

Data Mining

Assignment 2

Data Exploration and Preprocessing

Aya Lotfy Saeed 4

Dahlia Chehata 27

Table of contents

Dataset investigation	3
number of readings and attributes	3
classes	3
data description	4
dataset information	4
Dataset visualisation	5
Boxplot	5
plots for each class	6
Data exploration	8
Pearson's correlation	8
matrix	8
visualisation	9
covariance matrix	9
matrix	9
visualisation	10
What is the relation between the covariance matrix of the dataset and the Pearson's correlation matrix of it?	10
Histograms	11
Histograms for each class	11
bins 5, 10, 12	14
Preprocessing	17
Normalization	17
Min-max scaler	17
Z-score normalization	18
Dimensionality reduction	19
Feature Projection	19
Feature selection	23

Dataset investigation

The dataset used is the combination of the files :segmentation.data and segmentation.test

- **number of readings and attributes**

- the data file consists of 210 instances (rows) and 19 features/attributes (columns)
- the test file consists of 2100 instances (rows) with 19 features/attributes (columns)
- the merged data consists of 2310 instances (rows) with 20 features (columns)

- **classes**

- The class column has no header. We add a name to it and reset the index to be a zero based index
- The different class values are (['BRICKFACE', 'SKY', 'FOLIAGE', 'CEMENT', 'WINDOW', 'PATH', 'GRASS'])
- each unique value in class has 330 instances

Index	class	ION-CENTROID-X	ION-CENTROID-Y	ION-PIXEL-COUL	JRT-LINE-DENSIT	JRT-LINE-DENSIT	VEDGE-MEAN	VEDGE-SD	HEDGE-MEAN	HEDGE-SD	NTENSITY-MEAN	RAWRED-MEAN	RAWRED-SD
0	BRICKFACE	140	125	9	0	0	0.277778	0.062963	0.666667	0.311111	6.18518	7.33333	7.66667
1	BRICKFACE	27	68	9	0	0	1.38889	1.48519	1.77778	5.0963	21.5926	20.4444	28.8889
2	BRICKFACE	29	100	9	0	0	2.22222	1.36296	2.66667	3.64444	20.6296	20.8889	25.5556
3	BRICKFACE	23	113	9	0	0	0.722222	0.996296	2.77778	2.78518	14.7037	17.1111	16.6667
4	BRICKFACE	40	85	9	0.111111	0	1.11111	0.518519	2.5	0.655555	21.1111	20.7778	27.7778
5	BRICKFACE	77	78	9	0	0	2.22222	2.34074	2	5.86667	24.5926	24	31.1111
6	BRICKFACE	96	92	9	0.111111	0	3.33333	2.97778	1.77778	0.651851	22.8889	21.7778	29.7778
7	BRICKFACE	147	92	9	0	0	1.11111	0.429631	1.38889	2.24074	23.4444	22.7778	30
8	BRICKFACE	42	59	9	0	0	1.83333	3.5	2.05556	0.862963	21.8148	20.7778	28.8889
9	BRICKFACE	2	63	9	0	0	1.22222	0.562963	1.5	0.700001	18.0741	17.2222	23.3333
10	BRICKFACE	31	106	9	0	0	1.61111	0.72963	0.722222	1.17407	19.1111	19.5556	23.3333
11	BRICKFACE	33	104	9	0.111111	0	1.55556	0.962963	3.05556	5.04074	20.1852	20.2222	24.4444
12	BRICKFACE	81	98	9	0.111111	0	1.22222	0.42963	1.88889	3.22963	20.8889	21	25.5556
13	BRICKFACE	6	90	9	0	0	1.94444	0.596296	1.33333	1.82222	18.1111	18.7778	22.2222
14	BRICKFACE	76	81	9	0.111111	0	1.88889	1.67407	1.33333	2.4	22.7037	22.3333	28.8889
15	BRICKFACE	145	90	9	0	0	1	0.577778	2	0.399999	24.8519	23.6667	32.2222
16	BRICKFACE	133	67	9	0	0	1.38889	1.3963	2.11111	1.22963	27.1481	25.1111	35.5556
17	BRICKFACE	80	95	9	0	0	1.22222	1.00741	0.944444	0.551851	21.4074	21.3333	26.6667
18	BRICKFACE	91	115	9	0.111111	0	1.72222	1.44074	3.88889	2.82963	19.9259	20.5556	24.4444
19	BRICKFACE	75	107	9	0.111111	0	1.16667	0.344445	1.77778	1.54074	16.7037	18.7778	19.7778
20	BRICKFACE	121	60	9	0	0	2.27778	2.32963	2.88889	2.87407	26.7407	24.6667	35.5556

- data description

x - DataFrame

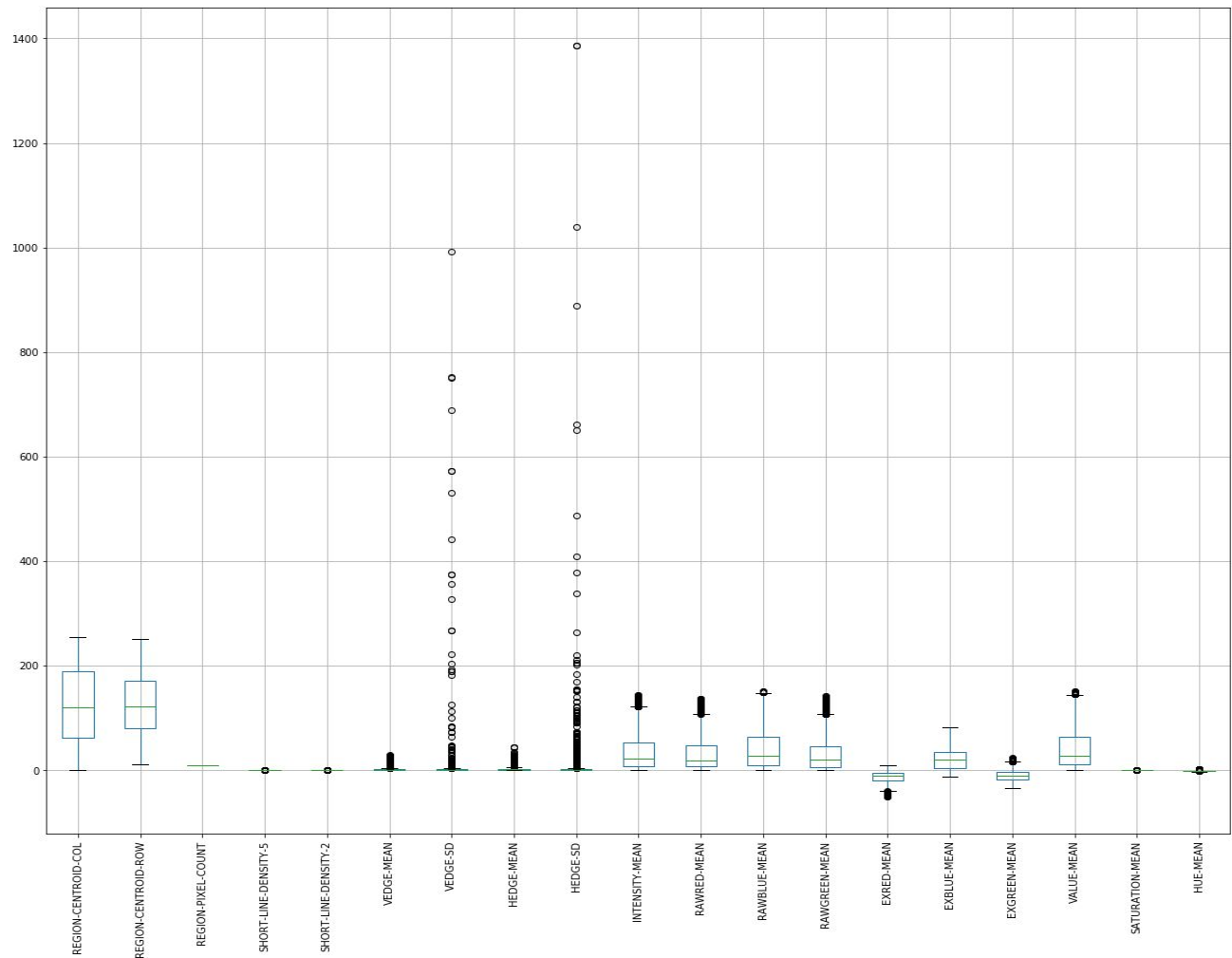
Index	ION-CENTROID-C	ION-CENTROID-F	ION-PIXEL-COU	ORT-LINE-DENSIT	ORT-LINE-DENSIT	VEDGE-MEAN	VEDGE-SD	HEDGE-MEAN	HEDGE-SD	NTENSITY-MEAN	RAWRED-MEAN	RAWBLUE-MEAN	RAWGR
count	2310	2310	2310	2310	2310	2310	2310	2310	2310	2310	2310	2310	2310
mean	124.914	123.417	9	0.0143338	0.0047138	1.89394	5.70932	2.42472	8.24369	37.0516	32.8213	44.1879	34.14
std	72.9565	57.4839	0	0.0401541	0.0242343	2.69891	44.8465	3.61008	58.8115	38.1764	35.0368	43.5275	36.36
min	1	11	9	0	0	0	0	0	-1.58946e-08	0	0	0	0
25%	62	81	9	0	0	0.722222	0.355555	0.77778	0.421637	7.2963	7	9.55556	6.027
50%	121	122	9	0	0	1.22222	0.833333	1.44444	0.962963	21.5926	19.5556	27.6667	20.33
75%	189	172	9	0	0	2.16667	1.80637	2.55556	2.18327	53.213	47.3333	64.8889	46.5
max	254	251	9	0.333333	0.222222	29.2222	991.718	44.7222	1386.33	143.444	137.111	150.889	142.5

- dataset information

```
In [52]: display (dataset.info())
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2310 entries, 0 to 2309
Data columns (total 20 columns):
class                2310 non-null object
REGION-CENTROID-COL  2310 non-null float64
REGION-CENTROID-ROW  2310 non-null float64
REGION-PIXEL-COUNT    2310 non-null int64
SHORT-LINE-DENSITY-5  2310 non-null float64
SHORT-LINE-DENSITY-2  2310 non-null float64
VEDGE-MEAN            2310 non-null float64
VEDGE-SD             2310 non-null float64
HEDGE-MEAN            2310 non-null float64
HEDGE-SD            2310 non-null float64
INTENSITY-MEAN        2310 non-null float64
RAWRED-MEAN           2310 non-null float64
RAWBLUE-MEAN         2310 non-null float64
RAWGREEN-MEAN        2310 non-null float64
EXRED-MEAN           2310 non-null float64
EXBLUE-MEAN          2310 non-null float64
EXGREEN-MEAN         2310 non-null float64
VALUE-MEAN            2310 non-null float64
SATURATION-MEAN       2310 non-null float64
HUE-MEAN             2310 non-null float64
dtypes: float64(18), int64(1), object(1)
memory usage: 361.0+ KB
None
```

- **Dataset visualisation**

- **Boxplot**



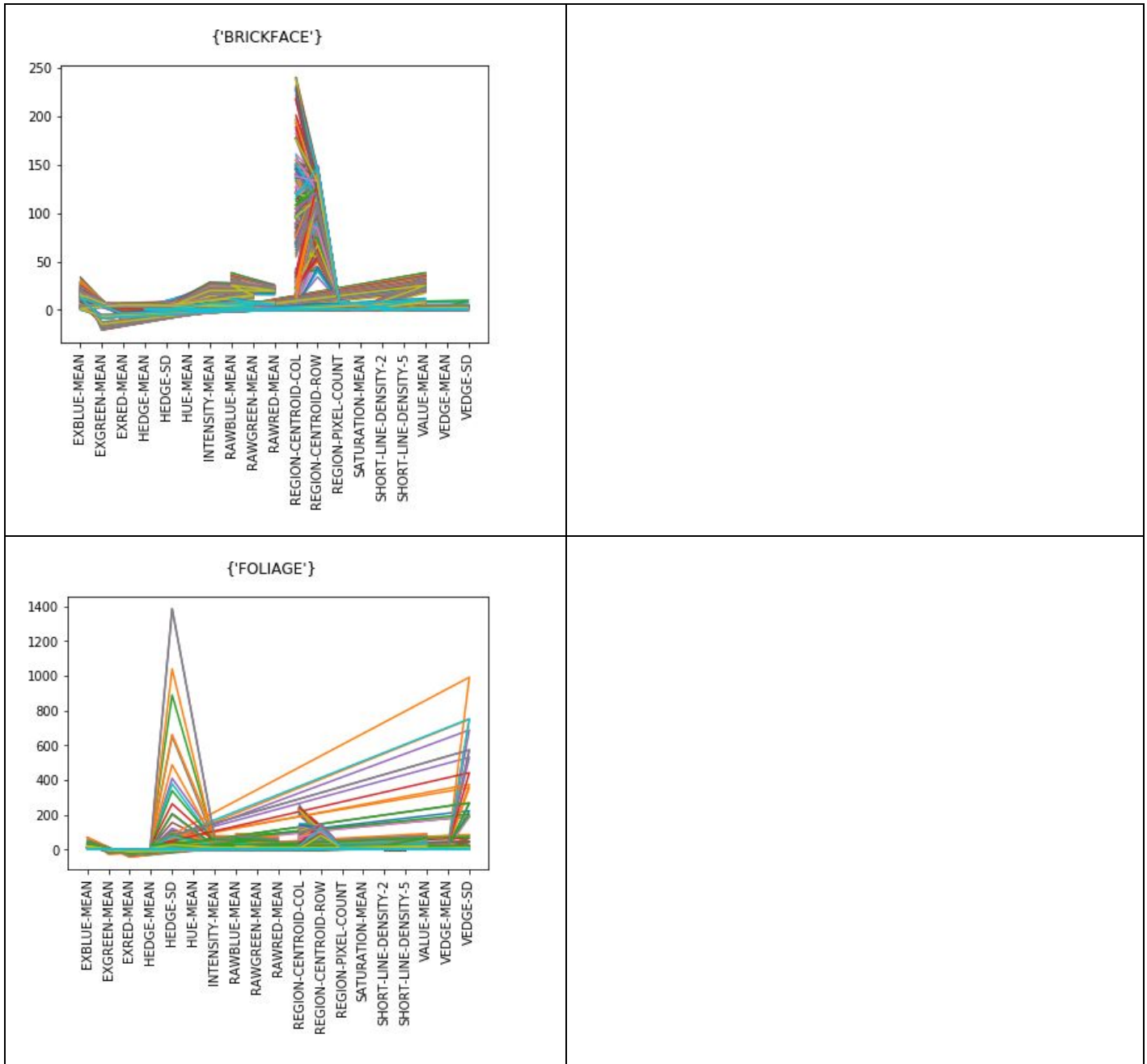
Observations:

- The dataset attributes are different in ranges
- *REGION-CENTROID-COL* have a normal distribution and no outliers
- *REGION-PIXEL-COUNT* has a zero standard deviation because it has fixed value = 9. (redundant dimension to be removed).
- *SHORT-LINE-DENSITY-5* and *SHORT-LINE-DENSITY-2* has a small standard deviation
- *VEDGE-MEAN* and *HEDGE-MEAN* have outliers that affects mean and standard deviation.
- *HEDGE-SD* and *VEDGE-SD* have very large outliers

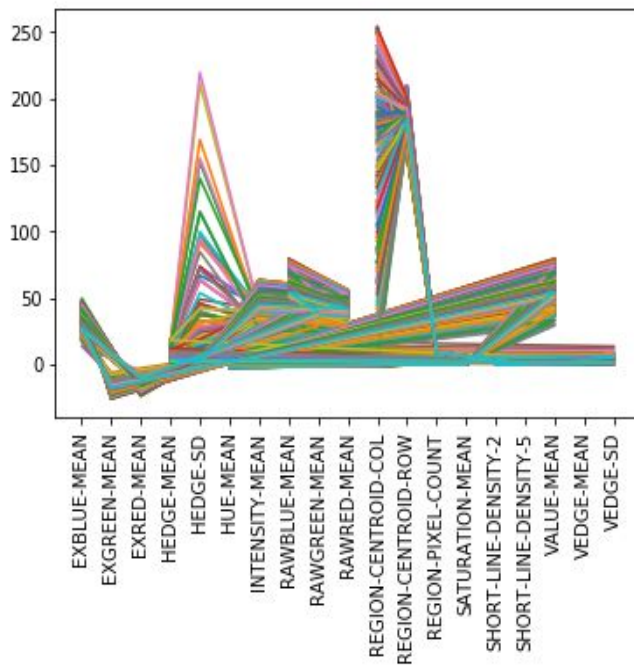
Conclusion

- The data needs to be normalized
- Outliers need to be removed
- Redundant attributes need to be removed

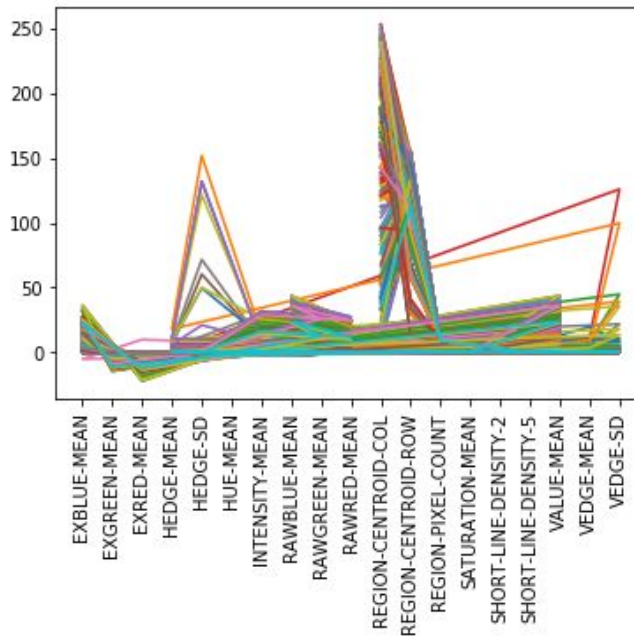
- plots for each class



{'PATH'}



{'WINDOW'}

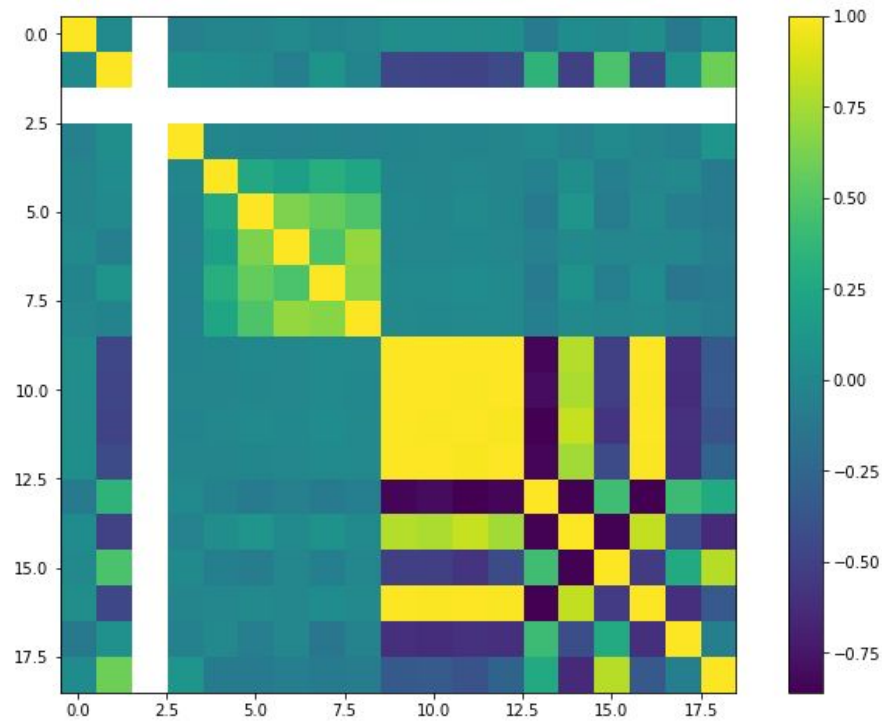


Data exploration

- Pearson's correlation
 - matrix

Index	ION-CENTROID-COL	ION-CENTROID-ROW	REGION-PIXEL-COUNT	SHORT-LINE-DENSITY-5	SHORT-LINE-DENSITY-2	VEDGE-MEAN	VEDGE-SD	HEDGE-MEAN	HEDGE-SD	INTENSITY-MEAN	RAWRED-MEAN	RAWBLUE-MEAN	RAWGREEN-MEAN
REGION-CENTROID-COL	1	0.0267683	nan	-0.0519617	-0.0159643	-0.0113042	0.0219603	-0.0189142	-0.00193879	0.0589574	0.054673	0.0581691	0.063
REGION-CENTROID-ROW	0.0267683	1	nan	0.0648913	0.0418694	0.0261463	-0.053578	0.105223	-0.0210774	-0.46524	-0.468009	-0.481521	-0.43
REGION-PIXEL-COUNT	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
SHORT-LINE-DENSITY-5	-0.0519617	0.0648913	nan	1	-0.00902435	-0.0202057	-0.0327814	-0.0212863	-0.0379961	-0.0182106	-0.0167553	-0.0213921	-0.01
SHORT-LINE-DENSITY-2	-0.0159643	0.0418694	nan	-0.00902435	1	0.262575	0.193728	0.303182	0.243155	-0.00691096	-0.0124706	0.00307818	-0.01
VEDGE-MEAN	-0.0113042	0.0261463	nan	-0.0202057	0.262575	1	0.637452	0.559491	0.488347	0.0051292	-0.00548196	0.0204975	-0.00
VEDGE-SD	0.0219603	-0.053578	nan	-0.0327814	0.193728	0.637452	1	0.471016	0.703049	0.00300641	-0.00213776	0.00678241	0.003
HEDGE-MEAN	-0.0189142	0.105223	nan	-0.0212863	0.303182	0.559491	0.471016	1	0.668179	0.0339725	0.0260589	0.0438457	0.029
HEDGE-SD	-0.00193879	-0.0210774	nan	-0.0379961	0.243155	0.488347	0.703049	0.668179	1	0.013518	0.00853753	0.0168992	0.014
INTENSITY-MEAN	0.0589574	-0.46524	nan	-0.0182106	-0.00691096	0.0051292	0.00300641	0.0339725	0.013518	1	0.998112	0.995809	0.995
RAWRED-MEAN	0.054673	-0.468009	nan	-0.0167553	-0.0124706	-0.00548196	-0.00213776	0.0260589	0.00853753	0.998112	1	0.990813	0.994
RAWBLUE-MEAN	0.0581691	-0.481521	nan	-0.0213921	0.00307818	0.0204975	0.00678241	0.0438457	0.0168992	0.995809	0.990813	1	0.984
RAWGREEN-MEAN	0.0633807	-0.437971	nan	-0.0156042	-0.013435	-0.00309891	0.00340993	0.0294059	0.014121	0.995842	0.994056	0.984659	1
EXRED-MEAN	-0.0068165	0.353175	nan	0.0280127	-0.0448293	-0.100457	-0.0491233	-0.0994335	-0.0561856	-0.830261	-0.794457	-0.855058	-0.82
EXBLUE-MEAN	0.0430985	-0.490219	nan	-0.036164	0.0609787	0.106744	0.0276592	0.0937381	0.0336465	0.792257	0.76997	0.844741	0.742
EXGREEN-MEAN	0.0140351	0.476421	nan	0.0331823	-0.0583623	-0.0801201	0.00239638	-0.0591112	-0.000666109	-0.509756	-0.507899	-0.573816	-0.42
VALUE-MEAN	0.0601893	-0.458388	nan	-0.0158859	-0.000145206	0.0181477	0.00480412	0.0422324	0.0148579	0.997385	0.992062	0.998644	0.990
SATURATION-MEAN	-0.108214	0.0815563	nan	-0.0432207	0.0162084	-0.0648269	0.0023061	-0.125955	-0.0241491	-0.60829	-0.616928	-0.595166	-0.60
HUE-MEAN	0.0392985	0.59293	nan	0.112989	-0.0829386	-0.0979591	-0.0615915	-0.0938031	-0.0699882	-0.329845	-0.328574	-0.384925	-0.26

