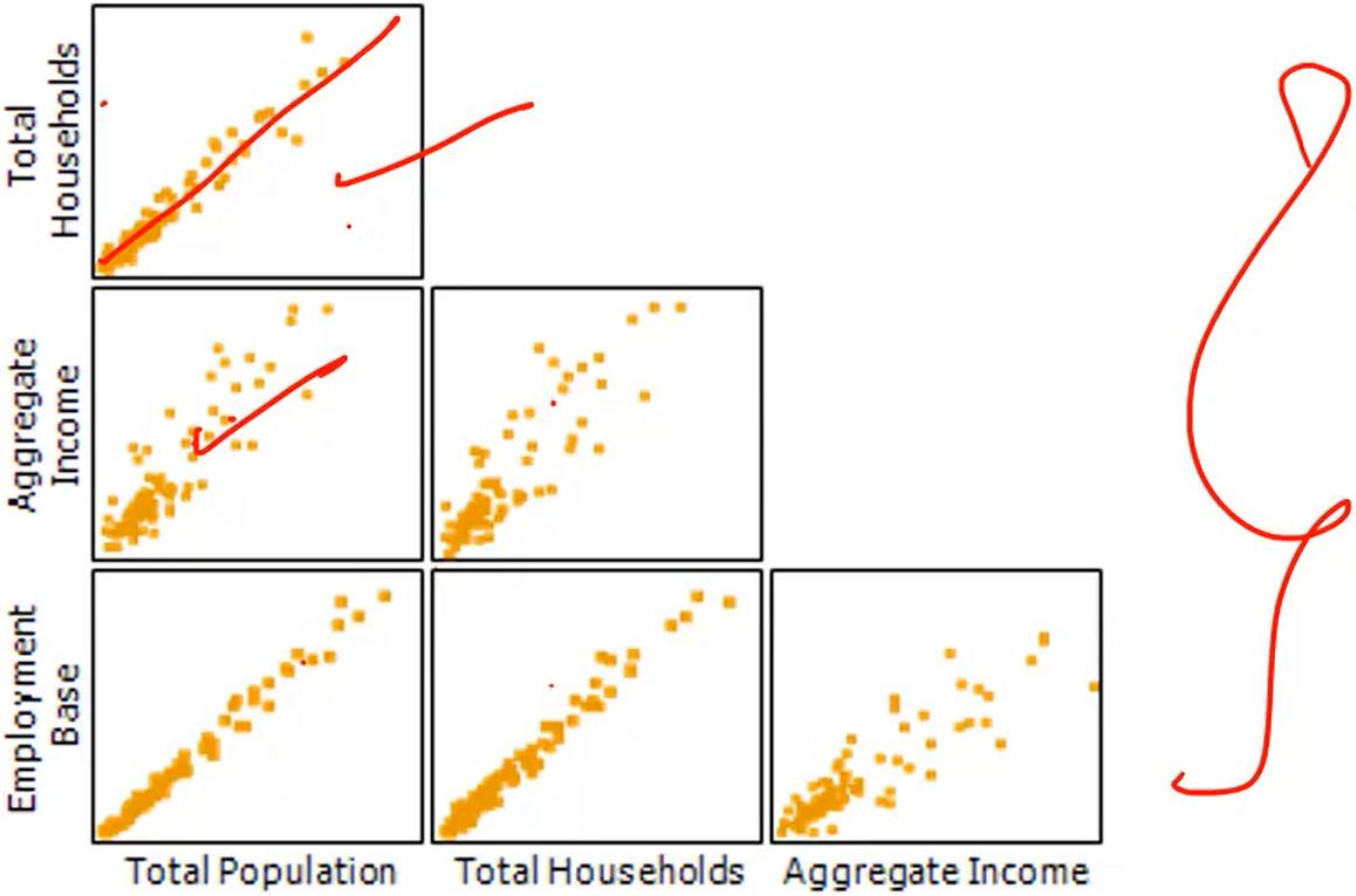
MULTI-COLINEAR | TESTING



MULTI-COLINEAR | TESTING

- An easy way to detect multicollinearity is to calculate correlation coefficients for all pairs of predictor variables.
- If the correlation coefficient, r, is exactly +1 or -1, this is called perfect multicollinearity.
- If r is close to or exactly -1 or +1, one of the variables should be removed from the model if at all possible
- Multicollinearity generally occurs when there are high correlations between two or more predictor variables.
- In other words, one predictor variable can be used to predict the other.
- This creates redundant information, skewing the results in a regression model.
- Examples of correlated predictor variables (also called multicollinear predictors) are a person's height and weight, age and sales price of a car, or years of education and annual income.

