Wrocław, 17.03.2023

**Linear Regression – practice**

**PART 1 (max 4 points if done until end of the classes, max 3 points if done until next classes)**

*You can use all notes/internet etc. Just be sure that you understand what you are doing if I ask some questions J*

*When you upload your solution indicate which tasks you completed.*

1. Read the data from the csv file named LifeExpectance to pandas dataframe
2. Prepare your data by splitting records into train dataset and test dataset. You can use functions available in libraries like train\_test\_split from sklearn, write your own function that will do that randomly. Eventually you can choose for example records with years 2003, 2008, 2013 as a test set. Show basic information about data:
   1. How many records is in each data set?
   2. Show histogram of life expectance and print statistic information about it (mean, standard deviation etc.)
   3. Find three countries with highest expectancy life
3. Use data from train set to fit three models (using simple regression) which are based on following parameters:
   1. GDP
   2. Total expenditure
   3. Alcohol
4. Find coefficients (slopes and intercepts) and scores of regression line for models from point 3. Show charts with data points from training set and regression lines. Show equation of regression line for each case on the chart.
5. Using models created in point 3 predict values of life expectance for data in test set. Find the average error for all three models as well as standard deviation for these predictions
6. Prepare report that sum up your activity: all information that is obtained from data and charts from previous points. Write short conclusions.

**PART 2 (additional 1 point if done until end of the classes; additional 0.5 point if done until next classes)**

1. Choose four parameters that are best to use in prediction of life expectance. Justify your choice in final report.
2. For your four parameters prepare model using multilinear regression fitting it on test data.
3. Print coefficients and score for the model. Predict values for test set and print statistical information about errors: average, standard deviation etc.
4. Compare results with that from part 1 point 5. Write conclusions about this comparison.