Statistical Learning: guidelines for final project

The final project of the course will consist of **a project done in teams of 2 or 3 people** (single person teams could also be exceptionally allowed after motivated request to the lecturer).

The project should go through the following steps:

1. **Obtaining data.**
   Explain how you obtained the data.

2. **Clean and filter data.**
   This process involves organizing and tidying up the data, removing what is not needed and identifying what is missing. In this process, you may also need to convert the data from one format to another and consolidate everything into one standardized format across all data.

3. **Explore data.**
   Once your data is ready to be used you will have to examine the data. Usually, in a work environment, your boss will just throw you a set of data and it is up to you to make sense of it. So it will be up to you to help them figure out the relevant questions and formalize them into "data science" questions.

   To achieve that, we will need to explore the data. First of all, you will need to inspect the data and its features. Recall also that different data types like numerical data, categorical data, ordinal data etc. require different treatments.

   Then, the next step is to compute descriptive statistics and utilise data visualisation to help to identify significant patterns and trends in the data.

4. **Model data.**
   In this step you may, for example, use regression and predictions for forecasting future values, and classification to identify groups.

   One of the first things you need to do in modelling data is to reduce the dimensionality of your data set. Not all your features or values are essential to predicting your model. What you need to do is to select the relevant ones that contribute to the prediction of results or apply shrinkage regressions.

5. **Interpreting data.**
   Interpreting data refers to the presentation of your analysis. Technical details are necessary but make an effort to make you presentation accessible also to a non-technical audience. You should deliver the results to answer the questions you asked when we first started the project, together with the actionable insights that you found through the process.

In this process, technical skills only are not sufficient. One essential skill you need is to be able to tell a clear and actionable story.

6. **VERY IMPORTANT REMARK:** Keep in mind that the project is aimed at evaluating what you have learnt from the Statistical Learning course. For this reason, in your project work you are allowed to use **only** methods and models presented in the course and that you can find either on the slides or on the R-script provided. For example, you are not allowed to use `Tidyverse` R-packages such as `ggplot` or `ggplot2` as well as any data analysis method that has not been presented during the classes.

7. **Deliveries.**
   Every team is supposed to produce:

   - A final report detailing the data collection, analysis, and results. This should also contain the `R` commands and the corresponding outputs. The use of Markdown with `knitr` is suggested. **The maximum length of the report is 25 pages.**
   - A slide presentation. The presentation time is 30 minutes for a group of 3 and 20 minutes for a group of 2.
   - Both the report and the slides are to be uploaded on a suitable "Assignment" on the Moodle page of the course **at least four days before the date of the exam**.
   - Please, register to the exam on uniweb at least one week before the exam and send an email to the lecturer with the names of the members of your group.