Creating a Cryptocurrency Trading Strategy with Google Trends Data

Peter Chettiar, Elaine Eu, Nicole Lim, Wei Qi Lim, Dickson Tan, Yong Wen Wong

November 16, 2021

1 Introduction

Given the recent publicity and excitement surrounding cryptocurrency due to its abnormally high returns (for example, Bitcoin had about 6800% returns from January 2017 to October 2021), predicting cryptocurrency market returns has become a natural source of interest. In addition, social media and online platforms have had a great impact on cryptocurrency returns, as can be seen from the surge of Bitcoin prices from $32,000 to $38,000 due to Elon Musk adding "#bitcoin" to his Twitter Bio.

As the market moves, investors' decisions are often driven by their emotions, especially in such a volatile market. Therefore, if the emotions of investors can be understood, a model to predict returns of the cryptocurrency market can be created.

To do so, a regression model was used to perform sentiment analysis. Linear regression was run using Google Trends data regarding search terms related to cryptocurrency and the prices of various cryptocurrencies to see if a significant linear relationship between the two could be established. Ultimately, this relationship was used to build and backtest a mean-reversion trading strategy.

Even though Google is not a conventional social media platform, there were two main reasons to use Google Trends data. Firstly, Google is the most widely used search engine. Therefore, the trends from search terms on Google would be the best representation of the overall sentiment in the market as opposed to other search engines. Secondly, data acquisition is easier as Google has a seamless user interface (UI) that allows the data to be pulled directly from the Google Trends website or via an Application Programming Interface (API).

2 Methodology

2.1 Data Selection

2.1.1 Search Terms

25 recurring search terms related to cryptocurrency that were found across various websites were selected, and are listed in Table 1.

2.1.2 Cryptocurrencies

The top twenty cryptocurrencies by market capitalization were selected as the initial cryptocurrencies of interest. Stablecoins pegged to a fiat currency or commodity were then excluded. 2 large-cap coins, 4 medium-cap coins, and 4 small-cap coins were then selected as the final cryptocurrency sample, and these are listed in Table 2.

2.2 Data Extraction

Cryptocurrency is a volatile asset class with extreme day-to-day price swings being common, hence it might be more ideal to look at daily data compared to weekly data. The prices of the various cryptocurrencies from 1 October 2017 – 30 September 2021 were obtained from Yahoo Finance, approximately 4 years worth of data.

Similarly, daily Google Trends data for the same time period was used. However, with a time period of 4 years, only weekly data was available, as Google changes the frequency of the data to weekly data for periods of more than 9 months. Many different methods to extract daily data

Search Term
Crackdown
Crypto
Cryptocurrency
Blockchain
Digital Currency
Crypto Mining
DeFi
Altcoins
Smart Contracts
Stablecoins
Binance
Bittrex
Coinbase
Vitalik
Satoshi
Cryptocurrency Regulation
Gary Gensler
Tesla
Elon Musk
Whales
Dencentralised Finance
Distributed Ledger
Halving
ICO
China Digital Currency

Table 1: Search Terms

	Name (Symbol)
Large-cap	Bitcoin (BTC)
	Ethereum (ETH)
Medium-cap	Cardano (ADA)
	Binance Coin (BNB)
	XRP (XRP)
	Dogecoin (DOGE)
Small-cap	Chainlink (LINK)
	Litecoin (LTC)
	Bitcoin Cash (BCH)
	Monero (XMR)

Table 2: Cryptocurrencies

	Training Set	Test Set
Large-cap	Bitcoin (BTC)	Ethereum (ETH)
Medium-cap	Binance Coin (BNB) XRP (XRP)	Cardano (ADA) Dogecoin (DOGE)
Small-cap	Litecoin (LTC) Bitcoin Cash (BCH) Monero (XMR)	Chainlink (LINK)

Table 3: Training and Test Set

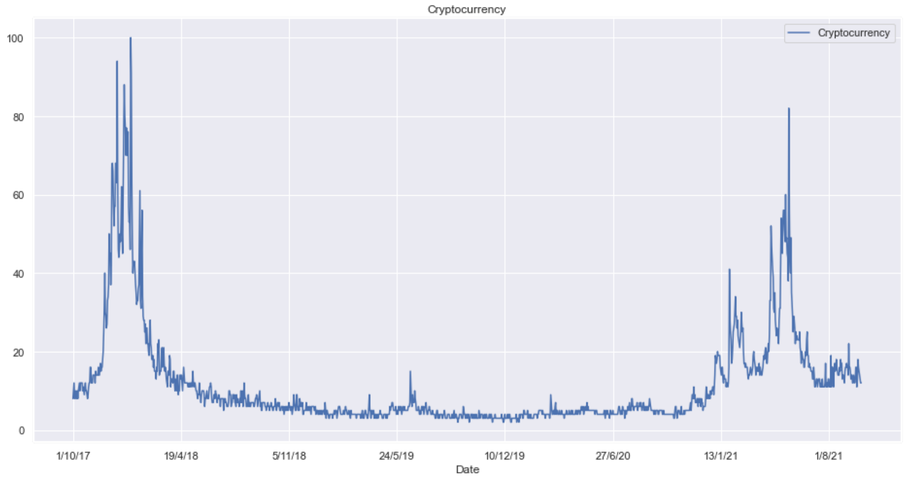


Figure 1: Google Trends Data for ”Cryptocurrency” in Sample Period

have been developed, and in this project, the “Overlapping method” (Tseng, 2019)[1] was used. This methodology involves extracting blocks of daily data within the specified time period, with time periods less than 269. Between 2 blocks of data, an overlap period will be used as a point of reference to combine and scale the 2 blocks of data, as Google Trends works on a relative scale basis. This process iterates until the required amount of daily data is obtained.

The data set was then split into a training and test set in a 6-4 split. To ensure a good representation of coins for the data set, both the training and test set include big, medium and small cap coins as shown in Table 3.

3 Exploratory Data Analysis and Statistical Analysis

3.1 EDA

Among the search terms, some have a positive correlation with Bitcoin price and some have a negative correlation. For example, the search term ”Cryptocurrency” (Figure 1) and the price of Bitcoin (Figure 2) have an apparent positive correlation for the sample period. This can be observed from a concurrent spike in Bitcoin price during the two major spikes in the search term. Another example would be the search term ”Crackdown” that has an inverse relationship with Bitcoin price.

Next, in order to make both time series stationary, the log returns were calculated. For illustrative purposes, the log returns versions of the Figures 1 and 2 (Figures 3 and 4) have been included, where different volatility patterns between the two can be observed. While price volatility peaks during the period that Bitcoin price itself peaked, the volatility of the Google Trends data decreased once prices began to exponentially increase and Bitcoin gained popularity.

The correlation matrix between the return of the training set cryptocurrencies and the various search terms is shown in Figure 5. This was done to check for multicollinearity between the search terms. Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of the resulting regression model, and thus the p-values used to identify independent variables that are statistically significant cannot be trusted. Generally, the search

[1] TSENG, Q. (2019, November 27). Reconstruct google trends daily data for extended period. Medium. Retrieved October 1, 2021, from https://towardsdatascience.com/reconstruct-google-trends-daily-data-for-extended-period-75b6ca1d3420.

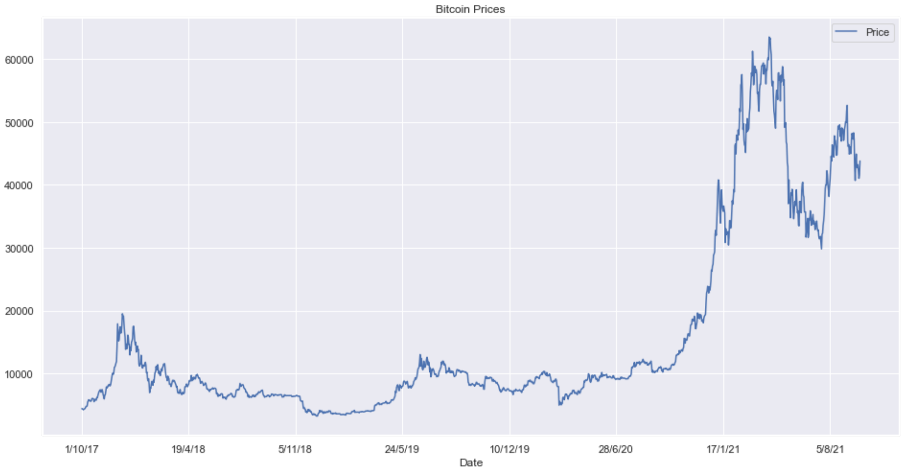


Figure 2: Bitcoin Prices in Sample Period

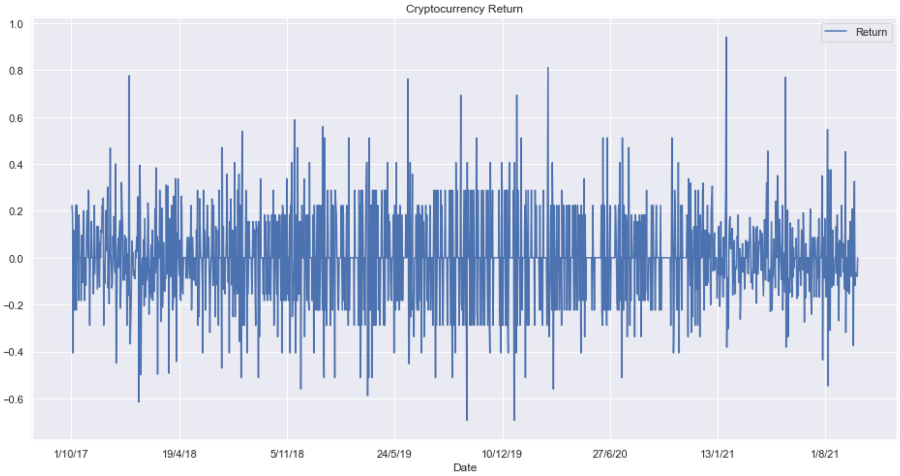


Figure 3: ”Cryptocurrency” Log Return in Sample Period

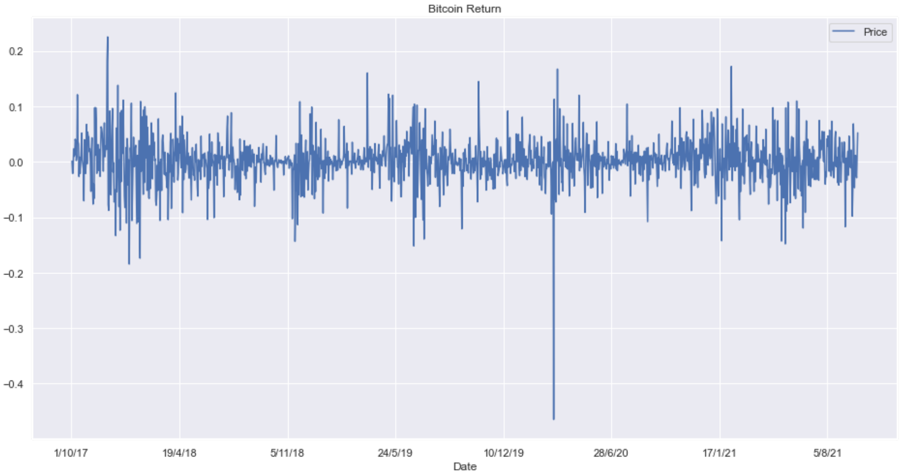


Figure 4: Bitcoin Daily Log Returns in Sample Period

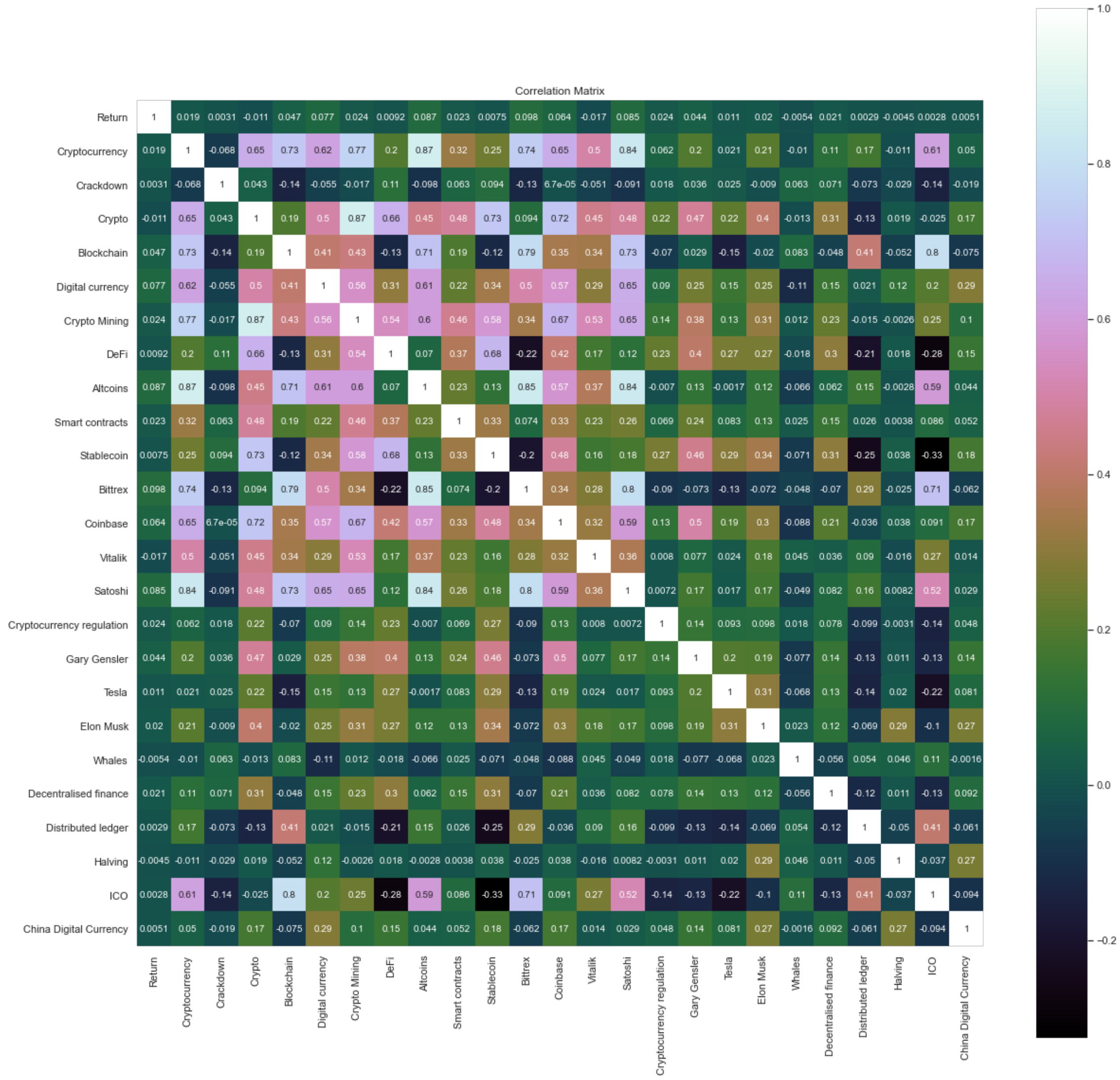


Figure 5: Correlation Matrix of Return vs Google Trends search terms

terms are not correlated with one another with a few exceptions. For example, the search terms ”Cryptocurrency” and ”Satoshi” are highly correlated with a correlation of 0.84. Hence, in order to address the issue of multicollinearity, we decided to run Principal Component Analysis (PCA).

3.2 Statistical Tests

3.2.1 Chow Test

A change in the parameters of the regression model over time could cause the results and forecast of the subsequent model to be unreliable, hence the Chow test was done to identify potential structural breaks in the data set. The selected time period for the test is based on Figure 6.

A potential structural break was observed near 29 June 2020. The Chow test was done based on this period, and a p-value of 0.999 was obtained, hence the null hypothesis that there are no break points in the data is not rejected.

3.2.2 Principal Component Analysis

As mentioned in Section 3.1, some of the search terms have a strong correlation with each other, such as “Cryptocurrency” and “Crypto”. Hence, Principal Component Analysis was conducted to serve two purposes: to reduce the number of dimensions, as well as to reduce collinearity between the search terms.

The first 10 principal components explained 55.14% of the variation. With the 10 principal components, an Ordinary Least Squares (OLS) regression model was run, with the results shown in Figure 7.

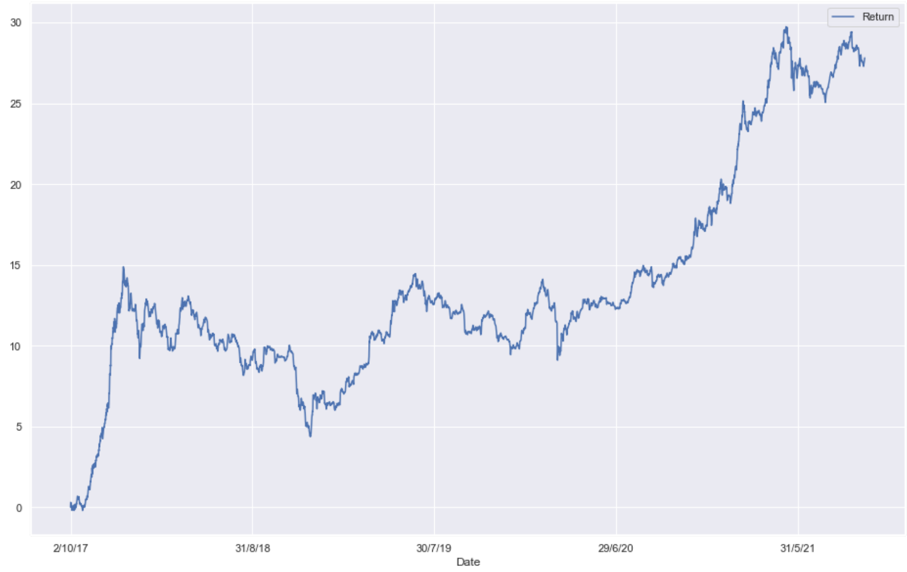
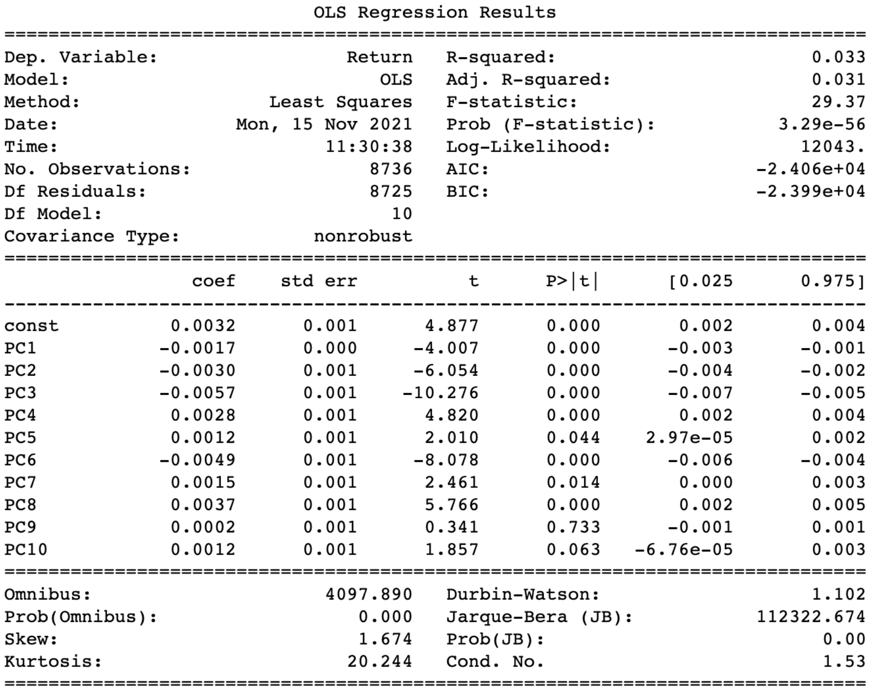
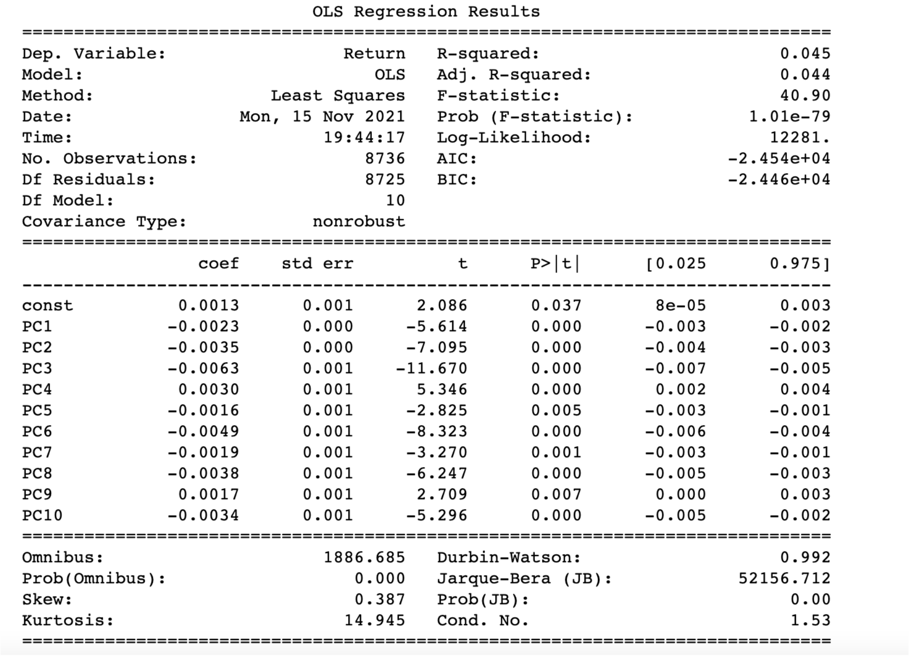


