Homework 3 - Section 11 - Aaron_Yang

Question 1

"""

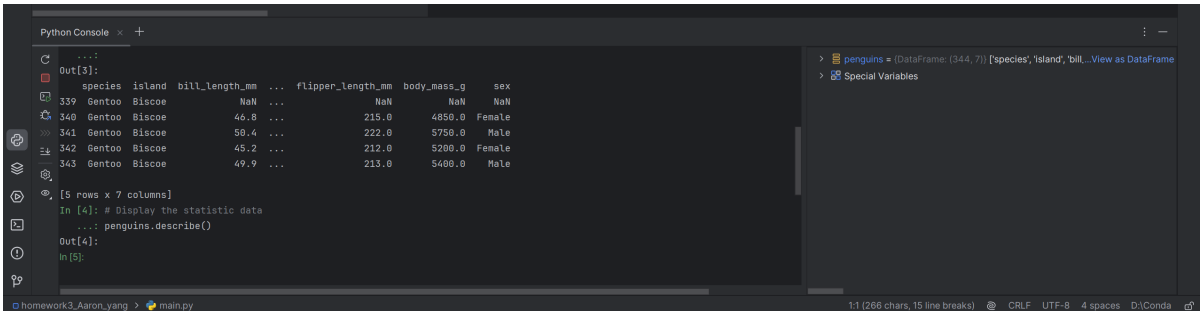
Load the 'penguins' dataset from the seaborn package. Display the last 5 observations. Display the dataset statistics. Hint: describe()

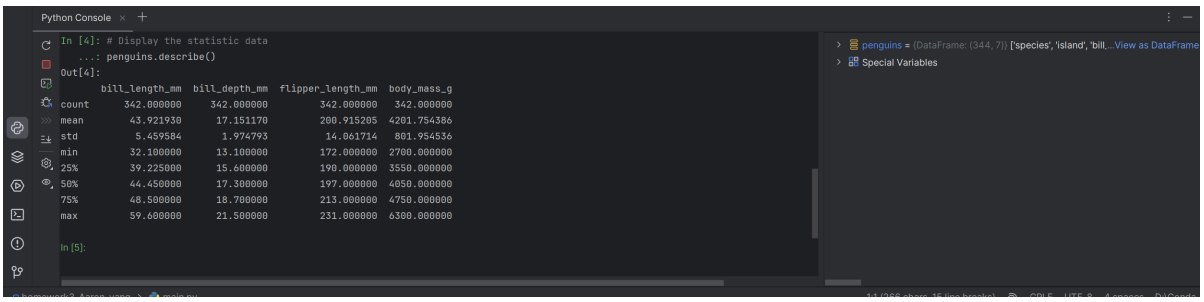
"""

Code

```
#%%
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

#%%
# Load the dataset
penguins = sns.load_dataset("penguins")

# Display the last 5 Observations
penguins.tail(5)

#%%
# Display the statistic data
penguins.describe()
```

Answer





Question 2

"""

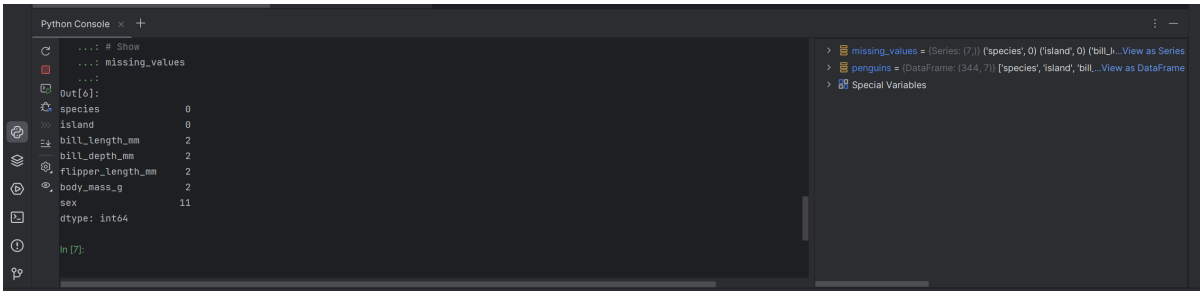
Dataset cleaning: Write a python program that check if the dataset is cleaned. If not, removed the missing observations from the data set. Display the portion of the code that perform the task here. Display the results that confirms the dataset is clean.

"""

Code

```
#%%
# Question 2
# Display the number of missing value
 penguins.isnull().sum()


#%%
# Clean the dataset
penguins = penguins.dropna()

# Show the missing values again
print(penguins.isnull().sum().sum())
```

Answer



As the screenshot shows, the dataset 'penguins' has been cleaned. The nan value count is zero.

Question 3

"""

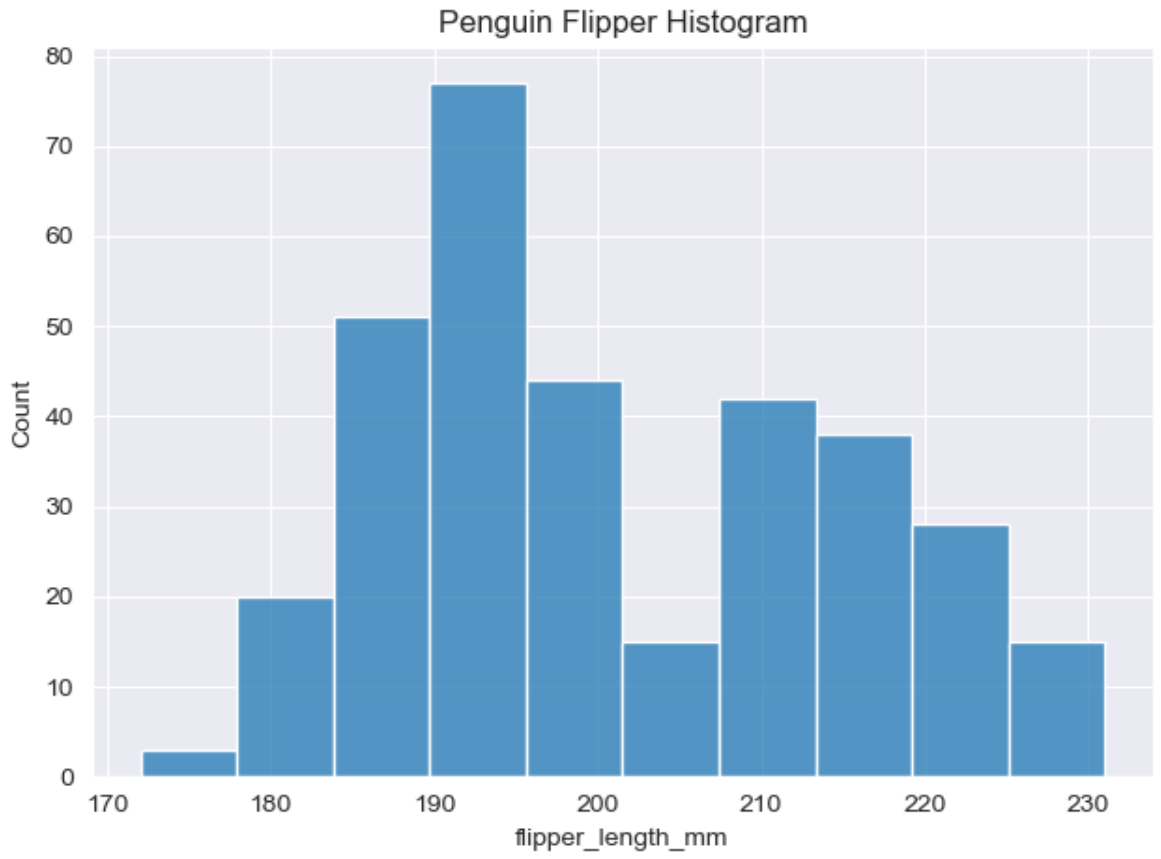
Using the seaborn package graph the histogram plot “flipper_length_mm”. Use the ‘darkgrid’ style as seaborn the theme. Write down your observation about the graph on the console.

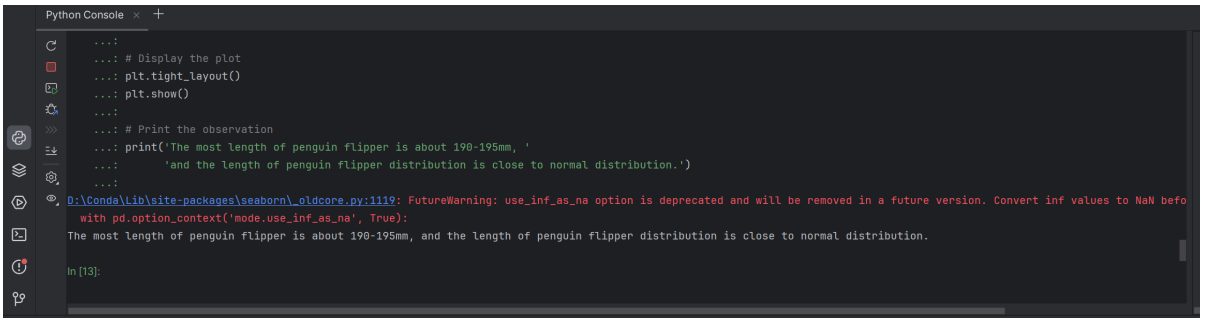
"""

Code

```
#%%
# Question 3
# Set the sns plot style
sns.set_style('darkgrid')

# Histogram plot for 'flipper_length_mm'
Q3_hist = sns.histplot(data=penguins['flipper_length_mm'])
Q3_hist.set_title('Penguin Flipper Histogram')

# Display the plot
plt.tight_layout()
plt.show()

# Print the observation
print('The most length of penguin flipper is about 190-195mm, '
      'and the length of penguin flipper distribution is close to normal
distribution.')
```

Answer





Question 4

"""

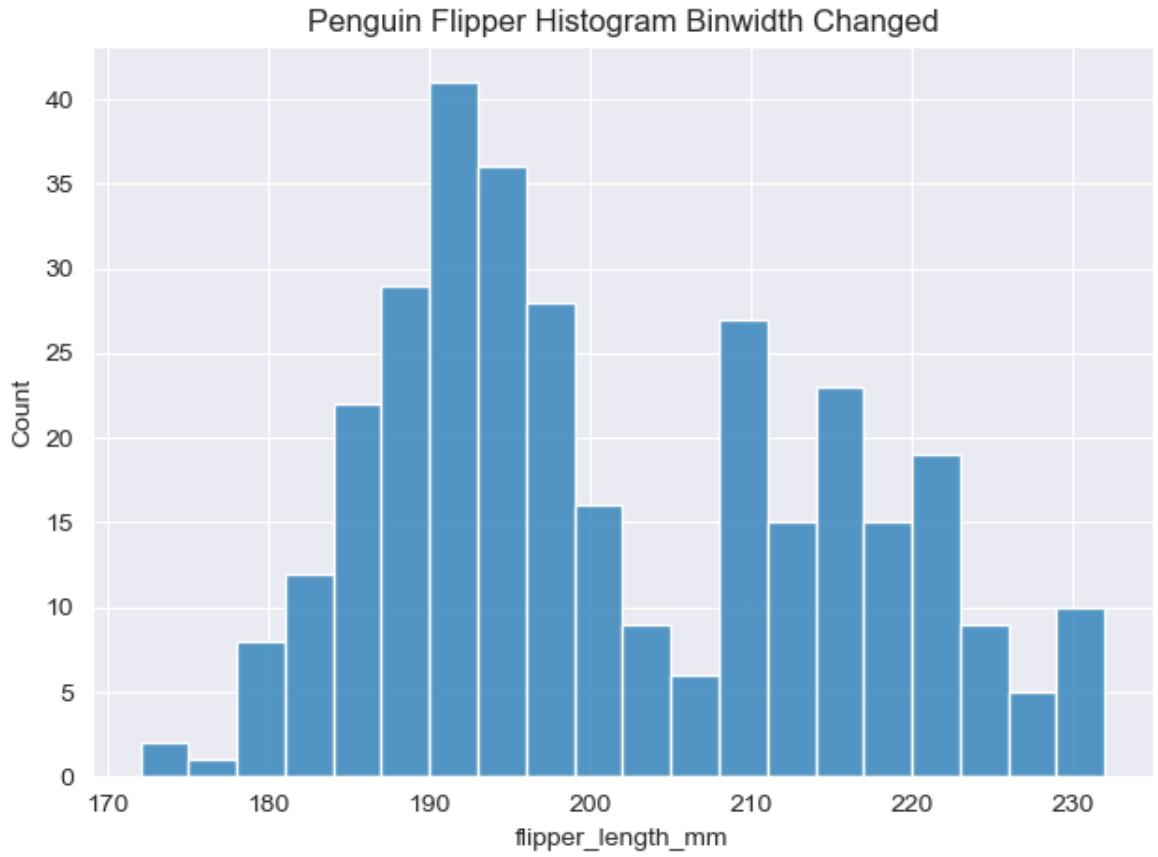
Change the bin width in the previous question to 3 and replot the graph. Write down your observation about the graph on the console. Hint: binwidth

"""

Code

```
#%%
# Question 4
# Set the sns plot style
sns.set_style('darkgrid')

# Histogram plot for 'flipper_length_mm' with binwidth = 3
Q3_hist = sns.histplot(data=penguins['flipper_length_mm'], binwidth=3)
Q3_hist.set_title('Penguin Flipper Histogram Binwidth Changed')

# Display the plot
plt.tight_layout()
plt.show()

# Print the observation
print('The most length of penguin flipper is about 190-192mm, '
      'while the least length of penguin flipper is 175-177mm.')
```

Answer



Question 5

"""

Change the bin numbers to 30 in the previous question and replot the graph. Write down your observation about the graph on the console. Hint: bins

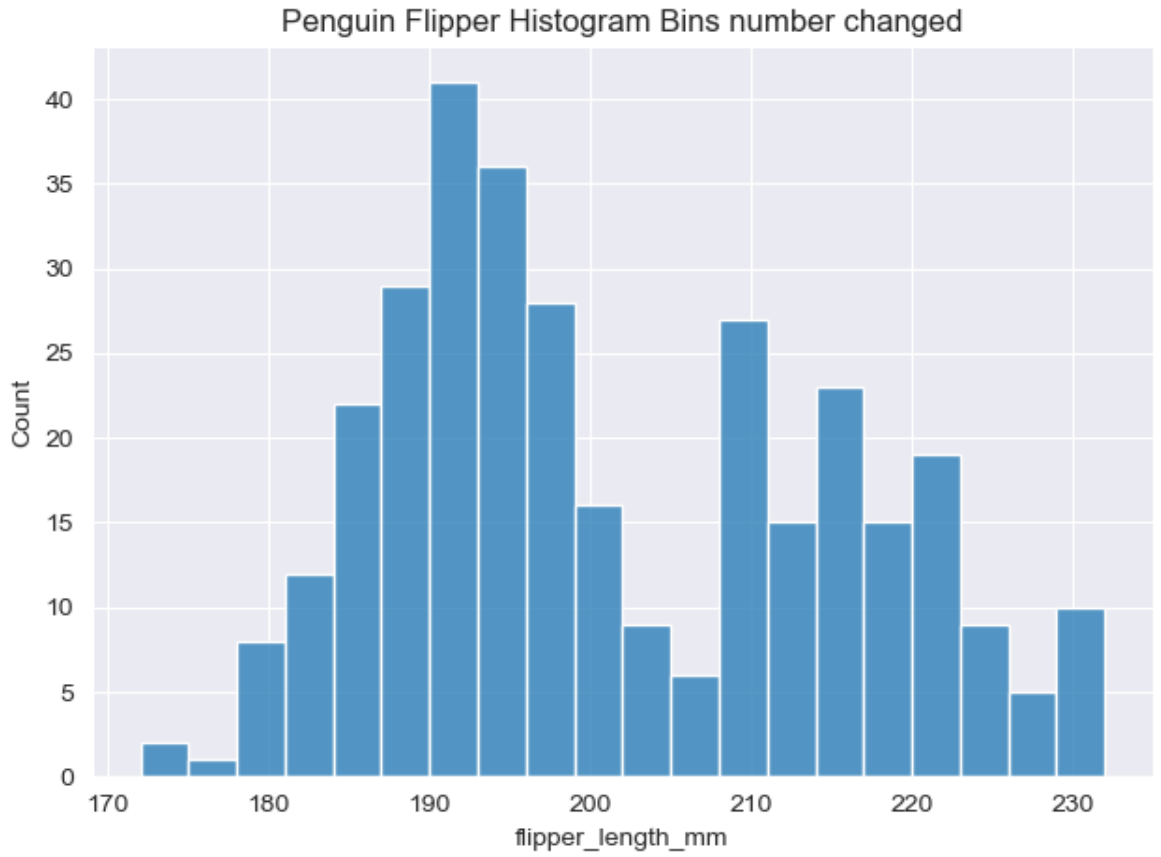
"""

Code

```
#%%
# Question 5
# Set the sns plot style
sns.set_style('darkgrid')

# Histogram plot for 'flipper_length_mm' with binwidth = 3
```

```
Q3_hist = sns.histplot(data=penguins['flipper_length_mm'], bins=30)
Q3_hist.set_title('Penguin Flipper Histogram Bins number changed')

# Display the plot
plt.tight_layout()
plt.show()

# Print the observation
print('The most length of penguin flipper is about 191mm, '
      'while the least length of penguin flipper is 172-173, 174, 177mm.')
```

Answer



Question 6

"""

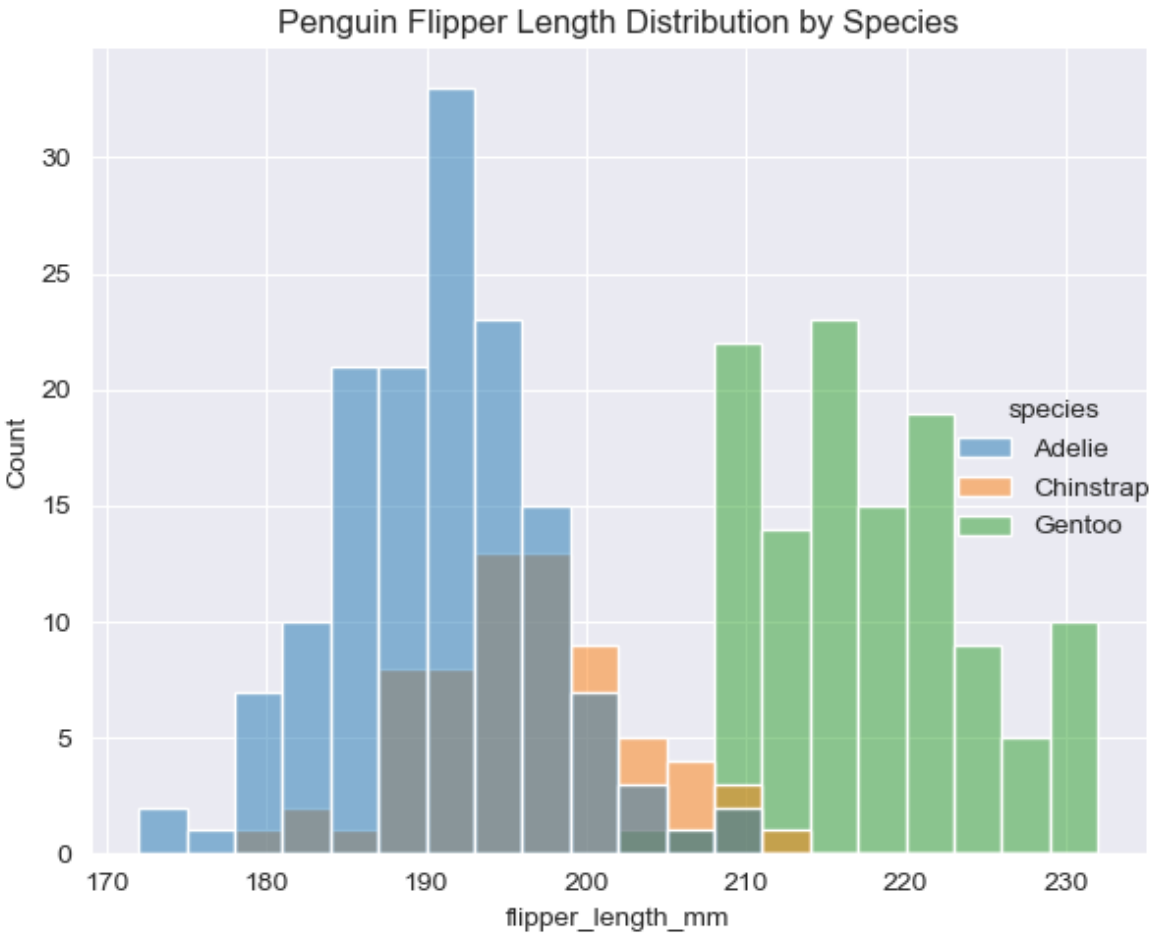
Using the seaborn “displot”, graph then histogram plot per the species. Hint: You need to use the ‘hue’ . Write down your observation about the graph on the console.

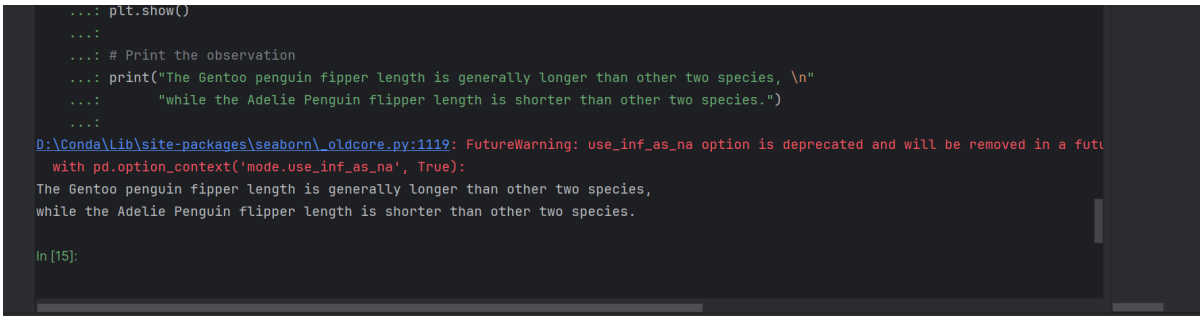
"""

Code

```
# Question 6
# Set the hue and displot
Q6_displot = sns.displot(data=penguins,
                         x='flipper_length_mm',
                         binwidth=3,
                         bins=30,
                         hue='species')

# Build the title for the displot
Q6_displot.set(title='Penguin Flipper Length Distribution by Species')

# Display the plot
plt.tight_layout()
plt.show()

# Print the observation
print("The Gentoo penguin fipper length is generally longer than other two
species, \n"
      "while the Adelie Penguin flipper length is shorter than other two
species.")
```

Answer





Question 7

"""

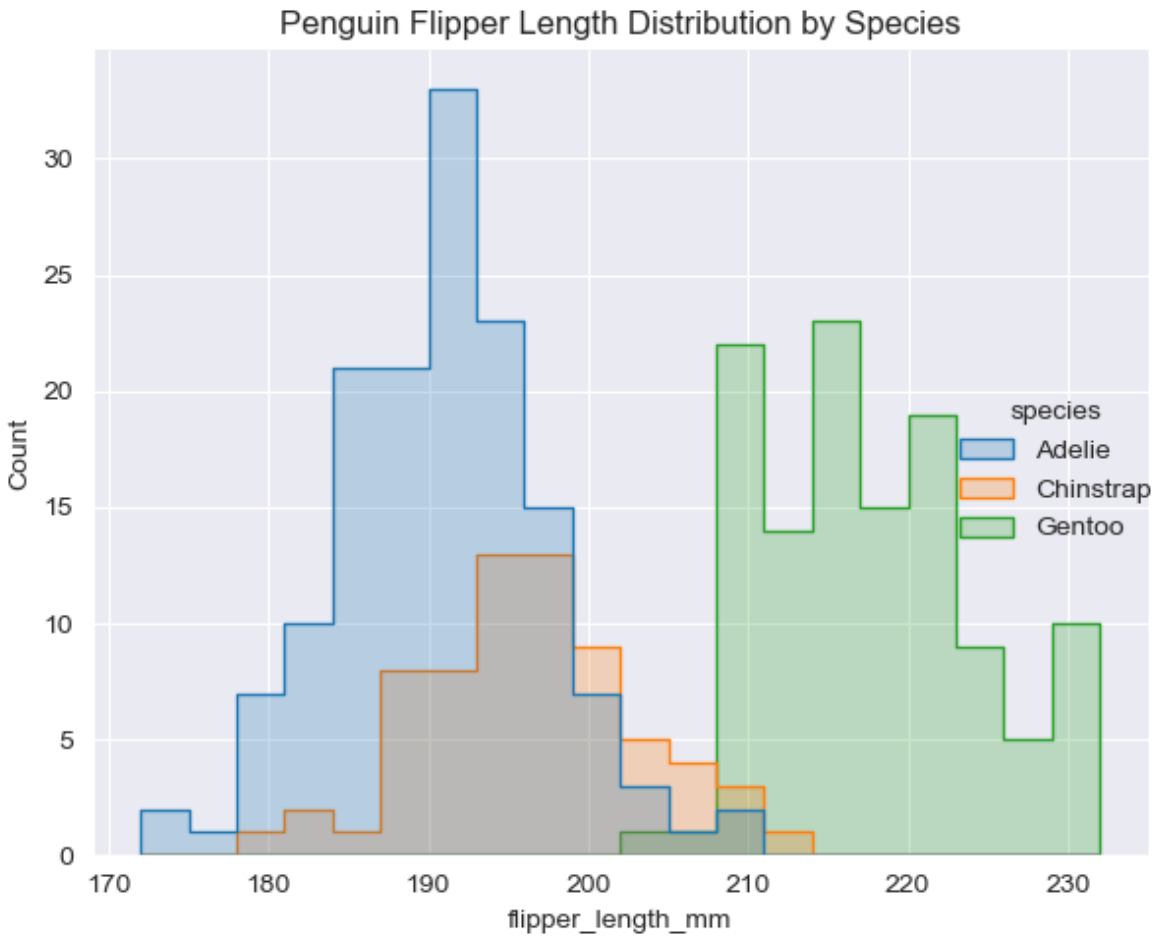
Re-graph the plot in the previous question with element='step'.

"""

Code

```
#%%
# Question 7
# Set the element is 'step'
Q7_displot = sns.displot(data=penguins,
                         x='flipper_length_mm',
                         binwidth=3,
                         bins=30,
                         hue='species',
                         element='step')

# Build the title for the displot
Q7_displot.set(title='Penguin Flipper Length Distribution with element')

# Display the plot
plt.tight_layout()
plt.show()

# Print the observation
print("Adelie penguins tend to have the shortest flipper lengths, mostly
clustering around 180-190 mm. \n "
      "Chinstrap penguins have flipper lengths that fall mainly between 190-200
mm. "
      "\nGentoo penguins have the longest flippers, with a distribution that is
centered around 210-220 mm.")
```

Answer




```
    ...:
    ...: # Build the title for the displot
    ...: Q7_displot.set(title='Penguin Flipper Length Distribution with element')
    ...:
    ...: # Display the plot
    ...: plt.tight_layout()
    ...: plt.show()
    ...:
    ...: # Print the observation
    ...: print("Adelie penguins tend to have the shortest flipper lengths, mostly clustering around 180-190 mm. \n "
    ...:       "Chinstrap penguins have flipper lengths that fall mainly between 190-200 mm. "
    ...:       "\nGentoo penguins have the longest flippers, with a distribution that is centered around 210-220 mm.")
    ...:
D:\Conda\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a futu
  with pd.option_context('mode.use_inf_as_na', True):
Adelie penguins tend to have the shortest flipper lengths, mostly clustering around 180-190 mm.
 Chinstrap penguins have flipper lengths that fall mainly between 190-200 mm.
Gentoo penguins have the longest flippers, with a distribution that is centered around 210-220 mm.

In [16]:
```

Question 8

"""

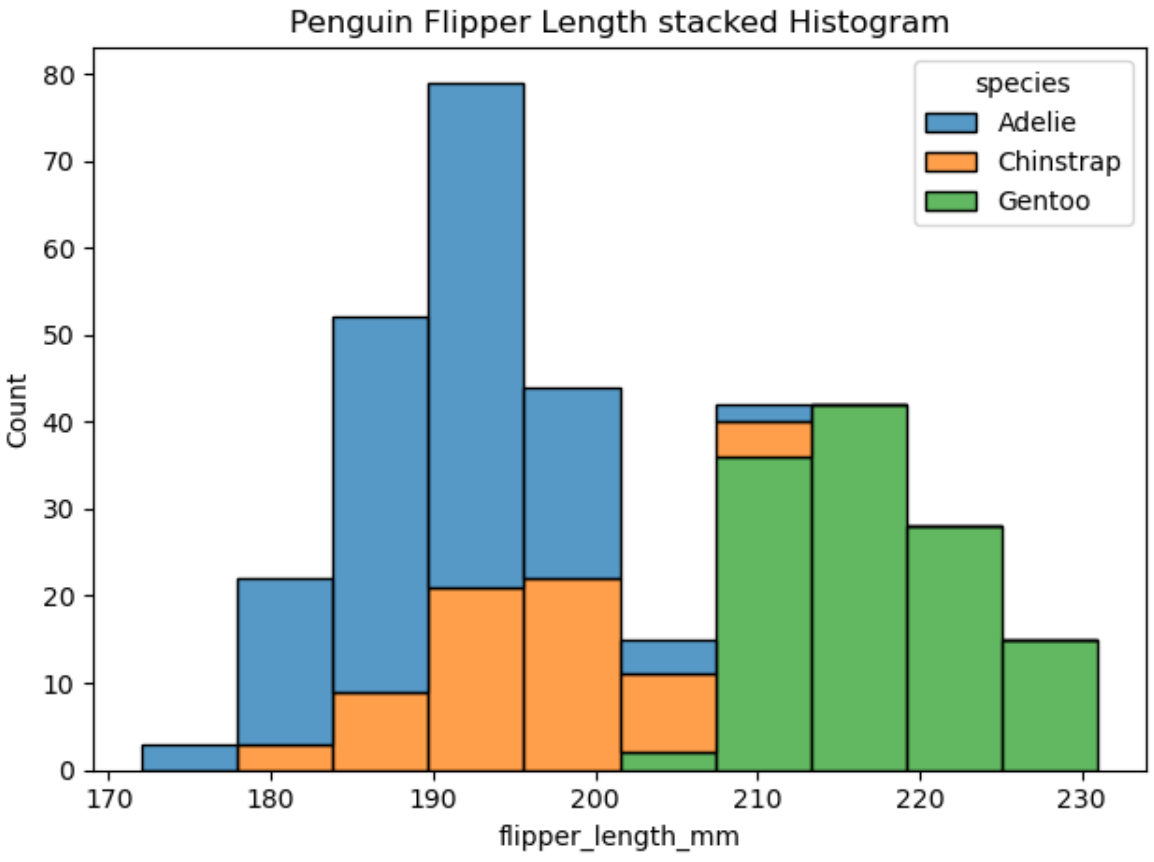
Using the seaborn package graph the 'stacked' histogram plot of 'flipper_length_mm' with respect to 'species'. Hint: multiple = 'stack'. Write down your observation about the graph on the console.

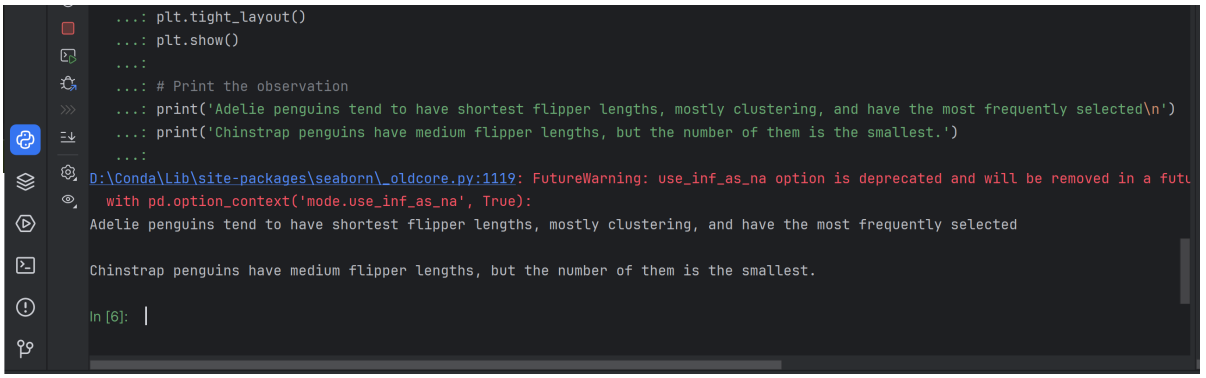
"""

Code

```
#%%
# Question 8
# Set the mutiple = stack
Q8_hist = sns.histplot(data=penguins,
                       x='flipper_length_mm',
                       hue='species',
                       multiple='stack')

# Set the title for Q8_hist
Q8_hist.set_title('Penguin Flipper Length stacked Histogram')

# Disply the Q8_hist
plt.tight_layout()
plt.show()

# Print the observation
print('Adelie penguins tend to have shortest flipper lengths, mostly clustering,
and have the most frequently selected\n')
print('Chinstrap penguins have medium flipper lengths, but the number of them is
the smallest.')
```

Answer





Question 9

"""

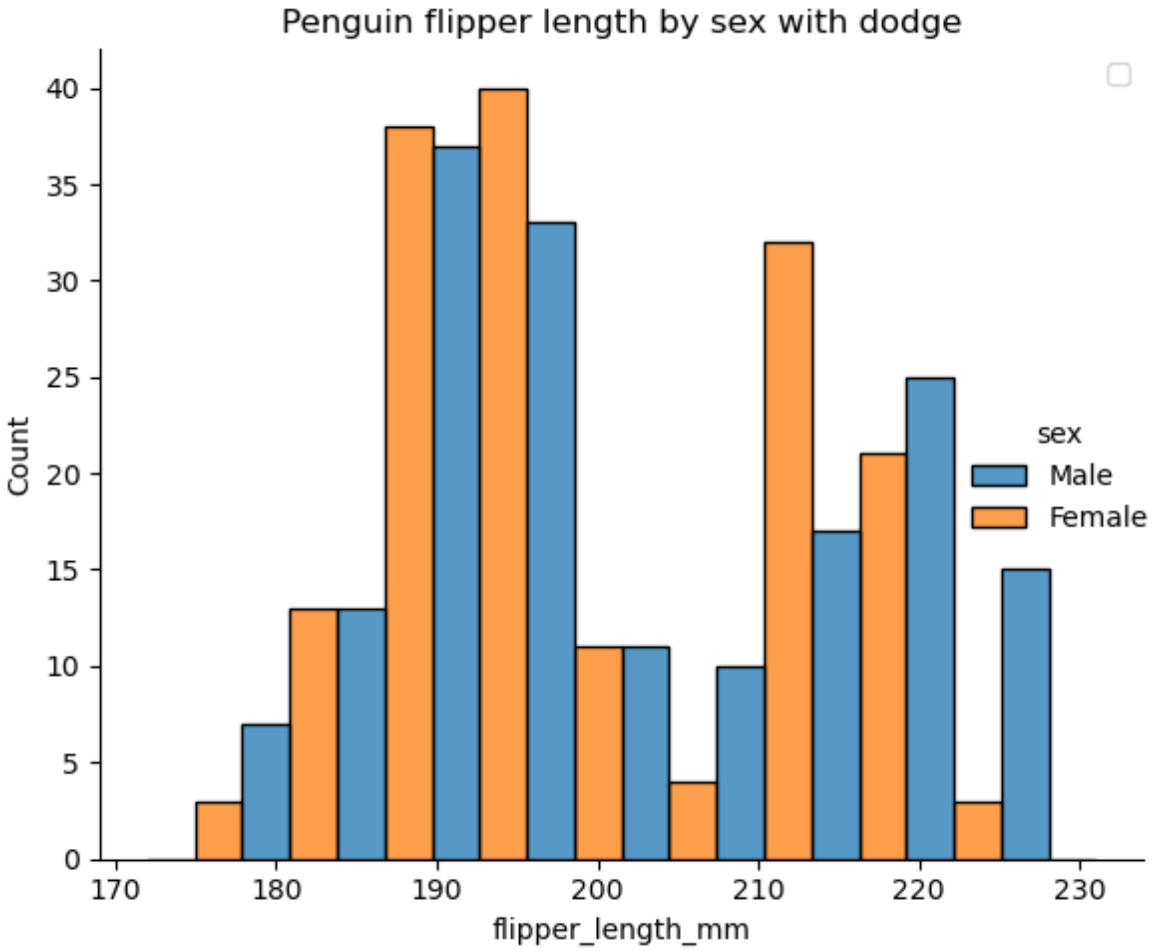
Using the seaborn package and 'displot', graph the histogram plot of 'flipper_lebgth_mm' with respect to 'sex' and use the option "dodge". Write down your observation about the graph on the console. Hint: multiple = 'dodge'.

"""

Code

```
#%%
# Question 9
Q9_hist = sns.histplot(data=penguins,
                       x='flipper_length_mm',
                       hue='sex',
                       multiple='dodge')
# Set the title
Q9_hist.set_title('Penguin flipper length by sex with dodge')

# Set the legend
Q9_hist.legend(loc='upper right', title='sex', labels=['Male', 'Female'])

plt.tight_layout()
plt.show()

# Print the observation
print("Female penguins tend to have longer flipper lengths than male penguins.")
```

Answer




```
    ...: print("Female penguins tend to have longer flipper lengths than male penguins.")
    ...:
D:\Conda\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be re
  with pd.option_context('mode.use_inf_as_na', True):
Female penguins tend to have longer flipper lengths than male penguins.

In [21]:
```

Question 10

"""

Using the seaborn package and 'displot', graph the histogram plot of 'flipper_lebgth_mm' in two separate figures (not shared axis) but in one single graph (one row two columns). What is the most frequent range of flipper length in mm for male and female penguins?

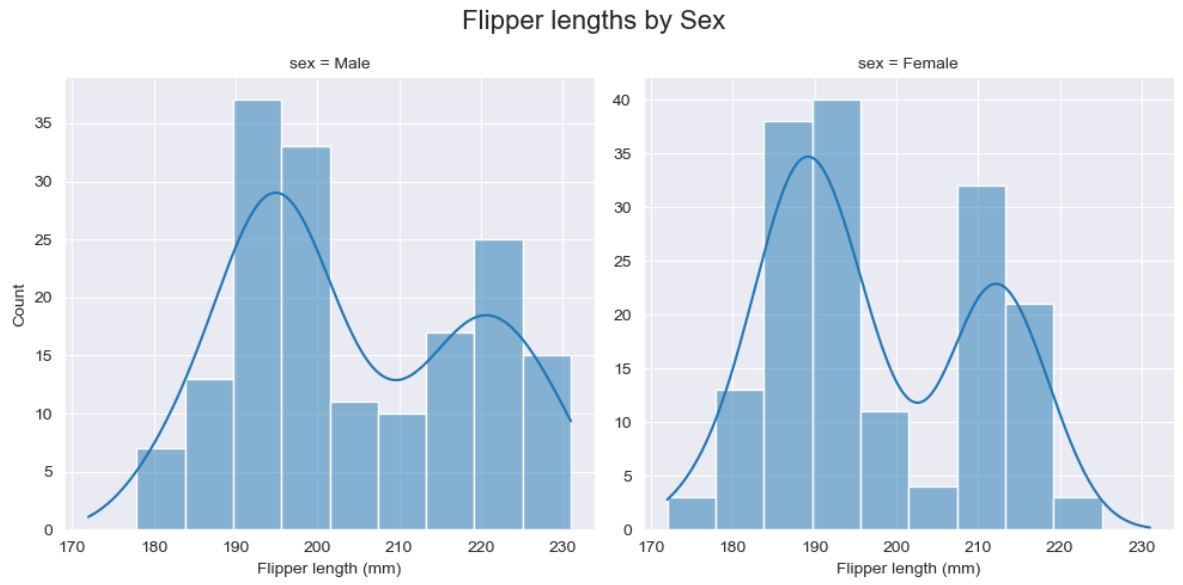
"""

Code

```
Q10_fig = sns.displot(
    data=penguins,
    x='flipper_length_mm',
    col='sex',
    kde=True,
    facet_kws={'sharey':False, 'sharex':False}
)
```

```
# Set the title and labels
Q10_fig.fig.suptitle('Flipper lengths by Sex', fontsize=16)
Q10_fig.set_axis_labels('Flipper length (mm)', 'Count')
Q10_fig.fig.subplots_adjust(top=0.8)

plt.tight_layout()
plt.show()

print("Most male penguins tend to have 190-200mm flipper lengths")
print("Most Female penguins tend to have 185-195mm flipper lengths")
```

Answer




```
    ....
    ...: print("Most male penguins tend to have 190-200mm flipper lengths")
    ...: print("Most Female penguins tend to have 185-195mm flipper lengths")
    ...:
Most male penguins tend to have 190-200mm flipper lengths
Most Female penguins tend to have 185-195mm flipper lengths

In [22]:
```

Question 11

"""

Using the seaborn package compare the distribution of ‘flipper_length_mm’ with respect to species in one graph (shared axis) in a normalized fashion. Which species has the larger flipper length and what is the approximate range? Hint: Use stat = ‘density’

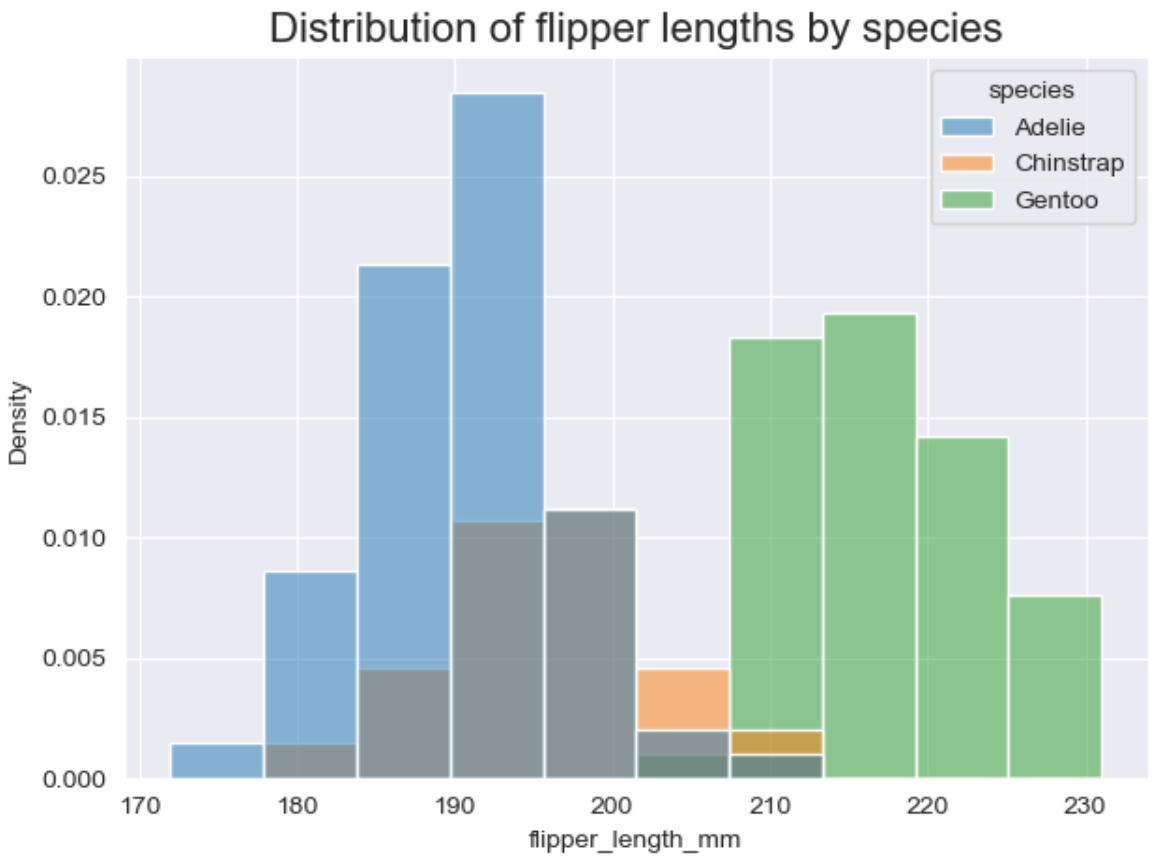
"""

Code

```
#%%
# Question 11
Q11_fig = sns.histplot(data=penguins,
                       x='flipper_length_mm',
                       stat='density',
```

```
                        hue='species',
                        legend=True,
                        kde=False)

# Set the title and labels
Q11_fig.set_title('Distribution of flipper lengths by species', fontsize=16)


# Show the plot
plt.tight_layout()
plt.show()

print("Adelie penguins tend to have 190-195 mm flipper lengths")
print('Chinstrap penguins tend to have 195-200mm flipper lengths')
print('Gentoo penguins tend to have 215-220mm flipper lengths')
```

Answer




```
    ...:
    ...: print("Adelie penguins tend to have 190-195 mm flipper lengths")
    ...: print('Chinstrap penguins tend to have 195-200mm flipper lengths')
    ...: print('Gentoo penguins tend to have 215-220mm flipper lengths')
    ...:
Adelie penguins tend to have 190-195 mm flipper lengths
Chinstrap penguins tend to have 195-200mm flipper lengths
Gentoo penguins tend to have 215-220mm flipper lengths

In [30]:
```

Question 12

"""

