# Important things to remember

## When cleaning data

1. Drop columns that have more than 30% null values.

## When processing data

1. If there is any need to do scaling, fit and transform on the training set. Thereafter, transform on the test set when you are assessing the model. DO NOT FIT TRANSFORM the test set. This is to prevent data leakage

# Modeling

1. Ensemble technique:
   1. Averaging (Bagging)
      1. **How?**
         1. We learn k base classifiers on k different samples of training data. These samples are independently created by resampling the training data using uniform weights
         2. In order to promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set
         3. The m models are fitted using the above m samples and combined by averaging the output (for regression) or voting (for classification)
         4. bagging methods work best with strong and complex models (e.g., fully developed decision trees)
      2. **Pros?**
         1. Bagging reduces variance
   2. Boosting
      1. How?
         1. Boosting methods which usually work best with weak models (e.g., shallow decision trees)
         2. Boosting takes a weak base learner and tries to make it a strong learner by re-training it on the misclassified samples.
      2. **Pros?**
         1. Boosting reduces bias
            1. Each weak learner of boosting has low variance and high bias because it use shallow and high bias base estimators
         2. Achieves higher performance than bagging when hyper-parameters are tuned properly
      3. **Cons?**
         1. Difficult and time consuming to properly tune hyper-parameters.
         2. Cannot be parallelized like bagging (bad scalability when huge amounts of data).
         3. More risk of overfitting compared to bagging
      4. **Different types of boosting:**
         1. *Adaboost* – is about re-weighting the preceding model’s error to subsequent iterations
         2. *Gradient Boosting* – is about fitting subsequent models to the residual of the last model
2. Random Forest:
   1. Differs from decision tree bagging in one way: at each candidate split in the learning process, the algorithms selects a random subset of the features. This is known as feature bagging
   2. The reason for doing this is due to correlation of trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the bagging base trees, causing them to become correlated. By selecting a random subset of the features at each split, we avoid this correlation between base trees, strengthening the overall model

# Models FAQ

1. SVM - <https://www.svm-tutorial.com>
2. Gradient Boosting - <http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides/gradient_boosting.pdf>
3. Gradient Descent - <https://machinelearningmastery.com/gradient-descent-for-machine-learning/>