San José State University
Department of Applied Data Science

DATA 201
Database Technologies for Data Analytics

Spring 2025
Sections 21 and 71
Instructor: Ron Mak

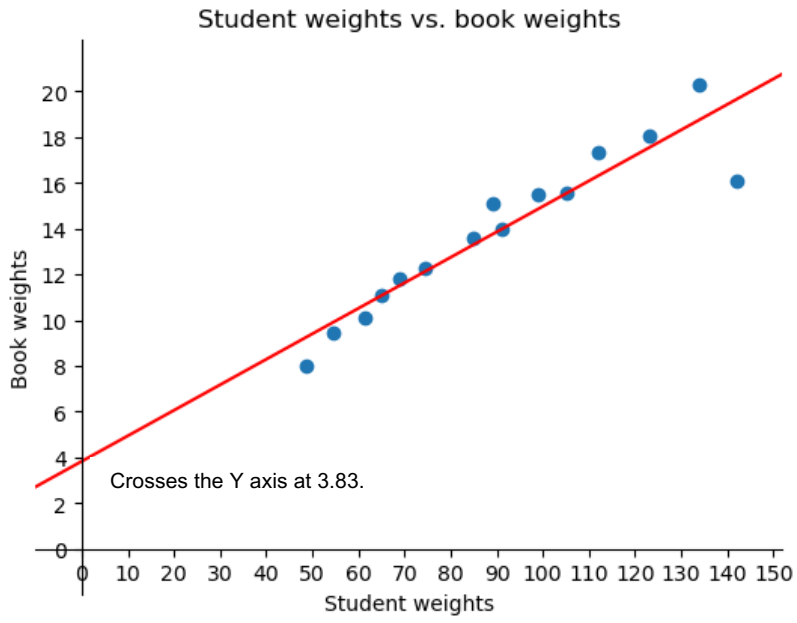
Assignment #9

Assigned: Thursday, March 27
Due: Thursday, April 10
Individual assignment, 110 points max

Linear regression calculations on the database server

The purpose of this assignment is to give you practice writing analytical code in SQL that runs on the database server, and you will query for their results. For this assignment, assume that you measured the weights (in pounds) of students and the weights (in pounds) of the books they're carrying. You want to investigate the relationship between the corresponding weights. Some sample data:

Student Weight (X)	Book Weights (Y)
48.50	8.00
54.50	9.44
61.25	10.08
65.00	11.07
69.00	11.81
74.50	12.28
85.00	13.61
89.00	15.13
91.23	14.00
99.00	15.47
105.00	15.53
112.00	17.36
123.00	18.07
134.00	20.29
142.00	16.06



One way to analyze the data is to calculate and plot the **linear regression line** of student weights vs. book weights. The regression line is shown above for the sample data. The linear regression line is defined by the equation

$$\hat{y} = mx + b$$

The value m in the regression equation is the slope of the regression line, which measures how much y changes for each unit change of x. The value b is the y-intercept, which is the value of y when it crosses the y axis at x = 0. The hat over the y indicates that the formula generates an estimated value for y. From the sample data, m = 0.11 and b = 3.83. **The values of the regression coefficients m and b determine the regression line.**

A formula for m is:

$$m = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

The formula for b is:

$$b = \bar{y} - m\bar{x}$$

where:

- n is the number of x and y pairs.
- $\sum x$ is the sum of all the x values (student weights) and $\sum y$ is the sum of all the y values (book weights).
- $\sum x^2$ is the sum of all the squares of the x values.
- $(\sum x)^2$ is the square of the sum of all the x values.
- $\sum xy$ is the sum of all the products each of which is a value of x multiplied by the corresponding value of y.
- $\bar{x}$ is the average of all the x values and $\bar{y}$ is the average of all the y values.

In class, you saw Python function **`calculate_slope_intercept(X, Y)`** which used Python statements and expressions to calculate and return the values of the regression coefficients m and b using these formulas.

But if the dataset large, such as containing millions of values of X and Y, you don't want to incur network traffic and latencies by downloading the data to the clientside in order to use a Python function. Instead, you should calculate the regression coefficients m and b using SQL code in the database server, and then only download those two values. To plot the regression line using the **`draw_graph()`** function given in class, you will also need to download the minimum and maximum values of X and Y.

Linear regression formulas in SQL

There are at least five ways to structure SQL code to implement the formulas that calculate the regression coefficients m and b given a database table of X and Y values:

1. Views
2. CTEs
3. Nested subqueries
4. Stored procedures
5. User-defined variables

For this assignment, chose four of the above five ways to implement your SQL code. Then use to query and download the regression coefficients m and b and the minimum and maximum values from each. The results should be the same each time and match the results from the Python function. Call the **`draw_graph()`** function once using the query results to create the regression line graph.

What to submit

A Jupyter notebook that contains:

- Python code that issues SQL code that implement four of the five ways described above.
- Python code that issues queries of each of the ways for the regression coefficients m and b and the minimum and maximum values. The results should be the same for each.
- A regression line graph generated from the queried values.

Rubric

Criteria	
Four out of the five ways to structure your SQL code and queries of each: • Way #1. • Way #2. • Way #3. • Way #4.	**100** • 25 • 25 • 25 • 25
One regression line graph based on the queried values.	**10**