**Palmer Penguins dataset analysis report**

For this assignment analysis of the public Palmer Penguin dataset was used. The data contains information of three species of Antarctic Penguins: Adelie, Chinstrap and Gentoo. The data was collected from 3 different islands in Antarctica by Dr Kristen Gorman and Palmer Station, Antarctica LTER.

After importing the data and viewing the raw data, the correlational relationships of the data was assessed using seaborn.heatmap() and seaborn.pairplot() (See Figure 1). The heatmap was useful to understand the strength of the correlations and the pair plot was useful to assess these correlations between the different species. Based on these plots it was obvious that year was not strongly correlated with any other variables. The remaining variables however all seem to have a relatively strong correlation with one another. Of these correlations bill\_depth\_mm vs bill\_length\_mm had the weakest correlation and flipper\_length\_mm vs body\_mass\_g appeared to have the strongest correlation. These correlations can start to give us an idea of the morphology of the penguins. For example, the strong correlation of flipper length with body mass infers that heavier penguins require longer flippers to support them. There also appears to be a relatively strong correlation between flipper\_length\_mm vs bill\_depth\_mm or bill\_length\_mm; however, the correlation with flipper length and bill\_length is positive, whereas flipper\_length and bill\_depth is negative. The direction and strength of these correlations are also the same for body mass vs bill\_length or bill\_depth. This is not surprising as given the strong correlation between flipper\_length and body mass. From these additional correlations we can build an even better idea of the penguin’s morphology. Heavier penguins, have longer flippers and longer bills, whereas lighter penguins have shorter flippers and deeper bills. Whilst this information is useful to understand the basic expectation of a penguin’s morphology it does not account for the different species of the penguin. Therefore, I decided to create a Nearest Neighbour model to see if any of the morphological characteristics within this dataset could be used to distinguish between the three species of penguins within this data.

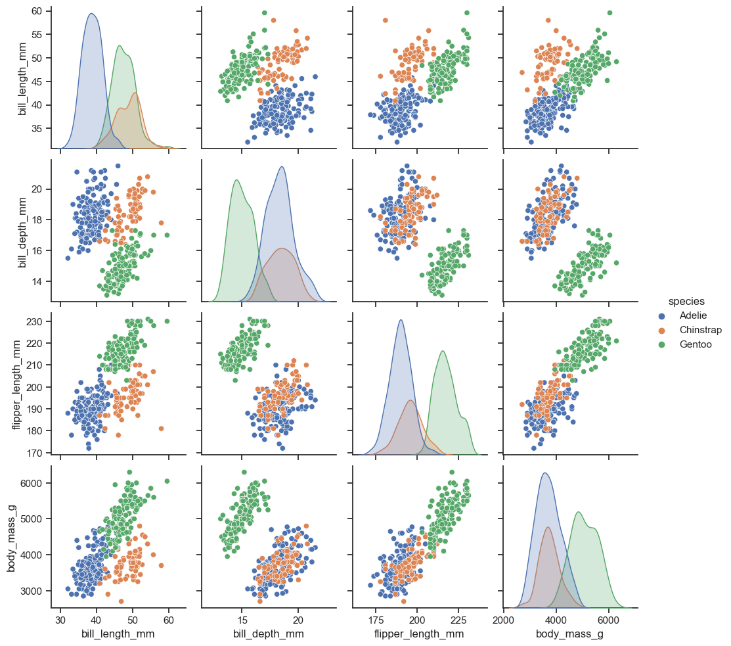
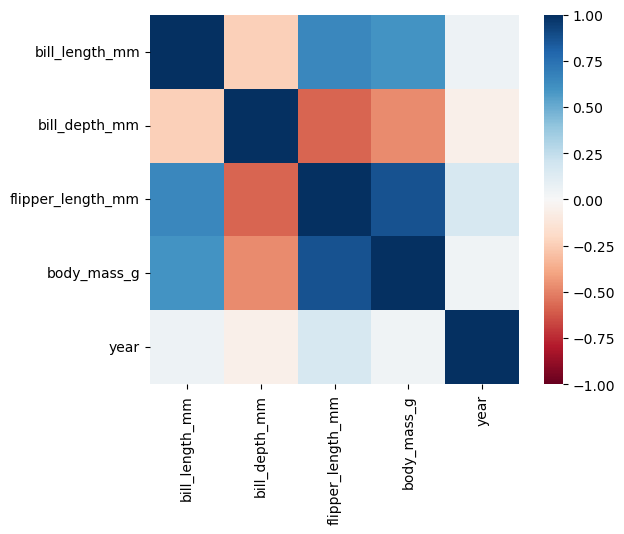


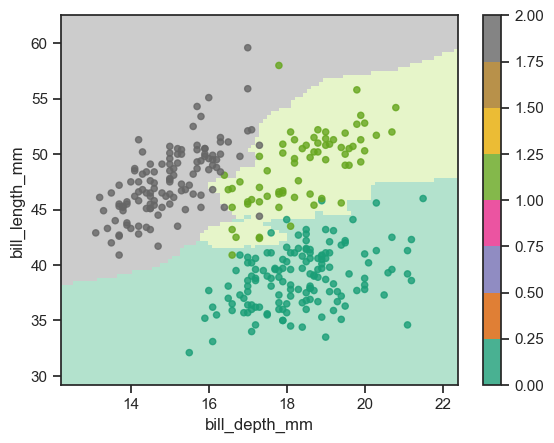
Figure 1: Correlation analysis of Palmer penguin data set. A: heatmap produced using seaborn.heatmap(). B: pairs plot produced using seaborn.pairplot()

A

B

To create a nearest neighbour model, data with minimal correlation and overlap are required. Based on the seaborn pairs plot produced I decided that bill depth (mm) and bill length (mm) where suitable for this model. Using a K value of 10 the model produced had a 0.94 confidence. Whilst this is a good score I decided to try and refine my model using the GridSearchCv function. This showed that to get the highest mean confidence of 0.97 K should equal 3. However, because I used a random state of 42 to create my models (in order to create reproducible results) using K=3 in the model actually reduced the accuracy of the model compared to K=10 by 0.01. In light of this with more time, the best way to find the most accurate model would be to test different K values on the efficacy of the model individually. Additionally, I also plotted the best outcome of the GridSearchCv which gave a model score of 1 (See Figure 2). This in theory is a 'perfect' model. However, as you can see the model does not even predict all of the current data points correctly (some of the green data points are in the grey background etc)! Thus, demonstrating that even the most effect nearest neighbour models are not 100% reliable. Therefore, with more time it would have been interesting to assess the confidence of this model for individual data points i.e., looking as data that falls within the middle of the predicted areas vs those the lie at the boundary edge. So that for each new penguin measured the confidence of the prediction could be more accurately determined.

Figure 2: Nearest Neighbour plot using the plot of highest confidence from found using the GridSearchCv function. Colour bar to the right indicates the species of the penguin 0 = Adelie, 1 = Chinstrap, 2 = Gentoo.



Further analysis that would be extremely interesting to conduct would involve statistically assessing the differences in each morphological characteristic within the data and seeing if this has any ecological relevance for each species. For example, as seen in this analysis, each species has a distinct beak morphology. It would be interesting to know if these differences were statistically significant and whether the ecology and lifestyles of the different species might be able to explain these differences in morphology.

For the full analysis written in Jupyter notebook please go to: