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

Student's Name: Srishti Ginjala **Roll Number: b19084**

Mobile No: 9440000900 **Branch: CSE**

1a)

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	3.000	9.000
2	Humidity (in g.m^{-3})	34.206	99.720	3.000	9.000
3	Pressure (in mb)	992.655	1037.604	3.000	9.000
4	Rain (in ml)	0.000	2470.500	3.000	9.000
5	Lightavgw/o0 (in lux)	0.000	10565.352	3.000	9.000
6	Lightmax (in lux)	2259.000	54612.000	3.000	9.000
7	Moisture (in %)	0	100	3.000	9.000

Inferences:

1)We don't get any outliers for temperature, moisture, lightmax, lightavgw/o0 and humidity after normalization. But for rain we still get huge number of outliers because the values are spread over a wide range and we calculated median without removing outliers. For pressure we get few outliers on both extremes.

2)If we don't normalize data, then the attributes with large values will dominate over the others. This does not mean that they are more important in prediction. So, after normalization, all the attributes are in a common scale and equally influence the model.

b.

Table 2 Mean and Standard Deviation Before and After Standardization

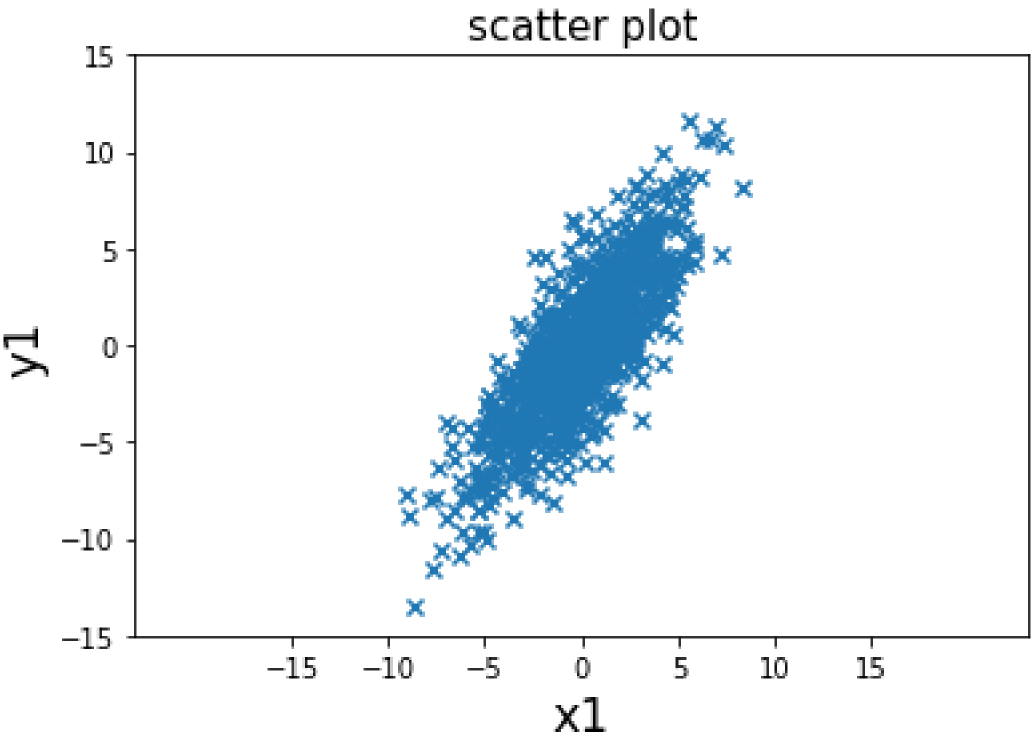
S. No.	Attribute	Before Standardization	After Standardization

		Mean	**Std. Deviation**	**Mean**	**Std. Deviation**
1	Temperature (in °C)	21.215	4.356	0	1
2	Humidity (in g.m^{-3})	83.480	18.210	0	1
3	Pressure (in mb)	1009.009	46.980	0	1
4	Rain (in ml)	10701.538	24852.255	0	1
5	Lightavgw/o0 (in lux)	4438.428	7573.163	0	1
6	Lightmax (in lux)	21788.623	22064.993	0	1
7	Moisture (in %)	32.386	33.653	0	1

Inferences:

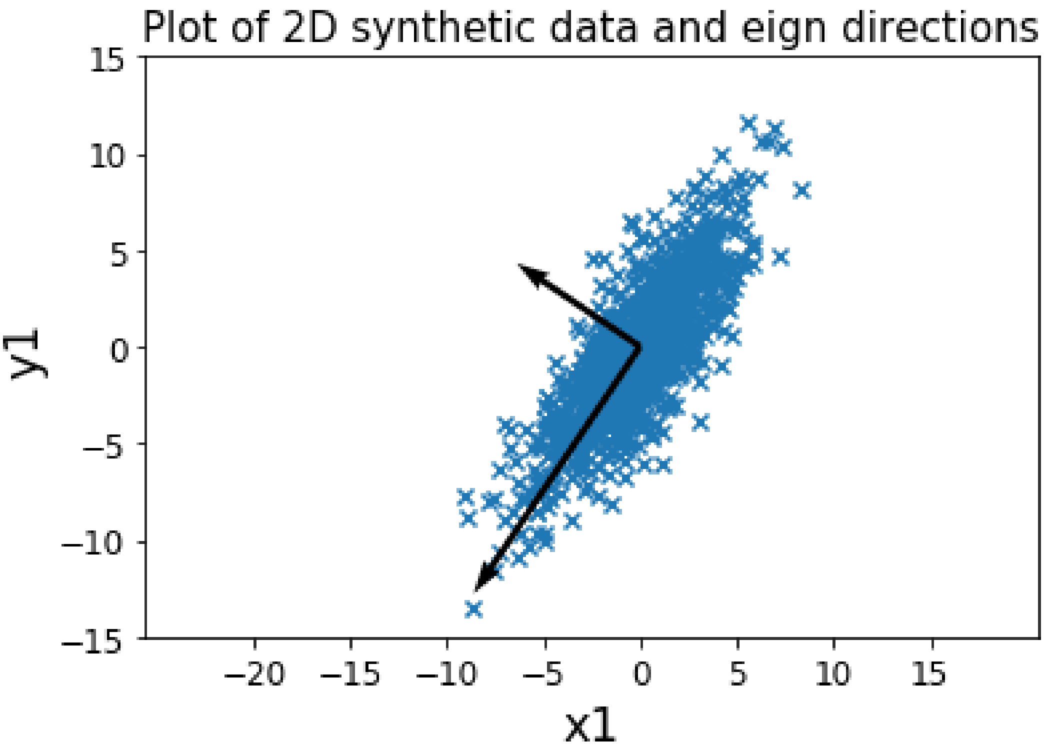
1)We standardize data to avoid out of bound error in test phase. The problem with normalization is that if the test cases exceed the maximum or less than minimum, our prediction may go wrong. So, to avoid that we standardize data with respect to their mean and standard deviation (z-score analysis)

2a)



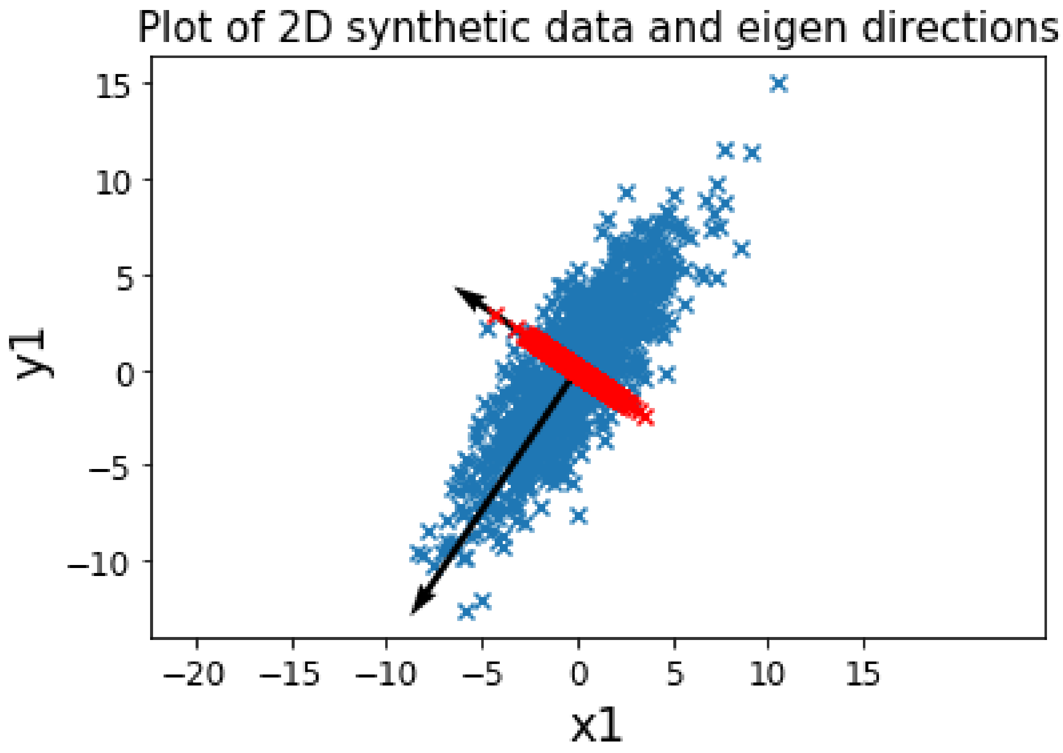
1) We infer that attribute 1 is positively correlated to attribute2 as the plot is linear with a high value of Pearson’s correlation coefficient.
2) The density of the attributes is more, hence they are highly correlated.

