ASSIGNMENT 3

Pratyush Tyagi 21BCE2747 VIT VELLORE

TASK 1

DATASET DOWNLODED

```
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    int64  
 1   island             344 non-null    int64  
 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        344 non-null    float64
 6   sex                344 non-null    int64  
dtypes: float64(4), int64(3)
memory usage: 18.9 KB
```

TASK 2

```
import pandas as pd
df = pd.read_csv('/content/penguins_size.csv')
df
```

	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	F
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	F
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	F
...	...	...	...	...	...	...	
339	Gentoo	Biscoe	NaN	NaN	NaN	NaN	
340	Gentoo	Biscoe	46.8	14.3	215.0	4850.0	F
341	Gentoo	Biscoe	50.4	15.7	222.0	5750.0	
342	Gentoo	Biscoe	45.2	14.8	212.0	5200.0	F
343	Gentoo	Biscoe	49.9	16.1	213.0	5400.0	

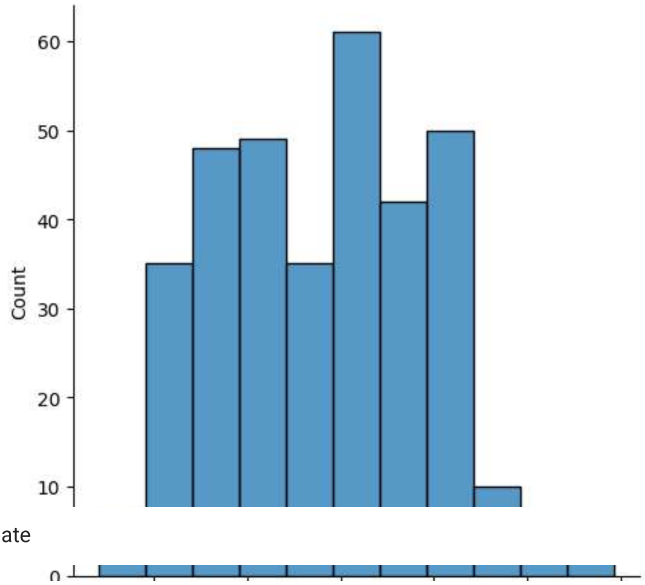
344 rows × 7 columns

TASK 3 Univariate

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

```
sns.displot(df.culmen_length_mm)
```

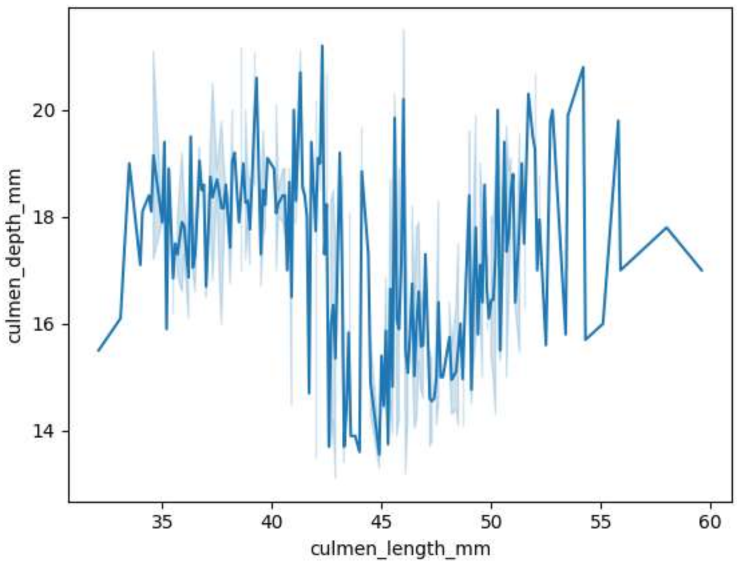
```
<seaborn.axisgrid.FacetGrid at 0x79d72e75ada0>
```



