CHAPTER 6
CLUSTERING

Learning Objectives:

1. *Understand basic concepts in cluster analysis.*
2. *Understand and apply K-Means method.*
3. *Know how to evaluate the clustering method.*
4. *Build a K-Means workflow in KNIME and interpret result.*

Content

1. Introduction

Clustering tries to find natural groupings in the data, presumably with some similarities. Each grouping discovered is called a **cluster**.

Clustering is known as an unsupervised algorithm. We have a set of features (or variables) $X_1, X_2, \ldots, X_p$ (independent variables) measured on *n* items (objects, observations, records, cases, or rows). When we cluster the items, we seek to partition them into distinct groups so that the items within each group are quite similar to each other, while items in different groups are quite different from each other.

For example:

- You need to identify people with similar patterns of past purchases so that you can tailor your marketing strategies.

- You have been assigned to group television shows into homogeneous categories based on viewer characteristics. This can be used for market segmentation.

- You want to cluster skulls excavated from archaeological digs into the civilisations from which they originated. Various measurements of the skulls are available.

- You are trying to examine patients with a diagnosis of depression to determine if distinct subgroups can be identified, based on a symptom checklist and results from psychological tests.

The goal of clustering is to identify a structure of natural groupings. Quite often, we do not even know the number of groups.

2. Concept

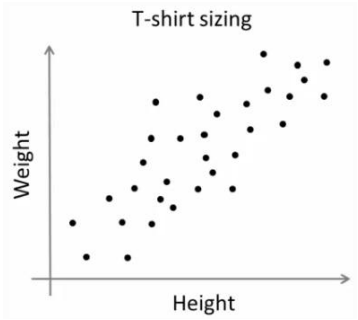
In cluster analysis, grouping is done on the basis of similarities or distances (i.e. dissimilarities). The inputs required are similarity measures or data from which similarities can be computed.

Generally, clustering analysis consists of two main steps.

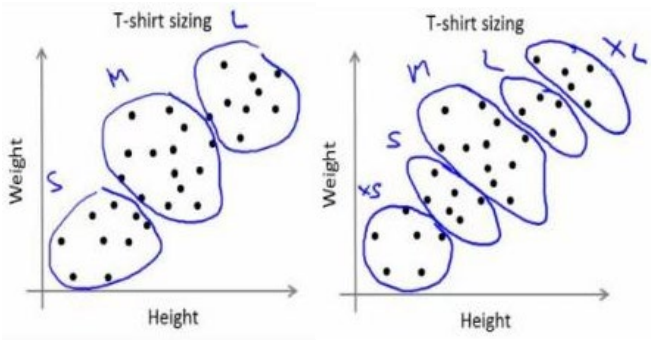
- Similarity measures to obtain distance matrix
  We must first develop or select a quantitative scale on which to measure similarity between objects.

- Clustering method
  Select a clustering procedure which can be a hierarchical algorithm or a non-hierarchical algorithm.

Suppose a company is going to release a new model of T-shirt to market. They need to manufacture models in different sizes to satisfy people of all sizes. So the company makes a plot of people's height and weight as seen below:



Intuitively, we can try to group the data as follows:



The above grouping of people into 3 groups or 5 groups can be done by clustering. We can visualize clusters in up to 3 dimensions but beyond that we have to rely on a more mathematical understanding.

3. Similarity Measures and Centroids

Most efforts to produce a rather simple group structure from a complex data set require a measure of *similarity*. There is often a great deal of subjectivity involved in the choice of a similarity measure. Important considerations include the nature of the variables (discrete, continuous, binary), scales of measurement (nominal, ordinal, interval, ratio), and knowledge of subject matter.

When items are clustered, proximity is usually indicated by some sort of distance. Some commonly used distance measures are:

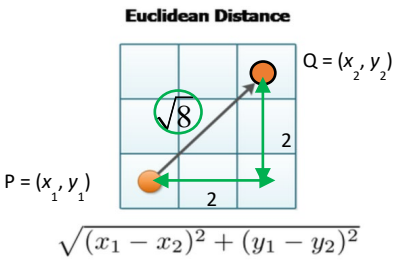
- Euclidean Distance (square root of the sum of the squares of the differences of the coordinates)

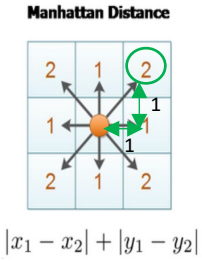
$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Manhattan Distance (sum of the absolute values of the differences of the coordinates)

$$d(x,y)=\sum_{i=1}^{n}|x_i-y_i|$$

where n is the number of variables, x_i and y_i are the variables of vector x and y respectively in the two dimensional vector space.



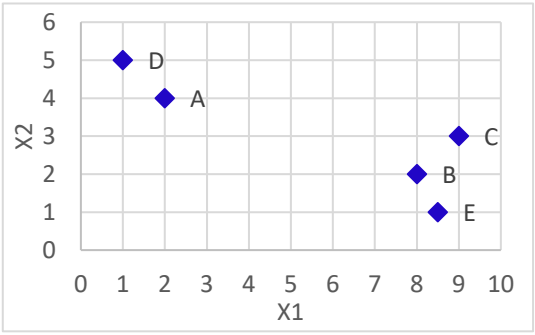


When items are clustered, each member of the cluster has more in common with other members of the same cluster than with members of the other groups. The most representative point within the group is called the *centroid*. Usually, this is the mean of the values of the points of data in the cluster.

Example 1: The daily expenditures on food X_1 and clothing X_2 of five persons are shown as follows:

(Since only two variables were involved in grouping, we can create a plot and visually inspect the clustering.)

Person	X_1	X_2
A	2	4
B	8	2
C	9	3
D	1	5
E	8.5	1



(a) Complete the following distance matrix based on the Euclidean distance :

	A	B	C	D	E
A	0	6.325	7.071	1.414	7.159
B					
C					
D					
E					

(b) Complete the following distance matrix based on the Manhattan distance:

	A	B	C	D	E
A	0	8	8	2	9.5
B					
C					
D					
E					

Answers:

Euclidean distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

$$d(A,B) = \sqrt{\sum_{i=1}^{2}(A_i - B_i)^2} = \sqrt{(2-8)^2 + (4-2)^2} = 6.325$$

Manhattan distance:

$$d(x,y) = \sum_{i=1}^{p} | x_i - y_i |$$

$$d(A,B) = \sum_{i=1}^{2} | A_i - B_i | \; = \; |2-8| + |4-2| \; = \; 8$$

Example 2: From *Example 1*, if each person is grouped into the following cluster:

Person	X_1	X_2	Cluster
A	2	4	1
B	8	2	2
C	9	3	2
D	1	5	1
E	8.5	1	2

Each cluster has a "**centroid**", which is just the mean of the attributes, X_1 and X_2. Compute the centroids of cluster 1 and cluster 2.

Answers:

Mean of $X_1 = (2+1)/2 = 1.5$

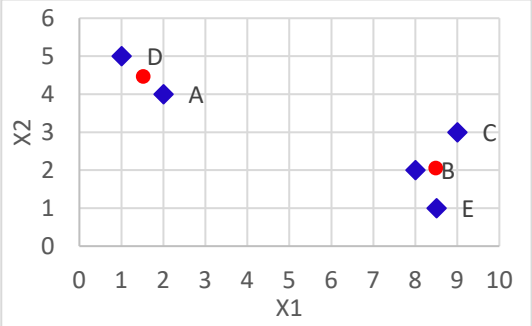
Mean of $X_2 = (4+5)/2 = 4.5$

Centroid of cluster 1 $= (1.5,\ 4.5)$

Mean of X_1 =

Mean of X_2 =

Centroid of cluster 2 =



4. Standardization of Variables for Clustering

Usually variables need to be standardized or normalized before use in clustering, often because different variables may be measured on different scales. In cases where the range of values differs widely from variable to variable, these different scales can influence the clustering results and it is common practice to standardize the data so that all variables are on the same scale. Common normalization methods include:

- *Z*-score standardization
- Min-max standardization

5. Partitional Clustering

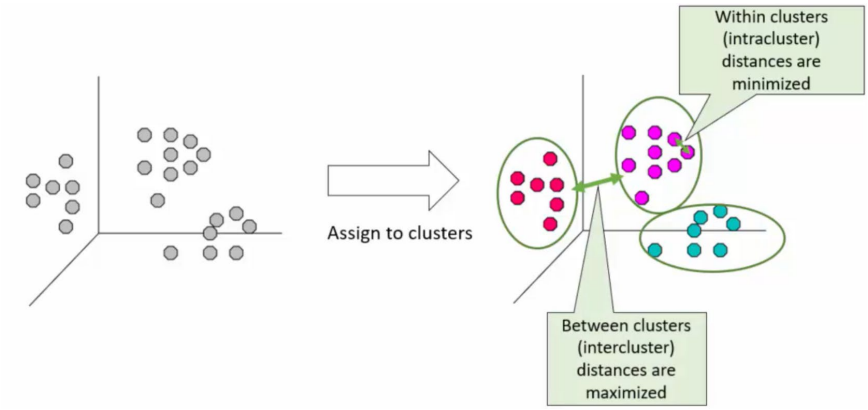
Partitional clustering method falls into the non-hierarchical clustering techniques which are designed to group items into a collection of K clusters. The number of clusters K may either be specified in advance or determined as part of the clustering procedure. The K-Means algorithm is one of the most popular partitional clustering methods.

Non-hierarchical methods start from either:

- an initial partition of items into groups, or
- an initial set of seed points, which will form the centroid of clusters.

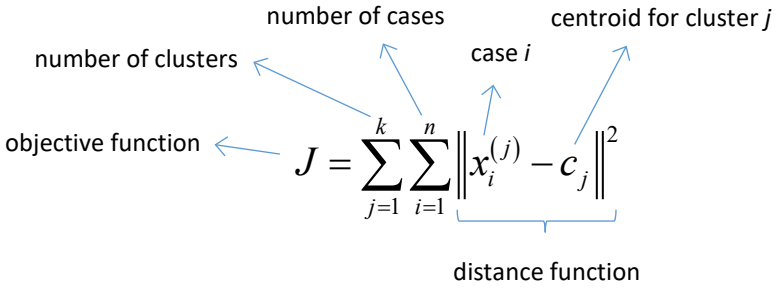
Good choices for starting configurations should be free of overt biases. One way to start is to randomly select seed points from among the items or to randomly partition the items into initial groups.

Generally, a way to achieve a good clustering is to minimize the within clusters distances (intra-cluster) and maximize the between clusters distances (inter-cluster) as illustrated in the following:



6. K-Means Clustering

K-Means clustering is a non-hierarchical clustering technique. It is an algorithm that attempts to find a user-specified number of clusters (k), which are represented by their centroids. K-Means clustering partitions n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:



6.1 Algorithm

1. Clusters the data into *k* groups where *k* is predefined.
2. Randomly select *k* points as initial cluster centers.
3. Assign data points to their closest cluster center according to the *distance* function.
4. Calculate the centroid or mean of all points in each cluster.
5. Repeat steps 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

K-Means is a relatively efficient method. However, we need to specify the number of clusters in advance and the final results are sensitive to initialization and often terminates at a local optimum. Unfortunately there is no global theoretical method to find the optimal number of clusters. A practical approach is to compare the outcomes of multiple runs with different *k* and choose the best one based on a predefined criterion. In general, a large *k* probably decreases the error but increases the risk of overfitting.

6.2 Terminating conditions

The K-Means algorithm runs iteratively until a terminating condition is met. Following are the common terminating conditions:

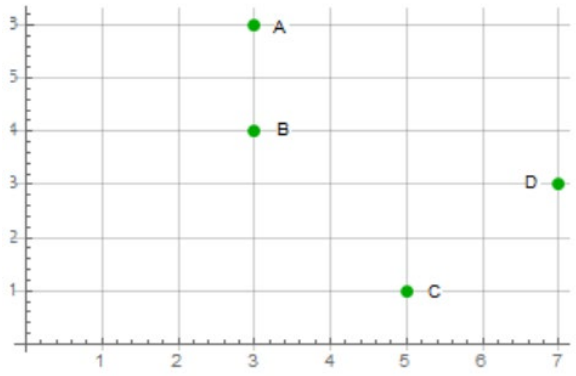
- When same points are assigned to each cluster in consecutive rounds
- Maximum number of iterations is reached
- Centroids do not change their positions
- Sum of the distances is minimized

Example 3: Use the K-Means method to cluster the following four items measured on two variables into 2 clusters:

Item	X_1	X_2
A	3	6
B	3	4
C	5	1
D	7	3

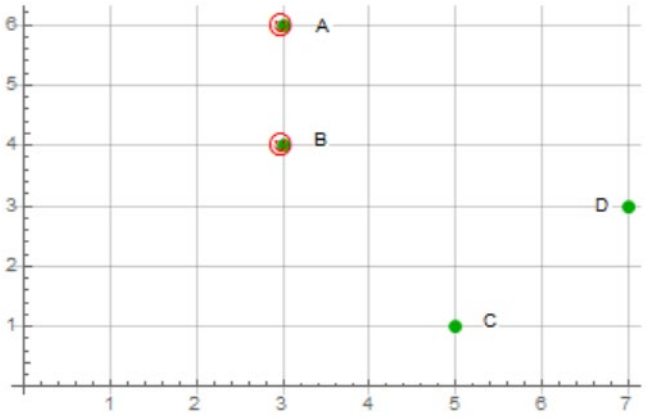
Solution:

The items can be visualized using a scatterplot as follows:



We arbitrarily choose A and B as the first centroids C1 and C2. That is:

$C1 = A = (3,6)$
$C2 = B = (3,4)$



We then compute the distance between each item to each centroid, using the Euclidean distance.

$d(A,C1) = d(A,A) = 0$	$d(A,C2) = d(A,B) = 2$
$d(B,C1) = d(B,A) = \sqrt{2^2 + 0^2} = 2$	$d(B,C2) = d(B,B) =$
$d(C,C1) = d(C,A) =$	$d(C,C2) = d(C,B) =$
$d(D,C1) = d(D,A) =$	$d(D,C2) = d(D,B) =$

The above leads us to the following:

- A stays with C1
- B goes with C2
- C goes with C2
- D goes with C2

We now have two clusters: (A) and (BCD)

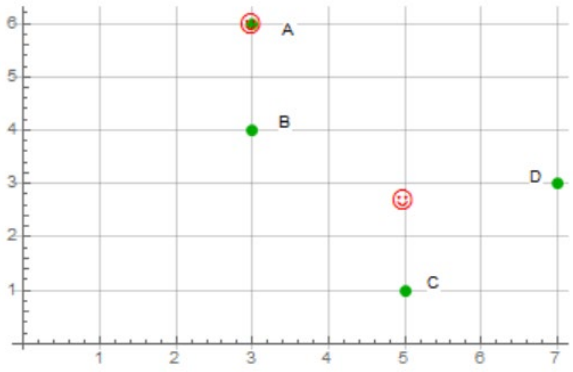
Iteration 1

New centroids are as follows:

$$C1 = A = (3,6)$$

$$C2 = \left(\frac{3+5+7}{3}, \frac{4+1+3}{3}\right)$$

$$= \left(5, \frac{8}{3}\right) = (5, 2.67)$$



Again we compute the distance between each item and the new centroids C1 and C2.

$d(A, C1) = 0$	$d(A, C2) = 3.88$
$d(B, C1) = 2$	$d(B, C2) = 2.40$
$d(C, C1) = 5.39$	$d(C, C2) =$
$d(D, C1) = 5.0$	$d(D, C2) =$

This leads us to:

- A stays with C1
- B goes with C1
- C stays with _____
- D stays with _____

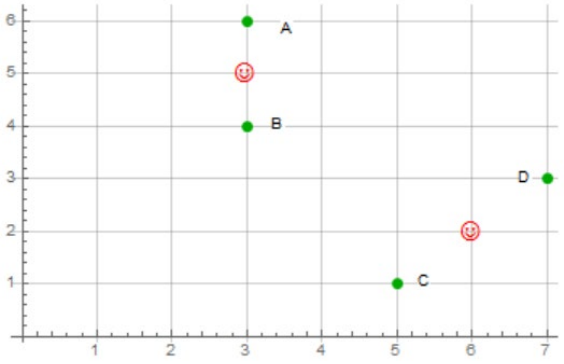
Now the new clusters are (AB) and (CD).

Iteration 2

New centroids are:

$$C1 = \left(\frac{3+3}{2}, \frac{6+4}{2}\right) = (3,5)$$

$$C2 =$$



Distance between each item and the new centroids C1 and C2 are as follows:

$d(A,C1)=1$	$d(A,C2)=$
$d(B,C1)=1$	$d(B,C2)=$
$d(C,C1)=4.47$	$d(C,C2)=$
$d(D,C1)=4.47$	$d(D,C2)=$

This leads us to:

- A stays with C1
- B stays with C1
- C stays with C2
- D stays with C2

Since there is no more changing of cluster membership, the computations stop and the final grouping is (AB) and (CD).

7. Evaluating the clustering method

The goal of clustering is to come up with meaningful clusters. Since there are many variations that can be chosen, it is important to make sure that the resulting clusters are valid, in the sense that they really create some insight. The major tasks of clustering evaluation include the following:

- **Understand the nature of the clusters formed.**
  We can do this by using visualization tools to analyse how the input attributes define the clusters formed and how they affect the clustering (e.g. scatterplots, parallel coordinates).

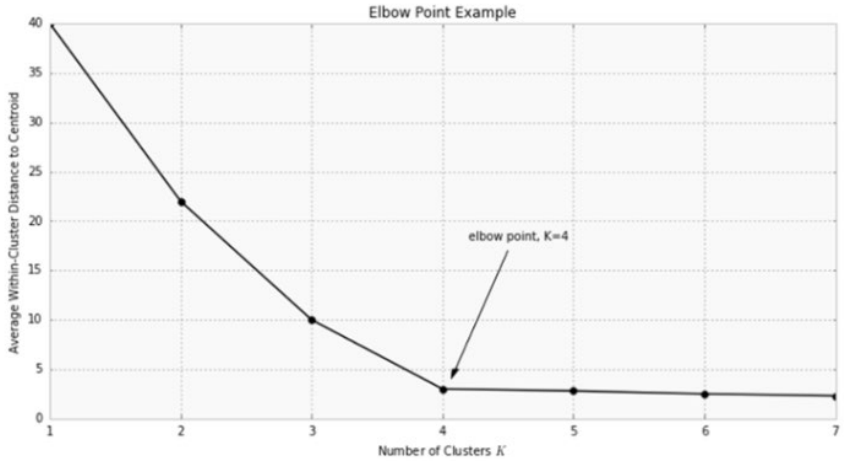
- **Measure the clustering quality.** We can use supervised mining techniques (e.g. decision trees) to derive rules for the clusters formed by performing supervised mining using the cluster ID as the output attribute and evaluate indicators in the report. Alternatively, we can examine how well the clusters are separated and how compact the clusters are (e.g. ratio of between-cluster variation to within-cluster variation).

- **Determine the 'right' number of clusters in a data set.**
  The appropriate number of clusters controls the proper granularity of cluster analysis. One method that is commonly used is the 'elbow' method (refer section 7.1). It is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. A heuristic for selecting the right number of clusters is to use the turning point in the curve of the sum of within-cluster variances with respect to the number of clusters.

7.1 Choosing K

There is no method for determining the exact value of *k*, but an accurate estimate can be obtained using the following 'elbow' method:

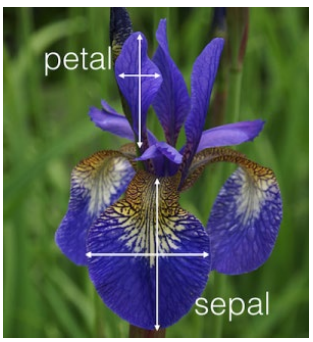
We can compute the mean distance between data points and their cluster centroid and compare the results across different values of *k*. Since increasing the number of clusters, *k*, reduces the distance to data points, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of *k* is plotted and the 'elbow point', where the rate of decrease sharply shifts, can be used to roughly determine *k*.



Other techniques exist for validating *k* and the list is not exhaustive: cross-validation method, the information criteria, the information theoretic jump method, the silhouette method, and the G-means algorithm. In addition, visualization of the resulting distribution of data points across groups provides insight into how the algorithm is splitting the data for each *k*. Generally, good clusters are compact (small sum of squares divided by number of records in the cluster) and well-separated from other clusters (centroid distance).

8. Build K-Means workflow in KNIME: Case study with UCI dataset

The iris flower dataset *iris.csv* is a multivariate dataset first introduced by the British statistician and biologist Ronald Fisher in 1936. The dataset consists of 50 samples for each of the three species of Iris (*Iris setosa*, *Iris versicolor*, and *Iris virginica*). Four attributes were measured for each sample: *sepal length*, *sepal width*, *petal length*, and *petal width* (all in cm). The table below gives a partial display of the dataset.



Row ID	D sepal le...	D sepal w...	D petal le...	D petal wi...	S class
Row0	5.1	3.5	1.4	0.2	Iris-setosa
Row1	4.9	3	1.4	0.2	Iris-setosa
Row2	4.7	3.2	1.3	0.2	Iris-setosa
Row3	4.6	3.1	1.5	0.2	Iris-setosa
Row4	5	3.6	1.4	0.2	Iris-setosa
Row5	5.4	3.9	1.7	0.4	Iris-setosa
Row6	4.6	3.4	1.4	0.3	Iris-setosa
Row7	5	3.4	1.5	0.2	Iris-setosa
Row8	4.4	2.9	1.4	0.2	Iris-setosa
Row9	4.9	3.1	1.5	0.1	Iris-setosa
Row10	5.4	3.7	1.5	0.2	Iris-setosa
Row11	4.8	3.4	1.6	0.2	Iris-setosa
Row12	4.8	3	1.4	0.1	Iris-setosa

Suppose we want to know if there are more than one distinct kinds of iris plants represented in these data. In other words, we want to label each measured flower as some species of iris. We assume that there are only 3 species of iris. Hence, we assume we don't know, a priori, which points belong in which ‘class’.

K-Means clustering then works as follows:

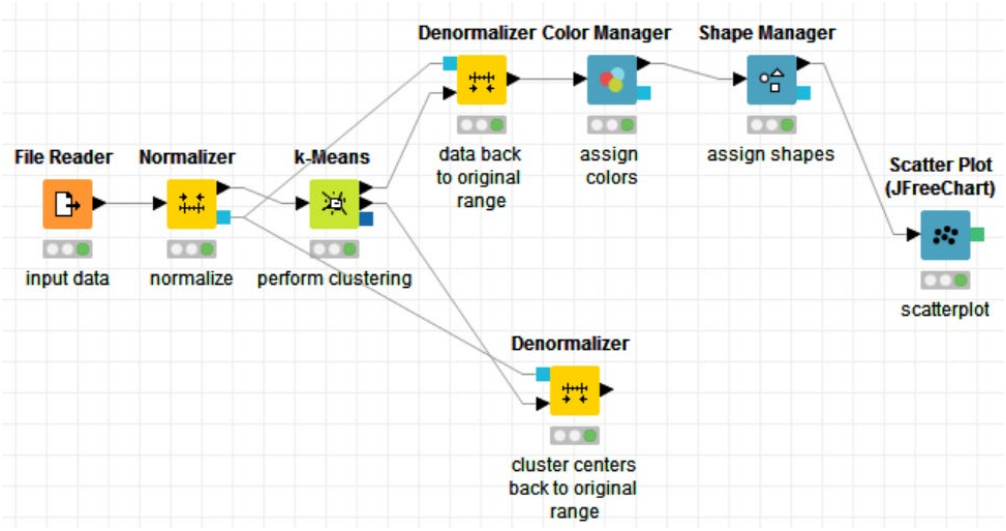
1. Decide how many clusters we want. Call this *k*.
2. Create *k* random cluster means (also called "centroids"). We can choose random values for each dimension for each of the *k* clusters or we can choose a random data point to represent each initial cluster mean.
3. For each measured flower (each row in the table of data), use the Euclidean distance to determine which cluster's mean is closest to the measurements. Assign this flower to that cluster. (Note it may have already been assigned to that cluster.)
4. Now that all flowers have been assigned (or reassigned) to clusters, recalculate the cluster means.
5. Go back to step 3 until no cluster assignments change.

We construct a KNIME workflow as follows:

KNIME generates the cluster centroids below:

Row ID	sepal le...	sepal w...	petal le...	petal wi...
cluster_0	0.707	0.451	0.797	0.825
cluster_1	0.441	0.307	0.576	0.549
cluster_2	0.196	0.591	0.079	0.06

Notice that we have used a Normalizer to standardize the variables. To return the variables to their original scale, we have to de-normalize them:

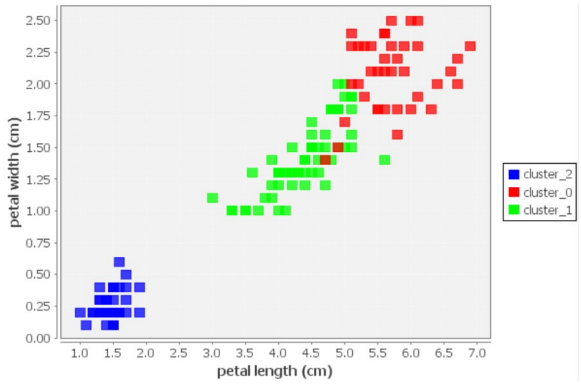


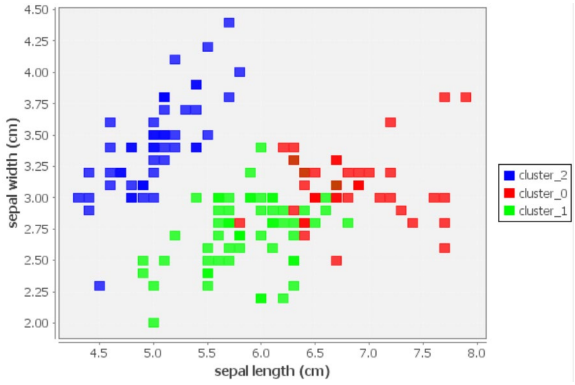
We are now ready to view the results and analyse the clusters generated.

The cluster centroids are:

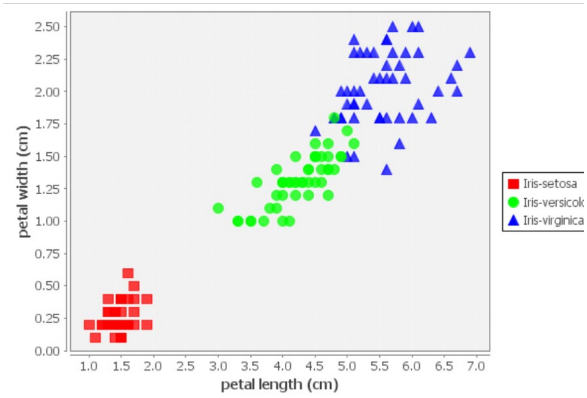
Row ID	sepal le...	sepal w...	petal le...	petal wi...
cluster_0	6.846	3.082	5.703	2.079
cluster_1	5.889	2.738	4.397	1.418
cluster_2	5.006	3.418	1.464	0.244

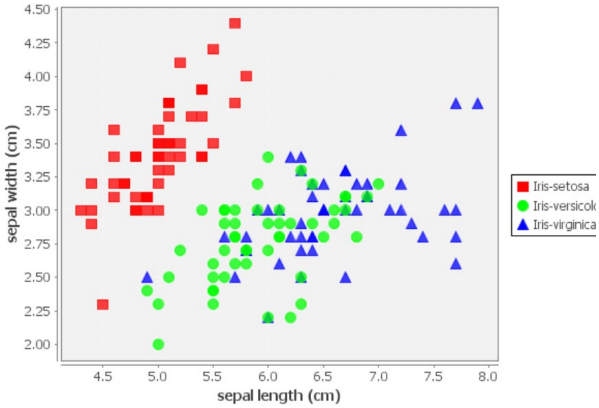
Some clusters generated are as follows:





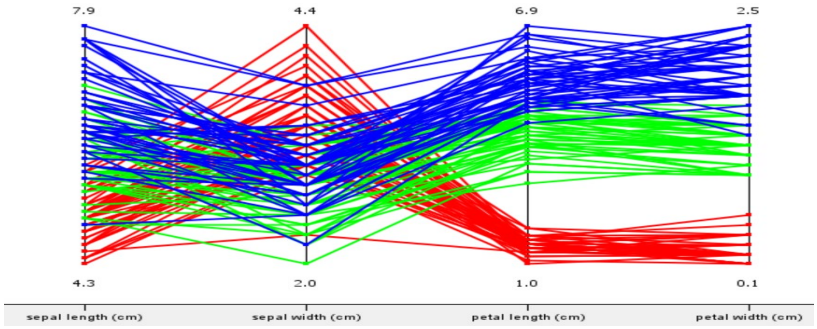
Compare the above with known clusters:





Notice that Iris-Setosa, with smaller petal width and petal length, are well separated from Versicolor and Virginica. However, Versicolor and Virginica do not appear to be easily separated based on measurements on petal length/width and sepal length/width.

A parallel coordinates plot below suggests a similar picture, that the species are reasonably well separated for petal length and petal width, but less well separated for sepal length and sepal width.



TUTORIAL 6

1. *Table 1.1* contains the customer database for the credit card operations of a bank. The database stores customer age, annual income ($) in thousands, and average monthly credit card expenditure ($) in thousands. The k-means algorithm with k=2 was applied to the data to find customer segments. The cluster membership is shown in the last column of the table.

Table 1.1			
Age	**Income**	**CCAvg**	**Cluster**
52.00	35.00	0.20	2
59.00	73.00	1.70	2
41.00	115.00	7.00	1
49.00	81.00	2.00	2
38.00	81.00	4.00	1

(a) Calculate the centroids of the two clusters (in the original units of the data in Table 1.1).

(b) Use the results of part (a) to assign a descriptive name to each of the customer segments. The names you choose should aid the marketing department in designing marketing campaigns directed at the customer segments.

(c) The clustering results above were obtained after the data was normalized using z-score standardization (subtracting the mean and dividing by the standard deviation). Table 1.2 contains the normalized data.

Table 1.2			
Age	**Income**	**CCAvg**	**Cluster**
0.50	-1.47	-1.06	2
1.32	-0.14	-0.49	2
-0.80	1.33	1.53	1
0.14	0.14	-0.37	2
-1.16	0.14	0.39	1

Explain why the clustering was done on normalized data.

(d) The database record for a new bank customer is shown below (normalized).

Age	**Income**	**CCAvg**
1.322	1.542	-0.0305

Assign the new bank customer to one of the two customer segments. Show the detailed calculations necessary. (Hint: you need to find the cluster centroids for Table 1.2.)

(e) In what way does the new bank customer differ most from the typical customers in the customer segment you selected in part (d)?

2. The dataset decathalon.txt contains the results of athletes in two decathlons. Each of the 41 records in the dataset represents the results of one athlete (some athletes appear twice in the dataset), while each column records the results for one of the 10 events of the decathlon.

   The dataset was clustered using k-means with k = 2. Only the variables representing the results for the long jump (in metres) and the 400m run (in seconds) were included in the clustering, and the data was normalized before k-means was applied. Results are below.

```
                          Within   Average   Maximum
                         cluster  distance  distance
            Number of    sum of      from      from
          observations  squares  centroid  centroid
Cluster1            21   16.359     0.753     2.075
Cluster2            20   25.545     1.014     2.673


Cluster Centroids

Variable   Cluster1  Cluster2
400m        -0.6450    0.6772
Long.jump    0.6848   -0.7190
```

   (a) Which of the two clusters is more compact?

   (b) Which of the two clusters is the winner of the decathlon most likely to belong to?

   (c) Assign the following new record (normalized) to one of the clusters.

New Record	
400m	**Long.jump**
-0.768	-0.221

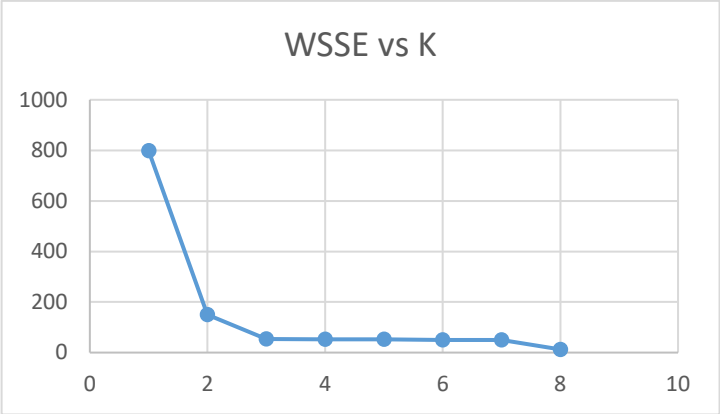
3. Given a set { 2, 3, 4, 10, 11, 12, 20, 25, 30 } is clustered using K-Means with $k = 2$.

   The centroids of the two clusters are initially chosen as Centroid_1 = 4 and Centroid_2 = 12 respectively.

   (a) For the first round, compute the following:
   (i) Euclidean distance between '10' and Centroid_1;
   (ii) Euclidean distance between '10' and Centroid_2;
   (iii) Hence, state the cluster for which '10' is assigned to and explain why.

   (b) After the first round, numbers '2', '3' and '4' are assigned to Cluster 1, while numbers '10', '11', '12', '20', '25' and '30' are assigned to Cluster 2.
   (i) Calculate the new centroid of Cluster 1.
   (ii) Calculate the new centroid of Cluster 2.

(c) The following is a graph of within-cluster sum of squared error (*WSSE*) vs *K* in the K-Means clustering process, which *K* is the best number of cluster? Explain in context.



4. [🖧] The Ecoli dataset (Resource: UCI Machine Learning Repository) contains protein localization sites. The meaning of the attributes are given as follows:

Attribute	**Meaning**
SeqName	Accession number for the SWISS-PROT database
mcg	McGeoch's method for signal sequence recognition
gvh	von Heijne's method for signal sequence recognition
lip	von Heijne's Signal Peptidase II consensus sequence score
chg	Presence of charge on N-terminus of predicted lipoproteins
aac	Score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins
alm1	Score of the ALOM membrane spanning region prediction program
alm2	Score of ALOM program after excluding putative cleavable signal regions from the sequence
Class	Localization site ('cp' (cytoplasm), 'im' (inner membrane), 'pp' (perisplasm), 'om' (outer membrane))

Perform K-Means clustering on the Ecoli dataset to uncover the clusters of its localization sites. The data is to be normalized using min-max (0.0-1.0) and random initialization of initial centroids using the seed '1234' to create the K-means model.

(a) What is the value of *k* to be used for the K-means model? Explain why.
(b) Explain why normalization is important.
(c) Write down the resulting cluster centroids for each cluster.
(d) Which cluster group contains the most number of records? State the number of records for that group.
(e) Based on the model created in (a), write down the cluster ID assigned to the following record:

	D mcg	D gvh	D lip	D chg	D aac	D alm1	D alm2
Row0	0.6	0.8	0.48	0.5	0.5	0.3	0.44

(f) Based on the parallel coordinates visualization for the cluster result, comment on the separability of 'cp' and 'im', and state any attributes that characterize the cluster.

5. [🖧] The wine data set (https://archive.ics.uci.edu/ml/datasets/wine) contains the results of a chemical analysis of wines grown in a specific area of Italy. It was suspected that three types of wine are represented in the 178 samples, with the results of 13 chemical analyses recorded for each sample. As the range of each attribute is different, it is important to normalize all the attributes first. Use the min-max (0.0-1.0) normalization method for this purpose. For initialization of the centroid, use the first *k* rows as the initial centroids. Use K-Means to uncover the types of wine. Interpret your results.

ANSWERS

1.

(a)

Cluster 1	Cluster 2
Age = (41 + 38)/2 = 39.5	Age = (52 + 59 + 49)/3 = 53.3
Income = (115 + 81)/2 = 98 thousands	Income = (35 + 73 + 81)/3 = 63 thousands
CCAvg = (7 + 4)/2 = 5.5 thousands	CCAvg = (0.2 + 1.7 + 2.0)/3 = 1.3 thousands

(b) Cluster 1 – high-income, high-spenders,
Cluster 2 – middle-income, low-spenders

(c) Clustering was done on normalized data so as not to let any attribute dominate the distance calculation due to large numerical values.

(d) Cluster 1 centroid
Normalized age = -0.980
Normalized Income = 0.736
Normalized CCAvg = 0.962

Euclidean distance from new record to centroid of cluster 1

$$= \sqrt{(1.322-(-0.980))^2+(1.542-0.736)^2+(-0.0305-0.962)^2} = 2.633$$

Cluster 2 centroid
Normalized age = 0.649
Normalized Income = -0.491
Normalized CCAvg = -0.641

Euclidean distance from new record to centroid of cluster 2

$$= \sqrt{(1.322-0.649)^2+(1.542-(-0.491))^2+(-0.0305-(-0.641))^2} = 2.227$$

The customer is assigned to Cluster 2. That is because the distance is nearer.

(e) The customer has a high income that is twice the average income in that cluster. The customer's average credit card spending (CCAvg) is also more than twice the average spending in that cluster.

2.

(a) Cluster 1. That is because the average distance from centroid, the within-cluster sum of squares and the maximum distance from centroid are lower that that of cluster 2.

(b) Cluster 1.
To win a 400 m race, the faster the better, hence the shorter time the better. Cluster 1 has a negative centroid, which means it has a time that is smaller than average.
To win a long jump, the longer the better. Cluster 1 has a positive centroid for long jump, which means it has a length that is longer than the average.

(c)

Euclidean distance from cluster 1 $= \sqrt{(-0.768-(-0.6450))^2+(-0.221-0.6848)^2}$
$= 0.914$

Euclidean distance from cluster 2 $= \sqrt{(-0.768-0.6772)^2+(-0.221-(-0.7190))^2}$
$= 1.529$

The record is assigned to cluster 1 due to a lower distance.

3. (a)
(i) 6
(ii) 2
(iii) Cluster_2, because distance to centroid_2 is smaller.
(b)

(i) Centroid of Cluster $1 = \frac{2+3+4}{3} = 3$

(ii) Centroid of Cluster $2 = \frac{10+11+20+25+30+12}{6} = 18$

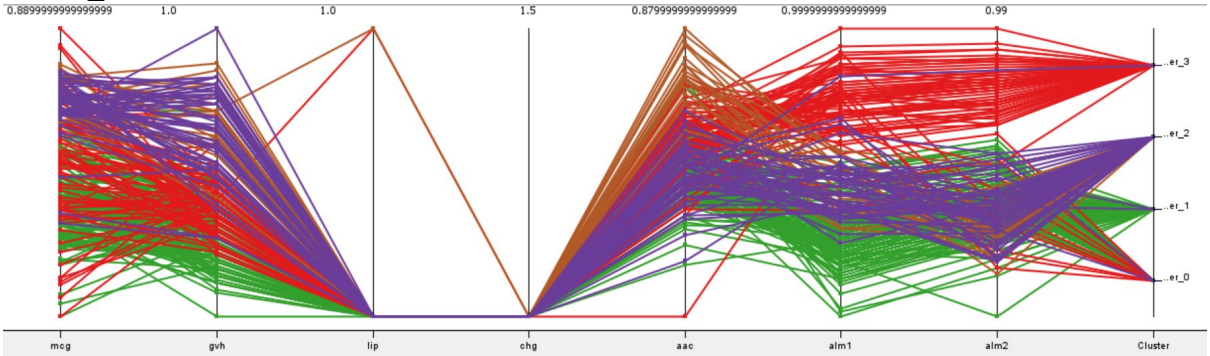
(c) k = 3. This is because from the plot, the rate of decrease of WSSE sharply shifts when k = 3 in the elbow plot. Hence, the best k is 3.

4.

(a) K is 4 because we are clustering into 4 classes.
(b) Normalization prevents any attribute with large numerical values from dominating the distance calculations.
(c)

Row ID	mcg	gvh	lip	chg	aac	alm1	alm2
cluster_0	0.45	0.375	0	0	0.487	0.448	0.463
cluster_1	0.368	0.269	0	0	0.526	0.25	0.367
cluster_2	0.754	0.656	0.014	0	0.592	0.444	0.347
cluster_3	0.577	0.398	0.015	0	0.632	0.765	0.793

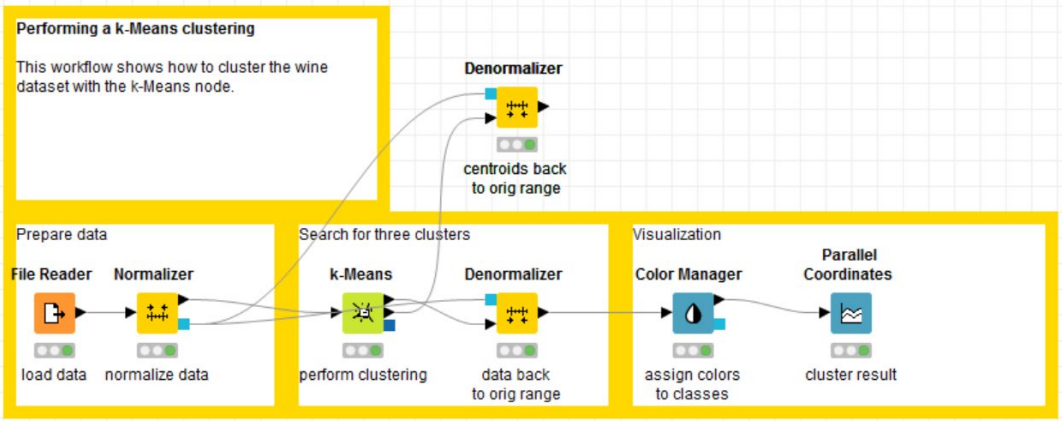
(d) cluster_1 with 102 records.
(e) cluster_2
(f)



Based on parallel coordinates plot, cp (green) is not well separated. Half of cp went to cluster 0 and the other half went to cluster 1. On the other hand, im seems to be well separated as majority went to cluster 3. cp cluster is characterised by lower than average alm1. im cluster is characterised by higher than average alm1 and alm2.

5.

Completed workflow for Q5 might appear as follows:

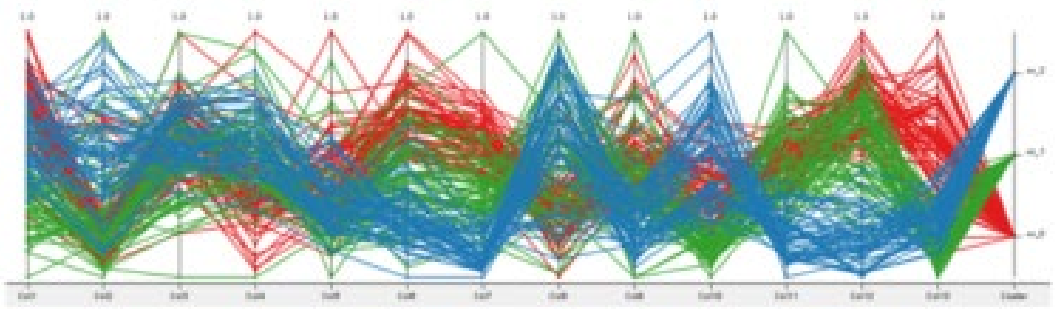


Interpretation of clustering result:

- The three wine types are characterized by the following profile of attribute values:

	Alcohol	Maclic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavonoids'	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue	OD280/OD315	Proline
cluster_0	0.691	0.238	0.576	0.353	0.398	0.65	0.555	0.291	0.478	0.348	0.481	0.688	0.572
cluster_1	0.309	0.238	0.476	0.495	0.255	0.421	0.358	0.451	0.378	0.142	0.469	0.561	0.16
cluster_2	0.554	0.507	0.566	0.549	0.312	0.243	0.101	0.607	0.232	0.508	0.172	0.156	0.243

- From the parallel coordinates, we observe that one of the cluster group (blue) is well-separated from the other two groups (green and red).



- We observe that for example, the cluster group_2 (blue) has low values of flavonoids (Col7), hue (Col11) and OD280/OD315 of diluted wines (Col12) compared to the other two groups.

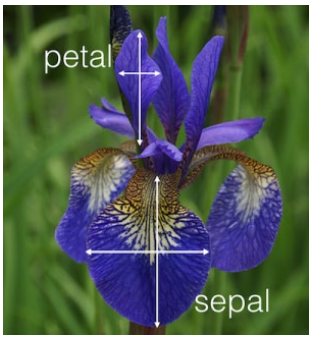
LAB 6 : K-Means Clustering using KNIME

Learning Objectives:

1. *Build a K-Means clustering model using KNIME.*
2. *Interpret cluster result through the use of visualization plots in KNIME.*

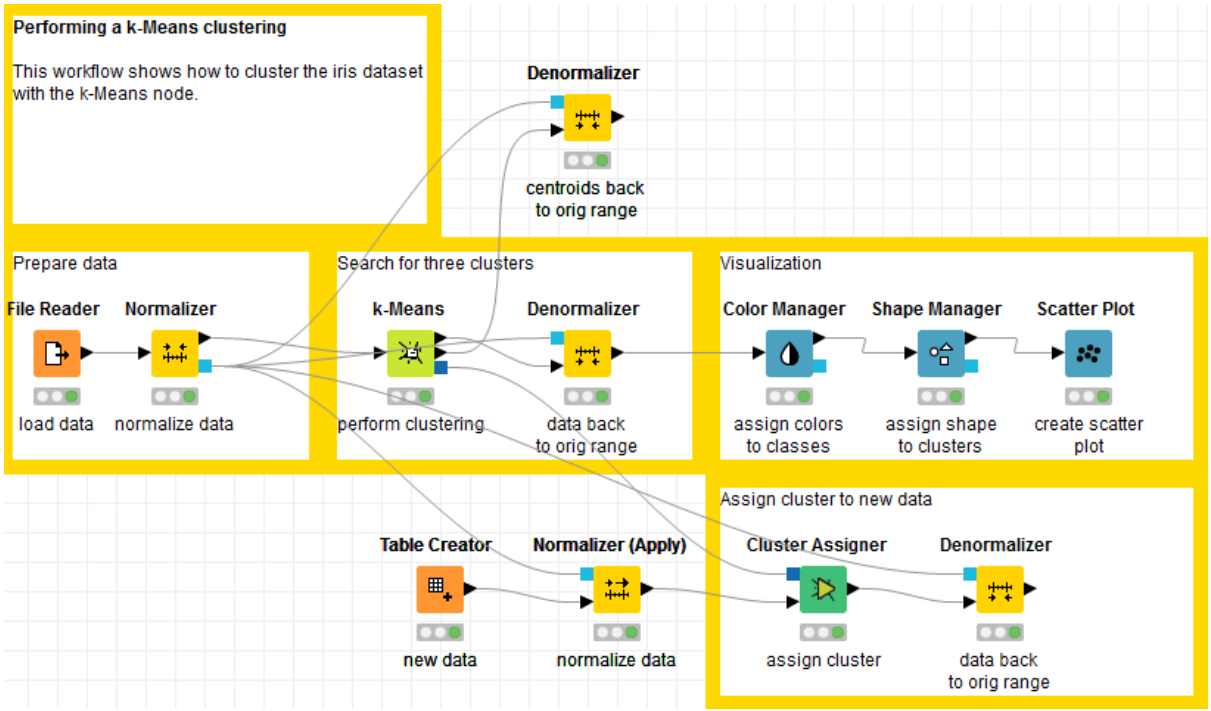
Task: The Iris Dataset

The iris flower dataset *iris.csv* is a multivariate dataset first introduced by the British statistician and biologist Ronald Fisher in 1936. The dataset consists of 50 samples for each of the three species of Iris (*Iris setosa*, *Iris versicolor*, and *Iris virginica*). Four attributes were measured for each sample: *sepal length*, *sepal width*, *petal length*, and *petal width* (all in cm).



Problem: We want to label each measured flower as some species of iris. For the purpose of this lab, let us ignore the 'class' column and assume we don't know, a priori, the class of the plant. Hence, we want to label each flower as one of the three species by using the four attributes: *sepal length*, *sepal width*, *petal length*, and *petal width* only.

The completed workflow is as follows:

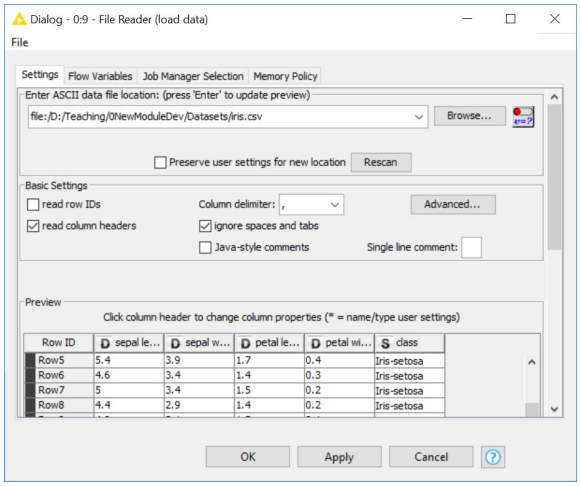


A. Create a KNIME Workflow

1. Launch KNIME.
2. In the upper menu bar, choose File > New ….
3. Choose New KNIME Workflow and click Next >.
4. Give the workflow a name.
5. Use the default destination LOCAL to save the workflow and click Finish.

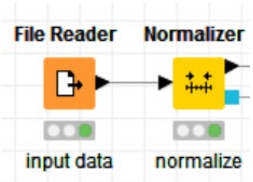
B. Import the Dataset

1. Drag and drop the File Reader node (IO > Read) onto the Workflow Editor. Alternatively, you can search for this node by typing "File Reader" in the search box in the Node Repository.
2. Right-click on the File Reader and click Configure.
3. Click Browse and select the location of the dataset *iris.csv*. Click OK.



C. Prepare the data set

We need to normalize the data.



1. Search for the Normalizer node (Manipulation > Column > Transform > Normalizer). Drag it to the Workflow Editor.
2. Right-click on the Normalizer node and click Configure. Choose Min-Max Normalization and ensure the Min value is 0 and Max value is 1.0. Make sure that all '*sepal length*', '*sepal width*', '*petal length*' and '*petal width*' are included to Normalize.
3. Connect the Normalizer node to the File Reader node.
4. Right-click and select all the nodes in the workflow and choose Execute.

CHECKPOINT

"We want to label each flower as one of the three species by using the four attributes: *sepal length*, *sepal width*, *petal length*, and *petal width* only."

#1. Which kind of data mining task describe the above task: unsupervised, or supervised?

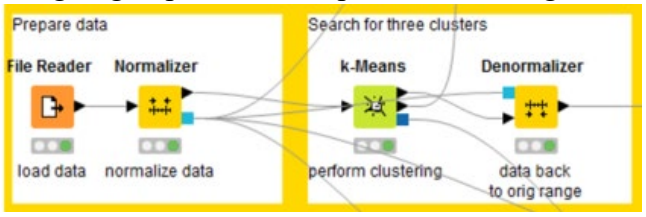
#2. What are the attributes used in the data mining task?

#3. Describe the data type of each of the attribute.

