Experiment 1: Exploratory Data Analysis

Name: Aryan Dali

UID: 2019120016

Class: TE EXTC

The following dataset used is a penguin dataset which has 3 species of penguins Adelie, Chinstrap and Gentoo in three different islands aka Torgoesen, Biscoe and Dream. It has the penguins body dimensions like culmen(beak) length , culmen depth, Flipper length and body mass.

```
import pandas as pd
import seaborn as sns
import·numpy·as·np
```

```
#importing the data and understanding it
data = pd.read_csv("/content/penguins_size.csv")

data.head(10)
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0
3	Adelie	Torgersen	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0
6	Adelie	Torgersen	38.9	17.8	181.0	3625.0
7	Adelie	Torgersen	39.2	19.6	195.0	4675.0
8	Adelie	Torgersen	34.1	18.1	193.0	3475.0

```
data.tail()
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_m
339	Gentoo	Biscoe	NaN	NaN	Na
340	Gentoo	Biscoe	46.8	14.3	215.
341	Gentoo	Biscoe	50.4	15.7	222.
342	Gentoo	Biscoe	45.2	14.8	212

```
data.shape
```

```
(344, 8)
```

```
data = data.drop([data.index[3], data.index[339]])
```

```
data.head()
```

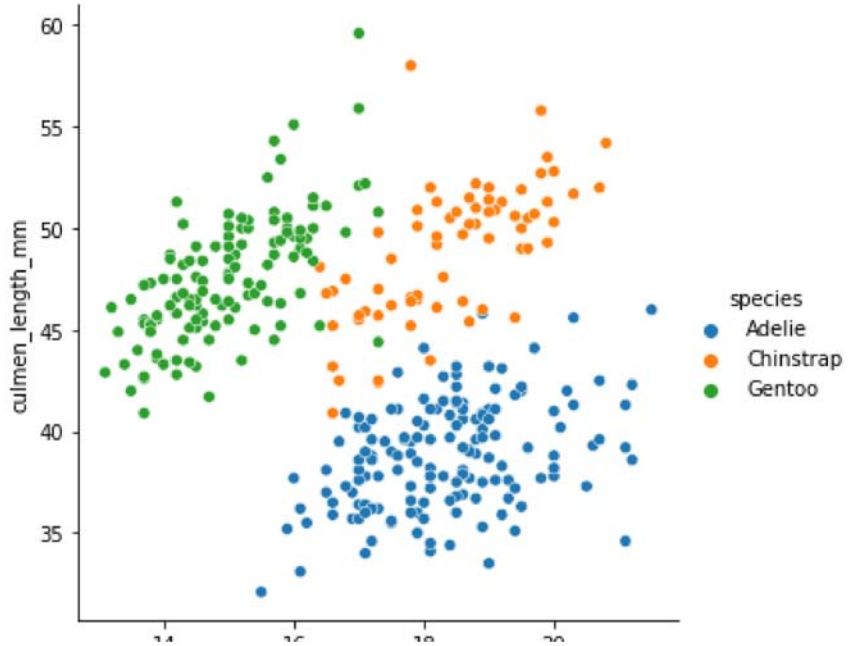
	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_m
0	Adelie	Torgersen	39.1	18.7	181.
1	Adelie	Torgersen	39.5	17.4	186.
2	Adelie	Torgersen	40.3	18.0	195.
4	Adelie	Torgersen	36.7	19.3	193

```
data.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

```
sns.relplot(y='culmen_length_mm', x='culmen_depth_mm',hue = 'species', data = data )
```

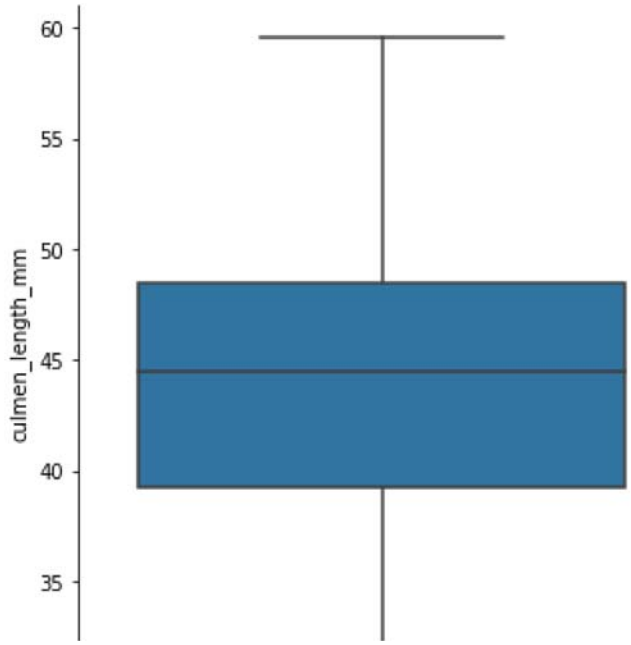
```
<seaborn.axisgrid.FacetGrid at 0x7f14757b1910>
```



From this plot we can see that culmen depth to culmen length height ratio is maximum for adelie balanced for Chinstrap and least for Getoo.

```
sns.catplot(y="culmen_length_mm", data=data, kind = 'box')
```

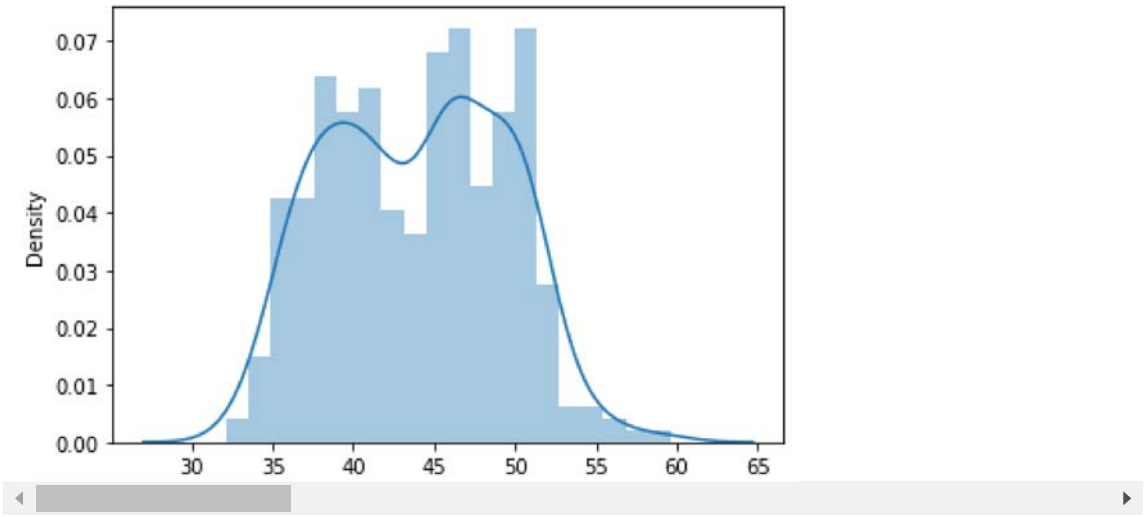
```
<seaborn.axisgrid.FacetGrid at 0x7f1474eadb10>
```



Here we can see the category plot of the culmen length the blue box shows the IQR and line inside the box shows the median.

```
sns.distplot(data['culmen_length_mm'], bins=20)
```

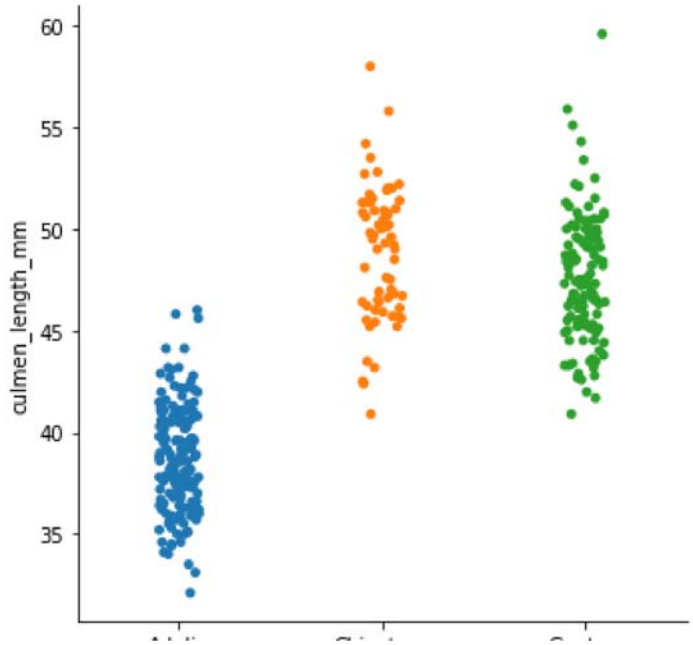
```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: Futu
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f1472073650>
```



The Histogram for the culmen legth is obtained and as we can see it is a bimodal plot.

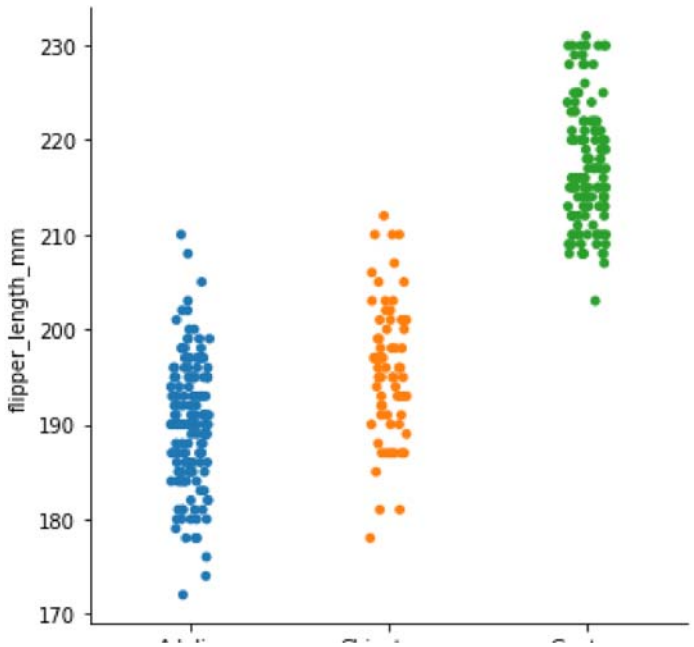
```
sns.catplot(y="culmen_length_mm", x="species", data=data)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f146f732510>
```




```
sns.catplot(y="flipper_length_mm", x="species", data=data)
```

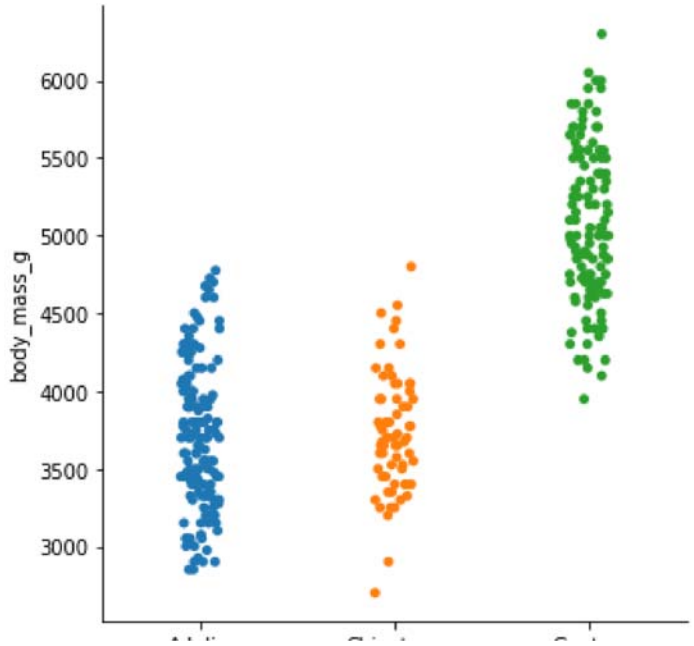
```
<seaborn.axisgrid.FacetGrid at 0x7f1471fee2d0>
```



Here we can see that adelie and chinstrap have almost the same flipper length while gentoo has the a much higher flipper length

```
sns.catplot(y="body_mass_g",·x="species",·data=data)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f146ff329d0>
```



Here we can see that adelie and chinstrap have almost the same body mass while gentoo has the a much higher body mass

Below we are dividing our data sets into three different datasets accrding to the species of the penguin.

```
adelie_data = data[data['species'] == 'Adelie'].reset_index(drop=True)
adelie_data.head()
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_m
0	Adelie	Torgersen	39.1	18.7	181.
1	Adelie	Torgersen	39.5	17.4	186.
2	Adelie	Torgersen	40.3	18.0	195.
3	Adelie	Torgersen	36.7	19.3	193

```
chinstrap_data = data[data['species'] == 'Chinstrap'].reset_index(drop=True)
chinstrap_data.head()
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm
0	Chinstrap	Dream	46.5	17.9	192.0
1	Chinstrap	Dream	50.0	19.5	196.0
2	Chinstrap	Dream	51.3	19.2	193.0
3	Chinstrap	Dream	45.4	18.7	188.0

```
gentoo_data = data[data['species'] == 'Gentoo'].reset_index(drop=True)
gentoo_data.head()
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm
0	Gentoo	Biscoe	46.1	13.2	211.0
1	Gentoo	Biscoe	50.0	16.3	230.0
2	Gentoo	Biscoe	48.7	14.1	210.0
3	Gentoo	Biscoe	50.0	15.2	218.0

```
adelie_data.describe()
```

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
count	151.000000	151.000000	151.000000	151.000000
mean	38.791391	18.346358	189.953642	3700.662252
std	2.663405	1.216650	6.539457	458.566126
min	32.100000	15.500000	172.000000	2850.000000
25%	36.750000	17.500000	186.000000	3350.000000
50%	38.800000	18.400000	190.000000	3700.000000
75%	40.750000	19.000000	195.000000	4000.000000
max	46.000000	21.500000	210.000000	4775.000000

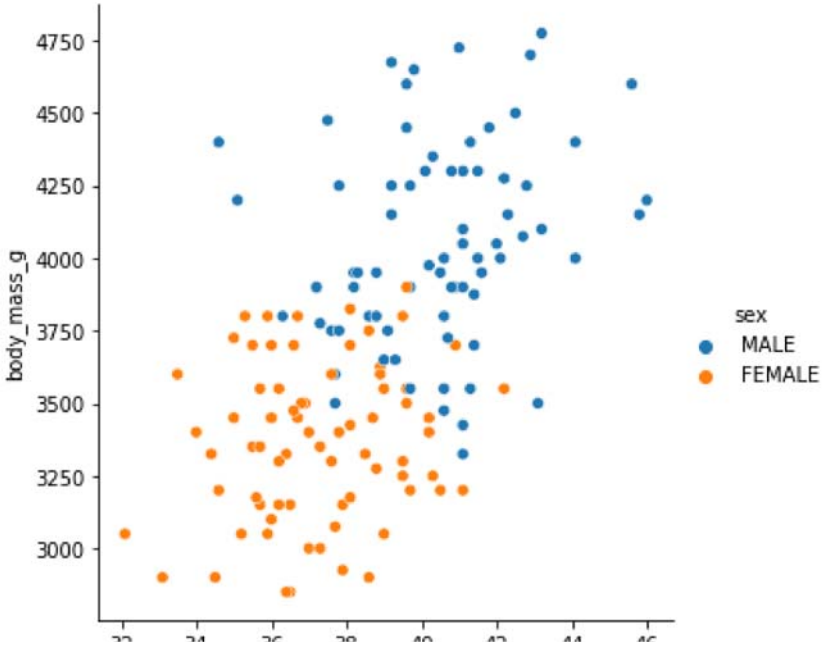
```
chinstrap_data.describe()
```

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
count	68.000000	68.000000	68.000000	68.000000
mean	48.833824	18.420588	195.823529	3733.088235
std	3.339256	1.135395	7.131894	384.335081
min	40.900000	16.400000	178.000000	2700.000000
25%	46.350000	17.500000	191.000000	3487.500000
50%	49.550000	18.450000	196.000000	3700.000000
75%	51.075000	19.400000	201.000000	3950.000000
max	58.000000	20.800000	212.000000	4800.000000

```
gentoo_data.describe()
```

