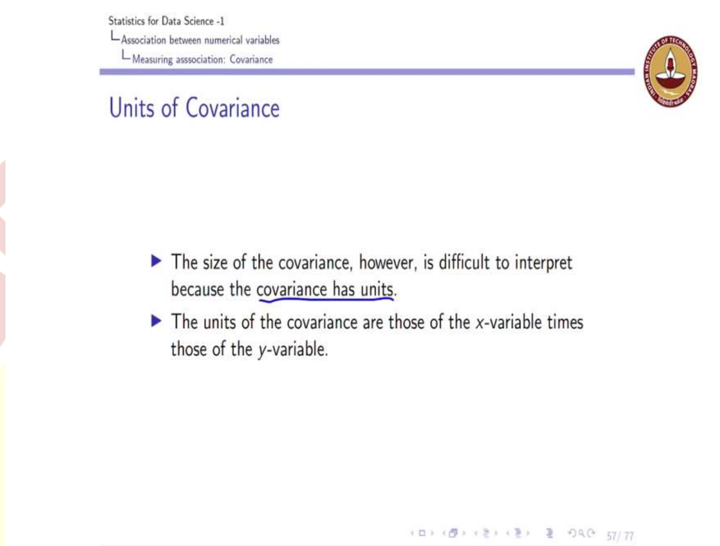
IIT Madras

ONLINE DEGREE

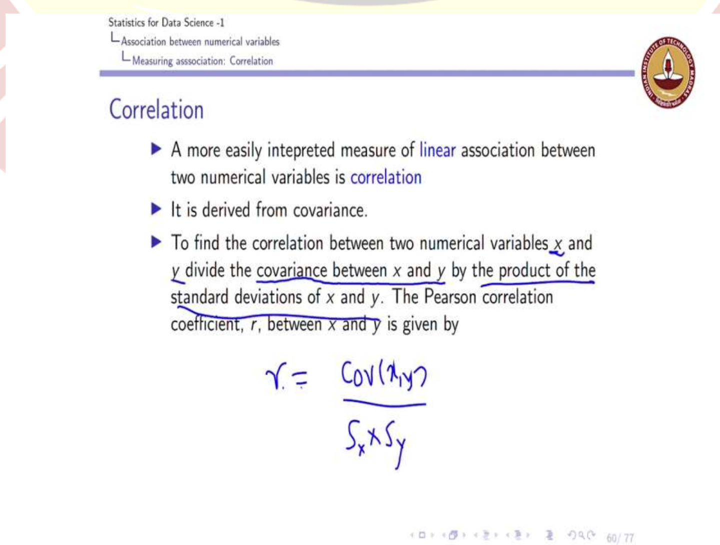
Statistics for Data Science - 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Management, Madras
Lecture 4.7
Association between two numerical variables – Correlation

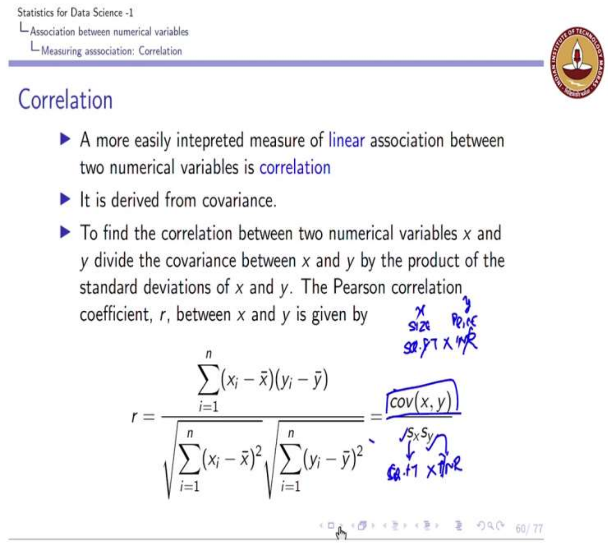
(Refer Slide Time: 0:24)

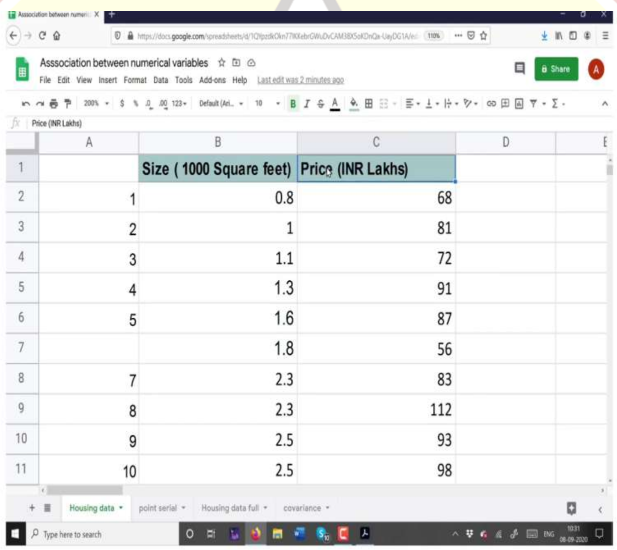


So, we understand now another measure of association again when I say association, I mean linear association between two numerical variables. You already seen the measure of covariance, but recall when we talked about covariance, we said that the covariance is difficult to interpret, because the covariance has units.

(Refer Slide Time: 0:41)







	A	B	C
1		Size (1000 Square feet)	Price (INR Lakhs)
2	1	0.8	68
3	2	1	81
4	3	1.1	72
5	4	1.3	91
6	5	1.6	87
7		1.8	56
8	7	2.3	83
9	8	2.3	112
10	9	2.5	93
11	10	2.5	98

So, we are going to another measure of association and this is what I term, the correlation between two numerical variables. What is a correlation? It is a, again it is another measure of linear association between two numerical variables, it is derived from the concept of covariance, we have already introduced the concept of covariance. Now, how do I find the correlation between two numerical variables, let me call those two numerical variables x and y.

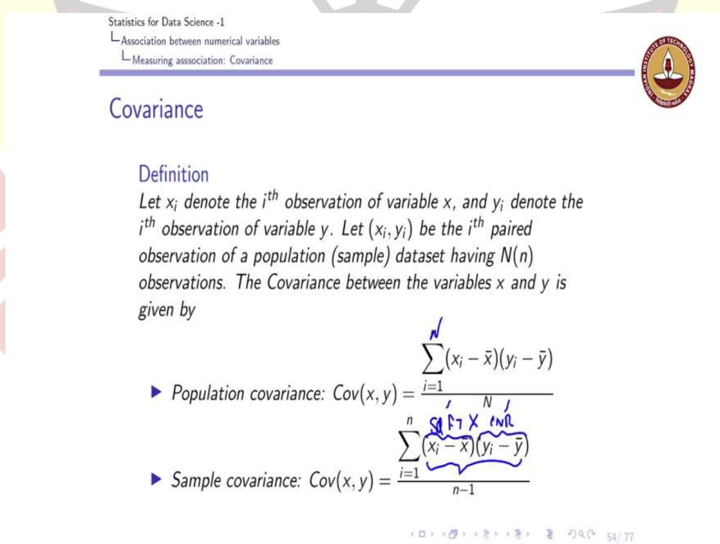
Again going back to our example, you can see that I have this example here, where I have size and price of a house, there are 15 observations. So, if you can look at this, let me zoom it a bit more. So, you can see that I have the size and price. So, x is my explanatory variable, which is the size of the house, y is my response variable which is the price of the house.

So, if I want to know the correlation between these two variables, I can represent it by r or some books represented by ρ. The Pearson correlation coefficient r between variables x and y, r_{xy} I can drop this x and y also. And just represent it by r, the Pearson correlation coefficient is given by r denote the covariance between x and y divided by the product of the standard deviations of x and y i.e. $r = \frac{Cov(x,y)}{s_x s_y}$

Since, I have already said that it is derived from the covariance, a more formal way of defining the correlation is the following, $r = \frac{\sum_{i=0}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=0}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=0}^{n}(y_i-\bar{y})^2}}$. This is my n here, I am summing it up in the case of my example, n equal to 15. Because I have the data on 15 homes I am looking at the sizes and prices of these 15 homes I divided by the square sum of deviations, this is the sum of the, this is the square deviation sum of square deviations of x and this is the sum of square deviations of y.

In other words, I also have that this is equivalently I can write that $r = \frac{cov(x,y)}{s_x s_y}$. So, this is how we compute our correlation between two numerical variables. Now, why is it important? Now, when you look at this term cov(x, y).

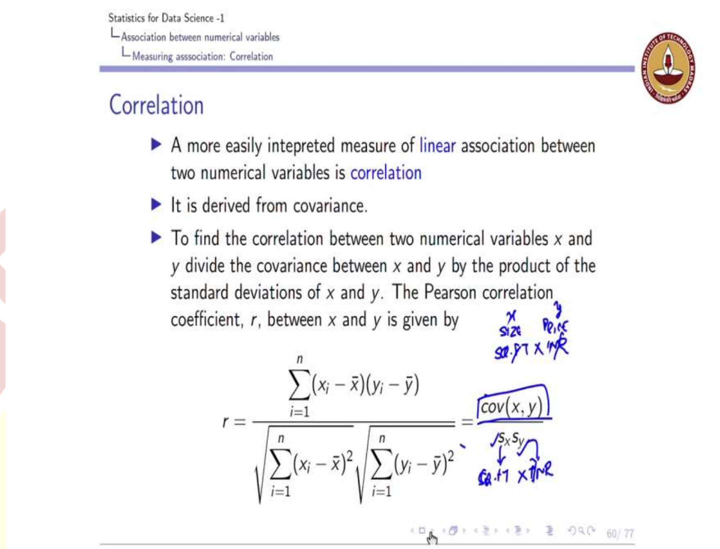
(Refer Slide Time: 4:21)



I know that the covariance term, again go back to the where we defined a covariance, we saw that when we looked at the covariance term, this is how we defined a covariance term, we looked at the deviation. Now, the deviation for example, if x is measured in terms of square feet, the deviation is also going to have the unit of square feet, y is measured in terms of INR.

The deviation also is in terms of the currency which is in lakhs of rupees. So, this correlation covariance term has a unit which is the product of the unit of this variable and this variable which is square feet into Indian National Rupees.

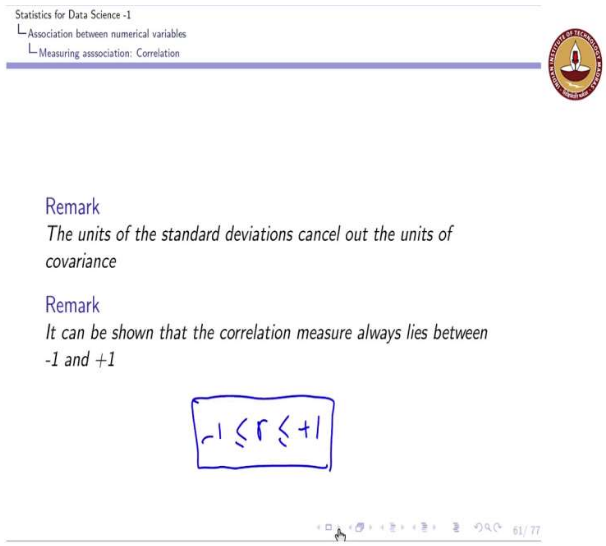
(Refer Slide Time: 5:14)



Now, when I look at the correlation for the same variable, I have defined the correlation as the following. Covariance, now, what are the units of covariance, covariance in the example, the units of this covariance is square feet, which is the unit of size which is my variable x, and then I have which is INR in lakhs of rupees, which is the unit of price, which is my response variable y.

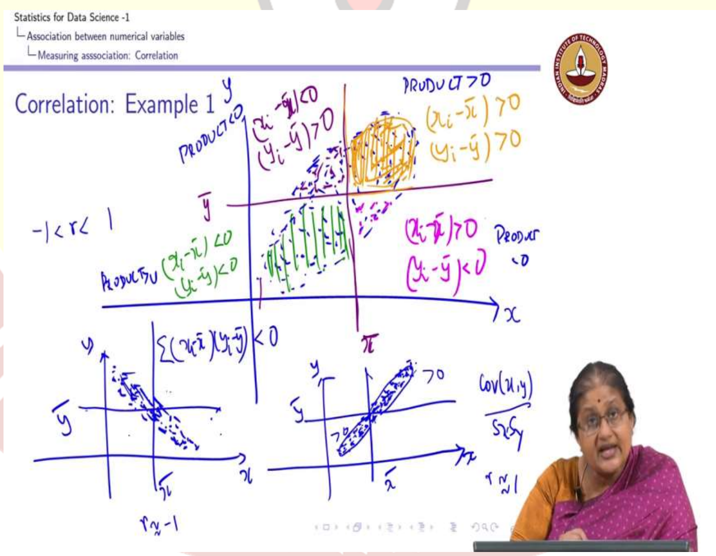
Now, if you look at the standard deviation of x, standard deviation of x is also going to have the same units as that of x, standard deviation of y is going to have the same units as that of y, which is in Indian National Rupees. So, you can see that the units cancel off when I talk about a correlation measure.

(Refer Slide Time: 6:09)



So, this measure of correlation is a unit less quantity. But, we need to also remember that it is a measure of linear association, it can be shown that this correlation coefficient always lies between +1 and −1. So now, we have a covariance measure and a correlation coefficient.

(Refer Slide Time: 6:46)



Now, how do we use this correlation coefficient to tell about the strength of the association between my variables? So, recall when we looked at the covariance matrix, we started by what we said was a scatter plot matrix. In the scatter plot matrix, I have my explanatory variable, which is on the x axis, I have my response variable which is on the y axis.

So, if I have a scatter of this kind between these two variables, I am just looking at one of the quadrant, but I could have in both the quadrant but I am just for exposition sake, I am just looking at this quadrant. And suppose I have my $\bar{x}$, so x varies from this point to this point. So, suppose this is my $\bar{x}$ that is the mean of my first variable.

