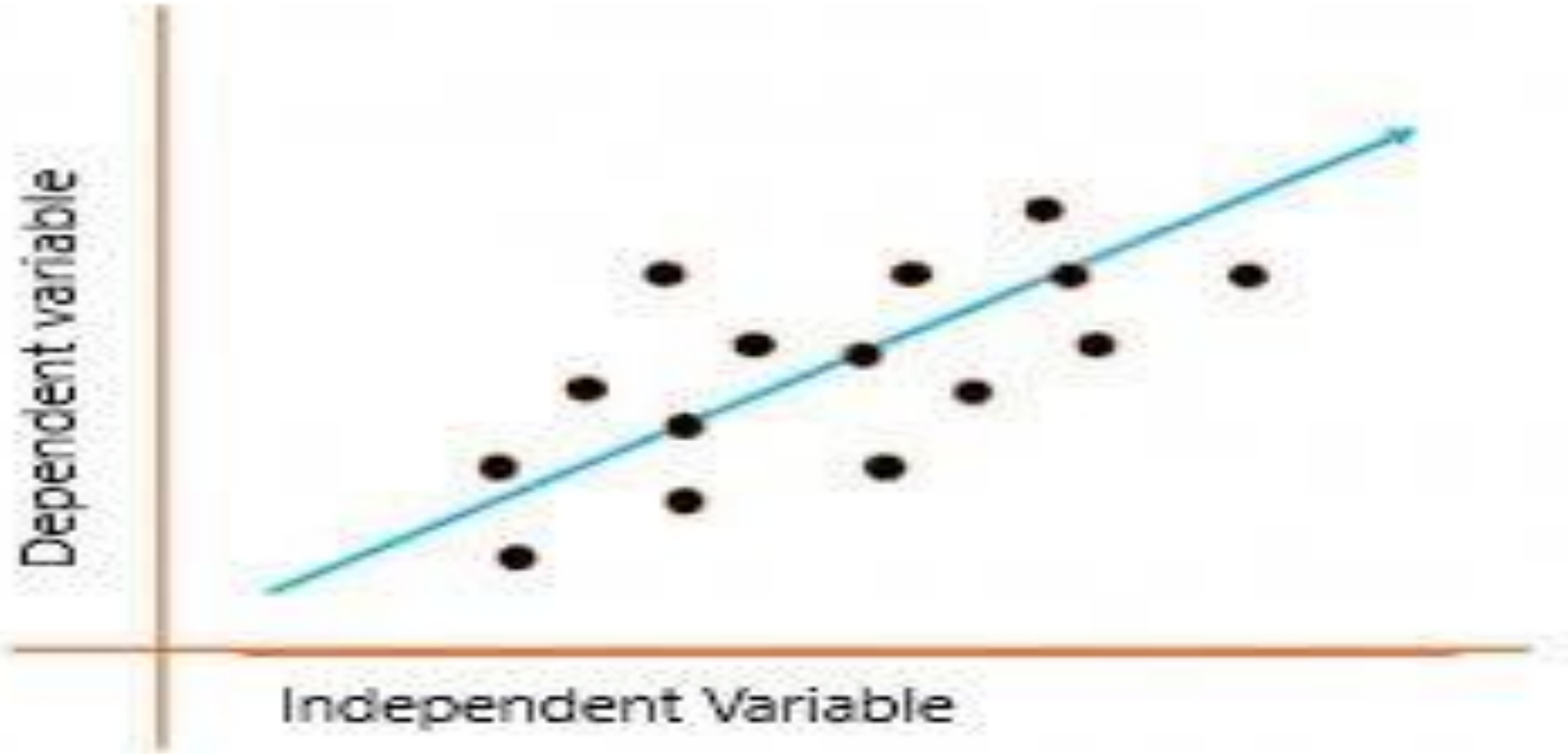
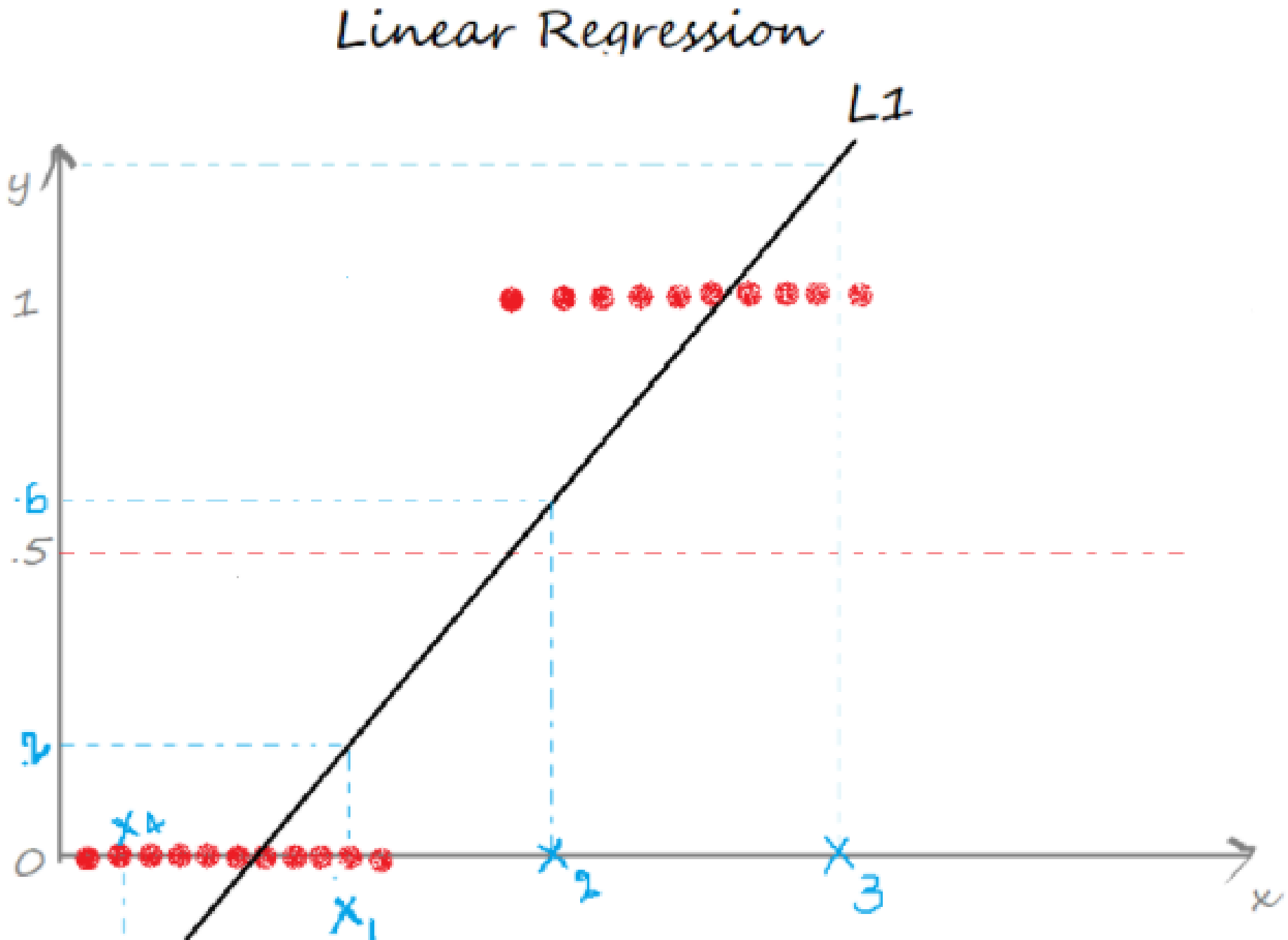
Logistic Regression


