Clustering

Assignment 3_G4

Ahmed Yousry **Bassel Hamshary**
Mohamed El-Namoury

uOttawa
Faculté de génie
Faculty of Engineering

JUNE 29, 2021
UNIVERSITY OF OTTAWA
Ottawa, Canada

Table of Contents

1. Implementation

1.1. Part one (Numerical)

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Figure 1: Distance Matrix

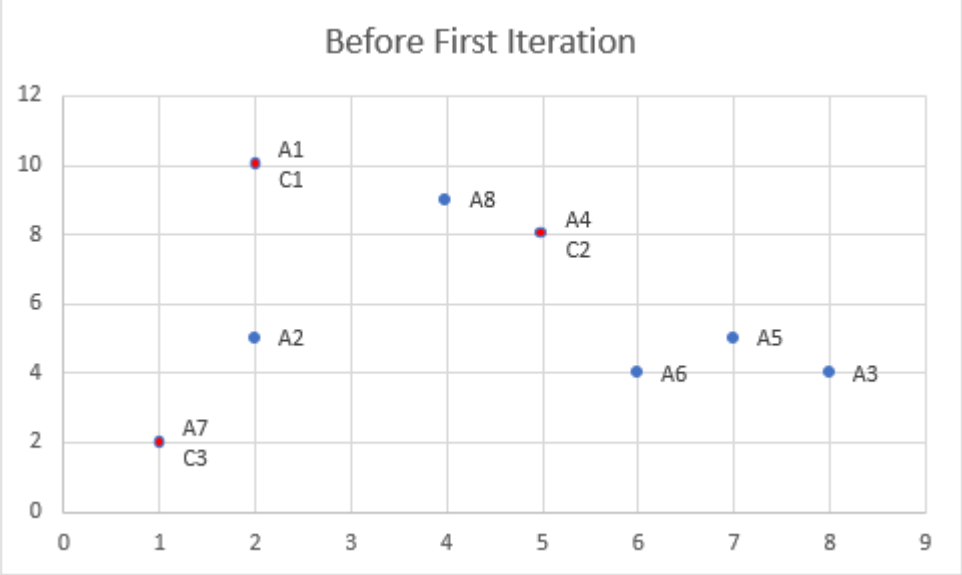
1.1.1. K-Means

➢ Question 1

- The Points:

A1= (2,10), A2= (2,5), A3= (8,4), A4= (5,8), A5= (7,5), A6= (6,4), A7= (1,2), A8= (4,9)

- The 3 clusters are A1= (2,10), A4= (5,8) & A7= (1,2)

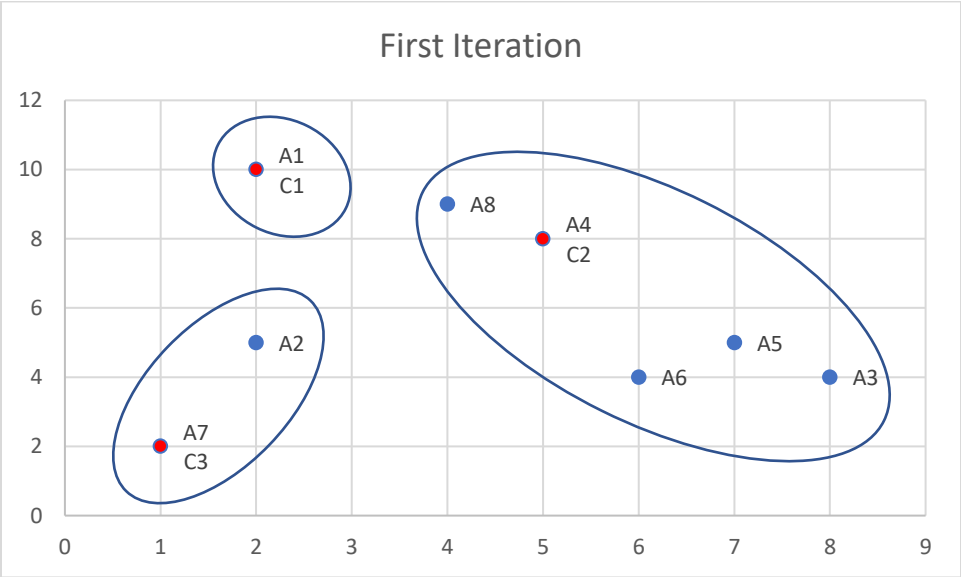


- 1st Iteration
  - Choosing a Cluster for each point (Depending on the minimum Euclidean Distance (ED))

$$Euclidean\ Distance(x,y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

A1	A2	A3	A4
ED to C1 = 0 ED to C2 = $\sqrt{13}$ ED to C3 = $\sqrt{65}$ A1 ϵ C1	ED to C1 = 5 ED to C2 = 4.24 ED to C3 = 3.16 A2 ϵ C3	ED to C1 = 6 ED to C2 = 5 ED to C3 = 7.28 A3 ϵ C2	ED to C1 = $\sqrt{13}$ ED to C2 = 0 ED to C3 = $\sqrt{52}$ A4 ϵ C2

