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

   Residuals are also knows as errors in linear regression.
   So,
   Assumptions of liner regression regarding residuals are:
   a) Error terms are normally distribute with mean equal to zero.
   b) Error terms are independent of each other
   c) Error terms have constant variance (Homoscedasticity)

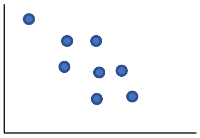
2. **What is the coefficient of correlation and the coefficient of determination?**
   Answer:

   Coefficient of correlation(r):
   a) Coefficient of correlation is the strength of the linear relationship between two random variable X and Y.
   b) It ranges from **-1 to +1**
   c) For r = -1 --->>> increase in X value leads to decrease in Y- value and vice-versa
      For r = +1 --->>> increase in X value leads to increase in Y- value and vice-versa
      For both +1 or -1 ----> X and Y are perfectly correlated
   d) For r = 0 , X and Y are not correlated

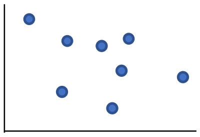
   Coefficient of Determination (R^2) :

   a) Coefficient of determination describes variance of Y which can be described by X
   b) It ranges from **0 to 1**. It can not be a negative value.
   c) **R^2 = r x r , r - Coefficient of correlation**
      R^2 - Coefficient of Determination

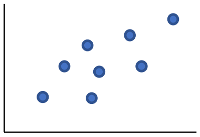
r = -0.85
R^2 = 0.72



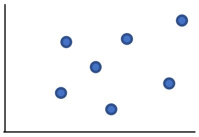
r= -0.57
R^2 = 0.32



r= +0.85
R^2 = 0.72



r= + 0.65
R^2 = 0.42



3. Explain the Anscombe’s quartet in detail.

Answer:

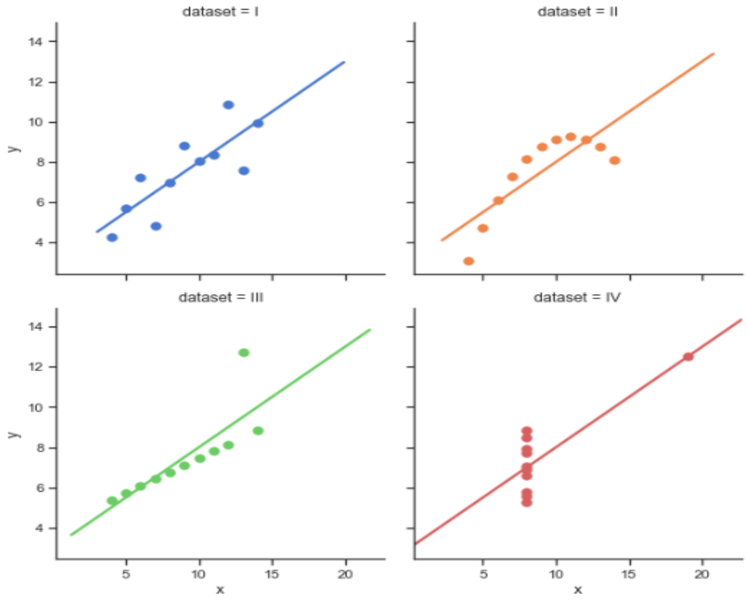
It is observed that data sets with similar summary statistics are actually not same in distribution. In this case it is very dangerous to consider only summary statistics (mean, median, mode, standard deviation, etc.) and ignore the data distribution. Because we can not determine the effect of one variable on other just from the summary statistics.

Ancombe’s Quatet is a group of four data sets which appear to be similar by a summary statistics but show different picture when plotted.

This is the group of data.

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.76	5.00	5.73	8.00	6.89

When these data are plotted : (As we can observe the difference of each data sets. Specially the distribution.)



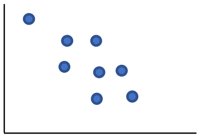
4. **What is Pearson’s R?**
   <u>Answer:</u>

Pearson’s R is also known as Pearson’s Correlation Coefficient or Coefficient of Correlation.

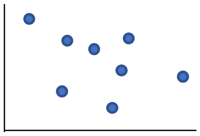
It is the measure of the strength of linear relationship between two quantitative variables, let’s say X and Y.

a) It ranges from -1 to +1
b) For r = -1 --->>> increase in X value leads to decrease in Y- value and vice-versa
   For r = +1 --->>> increase in X value leads to increase in Y- value and vice-versa
   For both +1 or -1 ----> X and Y are perfectly correlated
c) For r = 0 , X and Y are not correlated

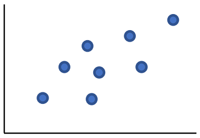
r = -0.75



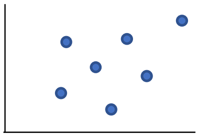
r= -0.50



r= +0.75



r= + 0.62



5. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   <u>Answer:</u>

- Scaling can be described as a process which can normalize data within a particular range.

Need of scaling:

- Let’s consider a real life scenario. If we take an example of a real estate data set. It ay contains area information(like 200 sqm,300 sqm ,etc)and number of windows (like 2 no.s , 3 no.s, etc).

- Scaling does not affect the regression model, but the coefficients are obtained from the model might be very large or very small. It would be annoying and problematic during model evaluation.

- So for better analysis and comparison scaling is required.

- Methods of scaling:

  a) Normalized scaling(MinMax Scaling):

  X(i) = $\frac{\mathrm{X(i)}-MIN(X)}{MAX(X)-MIN(X)}$, X(i) IS TO BE SCALLED i.e ith X of the data

  - ➢ MinMax Scaling brings everything into 0-1.
  - ➢ It takes care of the outliers

  b) Standardized Scaling:

  X(i) = $\frac{\mathrm{X(i)}-mean(X)}{SD(X)}$, X(i) IS TO BE SCALLED i.e ith X of the data

  SD – Standard Deviation

  - ➢ It scaled the data into a new set with mean = 0 and SD = 1

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

VIF – Variance Inflation Factor describes how one predictor variable can be described by other predictor variables. i.e. if R^2 is more , this variable / feature is correlated with other features.

VIF = $\frac{1}{1-R^2}$,

------ > R^2 = 1 ->>> VIF = infinity , Here there is a perfect correlation (Variables are highly correlated)
--------> R^2 = 0 ->>> VIF = 1, Variables are orthogonal to each other