Exercise 5: Evaluating the grain clustering

In the previous exercise, you observed from the inertia plot that 3 is a good number of clusters for the grain data. In fact, the grain samples come from a mix of 3 different grain varieties: "Kama", "Rosa" and "Canadian". In this exercise, cluster the grain samples into three clusters, and compare the clusters to the grain varieties using a cross-tabulation.

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)*.

You have the array `samples` of grain samples, and a list `varieties` giving the grain variety for each sample.

```
In [ ]: import pandas as pd

        seeds_df = pd.read_csv('datasets/seeds.csv')

        # extract the grain varieties from the dataframe
        varieties = list(seeds_df['grain_variety'])
        del seeds_df['grain_variety']

        samples = seeds_df.values
```

Step 2: Import KMeans

```
In [ ]:
```

Step 3: Create a `KMeans` model called `model` with `3` clusters.

```
In [ ]:
```

Step 4: Use the `.fit_predict()` method of `model` to fit it to `samples` and derive the cluster labels.

Calling `.fit_predict()` is the same as calling `.fit()` and then calling `.predict()`.

```
In [ ]:
```

Step 5: Create a DataFrame `df` with two columns named `'labels'` and `'varieties'`, using `labels` and `varieties`, respectively, for the column values. *(This has been done for you.)*

```
In [ ]:
```

Step 6: Use the `pd.crosstab()` function on `df['labels']` and `df['varieties']` to count the number of times each grain variety coincides with each cluster label. Assign the result to `ct`.

```
In [ ]:
```


Step 7: Display `ct` by evaluating it - and inspect your cross-tabulation! You'll see that your clustering is pretty good.

```
In [ ]:
```

