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

Student's Name: Shubham Shukla **Mobile No: 8317012277**

Roll Number: B20168 **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	13	5	12
2	plas	44	199	5	12
3	pres (in mm Hg)	38	106	5	12
4	skin (in mm)	0	63	5	12
5	test (in mu U/mL)	0	318	5	12
6	BMI (in kg/m^2)	18.2	50	5	12
7	pedi	0.078	1.191	5	12
8	Age (in years)	21	66	5	12

Inferences:

1. Outliers create the invariability in data, they are unusual values in our dataset, hence they need to be corrected.
2. We use median since they are unaffected statistic in presence of the outliers.
3. The range of different attribute are different initially but after after normalization they all came in same range, which increases our statistical analyses.

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	7.037	1.761	0	1
2	plas	8.507	1.375	0	1
3	pres (in mm Hg)	8.520	1.147	0	1
4	skin (in mm)	7.270	1.744	0	1
5	test (in mu U/mL)	6.341	1.709	0	1
6	BMI (in kg/m^2)	8.081	1.411	0	1
7	pedi	7.120	1.542	0	1

8	Age (in years)	6.830	1.720	0	1

Inferences:

1. Before, the statistical data values has difference in their ranges, hence one having high will empower the lower one but after they all have same statistic, so variation decreases.
2. Standardization is better than normalization, as in after case we can get out of bound error.

2 a.

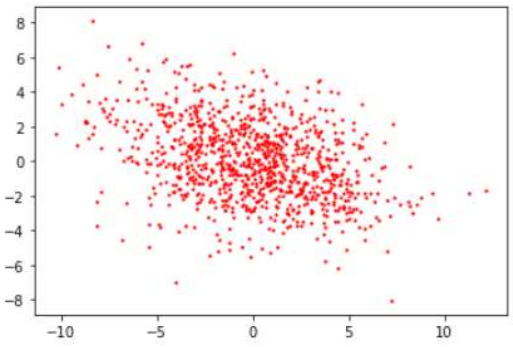

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. Both are slightly negatively correlated.
2. Both are spread about 0, 0. Hence most probably the mean should be around zero.

b.

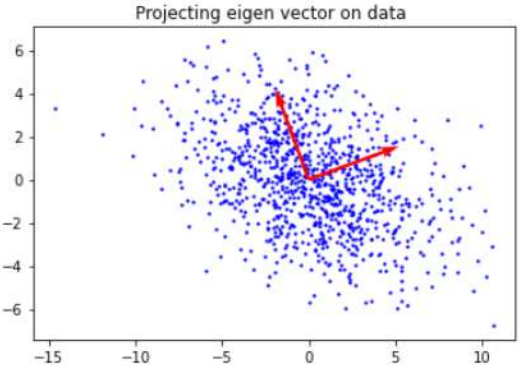

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. The data is more spread on first eigenvector as it is having greater eigenvalue.
2. At the intersection the data has more density showing that mostly data is spread about the intersection point of the vectors.

c.

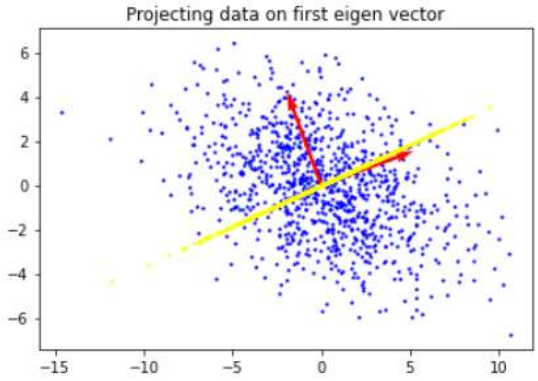

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

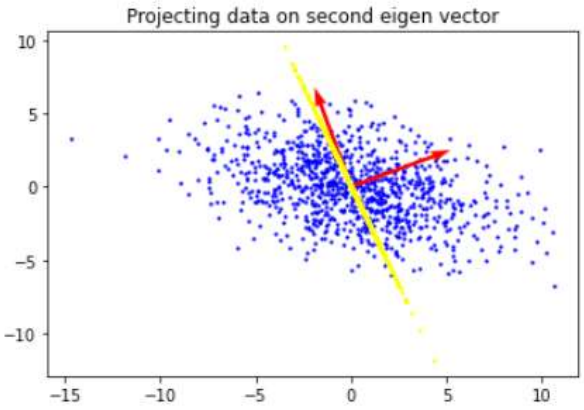

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. Seeing the distribution is seems that the difference between the eigenvalue are not much as distribution is much same, but yes whatever the difference is it is to first eigenvalue.
2. Eigenvalue 1 being more than that of eigenvalue 2 indicates that spread of projection of data will more in case of eigenvector 1. Hence the distribution follows the above pattern.

d. Reconstruction error = `23.278`

Inferences:

1. Since the reconstruction directly refers to the data loss, hence higher the error higher the loss and vice-versa.
2. Seeing the error it seems that the PCA is lossy method.

3 a.

