Assignment2

Lorenzo Ausiello

2023-10-17

PROBLEM 1

Introduction The purpose of this report is to analyze information about S&P 500 component stocks. The analysis aims to identify most relevant sectors, most frequent headquarter locations and years in which more stocks (still included) have been added.

Data Description The dataset includes columns such as Symbol, Security, GISC Sector, GICS Sub-Industry, Headquarters Location, Date added, CIK, Founded. It has been retrieved scraping Wikipedia website relative to the List of S&P 500 Companies.

Table 1: S&P500

Symbol	Security	GICS Sector	GICS Sub-Industry	Headquarters Location	Date Added	CIK	Founded
MMM	3M	Industrials	Industrial Conglomerates	Saint Paul, Minnesota	1957-03-04	0000066740	1902
AOS	A. O. Smith	Industrials	Building Products	Milwaukee, Wisconsin	2017-07-26	0000091142	1916
ABT	Abbott	Health Care	Health Care Equipment	North Chicago, Illinois	1957-03-04	0000001800	1888
ABBV	AbbVie	Health Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	0001551152	2013 (1888)
ACN	Accenture	Information Technology	IT Consulting & Other Services	Dublin, Ireland	2011-07-06	0001467373	1989
ADM	ADM	Consumer Staples	Agricultural Products & Services	Chicago, Illinois	1957-03-04	0000007084	1902

Data Cleaning The dates have been formatted correctly, and it has been giving to the variables useful format. The variable Headquarters Location has been divided in two different variables: Headquarters State and Headquarters City. This helped the analysis.

Data Analysis Below is the statistics and plots.

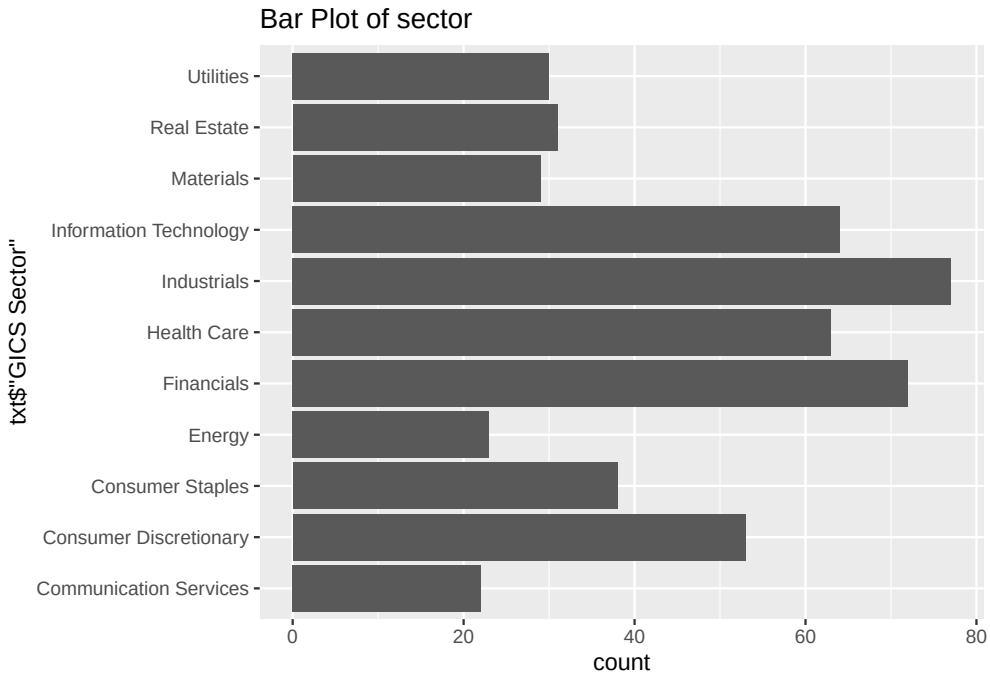
Table 2: Sectors S&P500

x
Communication Services
Consumer Discretionary
Consumer Staples
Energy
Financials
Health Care

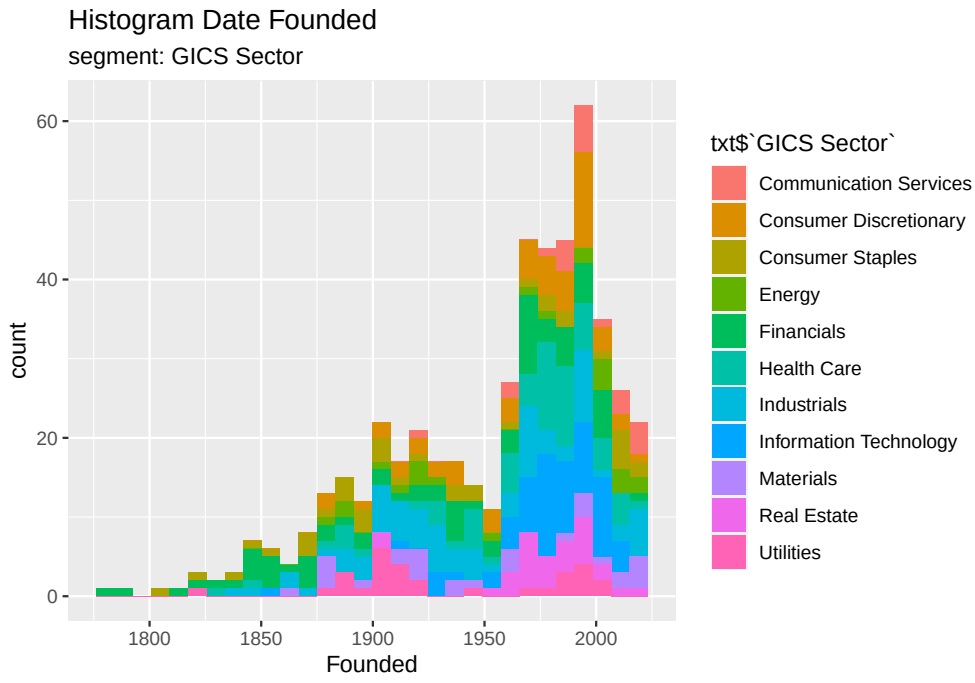
x
Industrials
Information Technology
Materials
Real Estate
Utilities

