**DATA SCIENCE**

**INTERSHIP REPORT**

**Group Members:**

Nithyalakshmi J

Monisha P

**Objective(s):**

* The target is to prepare an ML model which can predict the profit value of a company if the value of its R&D Spend, Administration Cost and Marketing Spend are given.

i) Construct Different Regression algorithms

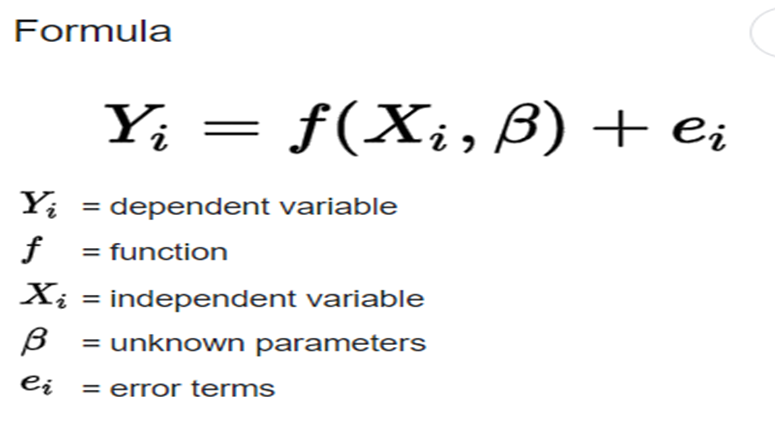
ii) Divide the data into train set and test set

iii) Calculate different regression metrics

iv) Choose the best model

* The model1 is successfully created with great R^2 value of 98.9%.
* The model2 is successfully created with great R^2 value of 98.8%.
* On comparing model 1 and model 2, model 1 is the best model

**Regression:**

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other

**Types of Regression:**

Every regression technique has some assumptions attached to it which we need to meet before running analysis. These techniques differ in terms of type of dependent and independent variables and distribution.

**1. Linear Regression**

It is the simplest form of regression. It is a technique in which the **dependent variable is continuous** in nature. The relationship between the dependent variable and independent variables is assumed to be linear in nature.

**2. Polynomial Regression**

It is a technique to fit a nonlinear equation by taking polynomial functions of independent variable.  
In the figure given below, you can see the red curve fits the data better than the green curve. Hence in the situations where the relation between the dependent and independent variable seems to be non-linear we can deploy **Polynomial Regression Models.**

**3. Logistic Regression**

In logistic regression, the dependent variable is binary in nature (having two categories). Independent variables can be continuous or binary. In multinomial logistic regression, you can have more than two categories in your dependent variable.

**4. Quantile Regression**

Quantile regression is the extension of linear regression and we generally use it when outliers, high skeweness and heteroscedasticity exist in the data.  
  
In linear regression, we predict the mean of the dependent variable for given independent variables. Since mean does not describe the whole distribution, so modeling the mean is not a full description of a relationship between dependent and independent variables. So we can use quantile regression which predicts a quantile (or percentile) for given independent variables.

**Linear Regression**

* a [scalar](https://en.wikipedia.org/wiki/Scalar_(mathematics)) response and one or more explanatory variables (also known as [dependent and independent variables](https://en.wikipedia.org/wiki/Dependent_and_independent_variables)).
* The most common form of **regression** analysis is linear **regression**
* **Linear regression** is a [linear](https://en.wikipedia.org/wiki/Linearity) approach for modelling the relationship between Linear regression models are often fitted using the [least squares](https://en.wikipedia.org/wiki/Least_squares) approach

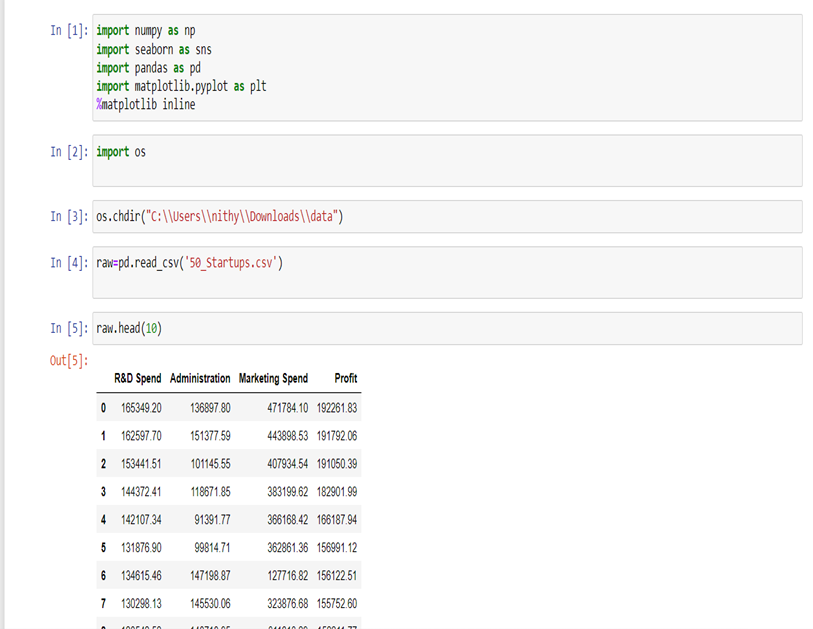
There are **four principal assumptions**which justify the use of linear regression models for purposes of inference or prediction:

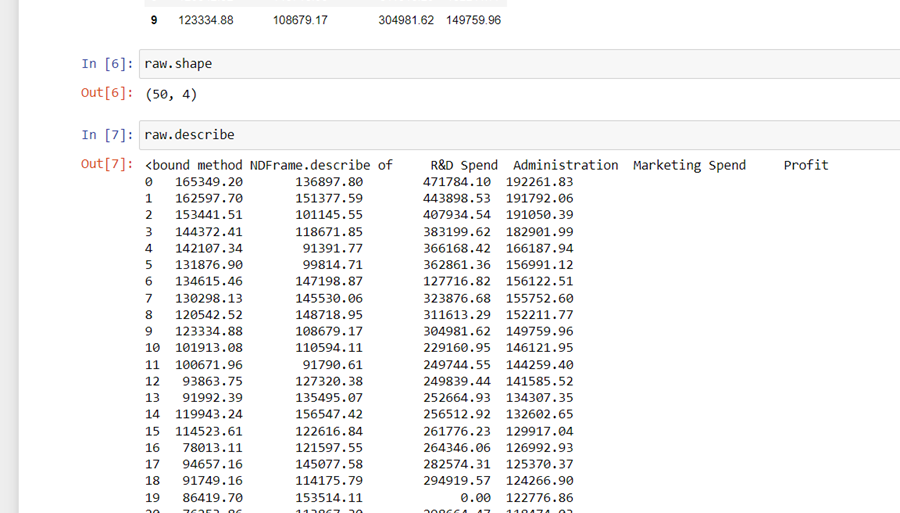
* **(i) linearity and additivity** of the relationship between dependent and independent variables:
* (a) The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed.
* (b) The slope of that line does not depend on the values of the other variables.
* © The effects of different independent variables on the expected value of the dependent variable are additive.
* **(ii) statistical independence** of the errors (in particular, no correlation between consecutive errors in the case of time series data)
* **(iii) homoscedasticity** (constant variance) of the errors
* (a) versus time (in the case of time series data)
* (b) versus the predictions
* © versus any independent variable
* **(iv) normality** of the error distribution.

1. Most applications fall into one of the following two broad categories:

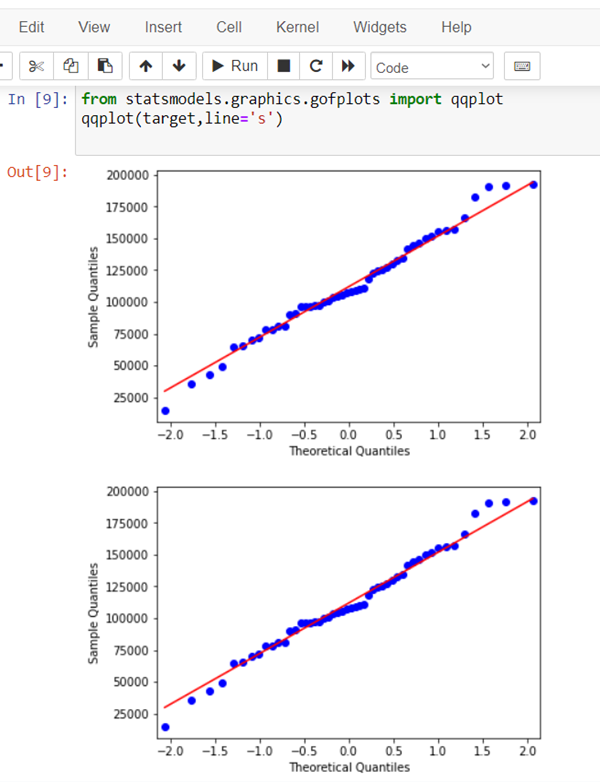
* If the goal is prediction, forecasting, or error reduction,linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
* If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

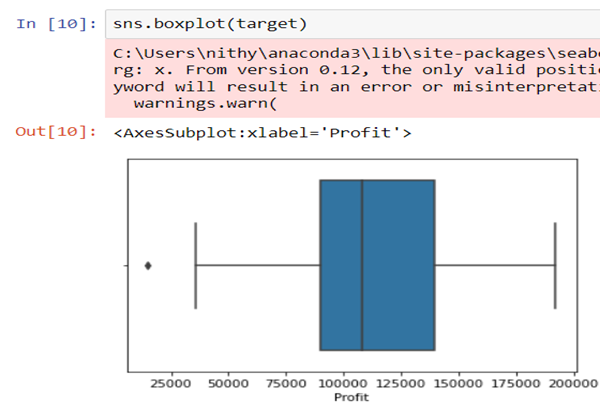
**LOADING DATASET:**

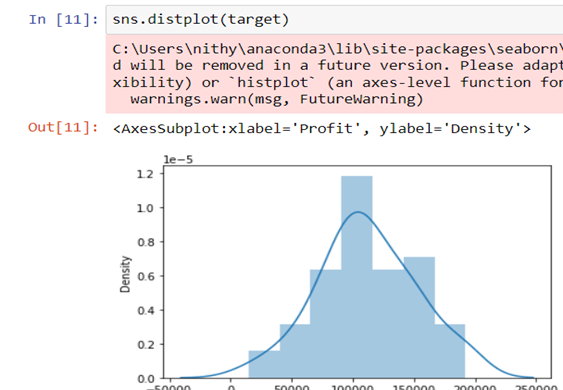




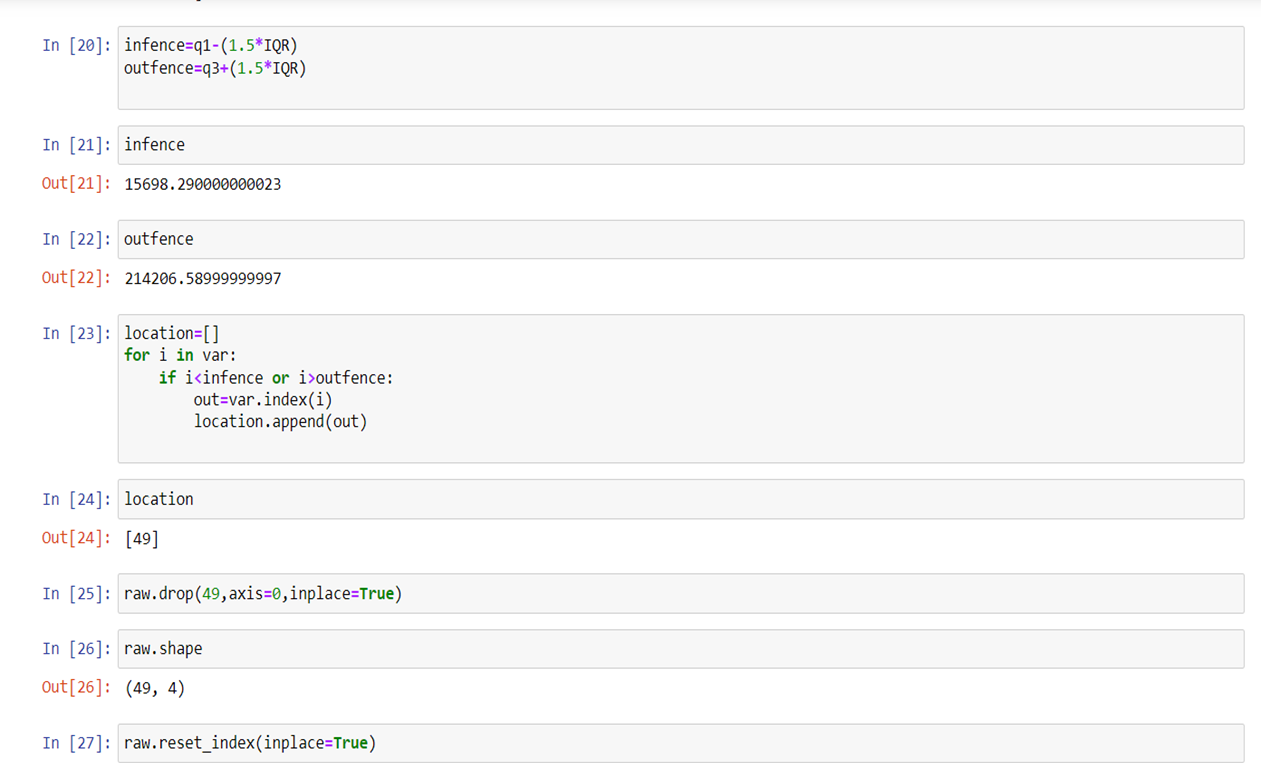
**DATA VISUALISATION:**



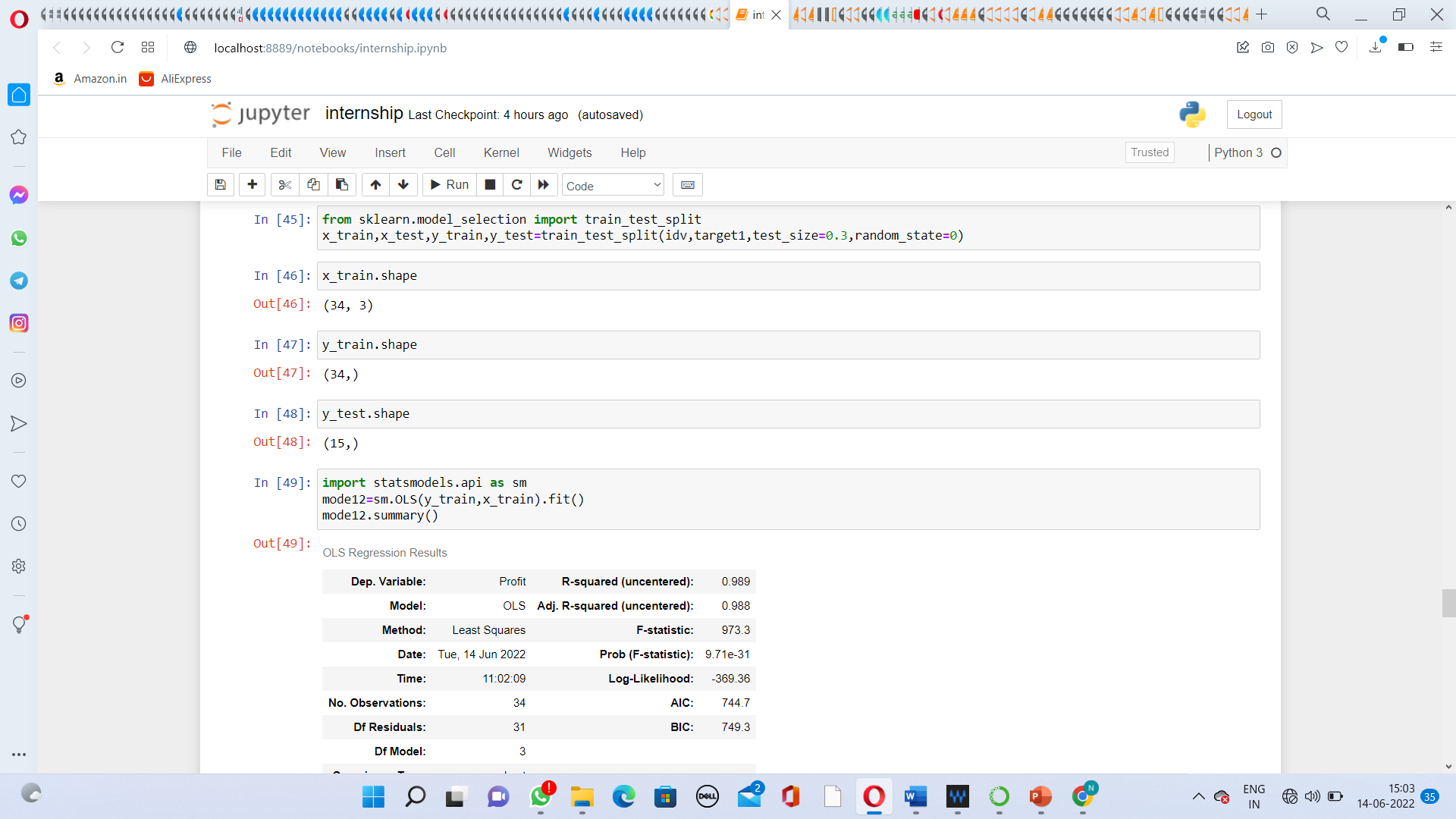




**DATA PREPROCESSING:**



**MODEL 1**



**MODEL 2**

