Assignment 4: Static Visualization of Penguin Data Using Seaborn

Grace Texana Long Torales - gtl1500@g.rit.edu

Summary

In this notebook, I explored the penguins dataset provided by seaborn. This dataset contains information about 3 species of penguin: Adelie, Chinstrap, and Gentoo. During my exploration, I was able to determine that Gentoo penguins tend to be the largest species of the 3. More subtly, they also tend to have slightly longer wing lengths. Furthermore, male penguins of each species tend to be larger than the females and also tend to have larger beaks than the females.

Additionally, the data described in the dataset is collected from 3 islands: Biscoe, Dream, and Torgersen. Biscoe is home to the largest penguin population, followed by Dream and then Torgersen. Each island has a roughly equal split of females and males. Gentoo are only present on Biscoe while Chinstrap are only present on Dream. Adelies are the only penguin species present on all 3 islands and are outnumber by the other penguin species on each island that they share. There does not seem to be much difference between the Adelie populations on each of the islands.

Introduction

In this report, I was primarily interested in finding what differences there were between each of the islands. I wanted to know the distributions of each species, the reasons for those distributions, and how they are changing due to climate change. But first, I had to learn more about the penguins themselves, which is where I began. My initial exploration was a look at the big picture, and I got progressively more specific from there.

To really gain insight about the penguins and their distributions, I'd ideally like additional relevant data, such as information about each of the island ecosystems, information about their proximity to each other, as well as the ages of the penguins, and things of that nature. The data here is enough to have an idea of the penguins' distributions, but not enough to explain why, nor to see how the ecosystems may be changing due to climate change, nor to form any hypotheses on how these distributions may be impacted by climate change. Luckily, there is a more detailed dataset available to analyze as well. Analyzing the more completed dataset might be my next step if I were to continue this project.

In [ ]:

```
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
import seaborn as sns
```

Dataset - Penguins

As previously stated, this dataset describes 3 species of penguin as well as 3 island habitats. There are 344 penguins in the dataset. In addiiton to species and island, the information collected for each penguin was bill length (mm), bill depth (mm), flipper length (mm), body mass (g), and sex. This data was collected by Dr. Kristen Gorman in the Palmer Archipelago of Antarctica.

In [ ]:
```
df = pd.DataFrame(sns.load_dataset("penguins"))

df.columns = ["Species", "Island", "Bill Length (mm)", "Bill Depth (mm)", "Flipper Len

df.sort_values(by=["Sex", "Species", "Island"], inplace=True)

df
```

Out[ ]:

	Species	Island	Bill Length (mm)	Bill Depth (mm)	Flipper Length (mm)	Body Mass (g)	Sex
20	Adelie	Biscoe	37.8	18.3	174.0	3400.0	Female
22	Adelie	Biscoe	35.9	19.2	189.0	3800.0	Female
25	Adelie	Biscoe	35.3	18.9	187.0	3800.0	Female
27	Adelie	Biscoe	40.5	17.9	187.0	3200.0	Female
28	Adelie	Biscoe	37.9	18.6	172.0	3150.0	Female
...	...	...	...	...	...	...	...
246	Gentoo	Biscoe	44.5	14.3	216.0	4100.0	NaN
286	Gentoo	Biscoe	46.2	14.4	214.0	4650.0	NaN
324	Gentoo	Biscoe	47.3	13.8	216.0	4725.0	NaN
336	Gentoo	Biscoe	44.5	15.7	217.0	4875.0	NaN
339	Gentoo	Biscoe	NaN	NaN	NaN	NaN	NaN

344 rows × 7 columns

Customization

I wanted my visualizations to be cohesive and consistent, so I used a pre-made color palette and manipulated it to suit my needs. Throughout this notebook, each variable is assigned one of the colors shown below. The color of each variable remains the same throughout.

In [ ]:
```
my_colors = sns.color_palette("hls", 8)
my_colors = my_colors[4:] + my_colors[:4]
my_colors = sns.color_palette(my_colors)
sns.set_theme(palette=my_colors)

sex_color_palette = sns.color_palette(my_colors[:2])
species_color_palette = sns.color_palette(my_colors[2:5])
```

```
island_color_palette = sns.color_palette(my_colors[5:])

my_colors
```

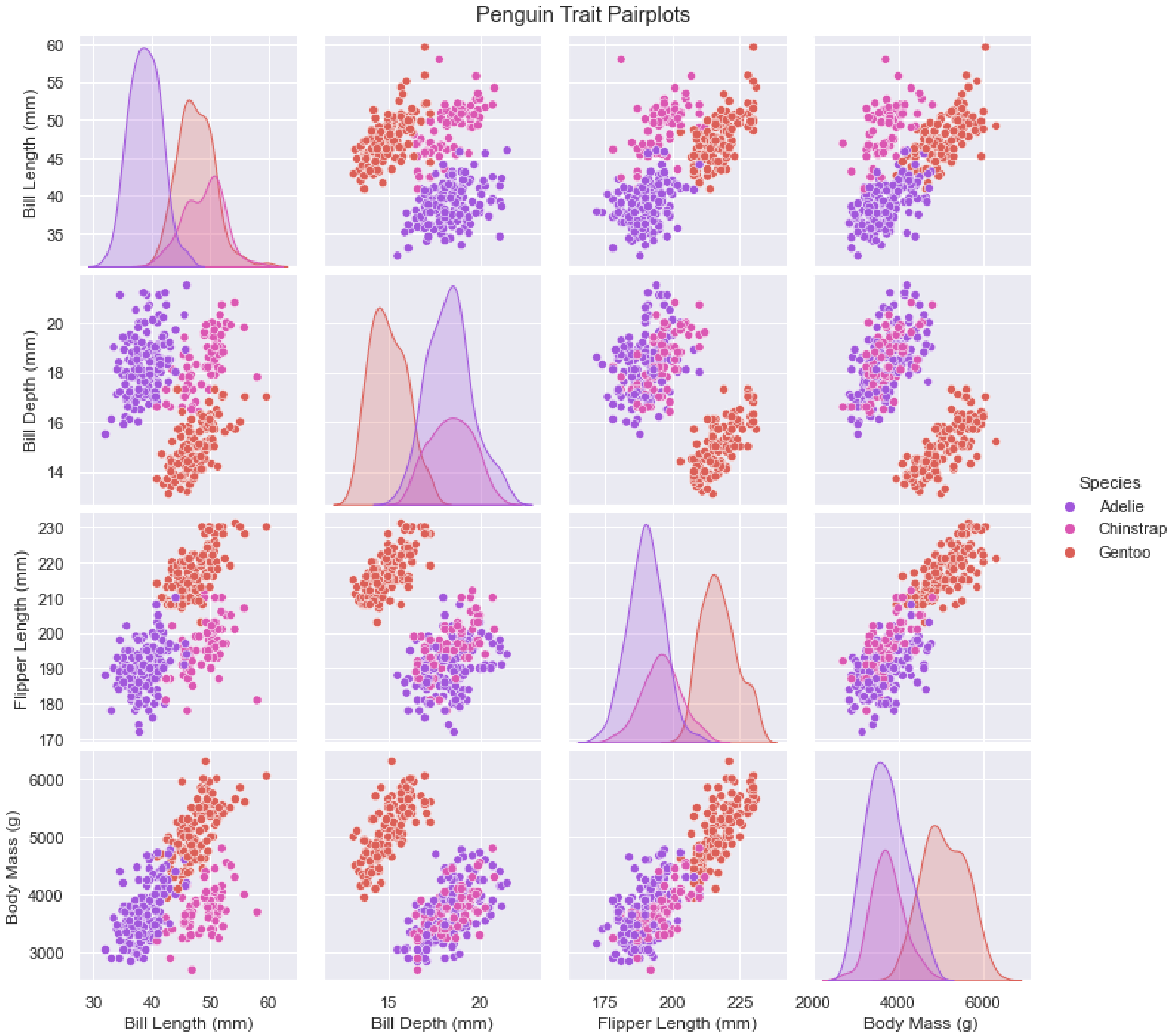
Out[ ]:

Preliminary Exploration

The next 2 visualizations are comparisons between each of the numeric variables within the dataset - color coded by species. In Penguin Trait Pairplots, each of the penguin species can clearly be seen to cluster together. Often, Adelie and Chinstrap physical traits are similar, as can be seen by their overlapping scatterplot clusters. Notably, body mass and flipper length seem to be linear related across all species. This is unsurprising, as one would expect a larger penguin to have larger wings. This relationship is further supported in the Penguin Trait Correlation Map in which body mass and wing length are shown to be highly correlated.

