Palmer Penguins Data Analysis Presentation

Zach Newby - November 27th 2023

Overview:

This is a Python data analysis project analyzing the Palmer Penguins Dataset. The purpose of this project was to learn how to perform data analysis using python libraries for my Applied Programming Course from Brigham Young University Idaho Online. (CSE 310). For it I used Python 3.12, and the Pandas and matplotlib libraries.

This is a guide document that gives an overview of the project, showing the code and graphs I created to answer the following questions:

1. On which of the three islands (Dream, Biscoe, and Torgenson) were Adelie penguins with long flippers (greater than the average length of the species) most common?
2. Is there a relationship (correlation) between bill length and depth? If so, what kind of relationship is it? (ie. direct correlation or indirect). If not, do any factors correlate to either?

For fun and demonstration purposes I also created two graphs comparing the populations of the penguins, one by island, and the other by species.

To see the python functions described in this document and recreate the graphs yourself, use *analysis_presentation.py* in the project Github repository. This will call *graph_populations()*, *answer_question_1()*, and *answer_question_2()* in that order, but you can comment them out to call them one at a time.

The Dataset:

The Dataset used in this project is the Palmer Penguins Dataset from Kaggle.com, it was compiled from research on Adelie, Chinstrap, and Gentoo penguins by Dr. Kristen Gorman and the Palmer Station and uploaded to Kaggle by Ashwani Rathee for data visualization and exploration purposes as an alternative to the Iris dataset. It consists of the basic characteristics studied by Dr. Gorman, consisting of species, the island they were found on, bill length (mm), bill depth(mm), flipper length(mm), body mass (g), and sex.

To properly analyze the data, we will download the dataset as a .csv file (penguins.csv) and convert it to a pandas dataframe at the start of *analysis_presentation.py* after importing all of the libraries.

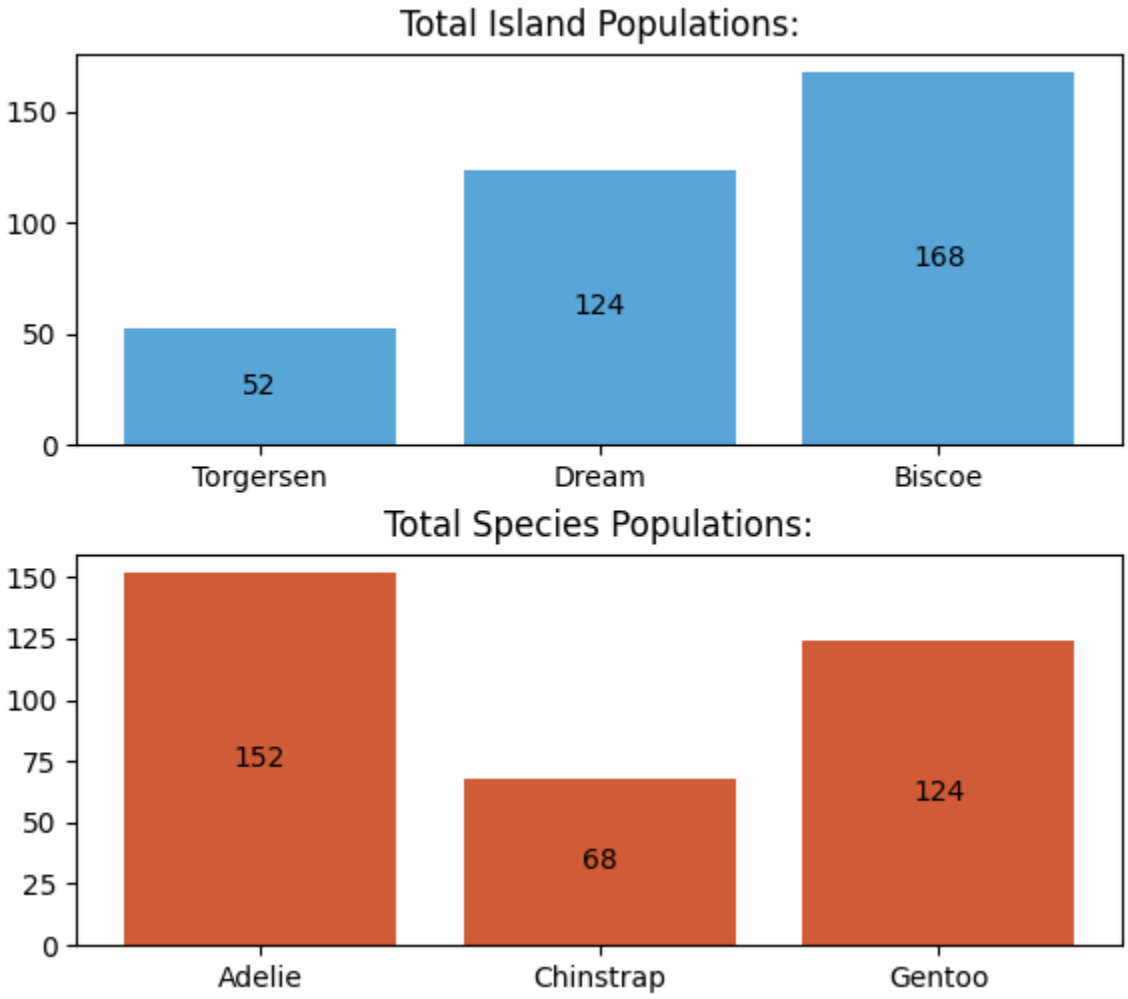
```
#Set this to global since we will be using it in multiple functions and not
modifying it.
global penguins
penguins = pd.read_csv("penguins.csv")

    ...
print(penguins)
```