Table 3: Summary statistics S&P500

Symbol	Security	GICS Sector	GICS Sub-Industry	Headquarters Location	Date Added	CIK	Founded	Headquarters City	Headquarters State
Length:502	Length:502	Industrials : 77	Health Care Equipment : 19	Length:502	Min. :1957-03-04	Length:502	Min. :1784	New York City: 40	California : 68
Class :character	Class :character	Financials : 72	Semiconductors : 15	Class :character	1st Qu.:1989-05-30	Class :character	1st Qu.:1911	Houston : 20	New York : 52
Mode :character	Mode :character	Information Technology: 64	Industrial Machinery & Supplies & Components: 14	Mode :character	Median :2007-05-21	Mode :character	Median :1970	Chicago : 16	Texas : 46
NA	NA	Health Care : 63	Application Software : 13	NA	Mean :2000-06-25	NA	Mean :1955	Atlanta : 15	Illinois : 32
NA	NA	Consumer Discretionary: 53	Electric Utilities : 13	NA	3rd Qu.:2016-06-24	NA	3rd Qu.:1994	Dallas : 10	Massachusetts: 23
NA	NA	Consumer Staples : 38	Multi-Utilities : 13	NA	Max. :2023-10-18	NA	Max. :2023	Dublin : 10	Ohio : 20
NA	NA	(Other) :135	(Other) :415	NA	NA's :12	NA	NA	(Other) :391	(Other) :261




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

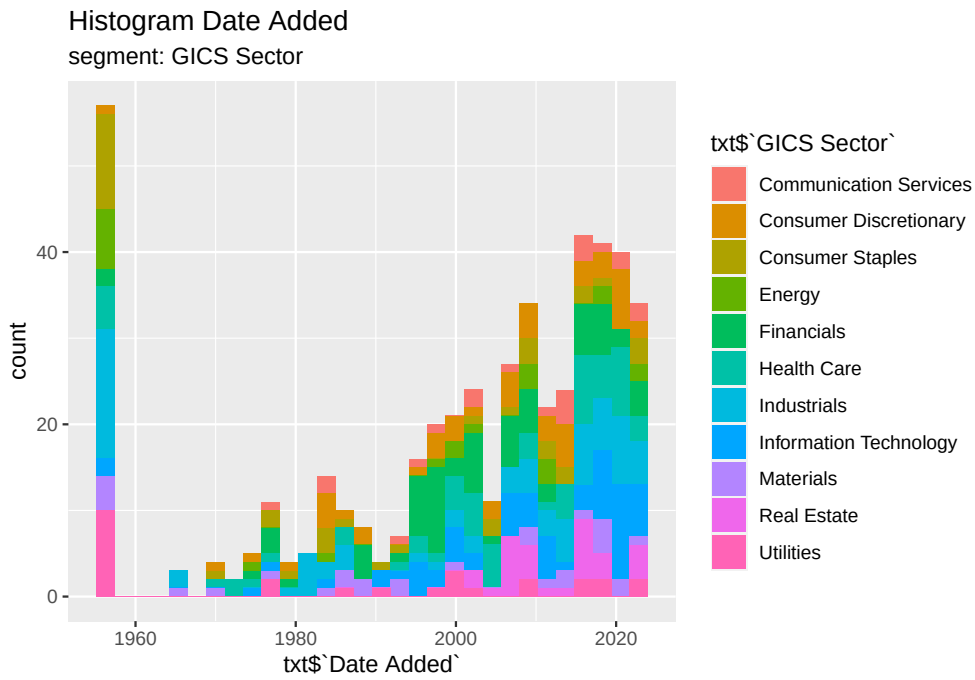


Table 4: Most frequent combinations: State and Sector

	State	Sector	Frequency
356	California	Information Technology	28
234	New York	Financials	20
194	Texas	Energy	16
256	California	Health Care	11
275	Massachusetts	Health Care	10
34	New York	Communication Services	10
456	California	Real Estate	7
344	Texas	Industrials	7
6	California	Communication Services	7
315	Illinois	Industrials	6

Summary statistics and Bar Plot show that S&P 500 stocks belong to 11 different sectors. The most frequent GICS Sector is represented by Industrials, followed by Financials, Health Care and IT.

The first stock added to the S&P 500, and still included, was added on 1957-03-04, when S&P 500 stock market index was introduced. The last stock has been added on 2023-10-02, and 50% of stocks have been added before 2007-05-2021.

New York City is the city where there are more S&P headquarters, and California is the state there are more headquarters.

Most frequent combinations of State and Sector are shown in the table and they may give some insights: California and IT is the more frequent combination (28), and thanks to the growth of Silicon Valley, almost the 50% of S&P IT Companies are based in this State. Then we have NYC and Financials (20), followed by Texas and Energy (16).

