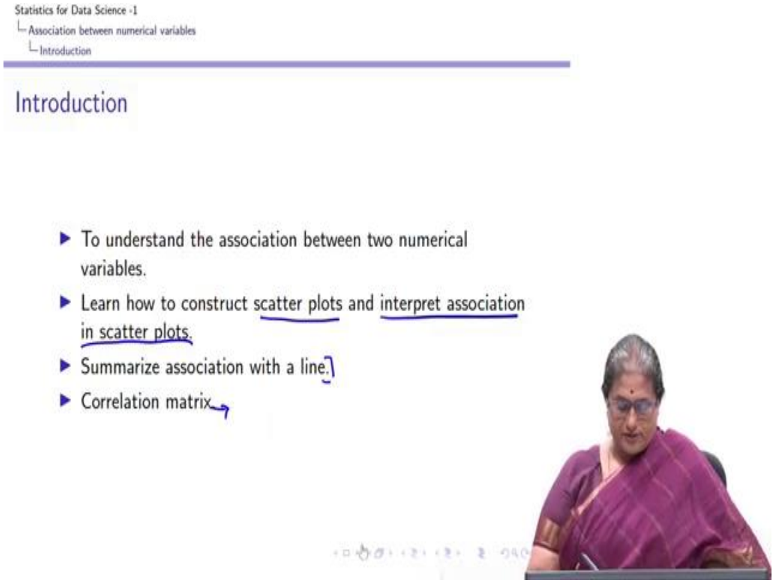
INDIAN INSTITUTE OF TECHNOLOGY MADRAS · सिद्धिर्भवति कर्मजा ·

IIT Madras

ONLINE DEGREE

Statistics for Data Science 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture 4.4
Association between Two Numerical Variables: Scatterplot

(Refer Slide Time: 00:16)



Statistics for Data Science -1
└ Association between numerical variables
 └ Introduction

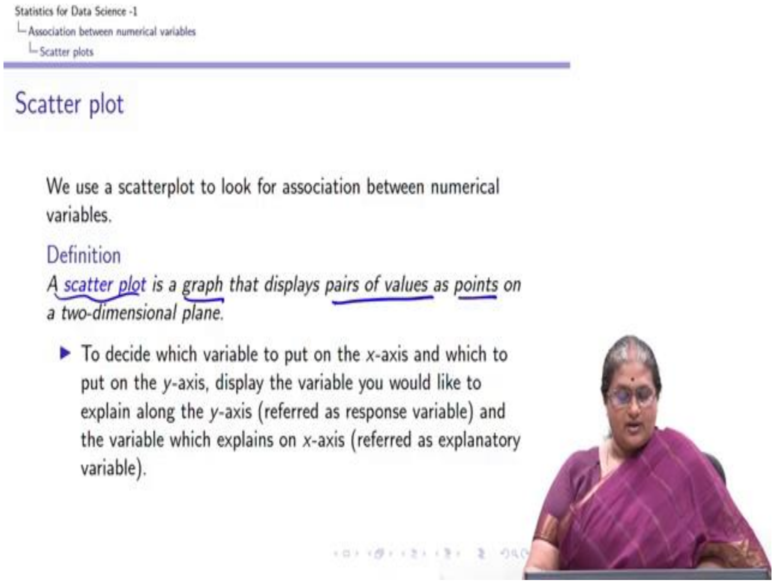
Introduction

- To understand the association between two numerical variables.
- Learn how to construct scatter plots and interpret association in scatter plots.
- Summarize association with a line.
- Correlation matrix

The next part of this lecture, we are going to understand how we describe the association between 2 numerical variables. What are we going to do in this case; we first introduced what is a scatter plot. And then we try and interpret the association between the 2 variables using the scatter plot. Recall, we interpreted the association between 2 categorical variables using the notion of a contingency table.

And here we are going to use a notion of a scatter plot. And then further, we will just briefly introduce how we summarize this association through a concept of a line. And then we will also introduce the notion of a correlation matrix so that this notion can be extended to understand association between more than 2 variables.

(Refer Slide Time: 01:14)



Statistics for Data Science -1
└Association between numerical variables
 └Scatter plots

Scatter plot

We use a scatterplot to look for association between numerical variables.

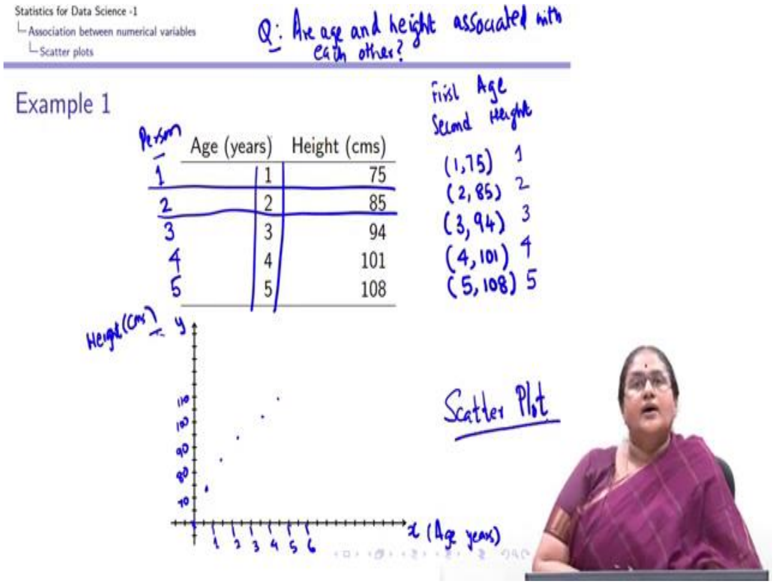
Definition

A scatter plot is a graph that displays pairs of values as points on a two-dimensional plane.

- To decide which variable to put on the x-axis and which to put on the y-axis, display the variable you would like to explain along the y-axis (referred as response variable) and the variable which explains on x-axis (referred as explanatory variable).

So what is a scatter plot? A scatter plot is defined as a graph that displays pairs of values as points on a 2-dimensional plane. So, what do we mean by this?

(Refer Slide Time: 01:33)



Statistics for Data Science -1
└Association between numerical variables
 └Scatter plots

Q: Are age and height associated with each other?

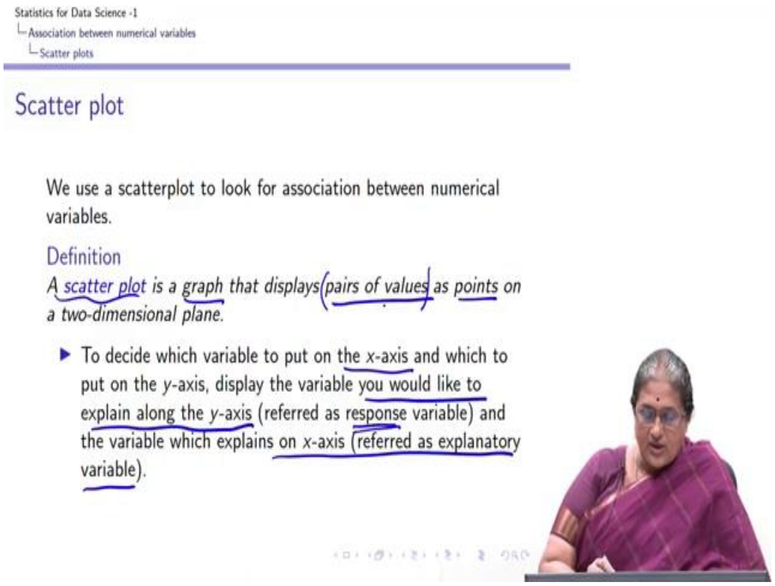
Example 1

Person	Age (years)	Height (cms)
1	1	75
2	2	85
3	3	94
4	4	101
5	5	108

First Age
Second Height

(1,75)	1
(2,85)	2
(3,94)	3
(4,101)	4
(5,108)	5

Scatter Plot



Statistics for Data Science -1
└Association between numerical variables
 └Scatter plots

Scatter plot

We use a scatterplot to look for association between numerical variables.

Definition

A scatter plot is a graph that displays (pairs of values) as points on a two-dimensional plane.

