Lab 2: Linear and polynomial regression

For this lab you will load a new dataset and a subset of that dataset. The data to be used is available in Canvas and has been downloaded from the SCB website (www.scb.se). The full dataset includes **mean income by region, age and year** while the provided subset contains the average income by age from 20 to 50 years only.
The objective of this lab is to conduct regression analysis. The focus will be on performing multiple linear regressions using the same dataset to assess their effectiveness in various segments. Additionally, you will explore alternative regression techniques to address the dataset, followed by a reflective analysis of these options.

Note that in addition to the Tasks described below the following exercises in the course book are mandatory for completing this lab (Chapter-Exercise): **10-11, 11-4, 13-1**.

Task 1: Load first a subset of the data to be analyzed

Create a function to load data from a csv file based on basic file I/O functions described in the lectures. (Do not use pandas or a csv library to load the data, though you can use pandas to save the result). Using that function, load the file **inc_subset.csv** that is available in Canvas. Inspect the values of the loaded data elements using the debugger and print them on screen (you should use both the debugger and the printout to verify that the csv loader behaves as expected).

This dataset contains information regarding the average income by age from 20 to 50 years. The output should look something like this:

	age	2020
0	20	129.242857
1	21	146.309524
2	22	164.000000
3	23	182.395238
4	24	202.076190
5	25	224.138095
6	26	242.619048
7	27	258.233333
8	28	271.528571
9	29	282.833333

Discussion: Discuss with your lab partner. When is it more appropriate to use the debugger for inspecting variable values, and when would you prefer to use a print function? Identify:

- One case when it is more convenient/efficient to use the debugger.
- One case when it is more convenient/efficient to use a printout.

Task 2: Linear regression on a data subset.

For this task you will perform a linear regression on the average yearly income and age data of the year 2020. Develop the model from scratch using the vectorized functions presented in Lecture 8 and chapter 4 of the ML course book (do not use sklearn).

Another approach you can use which does not use vectorized functions can be found here:

https://www.toppr.com/guides/maths-formulas/linear-regression-formula/ and

https://statisticsbyjim.com/regression/mean-squared-error-mse/

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

$y = a + bx$ $\qquad$ $a = \frac{\sum y - b(\sum x)}{n}$ $\qquad$ $MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n}$

2.1 Perform the linear regression on the provided dataset

- It's important that you divide your dataset in validation and train dataset, discuss and explain why.
- We suggest that you divide the dataset in a 80 –20 proportion especially in this small dataset.
- Discuss whether you should take a random sample in this case.

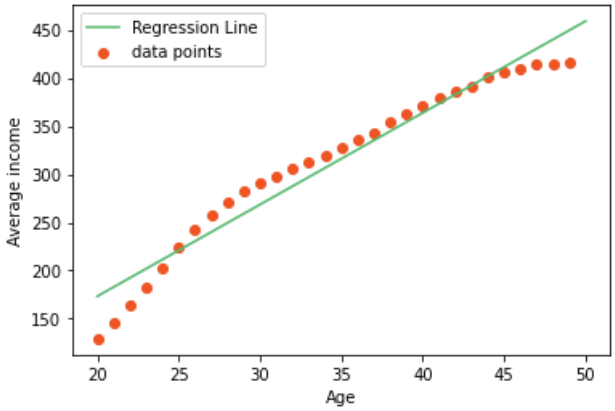
  Tip: You can use the pandas df.sample function to take a random sample of your dataset:

  validation _proportion=0.2 # 20% of the dataset

  validation _data = df.sample(frac= validation _proportion, random_state=42)

2.2 Create a scatter plot for your data and plot the regression line.

Checkpoint: A graph with age on the X axis should look like the following with a few datapoints from the data split.



2.3 Predict the values for the validation dataset and print both the real values and the predicted ones. How do they look?

2.4 Evaluate the model using MSE, implement the function that evaluates each of the real values against the predicted values (do not use sklearn MSE functions).

2.5 Explain the obtained MSE value and what does it mean.

Do you think linear regression is a good approach for this dataset?

Task 3: Linear regression on full dataset.

With the load function created earlier, load the second table **inc_utf.csv** that is available in Canvas. This dataset contains information regarding the average income by age and region.

Check that the data has been loaded successfully using the debugger and a printout on screen. The output should look something like this:

	region	age	2020
0	01 Stockholm county	16 years	4.6
1	01 Stockholm county	17 years	8.8
2	01 Stockholm county	18 years	17.8
3	01 Stockholm county	19 years	52.5
4	01 Stockholm county	20 years	112.0
5	01 Stockholm county	21 years	127.9
6	01 Stockholm county	22 years	147.1
7	01 Stockholm county	23 years	167.9
8	01 Stockholm county	24 years	191.8
9	01 Stockholm county	25 years	225.2
10	01 Stockholm county	26 years	257.1

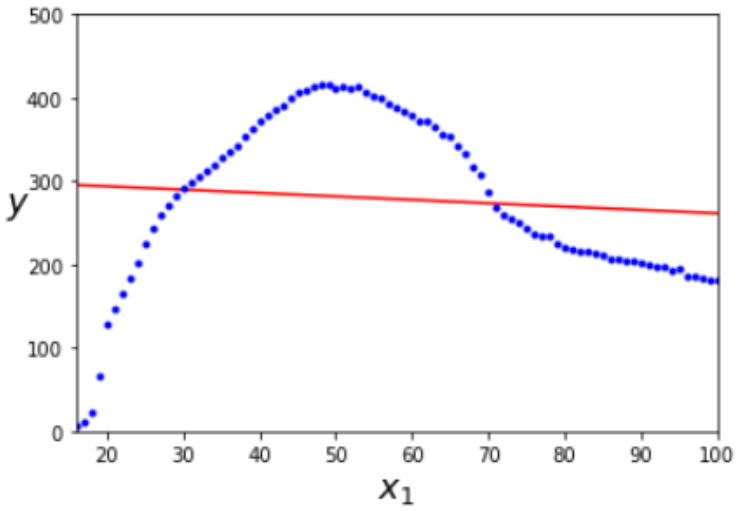
3.1 To execute the linear regression as you did in the previous task, you'll need to group the data by age groups to find the mean across all regions. A recommended approach is to utilize the

pandas group_by function. Inspect the data to see that now you have way more age groups that in the previous task. Additionally, some fields need to be cleaned e.g., the "years" column need to be converted to an integer and the text that is not number has to be removed.

3.2 Perform the linear regression on the provided dataset. Use validation and train data split.

3.3 Create a scatter plot for your data and plot the regression line.

Checkpoint: A graph with age on the X axis should look like the following.



3.4 Predict the values for the test dataset and print both the real values and the predicted ones. How do they look?

3.5 Evaluate the model using MSE as you did with the previous exercise.

3.6 Explain the obtained MSE value and compare it to the previous task.

Task 4: Reflections.

4.1 What differences can you see from the graphs, predicted values and MSE scores from both linear regressions?

4.2 How do you think this analysis can be improved considering the obtained results of task 3.

Task 5: Polynomial regression with hyperparameter tunning

In this lab you performed linear regression to find the relationship between age and income. However, the linear regression result may not be the optimal result as shown by the MSE value.

You will now use Scikit-learn to perform polynomial regression and evaluate if it is a better model.

5.1 Polynomial regression is a special case of linear regression. Perform polynomial regression. With the use of **PolynomialFeatures** from **sklearn.preprocessing** you will be able to increase the number of features the linear regression model is trained.

For additional information on how to perform this task check lecture 8 and page 128 of the ML book.

5.2 You may notice there is a Hyperparameter for this model, which is the polynomial degree, select at least 4 different values for this parameter and perform the polynomial regression for each. Use MSE to evaluate the model using the validation dataset. For example, you can choose a polynomial degree of 2, 3, 5 and 8.

5.3 In this data set you do not use a test set. However, in case you did, which dataset would you use to perform the hyperparameter tunning?

5.3. From your choices, which order of polynomial is best? Discuss and explain.

5.4 Graph the results of the polynomial regression line with the optimal degree found along with the linear regression line.

An example of the desired graph that ***is not done on your dataset*** but shows the type of graph you need to present is the following: