01_tabular_data_exploration_ex_01

June 25, 2023

1 Exercise M1.01

Imagine we are interested in predicting penguins species based on two of their body measurements: culmen length and culmen depth. First we want to do some data exploration to get a feel for the data.

What are the features? What is the target?

The data is located in `../datasets/penguins_classification.csv`, load it with `pandas` into a `DataFrame`.

[2]:

```python
# Write your code here.
import sklearn
sklearn.show_versions()

import pandas as pd
penguins = pd.read_csv("../datasets/penguins_classification.csv")
```

```
System:
    python: 3.9.10 (v3.9.10:f2f3f53782, Jan 13 2022, 17:02:14)  [Clang 6.0
(clang-600.0.57)]
executable: /Users/nirvanabear/.local/share/virtualenvs/scikit-learn-mooc-
btn2WeXi/bin/python3
   machine: macOS-10.16-x86_64-i386-64bit

Python dependencies:
          pip: 22.0.3
   setuptools: 60.9.3
      sklearn: 1.0.2
        numpy: 1.22.2
        scipy: 1.8.0
       Cython: 0.29.28
       pandas: 1.4.1
   matplotlib: 3.5.1
       joblib: 1.1.0
threadpoolctl: 3.1.0

Built with OpenMP: True
```

Show a few samples of the data.

How many features are numerical? How many features are categorical?

[3]:

```
penguins.head()
```

[3]:

```
   Culmen Length (mm)  Culmen Depth (mm) Species
0                39.1               18.7  Adelie
1                39.5               17.4  Adelie
2                40.3               18.0  Adelie
3                36.7               19.3  Adelie
4                39.3               20.6  Adelie
```

What are the different penguins species available in the dataset and how many samples of each species are there? Hint: select the right column and use the `value_counts` method.

[6]:

```
penguins["Species"].value_counts()
```

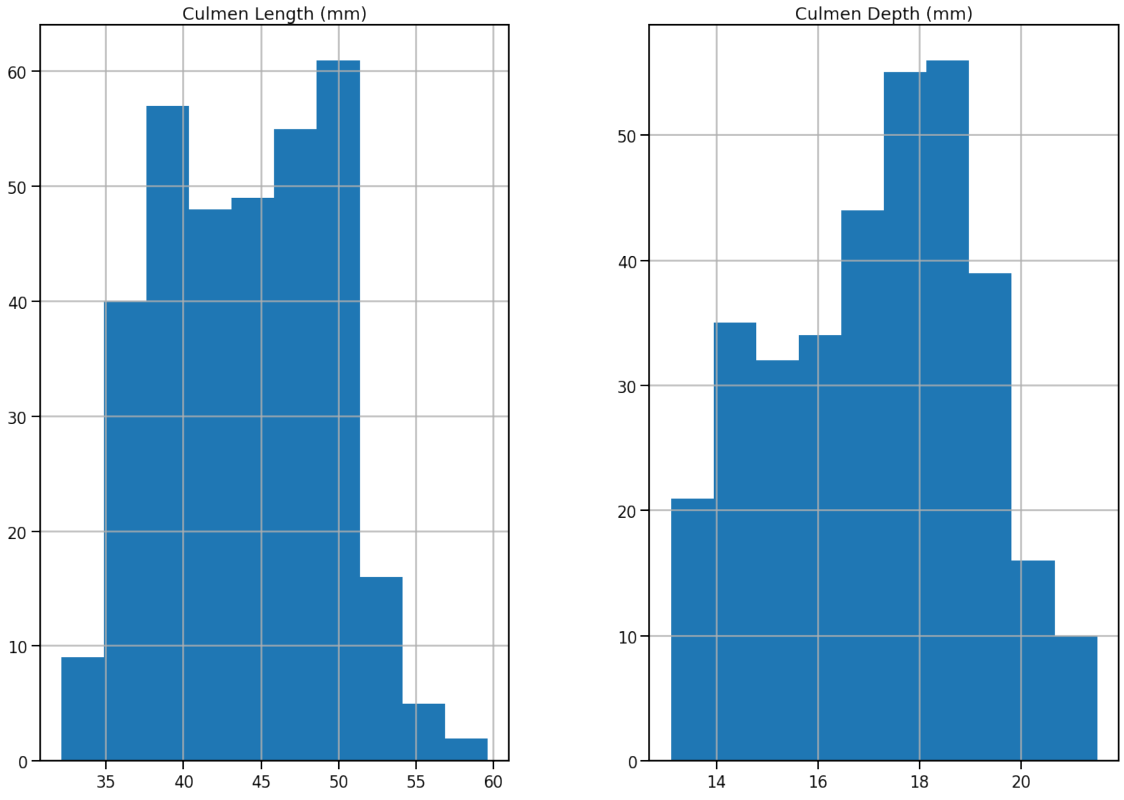
[6]:

```
Adelie       151
Gentoo       123
Chinstrap     68
Name: Species, dtype: int64
```

