IBM SPSS Data Preparation

Improve data preparation for more accurate results

Highlights

- Identify suspicious or invalid cases, variables, and data values

- View patterns of missing data

- Summarize variable distributions

- Prepare data for analysis more accurately, quickly

All researchers have to prepare their data prior to analysis. While IBM SPSS Statistics* includes tools for data preparation, sometimes you need more specialized techniques to get your data ready. With IBM® SPSS® Data Preparation*, you can easily identify suspicious or invalid cases, variables and data values; view patterns of missing data; summarize variable distributions; and more accurately work with algorithms designed for nominal attributes. This streamlines the data preparation process – so that you can get ready for analysis faster and reach more accurate conclusions. Choose from a completely automated data preparation procedure for the fastest results, or select from several other methods to help you handle more challenging datasets.

IBM SPSS Data Preparation is available for installation as client-only software but, for greater performance and scalability, a server-based version is also available.

Choose from several options for data preparation
The Validate Data procedure

Data validation has typically been a manual process. You might run a frequency on your data, print the frequencies, circle what needs to be fixed and check for case IDs. This is time consuming and, since every analyst in your organization could use a slightly different method, maintaining consistency from project to project may be a challenge.

To eliminate manual checks, use the Validate Data procedure. This procedure enables you to apply rules to perform data checks based on each variable's measure level (whether categorical or continuous). For example, if you're analyzing survey data that has variables on a five-point Likert scale, use the Validate Data procedure to apply a rule for five-point scales and flag all cases that have values outside of the 1-5 range. You can receive reports of invalid cases as well as summaries of rule violations and the number of cases affected, as well as specify validation rules for individual variables (such as range checks) and cross-variable checks (for example, "pregnant males").

* IBM SPSS Statistics and IBM SPSS Data Preparation were formerly called PASW® Statistics and PASW® Data Preparation.

This knowledge can help you determine data validity and remove or correct suspicious cases at your discretion prior to analysis.

Prepare data in a single step, automatically

Manual data preparation is a complex process that can account for as much as 40 to 90 percent of an analyst's time on a given project. When you need results quickly, the Automated Data Preparation (ADP) procedure helps you detect and correct quality errors and impute missing values in one efficient step. The ADP feature provides an easy-to-understand report with complete recommendations and visualizations to help you determine which data to use in your analysis.
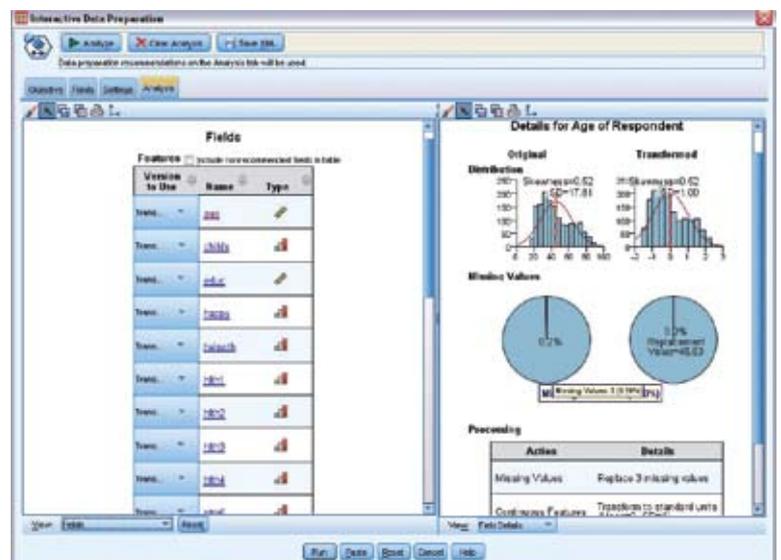
The Automated Data Preparation feature delivers recommendations and allows users to drill in and examine the recommendations.

The Anomaly Detection procedure

Prevent outliers from skewing analyses by using the Anomaly Detection procedure, which searches for unusual cases based upon deviations from similar cases and gives reasons for such deviations. You can flag outliers by creating a new variable. And, once you have identified unusual cases, you can further examine them and determine if they should be included in your analyses.

Optimal Binning

In order to use algorithms that are designed for nominal attributes (such as Naïve Bayes and logit models), you must bin your scale variables before model building. If scale variables aren't binned, algorithms such as multinominal logistic regression will take an extremely long time to process, or they might not converge, especially if you have a large dataset. In addition, the results you receive may be difficult to read or interpret.

Optimal Binning, however, enables you to determine cutpoints to help you reach the best possible outcome for algorithms designed for nominal attributes.

With this procedure, you can select from three types of binning for preprocessing data prior to model building:

- **Unsupervised** – Create bins with equal counts
- **Supervised** – Take the target variable into account to determine cutpoints. This method is more accurate than unsupervised; however, it is also more computationally intensive.
- **Hybrid approach** – Combines the unsupervised and supervised approaches. This method is particularly useful if you have a large amount of distinct values.

To share and re-use assets efficiently, protect them in ways that meet internal and external compliance requirements, and publish results so that a greater number of business users can view and interact with them, consider augmenting your IBM SPSS Statistics software with IBM® SPSS® Collaboration and Deployment Services*. More information about these valuable capabilities can be found at **www.spss.com/software/deployment/cds.**

Our suite of statistical software is now available in three editions: IBM SPSS Statistics Standard, IBM SPSS Statistics Professional and IBM SPSS Statistics Premium. By grouping essential capabilities, these editions provide an efficient way to ensure that your entire team or department has the features and functionality they need to perform the analyses that contribute to your organization's success.

* *IBM SPSS Collaboration and Deployment Services was formerly called PASW® Collaboration and Deployment Services.*

Features

Automated Data Preparation

Recommend steps to speed up model building and improve predictive power:

- Determine Objective: Balance speed and accuracy, Optimize for speed, Optimize for accuracy, or Customize analysis
- Prepare dates and times for modeling:
    - Compute elapsed time until a reference date
    - Compute elapsed time until a reference time
    - Extract cyclical time elements
- Exclude low-quality input fields:
    - Exclude fields with too many missing values
    - Exclude nominal fields with too many unique categories
    - Exclude categorical fields with too many values in a single category
- Adjust measurement levels:
    - Adjust measurement levels of numeric fields
- Prepare fields to improve data quality:
    - Outlier handling
    - Replace missing values
    - Reorder nominal fields
- Rescale Fields:
    - Analysis weight
- Continuous input fields Continuous target fields
- Transform Fields:
    - Using both categorical and/or continuous input fields
- Perform feature selection and construction
- Name fields:
    - Transformed and constructed fields
    - Computed durations
    - Extracted cyclical time elements
- Apply transformations to data

Validate data

Use the Validate Data procedure to validate data in the working data file: Basic checks: Specify basic checks to apply to variables and cases in your file.

For example, obtain reports that identify variables with a high percentage of missing values or empty cases:

- Maximum percentage of missing values
- Maximum percentage of cases in a single category
- Maximum percentage of cases with a count of 1
- Minimum coefficient of variation
- Minimum standard deviation
- Flag incomplete IDs
- Flag duplicate IDs
- Flag empty cases
- Standard rules: Describe the data, view single variable rules and apply them to analysis variables:
    - Description of data:
        - Distribution: Shows a thumbnail-size bar chart for categorical variables or histogram for scale variables
        - Minimum and maximum data values are shown
    - Single-variable rules:
        - Apply rules to individual variables to identify missing or invalid values, such as values outside a valid range
        - User-defined single-variable rules are also possible
- Custom rules: Define cross-variable rule expressions in which respondents'answers violate logic ("pregnant males,"for example)
- Output: Reports describing invalid data:
    - Casewise report, which lists the validation rule violations by case:
        - Specify the minimum number of violations needed for a case to be included in the report
        - Specify the maximum number of cases in the report
    - Standard validation rules reports:

- Summarize violations by analysis variable
- Summarize violations by rule
- Display descriptive statistics
- Save: Enables you to save variables that record rule violations and use them to help clean data and filter out bad cases:
    - Summary variables:
        - Empty case indicator
        - Duplicate ID indicator
        - Incomplete ID indicator
        - Validation rule violation (total count)
    - Indicator variables that record all validation rule violations

Identify unusual cases

The Anomaly Detection procedure searches for unusual cases, based upon deviations from their peer group, and gives reasons for such deviations:

- Specify variables to be used by the procedure with the VARIABLES subcommand. Specify categorical, continuous, and ID variables (to identify cases), and list variables that are excluded from the analysis.
- The HANDLEMISSING subcommand specifies the methods of handling missing values in this procedure:
    - Apply missing value handling. If this option is selected, grand means are substituted for missing values of continuous variables, and missing categories of categorical variables are combined and treated as a valid category. The processed variables are then used in the analysis. If this option is not selected, cases with missing values are excluded from the analysis.

– Create an additional Missing Proportion Variable and use it in the analysis. If chosen, an additional variable called the Missing Proportion Variable that represents the proportion of missing variables in each record is created, and this variable is used in the analysis. If it is not chosen, the Missing Proportion Variable is not created.

- The CRITERIA subcommand specifies the following settings:
  – Minimum and maximum number of peer groups
  – Adjustment weight on the measurement level
  – Number of reasons in the anomaly list
  – Percentage of cases considered as anomalies and included in the anomaly list
  – Number of cases considered as anomalies and included in the anomaly list
  – Cutpoint of the anomaly index to determine whether a case is considered as an anomaly
- Save additional variables to the working data file with the SAVE subcommand:
  – Anomaly index
  – Peer group ID
  – Peer group size
  – Peer group size in percentage
  – The variable, associated with a reason
  – The variable impact measure, associated with a reason
  – The variable value, associated with a reason
  – The norm value, associated with a reason
- Write the model to a specified filename as XML with the OUTFILE subcommand
- Control the display of the output results with the PRINT subcommand

You can print:
  – Case-processing summary
  – The anomaly index list, the anomaly peer ID list and the anomaly reason list
  – The Continuous Variable Norms table, if any continuous variable is used in the analysis, and the Categorical
  – Variable Norms, if any categorical variable is used in the analysis
  – Anomaly Index Summary
  – Reason Summary Table for each reason:
    ◦ Suppress all displayed output except the notes table and any warnings

Optimal Binning

Preprocess data using Optimal Binning, which categorizes one or more continuous variables by distributing the values of each variable into bins. This procedure is useful for reducing the number of values in the given binning input variables, which can greatly improve the performance of algorithms. When using certain Optimal Binning methods, a guide variable helps you determine the cutpoints, thereby maximizing the relationship between the guide variable and the binned variable.

- Select from the following methods:
  – Unsupervised binning via the equal frequency algorithm. This method uses the equal frequency algorithm to discretize the binning input variables. A guide variable is not required.
  – Supervised binning via the MDLP (Minimal Description Length Principle) algorithm. This method discretizes the binning input variables using the MDLP algorithm without any preprocessing. It is suitable for datasets with a small number of cases. A guide variable is required.

  – Hybrid MDLP binning. This involves preprocessing via the equal frequency algorithm, followed by the MDLP algorithm. This method is suitable for datasets with a large number of cases. A guide variable is required.
- Specify the following criteria:
  – How to define the minimum cutpoint for each binning input variable
  – How to define the maximum cutpoint for each binning input variable
  – How to define the lower limit of an interval
  – Whether to force merging of sparsely populated bins
  – Whether missing values are handled using listwise or pairwise deletion
- Save the following:
  – New variables containing binned values
  – Syntax to an IBM SPSS Statistics Base syntax file
- Control output results display with the PRINT subcommand. You can print:
  – The binning input variables' cutpoint sets
  – Descriptive information for all binning input variables
  – Model entropy for binned variables

System requirements

Requirements vary according to platform.

About IBM Business Analytics

IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of business intelligence, predictive analytics, financial performance and strategy management, and analytic applications provides clear, immediate and actionable insights into current performance and the ability to predict future outcomes. Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest productivity, confidently automate decisions and deliver better results.

As part of this portfolio, IBM SPSS Predictive Analytics software helps organizations predict future events and proactively act upon that insight to drive better business outcomes. Commercial, government and academic customers worldwide rely on IBM SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. By incorporating IBM SPSS software into their daily operations, organizations become predictive enterprises – able to direct and automate decisions to meet business goals and achieve measurable competitive advantage.

Business Analytics software