Exercise Sheet 1

1. Exercise: Handling of Tabular Data

(a) Use the `pandas` library to represent the following tabular data.

Patient ID	Age	Height [cm]	Weight [kg]
000345	45	167	67
000124	60	181	78
001758	22	158	57
000994	38	185	90
001233	36	164	72
001145	77	190	75
000222	65	180	110

(b) Add a new patient entry to the generated table.

Patient ID	Age	Height [cm]	Weight [kg]
001122	51	177	81

(c) Implement the function `func_norm(v)` that normalizes the values of a feature vector $\boldsymbol{v}$ so that its range is between 0 and 1, by performing

$$v_m^{(\mathrm{norm})} = \frac{v_m - v_{\min}}{v_{\max} - v_{\min}} \tag{1}$$

for each entry m of the feature vector.

(d) Why is normalization important in Machine Learning?

(e) Normalize each the feature of the table given in (a) and add them as new columns to the table.

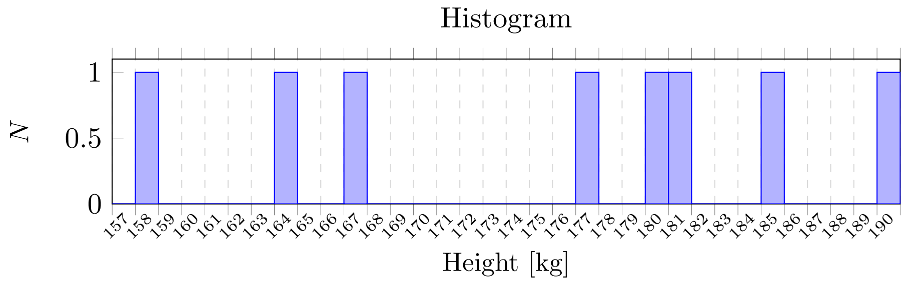
(f) Save the extended table to a file named `patients.csv`.

(g) Filter out the patients that are above the average age of all patients and save the table to a file named `youngest_patients.csv`.

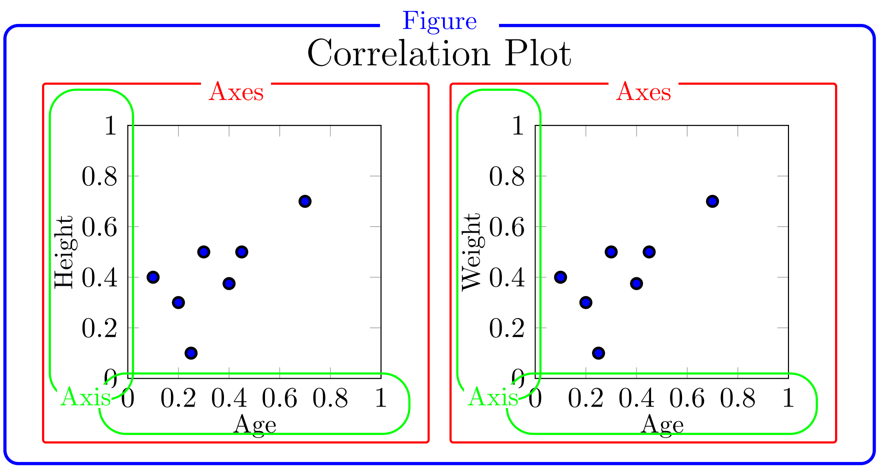
2. Exercise: Plots using `matplotlib`

(a) Load the stored `patients.csv` file from Exercise 1.

(b) Generate a histogram showing the height of all participants as presented below, by setting the number of the shown bars to match the total number of patients.



(c) Generate the following plot using the normalized features and the `matplotlib` library. Note, that it is not necessary to plot the colored rectangles. They are only intended to help you generate the plot.



(d) Compute the Pearson correlation coefficients for the feature combinations:

(i) Age ($\boldsymbol{v}_n$) and Height ($\boldsymbol{v}_l$)

(ii) Age ($\boldsymbol{v}_n$) and Weight ($\boldsymbol{v}_l$)

using

$$r_{nl} = \frac{\sum_{m=1}^{M}(v_{m,n} - \hat{\mu}_n)(v_{m,l} - \hat{\mu}_l)}{\sqrt{\sum_{m=1}^{M}(v_{m,n} - \hat{\mu}_n)^2 \sum_{m=1}^{M}(v_{m,l} - \hat{\mu}_l)^2}}, \tag{2}$$

with

$$\hat{\mu}_n = \frac{1}{M}\sum_{m=1}^{M} v_{m,v}, \qquad \hat{\mu}_l = \frac{1}{M}\sum_{m=1}^{M} v_{m,l}, \tag{3}$$

and interpret your results. What do the computed Pearson correlation coefficients say about how the features correlate with each other?

3. Exercise: Similarity Metrics

(a) The following feature vectors are given

$$\boldsymbol{v}_1 = \begin{bmatrix} 0.2 \\ 0.1 \\ 0.4 \\ -0.4 \end{bmatrix}, \qquad \boldsymbol{v}_2 = \begin{bmatrix} -0.1 \\ -0.1 \\ 0.8 \\ 0.5 \end{bmatrix}. \tag{4}$$

(b) Implement the functions

(i) `func_L1norm(v1, v2)`,
computing the L1-norm $s(\boldsymbol{v}_1, \boldsymbol{v}_2) = \sum_{m=1}^{M} |v_{m,1} - v_{m,2}|$.

(ii) `func_L2norm(v1, v2)`,
computing the L2-norm $s(\boldsymbol{v}_1, \boldsymbol{v}_2) = \sqrt{\sum_{m=1}^{M} (v_{m,1} - v_{m,2})^2}$.

(iii) `func_cosine_sim(v1, v2)`,
computing the Cosine-similarity $s(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1^{\mathrm{T}} \boldsymbol{v}_2}{||\boldsymbol{v}_1||_2 \cdot ||\boldsymbol{v}_2||_2}$.

(c) Compute the similarity metrics for the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ based on the implemented functions in `(b)` and print their results.