Exercises 6

1. Iris data set

- Download the file `iris.data` from the Iris data set available at https://archive.ics.uci.edu/dataset/53/iris.
  - In R, you can simply use `data(iris)`.
- Plot `Sepal length` (on X axis) vs. `Sepal width` (on Y axis).
- Create the scatter plots of all feature pairs (4 x 4).
  - (BONUS) Try the `pairplot` function available in the `seaborn` library (https://seaborn.pydata.org/).
- What can you see in the plots?

2. PCA plot

- Perform PCA for the data.
- Visualize the data on the principal components 1 - 2.
- Use colors to highlight each observation by the species.
- Are the species distinct?

3. PCA loadings

- Retrieve the loadings of the PCA components.
- What are the most important features in the component 1?
- What are the most important features in the component 2?

4. PCA scores

- Calculate correlations between the original features (axes).
- Calculate correlations between the PCA features (axes).
- Which features are the most correlated in each case?

5. Clustering

- Cluster the observations into three clusters.
- Visualize the clusters on top of PCA.
- Count how many times each species is assigned to each cluster.
- Does the unsupervised clustering correspond to the known species?