1. **What are the assumptions of linear regression regarding residuals?**

Ans:

In linear regression term Residual is difference between the observed value and predicted value.

The assumption of Linear regression regarding residuals are as below:

a. Residuals have normal distribution
b. Residuals are independent of each other
c. Residuals are homoscedastic which means they have constant variance

2. **What is the coefficient of correlation and the coefficient of determination?**

Ans:

<u>Coefficient of correlation:</u>

The strength of the relationship between 2 variables can be measured with coefficient of correlation denoted by 'r'.
coefficient of correlation can a value between -1 to 1

1:- shows strong positive relationship between 2 variables
0:- shows no relationship between 2 variables
-1:- shows string negative relationship between 2 variables.

Absolute value of the coefficient of correlation show strength of the relationship.
Example: relationship with coefficient of correlation value of 0.85 is stronger than the one with 0.50

<u>Coefficient of determination:</u>

The coefficient of determination is the proportion of variance in dependent variable which is predicted from the independent variable.
It is denoted by "R-Square".

The coefficient of determination is square of the correlation between predicted y value and actual y value.
It has value between 0-1

- coefficient of determination = 0 means we can not predict dependent variable from independent variable

- coefficient of determination = 1 means we can predict dependent variable from independent variable without any error

- value between 0 and 1 shows extent of prediction of the dependent variable from independent variable

3. Explain the Anscombe's quartet in detail.
Ans:

Anscombe's quartet is used to show the importance of plotting the data/data visualisation while analysing the data. Anscombe's quartet uses for data sets which have almost identical descriptive statistics for this experiment.

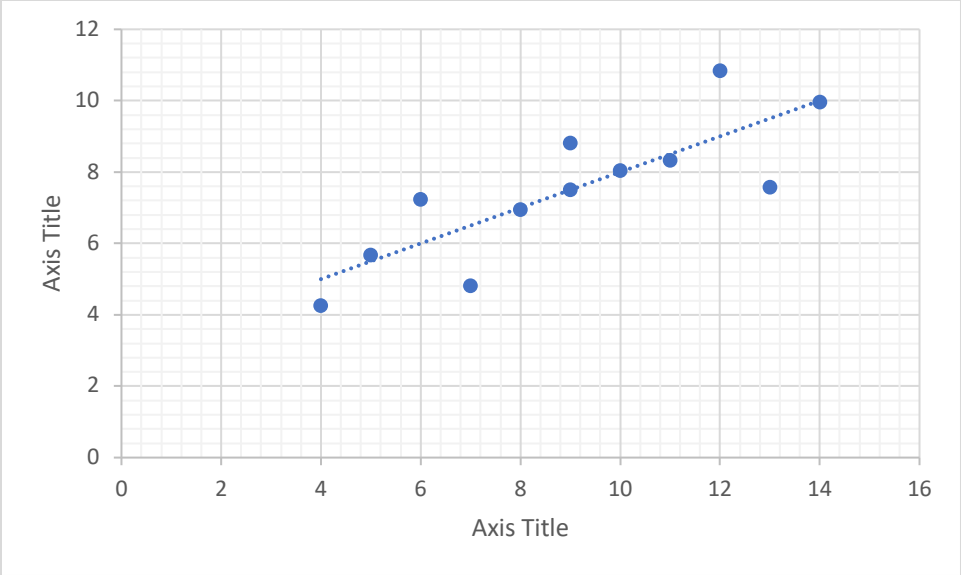
Consider below 4 datasets:

	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
mean	9	7.5009	9	7.5009	9	7.5	9	7.5009
STD	3.1623	1.937	3.1623	1.9371	3.1623	1.9359	3.1623	1.9361

All four datasets have almost same statistical properties of mean and standard deviation.

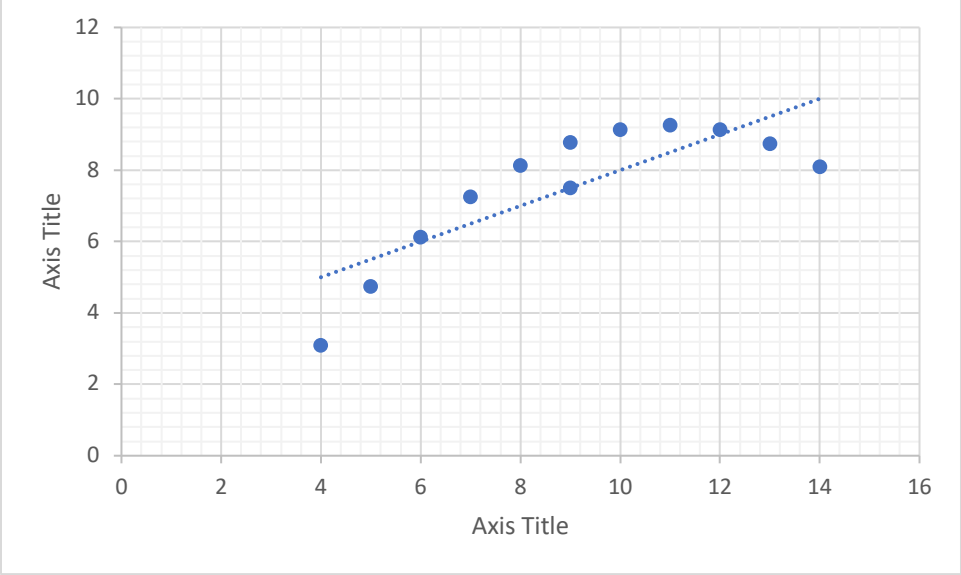
If we plot these 4 datasets they look as below:

Dataset I:



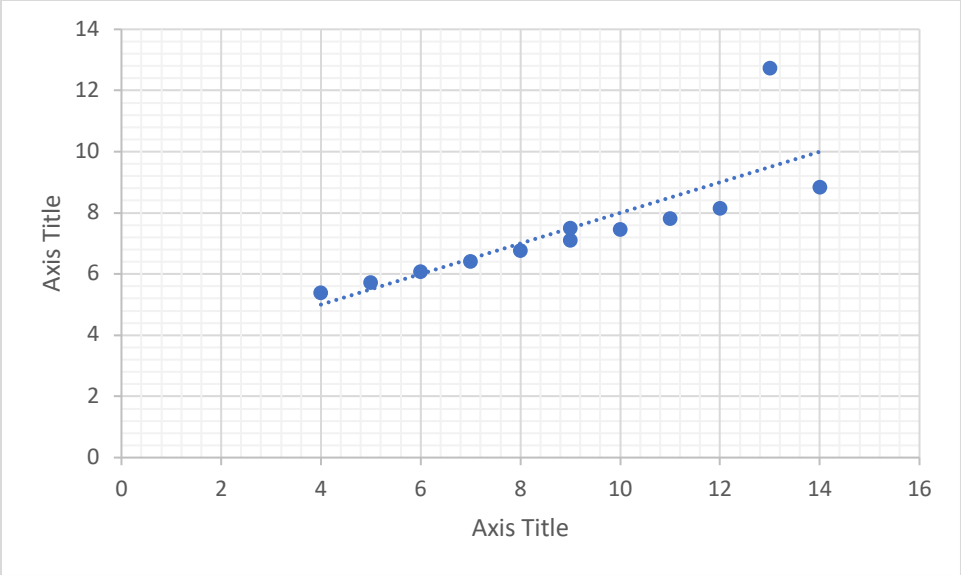
This data set seems to have linear regression with some variance.

Dataset II:



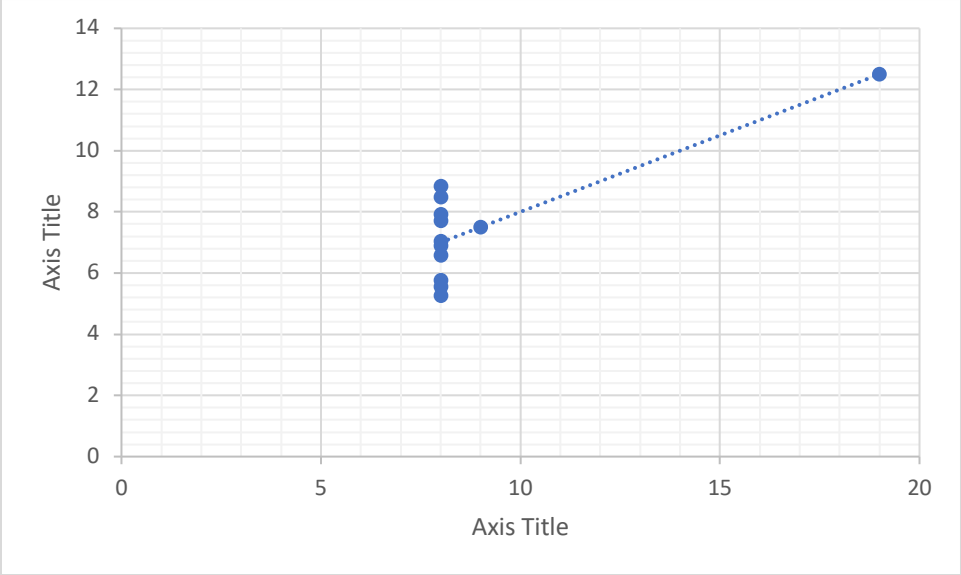
Dataset II fits a curve but does not seems to be following a perfect linear relationship

Dataset III:



Dataset III seems to be following in good linear relationship with one outlier

Dataset IV:



Dataset IV seems to have almost all X constant with one outlier.

So based on the visualisation of the dataset we could understand the data spread and relationships between variable but we if make out judgement by just looking at statistical data we could have made incorrect

assumptions. For this purpose the Anscombe's quartet plays a significant role.

4. What is Pearson's R?

Ans:

Pearson's R (Pearson's correlation coefficient) is the number between -1 and 1 which shows strength and direction of the relationship between two continuous variables.

1:- shows strong positive relationship between 2 variables
0:- shows no relationship between 2 variables
-1:- shows string negative relationship between 2 variables.

Absolute value of the coefficient of correlation show strength of the relationship.
Example: relationship with coefficient of correlation value of 0.85 is stronger than the one with 0.50

5. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:

Scaling of Feature scaling is a process of normalization of the independent variable during data pre-processing.

The reasons for feature scaling are as below:

a. If we have multiple independent variables with large ranges/scales can lead to introduction of bias
b. The data become more easy to interpret
c. For gradient decent method it leads to faster convergence

Normalized scaling: Normalized scaling is the process of scaling where we fit the data range between 0 and 1

Normalized scaling can be achieved with min-max scaling:

$Xi = (X - \min(x))/(\max(x) - \min(x))$

Where x is original value
Max(x) is maximum value of the feature

Min(x) is minimum value of the feature

Standardized scaling: Standardized scaling is rescaling of feature so that they have properties μ=0 and σ=1

Formula for Standardized scaling is:

X = (x – mean(x)) / sd (x)

sd = standard deviation

The difference between normalized scaling is normalized scaling rescales the data to a confined range whereas the standardised scaling changes/transformed the data to have statistical properties of μ=0 and σ=1

6. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   Ans:

   Variance inflation factor shows the relationship between all the independent variables. And the VIF show how strongly one independent variable is correlated with all other independent variables.

   Formula for VIF is = 1/(1 - R-square)

   So if the VIF infinite means that the variable has exact collinearity with another independent variable and so the variable is redundant.