|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h |
| A | 5 | 4 |  | 5 | 2 |  | 3 | 2 |
| B |  | 3 | 4 | 4 | 2 | 2 | 1 |  |
| C | 3 |  | 1 | 4 |  | 4 | 5 | 3 |

Q 1)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h |
| A | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| B | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| C | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |

a)

(A, B) = 3/5

(A, C) = 1/2

(B, C) = 6/7

b)

(A, B)

= 39/root(83)\*root(50) = 0.605

(A, C)

= 56/root(83)\*root(76) = 0.705

(B, C)

= 33/ root(50)\*root(76) = 0.535

c)

A ∩ B = 2

A U B = 5

(A, B) = 1 - 2/5 = 3/5

A ∩ C = 3

A U C = 6

(A, C) = 1 – 3/6 = 1/2

B ∩ C = 1

B U C = 7

(B, C) = 1 - 1/7 = 6/7

d)

(A, B)

= (2/root(4)\*root(3)) = 0.577

(A, C)

= (3/root(4)\*root(5)) = 0.671

(B, C)

= (1/root(3)\*root(5)) = 0.258

e)

Avg of A = 3.5 / Avg of B = 2.67 / Avg of C = 3.33

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h |
| A | 1.5 | 0.5 |  | 1.5 | -1.5 |  | -0.5 | -1.5 |
| B |  | 0.33 | 1.33 | 1.33 | -0.67 | -0.67 | -1.67 |  |
| C | -0.33 |  | -2.33 | 0.67 |  | 0.67 | 1.67 | -0.33 |

f)

(A, B)

= 4.99/root(9.5)\*root(7.3334) = 0.5978

(A, C)

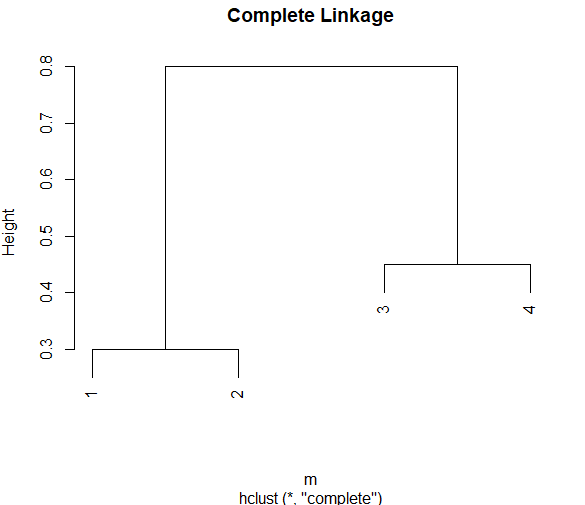
= 0.17/root(9.5)\*root(9.3334) = 0.0181

(B, C)

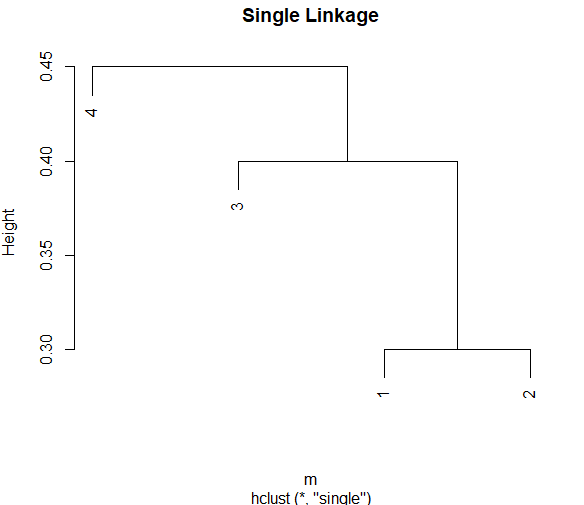
= -5.4456/root(7.3334)\*root(9.3334) = -0.6582

Q 2)

a)



b)



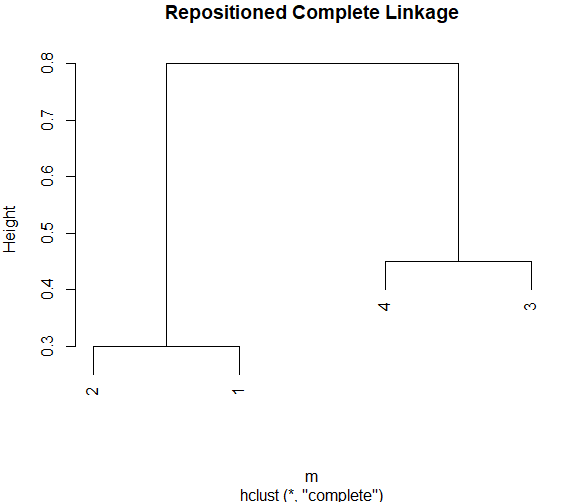
c)

