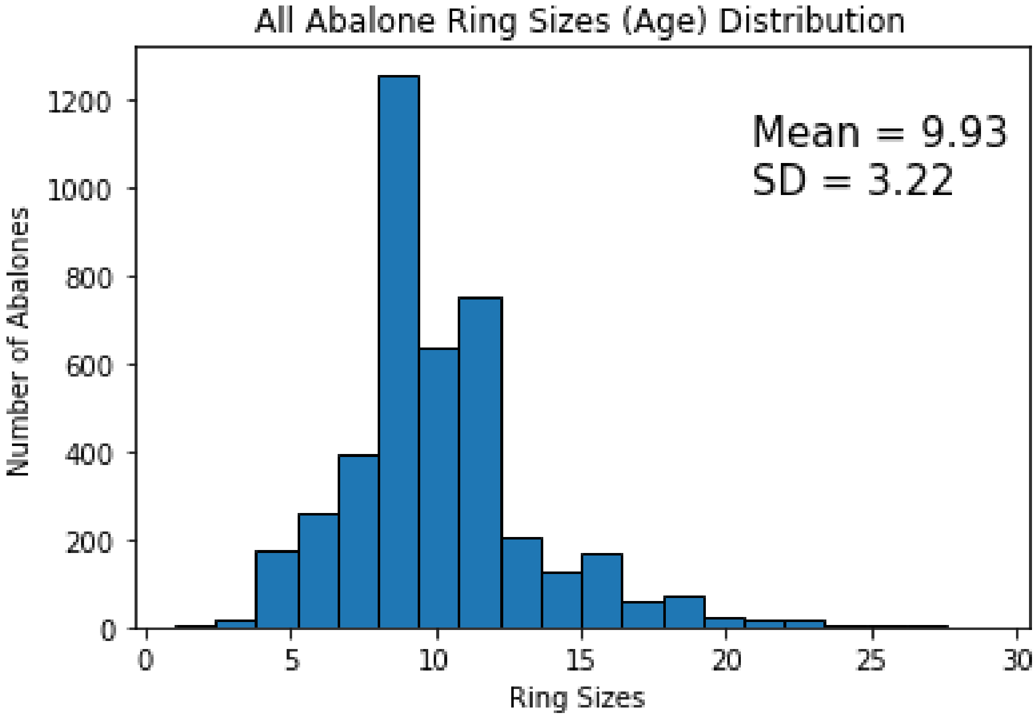
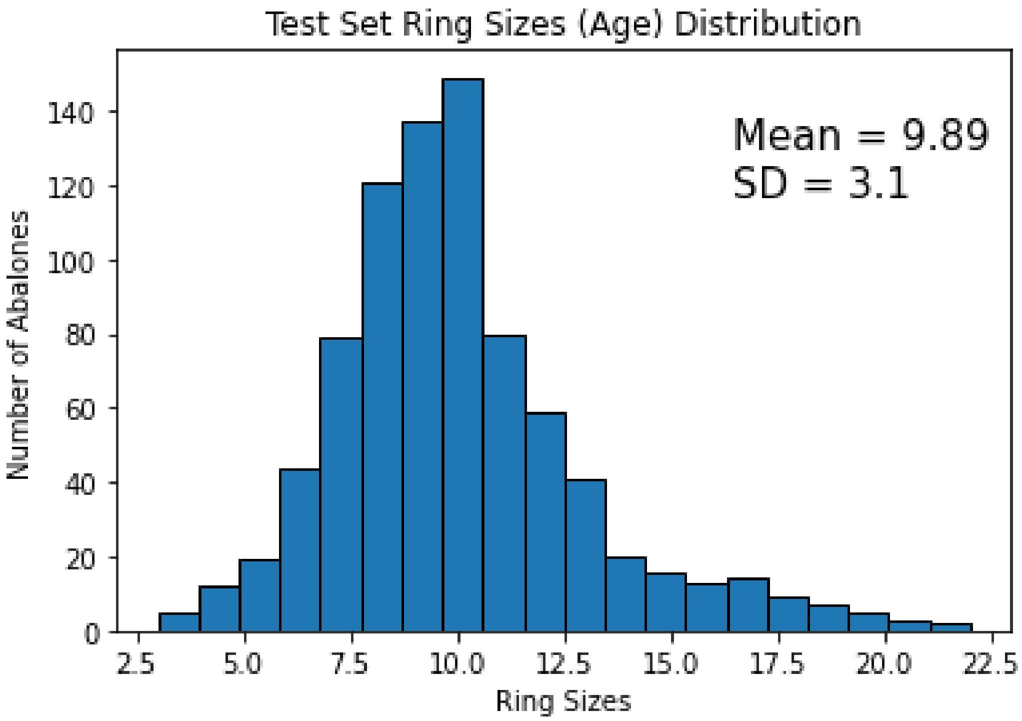
The age of an abalone can be found by cutting its shell and counting the number of rings on the shell. In the Abalone Dataset (abalone.txt), you can find the age measurements of a large number of abalones along with a lot of other physical measurements.
The physical measurements of an abalone include: "Sex", "Length", "Diameter", "Height", "Whole weight", "Shucked weight", "Viscera weight", "Shell weight", "Rings", among which “Sex” cannot be used as a features, but a label(class).
The goal of this exercise is to develop a model that can predict the age of an abalone based purely on the other physical measurements, except ‘sex’. This would allow researchers to estimate the abalone’s age without having to cut its shell and count the rings.

(First, plot the rings (ages) distribution of all abalones in the dataset and the randomly sampled 20% abalones Test Set, like: (hint: You don’t have to plot the rings (ages) distribution of the 80% Training Set.))





You'll be applying a **kNN** to find the closest prediction score possible. Please test various k in [3,5,7,9,11] to see the result.

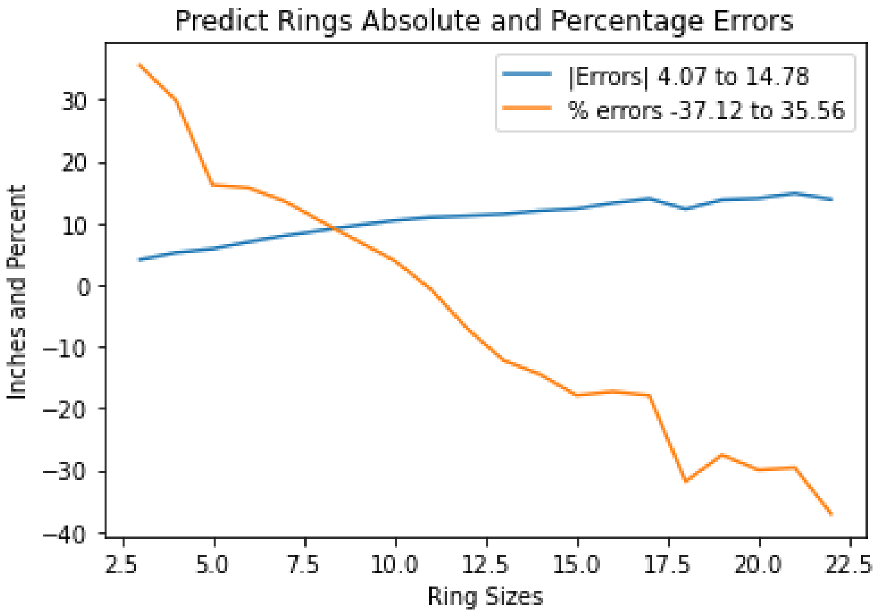
First, for each k pre-trains (hint: similar to findK function in chap2408.py) a model with the whole abalone samples and used it to get Coefficient of Determination R^2 and Root Mean Square Deviation of the pre-trained model, print the following messages and plots the corresponding figures like:

```
Pre-Training with Whole Examples Evaluation with k= 3
     Coefficient of Determination: rSquare(R²):    0.2352
     Root Mean Square Deviation Rmsd:    2.4959
```

Then, for each k uses all the samples of every ring size (age) in the Test Set to predict each of their ring size **by using kNN**, and collects all the **Absolute errors of the predictions**. On the other hand, also calculates the total **prediction percentage error** to the actual ring size(age) of all the abalones of each ring size. Print out the following messages:

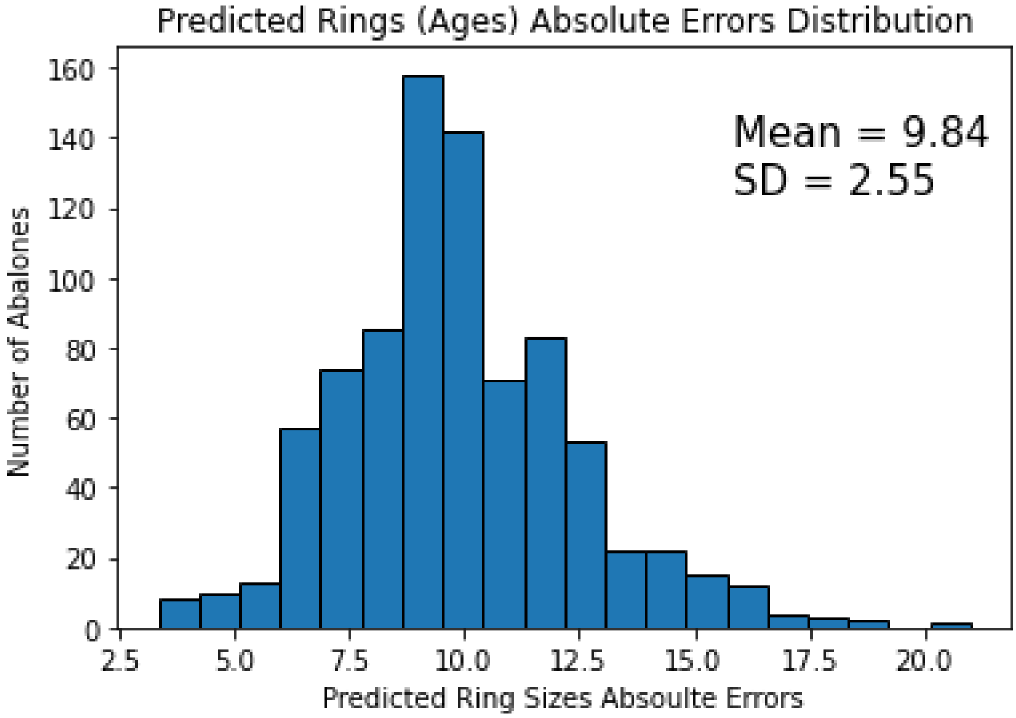
```
After Trained Testing Using Test Set with k= 3
     Coefficient of Determination: rSquare(R²):    0.5122
     Root Mean Square Deviation Rmsd:    2.2665
```

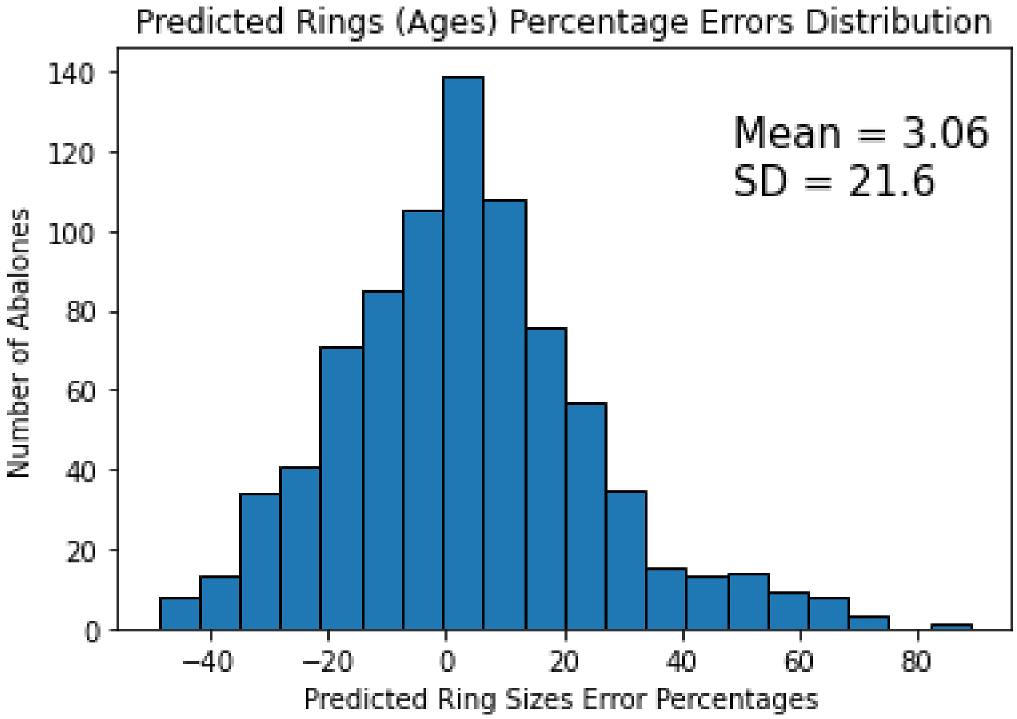
Repeat above works for k=[3,5,7,9,11] and plot the following three figures for each k.



(Hint: to produce the above figure, you have to count and save the number of abalones of each ring size(age) in **a list of lists** (hint: each list is for a ring size) for the Test Set)

(For each k, plot the histograms of the **Absolute Errors of the predictions** and **Percentages Errors of the predictions** of the test Set)





K=5 ….
K=7 ….
K=9 ….
K=11 ….

Finally, decide the value of k for the KNN clustering algorithm for the model and print:

The maximum R^2 value for trained model is 0.530154709013795 , happens in k = 9