Breakout Session 7: Model Selection

We will go back to using the bike data set (refer to Breakout Session 3 for information on the variables in the bike data set). We will work through several different concepts and possibilities within this data set (the data is located on the GitHub). First, let's naively try an automatic selection with the variables in the data set (remove the dteday variable for any analysis). We are trying to model the cnt of total rental bikes. First, we need to create a training and a test data set. We will use the seed of 18954 and partition the data into a 70/30 split.

1. Using the AIC criterion, do a forward selection on the training data set (with the data set as is…without the id variable). What happens?
2. Now also remove the casual and registered variables and try forward, backward and stepwise with AIC and BIC on the training data set. Comment on the similarities and differences you observe.