1) Consider the following var-cov matrix and find the PC's Also explain the proportion of variability explained by each PC and comment on the no of PC's to be retained if someone wishes to explain atleast 80% of the total system variability.

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

ans) By looking at Σ we can say that,

$|\Sigma| = 2(5-4) = 2 \neq 0$ i.e Σ is non singular.

Let $\lambda_1, \lambda_2, \lambda_3$ be the eigenvalue of Σ i.e we need to solve $|\Sigma - \lambda I| = 0$

$\therefore |\Sigma - \lambda I| = 0$

$$\Rightarrow \begin{vmatrix} 1-\lambda & -2 & 0 \\ -2 & 5-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{vmatrix} = 0$$

$$\Rightarrow (2-\lambda)[(1-\lambda)(5-\lambda) - 4] = 0$$

$$\Rightarrow (\lambda-2)[\lambda^2 - 6\lambda + 1] = 0$$

$$\Rightarrow \lambda = 2 \quad \text{or} \quad \lambda = \frac{6 \pm \sqrt{36-4}}{2} = 3 \pm 2\sqrt{2}$$

i.e $\lambda_1 = 3+2\sqrt{2}$, $\lambda_2 = 2$, $\lambda_3 = 3-2\sqrt{2}$

$\Rightarrow \lambda_1 = 5.828$, $\lambda_2 = 2$, $\lambda_3 = 0.172$

and $\lambda_1 > \lambda_2 > \lambda_3$

Now we have to find orthonormal eigenvectors corresponding to $\lambda_1, \lambda_2, \lambda_3$.

Let $\underset{\sim}{a}_i$ be the eigenvector corresponding to λ_i; $i = 1(1)3$

i.e $\underset{\sim}{a}_i = \begin{pmatrix} a_{i1} \\ a_{i2} \\ a_{i3} \end{pmatrix}$; $i = 1(1)3$

Now we have,

$$\Sigma \underset{\sim}{a}_1 = \lambda_1 \underset{\sim}{a}_1$$

$$\Rightarrow \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix} = \begin{pmatrix} 5.828\, a_{11} \\ 5.828\, a_{12} \\ 5.828\, a_{13} \end{pmatrix}$$

$$\Rightarrow -4.828\, a_{11} - 2a_{12} = 0 \quad \text{and} \quad -2a_{11} - 0.828\, a_{12} = 0 \quad \text{and} \quad a_{13} = 0$$

$$\Rightarrow a_{11} + 0.414\, a_{12} = 0$$

Now if $a_{11} = 1 \Rightarrow a_{12} = -2.415$

$$\therefore \underset{\sim}{a}_1 = \begin{pmatrix} 1 \\ -2.415 \\ 0 \end{pmatrix} \quad \text{and} \quad \|\underset{\sim}{a}_1\| = \sqrt{1^2 + (-2.415)^2} = 2.613$$

Normed

$$\therefore \underset{\sim}{e}_1 = \begin{pmatrix} 0.38 \\ -0.92 \\ 0 \end{pmatrix} \rightarrow \text{orthonormal eigenvector}$$

Hence the 1ST PC is $Y_1 = 0.38\, X_1 - 0.92 X_2$

Similarly for $\lambda_2 = 2$ we get,

$$\Sigma \underset{\sim}{a}_2 = \lambda_2 \underset{\sim}{a}_2$$

$$\Rightarrow \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} a_{21} \\ a_{22} \\ a_{23} \end{pmatrix} = \begin{pmatrix} 2\, a_{21} \\ 2\, a_{22} \\ 2\, a_{23} \end{pmatrix}$$

$$\Rightarrow a_{21} + 2a_{22} = 0 \quad \text{and} \quad -2a_{21} + 3a_{22} = 0 \quad \text{and} \quad 2a_{23} = 2a_{23}$$

$$\Rightarrow 2a_{21} = 3a_{22} \qquad \Rightarrow a_{23} = a_{23}$$

If $a_{21} = 0$ then $a_{22} = 0$ and $a_{23} = 1$

$$\therefore \underset{\sim}{a}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \|a_2\| = \sqrt{0^2 + 0^2 + 1^2} = 1$$

$$\therefore \underset{\sim}{e}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Hence the 2nd PC is $Y_2 = X_3$

Now, for $\lambda_3 = 0.172$ we have

$$\Sigma \underset{\sim}{a}_3 = \lambda_3 a_3$$

$$\Rightarrow \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} a_{31} \\ a_{32} \\ a_{33} \end{pmatrix} = \begin{pmatrix} 0.172\, a_{31} \\ 0.172\, a_{32} \\ 0.172\, a_{33} \end{pmatrix}$$

$\Rightarrow$ $0.828 a_{31} - 2a_{32} = 0$ and $-2a_{31} + 4.828 a_{32} = 0$

and $2a_{33} = 0.172\, a_{33}$

Solving the above we get $\underset{\sim}{a}_3 = \begin{pmatrix} 2.414 \\ 1 \\ 0 \end{pmatrix}$

and $\| a_3 \| = \sqrt{2.414^2 + 1^2 + 0^2} = 2.613$

Hence $\underset{\sim}{e}_3 = \begin{pmatrix} 0.923 \\ 0.383 \\ 0 \end{pmatrix}$

$\therefore$ The 3rd PC is $Y_3 = 0.923 X_1 + 0.383 X_2$

Now variance explained by the 1ST PC is $= \frac{5.828}{(5.828 + 2 + 0.172)}$

$= 0.728 = 72.8\%$.

