Setting up Environment

Creating Helper Functions

Visualizing Dataset

-

Generating Fake Data

* Split raw data into Train/Valid sets
  + Pros: no train data spill into valid data set
  + Cons: diversity of training model is not guaranteed
* Split data into Train/Valid sets after generating fake data
  + Pros: diversity of image samples
  + Cons: train data spill into valid data set causes overfitting
* Split data into Train/Valid sets according to class shares based on raw data class distribution and generate fake data (based on Train data partition) for modeling
  + Pros: guarantees diverse data modeling and prevents train data spill into valid data set
  + Cons: data split threshold may determine accuracy results for classes of low training samples
    - E.g., class of raw sample size 1 will have 0 validation samples to assess validation accuracy