**ASSIGNMENT 4**

**Problem Statement: Abalone Age Prediction**

**Description:-** Predicting the age of abalone from physical measurements. 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 age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

**Attribute Information:**

Given is the attribute name, attribute type, measurement unit, and a brief description. The number of rings is the value to predict: either as a continuous value or as a classification problem.

**Name / Data Type / Measurement Unit / Description**

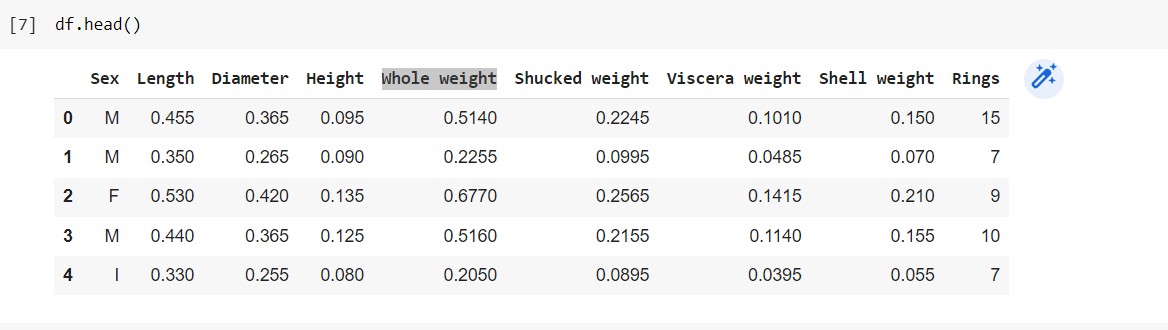
1. Sex / nominal / -- / M, F, and I (infant)
2. Length / continuous / mm / Longest shell measurement 3- Diameter / continuous / mm / perpendicular to length
3. Height / continuous / mm / with meat in shell
4. Whole weight / continuous / grams / whole abalone
5. Shucked weight / continuous / grams / weight of meat
6. Viscera weight / continuous / grams / gut weight (after bleeding) 8- Shell weight / continuous / grams / after being dried

9- Rings / integer / -- / +1.5 gives the age in years

**Building a Regression Model**

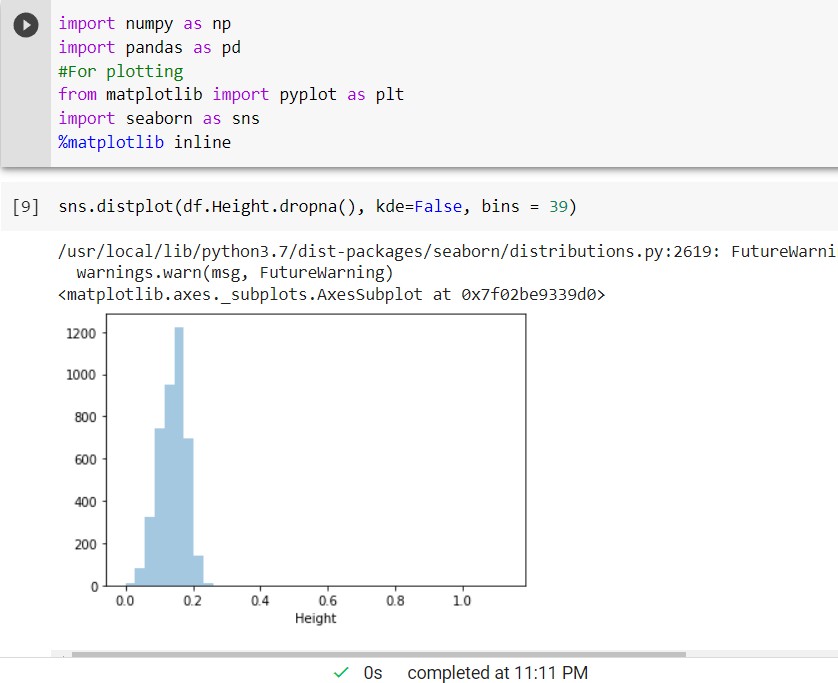
1. Download the dataset: Dataset
2. Load the dataset into the tool.



