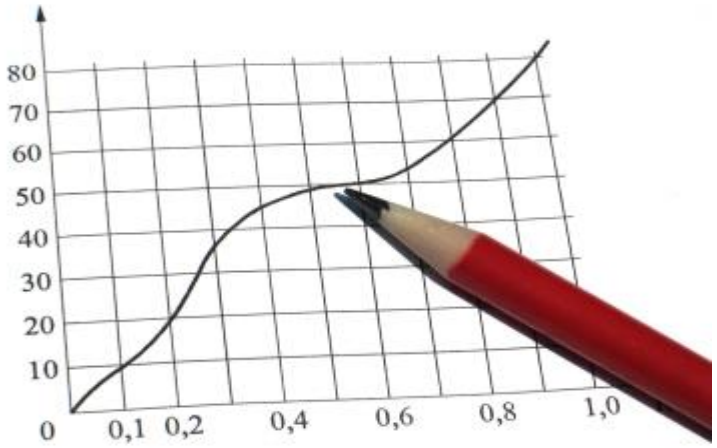
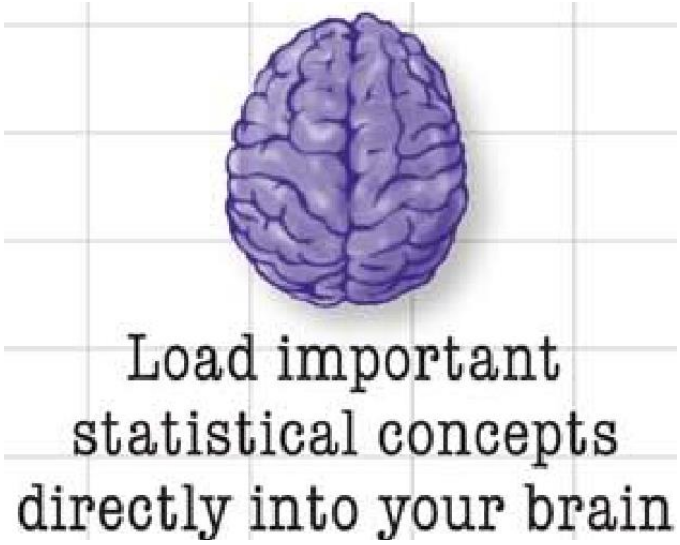
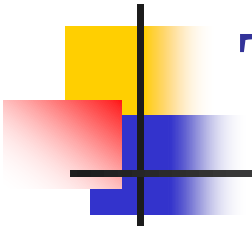


Correlation and Regression



Nikesh Bajaj

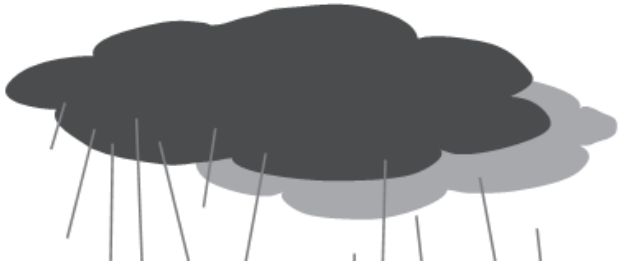
http://nikehsbajaj.in



Two Variable

- Till now, We dealt with one variable
- Let’s try to find out,
  - How two things are connected?
  - How they effect each other?

Concert and Weather



- *Guys: organizing concerts*
  Concert are best in open air
- *Ticket sales in summer look promising*

Let's Analyze and predict

- Sunshine and Attendance of audience

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100's)	22	33	30	42	38	49	42	55

  - ***Scenario***: Temperature is dipping, look like rain, guys want to predict attendance of audience for given hours of sun shine.
  - If attendance will be less than 3500, where ticket won't even cover the expenses they will cancel the event.

