```
#D.Vishaal
#21BRS1173
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df=pd.read_csv('penguins_size.csv')
```

```
df.head(15)
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_ma
0	Adelie	Torgersen	39.1	18.7	181.0	37
1	Adelie	Torgersen	39.5	17.4	186.0	38
2	Adelie	Torgersen	40.3	18.0	195.0	32
3	Adelie	Torgersen	NaN	NaN	NaN	
4	Adelie	Torgersen	36.7	19.3	193.0	34
5	Adelie	Torgersen	39.3	20.6	190.0	36
6	Adelie	Torgersen	38.9	17.8	181.0	36
7	Adelie	Torgersen	39.2	19.6	195.0	46
8	Adelie	Torgersen	34.1	18.1	193.0	34
9	Adelie	Torgersen	42.0	20.2	190.0	42
10	Adelie	Torgersen	37.8	17.1	186.0	33
11	Adelie	Torgersen	37.8	17.3	180.0	37
12	Adelie	Torgersen	41.1	17.6	182.0	32
13	Adelie	Torgersen	38.6	21.2	191.0	38
14	Adelie	Torgersen	34.6	21.1	198.0	4

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   species            344 non-null    object 
 1   island             344 non-null    object 
 2   culmen_length_mm   342 non-null    float64
 3   culmen_depth_mm    342 non-null    float64
 4   flipper_length_mm  342 non-null    float64
 5   body_mass_g        342 non-null    float64
 6   sex                334 non-null    object 
dtypes: float64(4), object(3)
memory usage: 18.9+ KB
```

```
df.describe()
```

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
count	342.000000	342.000000	342.000000	342.000000
mean	43.921930	17.151170	200.915205	4201.754386
std	5.459584	1.974793	14.061714	801.954536
min	32.100000	13.100000	172.000000	2700.000000
25%	39.225000	15.600000	190.000000	3550.000000
50%	44.450000	17.300000	197.000000	4050.000000

```
df.isnull().sum()
```

```
species               0
island                0
culmen_length_mm      2
culmen_depth_mm       2
flipper_length_mm     2
body_mass_g           2
sex                  10
dtype: int64
```

```
sns.distplot(df['culmen_length_mm'])
```

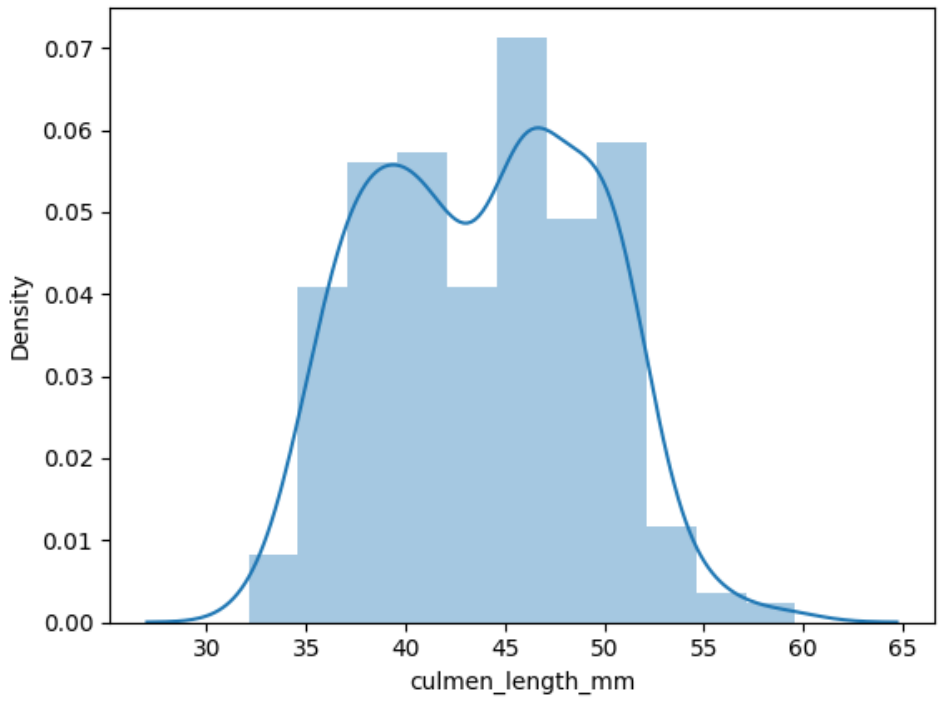
```
<ipython-input-14-87f900721a46>:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df['culmen_length_mm'])
<Axes: xlabel='culmen_length_mm', ylabel='Density'>
```




```
sns.distplot(df['culmen_depth_mm'])
```

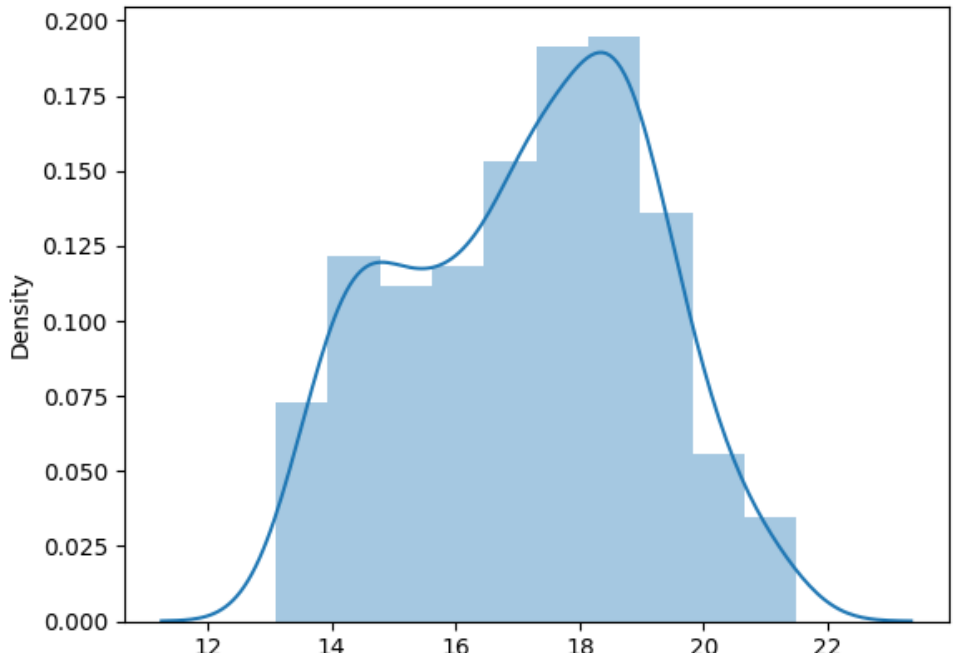
```
<ipython-input-15-9161f519b2fb>:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df['culmen_depth_mm'])
<Axes: xlabel='culmen_depth_mm', ylabel='Density'>
```




```
sns.distplot(df['flipper_length_mm'])
```

```
<ipython-input-16-25d29e01b18c>:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.
```

```
df['culmen_length_mm']=df['culmen_length_mm'].fillna(df['culmen_length_mm'].median())
df['culmen_depth_mm']=df['culmen_depth_mm'].fillna(df['culmen_depth_mm'].median ())
df['flipper_length_mm']=df['flipper_length_mm'].fillna(df['flipper_length_mm'].median())
df['body_mass_8']=df['body_mass_g'].fillna(df['body_mass_g'].median())
```

```
df.isnull().sum()
```

```
species              0
island               0
culmen_length_mm     0
culmen_depth_mm      0
flipper_length_mm    0
body_mass_g          2
sex                 10
body_mass_8          0
dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   species            344 non-null    object 
 1   island             344 non-null    object 
 2   culmen_length_mm   344 non-null    float64
 3   culmen_depth_mm    344 non-null    float64
 4   flipper_length_mm  344 non-null    float64
 5   body_mass_g        342 non-null    float64
 6   sex                334 non-null    object 
 7   body_mass_8        344 non-null    float64
dtypes: float64(5), object(3)
memory usage: 21.6+ KB
```

```
df['sex']=df['sex'].fillna(df['sex'].mode()[0])
df['sex']
```

