PENGUINS DATASET - PREPROCESSING

Short description about d datasest (e.g. number of samples, the domain, main statistics)

	calorie requirement	average sleep duration	bill_length_mm	\
count	344.000000	344.000000	337.000000	
mean	5270.002907	10.447674	45.494214	
std	1067.959116	2.265895	10.815787	
min	3504.000000	7.000000	32.100000	
25%	4403.000000	9.000000	39.500000	
50%	5106.500000	10.000000	45.100000	
75%	6212.750000	12.000000	49.000000	
max	7197.000000	14.000000	124.300000	

	bill_depth_mm	flipper_length_mm	body_mass_g	year
count	333.000000	336.000000	339.000000	342.000000
mean	18.018318	197.764881	4175.463127	2008.035088
std	9.241384	27.764491	858.713267	0.816938
min	13.100000	10.000000	882.000000	2007.000000
25%	15.700000	190.000000	3550.000000	2007.000000
50%	17.300000	197.000000	4050.000000	2008.000000
75%	18.700000	213.000000	4750.000000	2009.000000
max	127.260000	231.000000	6300.000000	2009.000000

Domanin : It’s a biological classification of species penguins,

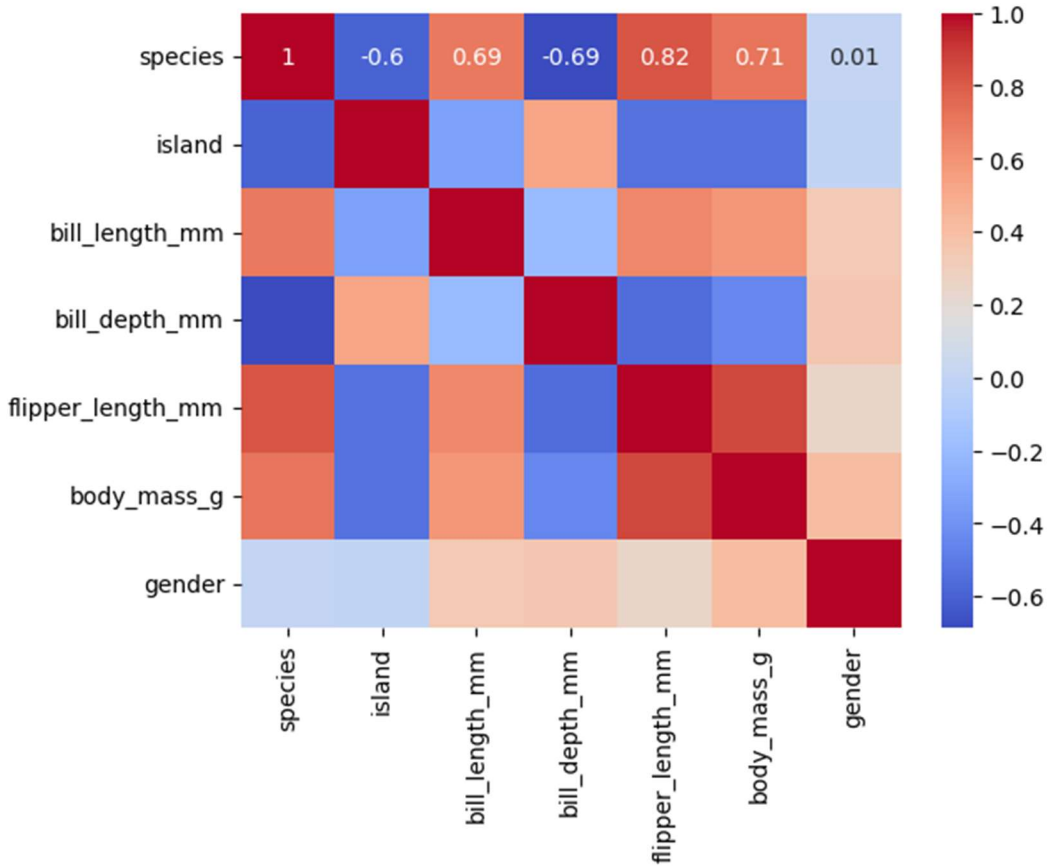
It contains 344 rows and 10 columns

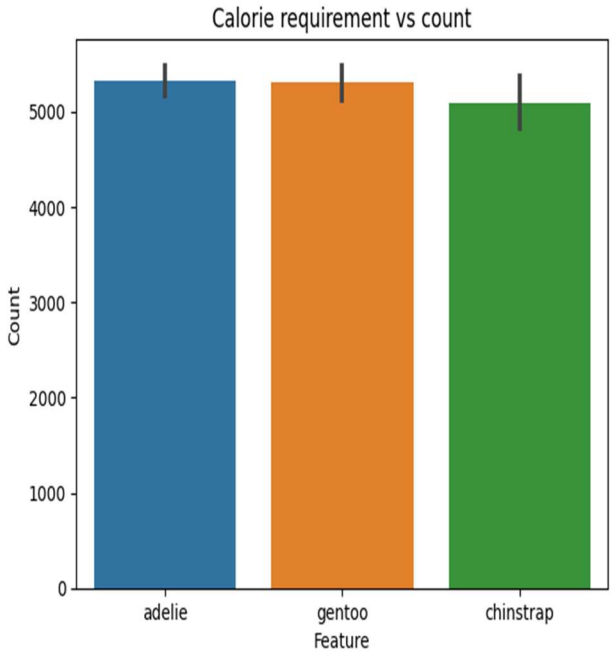
Short details of the methods you used for data preprocessing

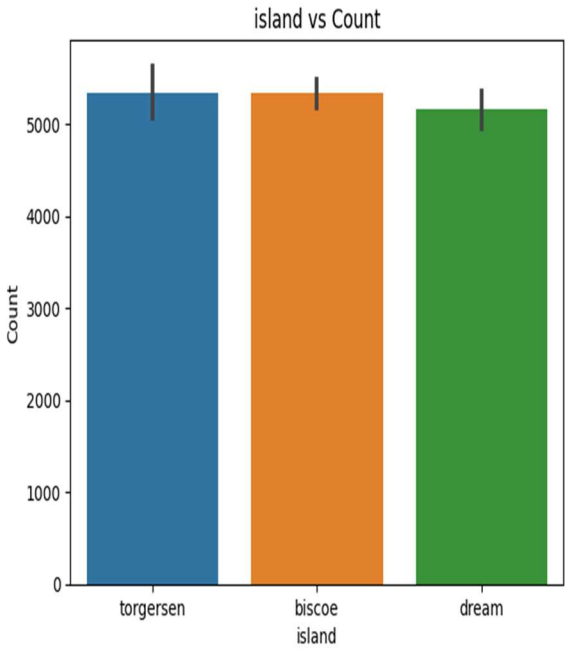
For preprocessing we used

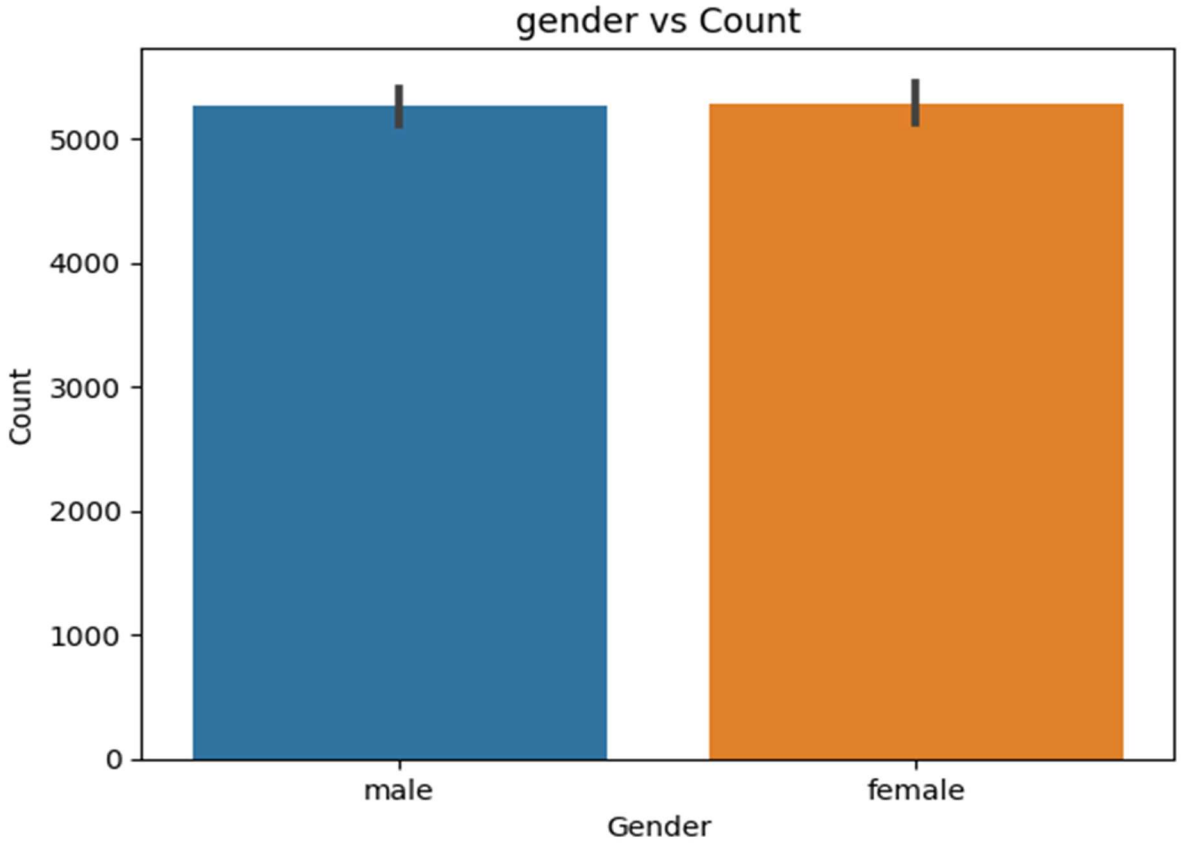
- mode imputaion for categorical data value_counts() and get the index0,
- mean imputation for numerical data
- find_quartile for finding the interquartile range
- corr() to find the correlation matrix
- Categorical().codes for converting the string values to categorical values
- Used fillna() to fill the computed values for mode where the column is empty
- Str.lower() to convert the string to lower case letter to handle the mismatched data
- Also used some basic functions like abs() to find the absolute value, df.mean() to find the mean and df.mode() to find the mode during preprocessing

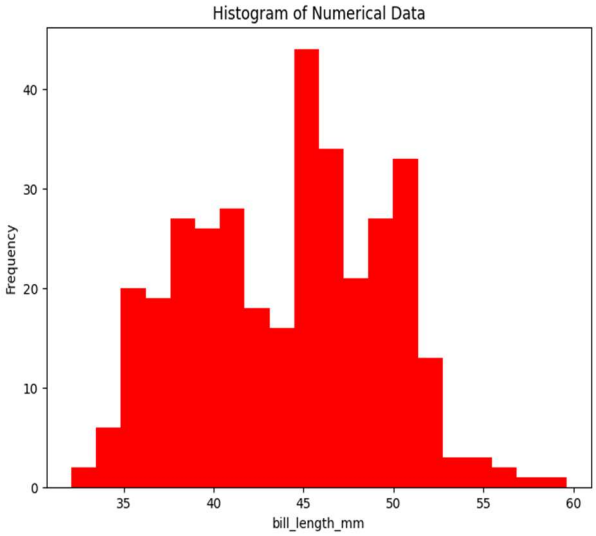
Graphs

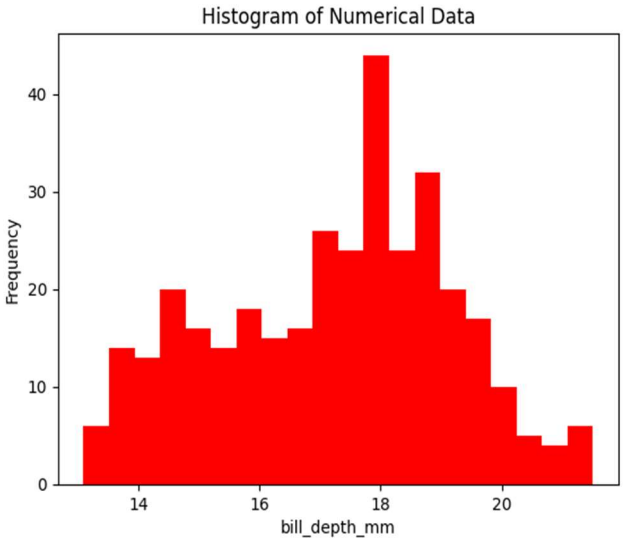


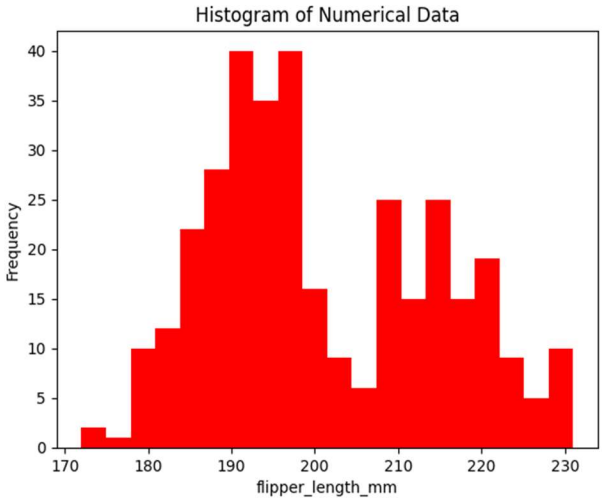

Calorie requirement vs count
Count
0
1000
2000
3000
4000
5000
adelie
gentoo
chinstrap
Feature

