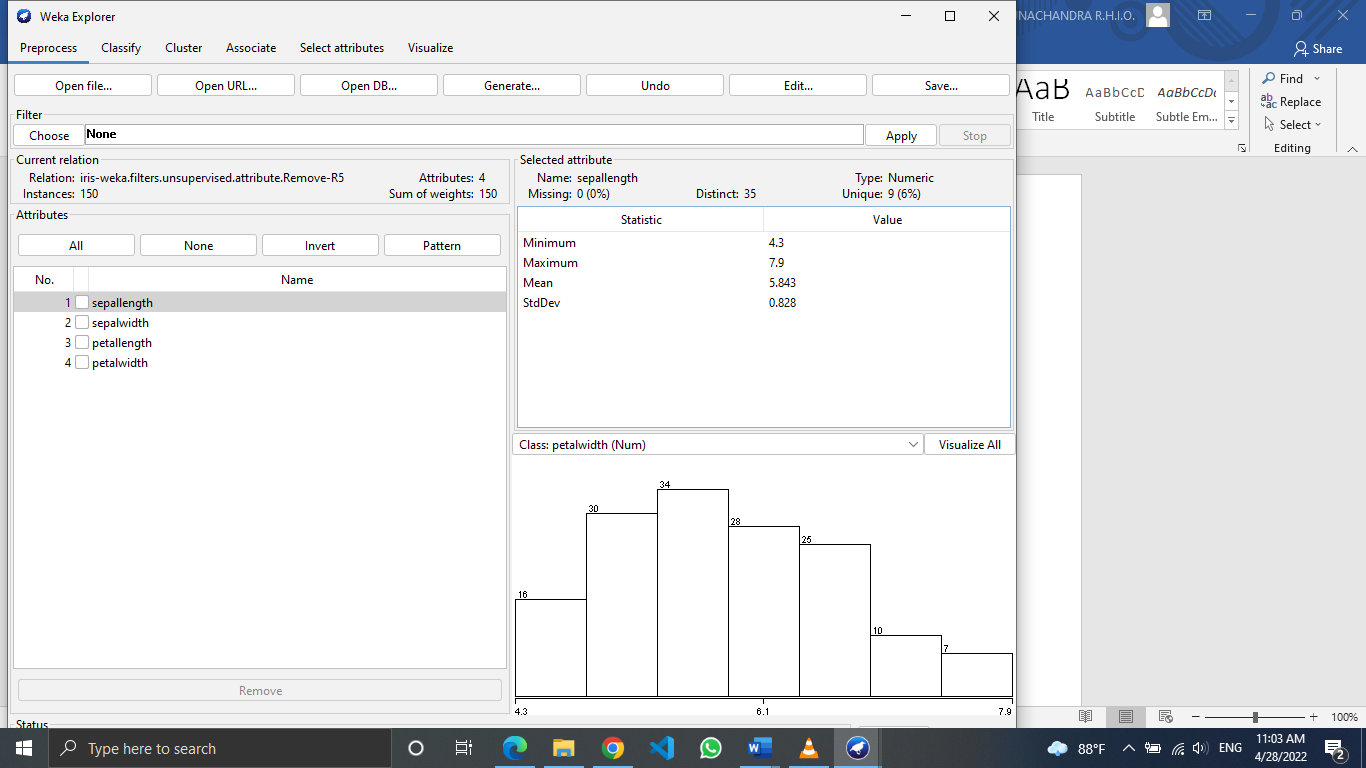
E/17/153 : Part 3 – Clustering

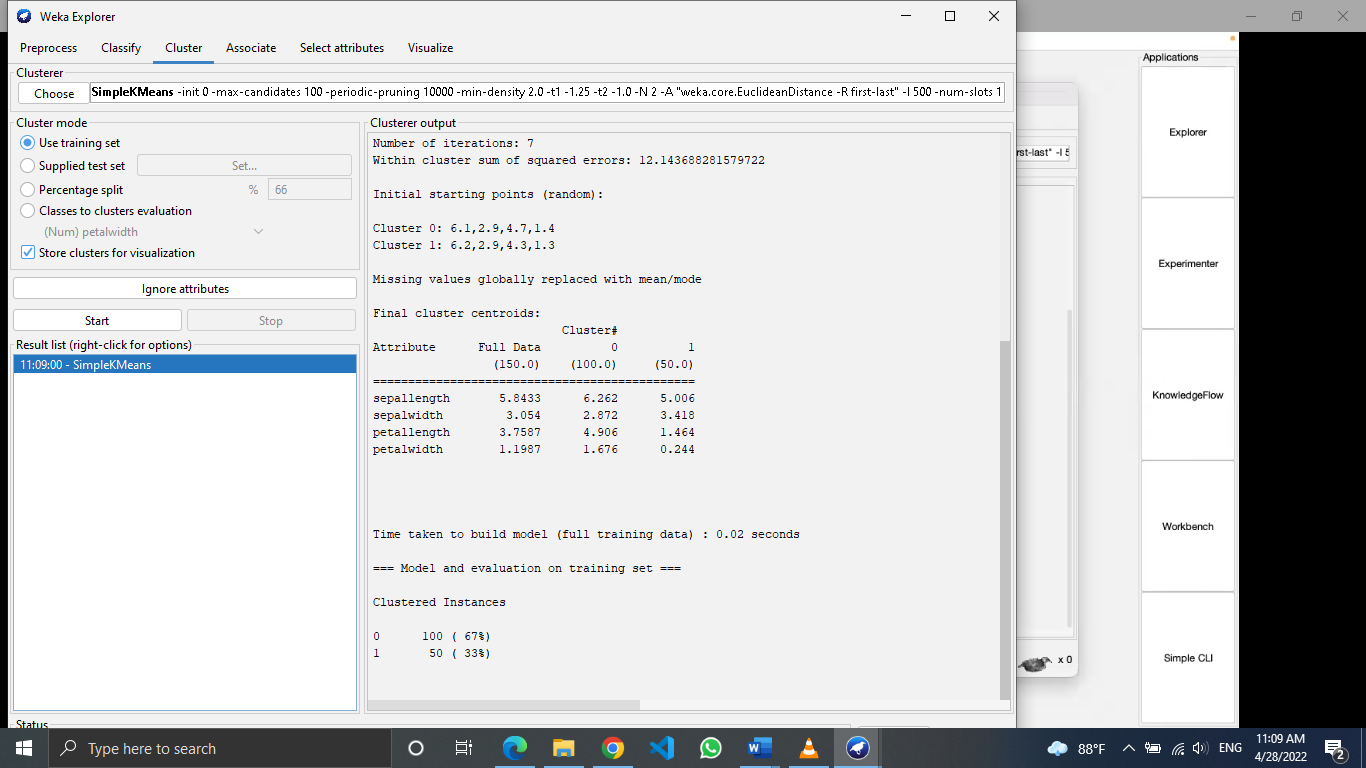
1. There are 150 instances and 4 attributes after removing the class attribute.