  - **What you can do with given data?**

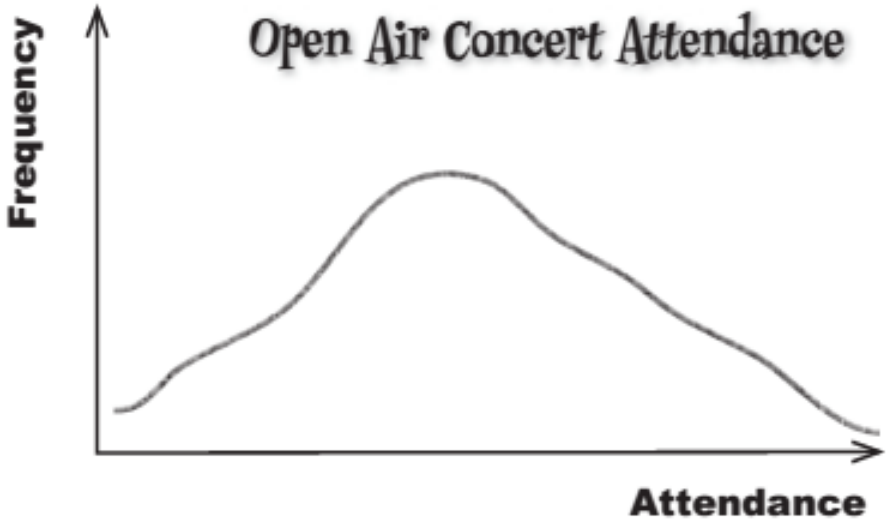
What sort of analysis you suggest?

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100's)	22	33	30	42	38	49	42	55

How would you go about modelling the connection between sets of data?

Exploring types of Data

- *Univariate Data*: Frequency or probability of one variable, e.g. weight, player's score etc. "One thing"
- It does not tell connection between two
- If



Exploring types of Data

- *Bivariate Data*: Values of two variable for each observation.

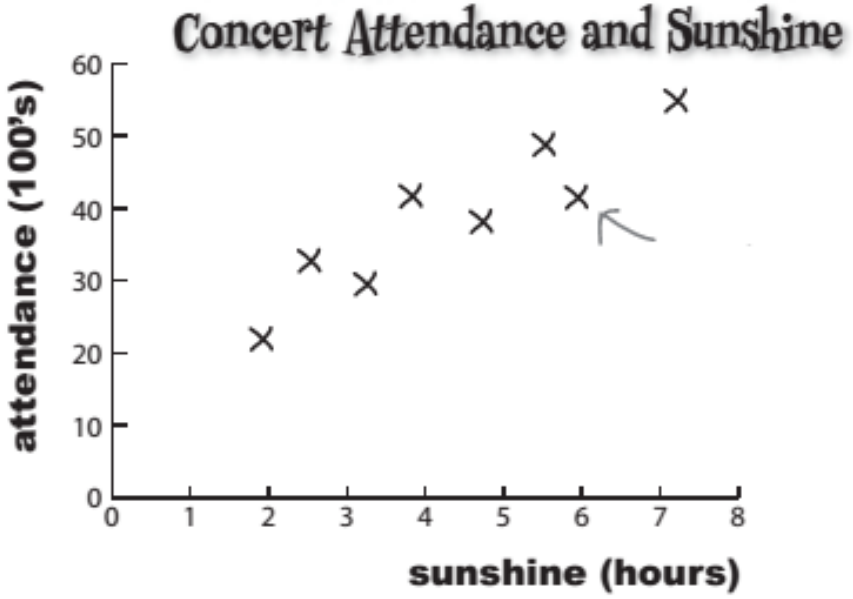
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100's)	22	33	30	42	38	49	42	55

  - *Independent or Explanatory variable*
    - *One of variable has been controlled in some way or used to explain other*
  - *Dependent or Response variable*
- *So Which is which for our example?*

Visualizing bivariate data

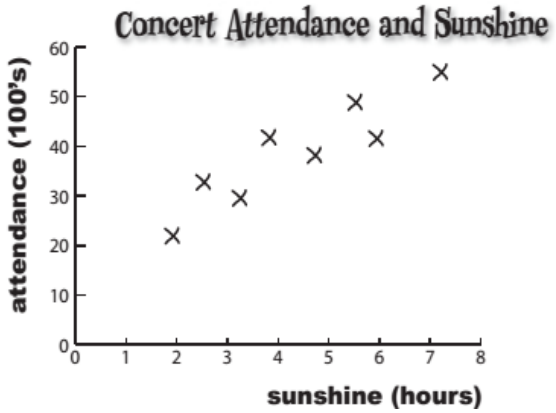
- *Scatter plot or scatter diagram:* **DOES IT HELPS?**

x (sunshine)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
y (attendance)	22	33	30	42	38	49	42	55



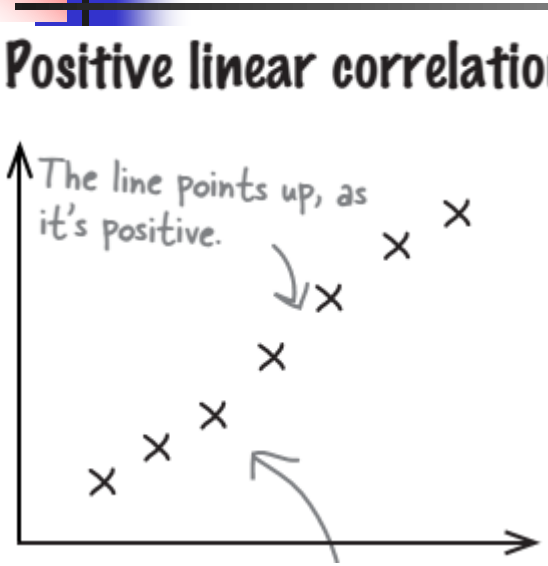
SO WHAT YOU CAN OBSERVE??

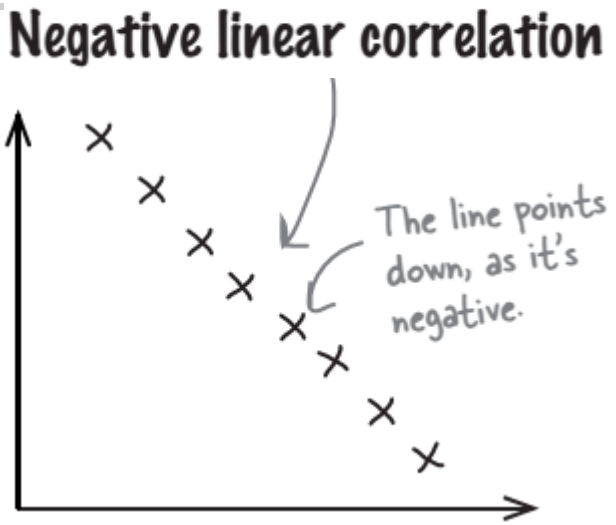
Correlation

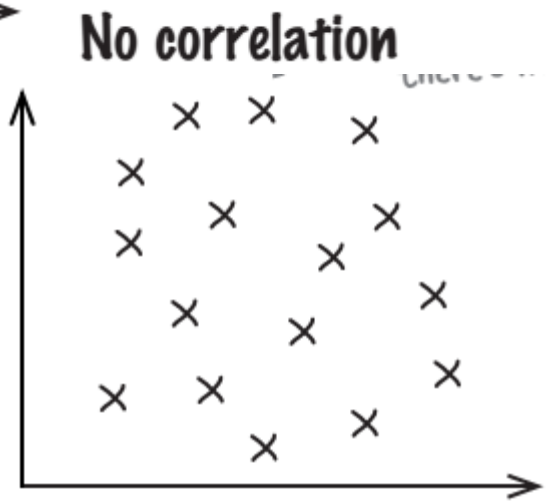


- Scatter diagram shows the correlation between two variable
- Correlation
  - Linear: If it is straight line, can be others