KINDS OF MULTICOLLINEARIT

Structural multicollinearity:

- This type occurs when we create a model term using other terms.
- In other words, it's a byproduct of the model that we specify rather than being present in the data itself.
- For example, if you square term X to model curvature, clearly there is a correlation between X and X^2.

Data multicollinearity:

- This type of multicollinearity is present in the data itself rather than being an artifact of our model. Observational experiments are more likely to exhibit this kind of multicollinearity.

#learnaiwithramisha

Variance Inflation Factor(VIF)

- A variance inflation factor(VIF)detects multicollinearity in regression analysis.
- Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model;
- it's presence can adversely affect your regression results.
- The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

Example: Multicollinearity

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0705118	56.23%	54.22%	50.48%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.155	0.132	1.18	0.243	
%Fat	0.00557	0.00409	1.36	0.176	14.93
Weight kg	0.01447	0.00285	5.07	0.000	33.95
Activity	0.000022	0.000007	3.08	0.003	1.05
%Fat*Weight kg	-0.000214	0.000074	-2.90	0.005	75.06

Example: Multicollinearity

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0705118	56.23%	54.22%	50.48%

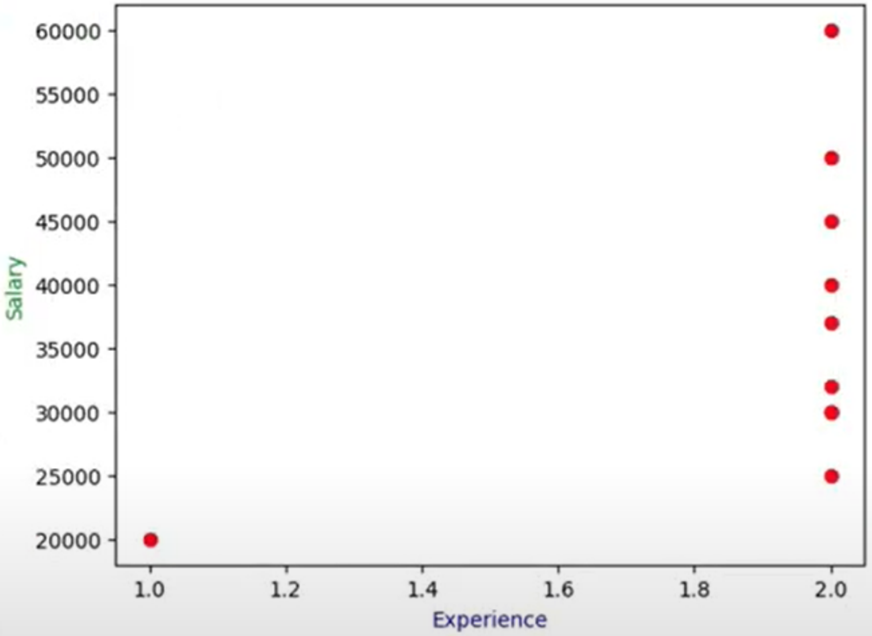
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.82161	0.00973	84.40	0.000	
%Fat S	-0.00598	0.00193	-3.10	0.003	3.32
Weight S	0.00835	0.00107	7.83	0.000	4.75
Activity S	0.000022	0.000007	3.08	0.003	1.05
%Fat S*Weight S	-0.000214	0.000074	-2.90	0.005	1.99

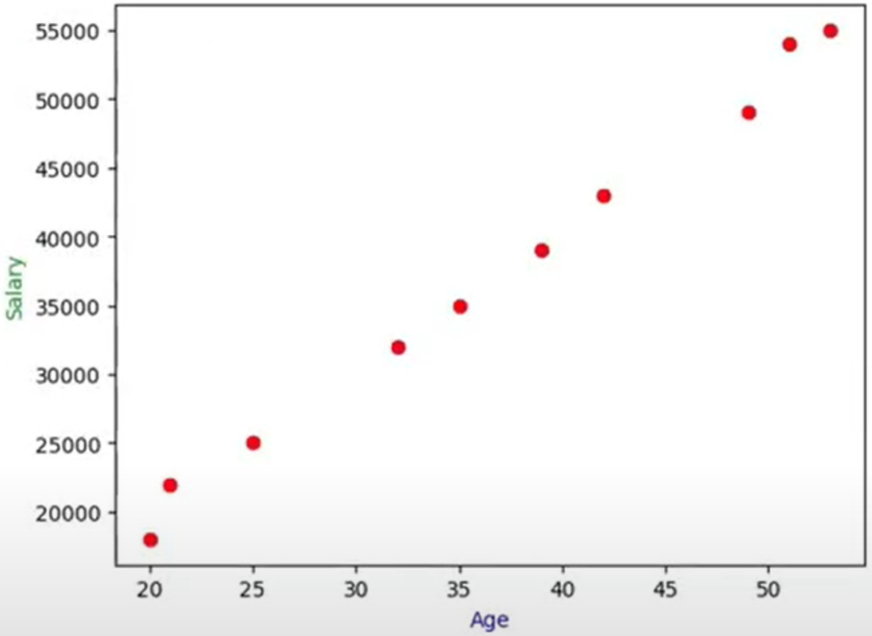
Experience	Age	Designation	Salary
2	25	Data Scientist	25000
2	35	Data Analyst	35000
2	39	HR	39000
2	49	Business Analyst	49000
1	20	Data Scientist	18000
2	51	Data Engineer	54000
2	42	Data Scientist	43000
2	53	Data Analyst	55000
2	32	HR	32000
2	21	Business Analyst	22000

3. Correlation Coefficient

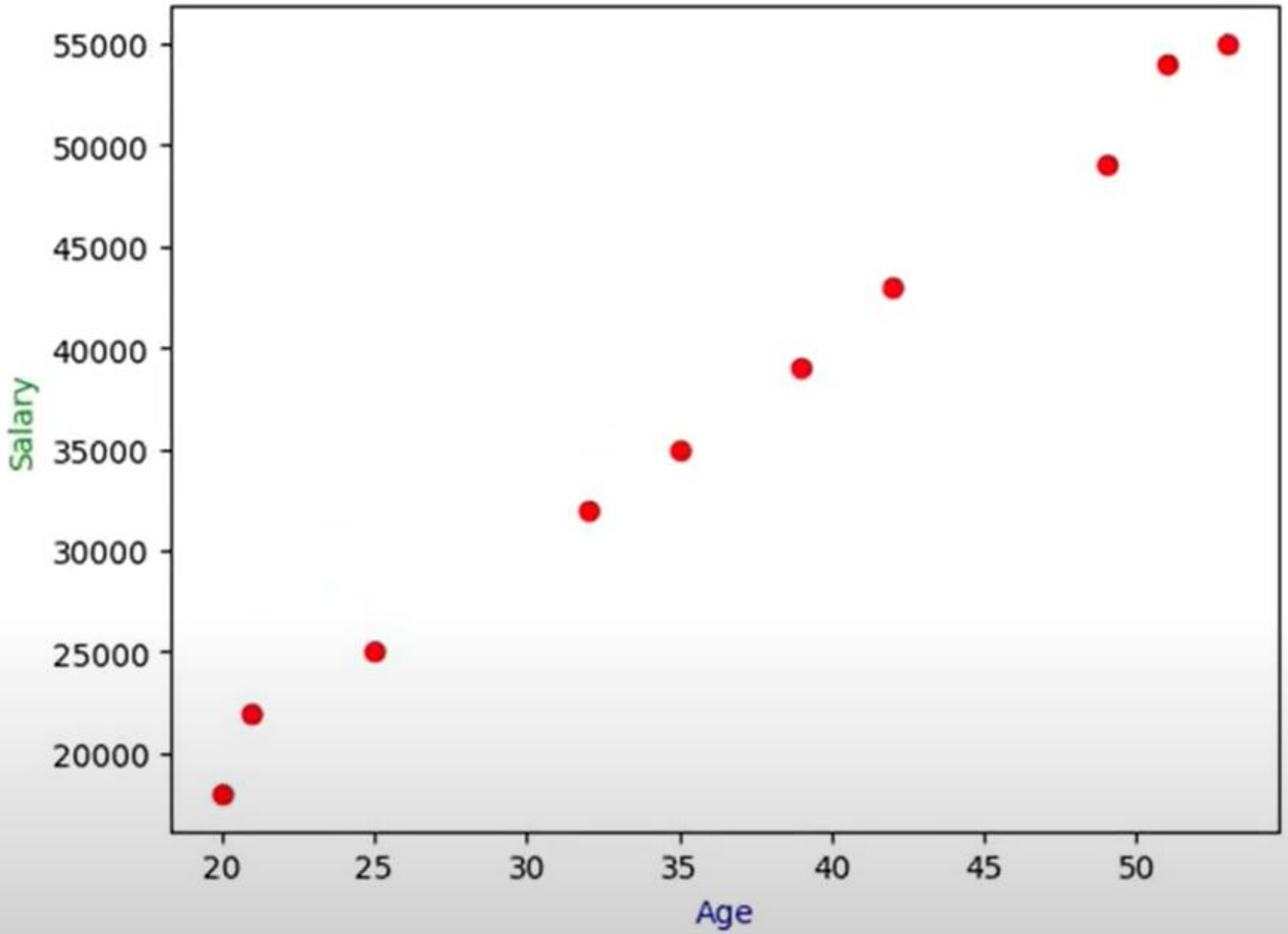
Experience	Salary
2	25000
2	35000
2	39000
2	49000
1	18000
2	54000
2	43000
2	55000
2	32000
2	22000



Age	Salary
25	25000
35	35000
39	39000
49	49000
20	18000
51	54000
42	43000
53	55000
32	32000
21	22000

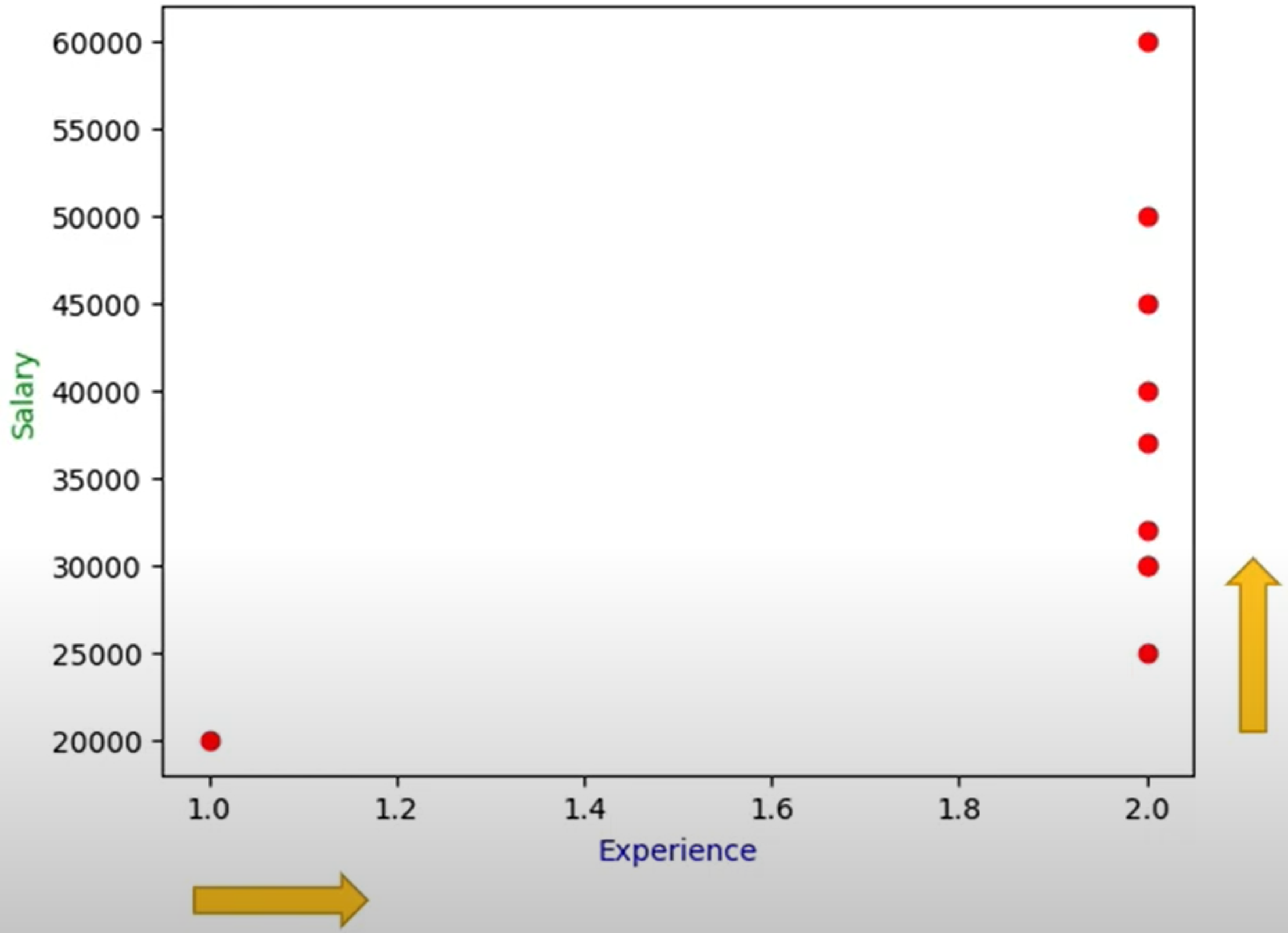


Age	Salary
25	25000
35	35000
39	39000
49	49000
20	18000
51	54000
42	43000
53	55000
32	32000
21	22000



Experience
2
2
2
2
1
2
2
2
2
2

Salary
25000
35000
39000
49000
18000
54000
43000
55000
32000
22000



Filter Method

3. Correlation Coefficient

If the **correlation** is lesser than the **Threshold** then drop those features.

Threshold = 0.5

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

Experience	Salary
2	25000
2	35000
2	39000
2	49000
1	18000
2	54000
2	43000
2	55000
2	32000
2	22000

Correlation = 0.13

Age	Salary
25	25000
35	35000
39	39000
49	49000
20	18000
51	54000
42	43000
53	55000
32	32000
21	22000

Correlation = 0.99

3. Correlation Coefficient

How to choose threshold value????

Experience	Age	Designation	Salary
2	25	Data Scientist	25000
2	35	Data Analyst	35000
2	39	HR	39000
2	49	Business Analyst	49000
1	20	Data Scientist	18000
2	51	Data Engineer	54000
2	42	Data Scientist	43000
2	53	Data Analyst	55000
2	32	HR	32000
2	21	Business Analyst	22000

Threshold = [0.05,0.1,0.15,0.2,0.25]

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma (x_i - \dot{x})^2 \Sigma (y_i - \dot{y})^2}}$$

What is Multicollinearity?

- Multicollinearity is a statistical phenomenon where two or more independent variables in a linear regression model are strongly correlated. It means that the variables have an almost perfect or exact relationship between them.
- For Example:

DOB	AGE
Diabetic	Glucose Level

Types of Multicollinearity

- ***Positive Correlation***
- Example:

age | WorkExperience

- ***Negative Correlation***
- Example:

age | YearsToRetire

Steps to Avoid Multicollinearity

- Set VIF value and remove variables above the value
- Use Regularization Techniques like Ridge, Lasso and ElasticNet
- Using Heatmap/Correlation Matrix detect the highly collinear variables and remove them manually