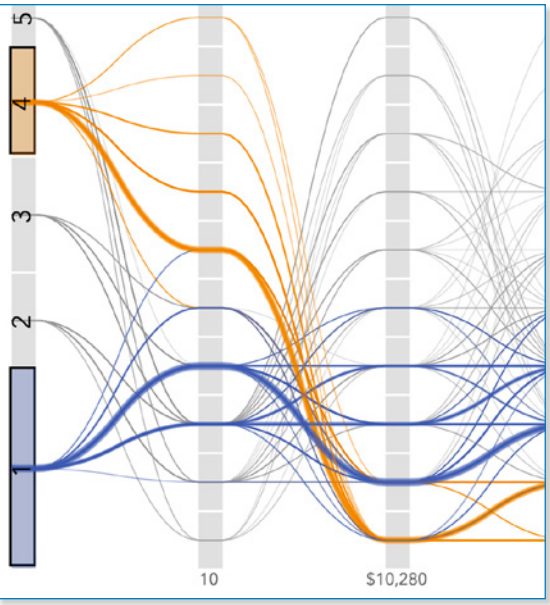
SAS® Visual Statistics on SAS® Viya®

Interactively create, refine and evaluate analytical models in a distributed, open environment for faster, better insights



What does SAS® Visual Statistics on SAS® Viya® do?

It enables you to interactively explore data, and build descriptive and predictive models using either a visual drag-and-drop interface or a programming interface. Data scientists can collaborate with business analysts to refine models for better insights. Distributed, in-memory processing shortens data exploration and model development time.

Why is SAS® Visual Statistics on SAS® Viya® important?

With SAS Visual Statistics, changing market conditions or shifts in customer behavior can be rapidly reflected in analytical models. Fast experimentation and iterative building and refinement of well-qualified models for each segment or group produces valuable insights that can be appropriately acted on. It enhances productivity and collaboration for analytics and business teams, while supporting governance and centralized administration for IT.

For whom is SAS® Visual Statistics on SAS® Viya® designed?

SAS Visual Statistics is primarily designed for use by statisticians, data scientists, programmers and citizen data scientists who need to build, refine and evaluate predictive models to get powerful insights.

Market conditions and customer preferences change fast, and most analytics software packages and architectures can't keep up. They just weren't designed to handle iterative analytical development or on-the-fly changes to predictive models. And technology silos for data preparation, exploration and predictive analytics make it hard to create the numerous and appropriate models needed. Too much time is spent waiting for models on multiple segments to run, and model evaluation is manual and labor intensive.

SAS Visual Statistics solves these issues. As an add-on to SAS Visual Analytics, it combines interactive data exploration with the ability to build and adjust predictive models on the fly. Users can choose a visual or programming interface to build models, depending on their preferences and skills. In-memory, distributed processing addresses complex analytical challenges with the ability to scale for any size problem.

Benefits

- **Surface new opportunities faster to beat your competition.** Data scientists and statisticians can operate on observations at a granular level using the most appropriate analytical modeling technique for the problem they are trying to solve. The result? Unprecedented speed in uncovering insights and finding new ways to increase revenue.
- **Improve the productivity of your analytics staff.** SAS Visual Statistics reduces manual experimentation and improves collaboration. Multiple users can visually interact with data – adding or changing variables, removing outliers, etc. – and instantly see how those changes affect predictive power. If a model does not meet the needs of business or domain experts, your analytics staff can quickly go back and refine it.
- **Dramatically shorten model development time and put better models into action sooner.** It's easy to build and refine models to target specific groups or segments, and run numerous scenarios simultaneously. Analytical professionals can ask more what-if questions and get relevant answers because refined models produce better results. Then put those results into action with automatically generated score code for more timely outcomes.
- **Empower users with the programming language of their choice.** Python, Java, R and Lua programmers can experience the power of SAS Visual Statistics without learning to program in SAS. Give them the flexibility to access trusted and tested SAS machine learning and statistical algorithms from other coding environments.
- **Gain ultimate scalability.** The distributed, in-memory environment means IT can easily scale up and out as data needs grow, numbers of users increase or they tackle more complex problems. Built-in failover management in distributed environments guarantees submitted jobs always finish and computing resources are available whenever needed.

Overview

SAS Visual Statistics provides a visual drag-and-drop interface for quickly creating descriptive and predictive models on data of any size. It also provides an interactive programming interface for those who want to code in SAS, or in other languages while taking advantage of powerful SAS statistical modeling and machine learning techniques. These analytical modeling techniques are used to predict outcomes that result in better, more targeted actions.

SAS Visual Statistics takes advantage of the SAS Viya engine for even faster results. SAS Viya brings new enhancements to the SAS Platform, including high availability, multi-tenancy, faster in-memory processing and native cloud support. Data and analytical workload operations are automatically distributed across the cores of a single server or the nodes of a massive compute cluster, taking advantage of parallel processing. It also provides a single integrated, scalable environment that can be easily managed, maintained and governed.

