Financial Data Science

G7 (not the currency)

Fernando Oktavianes
Gan Kai Heng Cassidy
Jiang Jin
Miti Nopnirapath
Stephen Kusrianto

Part I

Problem 1 - Signals for the last few days (6 to be exact) are zero

6 rows × 6 columns

Date	Signal	Open	High	Low	Close	Adj Close
2019-12-27	0.0	167.119995	167.119995	165.429993	165.860001	164.039063
2019-12-30	0.0	165.979996	166.210007	164.570007	165.440002	163.623688
2019-12-31	0.0	165.080002	166.350006	164.710007	165.669998	163.851135
2020-01-02	0.0	166.740005	166.750000	164.229996	165.779999	163.959946
2020-01-03	0.0	163.740005	165.410004	163.699997	165.130005	163.317093
2020-01-06	0.0	163.850006	165.539993	163.539993	165.350006	163.534668

Problem 2 - Adj close minimum value is negative

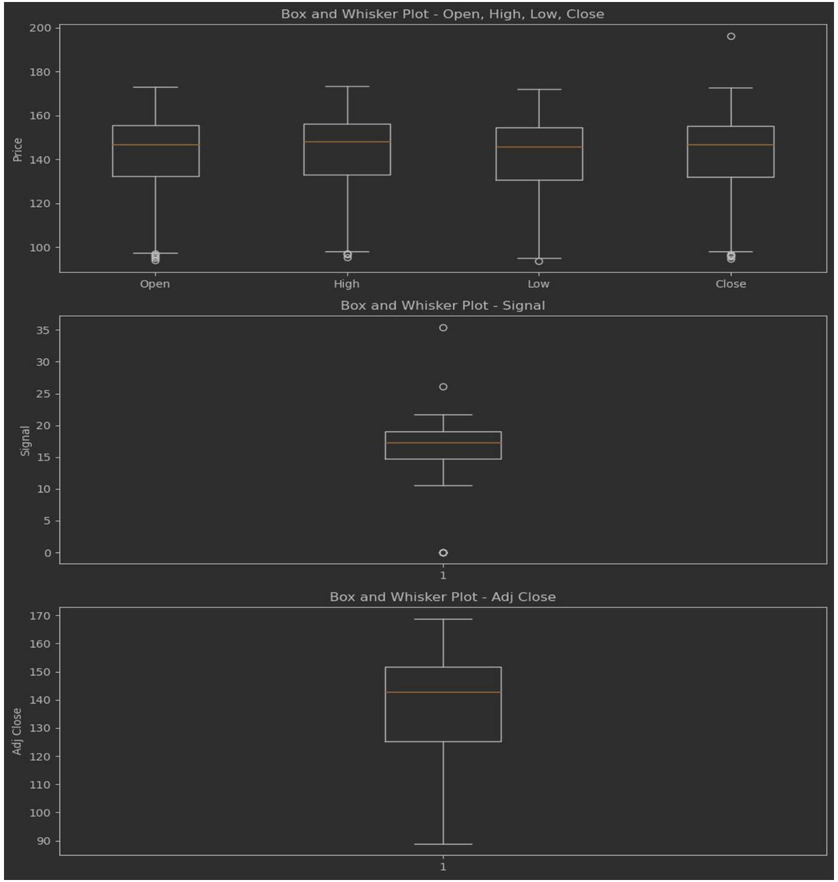
8 rows × 6 columns

	Signal	Open	High	Low	Close	Adj Close
count	1038.000000	1038.000000	1038.000000	1038.000000	1038.000000	1038.000000
mean	16.766190	141.847360	142.691801	140.907746	141.840973	136.341060
std	3.095783	18.475574	18.470255	18.404504	18.497010	21.427837
min	0.000000	94.080002	95.400002	93.639999	94.790001	-152.277847
25%	14.691150	132.132496	132.912495	130.542503	131.824993	125.290491
50%	17.298240	146.769997	147.959999	145.634995	146.885002	142.667732
75%	19.030890	155.367496	156.287495	154.422500	155.289993	151.798325
max	35.434147	172.789993	173.389999	171.949997	196.279999	168.842270

Part I

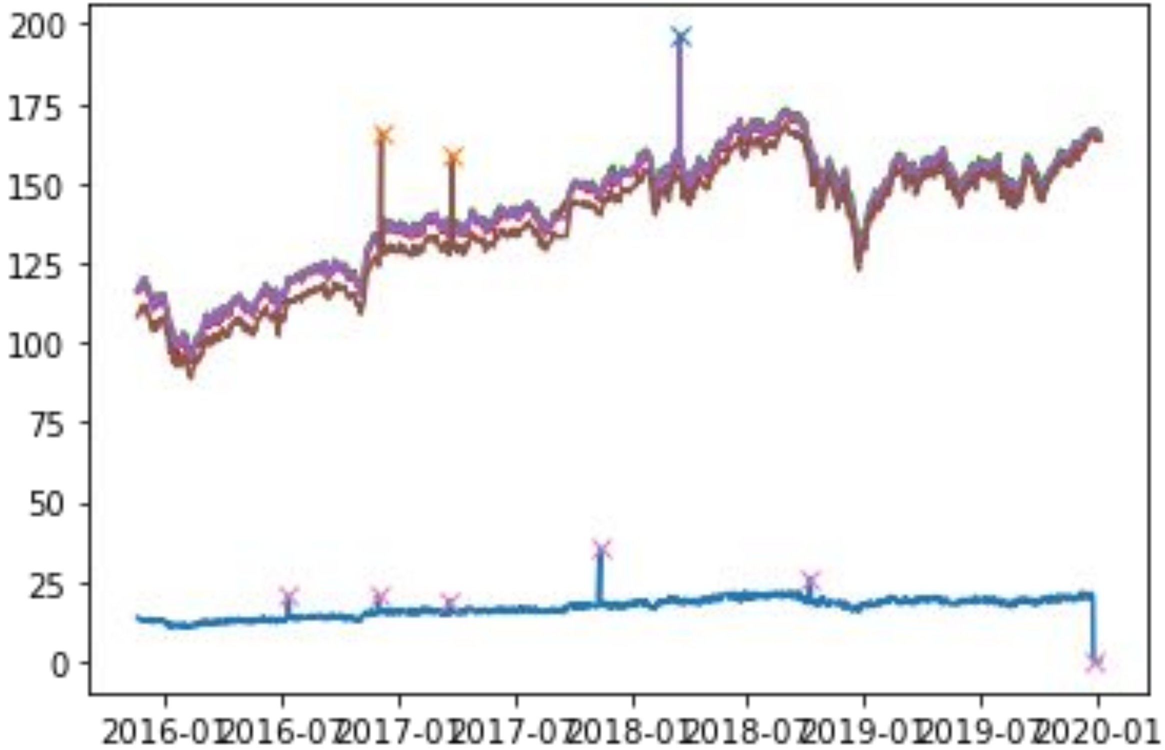
Problem 3

Outliers in the dataset



Part I

Assumption: Outliers is identifiable as an extreme value change, followed by an extreme reversion. With this logic, we can automatically detect outliers.



Part I

Problem 4

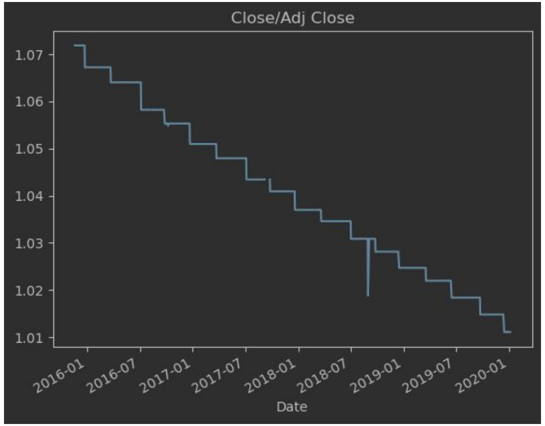
38 rows contain NaN values

10-20 38 rows × 4 columns

Date	Open	High	Low	Close
2017-07-04	NaN	142.600000	141.400003	142.200006
2017-07-05	141.699997	141.850006	140.699997	141.589996
2017-08-01	142.169998	142.199997	NaN	141.779999
2017-09-08	138.929993	139.770004	138.619995	NaN
2017-09-11	140.389999	140.919998	140.229996	NaN
2017-09-12	141.039993	141.690002	140.820007	NaN
2017-09-13	141.410004	142.220001	141.320007	NaN
2017-09-14	141.779999	142.160004	141.419998	NaN
2017-09-15	141.639999	142.470001	141.550003	NaN
2017-09-18	142.619995	143.809998	142.600006	NaN

Problem 5

Irregularity in ratio of Close/Adj Close observed



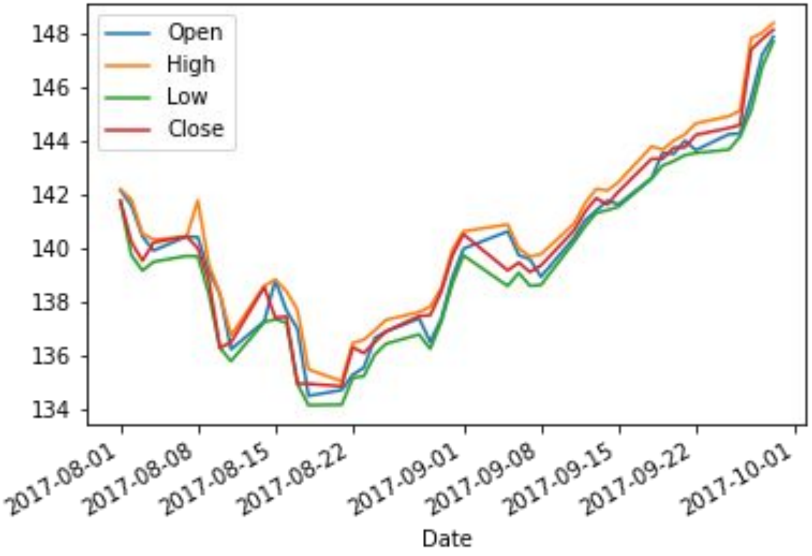
Part I

Problem 6 - Handle inconsistent data values

- Shifting high and low to max and min value
- Shift and scale so that Open-High-Low and Close of each rows are 0 mean and 1 variance
- Handle imputations
- Forward fill signal for last 6 values that are NaN

