Assignment template 5) Cluster Abalone data into distinct groups

This assignment follows on from 'Practical 8. Pandas data discovery - Cluster analysis'. The main objective is:

Cluster observations of Abalone into distinct groups and assess the profile of age in each group

The basic premise is that we should be able to group different observations of the Abalone shellfish into distinct clusters that should have a different age. This follows on from the challange of being able to predict the age of abalone from physical measurements. As noted in the original repository, _"the age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task."

The original dataset can be found atUCI Machine Learning repository (https://archive.ics.uci.edu/ml/datasets/abalone) but we have placed it in the data folder for ease of access.

Overview: You are tasked with:

- Loading a dataset of Abalone features into a Pandas dataframe.
- Expanding this to include a calculation of Albalone age as a new column in the dataframe.
- Present a boxplot of cluster member age that have been predicted using K-means for 5 clusters on all the original features.
- Repeat the above but with 7 clusters.

According to the variable definitions provided on the link given above, we are given 8 variabes:

Name	Data Type	Measurement Unit	Description
Sex	nominal	--	M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer	--	+1.5 gives the age in years

You need to create a new variable in our dataframe, 'age', which is calculated using the number of rings as per the information in the table. Based on the example given in class, we can break this exercise down into a number of steps:

- 1. Loading a dataset of Abalone profiles into a Pandas dataframe
- 2. Create a new column with values from calculating the age of abalone
- 3. Perform K-Means cluster analysis assuming 5 distinct clusters.
- 4. Assess any change in member properties by increasing the number of clusters to 7.

Please note:

We will discuss this in class, but aside from a working notebook we are also looking for the following:

- An associated narrative with each operation. This includes the following sections:

> Abstract

  - Summarise the project and main results

> Introduction and methdology

  - What is the challenge and how are you solving it?
  - What modules/functions are you using?

> Results

  - What is happening in each figure?

We also want to see adequate referencing around:

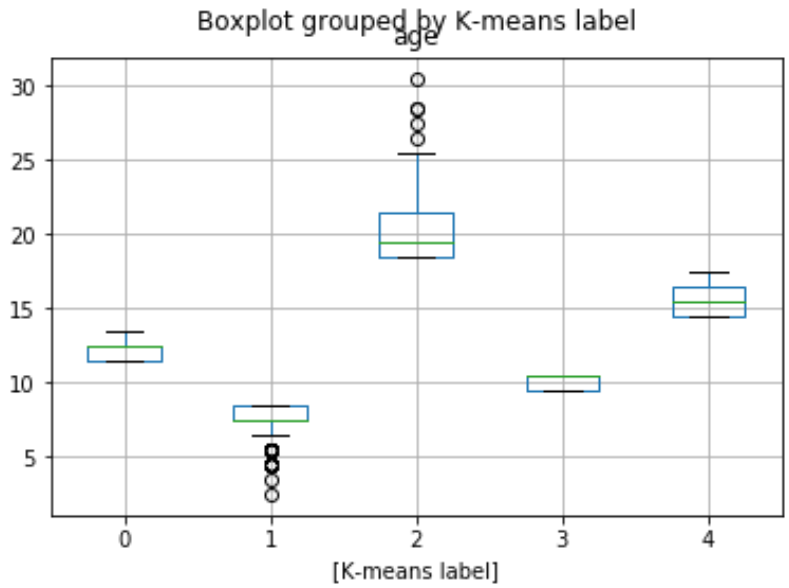
- What is the original source of the theory and/or data?
- Comments in the code boxes using the # symbol. Remember that someone might not know what each line of code does.

You may also want to consider a broader discussion around this challenge. For example:

- What could be improved?
- How do you know your results are correct? [i.e. what might improve trust in your simulations?]
- What if someone wanted to get in touch with you and re-use this code? Any restrictions on data?

To start, we reccomend you first get the code implementation working and then construct the narrative around it. Also please note that to add another code or markdown box, you can simple use the 'Insert' option on the main menu.

Your boxplot of cluster age should resemble the following figure:



Also please note that you can load your data, once the Pandas module has been loaded, bya:

```
data = pd.read_csv("data/abalone.csv")
```

Abstract

Introduction

Methodology

In [ ]:

```
#### -------- INSERT CODE HERE ----------


#### ------------------------------------
```

Results

references