Bivariate

```
sns.lineplot(x = df.culmen_length_mm,y=df.culmen_depth_mm)
```

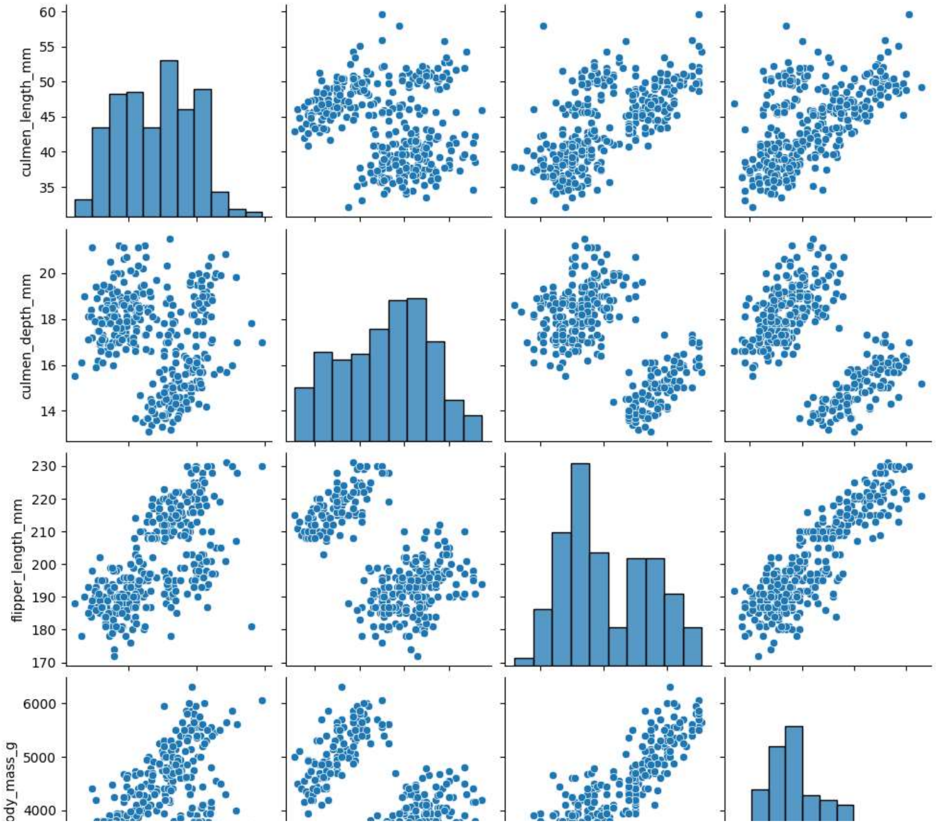
```
<Axes: xlabel='culmen_length_mm', ylabel='culmen_depth_mm'>
```



Multivariate

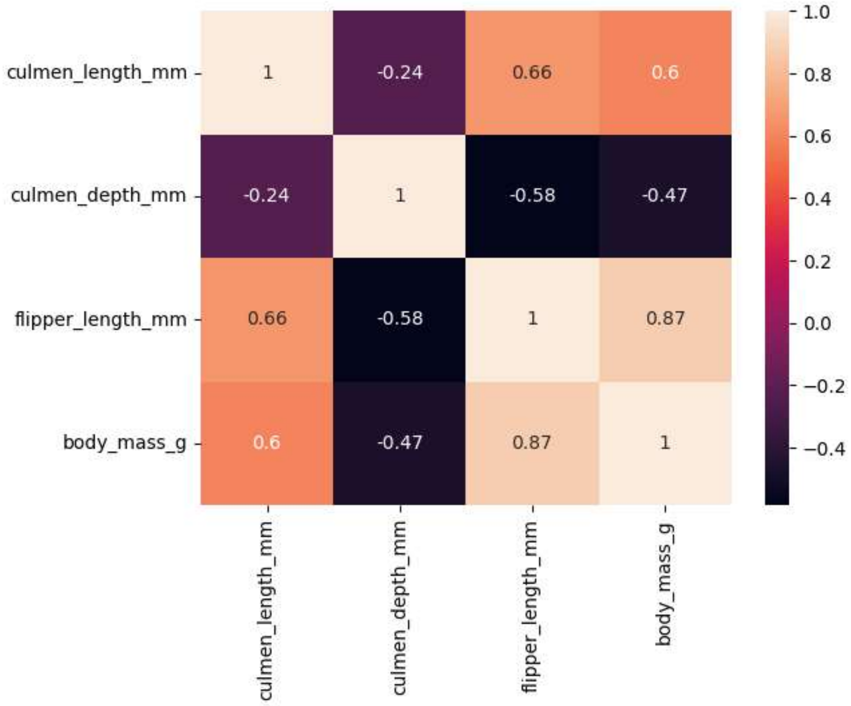
```
sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x79d72f6bae90>
```



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

```
<ipython-input-65-8df7bcac526d>:1: FutureWarning: The default value of numeric_only in DataF
  sns.heatmap(df.corr(),annot=True)
<Axes: >
```



TASK 4

```
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

TASK 5

```
df.isnull().any()
```

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

```
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['culmen_length_mm'].fillna(df['culmen_length_mm'].median(),inplace =True)
```

```
df['culmen_depth_mm'].fillna(df['culmen_depth_mm'].median(),inplace =True)
```

```
df['flipper_length_mm'].fillna(df['flipper_length_mm'].median(),inplace =True)
```

```
df['body_mass_g'].fillna(df['body_mass_g'].median(),inplace =True)
```

```
df['sex'].fillna(df['sex'].mode(),inplace =True)
```

```
df.describe()
```

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
count	344.000000	344.000000	344.000000	344.000000
mean	43.925000	17.152035	200.892442	4200.872093
std	5.443792	1.969060	14.023826	799.696532
min	32.100000	13.100000	172.000000	2700.000000
25%	39.275000	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

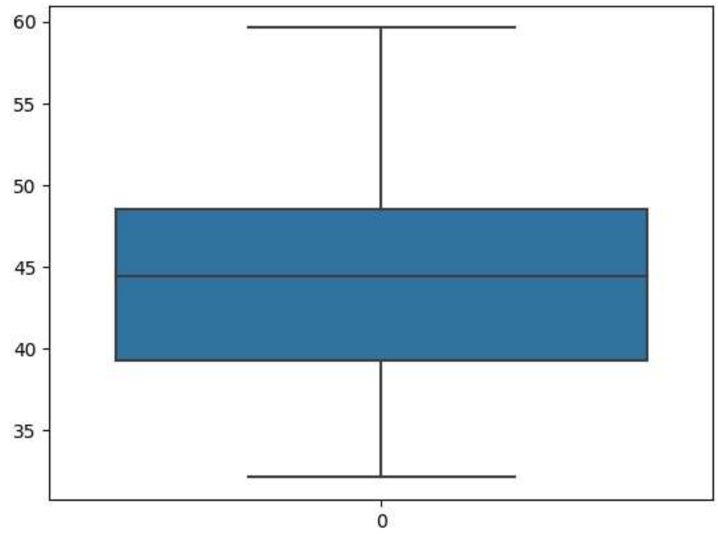
TASK 6 Using Z-score

```
df.head()
```

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

```
sns.boxplot(df.culmen_length_mm)
```

```
<Axes: >
```



```
from scipy import stats
```

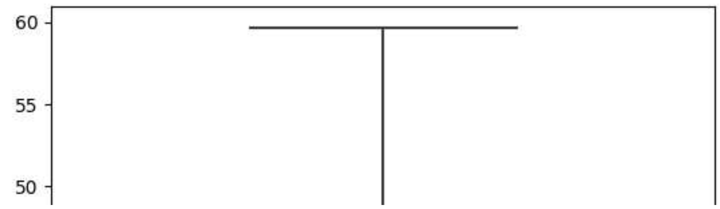
```
culmen_length_mm_zscore = stats.zscore(df.culmen_length_mm)
culmen_length_mm_zscore
```