Finally the histograms show when most companies have been founded and added to the index based on sector. 'Histogram Date Added' show that many companies still included have been added when S&P was introduced in 1957. It implies that almost 16% of the index (almost 80 companies) never changed. At the

time the most relevant sectors were Industrials and Utilities. Most IT companies have been obviously founded after 1970 and added after 1990. Before 1900, as foundation years, we have mainly Financial companies, that has been added mainly after 1980.

PROBLEM 2

Introduction The purpose of this report is to analyze information about S&P 500 component stocks. The analysis aims to identify most relevant similarities and distances within a subset of 100 stocks, with the goal of identify possible clusters.

Data Description The datasets include columns such as GISC Sector, GICS Sub-Industry, Headquarters Location as categorical data and 76 quantitative variable. It has been decided to select 10 quantitative variable considered most relevant: 'After Tax ROE', 'Cash Ratio', 'Current Ratio', 'Pre Tax Margin', 'Pre Tax ROE', 'Profit Margin', 'Quick Ratio', 'Total Assets', 'Total Liabilities' and 'Earnings Before Tax'. These quantitative data have been normalized from the beginning, because some variable could have had a greater impact than other on the similarity/distance indicators/functions.

```
##
## Caricamento pacchetto: 'dplyr'

## I seguenti oggetti sono mascherati da 'package:stats':
##
##     filter, lag

## I seguenti oggetti sono mascherati da 'package:base':
##
##     intersect, setdiff, setequal, union
```

Table 5: 100 Tickers of S&P500: Quantitative Data

	After.Tax.ROE	Cash.Ratio	Current.Ratio	Pre.Tax.Margin	Pre.Tax.ROE	Profit.Margin	Quick.Ratio	Total.Assets	Total.Liabilities	Earnings.Before.Tax
AAL	0.9964326	0.0359096	-0.6022602	-0.7146110	0.7709700	-0.5538632	-0.3932002	0.3868382	1.4689684	-0.7308349
AAP	0.0227973	-0.3028602	-0.2585534	-0.4967418	0.1290437	-0.6902827	-0.7747795	-0.5822532	-0.6339602	-0.2813987
AAPL	0.1177861	0.2101341	-0.0523294	1.5730156	0.1290437	1.4924294	0.1965134	4.7348810	3.4425995	7.6504586
ABBV	1.5901126	0.7037700	0.7296036	1.4640810	1.4128963	1.4924294	0.8816217	0.0415753	0.4266024	0.4721894
ABC	-0.1434331	-0.6125927	-0.6624089	-1.4771532	0.0467454	-1.5087997	-0.7921240	-0.2297613	0.0103796	-0.2496800
ABT	-0.3571579	0.1327010	0.2398214	-0.4967418	-0.4141248	0.2646538	0.2832359	0.4046557	0.0711209	-0.0548544

Table 6: 100 Tickers of S&P500: Categorical Data

	Ticker.symbol	Security	SEC.filings	GICS.Sector	GICS.Sub.Industry	Address.of.Headquarters	Date.first.added	CIK
15	AAL	American Airlines Group	reports	Industrials	Airlines	Fort Worth, Texas	2015-03-23	6201

	Ticker.symbol	Security	SEC.filings	GICS.Sector	GICS.Sub.Industry	Address.of.Headquarters	Date.first.added	CIK
6	AAP	Advance Auto Parts	reports	Consumer Discretionary	Automotive Retail	Roanoke, Virginia	2015-07-09	1158449
26	AAPL	Apple Inc.	reports	Information Technology	Computer Hardware	Cupertino, California	1982-11-30	320193
2	ABBV	AbbVie	reports	Health Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	1551152
19	ABC	AmerisourceBergen Corp	reports	Health Care	Health Care Distributors	Chesterbrook, Pennsylvania		1140859
1	ABT	Abbott Laboratories	reports	Health Care	Health Care Equipment	North Chicago, Illinois	1964-03-31	1800

Next, several distance and similarity functions have been defined and applied to find the extreme values for distance and similarities between the subset of tickers chosen. There are functions that allow to calculate the quantity required for all ticker pairs, and function that allow to calculate the top and bottom 10 values for each case.

First of all, *Lp-Norm function*, that is a distance function, has been defined and applied for p=1, p=2, P=3 and p=10. Below a subset of the resulting tables and the rank of the top and bottom 10 values.

Table 7: Lp Norm, Manhattan, p=1

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	0.000000	6.555363	21.87133	11.325790	7.245487	7.282026	10.347074	32.67331	6.906907	7.631324
AAP	6.555363	0.000000	23.36352	13.083871	3.806940	5.788546	6.063904	27.87461	5.646738	6.177404
AAPL	21.871328	23.363521	0.00000	19.713363	25.113360	20.178969	25.519339	39.24853	22.987418	22.573500
ABBV	11.325790	13.083871	19.71336	0.000000	14.833710	9.867733	11.318335	25.26539	13.235515	7.041974
ABC	7.245487	3.806940	25.11336	14.833710	0.000000	7.041328	8.901693	31.08085	3.792372	8.533572
ABT	7.282026	5.788546	20.17897	9.867733	7.041328	0.000000	5.495237	26.30574	4.026706	6.173710