A5 ED to C1 = 7.07 ED to C2 = 3.6 ED to C3 = 6.7 A5 ϵ C2	A6 ED to C1 = 7.21 ED to C2 = 4.12 ED to C3 = 5.38 A6 ϵ C2	A7 ED to C1 = $\sqrt{65}$ ED to C2 = $\sqrt{52}$ ED to C3 = 0 A7 ϵ C3	A8 ED to C1 = $\sqrt{5}$ ED to C2 = $\sqrt{2}$ ED to C3 = $\sqrt{58}$ A8 ϵ C2



C1: {A1}, C2: {A3, A4, A5, A6, A8}, C3: {A2, A7}

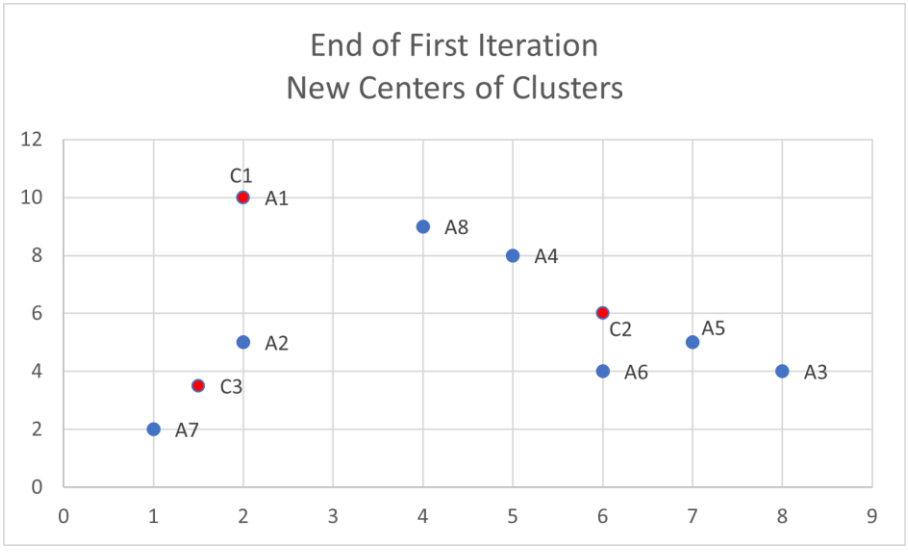
- Calculating the Coordinates of new Clusters:

$\left(\frac{\sum_{i=1}^{I} x_i}{I}, \frac{\sum_{i=1}^{I} y_i}{I}\right)$,Where I is the number of points that belongs to the cluster.

New C1 = (2,10)

New C2 = ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)

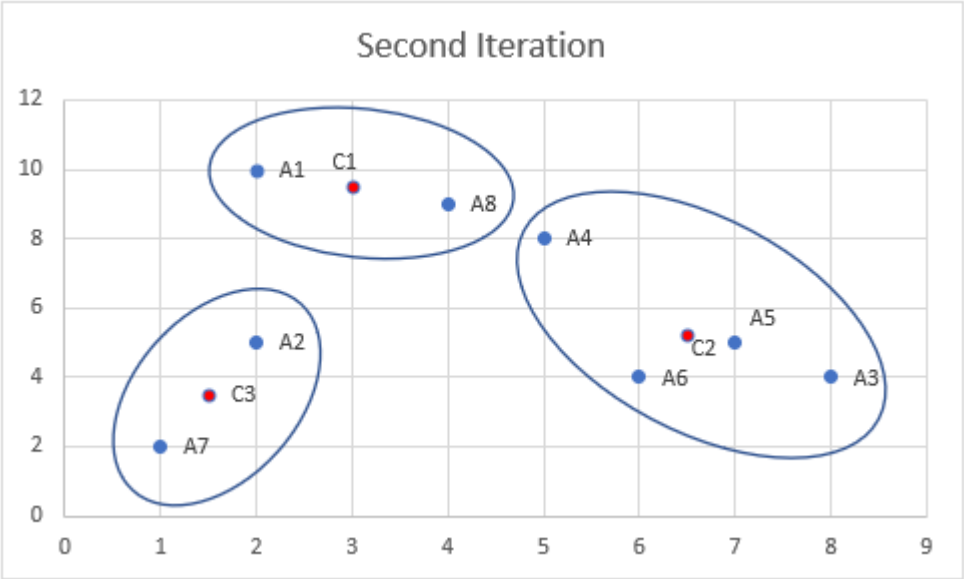
New C3 = ((2+1)/2, (5+2)/2) = (1.5, 3.5)