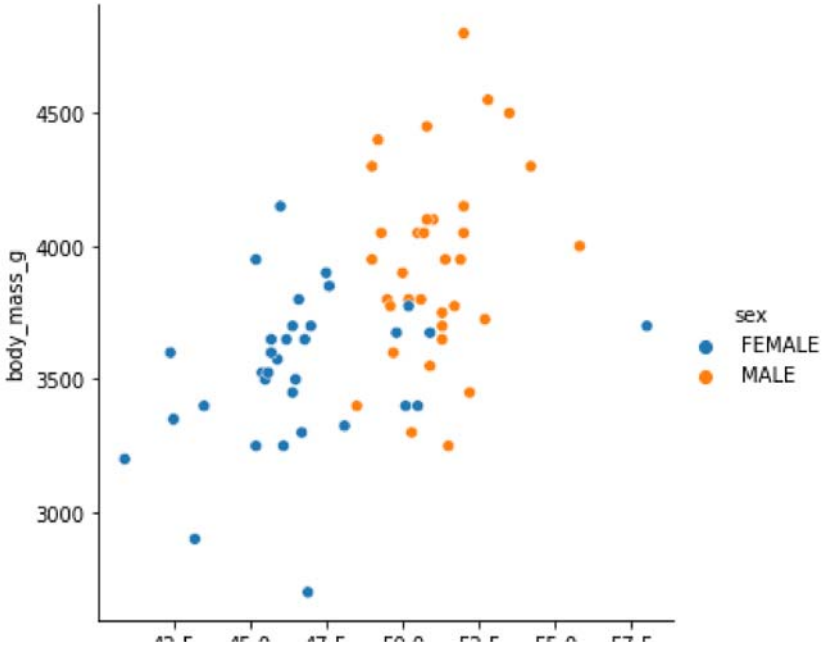
```
sns.relplot(x='culmen_length_mm', y='body_mass_g',hue = 'sex', data = adelie_data )
```

```
<seaborn.axisgrid.FacetGrid at 0x7f146fe2cc50>
```



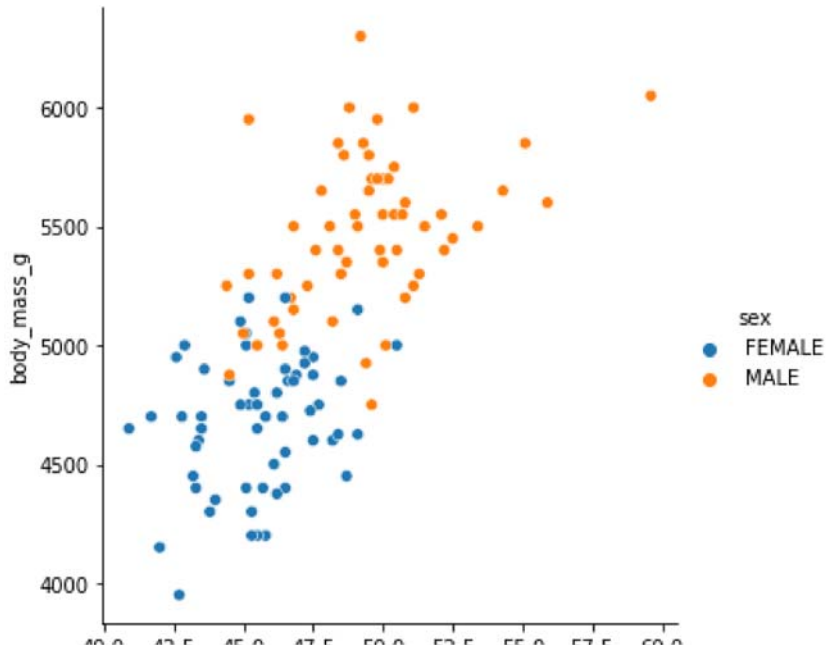

```
sns.relplot(x='culmen_length_mm', y='body_mass_g',hue = 'sex', data = chinstrap_data )
```

```
<seaborn.axisgrid.FacetGrid at 0x7f146fdddf90>
```




```
sns.relplot(x='culmen_length_mm', y='body_mass_g',hue = 'sex', data = gentoo_data )
```

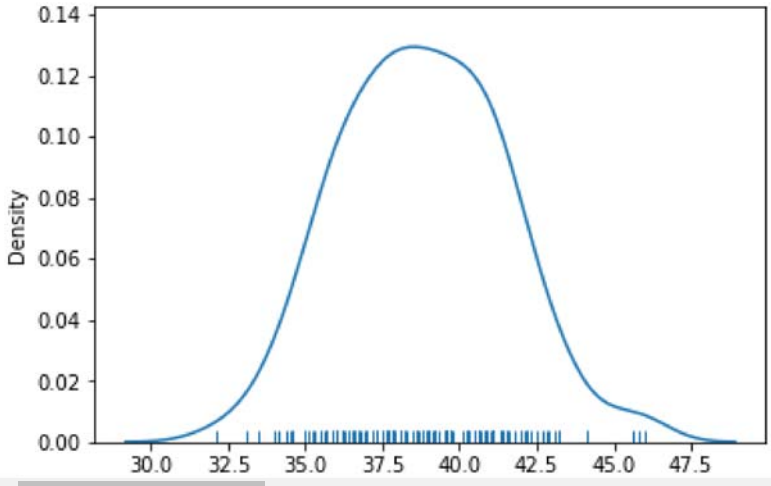
```
<seaborn.axisgrid.FacetGrid at 0x7f146fd580d0>
```



From the above three graphs we can see that Males generally have a larger Body mass than females for all three species.

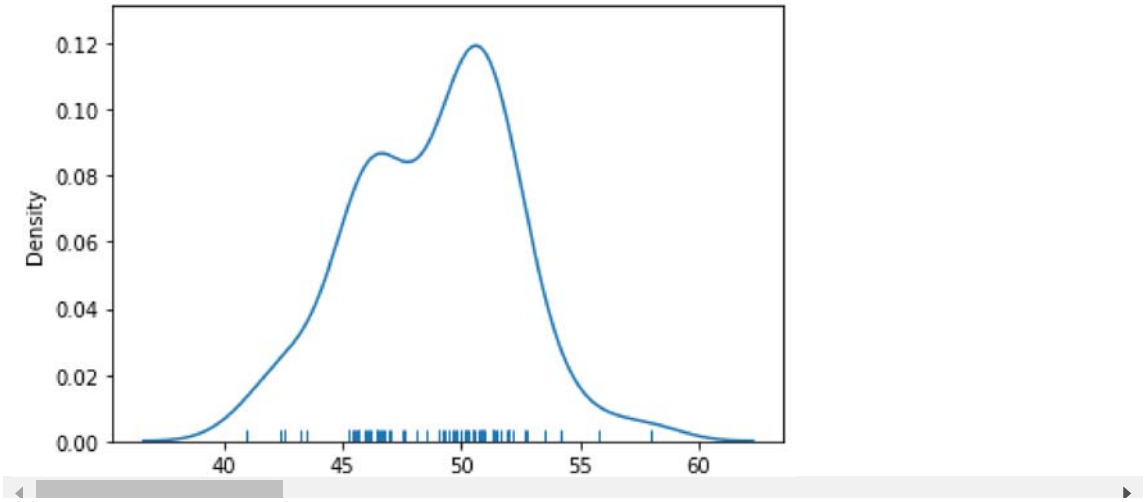
```
sns.distplot(adelie_data['culmen_length_mm'], bins=20, rug=True, hist=False)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: Futu
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2103: Futu
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f146fc5df90>
```



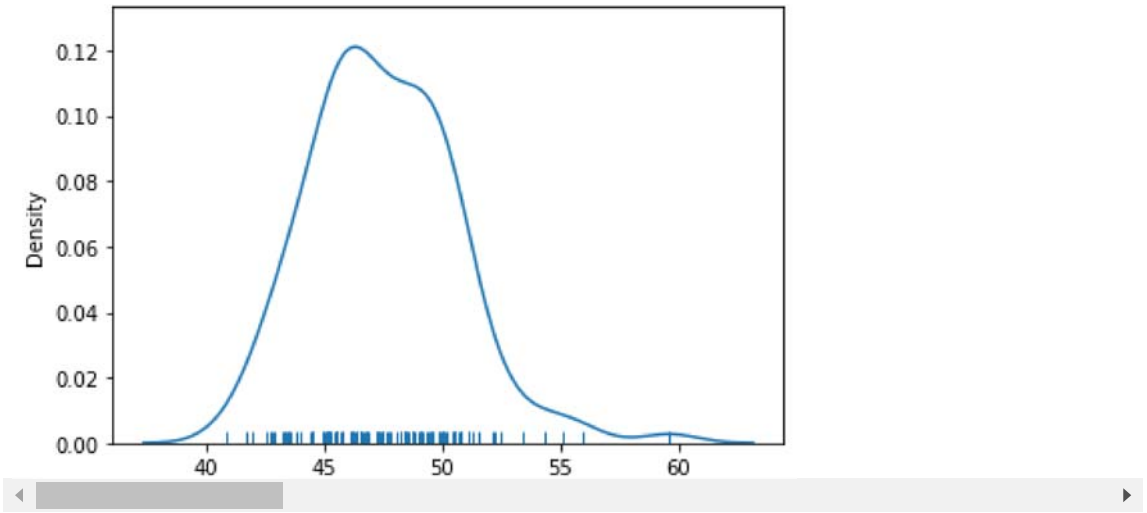

```
sns.distplot(chinstrap_data['culmen_length_mm'], bins=20, rug=True, hist=False)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: Futu
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2103: Futu
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f146fbf5f10>
```



