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

Student's Name: Titiksha Behal **Mobile No: 9811162646**

Roll Number: B20138 **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	0.000	13.000	5.0	12.0
2	plas	44.000	199.000	5.0	12.0
3	pres (in mm Hg)	38.000	106.000	5.0	12.0
4	skin (in mm)	0.000	63.000	5.0	12.0
5	test (in mu U/mL)	0.000	318.000	5.0	12.0
6	BMI (in kg/m^2)	18.200	50.000	5.0	12.0
7	pedi	0.078	1.191	5.0	12.0
8	Age (in years)	21.000	66.000	5.0	12.0

Inferences:

1. Outlier correction is important because outliers are extreme values present in the data due to some random error and lead to noise. So we replace all the outlier values with median of that attribute.
2. In normalization the values of an attribute are scaled so that they fall within a small specified range.
3. It is useful when different attributes have different ranges.

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	3.78255	3.26851	-0.0	1.0
2	plas	121.65625	30.41846	0.0	1.0
3	pres (in mm Hg)	72.19661	11.13946	0.0	1.0
4	skin (in mm)	20.43750	15.68833	0.0	1.0
5	test (in mu U/mL)	60.91927	77.58511	-0.0	1.0
6	BMI (in kg/m^2)	32.19896	6.40638	-0.0	1.0
7	pedi	0.42767	0.24500	0.0	1.0
8	Age (in years)	32.76042	11.04818	0.0	1.0

Inferences:

1. In standardization the data is scaled in such a manner that the transformed data has mean equal to 0 and standard deviation equals to 1.
2. This method is useful mainly when the data has gaussian distribution.

2 a.

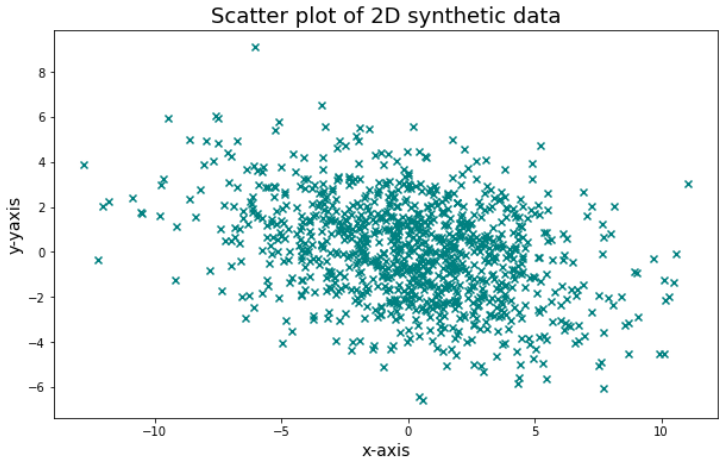

Figure 1. Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. Looking at the scatter plot we can see that both the attributes are negatively correlated because as x increases, y decreases.
2. Looking at the plot we can see that most of points are centered about (0,0)

b.

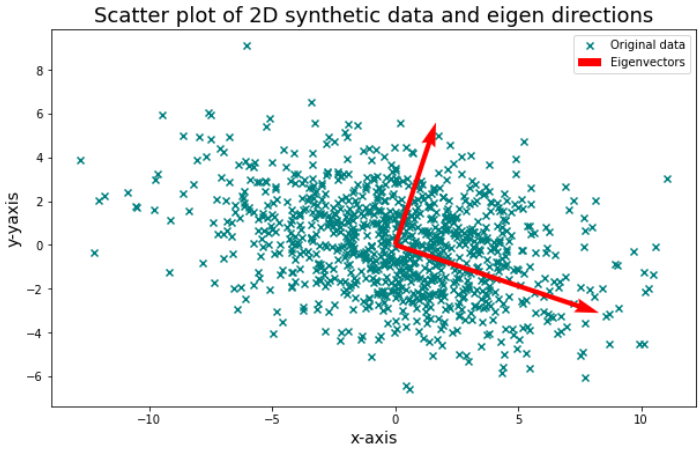

Figure 2. Plot of 2D synthetic data and Eigen directions

Inferences:

1. Eigenvalues represent the variance of the data in that direction. This means that variance of the data is more along the direction of eigen vector1.
2. We can see that both the eigen directions are perpendicular to each other.
3. The density of points is maximum near the intersection of the eigen axis and as we move away from it the density decreases.

c.

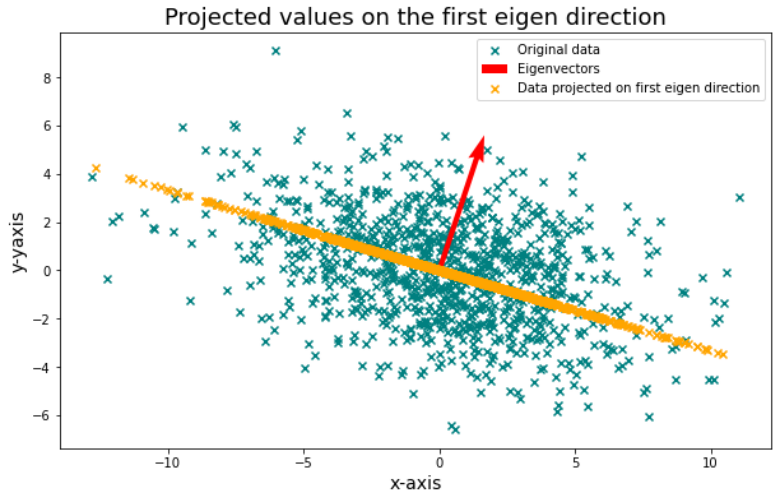

Figure 3. Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

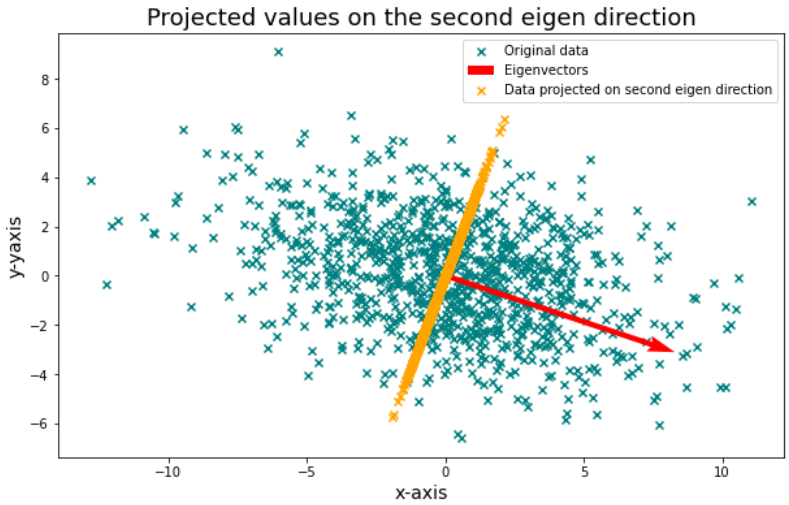

Figure 4. Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. The magnitude of the eigenvalue corresponding to eigen vector is greater than that corresponding to eigen vector 2.
2. The variance of the data is more along the first eigen direction as it has larger eigen value associated with it.

d. Reconstruction error = 1.27137 e-15

Inferences:

1. A smaller magnitude of reconstruction error implies a better quality of reconstruction as it means that the dimensionally reduced data can give us a good approximation of the original data.
2. Since the reconstruction error obtained in this case is extremely small which means that the projection of the data along the eigen vectors gives a good approximation of the data due to which the quality of reconstruction is high.

3 a.

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

Direction	Variance	Eigenvalue
1	1.99246	1.99506
2	1.85342	1.85584

Inferences:

1. It can be observed that the value of the variance along a particular direction is very close to the eigen value corresponding to that direction.
2. This verifies the fact that eigen values represent the variance of the data in that direction.

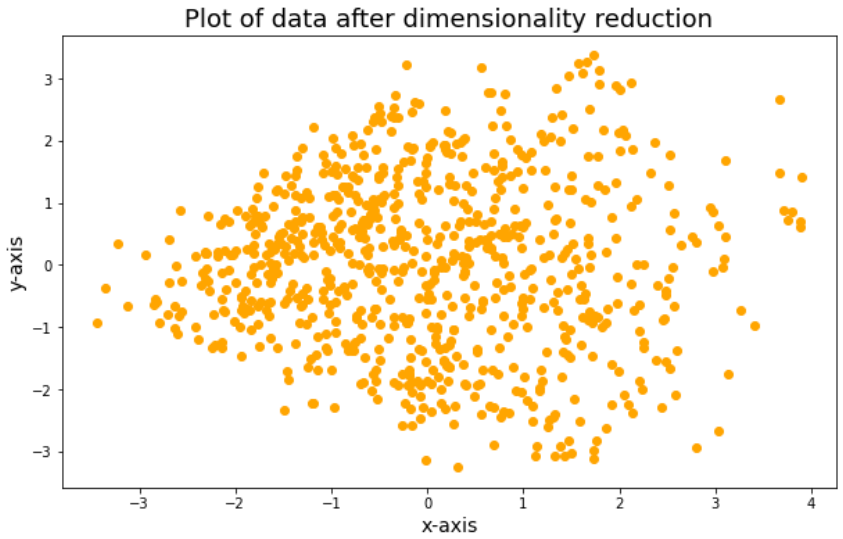

Figure 5. Plot of data after dimensionality reduction

Inferences:

1. The two attributes obtained after dimensionality reduction seem to be uncorrelated.
2. The density of points is maximum near x = -1 and x=0.

b.

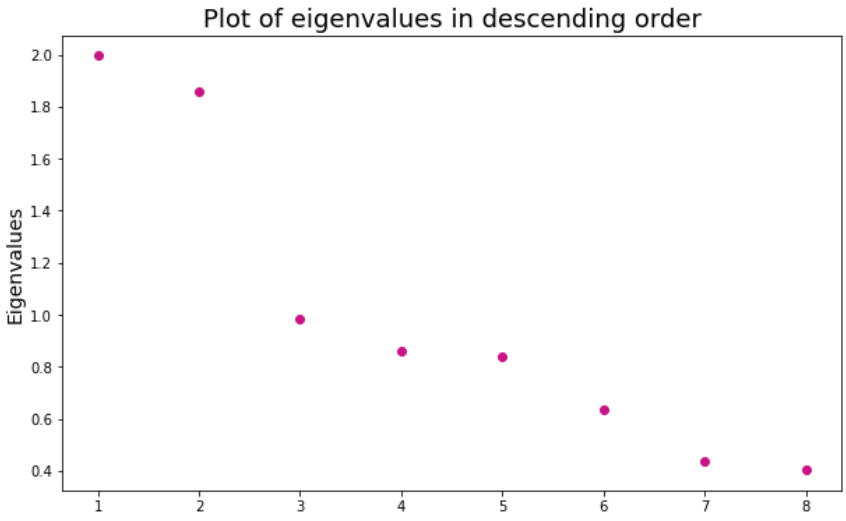

Figure 6. Plot of Eigenvalues in descending order

Inferences:

1. The eigen values decrease rapidly in the beginning then the decrease becomes more gradual.
2. Till l=3 they decrease rapidly after which they start decreasing in a more gradual manner.

c.

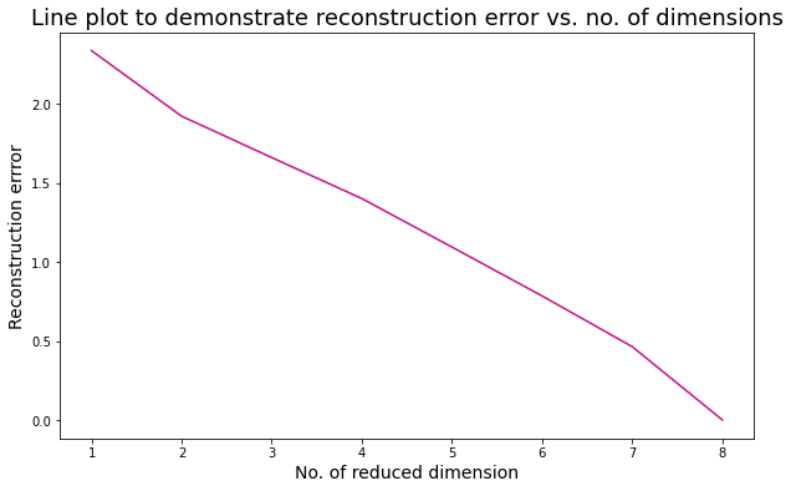

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. The smaller the value of reconstruction error, the better the quality of reconstruction. Which means that dimensionally reduced data can be used to give a good approximation of the original data.
2. We see that the as the value of l increases the reconstruction error decreases. This means that when we project the data onto a larger number of orthogonal eigenvectors the reduced data obtained gives a better approximation of the original data as compared to the data which is obtained for smaller values of l.

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

	x1	x2
x1	1.99506	0.00000
x2	0.00000	1.85584

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

	x1	x2	x3
x1	1.99506	0.00000	0.00000
x2	0.00000	1.85584	0.00000
x3	0.00000	0.00000	0.98316

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

	x1	x2	x3	x4
x1	1.99506	0.00000	0.00000	0.00000
x2	0.00000	1.85584	0.00000	0.00000
x3	0.00000	0.00000	0.98316	0.00000
x4	0.00000	0.00000	0.00000	0.85943

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

	x1	x2	x3	x4	x5
x1	1.99506	0.00000	0.00000	0.00000	0.00000
x2	0.00000	1.85584	0.00000	0.00000	0.00000
x3	0.00000	0.00000	0.98316	0.00000	0.00000
x4	0.00000	0.00000	0.00000	0.85943	0.00000
x5	0.00000	0.00000	0.00000	0.00000	0.83984

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

	x1	x2	x3	x4	x5	x6
x1	1.99506	0.00000	0.00000	0.00000	0.00000	0.00000
x2	0.00000	1.85584	0.00000	0.00000	0.00000	0.00000
x3	0.00000	0.00000	0.98316	0.00000	0.00000	0.00000
x4	0.00000	0.00000	0.00000	0.85943	0.00000	0.00000
x5	0.00000	0.00000	0.00000	0.00000	0.83984	0.00000
x6	0.00000	0.00000	0.00000	0.00000	0.00000	0.63724

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

	x1	x2	x3	x4	x5	x6	x7
x1	1.99506	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
x2	0.00000	1.85584	0.00000	0.00000	0.00000	0.00000	0.00000
x3	0.00000	0.00000	0.98316	0.00000	0.00000	0.00000	0.00000
x4	0.00000	0.00000	0.00000	0.85943	0.00000	0.00000	0.00000
x5	0.00000	0.00000	0.00000	0.00000	0.83984	0.00000	0.00000
x6	0.00000	0.00000	0.00000	0.00000	0.00000	0.63724	0.00000
x7	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.43471

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

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1.99506	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
x2	0.00000	1.85584	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
x3	0.00000	0.00000	0.98316	0.00000	0.00000	0.00000	0.00000	0.00000
x4	0.00000	0.00000	0.00000	0.85943	0.00000	0.00000	0.00000	0.00000
x5	0.00000	0.00000	0.00000	0.00000	0.83984	0.00000	0.00000	0.00000
x6	0.00000	0.00000	0.00000	0.00000	0.00000	0.63724	0.00000	0.00000
x7	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.43471	0.00000
x8	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.40516

Inferences:

1. All the off-diagonal elements are zero because they represent the correlation between the different attributes and after dimension reduction the different features are uncorrelated.
2. The diagonal elements on the other hand are non-zero because they represent the variance of the data along that direction.
3. We see that as the value of dimensions increases the value of the diagonal element(variance) decreases.
4. The values decrease because the features are arranged in the descending order of their corresponding eigen values which represent the variance of the data along that direction. Because of this as the values of l increases, eigen values decrease and hence the variance also decreases.
5. The element with highest value of variance that is x1 captures the data variations best.
6. About 3 components would give optimal reconstruction along with dimension reduction.
7. The magnitude of the first diagonal element in all the matrices ids the same as it represents the variance of the first feature i.e., the variance of the data along the first eigen direction.
8. The same is true for all the other diagonal elements as well.

d.

Table 11: Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1.001304	0.117845	0.209225	-0.096846	-0.108616	0.028376	0.004525	0.561499
plas	0.117845	1.001304	0.204806	0.060112	0.179800	0.228542	0.081720	0.274622
pres (in mm Hg)	0.209225	0.204806	1.001304	0.025678	-0.051022	0.271914	0.022525	0.326798
skin (in mm)	-0.096846	0.060112	0.025678	1.001304	0.473330	0.374213	0.152962	-0.101529
test (in mu U/mL)	-0.108616	0.179800	-0.051022	0.473330	1.001304	0.171727	0.198839	-0.073822
BMI (in kg/m^2)	0.028376	0.228542	0.271914	0.374213	0.171727	1.001304	0.123937	0.077770
pedi	0.004525	0.081720	0.022525	0.152962	0.198839	0.123937	1.001304	0.036156
Age (in years)	0.561499	0.274622	0.326798	-0.101529	-0.073822	0.077770	0.036156	1.001304

Inferences:

1. In this case the value of non-diagonal elements is not zero because the attributes are correlated to each other in the original data.
2. In this case all the diagonal elements have the same value which means all the attributes have the same variance. This is because this covariance matrix was obtained for standardized data because of which the standard deviation of all the attributes is equal to 1. Because of this there variance are also equal.