Data Science Assessment

In the assessment you will implement a linear model for a set of 20 x/y data points.

We assume that the data can be described by a straight line with the slope `a` through the origin.

$\hat{y} = a * x$.

Task 1

Read the x/y data points from the file `datapoints.csv` into Python

Task 2

Set the slope a to 10. Calculate $\hat{y}$ for every value of x.

Task 3

Calculate the Mean Squared Error (MSE) of $\hat{y}$ and y using the formula:

$MSE = \frac{1}{N} \sum(\hat{y}_i - y_i)^2$

Task 4

Find a value for `a` that gives the lowest possible MSE. Implement the following procedure:

- increase `a` by `0.1`
- re-calculate $\hat{y}$ using the modified `a`
- re-calculate the MSE
- check if the new MSE is smaller than the previous one
- if it is smaller, use the new value for a, otherwise discard it
- repeat the procedure 100 times
- print the final value for `a` and the MSE

Task 5

How could the algorithm be improved? Write down one or two ideas.

Hints

- the implementation must be done in Python
- do not use any existing linear regression functions
- you may use `pandas` or `numpy`
- you may use `matplotlib` to plot the data