

**IIT Madras**  
ONLINE DEGREE

**Statistics for Data Science 1**  
**Professor Usha Mohan**  
**Department of Management Studies**  
**Indian Institute of Technology, Madras**  
**Lecture – 4.8**

**Association between two numerical variables- Fitting a Line**

(Refer Slide Time: 0:16)

Statistics for Data Science -1

- Association between numerical variables
  - Fitting a line

Learning objectives

1. Summarize the linear association between two variables using the equation of a line.
2. Understand the significance of  $R^2$

0:16

So, we have seen that this correlation and covariance are measures of linear relationship or linear association between two numerical variables; it measures a strength of a association.

(Refer Slide Time: 0:34)

Statistics for Data Science -1

- Association between numerical variables
  - Fitting a line

Learning objectives

1. Summarize the linear association between two variables using the equation of a line. ?
2. Understand the significance of  $R^2$

66/78

So, when I say that the association is a linear association. The next natural question to ask is, can I summarize this linear association through a mathematical equation in particular? The question we are asking as can this linear relationship be summarized through a equation of a line.

(Refer Slide Time: 1:03)

Statistics for Data Science - I  
└ Association between numerical variables  
└ Fitting a line

Summarizing the association with a line

$(x, y)$

- ▶ The strength of linear association between the variables was measured using the measures of Covariance and Correlation.

67/78

So, the question we are asking here is, when I am saying that  $x$  and  $y$  are numerical variables and they have association which I expect to be linear.

(Refer Slide Time: 1:15)

Statistics for Data Science - I  
└ Association between numerical variables  
└ Fitting a line

Summarizing the association with a line

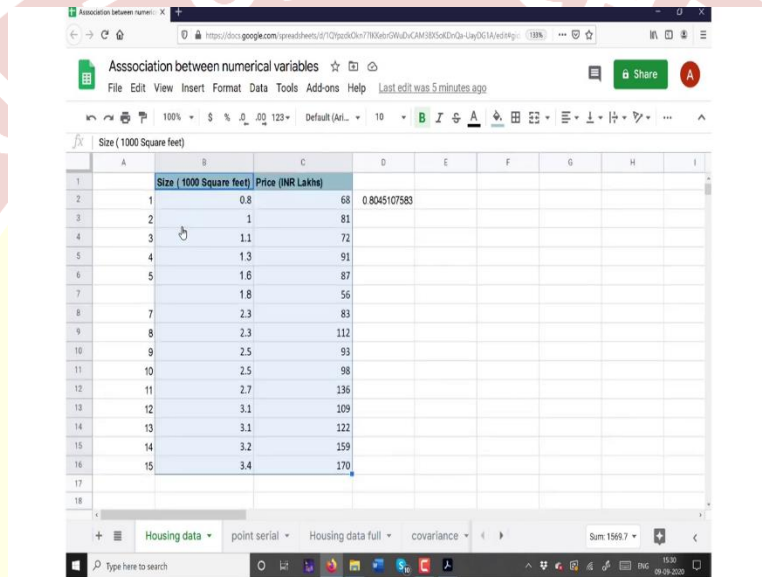
- ▶ The strength of linear association between the variables was measured using the measures of Covariance and Correlation.
- ▶ The linear association can be described using the equation of a line.

Best line of fit

67/78

Then the next question we are asking is can I described that association using the equation of a line? If yes, then how do I compute this equation of a line? The mathematics behind coming up with an equation of a line is beyond the scope of this particular course. You will be learning about how to come out with what we call the best line of fit in your future courses. But, what I want you to see at this point of time is, yes; the relationship can be summarized. How do I summarize this?