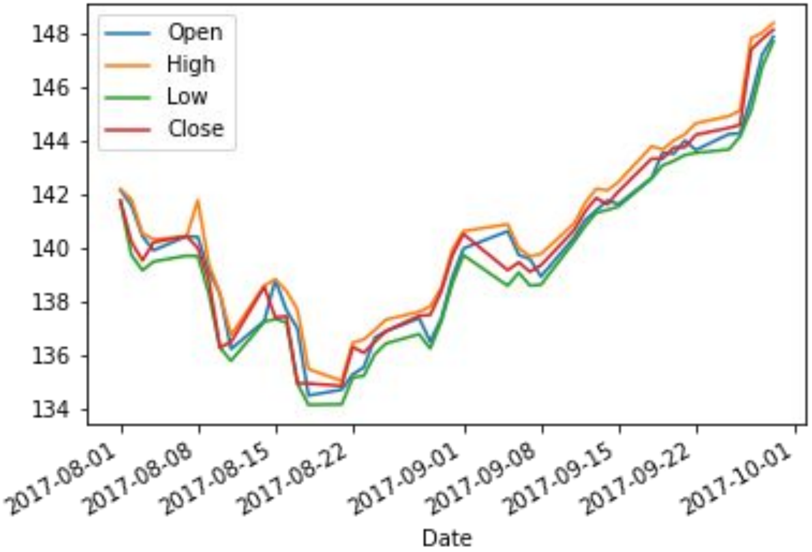
Part I

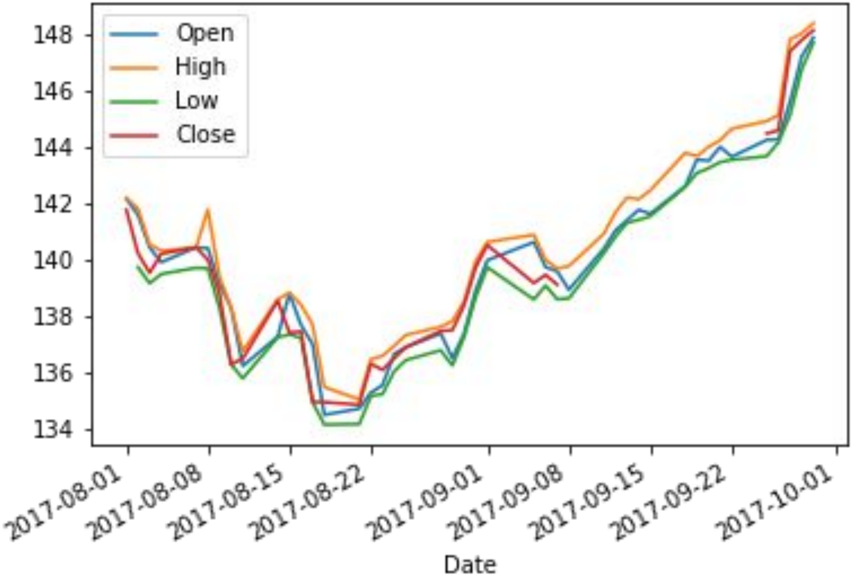
Can you spot what is wrong?



Part I

Some of it is completely made up! Imputed with RandomForest





Part I

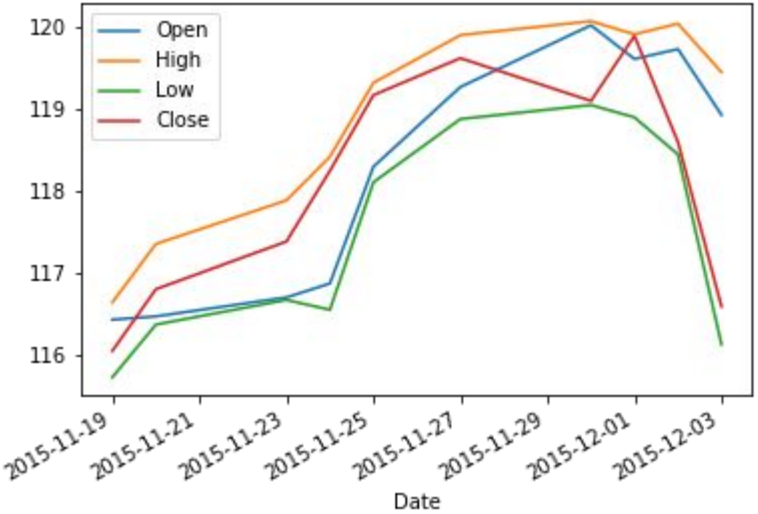
Imputation Process

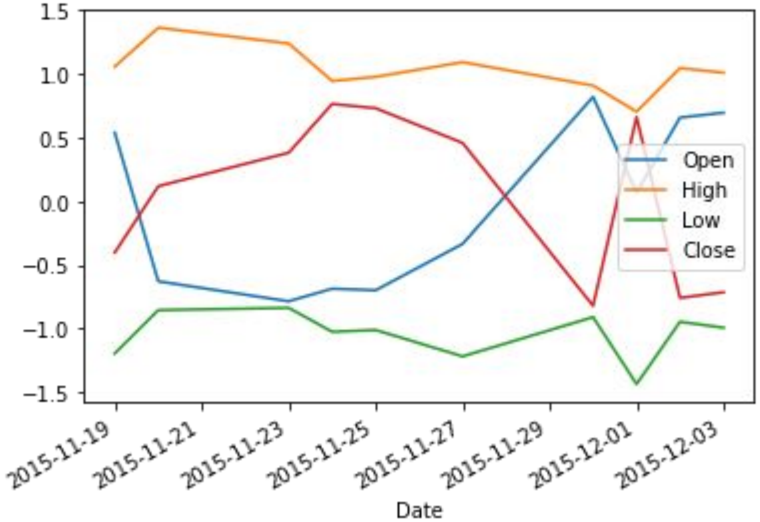
Standardize Open High Low Close to mean 0 and std 1.

- We save the original offset and original std so we can transform our data back.

Use Sklearn IterativeImputer with RandomForest to fill in missing values

Reverse the transformation by adding back offset and multiply by original standard deviation



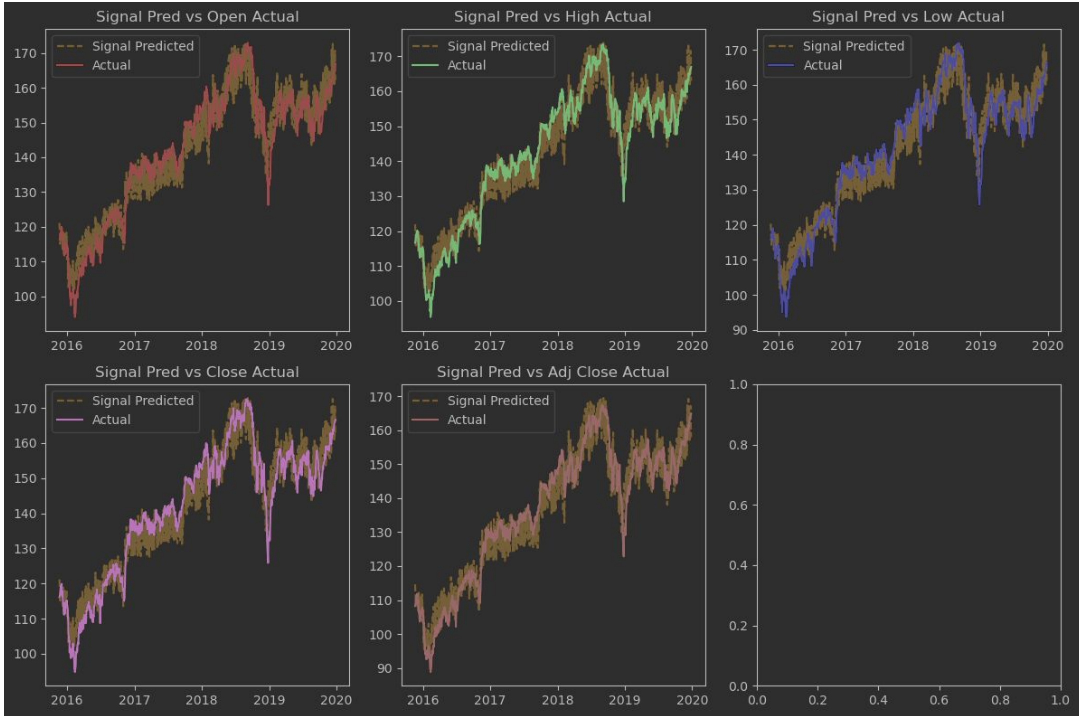


We transform from left values to right values by taking away the mean of OHLC. The values on the right can then be used to train randomforest on and fill other missing data without issues with out-of-range observation.

This process is reversible, and we can transform from right back to left after imputing

Part I

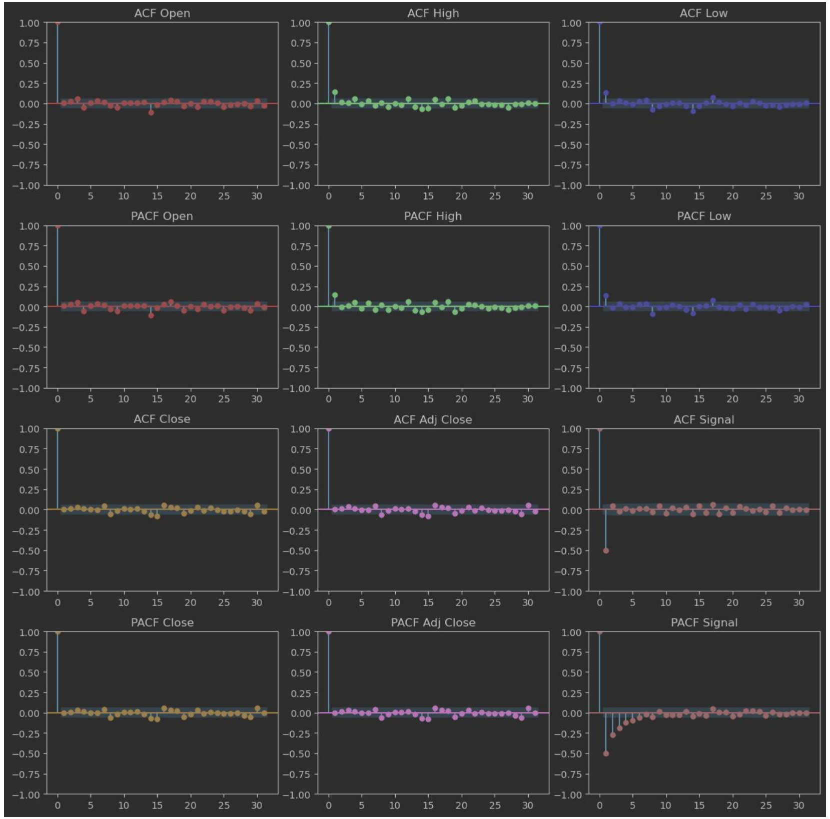
Signal Analysis



Part I

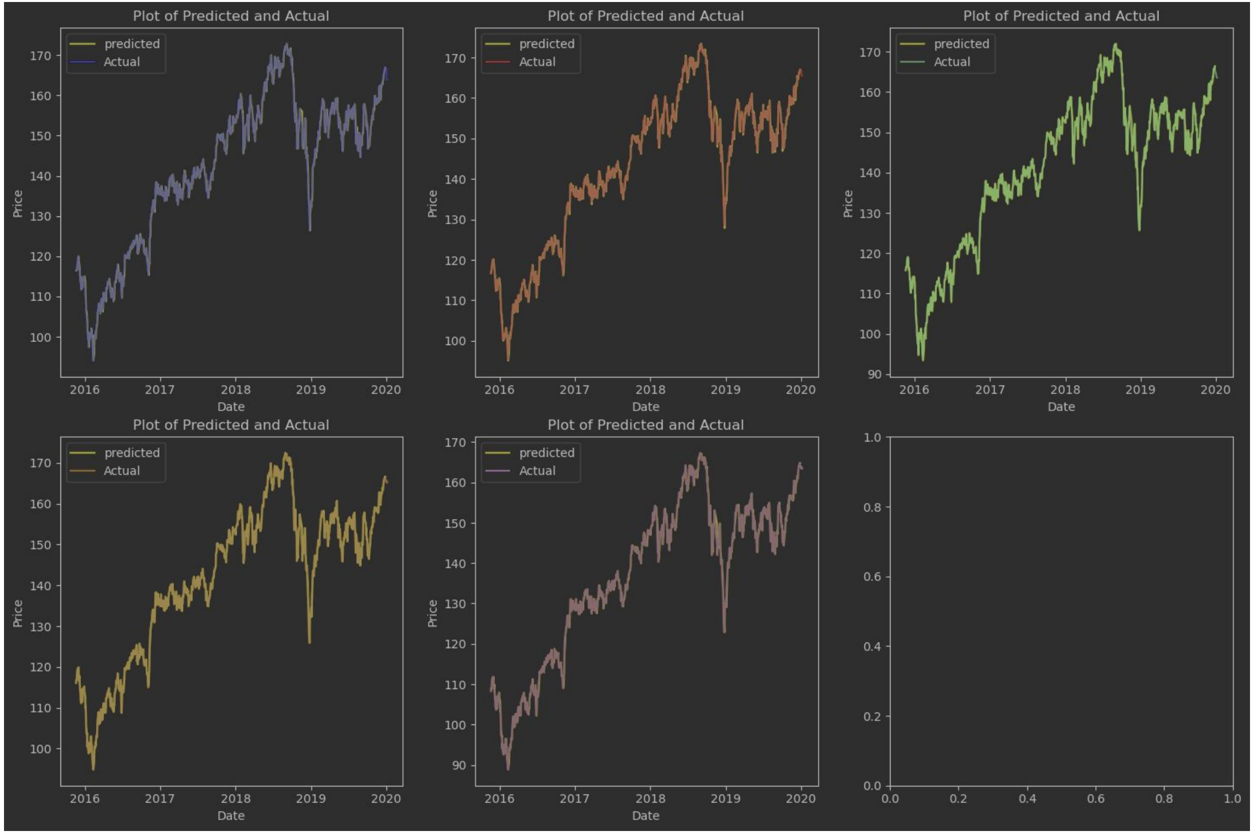
Evidence of MA (1) are in the signals given to us.

This is despite the fact that no such process appeared in the KLines/Candlestick data.



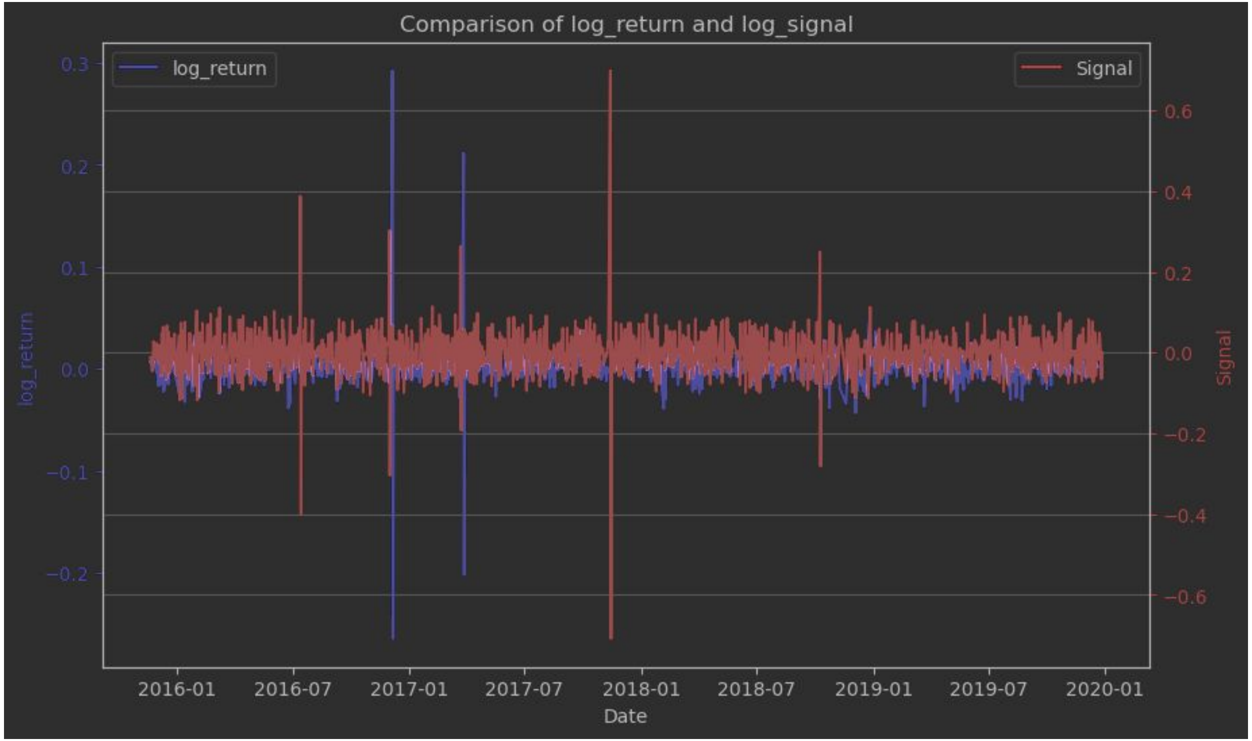
Part I

ARIMA (1,0,1) Modelling



Part I

Mean Squared Error: 0.00403771



Part II

Click to add text

Part III Clustering stocks with Analysts coverage

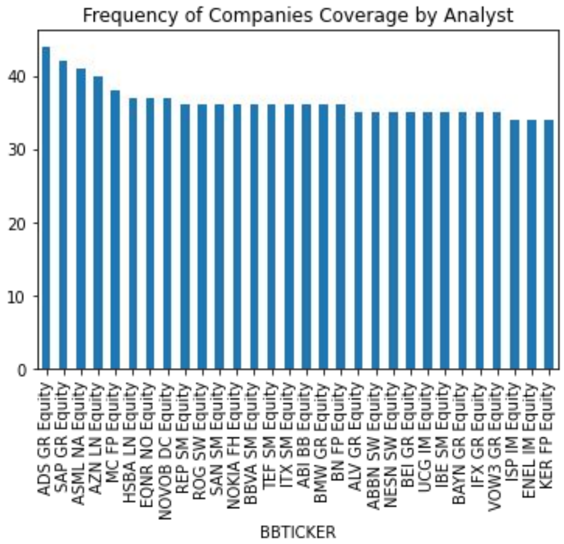
We are given a data of 8676 rows. We will only be looking at this four columns as they are the most useful for clustering stocks. The rest are either irrelevant or an expert's opinion, which is the last thing we want.

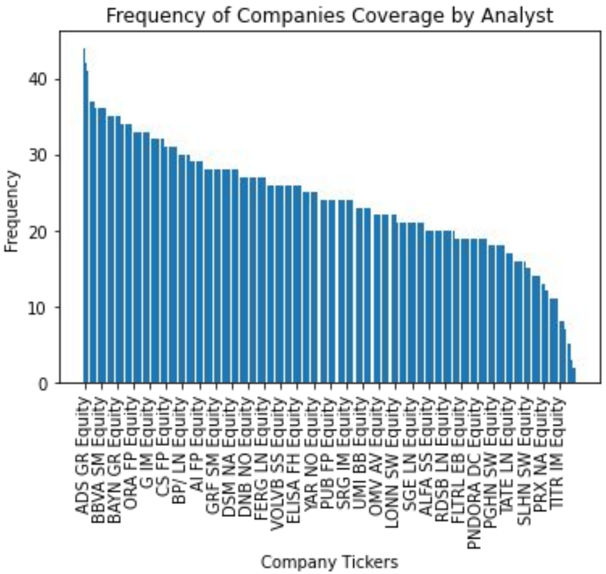
	ANALYST	BBTICKER	GICS_SECTOR_NAME	GICS_INDUSTRY_GROUP_NAME
0	Jamrgett	NESN SW Equity	Consumer Staples	Food, Beverage & Tobacco
1	Joneeney	NESN SW Equity	Consumer Staples	Food, Beverage & Tobacco
2	MarDeboo	NESN SW Equity	Consumer Staples	Food, Beverage & Tobacco
3	Niclberg	NESN SW Equity	Consumer Staples	Food, Beverage & Tobacco
4	Antpagna	NESN SW Equity	Consumer Staples	Food, Beverage & Tobacco
...	...	...	...	...
8671	Inghmidt	LHA GR Equity	Industrials	Transportation
8672	Xavaroen	BMW3 GR Equity	Consumer Discretionary	Automobiles & Components
8673	FraMaury	BMW3 GR Equity	Consumer Discretionary	Automobiles & Components
8674	RenWeber	UHRN SW Equity	Consumer Discretionary	Consumer Durables & Apparel
8675	Loiorvan	UHRN SW Equity	Consumer Discretionary	Consumer Durables & Apparel

Part III

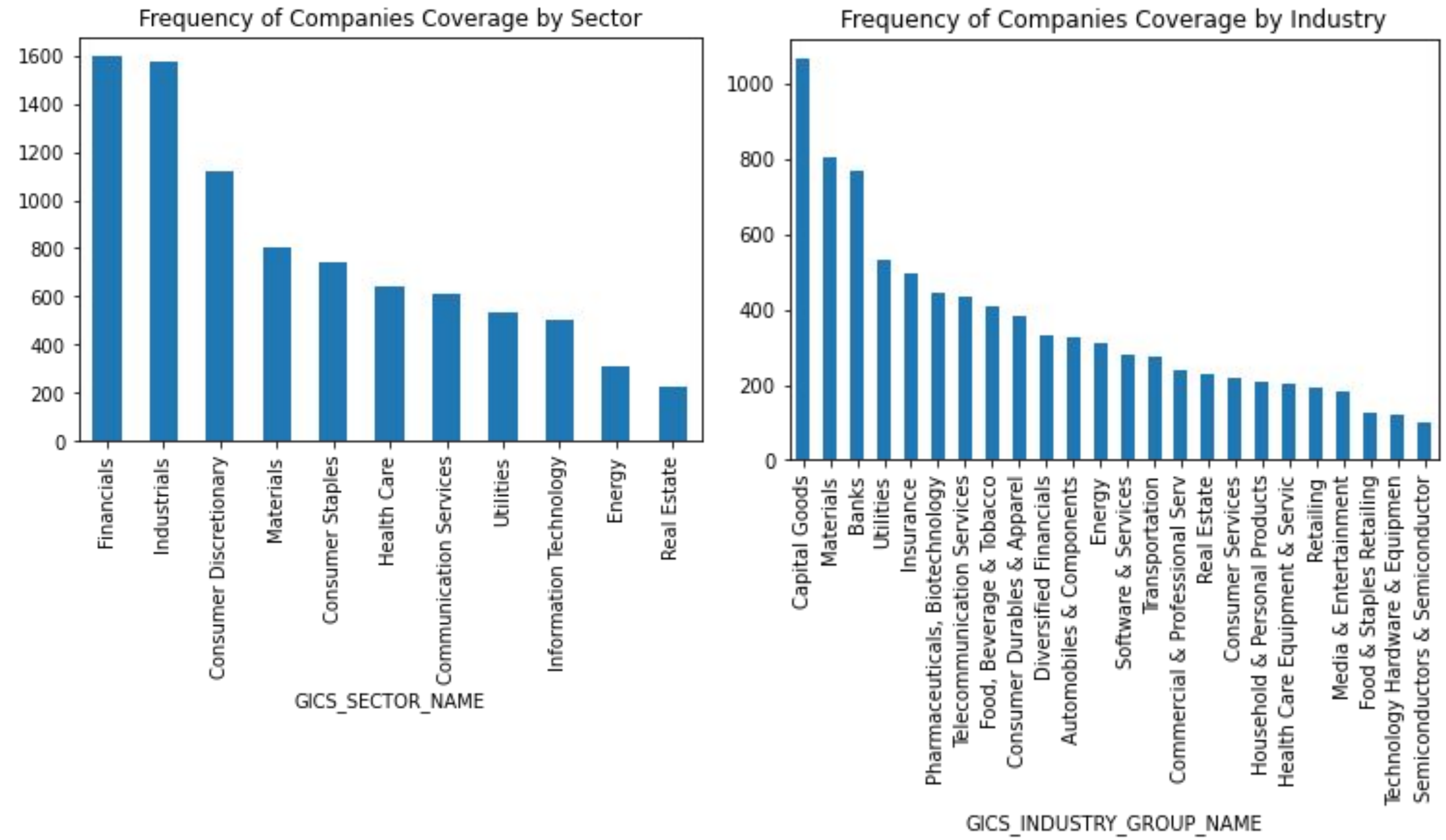
Q1 Which company has the higher analyst coverage?

Of the 360 stocks given to us, most (80-90%~) of the stocks will be covered by at least 20 analysts. Most of which is ADS GR.





Part III

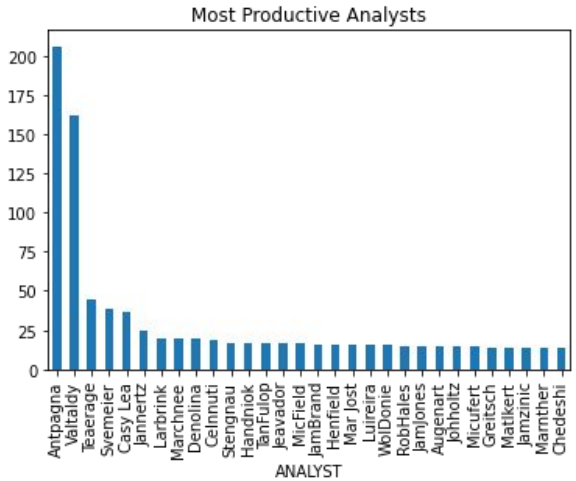


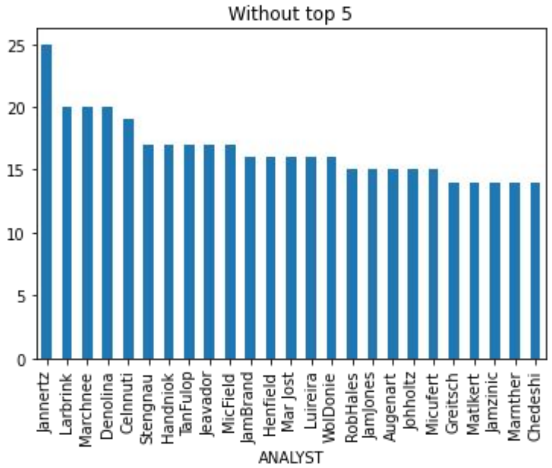
Frequency of Companies Coverage by Sector
1600
1400
1200
1000
800
600
400
200
0
Financials
Industrials
Consumer Discretionary
Materials
Consumer Staples
Health Care
Communication Services
Utilities
Information Technology
Energy
Real Estate
GICS_SECTOR_NAME
Frequency of Companies Coverage by Industry
1000
800
600
400
200
0
Capital Goods
Materials
Banks
Utilities
Insurance
Pharmaceuticals, Biotechnology
Telecommunication Services
Food, Beverage & Tobacco
Consumer Durables & Apparel
Diversified Financials
Automobiles & Components
Energy
Software & Services
Transportation
Commercial & Professional Serv
Real Estate
Consumer Services
Household & Personal Products
Health Care Equipment & Servic
Retailing
Media & Entertainment
Food & Staples Retailing
Technology Hardware & Equipmen
Semiconductors & Semiconductor
GICS_INDUSTRY_GROUP_NAME

Part III

Q2 Which analyst covers the most companies?

Some Analysts are extremely productive, covering 200+ stocks. However, if we exclude the top 5, many still covers a very reasonable number of 15+





Part III

Many analysts (544 out of 2065) only covered 1 stocks. They would add no value to our purpose of using them for clustering.

Statistics	Before winsorize	After winsorize 10-90
mean	4.20	4.01
std	6.60	1.62
median	3	4
mode	1	2

Part III

Q3 Clustering methodology

We will convert the analysts data into a vector of 1 and 0 for each stocks. Each element corresponds to a specific analysts, and 1 indicates that the stocks have been covered by the analysts.

In the figure to the right, we can see what this would look like with 5 analysts. 1COV GR is covered by Antpagna and Svemeir, and is not covered by the other 3.

Given that the values are bounded between 0 and 1, this should be suitable for Euclidean distance calculation.

