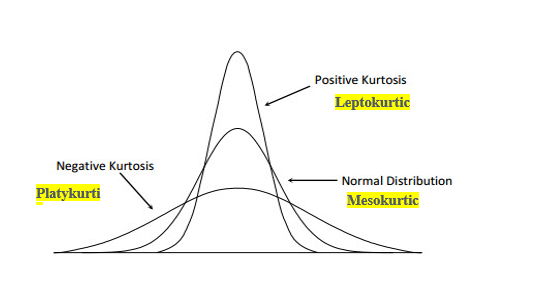
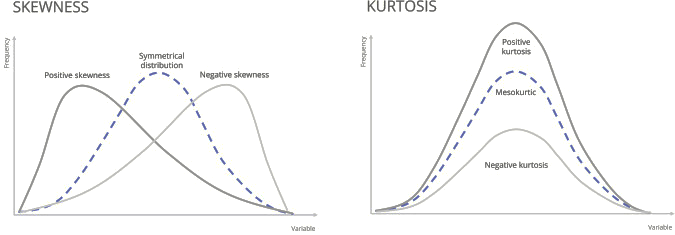
# DASC 522 guide to standardizing & transforming numeric features

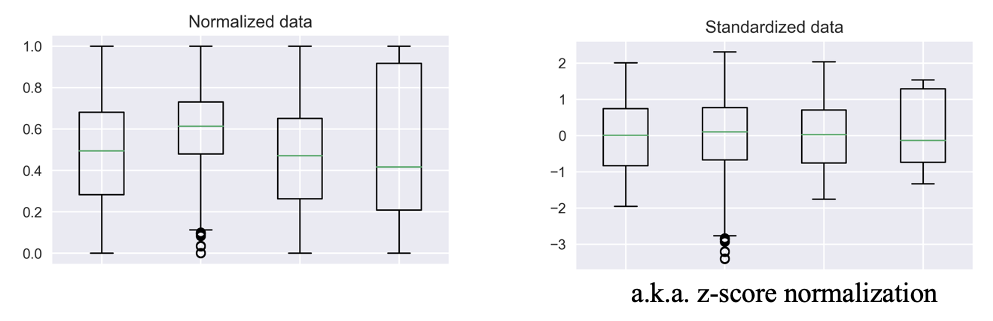
* A numeric feature may have any or all of the following issues:
  + Skew - peak of distribution is to the left or right
  + Kurtosis - peak of distribution is higher or lower than normal
  + Outliers - data far out in the tails of the distribution
  + Non-zero mean
  + Standard deviation other than 1
* Recommend only standardize/transform X features, not the y response variable



## Transform fixes skew & kurtosis

* Many options: log, square, square root, inverse etc. You could try all of them, although…
* Recommend to use the box-cox power transformer (yeo-johnson if feature has negative values) – they generally solve all skew & kurtosis issues, and standardize the data as well. Code excerpt in .ipynb file
* Only retain a transform if it helps model performance, even if the feature looks “more normal” when the transform is applied

## Standardize fixes mean, standard deviation & minimizes effect of outliers



* Min-max normalization (left graphic) isn’t good because outliers can induce kurtosis &/or skew
* Standardize (right graphic) gives mean=0 & standard deviation=1. In python accomplish with:
  + sklearn.preprocessing.StandardScaler() applied to a dataframe feature prior to using classical/statistical algorithms
  + a keras.layers.Normalization() layer in a NN. This performs z-score normalization a.k.a. standardization.
* Code excerpts in .ipynb file