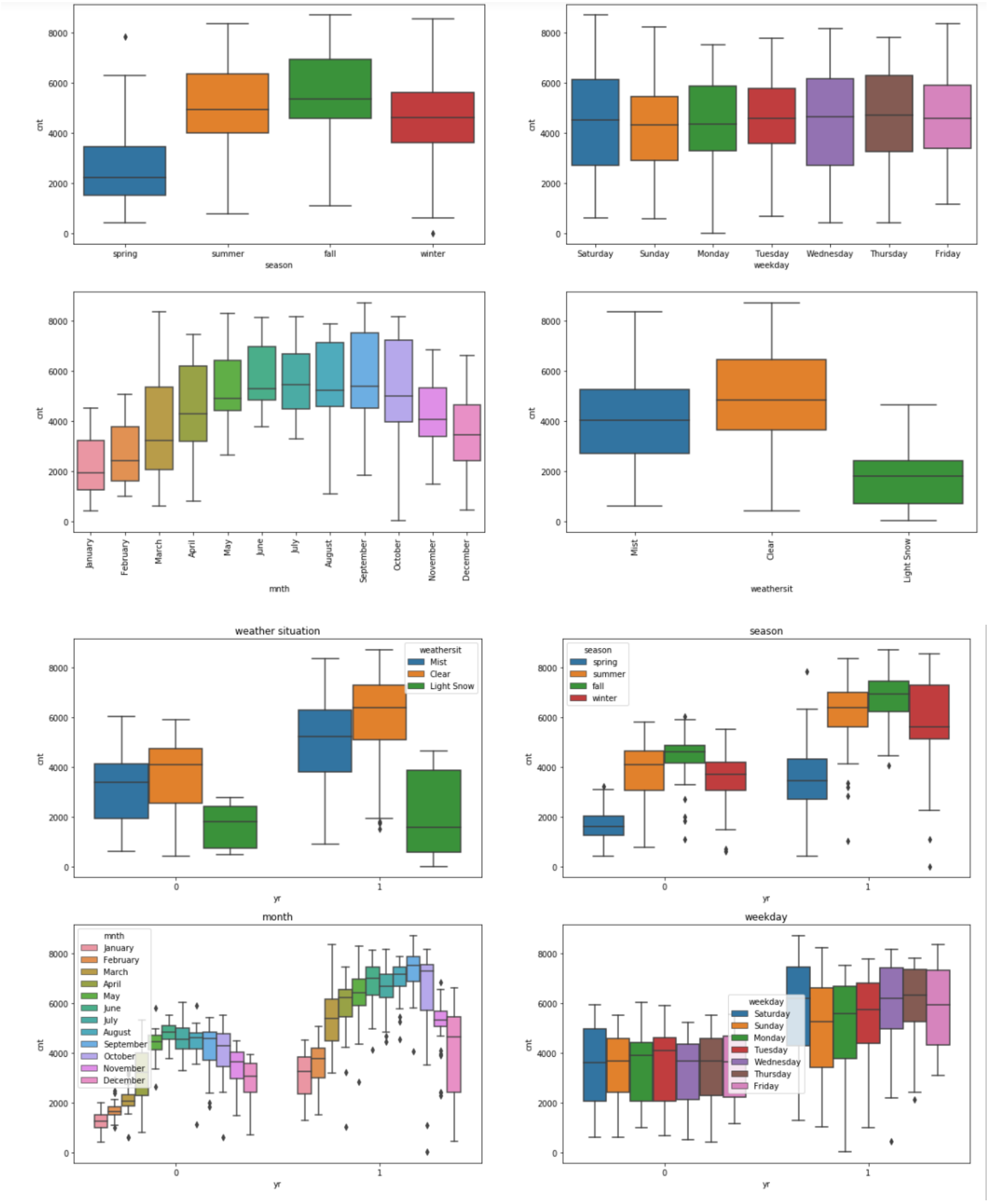
Ques 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: Below are the findings from the categorical variable's analysis on dependent variable cnt:



- Based on season bike sharing is in order of Fall>summer>winter>spring
  In fall season bike sharing is more than other seasons.
- Based on weather situation bike sharing is in order of clear>mist>light snow

In clear weather situation bike sharing is more

- Median for Thursday, Wednesday and Saturday is same
- In September Bike sharing is more
- In 2019 bike sharing is more than 2018
- in 2019 Bikes sharing is more for clear weather, in fall season and in September month

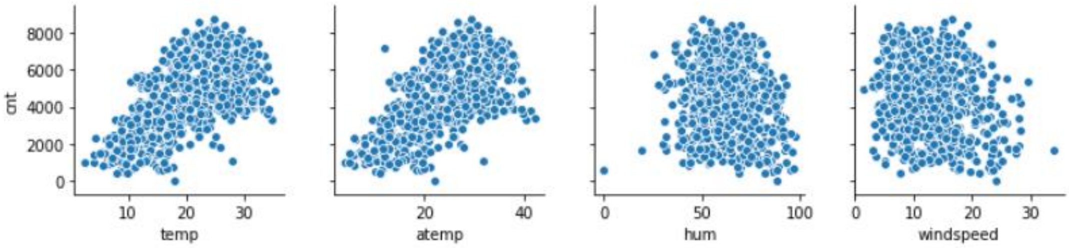
Ques 2: Why is it important to use drop_first=True during dummy variable creation?

Ans: During Dummy variable creation it's important to use drop_first = True because dummy variable will be correlated (redundant) and result multicollinearity. This may break our model.

For example, if we have variable Gender and having values male or Female in that case both will be corelated it will be either male or Female, hence its good to drop one dummy variable.

Ques 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

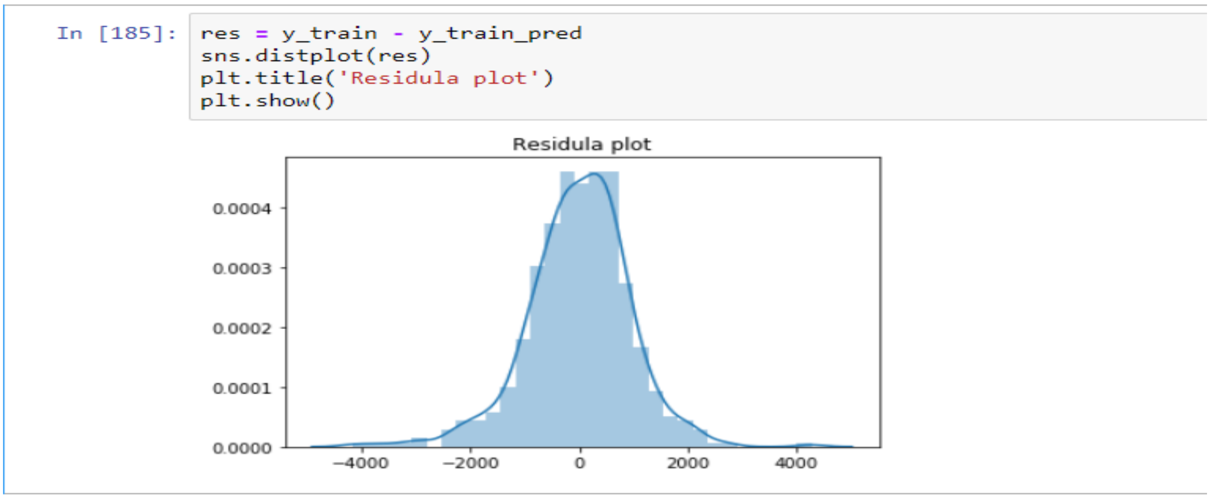
Ans: Based on below pair plot between numerical variables and target variables temp and atemp are having strong correlation.



Ques 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: Assumption validation of Linear Regression

1) **Distribution of error**
   Distribution of error is almost cantered around 0

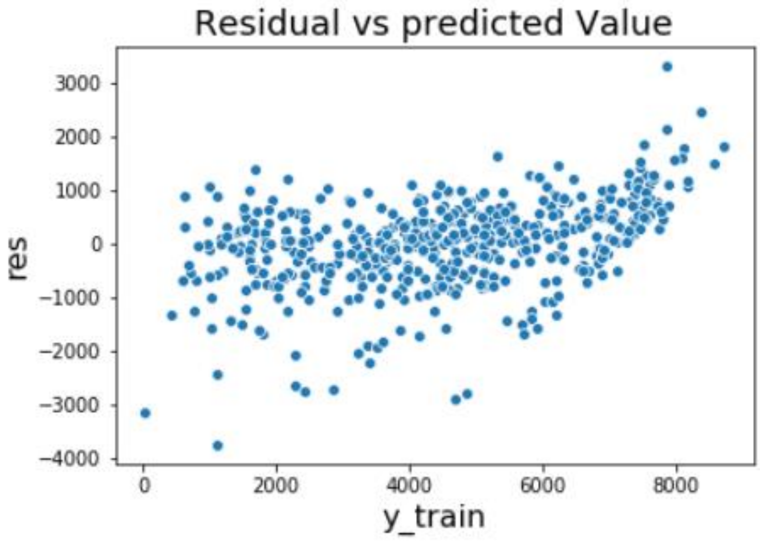


In [185]:

```
res = y_train - y_train_pred
sns.distplot(res)
plt.title('Residula plot')
plt.show()
```

