Taranis test (extended)

Julia-Suzana Dancu

March 2024

This is the second submission to the data modelling challenge. I continued to work on the dataset provided and tried my best to answer all the questions. I also found some errors in the first document I submitted which I'd like to correct here.

1 Part 1

Two separate plots (see Fig.: 1) on the 5 equities in the daily data were produced, in order to be at scale and better visualised. S&P500 has a continuous upward trend, whereas corn and soybean price are relatively stagnating. Crude oil dropped in 2015 and showed stagnation afterwards, whereas the US dollar index rose slightly in the second half of 2014 and stagnated afterwards.

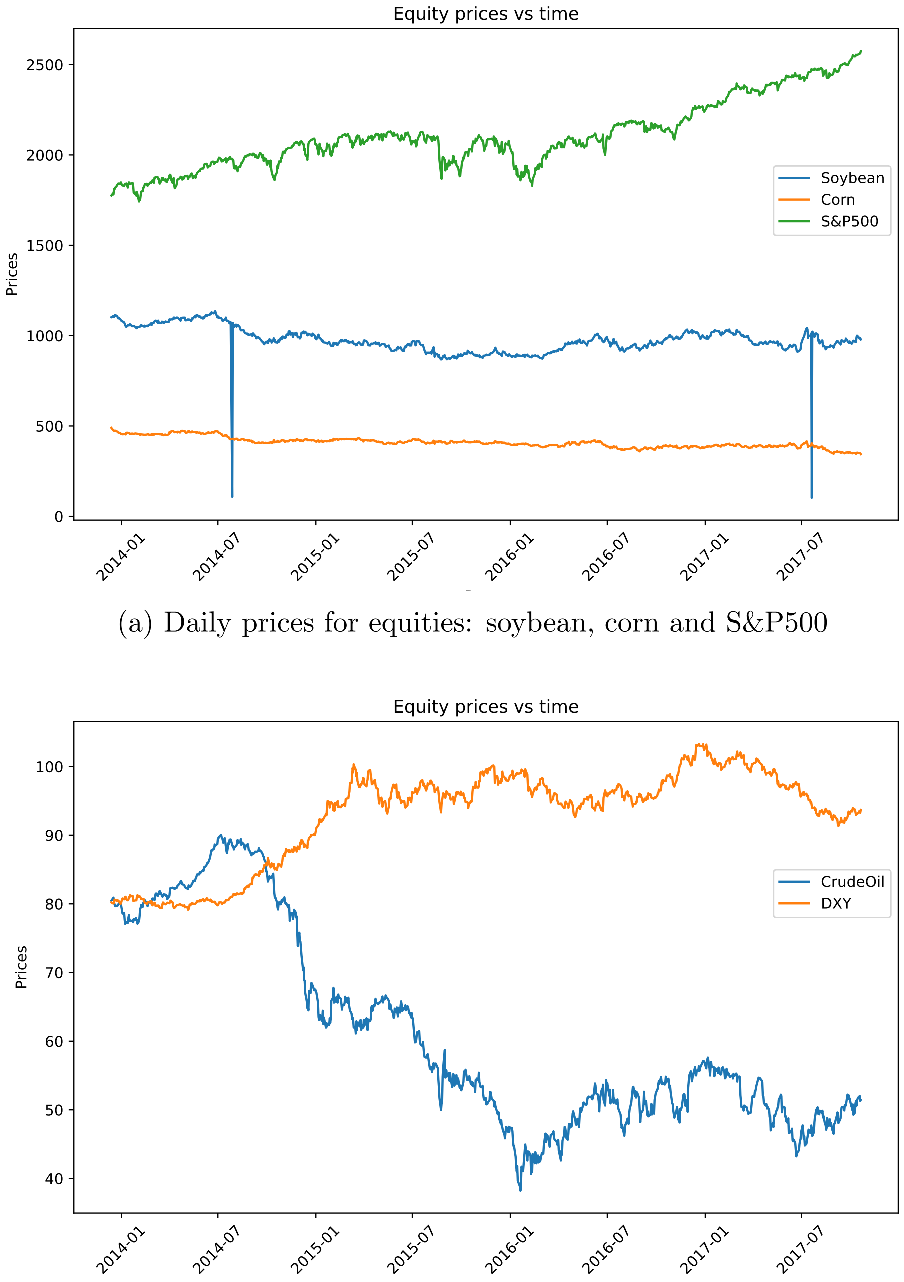
In case of the monthly data on stock to use for the USA (see Fig.: 2), there has been a significant outlier in 2014, but otherwise, it has been relatively stable.

2 Part 2

Three plots on the monthly averaged values for soybean and S&P500 versus stock to use (see Fig.: 3) were produced, seemingly showing little correlation in between. The outlier in the stock to use was removed from the dataset, resulting in less biased correlation measures and regression performance. The Pearson linear coefficients show relatively little (anti)correlation with each other for: S&P500 vs stock to use (0.25) and soybean vs S&P500 (-0.25). Soybean vs stock to use show stronger anticorrelation (-0.59), however. In the previous report I mentioned that the Pearson corr. coefficient might have been biased due to the outliers, which was the case.

3 Part 3

A correlation investigation was performed on the whole available dataset. First of all, the outlier in the stock to use was removed from the dataset, as it highly biased both the linear coefficient evaluation and multilinear regression model. Such linear methods are not robust against dealing with outliers. A series of scatter plots are provided in Fig.: 4 visualising potential (anti)correlation between variables, mostly linear at first sight: oil vs crude oil, corn vs soybean, crude oil vs soybean, crude oil vs S&P500 and US dollar index, S&P500 vs dollar index etc.. In addition, Figs.: 5-7 present the Pearson linear correlation coefficient, Spearman's rank correlation coefficient and the mutual information correlation coefficient, respectively. The Pearson correlation coefficient shows (anti)correlation relation between two variables if its absolute value is higher than 0.5 (roughly



(a) Daily prices for equities: soybean, corn and S&P500

(b) Daily prices for equities: crude oil and DXY

Figure 1: Daily prices for equities (grouped to scale).

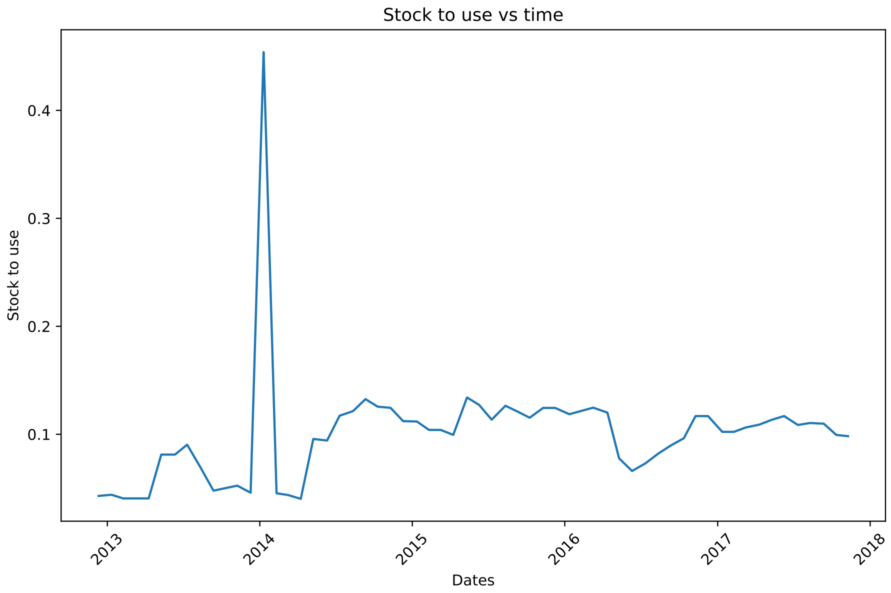


Figure 2: Monthly values for stock to use

speaking). Spearman's rank correlation coefficient is another method to quantify correlation relations between two variables including higher-order features (i.e. non-linear shapes). If its absolute value > 0.5, the two variables show higher order correlation. Similarly, Shannon's mutual information can quantify higher-order correlation, typically for values above ~ 0.3, as we used a normalised version ranging between 0 to 1. The code employed for its evaluation in this project was developed in a previous personal project with a colleague. Details on the mutual information equation are provided in Appendix 8.1.

Multilinear regression was performed on the monthly averaged values of equities and stock to use versus the monthly averaged values of soybean. Data was split 80-20% between training and testing. The obtained mean squared error value shows a reasonable outcome (~ 980) considering the price range of soybean ~ 1000. The mean absolute error is around 25 which is an excellent value for a linear fit with predicted values two orders of magnitude larger. The R-squared score (coefficient of determination) for the multilinear fit is 0.74, which is a good result (the closer it is to 1, the better). The p-values for S&P500, corn and US dollar index are low, so there is confidence in their outcome. However, the stock to use and crude oil fits show lower confidence levels. The outcome is presented in Fig. 8. A correlation plot between the predicted and true soybean price values in the test data set are shown in Fig.: 9.

4 Part 4

A small neural network was developed as an alternative to the multilinear regression algorithm and compared to it. Various architectures were tried out, from single-layer models with various amounts of nodes (4, 8, 16, 32), to two layer models (4×4, 8×4). Since the data set does not have many input variables and entry point, there is no point to use overly large neural networks, as they would easily overfit. Data was split 80-20% between training and testing, as for the previous model. Eventually, the 4×4 dense neural network showed the most robust performance during

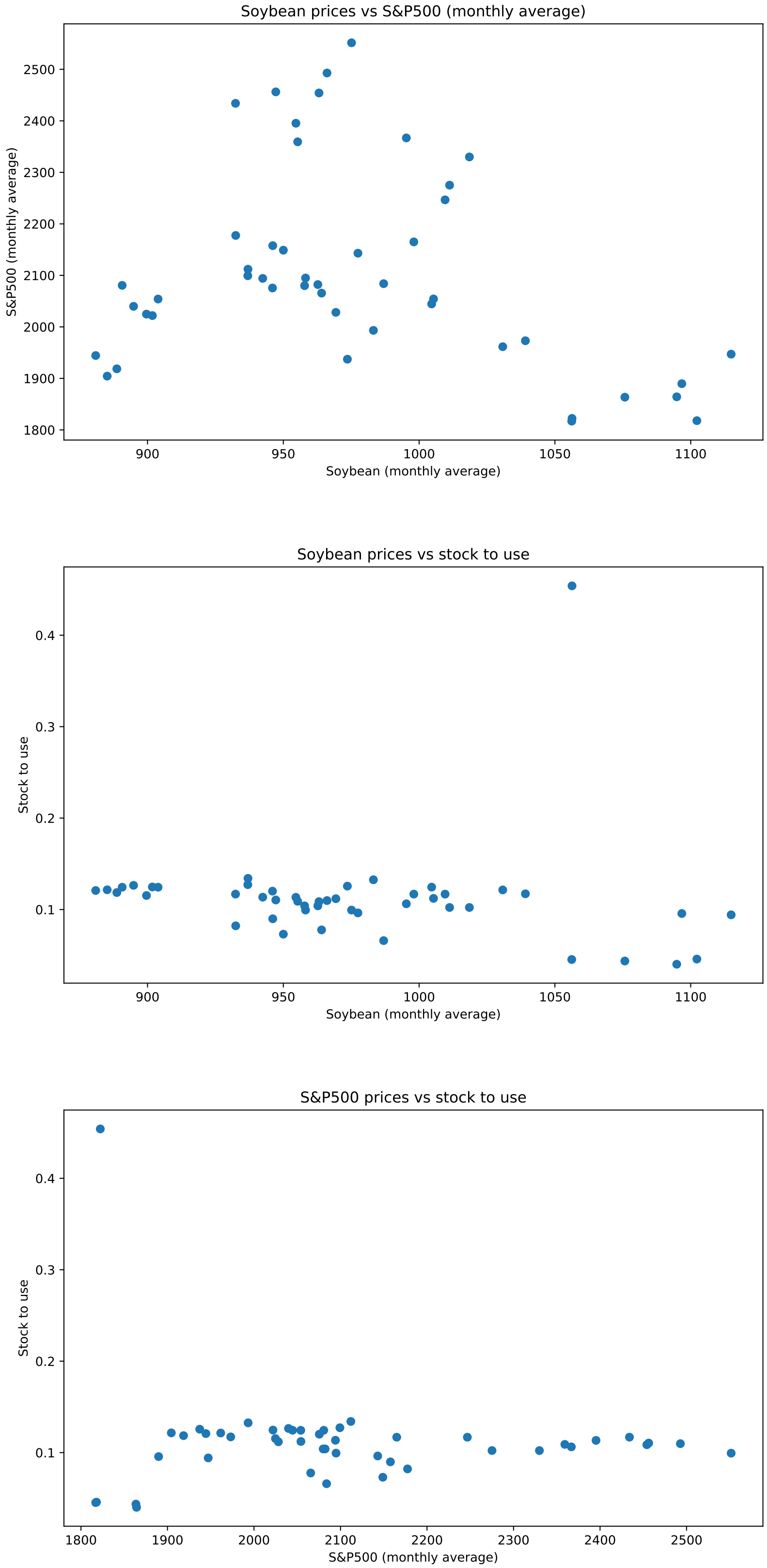


Figure 3: Correlation between monthly equity values.

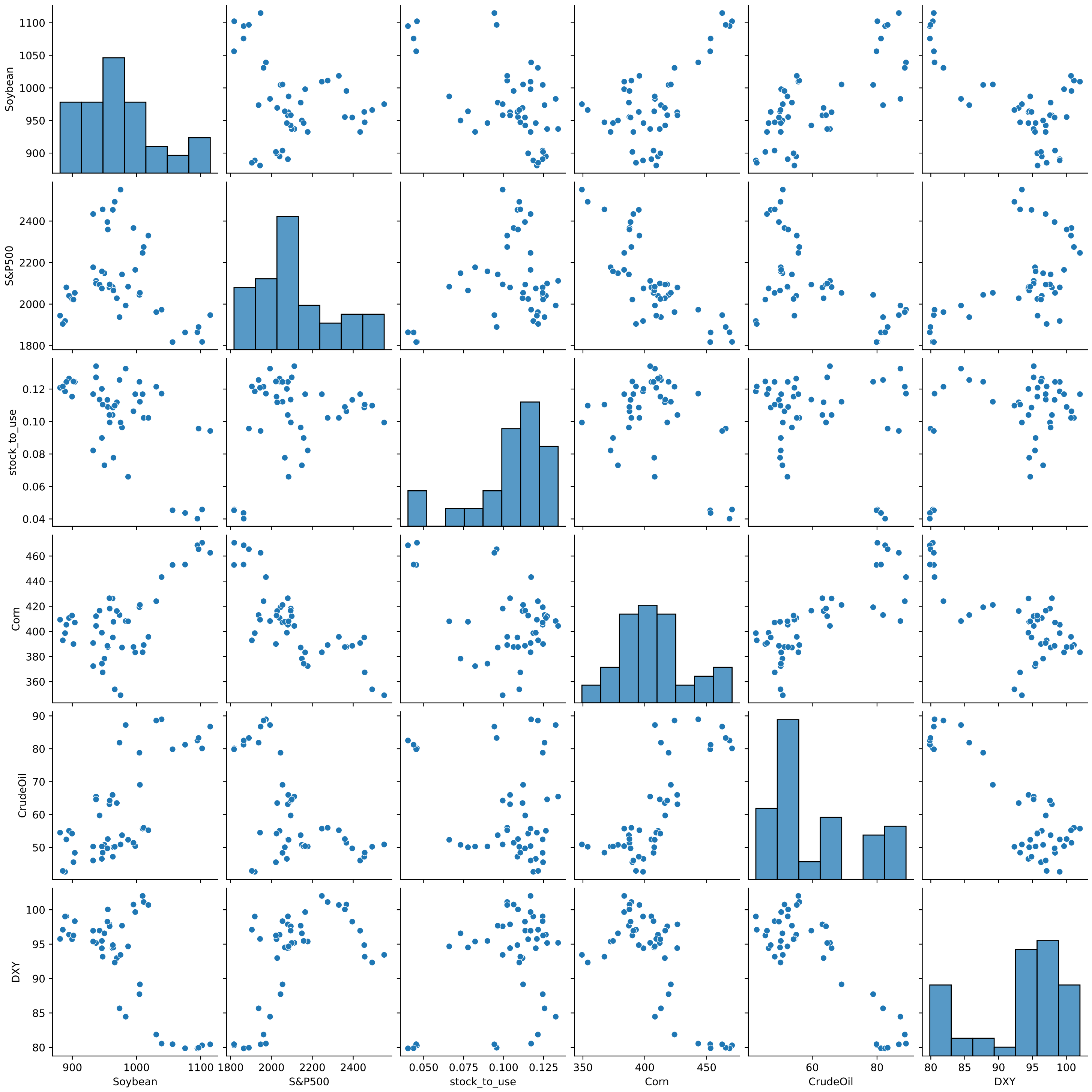


Figure 4: Correlation plots among input features used in the linear regression.

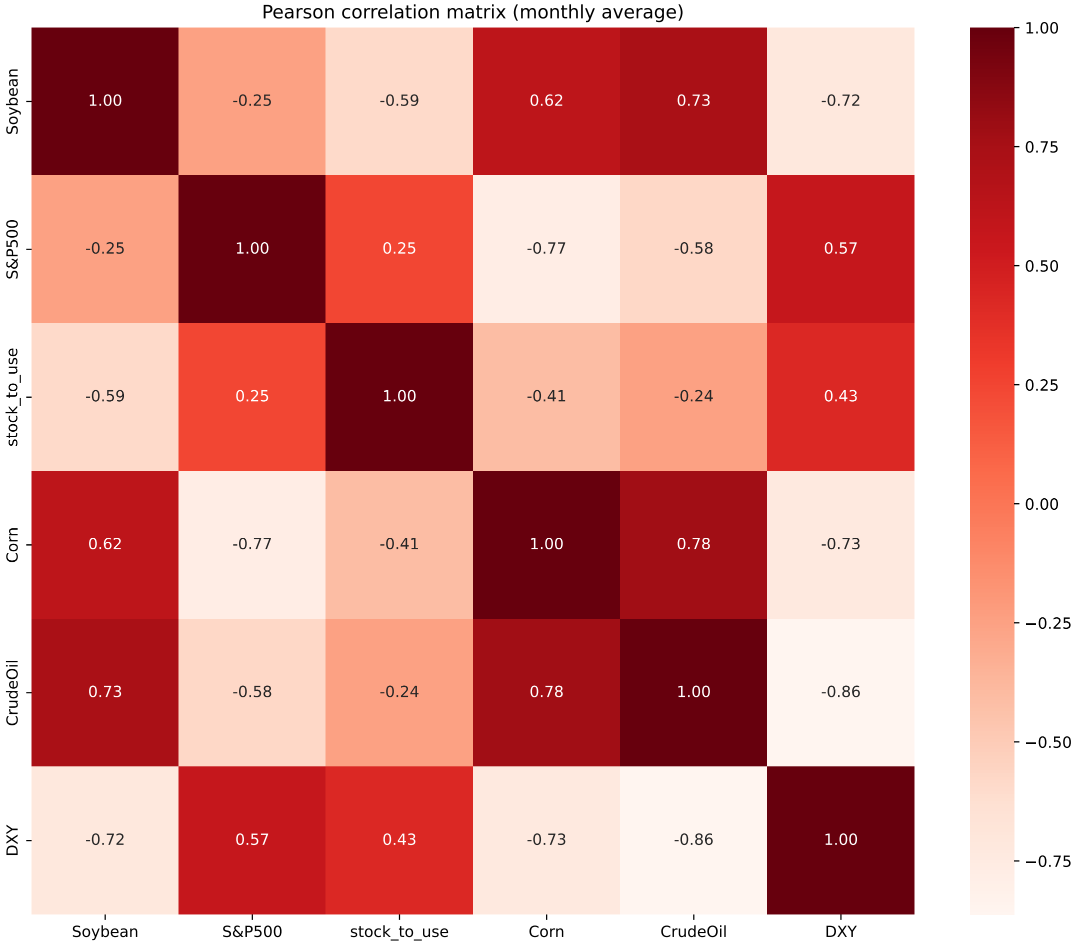
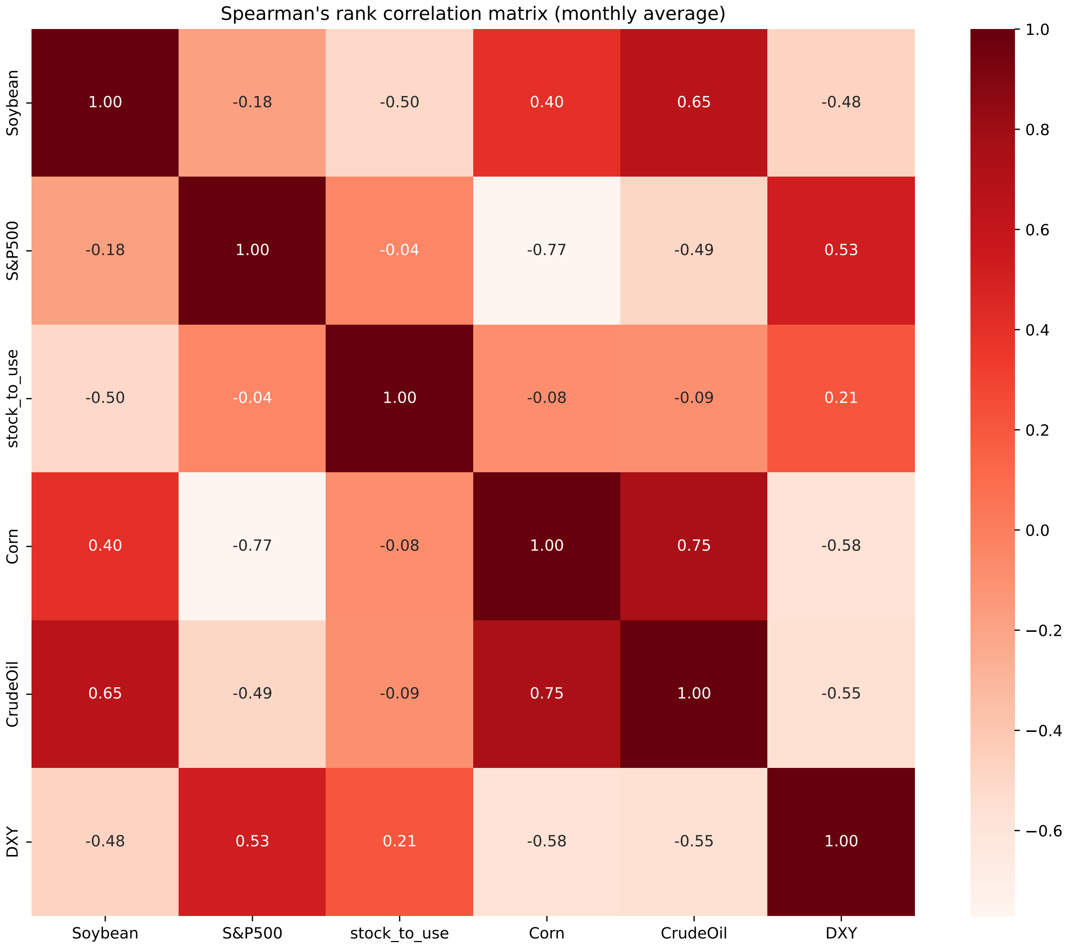


Figure 5: Pearson linear correlation coefficient values among input features used in the linear regression.



Spearman's rank correlation matrix (monthly average)

	Soybean	S&P500	stock_to_use	Corn	CrudeOil	DXY
Soybean	1.00	-0.18	-0.50	0.40	0.65	-0.48
S&P500	-0.18	1.00	-0.04	-0.77	-0.49	0.53
stock_to_use	-0.50	-0.04	1.00	-0.08	-0.09	0.21
Corn	0.40	-0.77	-0.08	1.00	0.75	-0.58
CrudeOil	0.65	-0.49	-0.09	0.75	1.00	-0.55
DXY	-0.48	0.53	0.21	-0.58	-0.55	1.00

Figure 6: Spearman's rank correlation coefficient values among input features used in the linear regression.

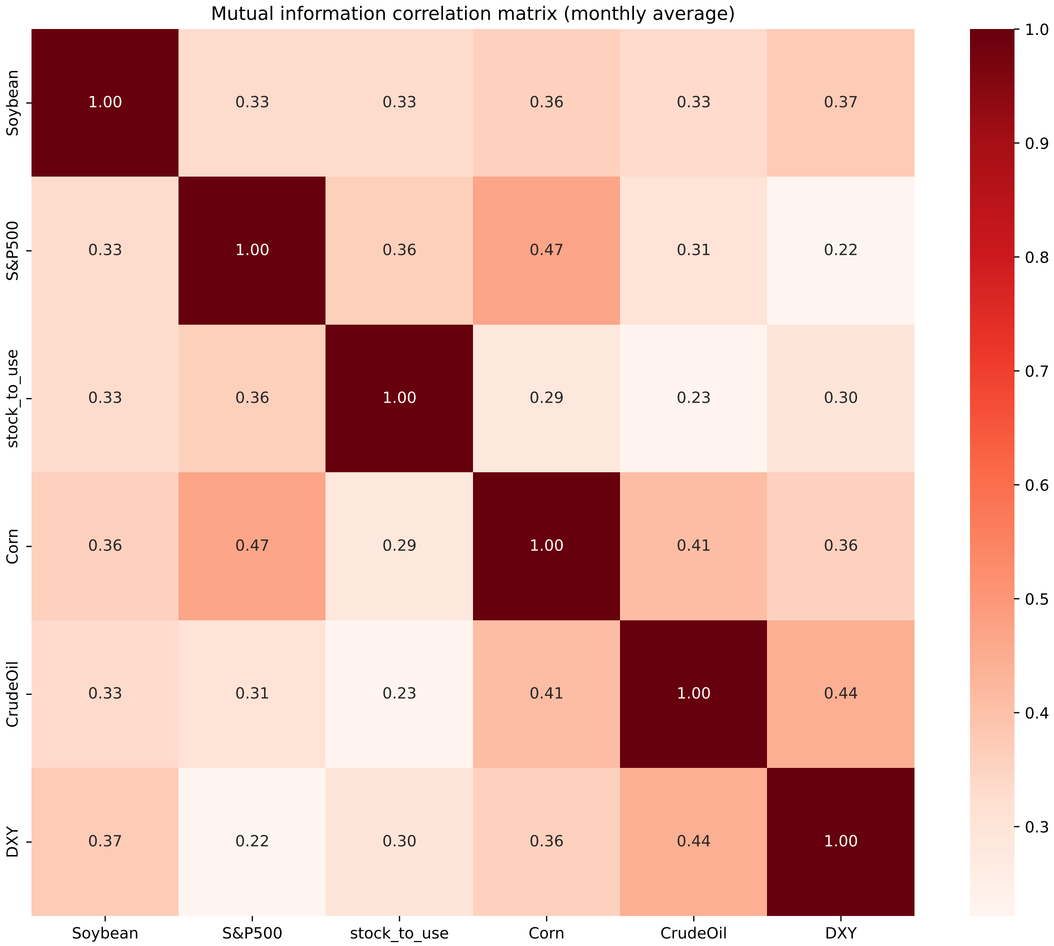
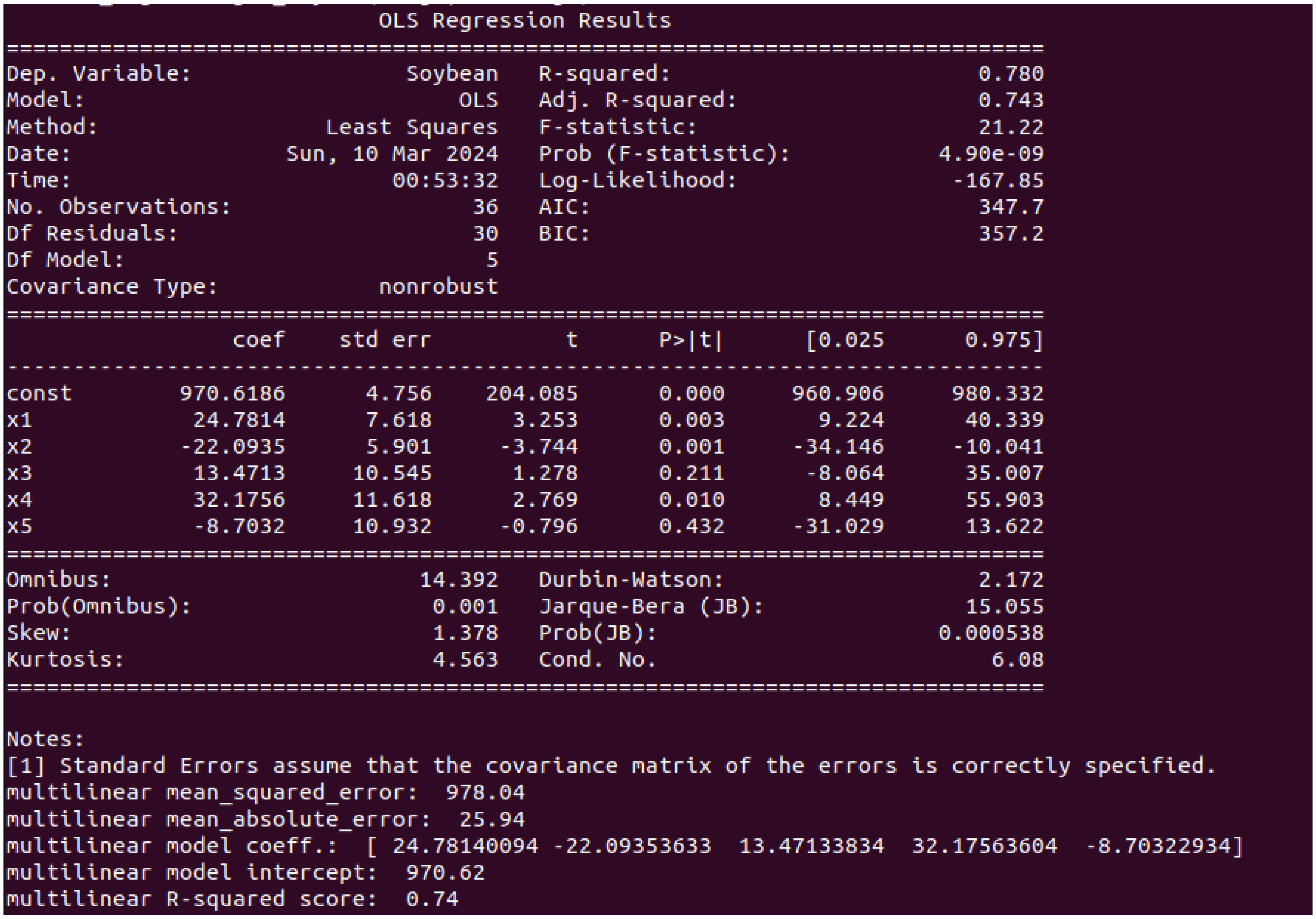


Figure 7: Mutual information correlation coefficient values among input features used in the linear regression.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Soybean   R-squared:                       0.780
Model:                            OLS   Adj. R-squared:                  0.743
Method:                 Least Squares   F-statistic:                     21.22
Date:                Sun, 10 Mar 2024   Prob (F-statistic):           4.90e-09
Time:                        00:53:32   Log-Likelihood:                -167.85
No. Observations:                  36   AIC:                             347.7
Df Residuals:                      30   BIC:                             357.2
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        970.6186      4.756    204.085      0.000     960.906     980.332
x1            24.7814      7.618      3.253      0.003       9.224      40.339
x2           -22.0935      5.901     -3.744      0.001     -34.146     -10.041
x3            13.4713     10.545      1.278      0.211      -8.064      35.007
x4            32.1756     11.618      2.769      0.010       8.449      55.903
x5            -8.7032     10.932     -0.796      0.432     -31.029      13.622
==============================================================================
Omnibus:                       14.392   Durbin-Watson:                   2.172
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               15.055
Skew:                           1.378   Prob(JB):                     0.000538
Kurtosis:                       4.563   Cond. No.                         6.08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
multilinear mean_squared_error:  978.04
multilinear mean_absolute_error:  25.94
multilinear model coeff.:  [ 24.78140094 -22.09353633  13.47133834  32.17563604  -8.70322934]
multilinear model intercept:  970.62
multilinear R-squared score:  0.74
```

Figure 8: Linear regression stats

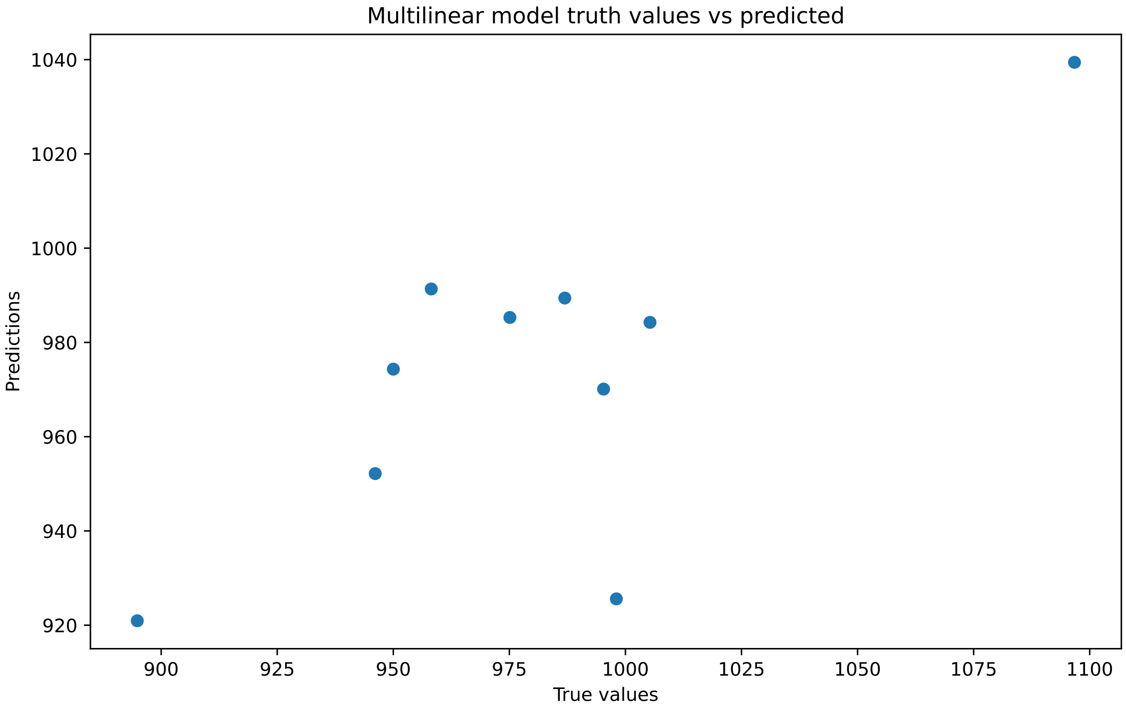


Figure 9: Linear regression predicted vs true soybean prices in test data set.

training and validation epochs (see Fig.: 10), as well as when its performance was evaluated on the test data set. An exponentially decaying learning rate was used (starting at 0.1), as keeping it constant was inducing large fluctuations in the training period. The algorithm was trained for 1000 epochs with the mean absolute error being the loss metric minimised. The mean square error of the algorithm on the test data set was 977.01, the mean absolute error (also test loss) was 27.99 and R-squared score was 0.74. These all match the performance of the multilinear model. Consequently, the conclusion is that the simple neural network model did not show better performance to the baseline multilinear regression model. It is quite surprising, as the correlation studies above showed potential higher-order correlations between several variables (S&P500 and corn, crude oil and dollar index etc.) which the neural network could have expolited in comparison to the linear regression. On the other hand, the data set does not have enough entries (less than 60) for even such small network to fully learn such features.

For longer time frame for the test, more advanced neural networks, such as LSTM could have been used for time-series analysis. However, since not much data was provided (a few hundred entry points), deep neural networks might not be adequate for this specific time series analysis, as it would overfit. Already the small neural network showed overfitting tendencies when its sized was increased by a few nodes.

5 Part 5

I would certainly support the suggestion of the manager on trying neural networks, as mentioned in the section above. Deep neural networks would not be possible to be performed, as they typically

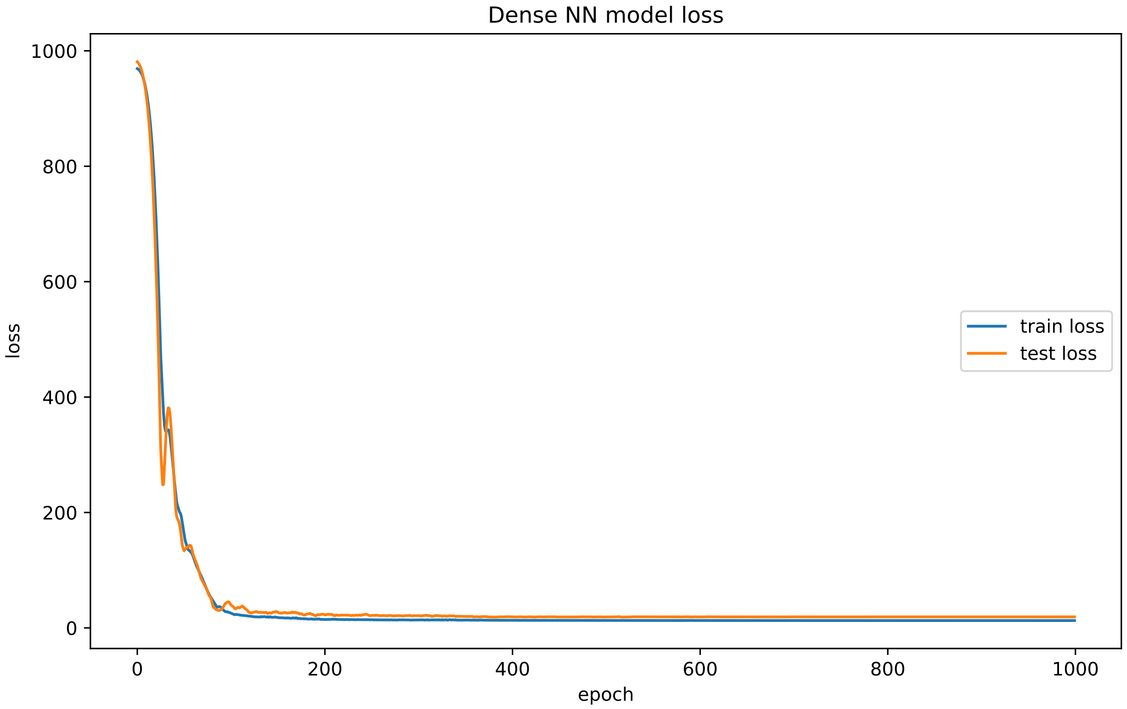


Figure 10: Dense neural network model loss progress during training over epochs.

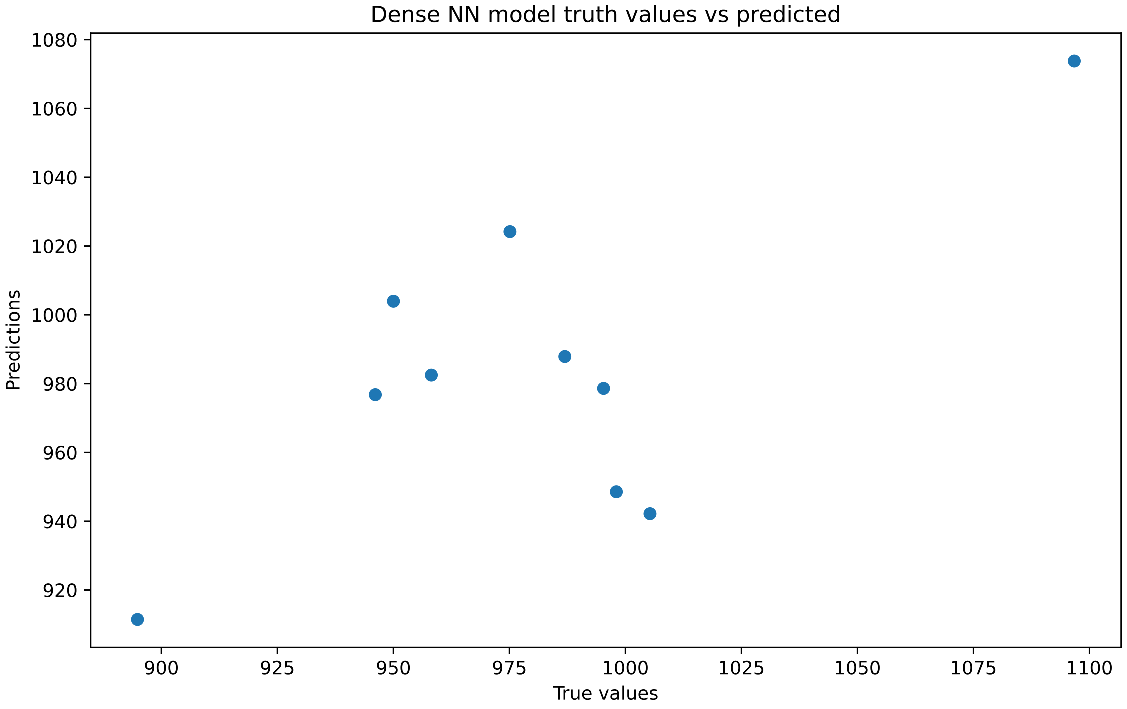


Figure 11: Dense neural network predicted vs true soybean prices in test data set.

use tens of thousands of parameters, whereas the available datasets have only a handful of entries. In order to avoid overfitting and keep model robustness, the number of parameters of employed machine learning models should always scale with the available data and number of variables within the data set.

The conclusion of the section above was that the use of a small-size dense neural networks did not show better performance to the baseline multilinear regression. However, it is always useful to check out alternative models and use linear regression as a baseline cross-check. If bigger datasets could be provided with multiple input features (possibly lower level than just the prices) and entry point, more complex algorithms would probably outperform linear regression.

6 Part 6

I would be wondering why my colleague uses only 3 variables when he has 100 available? Indeed, they have robust performance due to their low p-values but since he has so many more available variables, the rest should be exploited, as well? Or those give worse performance? He should certainly try to use neural networks to explore higher-order correlations in his available data set than linear regression, especially that his dataset is larger than the one available for this test.

7 Part 7

Firstly, both intraday time series were plotted with different scales, comparing their behaviour (see Fig.: 12). Then, the datasets were matched by the time series and plotted again (see Fig.: 13), this dataset being used for further correlation analysis. In addition, a correlation plot was produced showing clear linear correlation between brent and S&P500 prices (see Fig.: 14). The Pearson linear correlation coefficient proves this correlation with 0.86, as well as the mutual corrleation coefficient (0.41) reveals potential higher order correlation.

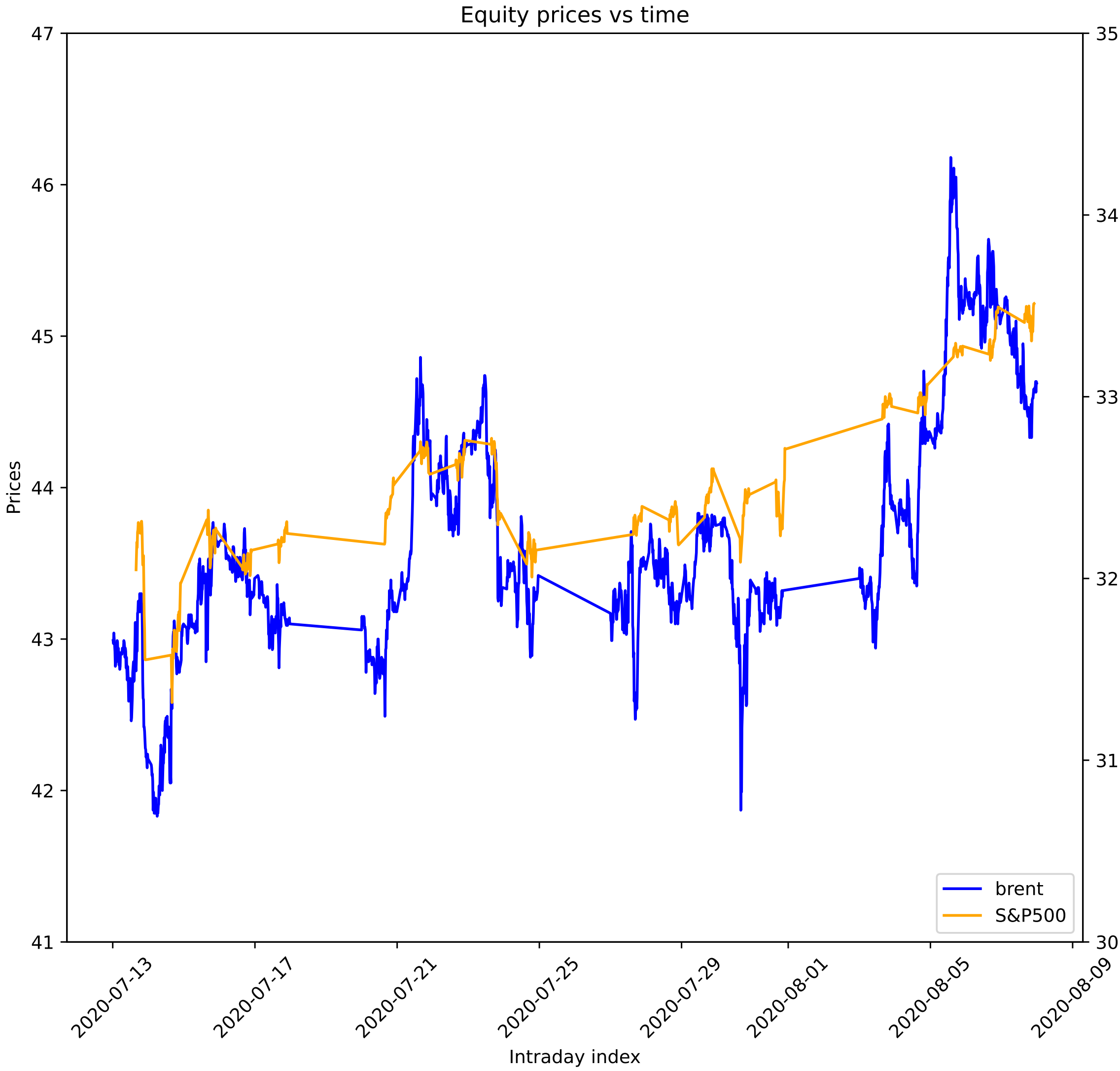


Figure 12: Brent and S&P500 prices over intraday time series.

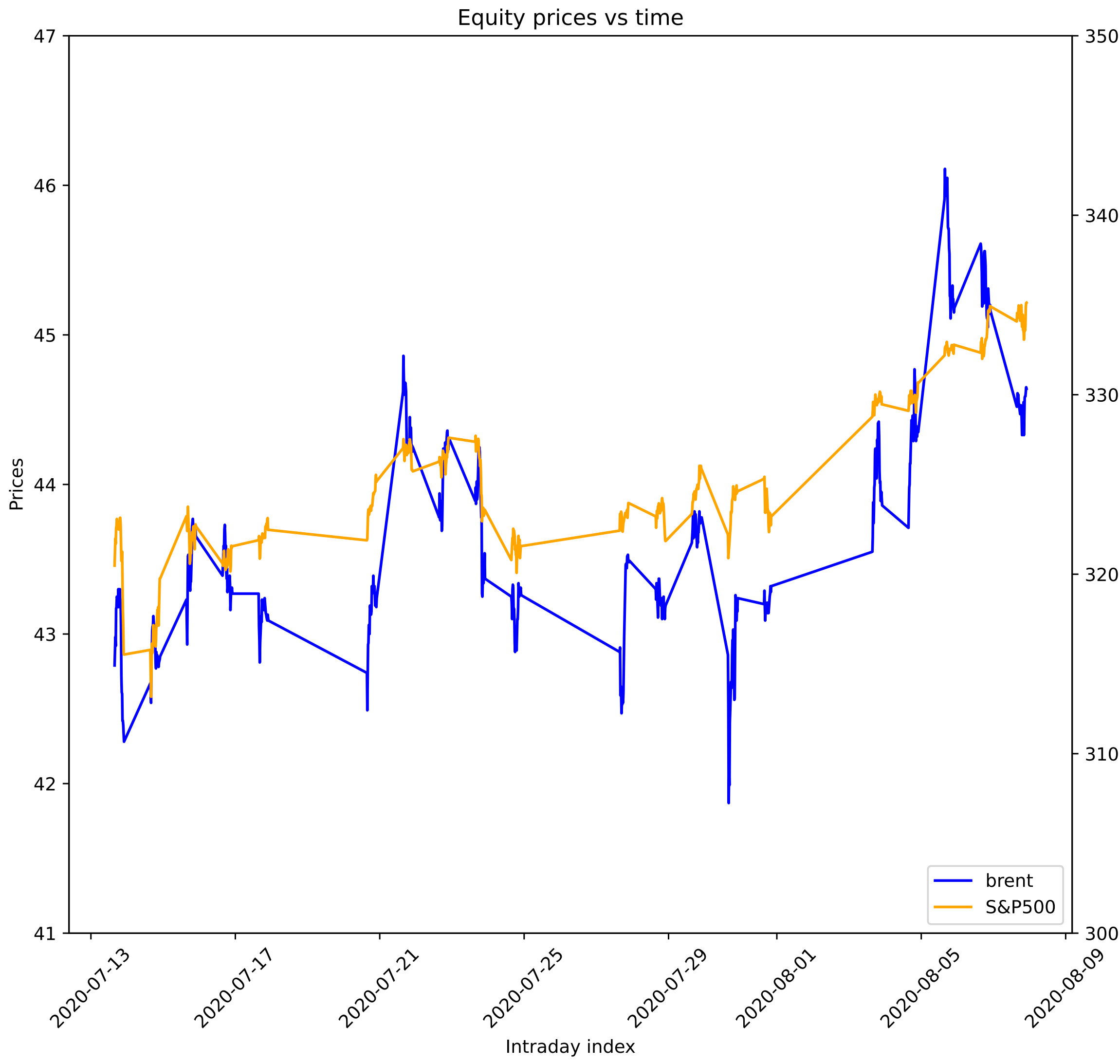


Figure 13: Brent and S&P500 prices over matching intraday time series.

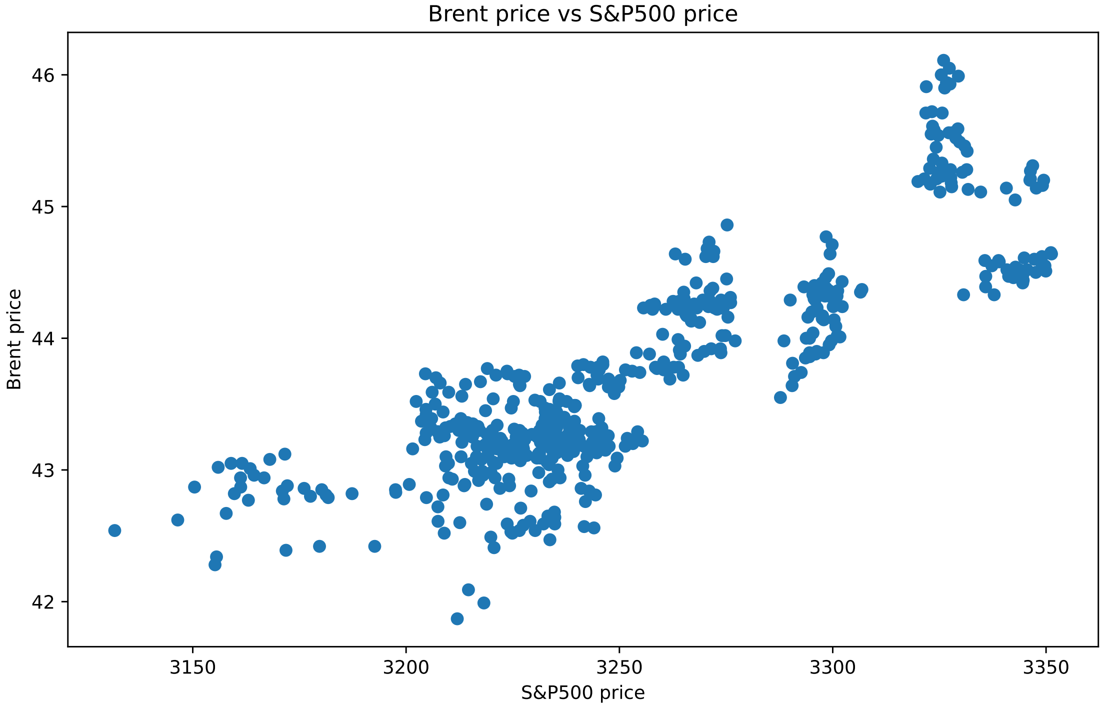


Figure 14: Brent and S&P500 prices showing clear linear correlation.

8 Appendices

8.1 Mutual information correlation coefficient

The mutual Shannon information entropy or M*-score*, is a non-linear measure of statistical dependency between two random variables (observables), X and Y. In the discrete case, this is defined as:

$$M = I(X,Y) = H(X) + H(Y) - H(X,Y) = \sum_i \sum_j f_{X,Y}(x_i, y_j) \log_2 \left(\frac{f_{X,Y}(x_i, y_j)}{f_X(x_i) f_Y(y_j)} \right), \tag{1}$$

where $f(X,Y)$ is the two-dimensional joint probability density mass function and $f(X), f(Y)$ are marginal densities. The function $H(X) = -\sum_i f(x_i) \log_2 f(x_i)$ denotes the Shannon entropy of X. By definition, when the random variables are statistically independent, the joint density factorises $f(X,Y) \equiv f(X)f(Y)$ and the mutual entropy will be zero. This quantity therefore provides a stronger, more generic statement than linear correlation being zero.

Discretised densities of continuous observables are estimated via unit mass normalised histograms. The number of bins in two dimensions is chosen automatically using the Scott rule, and the histogram range per dimension is determined so as to contain 99% of the empirical probability mass to minimise the impact of outliers on binning. The maximum possible M-score value depends on the underlying distributions. To take this into account, a typical additive normalisation is used, which scales the maximum possible value of the M-score to unity:

$$I_N(X,Y) = 2\frac{I(X,Y)}{H(X)+H(Y)}. \tag{2}$$

A simple exact relationship between the unnormalised mutual information, M, and Pearson linear coefficient, ρ, exists, assuming both marginal distributions are normally distributed:

$$M = -\frac{1}{2}\log\left(1-\rho^2\right). \tag{3}$$