Projekt na przedmiot Rachunek Prawdopodobieństwa i Statystyka 2022/2023 Analiza statystyczna cech pingwinów z wysp Archipelagu PalmeraPotrzebne importy

In [1]:

```
from palmerpenguins import load_penguins
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.linear_model import LinearRegression
from scipy.stats import norm, kurtosis
from scipy.stats import kstest
import scipy.stats
```

```
C:\Users\Dell\anaconda3\lib\site-packages\scipy\__init__.py:138: UserWarning: A NumPy version >=1.16.5 and <1.23.0 is required for this version of SciPy (detected version 1.23.4)
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion} is required for this version of "
```

Zgrywamy dataset i usuwamy puste pola

In [2]:

```
df = load_penguins()
df.dropna(inplace=True)
penguins = sns.load_dataset('penguins')
penguins.dropna(inplace=True)
```

Chcemy dowiedzieć się podstawowych informacji o naszym datasetcie. Sprawdzamy minimalne i maksymalne wartości, średnie, odchylenia standardowe oraz kwartyle zmiennych.

In [3]:

```
print(df.describe())
```

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

              year  
count   333.000000  
mean   2008.042042  
std       0.812944  
min    2007.000000  
25%    2007.000000  
50%    2008.000000  
75%    2009.000000  
max    2009.000000  
```

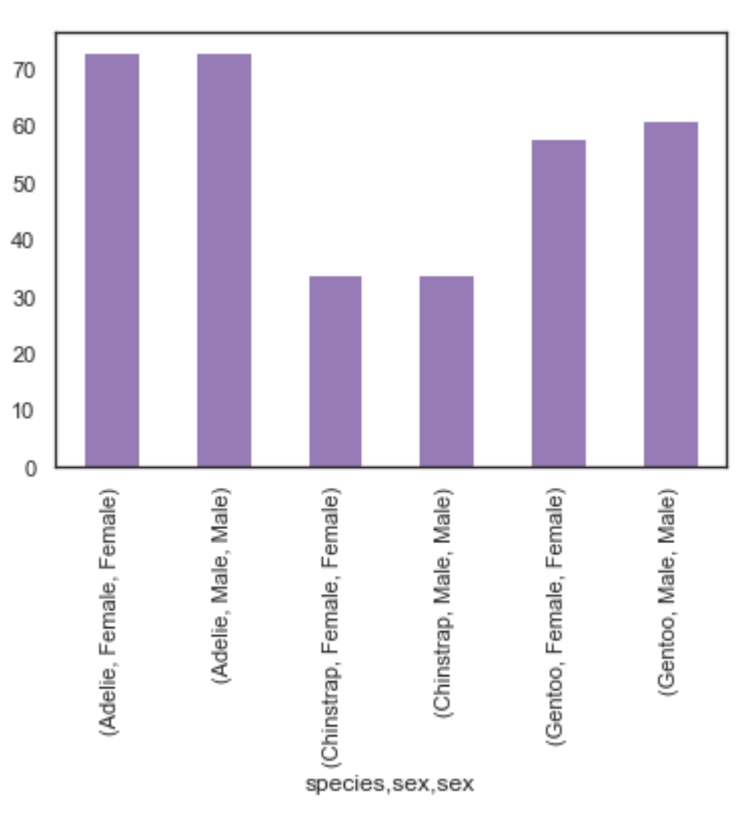
Pokazujemy wykres ilości pingwinów z danego gatunku i płci

In [64]:

```
penguins.groupby(['species','sex'])['sex'].value_counts().plot(kind='bar',color='#967BB6')
```

Out[64]:

```
<AxesSubplot:xlabel='species,sex,sex'>
```



Jak widać najwięcej jest pingwinów Adelie. Widać też, że jest po równo płci.Sprawdzamy korelacje i kowariancję zmiennych.

In [10]:

```
print(df.drop(['sex','year','species','island'],axis=1).cov())
print(df.drop(['sex','year','species','island'],axis=1).corr())
```

```
                   bill_length_mm  bill_depth_mm  flipper_length_mm  \
bill_length_mm          29.906333      -2.462091          50.058195   
bill_depth_mm           -2.462091       3.877888         -15.947248   
flipper_length_mm       50.058195     -15.947248         196.441677   
body_mass_g           2595.623304    -748.456122        9852.191649   

                     body_mass_g  
bill_length_mm       2595.623304  
bill_depth_mm        -748.456122  
flipper_length_mm    9852.191649  
body_mass_g        648372.487699  
                   bill_length_mm  bill_depth_mm  flipper_length_mm  \
bill_length_mm           1.000000      -0.228626           0.653096   
bill_depth_mm           -0.228626       1.000000          -0.577792   
flipper_length_mm        0.653096      -0.577792           1.000000   
body_mass_g              0.589451      -0.472016           0.872979   

                   body_mass_g  
bill_length_mm        0.589451  
bill_depth_mm        -0.472016  
flipper_length_mm     0.872979  
body_mass_g           1.000000  
```

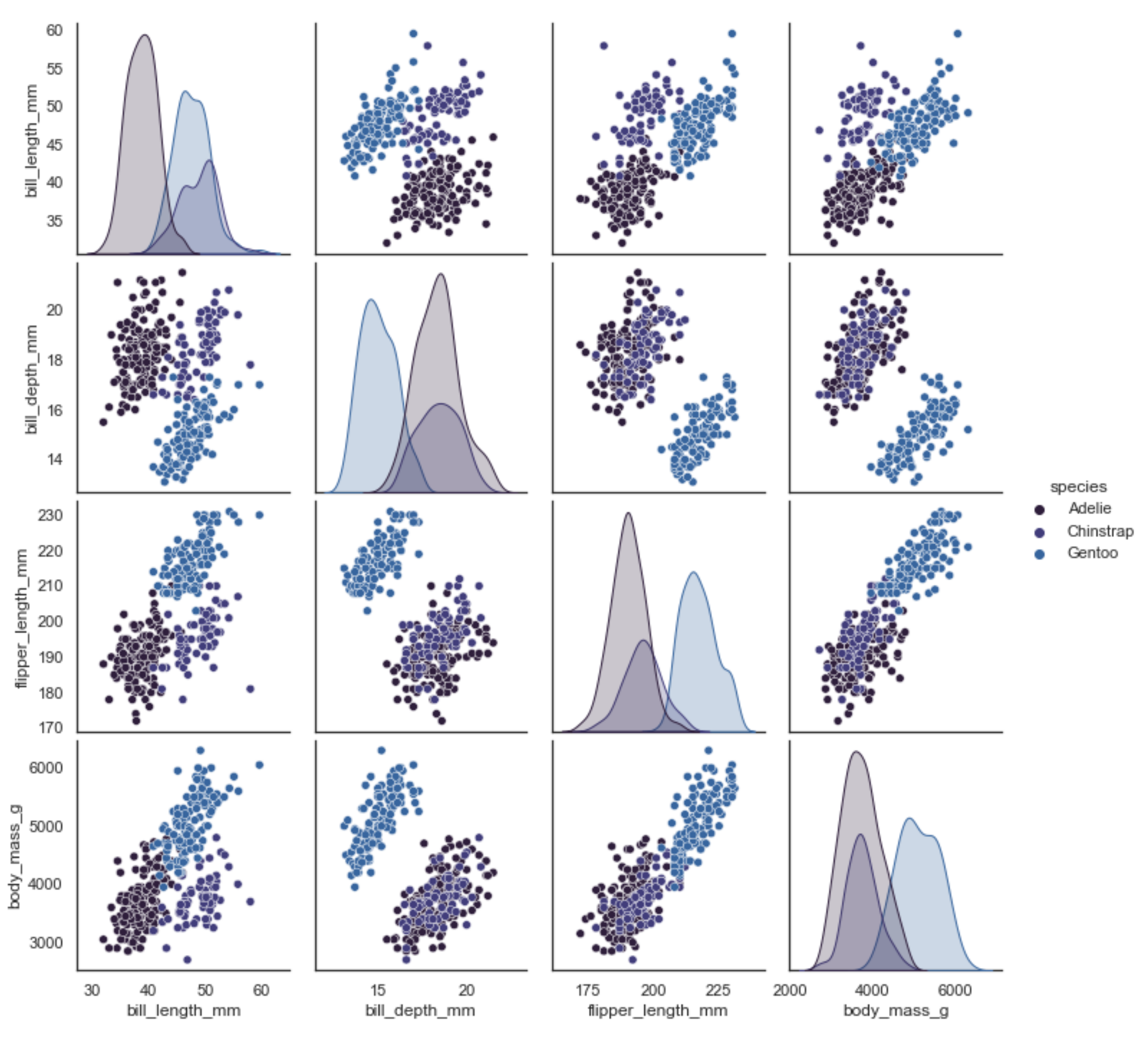
Pokazujemy wizualne rozłożenie danych ze względu na gatunek

In [47]:

```
sns.set(style="white", color_codes=True, palette='mako')
sns.pairplot(penguins,hue='species')
```

Out[47]:

```
<seaborn.axisgrid.PairGrid at 0x2a139af98e0>
```



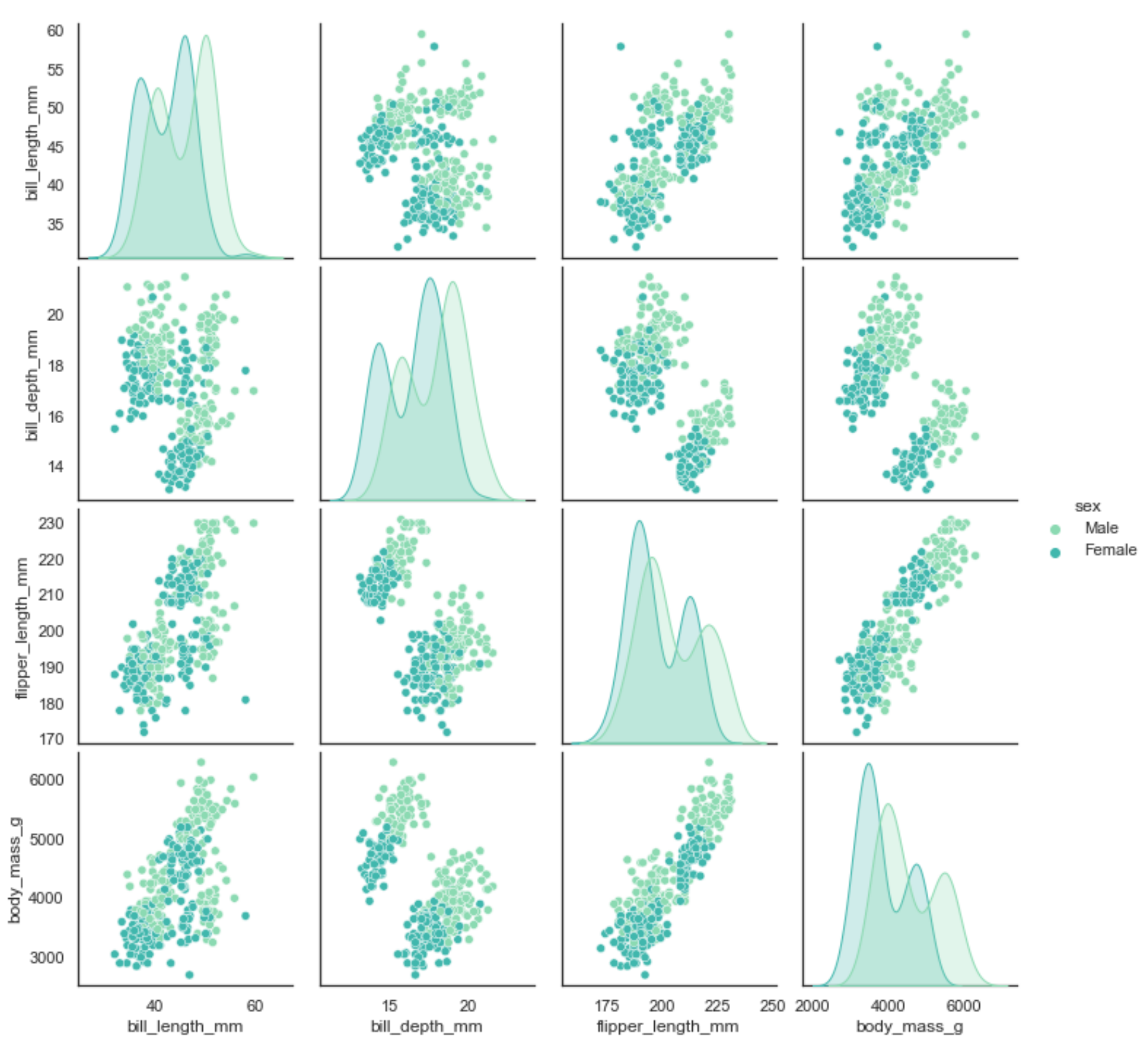
Pokazujemy wizualne rozłożenie danych ze względu na płeć

In [46]:

```
sns.set(style="white", color_codes=True, palette='mako_r')
sns.pairplot(penguins,hue='sex')
```

Out[46]:

```
<seaborn.axisgrid.PairGrid at 0x2a139247820>
```



In [13]:

```
df.drop(['sex','year','island','species'],axis=1,inplace=True)
```

In [ ]:

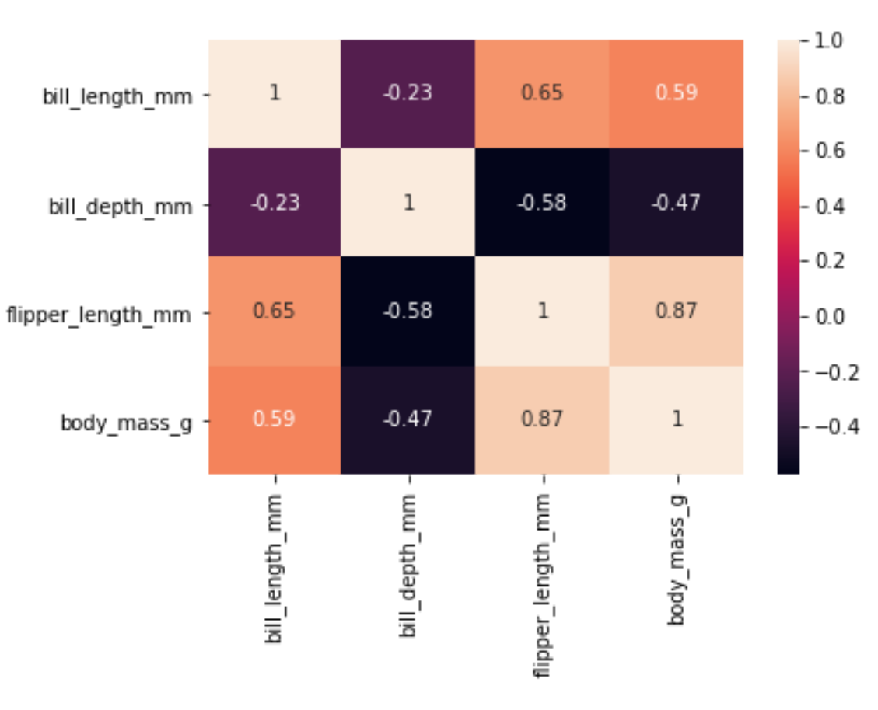
```
Pokazujemy wykres zależności pomiędzy zmiennymi
```

In [14]:

```
sns.heatmap(df.corr(),annot=True)
```

Out[14]:

```
<AxesSubplot:>
```



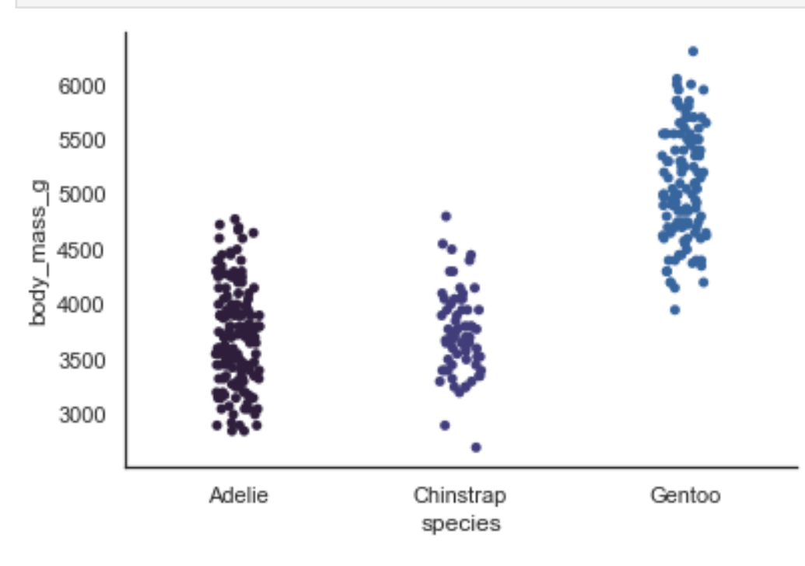
Widać, że masa pingwina i długość płetwy są mocno zależne od siebie, a długość i głębokość dziobu są mało zależne od siebie

In [ ]:

```
Pokazujemy wykres masy od gatunku
```

In [15]:

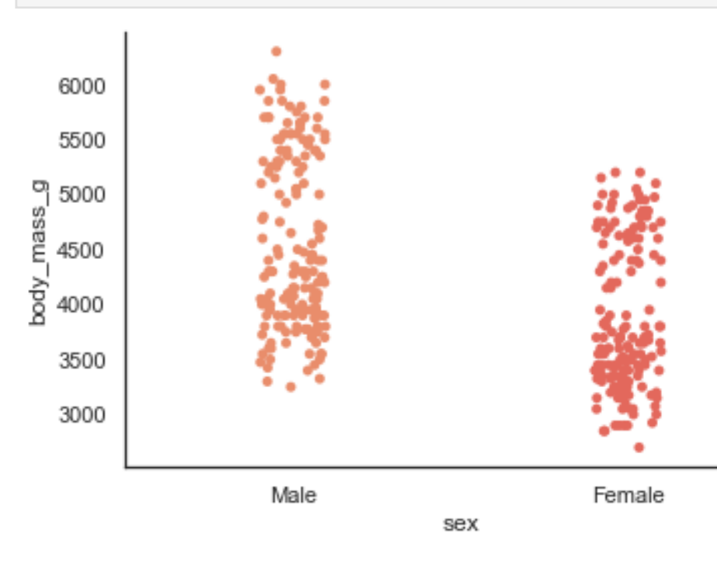
```
sns.set(style="white", color_codes=True, palette='mako')
sns.stripplot(x='species', y='body_mass_g', data=penguins)
sns.despine()
```



Pokazujemy wykres masy od płci

In [16]:

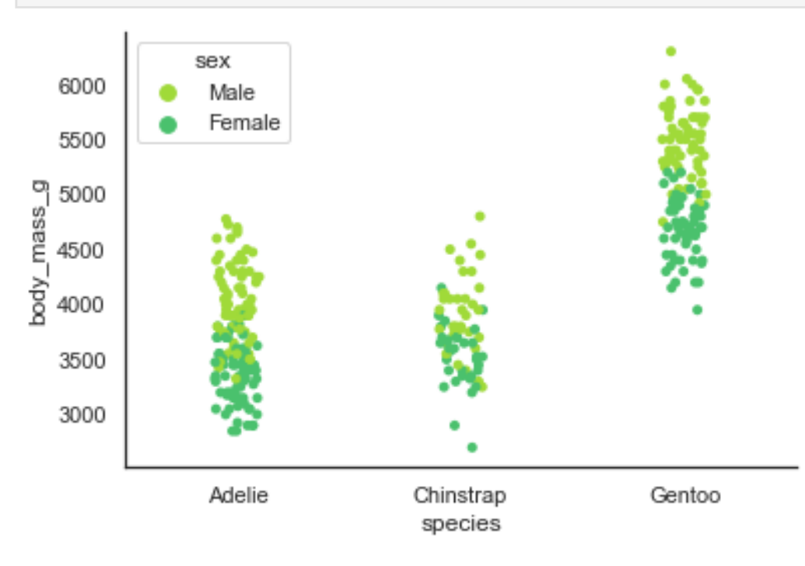
```
sns.set(style="white", color_codes=True, palette='flare')
sns.stripplot(x='sex', y='body_mass_g', data=penguins)
sns.despine()
```



Pokazujemy wykres masy od gatunku i płci

In [38]:

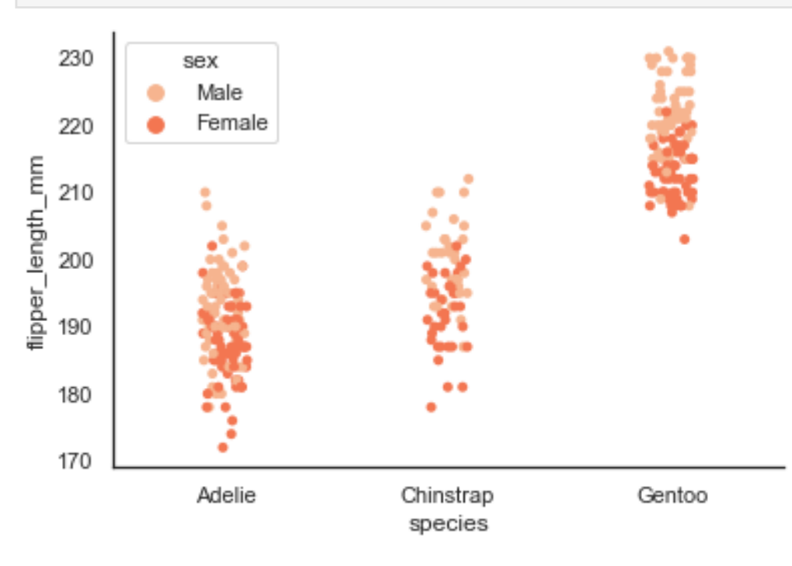
```
sns.set(style="white", color_codes=True, palette='viridis_r')
sns.stripplot(x='species', y='body_mass_g', hue='sex', data=penguins)
sns.despine()
```



Pokazujemy wykres długości płetw od płci i gatunku

In [39]:

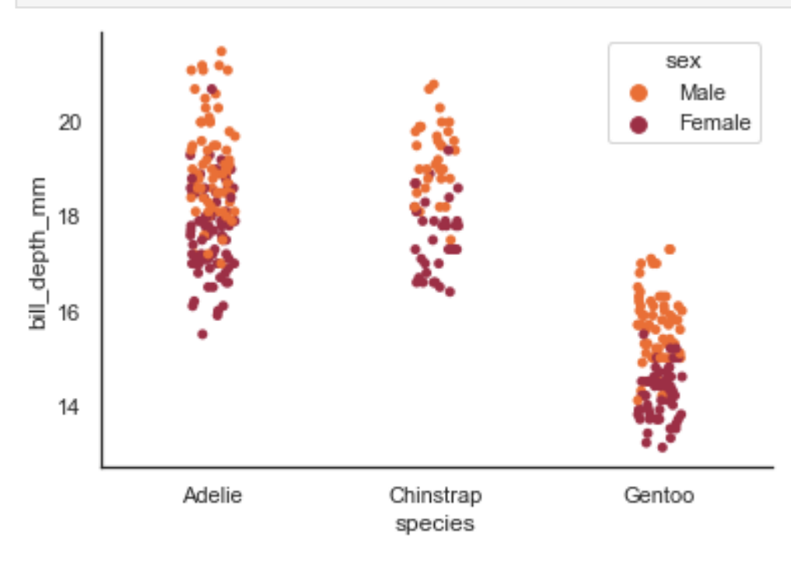
```
sns.set(style="white", color_codes=True, palette='rocket_r')
sns.stripplot(x='species', y='flipper_length_mm', hue='sex', data=penguins)
sns.despine()
```



Pokazujemy wykres głębokości dziobu od płci i gatunku

In [42]:

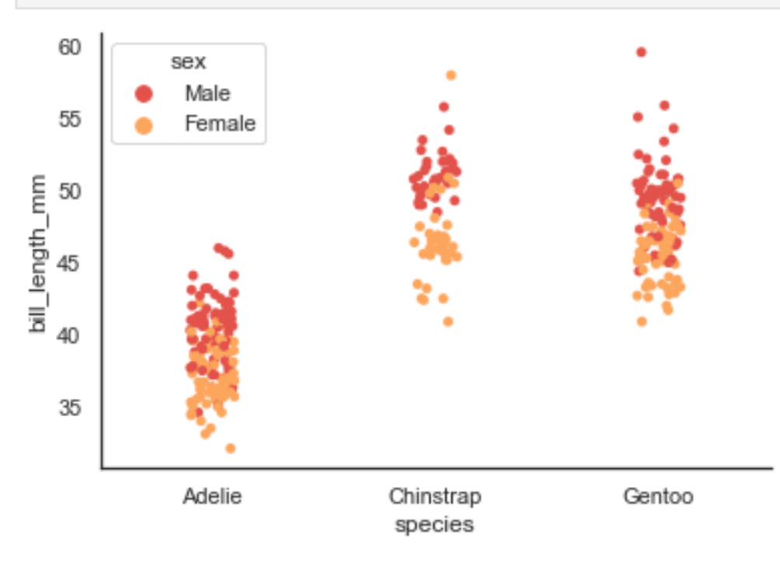
```
sns.set(style="white", color_codes=True, palette='icefire_r')
sns.stripplot(x='species', y='bill_depth_mm', hue='sex', data=penguins)
sns.despine()
```



Pokazujemy wykres długości dziobu od płci oraz gatunku

In [44]:

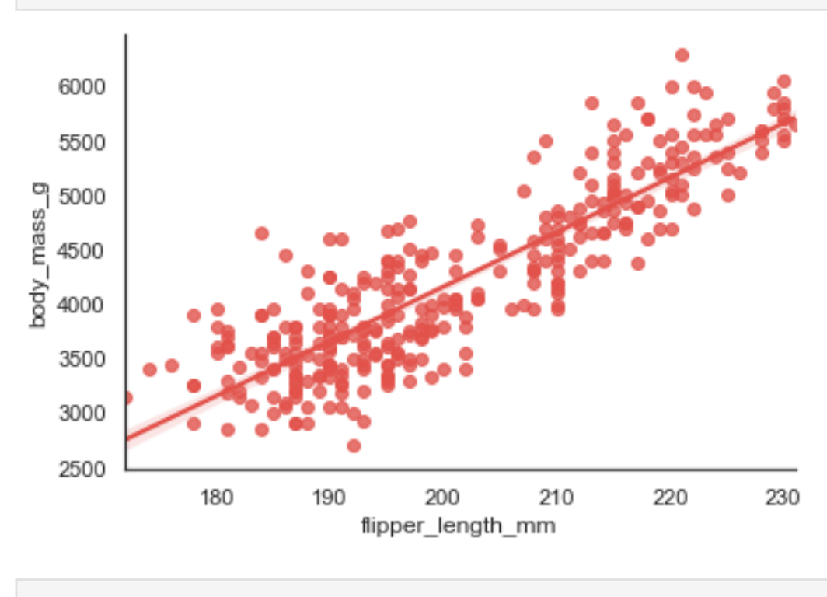
```
sns.set(style="white", color_codes=True, palette='Spectral')
sns.stripplot(x='species', y='bill_length_mm', hue='sex', data=penguins)
sns.despine()
```



Sprawdzamy czy istnieje korelacja, więc tworzymy wykres regresji długości płetw od masy i wyliczamy wartości współczynników regresji

In [59]:

```
sns.set(style="white", color_codes=True, palette='Spectral')
sns.regplot(x='flipper_length_mm',y='body_mass_g',data=penguins)
sns.despine()
```



In [33]:

```
X = df.drop(['bill_length_mm','bill_depth_mm'],axis=1)
Y = df['body_mass_g']
reg = LinearRegression().fit(X, Y)
print(reg.score(X,Y))
print(reg.coef_)
print(reg.intercept_)
```

```
1.0
[5.91596047e-15 1.00000000e+00]
-1.8189894035458565e-12
```

Sprawdzamy czy rozkład głębokości dziobu jest rozkładem normalnym

In [78]:

```
print(kurtosis(df['bill_depth_mm']))
```

```
-0.8965872511276172
```

In [79]:

```
print(df['bill_depth_mm'].skew())
```

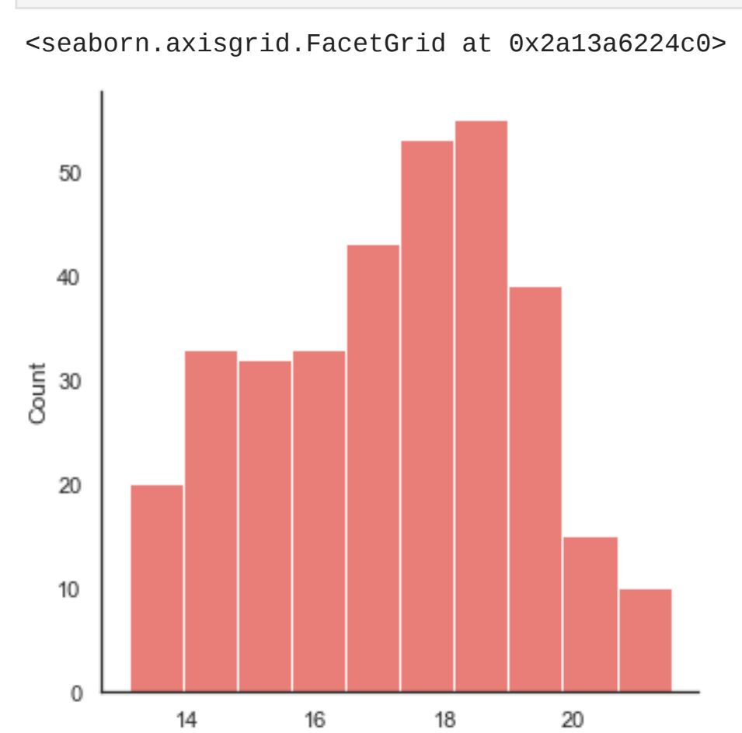
```
-0.1497202576146911
```

In [90]:

```
sns.displot(penguins['bill_depth_mm'])
```

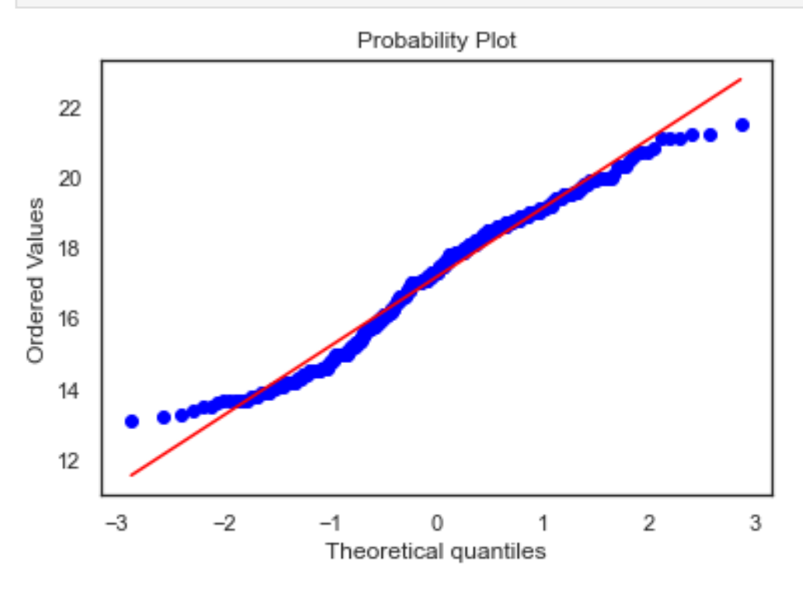
Out[90]:

```
<seaborn.axisgrid.FacetGrid at 0x2a13a6224c0>
```



In [87]:

```
scipy.stats.probplot(penguins['bill_depth_mm'],dist="norm", plot=plt)
plt.show()
```



In [88]:

```
print(kstest(penguins['bill_depth_mm'],'norm'))
```

```
KstestResult(statistic=1.0, pvalue=0.0)
```

Z testów wynika, że dystrybucja jest dystrybucją normalną