When we deal with a data set with multiple features, the first thing we should try is to do data reduction by using PCA (principle component analysis) method to convert a set of observations of possibly correlated variables into a set of values of [linearly uncorrelated](https://en.wikipedia.org/wiki/Correlation_and_dependence) variables.

“PCA is mostly used as a tool in [exploratory data analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis) and for making [predictive models](https://en.wikipedia.org/wiki/Predictive_modeling). It's often used to visualize genetic distance and relatedness between populations. PCA can be done by [eigenvalue decomposition](https://en.wikipedia.org/wiki/Eigendecomposition_of_a_matrix) of a data [covariance](https://en.wikipedia.org/wiki/Covariance) (or [correlation](https://en.wikipedia.org/wiki/Correlation)) matrix or [singular value decomposition](https://en.wikipedia.org/wiki/Singular_value_decomposition) of a [data matrix](https://en.wikipedia.org/wiki/Data_matrix_(multivariate_statistics)), usually after mean centering (and normalizing or using [Z-scores](https://en.wikipedia.org/wiki/Z-score)) the data matrix for each attribute.[[4]](https://en.wikipedia.org/wiki/Principal_component_analysis#cite_note-4) The results of a PCA are usually discussed in terms of *component scores*, sometimes called *factor scores* (the transformed variable values corresponding to a particular data point), and *loadings* (the weight by which each standardized original variable should be multiplied to get the component score).[[5]](https://en.wikipedia.org/wiki/Principal_component_analysis#cite_note-5)

PCA is the simplest of the true [eigenvector](https://en.wikipedia.org/wiki/Eigenvectors)-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. If a [multivariate dataset](https://en.wikipedia.org/w/index.php?title=Multivariate_dataset&action=edit&redlink=1) is visualised as a set of coordinates in a high-[dimensional](https://en.wikipedia.org/wiki/Dimension_(metadata)) data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a projection of this object when viewed from its most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.”

As for a data set with much complexity, a certain model such as SVM, LM, GLM can not give a perfect result. Thus we may need to attempt several potential model. Model selection is pretty important. It determines the accuracy of our data analysis. So, we need to compare different methods applied in one data set to choose a better one.

As for a data set with attributes expressed in letter values, there is no doubt that we need to convert letter values to numbers, because the clustering analysis methods don’t accept letter. If not, the clustering methods will not work.

In the clustering algorithms, the comparison is between records, rather than between attributes.

By doing this job, we find that analyzing data is very interesting, and we need to focus, and devote much more energy to master the techniques of big data.