Figure 6: Training Set Cryptocurrency Returns



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                 Return   R-squared:                       0.033
Model:                            OLS   Adj. R-squared:                  0.031
Method:                 Least Squares   F-statistic:                     29.37
Date:                Mon, 15 Nov 2021   Prob (F-statistic):           3.29e-56
Time:                        11:30:38   Log-Likelihood:                  12043.
No. Observations:                8736   AIC:                        -2.406e+04
Df Residuals:                    8725   BIC:                        -2.399e+04
Df Model:                          10                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0032      0.001      4.877      0.000       0.002       0.004
PC1           -0.0017      0.000     -4.007      0.000      -0.003      -0.001
PC2           -0.0030      0.001     -6.054      0.000      -0.004      -0.002
PC3           -0.0057      0.001    -10.276      0.000      -0.007      -0.005
PC4            0.0028      0.001      4.820      0.000       0.002       0.004
PC5            0.0012      0.001      2.010      0.044    2.97e-05       0.002
PC6           -0.0049      0.001     -8.078      0.000      -0.006      -0.004
PC7            0.0015      0.001      2.461      0.014       0.000       0.003
PC8            0.0037      0.001      5.766      0.000       0.002       0.005
PC9            0.0002      0.001      0.341      0.733      -0.001       0.001
PC10           0.0012      0.001      1.857      0.063   -6.76e-05       0.003
==============================================================================
Omnibus:                     4097.890   Durbin-Watson:                   1.102
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           112322.674
Skew:                           1.674   Prob(JB):                         0.00
Kurtosis:                      20.244   Cond. No.                         1.53
==============================================================================
```

Figure 7: First PCA Model



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Return   R-squared:                       0.045
Model:                            OLS   Adj. R-squared:                  0.044
Method:                 Least Squares   F-statistic:                     40.90
Date:                Mon, 15 Nov 2021   Prob (F-statistic):           1.01e-79
Time:                        19:44:17   Log-Likelihood:                 12281.
No. Observations:                8736   AIC:                        -2.454e+04
Df Residuals:                    8725   BIC:                        -2.446e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0013      0.001      2.086      0.037       8e-05       0.003
PC1           -0.0023      0.000     -5.614      0.000      -0.003      -0.002
PC2           -0.0035      0.000     -7.095      0.000      -0.004      -0.003
PC3           -0.0063      0.001    -11.670      0.000      -0.007      -0.005
PC4            0.0030      0.001      5.346      0.000       0.002       0.004
PC5           -0.0016      0.001     -2.825      0.005      -0.003      -0.001
PC6           -0.0049      0.001     -8.323      0.000      -0.006      -0.004
PC7           -0.0019      0.001     -3.270      0.001      -0.003      -0.001
PC8           -0.0038      0.001     -6.247      0.000      -0.005      -0.003
PC9            0.0017      0.001      2.709      0.007       0.000       0.003
PC10          -0.0034      0.001     -5.296      0.000      -0.005      -0.002
==============================================================================
Omnibus:                     1886.685   Durbin-Watson:                   0.992
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            52156.712
Skew:                           0.387   Prob(JB):                         0.00
Kurtosis:                      14.945   Cond. No.                         1.53
==============================================================================
```

Figure 8: Second PCA Model

From Figure 7, the adjusted R-squared tells us that the model explains only about 3.3% of the variation. In addition, the Jarque-Bera(JB) statistic implies that the sample data does not match a normal distribution, which matches the conclusion drawn from the skewness and kurtosis values. The low condition number implies that collinearity does not seem to be a problem with our data due to the Principal Components.

3.2.3 White Test

The White general test for heteroscedasticity was run to check for heteroscedasticity. From the p-value of $1.671e^{-38}$, the null hypothesis that the disturbances are homoscedastic is rejected.

3.2.4 Heteroskedasticity and Functional Form Tests

To solve the above issues, log returns were used and the entire process was repeated. The results from the resulting OLS model are in Figure 8.

In this second model, the adjusted R-squared increased and the omnibus and JB statistic significantly improved. White's test statistic of 7.6506 also improved from the previous value of 5.2816 although the p-value is still close to 0. Therefore, transforming the variable appears to have addressed the issues.

The Ramsey reset test was run to check if the wrong functional form of the explanatory variables was used. The p-value of $6.6462e^{-40}$ suggested that some higher order terms could be included to improve the model.

Further modification of the model was done by adding squared and cubed terms. The process was again repeated to obtain the following results seen in Figure 9.

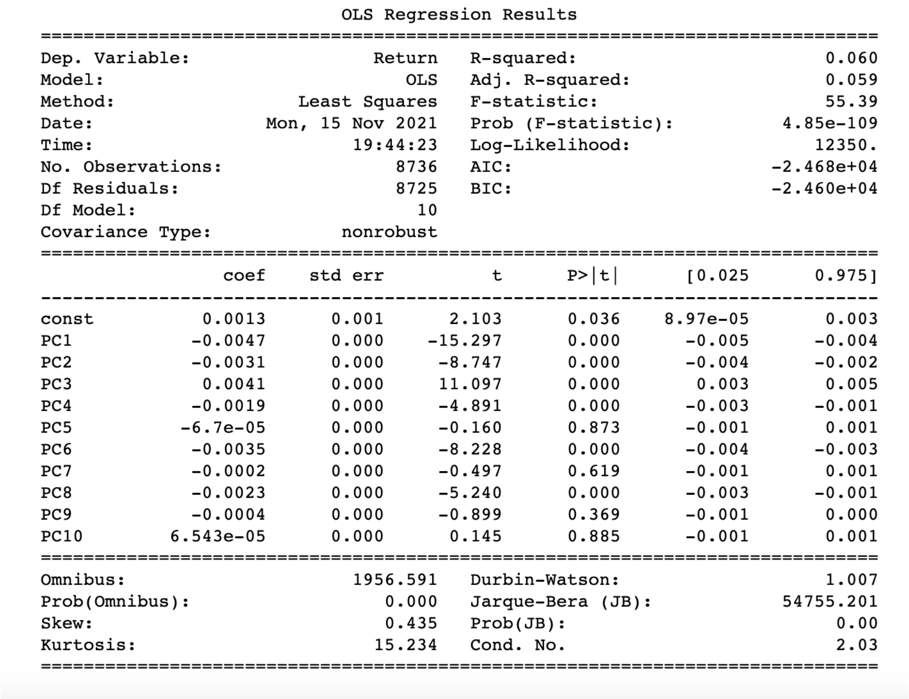
The new Ramsey reset test on the new model was run with a p-value of $1.7475^{e}xp^{-5}$ implying that higher order terms should be included in the model.

Finally, the insignificant coefficient estimates from the model (PC5,PC7,PC9,PC10) was removed, and the residual plot of the final OLS model was obtained (Figure 10).

From this, we can see that the residuals somewhat follow a normal distribution, but with fatter tails than a standard normal distribution, due to the existence of outliers. This is consistent with the skewness and kurtosis values given in the OLS table, which suggest that the it has excess postive kurtosis, and that our data follows a leptokurtic distribution.

This is not surprising, given the nature of cryptocurrencies, where extreme values are more commonly seen compared to other asset classes.

Therefore, Figure 11 is the final model used to predict returns for the test set.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Return   R-squared:                       0.060
Model:                            OLS   Adj. R-squared:                  0.059
Method:                 Least Squares   F-statistic:                     55.39
Date:                Mon, 15 Nov 2021   Prob (F-statistic):          4.85e-109
Time:                        19:44:23   Log-Likelihood:                 12350.
No. Observations:                8736   AIC:                        -2.468e+04
Df Residuals:                    8725   BIC:                        -2.460e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0013      0.001      2.103      0.036    8.97e-05       0.003
PC1           -0.0047      0.000    -15.297      0.000      -0.005      -0.004
PC2           -0.0031      0.000     -8.747      0.000      -0.004      -0.002
PC3            0.0041      0.000     11.097      0.000       0.003       0.005
PC4           -0.0019      0.000     -4.891      0.000      -0.003      -0.001
PC5          -6.7e-05      0.000     -0.160      0.873      -0.001       0.001
PC6           -0.0035      0.000     -8.228      0.000      -0.004      -0.003
PC7           -0.0002      0.000     -0.497      0.619      -0.001       0.001
PC8           -0.0023      0.000     -5.240      0.000      -0.003      -0.001
PC9           -0.0004      0.000     -0.899      0.369      -0.001       0.000
PC10        6.543e-05      0.000      0.145      0.885      -0.001       0.001
==============================================================================
Omnibus:                     1956.591   Durbin-Watson:                   1.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            54755.201
Skew:                           0.435   Prob(JB):                         0.00
Kurtosis:                      15.234   Cond. No.                         2.03
==============================================================================
```

Figure 9: OLS results after inclusion of squared and cubed terms

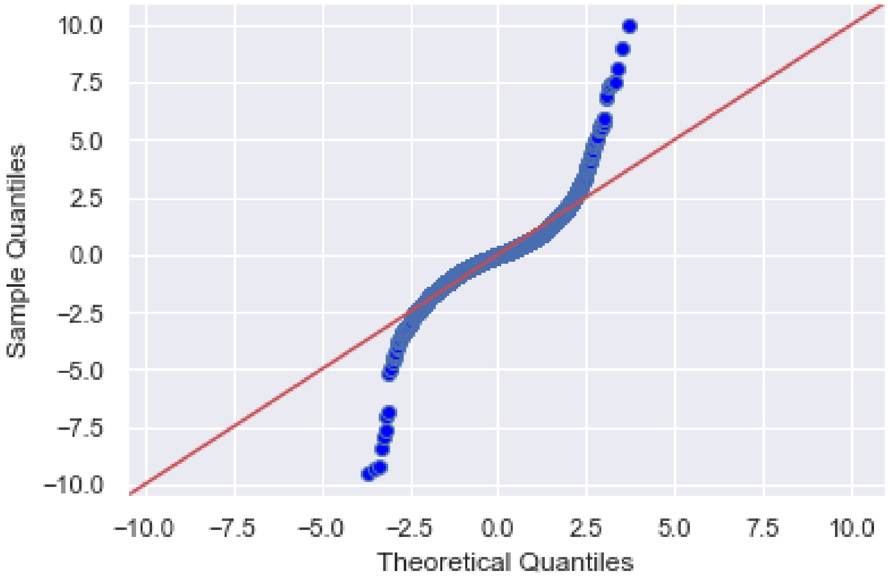
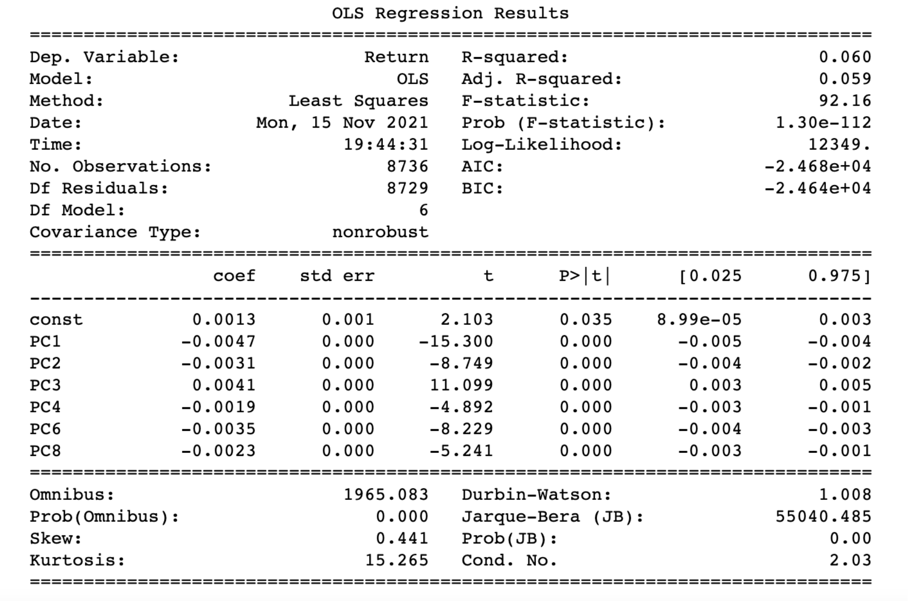


Figure 10: Theoretical Quantiles



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Return   R-squared:                       0.060
Model:                            OLS   Adj. R-squared:                  0.059
Method:                 Least Squares   F-statistic:                     92.16
Date:                Mon, 15 Nov 2021   Prob (F-statistic):          1.30e-112
Time:                        19:44:31   Log-Likelihood:                 12349.
No. Observations:                8736   AIC:                        -2.468e+04
Df Residuals:                    8729   BIC:                        -2.464e+04
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0013      0.001      2.103      0.035    8.99e-05       0.003
PC1           -0.0047      0.000    -15.300      0.000      -0.005      -0.004
PC2           -0.0031      0.000     -8.749      0.000      -0.004      -0.002
PC3            0.0041      0.000     11.099      0.000       0.003       0.005
PC4           -0.0019      0.000     -4.892      0.000      -0.003      -0.001
PC6           -0.0035      0.000     -8.229      0.000      -0.004      -0.003
PC8           -0.0023      0.000     -5.241      0.000      -0.003      -0.001
==============================================================================
Omnibus:                     1965.083   Durbin-Watson:                   1.008
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            55040.485
Skew:                           0.441   Prob(JB):                         0.00
Kurtosis:                      15.265   Cond. No.                         2.03
==============================================================================
```

