Regression analysis is a method that allows us to fill in missing data based on the expected value of the missing data based on the values present in the rest of the dataset. Generally, it’s an advanced form of the linear interpolation that we all learned in high school algebra. Using this method, as opposed to simply using the overall mean, to fill in the data leads to a more accurate model.

Principal component analysis is a dimensionality reduction technique often used during the data preprocessing step of a model’s pipeline. This method makes heavy use of linear algebra by linearly transforming the dataset into a new set of coordinate axes which has less dimensions than the initial dataset. This allows us to reduce the size of our models by only focusing on these new principal components.

Data cleaning is the collective term for the various methods that seek to either correct or delete incorrect or missing information and is a vital part of data preprocessing. Some methods that fall under this category include binning, regression, and clustering. These are important because we don’t always have enough data to allow throwing out every entry with missing or incorrect data. So, in order to maximize the effectiveness of our dataset we will likely need to clean it.

Data reduction is simply taking our initial dataset and transforming it into a simplified, easier to use form. Methods in this category include wavelet transforms and PCA. The main benefits of this application are that it allows the use of smaller models, helps avoid the curse of dimensionality, and also helps to eliminate irrelevant features and ignore noise in the data.