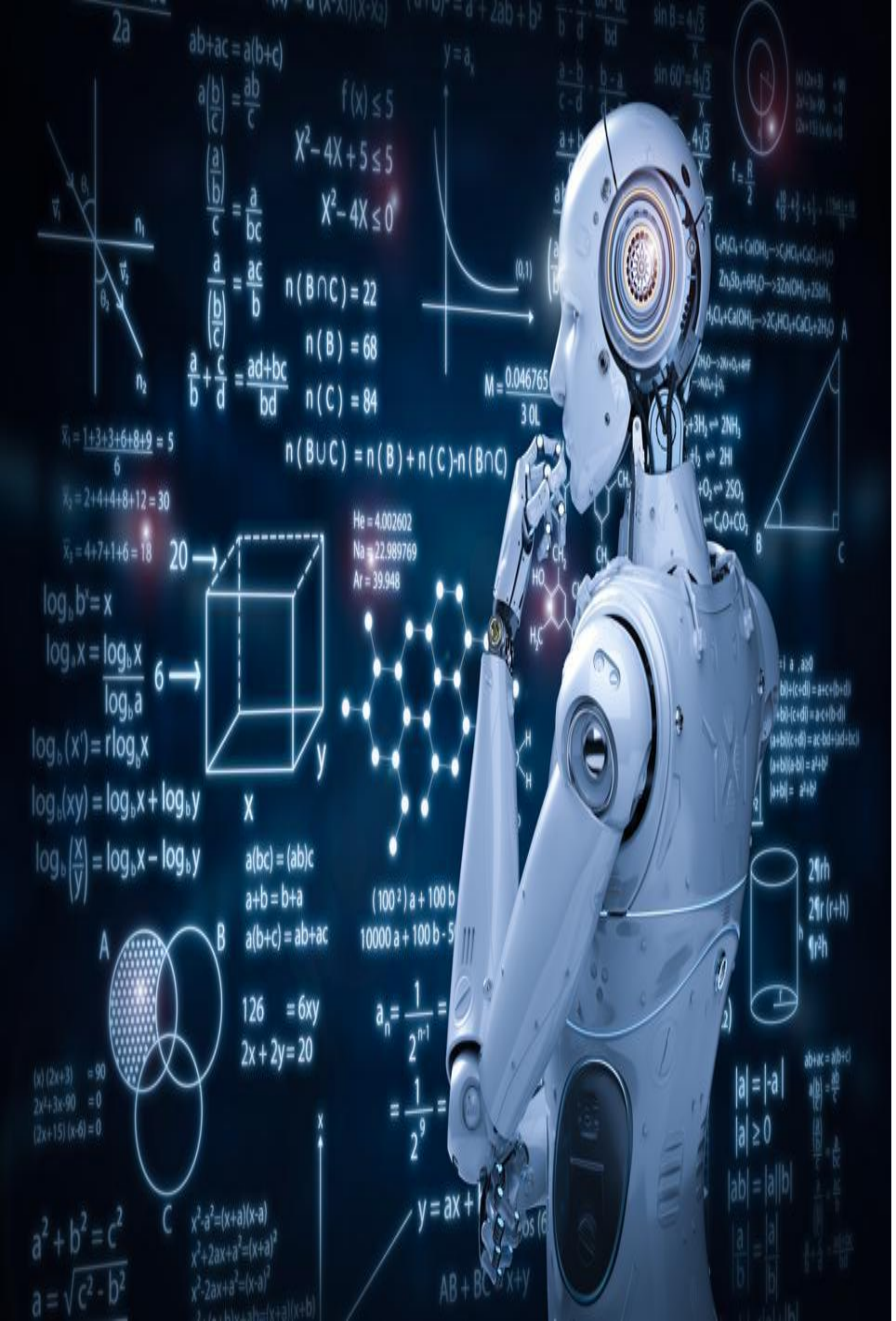


Practical Machine Learning

Day 5: Mar22 DBDA

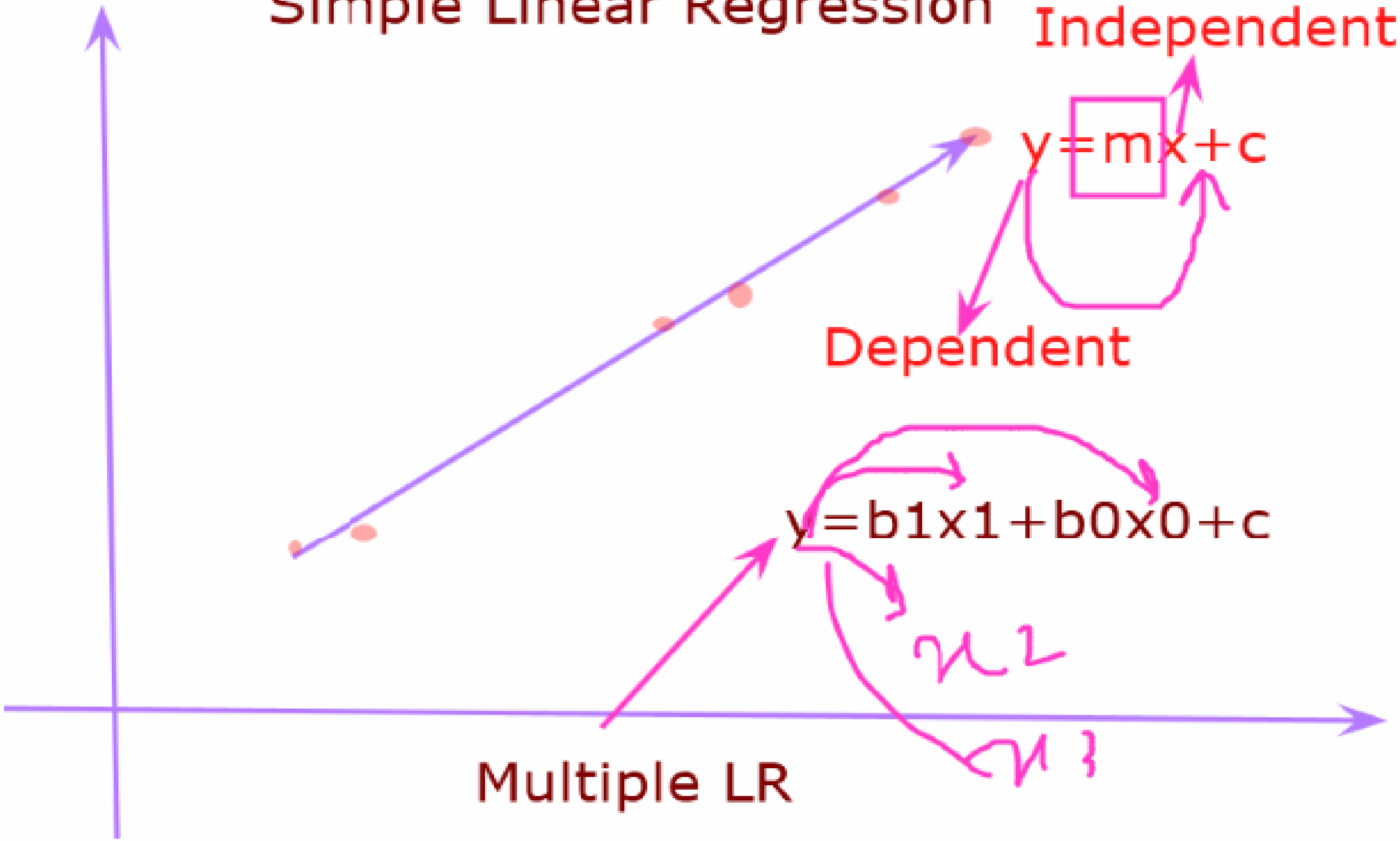
Kiran Waghmare

Agenda

- Regression
- Types of Regression

Who can see what you share here? Recording On

Simple Linear Regression



Independent

$y=m\boxed{x}+c$

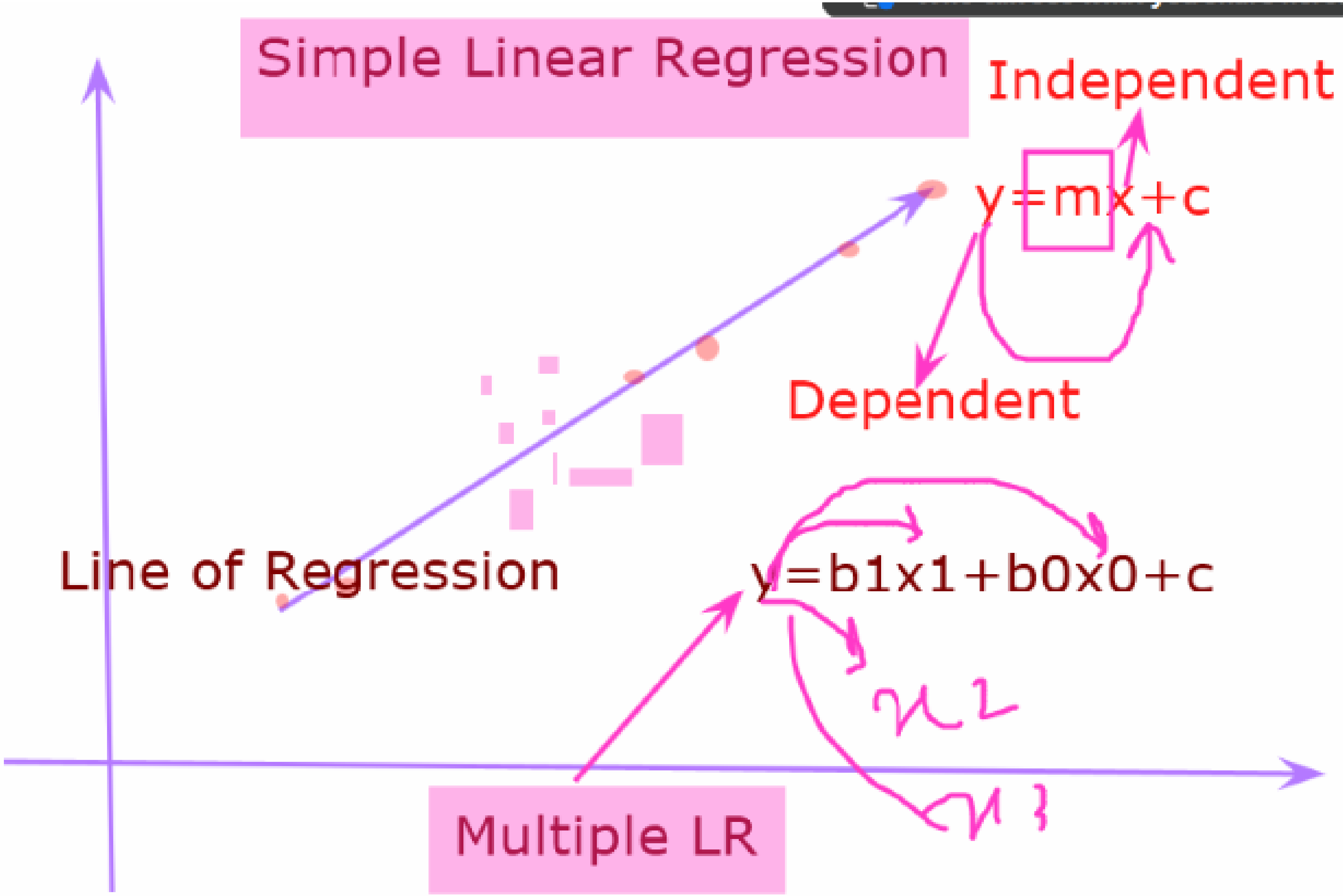
Dependent

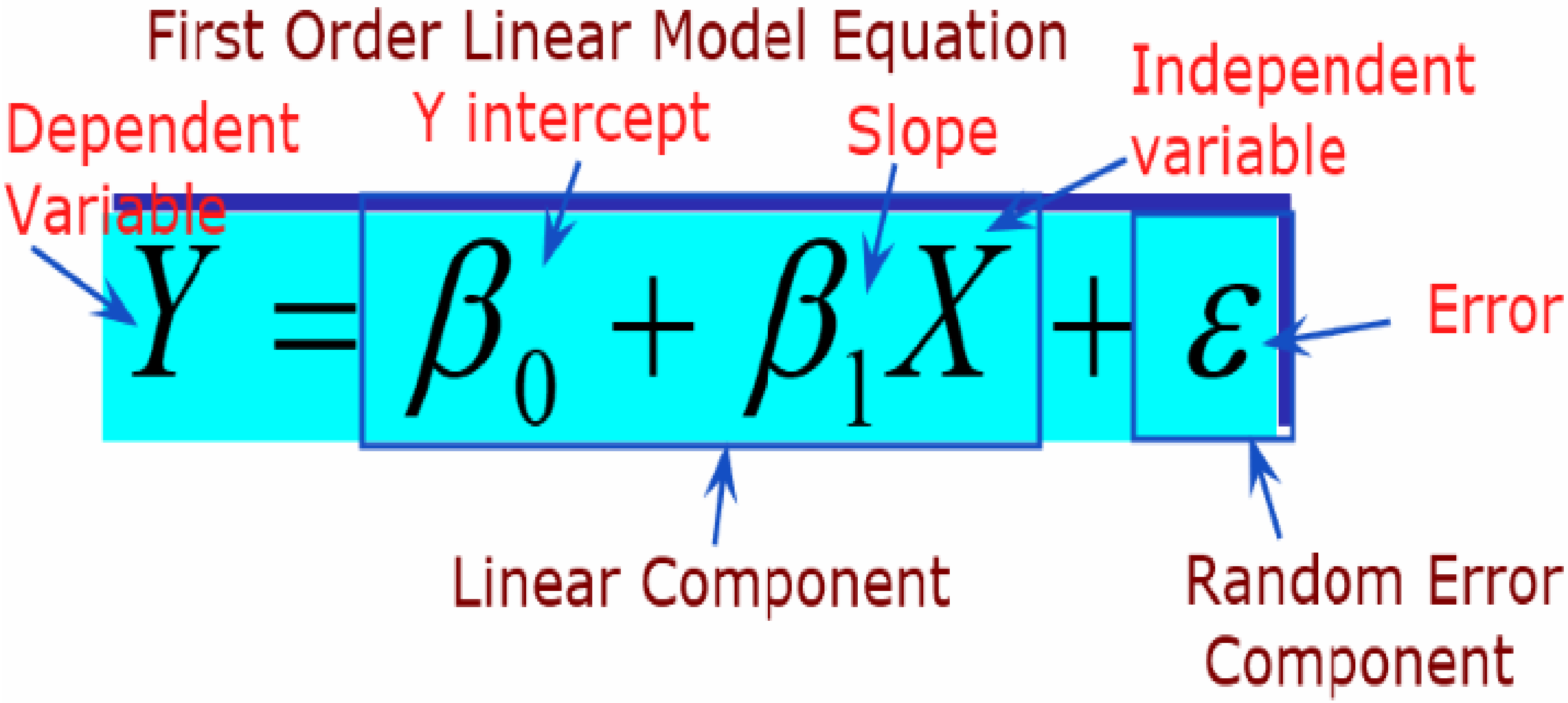
$y=b1x1+b0x0+c$

x2

x3

Multiple LR


Simple Linear Regression
Independent
y=mx+c
Dependent
Line of Regression
y=b1x1+b0x0+c
x2
x3
Multiple LR


First Order Linear Model Equation
Dependent Variable
Y intercept
Slope
Independent variable
$Y = \beta_0 + \beta_1 X + \varepsilon$
Error
Linear Component
Random Error Component

Who can see what you share here! Recording On

First Order Linear Model Equation

Dependent Variable

Y intercept

Slope

Independent variable

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Error

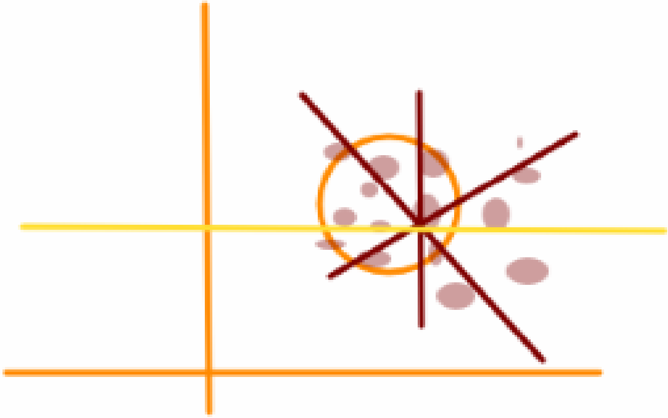
Linear Component

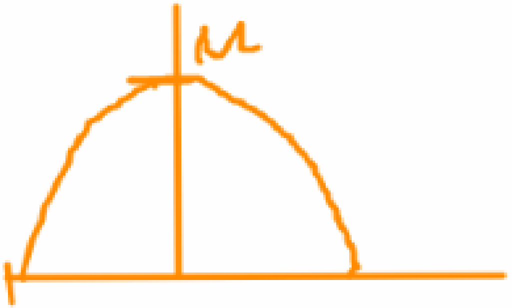
Random Error Component

Autocorrelation

r>0: positive
r<0:negative
r=0:no linear relationship

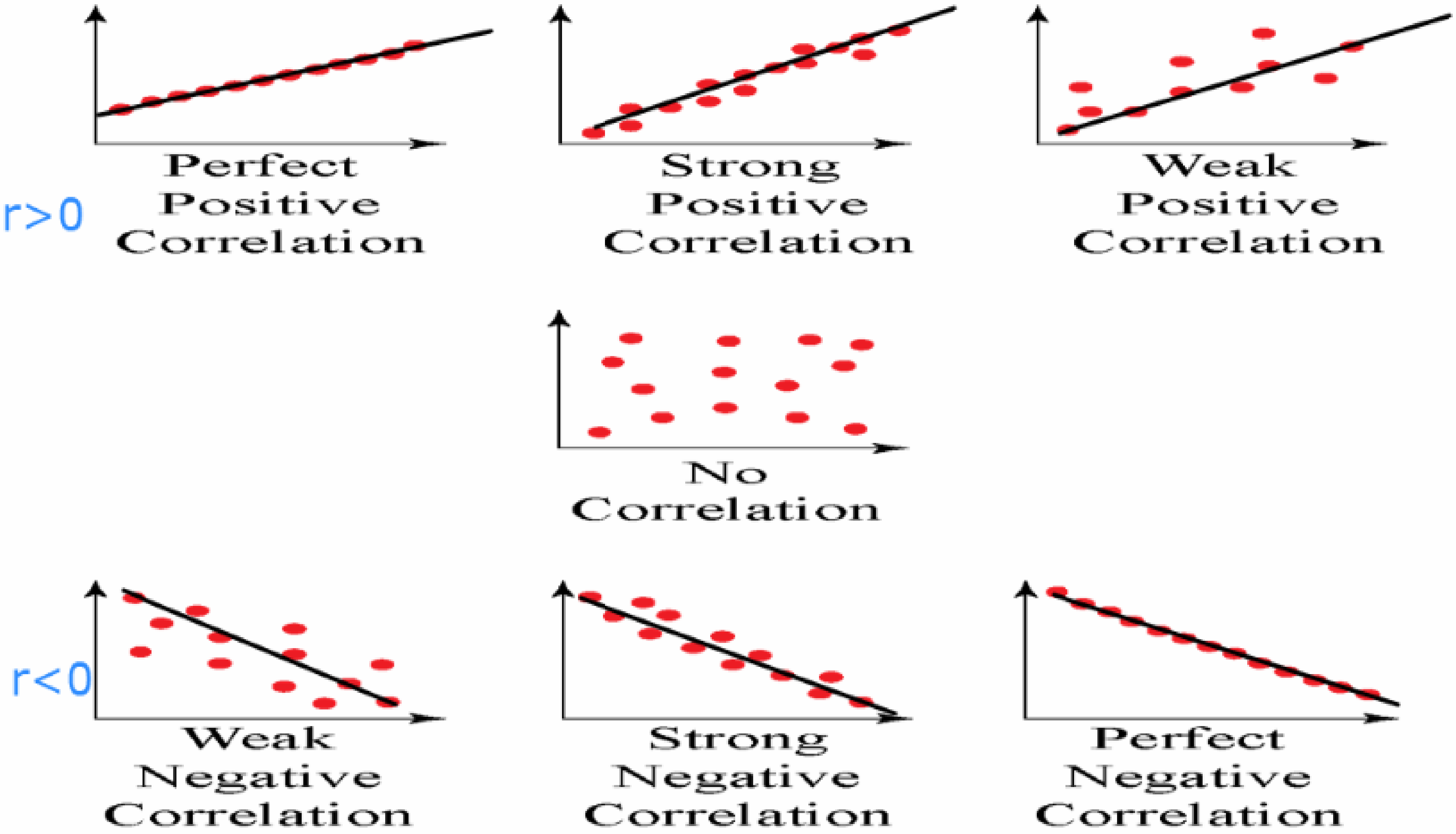
-1 to 1





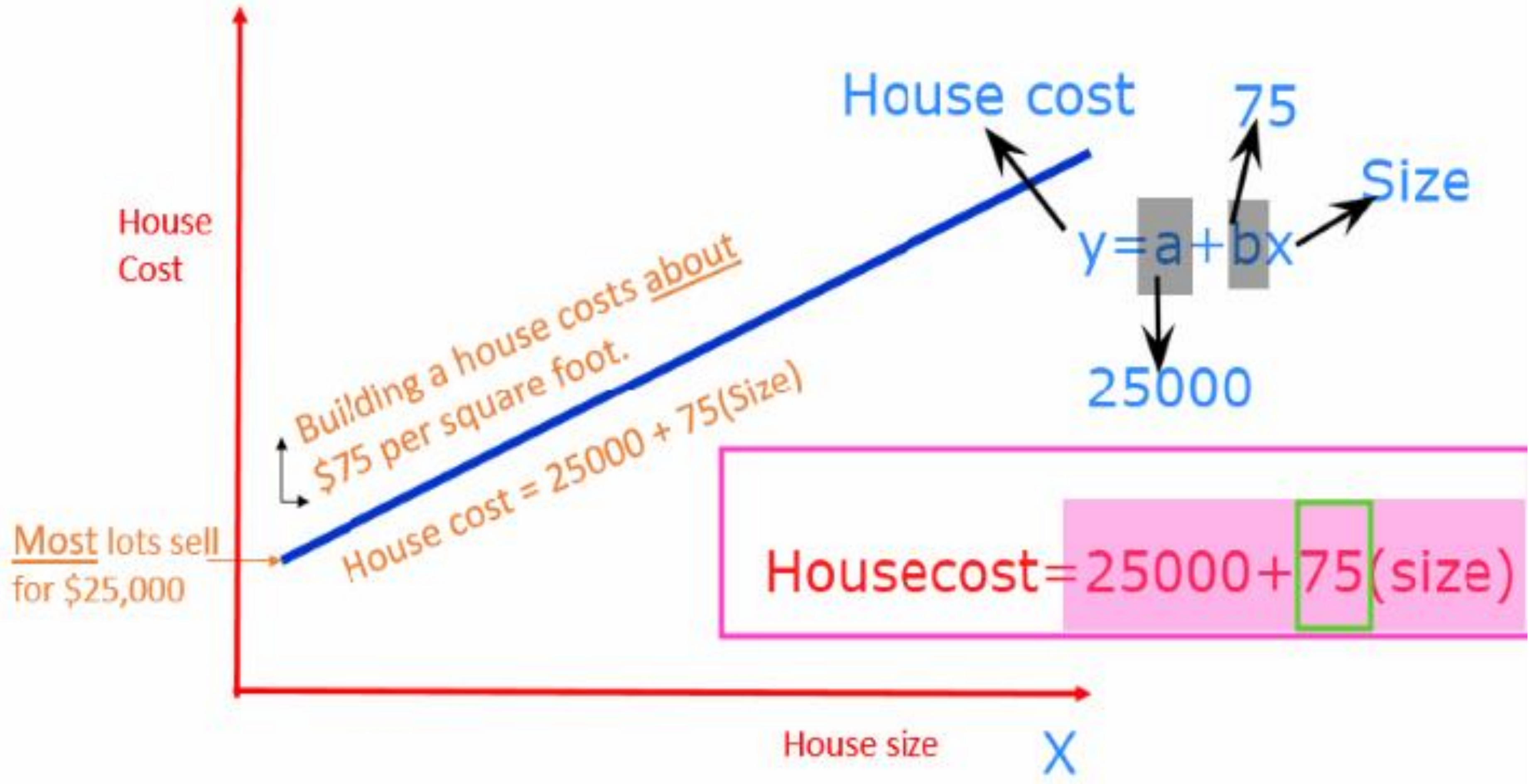
r R^2

Corelation Coefficient


r>0
Perfect Positive Correlation
Strong Positive Correlation
Weak Positive Correlation
No Correlation
r<0
Weak Negative Correlation
Strong Negative Correlation
Perfect Negative Correlation

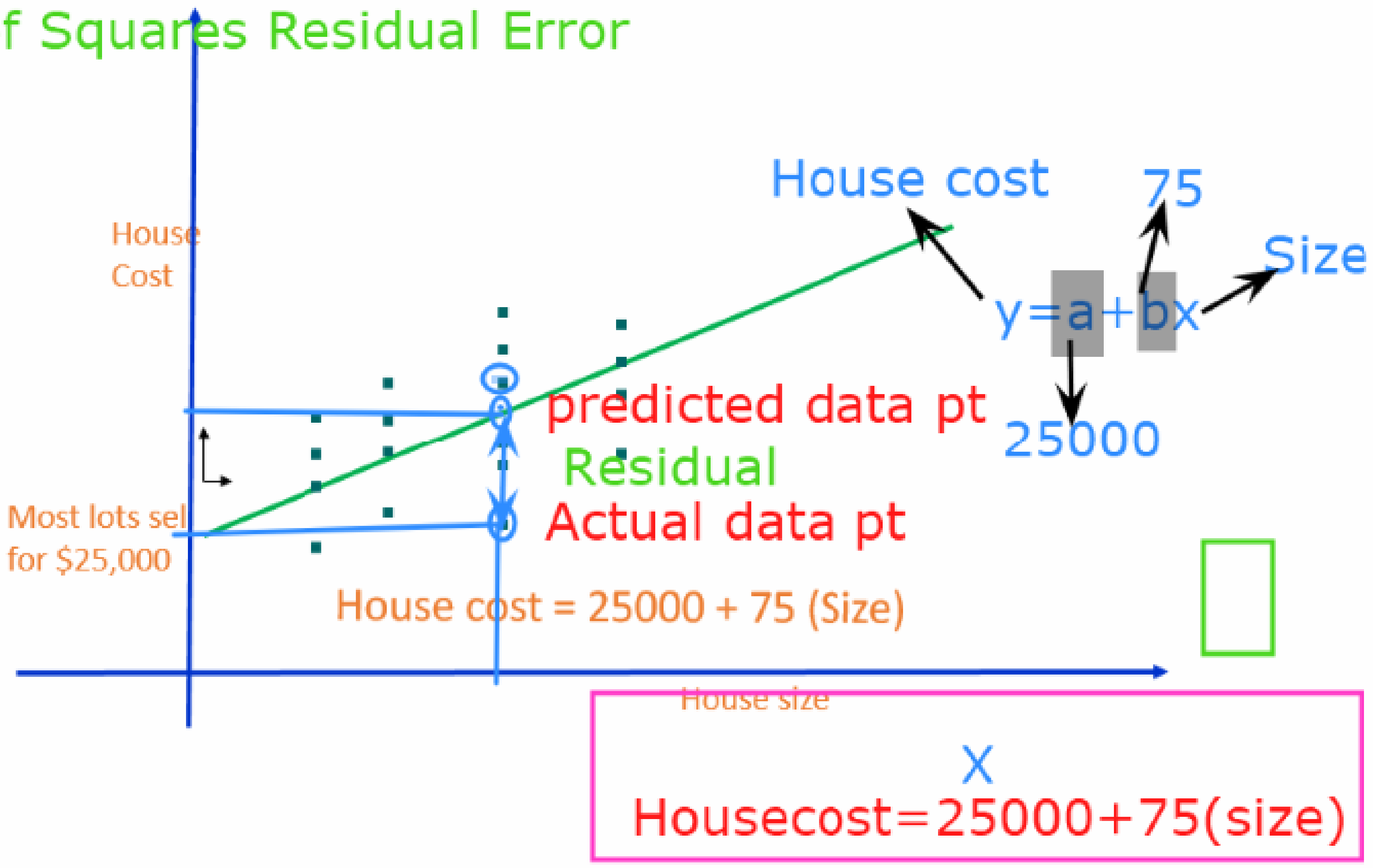
The Model

The model has a deterministic and a probabilistic components



However, house cost vary even among same size houses!

SSE: Sum of Squares Residual Error



Estimating the Coefficients

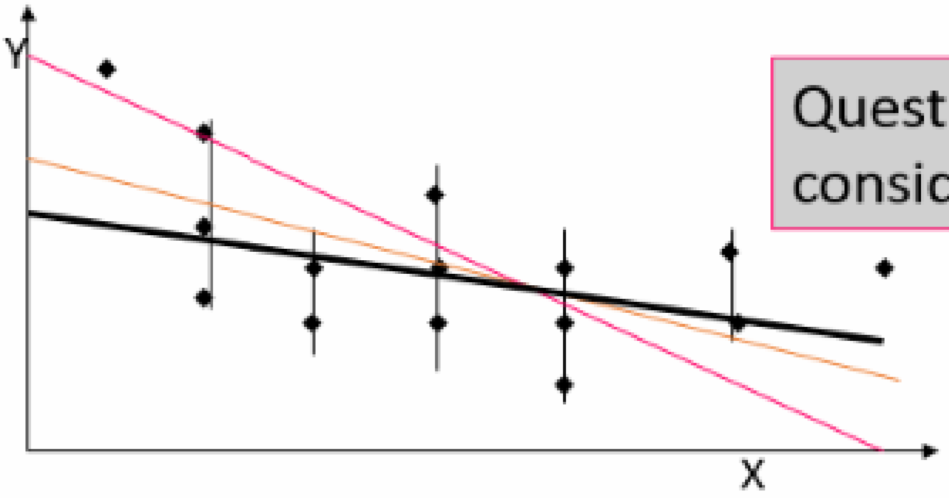
Cost function

minimum

Best fit line for Regression

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.

MSE:Mean Squared Error



Question: What should be considered a good line?

- $MeanSquaredError(mse) = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}\boxed{(y_i - x_i)^2}}$ $\hat{y}$
- $MeanAbsoluteError(mae) = \underbrace{\boxed{(\frac{1}{n})}}_{\text{no. of data points}} \underbrace{\boxed{\sum_{i=1}^{n}}}_{\text{sum of}} \boxed{|\underbrace{y_i}_{\text{Actual value}} - \underbrace{x_i}_{\text{Predicted value}}|}$

The Estimated Coefficients

To calculate the estimates of the line coefficients, that minimize the differences between the data points and the line, use the formulas:

$$b_1 = \frac{\text{cov}(X,Y)}{s_X^2} \left(= \frac{s_{XY}}{s_X^2} \right)$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

The regression equation that estimates the equation of the first order linear model is:

$$\hat{Y} = b_0 + b_1 X$$

The Least Squares (Regression) Line

A good line is one that **minimizes the sum of squared** differences between the points and the line.

Sum of Squares for Errors

- This is the sum of differences between the points and the regression line.
- It can serve as a measure of how well the line fits the data. SSE is defined by

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 .$$

  - A shortcut formula

$$\text{SSE} = (n-1)s_Y^2 - \frac{[\text{cov}(X,Y)]^2}{s_X^2}$$

```
In [30]: summ.summary()
```

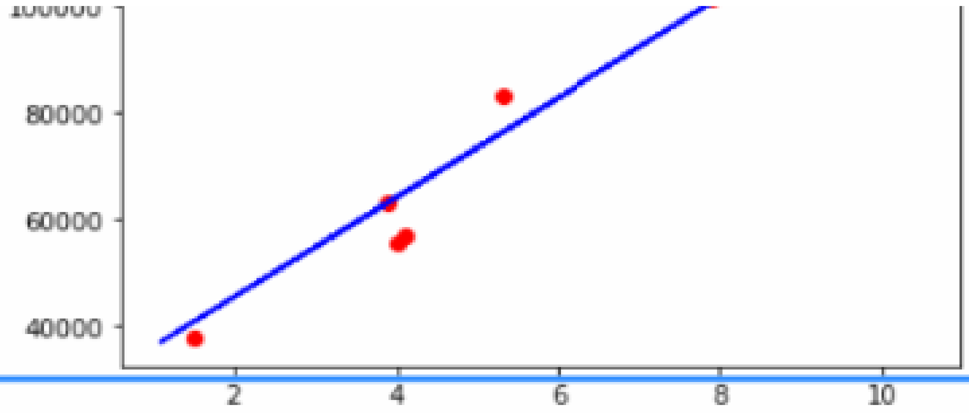
Out[30]:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.938
Model:	OLS	**Adj. R-squared:**	0.935
Method:	Least Squares	**F-statistic:**	273.2
Date:	Mon, 18 Jul 2022	**Prob (F-statistic):**	2.51e-12
Time:	15:32:03	**Log-Likelihood:**	-202.60
No. Observations:	20	**AIC:**	409.2
Df Residuals:	18	**BIC:**	411.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.682e+04	3033.148	8.841	0.000	2.04e+04	3.32e+04
x1	9345.9424	565.420	16.529	0.000	8158.040	1.05e+04

Omnibus:	2.688	**Durbin-Watson:**	2.684
Prob(Omnibus):	0.261	**Jarque-Bera (JB):**	1.386
Skew:	0.305	**Prob(JB):**	0.500
Kurtosis:	1.864	**Cond. No.**	11.7




```
In [16]: #Coefficient
         b=reg.coef_

In [17]: b

Out[17]: array([9345.94244312])

In [18]: #Intercept
         a=reg.intercept_

In [19]: a

Out[19]: 26816.19224403119

In [22]: reg.predict([[13]])

Out[22]: array([148313.44400462])
```

$y=a+bx$

$y=26816.19+9345.94(\text{Exp})$

13

Types of Linear Regression

- Linear regression can be further divided into two types of the algorithm:
- **Simple Linear Regression:**

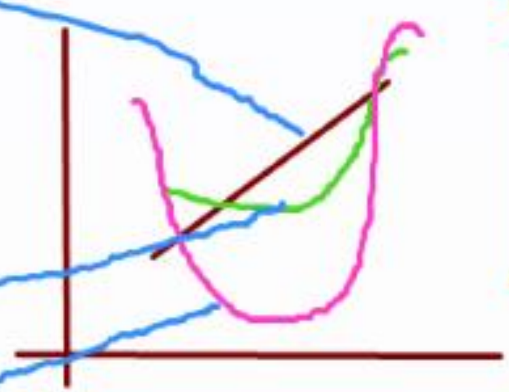
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:**

  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Who can see what you share here? Recording On

Simple Linear Regression

$$y = b_0 + b_1 \boxed{x_1}$$



Multiple Linear Regression

$$y = b_0 + b_1 \boxed{x_1} + b_2 \boxed{x_2} + \boxed{\ldots} + b_n \boxed{x_n}$$

Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

Who can see what you share here? Recording On

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code Commit

```
Out[36]: (50, 5)
```

```
In [37]: dataset.head()
```

Out[37]:

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

Feature Transform

New York	Cali	Florida
1	0	0
0	1	0

Encoding

OneHotEncoder

```
In [41]: X=dataset.iloc[:,:-1].values
         y=dataset.iloc[:,-1].values
```

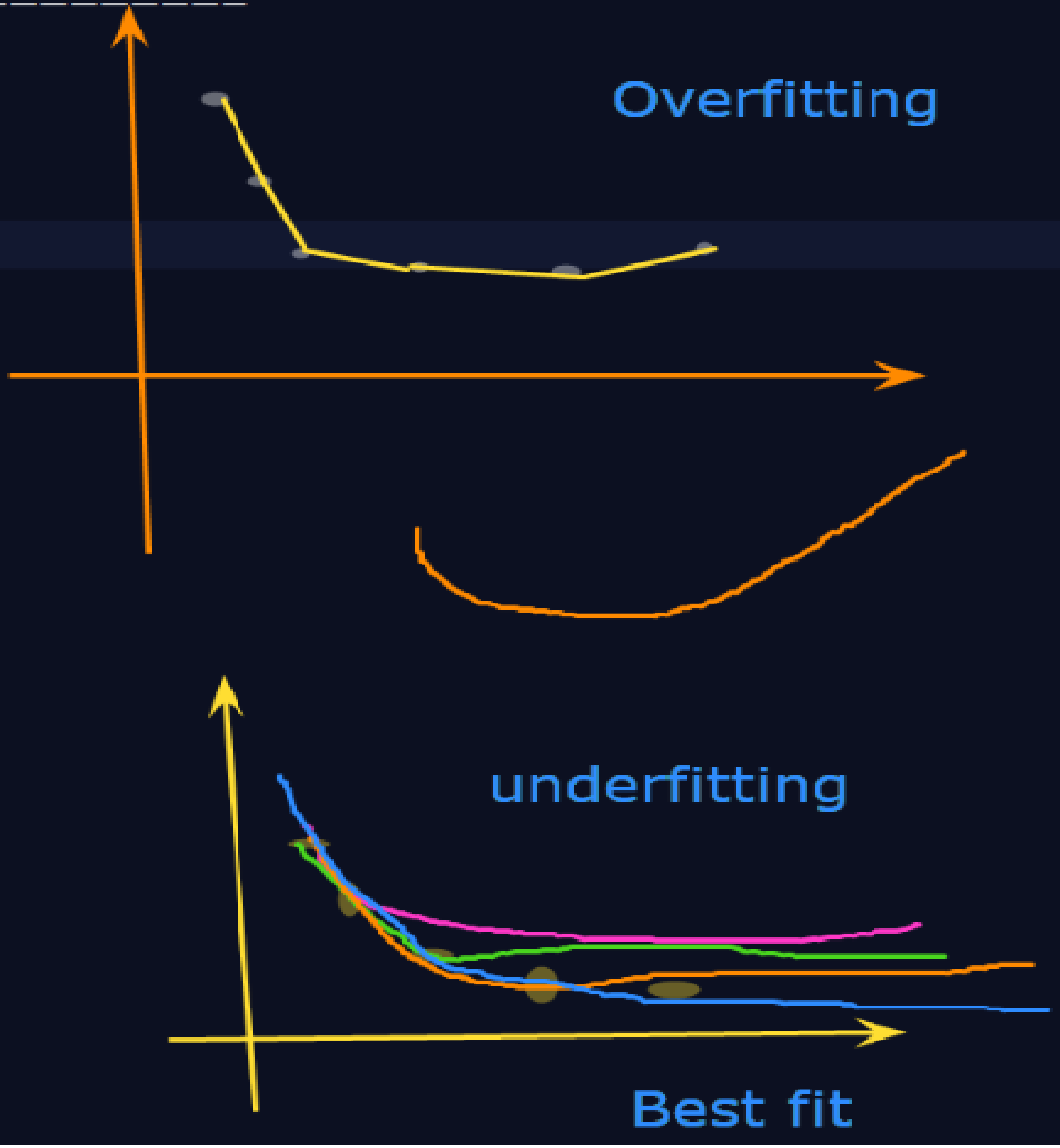
Convert cate---->num

```
In [43]: print(X)

[[165349.2 136897.8 471784.1 'New York']
 [162597.7 151377.59 443898.53 'California']
 [153441.51 101145.55 407934.54 'Florida']
 [144372.41 118671.85 383199.62 'New York']
 [142107.34 91391.77 366168.42 'Florida']
 [131876.9 99814.71 362861.36 'New York']
 [134615.46 147198.87 127716.82 'California']
```

Who can see what you share here? Recording On

```
==============================================
Day 5: Types of Regression
==============================================
Date: 18/07/2022
Topics:
------------------
     -Linear Regression
     -Polynomial Regression
     -Ridge Regression
     -Lasso Regression
     -ElasticNet Regression
     -Logistic Regression
```



α

Ridge=Loss+@$\boxed{||W||}$

Penalty

$\alpha = \cancel{0.01}$

$\alpha = 2$