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

Student's Name: Sumit Kumar

Branch: CSE

Roll Number: B19118

Mobile No: 7807327549

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	3.0	9.0
2	Humidity (in g.m^{-3})	34.205	99.72	3.0	9.0
3	Pressure (in mb)	992.654	1037.604	3.0	9.0
4	Rain (in ml)	0.0	2470.5	3.0	9.0
5	Lightavgw/o0 (in lux)	0.0	10565.352	3.0	9.0
6	Lightmax (in lux)	2259	54612	3.0	9.0
7	Moisture (in %)	0.0	100.0	3.0	9.0

Inferences:

1. The attribute **"Rain"** contains the maximum number of outliers whereas there is no any outlier in the attribute **"Lightmax"** and **"Moisture".**
2. After performing the maximum – minimum normalization, the data points get transformed into the value between 3-9.
3. On doing normalization the behavior of the data is not changing only scaling is changed.

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)	6.180	1.163	0	1.0
2	Humidity (in g.m^{-3})	7.559	1.608	0	1.0
3	Pressure (in mb)	5.955	0.817	0	1.0
4	Rain (in ml)	3.416	0.968	0	1.0
5	Lightavgw/o0 (in lux)	4.271	1.253	0	1.0

6	Lightmax (in lux)	5.238	2.529	0	1.0
7	Moisture (in %)	4.943	2.019	0	1.0

Inferences:

1. After standardization the mean is found to be zero and the standard deviation is found to be 1.
2. We are considering the gaussian distribution of the data points.

2 a.

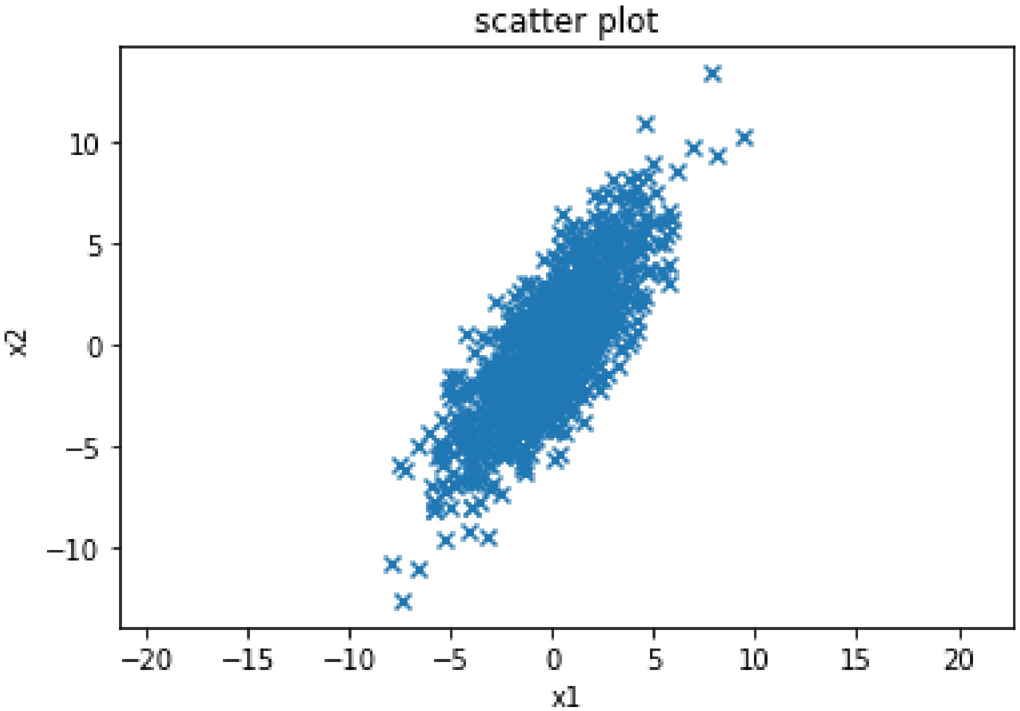

Figure 1 Scatter Plot of 2D Synthetic Data of 1000 samples

Inferences:

1. The correlation between the attribute x1 and x2 is high and positively correlated because as x1 is increasing, the value of x2 is also increasing.
2. The scatter plot is very dense around the mean 0.

b.

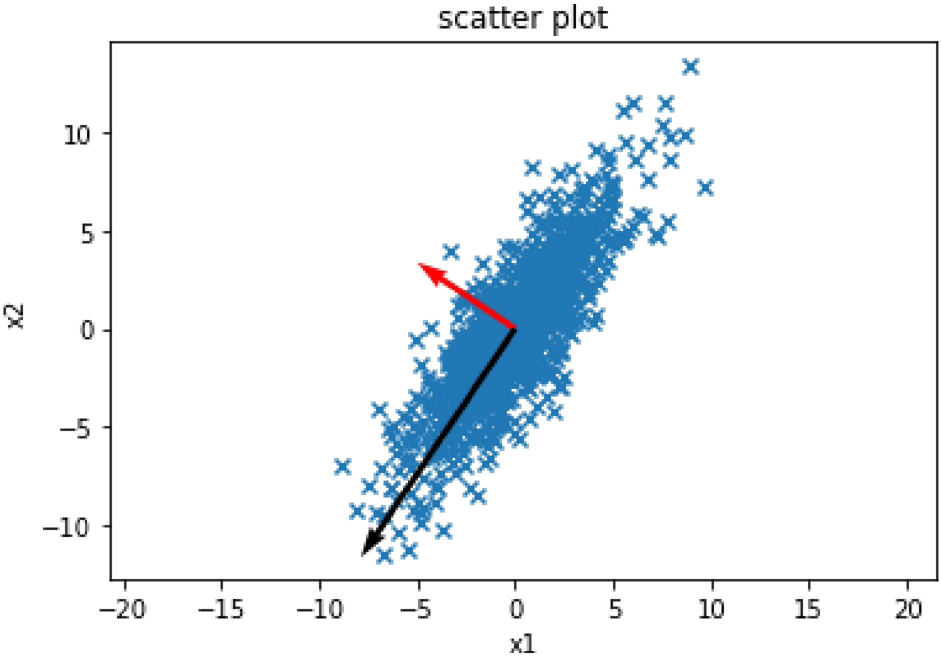

Figure 2 Plot of 2D Synthetic Data and Eigen Directions

Inferences:

1. The data is highly spread along the eigen value 18.169.
2. The eigen vector intersect at the origin (mean) where the data is very dense and the density decreases as we move along the 2nd eigen vector.
3. The variance along the eigen vector whose eigen value is 1.699 is less as compared to the other.

c.

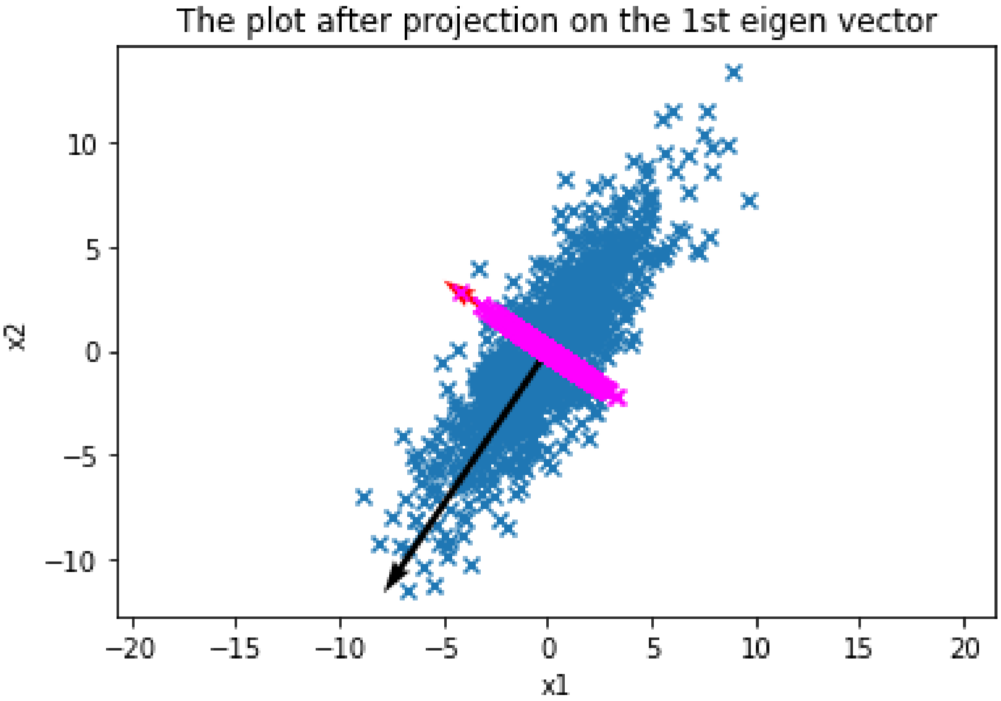

Figure 3 Projected Eigen Directions onto the Scatter Plot with 1st Eigen Direction highlighted

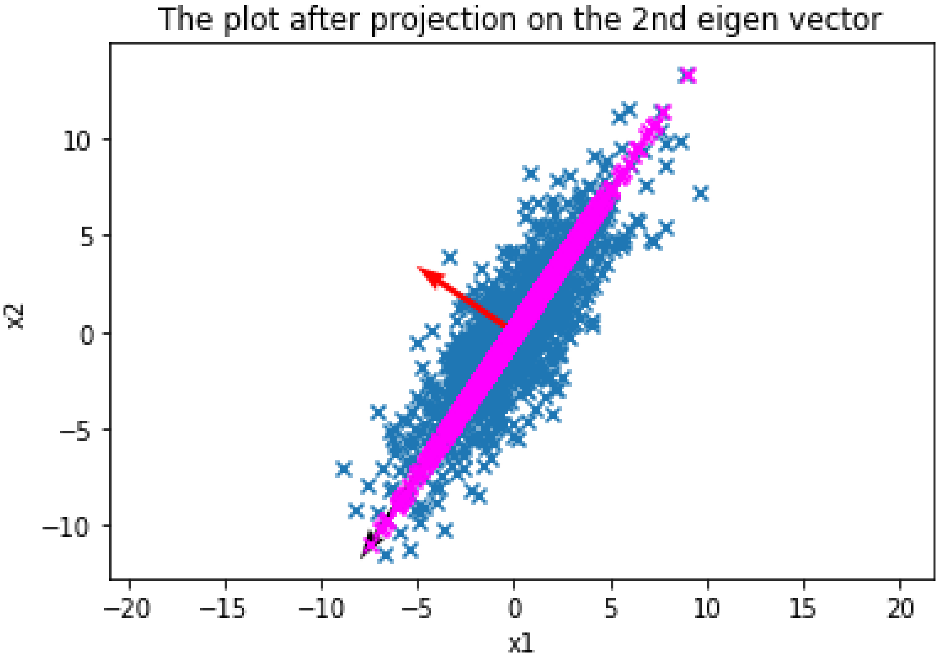

Figure 4 Projected Eigen Directions onto the Scatter Plot with 2nd Eigen Direction highlighted

Inferences:

1. The magnitude of the 1st eigen vector is **1.699** whereas the magnitude of the second eigen vector is **18.169**.
2. The figure3 shows that the projection the 1st eigen vector lies in small values because the magnitude of the eigen value is small.
3. The variance along the eigen vector whose eigen value is 1.699 is less as compared to the other.

d. Reconstruction Error = 1.135 * 10^-14

Inferences:

1. Here the actual dimension and the dimension after the reduction are same (i = 2), so root square mean error is almost zero. The error in reconstruction increases as we increase the dimension.

3 a.

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

Direction	Variance	Eigen Value
1	2.222	2.224
2	1.428	1.430

Inferences:

1. The variance and the eigen value is almost same for both the directions.
2. High eigen value means that the data is more spread along that eigen vector.