For the first cluster dendrogram at the height of 0.8 when i=4, I obtained complete linkage of (1,2) and (3,4).

d)

For the second cluster at the height of 0.45 when i=4, I obtained single linkage of (4), (3,(1,2))

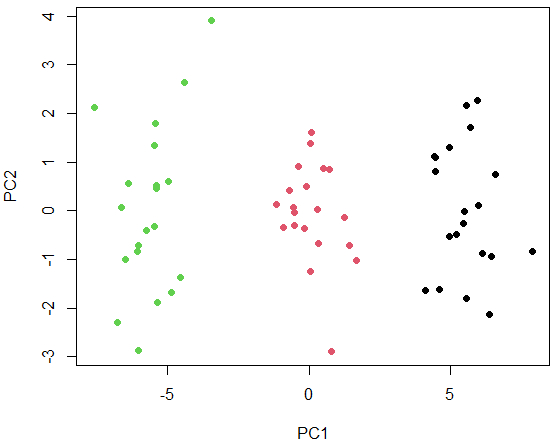
e)



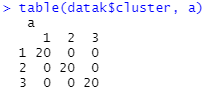
Keeping the dendrogram same, I flipped the numbers in the complete linkage.

Question 3

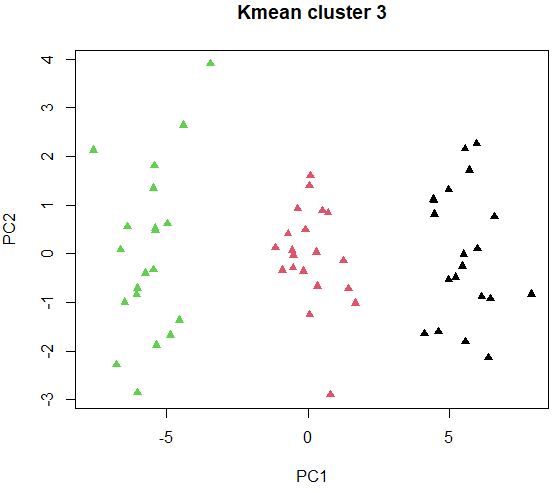
a)



b)

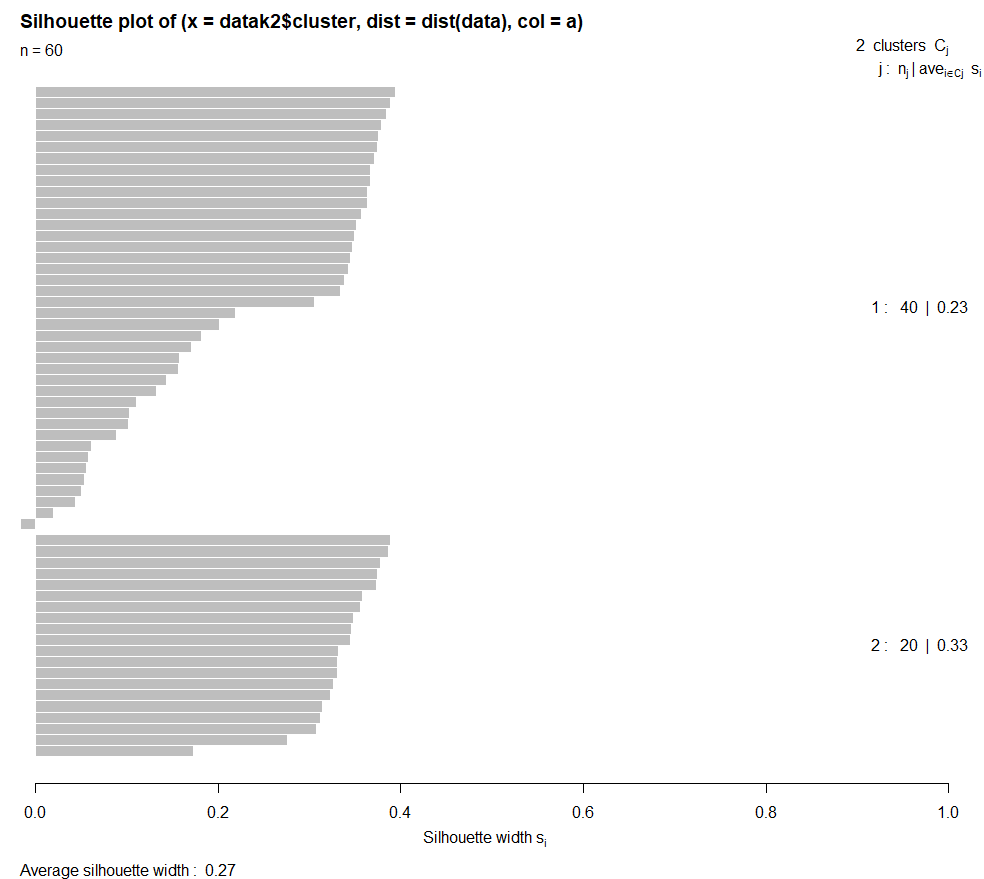
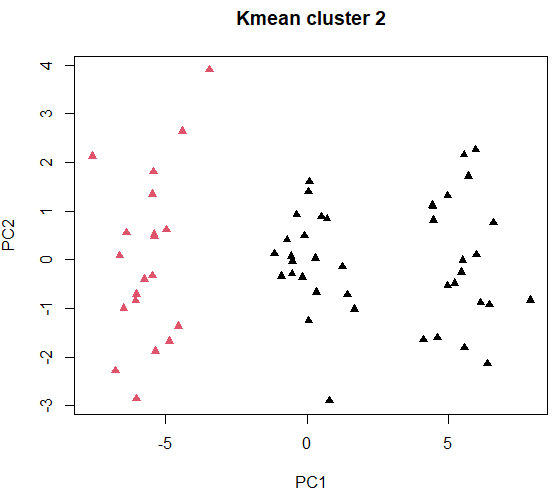


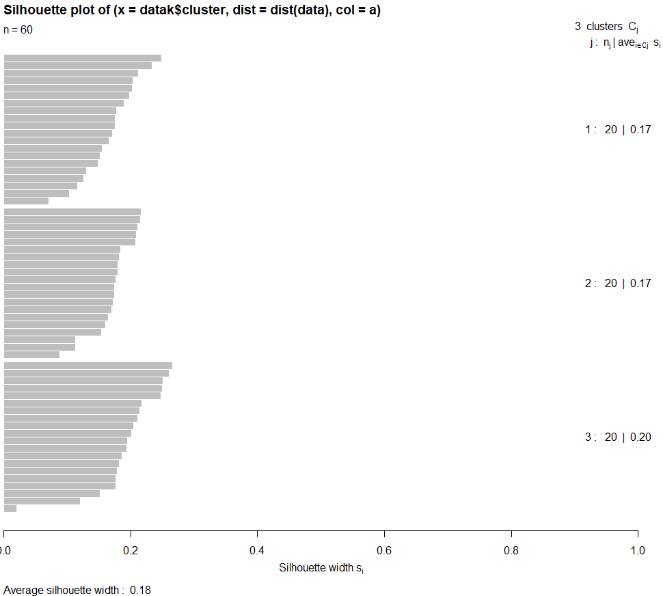
You can see that the data is evenly clustered into 20.



