{Use this .ipynb template to write the report, open it with Jupyter-lab}

Assignment-3: Data Visualization with Report in Jupyter Notebook, Python Packs and Markdown

{Ji Woong Kim}, ISTE-782, Fall 2022

Summary

This document species analyzes and specifies the penguins species(Adelie, Gentoo and Chinstrap) and their body features. Regarding the species, the Adeline consists of the most number of individuals. The most individuals live in the Biscoe island, and Gentoo species only resides in the Biscoe island. Close number of Chinstrap and Adelie both reside in the Dream island. And Adelie is the only species found from the Torgersen island. Regarding the body features, all three species show that the positive relationship between body mass and flipper length. Also, the Chinstrap species has the biggest bill out of three species.

{ **Your tasks:**

- Take a look at this doc and read the instructions.
- Choose a dataset from seaborn package's repository.
- Based on your goal, do explaratory data analysis (eda) with visual components in the Seaborn pack. You can use other packs to create static graphics or to clean/organize data.
- Practice the visualization codes in the first three workshops. Include some here if relevant to the goal.
- Browse the seaborn gallery (it is in the seaborn workshop). Probably including five modern and multivariate plots should be ok.
- Use Markdown to write a professional report that has topics, subtopics, table of contents, references, full sentences etc.
- Give a name to this .ipynb notebook for the report that includes your name and dataset name like {yusuf}-{iris}.
- Update/add topic titles, include Python codes and show in the report, remove my comments given in {}.
- Once you finish the work, run all cells, save and export it as pdf. Your submission on myCourses will include both the .ipynb notebook and the pdf file. Export as pdf will require some packages. However, you can use the web browser feature of print to make pdf. Make sure your pdf report shows all writing and visualizations.
- Write a paragraph about what you learnt and the impression you got about the capacist of the pack.
- And write the summary at the top. }

Table of Contents

Introduction

I want to highlight the species and its body characteristics. Each species would have its characteristic of flipper and bill. And I believe the correlation would be implied in the dataset. From this report, I aim to visualize the relationship between feature of bill, flipper and mass. Plus, I also want to show the correlation between habitats and species.

Data Set

In [305…

```
import seaborn as sns
import pandas as pd
import matplotlib as plt
import numpy as np
#import pandoc #you will need this installed and imported to generate pdf
from palmerpenguins import load_penguins
sns.set_style('whitegrid')
```

In [306…

```
penguins = load_penguins()
penguins.head()
```

Out[306…

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	female

Before Clearning dataset

In [274…

```
print("Is there NaN Value in 'species' columns           : ",
penguins['species'].isnull().values.any())
print("Is there NaN Value in 'island' columns            : ",
penguins['island'].isnull().values.any())
print("Is there NaN Value in 'bill_length_mm' columns    : ",
penguins['bill_length_mm'].isnull().values.any())
print("Is there NaN Value in 'bill_depth_mm' columns     : ",
penguins['bill_depth_mm'].isnull().values.any())
print("Is there NaN Value in 'flipper_length_mm' columns : ",
penguins['flipper_length_mm'].isnull().values.any())
print("Is there NaN Value in 'body_mass_g' columns       : ",
penguins['body_mass_g'].isnull().values.any())
print("Is there NaN Value in 'sex' columns               : ",
penguins['sex'].isnull().values.any())
print("Is there NaN Value in 'year' columns              : ",
penguins['year'].isnull().values.any())
```

```
Is there NaN Value in 'species' columns           :  False
Is there NaN Value in 'island' columns            :  False
Is there NaN Value in 'bill_length_mm' columns    :  True
Is there NaN Value in 'bill_depth_mm' columns     :  True
Is there NaN Value in 'flipper_length_mm' columns :  True
Is there NaN Value in 'body_mass_g' columns       :  True
Is there NaN Value in 'sex' columns               :  True
Is there NaN Value in 'year' columns              :  False
```

In [275…

```
print(penguins['species'].isna().sum())
print(penguins['island'].isna().sum())
print(penguins['bill_length_mm'].isna().sum())
print(penguins['bill_depth_mm'].isna().sum())
print(penguins['flipper_length_mm'].isna().sum())
print(penguins['body_mass_g'].isna().sum())
print(penguins['sex'].isna().sum())
print(penguins['year'].isna().sum())
```

```
0
0
2
2
2
2
11
0
```

After Clearning dataset

In [276…
```
penguins = penguins.dropna()
```

In [277…
```
print("Is there NaN Value in 'species' columns           : ",
penguins['species'].isnull().values.any())
print("Is there NaN Value in 'island' columns            : ",
penguins['island'].isnull().values.any())
print("Is there NaN Value in 'bill_length_mm' columns    : ",
penguins['bill_length_mm'].isnull().values.any())
print("Is there NaN Value in 'bill_depth_mm' columns     : ",
penguins['bill_depth_mm'].isnull().values.any())
print("Is there NaN Value in 'flipper_length_mm' columns : ",
penguins['flipper_length_mm'].isnull().values.any())
print("Is there NaN Value in 'body_mass_g' columns       : ",
penguins['body_mass_g'].isnull().values.any())
print("Is there NaN Value in 'sex' columns               : ",
penguins['sex'].isnull().values.any())
print("Is there NaN Value in 'year' columns              : ",
penguins['year'].isnull().values.any())
```

```
Is there NaN Value in 'species' columns           :  False
Is there NaN Value in 'island' columns            :  False
Is there NaN Value in 'bill_length_mm' columns    :  False
Is there NaN Value in 'bill_depth_mm' columns     :  False
Is there NaN Value in 'flipper_length_mm' columns :  False
Is there NaN Value in 'body_mass_g' columns       :  False
Is there NaN Value in 'sex' columns               :  False
Is there NaN Value in 'year' columns              :  False
```

In [278…
```
penguins.describe()
```

Out[278…

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	year
count	333.000000	333.000000	333.000000	333.000000	333.000000
mean	43.992793	17.164865	200.966967	4207.057057	2008.042042
std	5.468668	1.969235	14.015765	805.215802	0.812944
min	32.100000	13.100000	172.000000	2700.000000	2007.000000
25%	39.500000	15.600000	190.000000	3550.000000	2007.000000
50%	44.500000	17.300000	197.000000	4050.000000	2008.000000
75%	48.600000	18.700000	213.000000	4775.000000	2009.000000
max	59.600000	21.500000	231.000000	6300.000000	2009.000000

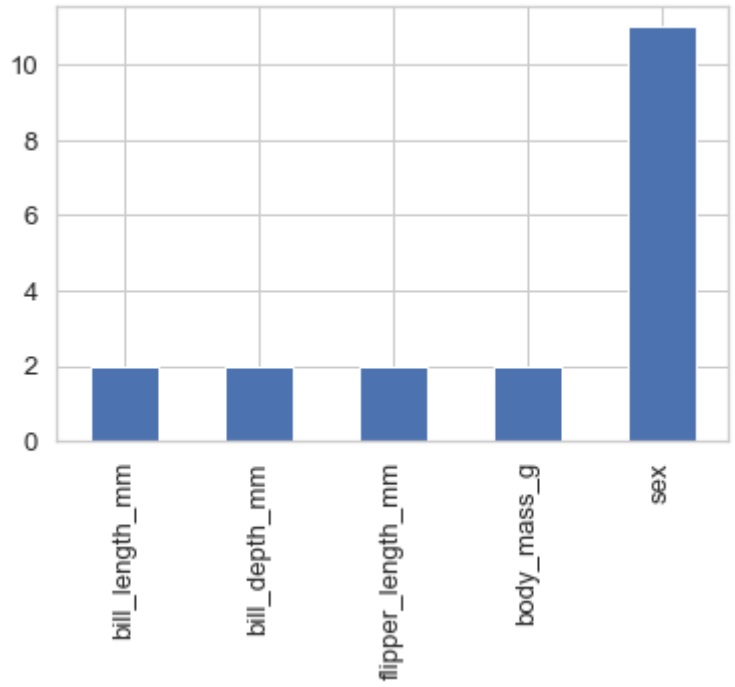
Data Visualization

Visualizing the Nan from the dataframe

In [279…
```
penguins = load_penguins()
```

In [280…
```
# penguins.isna().sum()[penguins.isna().sum()>0].plot(kind='bar')
penguins.isna().sum()[penguins.isna().sum() > 0].plot(kind='bar')
```