- visualisation



● covariance matrix

- matrix

covariance - DataFrame													
Index	ION-CENTROID-COL	ION-CENTROID-ROW	REGION-PIXEL-COUNT	SHORT-LINE-DENSITY-5	SHORT-LINE-DENSITY-2	VEDGE-MEAN	VEDGE-SD	HEDGE-MEAN	HEDGE-SD	INTENSITY-MEAN	RAWRED-MEAN	RAWBLUE-MEAN	RAWGREEN-MEAN
REGION-CENTROID-COL	5322.66	112.262	0	-0.152222	-0.0282257	-2.22583	71.8508	-4.98161	-8.31874	164.209	139.753	184.723	168.152
REGION-CENTROID-ROW	112.262	3304.39	0	0.149783	0.0583275	4.05643	-138.121	21.8359	-71.2565	-1020.98	-942.593	-1204.83	-915.53
REGION-PIXEL-COUNT	0	0	0	0	0	0	0	0	0	0	0	0	0
SHORT-LINE-DENSITY-5	-0.152222	0.149783	0	0.00161235	-0.78166e-06	-0.00218974	-0.0590317	-0.00308566	-0.0097287	-0.0279157	-0.0235725	-0.0373893	-0.0227852
SHORT-LINE-DENSITY-2	-0.0282257	0.0583275	0	-0.78166e-06	0.000587302	0.0171741	0.210548	0.0265247	0.346558	-0.00639388	-0.0105887	0.00324704	-0.01184
VEDGE-MEAN	-2.22583	4.05643	0	-0.00218974	0.0171741	7.2841	77.1549	5.45128	77.5139	0.528485	-0.51838	2.40798	-0.304143
VEDGE-SD	71.8508	-138.121	0	-0.0590317	0.210548	77.1549	2011.2	76.2572	1854.28	5.14721	-3.35901	13.2396	5.56103
HEDGE-MEAN	-4.98161	21.8359	0	-0.00308566	0.0265247	5.45128	76.2572	13.0327	141.864	4.6821	3.29607	6.88981	3.86041
HEDGE-SD	-8.31874	-71.2565	0	-0.0097287	0.346558	77.5139	1854.28	141.864	3458.79	30.3509	17.5921	43.2605	30.2001
INTENSITY-MEAN	164.209	-1020.98	0	-0.0279157	-0.00639388	0.528485	5.14721	4.6821	30.3509	1457.44	1335.05	1654.76	1382.5
RAWRED-MEAN	139.753	-942.593	0	-0.0235725	-0.0105887	-0.51838	-3.35901	3.29607	17.5921	1335.05	1227.58	1511.05	1266.53
RAWBLUE-MEAN	184.723	-1204.83	0	-0.0373893	0.00324704	2.40798	13.2396	6.88981	43.2605	1654.76	1511.05	1894.64	1558.58
RAWGREEN-MEAN	168.152	-915.53	0	-0.0227852	-0.01184	-0.304143	5.56103	3.86041	30.2001	1382.5	1266.53	1558.58	1322.5
EXRED-MEAN	-73.3683	235.168	0	0.0130295	-0.0125845	-3.1406	-25.5187	-4.15807	-38.2763	-367.157	-322.431	-431.123	-347.157
EXBLUE-MEAN	61.54	-551.527	0	-0.0284209	0.0289228	5.63848	24.2772	6.62315	38.7288	591.96	527.994	719.645	528.2
EXGREEN-MEAN	11.8283	316.359	0	0.0153915	-0.0163383	-2.49789	1.24144	-2.46507	-0.452533	-224.802	-205.563	-288.522	-180.8
VALUE-MEAN	188.478	-1130.98	0	-0.0273791	-0.00015104	2.10227	9.2474	6.54395	37.5056	1634.31	1491.9	1865.74	1545.5
SATURATION-MEAN	-1.80249	1.07035	0	-0.000396228	0.96796e-05	-0.0399454	0.0236118	-0.103814	-0.324256	-5.30187	-4.93495	-5.9146	-5.05
HUE-MEAN	4.43061	52.671	0	0.00781116	-0.00310606	-0.40856	-4.26846	-0.523308	-6.36077	-19.4593	-17.7901	-25.8918	-14.6

- visualisation

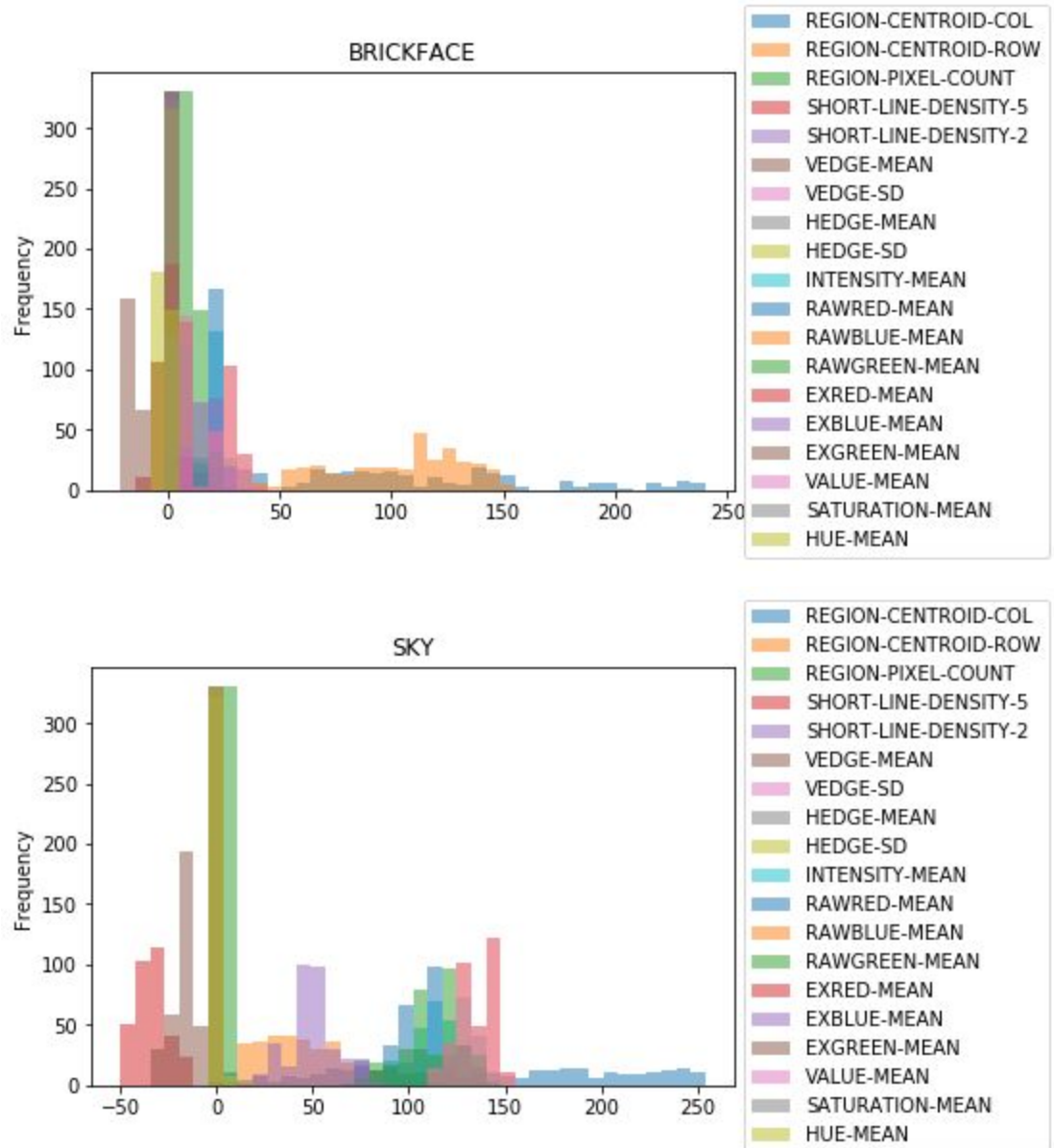
- **What is the relation between the covariance matrix of the dataset and the Pearson's correlation matrix of it?**

