In [1]:
```
# Load dataset and import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly .express as px
```

In [2]:
```
df = pd.read_csv(r"C:\Users\jayant soni\Downloads\penguins_lter (1).csv")
df.head()
```

Out[2]:

	studyName	Sample Number	Species	Region	Island	Stage	Individual ID	Clutch Completion	Date Egg	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Sex	Delta 15 N (o/oo)	Delta 13 C (o/oo)	Comments
0	PAL0708	1	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen	Adult, 1 Egg Stage	N1A1	Yes	11/11/07	39.1	18.7	181.0	3750.0	MALE	NaN	NaN	Not enough blood for isotopes.
1	PAL0708	2	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen	Adult, 1 Egg Stage	N1A2	Yes	11/11/07	39.5	17.4	186.0	3800.0	FEMALE	8.94956	-24.69454	NaN
2	PAL0708	3	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen	Adult, 1 Egg Stage	N2A1	Yes	11/16/07	40.3	18.0	195.0	3250.0	FEMALE	8.36821	-25.33302	NaN
3	PAL0708	4	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen	Adult, 1 Egg Stage	N2A2	Yes	11/16/07	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Adult not sampled.
4	PAL0708	5	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen	Adult, 1 Egg Stage	N3A1	Yes	11/16/07	36.7	19.3	193.0	3450.0	FEMALE	8.76651	-25.32426	NaN

In [3]:
```
#to detect outlier
df.describe()
```

Out[3]:

	Sample Number	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Delta 15 N (o/oo)	Delta 13 C (o/oo)
count	344.000000	342.000000	342.000000	342.000000	342.000000	330.000000	331.000000
mean	63.151163	43.921930	17.151170	200.915205	4201.754386	8.733382	-25.686292
std	40.430199	5.459584	1.974793	14.061714	801.954536	0.551770	0.793961
min	1.000000	32.100000	13.100000	172.000000	2700.000000	7.632200	-27.018540
25%	29.000000	39.225000	15.600000	190.000000	3550.000000	8.299890	-26.320305
50%	58.000000	44.450000	17.300000	197.000000	4050.000000	8.652405	-25.833520
75%	95.250000	48.500000	18.700000	213.000000	4750.000000	9.172123	-25.062050
max	152.000000	59.600000	21.500000	231.000000	6300.000000	10.025440	-23.787670

In [4]:
```
#Missing value imputation
df.isnull().sum()
```

Out[4]:
```
studyName             0
Sample Number         0
Species               0
Region                0
Island                0
Stage                 0
Individual ID         0
Clutch Completion     0
Date Egg              0
Culmen Length (mm)    2
Culmen Depth (mm)     2
Flipper Length (mm)   2
Body Mass (g)         2
Sex                  10
Delta 15 N (o/oo)    14
Delta 13 C (o/oo)    13
Comments            318
dtype: int64
```

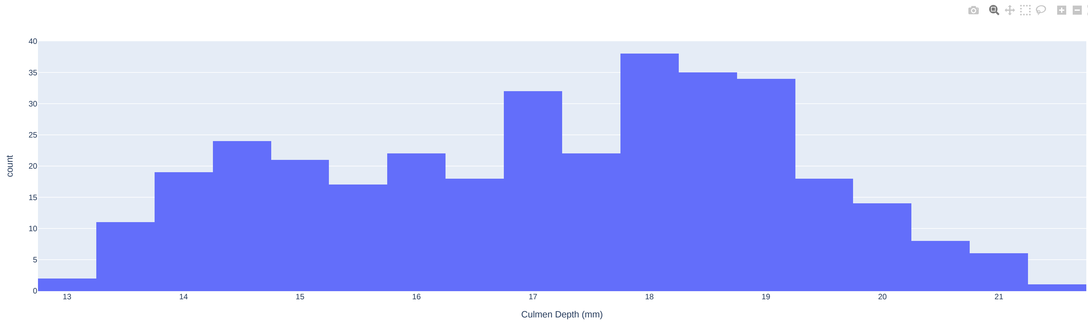
In [5]:
```
df.notnull().sum()
```

Out[5]:
```
studyName           344
Sample Number       344
Species             344
Region              344
Island              344
Stage               344
Individual ID       344
Clutch Completion   344
Date Egg            344
Culmen Length (mm)  342
Culmen Depth (mm)   342
Flipper Length (mm) 342
Body Mass (g)       342
Sex                 334
Delta 15 N (o/oo)   330
Delta 13 C (o/oo)   331
Comments             26
dtype: int64
```

