# Part 1: Preprocessing:

syntax to run the code:

python preprocessing.py <url> <output directory/>

example:

If you are using the cars dataset and output it to the Documents folder of your system, the syntax is as follows:

$ python preprocessing.py <https://archive.ics.uci.edu/ml/machine-learning-databases/car/car.data> C:/Users/pxg131330/Documents/

It will generate 2 files in the Documents/specified directory:

1. original\_dataframe.csv
2. standardized\_dataframe.csv

For categorical to numerical encoding purpose, we use power of 2 encoding method. Suppose an attribute , X1, have 4 values ‘huge’, ‘big’, ’medium’ and ’small’ . The encoding will be done in the following way:

1. Huge: 20 = 0001
2. Big: 21= 0010
3. Medium 22=0100
4. Small 23=1000

Same encoding scheme is used for class encoding also.

For standardization of numerical values, we used the following formula:

For output purpose, if the class value is numerical, we keep it as it is and if the class value is categorical, we convert it to the encoding scheme mentioned above.