```
0     -0.887622
1     -0.814037
2     -0.666866
3      0.096581
4     -1.329133
         ...
339    0.096581
340    0.528894
341    1.191161
342    0.234553
343    1.099179
Name: culmen_length_mm, Length: 344, dtype: float64
```

```
df_z = df[np.abs(culmen_length_mm_zscore)<=3]
```

```
sns.boxplot(df_z.culmen_length_mm)
```

```
<Axes: >
```



```
df_z.shape
```

```
(344, 7)
```

TASK 7/8

Label encoding

```
from sklearn.preprocessing import LabelEncoder


le = LabelEncoder()


df.species   = le.fit_transform(df.species  )
df.island  = le.fit_transform(df.island )
df.sex   = le.fit_transform(df.sex  )

df.head()
```

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

One hot encoding

```
df_main = pd.get_dummies(df,columns =['species'])
df_main.head()
```

	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	species_0
0	2	39.10	18.7	181.0	3750.0	2	1
1	2	39.50	17.4	186.0	3800.0	1	1
2	2	40.30	18.0	195.0	3250.0	1	1
3	2	44.45	17.3	197.0	4050.0	3	1
4	2	36.70	19.3	193.0	3450.0	1	1

```
df_main = pd.get_dummies(df,columns =['island'])
df_main.head()
```

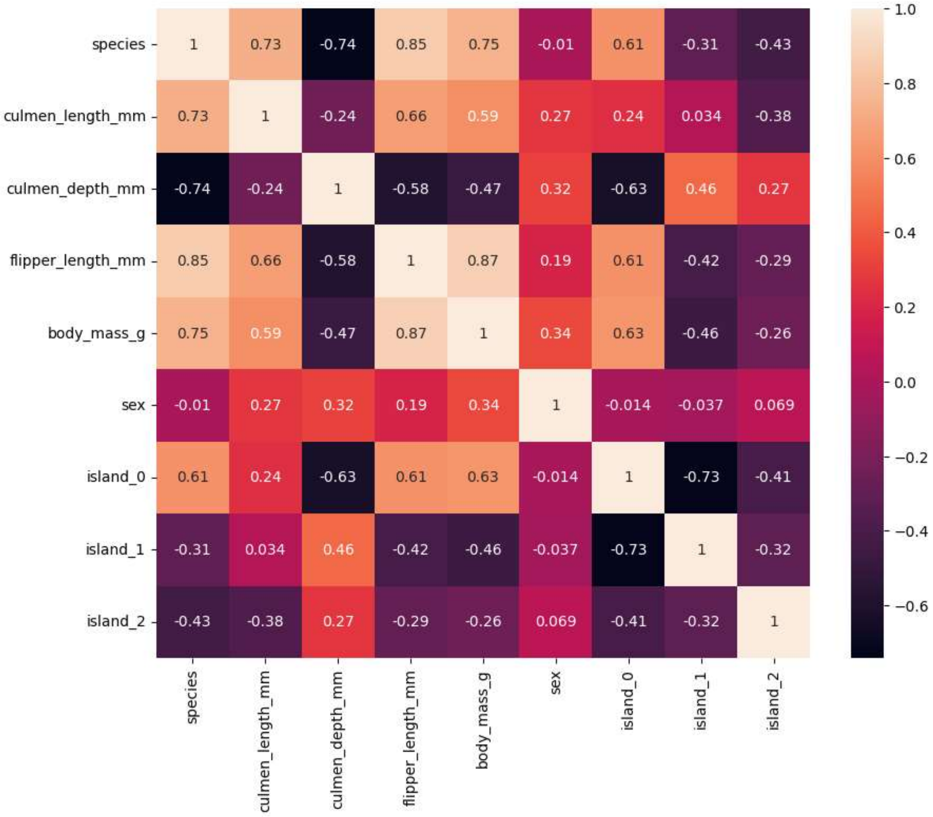
	species	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	island_0
0	0	39.10	18.7	181.0	3750.0	2	0
1	0	39.50	17.4	186.0	3800.0	1	0
2	0	40.30	18.0	195.0	3250.0	1	0
3	0	44.45	17.3	197.0	4050.0	3	0
4	0	36.70	19.3	193.0	3450.0	1	0

