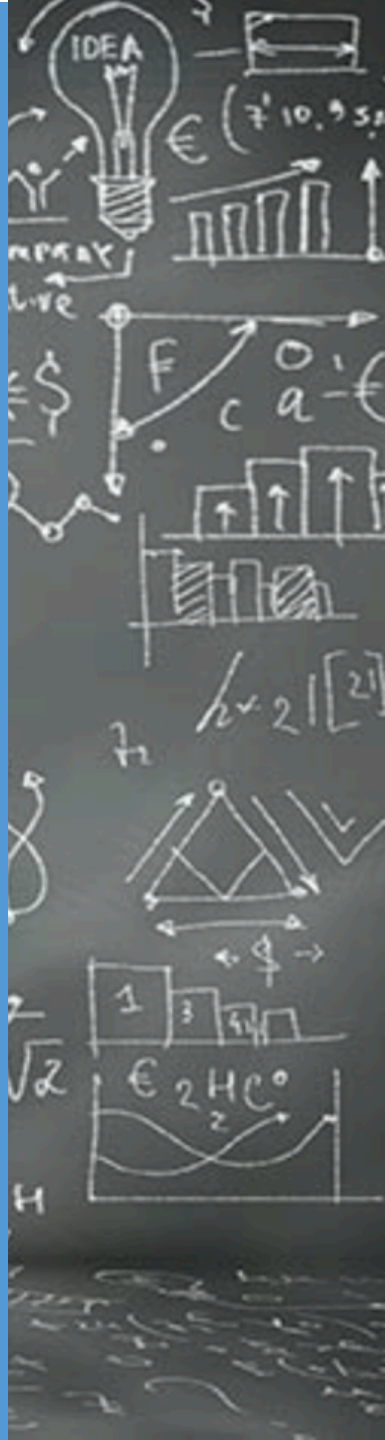
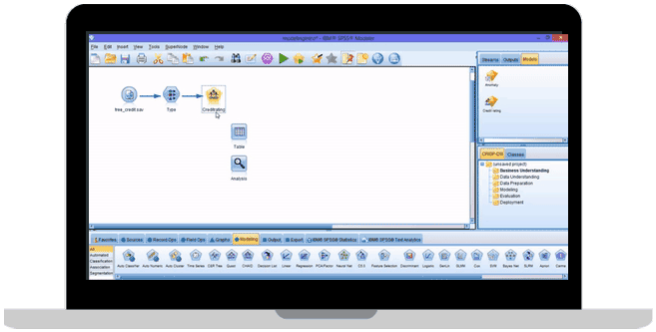
EDHEC BUSINESS SCHOOL

Consumer Insights & Big Data
Descriptive Statistics with IBM Statistics



Yann Gouedo
Data Scientist Leader – Machine Learning / Artificial Intelligence
Marketing / Risk / Fraud / Maintenance
IBM Certified Senior Data Scientist & IBM Certified Senior Architect

IBM SPSS Portfolio

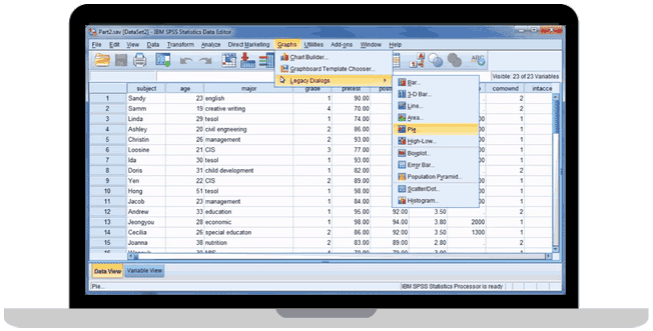


IBM SPSS Modeler

A predictive analytics platform that brings predictive intelligence to decisions made by individuals, groups, systems and the enterprise.

IBM SPSS Statistics

The world's leading statistical software used to solve such business and research problems by means of ad-hoc analysis and hypothesis testing



Statistical Analysis & Data Mining: Feeding Predictive Analytics

Top-Down Approach

- **A statistical approach involves**
  - forming a theory about a possible relationship
  - converting it to a hypothesis
  - testing that hypothesis using statistical methods
- **It is a manual, user-driven, top-down approach to data analysis**

Source: *DM Review*

Bottom-up Approach

- **Data mining involves the interrogation of the data and is performed by the data mining method rather than by the user**
- **It is a data-driven, self-organizing, bottom-up approach to data analysis that works on very large data sets**

"Statistical Modeling: The Two Cultures," Leo Breiman, *Statistical Science*, 2001, Vol.16 (3), pp.199-231.

Note that <u>Both</u> Approaches can Drive Predictive Analytics

IBM SPSS Modeler

OVERVIEW

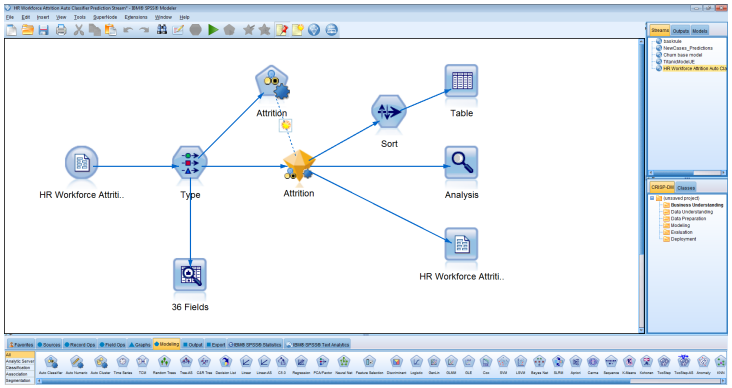
- Comprehensive predictive analytics workbench
- Easy-to-use, interactive interface without the need for programming
- Automated modeling and data preparation capabilities
- Access ALL data – structured and unstructured – from disparate sources
- Providing a range of advanced analytics - text analytics, entity analytics, social network analysis, etc.

CUSTOMER NEEDS

- Improve business outcomes through predictive intelligence
- Deploy predictive models in operational processes to take better decisions at the point of impact

TARGET AUDIENCE

- Technical Data Scientist
- Citizen Data Scientist
- Business Analyst



CUSTOMER EXAMPLES

- AB Volvo in Sweden reduces truck diagnostic time by up to 70 percent.
- Autobacs Seven Co. in Japan conducts more targeted promotional campaigns, increasing its conversion rates by more than 20 percent

WHY IBM ?

- Simplicity without sacrifice
- Code optional, open to open source (R, Python, SPARK)
- Deployment at scale

IBM SPSS Statistics

OVERVIEW

- Quickly understand large and complex datasets using advanced statistical procedures ensuring high accuracy to drive quality decision-making
- IBM SPSS Statistics is the world's leading statistical software which has been around for 40+ years

CUSTOMER NEEDS

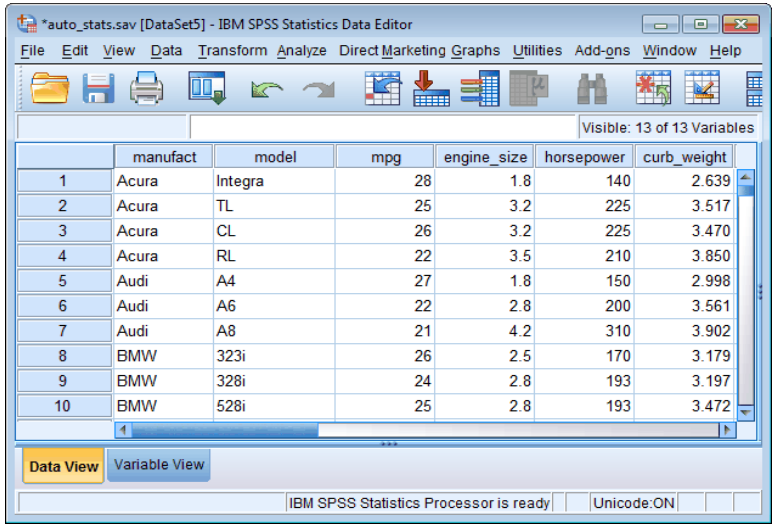
- We help people validate (or disprove!) assumptions faster and efficiently using the right statistical capability, at the right time.
- Organizations use IBM SPSS Statistics to understand data, analyze trends, forecast and plan so they can validate assumptions and drive accurate conclusions.

