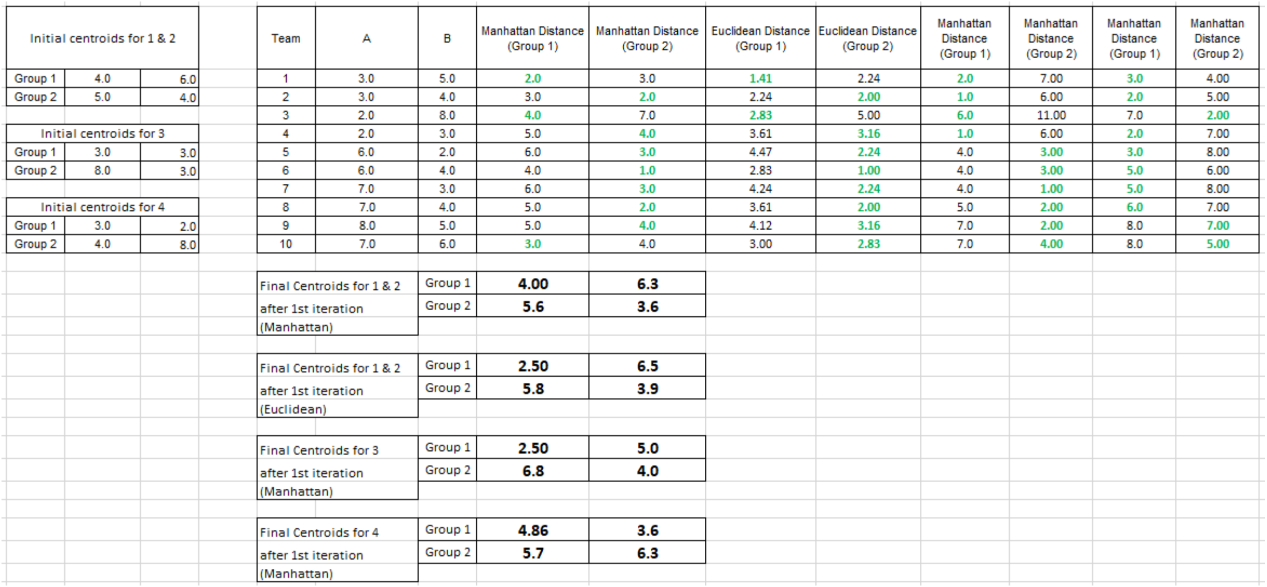
Task 1:

a, b, c, d) Results after one iteration on each of the centroids



Initial centroids for 1 & 2		
Group 1	4.0	6.0
Group 2	5.0	4.0

Initial centroids for 3		
Group 1	3.0	3.0
Group 2	8.0	3.0

Initial centroids for 4		
Group 1	3.0	2.0
Group 2	4.0	8.0

Team	A	B	Manhattan Distance (Group 1)	Manhattan Distance (Group 2)	Euclidean Distance (Group 1)	Euclidean Distance (Group 2)	Manhattan Distance (Group 1)	Manhattan Distance (Group 2)	Manhattan Distance (Group 1)	Manhattan Distance (Group 2)
1	3.0	5.0	2.0	3.0	1.41	2.24	2.0	7.00	3.0	4.00
2	3.0	4.0	3.0	2.0	2.24	2.00	1.0	6.00	2.0	5.00
3	2.0	8.0	4.0	7.0	2.83	5.00	6.0	11.00	7.0	2.00
4	2.0	3.0	5.0	4.0	3.61	3.16	1.0	6.00	2.0	7.00
5	6.0	2.0	6.0	3.0	4.47	2.24	4.0	3.00	3.0	8.00
6	6.0	4.0	4.0	1.0	2.83	1.00	4.0	3.00	5.0	6.00
7	7.0	3.0	6.0	3.0	4.24	2.24	4.0	1.00	5.0	8.00
8	7.0	4.0	5.0	2.0	3.61	2.00	5.0	2.00	6.0	7.00
9	8.0	5.0	5.0	4.0	4.12	3.16	7.0	2.00	8.0	7.00
10	7.0	6.0	3.0	4.0	3.00	2.83	7.0	4.00	8.0	5.00

Final Centroids for 1 & 2 after 1st iteration (Manhattan)			
	Group 1	**4.00**	**6.3**
	Group 2	**5.6**	**3.6**

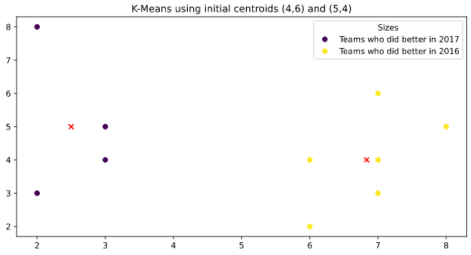
Final Centroids for 1 & 2 after 1st iteration (Euclidean)			
	Group 1	**2.50**	**6.5**
	Group 2	**5.8**	**3.9**

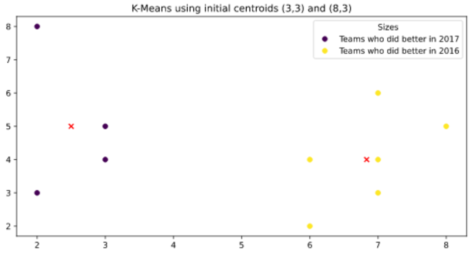
Final Centroids for 3 after 1st iteration (Manhattan)			
	Group 1	**2.50**	**5.0**
	Group 2	**6.8**	**4.0**

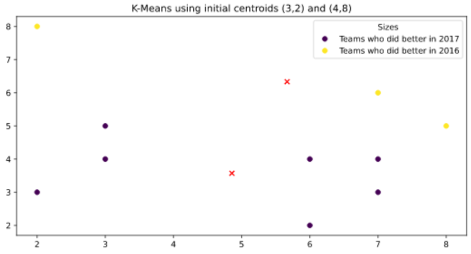
Final Centroids for 4 after 1st iteration (Manhattan)			
	Group 1	**4.86**	**3.6**
	Group 2	**5.7**	**6.3**

To access spreadsheet, please visit: https://github.com/lumalav/CAP5610/blob/master/HW5/Book1.xlsx

Clusters after all iterations using sklearn:



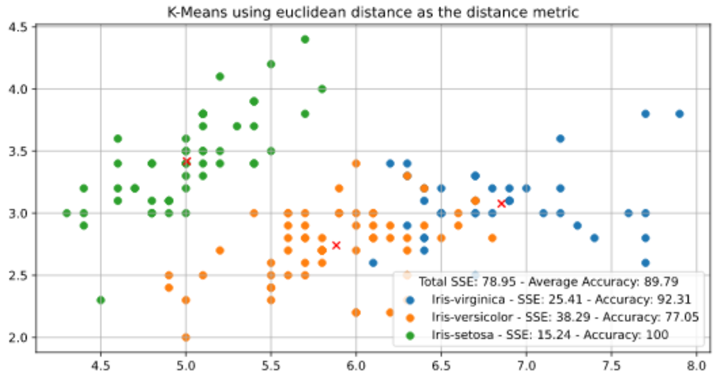


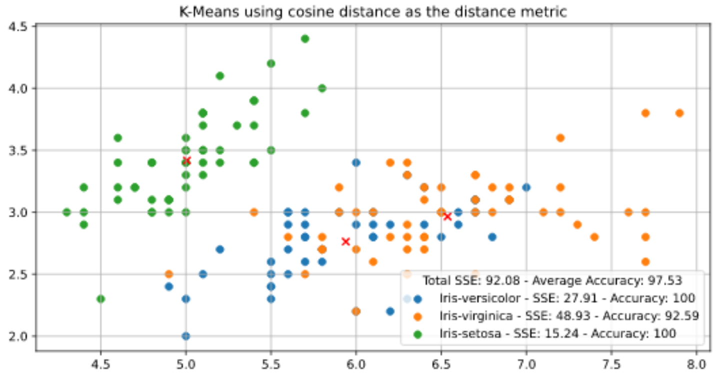


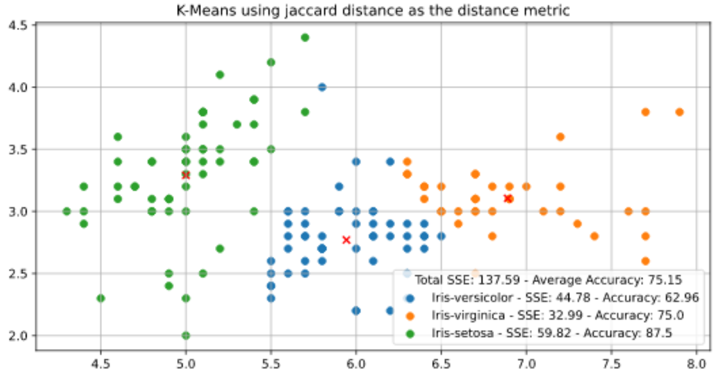
please refer the code at section #task 1 at: https://github.com/lumalav/CAP5610/blob/master/HW5/HW5.ipynb

Task 2:

1) Euclidean seems better since it has a lower SSE than the other 2 methods.
2) Even though it has a higher SSE, Cosine seems to have a better overall accuracy than the other 2 methods.
3) Jaccard seems to need more iterations than the other two methods.
4) a) Euclidean does best when there is no change in centroid position.
   b) Euclidean is also the best choice when SSE increases in the next iteration.
   c) Euclidean is also the best choice with the best SSE after 100 iterations.
   d) Jaccard requires more iterations.







please refer the code at section #task 2 at: https://github.com/lumalav/CAP5610/blob/master/HW5/HW5.ipynb

Task 3:

1) When the data has clusters of different sizes and densities it won't perform very well. Also, data with many outliers will change the position of the centroids. Additionally, as the numbers of dimensions increases with the data, PCA needs to be performed to reduce dimensionality in the data.

2) The approach that yields more consistent results is called K-means. Which steps are these:

a) Select one data point at random as an initial centroid.

b) Calculate the distance D(x) between the initial centroid and all other data points

c) Select next centroid from the remaining datapoints with probability proportional to D(x)^2

d) Repeat until all centroids have been assigned.

Task 4:

a) 2.1095
b) 0.9220
c) 1.4129
d) The group average is the one that is less susceptible to noise and less susceptible to outliers.