Key Features

Visual data exploration and discovery (available through SAS® Visual Analytics)

- Quickly interpret complex relationships or key variables that influence modeling outcomes within large data sets.
- Filter observations and understand a variable's level of influence on overall model lift.
- Detect outliers and/or influence points to help you determine, capture and remove them from downstream analysis (e.g., models).
- Explore data using bar charts, histograms, box plots, heat maps, bubble plots, geographic maps and more.
- Derive predictive outputs or segmentations that can be used directly in other modeling or visualization tasks. Outputs can be saved and passed to those without model-building roles and capabilities.

Visual interface access to analytical techniques

- Clustering:
  - K-means, k-modes or k-prototypes clustering.
  - Parallel coordinate plots to interactively evaluate cluster membership.
  - Scatter plots of inputs with cluster profiles overlaid for small data sets and heat maps with cluster profiles overlaid for large data sets.
  - Detailed summary statistics (means of each cluster, number of observations in each cluster, etc.).
  - Generate on-demand cluster ID as a new column.
  - Supports holdout data (training and validation) for model assessment.
- Decision trees:
  - Computes measures of variable importance.
  - Supports classification and regression trees.
  - Based on a modified C4.5 algorithm or cost-complexity pruning.
  - Interactively grow and prune a tree. Interactively train a subtree.
  - Set tree depth, max branch, leaf size, aggressiveness of tree pruning and more.
  - Use tree map displays to interactively navigate the tree structure.
  - Generate on-demand leaf ID, predicted values and residuals as new columns.

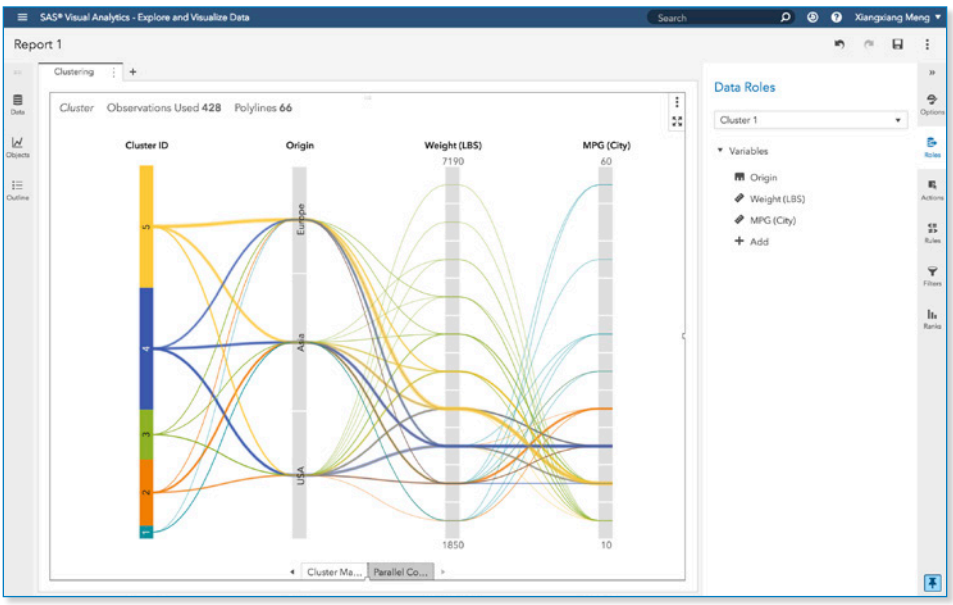


Figure 1: Use the point-and-click visual interface to build and refine clusters.

A visual data exploration and discovery environment

SAS Visual Statistics is an add-on to SAS Visual Analytics, meaning the products share the same interactive data exploration and predictive modeling interface. You get an integrated process for going from data to analytically derived insights.

Both the point-and-click interface and the programming interface let you easily identify predictive drivers among multiple exploratory variables, and visually discover and understand outliers and data discrepancies.

Visual data exploration makes it much easier to understand relationships in your data, derive new variables and select relevant variables to improve your model development efforts. Find out which variables are relevant as inputs to your model and which variables best define your segmentation strategy.

With integrated model building and visual data discovery, you can maintain an uninterrupted workflow, cycling quickly between hypotheses and verification. This will boost your modeling confidence, productivity and accuracy.

GUI-based access to predictive modeling techniques

With the visual web browser interface, it's a simple drag-and-drop process for citizen data scientists to experiment with and create powerful descriptive and predictive models.

And if faced with a backlog of modeling projects or the need to experiment with lots of granular segments, programmers and data scientists could find the visual GUI useful for creating models quickly.

Key Features (continued)

- Supports holdout data (training and validation) for model assessment.
- Supports pruning with holdout data.
- Supports autotuning.

