Summary of Mini Project 3

Statistics (Mean, Median, and Standard Deviation)

In [ ]:
```python
# 1. import Python packages
import polars as pl
import matplotlib.pyplot as plt
```

```
c:\Users\User\miniconda3\Lib\site-packages\numpy\_distributor_init.py:30: UserWarning: loaded more than 1 DLL from .l
ibs:
c:\Users\User\miniconda3\Lib\site-packages\numpy\.libs\libopenblas64__v0.3.21-gcc_10_3_0.dll
c:\Users\User\miniconda3\Lib\site-packages\numpy\.libs\libopenblas64__v0.3.23-246-g3d31191b-gcc_10_3_0.dll
  warnings.warn("loaded more than 1 DLL from .libs:"
```

In [ ]:
```python
# 2. Load the dataset and verify if it has been imported correctly.
penguins_df = pl.read_csv("penguins.csv")
print(penguins_df)
```

shape: (344, 9)

rowid	species	island	bill_length_mm	…	flipper_length_mm	body_mass_g	sex	year
i64	str	str	f64		i64	i64	str	i64
1	Adelie	Torgersen	39.1	…	181	3750	male	2007
2	Adelie	Torgersen	39.5	…	186	3800	female	2007
3	Adelie	Torgersen	40.3	…	195	3250	female	2007
4	Adelie	Torgersen	null	…	null	null	null	2007
…	…	…	…	…	…	…	…	…
341	Chinstrap	Dream	43.5	…	202	3400	female	2009
342	Chinstrap	Dream	49.6	…	193	3775	male	2009
343	Chinstrap	Dream	50.8	…	210	4100	male	2009
344	Chinstrap	Dream	50.2	…	198	3775	female	2009

In [ ]:
```python
# 3. Calculate mean, median, standard deviation of each columns
def calculate_stat():
    penguins_desc = penguins_df.describe()
```

```
    print(penguins_desc)


calculate_stat()
```

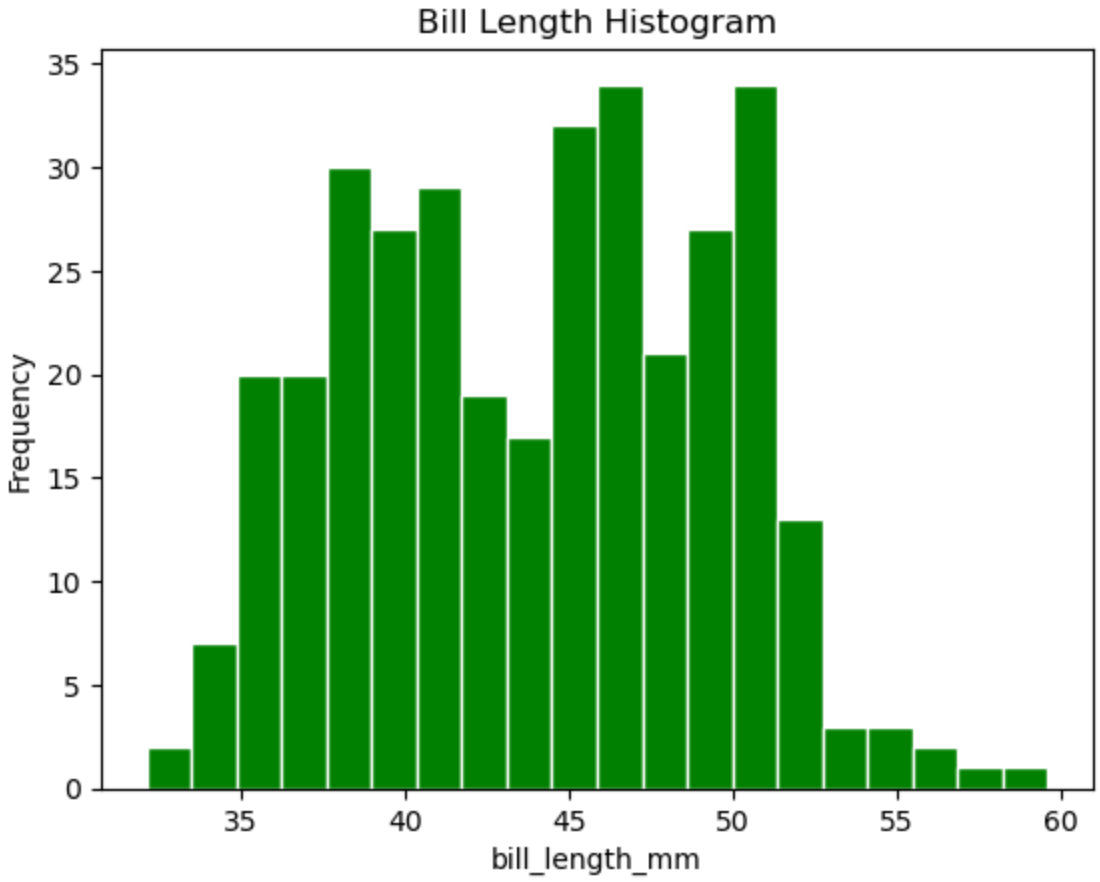
shape: (9, 10)

describe	rowid	species	island	…	flipper_length_mm	body_mass_g	sex	year
str	f64	str	str		f64	f64	str	f64
count	344.0	344	344	…	344.0	344.0	344	344.0
null_count	0.0	0	0	…	2.0	2.0	11	0.0
mean	172.5	null	null	…	200.915205	4201.754386	null	2008.02907
std	99.448479	null	null	…	14.061714	801.954536	null	0.818356
min	1.0	Adelie	Biscoe	…	172.0	2700.0	female	2007.0
25%	87.0	null	null	…	190.0	3550.0	null	2007.0
50%	173.0	null	null	…	197.0	4050.0	null	2008.0
75%	259.0	null	null	…	213.0	4750.0	null	2009.0
max	344.0	Gentoo	Torgersen	…	231.0	6300.0	male	2009.0

Data Visualization (Histogram)

In [ ]:
```
# 4. Make a histogram of 'bill_length_mm' column in penguins.csv
def build_histogram():
    plt.hist(penguins_df["bill_length_mm"], bins=20, color="green", edgecolor="white")
    plt.xlabel("bill_length_mm")
    plt.ylabel("Frequency")
    plt.title("Bill Length Histogram")
    plt.savefig("bill_length_hist.png")
    plt.show()
    return


build_histogram()
```


Bill Length Histogram
Frequency
bill_length_mm
0
5
10
15
20
25
30
35
35
40
45
50
55
60