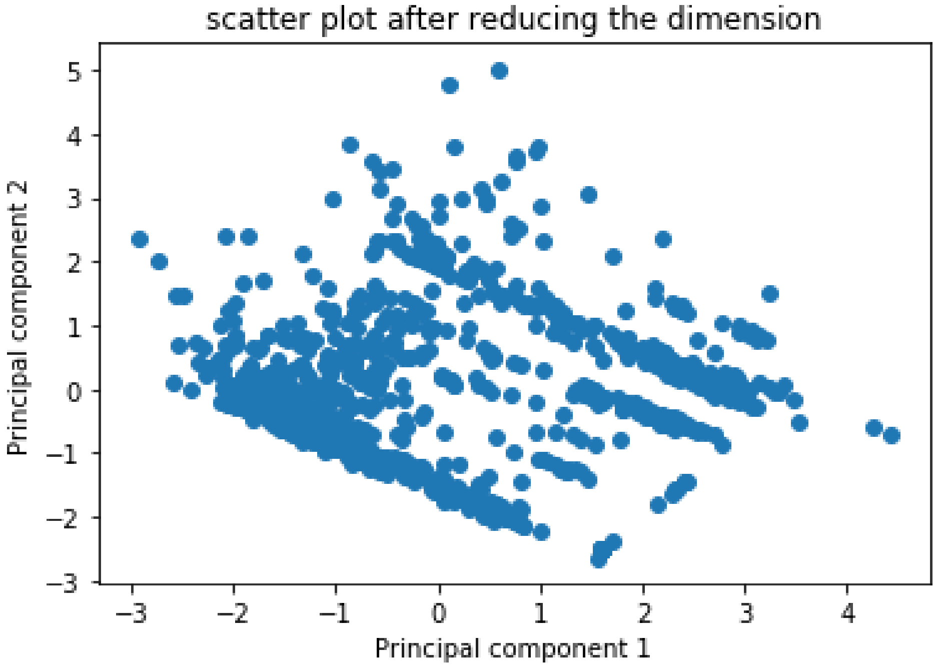

Figure 5 Plot of Landslide Data after dimensionality reduction

Inferences:

1. The data points is highly scattered after the reduction in dimensionality.

b.

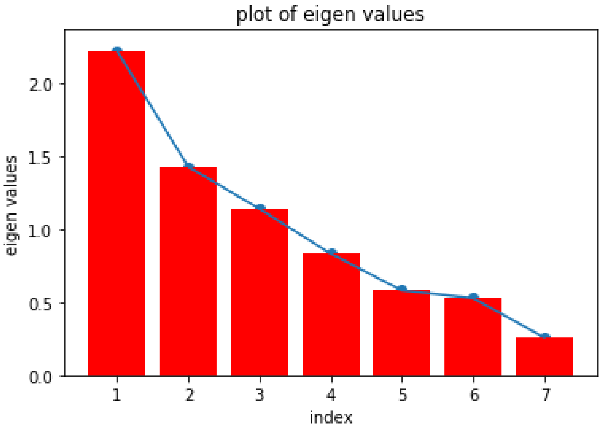

Figure 6 Plot of Eigen Values in descending order

Inferences:

1. The eigen value is decreasing gradually except from the 1st to 2nd . There is a sharp decrease in the eigen values from 1st to 2nd after that it decreases gradually.
2. After 2nd , the eigen value is decreasing substantially.

c.

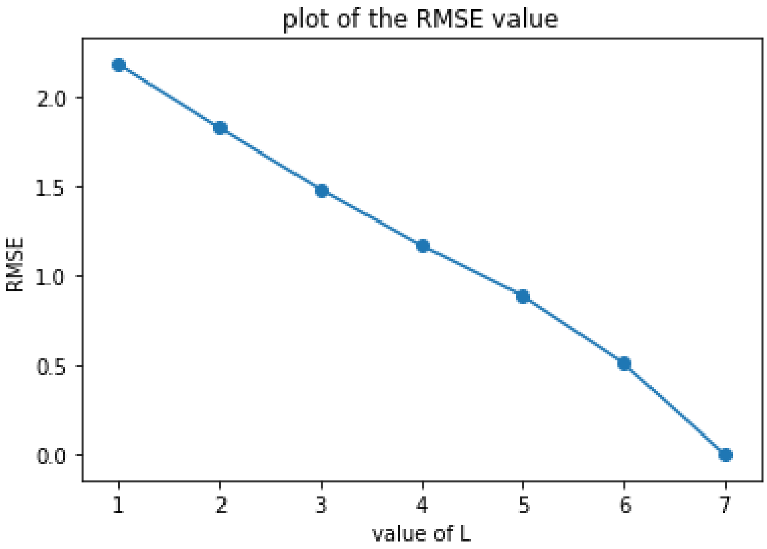

Figure 7 Line Plot to demonstrate Reconstruction Error vs. Components

Inferences:

1. The magnitude of the reconstruction is decreasing as the value of L increases.
2. It means that if we reduce the dimensionality to L which is same as the dimension of the original data then the reconstruction error is very less and we can say that the data is recovered more accurately if L increases.