Lab 2: Data aggregation and regression with SCB databases

In this lab you will implement more advanced statistical functions and reuse some of the functions from Lab 1. The data to be used is available in Canvas and has been downloaded from the SCB website (www.scb.se).

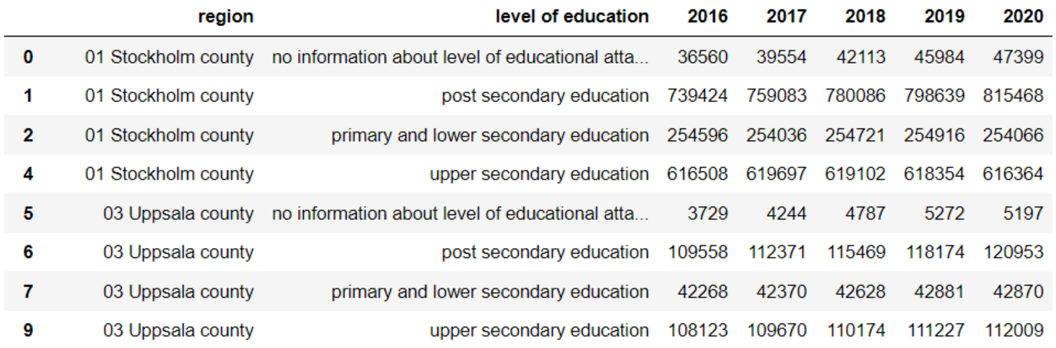
For this lab you will load 2 tables. The first table includes the **total population by level of education by region and year**. The second table includes **mean income by region, age and year.**

The goal of this Lab is to load the data, preprocess it and apply aggregation functions. Finally, you will create a function to perform a multiple linear regressions on the 2nd dataset and compare results.

Task 1: Load and inspect the data

Create a function to load data from a csv file based on the information provided in the lectures. Using that function, load the file **pop_year_trim.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). The output should look something like this:

Tip: Save the loaded rows in a Pandas Dataframe to help with the analysis.

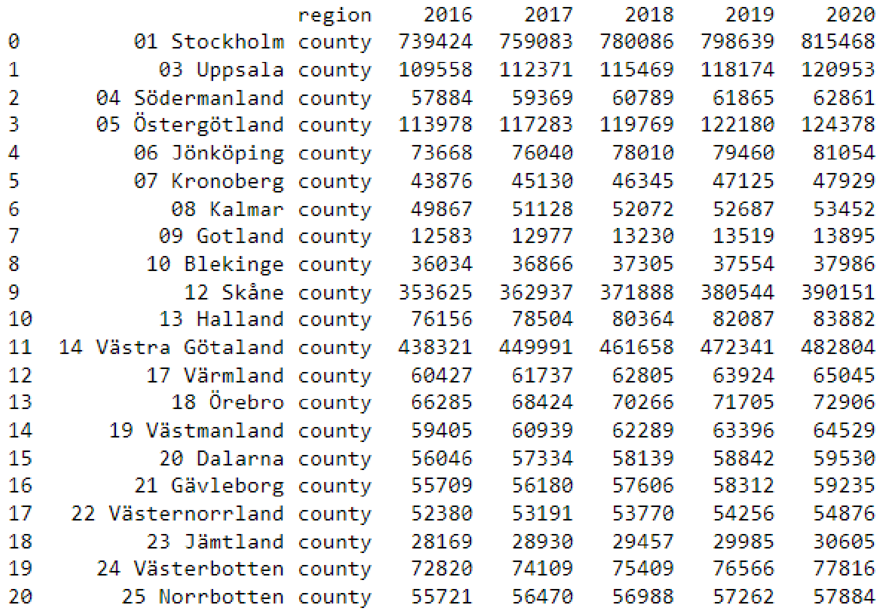


	region	level of education	2016	2017	2018	2019	2020
0	01 Stockholm county	no information about level of educational atta...	36560	39554	42113	45984	47399
1	01 Stockholm county	post secondary education	739424	759083	780086	798639	815468
2	01 Stockholm county	primary and lower secondary education	254596	254036	254721	254916	254066
4	01 Stockholm county	upper secondary education	616508	619697	619102	618354	616364
5	03 Uppsala county	no information about level of educational atta...	3729	4244	4787	5272	5197
6	03 Uppsala county	post secondary education	109558	112371	115469	118174	120953
7	03 Uppsala county	primary and lower secondary education	42268	42370	42628	42881	42870
9	03 Uppsala county	upper secondary education	108123	109670	110174	111227	112009

Task 2: Data aggregation:

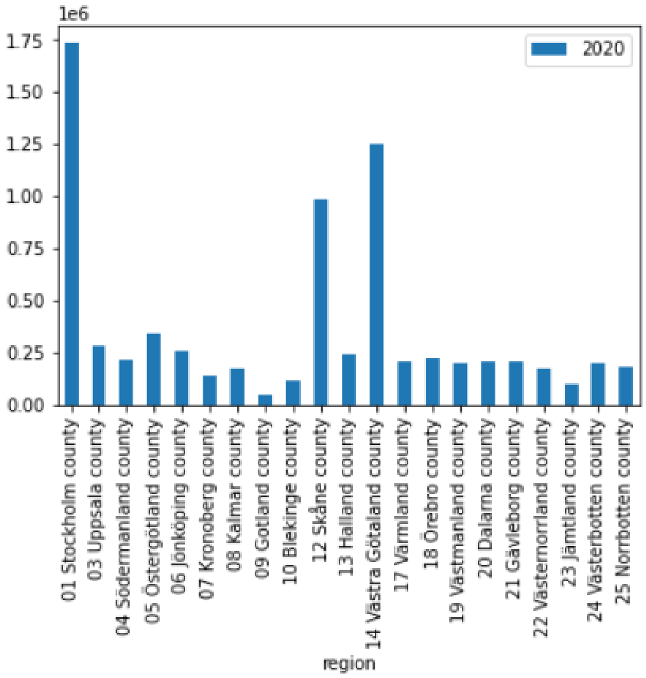
Now when you have loaded the data successfully you will re-use some of the functions from Lab 1 to perform data aggregation.

- Calculate the mean population with post-secondary education in the Norrbotten region in the last 5 years. Do this by filtering all the values in the education level to only be those of post-secondary education.
- The table produced by this task should look similar to this one:



```
                     region    2016    2017    2018    2019    2020
0          01 Stockholm county  739424  759083  780086  798639  815468
1            03 Uppsala county  109558  112371  115469  118174  120953
2       04 Södermanland county   57884   59369   60789   61865   62861
3       05 Östergötland county  113978  117283  119769  122180  124378
4           06 Jönköping county   73668   76040   78010   79460   81054
5           07 Kronoberg county   43876   45130   46345   47125   47929
6              08 Kalmar county   49867   51128   52072   52687   53452
7             09 Gotland county   12583   12977   13230   13519   13895
8            10 Blekinge county   36034   36866   37305   37554   37986
9               12 Skåne county  353625  362937  371888  380544  390151
10            13 Halland county   76156   78504   80364   82087   83882
11     14 Västra Götaland county  438321  449991  461658  472341  482804
12           17 Värmland county   60427   61737   62805   63924   65045
13             18 Örebro county   66285   68424   70266   71705   72906
14        19 Västmanland county   59405   60939   62289   63396   64529
15            20 Dalarna county   56046   57334   58139   58842   59530
16           21 Gävleborg county   55709   56180   57606   58312   59235
17      22 Västernorrland county   52380   53191   53770   54256   54876
18            23 Jämtland county   28169   28930   29457   29985   30605
19        24 Västerbotten county   72820   74109   75409   76566   77816
20         25 Norrbotten county   55721   56470   56988   57262   57884
```

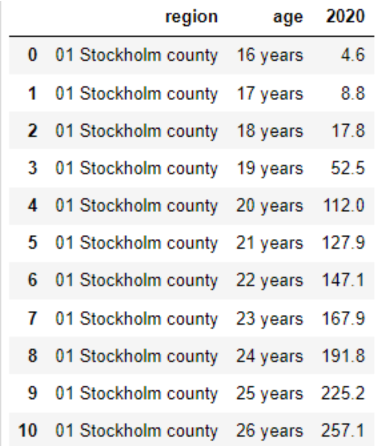
-
    - Using that value calculate the standard deviation as well. What does this value mean?
- Calculate the mean population for each region in Sweden in the last 5 years.
- Tip: To perform this task you will need to create a function that groups and sums all the levels of education per region, check how pandas group_by function works for inspiration.
  The average population in Norrbotten should be around 182,250 people. This value may differ from reality as this data only considers population between 16-75 years
- Finally let's visualize the data, for 2020 create a histogram with the regions and their populations.
- The histogram for all regions in 2020 should look like the following graph.



Task 3: Load a second data set

With the load function created, use it to load the second table **income_data.csv** that is available in Canvas. This dataset contains information regarding the average income by age.

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

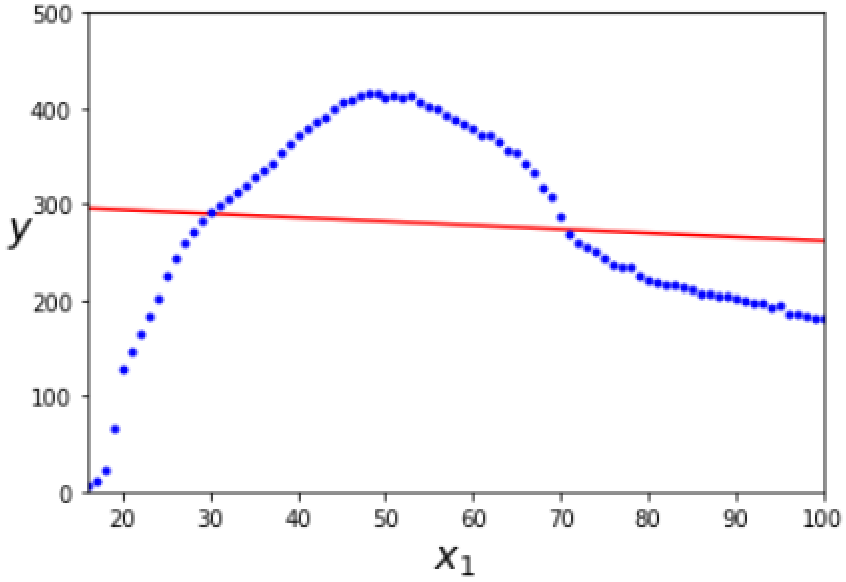
Discussion: Discuss with your lab partner. When is it more appropriate to use the debugger for inspecting variable values, and when would you prefer to use a print function? Identify:

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

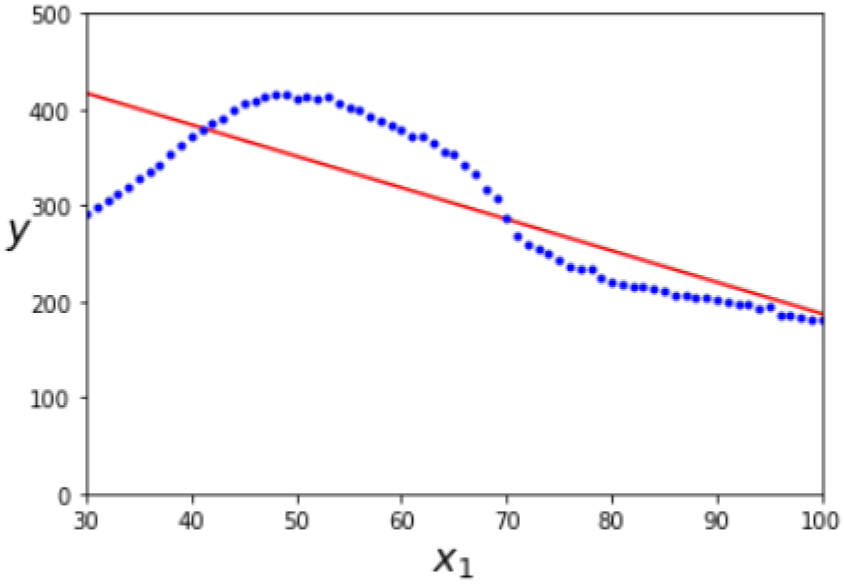
Task 4: Linear regression

For this task you will perform a linear regression on the average yearly income and age data of the year 2020. Based on the example shown in Lecture 8 and chapter 4 of the ML course book.

- Perform the linear regression on the provided dataset
- You will need to group the data once again and provide the mean income per region based on every age value.
- 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.



-
- What are the predicted y values for the following population points (35, 80)
- Evaluate the model using MSE, implement the function that evaluates each of the provided data points against the predicted value (do not use sklearn MSE functions). (Often while creating regression models, we would split our data into test and train samples but for the sake of this exercise we will evaluate against the complete data)
- Explain the obtained MSE value and what does it mean.
- Now take only into consideration the values from people above 30 years and perform the linear regression again along with the graph.
- Checkpoint: A graph for people above 30 years should look like the following.

- 

-
- Calculate once again the predicted y values for the following population points (35, 80)
- Calculate once again the MSE for this new regression and explain the obtained value.
- What differences can you see from the graphs, predicted values and MSE scores from both linear regressions?
- How do you think this analysis can be improved considered the obtained results.

You have now completed Lab 2.