TARGET AUDIENCE

- Analytic Professional/ Researcher
- Business Analyst

CUSTOMER EXAMPLES

- Meteolytix GmbH builds statistical models that provide daily sales forecasts for the retail and service sectors.
- Bank Alfalah Pakistan - improves the efficiency of credit application processing



	manufact	model	mpg	engine_size	horsepower	curb_weight
1	Acura	Integra	28	1.8	140	2.639
2	Acura	TL	25	3.2	225	3.517
3	Acura	CL	26	3.2	225	3.470
4	Acura	RL	22	3.5	210	3.850
5	Audi	A4	27	1.8	150	2.998
6	Audi	A6	22	2.8	200	3.561
7	Audi	A8	21	4.2	310	3.902
8	BMW	323i	26	2.5	170	3.179
9	BMW	328i	24	2.8	193	3.197
10	BMW	528i	25	2.8	193	3.472

WHY IBM ?

- The world's leading statistical software which has been around for 40+ years
- Menu-driven experience for the beginner
- Command syntax and programmable extensibility for the experienced statistician
- access to external programming (R, Python, Java, .Net)
- Flexible deployment from stand-alone to enterprise support

IBM® SPSS® Statistics ROI
IBM® SPSS® Modeler ROI

- Descriptive Analytics with IBM® SPSS® Statistics helps statisticians to validate statistical hypotheses by using efficient descriptive methods. IBM® SPSS® Statistics will not predict information but will describe data and validate statistical hypothesis.
  - Difficult to evaluate financial results of IBM SPSS Statistical because it will not allow statistics results to be integrated to operational IT (call center, campaigns, website…).

- Predictive Analytics with IBM® SPSS® Modeler helps connect data to effective action by drawing reliable conclusions about current conditions and future events. Predictive Analytics allows organisation to retain customers, increase customers portfolio, increase Marketing campaign effectiveness, reduce churn rate et fraud….
  - Some examples of financial results:
    - 35% reduction in mailing cost, 2X response rate, 29% more profit - FTBO
    - Reduced churn from 19 to 2% - Cablecom
    - 100% increase in campaign effectiveness - BT
    - 30 Million Euro in new revenue - AEGON

End users

- **Typical IBM® SPSS® Statistics users**
  - Usually have academic training
  - Need to perform hypothesis tests
  - Process Oriented – More likely to write Syntax and discuss routine statistical procedures
  - Focused on setting up analyses from beginning to end (methodology)
  - Primary concern – utilizing selected algorithm for analysis
  - Deep Diver – Often trained in a specific area of analysis
  - Hard-core analyst or researcher
  - A technical understanding of data
  - Preferred interface: Spreadsheet
- **Typical IBM® SPSS® Modeler users**
  - Usually have subject matter expertise such as a business analyst
  - Need to build models but don't necessarily understand the analytics
  - Graphically Oriented – Prefers to use graphical, point and click tools
  - Focused on uncovering quick, accurate and actionable insights
  - Primary concern – explore multiple analytical approaches using automated techniques for best results
  - Data Miner– Extract data from multiple data source, merge and derive new data for analysis
  - Empower Business User – Deep understanding of business issues and drivers
  - A business understanding of data
  - Preferred interface: Graphical Reporting

Editions IBM® SPSS® Statistics

Statistics Standard Bundle

- IBM SPSS Statistics Base
- IBM SPSS Advanced Statistics
- IBM SPSS Custom Tables
- IBM SPSS Regression

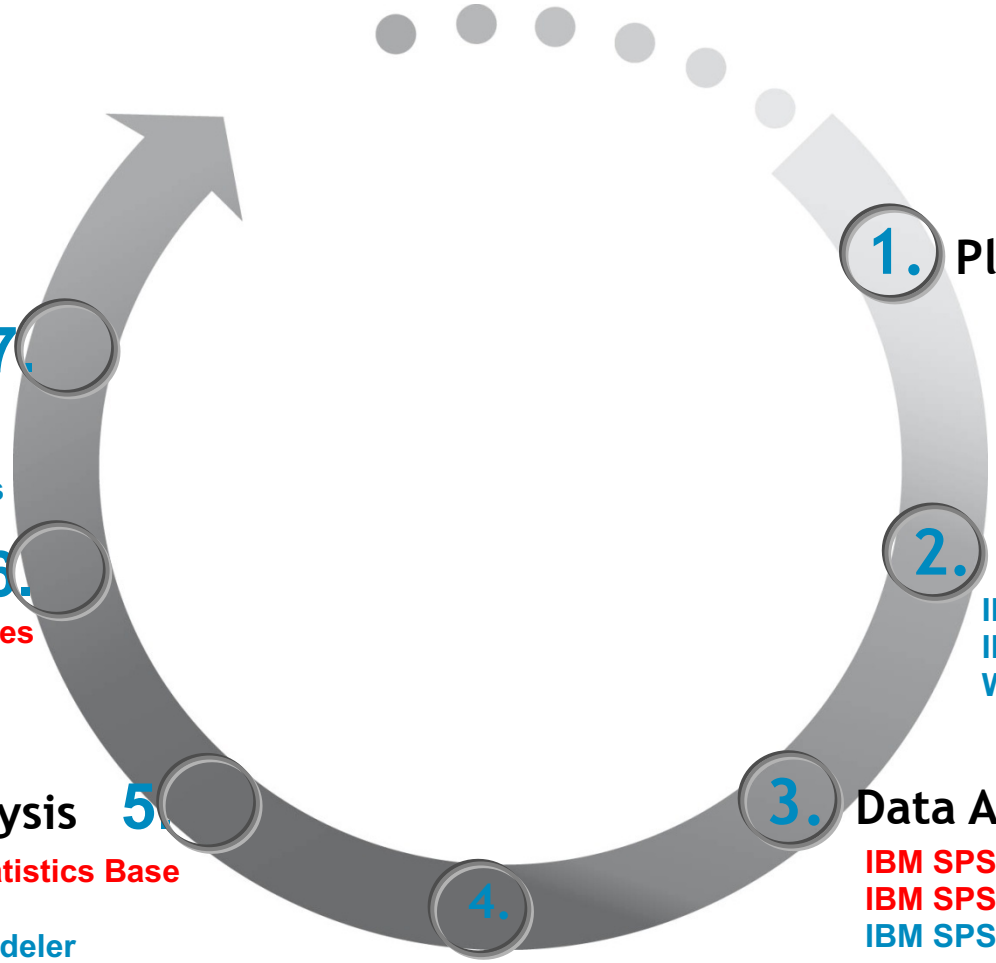
Statistics Professional Bundle

- IBM SPSS Statistics Base
- IBM SPSS Advanced Statistics
- IBM SPSS Custom Tables
- IBM SPSS Regression
- **IBM SPSS Data Preparation**
- **IBM SPSS Missing Values**
- **IBM SPSS Categories**
- **IBM SPSS Decision Trees**
- **IBM SPSS Forecasting**

Statistics Premium Bundle

- IBM SPSS Statistics Base
- IBM SPSS Advanced Statistics
- IBM SPSS Custom Tables
- IBM SPSS Regression
- IBM SPSS Data Preparation
- IBM SPSS Categories
- IBM SPSS Decision Trees
- IBM SPSS Forecasting
- **IBM SPSS Bootstrapping**
- **IBM SPSS Conjoint**
- **IBM SPSS Exact Tests**
- **IBM SPSS Neural Networks**
- **IBM SPSS Direct Marketing**
- **IBM Complex Samples**
- **IBM SPSS Viz Designer**
- **IBM SPSS Amos**
- **IBM SPSS SamplePower**

IBM SPSS Statistics and the Analytical Process



1. Planning

IBM SPSS SamplePower
IBM SPSS Conjoint
IBM SPSS Complex Samples

2. Data Collection

IBM SPSS Data Collection Data Entry
IBM SPSS Data Collection Interviewer Web

3. Data Access

IBM SPSS Missing Values
IBM SPSS Data Preparation
IBM SPSS Text Analytics for Surveys
IBM SPSS Modeler