	species	island	bill_length_mm	bill_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
..	...	...	...	...	...	...	...
339	Chinstrap	Dream	55.8	19.8	207.0	4000.0	male
340	Chinstrap	Dream	43.5	18.1	202.0	3400.0	female
341	Chinstrap	Dream	49.6	18.2	193.0	3775.0	male
342	Chinstrap	Dream	50.8	19.0	210.0	4100.0	male
343	Chinstrap	Dream	50.2	18.7	198.0	3775.0	female

[344 rows x 7 columns]

Population Graphs:

These two graphs don't help answer the questions, but I decided to include them anyway as a demonstration. To recreate them, go into *analysis_presentation.py* and call *graph_populations()*. This will call *calc_population_totals()* and save the result, (a dictionary containing the populations for each island and species) as a variable called pop_totals, and use that to create the two bar charts.



As you can see, Adelie penguins have the largest population (and are the only species found on all three islands in fact), while Biscoe Island has the largest population of penguins.

Question 1:

"*On which of the three islands (Dream, Biscoe, and Torgenson) were Adelie penguins with long flippers (greater than the average length of the species) most common?*"

Originally I was going to examine the flipper lengths of all three species of penguin, but only the Adelie penguins are found on all three islands. The function *answer_question_1()* contains several nested functions and performs the operations to call them and generate the graphs. We will work our way through this function step by step.

Step 1: Calculate the average flipper length:

First we calculate the average flipper length for each species and save it to a dictionary using *_calculate_median_flipper_lengths()*. While only the Adelie penguin average flipper length is important to us, we will save the other two average flipper lengths for later. For our purposes, a long flipper is above the average (median) length for the species.

