A Work Project presented as part of the requirements for the Award of a Master's degree in International Finance from the Nova School of Business and Economics and Universidad de los Andes.

CRYPTOCURRENCY FORECASTING

PAUL NEUPERT

Work project carried out under the supervision of:
Andrés Mora Valencia and Melissa Prado (as co-advisor)

12-09-2023

CRYPTOCURRENCY FORECASTING

Abstract:

This thesis aims to investigate the factors influencing daily cryptocurrency returns and assesses the feasibility of forecasting these returns using a diverse set of 25 globally traded assets as predictors between 2017 and 2022. Inspired by a New York Times article on cryptocurrency bubbles and market volatility, the investigation undergoes several statistical computations such as the Principal Component Analysis and the Complete Subset Regression to observe the data from distinct angles. Notably, the FX-rate USD/CNY and the Japanese NIKKEI225 index emerge as consistently influential predictors within the out-of-sample forecasting periods, however, the significance of the Out-of-Sample R-squared remains low. This research contributes to the understanding of cryptocurrency market dynamics by examining the impact of a wide range of predictors and aligns with recent academic findings regarding the challenges in forecasting cryptocurrencies. Further research should focus on market sentiments regarding potential price movements in the crypto space.

Keywords:

Forecasting, Complete Subset Regression, Cryptocurrency Market, Principal Component Analysis, Trading Strategy, Asset Prices.

This document would not have been possible without the encouragement and numerous advice from the professors Andrés Mora Valencia and Melissa Prado. To them, I express my warmest gratitude for their time, guidance, and patience.

1. Introduction

1.1 Motivation

A New York Times article published in January 2022 questions the existence of a bubble in the cryptocurrency space and the high degree of volatility of these assets (The New York Times, 2022). Unprecedent growth in the previous years and Bitcoin reaching its all-time high of $60.000 recorded on CoinMarketCap have raised doubts about its sustainability (figure 1). Since more and more institutions as well as private investors started deploying capital into exchanges like Binance and Kraken it has become crucial to investigate the drivers that may influence cryptocurrency predictability and their prices.

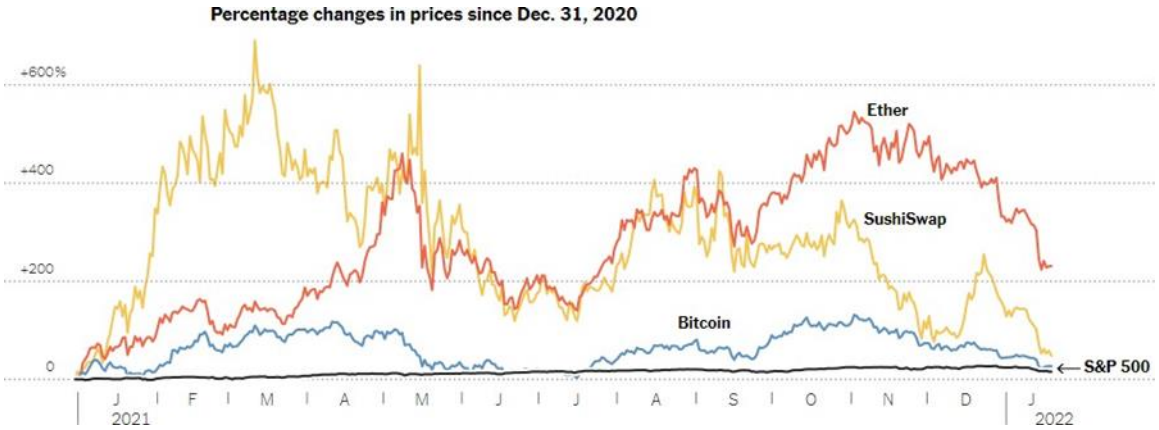

Figure 1 - Percentage changes in cryptocurrency prices since December 31, 2020.

The outbreak of the war in Ukraine accelerated inflation rates and commodity prices drastically and various cryptocurrencies including Bitcoin were hit heavily. Industry experts refer to this extended period of instability and economic downturn as the “crypto winter” (International Monetary Fund, 2022). Comprehending the drivers closely associated with the changes in cryptocurrency returns is essential for hedging risk occurrences and making informed financial decisions.

Besides, the permanent regulatory uncertainty and security breaches from time to time, a diverse spectrum of investment asset classes comes to mind which could affect a diversified

portfolio. War-related agricultural richness of Ukrainian soils and its importance within Europe, the energy-exhaustive coin mining Proof-of-Work concept, global stock indices, U.S.- American export restrictions on certain products altering the FX-market or metal-based commodities including gold which similar as Bitcoin claims to be a “safe haven” in time spans of economic downturns and lack of trust in centralized bank institutions (Nickel, 2023). Gauging the correlation and degree of prediction power of these assets in connection to a selected cryptocurrency portfolio is a key objective of the conducted research.

All the selected cryptocurrencies were publicly listed before December 2017 and the approximately 5-year period has been chosen due to its statistical significance in terms of length, its relevant pre-Covid timeframe and the fact that Bitcoin had a significant price rally in late 2017 reaching from $3.000 to a peak of close to $20.000 (Higgins, 2017).

1.2 Literature review

The degree of forecasting of cryptocurrency returns and their diverse factors is a subject that has piqued the interest of scholars and financial experts around the globe. Theoretical papers by Obeng & Attor (2022) and other reputable authors have examined the topic from different angles thoroughly. The following paragraphs will briefly summarize the current status quo and outline how the current analysis differs from recent results and applied methodologies.

Zargar and Kumar (2022) explore the impact of geopolitical events on the cryptocurrency market, primarily focused on the COVID-19 pandemic. They analyze pre- and post-COVID data to understand how it affected the volatility, liquidity and returns of several cryptocurrencies using a panel data regression approach. As the timeframe covered in this research also falls between 2017 and 2022, the paper provides beneficial insights on the impact of the virus outbreak. This thesis additionally comprises affects from other geopolitical events such as the war in Ukraine and the increasing rivalry between China and

the United States of America regarding trade and global influence. Thus, the contribution of this research will reveal whether the crypto market becomes more robust as these external perturbations occur.

Briola & Aste's (2022) paper investigates the correlation of logarithmic price returns among 25 liquid cryptocurrencies using Minimum Spanning Tree and Triangulated Maximally Filtered Graph. Their results reveal the evolutionary process of currencies and their mutual dependencies as well as a decreasing trend in overall correlation referred to as the "Epps Affect". The latter shared academic insight from the authors supports the decision of constructing a portfolio consisting of 10 liquid cryptocurrencies in this research, rather than focusing on each of them individually. Novel insights will be gained in terms of dependencies towards other financial assets.

Klose's research compares Bitcoin as well as Ethereum to gold using a GARCH-related model, investigating inter alia whether these currencies have properties connected to a store of value. Finally, both cryptocurrencies are pronounced to be more dependent on exchange rates of developed countries than gold (Klose, 2022). The core innovation of this work is further comparing cryptocurrencies to other existing liquid assets in the market, not just gold using different statistical techniques. The returns of gold will particularly be compared to the portfolio returns during extreme events.

Lastly, Yavuz et al. (2022) studies the market linkages between cryptocurrencies and conventional assets including stocks, bonds, the US-dollar representing the FX-market and gold using the Granger Causality Test. Their results are valuable since they reveal a two-way causality between the cryptocurrency and the bond market. This academic work extends the observation of causalities and potential linkages by including commodity asset classes.

These papers are relevant to the research as they contribute to a better understanding of the

cryptocurrency market, its drivers, properties, and interconnections. Seeing the employment of different methodologies, including statistical, machine-learning, and deep-learning approaches, garners significant understandings which will be further developed in the thesis.

2. Applied methodologies

The methodological approach of the underlying thesis is trifold:

i. *Performing an Ordinary Least Squares regression* – this parametric technique is considered widely used in econometrics and finance research (Greene, 2007). It assumes a certain function form of the relationship between the independent and dependent variables (Montgomery et al., 2012). Moreover, it provides the statistical significance of those coefficients (Holmes et al., 2017).

ii. *Performing a Principal Component Analysis* – as a multivariate technique its academic relevance comes from the identification of patterns and the reduction of the dimensionality of complex datasets (Jolliffe, 2002). Since 24 different predictor variables are present in the analysis, the application of this method aims to simplify the dataset and capture the most important information influencing the performance of the constructed cryptocurrency portfolio (Jolliffe & Cadima, 2016).

iii. *Performing a Complete Subset Regression* – a non-functional regression technique or variable selection technique intending to find the best line through all possible subsets of the data ranked by a goodness-of-fit-criterion (Elliott et al., 2013). Further, several 6-month rolling windows with a training period of 1-year will be used to measure the actual forecasting quality (Elliott et al., 2015).

Novel compared to other recent academic works is the inspection of the datasets from three completely different angles with the objective of making financially informed decisions in

the cryptocurrency market. These methodologies can be applied sequentially and bring an appropriate degree of sophistication to the research process.

3. Data characteristics

3.1 Data curation and asset classes

a) Cryptocurrency portfolio

The selected cryptocurrency portfolio includes 10 currencies which data has been collected from Yahoo Finance from December 2017 to November 2022 resulting in 1779 observation.

#	Crypto asset	Abbreviation	Average weight in portfolio	Asset's purpose in crypto space
1	Bitcoin	BTC	0.621	Digital currency, store of value
2	Ethereum	ETH	0.285	Smart contracts and decentralized apps
3	Litecoin	LTC	0.01	Peer-to-peer currency
4	MIOTA	IOTA	0.001	Internet-of-Things currency
5	Dogecoin	DOGE	0.018	Digital currency, peer-to-peer currency
6	Chainlink	LINK	0.006	Decentralized oracle network
7	Stellar	XLM	0.004	Cross-border payments
8	Monero	XMR	0.005	Cross-border payments
9	XRP	XRP	0.033	Cross-border payments
10	Cardano	ADA	0.017	Smart contracts

Figure 2 - Description crypto assets.

Based on figure 2, it is observable that BTC and ETH hold most of the weight (market capitalization) with around 90% of the entire portfolio. Further, given the defined purpose, the point can be made that the projects are rather classified as assets instead of pure currencies, e.g., as ETH provides a platform for the development of decentralized applications and the enablement of smart contracts. These 10 assets have been selected due to its existence on major cryptocurrency exchanges before 2018 as well as its popularity within the community and market capitalization. The annualized performance of the constructed portfolio as well as the individual cryptocurrencies during the investigated time

frame is depicted in figure 3 with the average revolving slightly above 0 and most assets peaking in 2020. The returns for 2017 have been excluded in the chart due to its short time span.

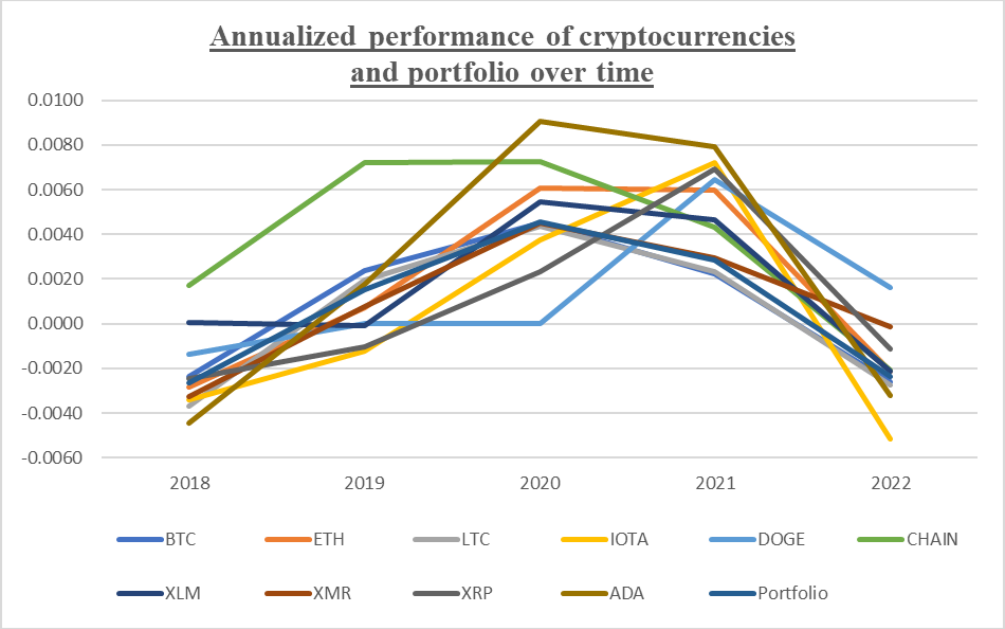

Figure 3 - Annualized performance of cryptocurrencies and portfolio over time.

b) Predictors

The data used in the study consists of the daily closing prices of 25 variables (figure 4) during the beforementioned period. The analysis performed within the thesis considers several asset classes as aforementioned in the introduction. Gold and other well-known commodities are prominently selected by researchers such as Klose (2022) and Yavuz et al. (2022) seeking commonalities with Bitcoin. Other investigators like Czeczeli & Vilonya (2022) have been including foreign stocks and the FX-market to explore and measure the comparisons towards external perturbations. The underlying thesis comprises the most common economic indicators found in scientific papers, however, focuses on a full-scope analysis of all different types of commodities available in the market.