4. Data Management and Preparation

IBM SPSS Statistics Base and Modules
IBM SPSS Modeler

5. Data Analysis

IBM SPSS Statistics Base and Modules
IBM SPSS Modeler

6. Reporting

IBM SPSS Custom Tables

7. Deployment

IBM SPSS Collaboration and Deployment Services

IBM® SPSS® Statistics vs IBM® SPSS® Modeler

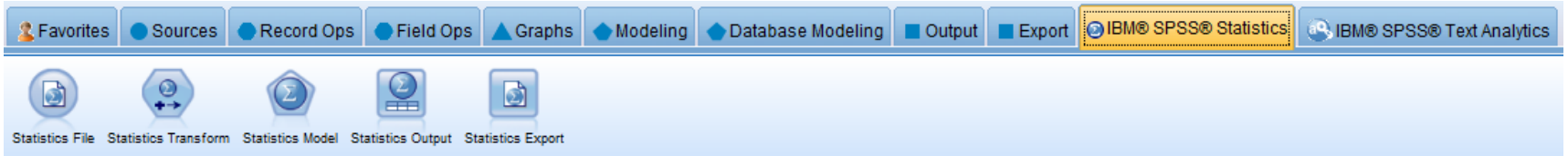
	IBM® SPSS® Statistics	IBM® SPSS® Modeler	IBM® SPSS® Modeler + IBM® SPSS® Statistics
Data Access	✓	✓	✓
Big Data	X	✓	✓
Design of experiments	✓	X	✓
Principal Component Analysis	✓	X	✓
Exacts Tests	✓	X	✓
Non parametric tests	✓	X	✓
Multiple imputations of missing values	✓	X	✓
Non linear regressions	✓	X	✓
Advanced regressison models	✓	X	✓
Cross tab reports	✓	X	✓
Data mining	X	✓	✓
Text mining	X	✓	✓
Use of SPSS Statistics syntax files	✓	X	✓
Easy of use	X	✓	✓
Performance	X	✓	✓

Summary
IBM® SPSS® Statistics vs IBM® SPSS® Modeler

	Statistics	Predictive Analytics
	IBM® SPSS® Statistics	IBM® SPSS® Modeler
Structure	Structured	Structured & Unstructured
Size	Small	Large, BigData
Generation	Planned	Transactional
Aim	Understand	Optimize business
Founded on	Concepts & theory	Technology & tool

Integration IBM® SPSS® Statistics / IBM® SPSS® Modeler

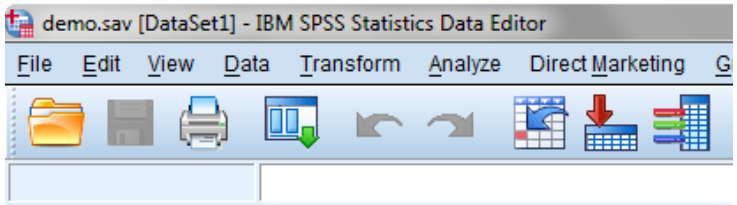
- IBM SPSS Statistics is integrated with IBM SPSS Modeler :
  - There is a dedicated IBM® SPSS® Statistics Palette within IBM®



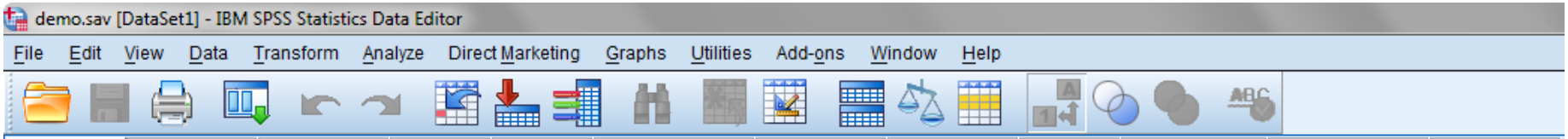
  - Allows Statistics models, transformations, output and syntax within the IBM® SPSS® Modeler Graphics User Interface ➔ Statistics + Predictive Analytics in the same box!
  - Uses IBM® SPSS® Statistics in the background to run analysis from the IBM® SPSS® Modeler interface
  - Requires an IBM® SPSS® Statistics license for the procedures

SPSS Interface

- Data view vs. Variable view
- Variable and value labels
- Variable types and measures



	age	marital	address	income
1	55	Married	12	$72
2	56	Unmarried	29	$153
3	28	Married	9	$28
4	24	Married	4	$26
5	25	Unmarried	2	$23
6	45	Married	9	$76
7	42	Unmarried	19	$40
8	35	Unmarried	15	$57
9	46	Unmarried	26	$24
10	34	Married	0	$89
11	55	Married	17	$72
12	28	Unmarried	3	$24



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	age	Numeric	4	0	Age in years	None	None	8	Right	Scale	Input
2	marital	Numeric	4	0	Marital status	{0, Unmarrie...	None	8	Right	Nominal	Input
3	address	Numeric	4	0	Years at curren...	None	None	8	Right	Scale	Input
4	income	Dollar	8	0	Household inco...	None	None	8	Right	Scale	Input
5	inccat	Numeric	8	0	Income categor...	{1, Under $2...	None	8	Right	Ordinal	Input

IBM® SPSS® Statistics Video

https://www.youtube.com/watch?v=SWdEtsMvSnM

IBM STATISTICS STANDARD

IBM® SPSS® Statistics Base

What is it?

