IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – III
Attribute normalization, standardization and dimension reduction of data

Student's Name: Garvit Verma **Mobile No: 8272840777**

Roll Number: B20098 **Branch:CSE**

1 a.

Table 1 Minimum and maximum attribute values before and after normalization

S. No.	Attribute	Before normalization		After normalization	
		Minimum	Maximum	Minimum	Maximum
1	pregs	0	13	5	12
2	plas	44	199	5	12
3	pres (in mm Hg)	38	106	5	12
4	skin (in mm)	0	63	5	12
5	test (in mu U/mL)	0	318	5	12
6	BMI (in kg/m^2)	18.2	50	5	12
7	pedi	0.078	1.191	5	12
8	Age (in years)	21	66	5	12

Inferences:

1. There is a need for outlier correction because these are the data values that lies far away from most part of the data, which may be due to mistake or variance.
2. All values above upper bound and all values below lower bound are considered as outliers. So, all the values greater than q3+1.5*iqr and less than q1-1.5*iqr are replaced by their respective medians.
3. Before normalization, there was a huge range gap between the smallest and largest value, so the bigger values used to overpower smaller values. So, there was difficulty in analysis. But after normalization, all values are in the range 5-12 so that the analysis would be better.

b.

Table 2 Mean and standard deviation before and after standardization

S. No.	Attribute	Before standardization		After standardization	
		Mean	Std. Deviation	Mean	Std. Deviation
1	pregs	3.782	3.270	0.0	1.0
2	plas	121.656	30.438	0.0	1.0
3	pres (in mm Hg)	72.196	11.146	0.0	1.0
4	skin (in mm)	20.437	15.698	0.0	1.0
5	test (in mu U/mL)	60.919	77.635	0.0	1.0

6	BMI (in kg/m^2)	32.198	6.410	0.0	1.0
7	pedi	0.427	0.245	0.0	1.0
8	Age (in years)	32.760	11.055	0.0	1.0

Inferences:

1. Before normalization, there was a huge range gap between the smallest and largest value, so the bigger values used to overpower smaller values. So, there was difficulty in analysis. But after standardization, all attributes have common mean 0 and standard dev 1. So, not much variation.

2 a.

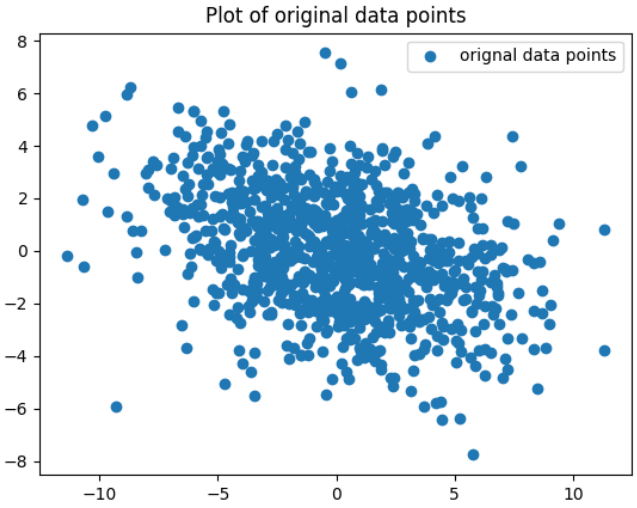

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. Attribute 2 is negatively related to attribute 1 according to graph as when attribute 1 increases, the value of attribute 2 decreases.
2. By observing the density of graph, the distribution of attributes seems symmetric and the mean tends to be 0.

b.

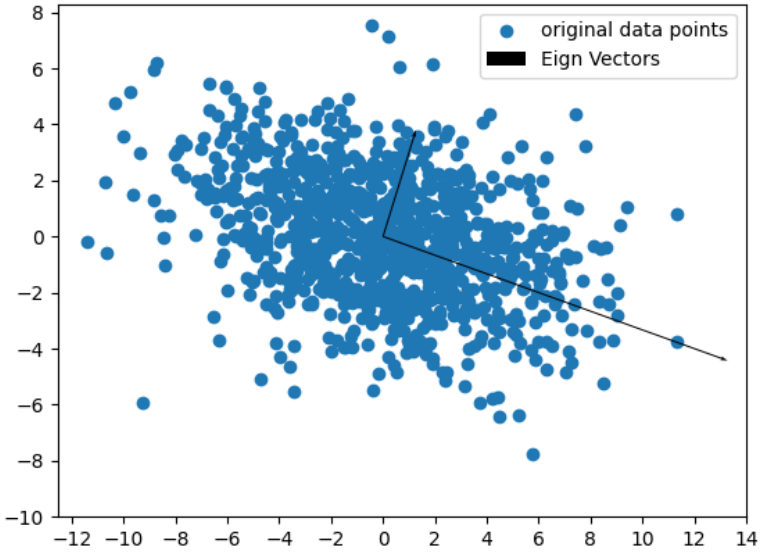

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. The spread along smaller eigen vector is small as compared to the spread along larger stretched eigen vector. It implies that the data spread is more linear with small spread in other's vector.
2. The density of points near the intersection is very dense which gradually decreases as spread increases i.e., the density decreases as we move away from point of intersection.

c.

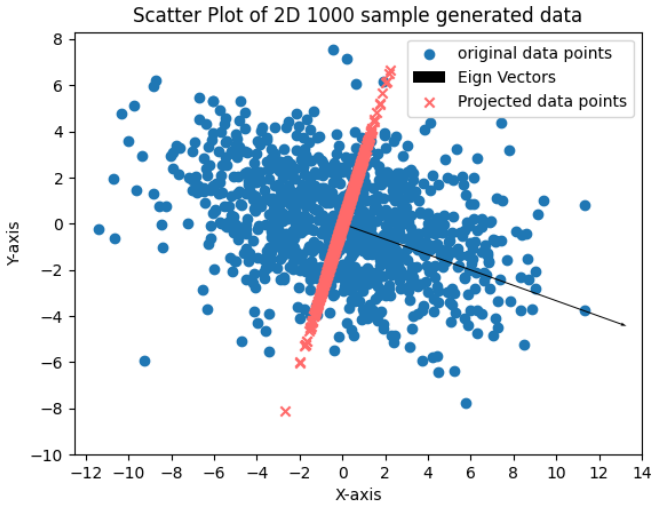

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

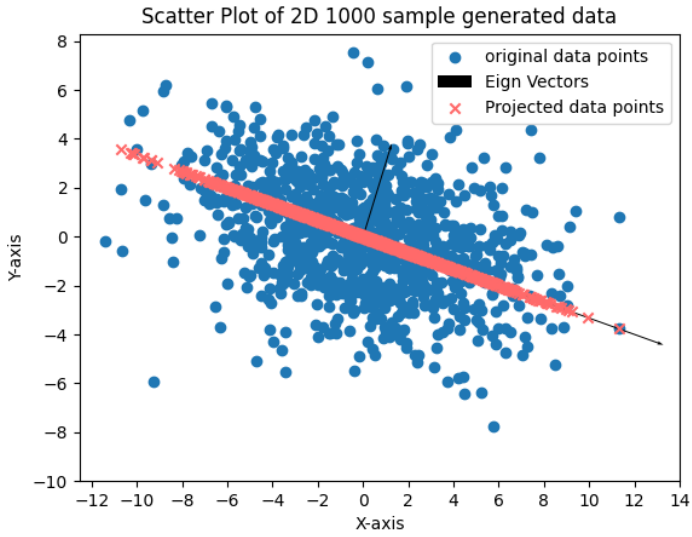

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. The magnitude of eigen values can be inferred from the variance along the eigen vectors. As we can see the variance along eigen vector (longer line) is more as compared to other eigen vector.

2. Along first eigen vector (small line) the variance is not very large, so the spread is not so much varying. But along second eigen vector (longer line) the variance is large and hence the spread is also large. So, the density is high near the intersection.

d. Reconstruction error = 0.0

Inferences:

1. More the reconstruction error leads to more loss of data. So, the reconstruction error must not be high.
2. Here the reconstruction error tends to 0, as the number of dimensions remains the same.

3 a.

Table 3 Variance and Eigenvalues of the projected data along the two directions

Direction	Variance	Eigenvalue
1	1.992	1.992
2	1.853	1.853

Inferences:

1. Eigen values is same as variance of the covariance matrix of transformed data.
2. Higher the value of eigen vector, more variance along that vector, so more strength along that direction. So, data will be more spread along first eigen vector.

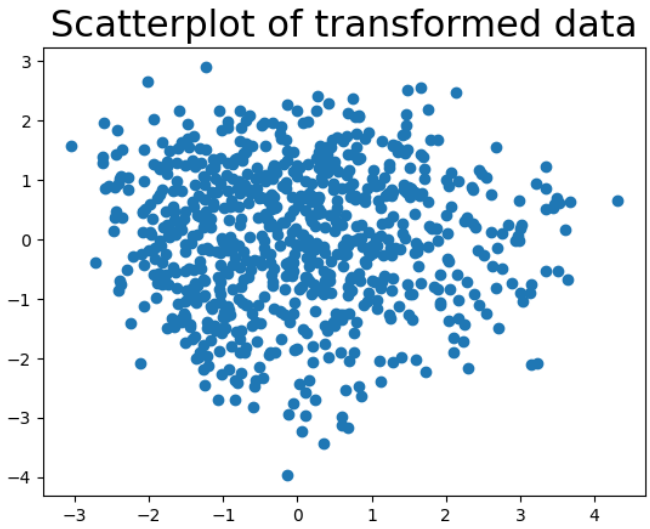

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. Attributes are negatively correlated to each other as one increases the other decreases.
2. The concentration is denser when both attributes are negative as PCA reduced the data in 2-dimension.

b.

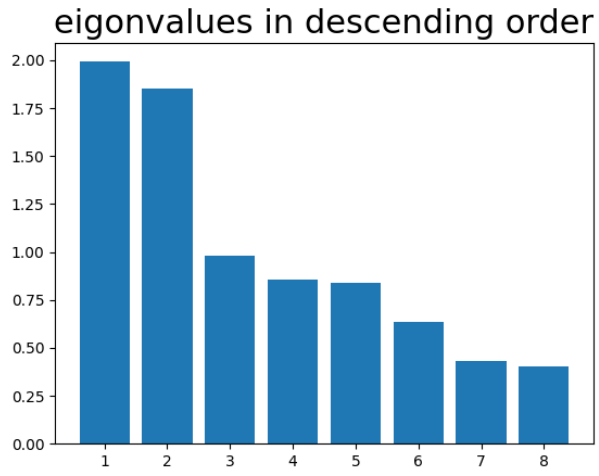

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. Eigen values drops significantly from second eigen value to third, and then gradually decreases.
2. At second eigen value, the rate of decrement is high.

c.

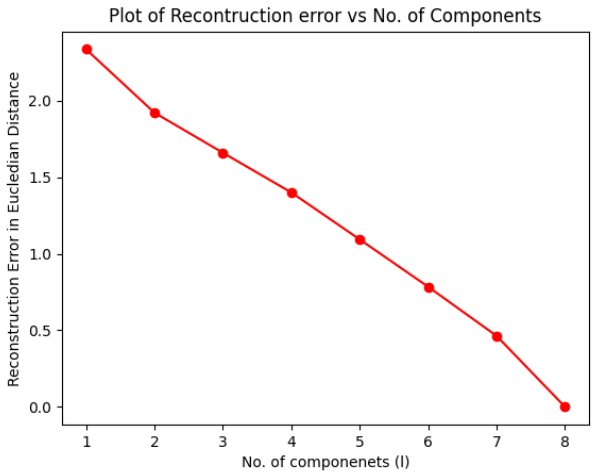

Figure 6 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. More the value of reconstruction error, less will be the quantity of reconstruction. As we drop the dimensions, Euclidian distance keep on increasing. So, at l=8, the reconstruction error is almost negligible.

