"Line of Best Fit" (from Statistics)

Write a program that reads a data set of two-dimensional points and calculates the "line of best fit" for that point set. Plot the data and resulting line graphically.

Line of Best Fit is also called **Regression Line** in Statistics. Take a look at example:

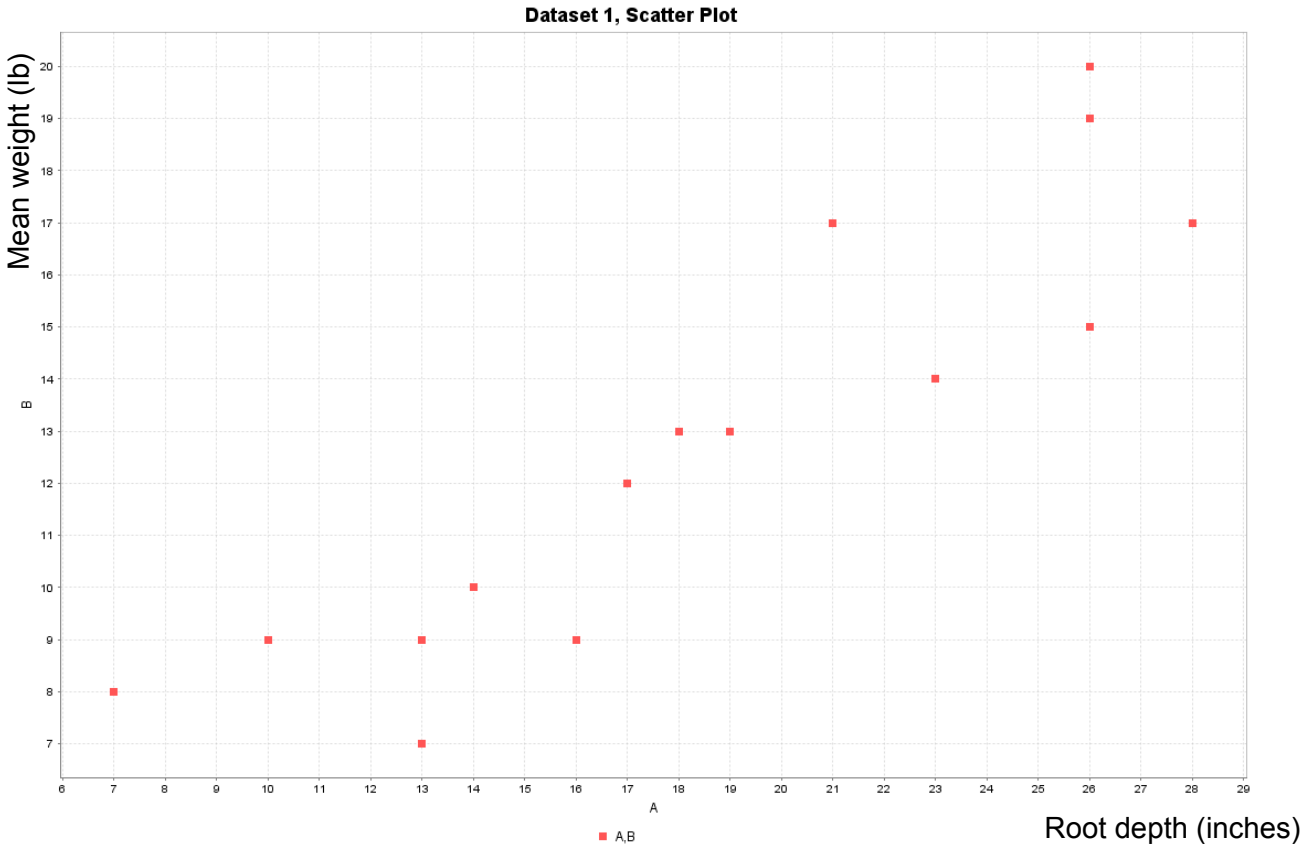
Example:
Is there a connection between the average weight of watermelons a vine produces and the root depth of the vine?
Suspected: vines with deeper roots have a better water supply, and thus larger average melons.
Large watermelon field, 15 vines are chosen at random. At the end of 8 weeks the watermelons are removed from each vine, weighed, and then the average weight (in pounds) is determined. Root depth of each vine is measured (in inches). Plot a scatter diagram. Find the Regression Line (if possible)

Root depth:	26	14	18	10	26	21	7	26	13	19	17	13	16	28	23
Mean weight:	20	10	13	9	19	17	8	15	9	13	12	7	9	17	14

