For this task you need to download **4** time series from the Yahoo!Finance website:
Any student should have their own unique set of time series!

Please collect available data for three years **2018-2020**

Please pay attention that for your analysis the time moments should be sorted from oldest to newest.

Use the daily closing price.

1. Data evaluation and elementary preprocessing. Analyse completeness of data. Are there missed data (besides weekends)? How many missed data points are in your time series? Are the dates of missed values the same for all your time series? What may be the reasons for missing? How can you handle the missed values in your data (explain at least three approaches)? Use the simple rule: fill in a missed value by the closest in time past existing value. Plot the results. Normalise to the z-score (zero mean and unit standard deviation). Plot the results. (15 marks)

3. Segmentation. Prepare the bottom-up piecewise linear segmentation for the transformed and normalised log-return time series. Use the following mean square errors tolerance levels: 1%, 5%, 10% (the thresholds of the mean square errors). Plot the results. Are the segments similar for different time series you analysed? (25 marks)

4. Prediction. Chose one of the transformed and normalised time series as a target $g(t)$ and other 3 as supporting data $d_1(t), d_2(t), d_3(t)$, where $t = 1, \dots, T$. Provide scatter diagrams of ($g(t)$,$g(t+1)$). Evaluate the error of the "next-day forecast", $\hat{g}(t+1) = g(t)$.

Use data for 2018 as the training set and find the predictor of $g(t+1)$ (the next day value) as a linear function Ψ of $g(t), d_1(t), d_2(t), d_3(t)$:

$$\hat{g}(t+1) = \Psi\big(g(t), d_1(t), d_2(t), d_3(t)\big) \qquad (1)$$

(linear regression). Evaluate the training set error. Use data for 2019 as a test set and evaluate the test set error for this set. Also, use data for 2020 as a test set and evaluate the test set error for this set. Compare these errors. Compare these errors to the errors of the "next-day forecast". Comment. Provide plots of $g(t)$, $\hat{g}(t)$, and the residual. Present the ($g(t)$, $\hat{g}(t)$) scatter diagram. (30 marks)

5. Adaptive predictors. For each given value of the "frame width", Δ=5, 10, 30, create and test the following adaptive predictor. For every $T > \Delta$ create the training set with Δ input vectors $\big(g(t), d_1(t), d_2(t), d_3(t)\big)$ ($t = T - \Delta, \dots, T$-1) and the corresponding outputs $g(t+1)$.
In more detail, the input vectors $\boldsymbol{x}_i$ and the output values y_i for a given T are

$$\boldsymbol{x}_1 = \big(g(T-\Delta), d_1(T-\Delta), d_2(T-\Delta), d_3(T-\Delta)\big), y_1 = g(T-\Delta+1)$$

$$\dots\dots\dots$$

$$\boldsymbol{x}_i = \big(g(T-\Delta+i-1), d_1(T-\Delta+i-1), d_2(T-\Delta+i-1), d_3(T-\Delta+i-1)\big),$$
$$y_i = g(T-\Delta+i)$$

Where i=1,2,…, Δ.
Find the linear regression (1) for each $T > \Delta$. Test this linear regression for the next time value, $t=T+1$.
In more detail, for each T there is one test example with the input vector $\boldsymbol{x}_{test}$ and output value y_{test}:

$$\boldsymbol{x}_{test} = \big(g(T), d_1(T), d_2(T), d_3(T)\big), y_{test} = g(T+1)$$

Please pay attention that this example does not belong to a training set for this value of T.

Find the residuals at these test time moments. Plot these residuals and the values $g(t)$, $\hat{g}(t)$. Present the $(g(t), \hat{g}(t))$ scatter diagram (t=T+1). Calculate the mean square error. Compare to the previous task. Comment. (30 marks)