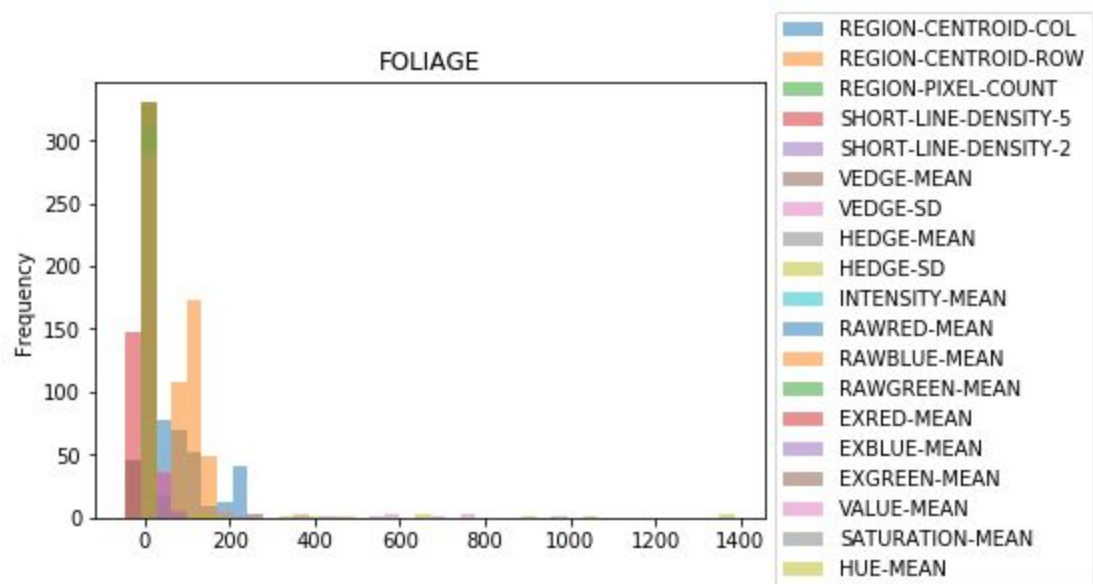
- Correlation is a special case of covariance where the matrix is standardized
- Both measures only linear relationship between two variables, i.e. when the correlation coefficient is zero, covariance is also zero.
- A measure used to indicate the extent to which two random variables change in tandem is known as covariance. A measure used to represent how strongly two random variables are related known as correlation.
- Covariance is a measure of correlation. On the contrary, correlation refers to the scaled form of covariance.
- The value of correlation takes place between -1 and +1. Conversely, the value of covariance lies between $-\infty$ and $+\infty$
- Correlation is dimensionless, i.e. it is a unit-free measure of the relationship between variables. Unlike covariance, where the value is obtained by the product of the units of the two variables.

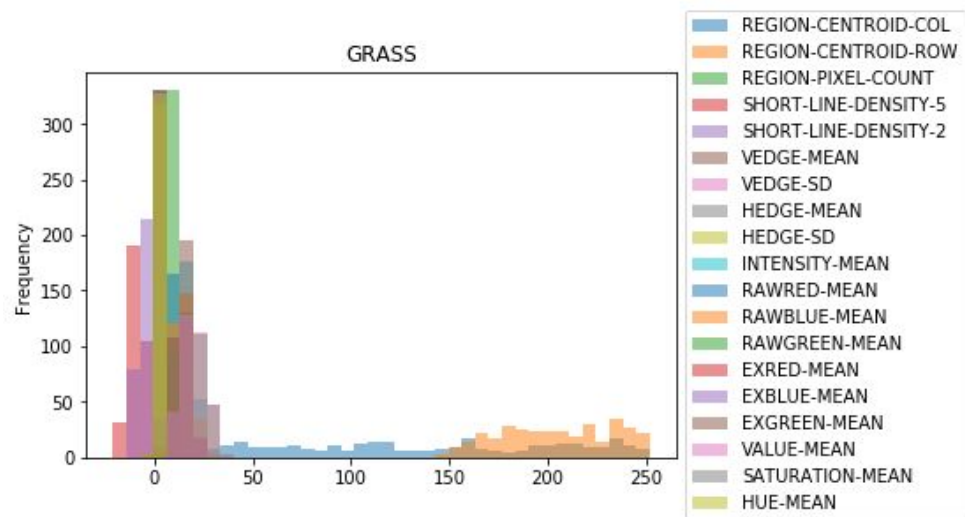
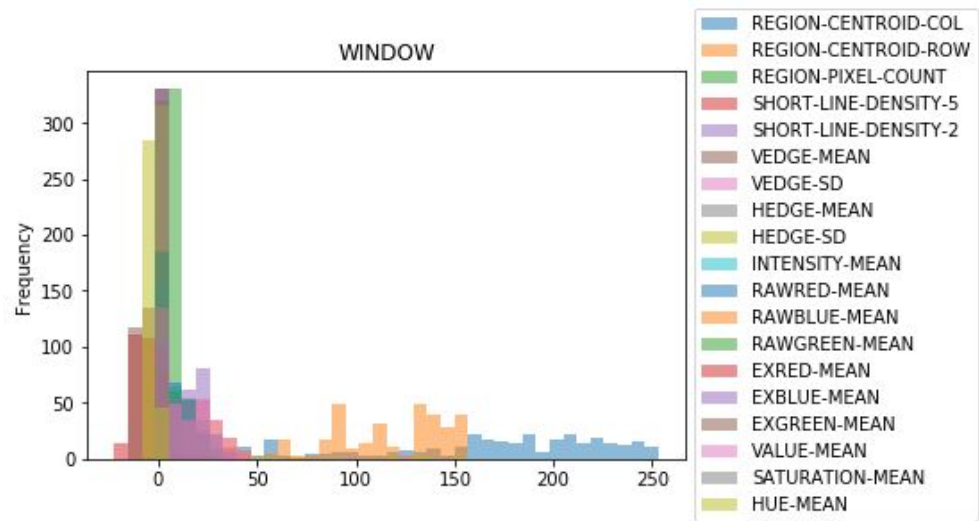
$$r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}$$

- Histograms

- Histograms for each class







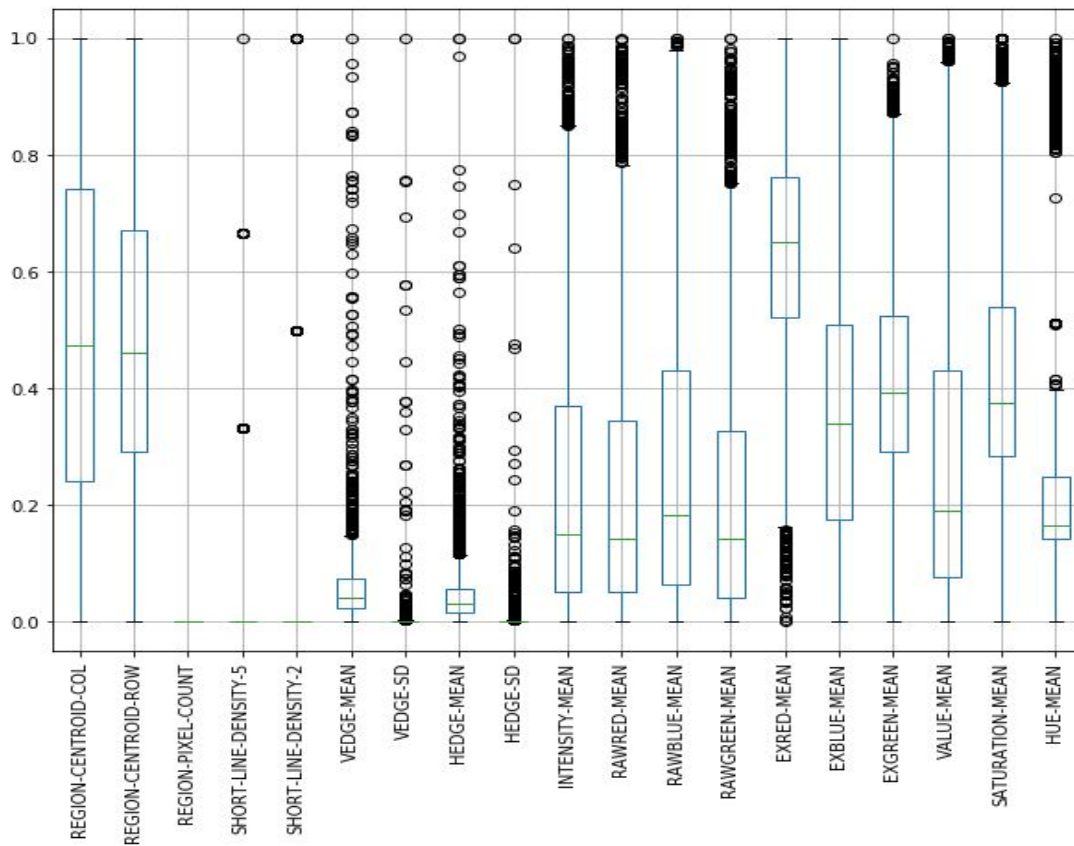
- bins 5, 10, 12

Preprocessing

- Normalization

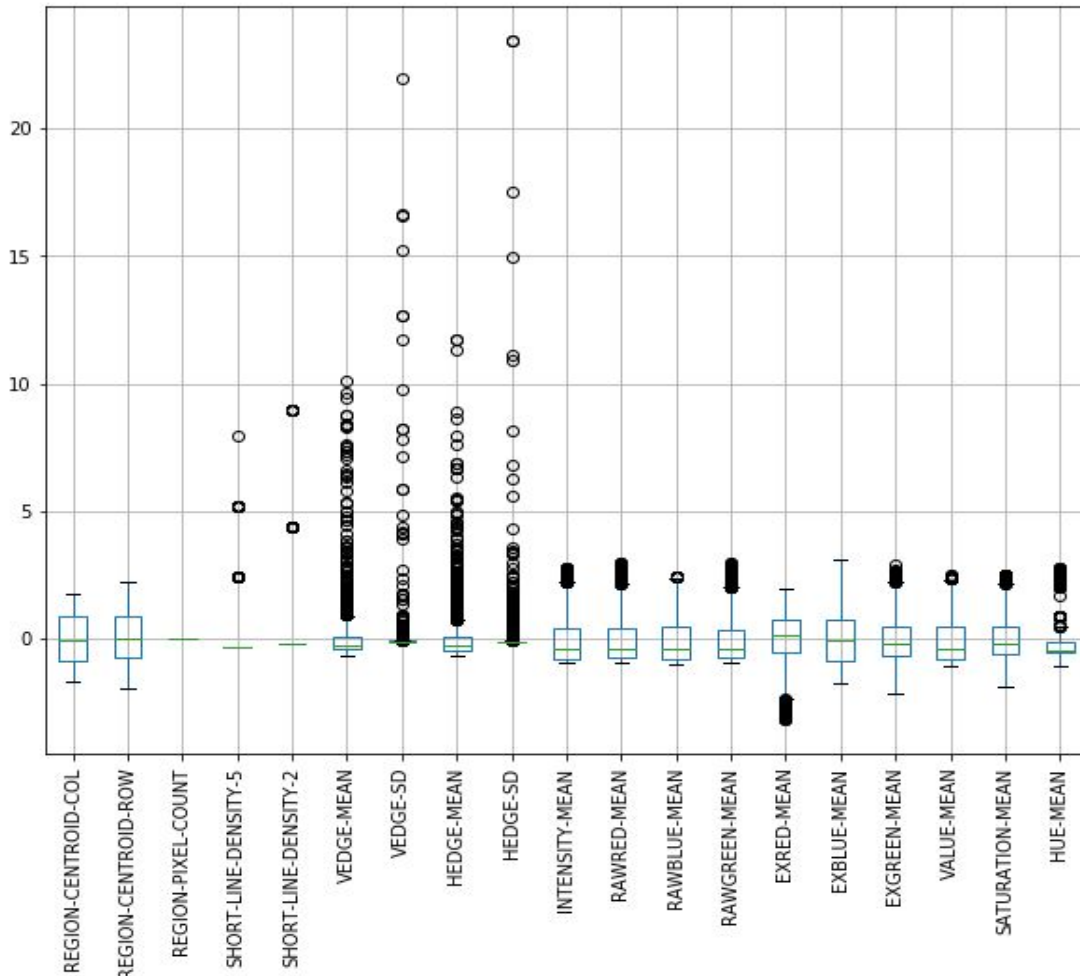
1. Min-max scaler

After the max-min normalization, all data are scaled to be in range from 0 to 1

[illegible]

2. Z-score normalization

All values almost have mean = 0, sd = 1. This is good for the normally distributed features.



The difference is that: the feature with high range will not dominate after normalization.

Z-score method preserves range (maximum and minimum) and introduces the dispersion of the serie (standard deviation / variance). If data follow a gaussian distribution, they are converted into a $N(0,1)$ distribution and the comparison between series (probabilities calculation) will be easier.

