First lab assignment WUM
starting date: April 15th
submission date: April 29th

Dataset: You will work with a sample of the data provided by Central Statistical Office of Poland. The sample is obtained from the Structure of Wages and Salaries by Occupations (SWS) database from October 2010 (a part of Z12 program, for more information see [1], [2], [3]). SWS database covers non-financial entities of the national economy that employ more than nine employees. Data on earnings from SWS are highly reliable because they are reported by corporate accounting departments. The original database contains data on wages and their components as well as selected characteristics of the companies and employees. The available variables for the needs of our project are:

- id – observation id;
- base – total of base salaries;
- bonus – statutory bonuses, awards and discretionary bonuses;
- overtime pay – overtime pay;
- other – remuneration in the form of employee remuneration, additional annual remuneration for employees of public sector entities, payments for participation in profit or in the balance sheet surplus in cooperatives;
- sector – economic sector (1 – public, 2 – private);
- section 07 – NACE section (1 – Public Administration and Defence; Compulsory Social Security, 2 – Education, 3 – Human Health and Social Work Activities);
- sex – the sex of the employee (1 – man, 2 – woman);
- education – highest educational level obtained by the employee (1 – doctorate, 2 – higher, 3 – post-secondary, 4 – secondary, 5 – basic vocational, 6 – middle school and below);
- contract – type of employment contract (1 – for an indefinite period, 2 – for a definite period);
- age – age of the employee as in 2010;
- duration total – total duration of employment;
- duration entity – duration of employment in the reporting entity;
- duration nominal – the time actually worked in nominal hours;
- duration overtime – time actually worked overtime.

Desired output: You are supposed to submit Jupyter notebook with the solutions, commentary, and results by Moodle. Please make sure your notebook opens and works in Google Colab, it will not be graded otherwise. They will be graded by lab assistants of respective groups.

Specific tasks to perform:

Task 1. Data description (3 points total) Download and load the data, describe and summarize them in a few sentences. Leading questions:

- how many observations are there in the sample? Discuss the structure of the dataset: how many quantitative and how many qualitative variables do we have? Are there any missing data? (0.5 point)
- Provide and describe appropriate frequency tables or descriptive statistics for the variables (take into account the type of the variables!) (0.5 point).
- Present and discuss (where appropriate) variables' distributions, e.g.. compare them with the normal, or other distribution by making histograms and plotting them together with a known density function . (2 points)

Task 2. Clustering (7 points total) Explore clustering of the samples. Using the clustering method of your choice and an appropriate distance measure to perform a clustering analysis of the data. (3 points) Describe your approach, in particular choice of variables and their transformations (coding, scaling), including justification of your choices. (2 points). Choose the optimal number of clusters based on the average silhouette score. (2 points)

Task 3. Classification (10 points total) Using the classification methods we have discussed, build a model that is predicting whether an employee has a higher education degree (education<=2) based on the other data available. (4 points). Use cross-validation to assess the quality of your model (2 points) and provide an assessment of the expected performance on the data not seen in training (2 points). Evaluate the relative importance of different variables in the model you have constructed. (2 points),

Task 4. Regression (10 points total) Using the regression methods we have discussed, build a regression model that predicts the base salary variable based on the other variables (4 points). Use appropriate methods to divide your data into training and testing subsets (2 points). Try to build a model that is not using unnecessary variables (2 points). Discuss the role (positive or negative) of all variables included in the model (2 points).