

# Assignment 1

- Due Monday by 8:00
- Points 0
- Available after 29 Oct at 15:00

## Purpose




The overall purpose of this assignment is to get familiar with the use of APIs, data wrangling, clustering and visualization. These components are important basic steps in any data exploration workflow.

This assignment shares similarities with exercise 2, so having completed this exercise is a good first step.

---

## Description

This assignment should be solved in groups of 3-4 students.

The goal of the assignment is to visualize how energy production in Sweden depends on the weather. To achieve this goal, your main task is to collect and clean weather data that is provided in an API by SMHI, which you can find [here](https://opendata.smhi.se/apidocs/metobs/index.html).  <https://opendata.smhi.se/apidocs/metobs/index.html>. To provide you with a small example of how an API can be called, a small workflow that collects data about the energy production from SCB is provided here: [SCB\\_Power\\_Production.knwf](https://ju.instructure.com/courses/12258/files/1841983?wrap=1) (<https://ju.instructure.com/courses/12258/files/1841983?wrap=1>)  ([https://ju.instructure.com/courses/12258/files/1841983/download?download\\_frd=1](https://ju.instructure.com/courses/12258/files/1841983/download?download_frd=1)). You will need to extend/change the API request to SCB slightly, you can read more about the SCB API [here](https://api.scb.se/)  <https://api.scb.se/>. To solve the assignment, you will have to extend this workflow by performing the following steps:

1. Go through and analyse the provided workflow. Add a relevant comment to each of the nodes.
2. Add one (or several) nodes in the workflow that collect data from the SMHI API. You should collect data concerning at least 3 different variables and it should come from at least 4 different weather stations. Motivate both the selection of the variables as well as the weather stations.

**- You should collect data from at least 12 months.**

3. Join the values from the different stations and create one table that is filled with the data that you fetched from the SMHI API.
4. Aggregate the data at a monthly level. When you do the aggregation, you should calculate more statistics than just the mean.
5. Join the weather data together with the data that is provided for the energy production from SCB.
6. Create **at least 3** interesting plots that show what show how the weather and the energy production are linked (**at least one plot for each variable**). As an example, are wind speed and wind energy production linked positively, negatively or not at all?
  - **Use at least 2 different kinds of plots or visualization techniques!**
  - One good suggestion would be to use line plots, but other kinds of plots and visualizations might be better (or worse). Pick something that you think describes the patterns in a good way!
7. Use the hierarchical clustering algorithm to find clusters in the data (where each data point is a month). Use the average value of each of the clusters to describe what a typical sample in that cluster represents.
8. Visualize the clusters.

**Things to note:**

You can use [this page](https://opendata.smhi.se/apidocs/metobs/demo.html)  (<https://opendata.smhi.se/apidocs/metobs/demo.html>) to navigate through the different parameters, stations and timings.

Not all stations have information about all parameters! If you get a "404" error or similar, it could be due to the station not having information about the latest day/month, or that it does not collect information about a parameter that you have selected.

The SCB API might have a problem if the API-URL is set to English (.../doris/en/ssd/...). To combat this, change the URL to ".../doris/sv/ssd/...".

---

## Submission Details

When submitting your solution, please use the Assignment group you belong to. To pass this assignment you should write a short text about your chosen approach. **This text must be targeted to a general audience**, i.e., it should not assume that the reader is proficient in artificial intelligence, data mining, product development etc. The exact format of your text is free, you do not need to mimic a scientific paper.

## The Report Structure

Your text must include at least these three parts:

- An introduction with a motivation.
- A description of your method, i.e., your KNIME workflow and what you have done. You should complement this part with an image of your KNIME workflow.
- A discussion where you present and reason about your results.

## The Report Text

Your text must follow these guidelines:

- The text must be in English.
- Each part of the report should consist of at least 200 words.
- The report cannot be more than 1000 words.
- The report must contain at least 4 figures and at most 10 figures
  - One of the figures should depict your workflow in KNIME.
  - The other 3 to 9 figures depict the data, statistics or other visualizations.

## Passing

This assignment is graded using only Pass or Fail. To pass, your text must be well-structured, and well-written and it should of course treat the chosen subject satisfactorily. To pass the assignments in the course (to get the points in Ladok), you must pass this assignment, along with the other remaining assignments in the course. See the course PM for further details on re-examination.

Submit your text in Canvas before the deadline as communicated through Canvas.

Questions about this assignment should be sent to: [Helena.lofstrom@ju.se](mailto:Helena.lofstrom@ju.se) (<mailto:Helena.lofstrom@ju.se>) or [Cecilia.sonstrod@ju.se](mailto:Cecilia.sonstrod@ju.se) (<mailto:cecilia.sonstrod@ju.se>)