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

#the data set "Penguins" has been downloaded from kaggle and uploaded as a file on the python notebook
```

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

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass
0	Adelie	Torgersen	39.1	18.7	181.0	375
1	Adelie	Torgersen	39.5	17.4	186.0	380
2	Adelie	Torgersen	40.3	18.0	195.0	325
3	Adelie	Torgersen	NaN	NaN	NaN	N
4	Adelie	Torgersen	36.7	19.3	193.0	345

```
df.shape
```

```
(344, 7)
```

```
df.info
```

```
<bound method DataFrame.info of     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
..      ...        ...               ...              ...                ...
339  Gentoo     Biscoe               NaN              NaN                NaN
340  Gentoo     Biscoe              46.8             14.3              215.0
341  Gentoo     Biscoe              50.4             15.7              222.0
342  Gentoo     Biscoe              45.2             14.8              212.0
343  Gentoo     Biscoe              49.9             16.1              213.0

     body_mass_g     sex
0         3750.0    MALE
1         3800.0  FEMALE
2         3250.0  FEMALE
3            NaN     NaN
4         3450.0  FEMALE
..           ...     ...
339          NaN     NaN
340       4850.0  FEMALE
341       5750.0    MALE
342       5200.0  FEMALE
343       5400.0    MALE

[344 rows x 7 columns]>
```

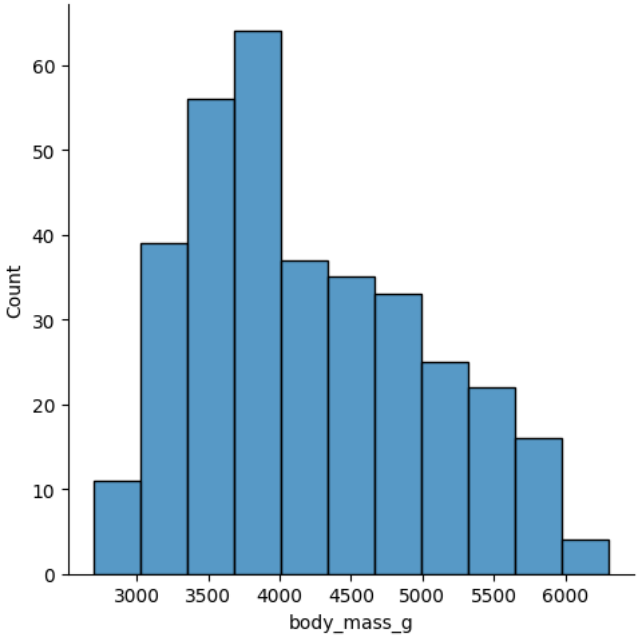
Univariate Analysis

```
#analysis of parameter 'culmen_length_mm'
sns.distplot(df['culmen_length_mm'])
```

```
<ipython-input-7-c5f9aeb44ede>:2: 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

  sns.distplot(df['culmen_length_mm'])
<Axes: xlabel='culmen_length_mm', ylabel='Density'>
```


```
#analysis of parameter 'body_mass_g'
sns.displot(df['body_mass_g'])
```

```
<seaborn.axisgrid.FacetGrid at 0x79a42db1da20>
```



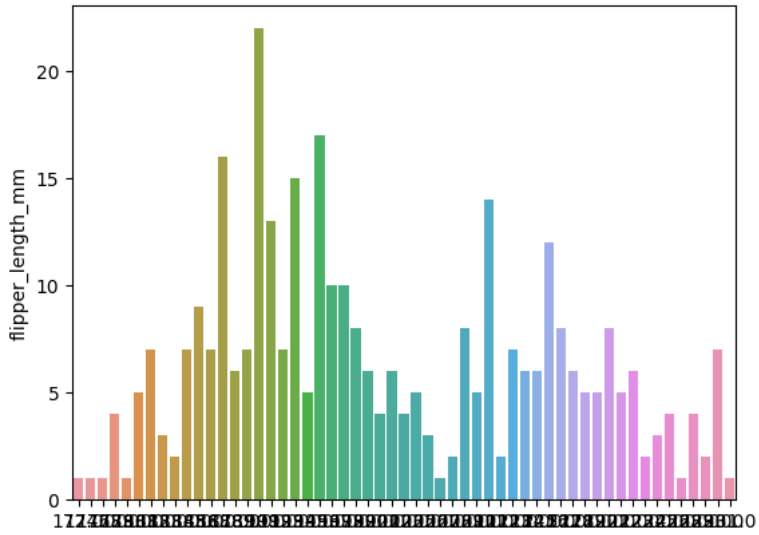

```
df['island'].value_counts()
```

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

```
#analysis of parameter 'island'
plt.pie(df['island'].value_counts(), labels=['Biscoe', 'Dream', 'Torgersen'], autopct='%1.1f%%', shadow=True, colors=['b', 'm', 'c'])
plt.title('Univariate analysis')
plt.show()
```

```
#analysis of parameter 'flipper_length_mm'
sns.barplot(x=df['flipper_length_mm'].value_counts().index, y= df['flipper_length_mm'].value_counts())
```

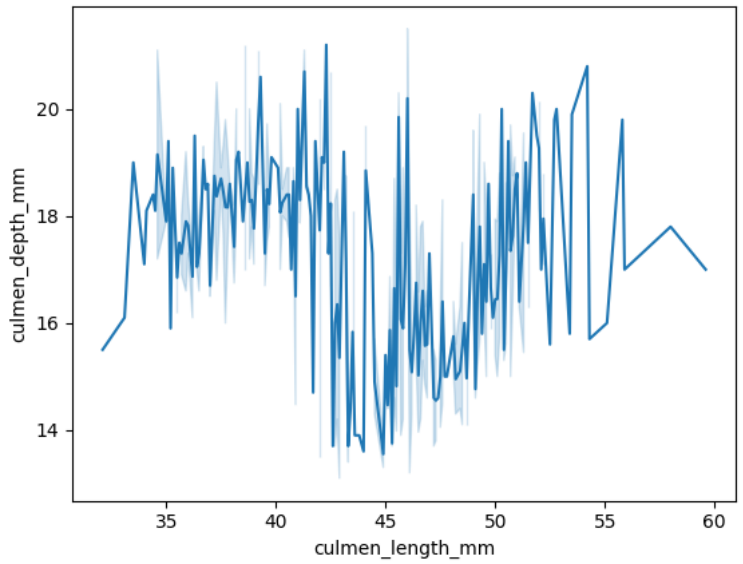
```
<Axes: ylabel='flipper_length_mm'>
```



Bivariate analysis

```
#analysis between parameters 'culmen_length_mm' and 'culmen_depth_mm'
sns.lineplot(x=df['culmen_length_mm'], y=df['culmen_depth_mm'])
```

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



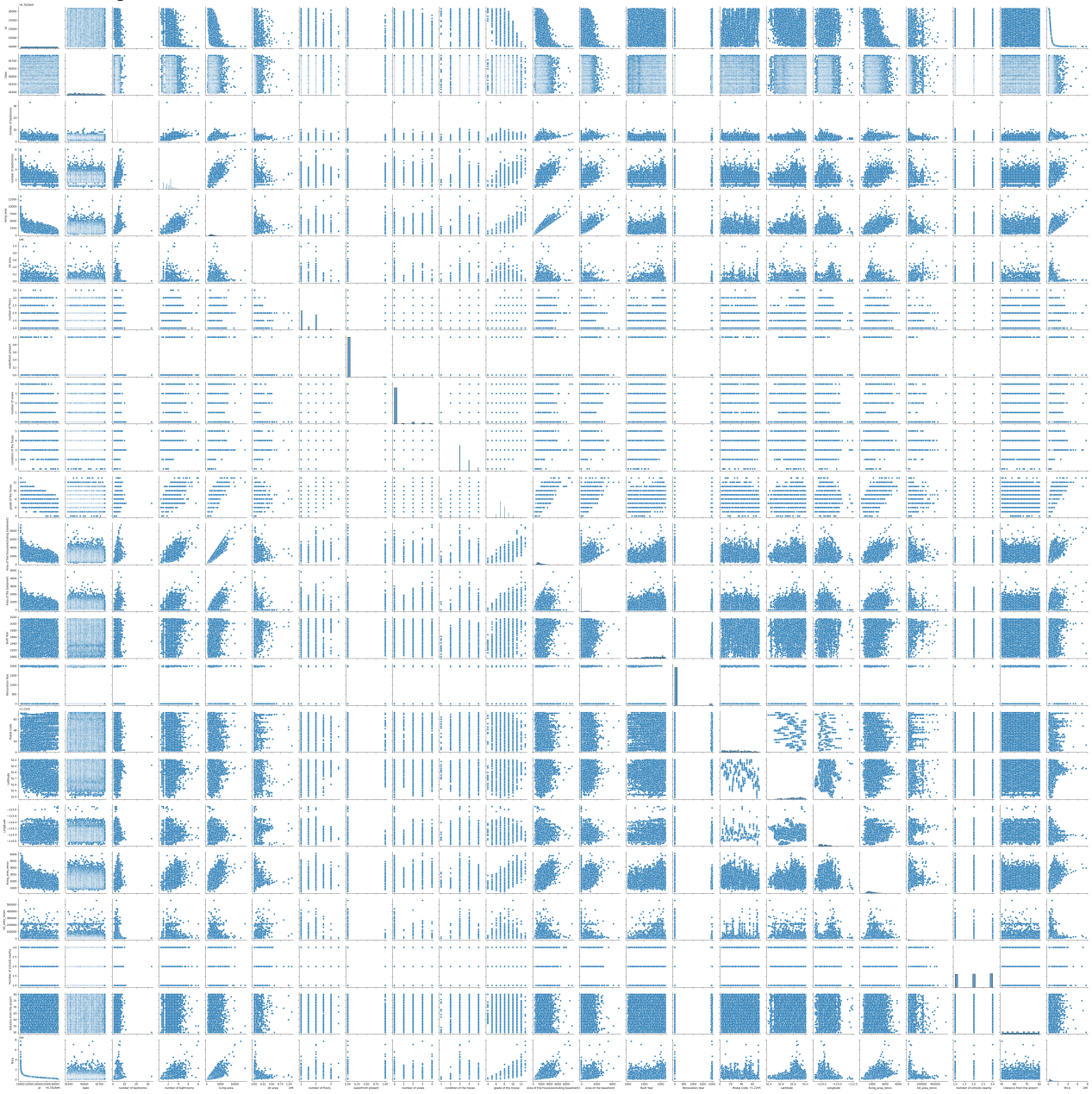

```
#analysis between parameters 'flipper_length_mm' and 'body_mass_g'
sns.scatterplot(x = df['flipper_length_mm'],y=df['body_mass_g'])
```

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

Multivariate Analysis

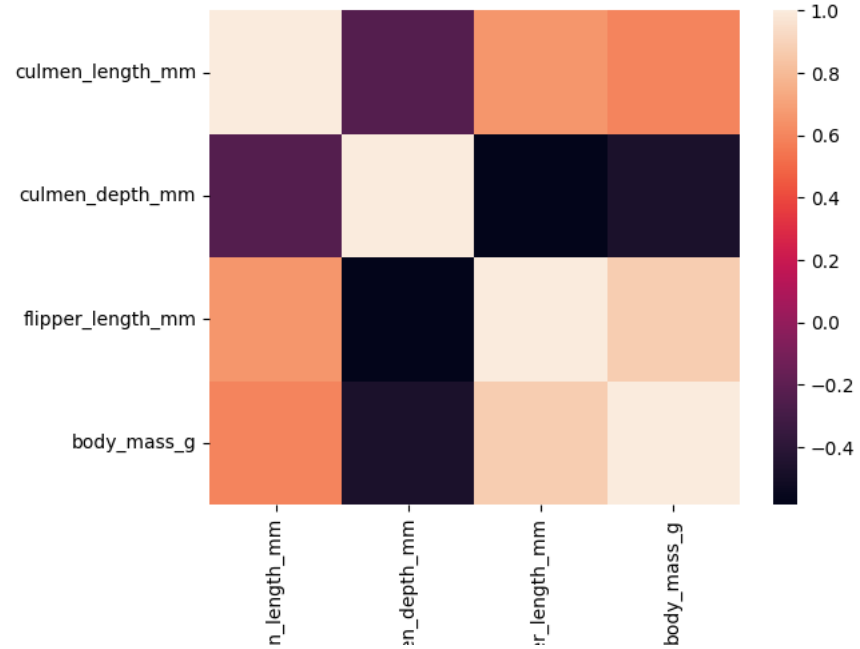
```
sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x7d90f1d8f250>
```



