\*note: it will work only for regression data sets. If you are not sure just go watch this video

YouTube: <https://www.youtube.com/watch?v=iIUq0SqBSH0>

When you are searching up for dataset, please give keyword “regression” so that it get you the data set for regression.

1. First start by taking in big data file into R
2. Do Some clean up (cleaning up data means that validate if each data cells are valid input, if it is not valid input, omit it by assigning NA or something else to indicate not valid)
3. Examine each variable(column) to which can be predicted and what not.
4. Identify possible relationships of two variables(columns)
5. If some other variable should/must be contained in our model,
   1. Test to see if adding extra variable(column) to the model makes significant improvement.
      1. If yes, create a model using it and go back to step 5. Or proceed to step 6
      2. If no, proceed to next step
6. Do some interpretation of the finding.
7. Do some visualization by plotting scatter plot or histogram. This step is effective to show audiences the relationships of variables(columns).