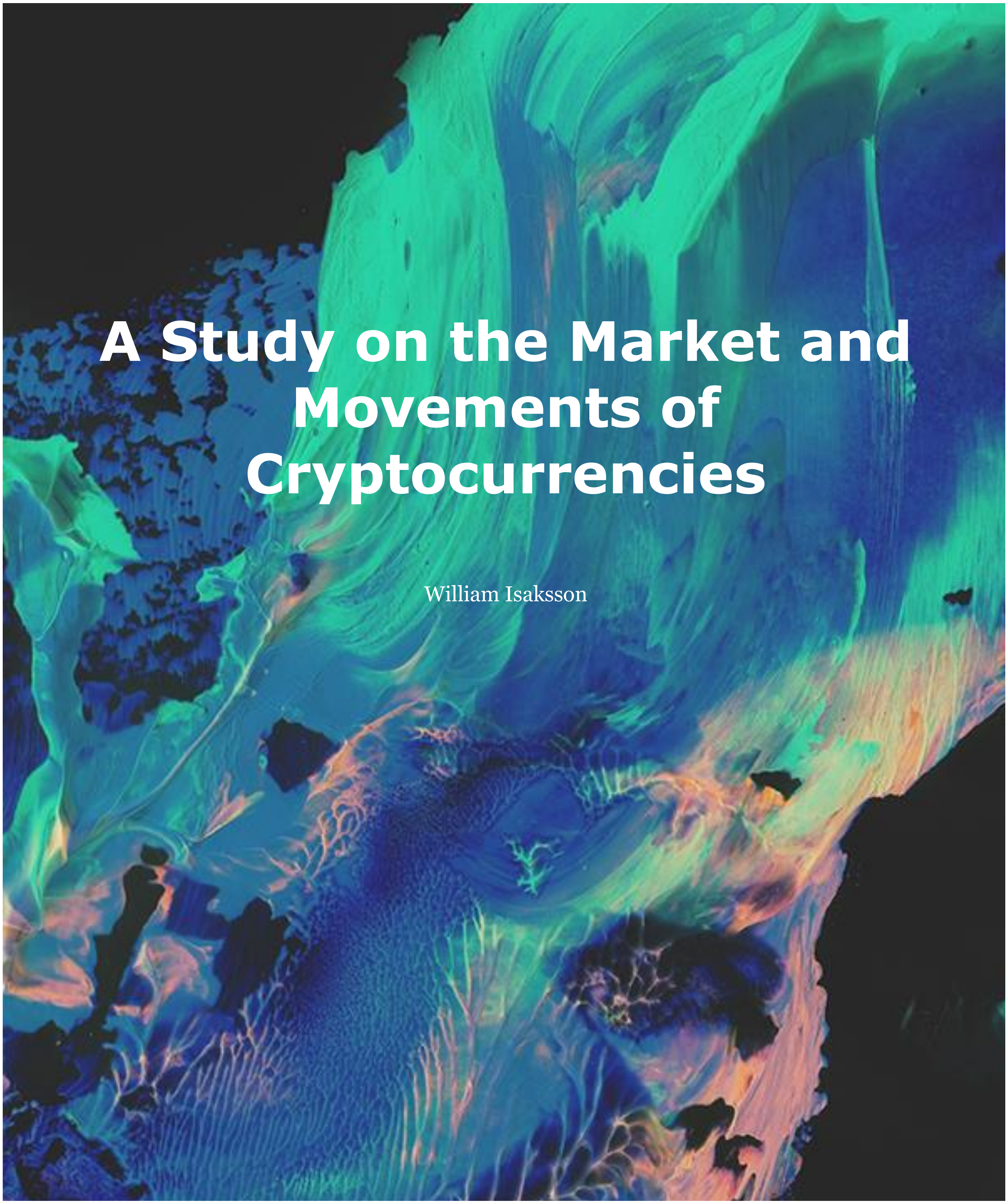
UMEÅ UNIVERSITY



A Study on the Market and Movements of Cryptocurrencies

William Isaksson

Master thesis, 30 ECTS

Master of Science in Industrial Engineering and Management, 300 ECTS

Department of Mathematics and Mathematical Statistics

Spring term 2022

Abstract

There has been much debate among investors on the benefits cryptocurrencies can have for portfolios and how their prices moves in the market. It is not difficult to see that cryptocurrencies are very volatile, yet that does not prevent investors from pouring tons of money in crypto-investments that either generate huge returns or catastrophic losses. One of the main challenges with cryptocurrencies is determining how they move with the rest of the market with assets such as stocks.

The objective with this thesis was to investigate whether or not crypto provides some diversification benefit and if individual cryptocurrencies move in the same manner with respect to each other. Of special interest was if there is a relationship between the cryptocurrency market and the stock market. The cryptocurrencies chosen for this project were compared mostly to the stocks in the, very information technology-sector focused, Nasdaq 100 index along with a few other assets.

This thesis was written in cooperation with Origin Group AB, an Umeå based startup firm specializing in development of cryptocurrency-related technologies, most notably blockchain. All data used comes from publicly available sources and mostly include prices for cryptocurrencies and stocks from which the daily and weekly returns were calculated. The main methods used for this thesis was four different portfolio strategies with different combinations of assets, Style-analysis, and principal component analysis.

The portfolio strategies showed some promise with varying tradeoffs between diversification and Sharpe-ratio but the results are a bit questionable due to the short investment period. The principal component analysis showed that the cryptocurrency price data is very noise and the currencies moves pretty much in unison in contrast to the industry sector divided Nasdaq 100, which seem to have a few more distinct directions of movement. The Style-analysis' inconclusive results show signs of a very noisy dataset and that there may not be a clear linear relationship between conventional asset returns and those of crypto.

Sammanfattning

Investerare har debatterat mycket gällande fördelarna kryptovalutor kan ha för portfolier och hur deras priser rör sig på marknaden. Det är inte svårt att se att kryptovalutor är väldigt volatila men det hindrar inte investerare från att lägga otroliga mängder pengar på krypto-investeringar som antingen genererar stora vinster eller katastrofala förluster. En av de sora utmaningarna med kryptovalutor är att förstå hur deras priser rör sig med resten av marknaden med tillgångar som aktier.

Målet med den här uppsatsen var att undersöka huruvida krypto ger några diversifieringsfördelar och om individuella kryptovalutor rör sig på samma sätt som varandra. Speciellt intressant var ifall det finns ett samband mellan kryptovalutmarknaden och aktiemarknaden. Kryptovalutorna utvalda för det här projektet jämfördes till mestadels med aktierna i det väldigt informationsteknologifokuserade Nasdaq 100 indexet tillsammans med några andra tillgångar.

Den här uppsatsen skrevs i samarbete med Origin Group AB som är ett Umeåbaserat startupföretag som specialiserat sig på att utveckla kryptovalutsrelaterade teknologier med särskilt fokus på blockchain. All data som använts kommer från allmänt åtkomliga källor och innehåller för det mesta priser för kryptovalutor och aktier från vilka dagliga och veckoliga avkastningar beräknades. De huvudsakliga metoderna använda för den här uppsatsen var fyra olika portfoliestrategier med olika kombinationer av tillgångar, Style-analys och principalkomponentsanalys.

Portfoliestrategierna visade sig dugliga med varierande avvägning mellan diversifiering och Sharpekvot men resultaten är någorlunda ifrågasättbara p.g.a. den korta investeringsperioden. Principalkomponentsanalysen visade att kryptovalutornas prisdata innehåller mycket brus och de olika valutorna rör sig nästan enhetligt i kontrast med Nasdaq 100, som innehåller aktier från flera olika industrier vilka rör sig åt flera distinkta håll. Style-analysens icke-övertygande resultat visar också på ett brusigt dataset samt att det kanske inte finns ett tydligt linjärt samband mellan konventionella tillgångars avkastningar och kryptovalutors.

Acknowledgements

Many thanks to my great supervisor, Markus Ådahl, at Umeå University for his excellent guidance and recommendations.

I would also like to thank my company supervisors, Nils Lundberg and Magnus Bostedt, at Origin Group AB for always being supportive and excited about my work.

Finally, kudos to my awesome friends and classmates in the MIT "bunker" at Umeå University for making my final term a great deal of fun!

Glossary

Bitcoin - The first cryptocurrency.
Blockchain - Distributed information “chained” together by a cryptographic hash function.
Byzantine Generals Problem - The difficulty in coming to agreement of the state of a peer-to-peer system when there is a lack of trust among its participants. **Cryptocurrency** A Peer-to-Peer Electronic Cash System (Nakamoto 2008, 1)
Cryptographic Hash Function - A function whose input is uncalculatable from the output.
Consensus - In the context of blockchain, consensus is the agreement that a new, specific block will be added to the chain.
Double Spending Problem - The risk of a malicious party taking control over the consensus mechanism for a cryptocurrency and spends the same money several times.
Fiat currency - Currency issued and backed by a government.
Lasso Regression - Linear regression with the lasso penalty for the weights added to the optimization problem.
Ledger - A list of transactions.
Linear Regression - A simple yet effective machine learning model that attempts to find a linear relationship between variables with features and a response variable.
Log-returns - The logarithm of regular returns.
Minimum Variance - An investment strategy that aims to keep the variance of a portfolio as small as possible.
Node - A network participant holding on to a copy of a blockchain which they frequently update.
Noise - Noise or white noise are disrupting and useless features in data.
NumPy - Library with mathematical tools that can be used with python.
Pandas - Library with an incredible useful data frame tool used with python.
Peer-to-peer network - A network of parties lacking centralization and hierarchies, where the state of the system is decided by the majority. All communication, decisions, and file sharing is done directly between the parties.
Principal Component Analysis - Extraction of the most important features in a dataset.
Proof-of-Stake - Another popular consensus mechanism. A network participant is given the responsibility of adding a new block to the blockchain at a probability equal to how much cryptocurrency they own.
Proof-of-Work - The consensus mechanism used by some of the largest cryptocurrencies. It gives a miner the right to add a new block to the blockchain and reap the reward.
Returns - The change in price from one time to the next.
Risk-Parity - An investment strategy that attempts to make every asset contribute to the same amount of risk to the portfolio.
Stablecoin - A currency functional in cryptocurrency environments pegged to real world currencies or assets. Arguably not a cryptocurrency.
Style-Analysis - Replication of an index or portfolio by forming a linear relationship between asset returns.
UTXO - Unspent transaction output, the cryptocurrency a specific user has the right to spend. Unlike account-based transaction systems, the UTXO-model treats money like physical cash and someone’s balance is the sum of the total UTXO on the blockchain that they have the right to spend.

Contents

1 Introduction

The market of *cryptocurrencies* is one of the most volatile and unpredictable markets there is. The ups and downs of these digital currencies sometimes seem sporadic and random. One simple explanation for this is the fact that most cryptocurrencies lack an obvious underlying asset from which to derive an intrinsic value from. This means that cryptocurrencies are seen as very speculative investments which is shown by the massive volatility of the market, a feature exacerbated by the huge influx of new cryptocurrencies adding to an already great amount of currencies in circulation. Many investors and analysts have tried to price cryptocurrencies fairly and to draw up investment strategies which predicts the market movements. The heavy correlation between the individual currencies and the somewhat lacking correlation between cryptocurrencies and the equity market, have made the cryptocurrency market extraordinary difficult to comprehend.

1.1 Background

1.1.1 Bitcoin and Blockchain

During the aftermath of the 2007-2008 financial crisis, an anonymous developer (or group of developers) under the pseudonym Satoshi Nakamoto created *Bitcoin*, the world's first purely *peer-to-peer* digital payments system functioning completely without any central party (Nakamoto 2008, 1). Bitcoin became the first of the cryptocurrencies and has an uninterrupted record of being the highest priced cryptocurrency since the first bitcoin was minted in 2009 (Coingecko 2022). The idea of Bitcoin was that it would function similarly to physical currency or gold, but digitally, i.e. that transactions can be made directly between two parties, and if participants wish: anonymously and untraceable. Another important feature is that transactions made using bitcoin are permanent and irreversible.

Various ideas of how currencies would work in a digital landscape has been thrown around since at least the 1990s. The way *fiat* currency (unbacked currency issued by governments) works is that financial institutions such as banks keep centralized *ledgers* of all transactions they process and so keeps track of the amount of money belonging to each account in the system. In order to remove the third party of the transaction, the centralized ledger has to become distributed. When the ledger is distributed, multiple parties have their own version of the transaction history recorded on the ledger and when a new transaction is made, every party updates their ledger.

By far the biggest challenge with ledgers in general is that they can easily be tampered with. Banks prevent this by using very secure systems that are difficult to access by intruders, which keep the ledger and bank accounts secure from manipulation. For a distributed ledger to work in the pre-digital era, every single party that own a copy of the ledger has to be completely honest and never make any transactional mistakes and somehow find a way to broadcast each update to every ledger-holder. At the dawn of the information age and with the rise of the internet, some of these problems could be solved. The internet allows updates to be sent out to each party so that they may update their ledger almost instantly. Cryptography allows incredibly secure transactions between anonymous people and as long as users keeps their passwords or private keys safe, no one can touch their funds. This means that distributed ledgers can be updated instantly and user's money can be kept secure with virtually unbreakable cryptography. It also means that no money can be created out of thin air since no more money than the amount available to every user can be spent and this is verified by everyone owning a copy of the ledger. Cryptography can also be used to keep the transaction history of the ledger unchangeable.

Despite advances in communication and cryptography, two major problems persisted which prevented peer-to-peer digital currencies from coming into existence. The first was the issue of broadcasting information on a network when there is difficulty knowing which actors are honest and what information is correct and has not been tampered with. The network needs to be able to form *consensus*, in our case on which version of the ledger is the correct one despite there being some faulty or malicious actors in it. This problem is often referred to as the *Byzantine general's problem* (Reischuk 1985, 1) [1]. The problem can be solved when a majority of the network's participants are honest and there is a way to reach consensus among them.

The second challenge was the *double-spending problem*, an accounting issue resulting from the fact that it is difficult for decentralized networks of parties to form consensus without a third or central party (Hoepman 2007, 152-153). For a peer-to-peer digital currency to function there must be some way for the parties involved to agree on which transactions are valid and which are not. Since there are many parties in the network owning a copy of the transaction history, money cannot be created out of nothing in a way not allowed by the network protocol, the money has to come from somewhere and be accounted for. But this type of verification does not prevent a party from spending the money they already posses twice or more. A dishonest actor may spend some money at one place, then immediately spend it somewhere else, thereby reaping a double reward for the same money and the network cannot keep up with which of the broadcast transactions are valid. There must be some way to form consensus on which transactions are accepted by the network despite the latency in the flow of information and the distrust among participants, without involving a central authority with extraordinary power over the network.

Nakamoto (2008) invented *blockchain* as a solution to the double-spending problem. Blockchain works by having transactions recorded on "blocks" which are chained together using a *cryptographic hash function* which makes it impossible to change the transaction history without ruining the entire blockchain. Users owning the entire transaction history/blockchain are called *nodes* and they are responsible for validating new blocks. To process transactions, Nakamoto (2008) proposed using *proof-of-work* to reach consensus on new blocks. Proof-of-work lets participants in the network use their computing power to *mine*, which is the process of using one's CPUs, graphic cards or other means of computation to generate random values until a value that is accepted by the Bitcoin protocol is found. The miner who manages to find a fitting value will add a new block with transactions to the blockchain and is rewarded in Bitcoin. The nodes will consider the longest blockchain as the correct one and miners will begin mining for a new block on this newer version. As long as the honest miners control more than 50% of the networks computing power, double spending is prevented and the network is considered to have solved the Byzantine general's problem since all honest actors are able to agree on which transactions are valid despite not necessarily

[1] The name is derived from the historical Byzantine Empire although the problem itself is fairly modern and related entirely to computing and communication.

trusting each other and despite the latency regarding information about new blocks. More miners and nodes equals a higher level of security for the network.

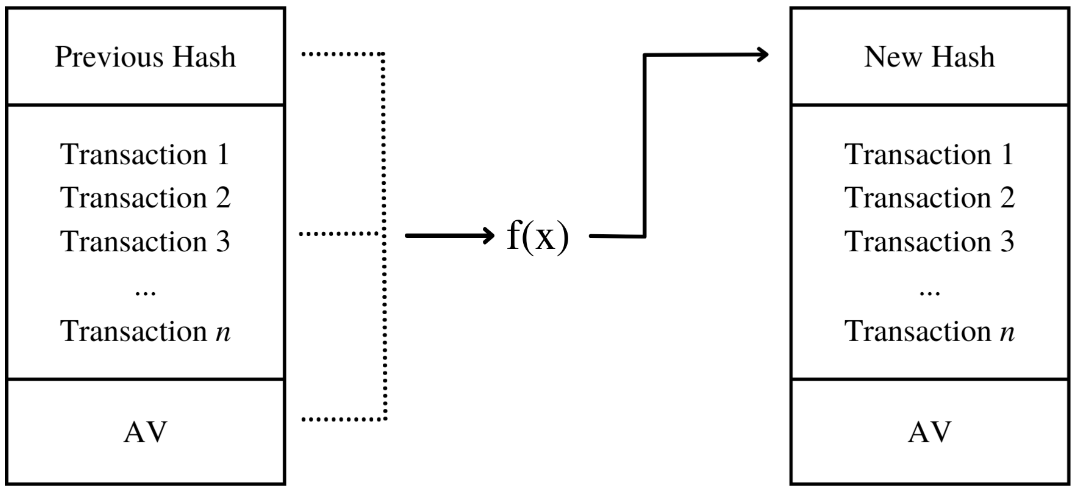


Figure 1: Overview of the blockchain

In figure 1 a visualization of the blockchain is presented. The function, f, represents a cryptographic hash-function which takes an input and outputs a hash which is a string of numbers and letters always of the same lenght despite the lenght of the input. The important feature of a good hash-function (such as SHA-256 in Bitcoins case) is that it is **impossible** to compute the input from the output. Another important feature of a great cryptographic hash-function is that a small change in the input completely changes the output. If any information in the blockchain is changed, the output of the hash-function between every block is changed and the chain "breaks". In a proof-of-work-blockchain, miners attempt to compute a hash output that contains certain features (usually a hash with a select number of zeros in the first characters). This is done by randomly guessing an arbitrary value (AV) which is added, together with all transaction information and the hash from the previous block, into the hash-function. The outputed hash is used as the header of the next block.

Basics of SHA-256:

$$f(x) = hash$$

$$\begin{aligned}
&\text{where} \\
&\quad f = \text{SHA-256} \\
&\quad x = \text{any information} \\
&hash = \text{256-bit hash output}^{2}
\end{aligned}$$

The mined arbitrary value can be verified by the network participants by simply putting in the value themselves in the AV-slot of the latest block to see that it does indeed fulfil the requirements of a valid hash. This is where the term proof-of-work makes the most sense: the miner broadcasts a proof of their work to the network which can be verified incredibly easy by everyone.

Keep in mind that it is the nodes, i.e., the participants with a copy of the blockchain that verifies the transactions. It is therefore very much in the miners best interest to only add valid transactions (and validate the blocks they mine on top themselves) or risk losing the reward and have their computing power wasted. The *Unspent Transaction Output* (UTXO) accounting model makes it easy for nodes to determine exactly which coins exist and who can spend them (River Financial n.d.). It is therefore impossible for a network participant to add a new block that does not have both a valid hash-header (a proof of their work) and valid transactions.

One obvious question arises from all of this: Why mine Bitcoin? We have already established that miners receives a reward from mining. This reward (at least after the first few years of bitcoin's history) usually just covers the cost of mining with a small profit since marginal revenue pretty much equals marginal cost. All the aforementioned rules do not make the network secure unless there are many miners enticed by the business of mining. In this sense, the network becomes more secure from the market forces of supply and demand: as long as it is profitable to mine Bitcoin, people will mine Bitcoin and the network becomes secure.

1.1.2 The Market and Purpose of Cryptocurrencies

In 2009 Bitcoin became the first cryptocurrency. The Bitcoin network would grow during the following years with more participants mining and trading the cryptocurrency, thereby increasing the security and decentralization of the network. It did not take long until the ideas of the cryptocurrency spread across the world and other developers and entrepreneurs followed suit by releasing their own cryptocurrencies. Various cryptocurrency exchanges started to appear as well which made it easier for investors to buy cryptocurrency with fiat/regular currency. By 2013 there were 66 different cryptocurrencies, a number that has since increased to over 10 000 in 2022 when including currencies that has been shut down (Best 2022). The money involved in cryptocurrency skyrocketed and their combined market capitalization peaked in November 2021 at \$2.9 trillion (Coingecko 2022). New cryptocurrencies such as Ethereum, Litecoin, Dogecoin, Cardano and

[2] The output of a cryptographic hash-function, in this case SHA-256, may look like: f(any information) = aa5ef7408a7614f188fee32c8768ecd62771931983b186ae5df95bf465805fdd

Ripple gained market caps of billions and Bitcoin reached over a trillion by itself.

The great number of cryptocurrencies entering the market may be explained by a few reasons:

- There is no general consensus on what the purpose of cryptocurrencies is. Some see them as nothing more than a storage of value while others view them as an instrument that will completely replace fiat currency as the main method of payment and a technology that may usurp the entire financial system. Others want to expand blockchain technology to other industries and view cryptocurrency as merely the first step in the rise of the blockchain.

- There is great disagreement on how the optimal cryptocurrency should look like. Almost every aspect of the cryptocurrency and blockchain technology has been hotly debated and many developers created coins that has features that they themselves consider optimal. The consensus mechanism for example has experienced much revision and the energy demanding proof-of-work have been replaced by many cryptocurrencies with options such as *proof-of-stake*.

- Low cost of entry with a great upside. The early investors in Bitcoin saw massive returns as the currency gained popularity and many have attempted to replicate these gains by investing early in new crypto. Many market participants raised millions with the so-called *initial coin offering* (ICO), an event where a new cryptocurrency is released to the public. $19.7 billion was raised through ICOs in 2018 (Liu, McConnel and Wang 2021, 1). There are few measures in place by governments to regulate the cryptocurrency market (although they have increased in recent years) and the cost to enter the market is low since much of the technology used in crypto is open-source and may be replicated by anyone.

Given that cryptocurrencies such as Bitcoin have properties resembling gold, investors may use them as a hedging instrument against market movements. However, as shown by Klein, Pham Thu and Walther (2018), Bitcoins market price behaves nothing like that of gold and correlates positively with market movements. At the moment, it is very difficult to asses cryptocurrencies' place as an investment. Unlike stocks or bonds, there are no underlying assets and cash-flows in crypto. The challenge to fairly price cryptocurrencies' intrinsic values are further complicated by the many cases of price manipulations, often caused by shady crypto-exchanges (Peterson 2021, 254-269). An investor may get a substantial trading advantage if the seemingly sporadic movements of cryptocurrencies' market prices can be more accurately decoded.

1.2 About Origin Group

Origin Group AB is a startup firm in Umeå which specializes in blockchain technology. Their vision is to apply blockchain in problem solving and to widen the use of the technology. They have proposed blockchain as a tool to solve problems related to workflow, supply chains, cybersecurity, data storage and handling, and more.

1.3 Problem

The problem at hand for investors in cryptocurrency is the volatility of the market and the future of the technology. If cryptocurrencies can be proven to move in a clear way with other publicly traded assets, this uncertainty can be mitigated. This can be shown in many ways and if no correlation can be found, there may be some great potential for cryptocurrencies as tools for diversification. Portfolios partly made up of cryptocurrencies could prove to have high return and be diversified.

1.4 Purpose of Thesis

This thesis will shed light on whether the crypto market is inherently random or if there is some underlying rationality in its movements. We will attempt to answer a few questions which are of importance to the problem:

- Can cryptocurrencies be used for portfolio-diversification?
- Are there similarities between the market structures of Nasdaq 100 and cryptocurrencies?
- Can the cryptocurrency market be replicated with a portfolio of stocks, bonds and gold?

1.5 Delimitations

Given the scale of the world of cryptocurrency, the thesis has its focus primarily on the largest cryptocurrencies. The technology of blockchain and crypto is ever changing and some things progress very fast while other factors remain the same. This thesis is not be too concerned with the social implications of crypto or what the technology's optimal use cases are, nor subjective questions like what really gives the currencies their value unless those things are measurable, comparable and of relevance to the subject at hand (although this topic is revisited briefly in the conclusion). The subject of cryptocurrencies is complicated and full of speculation, rumours, and misinformation. Therefore, this thesis only includes valid market data as a ground for all results and conclusions, as well as peer reviewed studies about relevant subjects among other sources. Hard facts about the different currencies are considered. Much of the inner workings of blockchain and cryptocurrencies was covered briefly in the introduction but not in great detail since the technology is very complicated and this thesis has its main focus on the general market and the cryptocurrencies' role there.

There is market data on cryptocurrencies spanning all the way back to 2013 but in order to make a portfolio consisting of a fair number of cryptocurrencies the time period covered by this thesis is significantly shorter. This is to make room for more currencies as well as using a time period where the market is more "mature". Equity data cover the same time period. The main data compared and analyzed together with cryptocurrency data are individual stocks from a chosen index, along with bonds and gold. More information about the market data is covered in the methodology.

2 Theory

2.1 Basic Theory

The returns, r, are the change in price, p, from one point in time to the next for example between days of weeks.

$$r_t = \frac{p_t}{p_{t-1}} - 1$$

and logarithmic returns:

$$y_t = ln(r_t + 1)$$

There are a few reasons for why we sometimes want to use the natural logarithm of the returns instead of regular returns. The most important is that since stock prices are often assumed to be log-normal (the prices' logarithms are normally distributed) it follows that the logarithms of the returns are also normally distributed.

For jointly distributed random variables (that is normally distributed returns in our case), the estimated variance for an asset's returns:

$$Var(\boldsymbol{x}) = \sigma_{\boldsymbol{x}}^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n}$$

and similarly the covariance between returns for two assets is defined as:

$$Cov(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

where $\bar{x}$ and $\bar{y}$ are the means of the vectors of returns $\boldsymbol{x}$ and $\boldsymbol{y}$. The covariance between a variable and itself is that variable's variance. In this thesis the covariance matrix will be denoted by Σ. Imagine we have n assets with returns $\boldsymbol{x_i}$, then the covariance matrix is defined as:

$$\Sigma = \begin{bmatrix} Cov(\boldsymbol{x_1}, \boldsymbol{x_1}) & \dots & Cov(\boldsymbol{x_1}, \boldsymbol{x_n}) \\ \vdots & \ddots & \vdots \\ Cov(\boldsymbol{x_n}, \boldsymbol{x_1}) & \dots & Cov(\boldsymbol{x_n}, \boldsymbol{x_n}) \end{bmatrix}$$

Correlation is as a quick way to determine a linear relationship between variables and is defined by this formula (Hastie et al. 2021, 70):

$$\rho_{\boldsymbol{x},\boldsymbol{y}} = \frac{Cov(\boldsymbol{x}, \boldsymbol{y})}{\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}}}$$

Whenever a mean value is calculated within the framework of this thesis, the mean in question is the arithmetic mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2.2 Portfolios

Four main investment strategies were used in this thesis for the portfolio evaluation. None of them includes shorting of assets, in other words, no weights are assigned negative values. For the minimum variance strategy and the risk parity strategy portfolios, variance is needed for the calculations which means that we need an estimation period of historical returns. This ”window in time” is moved forward one time-step every time the portfolio must be rebalanced. For all portfolios, the portfolio return is defined as $\boldsymbol{w}^T\boldsymbol{r}$ where $\boldsymbol{r}$ and $\boldsymbol{w}$ are the individual assets returns and weights respectively.

2.2.1 Buy and Hold - Equal Weights

This strategy involves investing an equally large amount in each asset with no rebalancing of the portfolio.

2.2.2 Market Capitalization Weights

The investment in each asset is proportional to the assets market capitalization relative to every other asset in the portfolio. For every asset in the portfolio, divide the estimated market capitalization of the asset with the sum of the estimated market capitalization of every other asset. This strategy involves updating the asset weights at every time step.

2.2.3 Minimum Variance Weights

A minimum variance strategy generally involves investing in such a way that the portfolio minimizes the risk. Since the variance for a random variable is defined as:

$$Var[X] = E[(X-\mu)^2]$$

the variance of the portfolio return, $\boldsymbol{w}^T\boldsymbol{r}$, with the weights, $\boldsymbol{w}$, and returns, $\boldsymbol{r}$, can be extended:

$$\begin{aligned}
Var[\boldsymbol{w}^T\boldsymbol{r}] &= E[(\boldsymbol{w}^T\boldsymbol{r} - \boldsymbol{w}^T\boldsymbol{\mu})^2] \\
&= E[(\boldsymbol{w}^T(\boldsymbol{r}-\boldsymbol{\mu}))(\boldsymbol{w}^T(\boldsymbol{r}-\boldsymbol{\mu}))^T] \\
&= E[\boldsymbol{w}^T(\boldsymbol{r}-\boldsymbol{\mu})(\boldsymbol{r}-\boldsymbol{\mu})^T\boldsymbol{w}] \\
&= \boldsymbol{w}^T E[(\boldsymbol{r}-\boldsymbol{\mu})(\boldsymbol{r}-\boldsymbol{\mu})^T]\boldsymbol{w} \\
&= \boldsymbol{w}^T\Sigma\boldsymbol{w}
\end{aligned}$$

where $\boldsymbol{\mu}$ is the mean of the returns over the estimation period and Σ is the covariance matrix of the returns over the estimation period. To update the portfolio weights, we solve a minimization problem where the variance of the portfolio at a given time is set as the objective function with the constraints dictating that all weights must be positive or equal to zero and that no value may be added or removed from the portfolio (when rebalancing the portfolio, we can only use the funds already present in it).

$$\min_{\boldsymbol{w}} \boldsymbol{w}^T \Sigma \boldsymbol{w}$$
$$\boldsymbol{w} \geq 0$$
$$\sum_{i=1}^{n} w_i = V$$

where V is the value of the portfolio. This optimization problem can be solved with quadratic programming.

2.2.4 Risk Parity Weights

The idea with a risk parity portfolio is that each asset contributes an equal amount of risk (Carli, Deguest and Martellini 2014, 24). Let the function $R(\boldsymbol{w})$ represent the risk of a portfolio with weights $\boldsymbol{w}$. Then the marginal risk contribution (MRC), i.e., the change in the portfolio risk given a change in $\boldsymbol{w}$, is:

$$MRC_i = \frac{\partial R(\boldsymbol{w})}{\partial w_i}$$

for every asset in the portfolio. Of course, risk is usually measured in volatility so for our intents and purposes we let $R(\boldsymbol{w})$ be the square root of the variance:

$$R(\boldsymbol{w}) = \sqrt{\boldsymbol{w}^T \Sigma \boldsymbol{w}}$$

which by rules of differentiation gives us this expression for marginal risk contribution:

$$MRC_i = \frac{(\Sigma w)_i}{\sqrt{\boldsymbol{w}^T \Sigma \boldsymbol{w}}}$$

The total risk contribution (TRC) is defined as:

$$TRC_i = w_i MRC_i$$

The point with a risk parity portfolio is that every asset stands for an equal amount of risk and with the total risk contribution defined we want:

$$TRC_i = TRC_j$$

to be true for every i and j. In essence, this means that for every i, we want the portfolio volatility to be equally split among the n asset allocations:

$$TRC_i = \frac{\sqrt{\boldsymbol{w}^T \Sigma \boldsymbol{w}}}{n} \iff TRC_i \frac{1}{\sqrt{\boldsymbol{w}^T \Sigma \boldsymbol{w}}} = \frac{1}{n}$$

Now we have a target function:

$$w_i \frac{(\Sigma w)_i}{\sqrt{\boldsymbol{w}^T \Sigma \boldsymbol{w}}} \frac{1}{\sqrt{\boldsymbol{w}^T \Sigma \boldsymbol{w}}} - \frac{1}{n} = 0$$

The risk parity allocation can be obtained by solving this unconstrained minimization problem where the sum of squares is minimized:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} (w_i \frac{(\Sigma \boldsymbol{w})_i}{\boldsymbol{w}^T \Sigma \boldsymbol{w}} - \frac{1}{n})^2$$

2.2.5 Portfolio Performance Metrics

Sharpe ratio The Sharpe ratio is a quick and comprehensive metric for measuring a portfolio of assets risk compared to its return. The ratio uses standard deviation (volatility) of returns as the metric for the assets risk. It is essentially a measurement of a portfolios manager's risk awareness. A portfolio invested entirely in crypto in 2016 would have yielded spectacular returns but probably with a low Sharpe ratio; the portfolio return was only possible by taking on a great risk and the market moved in the manager's favour. Therefore, the savvy investor would be looking for a great Sharpe ratio for his or her portfolio since that implies high returns compared to the risk.

The Sharpe ratio, S, is defined as the mean return of the portfolio over the investment period, $\bar{r}_p$ with the risk-free rate, r_f, subtracted, divided by the standard deviation of the returns during the investment period.

$$S = \frac{\bar{r}_p - r_f}{\sigma_p}$$

A portfolio with a yearly Sharpe ratio larger than one is generally considered good. Returns are also assumed to be normally distributed which is why the Sharpe ratio can be annualized from daily values. The assumed normal distribution is one of the weaknesses of the Sharpe ratio since assets not adhering to that distribution may exhibit great kurtosis (DeCarlo 1997, 1-3) with fat tails.

Effective Number of Constituents A useful metric for measuring a portfolios diversification is the effective number of constituents, ENC. The ENC is the inverse of the Herfindahl-Hirschman (sum of the portfolio weights squared) (EDHECinfra Docs n.d). For n assets the ENC at any time is:

$$ENC = \frac{1}{\sum_{i=1}^{n} w_i^2}$$

This tells roughly how many of the portfolio's asset that actually have a sizeable impact on the return. A low ENC indicates that just a few assets stand for nearly all the movement in price. Table 1 shows the difference in ENC for a diversified and less diversified portfolio.

Maximum Drawdown Maximum drawdown (MDD) is simple yet intuitive measure of the largest historical loss. All we need to do is look at the largest plummet in price an asset or portfolio have had historically and divide it with the highest price up to that point.

Table 1: The portfolio with a single constituent making up 70% of the value has a low ENC.

	w_1	w_2	w_3	w_4	ENC
Portfolio 1	0.25	0.25	0.25	0.25	4
Portfolio 2	0.7	0.1	0.1	0.1	1.923

$$MDD_T = \max_{t \in [0,T]} \frac{DD_t}{CM_t}$$

where DD_t is the largest drop in price from the historically largest price CM_t. This metric is useful over a longer time period but is lacking in the short term due to fewer highs and lows in the price history to analyze.

2.3 (Multiple) Linear Regression

Linear regression is a statistical method for supervised learning. It is useful for evaluating if there is a linear relationship between a number of variables and a response variable and which of these variables that matter the most (Hastie et al. 2021, 60). Linear regression fits a model by "training" on a dataset with variables and response variables to fit the model accordingly. The trained model may be used to predict future response variables using new data inputs.

Here the response variable *y* is approximately modeled (regressed) by *x*, estimated coefficients/weights *w*, and the error term *e*.

$$y = w_1 x_1 + w_2 x_2 + ... + w_j x_j + e$$

which can be written on the vector form:

$$\boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{w} + \boldsymbol{e}$$

where $\boldsymbol{x}$ is a vector with features from some dataset, $\boldsymbol{y}$ the response variables to the same dataset, and $\boldsymbol{w}$ is a vector of the weights.

The error term is defined as the difference between the actual value of the response variable, y, and the estimated, $\hat{y}$:

$$e_i = y_i - \hat{y}_i$$

The weights are estimated exactly using the *least squares method*. The least squares method is a solution to the minimization problem where the residual sum of squares are minimized:

$$\min_w \sum_{i=1}^{n} (\boldsymbol{w} \cdot \boldsymbol{x_i} - y_i)^2 = \min_w \sum_{i=1}^{n} e_i^2$$

Taking the gradient of this expression equal to zero from which we can derive this expression:

$$\boldsymbol{w} := (\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}^T \boldsymbol{y}$$

which is how the weights are updated in ordinary linear regression. These weights are then used to guess the next y_i and the error is used to update the weights again. This process is what we call training and by some arbitrary amount of training rounds, the model should be able to estimate a $\bar{y}_i$ from $\boldsymbol{x}_i$.

2.3.1 R-Squared Test and Bias

A way to evaluate the performance of our trained linear regression model is the R^2 statistic, commonly known as the R-squared test. It always takes on a value between zero and one and measures the proportion of variance in y that can be explained by x (Hastie et al. 2021, 70). The R-squared test is a metric that show how close the datapoints are to the fitted regression line and is the explained variation divided by the total variation (Minitab blog editor 2013).

$$R^2 = 1 - \frac{Var(e_i)}{Var(y_i)} = 1 - \frac{RSS}{TSS}$$

The explained variation is equal to the residual sum of squares (RSS) and the total variation is equal to the total sum of squares (TSS) (Hastie et al. 2021, 70):

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = RSS$$

$$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = TSS$$

where $\bar{y}_i$ is the mean of the response variable y_i.

A high R-squared value is generally considered good since it means that the model fits the data quite well. However, R-squared tells us nothing about the bias of the weights and a high R-squared does not always mean that the model is accurate nor does a low value mean that the model is bad. In fact, a model with high R-squared can be interpreted as overfitted (Minitab Blog Editor 2013). Indeed, Hastie et al (2021, 70) mentions that it is challenging to determine what a good R-squared value really is and that it largely depends on the problem at hand.

It is difficult to fit a model that has both low variance and low *bias*. If we estimate a function, $\hat{f}$, to fit a dataset then variance is the amount $\hat{f}$ would change its output if we used another dataset as input. Bias on the other hand is a more systematic type of error, which occurs when $\hat{f}$ consistently misses the mark and predicts an output that is off by some b. Imagine obtaining $\hat{f}$ by just drawing a line that goes through every datapoint in a dataset. This model would have low variance but probably fail spectacularly on other datasets with its high bias (Hastie et al. 2021, 35-36).

2.4 Lasso Regression

According to Hastie et al. (2021, 75), when performing linear regression, there are a few important questions that we are interested in answering. One of these is obviously how well linear regression model fits the data and as the reader is now aware, this will be partially answered by the R-squared test. Another important question is which features in $\boldsymbol{X}$ are actually useful in predicting the response variable $\boldsymbol{y}$. One method to decrease the number of predictors in our linear regression model is to add regularization in the form of a convex penalty term which in our case is the *lasso*. This term shrinks the weights towards zero at a rate decided by λ which may have the effect of decreasing variance. In the case of the lasso regularization, the method completely reduce some features weights to zero. This is how linear regression becomes *lasso regression* and the penalty/lasso is defined as:

$$\lambda \sum_{j=1}^{p} |w_j|$$

where $\lambda \geq 0$. This is added to the minimization problem:

$$\min_{w} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} w_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |w_j|$$

The tuning parameter, λ, is determined separately from the minimization problem and can be decided with a trial and error process known as cross validation (Hastie et al. 2021, 251) or by arbitrary selection, and its fixed value depends on the output and performance of the model. A λ close to zero will render the penalty term less effective and we will end up with a model resembling ordinary linear regression. A large λ will cause all weights to approach zero (Hastie et al. 2021, 237).

Lasso regularization leads to models which are simpler and more easily interpreted since the penalty removes entire features from the model instead of simply shrinking them close to zero. We may also end up with a model that is better at predicting response variables for new data sets since some unimportant features are removed. However, variance may be decreased with a penalty term such as the lasso but *bias* may increase. Furthermore, the ordinary least squares method is not sufficient to solve the minimization problem when the lasso term is in place with a positive λ. The method used by Scikit-Learn is a version of *coordinate descent* outlined by Friedman, Hastie and Tibshirani (2010, 1-5) which is a numerical optimization algorithm. The coefficient update rule is a bit complicated and outside of the scope of this paper but the bottom line is that the penalty term becomes a constraint for the minimization problem and by property of the lasso, leading some of them to be zero.

2.5 Style-Analysis

Style-analysis was defined by William F. Sharpe in 1988 as a statistical model to evaluate investment strategies regarding asset allocation and investment behaviors. Since the model is very intertwined with fund-analysis, the word “style” comes from the fact that it is indeed the investment/fund managers style that is being evaluated. While a fund manager may want to evaluate his or her funds returns sensitivity to different asset classes, in this thesis we are interested in using conventional assets to explain the return of a portfolio consisting solely of cryptocurrencies. The return y on time step t is explained by n asset returns, x, with sensitivities/weights, w. The unexplained returns are denoted by e.

$$y_t = w_1 x_{t1} + w_2 x_{t2} + ... + w_n x_{tn} + e_t$$

If the reader finds the expression above to be eerily similar to linear regression it is only because it absolutely is. Sharpe (1992) uses quadratic programming (least squares) to calculate the weights and the R-squared test as a performance metric of the model. Important to note is that if a weight is negative, the model has decided to sell that specific asset short. In some cases, negative weights are unwanted since in order to present a coherent and replicable investment strategy, having a lot of shorted assets can cause overcomplicated models which are to diffuse for an investor. Therefore, the constraint $w \geq 0$ is added to both the linear and lasso regression models.

2.6 Principal Component Analysis and Noise

A popular approach to reduce the number of features in a dataset is *principal components analysis* (PCA) (Hastie et al. 2021, 252-253). The main idea with PCA is to create a new set of features from $\boldsymbol{x}_i$ with fewer dimensions. These new features are called principal components and can be ranked from the largest to smallest. The largest (first) principal component stands for the direction in which the data vary the most (Hastie et al. 2021, 253), the second largest for the second most varying direction that is orthogonal to the first. Unlike linear and lasso regression which are supervised learning-models (they train on labeled datasets), principal component analysis requires no y_i and is therefore an unsupervised learning-model (Hastie et al. 2021, 253).

To determine the direction of the new feature space of principal components we calculate the eigenvectors with eigenvalues describing the variance of the data in the directions of the principal components. The eigenvectors determine the principal components directions (Raschka 2015). The eigenvalues and vectors are typically calculated from the covariance or correlation matrix, but in the field of finance (which we are in) the correlation matrix is most often used (Raschka 2015).

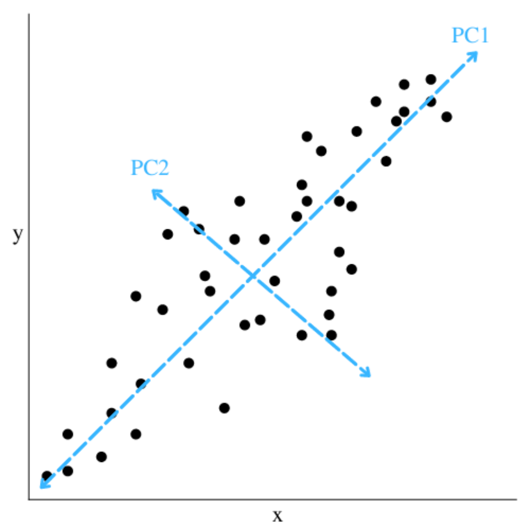


Figure 2: Principal components for a two dimensional dataset. PC1 has large corresponding eigenvalue and PC2 a smaller one.

Correlation and covariance matrices are square, symmetric, and made up of real numbers making them *real symmetric*:

$$A = A^T$$

By this property, every correlation and covariance matrix have eigenvalues which are real numbers (Fitzpatrick 2011). Eigenvalues are defined as:

$$AX = \boldsymbol{\lambda} X \iff (A - \boldsymbol{\lambda}\mathbb{1})X = 0$$

where $\boldsymbol{\lambda}$ (unrelated to the tuning parameter in lasso regularization) is an eigenvalue of A, $\mathbb{1}$ the identity matrix, and X is an eigenvector of A. To compute the eigenvalues and eigenvectors we take the determinant of this expression:

$$|A - \boldsymbol{\lambda}\mathbb{1}| = 0$$

This expression can be very hard to solve for a high dimensional A. NumPy's linear algebra functions uses the LAPAC routines (NumPy n.d) to calculate eigenvalues and vectors for real symmetric matrices, using a divide and conquer algorithm (LAPACK 2022).

According to Bouchaud and Potters (2000), the structure of the correlation matrix from market data is inherently dominated by *noise*. Noise can be defined as the unexplained variability in the data sample and is typically random and meaningless (Wigmore 2017). We can reduce this noise by isolating the k most important factors (principal components) from the total number of principal components, N. The rule for deciding when a factor (PC) is considered to be useless noise is according to Carli, Deguest, and Martellini (2014, 24) when the corresponding eigenvalue fall below this threshold:

$$\lambda^+ = 1 + \frac{N}{T} + 2\sqrt{\frac{N}{T}}$$

where T is the number of observations in the dataset the correlation matrix is calculated from. The formula itself is derived from random matrix theory (Carli, Deguest, and Martellini 2014, 24).

2.7 Linear Interpolation

Missing values in the datasets are replaced by interpolated values using *linear interpolation*. Since the data used in this thesis is two-dimensional (prices and equally spaced time stamps) the interpolation method is very intuitive: a missing value between two existing values is replaced by the average of those two values. Which means that values are calculated as equally spaced on the line between points (x_0, y_0) and (x_1, y_1). For cases where there a several missing values in a row, values are filled forward.

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0}$$

2.8 Theoretical limitations

It is difficult to use linear models for real-life problems since variables generated from situations occurring in reality are not as linear as we would like them to be, and they do not necessarily adhere to any known distribution. In the world of trading, we usually assume that variables are normally distributed which is very convenient as it allows for the use of every theory and model (risk parity, correlation, sharpe ratio) that assumes normality. In reality, prices may not be normally distributed.

If there is no linear relationship between X and Y, it does not matter how much data our linear or lasso regression has been trained on, an accurate estimate will not be found (Hastie et al. 2021, 35). This leads us to the bias-variance trade off; a model can not have both low variance and low bias when trained on real world data (Hastie et al. 2021, 35).

3 Methodology

3.1 Approach

A quantitative approach is the main methodology of the paper. This means that the results will be based entirely on analyses of collected quantitative data rather than on collected literary data. Although literature will complement the paper in many ways, the theoretical framework and results will only be used in regard to quantitative models.

3.2 Data

Literature has mainly been collected from online sources as well as from the library of Umeå university and material suggested by thesis-advisor. Special emphasis was put on literature originating from well known academic journals. Cryptocurrency market data comes from Coingecko.com which was retrieved using the sites free application programming interface and an unofficial Python API wrapper (Emmanouil 2018). Market data was collected through Yahoo! finance and its webpage's API using an unofficial Python API wrapper (Aroussi 2022). Quantitative data includes daily prices of cryptocurrencies, stock prices, bond prices, and the price of gold, spanning the period 2018-07-01 to 2022-05-15. The price at the closing time of the stock exchanges in New York (16.00 eastern time) is always used as the daily price for assets retrieved through Yahoo! finance. Coingecko retrieves its cryptocurrency data from various crypto-exchanges and these instruments are traded around the clock, every day including weekends and holidays. The daily prices of cryptocurrencies are taken at 00.00 UTC.

The 20 largest cryptocurrencies by market capitalization at the date 2018-07-01 were retrieved using the historical snapshots of prices at Coinmarketcap.com. The same cryptocurrencies were retrieved in the same manner once more 101 days before at 2018-02-06. This is due to the fact that when updating the weights in the minimum variance and risk parity portfolios we need a time frame to calculate variance, mean and covariance and this time frame was conveniently chosen to be 100 days (the extra day is present because from 100 prices only 99 returns can be calculated). By starting 100 days early, we are able to calculate the portfolio weights on day one. The cryptocurrencies used are the 20 largest by market capitalization but some of them at the date 2018-07-01 are excluded. Unused cryptocurrencies in the top 20 include stablecoins, coins that were not traded on 2018-02-06, and coins that have since gone defunct, i.e., they are no longer traded. These unused coins have been replaced with the next largest cryptourrencies outside the top 20 by market capitalization.

The stock data used for the analysis are the stocks that make up the Nasdaq 100-index. These are the market prices of a hundred companies in a handful of industries, most in that of information technology. The other assets used include three bond exchange traded funds; 20 year- and 7-10-year treasury bonds, and high-yield corporate bonds. The final asset used is gold. Bonds and gold is only used for the Style analysis. Out of the 102 stock prices (two of the companies are listed twice with different stock classes), ten were dropped from the analysis since they were listed on Nasdaq after 2018-02-06.

3.2.1 Data Processing

The Python-packages Pandas and NumPy were used to process and arrange data. Missing values in the data was filled using forward-linear interpolation. Returns and logarithmic returns of every crypto and asset type was calculated using Pandas built-in tools. Data results were presented using Pandas built-in tools for plotting as well as the python package Seaborn.

3.2.2 Data Limitations

Asset data from Yahoo! finance cover every trading-day in the desired time period. The cryptocurrency price data covering non-trading days from Coingecko have been removed and is not used in this thesis. The cryptocurrency prices and asset prices have a three-hour gap between their timestamps. This is because the daily prices for cryptocurrencies are only available at 00.00 UTC while the closing price of the assets used for this thesis are calculated at 16.00pm eastern time (the time trading closes in New York). 16.00pm is equal to 21.00 UTC which is three hours earlier than the calculation time for cryptocurrencies. This discrepancy in time may affect the results in different ways, especially if the cryptocurrency prices change drastically during the three-hour gap. On the other hand, the asset prices traded in New York does not change during the time the market is closed. This issue is further mitigated by the fact that only weekly prices are used in the Style-analysis, rendering the time gap less significant.

As stated before, cryptocurrencies are trade around the clock, every day. This means that the cryptocurrencies' prices can change during weekends and holidays while the stock prices remain unchanged. The cryptocurrencies can then react to events happening during the non-trading days which stocks will react to only when the market opens again. This may cause some issues for the results of this thesis, and it is something to keep in mind. Again, prices for every asset used are only retrieved from days when the market was open, and the main data used for the analyses are the change in price from one time step to the next (returns). This change in price hopefully captures the market activity happening on holidays and weekends well enough both for cryptocurrencies and other assets.

According to Peterson (2021), the cryptocurrency market is heavily affected by price manipulations which can make any legitimate analysis problematic in its nature. The data used in this thesis originated from several cryptocurrency exchanges and it is likely that some data may come from an exchange that at some point has been involved in price manipulation. This problem is not as serious for assets like equity since it is difficult for one stock to affect an entire index, while a manipulation of the Bitcoin price can be damaging to the entire cryptocurrency market.

3.3 Portfolio Analysis

For the portfolio analysis, 20 cryptocurrencies and 92 stocks are used in total. In total, eight portfolios have been created and evaluated with the four different strategies presented under the theory chapter:

- One using buy-and-hold-equal weights

- One using market capitalization weights
- Three using minimum variance
- Three using risk parity

Every portfolio uses daily returns both for crypto and conventional assets. The buy and hold portfolio with equal weights and the market cap proportional portfolio were intended as indexes representing the cryptocurrency market as a whole and therefore only includes cryptocurrency. The returns from the market cap proportional portfolio are used later on as the response variable in the Style-analysis.

As stated in the theory chapter about the portfolio strategies, no value can be added or removed from the portfolio and so every portfolio which are updated every time step must only use the value, V, already present in the portfolio. For the portfolios that are updated every trading day, we assume that there are no transactional costs. V is the sum of all percentage weights multiplied with their corresponding asset price. The initial investment, V_0, is always \$100. The python library SciPy includes functions for solving minimization problems and was utilized for the risk parity minimization problem is unconstrained and is solved by SciPy's method "SLSQP" which stands for "sequential least squares programming". A full explanation of this method is beyond the scope of this thesis but importantly, it does allow for unconstrained optimization and lets us choose an initial guess. At each time-step where the weights are updated using the risk parity strategy, the optimizer will start with each asset weight w_i is assigned $\frac{1}{\sigma_i}$ where the volatility σ_i is calculated from the previously discussed estimation period of historical returns (without this initial guess the optimizer wont converge). CVXPY is the python package used to solve the minimum variance optimization problem. The optimization method used is also SLSQP which is the standard method for minimization problems when using CVXPY and allows us to insert the constraints outlined in the theory (2.2.3).

The minimum variance and risk parity strategies were used for three sets of assets each. Since these portfolios use mathematically derived allocations (remember, small variance and equal risk), they have been tested on more asset combinations than the previous, more passive strategies. These two strategies were tested on portfolios with:

- Only cryptocurrency
- Only stocks
- Stocks and cryptocurrencies combined

The metrics used to evaluate portfolio performance is the Sharpe-ratio, effective number of constituents, and maximum drawdown. The Sharpe-ratio and maximum drawdown are calculated from the entire investment period and the ENC is taken as the arithmetic mean of the ENC for every time step. For the Sharpe-ratio calculation the risk-free rate chosen was the annual yield of ten year United States treasury bills. The portfolio values are saved in a Pandas data frame and the weight of each portfolio (eight in total) are saved data frames as well.

3.4 Style Analysis

Style-analysis was performed with linear regression and lasso regression with the R-squared test as evaluation and both with and without the $\boldsymbol{w} \geq 0$ constraint. The features, X, is a matrix of weekly returns from the same Nasdaq 100 stocks used in the portfolio analysis together with 20+ year treasury bills, seven to ten years treasury bills, high yield corporate bonds, and gold. $\boldsymbol{y}$ consists of logarithmic returns from the market cap weighted portfolio of cryptocurrencies and was chosen as the response variable because the Nasdaq 100 is weighted in a similar way where the largest stocks have the largest influence over the index. In this way, the style analysis attempt to create a linear model of assets that replicate the cryptocurrency portfolio at each time steps. For the lasso regression, the tuning parameter λ, is arbitrarily set to 0.001.

