## Reading Material on class imbalance: (good context for working people to describe the problem)

### 1. Pre-processing task:

TDS Blog with good introduction to data-preprocessing:

<https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825>

### 2. Handling imbalanced datasets:

<https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28>

I know it can be overwhelming to read fully, at least read the introduction, I think you should find it interesting!

## Assignment: evaluatable: play with MCQ

1. Which of the following are likely to be examples for class imbalance when predicting for following use cases? Mark all that are correct
2. Predicting Spam messages vs Important messages in your email
3. HIV Positive vs negative cases tested among people picked at random
4. Predicting the coin toss if it is heads or tails
5. Predicting fraudulent transactions from normal bank transactional activity
6. Which of the following are techniques for handling missing data?
7. Imputation (fill) by standard deviation of the variable for numeric variables
8. Multiple Imputation
9. Imputation (fill) by mode of the variable for categorical (non-numeric) variables
10. Regression Imputation
11. Which of the following is true about scaling/normalization/standardization of variables?
12. Scaling of variables brings all the variables to same scale, making the effects of variables on the outcome comparable
13. Scaling is important for techniques which involve measuring distances between variables.
14. Without scaling , it is wrong to build machine learning models using unscaled data.
15. If you build models using scaled data, it may require scaling back to original variables to interpret variables’ effect on outcome predicted.