NAVNEEL MONDAL **Reg No: 21BCE2654**

DATE: 14-09-2023

AI ML ASSIGNMENT-3

1. **Download the dataset:** penguins_size.csv is downloaded.

2. **Load The dataset:**

```
import numpy as np
import pandas as pd

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

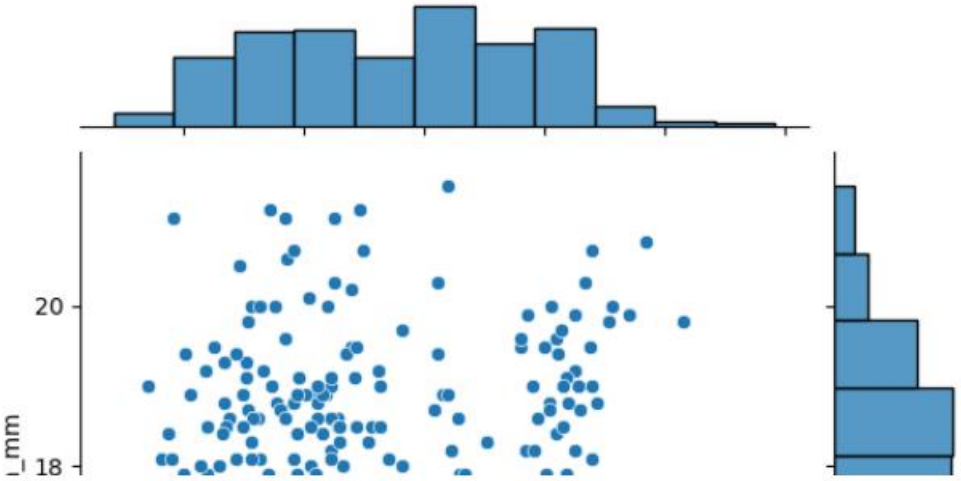
	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_
0	Adelie	Torgersen	39.1	18.7	181.0	
1	Adelie	Torgersen	39.5	17.4	186.0	
2	Adelie	Torgersen	40.3	18.0	195.0	
3	Adelie	Torgersen	NaN	NaN	NaN	
4	Adelie	Torgersen	36.7	19.3	193.0	

3.1. Perform Univariate Analysis

```


from matplotlib import rcParams
import seaborn as sns

sns.distplot(df.body_mass_g)


```

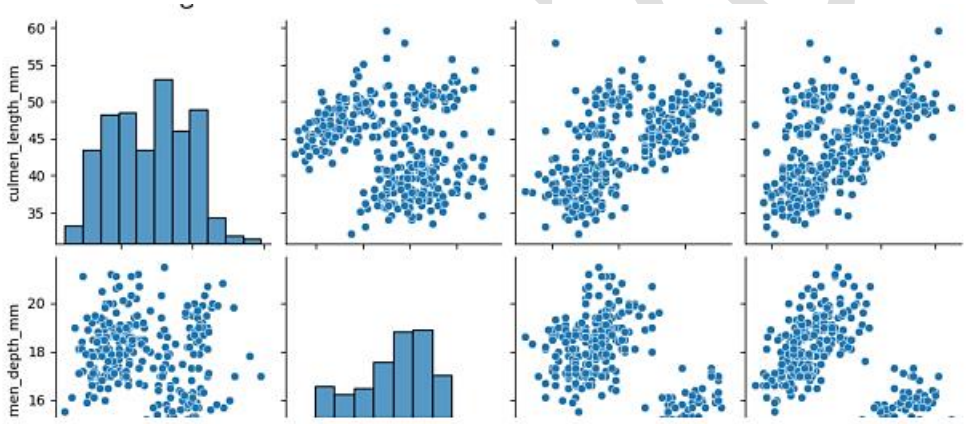
3.2. Perform Bivariate Analysis

```


sns.jointplot(x='culmen_length_mm', y='culmen_depth_mm', data=df)

```



3.3. Perform Multi-Variate Analysis

```
16
17  sns.pairplot(df)
18
19
```



4.Perform descriptive statistics on the dataset.

```
19
20  df.describe()
21
```

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_
count	342.000000	342.000000	342.000000	342

5. Check for Missing values and deal with them.

```
19
20  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
```

```
23
24  df.isnull().sum()
25
```

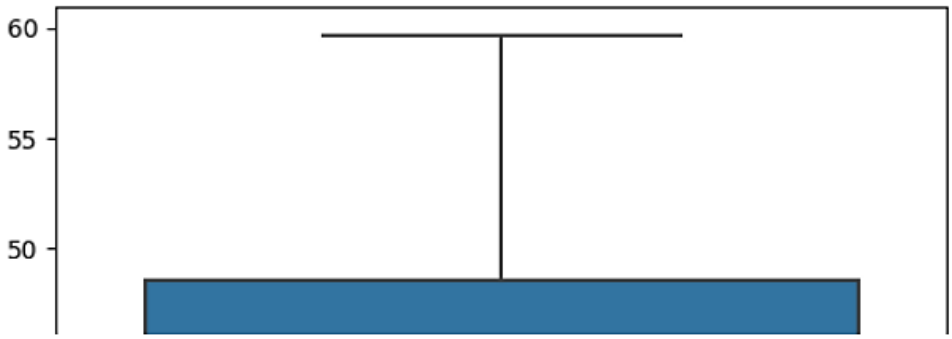
```
species              0
island               0
culmen_length_mm     2
culmen_depth_mm      2
flipper_length_mm    2
body_mass_g          2
sex                 10
dtype: int64
```

```
25
26  df.isnull().any()
27
```

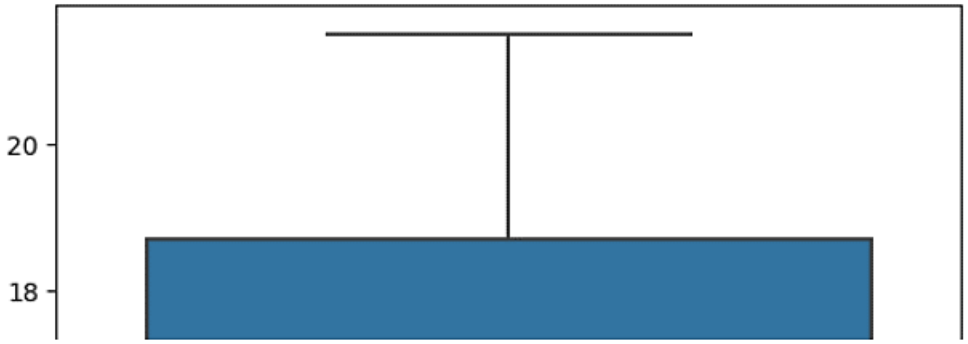
```
species              False
island               False
culmen_length_mm     False
culmen_depth_mm      False
flipper_length_mm    False
body_mass_g          False
sex                  False
dtype: bool
```

6. Find the outliers and replace the outliers.

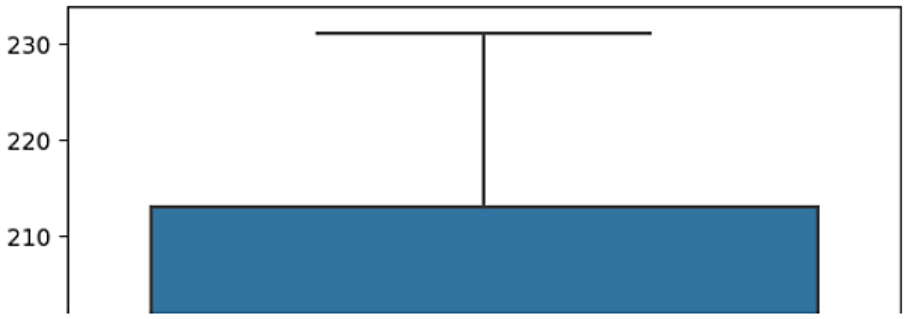
```
27
28  sns.boxplot(df.culmen_length_mm)
29
```



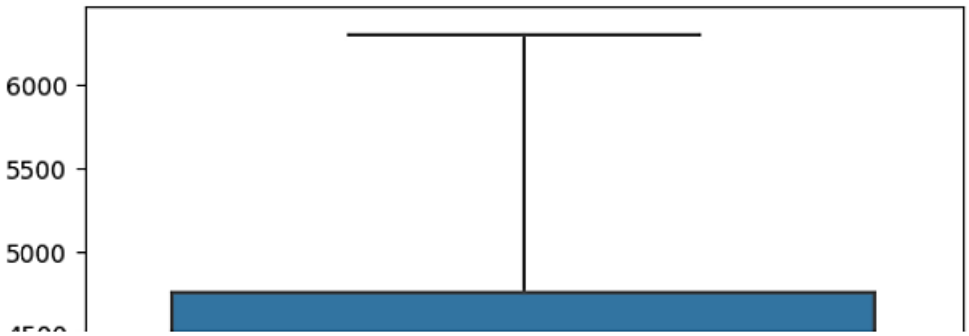

```
29
30  sns.boxplot(df.culmen_depth_mm)
31
```




```
31
32  sns.boxplot(df.flipper_length_mm)
33
```




```
33
34  sns.boxplot(df.body_mass_g)
35
```



Hence there are no outliers in the dataset.

7. Check for Categorical columns and perform encoding.

```
37  from sklearn.preprocessing import LabelEncoder
38  le = LabelEncoder()
39  df['sex'] = le.fit_transform(df['sex'])
40  df['species'] = le.fit_transform(df['species'])
41  df['island'] = le.fit_transform(df['island'])
42  df.head()
43
44
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_ler
0	0	2	39.10	18.7	
1	0	2	39.50	17.4	
2	0	2	40.30	18.0	
3	0	2	44.45	17.3	
4	0	2	36.70	19.3	

8. Check the correlation of independent variables with the target (TARGET IS SPECIES and remaining are independent).

```
45
46  df.corr().species.sort_values(ascending=False)
47
48
```

```
species              1.000000
flipper_length_mm    0.850819
body_mass_g          0.747547
culmen_length_mm     0.728706
sex                 -0.003823
island              -0.635659
culmen_depth_mm     -0.741282
Name: species, dtype: float64
```

9. Split the data into dependent and independent variables.

```
48
49  X=df.drop(columns=['species'],axis=1)
50  X.head()
51
```

	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	b
0	2	39.10	18.7	181.0	
1	2	39.50	17.4	186.0	
2	2	40.30	18.0	195.0	

```
52  Y=df['species']
53  Y.head()
54
55
```

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

10. Scaling the independent data.

```
55
56  from sklearn.preprocessing import MinMaxScaler
57  scale = MinMaxScaler()
58  X_scaled = pd.DataFrame(scale.fit_transform(X),columns=X.columns)
59  X_scaled.head()
60
```

	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	b
0	1.0	0.254545	0.666667	0.152542	
1	1.0	0.269091	0.511905	0.237288	
2	1.0	0.298182	0.583333	0.389831	
3	1.0	0.449091	0.500000	0.423729	
4	1.0	0.167273	0.738095	0.355932	

11. Split the data into training and testing.

```
61
62  from sklearn.model_selection import train_test_split
63  X_train,X_test,Y_train,Y_test = train_test_split(X_scaled,Y,test_size=0.2,random_state=0)
64
```

12. Check the training and testing data shape.

X_train.shape

(275, 6)

X_test.shape

(69, 6)

Y_train.shape

(275,)

Y_test.shape

(69,)