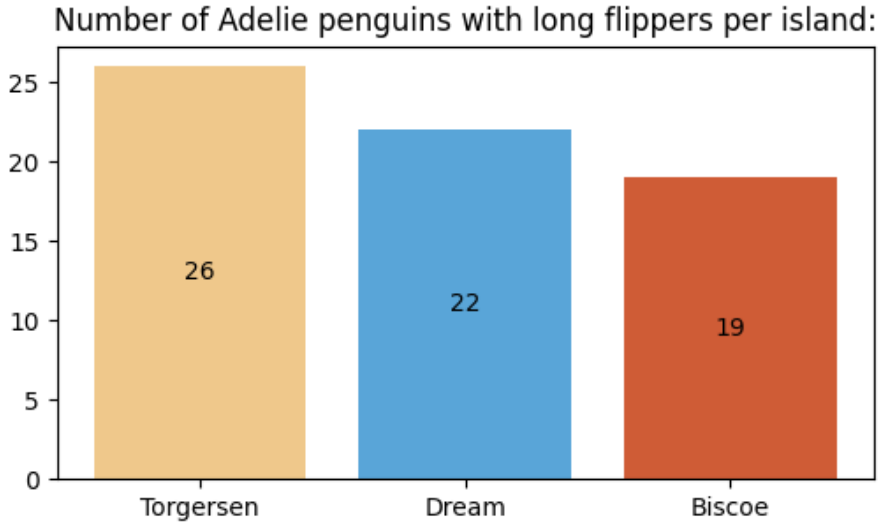
```
def _calculate_median_flipper_lengths():

        median_flipper_lengths = dict()
        #Get median flipper length of each species and save to dict
        median_flipper_lengths["adelie_median_flipper_length"] =
penguins[penguins["species"] == "Adelie"]["flipper_length_mm"].median()
        median_flipper_lengths["chinstrap_median_flipper_length"] =
penguins[penguins["species"] == "Chinstrap"]["flipper_length_mm"].median()
        median_flipper_lengths["gentoo_median_flipper_length"] =
penguins[penguins["species"] == "Gentoo"]["flipper_length_mm"].median()

        return median_flipper_lengths
```

The average (median) flipper length of Adelie penguins recorded in the dataset is **190mm.** We will then isolate all Adelie penguins with above average flippers and save them to a separate data frame:

```
#Isolate Adelie penguins with long flippers, and save to a separate data
frame
    adelie_with_long_flippers = penguins[(penguins["species"] == "Adelie")
& (penguins["flipper_length_mm"] >
median_flipper_lengths["adelie_median_flipper_length"])]
```

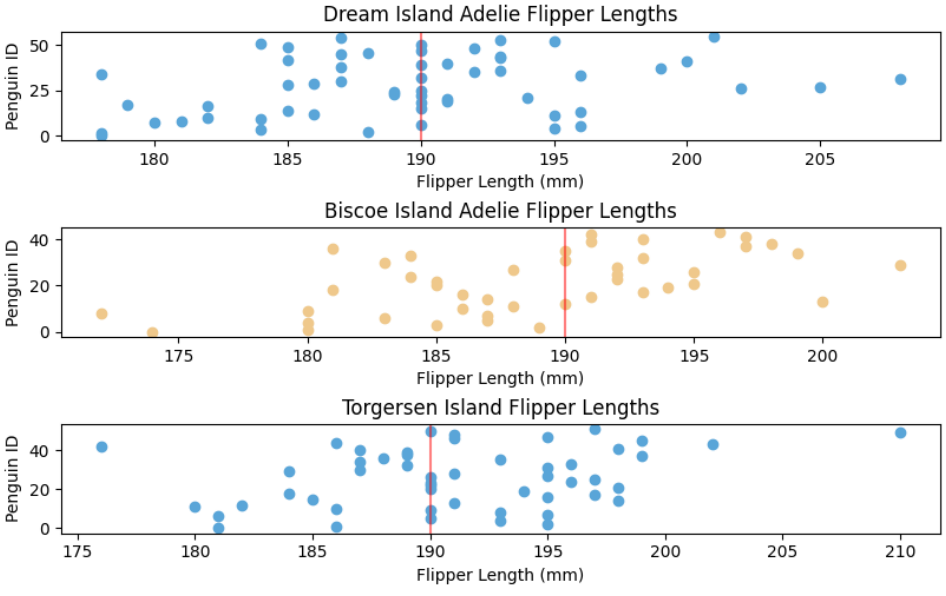
```
    #for demonstrational purposes, the penguins are sorted by flipper
length and the indexes are reset to match the ascending order of flipper
lengths
    adelie_with_long_flippers =
adelie_with_long_flippers.sort_values("flipper_length_mm")

adelie_with_long_flippers.set_index(pd.Series(range(0,len(adelie_with_long_
flippers))), inplace=True)
```

Then we call *_graph_long_flipper_adelies_pops()* to produce this graph:



This answers the first question - **Adelie penguins with long (greater than average length) flippers are found most often on Torgersen Island with 26 penguins, and least often on Biscoe.**

To help further illustrate that Torgersen has the most Adelie penguins with long flippers call *_graph_adelies_per_island()*

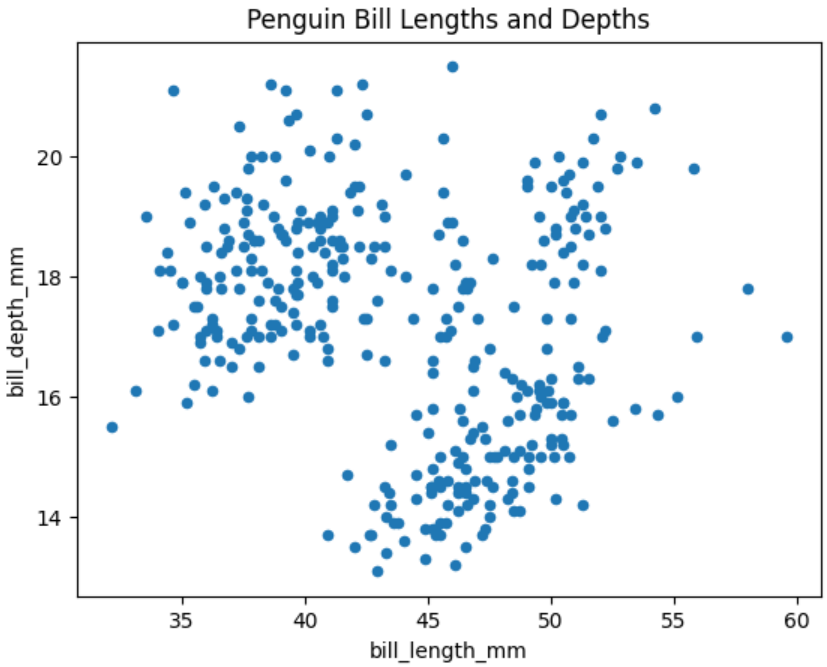


The redline in each scatter plot represents the average flipper length of 190mm. Torgersen has the most penguins to the right of the average length.

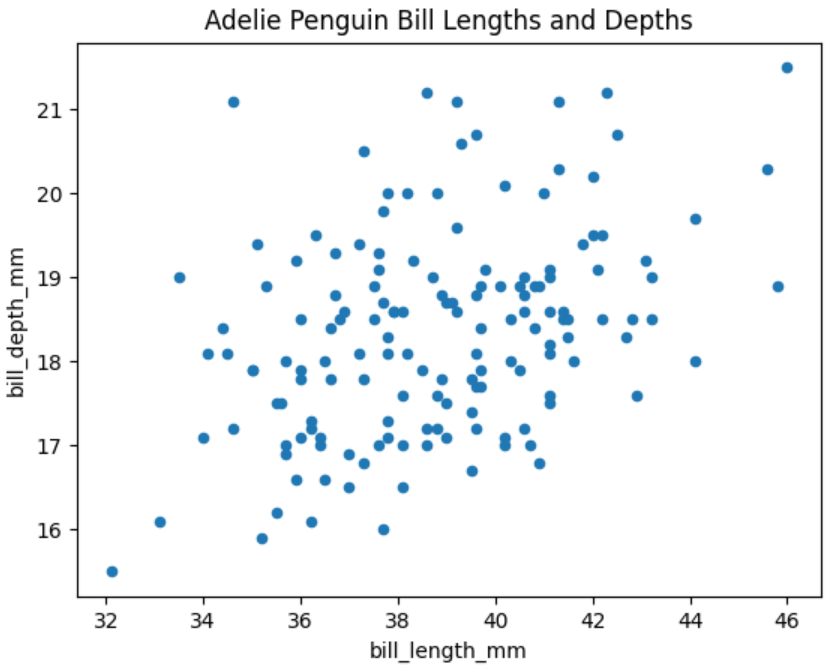
Question 2:

“Is there a relationship (correlation) between bill length and depth? If so, what kind of relationship is it? (ie. direct correlation or indirect). If not, do any factors correlate to either?”

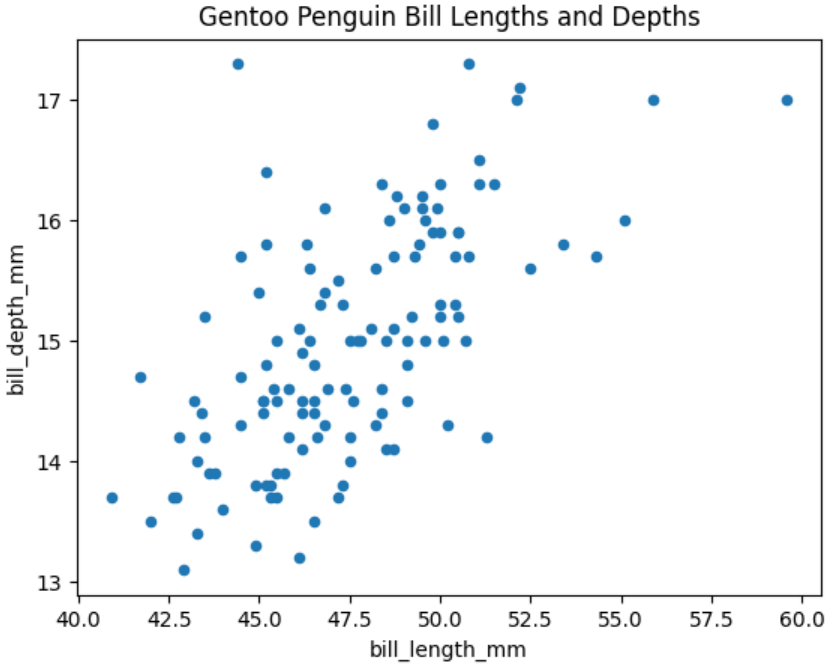
To answer we start by creating a scatterplot of all penguins with bill length as the x value, and bill depth as the y:



While the bill lengths do not appear to have either a positive or negative correlation to each other, as there is no clear trendline. However, one can see three clusters of similar bill length and depths, representing the three species in the dataset. Let's try isolating each species to view it in greater detail.

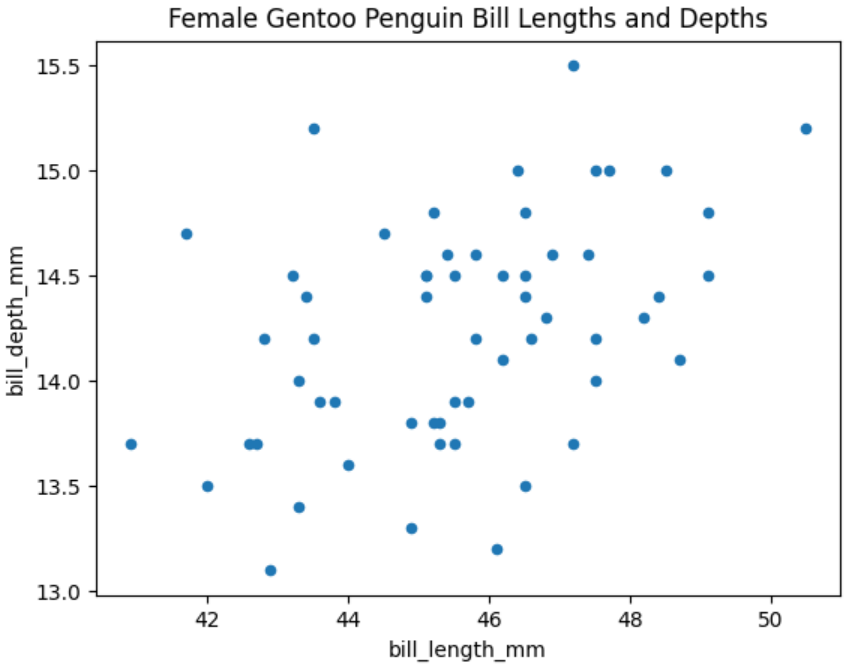


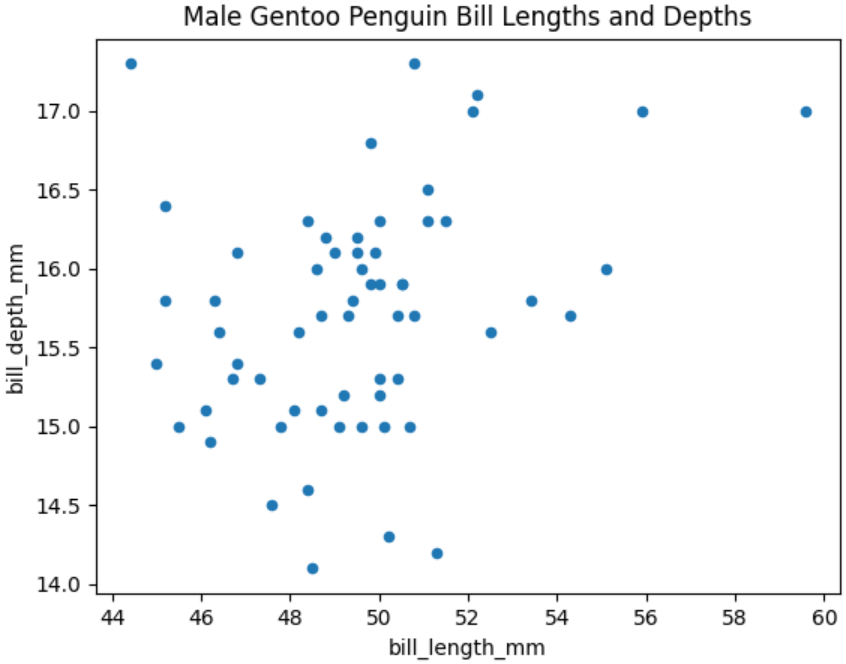
Adelie penguin bills are usually between 16.5 and 19.5 mm deep and between 35 and 42 mm long. There does not appear to be a correlation or trendline with the Adelie penguins, just a general range.



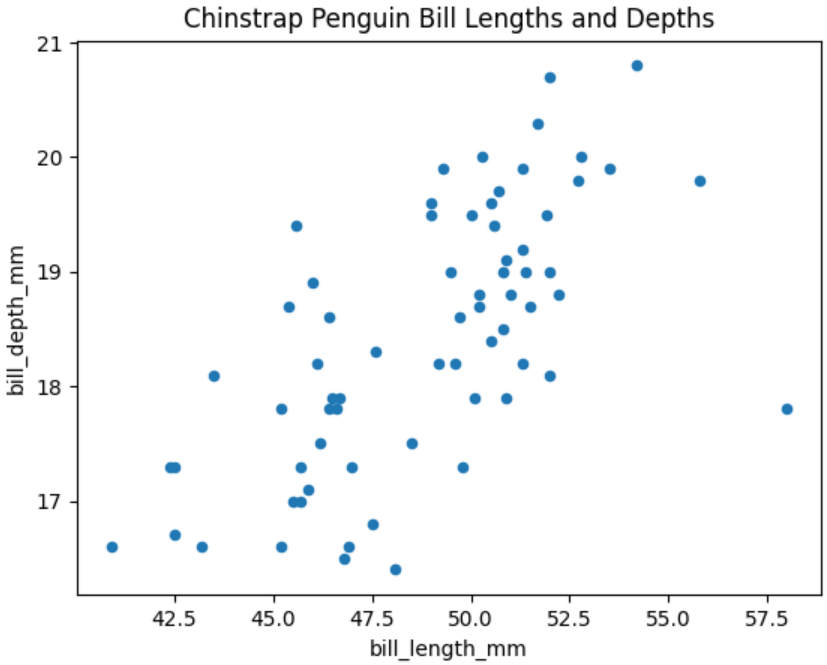
In Gentoos however, one can see a clear trendline, the longer the bill, the deeper it is. Most bills are between 42 and 51 mm long and between 13.5 and 16.5 mm deep, with a positive

correlation between the two. In fact, it almost looks as if there are two trendlines. If we graph male and female gentoos separately, we can see both genders have their own trendlines, with the males being slightly deeper (higher on the y-axis which is bill depth)
than the females



