IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – III
Attribute normalization, standardization and dimension reduction of data

Student's Name: Yash Sharma **Mobile No:** 8802131138

Roll Number: B20241 **Branch:** CSE

1 a.

Table 1 Minimum and maximum attribute values before and after normalization

S. No.	Attribute	Before normalization		After normalization	
		Minimum	Maximum	Minimum	Maximum
1	pregs	0.0	13.0	5.0	12.0
2	plas	44.0	199.0	5.0	12.0
3	pres (in mm Hg)	38.0	106.0	5.0	12.0
4	skin (in mm)	0.0	63.0	5.0	12.0
5	test (in mu U/mL)	0.0	318.0	5.0	12.0
6	BMI (in kg/m^2)	18.2	50.0	5.0	12.0
7	pedi	0.078	1.191	5.0	12.0
8	Age (in years)	21.0	66.0	5.0	12.0

Inferences:

1. The correction of outliers creates somewhat a normal distribution in some of the variables and thus makes transformations of other variables more effective.
2. In this method we replaced the outlier's values with the mean of the respective attributes. The process is quite justified as the median of a given data is quite robust (to outliers) and is least effected by outliers in the data (compared with mean), Thus, it is a better to use median while outlier correction.
3. After Min-Max normalization process, the minimum of the data is mapped to "5" and the maximum of the data is mapped to "12" while every other value gets mapped to a decimal between "5" and "12".

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	pregs	3.782552	3.270644	0	1
2	plas	121.6563	30.43829	0	1
3	pres (in mm Hg)	72.19661	11.14672	0	1
4	skin (in mm)	20.4375	15.69855	0	1
5	test (in mu U/mL)	59.56901	78.41532	0	1

6	BMI (in kg/m^2)	32.19896	6.410558	0	1
7	pedi	0.427044	0.245323	0	1
8	Age (in years)	32.76042	11.05538	0	1

Inferences:

1. In the above process of Standardization of the above data, it rescales all given attributes so that the transformed data have 0 mean and unit variance i.e., standard deviation of 1.
2. Standardization assumes that data is coming from Gaussian distribution and thus it became was necessary to replace all the outliers in previous step to make our data normally distributed (somewhat). However, it is not strictly needed to be true.

2 a.

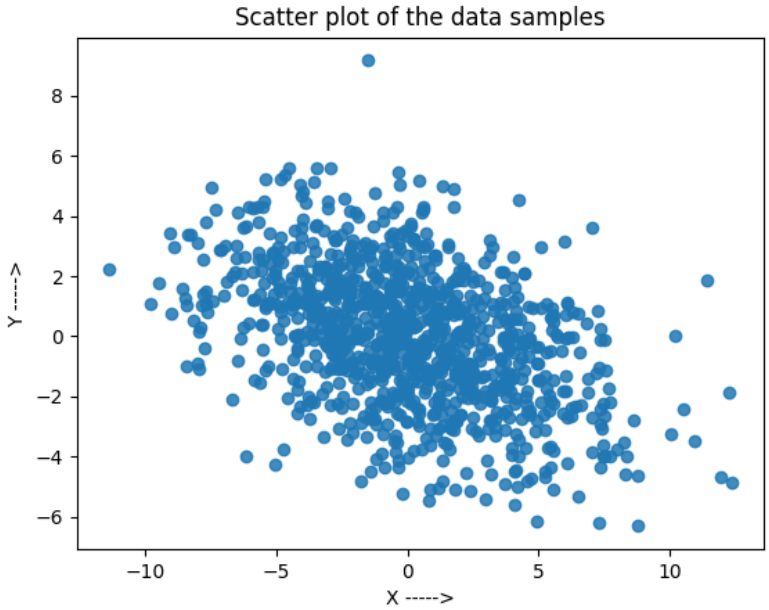

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. We may observe that as attribute 1 (X) increases proportionally attribute 2 (Y) is seen to decrease. Thus, we may conclude that they both are weakly/moderately negatively correlated.
2. A very low number of outliers are visible.
3. A cluster of points is clearly visible towards the center of given scatter plot.

b.