- To decide which variable to put on the *x*-axis and which to put on the *y*-axis, display the variable you would like to explain along the *y*-axis (referred as response variable) and the variable which explains on *x*-axis (referred as explanatory variable).

So, let us start with a very simple example. Suppose I am looking at 2 variables, what are the 2 variables I am looking at? The first variable I am looking at is age. And the second variable I am looking at is height. The people are the same. So I have an observation here. So, this is my first or this is the person, okay, the first person is of age 1 and 75 centimeters; person 2, person 3, person 4, person 5 and I am just looking at the age of this person.

So what this means, and the way I can interpret is, the person's 1 age is 1 and 75 centimeters, I can write this as an order pair 1, 75. For the person 2 age is 2 and height is 85 centimeters, age is measured in years, height is measured in centimeters, person 3 is 3 and 94 centimeters, person 4 is 4 and 101 and person 5 is 5 and 108 centimeters, okay. So, I have 2 variables, both of the variables are numerical.

And I am measuring these 2 variables age in units of years and height in units of centimeters. And the question we are asking here is are age and height associated with each other? This is the question we are asking. This is the question we are asking. Now to answer this question, we first try and plot it as a scatter plot on a 2-dimensional plane. Now when I am plotting it on a 2 dimensional plane, all of us know that this axis is called the x axis and this axis is called the y axis.

So the first thing which we need to decide is which of these 2 variables would go on the x axis and which variable would go on the y axis. The rule of thumb is when I want to understand the association between any 2 variables typically I might want to know whether one variable is being explained by the other variable.

In this example, I would want to know whether as a person grows older, I want to know what is the height or what is the association with the height does the height of an individual increase or does it decrease or is there any association at all? So the variable that goes on to your x axis that is also referred to as the explanatory variable.

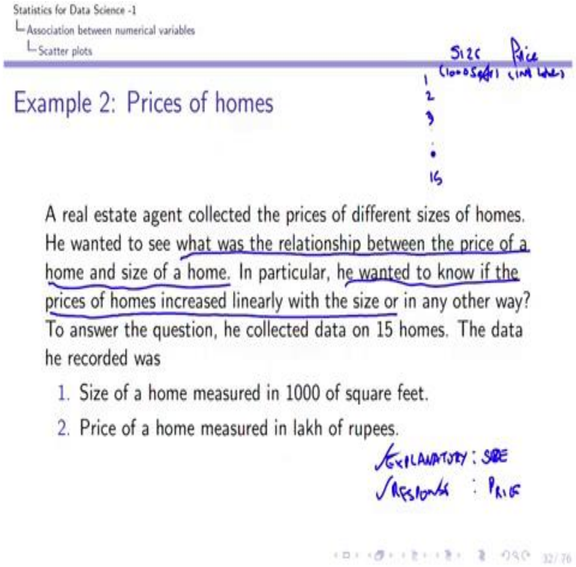
The variable that you are using to explain and the variable you would like to explain, which is height in this example, is also referred to as a response variable that is one my Y axis. So, I can write my response variable on my Y axis. In this case, it is the height which is measured in centimeters. And my explanatory variable is age which is again measured in years. So, once I have the age and height written then I will go and start writing on my x axis.

Remember, we are again talking about numerical data, there is an order to it, so, I have a start with the 0 yet I have a 1, I have a 2, I have a 3, I have a 4, I have a 5, I have a 6. I can start putting my data on my x axis and what is the data on my x axis? It is the age in years that is what is given to me. Now, when I look at the height, I can start and remember I can put a break here because I wants to start with 70 centimeters, I have 80, 90, 100, 110.

So, first age 75. So, this would be the point which is associated with my first point with 2 I have 85 which is this point, with 3 I have 94 which is this point, with 4 I have 101, which is again this point and with 5 I have 108, which is this point. So, in effect what we have done is, we have actually plotted the data that is a pair of values.

What are the pairs of values we have here? This is the first pair this is the second pair, this is the third pair, this is a fourth pair and this is the fifth pair, I have these 5 pairs of values, which I have represented as points on my 2-dimensional plane. So, this plot which I have constructed here is what we refer to as a scatter plot.

(Refer Slide Time: 07:20)



Statistics for Data Science -1
└ Association between numerical variables
  └ Scatter plots

Size (1000 Sqft) Price (in lakhs)

1
2
3
:
15

Example 2: Prices of homes

A real estate agent collected the prices of different sizes of homes. He wanted to see what was the relationship between the price of a home and size of a home. In particular, he wanted to know if the prices of homes increased linearly with the size or in any other way? To answer the question, he collected data on 15 homes. The data he recorded was

1. Size of a home measured in 1000 of square feet.
2. Price of a home measured in lakh of rupees.

✓ EXPLANATORY : SIZE
✓ RESPONSE : PRICE

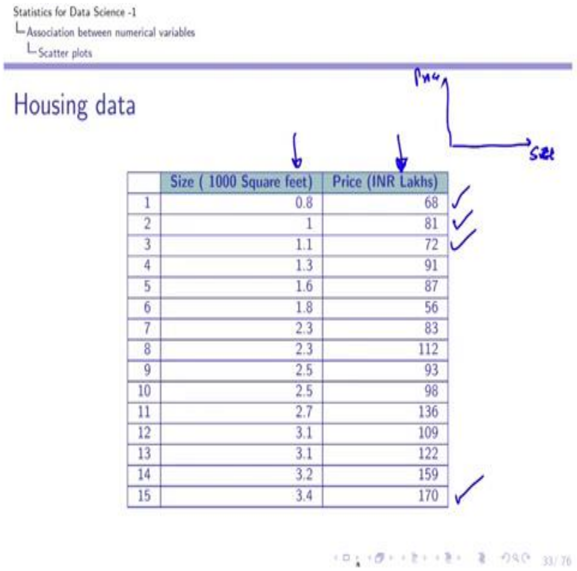
32 / 76

So, the scatter plot is an extremely powerful graph, which generally is just the scatter or it is a display of pairs of values of my variables. Now, let us look at another example. So, a real estate agent has collected the prices of different size of homes. So, what the real estate agent has done? He has gone he has collected 1, 2, 3 up to 15 homes. So, the real estate agent has collected on every home the size of the house, this is measured in thousands of square feet and the price of the house in lakhs of rupees. This is what I have as my data.

So what is it I am seeking? The seeking the question I am seeking an answer to is whether there is a relationship between the price of a home and size of home. Now in this example, you very clearly see that I want to see whether the prices vary according to the sizes. So, the natural explanatory variable in this case in this example is going to be the size of a house, whereas my response variable in this case is going to be the price of a house.

So, whenever you want to understand the association between 2 variables and you are interested in plotting or coming up with a graphical display in terms of a scatter plot. The first step is to recognize what is your explanatory variable and what is your response variable. So, next he wanted to know whether the prices of homes increase linearly with this size. We will come to answer this question in some time, but even before that, we want to see what is the data is recorded.

(Refer Slide Time: 09:29)



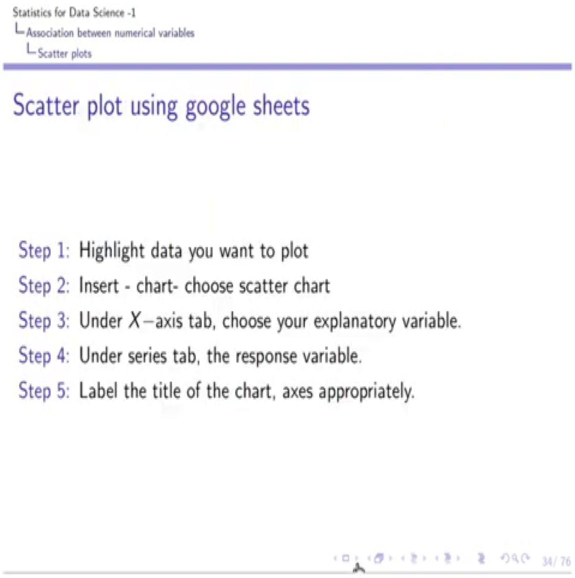
Statistics for Data Science -1
└Association between numerical variables
└Scatter plots

Housing data

	Size (1000 Square feet)	Price (INR Lakhs)	
1	0.8	68	✓
2	1	81	✓
3	1.1	72	✓
4	1.3	91	
5	1.6	87	
6	1.8	56	
7	2.3	83	
8	2.3	112	
9	2.5	93	
10	2.5	98	
11	2.7	136	
12	3.1	109	
13	3.1	122	
14	3.2	159	
15	3.4	170	✓

So, this is my data. So, I have a different data given here. So with this data, what we do is if I am just going to so I have 15 data points and I have these are the lakhs INR, lakhs or 68 lakhs is 800 square feet house is costing me 68 lakhs, 1000 square feet is 81 lakhs, 1100 square feet is 72 lakhs and 3400 square feet house is 170 lakhs. So what I do now I just so now I have, I need to plot a scatter plot. So on my x axis I am going to take the size. On the y axis, I take the price of the house plot a scatter plot, how do I plot a scatter plot.

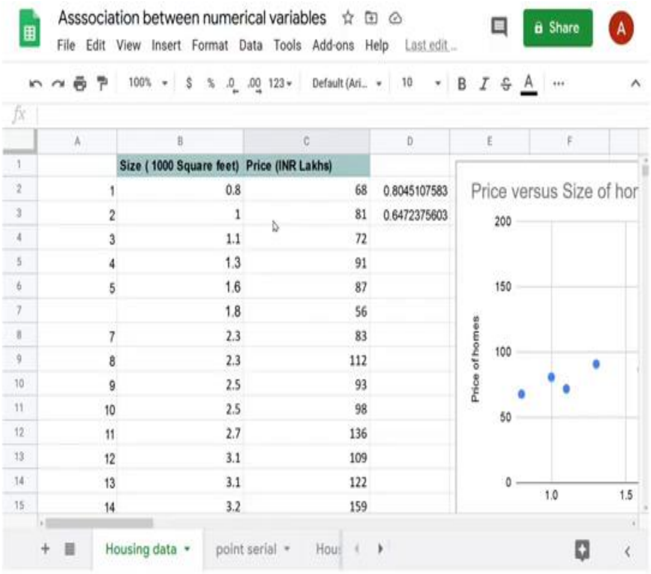
(Refer Slide Time: 10:16)



Statistics for Data Science -1
└Association between numerical variables
└Scatter plots

Scatter plot using google sheets

Step 1: Highlight data you want to plot
Step 2: Insert - chart- choose scatter chart
Step 3: Under X−axis tab, choose your explanatory variable.
Step 4: Under series tab, the response variable.
Step 5: Label the title of the chart, axes appropriately.

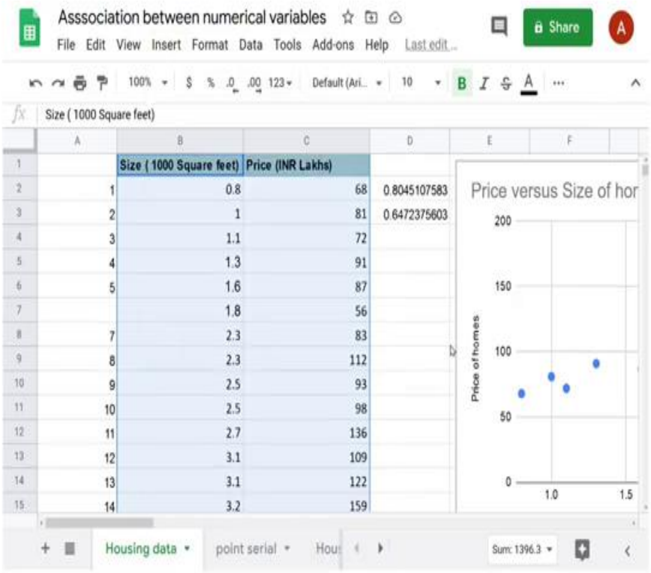


Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help Last edit...
Share
A
100%
Default (Ari...
10
Housing data
point serial
Hou

	A	B	C	D
1		Size (1000 Square feet)	Price (INR Lakhs)	
2	1	0.8	68	0.8045107583
3	2	1	81	0.6472375603
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	
12	11	2.7	136	
13	12	3.1	109	
14	13	3.1	122	
15	14	3.2	159	

So let us go back to our Google Sheets. This is the data given I have the same data given in my Google Sheet. So you can see that this is the same data I have 0.8, 0.1. 1.1 so I have 68, 81, 72 this is in square feet, and this is in lakhs of rupees. So the first thing I want to ask is can a plot a scatter plot, the Google Sheet gives me an easy way to plot a scatter plot. Let us go about and plot a scatter plot using a Google Sheet.

