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

Student's Name: ISHAAN GUPTA **Mobile No: 9179242114**

Roll Number: B20292 **Branch: MECHANICAL ENGINEERING**

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	2.420	5	12
8	Age (in years)	21	81	5	12

Inferences:

1. The need for outlier correction is that statistical patterns and conclusions might differ between analyses.
2. In min-max normalization minimum value is scaled to 5 and maximum value is scaled to 12 in this case.
3. All the minimum value changes to 5 and all the maximum value changes to 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	3.845	3.370	0	1
2	plas	120.895	31.973	0	1
3	pres (in mm Hg)	69.105	19.365	0	1
4	skin (in mm)	20.536	15.952	0	1
5	test (in mu U/mL)	79.799	115.244	0	1
6	BMI (in kg/m^2)	31.993	7.884	0	1
7	pedi	0.472	0.331	0	1
8	Age (in years)	33.241	11.760	0	1

Inferences:

1. In standardization mean is scaled to 0 and standard deviation is scaled to 12 in this case.

2 a.

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

Inferences:

1. Attribute 1 is negatively correlated to attribute 2 .
2. The density of points is high on origin.

b.

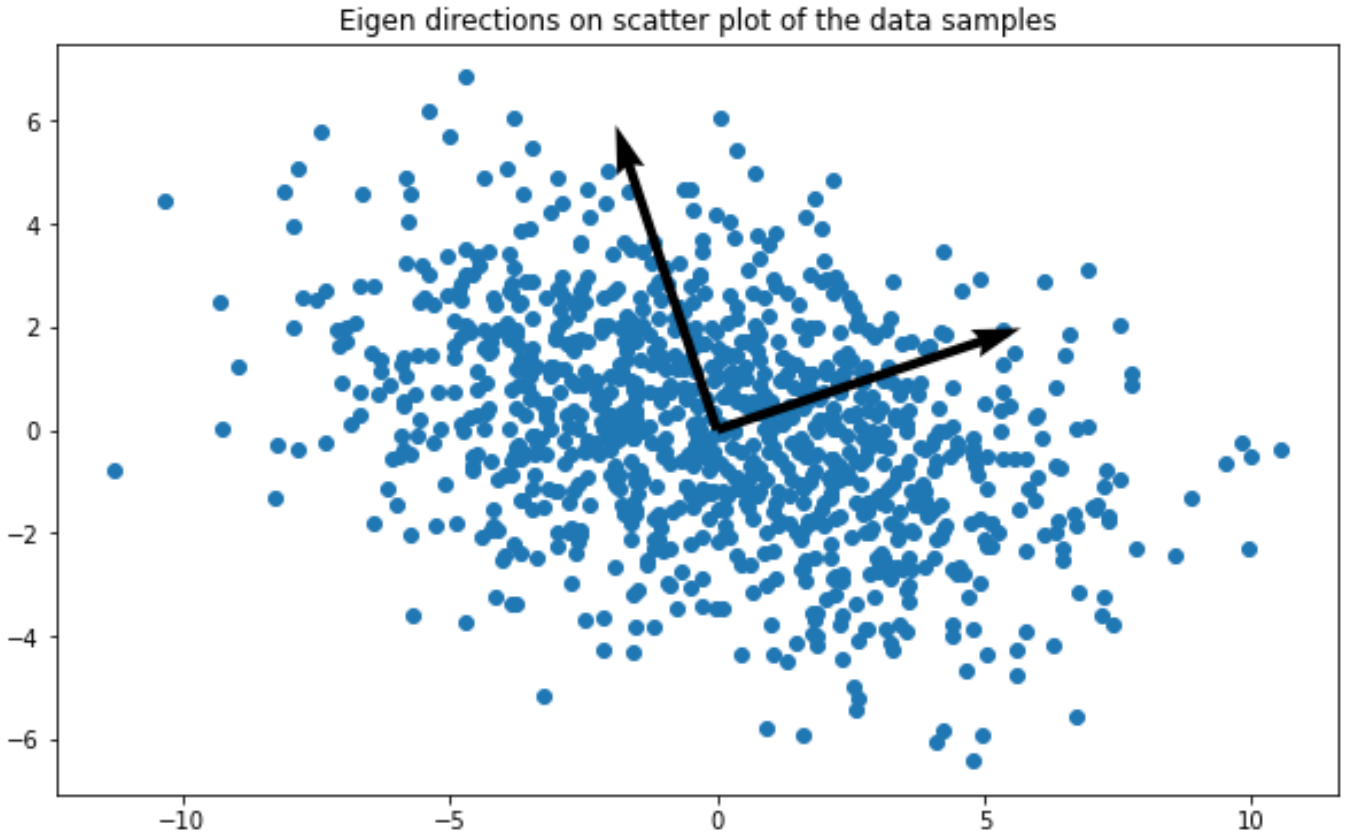

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. There is not much spreaded data based upon the magnitude of Eigenvalues.
2. The density of points is high on origin.

c.