In fact, the asset classes differ in FX-rates to the US-dollar including the Euro, the British

pound sterling, the Japanese yen, the Australian dollar, and the Chinese yuan. Based on Investopedia (Lee, 2023), these are considered the most traded currencies worldwide and the Chinese yuan was also selected due to the trade war during the Trump-presidency.

Further, the selected stock indices contain the Standard & Poor's 500 (500 largest US-based companies), the German DAX 40 (40 largest German-based companies), the Japanese NIKKEI 225 (225 largest Japan-based companies) and the British Finance Times Stock Exchange Index 100 (100 largest UK-based companies). Further, the investigation was appended by a South American index – the Colombian COLCAP 25 to reduce any type of selection bias.

The incorporated commodities mainly traded in future markets, constituting the core novelty of this analysis, can be split into four categories – grains, softs, metals, and energy. The only excluded asset class in terms of commodities is livestock e.g., cattle, sheep, and swine, due to the liquidity and its growing controversial characteristic.

Asset class	Assets	Amount of assets included
FX-rate to USD	Euro, British pound sterling, Japanese yen, Australian dollar, Chinese yuan	5
Stock indices	Standard and Poor's 500, DAX40, Nikkei225, Financial Times Stock Exchange 100 Index, COLCAP25	5
Grains	Corn, Wheat, Rough Rice, Soybean	4
Softs	Cocoa, Coffee, Cotton, Lumber, Rubber	5
Metals	Silver, Copper, Gold	3
Energy	Crude Oil, Brent crude oil, Natural gas	3
Total		25

Figure 4 - Data classification predictors.

3.2 Descriptive statistics

In general terms, the data seems reasonable since all asset classes have a mean and median value of around 0 (see figure 5). In terms of annualized returns Natural Gas achieved the best

performance on average while also being one of the most volatile assets.

Asset	Ann. Return 2018	Ann. Return 2019	Ann. Return 2020	Ann. Return 2021	Ann. Return 2022	Volatility	Max.	Min.
USDEUR	0.0001	0.0001	-0.0002	0.0002	0.0005	0.0038	0.0267	-0.0156
USDGBP	0.0002	-0.0001	-0.0001	0.0000	0.0007	0.0049	0.0435	-0.0293
USDJPY	-0.0001	0.0000	-0.0001	0.0003	0.0010	0.0039	0.0266	-0.0215
USDAUD	0.0003	0.0000	-0.0002	0.0002	0.0006	0.0073	0.1695	-0.1287
USDCNY	0.0001	0.0000	-0.0002	-0.0001	0.0005	0.0023	0.0160	-0.0181
SP500	-0.0002	0.0007	0.0006	0.0007	-0.0007	0.0113	0.0938	-0.1198
DAX40	-0.0005	0.0007	0.0002	0.0005	-0.0006	0.0112	0.1098	-0.1224
NIKKEI225	-0.0006	0.0011	0.0011	-0.0010	-0.0001	0.0243	0.1626	-0.5133
FTSE100	-0.0003	0.0003	-0.0003	0.0004	-0.0001	0.0093	0.0905	-0.1087
COLCAP25	-0.0003	0.0006	-0.0002	0.0000	-0.0004	0.0115	0.1328	-0.1503
Corn	0.0002	0.0002	0.0007	0.0007	0.0008	0.0142	0.0641	-0.1739
Wheat	0.0005	0.0004	0.0005	0.0006	0.0011	0.0179	0.2178	-0.1068
RoughRice	-0.0003	0.0008	0.0001	0.0005	0.0007	0.0138	0.1030	-0.2590
SoyBean	-0.0002	0.0002	0.0009	0.0001	0.0005	0.0112	0.0664	-0.1050
Cocoa	0.0008	0.0002	0.0002	0.0000	-0.0002	0.0154	0.1218	-0.0852
Coffee	-0.0006	0.0008	0.0002	0.0017	-0.0008	0.0177	0.1003	-0.0863
Cotton	-0.0001	-0.0001	0.0004	0.0011	-0.0012	0.0152	0.0701	-0.2388
Lumber	-0.0006	0.0007	0.0027	0.0015	-0.0030	0.0295	0.3070	-0.3348
Rubber	-0.0003	0.0004	0.0003	0.0004	-0.0014	0.0116	0.0910	-0.1220
Silver	-0.0002	0.0004	0.0014	-0.0002	-0.0005	0.0160	0.0930	-0.1165
Copper	-0.0005	0.0002	0.0007	0.0007	-0.0009	0.0118	0.0518	-0.0669
Gold	-0.0001	0.0005	0.0007	-0.0001	-0.0004	0.0079	0.0595	-0.0499
CrudeOil	-0.0006	0.0010	-0.0111	0.0014	0.0010	0.0830	0.2467	-3.0597
BrentCrude	-0.0005	0.0007	0.0018	0.0013	0.0011	0.0363	0.6683	-0.4339
NatGas	0.0002	-0.0006	0.0010	0.0015	0.0033	0.0341	0.4648	-0.2595

Figure 5 - Economic properties of predictors.

The quartiles and maximums as well as minimums reveal non-normal behaviors. Moreover, most of the predictors demonstrate a positive kurtosis and a negative skewness, meaning distributions with heavier tails and longer left tails (extreme values) than a normal distribution (see figure 17 in appendix).

The commodity “Crude Oil” stands out with the highest kurtosis of 1066.975, while the FX-rate USD/AUD exhibits the highest positive skewness of 3.980. Additionally, the FX-rate

“USDJPY” has the lowest kurtosis among the listed variables with 8.870 and the Japanese NIKKEI225 stock index ranks lowest skewness with -5.367 during the almost 5-year long period.

The only outlier, regarding the minimums, identified is crude oil, which circumstances will be explained in the following sentences. In April 2020 (18.04.2020 – 22.04.2020), the price of crude oil briefly turned negative, which means that sellers were paying buyers to take oil off their hands, rather than the other way around. This was an unprecedented event in the history of oil markets, and it was mainly due to the collapse in demand caused by the COVID-19 pandemic and a supply glut.

As lockdowns and travel restrictions were implemented around the world to contain the spread of the virus, demand for oil dropped sharply. At the same time, major oil-producing countries such as Saudi Arabia and Russia engaged in a price war and flooded the market with excess supply, which further depressed prices.

The combination of low demand and oversupply led to a situation where oil storage facilities were filling up quickly, and some traders were left with no place to store the excess oil they had purchased. This prompted panic selling, which in turn caused prices to drop precipitously, and briefly dip into negative territory.

3.3 Correlations among cryptocurrencies

Furthermore, a correlation analysis has been performed to analyze dependencies among the different cryptocurrencies excluding the constructed portfolio since the portfolio weights have already been described. Figure 6 exhibits the results in a condensed form.

On a holistic view, most of the occurrences can be seen between a positive correlation of 0.5 to 0.6 signaling that there is a strong dependence within the market. Unlike the correlation

matrix among the predictors attached in the appendix, not a single negative correlation can be found. Litecoin and Ethereum depict the highest correlation with 0.8113, while the Dogecoin and Cardano show the lowest correlation with 0.1645. While the correlations among the predictors are rather connected to trade agreements, geographic proximity and economic ties of the respective countries, the crypto market is highly dominated by Bitcoin, quite speculative in nature and driven by market sentiment especially in social media. More detailed correlation matrices are attached to the appendix both for the cryptocurrencies (figure 18) as well as for the predictors (figure 19-23).

Range	Occurrences
(0.1 to 0.2)	6
(0.2 to 0.3)	12
(0.3 to 0.4)	0
(0.4 to 0.5)	16
(0.5 to 0.6)	34
(0.6 to 0.7)	12
(0.7 to 0.8)	8
(0.8 to 0.9)	2

Figure 6 - Correlation results cryptocurrencies.

4. In-sample results and discussion

4.1 Analysis - Ordinary Least Squares

The analysis focuses on lagging the predictor variables by up to three days to find the "best" forecast version measured by the adjusted R-squared and the Bayesian Information Criterion (BIC). Further, the dependent variable – the portfolio will also be lagged by one and two days to check for autocorrelation within the data. The investigation further exhibits the variables with a P-value significance below 0.05 (figure 7). The latter is considered valuable for portfolio hedging purposes and to narrow down the number of predictors in the following steps. Below the regression specifications have been listed for lagging the data once keeping

in mind the portfolio structure and the predefined period.

$$Portfolio_t = \beta 0 + \beta 1 * Predictor1\_t - 1 + \beta 2 * Predictor2\_t - 1 + \ldots + \beta 24 * Predictor24\_t - 1 + \varepsilon$$

$$\begin{aligned} Portfolio_t = {} & -0.4452 * USDEUR\_t + 0.6051 * USDGBP\_t + -0.0202 * USDJPY\_t + 0.2672 * USDAUD\_t \\ & + -0.0106 * USDCNY\_t + 0.5826 * SP500\_t + 0.4124 * DAX40\_t + -0.0665 \\ & * NIKKEI225\_t + -0.1234 * FTSE100\_t + -0.0635 * COLCAP25\_t + 0.0109 * Corn\_t \\ & + 0.0218 * Wheat\_t + 0.0014 * RoughRice\_t + 0.0177 * SoyBean\_t + -0.0852 * Cocoa\_t \\ & + -0.001 * Coffee\_t + -0.0347 * Cotton\_t + -0.0098 * Lumber\_t + 0.1448 * Rubber\_t \\ & + 0.1848 * Silver\_t + 0.0163 * Copper\_t + -0.0419 * Gold\_t + -0.0139 * CrudeOil\_t \\ & + 0.021 * BrentCrude\_t + -0.0382 * NatGas\_t + \varepsilon \end{aligned}$$

Figure 8 reveals that the "best" results are being obtained by not lagging the explanatory variables. This version captures around 7% of the variance in terms of the adjusted R-squared. Generally, all the results in terms of adjusted R-squared are quite low and a negative BIC is also generally not a sign of a strong model adequacy. Nevertheless, lagging the response variables twice results in the lowest BIC score with -6196. For the predefined p-value significance of 0.05 the FX-rates for the US-dollar to the British pound sterling and to the Australian-dollar, the S&P 500, the DAX40, the NIKKEI225 and the commodity Brent Crude are included among the "best" version. The unlagged version being the strongest also indicates no substantial delay in communication between the markets, while the cryptocurrency market is "open" during the weekends as well.

The density plot of the residuals in figure 9 allows for further statistical interpretation of the OLS results:

i. *Normality:* Density plot follows a roughly symmetric, bell-shaped distribution. Meaning that the residuals are approximately normally distributed.

ii. *Skewness and Kurtosis:* The shape of the density plot of the residuals reveals a slight asymmetric (skewness) and peaking behavior (kurtosis). To be more specific, the slightly longer left tail indicates negative skewness and the relatively peaked distribution characteristic a high kurtosis.

iii. *Outliers:* Unusual and extreme values in the residuals can be identified on the

density plot. These outliers reflect the anomalies and extreme event occurrences within the timeframe.

iv. *Model Adequacy:* Although the density plot depicts that the residuals deviate only to a lower extent, the differences suggest that the model may not adequately capture the underlying patterns in the data.