Table 8: Lp Norm, Euclidean, p=2

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	0.000000	2.714416	10.214301	4.024501	2.576755	2.746156	3.856754	13.32985	2.579283	2.807610
AAP	2.714416	0.000000	10.867882	4.419173	1.570142	2.105051	2.794686	13.20923	2.028555	2.386097
AAPL	10.214301	10.867882	0.000000	9.370340	10.920274	9.792873	10.940275	15.69929	10.260899	10.381229
ABBV	4.024501	4.419173	9.370340	0.000000	5.456160	3.733921	4.339547	11.02356	5.158058	2.570646
ABC	2.576755	1.570142	10.920274	5.456160	0.000000	2.708216	3.444477	14.01716	1.354333	3.360033
ABT	2.746156	2.105051	9.792873	3.733921	2.708216	0.000000	2.091851	11.99441	2.002360	2.317160

Table 9: Lp Norm, p=3

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	0.000000	2.266049	8.881163	3.001143	1.909322	2.052841	2.860022	10.546372	1.931968	2.127955
AAP	2.266049	0.000000	9.034540	3.184606	1.239179	1.560499	2.294700	10.684586	1.503733	1.838483
AAPL	8.881163	9.034540	0.000000	7.973464	8.933848	8.377757	9.035994	11.818710	8.510932	8.806117

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
ABBV	3.001143	3.184606	7.973464	0.000000	4.128577	2.869574	3.280491	8.904597	3.981124	1.935012
ABC	1.909322	1.239179	8.933848	4.128577	0.000000	2.101640	2.657151	11.156271	1.007056	2.554758
ABT	2.052841	1.560499	8.377757	2.869574	2.101640	0.000000	1.577682	9.711093	1.744119	1.783089

Table 10: Lp Norm, p=10

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	0.000000	2.1031161	8.382479	2.276222	1.4728733	1.492784	2.139462	8.055548	1.4511776	1.683360
AAP	2.103116	0.0000000	7.947279	2.265122	0.9967509	1.126763	1.893754	8.222881	1.0904061	1.544860
AAPL	8.382479	7.9472793	0.000000	7.188573	7.9079889	7.707930	7.986948	8.329901	7.7104982	7.915553
ABBV	2.276222	2.2651217	7.188573	0.000000	3.1873718	2.143371	2.337112	6.914708	3.0597784	1.512062
ABC	1.472873	0.9967509	7.907989	3.187372	0.0000000	1.775352	2.046255	8.487691	0.7540072	1.903046
ABT	1.492784	1.1267633	7.707930	2.143371	1.7753517	0.000000	1.200396	7.527915	1.6373370	1.417620

Table 11: Bottom 10 values,Lp Norm, Manhattan

X1	X2	X3
ADI	CTL	32.73909
ADI	DUK	32.90937
ADI	CAT	33.03178
ADI	CHTR	34.02778
ADI	BA	35.68150
CLX	CVX	36.49639
AAPL	ADI	39.24853
AAPL	CLX	39.67247
ADI	CVX	42.50433
ADI	CLX	44.39724

Table 12: Top 10 values,Lp Norm, Manhattan

X1	X2	X3
CMS	ECL	0.6600276
AVY	CBG	0.8129164
CAG	ECL	1.0497356
APH	DOV	1.0505372
APH	CHD	1.0813943
APD	CMS	1.0943455
CAG	CMS	1.1268845
COG	EFX	1.1350945
AME	EFX	1.1662549
DNB	EFX	1.1995076

Table 13: Bottom 10 values,Lp Norm, Euclidean

X1	X2	X3
ADI	CTL	13.99630
ABC	ADI	14.01716
ADI	CCL	14.14445
ADI	BA	14.24712
ADI	CHTR	14.65384
AAPL	ADI	15.69929
ADI	CVX	16.08744
CLX	CVX	16.34844
AAPL	CLX	16.34906
ADI	CLX	18.34758

Table 14: Top 10 values,Lp Norm, Euclidean

X1	X2	X3
CMS	ECL	0.2871981
AVY	CBG	0.3144083
APH	CHD	0.4182369
APH	DOV	0.4412592
CAG	ECL	0.4475628
DNB	EFX	0.4946790
APD	CMS	0.5076617
CAG	CMS	0.5239675
AAP	BLL	0.5254082
AME	DGX	0.5405754

Table 15: Bottom 10 values,Lp Norm, p=3

X1	X2	X3
AMZN	CLX	11.84822
CCI	CLX	11.86882
BSX	CLX	11.87037
ADI	CVX	11.87243
CLX	DUK	11.90639
CLX	CTL	11.92395
CLX	DVN	11.93137
AAPL	CLX	12.74570
CLX	CVX	12.81818
ADI	CLX	13.98487

Table 16: Top 10 values,Lp Norm, p=3

X1	X2	X3
CMS	ECL	0.2368027
AVY	CBG	0.2447028
APH	CHD	0.3295663

X1	X2	X3
APH	DOV	0.3612358
CAG	ECL	0.3671762
DNB	EFX	0.3851992
AAP	BLL	0.3954382
AME	DGX	0.4031769
AAP	DG	0.4266672
COG	DNB	0.4299540

Table 17: Bottom 10 values,Lp Norm, p=10

X1	X2	X3
APC	CLX	10.00764
CLX	EA	10.04739
AMAT	CLX	10.05750
ADBE	CLX	10.05750
CCL	CLX	10.06767
AMZN	CLX	10.06836
BSX	CLX	10.09939
CCI	CLX	10.10041
CLX	CTL	10.13043
CLX	DVN	10.14135

Table 18: Top 10 values,Lp Norm, p=10

X1	X2	X3
AVY	CBG	0.2066055
CMS	ECL	0.2178866
APH	CHD	0.2752558
DNB	EFX	0.2848702
AME	DGX	0.3107451
CAG	ECL	0.3215539
APH	DOV	0.3269003
AAP	BLL	0.3271504
BBBY	DLTR	0.3293579
APD	CXO	0.3303501

The diagonals of the resulting Lp_norm tables are equal to 0, because each stock is identical to itself. Lp-Norm, in fact, takes values starting from 0 and it takes on higher values as distance between the pair of observations increases. Therefore, the bottom 10 values tables show the most different pairs of stock based on the 10 quantitative variables, and the top 10 values tables show the most similar ones. Lp-Norm show different ranks for different values of p. In fact, higer values of p implies to using more and emphasizing the dimensions where the two objects are the most dissimilar.

The second distance function computed is *Minkoswki distance*. This function allows to give more weight to the variables considered more relevant in the analysis. In this case, most relevant variables are: 'Current Ratio', 'Pre Tax Margin' and 'Earning Before Tax'. Below a subset of the resulting tables and the rank of the top and bottom 10 values.

Table 19: Minkowski, p=1

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	0.000000	3.020851	13.28070	7.487379	4.070610	4.399202	5.446039	19.80374	4.545389	4.051730
AAP	3.020851	0.000000	12.47786	8.291985	2.390587	2.942232	3.635577	17.68572	3.042343	3.746931
AAPL	13.280704	12.477862	0.00000	10.579157	14.418967	11.521397	14.240681	21.69426	13.623285	12.398326
ABBV	7.487379	8.291985	10.57916	0.000000	10.233090	6.961277	7.701556	15.35122	9.398601	4.639910
ABC	4.070610	2.390587	14.41897	10.233090	0.000000	4.552510	5.448411	19.68845	2.129026	5.747514
ABT	4.399202	2.942232	11.52140	6.961277	4.552510	0.000000	2.827691	16.11698	2.621779	4.006869

Table 20: Minkowski, p=2

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	0.000000	1.547323	8.340481	3.377762	1.835095	2.073591	2.676662	10.401954	2.0524442	1.914017
AAP	1.547323	0.000000	8.185706	3.569665	1.288824	1.417093	2.062879	10.293984	1.4064434	1.797472
AAPL	8.340481	8.185706	0.000000	7.140733	8.543965	7.675564	8.308084	11.993095	8.2029710	7.891125
ABBV	3.377762	3.569665	7.140733	0.000000	4.640011	3.248223	3.682885	8.639914	4.4536703	2.133771
ABC	1.835095	1.288824	8.543965	4.640011	0.000000	2.191000	2.661182	11.071863	0.9993356	2.793317
ABT	2.073591	1.417093	7.675564	3.248223	2.191000	0.000000	1.488952	9.428005	1.6701727	1.951239

Table 21: Minkowski, p=3

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	0.000000	1.430628	7.971538	2.715423	1.457432	1.670564	2.168675	8.951226	1.6356674	1.579534
AAP	1.430628	0.000000	7.710889	2.784282	1.110815	1.164977	1.825416	8.986796	1.1385220	1.459381
AAPL	7.971538	7.710889	0.000000	6.884795	7.787892	7.364355	7.746009	10.055473	7.5352593	7.586656
ABBV	2.715423	2.784282	6.884795	0.000000	3.743440	2.635004	2.962416	7.593449	3.6389831	1.723998
ABC	1.457432	1.110815	7.787892	3.743440	0.000000	1.833406	2.213911	9.485353	0.8225457	2.267090
ABT	1.670564	1.164977	7.364355	2.635004	1.833406	0.000000	1.272662	8.273049	1.5552712	1.630003

Table 22: Minkowski, p=10

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	0.000000	1.7905932	8.196634	2.233744	1.2723677	1.360616	1.842754	7.695457	1.3623756	1.446011
AAP	1.790593	0.0000000	7.760621	2.193600	0.9821901	1.014362	1.722868	7.777270	0.9599040	1.371645
AAPL	8.196634	7.7606206	0.000000	7.022396	7.7278099	7.535920	7.802053	8.076467	7.5386101	7.733231
ABBV	2.233744	2.1936004	7.022396	0.000000	3.1125386	2.087059	2.263828	6.630940	2.9887464	1.460563
ABC	1.272368	0.9821901	7.727810	3.112539	0.0000000	1.712647	1.925841	8.086075	0.7213086	1.860551
ABT	1.360616	1.0143620	7.535920	2.087059	1.7126471	0.000000	1.153786	7.206499	1.5800467	1.408962

Table 23: Bottom 10 values,Minkowski, p=1

X1	X2	X3
ALXN	CLX	20.29127
CF	CLX	20.75842
CLX	CSCO	20.91130

X1	X2	X3
CLX	CVX	21.10321
ADI	CHTR	21.27829
AMGN	CLX	21.34217
AAPL	ADI	21.69426
ADI	CVX	22.76563
AAPL	CLX	24.89714
ADI	CLX	30.21178

Table 24: Top 10 values,Minkowski, p=1

X1	X2	X3
CMS	ECL	0.4254782
AVY	CBG	0.4991666
CAG	ECL	0.5717725
APH	CHD	0.6918614
DVA	ECL	0.6993215
AMZN	DVN	0.7011795
APH	DOV	0.7246682
CAG	CMS	0.7290254
APD	CMS	0.7414319
AEP	ED	0.7981924

Table 25: Bottom 10 values,Minkowski, p=2

X1	X2	X3
CLX	CSCO	11.75066
CLX	CTL	11.80072
CLX	DVN	11.81092
ALXN	CLX	11.82092
AKAM	CLX	11.85889
CLX	EBAY	11.89253
AAPL	ADI	11.99310
CLX	CVX	12.82906
AAPL	CLX	13.42533
ADI	CLX	15.20404

