INDIAN INSTITUTE OF TECHNOLOGY MADRAS

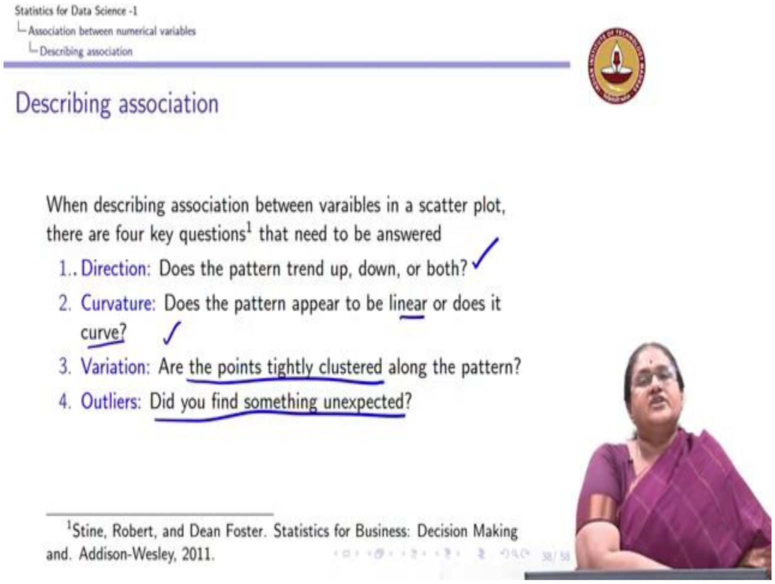
सिद्धिर्भवति कर्मजा

IIT Madras

ONLINE DEGREE

Statistics for Data Science 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture 4.5
Association between Two Numerical Variables: Describing Association

(Refer Slide Time: 00:15)



Statistics for Data Science -1
└Association between numerical variables
└Describing association

Describing association

When describing association between varaibles in a scatter plot, there are four key questions[1] that need to be answered

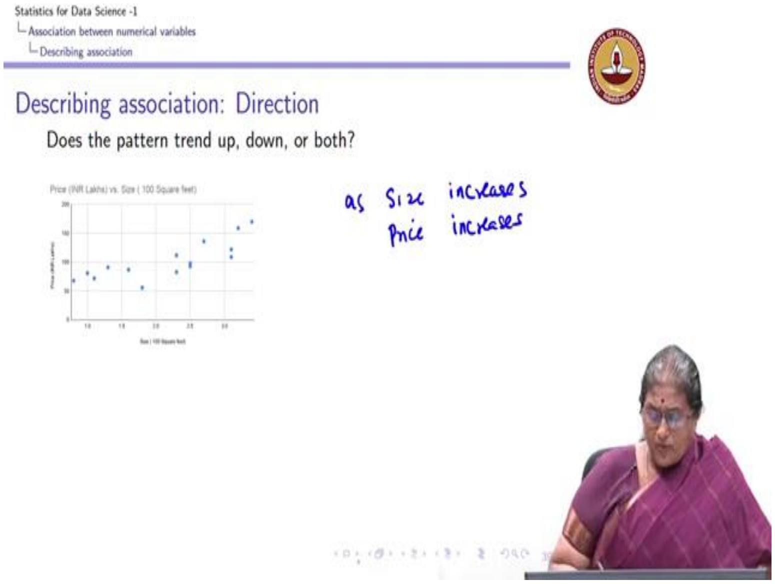
1. Direction: Does the pattern trend up, down, or both? ✓
2. Curvature: Does the pattern appear to be linear or does it curve? ✓
3. Variation: Are the points tightly clustered along the pattern?
4. Outliers: Did you find something unexpected?

[1] Stine, Robert, and Dean Foster. Statistics for Business: Decision Making and. Addison-Wesley, 2011.

38/58

So how do we describe the association between the 2 variables using a scatter plot? So when we are describing association between 2 variables, there are 4 key questions that I need to answer. The first question is, is there a direction? What do I mean by a direction? Does the pattern trend up or down or does it exhibit some sort of a trend? Is it linear or does it curve? The third is, are the points tightly clustered around the pattern or are they spread? Do I find anything unexpected? We will look at each one of these questions in detail now.