Table 4 Covariance matrix for dimensionally reduced data (l=2)

	x1	x2
x1	1.992	0
x2	0	1.853

Table 5 Covariance matrix for dimensionally reduced data (l=3)

	x1	x2	x3
x1	1.992	0	0
x2	0	1.853	0
x3	0	0	0.982

Table 6 Covariance matrix for dimensionally reduced data (l=4)

	x1	x2	x3	x4
x1	1.992	0	0	0
x2	0	1.853	0	0
x3	0	0	0.982	0
x4	0	0	0	0.858

Table 7 Covariance matrix for dimensionally reduced data (l=5)

	x1	x2	x3	x4	x5
x1	1.992	0	0	0	0
x2	0	1.853	0	0	0
x3	0	0	0.982	0	0
x4	0	0	0	0.858	0
x5	0	0	0	0	0.839

Table 8 Covariance matrix for dimensionally reduced data (l=6)

	x1	x2	x3	x4	x5	x6
x1	1.992	0	0	0	0	0
x2	0	1.853	0	0	0	0
x3	0	0	0.982	0	0	0
x4	0	0	0	0.858	0	0
x5	0	0	0	0	0.839	0
x6	0	0	0	0	0	0.636

Table 9 Covariance matrix for dimensionally reduced data (l=7)

	x1	x2	x3	x4	x5	x6	x7
x1	1.992	0	0	0	0	0	0
x2	0	1.853	0	0	0	0	0
x3	0	0	0.982	0	0	0	0
x4	0	0	0	0.858	0	0	0
x5	0	0	0	0	0.839	0	0
x6	0	0	0	0	0	0.636	0
x7	0	0	0	0	0	0	0.434

Table 10 Covariance matrix for dimensionally reduced data (l=8)

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1.992	0	0	0	0	0	0	0
x2	0	1.853	0	0	0	0	0	0
x3	0	0	0.982	0	0	0	0	0
x4	0	0	0	0.858	0	0	0	0
x5	0	0	0	0	0.839	0	0	0
x6	0	0	0	0	0	0.636	0	0
x7	0	0	0	0	0	0	0.434	0
x8	0	0	0	0	0	0	0	0.405

Inferences:

1. Off diagonal elements are 0 which means that after reduction, the entries are uncorrelated with each other.
2. The diagonal elements are non-zero and off-diagonal elements are 0, as after pca reduction, the entries become uncorrelated.
3. Eigen value decreases as we move from x1 to x8.
4. It shows that x1 column contains highest portion of variance or information and this portion gradually decreases as we move towards x8.
5. Element x1(1.992) or component X1 captures the highest variation.
6. According to the diagonal elements we have to take 7 components that will give the optimum reconstruction as it preserves around 95% of variation of the data
7. Values of topmost diagonal element remain same because this value is corresponding to the eigen value of covariance matrix of original data and original data is not changing.
8. Values of 2nd diagonal element remain same because this value is corresponding to the eigen value of covariance matrix of original data and original data is not changing.
9. 3rd, 4th, 5th, 6th, and 7th diagonal elements across covariance matrices are same.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1.000	0.117	0.208	-0.096	-0.108	0.028	0.004	0.560
plas	0.117	1.000	0.204	0.060	0.179	0.228	0.081	0.274
pres (in mm Hg)	0.208	0.204	1.000	0.025	-0.050	0.271	0.022	0.326

skin (in mm)	-0.096	0.060	0.025	1.000	0.472	0.373	0.152	-0.101
test (in mu U/mL)	-0.108	0.179	-0.050	0.472	1.000	0.171	0.198	-0.073
BMI (in kg/m^2)	0.028	0.228	0.271	0.373	0.171	1.000	0.123	0.077
pedi	0.004	0.081	0.022	0.152	0.198	0.123	1.000	0.036
Age (in years)	0.560	0.274	0.326	-0.101	-0.073	0.077	0.036	1.000

Inferences:

1. The values of off diagonal elements here are non-zero while after PCA l=8, the off-diagonal elements are of the order 10^ (-16) which tends to zero.
2. Here covariance of attributes with itself is unity but in PCA (l=8) values range from 1.992 to 0.405.
3. There is no increase or decrease observed in diagonal elements in above covariance matrix but increase in the covariance matrix with PCA (l=8).