Course Project Guideline 2

In this guideline, we want to point out a few considerations you need to take into account for your project to be successful.

In Phase 1, (II), in addition to the linear regression, it is **optional** to use other statistical model(s) and try to fit them on the train data and show the differences of the test data to train data errors.

You should not train an HMM on the whole training data set!
Based on the third question of the second group assignment, you should already have a sense of seasonality of the data to be able to choose meaningful time windows for your models. Use these time windows for the project.

In Phase 2, Approach 1, (II), which is about calculating the moving average, you can use an R package (Google it!) or you can implement it by yourself.

In Phase 2, Approach 2, what you need to do is to train an HMM (with the right number of states, strong feature(s) for training, the right time window etc.), then **use the same model** to calculate the log-likelihood of the test data.

- In order to fit a model on the training data, you should train a number of models and pick those which have **high log-likelihood** and also **low BIC** (try to use BIC because AIC doesn't penalize the number of parameters as strongly as BIC).

- To calculate the log-likelihood of test data with the depmixS4 package, you should extract and transfer the specification of the trained model (transition matrix, emission matrix, number of sates, feature etc.) to a new model and basically recreate the same model. Then calculate the log-likelihood of the test data with the model.
For more information look at the documents of the R package (depmixS4) and pay more attention to **page 4, 5** and **16 to 21**!