Juniorprofessur für Informatik – Beck

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Exercise Sheet 4

Information and Software Visualization (SoSe 2019)

Deadline: Wednesday, 15.05.2019, 14:00

Everyone is required to submit an individual solution. Group discussions are possible and encouraged, but each individual solution must be clearly distinguishable from others. The submission takes place via Moodle—upload your solution as a single zip file containing two R script files (one for each task) and a PDF file with short explanations on your choice of visualization (for task 1). This exercise makes use of the built-in datasets of R, "mtcars" and "iris". They are shipped with the standard download of R.

This exercise sheet includes 2 task with a total of 10 Points (10 Point $\widehat{=}$ 100%).

Task 1 *[Points: 6]*

The `iris` dataset includes measurements (in centimeters) of the variables sepal length, sepal width, petal length, and petal width for 150 flowers from each of the 3 species of Iris. The species are setosa, versicolor, and virginica. Choose an appropriate visualization (**one visualization per subtask**), and produce it in **R** for the following subtasks. Also, provide a brief justification of your choice in each case.

(a) *(2 Points)* Which of the Iris species have the least variance in terms of *petal length*? Provide a comparison of the variance across all species.

(b) *(2 Points)* Based on the values of *sepal length* and *sepal width*, are there any outliers in any of the Iris species?

(c) *(2 Points)* Provide an overview of pairwise correlations among the variables *sepal length*, *sepal width*, *petal length*, and *petal width*. Different species must be distinguishable too.

Task 2 *[Points: 4]*

The `mtcars` dataset contains values for 11 different properties of various car models. Cluster these car models into 5 groups using k-means clustering algorithm. Now, use the parallel coordinates plot to visualize the data and mark the clusters with distinct colors.