- **Dimensionality reduction**

- 1. **Feature Projection**

- chosen components_num = [1, 2, 4, 6, 8, 10, 13, 16, 19]

- [0.42341135]
 - [0.42341135 0.16203649]
 - [0.42341135 0.16203649 0.09959451 0.05857283]
 - [0.42341135 0.16203649 0.09959451 0.05857283 0.05197997 0.05050372]
 - [0.42341135 0.16203649 0.09959451 0.05857283 0.05197997 0.05050372 0.04041415 0.03120143]
 - [0.42341135 0.16203649 0.09959451 0.05857283 0.05197997 0.05050372 0.04041415 0.03120143 0.02999802 0.02195028]
 - [0.42341135 0.16203649 0.09959451 0.05857283 0.05197997 0.05050372 0.04041415 0.03120143 0.02999802 0.02195028 0.0142209 0.00993527 0.00616366]
 - [4.23411347e-01 1.62036489e-01 9.95945088e-02 5.85728321e-02 5.19799667e-02 5.05037229e-02 4.04141498e-02 3.12014310e-02 2.99980217e-02 2.19502844e-02 1.42209011e-02 9.93526978e-03 6.16366453e-03 1.74116798e-05 1.58780274e-16 1.30219428e-16]
 - [4.23411347e-01 1.62036489e-01 9.95945088e-02 5.85728321e-02 5.19799667e-02 5.05037229e-02 4.04141498e-02 3.12014310e-02 2.99980217e-02 2.19502844e-02 1.42209011e-02 9.93526978e-03 6.16366453e-03 1.74116798e-05 1.58780274e-16 1.30219428e-16 1.03964749e-16 9.55203032e-17 1.39078630e-34]

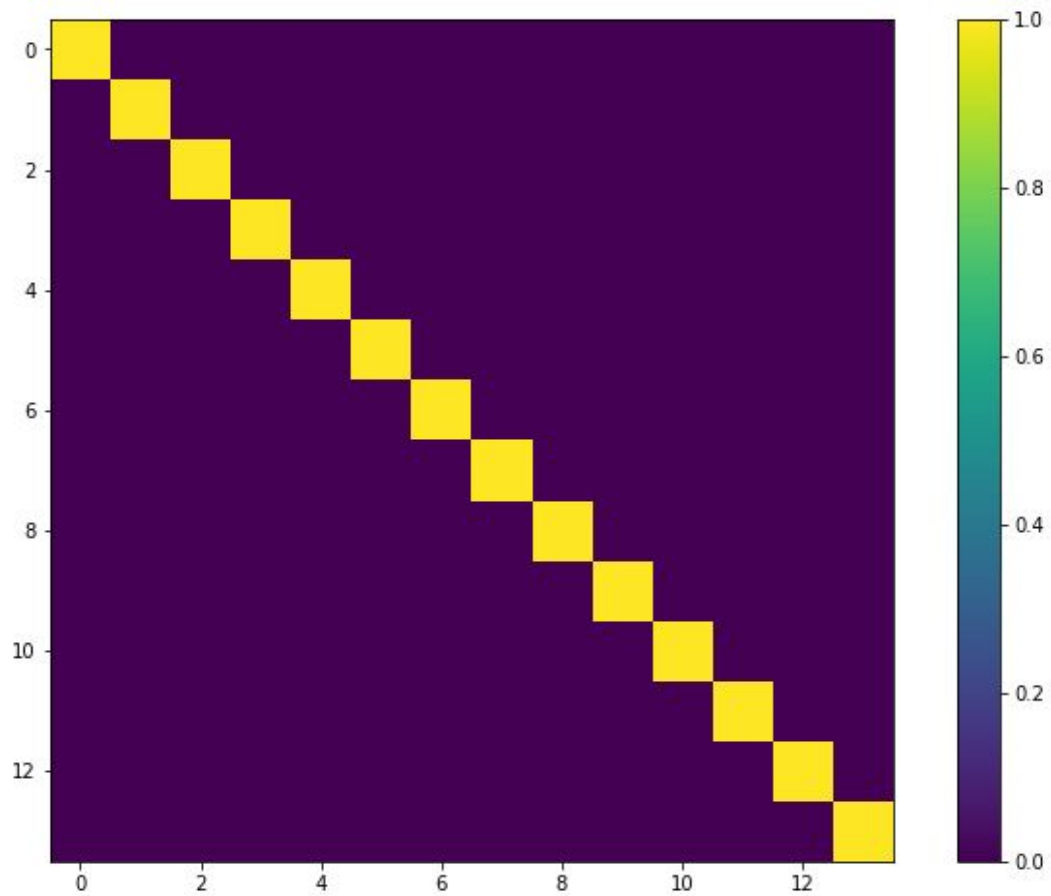
index	captured variance	component number
	sum	
0	0.423411	1
1	0.585448	2
2	0.743615	4
3	0.846099	6
4	0.917714	8
5	0.969663	10
6	0.999983	13
7	1.000000	16
8	1.000000	19

We notice that the last 3 components are nearly 1 so, we can take the first 14 components

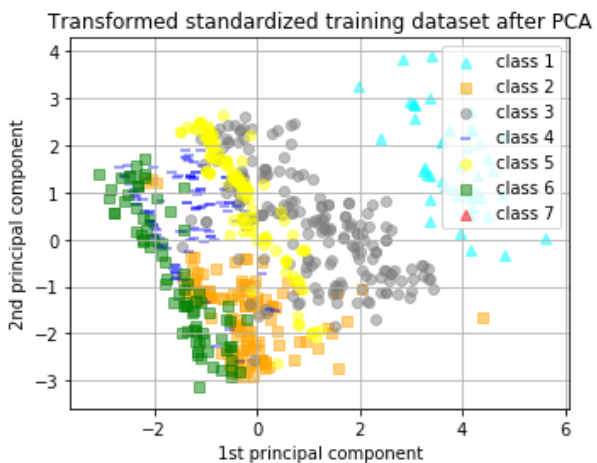
showing the matrix resultant after the PCA

Index	0	1	2	3	4	5	6	7	8	9	10	11	
0	-2.34101	-0.568638	-0.650548	-0.477064	-0.0547828	-0.136842	0.012842	-0.62045	0.276409	0.263032	-0.284534	-0.0399267	-0.
1	-0.612479	-0.191599	-1.42468	0.624449	-0.971095	0.178148	-0.291384	-0.656609	0.596323	-0.00602486	0.327241	-0.00252681	-0.
2	-0.986023	0.04876	-0.985904	0.647385	-0.896088	0.293113	-0.68396	-0.65327	0.726483	-0.0861032	-0.0956005	-0.060829	-0.
3	-1.78441	-0.21824	-0.815338	0.692126	-1.0088	0.378406	-0.564007	-1.03559	0.533698	0.11252	-0.387638	0.0935647	-0.
4	-0.840568	-0.284457	-0.870496	2.50717	0.477375	-1.11299	-0.0888781	-0.778695	0.351134	-0.0601369	0.0171944	0.107554	-0.
5	-0.497166	-0.0268167	-1.14837	0.190721	-0.512381	-0.0191702	-0.440232	-0.619748	0.764301	-0.0907055	0.101014	-0.135551	-0.
6	-0.616498	0.0286702	-0.721122	1.99571	0.987315	-1.40613	-0.139021	-0.186877	0.674429	-0.231794	-0.0196083	-0.251919	-0.
7	-0.606379	-0.345504	-0.985121	-0.466912	0.139852	-0.240579	-0.331362	-0.687787	0.523784	0.144244	0.0124836	-0.0631898	-0.
8	-0.556099	-0.0993556	-1.4741	0.489699	-0.853301	0.0663909	-0.261186	-0.618027	0.706274	-0.171118	0.324907	0.0338353	-0.
9	-0.920873	-0.297635	-1.42826	0.831645	-1.24293	0.268776	-0.190548	-0.732222	0.706257	-0.0779116	0.368993	0.0473832	-0.
10	-1.15722	-0.333159	-0.996147	0.598124	-0.889067	0.279689	-0.521061	-0.59979	0.835949	0.274237	-0.101258	-0.188172	-0.
11	-1.06347	-0.0946611	-0.529351	2.58179	0.414788	-1.01566	-0.28929	-0.806169	0.39359	-0.108707	-0.13508	0.06411	-0.
12	-0.985916	-0.336779	-0.550753	2.12227	0.81748	-1.26563	-0.0918333	-0.818186	0.47142	0.0291417	-0.12493	-0.0276746	-0.
13	-1.21033	-0.198951	-1.05523	0.825672	-1.16814	0.326482	-0.487676	-0.727645	0.963697	0.0178624	-0.00487703	-0.148549	-0.
14	-0.717101	-0.290457	-0.793273	2.16101	0.751873	-1.34338	0.023042	-0.628421	0.643131	-0.046785	0.00645112	-0.142351	-0.
15	-0.425876	-0.296811	-1.03836	-0.436665	0.151791	-0.208142	-0.377001	-0.662763	0.387249	0.0892526	0.0760996	0.0520209	-0.
16	-0.0626347	-0.205131	-1.27661	-0.330228	0.0147155	-0.248289	-0.265724	-0.58549	0.445431	-0.069919	0.325137	0.0492806	-0.
17	-0.886784	-0.387908	-0.969632	0.141301	-0.480332	0.0452911	-0.390777	-0.720113	0.745749	0.201338	-0.0124774	-0.0916619	-0.
18	-1.16836	0.0186746	-0.288599	2.069	0.952813	-1.21371	-0.388945	-0.859736	0.327125	-0.226374	-0.341839	0.12225	-0.
19	-1.5681	-0.386736	-0.398264	2.18156	0.73347	-1.23042	-0.124791	-0.995216	0.638688	0.0564065	-0.394325	-0.0724101	-0.
20	-0.0685063	0.0718931	-1.27563	-0.214001	-0.122107	-0.237934	-0.312053	-0.526885	0.548046	-0.353922	0.351167	0.0272227	-0.