```
sns.heatmap(df.corr())
```

```
<ipython-input-14-aa4f4450a243>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a f
  sns.heatmap(df.corr())
<Axes: >
```




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

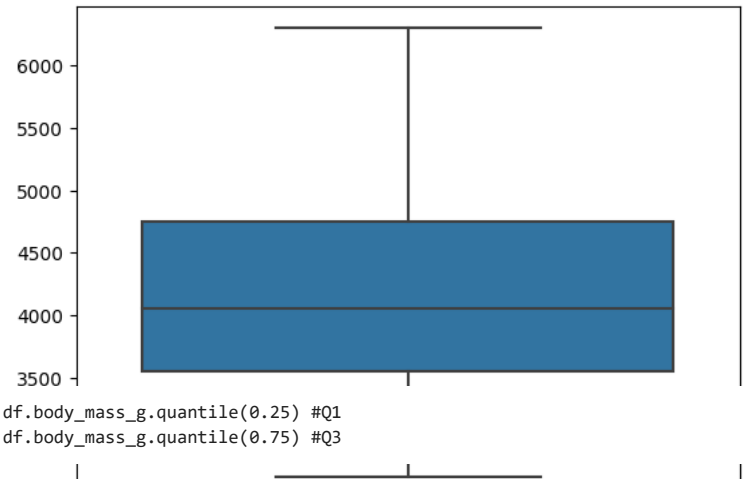
```
species             False
island              False
culmen_length_mm     True
culmen_depth_mm      True
flipper_length_mm    True
body_mass_g          True
sex                  True
dtype: bool
```

```
df.head()
```

	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
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE

```
sns.boxplot(df.body_mass_g)
```

```
<Axes: >
```




```
q1 = df.body_mass_g.quantile(0.25) #Q1
q3 = df.body_mass_g.quantile(0.75) #Q3
```

```
print(q1)
print(q3)
```

```
3550.0
4750.0
```

```
IQR = q3-q1
```

```
IQR
```

```
1200.0
```

```
upper_limit = q3+1.5*IQR
```

```
upper_limit
```

```
6550.0
```

```
lower_limit =q1-1.5*IQR
```

```
lower_limit
```

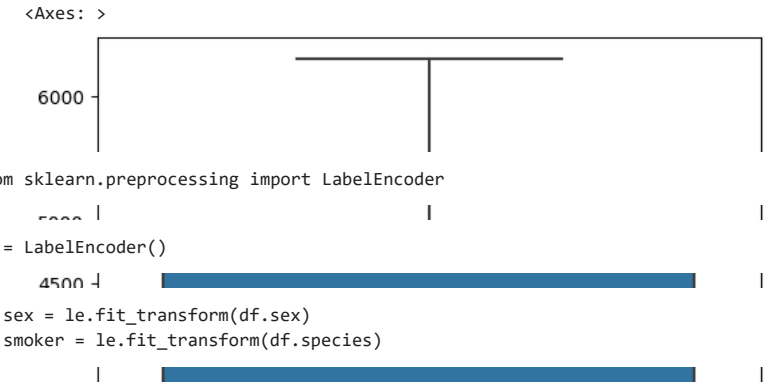
```
1750.0
```

```
df.median()
```

```
<ipython-input-17-6d467abf240d>:1: FutureWarning: The default value of numeric_only in DataFrame.median is deprecated. In a future
  df.median()
culmen_length_mm       44.45
culmen_depth_mm        17.30
flipper_length_mm     197.00
body_mass_g          4050.00
dtype: float64
```

```
df['body_mass_g'] = np.where(df['body_mass_g']>upper_limit,30,df['body_mass_g'])
```

```
sns.boxplot(df.body_mass_g)
```




```
from sklearn.preprocessing import LabelEncoder
```

```
le = LabelEncoder()
```

```
df.sex = le.fit_transform(df.sex)
df.smoker = le.fit_transform(df.species)
```

