Yasmin's PIC16B Portfolio

Yasmin Munoz

Blog About

Blog Post 0

In this post, I'll show how to create a helpful data visualization of the penguins data set.

Generate Synthetic Data

We'll start by reading and cleaning the data.

```
import pandas as pd
#Import the data and clean it
url = "https://raw.githubusercontent.com/PhilChodrow/PIC16B/master/datasets/
penguins = pd.read_csv(url)
penguins = penguins.dropna(subset = ["Body Mass (g)", "Sex"])
penguins["Species"] = penguins["Species"].str.split().str.get(0)
penguins = penguins[penguins["Sex"] != "."]

cols = ["Species", "Island", "Sex", "Culmen Length (mm)", "Culmen Depth (mm)"
penguins = penguins[cols]
```

Originally, I only imported the data using the following code block:

```
import pandas as pd
url = "https://raw.githubusercontent.com/PhilChodrow/PIC16B/master/datase
penguins = pd.read_csv(url)
```

> This lead me to getting an error when running my code, then I came to the realization that the data needs to be cleaned in order to avoid input errors.

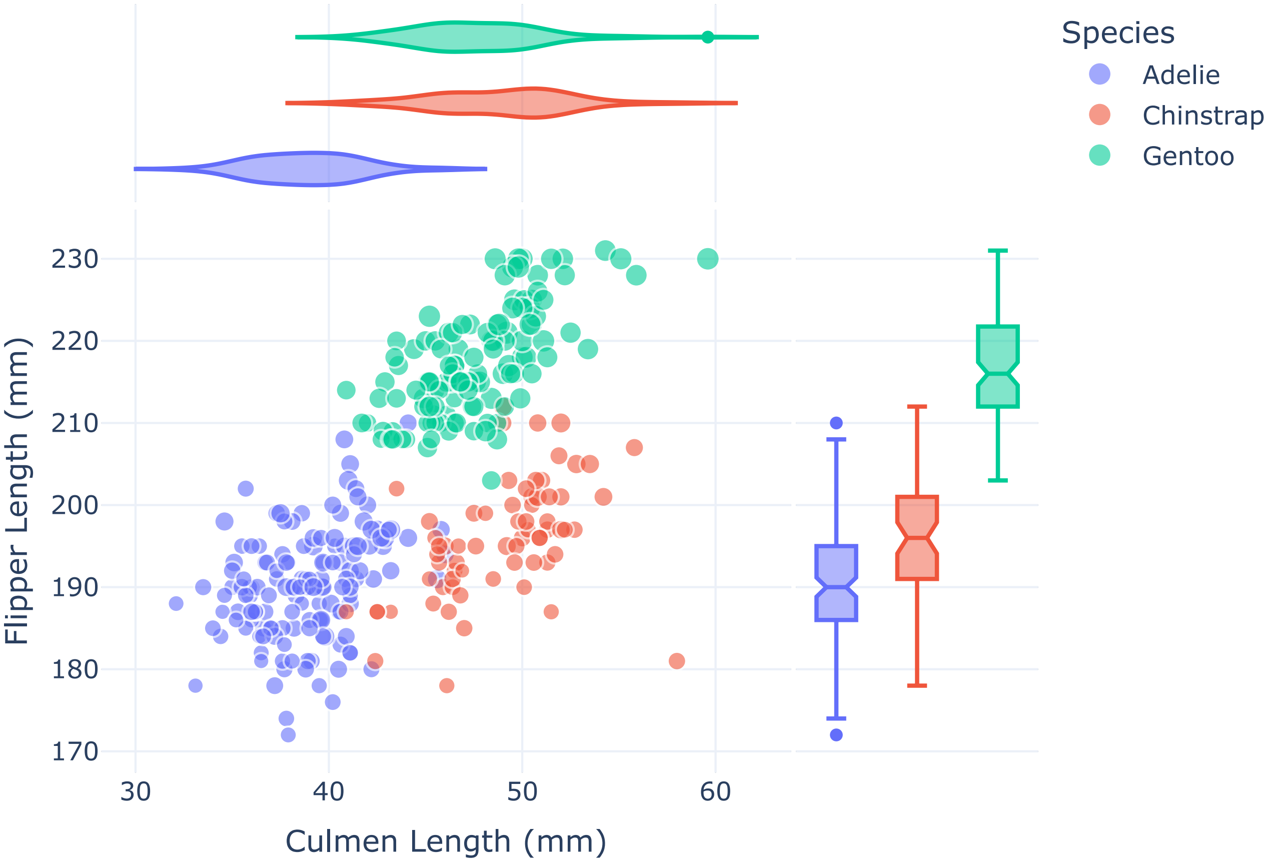
Make the plot

Next, we can create a scatter plot using `matplotlib`:

```python
from plotly import express as px
import plotly.io as pio
#I changed the template to a white plot in order to better decipher the plots
#with no background colors.
pio.templates.default = "plotly_white"
#Now create a scatter plot named fig using the penguins data.
fig = px.scatter(data_frame = penguins,
#I put the data from the Culmen Length column on the x-axis, but any data
#category you see fit works
                 x = "Culmen Length (mm)",
#Now assign the y-axis to be the flipper length column of data
                 y = "Flipper Length (mm)",
#The plots of the data are color coordinated by Species with the legend on
#the upper right specifying what species correlates to what color
                 color = "Species",
#The hover name specifies what the title of the data point will be when you
#'hover' over it, and the hover data
#is data that the graph does not depend on, but included in individual points
#when the curser is hovered over points
                 hover_name = "Species",
                 hover_data = ["Island", "Sex"],
#The size of each data point is correlated to the Body Mass of the specific
#penguin
```

```
                 size = "Body Mass (g)",
                 size_max = 8,
                 width = 600,
                 height = 400,
                 opacity = 0.6,
 #This creates a marginal graph along the y-axis that analyzes the distri-
 #butuion of Flipper length for each species of penguin using a box plot.
                 marginal_y = "box",
#This is making a marginal graph on the y-axis that only shows the
#distributoins of Culmen Length shown in a violin graph.
                 marginal_x = "violin")

#reduce whitespace
fig.update_layout(margin={"r":0,"t":0,"l":0,"b":0})
#show the plot
fig.show()
```



A thing I like about this graph is that the user is able to see the distribution of data from the x and y axis both alone in the marginal graphs and juxtaposed in the scatter plot. Another neat thing about this data visualization is the information that it carries and display when hovering over each data point.

Written on February 5, 2021