Vesa Ollikainen

Last updated: 17.1.2018

Assignment: Data manipulation

Learning goals

In this assignment, you:

1. install the machine learning software required on the course.

2. acquire the basic data manipulation skills in Python. This includes data importing, selection and reformatting as well as calculating the summary statistics and the correlation matrix.

Assignment

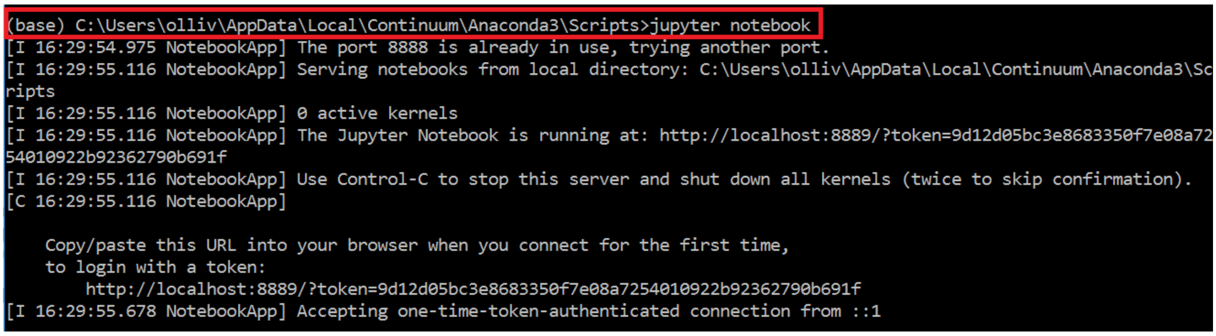
1. For software installations, follow Steps 1 to 3 in Sakari Lukkarinen's installation instructions in the Software folder in the Documents tab of the course's Oma workspace.

   In the end of Step 3, also install **`pandas`** and **`scikit-learn`**:

   ```
   conda install pandas
   conda install scikit-learn
   conda install python-graphviz
   ```

   After the installation, the Jupyter Notebook server can be started and a connection established by selecting Jupyter Notebook from the Anaconda3 program group in the Windows menu.

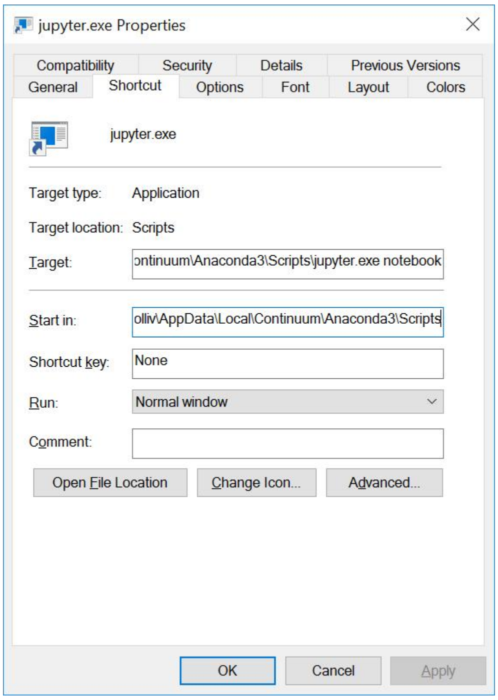
   Should this fail, you can manually start the Jupyter notebook server and connect to it by opening the Anaconda prompt, changing the default directory to the one where **`jupyter.exe`** is located and giving the following command:



```
(base) C:\Users\olliv\AppData\Local\Continuum\Anaconda3\Scripts>jupyter notebook
[I 16:29:54.975 NotebookApp] The port 8888 is already in use, trying another port.
[I 16:29:55.116 NotebookApp] Serving notebooks from local directory: C:\Users\olliv\AppData\Local\Continuum\Anaconda3\Sc
ripts
[I 16:29:55.116 NotebookApp] 0 active kernels
[I 16:29:55.116 NotebookApp] The Jupyter Notebook is running at: http://localhost:8889/?token=9d12d05bc3e8683350f7e08a72
54010922b92362790b691f
[I 16:29:55.116 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 16:29:55.116 NotebookApp]

    Copy/paste this URL into your browser when you connect for the first time,
    to login with a token:
        http://localhost:8889/?token=9d12d05bc3e8683350f7e08a7254010922b92362790b691f
[I 16:29:55.678 NotebookApp] Accepting one-time-token-authenticated connection from ::1
```

   Now, you can right-click on the Windows desktop and generate a shortcut the opens the Jupyter notebook server and connects to it.



