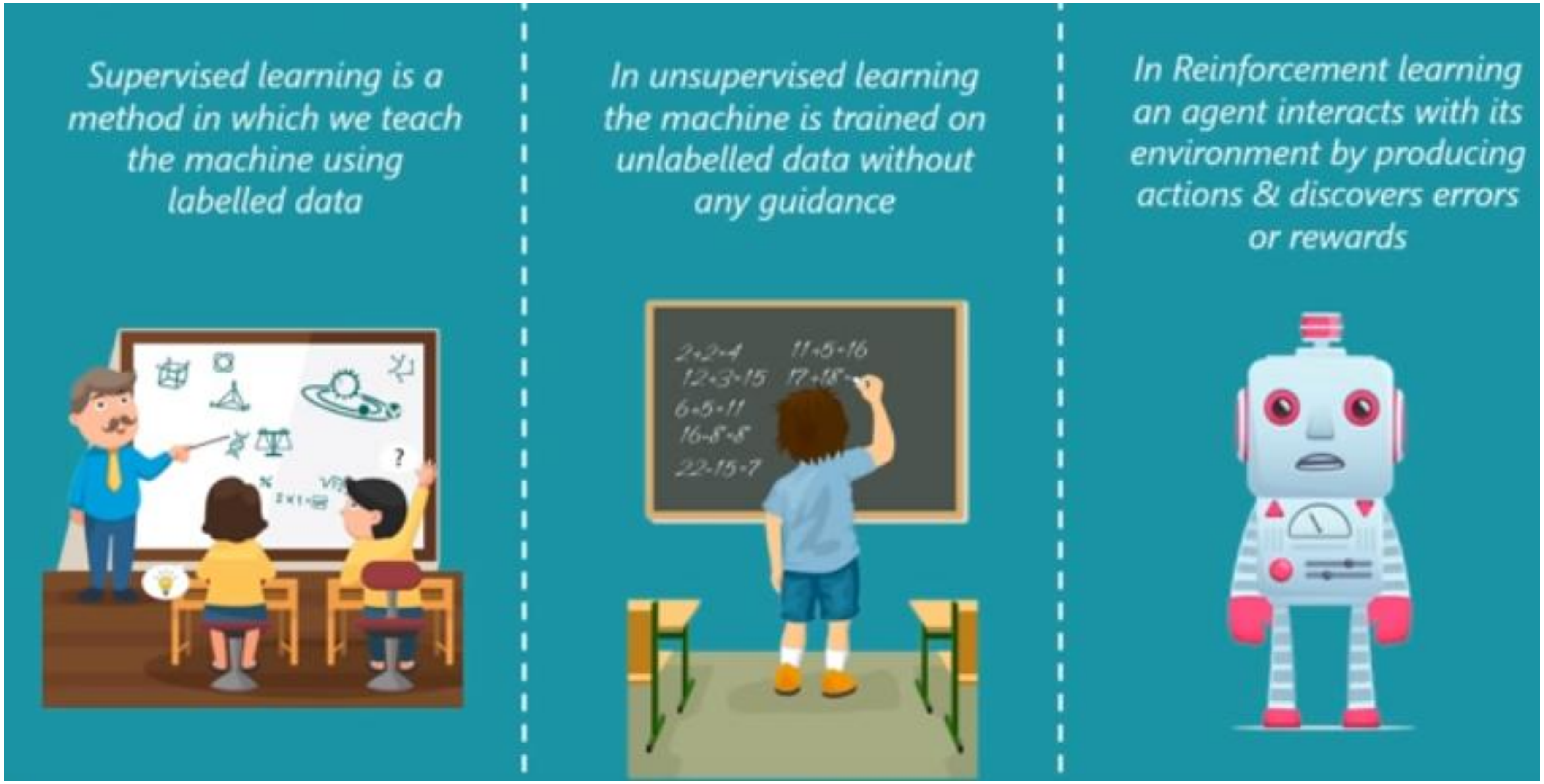
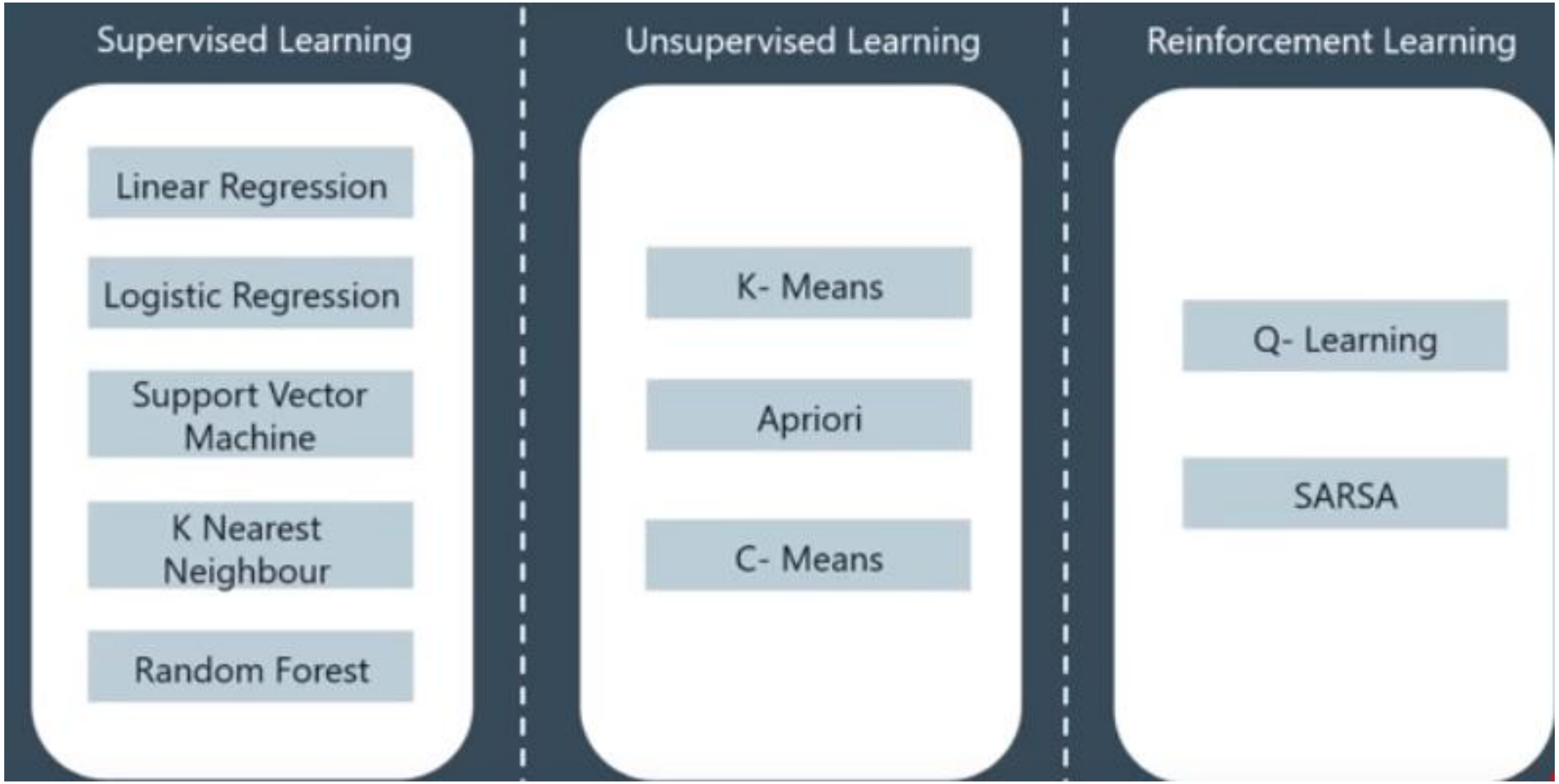
Regression

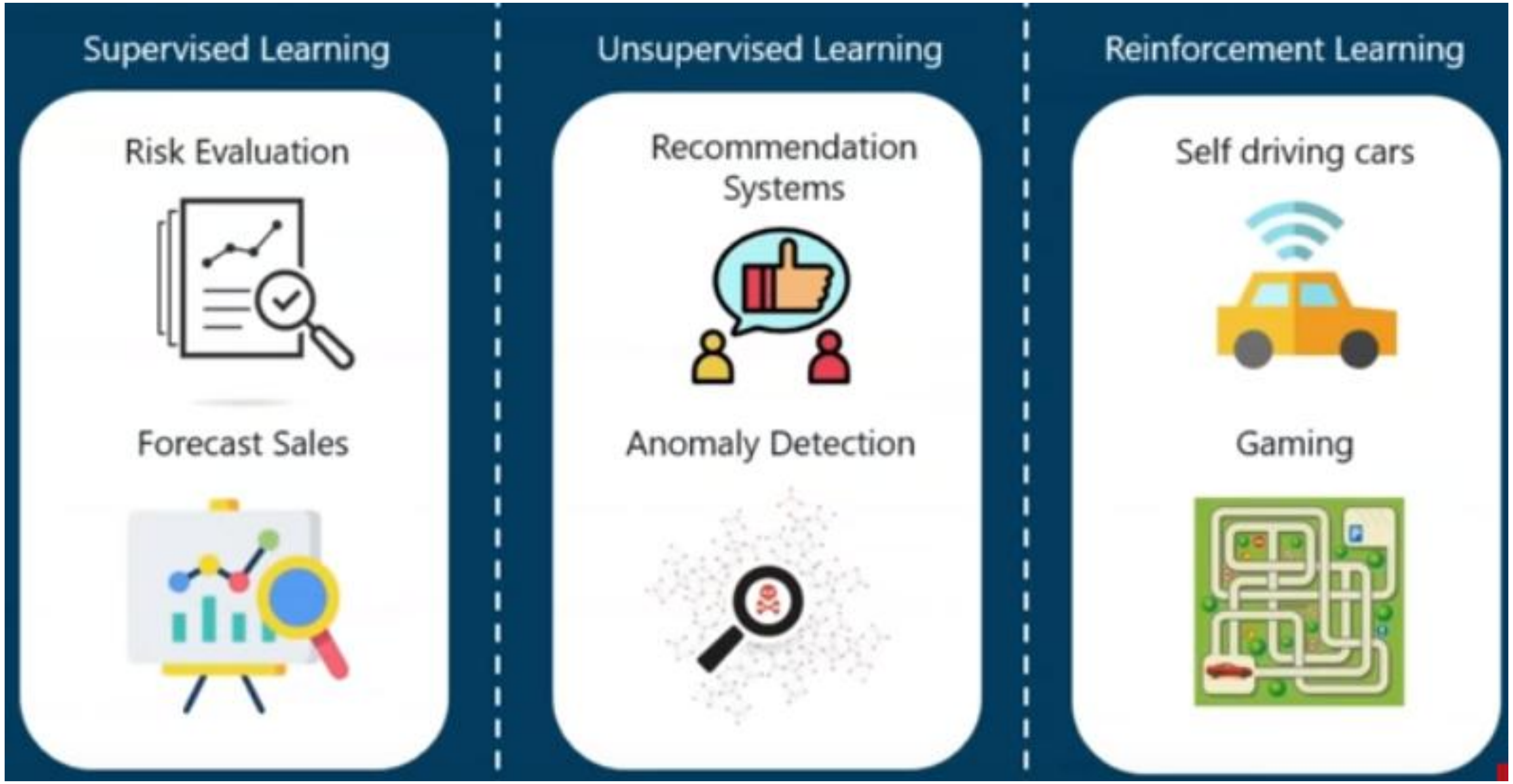

Supervised learning is a method in which we teach the machine using labelled data
In unsupervised learning the machine is trained on unlabelled data without any guidance
In Reinforcement learning an agent interacts with its environment by producing actions & discovers errors or rewards

Algorithim



Supervised Learning
Linear Regression
Logistic Regression
Support Vector Machine
K Nearest Neighbour
Random Forest
Unsupervised Learning
K- Means
Apriori
C- Means
Reinforcement Learning
Q- Learning
SARSA

Applications



What is Regression?

"Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable"

It is a technique used to:

- Estimate a relationship between variables
- Predict the value of one variable (dependent variable) on the basis of other variables (independent variables)

Example:
$Y = \beta_0 + \beta_1 x + \varepsilon$
Here, Y is a dependent variable, whereas β_0, β_1, x, and ε are independent variables.

Types Regression analysis

These are:

Regression Analysis

Simple Regression

Considers one quantitative and independent variable X to predict the other quantitative, but dependent, variable Y

Multiple Regression

Considers more than one quantitative and qualitative variable ($X_1 \ldots X_N$) to predict a quantitative and dependent variable Y

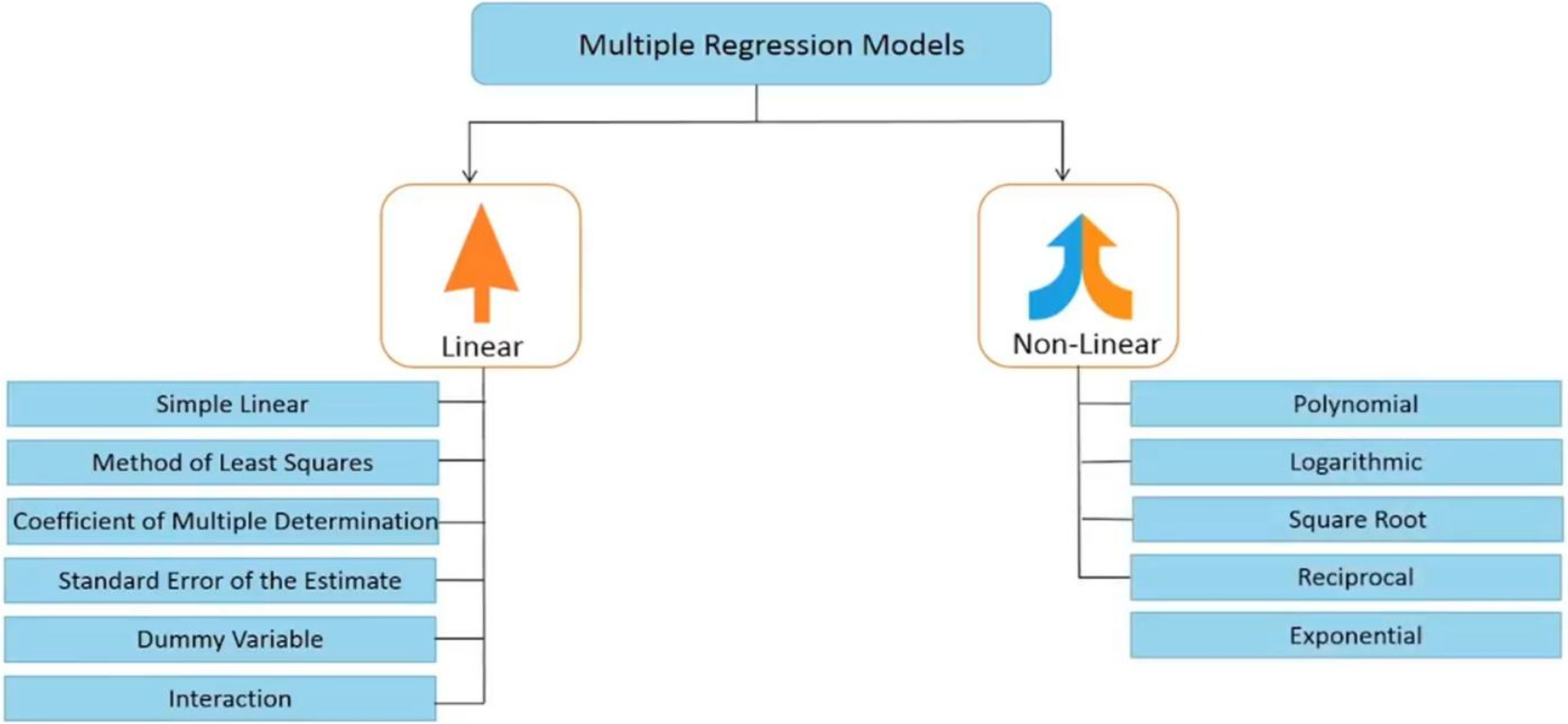
Simple Regression analysis

In this analysis:

- A relationship between a dependent variable and an independent variable is modeled
- The output is a function to predict the dependent variable on the basis of the values of independent variables
- A straight line is fit to the data

Multiple Regression Models

These are:



Uses of regression analysis

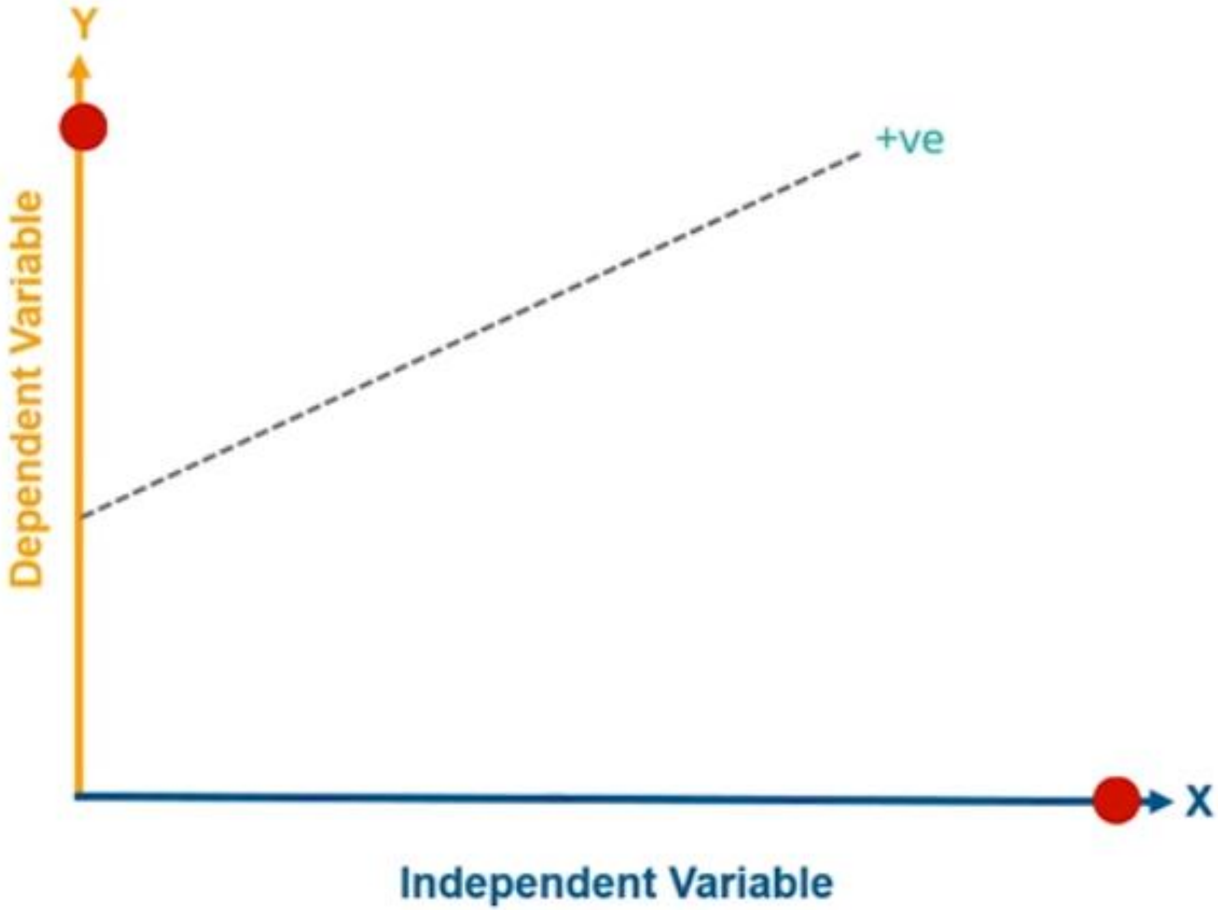
Three major uses for regression analysis are

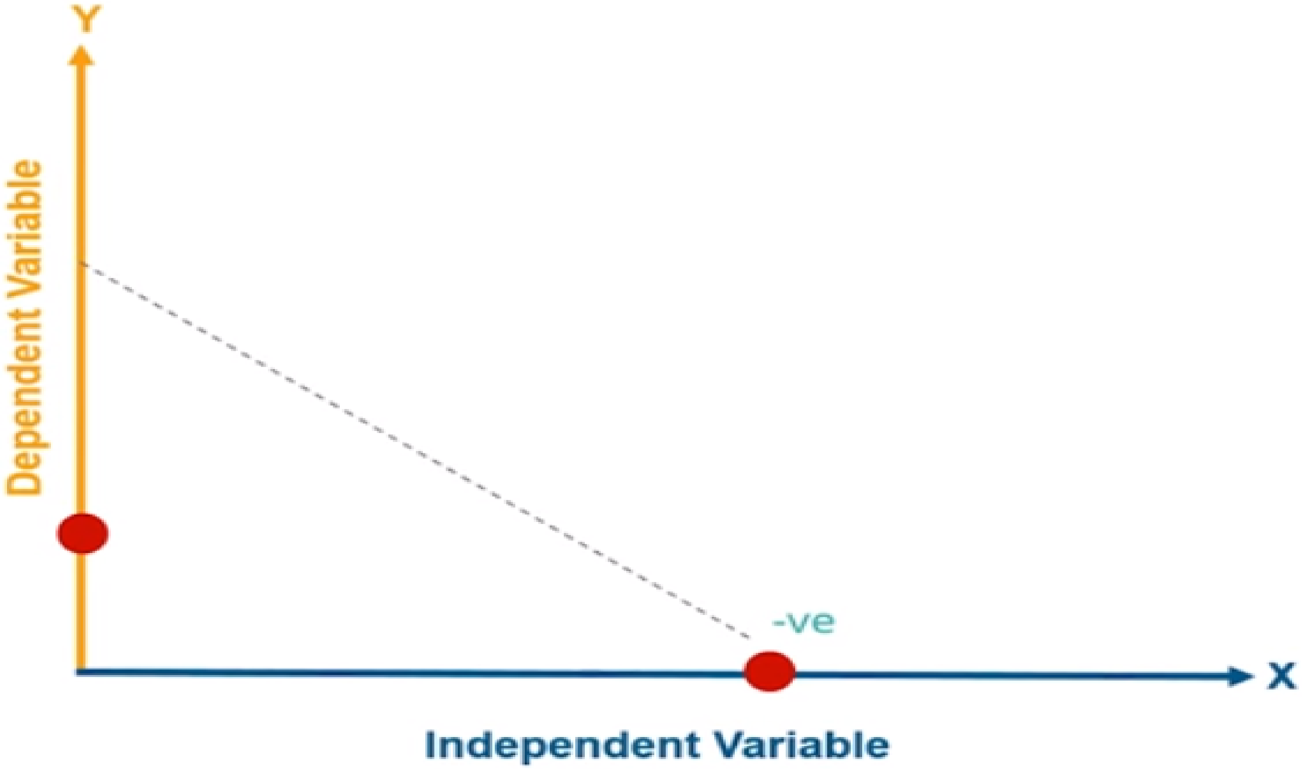
- Determining the strength of predictors
- Forecasting an effect, and
- Trend forecasting

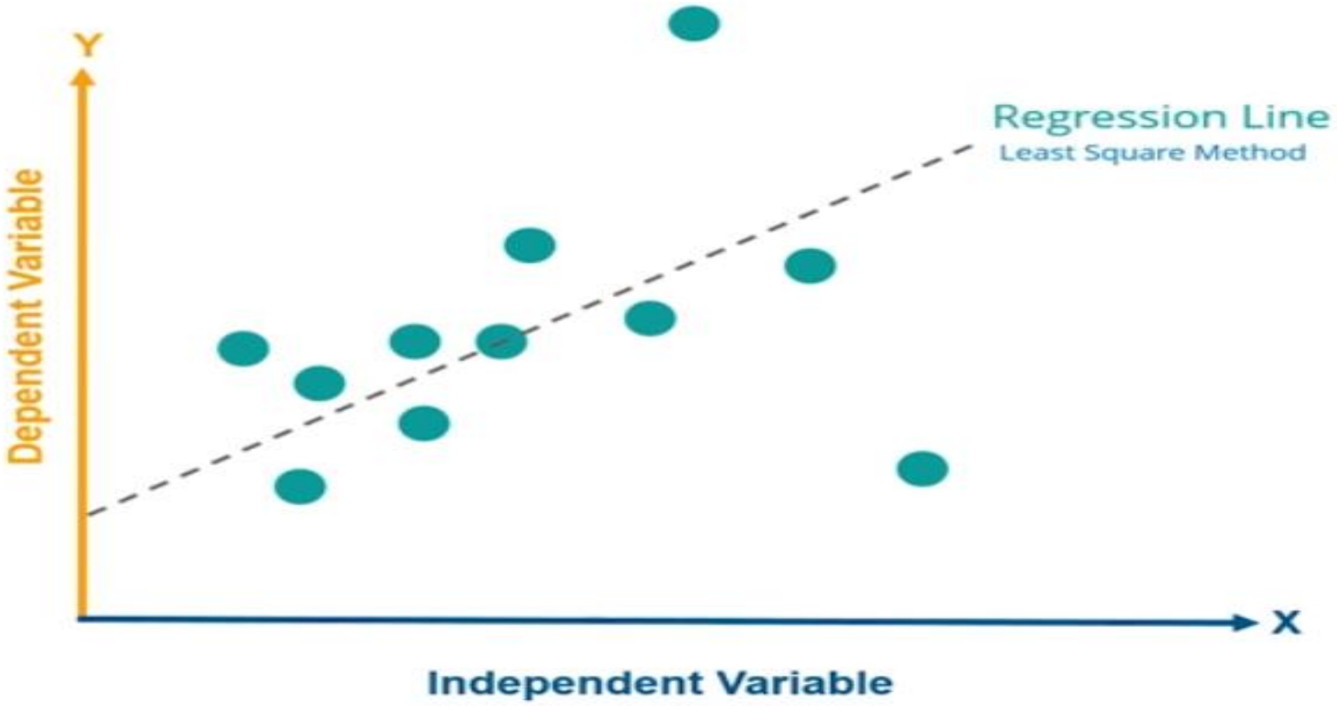
Where is linear regression used

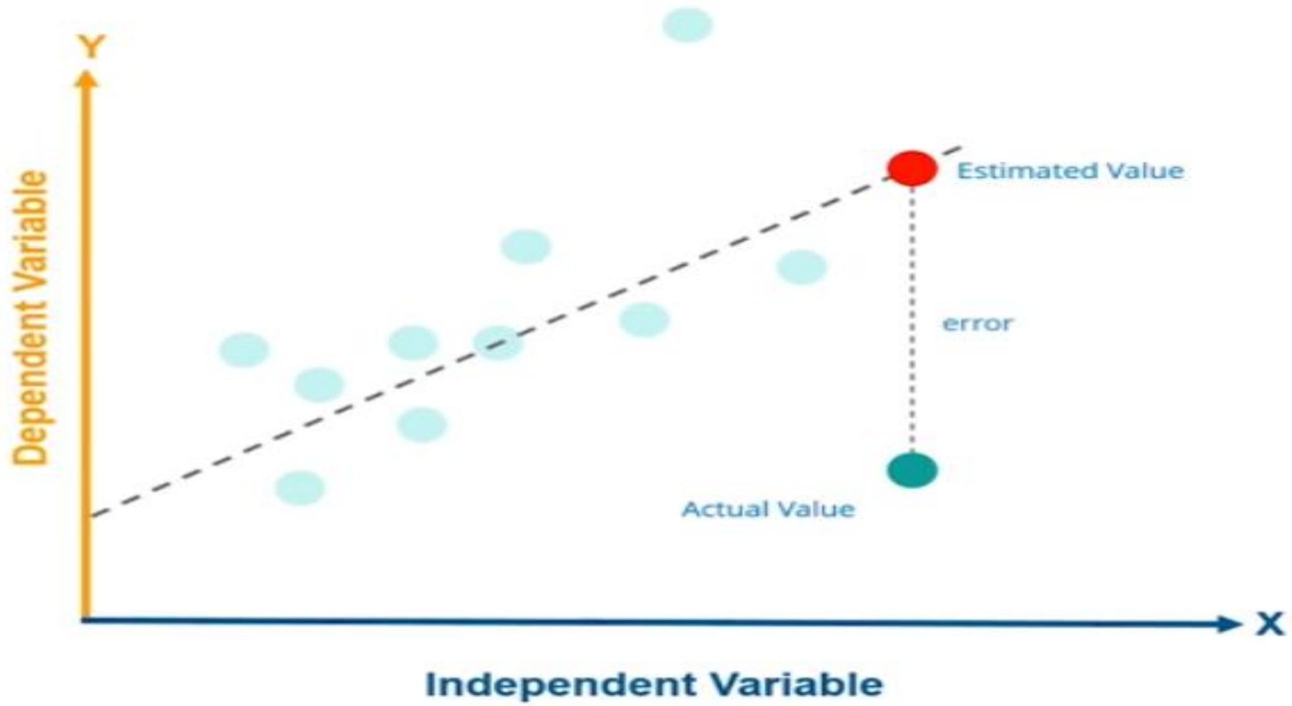
- Evaluating Trends and Sales Estimates
- Analyzing the Impact of Price Changes
- Assessment of risk in financial services and insurance domain

