Homework #1: Exploratory Analysis and Clustering

Following is the description of the first homework. **Please submit the homework report by 8.00 on Wednesday, December 16, 2020.** Carefully follow the submission instructions below. Please send any questions related to the homework through Slack.

Find any data from some publication, or retype it from any other data source, construct it on your own, or find it in your lab. The data set could be small, but it should contain at least 30 but no more than 200 samples (data instances) and at least four but no more than 20 features. It would be best if the data set is real, say, is relevant to your research or interest, comes from some publication, web page, or from some problem that you have worked on before. **Do not use any data sets that come with Orange or similar programs, or any standard machine learning data sets from web sites such as Kaggle or UCI ML Repository!**

Use any of the data exploration techniques that we have learned about in our first lecture, including those that we have not mentioned but that are exemplified in the lecture notes. Find any visualization that shows anything interesting in your data. Try not to use (only) scatterplot, but instead use box plot, visualization of distribution, sieve diagram, or some other visualization. Was your finding expected? Was it reported anywhere? Or is this something entirely new?

In the second part of the homework, perform hierarchical clustering. Your report should include a dendrogram and explanation of the clusters, say, how one clustering branch is different from another. Which features characterize the clustering best? Use some visualization (say, a box plot) in combination with the dendrogram to explain the results.

Submit the homework as a short report in PDF. The report should include a title of the homework ("Exploratory analysis and simple Orange workflows"), your name and email, and the following sections:

> **Introduction**, a three-line paragraph overview of your report;
>
> **Material** (one paragraph, five lines max) that reports on where and how did you get the data, what is its size (number of samples and features), and what and what type are the features;
>
> **Methods** (one paragraph, five lines max) on the methods you have used, preferably include a workflow;
>
> **Results and Discussion** (one paragraph, five lines max), show you results and visualizations. Explain and discuss the figures.

The report should not exceed two pages (this limit is strict!), use 11 pt Arial or similar sans-serif font. All figures should be captioned and referred to in the text. Make sure the figures and the text on them are readable. Include only the visualizations, not the screenshots that include images of windows, programs or, desktops. Do not skew the images by changing their aspect ratio.

Submit the report as a PDF document (do not send us Word documents!); name the PDF file as lastname-firstname-1.pdf (like: zupan-blaz-1.pdf) where lastname is your last name and firstname your first name. Email the report to bzupan@gmail.com with subject "DM-HW1" (omit the quotes, copy (!) this title to the subject line of the email to make sure there is a precise match).