MONASH University

MONASH BUSINESS SCHOOL

Comparative Analysis and Forecasting of Clean vs. Dirty Cryptocurrencies

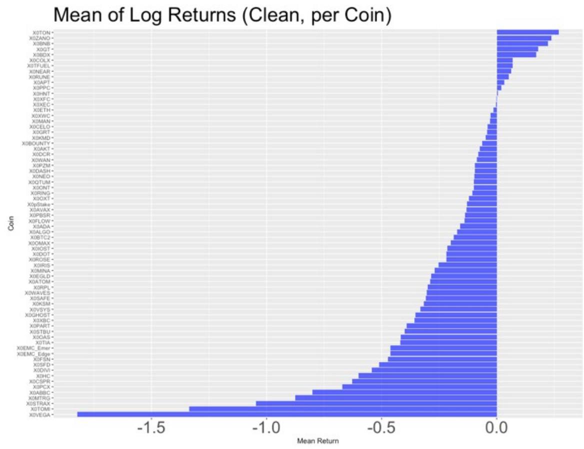
ETW3481_A1

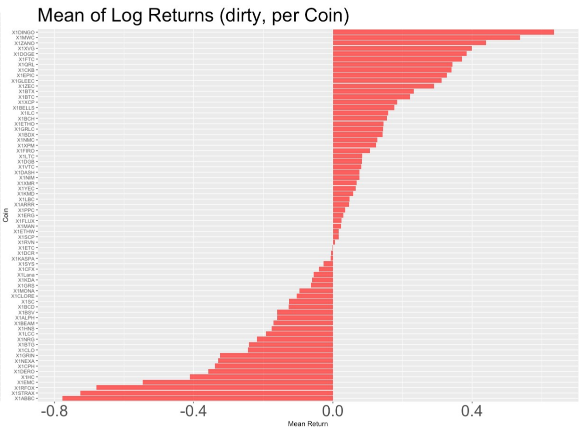
BYUN JIHOO

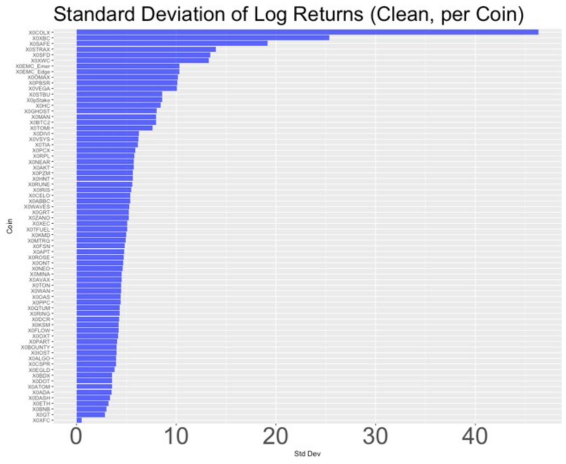
[Introduction]

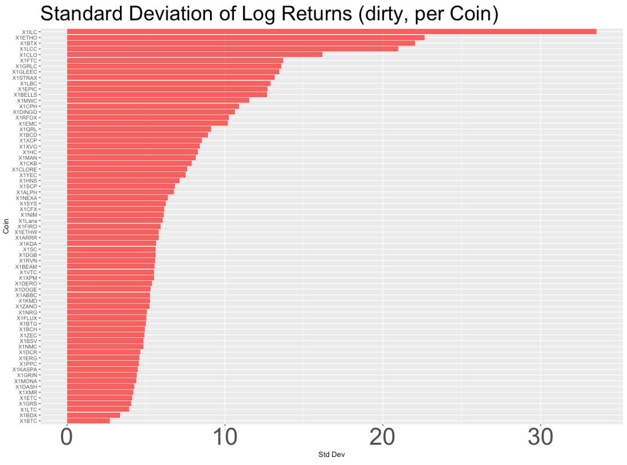
Clean and dirty cryptos are identified depending on energy consumption of the algorithms they use (Ben, & Lucey, 2022). Clean cryptos are the cryptos built based on algorithms with energy efficiency including ‘Proof of Stake’, ‘Ripple Protocol’, or ‘Stellar Protocol’ (Ben, & Lucey, 2022). Dirty cryptos are built on algorithms which requires massive energy to maintain their mining and activity of transactions which includes Bitcoin, Ethereum, or Monera (Ben, & Lucey, 2022).

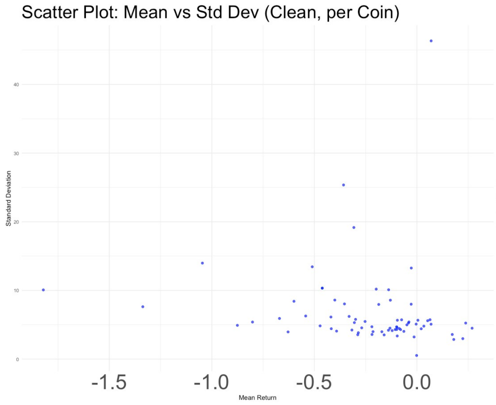
[Mean and Standard Deviations]



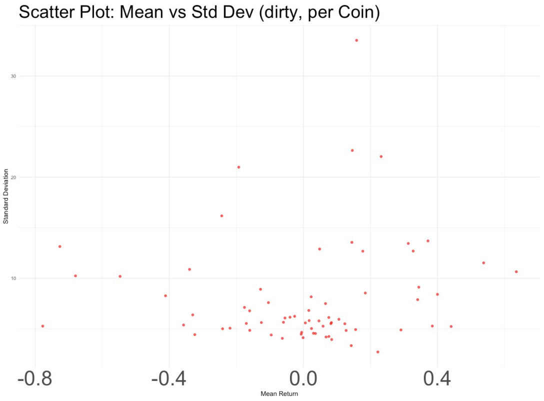








Scatter Plot: Mean vs Std Dev (Clean, per Coin)



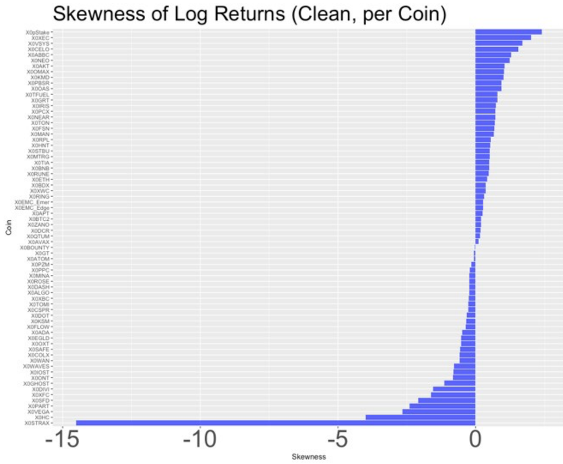
Scatter Plot: Mean vs Std Dev (dirty, per Coin)

As it can be seen from the plot, average return is higher in dirty cryptos, but standard deviation is deviated less in the dirty cryptos. From the scatter plot, dirty crypto shows more distribution for scatters.

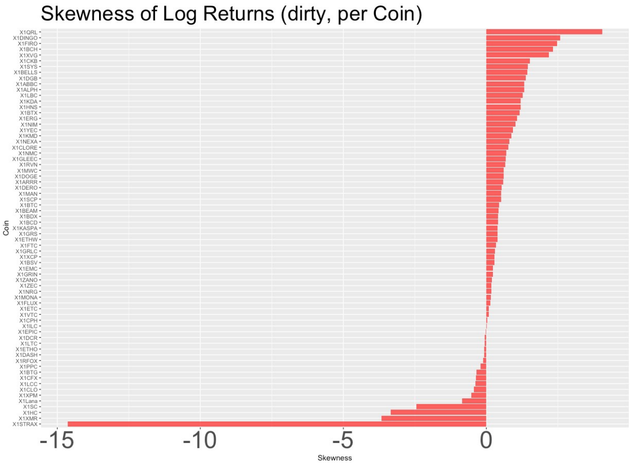
Clean dirty crypto may look more profitable in average data and it carries higher volatility than the clean crypto. Profit the clean crypto can generate is less than dirty crypto but loss the clean crypto can occur is larger than the dirty crypto and standard deviation itself is higher in the clean crypto.

Therefore, it is recommended to invest in the dirty crypto.

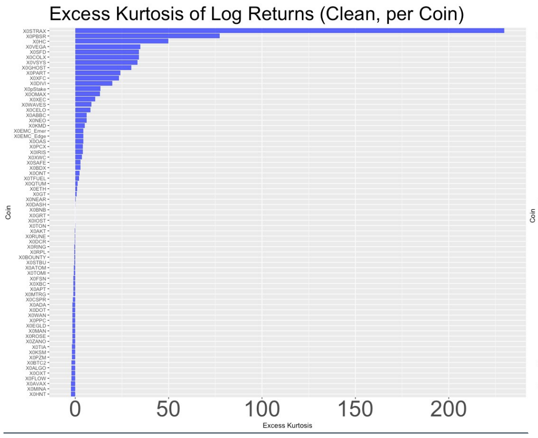
[Skewness and Excess Kurtosis]



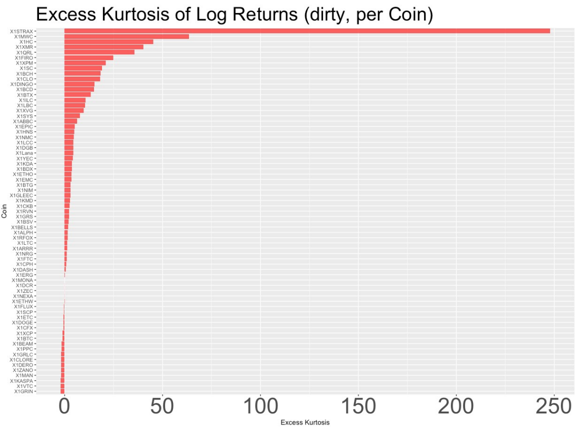
Skewness of Log Returns (Clean, per Coin)



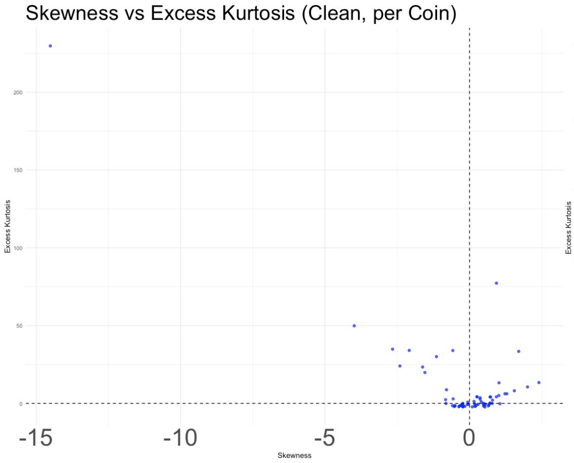
Skewness of Log Returns (dirty, per Coin)

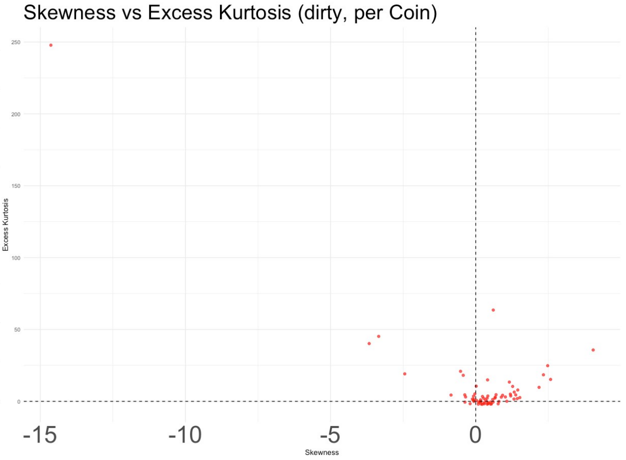


Excess Kurtosis of Log Returns (Clean, per Coin)



Excess Kurtosis of Log Returns (dirty, per Coin)



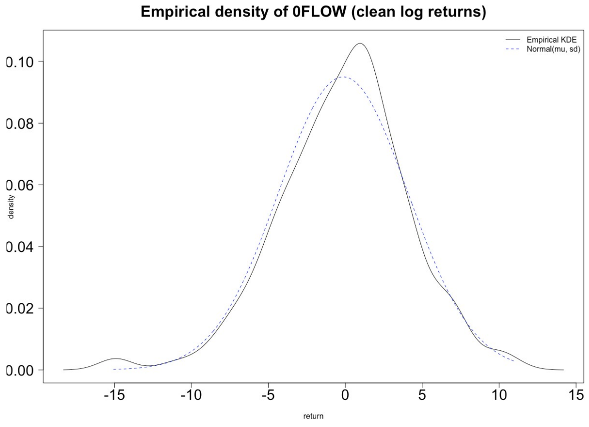


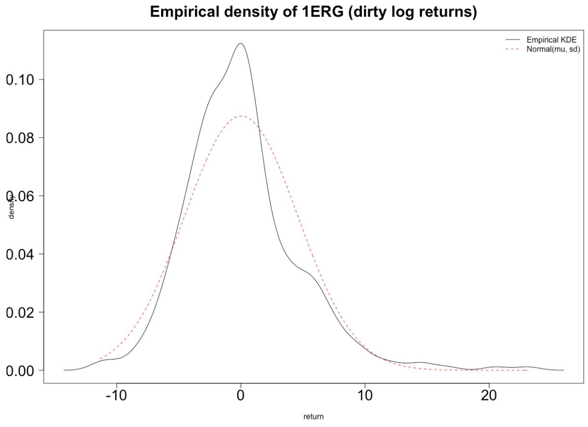
Implication of skewedness and excess kurtosis in investment return is significant. Clean crypto seems it has balance of positive and negative skewness. There will be balanced loss and gain. However, dirty crypto has a lot of positively skewed. In dirty cryptos, there will be numerous frequent small losses and few large gains.

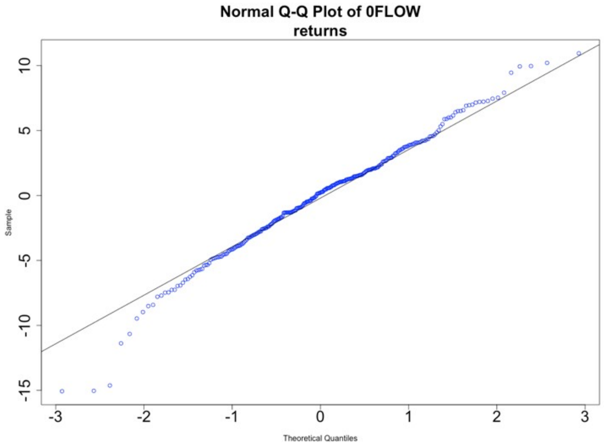
In excess kurtosis, clean crypto has high kurtosis while dirty crypto kurtosis show dramatic decrease. In dirty crypto, investors may not face extreme tail data much as they may face in the clean cryptos.