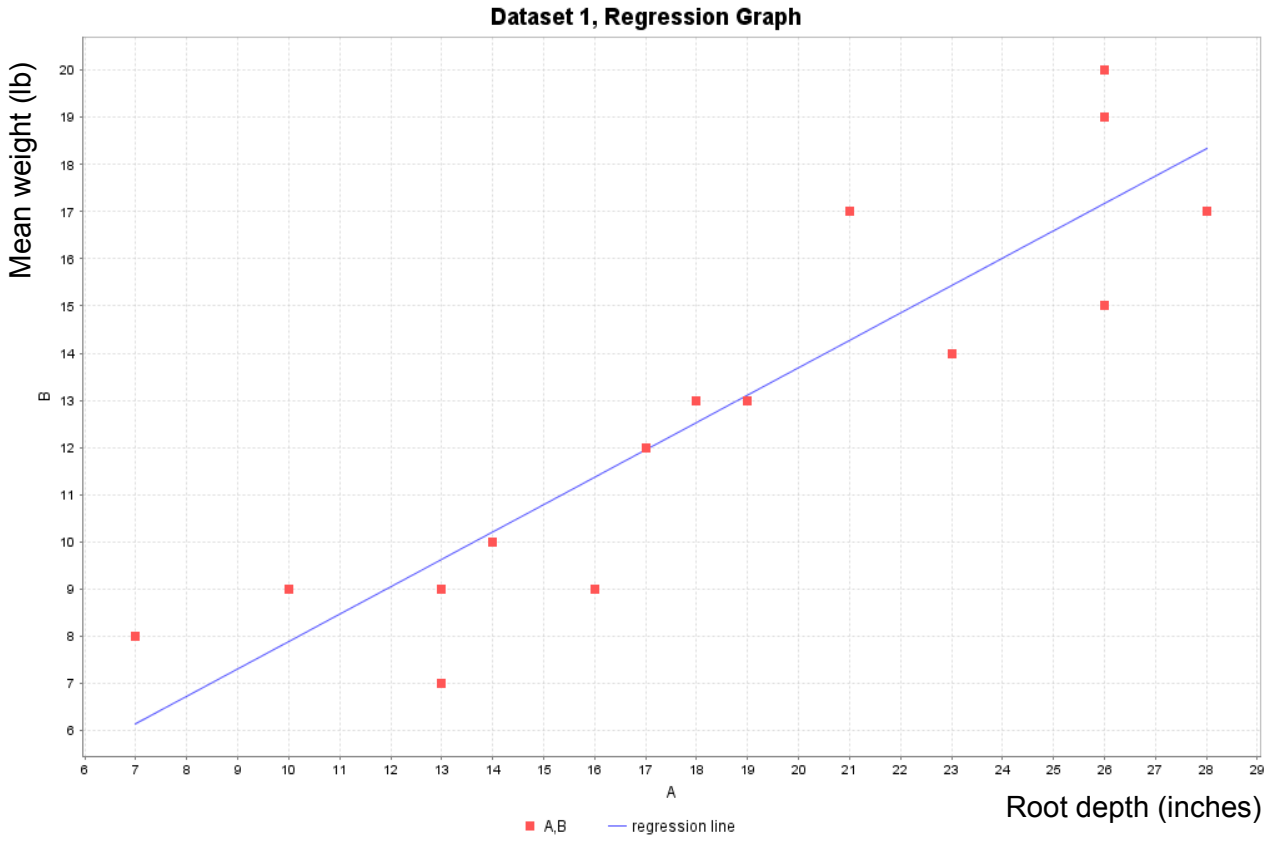
Let x (explanatory variable) be *root depth* and let y (response variable) be *mean weight*
y depends on x (we suspect that weight depends on root depth)

Therefore, *for each plant there is an ordered pair (x,y) : (root depth, mean weight)*

Let's plot the points (plot **scatter diagram**):



Can we draw a straight line, so that the points are close to it? (Regression Line)



First, we usually find the correlation coefficient (that tells us if it is possible to find the regression line). You don't need to implement this check in your project, but here are the formulas anyway.

Let's find correlation coefficient *r*. Recall that

Root depth:	26	14	18	10	26	21	7	26	13	19	17	13	16	28	23	**x**
Mean weight:	20	10	13	9	19	17	8	15	9	13	12	7	9	17	14	**y**

X	Y	X^2	Y^2	XY
26	20	676	400	520
14	10	196	100	140
18	13	324	169	234
10	9	100	81	90
26	19	676	361	494
21	17	441	289	357
7	8	49	64	56
26	15	676	225	390
13	9	169	81	117
19	13	361	169	247
17	12	289	144	204
13	7	169	49	91
16	9	256	81	144
28	17	784	289	476
23	14	529	196	322
$\Sigma x=277$	$\Sigma y=192$	$\Sigma x^2=5695$	$\Sigma y^2=2698$	$\Sigma xy=3882$

$$r=\frac{(n\sum xy-(\sum x)(\sum y))}{(\sqrt{n\sum x^2-(\sum x)^2}\sqrt{n\sum y^2-(\sum y)^2})}=\frac{(15*3882-277*192)}{(\sqrt{15*5695-277^2}\sqrt{15*2698-192^2})}=$$

$$= \frac{5046}{(\sqrt{8696}\sqrt{3606})} \approx \frac{5046}{(93.25*60.05)} \approx 0.901$$

Therefore, **r = 0.901** – close to 1, which means that there is a strong positive linear correlation between the root length and the watermelon weight.

Regression line has the following linear equation: **y = mx+b**, where slope **m** and y-intercept **b** can be found from the following formulas:

$$m = \frac{(n\sum xy - (\sum x)(\sum y))}{(n\sum x^2 - (\sum x)^2)} = \frac{(15*3882 - 277*192)}{(15*5695 - 277^2)} = \frac{5046}{8696} \approx 0.58$$

$$b = \bar{y} - b\,\bar{x} = \frac{\sum y}{n} - b\frac{\sum x}{n} = \frac{(\sum y - b\sum x)}{n} = \frac{(192 - \frac{5046}{8696}277)}{15} \approx 2.08$$

The equation of regression line is **y = 0.58x+2.08**

Comment to the example: n is the number of points (we have 15 points in this example)

Comments to the program:
For your program the input will be taken from a file. The file will be formatted so that each line describes a single point, denoted by its **x** and **y** coordinate values separated by a space.
Your output should report the equation of the line (**y = 0.58x+2.08** from our example), plot the points and draw the line (see the second figure: Regression Graph).

When you will be designing/developing your project refer to the Mastermind game:
Write a separate Input class, Output class, and RegressionLine class.
For output recall the example I showed on Lecture 14 (rec_coord_system.py).