(Refer Slide Time: 10:51)

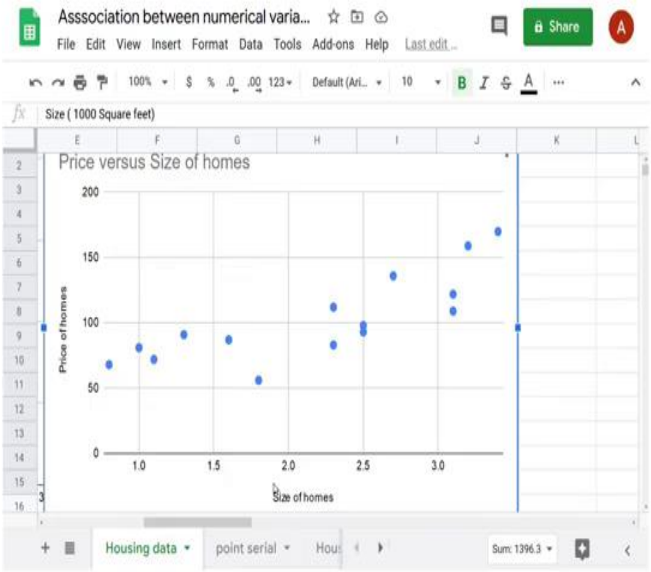


Asssociation between numerical variables

File Edit View Insert Format Data Tools Add-ons Help Last edit ...

Size (1000 Square feet)

	A	B	C	D
1		Size (1000 Square feet)	Price (INR Lakhs)	
2	1	0.8	68	0.8045107583
3	2	1	81	0.6472375603
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	
12	11	2.7	136	
13	12	3.1	109	
14	13	3.1	122	
15	14	3.2	159	

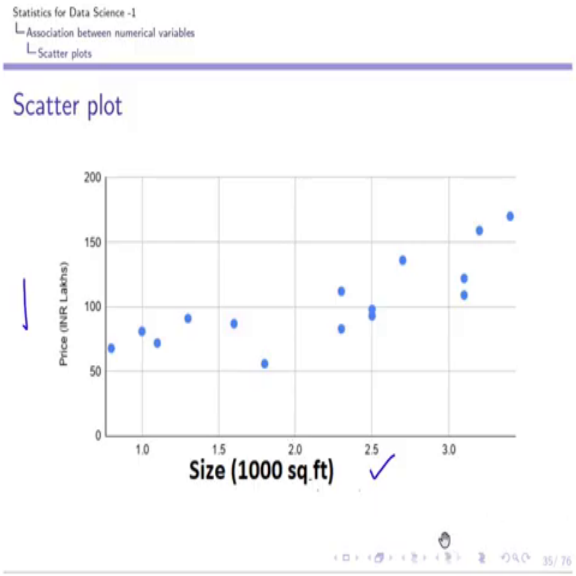


Asssociation between numerical varia...

File Edit View Insert Format Data Tools Add-ons Help Last edit ...

Size (1000 Square feet)

सिद्धिर्भवति कर्मजा



Statistics for Data Science -1
└Association between numerical variables
└Scatter plots

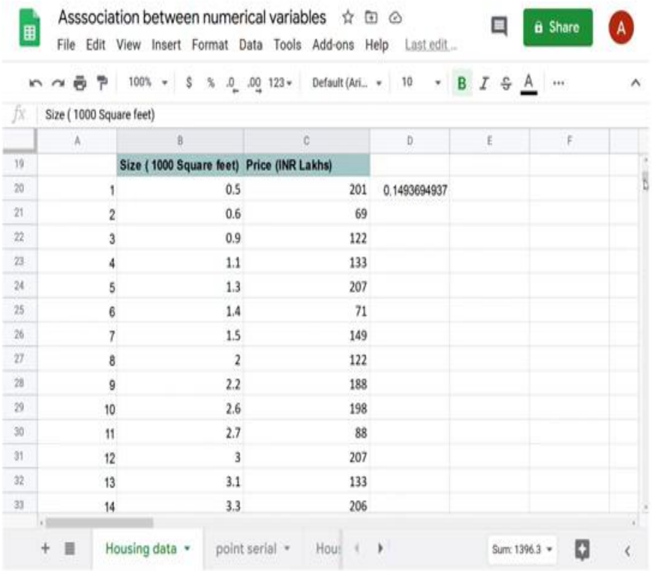
Scatter plot

So the first step is highlight the data you want to plot as a scatter plot. And I have highlighted that data insert chart. So if I go to insert and go for a chart, you can see that I have a scatter chart, which has appeared. And this is what it is under the x axis tab, the x axis (I have) it chooses the square feet, and it has automatically chosen in this case, because that was my first column available.

And in under the series tab, it gives me what is the response variable, which is again price in INR lakhs, and I can again, label, title and access appropriately. So you can see that I have a nice scatter plot between the size and the price which I got from my Google Sheet. So this is the scatter plot we have got, again, I have size on my x axis, and price on my Y axis. So this tells us how to construct a scatter plot.

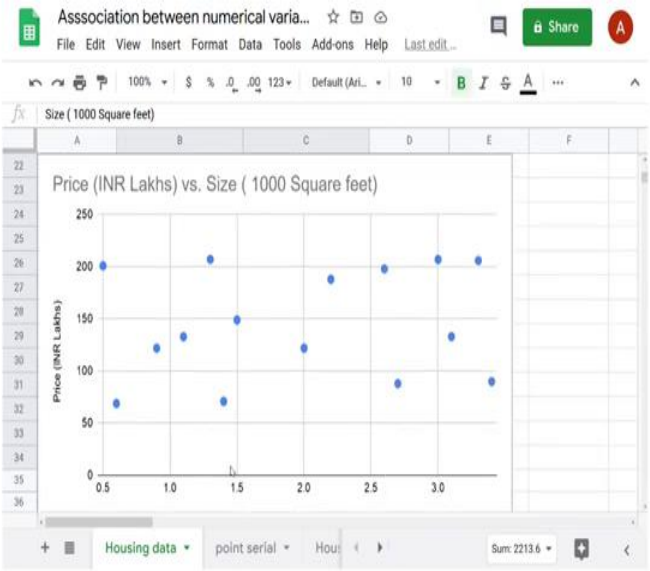
सिद्धिर्भवति कर्मजा

(Refer Slide Time: 12:05)


Asssociation between numerical variables
File Edit View Insert Format Data Tools Add-ons Help Last edit...
Share
A
100%
Default (Ari...
10
Housing data
point serial
Hou:
Sum: 1396.3

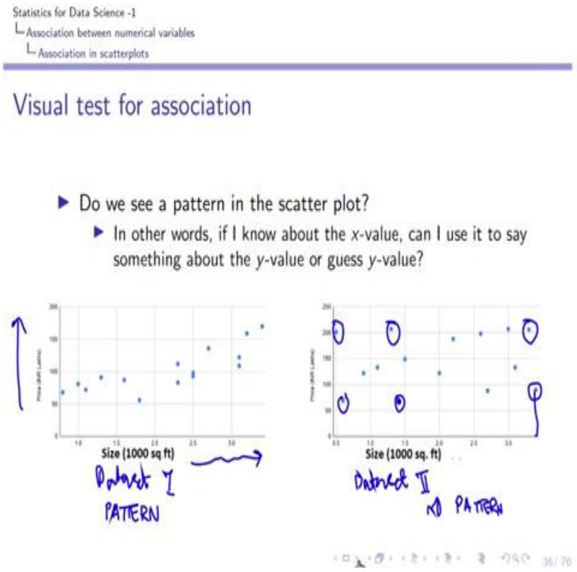
fx Size (1000 Square feet)

	Size (1000 Square feet)	Price (INR Lakhs)	
1	0.5	201	0.1493694937
2	0.6	69	
3	0.9	122	
4	1.1	133	
5	1.3	207	
6	1.4	71	
7	1.5	149	
8	2	122	
9	2.2	188	
10	2.6	198	
11	2.7	88	
12	3	207	
13	3.1	133	
14	3.3	206	


Asssociation between numerical varia...
File Edit View Insert Format Data Tools Add-ons Help Last edit...
Share
A
fx Size (1000 Square feet)
Price (INR Lakhs) vs. Size (1000 Square feet)
Price (INR Lakhs)
250
200
150
100
50
0
0.5
1.0
1.5
2.0
2.5
3.0
Housing data
point serial
Hou:
Sum: 2213.6

Now once given a scatter plot, let us look at another example. For example, if I go and look at this other example I have here, so the same price so I remove this, let me remove this graph, because I have already I delete this chart. Let me go to another example where I am again, talking about same the price and the size versus the price. This is the data I have, but if I am plotting a scatter plot between this data, you can notice a scatter plot of this kind, okay.

(Refer Slide Time: 12:47)



Statistics for Data Science -1
└Association between numerical variables
└Association in scatterplots

Visual test for association

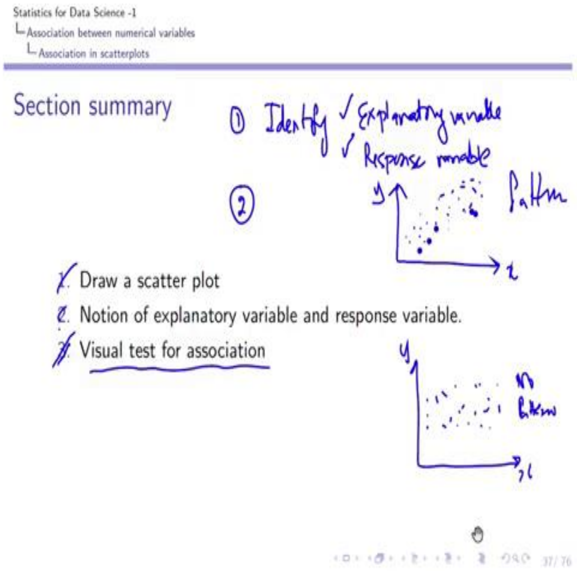
- Do we see a pattern in the scatter plot?
  - In other words, if I know about the *x*-value, can I use it to say something about the *y*-value or guess *y*-value?

So, let us copy the scatter plot. So, what we do is this is the scatter plot again 15 homes I have a scatter plot here. So this graph both the graphs actually plot size versus price. This is for my first data set which I call data set 1 and this is for the second data set which I refer to as data set 2, okay. So now, if you look at this in data set 1, I can observe that as the sizes of the homes are increasing, the prices are also exhibiting some sort of a trend or it increases.

Whereas in this I have smaller homes for example, if I look at this size of a home, I have bought a lower priced house and higher priced house and I have larger homes also this is a larger home because the size is a larger size, this size also I have a lower price and a higher price. A midsize home also has a lower price and a larger price. In other words, this data set 1 has a pattern which I can in some sense explain whereas in data set 2 I do not see any clear pattern.

So, the first thing which we want to understand is, remember, we wanted to know whether a scatter plot can actually help us understand association between numerical variables. So, we can see that from these 2 examples in the case of 1 example I can actually see a pattern whereas in the other case, I do not see any clear pattern. So the visual test for association tells us whether you can see a pattern or not see a pattern.

(Refer Slide Time: 14:54)



Statistics for Data Science -1
└Association between numerical variables
└Association in scatterplots

Section summary

① Identify ✓ Explanatory variable
✓ Response variable

②

1. Draw a scatter plot
2. Notion of explanatory variable and response variable.
3. Visual test for association

So, what we have learned so far as a first given data set identify what is your explanatory variable and what is your response variable. Given this, let your explanatory variable be on the x axis, response variable on the y axis, draw a scatter plot.

Once you have a scatter plot then I could have a scatter plot which is of this kind where I see a pattern or I could have a scatter plot which is of this kind where I really do not see any clear pattern between my x and y. This is referred to as a visual test for association and how do we actually draw conclusions from this visual test is what we are going to see next.

सिद्धिर्भवति कर्मजा