**A Multiple Regression Approach to Predicting Abalone Age Based on Physical Characteristics Alone**

**Kyle Culp**

School of Mathematical and Statistical Sciences (SOMSS)

Arizona State University, Tempe, Arizona

**Introduction**

Abalones are an endangered shellfish that are highly nutritious and eaten often in many places such as France in addition to the places where they are usually sourced from such as Japan and New Zealand. Their shells are not wasted or consumed either, they are instead used in art and jewelry, which clearly shows their value to the people and culture in these areas.

Traditionally, in order to determine the age of an abalone, it is necessary to perform the time-consuming process of cutting the shell open, staining it with a chemical to dye the rings and then counting the rings under a microscope. Each ring accounts for one year of life. This has two key downsides: the death of the abalone and the person’s time spent on this process. If we could instead take a few measurements of the abalone to accurately determine the age of the creature, then we would significantly decrease the time it takes to determine abalone age and only harvest certain abalone based on age.

This would be significant as currently we have a wasteful process that limits the amount of data yielded from a limited resource. With this increased set of data and simpler method we could easily determine:

* Which abalone to harvest
* Factors that affect growth rate of individual abalones
* How growth rate affects specific physical characteristics
* Whether sexual dimorphism is observed at all ages
* The climate of specific abalone
* A general rule of thumb to easily estimate the age of the abalone

Regression analysis has been of great use in predicting and modeling a variety of problems. Applying this technique to a biological problem is not a unique solution, however it is a good solution due to the ability to maximize the accuracy of the prediction and the variety of model types we can employ in order to determine the best prediction of the age of a specific abalone based on certain most important characteristics.

In this paper, we will explore the data in order to look at general trends in abalone populations and their characteristics, determine correlated characteristics, and to determine an equation via multiple regression to accurately and precisely determine the age of the abalone using measurements only.

**Related Work**

Abalone age prediction based on their characteristics has been worked on in the past by a few people looking to solve the same issue.

Notably, Runzue Guo, Junmin Luo and Wuqi Gao from the Xi’an Technological University in Xi’an China published a paper on this exact dataset in the Journal of Physics: Conference Series back in 2021. They came to the conclusion that their multiple regression model was able to accurately predict the age of abalone and discussed the significance of their results. They had identified that length is insignificant and that there were abnormal points that needed to be accounted for. This is the same conclusion reached by Harshitha Indumathi and Nishit Patel in their own works on Medium and Github respectively.

Harshitha focused on how well the Ordinary Least Squares (OLS) regression model was able to represent the abalone dataset and provide predictions. Concluding that OLS by itself is sufficient and other techniques such as Synthetic Minority Oversampling Technique (SMOTE) and Cross Validation were unnecessary and did not improve performance of the models, and that a robustness regression model was sufficient to handle the outliers.

Nishit was also focused on determining the best model to predict these and determined that there were actually errors in the dataset, most notably two infant abalone had heights of 0, and 153 abalones had whole weights smaller than the total of their shucked weight, viscera weight and shell weight combined. This should not be possible and is likely due to the heights not being recorded for the first instance, and the weights being improperly recorded for the second instance.

**Proposed Methodology**

In this paper, we shall draw upon the past works to inform our first steps in our Exploratory Data Analysis. We will visually find trends, find errors and clean up the dataset before we move into using those found trends to conduct joint probability distributions in order to more accurately quantify those trends and root out characteristics that are not going to be helpful to feed into the regression model. As part of this data cleaning effort, we will also consider adding, subtracting or changing certain measurements to better represent out data. As an example, we shall create an Age column where we have added 1.5 to the number of rings as this finds the accurate age of the abalone, which is the measurement we would like to determine. We should conclude this stage by performing Principal Component Analysis to verify that we have identified the most important and least important characteristics in the dataset.

Following this stage, we shall train a multiple regression model, both the Ordinary Least Squares (OLS) and Maximum Likelihood Estimation (MLE) models, in order to see how well they both predict the data. We will then consider Logistic Regression, K-means clustering, Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Spectral Graph Bipartitioning (SGB) and whether or not they would yield any new or necessary insights in this dataset or if they would not increase accuracy more than already found.

Below is a more concise description of our methodology:

1. Data Collection:

Collect the data from the Abalone Dataset provided on [Kaggle](https://www.kaggle.com/datasets/rodolfomendes/abalone-dataset/data) compiled by the [UCI ML Repository](https://archive.ics.uci.edu/dataset/1/abalone) using multiple data sources.

1. Exploratory Data Analysis:

Visualize characteristic trends using histograms. Use of Joint Probability distributions to determine likelihoods of certain physical characteristics occurring together. Conduct Principal Component Analysis (PCA) on the dataset to verify and highlight the most important variables and their relationships.

1. Model Training:

Conduct a Linear Regression Analysis to predict the age of the abalone. Logistic Regression will then be used to try to determine the probability an abalone with certain characteristics will be of a certain gender or age.

1. Evaluate and Improve Model Performance:

After developing the initial models, it is important to improve upon any issues found and evaluate which models to draw results from. This can include removing any variables found to be insignificant and reconducting a Regression Analysis based on different calculations.

**Experimental Setup and Results**

Initial Abalone Dataset

| **Characteristic** | **Description (units)** | **Type and Role** |
| --- | --- | --- |
| Sex | Male, Female, Infant | Categorical - Feature |
| Length | Longest Shell Measurement (mm) | Continuous - Feature |
| Diameter | Perpendicular to Length (mm) | Continuous - Feature |
| Height | With Meat In Shell (mm) | Continuous - Feature |
| Whole Weight | Whole Abalone (grams) | Continuous - Feature |
| Shucked Weight | Meat Only (grams) | Continuous - Feature |
| Viscera Weight | Gut Weight (after bleeding) (grams) | Continuous - Feature |
| Shell Weight | Shell Weight (grams) | Continuous - Feature |
| Rings | Add 1.5 to get age | Integer - Target |

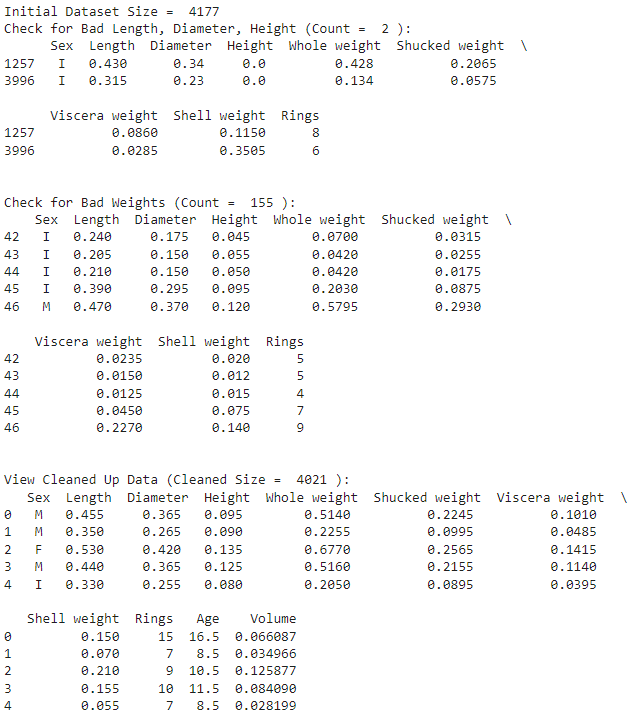
Looking at the characteristics provided, we can see some promising characteristics and anything we should note about them.

1. Sex will let us determine if the distribution of the other characteristics is different for male, female and infant abalones.
2. We can use Length, Diameter and Height to determine the Volume of the abalone in mm3 which we can use to get a whole visualization on the difference in volume of each abalone and not just in any one direction. It’s important to note that since Length is just the longest shell measurement, it’s not going to always be consistent which direction that measurement is in. This is not something we can account for, but it would be interesting to explore whether or not this has any effect on the prediction in future works. For now let us assume that it is always in the same direction and therefore has no effects on our conclusions.
3. We will also take a look at the weights and determine how much it correlates with the size of the abalone. Does the whole abalone get more weighty, is it only the shell and meat?

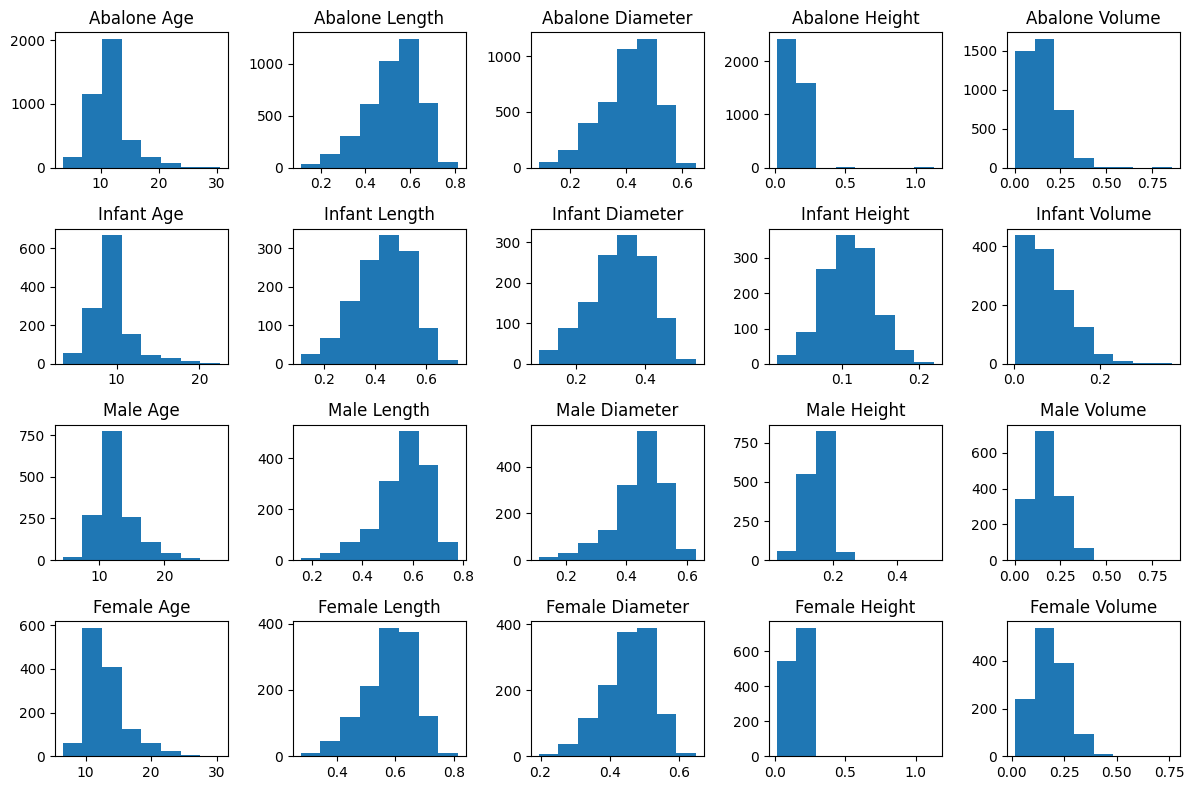
The first change I will make is to check for any errors in the data. I will do this by checking for any abalone with length, diameter, or height equal to 0. We find two cases of this. Similarly, I will look for any issues with the weights. Namely, any weights that are equal to 0 or any total weights that are less than the sum of the shucked, viscera and shell weights. We find 155 cases of this. Notably, it appears that the only error with weights is the latter case where the sum of the other weights is greater than the total weight measured. This should not be the case as the Viscera Weight is calculated after bleeding, which means that some weight should be lost and unaccounted for in our tables. Both of these issues were noticed in prior works, however they were not removed from the dataset. In this work, we will remove these errors from the data set as while they may be small amounts compared to the whole dataset, any error could throw off prediction results which is not the intended goal.