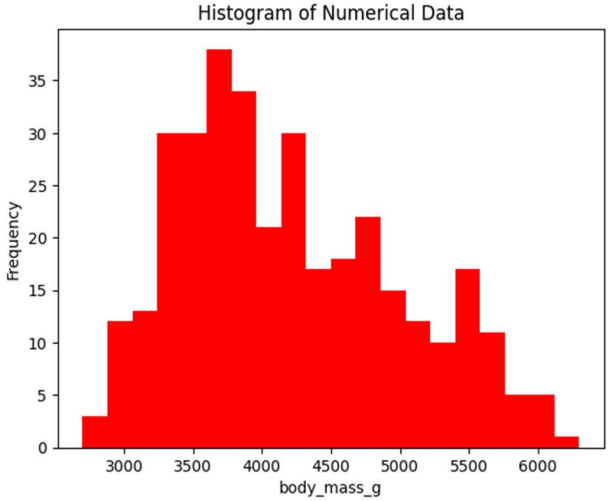

island vs Count
Count
0
1000
2000
3000
4000
5000
torgersen
biscoe
dream
island

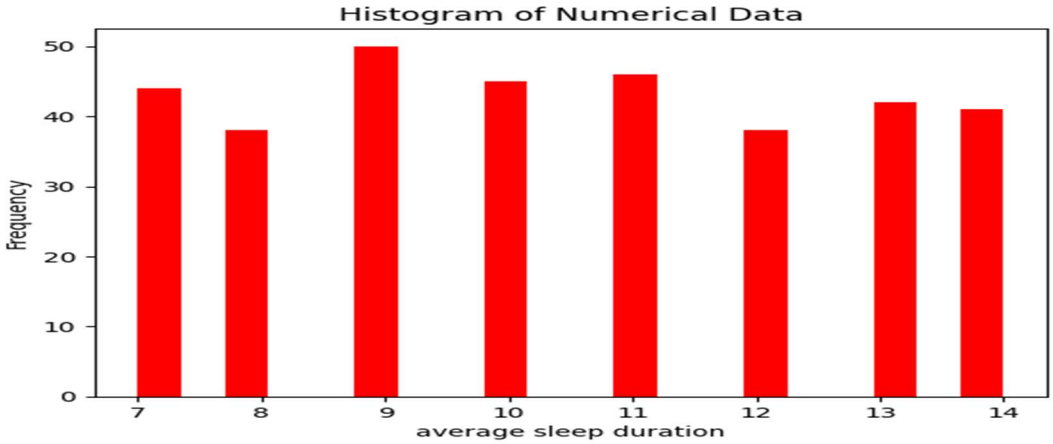

gender vs Count
Count
0
1000
2000
3000
4000
5000
male
female