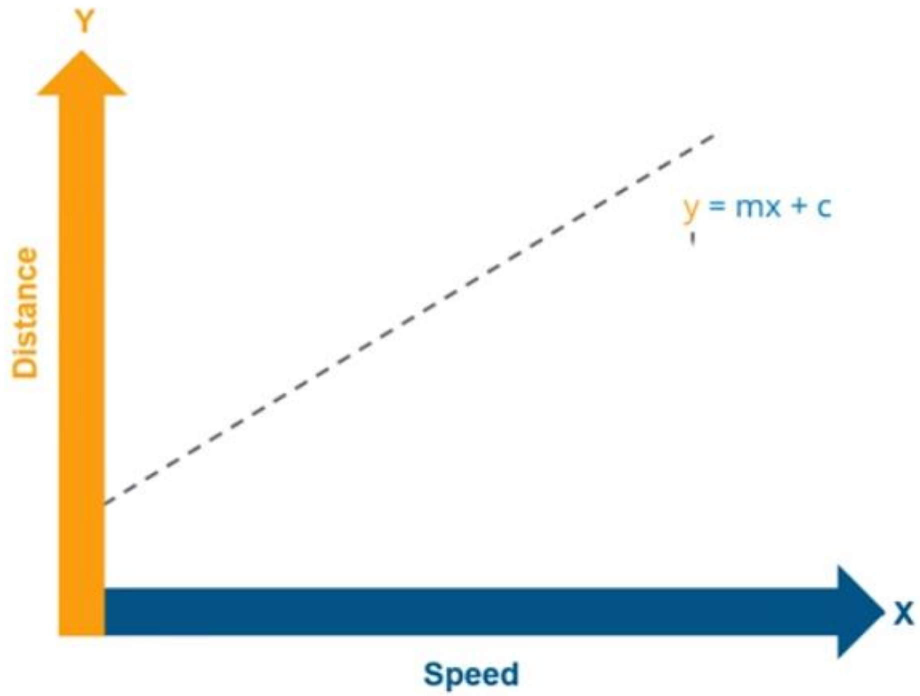
Linear regression algorithim

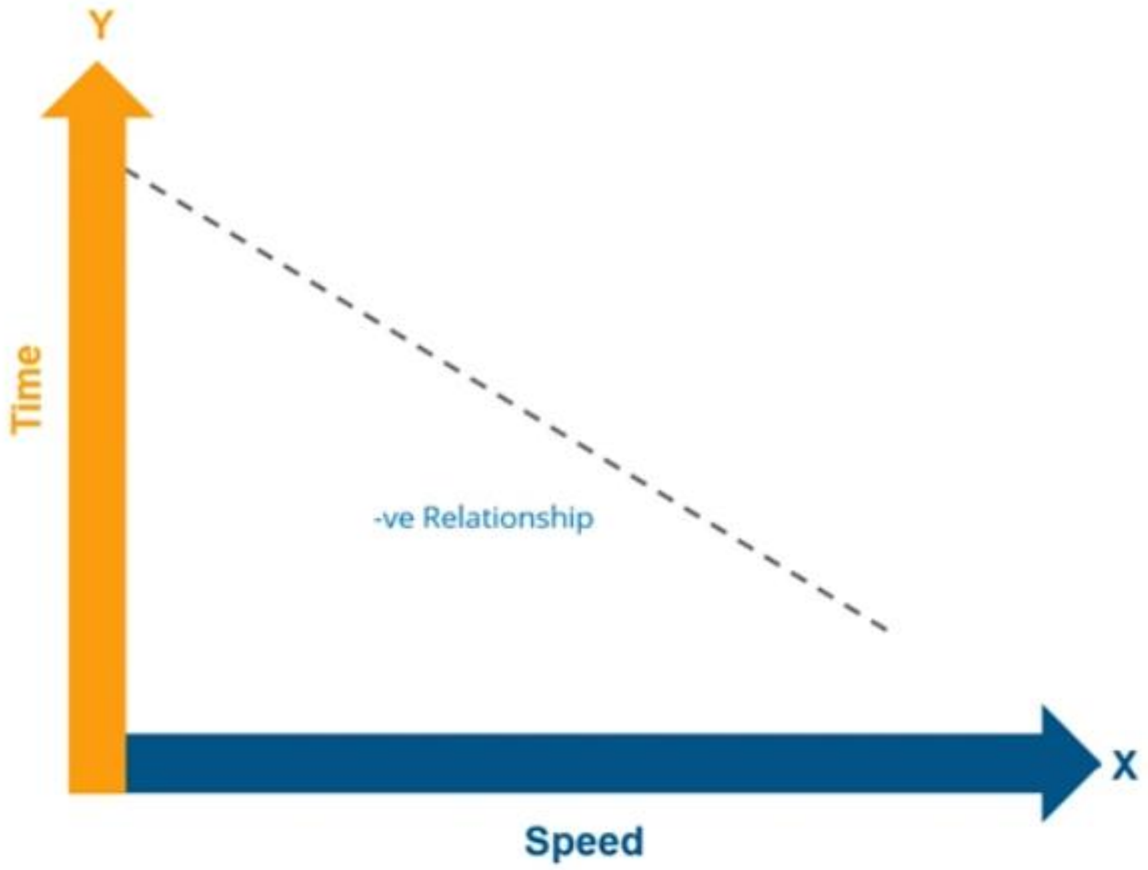


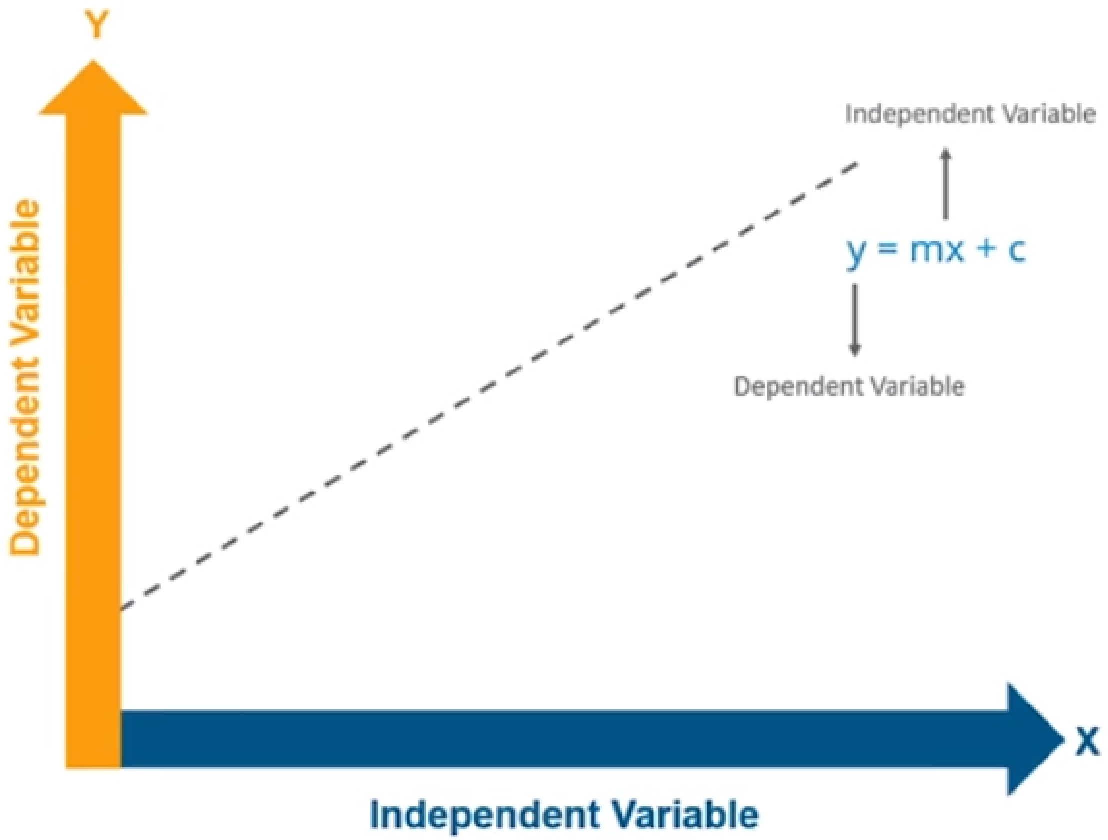

Y
Dependent Variable
-ve
X
Independent Variable

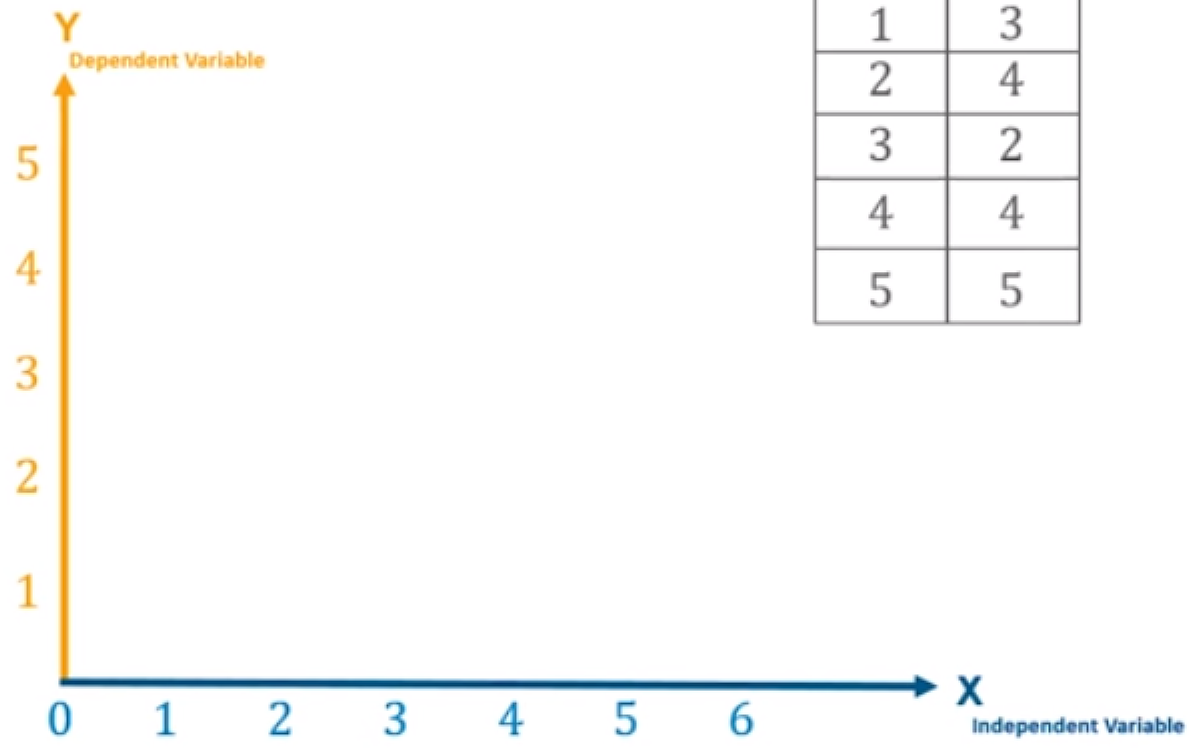

Y
Regression Line
Least Square Method
Dependent Variable
X
Independent Variable


Y
Dependent Variable
Estimated Value
error
Actual Value
X
Independent Variable

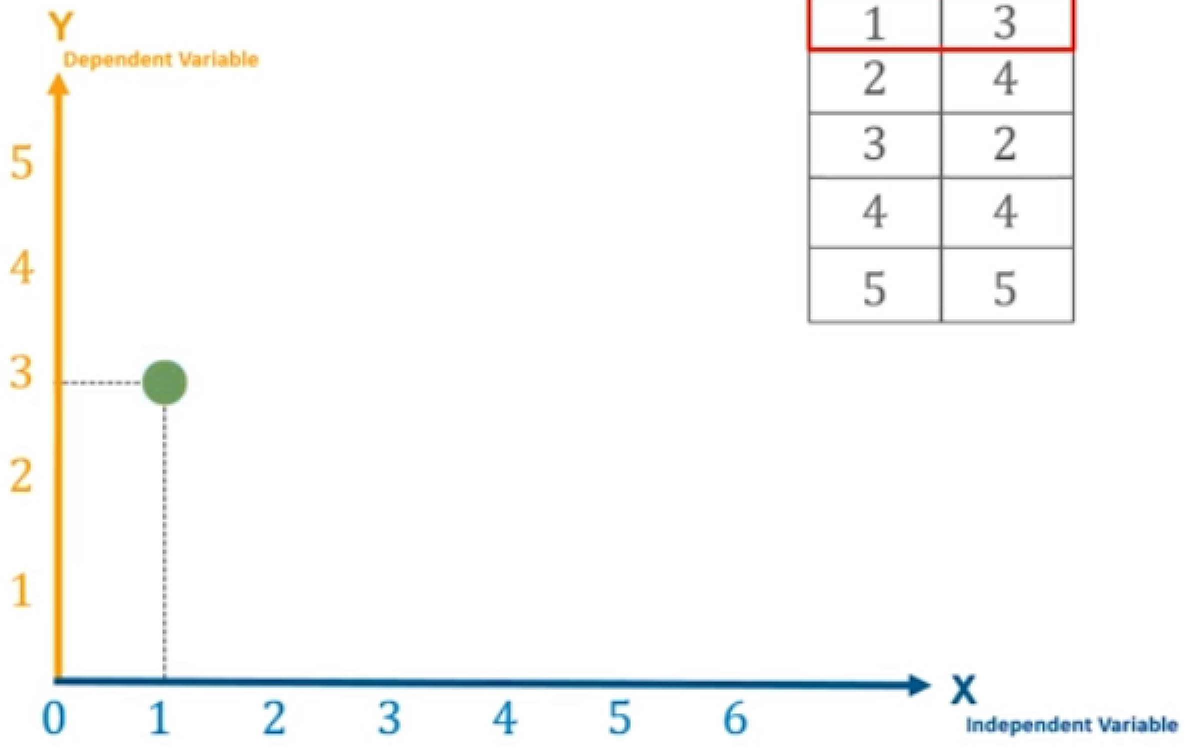

Y
Distance
y = mx + c
X
Speed


Y
Time
-ve Relationship
X
Speed

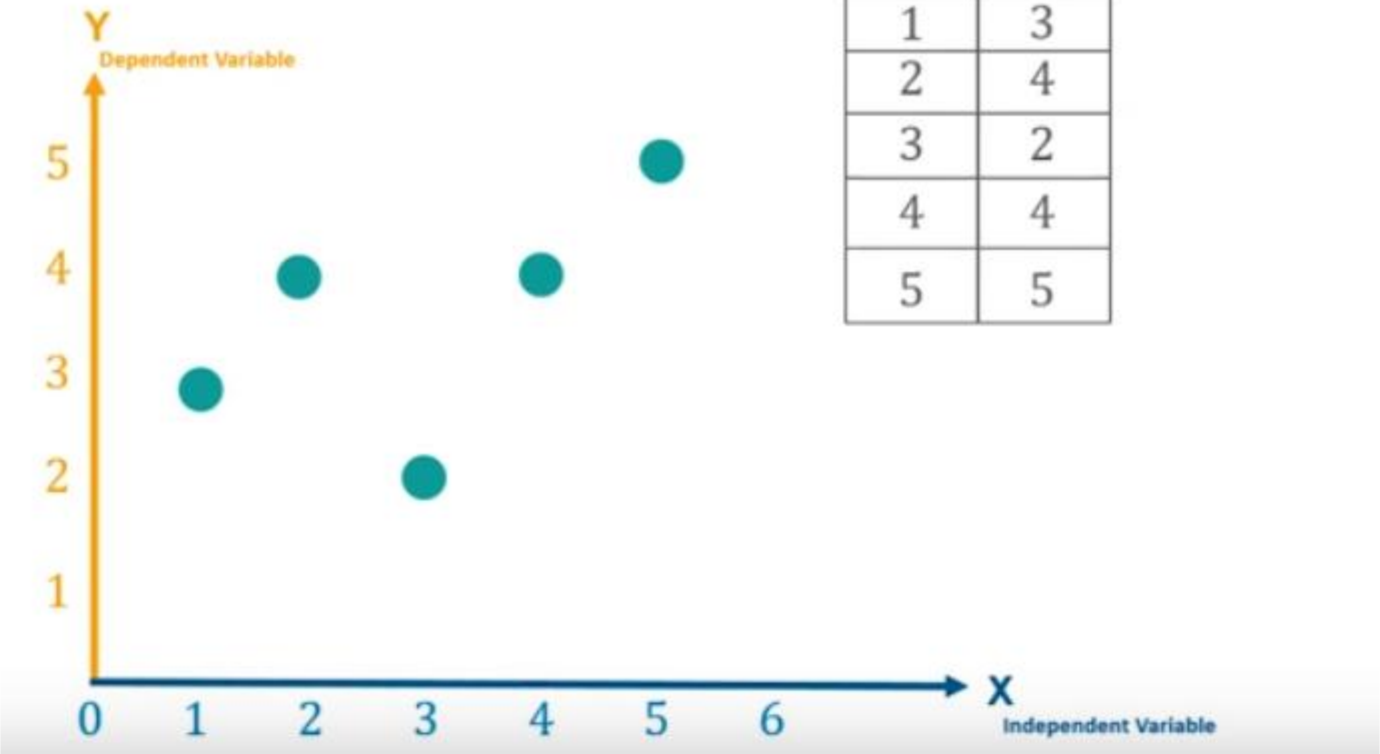

Y
Dependent Variable
Independent Variable
y = mx + c
Dependent Variable
X
Independent Variable



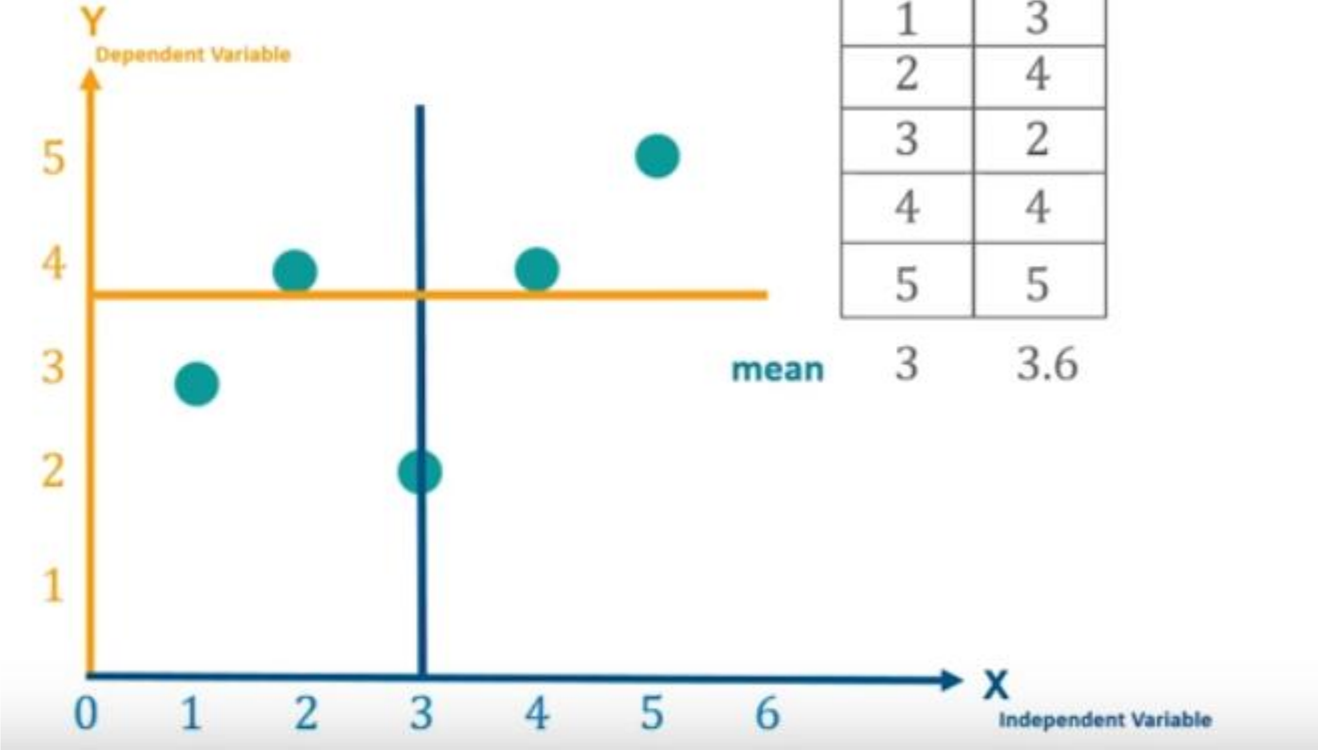
x	y
1	3
2	4
3	2
4	4
5	5



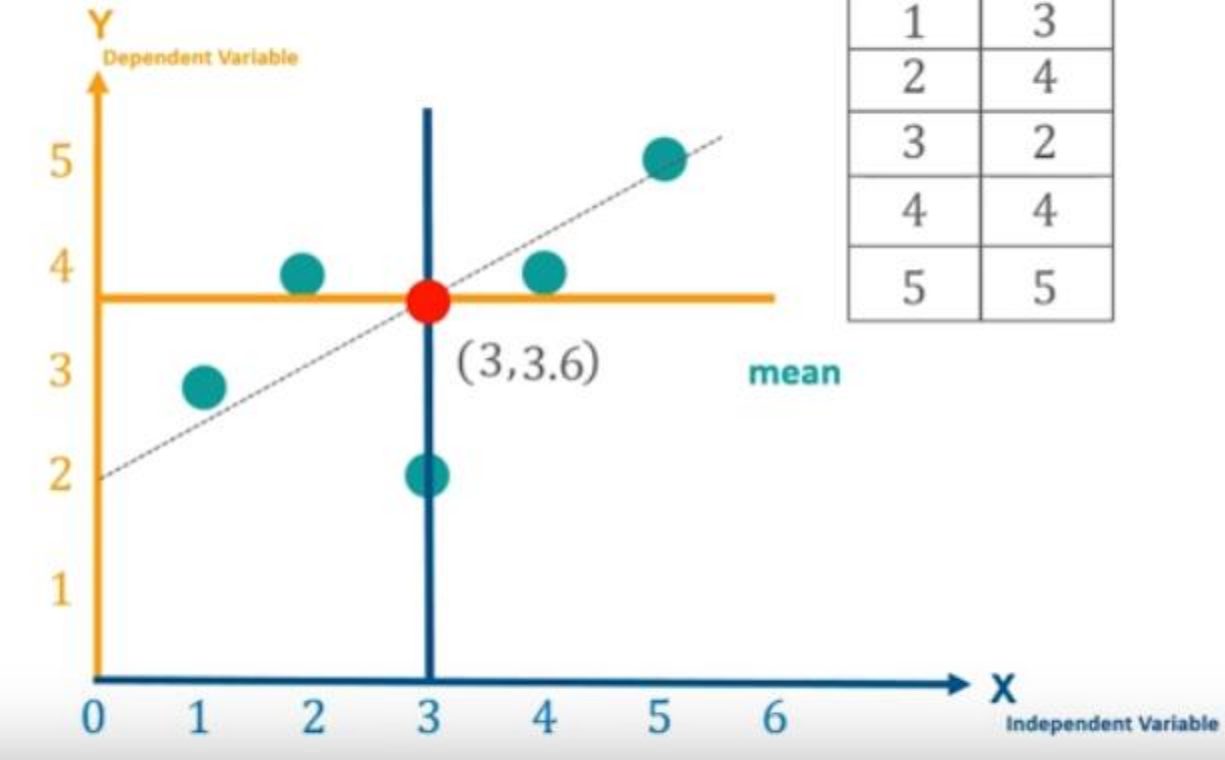
x	y
1	3
2	4
3	2
4	4
5	5



x	y
1	3
2	4
3	2
4	4
5	5

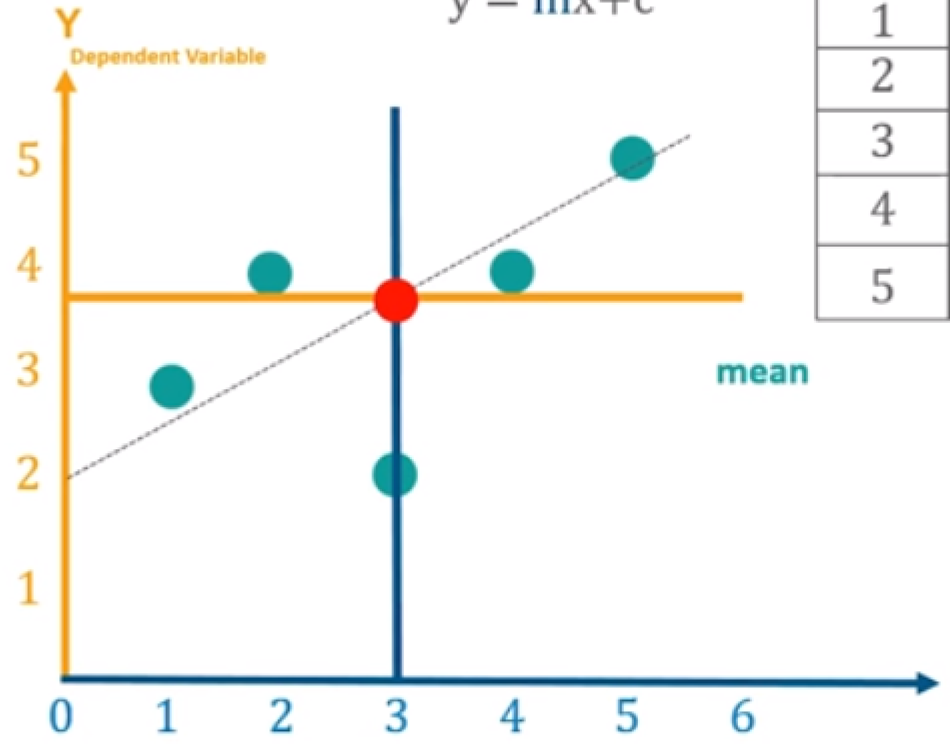


x	y
1	3
2	4
3	2
4	4
5	5
mean 3	3.6



x	y
1	3
2	4
3	2
4	4
5	5

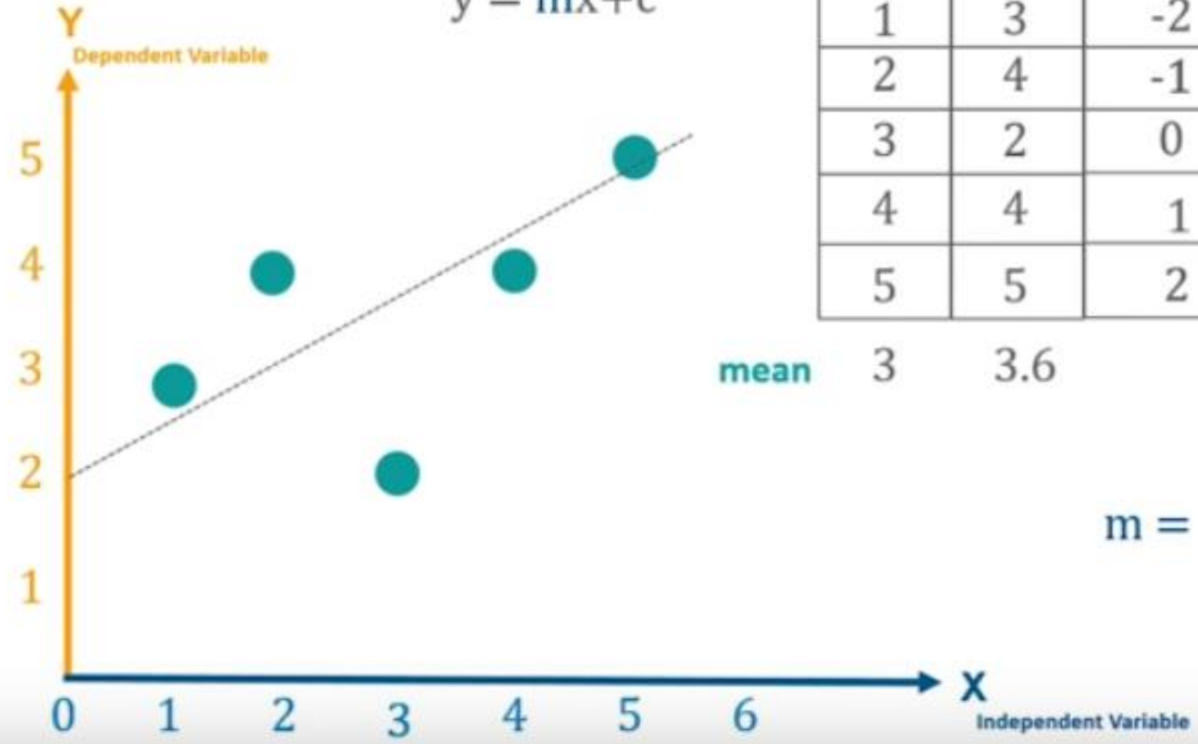
$y = mx + c$



x	y
1	3
2	4
3	2
4	4
5	5

$$m = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$y = mx + c$



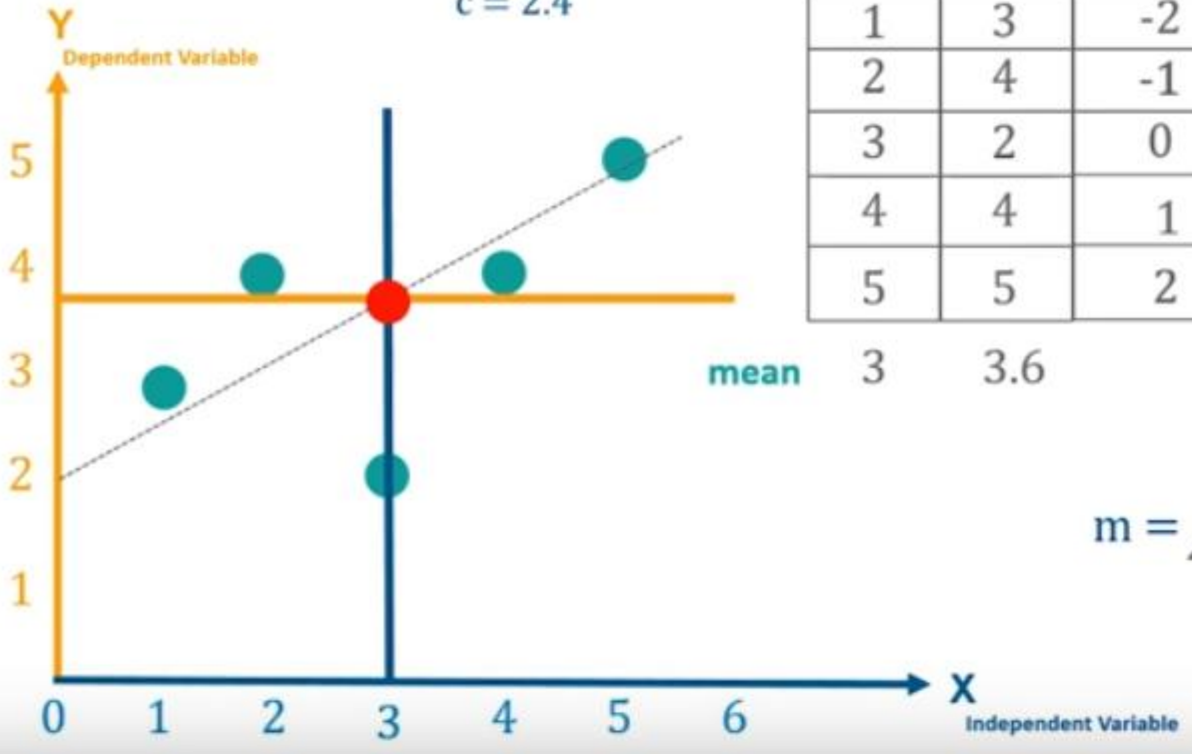
	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	1	3	-2	-0.6	4	1.2
	2	4	-1	0.4	1	-0.4
	3	2	0	-1.6	0	0
	4	4	1	0.4	1	0.4
	5	5	2	1.4	4	2.8
mean	3	3.6			$\Sigma = 10$	$\Sigma = 4$

$$m = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\Sigma (x - \bar{x})^2} = \frac{4}{10}$$

- $Y=mx+c$
- $Y=3.6$, $m=0.4$, $x=3$
- $3.6=0.4*3+c$
- $3.6-1.2=c$
- $C=2.4$

$y = mx + c$

$c = 2.4$



x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
mean 3	3.6			$\Sigma = 10$	$\Sigma = 4$

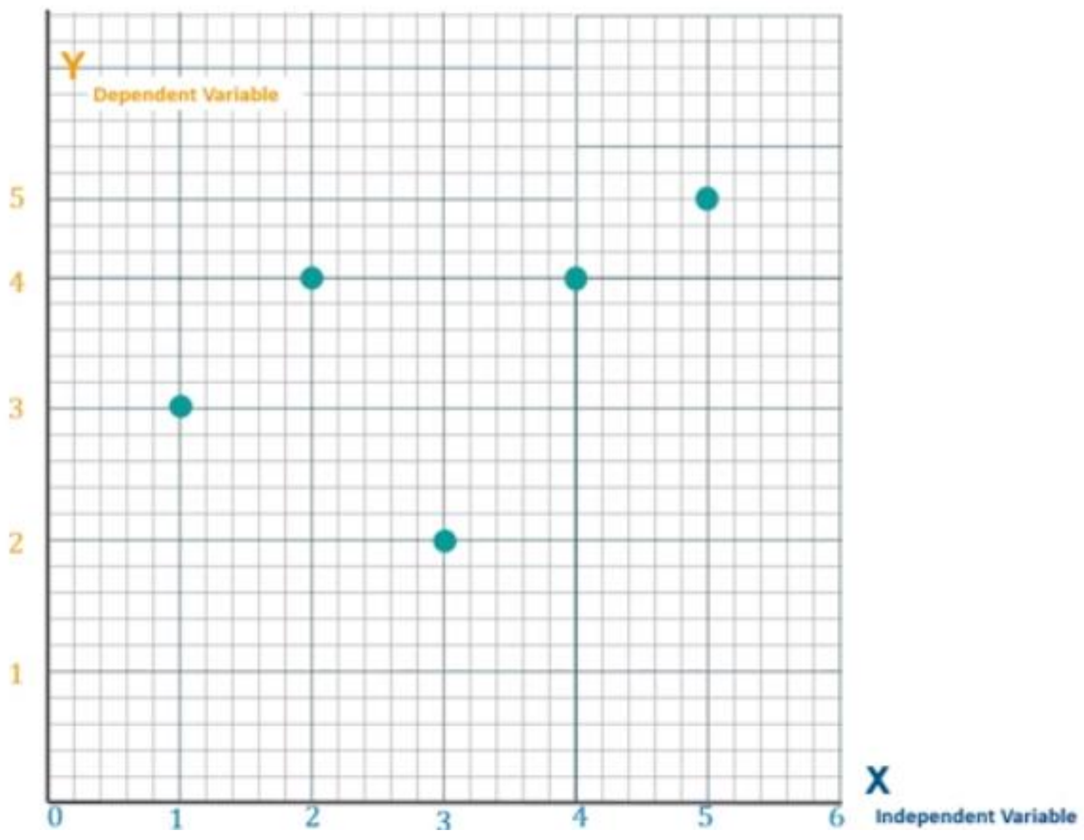
$$m = \sum \frac{(x - \bar{x})(y - \bar{y})}{(x - \bar{x})^2} = \frac{4}{10}$$