The second change I will make to this dataset is to add an age column since this is our true target and the Rings is only a measured characteristic that we use to infer the age. I will also be adding a Volume characteristic as mentioned before with the units in mm3 with the calculation being for the volume of an oval, which is the rough shape an abalone takes.

After making these changes and creating a cleaned up dataset, here are the resulting outputs:

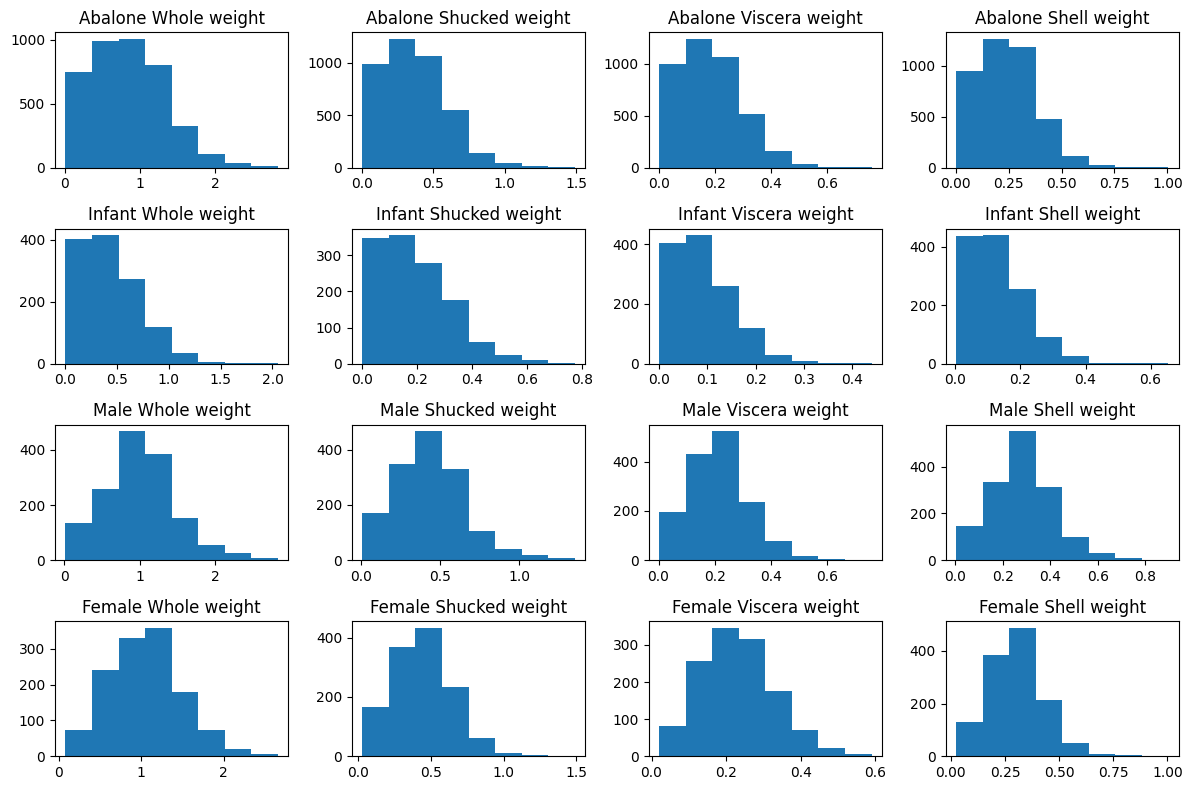


Now, with this cleaned dataset we can look at the distributions in the Age, Length, Diameter, Height, and Volume of the Abalones using a histogram to get general trends and to easily compare the shapes of the graphs:

****

Looking at the distributions, it’s interesting to note that the ages are not all that different between the Infant, Male and Female abalones. They tend to be below 20 and hover around 10 years of age. We can see that the Male and Female datasets are remarkably similar. Ignoring the Height graphs since we can clearly see outliers on them, we are able to conclude that the main difference between the Male and Female, or Adult, versus the Infant abalones is actually their total volumes. Infants tend towards the below 0.2 mm3 volumes, whereas the Adults tend to center around 0.25 mm3 with some outliers being much higher. From an initial glance, it looks like the differences between Adult and Infant here is really their volumes and not their ages, this could call into question how accurate the Sex Categorical variable is.

Before we move to clean up our data further, we should take a look at the weights:



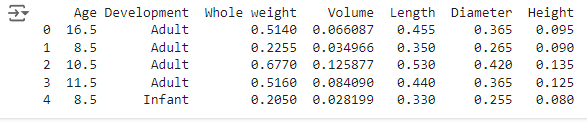
Similar to size, we see little differences between Male and Female abalone’s weights. We see that Female abalone tend to have more viscera than Males, but less meat and lighter shells. We do see here a difference in the Adult versus Infant abalone’s weights. We see that infants tend towards being lighter than 1 gram, whereas Adults tend to be around 1 gram.

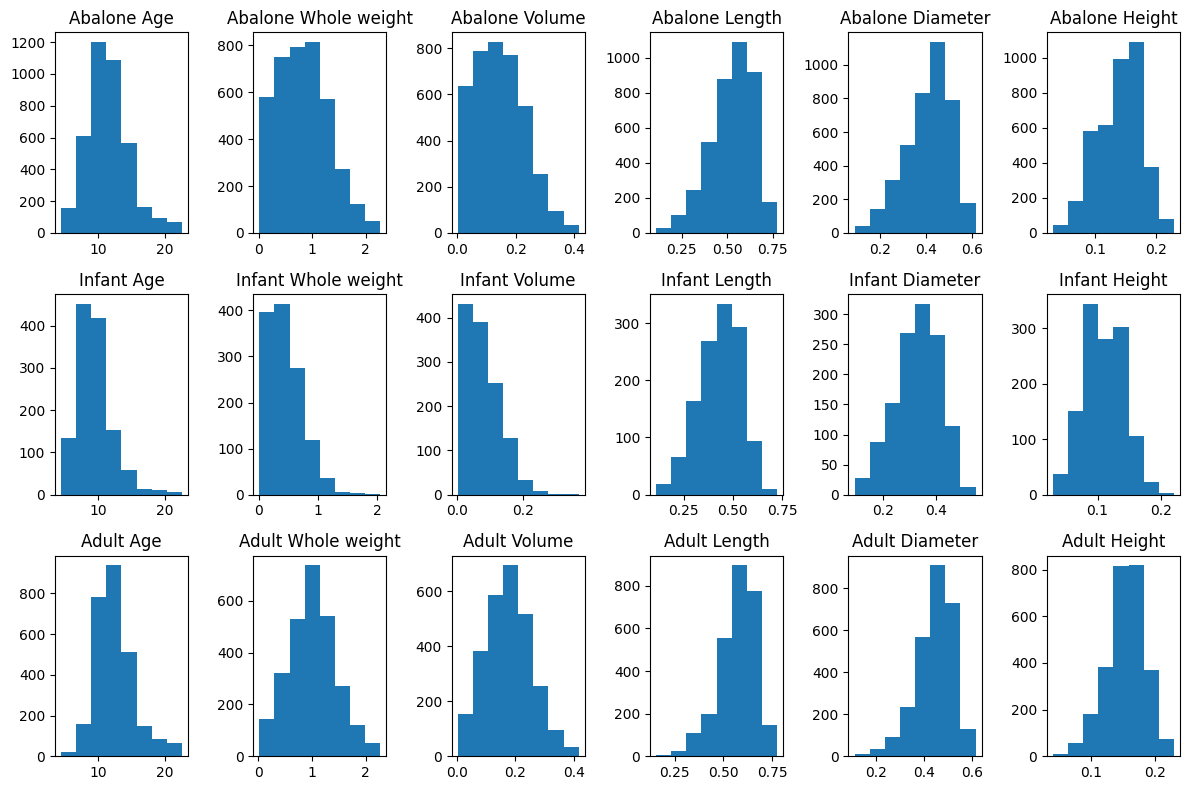
Using both of these results, we can conclude the next data cleaning steps we need to take.

1. Take the outliers out of the dataset. This should be an easy step to increase the accuracy of our model. Without this step, we will have some overestimations or underestimations for age since we will be using the Whole Weight in the final dataset..
2. Combine the Male and Female datasets into one Adult dataset, creating a new categorical variable Development containing Infant and Adult labels only
3. Create a dataset of only Development, Age, Length, Diameter, Height, Volume, and Whole Weight since Shucked, Viscera and Shell Weight are not determinable prior to harvesting the abalone and will only impede the accuracy of the regression model. Additionally, Sex has been changed into Development to more accurately fit our data and Rings is accounted for by Age, so it is not necessary to keep in the data set.