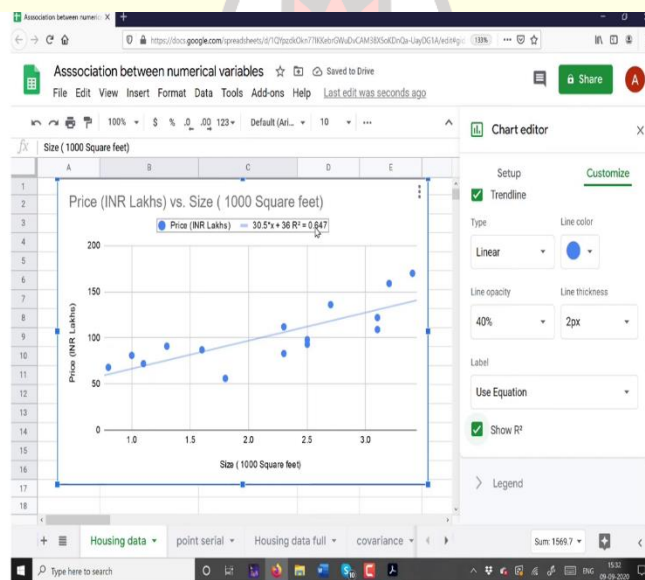
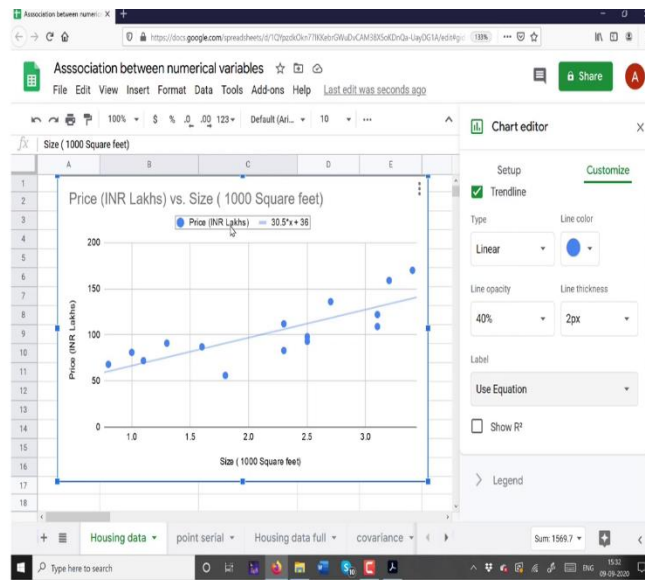
(Refer Slide Time: 2:06)



	A	B	C	D	E	F	G	H	I
1		Size (1000 Square feet)	Price (INR Lakhs)						
2	1	0.8	68	0.8045107583					
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						
12	11	2.7	136						
13	12	3.1	109						
14	13	3.1	122						
15	14	3.2	159						
16	15	3.4	170						

So, let us go back to our Google sheets; so you can see that I start with my first data here, which is my size wise is the price. I go to this data set. So, how do I find the equation of a line using a Google sheet?

(Refer Slide Time: 2:28)



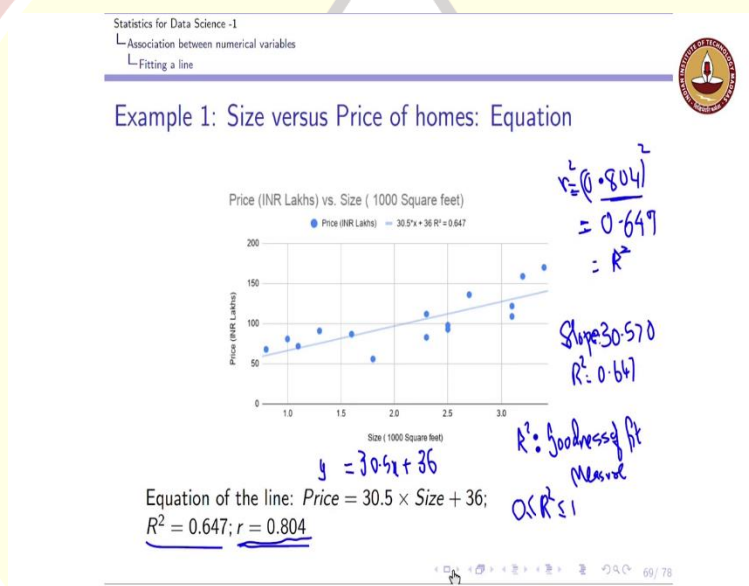
So, the first step is you open the scatter plot. So, I first go and I insert a scatter plot and this scatter plot is what I have here; so first I go and I insert this scatter plot. So, the next step is under the Customize tab, so I have a scatter plot here. This is scatter plot of size versus price; size is my explanatory variable, price is my response variable. Under the Customize tab click on Series, so under the Customize tab I click on series here; now within this click on trendline.