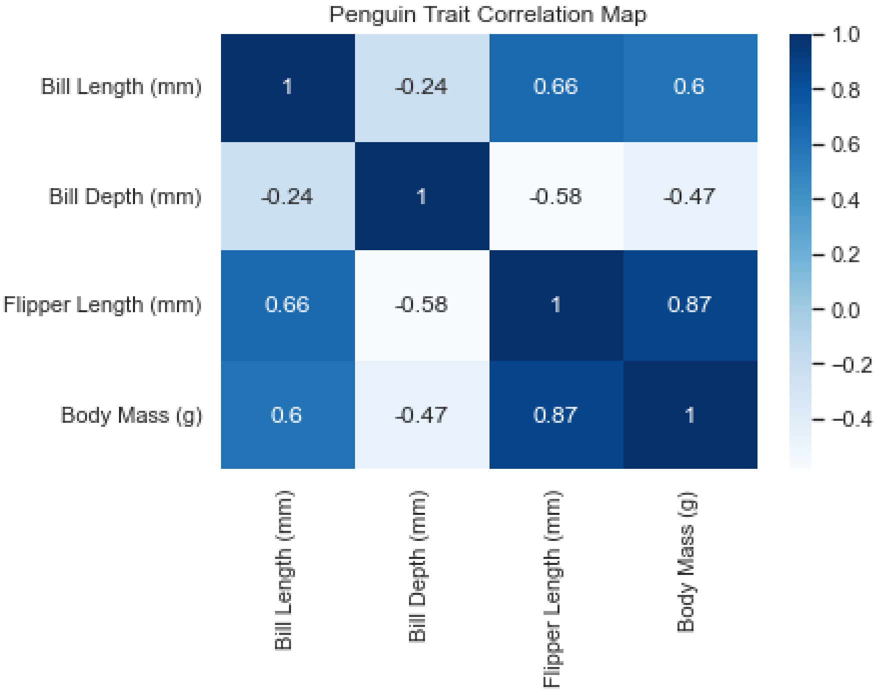
In [ ]:

```
g = sns.pairplot(df,
                 hue="Species",
                 palette=species_color_palette)

g.fig.suptitle("Penguin Trait Pairplots")
g.fig.subplots_adjust(top=0.95)
```



```
In [ ]: g = sns.heatmap(df.corr(),annot=True, cmap="Blues")

        g.set_title("Penguin Trait Correlation Map")
```

```
Out[ ]: Text(0.5, 1.0, 'Penguin Trait Correlation Map')
```



Differences Between Species

Upon seeing the strong correlation between body mass and wing length, I decided to analyze each separately. These plots further highlight the similarities between the Adelie and Chinstrap species in these respects. The Gentoo species is larger than the other 2 and also has slightly longer wings. Furthermore, males of each species are, on average, larger than females. They also have slightly longer wings, although only barely with regard to the Adelie.

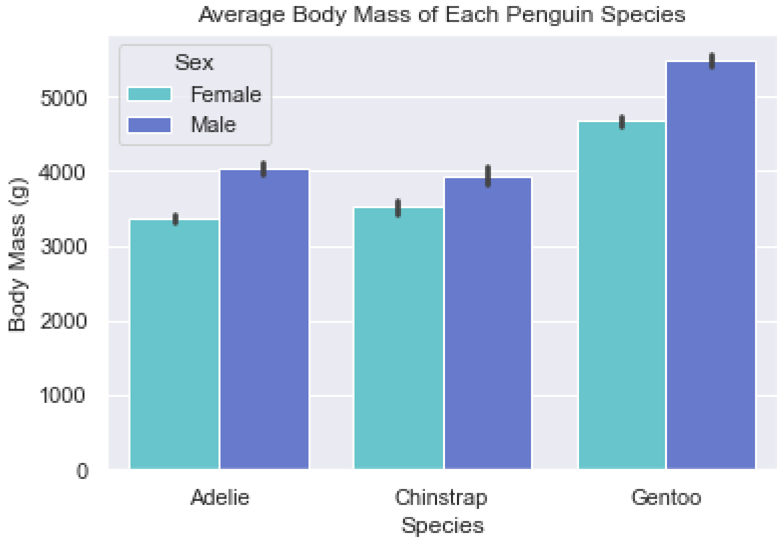
Additionally, I analyzed the beak sizes of each of the species. Once more, male beaks tend to be larger than female beaks. Adelie beaks tend to be short and thick. Chinstrap beaks tend to be long and thick. Gentoo beaks tend to be long and thin. Penguins with longer beaks tend to be larger, which is not surprising. One would expect that a larger penguin would be larger in every aspect.

In [ ]:

```
g = sns.barplot(df,
                x="Species",
                y="Body Mass (g)",
                hue="Sex",
                palette=sex_color_palette)

g.set_title("Average Body Mass of Each Penguin Species")
```

Out[ ]:

```
Text(0.5, 1.0, 'Average Body Mass of Each Penguin Species')
```

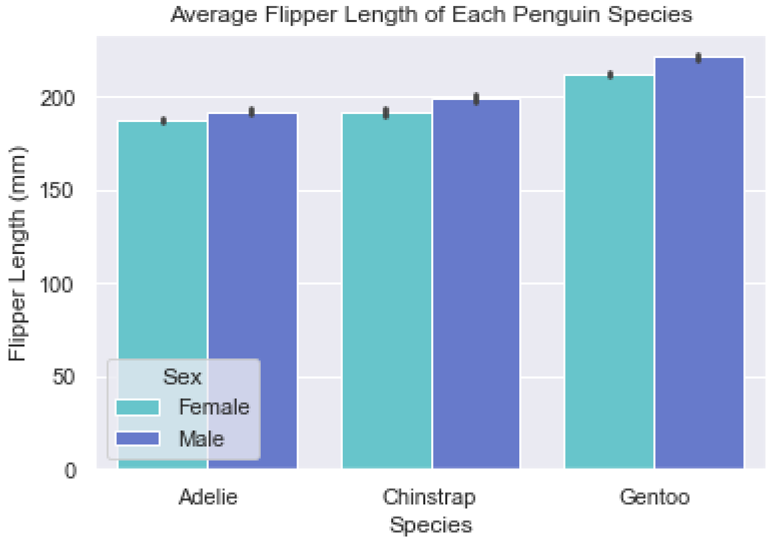


In [ ]:

```
g = sns.barplot(df,
                x="Species",
                y="Flipper Length (mm)",
                hue="Sex",
                palette=sex_color_palette)

g.set_title("Average Flipper Length of Each Penguin Species")
```

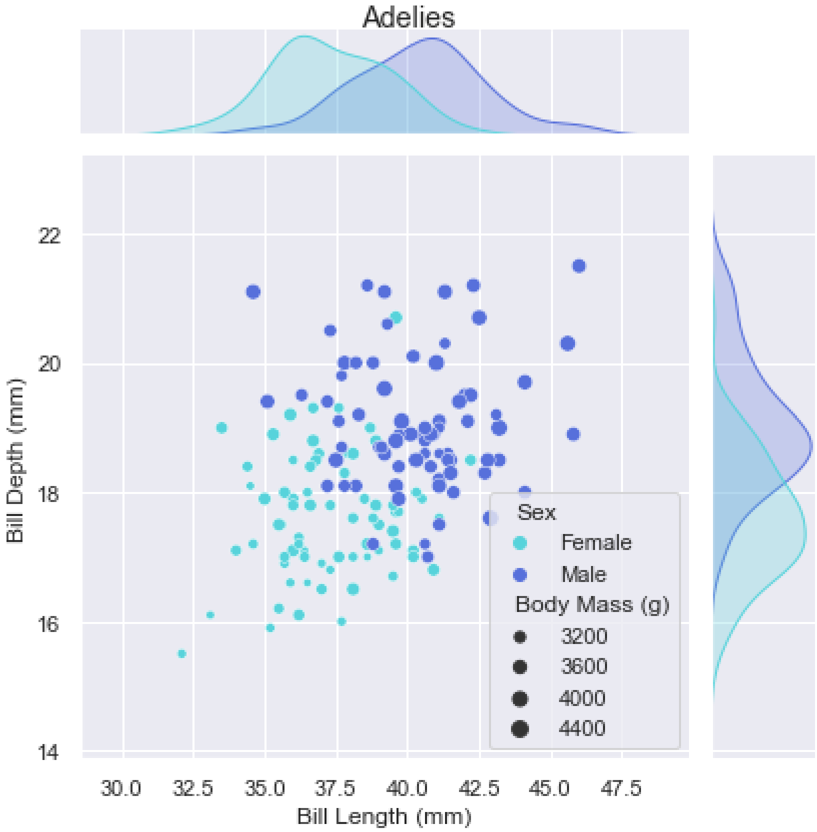
Out[ ]:

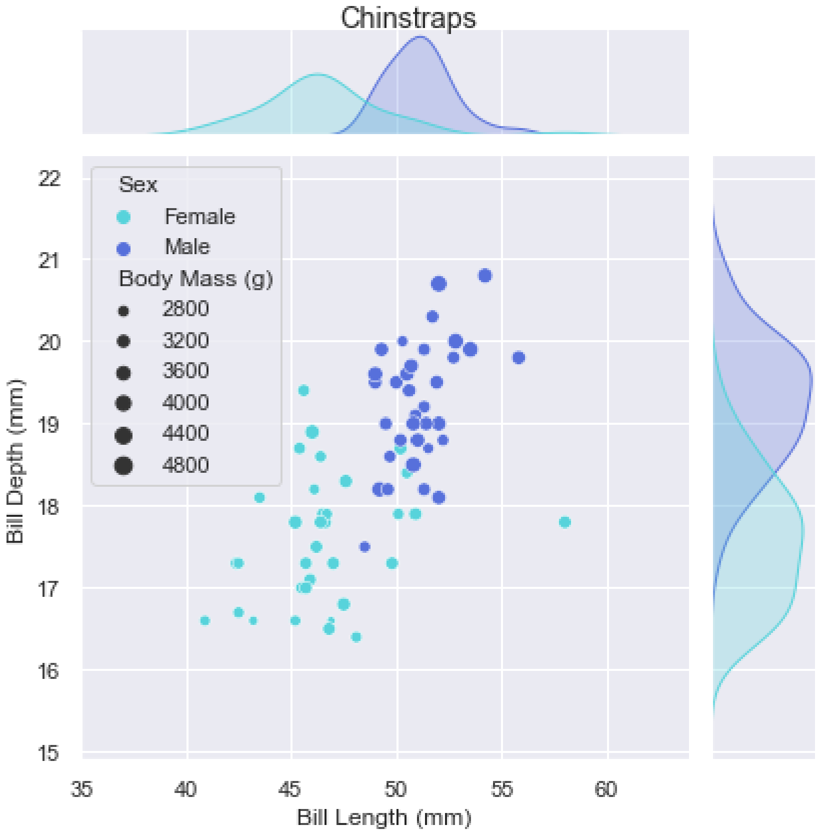
```
Text(0.5, 1.0, 'Average Flipper Length of Each Penguin Species')
```