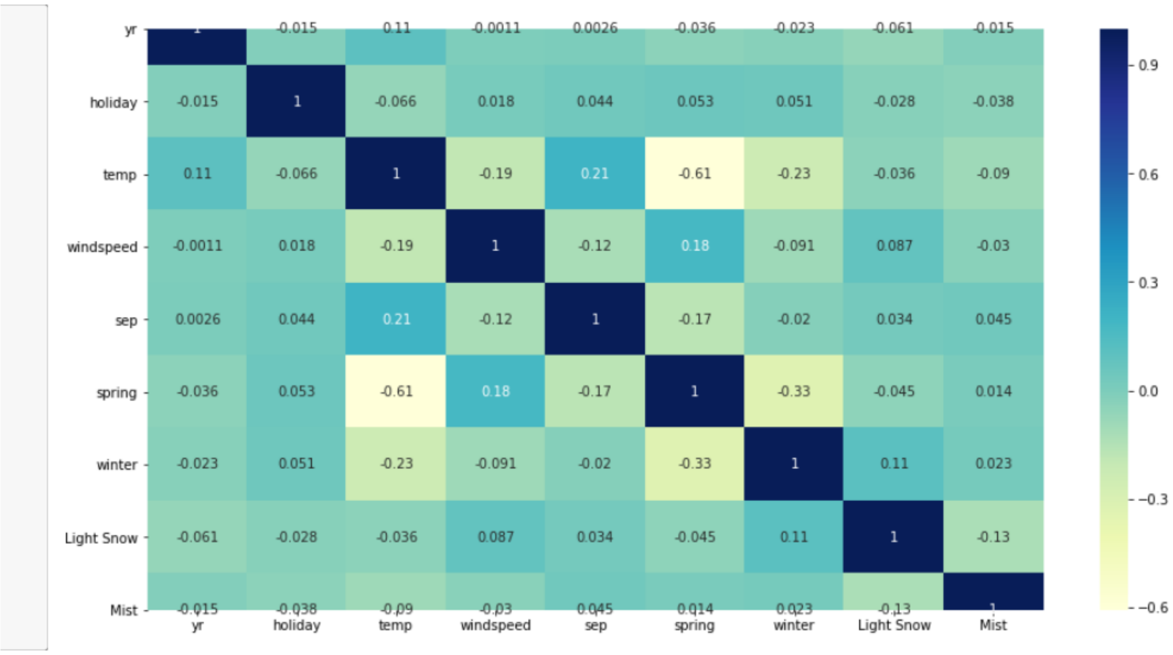
2) Validating Homoscedasticity

No clear pattern in the distribution, hence Homoscedasticity is not there



3) Validating Multi Collinearity

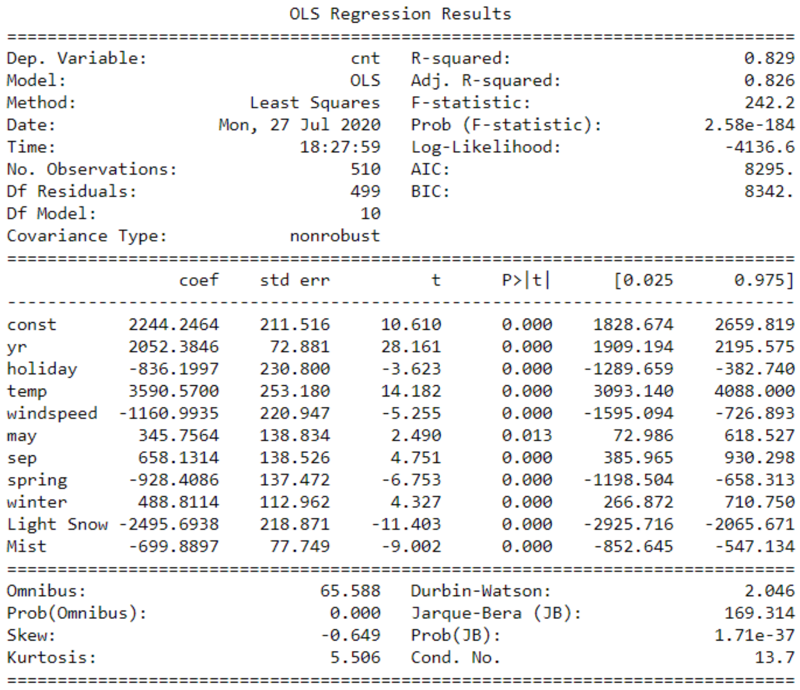
In Heat plot, the magnitude of the correlation coefficients are less than .80, hence no multicollinearity is there



Ques 5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: As per below model top 3 features contributing significantly are:

- tmp
- Light snow
- Yr

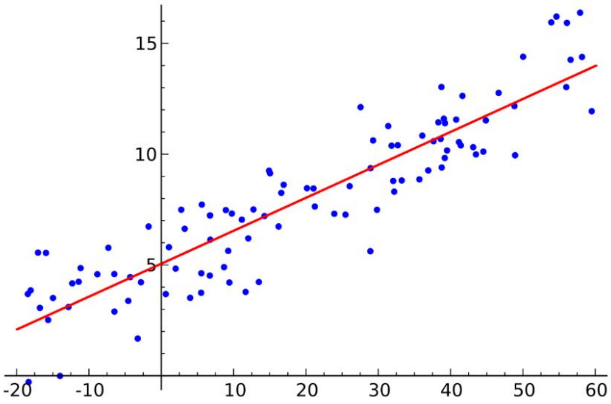


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.829
Model:                            OLS   Adj. R-squared:                  0.826
Method:                 Least Squares   F-statistic:                     242.2
Date:                Mon, 27 Jul 2020   Prob (F-statistic):          2.58e-184
Time:                        18:27:59   Log-Likelihood:                -4136.6
No. Observations:                 510   AIC:                             8295.
Df Residuals:                     499   BIC:                             8342.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2244.2464    211.516     10.610      0.000    1828.674    2659.819
yr          2052.3846     72.881     28.161      0.000    1909.194    2195.575
holiday     -836.1997    230.800     -3.623      0.000   -1289.659    -382.740
temp        3590.5700    253.180     14.182      0.000    3093.140    4088.000
windspeed  -1160.9935    220.947     -5.255      0.000   -1595.094    -726.893
may          345.7564    138.834      2.490      0.013      72.986     618.527
sep          658.1314    138.526      4.751      0.000     385.965     930.298
spring      -928.4086    137.472     -6.753      0.000   -1198.504    -658.313
winter       488.8114    112.962      4.327      0.000     266.872     710.750
Light Snow -2495.6938    218.871    -11.403      0.000   -2925.716   -2065.671
Mist        -699.8897     77.749     -9.002      0.000    -852.645    -547.134
==============================================================================
Omnibus:                       65.588   Durbin-Watson:                   2.046
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              169.314
Skew:                          -0.649   Prob(JB):                     1.71e-37
Kurtosis:                       5.506   Cond. No.                         13.7
==============================================================================
```

General Subjective Questions

Ques 1: Explain the linear regression algorithm in detail

Ans: Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range (Example rate, sales). In Linear Regression we train a model to predict the behaviour of data based on some variables. In the case of linear regression, the two variables which are on the x-axis and y-axis should be linearly correlated.



Linear Regression

The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation as **Y= β0 +β1 X**

The motive of the linear regression algorithm is to find the best values for **α** and **β0**

There are 2 main types:

Simple Linear Regression: Simple Linear regression uses formula **Y= β0 +β1 X** where **β1** and **β0** are the variables, algorithm will try to learn to produce most accurate predictions.

Multivariable Linear Regression: Multivariable Linear regression uses formula y=m1x1+m2x2+m3x3+…..+mnxn

There are 2 more important concept for Linear regression

1. **Cost function**
2. **Gradient descent**

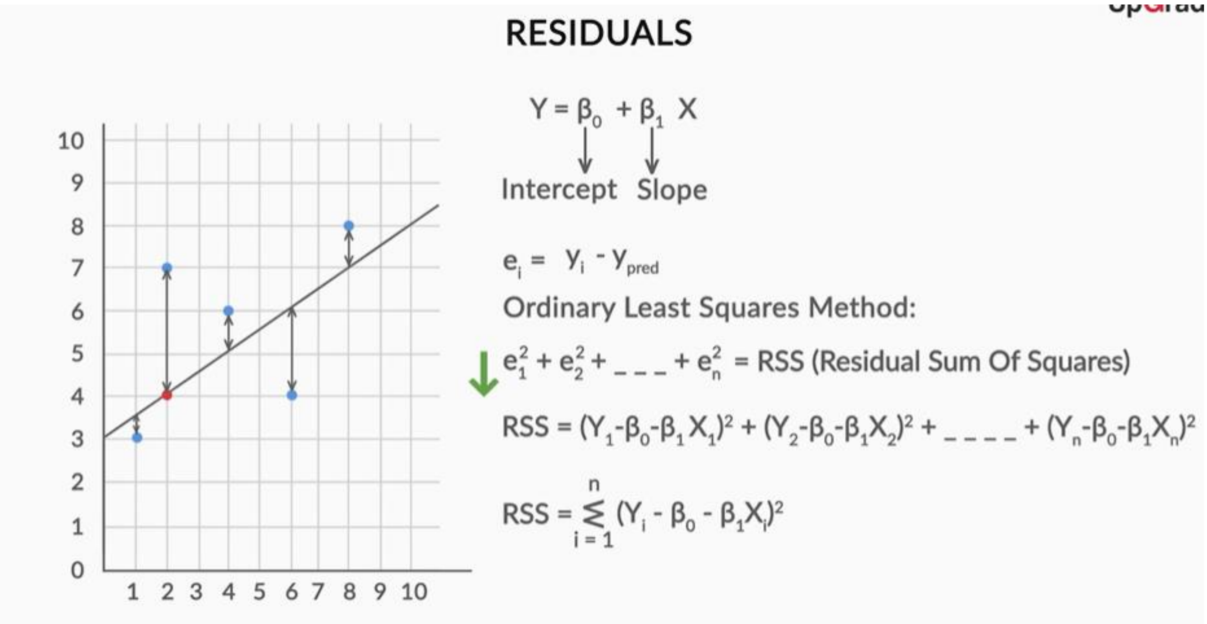
Simple Linear Regression is a statistical model, widely used in ML regression tasks, based on the idea that the relationship between two variables can be explained by the following **formula:**

$$Y= \beta0 \,_{+}\beta1 \, X$$

Where εi is the error term, and β0, β1 are the true (but unobserved) parameters of the regression. The parameter β1 represents the variation of the dependent variable when the independent variable has a unitary variation: namely, if my parameter is equal to 0.75, when my *x* increases by 1, my dependent variable will increase by 0.75. On the other hand, the parameter α represents the value of our dependent variable when the independent one is equal to zero.

Cost function:

The cost function helps us to figure out the best possible values for β0 and β1 which would provide the best fit line for the data points. Best fit line can be obtained by minimizing a quantity called RSS.



RESIDUALS

$$Y = \underset{\text{Intercept}}{\underset{\downarrow}{\beta_0}} + \underset{\text{Slope}}{\underset{\downarrow}{\beta_1}} X$$

$$e_i = y_i - y_{pred}$$

Ordinary Least Squares Method:

$$e_1^2 + e_2^2 + \_\_\_ + e_n^2 = \text{RSS (Residual Sum Of Squares)}$$

$$RSS = (Y_1-\beta_0-\beta_1X_1)^2 + (Y_2-\beta_0-\beta_1X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1X_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1X_i)^2$$

- Blue points represents Yi(y actual) and red point on line represent ypred(predicted y value from best fit line)
- ei represents error residuals
- RSS – Residual sum of squares
- After minimizing RSS value, we can get optimal value for β1 and β0

Gradient Descent:

- Is an optimize algorithm which figure out the best possible values for **β0** and **β1** which would provide the best fit line for the data points.
- Is an optimisation algorithm that optimise the objective function (Cost function for Linear Regression) to reach optimal solution.
- Gradient descent is a method of updating **β0** and **β1** to reduce the cost function RSS.

RSS is cost function

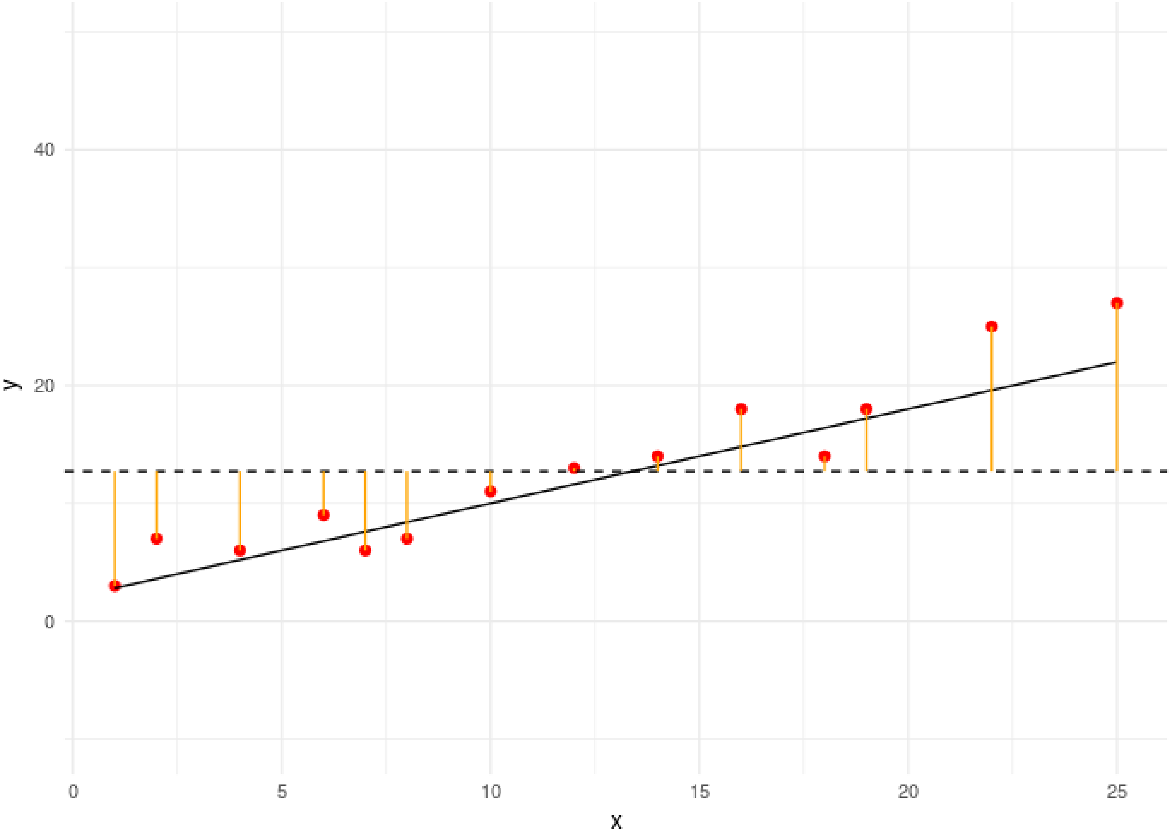
Strength of Simple linear regression:

Strength of the linear regression can be determined by R squared.

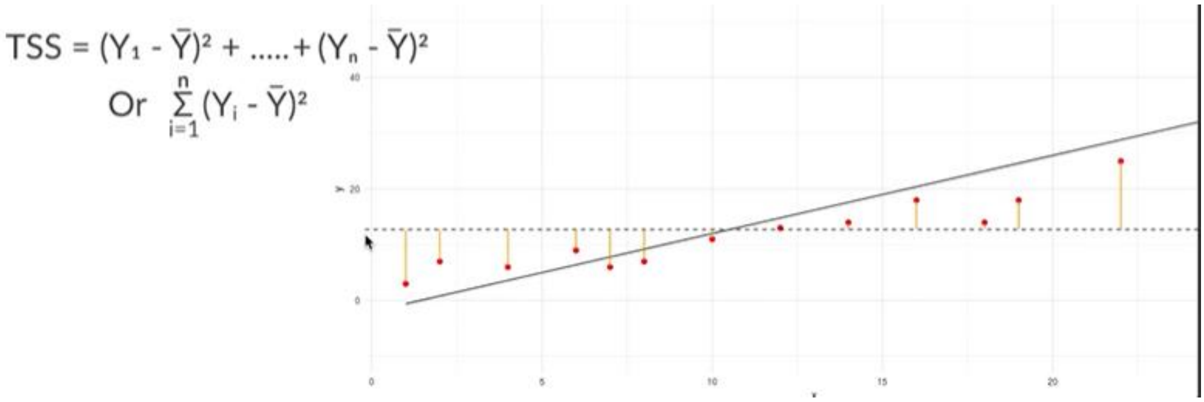
TSS

TSS stands for **total sum of square**. This represent the error of every basic model that can be build without having any independent variable, hence any model which has been build should be better than this.

- In below dotted line represent average of Y.



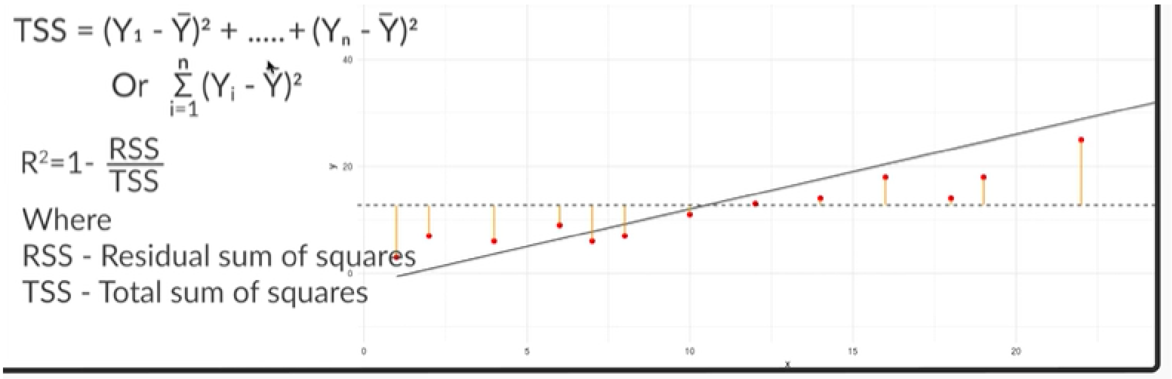
TSS is sum of square of difference between y actual and average of y. Formula for TSS is as Below:



$$TSS = (Y_1 - \bar{Y})^2 + \ldots\ldots + (Y_n - \bar{Y})^2$$

$$\text{Or} \ \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

R-squared (**R²**) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model



$TSS = (Y_1 - \bar{Y})^2 + \ldots.. + (Y_n - \bar{Y})^2$

Or $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$

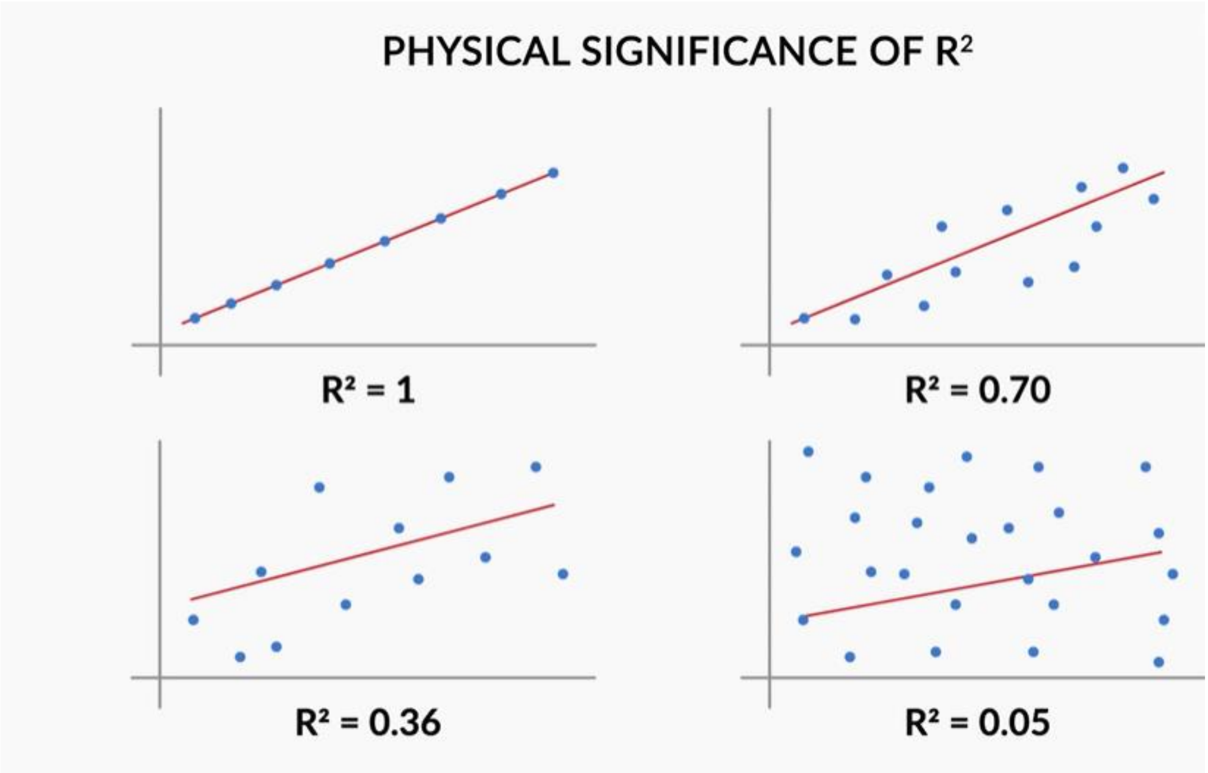
$R^2 = 1 - \frac{RSS}{TSS}$

Where

RSS - Residual sum of squares

TSS - Total sum of squares

Physical significance of R²



PHYSICAL SIGNIFICANCE OF R²

$R^2 = 1$

$R^2 = 0.70$

$R^2 = 0.36$

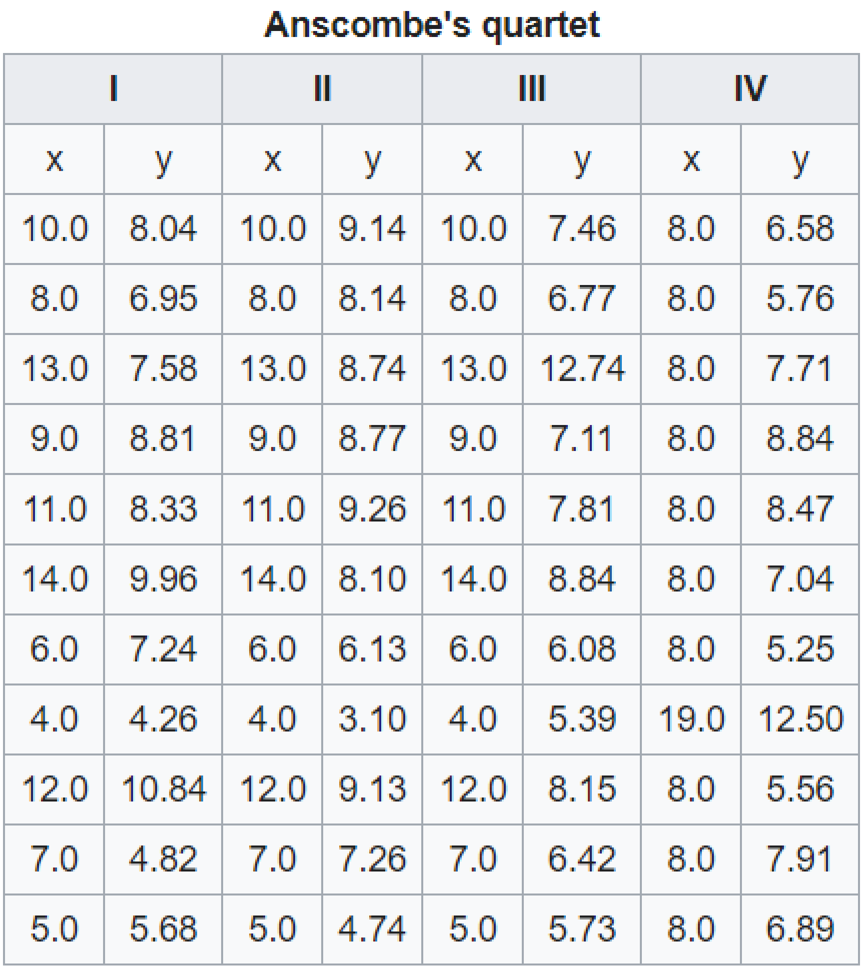
$R^2 = 0.05$

- Value of **R²** varies from 0 to 1. 1 is good fit and 0 is worst fit.
- Blue dots represent Data points
- First figure is explaining good linear regression as here RSS= 0 and all data points are on best fit line
- 2nd figure data points are around best fit line and is R² .70, which is good model
- 4th figure data points are noisier and **R²** is very less, which is not a good model.

Ques 2: Explain the Anscombe’s quartet in detail.

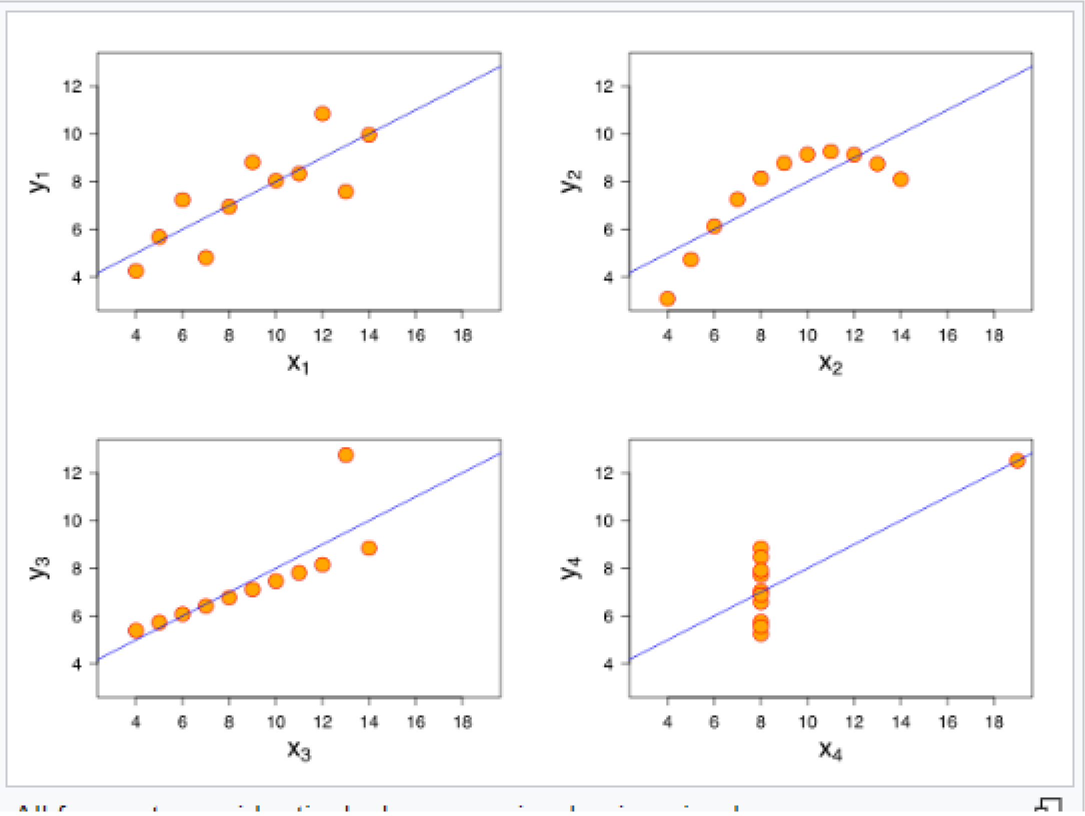
Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

He then presents a table of numbers. It contains four distinct datasets (hence the name *Anscombe’s Quartet*), each with statistical properties that are essentially identical: the mean of the *x* values is 9.0, mean of *y* values is 7.5, they all have nearly identical variances, correlations, and regression lines (to at least two decimal places).



Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



- The first scatter plot appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

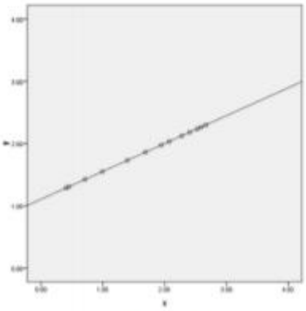
The datasets are as follows. The *x* values are the same for the first three datasets

Ques 3: What is Pearson’s R?

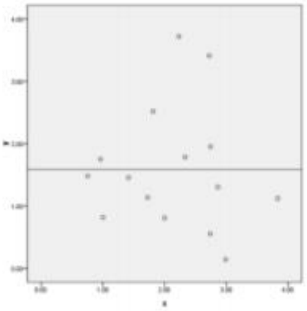
Ans: Pearson’s correlation coefficient is a statistical measure of the strength of a linear relationship between paired data, also referred as also referred to as **Pearson's *r***

- Positive values denote positive linear correlation;
- Negative values denote negative linear correlation;
- A value of 0 denotes no linear correlation;
- The closer the value is to 1 or –1, the stronger the linear correlation.

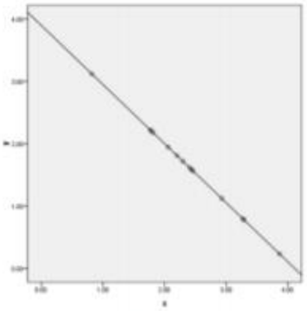
In the figures various samples and their corresponding sample correlation coefficient values are presented. The first three represent the “extreme” correlation values of -1, 0 and 1:



$r = -1$
perfect -ve correlation



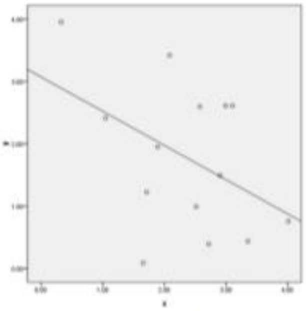
$r = 0$
no correlation



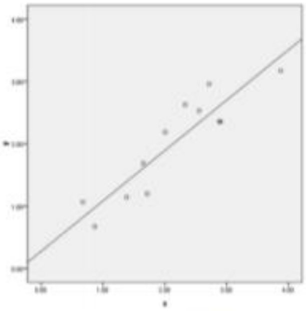
$r = 1$
perfect +ve correlation

When $r = \pm 1$ we say we have *perfect* correlation with the points being in a perfect straight line.

Invariably what we observe in a sample are values as follows:



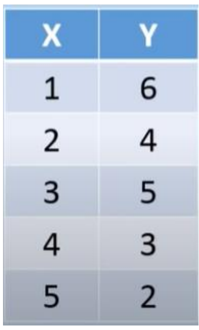
$r = -.45$
moderate -ve correlation



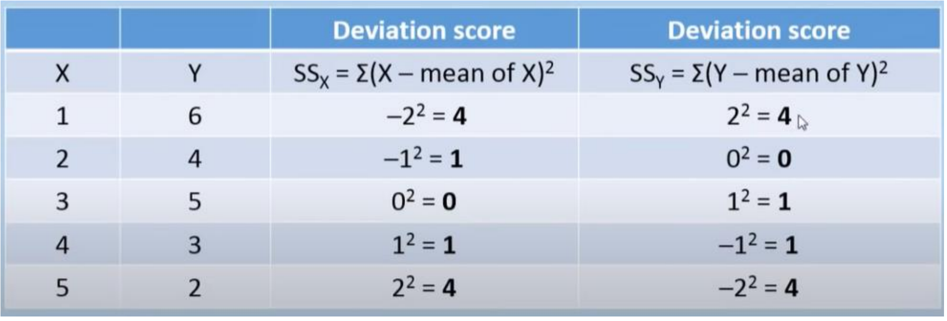
$r = .92$
very strong +ve correlation

Correlation coefficient calculation:

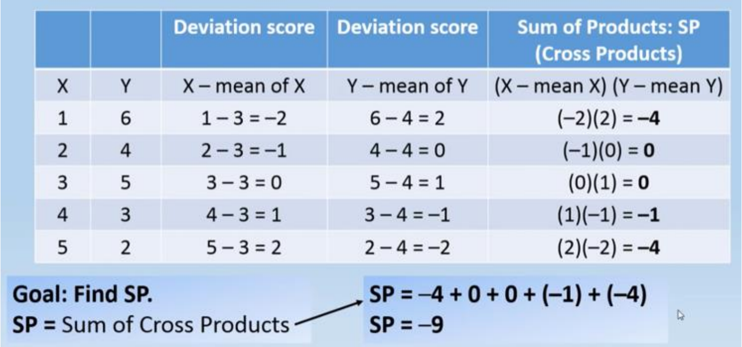
Example: We have data as below: 5 values are assigned to X and Y variable and will calculate Pearson's R



X	Y
1	6
2	4
3	5
4	3
5	2



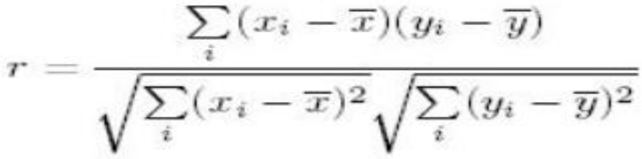
		Deviation score	Deviation score
X	Y	$SS_X = \Sigma(X - \text{mean of } X)^2$	$SS_Y = \Sigma(Y - \text{mean of } Y)^2$
1	6	$-2^2 = \mathbf{4}$	$2^2 = \mathbf{4}$
2	4	$-1^2 = \mathbf{1}$	$0^2 = \mathbf{0}$
3	5	$0^2 = \mathbf{0}$	$1^2 = \mathbf{1}$
4	3	$1^2 = \mathbf{1}$	$-1^2 = \mathbf{1}$
5	2	$2^2 = \mathbf{4}$	$-2^2 = \mathbf{4}$



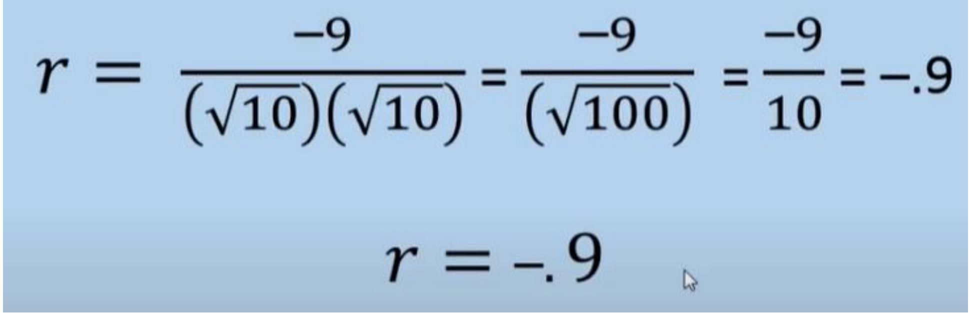
		Deviation score	Deviation score	Sum of Products: SP (Cross Products)
X	Y	X – mean of X	Y – mean of Y	(X – mean X) (Y – mean Y)
1	6	1 – 3 = –2	6 – 4 = 2	(–2)(2) = **–4**
2	4	2 – 3 = –1	4 – 4 = 0	(–1)(0) = **0**
3	5	3 – 3 = 0	5 – 4 = 1	(0)(1) = **0**
4	3	4 – 3 = 1	3 – 4 = –1	(1)(–1) = **–1**
5	2	5 – 3 = 2	2 – 4 = –2	(2)(–2) = **–4**

Goal: Find SP.
SP = Sum of Cross Products

SP = –4 + 0 + 0 + (–1) + (–4)
SP = –9



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



$$r = \frac{-9}{(\sqrt{10})(\sqrt{10})} = \frac{-9}{(\sqrt{100})} = \frac{-9}{10} = -.9$$

$$r = -.9$$

Ques 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:

Scaling: Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units

If scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Techniques to perform Scaling
Consider the two most important ones:

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

Ques 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: If VIF value is infinite then that means there is perfect collinearity. Data is having completely redundant variables.

VIF = 1 / (1 – R2) = Inf indicates R2= 1

An **infinite VIF** value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an **infinite VIF** as well).

Ques 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

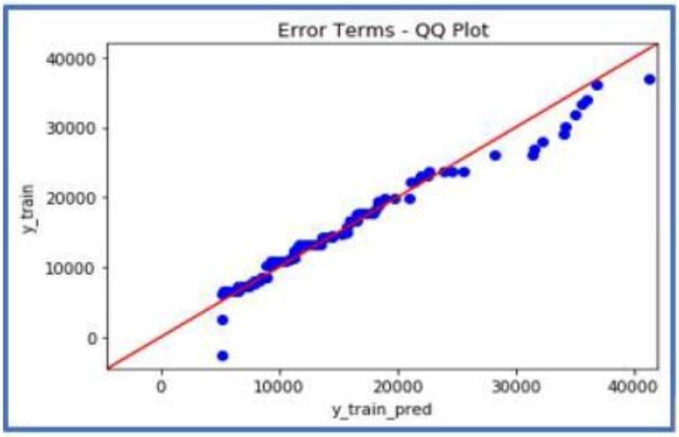
Quantile means the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

Advantage of Q-Q plot:

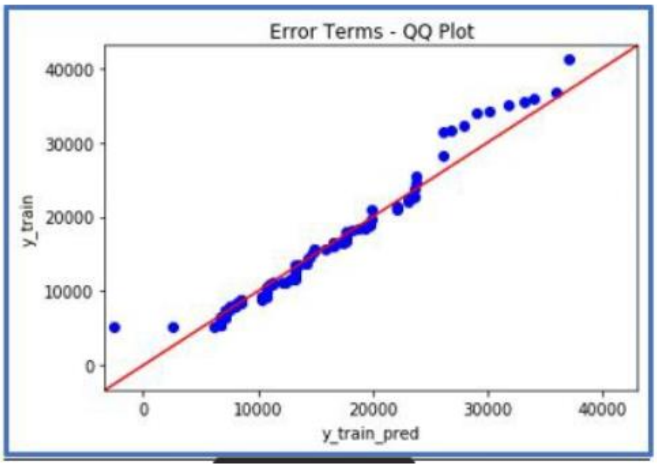
1) Sample size do not need to be equal
2) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot**.**

Interpretations:

1) **Similar distribution:** If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
2) **Y-Values< X-Values:** If y-quantiles are lower than the x-quantiles.



3) **X-Values< Y-Values:** If x-quantiles are lower than the y-quantiles.



4) **Different Distributions:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Importance of a Q-Q plot in linear regression:

Q-Q plot helps in a scenario of linear regression when we have training and test data set received separately:

1) Both the data sets are from populations with same distributions
2) Have common location and scale
3) Have similar distributional shape
4) Have similar tail behaviour