Asset	Coefficients	P-Value	Significant?
Intercept	0.0130	0.163	No
USDEUR	-0.0584	0.868	No
USDGBP	-0.7462	0.004	Yes
USDJPY	-0.1771	0.527	No
USDAUD	-0.3336	0.027	Yes
USDCNY	0.2588	0.542	No
SP500	-0.2930	0.009	Yes
DAX40	0.4253	0.010	Yes
NIKKEI225	0.3082	0.000	Yes
FTSE100	-0.1888	0.326	No
COLCAP25	-0.0477	0.628	No
Corn	-0.0336	0.705	No
Wheat	0.0136	0.829	No
RoughRice	-0.0367	0.593	No
SoyBean	0.0759	0.468	No
Cocoa	0.1002	0.113	No
Coffee	0.0229	0.681	No
Cotton	-0.0927	0.164	No
Lumber	0.0117	0.719	No
Rubber	0.0738	0.381	No
Silver	0.1420	0.157	No
Copper	0.0589	0.529	No
Gold	-0.1820	0.351	No
CrudeOil	-0.0070	0.574	No
BrentCrude	0.0630	0.030	Yes
NatGas	0.0109	0.695	No

Figure 7 – Asset coefficients and their significance.

Mode	Adjusted R-squared	BIC	P-value significance below 0.05
OLS regular	0.071	-6332.000	USDGBP, USDAUD, SP500, DAX40, NIKKEI225, Brent Crude
Explanatory variables lag 1	0.052	-6438.000	USDGBP, SP500, DAX40
Explanatory variables lag 2	0.002	-6197.000	DAX40, FTSE100, USDAUD
Explanatory variables lag 3	0.004	-6196.000	/
Response variable lag 1	0.022	-6235.000	Nikkei225, FTSE100, COLCAP25
Response variable lag 2	-0.001	-6188.000	Coffee

Figure 8 - OLS results.

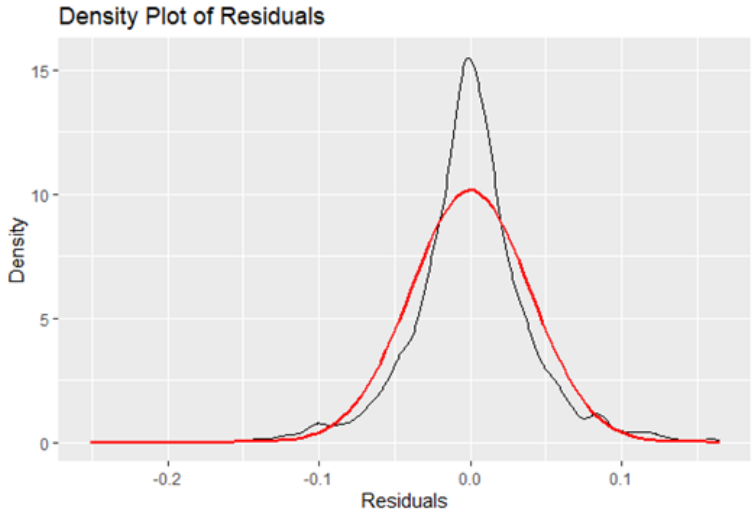

Figure 9 - Density plot of residuals.

4.2 Analysis - Principal Component Analysis

In the next step a PCA was conducted to reduce dimensionality and further explore the data structure. Since the calculations of the thesis are programmed in R, the "prcomp" function will be applied which uses the covariance matrix of the data by default to perform PCA and thus capture the relationships between the variables (Kassambara, 2017).

A crucial step in PCA is selecting the desired minimum number of PCs based on a pre-defined threshold, such as 60% or 80% (Kassambara, 2017). However, the result depicted in

figure 11 reveals a lack of common variation and multiple PCs within an approximate threshold of 60%. The 1st principal component (PC1) only captures 14% and the 2nd principal component (PC2) 10% of the total variation. This tendency is also depicted within the scree plot in figure 10. Although no obvious "elbow point" can be identified, one could argue that the eigenvalues start leveling off after the third or fourth component, meaning decreasing returns regarding the explained variance. Additional visualization regarding the scores is provided in figure 24 in the appendix.

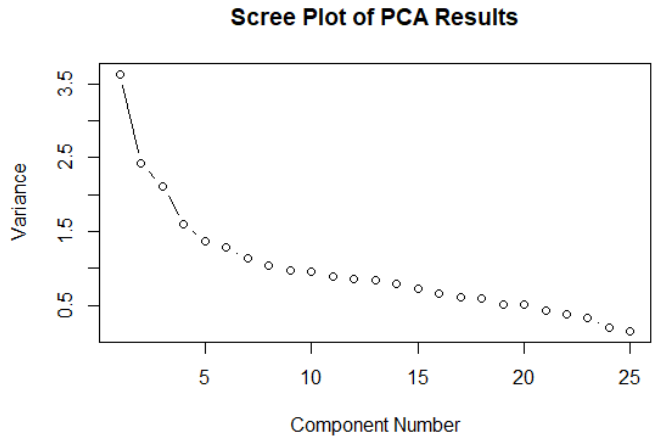

Figure 10 - Scree Plot of PCA results.

Due to the low obtained variance in the initial PCs, instead of using a threshold, a regression including the first three PCs was performed subsequently. The following key results have been acquired from the regression:

i. The coefficient for PC1 is statistically significant based on a p-value of 2.65e-05.

ii. The coefficient for PC2 is highly statistically significant with a p-value of 2.41e-10.

iii. Intercept and third PC are not statistically significant based on the p-value.

iv. Generally, quite low adjusted R-squared of 0.03011 explaining a quite small portion of the variation of the data. The F-statistic of 19.59 and the overall p-value of 1.707e-12 indicate that the model is overall statistically significant.

Component	Captured variance	Cumulated variance
PC1	0.145	0.145
PC2	0.097	0.242
PC3	0.084	0.326
PC4	0.064	0.390
PC5	0.055	0.445
PC6	0.051	0.496
PC7	0.046	0.542
PC8	0.041	0.583
PC9	0.039	0.622

Figure 11 - Captured variance of initial principal components.

After analyzing the loadings of the first two, significant PCs, it can be concluded that (i) the highest positive loadings from PC1 are coming from the FX-rates "USD/AUD" (0.1049) and "USDGBP" (0.0745), (ii) the highest negative loadings from PC2 are obtained from the stock indices DAX40 (-0.3651) and FTSE100 (0.3706), (iii) the highest positive loading for PC2 is coming from the commodity "Rubber" with 0.1026 and (iv) the highest negative loadings for PC2 are obtained from the FX-rates "USD/EUR" (-0.5185) and "USDGBP" (-0.5076).

Overall, although just a small percentage of the variance was captured by these two PCs, the implications from the loadings towards the research question are that these predictors indicate a stronger influence on the portfolio. In particular, the most prominent asset class emerging from the analysis is Foreign Exchange.

Despite the dimensionality reduction achieved through PCA, it's important to clarify that the outcomes of this analysis serve as a foundation for the subsequent variable selection process. However, these PCA results will be viewed as stand-alone findings that offer initial guidance and understanding of the relationships among the variables.

4.3 Analysis - Complete Subset regression

In this step the "best", unlagged, version obtained from the OLS process is being used to find

the strongest performing subsets. For the performance measurement the adjusted R-squared, the BIC and Mallow's Cp are being used as information criteria to consider the results from different kinds of angels. Further, the variables included in these best performing subsets are filtered as well as interpreted and the model will be re-estimated using the best subset to verify its robustness.

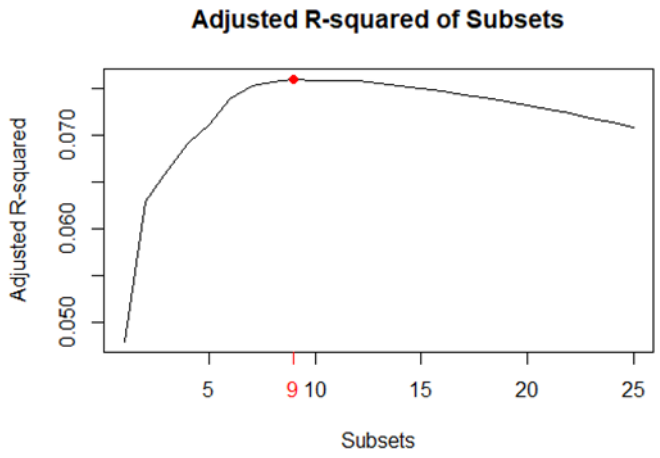

Figure 12 - Best subset based on the adjusted R-squared.

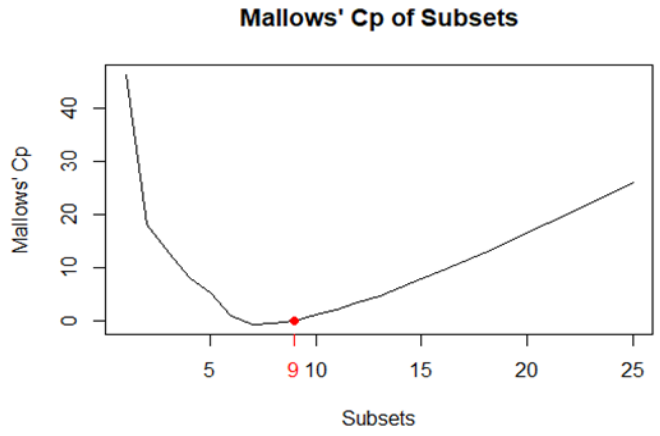

Figure 13 - Best subset based on Mallow's Cp.

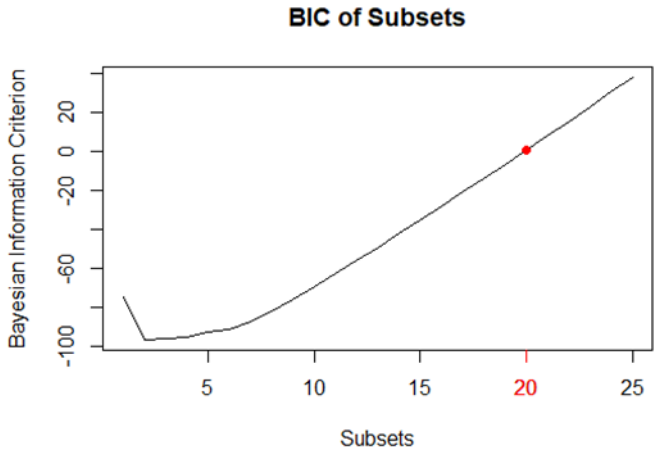

Figure 14 - Best subset based on BIC.

Figures 12, 13 and 14 illustrate the different dimensions of the possible results. While for the adjusted R-squared the optimal value is the highest, for Mallow's Cp and BIC it would be the smallest greater than zero. In Summary, regarding the adjusted R-squared and Mallow's Cp 9 out of 25 variables are included within the best subset (figure 15), while for the BIC 20 out of 25 variables form the best subset. Since USDGBP, USDAUD, SP500, DAX40, NIKKEI225, Cocoa, Cotton, Silver, BRENT CRUDE, are included in all these best subsets, it can be concluded that they are deemed most influential regarding the performance of the portfolio. Thus, the subset containing 9 variables will be re-estimated in the following step.

Indicator	Value	Variables included
Adjusted R-squared	0.076	USDGBP, USDAUD, SP500, DAX40, NIKKEI225, Cocoa, Cotton, Silver, BRENT CRUDE (9)
Mallow's Cp	0.168	USDGBP, USDAUD, SP500, DAX40, NIKKEI225, Cocoa, Cotton, Silver, BRENT CRUDE (9)
BIC	0.662	All (20) besides: USDEUR, Corn, Wheat, Lumber, NatGas

Figure 15 - Results CSR.

