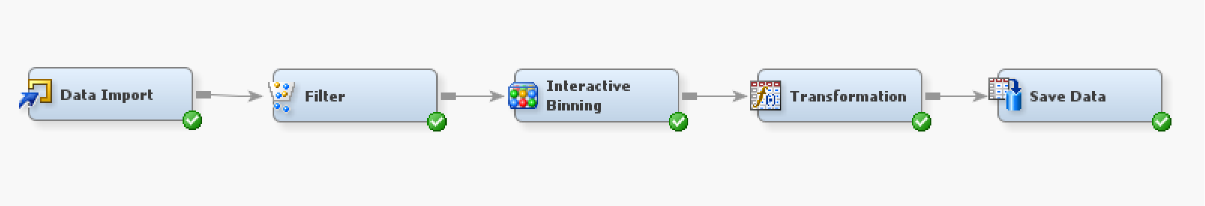
Data Cleaning



Filtering of outlier: The handling of outlier was done using the "Filter" node. Selection was based on the data distribution and the performance testing on a model. The filtering method that generates lowest misclassification rate was selected.

Variable Binning: Only the "DEP_DELAY_NEW" variable can be binned based on Gini Index. The method of binning is bucket. Bucket generates group by dividing the data into evenly spaced intervals based on differences between minimum and maximum values. The variable selection method is based on Gini statistic and the cut off is set at 20. The main purpose of binning is to handle noisy data and smoothing.

Data Transformation: There are two variables that require transformation based on their skewness and kurtosis, which is higher as compared to the other independent variables. "DEST_vsby" and "ORIGIN_vsby" were transformed using Square Root. The selection of the transformation was based on distribution patterns and model testing.