Table 26: Top 10 values,Minkowski, p=2

X1	X2	X3
CMS	ECL	0.2447596
AVY	CBG	0.2570698
CAG	ECL	0.3125985
APH	CHD	0.3442538
APH	DOV	0.3909833
DVA	ECL	0.4007898
APD	CMS	0.4187339
AAP	DG	0.4234347

X1	X2	X3
DNB	EFX	0.4269799
CRM	DRI	0.4316545

Table 27: Bottom 10 values,Minkowski, p=3

X1	X2	X3
CCL	CLX	10.80805
ADBE	CLX	10.81008
AMZN	CLX	10.81609
CCI	CLX	10.83281
BSX	CLX	10.84274
CLX	CTL	10.88462
CLX	DVN	10.89285
CLX	CVX	11.10873
AAPL	CLX	11.38592
ADI	CLX	12.40253

Table 28: Top 10 values,Minkowski, p=3

X1	X2	X3
AVY	CBG	0.2192860
CMS	ECL	0.2201962
CAG	ECL	0.2799438
APH	CHD	0.2928613
APH	DOV	0.3421739
AAP	DG	0.3489072
DNB	EFX	0.3499998
CRM	DRI	0.3522244
DVA	ECL	0.3528893
AAP	BLL	0.3574387

Table 29: Bottom 10 values,Minkowski, p=10

X1	X2	X3
CHK	CLX	9.768883
CLX	EA	9.812927
AMAT	CLX	9.825408
ADBE	CLX	9.825408
AMZN	CLX	9.832125
CCL	CLX	9.837941
CCI	CLX	9.858203
BSX	CLX	9.863766
CLX	CTL	9.895420
CLX	DVN	9.902173

Table 30: Top 10 values,Minkowski, p=10

X1	X2	X3
AVY	CBG	0.2019973
CMS	ECL	0.2155927
APH	CHD	0.2659223
DNB	EFX	0.2755594
CAG	ECL	0.2862246
AME	DGX	0.3028599
CCI	CNP	0.3072265
AAP	DG	0.3106725
CRM	DRI	0.3146820
DVA	ECL	0.3201024

Also in this case the diagonals of the resulting Lp_norm tables are equal to 0, because each stock is identical to itself. Minkowski distance takes values starting from 0 and it takes on higher values as distance between the pair of observations increases. Therefore, the bottom 10 values tables show the most different pairs of stock based on the 10 quantitative variables, and the top 10 values tables show the most similar ones. Minkowski distance show different ranks for different values of p as well.

Afterwards, dividing each variable in 4 equi-depth buckets, *Match-Based Similarity Computation* has been computed. Below a subset of the resulting table and the rank of the top and bottom 10 values.

Table 31: Matched Based Similarity Computation: 4 equi-depth buckets

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	10.00000002.712740	2.5886332	2.613589	1.8379008	0.9968404	1.0613657	0.0000000	2.8321144	1.951915	
AAP	2.7127399	10.000000	2.6909642	1.678795	2.4015482	1.2795031	2.4917979	0.4671445	1.3291925	2.347504
AAPL	2.5886332	2.690964	10.00000004.922933	0.9909379	1.7990143	0.8455858	1.7224908	1.3315142	1.829460	
ABBV	2.6135893	1.678795	4.9229331	10.000000	1.4560145	1.8271242	2.8760742	1.8871332	0.9859995	3.803294
ABC	1.8379008	2.401548	0.9909379	1.456014	10.00000000.9152611	0.0000000	0.9742351	1.6832298	2.507563	
ABT	0.9968404	1.279503	1.7990143	1.827124	0.9152611	10.00000003.7158739	0.0127669	4.2939824	2.137686	

Table 32: Bottom 10 values, Matched Based Similarity

X1	X2	X3
APA	APC	6.389351
ABBV	CF	6.485863
ALB	ALXN	6.497053
BCR	CF	6.594964
APD	ECL	6.614037
BDX	CELG	6.711435
AMZN	CTL	6.889358
AMGN	CSCO	7.330663
CMS	ECL	7.461483
AKAM	ALXN	7.470959

Table 33: Top 10 values, Matched Based Similarity

a	b	zeros
AAL	ADI	0
AAL	ALXN	0
AAL	ATVI	0
AAL	CAH	0
AAL	CERN	0
AAL	CTSH	0
AAL	DISCA	0
AAP	AMGN	0
AAP	BHI	0
AAP	CSCO	0

The diagonal of the resulting table is equal to 10, because each stock is identical to itself. Match-based similarity is, in fact, a similarity function and it takes on higher values as similarity between the pair of observations increases. The max value is 10. Therefore, the bottom 10 values tables show the most similar pairs of stock based on the 10 quantitative variables, and the top 10 values tables show the most different ones.

