ECE-657A Data and Knowledge Modeling and Analysis
Winter 2018

Assignment 1
Data Cleaning and Dimensionality Reduction

Prepared by

Samina Islam Eva
20704504
Souradip Dey
20741206
Md Rubayatur Rahim Bhuyian
20743420

UNIVERSITY OF
WATERLOO

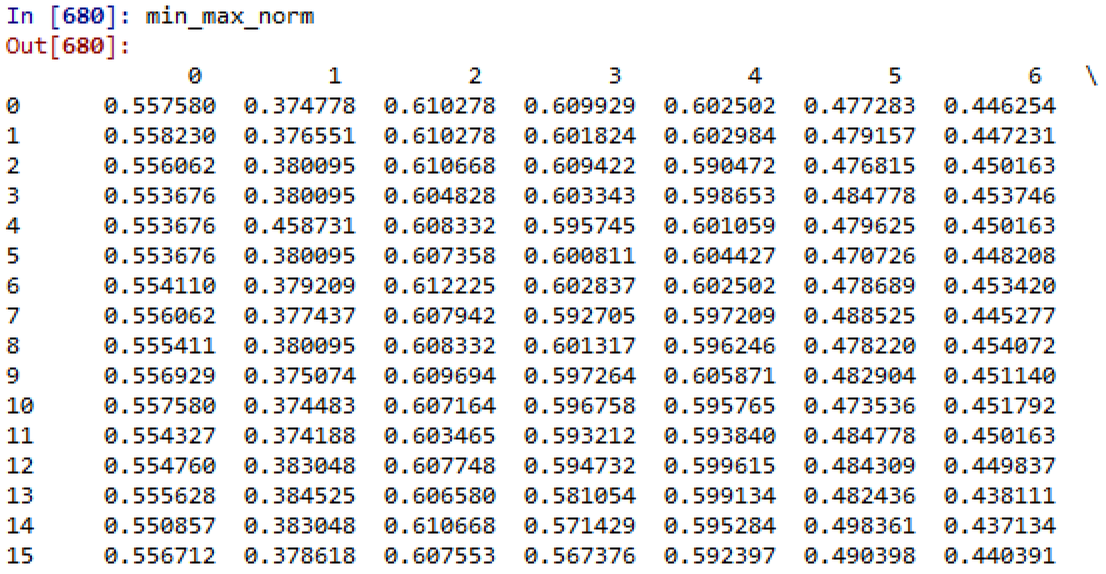
Department of Electrical & Computer Engineering

Part 1:

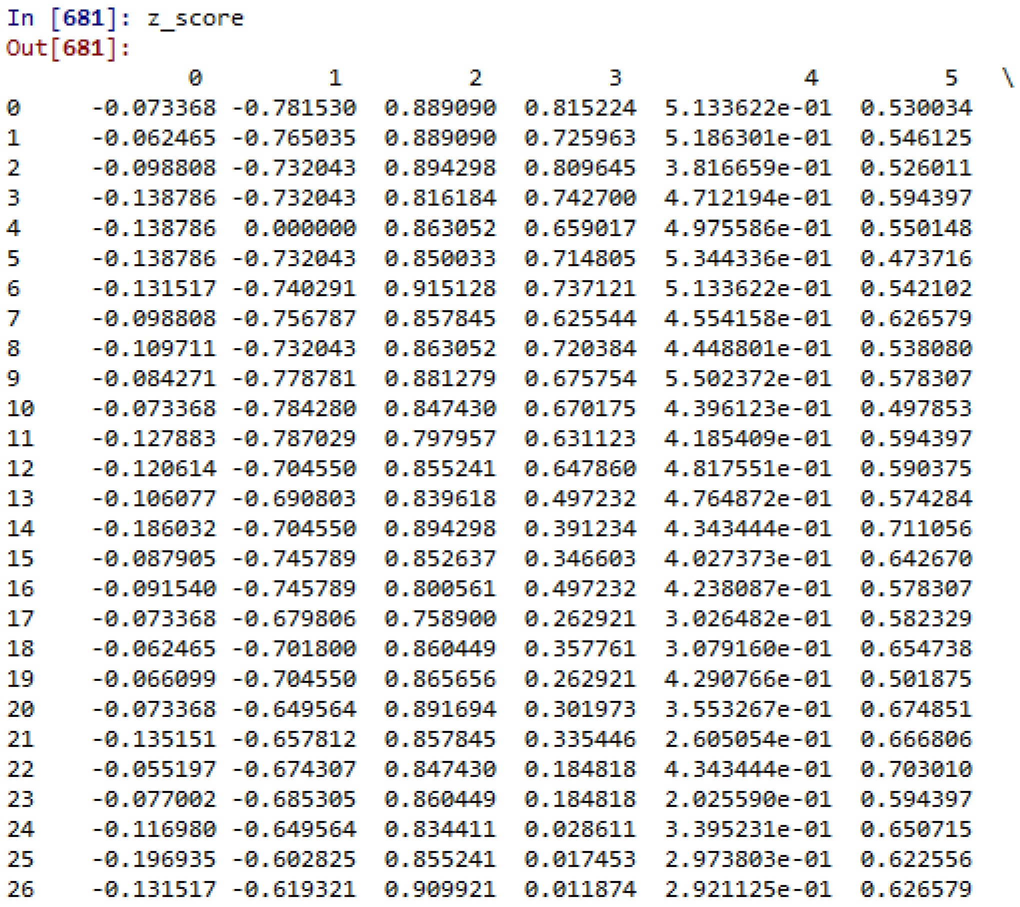
1. In the DatasetA, there are some attributes having missing values stated as "NAN" and also, for some attributes there are some outliers.

