Data Mining
Lab 01

Names:

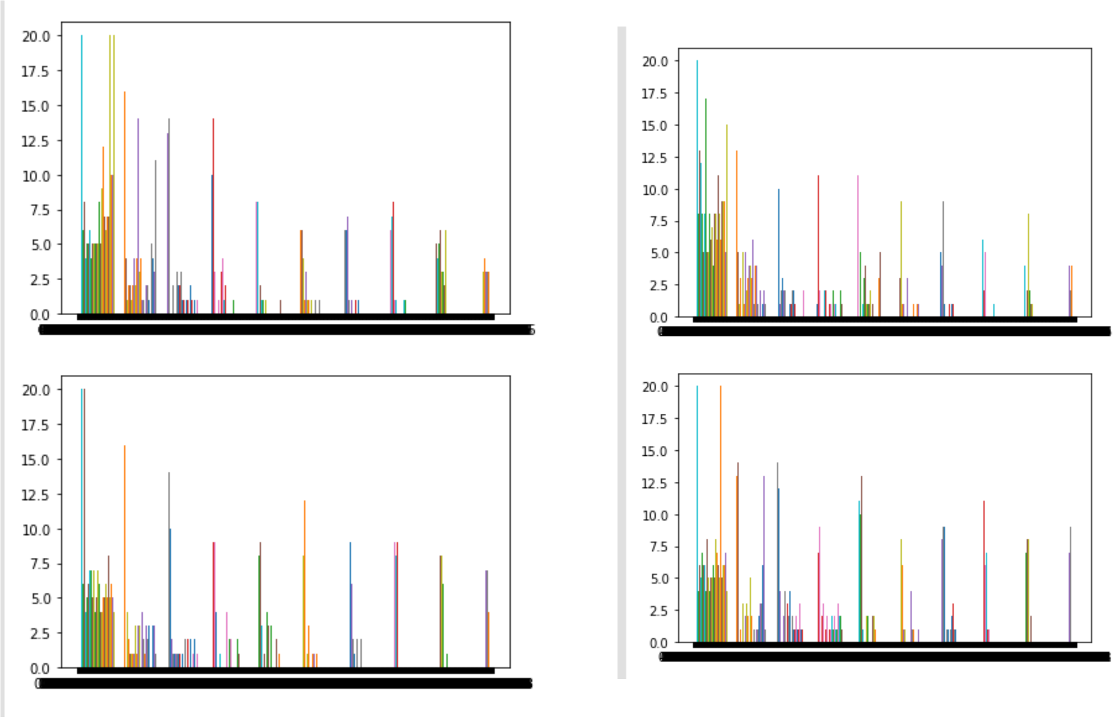
- Ahmed Mohamed EL-Bawab (08)
- Khalil Ismail Khalil (23)

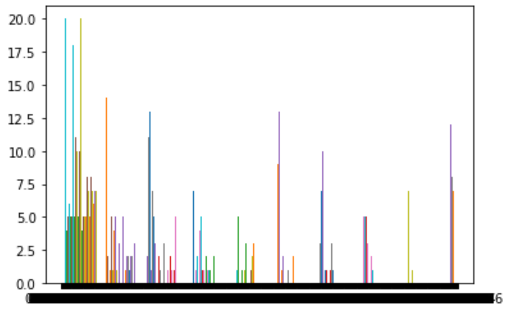
- Number of instances = 2310 instances.
- Number of attributes = 19 attributes.
- Number of classes = 7 classes.

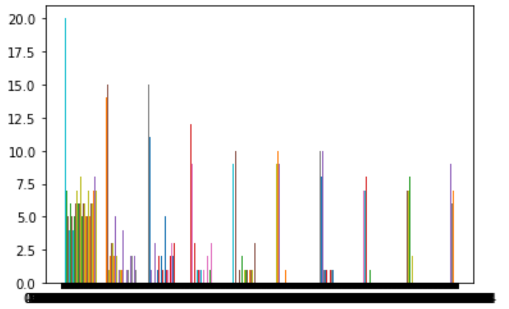
1)Data Exploration:

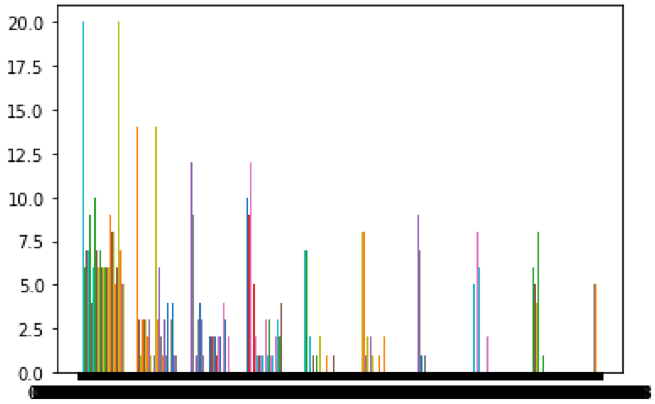
1)Histograms:

1)



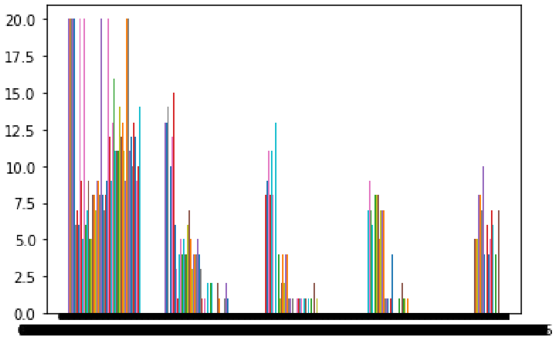

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

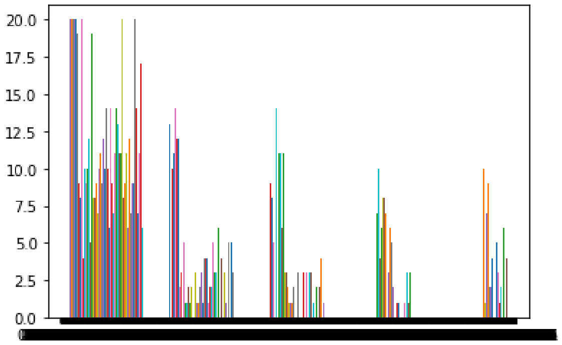

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

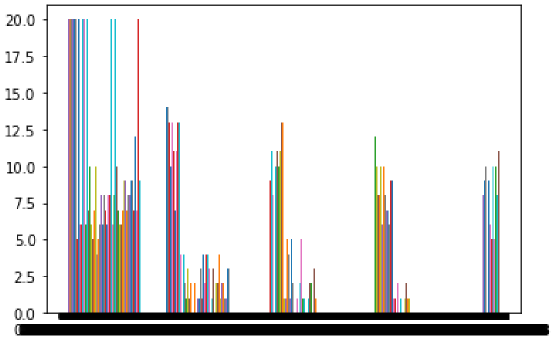

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

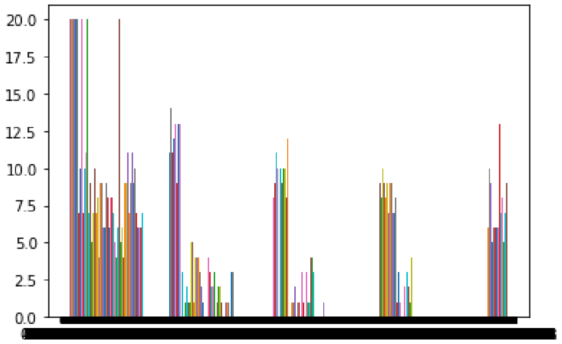
2)