Out[280…
```
<AxesSubplot:>
```



Data Clearning

In [281…
```
penguins = penguins.dropna()
```

In [282…
```
penguins.head()
```

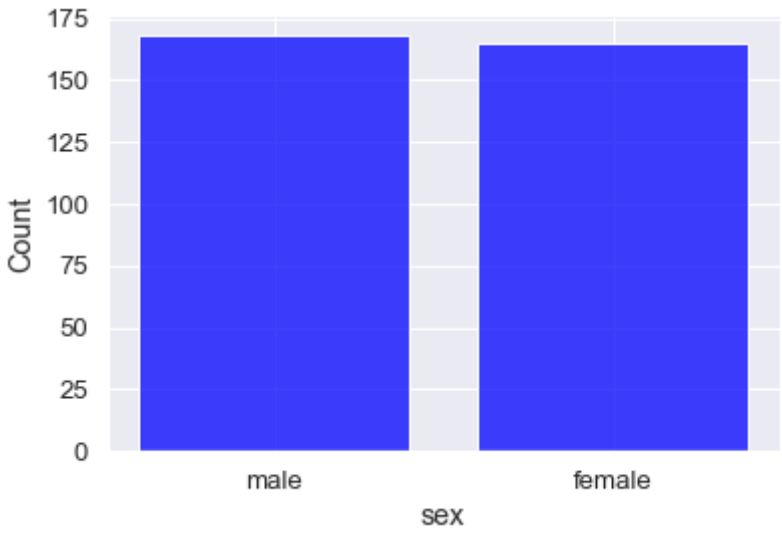
Out[282…

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	female
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	female
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0	male

Distribtion of Sex

In [283…
```
sns.set(font_scale = 1.2)
sns.histplot(data=penguins, x="sex", binwidth=3, shrink=.8, color =
"blue")
```

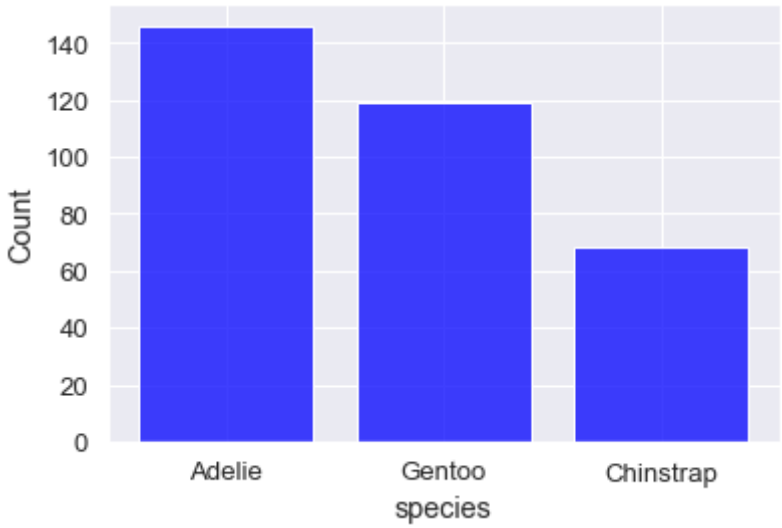
Out[283…
```
<AxesSubplot:xlabel='sex', ylabel='Count'>
```



Distribtion of Species

In [284…
```
sns.set(font_scale = 1.2)
sns.histplot(data=penguins, x="species", binwidth=3, shrink=.8, color
= "blue")
```

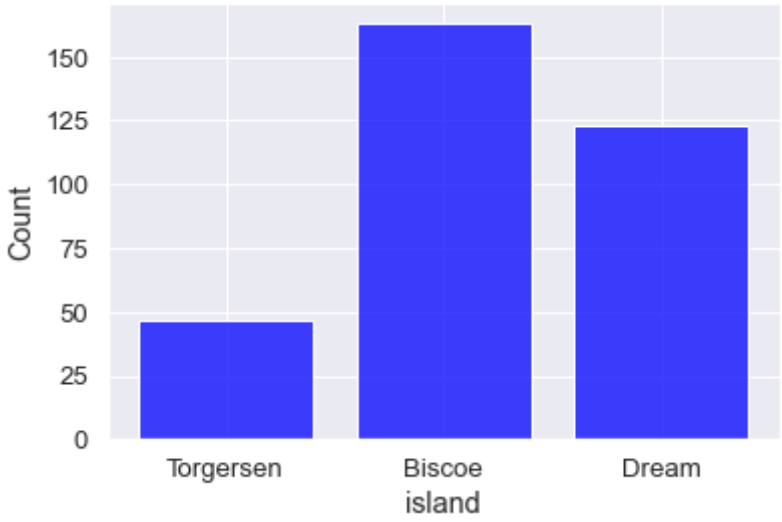
Out[284…
```
<AxesSubplot:xlabel='species', ylabel='Count'>
```



Distribtion of Islands

In [285…
```
sns.set(font_scale = 1.2)
sns.histplot(data=penguins, x="island", binwidth=3, shrink=.8, color
= "blue")
```

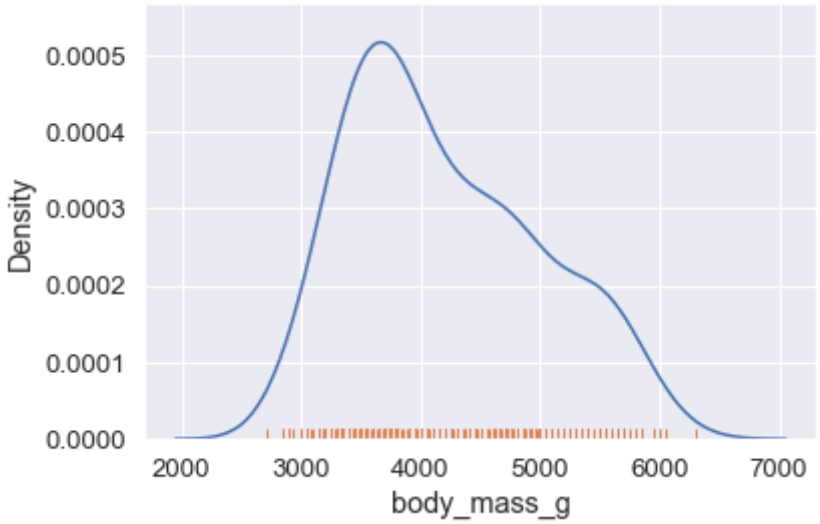
Out[285… <AxesSubplot:xlabel='island', ylabel='Count'>



Distribution of Body Mass

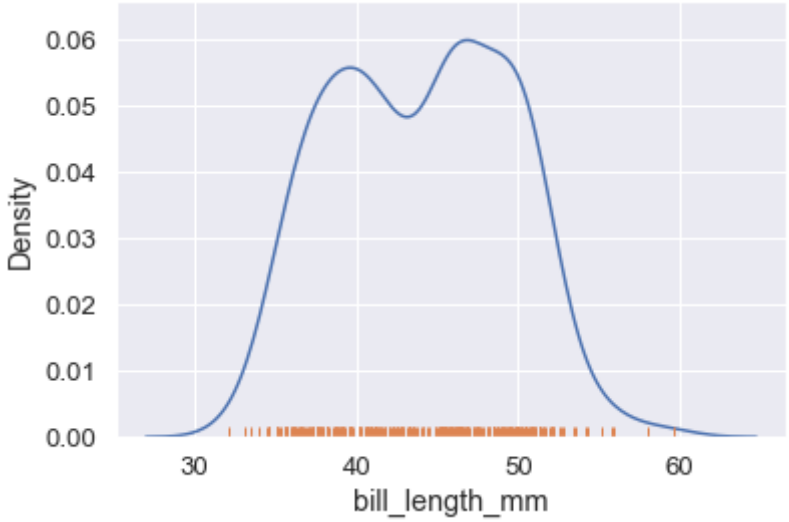
In [288…
```
sns.set(font_scale = 1.2)
sns.kdeplot(data=penguins, x="body_mass_g")
sns.rugplot(data=penguins, x="body_mass_g")
```

Out[288… <AxesSubplot:xlabel='body_mass_g', ylabel='Density'>



In [289…
```
sns.set(font_scale = 1.2)
sns.kdeplot(data=penguins, x="bill_length_mm")
sns.rugplot(data=penguins, x="bill_length_mm")
```

Out[289… <AxesSubplot:xlabel='bill_length_mm', ylabel='Density'>

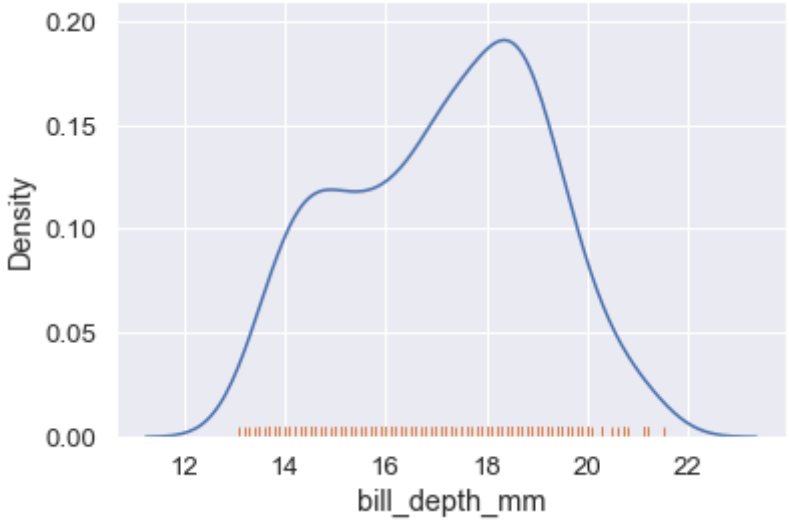


Distribution of Bill Depth

In [290…

```
sns.set(font_scale = 1.2)
sns.kdeplot(data=penguins, x="bill_depth_mm")
sns.rugplot(data=penguins, x="bill_depth_mm")
```

Out[290… <AxesSubplot:xlabel='bill_depth_mm', ylabel='Density'>



Relationship between Fliper and Body Mass

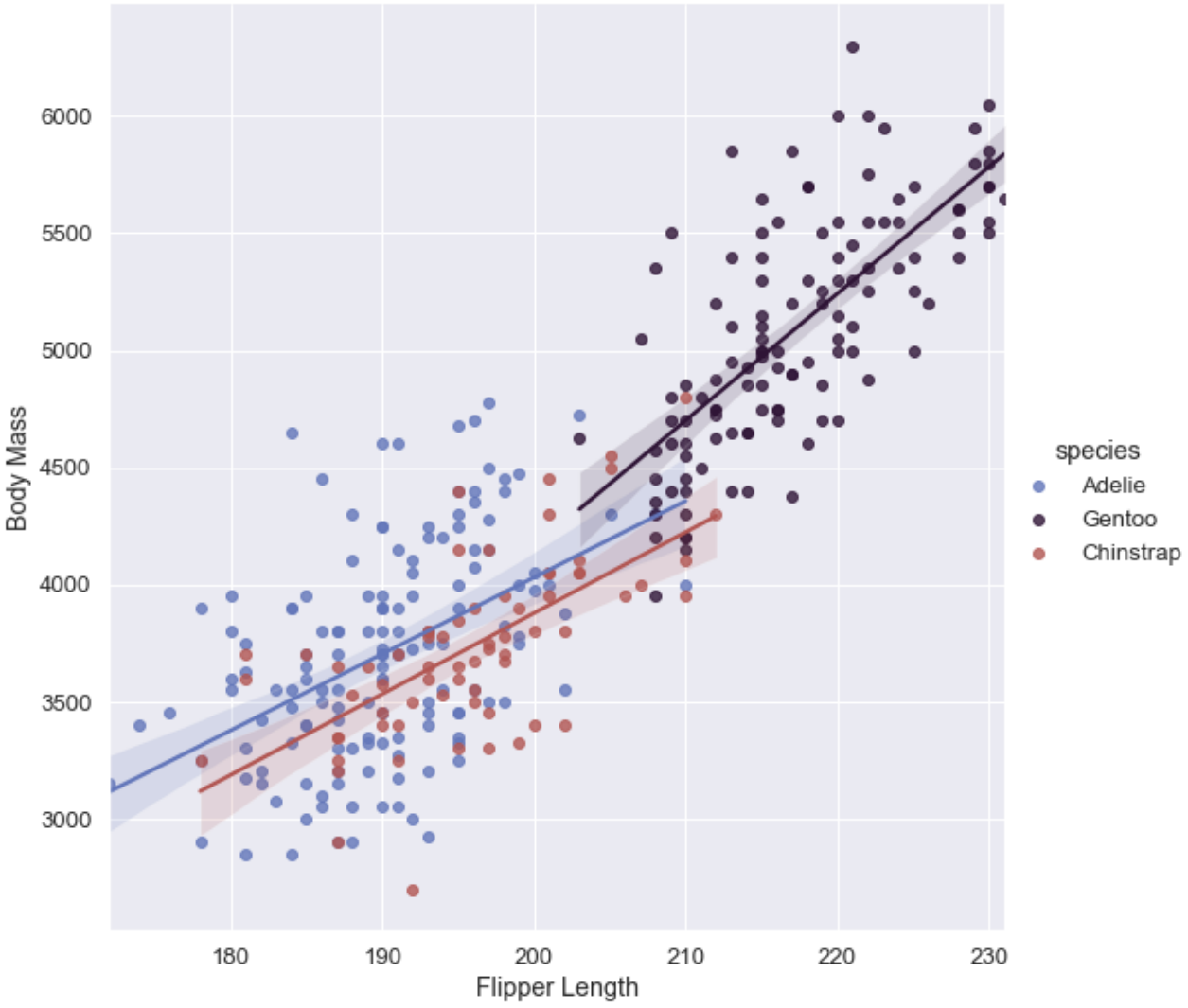
In [291…

```
sns.set(font_scale = 1.2)
plot = sns.lmplot(x="flipper_length_mm",
                  y="body_mass_g",
                  hue="species",
                  height=8,
                  data=penguins,
```

```
                    palette='twilight')
plot.set_xlabels('Flipper Length')
plot.set_ylabels('Body Mass')
```

Out [291…

```
<seaborn.axisgrid.FacetGrid at 0x7fc7d12d3e20>
```



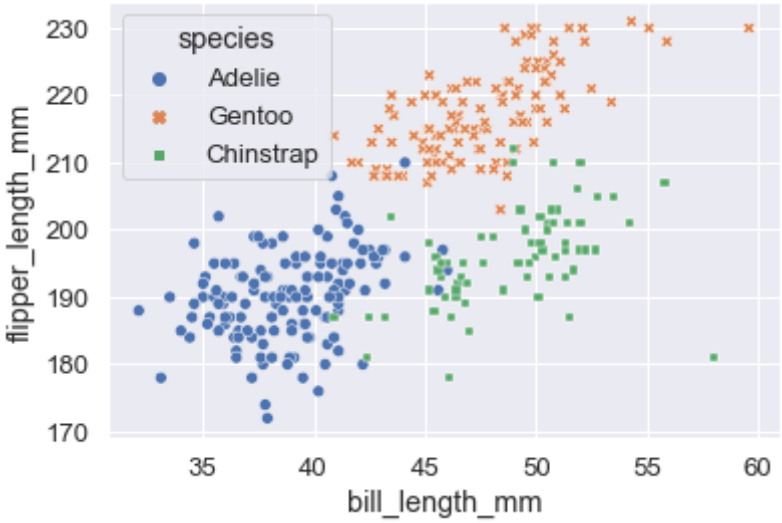
Relationship between Flipper Length and Bill Length

In [293…

```
sns.set(font_scale = 1.2)
sns.scatterplot(
    data=penguins, x="bill_length_mm", y="flipper_length_mm",
hue="species", style = "species", sizes=(20, 200), legend="full")
```

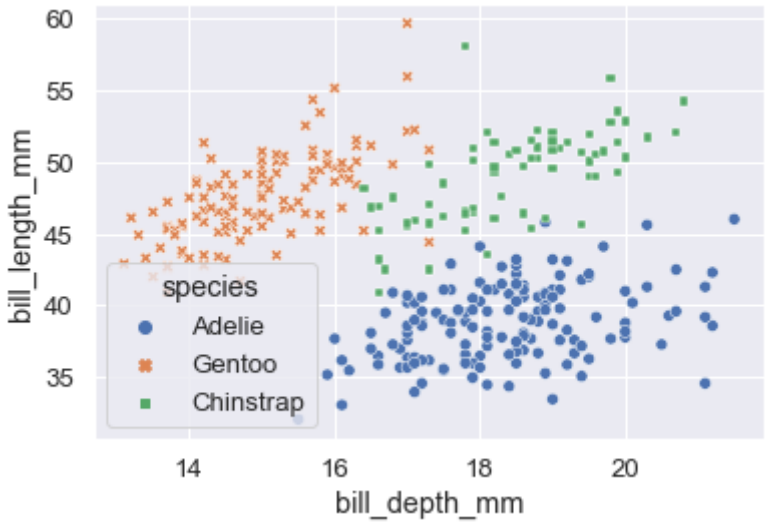
Out [293…

```
<AxesSubplot:xlabel='bill_length_mm', ylabel='flipper_length_mm'>
```



In [301…

```
sns.set(font_scale = 1.2)
plot = sns.scatterplot(
    data=penguins, x="bill_depth_mm", y="bill_length_mm",
hue="species", style = "species", sizes=(20, 200), legend="full")
```

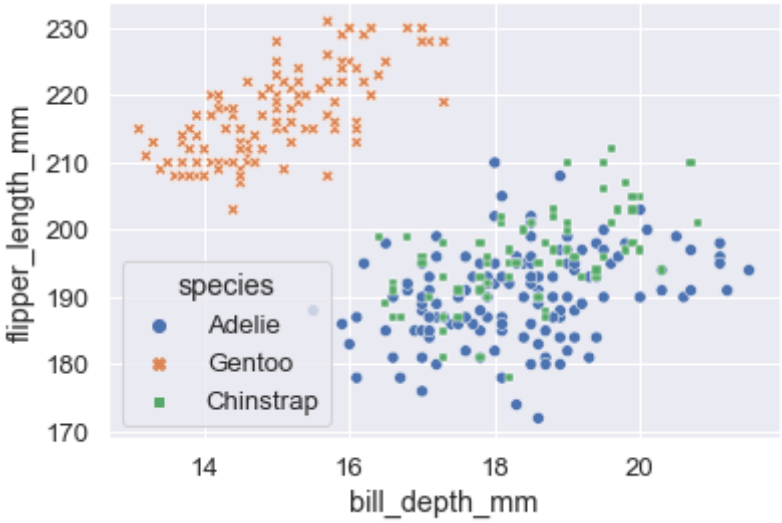


Relationship between Flipper Length and Bill Depth

In [298…

```
sns.set(font_scale = 1.2)
sns.scatterplot(
    data=penguins, x="bill_depth_mm", y="flipper_length_mm",
hue="species", style = "species",
    sizes=(20, 200), legend="full")
```

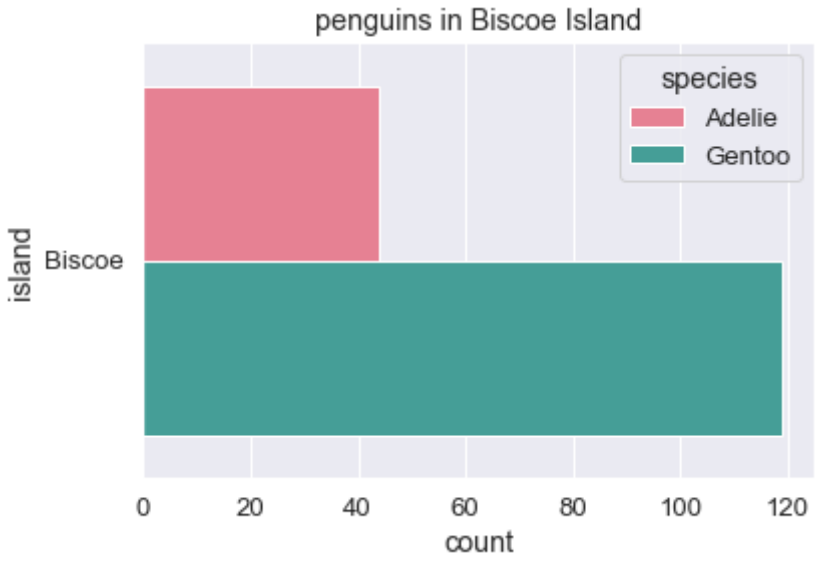
Out[298… <AxesSubplot:xlabel='bill_depth_mm', ylabel='flipper_length_mm'>



Habitat Information

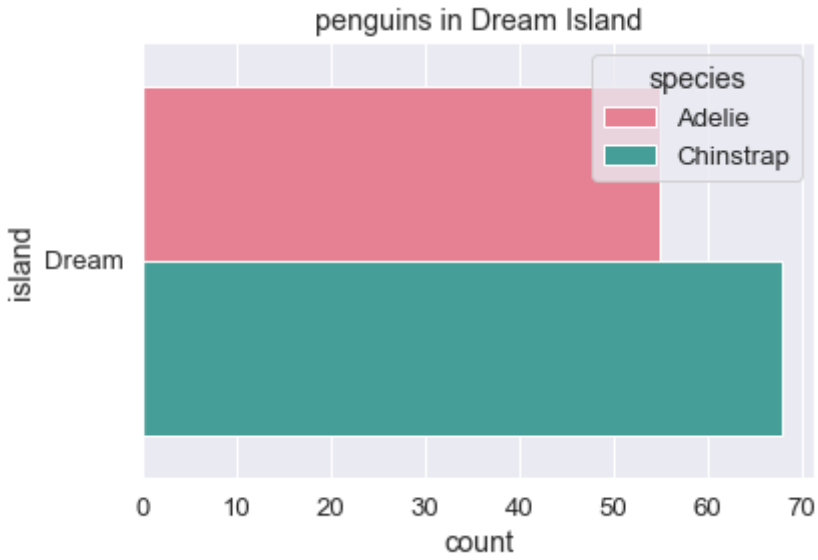
In [302…
```
island_spec = sns.countplot(data =
penguins[penguins["island"]=="Biscoe"], y = "island", hue =
"species", palette = 'husl')
island_spec.set(title = "penguins in Biscoe Island")
```

Out[302… `[Text(0.5, 1.0, 'penguins in Biscoe Island')]`



In [303…
```
island_spec = sns.countplot(data =
penguins[penguins["island"]=="Dream"], y = "island", hue = "species",
palette = 'husl')
island_spec.set(title = "penguins in Dream Island")
```

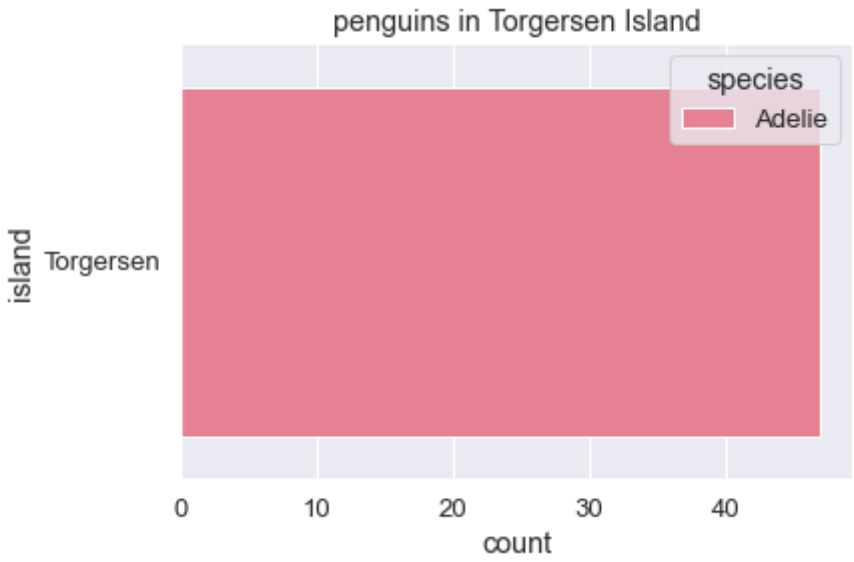
Out[303… `[Text(0.5, 1.0, 'penguins in Dream Island')]`



In [304…

```
island_spec = sns.countplot(data =
penguins[penguins["island"]=="Torgersen"], y = "island", hue =
"species", palette = 'husl')
island_spec.set(title = "penguins in Torgersen Island")
```

Out[304… `[Text(0.5, 1.0, 'penguins in Torgersen Island')]`



What I learnt

- On top of the technical parts, I tried to consider how to plot the graph as meaningful as possible.
- I tried to plot the distribution of species based on the island. I tried to preprocess the data by grouping the island. But I realized that there is a countplot that can plot the data using 'y' argument as island, and countplot itself can count the number of data.

- Also, I could learn the diverse function of pandas and Seaborn to preprocess and plot the data.

References

- Penguins Dataset (https://allisonhorst.github.io/palmerpenguins/)
- Seaborn Official Website (https://seaborn.pydata.org/index.html)