Regarding the model re-estimation using 9 variables, the best subsets per measurement slightly differ, meaning (i) based on the Adjusted R-squared the strongest subset is excluding

"Natural gas" and therefore considering 8 variables, (ii) Mallow's Cp is lowest leaving out "Cotton" and "Natural gas" and (iii) the BIC is positive for the first time not excluding any variables. Besides a tiny improvement in the Adjusted R-Squared, Mallow's Cp and BIC are worse than in the re-estimated model (figure X in appendix). In other words, simply reducing the number of variables has a visible impact on the results per measurement. Overall, the goodness of fit of the re-estimated model is still given and concrete coefficients are not obtained from this variable-selection model (figure 25 in appendix).

5. Out-of-sample forecasting

After having analyzed the data based on three different methods – OLS, PCA and CSR, in the in-sample period, this section is dedicated to the actual forecasting. To capture long-term trends and the dynamic nature of the cryptocurrency data using rolling windows is considered a sound approach to assess the model performance. Hereby, the previous 1-year period will be equal to the training data and the following six months will be deemed the forecast data. Hence, seven rolling windows will be determined and subsequently, other than in the in-sample period, the "Out-of-sample R-squared" will be measured comparing the naïve model with a model estimated by the CSR approach. Also, for each timeframe the significant variables will be filtered and observed throughout the approximately 5-year period.

As depicted in figure 16, the CSR model has performed a superior performance twice – between 02.06.2020 and 02.06.2020 compared to the naïve model. A positive OOS R-squared suggests that the CSR model's predictions have captured more of the variation in the actual values compared to the benchmark model.

Rolling window	Learning period	Forecast period	Significant variables based on adjusted R-squared	OOS R-squared
1	01.12.2017 – 01.12.2018	02.12.2018 – 02.06.2019	USDGBP, USDCNY, SP500, NIKKEI225, FTSE100, Corn, Wheat, Lumber, Silver, Gold	-0.099
2	01.06.2018 – 01.06.2019	02.06.2019 - 02.12.2019	USDJPY, USDCNY, SP500, DAX40, NIKKEI225, FTSE100, Wheat, Soybean, Coffee, Cotton, Silver, BrentCrude, NatGas	-0.077
3	01.12.2018 – 01.12.2019	02.12.2019 - 02.06.2020	USDJPY, USDCNY, DAX40, FTSE100, COLCAP25, Corn, Cocoa, Cotton, Gold, CrudeOil, BrentCrude	-7.452
4	01.06.2019 – 01.06.2020	02.06.2020 - 02.12.2020	USDEUR, USDJPY, USDAUD, USDCNY, SP500, DAX40, NIKKEI225, RoughRice, Soybean, Cocoa, Copper, Gold	0.007
5	01.12.2019 – 01.12.2020	02.12.2020 - 02.06.2021	USDEUR, USDJPY, USDAUD, USDCNY, SP500, DAX40, NIKKEI225, Wheat, Cocoa, Rubber, Gold, BrentCrude	0.091
6	01.06.2020 – 01.06.2021	02.06.2021 - 02.12.2021	USDAUD, USDCNY, SP500, DAX40, NIKKEI225, Corn, Wheat, Soybean, Cocoa, Coffee, Cotton, BrentCrude, Natgas	-0.113
7	01.12.2020 – 01.12.2021	02.12.2021 - 02.06.2022	USDEUR, USDAUD, USDCNY, NIKKEI225, FTSE100, Corn, Soybean, Cocoa, Cotton, Silver, Gold, CrudeOil	-0.012

Figure 16 - Out-of-sample results rolling windows.

In all the 7 rolling windows the FX-rate USDCNY has been included followed by the NIKKEI225 which has been selected 6 times. This indicates that these two variables are deemed the most influential based on the OOS adjusted R-squared. The 5th rolling window in the forecasting period 02.12.2020 to 02.06.2021 depicts the highest OOS R-squared with around 9% of the captured variance.

6. Limitations

After having performed regression analysis targeting different angels of the data, the sixth agenda point discusses how certain constraints undermine the results and learnings of this thesis regarding investment strategies. The limitations are three-fold covering the data characteristics in general terms and more specifically the portfolio as well as the predictors.

In the subsequent agenda point (i.e., 7) a reflection of the results will be debated separately including thoughts referring to the applied methodologies as well as the pre-discussed literature.

The data period was heavily influenced by a, partially simultaneously occur series of catastrophic and unusual events such as the COVID-19 pandemic beginning in early 2020, the outbreak of the war in Ukraine in 2022, the major trade conflict between the US and China in 2018-2019, climate emergencies as well as the troop withdrawal in Afghanistan to name a few. These occurrences made forecasting even more challenging and underlined the non-stationary, dynamic, and volatile behavior of cryptocurrency data. The formative collapse of FTX, a Bahamas-based cryptocurrency exchange, has been left out on purpose to obtain a certain degree of normality in the data.

Moreover, the thesis faced constraints in respect to the availability, reliability, and quality of the data. The relatively nascent cryptocurrency market where new cryptocurrency market was emerging almost daily in the past provides limited historical data and currencies that are popular today were not existing in 2017 which marks the beginning of the explored period. Further in terms of reliability, it is important to state that the cryptocurrency market is and has been susceptible to market manipulation through social media (e.g., Elon Musk Tweets) and the fact that a substantial share of coins is hold by few major entities.

Mentioning the predictor variables, it is reasonable to assert that, although a total of 25 dependent variables from different asset classes have been picked, the list could have been extended by e.g., African, Indian, or Chinese stock indices or other FX trading pairs with the US-Dollar. The collection process was primarily focused on the “Western World”. Live cattle were excluded as the only type of commodity because of the relatively low liquidity and the existence of certain moral and ethical issues associated with this investment class.

It is worth noting that the portfolio and its construction also faced certain limitations and inaccuracies, although minor. While the cryptocurrency exchanges are becoming more diversified and adding additional currencies on a weekly basis, the numbers reveal that the most popular and traded ones are Bitcoin and Ethereum. The constructed portfolio reflects this and 90% of the weight has been held by these two. In the beginning of the analyzed period the two currencies even held a higher percentage above 95%. Consequently, many of the other 8 currencies do not really influence the movement of the portfolio and a single instead of a portfolio approach could also have been chosen with results not heavily deviating from the current ones. Moreover, even though minor, the “Chainlink” as well as “Cardano” project depict a high degree of interconnectedness and collaboration with the “Ethereum” blockchain underlining the natural dependence of the projects in terms of technological advancements and financial returns.

7. Conclusions

The journey into the cryptocurrency forecasting space has been an inspiring endeavor, exploring the impact of various factors on cryptocurrency returns. The New York Times article from January 2022 sparked crucial debates on the existence of a cryptocurrency bubble and the market's inherent volatility. Considering unprecedented growth and Bitcoin's record-breaking surge, it became important to unravel the drivers influencing cryptocurrency predictability and prices.

The out-of-sample results revealed the FX-rate USD/CNY and the NIKKEI225 as consistently influential predictors among the 24 assets, based on the OOS adjusted R-squared values. Despite relatively low adjusted R-squared values overall, the inclusion of these variables played a significant role in explaining the variability in cryptocurrency returns.

Additionally, the PCA results indicated that PC1 held importance in predicting cryptocurrency returns, particularly influenced by the FX rates "USD/AUD" and "USD/GBP" with positive loadings. Furthermore, PC2 highlighted significant negative loadings from stock indices such as DAX40 and FTSE100, as well as positive loadings from commodities like "Rubber." These insights allowed us to gain a deeper understanding of the underlying patterns within the data.

The main findings of this work overlap with the following topics and its respective researchers:

i. *Long-Range Trends:* The exploration of long-range trends of cryptocurrency returns in this study aligns with the Hurst exponent used by Alexiadou et al. (2023). Both studies find evidence suggesting that cryptocurrency returns follow a random walk, making accurate forecasting challenging.

ii. *Uncertainty in cryptocurrency returns:* This investigation into the influence of different types of uncertainty on cryptocurrency returns echoes the work of Nguyen Quang et al. (2020) which explores how World Uncertainty and Global Economic Policy Uncertainty affect cryptocurrency portfolios. The findings presented in this study support the idea that increased global economic policy uncertainty negatively impacts cryptocurrency returns.

iii. *Cryptocurrencies vs. Gold:* Just like Klose's research (2022), this research contributes to the characterization of cryptocurrencies as speculative assets rather than stores of value.

Further research could focus on including more detailed types of predictors, such as the integration of market sentiment indicators. These are derived from social media, news

articles, Google search trends, as well as sentiment indices and could offer a comprehensive view of market participants' emotions and perceptions, thereby illuminating potential drivers of cryptocurrency price movements. Novel insights could be gained therefrom fostering a greater understanding of the multifaceted dynamics that underlie the cryptocurrency market.

References:

1. Alatorre, D., Gershenson, C., & Mateos, J. L. (2023, March 16). Stocks and cryptocurrencies: Antifragile or robust? A novel antifragility measure of the stock and cryptocurrency markets. *PLOS ONE*, 18(3), e0280487. https://doi.org/10.1371/journal.pone.0280487

2. Alexander, C., & Dakos, M. (2019). A Critical Investigation of Cryptocurrency Data and Analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3382828

3. Alexander, C., Heck, D., & Kaeck, A. (2021). The Role of Binance in Bitcoin Volatility Transmission. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3877949

4. Alexiadou, M., Sofianos, E., Gogas, P., & Papadimitriou, T. (2023, February 27). Cryptocurrencies and Long-Range Trends. *International Journal of Financial Studies*, 11(1), 40. https://doi.org/10.3390/ijfs11010040

5. Briola, A., & Aste, T. (2022, October 28). Dependency Structures in Cryptocurrency Market from High to Low Frequency. *Entropy*, 24(11), 1548. https://doi.org/10.3390/e24111548

6. Czeczeli, V., & Vilonya, M. (2022). Exchange Rate Developments of Cryptocurrencies Based on Event Study Analysis. *Pénzügyi Szemle = Public Finance Quarterly*, 67(2), 231–247. https://doi.org/10.35551/pfq_2022_2_5

7. Elliott, G., Gargano, A., & Timmermann, A. (2013, December). *Complete subset regressions. Journal of Econometrics*, 177(2), 357–373. https://doi.org/10.1016/j.jeconom.2013.04.017

8. Elliott, G., Gargano, A., & Timmermann, A. (2015, May). Complete subset

regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control*, 54, 86–110. https://doi.org/10.1016/j.jedc.2015.03.004

9. Higgins, S. (2017, December 29). From $900 to $20,000: The Historic Price of Bitcoin in 2017. https://www.coindesk.com/markets/2017/12/29/from-900-to-20000-bitcoins-historic-2017-price-run-revisited/

10. Greene, W. (2007, August 7). Econometric Analysis. 193. https://doi.org/10.1604/9780135132456

11. Hjort, N. L., & Claeskens, G. (2003, December). Frequentist Model Average Estimators. *Journal of the American Statistical Association*, 98(464), 879–899. https://doi.org/10.1198/016214503000000828

12. Holmes, A., Illowsky, B., & Dean, S. (2017, November 30). Introductory Business Statistics. 52-58.

13. International Monetary Fund. (September 2022). 'DeFi' and 'TradFi' Must Work Together. https://www.imf.org/en/Publications/Fandd/Issues/2022/09/22/can-blockchain-revolutionize-international-trade

14. Jolliffe, I. T. (2002, October 1). Principal Component Analysis. Springer. 1-74. https://doi.org/10.1007/b8486510.1007/978-0-387-22440-410.1007/b98835

15. Jolliffe, I. T., & Cadima, J. (2016, April 13). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

16. Kassambara, A. (2017, August 22). Practical Guide to Principal Component Methods in R. 17-50.

17. Klose, J. (2022, October 5). Comparing cryptocurrencies and gold - a system-GARCH-approach. *Eurasian Economic Review*, 12(4), 653–679. https://doi.org/10.1007/s40822-022-00218-4

18. Lee, R. (2023, July 26). The Top 8 Most Tradable Currencies. Investopedia. Retrieved August 31, 2023, from https://www.investopedia.com/trading/most-tradable-currencies/

19. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012, April 9). Introduction to Linear Regression Analysis (Vol. 821). 748-749.

20. Murray, K., Rossi, A., Carraro, D., & Visentin, A. (2023, January 29). On Forecasting Cryptocurrency Prices: A Comparison of Machine Learning, Deep Learning, and Ensembles. *Forecasting*, 5(1), 196–209. https://doi.org/10.3390/forecast5010010

21. Naeem, M. A., Mbarki, I., Suleman, M. T., Vo, X. V., & Shahzad, S. J. H. (2020, December 16). Does Twitter Happiness Sentiment predict cryptocurrency? *International Review of Finance*, 21(4), 1529–1538. https://doi.org/10.1111/irfi.12339

22. Nguyen Quang, B., Le, T., & Nguyen Phuc, C. (2020, October 9). Influences of uncertainty on the returns and liquidity of cryptocurrencies: Evidence from a portfolio approach. *International Journal of Finance & Economics*, 27(2), 2497–2513. https://doi.org/10.1002/ijfe.2283

23. Nickel, R. (2023, March 1). Insight: Soils of war: The toxic legacy for Ukraine's breadbasket. *Reuters*. https://www.reuters.com/world/europe/soils-war-toxic-legacy-ukraines-breadbasket-2023-03-01/

24. Obeng, C., & Attor, C. (2022, December 22). INTERCONNECTION AMONG CRYPTOCURRENCIES: USING VECTOR ERROR CORRECTION MODEL. *International Journal of Entrepreneurial Knowledge*, 10(2), 24–41.

https://doi.org/10.37335/ijek.v10i2.157

25. Panagiotidis, T., Stengos, T., & Vravosinos, O. (2018, December). On the determinants of bitcoin returns: A LASSO approach. *Finance Research Letters*, 27, 235–240. https://doi.org/10.1016/j.frl.2018.03.016

26. Sung, S. H., Kim, J. M., Park, B. K., & Kim, S. (2022, September 1). A Study on Cryptocurrency Log-Return Price Prediction Using Multivariate Time-Series Model. *Axioms*, 11(9), 448. https://doi.org/10.3390/axioms11090448

27. The New York Times. (2022, January 27). It's Hard to Tell When the Crypto Bubble Will Burst, or If There Is One. https://www.nytimes.com/2022/01/27/business/crypto-price-bubble.html

28. Yavuz, M. S., Bozkurt, G., & Boğa, S. (2022, December 19). Investigating the Market Linkages between Cryptocurrencies and Conventional Assets. *EMAJ: Emerging Markets Journal*, 12(2), 36–45. https://doi.org/10.5195/emaj.2022.266

29. Zargar, F. K., Dilip, K. (2022). COVID-19 and Cryptocurrency Market: Impact on Return, Volatility and Liquidity | *The Journal of Prediction Markets*; 16(2):19-38, 2022. | ProQuest Central. https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/pt/covidwho-2110692

Appendix:

a) Additional figures

Assets	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Skewness	Kurtosis
Portfolio	-0.3701	-0.0170	0.0012	0.0013	0.0208	0.1809	-0.4295	8.7346
USDEUR	-0.0156	-0.0013	0.0000	0.0001	0.0016	0.0267	0.3973	6.9593
USDGBP	-0.0293	-0.0017	0.0000	0.0001	0.0019	0.0435	0.4965	12.4431
USDJPY	-0.0215	-0.0011	0.0000	0.0002	0.0016	0.0266	0.1875	8.8701
USDAUD	-0.1287	-0.0022	0.0000	0.0001	0.0020	0.1695	3.9799	217.4761
USDCNY	-0.0181	-0.0007	0.0000	0.0001	0.0006	0.0160	0.1226	10.3322
SP500	-0.1198	-0.0016	0.0000	0.0003	0.0035	0.0938	-0.6607	22.5551
DAX40	-0.1224	-0.0023	0.0000	0.0001	0.0035	0.1098	-0.4512	21.6713
NIKKEI225	-0.5133	-0.0050	0.0000	0.0002	0.0067	0.1626	-5.3666	118.1377
FTSE100	-0.1087	-0.0019	0.0000	0.0000	0.0030	0.0905	-1.0478	23.9836
COLCAP25	-0.1503	-0.0021	0.0000	0.0000	0.0028	0.1328	-0.7207	49.7371
Corn	-0.1739	-0.0037	0.0000	0.0005	0.0052	0.0641	-1.3590	22.6360
Wheat	-0.1068	-0.0068	0.0000	0.0006	0.0054	0.2178	1.1639	18.4062
RoughRice	-0.2590	-0.0028	0.0000	0.0003	0.0038	0.1030	-3.5862	80.7386
SoyBean	-0.1050	-0.0032	0.0000	0.0003	0.0040	0.0664	-0.7958	13.7209
Cocoa	-0.0852	-0.0061	0.0000	0.0002	0.0066	0.1218	0.1528	7.2692
Coffee	-0.0863	-0.0073	0.0000	0.0003	0.0059	0.1003	0.5218	6.5268
Cotton	-0.2388	-0.0039	0.0000	0.0001	0.0045	0.0701	-2.1093	37.7442
Lumber	-0.3348	-0.0086	0.0000	0.0005	0.0115	0.3070	-0.5325	25.1221
Rubber	-0.1220	-0.0028	0.0000	0.0000	0.0029	0.0910	-0.9230	17.9738
Silver	-0.1165	-0.0042	0.0000	0.0002	0.0049	0.0930	-0.4032	12.9202
Copper	-0.0669	-0.0041	0.0000	0.0001	0.0042	0.0518	-0.2147	6.2667
Gold	-0.0499	-0.0017	0.0000	0.0002	0.0028	0.0595	-0.1691	11.7592
CrudeOil	-3.0597	-0.0037	0.0000	-0.0017	0.0082	0.2467	-30.0305	1066.9754
BrentCrude	-0.4339	-0.0034	0.0000	0.0009	0.0082	0.6683	4.2353	144.6106
NatGas	-0.2595	-0.0087	0.0000	0.0009	0.0105	0.4648	1.4039	27.9913

Figure 17 - Descriptive statistics of predictors and portfolio.

	BTC	ETH	LTC	IOTA	DOGE	CHAIN	XLM	XMR	XRP	ADA
BTC	1.000	0.779	0.740	0.625	0.242	0.540	0.518	0.714	0.502	0.494
ETH	0.779	1.000	0.811	0.668	0.252	0.621	0.554	0.709	0.596	0.559
LTC	0.740	0.811	1.000	0.635	0.257	0.539	0.524	0.684	0.577	0.508
IOTA	0.625	0.668	0.635	1.000	0.219	0.526	0.527	0.626	0.526	0.467
DOGE	0.242	0.252	0.257	0.219	1.000	0.240	0.165	0.249	0.200	0.169
CHAIN	0.540	0.621	0.539	0.526	0.240	1.000	0.445	0.515	0.457	0.425
XLM	0.518	0.554	0.524	0.527	0.165	0.445	1.000	0.513	0.585	0.497
XMR	0.714	0.709	0.684	0.626	0.249	0.515	0.513	1.000	0.509	0.491
XRP	0.502	0.596	0.577	0.526	0.200	0.457	0.585	0.509	1.000	0.485
ADA	0.494	0.559	0.508	0.467	0.169	0.425	0.497	0.491	0.485	1.000

Figure 18 - Detailed correlation crypto.

Range	Occurrences
(-0.3 to -0.2)	2
(-0.2 to -0.1)	16
(-0.1 to 0)	152
(0 to 0.1)	296
(0.1 to 0.2)	122
(0.2 to 0.3)	28
(0.3 to 0.4)	10
(0.4 to 0.5)	10
(0.5 to 0.6)	8
(0.6 to 0.7)	2
(0.7 to 0.8)	2
(0.8 to 0.9)	2

Figure 19 - Correlation results predictors.

	Portfolio	USDEUR	USDGBP	USDJPY	USDAUD	USDCNY
Portfolio	1.000	-0.103	-0.156	0.0018	-0.152	-0.049
USDEUR	-0.103	1.000	0.634	0.3874	0.418	0.265
USDGBP	-0.156	0.634	1.000	0.2899	0.438	0.242
USDJPY	0.002	0.387	0.290	1.000	0.167	0.125
USDAUD	-0.152	0.418	0.438	0.1666	1.000	0.223
USDCNY	-0.049	0.265	0.242	0.125	0.223	1.000
SP500	-0.029	-0.040	-0.004	-0.0158	-0.020	0.060
DAX40	0.075	0.007	-0.022	0.1504	-0.079	0.051
NIKKEI225	0.220	-0.080	-0.151	0.235	-0.210	-0.131
FTSE100	0.045	0.024	0.033	0.1493	-0.068	0.064
COLCAP25	0.023	-0.045	-0.097	-0.013	-0.081	-0.002
Corn	0.015	-0.003	-0.007	0.021	-0.042	-0.010
Wheat	0.025	0.063	-0.003	0.0328	-0.098	0.005
RoughRice	-0.012	0.023	-0.007	0.0316	0.015	0.021
SoyBean	0.036	-0.004	-0.017	-0.0055	-0.050	0.025
Cocoa	0.066	-0.040	-0.077	-0.0634	-0.044	0.001
Coffee	0.033	0.002	0.015	0.02	-0.025	0.013
Cotton	-0.001	-0.002	-0.039	0.0361	-0.016	-0.008
Lumber	0.016	-0.041	-0.006	0.0426	0.001	0.028
Rubber	0.054	-0.101	-0.112	0.0116	-0.127	-0.046
Silver	0.060	-0.010	-0.044	-0.0075	-0.050	0.005
Copper	0.062	-0.079	-0.052	-0.0285	-0.097	0.020
Gold	0.029	-0.052	-0.047	-0.0535	-0.019	0.013
CrudeOil	0.013	0.005	-0.019	0.0089	-0.031	0.003
BrentCrude	0.064	0.035	-0.009	0.033	-0.021	0.012
NatGas	0.023	-0.032	-0.028	-0.004	-0.011	-0.021

Figure 20 - Detailed correlation predictors part 1.

SP500	DAX40	NIKKEI 225	FTSE100	COLCAP 25	Corn
-0.029	0.075	0.220	0.045	0.023	0.0145
-0.040	0.0072	-0.080	0.024	-0.045	-0.0034
-0.004	-0.0223	-0.151	0.033	-0.097	-0.0068
-0.016	0.1504	0.235	0.149	-0.013	0.021
-0.020	-0.0786	-0.210	-0.068	-0.081	-0.0421
0.060	0.0514	-0.131	0.064	-0.002	-0.01
1.000	0.5949	-0.030	0.553	0.464	0.0502
0.595	1.000	0.128	0.838	0.411	0.0243
-0.030	0.1281	1.000	0.124	0.057	0.0422
0.553	0.8379	0.124	1.000	0.449	0.0726
0.464	0.4108	0.057	0.449	1.000	0.1236
0.050	0.0243	0.042	0.073	0.124	1.000
0.008	-0.0304	0.066	-0.001	0.091	0.5067
0.012	0.0349	-0.010	0.044	0.048	0.1134
0.080	0.111	0.019	0.123	0.130	0.5674
0.123	0.1425	0.020	0.092	0.102	-0.0136
0.076	0.0689	0.060	0.076	0.060	0.164
0.191	0.1962	0.053	0.202	0.157	0.1971
0.183	0.1552	0.011	0.151	0.118	0.0489
0.057	0.0937	0.036	0.115	0.078	0.0347
0.174	0.1645	0.079	0.162	0.221	0.1414
0.278	0.293	0.077	0.320	0.206	0.1498
0.076	0.034	0.043	0.041	0.108	0.1077
0.101	0.0491	0.009	0.058	0.120	0.0735
0.037	0.0606	0.013	0.077	0.136	0.1241
0.091	0.0971	0.043	0.087	0.036	0.0517

Figure 21 - Detailed correlation predictors part 2.

Wheat	Rough Rice	Soybean	Cocoa	Coffee	Cotton
0.025	-0.012	0.036	0.0661	0.033	-0.0007
0.063	0.023	-0.004	-0.0397	0.002	-0.0016
-0.003	-0.007	-0.017	-0.0767	0.015	-0.0386
0.033	0.032	-0.006	-0.0634	0.020	0.0361
-0.098	0.015	-0.050	-0.0442	-0.025	-0.0161
0.005	0.021	0.025	0.0013	0.013	-0.0076
0.008	0.012	0.080	0.1227	0.076	0.1908
-0.030	0.035	0.111	0.1425	0.069	0.1962
0.066	-0.010	0.019	0.0201	0.060	0.0533
-0.001	0.044	0.123	0.0917	0.076	0.2023
0.091	0.048	0.130	0.1023	0.060	0.1567
0.507	0.113	0.567	-0.0136	0.164	0.197
1.000	0.128	0.345	0.0444	0.116	0.1986
0.128	1.000	0.100	0.0242	0.056	0.0865
0.345	0.100	1.000	0.0015	0.183	0.1571
0.044	0.024	0.002	1.000	0.120	0.0643
0.116	0.056	0.183	0.1201	1.000	0.1534
0.199	0.087	0.157	0.0643	0.153	1
0.051	-0.020	0.064	0.0712	0.041	0.0589
0.025	0.003	0.056	0.0959	0.121	0.1393
0.123	-0.025	0.169	0.1065	0.134	0.1354
0.137	0.052	0.251	0.1006	0.195	0.2694
0.111	-0.015	0.144	0.0778	0.115	0.0773
0.021	-0.015	0.071	0.0388	0.061	-0.0047
0.101	0.040	0.129	0.1226	0.078	0.0958
0.031	0.005	0.041	-0.0146	0.057	0.0704

Figure 22 - Detailed correlation predictors part 3.

Lumber	Rubber	Silver	Copper	Gold	Crude Oil	Brent Crude	NatGas
0.016	0.0541	0.060	0.062	0.029	0.013	0.064	0.023
-0.041	-0.1009	-0.010	-0.079	-0.052	0.005	0.035	-0.032
-0.006	-0.1117	-0.044	-0.052	-0.047	-0.019	-0.009	-0.028
0.043	0.0116	-0.008	-0.029	-0.054	0.009	0.033	-0.004
0.001	-0.1273	-0.050	-0.097	-0.019	-0.031	-0.021	-0.011
0.028	-0.0464	0.005	0.020	0.013	0.003	0.012	-0.021
0.183	0.0573	0.174	0.278	0.076	0.101	0.037	0.091
0.155	0.0937	0.165	0.293	0.034	0.049	0.061	0.097
0.011	0.0361	0.079	0.077	0.043	0.009	0.013	0.043
0.151	0.1152	0.162	0.320	0.041	0.058	0.077	0.087
0.118	0.0777	0.221	0.206	0.108	0.120	0.136	0.036
0.049	0.0347	0.141	0.150	0.108	0.074	0.124	0.052
0.051	0.0245	0.123	0.137	0.111	0.021	0.101	0.031
-0.020	0.0026	-0.025	0.052	-0.015	-0.015	0.040	0.005
0.064	0.0561	0.169	0.251	0.144	0.071	0.129	0.041
0.071	0.0959	0.107	0.101	0.078	0.039	0.123	-0.015
0.041	0.1208	0.134	0.195	0.115	0.061	0.078	0.057
0.059	0.1393	0.135	0.269	0.077	-0.005	0.096	0.070
1.000	0.0183	0.069	0.108	0.044	0.041	0.075	-0.012
0.018	1.000	0.074	0.164	0.019	0.015	0.045	-0.023
0.069	0.0739	1.000	0.328	0.784	0.020	0.078	0.041
0.108	0.1639	0.328	1.000	0.227	0.089	0.179	0.085
0.044	0.019	0.784	0.227	1.000	-0.011	0.043	0.017
0.041	0.015	0.020	0.089	-0.011	1.000	0.392	-0.048
0.075	0.0446	0.078	0.179	0.043	0.392	1.000	0.040
-0.012	-0.0234	0.041	0.085	0.017	-0.048	0.040	1.000

Figure 23 - Detailed correlation predictors part 4.

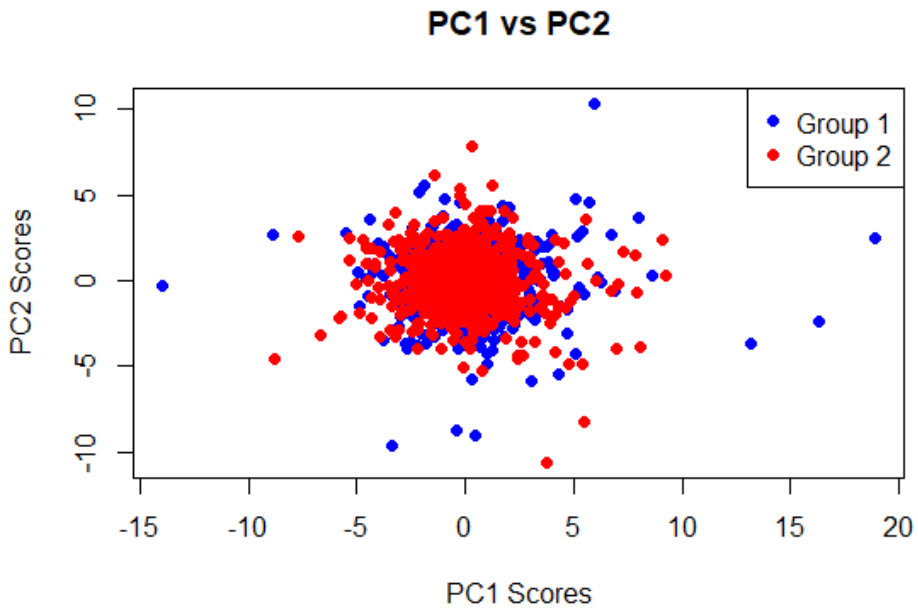

Figure 24 - PC1 vs. PC2 based on scores.

Indicator	Value	Previous value	Variables included	Not included anymore
Adjusted R-squared	0.074	0.076	USDGBP, USDAUD, SP500, DAX40, NIKKEI225, Cocoa, Cotton, Silver (8)	NatGas
Mallow's Cp	7.334	0.168	USDGBP, USDAUD, SP500, DAX40, NIKKEI225, Cocoa, Silver (7)	Cotton, NatGas
BIC	-0.711	0.662	USDGBP, USDAUD, SP500, DAX40, NIKKEI225, Cocoa, Cotton, Silver, Brent Crude (9)	/

Figure 25 – Re-estimated CSR.

b) R-codes

```
# Data & descriptive statistics

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/")

getwd()

OLSdata = read.table("Datafull1.txt", header = T)

OLSdata = as.data.frame(OLSdata)

install.packages("moments")

install.packages("tseries")

install.packages("timeSeries")

install.packages("PerformanceAnalytics")

library(timeSeries)

library(moments)

library(tseries)

library(PerformanceAnalytics)



# descriptive statistics of asset returns

summary(OLSdata)

OLSdata_summary <- sapply(OLSdata[-1], function(x) round(summary(x), 4))

OLSdata_summary

rownames(OLSdata_summary) <- c("Asset", "Min.", "1st Qu.", "Median", "Mean", "3rd
Qu.", "Max.")
```

```
Summary_descriptive_transposed <- t(OLSdata_summary)

Summary_descriptive_transposed


#Calculate Skewness and kurtosis

skew <- apply(OLSdata[-1], 2, skewness)

kurt <- apply(OLSdata[-1], 2, kurtosis)

skew

kurt


Summary_descriptive_transposed <- cbind(Summary_descriptive_transposed, Skewness =
skew, Kurtosis = kurt)

Summary_descriptive_transposed


#Extract data to excel

write.table(Summary_descriptive_transposed, file = "descriptivestatistics.csv", sep = ",",
dec = ".", row.names = TRUE, quote = FALSE)


#correlations among all variables (simple)

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/")

getwd()
```

```
OLSdata = read.table("Datafull1.txt", header = T)

OLSdata = as.timeSeries(OLSdata)



correlation <- cor(OLSdata)

correlation <- round(correlation, 4)

correlation



write.csv(correlation, file = "correlation.csv")

min(correlation)

max(correlation[correlation < 1])



#clustering correlations (holistic view)

#range

lower_bound <- -0.3

upper_bound <- 0.9

step <- 0.1

ranges <- c(seq(lower_bound, -step, by = step), 0, seq(step, upper_bound, by = step))



#cluster into groups

cor_clusters <- cut(correlation, breaks = ranges, include.lowest = TRUE)
```

```
#count of number for each group

cor_counts <- table(cor_clusters)

print(cor_counts)


#COrrelation with portfolio max and min

cor_column1 <- correlation[, 1:1]

cor_column1

cor_column1[1] <- NA


# Find the lowest correlation (excluding self-correlation)

min_corr_p <- min(cor_column1, na.rm = TRUE)

min_corr_p

max_corr_p <- max(cor_column1, na.rm = TRUE)

max_corr_p


#Clustering correlations with portfolio

lower_bound_p <- -0.2

upper_bound_p <- 0.24

step_p <- 0.04
```

```
ranges_p <- c(seq(lower_bound, -step, by = step), 0, seq(step, upper_bound, by = step))

cor_clusters_p <- cut(cor_column1, breaks = ranges_p, include.lowest = TRUE)

cor_counts_p <- table(cor_clusters_p)

print(cor_counts_p)



#OLS regression

install.packages("lmtest")

install.packages("ggplot2")

library(lmtest)

library(ggplot2)

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/")

getwd()

OLSdata = read.table("Datafull1.txt", header = T)

OLSdata = as.data.frame(OLSdata[-1])

OLSdata



OLSreg <- lm(Portfolio ~ ., data = OLSdata)

summary(OLSreg)



#Overview coefficients
```

```
install.packages("broom")

library(broom)

OLSregresults <- tidy(OLSreg)

print(OLSregresults, n = Inf)

write.csv(tidy(OLSreg), "OLS_results.csv", row.names = FALSE)

glance(OLSreg)



#Graphing the resuls of the OLS unlagged

# Scatter plot of actual vs. predicted values

ggplot(OLSdata, aes(x = Portfolio, y = OLSreg$fitted.values)) +

  geom_point(alpha = 0.4, color = "blue") +

  geom_abline(intercept = 0, slope = 1, linetype = "dashed") +

  labs(title = "Actual vs. Predicted Returns", x = "Actual Returns", y = "Predicted Returns")



# Density plot of residuals

ggplot(data.frame(resid = OLSreg$residuals), aes(x = resid)) +

  geom_density() +

  labs(x = "Residuals", y = "Density") +

  ggtitle("Density Plot of Residuals") +

  stat_function(fun = dnorm, args = list(mean = mean(OLSreg$residuals), sd =
```

```
sd(OLSreg$residuals)), color = "red", size = 1)


#lagging the data

#explag 1

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/lags")

getwd()

Explag1 = read.table("Explag1.txt", header = T)

Explag1 = as.timeSeries(Explag1)

Explag1OLS <- lm(Portfolio ~ ., data = Explag1)

summary(Explag1OLS)

glance(Explag1OLS)


#explag 2

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/lags")

getwd()

Explag2 = read.table("Explag2.txt", header = T)

Explag2 = as.timeSeries(Explag2)

Explag2OLS <- lm(Portfolio ~ ., data = Explag2)

summary(Explag2OLS)

glance(Explag2OLS)
```

```
#explag 3

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/lags")

getwd()

Explag3 = read.table("Explag3.txt", header = T)

Explag3 = as.timeSeries(Explag3)

Explag3OLS <- lm(Portfolio ~ ., data = Explag3)

summary(Explag3OLS)

glance(Explag3OLS)


#Resplag 1

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/lags")

getwd()

Resplag1 = read.table("Resplag1.txt", header = T)

Resplag1 = as.timeSeries(Resplag1)

Resplag1OLS <- lm(Portfolio ~ ., data = Resplag1)

summary(Resplag1OLS)

glance(Resplag1OLS)


#Resplag 2
```

```
setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/lags")

getwd()

Resplag2 = read.table("Resplag2.txt", header = T)

Resplag2 = as.timeSeries(Resplag2)

Resplag2OLS <- lm(Portfolio ~ ., data = Resplag2)

summary(Resplag2OLS)



glance(Resplag2OLS)

#PCA

install.packages("stats")

library(stats)

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/")

getwd()

OLSdata = read.table("Datafull1.txt", header = T)

OLSdata



OLSdata_df <- as.data.frame(OLSdata)

OLSdata_df <- OLSdata_df[ -c( 1, 2)]

OLSdata_df
```

```
# Standardizing the data with 0 mean and and std of 1 for better comparability

OLSdata_std <- scale(OLSdata_df)


#performing PCA

pca_results <- prcomp(OLSdata_std)

pca_results


#Selection of desired number of components -> capturing 80% of the total variance

cum_var <- cumsum(pca_results$sdev^2) / sum(pca_results$sdev^2)

cum_var


num_components <- which(cum_var >= 0.6)[1] #60 % threshold

num_components


cum_var_rel <- cum_var [1:num_components] #from 1 to the num_component the cum
variance

cum_var_rel


# Examination of loading of the variables for each retained component

loadings <- pca_results$rotation[, 1:9]
```

```
loadings <- round(loadings, 4)

loadings


write.table(loadings, file = "pca_results.csv", sep = ",", dec = ".", row.names = TRUE,

quote = FALSE)


# significance of PC as a data frame for multiplication later

cum_var_rel <- as.data.frame(t(cum_var_rel))


colnames(cum_var_rel) <- paste0("PC", 1:9) #replacing v1 with PC1

cum_var_rel <- t(cum_var_rel)

cum_var_rel


# Multiply loading of each variable times the variation impact

#into matrices

loadings_mx <- as.matrix(loadings)

loadings_mx


cum_var_rel_mx <- as.matrix(cum_var_rel)

cum_var_rel_mx
```

```
write.table(cum_var_rel_mx, file = "PCA_impact.csv", sep = ",", dec = ".", row.names =
TRUE, quote = FALSE)

#rename column
colnames(cum_var_rel_mx) <- "Cum_Var"

#multiplication
loadings_mx * as.vector(cum_var_rel_mx)

result_pca_asset <- rowSums(loadings_mx * as.vector(cum_var_rel_mx))
result_pca_asset <- as.table(result_pca_asset)
result_pca_asset

write.table(result_pca_asset, file = "pca_results_sum.csv", sep = ",", dec = ".", row.names =
TRUE, quote = FALSE)

#extract the scores; rank of the contributions
scores <- pca$x
scores
loadings
```

```
# PCA plots

# Biplots

biplot(pca)



# Scree plot

scree_vals <- pca_results$sdev^2

plot(1:length(scree_vals), scree_vals, type = "b", xlab = "Component Number", ylab =

"Variance", main = "Scree Plot of PCA Results")



# Scores plot

pc1_scores <- pca_results$x[, 1]

pc2_scores <- pca_results$x[, 2]

#Grouping PC1 and PC2

groups <- rep(c(1, 2), each = 50)

# Creating a scatter plot with different colors for each group

plot(pc1_scores, pc2_scores, type = "n", xlab = "PC1 Scores", ylab = "PC2 Scores", main =

"PC1 vs PC2")

# Adding points for each group with different colors

points(pc1_scores[groups == 1], pc2_scores[groups == 1], pch = 16, col = "blue")

points(pc1_scores[groups == 2], pc2_scores[groups == 2], pch = 16, col = "red")
```

```
# Adding a legend

legend("topright", legend = c("Group 1", "Group 2"), col = c("blue", "red"), pch = 16)



#Run regression using the first three principal components

pca_reg_prep <- pca_results$rotation[, 1:3]

pca_reg_prep <- round(pca_reg_prep, 4)

pca_reg_prep



# Extract the scores for the first three components

scores <- as.data.frame(pca_results$x[, 1:3])

scores



#PCA model

OLSdata[,2]

pca_model <- lm(OLSdata[,2] ~ scores$PC1 + scores$PC2 + scores$PC3)

summary(pca_model)



# loadings of statistically relevant principal components

pca_rel <- pca_results$rotation[, 1:2]

pca_rel
```

```
write.table(pca_rel, file = "pca_loadings.csv", sep = ",", dec = ".", row.names = TRUE,
quote = FALSE)


# Complete Subset Regression
install.packages("leaps")
library(leaps)


setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/")
getwd()
CSRdata = read.table("Datafull1.txt", header = T)
CSRdata_df <- as.data.frame(CSRdata)
CSRdata_df


subset_reg <- regsubsets(Portfolio ~ . - Date, data=CSRdata_df, nvmax=24,
method="exhaustive")
subset_reg
subset_summary <- summary(subset_reg)
subset_summary


#adjr2
```

```
best_subset_number <- which.max(summary(subset_reg)$adjr2)

best_subset_number



best_subset_adjr2 <- summary(subset_reg)$adjr2[best_subset_number]

best_subset_adjr2



# all subset adjr2 values

all_adjr2 <- subset_summary$adjr2

all_adjr2

#plot for adjr2 of subsets

plot(1:length(all_adjr2), all_adjr2, type = "l", xlab = "Subsets", ylab = "Adjusted R-
squared",

    main = "Adjusted R-squared of Subsets", xlim = c(1, length(all_adjr2)))



points(best_subset_number, best_subset_adjr2, col = "red", pch = 19)



axis(side = 1, at = best_subset_number, labels = best_subset_number, col.axis = "red",
col.ticks = "red")



# What variables are included in the best subset based on adjr2

best_subset_number <- which.max(summary(subset_reg)$adjr2)
```

```
best_subset_variables <- names(coef(subset_reg, best_subset_number)) [-1]

best_subset_variables



#cp

all_cp <- subset_summary$cp

positive_cp <- all_cp[all_cp > 0]

best_cp_number <- which(all_cp == min(positive_cp))

best_cp_number



best_subset_cp <- summary(subset_reg)$cp[best_cp_number]

best_subset_cp



#plot for cp of subsets

plot(1:length(all_cp), all_cp, type = "l", xlab = "Subsets", ylab = "Mallows' Cp",

    main = "Mallows' Cp of Subsets", xlim = c(1, length(all_adjr2)))



points(best_cp_number, best_subset_cp, col = "red", pch = 19)

axis(side = 1, at = best_cp_number, labels = best_cp_number, col.axis = "red", col.ticks =

"red")
```

```
# What variables are included in the best subset based on Cp

best_subset_variables_cp <- names(coef(subset_reg, best_cp_number)) [-1]

best_subset_variables_cp



#bic

all_bic <- subset_summary$bic

closest0_bic <- abs(all_bic)

closest0_bic

best_bic_number <- which(all_bic == min(closest0_bic))

best_bic_number



best_subset_bic <- summary(subset_reg)$bic[best_bic_number]

best_subset_bic



#plot for bic of subsets

plot(1:length(all_bic), all_bic, type = "l", xlab = "Subsets", ylab = "Bayesian Information

Criterion",

    main = "BIC of Subsets", xlim = c(1, length(all_adjr2)))



points(best_bic_number, best_subset_bic, col = "red", pch = 19)
```

```
axis(side = 1, at = best_bic_number, labels = best_bic_number, col.axis = "red", col.ticks =
"red")


# What variables are included in the best subset based on bic
best_subset_variables_bic <- names(coef(subset_reg, best_bic_number)) [-1]
best_subset_variables_bic


best_subset_variables_bic_not <- setdiff(names(CSRdata_df[-1:-2]),
names(coef(subset_reg, best_bic_number)))
best_subset_variables_bic_not


# Model re-estimation
# Selecting assets of best subset (9)
install.packages("leaps")
library(leaps)


setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/")


getwd()
CSRdata = read.table("Datafull1.txt", header = T)
```

```
CSRdata_df <- as.data.frame(CSRdata)

CSRdata_new <- CSRdata_df[c("Portfolio", "USDGBP", "USDAUD", "SP400",
"DAX40", "NIKKEI224", "Cocoa", "Cotton", "Silver", "NatGas")]
CSRdata_new

#Run CSR
subset_reg_2 <- regsubsets(Portfolio ~ ., data=CSRdata_new, nvmax=9,
method="exhaustive")
subset_reg_2
subset_summary_2 <- summary(subset_reg_2)
subset_summary_2

#adjr2
best_subset_number_2 <- which.max(summary(subset_reg_2)$adjr2)
best_subset_number_2

best_subset_adjr2_2 <- summary(subset_reg_2)$adjr2[best_subset_number_2]
best_subset_adjr2_2

# all subset adjr2 values
```

```
all_adjr2_2 <- subset_summary_2$adjr2

all_adjr2_2


#plot for adjr2 of subsets

plot(1:length(all_adjr2_2), all_adjr2_2, type = "l", xlab = "Subsets", ylab = "Adjusted R-
squared",

    main = "Adjusted R-squared of Subsets", xlim = c(1, 10))


points(best_subset_number_2, best_subset_adjr2_2, col = "red", pch = 19)

axis(side = 1, at = best_subset_number_2, labels = best_subset_number_2, col.axis = "red",
col.ticks = "red")



# What variables are included in the best subset based on adjr2

best_subset_variables_2 <- names(coef(subset_reg_2, best_subset_number_2)) [-1]

