

**Report on**

**Penguin Datasets EDA & its Predictive Model**

**Course Title: Bigdata Analysis LAB**

**Course Code: CSE 4460**

**Submitted by:**

**Md Abrar Saief Safat, 203014020**

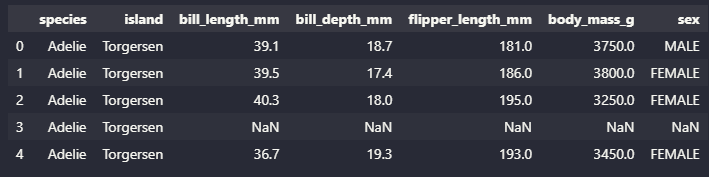
**Submitted to:**

**Fahim Morshed**

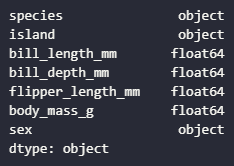
**Visiting Lecturer**

**School of Science & Engineering**

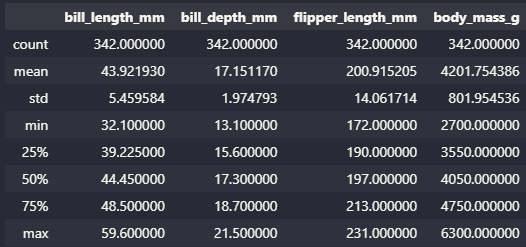
**Sample of dataset (first 5 rows):**

****

**Data types of all features:**

****

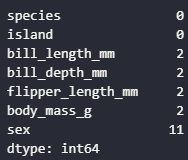
**Standard Deviation of all samples:**

****

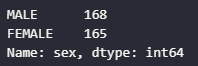
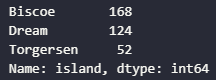
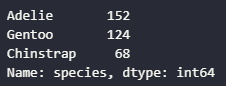
**Dimension of data:**

****

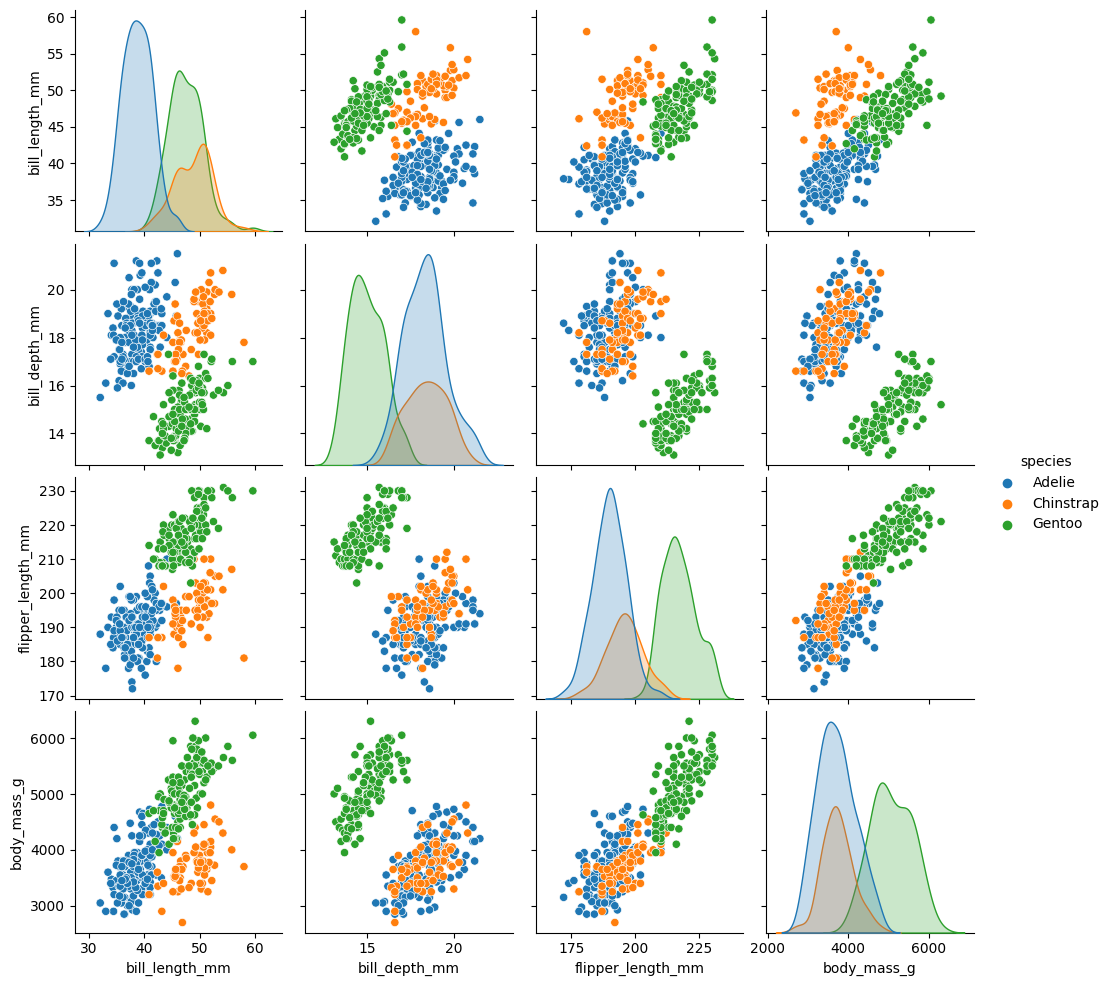
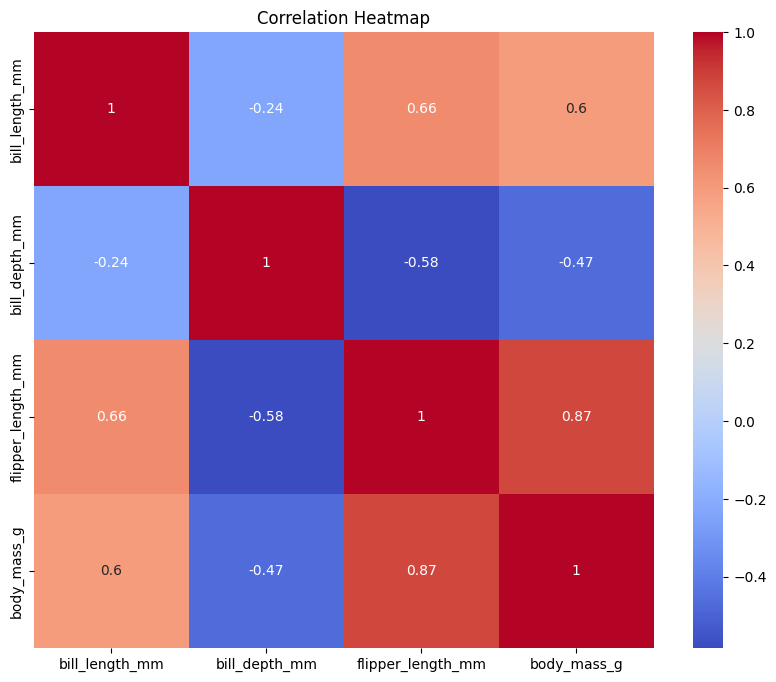
**Total null values in dataset:**

****

**Total feature counts for ‘sex’, ‘island’ and ‘species’ feature:**

**  **

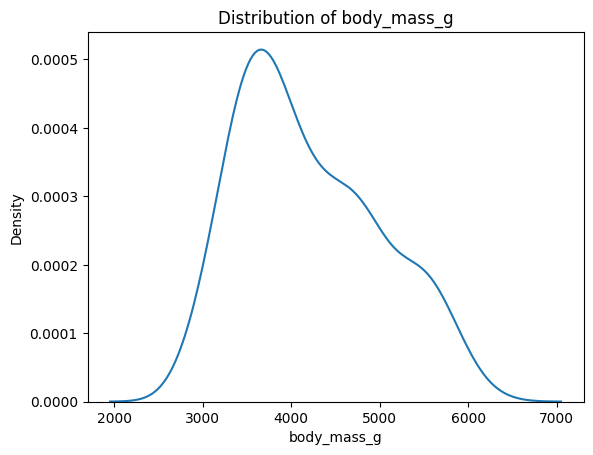
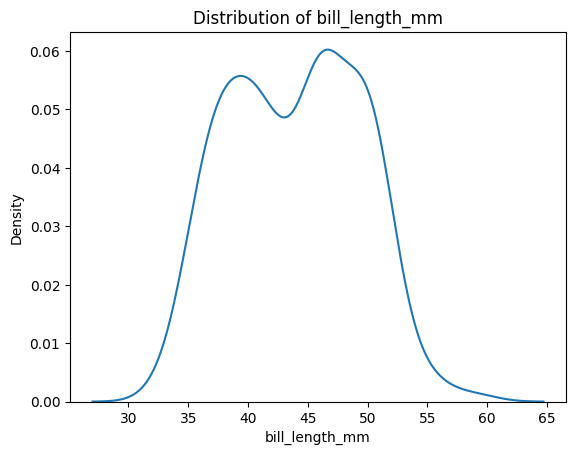
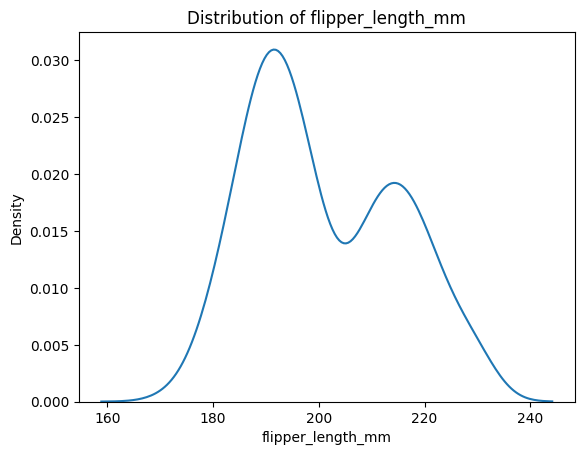
**Correlation Heatmap and Pair plot of all the features:**

****

**Conclusion from outputs:**

There is strong positive correlation between ‘flipper length’ and the ‘body mass’ of a penguin as well as a slightly less positive correlation between ‘bill length’ and ‘flipper length’ of a penguin. These outputs suggests that penguins with longer flippers tend to be heavier and the penguins with a deeper bill tend to have shorter flippers. Gentoo penguins, which cluster distinctly from Adelie and Chinstrap. Bill length, bill depth, flipper length and body mass contribute to the distinction of species. However, there is some overlap existing between Adelie and Chinstrap but Gentoo is clearly distinct.

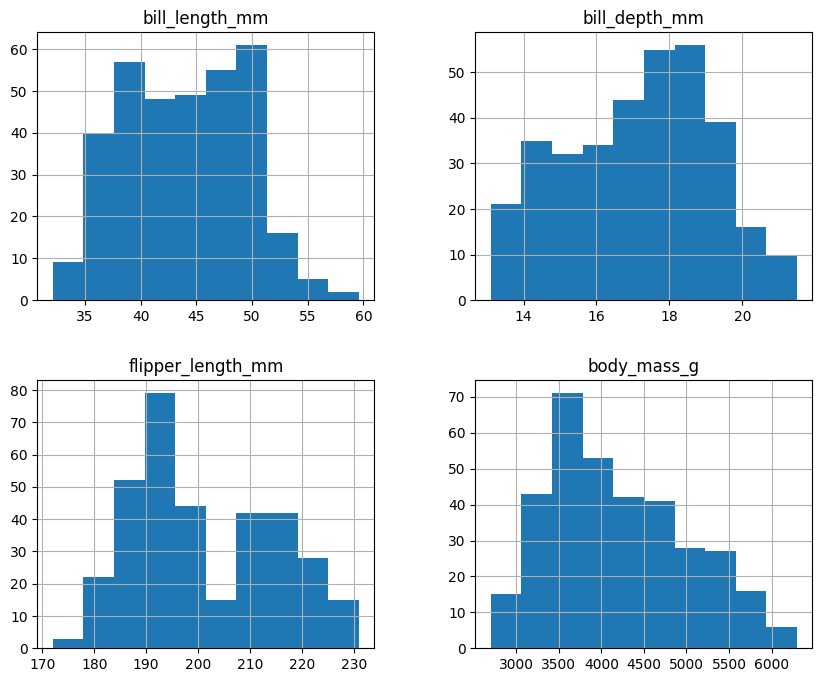
**KDE Plot: Distribution of ‘body mass’, ‘bill length’, and ‘flipper length’:**

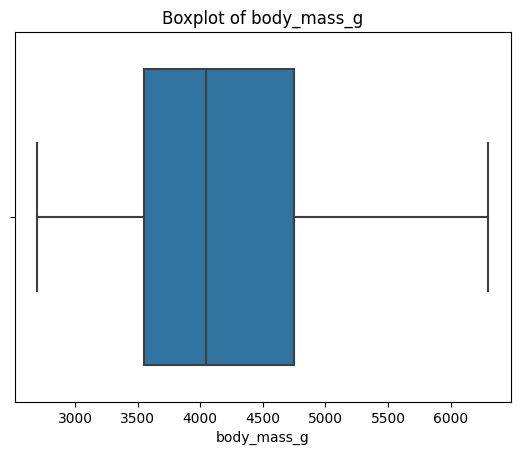
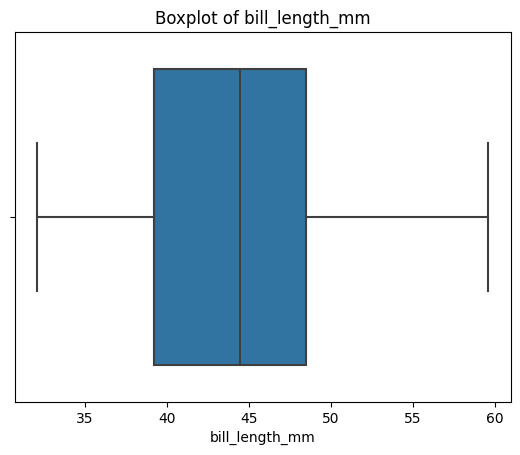
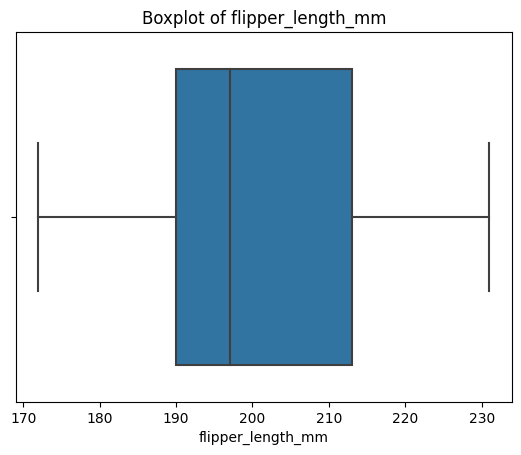


**Conclusion from outputs:**

For the ‘body mass’,the distribution is right skewed and most penguins have a body mass between 3kg and 5kg. For ‘bill length’, the distribution is bimodal, showing two different groups of bill length, which is likely due to different species and most bill lengths fall between 35mm and 55mm. For ‘flipper length’, the distribution is also bimodal, again suggesting two different species and most flipper lengths are between 180mm and 220mm.

**Histogram of all numeric features and boxplots:**

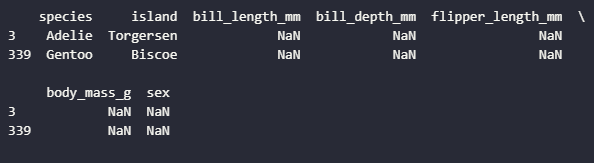




**Conclusion from outputs:**

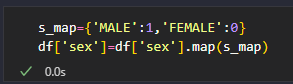
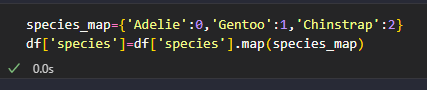
Every feature shows bimodal distinction that suggests two different groups/species. The data looks well distributed with no extreme outliers. The boxplots reinforce the outputs of no existing outliers.

**Missing samples identification:**



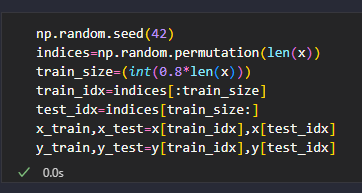
**Conclusion:**  
All the rows with more than 2 missing samples were removed as imputing them will distort the quality of the dataset, but the rows with 2 or less samples missing were imputed.

**Encoding:**



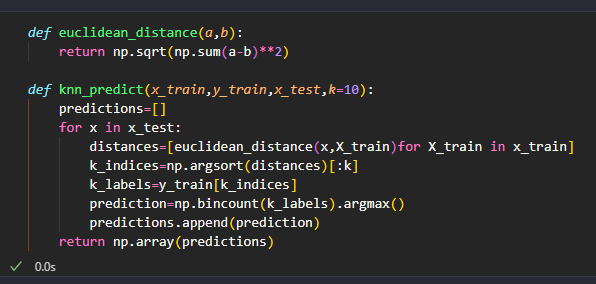
Preparation for down streaming data into the KNN, categorical encoding was applied for the species feature, binary encoding was done for the sex feature as well as one hot encoding for the island feature.

**Data splitting**



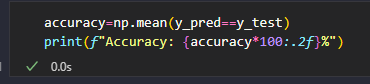
The data was split into: 80% for training and 20% for testing. This split proved to provide the best accuracy of the model.

**KNN Implementation and prediction**





**Evaluation:**

****

****

The accuracy is lower than expected. This may be due to some overlap existing between Adelie and Chinstrap as concluded from the pair plot.