Target type: Application

Target location: Scripts

Target: ontinuum\Anaconda3\Scripts\jupyter.exe notebook

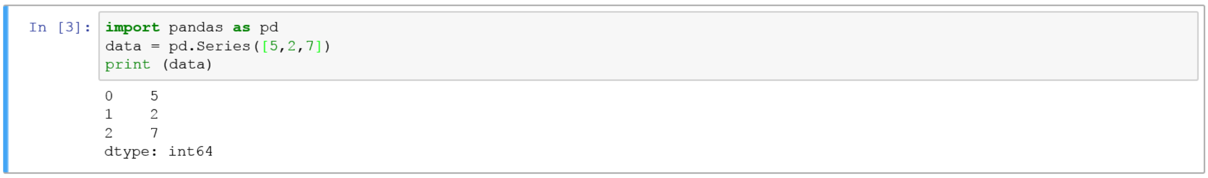
Start in: olliv\AppData\Local\Continuum\Anaconda3\Scripts

Shortcut key: None

Run: Normal window

Comment:

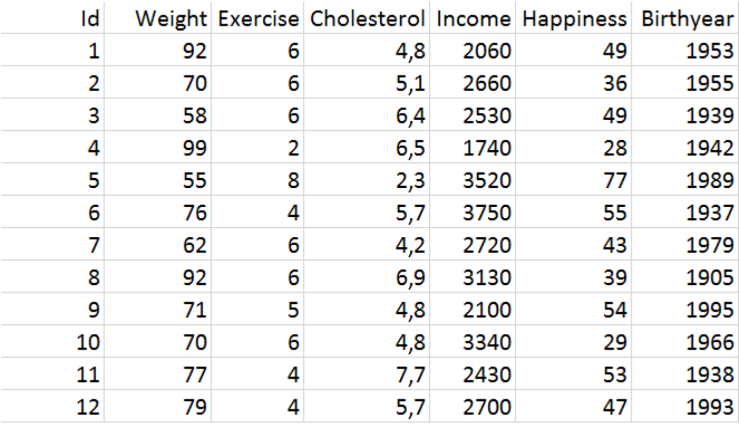
To verify the installation, create a new Jupyter3 workbook and run the following code snippet:



```
In [3]: import pandas as pd
        data = pd.Series([5,2,7])
        print (data)

        0    5
        1    2
        2    7
        dtype: int64
```

2. Using Python in Jupyter Notebook, construct a **pandas** data frame that contains the following observations:



Id	Weight	Exercise	Cholesterol	Income	Happiness	Birthyear
1	92	6	4,8	2060	49	1953
2	70	6	5,1	2660	36	1955
3	58	6	6,4	2530	49	1939
4	99	2	6,5	1740	28	1942
5	55	8	2,3	3520	77	1989
6	76	4	5,7	3750	55	1937
7	62	6	4,2	2720	43	1979
8	92	6	6,9	3130	39	1905
9	71	5	4,8	2100	54	1995
10	70	6	4,8	3340	29	1966
11	77	4	7,7	2430	53	1938
12	79	4	5,7	2700	47	1993

Next, expand your program to:

a) compute the basic statistics (count, mean, quartiles, etc.) summary for all variables.

b) iterate through the rows in the original data frame and produce the output below. For each individual, you should indicate whether his/her income is above or below the average computed from the data. Take the average programmatically from the computed basic statistics summary; it must not be hard-coded in the program.

```
Person 1: below average income.
Person 2: below average income.
Person 3: below average income.
Person 4: below average income.
Person 5: above average income.
…
```

3. Load the Chronic kidney disease data set **kd.csv** from the Data folder in the course's Oma workspace (Data source: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease). Your goal is to find out the minimum, maximum, and mean values and pairwise correlation coefficients of all numerical variables for the affected individuals in the data set.

   Tips:

   See the accompanying text document for an interpretation of the variables.

   The loading should initially fail, so find the cause and correct it. Check how the missing data is encoded and make sure the missing values are recorded as such.

   Filter the data set to include only the affected patients.

   For the remaining subset, print the basic statistics.

   Then, calculate the pairwise correlation coefficients (correlation matrix) between each pair of numerical variables. The correlation coefficients vary between minus and plus unity and show how interrelated the variable values are (-1 = strong negative correlation, 0 = uncorrelated variables; not related to each other; 1 = strong positive correlation).

   Visualize the correlation matrices with matplotlib. See, e.g. https://machinelearningmastery.com/visualize-machine-learning-data-python-pandas/

Deliverables

Your deliverable should include both the Python codes and the results.

The simplest way to produce them is to use Jupyter workbook and select **File** / **Download as** / **Notebook**.