```
df.head()
```

	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	2
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	1
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	1
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	3
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	1

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

	species	island	culmen_length_mm	culmen_depth_mm	body_mass_g	sex	flipper_length_mm_172.0	flipper_length_mm_174.0
0	Adelie	Torgersen	39.1	18.7	3750.0	2	0	0
1	Adelie	Torgersen	39.5	17.4	3800.0	1	0	0
2	Adelie	Torgersen	40.3	18.0	3250.0	1	0	0
3	Adelie	Torgersen	NaN	NaN	NaN	3	0	0
4	Adelie	Torgersen	36.7	19.3	3450.0	1	0	0

5 rows × 61 columns

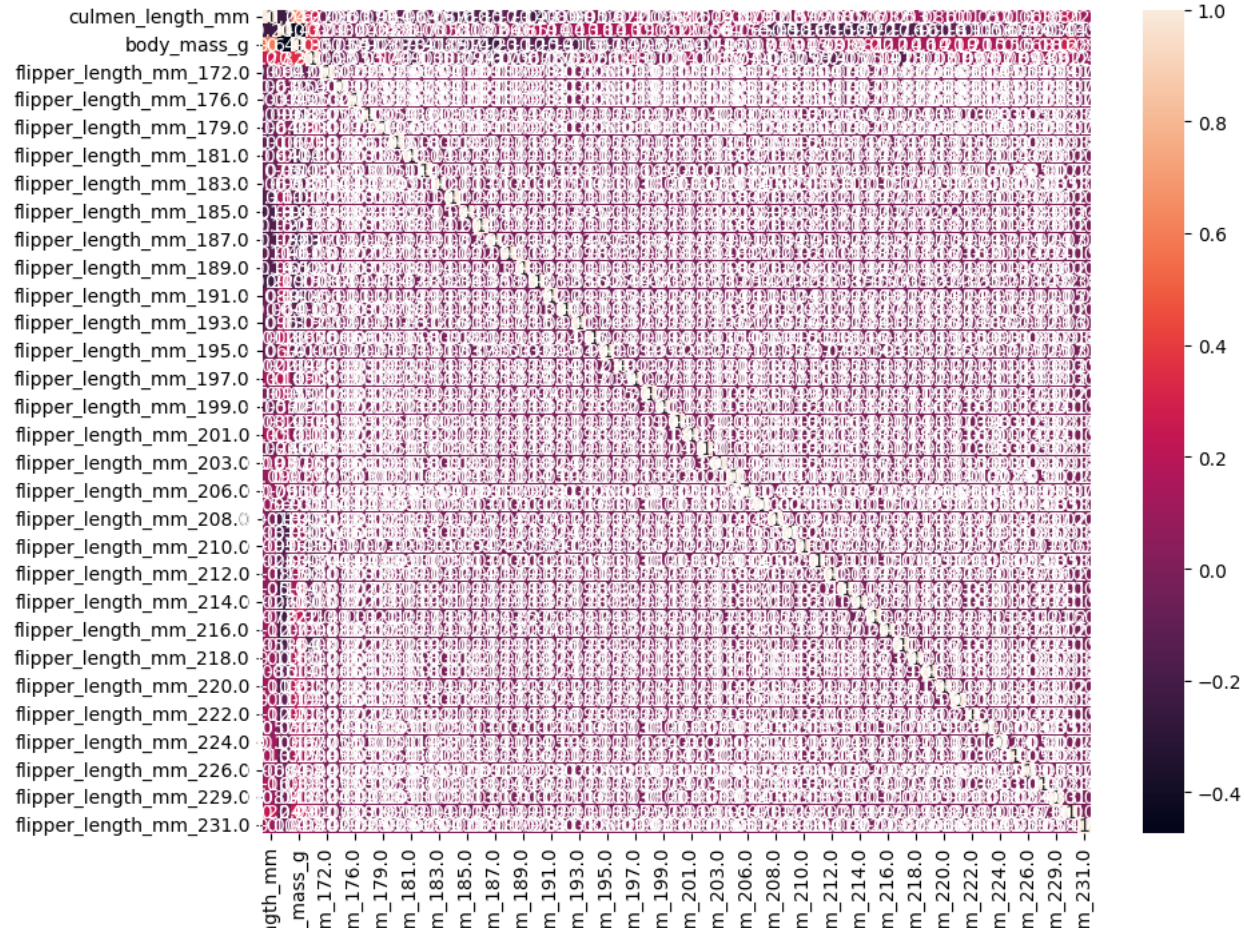
```
df_main.corr()
```

```
<ipython-input-31-b764c75a6398>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In
  df_main.corr()
```

	culmen_length_mm	culmen_depth_mm	body_mass_g	sex	flipper_length_mm_172.0	flipper_lengt
culmen_length_mm	1.000000	-0.235053	0.595110	0.269434	-0.059818	
culmen_depth_mm	-0.235053	1.000000	-0.471916	0.322860	0.039788	
body_mass_g	0.595110	-0.471916	1.000000	0.347376	-0.071125	
sex	0.269434	0.322860	0.347376	1.000000	-0.052544	
flipper_length_mm_172.0	-0.059818	0.039788	-0.071125	-0.052544	1.000000	
flipper_length_mm_174.0	-0.060812	0.031550	-0.054219	-0.052544	-0.002915	
flipper_length_mm_176.0	-0.036972	-0.004151	-0.050838	-0.052544	-0.002915	
flipper_length_mm_178.0	-0.098715	0.006831	-0.119107	-0.057009	-0.005857	
flipper_length_mm_179.0	-0.063792	0.048027	-0.082960	0.140773	-0.002915	
flipper_length_mm_180.0	-0.101035	0.059846	-0.073279	0.099223	-0.006557	
flipper_length_mm_181.0	-0.063840	0.046516	-0.139571	-0.103393	-0.007782	
flipper_length_mm_182.0	-0.075154	0.050036	-0.110829	-0.035306	-0.005064	
flipper_length_mm_183.0	-0.067135	0.005789	-0.085170	-0.005969	-0.004129	
flipper_length_mm_184.0	-0.160812	0.106208	-0.097021	0.007179	-0.007782	
flipper_length_mm_185.0	-0.178247	0.017966	-0.142928	-0.094299	-0.008850	
flipper_length_mm_186.0	-0.154372	0.022430	-0.120230	-0.029679	-0.007782	
flipper_length_mm_187.0	-0.166743	0.036382	-0.228725	-0.165506	-0.011925	
flipper_length_mm_188.0	-0.109208	0.046228	-0.112238	-0.050151	-0.007194	
flipper_length_mm_189.0	-0.144524	0.065366	-0.119585	-0.066536	-0.007782	
flipper_length_mm_190.0	-0.218144	0.187814	-0.148291	0.000866	-0.014114	
flipper_length_mm_191.0	-0.089152	0.137522	-0.144602	-0.056404	-0.010701	
flipper_length_mm_192.0	-0.060052	0.040233	-0.113138	-0.029679	-0.007782	
flipper_length_mm_193.0	-0.095412	0.131296	-0.150692	0.021556	-0.011529	
flipper_length_mm_194.0	0.011575	0.183387	-0.067195	0.055742	-0.006557	
flipper_length_mm_195.0	-0.067057	0.138704	-0.102984	-0.029803	-0.012311	
flipper_length_mm_196.0	0.021905	0.126635	-0.046977	0.048444	-0.009343	
flipper_length_mm_197.0	0.114864	0.190004	-0.052937	0.141369	-0.009343	
flipper_length_mm_198.0	-0.009494	0.130389	-0.056507	-0.012045	-0.008332	
flipper_length_mm_199.0	-0.056916	0.062041	-0.055223	-0.010400	-0.007194	
flipper_length_mm_200.0	0.032488	0.115788	-0.053729	0.040075	-0.005857	
flipper_length_mm_201.0	0.150621	0.130938	-0.011418	0.108854	-0.007194	
flipper_length_mm_202.0	-0.024383	0.066137	-0.074106	-0.008467	-0.005857	
flipper_length_mm_203.0	0.092905	0.087025	0.016465	0.055742	-0.006557	
flipper_length_mm_205.0	0.089927	0.104103	0.029163	0.076632	-0.005064	
flipper_length_mm_206.0	0.079250	0.064504	-0.017025	0.044114	-0.002915	
flipper_length_mm_207.0	0.091841	-0.000045	0.030960	-0.005969	-0.004129	
flipper_length_mm_208.0	0.001507	-0.187474	0.044958	-0.081101	-0.008332	
flipper_length_mm_209.0	0.037047	-0.177355	0.090999	-0.074702	-0.006557	
flipper_length_mm_210.0	0.098513	-0.164763	0.044697	-0.095089	-0.011121	
flipper_length_mm_211.0	0.025718	-0.126453	0.042932	-0.074417	-0.004129	
flipper_length_mm_212.0	0.090102	0.140316	0.008322	0.103393	0.007782	

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