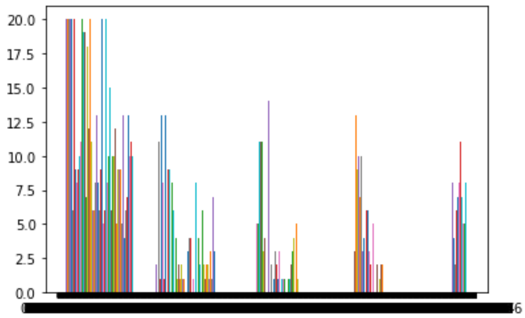
- at bins = 5,

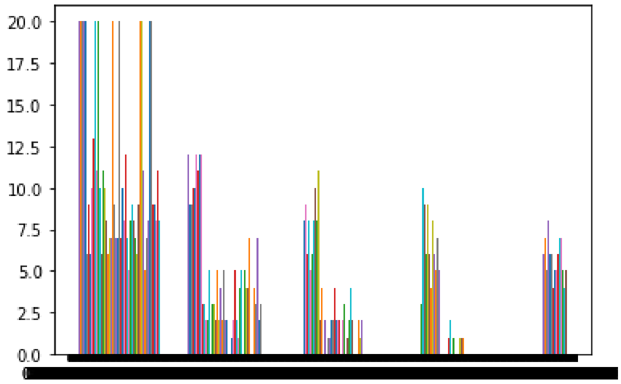

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

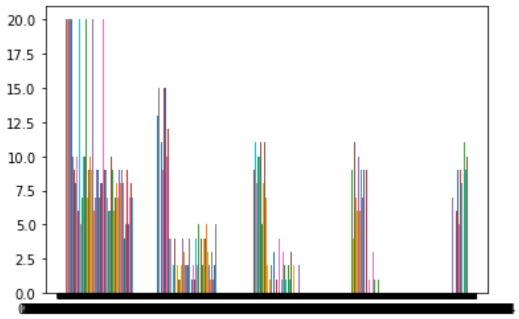

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0


20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

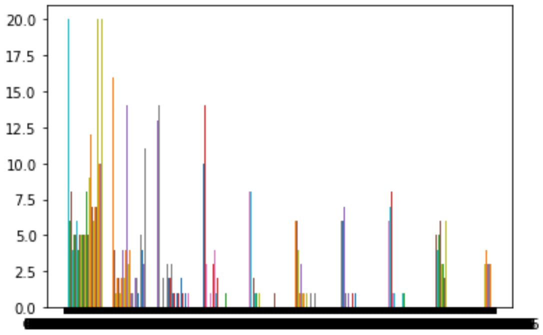

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

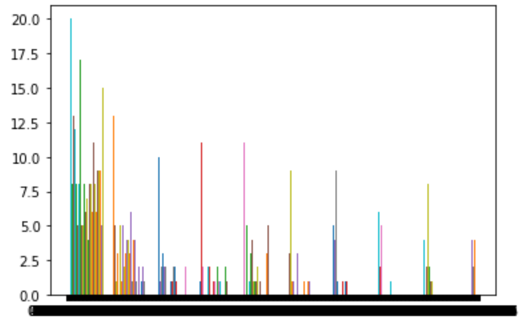

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

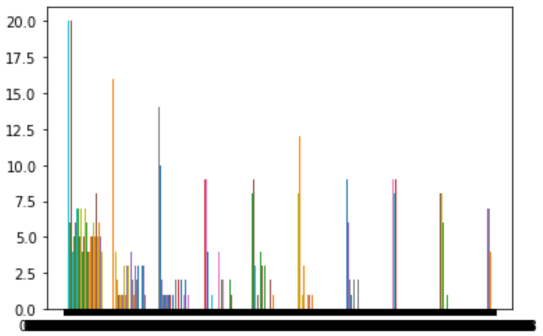

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

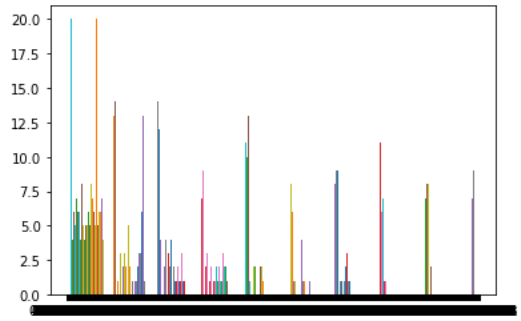

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

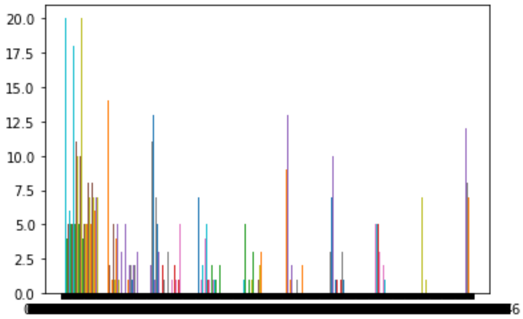
- at bins = 10,

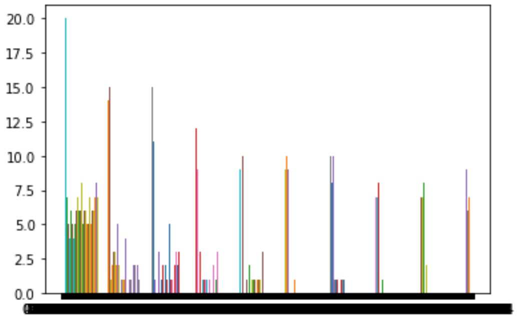

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

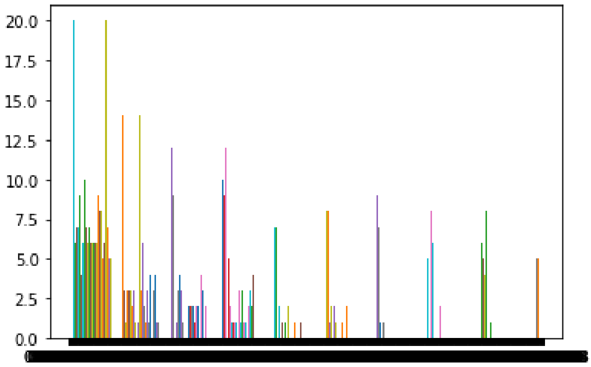

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0


20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

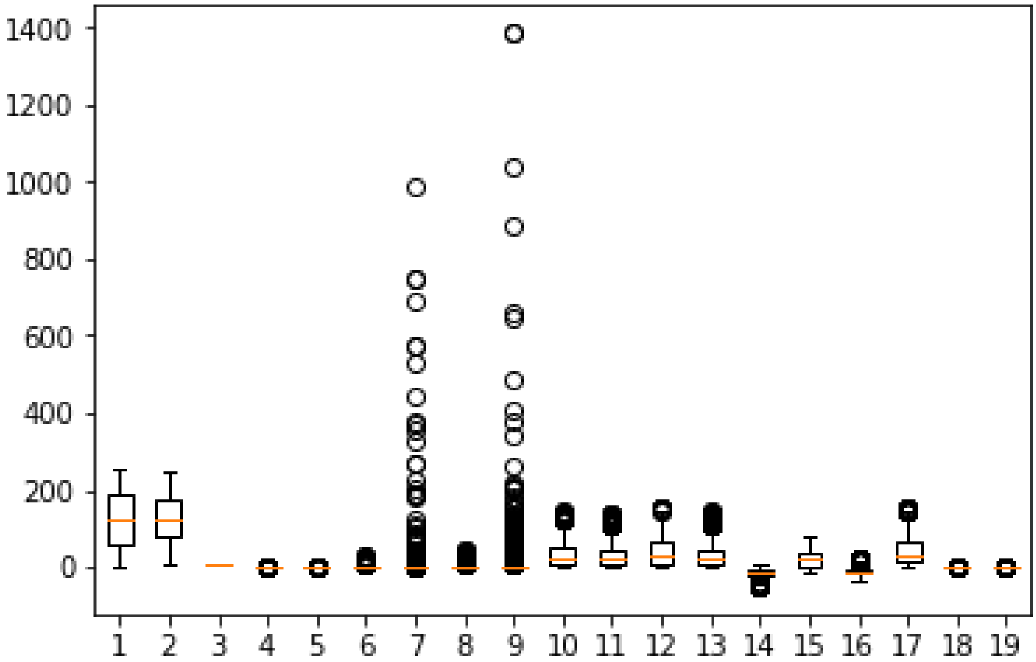

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0


20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0


20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

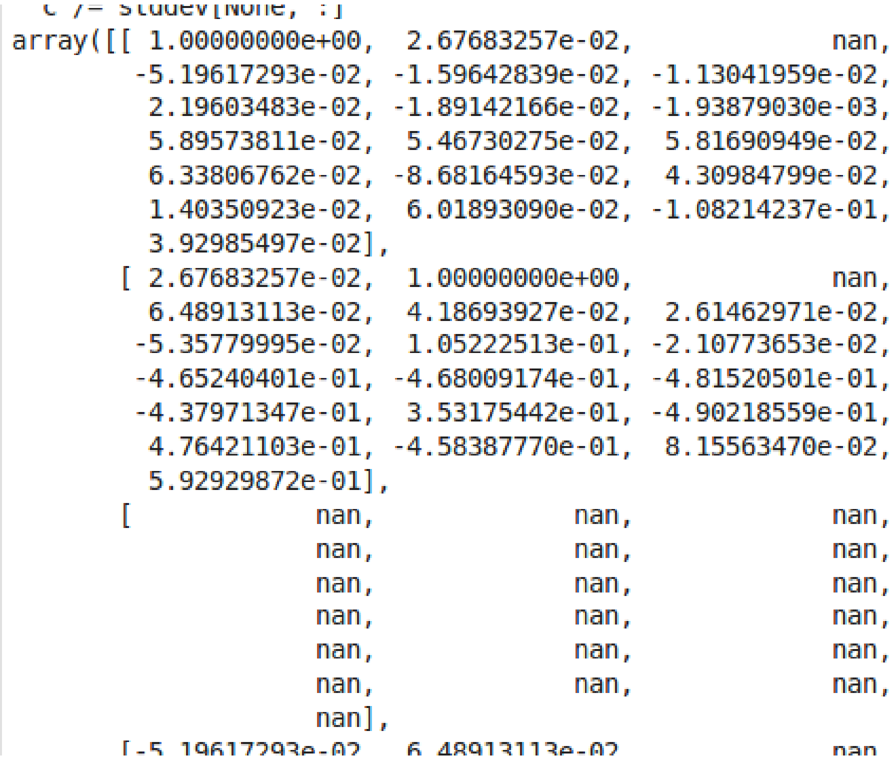

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

2)Boxplots:

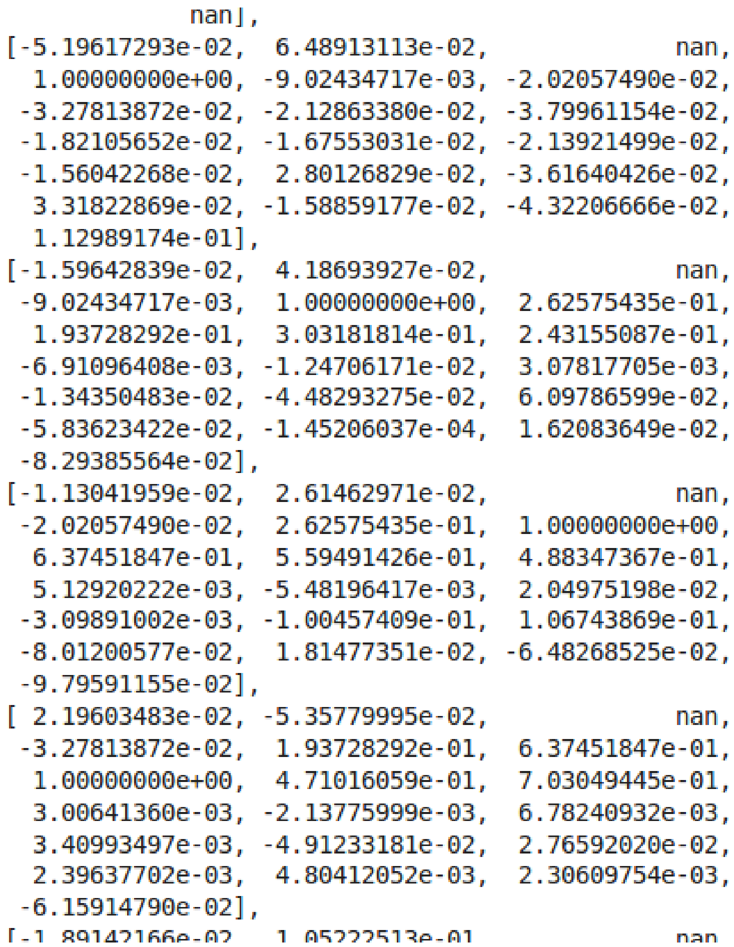

1400
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

3)Correlation Matrix:

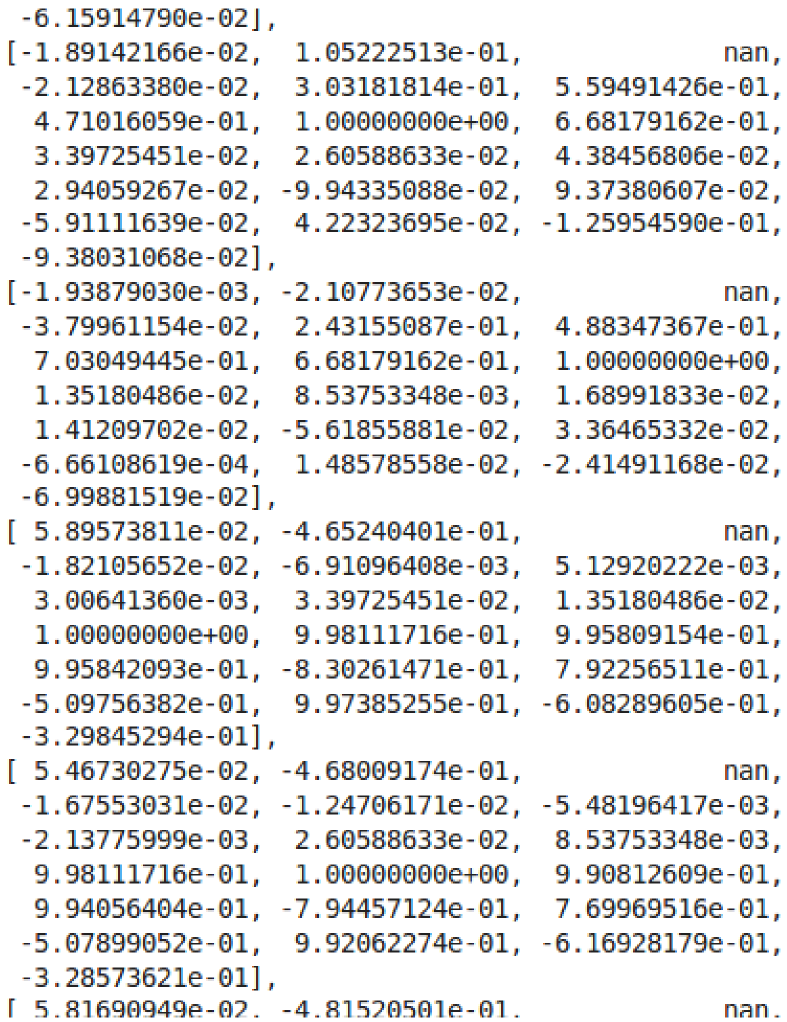
1)dxd symmetric matrix(d=19):