2. The missing values can be replaced by interpolation, linear regression, moving average. Here we used mean value to replace the outliers and missing values. Because of that the property of data (mean and standard deviation) remains same.

3. Normalizing the data using Min-Max normalization and Z_score normalization.

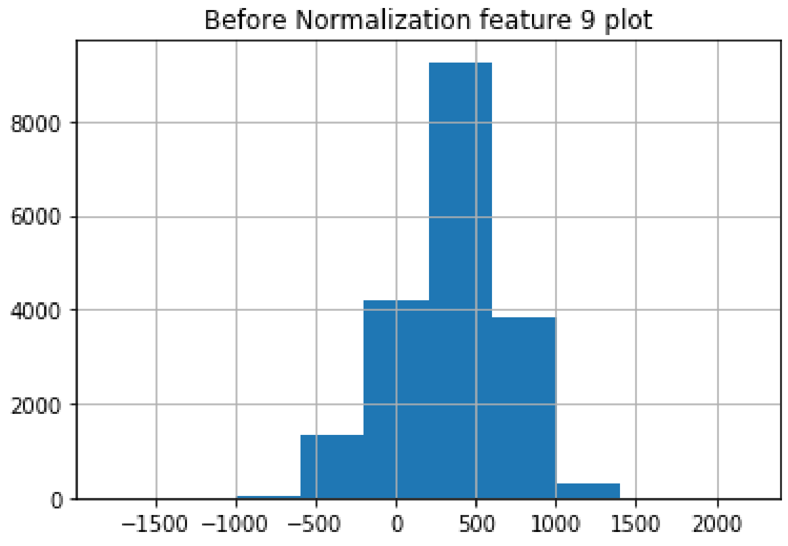


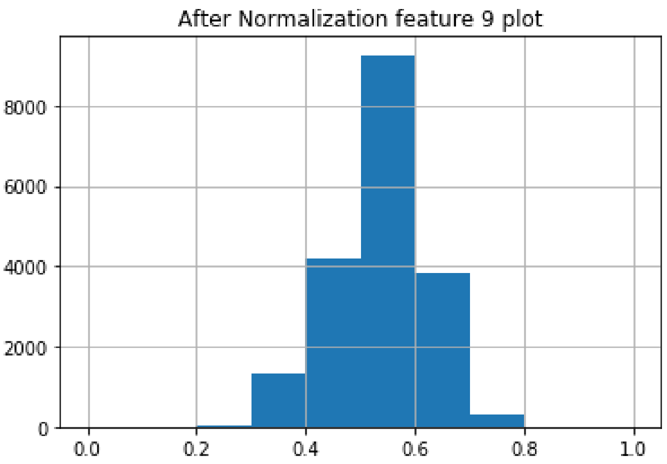
```
In [680]: min_max_norm
Out[680]:
             0         1         2         3         4         5         6  \
0     0.557580  0.374778  0.610278  0.609929  0.602502  0.477283  0.446254
1     0.558230  0.376551  0.610278  0.601824  0.602984  0.479157  0.447231
2     0.556062  0.380095  0.610668  0.609422  0.590472  0.476815  0.450163
3     0.553676  0.380095  0.604828  0.603343  0.598653  0.484778  0.453746
4     0.553676  0.458731  0.608332  0.595745  0.601059  0.479625  0.450163
5     0.553676  0.380095  0.607358  0.600811  0.604427  0.470726  0.448208
6     0.554110  0.379209  0.612225  0.602837  0.602502  0.478689  0.453420
7     0.556062  0.377437  0.607942  0.592705  0.597209  0.488525  0.445277
8     0.555411  0.380095  0.608332  0.601317  0.596246  0.478220  0.454072
9     0.556929  0.375074  0.609694  0.597264  0.605871  0.482904  0.451140
10    0.557580  0.374483  0.607164  0.596758  0.595765  0.473536  0.451792
11    0.554327  0.374188  0.603465  0.593212  0.593840  0.484778  0.450163
12    0.554760  0.383048  0.607748  0.594732  0.599615  0.484309  0.449837
13    0.555628  0.384525  0.606580  0.581054  0.599134  0.482436  0.438111
14    0.550857  0.383048  0.610668  0.571429  0.595284  0.498361  0.437134
15    0.556712  0.378618  0.607553  0.567376  0.592397  0.490398  0.440391
```



