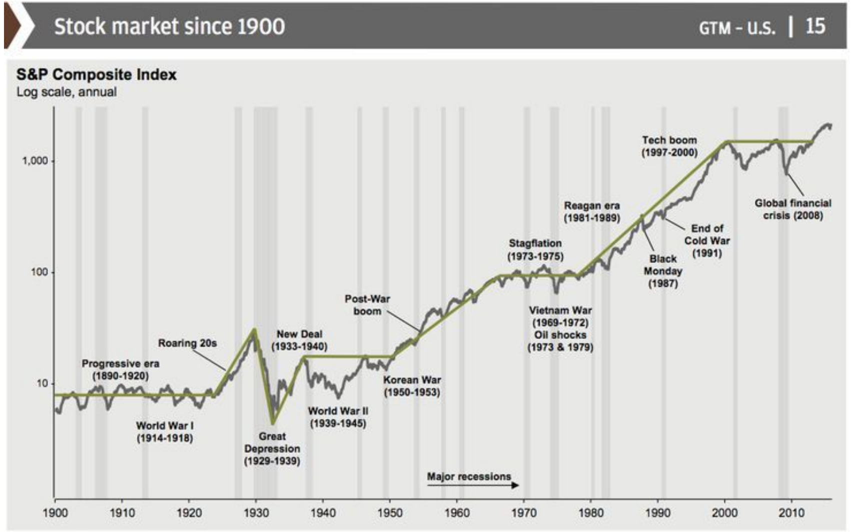
Clustering Fundamental Indicators of S&p 500 Listed Companies

Andrew Koller
Regis MSDS
03.06.2024

Introduction



Stock market since 1900

GTM – U.S. | 15

S&P Composite Index

Log scale, annual

Source: FactSet, NBER, Robert Shiller, J.P. Morgan Asset Management.
Data shown in log scale to best illustrate long-term index patterns.

The purpose of investing in the US equity market is to make money.

The best way to do this is by ‘buying low and selling high’..

Timing the market like this is exceedingly difficult.

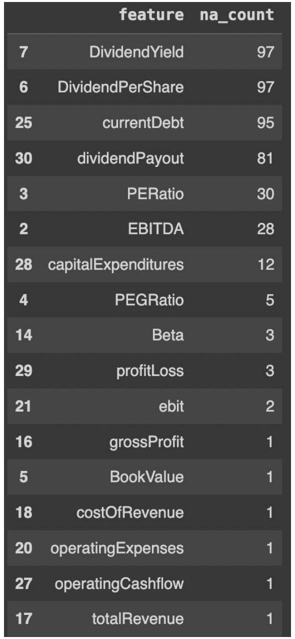
An alternative approach is to not time the market, but simply invest in well run companies.

Hypothesis

In the long run, well-run companies do well & you can use a companies fundamental data to identify well-run companies. **If well run companies do better in the long run, clustering on fundamental indicators should yield clusters of companies that have better or worse returns.**

Data

The data for this project includes income, cashflow, and balance sheet information for all S&P 500 companies for the most recent fiscal quarter. The data for this project was downloaded using a paid tier of the Alpha Vantage Python API. There was an attempt to impute any missing data with values from yfinance, and features missing from both sources were dropped from the analysis.



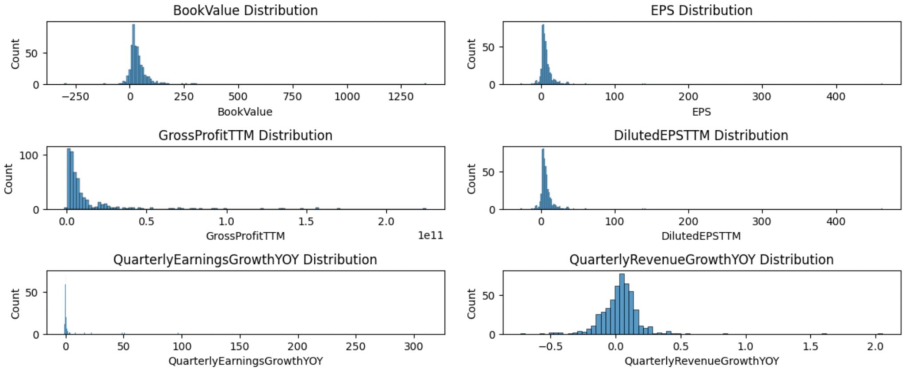
	feature	na_count
7	DividendYield	97
6	DividendPerShare	97
25	currentDebt	95
30	dividendPayout	81
3	PERatio	30
2	EBITDA	28
28	capitalExpenditures	12
4	PEGRatio	5
14	Beta	3
29	profitLoss	3
21	ebit	2
16	grossProfit	1
5	BookValue	1
18	costOfRevenue	1
20	operatingExpenses	1
27	operatingCashflow	1
17	totalRevenue	1

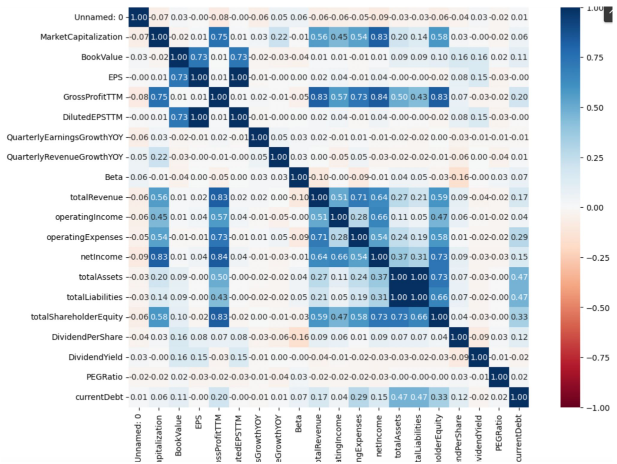
EDA & Feature Engineering

- The distributions of each feature were checked to ensure enough variation to draw clusters

- The correlation matrix shows that there are not many redundancies in the data

- All data was scaled between 0 and 1 using a min/max scaler

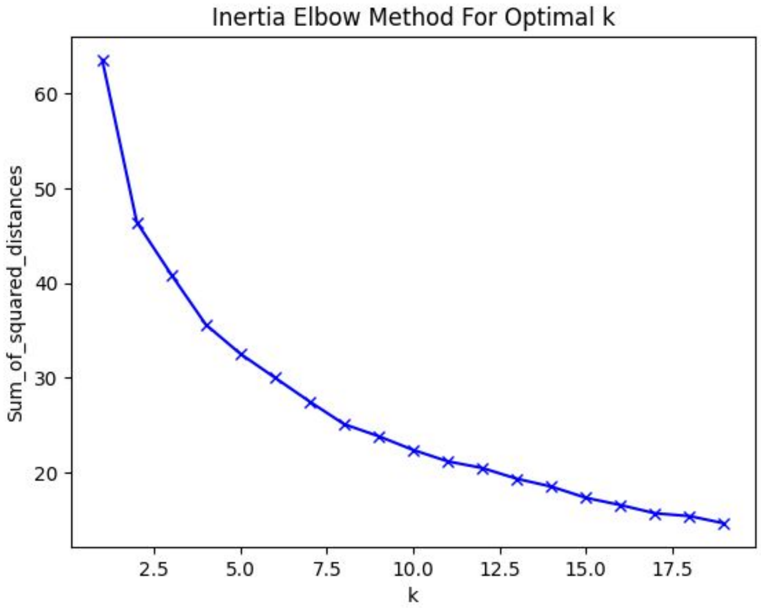




	Unnamed: 0	apitalization	BookValue	EPS	ssProfitTTM	itedEPSTTM	sGrowthYOY	eGrowthYOY	Beta	otalRevenue	atingIncome	ngExpenses	netIncome	totalAssets	talLiabilities	holderEquity	ndPerShare	videndYield	PEGRatio	currentDebt
Unnamed: 0	1.00	-0.07	0.03	-0.00	-0.08	-0.00	-0.06	0.05	0.06	-0.06	-0.06	-0.05	-0.09	-0.03	-0.03	-0.06	-0.04	0.03	-0.02	0.01
MarketCapitalization	-0.07	1.00	-0.02	0.01	0.75	0.01	0.03	0.22	-0.01	0.56	0.45	0.54	0.83	0.20	0.14	0.58	0.03	-0.00	-0.02	0.06
BookValue	0.03	-0.02	1.00	0.73	0.01	0.73	-0.02	-0.03	-0.04	0.01	0.01	-0.01	0.01	0.09	0.09	0.10	0.16	0.16	0.02	0.11
EPS	-0.00	0.01	0.73	1.00	0.01	1.00	-0.01	-0.00	0.00	0.02	0.04	-0.01	0.04	-0.00	-0.00	-0.02	0.08	0.15	-0.03	-0.00
GrossProfitTTM	-0.08	0.75	0.01	0.01	1.00	0.01	0.02	-0.01	-0.05	0.83	0.57	0.73	0.84	0.50	0.43	0.83	0.07	-0.03	-0.02	0.20
DilutedEPSTTM	-0.00	0.01	0.73	1.00	0.01	1.00	-0.01	-0.00	0.00	0.02	0.04	-0.01	0.04	-0.00	-0.00	-0.02	0.08	0.15	-0.03	-0.00
QuarterlyEarningsGrowthYOY	-0.06	0.03	-0.02	-0.01	0.02	-0.01	1.00	0.05	0.03	0.02	-0.01	0.01	-0.01	-0.02	-0.02	0.00	-0.03	-0.01	-0.01	-0.01
QuarterlyRevenueGrowthYOY	0.05	0.22	-0.03	-0.00	-0.01	-0.00	0.05	1.00	0.03	0.00	-0.05	0.05	-0.03	-0.02	-0.02	-0.01	-0.06	0.00	-0.04	0.01
Beta	0.06	-0.01	-0.04	0.00	-0.05	0.00	0.03	0.03	1.00	-0.10	-0.00	-0.09	-0.01	0.04	0.05	-0.03	-0.16	-0.00	0.03	0.07
totalRevenue	-0.06	0.56	0.01	0.02	0.83	0.02	0.02	0.00	-0.10	1.00	0.51	0.71	0.64	0.27	0.21	0.59	0.09	-0.04	-0.02	0.17
operatingIncome	-0.06	0.45	0.01	0.04	0.57	0.04	-0.01	-0.05	-0.00	0.51	1.00	0.28	0.66	0.11	0.05	0.47	0.06	-0.01	-0.02	0.04
operatingExpenses	-0.05	0.54	-0.01	-0.01	0.73	-0.01	0.01	0.05	-0.09	0.71	0.28	1.00	0.54	0.24	0.19	0.58	0.01	-0.02	-0.02	0.29
netIncome	-0.09	0.83	0.01	0.04	0.84	0.04	-0.01	-0.03	-0.01	0.64	0.66	0.54	1.00	0.37	0.31	0.73	0.09	-0.03	-0.03	0.15
totalAssets	-0.03	0.20	0.09	-0.00	0.50	-0.00	-0.02	-0.02	0.04	0.27	0.11	0.24	0.37	1.00	1.00	0.73	0.07	-0.03	-0.00	0.47
totalLiabilities	-0.03	0.14	0.09	-0.00	0.43	-0.00	-0.02	-0.02	0.05	0.21	0.05	0.19	0.31	1.00	1.00	0.66	0.07	-0.02	-0.00	0.47
totalShareholderEquity	-0.06	0.58	0.10	-0.02	0.83	-0.02	0.00	-0.01	-0.03	0.59	0.47	0.58	0.73	0.73	0.66	1.00	0.04	-0.03	-0.00	0.33
DividendPerShare	-0.04	0.03	0.16	0.08	0.07	0.08	-0.03	-0.06	-0.16	0.09	0.06	0.01	0.09	0.07	0.07	0.04	1.00	-0.09	0.03	0.12
DividendYield	0.03	-0.00	0.16	0.15	-0.03	0.15	-0.01	0.00	-0.00	-0.04	-0.01	-0.02	-0.03	-0.03	-0.02	-0.03	-0.09	1.00	-0.01	-0.02
PEGRatio	-0.02	-0.02	0.02	-0.03	-0.02	-0.03	-0.01	-0.04	0.03	-0.02	-0.02	-0.02	-0.03	-0.00	-0.00	-0.00	0.03	-0.01	1.00	0.02
currentDebt	0.01	0.06	0.11	-0.00	0.20	-0.00	-0.01	0.01	0.07	0.17	0.04	0.29	0.15	0.47	0.47	0.33	0.12	-0.02	0.02	1.00

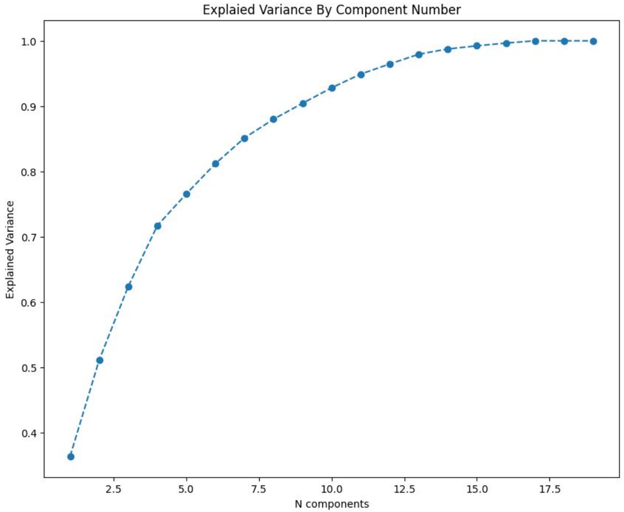
Optimal Number of Clusters

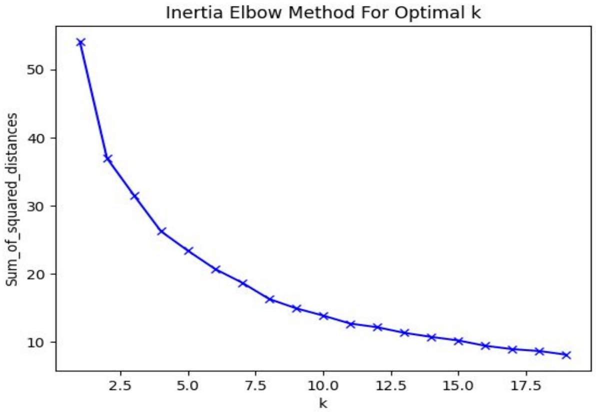
- The ‘Elbow Method’ was used to choose K
- Borth Distortion and Inertia were considered
- Both charts had ambiguous ‘elbows’ but 8 seemed like a good choice
- K = 8. Lower cluster numbers make little sense.



PCA + Elbow Method

- Distance measures in higher dimensional spaces become ambiguous
- Use PCA to reduce the dimensionality
- Choose appropriate number of components based on 80% of explained variance
  - Components = 7
- Elbow method with PCA yields 5 clusters





Clustering Results

Kmeans	N
0	220
1	33
2	8
3	61
4	168
5	4
6	1
7	5

kmeans_pca	N
0 192	192
1 219	219
2 9	9
3 6	6
4 74	74

Both models were successfully able to segment different equities off into different clusters. The Silhouette scores for the two models were .32 and .45 respectively. This suggests that there is a bit of overlap, and the clusters might not be very well defined, however doing PCA did yield more consistent clusters.

Cluster Returns

The returns for each cluster was calculated by finding the percent change in price for each each equity from Jan 2024 - Feb 2024. Each equity is assumed to carry an even weight, and the average percent change was calculated for each cluster.

The overall percent change for the S&P500 was ~5% over this time period

There are clusters which have better and worse returns than the overall index

Hypothesis is validated

Kmeans Cluster % Return

0	0.041109
1	0.036862
2	0.042866
3	0.055585
4	0.064135
5	0.079568
6	0.089605
7	0.028114

Kmeans w/ PCA Cluster % Returns

0	0.065213
1	0.042154
2	0.040002
3	0.055241
4	0.038837

Discussion

Kmeans is likely not the best clustering method for this type of analysis. Hierarchical clustering might be a better option.

The clustering DID work as intended and the hypothesis was validated.

The optimal number of clusters is a crap shoot. A combination of heuristics and industry knowledge is needed.