COMPUTER PRACTICE 2

CLUSTER ANALYSIS

OBJECTIVES

- Obtain practical skills of cluster analysis using Python libraries and self-developed scripts

GUIDELINES (CONSISTS FROM 11 POINTS)

1. Recall theoretical concepts of clustering, provided in topics Topic 5 and Topic 6.
2. Create Python Notebook for scripting and analysis running. You can use Jupyter Notebook or Google Colab as work environment.
3. Load the data set provided for the Computer Practice 2 in the course and read through its description.
4. Conduct hierarchical cluster analysis using Euclidean distance as affinity measure and applying the following linkage methods:
    a. Single linkage
    b. Complete linkage
    c. Average linkage
    d. Ward

    Produce the dendrograms for each of linkage methods, describe each of them in terms of number of clusters and their quality.
5. Standardize data and repeat the task N4 again. Compare the results and make final decision regarding the number of clusters.
6. Apply k-means method for selected data sample with information about clusters membership. Analyse:
    a. Number of observations in clusters
    b. Variable values of cluster centres
    c. ANOVA of cluster means
7. Repeat the same steps from the task N6 on standardized data.

    Make final decision about quality of cluster analysis based on k-means, including analysis the size of clusters; significance of variables for clustering based on ANOVA etc.
8. Provide an interpretation of the resulting clusters.
9. Save the cluster membership for each observation from data sample.
10. Create the report providing all generated results (dendrograms and numerical results), dendrograms description and interpretation, comparing analysis of different algorithms' results, numerical results description and interpretation, your intermediate and final conclusions. Your report should represent the logic of analysis, understanding of obtained results and ability to plan research and make decision regarding the next steps of analysis on the base of intermediate results.
11. Upload your report (.pdf) and Python Notebook (.ipynb) with all your scripts used for analysis and generated results to the learning environment.