Correlation

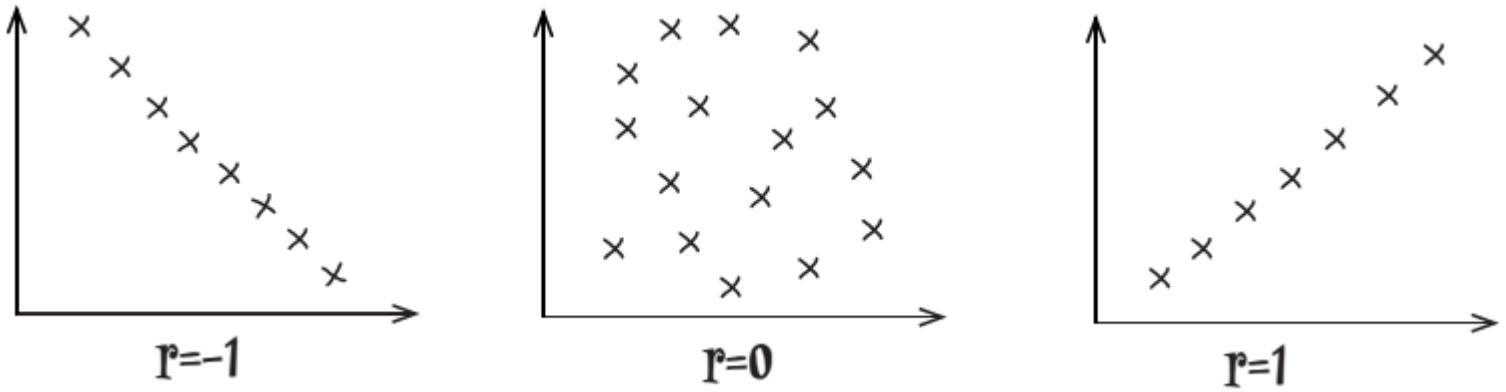




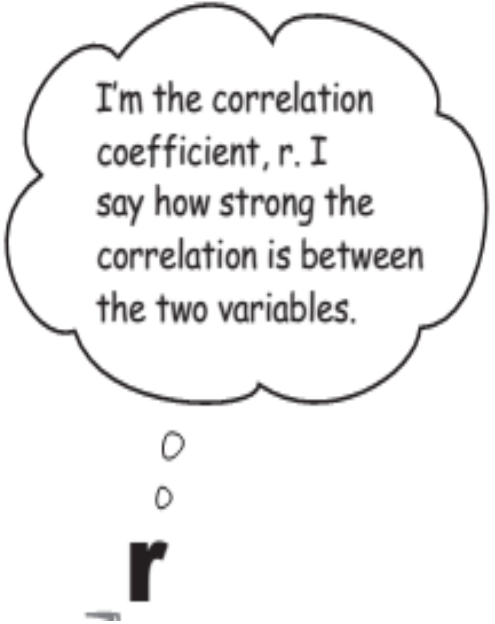


Correlation Coefficient r

- r tells you kind of correlation, positive, negative, perfect or no



Computing r



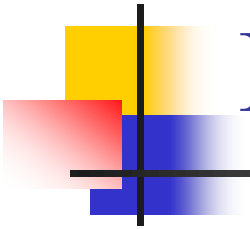
$$r = b \frac{s_x}{s_y}$$

$$b = \frac{\Sigma((x - \bar{x})(y - \bar{y}))}{\Sigma(x - \bar{x})^2}$$

$$s_x = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} \qquad s_y = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n - 1}}$$

Correlation and Causation

- If there is correlation between two variable Does that mean one caused the value of other??



Let’s See example

- One intern was given many scatter plot of..

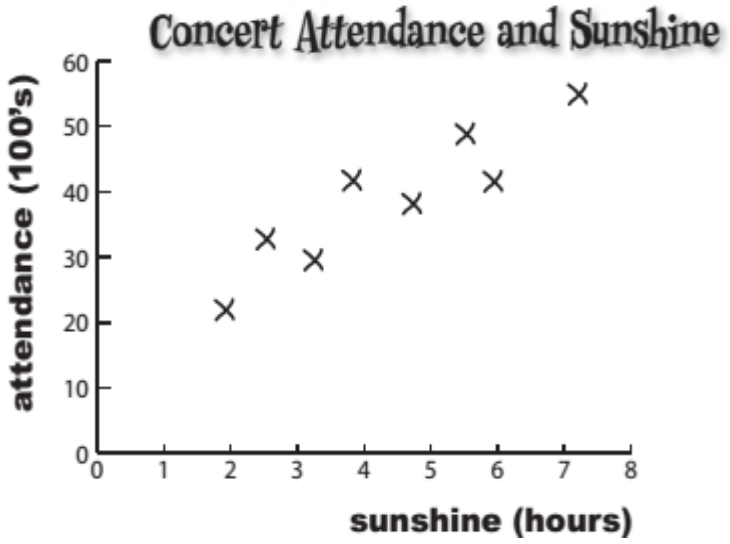
Correlation and Causation

- If there is correlation between two variable Does that mean one caused the value of other??