varcie explained by 2nd PC is $\frac{2}{(5.828 + 2 + 0.172)} = 0.25 = 25\%$.

variance explained by 3rd PC is $\frac{0.172}{(5.828 + 2 + 0.172)} = 2.15\%$.

and variance explained by

1ST and 2ND PC is $\frac{5.828 + 0.172}{(5.828 + 2 + 0.172)} = 0.75 = 75\%$.

$\therefore$ If one wants to explain 80% of the total system variability one can retain the 1ST and 3RD PC.

2) Solve the following var-cov matrix and find the PC's Also convert the given matrix into correlation matrix, find the PC's and interpret the results.

$$\begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}$$

ans) By looking at Σ we can say that $|\Sigma| = 84 \neq 0$ i.e Σ is non singular.

Let λ_1 and λ_2 be the eigenvalue of Σ i.e we need to solve $|\Sigma - \lambda I| = 0$

$$\Rightarrow \begin{vmatrix} 1-\lambda & 4 \\ 4 & 100-\lambda \end{vmatrix} = 0$$

$$\Rightarrow (1-\lambda)(100-\lambda) - 16 = 0$$

$$\Rightarrow \lambda^2 - 101\lambda + 84 = 0$$

$$\therefore \lambda = \frac{101 \pm \sqrt{101^2 - (4 \times 84)}}{2} = 100.16 \text{ and } 0.84$$

$\therefore \lambda_1 = 100.16$ and $\lambda_2 = 0.84$ and $\lambda_1 > \lambda_2$

Now to have to find the orthonormal eigenvectors corresponding to λ_1 and λ_2

Let $\underline{a}_i' = (a_{i1}, a_{i2})$; $i = 1(1)2$ be the eigenvector to λ_i

Now we have, for $\lambda_1 = 100.16$

$$\Sigma \underline{a}_1 = \lambda_1 \underline{a}_1$$

$$\Rightarrow \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} 100.16\, a_{11} \\ 100.16\, a_{12} \end{pmatrix}$$

$$\Rightarrow 4a_2 - 99.16\, a_1 = 0 \quad \text{and} \quad 4a_1 - 0.16 a_2 = 0$$

$$\Rightarrow a_1 - 0.04 a_2 = 0$$

Let $a_2 = 1$ then $a_1 = 0.04$

$$\therefore \underline{a}_1 = \begin{pmatrix} 0.04 \\ 1 \end{pmatrix} \text{ and } \|a_1\| = \sqrt{1^2 + 0.04^2} \simeq 1$$

$$\therefore \underline{e}_1 = \begin{pmatrix} 0.04 \\ 1 \end{pmatrix}$$

Hence the 1ST PC is $Y_1 = 0.04 x_1 + x_2$

Now for $\lambda_2 = 0.84$ we have

$$\Sigma \underline{a}_2 = \lambda_2 \underline{a}_2$$

$$\Rightarrow \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}\begin{pmatrix} a_{21} \\ a_{22} \end{pmatrix} = \begin{pmatrix} 0.84\, a_{21} \\ 0.84\, a_{22} \end{pmatrix}$$

$$\Rightarrow \quad 0.16\, a_1 + 4a_{12} = 0 \quad \text{and} \quad 4a_{21} + 99.16\, a_{22} = 0$$

If $a_1 = 1$ then $a_2 = -0.04$

$$\therefore \quad \underset{\sim}{a}_2 = \begin{pmatrix} 1 \\ -0.04 \end{pmatrix} \quad \text{and} \quad \|a_2\| = \sqrt{1^2 + 0.04^2} \approx 1$$

Hence $\underset{\sim}{e}_2 = \begin{pmatrix} 1 \\ -0.04 \end{pmatrix}$

$\therefore$ The 2nd PC is $Y_2 = x_1 - 0.04 x_2$

Now converting the given var-cov matrix into correlation matrix we get,

$$R = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$$

Now $\det R = 1 - 0.4^2 = 0.84 \neq 0$ i.e R is non singular. Proceeding in a similar way we get,

$$|R - \lambda I| = 0 \Rightarrow \begin{vmatrix} 1-\lambda & 0.4 \\ 0.4 & 1-\lambda \end{vmatrix} = 0$$

$$\Rightarrow (1-\lambda)^2 - 0.4^2 = 0$$

$$\Rightarrow (1-\lambda + 0.4)(1-\lambda - 0.4) = 0$$

$$\Rightarrow (1.4 - \lambda)(0.6 - \lambda) = 0$$

$$\therefore \lambda_1 = 1.4 \quad \lambda_2 = 0.6$$

and $\lambda_1 > \lambda_2$

Now for the first eigenvalue λ_1 we get

$$R\, \underset{\sim}{a}_1 = \lambda_1\, \underset{\sim}{a}_1$$

$$\Rightarrow \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}\begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} 1.4\, a_{11} \\ 1.4\, a_{12} \end{pmatrix}$$

$\Rightarrow 0.4\,a_{12} - 0.4\,a_{11} = 0$ and $0.4\,a_{21} - 0.4\,a_{12} = 0$

$\Rightarrow a_{11} = a_{12}$ $\Rightarrow a_{11} = a_{21}$

Hence $\underset{\sim}{a}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $\|\underset{\sim}{a}_1\| = \sqrt{1+1} = \sqrt{2}$

$\therefore \underset{\sim}{e}_1 = \begin{pmatrix} 0.707 \\ 0.707 \end{pmatrix}$

$\therefore$ The 1ST PC is $Y_1 = 0.707\,x_1 + 0.707\,x_2$

Similarly for $\lambda_2 = 0.6$ we get,

$R\,\underset{\sim}{a}_2 = \lambda_2\,\underset{\sim}{a}$

$\Rightarrow \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}\begin{pmatrix} a_{21} \\ a_{22} \end{pmatrix} = \begin{pmatrix} 0.6\,a_{21} \\ 0.6\,a_{22} \end{pmatrix}$

$\Rightarrow a_{21} + 0.4\,a_{22} = 0.6\,a_{21}$ and $0.4\,a_{21} + a_{22} = 0.6\,a_{22}$

$\Rightarrow 0.4\,a_{22} = +0.4\,a_{21}$ and $0.4\,a_{21} + 0.4\,a_{22} = 0$

$\Rightarrow a_{22} = +a_{21}$ $\Rightarrow a_{21} = -a_{22}$

Hence $\underset{\sim}{a}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ $\|a_2\| = \sqrt{1^2+1^2} = \sqrt{2}$

$\therefore \underset{\sim}{e}_2 = \begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix}$

$\therefore$ The 2nd PC is $Y_2 = 0.707\,x_1 - 0.707\,x_2$

$\therefore$ We can see that the PC's obtained from var cov matrix and correlation matrix is not the same. It may also be noted that the corr- matrix is cov matrix of the standardized variables. PC's from corr-matrix is more preferred since,

- In cov matrix the 1ST PC explains $\frac{100.16}{101} = 0.992 = 99.2\%$
- In corr matrix the 1ST PC explains $\frac{1.4}{2} = 0.707 \simeq 70\%$

Moreover in corr matrix both the variables x_1 and x_2 contribute equally whereas in case of cov matrix it is not the case.

Canonical correlation between the variates U and V

1) Consider the following correlation matrix and try to find a 1 factor model solution of it.

$$R = \begin{pmatrix} 1 & 0.705 & 0.731 \\ 0.705 & 1 & 0.716 \\ 0.731 & 0.716 & 1 \end{pmatrix}$$

ans) We know that for a 1 factor model solution, the correlation matrix can be written as,

$$R = \begin{pmatrix} l_{11}^2 & l_{11}l_{12} & l_{11}l_{13} \\ l_{21}l_{11} & l_{21}^2 & l_{21}l_{31} \\ l_{31}l_{11} & l_{31}l_{21} & l_{31}^2 \end{pmatrix} + \begin{pmatrix} \psi_1^2 & 0 & 0 \\ 0 & \psi_2^2 & 0 \\ 0 & 0 & \psi_3^2 \end{pmatrix} \quad \text{---(1)}$$

$$= LL' + \psi^2$$

where l_{ij} = loading of ith variable on jth factor

and ψ_i^2 = specific variance

Now comparing the given correlation matrix with the above form we get,

$$l_{11}l_{21} = 0.705 \qquad l_{11}l_{31} = 0.731 \qquad l_{21}l_{31} = 0.716$$

Now, $l_{11} = \frac{0.705}{l_{21}}$ $\therefore$ $l_{11}l_{31} = 0.731 \Rightarrow \frac{0.705\, l_{31}}{l_{21}} = 0.731$

Now multiplying we get,

$$\frac{0.705}{l_{21}} l_{31} \times l_{21}l_{31} = 0.716 \times 0.731$$

$$\Rightarrow l_{31}^2 = \frac{0.716 \times 0.731}{0.705}$$

$$\Rightarrow l_{31} = \pm 0.861$$

Taking $l_{31} = 0.861$ we get, $l_{21} = \frac{0.705 \times 0.861}{0.731} = 0.831$

and $l_{11} = \frac{0.705}{0.831} = 0.848$

If $l_{31} = -0.861$ then the sign of the values of l_{21} and l_{11} will be -ve

Now Putting the values of l_{11}, l_{21}, l_{31} in the model of (1) the values of ψ_1^2, ψ_2^2 and ψ_3^2 can be obtained.