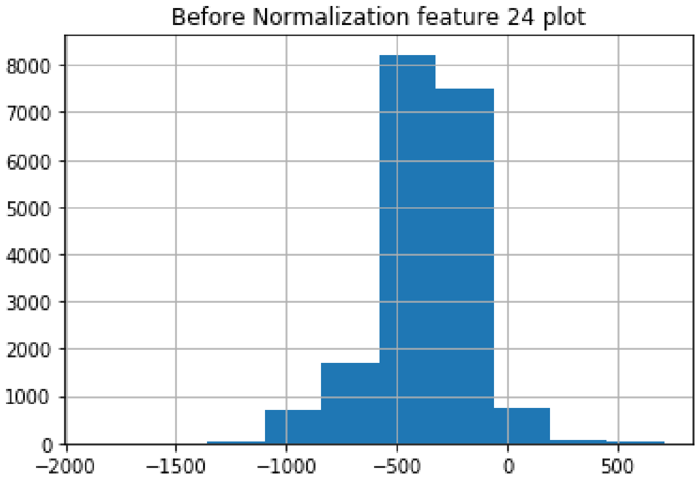
```
In [681]: z_score
Out[681]:
           0         1         2         3             4         5  \
0  -0.073368 -0.781530  0.889090  0.815224  5.133622e-01  0.530034
1  -0.062465 -0.765035  0.889090  0.725963  5.186301e-01  0.546125
2  -0.098808 -0.732043  0.894298  0.809645  3.816659e-01  0.526011
3  -0.138786 -0.732043  0.816184  0.742700  4.712194e-01  0.594397
4  -0.138786  0.000000  0.863052  0.659017  4.975586e-01  0.550148
5  -0.138786 -0.732043  0.850033  0.714805  5.344336e-01  0.473716
6  -0.131517 -0.740291  0.915128  0.737121  5.133622e-01  0.542102
7  -0.098808 -0.756787  0.857845  0.625544  4.554158e-01  0.626579
8  -0.109711 -0.732043  0.863052  0.720384  4.448801e-01  0.538080
9  -0.084271 -0.778781  0.881279  0.675754  5.502372e-01  0.578307
10 -0.073368 -0.784280  0.847430  0.670175  4.396123e-01  0.497853
11 -0.127883 -0.787029  0.797957  0.631123  4.185409e-01  0.594397
12 -0.120614 -0.704550  0.855241  0.647860  4.817551e-01  0.590375
13 -0.106077 -0.690803  0.839618  0.497232  4.764872e-01  0.574284
14 -0.186032 -0.704550  0.894298  0.391234  4.343444e-01  0.711056
15 -0.087905 -0.745789  0.852637  0.346603  4.027373e-01  0.642670
16 -0.091540 -0.745789  0.800561  0.497232  4.238087e-01  0.578307
17 -0.073368 -0.679806  0.758900  0.262921  3.026482e-01  0.582329
18 -0.062465 -0.701800  0.860449  0.357761  3.079160e-01  0.654738
19 -0.066099 -0.704550  0.865656  0.262921  4.290766e-01  0.501875
20 -0.073368 -0.649564  0.891694  0.301973  3.553267e-01  0.674851
21 -0.135151 -0.657812  0.857845  0.335446  2.605054e-01  0.666806
22 -0.055197 -0.674307  0.847430  0.184818  4.343444e-01  0.703010
23 -0.077002 -0.685305  0.860449  0.184818  2.025590e-01  0.594397
24 -0.116980 -0.649564  0.834411  0.028611  3.395231e-01  0.650715
25 -0.196935 -0.602825  0.855241  0.017453  2.973803e-01  0.622556
26 -0.131517 -0.619321  0.909921  0.011874  2.921125e-01  0.626579
```

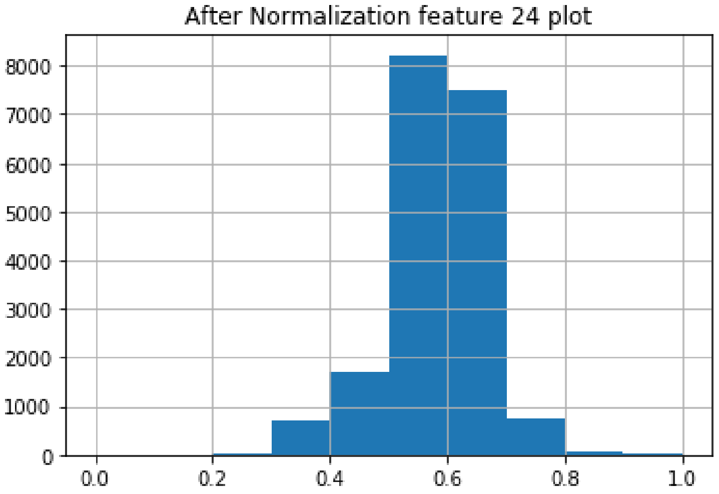
The before and after plot of feature 9 and feature 24 are given below.





Comment: After normalization, the distribution of the data remains same.



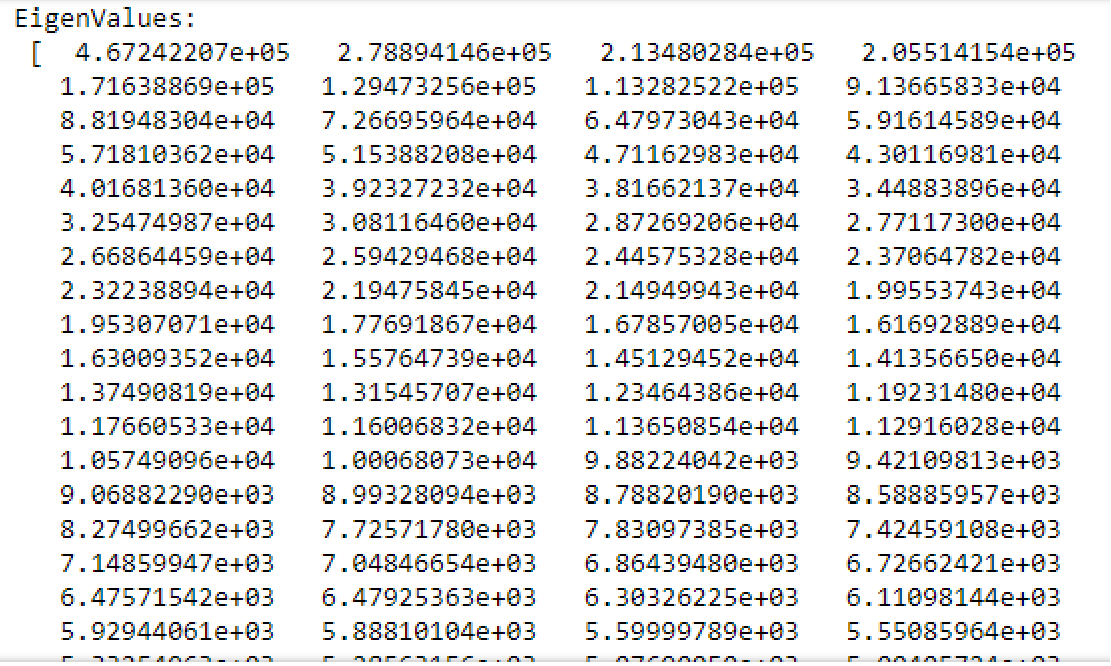


Comment: After Normalization, distribution of the data remains same.

Part 2:

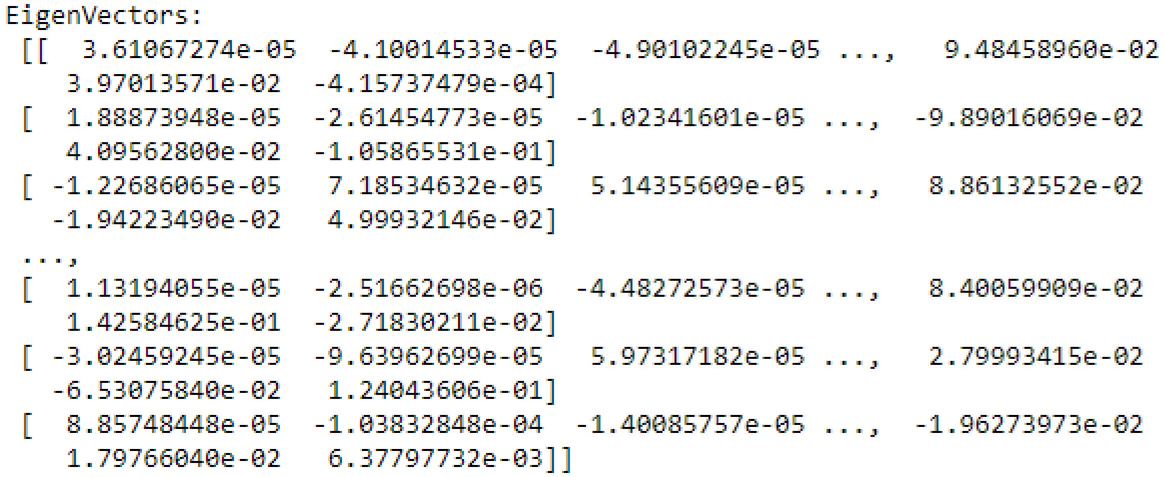
1. Using PCA as a dimensionality reduction technique we get the below eigen vectors and eigen Values.