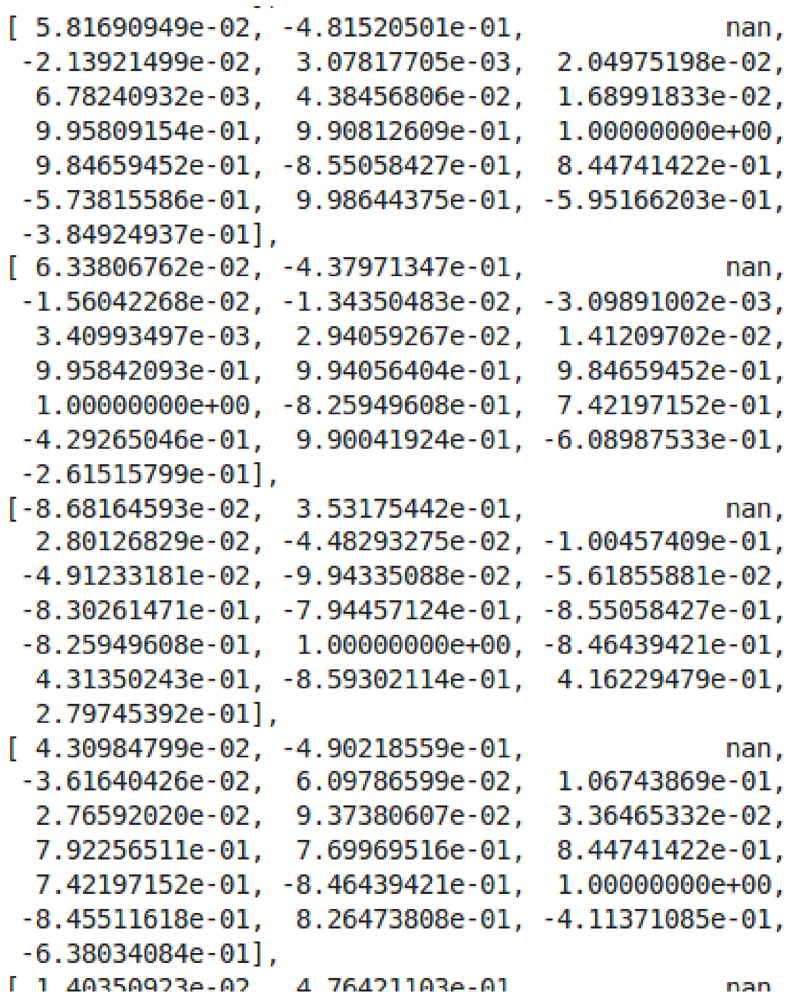
```
  c /= stdev[None, :]
array([[ 1.00000000e+00,  2.67683257e-02,             nan,
        -5.19617293e-02, -1.59642839e-02, -1.13041959e-02,
         2.19603483e-02, -1.89142166e-02, -1.93879030e-03,
         5.89573811e-02,  5.46730275e-02,  5.81690949e-02,
         6.33806762e-02, -8.68164593e-02,  4.30984799e-02,
         1.40350923e-02,  6.01893090e-02, -1.08214237e-01,
         3.92985497e-02],
       [ 2.67683257e-02,  1.00000000e+00,             nan,
         6.48913113e-02,  4.18693927e-02,  2.61462971e-02,
        -5.35779995e-02,  1.05222513e-01, -2.10773653e-02,
        -4.65240401e-01, -4.68009174e-01, -4.81520501e-01,
        -4.37971347e-01,  3.53175442e-01, -4.90218559e-01,
         4.76421103e-01, -4.58387770e-01,  8.15563470e-02,
         5.92929872e-01],
       [             nan,             nan,             nan,
                     nan,             nan,             nan,
                     nan,             nan,             nan,
                     nan,             nan,             nan,
                     nan,             nan,             nan,
                     nan,             nan,             nan,
                     nan],
       [-5.19617293e-02,  6.48913113e-02,             nan,
```



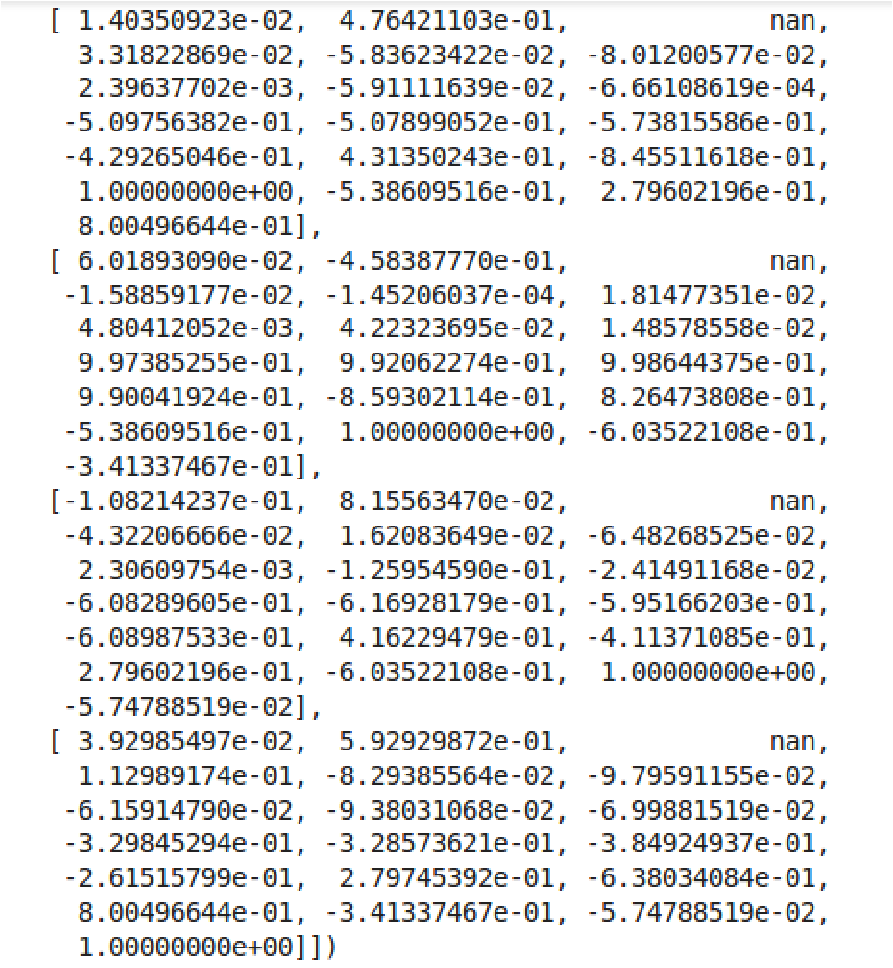
```
                     nan],
       [-5.19617293e-02,  6.48913113e-02,             nan,
         1.00000000e+00, -9.02434717e-03, -2.02057490e-02,
        -3.27813872e-02, -2.12863380e-02, -3.79961154e-02,
        -1.82105652e-02, -1.67553031e-02, -2.13921499e-02,
        -1.56042268e-02,  2.80126829e-02, -3.61640426e-02,
         3.31822869e-02, -1.58859177e-02, -4.32206666e-02,
         1.12989174e-01],
       [-1.59642839e-02,  4.18693927e-02,             nan,
        -9.02434717e-03,  1.00000000e+00,  2.62575435e-01,
         1.93728292e-01,  3.03181814e-01,  2.43155087e-01,
        -6.91096408e-03, -1.24706171e-02,  3.07817705e-03,
        -1.34350483e-02, -4.48293275e-02,  6.09786599e-02,
        -5.83623422e-02, -1.45206037e-04,  1.62083649e-02,
        -8.29385564e-02],
       [-1.13041959e-02,  2.61462971e-02,             nan,
        -2.02057490e-02,  2.62575435e-01,  1.00000000e+00,
         6.37451847e-01,  5.59491426e-01,  4.88347367e-01,
         5.12920222e-03, -5.48196417e-03,  2.04975198e-02,
        -3.09891002e-03, -1.00457409e-01,  1.06743869e-01,
        -8.01200577e-02,  1.81477351e-02, -6.48268525e-02,
        -9.79591155e-02],
       [ 2.19603483e-02, -5.35779995e-02,             nan,
        -3.27813872e-02,  1.93728292e-01,  6.37451847e-01,
         1.00000000e+00,  4.71016059e-01,  7.03049445e-01,
         3.00641360e-03, -2.13775999e-03,  6.78240932e-03,
         3.40993497e-03, -4.91233181e-02,  2.76592020e-02,
         2.39637702e-03,  4.80412052e-03,  2.30609754e-03,
        -6.15914790e-02],
       [-1.89142166e-02,  1.05222513e-01,             nan,
```



```
 -6.15914790e-02],
[-1.89142166e-02,  1.05222513e-01,             nan,
 -2.12863380e-02,  3.03181814e-01,  5.59491426e-01,
  4.71016059e-01,  1.00000000e+00,  6.68179162e-01,
  3.39725451e-02,  2.60588633e-02,  4.38456806e-02,
  2.94059267e-02, -9.94335088e-02,  9.37380607e-02,
 -5.91111639e-02,  4.22323695e-02, -1.25954590e-01,
 -9.38031068e-02],
[-1.93879030e-03, -2.10773653e-02,             nan,
 -3.79961154e-02,  2.43155087e-01,  4.88347367e-01,
  7.03049445e-01,  6.68179162e-01,  1.00000000e+00,
  1.35180486e-02,  8.53753348e-03,  1.68991833e-02,
  1.41209702e-02, -5.61855881e-02,  3.36465332e-02,
 -6.66108619e-04,  1.48578558e-02, -2.41491168e-02,
 -6.99881519e-02],
[ 5.89573811e-02, -4.65240401e-01,             nan,
 -1.82105652e-02, -6.91096408e-03,  5.12920222e-03,
  3.00641360e-03,  3.39725451e-02,  1.35180486e-02,
  1.00000000e+00,  9.98111716e-01,  9.95809154e-01,
  9.95842093e-01, -8.30261471e-01,  7.92256511e-01,
 -5.09756382e-01,  9.97385255e-01, -6.08289605e-01,
 -3.29845294e-01],
[ 5.46730275e-02, -4.68009174e-01,             nan,
 -1.67553031e-02, -1.24706171e-02, -5.48196417e-03,
 -2.13775999e-03,  2.60588633e-02,  8.53753348e-03,
  9.98111716e-01,  1.00000000e+00,  9.90812609e-01,
  9.94056404e-01, -7.94457124e-01,  7.69969516e-01,
 -5.07899052e-01,  9.92062274e-01, -6.16928179e-01,
 -3.28573621e-01],
[ 5.81690949e-02, -4.81520501e-01,             nan,
```

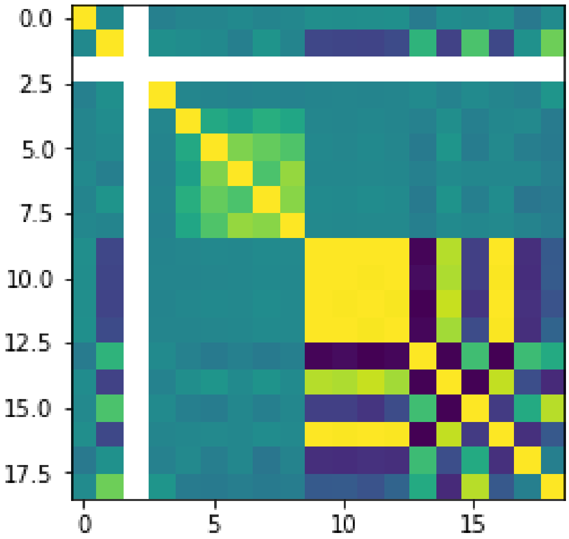


```
[ 5.81690949e-02, -4.81520501e-01,             nan,
 -2.13921499e-02,  3.07817705e-03,  2.04975198e-02,
  6.78240932e-03,  4.38456806e-02,  1.68991833e-02,
  9.95809154e-01,  9.90812609e-01,  1.00000000e+00,
  9.84659452e-01, -8.55058427e-01,  8.44741422e-01,
 -5.73815586e-01,  9.98644375e-01, -5.95166203e-01,
 -3.84924937e-01],
[ 6.33806762e-02, -4.37971347e-01,             nan,
 -1.56042268e-02, -1.34350483e-02, -3.09891002e-03,
  3.40993497e-03,  2.94059267e-02,  1.41209702e-02,
  9.95842093e-01,  9.94056404e-01,  9.84659452e-01,
  1.00000000e+00, -8.25949608e-01,  7.42197152e-01,
 -4.29265046e-01,  9.90041924e-01, -6.08987533e-01,
 -2.61515799e-01],
[-8.68164593e-02,  3.53175442e-01,             nan,
  2.80126829e-02, -4.48293275e-02, -1.00457409e-01,
 -4.91233181e-02, -9.94335088e-02, -5.61855881e-02,
 -8.30261471e-01, -7.94457124e-01, -8.55058427e-01,
 -8.25949608e-01,  1.00000000e+00, -8.46439421e-01,
  4.31350243e-01, -8.59302114e-01,  4.16229479e-01,
  2.79745392e-01],
[ 4.30984799e-02, -4.90218559e-01,             nan,
 -3.61640426e-02,  6.09786599e-02,  1.06743869e-01,
  2.76592020e-02,  9.37380607e-02,  3.36465332e-02,
  7.92256511e-01,  7.69969516e-01,  8.44741422e-01,
  7.42197152e-01, -8.46439421e-01,  1.00000000e+00,
 -8.45511618e-01,  8.26473808e-01, -4.11371085e-01,
 -6.38034084e-01],
[ 1.40350923e-02,  4.76421103e-01,             nan
```



```
[ 1.40350923e-02,  4.76421103e-01,             nan,
  3.31822869e-02, -5.83623422e-02, -8.01200577e-02,
  2.39637702e-03, -5.91111639e-02, -6.66108619e-04,
 -5.09756382e-01, -5.07899052e-01, -5.73815586e-01,
 -4.29265046e-01,  4.31350243e-01, -8.45511618e-01,
  1.00000000e+00, -5.38609516e-01,  2.79602196e-01,
  8.00496644e-01],
[ 6.01893090e-02, -4.58387770e-01,             nan,
 -1.58859177e-02, -1.45206037e-04,  1.81477351e-02,
  4.80412052e-03,  4.22323695e-02,  1.48578558e-02,
  9.97385255e-01,  9.92062274e-01,  9.98644375e-01,
  9.90041924e-01, -8.59302114e-01,  8.26473808e-01,
 -5.38609516e-01,  1.00000000e+00, -6.03522108e-01,
 -3.41337467e-01],
[-1.08214237e-01,  8.15563470e-02,             nan,
 -4.32206666e-02,  1.62083649e-02, -6.48268525e-02,
  2.30609754e-03, -1.25954590e-01, -2.41491168e-02,
 -6.08289605e-01, -6.16928179e-01, -5.95166203e-01,
 -6.08987533e-01,  4.16229479e-01, -4.11371085e-01,
  2.79602196e-01, -6.03522108e-01,  1.00000000e+00,
 -5.74788519e-02],
[ 3.92985497e-02,  5.92929872e-01,             nan,
  1.12989174e-01, -8.29385564e-02, -9.79591155e-02,
 -6.15914790e-02, -9.38031068e-02, -6.99881519e-02,
 -3.29845294e-01, -3.28573621e-01, -3.84924937e-01,
 -2.61515799e-01,  2.79745392e-01, -6.38034084e-01,
  8.00496644e-01, -3.41337467e-01, -5.74788519e-02,
  1.00000000e+00]])
```

