**Fault Detection and Sensor Diagnostics Using Control Charts**
**Week 3: Data Pretreatment**

Github Repository: Wind Turbines

The goals of this week's task are as follows:

- Feedback incorporation and adjustment of the initial pretreatment plan
- Division of data into calibration, validation and test partitions
- Data centering and scaling techniques
- Evaluation of extreme/ missing values and mitigating actions, data synchronisation (for time series), sampling
- Visualising pretreated data

From our data pretreatment plan, we had intended to do the following:

- *Discard the dataset for `WT3` given the above outlined issues*

  We have implemented this step using the `clear` command.

- *Drop one of the columns in the dataset for `WT2` in order to get an equal number of features for each of the datasets used in the analysis. This is because the dataset for `WT2` has one more feature than the other two datasets*

  Using the provided dataset hint, we have dropped the last variable in the `WT2` dataset. This variable represents an extra quality not measured for the other two remaining turbines (`WT14` and `WT39`).

- *Determine a strategy for handling the missing value*

  Since we are not using the time series for regression, we have dropped the single observation with the missing value in the dataset for `WT14`.

- Centre and scale the datasets

  We have also centred and scaled the three datasets individually using MATLAB's `zscore()` function.

Notably, we have incorporated some of the feedback into our pretreatment plan. It is also worthwhile to mention that given the nature of our project, we have not partitioned the data into calibration, validation and test sets. This is because we intend to calibrate the control charts using the healthy turbine PCA model and then project the faulty turbines into the model.

Finally, the pretreated data has been represented in the following boxplots:

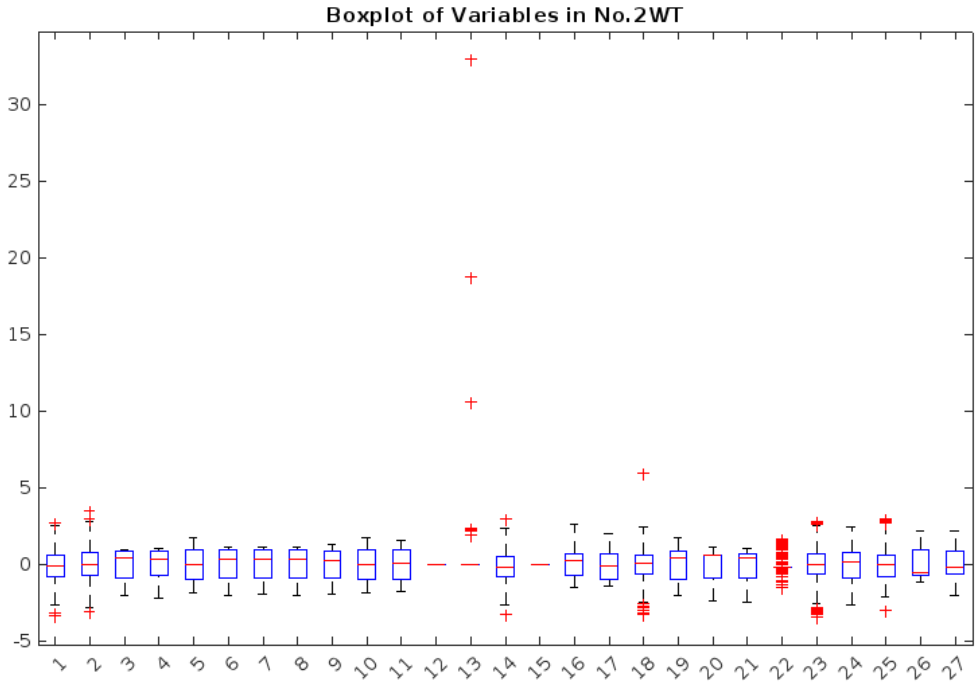


Figure 1: Boxplot of variables from turbine *WT2*

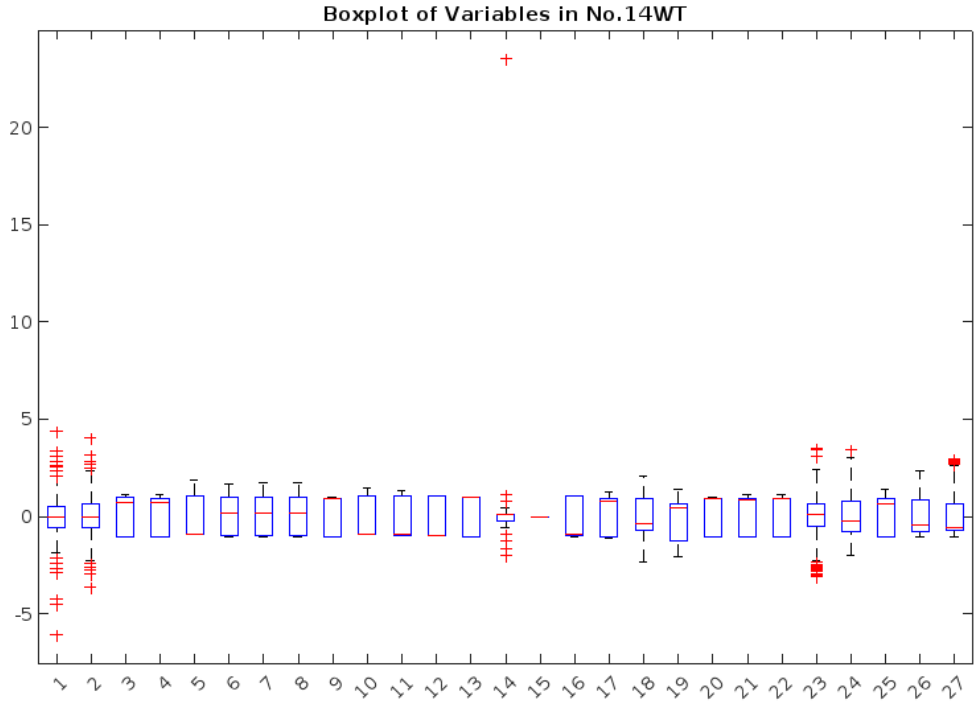


Figure 2: Boxplot of variables from turbine *WT14*

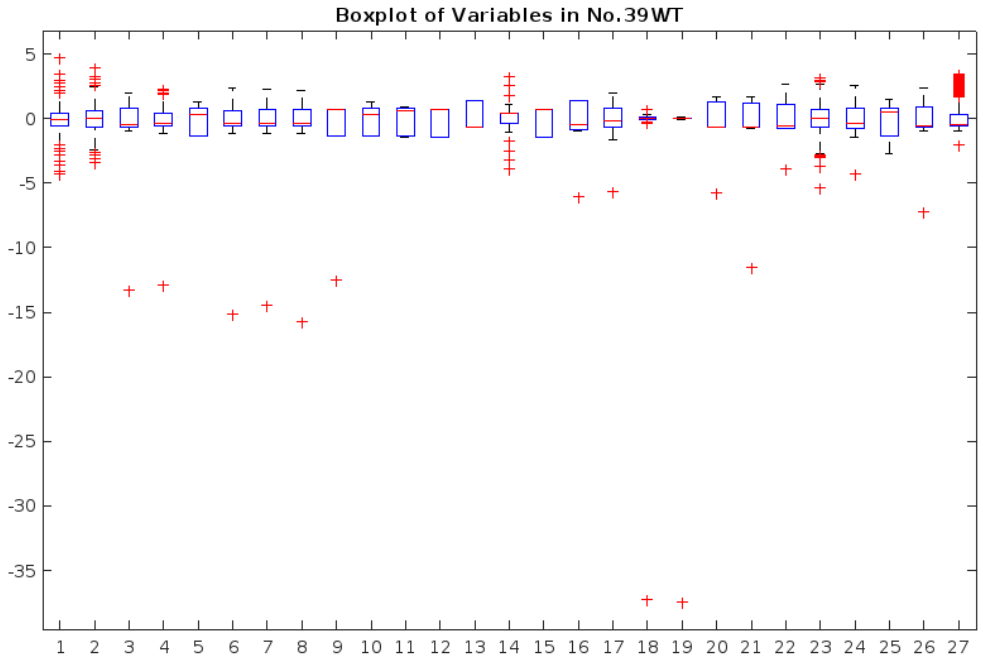


Figure 3: Boxplot of variables from turbine *WT39*