I clicked on series, I clicked on trendline; so, you can see a blue line which passes through my points. Now, the next thing is this trendline, so the question we ask is can I capture this linear relationship using the equation of a line? I can see that I can fit a line through the points passing

through the points. Now, further if I want to know what is the equation under the Label tab; so within this I have a Label tab. Under this Label tab click on this Use Equation; so it gives me the equation. You can see here the equation states price to equal  $30.5 \times x$ ;  $x$  here is my size +36.

This is of the type  $y = mx + c$ , which all of us know is the equation of a straight line. Now, further I can also ask it to report this  $R^2$ ; so it gives an  $R^2$  of 0.647. What is this  $R^2$  capture? This  $R^2$  actually captures the proportion of variance in my data set; that is captured by this line. Again to go into the mathematics of this  $R^2$  and how to derive it is beyond the scope of this course; but,  $R^2$  basically is also a measure of how good a fit is this line.

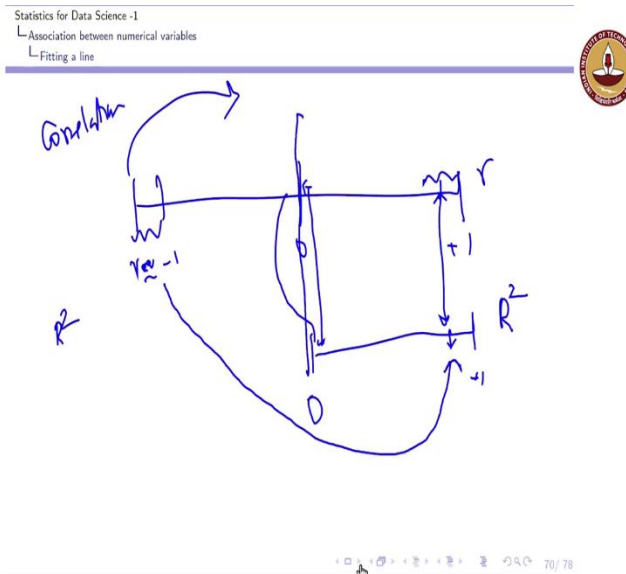
(Refer Slide Time: 5:01)



This  $R^2$  takes the value, so this  $R^2$  is also referred to a goodness of fit measure. It is an  $R^2$ , so this takes values between 0 and 1. The closer it is equal to 1, says that my fit is a good fit; the closer it is to 0, tells my fit is not a good fit for my data. So, now let us look at what is this  $R^2$  for different examples which we have looked. In the first example this is what I have demonstrated just now, I have price which is my  $y$ ; the response variable is  $30.5 \times x +$  my 36. So, the slope of this line is 30.5, you can see that the slope is positive; and your  $R^2$  is 0.647.

Recall my correlation coefficient was 0.804. An interesting observation is if I square this correlation coefficient; so you can see that I computed the correlation coefficient of my first data set to be 0.804. If I just square this term, I get 0.647 which is precisely this value of my  $R^2$ .

(Refer Slide Time: 6:45)



So, let us go back to our data here, so you can see that when I have a correlation which is closer to 1; my  $R^2$  which is the square of this would also be closer to 1. This is about  $r$ , the first thing is my  $R^2$  will only lie between 0 and 1; so, in a sense you can imagine to flip this along this 0. So, if I have a value which is here  $r$  value here; my  $R^2$  is going to be closer to 1. So, as my  $r$  goes to 0, my  $R^2$  will also tend to 0.

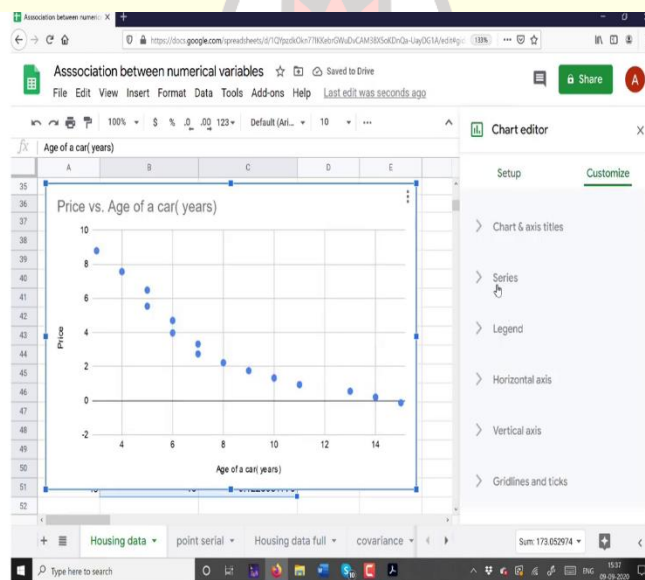


(Refer Slide Time: 7:35)

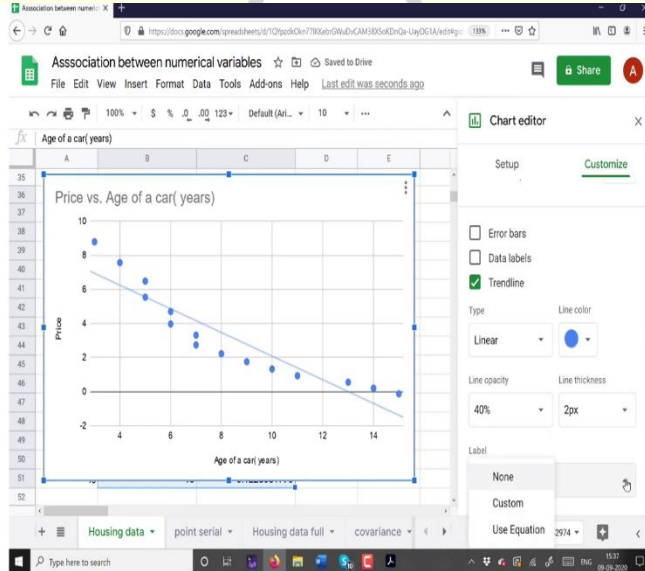
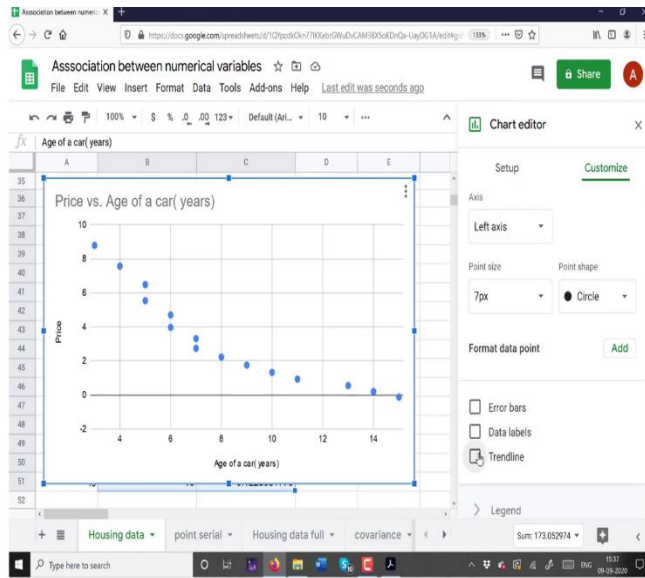
The screenshot shows a Google Sheet with the following data:

	Age of a car (years)	Price
1	3	8.8
2	4	7.576
3	5	6.49112
4	5	5.5472744
5	6	4.706128728
6	6	3.97431993
7	7	3.317668834
8	7	2.746371886
9	8	2.229343541
10	9	1.75952688
11	10	1.330790126
12	11	0.9377874095
13	13	0.5558750483
14	14	0.2036112903
15	15	-0.1226581775

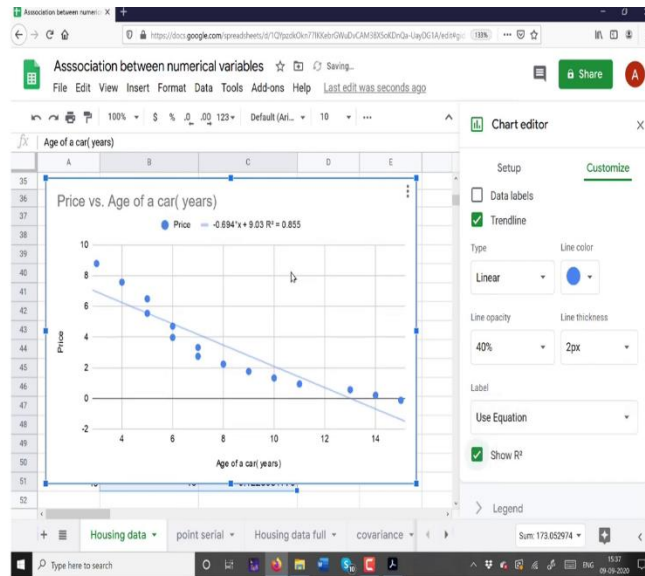
The formula bar shows the covariance calculation:  $\text{COVARIANCE.S}(\text{B37:B51}, \text{C37:C51})$  resulting in -0.9271053621.