➢ Question 2

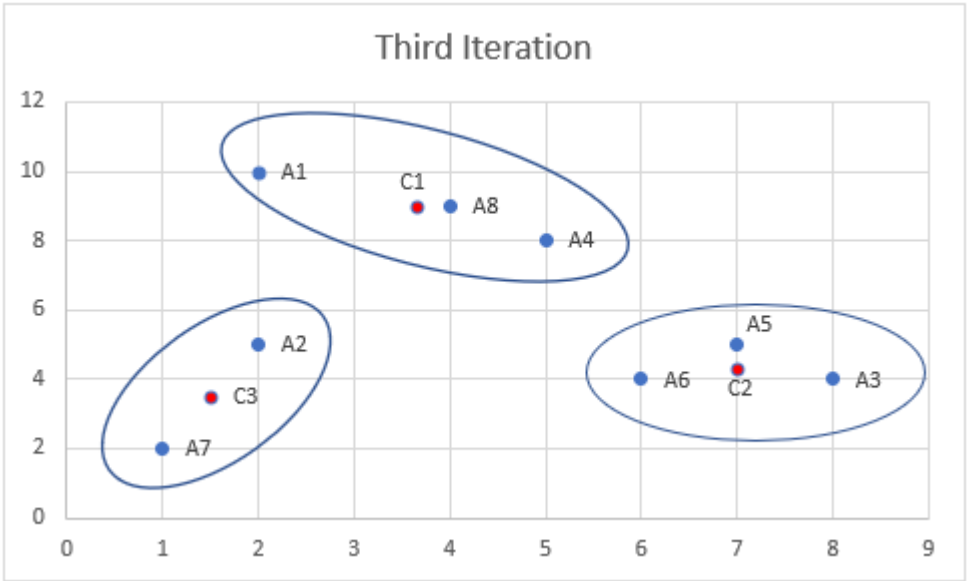
- Two more iterations are needed to converge.
  - 2nd Iteration

C1: {A1, A8}, C1 = (3, 9.5)
C2: {A3, A4, A5, A6}, C2 = (6.5, 5.25)
C3: {A2, A7}, C3 = (1.5, 3.5)



  - 3rd Iteration (Convergence)

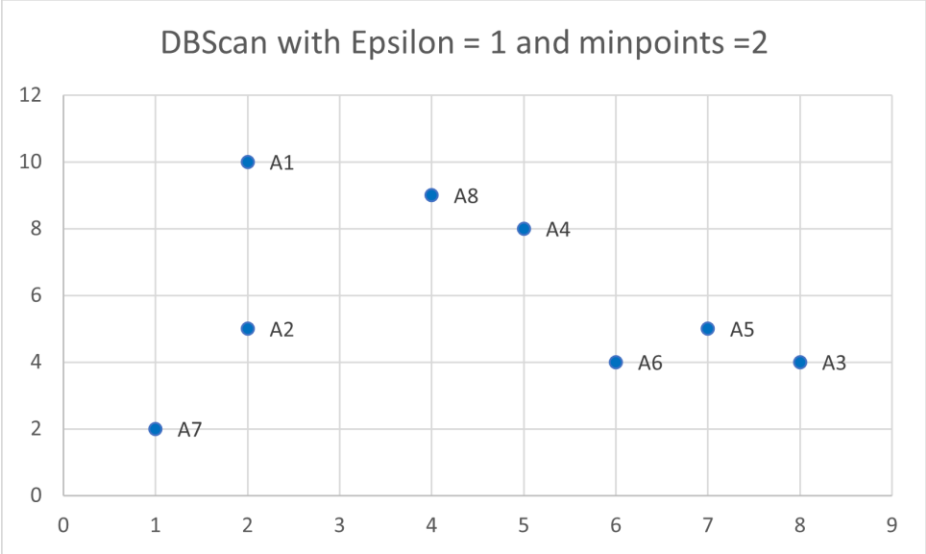
C1: {A1, A4, A8}, C1 = (3.66, 9)
C2: {A3, A5, A6}, C2 = (7, 4.33)
C3: {A2, A7}, C3 = (1.5, 3.5)



1.1.2. DBScan

- Question 1
  - **Epsilon = 4 / (1+3) = 1** **minpoints = 2**

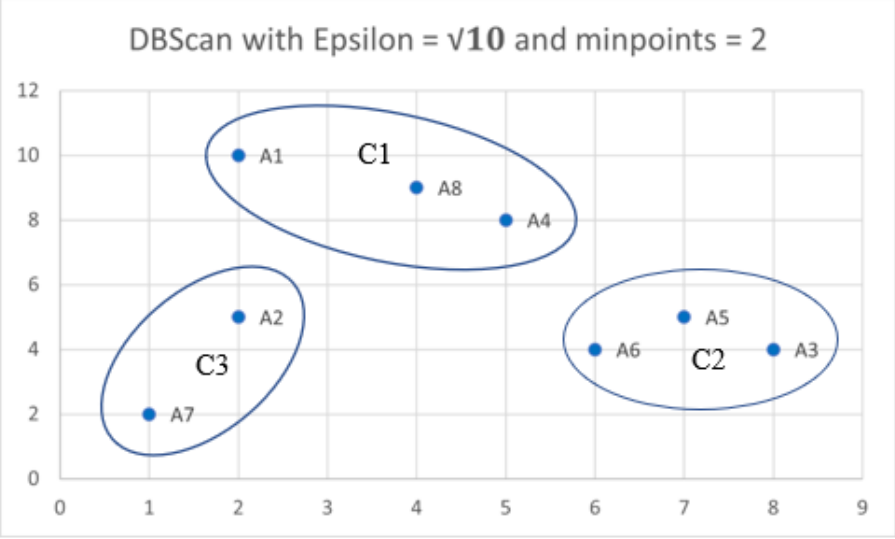
If Epsilon is equal to 1, so all the points will be categorized as noise, because epsilon is smaller than the Euclidean distance between each point and another one. Therefore, there will be no clusters.



- Question 2
  - **Epsilon = $\sqrt{10}$** **minpoints = 2**
  - **Calculating the neighborhood (N) depending on the distance matrix above.**

N(A1) = {A8}, N(A2) = {A7}, N(A3) = {A5, A6}, N(A4) = {A8}, N(A5) = {A3, A6}, N(A6) = {A3, A5}, N(A7) = {A2}; N(A8) = {A1, A4}

So, A1, A2, A4, and A7 are border points, while we have **3** clusters C1 = {A1, A4, A8}, C2 = {A3, A5, A6}, and C3 = {A2, A7}.