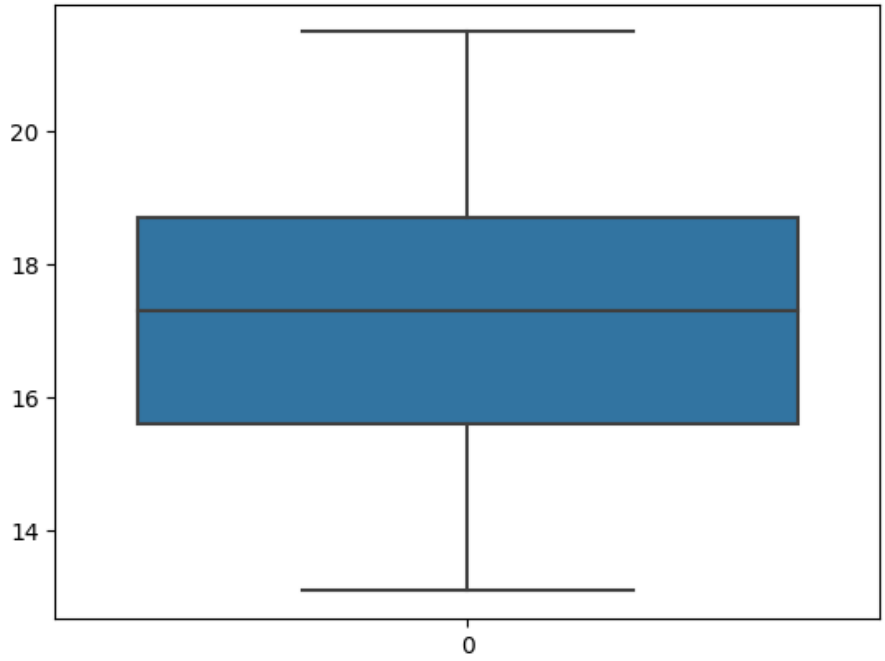
```
0       MALE
1     FEMALE
2     FEMALE
3       MALE
4     FEMALE
       ...
339     MALE
340   FEMALE
341     MALE
342   FEMALE
343     MALE
Name: sex, Length: 344, dtype: object
```

```
df.isnull().sum()
```

```
species              0
island               0
culmen_length_mm     0
culmen_depth_mm      0
flipper_length_mm    0
body_mass_g          2
sex                  0
body_mass_8          0
dtype: int64
```

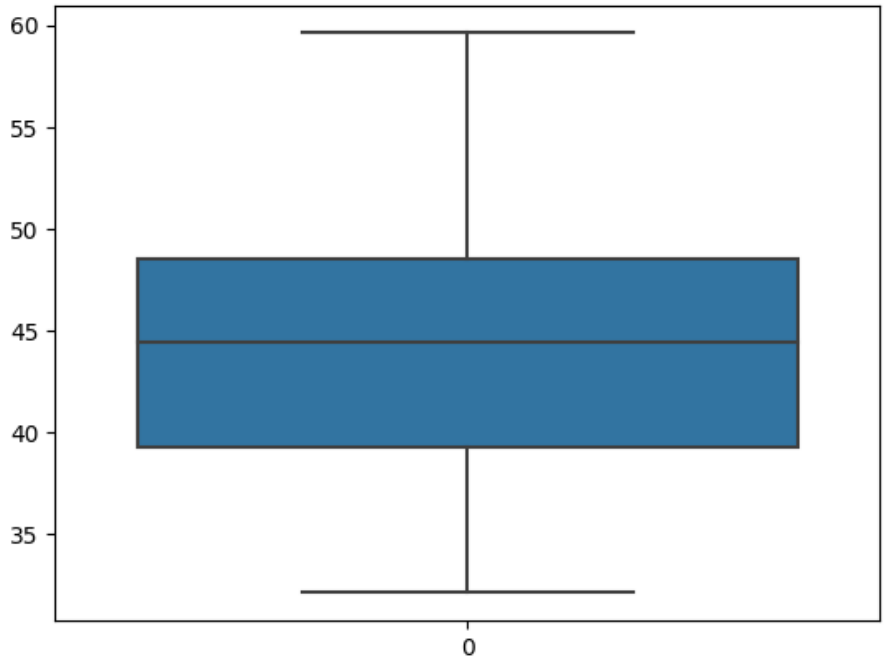
```
sns.boxplot(df['culmen_depth_mm'])
```

```
<Axes: >
```