```
<ipython-input-32-141015bc01e3>:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a f
  sns.heatmap(df_main.corr(),annot =True)
<Axes: >
```




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

```
<ipython-input-34-a602bfc050e4>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future
  df_main.corr().culmen_length_mm       .sort_values(ascending=False)
culmen_length_mm           1.000000
body_mass_g                0.595110
sex                        0.269434
flipper_length_mm_230.0    0.224423
flipper_length_mm_228.0    0.161196
flipper_length_mm_201.0    0.150621
flipper_length_mm_225.0    0.127772
flipper_length_mm_222.0    0.118755
flipper_length_mm_216.0    0.117191
flipper_length_mm_197.0    0.114864
flipper_length_mm_220.0    0.105481
flipper_length_mm_224.0    0.104307
flipper_length_mm_231.0    0.103090
flipper_length_mm_221.0    0.102289
flipper_length_mm_218.0    0.101842
flipper_length_mm_210.0    0.098513
flipper_length_mm_203.0    0.092905
flipper_length_mm_207.0    0.091841
flipper_length_mm_205.0    0.089927
flipper_length_mm_219.0    0.085755
flipper_length_mm_229.0    0.080586
flipper_length_mm_212.0    0.080102
flipper_length_mm_206.0    0.079250
flipper_length_mm_215.0    0.077936
flipper_length_mm_226.0    0.068323
flipper_length_mm_223.0    0.056669
flipper_length_mm_217.0    0.053389
flipper_length_mm_213.0    0.050121
flipper_length_mm_209.0    0.037047
flipper_length_mm_200.0    0.032488
flipper_length_mm_214.0    0.026426
flipper_length_mm_211.0    0.025718
flipper_length_mm_196.0    0.021905
flipper_length_mm_194.0    0.011575
flipper_length_mm_208.0    0.001507
flipper_length_mm_198.0   -0.009494
flipper_length_mm_202.0   -0.024383
flipper_length_mm_176.0   -0.036972
flipper_length_mm_199.0   -0.056916
flipper_length_mm_172.0   -0.059818
flipper_length_mm_192.0   -0.060052
```

```
flipper_length_mm_174.0   -0.060812
flipper_length_mm_179.0   -0.063792
flipper_length_mm_181.0   -0.063840
flipper_length_mm_195.0   -0.067057
flipper_length_mm_183.0   -0.067135
flipper_length_mm_182.0   -0.075154
flipper_length_mm_191.0   -0.089152
flipper_length_mm_193.0   -0.095412
flipper_length_mm_178.0   -0.098715
flipper_length_mm_180.0   -0.101035
flipper_length_mm_188.0   -0.109208
flipper_length_mm_189.0   -0.144524
flipper_length_mm_186.0   -0.154372
```

```
df_main.head()
```

	species	island	culmen_length_mm	culmen_depth_mm	body_mass_g	sex	flipper_length_mm_172.0	flipper_length_mm_174.0
0	Adelie	Torgersen	39.1	18.7	3750.0	2	0	0
1	Adelie	Torgersen	39.5	17.4	3800.0	1	0	0
2	Adelie	Torgersen	40.3	18.0	3250.0	1	0	0
3	Adelie	Torgersen	NaN	NaN	NaN	3	0	0
4	Adelie	Torgersen	36.7	19.3	3450.0	1	0	0

5 rows × 61 columns

```
y = df_main['culmen_depth_mm']
y
```

```
0      18.7
1      17.4
2      18.0
3       NaN
4      19.3
       ...
339     NaN
340    14.3
341    15.7
342    14.8
343    16.1
Name: culmen_depth_mm, Length: 344, dtype: float64
```

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

	species	island	culmen_length_mm	body_mass_g	sex	flipper_length_mm_172.0	flipper_length_mm_174.0	flipper_length_m
0	Adelie	Torgersen	39.1	3750.0	2	0	0	
1	Adelie	Torgersen	39.5	3800.0	1	0	0	
2	Adelie	Torgersen	40.3	3250.0	1	0	0	
3	Adelie	Torgersen	NaN	NaN	3	0	0	
4	Adelie	Torgersen	36.7	3450.0	1	0	0	

5 rows × 60 columns