Then, *Malahanobis distance*, that is a distance function, has been defined and applied. It is similar to the Euclidean distance (Lp-Norm with p=2), except that it normalizes the data on the basis of the inter-attribute correlations. Below a subset of the resulting table and the rank of the top and bottom 10 values

Table 34: Mahalanobis distance

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	0.000000	3.039934	15.30825	6.819974	3.285434	3.672795	5.921759	24.81424	2.959609	3.898272
AAP	3.039934	0.000000	17.33981	7.456745	2.457782	3.254715	4.631246	24.41741	2.663489	3.577095
AAPL	15.308254	17.339811	0.00000	14.557916	17.402988	15.118439	17.142018	26.74652	15.957657	16.236320
ABBV	6.819974	7.456745	14.55792	0.000000	9.396315	5.704647	6.019666	19.27141	8.318201	4.035590
ABC	3.285434	2.457782	17.40299	9.396315	0.000000	4.752397	6.461205	26.38099	2.163477	5.698669
ABT	3.672795	3.254715	15.11844	5.704647	4.752397	0.000000	3.014326	22.11675	3.138072	3.021882

Table 35: Bottom 10 values, Mahalanobis distance

X1	X2	X3
ADI	CAH	26.04387
ADI	AN	26.14893
ADI	CCL	26.19800
ADI	CTL	26.36829
ABC	ADI	26.38099
ADI	BA	26.56589
AAPL	ADI	26.74652
ADI	CHTR	27.33675
ADI	CVX	28.87872
ADI	CLX	30.39335

Table 36: Top 10 values, Mahalanobis distance

X1	X2	X3
CMS	ECL	0.2096961
AVY	CBG	0.3174856
CAG	CNP	0.3259933
APD	CMS	0.4763226
APH	CHD	0.4783077
AME	EFX	0.4968039
CHD	DOV	0.5045214
AKAM	ALXN	0.5127965
APH	DOV	0.5295863
AAP	DPS	0.5422299

The diagonal of the resulting table is equal to 0, because each stock is identical to itself. Mahalanobis distance in fact, takes values starting from 0 and it takes on higher values as distance between the pair of observations increases. Therefore, the bottom 10 values tables show the most different pairs of stock based on the 10 quantitative variables, and the top 10 values tables show the most similar ones.

As for Categorical data, since no ordering exists, it is more common to work with similarity functions matching different values. The first similarity function for the categorical attribute of the stocks of S&P 500 is the *Overlap measure*. Given two stocks, this measure represents the number of the attributes for which the two stocks has the same value. Since we consider 3 categorical variables (Sector, Sub Industry and Headquarters Location), the max value is 3. Below a subset of the resulting table and the rank of the top and bottom 10 values

Table 37: Similarity: Overlap Measure

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	3	0	0	0	0	0	0	0	0	0
AAP	0	3	0	0	0	0	0	0	0	0
AAPL	0	0	3	0	0	0	1	1	0	1
ABBV	0	0	0	3	1	2	0	0	0	0
ABC	0	0	0	1	3	1	0	0	0	0
ABT	0	0	0	2	1	3	0	0	0	0

Table 38: Bottom 10 values, Overlap Measure

	a	b	zeros
[9890,]	EBAY	CSCO	2
[9891,]	EBAY	CTXS	2
[9892,]	ECL	ALB	2
[9893,]	ED	AEP	2
[9894,]	ED	D	2
[9895,]	ED	DUK	2
[9896,]	EFX	AYI	2
[9897,]	EFX	DAL	2
[9898,]	EFX	DNB	2
[9899,]	APA	COG	3

Table 39: Top 10 values, Overlap Measure

a	b	zeros
AAL	AAP	0
AAL	AAPL	0
AAL	ABBV	0
AAL	ABC	0
AAL	ABT	0
AAL	ADBE	0
AAL	ADI	0
AAL	ADM	0
AAL	ADS	0
AAL	ADSK	0

The diagonal of the resulting table is equal to 3, because each stock is identical to itself. It takes values starting from 0 and it takes on higher values as similarity between the pair of observations increases. Therefore, the bottom 10 values tables show the most similar pairs of stock based on the 3 categorical variables, and the top 10 values tables show the most different ones.

However, this method does not consider relative frequencies among different attributes. Therefore, *Inverse frequency* and *Goodall measure* have been computed. In this cases, for example, if 2 stocks match a variable in a rare value, it counts more than matching in a common value. Below a subset of the resulting tables and the rank of the top and bottom 10 values.

Table 40: Similarity: Inverse frequency

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	11145.71	0.00	0.00	0.00000	0.00000	0.00000	0.00000	0.00000	0	0.00000
AAP	0.00	20044.44	0.00	0.00000	0.00000	0.00000	0.00000	0.00000	0	0.00000
AAPL	0.00	0.00	20044.44	0.00000	0.00000	0.00000	44.44444	44.44444	0	44.44444
ABBV	0.00	0.00	0.00	12534.60208	34.60208	2534.60208	0.00000	0.00000	0	0.00000
ABC	0.00	0.00	0.00	34.60208	11145.71319	34.60208	0.00000	0.00000	0	0.00000
ABT	0.00	0.00	0.00	2534.60208	34.60208	2934.60208	0.00000	0.00000	0	0.00000

Table 41: Bottom 10 values, Inverse frequency

	a	b	zeros
1	COL	DHR	2534.602
2	DE	CAT	2534.602
3	DGX	DVA	2534.602
5	DNB	EFX	2534.602
6	DOV	CMI	2534.602
9	ADBE	ADSK	2544.444
11	ATVI	EA	2544.444
12	BWA	DLPH	2544.444
13	CHTR	DISCA	2544.444
14	CMG	DRI	2544.444
15	DG	DLTR	2544.444

Table 42: Top 10 values, Inverse frequency

a	b	zeros
AAL	AAP	0
AAL	AAPL	0
AAL	ABBV	0
AAL	ABC	0
AAL	ABT	0
AAL	ADBE	0
AAL	ADI	0
AAL	ADM	0
AAL	ADS	0
AAL	ADSK	0

Table 43: Similarity: Goodall

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	2.9701	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	0.0000
AAP	0.0000	2.9773	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	0.0000
AAPL	0.0000	0.0000	2.9773	0.0000	0.0000	0.0000	0.9775	0.9775	0	0.9775
ABBV	0.0000	0.0000	0.0000	2.9706	0.9711	1.9707	0.0000	0.0000	0	0.0000
ABC	0.0000	0.0000	0.0000	0.9711	2.9701	0.9711	0.0000	0.0000	0	0.0000
ABT	0.0000	0.0000	0.0000	1.9707	0.9711	2.9682	0.0000	0.0000	0	0.0000

Table 44: Bottom 10 values, Goodall

	a	b	zeros
[9883,]	DUK	ED	1.9920
[9884,]	ED	AEP	1.9920
[9885,]	ED	D	1.9920
[9887,]	AEE	CMS	1.9927
[9888,]	AEE	CNP	1.9927
[9890,]	CMS	CNP	1.9927
[9893,]	CHK	CVX	1.9932
[9894,]	CHK	DVN	1.9932
[9897,]	ALB	ECL	1.9947
[9899,]	APA	COG	2.9886

Table 45: Top 10 values, Goodall

a	b	zeros
AAL	AAP	0
AAL	AAPL	0
AAL	ABBV	0
AAL	ABC	0
AAL	ABT	0
AAL	ADBE	0
AAL	ADI	0

a	b	zeros
AAL	ADM	0
AAL	ADS	0
AAL	ADSK	0

These above are similarity measures, and they takes values starting from 0 and it takes on higher values as similarity between the pair of observations increases. Therefore, the bottom 10 values tables show the most similar pairs of stock based on the 3 categorical variables, and the top 10 values tables show the most different ones. Inverse frequency and Goodall measure, used to calculate similarity for categorical data, consider relative frequencies among different attributes: matching a variable in a rare value counts more than matching in a common value.

Finally, we merged the quantitative dataset and the categorical dataset to create a unique mixed type data dataset relative to the 100 stocks of the S&P 500 subset. After that, *Overall similarity* has been computed. This measure allows to use the overlap approach to mixed data by adding the weights of the numeric and quantitative components. A weight lambda (for numerical data) equal to 0.6 has been chosen (1-lambda for categorical).

Table 46: Overall similarity using mixed type data

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	7.2	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0	0.0
AAP	0.0	7.2	0.6	0.0	0.0	0.6	0.0	0.0	0	0.0
AAPL	0.0	0.6	7.2	0.6	0.0	0.0	0.4	0.4	0	0.4
ABBV	0.0	0.0	0.6	7.2	0.4	0.8	0.0	0.0	0	0.0
ABC	0.0	0.0	0.0	0.4	7.2	0.4	0.0	0.0	0	0.0
ABT	0.0	0.6	0.0	0.8	0.4	7.2	0.0	0.0	0	0.0

Table 47: Top 10 values, overall similarity

a	b	zeros
AAL	AAP	0
AAL	AAPL	0
AAL	ABBV	0
AAL	ABC	0
AAL	ABT	0
AAL	ADI	0
AAL	ADM	0
AAL	ADS	0
AAL	ADSK	0
AAL	AEE	0

Table 48: Bottom 10 values, overall similarity

	a	b	zeros
3	AMAT	ADBE	1.6
4	AMAT	EA	1.6
5	APC	BHI	1.6
6	ATVI	AAPL	1.6

	a	b	zeros
7	AWK	D	1.6
9	CRM	EA	1.6
13	ALK	BBBY	1.8
15	CMS	ECL	1.8
17	ABC	CAH	2.0
18	AEP	ED	2.0

This above is a similarity measure, and it takes values starting from 0 and it takes on higher values as similarity between the pair of observations increases. Therefore, the bottom 10 values tables show the most similar pairs of stock based on the 13 variables, and the top 10 values tables show the most different ones. Below the *Overall normalized similarity* measure, that is a similarity measure.

Table 49: Overall normalized similarity using mixed type data

	AAL	AAP	AAPL	ABBV	ABC	ABT	ADBE	ADI	ADM	ADS
AAL	15.82418	0.0000000	0.0000000	0.0000000	0.000000	0.0000000	0.9609078	0.000000	0	0.000000
AAP	0.00000	15.8241804	0.9609078	0.0000000	0.000000	0.9609078	0.0000000	0.000000	0	0.000000
AAPL	0.00000	0.9609078	15.8241804	0.9609078	0.000000	0.0000000	2.0717007	2.071701	0	2.071701
ABBV	0.00000	0.0000000	0.9609078	15.8241804	2.071701	4.1434014	0.0000000	0.000000	0	0.000000
ABC	0.00000	0.0000000	0.0000000	2.0717007	15.824180	2.0717007	0.0000000	0.000000	0	0.000000
ABT	0.00000	0.9609078	0.0000000	4.1434014	2.071701	15.8241804	0.0000000	0.000000	0	0.000000

Table 50: Bottom 10 values, overall normalized similarity

	a	b	zeros
1	AMGN	BIIB	5.104309
2	APA	CXO	5.104309
3	APC	CXO	5.104309
5	BWA	DLPH	5.104309
6	CAG	CPB	5.104309
7	COL	DHR	5.104309
9	CRM	EBAY	5.104309
15	ABC	CAH	6.065217
16	AEP	ED	6.065217
19	APA	COG	6.215102

Table 51: Top 10 values, overall normalized similarity

a	b	zeros
AAL	AAP	0
AAL	AAPL	0
AAL	ABBV	0
AAL	ABC	0
AAL	ABT	0
AAL	ADI	0
AAL	ADM	0
AAL	ADS	0

a	b	zeros
AAL	ADSK	0
AAL	AEE	0