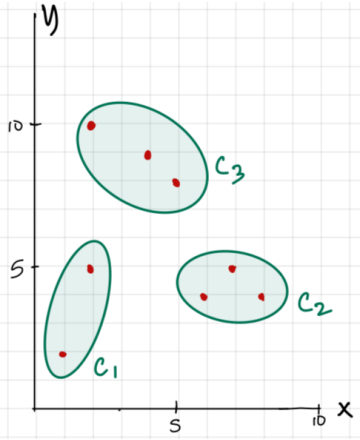
1a.

1a.)



3 clusters are required.

Cluster 1 (C1): {(2,4), (1,2)}

Cluster 2 (C2): {(6,4), (8,4), (7,5)}

Cluster 3 (C3): {(2,10), (4,9), (5,8)}

1b.

1b.) 3 clusters w/ centers {(2,5), (5,8), (4,9)}

Iteration 1:

- Assign points to nearest centroid:

  Cluster 1: (2,5), (1,2)

  Cluster 2: (5,8), (7,5), (6,4)

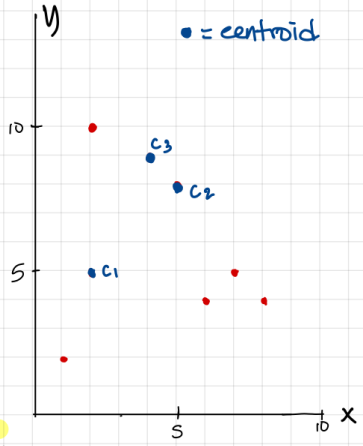
  Cluster 3: (2,10), (4,9), (8,4)

- Update centroids:

  Centroid 1: ((2+1)/2, (5+2)/2) = (1.5, 3.5)

  Centroid 2: ((5+7+6)/3, (8+5+4)/3) = (6, 5.67)

  Centroid 3: ((2+4+8)/3, (10+9+4)/3) = (4.7, 7.7)



1c.

1c.) Iteration 2:

- Assign points to nearest centroid:

  Cluster 1: (2,5), (1,2), (6,4), (7,5)

  Cluster 2: (5,8), (7,5), (6,4)

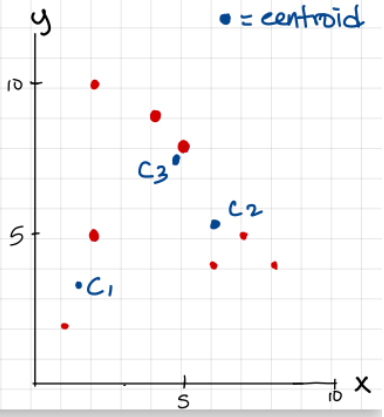
  Cluster 3: (2,10), (4,9), (8,4)

- Update centroids:

  Centroid 1: ((2+1+6+7)/4, (5+2+4+5)/4) = (4,4)

  Centroid 2: (6, 5.67)

  Centroid 3: (4.7, 7.7)



1d.

1d.) Iteration 3

The centroids will not change anymore, so we know that the algorithm has converged

Centroid 1: (4,4)
Centroid 2: (6, 5.67)
Centroid 3: (4.7, 7.7)

1e.

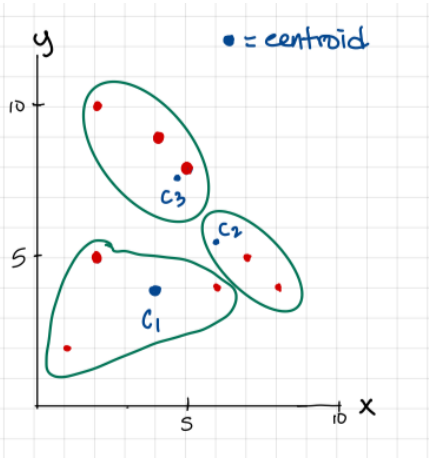
1e.) Cluster 1 was most the same as in part a.), after the 1st iteration, but had 2 points added to it after iteration 2.
Cluster 2 was changed only in iteration 1 after (8,4) was taken from it and placed in Cluster 3.
Cluster 3 was changed only in iteration 1 after (5,8) was taken from it and placed in Cluster 2.

1f.

1f.) 2 iterations are required for the clusters to converge

1g.

1g.)



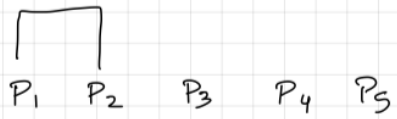
2.

2.) **Single Link (MIN)**

	P1	P2	P3	P4	P5
P1	1	0.10	0.41	0.55	0.35
P2	0.10	1	.64	.47	.98
P3	0.41	.64	1	.44	.85
P4	0.55	.47	.44	1	.76
P5	0.35	.98	.85	.76	1

Merge points with least similarity

	P1 ∪ P2	P3	P4	P5
P1 ∪ P2	1	0.41	0.47	0.35
P3	0.41	1	0.44	.85
P4	0.47	.44	1	.76
P5	0.35	.85	.76	1



	P1 ∪ P2 ∪ P5	P3	P4
P1 ∪ P2 ∪ P5	1	0.41	0.47
P3	0.41	1	0.44
P4	0.47		1

	P1 ∪ P2 ∪ P5 ∪ P3	P4
P1 ∪ P2 ∪ P5 ∪ P3	1	0.44
P4	0.44	1

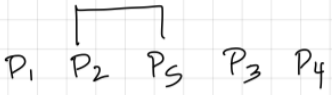
Final Dendrogram:

Complete Linkage (MAX)

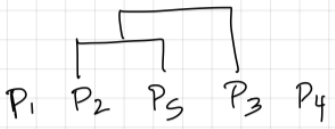
	P1	P2	P3	P4	P5
P1	1	0.10	0.41	0.55	0.35
P2	0.10	1	.64	.47	.98
P3	0.41	.64	1	.44	.85
P4	0.55	.47	.44	1	.76
P5	0.35	.98	.85	.76	1

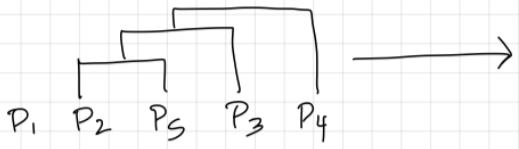
Merge points with the most similarity

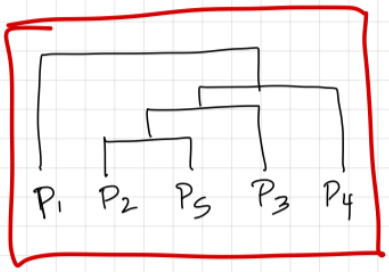
	P1	P2 U P5	P3	P4
P1	1	0.35	0.41	0.55
P2 U P5	0.35	1	0.85	0.76
P3	0.41	0.85	1	.44
P4	.55	0.76	.44	1



	P1	P2 U P5 UP3	P4
P1	1	0.41	0.55
P2 U P5 UP3	0.41	1	0.76
P4	0.55	0.76	1







3a.

3a.) User - User Collaborative Filtering

similarity measure = pearson
neighborhood size = |N| = 3

Customers

Products	1	2	3	4	5	6	7	8	9	10
1	5			8	5	4				5
2	8	7		5	?		7	5	5	
3		1			3	2			3	3
4		5	6				4	9		2
5	9		5	7		3			4	
6		6		4	6			5		4

Pearson similarity:

$$sim(x,y) = \frac{\sum_{ses_{xy}} (r_{xs} - \bar{r}_x)(r_{ys} - \bar{r}_y)}{\sqrt{\sum_{ses_{xy}} (r_{xs} - \bar{r}_x)^2} \cdot \sqrt{\sum_{ses_{xy}} (r_{ys} - \bar{r}_y)^2}}$$

S_{xy} = items rated by users x & y

1.) Find mean ratings for each user:

$\overline{user\ 1} = (5+8+9)/3 = 7.3$
$\overline{user\ 2} = (7+1+5+6)/4 = 4.75$
$\overline{user\ 3} = (6+5)/2 = 5.5$
$\overline{user\ 4} = (8+5+7+4)/4 = 6$
$\overline{user\ 5} = (5+3+6)/3 = 4.7$
$\overline{user\ 6} = (4+2+3)/3 = 3$
$\overline{user\ 7} = (7+4)/2 = 5.5$
$\overline{user\ 8} = (5+9+5)/3 = 6.3$
$\overline{user\ 9} = (5+3+4)/3 = 4$
$\overline{user\ 10} = (5+3+2+4)/4 = 3.5$

2.) Find similarity between us and other users

User 5 & 1 → S_{xy} = Product 1

$$sim(5,1) = \frac{(5-4.7)(5-7.3)}{\sqrt{(5-4.7)^2} \cdot \sqrt{(5-7.3)^2}} = \frac{-0.69}{0.69} = -1$$

User 5 & 2 → S_{xy} = Product 3 & 6

$$sim(5,2) = \frac{(3-4.7)(1-4.75) + (6-4.7)(6-4.75)}{\sqrt{(3-4.7)^2 + (6-4.7)^2} \cdot \sqrt{(1-4.75)^2 + (6-4.75)^2}}$$

$$= \frac{6.375 + 1.625}{8.46} = 0.946$$

User 5 & 3 → S_{xy} = no products that are rated by both of them.

$sim(5,3) = 0$

User 5 & 4 → S_{xy} = Products 1 & 6

$$sim(5,4) = \frac{(5-4.7)(8-6) + (6-4.7)(4-6)}{\sqrt{(5-4.7)^2 + (6-4.7)^2} \cdot \sqrt{(8-6)^2 + (4-6)^2}} = -0.53$$

User 5 & 6 → S_{xy} → Products 1 & 3

$$sim(5,6) = \frac{(5-4.7)(4-3) + (3-4.7)(2-3)}{\sqrt{(5-4.7)^2 + (3-4.7)^2} \cdot \sqrt{(4-3)^2 + (2-3)^2}} = 0.82$$

$sim(5,7) = 0$ → no products are shared by them

User 5,8 → S_{xy} → Product 6

$$sim(5,8) = \frac{(6-4.7)(5-6.3)}{\sqrt{(6-4.7)^2}\cdot\sqrt{(5-6.3)^2}} = -1$$

user 5,9 → S_{xy} → Product 3

$$sim(5,9) = \frac{(3-4.7)(3-4)}{\sqrt{(3-4.7)^2}\cdot\sqrt{(3-4)^2}} = 1$$

user 5, 10 → S_{xy} → Product 1, 3, 6

$$sim(5,10) = \frac{(5-4.7)(5-3.5)+(3-4.7)(3-3.5)+(6-4.7)(4-3.5)}{\sqrt{(5-4.7)^2+(3-4.7)^2+(6-4.7)^2}\cdot\sqrt{(5-3.5)^2+(3-3.5)^2+(4-3.5)^2}}$$

$$= \frac{1.95}{3.58} = 0.54$$

Top 3 users with the highest absolute correlation coefficients: users 1, 8, 9 (-1, -1, 1)

Prediction = $r_{xi} = \frac{\sum_{y \in N} S_{xy} \cdot r_{yi}}{\sum_{y \in N} S_{xy}} = \frac{-1(8)+(-1)(5)+1(5)}{-1+(-1)+1}$

S_{xy} = similarities

$$= \frac{-8-5+5}{-1} = 8$$

predicted rating of user 5 on product 2 = 8

3b.

3b.) Item – Item Collab. Filtering, $|N| = 2$, cosine similarity $= \frac{sim(x,y)}{\sqrt{\sum_i r_{xi}^2} \cdot \sqrt{\sum_i r_{yi}^2}}$

* treat unknowns as zeroes

Find similarities between product 2 and other products.

$sim(2,1) = \frac{8 \cdot 5 + 5 \cdot 8}{\sqrt{8^2+5^2} \cdot \sqrt{5^2+8^2}} = \frac{40+40}{89} = \frac{80}{89} = 0.89$

$sim(2,3) = \frac{7 \cdot 1 + 5 \cdot 3}{\sqrt{7^2+5^2} \cdot \sqrt{1^2+3^2}} = 0.809$

$sim(2,4) = \frac{7 \cdot 5 + 7 \cdot 4 + 5 \cdot 9}{\sqrt{7^2+7^2+5^2} \cdot \sqrt{5^2+4^2+9^2}} = 0.881$

$sim(2,5) = \frac{8 \cdot 9 + 5 \cdot 7 + 5 \cdot 4}{\sqrt{8^2+5^2+5^2} \cdot \sqrt{9^2+7^2+4^2}} = 0.984$

$sim(2,6) = \frac{7 \cdot 6 + 5 \cdot 4 + 5 \cdot 5}{\sqrt{7^2+5^2+5^2} \cdot \sqrt{6^2+4^2+5^2}} = 0.996$

$|N|$ = products 5, 6

$rating = \frac{0.984(0) + .996(6)}{0.984 + .996} = 3.018 \approx 3$

4a.

```
Optimal number of clusters (K): 2
```

4c.

```
WSSSE for K=2: 2338.7528589985814
WSSSE for K=3: 1956.2269857452052
WSSSE for K=4: 1695.2531197975961
WSSSE for K=5: 1518.2228986203206
WSSSE for K=6: 1450.3193373498814
WSSSE for K=7: 1271.9487840634365
WSSSE for K=8: 1247.457223452873
WSSSE for K=9: 1238.9891012552953
WSSSE for K=10: 1129.940404898411
```

The smallest WSSSE is when K = 10.

5.

```
Mean Squared Error (MSE): 0.8294242729268162
```

6c.

	precision	recall	f1-score	support
0	0.86	0.86	0.86	2495
1	0.86	0.87	0.86	2505
accuracy			0.86	5000
macro avg	0.86	0.86	0.86	5000
weighted avg	0.86	0.86	0.86	5000

86% accuracy - this is above average