best_subset_variables_2


#cp

all_cp_2 <- subset_summary_2$cp

all_cp_2

best_cp_number_2 <- which.min(all_cp_2)
```

```
best_cp_number_2


best_subset_cp_2 <- summary(subset_reg_2)$cp[best_cp_number_2]

best_subset_cp_2


#plot for cp of subsets

plot(1:length(all_cp_2), all_cp_2, type = "l", xlab = "Subsets", ylab = "Mallows' Cp",

    main = "Mallows' Cp of Subsets", xlim = c(1, 10))


points(best_cp_number_2, best_subset_cp_2, col = "red", pch = 19)


axis(side = 1, at = best_cp_number_2, labels = best_cp_number_2, col.axis = "red",

col.ticks = "red")


# What variables are included in the best subset based on Cp

best_subset_variables_cp_2 <- names(coef(subset_reg_2, best_cp_number_2)) [-1]

best_subset_variables_cp_2


#bic

all_bic_2 <- subset_summary_2$bic
```

```
all_bic_2


best_bic_number_2 <- which.max(all_bic_2)

best_bic_number_2


best_subset_bic_2 <- summary(subset_reg_2)$bic[best_bic_number_2]

best_subset_bic_2


#plot for bic of subsets

plot(1:length(all_bic_2), all_bic_2, type = "l", xlab = "Subsets", ylab = "Bayesian

Information Criterion", main = "BIC of Subsets", xlim = c(1, 10))


points(best_bic_number_2, best_subset_bic_2, col = "red", pch = 19)


axis(side = 1, at = best_bic_number_2, labels = best_bic_number_2, col.axis = "red",

col.ticks = "red")

# What variables are included in the best subset based on bic


best_subset_variables_bic <- names(coef(subset_reg, best_bic_number)) [-1]

best_subset_variables_bic
```

```
best_subset_variables_bic_not <- setdiff(names(CSRdata_df[-1:-2]),
names(coef(subset_reg, best_bic_number)))
best_subset_variables_bic_not

#complete subset regression applied on out-of-sample data using all predictors
install.packages("leaps")
library(leaps)

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project")

getwd()
CSRdata_OOS = read.table("Datafull_OOS.txt", header = T)
CSRdata_OOS_df <- as.data.frame(CSRdata_OOS)

subset_reg_OOS <- regsubsets(Portfolio ~ . - Date, data=CSRdata_OOS_df, nvmax=24,
method="exhaustive")
subset_reg_OOS
subset_summary_OOS <- summary(subset_reg_OOS)
subset_summary_OOS

#adjr2
```

```
best_subset_number_OOS <- which.max(summary(subset_reg_OOS)$adjr2)

best_subset_number_OOS


best_subset_adjr2_OOS <- summary(subset_reg_OOS)$adjr2[best_subset_number_OOS]

best_subset_adjr2_OOS


# all subset adjr2 values

all_adjr2_OOS <- subset_summary_OOS$adjr2

all_adjr2_OOS


#plot for adjr2 of subsets

plot(1:length(all_adjr2_OOS), all_adjr2_OOS, type = "l", xlab = "Subsets", ylab =

"Adjusted R-squared",

    main = "Adjusted R-squared of Subsets", xlim = c(1, length(all_adjr2_OOS)))


points(best_subset_number_OOS, best_subset_adjr2_OOS, col = "red", pch = 19)


axis(side = 1, at = best_subset_number_OOS, labels = best_subset_number_OOS , col.axis

= "red", col.ticks = "red")


# What variables are included in the best subset based on adjr2
```

```
best_subset_variables_OOS <- names(coef(subset_reg_OOS, best_subset_number_OOS))
[-1]

best_subset_variables_OOS


#cp

all_cp_OOS <- subset_summary_OOS$cp

positive_cp_OOS <- all_cp_OOS[all_cp_OOS > 0]

best_cp_number_OOS <- which(all_cp_OOS == min(positive_cp_OOS))

best_cp_number_OOS


best_subset_cp_OOS <- summary(subset_reg_OOS)$cp[best_cp_number_OOS]

best_subset_cp_OOS


#plot for cp of subsets

plot(1:length(all_cp_OOS), all_cp_OOS, type = "l", xlab = "Subsets", ylab = "Mallows'
Cp",
    main = "Mallows' Cp of Subsets", xlim = c(1, length(all_cp_OOS)))


points(best_cp_number_OOS, best_subset_cp_OOS, col = "red", pch = 19)


axis(side = 1, at = best_cp_number_OOS, labels = best_cp_number_OOS, col.axis = "red",
```

```
col.ticks = "red")


# What variables are included in the best subset based on Cp

best_subset_variables_cp_OOS <- names(coef(subset_reg_OOS, best_cp_number_OOS))
[-1]

best_subset_variables_cp_OOS


#bic

all_bic_OOS <- subset_summary_OOS$bic

closest0_bic_OOS <- abs(all_bic_OOS)

closest0_bic_OOS


best_bic_number_OOS <- which(all_bic_OOS == min(closest0_bic_OOS))

best_bic_number_OOS


best_subset_bic_OOS <- summary(subset_reg_OOS)$bic[best_bic_number_OOS]

best_subset_bic_OOS


#plot for bic of subsets

plot(1:length(all_bic_OOS), all_bic_OOS, type = "l", xlab = "Subsets", ylab = "Bayesian
Information Criterion",
```

```
    main = "BIC of Subsets", xlim = c(1, length(all_bic_OOS)))

points(best_bic_number_OOS, best_subset_bic_OOS, col = "red", pch = 19)


axis(side = 1, at = best_bic_number_OOS, labels = best_bic_number_OOS, col.axis =
"red", col.ticks = "red")


# What variables are included in the best subset based on bic

best_subset_variables_bic_OOS <- names(coef(subset_reg_OOS, best_bic_number_OOS))
[-1]

best_subset_variables_bic_OOS


# predict portfolio values based on best subset variables

#reference to in-sample dataframe and exclusion of best subset variables

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/")

getwd()

CSRdata = read.table("Datafull1.txt", header = T)

CSRdata_df <- as.data.frame(CSRdata)

CSRdata_df


CSRdata_IS <- CSRdata_df[c("Portfolio", "USDGBP", "USDAUD", "SP400", "DAX40",
"NIKKEI224", "Cocoa", "Cotton", "Silver", "NatGas")]
```

```
#Creation and fitting to linear regression model

model <- lm(Portfolio ~ ., data = CSRdata_IS)

model


#new dataframes exlcuding portfolio

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project")


CSRdata_OOS = read.table("Datafull_OOS.txt", header = T)

CSRdata_OOS

CSRdata_OOS_df <- as.data.frame(CSRdata_OOS)


CSRData_OOS_actualY <- CSRdata_OOS_df[c("Portfolio")]

colnames(CSRData_OOS_actualY)[colnames(CSRData_OOS_actualY) == "Portfolio"] <-

"AC"

CSRData_OOS_actualY


#predict the portfolio

predicted_portfolio <- predict(model, newdata = CSRdata_OOS)

predicted_portfolio
```

```
result_prediction <- data.frame(Portfolio = predicted_portfolio)

colnames(result_prediction)[colnames(result_prediction) == "Portfolio"] <- "PRED"

result_prediction <- round(result_prediction,4)

result_prediction



#comparison prediction and actuals

# Merge the predicted results with the actual results

result_prediction$Observation <- seq_len(nrow(result_prediction))

result_prediction$Observation



comparison <- cbind( CSRData_OOS_actualY, result_prediction)[-3]

comparison



#Analyzing the prediction results

#Mean squared error (MSE)

MSE <- apply(comparison, 1, function(row) mean((row["PRED"] - row["AC"])^2))

MSE <- as.data.frame(MSE)

colnames(MSE)[colnames(MSE) == "comparison$MSE"] <- "MSE"

MSE
```

```
#Root mean squared error (RMSE)

RMSE <- apply(comparison, 1, function(row) sqrt(mean((row["PRED"] - row["AC"])^2)))

RMSE <- as.data.frame(RMSE)

colnames(RMSE)[colnames(RMSE) == "comparison$RMSE"] <- "RMSE"

RMSE



# Calculate Mean Absolute Error (MAE)

MAE <- apply(comparison, 1, function(row) mean(abs(row["PRED"] - row["AC"])))

MAE <- as.data.frame(MAE)

colnames(MAE)[colnames(MAE) == "comparison$MAE"] <- "MAE"

MAE



# Calculate Mean Absolute Percentage Error (MAPE)

MAPE <- apply(comparison, 1, function(row) mean(abs((row["PRED"] - row["AC"]) /
row["AC"])) * 100)

MAPE <- as.data.frame(MAPE)

colnames(MAPE)[colnames(MAPE) == "comparison$MAPE"] <- "MAPE"

MAPE
```

```
#overall results all in one table

overall <- cbind(comparison, MSE, RMSE, MAE, MAPE)

overall



install.packages("writexl")

library(writexl)

write_xlsx(overall, path = "predictions.xlsx")



# Visualization of results

plot(dates[valid_rows], result_prediction$PRED[valid_rows], type = "l", col = "blue",

    xlab = "Timeline", ylab = "Returns", main = "Predicted vs Actual Portfolio Values",

    ylim = c(min(result_prediction$PRED[valid_rows],
CSRData_OOS_actualY$AC[valid_rows]),

            max(result_prediction$PRED[valid_rows],
CSRData_OOS_actualY$AC[valid_rows])))

lines(dates[valid_rows], CSRData_OOS_actualY$AC[valid_rows], type = "l", col = "red")



legend("topright", legend = c("Predicted", "Actual"), col = c("blue", "red"), lty = 1)

# Rolling windows

install.packages("dplyr")

library(dplyr)
```

```
setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project/data/")

getwd()

OLSdata = read.table("Datafull1.txt", header = T)

OLSdata # data with orignal values



#Respective time windows

train_start_1 <- as.Date("2017-12-01")

train_start_2 <- as.Date("2018-06-01")

train_start_3 <- as.Date("2018-12-01")

train_start_4 <- as.Date("2019-06-01")

train_start_4 <- as.Date("2019-12-01")

train_start_6 <- as.Date("2020-06-01")

train_start_7 <- as.Date("2020-12-01")



train_end_1 <- as.Date("2018-12-01")

train_end_2 <- as.Date("2019-06-01")

train_end_3 <- as.Date("2019-12-01")

train_end_4 <- as.Date("2020-06-01")

train_end_4 <- as.Date("2020-12-01")
```

```
train_end_6 <- as.Date("2021-06-01")

train_end_7 <- as.Date("2021-12-01")


forecast_start_1 <- as.Date("2018-12-02")

forecast_start_2 <- as.Date("2019-06-02")

forecast_start_3 <- as.Date("2019-12-02")

forecast_start_4 <- as.Date("2020-06-02")

forecast_start_4 <- as.Date("2020-12-02")

forecast_start_6 <- as.Date("2021-06-02")

forecast_start_7 <- as.Date("2021-12-02")


forecast_end_1 <- as.Date("2019-06-02")

forecast_end_2 <- as.Date("2019-12-02")

forecast_end_3 <- as.Date("2020-06-02")

forecast_end_4 <- as.Date("2020-12-02")

forecast_end_4 <- as.Date("2021-06-02")

forecast_end_6 <- as.Date("2021-12-02")

forecast_end_7 <- as.Date("2022-06-02")


# Subset the data for the training period
```

```
train_data_1 <- subset(OLSdata, Date >= train_start_7 & Date <= train_end_7, select = -
c(Date, 2))

train_data_1


# Fit a regression model using the training data

model_1 <- lm(train_data_1[, 1] ~ ., data = train_data_1)

model_1


# Subset the data for the forecast period

forecast_data_1 <- subset(OLSdata, Date >= forecast_start_7 & Date <= forecast_end_7,
select = -c(Date, 2))

forecast_data_1


#predict

predictions_naive_1 <- predict(model_1, newdata = forecast_data_1)

predictions_naive_1


# Getting forecast dates

forecast_dates_1 <- OLSdata$Date[forecast_start_7 <= OLSdata$Date & OLSdata$Date
<= forecast_end_7]

forecast_dates_1
```

```
# Creating a dataframe with the forecast dates and predictions

predictions_naive_1 <- data.frame(Date = rep(forecast_dates_1, each =
length(predictions_naive_1) / length(forecast_dates_1)), Predicted_naive_1 =
predictions_naive_1)


# View the predictions dataframe

predictions_naive_1

predictions_naive_1$Predicted_naive_1 <- round(predictions_naive_1$Predicted_naive_1,
4)

predictions_naive_1


#comparison

forecast_actuals_1 <- subset(OLSdata, Date >= forecast_start_7 & Date <= forecast_end_7,
select = c("Portfolio"))

colnames(forecast_actuals_1)[colnames(forecast_actuals_1) == "Portfolio"] <- "Actuals"

forecast_actuals_1

comparison_1 <- cbind(predictions_naive_1, forecast_actuals_1)

comparison_1


# Calculate mean squared error (MSE)
```

```
mse_bmk_1 <- mean((predictions_naive_1$Predicted_naive -
forecast_actuals_1$Actuals)^2)

mse_bmk_1

#CSR for first period

setwd("C:/Users/FUJITSU/OneDrive/Dokumente/2. Semester/work project")

getwd()

CSRdata_OOS = read.table("Datafull_OOS.txt", header = T)

CSRdata_OOS_df <- as.data.frame(CSRdata_OOS)

CSRdata_OOS_df



# Subset the data for the training period

train_data_csr_1 <- subset(OLSdata, Date >= train_start_7 & Date <= train_end_7, select =
-c(Date))

train_data_csr_1

subset_reg_1 <- regsubsets(Portfolio ~ ., data = train_data_csr_1, nvmax =
ncol(train_data_csr_1), method = "exhaustive")

subset_reg_1

subset_summary_1 <- summary(subset_reg_1)

subset_summary_1
```

```
#how many variables included

best_subset_1 <- which.max(summary(subset_reg_1)$adjr2)

best_subset_1



#which variables included

best_subset_var_1 <- names(coef(subset_reg_1, best_subset_1)) [-1]

best_subset_var_1



train_data_mod_1 <- train_data_csr_1[, c("Portfolio", best_subset_var_1)]

train_data_mod_1



# Fit a regression model using the training data

model_csr_1 <- lm(Portfolio ~ ., data = train_data_mod_1)

model_csr_1



#predict

predictions_csr_1 <- predict(model_csr_1, newdata = forecast_data_1)

predictions_csr_1



# Getting forecast dates
```

```
forecast_dates_1 <- OLSdata$Date[forecast_start_7 <= OLSdata$Date & OLSdata$Date
<= forecast_end_7]


# Creating a dataframe with the forecast dates and predictions

predictions_csr_1 <- data.frame(Date = rep(forecast_dates_1, each =
length(predictions_csr_1) / length(forecast_dates_1)), Predicted_csr_1 = predictions_csr_1)


# View the predictions dataframe

predictions_csr_1


predictions_csr_1$Predicted_csr_1 <- round(predictions_csr_1$Predicted_csr_1, 4)

predictions_csr_1 <- predictions_csr_1[,2]

predictions_csr_1


#add to comparison

comparison_1 <- cbind(predictions_naive_1,predictions_csr_1, forecast_actuals_1)

comparison_1


#Mse for model

mse_m_1 <- mean((predictions_csr_1 - forecast_actuals_1$Actuals)^2)

mse_m_1
```

```
mse_bmk_1



#Calculation OOS R-squared

OOS_R_1 <- 1- (mse_m_1/mse_bmk_1)

OOS_R_1
```