Learning Analytics Tools

Professor Ramkumar Rajendran

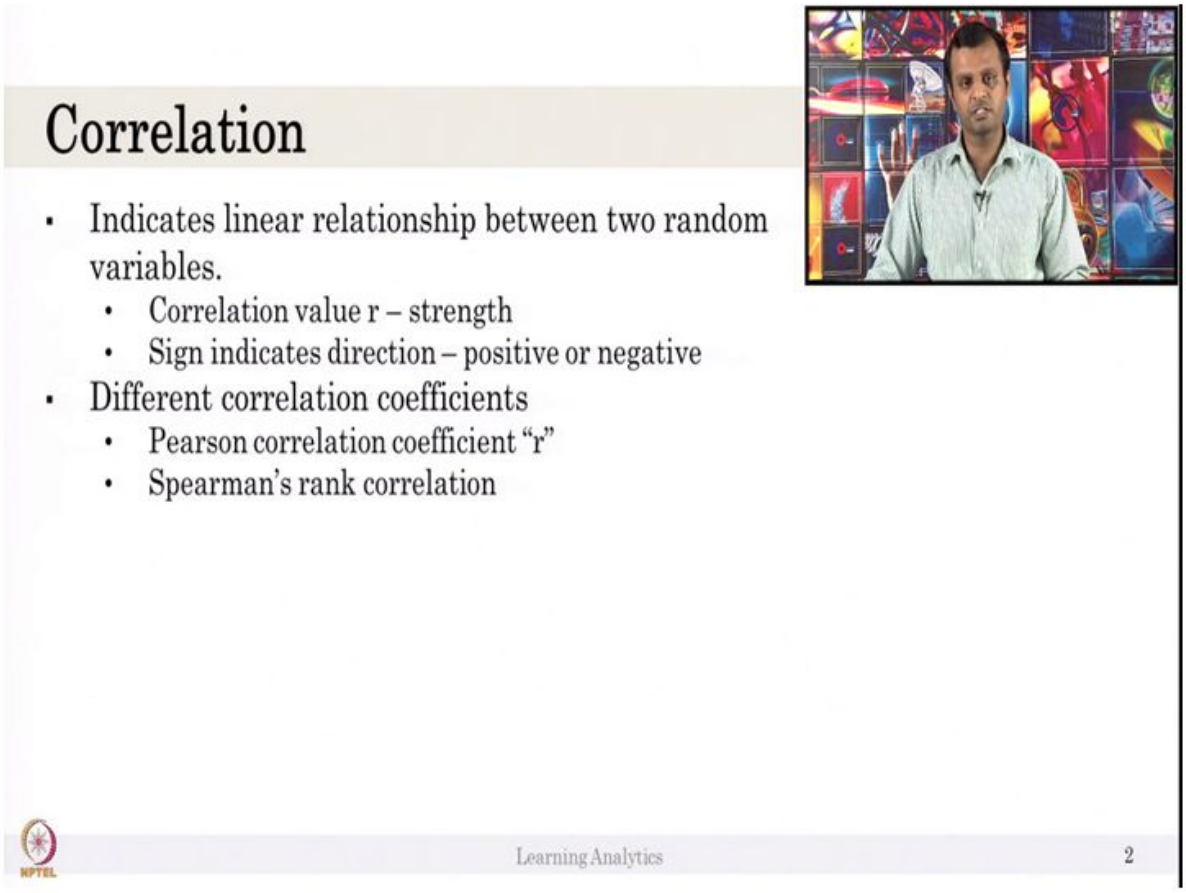
Educational Technology

Indian Institute of Technology, Bombay

Lecture 5.2

Correlation

(Refer Slide Time: 0:23)



In this video, let us talk about correlation. So what is the correlation? The correlation indicates the linear relationship between two random variables, what is the relationship between variable

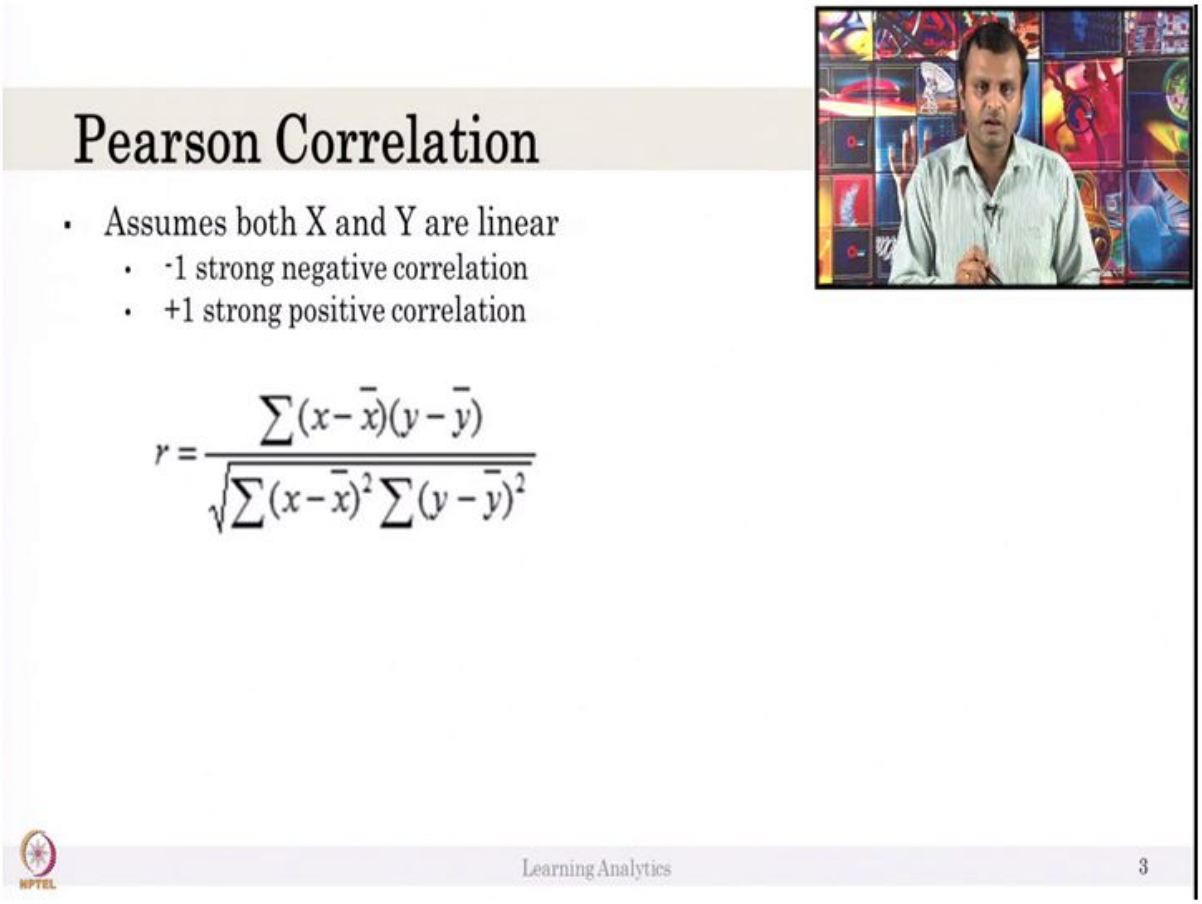
A and variable B. Why is it linear, that is a common correlation coefficient to use, that is why it is called a linear relationship. Let us look at other correlation types also.

So the correlation value(r) is to indicate the strength, so there is a value of r, correlation coefficient value equal to say 0.5 indicates if x varies, y varies in the magnitude of 0.5, the strength of the variations. Also, the sign indicates which direction the correlation is, is it direct or inverse. For example, if it is positive, r equal to plus 0.5 which indicates if x increases y also will increase.

If r is negative indicates both are inversely dependent. So, what happens, if x increases y will decrease, so negative indicates inverse proportional. So there are different correlation coefficients.

The first is the Pearson correlation coefficient, it is very common, used widely in excel or now the libraries too. So the definition we gave above is for the Pearson correlation coefficient. There is another correlation coefficient called the Spearman's rank correlation or Kendall's other correlation coefficient available.

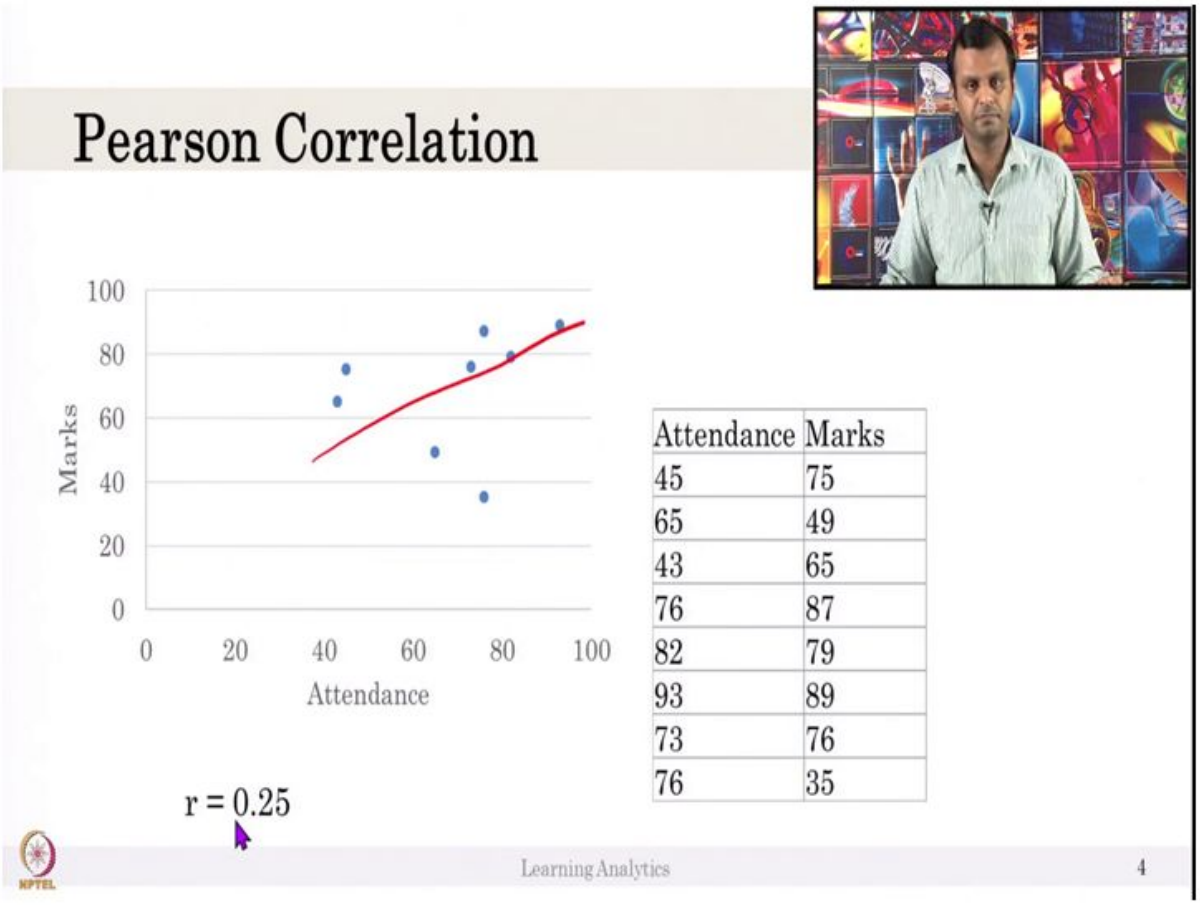
(Refer Slide Time: 1:47)



Let us look at the Pearson correlation in this video. Assume that both X and Y are linear, this is very important, this assumption that X and Y are having a linear relationship. Pearson correlation varies from -1 to +1 value. It indicates a strong negative correlation if it is -1 and +1 indicates a strong positive correlation. If it is 0, there is no correlation absolutely.

So, the Pearson correlation coefficient is computed using this formula. I thought of explaining this formula, but it is not needed. This formula can be simplified and you can compute the Pearson correlation coefficient easily using different tools available or free tools in the web.

(Refer Slide Time: 2:33)