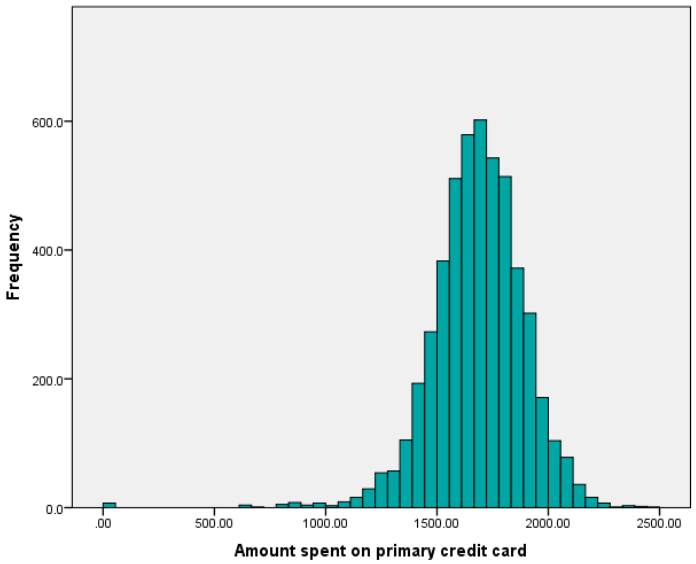
Statistical and data management package for analysts and researchers

When is it needed?

When working through data access, data management, data preparation, preparation, analysis and reporting

What it address?

Ability to generate decision making information quickly using statistics

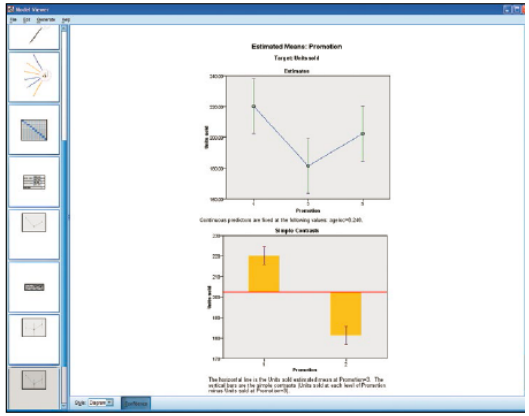


https://www.youtube.com/watch?v=OzAlvH8PbtI

IBM® SPSS® Statistics Advanced Statistics

What is it?

Helps you analyze complex relationships with sophisticated procedures



When is it needed?

- When working with data that has unique characteristics, such as nested-structure data or data with a covariance structure
- If you need to go beyond basic analysis and require extensive modeling flexibility

What's in it?

- Generalized Linear Models
- Generalized Estimating Equations
- General linear models
- Linear mixed models
- Generalized linear Mixed Models for Ordinal Targets
- Variance component estimation
- Survival analysis
- Loglinear analysis
- Generalized Linear Mixed Models
- Combination of Generalized Linear and Linear Mixed Models

https://www.youtube.com/watch?v=8YXs7mioo18

IBM® SPSS® Statistics Custom Tables

What is it?

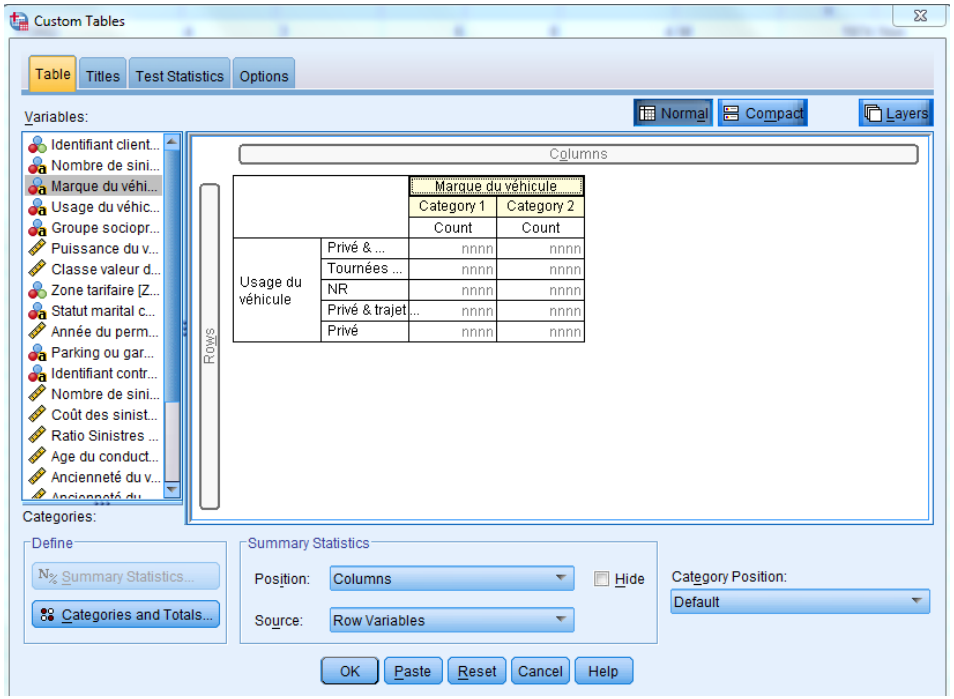
A reporting module that produces high-quality tabular reports

