Unsupervised Learning:
Stock Market Clustering with K-Means and Gaussian Mixture Algorithms.

GAUKHAR
JAVAROVA

AGENDA

Data Introduction

Feature engineering

Clustering methods

Conclusion

What would you invest in?

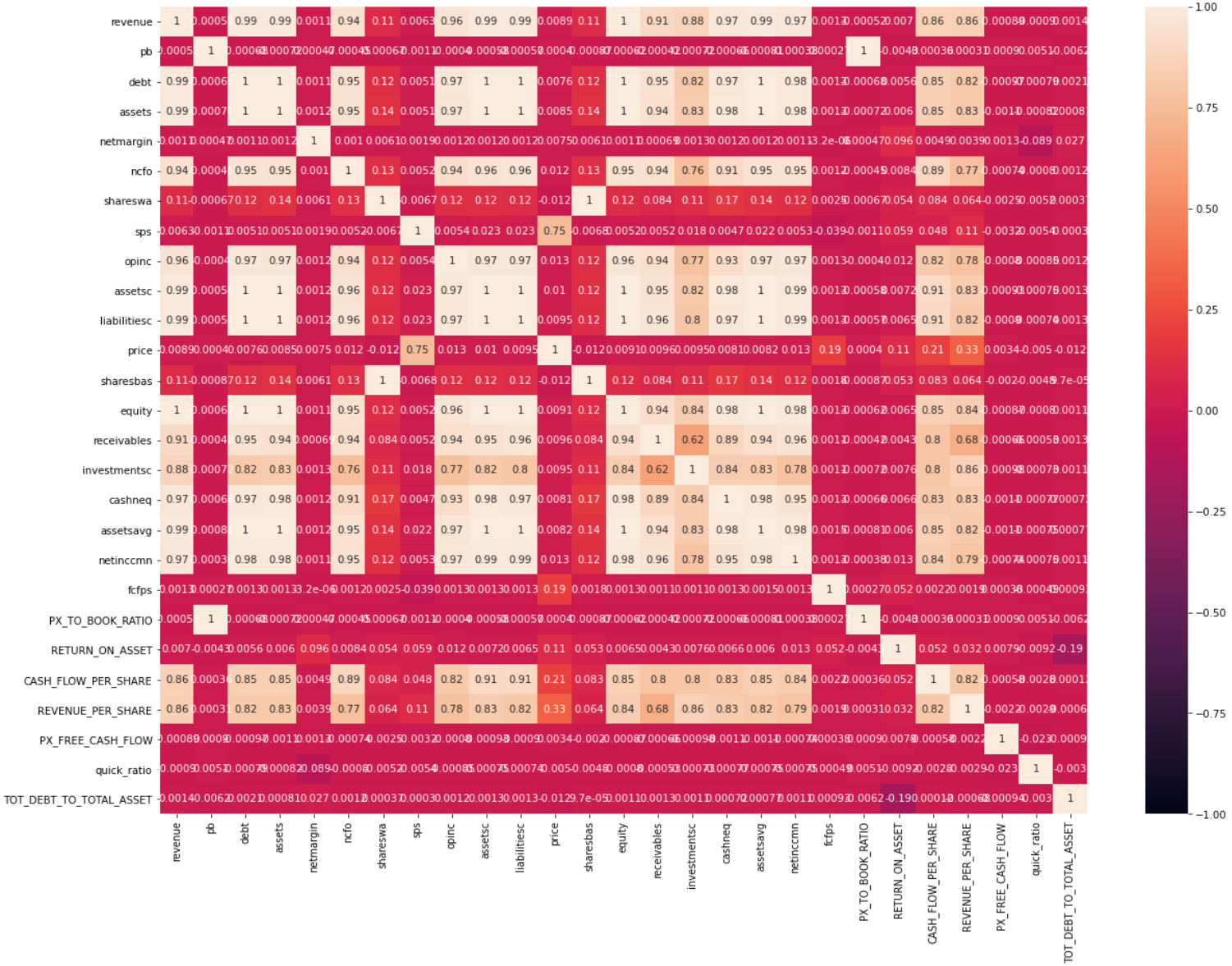
Data introduction

1486 companies were analyzed

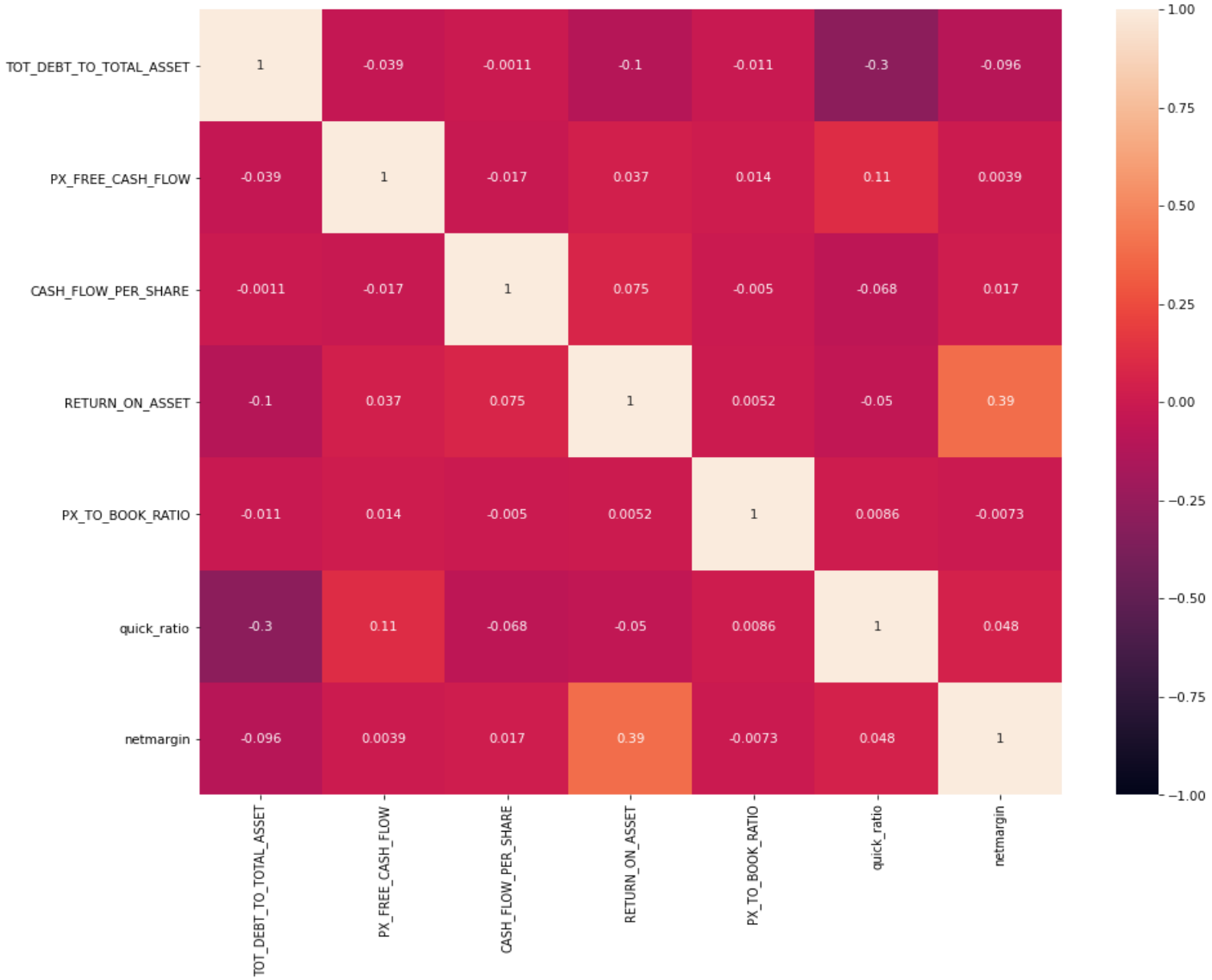
Indicators' descriptions can be found by the link.

Data includes ticker name, price, net margin, pb, calendar date

Raw data can be found here at QUANDL.COM

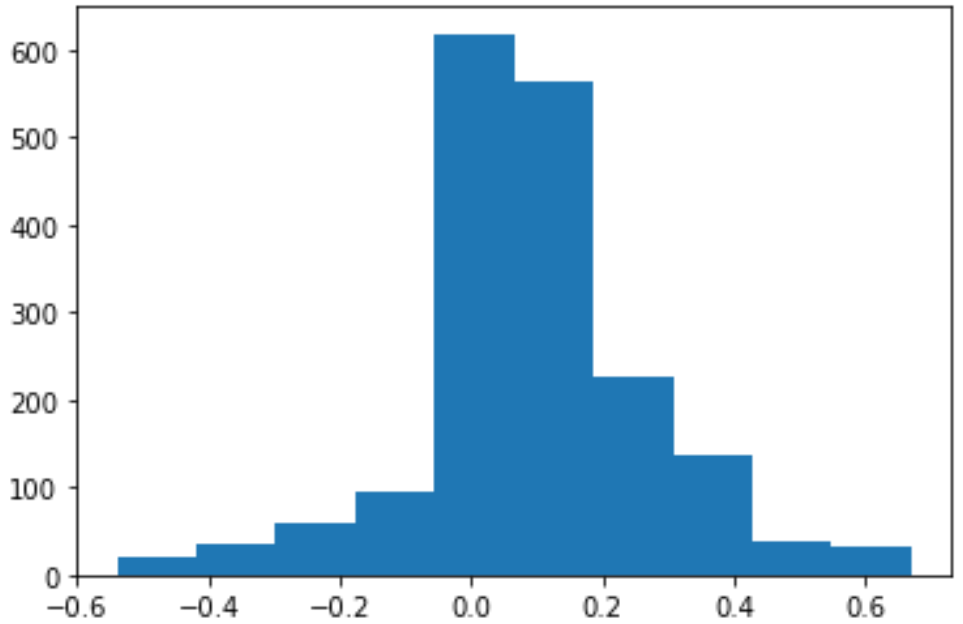


Feature Engineering



Cleaned Ratio Metrics

RETURN_ON_ASSET	PX_TO_BOOK_RATIO	quick_ratio	netmargin
0.758589	0.599277	0.022740	0.993041
0.730420	0.602942	0.077227	0.995603
0.663086	0.591529	0.005525	0.995560
0.724743	0.599518	0.021884	0.995614
0.732240	0.599128	0.027504	0.995485
0.744415	0.601173	0.023683	0.995578
0.787270	0.066328	0.014677	0.995804
0.721614	0.601944	0.031743	0.995504

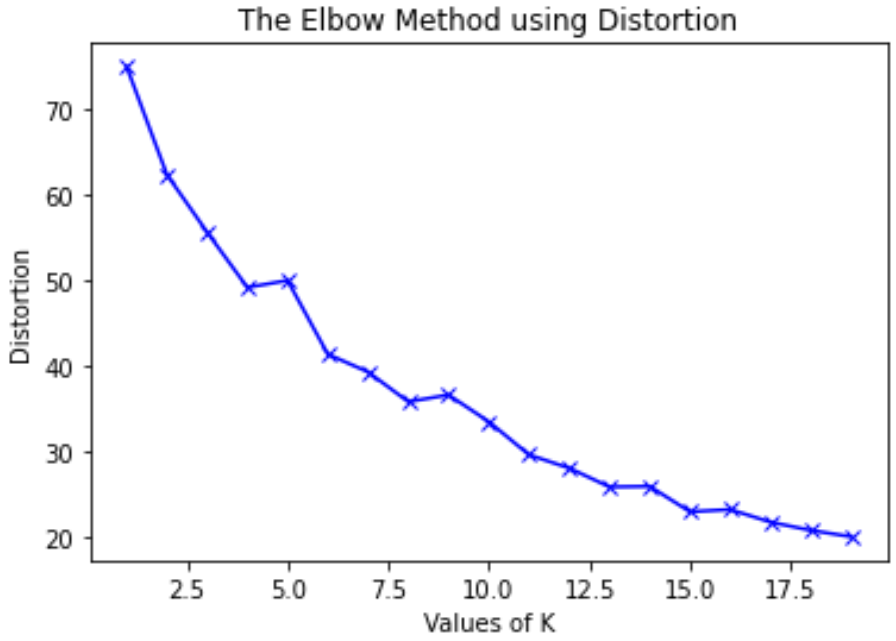


Z – score for ‘netmargin’ feature normalization is plotted

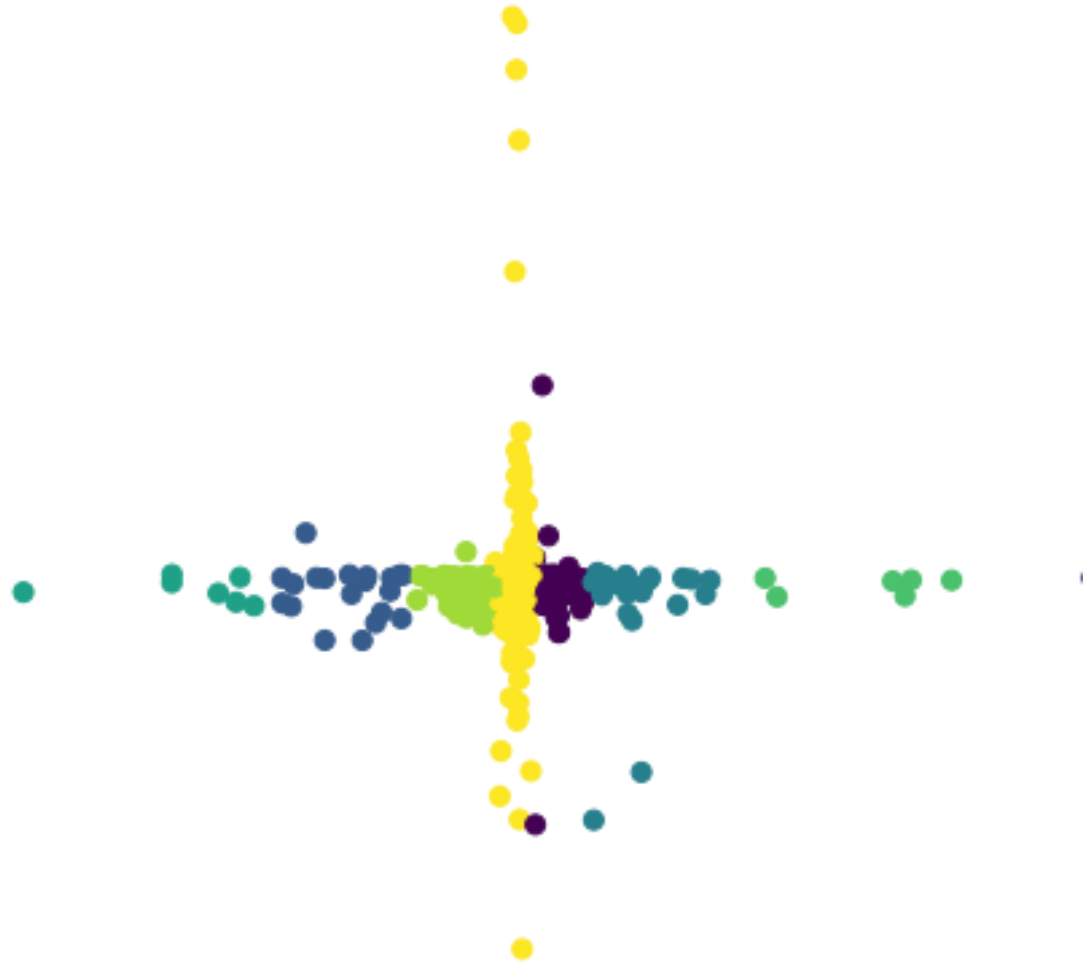
Scaled the data by StandardScaler

	TOT_DEBT_TO_TOTAL_ASSET	PX_FREE_CASH_FLOW	CASH_FLOW_PER_SHARE	RETURN_ON_ASSET	PX_TO_BOOK_RATIO	quick_ratio	netmargin
65	-0.683492	-0.060678	-0.390175	-0.398072	-0.032159	0.319745	-0.392442
86	-1.398403	0.355960	-0.383500	-0.318866	0.331491	0.953047	0.021462
159	0.264702	0.478250	-0.351998	-2.031868	0.225444	-0.100936	-1.113123
180	-0.209057	-0.018743	0.589083	0.902300	0.041992	-0.625693	0.342847
201	-0.005722	0.018594	0.008144	0.024113	-0.039413	-0.550184	0.386672

Elbow Method for the right number of clusters.



