IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – III
Attribute Normalization, Standardization and Dimension Reduction of Data

Student's Name: Jai Prakash Yadav **Mobile No: 930687178**

Roll Number: B19247 **Branch: Mechanical Engineering**

1 a.

Table 1 Minimum and Maximum Attribute Values Before and After Min-Max Normalization

S. No.	Attribute	Before Min-Max Normalization		After Min-Max Normalization	
		Minimum	Maximum	Minimum	Maximum
1	Temperature (in °C)	10.085	31.375	3.000	9.000
2	Humidity (in $g.m^{-3}$)	34.205	99.720	3.000	9.000
3	Pressure (in mb)	992.654	1037.604	3.000	9.000
4	Rain (in ml)	0.000	2470.500	3.000	9.000
5	Lightavgw/o0 (in lux)	0.000	10565.3520	3.000	9.000
6	Lightmax (in lux)	2259.000	54612.000	3.000	9.000
7	Moisture (in %)	0.000	100.000	3.000	9.000

Inferences:

1. Inference related to outliers
2. Infer upon what happens before and after normalization

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	Temperature (in °C)	21.369	4.125	0.000	1.000
2	Humidity (in $g.m^{-3}$)	83.992	17.565	0.000	1.000
3	Pressure (in mb)	1014.760	6.121	0.000	1.000
4	Rain (in ml)	168.400	399.689	0.000	1.000
5	Lightavgw/o0 (in lux)	2197.392	2220.820	0.000	1.000
6	Lightmax (in lux)	21788.623	22064.993	0.000	1.000
7	Moisture (in %)	32.386	33.653	0.000	1.000

Inferences:

1. Before standardization, the mean and standard deviation of the attributes are different for each attribute. In this process the values of an attribute are normalized based on mean and standard deviation. After standardisation, the mean become 0 and standard deviation 1 for all attributes in the data.
2. This technique is more effective when the distribution of attribute is gaussian.

2 a.

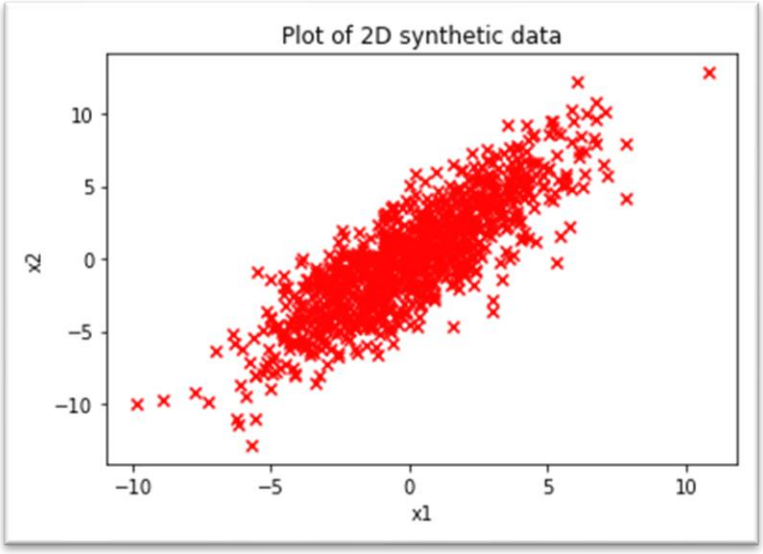

Figure 1 Scatter Plot of 2D Synthetic Data of 1000 samples

Inferences:

1. From the graph it is clearly visible that as the value of x1 increase value of x2 also increase. This implies that both x1 and x2 are highly correlated.
2. From the distribution points it is clearly seen that it is Gaussian bivariate distribution, where the points are highly densed near the origin.
3. Variance of x1 is less as compare to that of x2.

b.

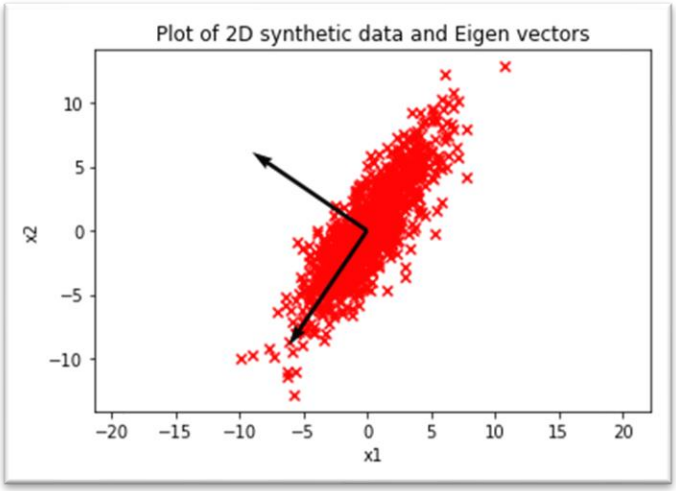

Figure 2 Plot of 2D Synthetic Data and Eigen Directions

Inferences:

1. For x1(Eigen value: 1.756, Eigen vector: [-0.833, 0.553]) and for x2 (Eigen value: 19.799, Eigen vector: [-0.553 -0.833]). Hence as the eigen value of x1 is less than as compared to x2 this implies that the data is more spread across x2 as compared to x1.
2. There is larger density of points near the origin because as the mean of data coincides with the origin.

c.

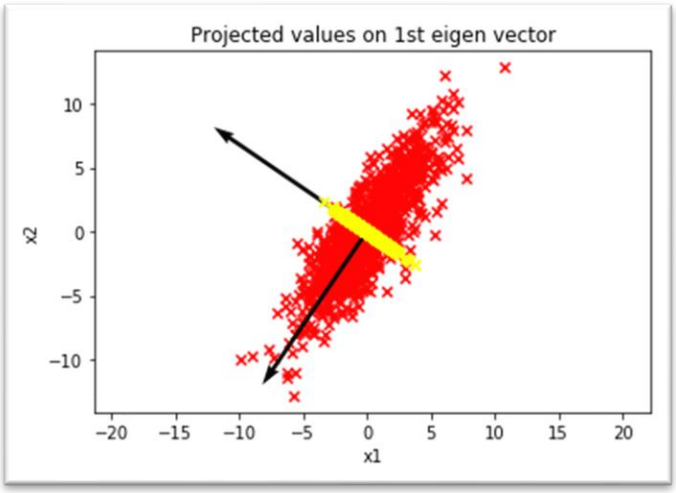

Figure 3 Projected Eigen Directions onto the Scatter Plot with 1st Eigen Direction highlighted

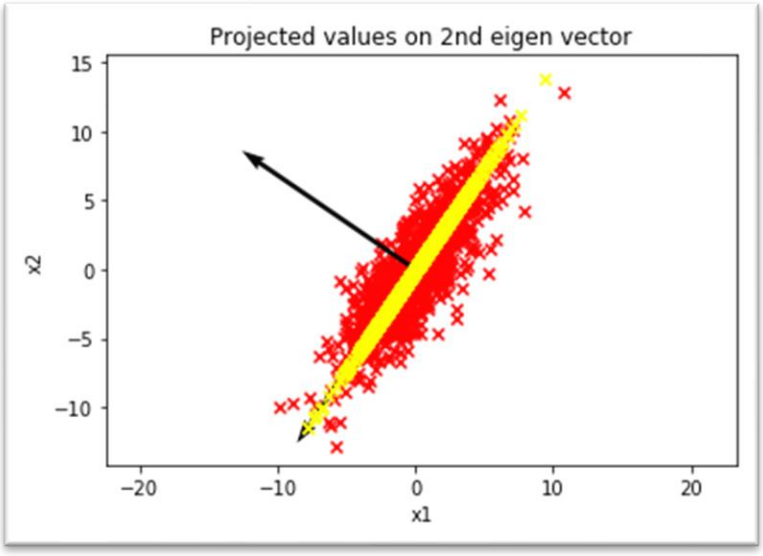

Figure 4 Projected Eigen Directions onto the Scatter Plot with 2nd Eigen Direction highlighted

