LAB 1 - Simple Linear Regression

In this lab, we will perform simple linear regression in Python. To do this, we will need the .csv file dataset provided in Blackboard.

The .csv file contains information related to 100 imaginary football players. We will investigate the "Age", "Height" and "Salary" columns. The first two shall correspond to x and the latter shall correspond to y, in the upcoming mathematical equations. We will try and observe if there is a linear correlation between "Salary" and the other two. Simple linear regression is going to be the method of choice for this task. Please follow these instructions:

- (15 pts) Read the .csv file, extract the "Age", "Height" and "Salary" columns into three different arrays. When extracting these values to vectors, don't use standard lists. We will use `numpy` arrays since it is much easier to do element-wise mathematical operations with them, unlike standard lists.

- Implement the following function:

  - (35 pts) The first function takes in x and y as parameters, then computes and returns the regression coefficients (b_0, b_1). x and y correspond to singular columns of a dataset, such as the "Salary" column.

    The calculation of b_0 and b_1 consists of simple mathematical equations, where:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1\bar{x}$$

    $\bar{x}$ and $\bar{y}$ refer to the *average* values of the x and y datasets, respectively.

    Note: You can return multiple entities in a Python function, without needing to put these entities in a container.

- (15 pts) We will create two linear regression models:

  Model 1: Simple linear regression model where x is "Age", y is "Salary". Call the function you've implemented (see above) and store the coefficients.

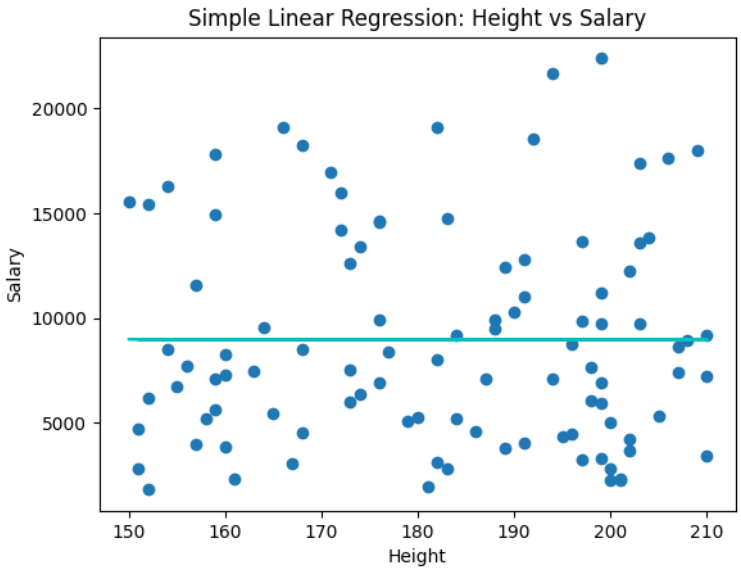
Model 2: Simple linear regression model where x is "Height", y is "Salary". Call the function you've implemented (see above) and store the coefficients.

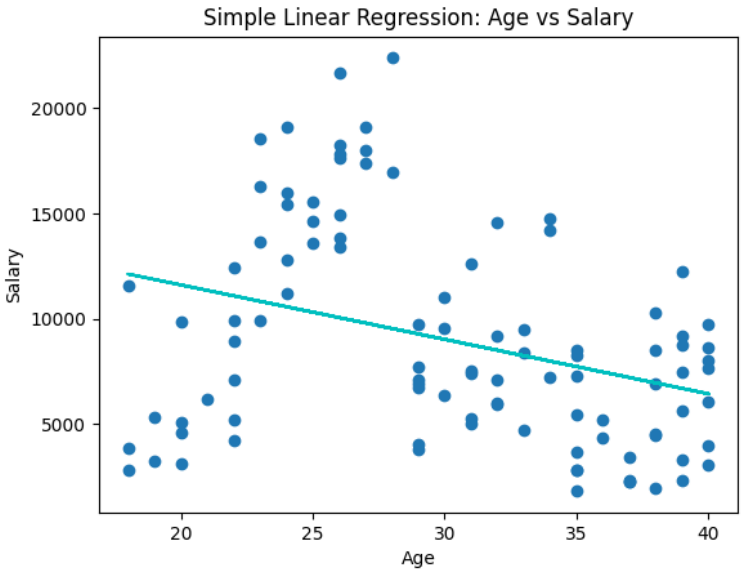
- (35 pts) Plotting:
  For each of your models, complete the following:
  - First, plot y against x as a scatter plot.
  - Then, calculate the regression line vector by using the coefficients of the model:

$$\hat{y} = b_1 x + b_0$$

  - Plot the regression line $\hat{y}$ against x as a line plot.

  - Each model should have its own plot in a separate figure.
  - Finally, once both models have been plotted, show the results. Here's how they look like:





IMPORTANT NOTE: The instructions specifically ask you to manually code the coefficient calculations. Bypassing these calculations (via using an external library, `sklearn` being the most popular one) will mean that the calculations are not done, therefore corresponding sections will yield 0 points in total.