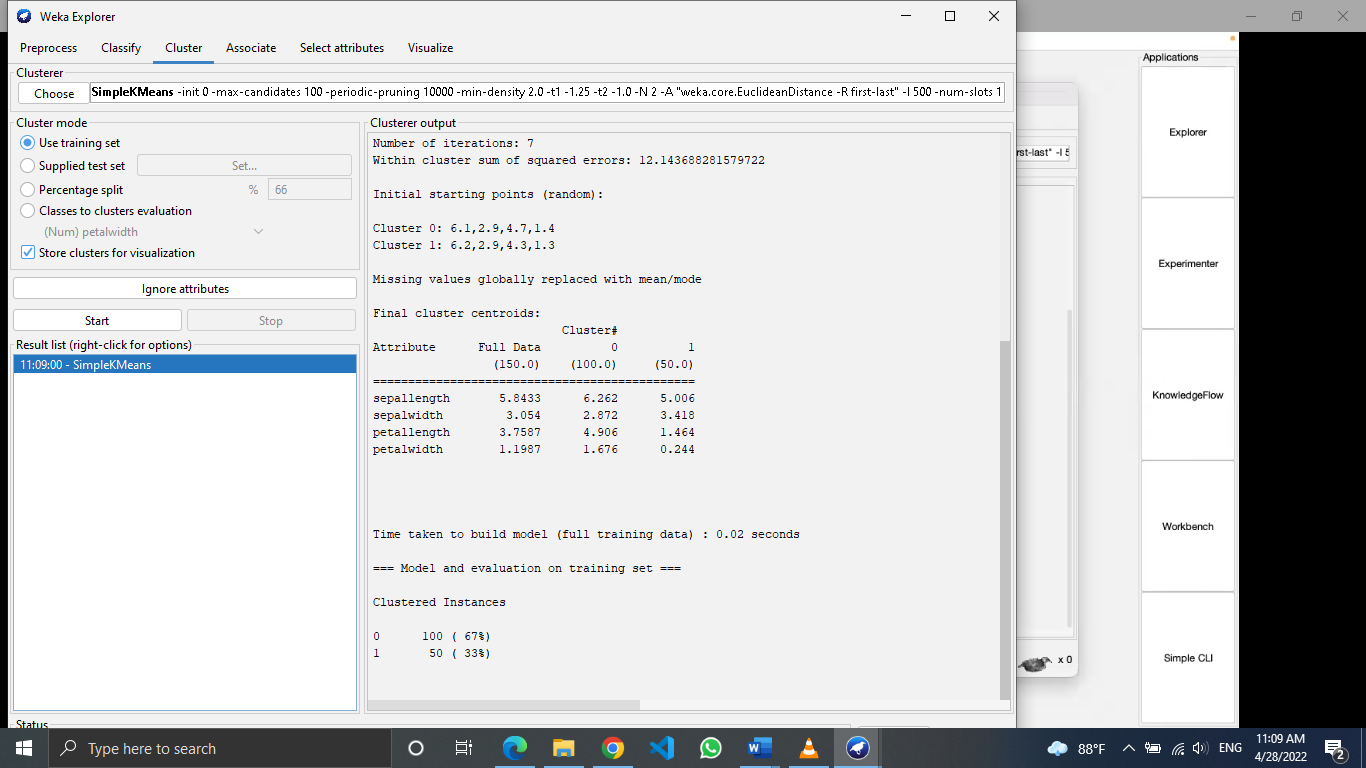


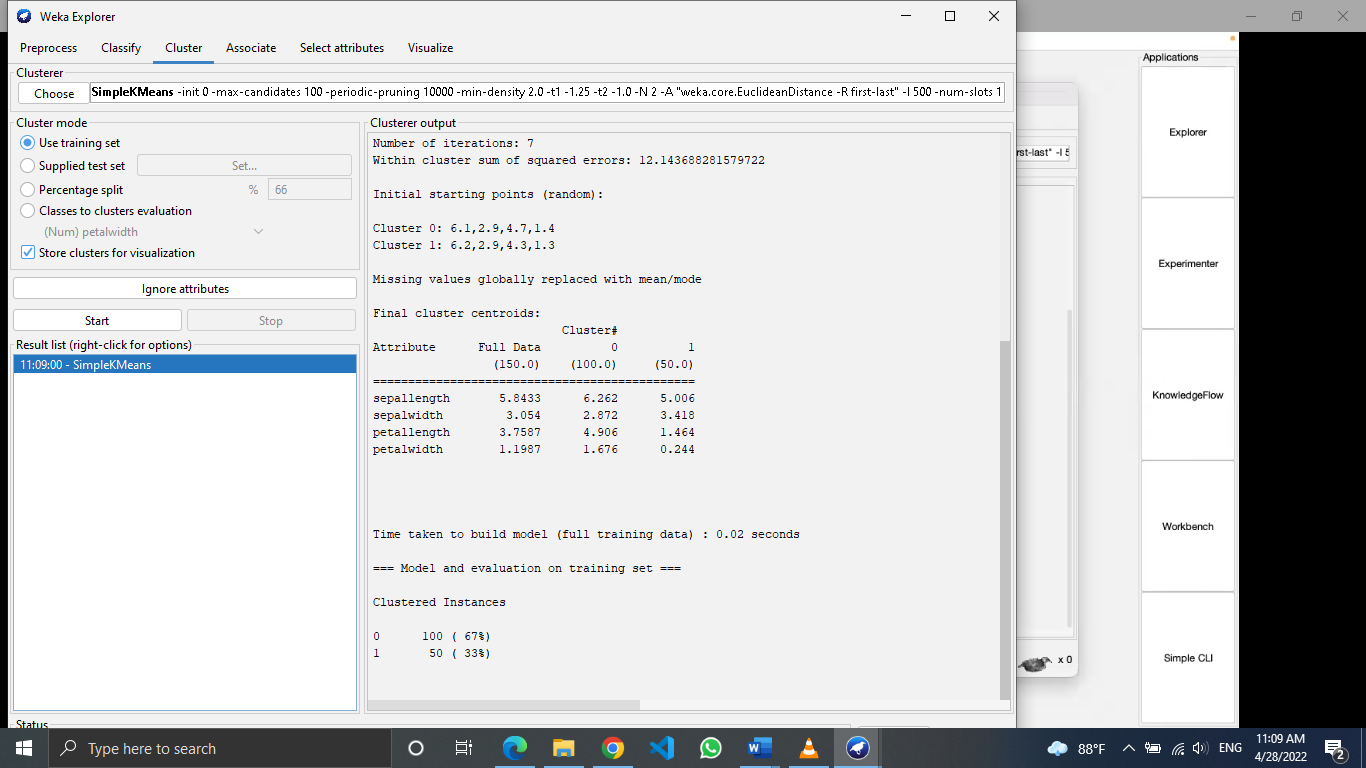
3. Seed

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





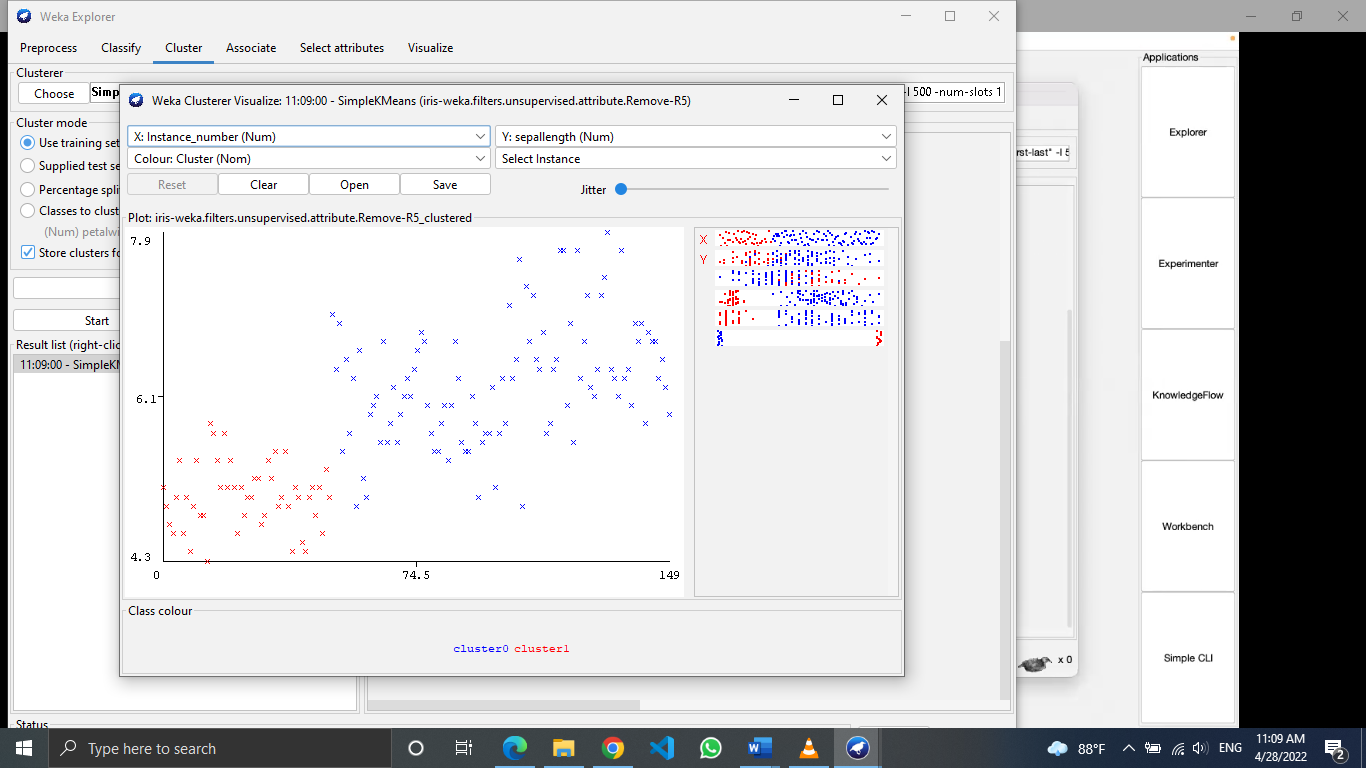


100 out if 150 instances are clustered to cluster 0 and 50 are categorized to cluster 1.

