# Exercise 7: Clustering the fish data

Now use your standardization and clustering pipeline from the previous exercise to cluster the fish by their measurements, and then create a cross-tabulation to compare the cluster labels with the fish species.

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, extracting the species of the fish as a list `species` *(done for you)*

```
In [ ]: import pandas as pd

        df = pd.read_csv('datasets/fish.csv')

        # remove the species from the DataFrame so only the measurements are left
        species = list(df['species'])
        del df['species']
```

**Step 2:** Build the pipeline as in the previous exercise *(filled in for you).*

```
In [ ]:
```

**Step 3:** Fit the pipeline to the fish measurements `samples`.

```
In [ ]:
```

**Step 4:** Obtain the cluster labels for `samples` by using the `.predict()` method of `pipeline`, assigning the result to `labels`.

```
In [ ]:
```

**Step 5:** Using `pd.DataFrame()`, create a DataFrame `df` with two columns named `'labels'` and `'species'`, using `labels` and `species`, respectively, for the column values.

```
In [ ]:
```

**Step 6:** Using `pd.crosstab()`, create a cross-tabulation `ct` of `df['labels']` and `df['species']`.

```
In [ ]:
```

**Step 7:** Display your cross-tabulation, and check out how good your clustering is!

```
In [ ]:
```