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

Student's Name: Prashant Kumar **Mobile No: 8700350173**

Roll Number: B19101 **Branch: CSE**

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)	10.0851	31.375	3	9
2	Humidity (in $g.m^{-3}$)	34.206	99.720	3	9
3	Pressure (in mb)	992.654	1037.604	3	9
4	Rain (in ml)	0	2770.5	3	9
5	Lightavgw/o0 (in lux)	0	10565.352	3	9
6	Lightmax (in lux)	2259	54612	3	9
7	Moisture (in %)	0	100	3	9

Inferences:

1. The outliers are replaced by the median computed from non-outliners.
2. As we can see that the minimum and maximum value of Lightmax(in lux) and Lightavgw/o0 (in lux) have very large value as compared to the other attributes. So, when we calculate the Euclidean distance from a point one attribute will dominate over other attributes. So, after min-max normalization all attributes have same range due to which no one will dominate over another.

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	1
2	Humidity (in $g.m^{-3}$)	83.992	17.566	0	1
3	Pressure (in mb)	1014.761	6.121	0	1
4	Rain (in ml)	168.400	399.689	0	1
5	Lightavgw/o0 (in lux)	2197.392	2220.820	0	1
6	Lightmax (in lux)	21788.623	22064.993	0	1
7	Moisture (in %)	32.386	33.653	0	1

Inferences:

1. There are some drawbacks of min-max normalization as if the test data contain values out of range then this will lead to outflow of data.
2. So, we normalize the data using Z-score normalization which normalize the data on the basis of mean and standard deviation due to which after normalization we can see that the mean tends to zero and standard deviation has value equal to 1.

2 a.

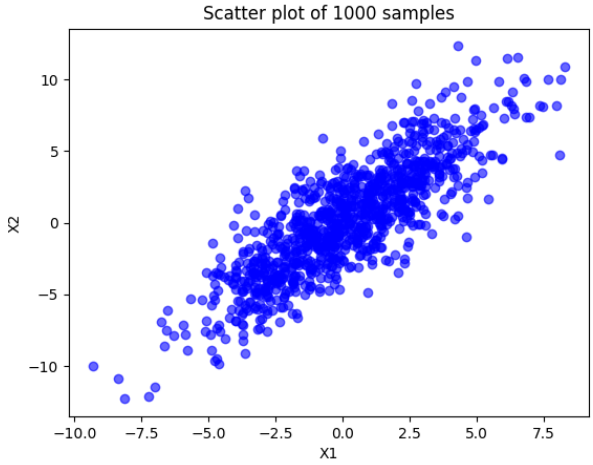

Figure 1 Scatter Plot of 2D Synthetic Data of 1000 samples

Inferences:

1. As X1 increases X2 also increase that means both are positively correlated. On computing correlation coefficient, it is 0.85 which also shows that both are highly corelated
2. Scatter Plot of 2D Synthetic Data of 1000 samples is centered at mean1 and mean2 ([0,0]) and dispersed according to the covariance matrix ([[5,10], [10,13]]).

b.

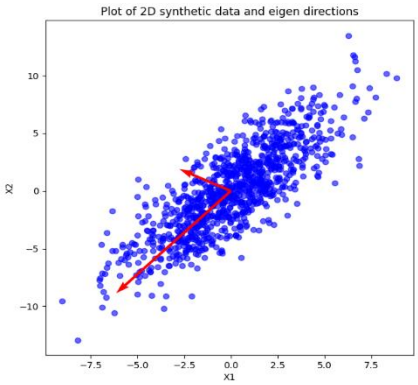

Figure 2 Plot of 2D Synthetic Data and Eigen Directions

Inferences:

1. Eigen Values are: 1.761, 19.341.
2. Eigen Vector are: [0.553 -0.834], [-0.834, -0.553].
3. Hence the higher the eigen value, then data will more spread towards the corresponding vector.
4. Density where the eigen vector intersect is more and decrease as we go away toward the axis.

c.