This clean up is important because it’ll increase the accuracy of our model, and most importantly, we will have isolated the only variables that we can realistically rely on to develop our model. Without this, we would see some errors on edge cases and some drift for all of our estimates that the regression model will produce.

We will remove heights greater than 0.23 mm and less than 0.03 mm, remove any Whole weights that are greater than 2.3 grams, and lastly any ages greater than 23. This is all the trimming we need to do in order to clean this data. With these changes, our entire dataset looks like this:

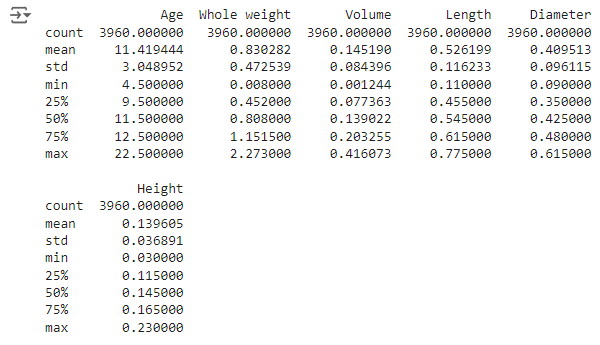




Now we can look at some probabilities that we can use in our conclusions. Looking for Abalones that are 15 years of age or older, we find this:

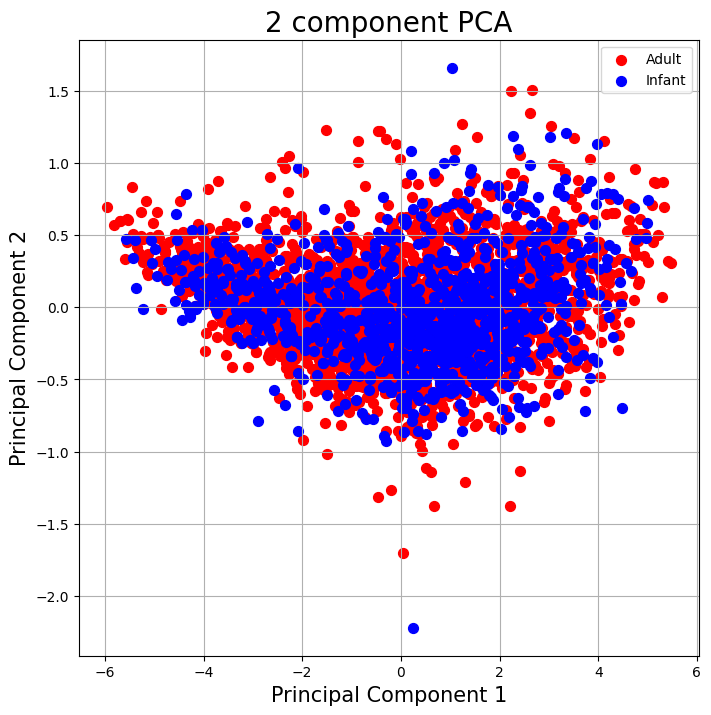
| **Age** | **Weigh >**  **1 gram** | **Volume >**  **0.2 mm3** | **Longer than >**  **0.5 mm** | **Diameter >**  **0.5 mm** | **Height >**  **0.2 mm** |
| --- | --- | --- | --- | --- | --- |
| **5+** | 35.48% | 26.34% | 62.75% | 15.98% | 3.01% |
| **10+** | 33.46% | 25.48% | 54.29% | 15.63% | 3.01% |
| **15+** | 6.41% | 5.13% | 9.75% | 3.21% | 1.06% |
| **20+** | 1.19% | 0.88% | 1.67% | 0.66% | 0.18% |

From this table, we can easily tell that all of these values are correlated, however height might be the least affected from a 5 to a 10 year old Abalone since the chance that an abalone being taller than 0.2 mm did not change between these years. Diameter is similar, but the probability decreased by 0.35% so there was a small growth within the 5-10 age range.

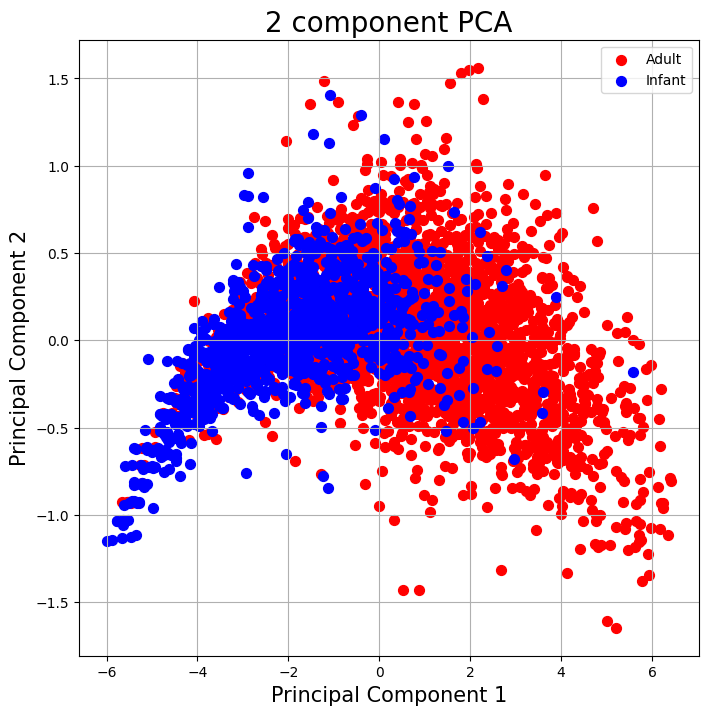


Looking at some summary details of the dataset, including maximum, minimum and mean, we can see the actual values which a histogram does not show you.

