**Assignment 1: K-Means Clustering**

**Team:** White Walkers

**Student Names:**

Sourabh Namilikonda (SJSU ID: 012414948)

Hemambujam Veeraraghavan (SJSU ID: 012482392)

Pratik Dhumal (SJSU ID: 012457185)

Hrishikesh Rendalkar (SJSU ID: 011485994)

**Business Objective:**

Abalone are a family of small to very large [sea snails](https://en.wikipedia.org/wiki/Sea_snail) [1], which is a delicacy and who’s shell have been used for jewelry, for ages. Traditionally, to estimate the age of an Abalone:

1. There is the counting of external growth checks on the shell.
2. And the counting of internal growth checks within sections of the shell.

This estimation is time consuming and painstaking. We, as part of assignment 1, decided to minimize this estimation period by calculating the age of an Abalone using its physical attributes.

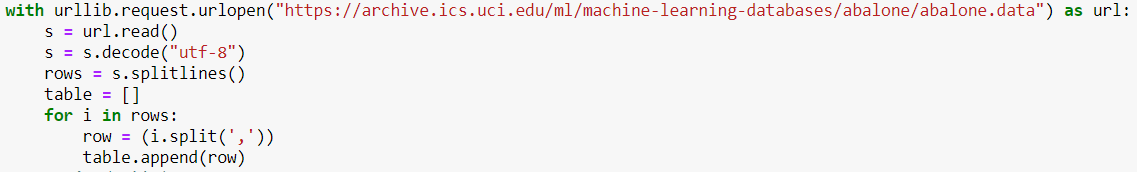
**Dataset**:

We came across the dataset that provided us with sample data to calculate the age using the height and weight of an Abalone. This dataset had ample data, with about 4000 sample data, which served to be useful for the K-Means clustering. The dataset can be accessed at:

<https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/>

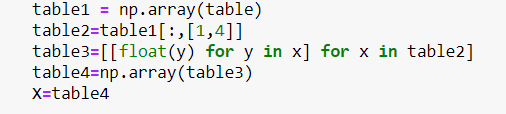
**Data Wrangling:**

Using data from the above-mentioned dataset, we converted data from a plain text format to a list of lists in Python, that can be used for our assignment purpose. Below is a code snippet on how data was accessed:

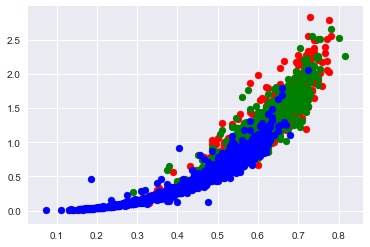


**Data Preparation:**

Using the NumPy library in python, the data was prepped using the following code snippet:



Plotting the prepared dataset before performing K-means clustering, just to ensure we could choose the right K value, looked as thus:



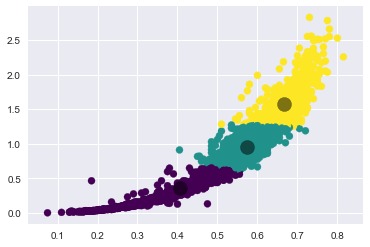
**K-Mean Clustering:**

Using the dataset, we wanted to make a K-Mean, where given the height and width of Abalone, we would be able to classify the input into:

1.Infant

2. Juvenile

3. Adult



**Data Interpretation:**

From the above image, we can interpret that the age of an abalone can be directly proportional to its height and weight. The three centroids segregate the dataset into three sub-sets making it easy to understand the data better from its visualization.

**Business Solution:**

With this K-means clustering, it solves the problem of having to check on the internal and external growth of shells on an abalone. This also save a lot of time, with just measuring the height and weight of an abalone will be able to plot it on the graph which would give an estimate on the age of the abalone.

**References:**

[1] <https://en.wikipedia.org/wiki/Abalone>