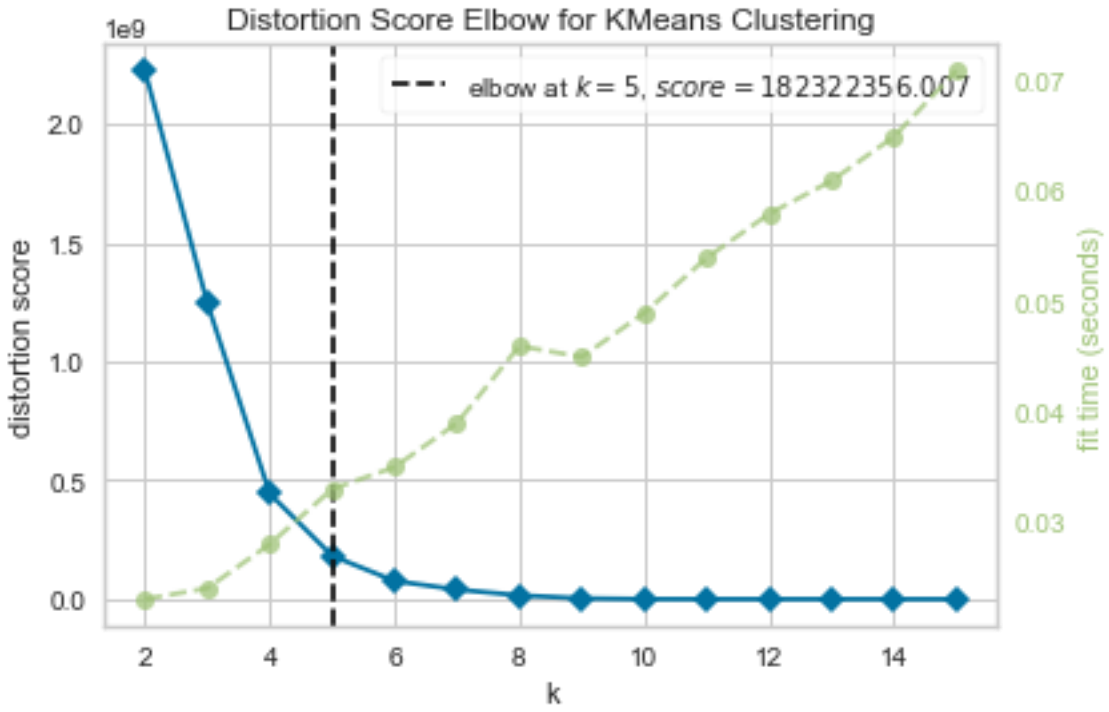


Comparing the result with the K-Means, we will find that both methods give the same results, but DBScan was easier as it does not require iteration like the K-Means.

1.2. Part Two (Programming)

1.2.1. Choose the Best Number of K for K-Means Algorithm

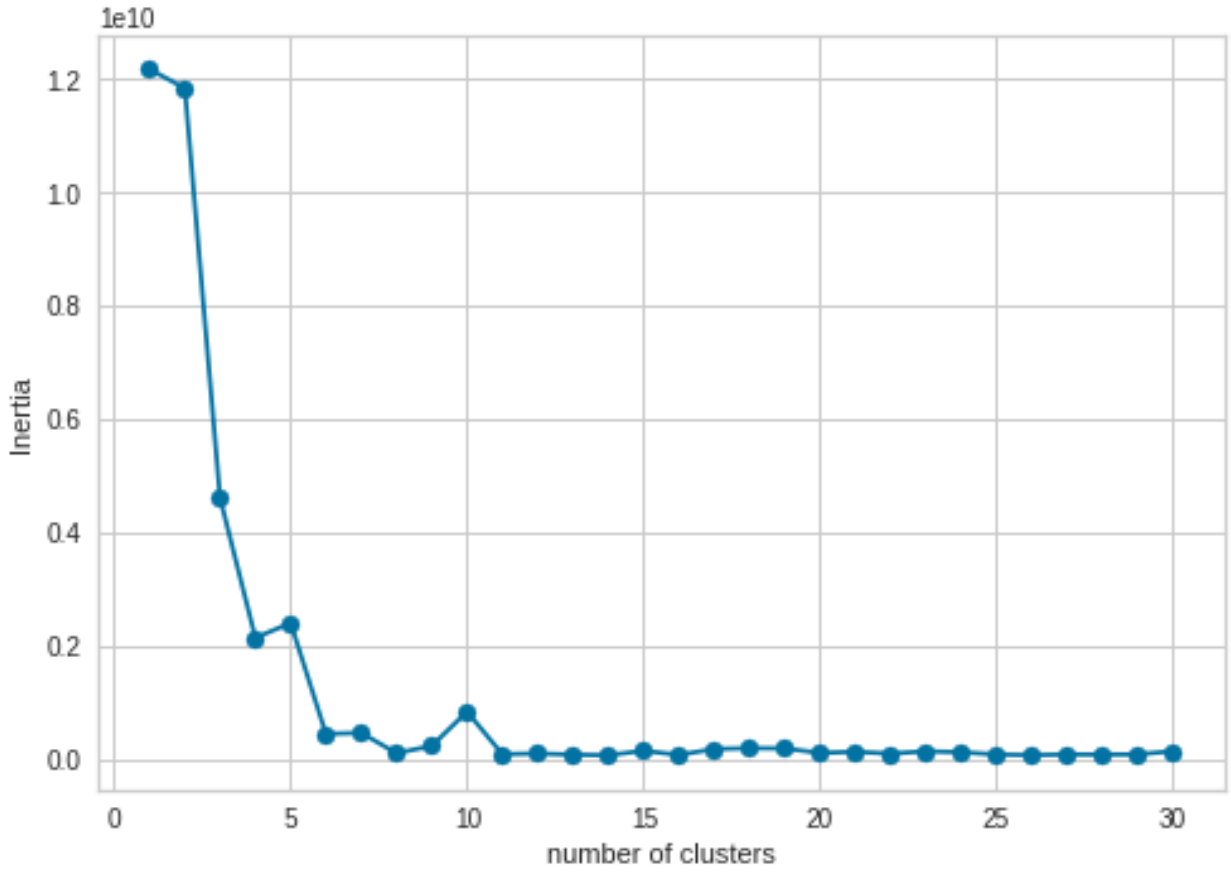
- Using the Elbow Rule for K_Means:



- The optimal number of clusters for k-means is **5** clusters.

