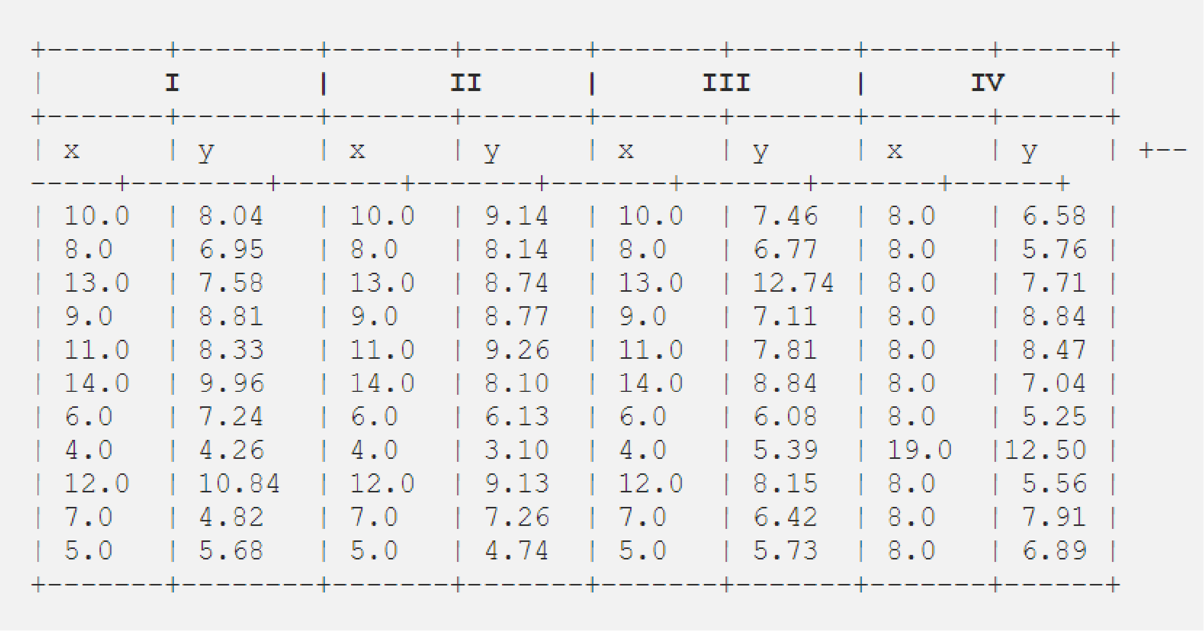
Q-1) Explain the Anscombe's quartet in detail .

Ans) There was great statistician name Francis John Anscombe, who was teaching statistics to its students where he observed a very dangerous pattern among the people of not plotting the visualizations of the data set instead they relied heavily on the descriptive statistics result which are based on sum metric's like mean, standard deviation etc. because according to those new learners they believe that numerical calculations are always correct.

But Francis Anscombe wanted to break this misconception he tried explaining those new learners as well as to the seniors statistician, but all of ignored them completely.

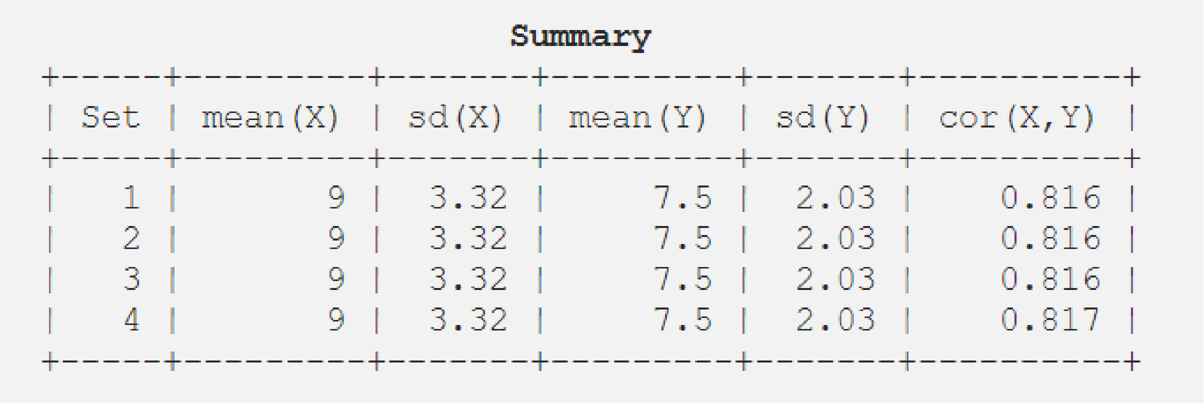
To proof his statement that "visualizations too is important along with the numerical summary statistics of the dataset".

He then came out with the four datasets which has exactly similar numerical summary statistics figures or values which as below:-



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

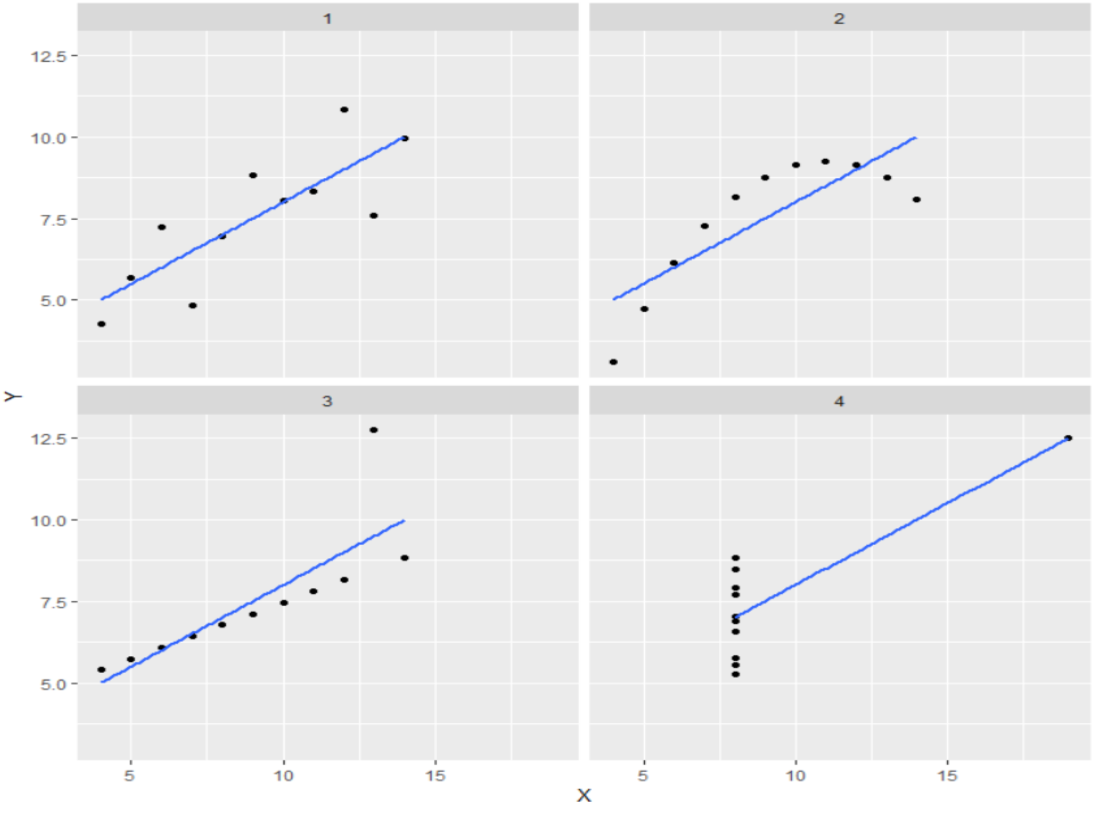
He presented the above dataset to the senior statistician and asked to analyze it but to the fact that these so called senior statistician where stunned to see the same results.



Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

On seeing them stunned he asked them to plot the above graphs of the above dataset, when they plotted the graph they were shocked to see all the graphs different story of the dataset, have a look upon the plots of the data set:



Anscombe's quartet — Plot

#) The first scatter plot (top left) appeared to be simple linear relationship, corrosponding to 2 variables correlated & following the linear assumption of normality.

#) The second scatter plot (top right) fits a neat curve but doesn't follow a linear relationship.

#) The third scatter plot (bottom left) exhibits the linear relationship among the data poins but the presence of the outlier has signficant impact on the model, if outlier was removed then might have probably fit the linear model.

#) The fourth scatter plot (bottom right) does not fit any kind of linear model, but the presence of single outlier keeps it from going off.

Q-2) What is Pearson's R ?

Ans)

#) Test statistics that measures the statistical relationship, or association between two continuous variable.

#) known as the best method of measuring the association between variables of interest because it is based on the method of covariance.

#) It gives information about the magnitude of the association, or correlation as well as the direction of the relationship.

#) The value of the pearson’s correlation coefficient varies between +1 & -1.

#) A value of +1 & -1 indicates a perfect degree of association between 2 variables.

#) As the correlation coefficient value goes towards 0, the relationship between 2 variables will be weaker.

#) The direction of relationship is indicated by the sign of the coefficient, +ve indicates positive relationship & -ve indicates negative relationship.

Assumptions :-

1) For Pearson’s correlation, both variables should be normally distributed i.e normal distribution describes how the values of a variable are distributed.
2) There should be no significant outliers. Including outliers in the analysis can lead to misleading results as Pearson’s correlation coefficient is very sensitive to outliers, which can have large effect on the line of best fit.
3) The 2 variables have a linear relationship.
4) The observations are paired observations i.e. for every observation of independent variable there should be dependent variable.
5) Homosedasiticity means equal variances i.e datapoints should be present on both sides of the line of best fit.

Q-3) What is scaling? Why is scaling performed? What is the difference between normalized & standard scaling ?

Ans)

#) It is step of data preprocessing which is applied to independent variables of features of the data. It basically helps to normalize the data within the particular range.

#) It is performed to handle highly varying magnitudes or values or units.

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

#) Example :- If an algorithim is not using feature scaling method then it can consider the value 3000 meter to be greater than 5km but that’s actually not true & in this case the algo will give wrong predictions so we do feature scaling to bring all the values to same magnitude and scale.

There are **two most important feature scaling techniques:**

1) Normalized scaling
2) Standard scaling

Normalized scaling :- It is a technique in which values are shifted and rescaled so that they end up ranging between 0 & 1. It is also known as min max scaling:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

X_{max} & X_{min} are the min & max function of the feature

When X value is minimum in the column, numerator will be 0 hence X` will be 0.

When X value is maximum in the column, numerator will be equal to denominator & value of X` is 1.

When X value is in between maximum & minimum then value of X` will be in between 0 & 1.

Standardized scaling :- is the technique where the values are centred around mean with a unit standard deviation. This means that the mean attribute becomes zero & the resultant distribution has a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

#) µ is the mean of the feature variable,

#) σ is the standard deviation of the feature values.

#) It is good practice to fit the scaler on the training data & then use it to transform the testing data. This would avoid any data leakage during the model testing process. Also, the scaling of the target values is generally not required.

Q-4) You might have observed that sometimes the value of VIF is infinite. Why does this happen ?

Ans) In VIF each feature is regression against all other features. If R^2 is more which means this feature is correlated with other features.

VIF = 1 / (1- R^2)

When R^2 reaches 1 VIF reaches infinity.

Q-5) What is Q-Q plot ? Explain the use & importance of a Q-Q plot in linear regression ?

Ans)

#) The quantile – quantile (q-q) plot is a graphical technique to determine if two data set come from populations with a common distribution.

#) It is a plot of quantiles of the first data set against the quantiles of the 2nd data set.

The advantage of the q – q plot are :

1) The sample size do not need to be equal.
2) Many distributional aspects can be simultaneously tested.

Importance :-

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so then location & scale estimators can pool both data sets to obtain estimates of the common location & scale. If two sample do differ, it is also useful to gain some understanding of the differences. The q – q plot can provide more insight into the nature of the difference than analytical methods such as chi – square etc sample tests.

Q-6) Explain linear regression in detail ?

Ans) It is used for finding the relationship between target & predictor varible.

#) 2 kind of linear regression :- simple linear regression & multiple linear regression.

#) Simple Linear regression :- useful for finding relationship between 2 continuous variable.

#) One is predictor/independent variable & other is dependent/target variable.

#) The main idea is to obtain the best fit line where the prediction error is as small as possible.

#) Residual/Error is the difference between the predicted values which lies on the best fit line and the actual values of the dependent variable.

Equation is :- y^ = $b_o + b_1x$ where b_o -> intercept & b_1 -> slope

Y -> target / dependent variable; x -> independent / predictor variable.

#) Residual sum squares = $(y_i - y_{pred})^2$

#) By minimizing the RSS we will get the best fit line. (RSS is also known as the cost function)

#) To determine the strength of linear regression model there are 2 methods :-

1. R2 or coefficient of determination
2. RSE (Residual standard error)

#) R2 gives the measure of how well the data variation is explained by the relationship between variables. It varies between 0 & 1. Higher the R2 better the model.

#) R2 = 1 – (RSS / TSS), where RSS -> Residual sum of square, TSS -> Total sum of square.

#) TSS is sum of errors of the data points from the mean of response variable.

@) Assumptions in Linear regression :-

1. Linear relationship between independent & dependent variable.
2. Error terms when plotted should be normally distributed.
3. Error terms should be independent of each other.
4. Homoscedasticity should be present.

@) Parameters to assess the model are :-

t statistics gives the significance of the coefficient

f statistics gives the significance of the model.

r squared gives to what extent is the model fit.