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

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	9
2	Humidity (in $g.m^{-3}$)	34.206	99.72	3	9
3	Pressure (in mb)	992.655	1037.604	3	9
4	Rain (in ml)	0.0	2470.5	3	9
5	Lightavgw/o0 (in lux)	0.0	10565.352	3	9
6	Lightmax (in lux)	2259	54612	3	9
7	Moisture (in %)	0.0	100.0	3	9

Inferences:

1. Inference related to outliers : After replacing all the outliers with median of the data, in some cases no outliers is present but for some cases outliers are still present.
2. Before normalization :The min max values of each and every attribute is different.
3. After normalization : The min max values are same for each cases.

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)	21.37	4.125	-0.0	1.0
2	Humidity (in $g.m^{-3}$)	83.992	17.566	0.0	1.0
3	Pressure (in mb)	1014.761	6.121	-0.0	1.0
4	Rain (in ml)	168.4	399.689	-0.0	1.0
5	Lightavgw/o0 (in lux)	2197.392	2220.820	-0.0	1.0
6	Lightmax (in lux)	21788.623	22064.993	0.0	1.0
7	Moisture (in %)	32.386	33.653	0.0	1.0

Inferences:

1. Before standardization : The values of both mean and standard deviation are quite varying i.e the range of both values is higher.
2. After standardization : The values of mean is equal to 0 and that of of standard deviation equals to 1 for each and every attribute .

2 a.

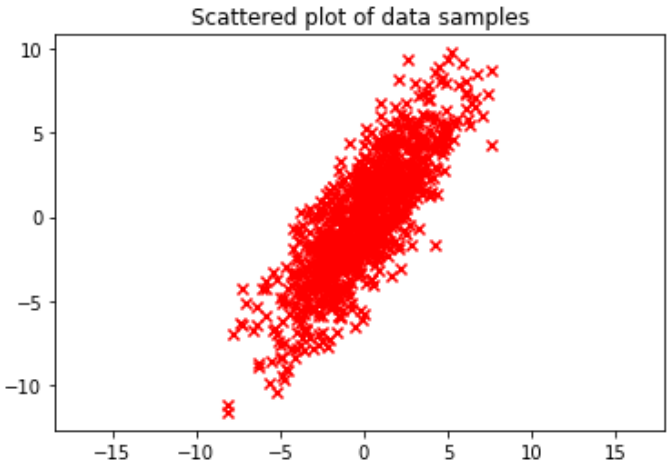

Figure 1 Scatter Plot of 2D Synthetic Data of 1000 samples

Inferences:

1. The data points of both the data sample makes a straight line with positive slope showing a **positive correlation** between them.
2. Density of plot is mainly high in the middle of the scattered plot i.e. the range of value of both x and y is approximately from -5 to +5.

b.

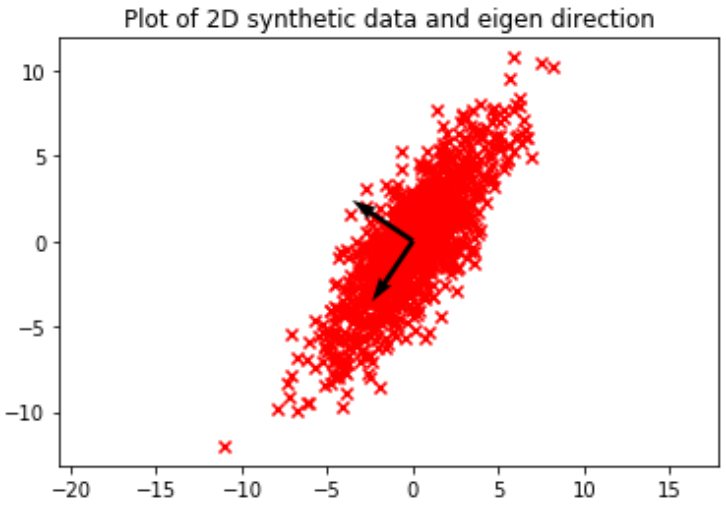

Figure 2 Plot of 2D Synthetic Data and Eigen Directions

Inferences:

1. Spread of data is mainly based upon the magnitude of Eigen values as the spread is less in center as compare to the ends of the eigen vector.
2. Density of points near the intersection of Eigen axes is quite high as compare to the end as it is gradually decreasing by going away from the intersection as the spread increases towards the pointer of the eigen directions.

c.

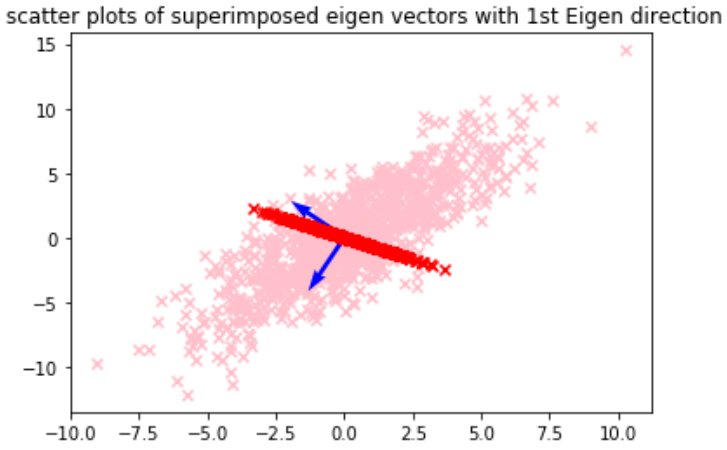

Figure 3 Projected Eigen Directions onto the Scatter Plot with 1st Eigen Direction highlighted

