Do It Skills
do it. enjoy it.

# 9 PM EST, MONDAY 11 NOVEMBER 2024

## SALARY EXPERIENCE MACHINE LEARNING SLIDES

Do It Skills
do it. enjoy it.

# LINEAR REGRESSION

LEARN EASY. ENJOY EASY. EXPERTISE EASY

Do It Skills
do it. enjoy it.

# PREDICT APPLICANTS SALARY -

USING LINEAR REGRESSION ALGORITHM

# LOAN APPLICATIONS – LET US PREDICT "SALARY" WHEN THE INPUT IS "YEARS OF EXPERIENCE"

You distribute Loans. When you distribute loans, Salary of a person or Income of an employee is an important component. Applicants will fill in their details of Income , they will also submit tax returns etc.

However you want to build a Machine Learning model that will Predict salary for years experience

| Exp | $ Salary |
|---|---|
| 2 | 15 |
| 3 | 28 |
| 5 | 42 |
| 13 | 64 |
| 8 | 50 |
| 16 | 90 |
| 11 | 58 |
| 1 | 8 |
| 9 | 54 |

What will be Salary if the Experience is 15?

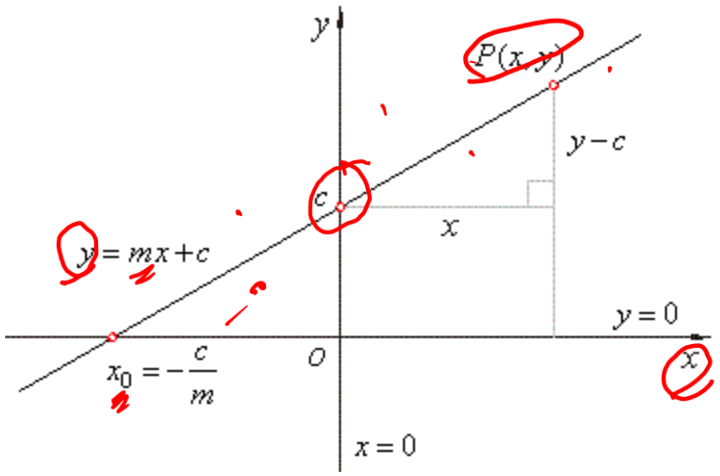
# LINEAR REGRESSION

Linear regression models are used to show or predict the relationship between two variables or factors. The factor that is being predicted (the factor that the equation solves for) is called the dependent variable. The factors that are used to predict the value of the dependent variable are called the independent variables.

In linear regression, each observation consists of two values. One value is for the dependent variable and one value is for the independent variable. In this simple model, a straight line approximates the relationship between the dependent variable and the independent variable.

When two or more independent variables are used in regression analysis, the model is no longer a simple linear one. This is known as multiple regression.

# LINEAR REGRESSION - EQUATION




**y = mx + c**

The slope of the line is ***m***, and ***c*** is the intercept (the value of ***y*** when ***x*** = 0).

Speed
=

→ Statistics model 18 century

→ Until 2014 — SPSS (IBM) → 
Libraries — SASS → 
MathLab ]

R open source → ML
→ 2000 year

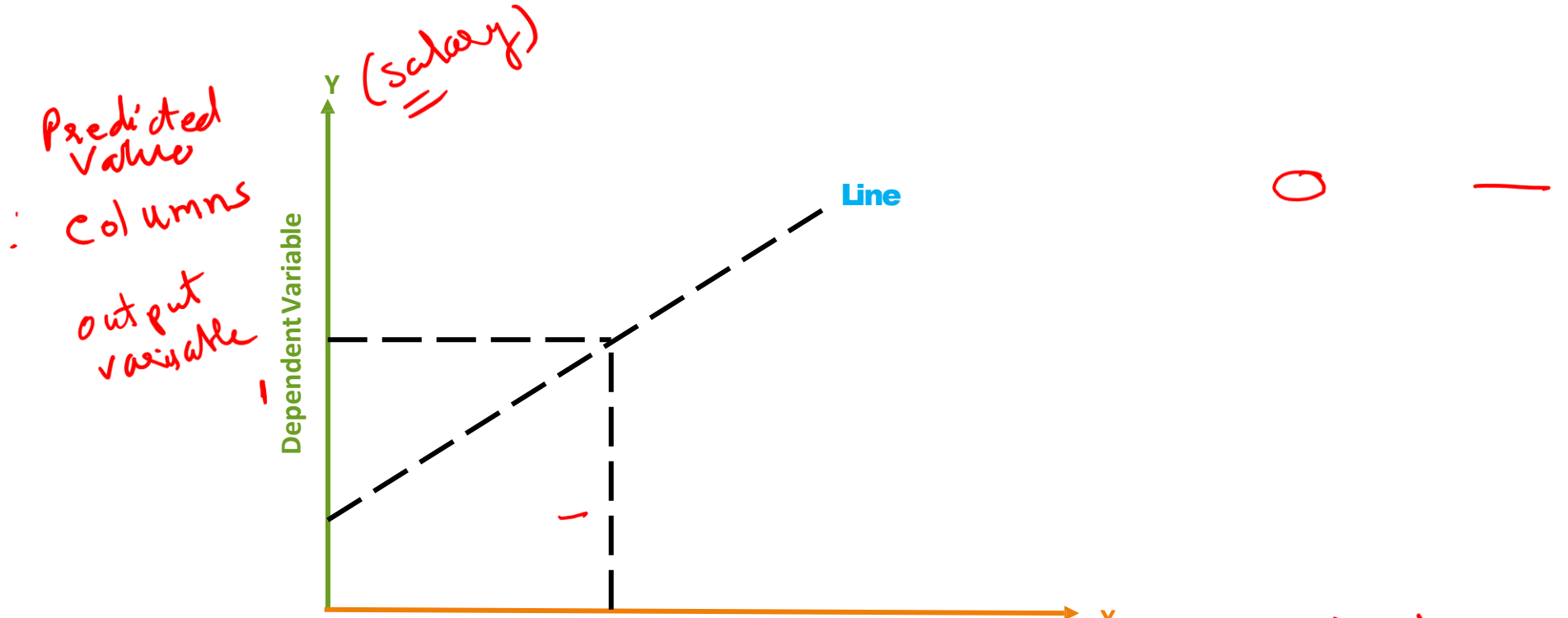
S
↳ Statistical

S

R → =

→ Python → 60,000 libraries
=
↳ Linear Reg → Scikit learn.

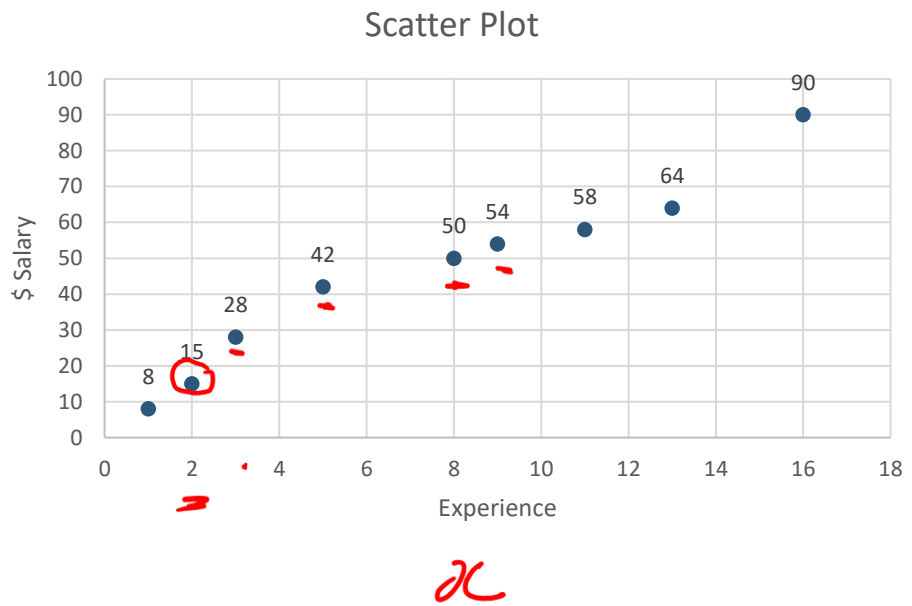
# LINEAR REGRESSION – INDEPENDENT AND DEPENDENT VARIABLES



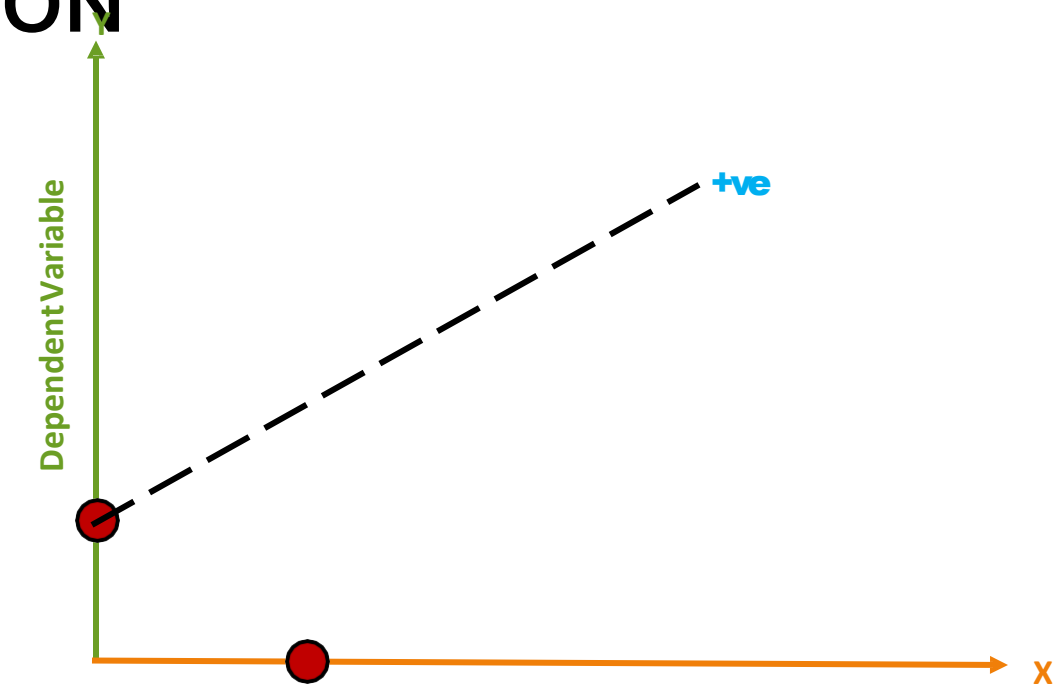

Do It Skills
do it. enjoy it.

# LINEAR REGRESSION – INDEPENDENT AND DEPENDENT VARIABLES

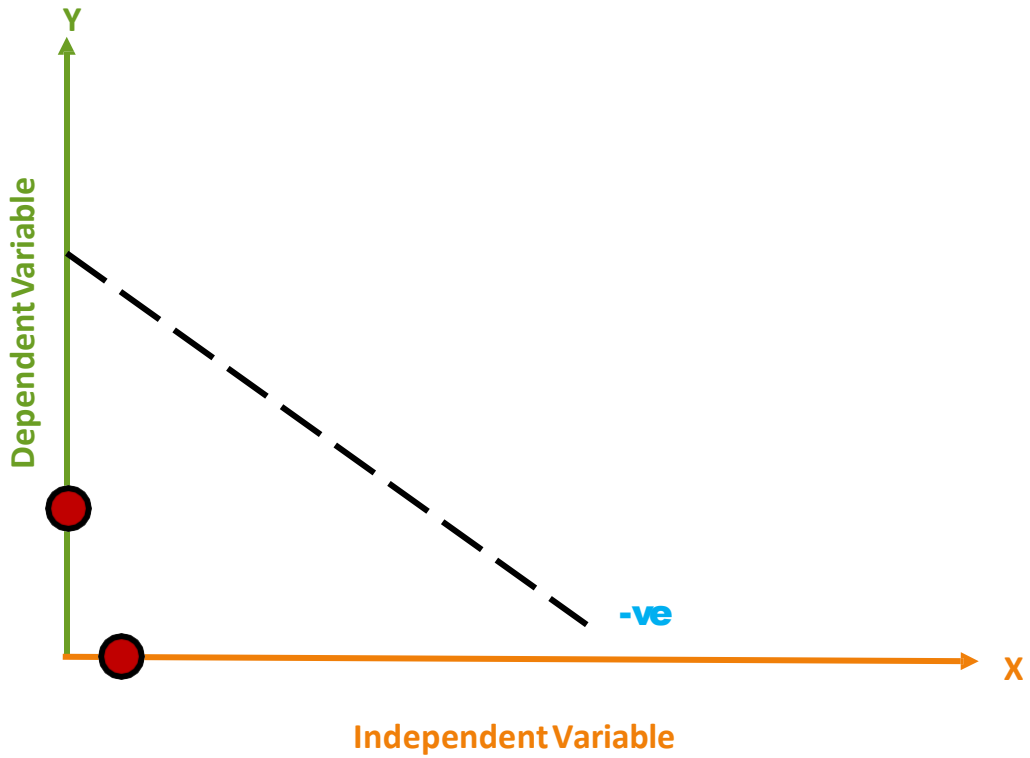
| Exp | $ Salary |
|---|---|
| 2 | 15 |
| 3 | 28 |
| 5 | 42 |
| 13 | 64 |
| 8 | 50 |
| 16 | 90 |
| 11 | 58 |
| 1 | 8 |
| 9 | 54 |