And suppose this is my $\bar{y}$, suppose this is my $\bar{x}$, and this is my $\bar{y}$, we further said that we could split this entire scatter into four pieces. In this piece, if I am going to have all my points orange. This is this piece, that is where my, so if I am going to share this region or this points, so all the points in this region are basically points where my $x_i - \bar{x}$ and my $y_i - \bar{y}$ are both greater than 0 because both all the points are greater both my x point and my y point are greater than the respective means.

Now, in this green area, green shaded area, my $x_i - \bar{x}$ and my $y_i - \bar{y}$ are both less than 0 because both the x_i points and the y_i points are less than their respective means. Now, if I look at the purple area, which is this scatterplot, I find the y_i's are greater than the $\bar{y}$, whereas my x_i is less than the $\bar{x}$. And in this final area, which is say, purple the smaller area here, I have here $x_i - \bar{x} > 0$, whereas $y_i - \bar{y} < 0$.

Now, let us understand why this correlation metric or this correlation coefficient becomes very important. Now, suppose I have a dataset of this kind where I have a tightly clustered data. So, remember when we talked about association, we wanted to see the direction. And whether it is tightly clustered, whether there is a variability all these points, suppose it is a tightly clustered data, for the same data point, if I continue with the exercise I have done before, this is my $\bar{x}$, this is my $\bar{y}$.

So, you can see that, in this I have very little points which are here, but majority of the points lie in this area, and in this area, the product. So, when I look at the product of the deviation, it is greater than 0 in this quadrant, the product of the deviation is again greater than 0 in this quadrant, hence the sum of the product of deviation, which is $\sum(x_i - \bar{x})(y_i - \bar{y})$, because a product is always greater than 0, the sum is going to be greater than 0. Hence, this type of a pattern will always result in a positive covariance.

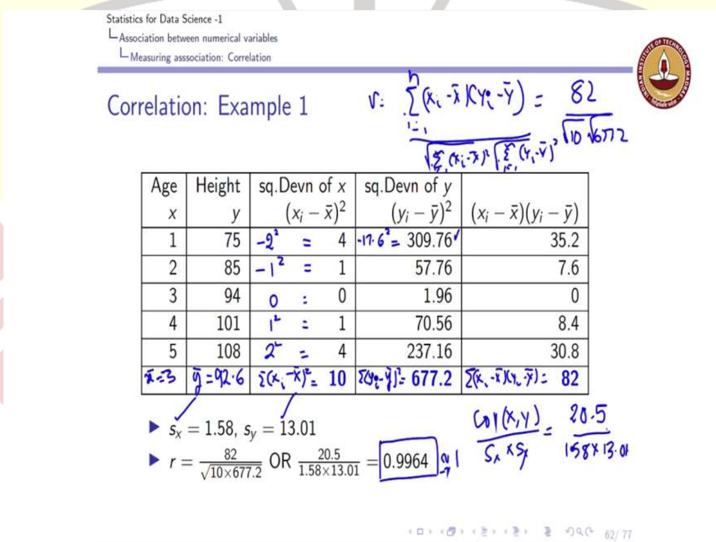
Whereas if I have my data, which is of this kind now, I have my x here, I have my y here, and my data is of this kind. Suppose, this is my scatterplot, again this is my $\bar{x}$, this is my $\bar{y}$, this is my x, this is my y, this is my $\bar{x}$, I have the same $\bar{y}$, now if you notice, this scatterplot, you can see that the product here I have a product.

So, if I look at the product of these coefficient, the product of the deviations is greater than 0, here the product is going to be greater than 0 in this quadrant, here the product is going to be less than 0. Here, also the product is going to be less than 0. So, what we notice here is the product of the deviation.

So, $(x_i - \bar{x})(y_i - \bar{y})$ in this portion is going to be less than 0, in this portion is also less than 0, I do not have any points in these two areas. So, $\sum(x_i - \bar{x})(y_i - \bar{y}) < 0$. And this resulted in a negative covariance measure.

Now, if I am going to divide them by their respective standard deviations, I get my correlation measure. So, the correlation measure always is between -1 and +1. How do I interpret this correlation measure? If there is a, so, if the data is of this kind, my covariance I know is going to be positive I divided by the standard deviations, and I see that if my data is of this kind, then my correlation measure is very close to 1. Whereas, if my data is this kind, my correlation measure is very close to -1, so this is a perfect linear relationship between x and y in the positive direction. This is a perfect linear relationship between x and y in the negative direction.

(Refer Slide Time: 14:02)



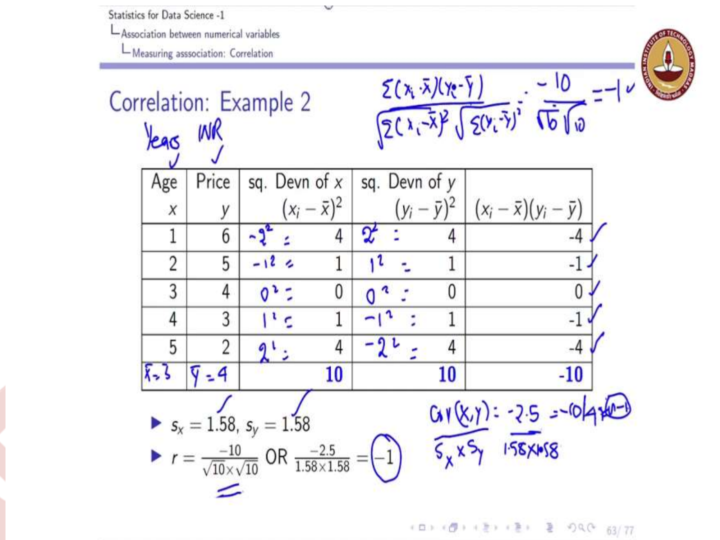
Before I go to this dataset, let us work on a small hypothetical example. Again, I had x, which was my age, y, which was the height I wanted to explain whether as people grow older, their heights increase or decrease, I wanted to know what is the association, so we found out what was the deviation, so I knew the mean here was 3, we computed the mean here, this is what we already did. This is what we did in our, so my mean was a 3, and the mean here was a

92.6, so the deviation was 1 - 3, which was - 2, the deviation was 2 - 3 a - 1, 4 equal to -2^2 , -1^2 is 1, so this is the sum or the square deviation is what is given here.

Similarly 0, 1, and 2, this is giving me the square deviation, similarly 75 - 92.6; I also computed that as -17.6. - 17.6^2 is equal to 309.76. And this is giving me the square deviation of each of my y observation. This is the sum of square deviations. This is the sum of square deviations of y. This is the sum of the product of the deviations, so if I apply my first formula, the correlation coefficient $r = \frac{\sum_{i=0}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=0}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=0}^{n}(y_i-\bar{y})^2}}$. Sum of total sum is equal to 82, divided by $\sqrt{10} \times \sqrt{677.2}$.

Using the second formula, $r = \frac{Cov(x,y)}{s_x s_y}$. What is the covariance of this value, we have already computed the covariance, we know the covariance is 20.5 covariance between x and y, this is something which you have already computed, I compute what is the s_x, which is 1.58, s_y, which is 13.01. I divide by 1.58 × 13.01 and get 0.9964. Hence, this 0.9964, which is very very close to 1 captures the strength of the linear relationship between age and height. So, the strength of the linear relationship, this is a positive very strong linear relationship between age and height.

(Refer Slide Time: 17:23)

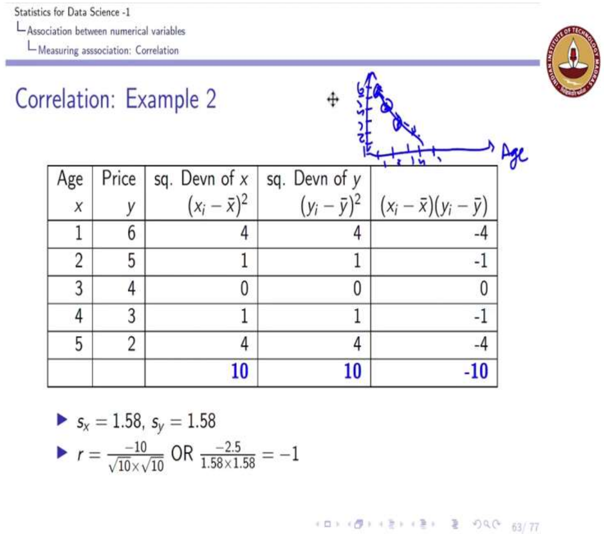


Now, let us go to the next example. In the next example, I looked at the age of a car and the price at which it is being sold. We know as cars get older, the price at which they are being sold comes down. So, again my price is recorded in lakhs of Indian rupees. So, as the age of the car, as my car gets older, my price comes down. Again, here I compute, so this was again a 3. So, you can go back this is what we have here, I had a 3, this the mean, so $\bar{y}$ was 4, $\bar{x}$ was 3, I compute the deviation, so I have a - 2, - 1, 0, 1, 2, and this is the square of the deviations which I have here.

Here again I have a 2, 1, 0, - 1, - 2, again the square is 4, 1, 0, -2^2 is 4 the product is - 4, so this is a – 4, - 2 × - 2, which is a 4, - 1 × 1, which is a - 1. So, again, the numerator term, which is the sum of the product, cross products are equal to 10, then I have the square root of the sum of x deviations, which is equal to $\sqrt{10}$, the $\sqrt{(y_i - \bar{y})^2}$, this is again $\sqrt{10}$.

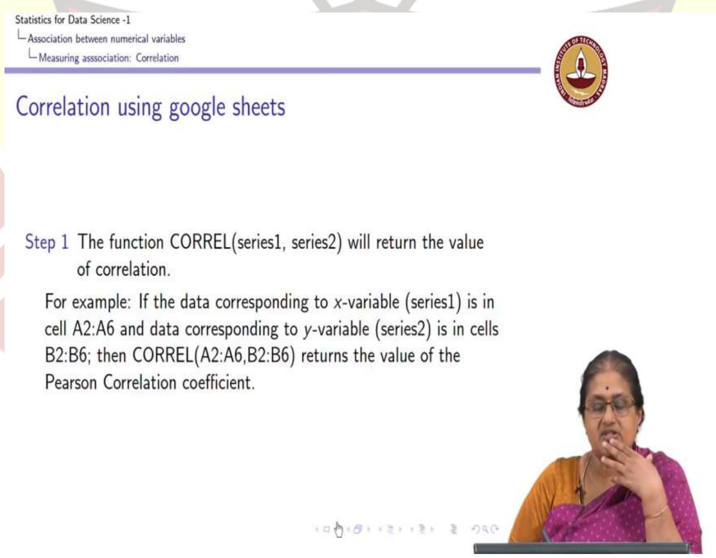
Hence, I have my first correlation metric which is $\frac{-1}{\sqrt{10}\sqrt{10}}$, which is - 1, we already have computed the covariance as - 2.5, which is – 10/4 again recall, remember it is – 10/4, when I am computing the sample covariance where, I am dividing it by n - 1 instead of the total number of observations. So, hence I am taking it to be a - 2.5. The standard deviation is 1.58. And I can see that $\frac{-2.5}{1.58 \times 1.58}$; I am going to get something which is close to - 1.

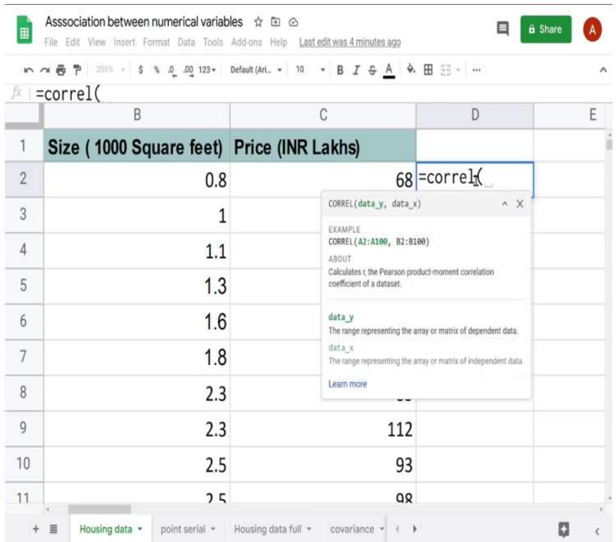
(Refer Slide Time: 20:07)

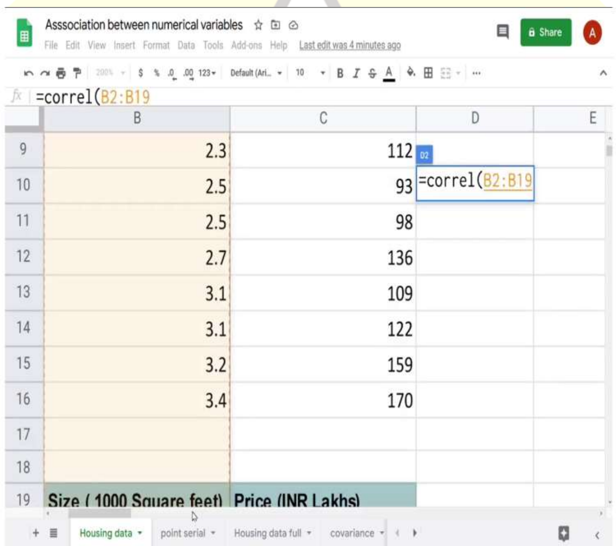


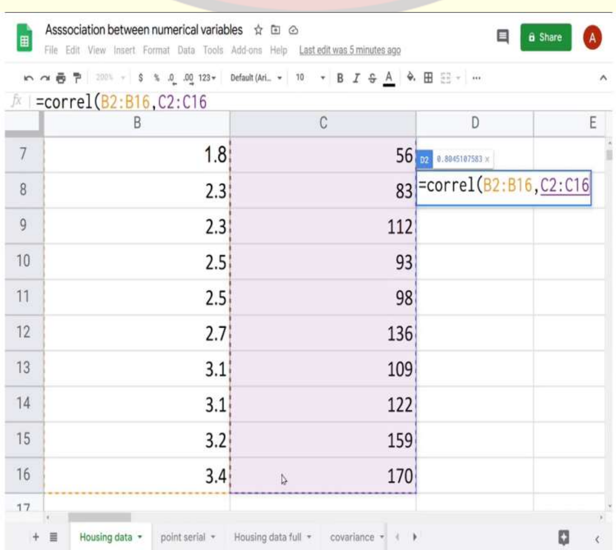
So, we have already seen from the scatter plots of both of these, this has a, so with age which is 1, 2, 3, 4, and 5; 3, 4, 5, 6. I have these as my points, 1, 2, 3, 4, 5, 6, 3 is a 4, 4 is a 3 and 5 is a 2. So, you can see that, as my car gets older, the price drops. And there is a perfect negative correlation between age and price, so I do not have to specify the units when I am talking about a correlation measure.

