k Clustering
Exercises

Exercise 1. (and k-Means Clustering)
By clustering a dataset, we try to find groups of instances that the set is composed of. Often, these groups can reveal insights on the processes of how the data was created – processes that are hidden or unclear beforehand. Imagine the following (toy example) situation: In the dataset in the table below, instances are given with two features x and y. The data is created by two processes (1 and 2). In a real setting, these processes would be unknown and the clustering algorithm is supposed to find appropriate groups. Consider the following dataset (at first without the actual process information).

x	3	3	4	4	5	6	7	7	8	9	1	2	2	3	4	5	5	6	7	7
y	1	2	2	3	3	4	4	6	5	7	3	4	5	6	6	7	8	8	8	9
process	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2

1. Based only on the features, how many clusters would you expect?
2. Knowing the generating processes, how many clusters do you expect?
3. What could be possible obstacles that might lead to clusterings that do not fit the processes?
4. Cluster the dataset using k-means with $k = 2, 3, 4$.
5. When clustering the data with k-means, which problems occur?
6. Following your insight on k-means, what do expect regarding conclusions about the two processes?
7. Based on the observations of k-means with $k = 2$, what would we recommend?

Exercise 2. (k-Means on IRIS)
In this task, we use the IRIS dataset for clustering. For that purpose load the data from the file (`iris.csv`). IRIS is a dataset where the instances are specimen of IRIS (plants). The features are petal and sepal lengths and widths.

1. Create a brief initial data analysis.
2. Visualize the data.
3. Scale the data. Because we are going to use feature selection by variance below, use a scaler that does not equalize the variance.
4. Run k-means clustering on the IRIS dataset using k from 2 through 10.

5. Inspect the resulting structures using the silhouette coefficient. Plot your result.

6. What is your conclusion regarding the choice of k?

7. When working with multidimensional data, sometimes it omitting dimensions can be helpful. For example, when we expect them to be irrelevant for the clustering that we want to create. Often, we do not know beforehand what will be relevant. To reduct the number of features, then automatic feature selection methods can be used. A simple feature selection is selecting features with high variance and discarding low-variance features. Compute the variance per feature (on the scaled data). Set a threshold such that only two features remain in the dataset. Use the `VarianceThreshold` class from `sklearn` to transform the dataset accordingly. Which two features stay in the dataset, which are discarded?

8. Compare the silhouette coefficients for different k when running k-means on all features vs. only on the selected subset of features.

9. Run k-means with the most successful combination (features and k) and plot the data colored by cluster together with the cluster centers. Also create a silhouette plot. Interpret the results.

10. Repeat the same task but set $k = 3$.

11. Compare the results of $k = 2$ and $k = 3$ in a confusion matrix.

12. Cluster Descriptions
    a) Group the data by cluster (resulting from the strongest combination of k and feature selection above). Compare the features' min, max, and mean per cluster.
    b) Use box plots and violin plots to compare the feature ranges in the clusters per feature.

Hints: Take a look at the documentation of `sklearn.cluster.KMeans`. For the silhouette plots take a look at the Python module yellowbrick and specifically at the class `yellowbrick.cluster.SilhouetteVisualizer`. For the confusion matrix, first read up on what they are usually used for. Then apply the implementation `ConfusionMatrixDisplay` from `sklearn.metrics`.