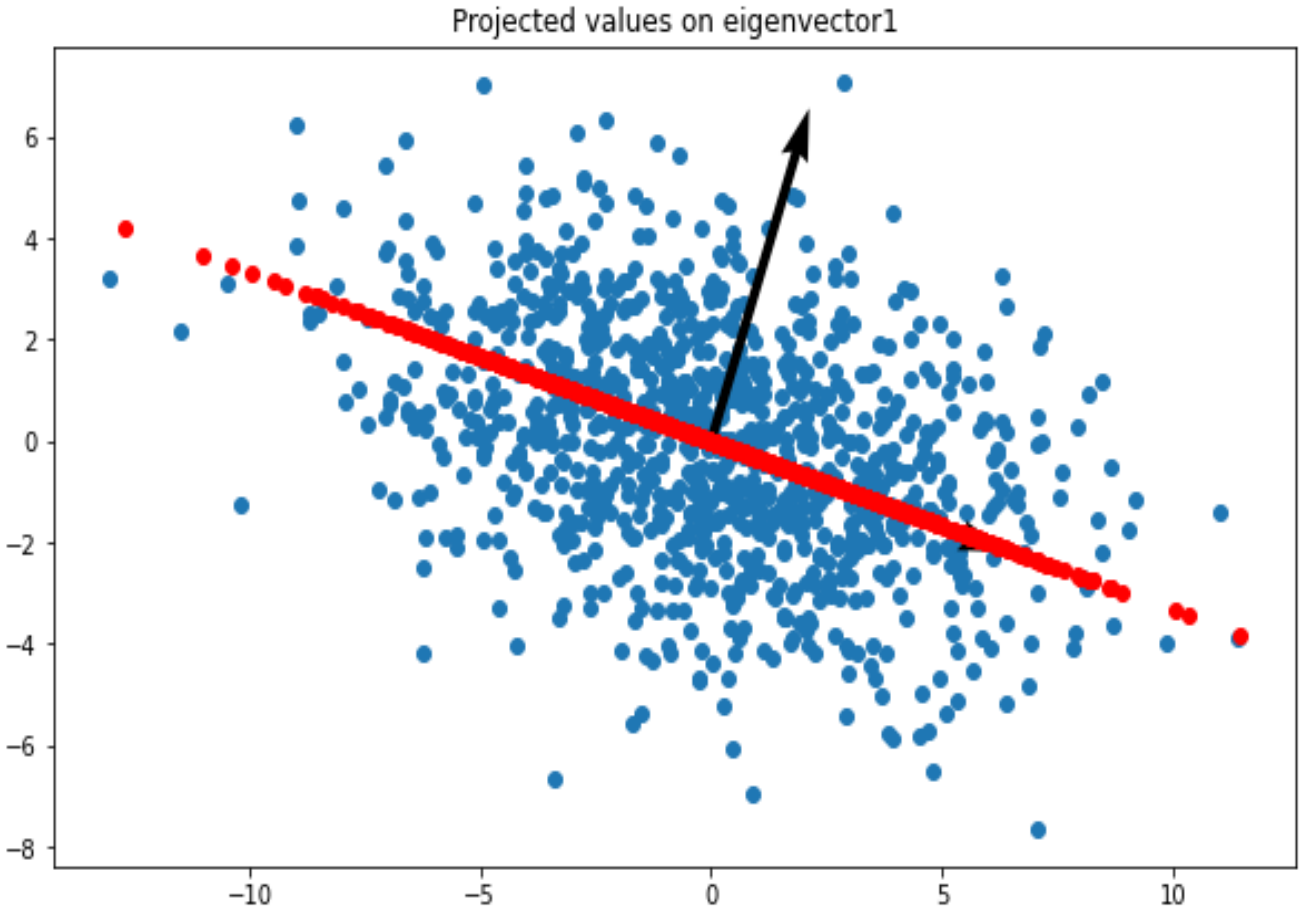

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

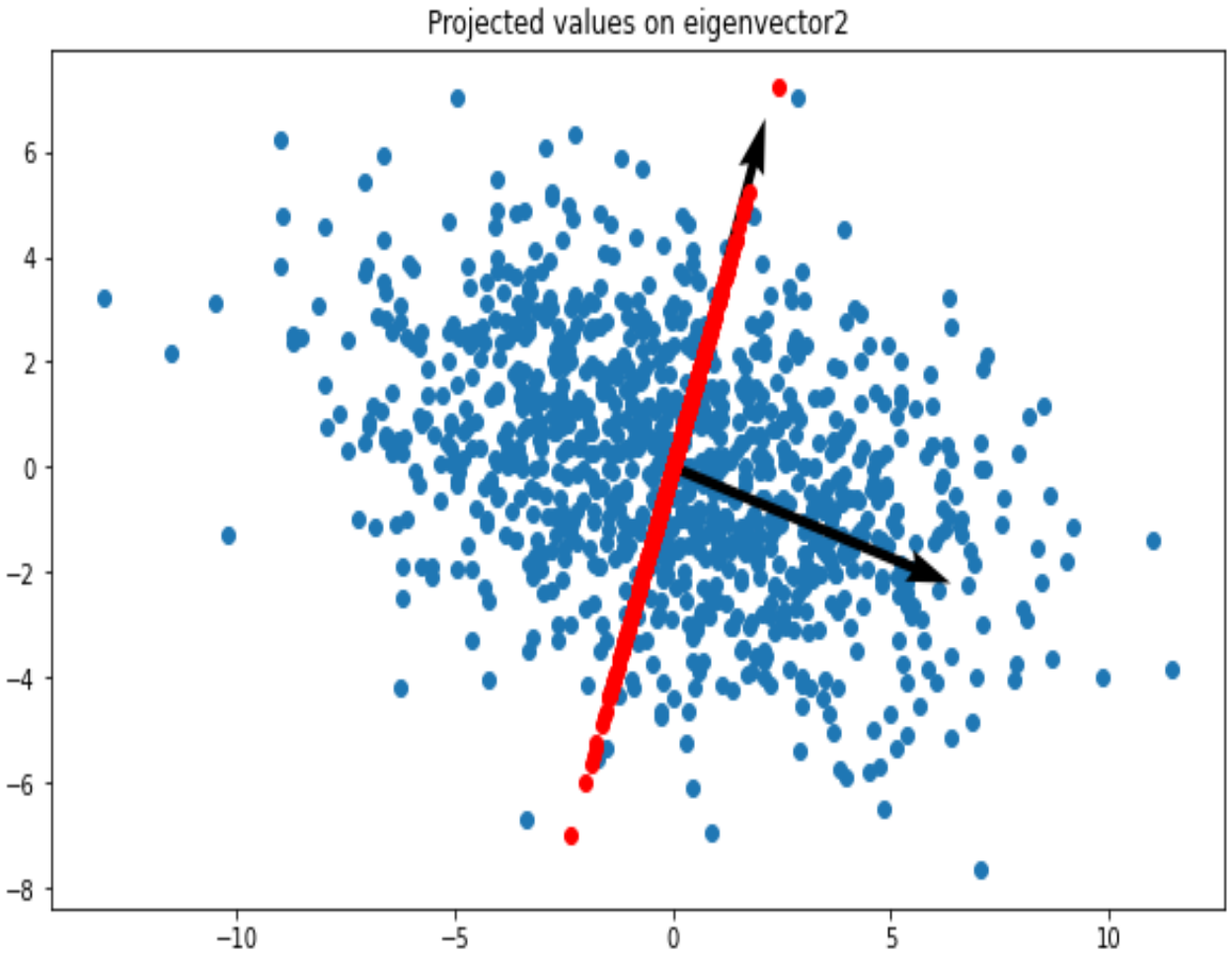

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. Compare and contrast the magnitude of Eigenvalues
2. Infer variance of data along the Eigen axes from spread & density of points and relate it to the magnitude of Eigenvalues.

d. Reconstruction error = 3.1463713616948717e-16

Inferences:

1. Infer how the magnitude of reconstruction error affects the quality of reconstruction.
2. Inference 2(You may add or delete the number of inferences)

3 a.

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

Direction	Variance	Eigenvalue
1	2.091	2.094
2	1.728	1.731

Inferences:

1. Variance 1 is more than Variance 2.

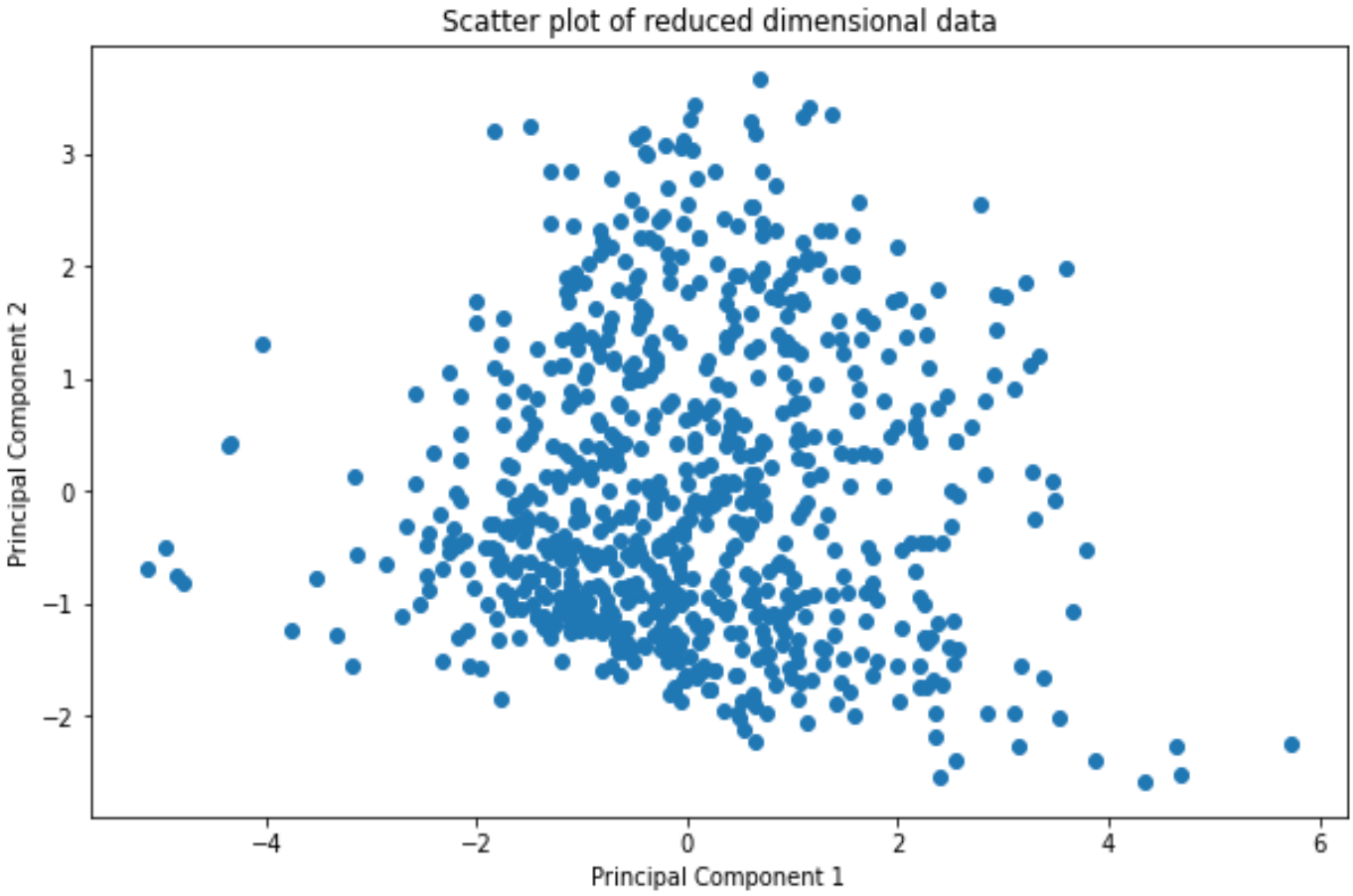

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. Infer the correlation between the two attributes obtained after dimensionality reduction from the spread of data points
2. Inference 2(You may add or delete the number of inferences)
   Note: The scatter plots above are for illustration purposes. Replace it with the scatter plot obtained by you. Rename x-axis legend with x1 and y-axis legend with x2.

b.

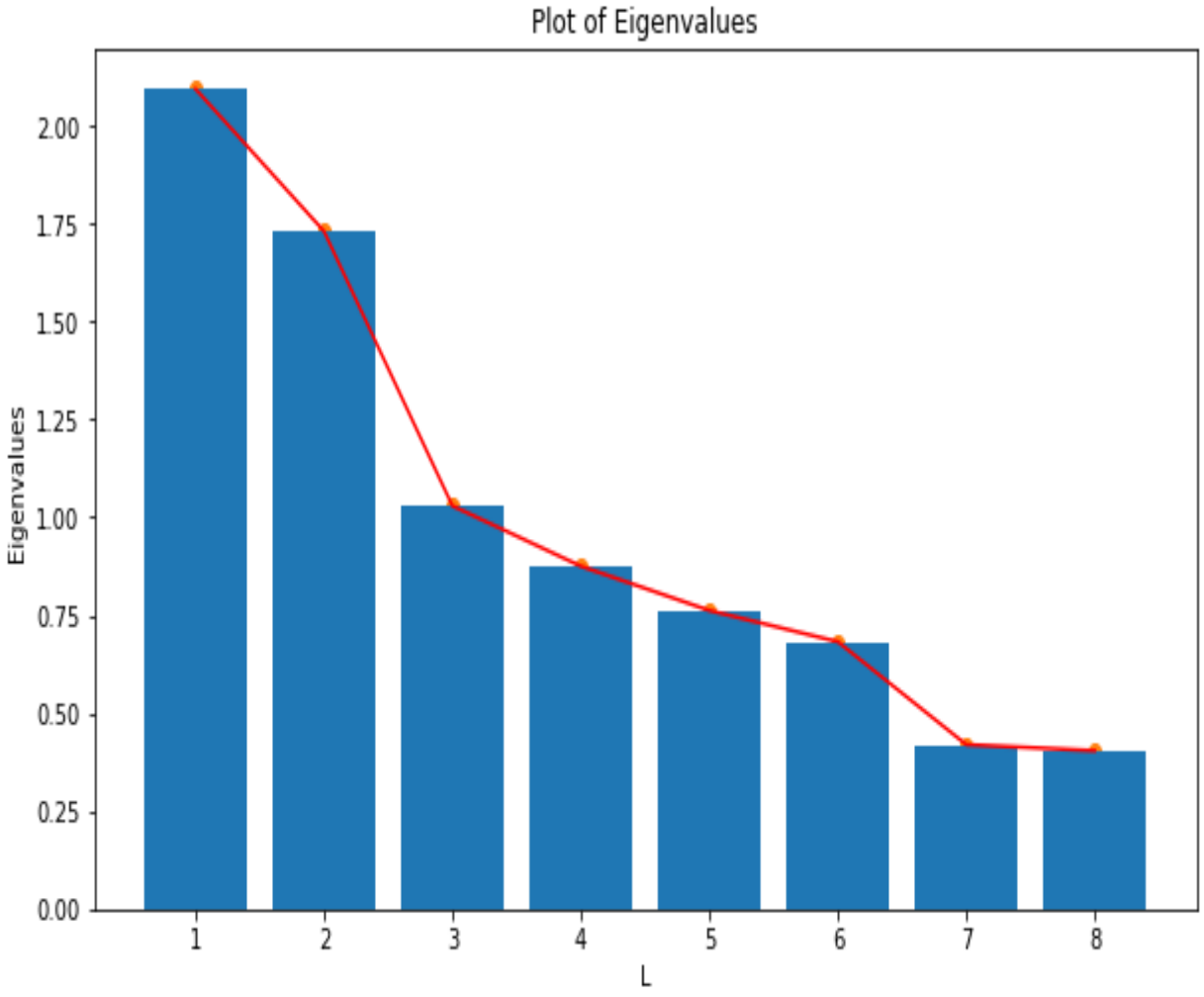

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. Infer whether the subsequent Eigenvalues decrease gradually or rapidly
2. Identify the Eigenvalue from where the rate of decrease changes substantially
3. Inference 3(You may add or delete the number of inferences)
   Note: The plot above is for illustration purposes. Replace it with the plot obtained by you. Rename x-axis legend with Eigenvalues and y-axis legend with magnitude.

c.

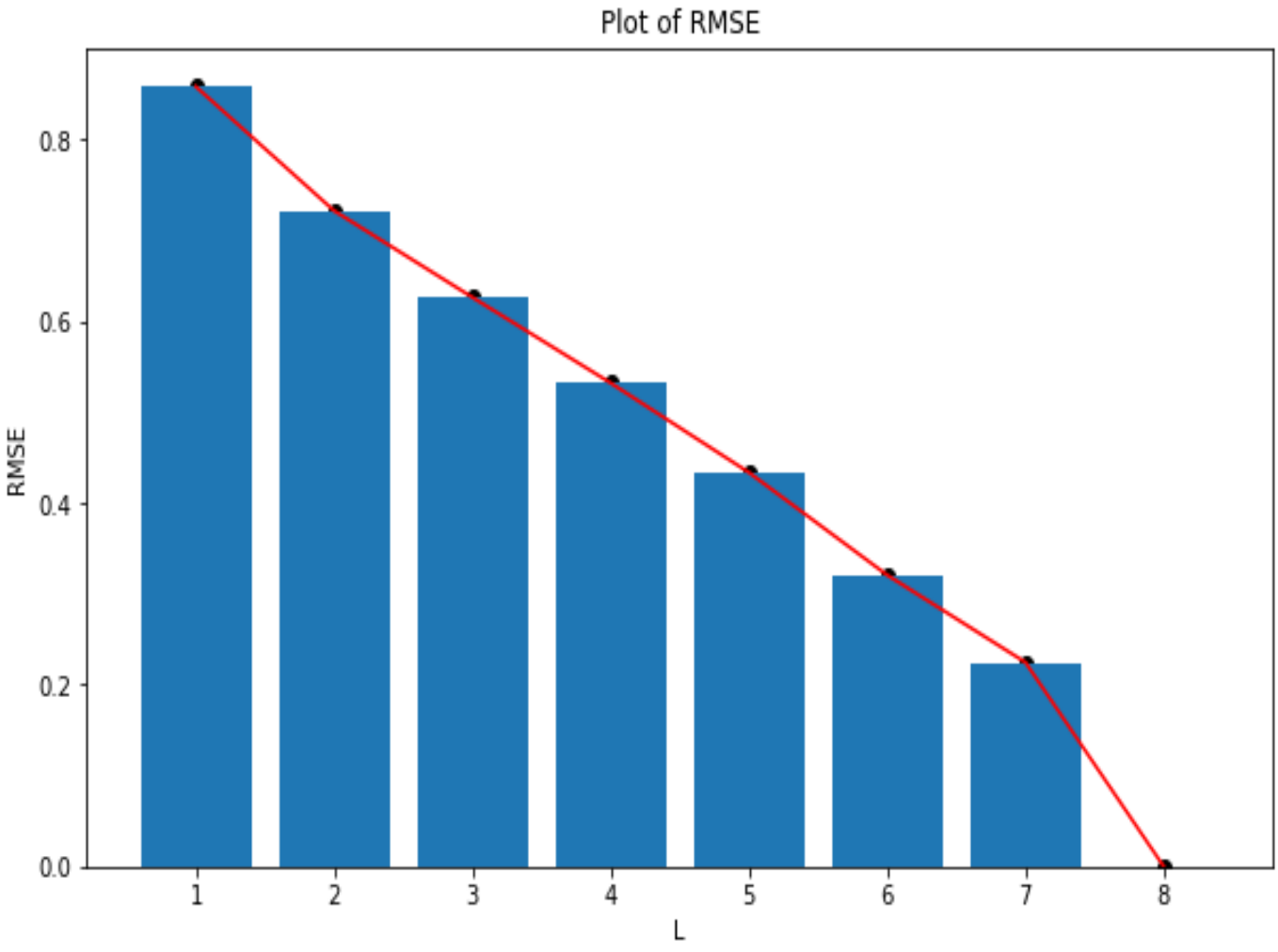

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. Infer how the magnitude of reconstruction error affects the quality of reconstruction.
2. Inference 2(You may add or delete the number of inferences)
   Note: The plot above is for illustration purposes. Replace it with the plot obtained by you. Rename x-axis legend with No. of components and y-axis legend with Reconstruction error.