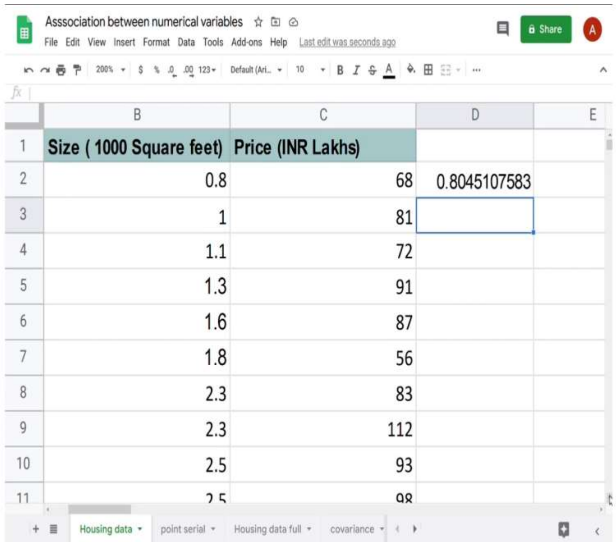
(Refer Slide Time: 20:46)




Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help Last edit was 4 minutes ago
Share
=correl(
Size (1000 Square feet)
Price (INR Lakhs)
0.8
68
=correl(
1
1.1
1.3
1.6
1.8
2.3
2.3
112
2.5
93
CORREL(data_y, data_x)
EXAMPLE
CORREL(A2:A100, B2:B100)
ABOUT
Calculates r, the Pearson product-moment correlation coefficient of a dataset.
data_y
The range representing the array or matrix of dependent data.
data_x
The range representing the array or matrix of independent data.
Learn more
Housing data
point serial
Housing data full
covariance

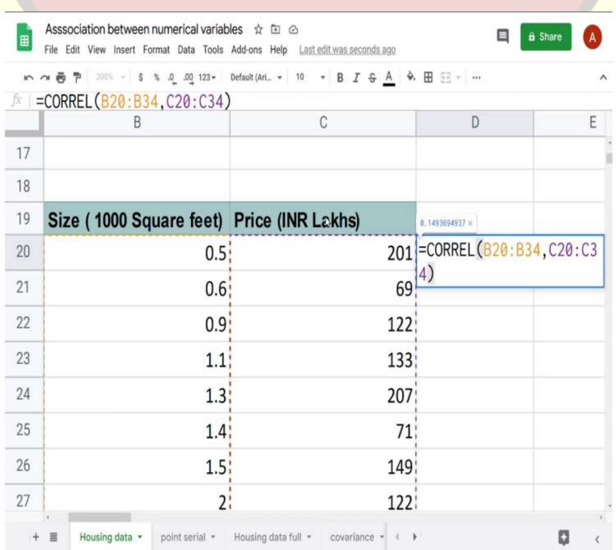

Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help Last edit was 4 minutes ago
Share
=correl(B2:B19
2.3
112
2.5
93
=correl(B2:B19
2.5
98
2.7
136
3.1
109
3.1
122
3.2
159
3.4
170
Size (1000 Square feet)
Price (INR Lakhs)
Housing data
point serial
Housing data full
covariance

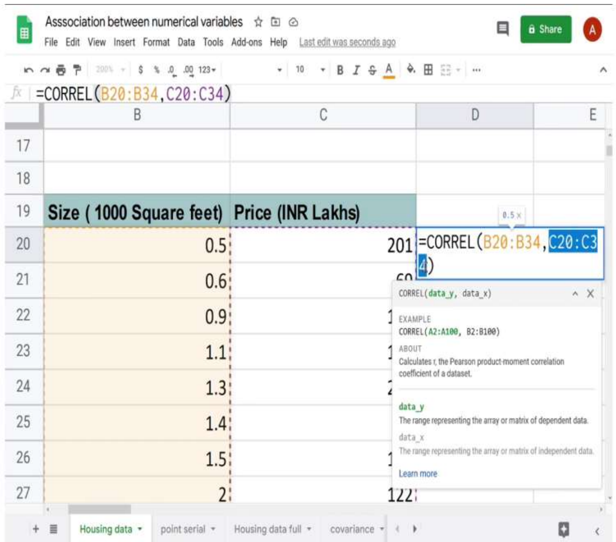

Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help Last edit was 5 minutes ago
Share
=correl(B2:B16,C2:C16
1.8
56
2.3
83
=correl(B2:B16,C2:C16
2.3
112
2.5
93
2.5
98
2.7
136
3.1
109
3.1
122
3.2
159
3.4
170
Housing data
point serial
Housing data full
covariance

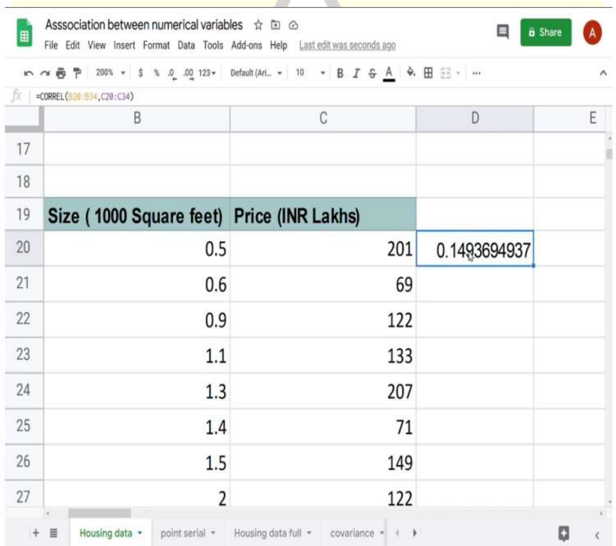


So, now let us use, how do we compute correlation using Google Sheets. So, to use the Google Sheets, what we do is first we go to any cell. Let me enlarge it a bit. I go to any cell here. So, what it ask for is data y, and data x, what I need to specify is what is my 1st data, which is my x data which is size, in this case, and my second data, which is my price, which is my C2 to your C16. So, the correl measure returns what is called the correlation coefficient, or the Pearson correlation coefficient between the two variables. So, we can see that the Pearson's correlation coefficient between the variable size and price is 0.804.

