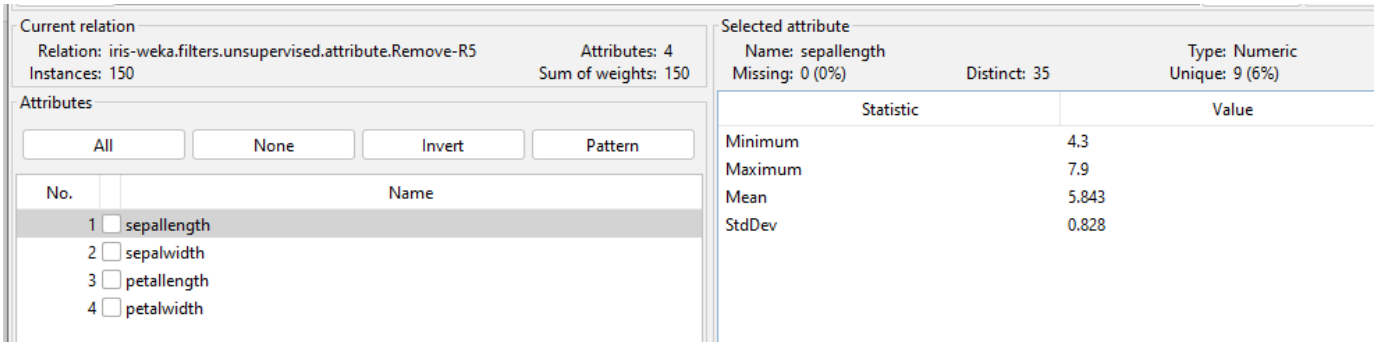
E/17/153 : Part 3 – Clustering

1. There are 150 instances and 4 attributes after removing the class attribute.



3. Seed

Seed means the growing point of the cluster. In seed-based clustering techniques, it is important to choose an appropriate seed. The performance of seed based algorithms are dependent on initial cluster center selection and the optimal number of clusters in a dataset. K-means is a widely used such algorithm and it is sensitive to initial seed selection of cluster centers.

4. Observations

```
Number of iterations: 7
Within cluster sum of squared errors: 12.143688281579722
```

```
Final cluster centroids:
                        Cluster#
Attribute      Full Data          0          1
                 (150.0)    (100.0)     (50.0)
=============================================
sepallength       5.8433      6.262      5.006
sepalwidth         3.054      2.872      3.418
petallength       3.7587      4.906      1.464
petalwidth        1.1987      1.676      0.244
```

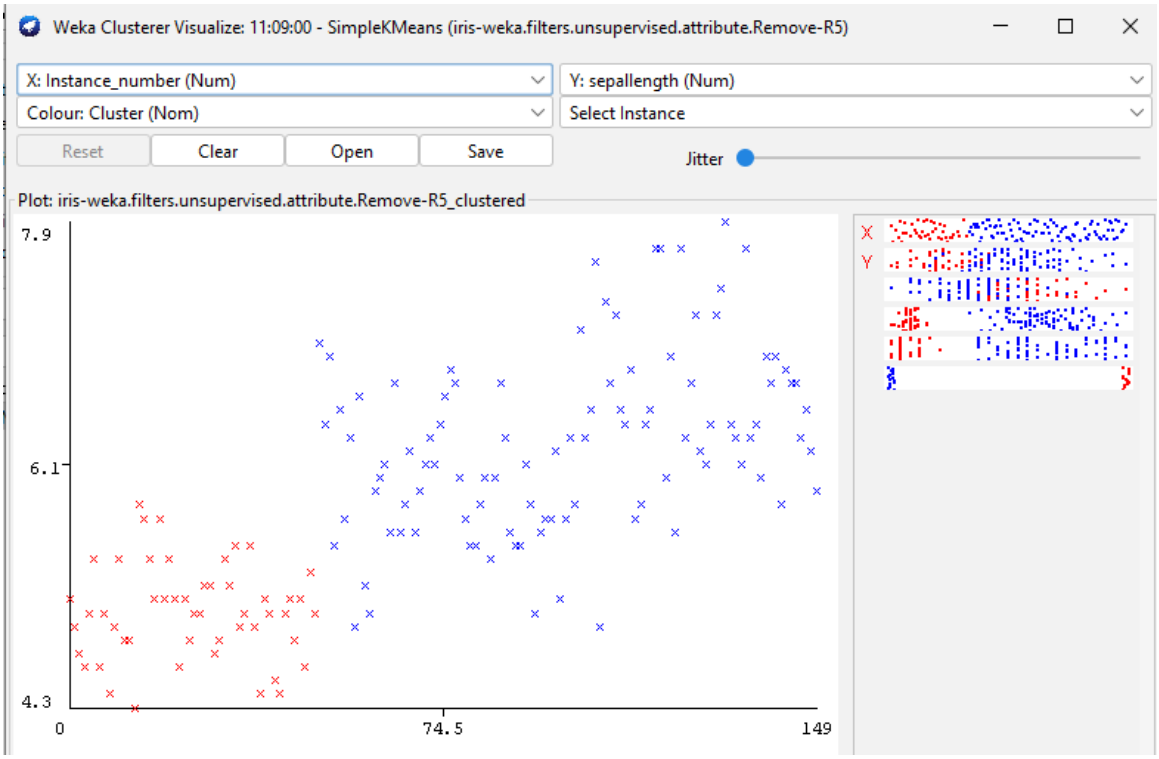
```
Clustered Instances

0       100 ( 67%)
1        50 ( 33%)
```

100 out if 150 instances are clustered to cluster 0 and 50 are categorized to cluster 1.

Each cluster centroid is represented by a mean vector. This vector can be used to describe a cluster.

5. Cluster visualization

As shown in the figure, data points were clustered into two clusters. They are visualized in red and blue points. Instance number and sepal length has been taken as the x axis and y axis respectively.



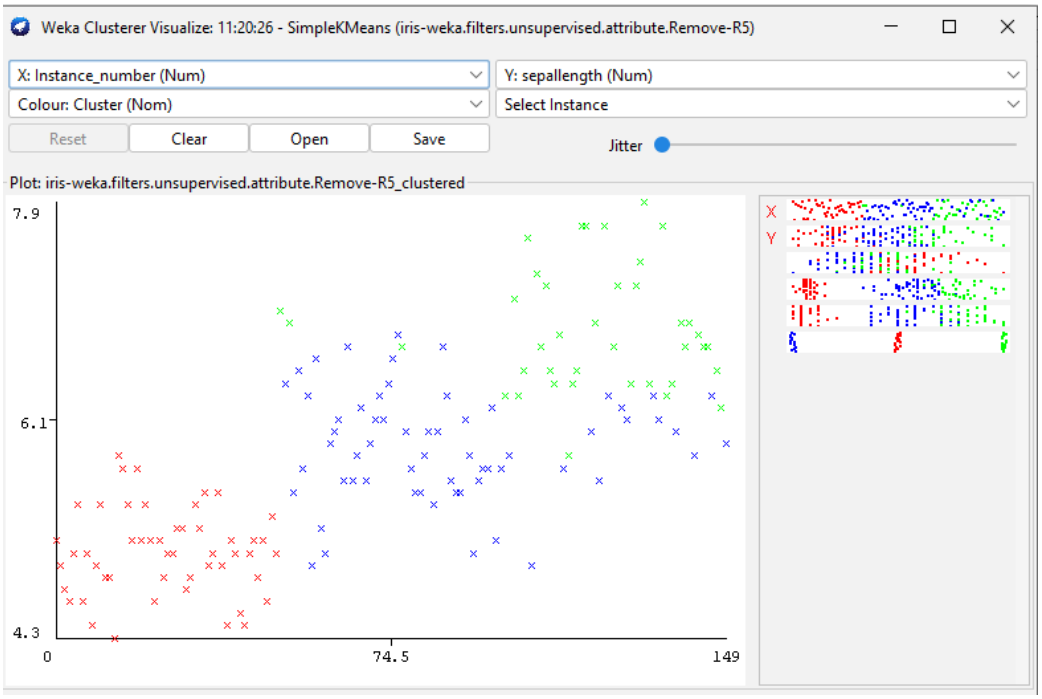
6. Contents of the ARFF file

It is generally made of two parts. The first part describes the data structure, that is to say the rows which begin by @attribute and the second part comprises the raw data, which follows the expression @data

7. When k =3

```
Number of iterations: 6
Within cluster sum of squared errors: 6.998114004826762
```

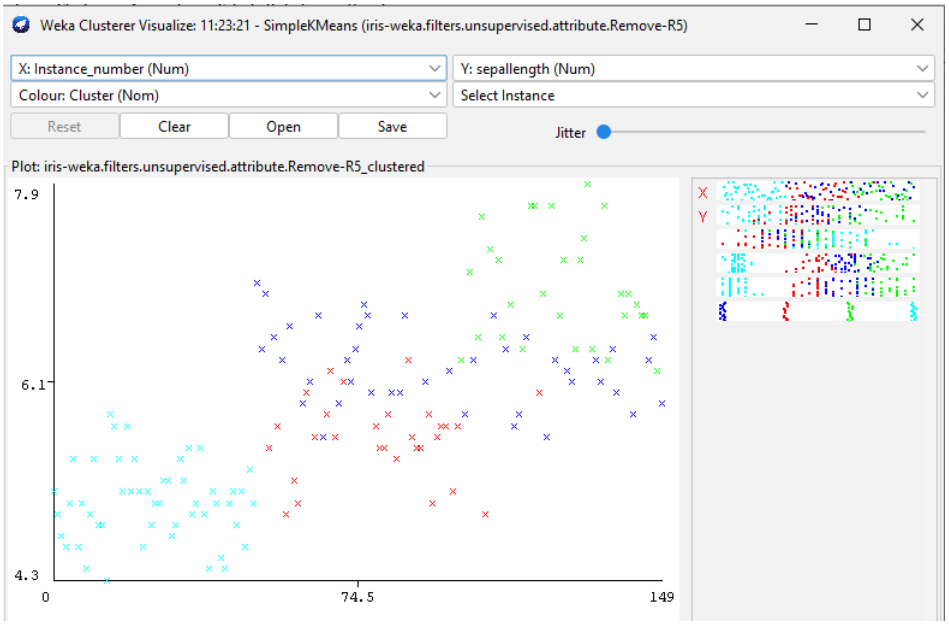
```
                           Cluster#
Attribute      Full Data          0          1          2
                 (150.0)     (61.0)     (50.0)     (39.0)
=============================================================
sepallength       5.8433     5.8885      5.006     6.8462
sepalwidth         3.054     2.7377      3.418     3.0821
petallength       3.7587     4.3967      1.464     5.7026
petalwidth        1.1987      1.418      0.244     2.0795



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)
```



When k = 4

```
Number of iterations: 4
Within cluster sum of squared errors: 5.532831003081898
```

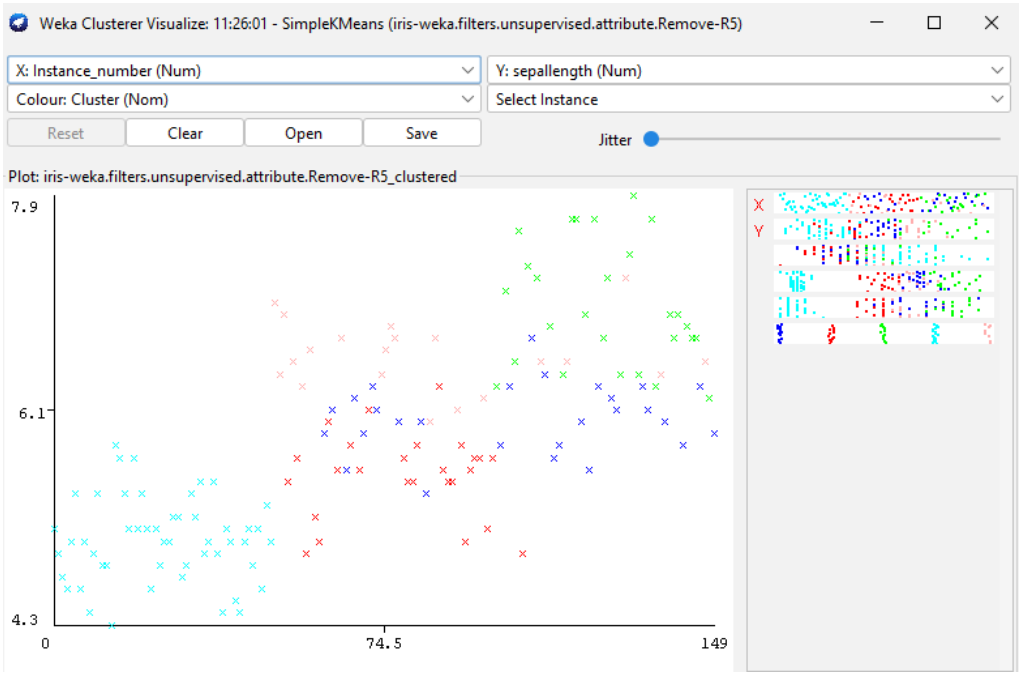
```
Attribute       Full Data          0          1          2          3
                  (150.0)     (42.0)     (29.0)     (29.0)     (50.0)
============================================================================
sepallength       5.8433       6.25     5.5828     6.9586      5.006
sepalwidth         3.054        2.9      2.569     3.1345      3.418
petallength       3.7587     4.8738     4.0034     5.8552      1.464
petalwidth        1.1987     1.6405      1.231     2.1724      0.244



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       42 ( 28%)
1       29 ( 19%)
2       29 ( 19%)
3       50 ( 33%)
```



When k = 5

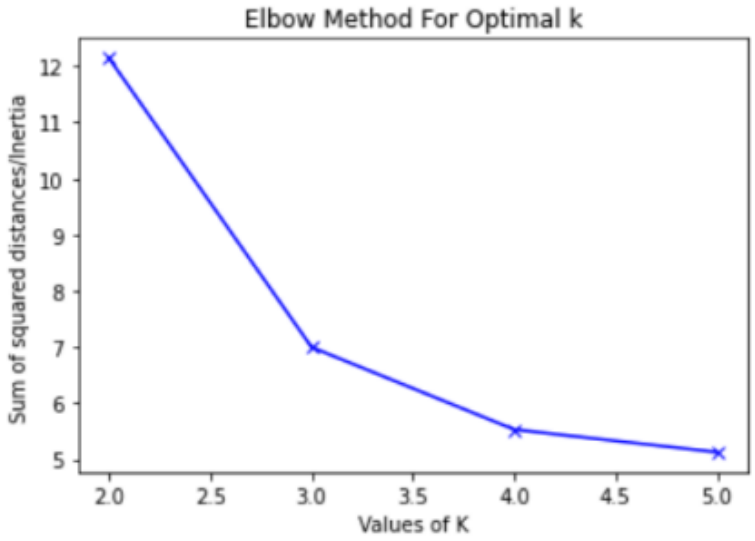
```
Number of iterations: 9
Within cluster sum of squared errors: 5.130784647061167
```

```
                         Cluster#
Attribute      Full Data          0          1          2          3          4
                 (150.0)     (27.0)     (26.0)     (27.0)     (50.0)     (20.0)
=============================================================================
sepallength       5.8433     6.0296       5.55     6.9667      5.006       6.55
sepalwidth         3.054     2.7556     2.5808      3.137      3.418       3.05
petallength       3.7587     4.9444     3.9269     5.8852      1.464      4.805
petalwidth        1.1987     1.7037        1.2        2.2      0.244       1.55



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       27 ( 18%)
1       26 ( 17%)
2       27 ( 18%)
3       50 ( 33%)
4       20 ( 13%)
```



K value	Within cluster sum of squared errors
2	12.14368828
3	6.998114004
4	5.532832003
5	5.130784647

The elbow method is a well-known method for determining the optimal k value. When cluster sun of squared errors is plotted against the values of k, there we can see an elbow shape.



This shows the optimal k value as 3 and as we already know there are 3 real clusters for the dataset, it is exact that optimal k value is 3.

8.

```
Number of iterations: 6
Within cluster sum of squared errors: 6.998114004826762
```

Out of 150 instances, 61 was clustered into cluster 0, 50 into cluster 1 and 39 into cluster 2. This gives 41%, 33% and 26% of percentages respectively.

```
                         Cluster#
Attribute      Full Data          0          1          2
                 (150.0)     (61.0)     (50.0)     (39.0)
=========================================================
sepallength       5.8433     5.8885      5.006     6.8462
sepalwidth         3.054     2.7377      3.418     3.0821
petallength       3.7587     4.3967      1.464     5.7026
petalwidth        1.1987      1.418      0.244     2.0795



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)
```

Then a class value is assigned to each cluster. Class Iris versicolor id assigned for cluster 0, Iris-setosa is assigned to cluster 1 and Iris-verginica is assigned to cluster 2. All instances of Iris-setosa have been correctly classified. But 3 out of 50 Iris-versicolor instances have been incorrectly categorized to Iris-virginica. 14 out of 50 instances of Iris-verginica has been incorrectly classified to Iris-versicolor. Hence, 17.0(11.333%) instances are incorrectly clustered.

```
Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  0 50  0 | Iris-setosa
 47  0  3 | Iris-versicolor
 14  0 36 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances :       17.0     11.3333 %
```