Values of K chosen number was 8

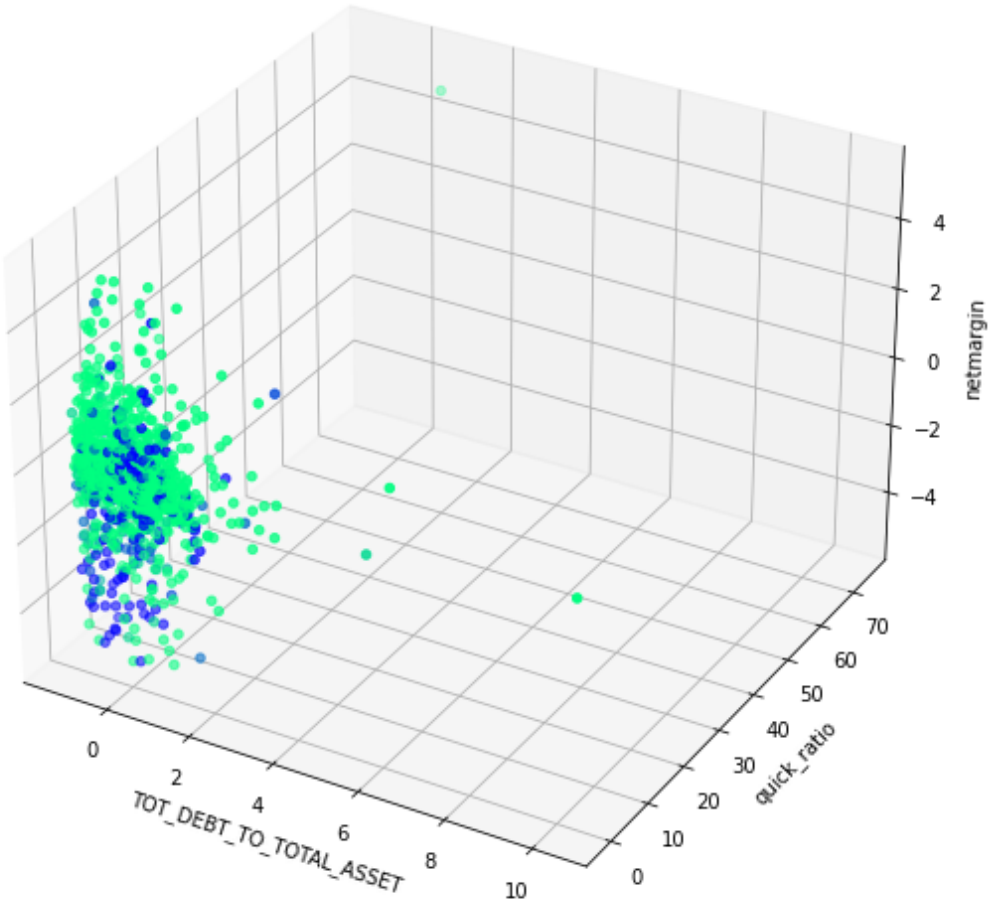


Principal Component Analyses

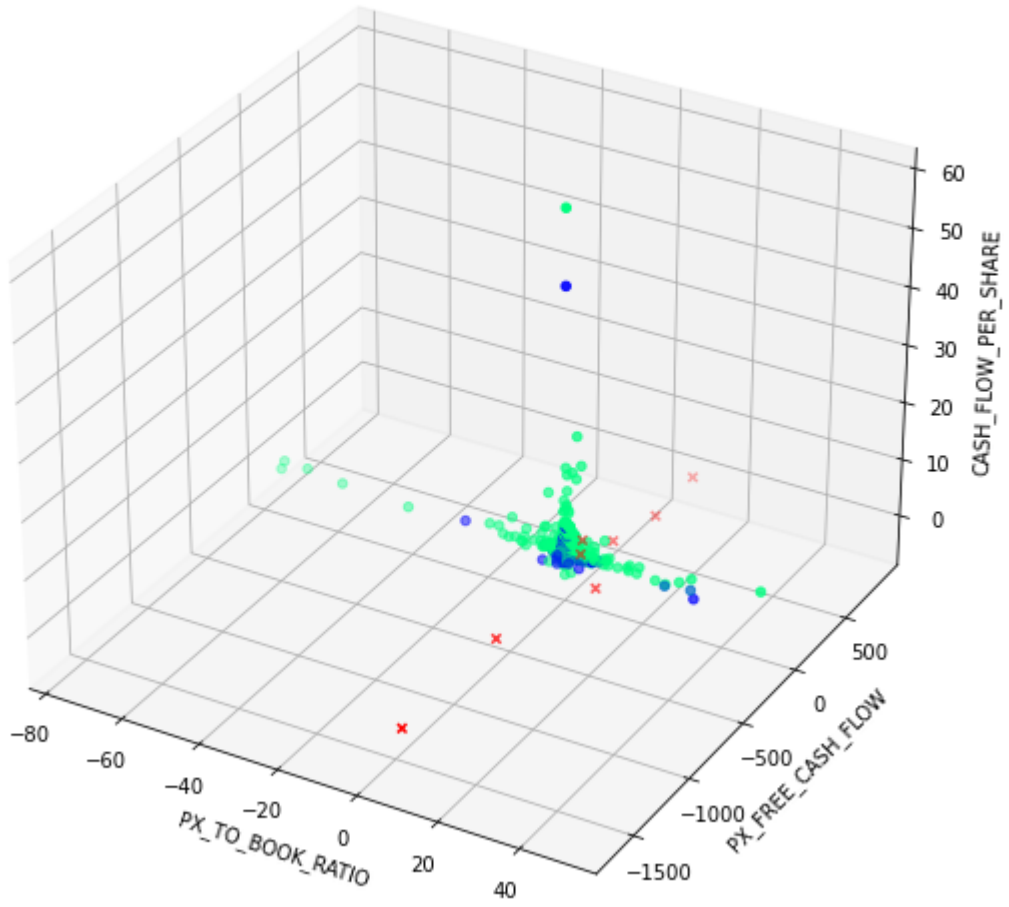
K-means clustering

label	TOT_DEBT_TO_TOTAL_ASSET	PX_FREE_CASH_FLOW	CASH_FLOW_PER_SHARE	PX_TO_BOOK_RATIO	quick_ratio	netmargin
0	0.327349	-49.395351	1.309120	5.300146	1.704368	-0.034025
1	0.221496	-1644.023026	0.562477	5.940000	1.255016	-0.079750
2	0.352457	450.224064	0.460211	10.959150	1.855371	0.015050
3	0.262885	-253.329800	0.913730	14.770100	3.037815	-0.030867
4	0.234888	846.430599	0.303203	10.003286	2.283584	-0.037143
5	0.192617	-835.746015	0.881952	5.988500	2.721528	-0.059667
6	0.312703	150.385770	0.695397	8.173250	2.111688	0.012517
7	0.346103	22.781756	2.200923	3.738358	1.532843	0.077911

