Exercise 4: How many clusters of grain?

In the video, you learnt how to choose a good number of clusters for a dataset using the k-means inertia graph. You are given a dataset of the measurements of samples of grain. What's a good number of clusters in this case?

This dataset was obtained from the UCI.

From the course *Transition to Data Science*. Buy the entire course for just $10 for many more exercises and helpful video lectures.

Step 1: Load the dataset *(written for you)*.

In [ ]:
```
import pandas as pd

seeds_df = pd.read_csv('datasets/seeds.csv')
# forget about the grain variety for the moment - we'll use this later
del seeds_df['grain_variety']
```

Step 2: Display the DataFrame to inspect the data. Notice that there are 7 columns - so each grain sample (row) is a point in 7D space! Scatter plots can't help us here.

In [ ]:

Step 3: Extract the measurements from the DataFrame using its `.values` attribute:

In [ ]:

Step 4: *(Written for you)*. Measure the quality of clusterings with different numbers of clusters using the inertia. For each of the given values of `k`, perform the following steps:

- Create a `KMeans` instance called `model` with `k` clusters.
- Fit the model to the grain data `samples`.
- Append the value of the `inertia_` attribute of `model` to the list `inertias`.

In [ ]:

Step 5: Plot the inertia to see which number of clusters is best. Remember: lower numbers are better!

In [ ]:

Excellent work! You can see from the graph that 3 is a good number of clusters, since these are points where the inertia begins to decrease more slowly.

In [ ]:

