Data Science Project
Sub-Project III
Attribute Normalization, Standardization and Dimension Reduction of Data

Student's Name: Anmol Bishnoi **Mobile No: 7042845211**

Roll Number: B19069 **Branch: CSE**

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)	992.654	1037.604	3	9
4	Rain (in ml)	0	2470.50	3	9
5	Lightavgw/o0 (in lux)	0	10565.352	3	9
6	Lightmax (in lux)	2259.0	54612.0	3	9
7	Moisture (in %)	0	100.0	3	9

Inferences:

1. Viewing the boxplots from the code, it can be observed *rain* has huge number of outliers, *pressure* and *humidity* have few outliers whereas other attributes have almost no outliers.
2. Before normalization, the values having bigger values would overpower and skew any calculations in their favor. So, the analysis will be more partial. Now after normalization, each value is not between 3 to 9, so they will have equal weightage in the analysis.

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)	21.3694	4.125	0	1.0
2	Humidity (in g.m^{-3})	83.991	17.565	0	1.0
3	Pressure (in mb)	1014.793	6.120	0	1.0
4	Rain (in ml)	171.466	398.460	0	1.0
5	Lightavgw/o0 (in lux)	2237.899	2206.422	0	1.0
6	Lightmax (in lux)	21788.623	22064.99	0	1.0
7	Moisture (in %)	21.3694	4.125	0	1.0

Inferences:

1. Before normalization, the values having bigger values would overpower and skew any calculations in their favor. So, the analysis will be more partial. Now after normalization, each value is not between 3 to 9, so they will have equal weightage in the analysis.

2 a.

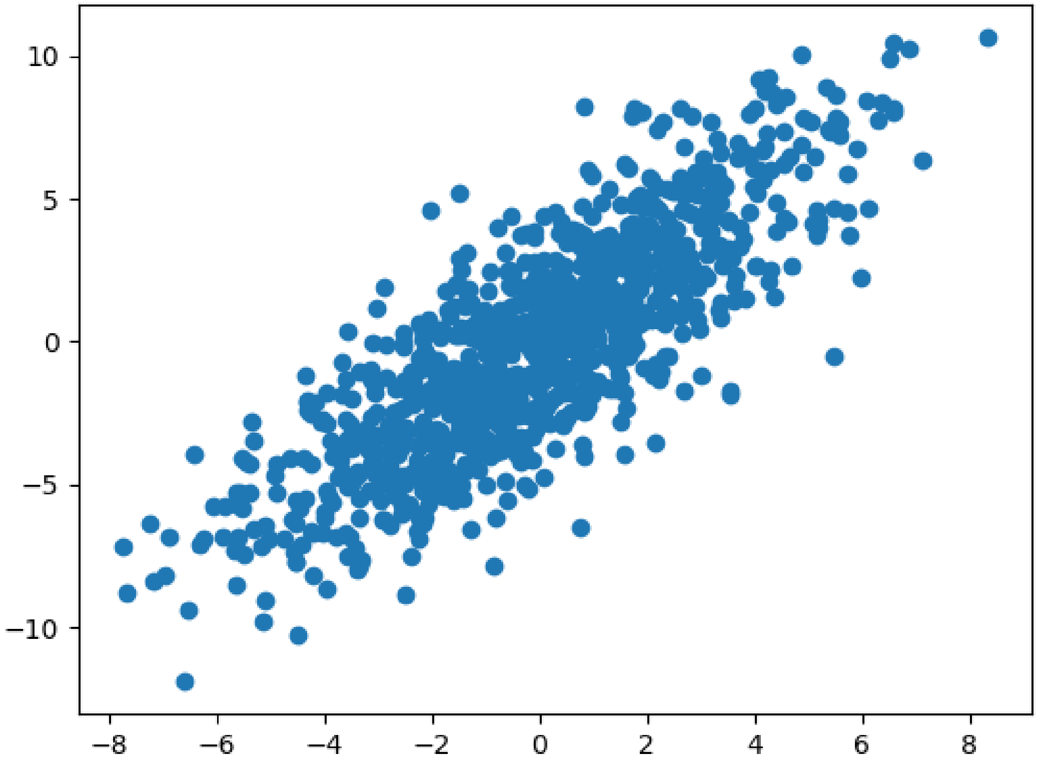


Figure 1 Scatter Plot of 2D Synthetic Data of 1000 samples

Inferences:

1. Attribute 2 is positively related to the Attribute 1 according to the graph. The covariance will be positive.

2. Seeing the density of the graph, the distribution of both the attributes seem to be symmetric. The mean of both the attribute is approximately 0.

b.

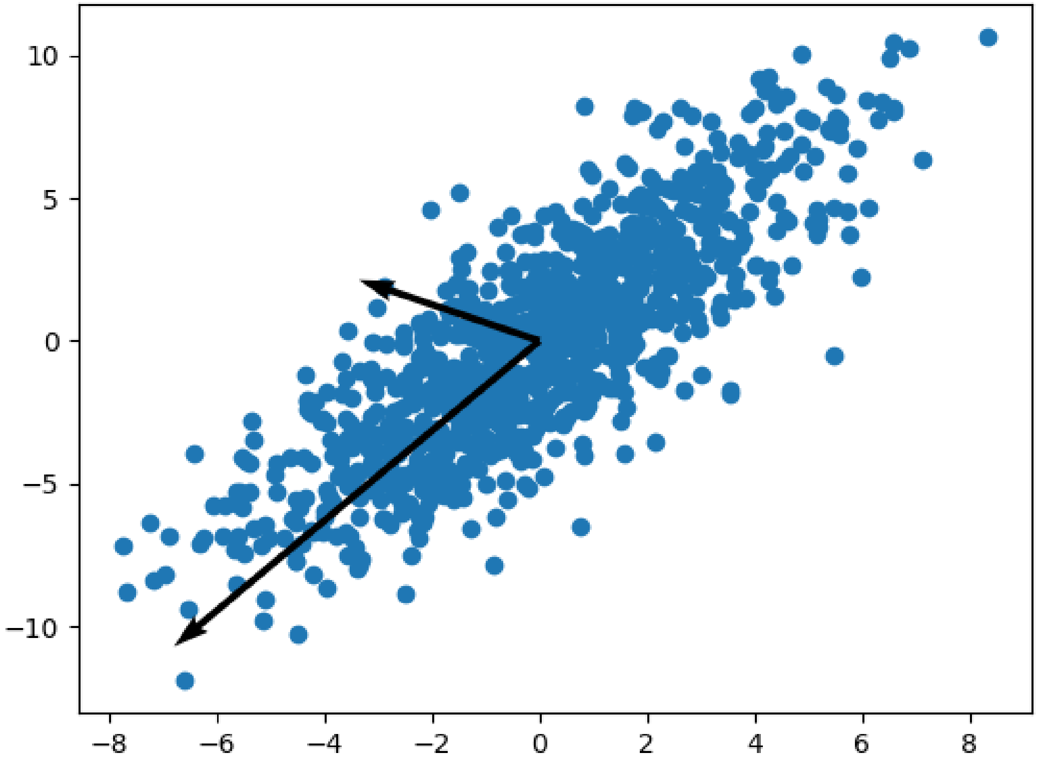


Figure 2 Plot of 2D Synthetic Data and Eigen Directions

Inferences:

1. The eigenvalues in this case are 1.667 and 19.476. The same can be clearly observed from the above figure as the spread along the direction of the first eigenvector is not much when compared to the spread along the second eigenvector.
2. The density of points near the intersection of axis is very dense, and it gradually decreases as the spread increases. In other word, the number of points decreases as move far from the center and the data properly characterizes a Normal Distribution along both the eigenvectors.

c.

