Introduction to Data Science

Semester Project



Session: 2021-2025

**Submitted To:**

Dr. Irfan Yousaf

**Submitted by:**

Zain Ali

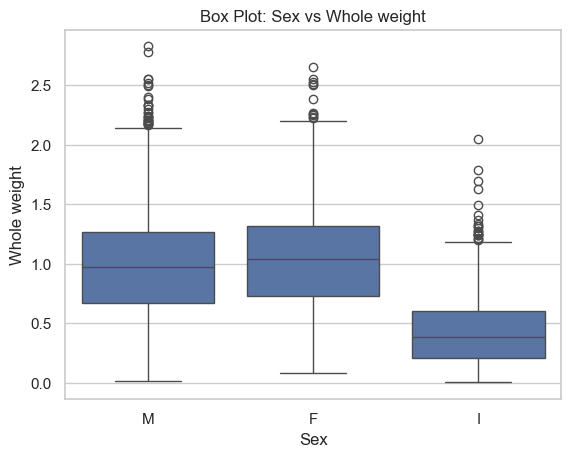
**Registration No:**

2021-SE-23

Department of Computer Science, New Campus **University of Engineering and Technology Lahore, Pakistan**

# Bivariate Analysis

## Box Plot



2021-SE-23

Figure 1

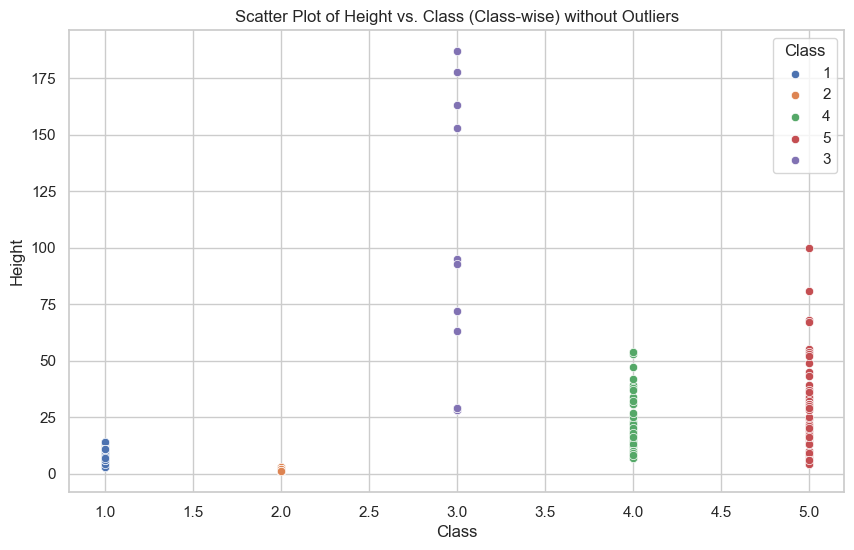
The box plot shows the distribution of whole weight for two groups, labeled "M" and "F". Without knowing more about the context of the data, it is impossible to say for sure what these groups represent, but they likely correspond to sex.

The median whole weight is higher for the "M" group than for the "F" group. The box extends further upward from the median for the "M" group than it does for the "F" group, which suggests that there may be more outliers (data points that fall outside the whiskers) in the upper tail of the distribution of whole weight for the "M" group.The interquartile range (IQR), the range that contains the middle 50% of the data, is larger for the "M" group than it is for the "F" group. This suggests that there is more variability in whole weight in the "M" group.

There are outliers in both groups. An outlier is a data point that falls outside the whiskers of the box plot. The whiskers extend to 1.5 times the IQR above and below the quartile.

Overall, the box plot suggests that there may be a sex difference in whole weight, with males tending to have higher whole weight than females. However, it is important to note that this is just a preliminary analysis and more data would be needed to draw any firm conclusions.

## Scatter Plot



2021-SE-23

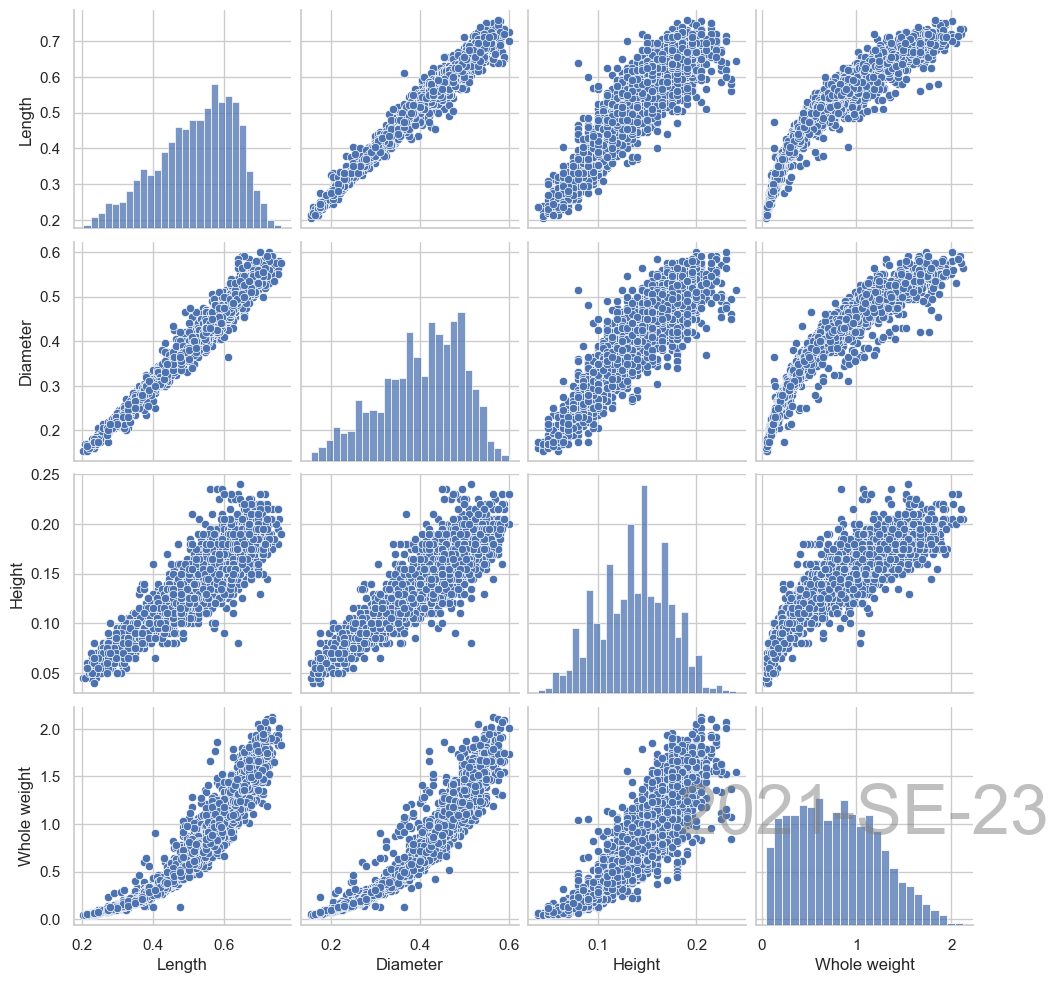
Figure 2

This scatter plot depicts the relationship between two variables. Each dot represents a single data point, with its horizontal position on the x-axis corresponding to the value of one variable and its vertical position on the y-axis indicating the value of the other. By examining the distribution of these dots, we can begin to uncover trends or patterns in how the two variables are related.

There does not appear to be a strong correlation between height and class. There is no clear linear relationship between the two variables. This means that we cannot predict the height of a student very well based on their class.

Overall this scatter plot reveals no strong association between height and class. There is no clear pattern in the distribution of the data points, suggesting that students’ heights are fairly evenly spread across classes.

# Multivariate Analysis

**Pair Plot - Multivariate**

2021-SE-23

Figure 3

This pairplot displays histograms for every variable on the diagonal in addition to scatter plots for each pair of variables (Length vs. Diameter, Length vs. Height, and Diameter vs. Height). It enables the simultaneous visualisation of relationships between several variables.

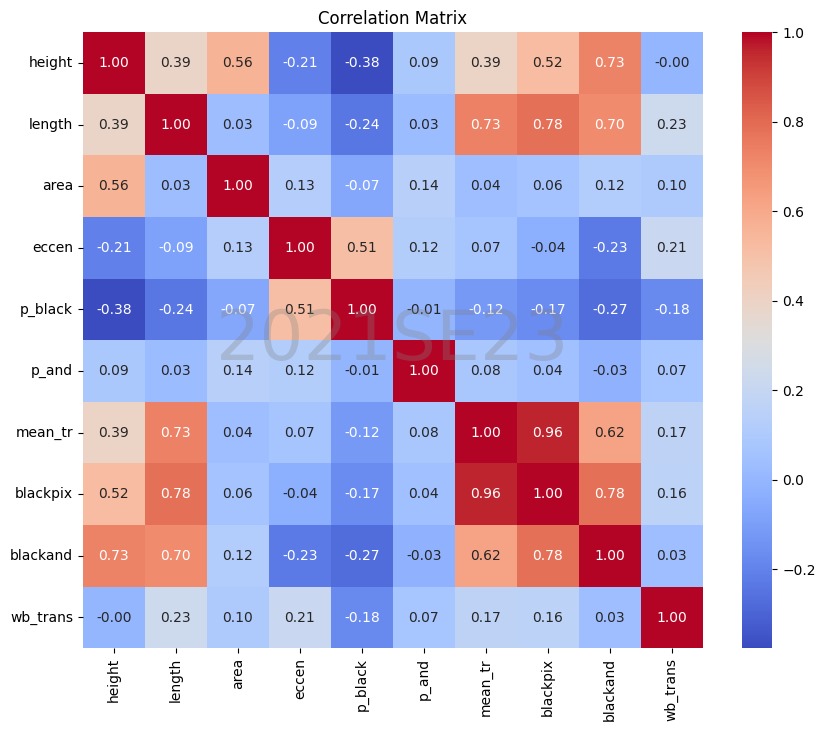
The correlations between variable pairs are revealed by the scatter plots.

For instance, a substantial positive association can be seen in the scatter plot of length vs. diameter, suggesting that an abalone's diameter tends to rise as its length does.

The distributions of the different variables are displayed in the histograms along the diagonal, revealing information about their spread and central tendency.

The relationship between pairs of characteristics is represented by offdiagonal elements, whilst the distribution of each feature is represented by diagonal elements.

## Heat Map



2021-SE-23

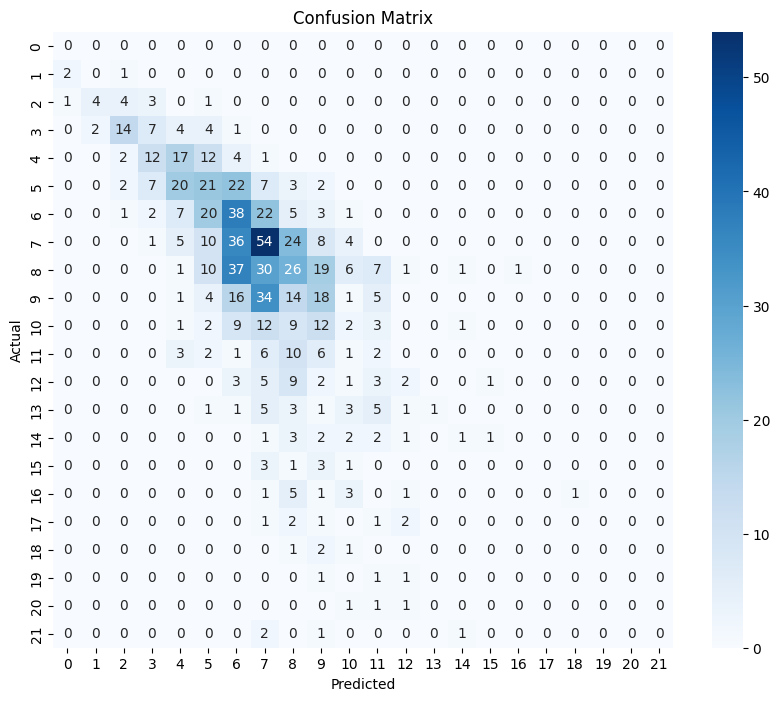
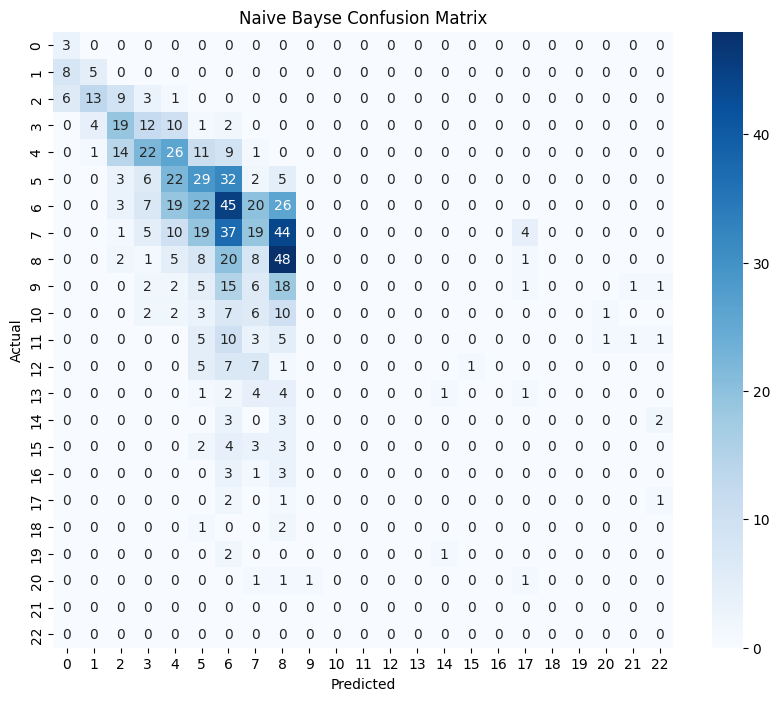
Figure 4

The heatmap displays the relationship between various dataset features that are used to classify blocks. The correlations are broken down as follows:

* Height, length, and area: There is a positive association between these three characteristics. Accordingly, blocks that are taller typically have longer and larger areas, and vice versa. Given that the area is determined by multiplying the height and length, this makes basic sense.
* Eccentricity: There is a slight positive association and a slight negative correlation between eccentricity and length. A block with a higher eccentricity value is longer than it is tall since eccentricity is defined as the length to height ratio.
* Average number of white-black transitions: The area and the amount of black pixels (both at initial and after applying RLSA) show a little positive link with this attribute. This indicates that there are typically more transitions between black and white pixels in blocks with a greater area and more black pixels.

It's crucial to recognise that the heatmap's associations are not very strong. This indicates that none of the qualities have a perfectly linear relationship with one another.

# Abalone\_Dataset Results

2021-SE-23

2021-SE-23

Figure 5 Figure 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-Score |
| For KNN Model | 0.725 | 0.730 | 0.718 | 0.722 |
| For Naïve Bayes Model | 0.763 | 0.763 | 0.759 | 0.759 |

### Results:

In the figure 5 the real sexes (M, F, and I) are represented by the rows in the particular confusion matrix, while the predicted sexes by the KNN classifier are represented by the columns. Each cell in the table indicates how many abalone are in the certain sex.

For instance, the cell labelled "M 126" in the upper left corner demonstrates that 126 abalone were in fact male (M), as predicted by the KNN classifier, and this information was accurate.

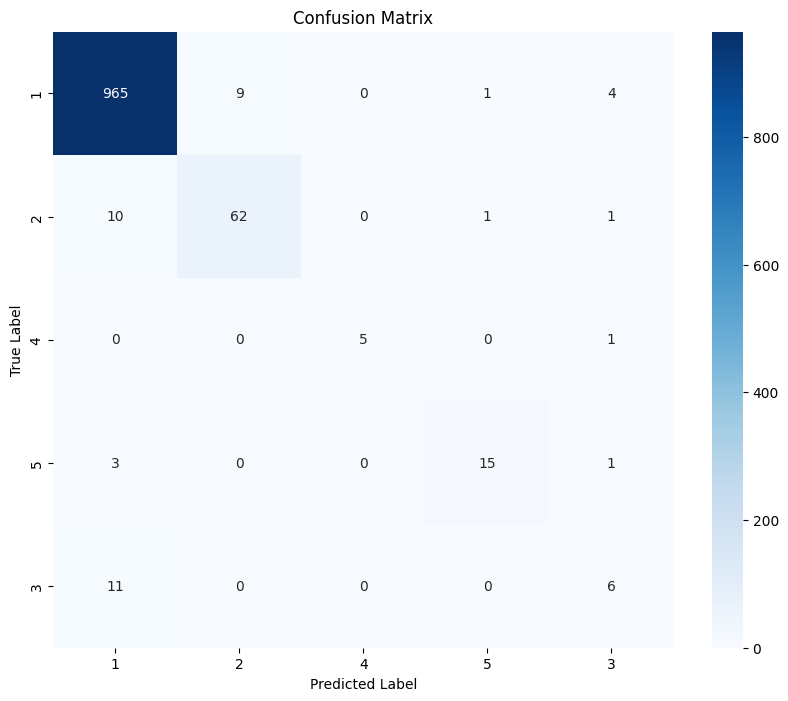
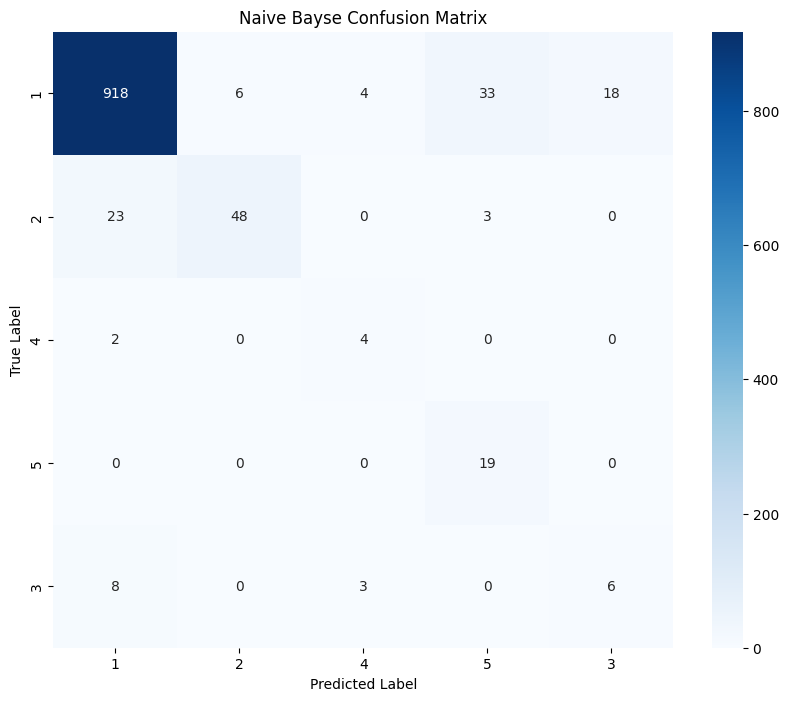
Given the low accuracy (about 55%) and low F1-score, it appears that the KNN classifier is having trouble identifying the sex of the abalone in this dataset.

In the figure 6 confusion matrix demonstrates that the Naive Bayes classifier is likewise having trouble correctly identifying the abalone in this dataset by sex (M, F, and I).

The high values in a confusion matrix should preferably be concentrated on the diagonal. This would suggest that the majority of the abalone sex is being accurately classified by the classifier.

But the low accuracy (Accuracy: 0.5239) indicates that the classifier frequently assigns incorrect gender classifications. It's reasonable to conclude from these measures that the Naive Bayes classifier is not the best option for identifying abalone sex in this dataset.

# Pages Dataset Results

2021-SE-23

2021-SE-23

Figure 7 Figure 8

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-Score |
| For KNN Model | 0.961 | 0.959 | 0.961 | 0.960 |
| For Naïve Bayes Model | 0.908 | 0.934 | 0.908 | 0.917 |

### Results

In the figure 7 the expected classes are represented by the columns in the confusion matrix, while the actual classes are represented by the rows.   
The table's cells each display the total amount of data points associated with a specific class. For instance, the cell labelled "0 965" in the upper left corner demonstrates that 965 data points were truly class 0, despite the KNN classifier's accurate prediction that they were class 0.   
The confusion matrix's diagonal displays the quantity of data points that were successfully classified. The KNN classifier properly identified 1078 out of 1088 data points in this instance because the total of the diagonal elements is 1078 (1078 / 1088 = 0.99). The Accuracy: 0.9616 value at the top of the matrix also indicates this.   
Additional measurements are

In the figure 8 the actual classes are represented by the rows in the particular confusion matrix, while the predicted classes are represented by the columns. Each cell in the table indicates how many text excerpts are in that specific class. As an illustration, the cell labelled "0 918" in the upper left corner reveals that 918 text samples were actually class 0, despite the Naive Bayes classifier's accurate prediction that they were class 0.

Although the recall (0.91) is somewhat greater than the precision (0.93) of the Naive Bayes classifier, the model may be better at avoiding misclassifying irrelevant snippets than locating all the relevant ones, despite the classifier's high accuracy.