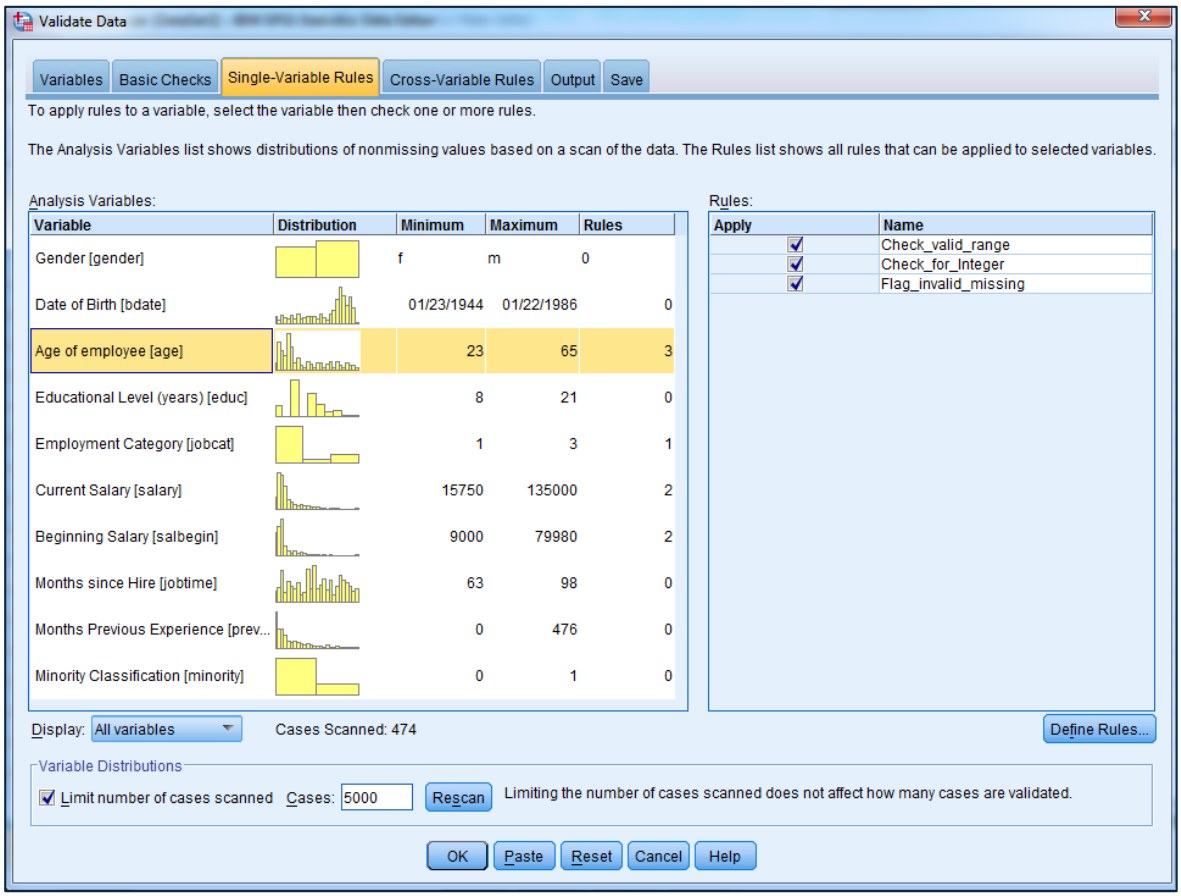
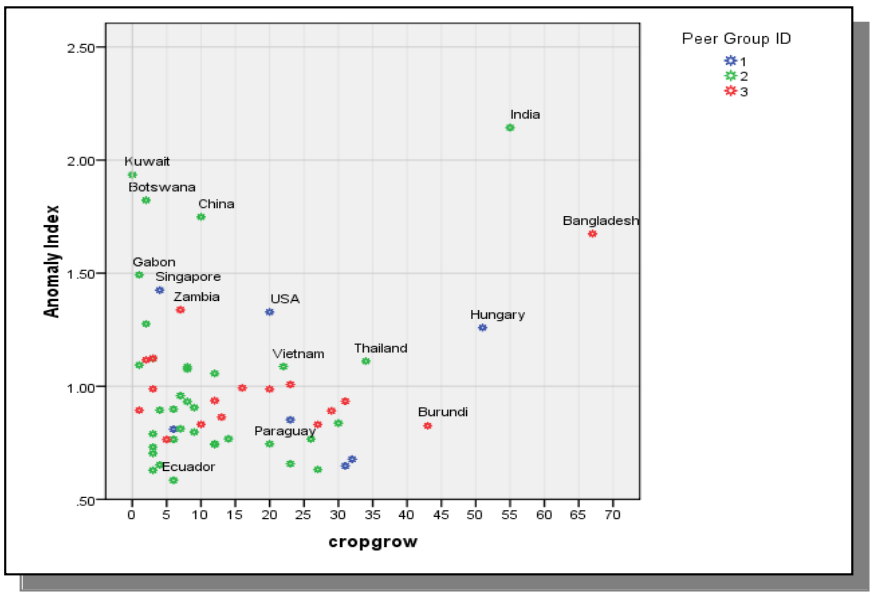
IBM SPSS Data Preparation

- The IBM SPSS Data Preparation module allows users to identify data errors or unusual cases in their datasets. Using a combination of basic checks, validation rules or anomaly detection algorithms, the Data Preparation module will generate new variables or output reports that identify problematic cases or unusual records.
- The *Basic Checks* dialog is designed to identify records with a high percentage of missing values, a high degree of variability or conversely, too little variability as well as incomplete id fields or duplicate records.
- The *Single Variable Rules* dialog provides a graphical overview of each of the fields and the capability to create validation rules for individual fields. An example of this would be a rule that ensures a field is an integer (i.e. no decimal places) such as age.
- The *Cross Variable Rules* dialog provides the capability to create rules that ensure that the values in combinations of variables do not contradict each other or imply errors in the data. An example would be a cross-variable rule that ensures that all car drivers are at least 17 years old.
- IBM SPSS Data Preparation allows these rules to be saved for re-use with new data files. In fact the module contains a number of pre-defined validation rules that check for common errors in datasets.



Screenshot show Validate Data dialog displaying single-variable rules tab.

- Identifying Unusual Cases
  - IBM SPSS Data Preparation provides an *Anomaly Detection* procedure which searches for unusual cases based on deviations from the norms of their cluster groups. The procedure is designed to quickly detect unusual cases for data-auditing purposes in the exploratory data analysis step, prior to any inferential data analysis. This algorithm is designed for generic anomaly detection, but it could be used for applications that are not strictly focussed on data cleaning such as fraud detection.
- Optimal Binning
  - The *Optimal Binning* procedure automatically creates bandings in one or more numeric variables by distributing the values of each variable into bins. The optimal aspect of this procedure is with respect to the procedure's use of a categorical guide variable that "supervises" the binning process.
  - An example of optimal binning could be a situation where a researcher wishes to create a variable containing income bands. Rather than simply manually defining the thresholds of the income groups, the researcher could use optimal binning to automatically create income bands *with respect to* a second variable such as occupational category (i.e. skilled manual, professional etc).



Graphical output from the results of running the Identify Unusual Cases procedure