1.        (0.2p) Divide the dataset into calibration partition and test partition and present the reasoning behind your partition.

2.       (0.2p) Pretreat the dataset: remove the rows missing values, if there are any. If it is more convenient to remove a variable that generally has little entries, do so. Center and scale the calibration partition, and use the mean and standard deviation of the calibration partition to center the test data. If there is any categorical data (values that do not represent numbers on a continuous scale with a known step), we drop it from the data.

3.        (0.2p) Calibrate a PLS and a PCR model for the calibration partition.

4.       (0.4p) Analyze the PLS and PCR models comparatively. Plot and comment on the plot the explained variance in X and the explained variance in Y for the two methods. What can you observe?

5.       (0.5p) Decide on the optimal number of latent variables, both for PLS and PCR. For deciding, use MSE plots for the test partition and the Q2 indicator.

6.       (0.5p) Using bar plots of the regression coefficients (both for PLS and PCR), decide on the importance of the original variables to the model. Comment on the differences between the models.

7.       (0.5p)  Interpret a PLS triplot with X-side weights(W), Y-side loadings(Q), and x-side scores (T) (or alternatively T, P, Q).

8.       (0.5p) Calibrate PLS and PCR models with the most important variables in prediction selected at the previous step (you can choose a number). Does anything change in the model performance, when considering Q2 and MSE indicators? Comment on the differences between models.

9.       (0.5p) Plot the true values and the predicted values against each other. Can you say that the model has a better performance on a certain interval of values?

10.      (0.5p) Plot the residuals as a function of the true value. Do you have larger residuals at the extremities of the interval? Comment on the differences between models.