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

Student's Name: Pooja Patidar **Mobile No:** 8516921968

Roll Number: B19255 **Branch:**Mechanical

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	100.00	3.0	9.0

Inferences:

1. Outliers are replaced by median of the remaining data.
2. Before normalization, the range of attributes were very different and varies with very high number. But after normalization the range of all attributes become same i.e., min (3.0) and max (9.0).
3. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

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	0.0	1.0
2	Humidity (in $g.m^{-3}$)	83.992	17.565	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 were different. But after standardization, the mean becomes 0 and the standard deviation becomes 1 for all attributes in the data.
2. The method of normalization is useful when actual minimum and maximum are unknown and when the outliers dominant the min – max normalization.

2 a.

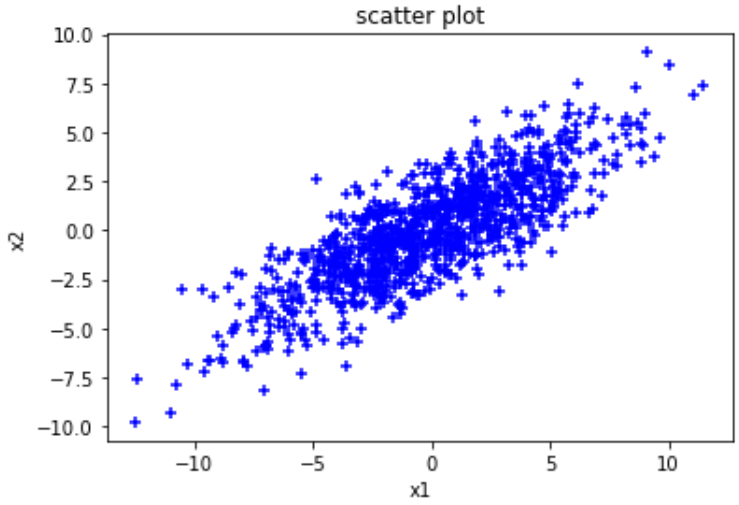

Figure 1 Scatter Plot of 2D Synthetic Data of 1000 samples

Inferences:

1. According to graph, attributes seems to be strong positive correlation because as x1 increases x2 also increases.
   On computation we got below correlation matrix
   [[1.0 0.8269]
   [0.8269 1.0]]
2. The data is highly dense at origin which implies that the data is mean subtracted data.
3. The graph shows the Gaussian bivariant distribution i.e., have 2-dimension matrix.

b.

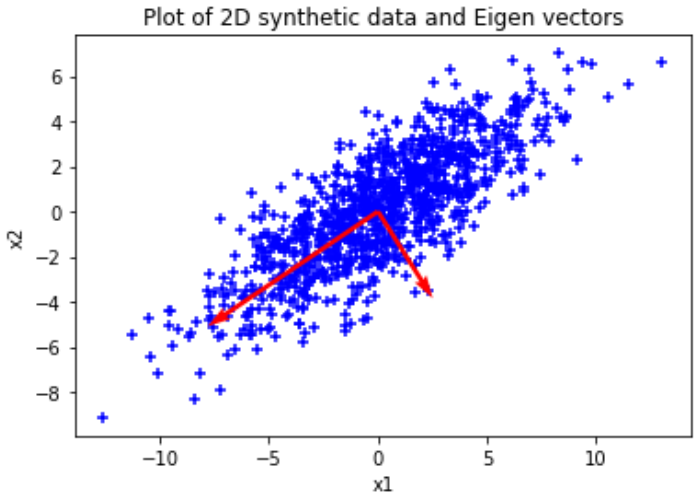

Figure 2 Plot of 2D Synthetic Data and Eigen Directions

Inferences:

1. Eigen values: 1.909 19.667
2. Eigen vectors: [-0.829 -0.558] [0.558 -0.829]
3. We observe that spread of data is more across eigenvector having value 19.667. It implies that larger the eigenvalue more the spread of data along the corresponding eigenvector.

c.

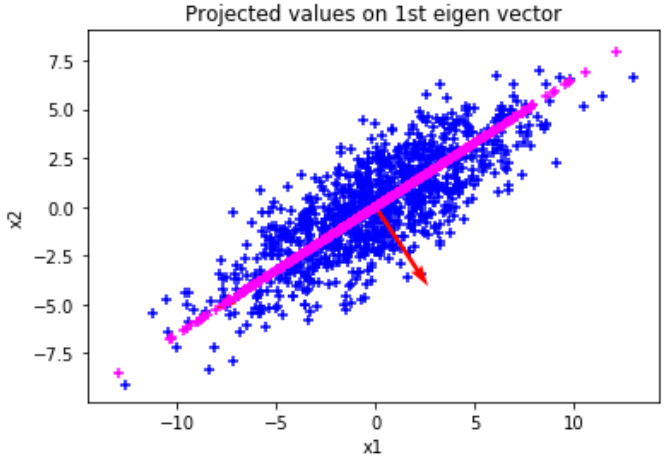

Figure 3 Projected Eigen Directions onto the Scatter Plot with 1st Eigen Direction highlighted

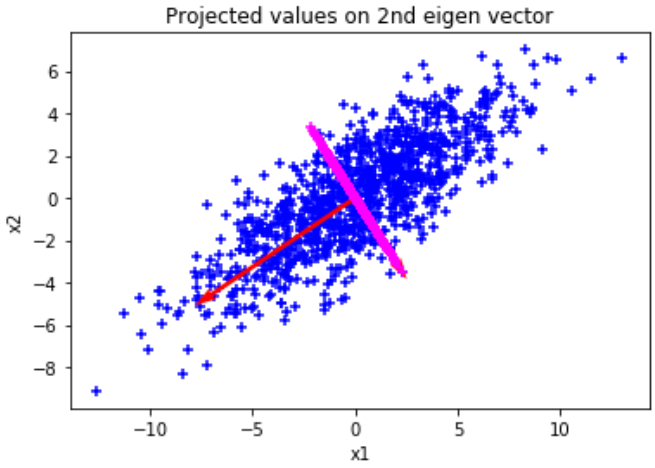

Figure 4 Projected Eigen Directions onto the Scatter Plot with 2nd Eigen Direction highlighted

Inferences:

1. | | |
|---|---|
| Eigenvalue 1 | 1.909 |
| Eigenvalue 2 | 19.667 |

   Eigenvalue is directly proportional to the variance of projection.
2. Larger the eigen value, larger the information content in the direction of corresponding eigen vector.

d. Reconstruction Error = 0.000

Inferences:

1. Magnitude to reconstruction error is directly related to the loss of information in compressed data.
2. If the original data can be reconstructed from compressed data without any loss of information, 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	2.222	2.2246
2	1.428	1.4300

Inferences:

1. Variance of the projected data along the 2 directions are same as the eigen values along that direction.

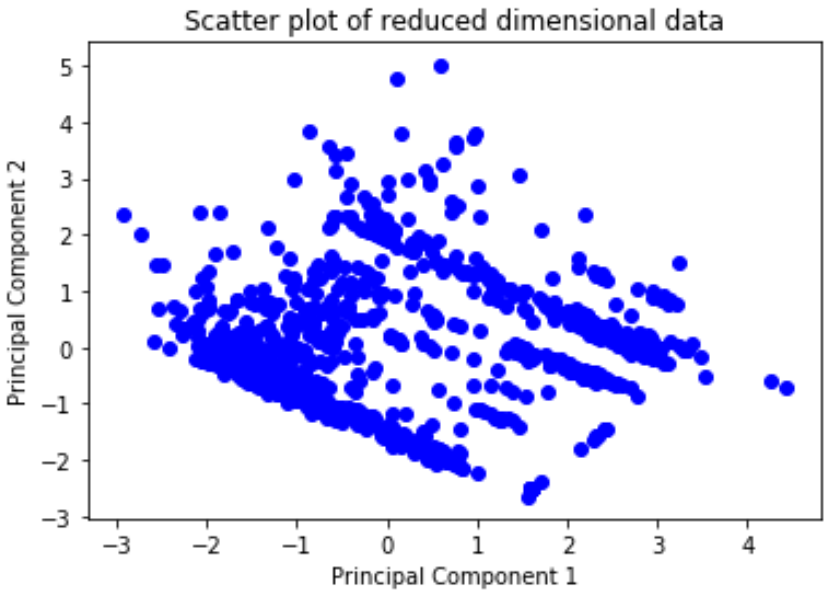

Figure 5 Plot of Landslide Data after dimensionality reduction

Inferences:

1. The density around the median of reduced data is very high and reduces as we move away. The reduced data seem to follow a skewed Gaussian distribution. The variance of each attribute of the reduced data is the eigen value corresponding to it.
2. Along principal component 1 data ranges from -3 to 5 and along principal component 2 it ranges from -2 to 5. After standardization mean of the data was shifted to 0 which explains the high density of points near (0,0).
3. The reduced data is uncorrelated

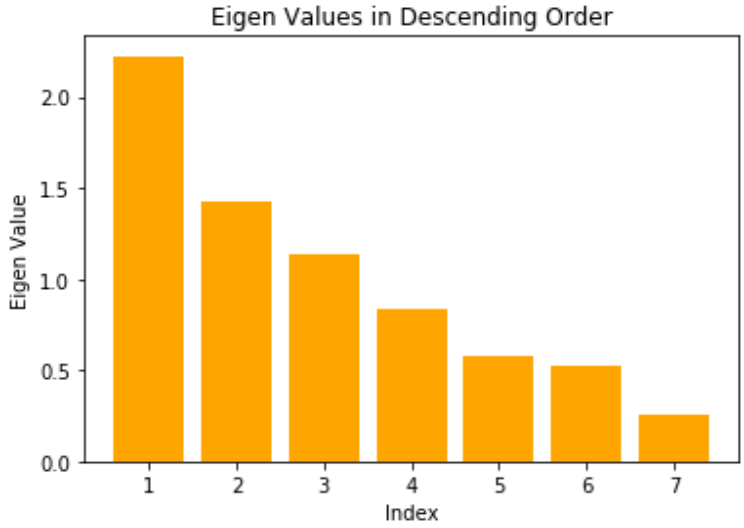

Figure 6 Plot of Eigen Values in descending order

Inferences:

1. Eigenvalue decreases rapidly from 1 to 2 but after that the decrement is quite slow.
2. Decrease change is substantial from 1 to 2.

c.

Figure 6 Line Plot to demonstrate Reconstruction Error vs. Components

Inferences:

1. As the value of x increases from left to right, we keep on introducing more dimensions to project data. And with increase in number of dimensions, data projection become more and more precise which explains the decreasing RMSE value from left to right.
2. If the original data can be reconstructed from compressed data without any loss of information, the data reduction is called lossless.
3. If only an approximation of the original data can be reconstructed from compressed data, then the data reduction is called lossy.
4. One of the popular and effective methods of lossy dimensionality reduction is principal component analysis (PCA)