Visualization of clustered data with 2 clusters



Visualization of clustered data with 5 clusters

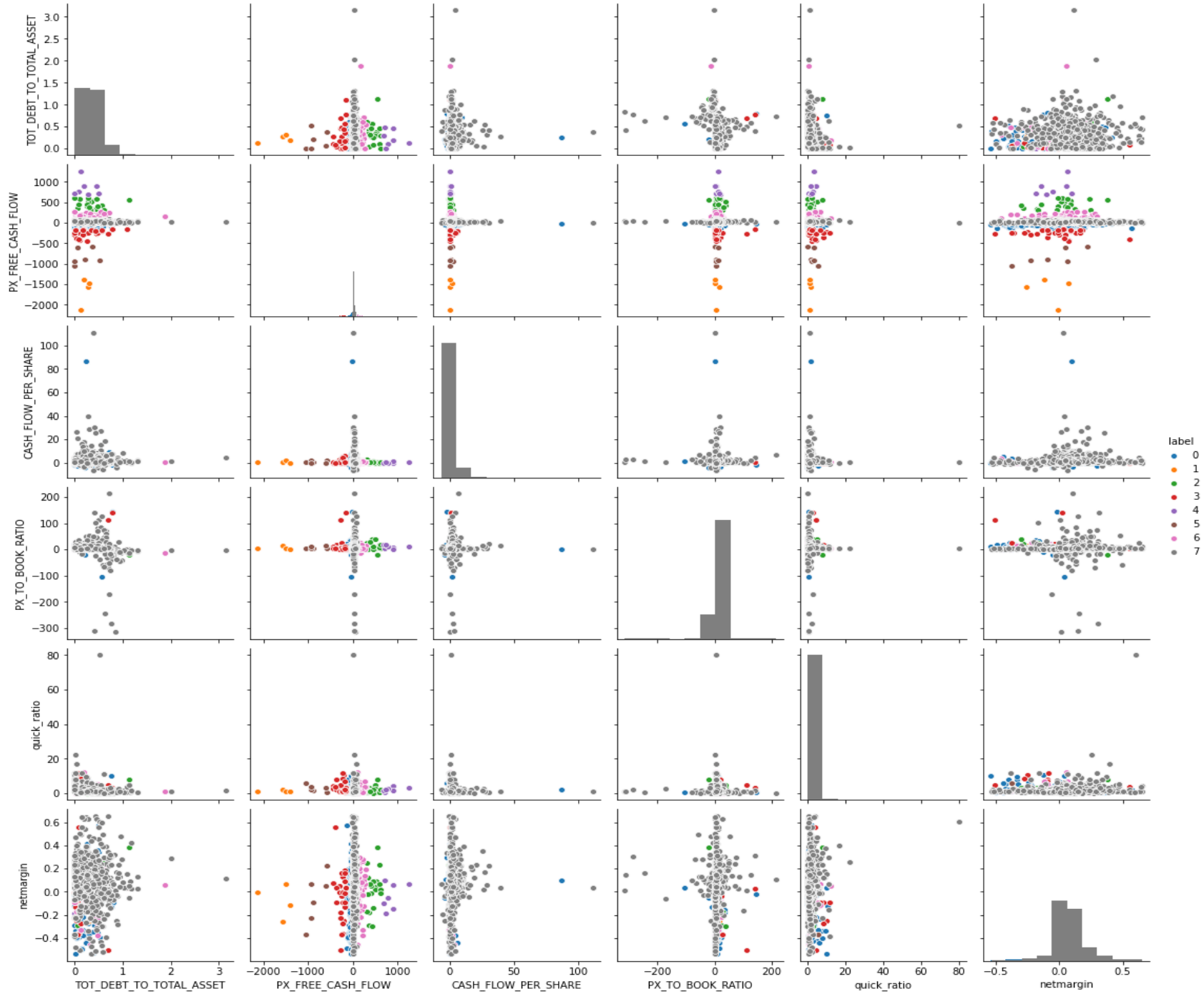


K-means labels

7	1088
0	158
6	60
3	30
2	20
4	7
5	6
1	4

K- means Silhouette Score

0.6195445780322226


TOT_DEBT_TO_TOTAL_ASSET
PX_FREE_CASH_FLOW
CASH_FLOW_PER_SHARE
PX_TO_BOOK_RATIO
quick_ratio
netmargin
label
0
1
2
3
4
5
6
7

