PIC16B HW0

PIC16B HW

AUTHOR
Zhe Shen

PUBLISHED
January 22, 2023

PIC 16B HW0

In this blog, we will visualize the penguin data.

These data were provided by Dr. Kristen Gorman and Palmer Station, Antarctica LTER. You can visit data by this link: https://raw.githubusercontent.com/PhilChodrow/PIC16B/master/datasets/palmer_penguins.csv

I am very grateful that their work has not only contributed to penguin conservation, but also provided rich data for my programming courses.

First of all, let's import data and any libraries maybe we need to use there. We use the function **pd.read_csv()** to read a CSV file. In this case, we read data from palmer_penguins.cvs into pandas data frame and named it **data**.

```
# import libraries and data
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

url = "https://raw.githubusercontent.com/PhilChodrow/PIC16B/master/datasets/palmer_penguins.cs
data = pd.read_csv(url)
```

Species distribution

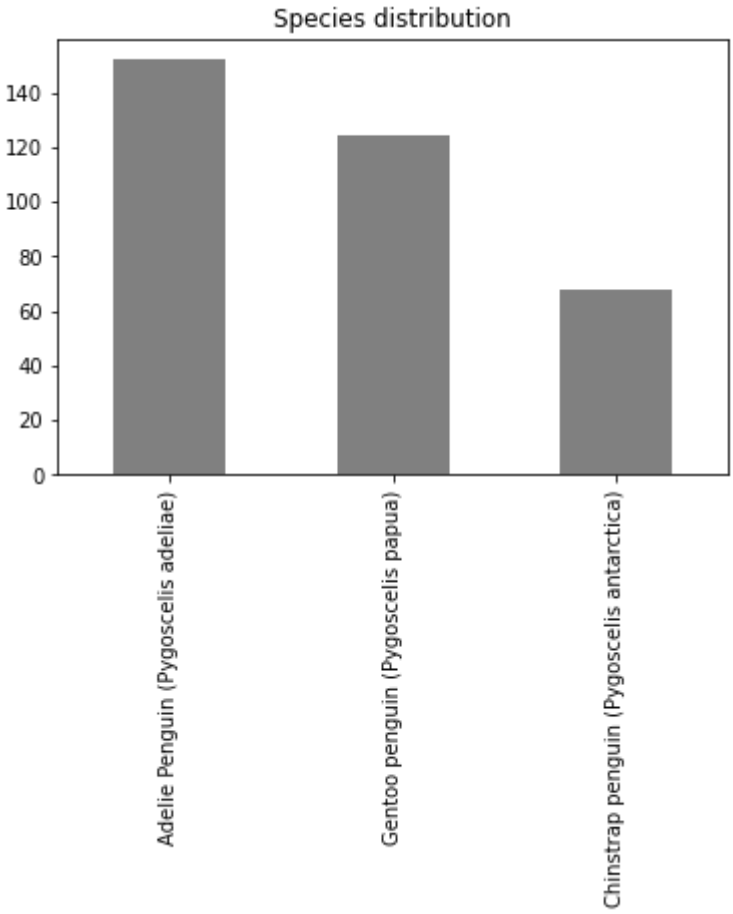
Bar plot is the easy and common visualization. We will use **dataframe.plot.bar()** to creat the plot. And we need use **data** to replace **dataframe**. And if we want calculate and show specific topic like Species distribution. We can add **Species** and **value_counts()** before **plot.bar()**. And use dot **.** to separate.

In addition, we can set parameters of **plot.bar()** by add code in **()**. The main parameters as follows: - x – What you want to have your bars be. This is just a label. - y – What you want the values or height of your bar plot to

be. Often be numbers. - color – What you want the color of plot to be. In this case we use gray as the color. - kwargs – Any general parameters you want add.

```
#Use bar plot to show 3 species distribution.
data.Species.value_counts().plot.bar(x = "Species",y = "number", color = "gray")
#Set title
plt.title('Species distribution')

plt.show()
```



Violin plots visualization

Violin plots is very beautiful plots in my opinion, and I try use it to show the distribution of Culmen Length in different species.

Before we start this part, I want to emphasize that in order to use violin plot we need to import **seaborn**(done on the top of these pages). This is a visual library based on **matplotlib**. And it is very like **plot.bar()**. it takes some basic parameters as: - x – What you want to have your bars be. - y – What you want the values or height of your bar plot to be. - data - What dataset you want plotting.

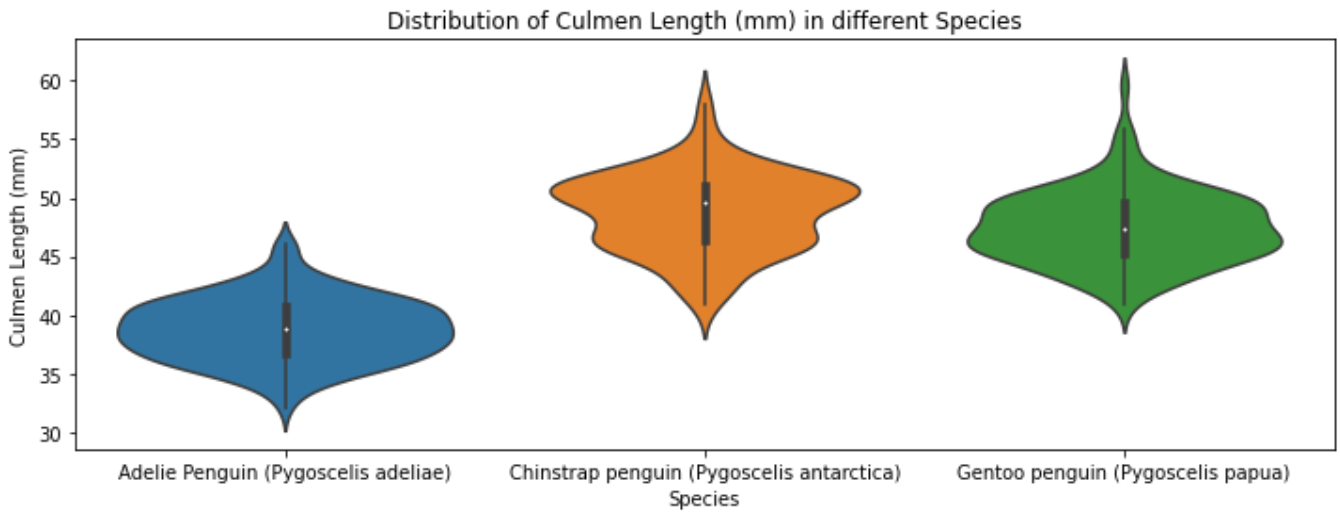
```
# Use violion polt to show the distribbution of Culmen Length in 3 species

#Setting image size
plt.figure(figsize=(12, 4))

#Setting x-axis, y-axis and find the data in set.
```

```
sns.violinplot(x='Species', y= 'Culmen Length (mm)', data = data)

#Add more information on the image.
plt.title('Distribution of {} in different Species'.format( 'Culmen Length (mm)'))
plt.show()
```



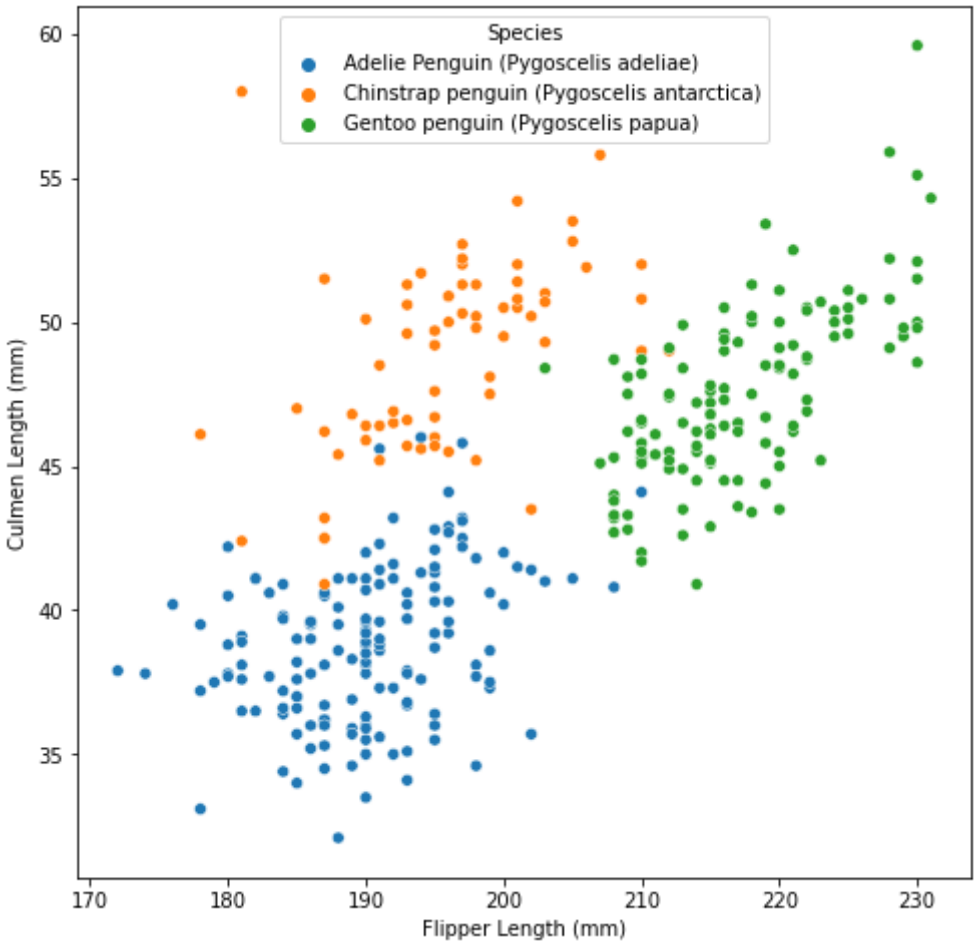
I try to pick up 2 features for more visualization by scatter plot

Scatter plot is a commom visualization. But what I've done here is a combination of lists and visualizations. This may be useful in complex visualization tasks. Also, I've added a **hue** parameter here to distinguish the data between different species. The general parameter introduce as follows: - x – What you want to have your bars be. - y – What you want the values or height of your bar plot to be. - hue - Use different colors to distinguish different types of data. - data - What dataset you want plotting.

```
#Setting features
quantitative_features = ['Flipper Length (mm)', 'Culmen Length (mm)']

#Setting image size
plt.figure(figsize=(8, 8))

#Show the plot
sns.scatterplot(x=quantitative_features[0], y=quantitative_features[1], hue='Species', data=da
plt.show()
```


Species
Adelie Penguin (Pygoscelis adeliae)
Chinstrap penguin (Pygoscelis antarctica)
Gentoo penguin (Pygoscelis papua)
Culmen Length (mm)
Flipper Length (mm)
60
55
50
45
40
35
170
180
190
200
210
220
230