- Logistic regression:
  - Models for binary data with logit and probit link functions.
  - Influence statistics.
  - Supports forward, backward, stepwise and lasso variable selection.
  - Variable selection, including iteration plot.
  - Frequency and weight variables.
  - Residual diagnostics.
  - Summary table includes model dimensions, iteration history, fit statistics, convergence status, Type III tests, parameter estimates and response profile.
  - Generate on-demand predicted labels and predicted event probabilities as new columns. Adjust the prediction cutoff to label an observation as event or nonevent.
  - Supports holdout data (training and validation) for model assessment.

- Linear regression:
  - Influence statistics.
  - Variable selection, including iteration plot.
  - Supports forward, backward, stepwise and lasso variable selection.
  - Frequency and weight variables.
  - Residual diagnostics.
  - Summary table includes overall ANOVA, model dimensions, fit statistics, model ANOVA, Type III test and parameter estimates.
  - Generate on-demand predicted values and residuals as new columns.
  - Supports holdout data (training and validation) for model assessment.

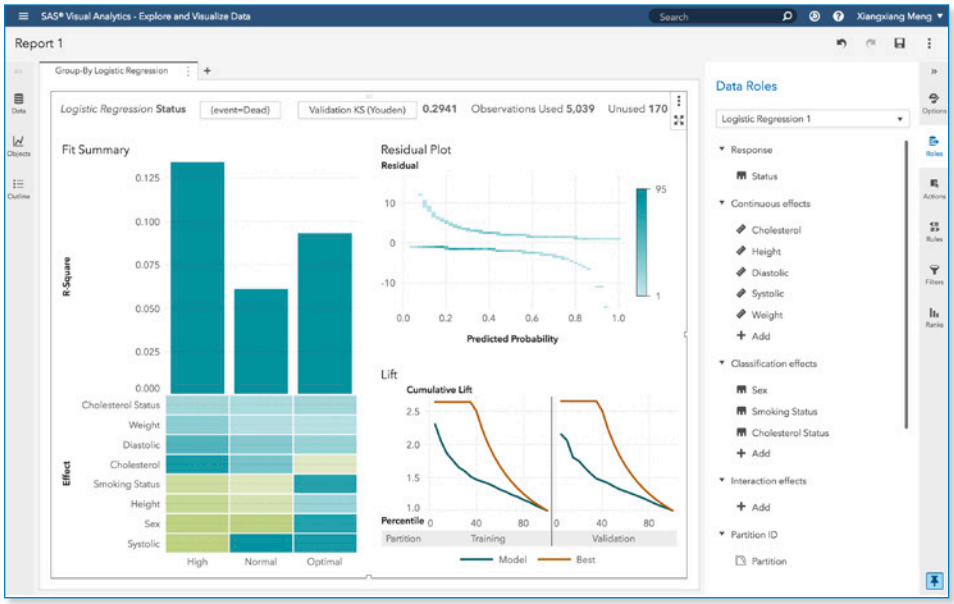


Figure 2: Quickly build and refine logistic regression models using the visual interface. More advanced features are available in the programming interface.

The visual interface provides point-and-click access to:

- Linear regression.
- Logistic regression.
- Generalized linear models.
- Generalized additive models.
- Nonparametric logistic regression.
- Clustering.
- Decision trees.

When working with large and complex data, dimension reduction techniques like clustering and decision trees can improve modeling accuracy. You can explore and evaluate segments for further analysis using k-means clustering, scatter plots and detailed summary statistics. Decision trees can be built for both classification and regression. After creating a decision tree, you can interactively prune trees and train subtrees.

SAS Visual Statistics users can copy their models into Model Studio and continue the data mining or machine learning processes in SAS Visual Data Mining and Machine Learning.

Open, code-based model development

While the visual GUI of SAS Visual Statistics is powerful and appealing, many statisticians, data scientists and quantitative specialists prefer to code their own predictive models and take advantage of more options to fine-tune the models.

Analytical actions running in SAS Visual Statistics can be programmatically accessed from the SAS Studio programming interface, or can be called from other languages such as Python, R, Lua and Java.

Key Features (continued)

- Generalized linear models:
  - Distributions supported include beta, normal, binary, exponential, gamma, geometric, Poisson, Tweedie, inverse Gaussian and negative binomial.
  - Supports forward, backward, stepwise and lasso variable selection.
  - Variable selection, including iteration plot.
  - Offset variable support.
  - Frequency and weight variables.
  - Residual diagnostics.
  - Summary table includes model summary, iteration history, fit statistics, Type III test table and parameter estimates.
  - Informative missing option for treatment of missing values on the predictor variable.
  - Generate on-demand predicted values and residuals as new columns.
  - Supports holdout data (training and validation) for model assessment.
- Generalized additive models:
  - Distributions supported include normal, binary, gamma, Poisson, Tweedie, inverse Gaussian and negative binomial.
  - Supports one- and two-dimensional spline effects.
  - GCV, GACV and UBRE methods for selecting the smoothing effects.
  - Offset variable support.
  - Frequency and weight variables.
  - Residual diagnostics.
  - Summary table includes model summary, iteration history, fit statistics and parameter estimates.
  - Supports holdout data (training and validation) for model assessment.
- Nonparametric logistic regression:
  - Models for binary data with logit, probit, log-log and c-log-log link functions.
  - Supports one- and two-dimensional spline effects.
  - GCV, GACV and UBRE methods for selecting the smoothing effects.
  - Offset variable support.
  - Frequency and weight variables.

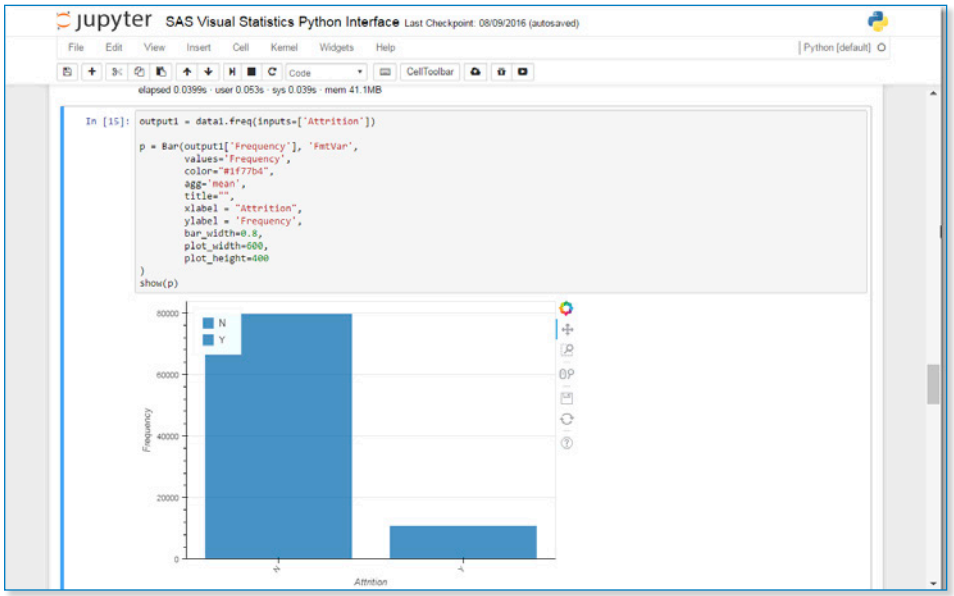


Figure 3: Use open source languages, including Python, to build descriptive and predictive models.

In addition, public REST APIs enable coders to add proven SAS Analytics to existing applications.

This puts the power of SAS in the hands of programmers who may not be familiar with SAS – but know it's the best analytics software available.

SAS Visual Statistics on SAS Viya provides programming access to many tested and proven SAS algorithms running in the distributed in-memory environment, including those for:

- Data manipulation.
- Variable binning.
- Missing value imputation.
- Supervised and unsupervised variable selection.
- Clustering using k-means and k-modes.
- Decision trees.
- Principal component analysis.
- Linear and logistic regression.
- Nonlinear regression.
- Generalized linear regression.
- Ordinary least squares.
- Partial least squares.
- Quantile regression.
- Generalized additive models.
- Proportional hazard regression.
- Statistical process control.
- Descriptive statistics.
- Model assessment.

SAS Visual Statistics on SAS Viya also includes SAS/STAT® procedures and SAS/GRAPH®.

This open programming environment provides flexibility for data scientists and statisticians, based on their programming skills and preferences, to easily access the power of SAS for data manipulation and advanced analytics.

Key Features (continued)

- Residual diagnostics.
- Summary table includes model summary, iteration history, fit statistics and parameter estimates.
- Supports holdout data (training and validation) for model assessment.

Programming access to analytical techniques

- Programmers and data scientists can access SAS Viya (CAS server) from SAS Studio using SAS procedures (PROCs) and other tasks.
- Programmers can execute CAS actions using PROC CAS or use different programming environments like Python, R, Lua and Java.
- Users can also access SAS Viya (CAS server) from their own applications using public REST APIs.
- Provides native integration to Python Pandas DataFrames. Python programmers can upload DataFrames to CAS and fetch results from CAS as DataFrames to interact with other Python packages such as Pandas, matplotlib, Plotly, Bokeh, etc.
- Includes SAS/STAT® procedures and SAS/GRAPH®.
- Principal component analysis (PCA):
  - Performs dimension reduction by computing principal components.
  - Provides the eigenvalue decomposition, NIPALS and ITERGS algorithms.
  - Outputs principal component scores across observations.
  - Creates scree plots and pattern profile plots.
- Decision trees:
  - Supports classification trees and regression trees.
  - Supports categorical and numerical features.
  - Provides criteria for splitting nodes based on measures of impurity and statistical tests.
  - Provides the cost-complexity and reduced-error methods of pruning trees.
  - Supports partitioning of data into training, validation and testing roles.
  - Supports use of validation data for selecting the best subtree.
  - Supports the use of test data for assessment of final tree model.
  - Provides various methods of handling missing values, including surrogate rules.
  - Creates tree diagrams.
  - Provides statistics for assessing model fit, including model-based (resubstitution) statistics.
  - Computes measures of variable importance.
  - Outputs leaf assignments and predicted values for observations.
- Clustering:
  - Provides the k-means algorithm for clustering continuous (interval) variables.
  - Provides the k-modes algorithm for clustering nominal variables.
  - Provides various distance measures for similarity.
  - Provides the aligned box criterion method for estimating the number of clusters.
  - Outputs cluster membership and distance measures across observations.
- Linear regression:
  - Supports linear models with continuous and classification variables.
  - Supports various parameterizations for classification effects.
  - Supports any degree of interaction and nested effects.
  - Supports polynomial and spline effects.
  - Supports forward, backward, stepwise, least angle regression and lasso selection methods.

Dynamic group-by processing

With SAS Visual Statistics, many users can concurrently build numerous models and process results for each group or segment without having to sort or index data each time. The grouping variables, or their properties, can change from one action to the next, and groups are processed without shuffling or reordering the data.

This means more models can be quickly created for more segments or groups on the fly without additional processing overhead. The result? Models that meet the unique needs of individual segments or groups.

Model comparison and assessment

After models have been created, they can be easily compared and assessed using a variety of statistical comparison summaries such as lift charts, ROC charts, concordance statistics and misclassification tables on one or more models from either the visual or programming interface.

And, from the visual interface, an interactive slider lets you manipulate cutoff thresholds so you can easily and visually evaluate lift at different percentiles. Combine model fitting with model diagnostics to quickly see and understand impacts on performance.

Model assessment features let you compare models to identify the ones that produce the best lift and ROI.

Key Features (continued)

- Supports information criteria and validation methods for controlling model selection.
- Offers selection of individual levels of classification effects.
- Preserves hierarchy among effects.
- Supports partitioning of data into training, validation and testing roles.
- Provides a variety of diagnostic statistics.
- Generates SAS code for production scoring.

- Logistic regression:
  - Supports binary and binomial responses.
  - Supports various parameterizations for classification effects.
  - Supports any degree of interaction and nested effects.
  - Supports polynomial and spline effects.
  - Supports forward, backward, fast backward and lasso selection methods.
  - Supports information criteria and validation methods for controlling model selection.
  - Offers selection of individual levels of classification effects.
  - Preserves hierarchy among effects.
  - Supports partitioning of data into training, validation and testing roles.
  - Provides variety of statistics for model assessment.
  - Provides variety of optimization methods for maximum likelihood estimation.
- Generalized linear models:
  - Supports responses with variety of distributions, including binary, normal, Poisson and gamma.
  - Supports various parameterizations for classification effects.
  - Supports any degree of interaction and nested effects.
  - Supports polynomial and spline effects.
  - Supports forward, backward, fast backward, stepwise and group lasso selection methods.
  - Supports information criteria and validation methods for controlling model selection.
  - Offers selection of individual levels of classification effects.
  - Preserves hierarchy among effects.

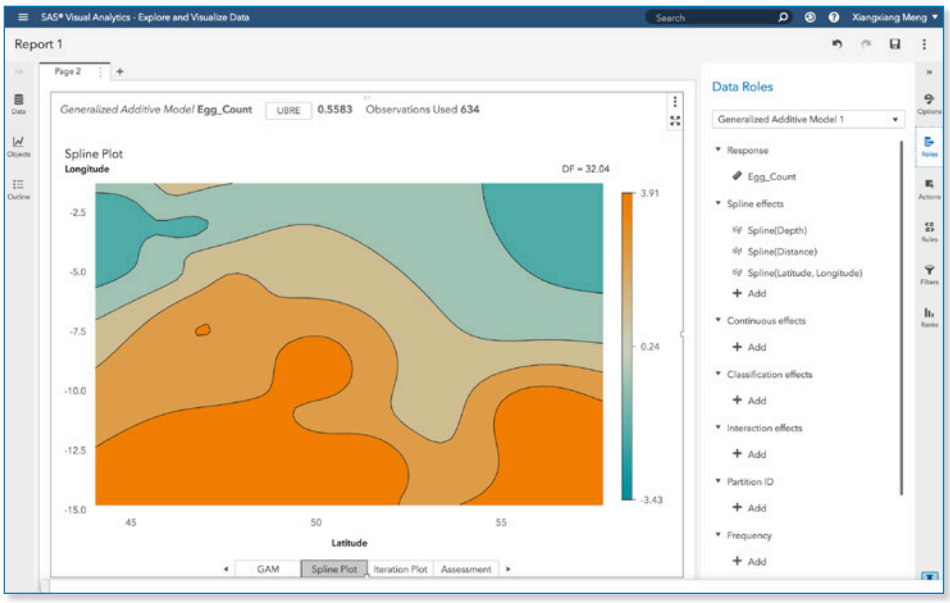


Figure 4: Generalized additive models have been added to SAS Visual Statistics. Build them using the visual interface or with code.

Model scoring

After you have determined which model performs best, your champion model can be easily applied against new data.

Moving all of the data preparation tasks and algorithms of a sophisticated model from a development environment to an operational system is usually one of the most difficult aspects of predictive modeling and machine learning.

With SAS Visual Statistics, you can export your models as SAS DATA step code and easily apply them to new data. Putting your predictive models into production produces the insights needed for making better decisions and taking optimal actions.

Distributed, in-memory analytical processing with SAS® Viya®

SAS Visual Statistics runs on SAS Viya, a new high-performance runtime engine that uses in-memory analytical processing to provide answers to a range of business questions in a single, scalable and governed environment. It takes advantage of distributed, parallel processing for blazingly fast speed, dramatically reducing data exploration and model development time.

SAS Viya delivers high availability and scalable processing so IT can scale computing capacity up and out to meet the needs of more users who are dealing with more data and increasingly complex analytical problems.

Key Features (continued)

- Supports partitioning of data into training, validation and testing roles.
- Provides variety of statistics for model assessment.
- Provides a variety of optimization methods for maximum likelihood estimation.

- Nonlinear regression models:
  - Fits nonlinear regression models with standard or general distributions.
  - Computes analytical derivatives of user-provided expressions for more robust parameter estimations.
  - Evaluates user-provided expressions using the ESTIMATE and PREDICT statements (procedure only).
  - Requires a data table that contains the CMP item store if not using PROC NLMOD.
  - Estimates parameters using the least squares method.
  - Estimates parameters using the maximum likelihood method.
- Quantile regression models:
  - Supports quantile regression for single or multiple quantile levels.
  - Supports multiple parameterizations for classification effects.
  - Supports any degree of interactions (crossed effects) and nested effects.
  - Supports hierarchical model selection strategy among effects.
  - Provides multiple effect-selection methods.
  - Provides effect selection based on a variety of selection criteria.
  - Supports stopping and selection rules.
- Predictive partial least squares models:
  - Provides programming syntax with classification variables, continuous variables, interactions and nestings.
  - Provides effect-construction syntax for polynomial and spline effects.
  - Supports partitioning of data into training and testing roles.
  - Provides test set validation to choose the number of extracted factors.
  - Implements the following methods: principal component regression, reduced rank regression and partial least squares regression.

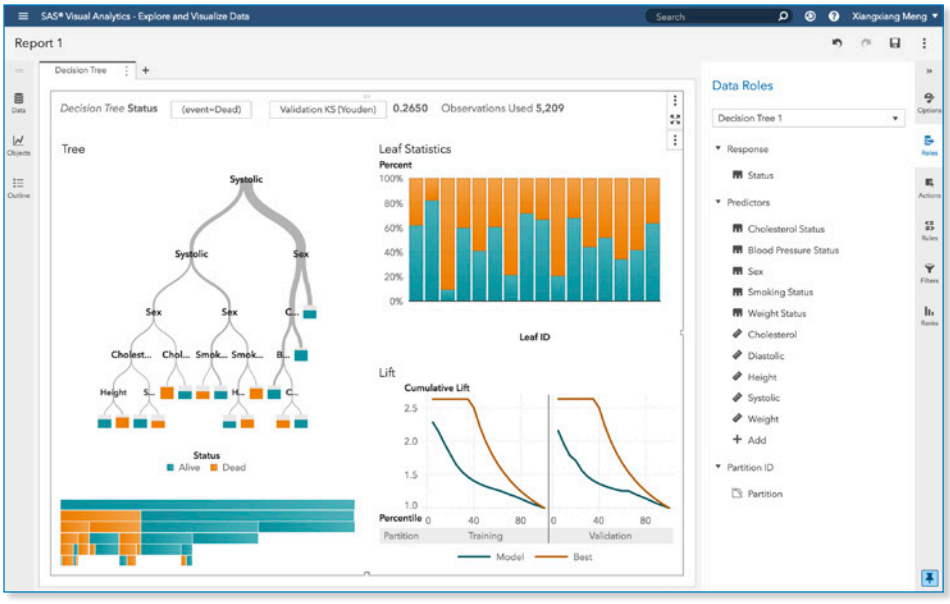


Figure 5: Decision trees illustrate the most likely outcomes and consequences. Build them using the visual interface or with code.

You need analytical processing power you can count on. The fault-tolerant design of SAS Viya automatically detects server failure, even in multiplatform processing environments, and redistributes processing as needed. It also manages several copies of data on the processing cluster. If a node in the cluster becomes unavailable or fails, the required data is retrieved from another block to quickly continue processing.

These self-healing mechanisms ensure high availability for uninterrupted processing and automated recovery. Support for multitenancy allows multiple tenants to share resources, even though each tenant is logically isolated.

Concurrent access to data in memory

SAS Viya enables you to load and persist data in memory on demand, and execute multipass analytical computations. All data, tables and objects are held in memory as long as required, whether it's for interactive visual investigations or advanced analytical processing. Many users can collaborate to explore the same raw data and build models simultaneously.

