IC 272: DATA SCIENCE - III

LAB ASSIGNMENT – III

Attribute Normalization, Standardization and Dimension Reduction of Data

Student's Name: Gajraj Singh Chouhan **Mobile No: +91-9351159849**

Roll Number: B19130 **Branch: DSE**

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 $g.m^{-3}$)	34.205	99.72	3	9
3	Pressure (in mb)	992.654	1037.60	3	9
4	Rain (in ml)	0	2470.5	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. Total number of outliers : 419
2. After normalization we have scaled all the data between 3 and 9. So the minimum and maximum values now becomes 3 and 9.

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)	21.36938	4.12316	0	1
2	Humidity (in $g.m^{-3}$)	83.99102	17.55606	0	1
3	Pressure (in mb)	1014.79313	6.11695	0	1
4	Rain (in ml)	171.46668	398.24884	0	1
5	Lightavgw/o0 (in lux)	2237.89984	2205.25512	0	1
6	Lightmax (in lux)	21788.62328	22053.31540	0	1
7	Moisture (in %)	32.38605	33.63543	0	1

Inferences:

1. After standardization the mean and std. deviation becomes 0 and 1 respectively.

2. a.

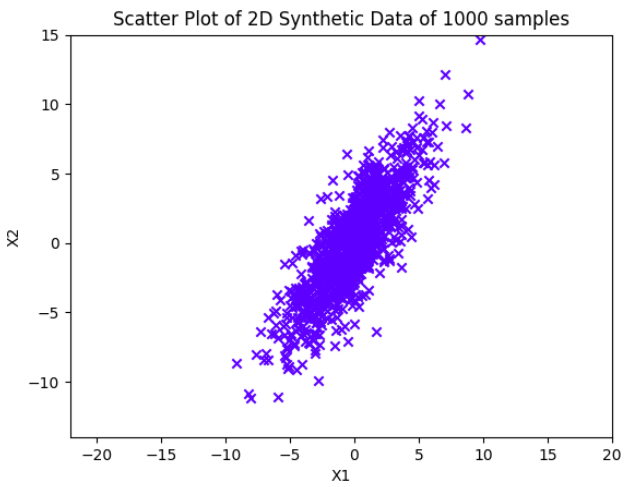


Figure 1 Scatter Plot of 2D Synthetic Data of 1000 samples

Inferences:

1. Their correlation is positive as we can see in the graph, as X1 is increasing, X2 is also increasing.
2. The relationship between x1 and x2 is strong
3. The data is closely packed and very close to each other. In other words, if we fit a line, our data is closely populated around that line

b.

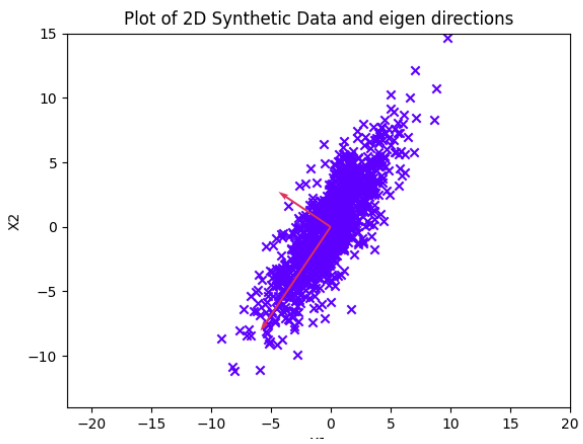


Figure 2 Plot of 2D Synthetic Data and Eigen Directions

Inferences:

Eigen Values	
1.699 (e1)	18.169 (e2)

1. Inference based on density of points near the intersection of Eigen axes and gradually away from it.
2. Density of points near the intersection of Eigen axes is larger and gradually decreases away from it.
3. Given eigen vectors are perpendicular and they form the basis

c.

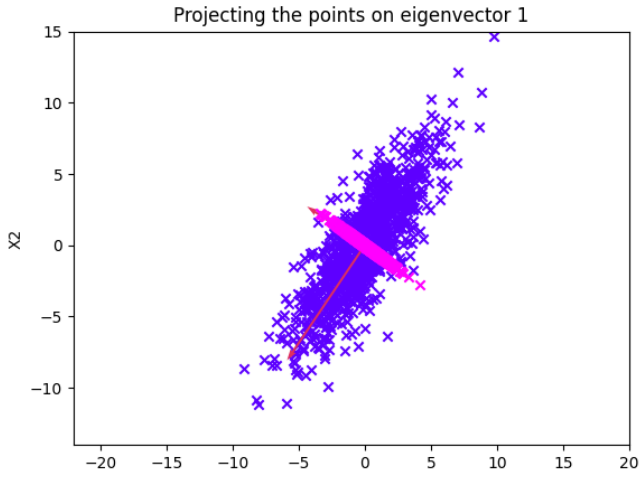


Figure 3 Projected Eigen Directions onto the Scatter Plot with 1st Eigen Direction highlighted

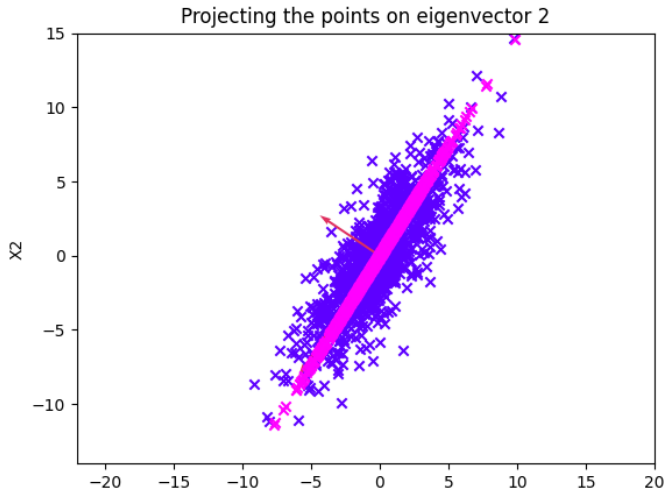


Figure 4 Projected Eigen Directions onto the Scatter Plot with 2nd Eigen Direction highlighted

Inferences:

1. The eigen value for the high variance eigen vector will be higher.
2. Infer variance of data along the Eigen axes from spread & density of points and relate it to the magnitude of Eigen values.

d.

1. Reconstruction Error = 3.6369203394438965e-16 ~= 0

Inferences:

1. As the reconstruction error increases the quality of reconstruction will be decreasing.
2. Reconstruction error is almost nearly equal to zero
3. If we use PCA from sklearn and check the variance we can see more than 91% [0.9114813 0.0885187] variance can be recovered and it could be more beneficial to use that.

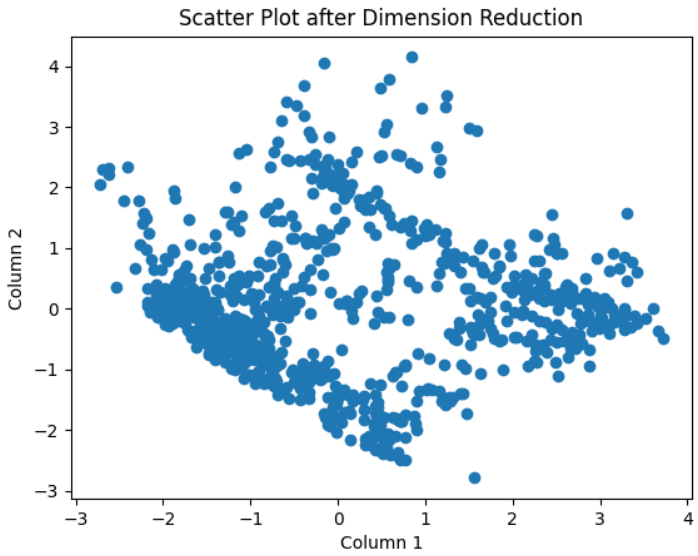
3 a.

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

Direction	Variance	Leading Eigen Values from the Data
1	2.278	2150.53
2	1.387	1310.26

Inferences:

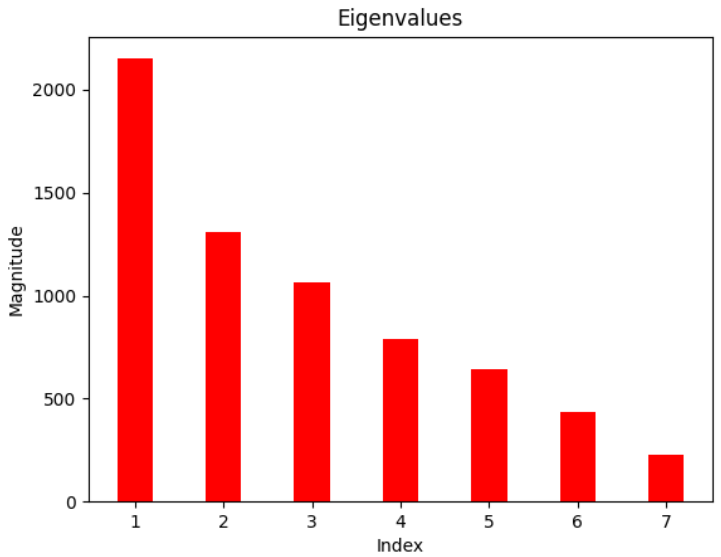
1. Spread of data is more along the higher variance eigen vector.



Inferences:

1. After dimensionality reduction, Now the data is Negatively correlated
2. These two attributes are representing our entire old data with only two dimensions but with compromise of some loss.

b.

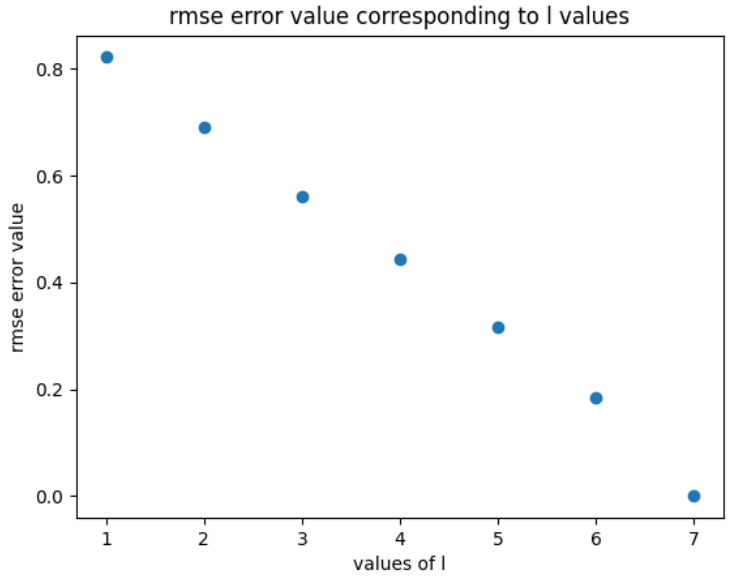


EigenValues from the Correlation Matrix

Inferences:

1. The eigen values are in decreasing order.
2. After the largest eigenvalue (index #1), the rate of decrease has gone up as there has been a drop of almost 1000.

c.



Inferences:

1. As the number of dimension the data is being reduced gets increased, the error is getting lower. This is expected as sum of variance will be increasing and then we can accurately represent the data with lower error.