Inferences:

1. Eigen Values of x1(1.756) is greater than that of x2(19.799). and the eigen vectors of both x1 and x2 are perpendicular to each other hence they are orthonormal vectors.
2. The magnitude of eigen vector is directly proportional to spread that is variance. Also, from the graph as eigen value of x1 is less than x2, the spread along eigen vector x1 is less as compared to x2.

d. Reconstruction Error = 0.000

Inferences:

1. Lower the reconstruction error implies lower the loss of information in compressed data.
2. Here in given data case RMSE comes out to 0.00. This means data reduction is lossless.

3 a.

Table 3 Variance and Eigen Values of the projected data along the two directions

Direction	Variance	Eigen Value
1	2.1999680069665772	2.2022984815502253
2	1.4193223133447708	1.4208258327445005

Inferences:

1. The variance of the data comes out to be almost equal to eigen value. This indicates that there is no loss of data while reduction and data still contain valuable information.

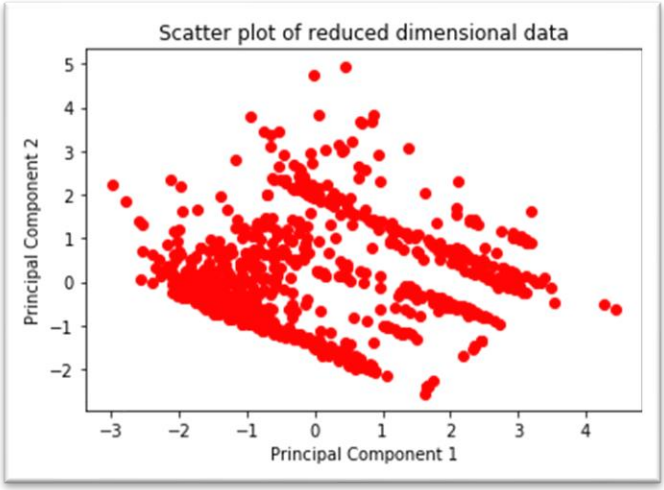

Figure 5 Plot of Landslide Data after dimensionality reduction

Inferences:

1. The reduced data seems less correlated as the points scattered all around. From the distribution it seems it is positively skewed as the median of the reduced data is less than mean.
2. Also, from graph as the density around the median is very high it implies that the distribution follows skewed gaussian distribution.

b.

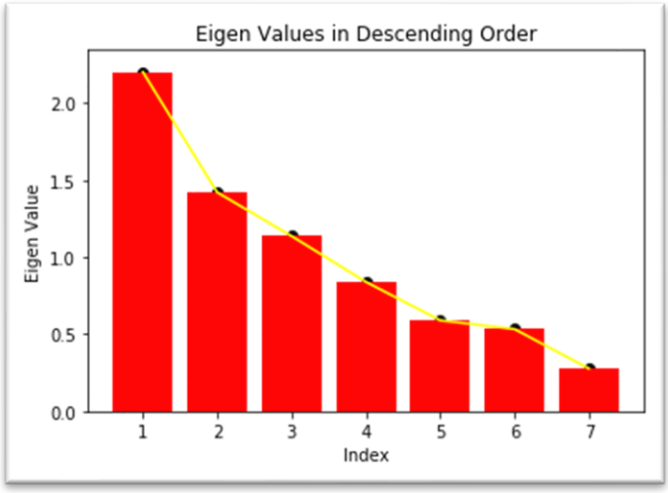

Figure 6 Plot of Eigen Values in descending order

Inferences:

1. There is gradual decrease in eigenvalue
2. There is rapid decrease in eigen value from 2.5 to 1.4 that is from eigen value 1 to eigen value 2.

c.

Figure 7 Line Plot to demonstrate Reconstruction Error vs. Components

Inferences:

1. Magnitude of reconstruction error decreases as the quality of reconstruction increases.
2. As the L tends to D (actual number of dimensions) the RMSE value tends to 0.
3. Lesser the reconstruction error lesser the data loss