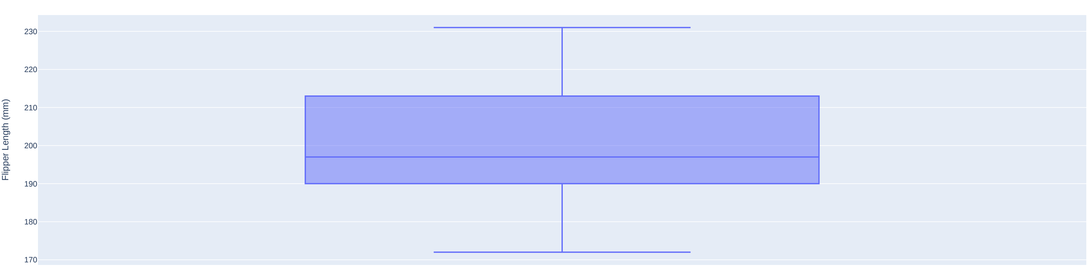
In [6]:
```
df.isna().mean()*100
```

Out[6]:
```
studyName            0.000000
Sample Number        0.000000
Species              0.000000
Region               0.000000
Island               0.000000
Stage                0.000000
Individual ID        0.000000
Clutch Completion    0.000000
Date Egg             0.000000
Culmen Length (mm)   0.581395
Culmen Depth (mm)    0.581395
Flipper Length (mm)  0.581395
Body Mass (g)        0.581395
Sex                  2.906977
Delta 15 N (o/oo)    4.069767
Delta 13 C (o/oo)    3.779070
Comments            92.441860
dtype: float64
```

In [7]:
```
fig = px.histogram(df, x='Culmen Depth (mm)')
fig.show()
```



In [8]:
```
fig=px.box(df,y=('Flipper Length (mm)'))
fig.show()
```



In [9]:
```
def find_outliers_IQR(df):
    q1=df.quantile(0.25)
    q3=df.quantile(0.75)
    IQR=q3-q1
    outliers = df[((df<(q1-1.2*IQR)) | (df>(q3+1.2*IQR)))]
    return outliers
```

In [10]:
```
outliers = find_outliers_IQR(df["Culmen Depth (mm)"])
print("number of outliers: "+ str(len(outliers)))
print("max outlier value: "+ str(outliers.max()))
print("min outlier value: "+ str(outliers.min()))
outliers
```

Out[10]:
```
number of outliers: 0
max outlier value: nan
min outlier value: nan
Series([], Name: Culmen Depth (mm), dtype: float64)
```

In [11]:
```
#to find the correlation amang
df.corr(method ='pearson')
```

Out[11]:

	Sample Number	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Delta 15 N (o/oo)	Delta 13 C (o/oo)
Sample Number	1.000000	-0.236356	-0.022352	0.040849	-0.007042	0.006952	-0.488690
Culmen Length (mm)	-0.236356	1.000000	-0.235053	0.656181	0.595110	-0.059759	0.189025
Culmen Depth (mm)	-0.022352	-0.235053	1.000000	-0.583851	-0.471916	0.605874	0.429933
Flipper Length (mm)	0.040849	0.656181	-0.583851	1.000000	0.871202	-0.507787	-0.376223
Body Mass (g)	-0.007042	0.595110	-0.471916	0.871202	1.000000	-0.537888	-0.374638
Delta 15 N (o/oo)	0.006952	-0.059759	0.605874	-0.507787	-0.537888	1.000000	0.570615
Delta 13 C (o/oo)	-0.488690	0.189025	0.429933	-0.376223	-0.374638	0.570615	1.000000

In [137]:
```
df.corr(method ='kendall')
```

Out[137]:

	Sample Number	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Delta 15 N (o/oo)	Delta 13 C (o/oo)
Sample Number	1.000000	-0.141843	-0.029314	0.040947	0.009671	0.009710	-0.254544
Culmen Length (mm)	-0.141843	1.000000	-0.122850	0.483345	0.433359	-0.064553	0.097830
Culmen Depth (mm)	-0.029314	-0.122850	1.000000	-0.281894	-0.195070	0.424881	0.293145
Flipper Length (mm)	0.040947	0.483345	-0.281894	1.000000	0.660467	-0.316815	-0.230067
Body Mass (g)	0.009671	0.433359	-0.195070	0.660467	1.000000	-0.372535	-0.254730
Delta 15 N (o/oo)	0.009710	-0.064553	0.424881	-0.316815	-0.372535	1.000000	0.362955
Delta 13 C (o/oo)	-0.254544	0.097830	0.293145	-0.230067	-0.254730	0.362955	1.000000

In [ ]:

In [ ]:

In [ ]: