Project:
Hedging against risk in the cryptocurrency market

MS-E2112 Multivariate Statistical Analysis

Esa Turkulainen
432296

30.03.2018

Motivation and research questions

The recent rapid growth in the popularity of the cryptocurrency market among investors has introduced many kinds of new players to the scene. The first movers on the scene were the technology enthusiasts, but soon after came the pundits and industrial investors (especially from Japan). 2017 saw a great deal of over-heating of the markets, increasing its volatility (greatly) for the first time in 3 years. "A bubble" is always deemed after the fact, but a certain colloquial definition is usually thrown around: when your taxi driver starts to chat you about his cryptocurrency investments, we may be at the terminus of increasing valuation. And so, through the first quarter of 2018 the entire market has been on the decline.

Because the cryptocurrency market is still very new and clearly in growing pains, its volatility is intimidatingly large to the majority of investors. But as new coins and investors are still coming to the scene, it might be interesting to see if there are any opportunities for information arbitrage. Particularly, would it be possible to build such a cyrptocurrency portfolio that its assets would balance out some of the risks? Do some coins move together in some direction while others move the other way? If so, does the coin's market cap have an effect?

The data

Not wanting to scrape all the data for the +1,000 cryptocurrencies in existence by hand, I set out to search a proper data set for the analysis. I found data exactly for my needs on Google's open data science platform Kaggle, by the user Jesse Vent (accessible from: https://www.kaggle.com/jessevent/all-crypto-currencies). The data set includes 649,051 observations of 13 variables:

1. Unique name
2. Ticker symbol
3. Given name
4. Date of observation
5. Rank by market capitalization
6. Daily open price
7. Daily high
8. Daily low
9. Daily close price
10. Daily trading volume
11. Close ratio
12. Spread

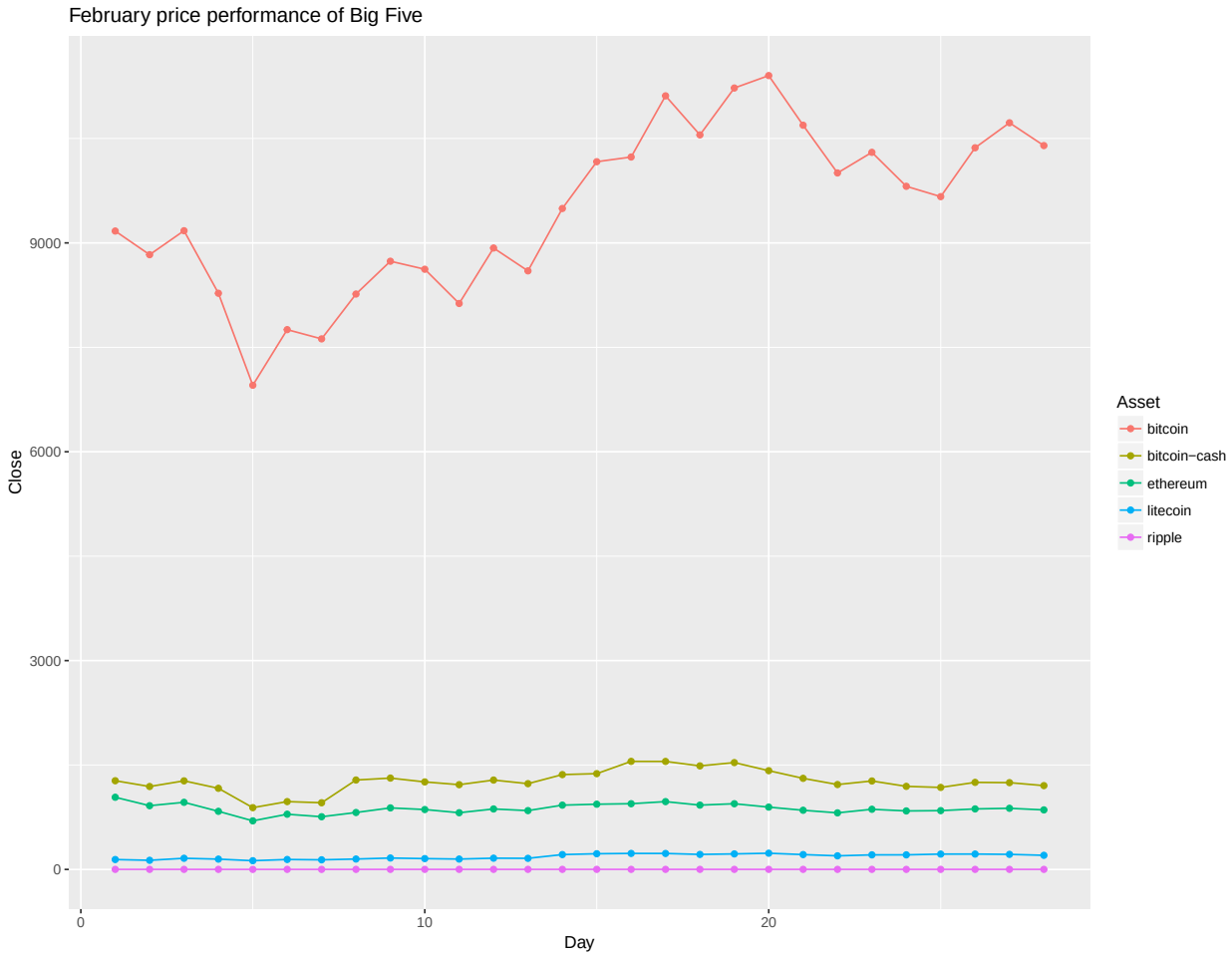
and it contained data for a total of 1,382 cryptocurrencies.