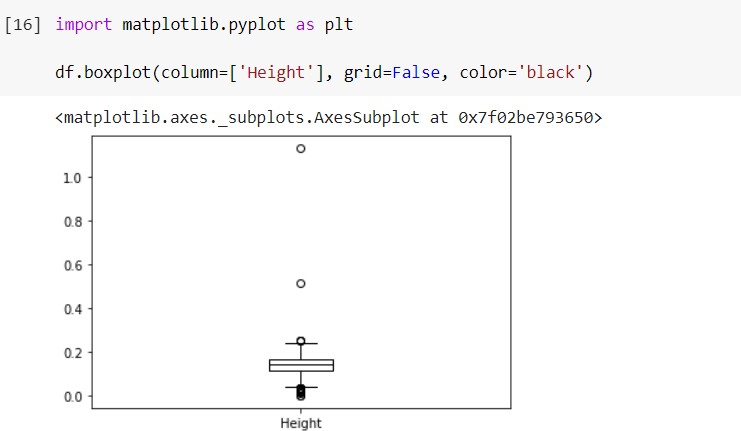


1. Perform Below Visualizations.

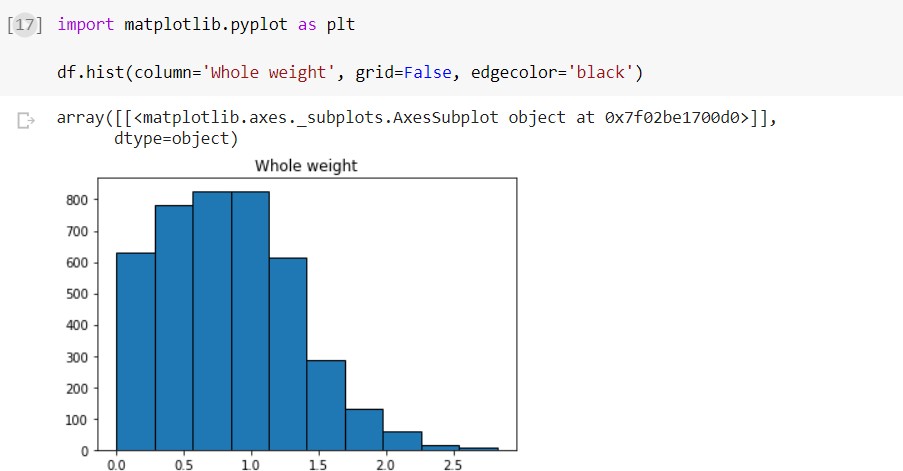
# Univariate Analysis



Box plot:

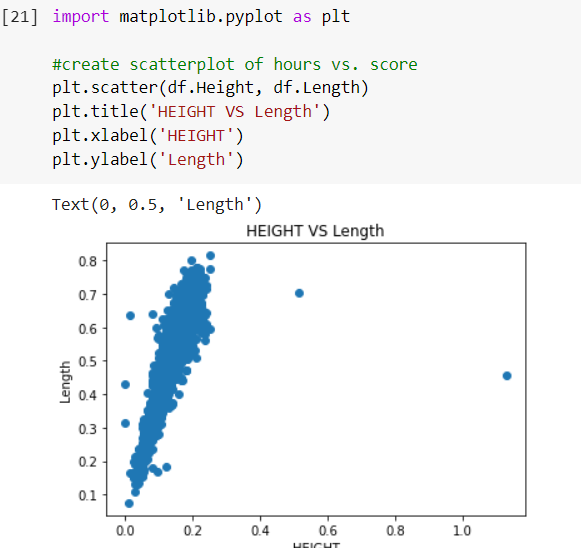


**Histogram:**



# Bi-Variate Analysis

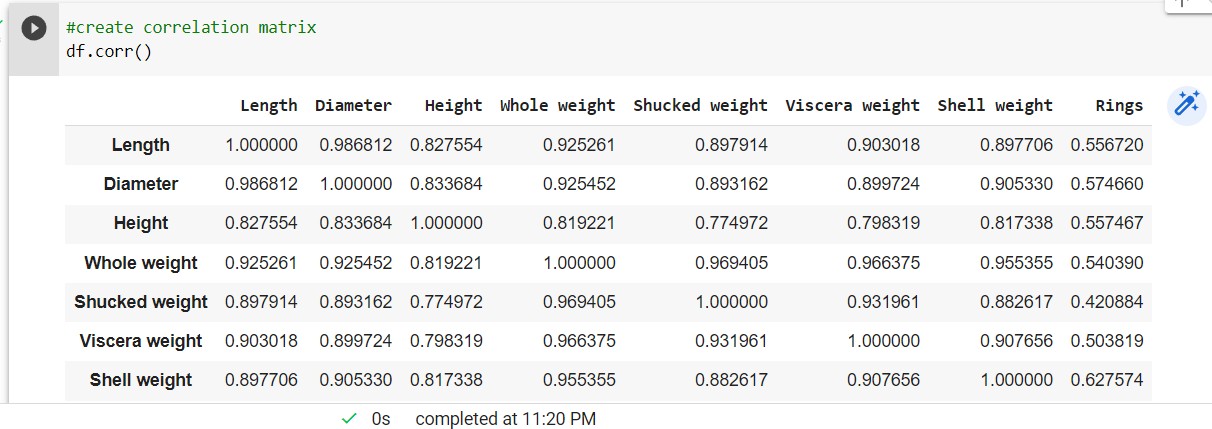
1. **Scatterplots**



1. **Correlation Coefﬁcients**

A Pearson Correlation Coefficient is a way to quantify the linear relationship between two variables.

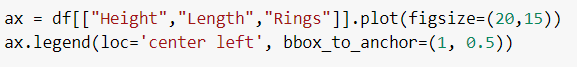
We can use the corr() function in pandas to create a correlation matrix:

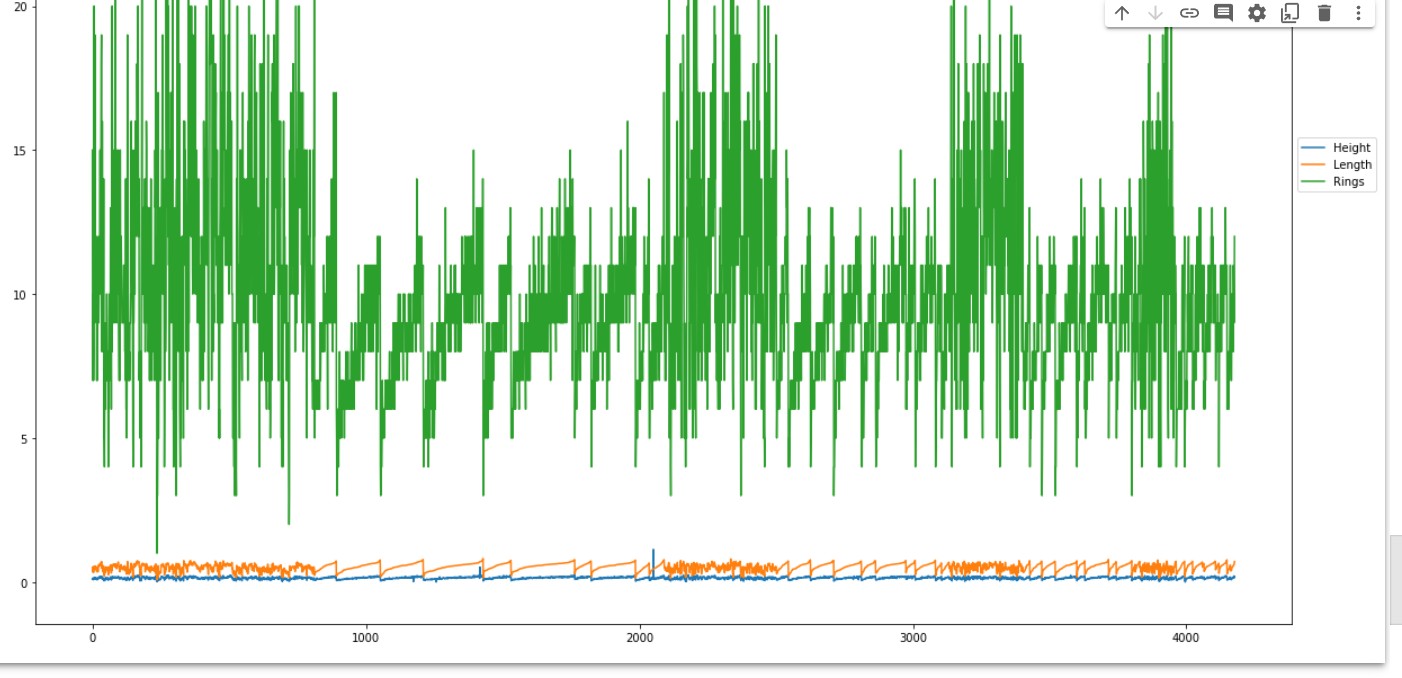


* + **Multi-Variate Analysis**

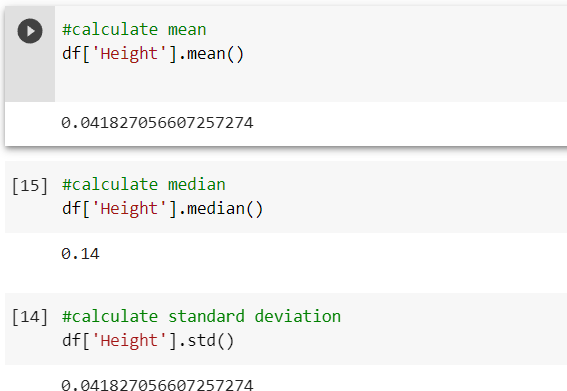
Profile Plot

**Profile plot, used to shows the variation in each of the variables, by plotting the value of each of the variables for each of the samples.**

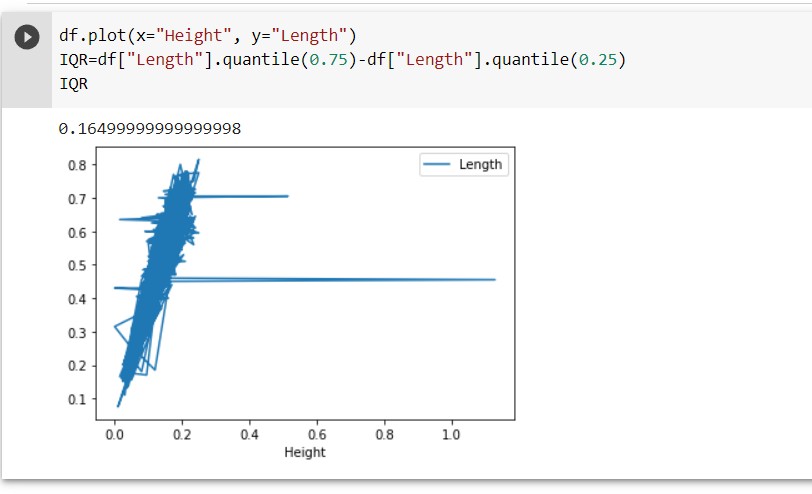




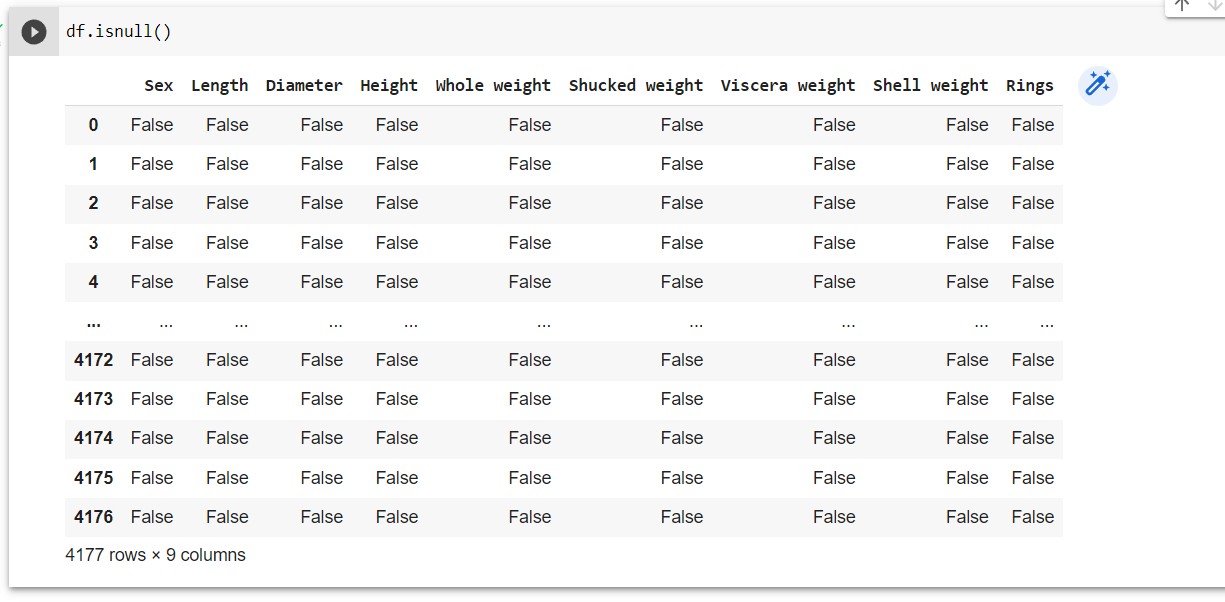
1. Perform descriptive statistics on the dataset.

Mean median and standard deviation.

IQR(difference between 75% and 25% quartile)



1. Check for Missing values and deal with them.



No missing values.

1. Find the outliers and replace them outliers

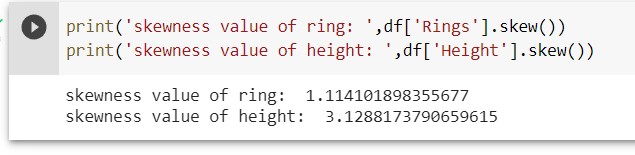
# Outliers Identification

There are different ways and methods of identifying outliers, but we are only going to use some of the most popular techniques:

* + Skewness

# Skewness

the skewness value should be within the range of -1 to 1 for a normal distribution, any major changes from this value may indicate the presence of outliers.



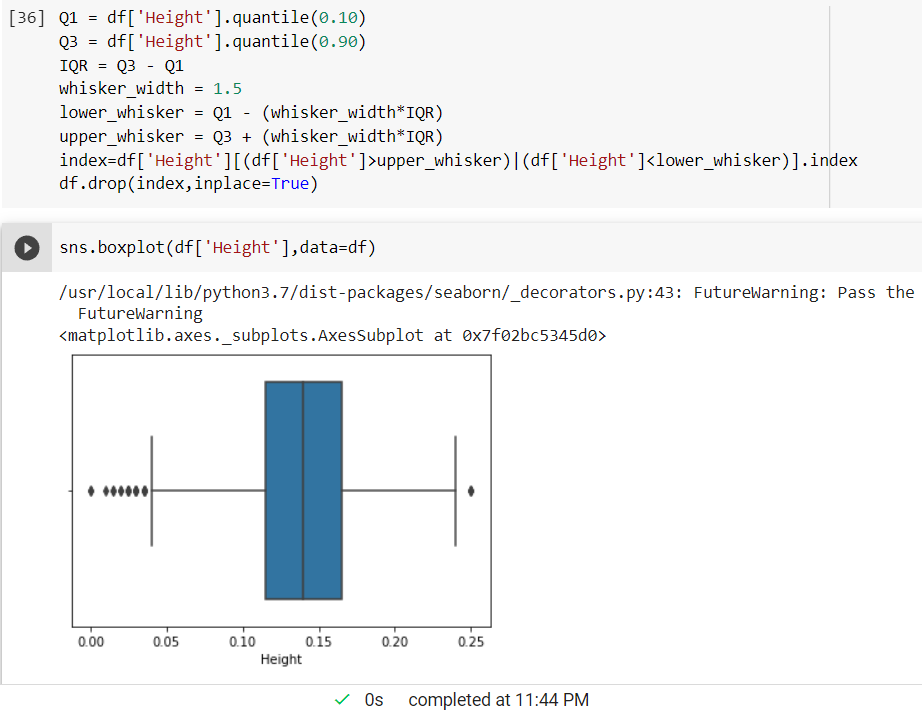
value of 3.12 shows the variable has been rightly skewed,

indicating the presence of outliers.

# Outliers Treatment

* + Trimming.

in this method, we removed and completely drop all the outliers, the line of code below creates an index for all data points and drop the index values.

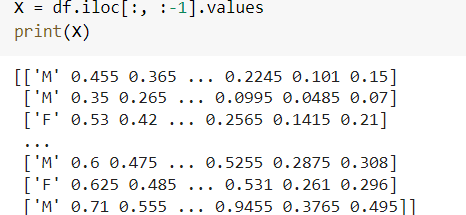


1. Check for Categorical columns and perform encoding.

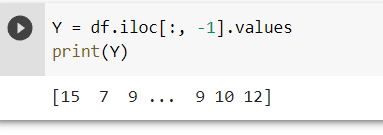
There are no categorial data:

1. Split the data into dependent and independent variables

**Splitting the Dataset into the Independent Feature Matrix:**



# Extracting the Dataset to Get the Dependent Vector



. 9. Scale the independent variables

It is a pre-processing step. This technique used to

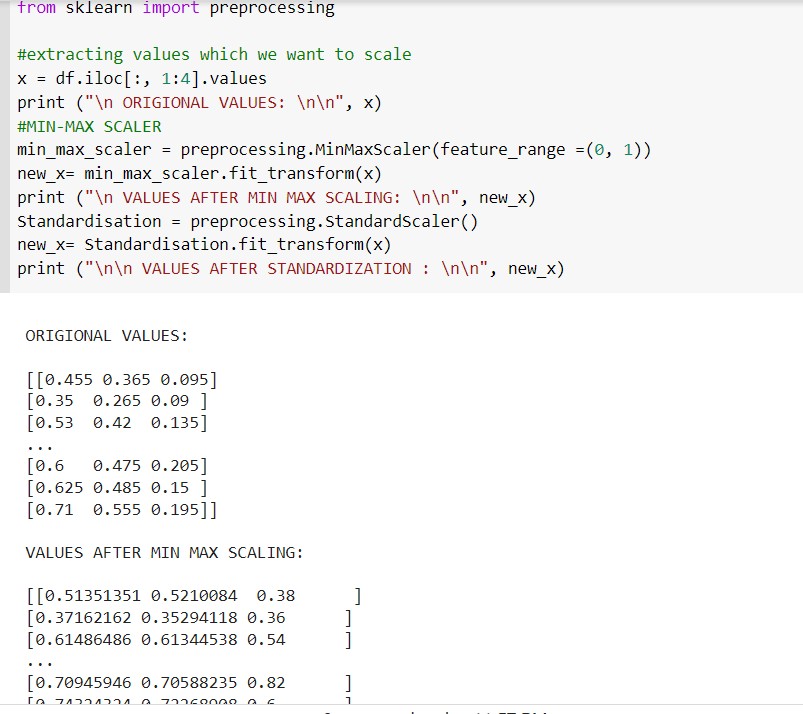
**normalize** the range of independent variables. Variables that are used to determine the target variable are known as features.

## MIN-MAX SCALING

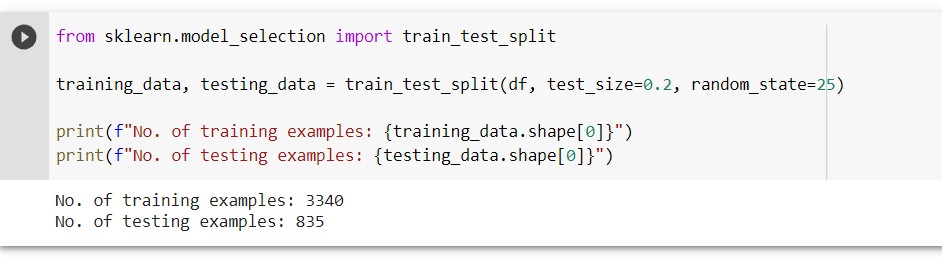
In min-max scaling or min-man normalization, we re-scale the data to a range of [0,1] or [-1,1].

## STANDARDIZATION

In this, we scale the features in such a way that the distribution has mean=0 and variance=1.

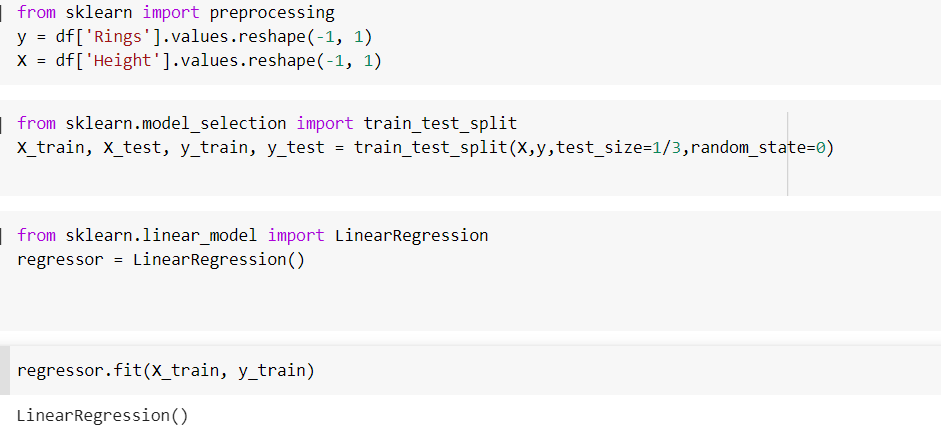


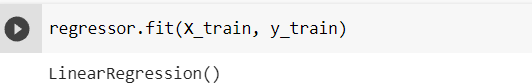
1. Split the data into training and testing

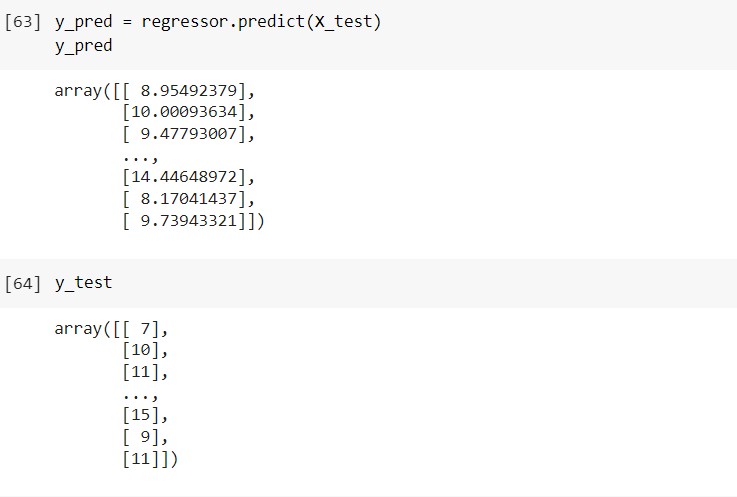


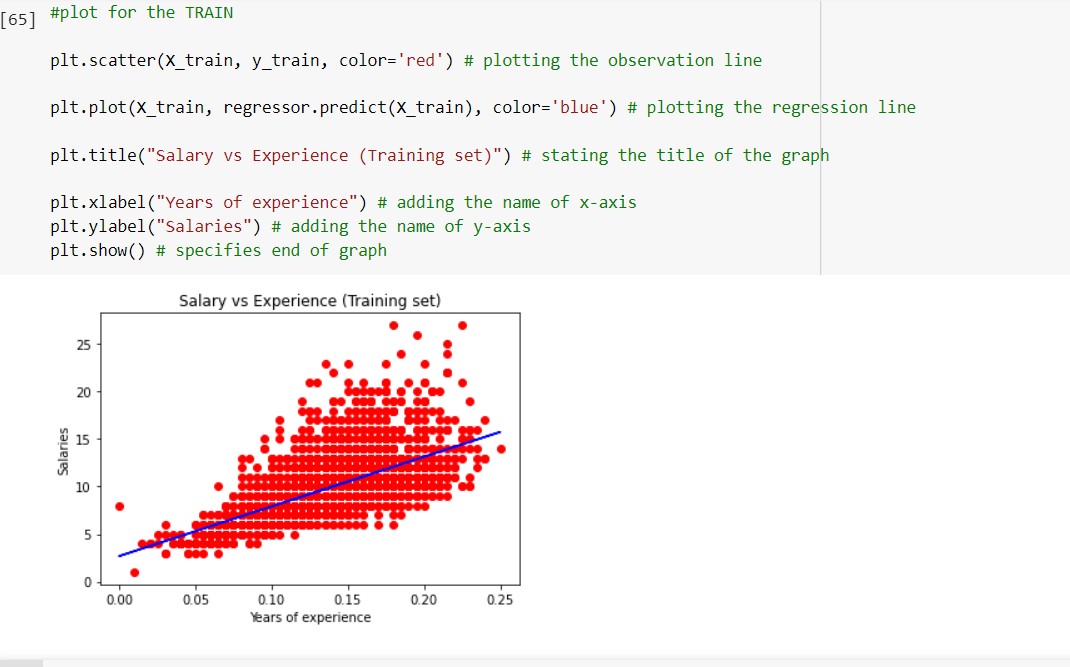
1. Build the Model

USING LINEAR REGRESSION :



1. Train the Model

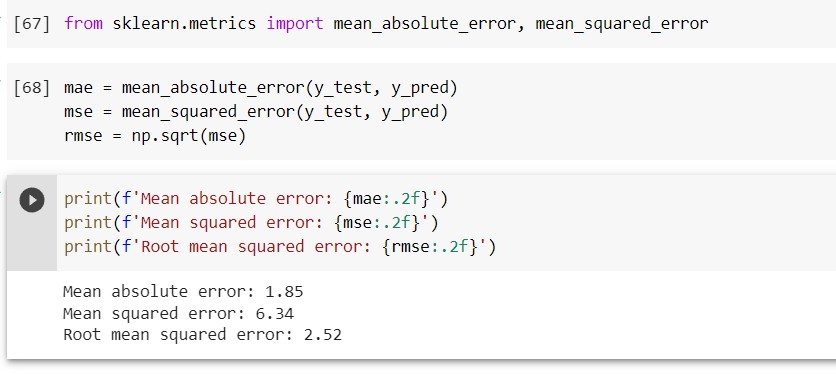




1. Test the Model



1. Measure the performance using Metrics



.