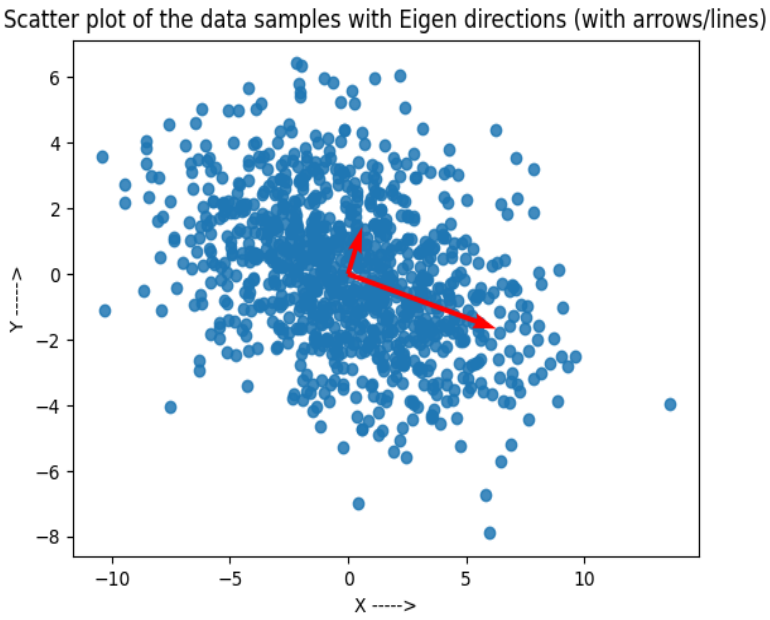

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. The higher is the eigenvalue, the higher will be the variance along a covariance matrix's eigenvector direction (principal component). That way we may infer about the spread of data by just looking at eigenvalues.
2. A high density of points is seen at the point of intersection of these two eigen axes and density gradually decreases as we go away from it.
3. Infer the spread of data based upon the magnitude of Eigenvalues.

c.