Inspection of missing observations was quick to reveal that many "coins" lacked volume data and that because new coins were constantly being introduced to the market, the chosen time series window needed to be cleaned to ensure that all coins under inspection shared the same number of days.

Exploring the variables

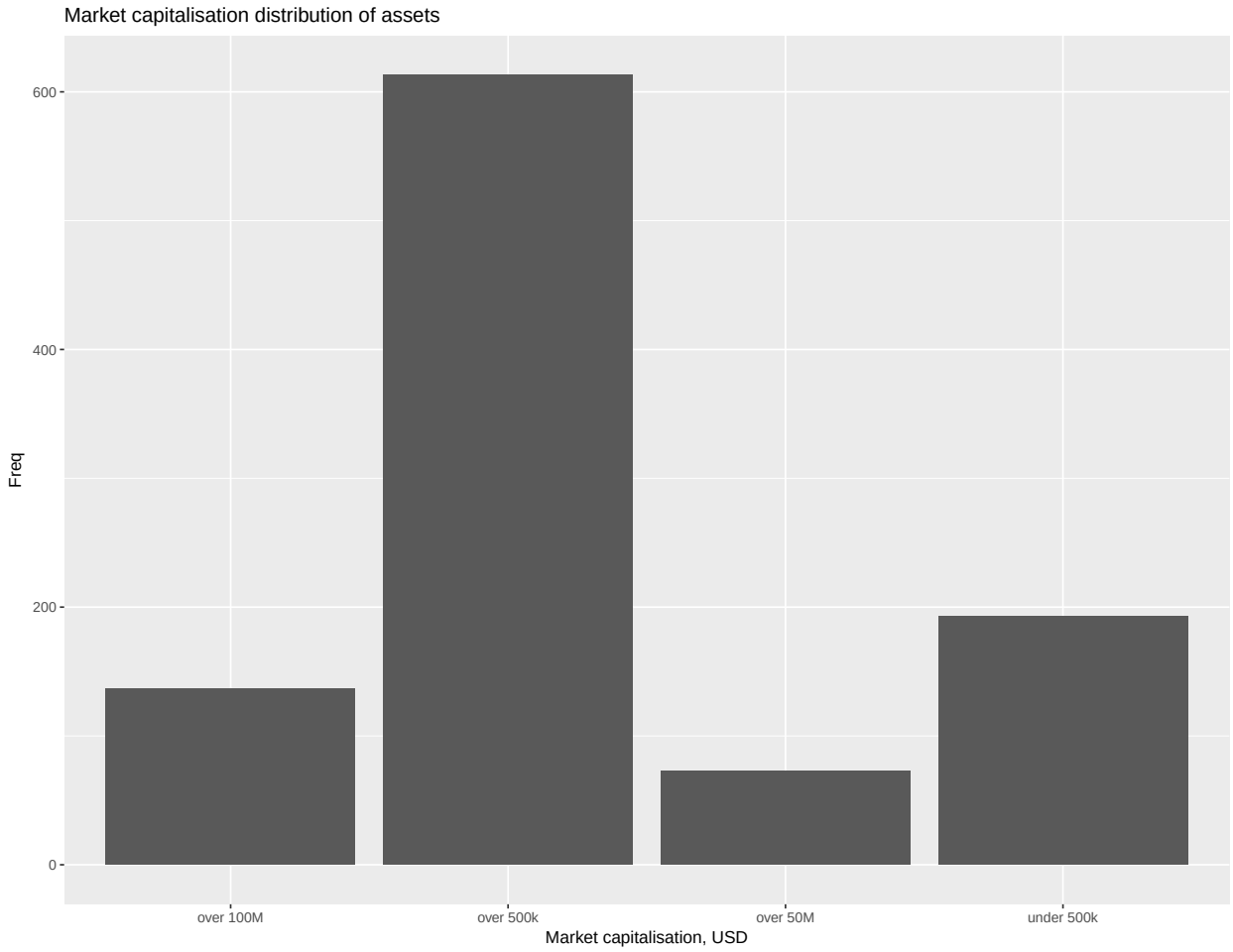
First some simple univariate and bivariate explorations to get to know our data.

A month of price development



The above graph depicts the price development of the 5 largest (by market capitalization) cryptocurrencies throughout February. All further analyses will also be performed on the period of 28 days in February as this time window enabled the inspection of 1016 different intact time series. It is clear from the graph that Bitcoin is dominating the market, and with quite dizzying price movements (for a traditional investor) with a low-high difference of over $1,000. The graph also draws attention to the fact that Ripple manages to remain as one of the 5 biggest with an asset valuation of under 1 dollar, highlighting the very different nature between these two instruments (even though both are deflationary).

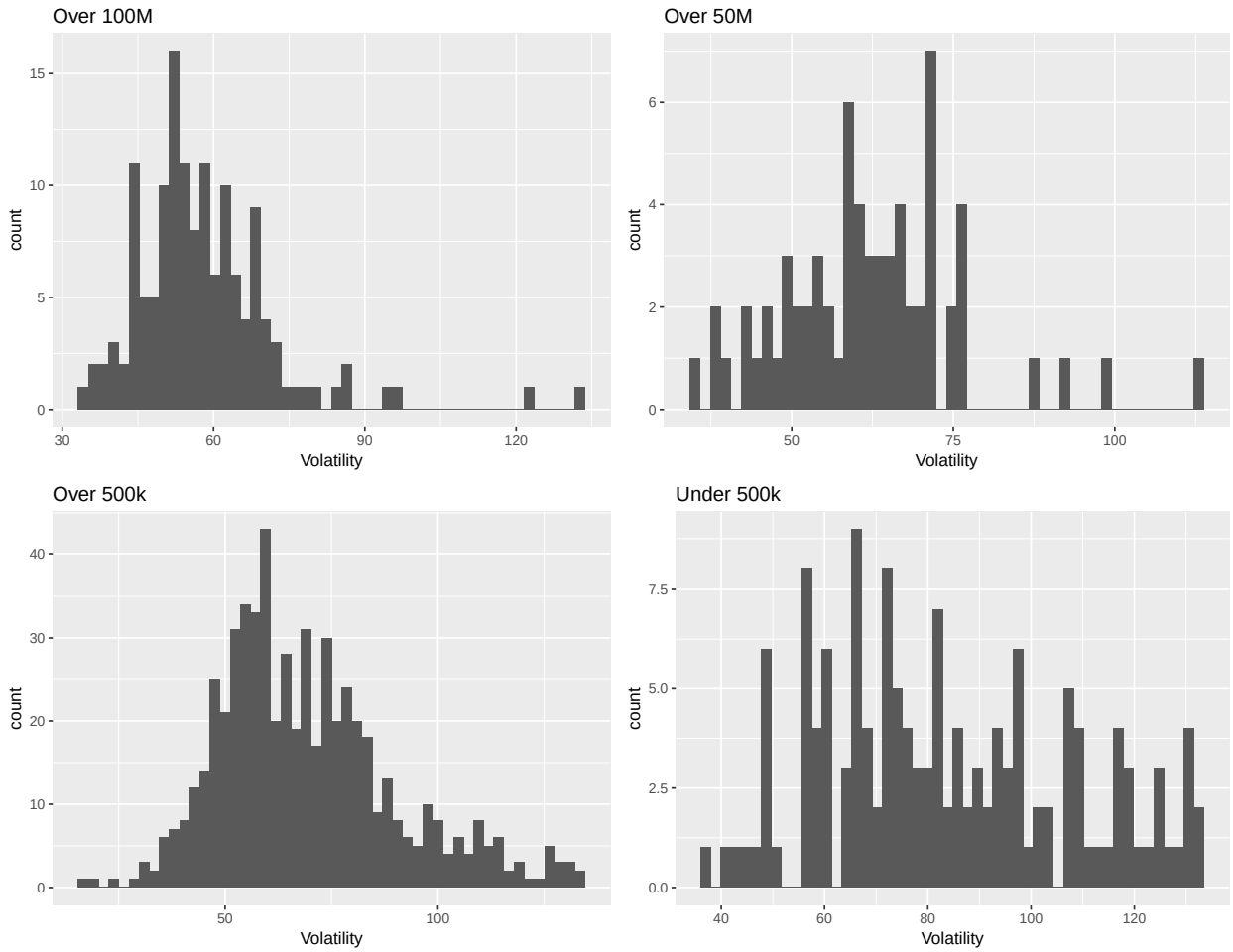
Goliaths and Davids



With over 1300 "coins" now on the market (and 1016 within our time window), one becomes curious of their nature. Are they all created by people who have seen an opportunity in a get-rich-quick-scheme? Are they carbon copies of each other in terms of market behavior? Or are they classifiable into certain categories and do they behave accordingly on the market?

The most straightforward way for checking if asset trends can be separated, without getting too specific, is to check it via asset market capitalization. Under the hypothesis of efficient markets we should see less opportunities for (information) arbitrage and possibly less volatility in markets with lots of capital. Now, the market capitalization in cryptocurrency markets is not voluminous enough to satisfy the assumptions of efficient market hypothesis, but it still might make a difference in daily returns. Above graph presents the distribution of assets into "massive" (over $100M), "large" (over $50M), "mid" (over $500k), and to "small" (under $500k) market capitalization categories. It's evident that there aren't that many massive-tier assets when compared to mid-tier and smaller assets.

Risky business



To see if our assumption about asset volatility in relation to market cap holds true, historical volatility through February of the assets was calculated by taking the standard deviations of daily log returns. Then, to improve analysis robustness, outliers were weeded out and the remaining assets (930) were grouped into 50 bins by their volatility, inside their market cap reference group. The volatility distribution is very non-normal, with closest to normal (given by a Shapiro-Wilk test) being the category "large" (over 50M). Although all categories exhibit a considerable tail towards higher volatilities, it would seem that the two categories with largest market caps contain the safest assets in terms of volatility. Assets with over $100 million market caps have a volatility mean of 58%, and the rest in decreasing order of capitalization: 62%, 69%, 84%. This seems to lend credibility to our assumption above.

Analysis

Method

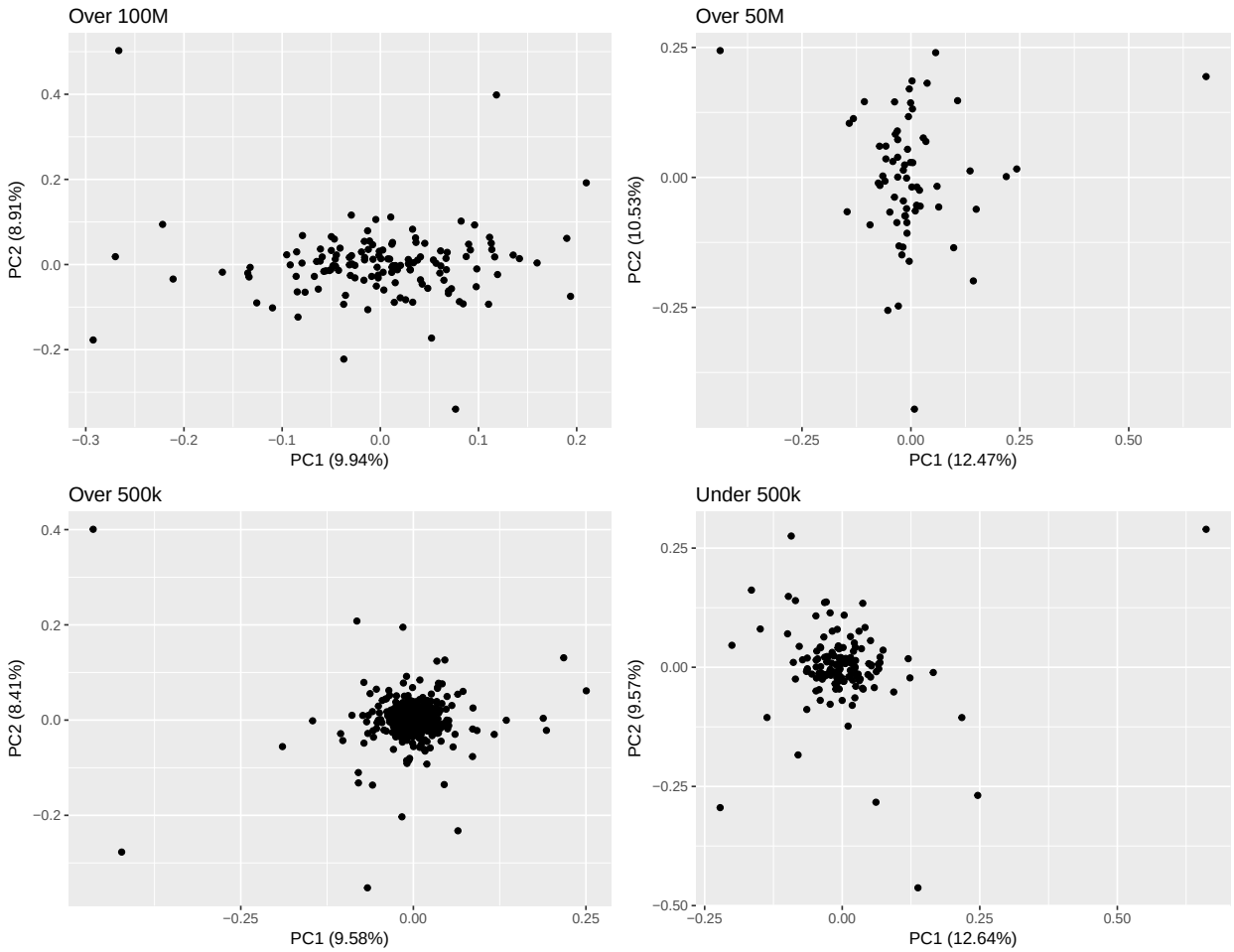
To extend our analysis into asset market behavior, we will explore the question of trend similarity with PCA. The idea behind this approach is that on a day of depressed or excited overall market sentiment the assets will exhibit same kind of behavior, and we will not gain much information by looking at the time series on those days. On the other hand, a day with a lot of variance in the

behavior of the assets could be much more valuable in determining if certain groups of coins tend towards similar movement patterns. This assumes that there is external information in the system that affects assets selectively, such as a bad day in fiat markets affecting the price of cryptopayment tokens but not for example ERC20 multipurpose tokens. In practice we perform the following:

1. Calculate daily log returns for each asset. Each day is thus a variable in our analysis.
2. Run a principal component analysis with centering and scaling of the variables.
3. Try to extract groups either inside categories or between them.