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

Direction	Variance	Eigenvalue
1	1.992	1.992
2	1.853	1.853

Inferences:

1. The variance and eigen values along the corresponding direction are around exact same.
2. It shows that the all the variables has same standard deviation thus all the variables have same weight.

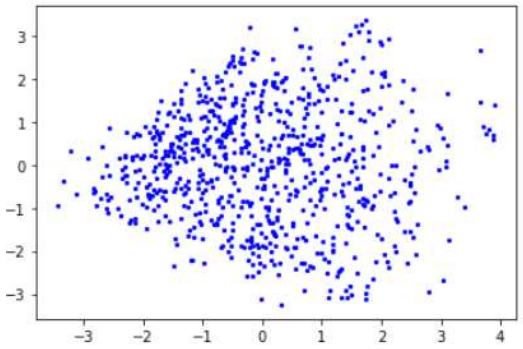

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. We are not getting any Inference about correlation on seeing the spread.
2. As the first attribute increases then we can see that spread of second also increases.

b.

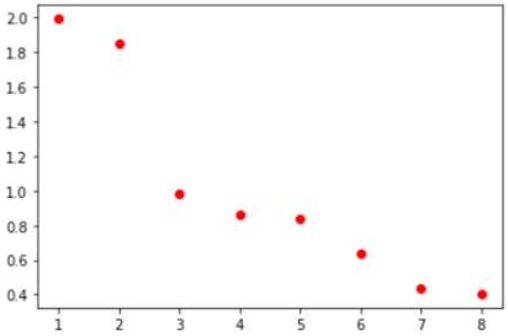

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. The eigen values decreases linearly
2. After the second eigen values, the left eigen values decreases very highly.

c.

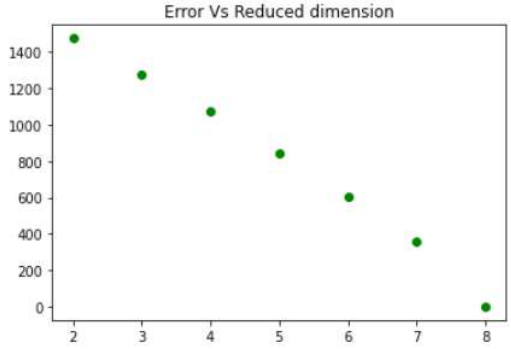

Figure 7 Line plot to demonstrate reconstruction error vs. components

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	1.992	0
x2	0	1.853

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

	x1	x2	x3
x1	1.992	0	0
x2	0	1.853	0
x3	0	0	9.819

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

	x1	x2	x3	x4
x1	1.992	0	0	0
x2	0	1.853	0	0
x3	0	0	9.819	0
x4	0	0	0	0.858

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

	x1	x2	x3	x4	x5
x1	1.992	0	0	0	0
x2	0	1.853	0	0	0
x3	0	0	9.819	0	0
x4	0	0	0	0.858	0
x5	0	0	0	0	0.839

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

	x1	x2	x3	x4	x5	x6
x1	1.992	0	0	0	0	0

x2	0	1.853	0	0	0	0
x3	0	0	9.819	0	0	0
x4	0	0	0	0.858	0	0
x5	0	0	0	0	0.839	0
x6	0	0	0	0	0	0.636

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

	x1	x2	x3	x4	x5	x6	x7
x1	1.992	0	0	0	0	0	0
x2	0	1.853	0	0	0	0	0
x3	0	0	9.819	0	0	0	0
x4	0	0	0	0.858	0	0	0
x5	0	0	0	0	0.839	0	0
x6	0	0	0	0	0	0.636	0
x7	0	0	0	0	0	0	0.434

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

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1.992	0	0	0	0	0	0	0
x2	0	1.853	0	0	0	0	0	0
x3	0	0	0.982	0	0	0	0	0
x4	0	0	0	0.858	0	0	0	0
x5	0	0	0	0	0.839	0	0	0
x6	0	0	0	0	0	0.636	0	0

x 7	0	0	0	0	0	0	0.434	0
x 8	0	0	0	0	0	0	0	0.404

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. It is decreasing.
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. For the lower variance, we will get less error hence we can say that lower component captures more reconstruction of the data.
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.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1.000	0.118	0.208	-0.097	-0.108	0.028	0.005	0.561
plas	0.118	1.000	0.205	0.060	0.180	0.228	0.081	0.274
pres (in mm Hg)	0.208	0.205	1.000	0.025	-0.050	0.271	0.022	0.326
skin (in mm)	-0.097	0.060	0.025	1.000	0.473	0.373	0.153	-0.101
test (in mu U/mL)	-0.108	0.180	-0.050	0.473	1.000	0.171	0.198	0.073
BMI (in kg/m^2)	0.028	0.228	0.271	0.373	0.171	1.000	0.123	0.777
pedi	0.005	0.081	0.022	0.153	0.198	0.123	1.000	0.036
Age (in years)	0.561	0.274	0.326	-0.101	0.073	0.777	0.036	1.000

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.
2. 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.
3. Initially it was more than real but then it decrease to real one.
4. More spread is in first two components while there is less spread capturing in the last components.