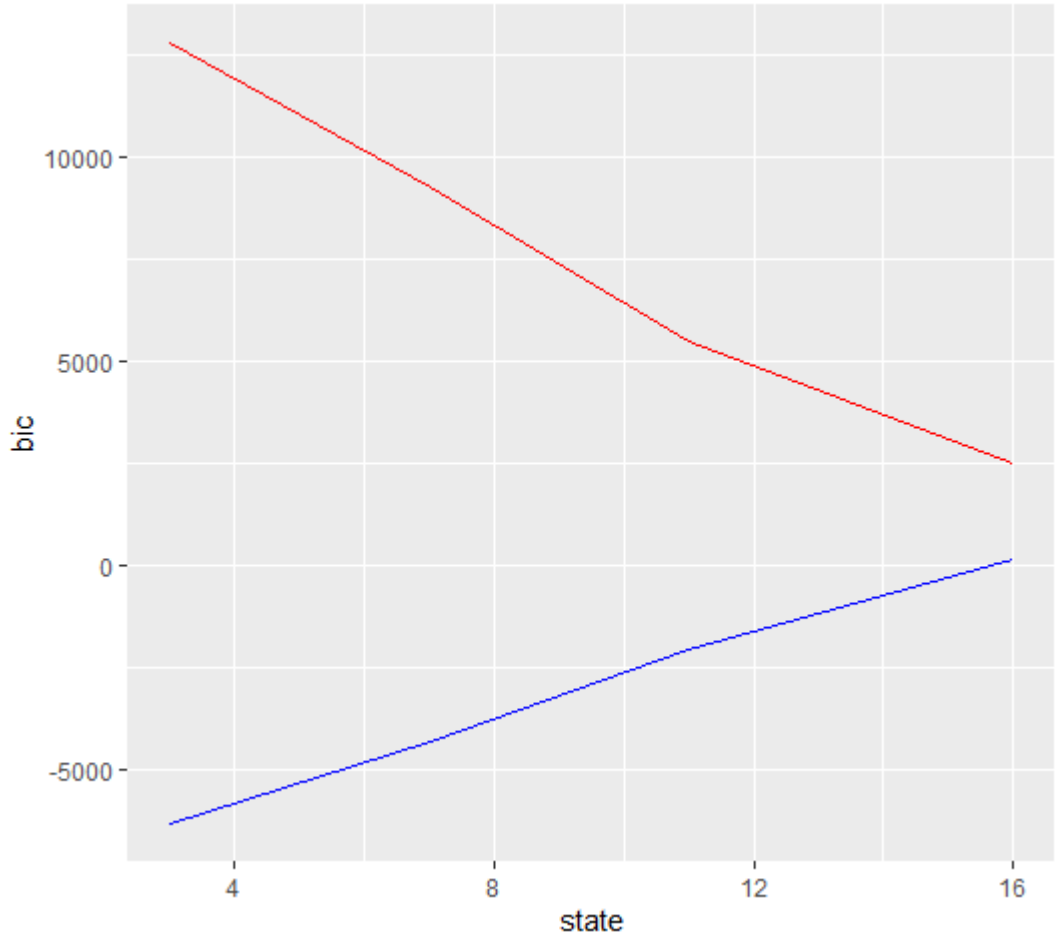
Assignment 3

For every HMM we tested our group chose global active power and used scale() on it. Originally we chose Monday over the time period 8am - 12pm. We chose 8 to 12 since most people would be waking up leading to a nice consumption of power in the household. This initial dataset gave HMM's with positive log-likelihood and a BIC that was only decreasing.

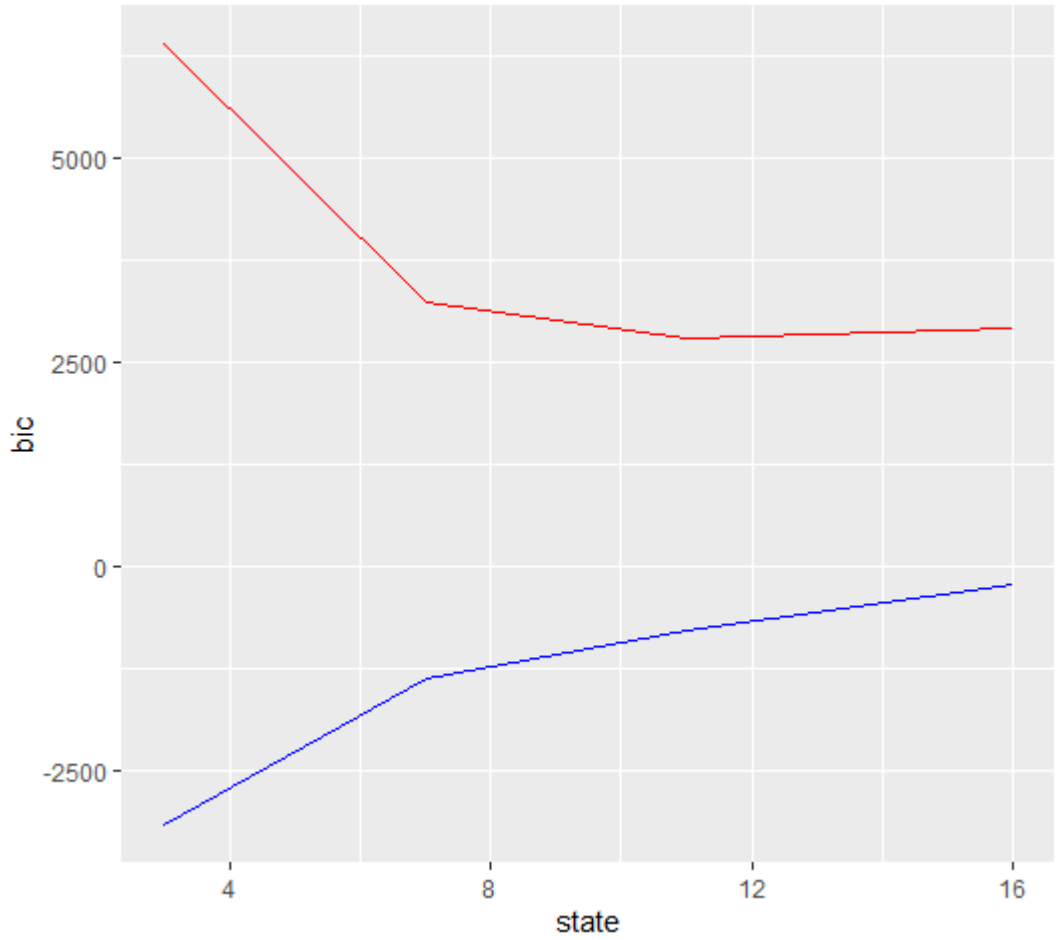


We found this outcome to be weird and not ideal since log-likelihood is not supposed to be positive and BIC should increase since BIC is used as a measure of the complexity of an HMM. From there we started testing different time windows to see if this was a common occurrence. Later we found that 4 hours was a much larger dataset leading to longer times to obtain HMM's and decided to use a 2 hour window instead so we could test more HMM's faster.

After testing many different time windows we found that the time window 9am to 11am had relatively negative log-likelihood but has less prominent minimum BIC. The time window itself seemed to have a distinct pattern. For these reasons we settled with this time window as our final decision.

When testing different Hidden Markov Models we started with low states and got very low log likelihood. This is inline with what is to be expected when starting with low states. We then skipped a few states to see noticeable differences in log likelihood with each test using states 3, 7, 11, 16.

# of states	3	7	11	16
log-likelihood	-3138.234	-1354.584	-780.6573	-204.3333
BIC	6399.077	3252.151	2804.92	2922.151



Here the blue line represents the BIC and the red line represents the Log-Likelihood. This gave us a better idea of which states we want to focus on rather than creating an HMM for all states from 3 -16. Here it seems that the BIC would dip around state 11 before rising again so we chose consecutive numbers in that range.

We compared the states 11, 12, 13, 14, 15 and determined state 13 to have the lowest complexity to highest log likelihood. We chose to omit states 14 and 15 since they gave a positive log-likelihood which is possibly an error in R itself.State 13 has the lowest BIC which means the model doesn't overfit and the largest log-likelihood which means the model predicts the observation very well. Looking at the graph we concluded that around state 13 the complexity starts rising again which is why we chose state 13 as well.

# of states	11	12	13	14	15
log-likelihood	-780.6573	-687.2149	-44.43023	105.479	49.16805
BIC	2804.92	2836.98	1787.87	1742.028	2126.141