(Refer Slide Time: 21:57)

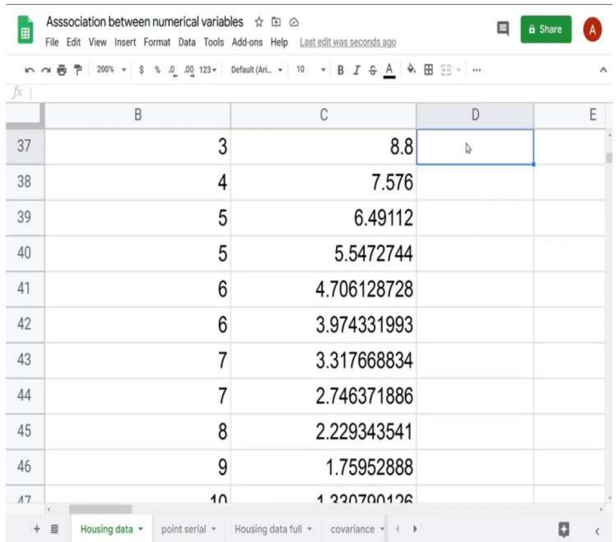


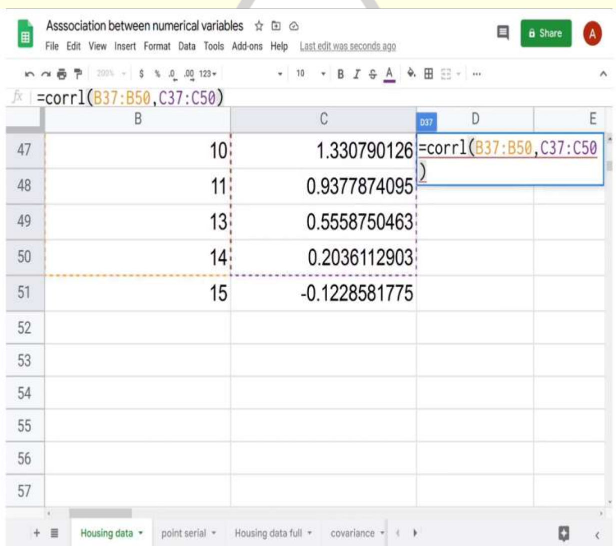


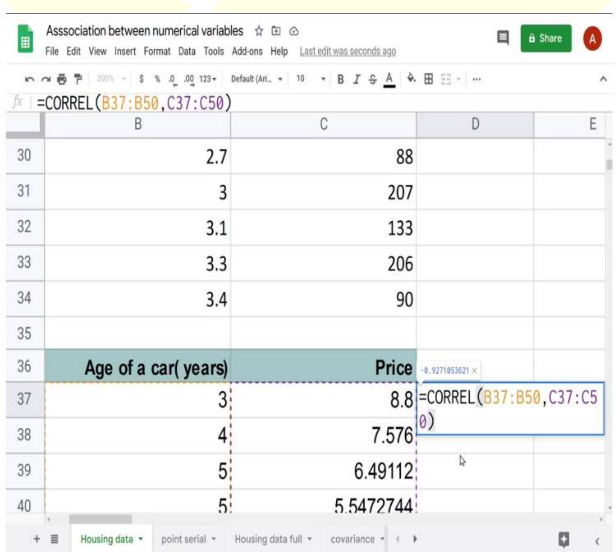


Similarly, let us look at the correlation coefficient between the next dataset. Again, what we had in that data set was a size and price for a different data set. Again, I find what is the correlation coefficient; my data here is B20 to B34, and C20 to C34, and I immediately see that the correlation coefficient is 0.149, which is very close to 0.

(Refer Slide Time: 22:29)