Plot histograms for the numerical features

[7]:

```
_ = penguins.hist(figsize=(20, 14))
```



Show features distribution for each class. Hint: use `seaborn.pairplot`

[9]:

```
penguins.shape[0]
```

[9]:

```
342
```

[10]:

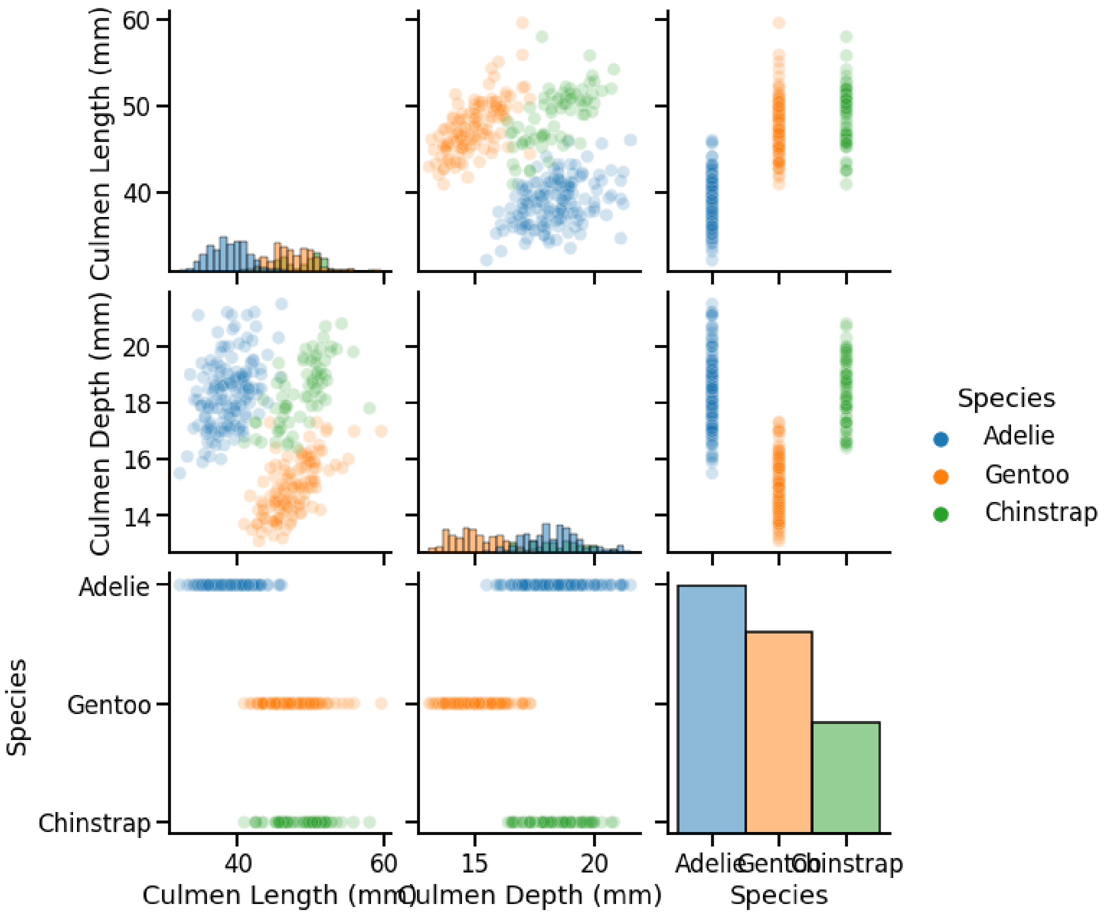
```
penguins.columns
```

[10]:

```
Index(['Culmen Length (mm)', 'Culmen Depth (mm)', 'Species'], dtype='object')
```

[13]:

```
import seaborn as sns
m = 342
columns = ['Culmen Length (mm)', 'Culmen Depth (mm)', 'Species']
_ = sns.pairplot(
    data=penguins[:342],
    vars=columns,
    hue="Species",
    plot_kws={"alpha": 0.2},
    height=3,
    diag_kind="hist",
    diag_kws={"bins": 30},
)
```



Looking at these distributions, how hard do you think it will be to classify the penguins only using `"culmen depth"` and `"culmen length"`?