$\therefore \psi_1^2 = 1 - l_{11}^2 = 1 - (0.848)^2 = 0.280$

$\psi_2^2 = 1 - l_{21}^2 = 1 - (0.831)^2 = 0.309$

$\psi_3^2 = 1 - l_{31}^2 = 1 - (0.861)^2 = 0.258$

$\therefore$ Given the correlation matrix a one factor model solution exists and can be written as,

$$x_1 = 0.848 F_1 + \epsilon_1$$
$$x_2 = 0.831 F_1 + \epsilon_2$$
$$x_3 = 0.861 F_1 + \epsilon_3$$

2) Consider the var-cov matrix and find one factor model solution if it exists using the principal component method

$$\Sigma = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}$$

ans) Let us consider the one factor model as,

$$x_1 = l_{11} F_1 + \epsilon_1$$
$$x_2 = l_{21} F_1 + \epsilon_2$$

Now we know, $\Sigma = \begin{pmatrix} l_{11}^2 & l_{11} l_{21} \\ l_{21} l_{11} & l_{22}^2 \end{pmatrix} + \begin{pmatrix} \psi_1^2 & 0 \\ 0 & \psi_2^2 \end{pmatrix}$

$$= LL' + \Psi$$

Comparing the variance cov matrix with the above model we get, $l_{11} l_{21} = 4 \Rightarrow l_{11} = \frac{4}{l_{21}}$

So here it may not be easy to find the one factor

model solution by this process. We consider the PCA method to solve this.

Using spectral decomposition we can say,

$$\Sigma = \lambda_1 \underset{\sim}{e}_1 \underset{\sim}{e}_1' + \lambda_2 \underset{\sim}{e}_2 \underset{\sim}{e}_2' + \cdots + \lambda_p \underset{\sim}{e}_p \underset{\sim}{e}_p' \quad \text{--- (1)}$$

where λ_i = eigenvalues and $\underset{\sim}{e}_i$ = corresponding eigenvectors $i = 1(1)p$

or, $$\Sigma = \left(\sqrt{\lambda_1}\underset{\sim}{e}_1 \; \sqrt{\lambda_2}\underset{\sim}{e}_2 \cdots \sqrt{\lambda_p}\underset{\sim}{e}_p\right)\begin{pmatrix} \sqrt{\lambda_1}\underset{\sim}{e}_1' \\ \sqrt{\lambda_2}\underset{\sim}{e}_2' \\ \vdots \\ \sqrt{\lambda_p}\underset{\sim}{e}_p' \end{pmatrix}$$

$\therefore$ The given matrix can be written as,

$$\begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix} = \left(\sqrt{100.16}\begin{pmatrix} 0.04 \\ 1 \end{pmatrix} \quad \sqrt{0.84}\begin{pmatrix} 1 \\ -0.04 \end{pmatrix}\right)\begin{pmatrix} \sqrt{100.16}\,(0.04 \;\; 1) \\ \sqrt{0.84}\,(1 \;\; -0.04) \end{pmatrix}$$

where $\sqrt{\lambda_1} = \sqrt{100.16}$ $\quad \sqrt{\lambda_2} = \sqrt{0.84}$

and $\underset{\sim}{e}_1 = \begin{pmatrix} 0.04 \\ 1 \end{pmatrix}$ $\quad e_2 = \begin{pmatrix} 1 \\ -0.04 \end{pmatrix}$

$$= \begin{pmatrix} 0.400 & 0.917 \\ 10.008 & -0.037 \end{pmatrix}\begin{pmatrix} 0.400 & 10.008 \\ 0.917 & -0.037 \end{pmatrix}$$

$$\simeq \begin{pmatrix} 0.400 \\ 10.008 \end{pmatrix}\begin{pmatrix} 0.400 & 10.008 \end{pmatrix} + \begin{pmatrix} \psi_1^2 & 0 \\ 0 & \psi_2^2 \end{pmatrix}$$

$$\simeq \begin{pmatrix} l_{11} \\ l_{12} \end{pmatrix}\begin{pmatrix} l_{11} & l_{12} \end{pmatrix} + \begin{pmatrix} \psi_1^2 & 0 \\ 0 & \psi_2^2 \end{pmatrix}$$

$\therefore$ It can be seen that $l_{11} = 0.400$ and $l_{12} = 10.008 \simeq 10$

Now for obtaining ψ_1^2 and ψ_2^2 we can proceed as follows:

$$\Sigma = \begin{pmatrix} l_{11}^2 & l_{11}l_{21} \\ l_{21}l_{11} & l_{22}^2 \end{pmatrix} + \begin{pmatrix} \psi_1^2 & 0 \\ 0 & \psi_2^2 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} l_{11}^2 & l_{11} l_{21} \\ l_{21} l_{11} & l_{22}^2 \end{pmatrix} + \begin{pmatrix} \psi_1^2 & 0 \\ 0 & \psi_2^2 \end{pmatrix}$$

∴ Comparing the above we can say that,

$$\sigma_{11} = l_{11}^2 + \psi_1^2 \Rightarrow \psi_1^2 = \sigma_{11} - l_{11}^2$$

$$\sigma_{22} = l_{21}^2 + \psi_2^2 \Rightarrow \psi_2^2 = \sigma_{22} - l_{21}^2$$

Hence we get, $\psi_1^2 = 1 - 0.400^2 = 0.84$

and, $\psi_2^2 = 100 - (10)^2 \simeq 0$

∴ The one factor model solution is,

$$x_1 = 0.4 F_1 + \epsilon_1$$

$$x_2 = 10 F_1 + \epsilon_2$$

3) Consider the distance matrix given below. Perform single, complete and average linkage clustering on the distance matrix.

