FACHHOCHSCHULE KIEL
University of Applied Sciences

Mathematics and Multivariate Statistics **Prof. Dr. Stephan Doerfel**

MADS-MMS SoSe 2025

k Clustering
Exercises

📊 Please join the ↗ Particify Room.

Exercise 1. (📊, 📝 + 💻 k-Means Clustering)
By clustering a dataset, we try to find groups of instances that the set is composed of. Imagine the following (toy example) situation: In the dataset in the table below, instances are given with two features x and y.

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

1. 📊 Based only on the features, how many clusters would you intuitively expect?

2. Cluster the dataset using k-means with $k = 2, 3, 4$.

3. 📊 Do the resulting clusterings fit to those intuitively expected?

4. 📊 Which properties of k-means are responsible for the result?

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 and according preprocessing. Among others:
   a) How many instances are in the dataset?
   b) How many columns are in the raw data?
   c) How many columns contain meaningful features?

2. 📊 Statistically summarize and 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 using a random state of 1 in every turn.

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 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 reduce the number of features, automatic feature selection methods are an option. A simple example 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. Discuss in which way the resulting clusterings (with and without feature selection) are comparable!

9. Run k-means with the most successful k on the selected features 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 cross table.

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`.