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

Student's Name: Ankit Pal Singh **Mobile No: 9149024234**

Roll Number: B20181 **Branch:EE**

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	0.000	13.000	5	12
2	plas	44.000	199.000	5	12
3	pres (in mm Hg)	38.000	106.000	5	12
4	skin (in mm)	0.000	63.000	5	12
5	test (in mu U/mL)	0.000	318.000	5	12
6	BMI (in kg/m^2)	18.200	50.000	5	12
7	pedi	0.078	1.191	5	12
8	Age (in years)	21.000	66.000	5	12

Inferences:

1. Outliers are those data which are far more from the data distribution , so they are of no use . that is why we need outlier correction like in BMI (it should never be 0).
2. The outliers are found using IQR and methods related to that , and these outliers are replaced by median of the attributes.
3. Before normalization , the data is not in a particular range like for some attributes the range is very high and for some it's not that much so in order to avoid these problems normalization is done by creating new values that maintain the general distribution and ratios in the source data.

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	121.656250	30.438286	0	1
3	pres (in mm Hg)	72.196615	11.146723	0	1
4	skin (in mm)	20.437500	15.698554	0	1
5	test (in mu U/mL)	60.897135	77.644418	0	1

6	BMI (in kg/m^2)	32.198958	6.410558	0	1
7	pedi	0.427667	0.245162	0	1
8	Age (in years)	32.760417	11.055385	0	1

Inferences:

1. Before Standardization the mean and standard deviation is not constant for the attributes , it is done for the data having Gaussian distribution so that machine learning algorithms can be better.
2. The mean and standard deviation after standardization is 0 and 1 respectively.

2 a.

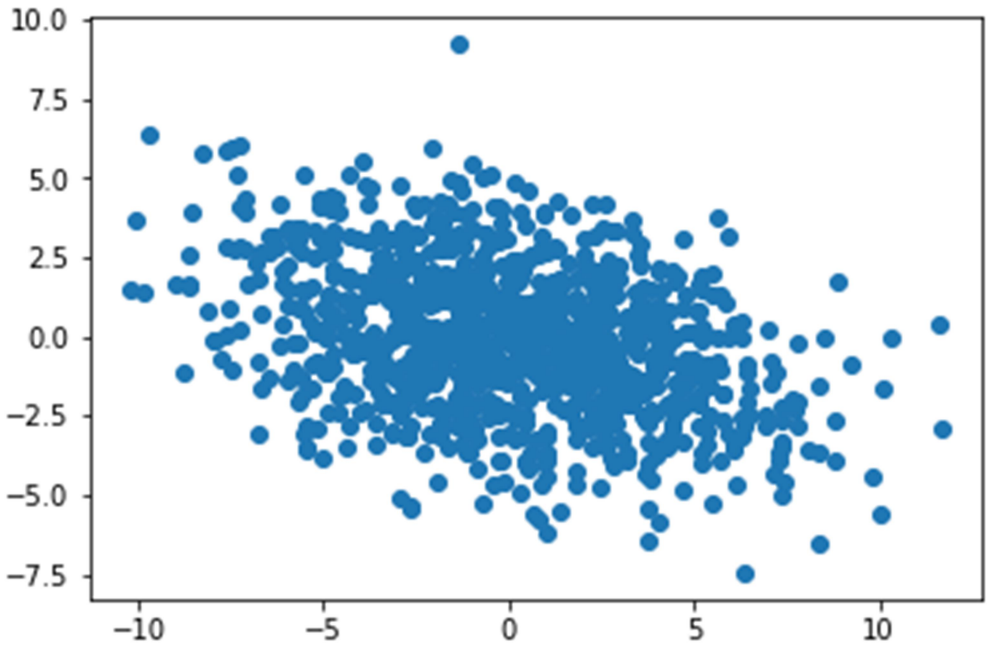

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. Attribute 1 is negatively correlated with Attribute 2 because as x values increases the y values decreases , but the correlation is very weak.
2. When attribute 1 lies in between -5 to 5 then attribute y lies in between -2.5 to 5.0.

b.

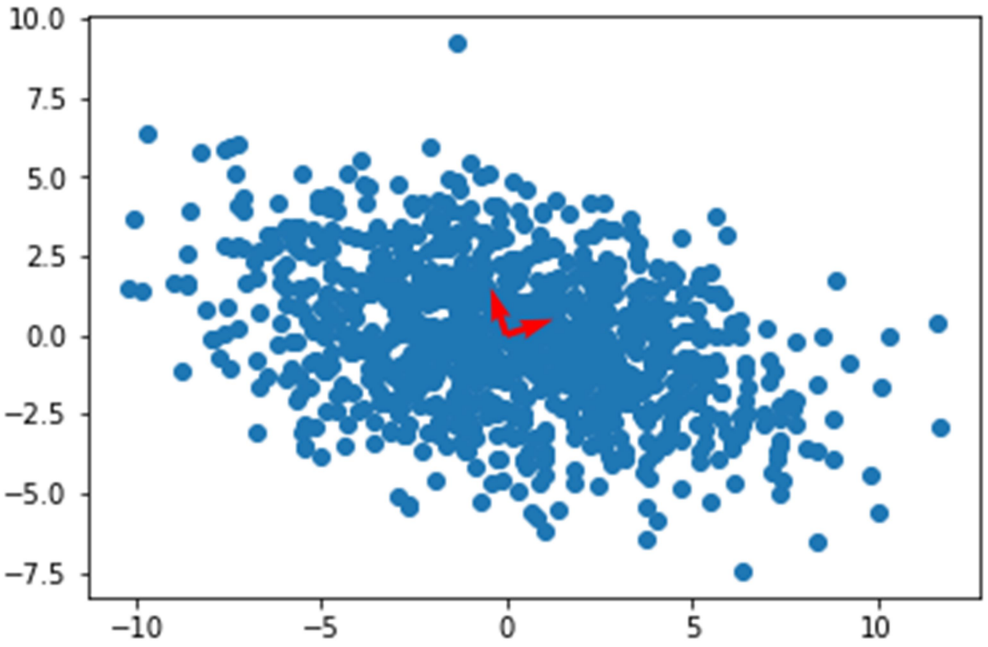

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. As the eigen values represents the stretching /compressing of linear transformation , since in the region between -5 to 5 on x axis eigen values lies so most of the data falls in that region.
2. As the eigen values comes closer the density of scatter plot increases and when the eigen values get away the density of scatter plot decreases.

c.

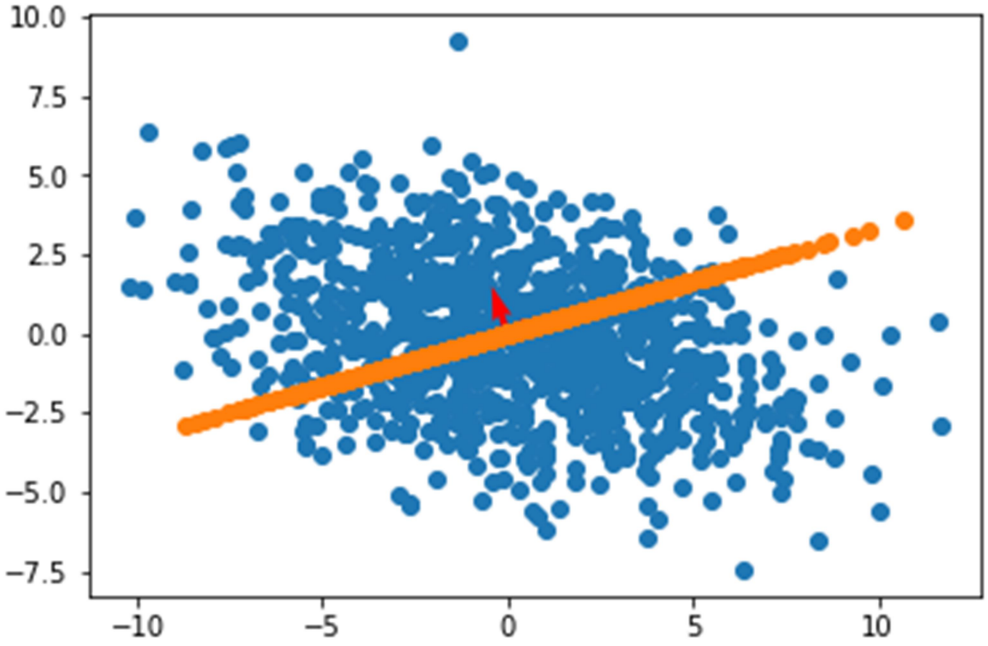

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

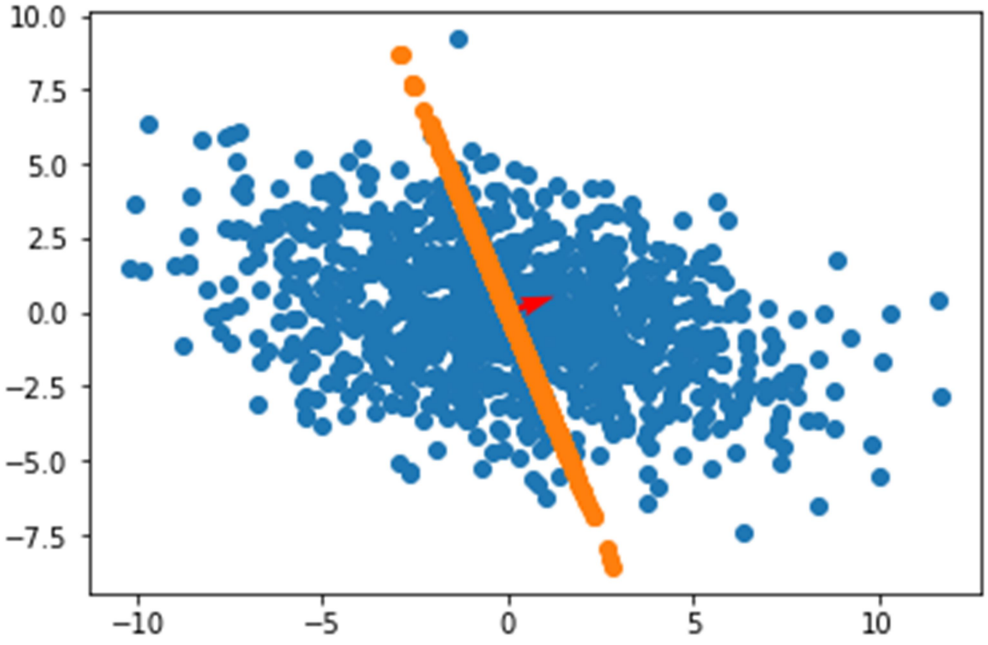

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. The magnitude of first eigen value is more as compared to magnitude of second eigen value .
2. It is possible to reconstruct the good approximation of original data as reconstruction is very low.

d. Reconstruction error =0.000

Inferences:

1. The magnitude of reconstruction error nearly approaches to 0 so data can be reconstructed without much loss.
2. The loss from the data is very low.

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 value of eigen values and variance are nearly same.