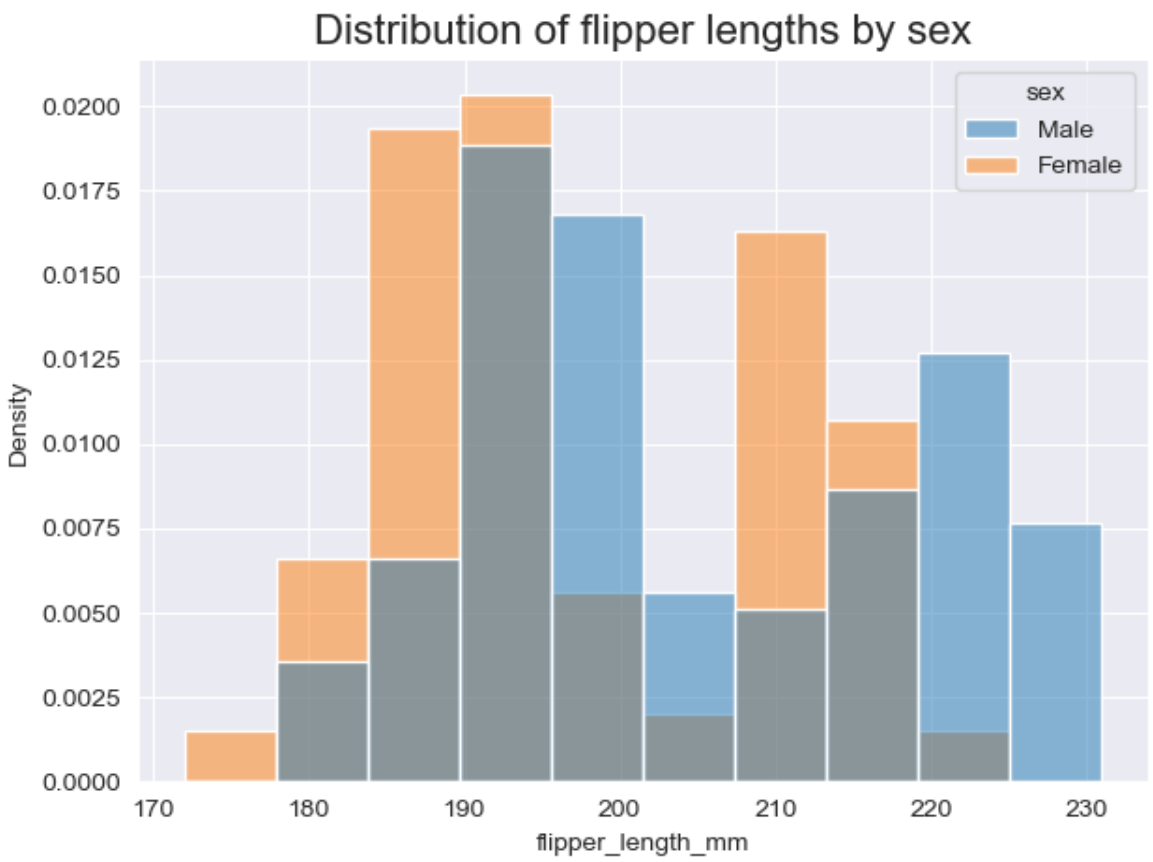
Using the seaborn package compare the distribution of ‘flipper_length_mm’ with respect to sex in one graph (shared axis) in a normalized fashion. Which sex has the larger flipper length and what is the approximate flipper length? Hint: Use stat = ‘density’

"""

Code

```
#%%
# Question 12
Q12_fig = sns.histplot(data=penguins,
                       x='flipper_length_mm',
                       stat='density',
                       hue='sex',
                       legend=True,
                       kde=False)

# Set the title and labels
Q12_fig.set_title('Distribution of flipper lengths by sex', fontsize=16)


# Show the plot
plt.tight_layout()
plt.show()

print("Male penguins tend to have 190-195 mm flipper lengths")
print('Female penguins tend to have 190-195mm flipper lengths')
```

Answer




```
    ...: print('Female penguins tend to have 190-195mm flipper lengths')
    ...:
Male penguins tend to have 190-195 mm flipper lengths
Female penguins tend to have 190-195mm flipper lengths

In [32]:
```

Question 13

"""

Using the seaborn package compare the distribution of 'flipper_length_mm' with respect to species in one graph (shared axis) in a normalized fashion that the bars height sum to 1. Which flipper length and species is more probable? Hint: Use stat = 'probability'

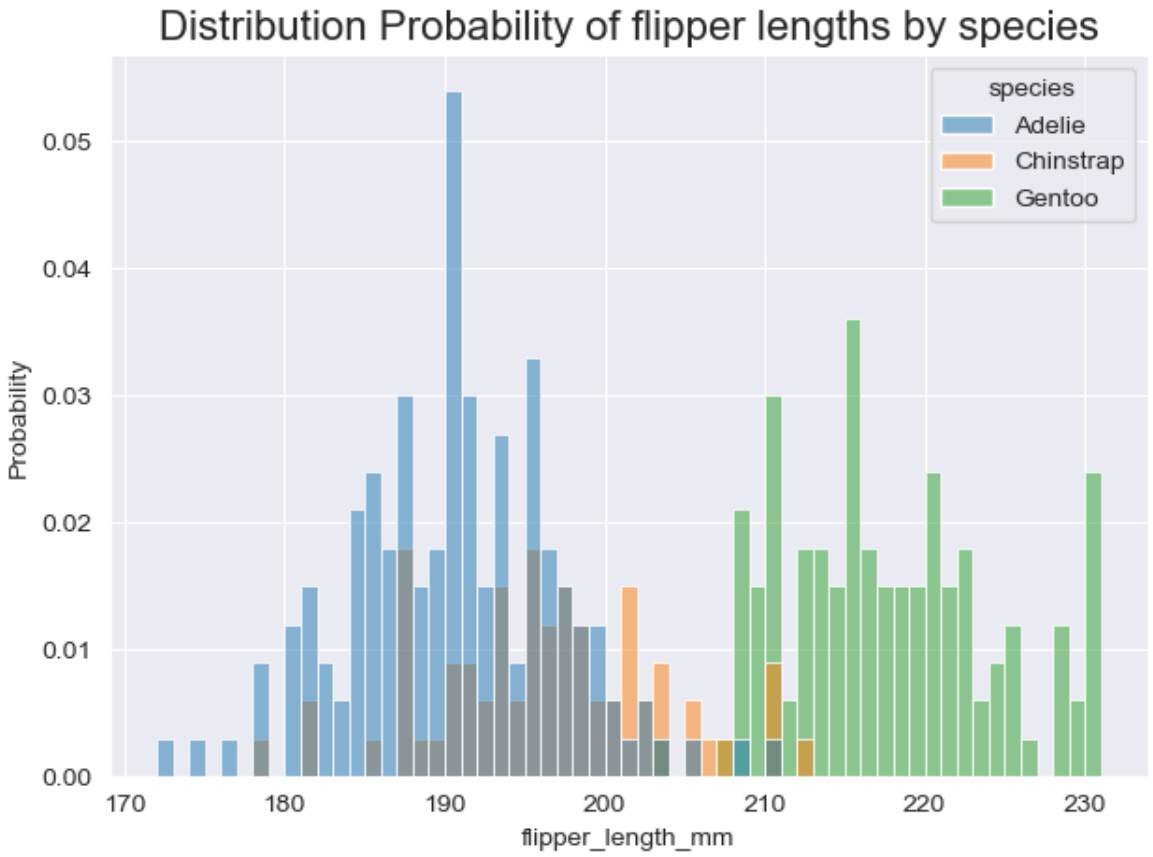
"""

Code

```
#%%
# Question 13
Q13_fig = sns.histplot(data=penguins,
                       x='flipper_length_mm',
                       stat='probability',
                       binwidth=True,
                       hue='species',
                       legend=True,
```

```
                        kde=False)

# Set the title and labels
Q13_fig.set_title('Distribution Probability of flipper lengths by species',
fontsize=16)


# Show the plot
plt.tight_layout()
plt.show()

print("Adelie Male penguins who has 191 mm flipper lengths is more probable")
```

Answer




```
    ...: plt.show()
    ...:
    ...: print("Adelie Male penguins who has 191 mm flipper lengths is more probable")
    ...:
Adelie Male penguins who has 191 mm flipper lengths is more probable

In [35]:
```

Question 14

"""

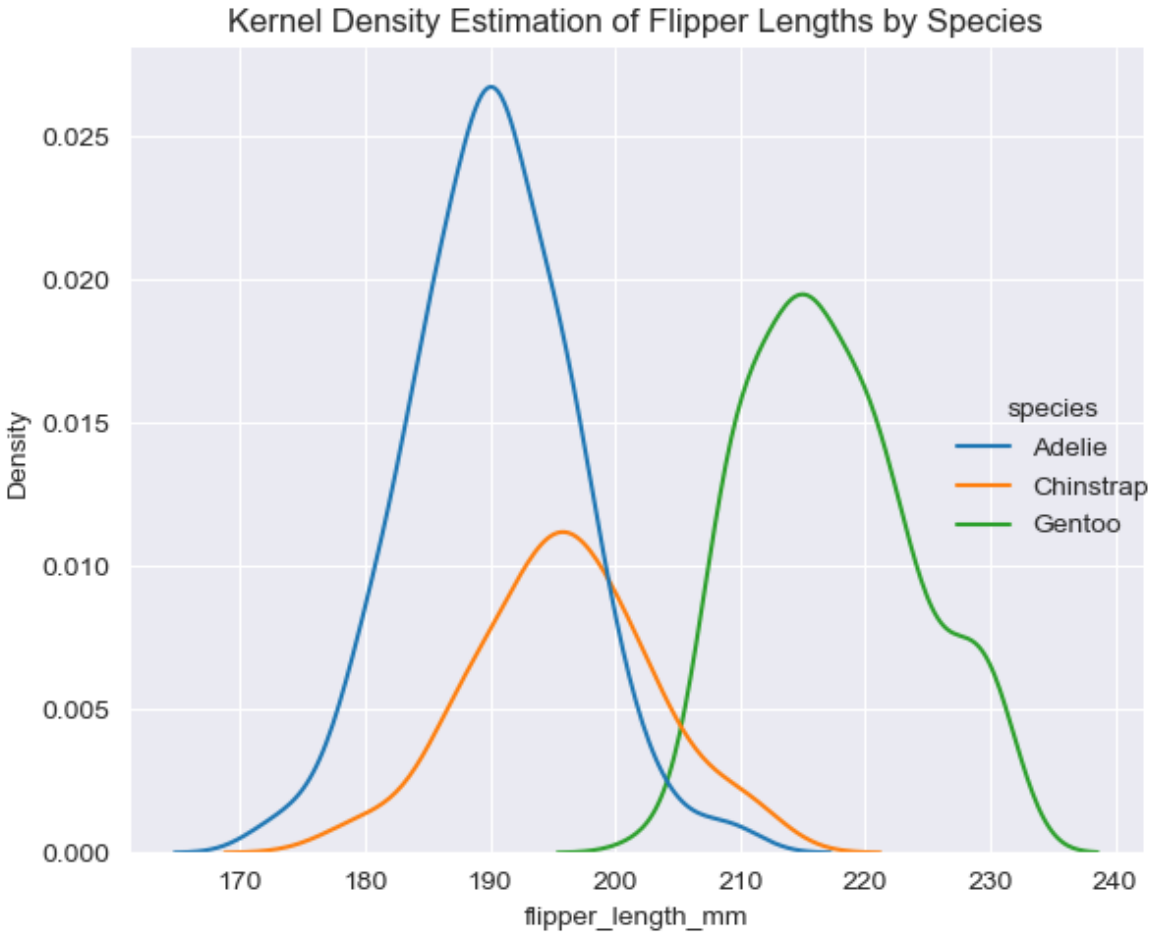
Using the seaborn package estimate the underlying density function of flipper length with respect to ‘species’ and the kernel density estimation. Plot the result. Hint: hue = ‘species’, kind = ‘kde’

"""

Code

```
#%%
# Question 14
Q14_fig = sns.displot(
    data=penguins,
    hue='species',
    kind='kde',
    x='flipper_length_mm',
    legend=True,
    fill=True
)

# Set the title and labels
Q14_fig.set(title='Kernel Density Estimation of Flipper Lengths by Species')
plt.tight_layout()
plt.show()
```

Answer



Question 15

"""

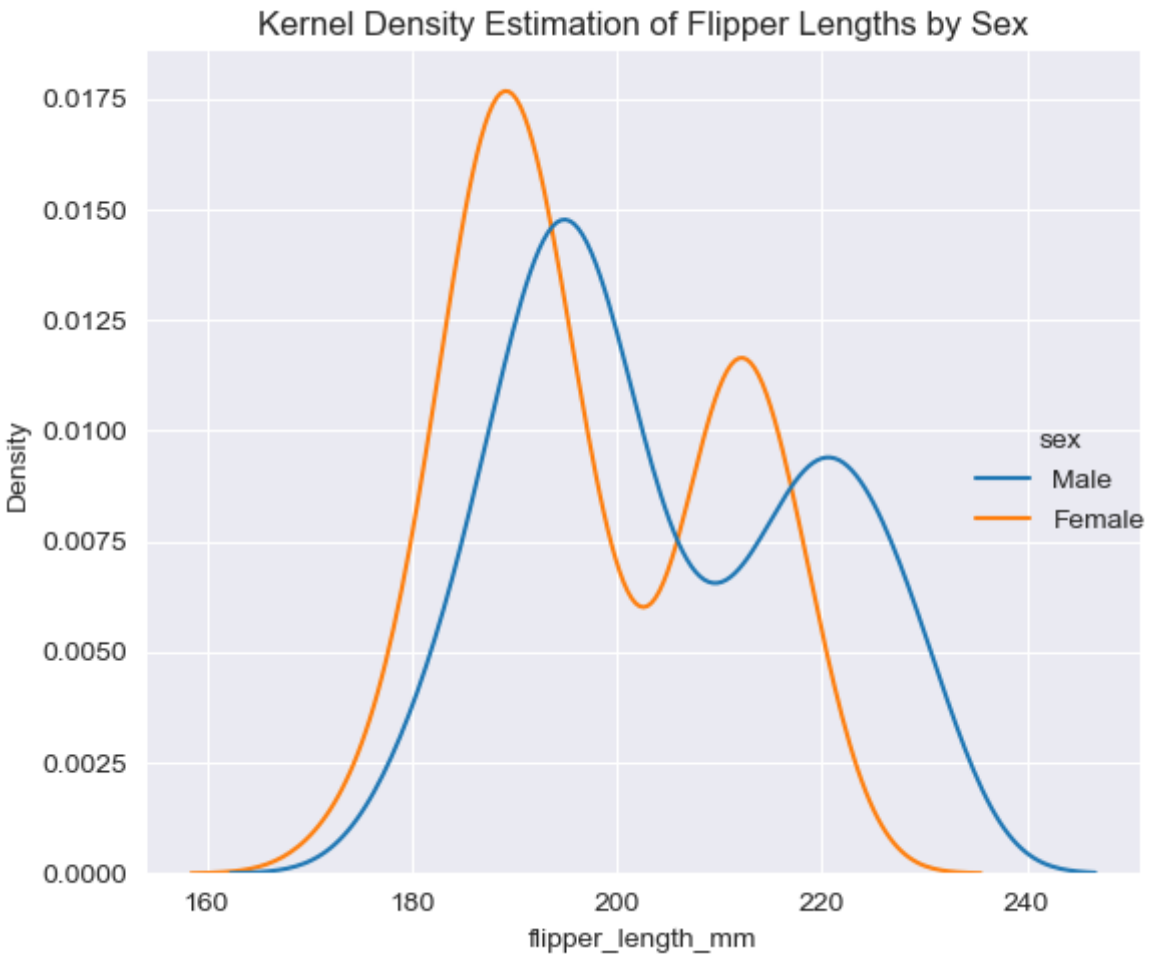
Using the seaborn package estimate the underlying density function of flipper length with respect to ‘sex’ and the kernel density estimation. Plot the result. Hint: hue = ‘sex’, kind = ‘kde’

"""

Code

```
#%%
# Question 15
Q15_fig = sns.displot(
    data=penguins,
    hue='sex',
    kind='kde',
    x='flipper_length_mm',
    legend=True,
    fill=True
)

# Set the title and labels
Q15_fig.set(title='Kernel Density Estimation of Flipper Lengths by Sex')

plt.tight_layout()
plt.show()
```

Answer



Question 16

"""

Repeat question 14 with argument multiple = ‘stack’

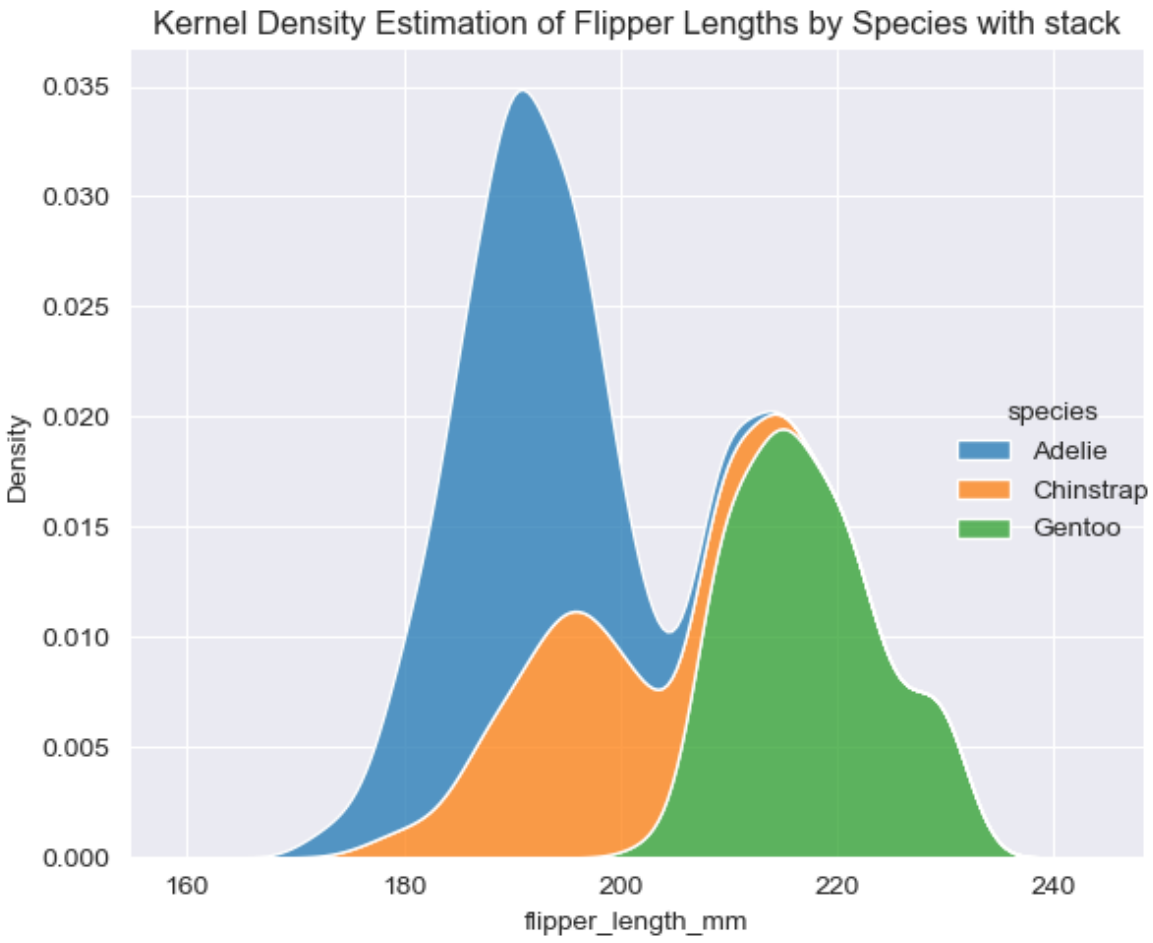
"""

Code

```
#%%
# Question 16
Q16_fig = sns.displot(
    data=penguins,
    hue='species',
    kind='kde',
    x='flipper_length_mm',
    legend=True,
    fill=True,
    multiple='stack'
)

# Set the title and labels
Q16_fig.set(title='Kernel Density Estimation of Flipper Lengths by Species with
stack')
plt.tight_layout()
```

```
plt.show()
```

Answer



Question 17

"""

Repeat question 15 with argument multiple = ‘stack’

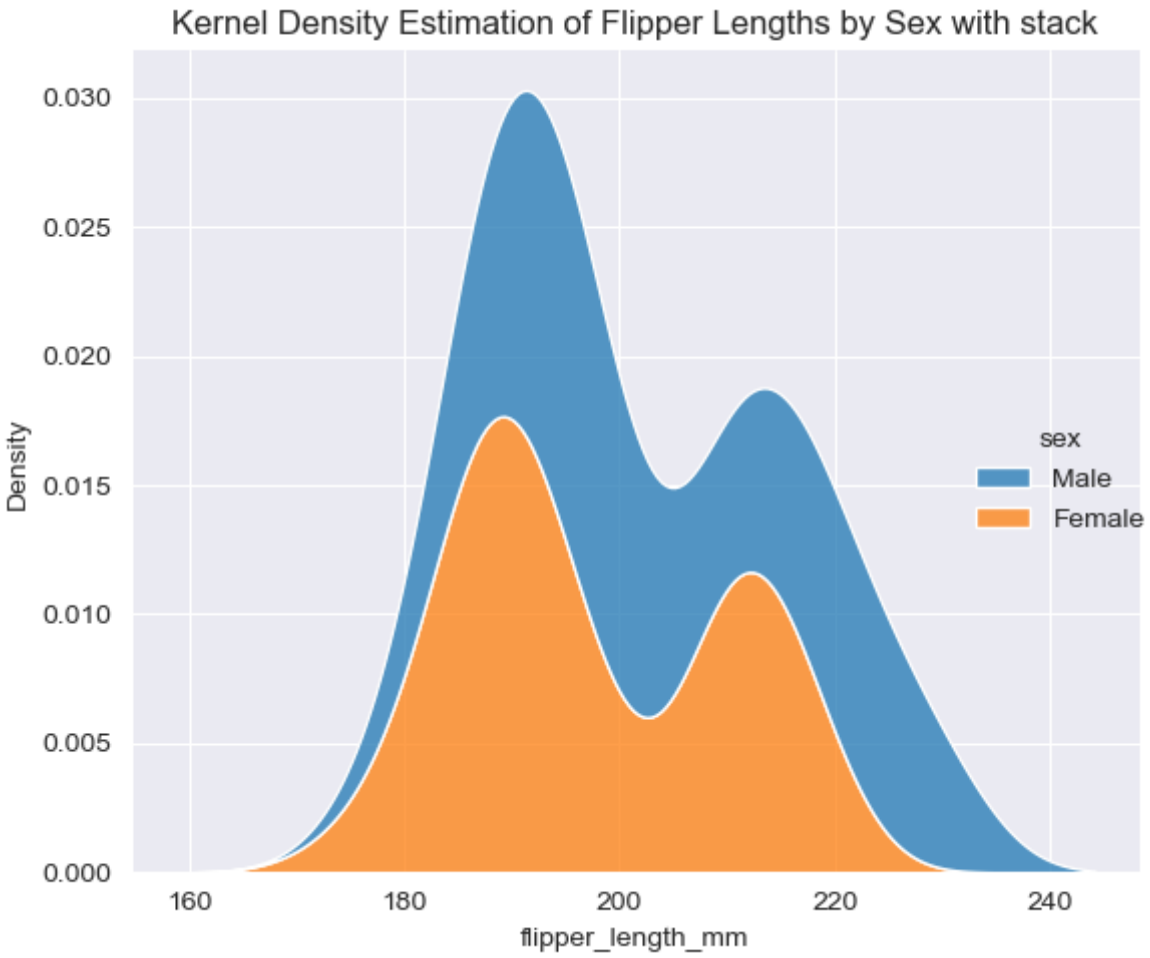
"""

Code

```
#%%
# Question 17
Q17_fig = sns.displot(
    data=penguins,
    hue='sex',
    kind='kde',
    x='flipper_length_mm',
    legend=True,
    fill=True,
    multiple='stack'
)
```

```
# Set the title and labels
Q17_fig.set(title='Kernel Density Estimation of Flipper Lengths by Sex with
stack')

plt.tight_layout()
plt.show()
```

Answer



Question 18

"""

Repeat question 14 with argument fill = True. Write down your observations about the graph.

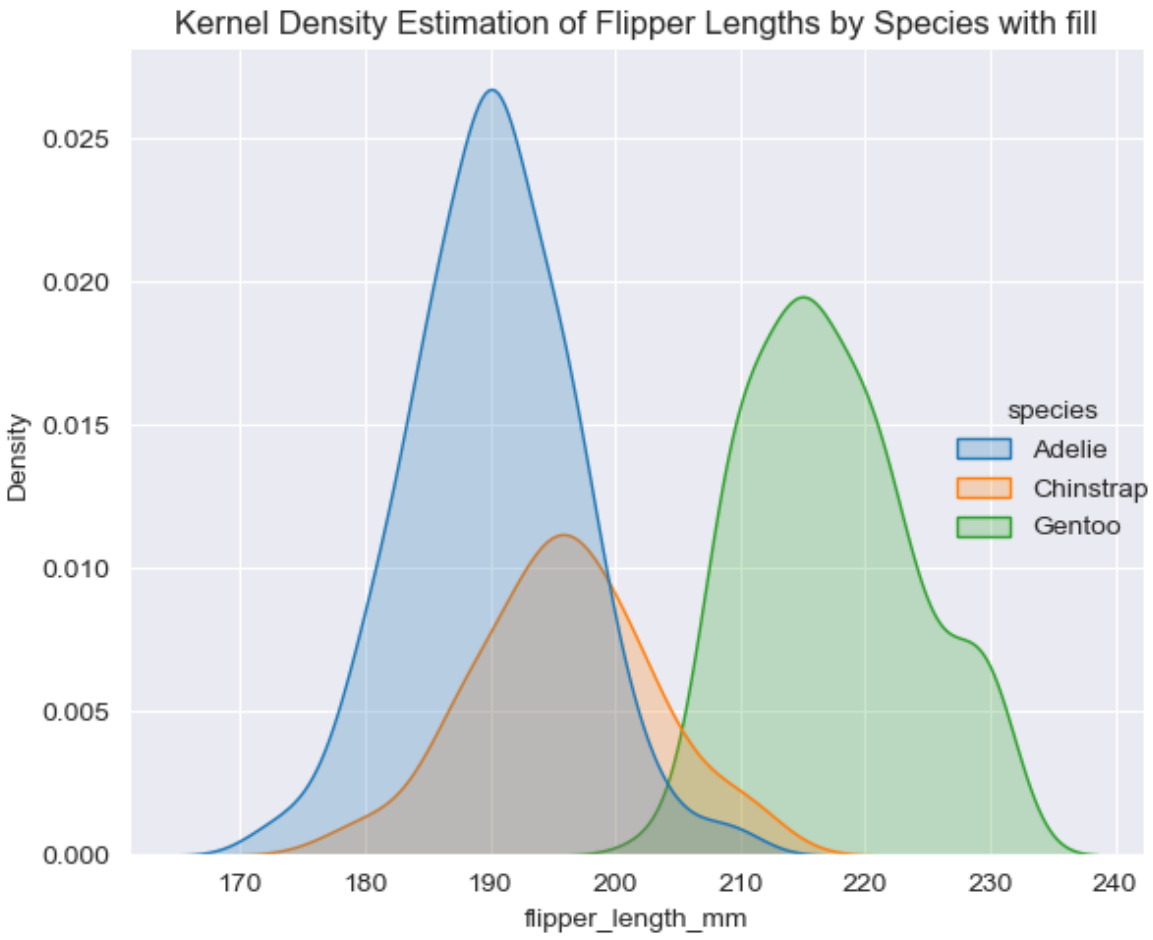
"""

Code

```
#%%
# Question 18
Q18_fig = sns.displot(
    data=penguins,
```

```
    hue='species',
    kind='kde',
    x='flipper_length_mm',
    legend=True,
    fill=True
)

# Set the title and labels
Q18_fig.set(title='Kernel Density Estimation of Flipper Lengths by Species with fill')
plt.tight_layout()
plt.show()

# Print the observation
print("The peak of Adelie penguins flipper length is 190mm, the density is 0.025")
print('The peak of Chinstrap penguins flipper length is 196mm and the density is about 0.012')
print("The peak of Gentoo penguins flipper length is 215mm, the density is about 0.018")
```

Answer



```
    ...: print('The peak of Chinstrap penguins flipper length is 196mm and the density is about 0.01
    ...: print("The peak of Gentoo penguins flipper length is 215mm, the density is about 0.018")
    ...:
The peak of Adelie penguins flipper length is 190mm, the density is 0.025
The peak of Chinstrap penguins flipper length is 196mm and the density is about 0.012
The peak of Gentoo penguins flipper length is 215mm, the density is about 0.018
```

Question 19

"""

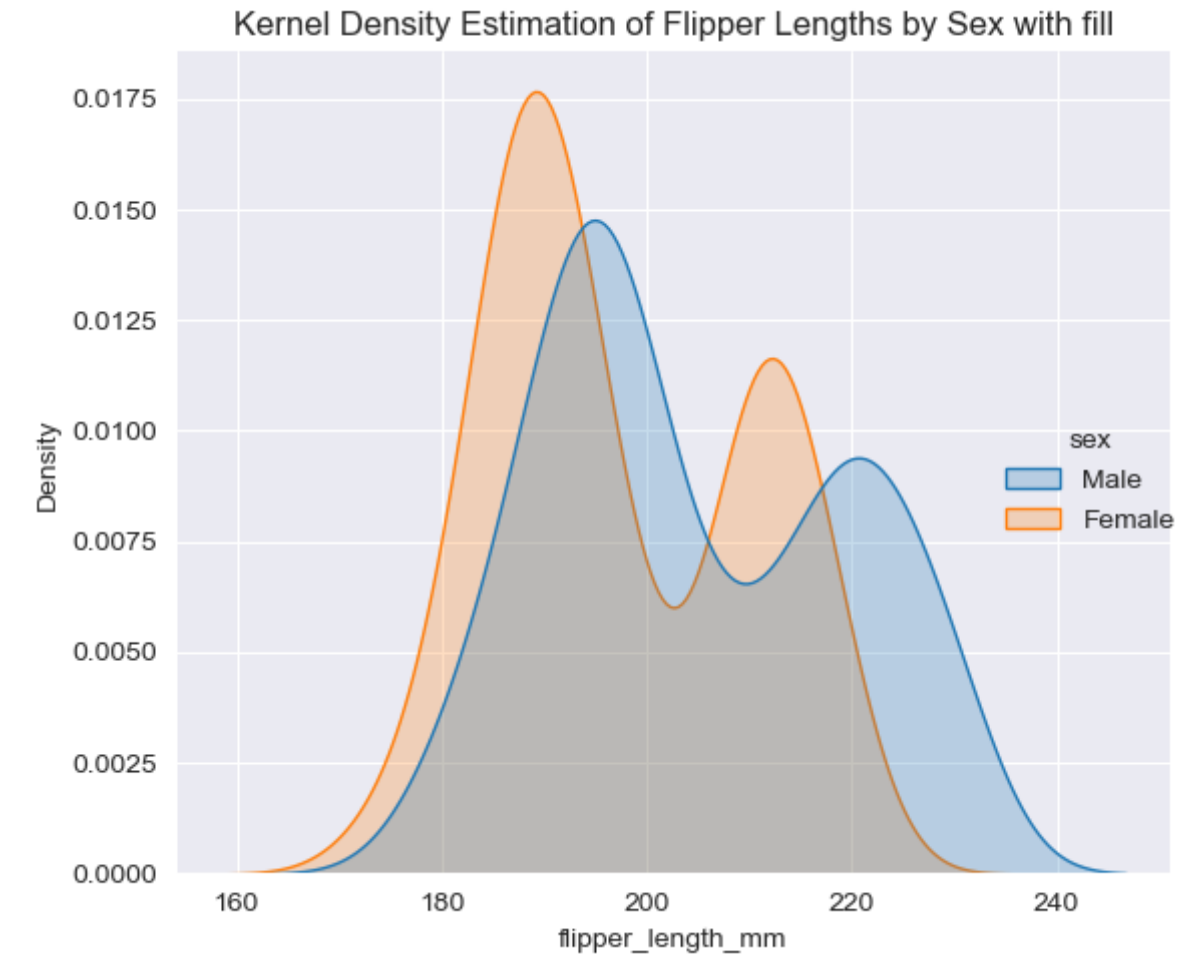
Repeat question 15 with argument fill = True. Write down your observations about the graph.