Eigen Values:



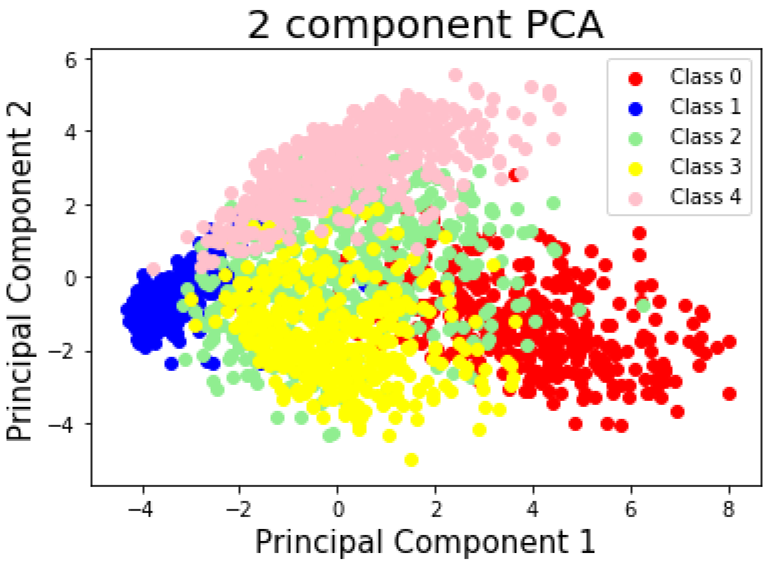
```
EigenValues:
 [  4.67242207e+05   2.78894146e+05   2.13480284e+05   2.05514154e+05
    1.71638869e+05   1.29473256e+05   1.13282522e+05   9.13665833e+04
    8.81948304e+04   7.26695964e+04   6.47973043e+04   5.91614589e+04
    5.71810362e+04   5.15388208e+04   4.71162983e+04   4.30116981e+04
    4.01681360e+04   3.92327232e+04   3.81662137e+04   3.44883896e+04
    3.25474987e+04   3.08116460e+04   2.87269206e+04   2.77117300e+04
    2.66864459e+04   2.59429468e+04   2.44575328e+04   2.37064782e+04
    2.32238894e+04   2.19475845e+04   2.14949943e+04   1.99553743e+04
    1.95307071e+04   1.77691867e+04   1.67857005e+04   1.61692889e+04
    1.63009352e+04   1.55764739e+04   1.45129452e+04   1.41356650e+04
    1.37490819e+04   1.31545707e+04   1.23464386e+04   1.19231480e+04
    1.17660533e+04   1.16006832e+04   1.13650854e+04   1.12916028e+04
    1.05749096e+04   1.00068073e+04   9.88224042e+03   9.42109813e+03
    9.06882290e+03   8.99328094e+03   8.78820190e+03   8.58885957e+03
    8.27499662e+03   7.72571780e+03   7.83097385e+03   7.42459108e+03
    7.14859947e+03   7.04846654e+03   6.86439480e+03   6.72662421e+03
    6.47571542e+03   6.47925363e+03   6.30326225e+03   6.11098144e+03
    5.92944061e+03   5.88810104e+03   5.59999789e+03   5.55085964e+03
```

Eigen Vectors:



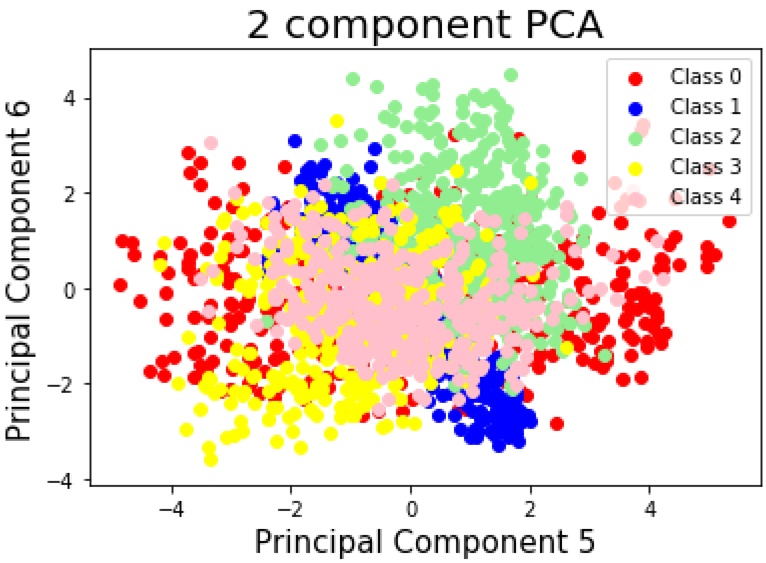
```
EigenVectors:
 [[  3.61067274e-05  -4.10014533e-05  -4.90102245e-05 ...,   9.48458960e-02
     3.97013571e-02  -4.15737479e-04]
  [  1.88873948e-05  -2.61454773e-05  -1.02341601e-05 ...,  -9.89016069e-02
     4.09562800e-02  -1.05865531e-01]
  [ -1.22686065e-05   7.18534632e-05   5.14355609e-05 ...,   8.86132552e-02
    -1.94223490e-02   4.99932146e-02]
  ...,
  [  1.13194055e-05  -2.51662698e-06  -4.48272573e-05 ...,   8.40059909e-02
     1.42584625e-01  -2.71830211e-02]
  [ -3.02459245e-05  -9.63962699e-05   5.97317182e-05 ...,   2.79993415e-02
    -6.53075840e-02   1.24043606e-01]
  [  8.85748448e-05  -1.03832848e-04  -1.40085757e-05 ...,  -1.96273973e-02
     1.79766040e-02   6.37797732e-03]]
```

2. A two-dimensional representation of the data points based on the the 1st and 2nd PCA is given below. There were 784 dimensions in the DatasetB, which is reduced into two dimensions using PCA.

