```
# Assignment_3
# Name: Keshav Goyal
# Roll No: 21BEC2297
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import rcParams
import seaborn as sns
import numpy as np
```

```
df = pd.read_csv('/content/penguins_size.csv') # Importing the dataset
```

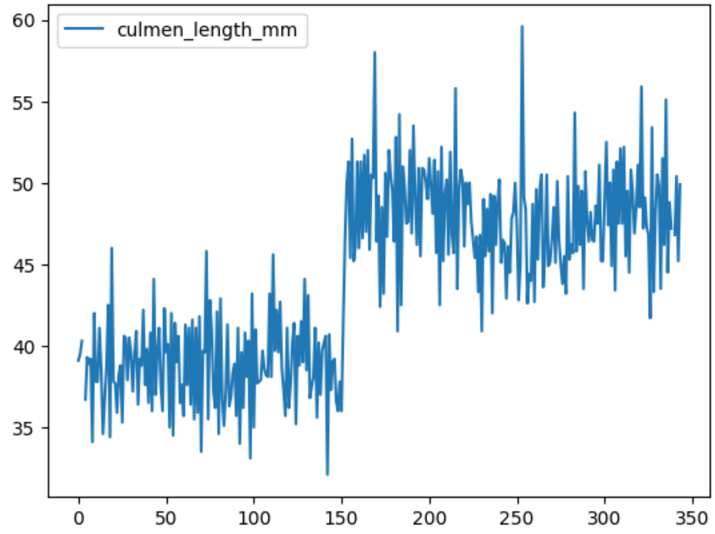
```
df
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_
0	Adelie	Torgersen	39.1	18.7	181.0	3750
1	Adelie	Torgersen	39.5	17.4	186.0	3800
2	Adelie	Torgersen	40.3	18.0	195.0	3250
3	Adelie	Torgersen	NaN	NaN	NaN	Na
4	Adelie	Torgersen	36.7	19.3	193.0	3450
...	...	...	...	...	...	
339	Gentoo	Biscoe	NaN	NaN	NaN	Na
340	Gentoo	Biscoe	46.8	14.3	215.0	4850
341	Gentoo	Biscoe	50.4	15.7	222.0	5750
342	Gentoo	Biscoe	45.2	14.8	212.0	5200
343	Gentoo	Biscoe	49.9	16.1	213.0	5400

344 rows × 7 columns

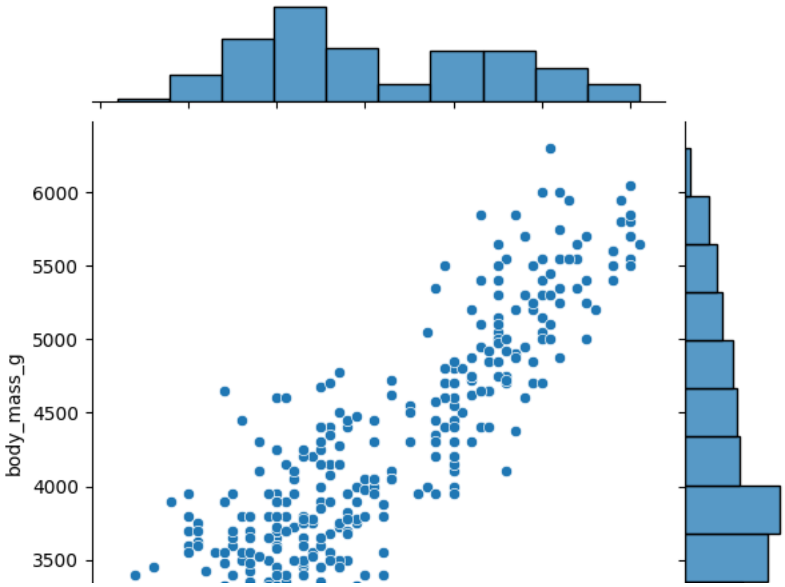
```
# Univariate Analysis
df.loc[:, ['culmen_length_mm']].plot()
```

```
<Axes: >
```



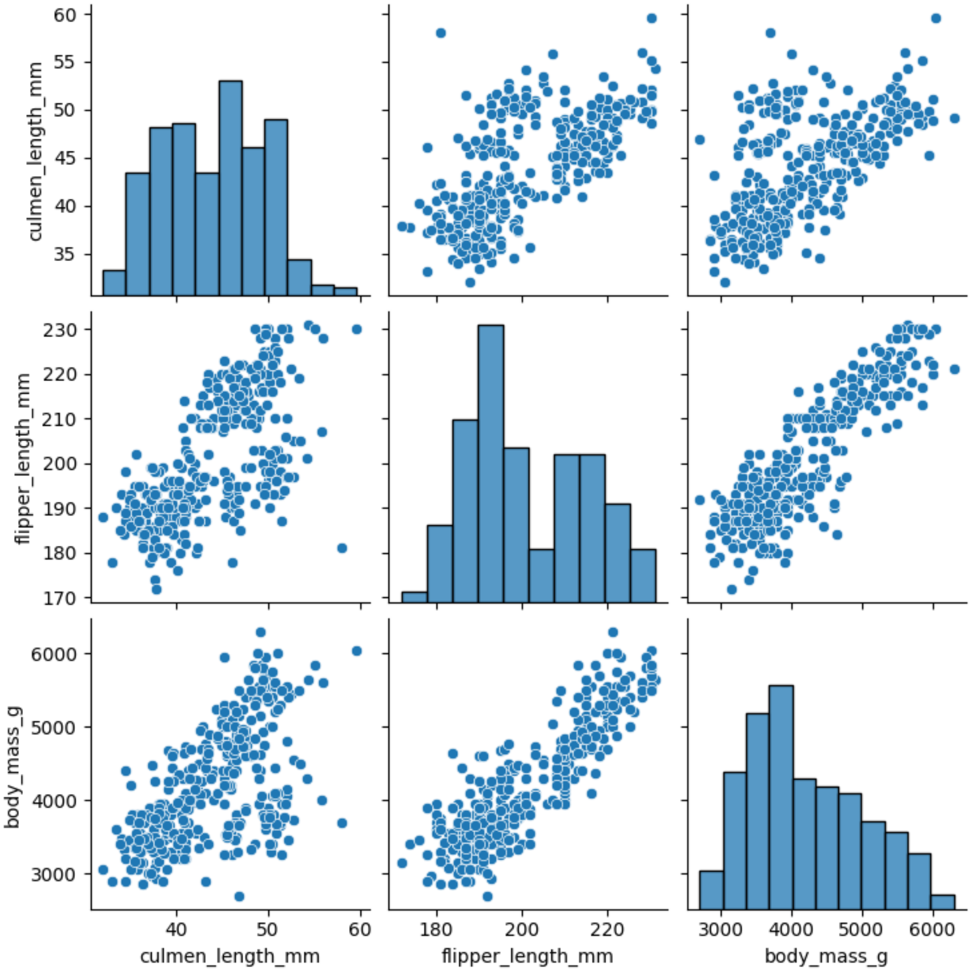
```
# Bi- Variate Analysis
sns.jointplot(x='flipper_length_mm', y='body_mass_g',data=df)
```

```
<seaborn.axisgrid.JointGrid at 0x7f933f4bee90>
```



```
# Multi-Variate Analysis
sns.pairplot(df.loc[:,['culmen_length_mm','flipper_length_mm','body_mass_g']])
```

```
<seaborn.axisgrid.PairGrid at 0x7f933d2fb820>
```



```
df.describe() # Descriptive statistics
```

	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

```
df.isnull().any() # checking is there any null values in our dataset
```

```
species              False
island               False
culmen_length_mm      True
culmen_depth_mm       True
flipper_length_mm     True
body_mass_g           True
sex                   True
dtype: bool
```

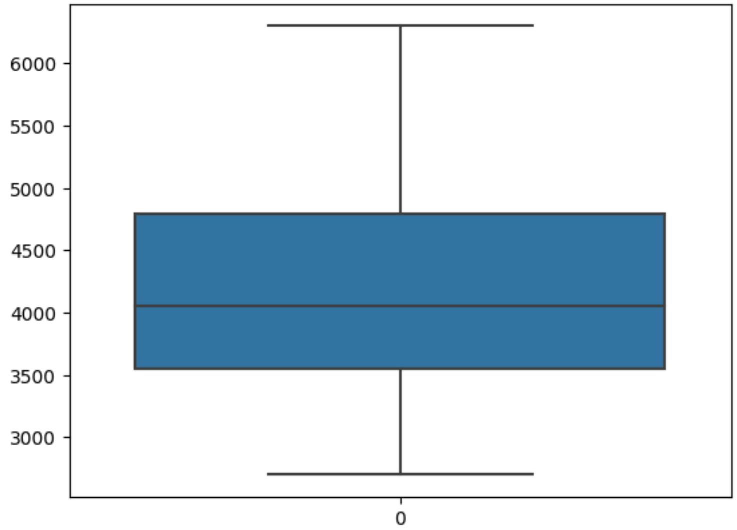
```
# Deleting rows with Null values
df=df.dropna()
df
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0	MALE
...	...	...	...	...	...	...	...
338	Gentoo	Biscoe	47.2	13.7	214.0	4925.0	FEMALE
340	Gentoo	Biscoe	46.8	14.3	215.0	4850.0	FEMALE
341	Gentoo	Biscoe	50.4	15.7	222.0	5750.0	MALE
342	Gentoo	Biscoe	45.2	14.8	212.0	5200.0	FEMALE
343	Gentoo	Biscoe	49.9	16.1	213.0	5400.0	MALE

334 rows × 7 columns

```
# Outlier detection and removal
q1 = df.body_mass_g.quantile(0.25) #Q1
q3 = df.body_mass_g.quantile(0.75) #Q3
IQR = q3-q1
upper_limit = q3+1.5*IQR
lower_limit =q1-1.5*IQR
df.median()
df['body_mass_g'] = np.where(df['body_mass_g']>upper_limit,4050,df['body_mass_g'])
df['body_mass_g'] = np.where(df['body_mass_g']<lower_limit,4050,df['body_mass_g'])
sns.boxplot(df.body_mass_g)
```

```
<ipython-input-34-8d2012ab2219>:7: FutureWarning: The default value of numeric_only in DataFrame.median is deprecated. In a
  df.median()
<Axes: >
```



✓ 0s completed at 11:40 PM