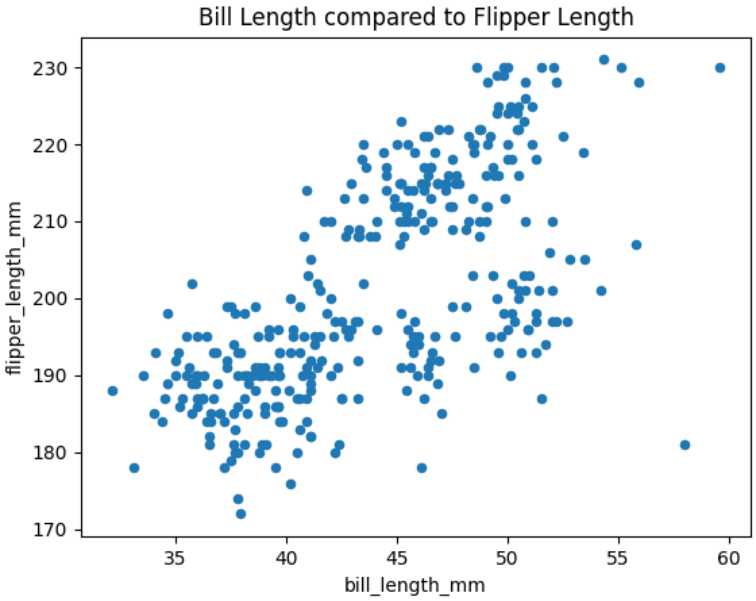


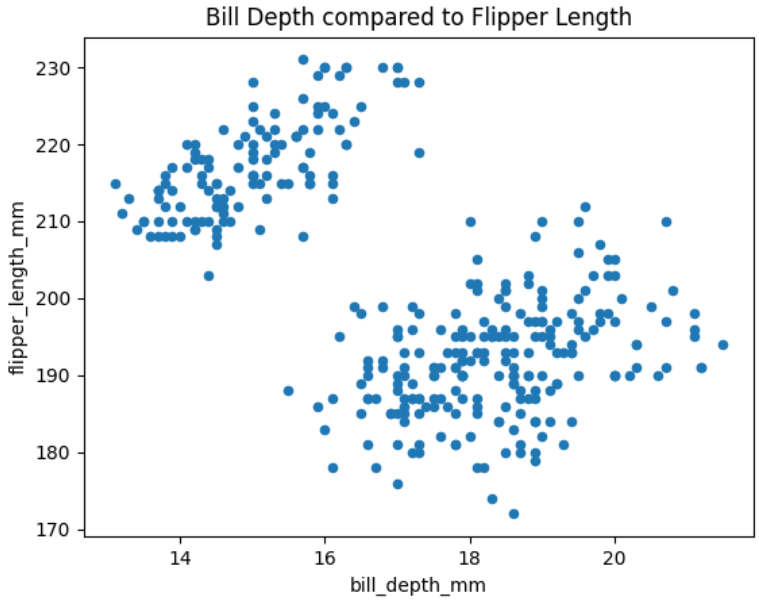
Moving on to chinstraps:



Chinstrap Penguins have their own trendline, with most bills being between 45 and 52 mm long, and 17 and 20 mm deep. There does not appear to be two separate trend lines between sexes though. This may be a result of the Chinstraps having a smaller population

While I did not go into further detail asking what other factors correlate to bill length and depth as they do appear to have a relationship, I created *_calculate_median_bill_lengths_lengths()* and *_calculate_median_flipper_lengths()* to test if flipper length affects either.





Both bill length and depth seem to be correlated to flipper length, and exhibit the clumping behavior as seen in the all penguins length and depth graph,

This gives evidence for other penguin characteristics having a correlation between bill length and depth.

Conclusion: Bill length and depth are positively correlated in gentoo and chinstrap penguins, with gentoos having two separate trend lines for each gender, while no direct correlation can be found for the Adelies. Each species has its own general range of bill length and depth. Flipper length appears to have direct, positive correlations between bill length and bill depth. While age is not recorded within the dataset, I suspect it has a correlation to bill length and depth as penguin bills grow with age.

This concludes the data analysis presentation. Feel free to look through the files in the practice_files folder to see how I developed the presentation code.

References and Links

- Project Github Repository - https://github.com/Zachary-P-Newby/data_analysis_project
- Palmer Penguin Dataset - https://www.kaggle.com/datasets/ashkhagan/palmer-penguins-datasetalternative-iris-dataset/
- Ashwani Rathee Kaggle Profile - https://www.kaggle.com/ashkhagan
- Pandas Docs - https://pandas.pydata.org/docs/index.html#
- Matplotlib docs - https://matplotlib.org/stable/