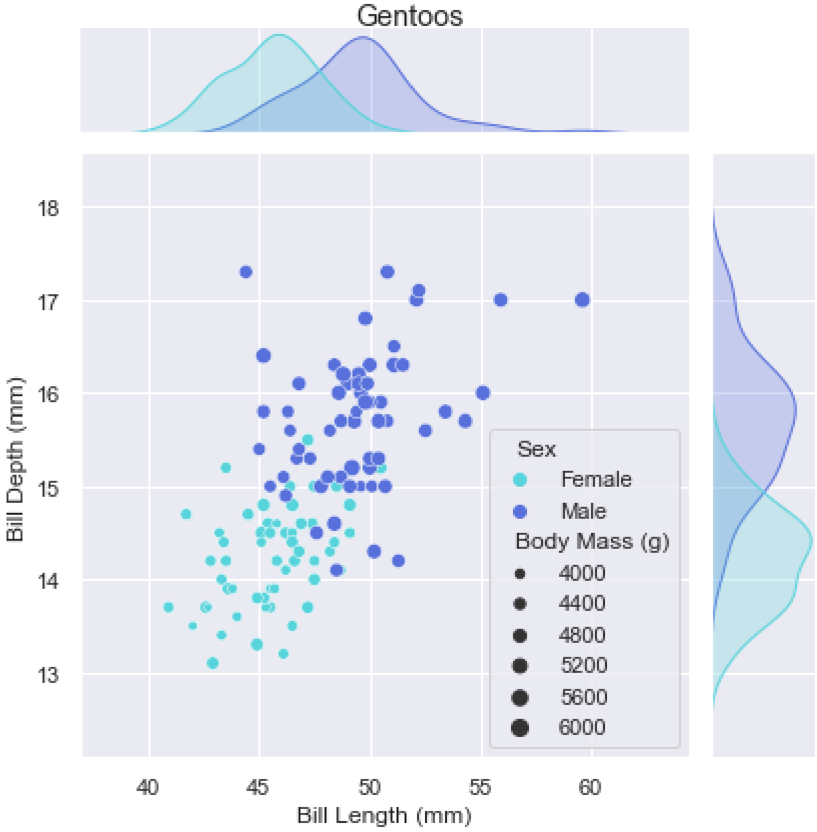


In [ ]:

```
species = df.Species.unique()

for spec in species:
    sub = df[(df.Species == spec)]

    g = sns.jointplot(sub,
                      x="Bill Length (mm)",
                      y="Bill Depth (mm)",
                      hue="Sex",
                      palette=sex_color_palette)

    g.ax_joint.cla()
    sns.scatterplot(sub,
                    x="Bill Length (mm)",
                    y="Bill Depth (mm)",
                    hue="Sex",
                    size="Body Mass (g)",
                    palette=sex_color_palette)

    g.fig.suptitle(spec + "s")
    g.fig.subplots_adjust(top=0.95)
```







Differences Between Islands

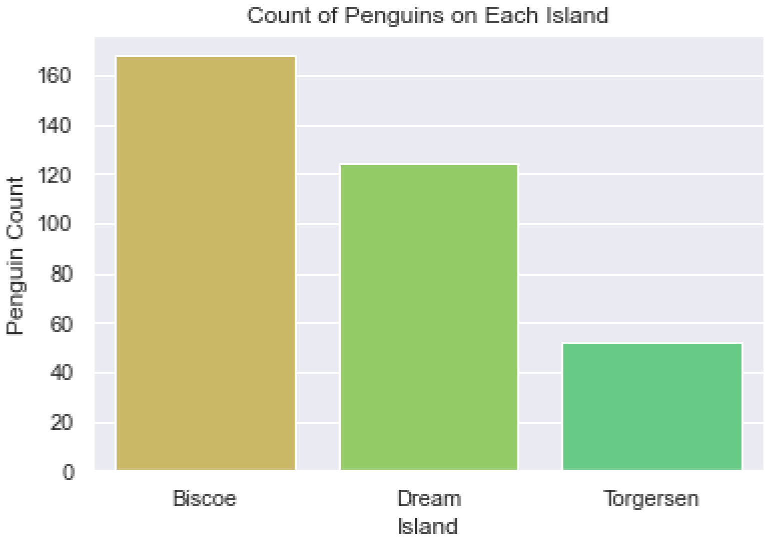
After gaining some insight on each penguin species, I was able to move on to my ultimate goal - comparison between the islands. Biscoe holds more penguins than Dream, which holds more than Torgersen. Each island has an approximately even split of males and females. Gentoo penguins only live on Biscoe, Chinstrap penguins only live on Dream, and Adelie penguins live on all 3 islands. There are more Gentoo than Adelie on Biscoe, and more Chinstrap than Adelie on Dream. The numbers of Adelie on each island are approximately equal. Without knowing more about each of the islands, it is hard to say what causes these different species distributions.

In [ ]:

```
g = sns.countplot(df,
                  x="Island",
                  palette=island_color_palette)

g.set_title("Count of Penguins on Each Island")

plt.ylabel("Penguin Count")
```

Out[ ]:

```
Text(0, 0.5, 'Penguin Count')
```

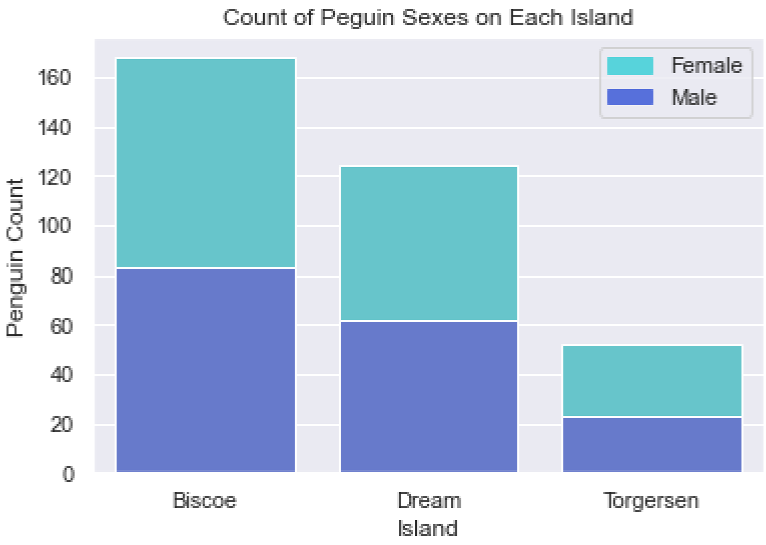


In [ ]:

```
plt.figure()
plt.title("Count of Peguin Sexes on Each Island")

females = sns.countplot(df,
                        x="Island",
                        color=sex_color_palette[0])

males = sns.countplot(df[df.Sex == "Male"],
                        x="Island",
                        color=sex_color_palette[1])

# Legend
female_legend = mpatches.Patch(color=sex_color_palette[0],
                               label="Female")
male_legend = mpatches.Patch(color=sex_color_palette[1],
                               label="Male")
plt.legend(handles=[female_legend, male_legend])

plt.ylabel("Penguin Count")
```

Out[ ]:

```
Text(0, 0.5, 'Penguin Count')
```

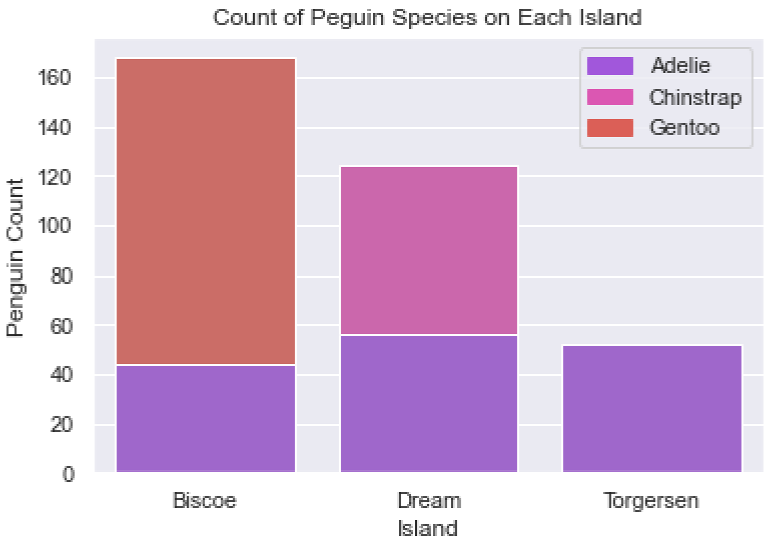


In [ ]:

```
plt.figure()
plt.title("Count of Peguin Species on Each Island")

gentoo = sns.countplot(df,
                       x="Island",
                       color=species_color_palette[2])

chinstrap = sns.countplot(df[(df.Species == "Chinstrap") |
                          (df.Species == "Adelie")],
                          x="Island",
                          color=species_color_palette[1])

adelie = sns.countplot(df[df.Species == "Adelie"],
                       x="Island",
                       color=species_color_palette[0])

# Legend
adelie_legend = mpatches.Patch(color=species_color_palette[0],
                               label="Adelie")
chinstrap_legend = mpatches.Patch(color=species_color_palette[1],
                                  label="Chinstrap")
gentoo_legend = mpatches.Patch(color=species_color_palette[2],
                               label="Gentoo")
plt.legend(handles=[adelie_legend, chinstrap_legend, gentoo_legend])

plt.ylabel("Penguin Count")
```

Out[ ]:

```
Text(0, 0.5, 'Penguin Count')
```



Variations in Adelie Between Islands

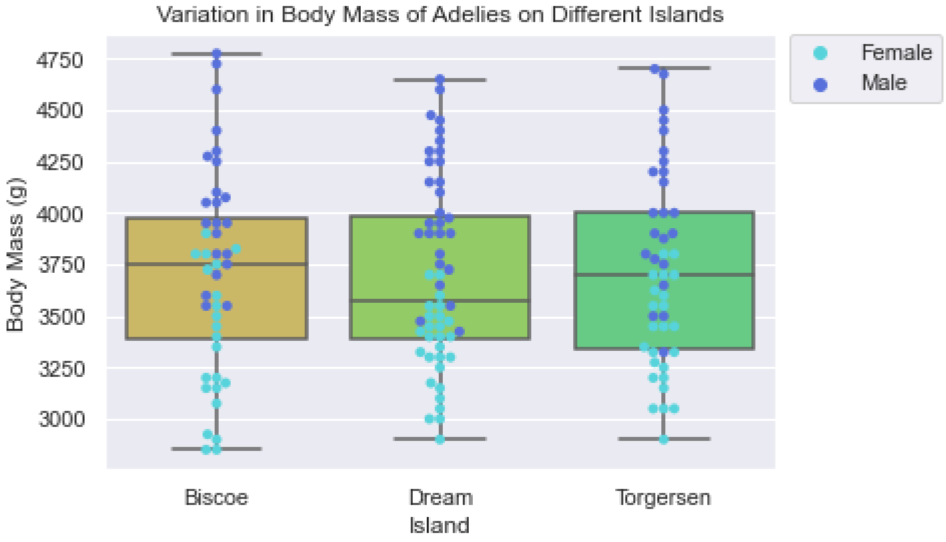
Because Adelie are the only penguins present on all 3 islands, analyzing differences between the island populations could provide clues about each ecosystem. Analyzing the data this way, I was not able to find any significant differences between the Adelie on each island. This suggests that the islands are approximately equivalent habitats for the Adelie. However, there must be some explanation for the different distributions between penguin species and, unfortunately, this analysis did not provide any clues. Perhaps the total number of penguins on each island is the only clue regarding the differences between ecosystems that can be gleaned from this dataset. If I knew the size of each island, I might be able to gain more insight.

In [ ]:

```
plt.figure()
plt.title("Variation in Body Mass of Adelies on Different Islands")

sns.boxplot(df[df.Species == "Adelie"],
                   x="Island",
                   y="Body Mass (g)",
                   palette=island_color_palette)

sns.swarmplot(df[df.Species == "Adelie"],
                   x="Island",
                   y="Body Mass (g)",
                   hue="Sex",
                   palette=sex_color_palette)

plt.legend(bbox_to_anchor=(1.02, 1), loc=2, borderaxespad=0.)
```

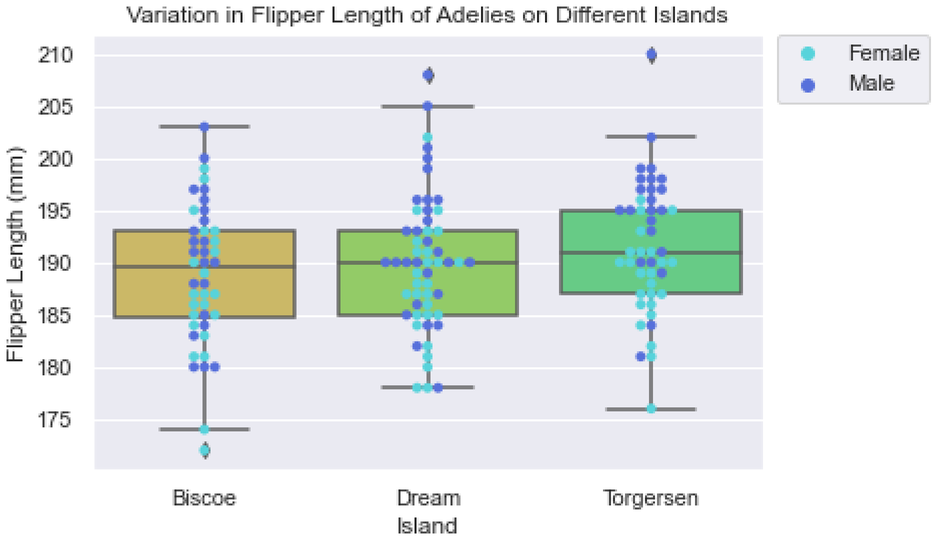
Out[ ]:

```
<matplotlib.legend.Legend at 0x1e427326fd0>
```



```
In [ ]: plt.figure()
        plt.title("Variation in Flipper Length of Adelies on Different Islands")

        sns.boxplot(df[df.Species == "Adelie"],
                          x="Island",
                          y="Flipper Length (mm)",
                          palette=island_color_palette)

        sns.swarmplot(df[df.Species == "Adelie"],
                          x="Island",
                          y="Flipper Length (mm)",
                          hue="Sex",
                          palette=sex_color_palette)

        plt.legend(bbox_to_anchor=(1.02, 1), loc=2, borderaxespad=0.)
```