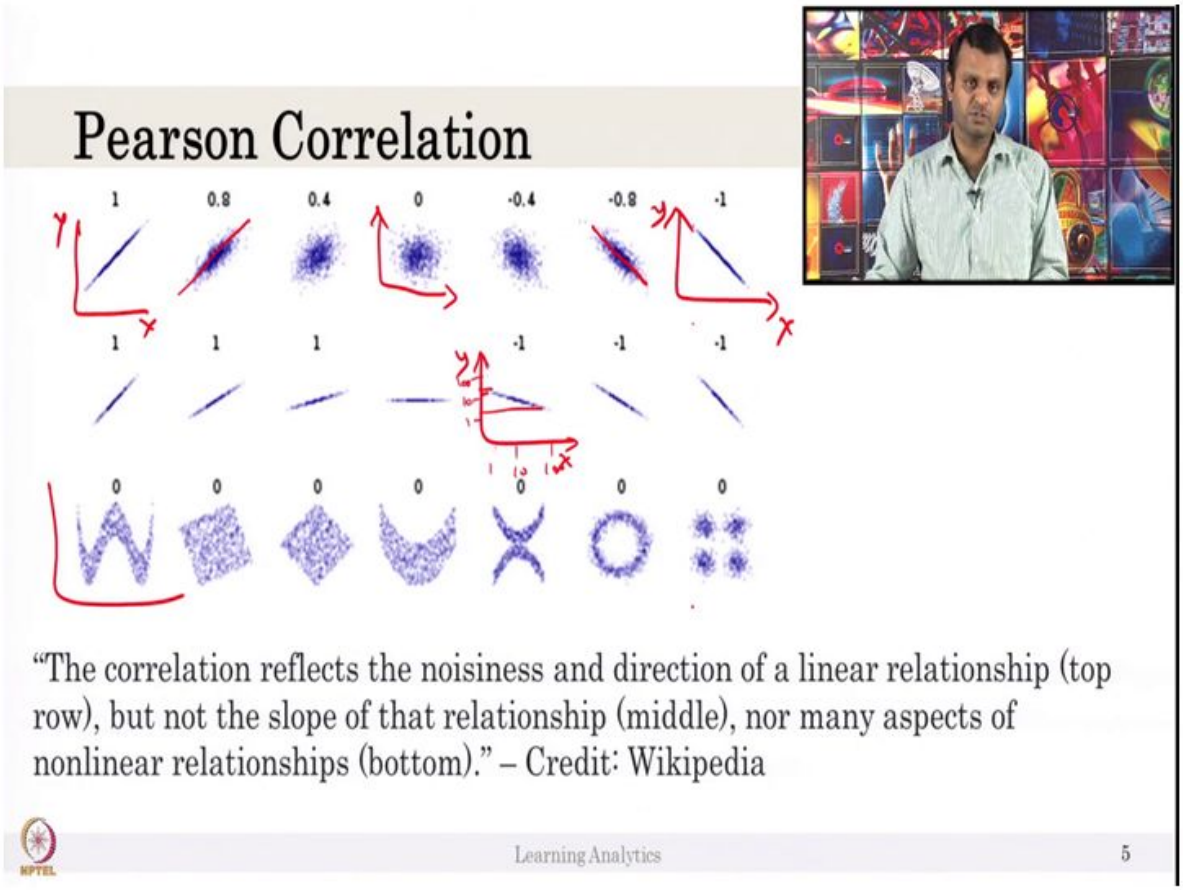
Attendance	Marks
45	75
65	49
43	65
76	87
82	79
93	89
73	76
76	35

Let us look at the Pearson correlation. Let us look at the Pearson correlation coefficient. If you remember, we had marks and attendance of 60 students. Say, we have some eight students data here and we have plotted the attendance and the marks.

Do you see there is a correlation between attendance and marks, it seems to be, right? So, if the attendance increase, somehow the performance increases except for these two values. This can be, like we do not know why it happened, but seems like there is a relationship. So the correlation coefficient will be positive because if attendance increases the marks increase, so it is a positive correlation coefficient.

It is a kind of, it is a positive correlation coefficient and it is strong, it is not very weak because the line is actually trying to match all the points. So, the value is 0.25, that is because we have two values which are well below this line, but it is a positive, okay? So +0.25.

(Refer Slide Time: 3:47)



So, this figure is from Wikipedia and let us look at this figure to understand what is the Pearson correlation coefficient. +1 indicates there is a strong positive correlation. If X increases Y also increases definitely, so if X increases, Y increases. -1 indicates a strong negative correlation. When X increases Y decreases and this is a noisy data but it is still fitting around the line. The value is -0.8. This value 0.8 is a very strong correlation by the way.

So, 0.4 is the weak correlation coefficient, that is not really a good correlation, but 0.4 is weak correlation and 0 indicates there is no correlation absolutely, there is no correlation between X and Y. We cannot say if X increases, Y also increases because you see here X is very low and Y is high and X is very high here, still Y is low, so we cannot say any definite correlation between these two values.