Figure 11: Final OLS Model

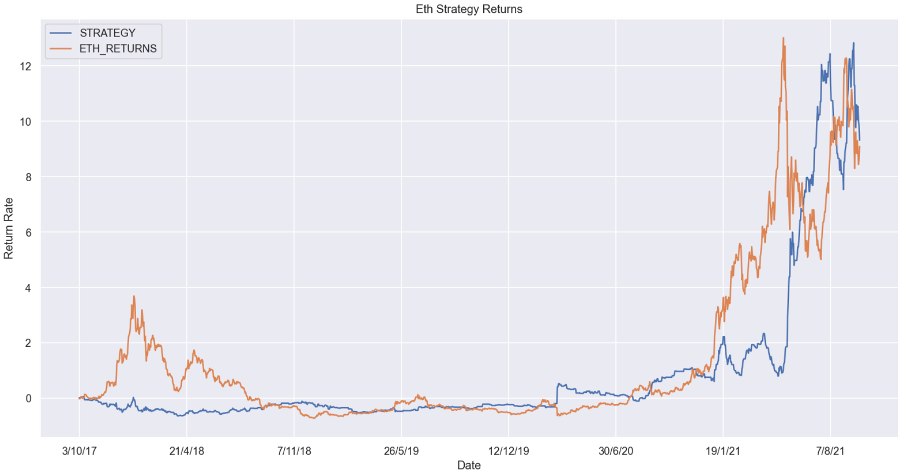


Figure 12: ETH strategy vs returns

4 Trading Strategy and Back-testing

A mean-reversion approach was decided to be implemented as a trading strategy. The fitted model predicted returns of the each of the coins in the test set for the sample period. Subsequently, the 25th and 75th percentiles were chosen as the relative threshold to generate trade signals (i.e., if the trade signal is above the 75th percentile, a short position is taken, and if the trade signal is below the 25th percentile, a long position is taken) in hopes that the predicted log returns will revert to its mean value.

The trading strategy was backtested for the coins in the test set and their individual ”buy-and-hold” positions was used as the benchmark. Figures 12, 13, 14 and 15 are the respective back-testing plots of the strategy performance against the benchmark.

Based on the backtesting results, it is observed that the trading strategy works well for large-cap coins such as Ethereum with a Sharpe ratio of 1.09 and a total return of 932% which exceeded the actual return of 909%.

However, the prediction does not work well for small/medium coins such as Cardano (Sharpe ratio: 0.84 and strategy return: 265%, compared to the actual return of 8,053%), Chainlink (Sharpe Ratio: 1.15 and strategy return: 1613%, compared to the actual return of 6,623%) and Dogecoin (Sharpe ratio: 1.01, strategy return: 2,163%, compared to the actual return of 18,724%) as shown in the performance metrics summary (Table 4).

Hence, the strategy could be profitable for cryptocurrencies that have a bigger market cap

