Name:Shruthi Gangam
Registration number:21BCB0063
Campus: Vellore

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sea

df=pd.read_csv('/content/penguins_size.csv')

df.shape

(344, 7)

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
75%           48.500000        18.700000         213.000000
4750.000000
max           59.600000        21.500000         231.000000
6300.000000

df.head(5)
```

```
  species     island  culmen_length_mm  culmen_depth_mm
flipper_length_mm  \
0  Adelie  Torgersen              39.1             18.7
181.0
1  Adelie  Torgersen              39.5             17.4
186.0
2  Adelie  Torgersen              40.3             18.0
195.0
3  Adelie  Torgersen               NaN              NaN
NaN
4  Adelie  Torgersen              36.7             19.3
193.0

   body_mass_g     sex
0       3750.0    MALE
1       3800.0  FEMALE
2       3250.0  FEMALE
3          NaN     NaN
4       3450.0  FEMALE
```

Visualization

Univariant Analysis

```
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
```

```
df.fillna(df.median(),inplace=True)
```

```
<ipython-input-49-e1b8fa6d4ecf>:1: FutureWarning: The default value of
numeric_only in DataFrame.median is deprecated. In a future version,
it will default to False. In addition, specifying 'numeric_only=None'
is deprecated. Select only valid columns or specify the value of
numeric_only to silence this warning.
  df.fillna(df.median(),inplace=True)
```

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

```
species               0
island                0
culmen_length_mm      0
culmen_depth_mm       0
```

```
flipper_length_mm     0
body_mass_g           0
sex                  10
dtype: int64
```

```
df.sex.value_counts()
```

```
MALE      168
FEMALE    165
.           1
Name: sex, dtype: int64
```

```
df.sex.fillna('MALE',inplace=True)
```

```
df.sex.replace(to_replace='.', value='MALE',inplace=True)
```

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

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

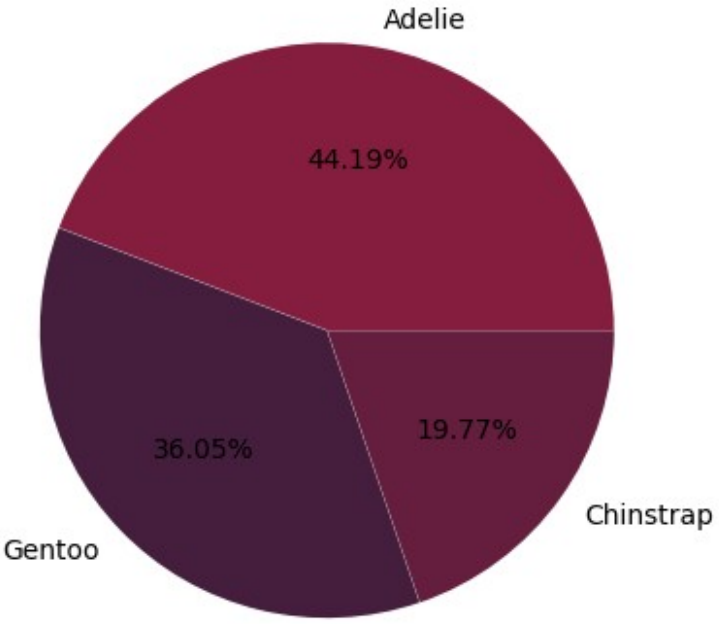
Visualization

Univariant

```
df.species.value_counts()
```

```
Adelie       152
Gentoo       124
Chinstrap     68
Name: species, dtype: int64
```

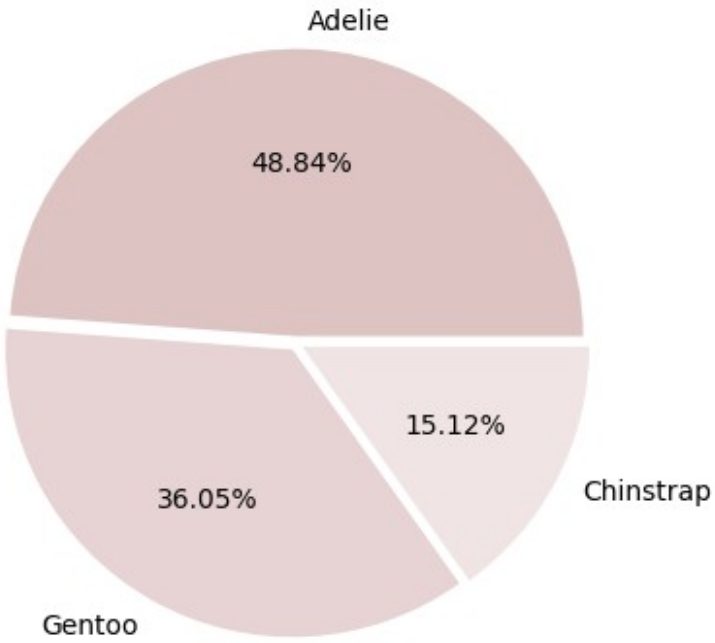
```
plt.pie(df['species'].value_counts(),autopct='%0.2f%
%',colors=['#851e3e','#451e3e','#651e3e'],labels=df['species'].value_c
ounts().keys(),explode=[0.001,0.001,0.001])
plt.show()
```




```
df.island.value_counts()
```

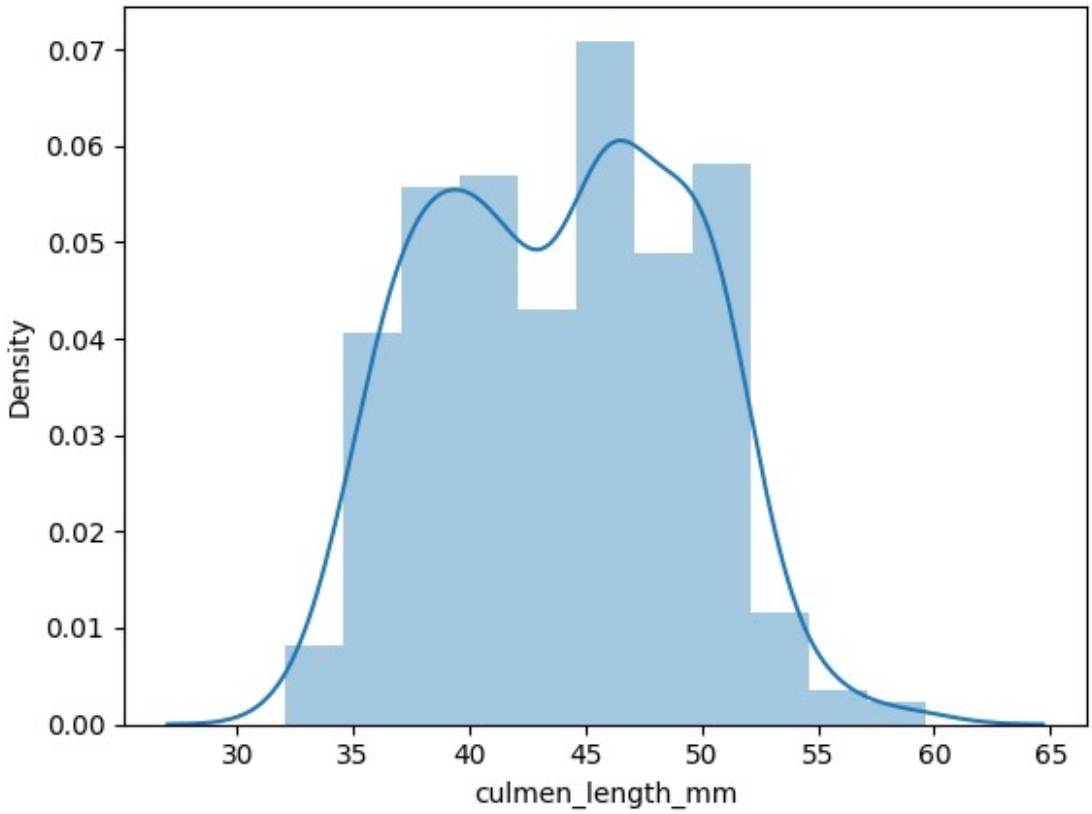
```
Biscoe       168
Dream        124
Torgersen     52
Name: island, dtype: int64
```

```
plt.pie(df['island'].value_counts(),autopct='%0.2f%
%',colors=['#dec3c3','#e7d3d3','#f0e4e4'],labels=df['species'].value_c
ounts().keys(),explode=[0.025,0.025,0.025])
plt.show()
```



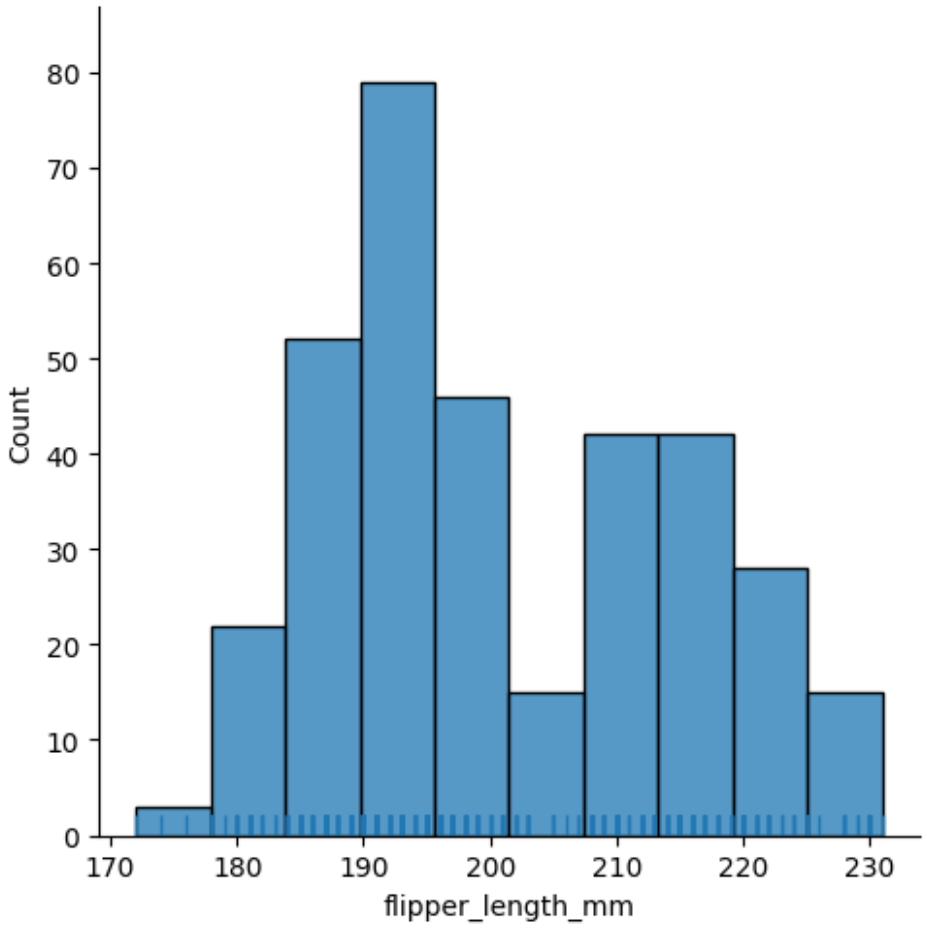

```
sea.distplot(df['culmen_length_mm'])
```

```
<ipython-input-59-f3489f6ab27f>: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

  sea.distplot(df['culmen_length_mm'])

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



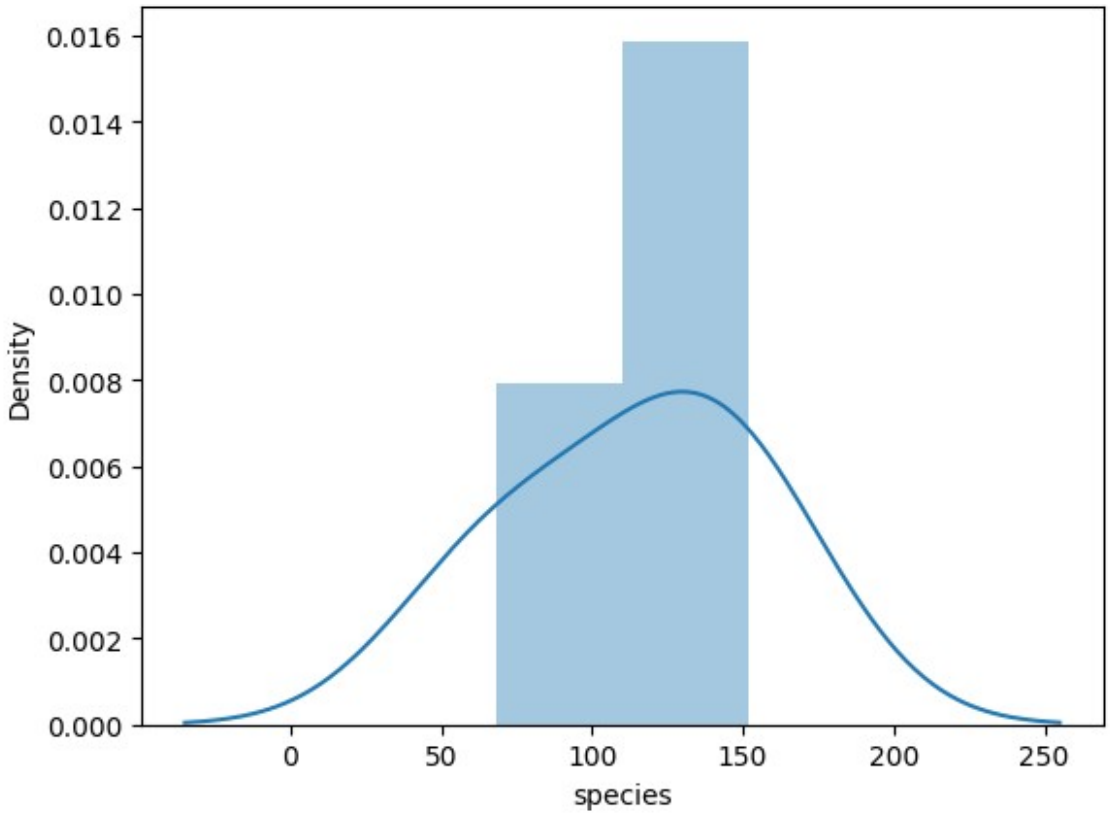

```
sea.displot(df.flipper_length_mm,rug=True)
```

```
<seaborn.axisgrid.FacetGrid at 0x7d471e6f6ce0>
```




```
sea.distplot(df.species.value_counts())
```

```
<ipython-input-61-c68065d769b1>: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

  sea.distplot(df.species.value_counts())

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



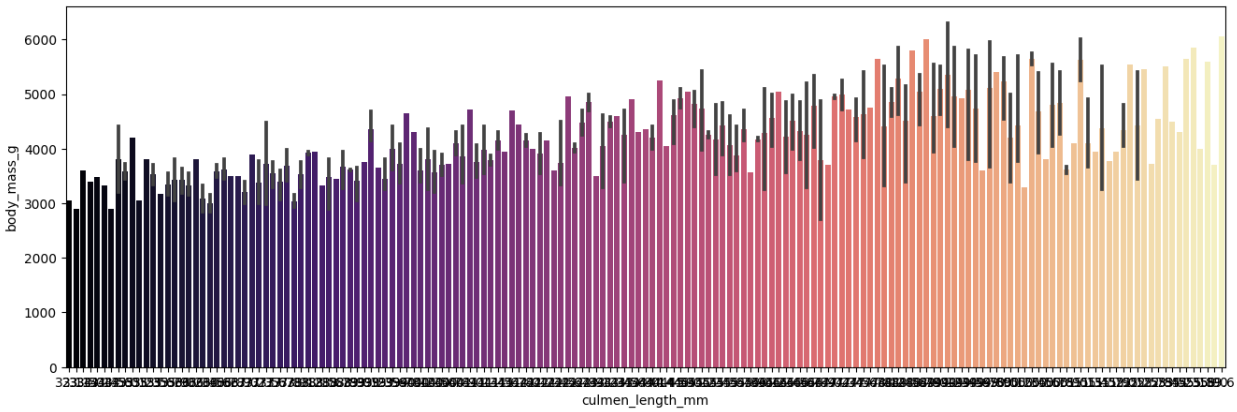

