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. First, install Anaconda: https://www.anaconda.com/products/individual.

   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 those individuals that are affected by CKD.

   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 patients affected by the disease.

   For the resulting 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 seaborn. See https://seaborn.pydata.org/

Deliverables

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

Please submit the answer as a downloaded HTML document. In Jupyter workbook, select **File / Download as / HTML (.html)**.