$m = 0.4$

$c = 2.4$

$y = 0.4x + 2.4$

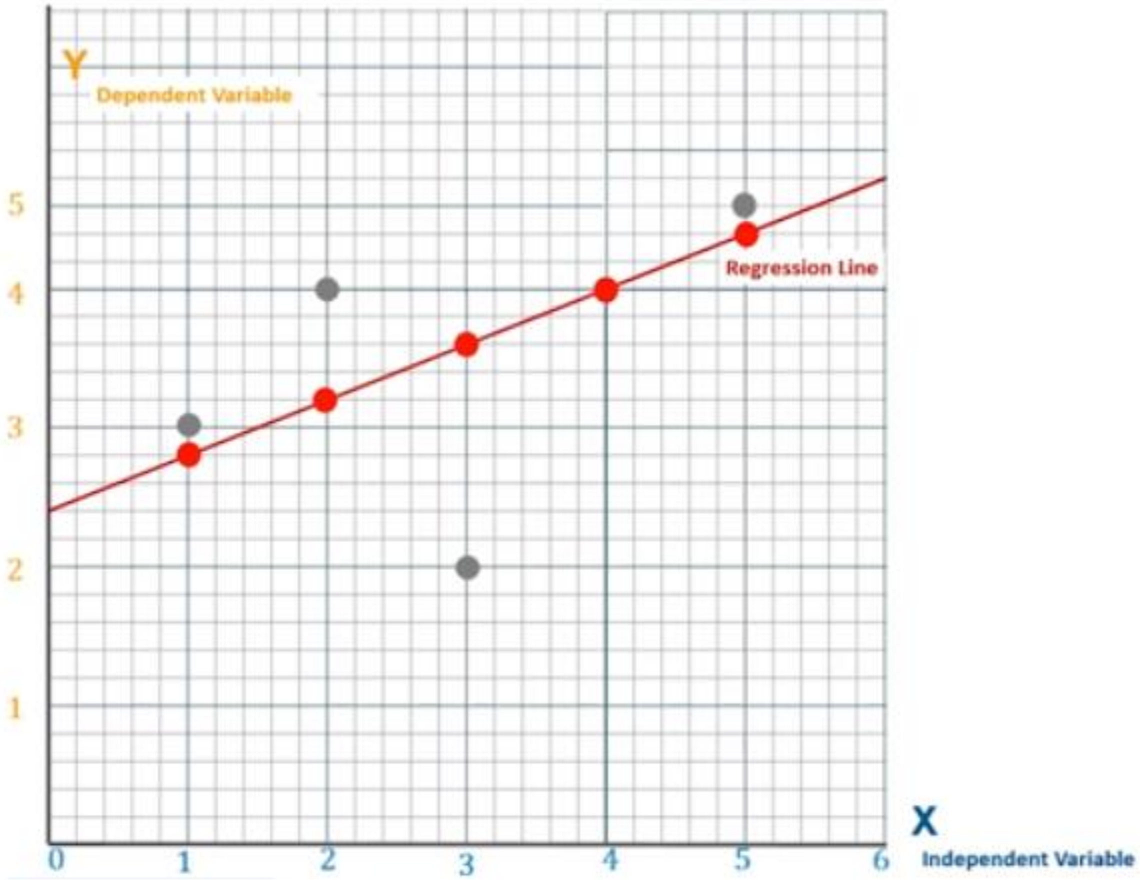
Mean Square Error



$m = 0.4$
$c = 2.4$
$y = 0.4x + 2.4$

For given m = 0.4 & c = 2.4, lets predict values for y for x = {1,2,3,4,5}

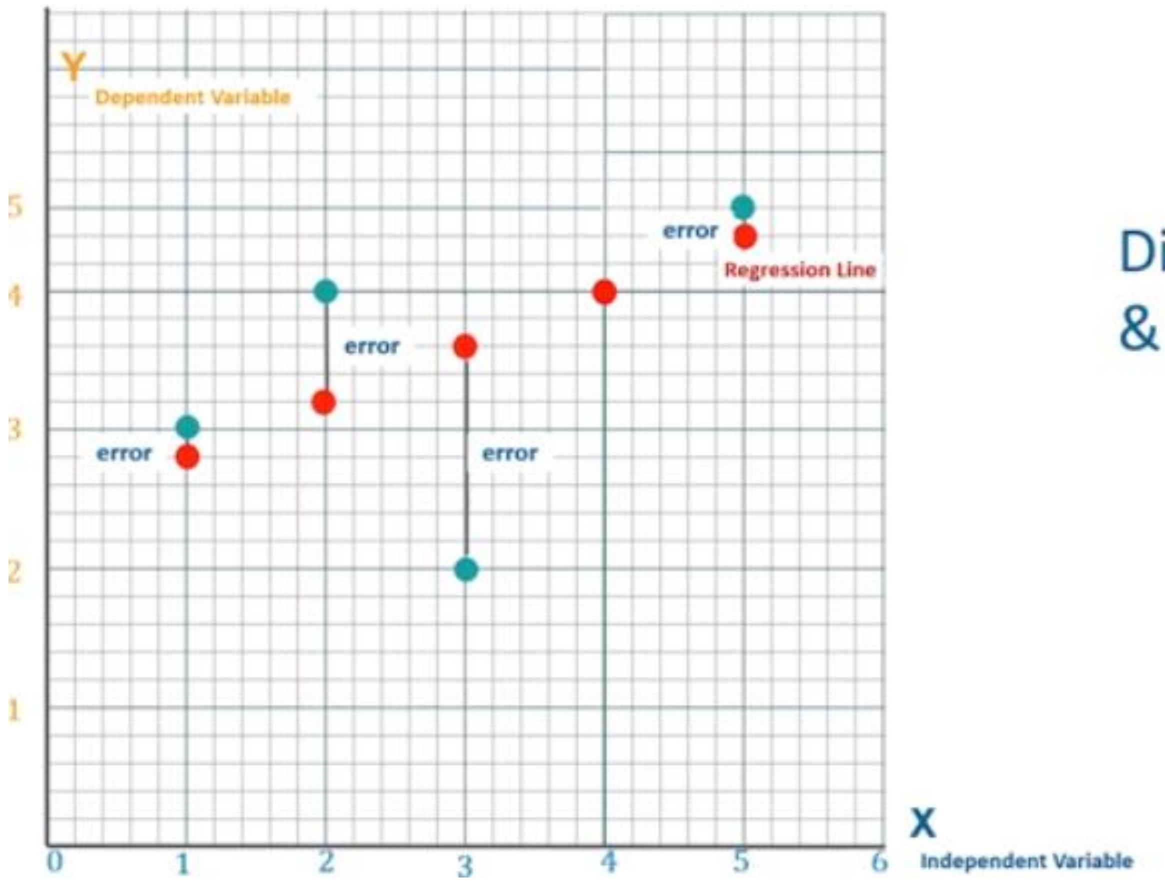
y = 0.4 x 1+ 2.4 = 2.8
y = 0.4 x 2+ 2.4= 3.2
y = 0.4 x 3+ 2.4= 3.6
y = 0.4 x 4+ 2.4= 4.0
y = 0.4 x 5+ 2.4= 4.4



$m = 0.4$
$c = 2.4$
$y = 0.4x + 2.4$

For given $m = 0.4$ & $c = 2.4$, lets predict values for y for $x = \{1,2,3,4,5\}$

$y = 0.4 \times 1 + 2.4 = 2.8$
$y = 0.4 \times 2 + 2.4 = 3.2$
$y = 0.4 \times 3 + 2.4 = 3.6$
$y = 0.4 \times 4 + 2.4 = 4.0$
$y = 0.4 \times 5 + 2.4 = 4.4$

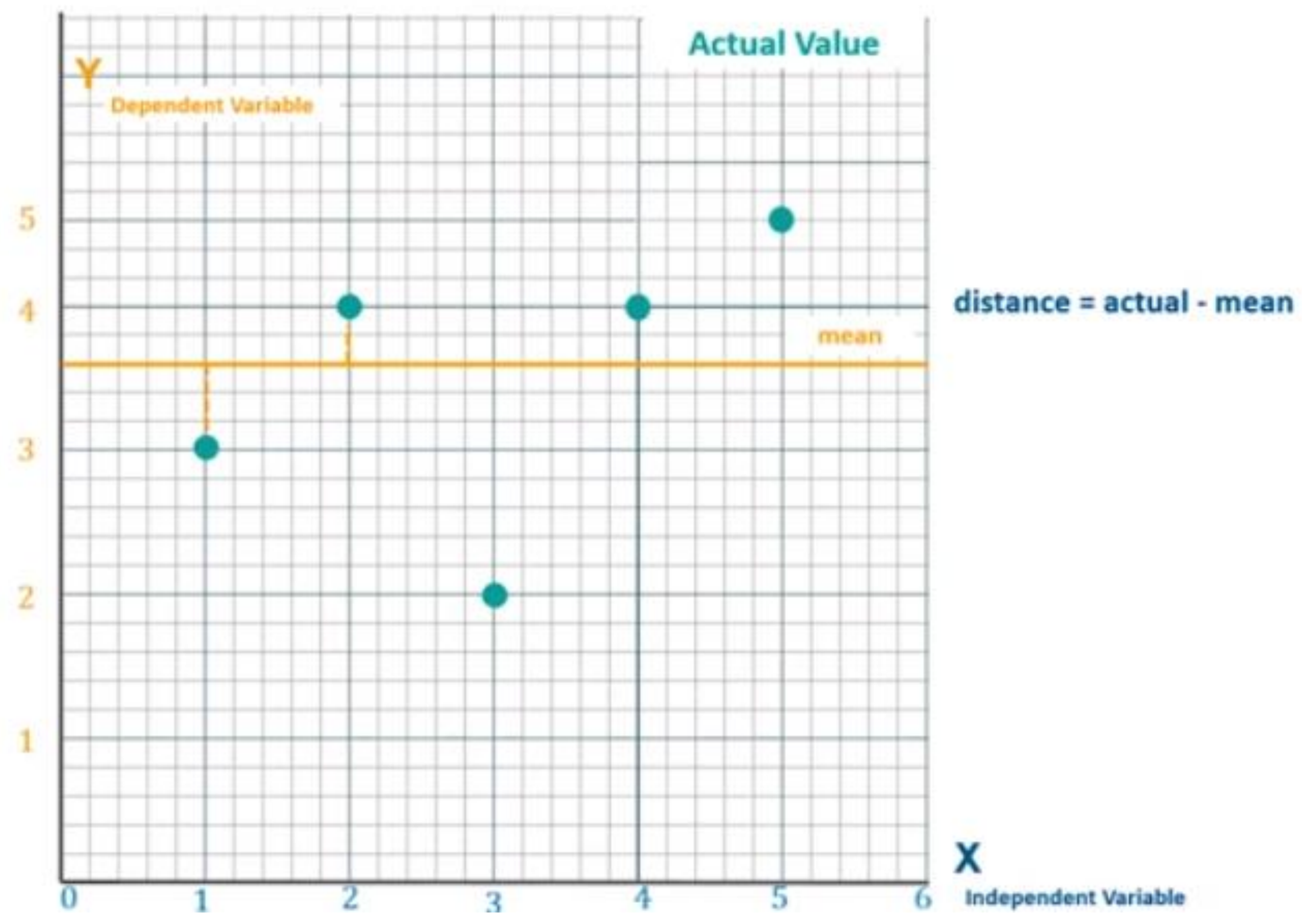


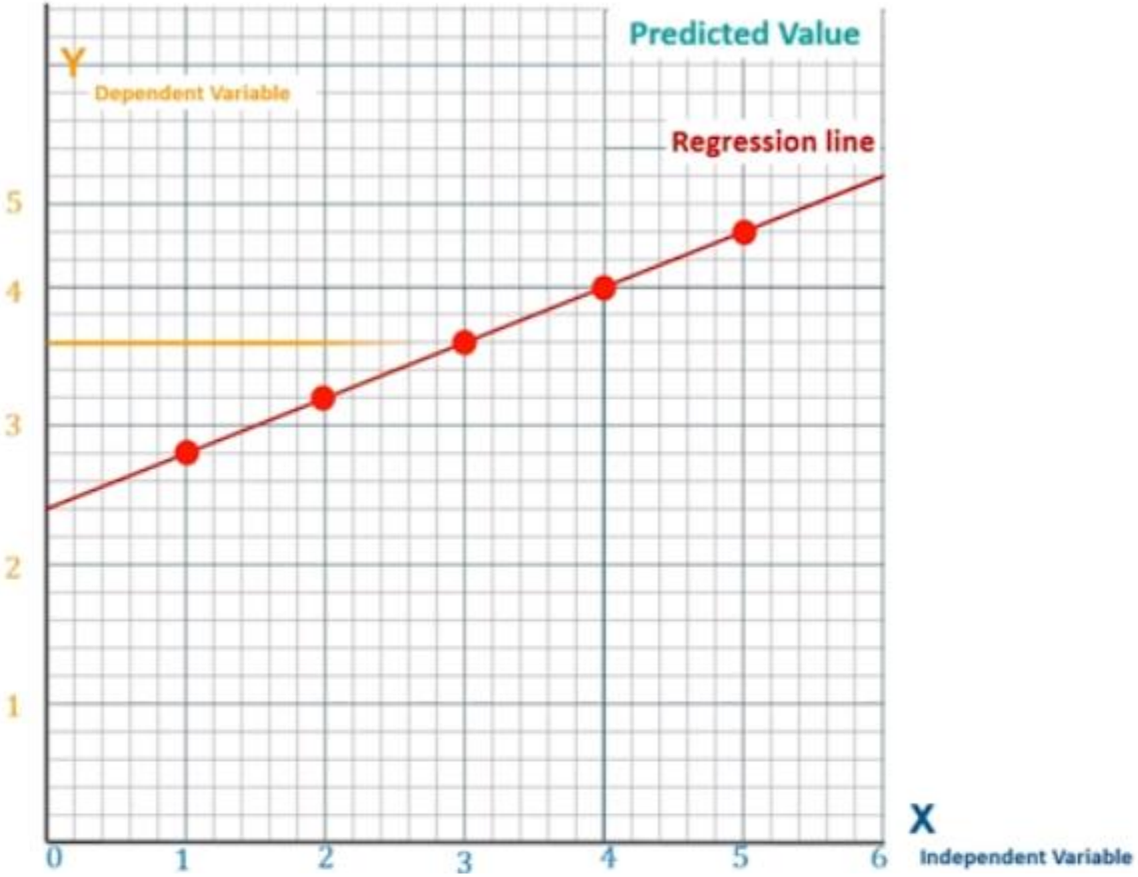
Distance between actual & predicted value

R Square:-

- **R-squared** value is a statistical measure of how close the data are to the fitted regression line
- It is also known as **coefficient of determination**, or the **coefficient of multiple determination**

Calculation of R square





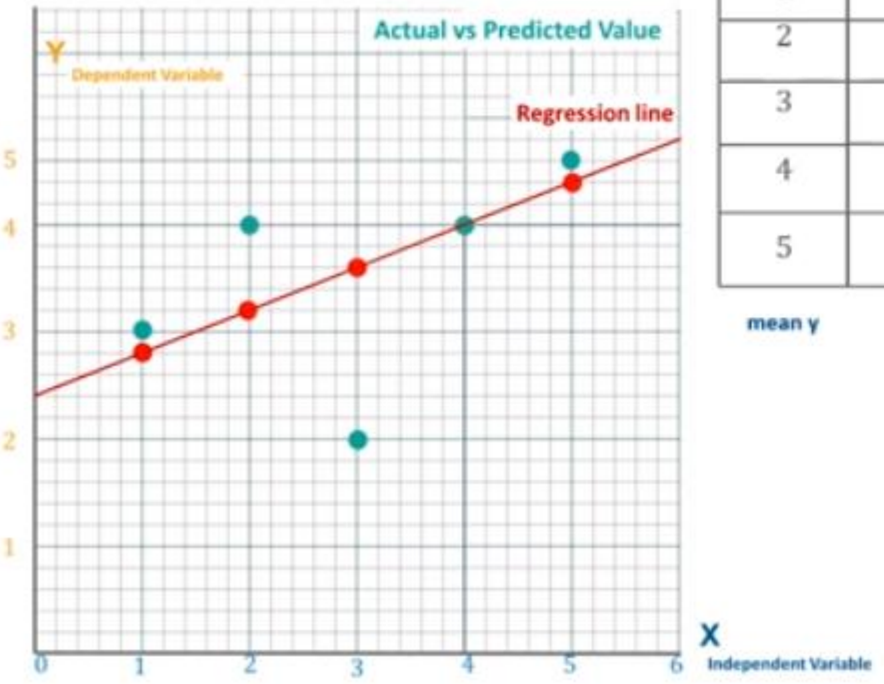
x	y_p
1	2.8
2	3.2
3	3.6
4	4.0
5	4.4

Distance actual - mean

vs

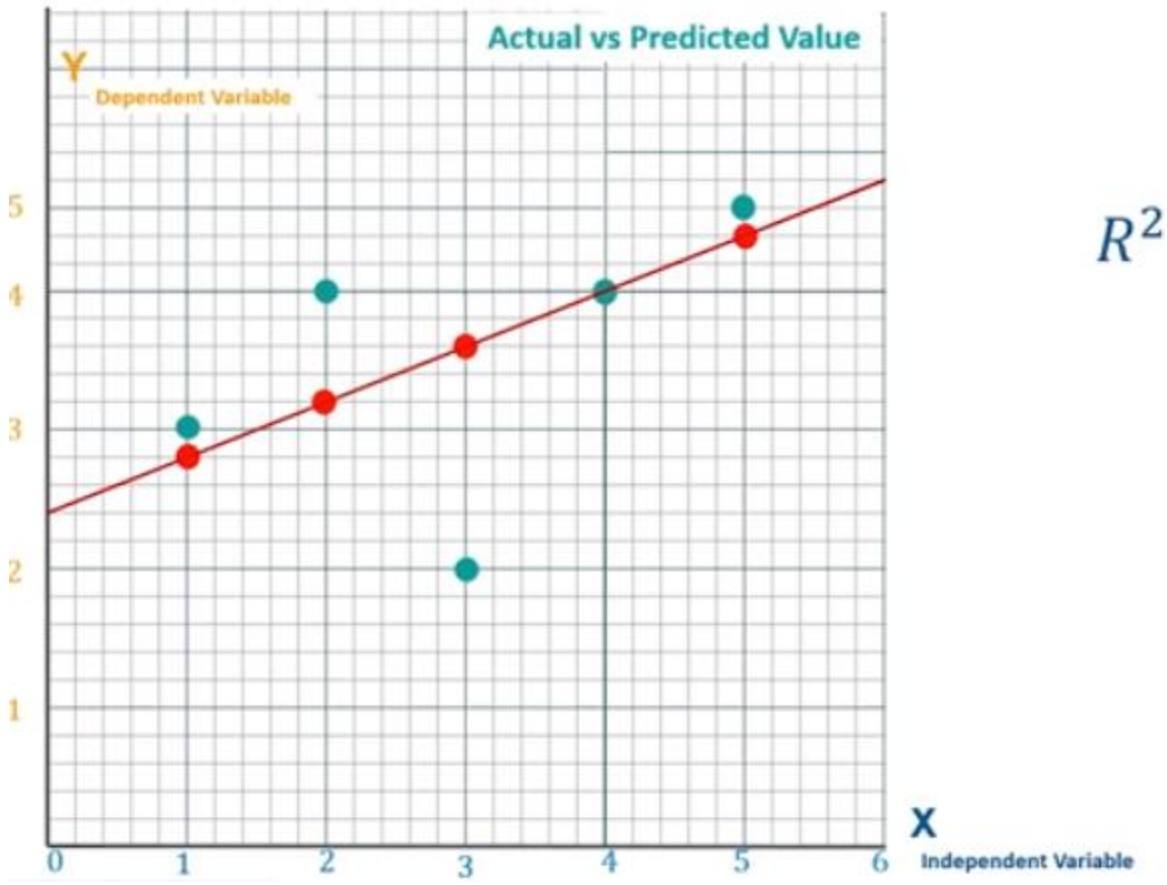
Distance predicted - mean

This is nothing but $R^2 = \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$



x	y	$y - y$	$(y - y)^2$	y_p	$(y_p - y)$	$(y_p - y)^2$
1	3	− 0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	−1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64
mean y	3.6		Σ 5.2			Σ 1.6

$$R^2 = \frac{1.6}{5.2} = \frac{\Sigma\left((y_p - \bar{y})\right)^2}{\Sigma (y - \dot{y})^2}$$



Actual vs Predicted Value
Y
Dependent Variable
X
Independent Variable
0
1
2
3
4
5
6
1
2
3
4
5

$$R^2 \approx 0.3$$