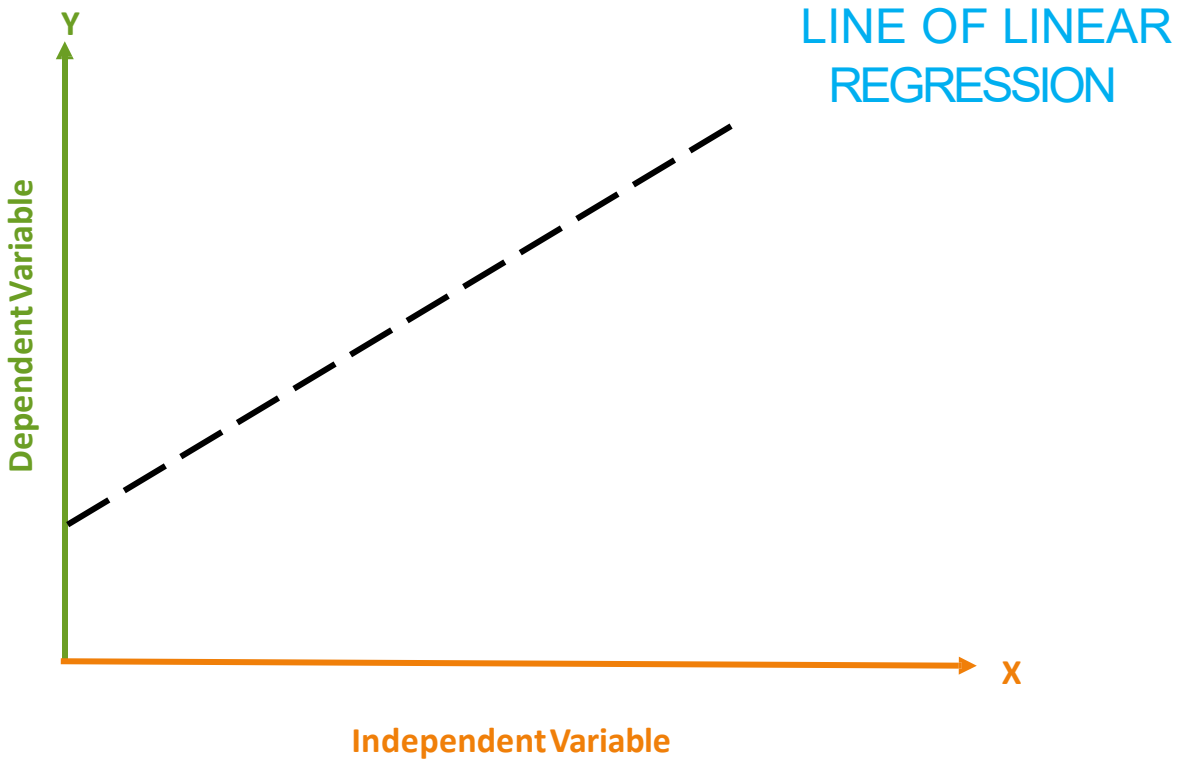
# POSITIVE RELATION



# NEGATIVE RELATION

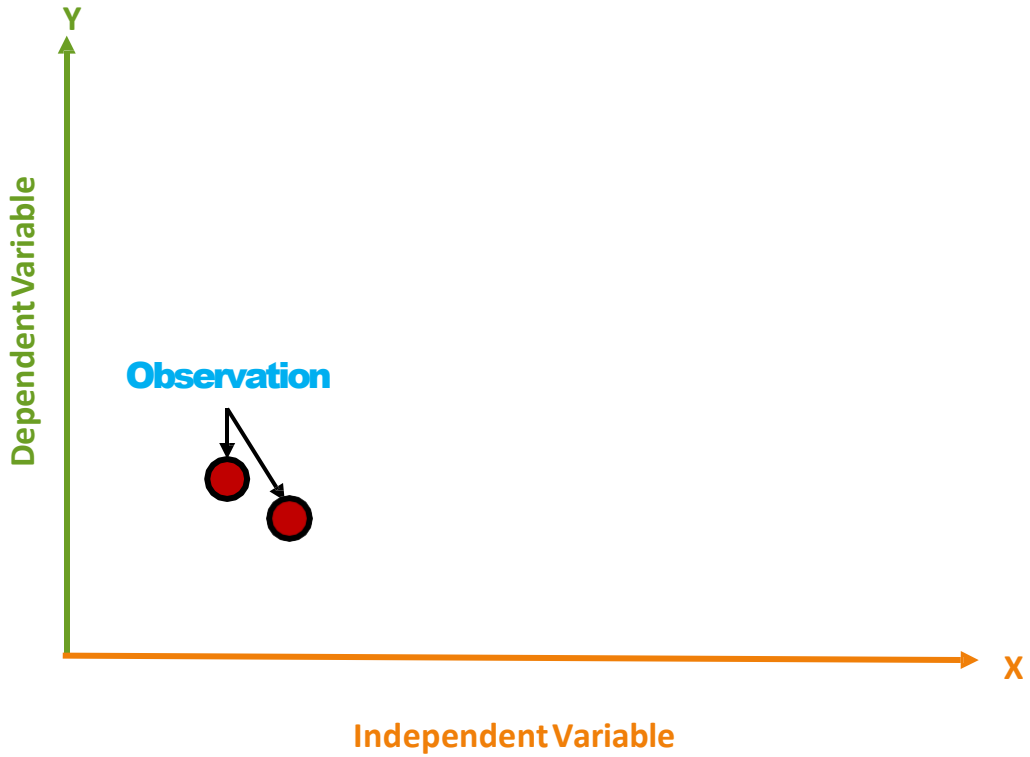


# LINEAR REGRESSION

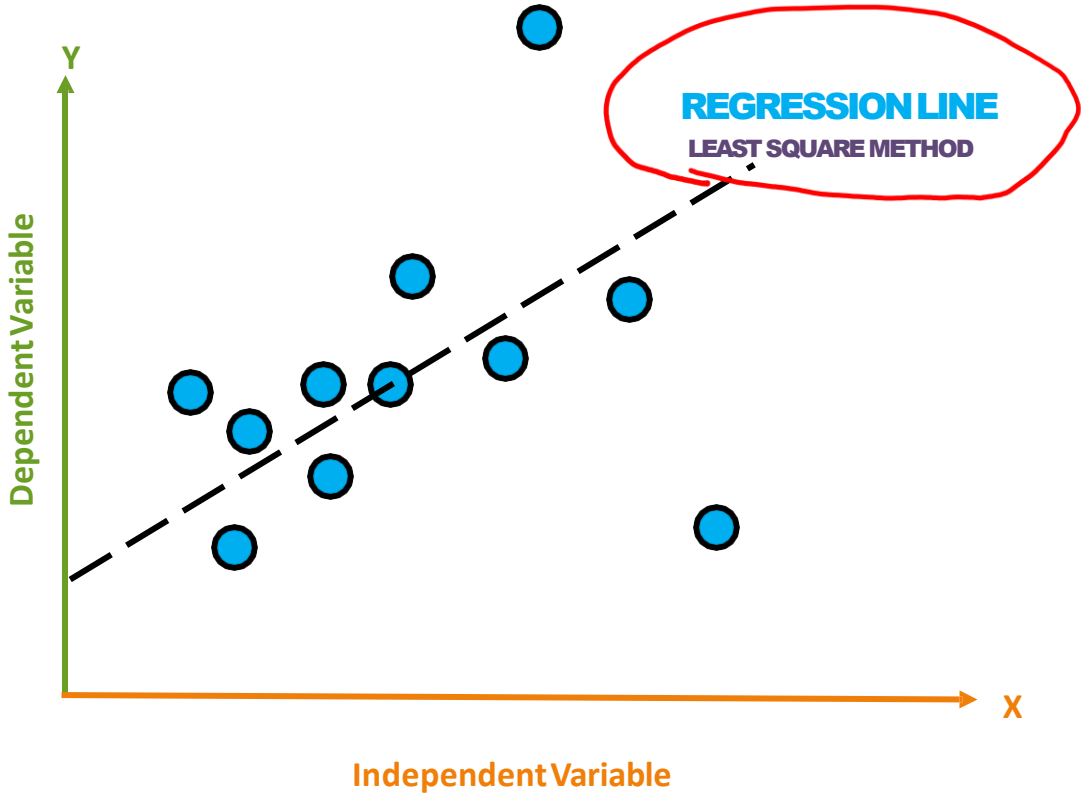



Do It Skills
do it. enjoy it.

# DATA POINTS - OBSERVATIONS

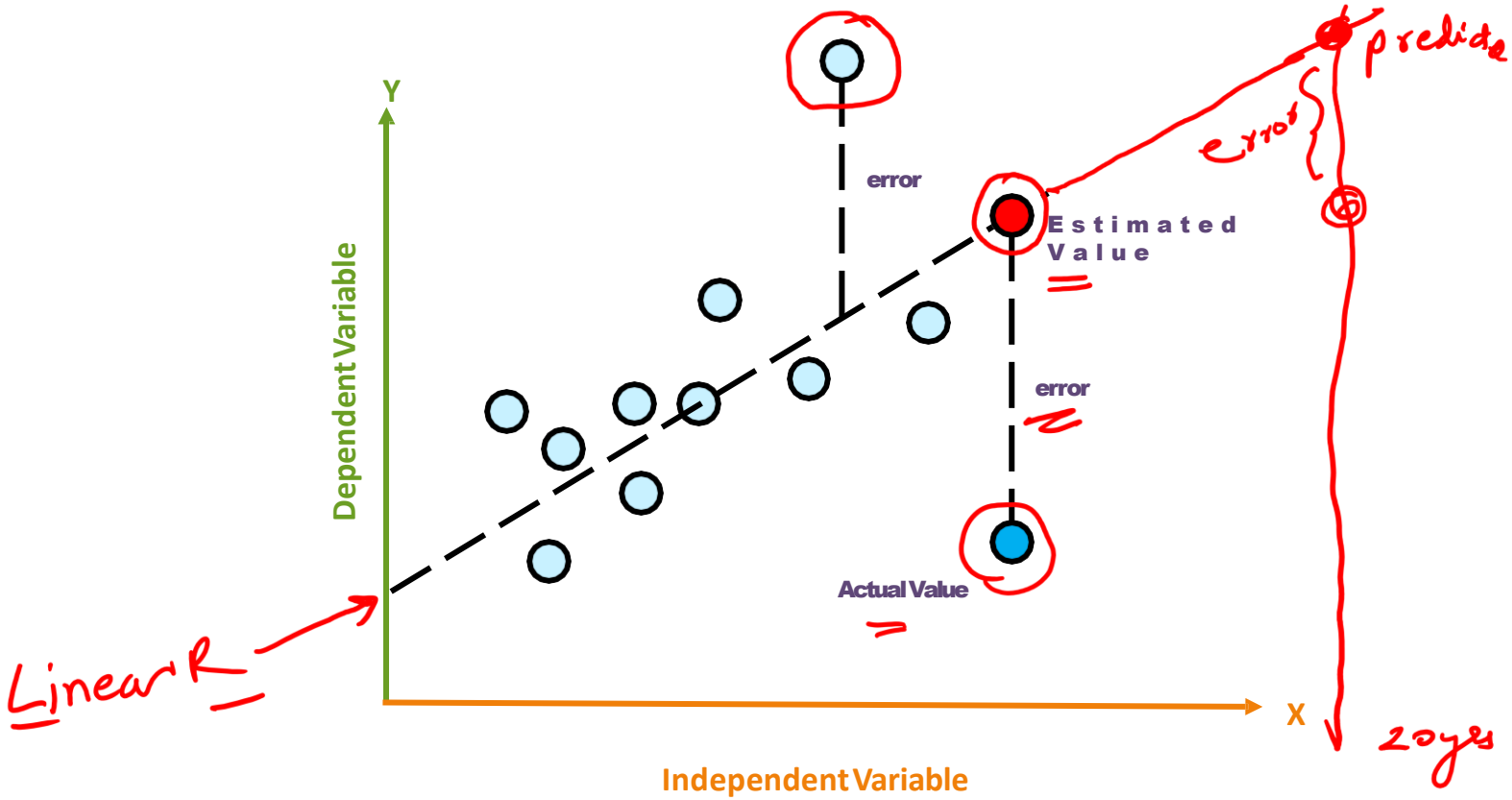


# PLOTTING THE BEST FIT LINE – USING THE LEAST SQUARE METHOD (ORDINARY LEAST SQUARES)



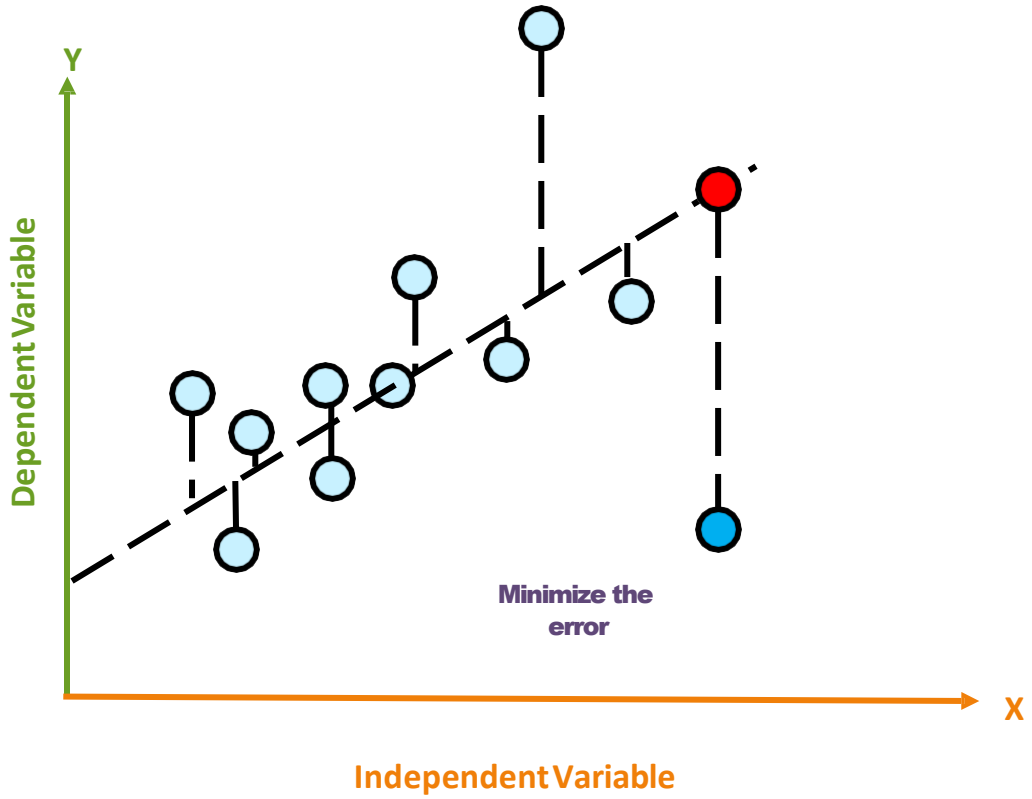

Do It Skills
do it. enjoy it.

# ERROR: ESTIMATED VALUE - PREDICTED VALUE



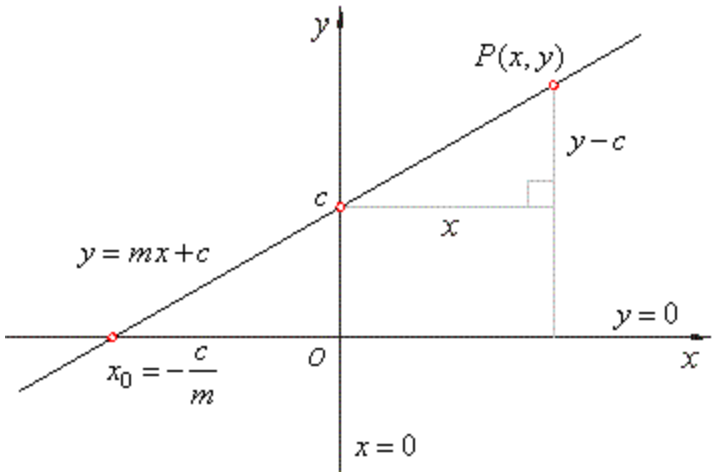

Do It Skills
do it. enjoy it.

# MINIMIZE THE ERROR




Do It Skills
do it. enjoy it.

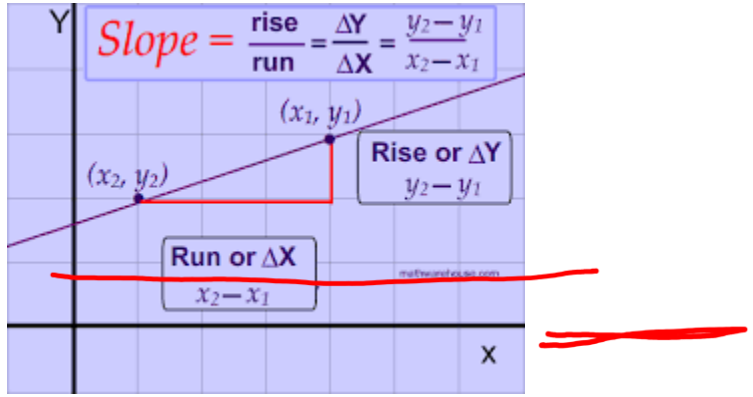
# HOW TO CREATE LINEAR REGRESSION – EQUATION TO PREDICT SALARY

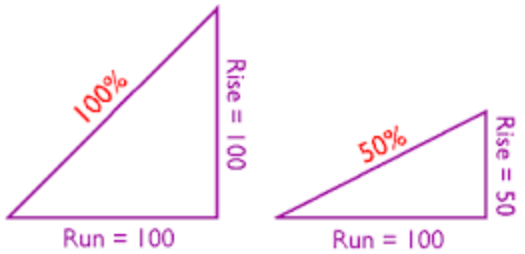



**y = mx + c**

The slope of the line is ***m***, and ***c*** is the intercept (the value of ***y*** when ***x*** = 0).

# CALCULATING THE SLOPE




**Percent slope**
= (rise / run) × 100
= (100 ÷ 100) × 100
= 100

**Percent slope**
= (rise / run) × 100
= (50 ÷ 100) × 100
= 50

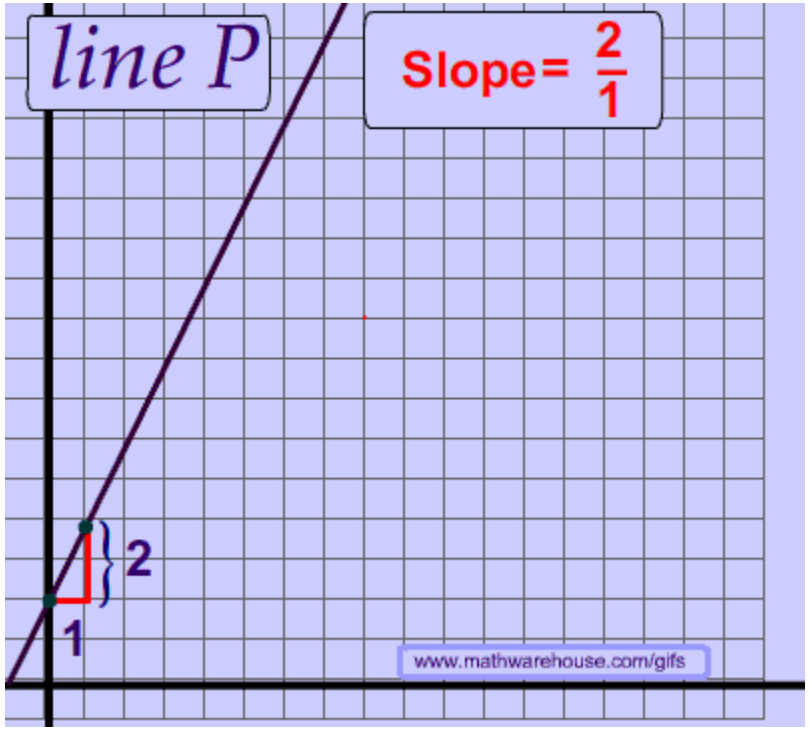
0

$y, x = 0$

$y = 1 \quad x = 0$

Do It Skills
do it. enjoy it.

# HOW A SLOPE WILL DRAW THE LINE

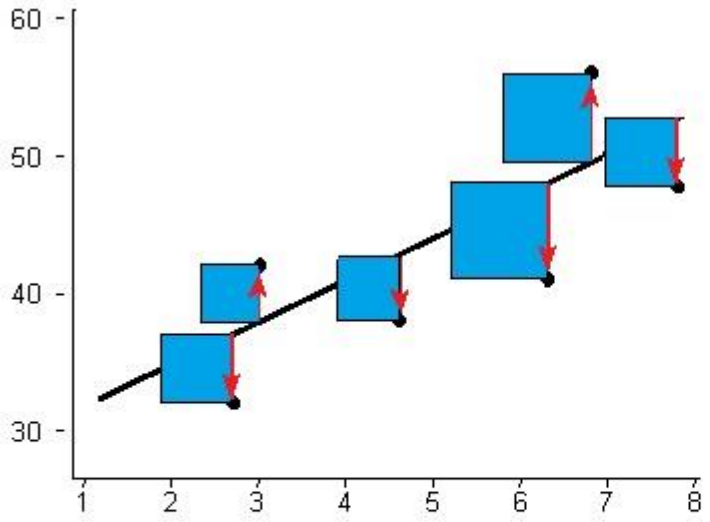



4

# LEAST SQUARES REGRESSION LINE

The Least Squares Regression Line is the line that makes the vertical distance from the data points to the regression line as small as possible. It's called a "least squares" because the best line of fit is one that minimizes the variance (the sum of squares of the errors).

This can be a bit hard to visualize but the main point is you are aiming to find the equation that fits the points as closely as possible.

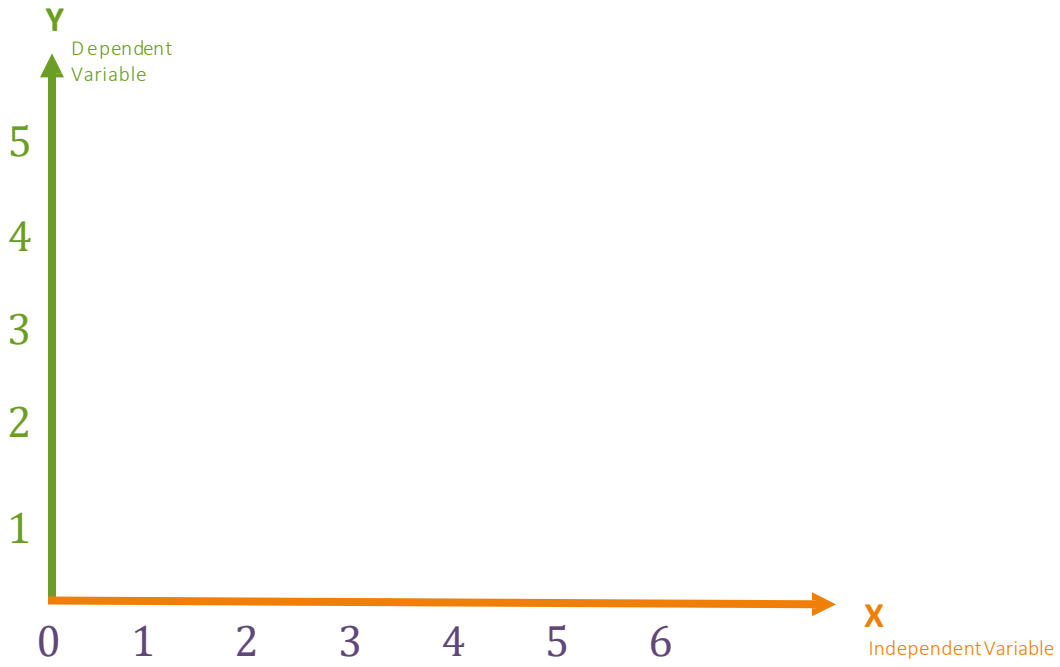


Do It Skills
do it. enjoy it.

# PREDICT SALARY

USING LINEAR REGRESSION EQUATION

# PREDICT THE SALARY WHEN INPUT IS EXPERIENCE




**X** Independent Variable
Exp

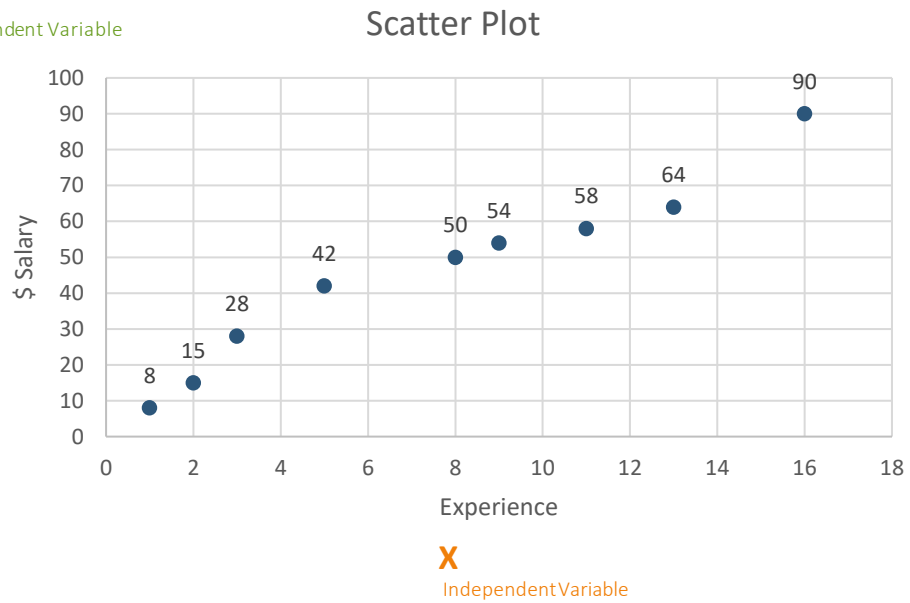
**Y** Dependent Variable
$Salary

| Exp | $ Salary |
|---|---|
| 2 | 15 |
| 3 | 28 |
| 5 | 42 |
| 13 | 64 |
| 8 | 50 |
| 16 | 90 |
| 11 | 58 |
| 1 | 8 |
| 9 | 54 |

Do It Skills
do it. enjoy it.

# PREDICT THE SALARY WHEN INPUT IS EXPERIENCE

Y
Dependent Variable




X
Independent Variable

**X** Independent Variable
Exp

**Y** Dependent Variable
$Salary

| Exp | $ Salary |
|---|---|
| 2 | 15 |
| 3 | 28 |
| 5 | 42 |
| 13 | 64 |
| 8 | 50 |
| 16 | 90 |
| 11 | 58 |
| 1 | 8 |
| 9 | 54 |

Do It Skills
do it. enjoy it.

# SUM OF SQUARED ERRORS

In order to fit the best intercept line between the points in the above scatter plots, we use a metric called “Sum of Squared Errors” (SSE) and compare the lines to find out the best fit by reducing errors. The errors are sum difference between actual value and predicted value.

To find the errors for each dependent value, we need to use the formula below.

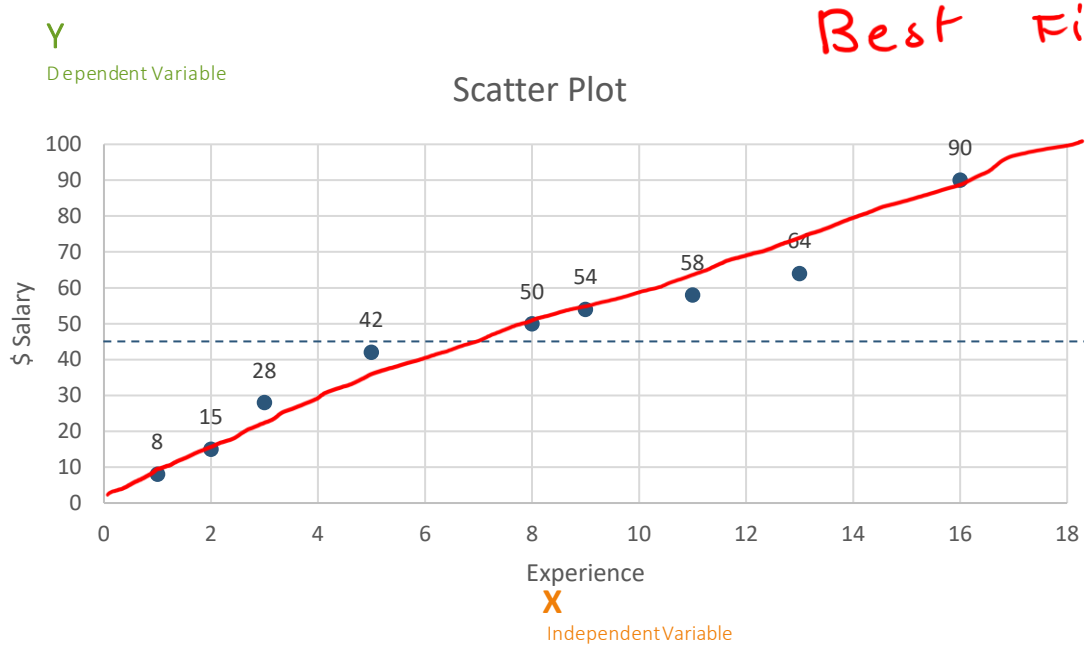
$$SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$y_i$ = Dependent Variables (Salary)

$\bar{y}$ = Average of Dependent Variables

# MEAN LINE OR THE WORST FIT LINE

Best Fit line

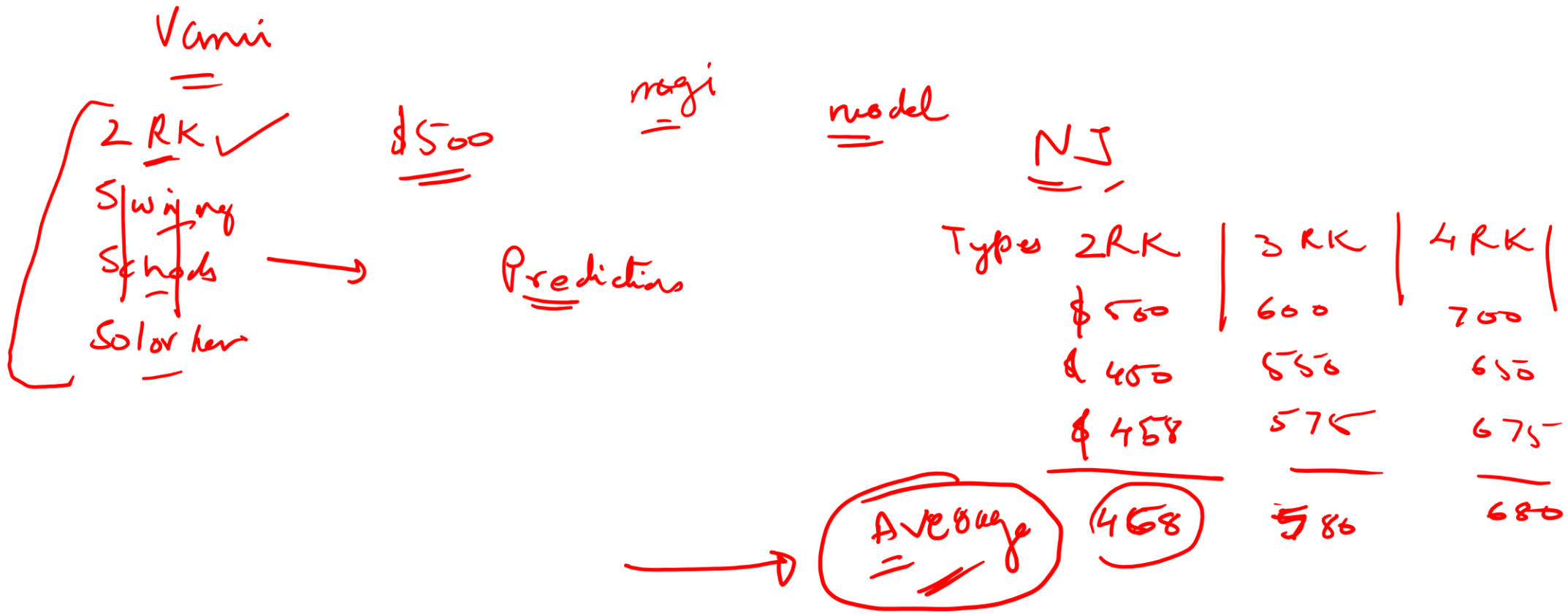


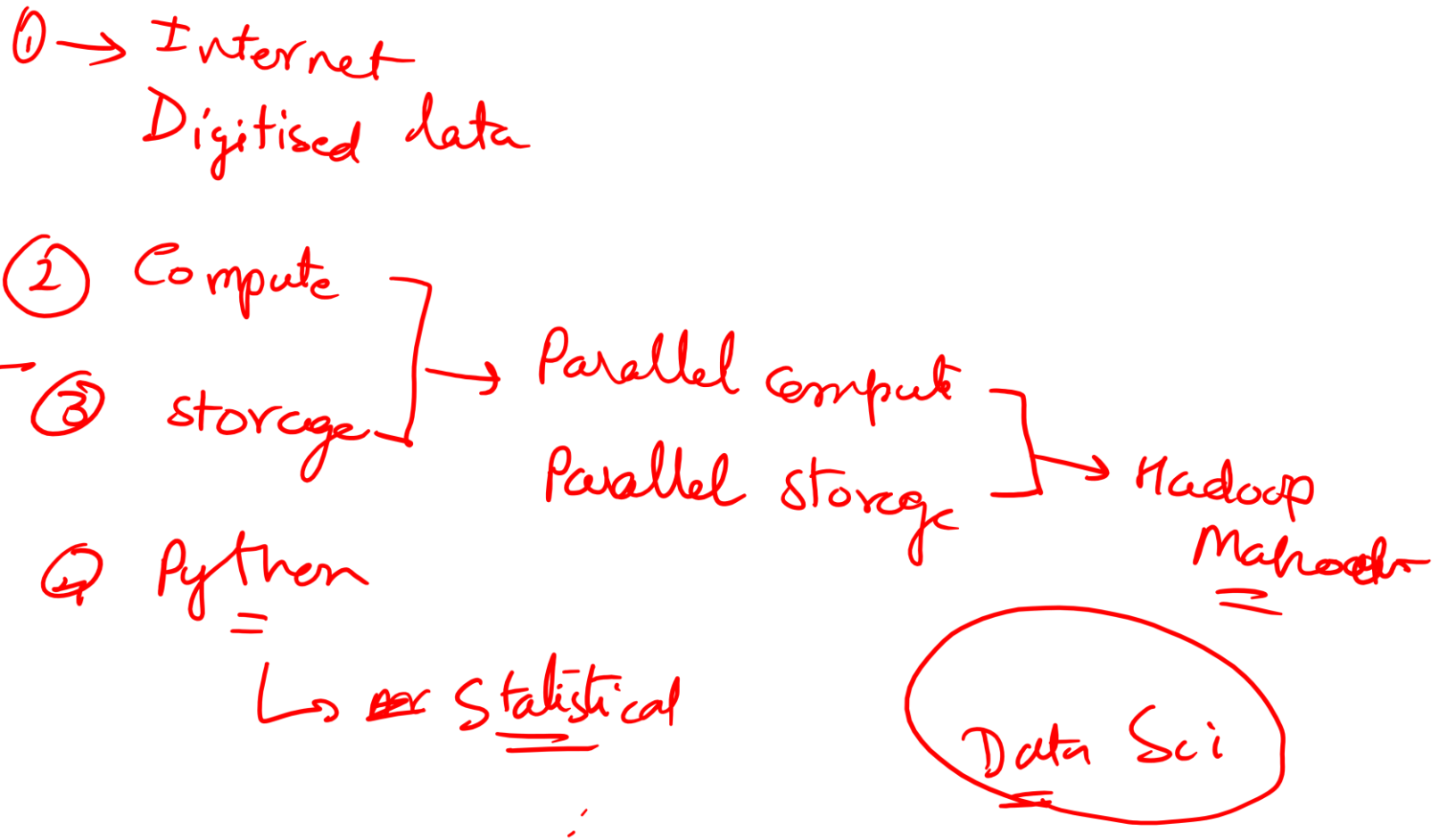

45.4-mean line

| x | y |
|---|---|
| Exp | $ Salary |
| 2 | 15 |
| 3 | 28 |
| 5 | 42 |
| 13 | 64 |
| 8 | 50 |
| 16 | 90 |
| 11 | 58 |
| 1 | 8 |
| 9 | 54 |
| | $\bar{y}$=45.444 |

We find the "Sum of Squared Errors" (SSE) for a Mean line

Do It Skills
do it. enjoy it.


Vanni
2 RK
Swiming
Schools
Solar heat
$500
Predicting
nagi
model
NJ
Types 2RK | 3RK | 4RK
$ 500 | 600 | 700
$ 450 | 550 | 650
$ 458 | 575 | 675
Average
468
586
680


0 → Internet
Digitised data
1 Compute
3 storage
Parallel compute
Parallel storage
Hadoop
Mahoodh
4 Python
Statistical
Data Sci

# SUM OF SQUARED ERRORS

$$SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

$y_i$ = *Dependent Variables (Salary)*

$\bar{y}$ = *Average of Dependent Variables*

The sum of squared errors SSE output is *5226.22* when the line is mean line.

| x | y | E=y-ȳ | Error Sqaure |
|---|---|---|---|
| Exp | $ Salary | Error | Error^2 |
| 2 | 15 | 15-45.444=-30.44 | 926.84 |
| 3 | 28 | 28-45.444=-17.44 | 304.29 |
| 5 | 42 | 42-45.444=-3.44 | 11.86 |
| 13 | 64 | 64-45.444=18.56 | 344.33 |
| 8 | 50 | 50-45.444=4.56 | 20.76 |
| 16 | 90 | 90-45.444=44.56 | 1985.24 |
| 11 | 58 | 58-45.444=12.56 | 157.65 |
| 1 | 8 | 8-45.444=-37.44 | 1402.05 |
| 9 | 54 | 54-45.444=8.56 | 73.21 |
| | ȳ=45.444 | | SSE=5226.22 |

Do It Skills
do it. enjoy it.

# FIND THE BEST FIT OF LINE

The mean line gave us SSE *5226.22*

We have to find a line that will bring down the SSE value.

We need to find the best fit of line intercept, we need to apply a linear regression model to reduce the SSE value at minimum as possible. To get the best fit line we need to identify a slope intercept, we use the equation

$y=mx+b$

m is the slope
b is intercept
x → independent variables
y → dependent variables

| x | y | E=y-ȳ | Error Sqaure |
|---|---|---|---|
| Exp | $ Salary | Error | Error^2 |
| 2 | 15 | 15-45.444=-30.44 | 926.84 |
| 3 | 28 | 28-45.444=-17.44 | 304.29 |
| 5 | 42 | 42-45.444=-3.44 | 11.86 |
| 13 | 64 | 64-45.444=18.56 | 344.33 |
| 8 | 50 | 50-45.444=4.56 | 20.76 |
| 16 | 90 | 90-45.444=44.56 | 1985.24 |
| 11 | 58 | 58-45.444=12.56 | 157.65 |
| 1 | 8 | 8-45.444=-37.44 | 1402.05 |
| 9 | 54 | 54-45.444=8.56 | 73.21 |
| | ȳ=45.444 | | SSE=5226.22 |

Actual

| Height | Model 1 | | | Model 2 | | |
|---|---|---|---|---|---|---|
| | Predicted | Error | Absolute error | Predicted | Error | Square error |
| 165 | 166 | -1 | 1 | 163 | -2 | 4 |
| 168 | 168 | 0 | 0 | 168 | 0 | 0 |
| 172 | 171 | -1 | 1 | 170 | -2 | 4 |
| 167 | 168 | 1 | 1 | 169 | 4 | 4 |
| 169 | 170 | 1 | 1 | 171 | 4 | 4 |
| Total | | 0 | 4 | | 0 | 16 |

# 2 METHODS TO FIND THE BEST FIT OF LINE

**1. Ordinary Least Square method :**will work for both univariate dataset and multi-variate dataset.

Univariate dataset which is single independent variables and single dependent variables.
Multi-variate dataset contains a single independent variables set and multiple dependent variables sets

2. **Gradient Descent** machine learning algorithm is applied on Multi-variate datasets

Do It Skills
do it. enjoy it.

# ORDINARY LEAST SQUARES (OLS) METHOD

We will use Ordinary Least Squares method to find the best line intercept (b) slope (m)

y=mx+b

To use OLS method, we apply the below formula to find the equation

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

$$b = \bar{y} - m * \bar{x}$$

$x$ = independent variables

$\bar{x}$ = average of independent variables

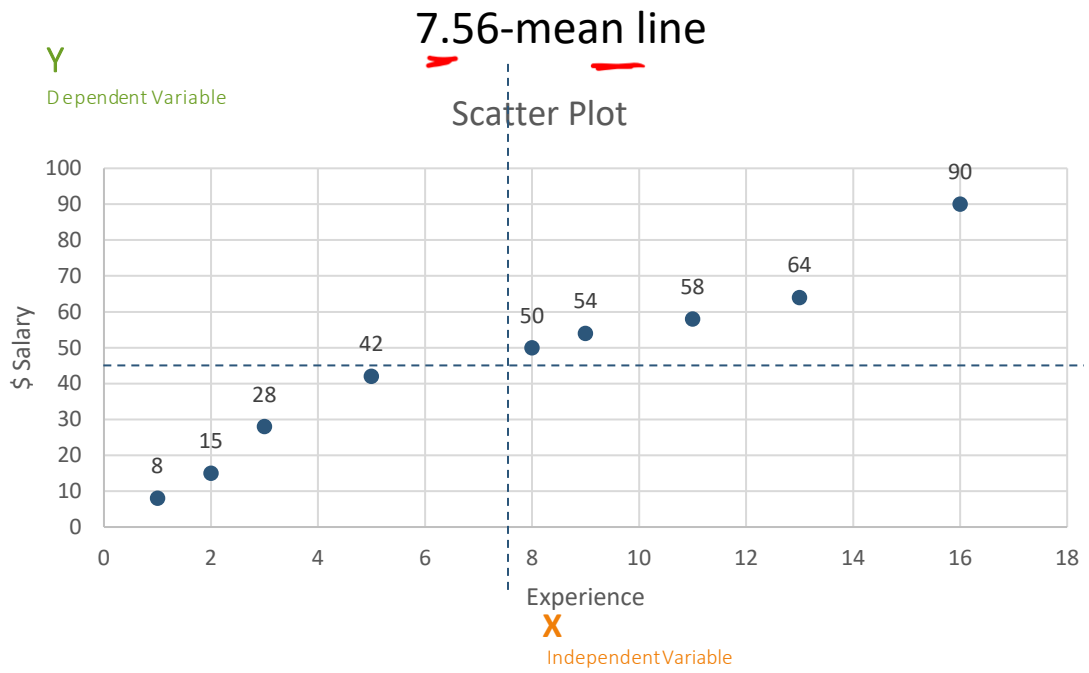
$y$ = dependent variables

$\bar{y}$ = average of dependent variables

| x | y |
|---|---|
| Exp | $ Salary |
| 2 | 15 |
| 3 | 28 |
| 5 | 42 |
| 13 | 64 |
| 8 | 50 |
| 16 | 90 |
| 11 | 58 |
| 1 | 8 |
| 9 | 54 |
| $\bar{x}$=7.56 | $\bar{y}$=45.444 |

Do It Skills
do it. enjoy it.

# MEAN OF THE DATA POINTS




45.4-mean line

| x | y |
|---|---|
| Exp | $ Salary |
| 2 | 15 |
| 3 | 28 |
| 5 | 42 |
| 13 | 64 |
| 8 | 50 |
| 16 | 90 |
| 11 | 58 |
| 1 | 8 |
| 9 | 54 |
| $\bar{x}$=7.56 | $\bar{y}$=45.444 |

We find the "Sum of Squared Errors" (SSE) for a Mean line

Do It Skills
do it. enjoy it.

# ORDINARY LEAST SQUARES (OLS) METHOD

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

$$b = \bar{y} - m * \bar{x}$$

$x$ = independent variables

$\bar{x}$ = average of independent variables

$y$ = dependent variables

$\bar{y}$ = average of dependent variables

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

Calculate Slope:

m = 1037.8 / 216.19

m = 4.80

| x | y | x-x̄ | y-ȳ | (x-x̄)(y-ȳ) | (x-x̄)^2 |
|---|---|---|---|---|---|
| Exp | $ Salary | | | | |
| 2 | 15 | 2-7.56=-5.56 | 15-45.444=-30.44 | 169.27 | 30.91 |
| 3 | 28 | 3-7.56=-4.56 | 28-45.444=-17.44 | 79.54 | 20.79 |
| 5 | 42 | 5-7.56=-2.56 | 42-45.444=-3.44 | 8.82 | 6.55 |
| 13 | 64 | 13-7.56=5.44 | 64-45.444=18.56 | 100.94 | 29.59 |
| 8 | 50 | 8-7.56=0.44 | 50-45.444=4.56 | 2.00 | 0.19 |
| 16 | 90 | 16-7.56=8.44 | 90-45.444=44.56 | 376.05 | 71.23 |
| 11 | 58 | 11-7.56=3.44 | 58-45.444=12.56 | 43.19 | 11.83 |
| 1 | 8 | 1-7.56=-6.56 | 8-45.444=-37.44 | 245.63 | 43.03 |
| 9 | 54 | 9-7.56=1.44 | 54-45.444=8.56 | 12.32 | 2.07 |
| x̄=7.56 | ȳ=45.444 | | | 1037.78 | 216.22 |

Do It Skills
do it. enjoy it.

# ORDINARY LEAST SQUARES (OLS) METHOD

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

$$b = \bar{y} - m * \bar{x}$$

$x$ = independent variables

$\bar{x}$ = average of independent variables

$y$ = dependent variables

$\bar{y}$ = average of dependent variables

| x | y | x-x̄ | y-ȳ | (x-x̄)(y-ȳ) | (x-x̄)^2 |
|---|---|---|---|---|---|
| Exp | $ Salary | | | | |
| 2 | 15 | 2-7.56=-5.56 | 15-45.444=-30.44 | 169.27 | 30.91 |
| 3 | 28 | 3-7.56=-4.56 | 28-45.444=-17.44 | 79.54 | 20.79 |
| 5 | 42 | 5-7.56=-2.56 | 42-45.444=-3.44 | 8.82 | 6.55 |
| 13 | 64 | 13-7.56=5.44 | 64-45.444=18.56 | 100.94 | 29.59 |
| 8 | 50 | 8-7.56=0.44 | 50-45.444=4.56 | 2.00 | 0.19 |
| 16 | 90 | 16-7.56=8.44 | 90-45.444=44.56 | 376.05 | 71.23 |
| 11 | 58 | 11-7.56=3.44 | 58-45.444=12.56 | 43.19 | 11.83 |
| 1 | 8 | 1-7.56=-6.56 | 8-45.444=-37.44 | 245.63 | 43.03 |
| 9 | 54 | 9-7.56=1.44 | 54-45.444=8.56 | 12.32 | 2.07 |
| x̄=7.56 | ȳ=45.444 | | | 1037.78 | 216.22 |

Calculate the intercept(b)

b = 45.44 - 4.80 * 7.56 = 9.15

b=9.15

Hence:

y = mx + b

→ 4.80x + 9.15

y = 4.80x + 9.15

model

Do It Skills

do it. enjoy it.

# BEST FIT LINE USING (OLS) METHOD

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

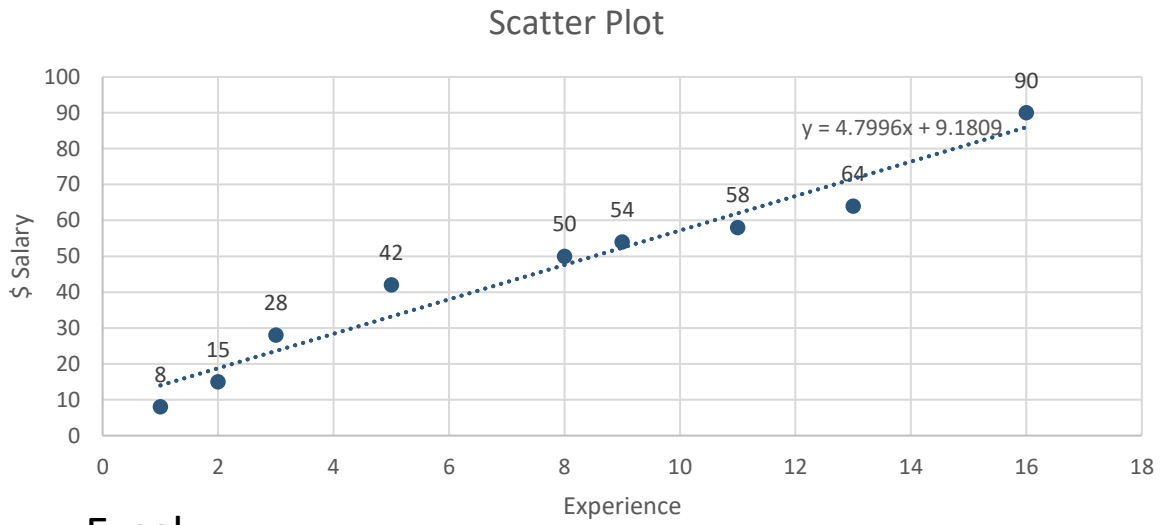
$$b = \bar{y} - m * \bar{x}$$

$x$ = independent variables

$\bar{x}$ = average of independent variables

$y$ = dependent variables

$\bar{y}$ = average of dependent variables




**OLS Method:**

m = 1037.8 / 216.19

m = 4.80

b = 45.44 - 4.80 * 7.56 = 9.15

Hence, y = mx + b → 4.80x + 9.15

y = 4.80x + 9.15

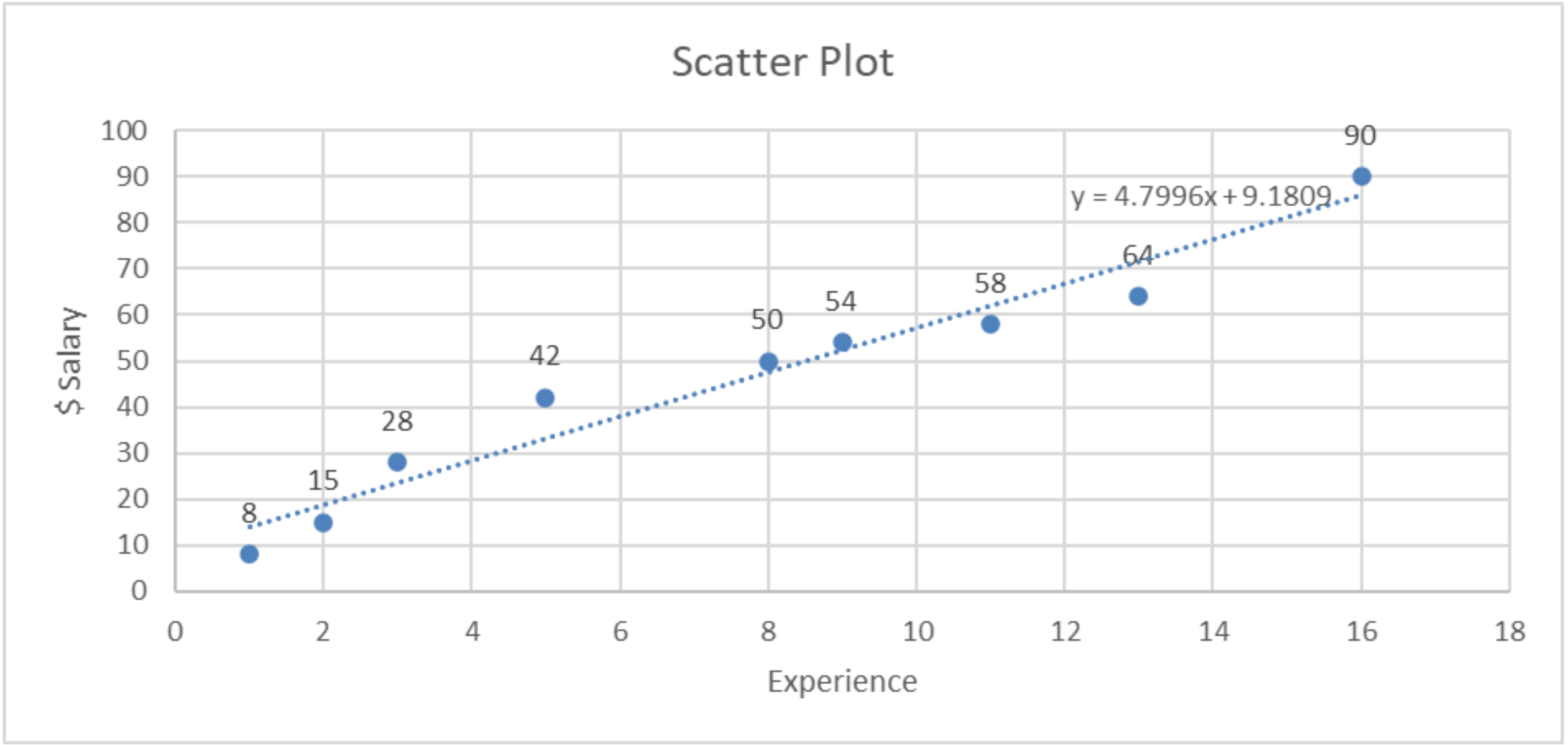
Excel:

y = 4.79x + 9.18

Do It Skills

do it. enjoy it.

# BEST FIT LINE USING (OLS) METHOD




Excel:
y = 4.79x + 9.18

**OLS Method:**
y = 4.80x + 9.15

Do It Skills
do it. enjoy it.

# SUM OF SQUARED ERRORS – PREDICTED OUTPUT

$$SSE = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$y_i$ = *Dependent Variables (Salary)*

$\bar{y}$ = *Average of Dependent Variables*

The sum of squared errors SSE output is *245.38* for the predicted line.

Now Sum of Squared Error got reduced significantly from 5226.19 to 245.38.

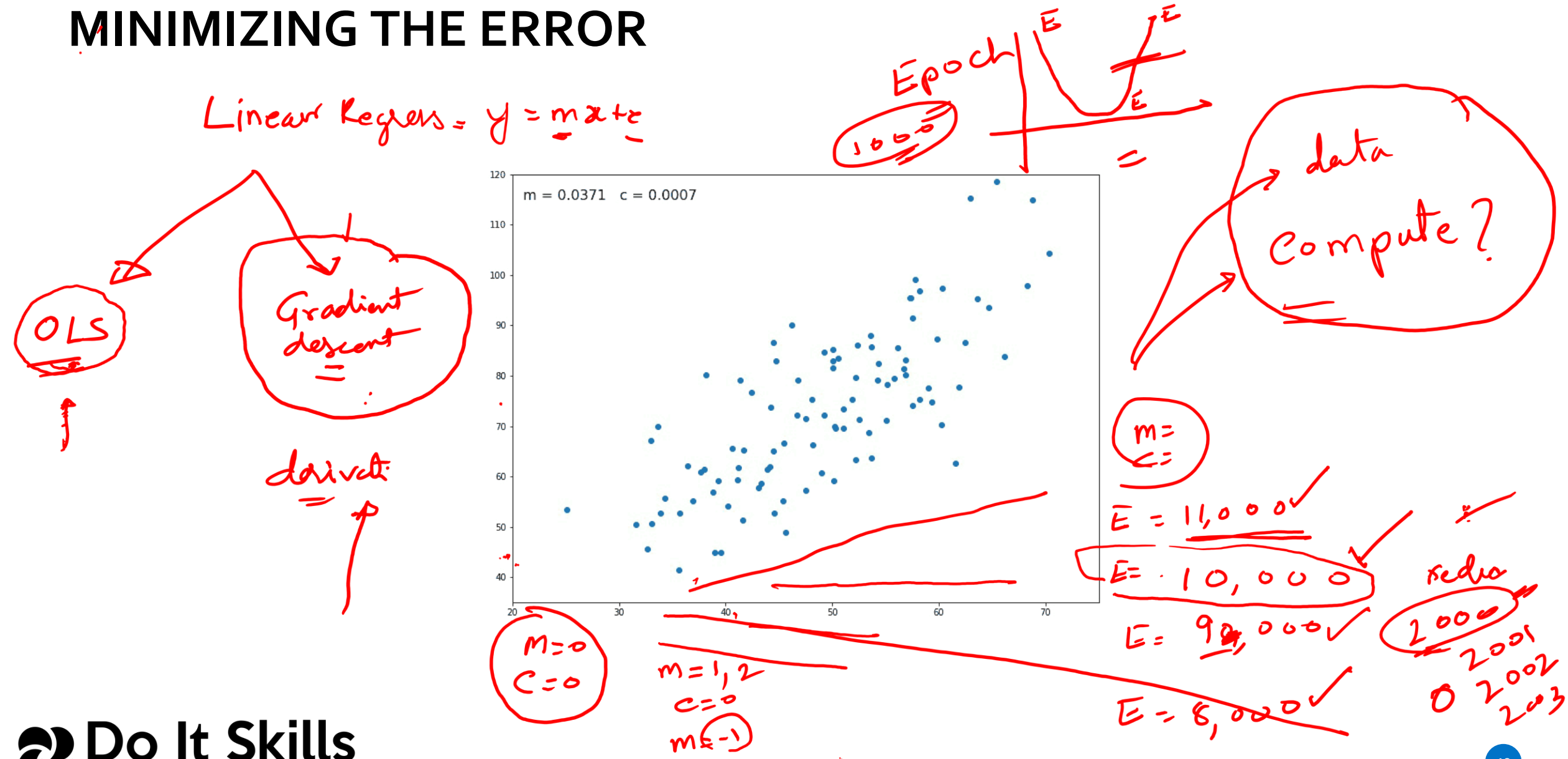
| | | ŷ=mx+b | | |
|---|---|---|---|---|
| x | y | ŷ=4.79x+9.18 | y-ŷ | (y-ŷ)^2 |
| Exp | $ Salary | Predicted (ŷ) | Error | Error^2 |
| 2 | 15 | 18.76 | -3.76 | 14.14 |
| 3 | 28 | 23.55 | 4.45 | 19.80 |
| 5 | 42 | 33.13 | 8.87 | 78.68 |
| 13 | 64 | 71.45 | -7.45 | 55.50 |
| 8 | 50 | 47.5 | 2.5 | 6.25 |
| 16 | 90 | 85.82 | 4.18 | 17.47 |
| 11 | 58 | 61.87 | -3.87 | 14.98 |
| 1 | 8 | 13.97 | -5.97 | 35.64 |
| 9 | 54 | 52.29 | 1.71 | 2.92 |
| x̄=7.56 | ȳ=45.444 | | | 245.38 |
| | | | | SSE |

Do It Skills
do it. enjoy it.

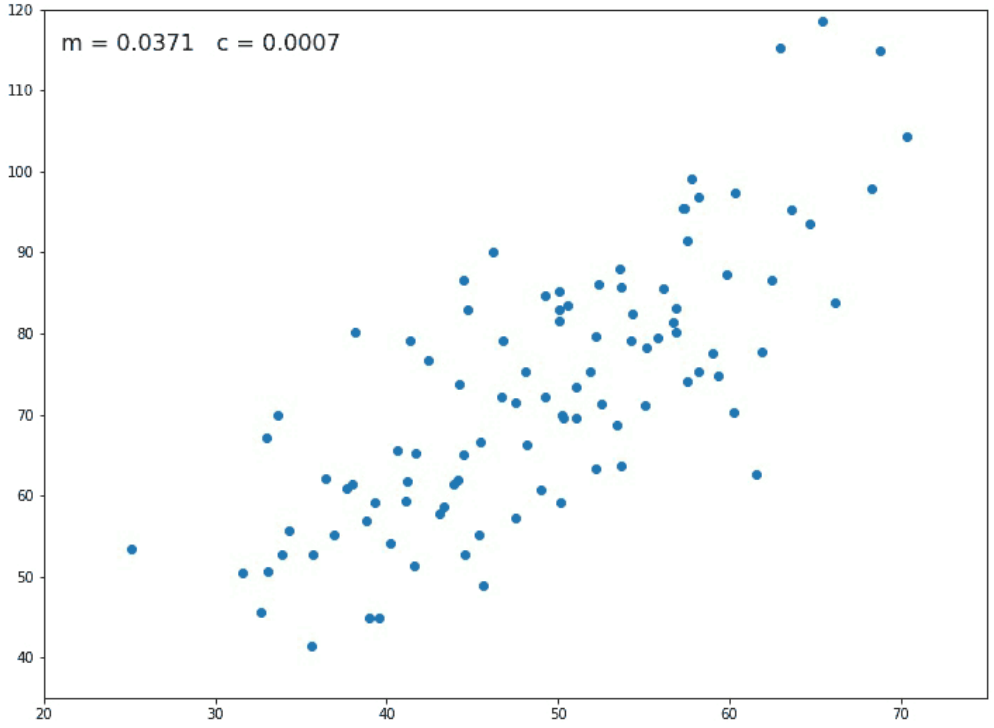
# GRADIENT DESCENT

**Gradient Descent** is an optimization algorithm used for minimizing the cost function in various machine learning algorithms. It is basically used for updating the parameters of the learning model.

# FINDING THE BEST FIT LINE: USING GRADIENT DESCENT, MINIMIZING THE ERROR

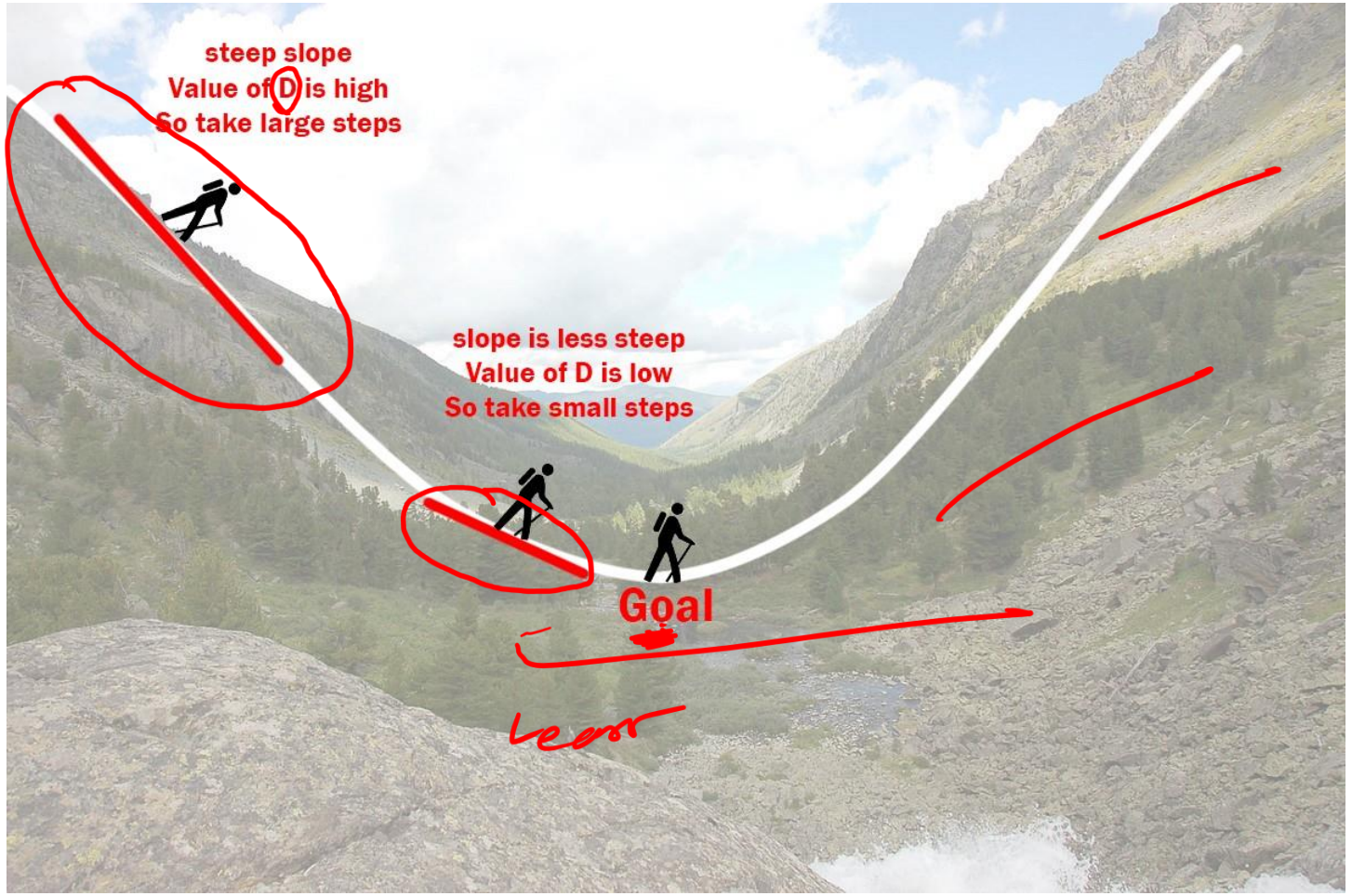


# FINDING THE BEST FIT LINE: USING GRADIENT DESCENT, MINIMIZING THE ERROR




Do It Skills
do it. enjoy it.

# HOW GRADIENT DESCENT WORKS




Do It Skills
do it. enjoy it.

# R SQUARE

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

**(SSres) Residual sum of squared errors** of our regression model

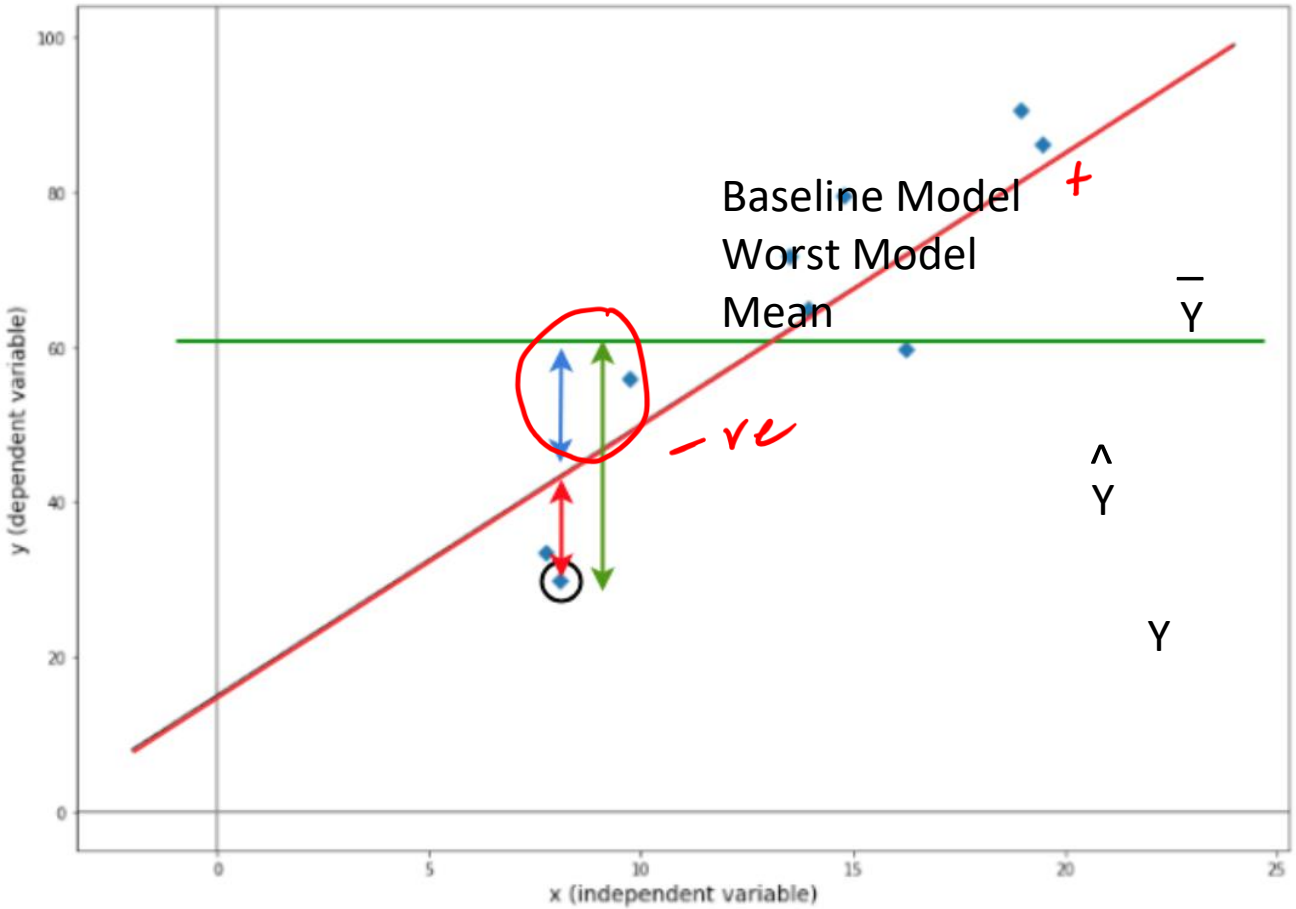
**actual y** value = 5 but we had **predicted y^** would be 6 then the **residual squared error** 1 and we would add that to the rest of the **residual squared errors** (SSres) for the model.

**(SStot) Total sum of squared errors** - This is comparing the **actual y** values to our **baseline model the mean $\overline{y}$.**

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

$$R^2 = \frac{\sum (\text{Predicted Distance} - \text{Mean})^2}{\sum (\text{Actual Distance} - \text{Mean})^2}$$

$$R^2 = \frac{\sum (y_p - \overline{y})^2}{\sum (y - \overline{y})^2}$$



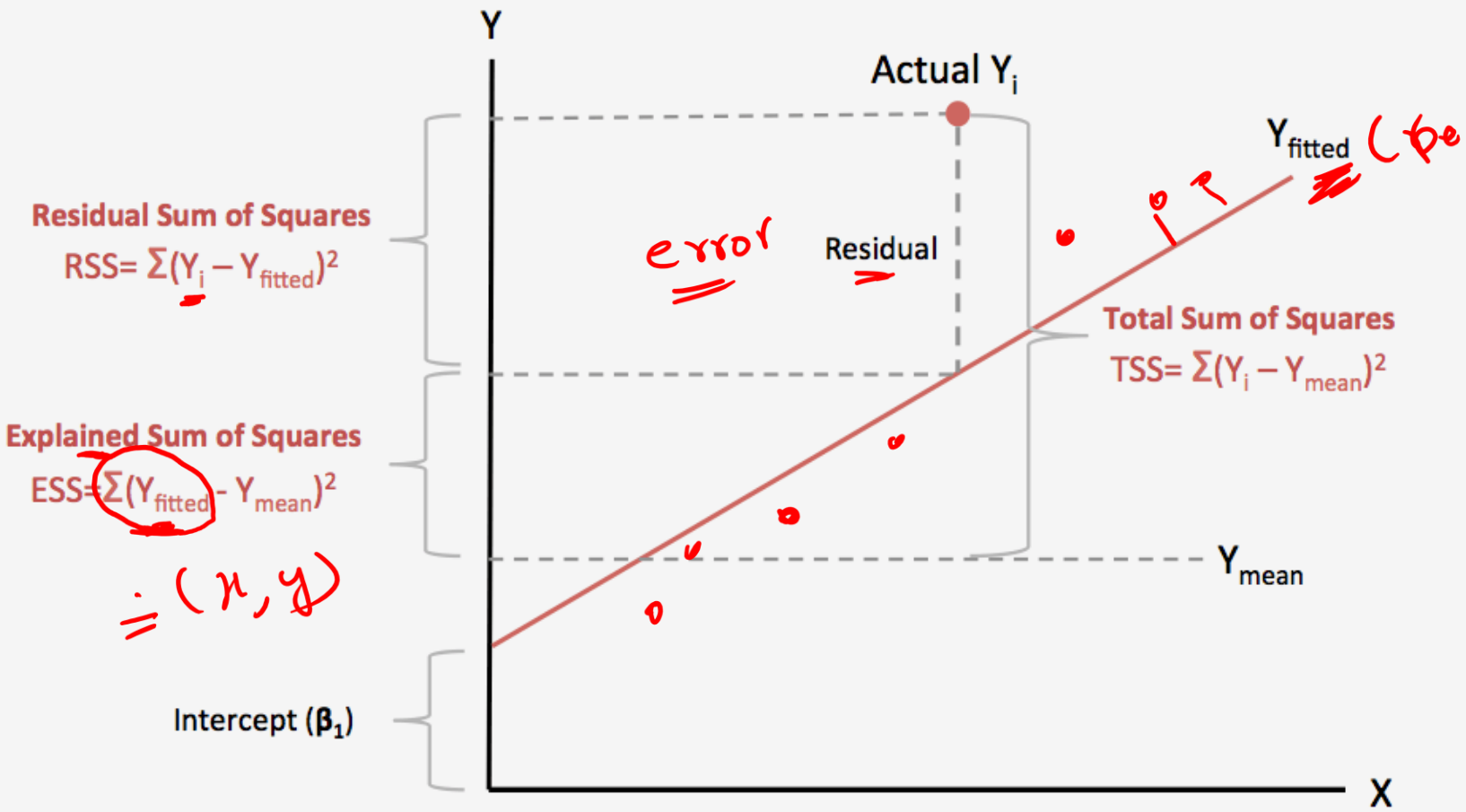

$\overline{Y}$ Mean

$\hat{Y}$ Predicted

Y Actual

Do It Skills
do it. enjoy it.

# R-Squared Explanation




$$R_{Sq} = 1 - \frac{RSS}{TSS}$$

# GOODNESS OF FIT – R SQUARE (HOW GOOD IS OUR MODEL)

| | | ŷ=mx+b |
|---|---|---|
| x | y | ŷ=4.79x+9.18 |
| Exp | $ Salary | Predicted (ŷ) |
| 2 | 15 | 18.76 |
| 3 | 28 | 23.55 |
| 5 | 42 | 33.13 |
| 13 | 64 | 71.45 |
| 8 | 50 | 47.5 |
| 16 | 90 | 85.82 |
| 11 | 58 | 61.87 |
| 1 | 8 | 13.97 |
| 9 | 54 | 52.29 |
| x̄=7.56 | ȳ=45.44 | |
| | | |

$$R^2 = \frac{\sum(\text{Predicted Distance} - \text{Mean})^2}{\sum\ (\text{Actual Distance} - \text{Mean})^2}$$

$$R^2 = \frac{\sum\ (y_p - \bar{y})^2}{\sum\ (y - \bar{y})^2}$$

Do It Skills
do it. enjoy it.

# GOODNESS OF FIT – R SQUARE (HOW GOOD IS OUR MODEL)

| | | ŷ=mx+b | | | | |
|---|---|---|---|---|---|---|
| x | y | ŷ=4.79x+9.18 | y-ȳ | (y-ȳ)^2 | ŷ-ȳ | (ŷ-ȳ)^2 |
| Exp | $ Salary | Predicted (ŷ) | | | | |
| 2 | 15.00 | 18.76 | 15-45.44=-30.44 | 926.84 | 18.76-45.44=-26.68 | 712.04 |
| 3 | 28.00 | 23.55 | 28-45.44=-17.44 | 304.29 | 23.55-45.44=-21.89 | 479.35 |
| 5 | 42.00 | 33.13 | 42-45.44=-3.44 | 11.86 | 33.13-45.44=-12.31 | 151.63 |
| 13 | 64.00 | 71.45 | 64-45.44=18.56 | 344.33 | 71.45-45.44=26.01 | 676.31 |
| 8 | 50.00 | 47.5 | 50-45.44=4.56 | 20.76 | 50.00-45.44=2.06 | 4.23 |
| 16 | 90.00 | 85.82 | 90-45.44=44.56 | 1985.24 | 90.00-45.44=40.38 | 1630.22 |
| 11 | 58.00 | 61.87 | 58-45.44=12.56 | 157.65 | 61.87-45.44=16.43 | 269.81 |
| 1 | 8.00 | 13.97 | 8-45.44=-37.44 | 1402.05 | 13.97-45.44=-31.47 | 990.61 |
| 9 | 54.00 | 52.29 | 54-45.44=8.56 | 73.21 | 52.29-45.44=6.65 | 46.87 |
| x̄=7.56 | ȳ=45.44 | | | 5226.22 | | 4961.07 |

Predicted (ŷ-ȳ)^2 4961.07
Actual (y-ȳ)^2 5226.22

# GOODNESS OF FIT – R SQUARE (HOW GOOD IS OUR MODEL)

| | | ŷ=mx+b | | | | |
|---|---|---|---|---|---|---|
| x | y | ŷ=4.79x+9.18 | y-ȳ | (y-ȳ)^2 | ŷ-ȳ | (ŷ-ȳ)^2 |
| Exp | $ Salary | Predicted (ŷ) | | | | |
| 2 | 15.00 | 18.76 | 15-45.44=-30.44 | 926.84 | 18.76-45.44=-26.68 | 712.04 |
| 3 | 28.00 | 23.55 | 28-45.44=-17.44 | 304.29 | 23.55-45.44=-21.89 | 479.35 |
| 5 | 42.00 | 33.13 | 42-45.44=-3.44 | 11.86 | 33.13-45.44=-12.31 | 151.63 |
| 13 | 64.00 | 71.45 | 64-45.44=18.56 | 344.33 | 71.45-45.44=26.01 | 676.31 |
| 8 | 50.00 | 47.5 | 50-45.44=4.56 | 20.76 | 50.00-45.44=2.06 | 4.23 |
| 16 | 90.00 | 85.82 | 90-45.44=44.56 | 1985.24 | 90.00-45.44=40.38 | 1630.22 |
| 11 | 58.00 | 61.87 | 58-45.44=12.56 | 157.65 | 61.87-45.44=16.43 | 269.81 |
| 1 | 8.00 | 13.97 | 8-45.44=-37.44 | 1402.05 | 13.97-45.44=-31.47 | 990.61 |
| 9 | 54.00 | 52.29 | 54-45.44=8.56 | 73.21 | 52.29-45.44=6.65 | 46.87 |
| x̄=7.56 | ȳ=45.44 | | | 5226.22 | | 4961.07 |

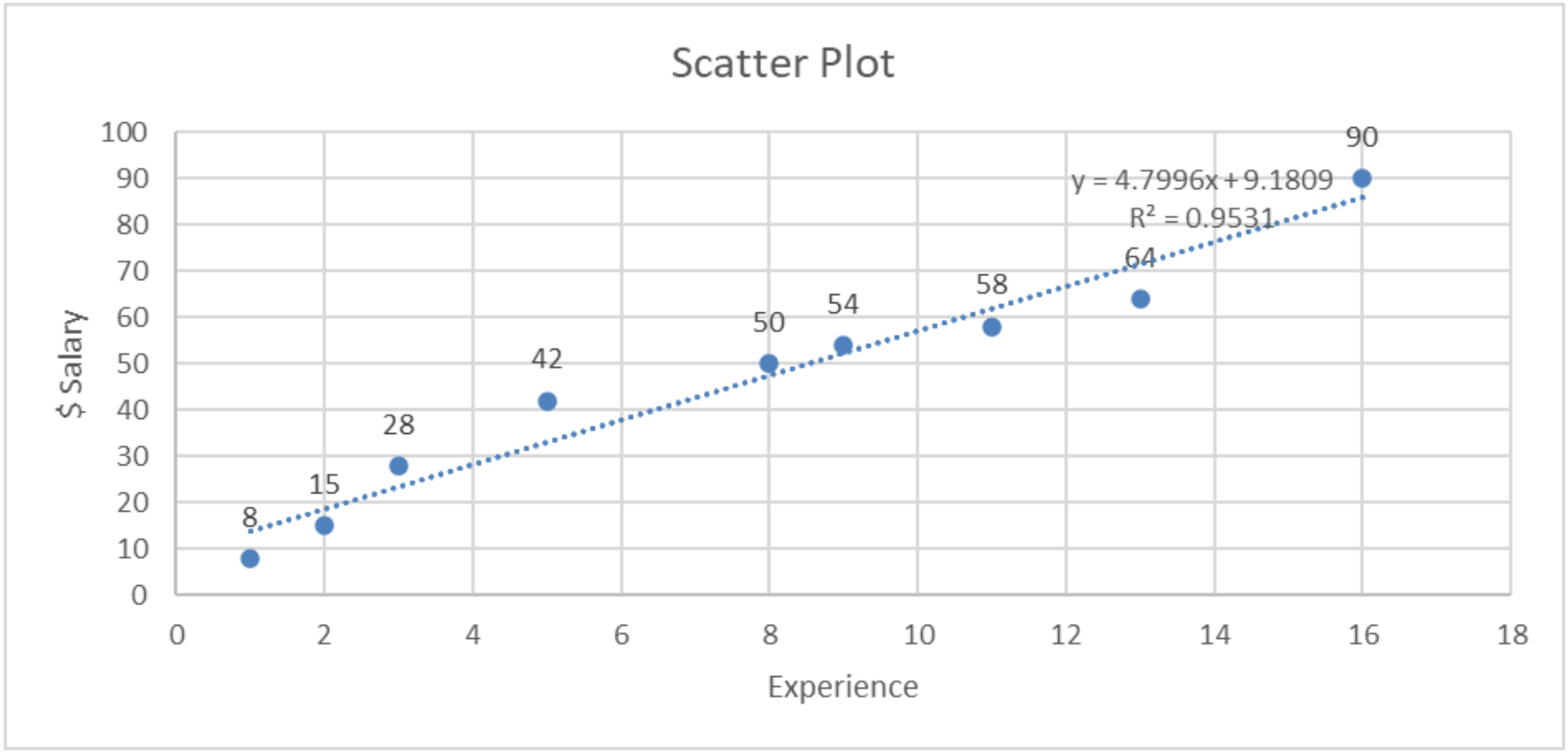
Predicted (ŷ-ȳ)^2 4961.07
Actual (y-ȳ)^2 5226.22
R Square 0.949265435

$$R^2 = \frac{\sum(\text{Predicted Distance} - \text{Mean})^2}{\sum(\text{Actual Distance} - \text{Mean})^2}$$

$$R^2 = \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$$

Do It Skills
do it. enjoy it.

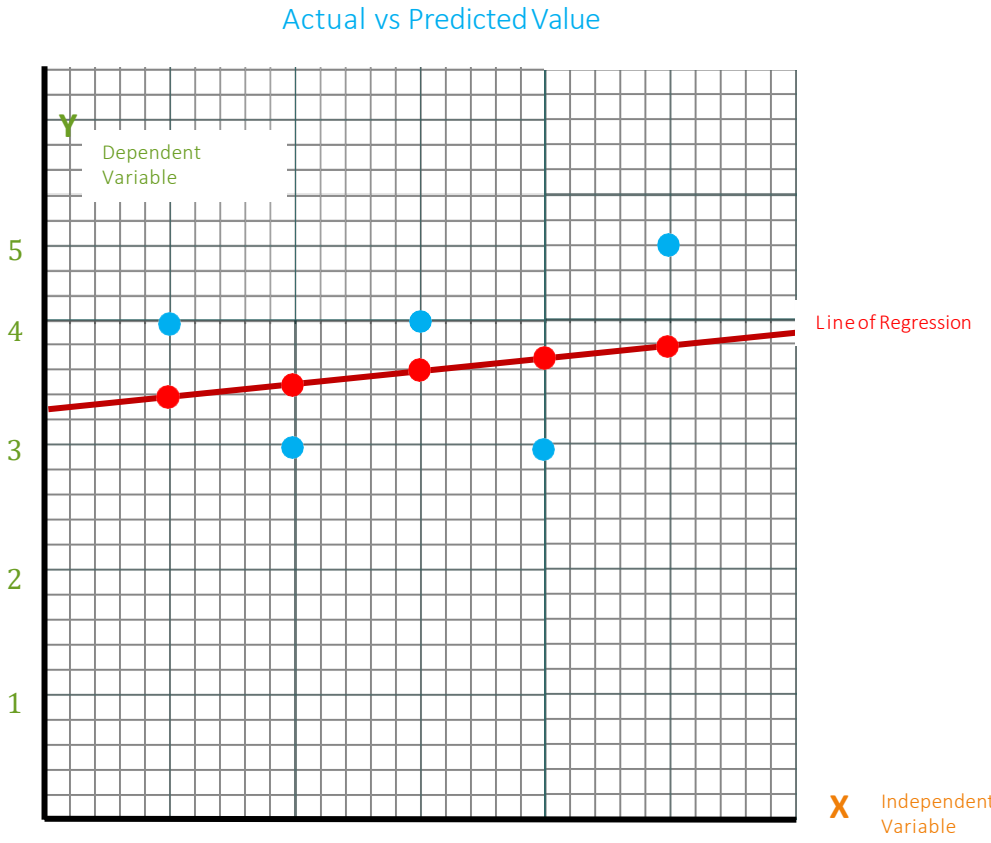
# GOODNESS OF FIT – R SQUARE (HOW GOOD IS OUR MODEL)




| | | |
|---|---|---|
| Predicted | (ŷ-ȳ)^2 | 4961.07 |
| Actual | (y-ȳ)^2 | 5226.22 |
| | R Square | 0.949265435 |

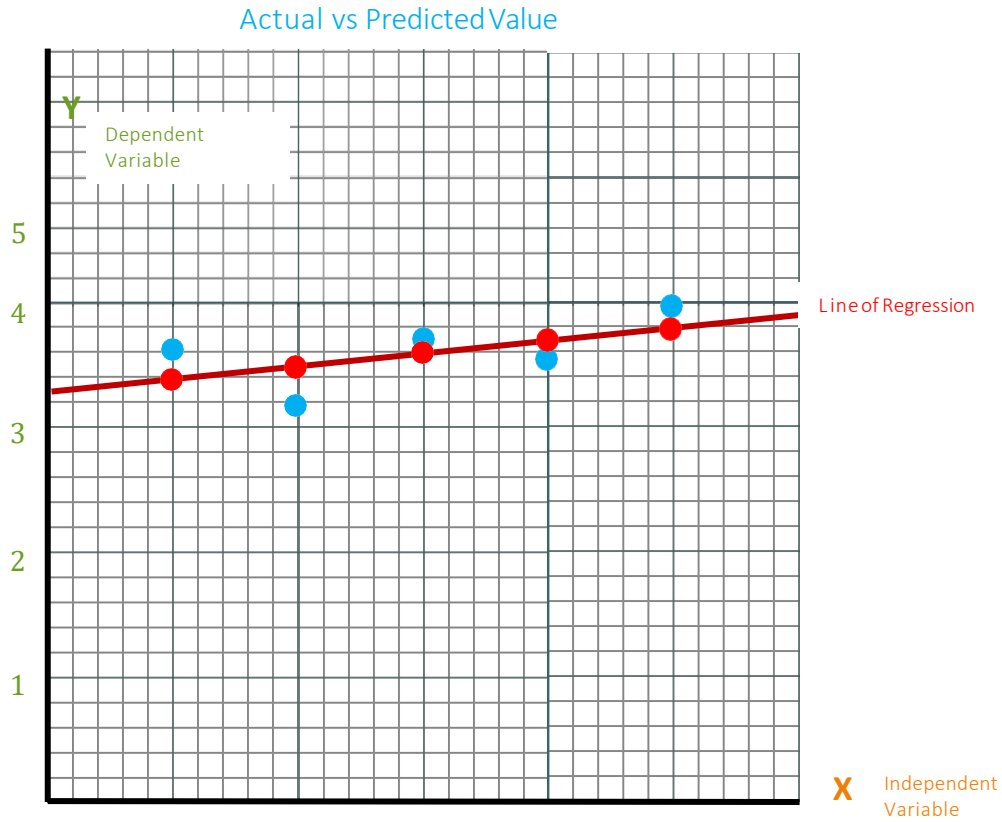
Do It Skills
do it. enjoy it.

# EXAMPLES OF BETTER REGRESSION – BETTER R SQUARE



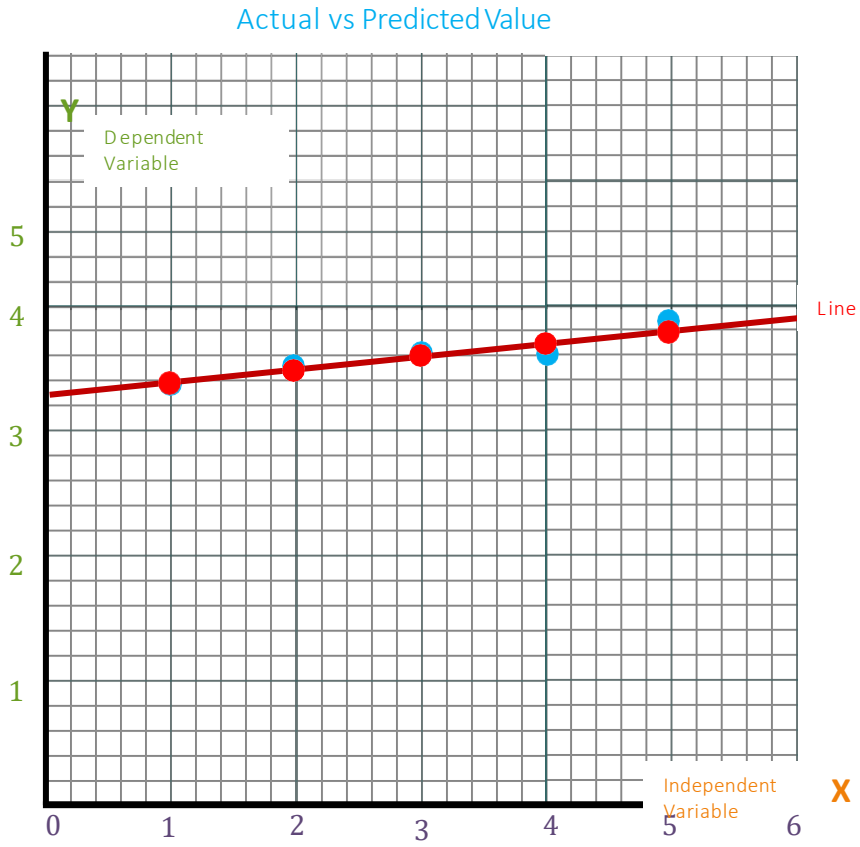

$$R^2 \approx .19$$

# EXAMPLES OF BETTER REGRESSION – BETTER R SQUARE




$$R^2 \approx 0.9$$

# EXAMPLES OF BETTER REGRESSION – BETTER R SQUARE



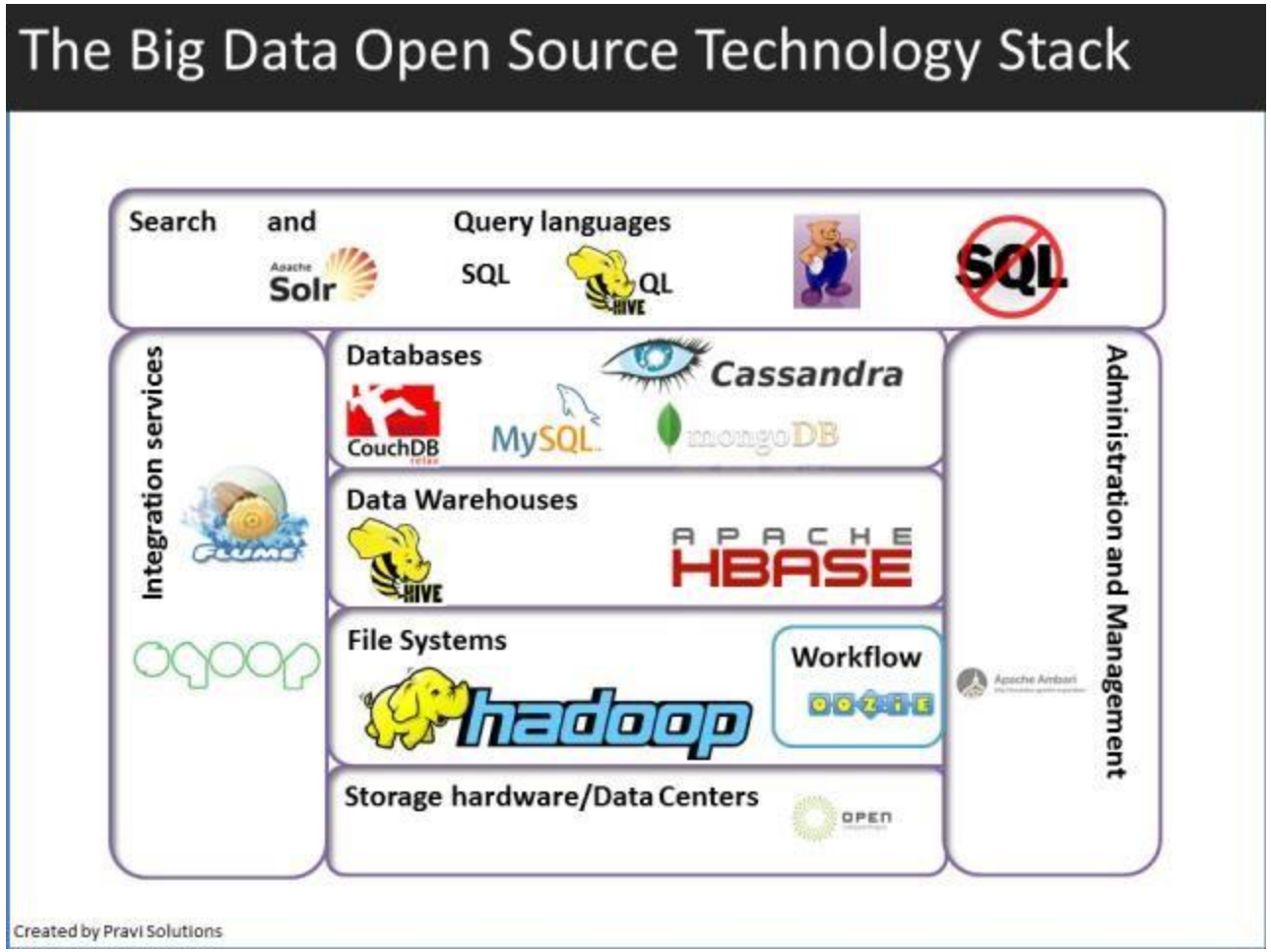

$$R^2 \approx 1$$

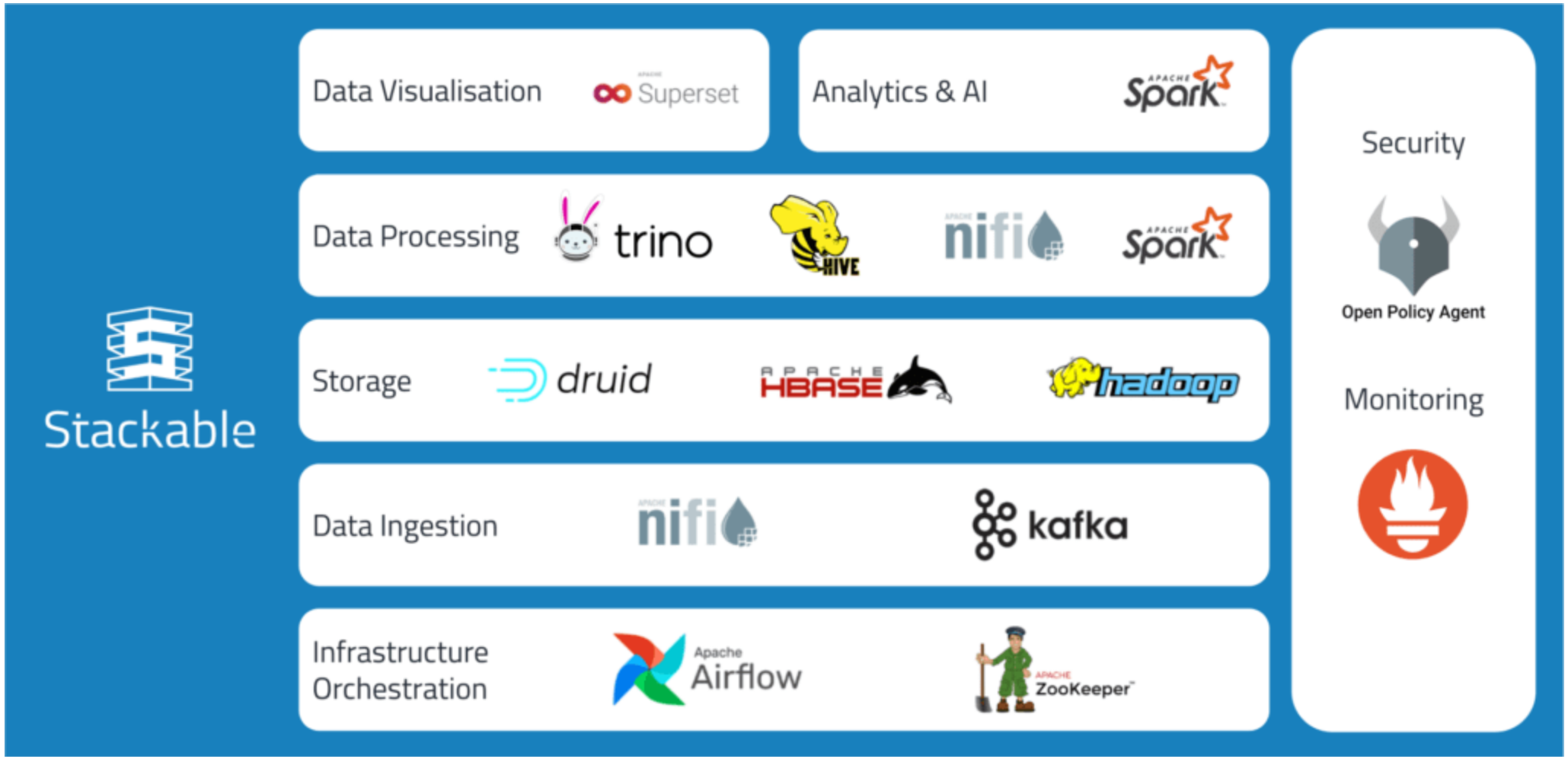
Do It Skills
do it. enjoy it.

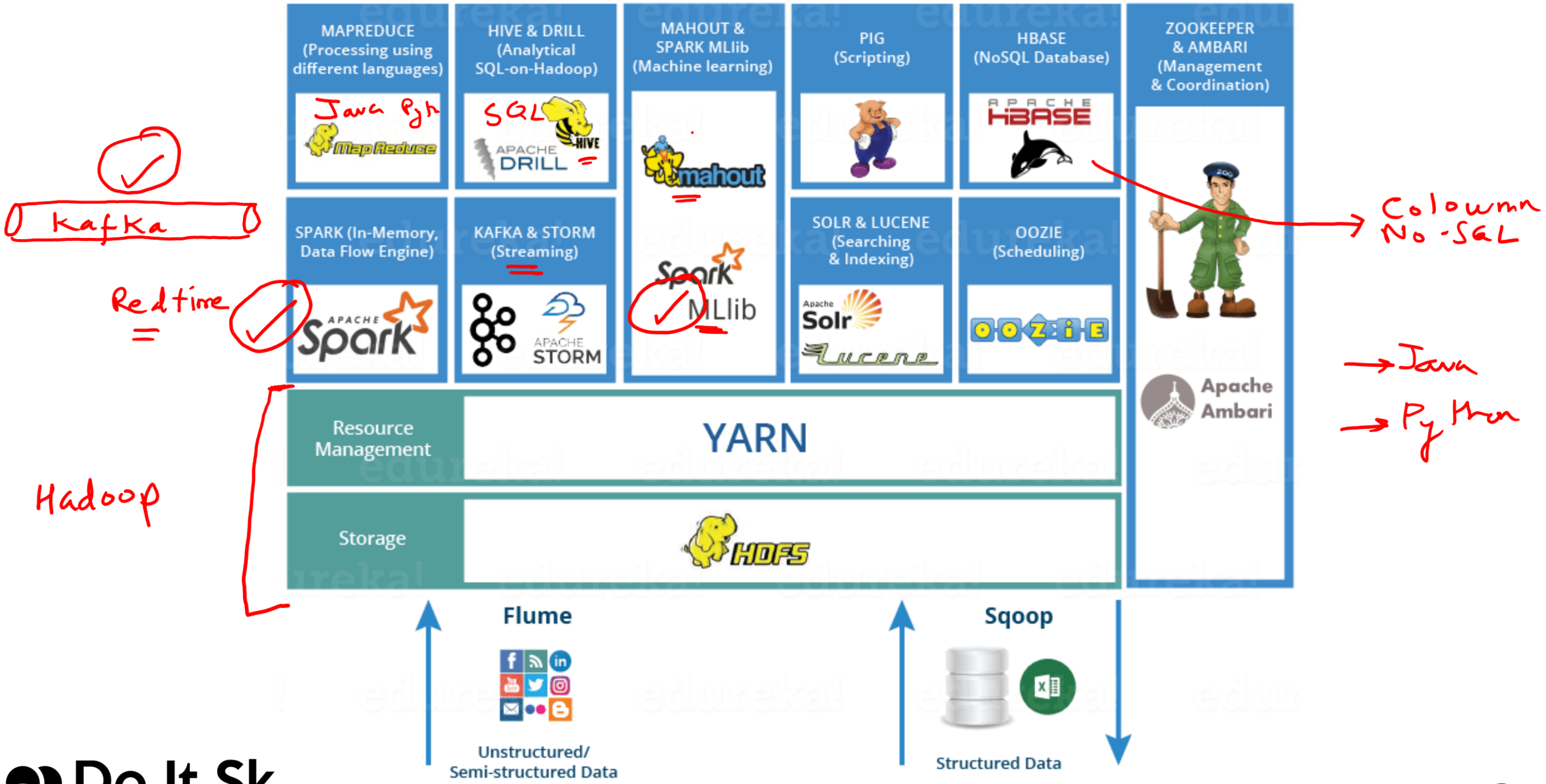
Do It Skills
do it. enjoy it.

# 9 PM EST, TUESDAY
# 12 NOVEMBER 2024

## BIG DATA, HADOOP ECOSYSTEM, ANALYTICS, BOSTON HOUSING PRICE


The Big Data Open Source Technology Stack
Search and
Apache Solr
Query languages
SQL
QL
SQL
Integration services
Databases
Cassandra
CouchDB
MySQL
mongoDB
Data Warehouses
HIVE
APACHE HBASE
File Systems
hadoop
Workflow
OOZIE
Storage hardware/Data Centers
OPEN
Administration and Management
Created by Pravi Solutions


Stackable
Data Visualisation
Superset
Analytics & AI
Spark
Data Processing
trino
HIVE
nifi
Spark
Storage
druid
HBASE
hadoop
Data Ingestion
nifi
kafka
Infrastructure Orchestration
Apache Airflow
ZooKeeper
Security
Open Policy Agent
Monitoring

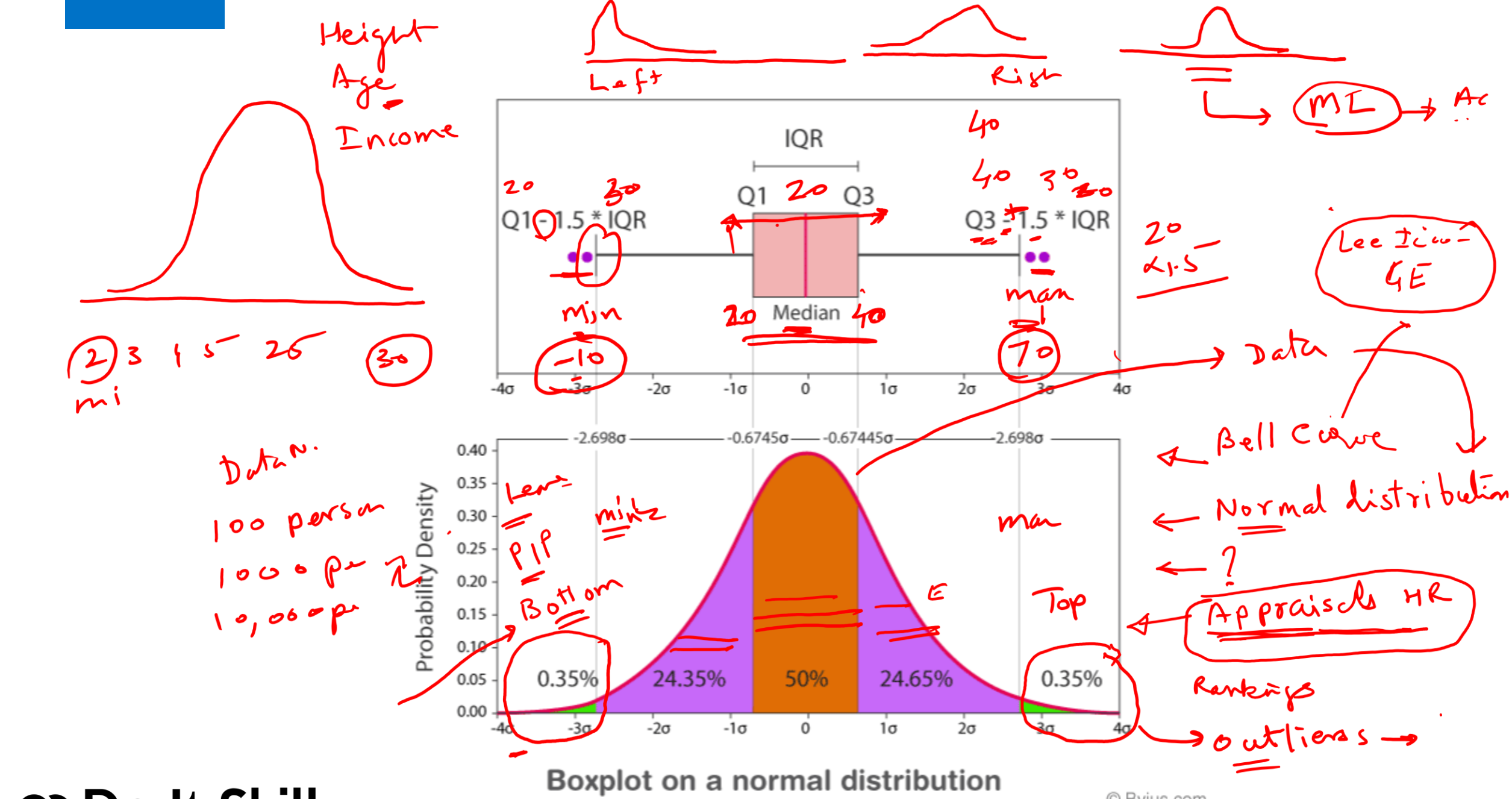

MAPREDUCE (Processing using different languages)
HIVE & DRILL (Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib (Machine learning)
PIG (Scripting)
HBASE (NoSQL Database)
ZOOKEEPER & AMBARI (Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM (Streaming)
SOLR & LUCENE (Searching & Indexing)
OOZIE (Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data
Apache Ambari
Kafka
Realtime
Hadoop
Column No-SQL
Java
Python

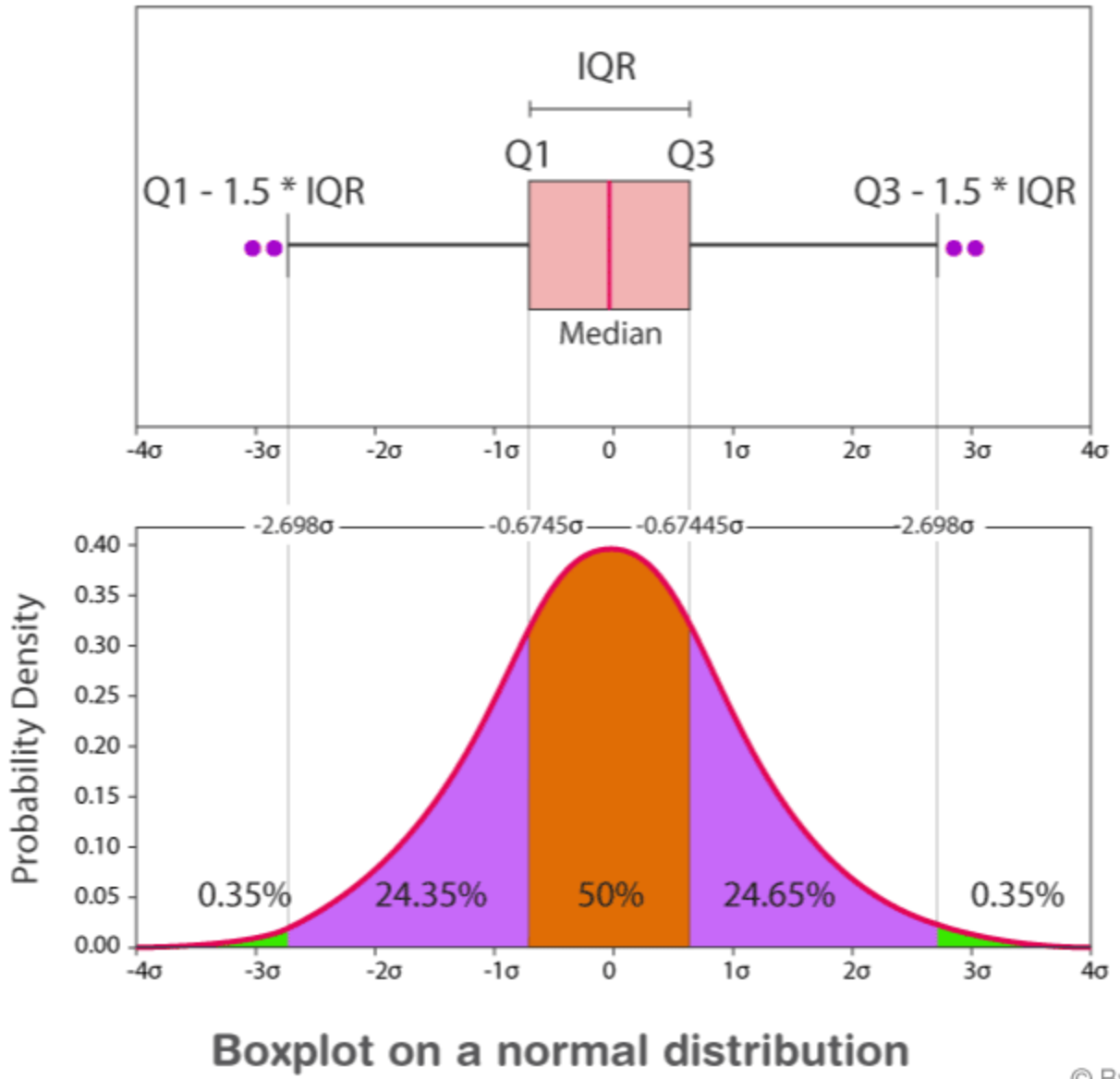
Do It Skills
do it. enjoy it.

# 9 PM EST, WEDNESDAY
# 13 NOVEMBER 2024

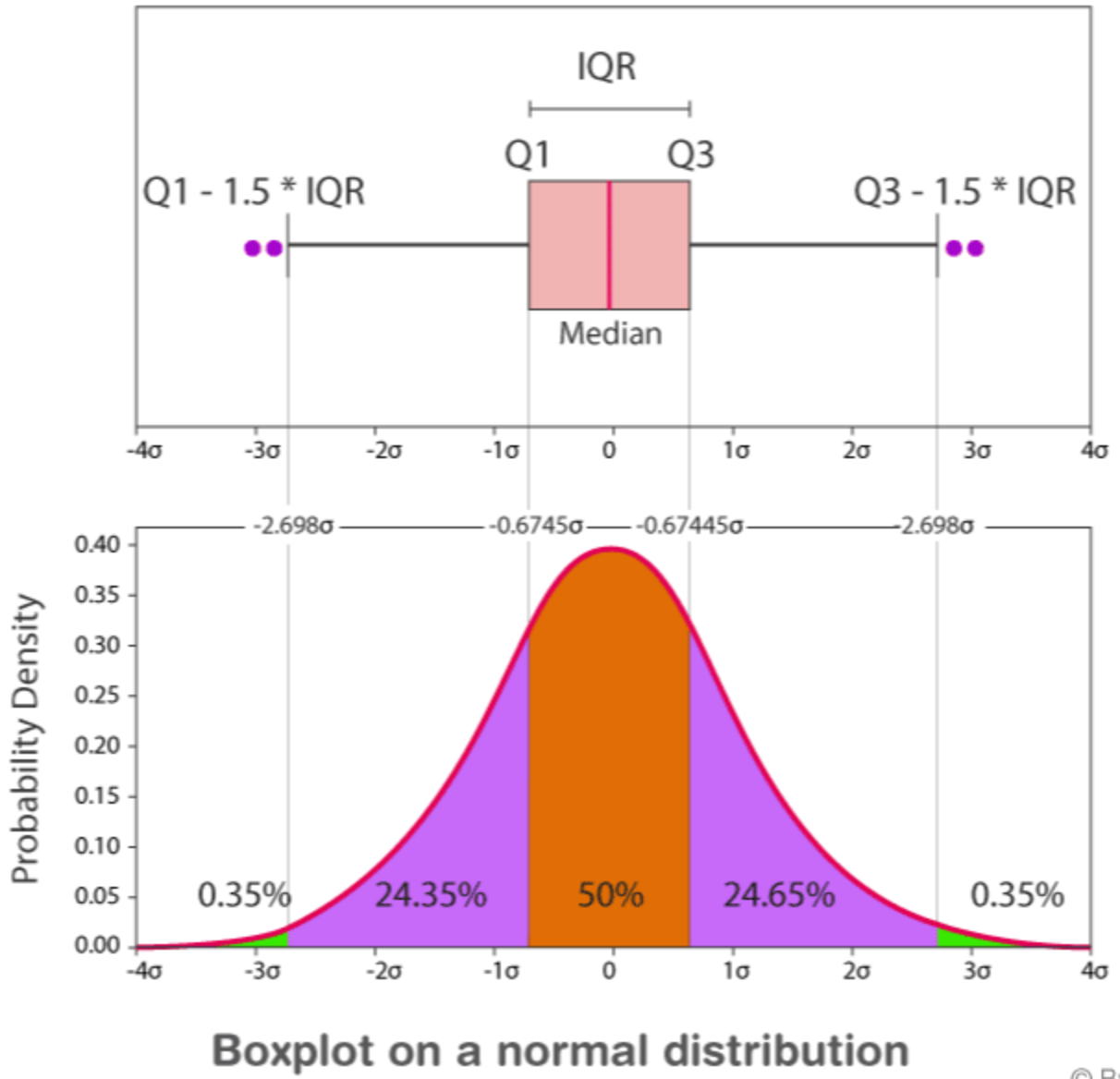
## BOSTON HOUSING PRICE

Outliers
Interquartile range ( IQR )
Outliers
" Minimum "
( Q1 - 1.5 * IQR )
Q1
( 25th percentile )
Median
Q3
( 75th percentile )
" Maximum "
( Q3 - 1.5 * IQR )
-4
-3
-2
-1
0
1
2
3
4
Different parts of boxplot
© Byjus.com


Height
Age
Income
Left
Right
ML
IQR
Q1
Q3
Q1 - 1.5 * IQR
Q3 + 1.5 * IQR
Median
min
max
-10
70
20
40
Data
Bell Curve
Normal distribution
Lee Jicu-
GE
Appraisals HR
Rankings
outliers
Top
Bottom
-2.698σ
-0.6745σ
-0.67445σ
2.698σ
Probability Density
0.35%
24.35%
50%
24.65%
0.35%
Boxplot on a normal distribution
© Byjus.com




Boxplot on a normal distribution




Boxplot on a normal distribution

Do It Skills
do it. enjoy it.

# 9 PM EST, THURSDAY
# 14 NOVEMBER 2024

# ASSIGNMENT FOR 14 NOVEMBER:
# 1)RUN THE BOSTON HOUSING PRICING AND EXPLAIN
# 2)HOW TO IMPROVE THE MODEL