# Programming assignment #4

Course: CHE1147H - Data Mining in Engineering

## 1 Chemical composition of pottery

You work for a pottery manufacturing company that produces two different types of products A and B. Product A uses raw material from Llanedyrn and Product B uses raw material from Isle Thorns and Ashley Rails[1]. Your company is informed that Llanedyrn will be closing soon for maintenance and your entire production of product A is at risk.

You received a potential new source of raw material from site Caldicot and you analyzed two samples to compare them to your existing samples from the other three sites. As the new data scientist of the company you are asked to look into the data and give your recommendation regarding the suitability of raw material from Caldicot as a replacement for Llanedyrn.

The ask *"look into the data"* normally calls for an unsupervised learning exercise, since there is no clear output you are asked to predict. You will investigate the multivariate chemical composition of the raw material from four different sources using Principal Component Analysis.

### 1.1 Data exploration with PCA

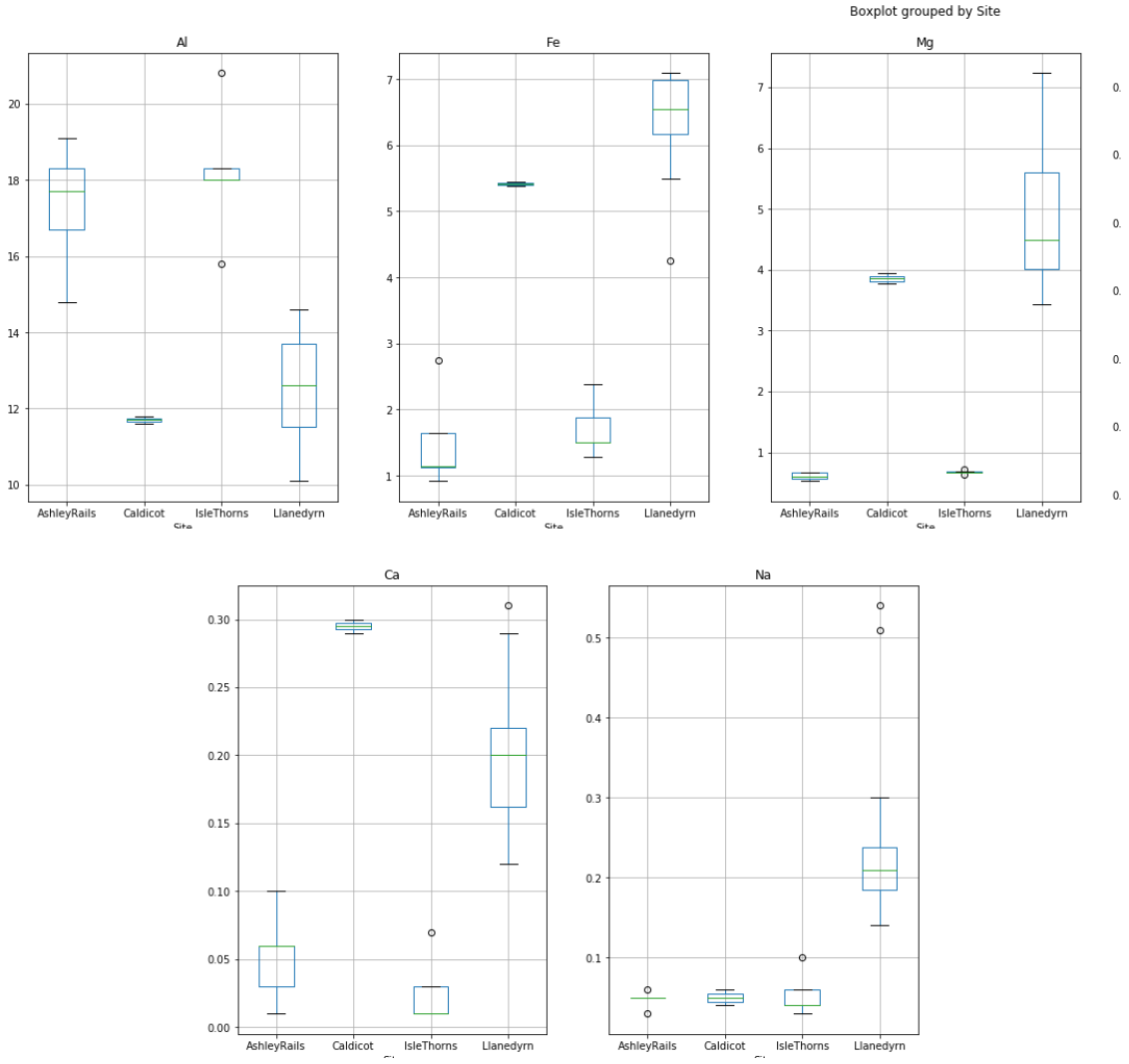
1. Import and view the data. How many columns do you have? Which columns will you use in your PCA?
2. Pre-process the data and perform PCA with 3 PCs.
3. Plot the cumulative explained variance graph. What percent of the variance do the first 2 and 3 components describe?
4. Plot the scores-loadings graph for PC1-PC2. Visualize the different sites with a different colour or symbol.
5. How does the map of scores-loadings explains the reason that your company uses the raw material from Isle Thorns and Ashley Rails to manufacture Product B?

[1]These are sites in Great Britain (or Game of Thrones if you prefer). The data come from ancient pottery findings and the problem is fictitious.

6. Is the raw material from Caldicot a good replacement for Llanedyrn? Yes or no and why?

7. What are the biggest differences in the two big clusters? How are the two samples from the candidate Caldicot different than the Llanedyrn samples?

8. Confirm the answers by producing the boxplot of the 5 variables grouped by the site of the raw material shown below.




**Final note:** In this problem, we reduced the number of variables from 5 to 2 in order to visualize the characteristics captured in the 5 variables. With more than 5 variables you realize that it becomes difficult to visualize and compare the different samples. Dimension reduction methods like PCA are crucial to understand multivariate data. Conventional statistical analysis like the boxplot shown here do not show the correlations between the variables which are simply captured in the PCA plots you created.

# 2 Batch data analysis

In this problem, we will look into batch data; dynamic time-series of a finite duration [2]. Batch manufacturing processes are very common in chemical, pharma, bioengineering and semiconductor industries such as baker's yeast production, beer brewing and vaccines production.

In theory, a reactor is designed with temperature, pressure, level, pH control and multiple sensors that measure these variables among others. A perfect batch (again in theory) is one that is tightly controlled to the specifications and as a result the productivity and quality of the final product is optimized.

In real life, a typical batch is run from a few hours up to a week or two and a lot of things can go wrong during this period. There is always variability either because the process is very sensitive to minor fluctuations in some variables or the control of some variables failed for a period of time.

In a company that implements Data Analytics or Multivariate Statistical Process Control (MSPC) monitoring is typically implemented with the following steps:

1. Identify a number of reference, perfect **historical** batches (15-20), both in terms of high productivity/quality and minimum anomalies or fluctuations around the setpoints.
2. Create a PCA model of the perfect batches identified. This is your **model**.
3. Every time your site is running a new batch, **fit** your data **online** or as soon as your data infrastructure allows you to do so. Fitting will tell you whether your batch is similar to the perfect batches or it is deviating from the reference behaviour.

Next, you will follow these steps to build a Batch Statistical Process Control and implement it to monitor a new batch (we will assume that you got the data at the end of the batch and fit them to the model). The dataset is from a baker's yeast production facility in Solna, Sweden capturing the last step of the fermentation.

## 2.1 Build a Batch Statistical Process Control model

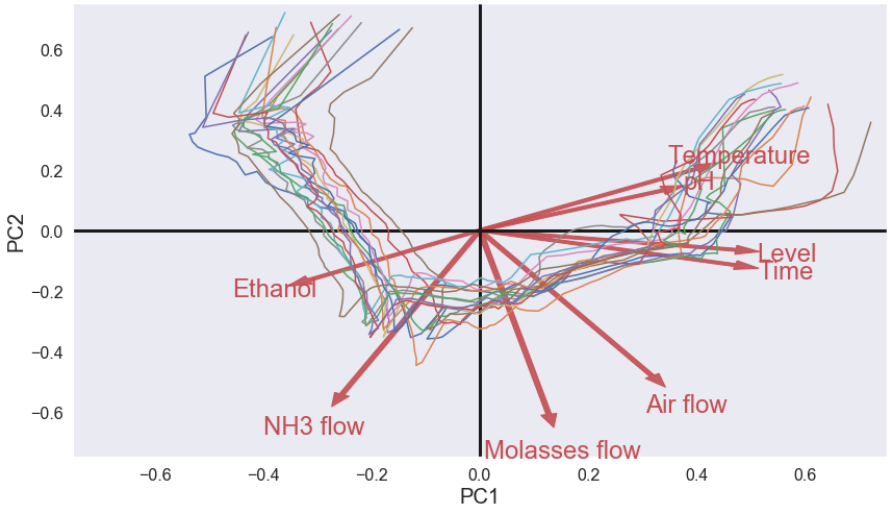
1. Import the data from 'bakers_yeast_reference_batches.xlsx'. Identify how many batches are in the data. What is the duration of each batch and how many data points are there per batch? How many variables are measured (including time)?
2. Plot the variables time-profiles in a 2x4 subplot. Inspect the graphs (don't just plot them). Look for potential outliers. Which variables have the largest variability? Which variables are tightly controlled?

[2]The data are taken from Chapter 16 of **Multi- and Megavariate Data Analysis** from Umetrics Academy. The problem in this assignment though is reformulated and is not the same as the one described in the book.

3. Select the features (including the Time column), pre-process the data and perform PCA with 5 principal components. Extract the scores and loadings.

4. In order to plot the scores-loading plot, you need to pivot the scores BatchID with index 'Time' (use pandas pivot_table).

5. Plot the scores-loadings plot with one line per batch (this is why the pivot in the previous step was needed). The output should look similar to the plot below. You may choose a different scaling, but the trend should be the same as this graph.




6. Explain this graph. In which quarter do the batches start and end? What happens at the kink where the direction of the lines changes? Can you tell from this graph which variables do not change in the first phase and which in the second phase?

7. Plot the cumulative explained variance. How much variance do the first two principal components capture?

## 2.2 Use the model to monitor running batches

The goal of building an unsupervised model is to monitor the running batches. Your site runs two reactors in parallel and here you will fit the data from these two reactors to the model previously built and identify potential problems and outliers[3].

1. Load the data from the file 'todays_batches.xlsx' and repeat the same procedure as in the steps 3-4 of the previous section with the exception of the PCA modeling. Here, instead of fit the data to the model and transform, you will only transform them with the model object you created in the previous section.

[3]Ideally, in most industries you have the data available online and you get a new data point every minute or so. Then you fit every coming point to the model and overlay it with the graph from the previous section.

2. Plot the same scores-loadings plot for the data in the batches you used to develop the model with solid lines. Overlay the new incoming data from the two current batches with dashed lines and two different colours to distinguish them. Also, add a legend for the two batches so that the viewer can distinguish them.

3. Do the batches show behaviour similar to that of the reference ones or there are outliers indicating potential problems?

# 3 A PAT application

In 2004, the Food and Drug Administration (FDA) released a guidance to encourage innovation in pharmaceutical development, manufacturing and quality assurance. The framework, called Process Analytical Technology (PAT), combines MultiVariate Data Analysis (MVDA), Design of Experiments (DoE) and process analytical chemistry methods such as UV, IR, NIR, NMR (fast, precise, online and preferably non-invasive methods).

Here, we use PCA and PLS, two classic MVDA approaches to assess the feasibility of using NIR to measure a critical quality component of a wood product, the composition of wood fiber. In the first part, we will perform unsupervised data exploration with PCA. In the second part, we will run PLS, a supervised learning algorithm, to quantify the outputs of interest (i.e. wood composition).

## 3.1 Description of the process

Sawdust from industrial sawing of birch, pine and spruce is mixed at different ratios to manufacture a final wood product. The NIR spectra for 15 samples of known composition of the three different types of wood were used to train the model. Each sample was tested twice (resulting in 30 rows) and NIR spectra data was collected in the range between 1100-2300 nm. Thus, the input X-table is (30 rows)x(1201 columns) and the output Y-table is (30 rows)x(3 columns), where the three outputs/compositions are recorded as percent proportion of spruce, pine and birch. Data for the training of the model are in the file **sawdust_train.xlsx**

In order to assess the accuracy of the model developed, we have a test data set of NIR data (X) and proportions Y different than the training data set. The test data set is not "seen" by the model during training. The proportions in the test set were selected to be different than those in the train set to assess the predictive power of the model under realistic conditions. There were 12 samples measured twice, therefore the number of rows in the X and Y-tables is 24. The test data are in the file **sawdust_pred.xlsx**

## 3.2 Task #1: Unsupervised learning, PCA of NIR data

A. Generate the plot of all the **training set** raw NIR data.

B. Pre-process the NIR data and perform PCA for 5 components.

C. Plot the cumulative explained variance versus the number of components. How many components do we need to capture 99% of the variance?

D. Plot the PCA scores for the first two components. Comment on any outliers, trends, clusters etc (if any).

E. Plot the loadings of the first two components and comment on which wavelengths seem to be important for different PCs. **Don't** describe the graphs in words. In two sentences, describe any significant trends or signals (if any).

F. Apply the dimensionality reduction with the PCA model you developed in step B for the **test set** using the **transform** method.

G. Create a scatter plot of the scores in PC1, PC2 of all 54 points in the train and test data set, coloured by the set they belong to (suggestion: use sns.scatterplot). Why is it important to overlap the scores of both train and test set in one graph? What are you checking to see?

## 3.3 Task #2: Supervised learning, PLS model

A. Run the PLS regression to model all 3 outputs with the scaled train set spectra. Increase the number of components to achieve $R^2$ value of at least 0.92 (use PLSRegression's method score to estimate $R^2$).

B. Predict the output values for the inputs in the test set and compare them by plotting the 24 actual vs predicted values in a 1x3 subplot (for spruce, pine and birch, respectively). Also, include the y=x line for reference.

**Submit three .ipynb files, one for each section:**
potery_LastName_FirstName
batch_LastName_FirstName
PAT_LastName_FirstName