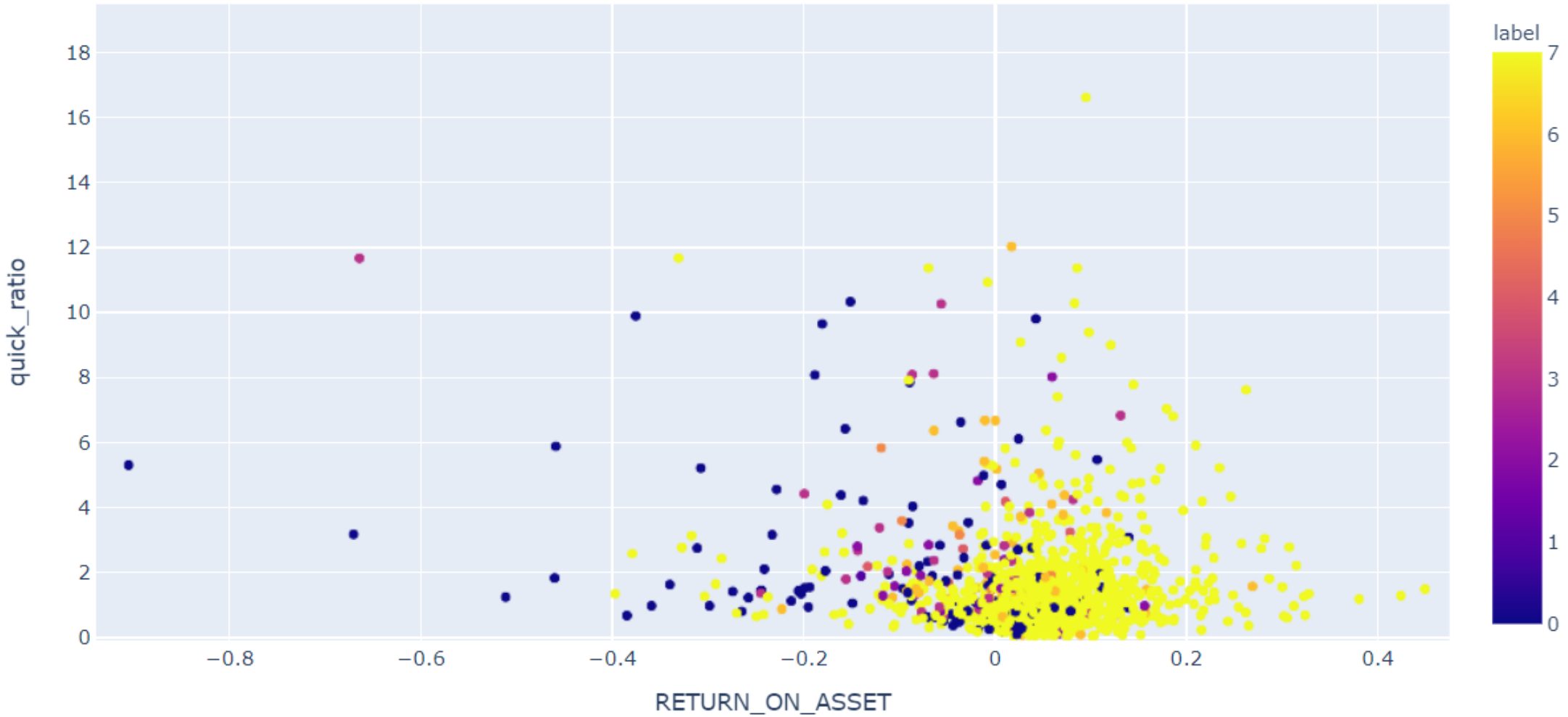
Gaussian Mixture Models Clustering

Gaussian silhouette score
0.5231161145540553

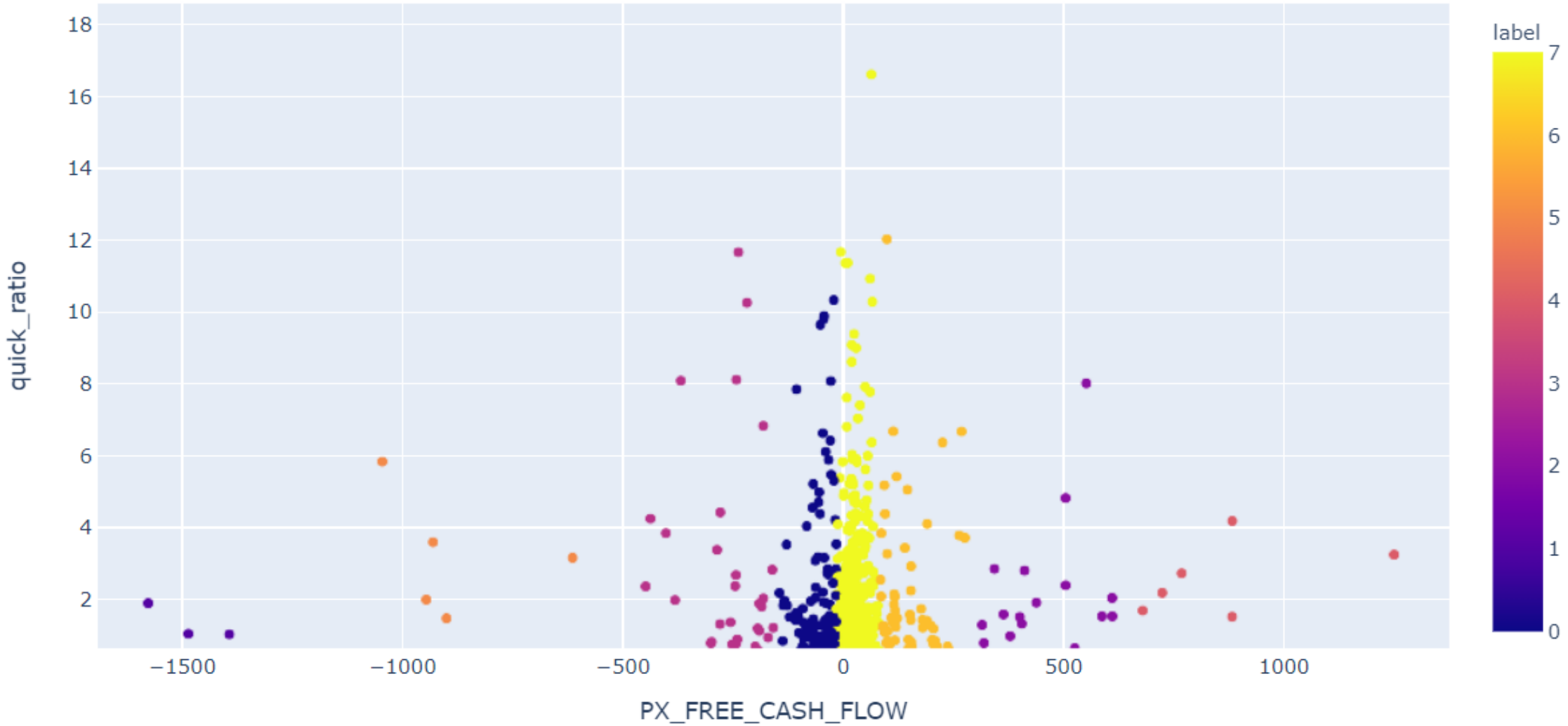
Number of components 8

Covariance Full

What would you invest in?



1. Profitable and quick to liquidate?
2. Not profitable and hard to liquidate?



1. Overpriced and hard or easy to liquidate shares?
2. Undervalued and easy or hard to liquidate shares?

CONCLUSION

- **Cluster 0** – price to free cash flow is low, stocks are undervalued (ex. AAL)

- **Cluster 1** – extremely undervalued, with good fundamentals. Good to research it more
- **Cluster 3** – undervalued, has debt

- **Cluster 2** – overpriced, little to no debt
- **Cluster 4** – same as 2

- **Cluster 5** – overvalued, no debt, increasing revenue per share (good outliers)

- **Cluster 6** – pb is high, revenue per share is highly profitable, overpriced. (ex. Zoom)

- **Cluster 7** – majority of stocks, pb is high, revenue is growing over time. Considered as long-term value stocks.