420-SN1-RE Programming in Science - Lab Exercise 16

October 23, 2024

Introduction

Goals for this lab:

- Work with pandas
- Create scatter plots

The data for this lab

The data for this lab, `penguins.csv`, is a comma-separated value file that contains information collected about penguins in the Palmer Archipelago in Antarctica. You can find more information and citations here https://archive.ics.uci.edu/dataset/690/palmer+penguins-3. We have modified the data by removing any incomplete records.

Open the file using a text editor (such as IDLE or Notepad). You can see that the file consists of 8 columns: There are a total of 333 records (rows) in this file. There are no missing values.

Column name	Type	values or units
species	str	Adelie, Gentoo, or Chinstrap
island	str	Torgersen, Biscoe, or Dream
bill_length_mm	float	millimetres
bill_depth_mm	float	millimetres
flipper_length_mm	int	millimeters
body_mass_g	int	grams
sex	str	male or female
year	int	year data was collected

We have also provided the `avgtemp.txt` file used in the lecture notes. It contains the same data as was used in lab 9.

Exercise 0

Do this part of the lab in the IDLE shell. You do not need to submit anything for this exercise.

1. **Import pandas:**

   ```
   import pandas as pd
   ```

   This is the most standard way to import the package.

2. **Read the file:**

   ```
   df = pd.read_csv('penguins.csv')
   ```

   This should read the file into a `DataFrame` and associate that with the name `df`.

3. **Check the shape**

   ```
   print(df.shape)
   ```

   This will print a `tuple` that contains the number of rows and columns in the `DataFrame`.

4. **Examine the column names**

   ```
   print(df.columns)
   ```

   This should print the names of the columns as shown in the file's header.

5. **Examine the row index**

   ```
   print(df.index)
   ```

   This should print the row indices, which will be an object similar to a standard Python range.

6. **Examine a Series**

   ```
   print(df['flipper_length_mm'])
   ```

   This will print a summary of the 333 different values associated with the flipper length column. Note that the following will produce the identical result:

   ```
   print(df.flipper_length_mm)
   ```

   Pandas allows you to using an index with a key to access a column, but you can also use the name of the column after a period. Each method is useful in different situations.

7. **Get the list of possible species**

   ```
   print(df.species.unique())
   ```

   The `unique()` method returns an iterable of all of the unique values found in the column.

8. **Create a `DataFrame` consisting of Adelie penguins only**

   ```
   a = df[df.species == 'Adelie']]
   print(a.shape, a.species.unique())
   ```

   Note that the shape tells you how many records there are for Adelie penguins.

9. **Compare the average body mass for male vs female penguins**

   ```
   m = df[df.sex == 'male']
   w = df[df.sex == 'female']
   from statistics import mean
   print(mean(m.body_mass_g), mean(w.body_mass_g))
   ```

10. **Compare the average flipper length for the three species**

    ```
    for species_name in df.species.unique():
        subset = df[df.species == species_name]
        print(species_name, mean(subset.flipper_length_mm))
    ```

    Which species has the longest flippers, on average?

11. **Compare the average body mass for the three species**

    ```
    for species_name in df.species.unique():
        subset = df[df.species == species_name]
        print(species_name, mean(subset.body_mass_g))
    ```

    Which species is the heaviest, on average?

12. **Create and show a scatter plot of the bill dimensions**

    ```
    import matplotlib.pyplot as plt
    plt.scatter(df.bill_depth_mm, df.bill_length_mm)
    plt.show()
    ```

13. **Answer some other questions**

    See if you can use what you've learned to answer these questions:

    (a) How many male penguins are in the file?

    (b) How many male penguins live on Biscoe island?

    (c) What is the average flipper length of the Adelie penguins on Biscoe island?

Exercise 1

Create a Python file named `lab16ex1.py` that does the following:

1. Use `pandas` to read the `penguins.csv` data.

2. Print a table showing the average body mass of the three species , broken out by sex. Your table should look like the following:

```
          male    female
Adelie    4043.49 3368.84
Gentoo    5484.84 4679.74
Chinstrap 3938.97 3527.21
```

Try to use values read from the file for the sex and the species, rather than "hard-wiring" these values in your code.

Exercise 2

Create a Python file named `lab16ex2.py` that does the following:

1. Use `pandas` to read the `penguins.csv` data.

2. Create a separate scatter plot that shows the bill depth vs. bill length broken out for each species. You can call `scatter` multiple times to create a scatter plot for multiple sets of X and Y values.

3. Use the call `plt.xlabel(s)` to set an appropriate label for the X axis.

4. Use the call `plt.ylabel(s)` to set an appropriate label for the Y axis.

5. Use the `plt.legend(ls)` to give a list of strings to use as a "legend" for the chart.

Your finished chart should resemble Figure 1.

As with exercise 1, try to use values read from the file for the species, rather than "hard-wiring" these values in your code. This will not always be possible, but it's a good goal to have.

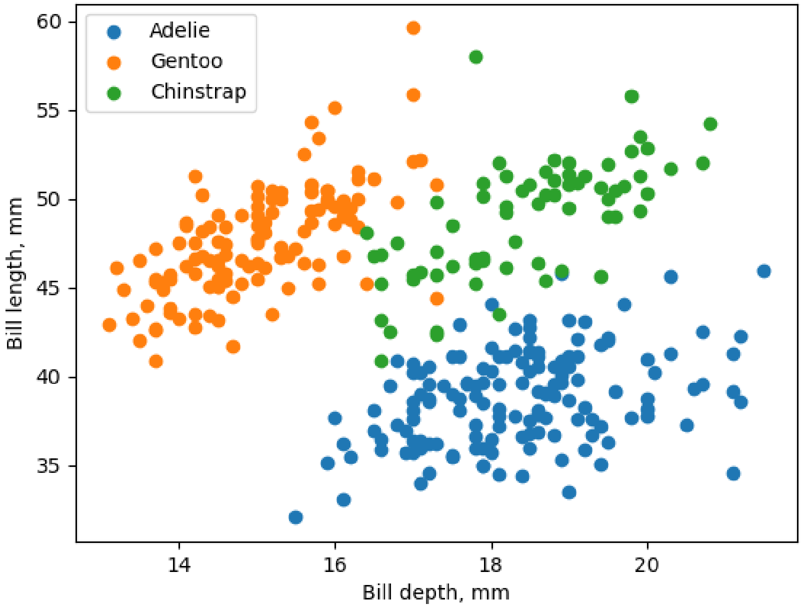

Figure 1: The finished scatter plot

Submission requirements

Submit your two completed Python files, `lab16ex1.py` and `lab16ex2.py`. Place them in a ZIP file and upload them to Omnivox.