"""

Code

```
#%%
# Question 19
Q19_fig = sns.displot(
    data=penguins,
    hue='sex',
    kind='kde',
    x='flipper_length_mm',
    legend=True,
    fill=True
)

# Set the title and labels
Q19_fig.set(title='Kernel Density Estimation of Flipper Lengths by Sex with
fill')

plt.tight_layout()
plt.show()

# Print the observation
print("The peak of female penguins flipper length is about 190mm, the density is
0.0176")
print("The peak of male penguins flipper length is about 195mm, the density is
0.0148")
```

Answer




```
...: print("The peak of female penguins flipper length is about 190mm, the density is 0.0176")
...: print("The peak of male penguins flipper length is about 195mm, the density is 0.0148")
...:
The peak of female penguins flipper length is about 190mm, the density is 0.0176
The peak of male penguins flipper length is about 195mm, the density is 0.0148

In [34]:
```

Question 20

"""

Plot the scatter plot and the regression line in one graph for the x-axis is 'bill_length_mm' and y-axis is 'bill_depth_mm'. How the 'bill_length_mm' and 'bill_depth_mm' are correlated?

"""

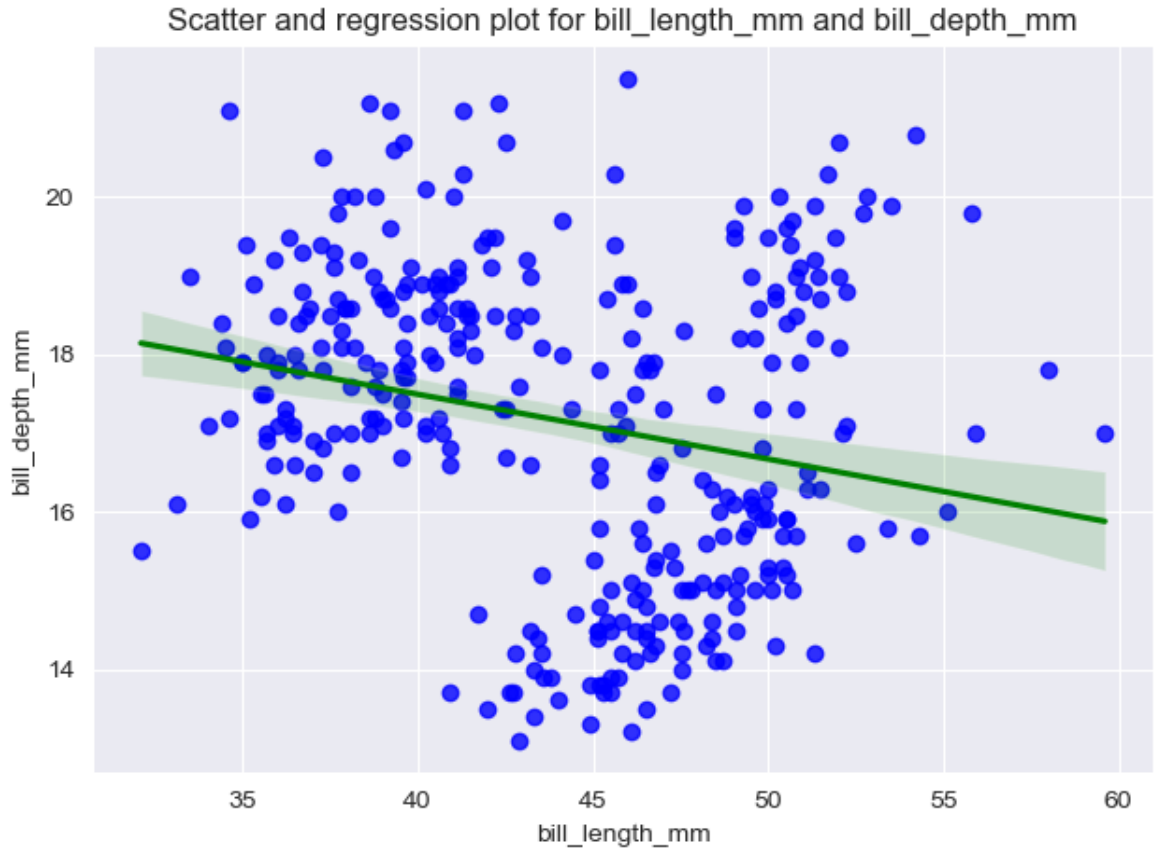
Code

```
#%%
# Question 20
# Create a scatter and regression plot for bill_length and bill_depth
Q20_fig = sns.regplot(
    data=penguins,
    x='bill_length_mm',
    y='bill_depth_mm',
    scatter_kws={'color':'blue'},
```

```
    line_kws={'color':'green'}
)

# Set title and xlabel, ylabel
Q20_fig.set(title='Scatter and regression plot for bill_length_mm and
bill_depth_mm',
            xlabel='bill_length_mm',
            ylabel='bill_depth_mm')


# Show the plot
plt.tight_layout()
plt.show()

correlation = penguins['bill_length_mm'].corr(penguins['bill_depth_mm'])
print(f'The correlation between bill_length_mm and bill_depth_mm is
{correlation:.2f}')
print('bill_length_mm and bill_depth_mm is negative correlation')
```

Answer

```
    ...: print(f'The correlation between bill_length_mm and bill_depth_mm is {corre
    ...: print('bill_length_mm and bill_depth_mm is negative correlation')
    ...:
The correlation between bill_length_mm and bill_depth_mm is -0.23
bill_length_mm and bill_depth_mm is negative correlation

In [42]:
```



Question 21

"""

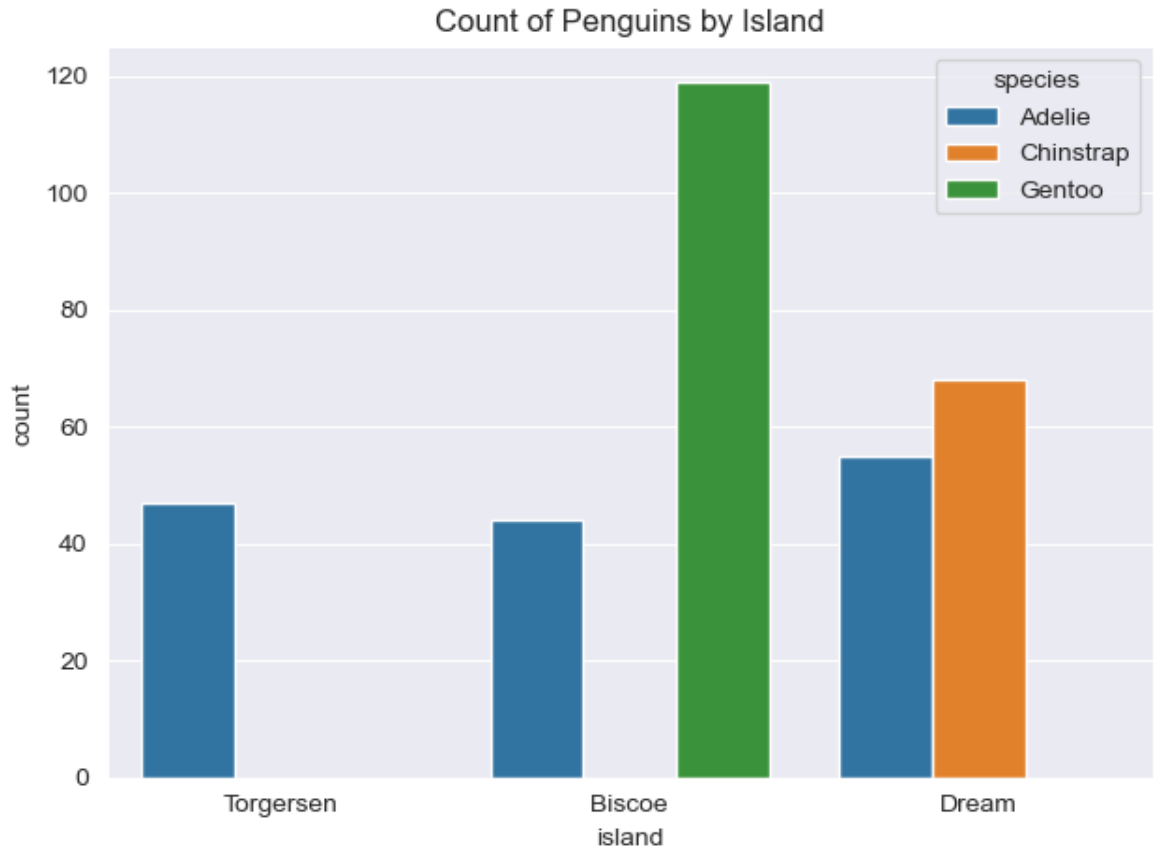
Using the count plot, display the bar plot of the number penguins in different islands using the hue = species. Write down your observations about the graph?

"""

Code

```
#%%
# Question 21
Q21_fig = sns.countplot(
    data=penguins,
    x='island',
    hue='species'
)

# Set the title and labels
Q21_fig.set(title='Count of Penguins by Island', xlabel='island', ylabel='count')

# Show the plot
plt.tight_layout()
plt.show()

# print the observation
print("Chinstrap penguins only live in the Dream island,\n Gentoo penguins only
live in the Biscoe island,\n and Adelie penguins live in the all three islands")
```

Answer




```
    ...: print("Chinstrap penguins only live in the Dream island,\n
    ...:
Chinstrap penguins only live in the Dream island,
 Gentoo penguins only live in the Biscoe island,
 and Adelie penguins live in the all three islands

In [51]:
```

Question 22

"""

Using the count plot, display the bar plot of the number of male and female penguins [in the dataset] using the hue = species. Write down your observations about the graph?

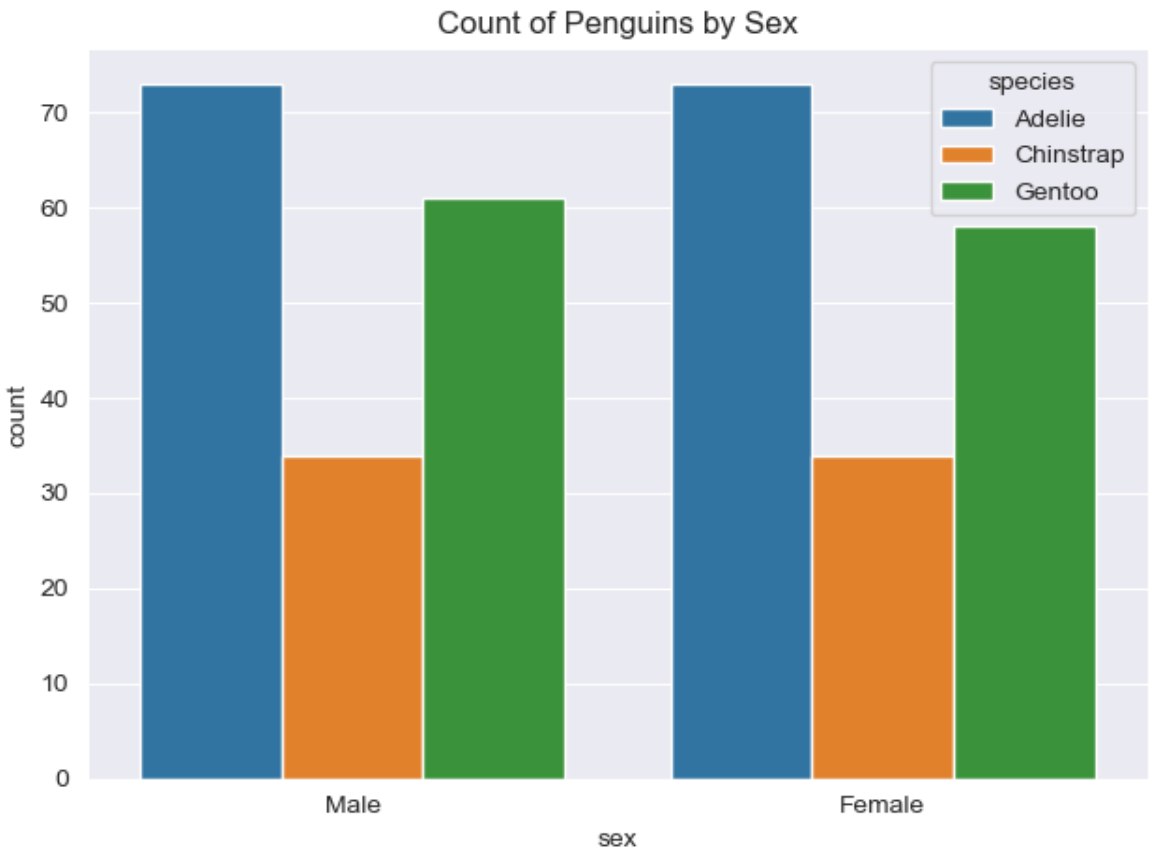
"""

Code

```
#%%
# Question 22
Q22_fig = sns.countplot(
    data=penguins,
    x='sex',
    hue='species'
)
```

```
# Set the title and labels
Q22_fig.set(title='Count of Penguins by Sex', xlabel='sex', ylabel='count')

# Show the plot
plt.tight_layout()
plt.show()

# print the observation
print("Each category penguins have same proportion male and female")
```

Answer




```
    ....
    ...: # print the observation
    ...: print("Each category penguins have same proportion male and female")
    ...:
Each category penguins have same proportion male and female

In [54]:
```

Question 23

"""

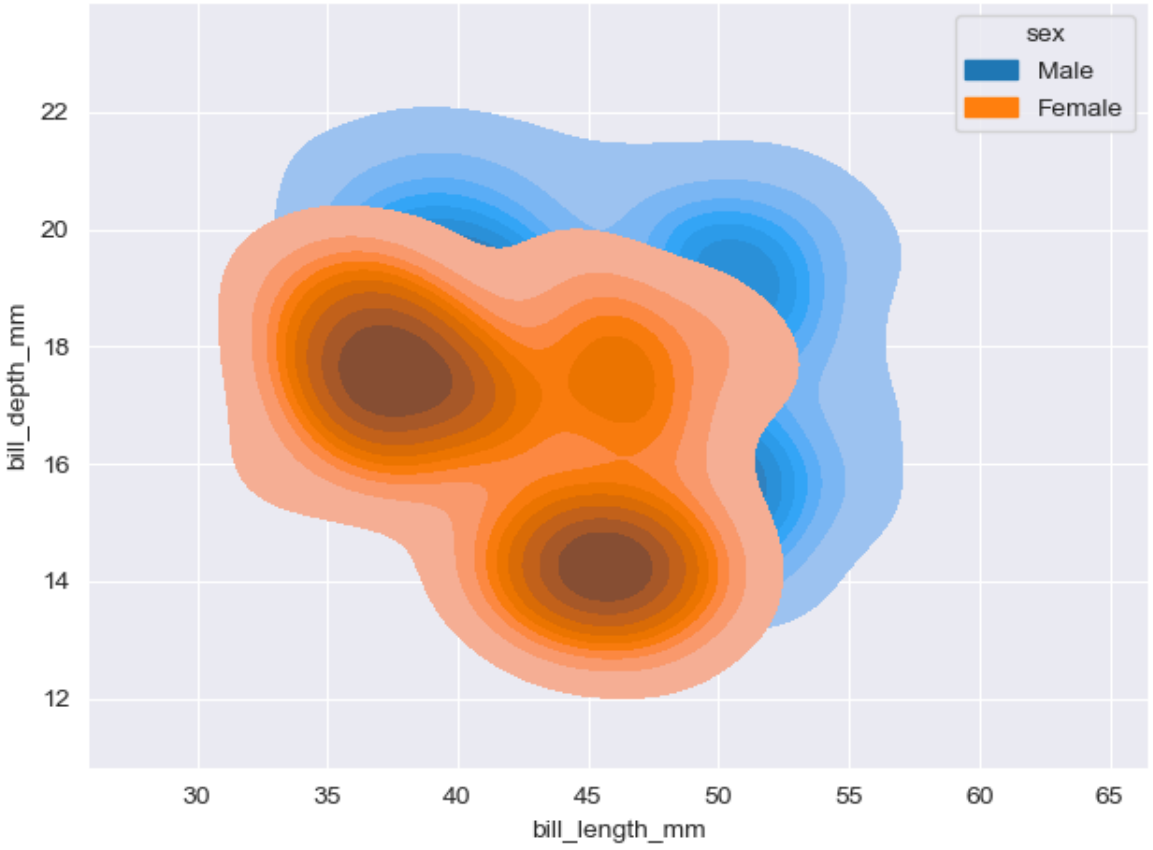
Plot the bivariate distribution between 'bill_length_mm' versus 'bill_depth_mm' for male and female.

