```
[1] import numpy as np
    import pandas as pd
    import seaborn as sns
    import matplotlib.pyplot as plt
    from matplotlib.pyplot import figure
    import matplotlib.gridspec as gridspec
    sns.set()
```

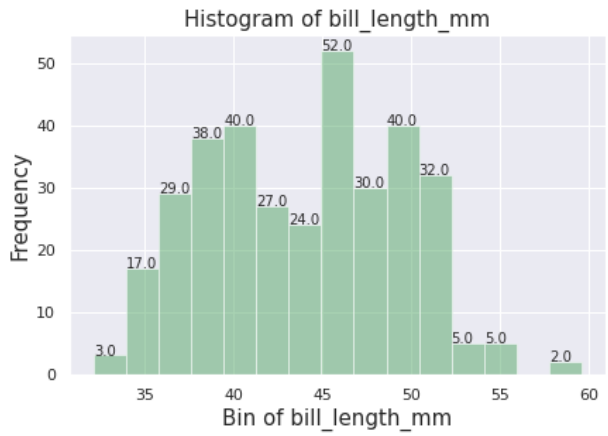
```
penguins = sns.load_dataset('penguins')
display(penguins.head())
print('\n', penguins.shape, '\n')
print(f"Min Bill Length: {penguins.bill_length_mm.min()} and Max Bill Length: {penguins.bill_length_mm.max()}")
print(f"Min Bill Depth: {penguins.bill_depth_mm.min()} and Max Bill Depth: {penguins.bill_depth_mm.max()}")
print(f"Min Flipper Length: {penguins.flipper_length_mm.min()} and Max Bill Flipper: {penguins.flipper_length_mm.max()}")
print(f"Min Body Mass: {penguins.body_mass_g.min()} and Max Body Mass: {penguins.body_mass_g.max()}\n")
print(f'Number of Unique Species : {list(penguins.species.unique())}')
print(f"<<<Count of Species>>> \n{penguins.species.value_counts()}\n")
print(f'Number of Unique Island : {list(penguins.island.unique())}')
print(f"<<<Count of Island>>> \n{penguins.island.value_counts()}\n")
print(f'Number of Unique Sex : {list(penguins.sex.unique())}')
print(f"<<<Count of Sex>>> \n{penguins.sex.value_counts()}")
```

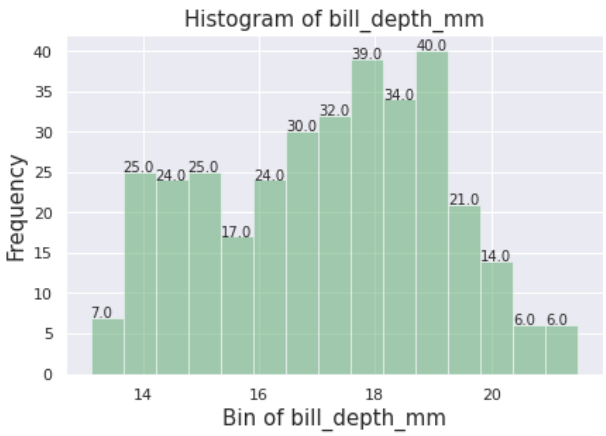
	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0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

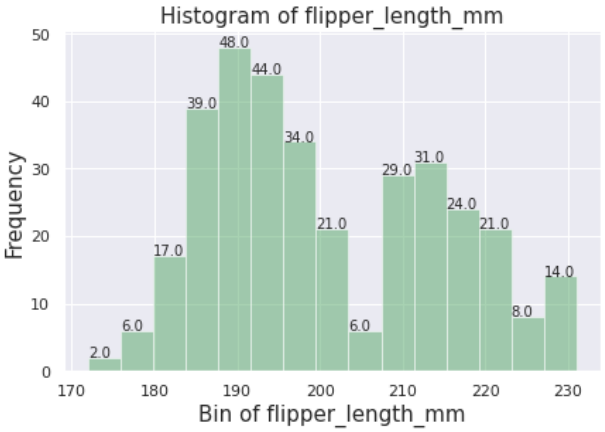
```
 (344, 7)

Min Bill Length: 32.1 and Max Bill Length: 59.6
Min Bill Depth: 13.1 and Max Bill Depth: 21.5
Min Flipper Length: 172.0 and Max Bill Flipper: 231.0
Min Body Mass: 2700.0 and Max Body Mass: 6300.0

Number of Unique Species : ['Adelie', 'Chinstrap', 'Gentoo']
<<<Count of Species>>>
Adelie       152
Gentoo       124
Chinstrap     68
Name: species, dtype: int64

Number of Unique Island : ['Torgersen', 'Biscoe', 'Dream']
<<<Count of Island>>>
Biscoe       168
Dream        124
Torgersen     52
Name: island, dtype: int64

Number of Unique Sex : ['Male', 'Female', nan]
<<<Count of Sex>>>
Male      168
Female    165
Name: sex, dtype: int64
```

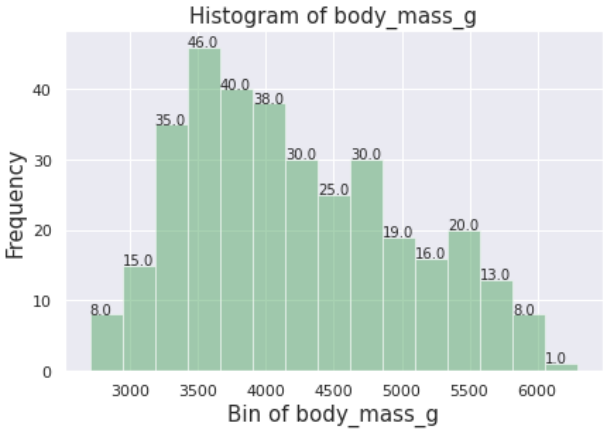
```
pen_copy = penguins.copy()
for i in pen_copy.columns:
  for j in range(len(pen_copy[i])):
    if i == 'sex':
      if pen_copy[i][j] not in ['Male', 'Female']:
        pen_copy[i][j] = 'Unknown'
    else:
      if pd.isna(pen_copy[i][j]):
        pen_copy[i][j] = '{:.1f}'.format(pen_copy[i].mean())
```

```
figure(figsize=(15, 15))

plot = 320
position = 1
for i in ["bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g"]:
  if (position % 3 == 0):
    position = 1
    plot += 3
  else:
    position += 1
    plot += 1
  plt.subplot(plot)
  plt.title(f"Histogram of {i}", size = 15)
  plt.xlabel(f"Bin of {i}", size = 15)
  plt.ylabel("Frequency", size = 15)
  density , bins, _ = plt.hist(pen_copy[i], bins = 15, alpha=0.5, color = "g")
  for x, y in zip(bins, density):
      if y != 0:
          plt.text(x, y+0.05, y, fontsize=10)

plt.show()
```









Bill_leng_mm: Nhiều ở các khoảng 40-42, 45-47, 49-50, càng xa các khoảng này thì tần số càng giảm.

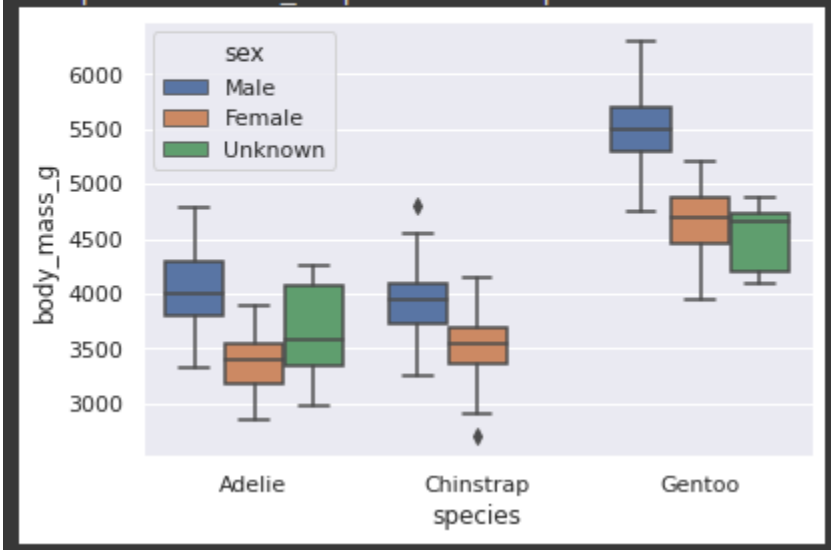
Bill_depth_mm: Tần số lệch phải, nhiều trong khoảng 17.5-19.

Flipper_length_mm: Biểu đồ có dạng chữ M, lệch trái, nhiều ở khoảng 190 và 215.

Body_mass_g: Lệch trái, nhiều ở khoảng 3500.

```
# boxplot
sns.boxplot(x="species", y="body_mass_g", hue="sex", data=pen_copy)
```

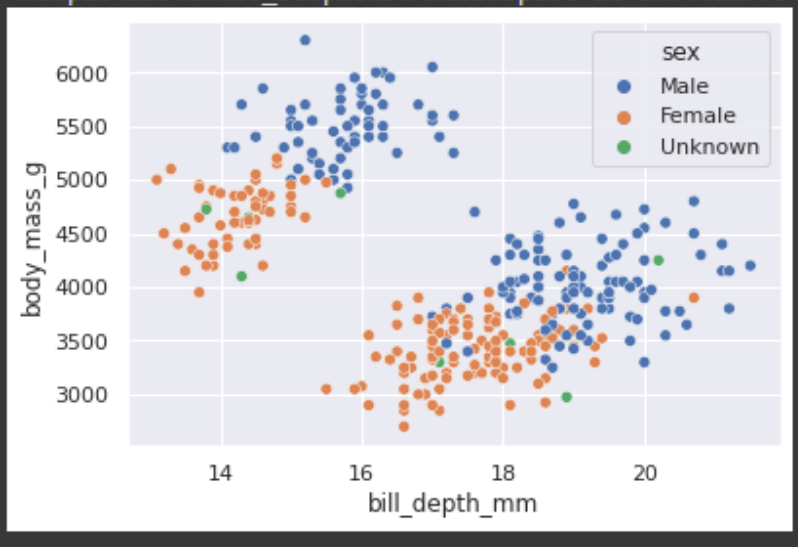
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8cfae9ea90>
```



Nhìn chung, con đực sẽ có body mass lớn hơn con cái. Body mass của loài Gentoo cao hơn hẳn 2 loài còn lai.