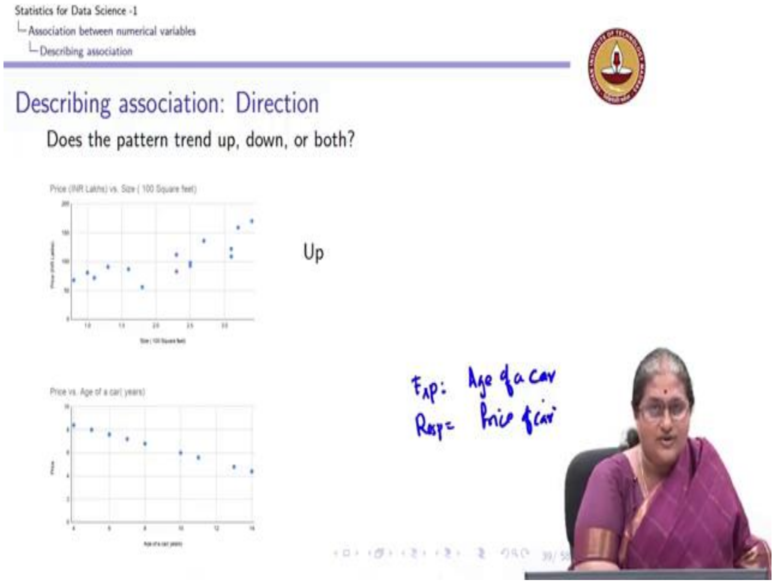
(Refer Slide Time: 01:05)



Statistics for Data Science -1
└ Association between numerical variables
  └ Describing association

Describing association: Direction

Does the pattern trend up, down, or both?



Statistics for Data Science -1
└ Association between numerical variables
  └ Describing association

Describing association: Direction

Does the pattern trend up, down, or both?

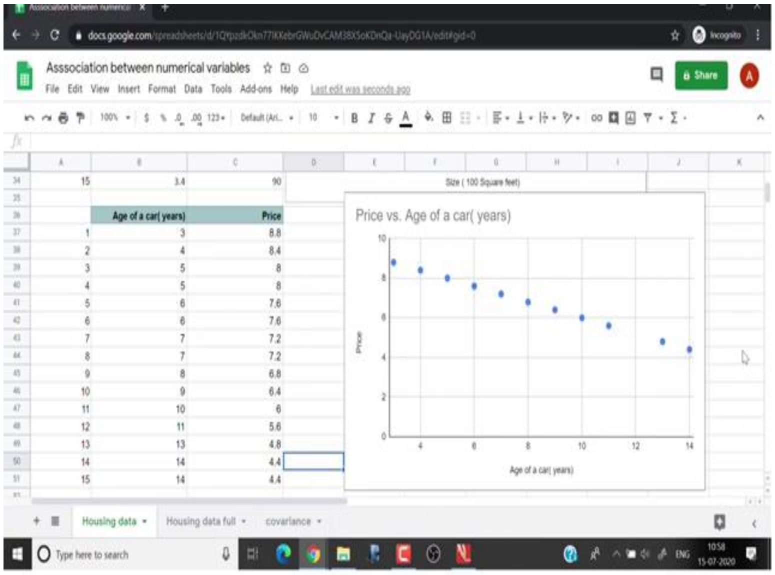
Up

Price (INR Lakhs) vs. Size (100 Square feet)
Price vs. Age of a car(years)
Exp: Age of a car
Resp= Price of car

So, the first question we want to know is, does the pattern trend up or down? Let us look at an example again, if you look, go back to this example where I had plotted the size of a whole house on my x axis, and I wanted to know the price versus size, we see that there is a pattern where I can easily say that as the sizes of my homes increase.

As size increases, price increases. I can see there is an upward trend, so the trend is up, okay. So, let us look at another example. We know that as car ages, so here instead of looking at an age of a person, my explanatory variable is age of a car. And my response variable is the price of the car. So what I mean by this is as a car becomes older or the older the car is the price I am going to get for that car again reduces.