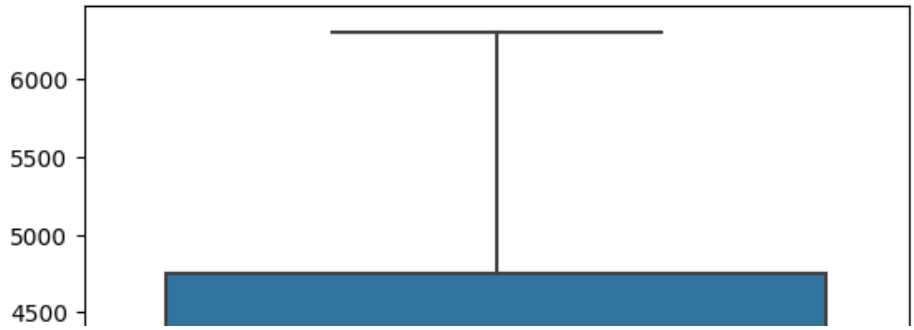

```
sns.boxplot(df['culmen_length_mm'])
```

```
<Axes: >
```



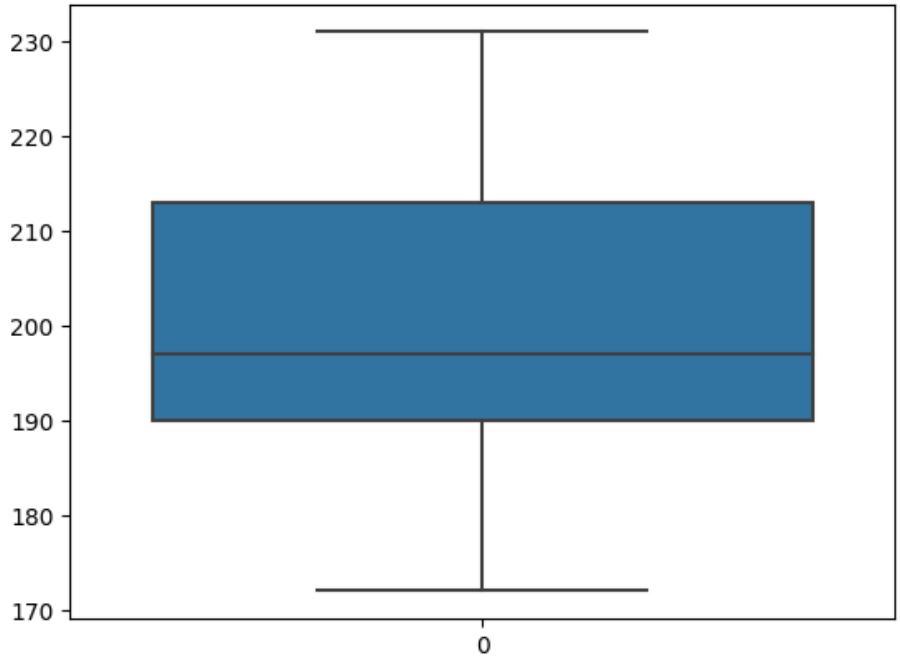

```
sns.boxplot(df['body_mass_g'])
```

```
<Axes: >
```




```
sns.boxplot(data=df['flipper_length_mm'])
```

```
<Axes: >
```




```
from sklearn import preprocessing


label_encoder=preprocessing.LabelEncoder()


df['species']=label_encoder.fit_transform(df['species'])
df.tail(10)
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	body_
334	2	Biscoe	46.20	14.1	217.0	4375.0	FEMALE	

```
df['island']=label_encoder.fit_transform(df['island'])
df['island'].unique()
```

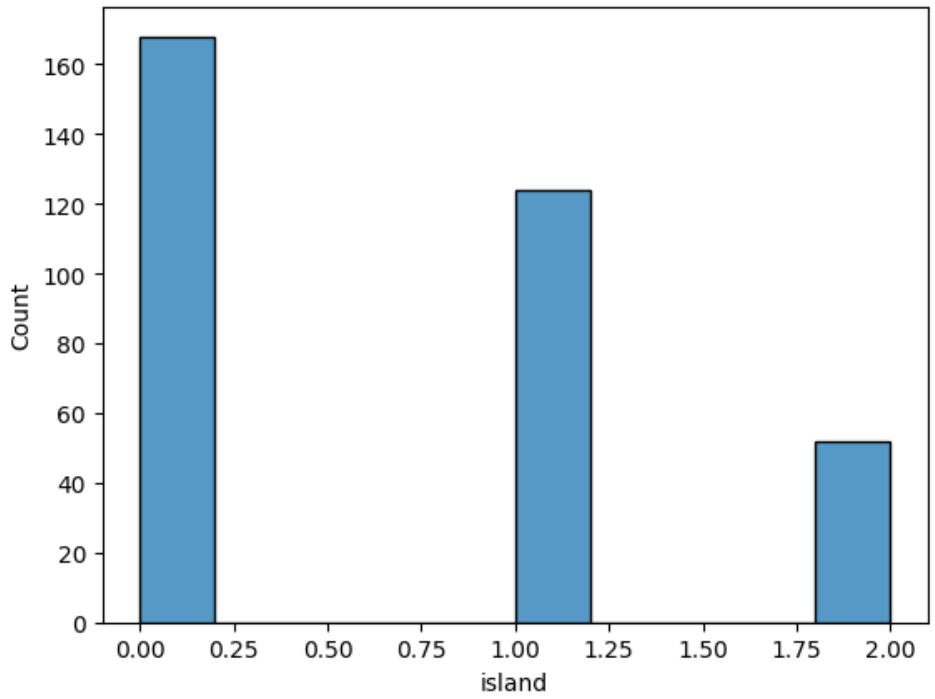
```
array([2, 0, 1])
```

```
df.head()
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	body_ma
0	0	2	39.10	18.7	181.0	3750.0	MALE	37
1	0	2	39.50	17.4	186.0	3800.0	FEMALE	38
2	0	2	40.30	18.0	195.0	3250.0	FEMALE	32
3	0	2	44.45	17.3	197.0	NaN	MALE	40
4	0	2	36.70	19.3	193.0	3450.0	FEMALE	34

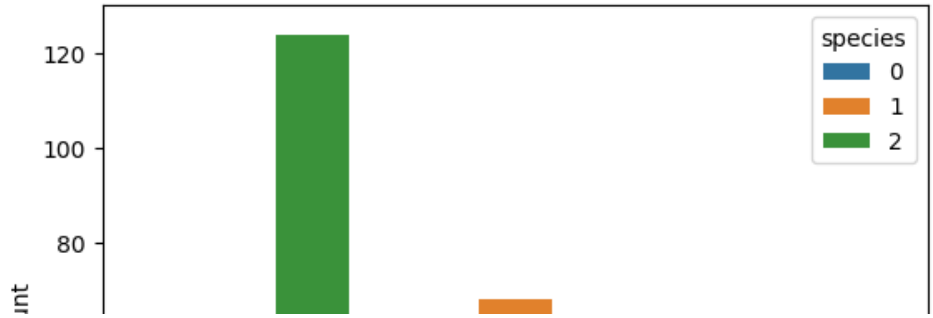
```
sns.histplot(data=df['island'])
```

```
<Axes: xlabel='island', ylabel='Count'>
```