Gender

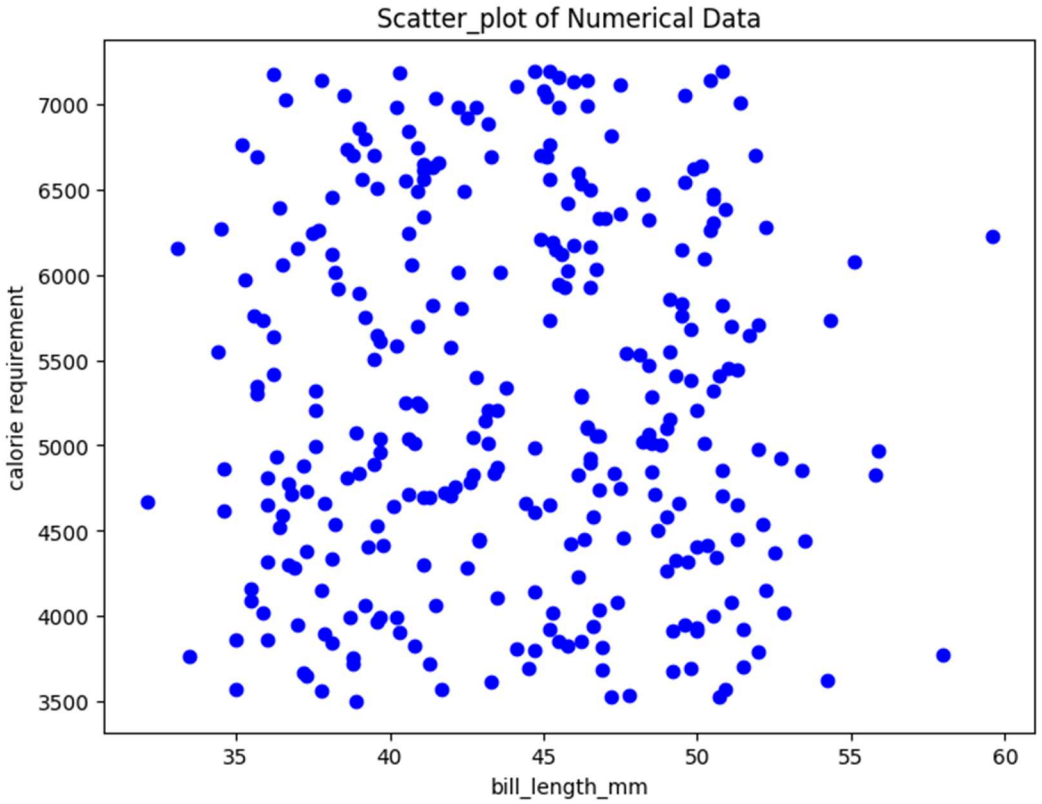

Histogram of Numerical Data
Frequency
bill_length_mm

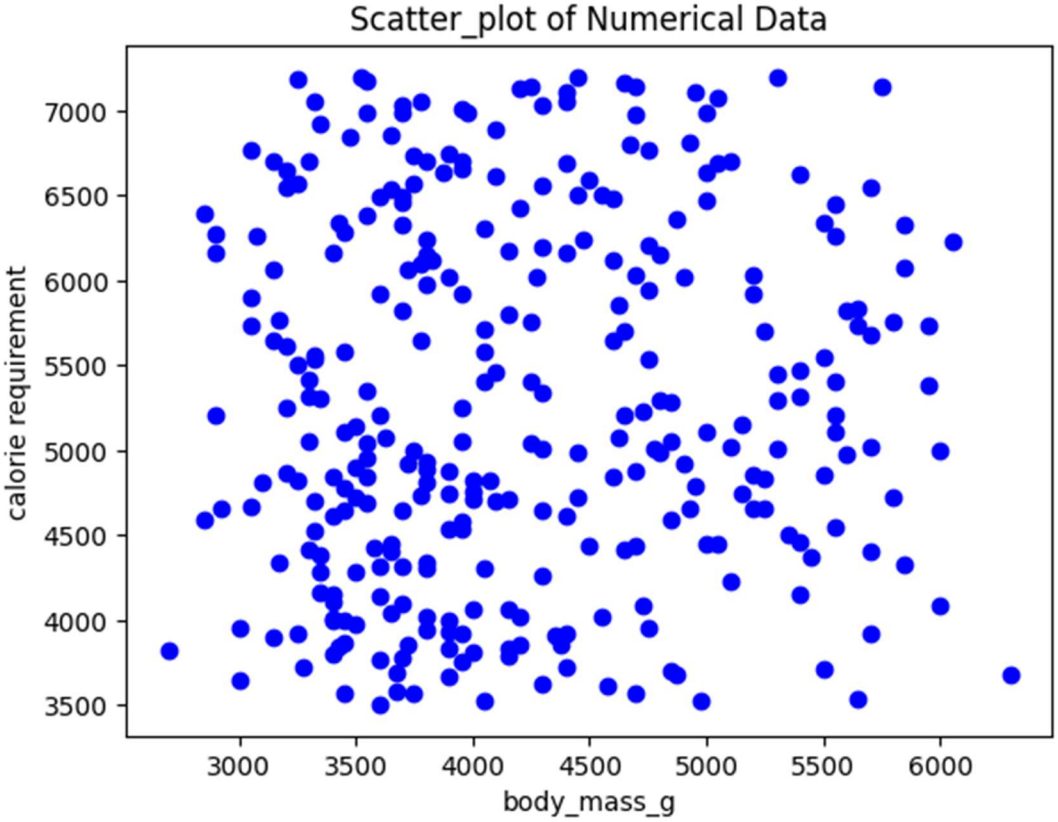

Histogram of Numerical Data
Frequency
bill_depth_mm

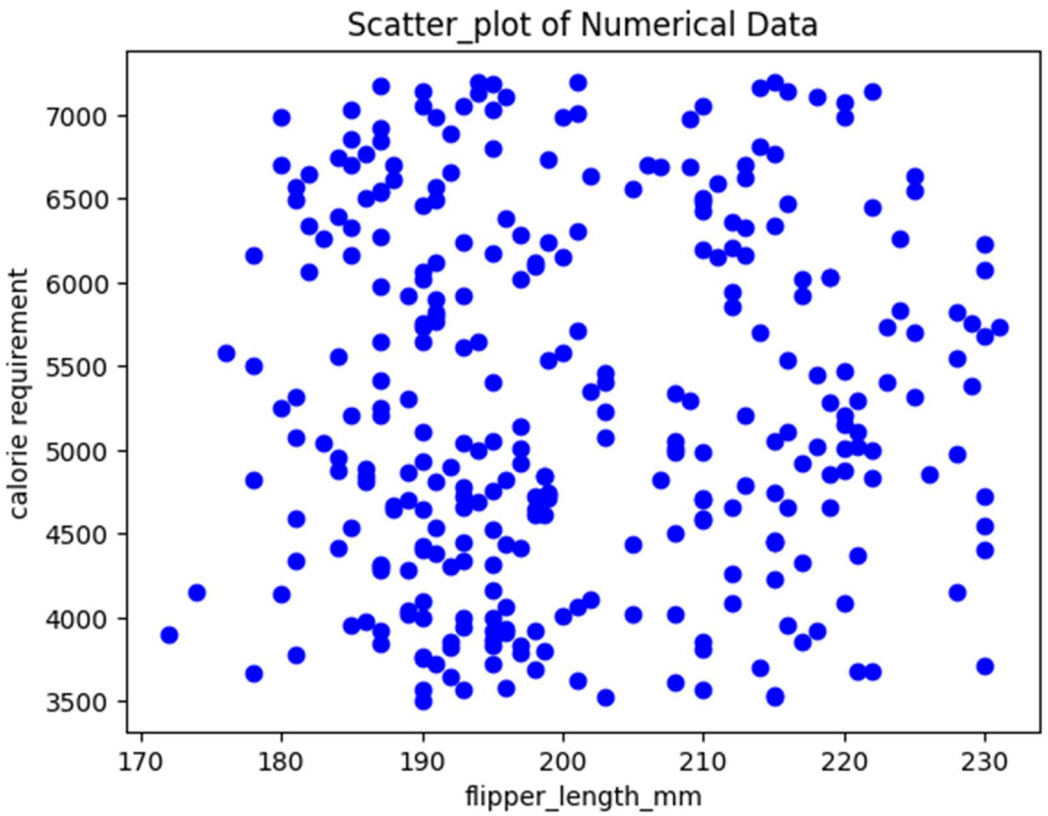

Histogram of Numerical Data
Frequency
flipper_length_mm

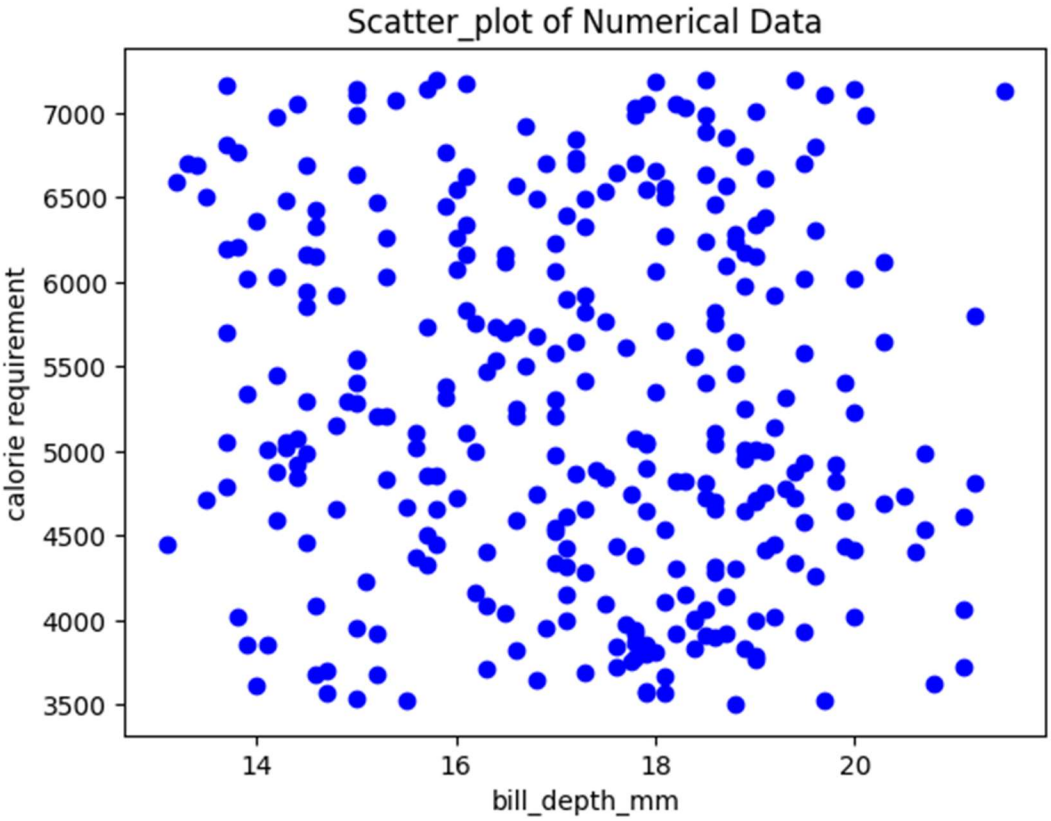

Histogram of Numerical Data
Frequency
body_mass_g

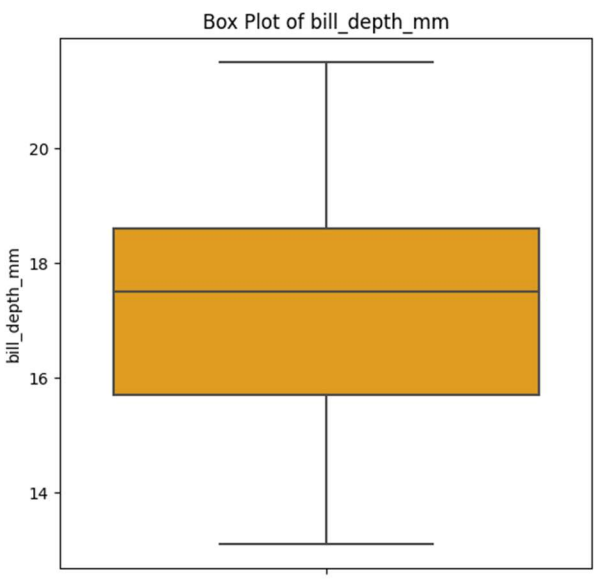

Histogram of Numerical Data
Frequency
average sleep duration

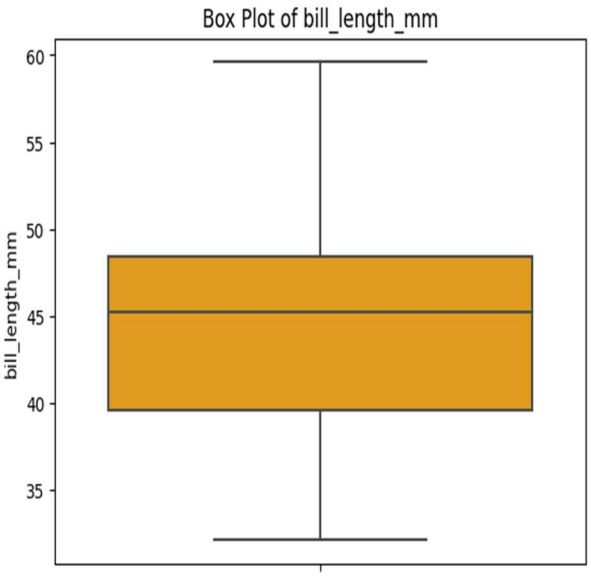

Scatter_plot of Numerical Data
calorie requirement
bill_length_mm
7000
6500
6000
5500
5000
4500
4000
3500
35
40
45
50
55
60

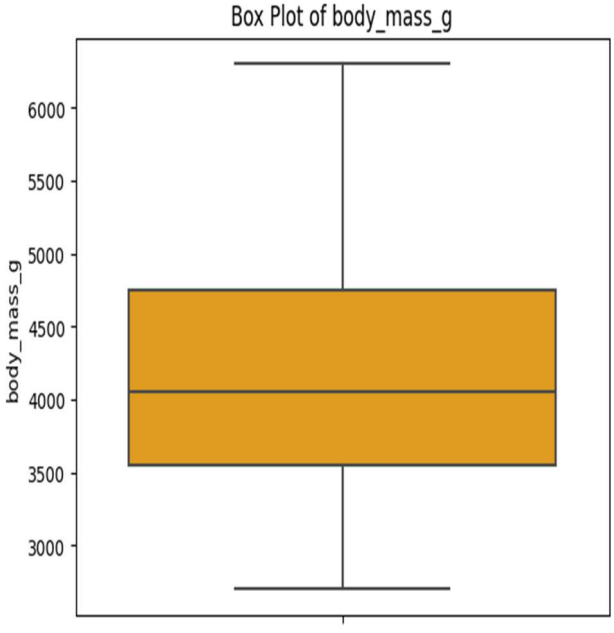

Scatter_plot of Numerical Data
calorie requirement
body_mass_g
7000
6500
6000
5500
5000
4500
4000
3500
3000
3500
4000
4500
5000
5500
6000

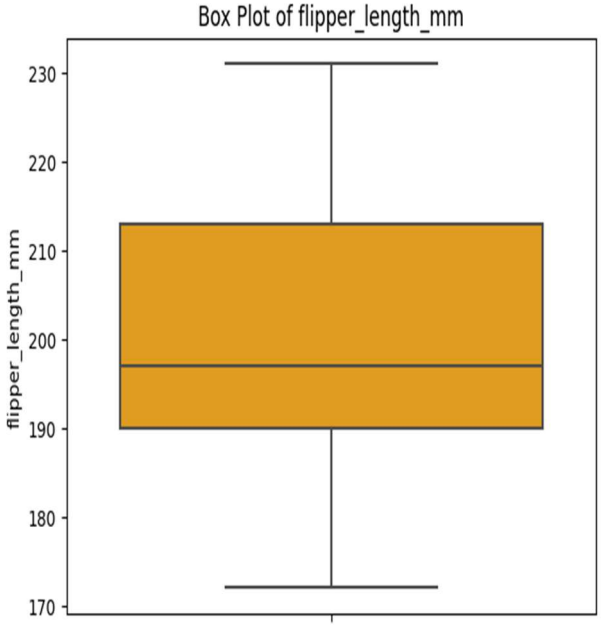

Scatter_plot of Numerical Data
calorie requirement
flipper_length_mm
7000
6500
6000
5500
5000
4500
4000
3500
170
180
190
200
210
220
230


Scatter_plot of Numerical Data
calorie requirement
bill_depth_mm
7000
6500
6000
5500
5000
4500
4000
3500
14
16
18
20


Box Plot of bill_depth_mm
bill_depth_mm
20
18
16
14


Box Plot of bill_length_mm
bill_length_mm
60
55
50
45
40
35


Box Plot of body_mass_g
body_mass_g
6000
5500
5000
4500
4000
3500
3000


Box Plot of flipper_length_mm
flipper_length_mm
230
220
210
200
190
180
170

- In visualization graphs we could have a prcise understanding on numeric data
- Most of the penguins have bill length between 38 to 48 mm
- Most of the penguins have bill depth between 15.5 to 18.5 mm
- Most of the penguins have body mass between 3500 to 4700 g
- Most of the penguind have flipper length duration between 190 to 215 mm

For categorical data

- Adelie species has the highest calorie requirement
- There are more male penguins than female
- Most of the penguins live in Torgersion island