The time steps are every week from 2018-07-01 to 2022-05-15 with the first 100 weeks used to train the first model which means that the first allocations are calculated around July 2020. Every week the "window" is moved forward one week and the model is retrained from scratch with the next 100 weeks. The logarithmic returns are weekly which means that there are five trading-days between p_t and p_{t-1}. The python package Scikit-Learn was used for the regression and for the R-squared test. The R-squared test uses the model's prediction of the same period the model was trained, that is, the model is trained on y_i and predicts $\hat{y}_i$. This will tell us how much the returns of the cryptocurrency market can be explained by the returns of the assets we have chosen (Sharpe 1992).

The allocations of the Style-analysis (the weights from the lasso and linear regressions) is what we are most interested in since it allows us to see how the stock composition changes over time and which stocks and industry sectors are most related to the cryptocurrency market. The allocations are presented with area plots with the weekly dates as the x-axis and the investment weights in the y-axis. Stocks with really small coefficients will not be visible in the area plot and are therefore not presented in the plot legends. The R-squared tests are plotted as lines, also with the weekly dates as the x-axis.

Keep in mind that Style-analysis does not find optimal investment strategies, rather it simply tries to replicate the market movement of one asset/index/portfolio with linear combinations of others. What our Style-analysis basically says is: "during these one hundred weeks, this combination of assets returns most closely resembled the returns of the cryptocurrency market".

3.5 Principal Component Analysis

Principal component analysis can be used for risk management to mitigate financial losses resulting from macroeconomic factors and to know which large movements in the market to diversify or hedge against. Although the correlation matrix is fairly useful for examining how assets move together, a big part of the matrix may be dominated by noise and is therefore not adequate for risk management (Bouchaud and Potters 2000, 83-84). For the principal component analysis in this thesis, the data used is the same as in the portfolio analysis. That is, daily returns for the Nasdaq 100 stocks and cryptocurrencies from 2018-07-01 forward.

Using the python packages Numpy and Pandas, two sets of eigenvalues and eigenvectors were calculated from the correlation matrix of the daily returns of cryptocurrencies and from the daily returns of the Nasdaq 100 stocks. By ordering the eigenvalues and their corresponding eigenvectors from largest eigenvalue to the smallest, we get the three most significant components. Often when performing principal component analysis, a few of the largest principal components stand for nearly all the total variance of the returns. Therefore, for this thesis, the three largest principal components directions for each stock or crypto will be plotted with the magnitude in the y-axis and the asset in the x-axis. For the Nasdaq 100 stocks, the seven major industries of the index is grouped by color for us to see if the principal components directions vary between industries. According to Bouchaud and Potters (2000, 83), the largest eigenvalue with its corresponding eigenvector (PCA) is the market itself since it affects all assets approximately the same, the direction components are almost equal. This means that this component stands for most of the movement in price for this group of assets (crypto or stocks in our case). The second and third largest principal components could be other macroeconomic significant factors such as sector specific movement or changes in rates, but these are not easily derived so we will keep them in the plots with the assumption that the largest principal component corresponds to the market movement and leave the other PCs up for interpretation.

In order to filter out insignificant noise from the correlation matrix, we see which eigenvalues from crypto and Nasdaq 100 stocks fall below and above the threshold, λ^+, defined under chapter 2.6. Crypto and Nasdaq 100 get different λ^+ since their return matrices are of different dimensions (same number of days but different number of assets). This is presented with a "rugplot" using the Seaborn package in python together with a kernel density estimation of the eigenvalues. The eigenvalues are spread out along the x-axis along with the threshold, λ^+, dividing the noise from the interpretable eigenvalues.

3.6 Validity

The validity of the results can be defined as how well the model's output match reality. That is, if there is a connection between some stocks and the cryptocurrency market and that they indeed move in a similar way over an arbitrary time period. Quantitative models and results are prone to many errors. These errors include overfitting, underfitting, high bias and variance and a low signal to noise ratio (predictability within a system (AQR n.d)). Another challenge to obtaining good results is the lack of market data. The cryptocurrency market data can also be unreliable as mentioned in the background. By using well known and peer reviewed theory and models the results does arguably have a great degree of validity. References and comparisons to previous literature in the field does also provide legitimacy to the results. That being said, financial data is very noisy which is true both for cryptocurrencies and traditional assets, and this noisiness is shown clearly in the principal component analysis. The data used for this study is otherwise regarded as of great quality and it is this author's belief that short term, i.e., hourly changes in prices will not affect the model's performance and result much. In the Style-analysis, the returns are weekly which further mitigate this three-hour gap.

The portfolio results are generally harder to validate. Portfolio theory and methods have almost exclusively been written for stocks and bonds backed by underlying assets and not for cryptocurrencies which value is based upon much more unclear factors. Most people vaguely knowledgeable about the financial markets know that historical return for every type of asset does not guarantee the same performance in the future and the lack of historical data for cryptocurrencies only magnify this issue. This is why the style analysis and principal component analysis complement the portfolio analysis to shed light on how the cryptocurrencies behave and what industry they might be most similar to or if they can be used as a diversification tool.

4 Results

Under this section the results from the main research are presented.

4.1 Portfolio Results

The three metrics used in this thesis to evaluate the portfolios are: the Sharpe ratio, effective number of constituents and maximum drawdown. These are presented in Table 2 below. The row labels include the portfolio strategy and the asset combination (only cryptocurrency, only Nasdaq 100-stocks or a combination of the two).

Table 2: Portfolio metrics for the eight portfolios

	Sharpe Ratio	Average ENC	Maximum Drawdown
Equal Initial Weights - Only Crypto	0.624	11.364	-77.5%
Market Cap Weights - Only Crypto	0.781	2.169	-67.6%
Minimum Variance - Only Crypto	1.062	2.263	-66.9%
Risk-Parity - Only Crypto	0.625	19.212	-75.5%
Minimum Variance - Only Nasdaq 100	1.135	79.969	-28.9%
Risk-Parity - Only Nasdaq 100	0.799	83.925	-29.3%
Minimum Variance - Nasdaq 100 and Crypto	1.502	8.637	-30.4%
Risk-Parity - Nasdaq 100 and Crypto 100	0.974	93.780	-29.8%

Below in figures 3a-3h, are the eight portfolios plotted together with the Nasdaq 100-index (the ticker for the index is ˆNDX) as a single price, each with \$100 as the initial investment.

Figure 3a: Initial equal weights - only cryptocurrency

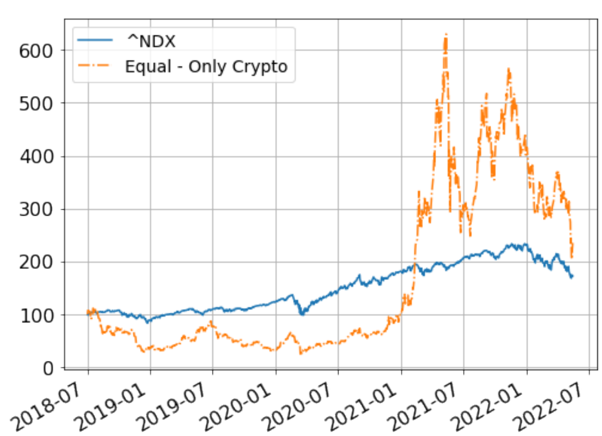


Figure 3b: Market capitalization weights - only cryptocurrency

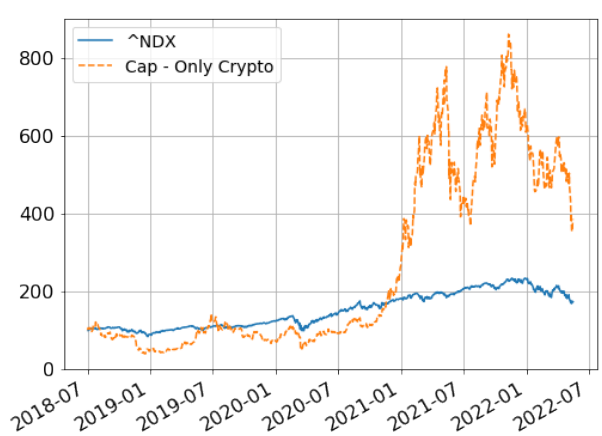


Figure 3c: Minimum variance - only cryptocurrency

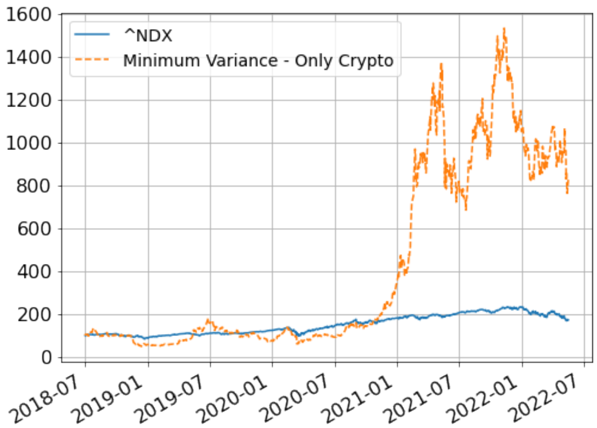


Figure 3d: Risk-parity - only cryptocurrency

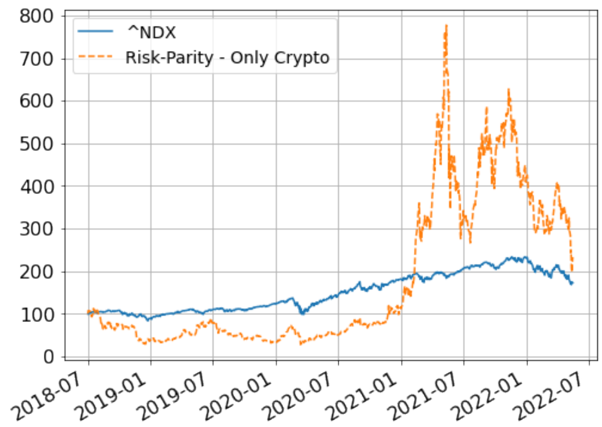


Figure 3e: Minimum variance - only Nasdaq 100

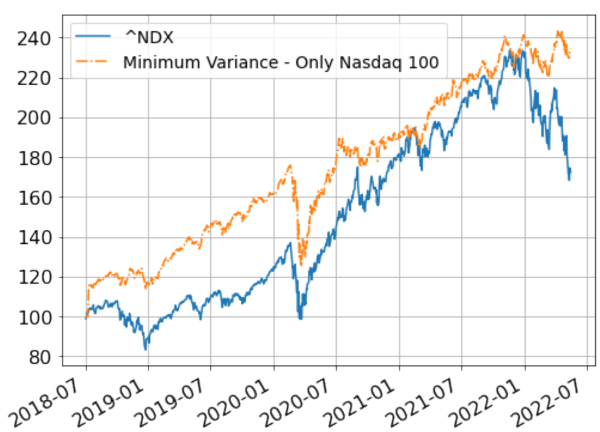


Figure 3f: Risk-parity - only Nasdaq 100

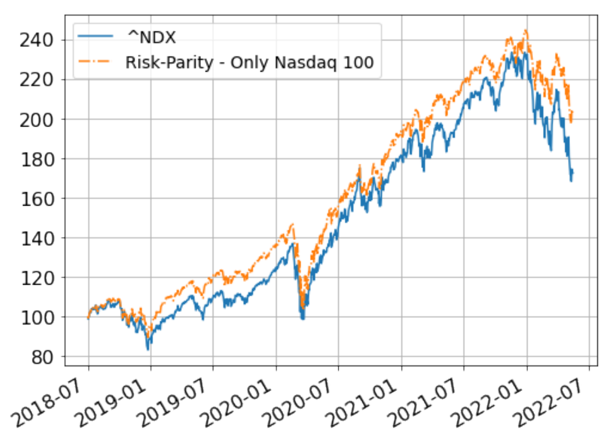


Figure 3g: Minimum variance - cryptocurrency and Nasdaq 100

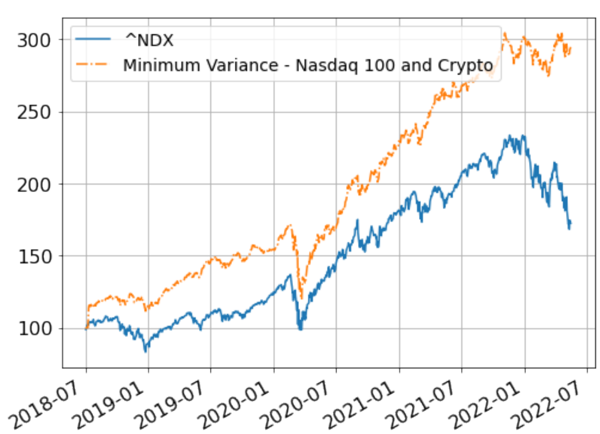
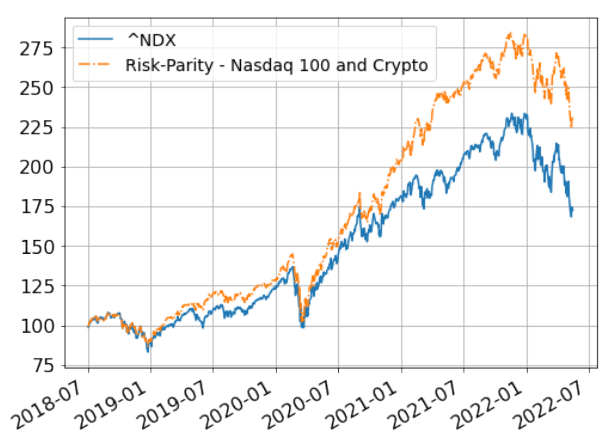


Figure 3h: Risk-parity - cryptocurrency and Nasdaq 100



The area plots in figure 4 and 5 show us the allocations of portfolio value for the portfolios that invested in both stocks and cryptocurrencies. On average, cryptocurrencies made up 5.3% of the portfolio for minimum variance and 6.4% for risk-parity. Bitcoin made up on average 17.4% and 7.5% of the total cryptocurrency investment in the aforementioned portfolios respectively.