(Refer Slide Time: 02:32)

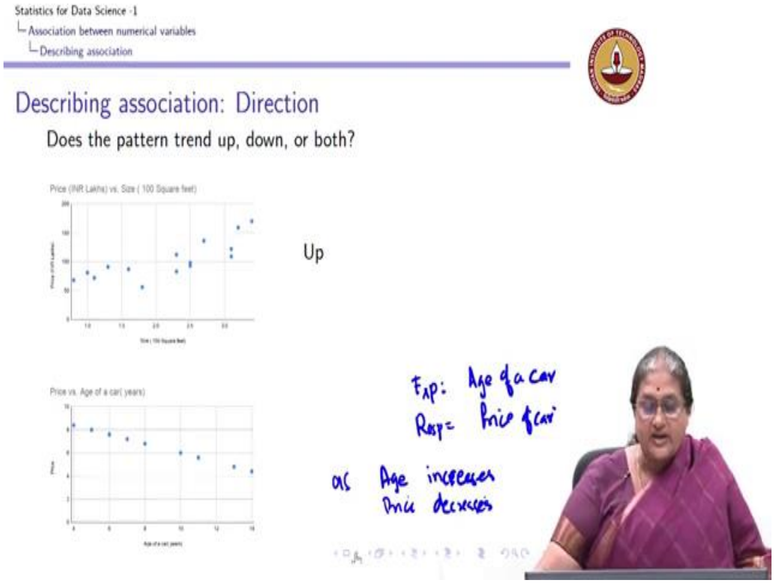


Asssociation between numerical variables

	A	B	C
34	15	3.4	90
35			
36		Age of a car(years)	Price
37	1	3	8.8
38	2	4	8.4
39	3	5	8
40	4	5	8
41	5	6	7.6
42	6	6	7.6
43	7	7	7.2
44	8	7	7.2
45	9	8	6.8
46	10	9	6.4
47	11	10	6
48	12	11	5.6
49	13	13	4.8
50	14	14	4.4
51	15	14	4.4

So let us look at a data here. So you can look at the data for a 3 year old car if I am getting the 8.8 lakhs again prices in lakhs of rupees for a 4 year old car I might get 8.4, for a 5 I might get 8, for 6 I would get 7.6 lakhs. So you can see that as the car becomes older my prices showing is coming down.

(Refer Slide Time: 02:58)

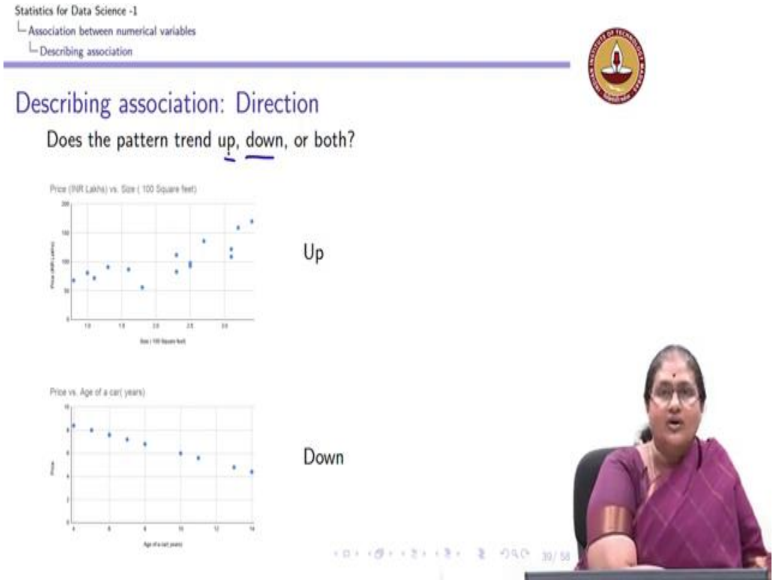


Statistics for Data Science -1
└ Association between numerical variables
└ Describing association

Describing association: Direction

Does the pattern trend up, down, or both?

Price (INR Lakhs) vs. Size (100 Square feet)
Up
Price vs. Age of a car(years)
Exp: Age of a car
Resp= Price of car
as Age increases
Price decreases



Statistics for Data Science -1
└ Association between numerical variables
└ Describing association

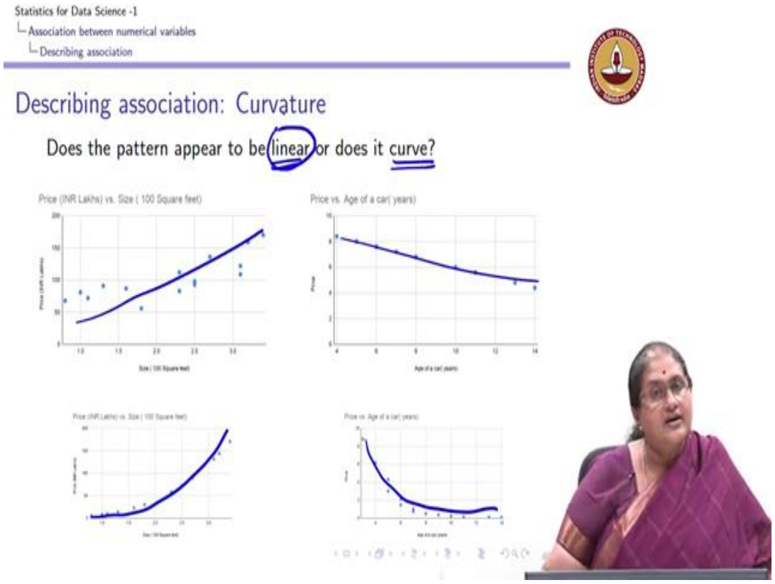
Describing association: Direction

Does the pattern trend up, down, or both?

Price (INR Lakhs) vs. Size (100 Square feet)
Up
Price vs. Age of a car(years)
Down
39 / 58

And that is what is shown by the scatter plot here. So what is the question or what is the pattern we see here? As the age increases, I see the price decreases. In this case I saw as size increases price increases. So here I can describe my pattern to have a decreasing or down trend. So the first thing which we need to see when we look at our scatter plot says whether it is showing an upward direction or a downward direction.

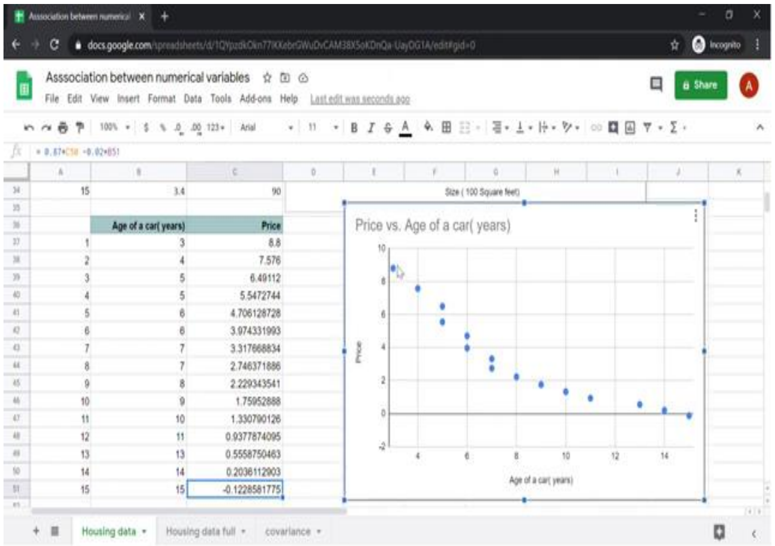
(Refer Slide Time: 03:45)



Statistics for Data Science -1
└ Association between numerical variables
  └ Describing association

Describing association: Curvature

Does the pattern appear to be linear or does it curve?



Asssociation between numerical variables

= 0.87*C50 -0.02*B51

	A	B	C
34	15	3.4	90
35			
36		Age of a car(years)	Price
37	1	3	8.8
38	2	4	7.576
39	3	5	6.49112
40	4	5	5.5472744
41	5	6	4.706128728
42	6	6	3.974331993
43	7	7	3.317668834
44	8	7	2.746371886
45	9	8	2.229343541
46	10	9	1.75952888
47	11	10	1.330790126
48	12	11	0.9377874095
49	13	13	0.5558750463
50	14	14	0.2036112903
51	15	15	-0.1228581775

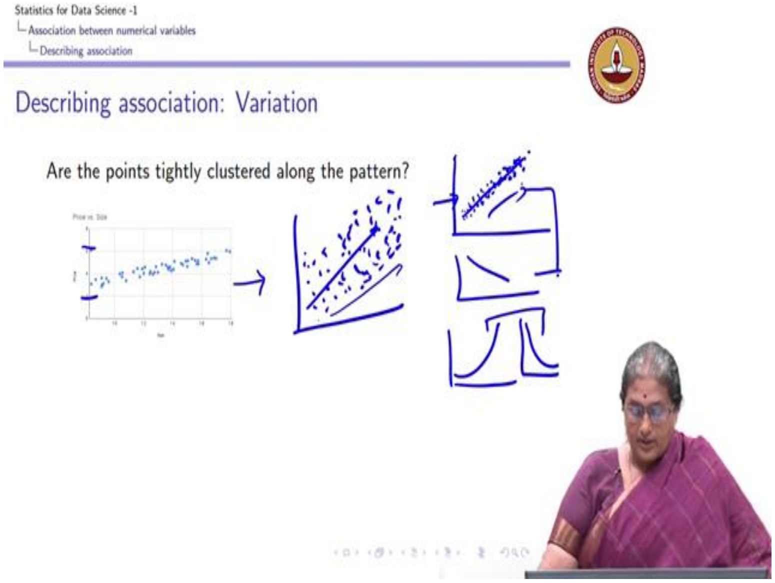
The next thing is, does the pattern appear to be linear or does it curve? Is it a linear pattern or is it a curved pattern. So let us look at the second example here. So now what you can see here is when I look at the price of the car versus age of a car here, now that you see notice a different pattern in this case.

So here you can see that so what you will notice here is even though the price of a car is decreasing, or it is showing a downward pattern as the car gets older, but you can see here there is a steep curvature, and it is not appearing linear as it was appearing earlier.

So in this case, it appeared to be a linear trend whereas here it is appearing to have curvature. Similarly, when I look at price of a house versus size of a house this graph appears to be a linear upward trend that is as my size increases, the price appears to increase linearly in terms of the size. Whereas here you can see that there is a curvature or I can say that as my size increases, the price increases according to a curved pattern.

So the next question we ask is, look at the scatter plot and check whether it is linear, or whether it is a curve. For this lecture, we are focusing on linear relationships.

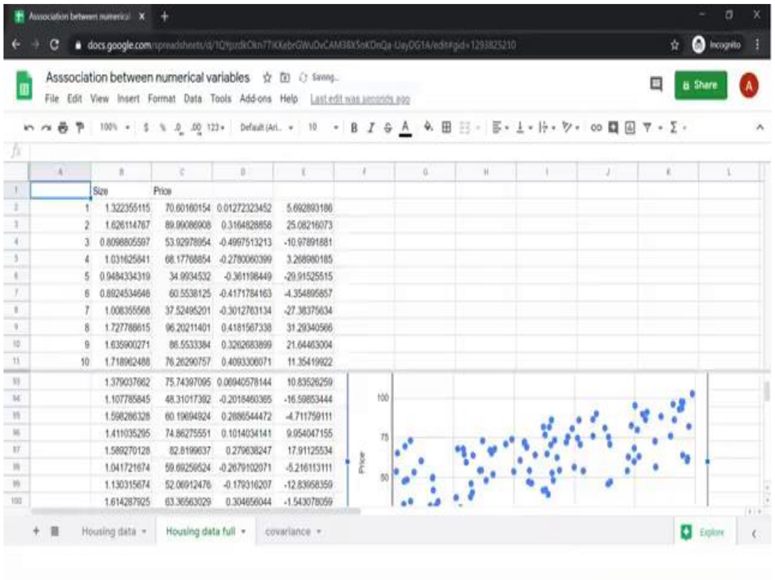
(Refer Slide Time: 05:32)



Statistics for Data Science -1
└Association between numerical variables
  └Describing association

Describing association: Variation

Are the points tightly clustered along the pattern?



Asssociation between numerical variables

	A	B	C	D	E
1		Size	Price		
2	1	1.322355115	70.60160154	0.01272323452	5.692893186
3	2	1.626114767	89.99086908	0.3164828858	25.08216073
4	3	0.8098805597	53.92978954	-0.4997513213	-10.97891881
5	4	1.031625841	68.17768854	-0.2780060399	3.268980185
6	5	0.9484334319	34.9934532	-0.361108449	-29.91525515
7	6	0.8924534646	60.5538125	-0.4171784163	-4.354895857
8	7	1.008355568	37.52495201	-0.3012763134	-27.38375634
9	8	1.727788615	96.20211401	0.4181567338	31.29340566
10	9	1.635900271	86.5533384	0.3262683899	21.64463004
11	10	1.718902488	76.26290757	0.4093306071	11.35419922
93		1.379037682	75.74397095	0.06940578144	10.83526259
94		1.107785845	48.31017392	-0.2018460365	-16.59853444
95		1.598286328	60.19694924	0.2886544472	-4.711759111
96		1.411035295	74.86275551	0.1014034141	9.954047155
97		1.589270128	82.8199637	0.279638247	17.91125534
98		1.041721674	59.69258524	-0.2679102071	-5.216113111
99		1.130315674	52.06912476	-0.179316207	-12.83958359
100		1.614287925	63.36563029	0.304656044	-1.543078059

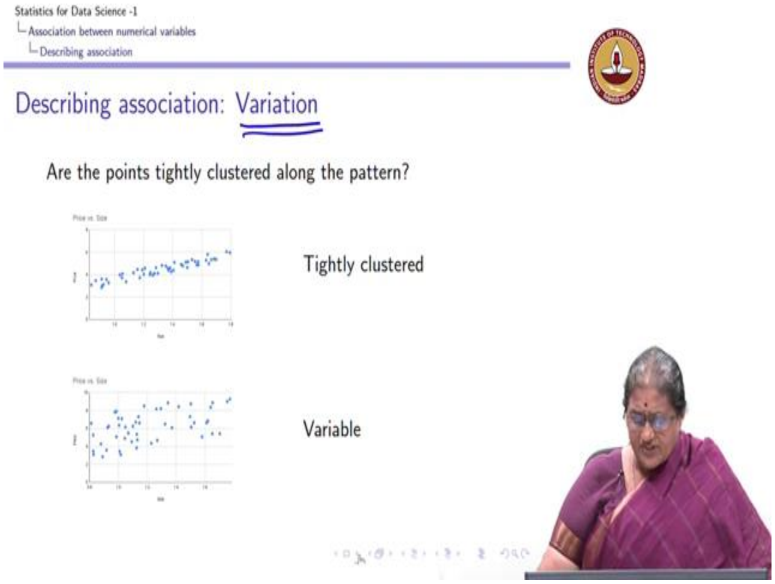
The next question which we want to answer is whether points are tightly clustered around the pattern? Now, if you look at this example, again, I have actually about 100 homes here. I have done the same thing, we have about 100 homes, okay. There are 100 homes I have taken the data of 100 homes here. So, you can see that the data we are talking about is 100 homes, and they are actually tightly clustered around each other.

So, if you look at this scatter plot which we are describing here, these are about 100 homes, and you can see that the data is tightly clustered in the sense that each of these 100 homes are between this range of my prices which could be between 2, 20 or 2 lakhs to 60 lakhs depending on what is the unit of the prices.

So, this is what we refer to as tightly clustered along the pattern. So, what do we mean by a pattern? First, we saw whether it was an upward pattern or a downward pattern, then we checked whether it was a curved pattern or whether it was a linear pattern, this is these 2 are linear patterns, whereas these 2 are curved patterns. Now, we are interested in knowing that along this pattern is my data tightly clustered.

This is the first question or along the pattern is my data variable very high. So, this first example gave me an example of a tightly clustered data whereas the second example is giving me an example of more variable data, both the cases I have an increasing trend which appears linear whereas this is more tightly clustered along the pattern, whereas this is more variable along the pattern.

(Refer Slide Time: 07:44)



Statistics for Data Science -1
└Association between numerical variables
└Describing association

Describing association: Variation

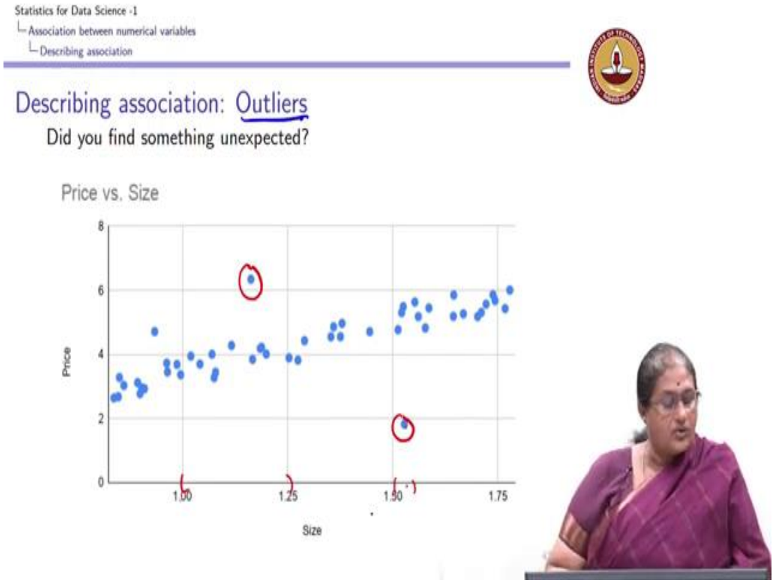
Are the points tightly clustered along the pattern?

Tightly clustered

Variable

So, that is the next thing which we check this example was tightly clustered. Whereas you can see that this example is more variable or it is not tightly clustered. So, this is the third thing which is check is, how varied is one of the variable or what is the variation exhibited in the variable? This is the next important thing we checked.

(Refer Slide Time: 08:13)



Statistics for Data Science -1
└Association between numerical variables
└Describing association

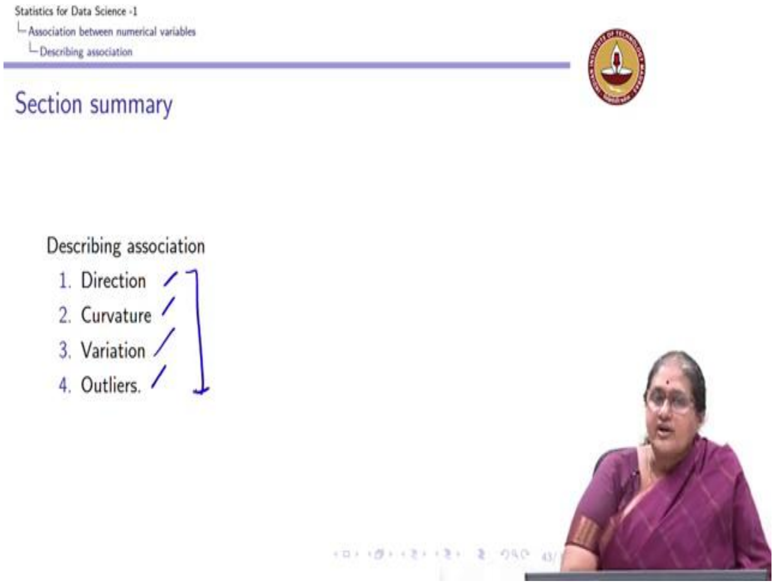
Describing association: Outliers

Did you find something unexpected?

The last thing which we check is what we refer to as the presence of an outlier. Again refer to the housing data. Now, look at these 3 points. So, there is this in fact, look at these 2 points which I am circling in red and there is this point, this is not an outlier, but now for now, let me just focus on these 2 points, you can see that all the other points are actually behaving according to a particular pattern whereas these 2 points are away from the regular pattern does exhibited by the other points.

Now, if you look at this point, this tells us about a house which is between 1000 square feet and say 1250 square feet which is actually priced higher than the usual houses in this interval. And what this says is a large sized house which is priced lower than the smallest house also. So, these 2 points are referred to as outliers. So, outlier means or it is not following the pattern which other points exhibit.

(Refer Slide Time: 09:52)



Statistics for Data Science -1
└Association between numerical variables
└Describing association

Section summary

Describing association

1. Direction
2. Curvature
3. Variation
4. Outliers.

So, when we are looking at association, in summary, the key things which we are trying to look at when we want to talk about association is. First we understand the direction whether it is an upward direction or a downward direction, whether the plot has an upward or a downward direction? Whether it is linear or it is curved? Whether it exhibits variation whether it is tight or whether it is varied?

And finally, whether there are presence of outliers? So, by this time we should know how to plot a scatter plot and look for the association between the variables through visual inspection when we are doing visual inspection these are the 4 key things which we need to take into account.