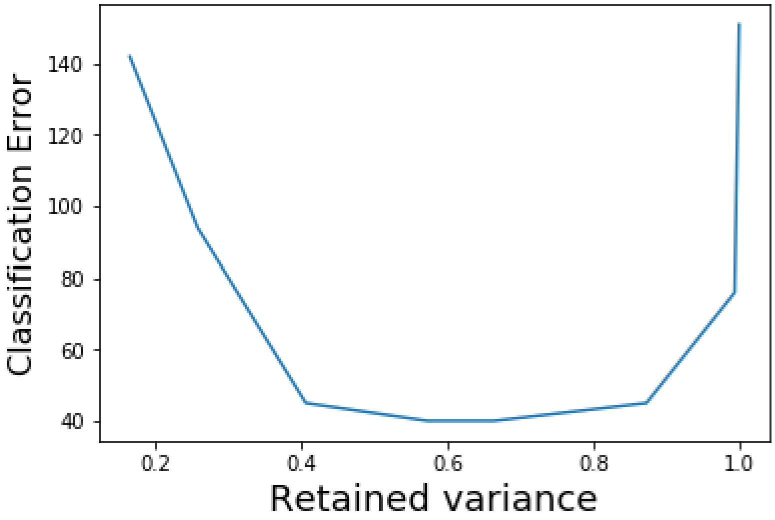


From the above figure we can see that the overlapping is present is the data. But still the classes are differentiable.

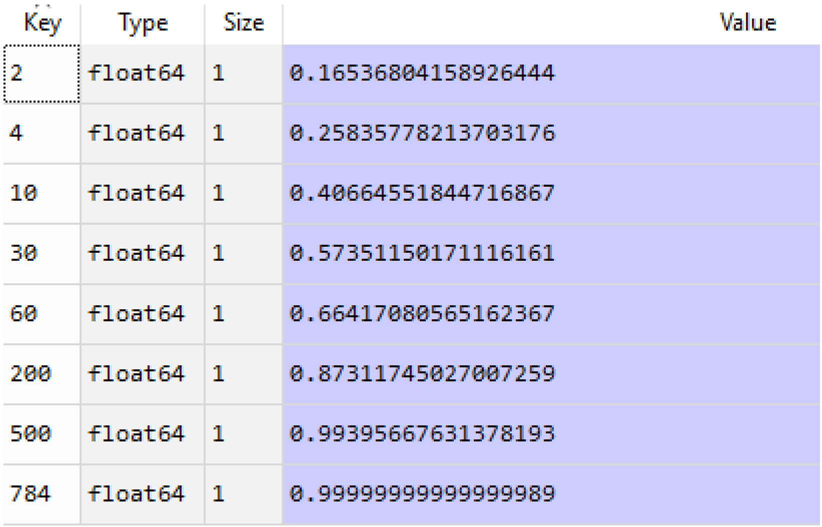
3. From the figure we can see that for the PCA5 and PCA6 the overlap is greater than previous case as they contain less information. Moreover, the classes are less differentiable.



4. Classification error for the 8 sets against the retained variance of each case.

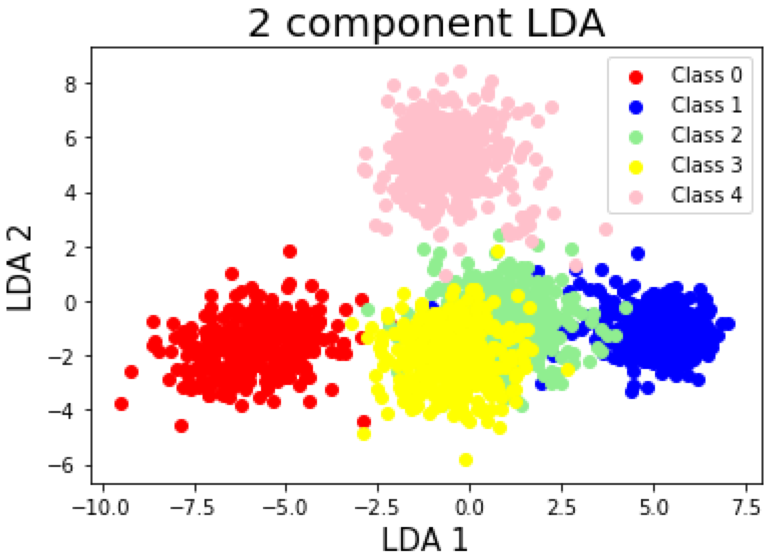


From the figure we can see that with the increase of retained variance, error initially decreased. But with higher retained variance the error increased which was unexpected. Below table shows the accuracy measurement for different PCA.



