Homework 8

PCA and Clustering

The aim of the homework is to use k-means clustering and hierarchical clustering methods with and without PCA for detecting interesting clusters in data. Use the dataset "hw8_study_book_nr.csv", where `study_book_nr` is your study book number (ask from me if you do not know it).

Exercises

1) Produce scatterplot of your data set with respect to the pair of variables which show most clearly some different groups of observations and also a scatterplot of points with respect to the first two principal components. Based on those plots, how many clusters there seems to be in the data?
2) Determine the best number of clusters for k-means method in the case of unscaled data by using elbow rule. Produce corresponding plot of clusters (showing cluster membership by color) using the same variables you chose in the first exercise.
3) Find the best number of clusters for k-means method in the case of scaled data and average silhouette method. Produce corresponding plot of clusters.
4) Determine the best number of clusters (based on largest gap between appearance of new clusters) for hierarchical clustering of the data when Minkowski's distance with $p = 4$ is used (see the help of `dist` function) is used. Produce the corresponding plot of clusters
5) Find the best number of clusters by using average silhouette method for the data when correlation based distance is used in the hierarchical clustering. Produce the corresponding plot of clusters.

Please submit the .R/.RMD code for your solution together with .html output showing the code, explanations and results. **This time submissions with missing components (no R/Rmd code or no .html/.pdf file) receive 0 points. Please make sure that your comments/decisions agree with the output of your code!**

I remind you that you are expected to solve the homework problems by yourself, sharing solutions (and copying parts of other student's solutions) is not allowed.