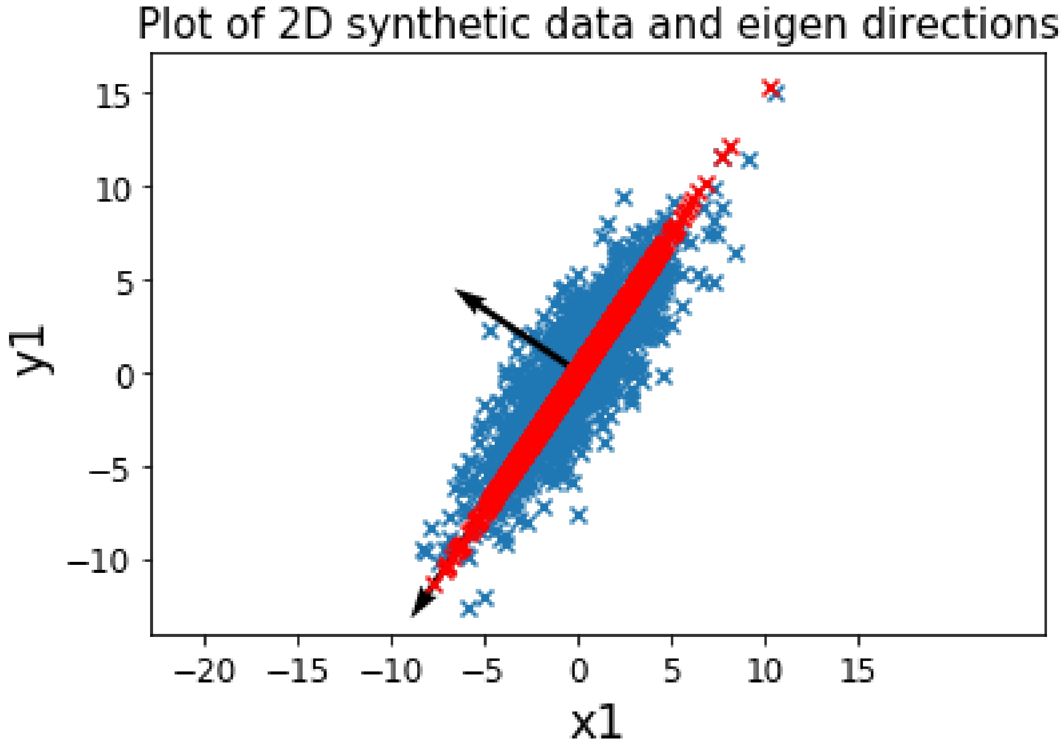
2b)



1) The eigen values are: 1.700,18.169. We can infer that higher the eigen value more is the variance along that direction which means more information content of data and lesser the eigen value, spread of data is less.
2) The density of points near the intersection of Eigen axes is more and decreases gradually away from it because the data was plotted according to normal distribution, so probability of finding data points decreases away from the mean value (that is the intersection of Eigen axes).

