Exercise 2: Data understanding and visualization

In this exercise we use the Iris dataset. The dataset consists of 50 samples from each of three species of Iris flowers (*Iris setosa, Iris virginica* and *Iris versicolor*). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

Matplotlib tutorials can be found here: https://matplotlib.org/

1. Load the Iris dataset from http://archive.ics.uci.edu/ml/datasets/Iris

2. Calculate the mean, median and standard deviation for all attributes of *Iris setosa.*

3. Select one attribute for species *Iris setosa*. Plot four histograms of the attribute using the methods intoduced in the lectures (Sturges, Scott, Square root, and Freedman-Diaconis) to determine the number of bins. Compare the results.

Next we would like to see if the distributions of the values for different flower species vary in some way. Plot histograms of your attribute for *Iris virginica* and *Iris versicolor.* Do these histograms give any indication about the feasibility or infeasibility of classifying the flower species?

4. Produce boxplot plots of attributes. Can you see any outliers?

5. Calculate Pearson's correlation tables for the attributes. These numbers may give us some indication about simple (linear) relationships between features. Calculate also Spearman's rho and Kendall's tau values. What is the purpose of these values?

o What features correlate linearly with each other?

o Does this exclude any other sort of correlation?

Produce scatter plots of interesting features.

6. Principal component analysis (PCA) with and without z-score normalization: project the data to the first two principal components. Visualize the result as a scatter plot. What is the proportion of variance explained in projections? What can be observed?

Bonus question (1p):

Familiarize yourself with some other dimensionality reduction method than PCA. Describe the basic principles of the method. Project the data to two dimensions using this method, present the scatter plot and compare the results to PCA.