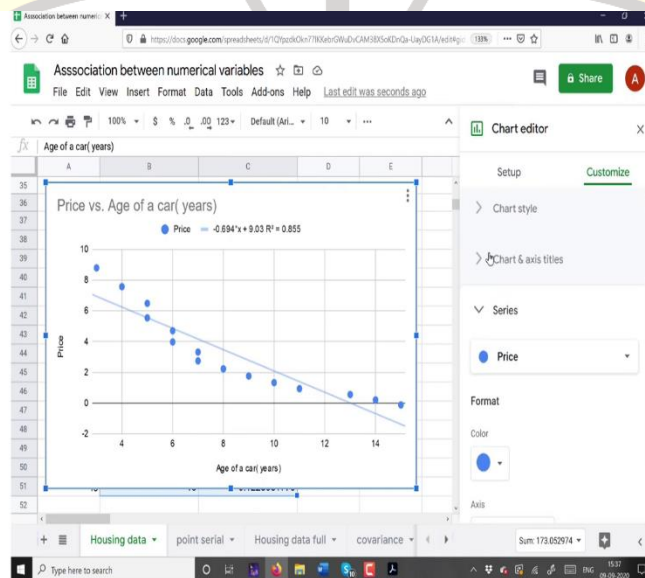


सिद्धिर्भवति कर्मजा



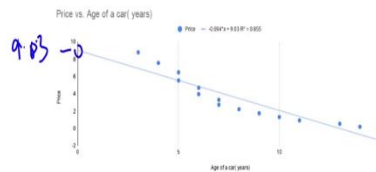
Let us look at an example to illustrate what we have just discussed. So, if I go and look at the same thing for this data set, we know about the age of a car and the price of a car. This is the next data set. Again I go and I click on a scatter plot, again I go to my Customize option; under series I plotted a trendline. And again I ask for a equation with a  $R^2$ .

(Refer Slide Time: 8:14)



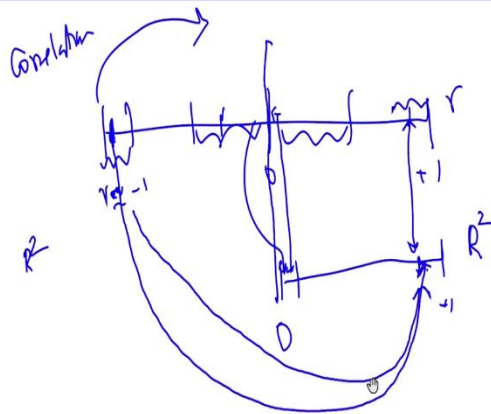


## Example 2: Age versus Price of cars: Equation



Equation of the line:  $\text{Price} = -0.694 \times \text{Age} + 9.03$ ;  
 $R^2 = 0.855$ ;  $r = -0.9247$

71/78



70/78

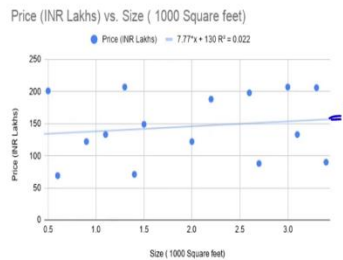
So, here what you noticed? You notice that the price is  $-0.694x + 3$ . With again an  $R^2$  of 0.855; which is price equal to  $-0.694$ . So, my slope of the line is negative, which is 0.694; the intercept is 9.03. And then we also see that the  $R^2$  is 0.855; even though your  $r$  was  $-0.92$ . So, my  $r$  in this case or my correlation in this case was actually here; the correlation was closer to  $-1$ . It was very strong negative linear relationship; my  $R^2$  is closer 1.