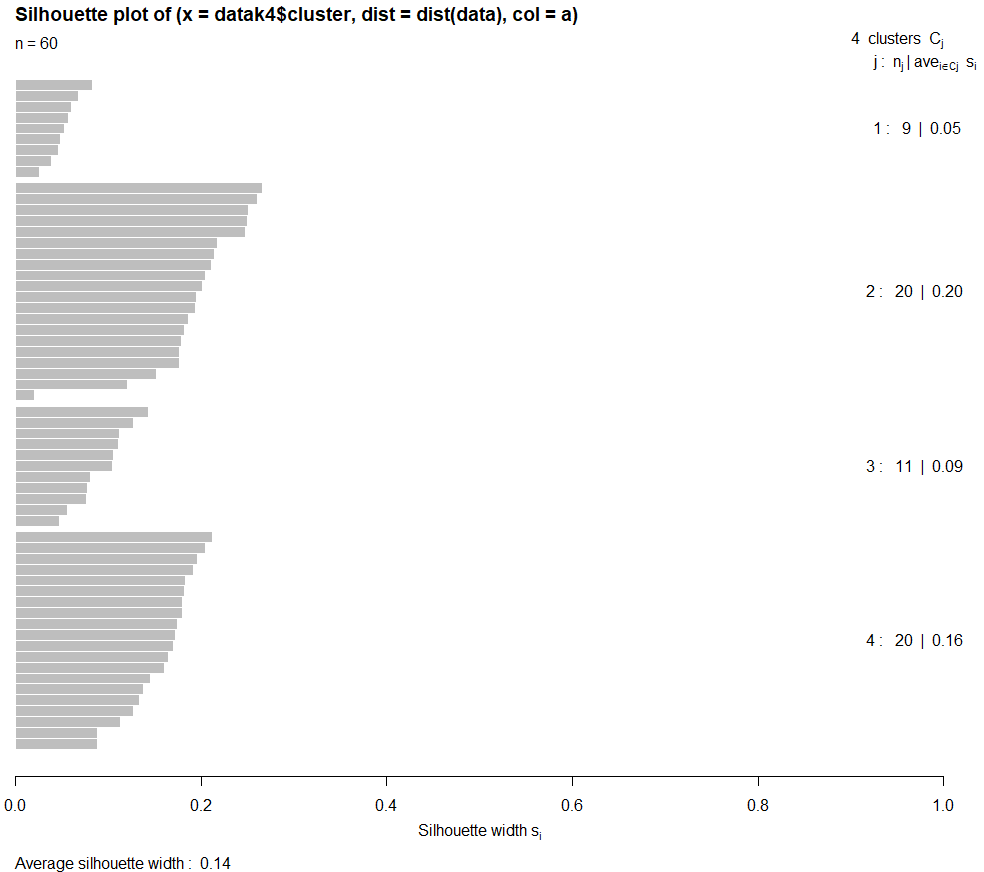
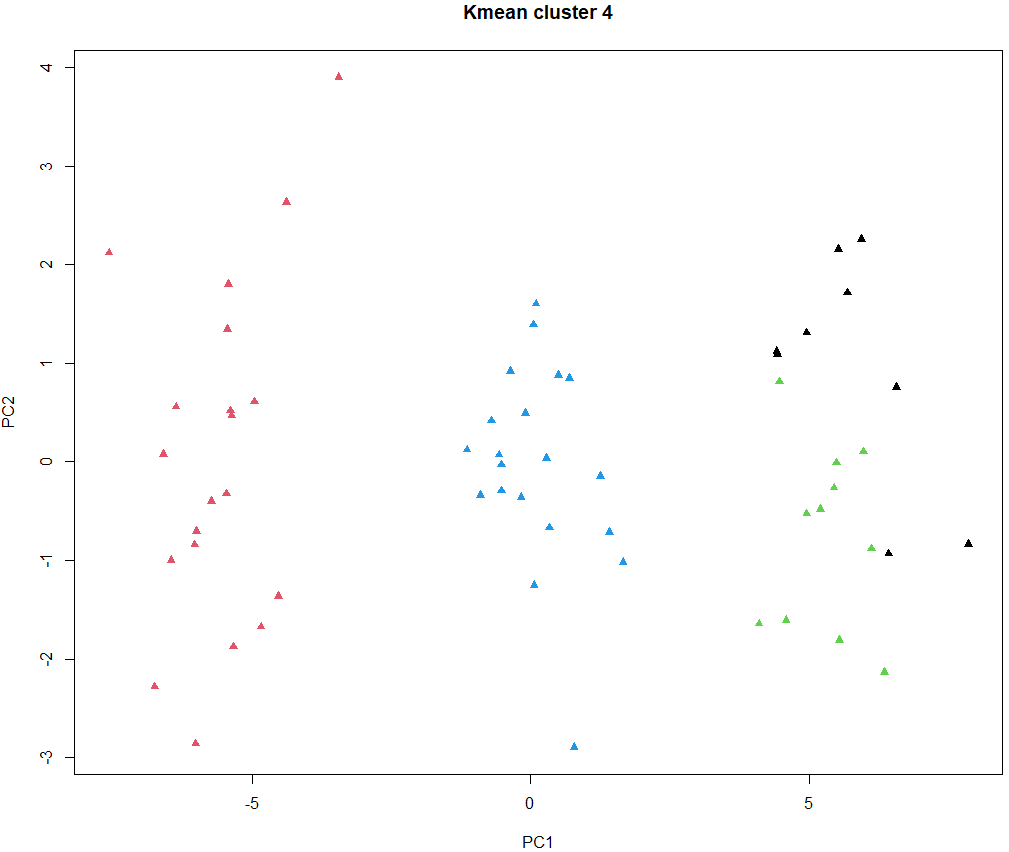
c)

K = 2

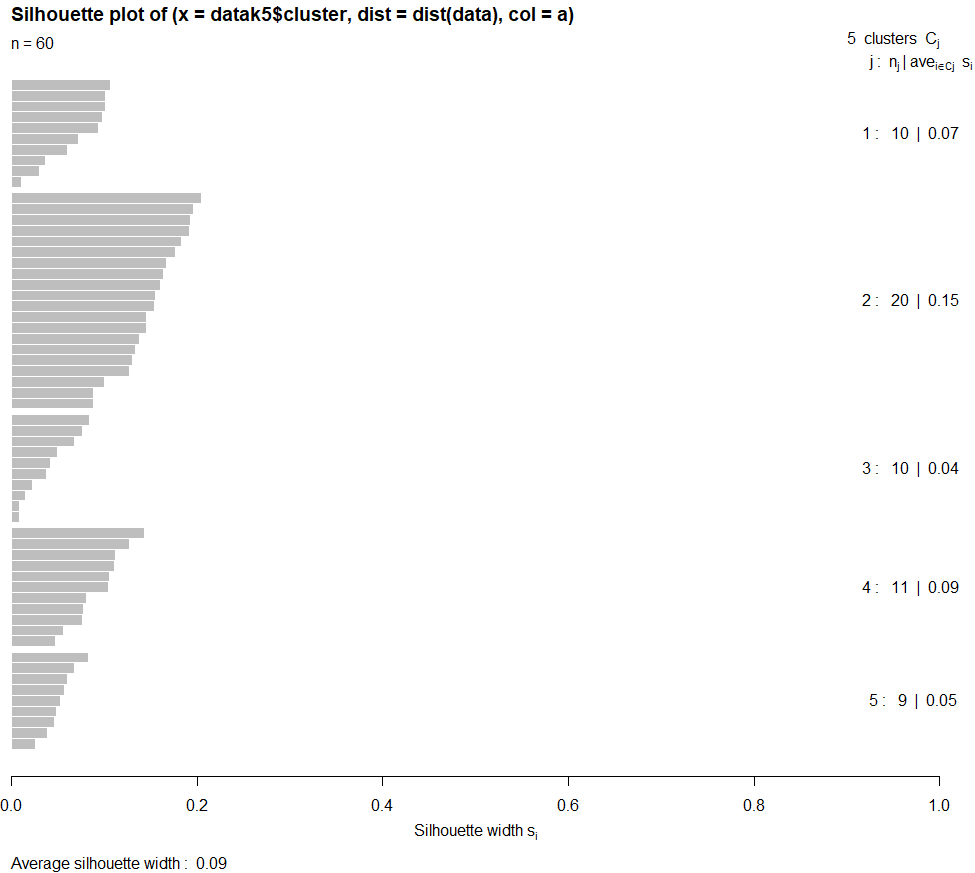
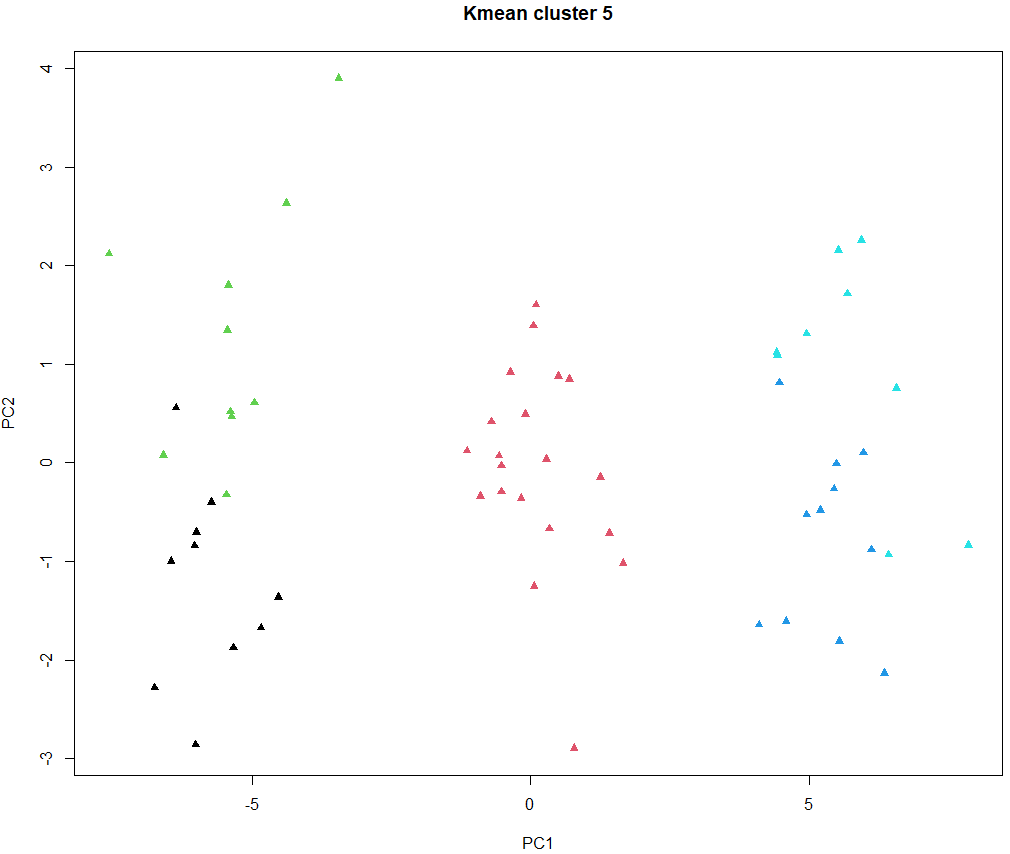
 