Key	Type	Size	Value
2	float64	1	0.16536804158926444
4	float64	1	0.25835778213703176
10	float64	1	0.40664551844716867
30	float64	1	0.57351150171116161
60	float64	1	0.66417080565162367
200	float64	1	0.87311745027007259
500	float64	1	0.99395667631378193
784	float64	1	0.99999999999999989

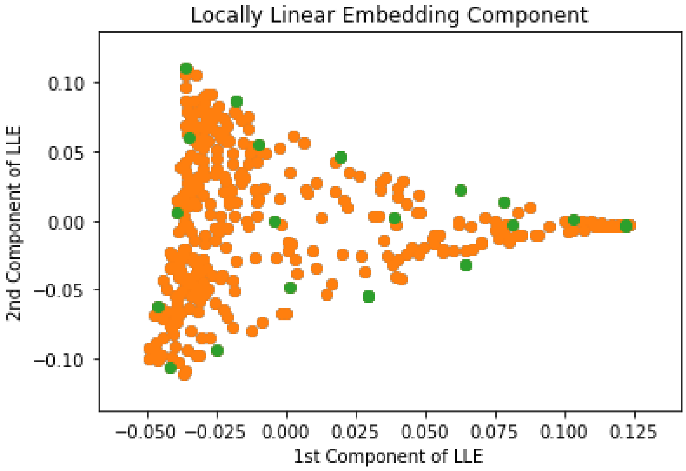
5. The data points have been plotted using 1st and 2nd LDA component and it has been displayed with each class with different color.

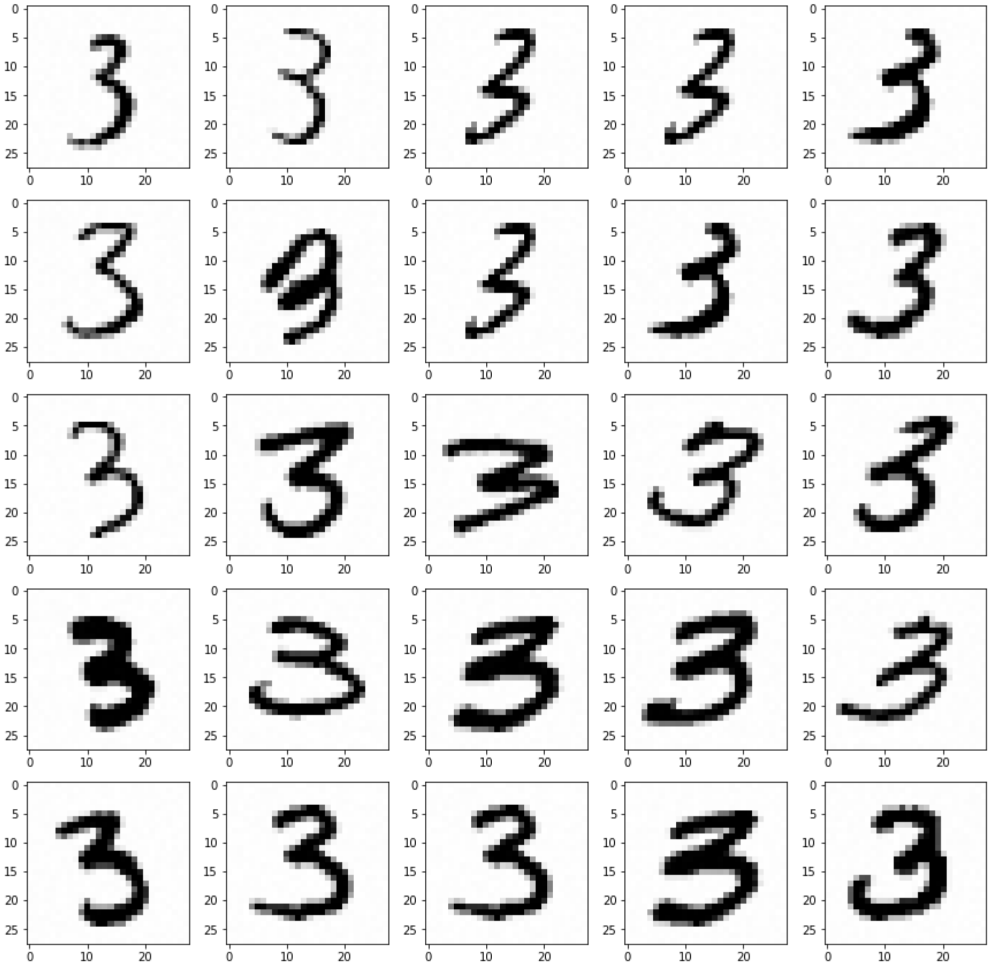


From the results obtained from PCA and LDA we can say that LDA technique is better because the classes are less overlapping compared to PCA and highly distinguishable.