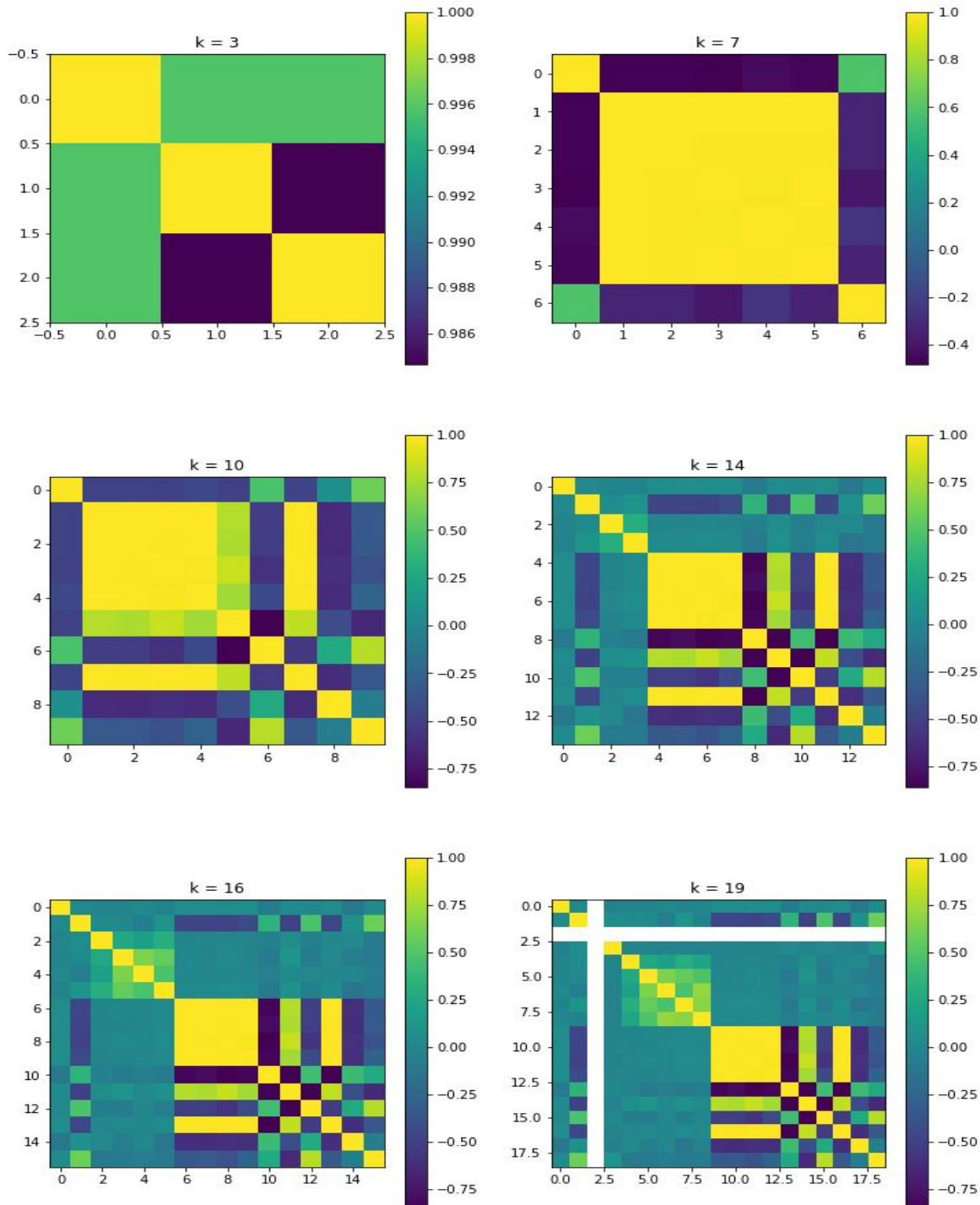
Visualization:



The visualisations shows that the correlation between each 2 attributes is 0.
So we have chosen the best uncorrelated features to avoid redundancy



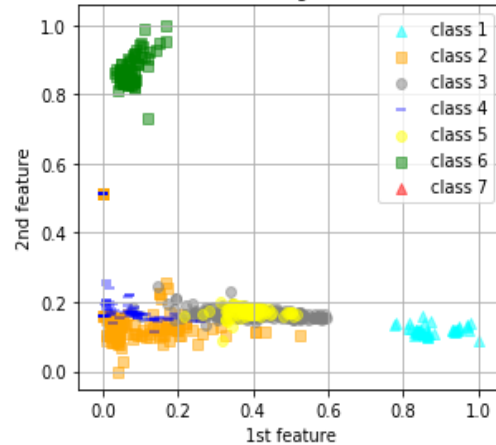
2. Feature selection



The first 5 attributes are very correlated as shown at $k=4$ and $k=7$ starting from index 1. Also the 3rd attribute appears only when $k=19$ which means that it is the worst attribute as we expected, due to the zero standard deviation . It gives no information.

K = 2

Transformed standardized training dataset after feature selection



Correlation matrix

- **K =5**

```
[[0.04311903 0.05357143 0.05081001 0.05154639 0.32532112]
 [0.0464756  0.06087663 0.05154639 0.05596466 0.35865966]
 [0.04260263 0.0551948  0.04786451 0.05007364 0.35169946]
 ...
 [0.78156473 0.71022725 0.86450659 0.86450659 0.09970969]
 [0.76658921 0.71266236 0.84462441 0.84462441 0.11727551]
 [0.7384457  0.6712662  0.82768774 0.82768774 0.1120159 ]]
```

- **k=9**

```
[[0.475    0.04311903 0.05357143 ... 0.43220337 0.05154639 0.32532112]
 [0.50833333 0.0464756  0.06087663 ... 0.4237288  0.05596466 0.35865966]
 [0.53333333 0.04260263 0.0551948  ... 0.43644067 0.05007364 0.35169946]
 ...
 [0.02916667 0.78156473 0.71022725 ... 0.38559321 0.86450659 0.09970969]
 [0.17916667 0.76658921 0.71266236 ... 0.29237287 0.84462441 0.11727551]
 [0.29166667 0.7384457  0.6712662  ... 0.30084744 0.82768774 0.1120159 ]]
```

- **k=19**

```
[[0.5515873 0.475    0.    ... 0.05154639 0.5456349 0.32532112]
 [0.74206349 0.50833333 0.    ... 0.05596466 0.53858024 0.35865966]
 [0.41269841 0.53333333 0.    ... 0.05007364 0.5326279 0.35169946]
 ...
 [0.59920635 0.02916667 0.    ... 0.86450659 0.2546684 0.09970969]
 [0.32539683 0.17916667 0.    ... 0.84462441 0.23450433 0.11727551]
 [0.45238095 0.29166667 0.    ... 0.82768774 0.26421982 0.1120159 ]]
```

High variance for larger k so the features.