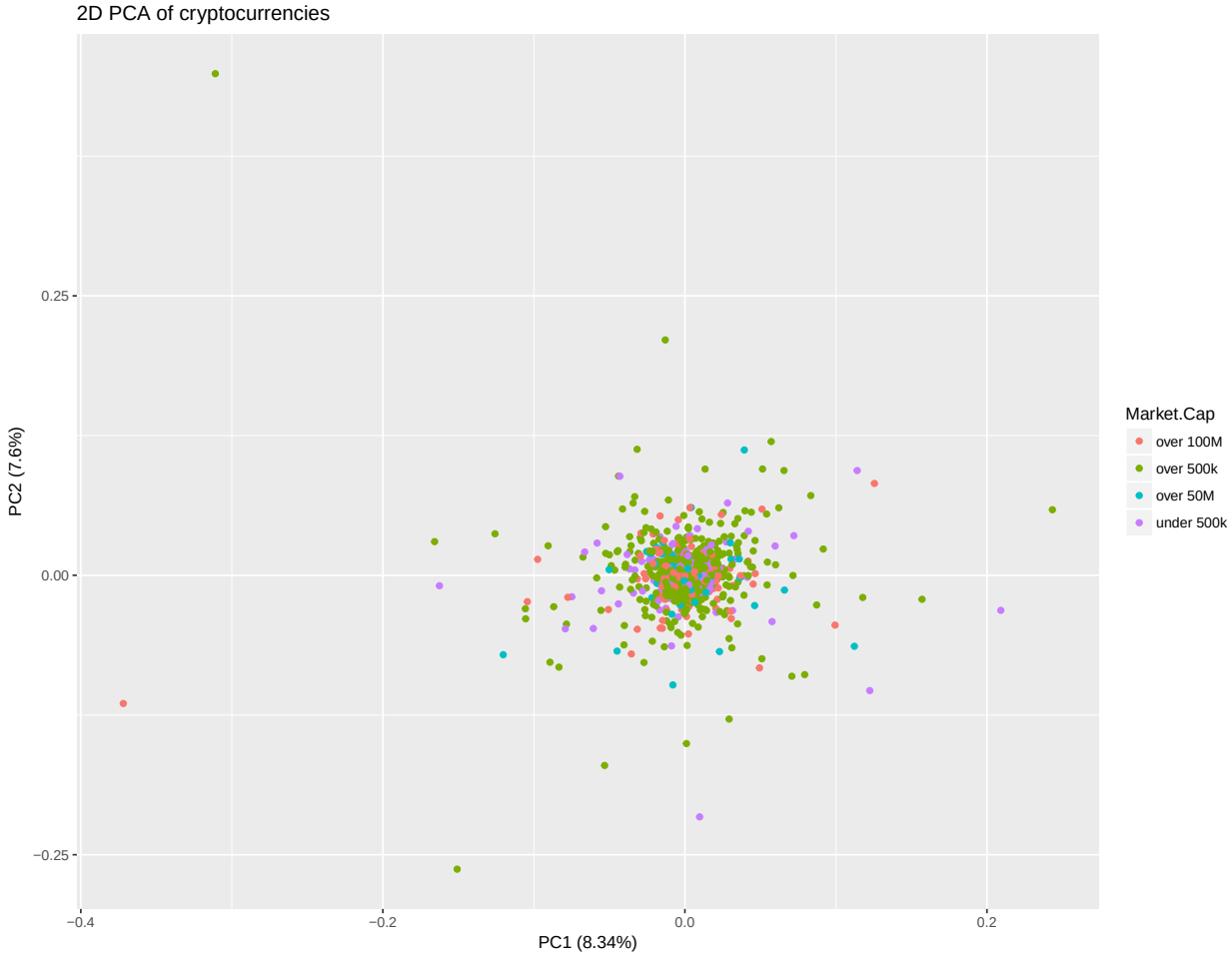
As there are no simple ways of determining the statistical significance of extracted group boundaries, this analysis remains mainly on an exploratory level.

