**Abalone Case Study /Blog Submission**

The purpose of this article is to summarize the predictive modeling process, from exploring the data to deploying the prediction. This includes the thought process behind various decisions during the process, be it selecting the features that could help with better prediction accuracy or engineering certain features to help get more details out of the data.  In this analysis we seek to understand the distribution of the dataset attributes, as well as the relationship between them.

The analysis is divided in six sections: on section 1 we briefly present what an abalone is. On section 2 we present the Abalone Dataset and their attributes. On section 3 we will perform the analysis of each attribute individually. On section 4 we seek for correlations between the attributes and how the segmentation affects the results. On section 5 we will build our Model based on observation & analysis done. Finally, on section 6 we present our conclusions.

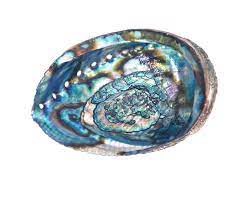
**Table of Contents:**

1.      Problem Definition  
2.      Data Analysis  
3.      EDA Concluding Remarks  
4.      Pre-processing Pipeline  
5.      Building Machine Learning Models  
6.     Concluding Remarks

**Problem Definition:** 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. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.Based on the information given in our dataset we have to predict the rings of each abalone which will lead us to the age of that abalone.

**What is an Abalone?**

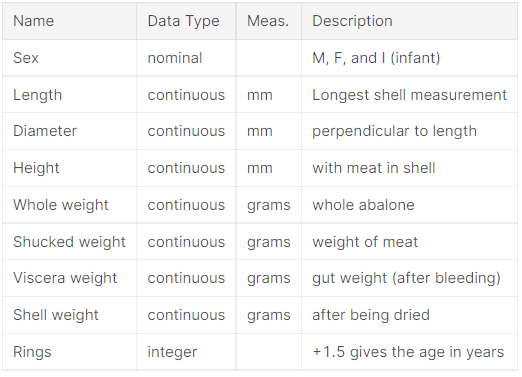
Abalones are marine snails. Their taxonomy puts them in the family Haliotidae which contains only one genus, Haliotis, which once contained six subgenera. These subgenera have become alternate representations of Haliotis. Abalones can be found along coasts of almost every continent. Usually, abalones are consumed as food all around the world, by different cultures. However, the bright and variety of colors of the interior side of their shells makes them an valuable object of adornment and decoration.

****

**Data Analysis**:We will be using the already collected Abalone dataset to see the algorithms in action. The first step in knowing the data is to know what it contains. This means understanding the type (continuous numeric, discrete numeric or categorical) and meaning of each feature and noting down the number of instances and features in the dataset.

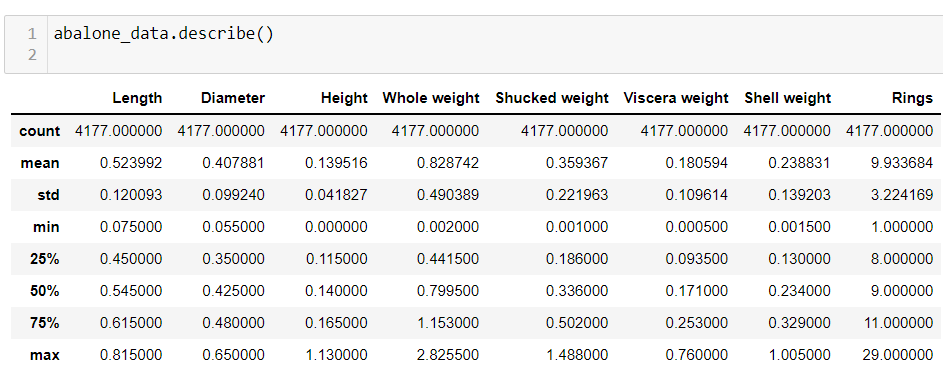
*A brief aside on the motivation behind collecting the dataset. Abalone is a type of consumable snail whose price varies as per its age and as mentioned here: The aim is to predict the age of abalone from physical measurements. The age of abalone is traditionally determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age.* *Thus, one possible solution is predict the number of rings of an abalone from characteristics like height, diameter, lenght and weight measurements.*

The Abalone Dataset is composed of the following attributes:



In this study, each one of these attributes were analyzed, as well as the relationships between them.

Now we will import necessary libraries to know the dataset.

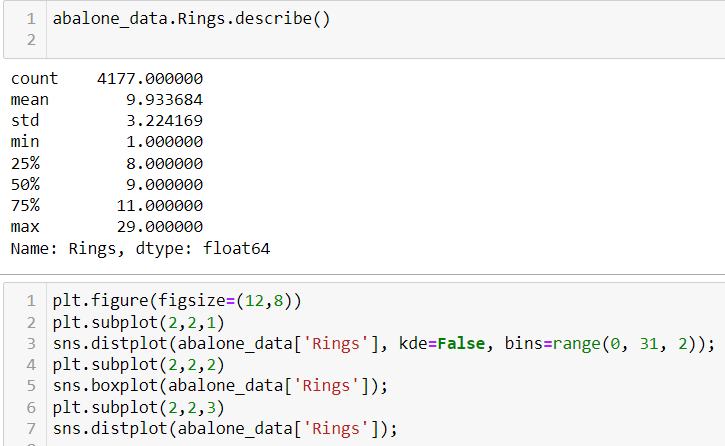


U**nivariate Analysis**

In this section the distribution of each attribute is analyzed individually. We start analyzing the distribution of the target attribute *Rings*. The rest of the attributes are divided in groups for convenience of the analysis: a group called *Size*, containing attributes that represents the dimensions of an abalone, a group *Weight*, containing the different weight attributes and a third group composed only of the *Sex* attribute. The continous or quantitative attributes were analyzed using histograms and boxplots, while categorical attributes were analyzed using barplots.

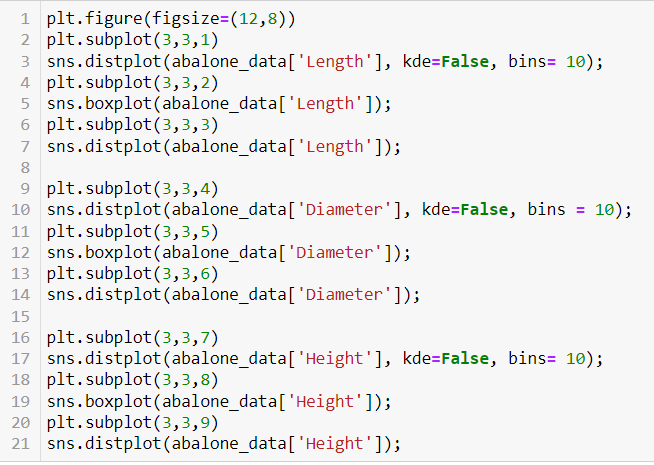
### **The Target Attribute**

The analysis shows that the Ring attribute values ranges from 1 to 29 rings on an abalone specimen. However, the most frequent values of Rings are highly concentrated around the median of the distribution, so that, the 2nd and 3rd quartiles are defined in a range of less than 1 std deviation. We observe that its possible to approximate the distribution of this attribute to a normal curve.

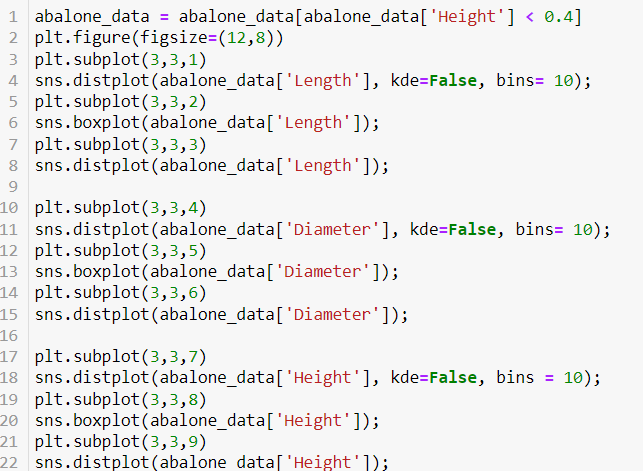


### **Size Attributes**

In this section, we analyze the attributes that represents the dimensions of an abalone. These attributes are Lenght, Diameter and Height. For each of these attributes we plotted two histograms and their respective boxplot. The first histogram is a density histogram and a kernel density estimate, and the second is the absolute frquencies of the attribute, with some adjustments to ticks and bins. Once more, we observe an approximate normal distribution. However we notice the high peak formed by the Height histogram. Analyzing the Height boxplot we conclude that the high peak is formed due the presence of two observations that lie far beyond the central positions of the distribution.



From the above, we filter these outliers in order to obtain a more realistic visualization of the distribution of the Height attribute



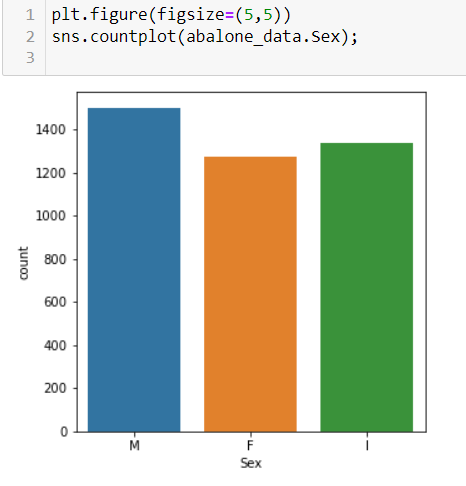
### **Weight Attributes**

The weight attributes were analyzed following a similar approach to the Size attributes analysis. A similar distributions were observed, however, for the weight attributes the bell curve is a little larger.



### **Sex**

The Sex attribute is a categorical variable for which the possibles values are: M for Male, F for Female and I of Infant (an abalone which is not adult). We analyzed the count of each category with a bar plot, and concluded that relative to this attribute, the dataset is balanced.

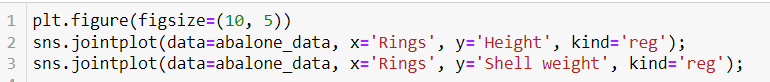


Multivariate Analysis

In this section we analyzed how the dataset attributes are related and how the independent variables influences the target variable. Our first step in the multivariate analysis was to visualize the correlation matrix in a heatmap:

How the independent variables influences the target variable.We need to visualize the correlation matrix in a heatmap

Analyzing the correlation matrix, we notice that *Height* and *Shell weight* are the attributes that most correlates to *Rings*. Therefore, we concentrated the multivariate analysis on the correlation of these two attributes with *Rings*:

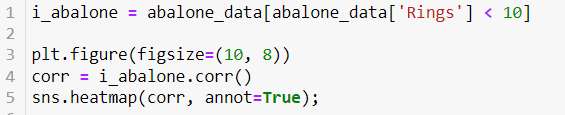


Height and Shell weight are the attributes that most correlates to Rings

Correlation of these two attributes with Rings

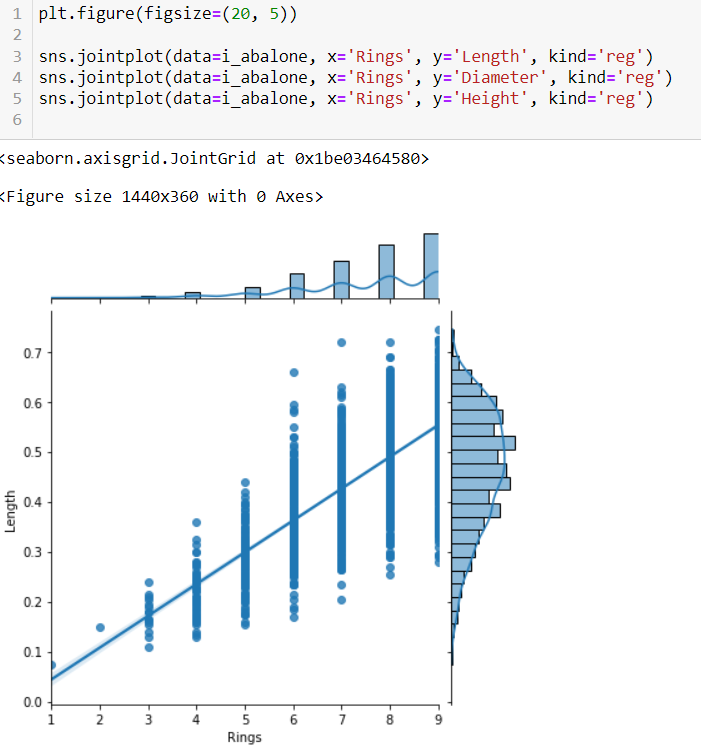
### **How correlation varies with the number of rings?**

Based on the previous analysis, we decided to investigate the variation of the correlation regarding the number of rings in more detail. We tested for many values, and found that the region delimited by Rings < 10 has greater correlation between the independent attributes and the target variable.

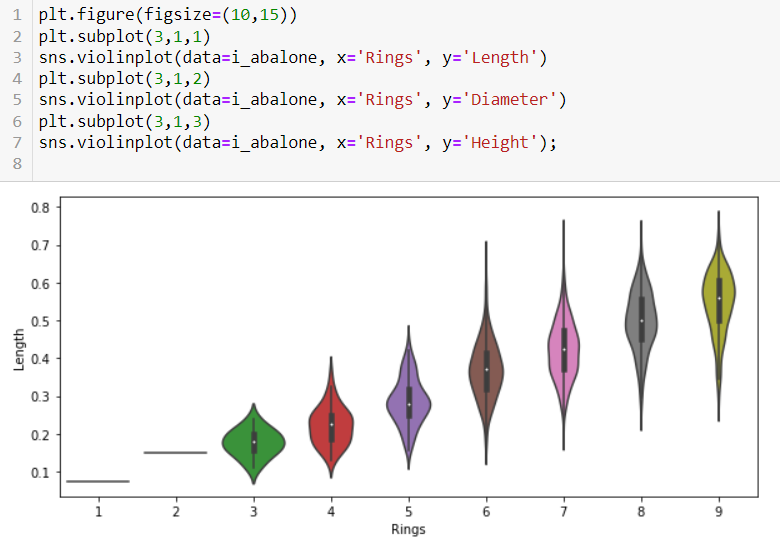


In the following plots we observe the positive correlation between the *Size* attributes and *Rings*

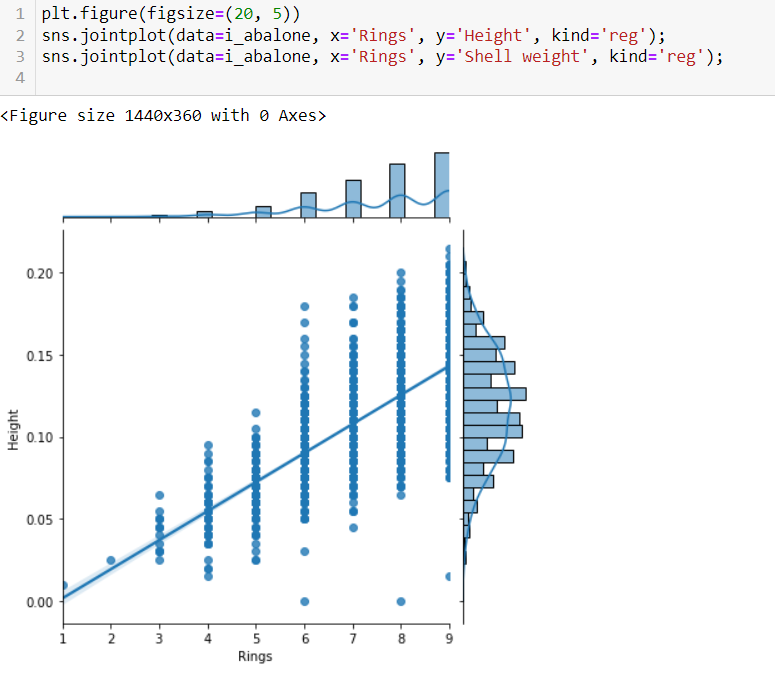
**Lets see the correlation between size attributes and Rings.**

****

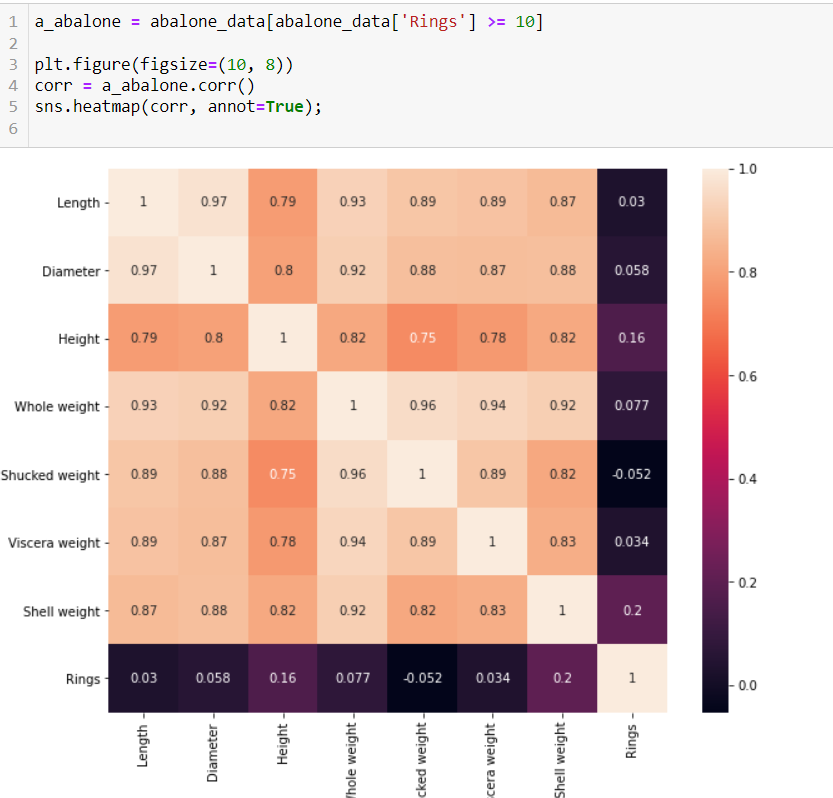
The violin plots bellow show that the median of *Size* attributes increases as instances are grouped by *Rings*:

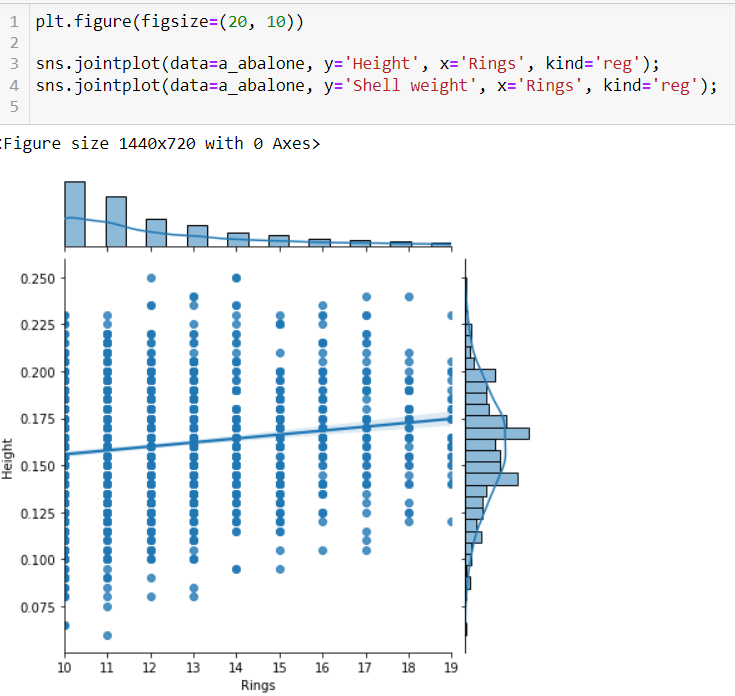


**Comparing Height and Shell weight to Rings:**



With more than 10 Rings we observe that correlation decays drastically to near 0 (zero)



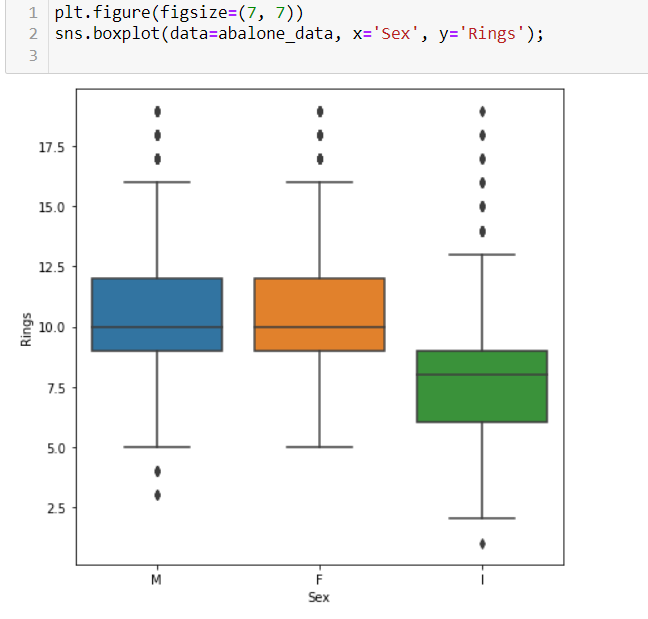


These results suggest that abalones grows in size and weight until a certain age, near 10 years. After this age their size and age remains stable.

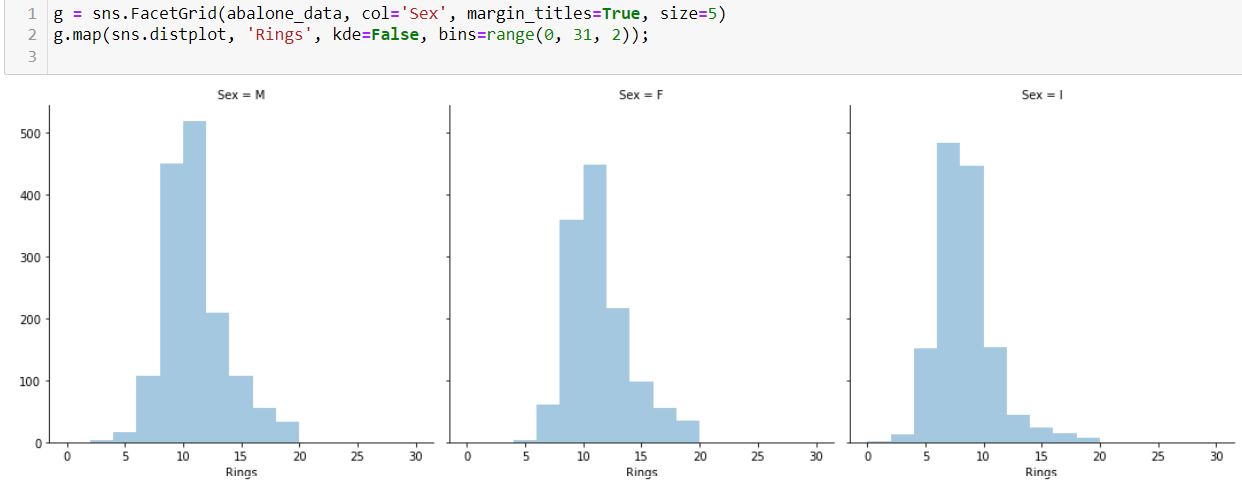
**Influence of Sex on attributes**

Finally, we analyze how the Sex category influences the distribution of variables Rings, Height and Shell weight. Our objective is learn if the different categories of abalones have different distribution parameters or even form. To accomplish this objective, we visualize the distribution of each one of these parameters in relation to Sex. Finally, we analyze how the different Sex categories of abalones influence the correlation of Rings, Height and Shell weight.

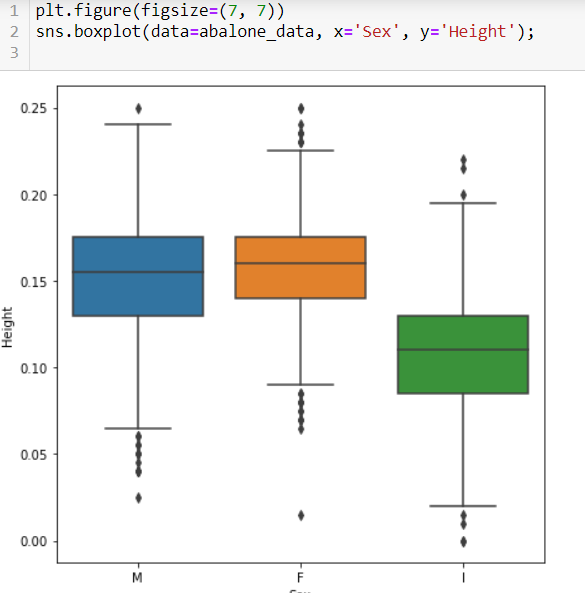
Our first step is to analyze how *Sex* categories influence the number of *Rings*. We observe that the median of *Rings* for the *I* category is lower than the median for *M* and *F* categories.

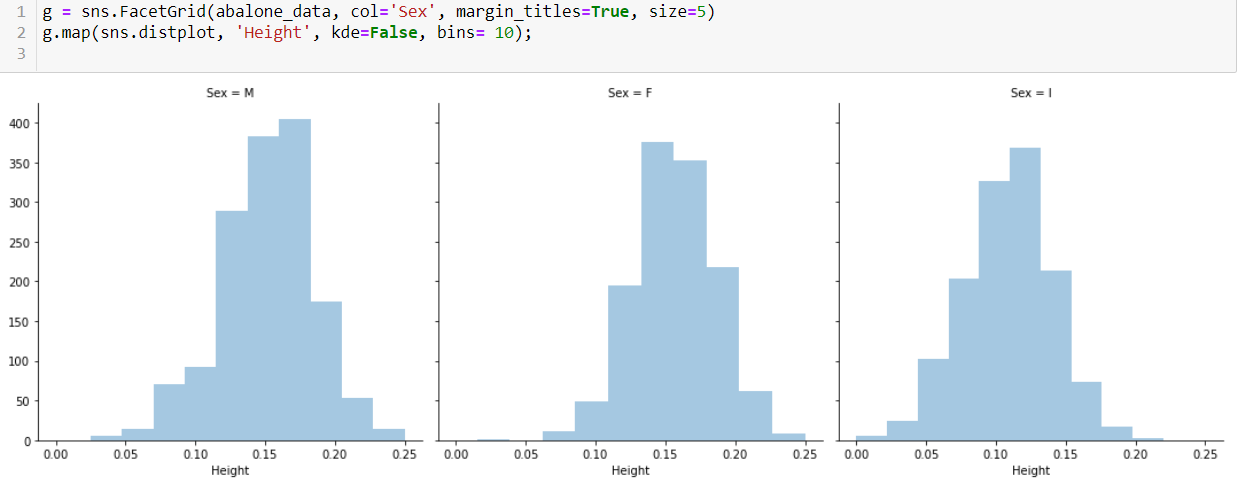


Infants have lower number of rings as compared to males and females

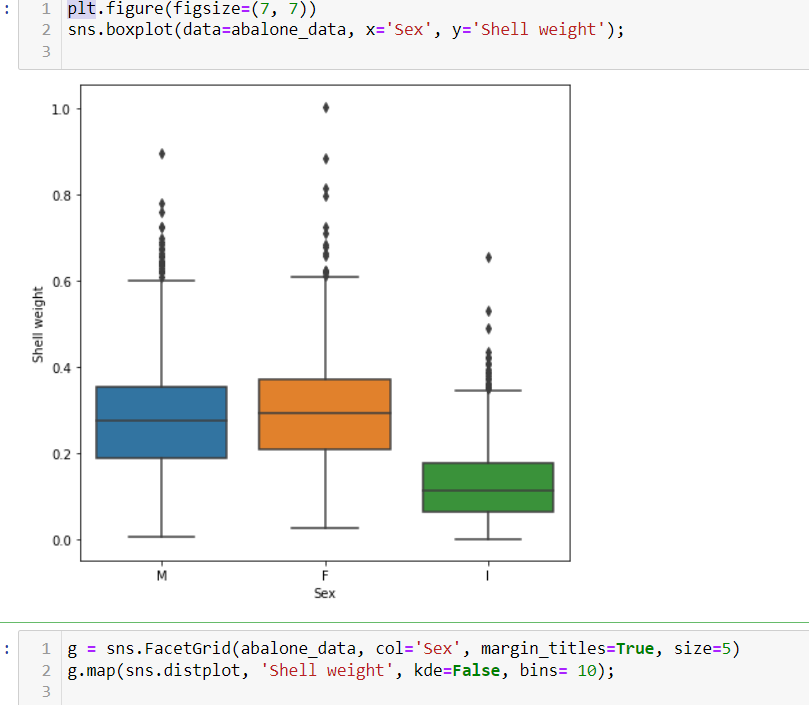


**For shell height**

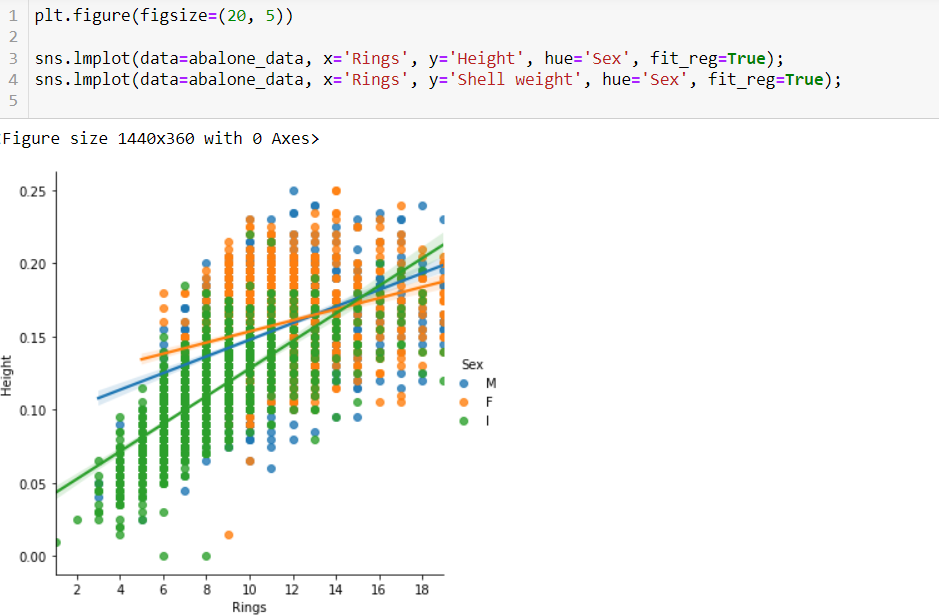
****

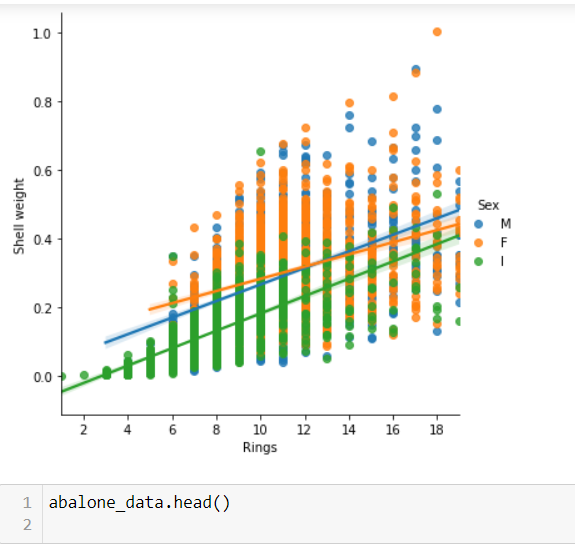


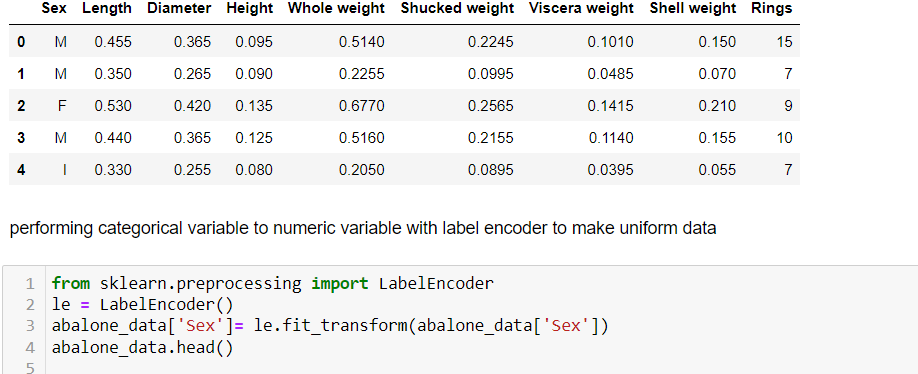
**For weight**

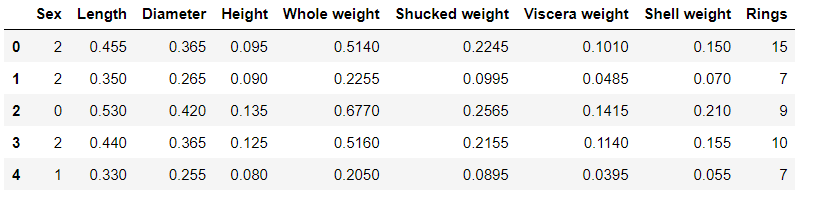


Finally, we analyze how the categories influence the correlations *Ring* x *Height* and *Ring* x *Shell weight*. We already concluded that these attributes has stronger correlation for lower values of *Rings*. Because *Infant* abalones have lower values of *Rings*, the consequence is that *Height* and *Shell weight* have stronger correlation to rings. Observing the regression curve for the *Infant* category, we notice that its inclination is closer to 45°.



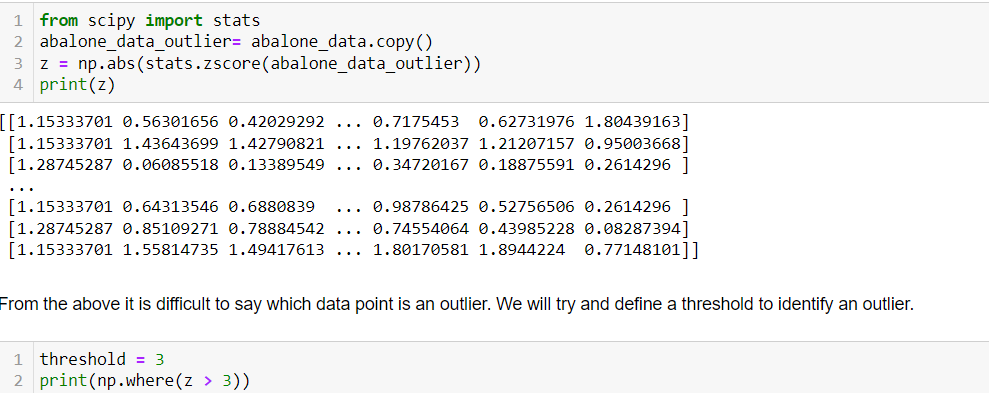


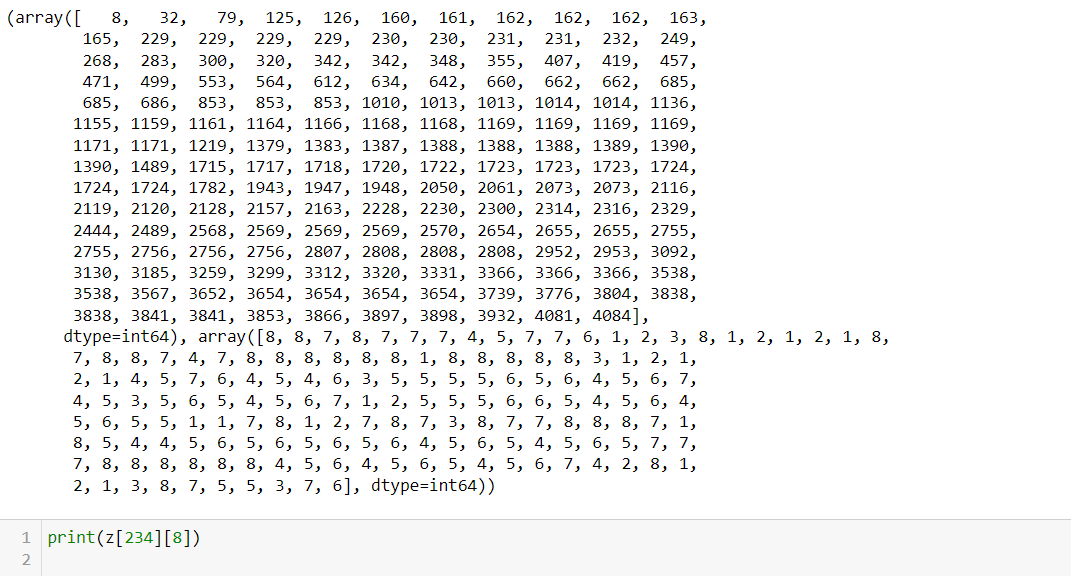


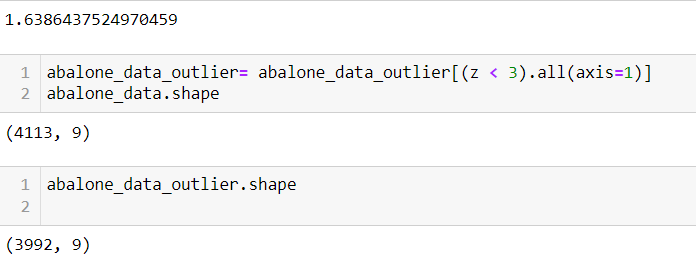


**Outlier detection and removal**

**Z score method**

****

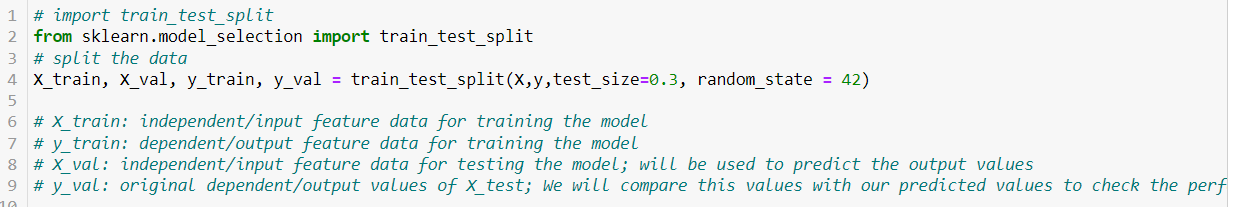
****

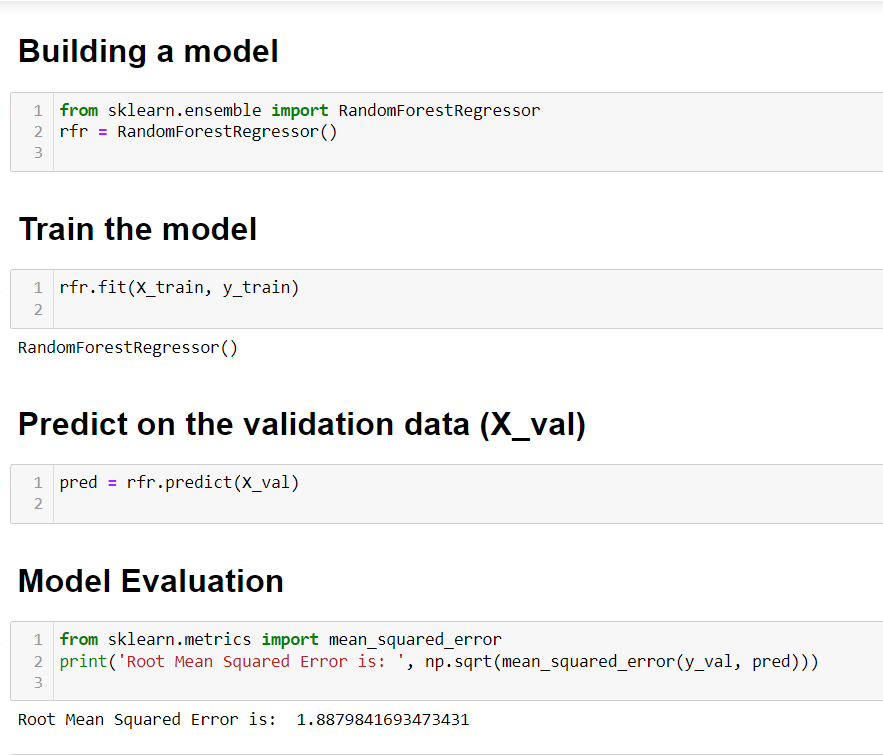
****

**Separating Input Features and Output Features**

****

**Splitting the data into Train and Validation Set**

****

****

**Conclusion**

* By observing the correlation between the target attribute Rings and the indepent variables, we conclude that it is possible to build a model to predict the target value in function of the independent attributes.
* The weight of the Abalones varies proportional to their sizes
* There's no significant differences in size, weight and numbers of rigns between male/female abalones
* The Infant Abalones groups presents lower mean values of size, weight and number of rings
* The weight and height of abalones varies accordingly to age until the adult age, after adult life size and weight stops varying, and after 16.5 years (15 rings) these measurements aren't correlated.