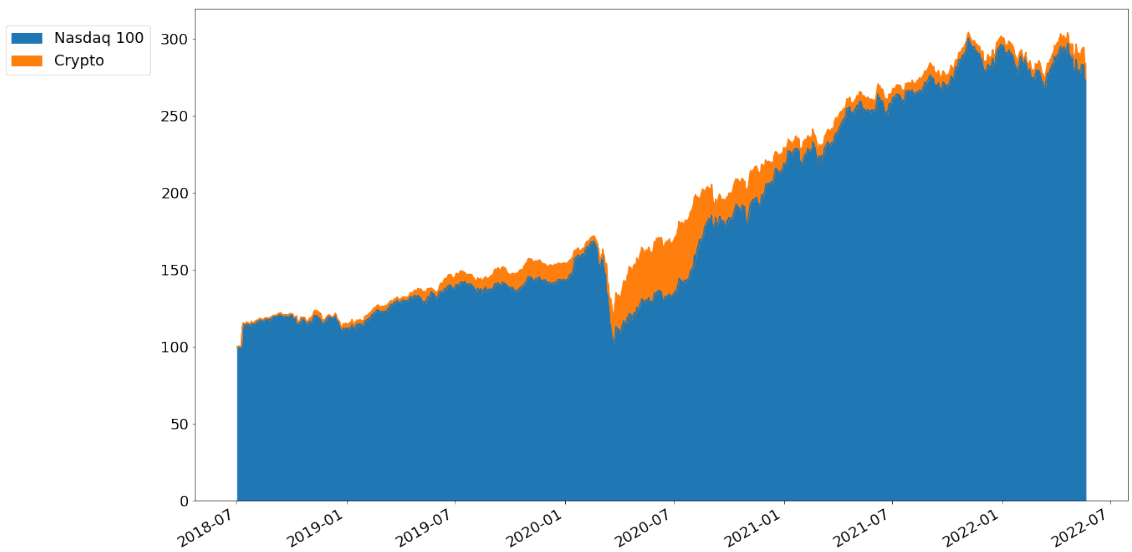


Figure 4: Allocations of crypto and stocks for the minimum variance portfolio with both cryptocurrency and Nasdaq 100

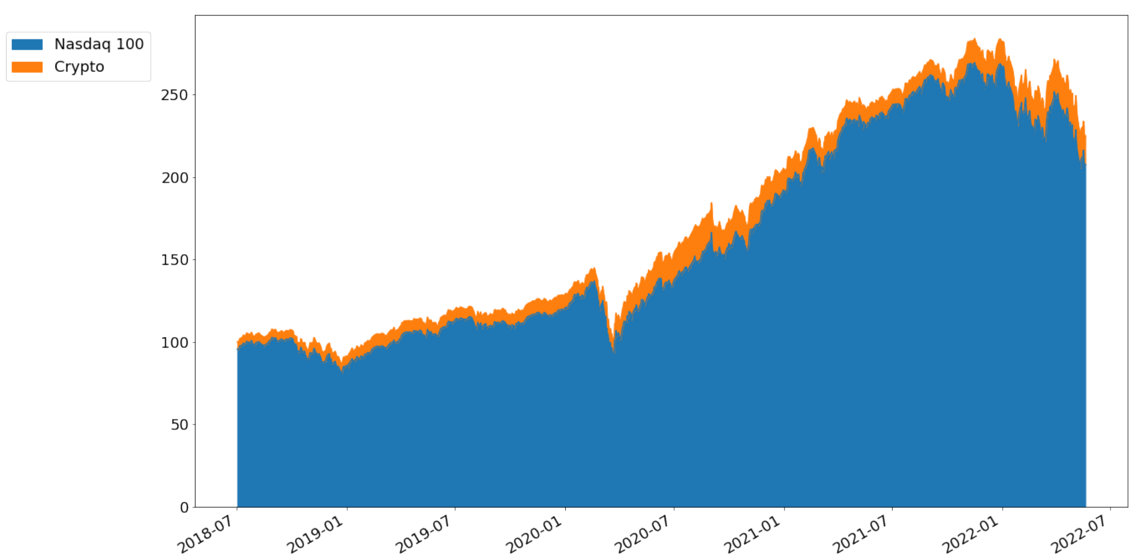


Figure 5: Allocations of crypto and stocks for the risk-parity portfolio with both cryptocurrency and Nasdaq 100

4.2 Style-analysis Results

The Style-analysis was entirely performed with linear regression and lasso regression, both with and without shorting. That leaves us with four models. Only the lasso regressions and the linear regression with no shorting are presented here but the R-squared test includes all four models over the entire time span (2018-07-02 to 2022-05-15). Keep in mind that the lasso regression with shorting allowed becomes almost identical to the lasso regression with no shorting after 2021 and falls under its line in the R-squared plot.

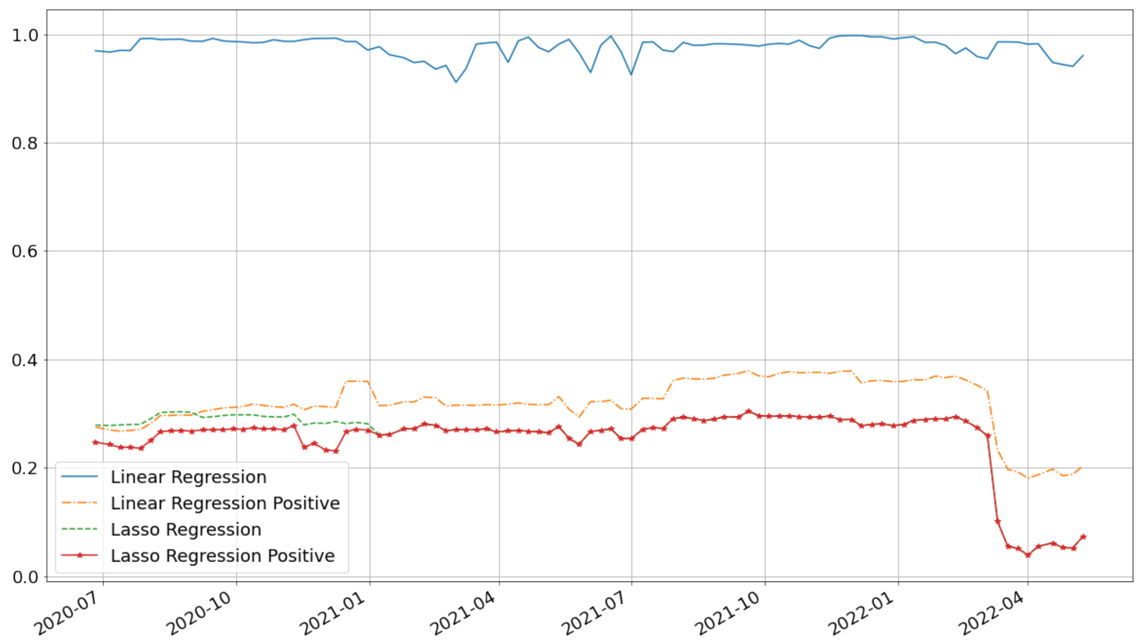


Figure 6: R-squared test

The Style allocations are presented in the area plots below. The stock tickers in the legend are ordered from largest to smallest area. Stocks with no allocations and with very tiny areas are omitted from the legend.

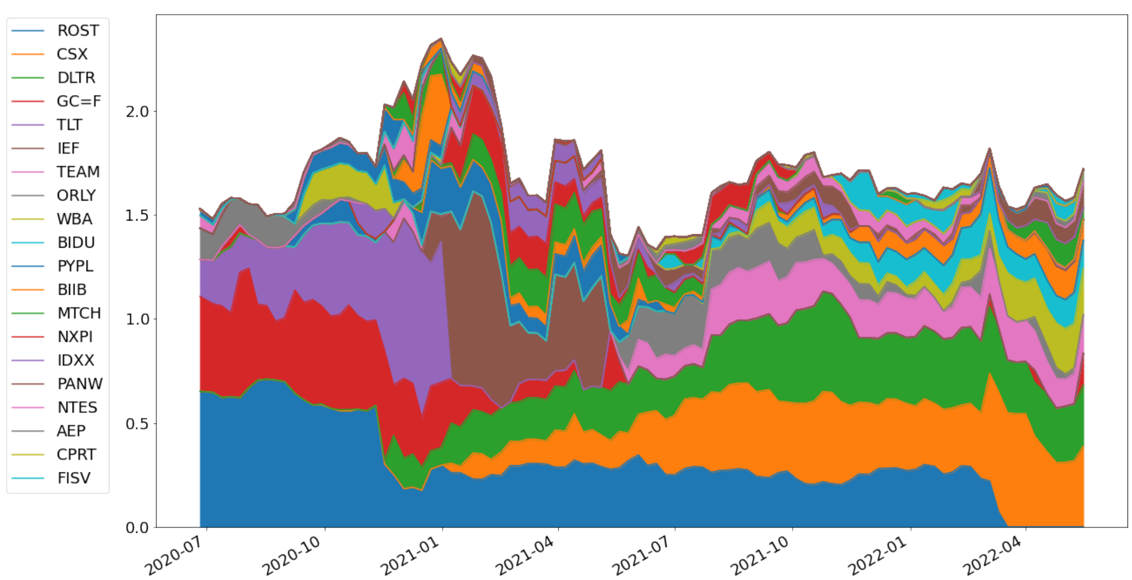


Figure 7: Linear regression allocations over time, no shorting

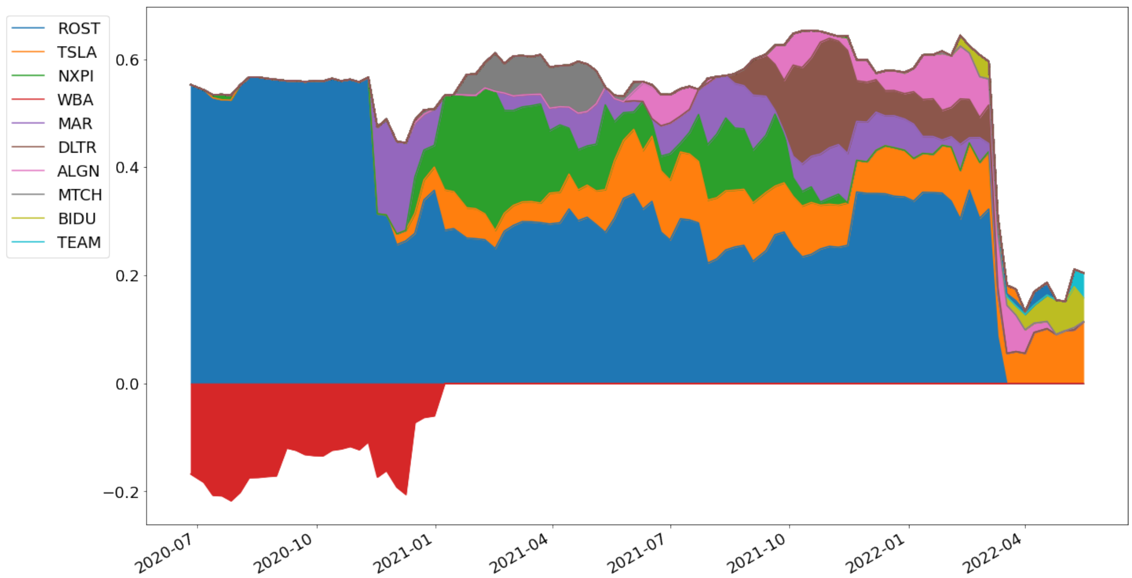


Figure 8: Lasso regression allocations

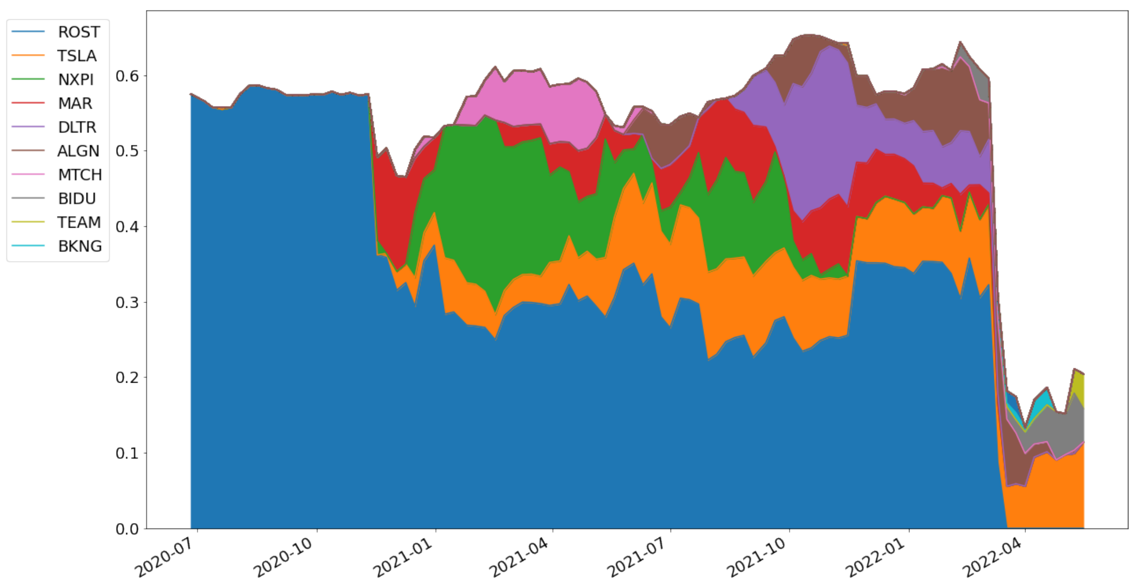


Figure 9: Lasso regression allocations, no shorting

The companies with the largest areas are ROST (Ross stores Inc, discount department stores), TSLA (Tesla, car and energy corporation), WBA (Walgreens Boots Alliance, pharmaceuticals), MAR (Marriot International, hotels), NXPI (NXP Semiconductors, semiconductors), CSX (CSX Transportation, railroads), and DLTR (Dollar Tree, discount variety stores). GC=F is the price of gold.

4.3 PCA Results

Principal components retrieved from the cryptocurrencies' and Nasdaq 100 stocks' correlation matrices, are plotted with the assets in the x-axis. The stocks are grouped by industry for Nasdaq 100.

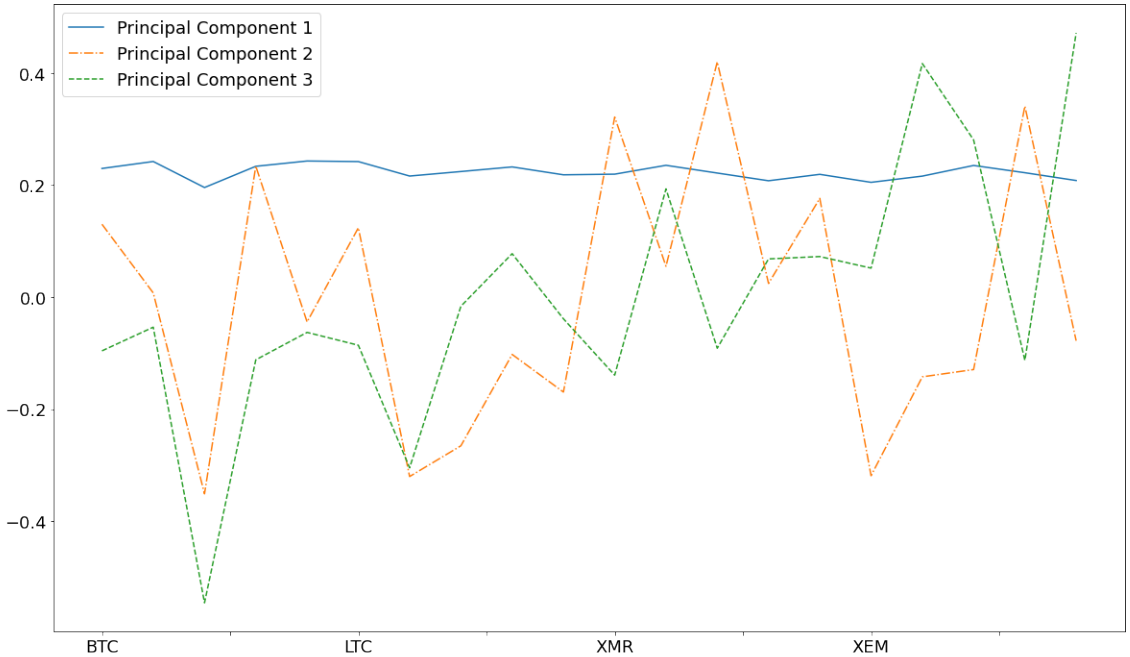


Figure 10: Principal components for the cryptocurrencies

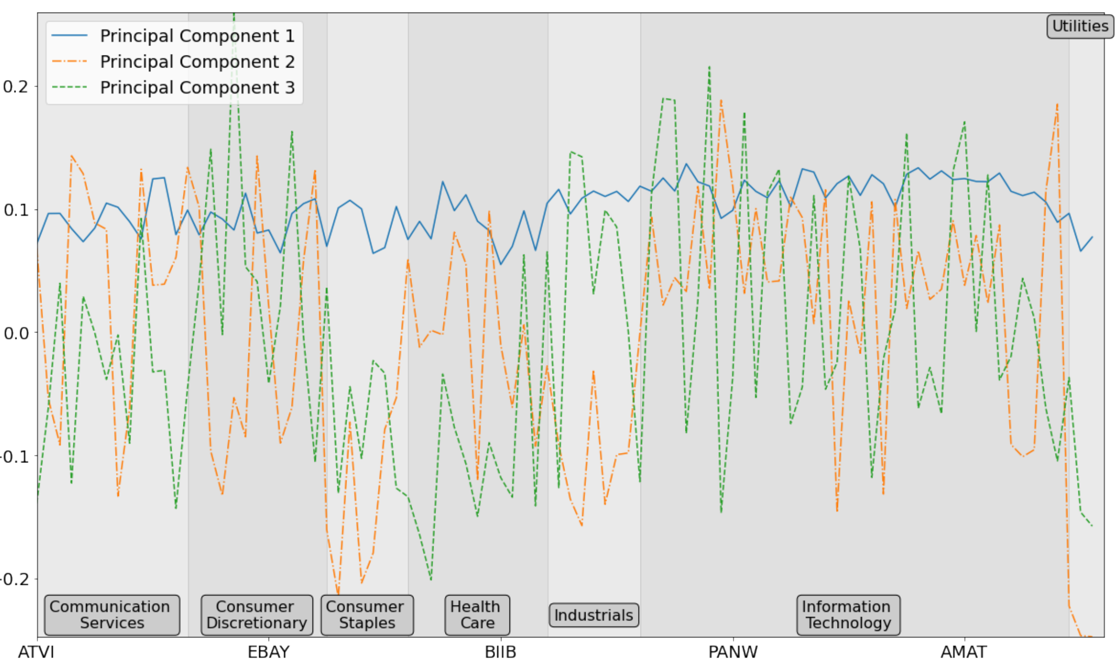


Figure 11: Principal components for the stocks grouped by industry sector

The rug plots below contain the eigenvalues scattered along the x-axis with the noise-threshold, λ^+, represented by the larger red line. The density of the eigenvalues is shown by the filled curve.

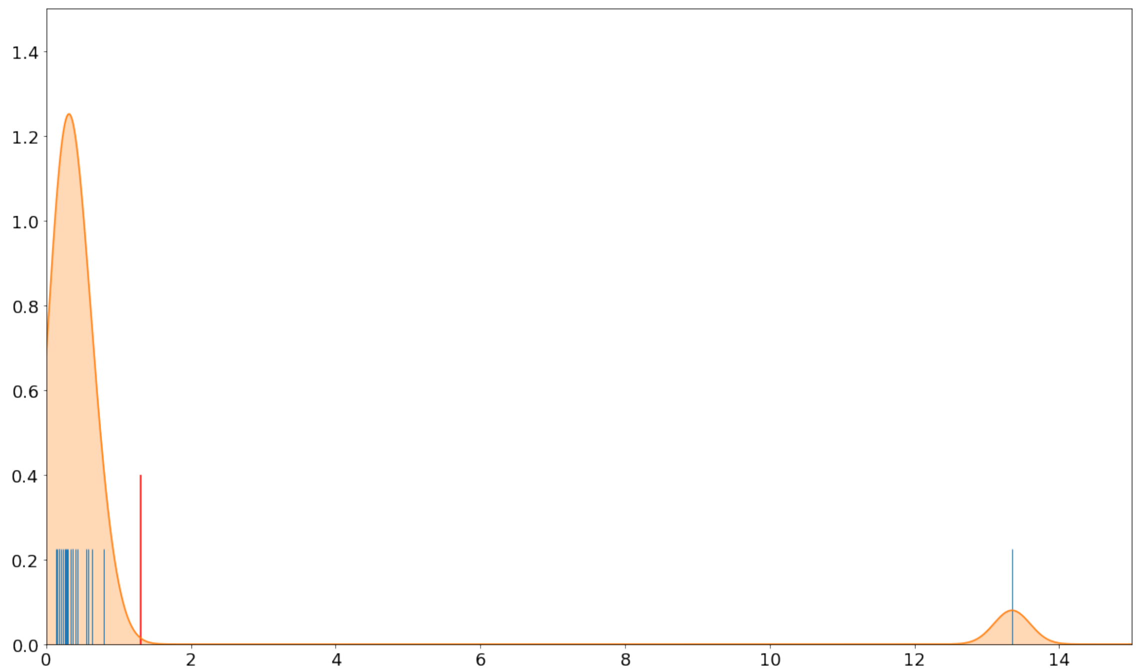


Figure 12: Rug plot for the cryptocurrencies

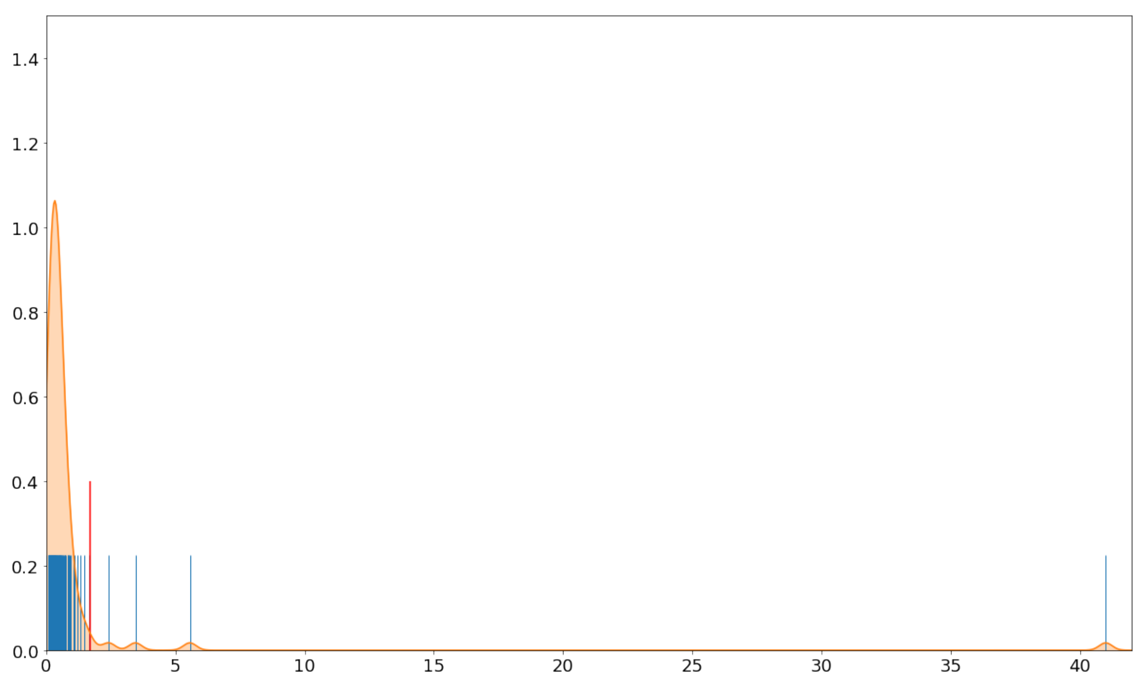


Figure 13: Rug plot for the Nasdaq 100 stocks

5 Discussion

By looking at the results from the different portfolios we see that all portfolios invested entirely in cryptocurrency (Figure 3a, 3b, 3c and 3d) exhibit similar behaviour and their trajectories are very similar to each other. The goal of taking the 20 largest cryptocurrencies in 2018 was to create a fair index of the market but it seems that rebalanced themselves performed way better than simply buying and holding. What happened here is that many of the currencies that had a large market cap in 2018 did not survive the competition and the few largest at the time turned out to be the best investment. This may be why the minimum variance and market capitalization weighted portfolios performed so well: they both invested heavily in Bitcoin which just kept on growing while the equal weight portfolio, which never updates its allocations, was deeply invested in coins like EOS which have since plummeted in price. This is show by the ENC which sits at just over 2 for those two portfolios. The greater diversification of the risk-parity and equal weights portfolios as shown by their ENC, did not at all benefit their return and their subsequent Sharpe ratio is lower than the minimum variance and market cap weighted portfolios despite them being not nearly as diversified in different cryptocurrencies. Minimum variance even has an annual Sharpe ratio above one for both of its iterations, which is considered good.

But that was just portfolios invested entirely in cryptocurrencies, what about stocks? There is not really any use having a portfolio with equal weight or one that updates its weights by market capitalization for the Nasdaq 100 stocks, one can just invest in the index. By looking at the plots of risk-parity and minimum variance portfolios (Figure 3e and 3f) we see that risk-parity follows the index quite closely, albeit performing slightly better while minimum variance clearly outperforms the standard Nasdaq 100 index (although it seems like it got "lucky" early on by investing heavily in a stock that grew a lot in a single day).

The main purpose of the portfolio analysis was to evaluate if there is a diversification benefit from investing both in stocks and cryptocurrency. The minimum variance strategy and risk-parity strategy was tried for both cryptocurrency and Nasdaq 100 stocks and we see in Figure 3g and Figure 3h that both strategies' movements are similar to the Nasdaq 100 index but benefit greatly from the 2021 crypto-bull run even. While the risk-parity strategy seems to follow the steep drop in price of cryptocurrencies in March-May, 2022, the minimum variance strategy seems better equipped to deal with those losses and has not dropped as much in 2022 as neither the Nasdaq 100 index or the cryptocurrency market as a whole. Indeed, this portfolio has a high Sharpe ratio of 1.502 indicating that it reaps both great returns while minimizing the risk. However, this comes at a cost: its ENC sits at a just under 9 which is not a lot considering there are around 110 assets to choose from and a sharp drop in price in one of its allocations may cause severe losses. Minimum variance is entirely dependent on historical variances which says nothing about the future movements. Regarding maximum drawdown, the loss in percentage is very similar for the portfolios invested only in cryptocurrency. This is also true for the two portfolios invested only in stocks and the two invested in both stocks and crypto. What is notable here is that the Sharpe ratio for the portfolios with both Nasdaq 100 stocks and cryptocurrencies are better than that of the portfolios with the same strategy but with only Nasdaq 100 stocks, despite their maximum drawdowns being very similar. It seems that for investors looking for a great Sharpe ratio, cryptocurrencies are useful. However, the similar maximum drawdowns may stem from the market crash in 2020 which neither minimum variance or risk-parity proved to be much better at enduring than the other.

Recall from section 3.4 that the point of the Style-analysis was to evaluate if the cryptocurrency market can be replicated by a portfolio of stocks from the Nasdaq 100. The challenge arising from this was how this linear relationship could be formed in a presentable way, that is without the coefficients exploding and the model becoming too hard to comprehend and deduct conclusions from. This is what happened with the Style-analysis when regular linear regression was used. Although its R-squared score was very good as seen in Figure 6, for pretty much the entire time period, the weights went far up in both the positive and negative direction. For example, for some 100 weeks the model wants to short one class of Google stock and invest in the other, both by a lot. Then for the next 100 weeks (remember the time window overlaps so the returns are 99% the same as the previous time frame for every time frame) the model does the exact opposite and invests heavily in one and short the other. This is a highly suspect result which clearly indicates overfitting and regular regression is probably not up to the task at hand.

This leaves us with three Style-analyses which are more interpretable as we can see in Figure 7-9. Fortunately for us, lasso regression is used to prevent weights from exploding and it seems that a strong linear relationship was formed with ROST along with some other stocks in the consumer discretionary industry sector, namely TSLA, DLTR, and MAR. Linear regression without shorting seemed to also draw this conclusion, that cryptocurrency follows those stocks strongly, along with some smaller weights for a number of other stocks and bonds. Notable for that model was that gold (GC=F) received strong weights in the first half of 2020 since crypto (specifically Bitcoin) has been dubbed "digital gold" by some.

Unfortunately, the R-squared score for Style-analysis with both positive and unconstrained lasso regression, and constrained linear regression, was very low for the entire time period. It seems that no model is both comprehensive and have a decent goodness-of-fit in our case. Without a great R-squared score it is very difficult to say if the cryptocurrency market is sensitive to the assets chosen for this thesis. This weak relationship becomes even more clear in 2022 when crypto prices started to drop and the lasso regression struggled to find weights for stocks that followed this drop (linear regression without shorting performed a little better here), seen both in the area plots and the R-squared plot.

The main goal of the principal component analysis was to filter analyze how the returns from crypto and stocks independently, move when the noise filtered out. The principal component with the largest eigenvalue can be viewed as the market which direction sits around 0.2 and 0.1 for crypto and stocks respectively (Figure 10 and 11) and explains the largest variation in price for our assets. By looking at the rug plots in Figure 12 and 13, we see that most of the principal components are concidered as noise and leaves us only with one significant eigenvalue and principal component for the cryptocurrencies and four for the Nasdaq 100 stocks. This tells us that the cryptocurrencies can all be viewed to be a part of the same industry while the Nasdaq 100 stocks make up at least seven distinct industries. This is why the second and third principal components could stand for industry specific movements. There seem to be a *slight* difference in the principal components values between the sectors as seen in the PCA plot for Nasdaq 100 in Figure 11, especially in the consumer staples sector; this movement is not very clear and more research is needed to decode the secondary principal components.

6 Conclusion

First and foremost, it is safe to say that the cryptocurrencies we have studied move pretty much in the same direction most of the time, with different magnitudes. However, some cryptocurrencies have seen massive returns while others have crashed to a price close to zero and it is possible that our methods have failed to capture, or at least make sense of, why some currencies drop so fast and suddenly. Then again, the crashes of individual currencies have not been the focus of this thesis. The little variation there is in movement directions are mostly indescribable noise or perhaps very specific events related to individual coins only. We have seen with the PCA that there is only really one significant factor that stands for most of the movement. Cryptocurrencies' move almost uniformly and it is probably futile to try and obtain a diversified portfolio with only cryptocurrencies. If one wish to diversify a portfolio of stocks, it seems that the largest cryptos with the least variance is the best option. In fact, the minimum variance portfolio strategy always invests most heavily in the least volatile cryptocurrency, Bitcoin, both for the only cryptocurrency portfolio and the portfolio where crypto and stocks are combined. Every cryptocurrency used in this thesis (Bitcoin included) are extremely volatile compared to stocks and cryptocurrencies on average only made up a relatively small part of the total portfolio value in the portfolios that invested in both cryptocurrencies and Nasdaq 100 (see figure 4 and 5). Still, it seems as if they did improve the Sharpe ratio for the minimum variance and risk-parity portfolios and provides some diversification. In figure 4 we see that cryptocurrencies make up as much as 20% of the portfolio during the market crash of 2020 since many cryptocurrencies were less volatile than many stocks. However, the massive hype around crypto and the currencies' huge ups and downs bring the returns to some absurd levels which we may not see again in the future.

Although the cryptocurrencies did grow explosively in 2021 around the same time the stock market became extremely bullish on its own, the Style-analysis showed that stocks and cryptocurrencies does not move dependently on each other. They neither move convincingly in the same or the opposite direction, at least for the weekly returns. It is true that financial data, and data in general that is dependent on human actions and choices, is very noisy and creating models with a good bias-variance tradeoff is very difficult. We theorized that cryptocurrencies would perhaps move similarly to information technology, a sector that also saw rapid growth in 2021 but the Style-analysis did not show any clear signs of that. Cryptocurrencies perhaps move with consumer discretionary stocks (Tesla also grew a lot in 2021) as shown with the Style analysis (Figure 7-9) but the R-squared score was so low for most of the time that this is not certain.

One of the main takeaways of this thesis is therefore that the analyses should be redone in some years and see how the results hold up after the crypto market has experienced a bear market to see how the Sharpe ratios hold up and if they coins are still as correlated in price movement. Also, since a few of the 20 cryptocurrencies we studied have lost most of their market value since 2018, future work should perhaps exclude coins that drop below some arbitrary market capitalization since small-cap coins may have misleading market data (for example too high or low volatility). Other stocks and assets than the ones we used could perhaps be better choices for our analyses, especially for the Style-analysis.

6.1 About the Future of Crypto

As stated before, the cryptocurrency market is simply too young for any detailed market analysis and only very general conclusions can be drawn from the results of this thesis. This issue will persist unless the market displays some "maturity" at some point or if the currencies reach a terminal value. One of the great challenges remaining for crypto-analysts is finding a reasonable pricing model, but that may be impossible considering most of the cryptocurrencies have, at least at the moment, no underlying assets with intrinsic values. That is not to say that the cryptocurrencies are without intrinsic value themselves, peer-to-peer digital payment systems are clearly in demand and have use cases; they are just extremely hard to price fairly. The blockchain technology itself may have a very real intrinsic value and has seen some use cases outside the cryptocurrency sector. Peer-to-peer digital payment systems *does* have value as a technology; the issue arises when trying to price them since they must have a market value to be useful. If they were useless then they would not have a market value.

It also does not help the case of market analysts that the cryptocurrency market very much seem like it is in a price bubble and is constantly experiencing crashes, scandals, price manipulations, and investors being scammed by shady start up-coins promising huge returns. Again, the crypto market may be in a bubble and as we have seen, many of even the largest coins does not survive even for a few years. As of May 2022, the cryptocurrency market is experiencing a big decline, but the most popular coins like Bitcoin and Ethereum will likely not lose all of their value. In this authors opinion, the most likely scenario which we have already seen play out in this analysis, is that a few very large cryptocurrency projects will survive, and most if not all of the 10 000+ coins there is will fade in to obscurity. Still, only time will tell and the next few years will no doubt be very interesting.

References

1. Ante, Lennart; Von Meduna, Marc and Steinmetz, Fred. 2021. Ownership, uses and perceptions of cryptocurrency: Results from a population survey. *Technological Forecasting Social Change* 173: 1-19.

2. Aroussi, Ran. 2022. Download market data from Yahoo! Finance's API (Version 0.1.70). *GitHub*. `https://github.com/ranaroussi/yfinance`.

3. AQR. N.d. *Machine Learning: Why Finance Is Different*. `https://www.aqr.com/Learning-Center/Machine-Learning/Machine-Learning-Why-Finance-Is-Different` (Retrieved 2022-05-16).

4. Best, Raynor de. 2022. Number of cryptocurrencies worldwide from 2013 to February 2022. `https://www.statista.com/statistics/863917/number-crypto-coins-tokens/` (Retrieved 2022-02-24).

5. Bouchaud, Jean-Philippe and Potters, Marc. 2000. *Theory of financial risks: from statistical physics to risk management*. Cambridge University Press.

6. Bouri, Elie; Saeed, Tareq and Vinh Vo, Xuan. 2021. Return equicorrelation in the cryptocurrency market: Analysis and determinants. *Finance Research Letters* 38: 1-9.

7. Carli, Tiffanie; Deguest, Romain and Martellini, Lionel. 2014. Improved Risk Reporting with Factor-Based Diversification Measures. *EDHEC-Risk Institute*.

8. Coingecko. 2022. `https://www.coingecko.com/` (Retrieved 2022-02-03).

9. Coinmarketcap. 2022. `https://coinmarketcap.com/` (Retrieved 2022-02-03).

10. DeCarlo, Lawrence T. 1997. On the Meaning and Use of Kurtosis. *Psychological Methods* 2(3): 292-307.

11. EDHECinfra Docs. N.d. *Effective number of constituents*. `https://docs.edhecinfra.com/display/IM/Effective+number+of+constituents` (Retrieved 2022-05-01).

12. Emmanouil, Christoforou. 2018. CoinGecko API wrapper (Version 2.2). *GitHub*. `https://github.com/man-c/pycoingecko#readme`

13. Fitzpatrick, Richard. 2011. *Matrix Eigenvalue Theory*. `https://farside.ph.utexas.edu/teaching/336k/Newton/node66.html` (Retrieved 2022-05-07).

14. Friedman, Jerome; Hastie, Trevor and Tibshirani, Rob. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*: 33(1).

15. Hastie, Trevor; James, Gareth; Tibshirani, Robert and Witten, Daniela. 2021. *An Introduction to Statistical Learning*. Springer Science Business Media: 103.

16. Hoepman, Jan-Henk. 2007. Distributed Double Spending Prevention. *Security Protocols* 2007: 152-165.

17. Klein, Tony; Pham Thu, Hien and Walther, Thomas. 2018. Bitcoin is not the New Gold – A Comparison of Volatility, Correlation, and Portfolio Performance. *International Review of Financial Analysis* 59: 105-116.

18. LAPACK. 2022. *LAPACK: Linear Algebra PACKage*. `http://www.netlib.org/lapack/explore-html/df/d9a/group__complex16_h_eeigen_ga9b3e110476166e66f2f62fa1fba6344a.html` (Retrieved 2022-05-07).

19. Li, Yuxi; Lv, Zhihan and Qiao, Liang. 2021. An Optimized Byzantine Fault Tolerance Algorithm for Consortium Blockchain. *Peer-to-Peer Networking and Applications* 14:2827-2839.

20. Liu, Baixiao; McConnel, John J. and Wang, Jingfang. 2021. The ICO Rating Game: Biased Ratings of Crypto Experts in Initial Coin Offerings. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3776935`.

21. Minitab Blog Editor. 2013. *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?* `https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit` (Retrieved 2022-04-25).

22. Nadal, Scott and King, Sunny. 2012. PPCoin: Peer-to-Peer Crypto-Currency with Proof-of-Stake.

23. Nakamoto, Satoshi. 2008. Bitcoin: A Peer-to-Peer Electronic Cash System. Bitcoin.org.

24. NumPy. N.d. *numpy.linalg.eigh*. `https://numpy.org/doc/stable/reference/generated/numpy.linalg.eigh.html` (Retrieved 2022-05-16).

25. Peterson, Timothy F. 2021. To the Moon: A History of Bitcoin Price Manipulation. *Journal of Forensic and Investigative Accounting* 13(2): 254-272.

26. Raschka, Sebastian. 2015. *Principal Component Analysis in 3 simple steps*. `https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html#1---eigendecomposition---computing-eigenvectors-and-eigenvalues` (Retrieved 2022-05-07).

27. Reischuk, Rüdiger. 1985. A New Solution for the Byzantine Generals Problem. *Information and Control* 64: 23-42.

28. River Financial. No date. *Bitcoin's UTXO Model*. `https://river.com/learn/bitcoins-utxo-model/` (Retrieved 2022-03-22).

29. Sharpe, William F. 1992. Asset Allocation: Management Style And Performance Measurement. *Journal of Portfolio Management*: 7-19.

30. Wigmore, Ivy. 2017. *Statistical Noise*. `https://www.techtarget.com/whatis/definition/statistical-noise` (Retrieved 2022-05-20).