Cryptocurrency Trading Data Analysis
Di Lu (QFDLU)   12/12/2018

Part 1 : Motivation

(a) Brief Introduction:

-      Crypto currency is a new type of currency circulating in the market. With Bitcoin invented in 2009, now there emerges over 2,000 types of crypto currencies and has a total market value more than 110 billion [https://coinmarketcap.com]. The price movement is exciting like riding with roller coaster, but people really don't know much about this virtual asset.

-      In this project, I will analyze the historical crypto trading dataset, to portrait its dynamic landscape, dig into features of crypto currencies, analyze if any patterns in their price movement.

(b) Interested Research Questions:

1.   How is the market of crypto currency emerging? Which currency has a leading position in the market? Do they survive or die over time?

2.   Explore the statistical characteristics of daily log return, how can we model it with distribution? Are different currencies show similar pattern?

3.   Crypto currency are virtual assets, do they have correlation with real world financial assets, e.g. USD, Gold price, stock indices. Can we use regression models to explain the dynamics of Crypto by real assets?

4.   What is the best currency to buy? Use Machine Learning techniques to select currencies with best performance.

Part 2 : Dataset Description

(a) Source:

-      From kaggle website, download dataset crypto-markets.csv, data updated on Dec 01, 2018
-      URL: https://www.kaggle.com/jessevent/all-crypto-currencies
-      Observations: 942,297, Variables: 13, Crypto Tokens: 2,071
-      From Yahoo finance and quandl, use pandas_datareader to download daily data from 2017-01-01 to 2018-11-30 for USD, gold price and stock index. Merge with crypto dataset by date.

(b) Variable:

-      'slug' is the unique symbol for each crypto, introduced to fix duplicate coins sharing 'symbol' or 'name'
-      Historical 'open', 'high', 'low', 'close' values for each crypto from 2013-04-28 to 2018-11-30
-      'ranknow' is the ranking of all currencies based on its market cap at 2018-11-30
-      'volume' is trading volume of one currency
-      'market' is the total market size = units * USD price per unit of currency
-      'close ratio' = (Close-Low)/(High-Low)
-      'spread' = (Close-Low)/Close (I modified by dividing Close to scale the wide range of different prices)

### (c) Data Manipulation and add new variables:

-      The original variables is raw daily trading data for crypto currencies. To analyze its dynamic over time or compare one currency over another, I performed log scaling: Crypto currency is growing exponentially fast from 2014 to 2017. Log scale is at a better position to describe its price/market size/trading volume.

- Handle missing data：missing data, such as N/A, inf, are handled by filling in 0.
- Add birth time: number of trading days since it came into existence in the market

## Part 3 : Analysis Method & Results

1. The emerging of crypto currency market can be visualized (with seaborn lineplot) by trading volume, market cap over time to determine if there are stages of development. Plot piechart for market composition, ratio of each currency of total market size.

Emerging of crypto currencies market: The emerging of crypto currency market can be visualized (with seaborn lineplot) by plotting trading volume, market cap over time. We can see three stages of development, the market comes into existence in 2013, booms in late 2017, with market size almost reaching 1 trillion, then followed by a crash during first half of 2018 up to now.
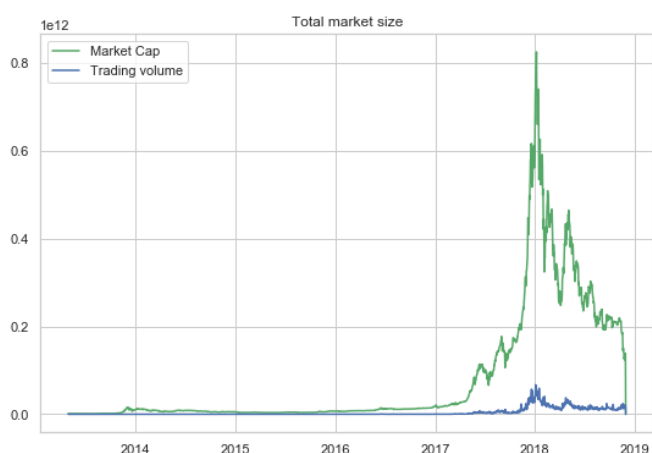

Leading market competitors: Selecting the top 5 cryptos by 'ranknow', then plot their market size over time. We can see the dynamics: top 5 competitors are relatively stable over time. They emerge from earlier 2014 to 2017. Bitcoin is leading the market, ripple recently surpasses ethereum to be at the second place.

Market composition: Take a snapshot on 2018-11-30, calculate the ratio of each currency to total market size, we can make a piechart. It shows Bitcoin is taking more than 50% of the market size. Top 10 cryptos compromise 85% of the total market.
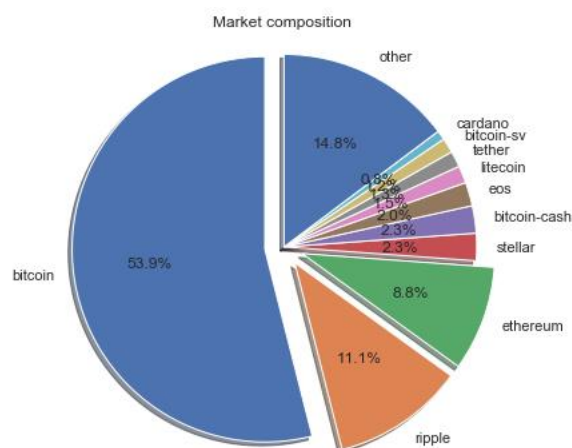
2. Start by studying the case of Bitcoin. Calculate its daily log return, check if normal distribution is a good fit (seaborn displot and QQ-plot). Performance Exploratory Data Analysis for other features. Study market size factor's effect on return.

Statistical characteristics of daily log return: Plot histogram of log return of Bitcoin, find it is centered around 0 and spread both sides. QQ-plot shows it is mostly symmetric and has much heavier tail than normal distribution, which coincides with it has skewness 0.19 and excess kurtosis 4.94.
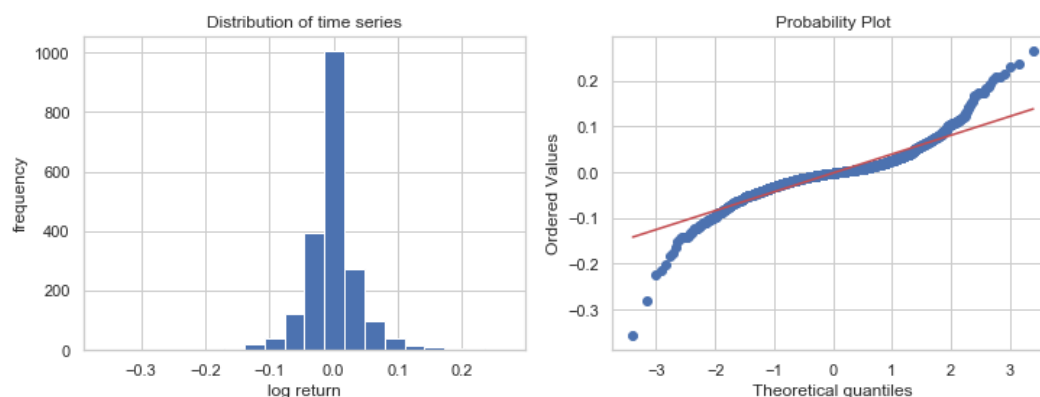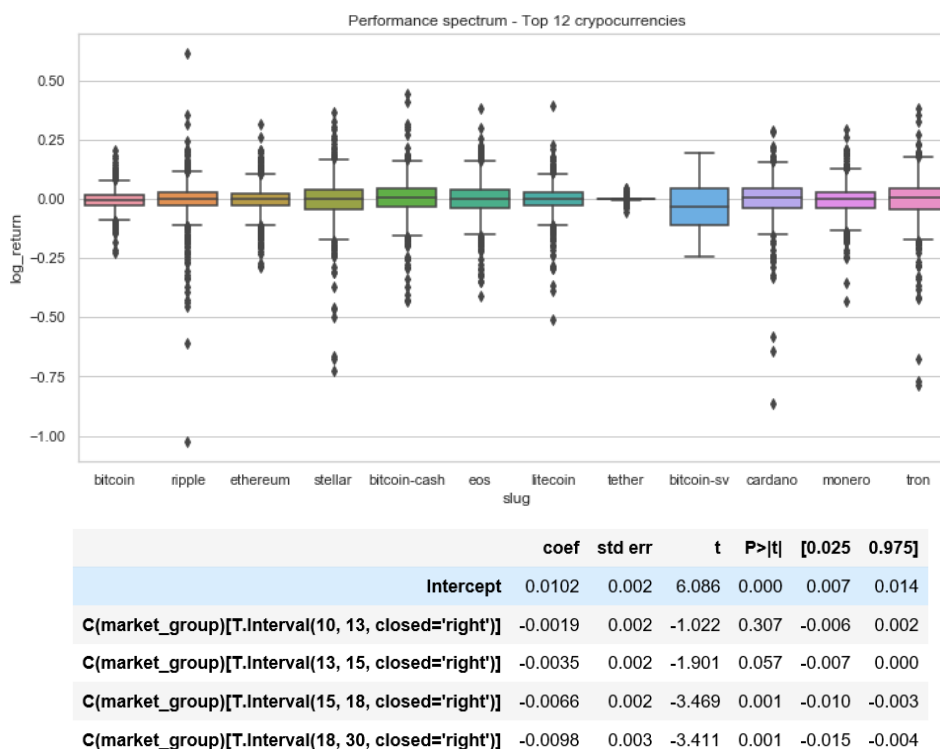

EDA for trading data: log_volume and log_market show patterns of bimodal distribution, spread looks like following exponential distribution and close_ratio is roughly uniform distributed.
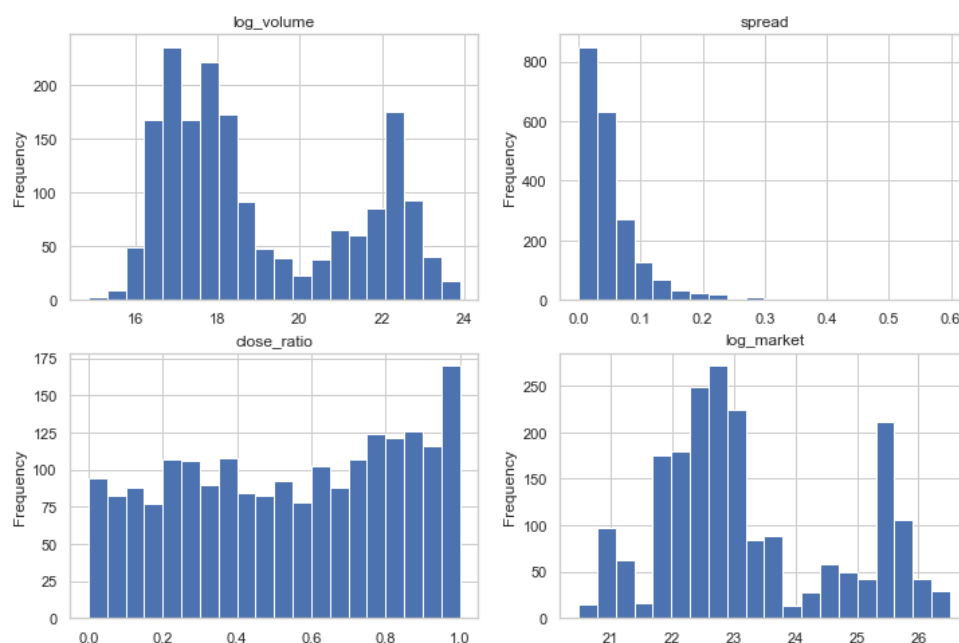
Performance analysis: Look at the boxplot for distribution of top12 cryptos and continue to analyze which factor has influence on the performance. For market size, is the larger the better? Traditional financial theory points out smaller companies which are more volatile will ask for more risk compensation and higher return. For crypto currencies, the rule also stands. Discretize the market size into five group, and carry out regression of log return on categorical variable - size factor, larger size cryptocurrencies

show significantly lower returns.

Performance spectrum - Top 12 crypocurrencies

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0102	0.002	6.086	0.000	0.007	0.014
C(market_group)[T.Interval(10, 13, closed='right')]	-0.0019	0.002	-1.022	0.307	-0.006	0.002
C(market_group)[T.Interval(13, 15, closed='right')]	-0.0035	0.002	-1.901	0.057	-0.007	0.000
C(market_group)[T.Interval(15, 18, closed='right')]	-0.0066	0.002	-3.469	0.001	-0.010	-0.003
C(market_group)[T.Interval(18, 30, closed='right')]	-0.0098	0.003	-3.411	0.001	-0.015	-0.004

3. Collect time series for USD, Gold price, stock indices daily market data with the same time window. Compute the correlation matrix and use heatmap for visualization. Try different regression models to explain Crypto currency price movement.

Correlation matrix: Select the top 3 cryptos and compare their return with 3 real world assets. The correlation between cryptos are high around 0.4-0.5, while the correlation with real world assets are very low, usually less than 0.1, considered weak correlation.

OLS Regression Results

Dep. Variable:	bitcoin	R-squared:	0.017
Model:	OLS	Adj. R-squared:	0.011
Method:	Least Squares	F-statistic:	2.712
Date:	Mon, 10 Dec 2018	Prob (F-statistic):	0.0445
Time:	20:53:55	Log-Likelihood:	751.38
No. Observations:	473	AIC:	-1495.
Df Residuals:	469	BIC:	-1478.
Df Model:	3		
Covariance Type:	nonrobust		

|	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0014	0.002	-0.619	0.536	-0.006	0.003
Gold_Price	-0.4170	0.365	-1.143	0.254	-1.134	0.300
Stock_Index	-0.8115	0.308	-2.631	0.009	-1.417	-0.205
USD_Index	-0.3659	0.617	-0.593	0.554	-1.579	0.847

Regression model: Explain the return of bitcoin with price change of 3 real world assets. The joint effect

is significant at 5% level (p-value 0.0445). The coefficient are all negative signs, indicating the opposite direction of the price movement between virtual and real world assets. Only stock index return has a significant effect on bitcoin return (p-value 0.009).

4. Given the metrics like market cap, trading, volume, spread, survival time, use machine learning algorithms to select the best currency. By K-Means clustering, select groups of currency yield higher returns. Also, may use dimensional reduction to extract features.

K-means Clustering: perform clustering (projected onto two dims) and tune parameter (number of cluster centers). By rule of thumb with n = 2,071, it is suitable to be 32 centers, inertia has elbow around 9 centers, highest silhouette score occurs at centers 3. I take 3 as number of cluster centers and plot as left below. The clusters are not obvious and number of points in each group may be not well balanced. The multidimensional scaling shows similar results, cryptos like 'vechain', 'status', 'iota' are isolated from others and most points cluster together. However, when I try to perform a T-test for performance difference between cluster: group 1 has a significant higher return 9.27% than group 0, and group 2 has a significantly lower return – 0.47% than group 0. The implication for trading is that to buy group 1 and to sell group 0 to grasp the absolute return from cluster difference.
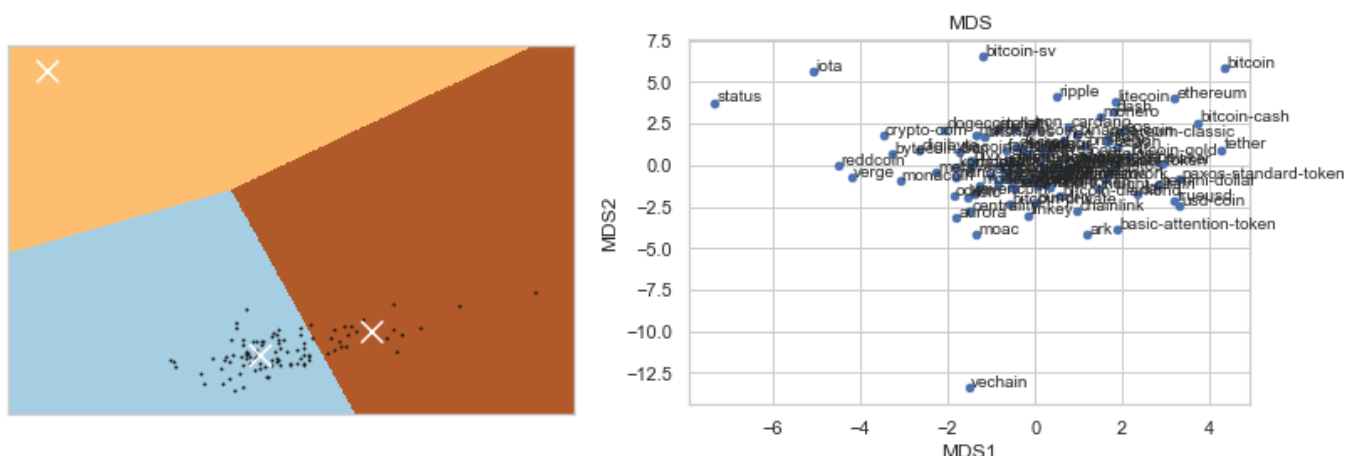
Exploratory Factor Analysis: We have measured many features of characteristics of cryptos, but did not look into the relationship between features. We can infer from the factor analysis that factor 1 is mostly related to std and mean, which is the first and second moments of log return, namely performance factor. Factor 3 is mostly related to log_market and log_volume, namely size factor. Factor 2 is more related to spread and kurtosis, negatively related to mean and skewness, namely maybe risk factor. The Principal Component Analysis gives similar result.
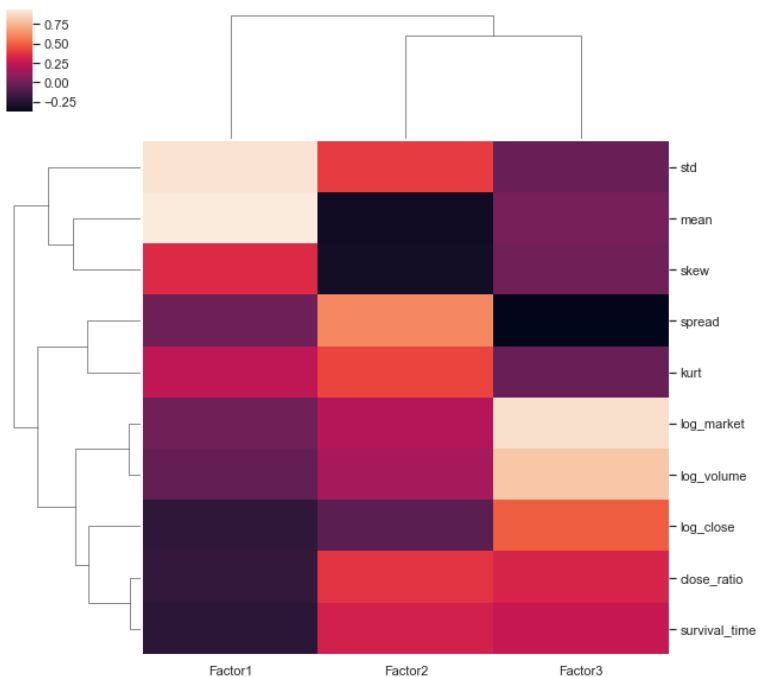
Part 4 : Summary and Conclusions

1. The market for cryptocurrencies booms in late 2017. Bitcoin, ripple, ethereum and other top 10 cryptos takes more than 75% of total market size.
2. Trading cryptocurrencies has high tail risk and extreme volatile times. Some differences between returns can be explained by size group factor, but the dynamics of cryptocurrencies have a weak correlation with real world financial assets.
3. By K-means clustering, I differentiate 3 groups with significant performance difference. With dimension reduction techniques, I am able to spot special currencies and extract three major factors, i.e. performance, size and risk to describe characteristics of cryptos.

Appendix :

1. Jupyter notebook (with working code, data manipulation process)
2. Dataset : crypto-markets-ver17.csv
3. Project proposal
4. The material are packed in crypto-trading-analysis_qfdlu.zip. Also available on URL: https://github.com/dlu-umich/Data-Manipulation-and-Analysis/tree/master