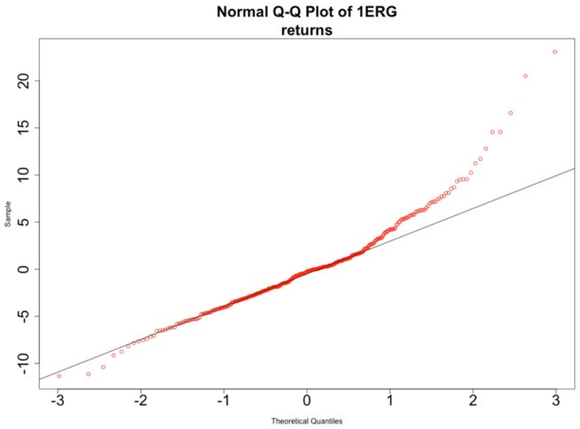
In the scatter plot, dirty crypto excess kurtosis remains lower than clean cryptos while its skewness is usually positive. Clean crypto investors will have similar change to loss of gain but will encounter extreme values in the tail but dirty crypto investors will have stable pattern of frequent small losses and few large gains.

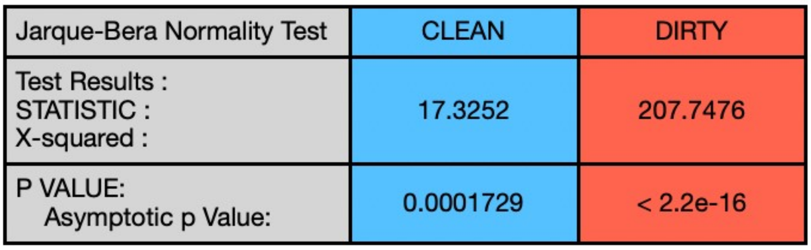
[Distribution and Normality Test]











Jarque-Bera Normality Test	CLEAN	DIRTY
Test Results : STATISTIC : X-squared :	17.3252	207.7476
P VALUE: Asymptotic p Value:	0.0001729	< 2.2e-16

$$\boxed{H_0: Distribution\ is\ Normal}$$

$$\boxed{H_1: Distribution\ is\ Normal}$$

$$\boxed{95\%\ confidence\ level}$$

$$\boxed{Clean\ p < 0.05, Dirty\ p < 0.05}$$

$$\boxed{Therefore, reject\ the\ NULL\ hypothesis\ at\ both\ of\ Clean\ and\ Dirty\ cryptos.}$$

According to the empirical density plot and normal distribution plot, clean crypto is positively skewed across the all cryptos while dirty crypto is negatively skewed across the all cryptos. It is unlike what has been observed from the skewness plot above.

Clean crypto investors may encounter small frequent gains and few large losses while dirty crypto investors may encounter small frequent losses and few large gains.

The QQ-plot of both clean and dirty crypto is hard to be said normal because the scatters are not aligning the black line. However, clean crypto is more normal than the dirty cryptos.

In JB normality test, both p-value is less than 0.05 and reject the null hypothesis which asserts the data is normal.

[Key Differences]

The key differences between the dirty and clean crypto return is mean and skewness. Although dirty crypto will show few gains and frequent losses, those few gains seem large enough because there are more positive return mean values in the dirty cryptos than the clean cryptos. However, despite of frequent gains of clean crypto, few losses seems to be dangerously large so that most of clean crypto is not generating profit and their loss is large than dirty crypto loss. Additionally, as excess kurtosis in clean crypto is larger than the dirty crypto, high volatility in frequent gain scenarios is expected.

[Part 2 Introduction]

0RPL has been chosen from clean crypto. Among all clean crypto, 0RPL ends the most recently while it has at least 2000 continuous observations without NA. To produce accurate and practical prediction, using the most recent data is considered as strategic and 0RPL was qualified.

1BTC has been chosen from the dirty crypto. In fact, 1BTC, 1LTC, and 1NMC had more than 2000 continuous observations without NA and ending at the same date which is the most recent. However, BTC is considered as the more representative crypto of dirty crypto. Considering its symbolism and impact in the crypto market, 1BTC has been selected.

Therefore, the most recent and the most representative cryptos are: ‘0RPL’ from the clean crypto and ‘1BTC’ from the dirty crypto.

[Model Briefing]

AR(1) assumption is white-noise assumption and past value remains its impact. The model is conditional on the past data which is rt-1. It is strong since the model indicates the serial-correlation which handles time series of financial data. However, if the data is non-stationary or non-linear, or is not serially correlated, forecast gets harder. Besides, rt-1 data can be insufficient to estimate $\boxed{\varepsilon_t^2}$ and the model should be extended to care correlation if residual ACF shows additional serial correlations.

MA(1) model assumes weakly stationary while past error remains in impact since they are linear combinations of white-noise sequence. In MA(1) model, ACF after lag 1 is zero, relies on past error, and invertible. It is finite memory model therefore they are linearly related to it first q-lagged value. MA(1) as strength since it easily track financial return data by eliminating random fluctuation and is computationally efficient (Jason, 2025). However, it requires to save error data which is costly and ignores complex relationship to characterize mean value rather than capturing natural fluctuations (Jason, 2025).

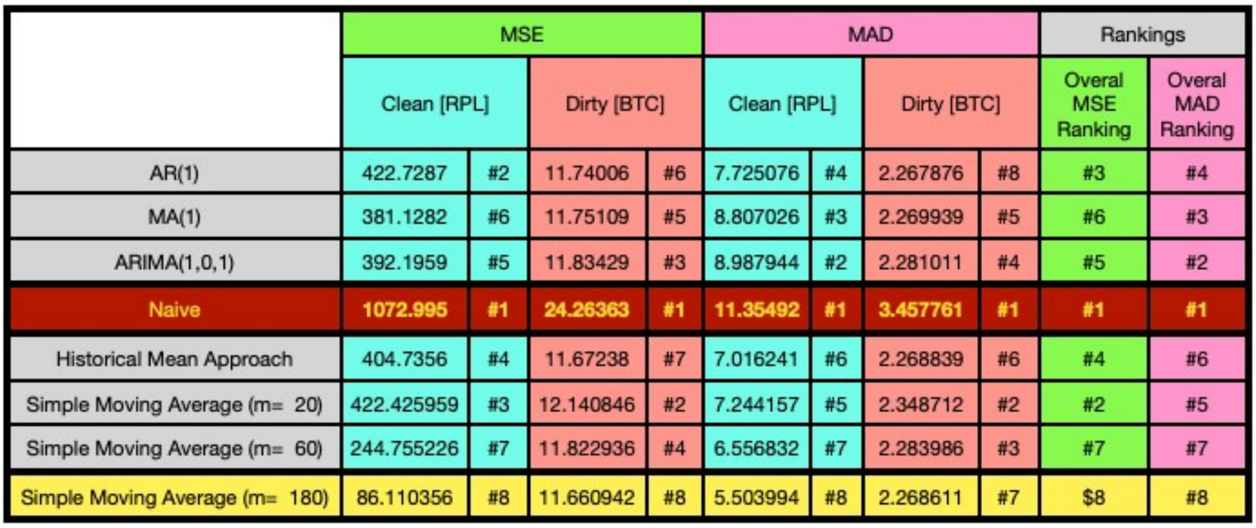
ARIMA(1,0,1) model assumes time series data is stationary. It make forecast using both past observations and past fluctuations. ARIMA(1,0,1) is efficient and flexible to forecast various behaviors and pattern but, cannot handle non-linear data or sudden shock.

Naïve model assumes that the data is conditionally independent each others. It is using the probabilistic nature for forecasting. Naïve model is suitable for small sample size data for light forecasting while its forecasting result is not highly reliable and assumption is nor usual in reality.

Historical mean approach model assumes that the data mean will not change a lot but will remain constant. It is utilizing previous period average return to forecast current period return. Historical mean approach model is easy to calculate and use as benchmarking model while it is not reliable while it gets heavy effect from outliers

Simple Moving Average Model assumes the innovation measurement period and moving average period is one month, three months, and six months. It carries error shock from fixed period past time over number of periods using closing prices (PSU, n.d).. It is easy and efficient to calculate and the result is smoothen but does not reflect seasonality or trend in the model and does not reflect current market condition since it relies on the past data (PSU, n.d).