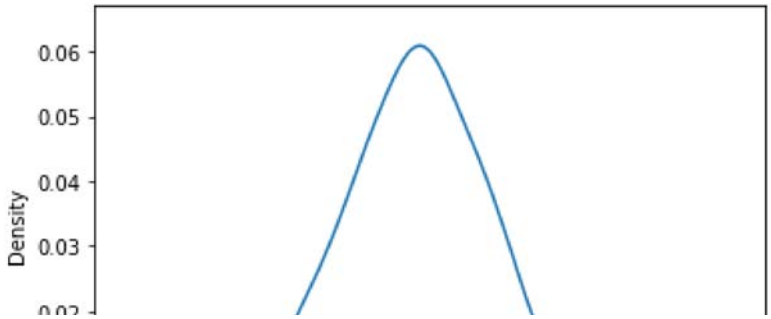

```
sns.distplot(gentoo_data['culmen_length_mm'], bins=20, rug=True, hist=False)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: Futu
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2103: Futu
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f146fb5ff90>
```



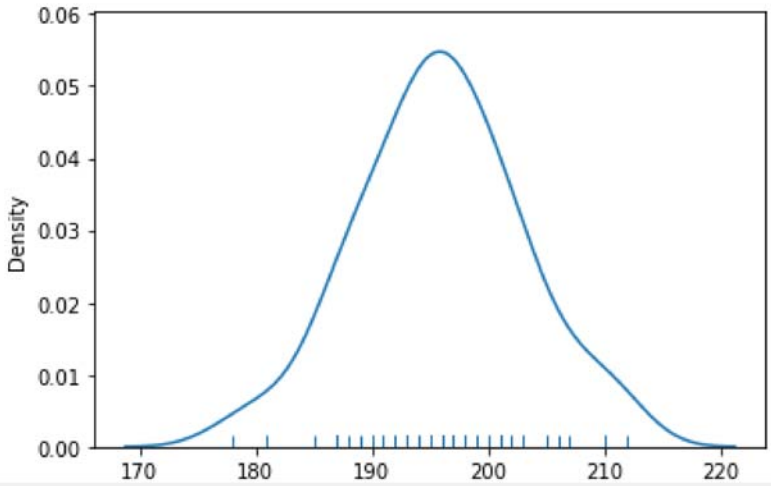

```
sns.distplot(adelie_data['flipper_length_mm'], bins=20, rug=True, hist=False)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: Futu
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2103: Futu
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f1472073910>
```



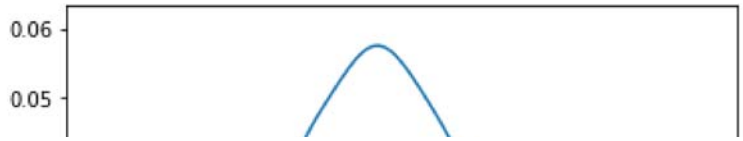

```
sns.distplot(chinstrap_data['flipper_length_mm'], bins=20, rug=True, hist=False)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: Futu
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2103: Futu
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f146fb68cd0>
```




```
sns.distplot(gentoo_data['flipper_length_mm'], bins=20, rug=True, hist=False)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: Futu
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2103: Futu
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f146fac46d0>
```



The Distribution plot for Culmen length and flipper length has been plotted for all three species.


```
gentoo_data.describe()
```

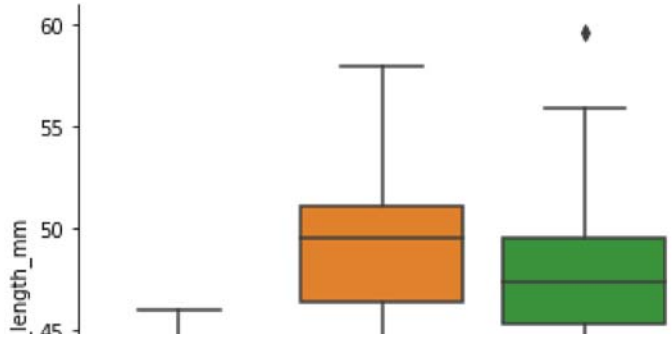
	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
count	123.000000	123.000000	123.000000	123.000000
mean	47.504878	14.982114	217.186992	5076.016260
std	3.081857	0.981220	6.484976	504.116237
min	40.900000	13.100000	203.000000	3950.000000
25%	45.300000	14.200000	212.000000	4700.000000
50%	47.300000	15.000000	216.000000	5000.000000
75%	49.550000	15.700000	221.000000	5500.000000
max	59.600000	17.300000	231.000000	6300.000000

The plot below shows that culmen length for adelie is the lowest then Gentoo and highest is for Chinstrap penguins.