- "Not always"
- Let's see example

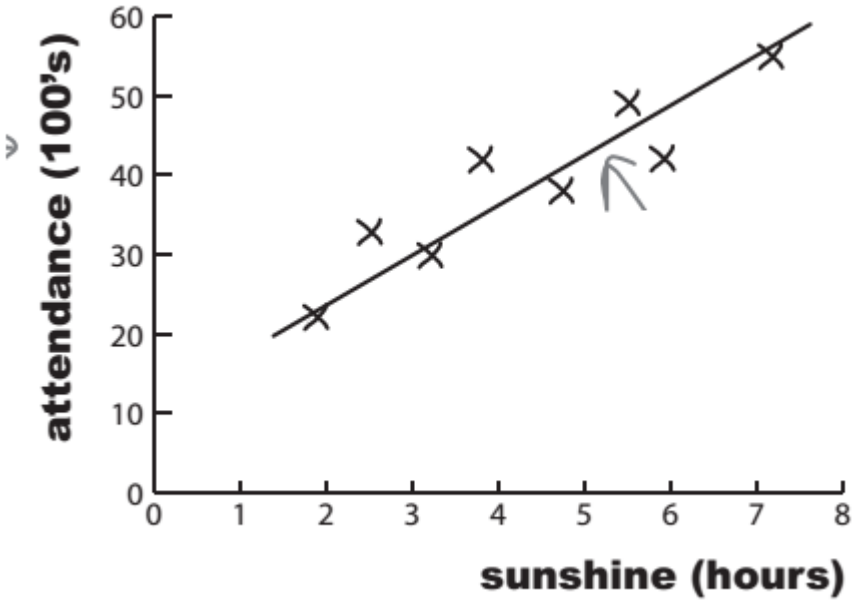
So for Concert

- Sunshine effect Attendance very much
- Good but
- *What about attendance of 3500 people??*