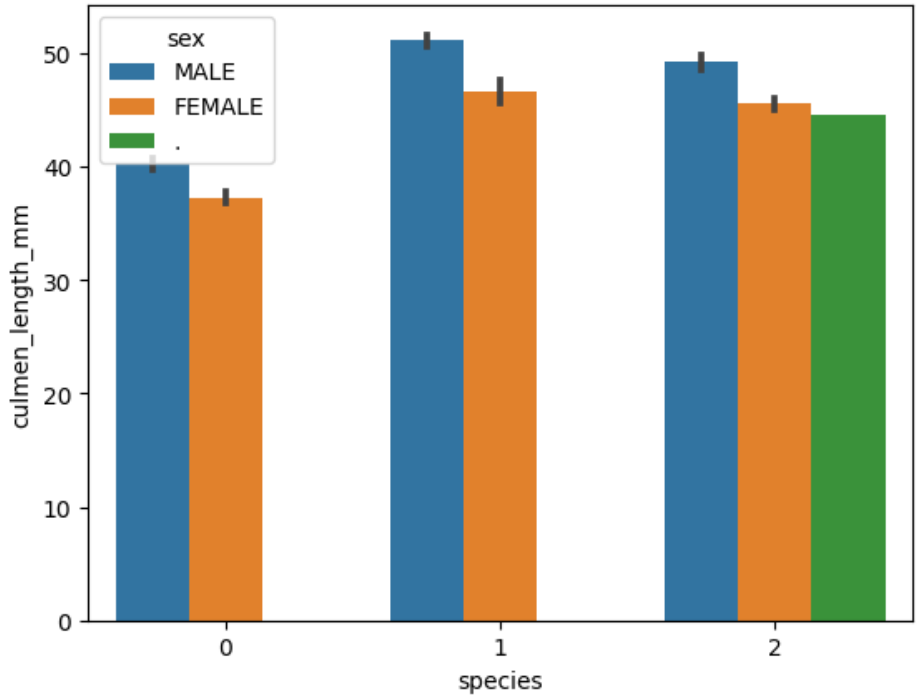

```
sns.countplot(x='island',hue='species',data=df)
```

```
<Axes: xlabel='island', ylabel='count'>
```



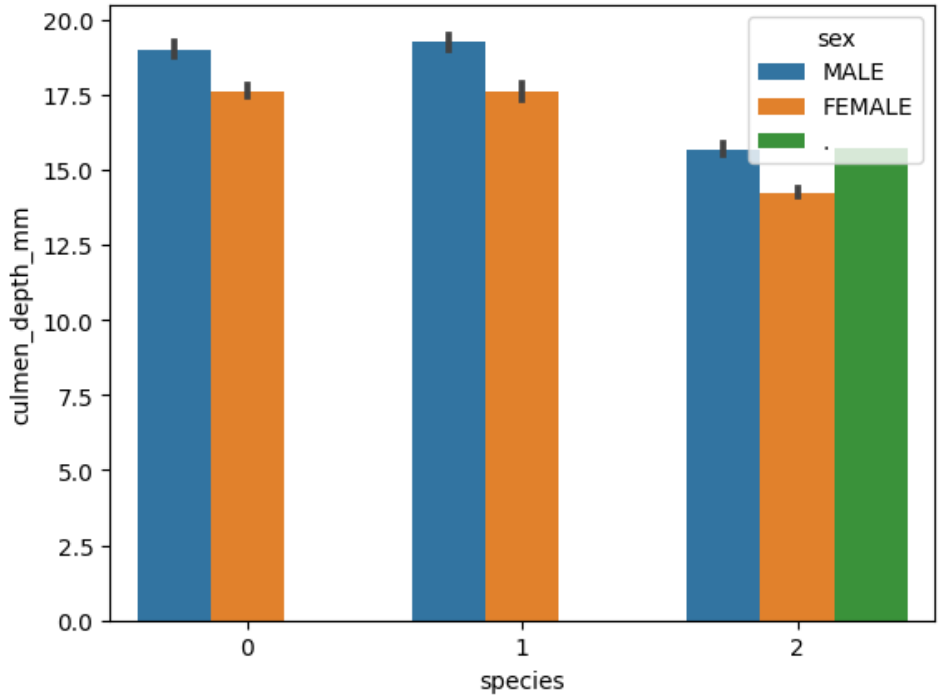

```
sns.barplot(x='species',y='culmen_length_mm',hue='sex',data=df)
```

```
<Axes: xlabel='species', ylabel='culmen_length_mm'>
```



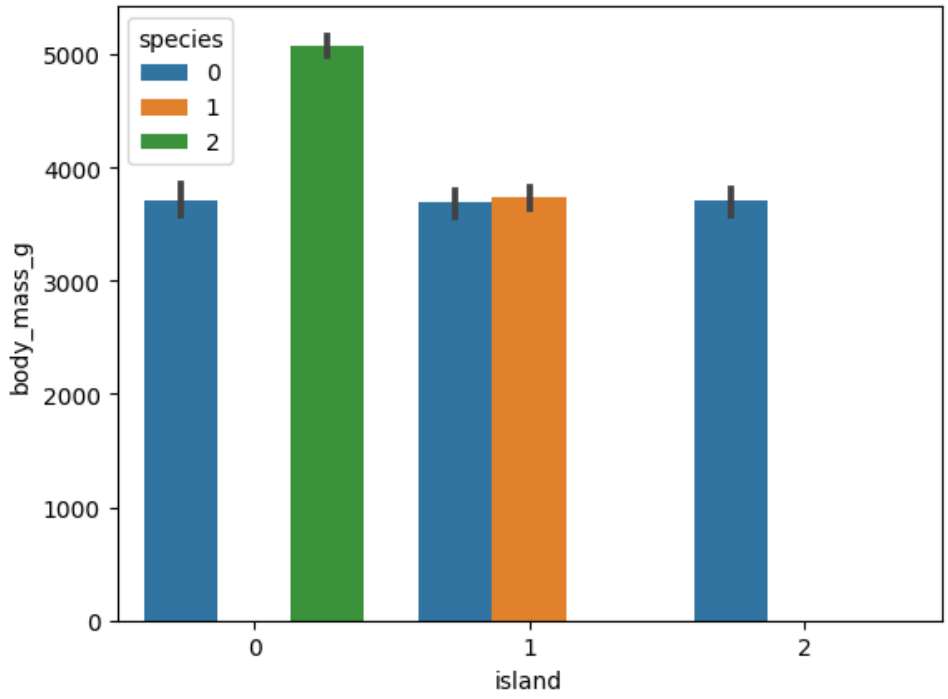

```
sns.barplot(x='species',y='culmen_depth_mm',hue='sex',data=df)
```

```
<Axes: xlabel='species', ylabel='culmen_depth_mm'>
```



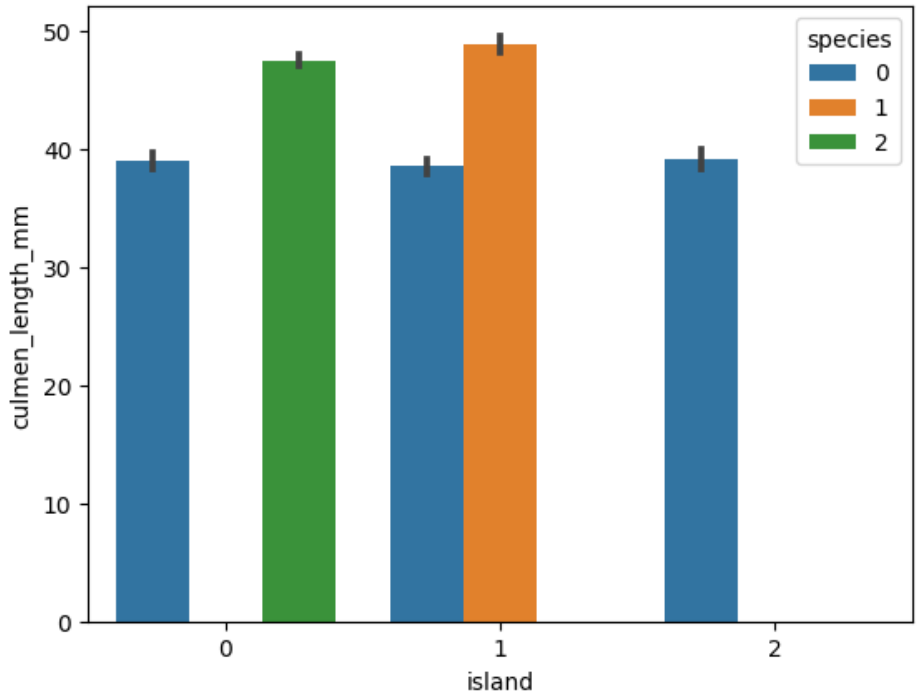
```
sns.barplot(x='island',y='body_mass_g',hue='species',data=df)
```

```
<Axes: xlabel='island', ylabel='body_mass_g'>
```