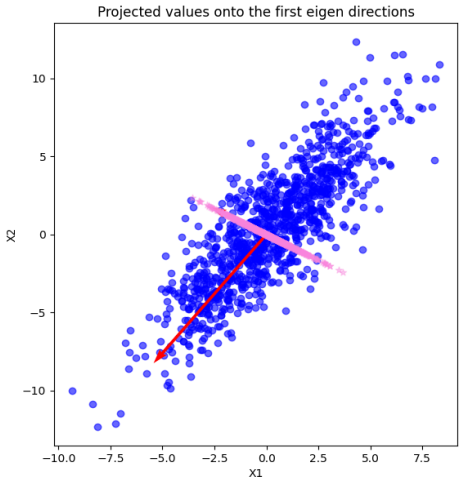

Figure 3 Projected Eigen Directions onto the Scatter Plot with 1st Eigen Direction highlighted

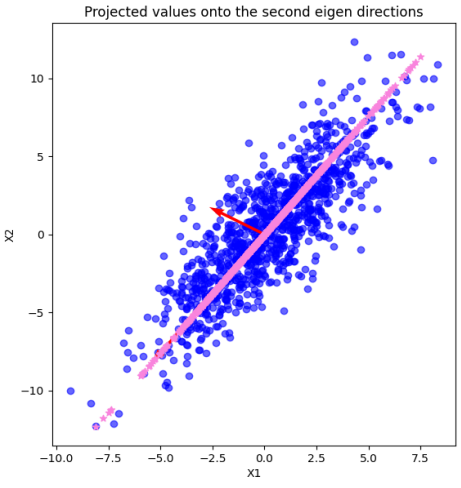

Figure 4 Projected Eigen Directions onto the Scatter Plot with 2nd Eigen Direction highlighted

Inferences:

1. Eigen Values are: 1.761, 19.341.
2. Higher the eigen values then there is more spread towards the corresponding eigen vector.
3. Hence, we can say that the eigen vector corresponding to higher eigen value have more information content

d. Reconstruction Error = 4.526e-31

Inferences:

1. More is the reconstruction error will lead to more loss of information that we reconstruct from compressed data.
2. Reconstruction error is one of the important parameters to check that the data loss is minimum or not.

3 a.

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

Direction	Variance	Eigen Value
1	2.199	2.203
2	1.419	1.421

Inferences:

1. The eigen values and the variance that we get are approximately same.
2. Both eigen values seems large. Hence the information content was also larger.

Figure 5 Plot of Landslide Data after dimensionality reduction

Inferences:

1. In the figure 5 we can see that the both dimensions are uncorrelated.
2. The plot looks like that most of the value are dense toward the median of each dimension i.e. the data is skewed.

b.

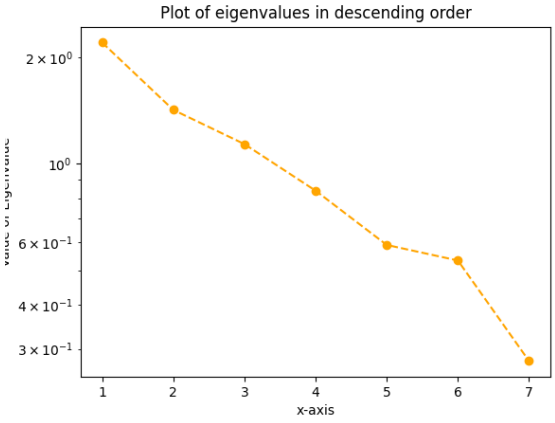

Figure 6 Plot of Eigen Values in descending order

Inferences:

1. According to graph, in the starting there is rapid decrease in eigen values but at the end the rate get decrease.
2. In the starting the rate of decrease changes is substantially but after 5th eigen value the rate of decrease is less.

c.

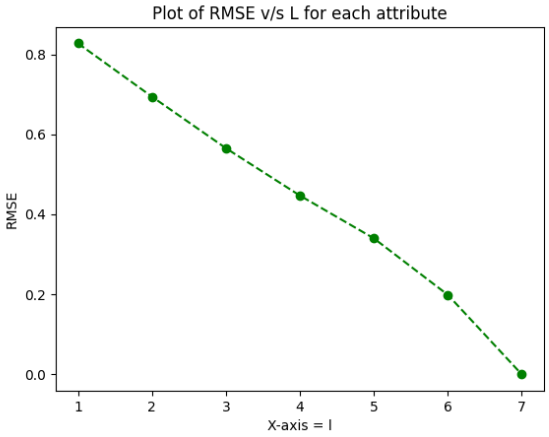

Figure 7 Line Plot to demonstrate Reconstruction Error vs. Components

Inferences:

1. As the magnitude of reconstruction error increase, then the quality of reconstruction decreases, Thus, we can say that reconstruction error is inversely proportional to quality of reconstruction
2. If the reconstruction error greater than zero then the data is called lossy and when the reconstruction error is zero then data is called lossless.