```
df_main.corr()
```

	species	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass
species	1.000000	0.728706	-0.741282	0.850819	0.7475
culmen_length_mm	0.728706	1.000000	-0.235000	0.655858	0.5949
culmen_depth_mm	-0.741282	-0.235000	1.000000	-0.583832	-0.4719
flipper_length_mm	0.850819	0.655858	-0.583832	1.000000	0.8712
body_mass_g	0.747547	0.594925	-0.471942	0.871221	1.0000
sex	-0.010379	0.265490	0.317521	0.189194	0.3374
island_0	0.610710	0.238628	-0.630421	0.609679	0.6254
island_1	-0.311589	0.033525	0.455266	-0.419241	-0.4587
island_2	-0.434574	-0.377934	0.269497	-0.288840	-0.2578

```
plt.figure(figsize=(10,8))
sns.heatmap(df_main.corr(),annot =True)
```

```
<Axes: >
```



```
df_main.corr().species.sort_values(ascending=False)
```

```
species              1.000000
flipper_length_mm    0.850819
body_mass_g          0.747547
culmen_length_mm     0.728706
island_0             0.610710
sex                 -0.010379
island_1            -0.311589
```

```
island_2            -0.434574
culmen_depth_mm     -0.741282
Name: species, dtype: float64
```

```
df_main.sex.value_counts()
```

```
2    168
1    165
3     10
0      1
Name: sex, dtype: int64
```

```
df_main.head()
```

	species	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	island_0
0	0	39.10	18.7	181.0	3750.0	2	0
1	0	39.50	17.4	186.0	3800.0	1	0
2	0	40.30	18.0	195.0	3250.0	1	0
3	0	44.45	17.3	197.0	4050.0	3	0
4	0	36.70	19.3	193.0	3450.0	1	0

TASK 9

```
y = df_main['culmen_length_mm']
y
```

```
0      39.10
1      39.50
2      40.30
3      44.45
4      36.70
       ...
339    44.45
340    46.80
341    50.40
342    45.20
343    49.90
Name: culmen_length_mm, Length: 344, dtype: float64
```

```
X =df_main.drop(columns =['culmen_length_mm'],axis =1)
X.head()
```

	species	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	island_0	island_1	island_
0	0	18.7	181.0	3750.0	2	0	0	
1	0	17.4	186.0	3800.0	1	0	0	
2	0	18.0	195.0	3250.0	1	0	0	
3	0	17.3	197.0	4050.0	3	0	0	
4	0	19.3	193.0	3450.0	1	0	0	

TASK 10

```
from sklearn.preprocessing import MinMaxScaler
scale =MinMaxScaler()
```

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

	species	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	island_0	island_1	is
0	0.0	0.666667	0.152542	0.291667	0.666667	0.0	0.0	
1	0.0	0.511905	0.237288	0.305556	0.333333	0.0	0.0	

TASK 11

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X_scaled,y,test_size=0.3,random_state=10)
```

```
X_train.shape
```

```
(240, 8)
```

```
X_train.head()
```

	species	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	island_0	island_1
258	1.0	0.059524	0.610169	0.458333	0.333333	1.0	0.0
332	1.0	0.250000	0.694915	0.541667	0.333333	1.0	0.0
121	0.0	0.797619	0.440678	0.222222	0.666667	0.0	0.0
61	0.0	0.952381	0.389831	0.472222	0.666667	1.0	0.0
70	0.0	0.702381	0.305085	0.250000	0.333333	0.0	0.0

TASK 12

```
y_train.shape
```

```
(240,)
```

```
X_test.shape
```

```
(104, 8)
```

✓ 0s completed at 11:51 PM ● ✕