PREDICTING THE AGE OF ABALONE FROM PHYSICAL MEASUREMENTS USING LOGISTIC REGRESSION.

A report by

CHERUBIN MAKEMBELE
STUDENT NO. **220418357**

As an Assessment for the Module
MACHINE LEARNING AND DATA SCIENCE
Within the Qualification
ADVANCED DIPLOMA IN MATHEMATICAL SCIENCES

LECTURER: Mrs. Tayla Wannenberg

Cape Peninsula University of Technology
March 2023

DECLARATION

I, CHERUBIN MAKEMBELE, declare that the contents of this research report represent my work. I know that plagiarism is wrong. Plagiarism is using another's work and pretending that it is one's own. Each contribution to, and quotation in, this report from the work(s) of other people has been attributed and has been cited and referenced.

I have not allowed, and will not allow anyone, to copy my work to pass it off as his or her work.

Date of Declaration: 12 March 2023

1. Introduction:

Abalone, also known as sea ear shell, is an enormous gastropod mollusc or snail frequently found in chilly coastal waters in North America, New Zealand, Australia, South Africa, and Japan. The aquatic creature belongs to the Haliotid family and has a single spiral-shaped shell with numerous small holes around the borders on top, along with a huge foot for clinging to rocks and eating algae. The meat of abalone is incredibly rich and prized, and it is regarded as a culinary delicacy. abalone is generally sold alive in the shell, frozen, or canned. It is usually cut into thick steaks and pan-fried though it can be eaten raw or incorporated into other meals. The pearl that protects the abalone's shells is another factor in its popularity. (Watson, 2022)

The popularity of abalone resulted in overfishing, which put the species in danger of going extinct. Age and the economic worth of abalone are positively connected. The number of rings on an abalone is crucial in determining its age. Abalone is classified according to gender male; female and infant in reference to age even though its age is binary. Researchers use a microscope and manual counting to determine the number of rings. After acquiring the required number of rings, the age is determined by adding 1.5. For example, an abalone with seven rings is 8.5 years old. (Watson, 2022)

2. Aim:

This work aims to create a classification model that can distinguish between young and older abalone with accuracy based on their age and other physical measurements. The model will specifically try to categorize abalone as either young (age <= 11) or old (age >11), which can subsequently help determine the abalone's economic value of an abalone.

3. Methods:

3.1. Data Source & Description

The data Abalone_1(first dataset) and Abalone_2 (second dataset) originate from the UC Irvine Machine Learning Repository published in 1995. The UCI Machine Learning Repository is a collection of databases, domain theories and data generators destined for machine learning empirical analysis (Warwick, J.N. et al. 1994). The datasets have variables which are:

- **Gender (first dataset) and Sex (second dataset):**

  Male, Female, or Infant

- **Length:**

Longest shell measurements

- **Diameter:**

The distance across the widest part of the shell in a straight line.

- **Height:**

The shell's height or depth is measured vertically from the apex (top) of the shell to the base (bottom).

- **Whole Weight:**

The weight of the entire animal, including the shell, flesh, and any other parts present

- **Shucked Weight:**

the shell is typically removed, and the meat is extracted. The shucked weight is then determined by weighing the meat alone.

- **Viscera Weight:**

The weight of the abalone's internal organs is measured after shucking.

- **Shell Weight:**

Refers to the empty shell's weight.

- **Rings**

Refers to the ring-like patterns found on the abalone's shell.

3.2. Data Preparation

The software used throughout this investigation to implement the supervised classification machine learning model is RStudio, which is a free, open-source, and multiplatform integrated development environment (IDE). RStudio uses R programming language for data processing and statistical analysis (Barnier, 2011).

The datasets were imported into RStudio's IDE for manipulation. They had to be merged since the study was to be conducted using the datasets combined. However, some discrepancies were to be dealt with before conducting a sound analysis.

- For starters, as mentioned in (3.1) Gender and Sex variables represent the same variable, ergo they must be standardized to the same name. the name Gender has been chosen for the variable.
- Secondly, in Abalone_1 there was an observation that pointed to a diameter of 24 whereas all other observations were between 0 and 1 so we had to change the 24 to Non-Applicable (NA) to be imputed later.
- Finally, Imputing needed to be done before performing any analysis to ensure a good model.

  Variables imputation was done as follows:

  - numerical missing observations were replaced with the mean of the column's observation without the observations that were missing (Roy, 2023).
  - Categorical missing observations were replaced by the mode of the column's (variable) observations that were missing (Roy, 2023).

3.3. Data Descriptive Statistic

The two sets of data were subjected to the descriptive statistics listed below to get insight into the data, including the mean, median, distribution, etc. according to the variables' appearance order. According to the variable classification that was created as the response variable, the hypotheses were as follows:

- **The null hypothesis (H_0)** states that the abalones were classified as young, implying that there is no significant difference or evidence to suggest that they are old.
- **The alternative hypothesis (H_1)** states that the abalones were old, suggesting that there is significant evidence to support the claim that they are not young (Simplilearn, 2023).

3.3.1. Gender

Following the dataset's combination and cleaning, the variable Gender displayed the characteristics depicted in Figure 1: Females made up 31.25% of the population, Infants made up 32.13%, and Males made up 36.62%, with frequencies of 1307, 1344, and 1532, respectively. Males were the majority, or most represented, gender, with a share of 36.62%.

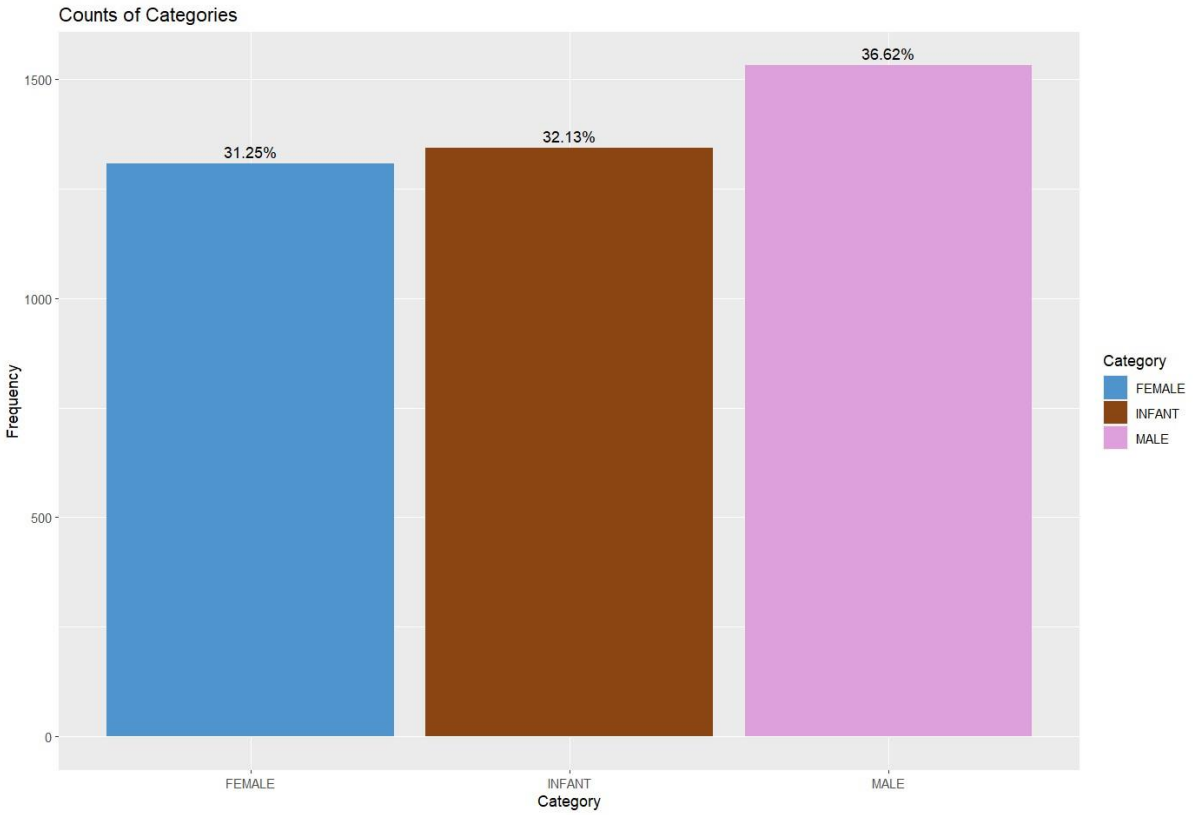

Figure 1 Gender

3.3.2. Length

The length followed a left-skewed distribution, as shown in Figure 2, where the mean=0.5240994 and median= 0.545. the minimum value = 0.075 and the maximum = 0.815 and there were no extreme outliers (Figure 3).