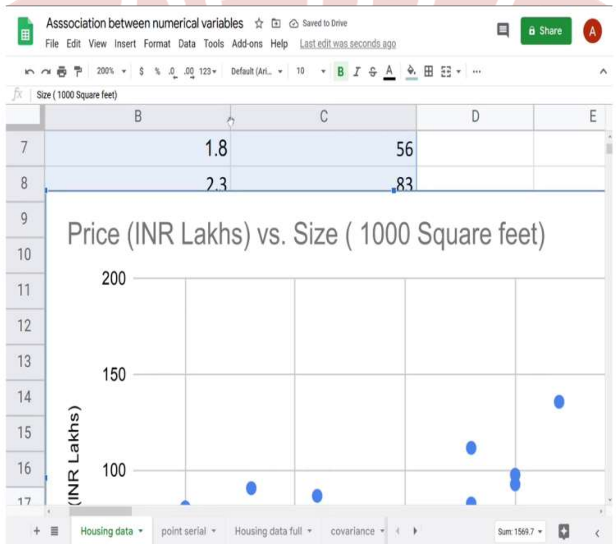


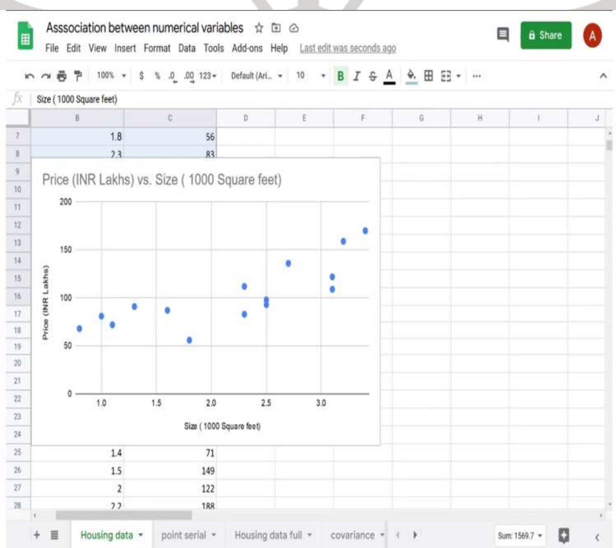


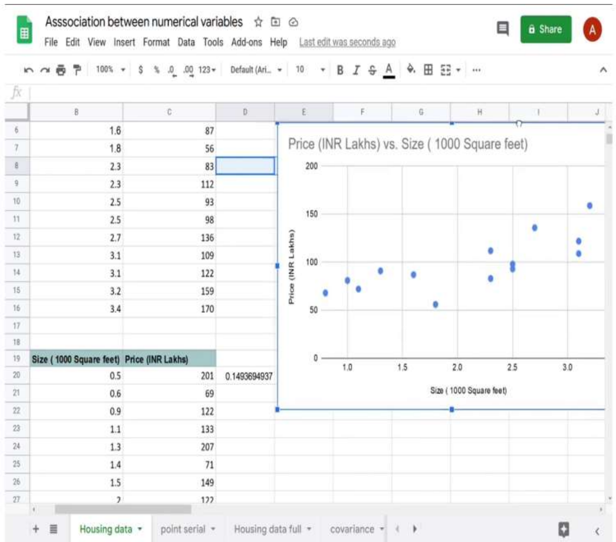


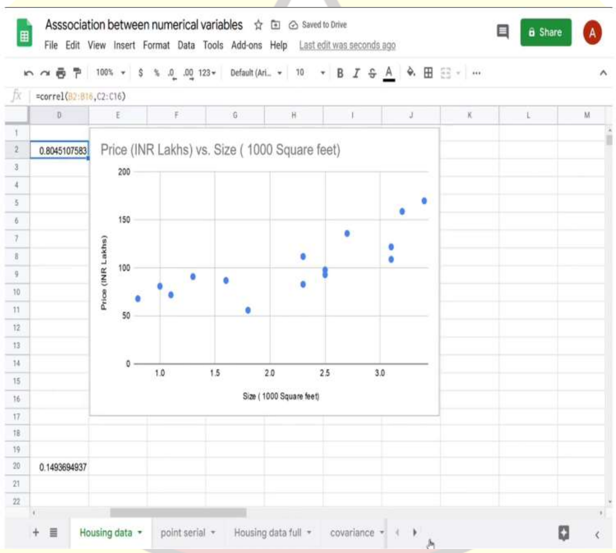
The third dataset was age of a car versus price of a car; I repeat the same here also. And I look at the correlation between the two variables here, which is going to be the age of a car and the price of a car. So this is B, so I just look at it, so let me look up to be B50 alone. And I notice that this correlation coefficient, the correl, the correl between these two is, what I am going to notice here, is given by minus. So, if you look at the data, you see it a - 0.927, which is also a large negative value close to - 1.

(Refer Slide Time: 23:27)



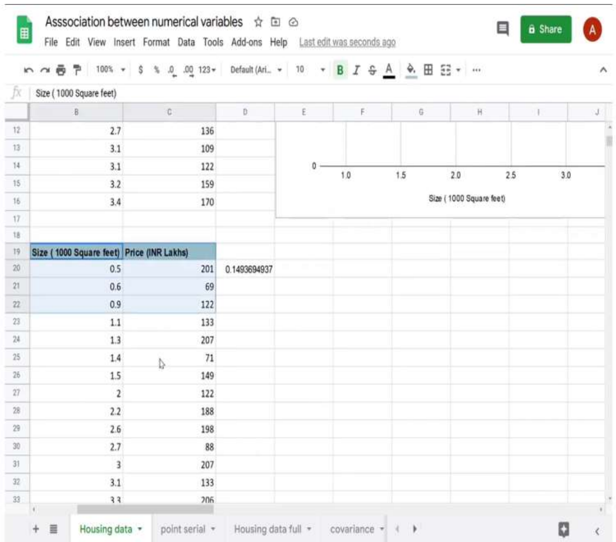


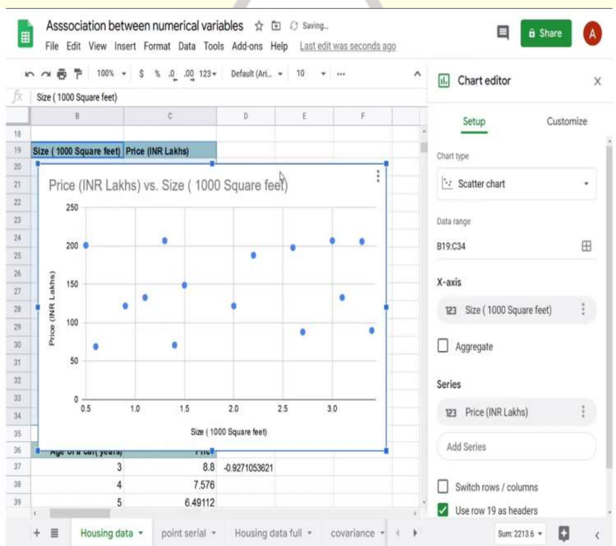


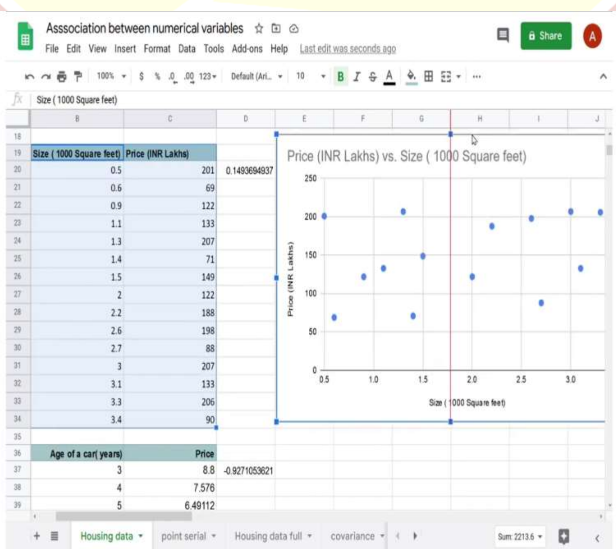


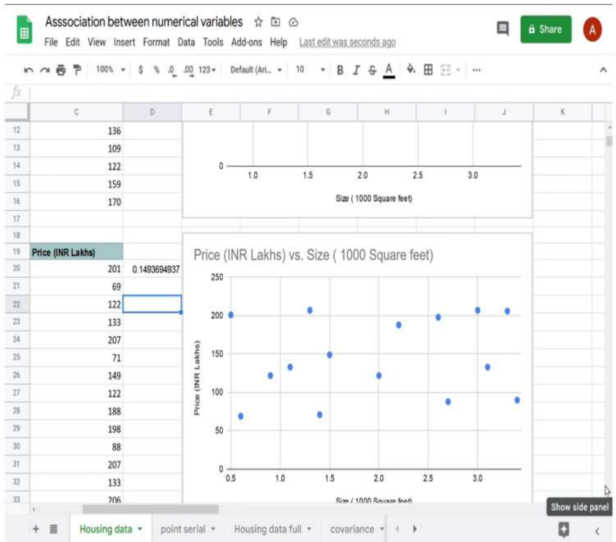
Now, one way to explain what is happening here is by looking at the scatter plots, so let me look at the scatter plot between my first dataset. So, again, I go, I plot a scatter plot between my first dataset. And, what I notice in my first dataset and the scatterplot is, there is a reasonable linear relationship between my x variable which was the size here, and my y variable, which was the price and this strength of this linear relationship is what is captured by my correlation, which is given by 0.804.

(Refer Slide Time: 24:11)



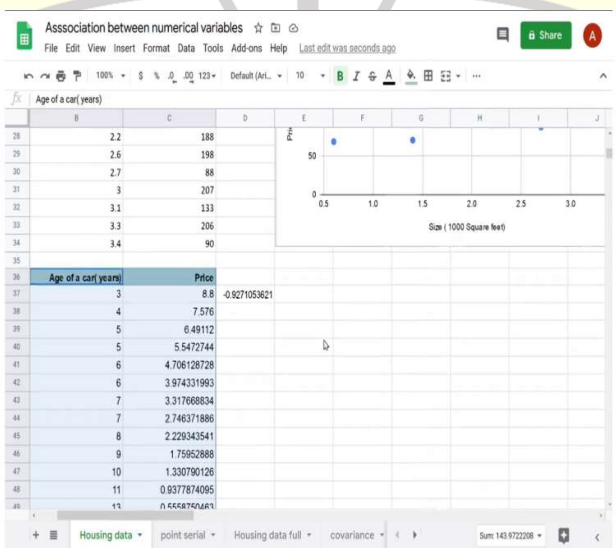


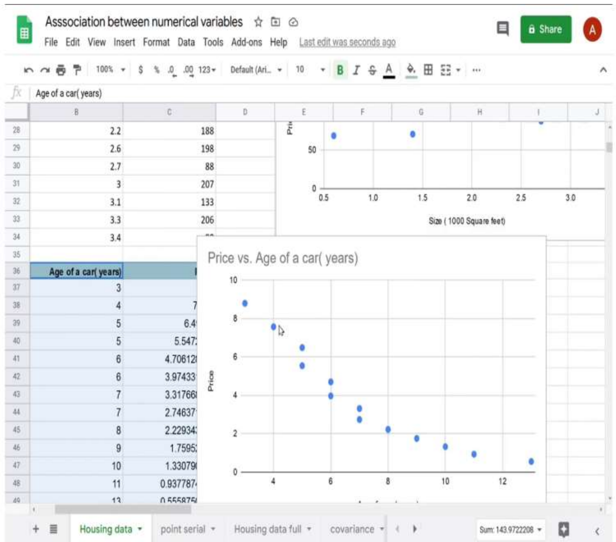


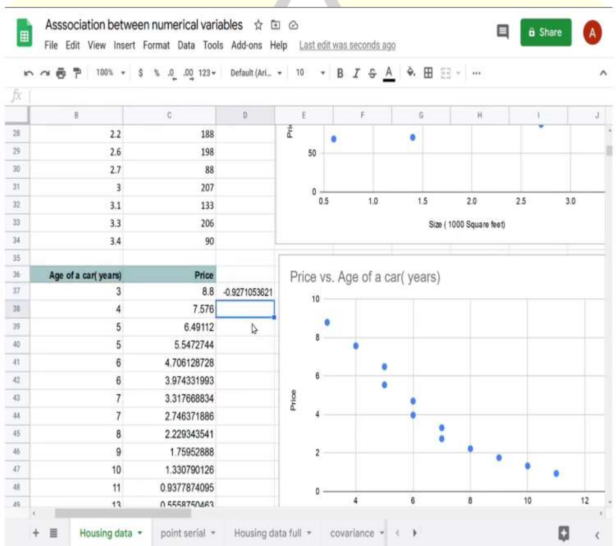


When I do the same scatterplot for the next dataset, I again go back and do the scatterplot for the next dataset. What I observe in this case is that even though the variables are the same, what we observe in this case is, the scatter which was, there was a pattern which was evident in the first dataset. I do not see that pattern here. And there is no pattern here, which is reflected in my correlation coefficient which is very low and close to 0, which is a 0.149.

(Refer Slide Time: 24:53)

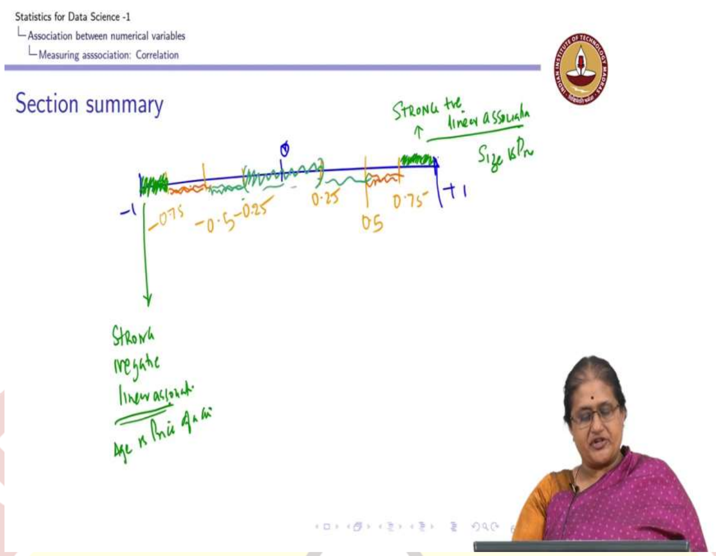






If I continue and do the same thing and plot a scatter in this scatterplot, I see a negative decrease in trend, and this quantifying or the strength of this negative relationship is quantified by this number -0.927. Hence, we can see that the correlation coefficient quantifies the strength of the linear relationship between my variables.

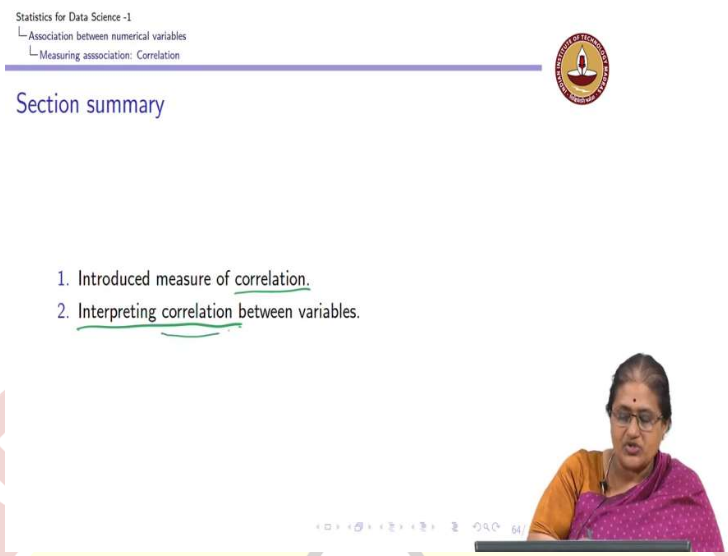
(Refer Slide Time: 25:27)



In other words, based on the correlation measure, I can tell the following that a the correlation measure lies between + 1 and - 1. So, I can have, so it can take any value. So, if it is this is a 0.5, this is a - 0.5, this could be a 0.75, - 0.75, this could be a - 0.25, this could be a plus 0.25, this could be a plus 0.75. So, I can start telling the following that if I have data, my correlation measure is between 0.75 and 1 then I could say that this indicates a strong positive linear association.

Similarly, if I have a correlation measure which is in this range, this could indicate a strong negative linear association. This association was my first example of size versus price of a home, this was the example of, this kind of relationship was age of a car versus price of a car, and depending, so this is a very strong, this portion could be just a portion where I have a reasonable positive and reasonable negative, this could have, this could indicate a weak, and between these two, this range could indicate no association. These are just indicator measures to say whether or to interpret; how strong is my linear relationship.

(Refer Slide Time: 27:42)



So, once we do this, we come to an end of this section, first we introduced the measure of correlation and then we saw how to interpret the strength of a linear relationship by using this measure of correlation by seeing where the measure lies.