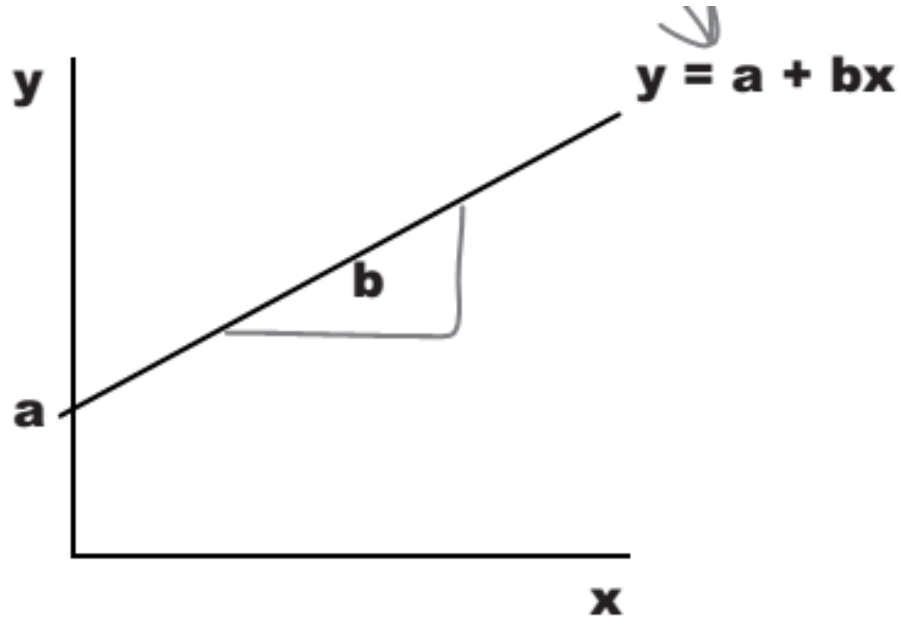


Predict the Attendance

- Line of Best Fit



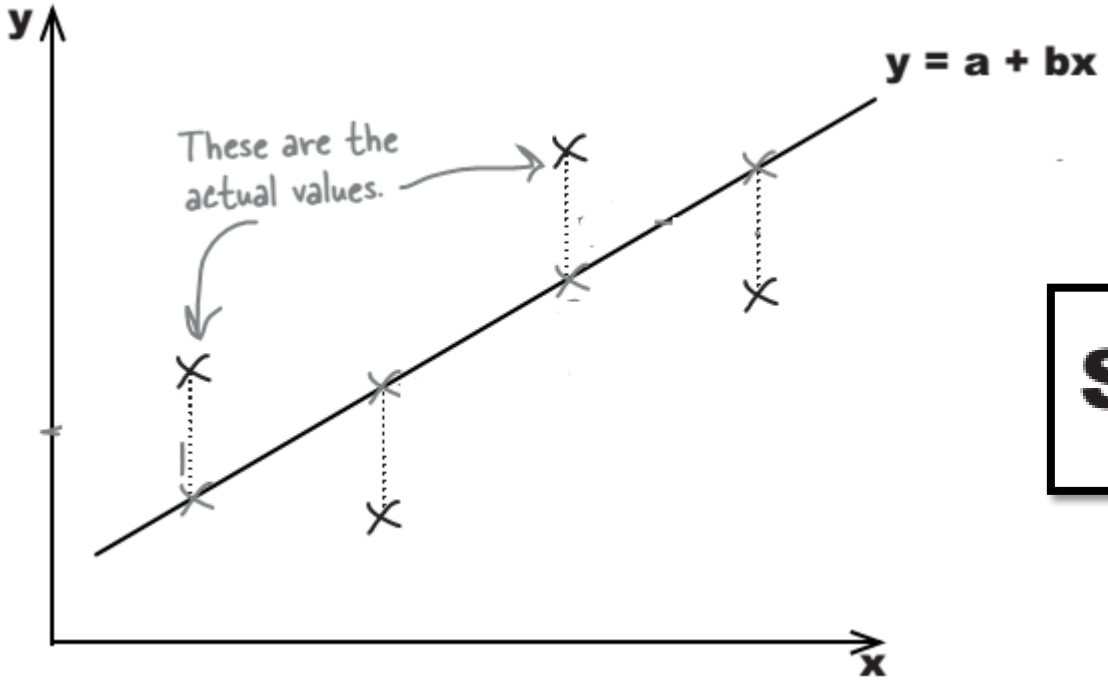
Find a Line y = a + bx



Line with minimum Error

$\Sigma(y_i - \hat{y}_i)$

- Error



$$\Sigma(y_i - \hat{y}_i)$$

$$SSE = \Sigma(y - \hat{y})^2$$

Let's find Line y= a + bx

- b: Steepness of line, Slop

$$b = \frac{\Sigma((x - \bar{x})(y - \bar{y}))}{\Sigma(x - \bar{x})^2}$$

No need proof right now

x (sunshine)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
y (attendance)	22	33	30	42	38	49	42	55

- Find b = ?

What about ‘a’ ???

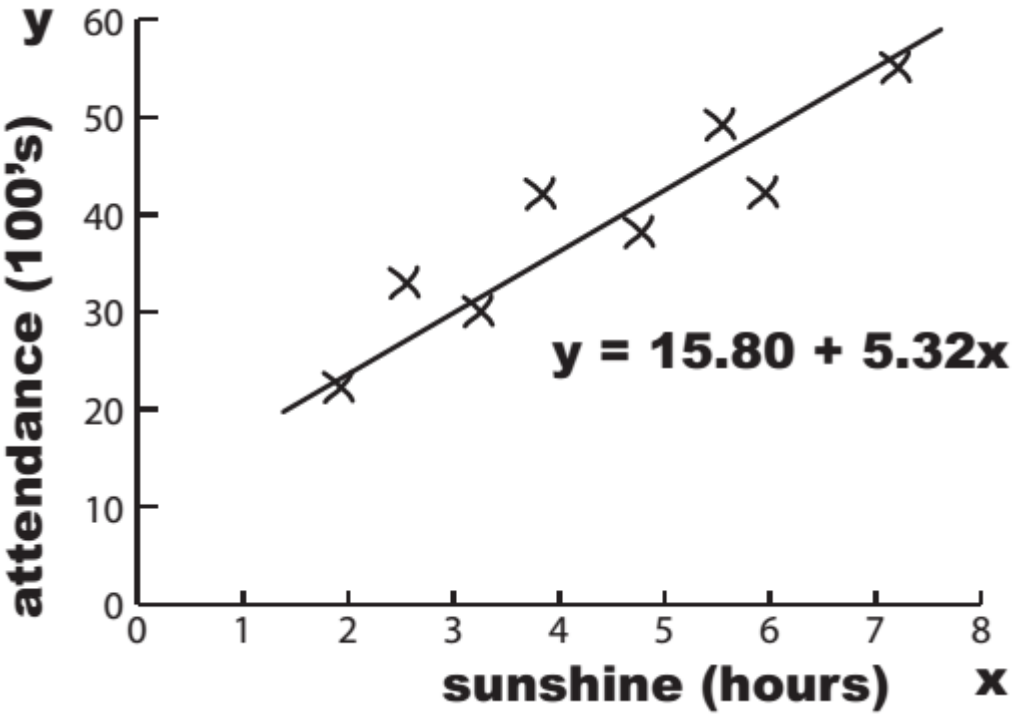
- How to compute?
- $y = a + bx$

$$\bar{y} = a + b\bar{x}$$

$$\mathbf{a = \bar{y} - b\bar{x}}$$

Solution (*Linear Regression*)

- Line of best fit



Now Answer Concert guys

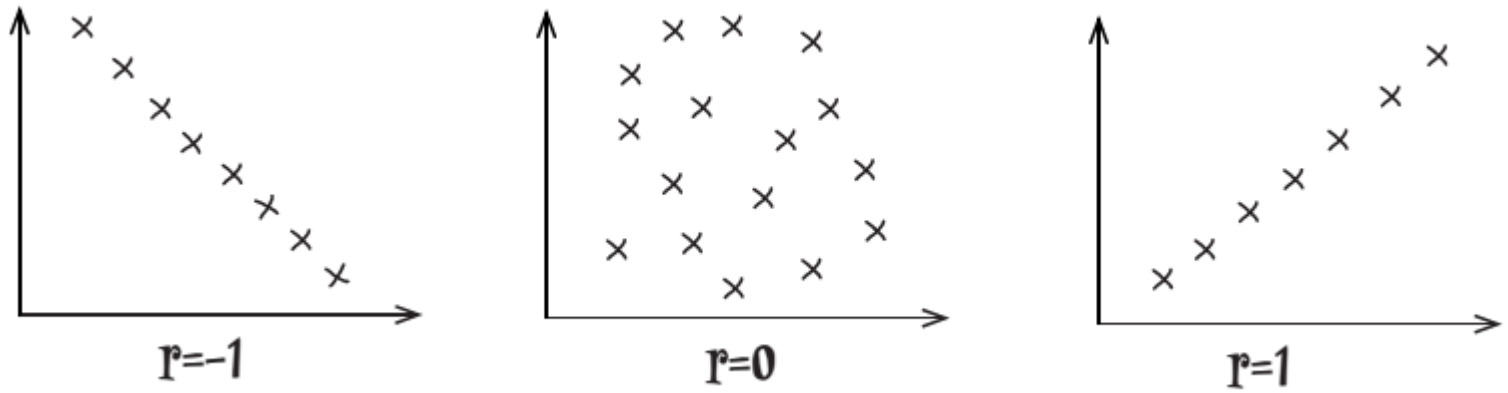
- y = 15.8 + 5.32x
  - Q1. When predicted sunshine is 6 Hours what would be attendance of audience in concert?
  - Q2. what should be sunshine hours for at least audience of 3500

Answers

- Ans 1: y =47.72 means 4772 people
- Ans 2: x=3.61 Hours

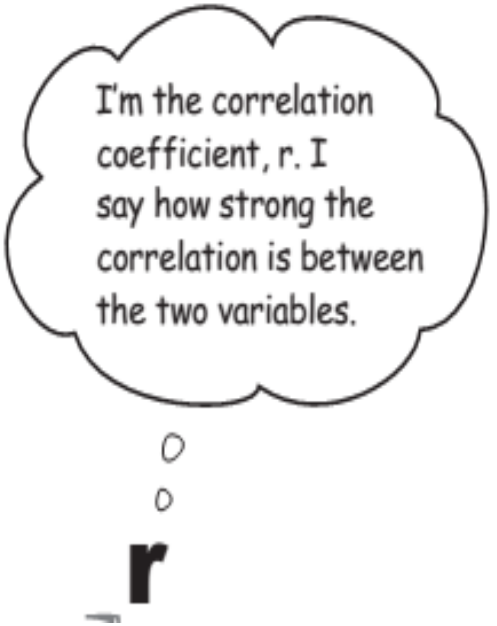
Correlation Coefficient r

- r tells you kind of correlation, positive, negative, perfect or no



Computing r

$$r = b \frac{s_x}{s_y}$$



$$s_x = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

$$s_y = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n - 1}}$$

Compute r (Correlation Coeff.)