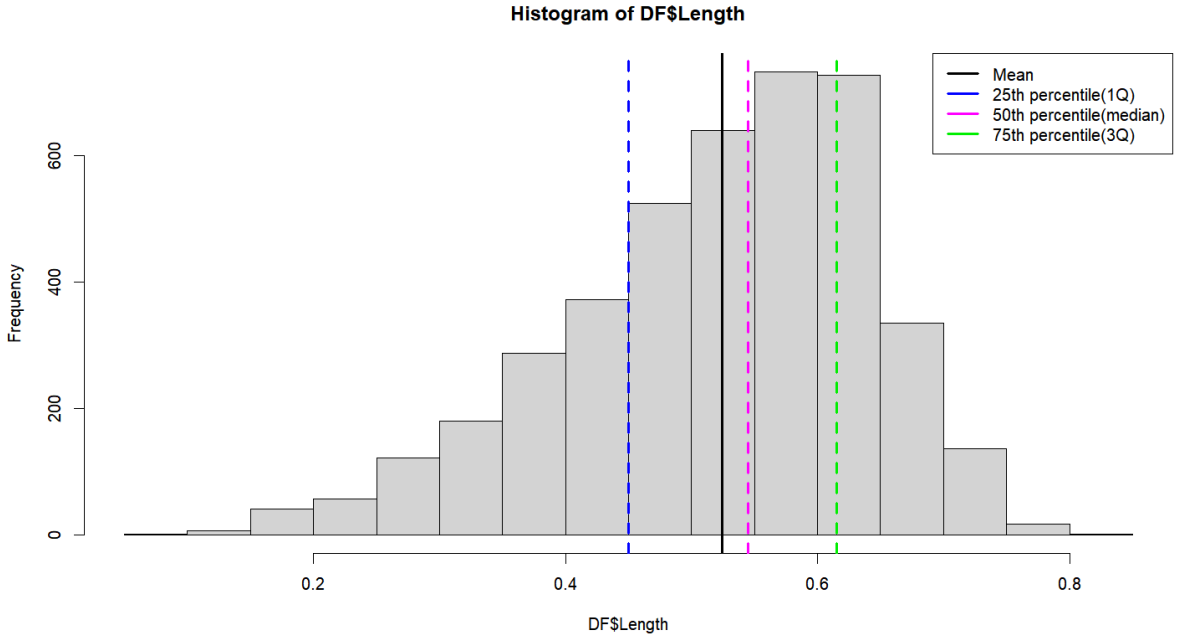

Figure 2 Length

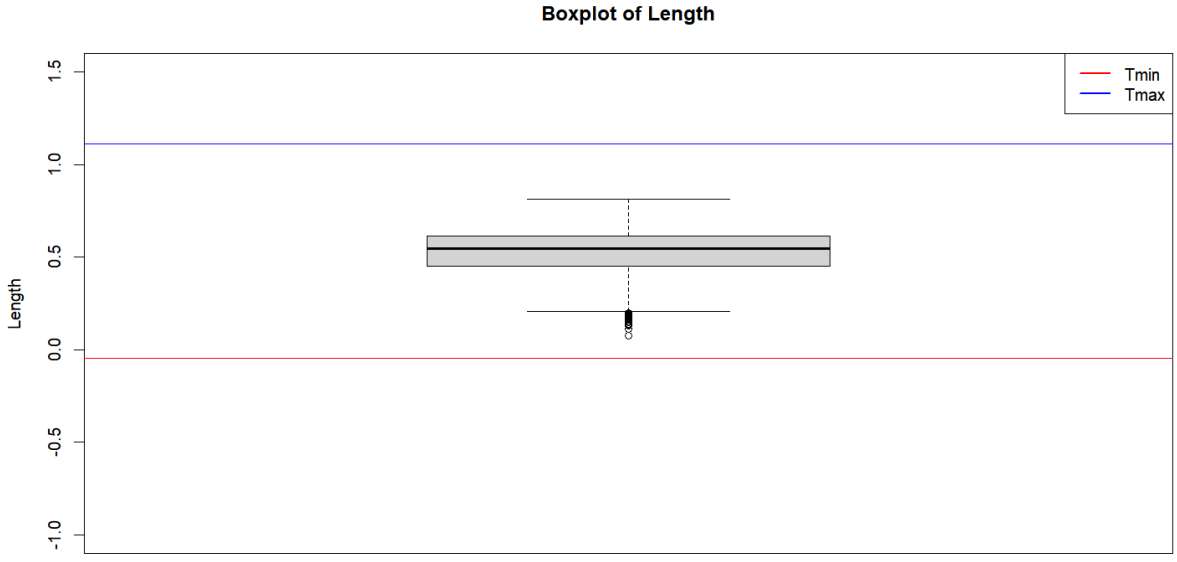

Figure 3 Boxplot Length

3.3.3. Diameter

The diameter followed a slightly left-skewed distribution with a median of 0.425 and a mean of 0.4080059 as per Figure 4. Additionally, the minimum value was 0.055 and the maximum value 0.65 and there were no extreme outliers (Figure 5).

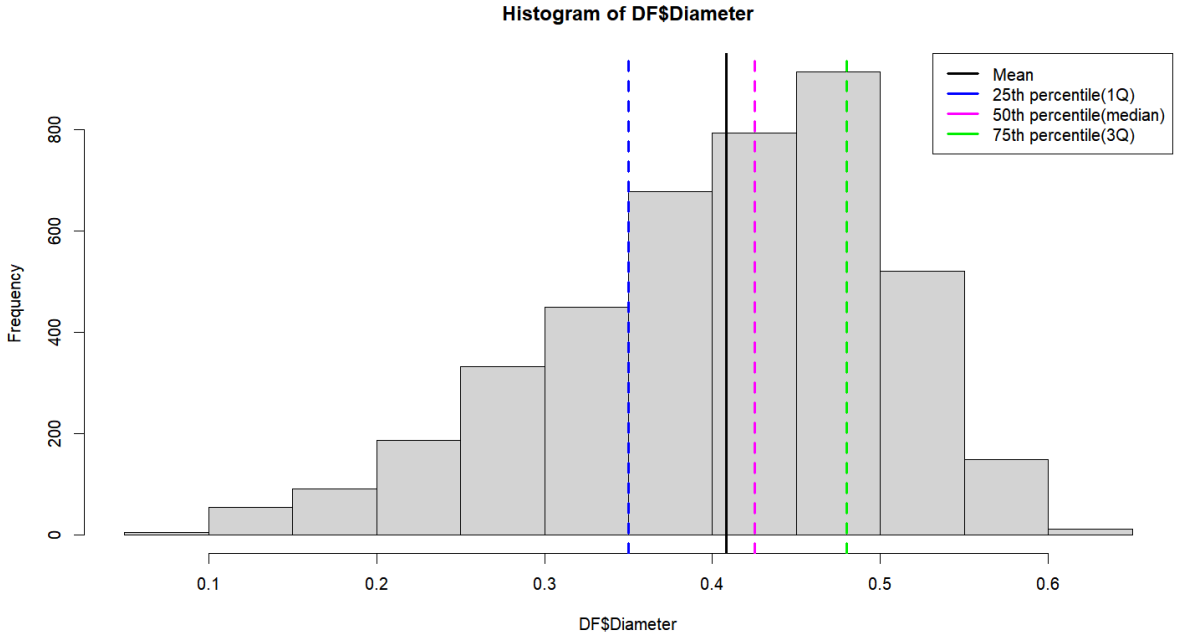

Figure 4 Diameter

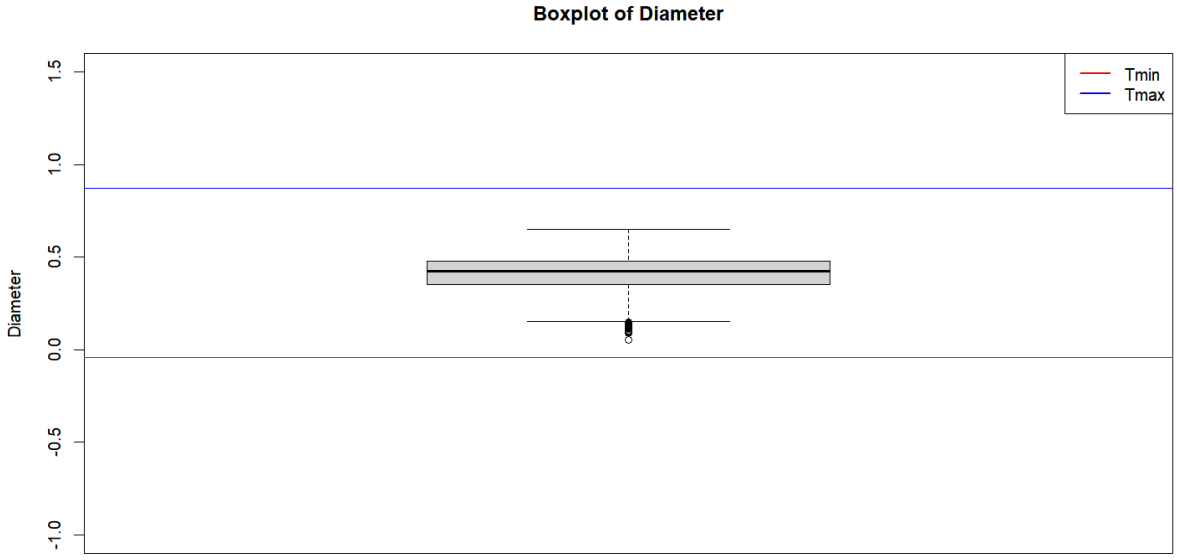

Figure 5 Boxplot Diameter

3.3.4. Height

The diameter followed a normal with a median and a mean of approximately 0.14 as per Figure 5. The minimum value was 0 and the maximum value was 2 and there were extreme outliers that we proceeded to remove in pursuit of a sound model (Figure 7 and Figure 8).