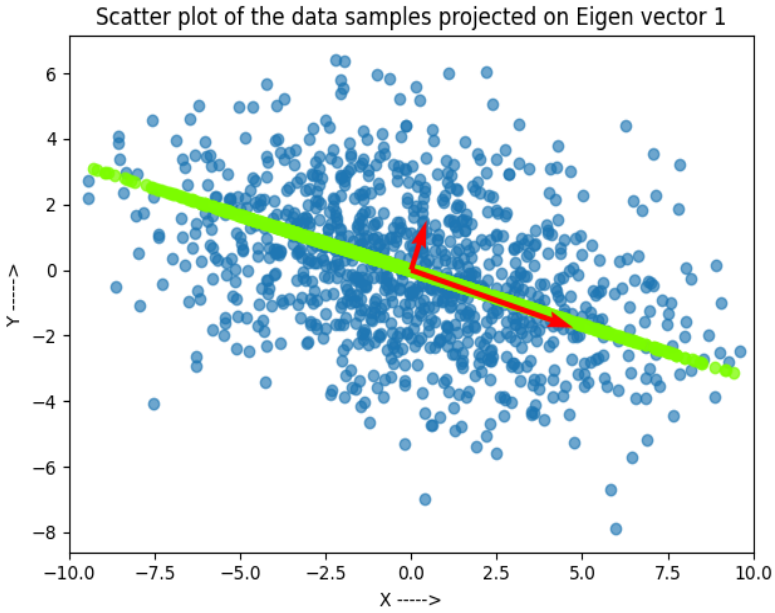

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

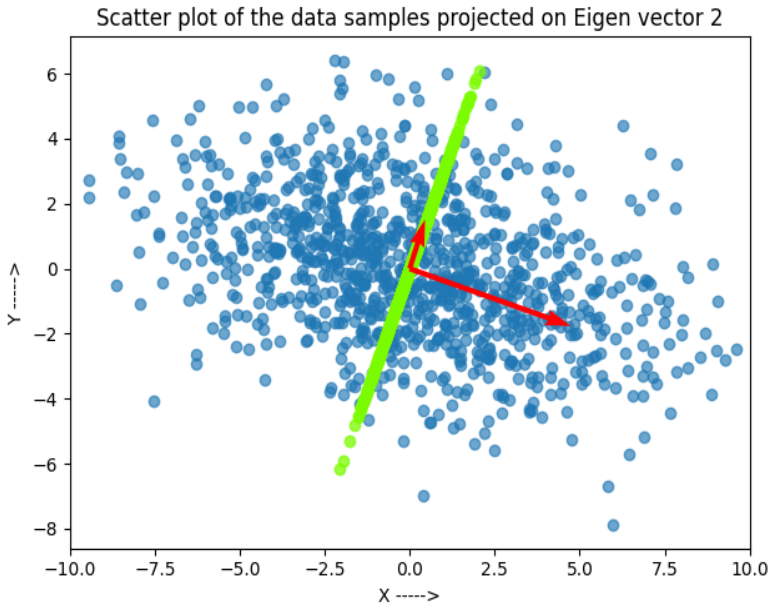

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. It can be clearly observed that Eigenvalue of vector 2 is high as compared to Eigenvalue of Eigen vector 1.

2. It may be observed that variance of given data along Eigen vector 2 is considerably high as compared to variability of same data along Eigen vector 1. An eigenvalue is the total amount of variance in the variables in the dataset explained by the common factor. So, as the eigen value of vector 2 is high thus more variance in the given data is seen along it's direction.

d. Reconstruction error = 0

Inferences:

1. In simple terms Reconstruction error is the distance between the original data point and its projection onto a lower-dimensional subspace (its 'estimate'). So, if Reconstruction error is low this infers that our reconstructed data is much closer to original data or in other terms our quality of reconstruction is quite good. On the other hand, if it is high generally infers that quality of reconstruction is not that good.
2. Here we may infer that as we used all the available eigenvectors to reconstruct the data, hence our reconstructed data is much like original data and that's why our reconstruction error is very small or technically zero (if rounded up to 3 decimals).

3 a.

Table 3 Variance and Eigenvalues of the projected data along the two directions

Direction	Variance	Eigenvalue
1	1.987	1.987
2	1.838	1.838

Inferences:

1. The Eigenvalues along the two directions exactly catches the variance of data along the directions. They both have the same value. Since we define eigenvectors as unit vectors then it falls out naturally that they are the variance of that vector in the data. If we calculate the scores by projecting the eigenvectors onto the data then the scores, since they have been multiplied by unit vectors, take on the total variance that is captured within the data by each unit vector.

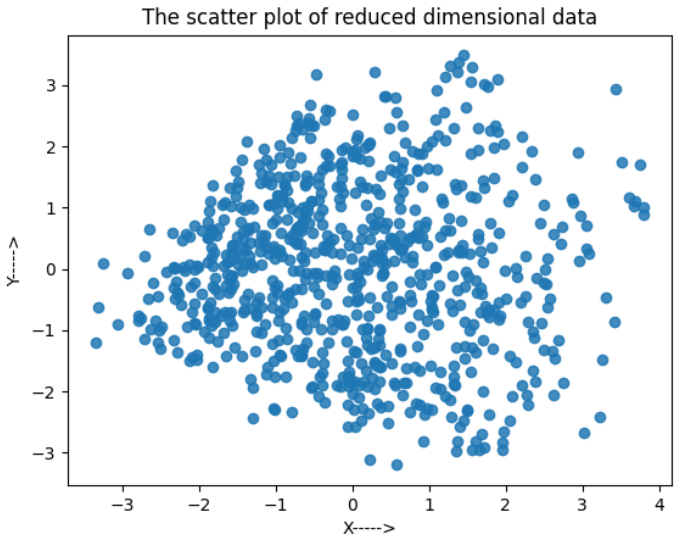

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. From the given scatter plot, the data points are scatter everywhere and thus no clear relation between the two reduced dimension is visible. Thus, we may conclude that there is either very low or no correlation between the two attributes.
2. The two attributes seem to be uncorrelated.

b.

