# Project description for report 1

Objective: The objective of this report is to apply the methods you have learned in the first section of the course,*"Data: Feature extraction, and visualization"* on your own data set to get a basic understanding of your data prior to the further analysis (project report 2 and 3).

Material: You can use the 02450Toolbox on Campusnet to see how the various methods learned in the course are used in Matlab, R or Python. In particular, you should review exercise 1 to 4 in order to see how the various tasks can be carried out.

---

**Handin Checklist**

- Specify **names *and* study numbers** of each group member on the front page
- According to the DTU regulations, each students contribution to the report must be clearly specified. Therefore, for each section, specify which student was responsible for it (use a list or table). **A report must contain this documentation to be accepted**. The responsibility assignment must be individualized[1]
- Your handin should consist of a `.pdf` file containing the report, and the code you have used as one (or more) files with the extension `.py`, `.R` or `.m`. The reports are not evaluated based on the quality of the code (comments, etc.), however we ask the code is included to avoid any potential issues of illegal collaboration between groups. Please do not compress or convert these files.
- Reports are evaluated based on how well they address the questions below. Therefore, to get the best evaluation, address all questions
- Use the group handin feature on campusnet. **Do not upload separate reports for each team member as this will lead to duplicate work and unhappy instructors**
- **Deadline for handin is no later than 3 March at 13:00**. Late handins will not be accepted under normal circumstances

[1]For reports made by 3 students: Each section must have a student who is 40% or more responsible. For reports made by 2 students: Each section must have a student who is 60% or more responsible.

**Description**

Understanding the data you are trying to model well is very important. You can apply very sophisticated machine learning methods but if you are not aware of potential issues with the data the further modeling will be difficult. Thus, the aim of this first project is to get a thorough understanding of your data and describe how you expect the data can be used in the later reports.

Report 1 should cover what you have learned in the lectures and exercises of week 1 to 4 covering the section *"Data: Feature extraction, and visualization"* . You should consider yourself as a new employee in a company who has just been given a data set. Your job is to make a useful description of the data set for your co-workers and make some basic plots. In particular, the report ***must*** include the following items and the report will be evaluated based on how it addresses each of the questions asked below and an overall assessment of the report quality. For readability and brevity consider not using one subsection for each item.

1. **A description of your data set.**

   - Explain what your data is about. I.e. what is the overall problem of interest?
   - Provide a reference to where you obtained the data.
   - Summarize previous analysis of the data. (i.e. go through one or two of the original source papers and read what they did to the data and summarize their results).
   - You will be asked to apply (1) Classification, (2) a regression, (3) a clustering, (4) an association mining, and (5) an anomaly detection on your data in the later reports. For now, we want you to consider how this should be done. Therefore:
     Explain, in the context of your problem of interest, what you hope to accomplish/learn from the data using these techniques?.
     Explain which attribute you wish to predict in the regression based on which other attributes? Which class label will you predict based on which other attributes in the classification task?
     If you need to transform the data in order to carry out these tasks, explain roughly how you plan to do this.

   One of these tasks (1)–(5) is likely more relevant than the rest and will be denoted the **main machine learning aim** in the following. The purpose of the following questions, which asks you to describe/visualize the data, is to allow you to reflect on the feasibility of this task.

2. **A detailed explanation of the attributes of the data.**

- Describe if the attributes are discrete/continuous, Nominal/Ordinal/Interval/Ratio,
- Give an account of whether there are data issues (i.e. missing values or corrupted data) and describe them if so.
- Include basic summary statistics of the attributes.

If your data set contains many similar attributes, you may restrict yourself to describing a few representative features (apply common sense).

3. **Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA).**
Touch upon the following subjects, use visualizations when it appears sensible. *Keep in mind the ACCENT principles and Tufte's guidelines when you visualize the data.*

   - Are there issues with outliers in the data,
   - do the attributes appear to be normal distributed,
   - are variables correlated,
   - does the primary machine learning modeling aim appear to be feasible based on your visualizations.

   There are three aspects that needs to be described when you carry out the PCA analysis for the report:

   - The amount of variation explained as a function of the number of PCA components included,
   - the principal directions of the considered PCA components (either find a way to plot them or interpret them in terms of the features),
   - the data projected onto the considered principal components.

   If your attributes have different scales you should include the step where the data is standardizes by the standard deviation prior to the PCA analysis.

4. **A discussion explaining what you have learned about the data.**
Summarize here the most important things you have learned about the data and give also your thoughts on whether your primary machine learning aim appears to be feasible based on your visualization.

## Collaboration

The usual DTU rules for collaboration applies for the reports. The main rule is that if you hand in a report, you must have authored or co-authored the content of the report for this assignment, and if your report contains text you did not write, then it must be with attribution. Notice in particular:

- If you are taking the course again, you are allowed to re-use content from a report that you previously authored or co-authored.
- If you are authoring a report together with a person who has previously taken the course, you cannot re-use that report since you did not originally author it. We recommend that you simply choose another dataset and re-write the text such that the new report can be considered original joint work by both authors.
- You are of course allowed to use the scripts, etc. supplied in this course for the reports.

The report should be 5-10 pages long (and no longer!) including figures and tables and give a precise and coherent introduction to and overview of the dataset you have chosen.

## Transferring/reusing reports from previous semesters

If you are retaking the course, you are allowed to reuse your previous report. You can either have the report transferred in it's entirety, or re-work sections of the report and have it evaluated anew.

To have a report transferred, *do absolutely nothing*. Reports from previous semesters are automatically transferred. Therefore, please do not upload old reports to campusnet as this will lead to duplicate work. As a safeguard, we will contact all students who are missing reports shortly after the exam.

If you wish to redo parts of a report you have already handed in as part of a group in a previous semester, then to avoid any issues about plagiarism please keep attribution to the original group members for those sections you choose not to redo.