D =	A	B	C	D	E
A	1	0.8	0.4	0.2	0.1
B	0.8	1	0.6	0.3	0.2
C	0.4	0.6	1	0.7	0.5
D	0.2	0.3	0.7	1	0.9
E	0.1	0.2	0.5	0.9	1

ans) We follow the given algorithm to perform clustering

(1) Start with n clusters each containing 1 item
(2) Club 2 most close clusters into one cluster. At this step there would be atmost (n-1) clusters
(3) Recompute or update the distances between the cluster
(4) Repeat step 2 and 3 until one single cluster is obtained

☒ Single clustering: we consider smallest distance i.e $d(UV)W = \min\{d_{UW}, d_{VW}\}$

There are 5 clusters A, B, C, D, E. Distance between A and E is minimum so they will be clubbed. The new cluster is,

	AE	B	C	D
AE	1			
B	(0.2)	1		
C	0.4	0.6	1	
D	0.2	0.3	0.7	1

$\min\{d(A,B), d(E,B)\} = \min\{0.8, 0.2\} = 0.2$

$\min\{d(A,C), d(E,C)\} = \min\{0.4, 0.5\} = 0.4$

$\min\{d(A,D), d(E,D)\} = \min\{0.2, 0.9\} = 0.2$

Now distance between AE and B is minimum so they are clubbed together.

	ABE	C	D
ABE	1		
C	0.4	1	
D	(0.2)	0.7	1

$\min\{d(AC), d(B,C), d(E,C)\} = 0.4$

$\min\{d(A,D), d(B,D), d(E,D)\} = 0.2$

Now distance between ABE and D is minimum

	ABED	C
ABED	1	
C	0.4	1

$\min\{d(AC), d(BC), d(D,C), d(EC)\} = 0.4$

Finally C will be clubbed with ABED and the process ends with one cluster.

- Complete linkage: We consider maximum distance here i.e $d(UV)_W = \max\{d_{UW}, d_{VW}\}$

Here distance between AE is minimum so they will be merged.

	AE	B	C	D
AE	1			
B	0.8	1		
C	0.5	0.6	1	
D	0.9	(0.3)	0.7	1

$\max\{d(A,B), d(E,B)\} = \max\{0.8, 0.2\} = 0.8$
$\max\{AC, EC\} = 0.5$
$\max\{AD, ED\} = 0.9$

Now distance between B and D is minimum so they are merged

	AE	BD	C
AE	1		
BD	0.9	1	
C	(0.5)	0.7	1

$\max\{B(AE), D(AE)\} = \max\{0.8, 0.9\} = 0.9$
$\max\{AC, EC\} = \max\{0.4, 0.5\} = 0.5$
$\max\{BC, D,C\} = \max\{0.6, 0.7\} = 0.7$

Now distance between C and AE is minimum so they are merged.

	AEC	BD
AEC	1	
BD	0.9	1

$\max\{B(AEC), D(AEC\} = \max\{0.9, 0.7\} = 0.9$

Finally BD will be merged with AEC and the process ends here with one cluster.

Average linkage: We take average distance between the cluster i.e

$$d_{(UV)W} = \sum_i \sum_k d_{ik} / N_{(UV)} N_W$$

Here distance between A and E is minimum so they are merged

	AE	B	C	D
AE	1			
B	0.5	1		
C	0.45	0.6	1	
D	0.55	0.3	0.7	1

$\text{avg}\{AB, EB\} = 0.5$
$\text{avg}\{AC, EC\} = 0.45$
$\text{avg}\{AD, ED\} = 0.55$

Now distance between B and D is minimum so they are merged together

	AE	BD	C
AE	1		
BD	0.95	1	
C	0.45	0.65	1

$\text{avg}\{B(AE), D(AE)\} = 0.85$
$\text{avg}\{AC, ED\} = 0.45$
$\text{avg}\{BC, DC\} = 0.65$

Now distance between AE and C is minimum so they are merged.

	AEC	BD
AEC	1	
BD	0.8	1

$\text{avg}\{B(AED), D(AED)\} = 0.8$

Finally BD will be merged with AED and the process stops with one cluster

2) Let there be 5 objects A B C D E where each of them has the coordinate as follows:

A(2,3), B(5,1), C(1,1), D(6,7), E(2,3)

Also let 2 clusters to be formed using K means method.

ans) Let us make 2 arbitary clusters C_1 and C_2 with the following members,

$C_1 = (A, B, C)$ and $C_2 = (D, E)$

Centroid of $C_1 = \left(\frac{2+5+2}{3}, \frac{3+1+2}{3}\right) = (3, 2)$

Centrod of $C_2 = \left(\frac{6+2}{2}, \frac{7+3}{2}\right) = (4, 5)$

Now, distance of each obs from C_1 and C_2 are to be computed,

$d(A, C_1) = [(3-2)^2 + (2-3)^2]^{1/2} = \sqrt{2}$
$d(A, C_2) = [(2-4)^2 + (3-5)^2]^{1/2} = \sqrt{8}$
$\therefore$ A stays in C_1