x (sunshine)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
y (attendance)	22	33	30	42	38	49	42	55

$$b = 5.32$$

Answer

x (sunshine)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
y (attendance)	22	33	30	42	38	49	42	55

$b = 5.32$, $s_x = 1.81$, and $s_y = 10.56$,

$$r = bs_x / s_y$$
$$= 5.32 \times 1.81 / 10.56$$
$$= 0.91 \text{ (to 2 decimal places)}$$

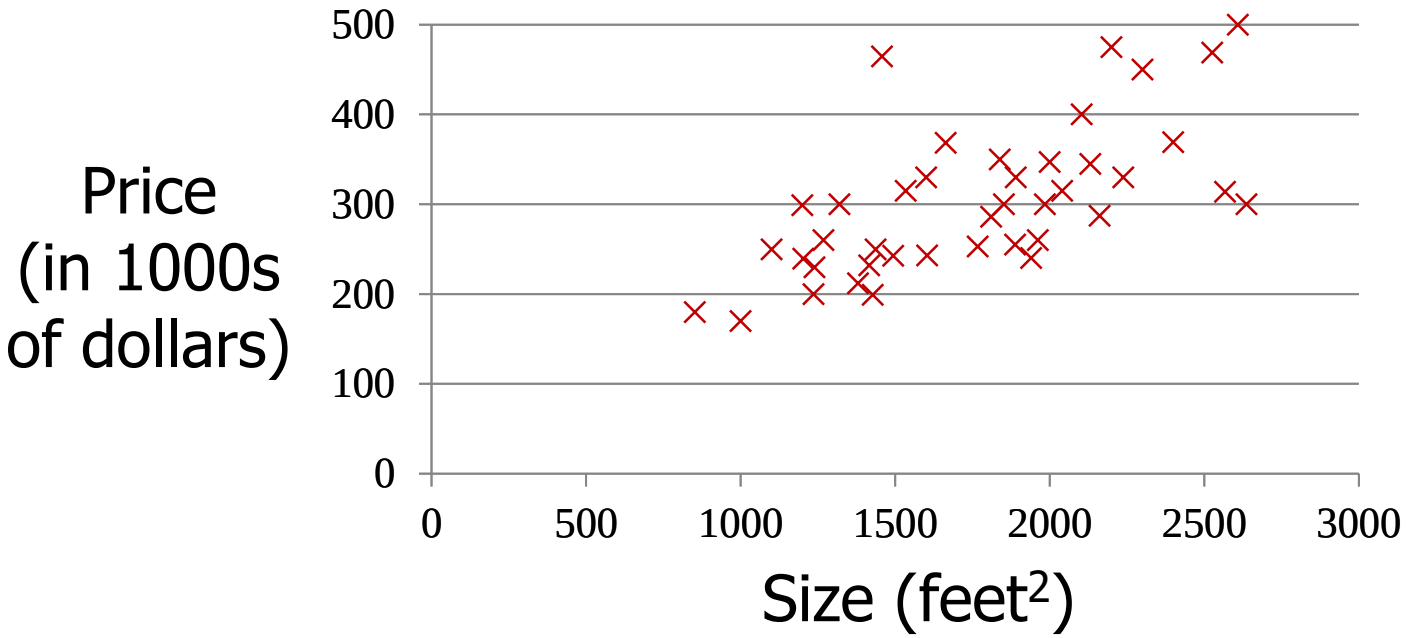
$$s_x = \sqrt{(23.02/7)}$$
$$= \sqrt{3.28857}$$
$$= 1.81 \text{ (to 2 decimal places)}$$

$$s_y = \sqrt{(780.875/7)}$$
$$= \sqrt{111.55357}$$
$$= 10.56 \text{ (to 2 decimal places)}$$

Exercise

Radiation exposure (minutes)	3	3.5	4	4.5	5	5.5	6	6.5	7
Weight (tons)	14	14	12	10	8	9.5	8	9	6

Housing Prices (Portland, OR)

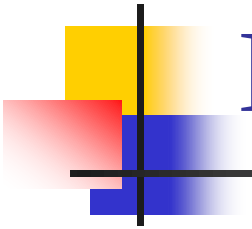


Supervised Learning

Given the "right answer" for each example in the data.

Regression Problem

Predict real-valued output



Links for Reading

- http://www.statistics.com/