2)

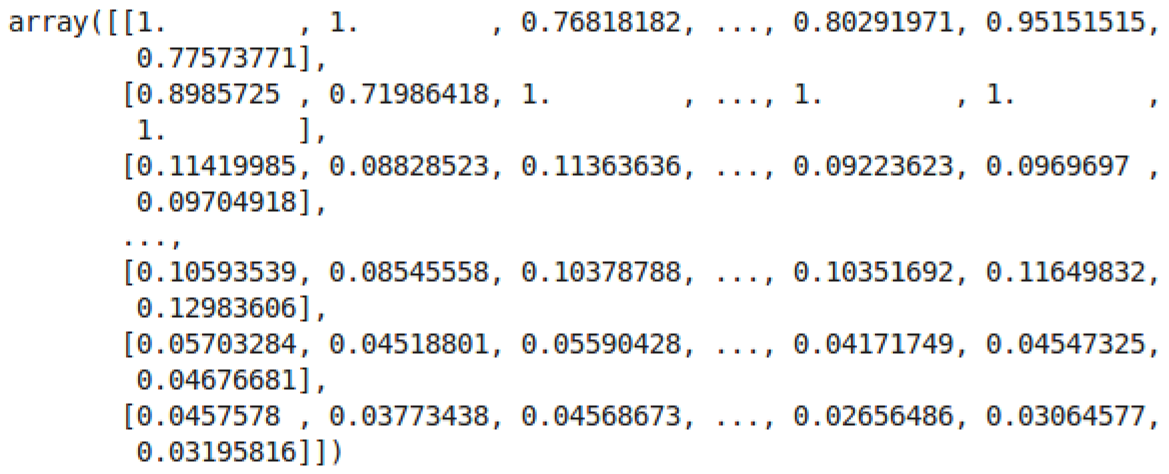


2)Preprocessing:

1)Normalization:

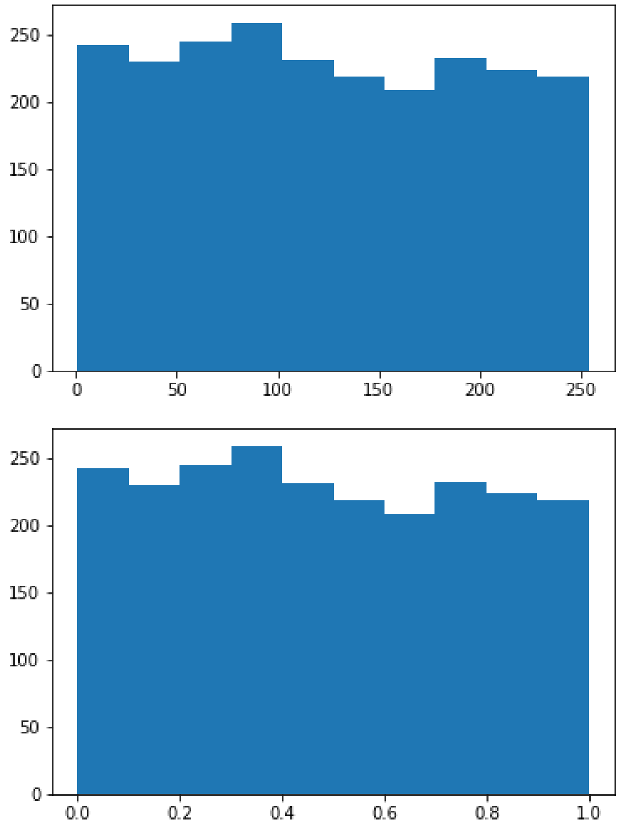
1)Min-max scaler:

- normalizedMinmaxList:

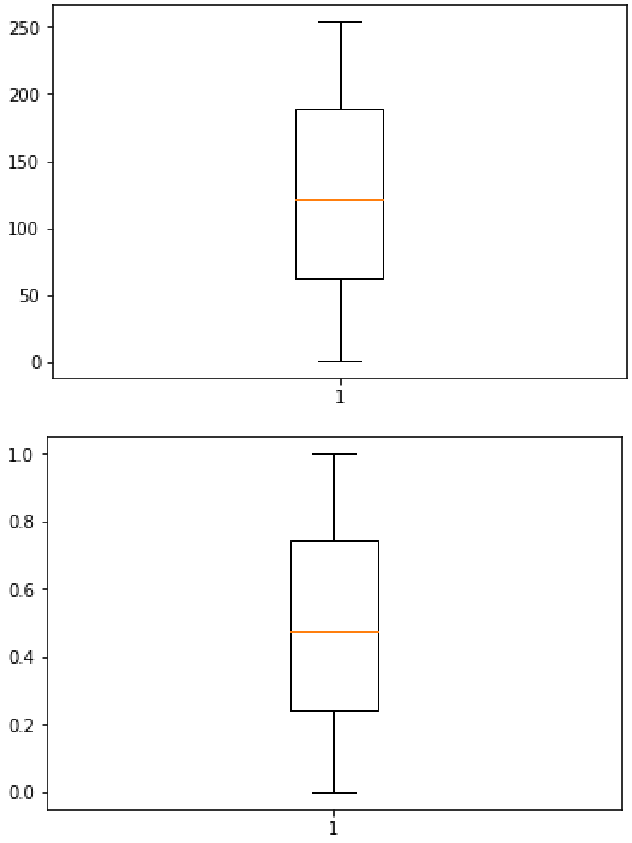


```
array([[1.        , 1.        , 0.76818182, ..., 0.80291971, 0.95151515,
        0.77573771],
       [0.8985725 , 0.71986418, 1.        , ..., 1.        , 1.        ,
        1.        ],
       [0.11419985, 0.08828523, 0.11363636, ..., 0.09223623, 0.0969697 ,
        0.09704918],
       ...,
       [0.10593539, 0.08545558, 0.10378788, ..., 0.10351692, 0.11649832,
        0.12983606],
       [0.05703284, 0.04518801, 0.05590428, ..., 0.04171749, 0.04547325,
        0.04676681],
       [0.0457578 , 0.03773438, 0.04568673, ..., 0.02656486, 0.03064577,
        0.03195816]])
```

- Histograms before and after normalization:

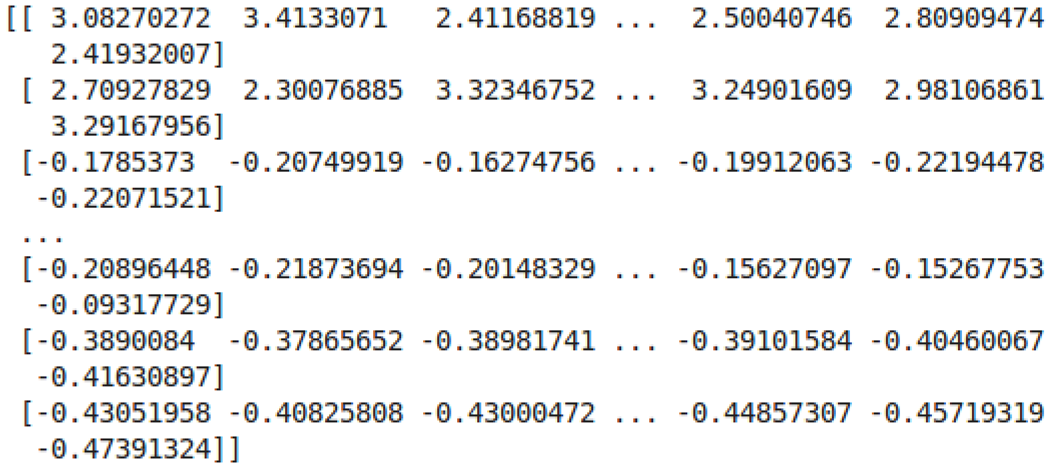


- Boxplots before and after normalization:



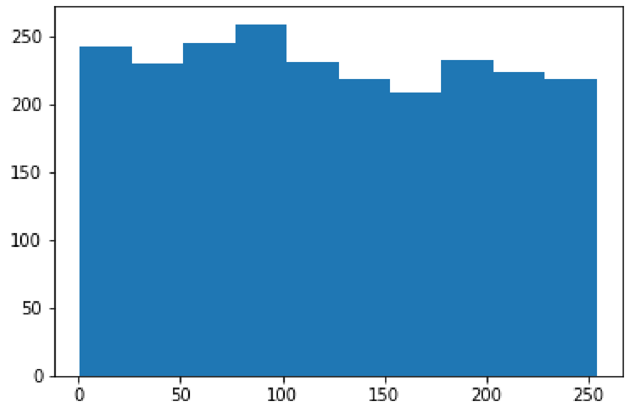
2)Z-score:

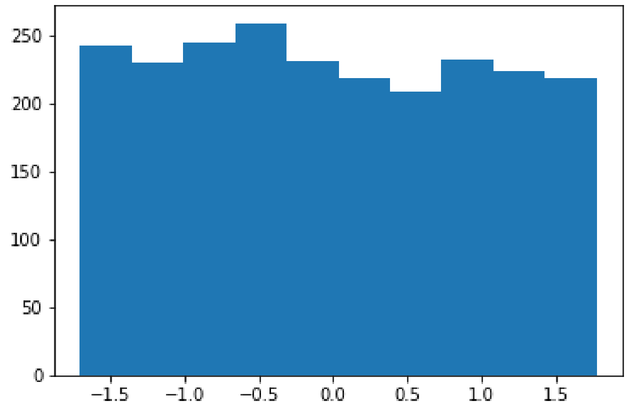
- normalizedZscoreList:



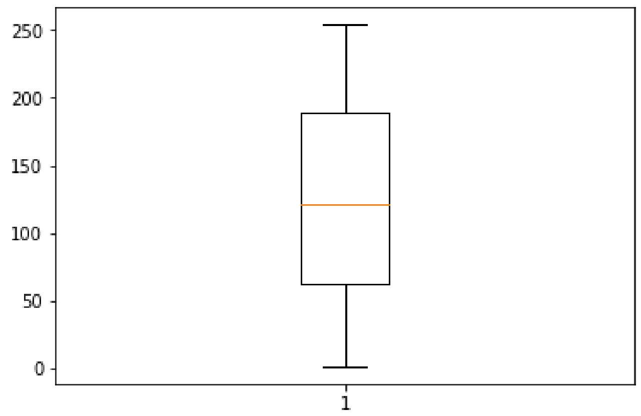
```
[[ 3.08270272  3.4133071   2.41168819 ...  2.50040746  2.80909474
   2.41932007]
 [ 2.70927829  2.30076885  3.32346752 ...  3.24901609  2.98106861
   3.29167956]
 [-0.1785373  -0.20749919 -0.16274756 ... -0.19912063 -0.22194478
  -0.22071521]
 ...
 [-0.20896448 -0.21873694 -0.20148329 ... -0.15627097 -0.15267753
  -0.09317729]
 [-0.3890084  -0.37865652 -0.38981741 ... -0.39101584 -0.40460067
  -0.41630897]
 [-0.43051958 -0.40825808 -0.43000472 ... -0.44857307 -0.45719319
  -0.47391324]]
```

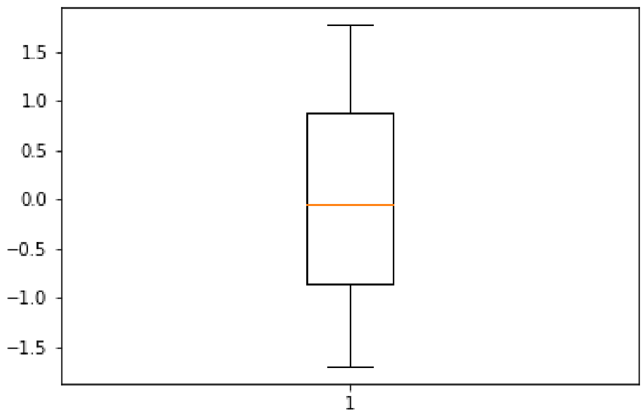
- Histograms before and after normalization:





- Boxplots before and after normalization:



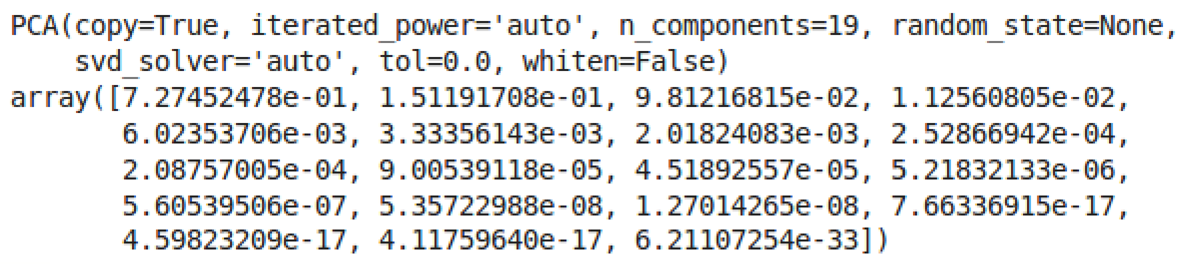


1)Dimensionality reduction:

1)Feature Projection:

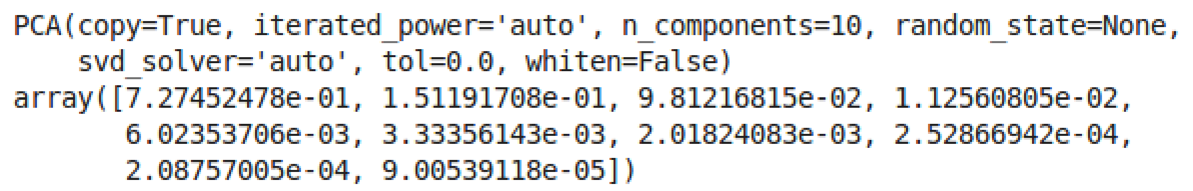
- Variance_ratio after applying PCA with:

components = 19,



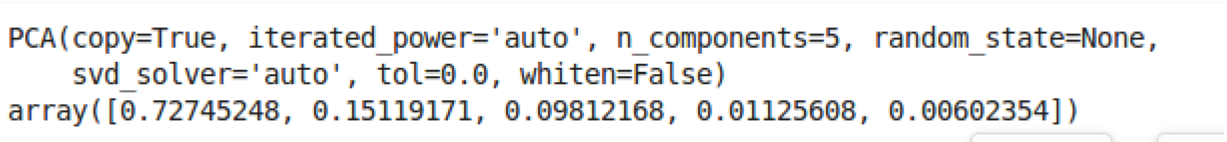
```
PCA(copy=True, iterated_power='auto', n_components=19, random_state=None,
    svd_solver='auto', tol=0.0, whiten=False)
array([7.27452478e-01, 1.51191708e-01, 9.81216815e-02, 1.12560805e-02,
       6.02353706e-03, 3.33356143e-03, 2.01824083e-03, 2.52866942e-04,
       2.08757005e-04, 9.00539118e-05, 4.51892557e-05, 5.21832133e-06,
       5.60539506e-07, 5.35722988e-08, 1.27014265e-08, 7.66336915e-17,
       4.59823209e-17, 4.11759640e-17, 6.21107254e-33])
```

components = 10,



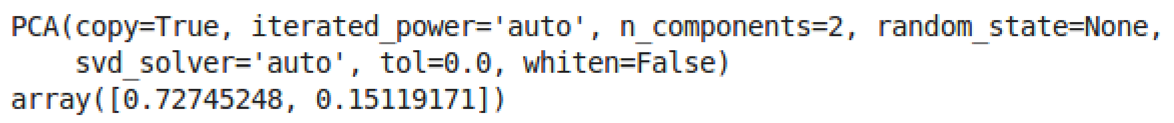
```
PCA(copy=True, iterated_power='auto', n_components=10, random_state=None,
    svd_solver='auto', tol=0.0, whiten=False)
array([7.27452478e-01, 1.51191708e-01, 9.81216815e-02, 1.12560805e-02,
       6.02353706e-03, 3.33356143e-03, 2.01824083e-03, 2.52866942e-04,
       2.08757005e-04, 9.00539118e-05])
```

components = 5,



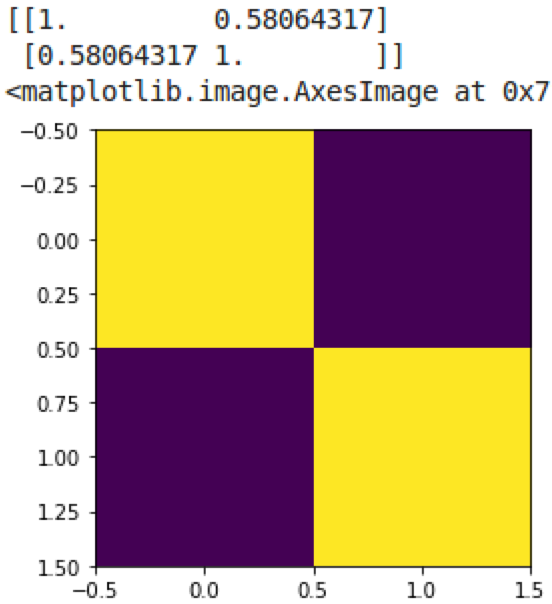
```
PCA(copy=True, iterated_power='auto', n_components=5, random_state=None,
    svd_solver='auto', tol=0.0, whiten=False)
array([0.72745248, 0.15119171, 0.09812168, 0.01125608, 0.00602354])
```

components = 2,



```
PCA(copy=True, iterated_power='auto', n_components=2, random_state=None,
    svd_solver='auto', tol=0.0, whiten=False)
array([0.72745248, 0.15119171])
```

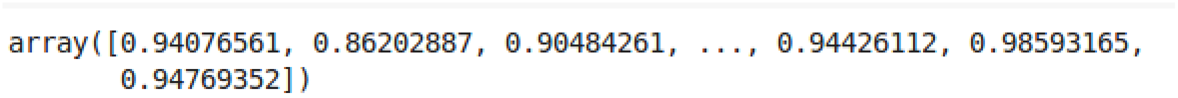
- Correlation matrix of your dataset after applying PCA:



```
[[1.         0.58064317]
 [0.58064317 1.        ]]
<matplotlib.image.AxesImage at 0x7f83d9d08908>
```

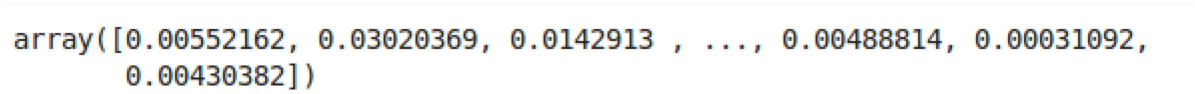
2)Feature selection:

- Pvalues after applying SelectKBest:



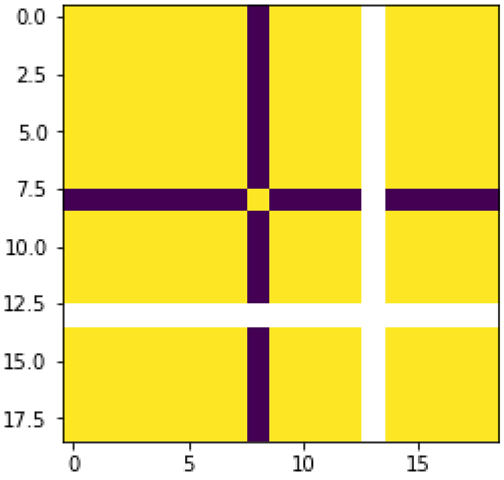
```
array([0.94076561, 0.86202887, 0.90484261, ..., 0.94426112, 0.98593165,
       0.94769352])
```

- Scores after applying SelectKBest:



```
array([0.00552162, 0.03020369, 0.0142913 , ..., 0.00488814, 0.00031092,
       0.00430382])
```

- Correlation matrix of your dataset after applying SelectKBest:



0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0
5
10
15