So, you can see that this goodness of fit measure takes a values between 0; and 1 but it does not tell me the direction of the relationship. For the direction, whether it is a positive relationship or negative relationship. I either look at the correlation coefficient or the sign of the slope; in this case the slope is a negative number.

(Refer Slide Time: 9:40)



### Example 3: Size versus Price of homes: Equation



Equation of the line:  $Price = 7.77 \times Size + 130$ ;  
 $R^2 = 0.022$ ;  $r = 0.149$

72/78

Now, let us look at the third example. The third example my  $r$  was very negligible; so it says that almost no relation. It is giving me the equation of a line; but what you have noticed about this line, it is almost parallel to my  $x$  axis. But, and your  $R^2$  is closer to 0. So, when your  $R^2$  is closer to 0; the goodness of the fit measure is also very low. So, when your  $R^2$  is close to 0; it quantifies the strength of the relationship. And you can say that my line is not actually describing this association very well.

(Refer Slide Time: 10:26)



### Section summary

$$y = mx + c$$

Response =  $\frac{mx + c}{\text{Explanatory}}$

$m > 0 \rightarrow \text{+ve}$   
 $m < 0 \rightarrow \text{-ve}$

1. Equation of a line describing linear relationship between two variables.
2. Interpreting slope,  $R^2$  of the line.

Association between Categorical	Association between Numerical
1. Contingency table	1. Scatter plot - visual
2. Relative frequencies	2. Covariance
3. Chi-square test	3. Equation of line, $R^2$



So, in summary what we have looked that is how do we obtain the equation of a line; we are not again gone into the mathematics of trying to find, what is a line of best fit. However, we got this equation of a line through our Google sheets. And when we looked at the equation of a line; so I get it as  $y = mx + c$ ,  $y$  is my response variable,  $x$  is my explanatory variable.

And  $m$  is the slope of the line,  $c$  is the intercept. The sign of the slope, so if  $m > 0$ ; then this says that I have a positive relationship.  $m < 0$ , it says that I have a negative relationship; and we also define what is  $R^2$ , which is a goodness of fit measure.  $R^2$  lies between 0 and 1 and if it is closer to 1; we are getting equations of the line. It says that the fit, I am talking about when it comes to a fit; I mean that the line is not capturing the variability in the data as much as in the other case. So, the proportion of variability in my data set that is captured by this line is very low, if  $R^2$  is closer to 0. And it is pretty high, if the  $R^2 = 1$ .

So, with this we actually have seen the following. In this section we started by looking at the association between two numerical variables; earlier we looked at association between categorical variables. Now, in this case the key thing is, we first learned about how we set up what we called is contingency table. Here we started with a scatter plot to look at the association, and then we introduced the notion of relative frequency here.

We looked at row relative frequency and column related frequency. If they are the same for all rows and columns, then after we set. When we looked at association between numerical variables, we started by looking at a scatter plot. From here we wanted to just have a visual inspection, within the visual inspection we identified what was the direction.

Whether it is a positive trend or a negative trend, whether it is a curve or a line, whether there they are tight, whether they clustered, or the presence of outliers; these are the four things which we looked at. We finally then we said that okay I am focus on a linear association between variables.

Now, if I want to know the strength of this association; I introduced two main numerical measures, which are covariance and correlation measures. And finally we also looked at how to summarize or describe this linear relation through a equation of a line. We introduced the concept of  $R^2$  which is nothing but the goodness of the fit measure.

(Refer Slide Time: 14:16)

Statistics for Data Science -1  
 Association between numerical variables  
 Fitting a line

Section summary

Variable

Categorical Numerical

- Equation of a line describing linear relationship between two variables.
- Interpreting slope,  $R^2$  of the line.

Association between Categorical	Association between Numerical
1. Contingency table	1. Scatter plot - Visual
2. Relative frequencies	2. Covariance
3. Chi-square test	3. Equation of line, $R^2$

Direction: Curve-Linear, Tight, Outliers

So, now if you look at variables, again you go back to your where we started from. And we saw that when we look at variables; I can broadly classify my variables or my data as categorical data and I can numerical data. So, this portion looked at association when both my variables or pair of variables are categorical. This looked at what happens when pair of variables are numerical in nature.