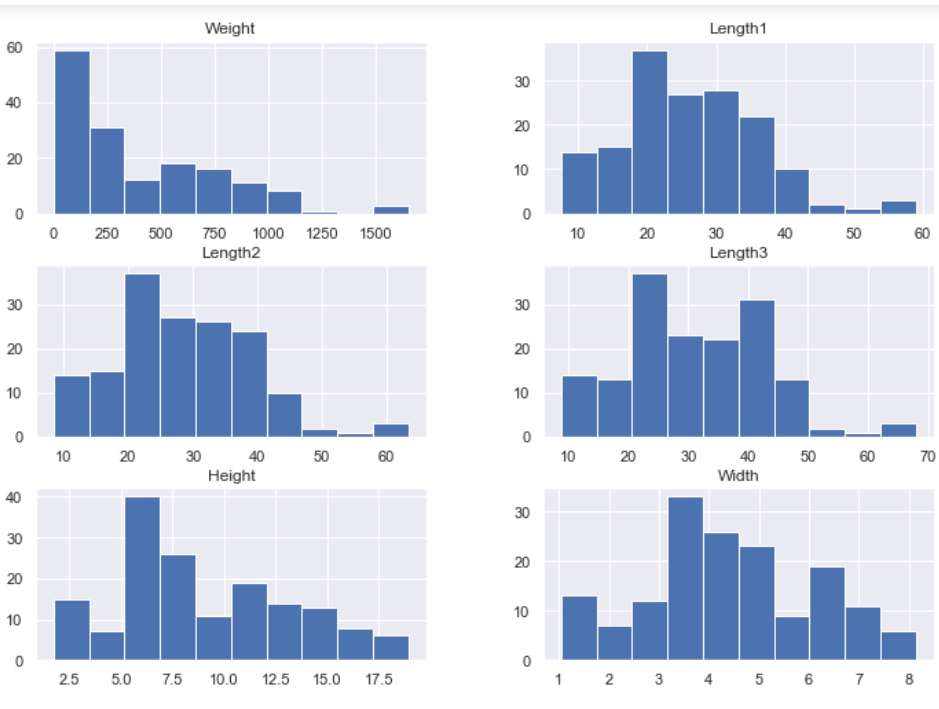
MACHINE LEARNING 1 LAB

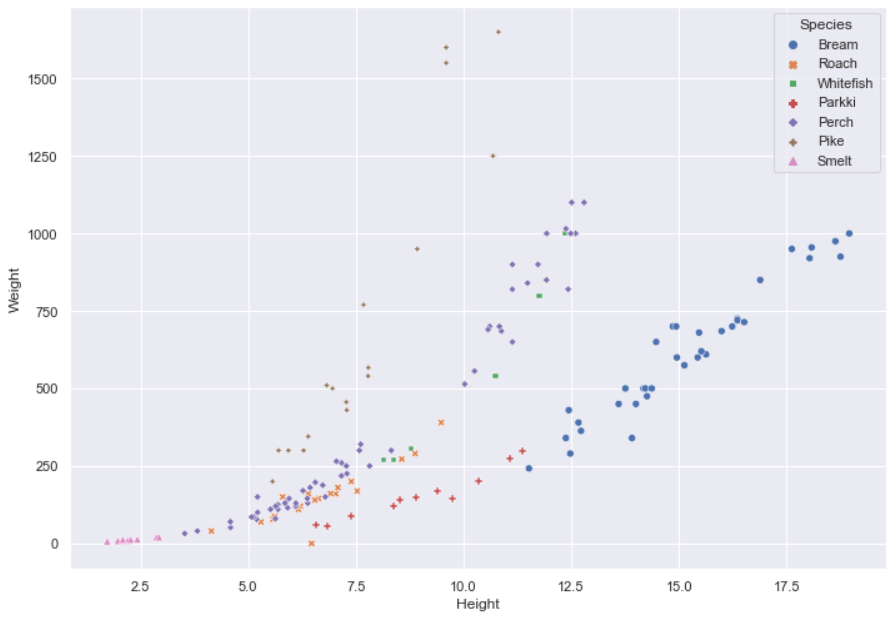
ASSIGNMENT 2

PART B

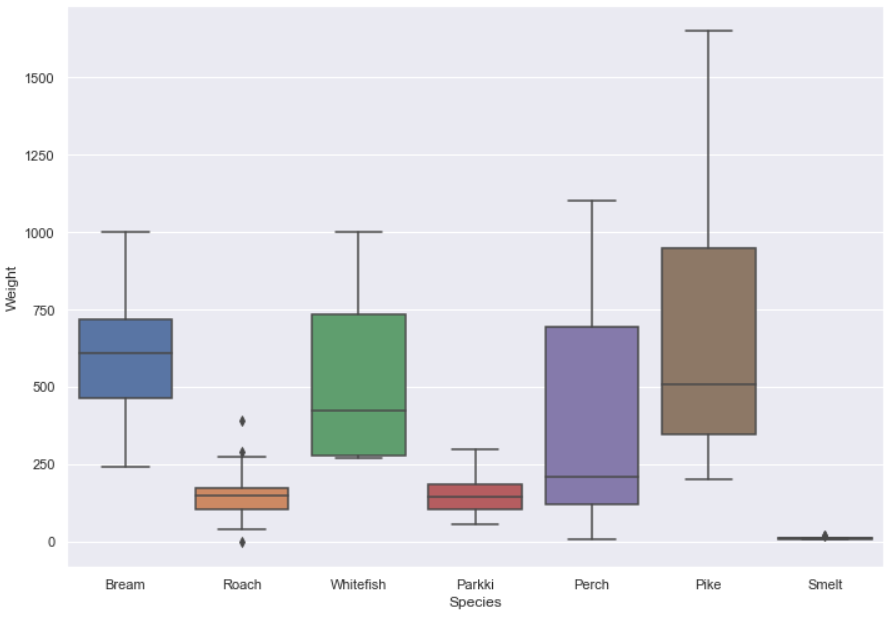
1. Document 5-6 key insights from EDA and support each point with a visualization.



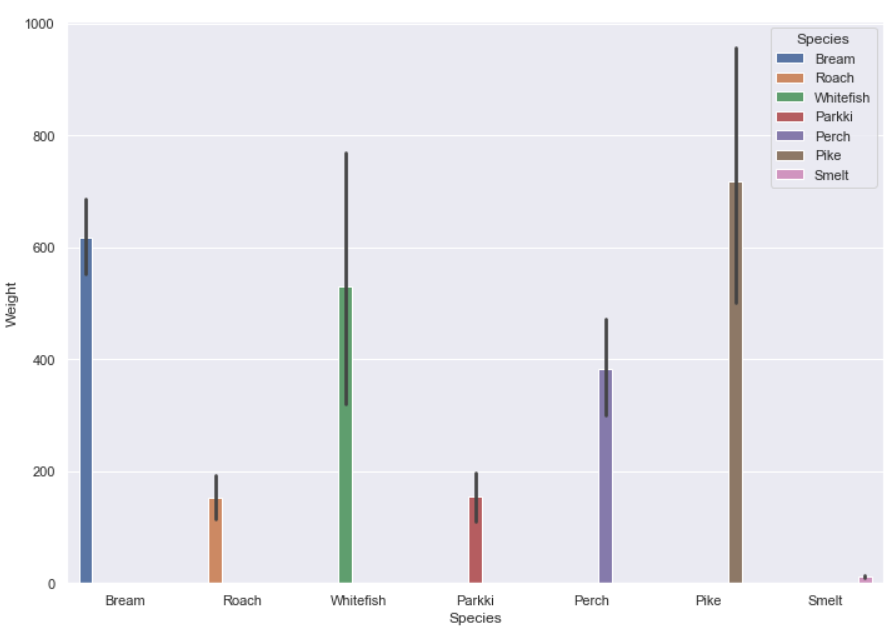
Variables are normally distributed



Height and weight are linearly depending



Brown and blue have more median weight than rest



Pike fishes are having more weights than others.

2.Answer the following questions:

i.What are the assumptions of linear regression?

**1. Linear relationship:** There exists a linear relationship between the independent variable, x, and the dependent variable, y.

**2. Independence:**The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

**3. Homoscedasticity:**The residuals have constant variance at every level of x.

**4. Normality:**The residuals of the model are normally distributed.

ii.How can we evaluate a Regression model? Define each metric and its interpretation.

* + MAE – Mean Absolute Error
    - The magnitude of difference between the prediction of an observation and the true value of that observation
  + RMSE - Root of the Mean of the Square of Errors
    - It shows how far predictions fall from measured true values using Euclidean distance
  + MSE – Mean Square Error
    - Average of the squared error that is used as the loss function for least squares regression
  + R2 – Coefficient of Determination
    - The proportion of the variation in the dependent variable that is predictable from the independent variable
  + Adjusted R2
    - Modified version of R-squared that has been adjusted for the number of predictors in the model

iii.Can R squared be negative?

It is possible to get a negative R-square for equations that do not contain a constant term. Because R-square is defined as the proportion of variance explained by the fit, if the fit is actually worse than just fitting a horizontal line then R-square is negative.

iv.What is dummy variable trap?

The Dummy Variable Trap occurs when two or more dummy variables created by one-hot encoding are highly correlated (multi-collinear). This means that one variable can be predicted from the others, making it difficult to interpret predicted coefficient variables in regression models. In other words, the individual effect of the dummy variables on the prediction model can not be interpreted well because of multicollinearity.

v.Is One Hot Encoding different from Dummy Variables?

One hot encoding can make values more than 2 while labeling like [1,0,0], [0,1,0], while dummy variables are only 0 and 1.

vi.How is polynomial regression different from linear regression?

Linear regression is linear in the parameters, not the covariates. You can make any transformations you want of them and still have a linear model. As such, polynomial regression is just a special case of linear regression.