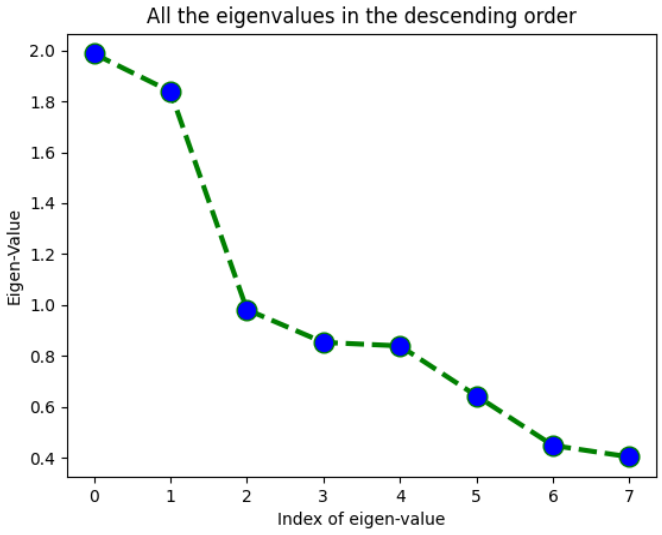

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. The subsequent Eigenvalues decrease gradually.
2. From Eigenvalue 2 the rate of decrease for given two eigenvalues changes substantially.

c.

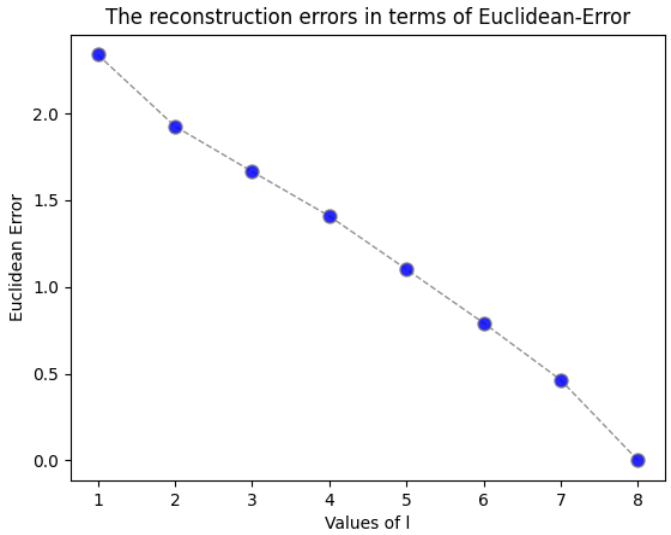

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. In simple terms Reconstruction error is the distance between the original data point and its projection onto a lower-dimensional subspace (its 'estimate'). So, if Reconstruction error is low this infers that our reconstructed data is much closer to original data or in other terms our quality of reconstruction is quite good. On the other hand, if it is high generally infers that quality of reconstruction is not that good.

Table 4 Covariance matrix for dimensionally reduced data (l=2)

	x1	x2
x1	1.987	0
x2	0	1.838

Table 5 Covariance matrix for dimensionally reduced data (l=3)

	x1	x2	x3
x1	1.987	0	0
x2	0	1.838	0

x3	0	0	0.982

Table 6 Covariance matrix for dimensionally reduced data (l=4)

	x1	x2	x3	x4
x1	1.987	0	0	0
x2	0	1.838	0	0
x3	0	0	0.982	0
x4	0	0	0	0.854

Table 7 Covariance matrix for dimensionally reduced data (l=5)

	x1	x2	x3	x4	x5
x1	1.987	0	0	0	0
x2	0	1.838	0	0	0
x3	0	0	0.982	0	0
x4	0	0	0	0.854	0
x5	0	0	0	0	0.840

Table 8 Covariance matrix for dimensionally reduced data (l=6)

	x1	x2	x3	x4	x5	x6
x1	1.987	0	0	0	0	0
x2	0	1.838	0	0	0	0
x3	0	0	0.982	0	0	0
x4	0	0	0	0.854	0	0
x5	0	0	0	0	0.840	0
x6	0	0	0	0	0	0.644

Table 9 Covariance matrix for dimensionally reduced data (l=7)

	x1	x2	x3	x4	x5	x6	x7
x1	1.987	0	0	0	0	0	0
x2	0	1.838	0	0	0	0	0
x3	0	0	0.982	0	0	0	0
x4	0	0	0	0.854	0	0	0
x5	0	0	0	0	0.840	0	0
x6	0	0	0	0	0	0.644	0
x7	0	0	0	0	0	0	0.449

Table 10 Covariance matrix for dimensionally reduced data (l=8)

	x1	x2	x3	x4	x5	x6	x7	x8

x1	1.987	0	0	0	0	0	0	0
x2	0	1.838	0	0	0	0	0	0
x3	0	0	0.982	0	0	0	0	0
x4	0	0	0	0.854	0	0	0	0
x5	0	0	0	0	0.840	0	0	0
x6	0	0	0	0	0	0.644	0	0
x7	0	0	0	0	0	0	0.449	0
x8	0	0	0	0	0	0	0	0.405

Inferences:

1. All the off-diagonal elements are zero in magnitude since Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. That's, why the resulting vectors are an uncorrelated orthogonal basis set.
2. We observe that all diagonal elements are non-zero while all diagonal elements are non-zero. Since, the goal of principal components analysis (PCA) is to explain the maximum amount of variance with the fewest number of principal components and the diagonal elements represent the same.
3. We see the diagonal values are following a descending trend in their values. The most initial diagonal element has highest value while the last one has least in any given covariance matrix.
4. The observed order of diagonal-elements has a very important explanation w.r.t to analysis of the data. This is observed since PCA technique projects data on eigenvectors which our sorted based upon their eigen values and since eigen value magnitude directly represents the importance principal component (that explains maximum amount of variance) thus as we go further in this order the eigen-value magnitude decreases consequently we see decrease in value of variance (diagonal elements).
5. It can clearly be observed that as the initial element of all covariance matrix has maximum value consequently have highest eigenvalue thus represents or captures the variance/ data variations the best (as it contains the most important principal component).
6. We can clearly observe from the given matrices that initial 2 diagonal elements have very good magnitude of variance as compared to the rest of the elements thus these 2 may be assumed to contains most of the variation of the given data in the best manner. Thus, 2 components i.e., l=2 shall give optimum reconstruction along with dimensionality reduction.
7. The magnitude of the 1st diagonal element (topmost left corner) in each of the obtained covariance matrices is similar across all the matrices. This happens because the highest eigenvalues across all 8 others is used for obtaining column 1. So, every time PCA is performed over this dataset then for obtaining first reduced column every time the data is projected over the same eigenvector with highest eigenvalue to capture most of the data variations (variance).
8. The magnitude of the 2nd diagonal element in each of the obtained covariance matrices is similar across all the matrices. This happens because the second highest eigenvalues across all 8 others is used for

obtaining column 2. So, every time PCA is performed over this dataset then for obtaining second reduced column every time the data is projected over the same eigenvector corresponding to second highest eigenvalue to capture most of the data variations (variance).

9. The magnitude of 3rd, 4th, 5th, 6th, and 7th diagonal elements across covariance matrices is same across all the matrices. This is observed because every time PCA is performed, it uses the same set of eigenvectors corresponding to the eigenvalues sorted in descending order. Thus, as the order is the property of data set given thus it remains unchanged and same set of vectors are chosen every time to dimensional reduction and obtain reduced columns in their order they appear in sorted list.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1	0.118	0.209	-0.097	-0.110	0.030	0.005	0.560
plas	0.118	1	0.204	0.060	0.157	0.228	0.0804	0.274
pres (in mm Hg)	0.209	0.204	1	0.026	-0.049	0.271	0.022	0.326
skin (in mm)	-0.096	0.060	0.025	1	0.455	0.374	0.1515	-0.101
test (in mu U/mL)	-0.108	0.157	-0.049	0.455	1	0.164	0.193	-0.076
BMI (in kg/m^2)	0.028	0.228	0.271	0.374	0.164	1	0.123	0.078
pedi	0.004	0.080	0.22	0.151	0.193	0.123	1	0.036
Age (in years)	0.561	0.274	0.326	-0.101	-0.076	0.078	0.036	1

Inferences:

1. It can be easily observed that the off-diagonal values are non-zero which clearly states that the all the 8 attributes are correlated with each other or show some dependency on each other. While when we compare it with the covariance matrix obtained after PCA l=8 reduction we see all off-diagonal elements were zero which was also expected as PCA result in columns or attributes that are uncorrelated to allow diagonal elements (columns) to represent the data variation in best way possible.
2. We can observe that some initial diagonal elements in PCA reduced (l=8) data frame captures much more variation in the data than the original data due to their very high values (nearly twice). Thus, we may conclude that the diagonal elements of reduced data contain more good picture of variance of original data as compared to original data itself.
3. We may observe that there is no specific order in the covariance matrix of original data frame moreover they all are unity which is very different what we see in the reduced data frame where there is a definite order (Descending) of diagonal elements as they are sorted in order of eigenvalues and order in which how much variance an attribute may capture.