[Accuracy index comparison: Introduction]



	MSE				MAD				Rankings	
	Clean [RPL]		Dirty [BTC]		Clean [RPL]		Dirty [BTC]		Overal MSE Ranking	Overal MAD Ranking
AR(1)	422.7287	#2	11.74006	#6	7.725076	#4	2.267876	#8	#3	#4
MA(1)	381.1282	#6	11.75109	#5	8.807026	#3	2.269939	#5	#6	#3
ARIMA(1,0,1)	392.1959	#5	11.83429	#3	8.987944	#2	2.281011	#4	#5	#2
Naive	1072.995	#1	24.26363	#1	11.35492	#1	3.457761	#1	#1	#1
Historical Mean Approach	404.7356	#4	11.67238	#7	7.016241	#6	2.268839	#6	#4	#6
Simple Moving Average (m= 20)	422.425959	#3	12.140846	#2	7.244157	#5	2.348712	#2	#2	#5
Simple Moving Average (m= 60)	244.755226	#7	11.822936	#4	6.556832	#7	2.283986	#3	#7	#7
Simple Moving Average (m= 180)	86.110356	#8	11.660942	#8	5.503994	#8	2.268611	#7	$8	#8

Mean Squared Error gives heavier penalties in larger error by squaring them. It brings significance when the large error is costly high. Meaning that, as error increase, MSE will increase exponentially.

Mean Absolute Error which takes error equally and is robust index which gets less impact from outliers and provide easier interpretation regarding the data itself.

[Accuracy index comparison: RPL-clean crypto]

Among the clean crypto currency RPL, Naïve-model showed the highest MSE and MAE among the other models. The index is extremely larger than the other models meaning the naïve model is not performing its forecasting meaningfully.

The return rate may fluctuate a lot therefore previous lag value does not reflect the current time value in the forecasting. Besides, large difference between those two indices implies existence of big outliers in the error since MSE squares it error value and it make the index more sensitive to outlier observations.

Phenomenon of inconsistent ranking of MSE and MAE for the other models indicates their unstable distribution of outliers and implies highly potential rapid increase/decrease of the return rate data.

One outperforming model among all eight models is `Simple Moving Average` model which sets m=180. It presents dramatic small value for both MSE and MAE while gap between the two indices is the smallest in `SMA` model. Remarkably small MSE proves that the model forecasting is reliable since it presents small outliers and most of observations are distributed narrowly. Additionally, due to the smoothening effect of `SMA` nature, rapid fluctuation which gave a heavy penalty in the Naïve model has been managed in the SMA model and finally showed the best index value.

Reviewing the other models which forecasts next period data using one lag data such as AR(1), MA(1), ARIMA(1,0,1) and Historical Mean Approach model, the return data is likely to be a noise pattern which shows frequent fluctuation so that one lag previous data does not reflect current period data well.

[Accuracy index comparison: BTC-dirty crypto]

As it has been observed in the clean crypto, Naïve model showed the highest MSE and MAD value. However, compare to the case in the clean crypto, MSE gap between Naïve and the other model is not large as it was in the clean crypto in the scale of value itself and ratio. Therefore, the outlier in the dirty crypto is not expected to be extreme as it was in the clean crypto. Therefore, log return forecast is more reliable in dirty crypto than the clean crypto.

The Mean Absolute Error value shows the highest value in the Naïve model again. However, it is not big as dramatic as it was in the clean crypto.

Considering the nature of Naïve model mechanics, current period actual return rate data covers next period return rate forecast better in the dirty crypto than it does in the clean crypto.