$d(B, C_1) = [(5-3)^2 + (1-2)^2]^{1/2} = \sqrt{5}$
$d(B, C_2) = [(5-1)^2 + (1-5)^2]^{1/2} = \sqrt{17}$
$\therefore$ B stays in C_1

$d(C, C_1) = [(3-2)^2 + (2-1)^2]^{1/2} = \sqrt{1}$
$d(C, C_2) = [(4-1)^2 + (5-2)^2]^{1/2} = \sqrt{13}$
$\therefore$ C stays in C_1

$d(D, C_1) = [(6-3)^2 + (7-2)^2]^{1/2} = \sqrt{34}$
$d(D, C_2) = [(6-4)^2 + (5-7)^2]^{1/2} = \sqrt{8}$
$\therefore$ D stays in C_2

$d(E, C_1) = [(2-3)^2 + (2-3)^2]^{1/2} = \sqrt{2}$
$d(E, C_2) = [(2-4)^2 + (3-5)^2]^{1/2} = \sqrt{8}$
$\therefore$ E goes to C_1 as $d(E, C_1)$ is less than $d(E, C_2)$

Now the clusters are $C_1 = (A, B, C, E)$ and $C_2 = D$

i.e E is reassigned to C_1 and others stay in theirs earlier clusters. Again the cluster centroid is computed.

∴ The new centroid of $C_1 = \left(\frac{1}{4}(2+5+2+2), \frac{1}{4}(3+1+2+3)\right)$

$= (2.75, 2.25)$

and centroid of $C_2 = (6,7)$

Again the distance of each obs from C_1 and C_2 is computed

$d(A,C_1) = [(2.75-2)^2 + (2.25-3)^2]^{1/2} = 1.06$

$d(A,C_2) = [(2-6)^2 + (3-7)^2]^{1/2} = 5.65$

$d(B,C_1) = [(5-2.75)^2 + (2.25-1)^2]^{1/2} = 2.57$

$d(B,C_2) = [(5-6)^2 + (1-7)^2]^{1/2} = 6.08$

$d(C,C_1) = [(2-2.75)^2 + (2-2.25)^2]^{1/2} = 0.79$

$d(C,C_2) = [(2-6)^2 + (2-7)^2]^{1/2} = 6.40$

$d(D,C_1) = [(6-2.75)^2 + (7-2.25)^2]^{1/2} = 5.75$

$d(D,C_2) = [(6-6)^2 + (7-7)^2]^{1/2} = 0$

$d(E,C_1) = [(2-2.75)^2 + (3-2.25)^2]^{1/2} = 1.06$

$d(E,C_2) = [(2-6)^2 + (3-7)^2]^{1/2} = 5.65$

Now it is observed in the latest composition of C_1 and C_2 that each object is at the nearest centroid of the clusters. Hence the final 2 clusters made from the above 5 elements are as follows.

$C_1 = (A, B, C, E)$, $C_2 = (D)$

1) Consider the var-cov matrix given below and find out the canonical correlation.

$$R = \begin{pmatrix} 1 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1 & 0.3 & 0.4 \\ 0.5 & 0.3 & 1 & 0.2 \\ 0.6 & 0.4 & 0.2 & 1 \end{pmatrix}_{4\times 4}$$

ans) From the given matrix we can say that

$$\Sigma_{11} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix} \qquad \Sigma_{22} = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$$

$$\Sigma_{12} = \begin{pmatrix} 0.5 & 0.6 \\ 0.3 & 0.4 \end{pmatrix} \qquad \Sigma_{21} = \begin{pmatrix} 0.5 & 0.3 \\ 0.5 & 0.4 \end{pmatrix}$$

Our objective is to find canonical correlation between the variables $\underset{\sim}{x}^{(1)}$ and $\underset{\sim}{x}^{(2)}$

where $\underset{\sim}{x}_{(1)} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ $\underset{\sim}{x}_{(2)} = \begin{pmatrix} x_3 \\ x_4 \end{pmatrix}$ and $\underset{\sim}{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$

We take linear combination such as, $U = \underset{\sim}{a}' \underset{\sim}{x}^{(1)}$ and $V = \underset{\sim}{b}' \underset{\sim}{x}_{(2)}$ such that $\underset{\sim}{a}$ and $\underset{\sim}{b}$ are so chosen that correlation coefficient between U and V is maximum

The canonical correlation can be found by calculating the eigenvalue of the matrix,

$$\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \quad \text{and} \quad \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$

Now, $\Sigma_{22}^{-1} = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}^{-1}$, Now $\det \Sigma_{22} = 1 - 0.2^2 = 0.96$

$$\Sigma_{22}^{-1} = \frac{1}{|\Sigma_{22}|} \operatorname{adj} \Sigma = \begin{pmatrix} 1.04 & -0.208 \\ -0.208 & 1.04 \end{pmatrix}$$

similarly $\Sigma_{11}^{-1} = \frac{1}{|\Sigma_{11}|} \text{adj}\,|\Sigma_{11}|$ $\quad$ det $|\Sigma_{11}| = 1 - 0.4^2$