```
df.species.value_counts()
```

```
Adelie       152
Gentoo       124
Chinstrap     68
Name: species, dtype: int64
```

Bivariant Analysis

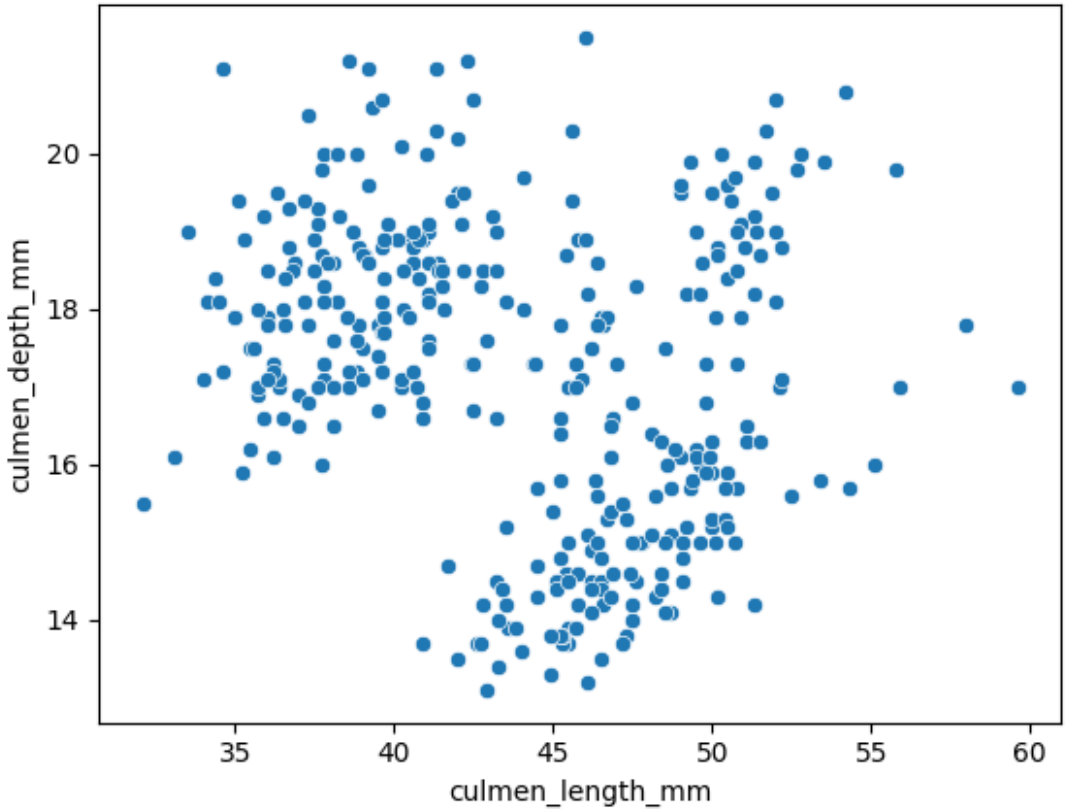
```
plt.figure(figsize=(16,5))
sea.barplot(x='culmen_length_mm',y='body_mass_g',data=df,palette='magma')
```

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



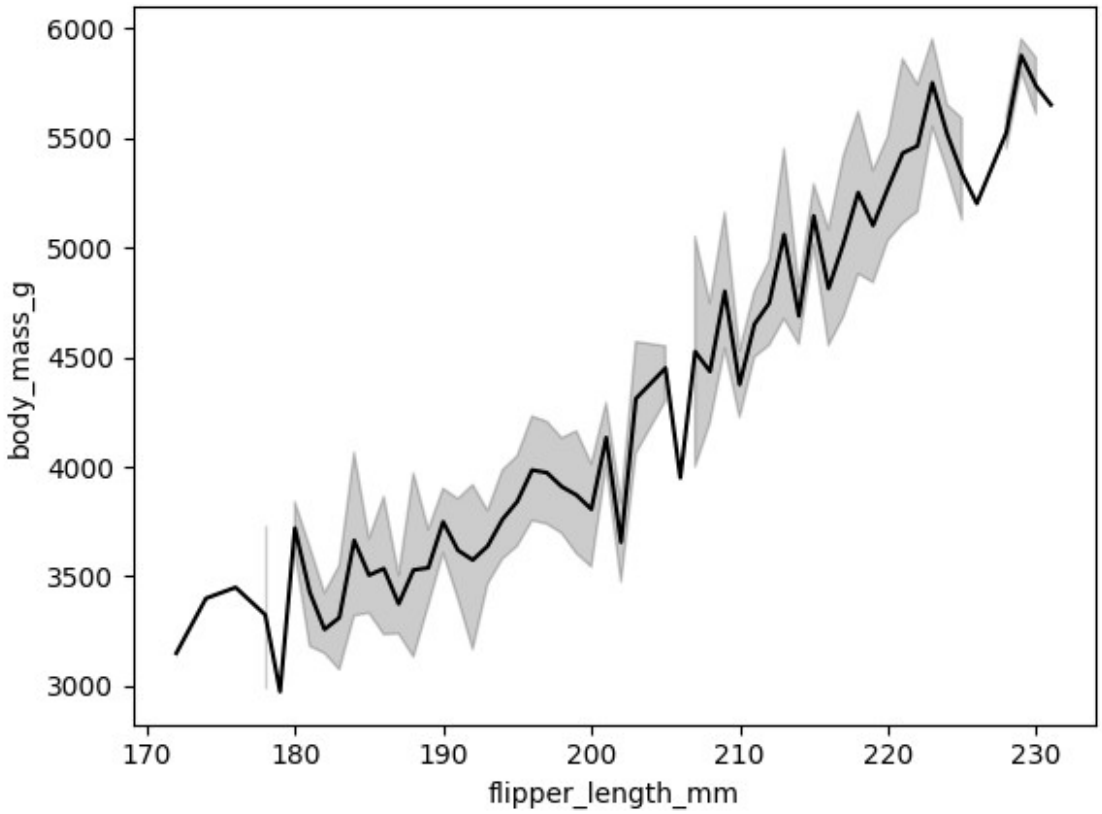

```
sea.scatterplot(x='culmen_length_mm',y='culmen_depth_mm',data=df)
```

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




