**Dataset Description:**

The Seeds dataset contains 210 rows of measurements of geometrical properties of kernels belonging to three different varieties of wheat. The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each. Every observation had 7 different features on which the observation was tagged in one of the three varieties.

**Attribute Information:**

To construct the data, seven geometric parameters of wheat kernels were measured:

1. area A,

2. perimeter P,

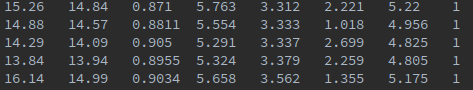
3. compactness C = 4\*pi\*A/P^2,

4. length of kernel,

5. width of kernel,

6. asymmetry coefficient

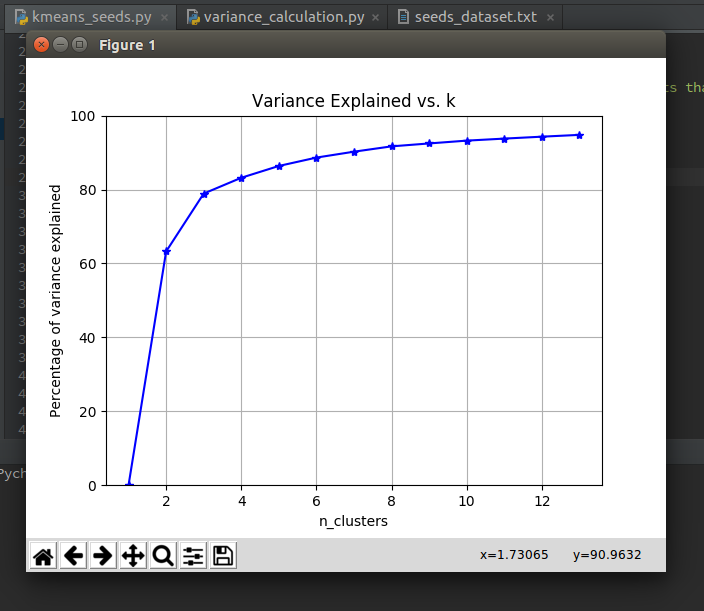
7. length of kernel groove.



The screenshot above is from the dataset. There are a total of 8 columns, the last column being the label/variety.

As we were performing unsupervised learning, we didn’t account the label column and the first 7 attributes were accounted for in k-means.

In an ideal world, the k-means would produce the same number of clusters as the actual varieties; we did a variance check that gave us a graphical illustration of how many clusters would be ideal for running k-means on this dataset. The following graph also told us that 3 clusters would be ideal.



**K-means Accuracy:**

After running k-means with scikit-learn’s clustering algorithm, the results were quite accurate.

The following centroids were computed by setting clusters=3, init= ‘k-means++’ which selects initial cluster centers for k-mean clustering in a smart way to speed up convergence.

[[ 18.72180328 16.29737705 0.88508689 6.20893443 3.72267213

3.60359016 6.06609836] //Cluster 0

[ 11.96441558 13.27480519 0.8522 5.22928571 2.87292208

4.75974026 5.08851948] //Cluster 1

[ 14.64847222 14.46041667 0.87916667 5.56377778 3.27790278

2.64893333 5.19231944]] //Cluster 2

The following labels were computed by k-means:

[2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2

0 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 2 2 2 2 2 1 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 2 0 2 0 0 0 0 0 0 0 2 2 2 2 0 2 2 2 1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1]

The list contains 210 records. As in the input dataset, first 70 rows belonged to one variety, next 70 belonged to another variety and so was the case for the last 70.

Out of the first 70 input, 9 inputs were labeled incorrect. Out of the Next 70 inputs, 10 were labeled incorrect. Out of the last 70, only 2 were labeled incorrect.

So out of a total of 210, 21 were labeled incorrect.

**Accuracy of k-means = 90%**

The output of k-means is variant because of k-means++ centroid selection but it always yields an accuracy ~90%.