```
# scatter
sns.scatterplot(x="bill_depth_mm", y="body_mass_g", hue="sex", data=penguins)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fca17f34290>
```



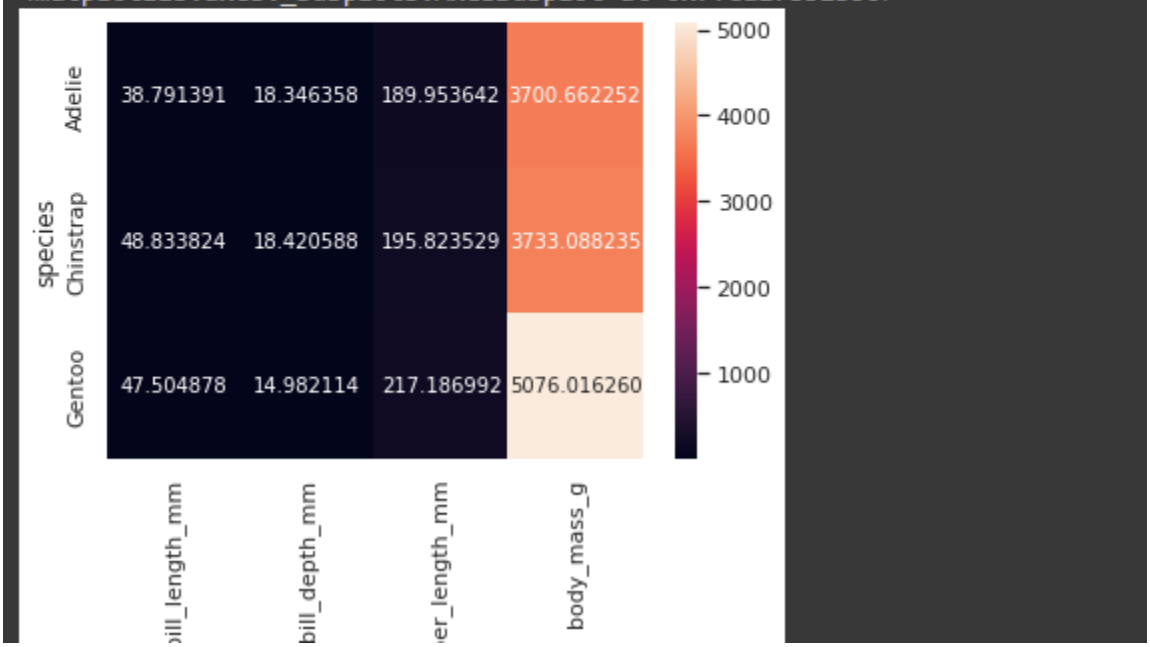
Có thể thấy biểu đổ chia ra làm 2 nhóm. Nhóm có bill_depth_mm trong khoảng 16.5 trở về trước có body_mass_g trung bình cao hơn nhóm có bill_depth_mm trong khoảng 16.5 trở về sau.

Con cái sẽ có bill_depth_mm và body_mass_g nhỏ hơn con đực trong cả 2 nhóm.

```
# heatmap
pen_groupby = pen_copy.groupby('species').mean()
display(pen_groupby)
sns.heatmap(pen_groupby, annot=True, fmt="f")
```

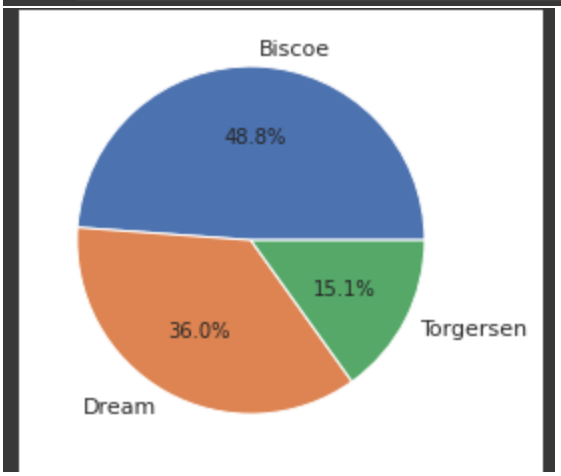
species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
Adelie	38.791391	18.346358	189.953642	3700.662252
Chinstrap	48.833824	18.420588	195.823529	3733.088235
Gentoo	47.504878	14.982114	217.186992	5076.016260

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fca17832350>
```



```
# pie
query = pen_copy.groupby('island', as_index = False).count()
display(query)
plt.pie(query['species'], labels = query['island'], autopct='%1.1f%%')
```

	island	species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Biscoe	168	167	167	167	167	168
1	Dream	124	124	124	124	124	124
2	Torgersen	52	51	51	51	51	52



So sánh tỉ lệ về số lượng cá thể giữa các island.