```
sea.lineplot(x='flipper_length_mm',y='body_mass_g',data=df,color='blac
k')
```

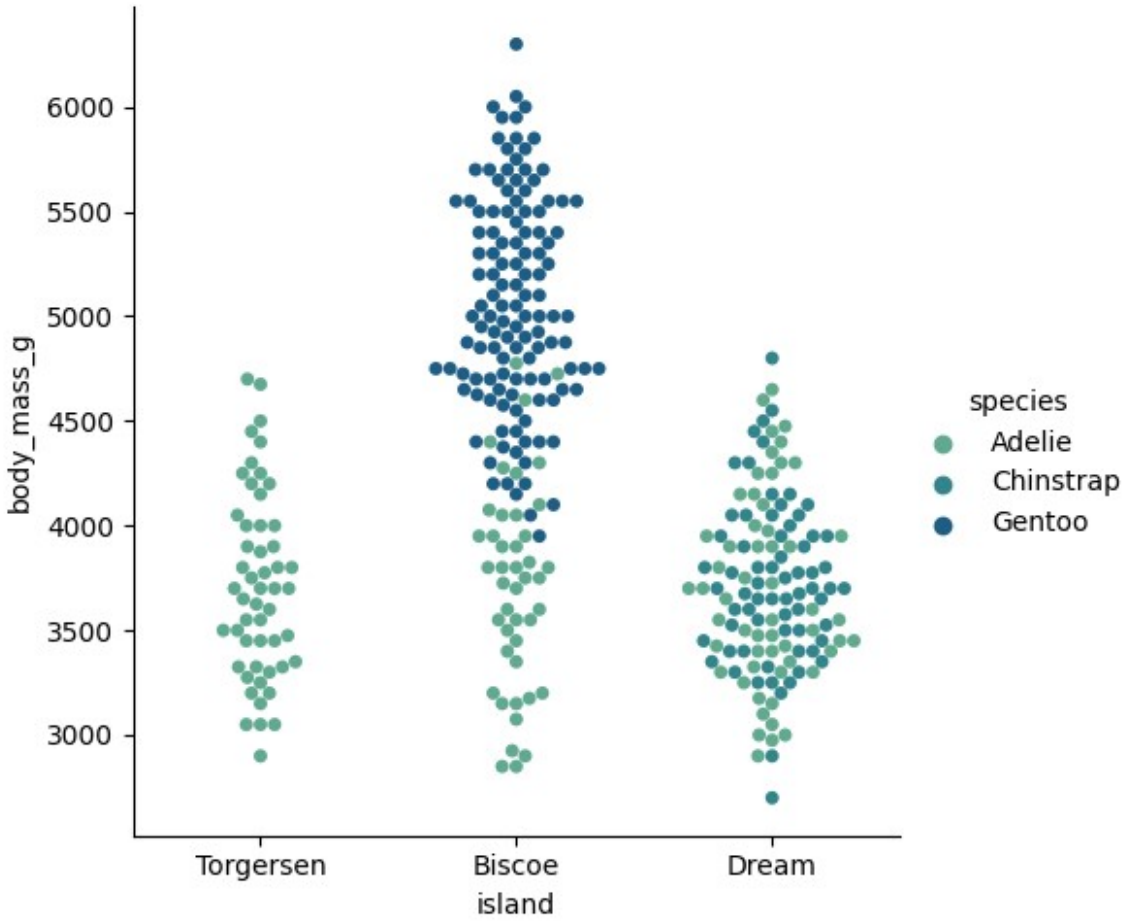
```
<Axes: xlabel='flipper_length_mm', ylabel='body_mass_g'>
```



Multivariant Analysis

```
sea.catplot(data=df,x='island',y='body_mass_g',hue='species',kind='swarm',palette='crest')
```

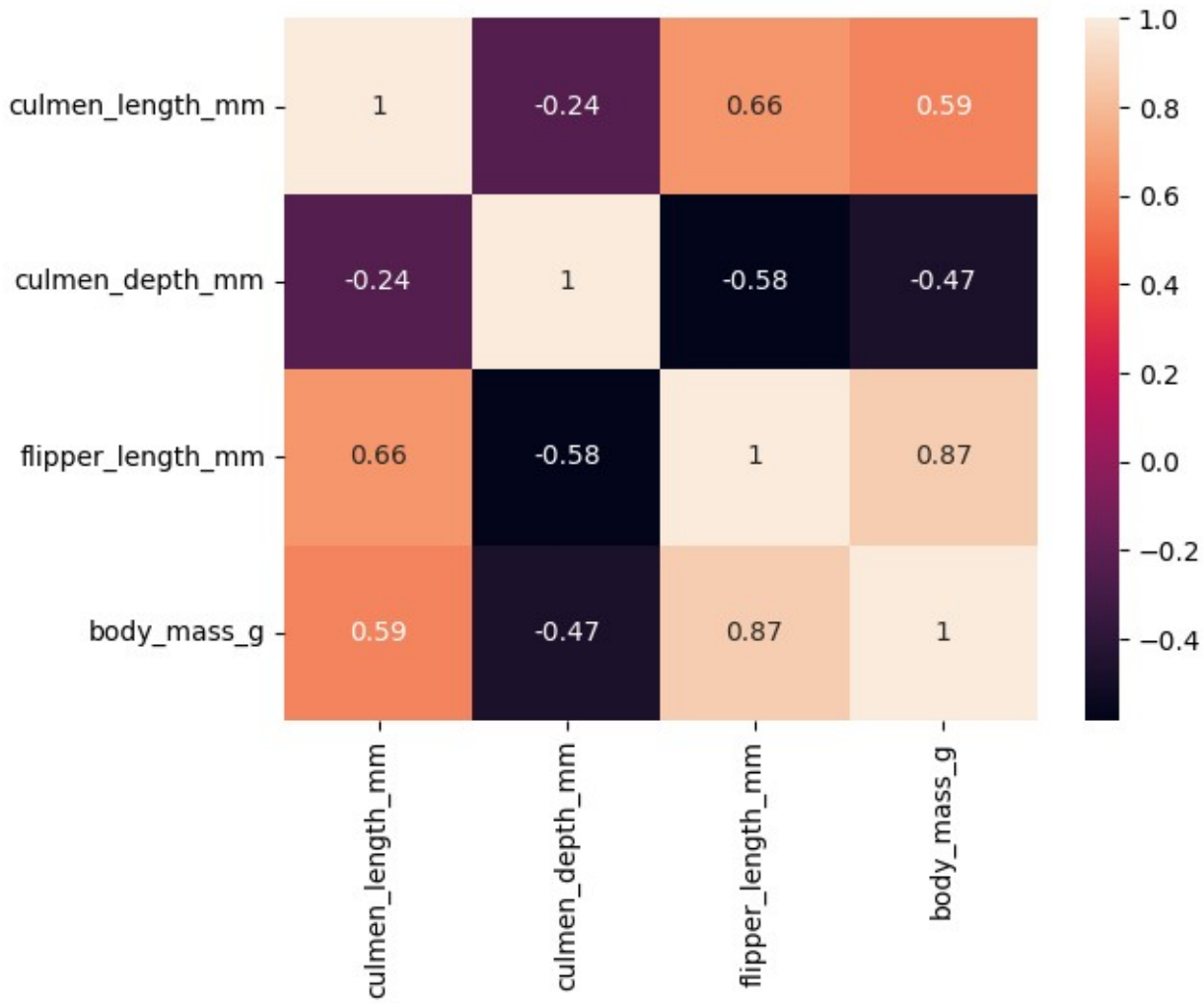
```
<seaborn.axisgrid.FacetGrid at 0x7d471ed7ead0>
```




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

```
<ipython-input-67-c38365734727>:1: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it
will default to False. Select only valid columns or specify the value
of numeric_only to silence this warning.
  sea.heatmap(df.corr(),annot=True)
```

```
<Axes: >
```



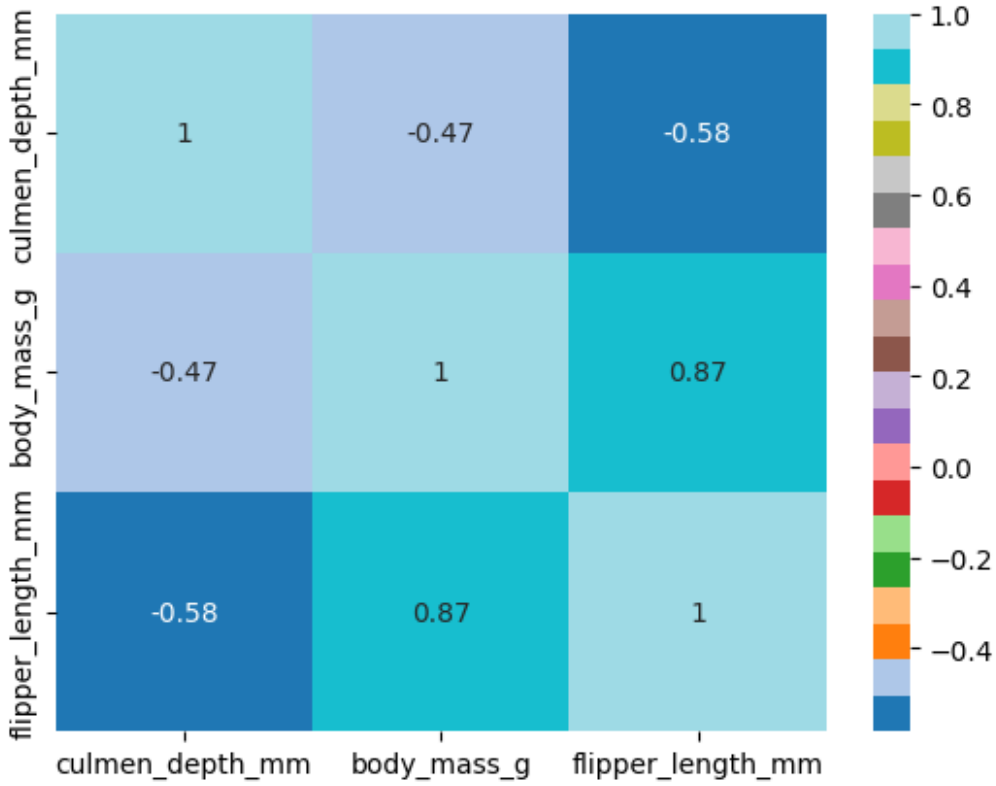

```
#target corelation
p=df.corr()['flipper_length_mm'].sort_values()
p
```

```
<ipython-input-68-b8b8ad29e0dd>:2: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it
will default to False. Select only valid columns or specify the value
of numeric_only to silence this warning.
  p=df.corr()['flipper_length_mm'].sort_values()
```

```
culmen_depth_mm     -0.583832
culmen_length_mm     0.655858
body_mass_g          0.871221
flipper_length_mm    1.000000
Name: flipper_length_mm, dtype: float64
```

```
sea.heatmap(df[['culmen_depth_mm' ,'body_mass_g','flipper_length_mm' ]
].corr(),annot=True,cmap='tab20')
```

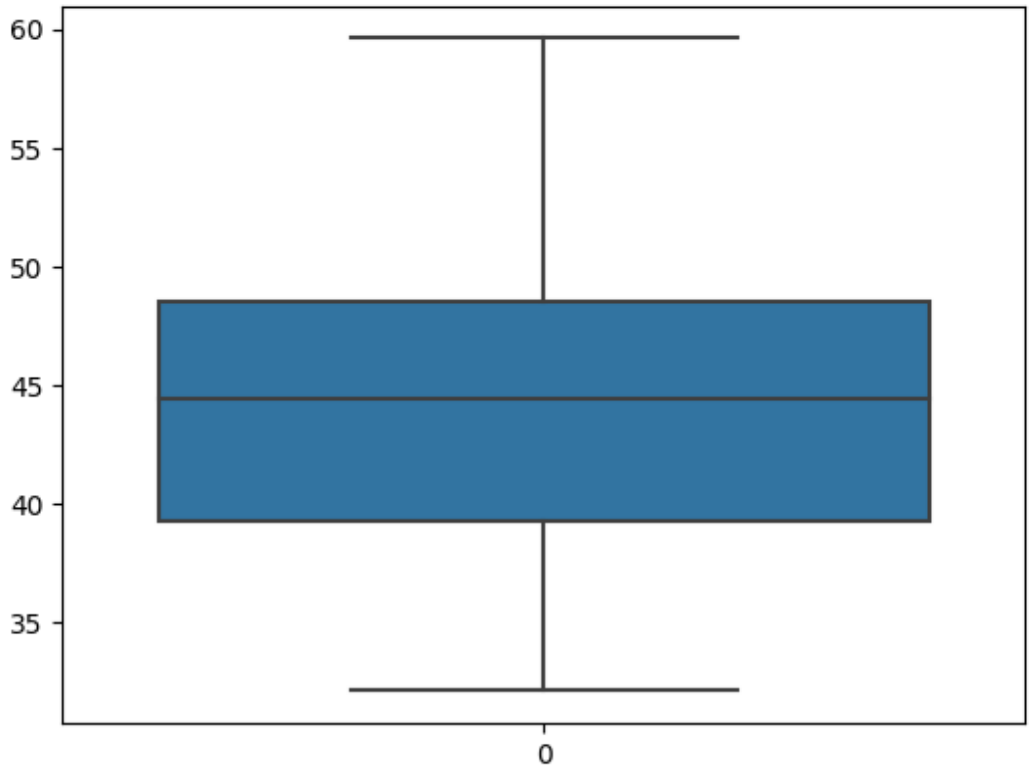
```
<Axes: >
```



Outliers Checking

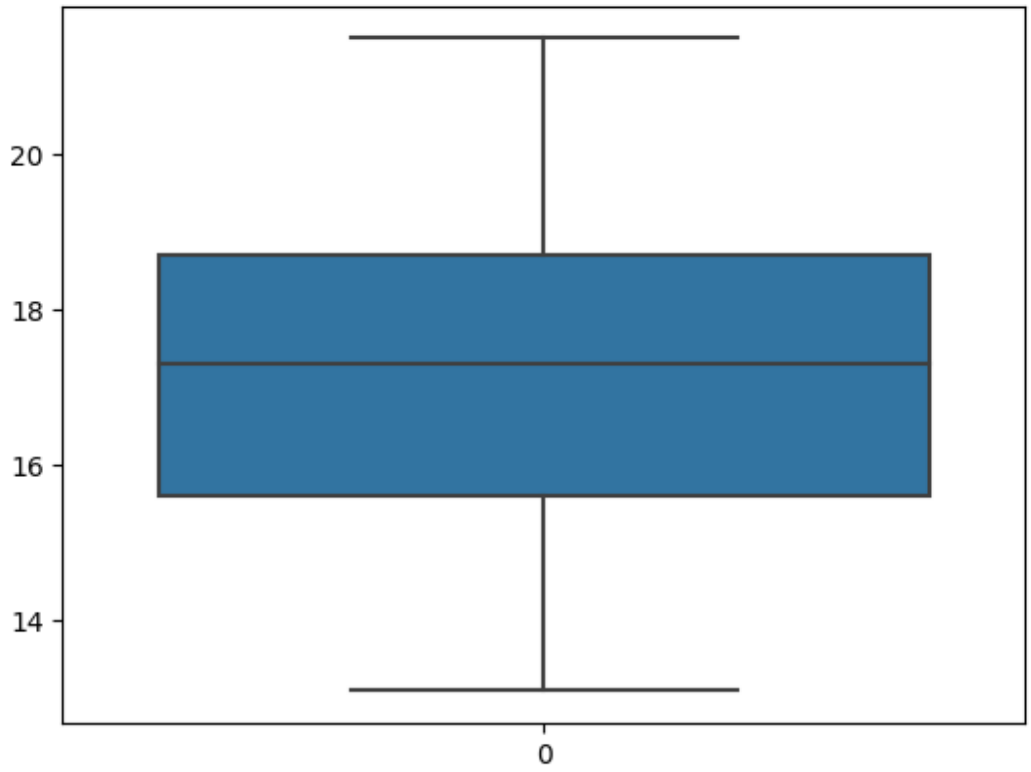
```
sea.boxplot(df.culmen_length_mm)
```