Key Features (continued)

- Generalized additive models:
  - Fit generalized additive models based on low-rank regression splines.
  - Estimates the regression parameters by using penalized likelihood estimation.
  - Estimates the smoothing parameters by using either the performance iteration method or the outer iteration method.
  - Estimates the regression parameters by using maximum likelihood techniques.
  - Tests the total contribution of each spline term based on the Wald statistic.
  - Provides model-building syntax that can include classification variables, continuous variables, interactions and nestings.
  - Enables you to construct a spline term by using multiple variables.
- Proportional hazard regression:
  - Fit the Cox proportional hazards regression model to survival data and perform variable selection.
  - Provides model-building syntax with classification variables, continuous variables, interactions and nestings.
  - Provides effect-construction syntax for polynomial and spline effects.
  - Performs maximum partial likelihood estimation, stratified analysis and variable selection.
  - Partitions data into training, validation and testing roles.
  - Provides weighted analysis and grouped analysis.
- Statistical process control:
  - Perform Shewhart control chart analysis.
  - Analyze multiple process variables to identify processes that are out of statistical control.
  - Adjust control limits to compensate for unequal subgroup sizes.
  - Estimate control limits from the data, compute control limits from specified values for population parameters (known standards) or read limits from an input data table.
  - Perform tests for special causes based on runs patterns (Western Electric rules).
  - Estimate the process standard deviation using various methods (variable charts only).
  - Save chart statistics and control limits in output data tables.

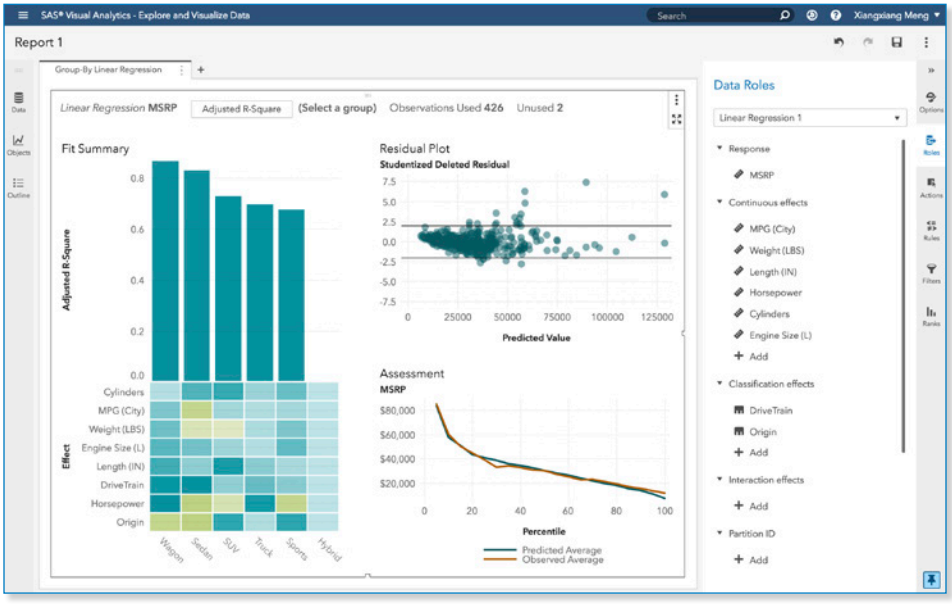


Figure 6: Dynamic group-by processing computes results for each group, partition or segment without having to sort or index data each time.

In-memory processing eliminates unnecessary and expensive data shuffling during iterative steps or requests. This means that the results from changes to models (e.g., adding new variables or removing outliers) are instantly visible.

Flexible deployment options

SAS Visual Statistics offers deployment options for organizations with any size data, different workloads and varying performance requirements. Deployments can scale from a single-machine environment for departmental workgroups or small to midsize businesses up to large distributed systems with hundreds of nodes in a cluster for large organizations.

You can deploy SAS Visual Statistics wherever makes the most sense for your organization:

- On-site:
  - Single-machine mode to support the needs of small to midsize organizations.
  - Distributed mode to meet growing data, workload and scalability requirements.

- Cloud deployments:
  - Enterprise hosting.
  - In a private cloud via technologies such as Cloud Foundry platform as a service (Paas) to support multiple cloud providers.
  - In public clouds, including Amazon Web Services and Microsoft Azure.
  - You can also access this software via the predeployed and preconfigured managed software-as-a-service offerings provided by SAS.

Key Features (continued)

- Independent component analysis:
  - Extracts independent components (factors) from multivariate data.
  - Maximizes non-Gaussianity of the estimated components.
  - Supports whitening and dimension reduction.
  - Produces an output data table that contains independent components and whitened variables.
  - Implements symmetric decorrelation, which calculates all the independent components simultaneously.
  - Implements deflationary decorrelation, which extracts the independent components successively.

- Linear mixed models:
  - Supports many covariance structures, including variance components, compound symmetry, unstructured, AR(1), Toeplitz, factor analytics, etc.
  - Provides specialized dense and sparse matrix algorithms.
  - Supports REML and ML estimation methods, which are implemented with a variety of optimization algorithms.
  - Provides inference features, including standard errors and t tests for fixed and random effects.

- Model-based clustering:
  - Models the observations by using a mixture of multivariate Gaussian distributions.
  - Allows for a noise component and automatic model selection.
  - Provides posterior scoring and graphical interpretation of results.

Descriptive statistics

- Distinct counts to understand cardinality.
- Box plots to evaluate centrality and spread, including outliers for one or more variables.
- Correlations to measure the Pearson's correlation coefficient for a set of variables. Supports grouped and weighted analysis.
- Cross-tabulations, including support for weights.

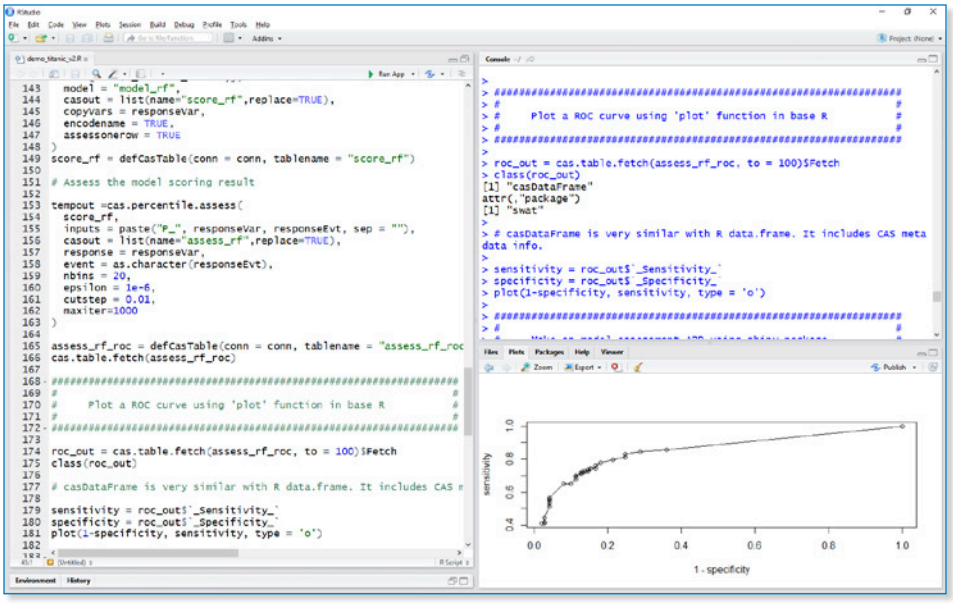


Figure 7: You can call analytical procedures and actions in SAS Visual Statistics using R, as well as other open source programming languages.

TO LEARN MORE »

To learn more about SAS Visual Statistics, download white papers, view screenshots and see other related material, please visit sas.com/visualstatistics.

Key Features (continued)

- Contingency tables, including measures of associations.
- Histograms with options to control binning values, maximum value thresholds, outliers and more.
- Multidimensional summaries in a single pass of the data.
- Percentiles for one or more variables.
- Summary statistics such as number of observations, number of missing values, sum of nonmissing values, mean, standard deviation, standard errors, corrected and uncorrected sums of squares, min and max, and the coefficient of variation.
- Kernel density estimates using normal, tri-cube and quadratic kernel functions.
- Constructs one-way to *n*-way frequency and cross-tabulation tables.

Group-by processing

- Build models, compute and process results on the fly for each group or segment without having to sort or index the data each time.
- Build segment-based models instantly (i.e., stratified modeling) from a decision tree or clustering analysis.

Model comparison, assessment and scoring

- Generate model comparison summaries such as lift charts, ROC charts, concordance statistics and misclassification tables for one or more models.
- Interactively slide the prediction cutoff for automatic updating of assessment statistics and classification tables.
- Interactively evaluate lift at different percentiles.
- Export models as SAS DATA step code to integrate models with other applications. Score code is automatically concatenated if a model uses derived outputs from other models (e.g., leaf ID, cluster ID, etc.).

SAS® Viya® in-memory runtime engine

- An in-memory server called CAS (SAS Cloud Analytic Services) performs processing in memory and distributes processing across nodes in a cluster.
- User requests (expressed in a procedural language) are translated into actions with necessary parameters to process in a distributed environment. The result set and messages are passed back to the procedure for further action by the user.
- Data is managed in blocks and loaded in memory on demand. If tables exceed the memory capacity, the server caches the blocks on disk. Data and intermediate results are held in memory as long as required, across jobs and user boundaries.
- An algorithm determines the optimal number of nodes for a given job.
- Communication layer supports fault tolerance. Remove or add nodes from or to a server while it is running. All components in the architecture can be replicated to support high availability.
- Products can be deployed in multitenant mode, allowing for a shared software stack to support securely isolated tenants.