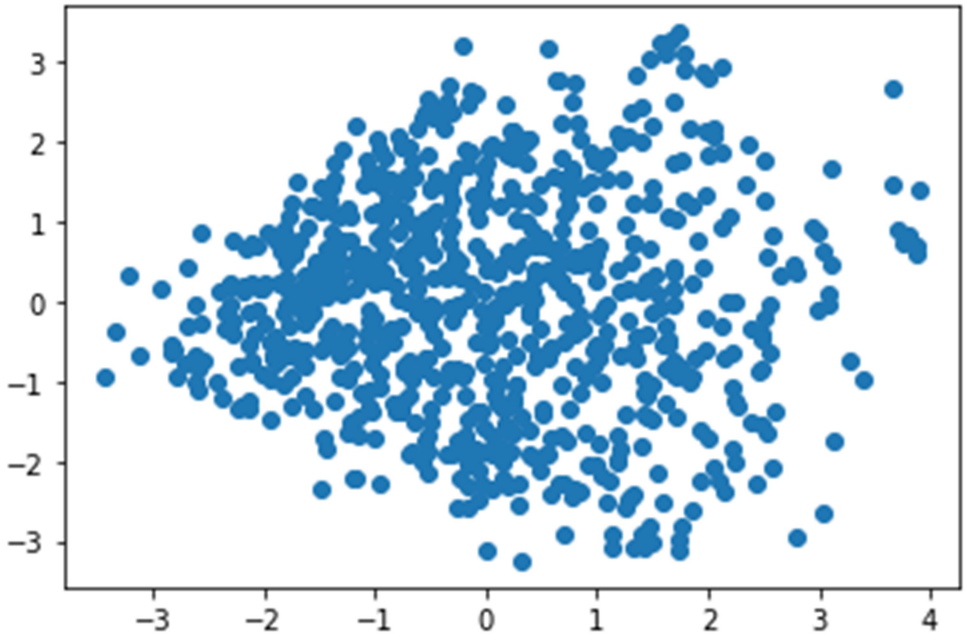

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. After the dimensionality reduction it can be seen that there is no correlation between attributes.
2. Also after dimensionality reduction the data is more spreaded more on scatter plot and it is not concentrated in any interval.

b.

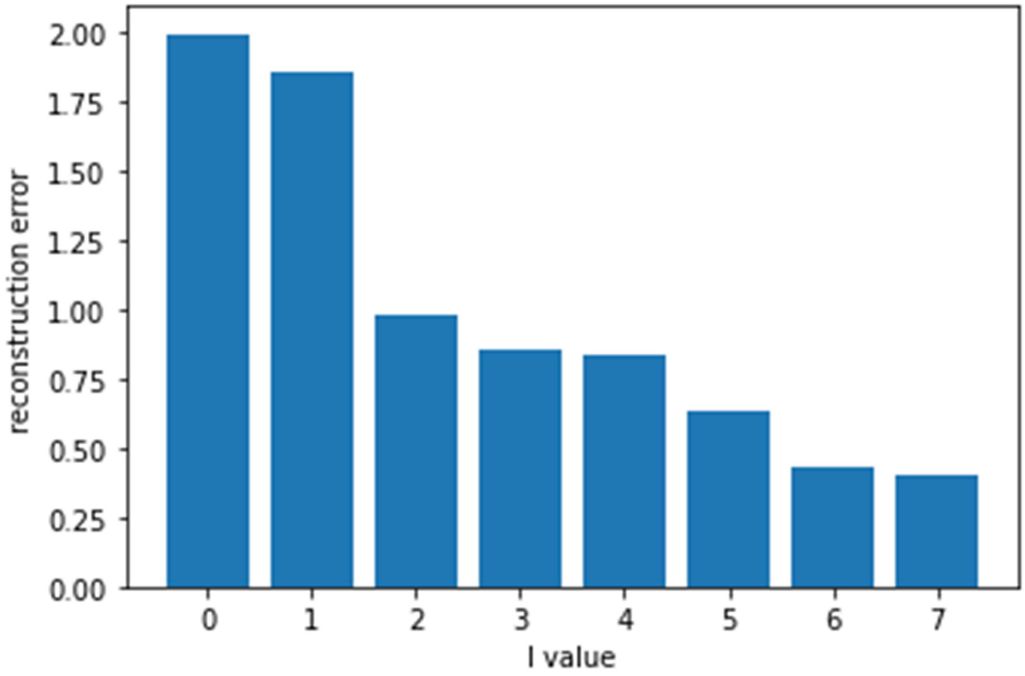

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. The subsequent eigenvalues are decreasing gradually , when arranged in descending order.
2. The eigenvalue at l =1 is the value after that the eigenvalues decreasing substanlly.

c.

```
[[ 1.99235918e+00 -2.25446201e-15]
 [-2.25446201e-15  1.85321993e+00]]
```

Table 4 Covariance matrix for dimensionally reduced data (l=3)

```
[[ 1.99235918e+00 -1.30867358e-15  2.40861944e-16]
 [-1.30867358e-15  1.85321993e+00 -4.26429861e-16]
 [ 2.40861944e-16 -4.26429861e-16  9.81883757e-01]]
```

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

```
[[ 1.99235918e+00 -1.56401040e-15  9.55052349e-16 -1.09430066e-16]
 [-1.56401040e-15  1.85321993e+00  6.14313757e-16 -2.83779953e-16]
 [ 9.55052349e-16  6.14313757e-16  9.81883757e-01 -6.54988160e-17]
 [-1.09430066e-16 -2.83779953e-16 -6.54988160e-17  8.58210408e-01]]
```

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

```
[[ 1.99235918e+00 -1.73988015e-16  3.34080149e-16  7.19256611e-16
   1.37511326e-16]
 [-1.73988015e-16  1.85321993e+00  1.58717020e-16  1.26401859e-16
  -1.94976586e-16]
 [ 3.34080149e-16  1.58717020e-16  9.81883757e-01 -6.81042938e-17
  -1.03097308e-15]
 [ 7.19256611e-16  1.26401859e-16 -6.81042938e-17  8.58210408e-01
  -3.05600829e-16]
 [ 1.37511326e-16 -1.94976586e-16 -1.03097308e-15 -3.05600829e-16
   8.38781538e-01]]
```

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

```
[[ 1.99235918e+00  9.82844112e-17 -6.45579490e-16  4.32509309e-16
   2.96300721e-16 -1.06774831e-16]
 [ 9.82844112e-17  1.85321993e+00  5.68066527e-16 -4.05296541e-17
   3.29448188e-16  6.60868578e-17]
 [-6.45579490e-16  5.68066527e-16  9.81883757e-01 -3.08170120e-16
   5.39333897e-16 -3.30750927e-17]
 [ 4.32509309e-16 -4.05296541e-17 -3.08170120e-16  8.58210408e-01
   1.54338370e-16  2.77049135e-16]
 [ 2.96300721e-16  3.29448188e-16  5.39333897e-16  1.54338370e-16
   8.38781538e-01 -1.43839561e-16]
 [-1.06774831e-16  6.60868578e-17 -3.30750927e-17  2.77049135e-16
  -1.43839561e-16  6.36524102e-01]]
```

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

```
[[ 1.99235918e+00 -2.40036876e-15  2.58810791e-16  2.21465610e-16
  -2.47520387e-17  8.07200533e-17  4.10507496e-16]
 [-2.40036876e-15  1.85321993e+00  4.00085585e-16  2.21103738e-17
   5.42663119e-16 -2.14381967e-16 -2.45349156e-17]
 [ 2.58810791e-16  4.00085585e-16  9.81883757e-01  1.39610183e-16
  -3.73089941e-16 -5.85816344e-16  8.75730026e-17]
 [ 2.21465610e-16  2.21103738e-17  1.39610183e-16  8.58210408e-01
  -5.99296073e-16  6.12400359e-17 -1.09212943e-16]
 [-2.47520387e-17  5.42663119e-16 -3.73089941e-16 -5.99296073e-16
   8.38781538e-01  2.67516072e-16  6.96241558e-17]
 [ 8.07200533e-17 -2.14381967e-16 -5.85816344e-16  6.12400359e-17
   2.67516072e-16  6.36524102e-01 -2.18937030e-16]
 [ 4.10507496e-16 -2.45349156e-17  8.75730026e-17 -1.09212943e-16
   6.96241558e-17 -2.18937030e-16  4.34392150e-01]]
```

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

```
[[ 1.99235918e+00 -2.40905369e-15  2.62863757e-16  2.05253748e-16
  -6.07944811e-18  8.41940236e-17  3.91690157e-16 -3.43416444e-17]
 [-2.40905369e-15  1.85321993e+00  4.00085585e-16  2.21103738e-17
   5.42663119e-16 -2.14381967e-16 -2.45349156e-17 -5.26790512e-16]
 [ 2.62863757e-16  4.00085585e-16  9.81883757e-01  1.39610183e-16
  -3.73089941e-16 -5.85816344e-16  8.75730026e-17  3.83448524e-16]
 [ 2.05253748e-16  2.21103738e-17  1.39610183e-16  8.58210408e-01
  -5.99296073e-16  6.12400359e-17 -1.09212943e-16  4.10272280e-17]
 [-6.07944811e-18  5.42663119e-16 -3.73089941e-16 -5.99296073e-16
   8.38781538e-01  2.67516072e-16  6.96241558e-17  6.47040548e-16]
 [ 8.41940236e-17 -2.14381967e-16 -5.85816344e-16  6.12400359e-17
   2.67516072e-16  6.36524102e-01 -2.18937030e-16 -1.87581960e-16]
 [ 3.91690157e-16 -2.45349156e-17  8.75730026e-17 -1.09212943e-16
   6.96241558e-17 -2.18937030e-16  4.34392150e-01 -9.41319310e-17]
 [-3.43416444e-17 -5.26790512e-16  3.83448524e-16  4.10272280e-17
   6.47040548e-16 -1.87581960e-16 -9.41319310e-17  4.04628935e-01]]
[[ 1.99235918e+00 -2.40905369e-15  2.62863757e-16  2.05253748e-16
  -6.07944811e-18  8.41940236e-17  3.91690157e-16 -3.43416444e-17]
 [-2.40905369e-15  1.85321993e+00  4.00085585e-16  2.21103738e-17
   5.42663119e-16 -2.14381967e-16 -2.45349156e-17 -5.26790512e-16]
 [ 2.62863757e-16  4.00085585e-16  9.81883757e-01  1.39610183e-16
  -3.73089941e-16 -5.85816344e-16  8.75730026e-17  3.83448524e-16]
 [ 2.05253748e-16  2.21103738e-17  1.39610183e-16  8.58210408e-01
  -5.99296073e-16  6.12400359e-17 -1.09212943e-16  4.10272280e-17]
 [-6.07944811e-18  5.42663119e-16 -3.73089941e-16 -5.99296073e-16
   8.38781538e-01  2.67516072e-16  6.96241558e-17  6.47040548e-16]
 [ 8.41940236e-17 -2.14381967e-16 -5.85816344e-16  6.12400359e-17
   2.67516072e-16  6.36524102e-01 -2.18937030e-16 -1.87581960e-16]
 [ 3.91690157e-16 -2.45349156e-17  8.75730026e-17 -1.09212943e-16
   6.96241558e-17 -2.18937030e-16  4.34392150e-01 -9.41319310e-17]
 [-3.43416444e-17 -5.26790512e-16  3.83448524e-16  4.10272280e-17
   6.47040548e-16 -1.87581960e-16 -9.41319310e-17  4.04628935e-01]]
```

Inferences:

d.

Table 9 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1.	0.117	0.208	-0.096	-0.108	0.0283	0.0045	0.5607
plas	0.117	1.	0.204	0.060	0.179	0.228	0.0816	0.274
pres (in mm Hg)	0.208	0.2045	1.	0.0256	-0.050	0.271	0.0224	0.3263
skin (in mm)	-0.0967	0.060	0.025	1.	0.4724	0.373	0.152	-0.101
test (in mu U/mL)	-0.108	0.179	-0.050	0.472	1.	0.171	0.198	-0.0737
BMI (in kg/m^2)	0.028	0.228	0.271	0.373	0.171	1.	0.123	0.077
pedi	0.004	0.081	0.0224	0.1527	0.198	0.123	1.	0.036
Age (in years)	0.560	0.274	0.326	-0.101	-0.073	0.077	0.0361	1.

Inferences:

1. After PCA l = 8 the values are increasing.
2. The magnitude of diagonal values are equal to 1 as they are variance between same attributes.