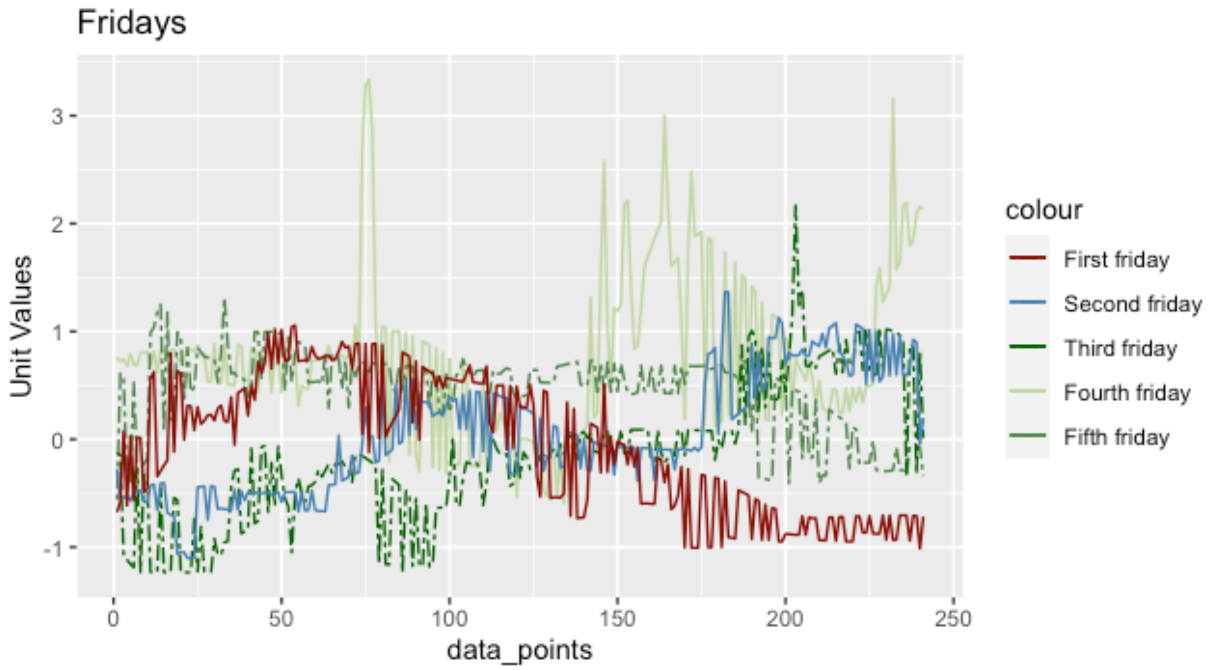
Group Assignment 3
Group 15
Karim Khoja
Bhakti Bhanushali

We were tasked with selecting a variable from the dataset and over a period of 3-4 hours on a weekday and evaluate the hidden Markov model to conclude the number of states based on the log likelihoods and the BIC.

Hidden Markov models are probabilistic frameworks where the observed data are modelled as a series of outputs generated by one of several hidden internal states.

Firstly, we selected Friday from 6:00 PM to 10:00 PM as the weekday of interest. Intuitively, on Fridays people are usually resting at home and using appliances that resulting in a possible pattern over the time frame. We also plotted the data points from this time frame over 5 Fridays of the year and we observed a certain pattern.



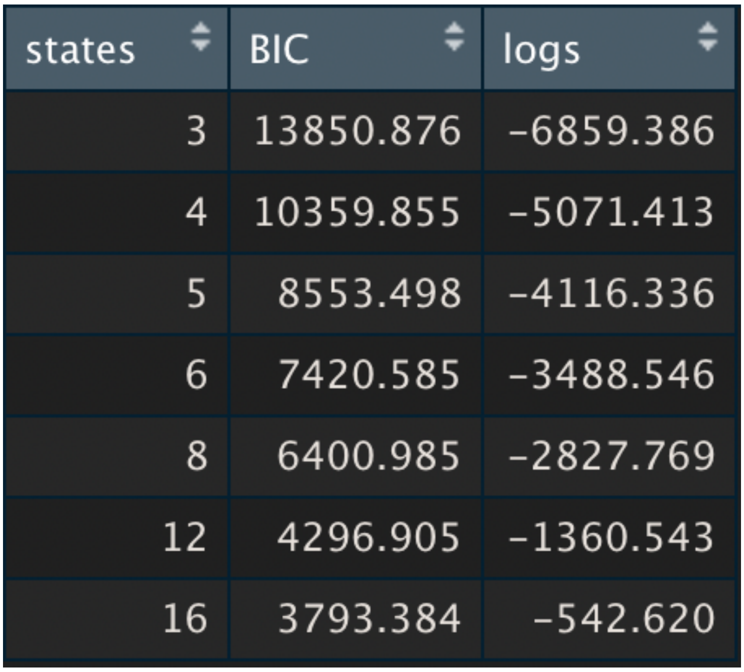
Next, we extracted 'Global active power' from the dataset and pre-processed it by scaling it. Doing so makes the values of the datapoints closer together which makes the algorithm more accurate and trains the model faster. If they are far away then the model will take time to find patterns and understand the data.

Lastly, we assume that the dataset follows the Markov process .i.e. the output at the present state only depends on the value at the previous state and not the ones before that. We trained our dataset using the Hidden Markov Model with different states ranging from 3 to 16. Since the states are hidden When we cannot observe the state themselves but only the result of some probability function(observation) of the states we utilize HMM.

In our experiment our observations are the global active power datapoints.

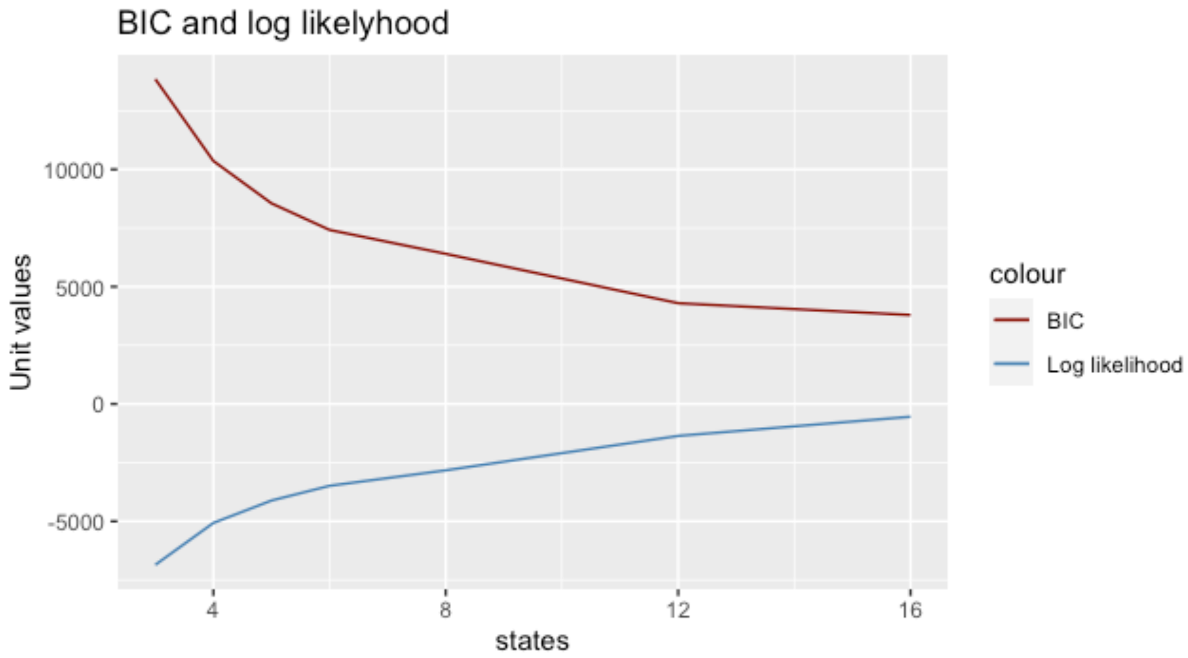
We set our starting point as 2 and then used the depmix function on our response variable with ntimes as a vector with 52 entries of 241. This indicates that our response variable is not one huge data collection but groups of data collection over different periods of time. It returns several values from which we use the log likelyhood and the BIC. The former is the log of how likely the observation is given a specific number of state and the latter is Bayesian information criterion. It is a measure of the complexity of the model. We want our model to have a high likelihood which can be done by increasing the number of parameters, making the model more complex, but this results in overfitting. Consequently, a more complex model is good for the training dataset but not the test dataset. To solve this problem we use the BIC. BIC introduces a penalty term for the number of parameters in the model. Therefore, we want a model with a high log likelyhood but a low BIC (low complexity) .

We picked different nstates : 3,4,5,6,8,12 and 16. The respective BIC and Log likelihood are presented in the table below.



states	BIC	logs
3	13850.876	-6859.386
4	10359.855	-5071.413
5	8553.498	-4116.336
6	7420.585	-3488.546
8	6400.985	-2827.769
12	4296.905	-1360.543
16	3793.384	-542.620

The respective plot of the BIC and the log likelihood for the states is given below.



It can be observed that as we increase the number of states the log likelihood improves as more number of parameters are added but the BIC reduces. The pattern of the graph indicates that the BIC and log likelihood will converge at some state after 16 but that was beyond the scope of the assignment. If we look only at this dataset and want to conclude the number of states that best fit it then 16 would be the ideal choice since the log likelihood is -542.620 and the BIC is 3793.384 both being the highest and lowest respectively. But, this might be an overfitted and thus might generate a lower log likelihood for test data. Since testing it on test data to find the overall best model was out of the scope of the assignment, we assume that state 12 might be a better fit. It has a log likelihood of -1360.543 and a BIC of 4396.905. Further computation on test data is required to say the same with certainty.