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

5. Cluster visualization

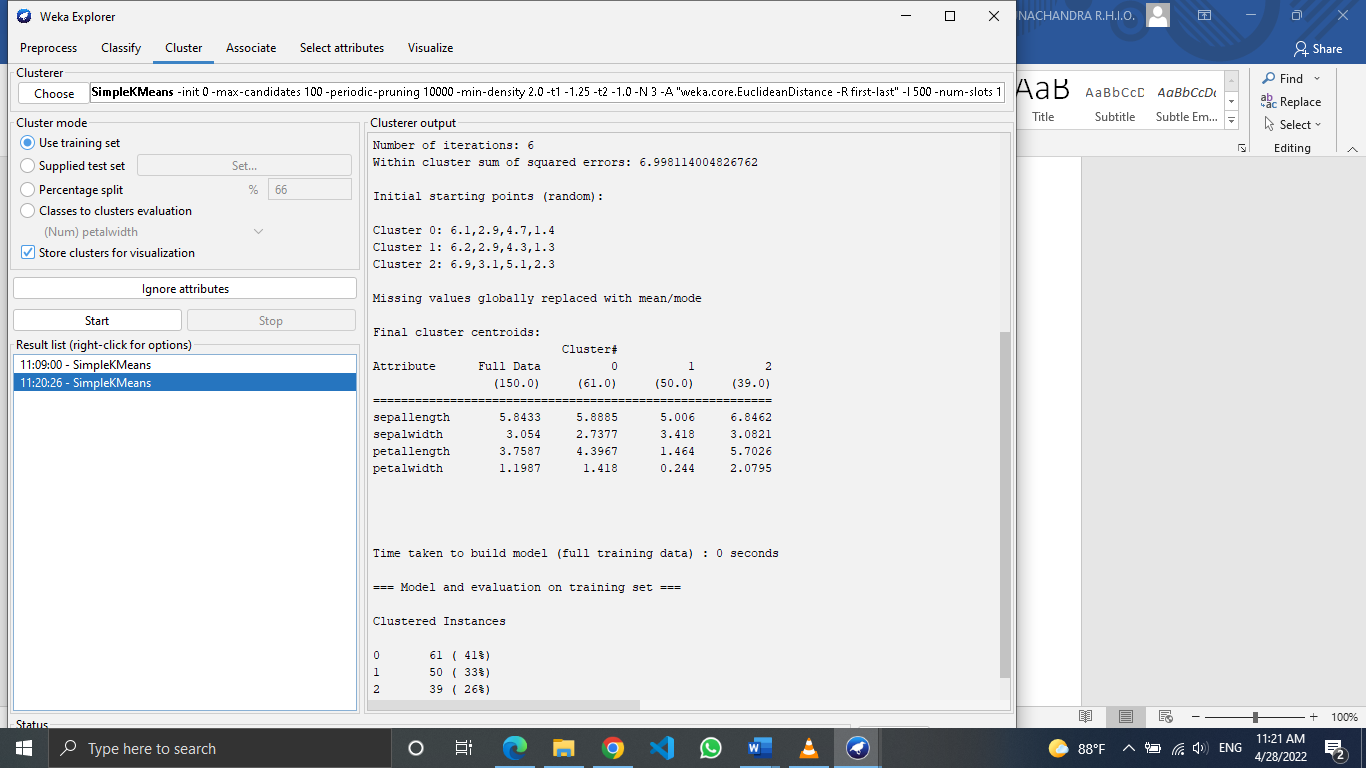
As shown in the figure, data points were clustered into two clusters. They are visualized in red and blue points. Instance number and sepal length has been taken as the x axis and y axis respectively.

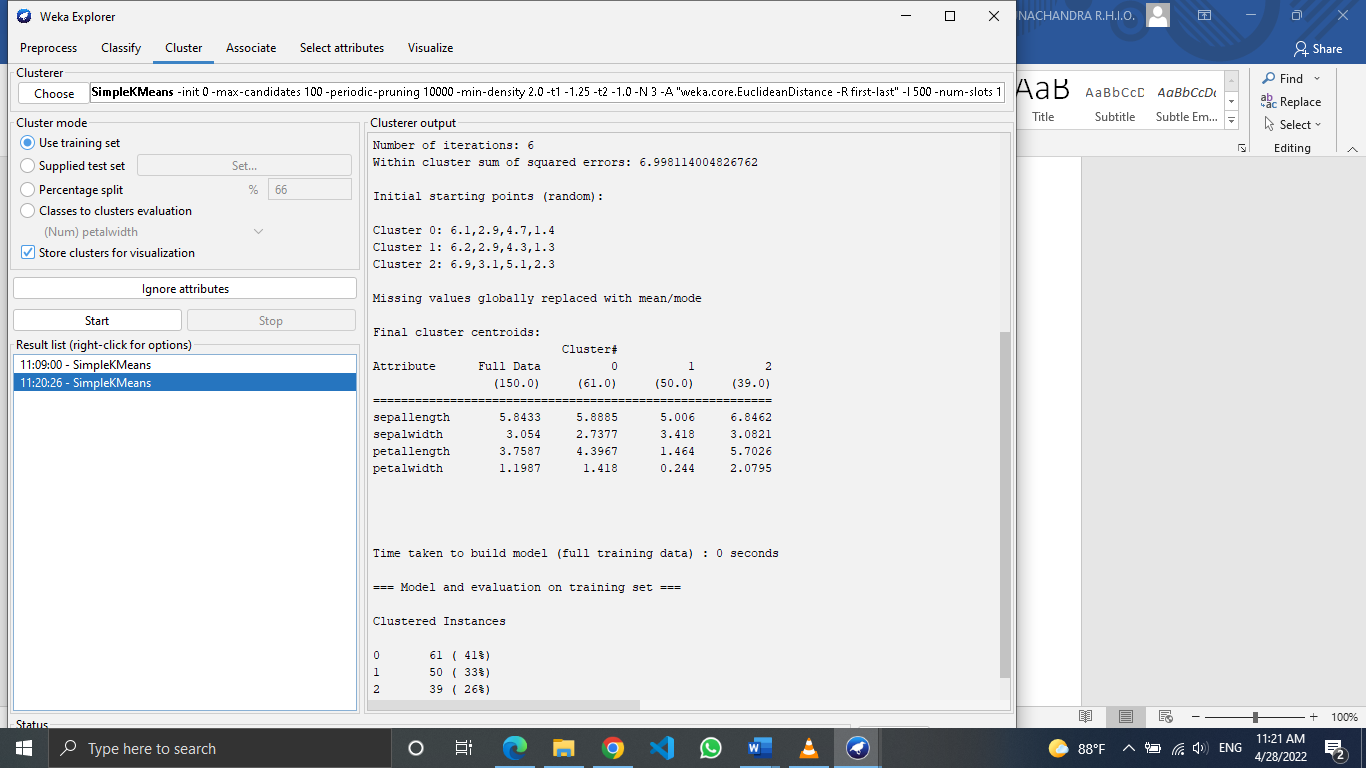


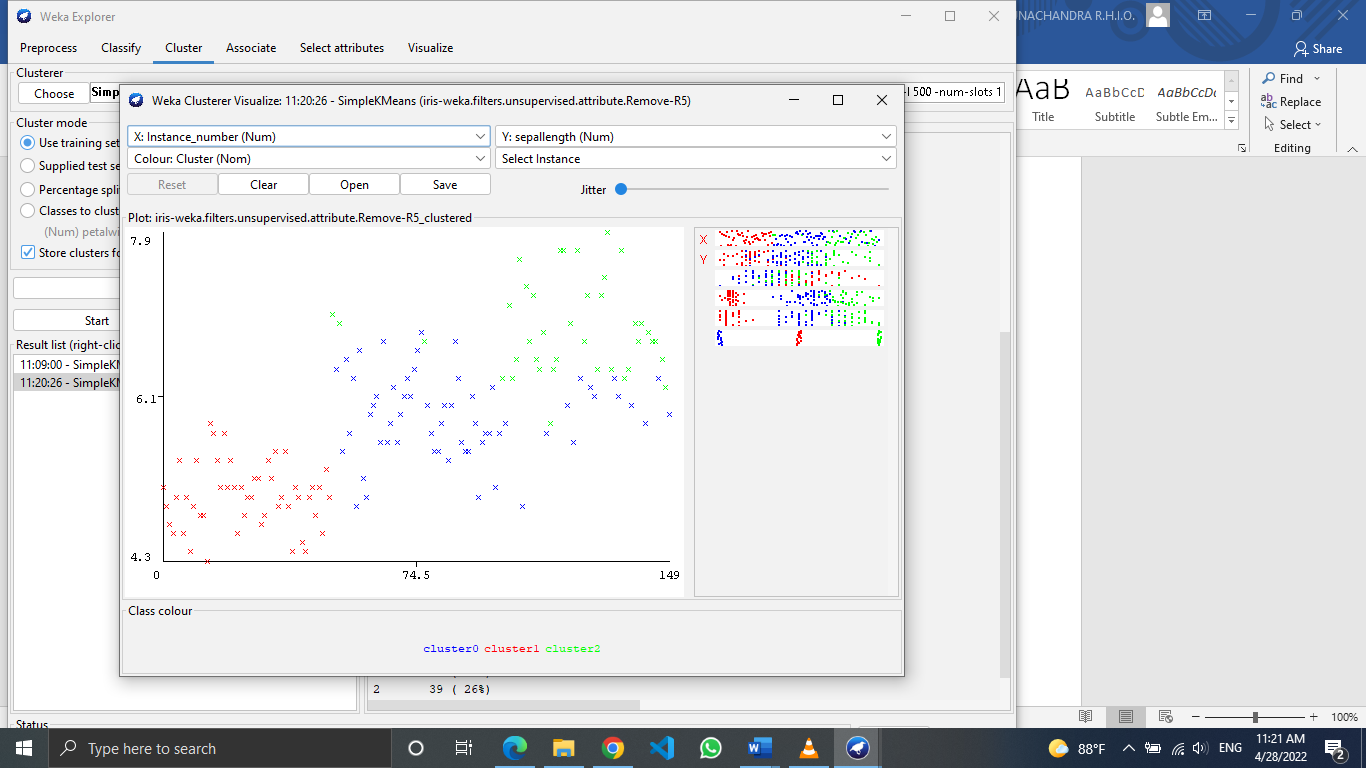
6. Contents of the ARFF file

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

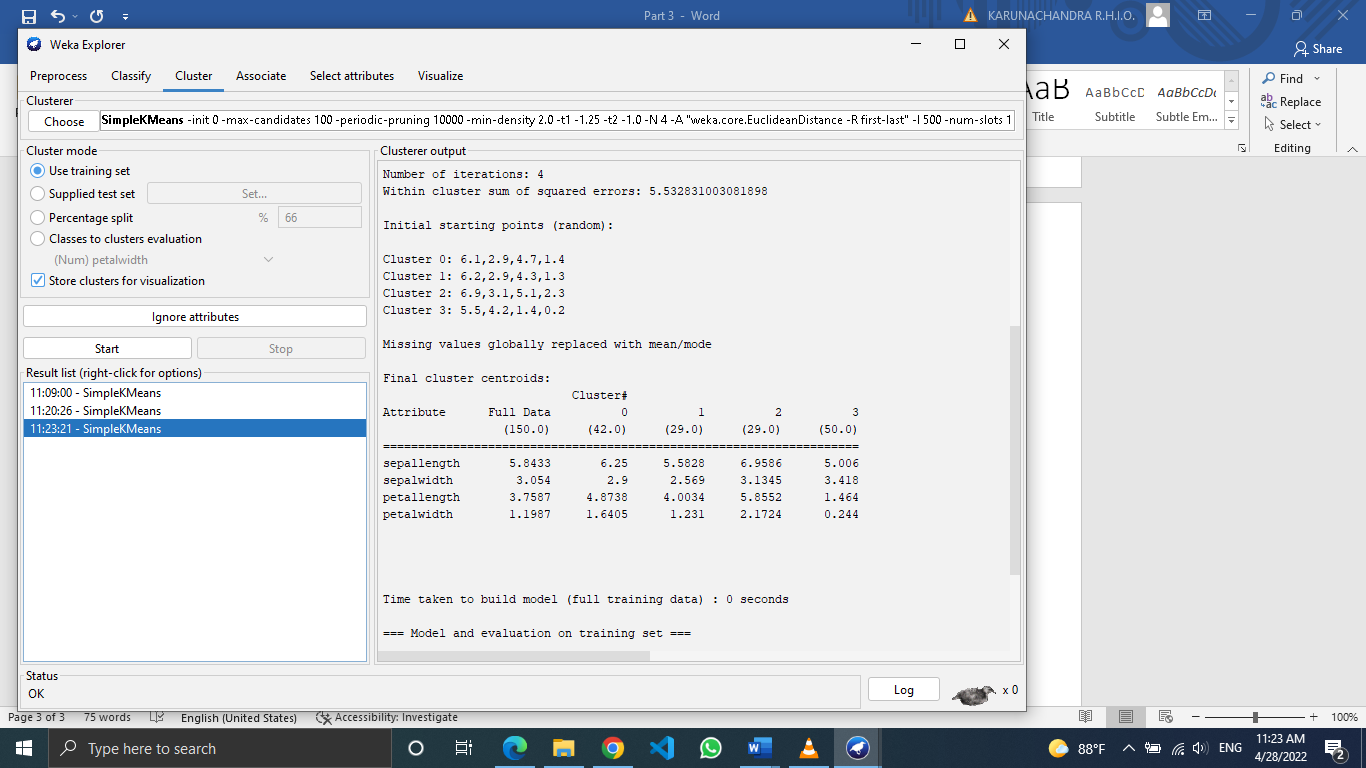
7. When k =3

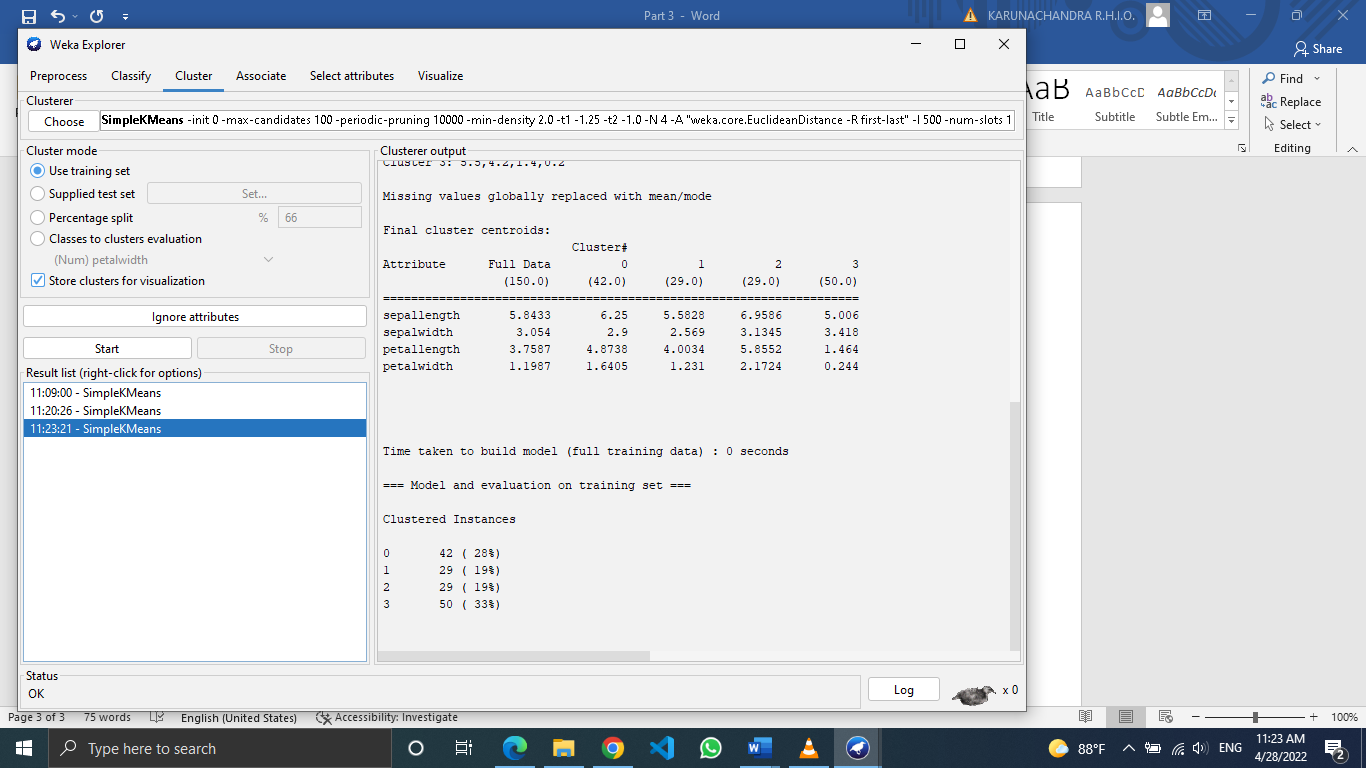


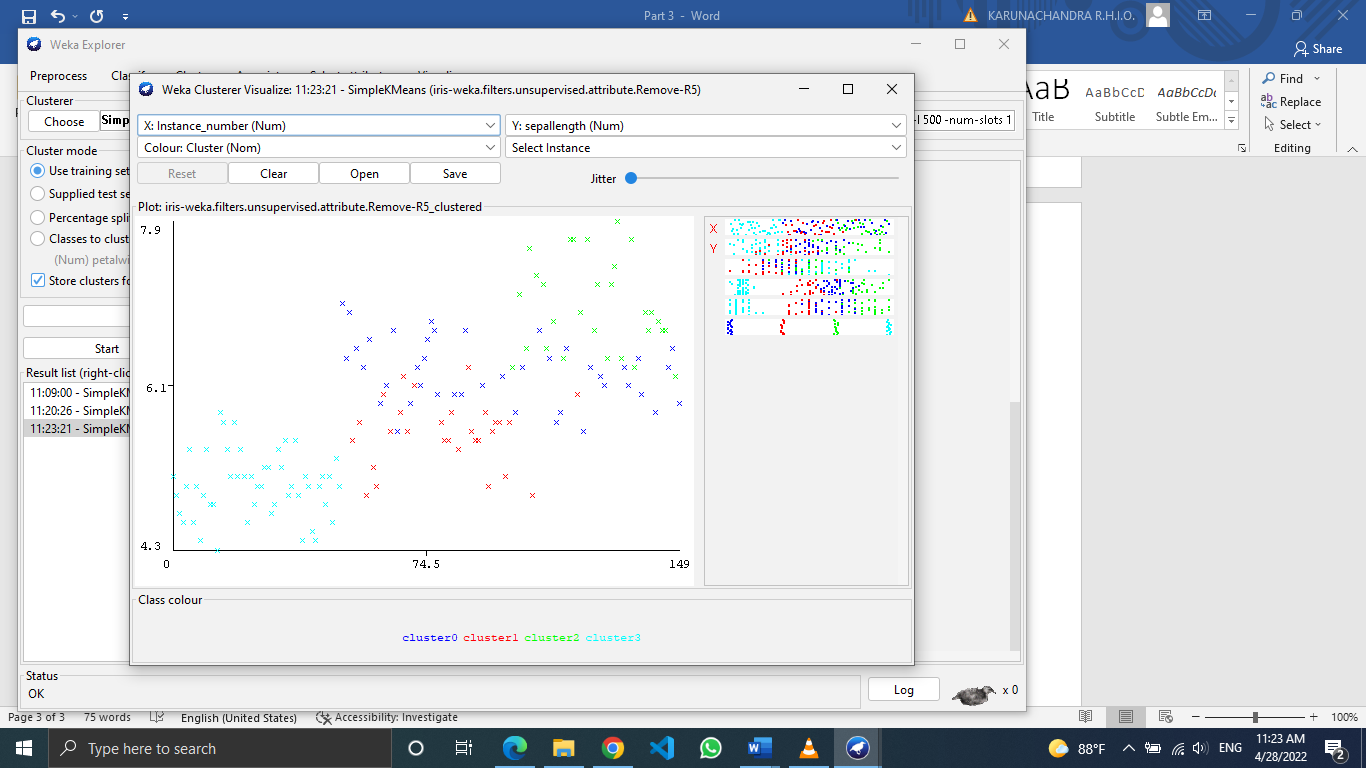




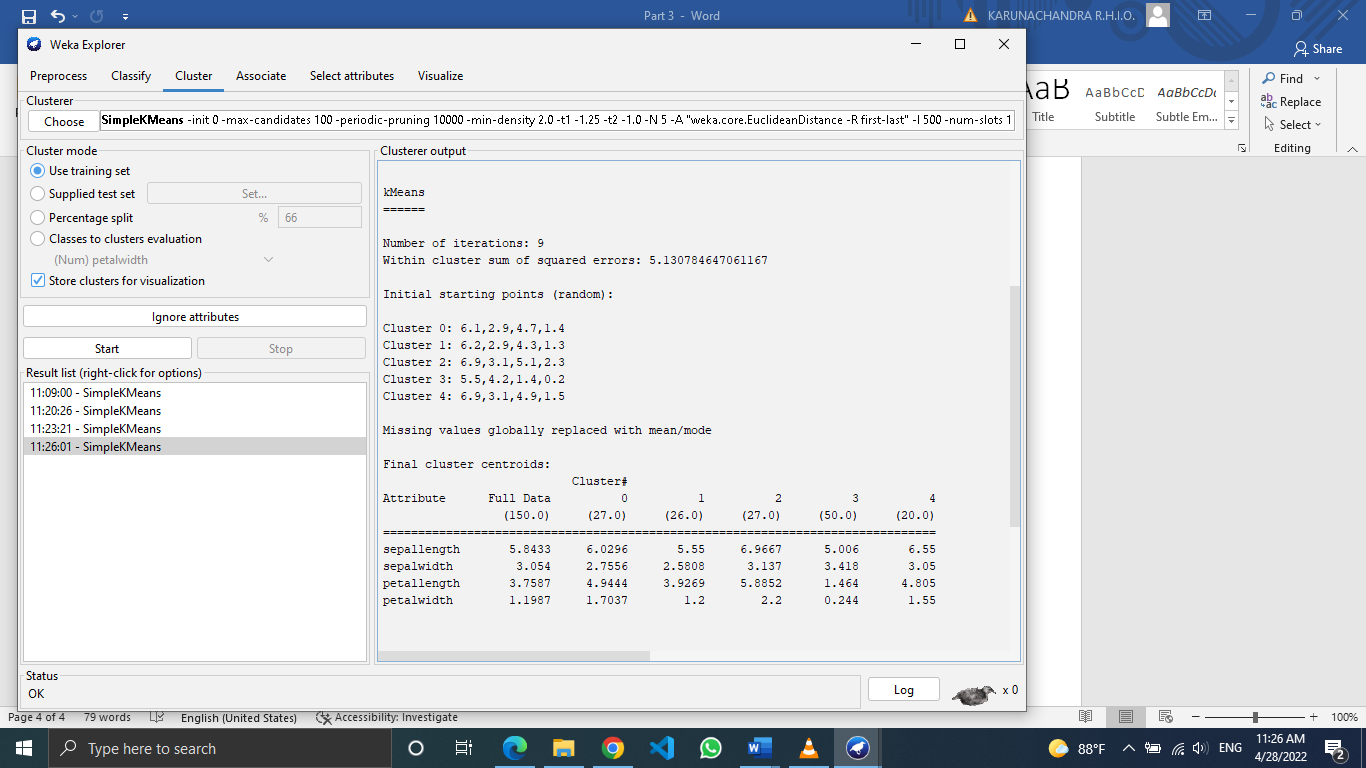
When k = 4

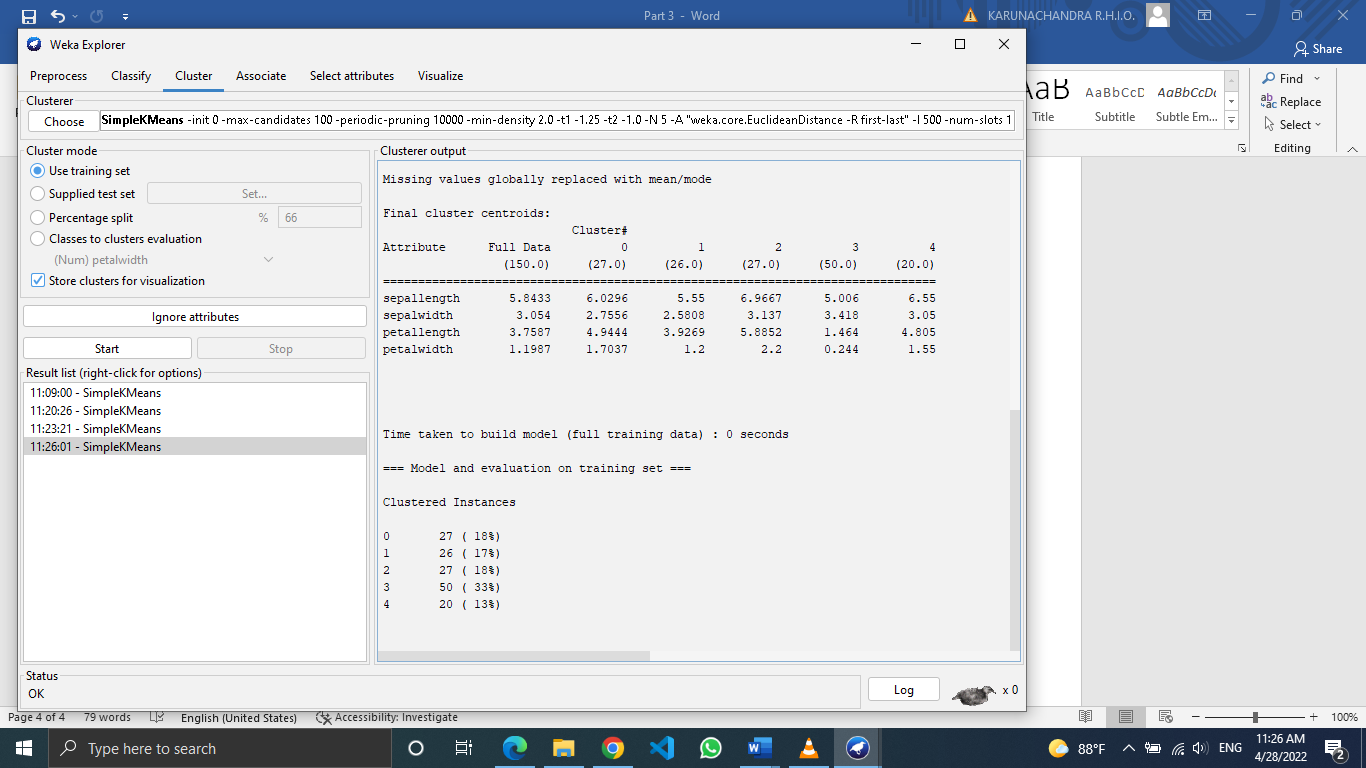


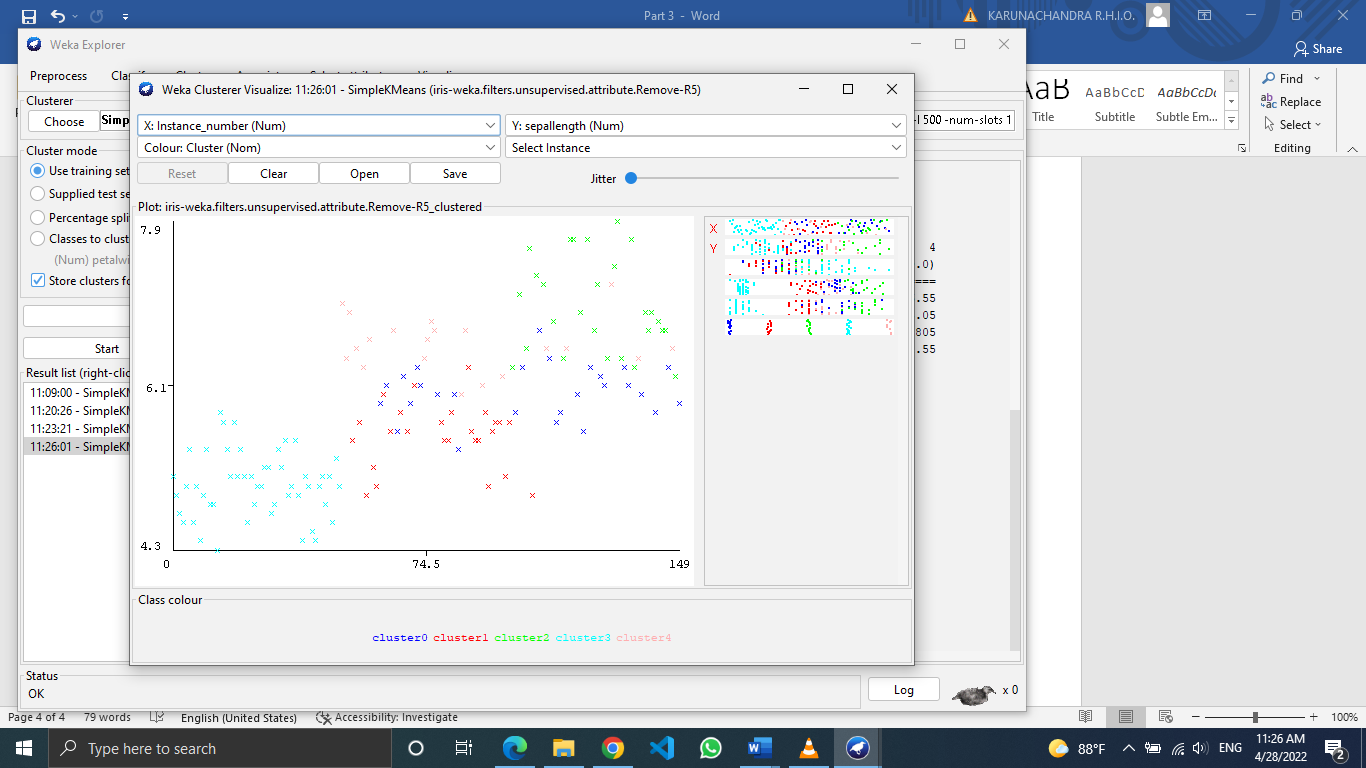




When k = 5

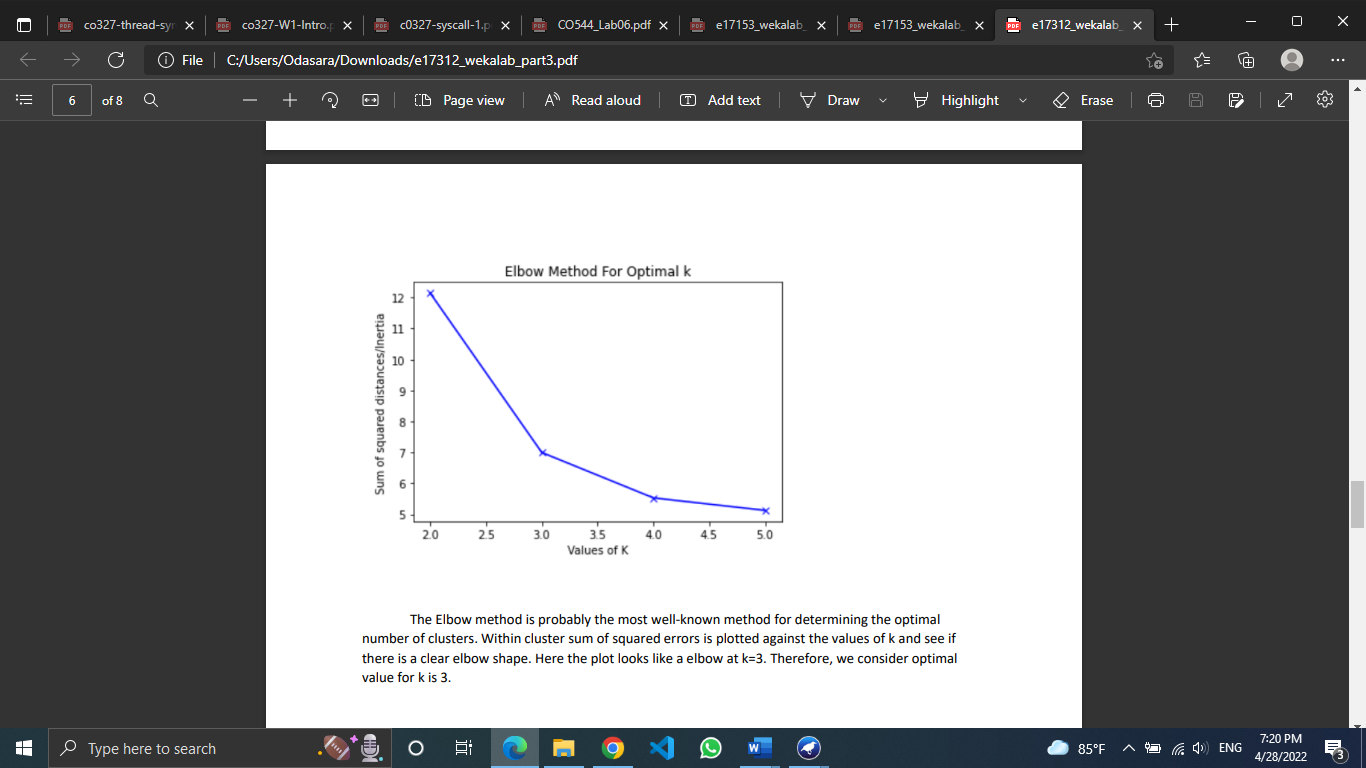






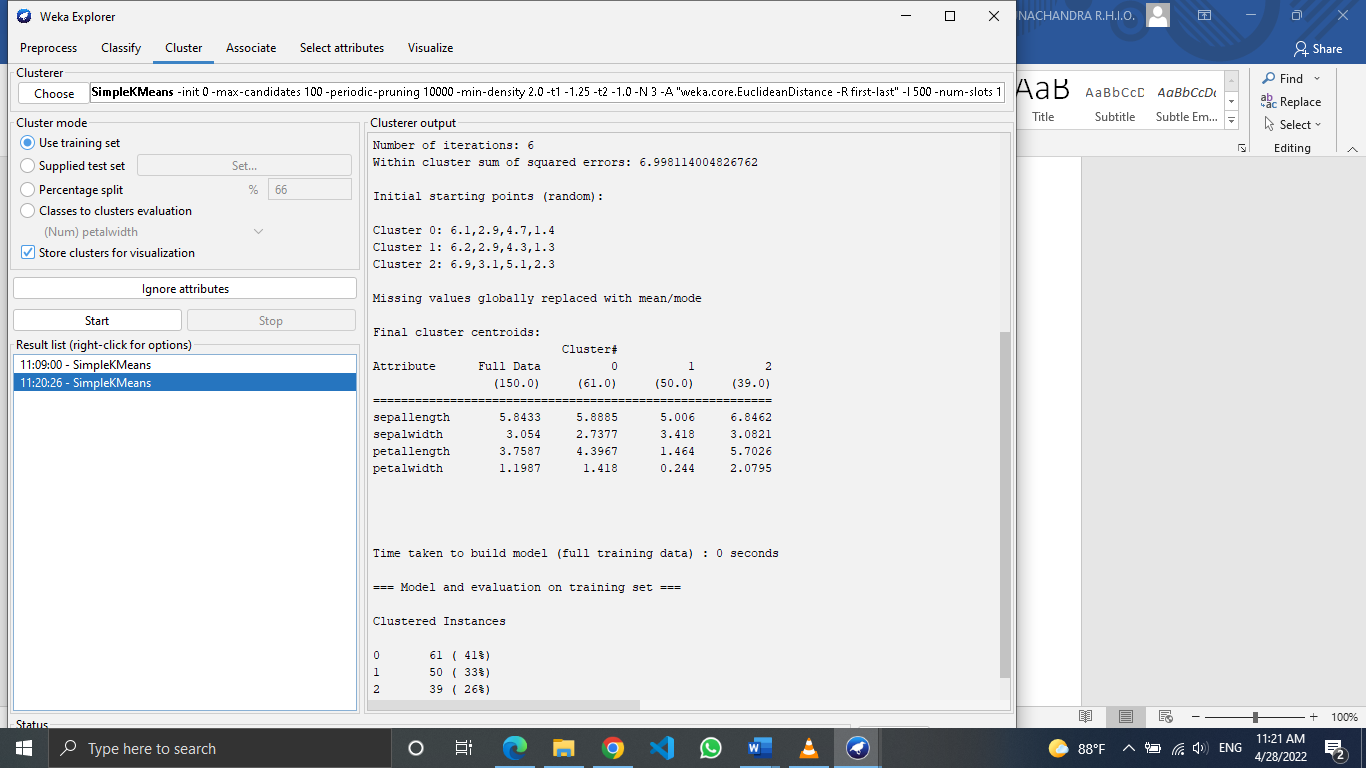
|  |  |
| --- | --- |
| K value | Within cluster sum of squared errors |
| 2 | 12.14368828 |
| 3 | 6.998114004 |
| 4 | 5.532832003 |
| 5 | 5.130784647 |

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.

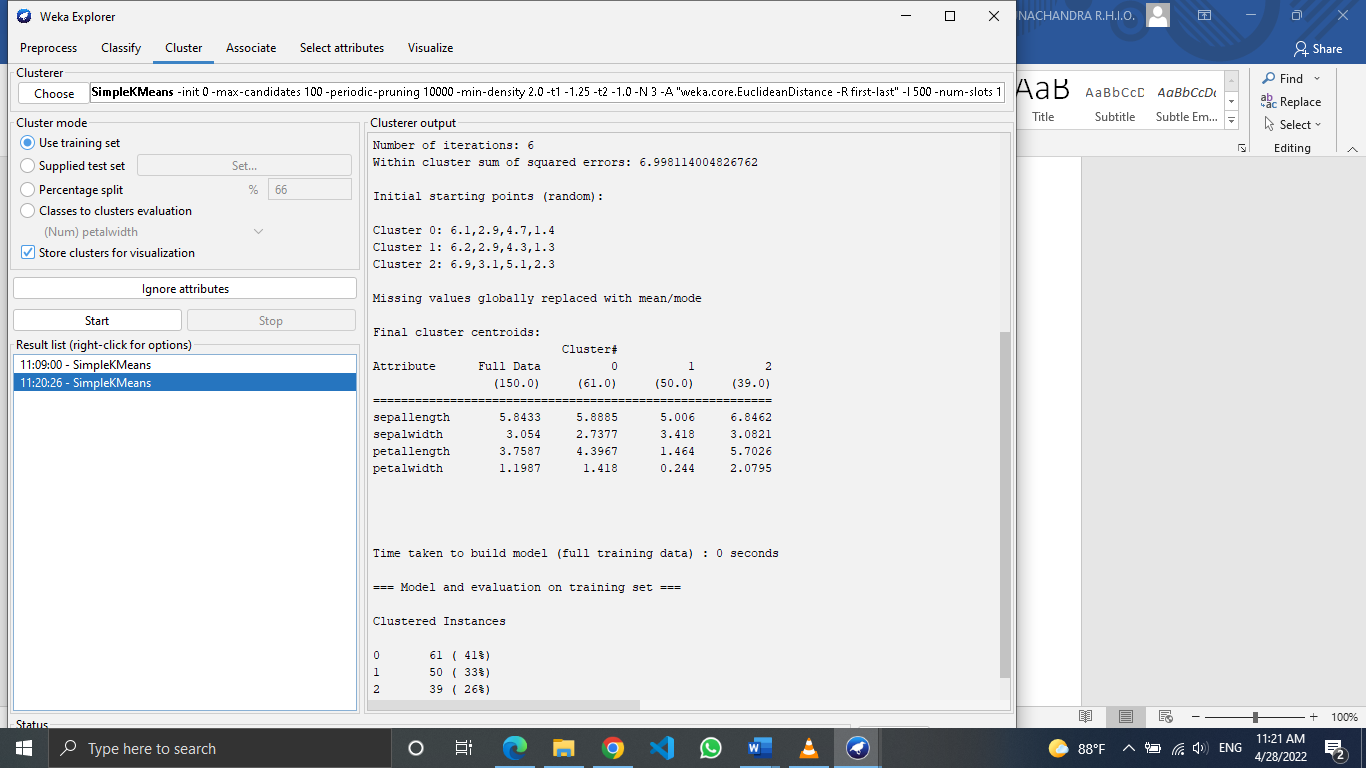


This shows the optimal k value as 3 and as we already know there are 3 real clusters for the dataset, it is exact that optimal k value is 3.

8.



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.



Then a class value is assigned to each cluster. Class Irisversicolor 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.