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

	0	1
0	2.094379945288804	1.4683314682917324e-15
1	1.4683314682917324e-15	1.7312101406197233

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

	0	1	2
0	2.0943799452888046	-2.0380626058307957e-16	-2.408619443254577e-16
1	-2.0380626058307957e-16	1.731210140619726	-3.5318698566954133e-17
2	-2.408619443254577e-16	-3.5318698566954133e-17	1.029629869184153

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

	0	1	2	3
0	2.094379945288807	-3.242372327458084e-17	4.6319604677972634e-17	1.5748665590510694e-16
1	-3.242372327458084e-17	1.7312101406197253	1.030611204084891e-16	-1.1579901169493159e-16
2	4.6319604677972634e-17	1.030611204084891e-16	1.029629869184154	5.141476119254962e-16
3	1.5748665590510694e-16	-1.1579901169493159e-16	5.141476119254962e-16	0.8755290438080346

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

	0	1	2	3	4

0	2.094379945288805	1.2043097216272884e-16	-2.7791762806783576e-17	6.484744654916168e-17	9.263920935594527e-17
1	1.2043097216272884e-16	1.7312101406197218	7.179538725085758e-17	-7.411136748475621e-17	-1.5285469543730968e-16
2	-2.7791762806783576e-17	7.179538725085758e-17	1.0296298691841454	-1.0190313029153979e-16	1.4011680415086722e-16
3	6.484744654916168e-17	-7.411136748475621e-17	-1.0190313029153979e-16	0.8755290438080354	-1.8991037917968778e-16
4	9.263920935594527e-17	-1.5285469543730968e-16	1.4011680415086722e-16	-1.8991037917968778e-16	0.7623443855511708

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

	0	1	2	3	4	5
0	2.094379945288804	-4.029805606983619e-16	5.558352561356715e-17	4.631960467797263e-18	4.307723235051455e-16	-2.223341024542686e-16
1	-4.029805606983619e-16	1.7312101406197244	8.684925877119868e-17	-4.631960467797263e-18	3.705568374237810 5e-17	-3.705568374237810 5e-17
2	5.558352561356715e-17	8.684925877119868e-17	1.0296298691841534	-3.937166397627673 7e-16	8.858624394662266e-17	-4.684070023059982e-16
3	4.631960467797263e-18	-4.631960467797263e-18	-3.937166397627673 7e-16	0.8755290438080335	-5.071996712238003 5e-16	1.7717248789324 53e-16
4	4.307723235051455e-16	3.705568374237810 5e-17	8.858624394662266e-17	-5.071996712238003 5e-16	0.7623443855511717	4.342462938559934e-19
5	-2.223341024542686e-16	-3.705568374237810 5e-17	-4.684070023059982e-16	1.7717248789324 53e-16	4.342462938559934e-19	0.6826283879464935

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

	0	1	2	3	4	5	6

0	2.0943799452888037	5.465713352000771e-16	-2.223341024542686e-16	4.168764421017537e-17	-9.263920935594526e-18	-3.9371663976276737e-16	1.551706756712083e-16
1	5.465713352000771e-16	1.7312101406197242	2.1422817163562342e-17	1.8527841871189053e-17	3.7518879789157833e-16	-7.874332795255347e-17	7.874332795255347e-17
2	-2.223341024542686e-16	2.1422817163562342e-17	1.0296298691841554	-3.705568374237811e-16	-3.022354205237714e-16	-1.812254533025679e-16	3.1265733157631526e-17
3	4.168764421017537e-17	1.8527841871189053e-17	-3.705568374237811e-16	0.8755290438080354	5.558352561356715e-17	2.874710465326676e-16	-3.010774304068221e-17
4	-9.263920935594526e-18	3.7518879789157833e-16	-3.022354205237714e-16	5.558352561356715e-17	0.7623443855511708	-1.881733940042638e-18	-9.958715005764116e-17
5	-3.9371663976276737e-16	-7.874332795255347e-17	-1.812254533025679e-16	2.874710465326676e-16	-1.881733940042638e-18	0.6826283879464933	-2.084382210508768e-17
6	1.551706756712083e-16	7.874332795255347e-17	3.1265733157631526e-17	-3.010774304068221e-17	-9.958715005764116e-17	-2.084382210508768e-17	0.4198161797057532

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

	0	1	2	3	4	5	6	7
0	2.094379945288803 7	5.465713352000771e-16	-2.223341024542686e-16	4.168764421017537e-17	-9.263920935594526e-18	-3.937166397627673 7e-16	1.551706756712083e-16	8.337528842035074e-17

1	5.4657133 52000771 e-16	1.7312101 40619724 2	2.1422817 16356234 2e-17	1.8527841 87118905 3e-17	3.7518879 78915783 3e-16	-7.8743327 95255347 e-17	7.8743327 95255347 e-17	-2.4317792 45593563 2e-17
2	-2.2233410 24542686 e-16	2.1422817 16356234 2e-17	1.0296298 69184155 4	-3.7055683 74237811 e-16	-3.0223542 05237714 4e-16	-1.8122545 33025679 2e-16	3.1265733 15763152 6e-17	2.1770214 19864713 7e-16
3	4.1687644 21017537 e-17	1.8527841 87118905 3e-17	-3.7055683 74237811 e-16	0.8755290 43808035 4	5.5583525 61356715 e-17	2.8747104 65326676 e-16	-3.0107743 04068221 e-17	-3.2423723 27458084 e-17
4	-9.2639209 35594526 e-18	3.7518879 78915783 3e-16	-3.0223542 05237714 4e-16	5.5583525 61356715 e-17	0.7623443 85551170 8	-1.8817339 40042638 2e-18	-9.9587150 05764116 e-17	-1.1579901 16949315 9e-17
5	-3.9371663 97627673 7e-16	-7.8743327 95255347 e-17	-1.8122545 33025679 2e-16	2.8747104 65326676 e-16	-1.8817339 40042638 2e-18	0.6826283 87946493 3	-2.0843822 10508768 5e-17	-1.6211861 63729042 e-17
6	1.5517067 56712083 e-16	7.8743327 95255347 e-17	3.1265733 15763152 6e-17	-3.0107743 04068221 e-17	-9.9587150 05764116 e-17	-2.0843822 10508768 5e-17	0.4198161 79705753 2	-8.1059308 1864521e-18
7	8.3375288 42035074 e-17	-2.4317792 45593563 2e-17	2.1770214 19864713 7e-16	-3.2423723 27458084 e-17	-1.1579901 16949315 9e-17	-1.6211861 63729042 e-17	-8.1059308 1864521e-18	0.4044620 47895868 3

Inferences:

1. Off-diagonal elements tend to 0 as eigen vectors are orthonormal.
2. Attributes are uncorrelated.
3. Diagonal values are in decreasing order.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1.000000000000000029	0.12945867149927276	0.14128197740714027	-0.08167177444900726	-0.07353461435162811	0.01768309072783645	-0.033522672962613 2	0.544341228402340 1
plas	0.129458671499272 76	0.99999999999999993	0.152589586568664 42	0.05732789073817688	0.33135710992020867	0.221071069458983 08	0.137337299828370 67	0.263514319824333 6
pres	0.141281977407140 27	0.15258958656866442	1.000000000000000013	0.20737053840307038	0.08893337837319289	0.281805288849910 9	0.041264947930098536	0.239527946421363 66
skin	-0.081671774449007 26	0.05732789073817688	0.207370538403070 38	0.99999999999999962	0.43678257012001 25	0.392573204159037 9	0.183927572954162 76	-0.113970262367741 38
test	-0.073534614351628 11	0.33135710992020867	0.088933378373192 89	0.43678257012001 25	0.99999999999999953	0.197859056493100 82	0.185070929168098 75	-0.042162954735376 79
BMI	0.017683090727830645	0.22107106945898308	0.281805288849910 9	0.39257320415903 79	0.19785905649310082	1.000000000000000018	0.140646952545105 34	0.036241870092294085
pedi	-0.033522672962613 2	0.13733729982837067	0.041264947930098536	0.18392757295416276	0.18507092916809875	0.140646952545105 34	1.000000000000000009	0.033561312434805576
Age	0.544341228402340 1	0.26351431982433 36	0.239527946421363 66	-0.11397026236774138	-0.04216295473537679	0.036241870092294085	0.033561312434805576	1.000000000000000016

Inferences:

1. Off-diagonal values are weakly correlated , therefore , they are non 0.
2. The magnitudes of diagonal values are 1 as they are standardized.