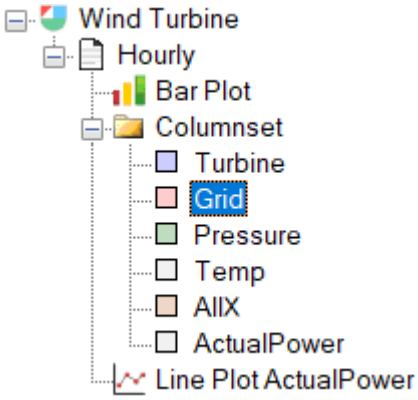
Open the file “WindTurbineData.unsb”

The data set has hourly data for one year for one wind turbine.



ActualPower is the generated power and is the response variable.
All x-variables are defined in the column set AllX.
The other column sets that are to be used as X are subsets of this set.

The objectives with the analysis:

Train and test various models for the ActualPower.

- Start with PCA and see if there are any time dependencies (hint: show Month as category variable in Sample Grouping)
- Try both with PCR and PLSR. Are there differences correlation structure in the correlation loadings for the two methods?
- Make models with selected column subsets and the column set “AllX”; compare results
- Find the optimal number of PCs/Factors
- Identify outliers and remove them if it can be justified. Does the RMSE improve after removing them?
- Test different validation schemes, compare RMSE
- Select a subset of the variables by marking in the correlation loadings plot and/or regression coefficients plot. What is the least number of variables to keep and still have the same RMSE as with all variables (or the best subset of X)?

Hints:

- Plot selected variables over time
- Show scores both in 2D/3D and as line plots. Mark suspicious samples in 2D and see where they are on the timeline.
- Choose “Tasks - Recalculate” to make models with/without samples and/or variables. Or right-click the model name in the overview to the left.

You will work in groups of three or four. For those of you who are experienced Unscrambler users by now (!?), I’ll ask you to take the lead in each group.