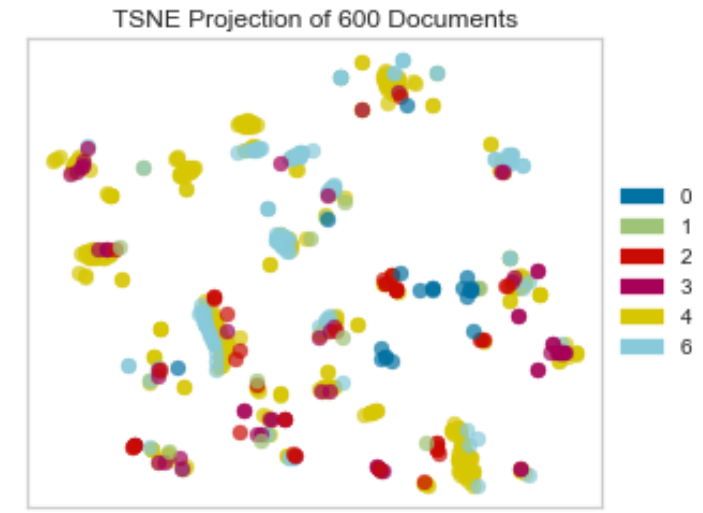
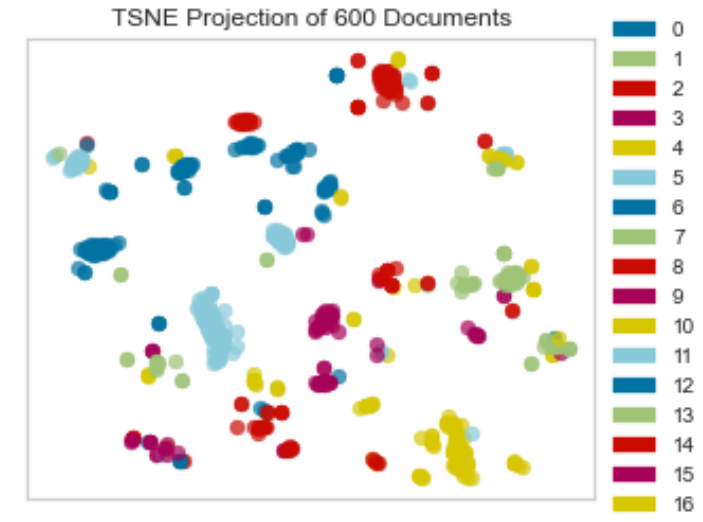
1.2.2. Choose the Best Number of Neurons for SOM Algorithm

- Using the Elbow Rule for SOM:



- The optimal number of neurons for SOM is 7 neurons.

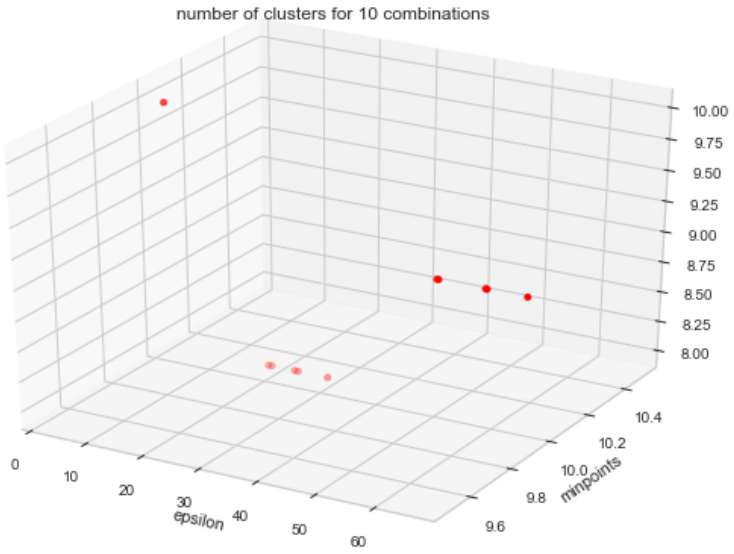
- **Use T-SNE method to visualize the SOM clusters, obtained from the previous step, on a 2D figure. Use different color code for each cluster.**

	**T-SNE vs SOM Predictions**
	**T-SNE vs Actual Labels**

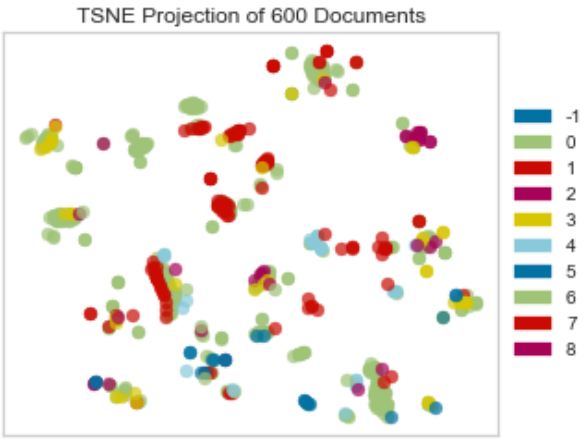
1.2.3. Tune the epsilon and minpoints to obtain 10 clusters.

- Plot the epsilon and minpoints values as x-y axes using a 3D figure

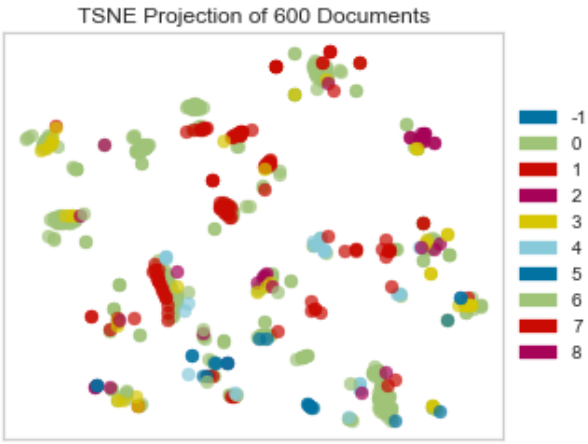
	Epsilon	Minpoints	Number of clusters
0	65.55	10	9
2	50.25	10	9
3	58.74	10	9
4	58.5	10	9
5	50.56	10	9
36	31.35	10	8
38	21.39	10	8
40	26.1	10	8
42	20.79	10	8
43	25.48	10	8
65	3.08	10	10



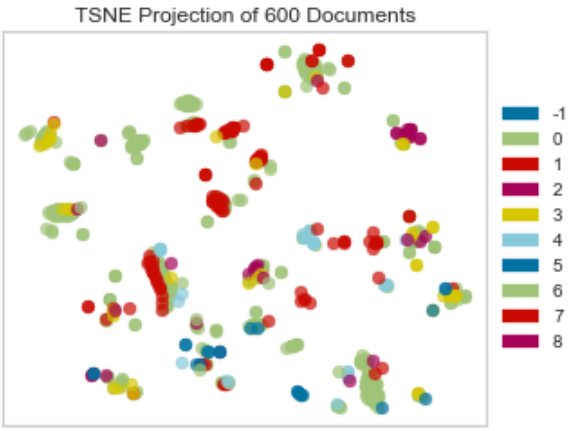
- After determining the parameters for DBSCAN with 10 clusters, plot DBSCAN clusters.



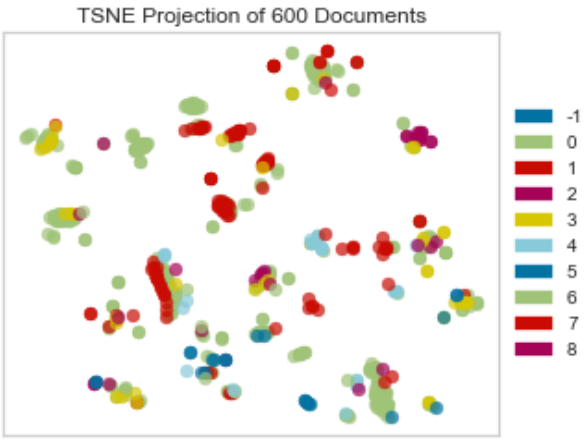
Epsilon = 65.55 Minpoints = 10



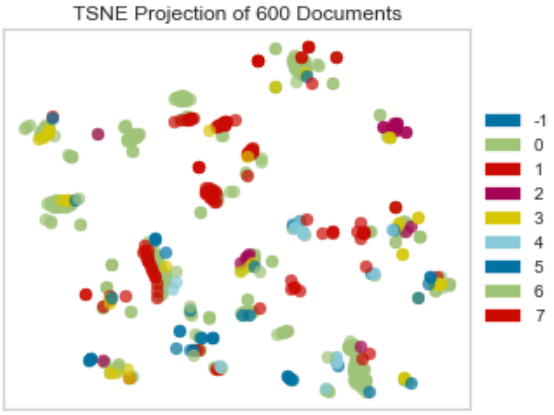
Epsilon = 50.25 Minpoints = 10



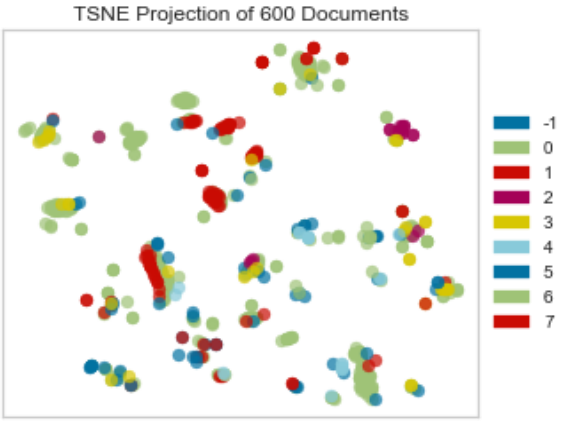
Epsilon = 58.74 Minpoints = 10



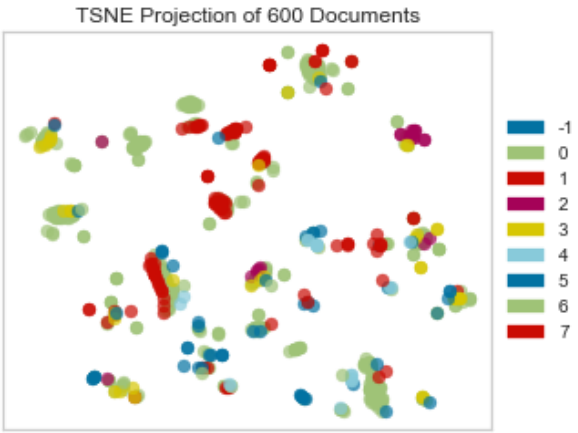
Epsilon = 58.5 Minpoints = 10



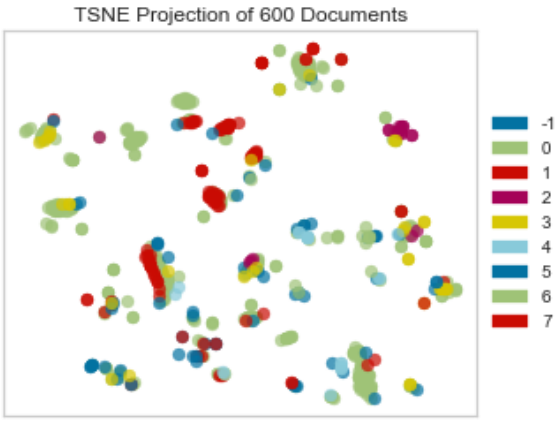
Epsilon = 31.35 Minpoints = 10



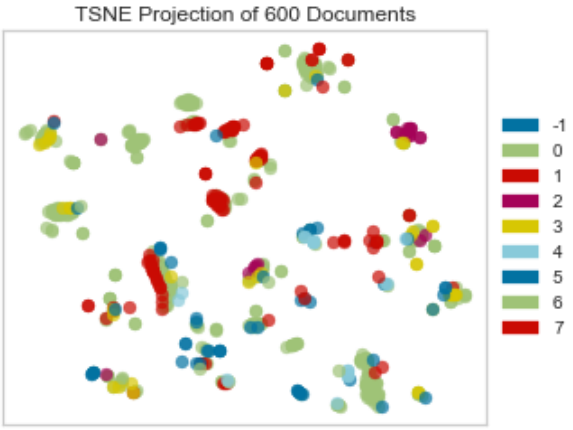
Epsilon = 21.39 Minpoints = 10



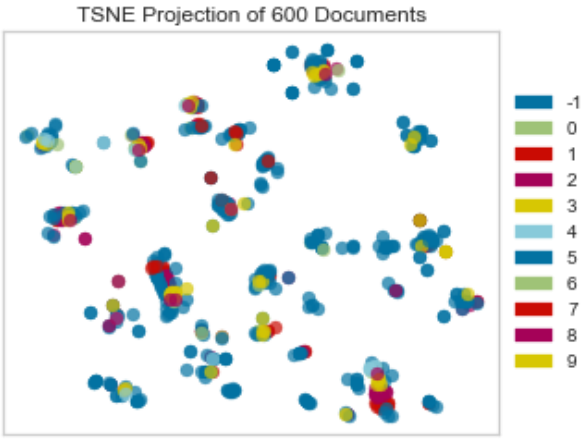
Epsilon = 26.1 Minpoints = 10



Epsilon = 20.79 Minpoints = 10



Epsilon = 25.48 Minpoints = 10



Epsilon = 3.08 Minpoints = 10