Throwing our dataset into a Principal Component Analysis (PCA) algorithm, removing Age since it’s our actual target and Development because it’s a text value, we are able to see that one component explains 94.64% of our variance and a second explains 2.602% of our variance. These two components give the following graph.

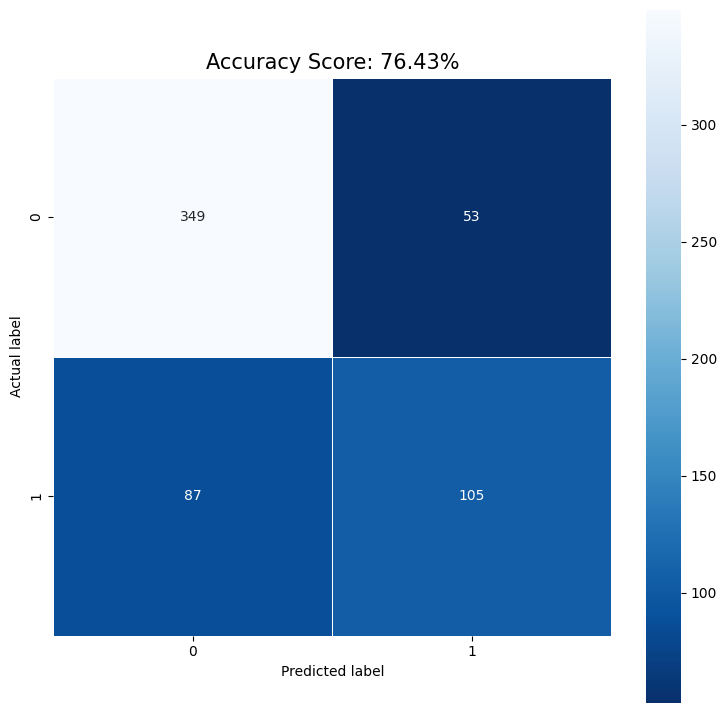


Comparing this to our original dataset, we see that the two Principal Components explain 90.79% and 3.99% of our dataset’s variance, and producing this graph:

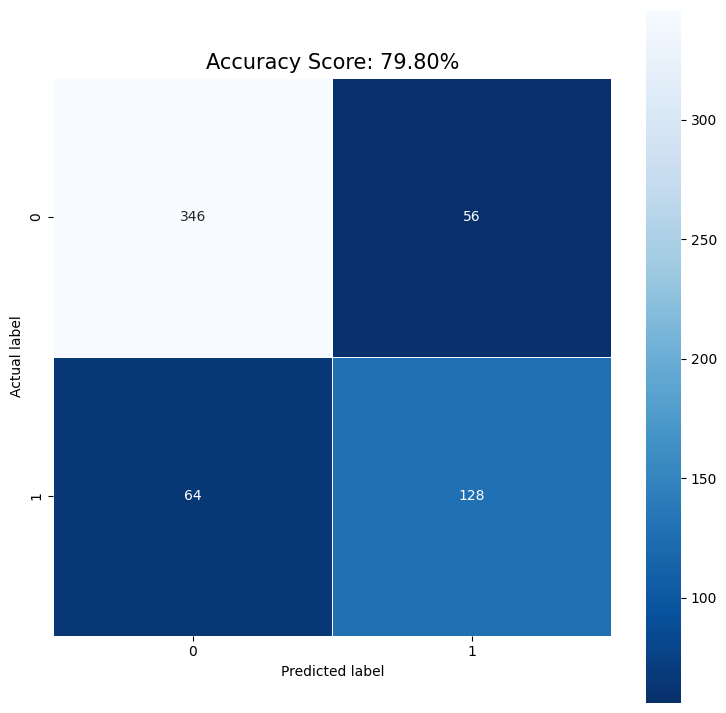


The two graphs look incredibly different, and it’s clear that there is less explained by the two principal components, which means our data cleaning efforts have been successful at reducing the variability and pinpointing specific variables which should be focused on. This should result in a more accurate regression model.

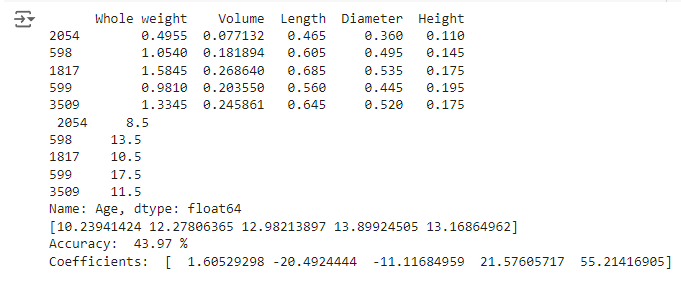
Putting the cleaned data into a PCA that aims to account for 95% of the variance before putting it into a Logistic Regression model, we get an accuracy of 76.43%. The 0 label is for Infants, whereas the 1 is for Adults.



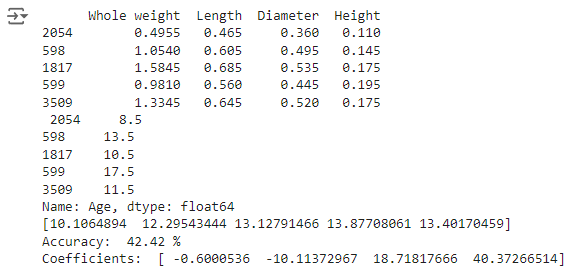
Instead of putting it into a PCA and putting our cleaned data into the Logistic Regression model immediately, we see an accuracy of 79.80%.



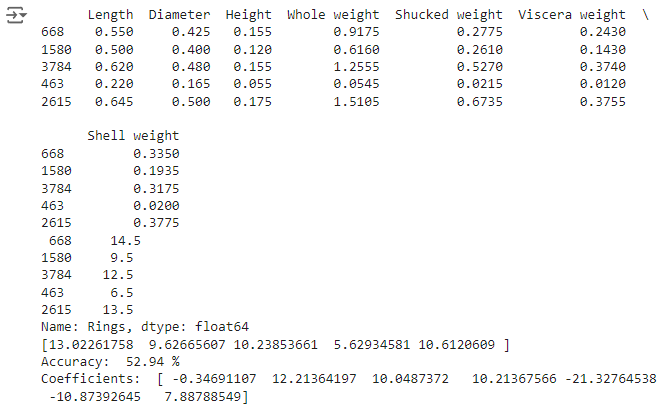
Moving to a Linear Regression model to predict ages, we are see that the cleaned data set has an accuracy of 43.97%, guessing that the age of a set of 5 abalones are 10.2, 12.3, 13, 13.9, and 13.2 years old when in reality they are 8.5, 13.5, 10.5, 17.5, and 11.5 years old.



To adjust or explain some of the inaccuracy, let’s remove the Volume variable since it’s related to the Length, Diameter and Height variables. This leads to an accuracy decrease to 42.42%.



When using our original dataset however, we see that we have an accuracy of 52.94%.



**Comparison**

As compared to our study, previous studies have found more success with their regression models. Runzue Guo, Junmin Luo and Wuqi Gao found that they had a strong relationship between the Abalone’s age and the multiple independent variables they used. In their paper, they decided on using the Rings, Diameter, Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight data. In this paper, we only used Whole Weight, Volume, Length Diameter and Height in the cleaned dataset, and in the whole dataset we differed from them by removing Rings, and including Height and Length. Running a regression model with their variables, we get a 100% accurate model.

****

This difference is likely due to the inclusion of the rings variable, which is the target of the dataset. Additionally, the Shucked, Viscera and Shell Weight data were removed from the cleaned dataset because they would not be measurable until the Abalone had been harvested, killed and prepared, which would defeat the purpose of the regression model as there would only be the staining and counting rings step left to determine the age of the Abalone.

Harshitha similarly concluded that their model was accurate, and notably that nothing more than a robustness regression model was enough to handle the outliers. They also went with a different route than the rest, including me, when dealing with the sex variable. They made two variables, s1 and s2, which corresponded to S1 = 1, S2 = 0 if Male and S1 = 0, S2 = 1 if Female. Infants had S1 = S2 = 0. Unfortunately it is impossible to determine which variables they used in the model, however there is no mention of removing any so I will assume they only removed the Rings variable so as to not give the model the answer in the training data.

Nishit Patel is another relevant study that found their model has accurately predicted the abalone’s age and number of rings. They also developed 3 candidate models to test. The first model consisted of the Sex, Height, Diameter, Whole, Viscera, Shucked and Shell Weight of the Abalone. The second contained the Sex, Length, Height, Diameter, Whole, Viscera, Shucked and Shell Weight of the Abalone. The third contained the Length, Height, Diameter, Whole, Viscera, Shucked, Shell Weight and Infancy status. They had found that their third model was their most accurate model. Similar to above, our main difference was the exact variables we included in our regression models vs the ones they included in theirs. Unlike the previous paper, I could not find an issue with their work and can only attribute the inaccuracies in this paper to poor data cleaning or model development.

Another difference between this paper and the rest is what we excluded from the data set. I removed any values that did not make sense and removed outliers from the cleaned dataset to remove errors or skews due to those outliers. This is different from the rest as they either make no mention of removing the errors or outliers or do mention that they were not significant enough to remove.

**Conclusion**

This paper failed to accomplish what it set out to do, however it has served as a learning experience and potentially highlights errors within the dataset provided. There has not been an accurate linear regression model developed and we were unable to draw any complete conclusions from the data as there was significant overlap in the characteristics of interest. The logistic regression model highlights the significant overlap between all of the characteristics that an Adult and an Infant Abalone share, which makes it difficult to guess whether or not an Abalone is one or the other. Additionally, there were outliers noted early on that could only be attributed to human error and ambiguity of the datasets variables. The length variable is not always measured from the same side, it is instead the longest measurement which could have an effect on the data. Additionally, it’d be a good idea to include the species of the Abalone for which the data comes from as that could make a difference in the expected age or other characteristics we would use to find the age. Lastly, there could be more data included in the dataset. The total of 4177 data points is a limiting amount to create as accurate as possible models with. In a real world scenario or for an actual model to provide to workers harvesting Abalone, it’d be ideal to have more data to give guesses based on.

Regarding improvements to this paper’s methodology or future ideas to explore, there is evidently an issue with how the models were developed, in either the data cleaning, or in what was fed into the linear regression model. This could be improved upon in future iterations, such as following what Harshitha did and making the Development variable a 1 or a 0 for Adults or Infants since Harshitha found this to be a significant enough variable to rely on. Additionally, plugging the data into an Artificial Neural Network (ANN) might be a good idea since each variable can be given a weight by the ANN which gets updated until it finds an optimal value and can continue learning through more tests and training data being provided. A K-means clustering or Support Vector Machine (SVM) algorithm could also be used to identify groups of Abalone who share a lot of characteristics in common with each other. Ideally this would be used to help clean and assign the data additional characteristics that may prove helpful in other goals than a regression model.

Both of the accuracy values for the Logistic Regression model on Adult vs Infant Abalone are bad, not even correctly guessing whether the data predicts that the Abalone is an Adult or an Infant > 20% of the time. Normally, this would mean we need to improve our data or improve the model itself, however in this case it might be due to the very similar characteristics of Adult and Infant Abalone. While there certainly is a difference between the two, their characteristics have a lot of overlap in each one of the variables we provided to the Logistic Regression models and the PCA. This significant overlap provides that ~20% error that we see as there are Abalone characteristics that would be characterized as both Adult and Infant depending on the study that the data is from. I would suggest an improvement to this dataset is to go through the entire dataset and make a basis for what is considered an Infant and what is an Adult before changing the Sex variable. This would be a minor improvement. We do however see that the results of our trimming has created a training dataset that is more accurate than the PCA result of that dataset.

Lastly, regarding the inaccuracy of the Linear Regression model for the Age and Ring count of the Abalone, we saw less than 50% Accuracy on models that did not include characteristics that required harvesting of the abalone and only 52.94% when we included those characteristics. This shows that there was either an error with the models or the data was insufficient. As at least one paper has drawn an accurate conclusion based on the same data, and another found similar results though the exact model cannot be confirmed, we shall conclude that the models themselves were of poor quality and that the data, if used correctly could create a sufficient model.

Unfortunately this means that the paper has not accomplished it’s originally stated goals of determining:

* Which abalone to harvest
* Factors that affect growth rate of individual abalones
* How growth rate affects specific physical characteristics
* Whether sexual dimorphism is observed at all ages
* A general rule of thumb to easily estimate the age of the abalone

Due to this, another study should be conducted taking into account the errors made in this one and the suggestions discussed in order to make an accurate model and explore what the data can be used to find in addition to these original goals.

**Acknowledgements**

The author acknowledges support from Dr. Haiyan Wang in introducing the materials and techniques used in this paper.

**Author Contributions**

Kyle Culp was the sole author of this paper, writing all the code used to generate the plots, creating the experimental methodology and analyzing the data to come up with the conclusion.

**Data Availability**

The Abalone Dataset provided on [Kaggle](https://www.kaggle.com/datasets/rodolfomendes/abalone-dataset/data) compiled by the [UCI ML Repository](https://archive.ics.uci.edu/dataset/1/abalone) using multiple data sources.

The python code used to generate all of these graphs and outputs is listed on [Kyle Culp’s MAT422 GitHub repository](https://github.com/KyleEtera/MAT422/blob/d656c1c09a3e8a8edbd4548adbf547f27f88596e/MAT422_Final_Project_Abalone_Age_Prediction.ipynb).

**References**

**(1)** Guo, Runze, et al. “A new method of measuring the age of abalone based on data

visualization analysis.” *Journal of Physics: Conference Series*, vol. 1744, no. 4, 1 Feb. 2021, p. 042181, <https://doi.org/10.1088/1742-6596/1744/4/042181>.

**(2)** nishitpatel01. “NISHITPATEL01/Predicting-Age-of-Abalone-Using-Regression:

Predicting the Age of Abalone Using Multiple Regression in R.” *GitHub*, [github.com/nishitpatel01/predicting-age-of-abalone-using-regression](http://github.com/nishitpatel01/predicting-age-of-abalone-using-regression). Accessed 27 Oct. 2024.

**(3)** Indumathi, Harshitha. “Abalone Age Prediction.” *Medium*, Medium, 26 Feb. 2021,

[harshithaindumathi.medium.com/abalone-age-prediction-607ecda3fa3b](http://harshithaindumathi.medium.com/abalone-age-prediction-607ecda3fa3b).