2c)





Inferences:

1) The magnitude of eigen values are 1.700 and 18.169. We observe that the eigen values differ hugely. This tells us that the data is more spread in a

single direction than the direction whose eigen value is less (as is indicated by the red line).

2) Eigen values indicate the strength or magnitude of eigen vectors. Greater the magnitude of eigen value means spread is more along that direction and information is preserved when projected along that direction and vice versa.

2d)

Reconstruction errors are 19.798 and 1.644.

Inference:

1)More reconstruction error means that the reconstruction is less accurate.

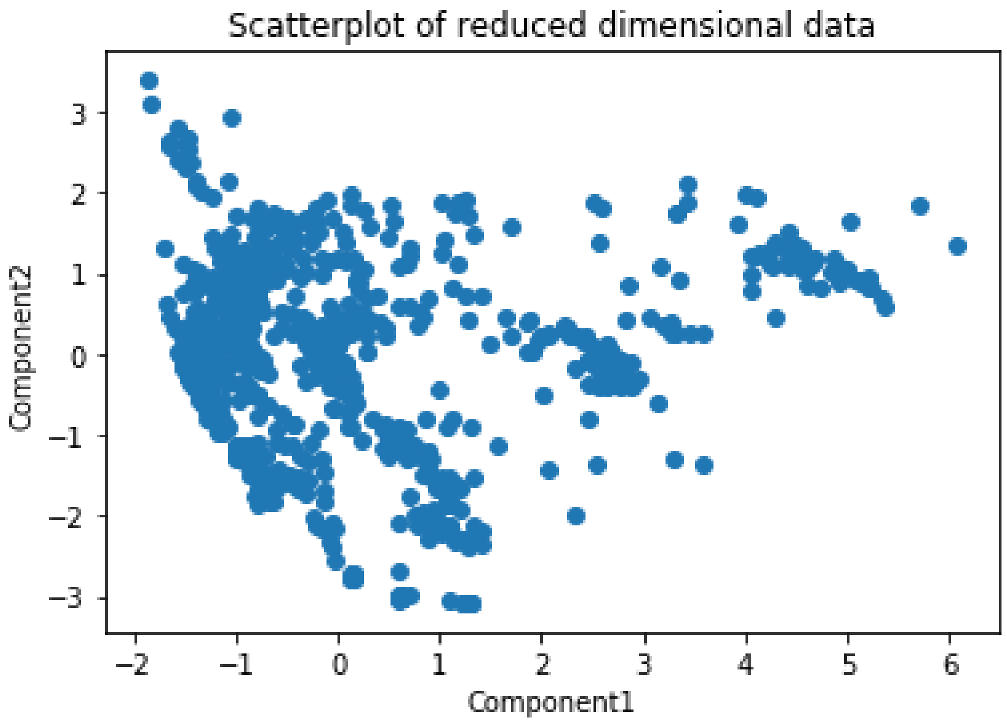
3a)

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

Direction	Variance	Eigen Value
1	2.467	2.467
2	1.683	1.683

We infer that eigen values are the same as the variance along the two directions.

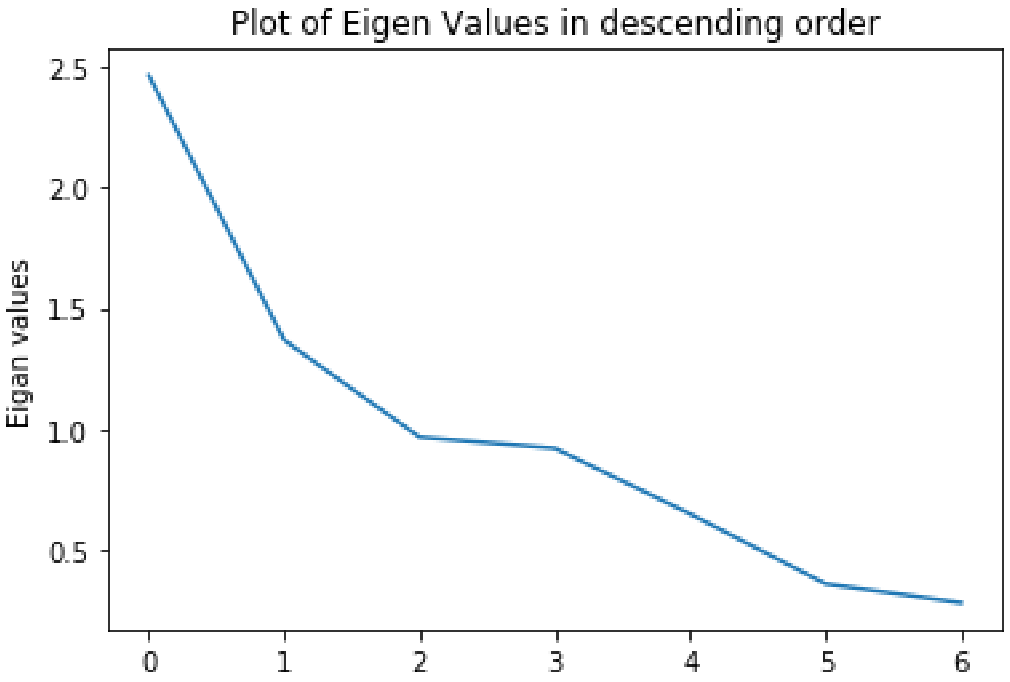
3b)



Inferences:

1)We projected the 7 dimensional data onto 2 dimensions which have the highest variance that is greatest eigen values. So, the variance or information content is mostly preserved .

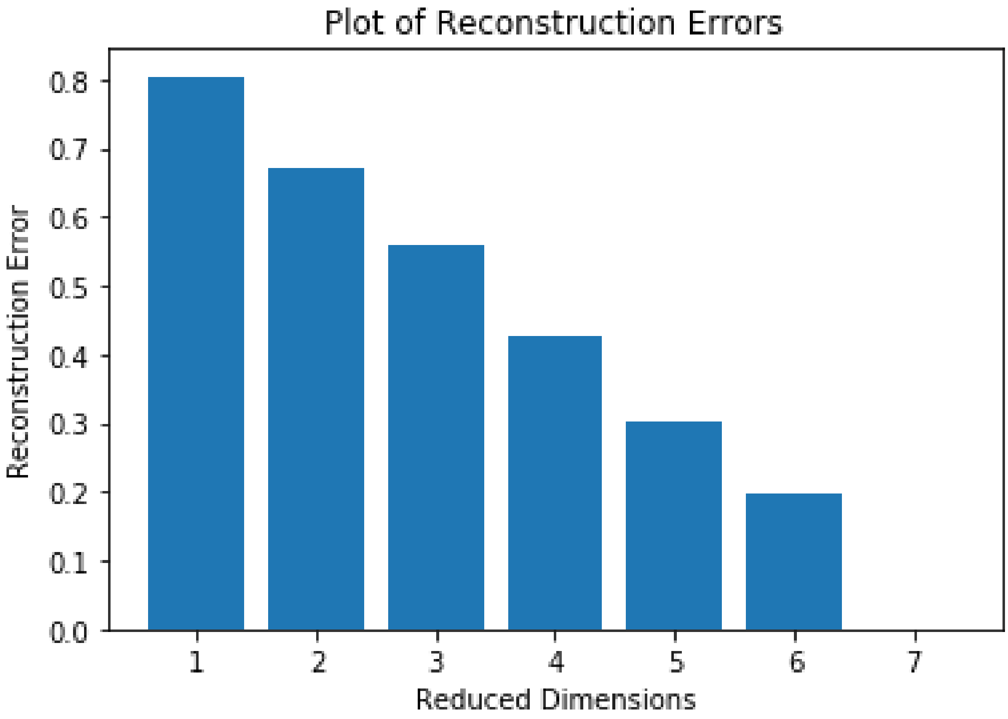
3b)



1)The subsequent eigen values do not decrease very rapidly according to the figure.

2)The rate of decrease falls rapidly from 3rd eigen vector (approximately 1).

3c)



1)The reconstruction errors are [0.804, 0.672, 0.56, 0.427, 0.301, 0.199, 0.0]

2)Hence, we observe that as the number of components increases, the root mean square error decreases. And for l=7, we don’t get any error as it is the original data itself. Hence as magnitude of reconstruction error reduces, we get a more accurate plot.

Thank You