

**Faculty Of Engineering And Technology**

**Department of Electrical and Computer Engineering**

**Intelligent Systems Lab (ENCS5141)**

**Case study #1**

Instructor: Dr. Mohammad Jubran

TA : Hanan Awwadeh

Section: 3

Made By:

Tariq Quraan 1190579

**November 2023**

# **Abstract**

This report aims to study and focus on the concepts of data engineering like data cleaning & wrangling, data integration, data transformation, and Exploratory Data Analysis to understand the data and to achieve the best data that suits the model with the desired output with the best accuracy. Also, to train a model with the original data and data after enhancement and to compare the results to observe the effect of the enhancement on the data. The data will be enhanced by replacing the noisy and null values with a new value that is more sensible and all the null rows in the data will be removed which will eventually enhance the model performance and accuracy.

# Introduction

## Background

Data engineering is one of the challenges that faces the programmers when dealing with machine learning models. It is the science of studying, understanding and modifying the data in a way that aims to improve the machine learning model’s performance. There are many sectors in data science which are data visualization, data cleaning, feature engineering, …etc.

### Data Visualization

Data visualization is the process of transforming data into visual representations, such as charts, graphs, and maps. It is a powerful tool that can be used to communicate information clearly and concisely.[1] Data visualization is used to explore and understand the data and to show the features and their distribution. Furthermore, it can show the outliers and the minimum and maximum values of each feature. Some data can not be explained for stakeholders but when shown in graphs generated by data visualization techniques, things can be explained and will become more clear. Another main use of data visualization is validation, where machine learning model outputs are plotted to observe the target values of the model then validate the output and adjust based on the validation.

### Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.[2] By using the Descriptive Statistics,[1][3] this will help in identify missing values, outliers, and inconsistencies that need to be addressed then replace them with new values like mean/ mode/ median to insure the best data for the model.

### Feature Selection

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.[4] The goal of feature selection is to improve the performance of machine learning models by reducing the dimensionality of the data, eliminating irrelevant or redundant features, and enhancing model interpretability and generalization.[1] There are many methods for feature selection which are Wrapper methods (forward, backward, and stepwise selection), Filter methods (ANOVA, Pearson correlation, variance thresholding), and Embedded methods (Lasso, Ridge, Decision Tree).[5] Filtering methods are methods that removes some features based on a factor. One famous filtering method is variance filtering, another one is the Chi-squared test which relies on the dependency between two features. There are more methods but these are the most popular ones.

### Data Transformation

the process of converting, cleansing, and structuring data into a usable format that can be analyzed to support decision making processes, and to propel the growth of an organization. Data transformation is used when data needs to be converted to match that of the destination system.[6] For example, some machine learning models require the data to be numerical instead of being categorical, so a data transformation technique called Encoding can be used. Another data transformation technique is scaling (normalization) where the numeric data can be changed with new values based on a scale without changing their relative relationship. This can help alot with algorithms that are sensitive to large magnitude features. The last data transformation technique is discretization which is the operation of converting the continuous numerical features to categorical values which can be used in models that require categorical values.

## context of the case study

The case study requires applying the concepts of data engineering on a dataset called the Penguins dataset. First the data should be displayed to check out the features and the type of each column. Then, the data cleaning methods should be applied to get rid of empty records and deal with outliers and null columns. After that, the data should be prepared for the machine learning algorithm, the features should be minimized so the features selection methods should be applied to select the features that have the best effect on the target values. Also, the categorical value should be encoded to numerical values so they can fit in the random forest algorithm which takes no categorical features. The data should be passed to a scaling method to ensure the best consistent scaling across variables. The data should be split to training data and testing data so it can be passed to the model for evaluating. Finally the dimensionality of the data should be reduced more by applying the suitable dimensionality reduction technique. The data should be evaluated before and after reprocessing to compare the results and see the effect of preprocessing techniques to ensure that the performance has gotten better.

## Dataset

The Penguins dataset is a dataset that is provided from seaborn library. It contains many features that can be studied and analyzed. It has 3 penguin types which are Adelie, Gentoo and Chinstrap that are living on 3 different islands which areDream, Biscoe, and Torgersen. The penguins dataset has another categorical feature which is sex that has the values of Male and Female. The dataset has numerical columns which are bill\_length\_mm, bill\_depth\_mm, flipper\_length\_mm and body\_mass\_g. These numerical values have a big role in the training of the model.

## Main Objectives

* To observe and explore the dataset by printing it then plotting the features with the right data visualization techniques.
* To check for null records and remove them from the dataset.
* To deal with outliers and missing values and replace them with the right strategy of replacement.
* To drop the features with the lowest effect from the dataset to reduce the models work.
* To encode the categorical values and convert them to numerical values.
* To normalize the data by using scalability techniques.
* To use a feature reduction method and minimize the number of features.
* To evaluate the model with the testing data before and after data reprocessing to ensure the increase in performance.

# Procedure & Discussion

The goal of this experiment is to train a machine learning model and evaluate the penguins dataset after being prepared with data engineering methods. There are many steps that were taken before passing the data to the model which will be shown discussed in details below.

## Loading The Dataset

The data set was loaded from the seaborn library to a dataframe and printed. Check figure 1.

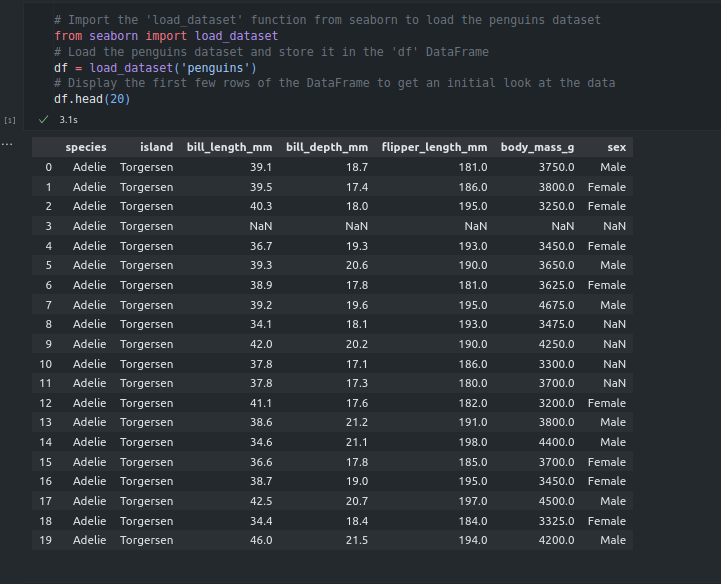


Figure 1: loading the dataset

As observed in figure 1, The penguins dataset was loaded, It has 7 columns where each column represents a feature. But some values were NaN which will be resolved in the upcoming steps.

## Data Exploration

The data was plotted using different visualization methods and the descriptive statistics were printed. Check figures 2, 3.

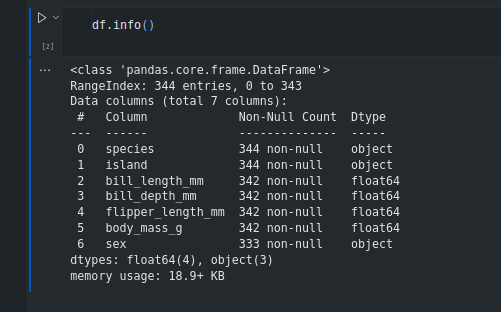


Figure 2: Displaying general info of the data

As observed in figure 2, by using df.info() the number of records will be displayed along with the number of non null values of each feature.

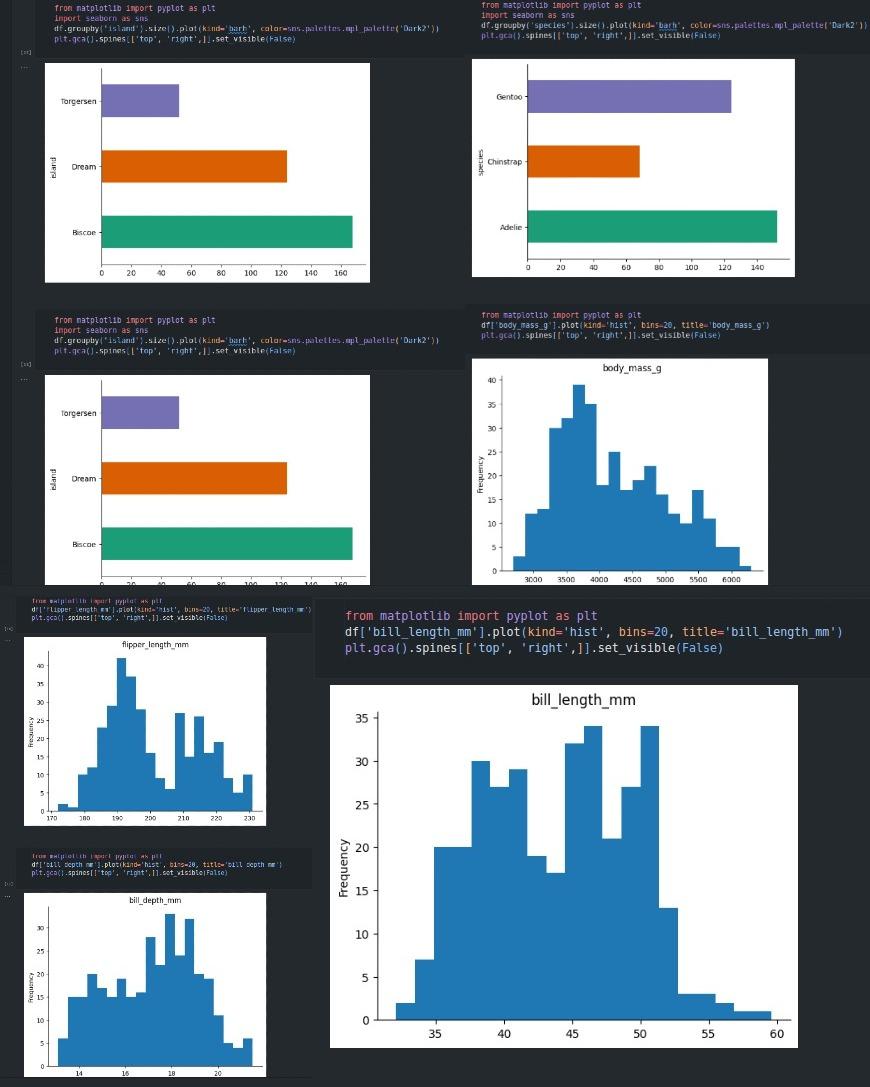


Figure 3: Displaying general info of the data

As shown in figure 3, the distribution of some of the features was shown in different types of figures that explains the ranges of each feature and how the data is distributed.

## Addressing & Resolving Quality Issues

In figure 4 below, the function .isnull() was used to return the records with one or more null values in its row. The .sum() function counts the values that were returned from the .isnull() function.

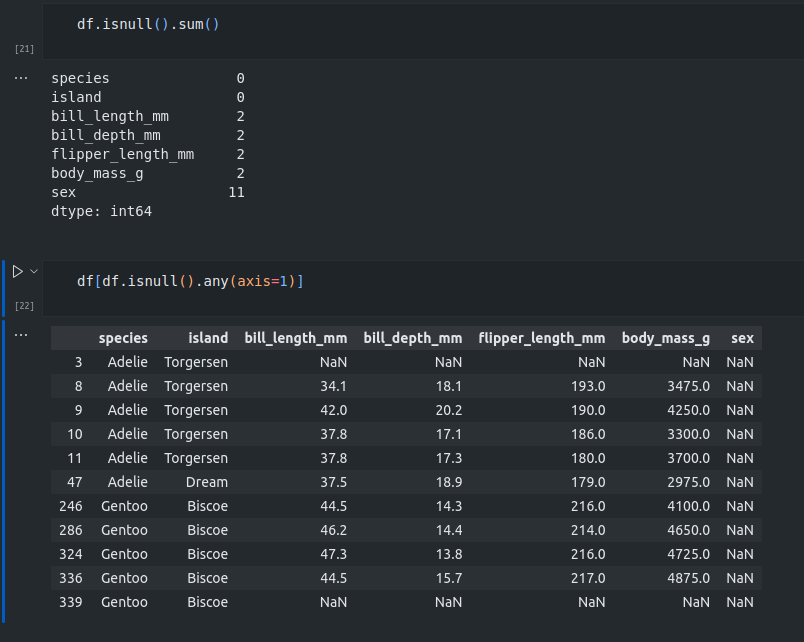


Figure 4: Displaying the null records

As observed in figure 4, the count of the null values was addressed along with the records that hold any null values. The sex attribute has the most null values as there are 11 records with a null sex feature.

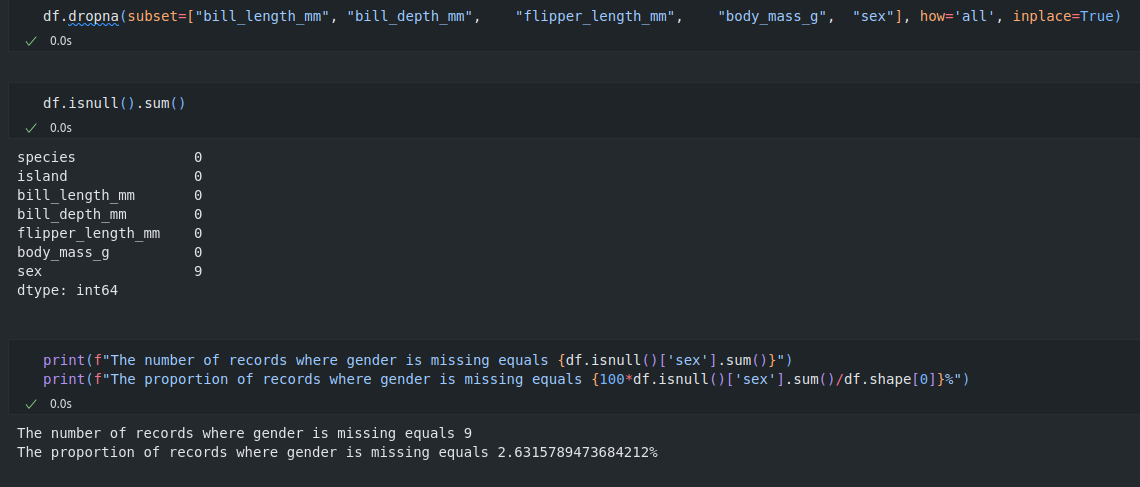


Figure 5: Displaying the null records

In figure 5, the records that have almost all the features empty were dropped because they won't affect the model’s performance. Two records were dropped from the dataset. But the number of records that have the sex feature as null is still more than zero. The data was splitted to thirds based on the island to ensure the similarity in features in each third. Check figure 6.

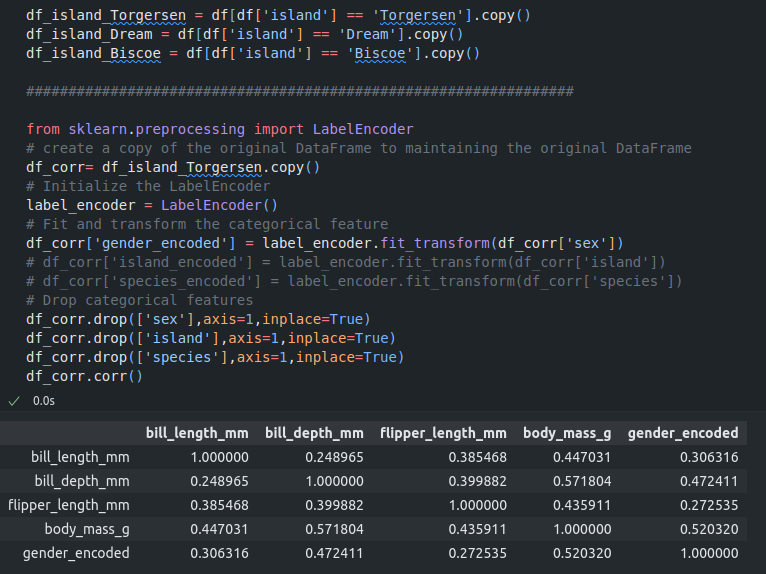


Figure 6: Splitting The data based on island to thirds

As shown in figure 6, the correlation between the features of the Torgersen third was displayed after encoding the sex feature and dropping the other categorical features. The body\_mass\_g feature shows the highest correlation with the gender\_encoded feature. Note that the categorical values were not dropped from the original Torgersen dataframe but from a copied one. A graph was plotted to show the relation between gender and body mass attributes for the Torgersen third of the data. Check figure 7.

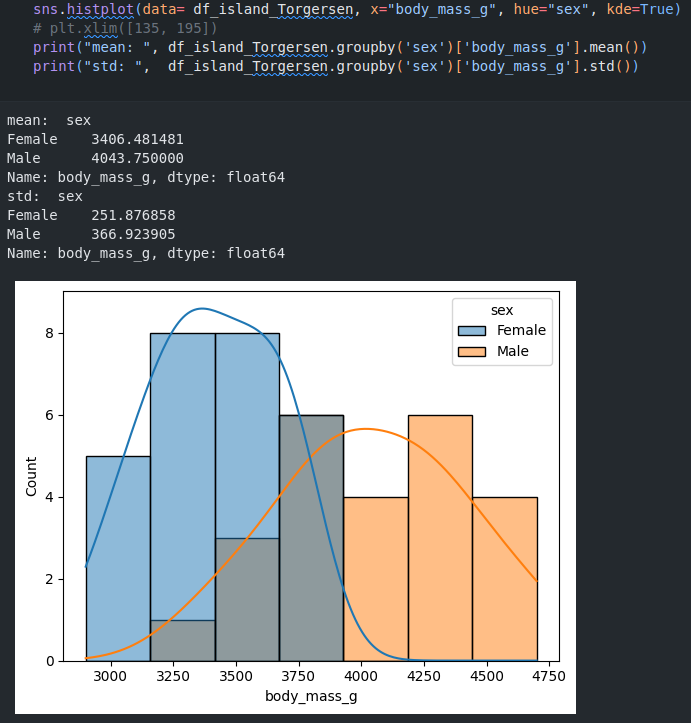


Figure 7: body\_mass\_g and sex relation graph for the Torgersen third

As shown in figure 7, the males have a higher body mass than females in general. After almost 3800 grams the majority gender are males. However, below 3800 grams the majority gender are females.

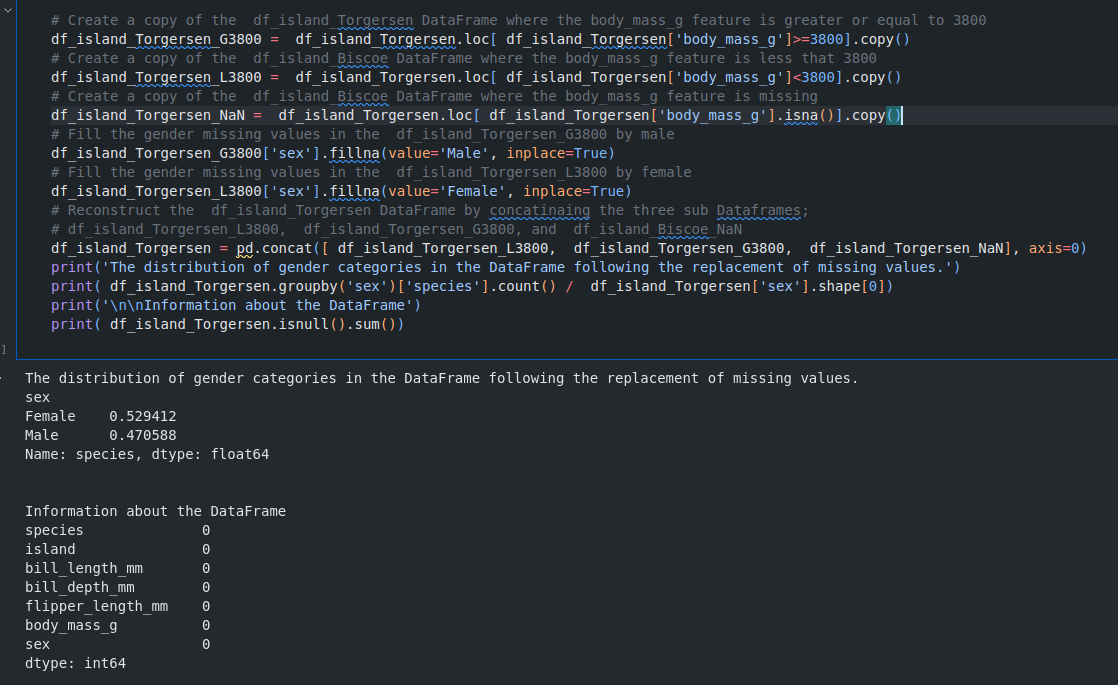


Figure 8: Replacing missing values for the Torgersen third

In figure 8, the Torgersen third of data was splitted for another 3 dataframes. One is the data that have the body\_mass\_g feature more than or equals 3800. Data that has it under 3800. Data that has the feature as NaN (which are non in this case). The NaN third sex’s values were replaced based on their body\_mass\_g where the penguins that had the body mass higher or equal 3800 were considered as males, others were considered as females. The three thirds were merged back to become one Torgersen third. The sex’s distribution was printed at the end of the replacement operation for this third. The other two thirds were treated the same, but the Dream island’s third had the highest correlation between gender and the bill\_depth\_mm attribute and the null gender records were replaced using the same method as the Torgersen third but according to bill\_depth\_mm instead of the body mass. Check figure 9.

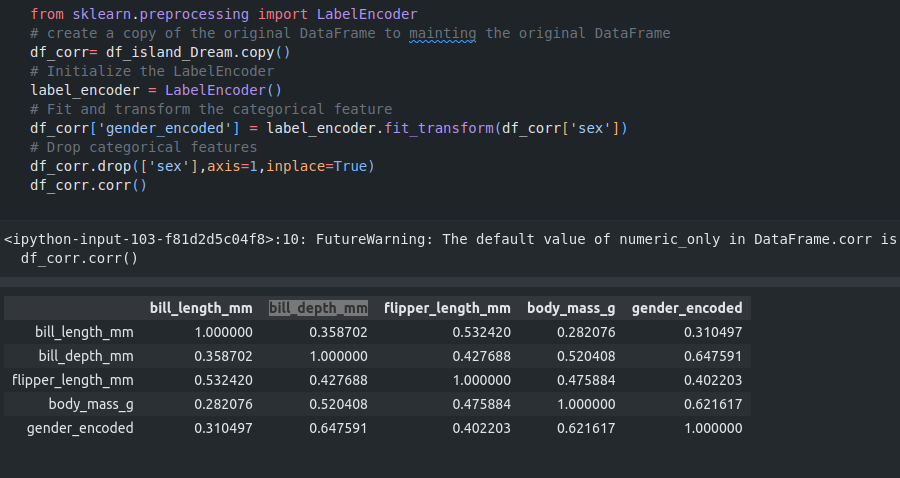


Figure 9: Correlation between features for Dream’s third

As observed, despite the high correlation between gender and body mass, the bill depth feature had the highest correlation with the gender feature. The replacement of the null values of the gender were done based on the bill depth. Check figure 10.

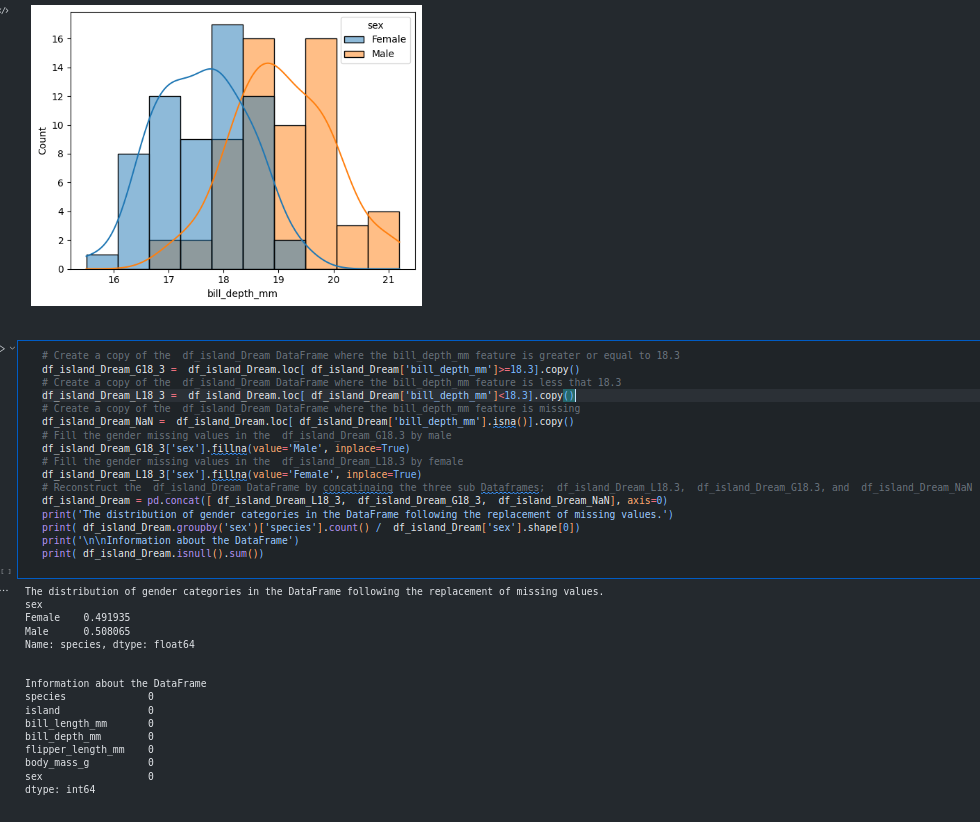


Figure 10: The replacement of empty records of the gender

As observed in figure 10, the dream’s third of the data has no more null values after replacing the null based on bill depth attribute.

After replacing the null records, the outliers of each third should be dealt with using the appropriate method. After displaying the distribution of each feature using the boxplot() visualizing method, the Torgersen third showed that the flipper\_length\_mm feature has one outlier that should be resolved. Check figure 11.

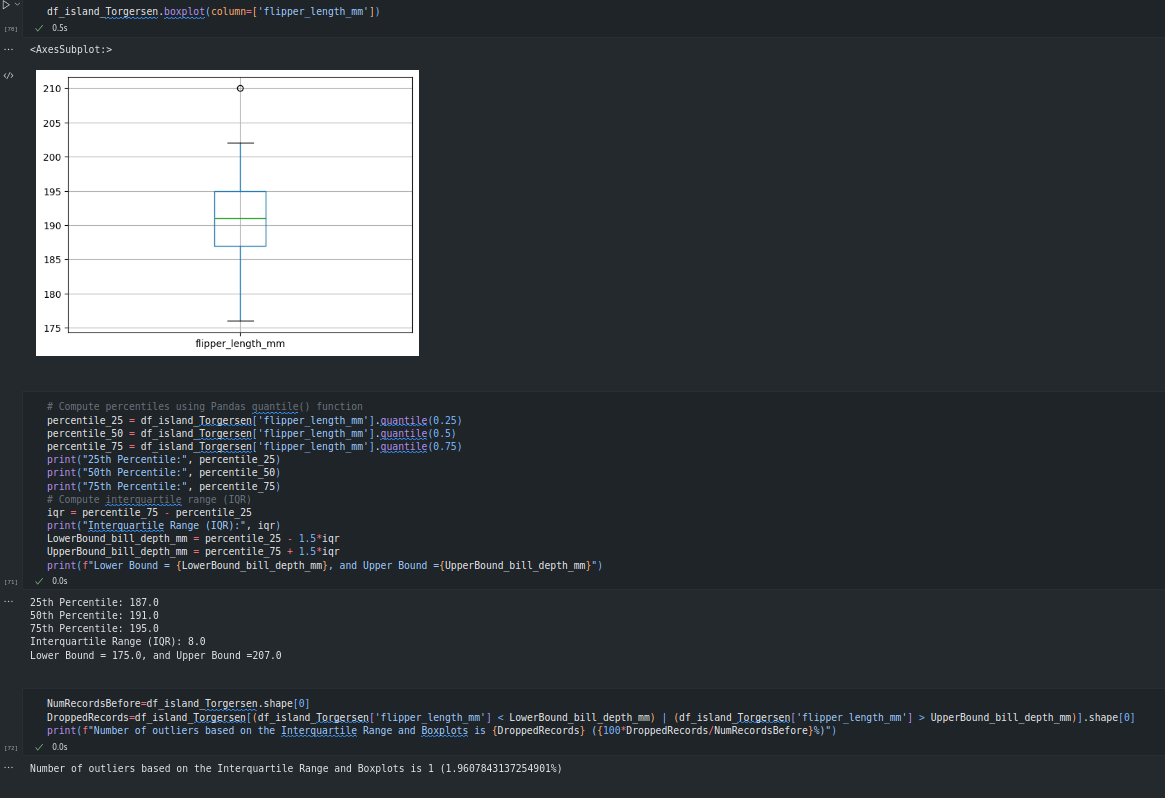


Figure 11: The outliers of flipper\_length\_mm

As shown in figure 11, the maximum value is more than the upper bound which is 1.5\*interquartile range. Which made it an outlier that should be resolved.

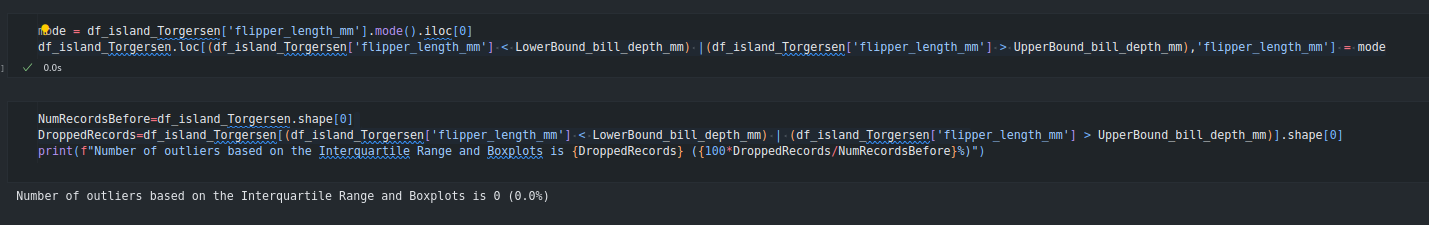


Figure 12: Replacement of outliers for flipper\_length\_mm attribute for Torgersen’s third

As shown in figure 12, the outlier value was replaced with the mode to reduce the effect of the new value and to minimize the change in the median and the mode of this feature. The outliers of other features for the three islands were dealt with by using the same method.

After dealing with the outliers, the 3 thirds were merged back to one dataframe called df for further analysis as shown in figure 13.

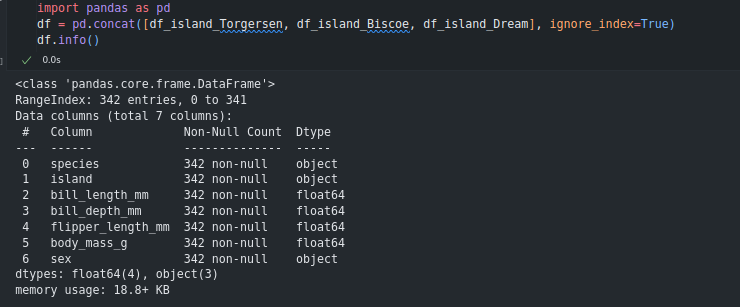


Figure 13: Merging the data back to one dataframe

## 4. Encoding Categorical Values

The categorical values should be encoded and converted to numerical values which helps in feature selection techniques and later on the model training.

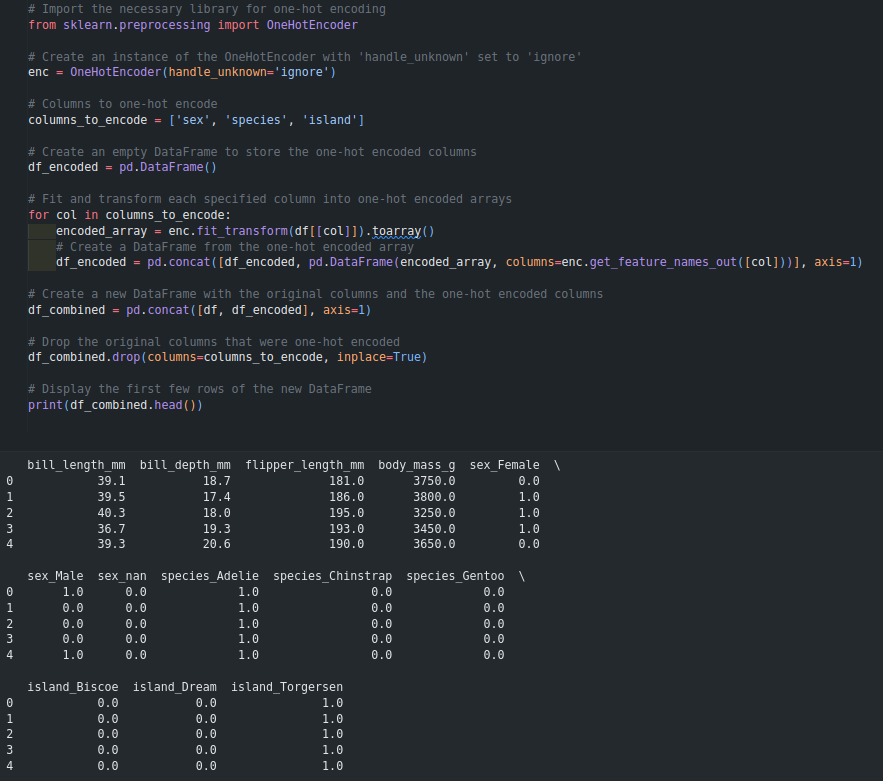


Figure 14: Encoding categorical features

As shown in figure 14, the categorical features were encoded using one hot encoding then later displayed.

## 5. Feature Selection

Feature selection is important to minimize the features by removing the ones with the lowest effect on the target value.

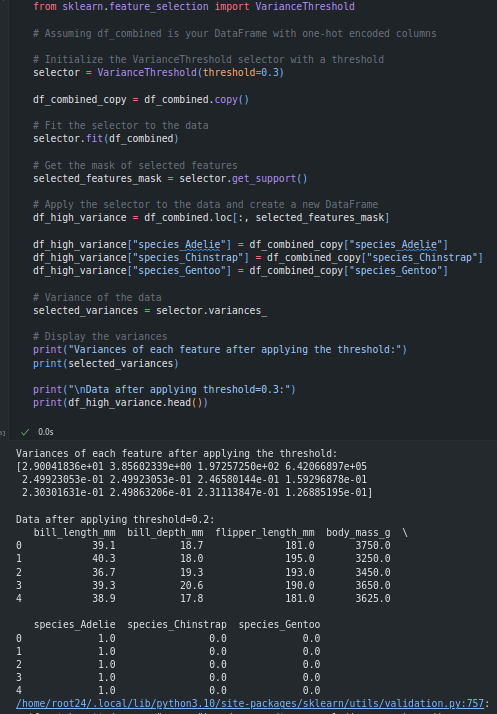


Figure 15: Applying Feature Selection with variance threshold

As shown in figure 15, a threshold of 0.3 was applied on the data and the features with the lowest variance were eliminated.

## 6. Normalizing the data

Normalization was applied to reduce the effect of the high values on the model.

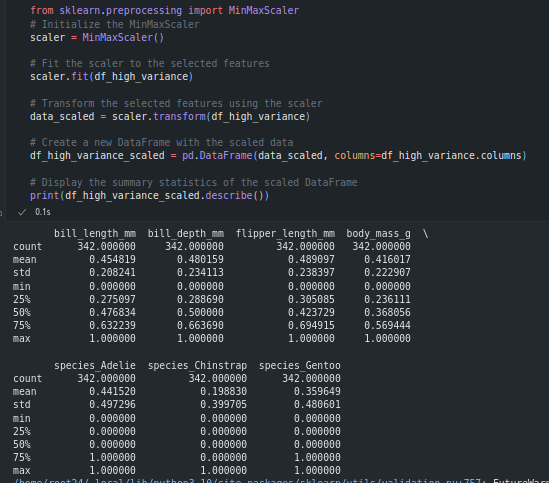


Figure 16: Scaling the data

As observed in figure 16, the data was normalized using MinMaxScaler(), the data description was printed afterwards to show the effect on scaling.

## 7. Reduce data dimensionality and evaluating

The dimensionality reduction is used to reduce the number of features without losing their effect by merging them into a smaller number of features. Which was done on the scaled data.

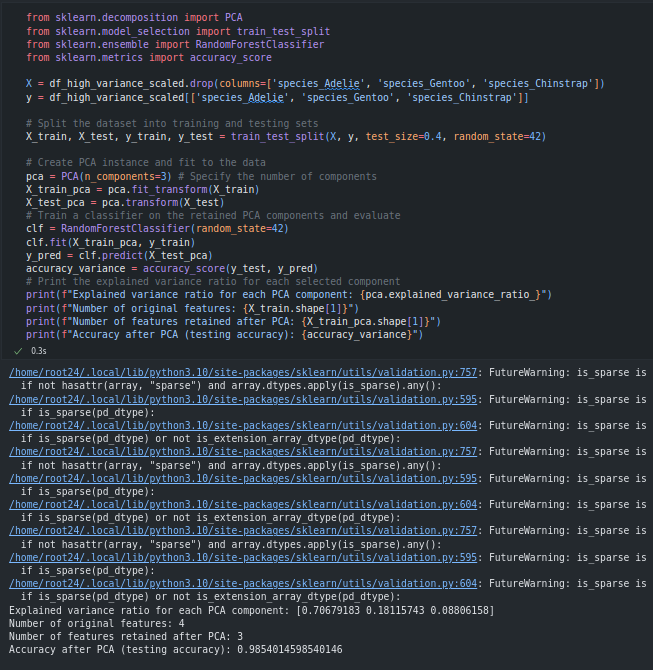


Figure 16: Scaling the data

As shown in figure 16, the data was splitted to training and testing sets after applying PCA data reduction technique and reducing the number of features to 3. By taking the species as target value. The model came up with an accuracy of 98.5%.

## 8. Evaluating the raw data

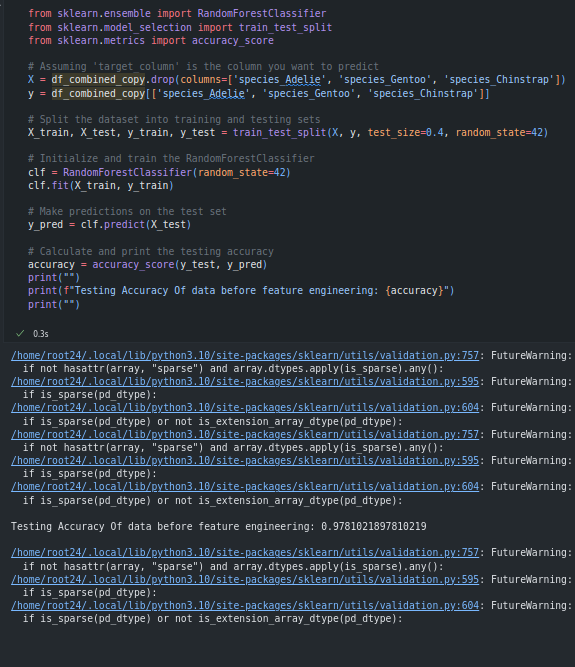


Figure 16: Evaluating the raw data

As observed, The raw data came up with a lower accuracy which is what was expected due to the enhancement of the data because of data reprocessing techniques that were applied.

# Conclusion

After applying data engineering methods on the penguins dataset, the data reprocessing techniques showed enhancement in the models accuracy which proved its effectiveness despite working on a dataset that does not have lots of features. The data was cleaned, the features were reduced and the accuracy was increased. All goals were accomplished, lots of lessons were learned. Data engineering methods were used and memorized due to the practical use of them on this dataset.

# **References**

1. The lab manual
2. [Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data](https://www.tableau.com/learn/articles/what-is-data-cleaning)
3. [Descriptive Statistics: Definition, Overview, Types, Example](https://www.investopedia.com/terms/d/descriptive_statistics.asp)
4. [Feature Selection In Machine Learning [2023 Edition] - Simplilearn](https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning#:~:text=ProgramExplore%20Program-,What%20is%20Feature%20Selection%3F,you%20are%20trying%20to%20solve).
5. [How to Choose a Feature Selection Method For Machine Learning - MachineLearningMastery.com](https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/)
6. [What is Data Transformation? | TIBCO Software](https://www.tibco.com/reference-center/what-is-data-transformation#:~:text=Data%20transformation%20is%20the%20process,that%20of%20the%20destination%20system)