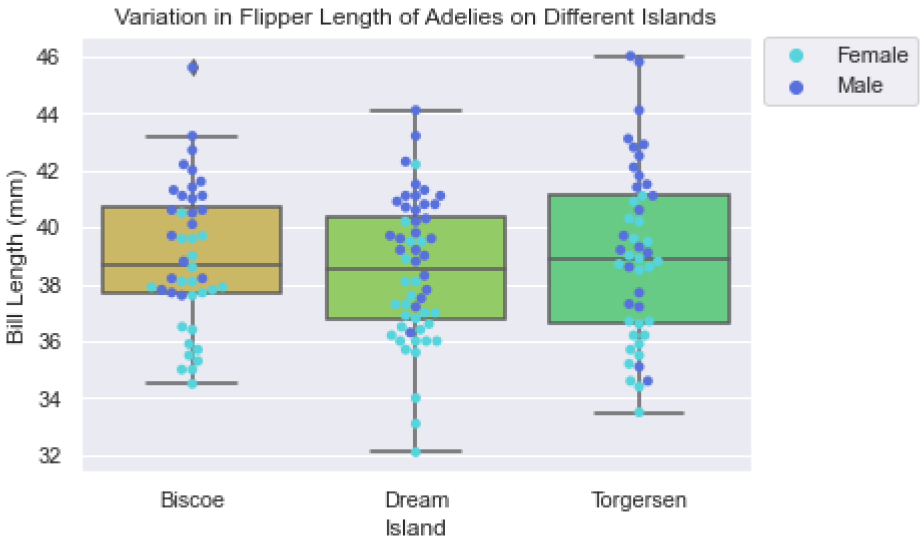
```
Out[ ]: <matplotlib.legend.Legend at 0x1e427375ca0>
```



```
In [ ]: plt.figure()
        plt.title("Variation in Flipper Length of Adelies on Different Islands")

        sns.boxplot(df[df.Species == "Adelie"],
                          x="Island",
                          y="Bill Length (mm)",
                          palette=island_color_palette)

        sns.swarmplot(df[df.Species == "Adelie"],
                          x="Island",
                          y="Bill Length (mm)",
                          hue="Sex",
                          palette=sex_color_palette)

        plt.legend(bbox_to_anchor=(1.02, 1), loc=2, borderaxespad=0.)
```

Out[ ]:

```
<matplotlib.legend.Legend at 0x1e4286f48b0>
```

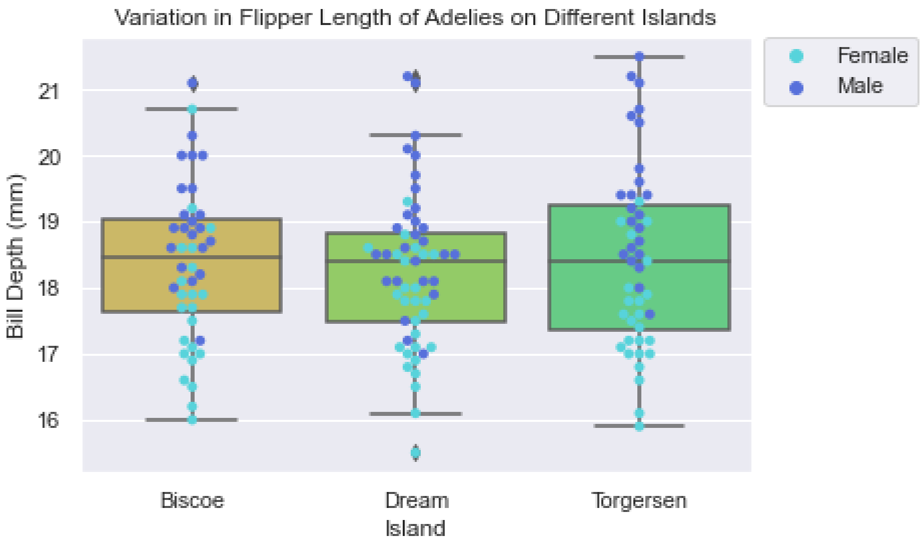


In [ ]:

```
plt.figure()
plt.title("Variation in Flipper Length of Adelies on Different Islands")

sns.boxplot(df[df.Species == "Adelie"],
                   x="Island",
                   y="Bill Depth (mm)",
                   palette=island_color_palette)

sns.swarmplot(df[df.Species == "Adelie"],
                   x="Island",
                   y="Bill Depth (mm)",
                   hue="Sex",
                   palette=sex_color_palette)

plt.legend(bbox_to_anchor=(1.02, 1), loc=2, borderaxespad=0.)
```

Out[ ]:

```
<matplotlib.legend.Legend at 0x1e425705370>
```



References

Gorman KB, Williams TD, Fraser WR (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). PLoS ONE 9(3):e90081. https://doi.org/10.1371/journal.pone.0090081