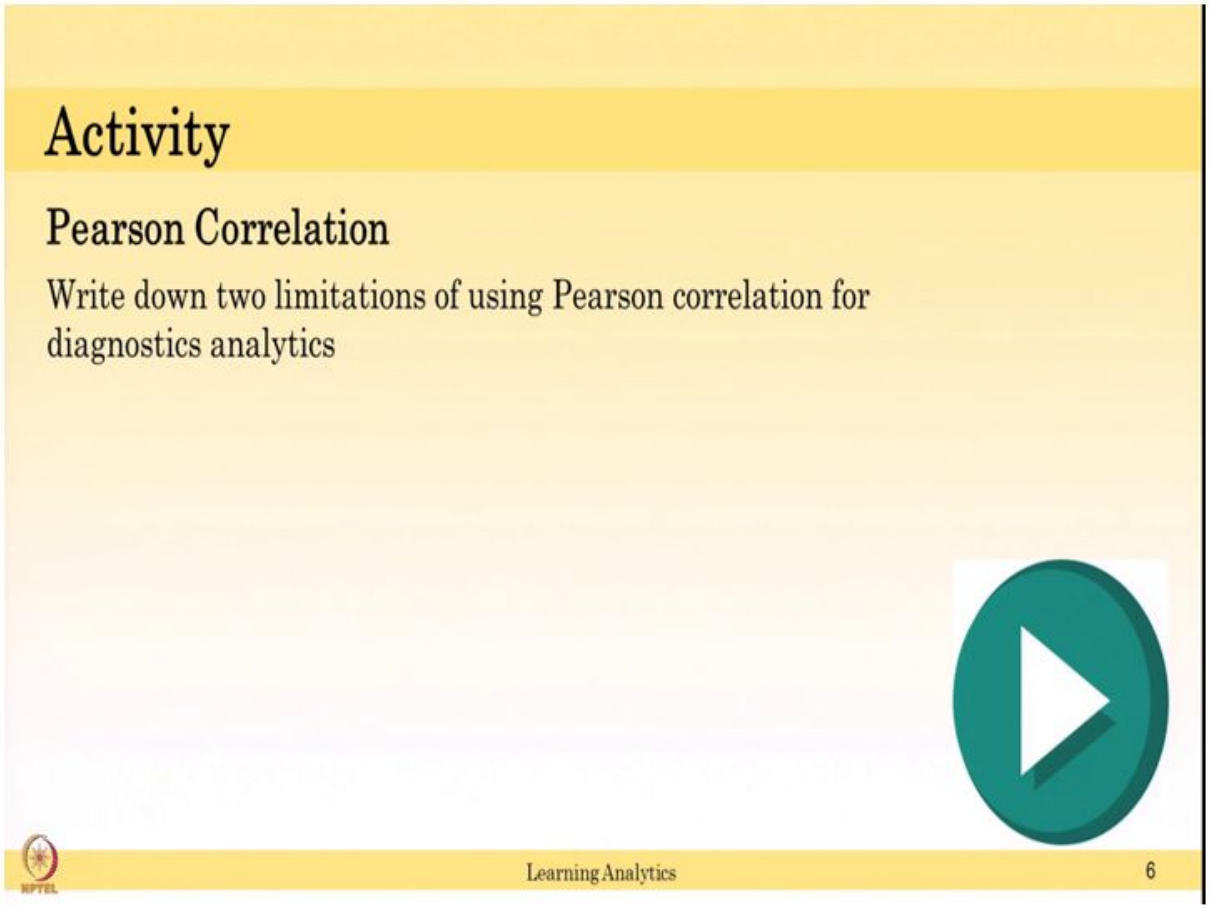
Let us look at this line, it is very interesting. So, all these values you know, except this, all these are +1. So, we saw that +1 indicates a positive and strong correlation, but it is not telling anything about the slope, that is very very important. So -1 does not say anything about slope. For example, in negative correlation, if X increases in the magnitude of say 1, 10, 100, Y might increase, Y might decrease by say 8, 220 or something like that, so not in the same scale, right? That is what it indicates.

So, Pearson correlation will not tell about the slope, it tells it is positive, whether the direction is positive and it is strong or not strong, that is it, magnitude is what way is it going, not "strong" or "not strong".

Let us look at the last row, it is very interesting this row. So definitely all of these correlation coefficients are zero and it indicates that there is no linear relationship between X and Y, but there is a relationship which is nonlinear. For example, there is a good relationship with this particular value, this might be high, there is some nonlinear relationship is existing in the system, the Pearson correlation coefficient may not tell this relationship because it has no idea what is the relationship between X and Y, but there is a nonlinear relationship axis between X and Y, there is some kind of relationship.

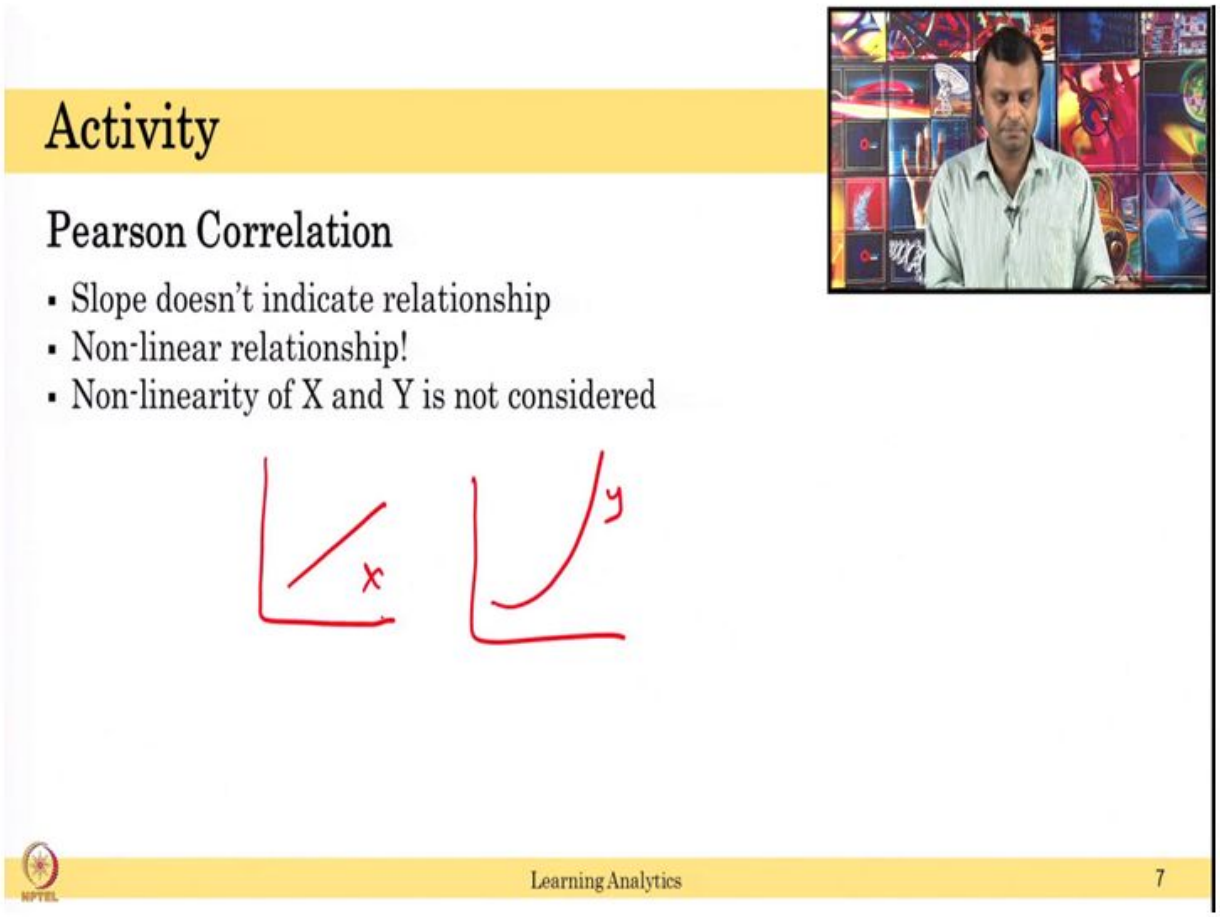
If you apply some predictive analytics definitely it will classify very simply using this data because there is a nonlinear relationship, but Pearson correlation coefficient cannot tell this, that is how the graph. So, this chart is very important to understand Pearson correlation coefficient, also it's drawbacks, for example, the slope is not indicated in the Pearson correlation coefficient value, also the nonlinear relationship between X and Y is not captured in the Pearson correlation coefficient.

(Refer Slide Time: 7:25)



So, you saw what is Pearson correlation coefficient and you saw the range of its value, +1 to _1 and what plus one indicates and you saw what are the limitations, so can you write down two limitations that we discussed in the last slide, using Pearson correlation coefficient for diagnostic analytics. So, yes, please write down two limitations after writing it on, please continue.

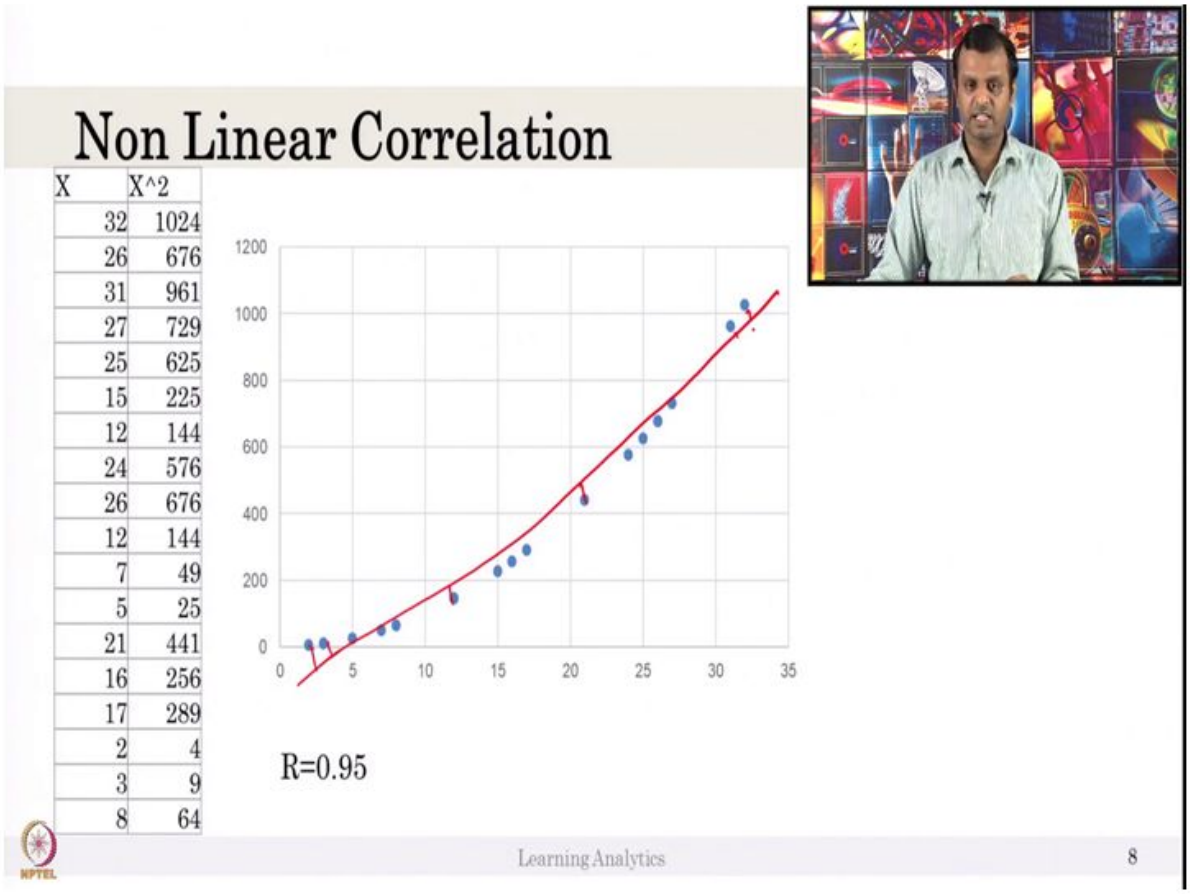
(Refer Slide Time: 7:53)



So, the slope is not captured in the Pearson correlation coefficient, also nonlinear relationships are not captured and non-linearity of X and Y is not considered. For example, X can be a linear relation, but Y is not a linear variable. For example, we will discuss this in detail in the next video.

Pearson will tell you whether if X increases, Y increases or not, we will discuss this in detail with an example in the next video.

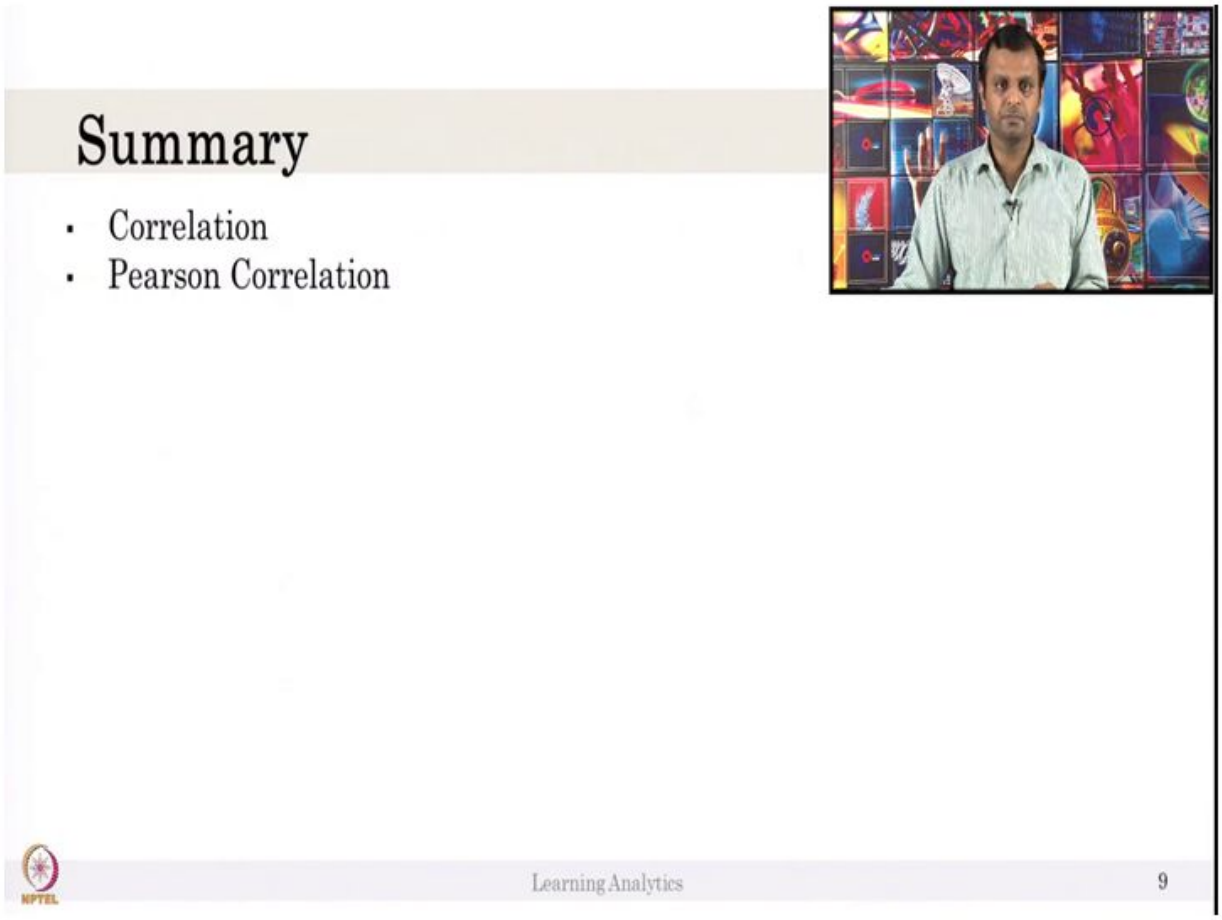
(Refer Slide Time: 8:54)



So, let us see what is a nonlinear correlation. If you see this X and x^2 if I plot this value, X versus x^2 you know that x^2 is increasing, so that Y value is actually like this. It is not a linear relationship, but the correlation coefficient is 0.95 the reason is there is a relationship between X and Y. If X increases, the Y increases and if X and Y increase there is a strong correlation, a positive correlation coefficient.

In general, it should be 1, you know. It should be 1, it is an exactly good correlation between X and Y, but Pearson cannot find that, Pearson's formula. Hope you understand what I mean by Pearson do not check the linearity of both X and Y, so Y can be a nonlinear function available, still, it cannot identify that relationship.

(Refer Slide Time: 10:47)



So, in this video we talked about what is correlation and what is Pearson correlation and what are the drawbacks in Pearson correlation, thank you.