```
<Axes: >
```



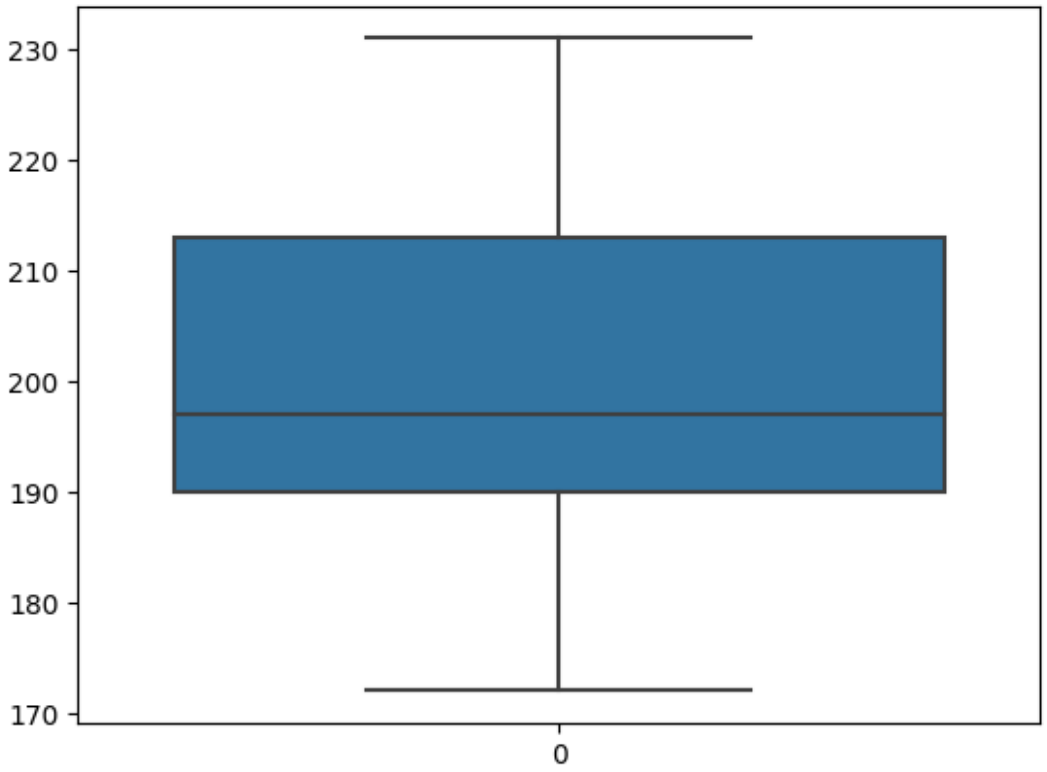

```
sea.boxplot(df.culmen_depth_mm)
```

```
<Axes: >
```



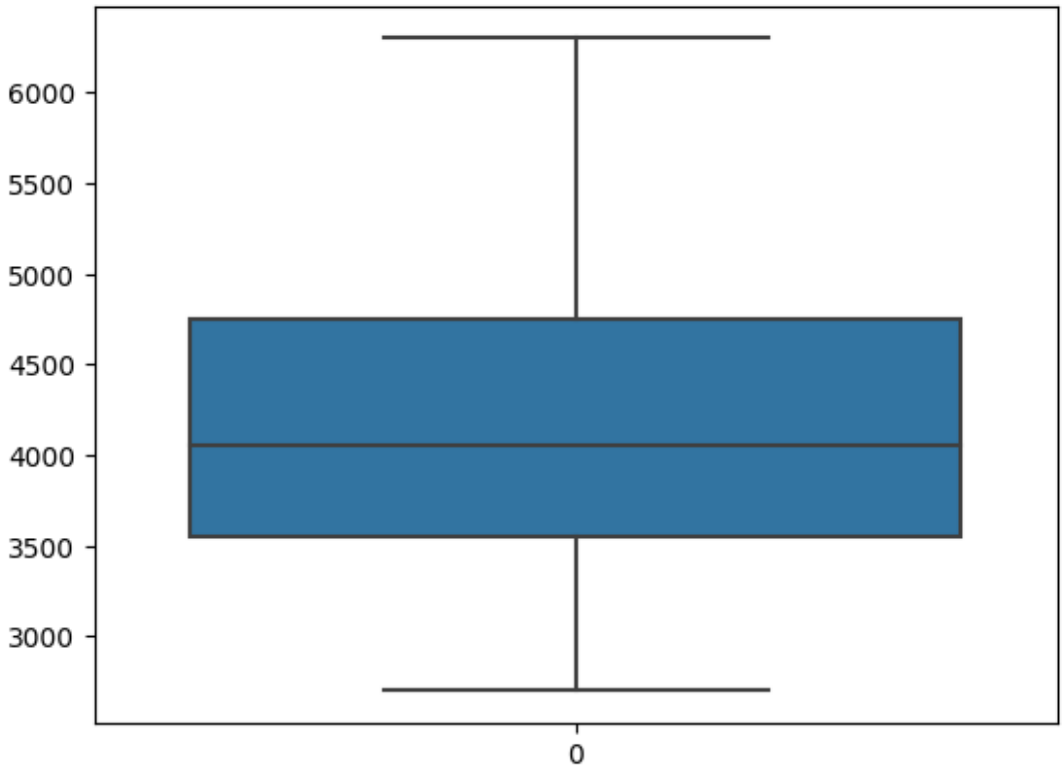

```
sea.boxplot(df.flipper_length_mm)
```

```
<Axes: >
```




```
sea.boxplot(df.body_mass_g)
```

```
<Axes: >
```



Independent and dependent variable split

```
x=df.drop("species",axis=1)

y=df.species
```

Categorical column's encoding

```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
x['island']=le.fit_transform(df['island'])
x['sex']=le.fit_transform(df['sex'])

x.head()
```

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

```
3450.0
```

```
   sex
0    1
1    0
2    0
3    1
4    0
```

Scaling

```
from sklearn.preprocessing import StandardScaler
s=StandardScaler()
x_scaled=pd.DataFrame(s.fit_transform(x),columns=x.columns)
x_scaled.head()
```

```
     island  culmen_length_mm  culmen_depth_mm  flipper_length_mm  \
0  1.844076         -0.887622         0.787289          -1.420541
1  1.844076         -0.814037         0.126114          -1.063485
2  1.844076         -0.666866         0.431272          -0.420786
3  1.844076          0.096581         0.075255          -0.277964
4  1.844076         -1.329133         1.092447          -0.563608

   body_mass_g       sex
0    -0.564625  0.960098
1    -0.502010 -1.041561
2    -1.190773 -1.041561
3    -0.188936  0.960098
4    -0.940314 -1.041561
```

Train-test spliting

```
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=11)

y_train.value_counts()
```

```
Adelie       126
Gentoo        96
Chinstrap     53
Name: species, dtype: int64
```

```
y_train.shape
```

```
(275,)
```

```
x_train.shape
```

```
(275, 6)
```

```
x_test.shape
```

```
(69, 6)
```

```
y_test.shape
```

```
(69,)
```