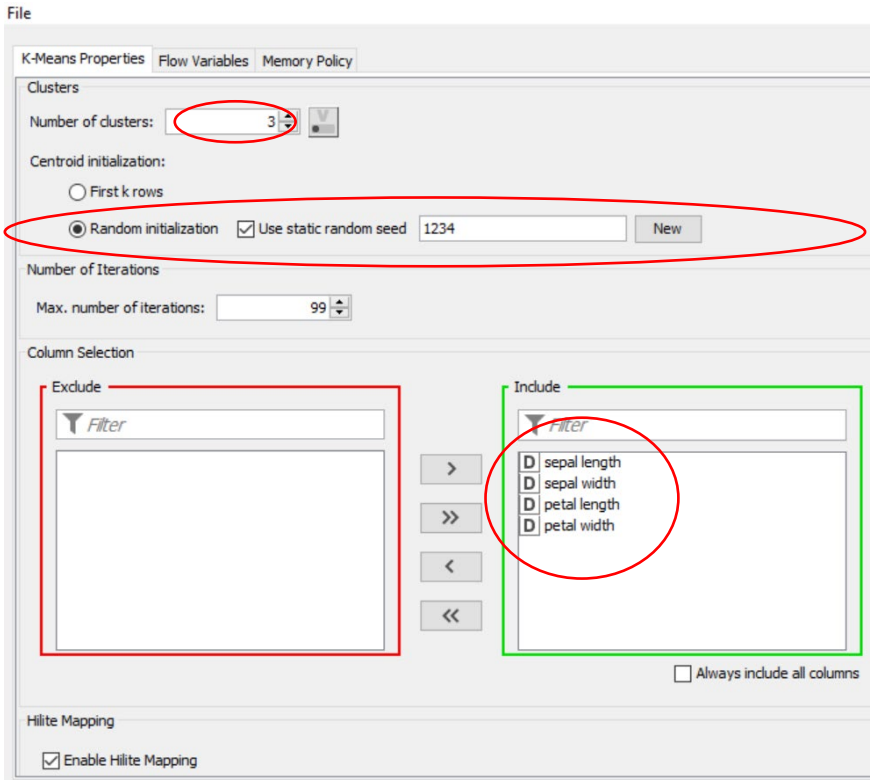
#4. Why do we need to perform normalization?

D. Clustering with K-Means

We are going to perform unsupervised learning on this dataset using K-Means algorithm.

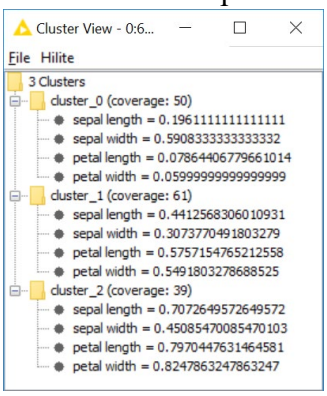


1. Create the workflow above with the indicated nodes. (k-Means can be found under Analytics > Mining > Clustering)
2. Right-click on the k-Means node and click Configure. Select the following parameters for k-Means as follows:



3. Right-click and select all the nodes in the workflow and choose Execute.

4. We can connect the Denormalizer node to the k-Means node by parsing the model computed from the Normalizer node to return the data back to original range for both the 'Labeled input' and 'Clusters' output from k-Means. Execute.
5. Right-click and choose View: Cluster View. Expand the list to show the centroid.



CHECKPOINT

K-Means node configuration:

#1. Why do we set the number of clusters to be 3?

#2. Which setting in the configuration is for Step 2 of the K-Means clustering algorithm "Randomly select k points as initial cluster centers"? Do you think the centroid initialization will affect the cluster result?

#3. What are the attributes used for clustering?

Explore the output ports of the K-Means node to answer the following:

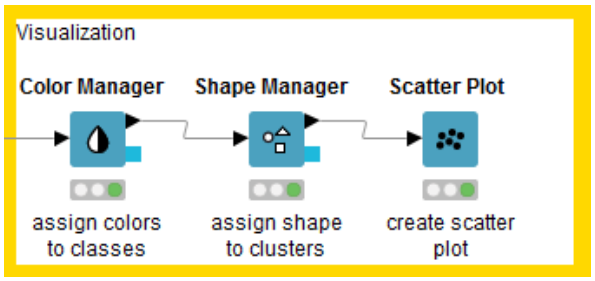
#4. What are the centroids of the *k* resulting clusters?

#5. How many records are there in each of the *k* resulting clusters?

#6. Identify which cluster is the following records allocated to: Row 35, Row 70, Row 130.

E. Visualize the cluster result

We use the scatter plot to visualize the cluster result.
The workflow is as follows:



1. Search for the Color Manager node (Views > Property). Drag it to the Workflow Editor.
2. Connect the Color Manager node to the K-Means output. You may use the default setting of the Color Manager node. Select 'class' to assign colors to class.
3. Search for the Shape Manager node (Views > Property). Drag it to the Workflow Editor.
4. Connect the Shaper Manager node after the Color Manager node. You may use the default setting of the Shape Manager node. Select 'Cluster' to assign shapes to cluster.
5. Search for the Scatter Plot node (Views). Drag it to the Workflow Editor.
6. Connect the Scatter Plot node to the Shape Manager node. You may use the default setting of the Scatter Plot node.
7. Right-click and select all the nodes in the workflow and choose Execute.
8. Right-click and choose: View: Scatter Plot. Go to the 'Column Selection' tab to explore the different scatterplots of different pairs of attributes.

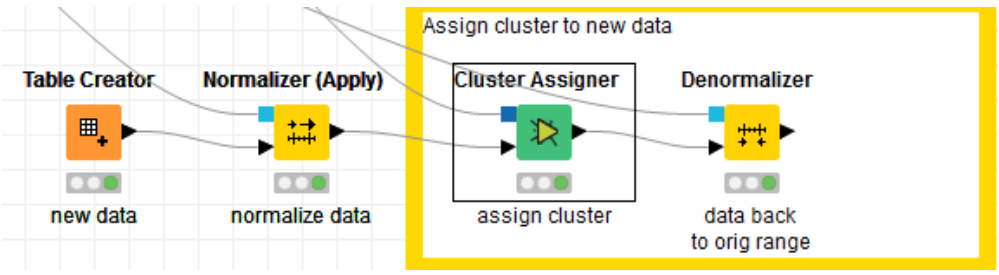
Investigative tasks

Explore the following pair of attribute in a scatterplot and interpret the cluster result:

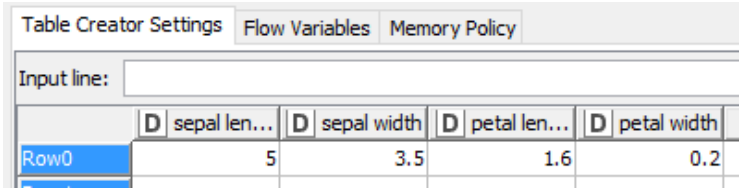
(a) X Column: sepal length Y Column: sepal width
(b) X Column: petal length Y Column: petal width

F. Assign cluster to new data

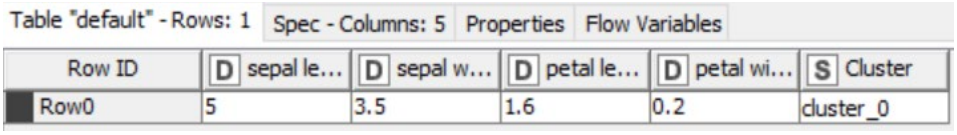
We use the Cluster Assigner to assign cluster to new data. The workflow is as follows:



1. Use the Table Creator node to create a new data record as follows:



   Note: Ensure that the name of the variable tallies with that of the dataset used to perform clustering.
2. Connect the Normalizer (Apply) node to the Table Creator node where the model from the Normalizer is connected to the first input port and table created from Table creator node as the table input port.
3. Connect the Cluster Assigner node with the k-Means model and the normalized data from the Normalizer (Apply) node.
4. Connect the Denormalizer node to return the data back to original range.
5. Right-click and select all the nodes in the workflow and choose Execute.
6. Right-click and choose Denormalized output. The result will be as follows:



CHECKPOINT

#1. Suppose a new Iris plant has the measurements sepal length=5.0 cm, sepal width=3.5 cm, petal length =1.6 cm, and petal width=0.2 cm.

Which of the three clusters is the new plant most likely to belong to?