From the below three category plots we can see that the islands do not really affect any body part dimensions .

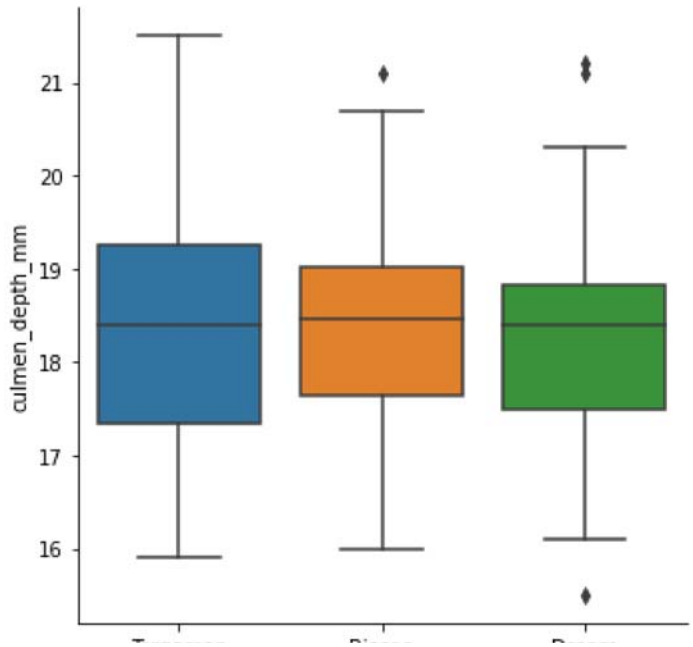
```
sns.catplot(y="culmen_length_mm", x="species", kind = 'box', data=data)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f146fa12ad0>
```




```
sns.catplot(y="culmen_depth_mm", x="island", kind = 'box', data=adelie_data)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f146fa16e10>
```

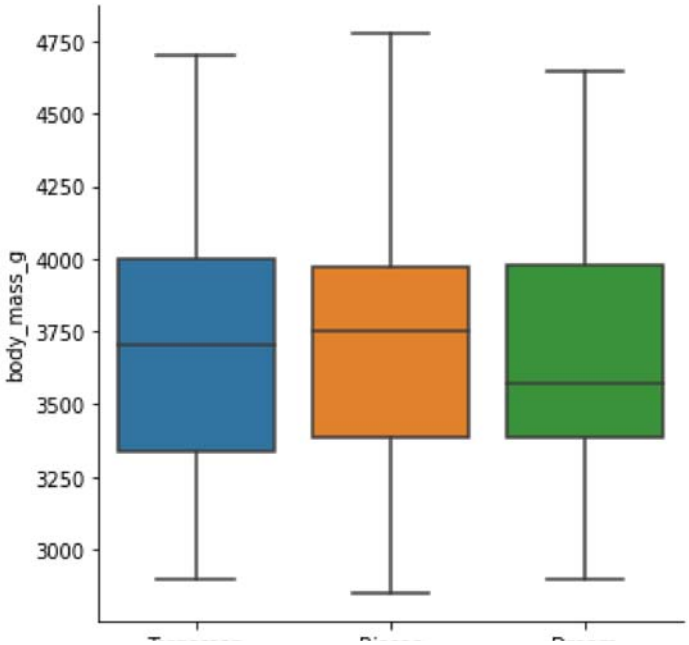



```
sns.catplot(y="flipper_length_mm", x="island", kind = 'box', data=adelie_data)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f146f925190>
```


```
sns.catplot(y="body_mass_g", x="island", kind = 'box', data=adelie_data)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f146f8123d0>
```



The Finding from this experiments were that

1. Culmen depth to culmen length height ratio is maximum for adelie balanced for Chinstrap and least for Getoo.
2. Adelie and chinstrap have almost the same flipper length while gentoo has the a much higher flipper length
3. Adelie and chinstrap have almost the same body mass while gentoo has the a much higher body mass
4. Males generally have a larger Body mass than females for all three species.
5. Culmen length for adelie is the lowest then Gentoo and highest is for Chinstrap penguins.
6. From the below three category plots we can see that the islands do not really affect any body part dimensions .