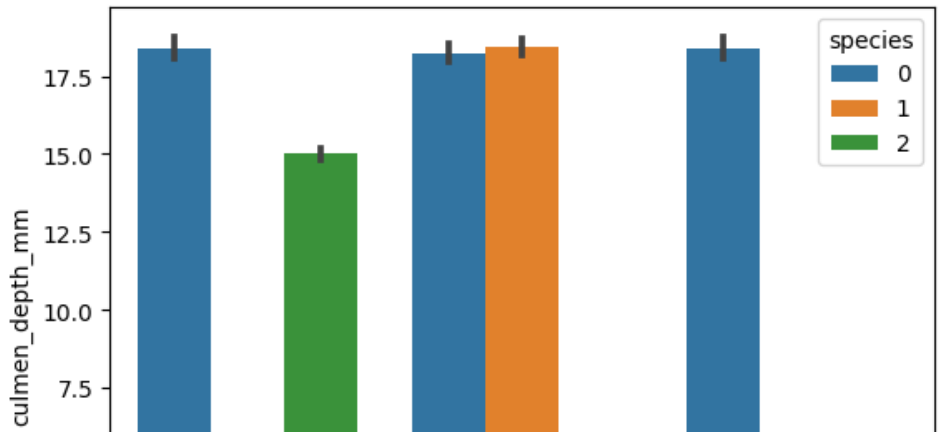

```
sns.barplot(x='island',y='culmen_length_mm',hue='species',data=df)
```

```
<Axes: xlabel='island', ylabel='culmen_length_mm'>
```




```
sns.barplot(x='island',y='culmen_depth_mm',hue='species',data=df)
```

```
<Axes: xlabel='island', ylabel='culmen_depth_mm'>
```




```
X=df.drop('species',axis=1)
y=df['species']
X.head()
```

	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	body_mass_8
0	2	39.10	18.7	181.0	3750.0	2	3750.0
1	2	39.50	17.4	186.0	3800.0	1	3800.0
2	2	40.30	18.0	195.0	3250.0	1	3250.0
3	2	44.45	17.3	197.0	NaN	2	4050.0
4	2	36.70	19.3	193.0	3450.0	1	3450.0

```
y.head()
```

```
0    0
1    0
2    0
3    0
4    0
Name: species, dtype: int64
```

```
df['sex']=label_encoder.fit_transform(df['sex'])
df['sex']
```

```
0      2
1      1
2      1
3      2
4      1
      ..
339    2
340    1
341    2
342    1
343    2
Name: sex, Length: 344, dtype: int64
```

```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
```

```
X_scaled=pd.DataFrame(sc.fit_transform(X),columns =X.columns)
X_scaled.head()
```

	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	body_mass_8
0	1.844076	-0.887622	0.787289	-1.420541	-0.564142	0.960230	-0.564625
1	1.844076	-0.814037	0.126114	-1.063485	-0.501703	-1.017729	-0.502010
2	1.844076	-0.666866	0.431272	-0.420786	-1.188532	-1.017729	-1.190773

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X_scaled,y,test_size=0.2,random_state=42)
X_train.shape
```

```
(275, 7)
```

```
y_train.shape
```

```
(275,)
```

```
X_test.shape
```

```
(69, 7)
```

```
y_test.shape
```

```
(69,)
```