ANALYST	Antpagna	Casy Lea	Svemeier	Teaerage	Valtaldy
BBTICKER					
1COV GR Equity	1	0	1	0	0
AAL LN Equity	0	0	1	0	1
ABBN SW Equity	0	0	0	0	1
ABF LN Equity	1	0	0	0	1
ABI BB Equity	1	0	0	1	1
...	...	...	...	...	...
WPP LN Equity	0	1	0	0	1
WRT1V FH Equity	1	0	0	0	0
WTB LN Equity	1	0	0	0	1
YAR NO Equity	1	0	0	0	0
ZURN SW Equity	0	1	0	0	1

300 rows × 5 columns

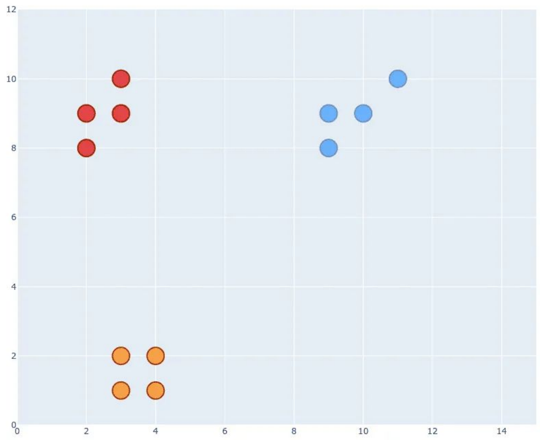
Part III

t-SNE (t-distributed Stochastic Neighbor Embedding)

The essence of this clustering method is that it calculates Euclidean distance (in the default implementation) between each samples and use that distance as the ‘attraction factor’ between each points.

It then instantiates a completely random distribution in a lower dimension space, and then “shake things around” and let the attraction factor make the samples cluster with each other in a random (stochastic) manner.

Part III



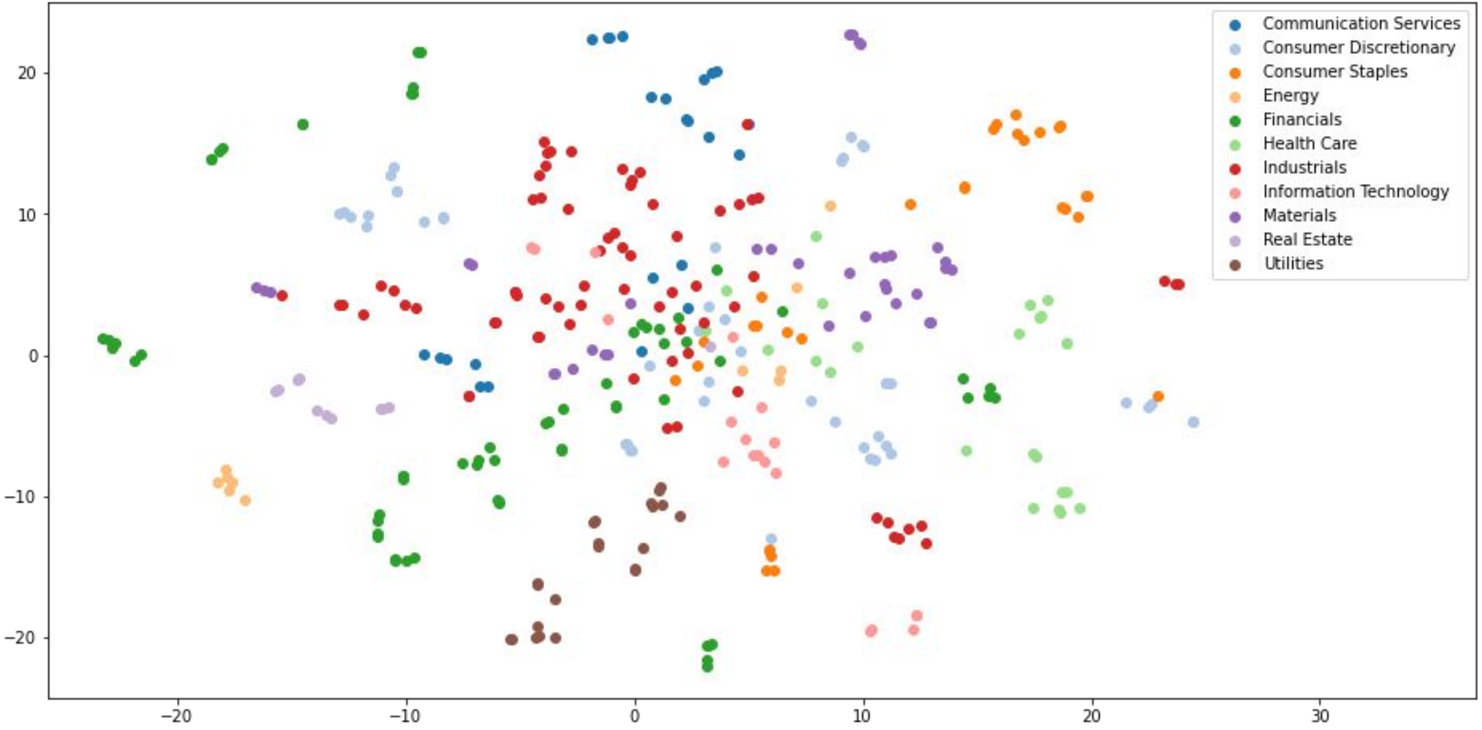
From a 2D space, it can calculate how close each samples should be between each other. It use this knowledge to produce a similar distribution in 1D space.

This can be generalized into N number of dimension.

https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a

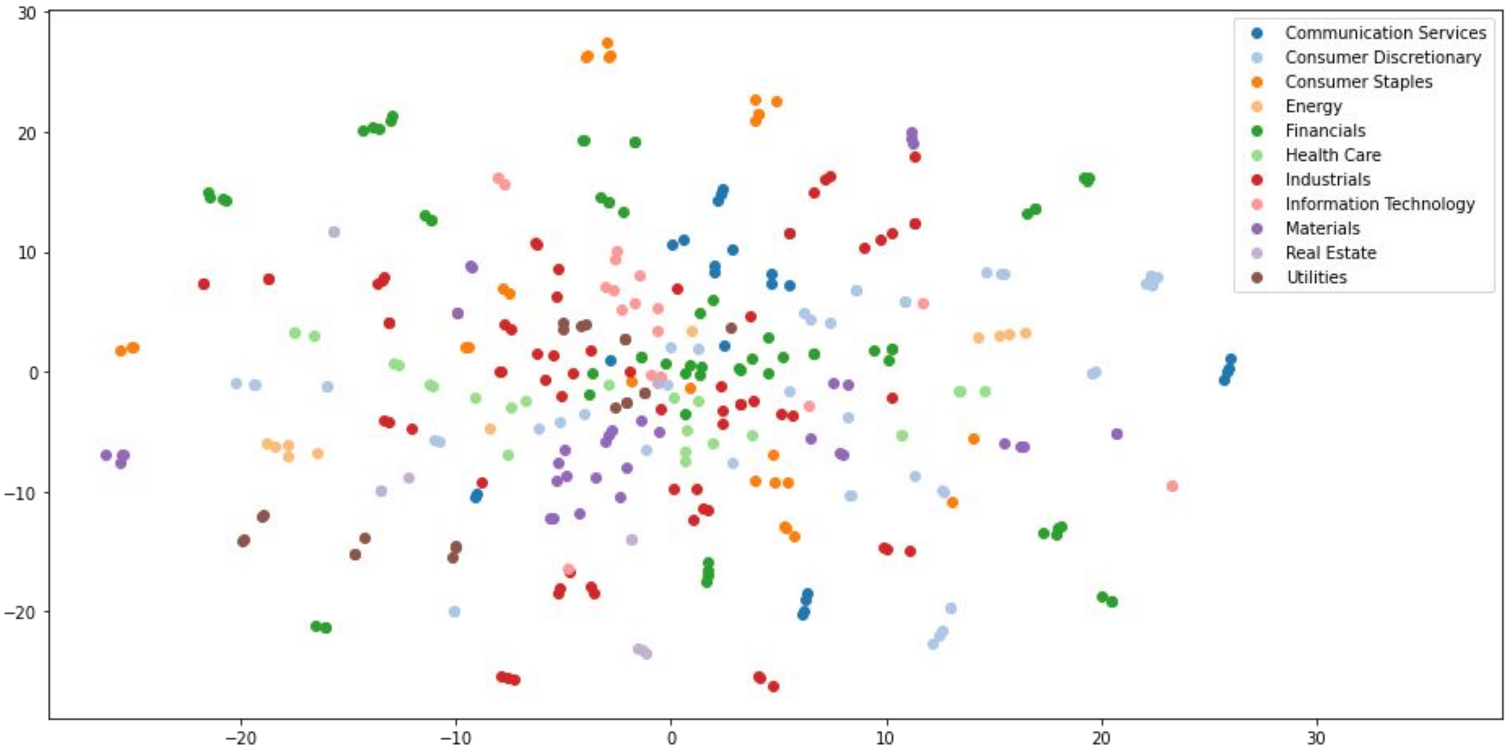
Part III

Q3a All Analyst



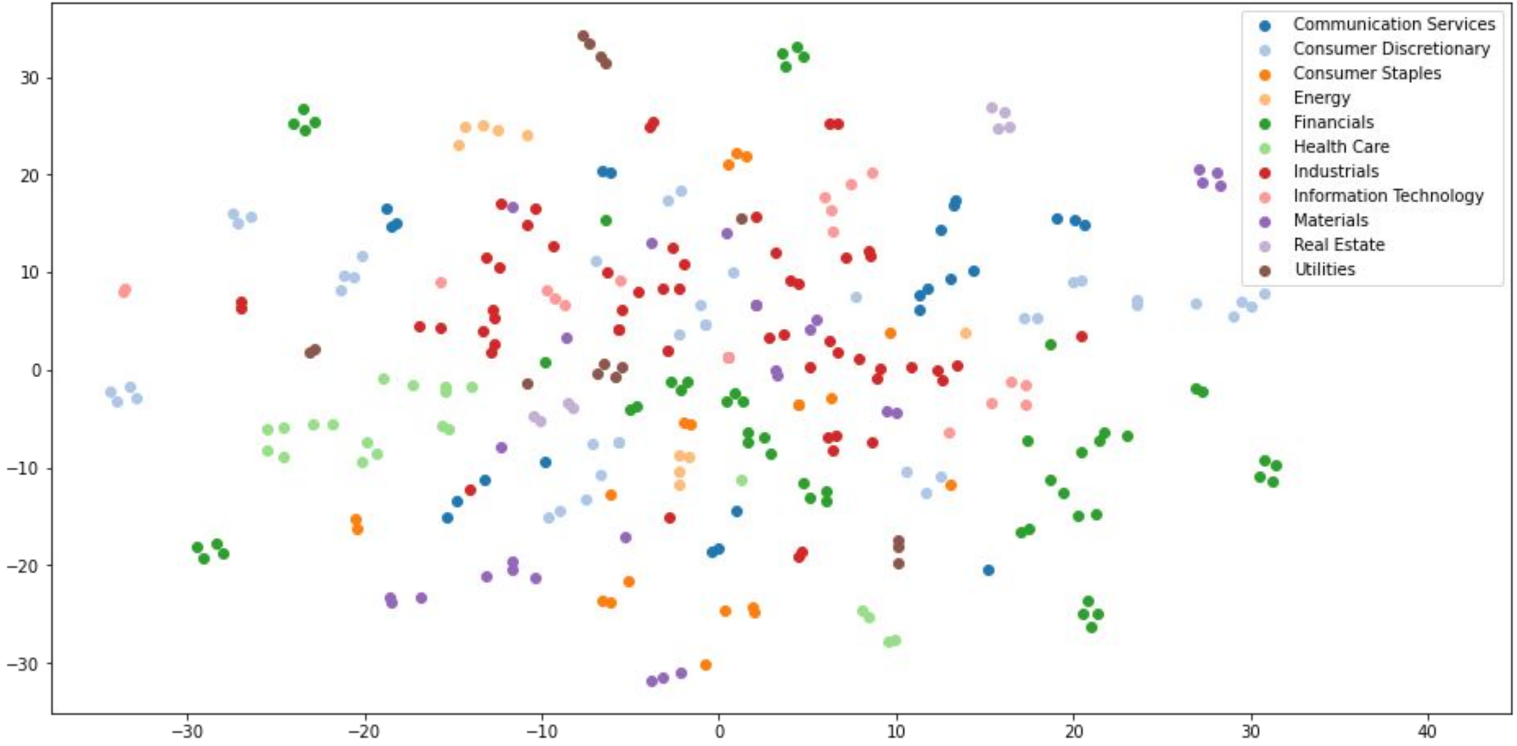
Part III

Q3b Winsorized and +- 1 S.D.



Part III

Q3c Analysts with 4 coverage



Part III

Q3+ TruncatedSVD

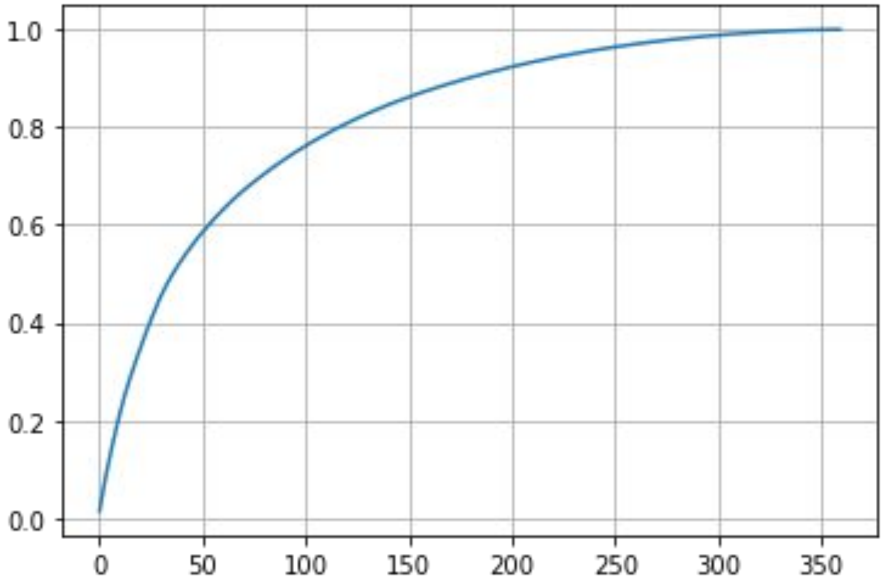
Instead of assuming that Analysts with coverage far from the mean provide less/noisy information, we can let the data speak for itself.

With all analyst, we have 2065 columns against 360 rows, which is somewhat irrational. Fortunately, its either 0 or 1 and therefore “sparse”. TruncatedSVD is specialized in decomposing sparse matrices and we will be using sklearn’s TruncatedSVD to decompose analysts information for us.

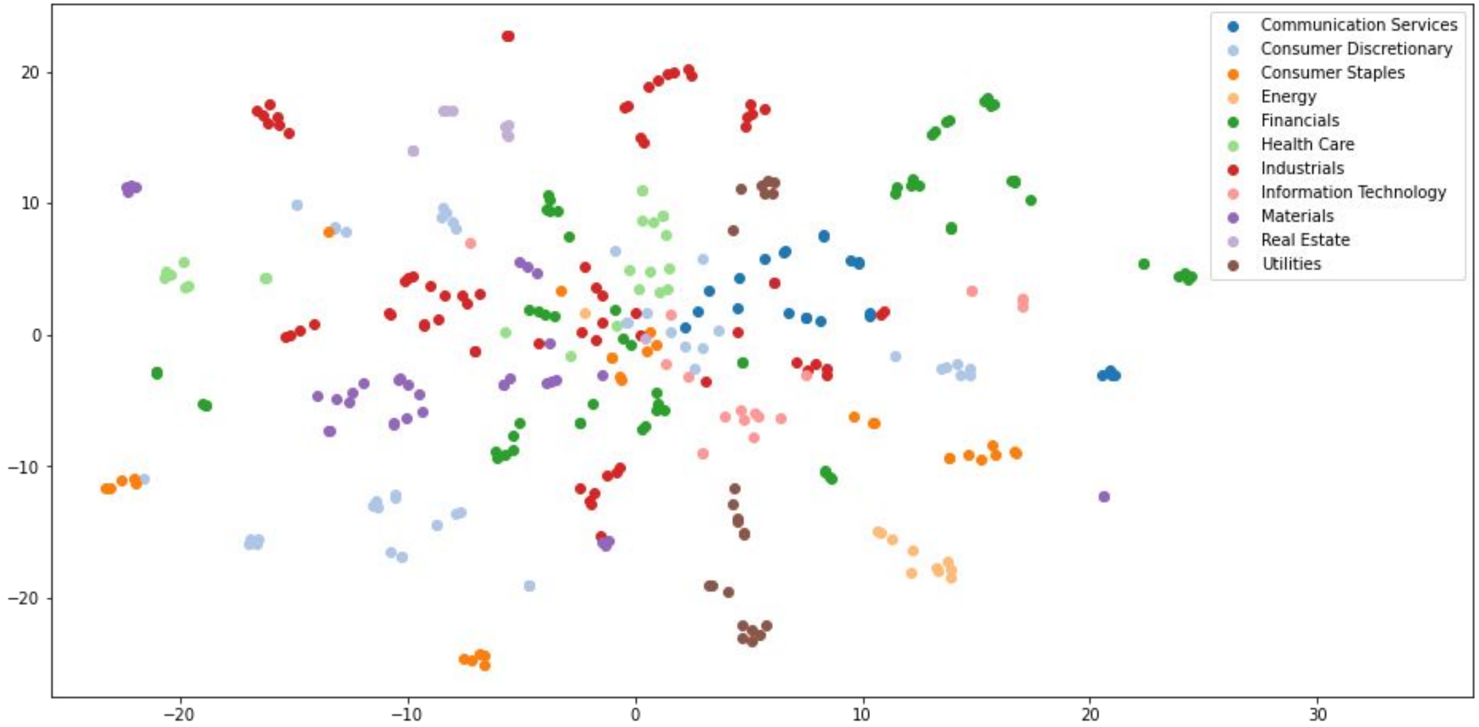
Part III

Explained Variance vs Components

At about 100 components, we will retain about 75% of all variance.



Part III TruncatedSVD, Select Top 100, then TSNE



Part III

Q4-6

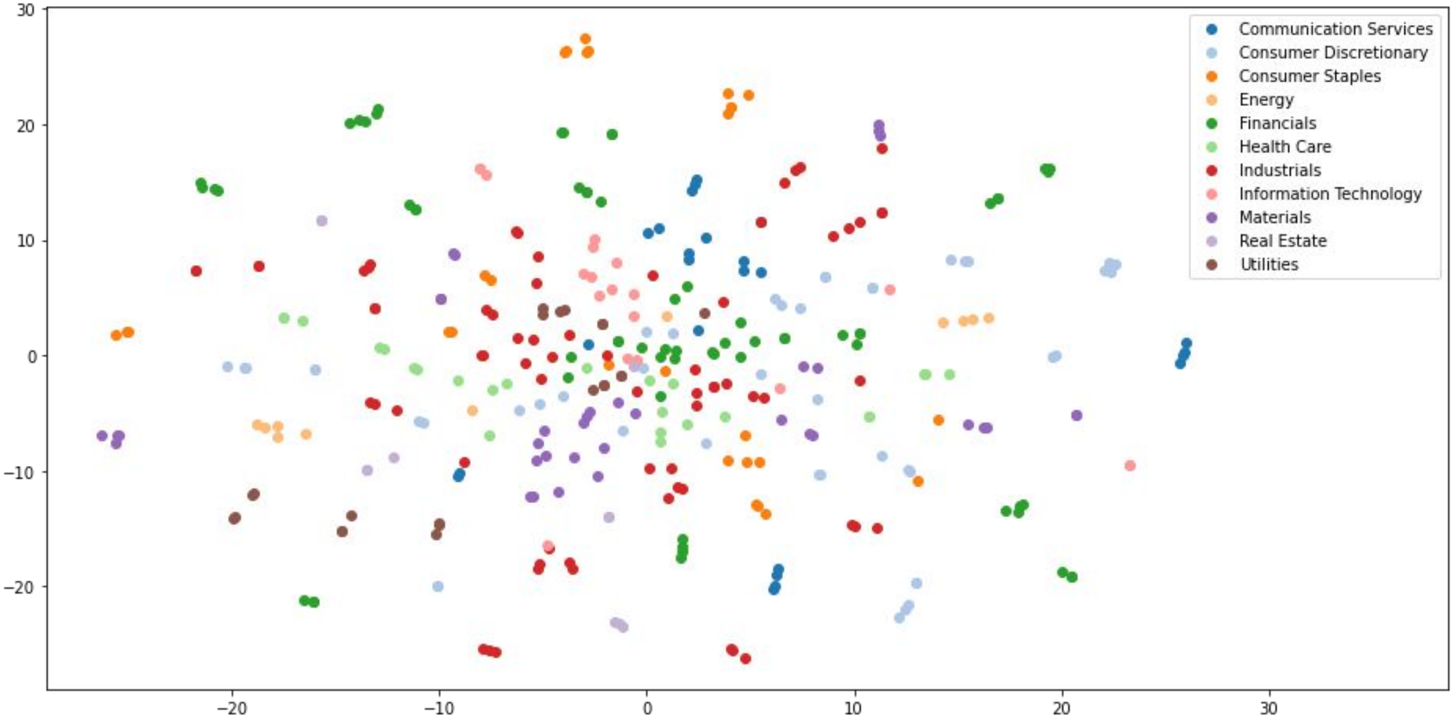
4. The most heterogenous are Financials and Industrials,

5. The most homogenous are Energy and Real Estate which almost purely form its own group. Although utilities sector is also a notable mention.

6. One way to determine 'outliers' would be to see which companies tends to have an ill-defined clusters. These could be the one that tends to 'join up' with other clusters or those that appears in the center of T-SNE plot which seems to be reserved for the companies that can't be clustered. Financials and Industrials are often the outlier everywhere, which makes sense as both tends to have their cashflow dictated by other sector rather than their own (Financial and Industrial companies would be exposed to the risk factor of their clients). Materials are also similar in this regard, but not as bad.

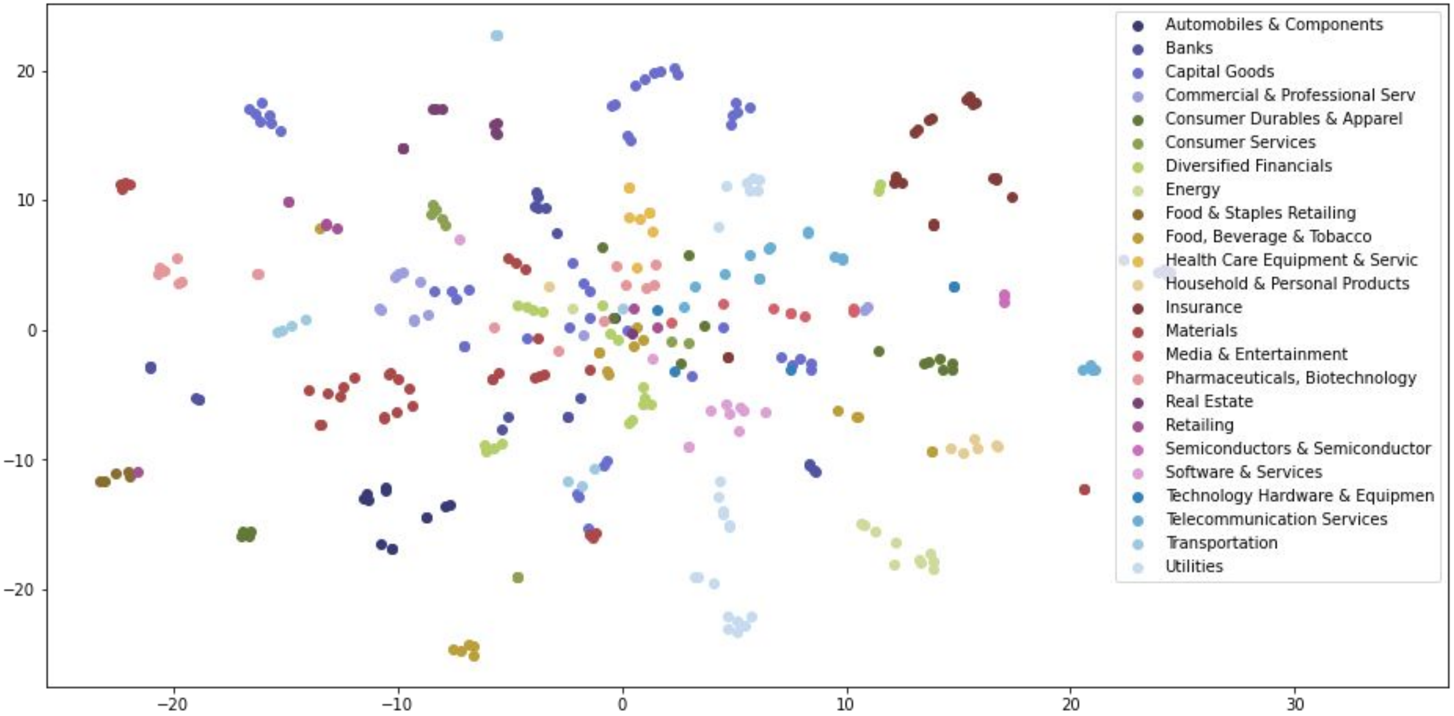
Part III

For comparison Winsorized and +- 1 S.D.



Part III

This makes sense in sub-industry too



Part III

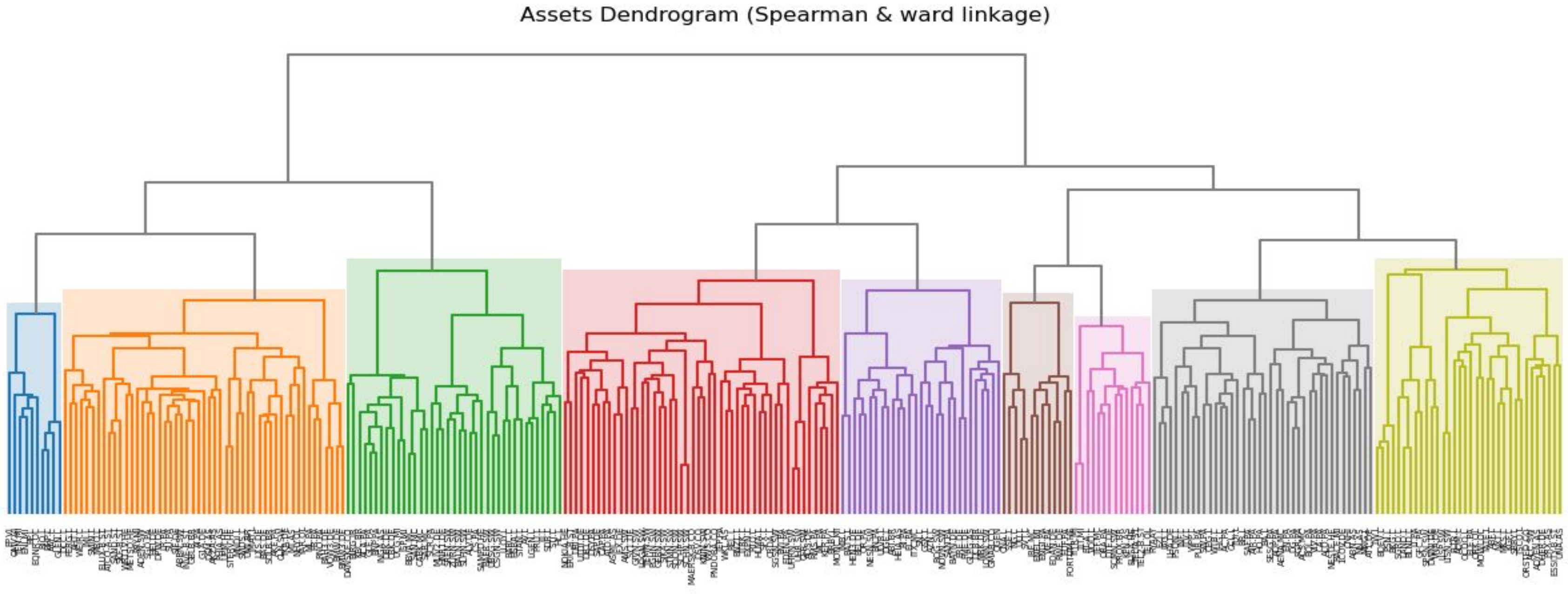
Additional Sections

Portfolio Optimization using HRP and HERC

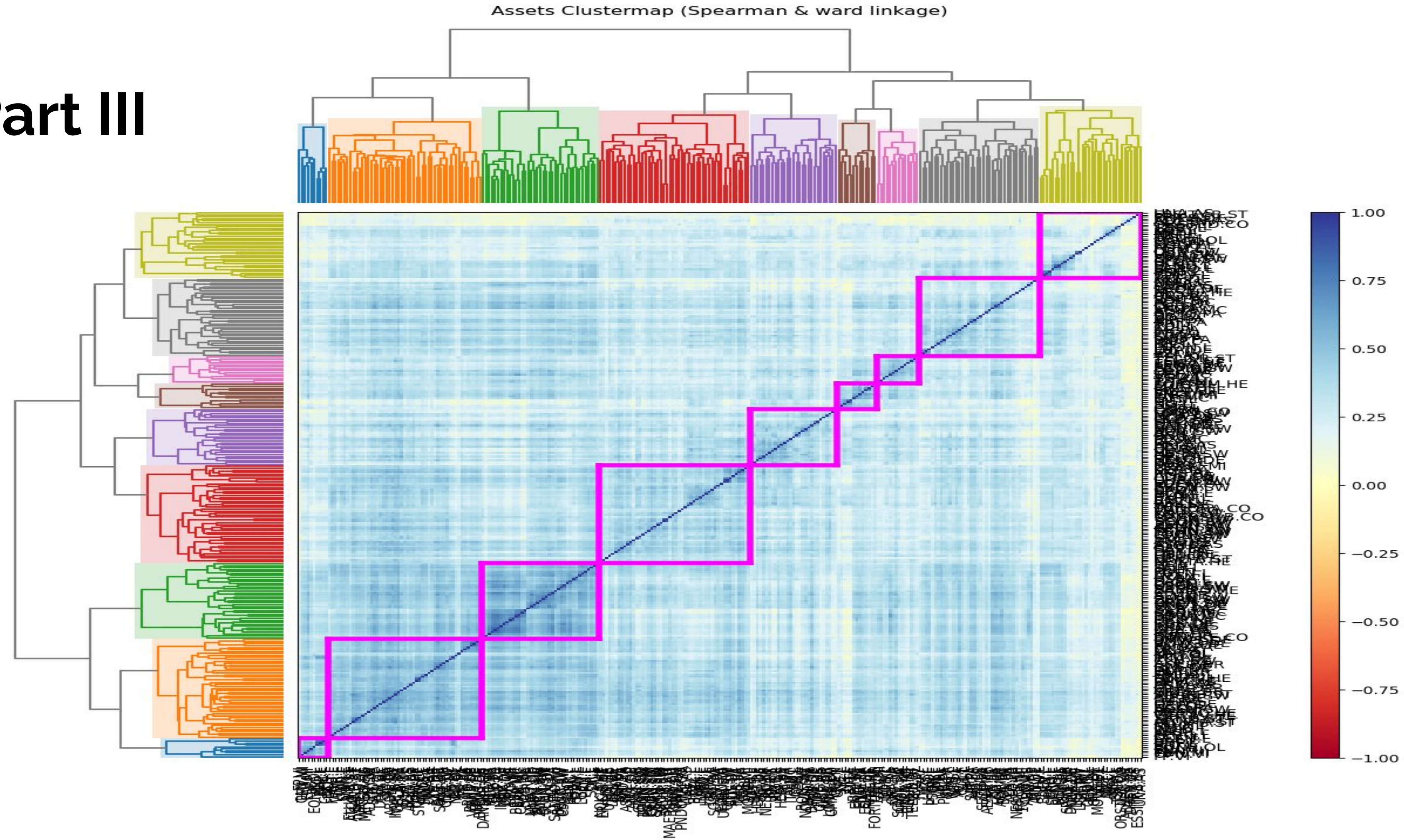
- Hierarchical Tree Clustering
- HRC vs HERC

Part III

Hierarchical Tree Clustering

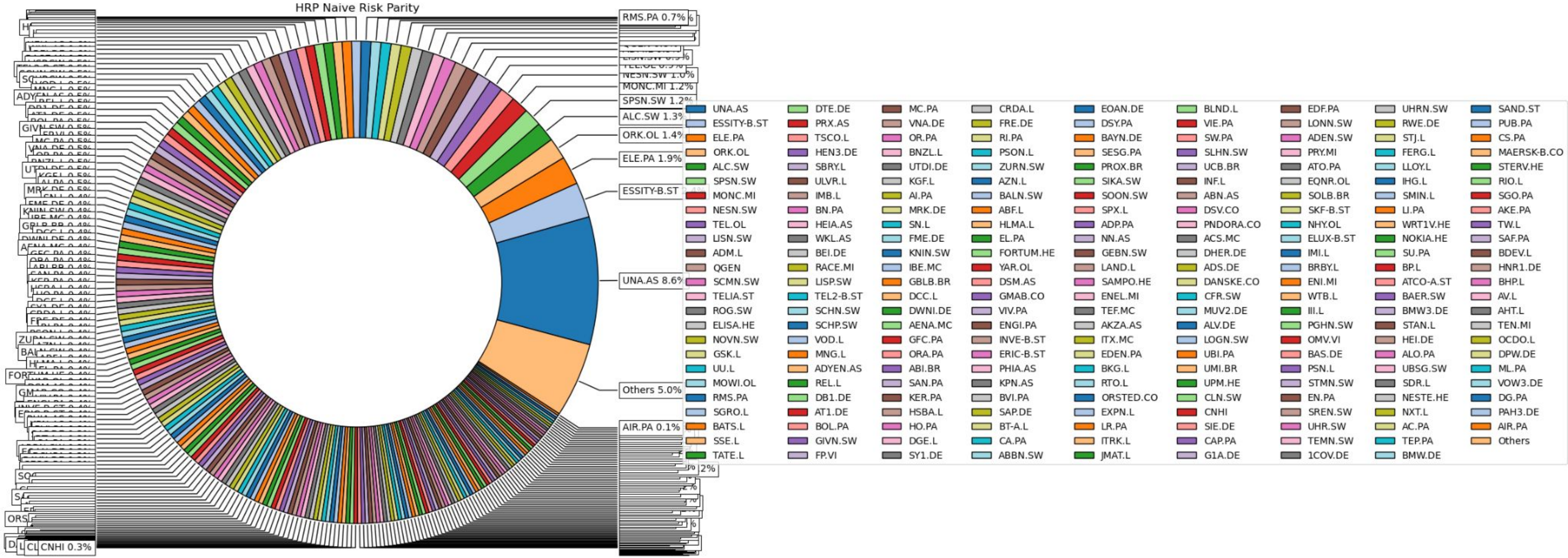


Part III


Assets Clustermap (Spearman & ward linkage)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Part III

HRP (Hierarchical Risk Parity)



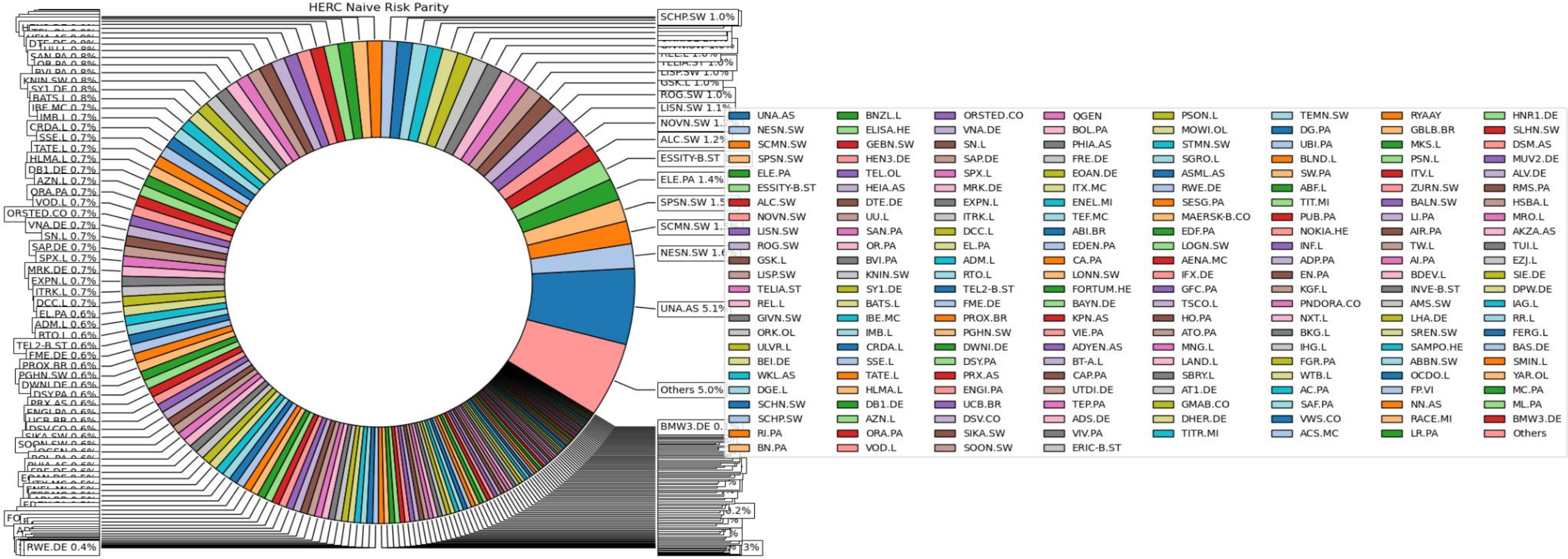
Part III

HRP Weights

	UNA.AS	ESSITY-B.ST	ELE.PA	ORK.OL	ALC.SW	SPSN.SW	MONC.MI	NESN.SW	TEL.OL	LISN.SW
weights	8.55%	2.36%	1.90%	1.44%	1.32%	1.22%	1.16%	0.99%	0.93%	0.93%

	mean	std	min	25%	50%	75%	max
weights	0.36%	0.56%	0.00%	0.18%	0.27%	0.40%	8.55%

Part III HERC(Hierarchical Equal Risk Contribution)



Part III

HERC weights

	UNA.AS	NESN.SW	SCMN.SW	SPSN.SW	ELE.PA	ESSITY-B.ST	ALC.SW	NOVN.SW	LISN.SW	ROG.SW
weights	5.09%	1.63%	1.53%	1.49%	1.43%	1.41%	1.21%	1.19%	1.13%	1.04%

	mean	std	min	25%	50%	75%	max
weights	0.36%	0.45%	0.00%	0.07%	0.18%	0.57%	5.09%

The End!