Dependent variable
Independent Variable

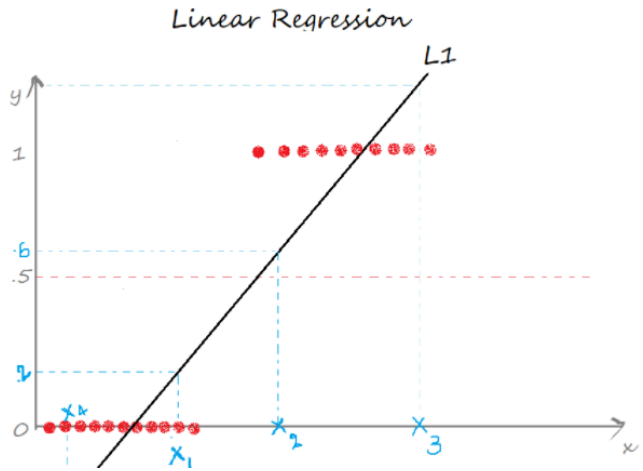

Linear Regression
L1
y
1
.6
.5
.2
0
x
x_4
x_1
x_2
x_3

- In Linear regression, we draw a straight line(the best fit line) L1 such that the sum of distances of all the data points to the line is minimal. The equation of the line L1 is y=mx+c, where m is the slope and c is the y-intercept.
- We define a threshold T = 0.5, above which the output belongs to class 1 and class 0 otherwise.
-

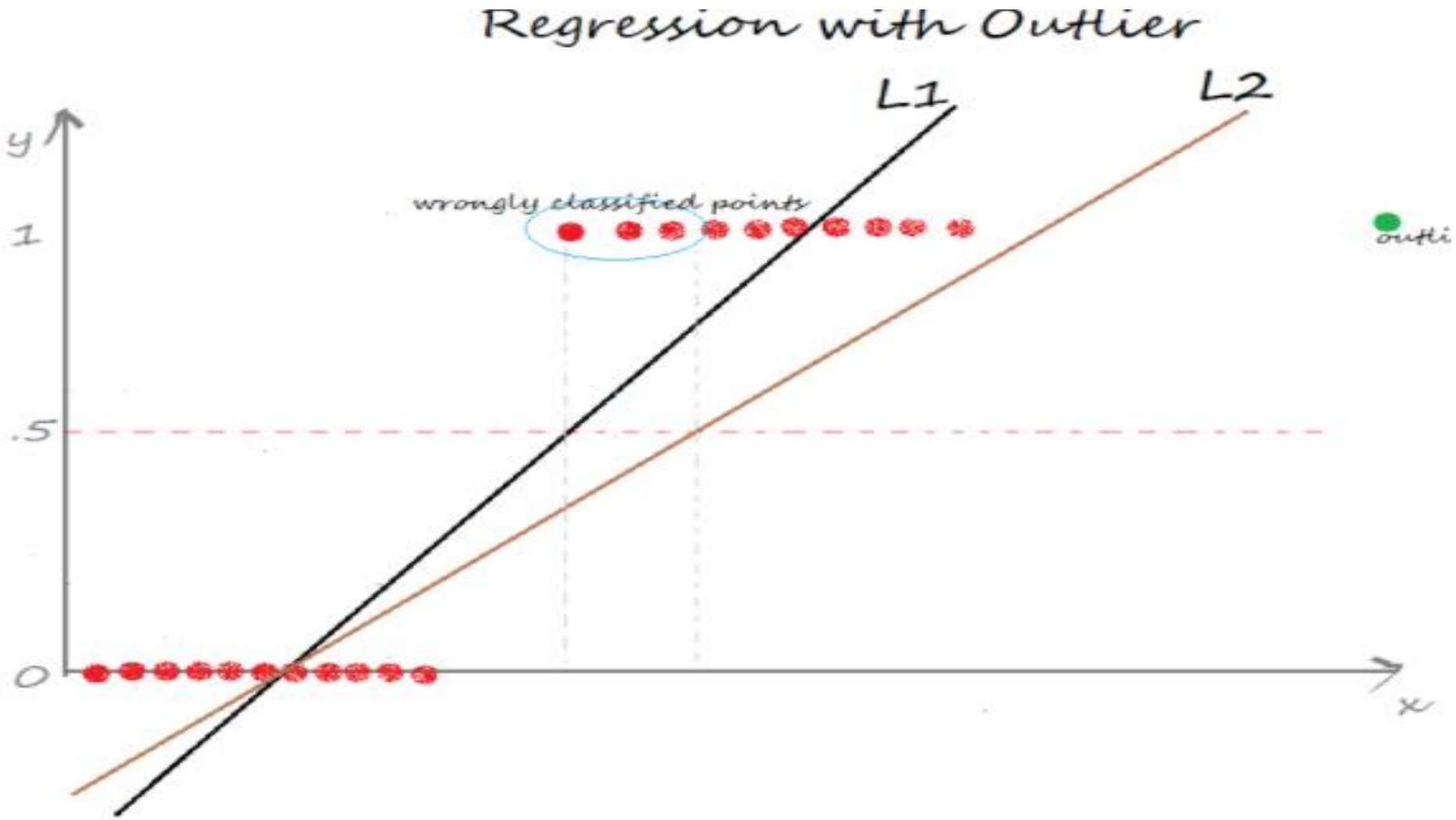
$$y = mx + c, \text{ Threshold } T = 0.5$$

$$y = \begin{cases} 1, & mx + c \geq 0.5 \\ 0, & mx + c < 0.5 \end{cases}$$

- **Case 1:** the predicted value for x1 is ≈0.2 which is less than the threshold, so x1 belongs to class 0.
- **Case 2:** the predicted value for the point x2 is ≈0.6 which is greater than the threshold, so x2 belongs to class 1.
- So far so good, yeah!
- **Case 3:** the predicted value for the point x3 is beyond 1.
- **Case 4:** the predicted value for the point x4 is below 0.
- The predicted values for the points x3, x4 exceed the range (0,1) which doesn't make sense because the probability values always lie between 0 and 1. And our output can have only two values either 0 or 1. Hence, this is a problem with the linear regression model



- Now, introduce an outlier and see what happens. The regression line gets deviated to keep the distance of all the data points to the line to be minimal.
-



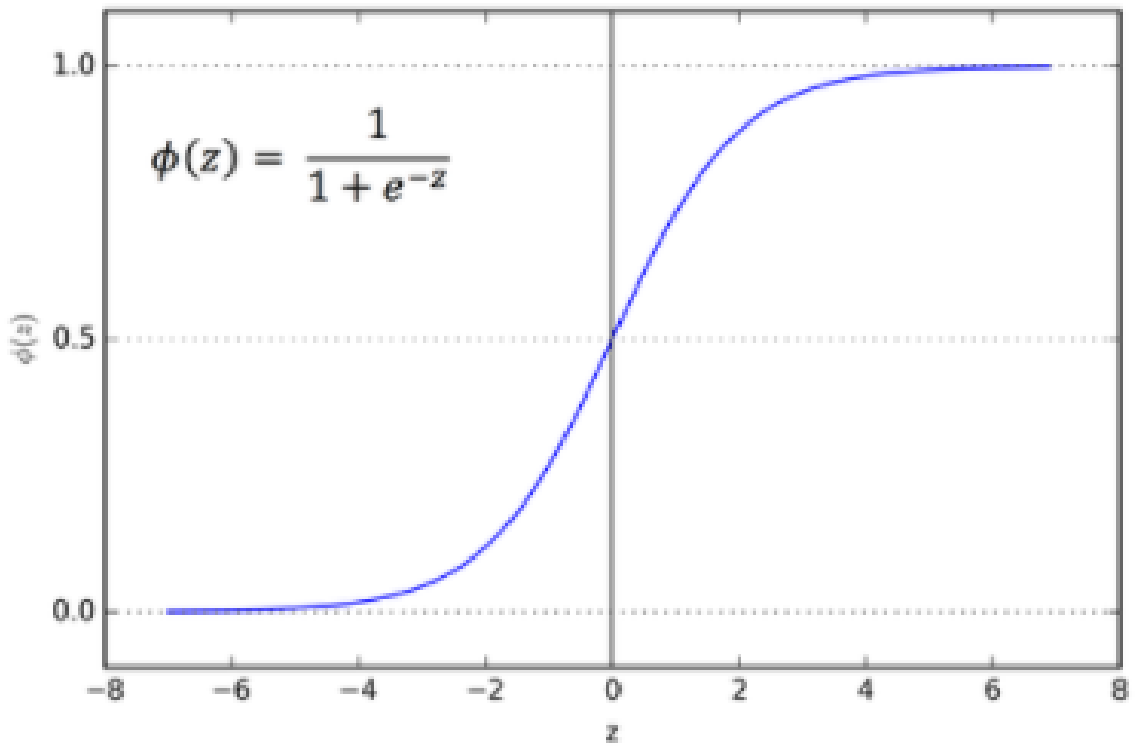
- L2 is the new best-fit line after the addition of an outlier. Seems good till now. But the problem is, if we closely observe, some of the data points are wrongly classified. Certainly, it increases the error term ☹ This again is a problem with the linear regression model.

- The two limitations of using a linear regression model for classification problems are:

- the predicted value may exceed the range (0,1)
- error rate increases if the data has outliers
- There definitely is a need for Logistic regression here

Sigmoid Function:

The sigmoid function is useful to map any predicted values of probabilities into another value between 0 and 1.



Type of Logistic Regression:

On the basis of the categories, Logistic Regression can be classified into three types:

- **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
- **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".