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

Student's Name: Ankit **Mobile No: 9053437219**

Roll Number: B19236 **Branch:Mech**

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})	34.206	99.720	3.0	9.0
3	Pressure (in mb)	992.654	1037.604	3.0	9.0
4	Rain (in ml)	0.000	2470.500	3.0	9.0
5	Lightavgw/o0 (in lux)	0.000	10565.352	3.0	9.0
6	Lightmax (in lux)	2259.000	54612.000	3.0	9.0
7	Moisture (in %)	0.000	100.000	3.0	9.0

Inferences:

1. Outliers are replaced with median of the remaining data. It is necessary as outliers affects the range of the data.
2. Before normalization the range of attributes were different. After normalization the range of all attributes becomes the same (3-9).
3. Normalization helps to prevent attributes with large ranges from overweighting attributes with smaller attributes.

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)	21.370	4.125	0.0	1.0
2	Humidity (in g.m^{-3})	83.992	17.566	0.0	1.0
3	Pressure (in mb)	1014.760	6.121	0.0	1.0
4	Rain (in ml)	168.400	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 mean and standard deviation of the attributes are of different values. After standardization the mean becomes 0 and the standard deviation becomes 1 for all attributes in the data.
2. It is useful when the actual minimum and maximum of attribute are unknown.

2 a.

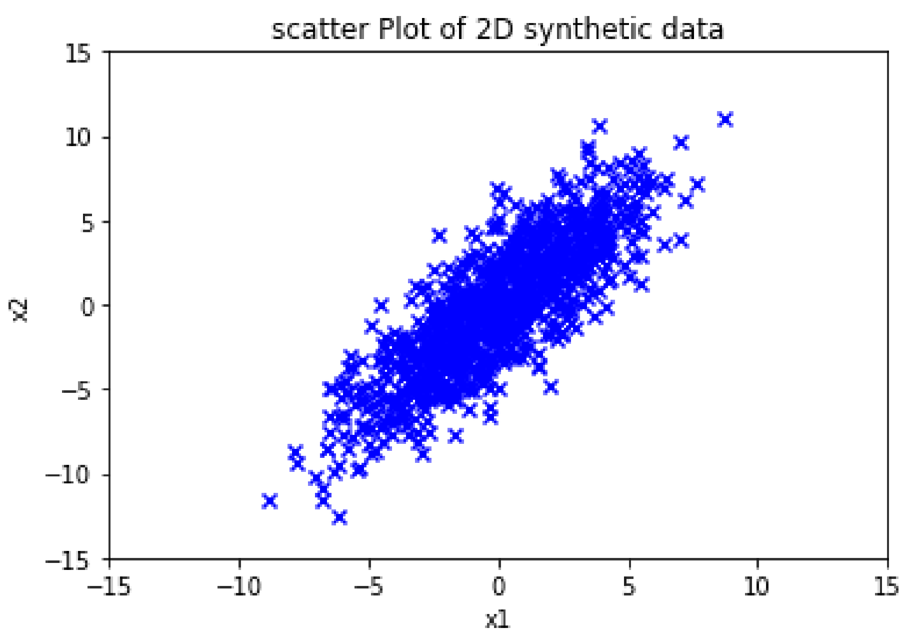

Figure 1 Scatter Plot of 2D Synthetic Data of 1000 samples

Inferences:

1. Attributes seem to have a strong positive correlation. As the value of x1 increases x2 also increases. On computation the Pearson's correlation coefficient is around 0.82 same as expected from the covariance matrix.
2. The plot has high density around origin (mean value). The variance of x1 is smaller than that of x2 as expected
3. The plot shows the Gaussian bivariate distribution.

b.

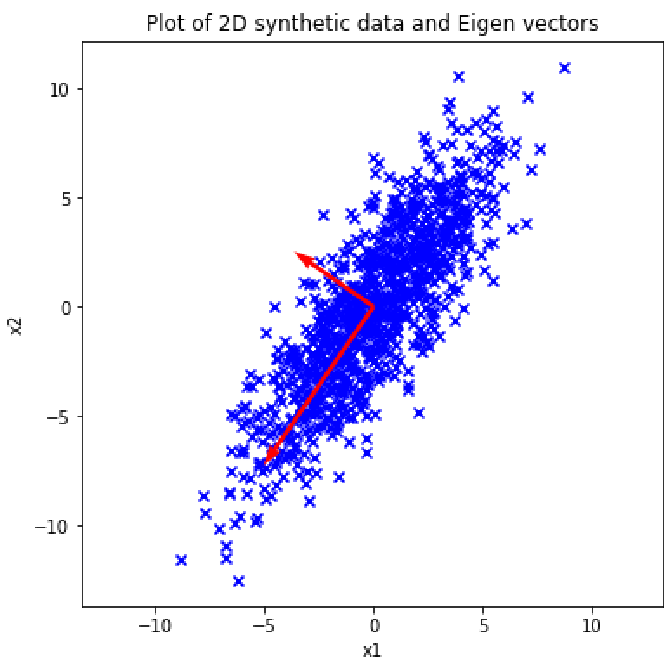

Figure 2 Plot of 2D Synthetic Data and Eigen Directions

Inferences:

1. Eigen value1: 1.703, Eigen vector1: [-0.833, 0.554]
2. Eigen value2: 19.688, Eigen vector2: [-0.554, -0.833]
3. We can observe that the spread of the data is more across 2nd eigen vector than the 1st. This is because the magnitude of eigen value 2 is greater than eigen value 1.
4. The plot has high density around the origin (point of intersection of eigen axes), since it is the mean of the distribution as we move away along 2 axes the density decreases.
5. Larger the eigen value, larger the distribution of data along the corresponding eigen vector.

c.

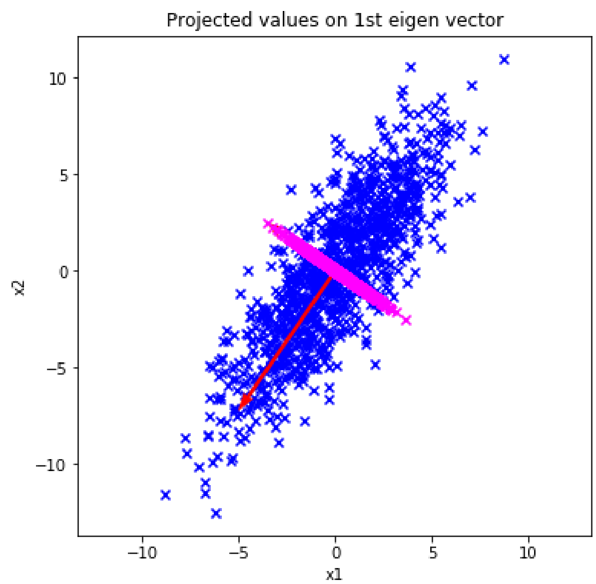

Figure 3 Projected Eigen Directions onto the Scatter Plot with 1st Eigen Direction highlighted

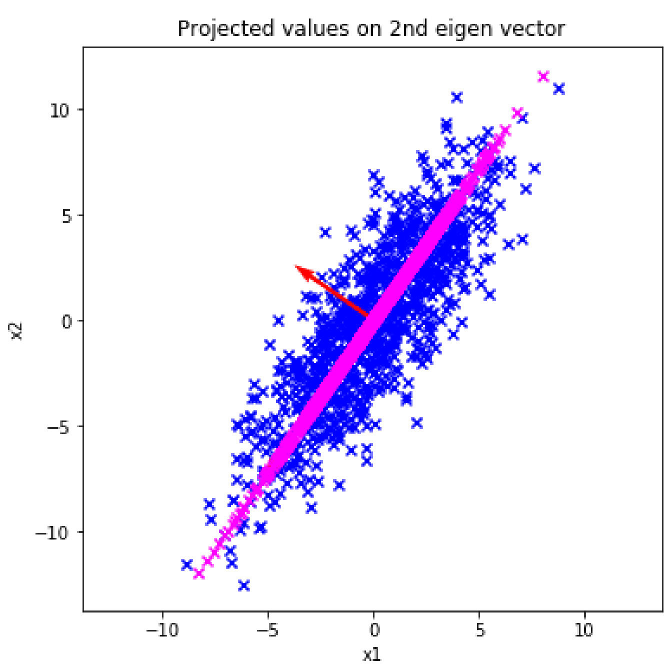

Figure 4 Projected Eigen Directions onto the Scatter Plot with 2nd Eigen Direction highlighted

Inferences:

1. Eigen value1: 1.703, Eigen value2: 19.688. Eigen value ∝ variance in projection.
2. Variance of data along eigen vector2> variance of data along eigen vector2. Variance along eigen vector ∝ spread of data along eigen vector 1/∝ density of data. Variance along eigen vector = Magnitude of eigen value.
3. Larger the eigen value, larger the information content in the direction of corresponding eigen vector.

d. Reconstruction Error = 0.000

Inferences:

1. Magnitude of reconstruction error ∝ Loss of information in compressed data.
2. Reconstruction Error= 0 <=> The data reduction is called lossless.

3 a.

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

Direction	Variance	Eigen Value
1	2.1999	2.2022
2	1.4193	1.4208

Inferences:

1. Eigenvalues and variances of the directions of projection in this reduced data are numerically very close meaning eigenvalues signify the spread/variance of data around a direction of projection.

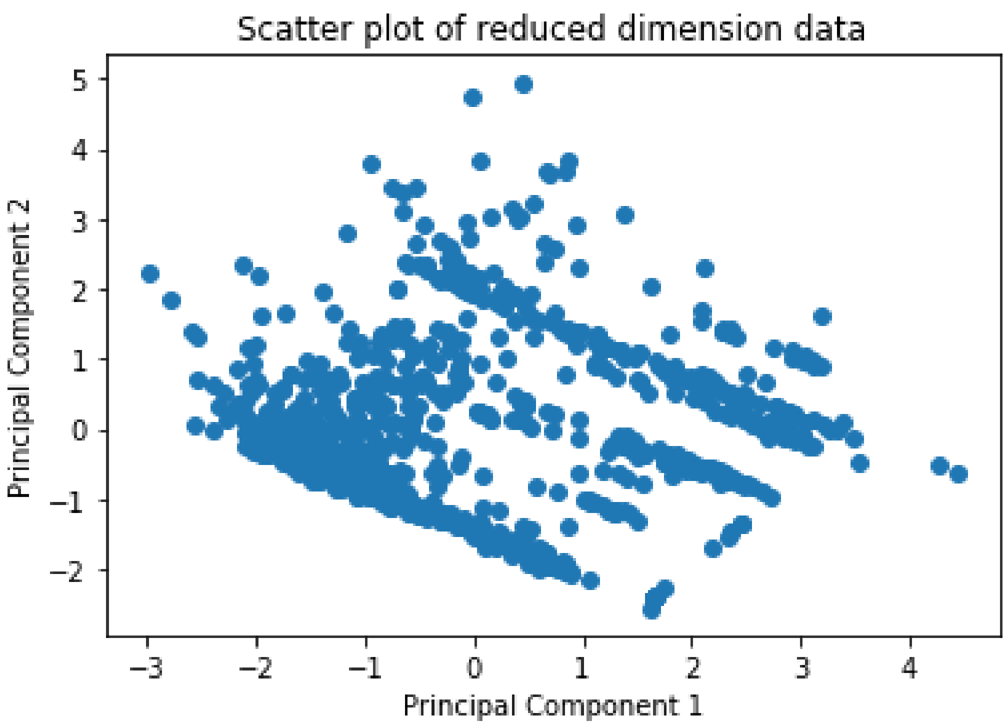

Figure 5 Plot of Landslide Data after dimensionality reduction

Inferences:

1. Since the number of eigendirections and the original dimensions of the data are the same, no actual dimension reduction has been performed. The data points only have been projected onto a new basis.
2. Therefore, the MSE calculated for this instance is vanishingly close to zero as the "reduced data" takes up the same number of dimensions as previous data.
3. From the plot the median of both attributes of the reduced data seem to be less than the mean (positively skewed).
4. The reduced data is uncorrelated.

b)

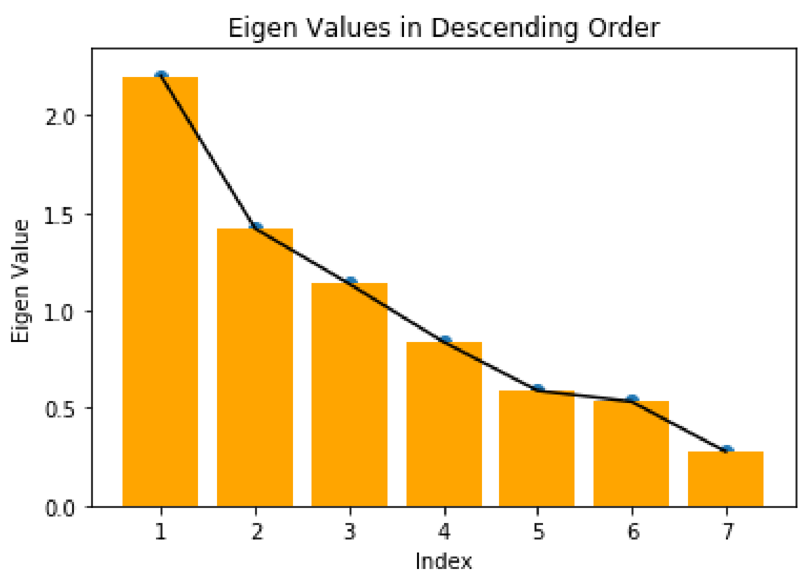

Figure 6 Plot of Eigen Values in descending order

Inferences:

1. Eigen values decrease gradually.
2. Highest rate of decrease is from eigen value 1 to eigen value 2.

c.

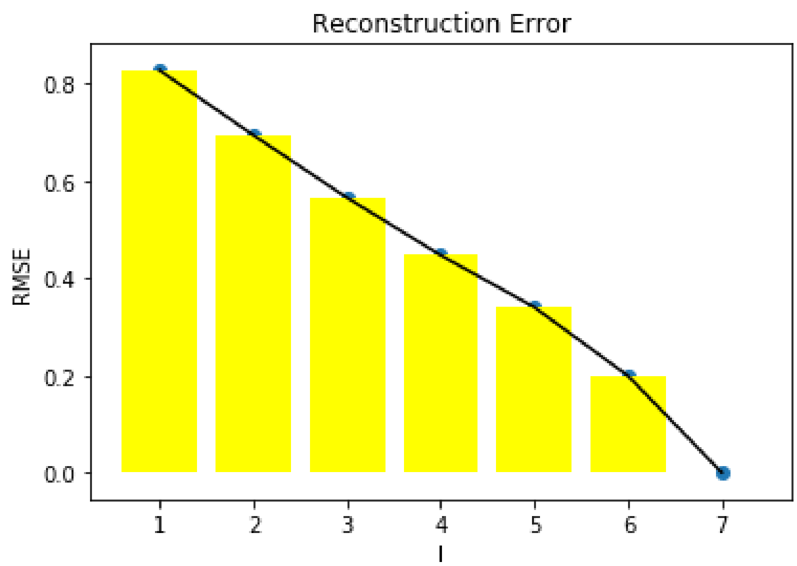

Figure 7 Line Plot to demonstrate Reconstruction Error vs. Components

Inferences:

1. Magnitude of reconstruction error 1/∝ the quality of reconstruction.
2. As l -> d, reconstruction error -> 0.
3. Reconstruction error = 0 => Data reduction is lossless. If reconstruction error ≠ 0 => The data reduction is called lossy.