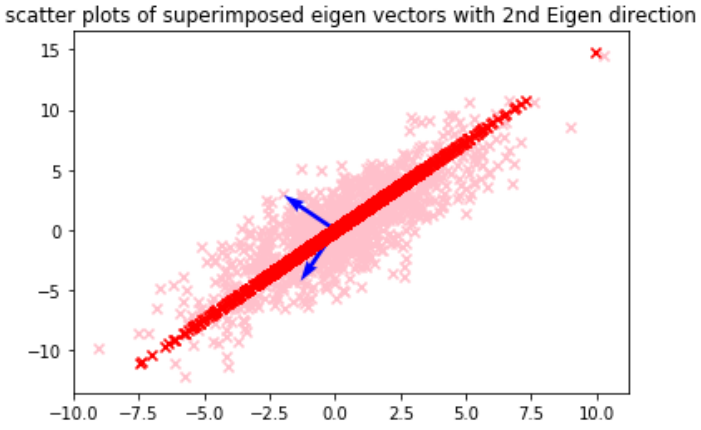

Figure 4 Projected Eigen Directions onto the Scatter Plot with 2nd Eigen Direction highlighted

Inferences:

1. Magnitude of Eigen Values is quite small of 1st data sample as compare to the other one resulting in less number of projection on 1st eigen direction as compare to the other one and vice versa.
2. Variance of data along the 1st Eigen axes from spread is smaller as the correlation is less as compare to the 2nd Eigen axes & density of points in the center of the intersection is high as the compare to the pointing direction and the magnitude of projected values is higher for the Eigen values having higher magnitude .

d. Reconstruction Error is :

For 1st column = 3.332

For 2nd column = 7.945

Inferences:

1. The magnitude of reconstruction error affects the quality of reconstruction as higher the value of reconstruction error more be the deviation from the original data happen result in reducing the quality of reconstruction.

3 a.

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

Direction	Variance	Eigen Value
1	2.202	2.202
2	1.421	1.421

Inferences:

1. The variance of the projected data along the two directions with the Eigen values of the two directions of projection is equal for both the directions.

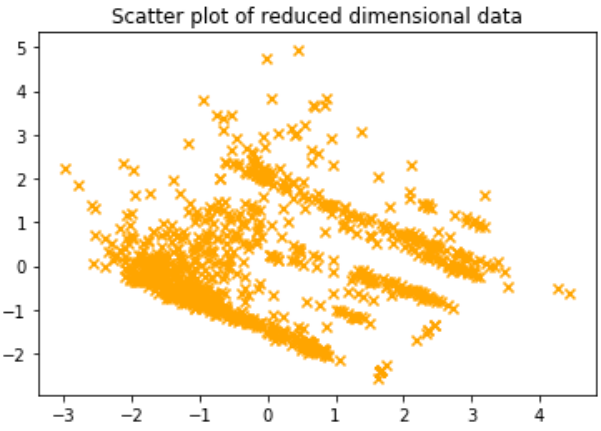

Figure 5 Plot of Landslide Data after dimensionality reduction

Inferences:

1. The spread of data points increases and density of data points decreases after dimensionality reduction resulting in higher spread at the center also resulting in less correlation between both reduce sample attribute.

b.

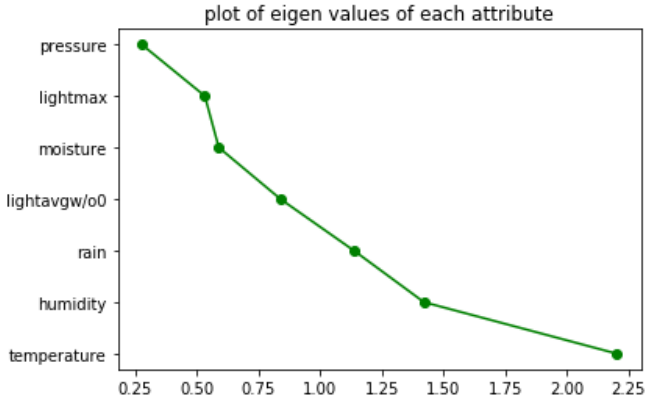

Figure 6 Plot of Eigen Values in descending order

Inferences:

1. The subsequent Eigen values decrease gradually.
2. From the eigen value of humidity (i.e. 1.5) the rate of decrease changes substantially

c.

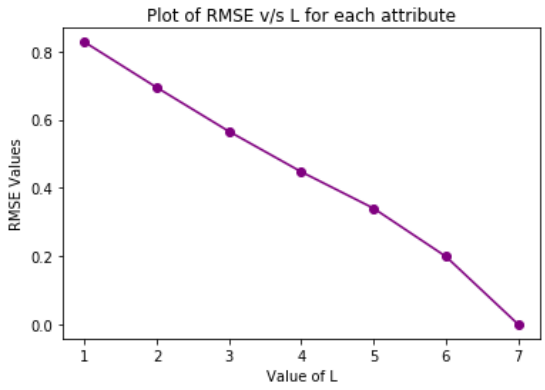

Figure 7 Line Plot to demonstrate Reconstruction Error vs. Components

Inferences:

1. Higher the magnitude of reconstruction error lower the quality of reconstruction as in this case the value of reconstruction error is quite small resulting in a good quality of reconstruction.