"""

Code

```
#%%
# Question 23
Q23_fig = sns.kdeplot(
    data=penguins,
    x='bill_length_mm',
    y='bill_depth_mm',
    hue='sex',
    fill=True,
    grid=False
)

plt.tight_layout()
plt.show()
```

Answer



Question 24

"""

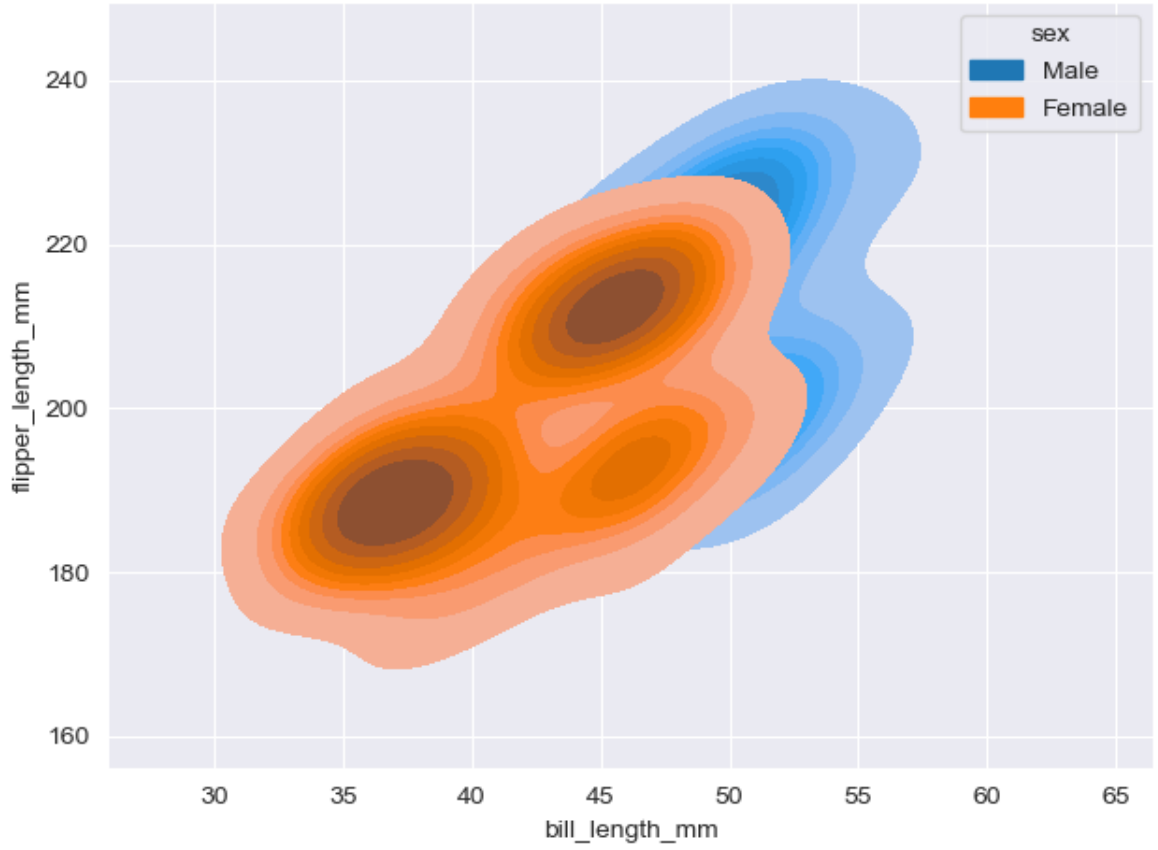
Plot the bivariate distribution between 'bill_length_mm' versus 'flipper_length_mm' for male and female. Final plot like question 23.

"""

Code

```
#%%
# Question 24
Q24_fig = sns.kdeplot(
    data=penguins,
    x='bill_length_mm',
    y='flipper_length_mm',
    hue='sex',
    fill=True,
    grid=False
)

plt.tight_layout()
plt.show()
```

Answer



Question 25

"""

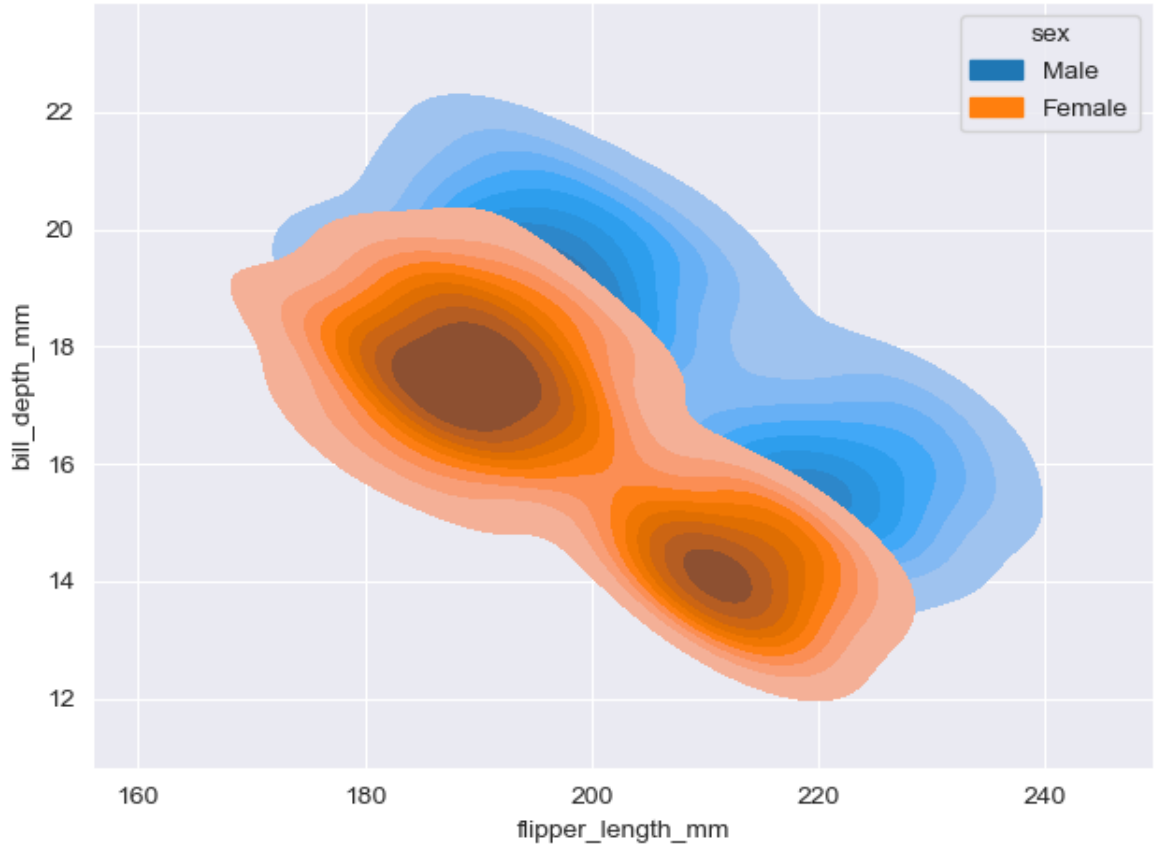
Plot the bivariate distribution between 'flipper_length_mm' versus 'bill_depth_mm' for male and female. Final plot like question 23.

"""

Code

```
#%%
# Question 25
Q25_fig = sns.kdeplot(
    data=penguins,
    x='flipper_length_mm',
    y='bill_depth_mm',
    hue='sex',
    fill=True,
    grid=False
)

plt.tight_layout()
plt.show()
```

Answer



Question 26

"""

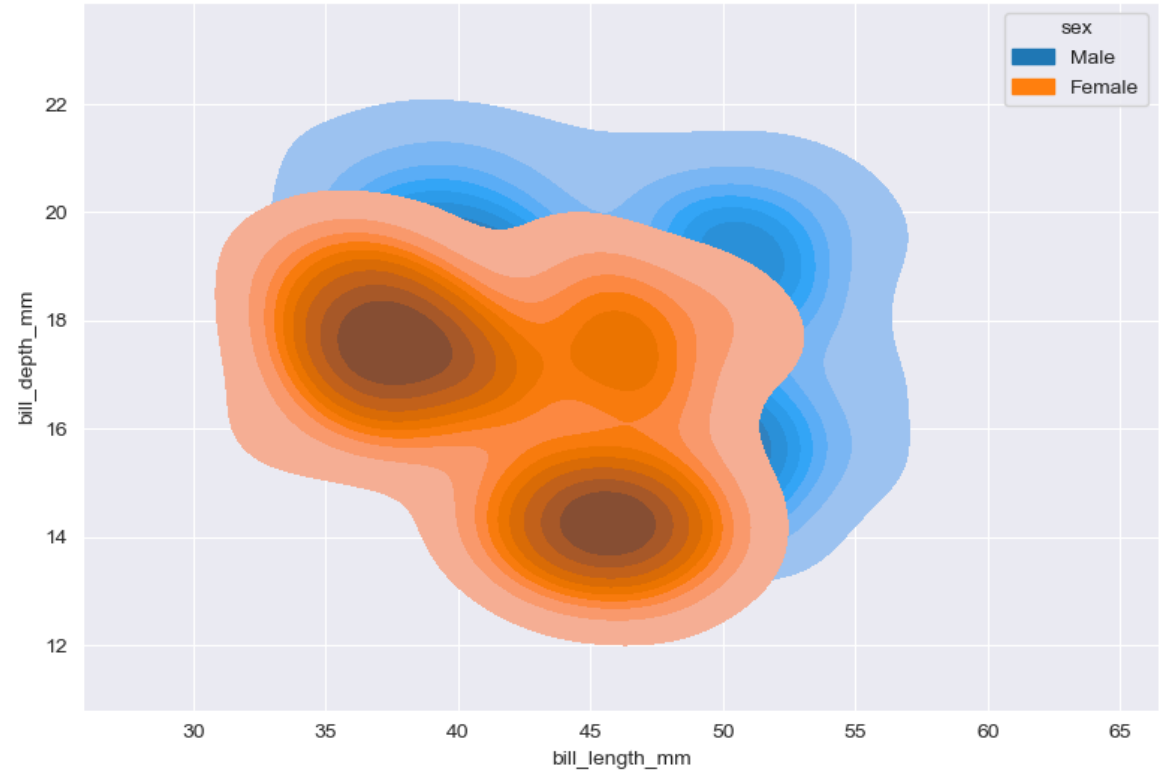
Using subplot, plot the last 3 questions in one graph with 3 rows and 1 column. Figure size = (8,16). Write down your observations about the plot in the last 3 questions.

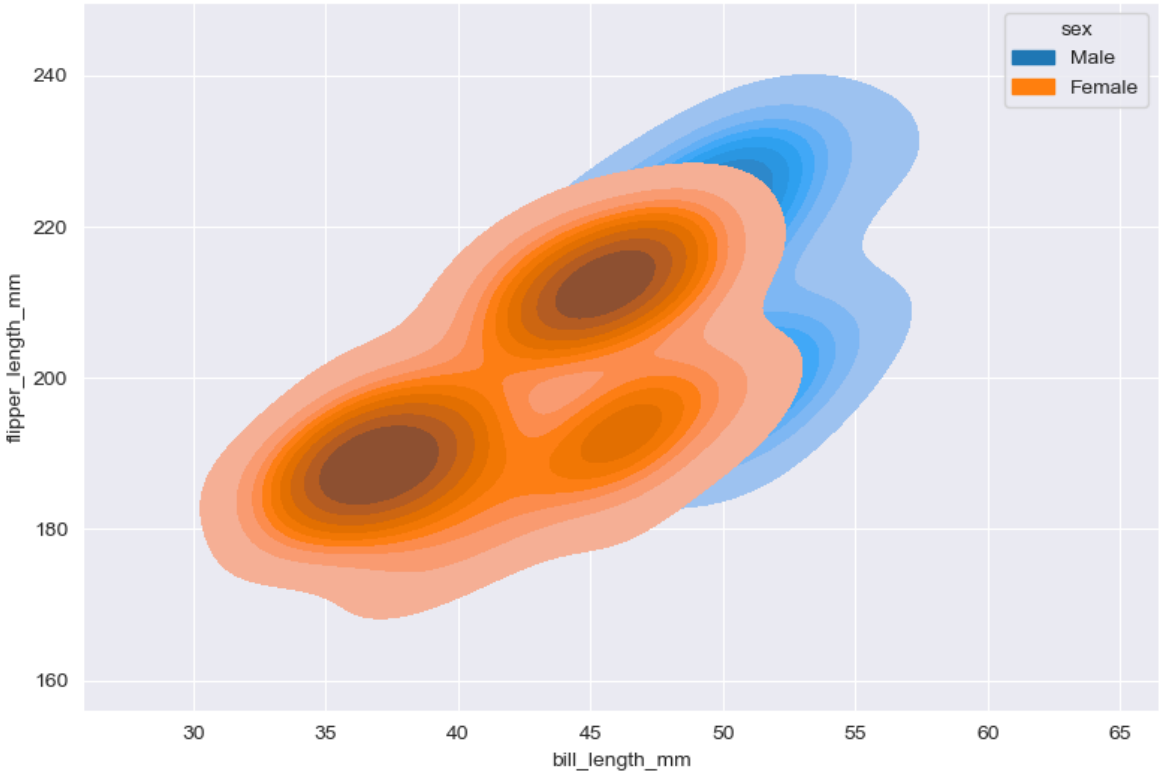
"""

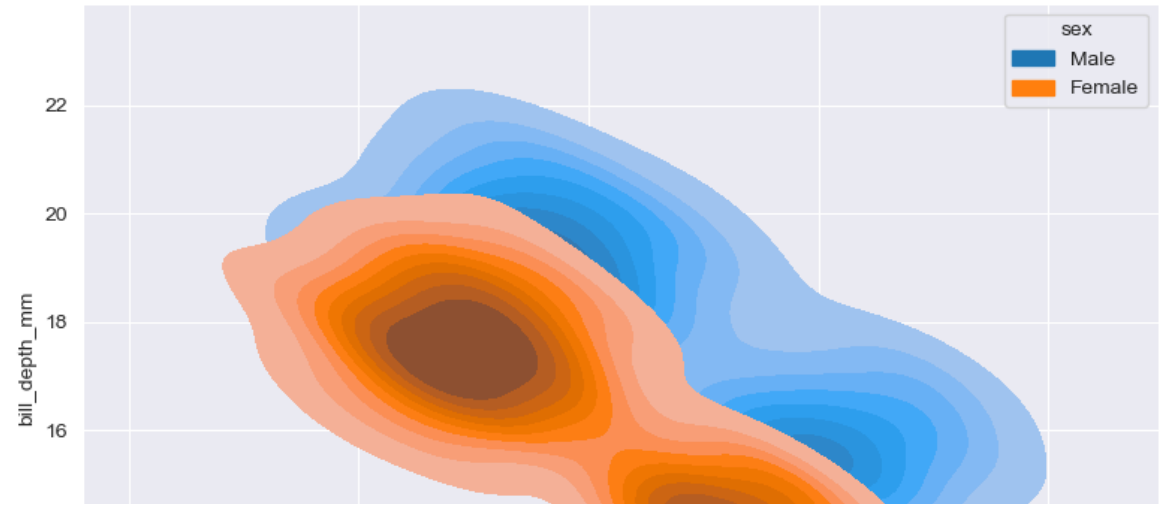
Code

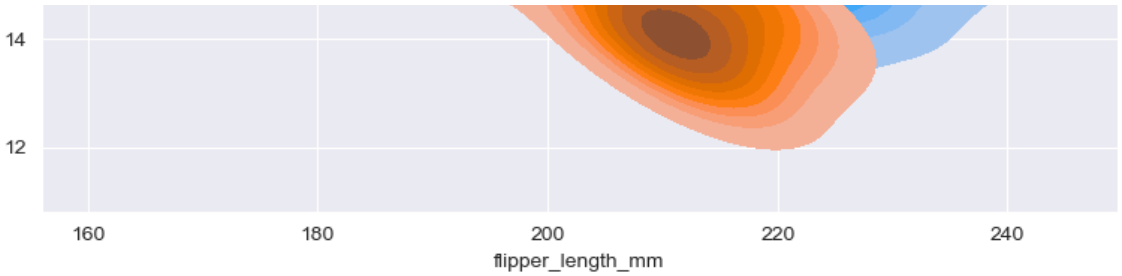
```
#%%
# Question 26
fig_26, axes_26 = plt.subplots(3, 1, figsize=(8, 16))

sns.kdeplot(
    data=penguins,
    x='bill_length_mm',
    y='bill_depth_mm',
    hue='sex',
    fill=True,
    grid=False,
    ax=axes_26[0]
)

sns.kdeplot(
    data=penguins,
    x='bill_length_mm',
    y='flipper_length_mm',
    hue='sex',
    fill=True,
    grid=False,
    ax=axes_26[1]
)

sns.kdeplot(
    data=penguins,
    x='flipper_length_mm',
    y='bill_depth_mm',
    hue='sex',
    fill=True,
    grid=False,
    ax=axes_26[2]
)

plt.tight_layout()
plt.show()

# Print the observation
print("For general, male penguins tend to have longer flipper, bill and deeper
bill,\n"
      "and longer bill longer flipper.\n "
      "Beside that, longer flipper tend to have lighter bill,\n and longer bill
tend to have lighter bill.")
```

Answer


sex
Male
Female
bill_depth_mm
22
20
18
16
14
12
30
35
40
45
50
55
60
65
bill_length_mm


sex
Male
Female
flipper_length_mm
240
220
200
180
160
30
35
40
45
50
55
60
65
bill_length_mm


sex
Male
Female
bill_depth_mm
22
20
18
16




```
    ...:       "and longer bill longer flipper.\n "
    ...:       "Beside that, longer flipper tend to have lighter bill,\n and longer bill tend to have lighter bill
    ...:
For general, male penguins tend to have longer flipper, bill and deeper bill,
and longer bill longer flipper.
 Beside that, longer flipper tend to have lighter bill,
 and longer bill tend to have lighter bill.
```

Question 27

"""

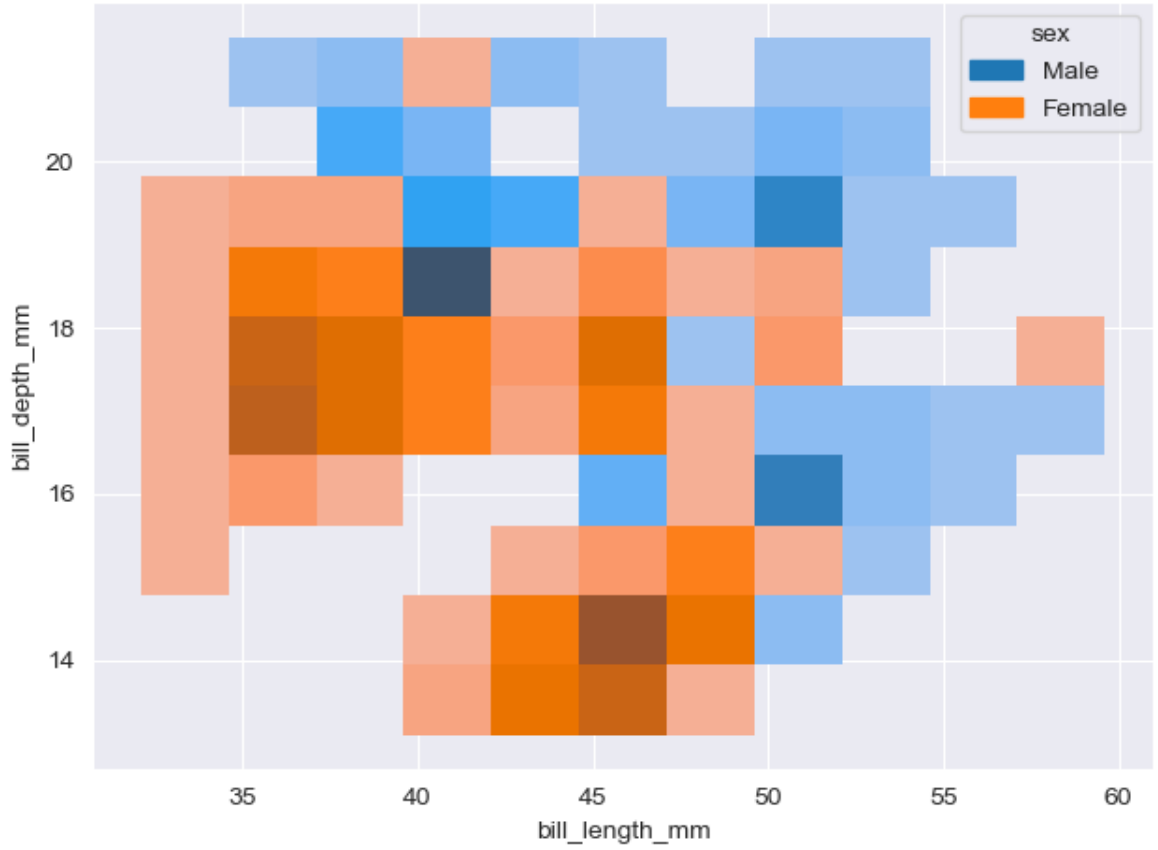
Graph the bivariate distributions between “bill_length_mm” versus “bill_depth_mm” for male and female.

"""

Code

```
#%%
# Question 27
Q27_fig = sns.histplot(
    data=penguins,
    x='bill_length_mm',
    y='bill_depth_mm',
    hue='sex'
)


plt.tight_layout()
plt.show()
```

Answer



Question 28

"""

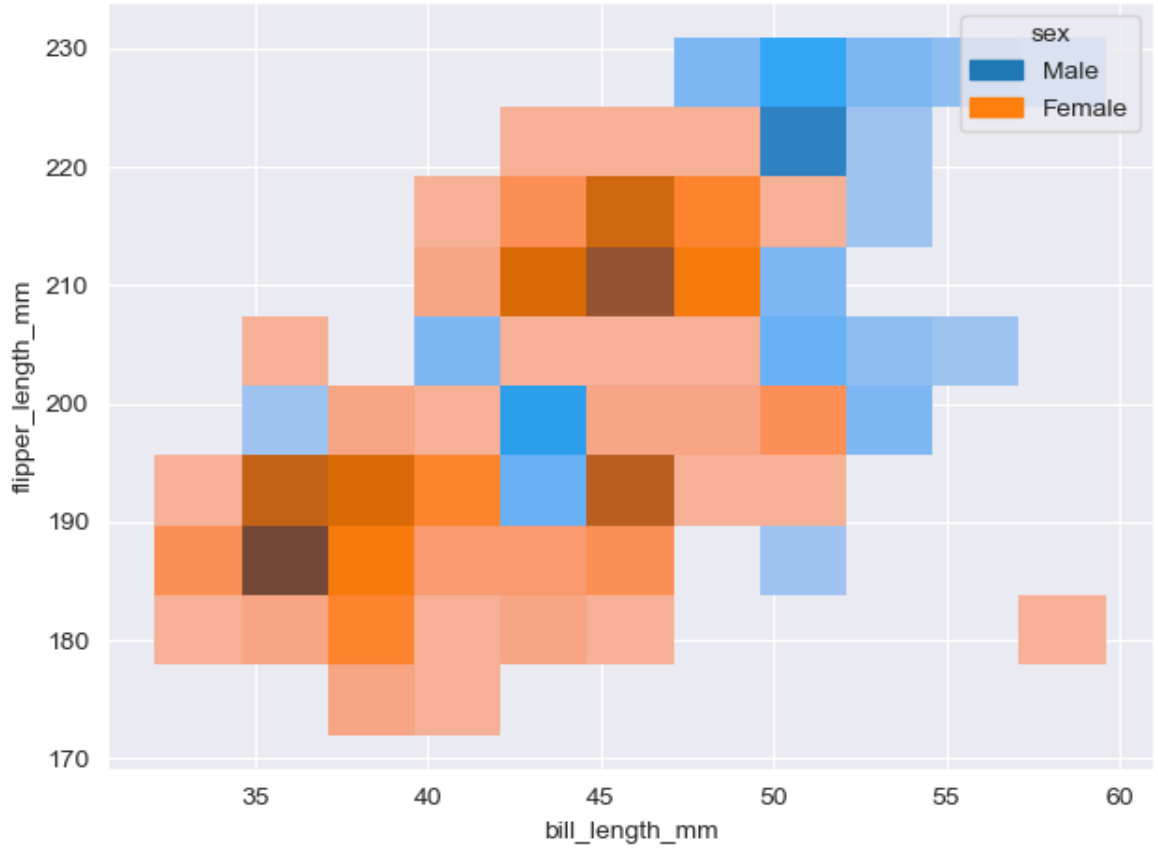
Graph the bivariate distributions between 'bill_length_mm' versus 'flipper_length_mm' for male and female. Final plot like question 27.

"""

Code

```
#%%
# Question 28
Q28_fig = sns.histplot(
    data=penguins,
    x='bill_length_mm',
    y='flipper_length_mm',
    hue='sex'
)


plt.tight_layout()
plt.show()
```

Answer



Question 29

"""

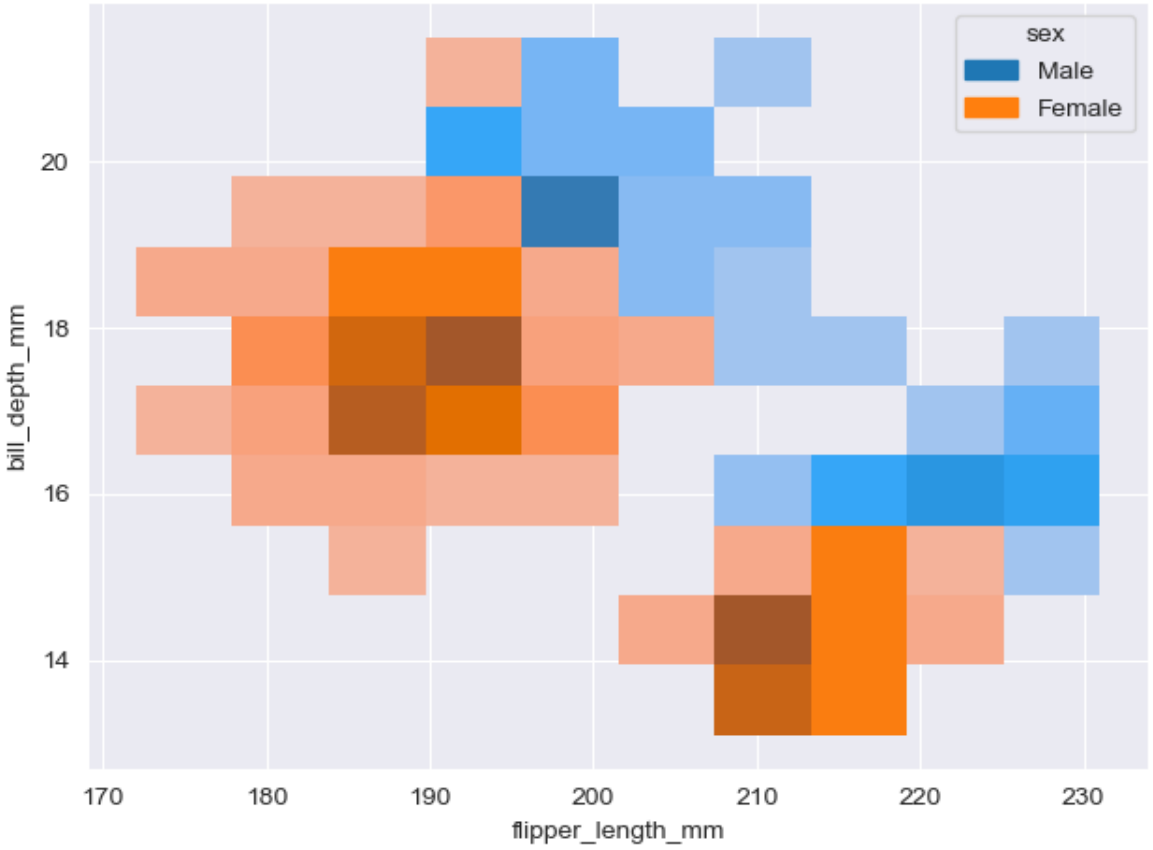
Graph the bivariate distributions between 'flipper_length_mm' versus 'bill_depth_mm' for male and female. Final plot like question 27.

"""

Code

```
#%%
# Question 29
Q29_fig = sns.histplot(
    data=penguins,
    x='flipper_length_mm',
    y='bill_depth_mm',
    hue='sex'
)


plt.tight_layout()
plt.show()
```

Answer


sex
Male
Female
bill_depth_mm
20
18
16
14
170
180
190
200
210
220
230
flipper_length_mm