$= 0.84$

$$= \begin{pmatrix} 1.19 & -0.476 \\ -0.476 & 1.19 \end{pmatrix}$$

$$\therefore \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$$

$$= \begin{pmatrix} 1.04 & -0.208 \\ -0.208 & 1.04 \end{pmatrix} \begin{pmatrix} 0.5 & 0.3 \\ 0.6 & 0.4 \end{pmatrix} \begin{pmatrix} 1.19 & -0.476 \\ -0.476 & 1.19 \end{pmatrix} \begin{pmatrix} 0.5 & 0.6 \\ 0.3 & 0.4 \end{pmatrix}$$

$$= \begin{pmatrix} 247/625 & 143/625 \\ 13/25 & 221/625 \end{pmatrix} \begin{pmatrix} 1.19 & -0.476 \\ -0.476 & 1.19 \end{pmatrix} \begin{pmatrix} 0.5 & 0.6 \\ 0.3 & 0.4 \end{pmatrix}$$

$$= \begin{pmatrix} 0.361 & 0.084 \\ 0.450 & 0.173 \end{pmatrix} \begin{pmatrix} 0.5 & 0.6 \\ 0.3 & 0.4 \end{pmatrix}$$

$$= \begin{pmatrix} 0.2057 & 0.2502 \\ 0.2769 & 0.3392 \end{pmatrix}$$

Now the eigenvalues are given by,

$|\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \lambda^2 I| = 0$ $\quad$ [Consider λ^2 as λ]

$$\Rightarrow \begin{vmatrix} 0.2057 - \lambda & 0.2502 \\ 0.2769 & 0.3392 - \lambda \end{vmatrix} = 0 \Rightarrow (0.2507 - \lambda)(0.3392 - \lambda) - (0.2502 \times 0.2769)$$

$$\Rightarrow 0.0850 - 0.2507\lambda - 0.3397\lambda + \lambda^2 - 0.069 = 0$$

$$\Rightarrow 0.0157 - 0.5899\lambda + \lambda^2 = 0$$

$$\therefore \lambda = \frac{0.5899 \pm \sqrt{0.5899^2 - (4 \times 0.0157)}}{2}$$

$\therefore \lambda_1^2 = 0.0009$, $\lambda_2^2 = 0.545$

$\Rightarrow \lambda_1 = 0.03$, $\lambda_2 = 0.738$

Hence the largest eigenvalue $\lambda_2 = 0.738$ is the canonical correlation between the variables U and V.

2) Consider the data given below. Test equality of mean vectors of all the groups

(24, 105)	(15, 203)	(2, 215)
(10, 210)	(11, 210)	(7, 194)
(5, 117)	(9, 234)	(6, 201)
(30, 200)	(5, 211)	(5, 198)
(15, 203)	(10, 200)	(3, 207)
↓ G3	↓ G2	↓ G1

ans) Our objective is to test $H_0: \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \underset{\sim}{\mu}_3$, where

$\underset{\sim}{\mu}_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix}$; $i = 1, 2, 3$

The model is, $\underset{\sim}{x}_{ij} = \underset{\sim}{\mu} + \underset{\sim}{\tau}_i + \underset{\sim}{\varepsilon}_{ij}$; $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$

where, $\underset{\sim}{\varepsilon}_{ij} \sim N(\underset{\sim}{0}, \Sigma)$ and $\sum_i n_i \underset{\sim}{\tau}_i = 0$

$\underset{\sim}{\mu}$ is the overall mean effect

$\underset{\sim}{\tau}_i$ is the ith treatment effect

The overall model can be written as,

$$\underset{\sim}{x}_{ij} = \underset{\sim}{\bar{x}} + (\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}}) + (\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i)$$

↓ Obs; ↓ overall sample mean; ↓ estimated treatment effect; ↓ error

Now following the above idea we have,

$\underset{\sim}{\bar{x}}_1 = \begin{bmatrix} 16 \\ 146.6 \end{bmatrix}$ $\underset{\sim}{\bar{x}}_2 = \begin{bmatrix} 10 \\ 210 \end{bmatrix}$ $\underset{\sim}{\bar{x}}_3 = \begin{bmatrix} 5.8 \\ 201.4 \end{bmatrix}$

and $\underset{\sim}{\bar{x}} = \begin{bmatrix} 10.6 \\ 186 \end{bmatrix}$

Consider the 1ST obs (1ST variable) in the pair of variables. We can write them as,

$$\begin{pmatrix} 20 & 10 & 5 & 30 & 15 \\ 15 & 11 & 9 & 5 & 10 \\ 8 & 7 & 6 & 5 & 3 \end{pmatrix} = \begin{pmatrix} 10.6 & 10.6 & 10.6 & 10.6 & 10.6 \\ 10.6 & 10.6 & 10.6 & 10.6 & 10.6 \\ 10.6 & 10.6 & 10.6 & 10.6 & 10.6 \end{pmatrix} +$$

$$\begin{pmatrix} 4 & -6 & -11 & 14 & -1 \\ 5 & 1 & -1 & -5 & 0 \\ 2.2 & 1.2 & 0.2 & -0.8 & -2.8 \end{pmatrix} + \begin{pmatrix} 5.4 & 5.4 & 5.4 & 5.4 & 5.4 \\ -0.6 & -0.6 & -0.6 & -0.6 & -0.6 \\ -4.8 & -4.8 & -4.8 & -4.8 & -4.8 \end{pmatrix}$$

$\Rightarrow$ observation = mean + residual + treatment effect

i.e $SS_{obs} = SS_{mean} + SS_{Res} + SS_{trt}$

Now $SS_{obs} = 20^2 + 10^2 + 5^2 + 30^2 + \dots + 5^2 + 3^2 = 2385$

$SS_{mean} = 15 \times 10.6^2 = 1685.4$

$SS_{trt} = (5 \times 5.4^2) + (5 \times -0.6^2) + (5 \times -4.8^2) = 262.8$

$SS_{res} = 4^2 + 6^2 + 11^2 + \dots + 0.8^2 + 2.8^2 = 436.8$

Similarly for the 2nd variable we can write,

$$\begin{pmatrix} 105 & 110 & 115 & 200 & 203 \\ 205 & 210 & 220 & 215 & 200 \\ 215 & 190 & 201 & 198 & 203 \end{pmatrix} = \begin{pmatrix} 186 & 186 & 186 & 186 & 186 \\ 186 & 186 & 186 & 186 & 186 \\ 186 & 186 & 186 & 186 & 186 \end{pmatrix}$$

$$+ \begin{pmatrix} -41.6 & -36.6 & -31.6 & 53.4 & 56.4 \\ -5 & 0 & 10 & 5 & -10 \\ 13.6 & -11.4 & -0.4 & -3.4 & 1.6 \end{pmatrix}$$

$$+ \begin{pmatrix} -39.4 & -39.4 & -39.4 & -39.4 & -39.4 \\ 24 & 24 & 24 & 24 & 24 \\ 15.4 & 15.4 & 15.4 & 15.4 & 15.4 \end{pmatrix}$$

$\Rightarrow$ observation = mean + residual + treatment effect

i.e $SS_{obs} = SS_{mean} + SS_{res} + SS_{trt}$

Now, $SS_{Obs} = 105^2 + 110^2 + \cdots + 198^2 + 203^2 = 54144.8$

$SS_{mean} = 15 \times 186^2 = 518940$

$SS_{trt} = (5 \times -39.4^2) + (5 \times 24^2) + (5 \times 15.4^2) = 11827.6$

$SS_{Res} = 41.6^2 + 36.6^2 + 31.6^2 + \cdots + 3.4^2 + 1.6^2 = 10680.4$

Now proceeding row by row for the 2 variables the cross product are computed as below.

$SSP_{Mean} = 15 \times (186 \times 10.6) = 29574$

$SSP_{trt} = 5(39.4 \times 5.4) + 5(0.6 \times 24) + 5(4.8 \times 15.4) = 1505.4$

$SSP_{Res} = (4 \times -39.4) + (-6 \times -36.6) + \cdots + (-2.8 \times 1.6) = 1046.4$

Total $SP = (20 \times 105) + (10 \times 110) + \cdots + (3 \times 203) = 29115$

Total (corrected) cross product = Total $SP - SSP_{mean}$

$= 29115 - 29574 = -459$

The Manova Table is given by,

MANOVA TABLE

Source	Matrix of SS and CP	d.f
treatment	$\begin{bmatrix} 262.8 & 1505.4 \\ 1505.4 & 11827.6 \end{bmatrix}$	2
residual	$\begin{bmatrix} 436.8 & 1046.4 \\ 1046.4 & 10680.4 \end{bmatrix}$	6
Total	$\begin{bmatrix} 699.6 & 2551.8 \\ 2551.8 & 22508 \end{bmatrix}$	8

Therefore Wilk's Lambda is,

$$\Lambda = \frac{|WSS|}{|WSS + BSS|}$$

$$\Lambda = \frac{\begin{vmatrix} 436.8 & 1046.4 \\ 1046.4 & 10680.4 \end{vmatrix}}{\begin{vmatrix} 699.6 & 2551.8 \\ 2551.8 & 22508 \end{vmatrix}} = \frac{3570245.76}{9234913.56} = 0.387$$

Now from the distribution of Wilk's Lambda we have,

No of variables = 2, No of groups = 3

Hence, $\left(\frac{\sum_{i=1}^{k} n_i - k - 1}{k-1}\right)\left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right)$

$$= \left(\frac{9-3-1}{3-1}\right)\left(\frac{1-\sqrt{0.387}}{\sqrt{0.387}}\right)$$

$$= 1.519 \sim F\left(2(k-1), 2\left(\sum_{i=1}^{k} n_i - k - 1\right)\right) = F(4, 10)$$

Now we can see that,

$$F_{obs} = 1.519 < F_{0.05}(4, 10) = 3.48$$

Therefore, we accept the null hypothesis that there is no difference between the groups w.r.t to the mean vectors.