Reviewing AR (1), MA (1), ARIMA (1,0,1) and Historical Mean Approach which forecasting next period value using one lag period data, dirty crypto seems is it not fluctuating and generating noise severely as it is in the clean data. One lag previous data is well reflecting the next period data, at least how well they reflect is not significantly different across all models including the SMA (180) model.

[Outperforming Model Further Discussion]

Among the other models, Simple Moving Average model which sets m equal to 180 performed the best in both of clean and dirty crypto. It showed the least value of MSE and MAD indices.

Since the `m` is set 180, the model takes previous 180 observations as one set and utilize them to make forecasting. The dramatic outperformance of the model is revealed in clean crypto return forecasting. As it has been mentioned above, the clean crypto data seems noise data which shows frequent but constant interval fluctuation. Therefore, while taking 180 days a one set, it naturally smoothens the noise of the data and outliers' impact significantly decreased in the forecasting performance.

In dirty crypto, the model would operate in the same beneficial way which it did in the clean crypto. However, its outperforming is not as remarkable as it it in the clean crypto forecasting. However, accounting 180 days as one set managed noise of data and managed outlier impact.

To identify reason of difference between MSE and MAD indices, outliers would play a significant role due to squaring characteristics of MSE and equally treating behaviour of MAD. More outlaying data is giving heavier penalty and penalty degree is increasing exponentially.

[Comparison between complex models and the other models]

Unexpectedly, complex models including AR(1), MA(1), and ARIMA(1,0,1) did not perform better than the simple model. Except the Naïve model which performed the worst in both assets, it is hard to conclude complex model has better performance.

Ironically, Simple Moving Average model with m=180 showed the best performance and even better than the complex models.

The expected reason is the nature of crypto assets. Crypto asset itself is volatile asset which fluctuates frequently. Due to this nature, one previous data may not have meaningful impact on forecasting next period data. Therefore, SMA(180) which covers the longest period and smoothens those noise like fluctuation could perform the best. However, if those complex models get developed and start using more lagged data, the result may change. To sum up, complex model functionality itself could be high enough but the data they cover was not sufficient to generate reliable forecasting.

[Insights summary]

Reviewing the log return rate mean, standard deviation, skewness and excess kurtosis, it is recommended to invest in the dirty crypto. Clean cryptos will being frequent gains but few losses are dangerously big while dirty crypto provides few big gains which mitigate frequent small losses.

As dirty crypto is consuming more energy in the algorithm operation, dirty cryptos are expected to be higher functioning based on larger capital invested. Considering the nature of crypto currencies which is a new legacy of capitalism with decentralization concept, return on crypto might be affected amount of capital invested in the crypto itself.

Additionally, it has been revealed that crypto itself is still volatile. Although variation range is similar across the time, variation frequency is significantly frequent. To generate meaningful forecast for crypto, length of data should be able to cover all those variations into training.

Referring to Means Square Error index, there are a lot of outliers detected, especially in the clean crypto. Consuming less energy could be reason of instability or late response to sudden shock or change of algorithms.

If investor is a brave risk taker and expecting high return when outlier occur in positive value, clean crypto might be recommended however, for safer and stable investment, dirty cryptos are recommended. In fact, dirty crypto includes major cryptos such as Bitcoin, or Ethereum which is considered as traditional and located in comparatively stable position.

[References]

Ren, B., & Lucey, B. (2022b, March 26). *Do clean and dirty cryptocurrency markets herd differently?*. Finance Research Letters. https://www.sciencedirect.com/science/article/pii/S1544612322001076#:~:text=Similar%20to%20Ren%20and%20Lucey,of%20energy%2Defficient%20consensus%20algorithms%2C

Fernando, J. (2025, August 26). *Moving average (MA): Purpose, uses, formula, and examples*. Investopedia. https://www.investopedia.com/terms/m/movingaverage.asp

PSU. (n.d.). *Moving average model overview*. Moving Average Model Overview | METEO 820: Time Series Analytics for Meteorological Data. https://www.e-education.psu.edu/meteo820/node/558