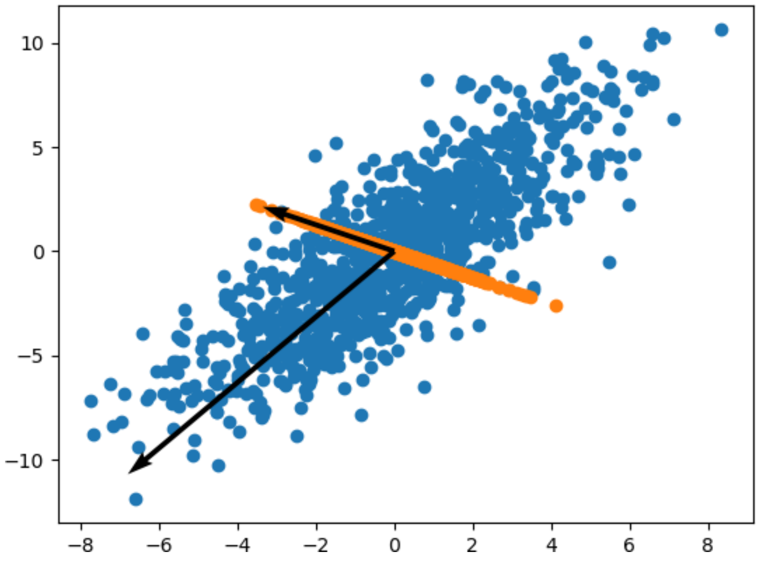


Figure 3 Projected Eigen Directions onto the Scatter Plot with 1st Eigen Direction highlighted

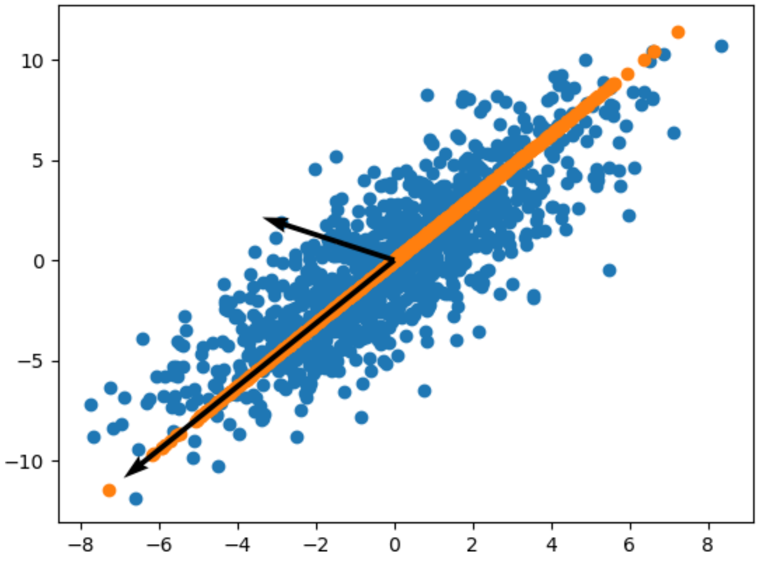


Figure 4 Projected Eigen Directions onto the Scatter Plot with 2nd Eigen Direction highlighted

Inferences:

1. The values of eigenvalues can be inferred from the variance along the directions. As one can see that the variance along the second eigenvector (long line) is more than the first eigenvector.
2. Regarding the density, along the first Eigen vector (smaller line), the variance is not very large, so the spread is not so much varying. However, along the second Eigen Vector, the variance is high, so the spread is more, so the density actually is high near the origin and spread is large.

d. Reconstruction Error = 0.0

Inferences:

1. More the reconstruction error, more loss in the nature of data. So, the reconstruction error must not be very high.
2. Here the reconstruction error was almost negligible. This is because we did not do any data reduction whatsoever. We would have seen some reduction error, had we reconstructed data from the projected data along fewer than the total number of eigenvectors.

3 a.

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

Direction	Variance	Eigen Value
1	2.2247	2.2247
2	1.43	1.43

Inferences:

1. From the above table, it is clear that the eigenvalues are equal to the variances in the covariance matrix of transformed data.

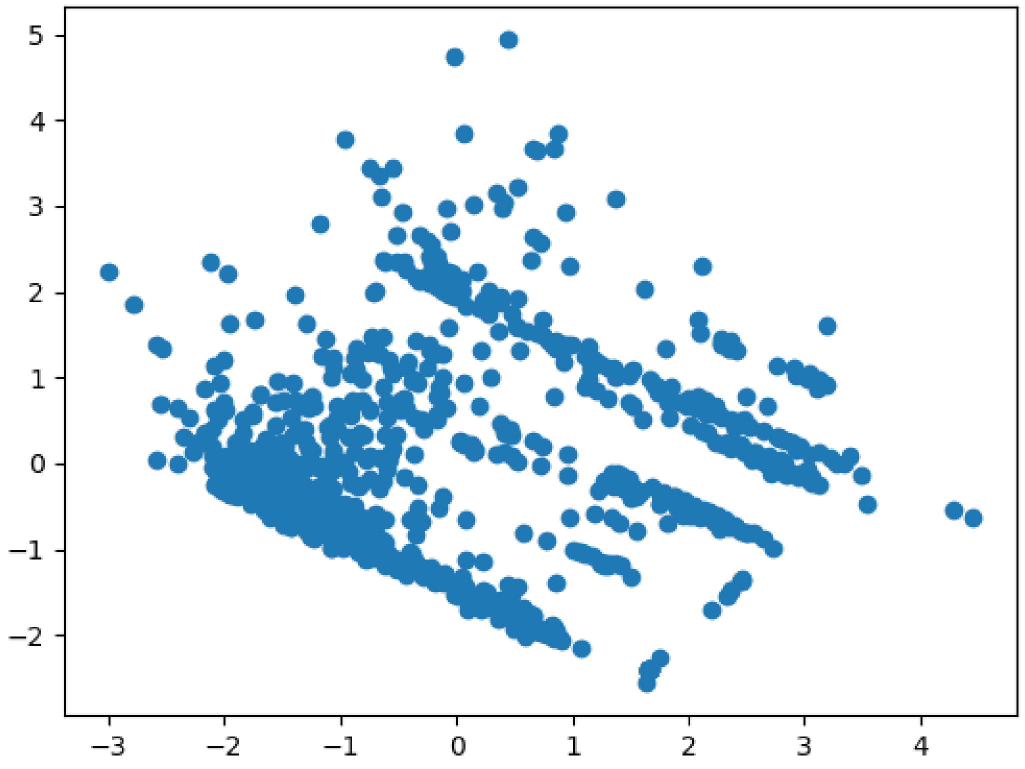
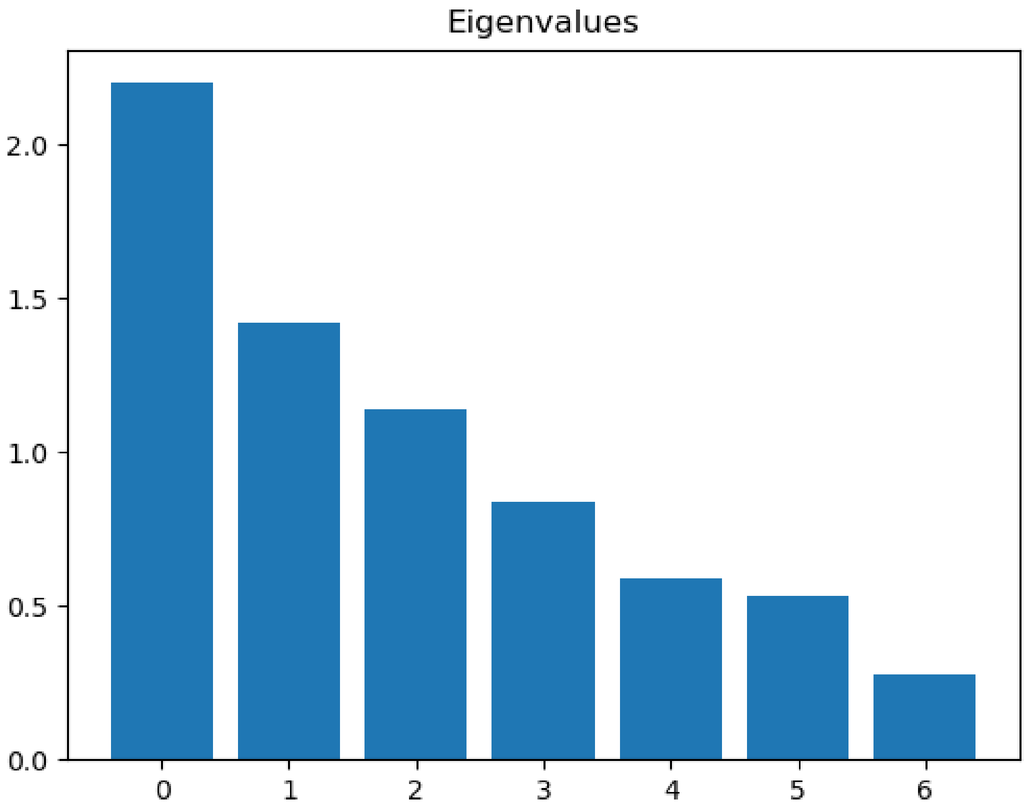