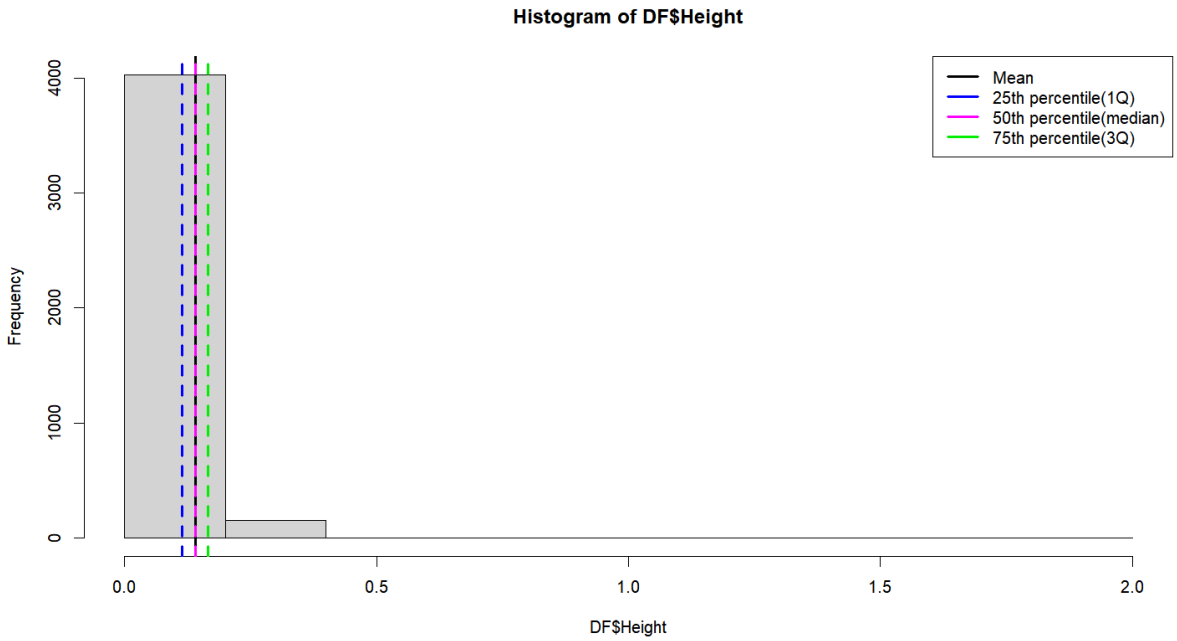

Figure 6 Height

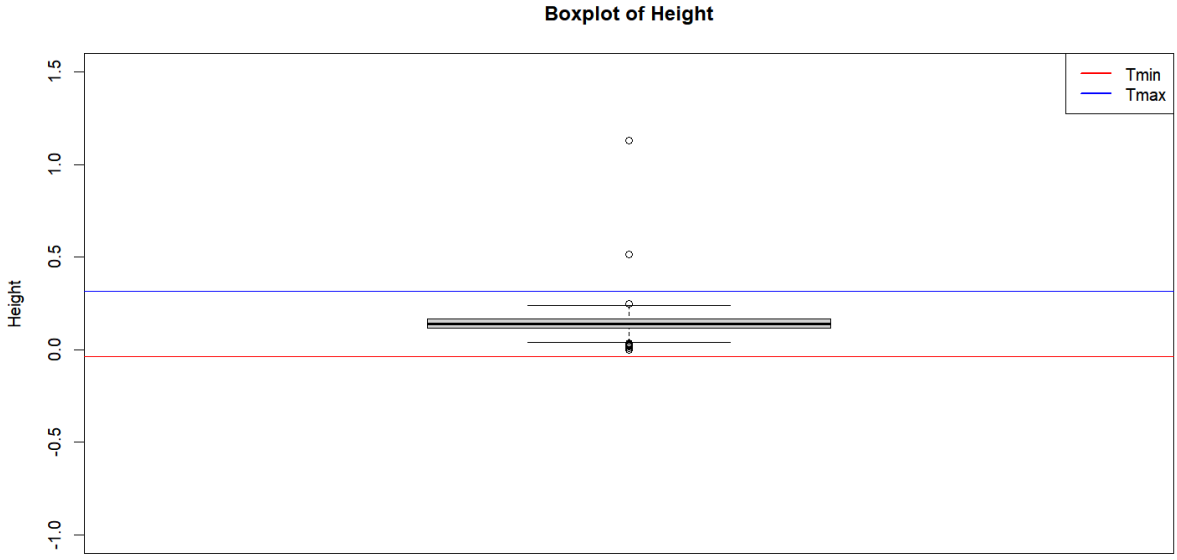

Figure 7 Height Boxplot 1

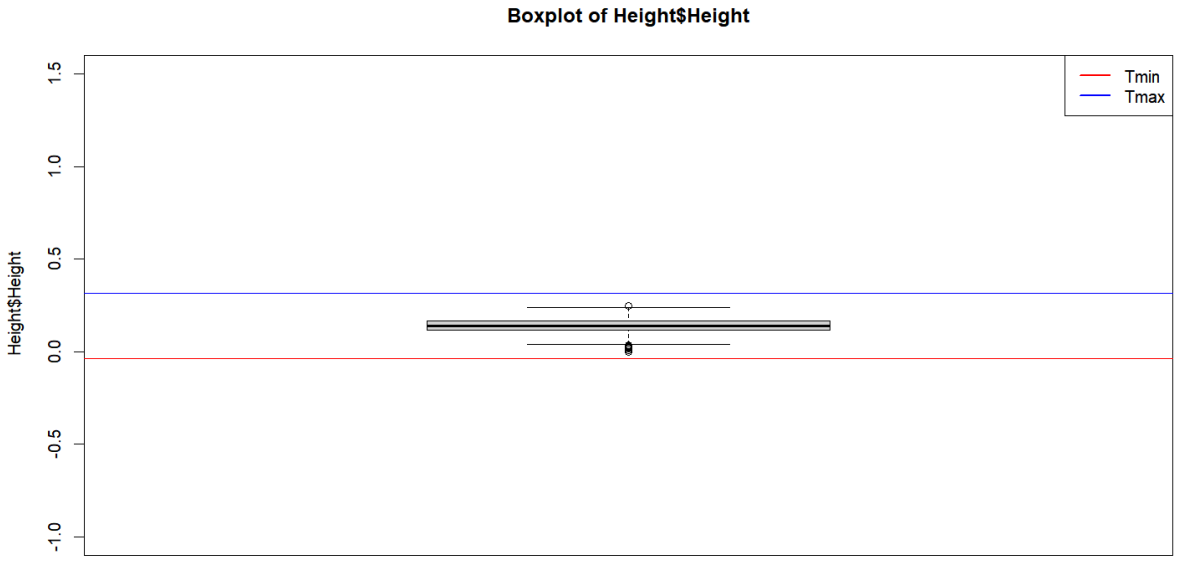

Figure 8 Height Boxplot 2

3.3.5. Whole weight

The median of the whole weight was 0.8 and the mean was 0.8292303 and the graph shows right skewness. The minimum value is 0.002 and the maximum is 2.8255 and there were no extreme outliers (Figure 9 & Figure 10).

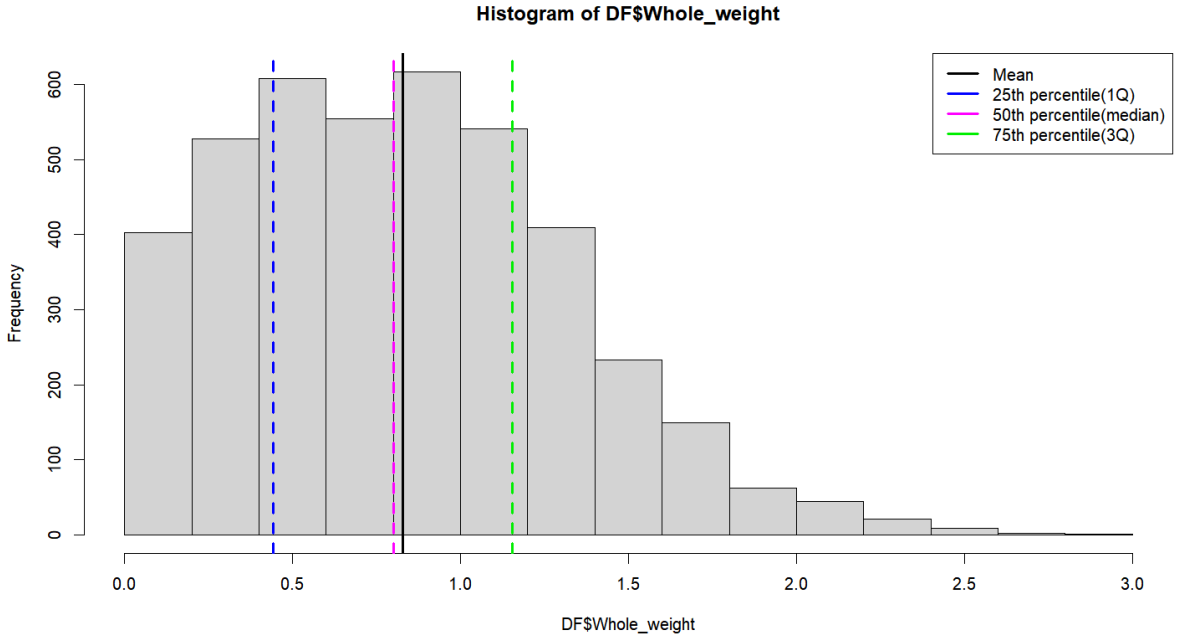

Figure 9 Whole Weight

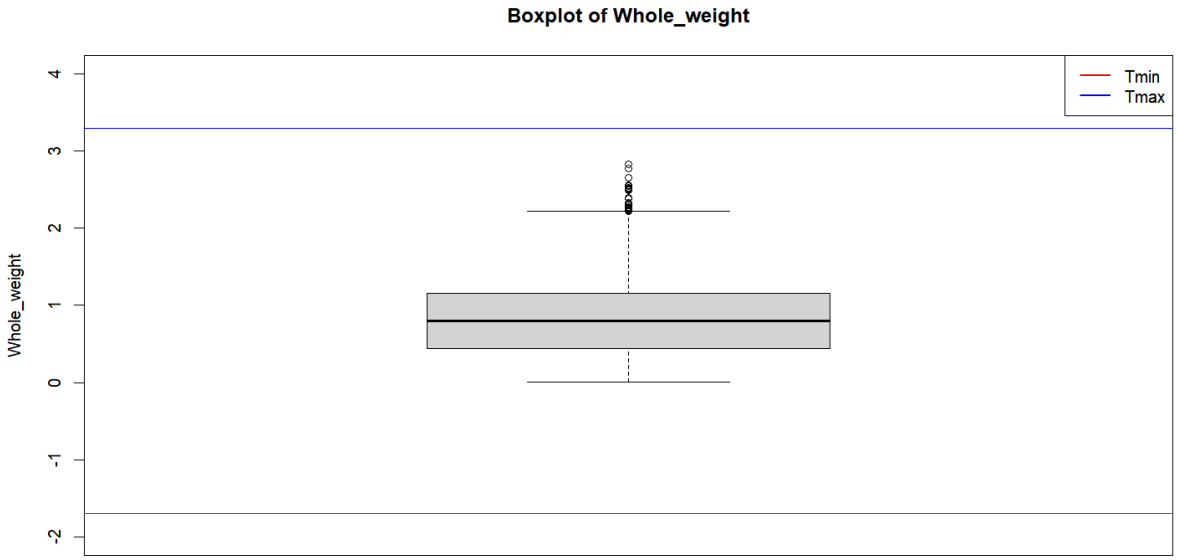

Figure 10 Boxplot Whole Weight

3.3.6. Shucked weight

The graph in Figure 11 shows a slight right skew with a median of 0.336 and a mean of 0.3597012. The minimum value is 0.001 and the maximum is 1.488 and there were no more extreme outliers (Figure 12 & Figure 13)

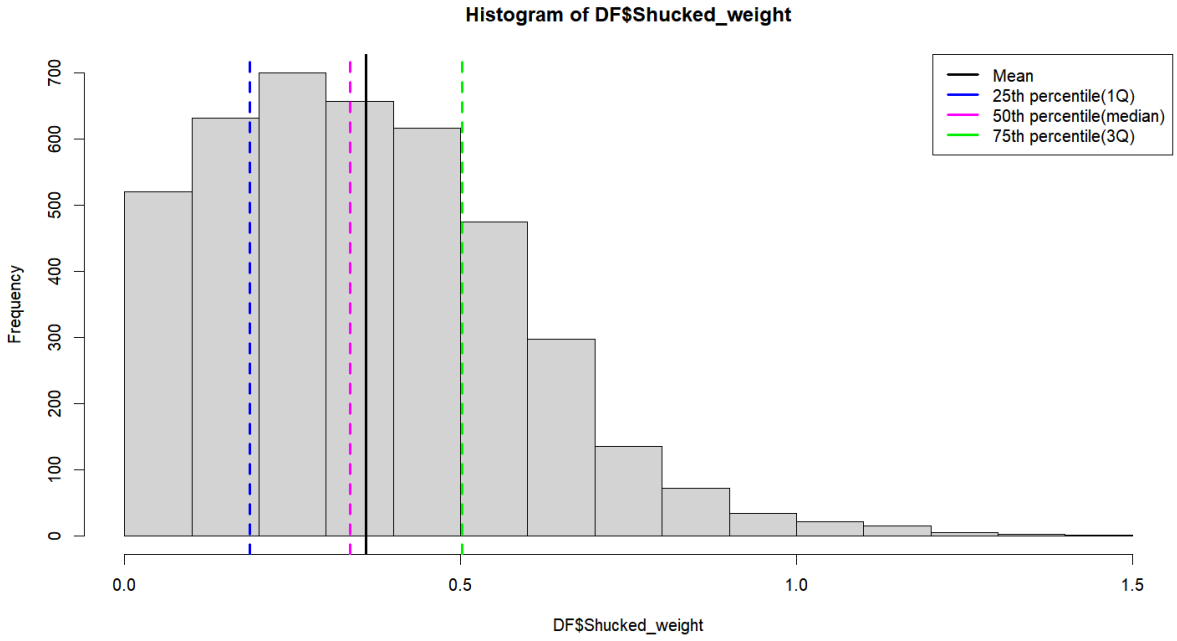

Figure 11 Shucked Weight

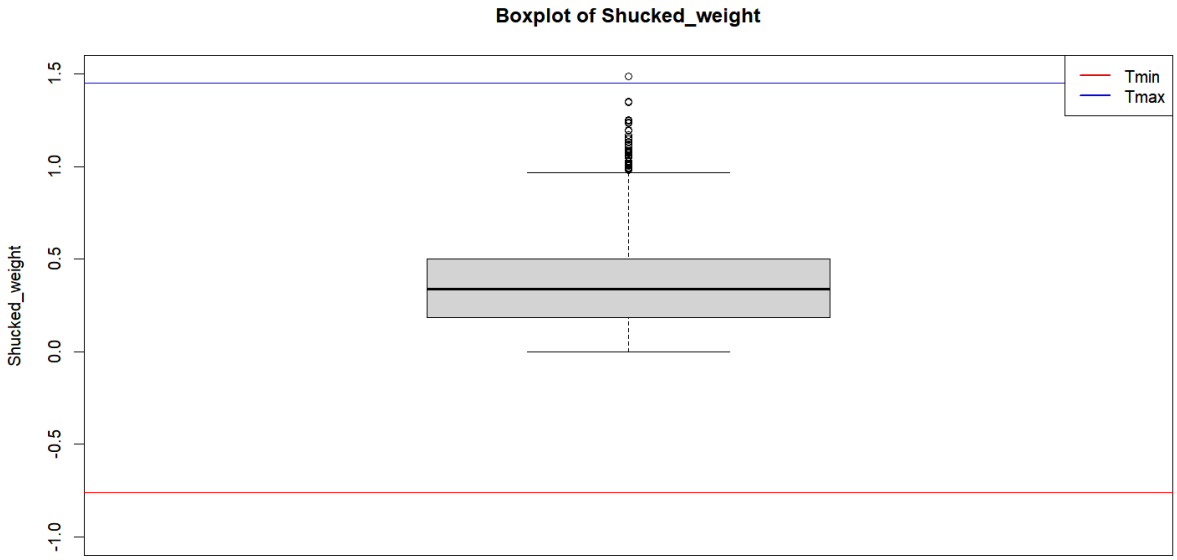

Figure 12 Boxplot Shucked Weight 1

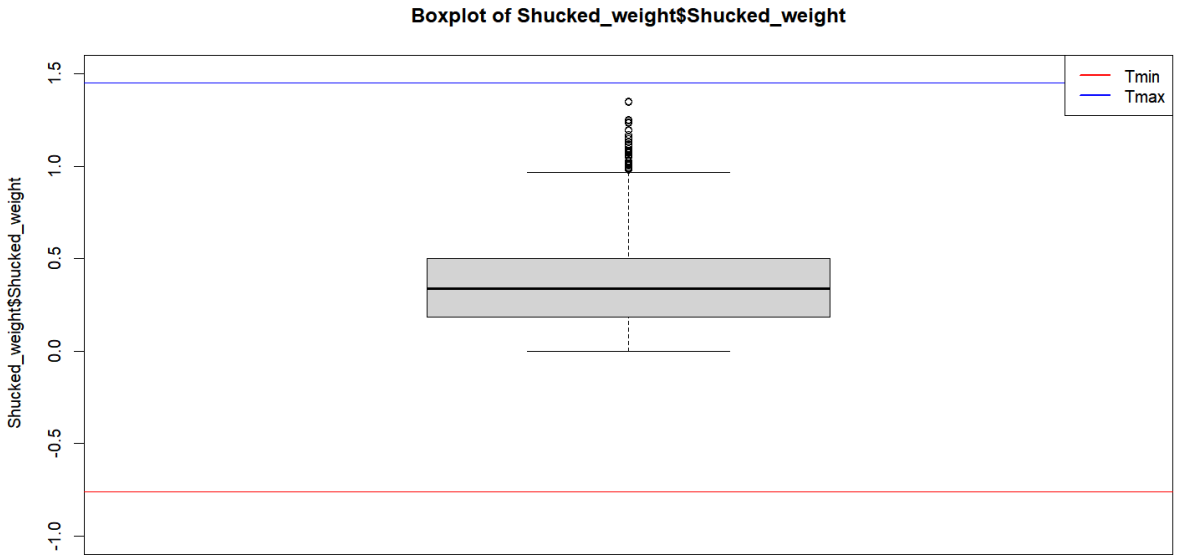

Figure 13 Boxplot Shucked Weight 2

3.3.7. Viscera weight:

Shucked weight displayed a light right skew with a median of 0.17 and a mean of 0.1807369 with a minimum value is 5e-04 and a maximum is 0.76 and there were extreme outliers that were removed to better the model (Figure 14, Figure 15 and Figure 16).

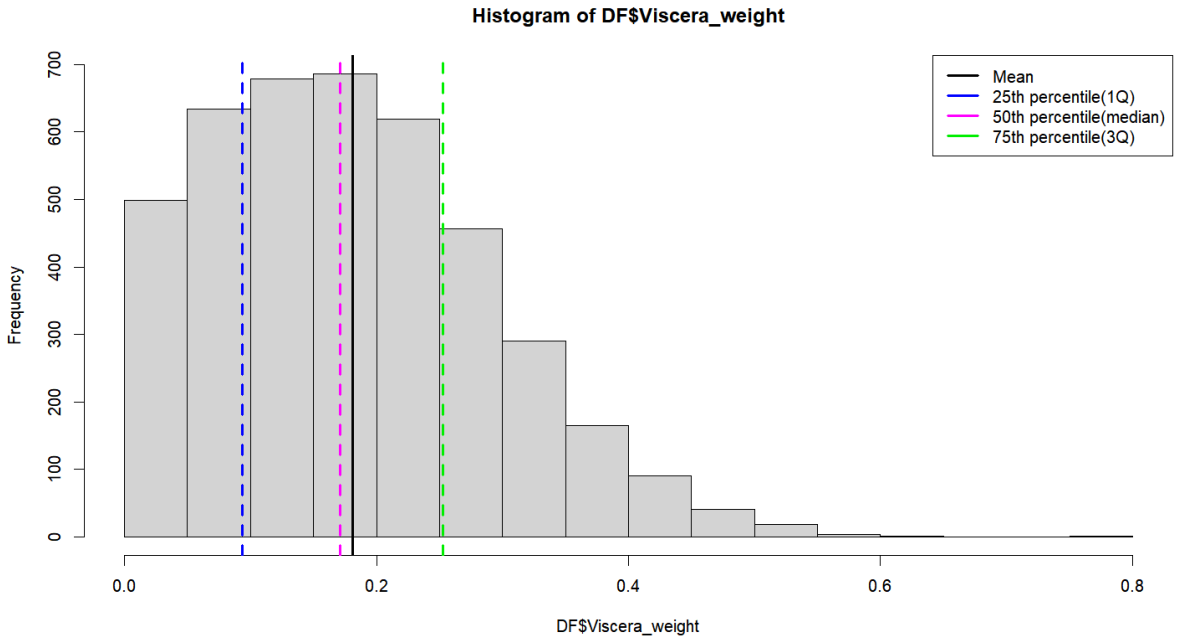

Figure 14 Viscera Weight

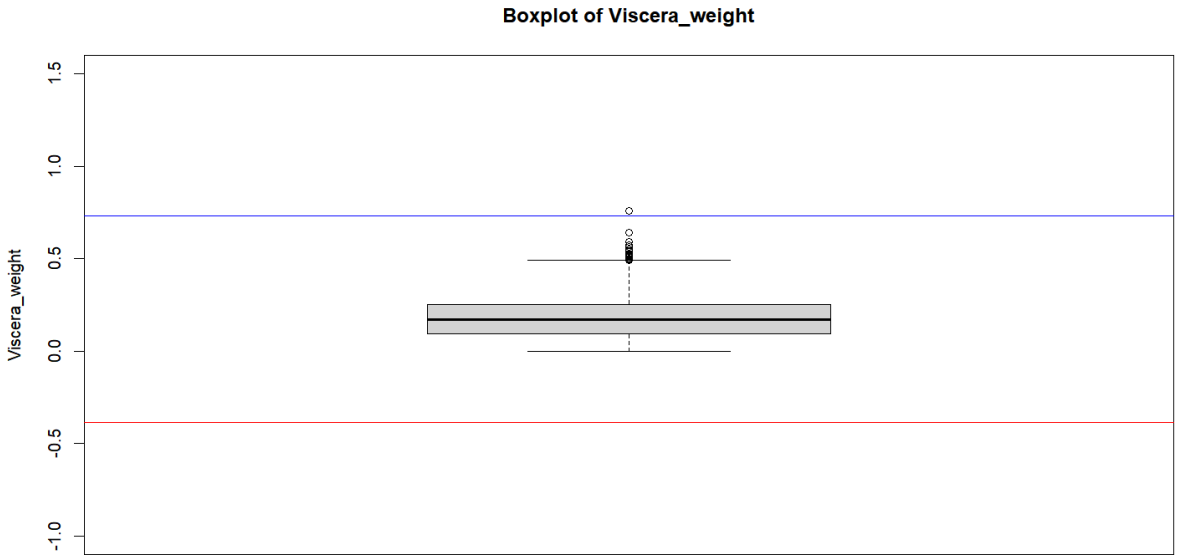

Figure 15 Viscera Weight Boxplot 1

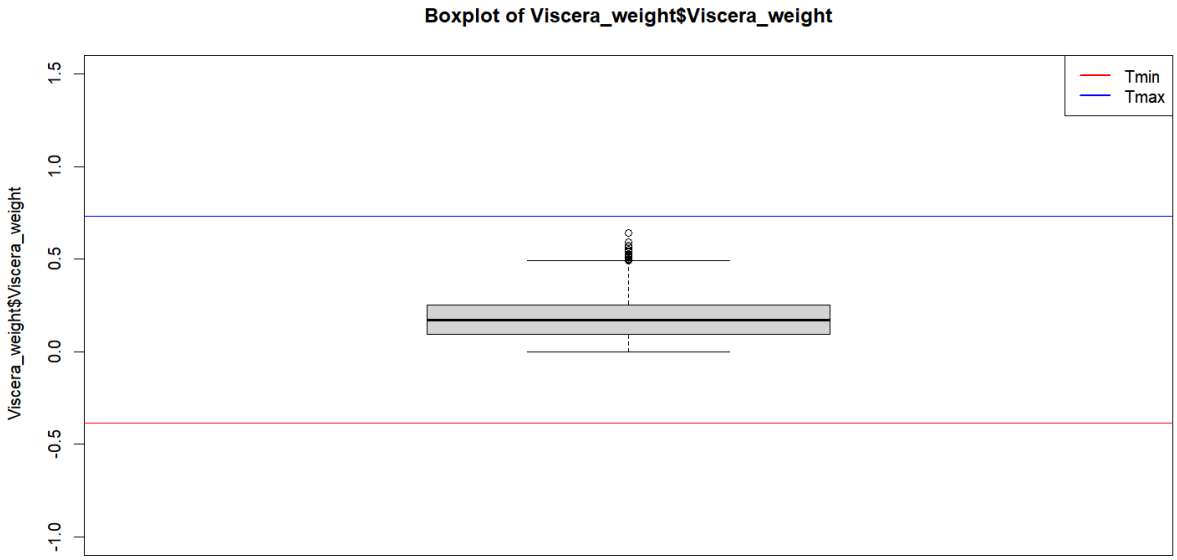

Figure 16 Viscera Weight Boxplot 2

3.3.8. Shell weight:

Shell weight displayed a nearly normal distribution with a median of 0.234, a mean of 0..2388394, a minimum value of 0.0015 and a maximum of 1.005 and there were extreme outliers that were removed (Figure 17, Figure 18 and Figure 19).

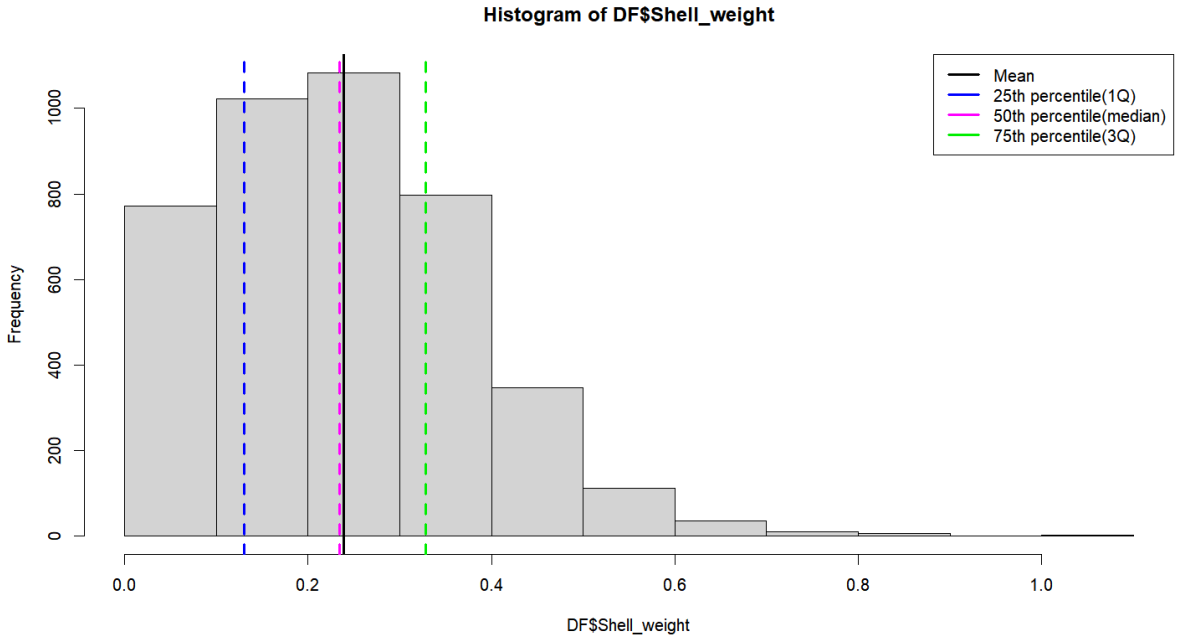

Figure 17 Shell Weight

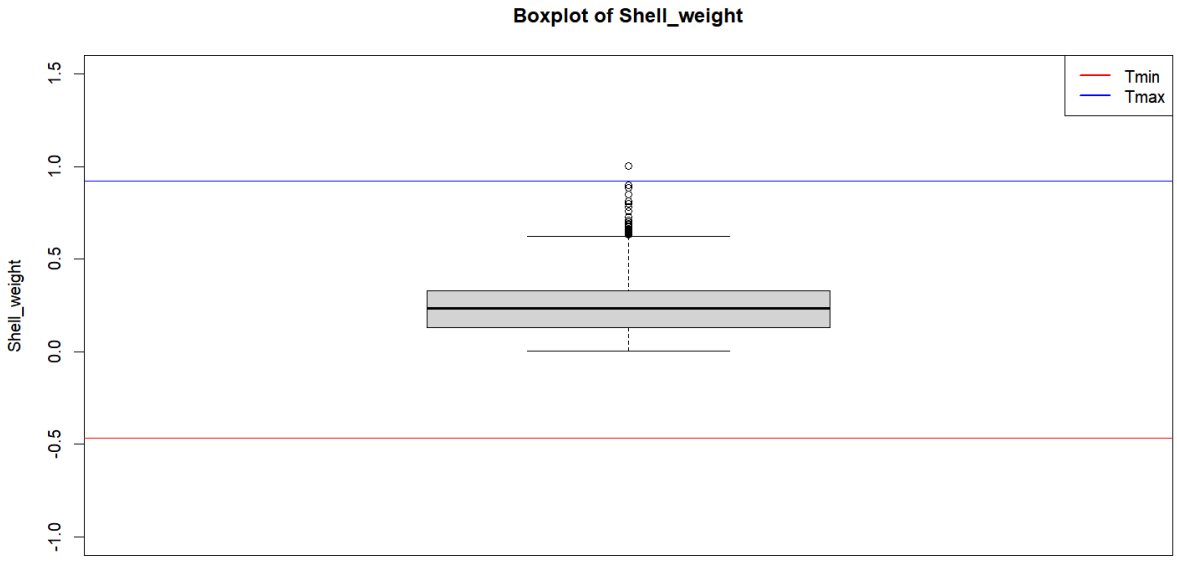

Figure 18 Shell Weight Boxplot 1

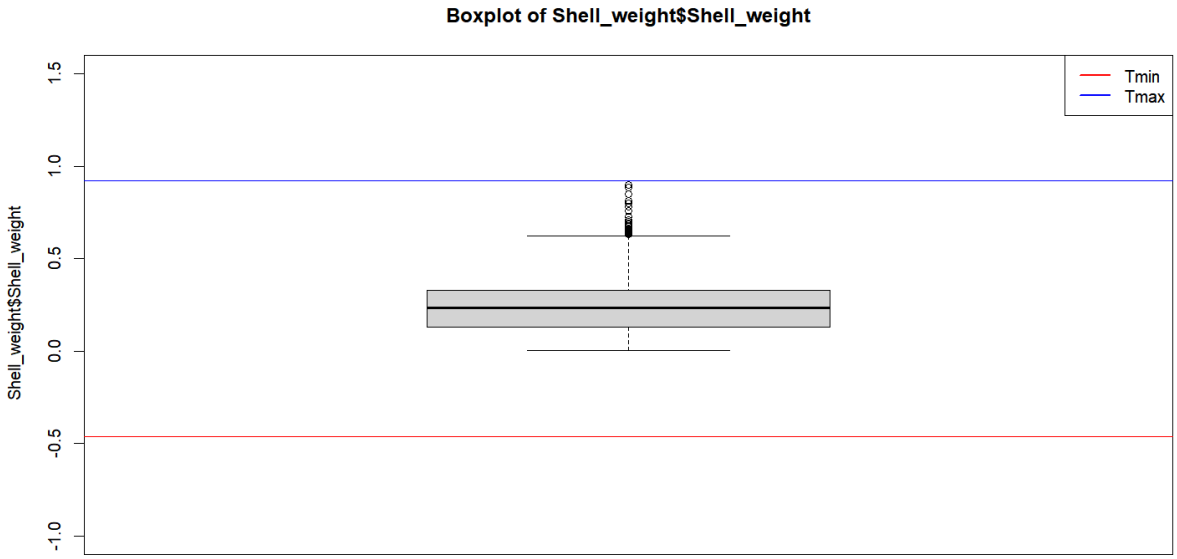

Figure 19 Shell Weight Boxplot 2

3.3.9. Rings

Rings or Age had a median 9 and a mean of 9.97204 and minimum value is 1 and maximum is 99 and there were extreme outliers.

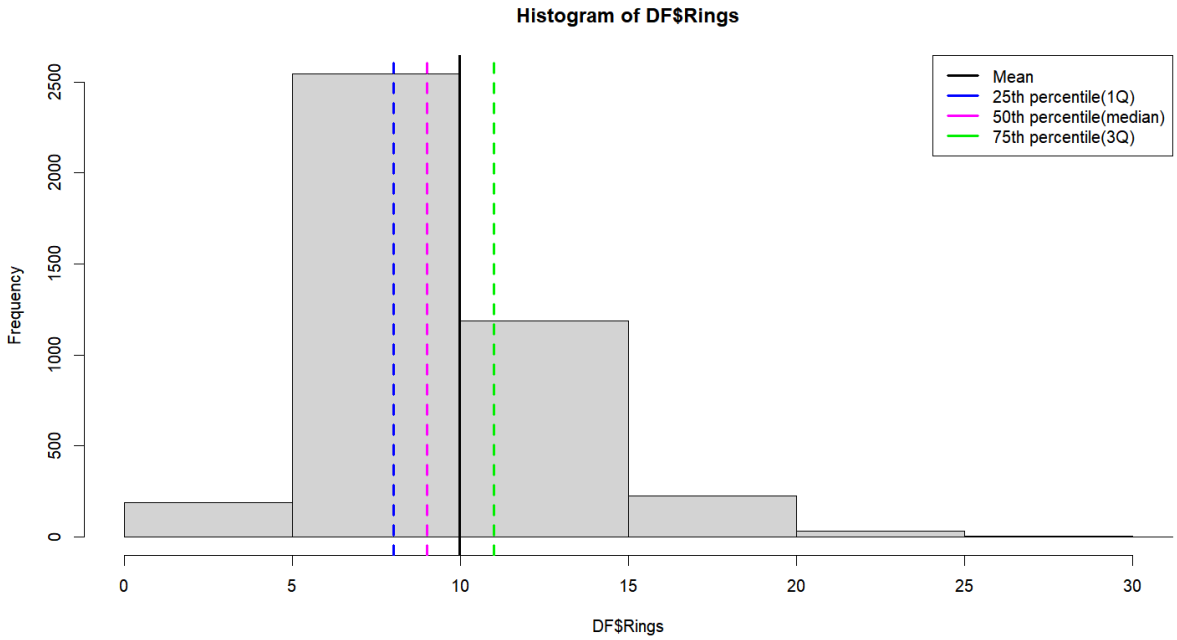

Figure 20 Rings

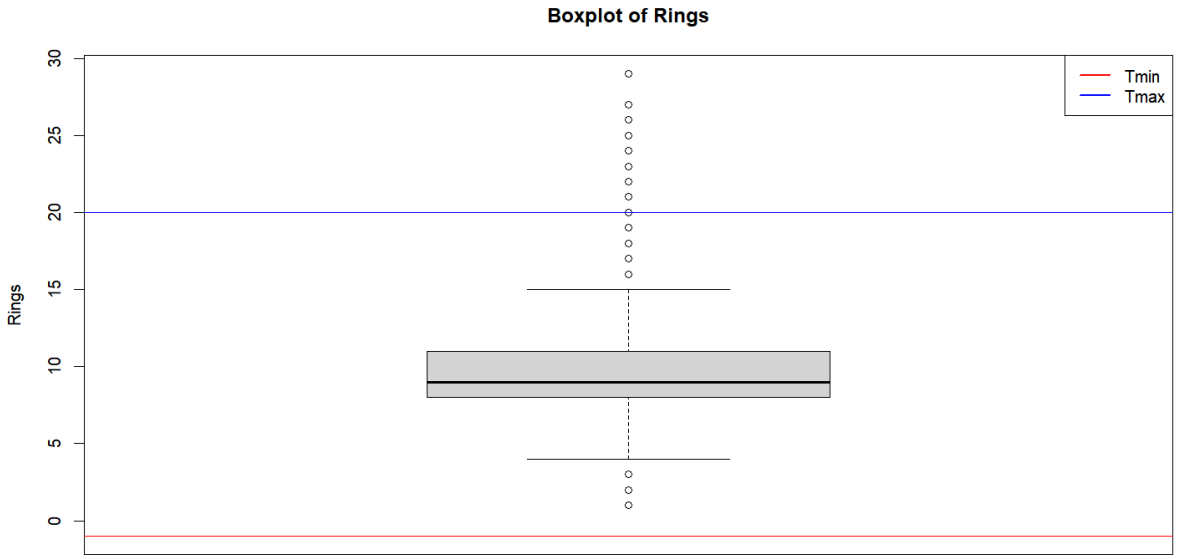

Figure 21 Rings Boxplot 1

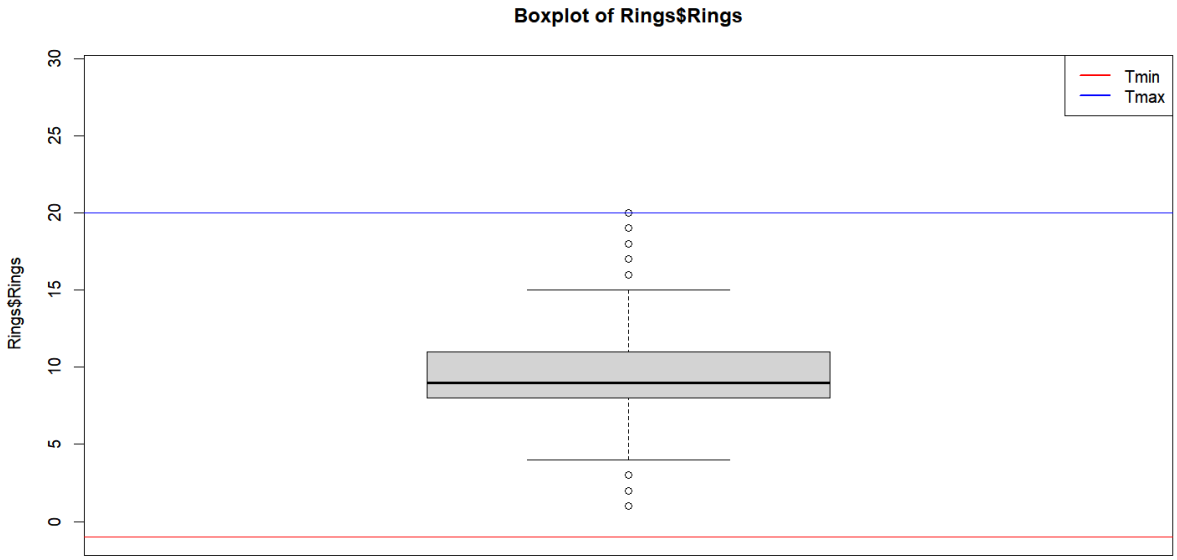

Figure 22 Rings Boxplot 2

3.4. Data Analysis

The machine learning model required for the problem is a classification model. For problems involving linear and binary classification, logistic regression is an easy-to-use technique that produces great results in classes that can be separated linearly. The best way to think of logistic regression, according to Edgar et al. (2020), the best approach to think of logistic regression is as a linear regression applied to classification problems. The logistic regression uses:

- the logistic function $\frac{dP_{(t)}}{dt} = rP_{(t)}(1 - \frac{P(t)}{K})$ which is a forecasting tool with the solution $P(t) = \frac{K}{1+\left(\frac{K}{P(0)}-1\right)e^{rt}}$ where:
    - $\frac{dP_{(t)}}{dt}$ = Population variation over time
    - $P_{(t)}$ = Population at time "t"
    - t = Time
    - r = Growth rate
    - K = Carrying capacity
    - e = The natural logarithm base or Euler's number
- The logistic function is $\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon_i$
- The inverse function is $p = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$

- Where p is the dependent or response variable with a binary value.
- β_0 is the intercept or the value of the response variable if independent values x_1, x_2 & x_n are of value 0
- x_1, x_2 & $x_n$are the independent or explanatory variables.
- β_1, β_2 & β_n are the linear components and the coefficients of explanatory variables.
- ε_i is the random error term (Edgar et al., 2020)

4. Results

Before implementing the logistic model, it was necessary to clean the data, check for extreme outliers and z-score outliers, and verify the logistic regression assumptions, such as multicollinearity, linearity of the independent variables and log odds, etc. For the assignment, the creation of a new variable that serves as the model's label was of need. That variable is Classification. The merged dataset presented the following variables:

4.1. Gender

Which is the abalone's gender which includes younger abalones referred to as "infants". The variable is categorical, and the dataset noted the ratios depicted in Figure 23 below, with male abalones having a mode of 36.62%.

A chi-squared test was conducted to determine whether the variable Gender was independent of the Classification response variable, and the resulting p-value of 2.2e-16<5% indicates that we should keep this variable for the final model (*8.1 - the chi-square test of independence: Stat 500*).

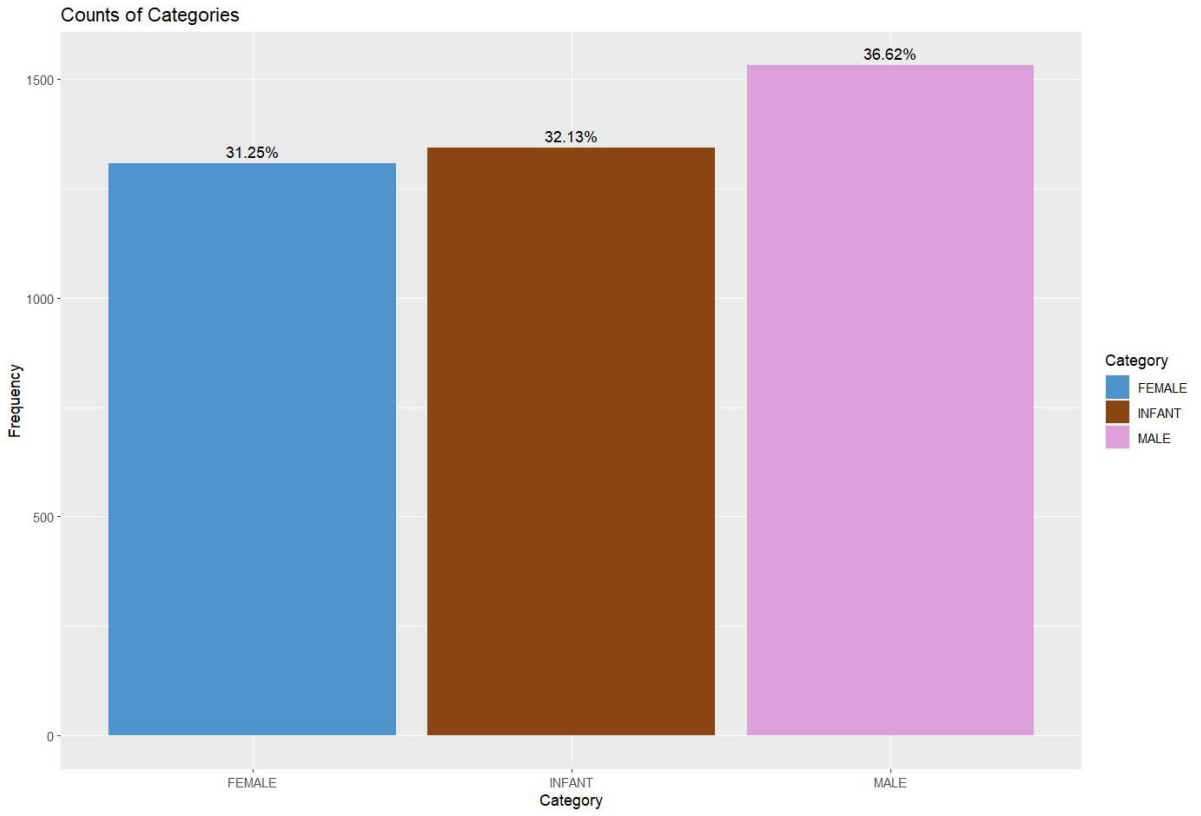

Figure 23

4.2. Length

Which is the longest shell measurement, thus a continuous variable. According to the box Tidwell test, length is linear to the log odds of classification and on that level, the variable is retained. The test's p-value was 8.215e-07, which is 0.05. The correlation matrix in Figure 24 demonstrates the high correlation (multicollinearity) of Length and various other variables on the second level therefore we do not include Length for the final model (Frost, 2023) .

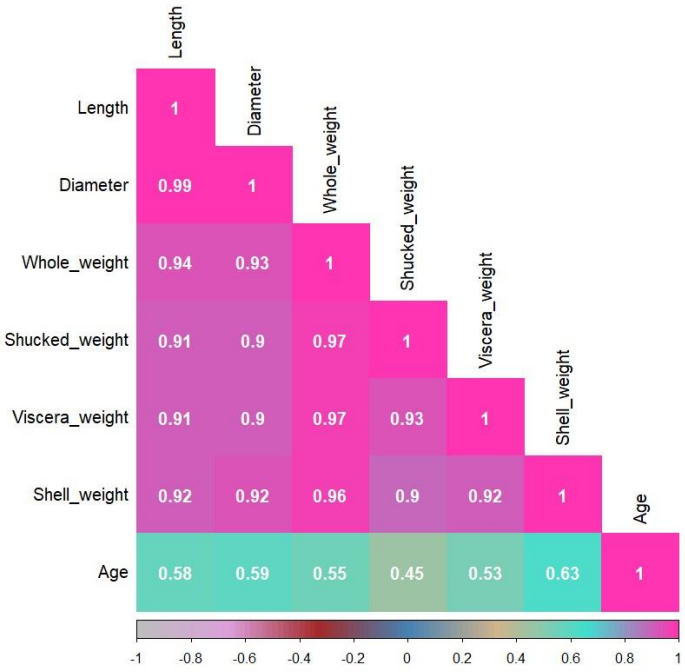

Figure 24

4.3. Diameter

Which is a measurement perpendicular to the length, thus a continuous variable. The Diameter is linear to the log odds of classification, and on that level, the variable is preserved, according to the box Tidwell test. The p-value for the test was 8.215e-07 < 0.05. We do not include the Diameter in the final model since the correlation matrix in Figure 24 shows that the Diameter and numerous other variables are multicollinear on the second level (Frost, 2023).

4.4. Height, Whole_weight, Shucked_weight, Viscera_weight, Shell_wheight, Rings, Age

Viscera weight is the weight of the intestines (after bleeding), Whole weight is the weight of the entire abalone, Shucked weight is the weight of the meat, Shell weight is the weight after being dried, Height is the measurement of the entire abalone and Rings: Age is expressed as +1.5 years. According to Figure 3, all those continuous variables will explain the same variation or phenomenon in a generalized linear model. Therefore, keeping either of them alone will work for the model (PPA 696 RESEARCH METHODS). Age is chosen as the investigation's focus.

Once the predictors had been determined, the final model contained only three variables: Gender, Age, and Shell_weight. None of these indicators were found to be statistically significant in determining the classification of abalone because their p-values were all higher than 0.05. Additionally, it is clear from the graph that the data is not entirely random.

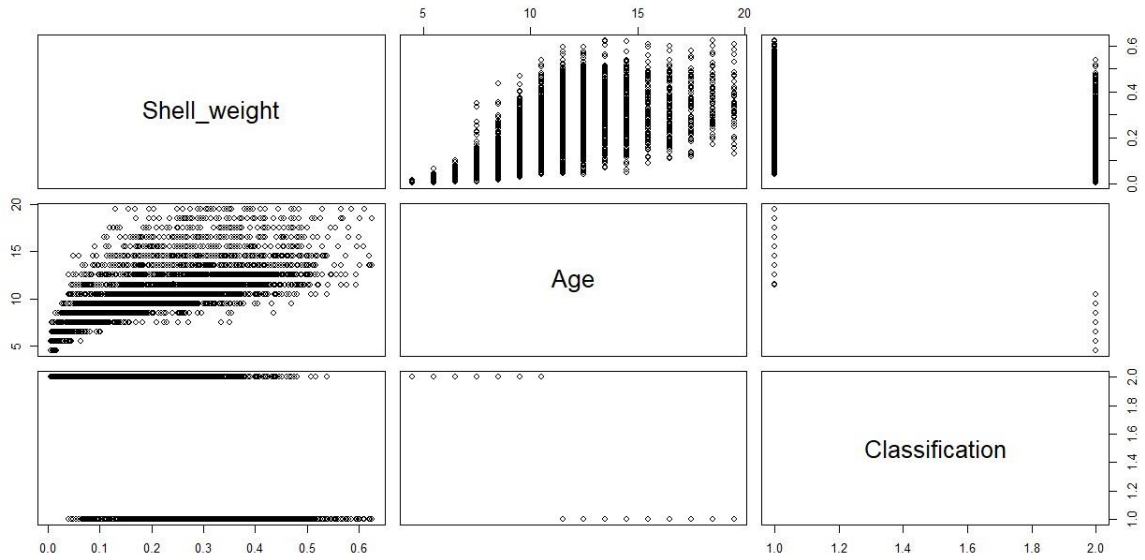

Figure 25 Data Scatter Plot

5. Conclusion

The linear regression analysis was performed to examine the relationship between Classification and predictors narrowed down to Gender, Age, and Shell_weight. Nevertheless, the results showed that none of the predictors were statistically significant at the 0.05 level in predicting the classification.

Multicollinearity was present in every numerical predictor that was approved for the model. Additionally, Gender was dependent on Classification, and the fact that it was a 3-level component despite being binary may be one of the reasons why it is not significant in creating an appropriate classification machine learning model, along with the numerical predictors. To build a machine learning model that determines if an abalone is young or elderly, additional analyses may investigate new predictors, more observations, a different dataset, or even a different classification method.

6. References

Barnier, J. and Me:, J.B.P.F. (2011) Rstudio, UN Environnement de Développement Pour R, QUANTI / Sciences sociales. https://quanti.hypotheses.org/488 [11 May 2023].

Edgar, T.W., Manz, D.O. and Subasi, A. 2020. Logistic regression, Logistic Regression - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/computerscience/logistic-regression [21 June 2023].

Frost, J. (2023) Multicollinearity in regression analysis: Problems, detection, and solutions, Statistics By Jim. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/ [21 June 2023].

PPA 696 RESEARCH METHODS (no date) Multiple regression. Available at: https://home.csulb.edu/~msaintg/ppa696/696regmx.htm#:~:text=When%20two%20variables%20are%20highly,by%20the%20second%20independent%20variable. [21 June 2023].

Roy, B. 2023. All about missing data handling, Medium. https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184 [21 June 2023].

Simplilearn. 2023. What is P-value in statistical hypothesis?: Simplilearn, Simplilearn.com. https://www.simplilearn.com/tutorials/statistics-tutorial/p-value-in-statistics-hypothesis#:~:text=A%20P%2Dvalue%20less%20than,it%20is%20not%20statistically%20significant

The Pennsylvania State University. 2023. 8.1 - the chi-square test of independence: Stat 500 PennState: Statistics Online Courses. https://online.stat.psu.edu/stat500/lesson/8/8.1 [21 June 2023].

Warwick, J.N., Tracy L.S., Simon R.T., Andrew J.C. and Wes B.F.. 1994. “UCI Repository of machine learning databases.” http://www.ics.uci.edu/~mlearn/MLRepository.html.

Watson, M. 2022. Rich, tender abalone is a delicious treat from the sea. The Spruce Eats. https://www.thespruceeats.com/about-abalone-2215715 [11 May 2023].