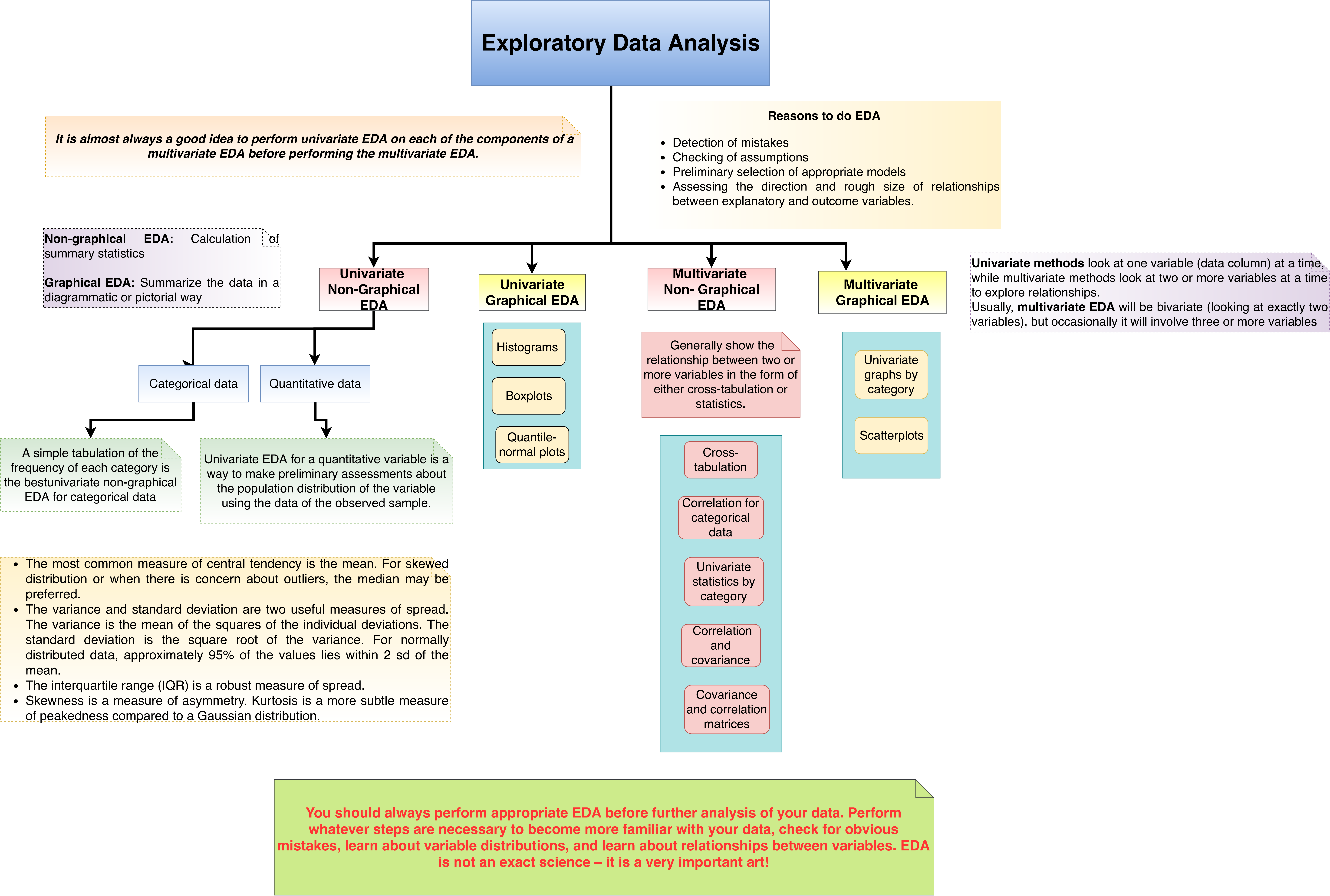

Exploratory Data Analysis

Reasons to do EDA

- Detection of mistakes
- Checking of assumptions
- Preliminary selection of appropriate models
- Assessing the direction and rough size of relationships between explanatory and outcome variables.

It is almost always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.

Non-graphical EDA: Calculation of summary statistics

Graphical EDA: Summarize the data in a diagrammatic or pictorial way

Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships.
Usually, **multivariate EDA** will be bivariate (looking at exactly two variables), but occasionally it will involve three or more variables

Univariate Non-Graphical EDA

Categorical data

A simple tabulation of the frequency of each category is the bestunivariate non-graphical EDA for categorical data

Quantitative data

Univariate EDA for a quantitative variable is a way to make preliminary assessments about the population distribution of the variable using the data of the observed sample.

- The most common measure of central tendency is the mean. For skewed distribution or when there is concern about outliers, the median may be preferred.
- The variance and standard deviation are two useful measures of spread. The variance is the mean of the squares of the individual deviations. The standard deviation is the square root of the variance. For normally distributed data, approximately 95% of the values lies within 2 sd of the mean.
- The interquartile range (IQR) is a robust measure of spread.
- Skewness is a measure of asymmetry. Kurtosis is a more subtle measure of peakedness compared to a Gaussian distribution.

Univariate Graphical EDA

Histograms

Boxplots

Quantile-normal plots

Multivariate Non- Graphical EDA

Generally show the relationship between two or more variables in the form of either cross-tabulation or statistics.

Cross-tabulation

Correlation for categorical data

Univariate statistics by category

Correlation and covariance

Covariance and correlation matrices

Multivariate Graphical EDA

Univariate graphs by category

Scatterplots

You should always perform appropriate EDA before further analysis of your data. Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn about variable distributions, and learn about relationships between variables. EDA is not an exact science – it is a very important art!