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 [11]: 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 [12]: samples = df.values

         from sklearn.pipeline import make_pipeline
         from sklearn.preprocessing import StandardScaler
         from sklearn.cluster import KMeans

         scaler = StandardScaler()
         kmeans = KMeans(n_clusters=4)
         pipeline = make_pipeline(scaler, kmeans)
```

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

```
In [13]: pipeline.fit(samples)

Out[13]: Pipeline(steps=[('standardscaler', StandardScaler(copy=True, with_mean=True, with_
             n_clusters=4, n_init=10, n_jobs=1, precompute_distances='auto',
             random_state=None, tol=0.0001, verbose=0))])
```

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

```
In [14]: labels = pipeline.predict(samples)
```

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 [8]: df = pd.DataFrame({'labels': labels, 'species': species})
```

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

```
In [9]: ct = pd.crosstab(df['labels'], df['species'])
```

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

```
In [10]: ct
```

```
Out[10]: species  Bream  Pike  Roach  Smelt
         labels
         0            1     0     19      1
         1           33     0      1      0
         2            0     0      0     13
         3            0    17      0      0
```