Figure 5 Plot of Landslide Data after dimensionality reduction

Inferences:

1. The spread of the data reduced and hence the density increased after dimensionality reduction because some amount of data, that would have caused further spread is now lost.



b.

Figure 6 Plot of Eigen Values in descending order

Inferences:

1. It drops significantly from first Eigen Value to second, and then it gradually decreases.
2. At the first Eigen Value, it gets the highest drop.

c.

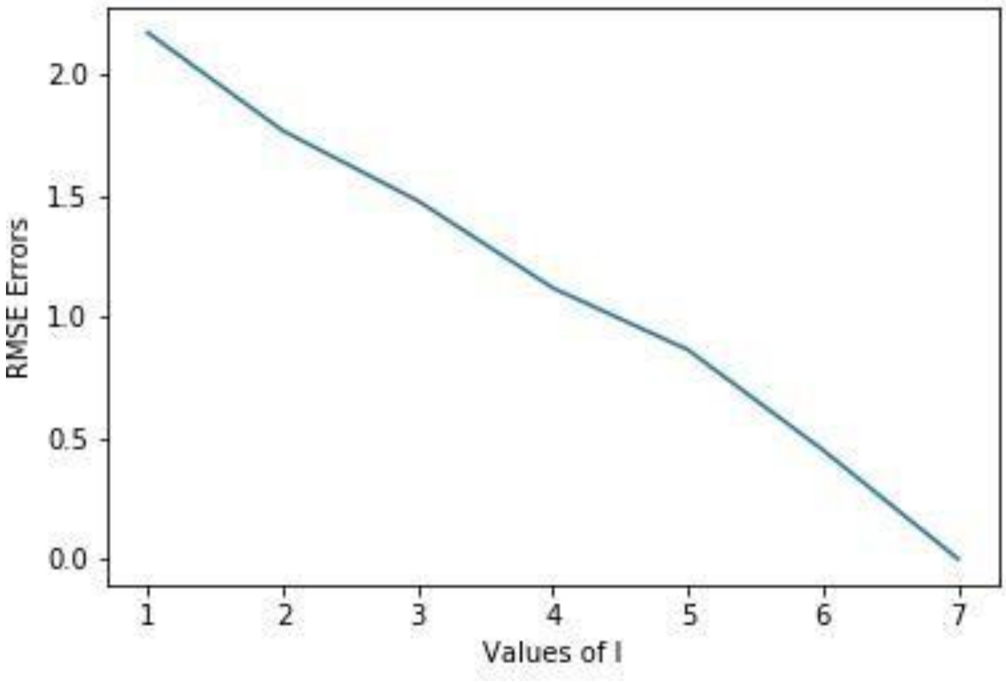


Figure 7 Line Plot to demonstrate Reconstruction Error vs. Components

Inferences:

1. More the magnitude of reconstruction error, lesser the quality of reconstructions. As we can see the RMSE increases, as we keep dropping the dimensions.
2. At l = d = 7, the reconstruction error is almost negligible.