Part 3:

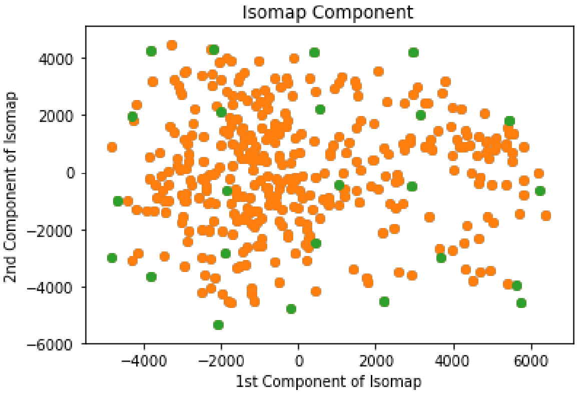
1.

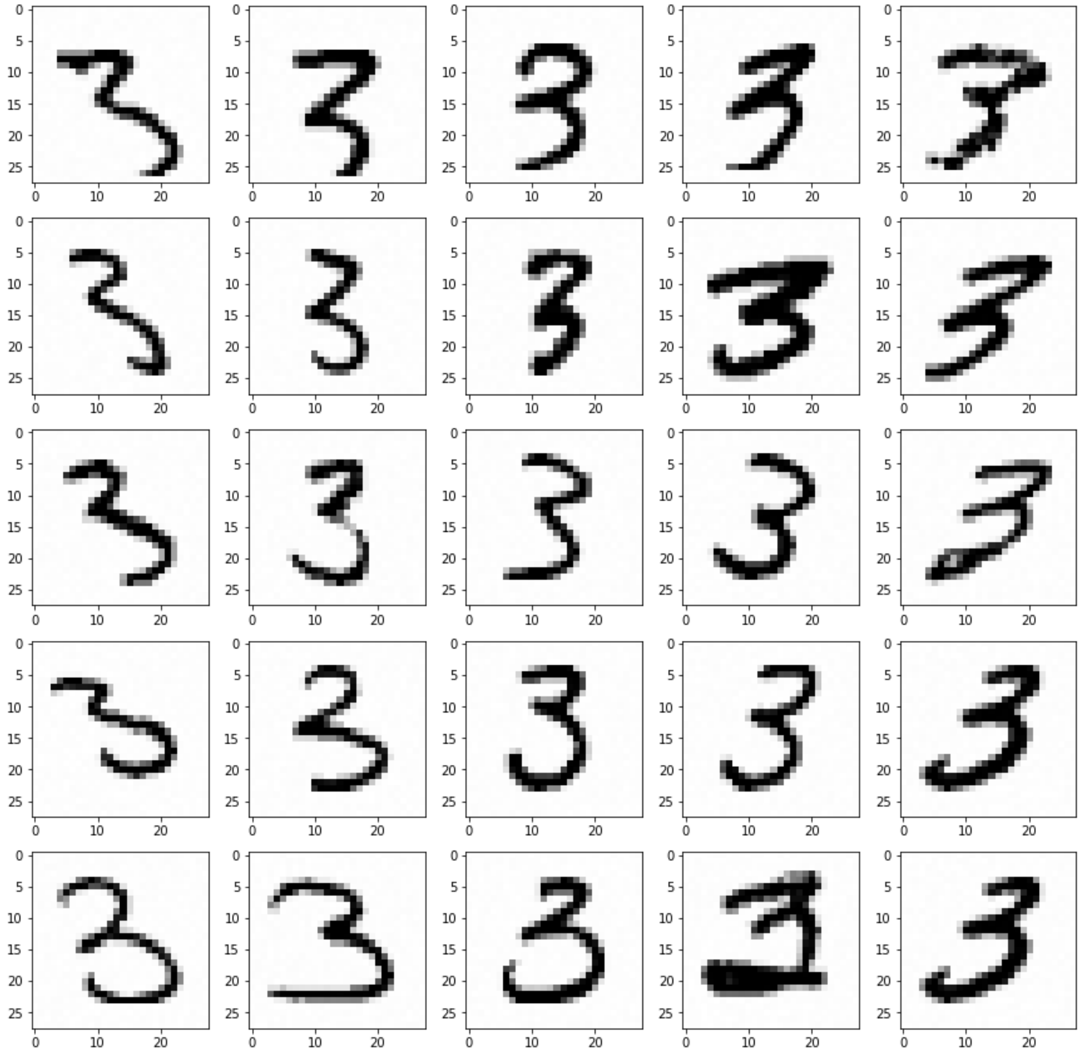




Alignment and symmetry were captured through this process.

2.





A significant change in tilt angle were captured as we move from left to right of the data. Also when we move from top to bottom we can see that it was able to capture the asymmetry of the digit.

The changes were better captured in ISOmap that LLE. Patterns are globally based.

3.

Given below the average accuracy of multiple runs for both cases

LLE average accuracy = **0.745161290323**

ISOmap average accuracy = **0.86935483871**

With the increase of the iteration number error accuracy level was not changing significantly. So, we can say for our case 100 iterations were sufficient.

PCA accuracy = **0.2583577821459**

LDA accuracy = **0 .9941**

From the results above we can conclude that LDA achieves higher accuracy among all.

Reference:

1. https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/

2. https://blog.paperspace.com/p/a6ee6e43-8af7-4de4-85fc-5bc8d90c789e/

3. http://scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html