K = 3

K = 4

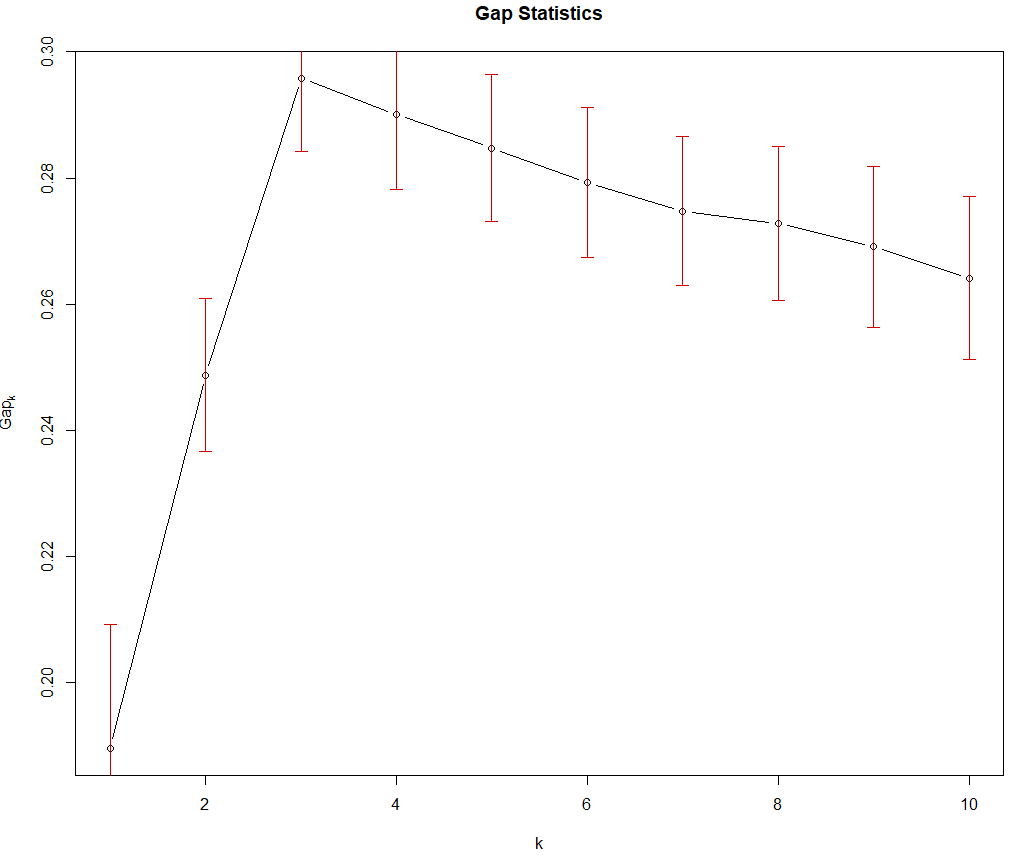
 

K = 5

The graphs shown are performed in order to compare and find the optimal K value for clustering. First of all, K = 2 has one that is pointing to minus, K = 4, 5 have components that are clustered below average. Therefore, K = 3 is the optimal K for silhouette plots.

d)



For gap statistics, K is maximized when K = 3 and immediately decreases after. Therefore, it is the optimal K