Output



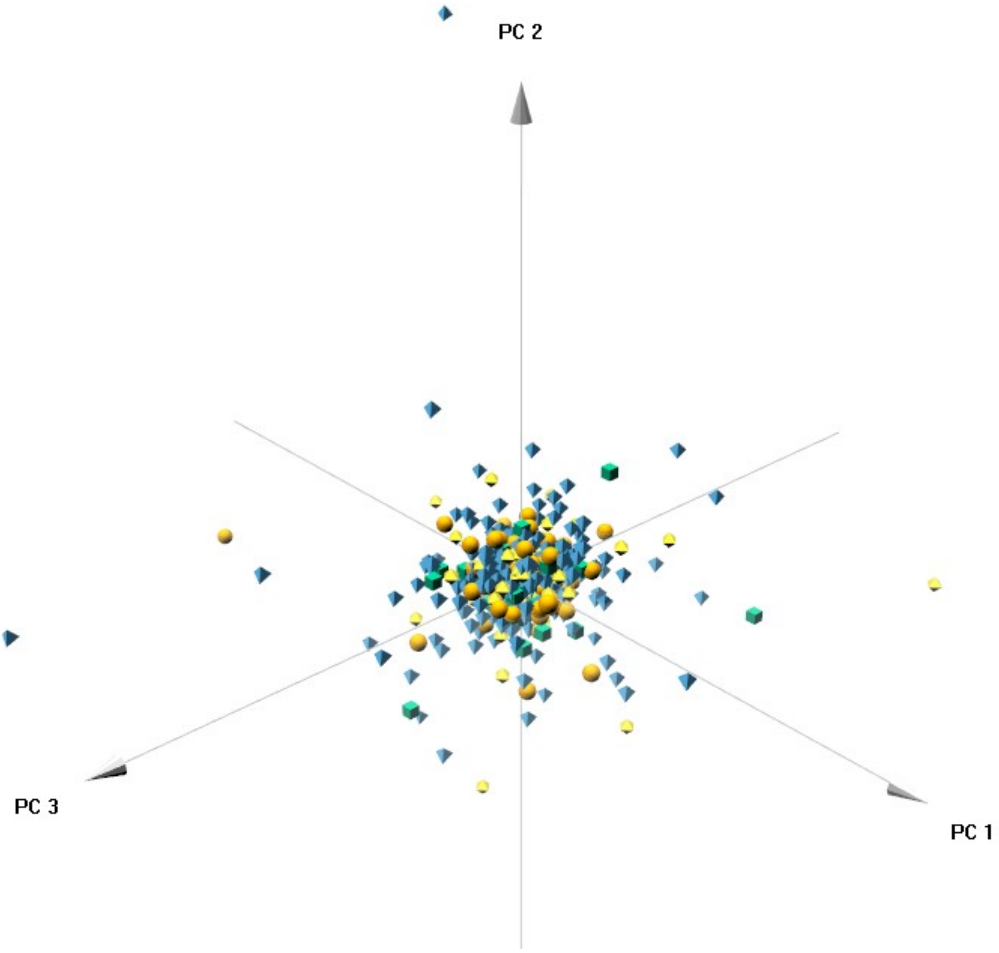
Here is a multiple of 4 different principal component analyses, one for each of the market cap categories. Using only two of the newly extracted principal components we detect few glaring outliers, but no major differentiation. The resulting components for each of the analyses are quite close to each other in their explanatory power as 75% of variance is always explained cumulatively at the 14th component. This indicates that daily log return vectors might not be as easily reduced as one might err to think. It could be, for example, that there were no remarkable differentiating

events or news that could have revealed inter-dependencies between the assets during February. Retaining 75% of information content with a 50% dimension reduction is something, but not much. We would have to use 20 principal components to retain 90% of the original information.



Above is a plot of 2 components from a singular principal component analysis performed with all assets. The colors represent the market capitalization of each coin. Judging by the color, no visible grouping has emerged, which might be expected as the first two components explain only 15.9% of the variance. Again, there are few significant outliers that our group coloring helps to identify as mid to small-tier assets in terms of market capitalization. The lonely massive outlier on the far left is Emercoin and the mid-tier outlier at the top of the graph is ExclusiveCoin. They seem to be nothing special in nature.

Lastly, a 3D-representation with an added third principal component was tried, but it does not seem to reveal any clusters either. Snapshot of the 3D-viewer on the next page.



Conclusions and discussion

The aim of this project was to explore the cryptocurrency market and to see if assets could be clustered into usable portfolio "baskets" by their daily behavior. Succeeding in this would give an edge to technical analysis of markets, as signals from other assets would help manage some other assets without actual external information (news, world events). Specifically, one could have hedged one's portfolio against the immense risks of the cryptocurrency market by choosing assets with different behavior profiles. The analyses done here were able to indicate lower volatility in categories of large market capitalization and to point out several outliers in the data. All of the outliers were assets that had had especially tumultuous February and the grouping by color shows that majority of the outliers were indeed mid-tier or small-tier assets, a phenomenon quite possibly attributable to aggressive trend shifts inherent in low market cap assets (both due to organic market sentiment and so called pump-and-dump schemes). The poor dimensionality reducing efficiency and lack of visible clusters given by our PCA suggests that the chosen approach is too naive for extracting any meaningful features. This is even though the variables were screened for outliers.

It is yet possible that this approach is viable in extracting features in groups of coins. It is likely that February was not a particularly eventful month and as such not suitable for this analysis. It could also improve segmentation results if there was some initial screening of coins with certain

characteristics. For example the data set could consist of solely inflationary and deflationary coins. Or they could be grouped by proof-of-stake and proof-of-work characteristics. Finally, it would be useful to have time series of *Coin Days Destroyed* and daily volatility also as variables. Coin days destroyed is a measure of how many days some transactioned coins have been sitting in someone's wallet and it could give a more meaningful glimpse into collective market sentiment over a coin.