When is it needed?

When needing to summarize your results so they are easily understood and meaningful to a large audience

What's in it?

- Table preview builder
- Flexible table types
- Controlled output during
- table creation
- Inferential statistics
- Powerful syntax



https://www.youtube.com/watch?v=0GBulmBkYvU

IBM® SPSS® Statistics Regression

What is it?

Helps you achieve more modeling control and power with logistic regression, nonlinear regression and other advanced modeling tools

When is it needed?

- When needing advanced predictive capabilities
- When working with categorical data
- If building predictive models but Ordinary Least Squares Regression is too limiting

What's in it?

- Multinomial logistic regression
- Binary logistic regression
- Unconstrained nonlinear regression
- Constrained nonlinear regression
- Weighted least squares
- Two-stage least squares
- PROBIT

https://www.youtube.com/watch?v=Fyz-tFXjNRg

IBM STATISTICS PROFESSIONAL

IBM® SPSS® Statistics Data Preparation

What is it?

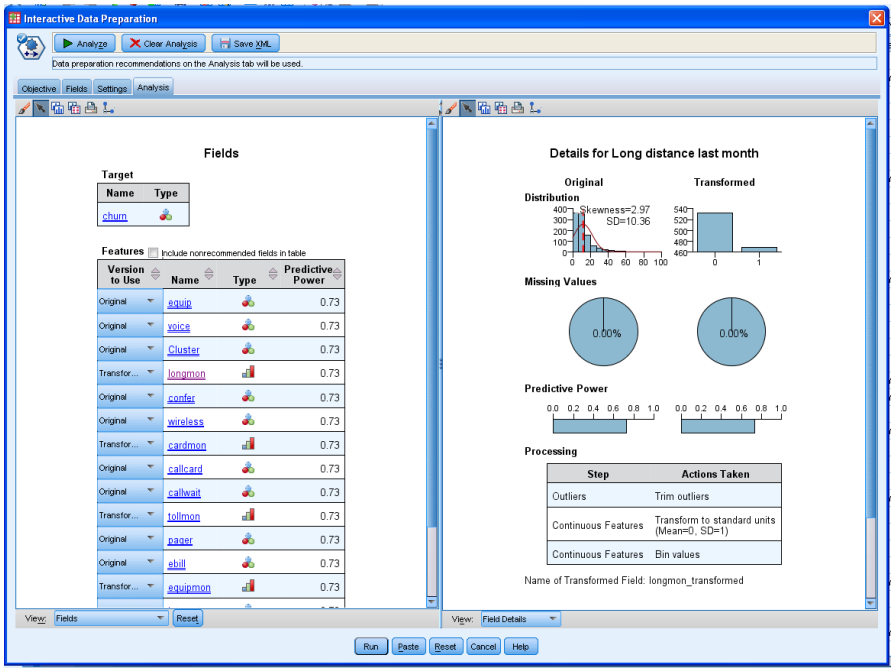
Get data ready for subsequent analyses

When is it needed?

To streamline the data validation process, eliminate labor-intensive manual checks and reach more accurate conclusions

What's in it?

- Streamline the process of validating data before analyzing it
- Anomaly detection to identify unusual cases in a multivariate setting
- Optimal cut-points to categorize continuous predictors
- Automated Data Preparation



https://www.youtube.com/watch?v=jMWb2AMC31s

IBM® SPSS® Statistics Missing Values

What is it?

