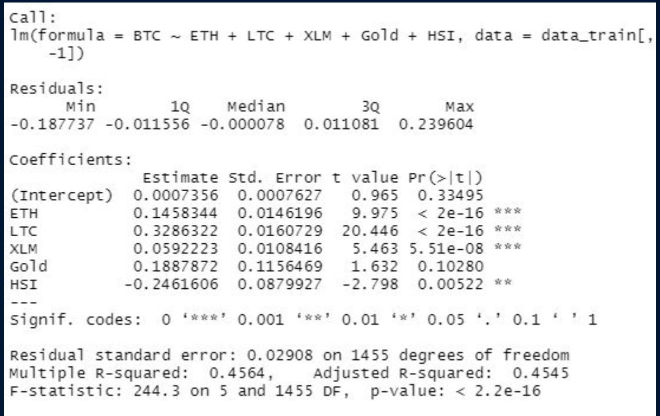
```
Call:
lm(formula = BTC ~ ., data = data_train[, -1])

Residuals:
      Min        1Q    Median        3Q       Max 
-0.189617 -0.011329 -0.000081  0.011165  0.238505 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.0006727  0.0007663   0.878  0.38011    
ETH          0.1436637  0.0147677   9.728  < 2e-16 ***
XRP          0.0115330  0.0136271   0.846  0.39751    
LTC          0.3262350  0.0163266  19.982  < 2e-16 ***
XLM          0.0539531  0.0123486   4.369 1.34e-05 ***
Palladium    0.0329313  0.0572557   0.575  0.56527    
Gold         0.1886868  0.1220119   1.546  0.12221    
HSI         -0.2731082  0.1037617  -2.632  0.00858 ** 
N225         0.0071967  0.0912193   0.079  0.93713    
SP500        0.0368020  0.1382834   0.266  0.79017    
NVIDIA       0.0168442  0.0418647   0.402  0.68749    
AMD          0.0001704  0.0266837   0.006  0.99490    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02912 on 1449 degrees of freedom
Multiple R-squared:  0.457,	Adjusted R-squared:  0.4529 
F-statistic: 110.9 on 11 and 1449 DF,  p-value: < 2.2e-16
```

MODEL WITH ALL VARIABLES

R^2 = 0.735043

Stepwise Regression

- So we have seen that several variables have got high p-values: this means that improvements may be possibles through simpler models that use fewer features for regression.



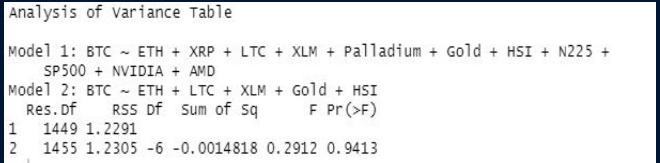
```
Call:
lm(formula = BTC ~ ETH + LTC + XLM + Gold + HSI, data = data_train[,
    -1])

Residuals:
      Min        1Q    Median        3Q       Max
-0.187737 -0.011556 -0.000078  0.011081  0.239604

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0007356  0.0007627   0.965  0.33495
ETH          0.1458344  0.0146196   9.975  < 2e-16 ***
LTC          0.3286322  0.0160729  20.446  < 2e-16 ***
XLM          0.0592223  0.0108416   5.463 5.51e-08 ***
Gold         0.1887872  0.1156469   1.632  0.10280
HSI         -0.2461606  0.0879927  -2.798  0.00522 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02908 on 1455 degrees of freedom
Multiple R-squared:  0.4564,	Adjusted R-squared:  0.4545
F-statistic: 244.3 on 5 and 1455 DF,  p-value: < 2.2e-16
```

STEPWISE BOTH DIRECTION

R^2 = 0.742909

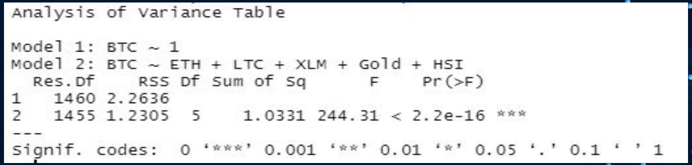
There are no big improvements in performance over the full model, but this is better chosen due to the smaller number of variables.

full model vs simpler model



```
Analysis of Variance Table

Model 1: BTC ~ ETH + XRP + LTC + XLM + Palladium + Gold + HSI + N225 +
    SP500 + NVIDIA + AMD
Model 2: BTC ~ ETH + LTC + XLM + Gold + HSI
  Res.Df    RSS Df  Sum of Sq      F Pr(>F)
1   1449 1.2291
2   1455 1.2305 -6 -0.0014818 0.2912 0.9413
```

no predictor vs simpler model

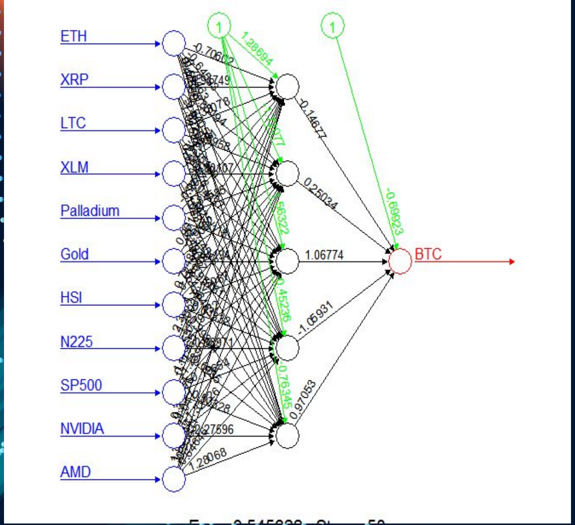


```
Analysis of Variance Table

Model 1: BTC ~ 1
Model 2: BTC ~ ETH + LTC + XLM + Gold + HSI
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1460 2.2636
2   1455 1.2305  5    1.0331 244.31 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Neural Network

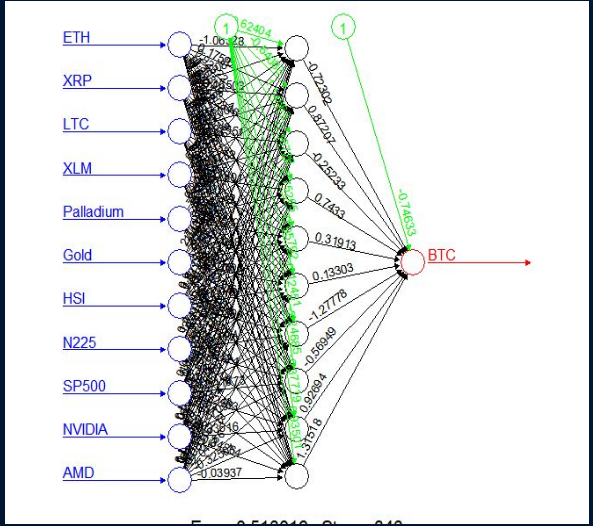
Neural network with one hidden layer.

5 HIDDEN NEURONS



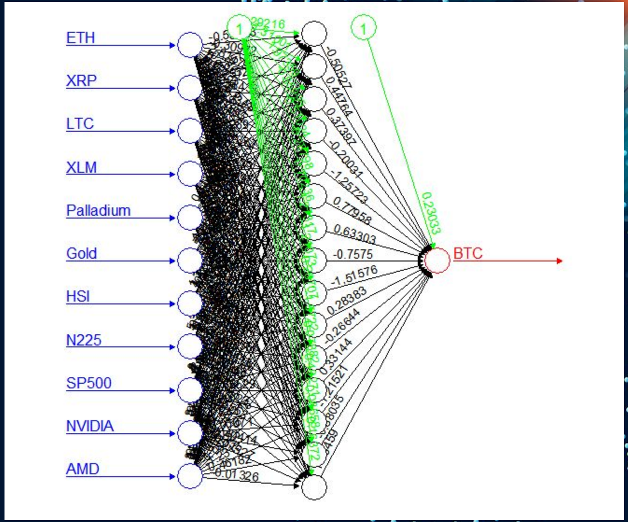
R^2 = 0.7686203

10 HIDDEN NEURONS



R^2 = 0.7718492

15 HIDDEN NEURONS

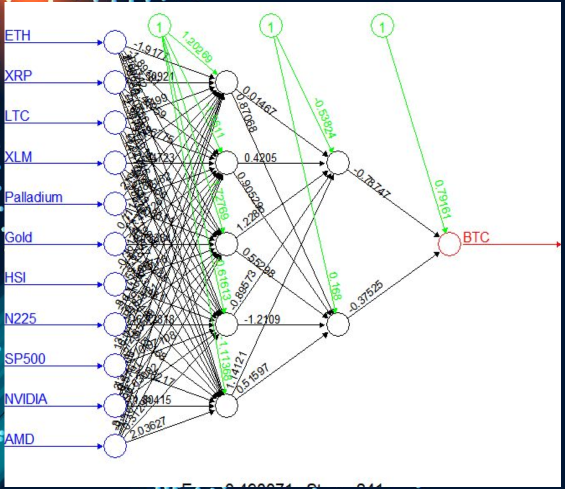


R^2 = 0.7716683

