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

Student's Name: Mayank Bansal **Mobile No: +919636993445**

Roll Number: B20156 **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	17	5	12
2	plas	0	199	5	12
3	pres (in mm Hg)	0	122	5	12
4	skin (in mm)	0	99	5	12
5	test (in mu U/mL)	0	846	5	12
6	BMI (in kg/m^2)	0	67.1	5	12
7	pedi	0.078	2.42	5	12
8	Age (in years)	21	81	5	12

Inferences:

1. Outliers increase the variability in our data, which decreases statistical power. So, we need to correct them.
2. We replace the extreme values or outliers with median values so as to maintain the pattern within the main dataset and also it is unaffected by the outliers.
3. Before normalization, the values having bigger values use to overpower the ones with smaller values. So, the analysis will be more partial. After normalization, each value is between 5 to 12, so normalization will make sure that all of our data looks and reads the same way across all records.

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	3.85	3.37	0	1
2	plas	120.89	31.97	0	1
3	pres (in mm Hg)	69.11	19.36	0	1
4	skin (in mm)	20.54	15.95	0	1
5	test (in mu U/mL)	79.8	115.24	0	1

6	BMI (in kg/m^2)	31.99	7.88	0	1
7	pedi	0.47	0.33	0	1
8	Age (in years)	33.24	11.76	0	1

Inferences:

1. After standardization, every value has a common mean of 0 with variance 1.0
2. Standardization is better than normalization because there will be no out of bound error in it

2 a.

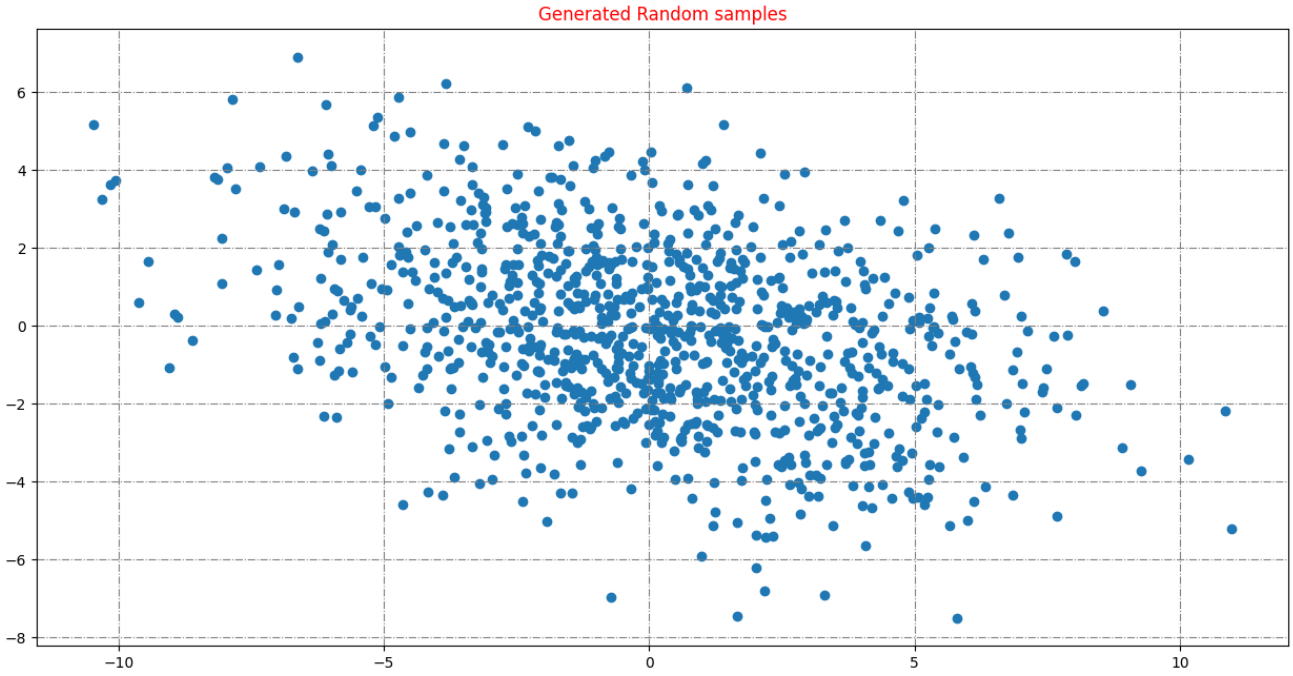

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. Both the attributes are negatively correlated to each other.
2. Both the attributes are symmetrically distributed about their mean and it is denser around and near to mean.

b.

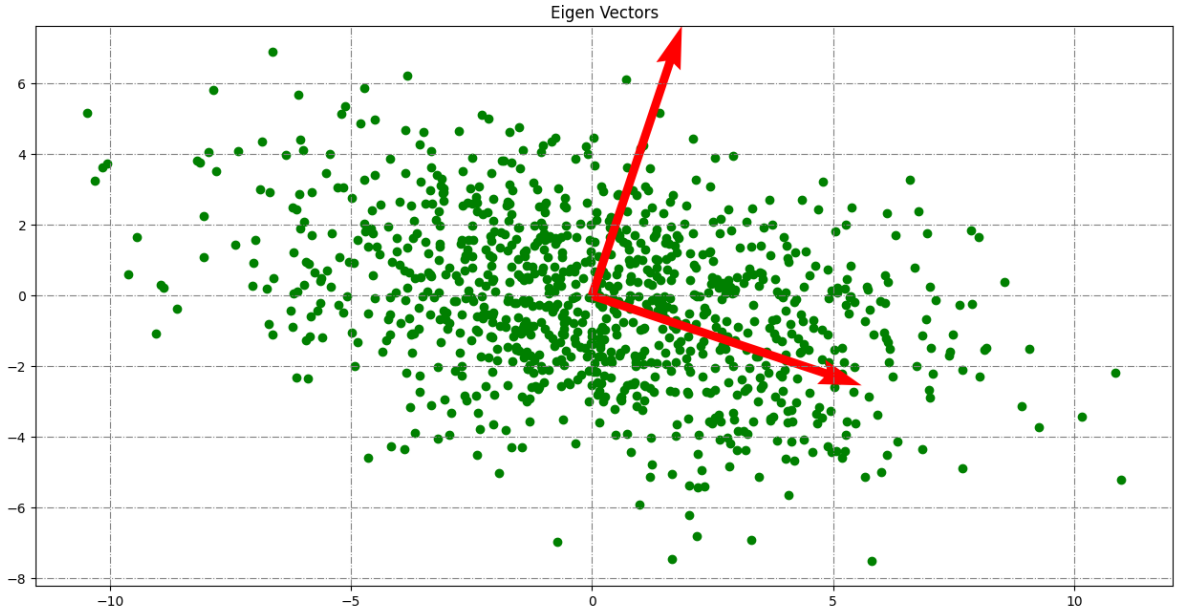

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. Eigenvalue 1 has more spread in case of projections. Eigenvalue 1 being more inclined to x axis implies that data has more spread on x axis.
2. Data is more concentrated around mean of the data and density gradually fades away. Both the Eigenvectors meet at the mean of the data i.e. [0,0].

c.

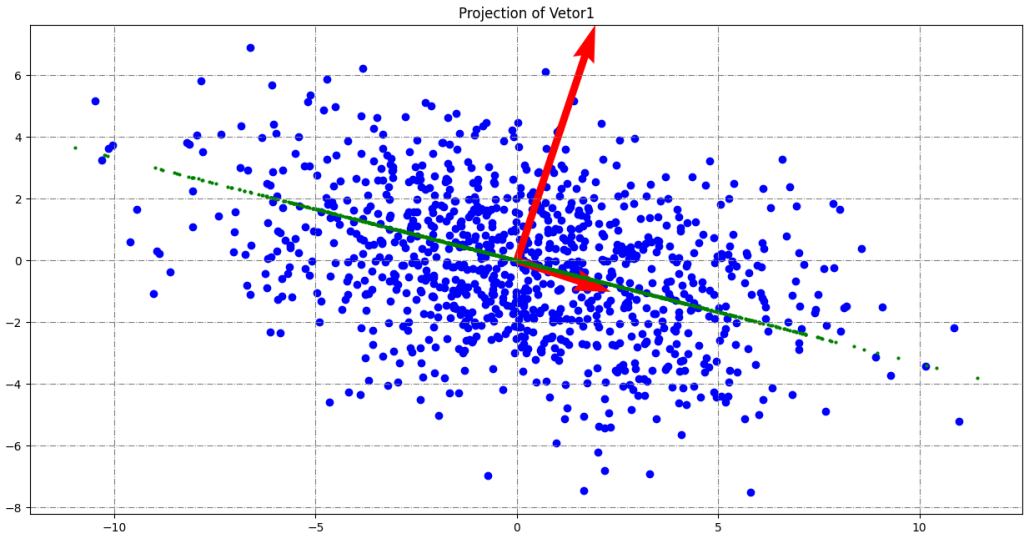

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

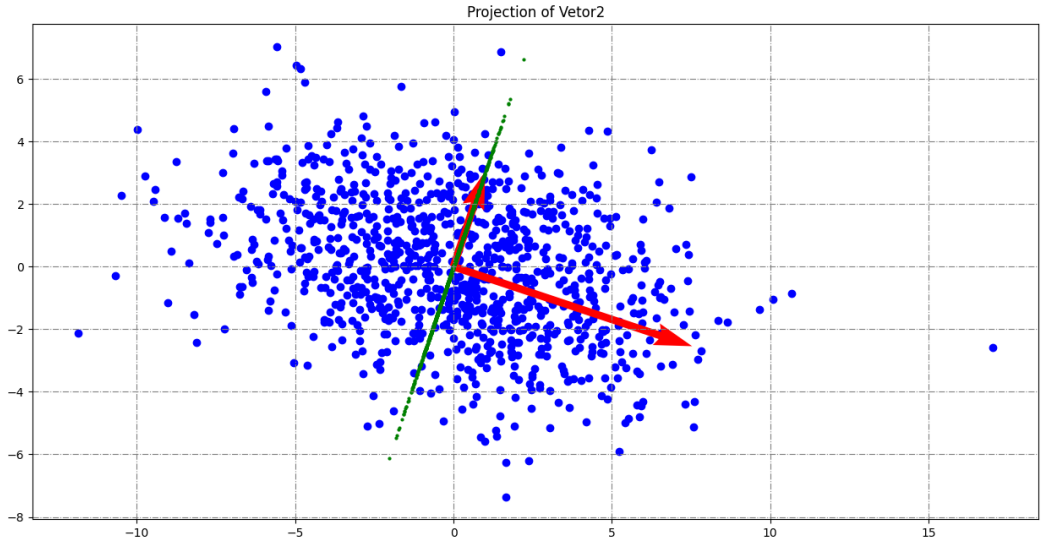

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. Eigenvalue 1 is more than that of Eigenvalue 2 and both eigenvalues indicate the spread of data in the above pattern
2. Eigenvalue 1 being more than that of eigenvalue 2 indicates that spread of projection of data will me more in case of eigenvector 1. Hence the distribution follows the above pattern.

d. Reconstruction error = 0

Inferences:

1. In this case reconstruction error of all the tuples is 0 because we try to project a 2d data on a 2d dimensional axis. In this case we lose no data and we obtain the data successfully back. Hence RE=0

3 a.

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

Direction	Variance	Eigenvalue
1	0.8592306782080319	1.8159863276720143
2	0.6509093278792369	1.3554399938718875

Inferences:

1. Here it is clear that variance and eigenvalue are not much deviated in the second case as much as they are in the first case.

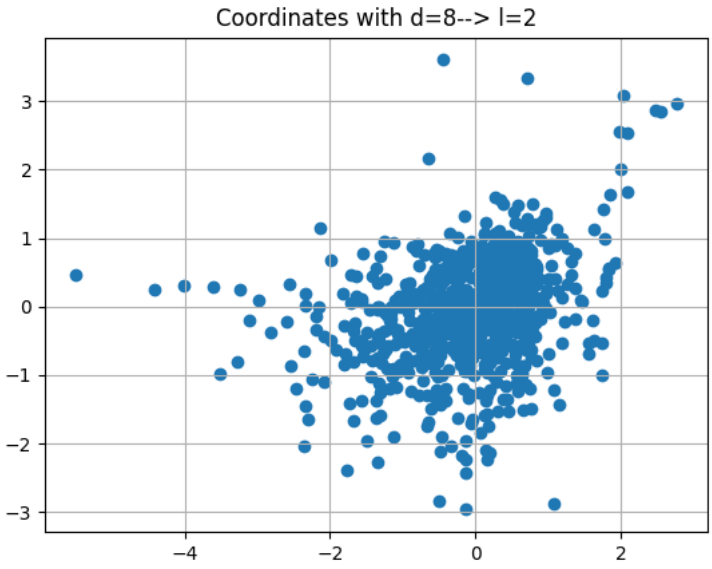

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. Variance is not very high and that is very logical as well as the reduced data that we obtain must be very less correlated as we have converted a big dimensionality data to a very low dimensionality data.

b.

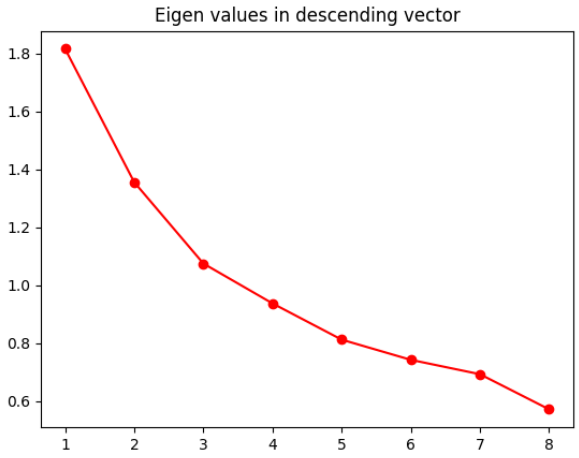

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. Plotted Eigen Values are = [1.81598633 1.35543999 1.07455153 0.93685112 0.81171521 0.74200394 0.69235368 0.5710982]
2. The value of Eigen value decreases very fastly from 1st and 2nd value

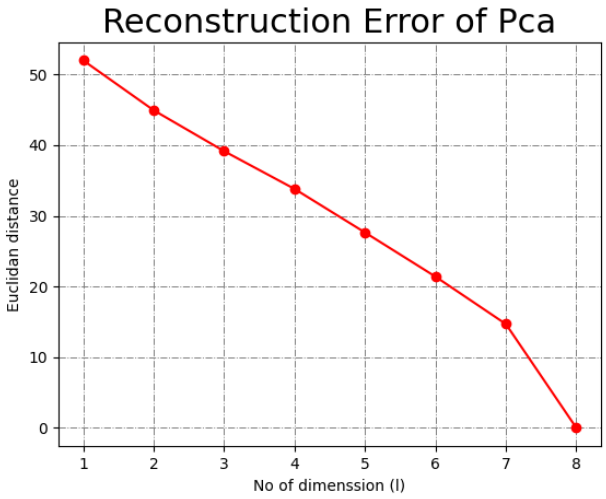

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1 Reconstruction shows the how much your data is more near to ideal one i.e. near to lossless data.

2 Hence it's less values shows power of PCA.

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

	x1	x2
x1	1022.476	0
x2	0	684.15

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

	x1	x2	x3
x1	1022.476	0	0
x2	0	684.15	0
x3	0	0	486.73

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

	x1	x2	x3	x4
x1	1022.476	0	0	0

x2	0	684.15	0	0
x3	0	0	486.73	0
x4	0	0	0	390.12

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

	x1	x2	x3	x4	x5
x1	1022.476	0	0	0	0
x2	0	684.15	0	0	0
x3	0	0	486.73	0	0
x4	0	0	0	390.12	0
x5	0	0	0	0	378.978

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

	x1	x2	x3	x4	x5	x6
x1	1022.476	0	0	0	0	0
x2	0	684.15	0	0	0	0
x3	0	0	486.73	0	0	0
x4	0	0	0	390.12	0	0
x5	0	0	0	0	378.978	0
x6	0	0	0	0	0	306.349

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

	x1	x2	x3	x4	x5	x6	x7
x1	1022.476	0	0	0	0	0	0
x2	0	684.15	0	0	0	0	0
x3	0	0	486.73	0	0	0	0
x4	0	0	0	390.12	0	0	0
x5	0	0	0	0	378.978	0	0
x6	0	0	0	0	0	306.349	0
x7	0	0	0	0	0	0	242.839

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

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1022.476	0	0	0	0	0	0	0
x2	0	684.15	0	0	0	0	0	0
x3	0	0	486.73	0	0	0	0	0
x4	0	0	0	390.12	0	0	0	0
x5	0	0	0	0	378.978	0	0	0
x6	0	0	0	0	0	306.349	0	0
x7	0	0	0	0	0	0	242.462	0
x8	0	0	0	0	0	0	0	215.839

Inferences:

1. We can see that the values at diagonal decreasing because of the eigen values to the corresponding down is decreasing and it shows variance of the data.
2. It shows that no component is related to another one, this is because of the projected data on orthogonal vectors.
3. The trend shows a regular decrease.
4. Since the eigenvalues is becoming small for lower component, and eigenvalue shows the variance of the data.
5. The first component captures data variations the best.
6. Because there are total 8 attributes we can assume n/2 components ie 4 components optimum for data reduction and reconstruction.
7. It is same as the component is along a vector which does not depend on any other component, hence when they come in picture the values of the previous data does not change.
8. Since are all the orthogonal vectors hence their no component is going to disturb other one in terms of stats or data, hence when new component is adding previous one doesn't affect.
9. They are all same.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1	0.03	0.064	-0.044	-0.015	-0.00969	-0.0164	0.245
plas	0.03	1	0.102	0.0288	0.295	0.154	0.116	0.172
pres (in mm Hg)	0.0645	0.102	1	0.187	0.070	0.200	0.0274	0.0718
skin (in mm)	-0.044	0.0288	0.187	1	0.216	0.338	0.1938	-0.102
test (in mu U/mL)	-0.015	0.295	0.07	0.216	1	0.153	0.122	0.053
BMI (in kg/m^2)	-0.0096	0.154	0.2007	0.338	0.1536	1	0.1406	-0.0116
pedi	-0.016	0.116	0.027	0.1938	0.1222	0.1406	1	0.0256
Age (in years)	0.245	0.1726	0.0718	-0.102	0.05	-0.0116	0.0256	1

Inferences:

1 In the original data we get that the attributes has their correlated in between them but in the dimensionality reduction data there is no relation between the components.

1 For the first two component the variance is greater than the real one which shows that it captures more spread of the data while other component has less than variance than the real data.

2 Initially it was more than real but then it decrease to real one.

3 More spread is in first two components while there is less spread capturing in the last components.