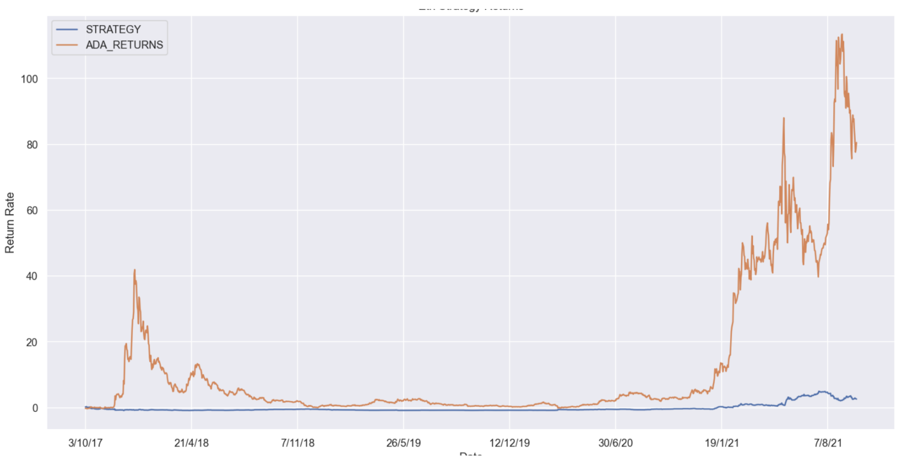


Figure 13: ADA strategy vs returns

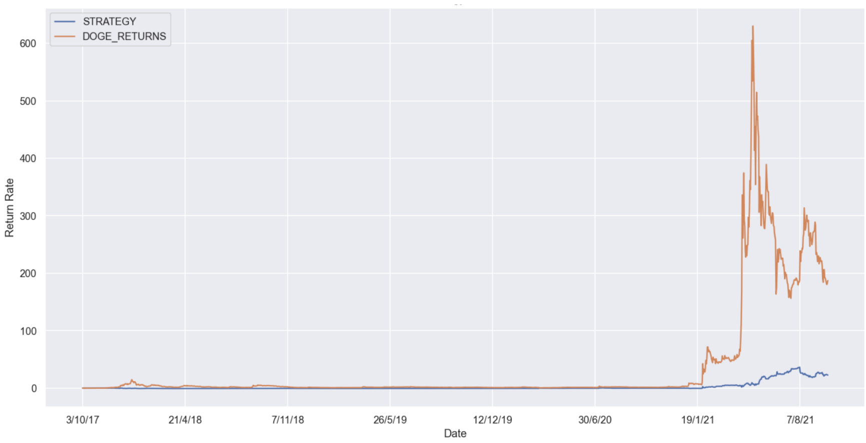


Figure 14: DOGE strategy vs returns

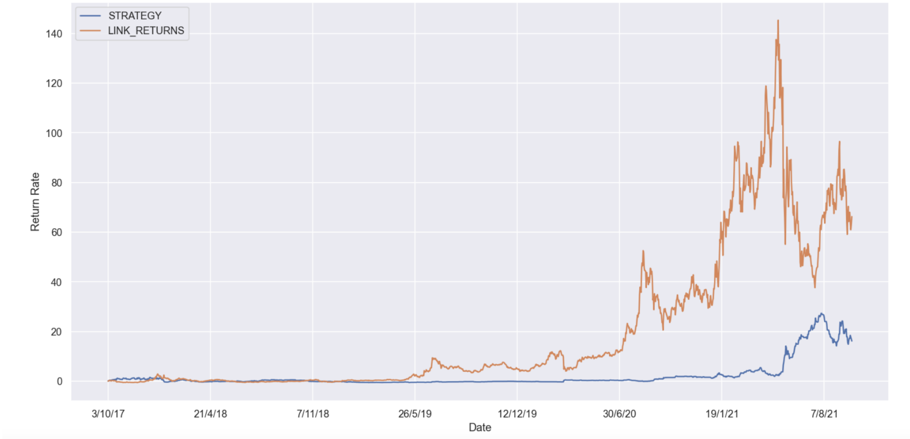


Figure 15: LINK strategy vs returns

	Returns (%)	Strategy (%)	Sharpe Ratio
Ethereum (ETH)	909.0525	932.0092	1.08713
Cardano (ADA)	8053.8330	264.7405	0.84021
Dogecoin (DOGE)	18724.3321	2263.3713	1.01930
Chainlink (LINK)	6623.2533	1613.3220	1.14989

Table 4: Performance Metrics

like Ethereum (18.51%). This could be due to the fact that popular coins have more commercial interest amongst investors. Any news relating to the cryptocurrency space in general would spur interest in these coins and hence more Google searches. Price movement in smaller coins on the other hand might be more affected by specific news. Hence this could lead to a higher correlation between trading of popular coins like Ethereum and Google Trends which makes it more suitable for the trading strategy.

5 Conclusion

5.1 Further Improvements

Firstly, the buy and sell signal threshold can be calibrated. In the current trading strategy, the 25th and 75th percentiles of predicted log return were set as the threshold. This may be a good reference point for buy and sell signals, but it may not be the optimal threshold value if higher returns are to be achieved from the market. By optimizing the threshold values, the trading strategy can be improved, especially for small-cap cryptocurrencies.

Secondly, trades can be made within a shorter time frame such as in hours or minutes. As cryptocurrencies are volatile, there is a possibility that there is a failure to trade at opportune moments that are presented. Worse the investor might be exposed to greater risk if trading was solely based on the last 24-hour prediction which may be deemed irrelevant or outdated, given the speculative nature of cryptocurrencies and its volatility. Moreover, a “take-profit and stop-loss” stance should be introduced in the trading strategy as it would help cap losses if the market moves against us.

Thirdly, more search terms and relevant keywords could be used during the data mining process. The search terms used in this project are limited and may not be sufficient to capture all or the majority of the information required to do the prediction. Hence, this would affect the accuracy of the results in the statistical analysis. Therefore, other platforms such as Reddit could be tapped on to obtain more data to buttress the accuracy of our results. This in turn can help to cross validate the results.

5.2 Conclusion

In conclusion, the usage of data from Google Trends data to generate signals to trade on cryptocurrencies has been explored, and a simplified mean reversion approach has been presented as a trading strategy. Though impressive positive results were achieved especially for ETH, the model does not fit well for small-cap cryptocurrencies. This has prompted several suggestions for the improvement of results as discussed in Section 5.1, involving the selection of an optimal choice of model for prediction, employing reasonable, rational judgement during the data mining process and crafting a more sophisticated trading strategy. As such, the results and discussion presented in this report have hopefully provided insights on how search terms derived from Google Trends can be used to generate positive returns from cryptocurrencies.