Neural Network

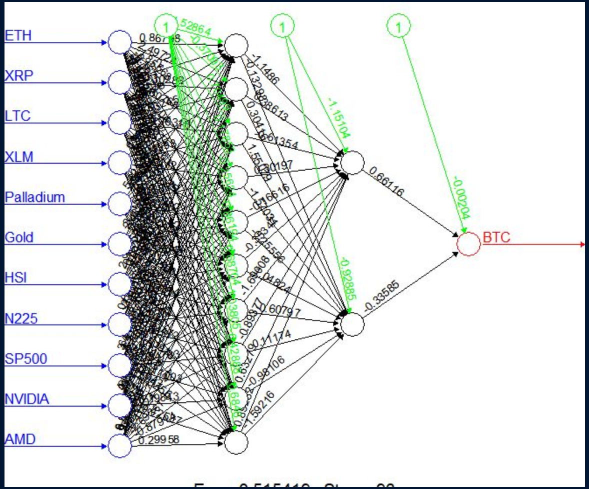
Neural network with two hidden layer, the second layer with 2 neurons.

5 HIDDEN NEURONS



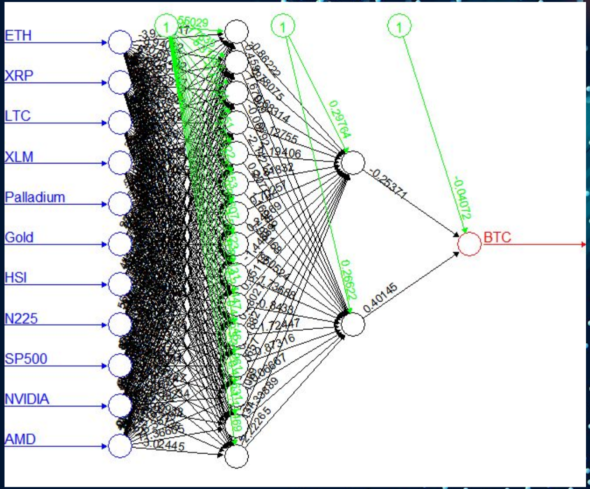
R^2 = 0.677365

10 HIDDEN NEURONS



R^2 = 0.6996827

15 HIDDEN NEURONS

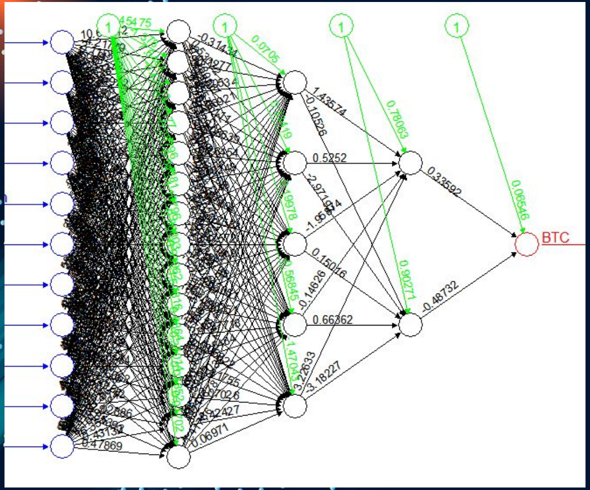


R^2 = 0.4682243

Neural Network

Neural network with 3 hidden layer, with 5 and 2 neurons in the second and third layers.

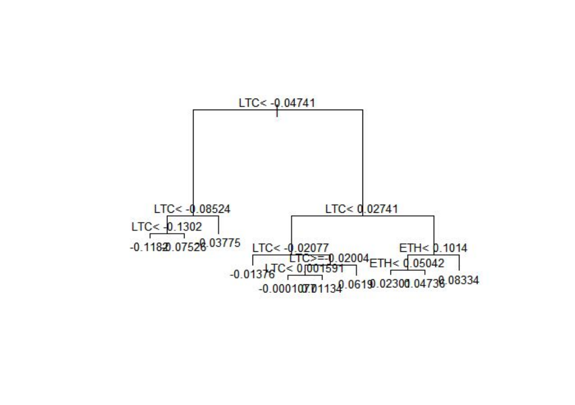
15 HIDDEN NEURONS



R^2 = 0.3397162

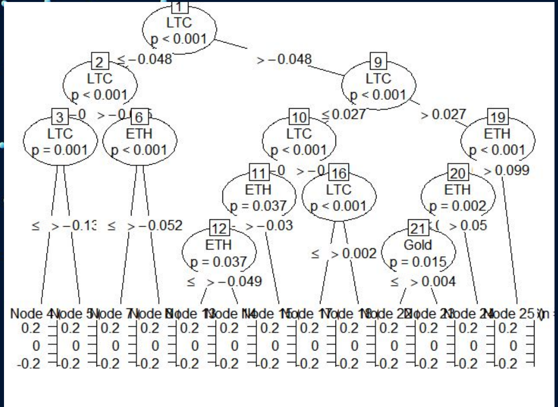
- We can see that by increasing the width or depth by hidden the layers, the result it is often a worse performance.
- Networks with multiple layers have a longer execution time.
- Among the various network configurations tested, the one with a hidden layer with 5 nodes reports the best performances, in terms of model accuracy and execution time.

Tree model Comparison



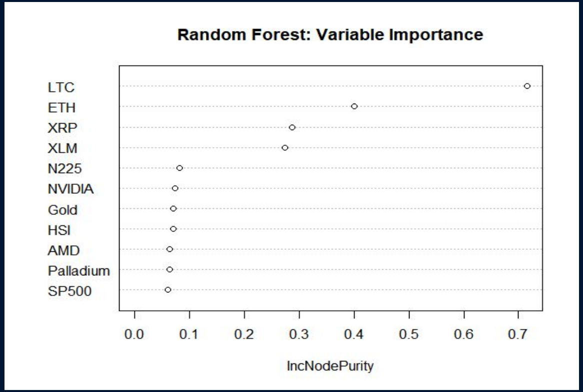
CART TREE

R^2 = 0.526549



CONDITIONAL TREE

R^2 = 0.5516889



RANDOM FOREST

R^2 = 0.6204256

- The random forest returns slightly better results than those reported by the other two trees tested, but still unsatisfactory. We can see as we had already seen even with linear models that the most important variables for BTC prediction are cryptocurrencies.

Conclusions

- We have seen that the models that provided the best performance were the neural network and linear regression after making the selection of the features.

- Tree regression models performed the worst.

- We have seen how cryptocurrencies tend to have a strong correlation between them, and very low compared to other variables, so it is not easy to be able to build a model with good performance that does not also use these variables.

- Further improvements may be possible with more data.

Thank you for your attention