(a)

x : (5, 3, 7, 9)
y : (7, 3, 6, 7)

i. The hamming distance between x and y:
diff (a,b) returns the minimum number of *substitutions* required to change a into b.
d = diff((5, 3, 7, 9), (7, 3, 6, 7)) = 3

ii. The Manhattan (L1) distance between two points x = (x1,…,xn) and y=(y1,…,yn) in *n*-dimensional space is the sum of the distances in each dimension.
L1 = |y1 - x1| + |y2 - x2| + |y3 - x3| + |y4 - x4| = |7 - 5| + |3 - 3| + |6 - 7| + |7 - 9| = 2 + 0 + 1 + 2 = 5

iii. Euclidean distance (L2) is a measure of the true straight line distance between two points in Euclidean space.
L2 = sqrt((y1 - x1)^2 + (y2 - x2)^2 + (y3 - x3)^2 + (y4 - x4)^2) = sqrt((7 - 5)^2 + (3 - 3)^2 + (6 - 7)^2 + (7 - 9)^2) = sqrt(4 + 0 + 1 + 4) = sqrt(9) = 3

(b)

Clustering using a distance measure such as Euclidean distance would be problematic due to the fact that there is a large difference in the scale of measurements between the height of a person and the length of the eyelashes. The height can be equal to thousands of millimetres, while the length of eyelashes is only tens. So when carrying out clustering using Euclidean distance, any difference between length of eyelashes in samples will be negligible compared to the differences in heights, and the clustering will reflect this.

To correct this, we can scale the data, normalising it so that differences in heights and length of eyelashes are treated equally.

(c)

	A	D	G
B	5	4.24	3.16
C	8.49	5	7.28
E	7.07	3.61	6.70
F	7.21	4.12	5.38
H	2.24	1.41	7.61

Point	Assign
A	A
B	G
C	D
D	D
E	D

F	D
G	G
H	D

A	x = 2/1 = 2	(2, 10)
	y = 10/1 = 10	
D	x = (8+7+6+4+5)/5 = 6	(6,6)
	y = (4+5+4+9+8)/5 = 6	
G	x = (2+1)/2 = 1.5	(1.5, 3.5)
	y = (5+2)/2 = 3.5	

Exercise 2

(a)

	Single	Complete	Average
Euclidean distance	29.3706%	44.4056%	30.0699%
Manhattan distance	29.3706%	44.4056%	29.7203%

The method which provides the best accuracy is using either of the distances with a single linkage type.

(b)

Seed number	Incorrectly clustered instances
1	34.2657%
2	31.1189%
3	37.4126%
4	30.4196%
5	24.4755%
6	29.3706%
7	48.951%
8	37.0629%
9	31.4685%
10	24.8252%

We find a large range of accuracies, with the worst seed incorrectly classifying roughy twice as many instances (49% compared with 24%). The fact that the accuracy depends so highly on the initial point used for clustering (which is affected by the generator seed) suggests that the method is not reliable.

(c)

It is possible to get 2 clusters when epsilon is in [1.5, 1.7] and MinPoints = 5 or minPoints = 6. However, these clusters have not the same size. For example, having epsilon = 1.7, minPoints = 5 we got:

0 245 (96%)
1 10 (4%)

So, we can conclude that DBSCAN doesn't work as we expected. DBSCAN is good to separate clusters of high density from clusters of low density. Given that DBSCAN is a density based clustering algorithm, it does a great job of seeking areas in the data that have a high density of observations, versus areas of the data that are not very dense with observations. However, it struggles with clusters of similar density.

(d)

The results of classifications and clustering are pretty much the same(<80%). So, both clustering and classification perform not very good in predicting class value in this dataset.

Exercise 3
(a)

Dealership, M5

(b)

Seed = 1
- Cluster 0 : M5 = 1, Purchase = 0
- Cluster 1 : Dealership = 0, Showroom = 1, ComputerSearch = 0, M5 = 0, 3Series = 1, Z4 = 1
- Cluster 2 : Showroom = 1, M5 = 0, 3Series = 1
- Cluster 3 : Dealership = 1, 3Series = 0
- Cluster 4 : M5 = 1, Financing = 1, Purchase = 1

Seed = 2
- Cluster 0 : ComputerSearch = 0, M5 = 1, Financing = 1
- Cluster 1 : Showroom = 1, 3Series = 1, Purchase = 0
- Cluster 2 : ComputerSearch = 1, Financing = 1
- Cluster 3 : Dealership = 0, Showroom = 1, M5 = 0, 3Series = 1, Financing = 1, Purchase = 1
- Cluster 4 : Dealership = 1, Financing = 0, Purchase = 0

Seed = 3
- Cluster 0 : Dealership = 0, Showroom = 1, M5 = 0, 3Series = 1, Purchase = 0
- Cluster 1 : Dealership = 0, Showroom = 1, 3Series = 1, Financing = 1
- Cluster 2 : M5 = 1, Z4 = 0, Financing = 1
- Cluster 3 : Dealership = 1, Financing = 0, Purchase = 0
- Cluster 4 : Dealership = 1, Financing = 1

We find quite a lot of consistency. The following attributes appear together in clusters:
Dealership = 0, Showroom = 1, 3Series = 1, Financing = 1.
Dealership = 0, Showroom = 1, M5 = 0, 3Series = 1
Showroom = 1, 3Series = 1, Purchase = 0
Dealership = 1, Financing = 0, Purchase = 0
M5 = 1, Financing = 1
Financing = 1, Purchase = 1

Several of these make sense intuitively; for example, Dealership = 1, Financing = 0, Purchase = 0 appears to represent people who visited the dealership only to look at the cars rather than intending to buy one.

(c)

Seed = 1
- Cluster 0 : Dealership = 0, Showroom = 1, M5 = 0, 3Series = 1
- Cluster 1 : Dealership = 1, Financing = 0, Purchase = 0
- Cluster 2 : Dealership = 1, Showroom = 1, ComputerSearch = 0
- Cluster 3 : ComputerSearch = 1, 3Series = 0, Financing = 1
- Cluster 4 : M5 = 1, Z4 = 0, Financing = 1, Purchase = 1

Seed = 2
- Cluster 0 : Showroom = 1, M5 = 0, Purchase = 0
- Cluster 1 : Dealership = 0, Showroom = 1, ComputerSearch = 0, 3Series = 1
- Cluster 2 : Showroom = 1, M5 = 1
- Cluster 3 : Showroom = 0, Financing = 0, Purchase = 0
- Cluster 4 : Dealership = 1, Showroom = 0, ComputerSearch = 1, Financing = 1

Seed = 3
- Cluster 0 : 3Series = 0
- Cluster 1 : Dealership = 0, Showroom = 1, 3Series = 1, Purchase = 0
- Cluster 2 : Financing = 1, Purchase = 1
- Cluster 3 : Dealership = 1, ComputerSearch = 1, Purchase = 0
- Cluster 4 : 3Series = 1, Financing = 1

Again we find several groups of attributes which regularly appears together in clusters, namely:

Dealership = 0, Showroom = 1, 3Series = 1
Dealership = 1, ComputerSearch = 1
ComputerSearch = 1, Financing = 1
3Series = 1, Financing = 1
Dealership = 1, Purchase = 0
Financing = 0, Purchase = 0
Showroom = 1, ComputerSearch = 0
Financing = 1, Purchase = 1
Showroom = 1, M5 = 0
Showroom = 1, Purchase = 0

Again we can explain some of these intuitively. For example, Dealership = 0, Showroom = 1, 3Series = 1 suggests that people who went to the showroom but not the dealership tended to look at the 3-series.

Comparing results of EM and K-means we see that there are some consistences:

Dealership = 0, Showroom = 1, 3Series = 1
3Series = 1, Financing = 1
Dealership = 1, Purchase = 0
Financing = 0, Purchase = 0
Financing = 1, Purchase = 1
Showroom = 1, M5 = 0
Showroom = 1, Purchase = 0