Helps you overcome missing data problems, resulting in better models and more usable data sets

When is it needed?

Whenever your data set contains missing data – questions without answers or variables without observations

What's in it?

- Pattern analysis and reporting
- Summary statistics
- Missing value estimation algorithms
- Data management
- Multiple imputation

https://www.youtube.com/watch?v=nnEZymh05Ng

IBM® SPSS® Statistics Categories

What is it?

Enables you to understand groupings in your data and predict key outcomes

When is it needed?

- When analyzing categorical data, such as market segments, political parties or biological species or continuous data where you want to use some advanced regression techniques such as flexible functional forms or regularization techniques such as ridge regression, lasso, or elastic net.

What's in it?

- Multi-Dimensional Scaling of Proximity Data
- Principal Components Analysis
- Correspondence Analysis
- Categorical Regression Analysis via optimal scaling
- Homogeneity Analysis via alternating least squares (also known as Multiple Correspondence Analysis)
- Canonical Correlation Analysis of two or more sets of variables via alternating least squares

https://www.youtube.com/watch?v=r4ZTaDoeZgE

IBM® SPSS® Statistics Decision Trees

What is it?

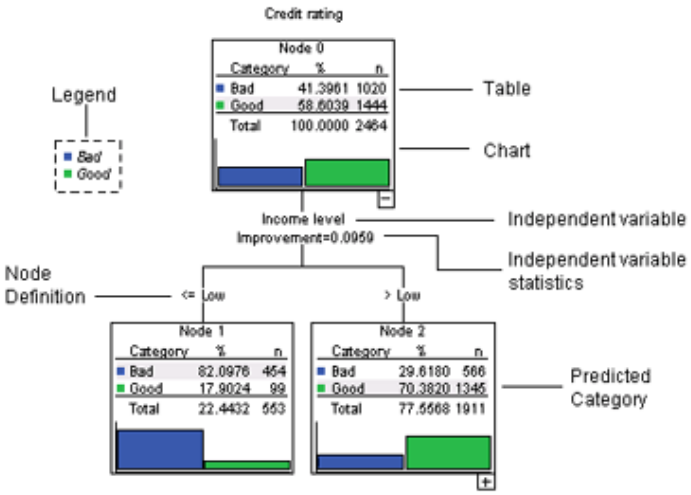
Help you better identify groups, discover relationships between groups, and predict future events

When is it needed?

- When working with data that has complicated, non-linear relationships
- Discover relationships that classical statistical techniques might not find

What's in it?

- CHAID
- Exhaustive CHAID
- CR&T
- QUEST



https://www.youtube.com/watch?v=I-LBKbbJTDM

IBM® SPSS® Statistics Forecasting

What is it?

Helps you predict the future with powerful time-series analysis

When is it needed?

- If you need to analyze historical information, forecast future trends, or predict events
- When seasonality is a consideration in your forecasting

What's in it?

- Box-Jenkins analysis
- Procedures for seasonal factors
- Expert Modeler
- Estimate up to four parameters in 12 different models for exponential smoothing
- Trend regressions
- Regression models with first-order autoregressive errors
- Decompose a time series into its harmonic components

https://www.youtube.com/watch?v=q_U_Lh5Byr4

IBM STATISTICS PREMIUM

IBM® SPSS® Statistics Bootstrapping

What is it?

Enables you to assess the stability of your statistics so you can be confident in their reliability

When is it needed?

When dealing with smaller samples or when you have outliers in your sample

What's in it?

- Framework to bootstrap a variety of statistical algorithms:
  - Estimates the sampling distribution of an estimator by re-sampling *with replacement.*
  - Derives estimates of standard errors and confidence intervals of a population parameter like a mean, median, proportion, odds ratio, correlation coefficient or regression coefficient.

https://www.youtube.com/watch?v=16myQbWsUdk

IBM® SPSS® Statistics Conjoint

What is it?

Helps you measure how individual product attributes affect consumer preferences

When is it needed?

When developing new products, to help determine the most important features and attributes to your customers

What's in it?

- Orthoplan
- Plancards
- Conjoint

https://www.youtube.com/watch?v=PjWQIrv7Jpc

IBM® SPSS® Statistics Exact Tests

What is it?

A resource when data is limited and collecting more is not an option

When is it needed?

When investigating relationships among categorical variables or using nonparametric analysis techniques, and have a limited amount of data

What's in it?

- Over 30 exact testing methods for nonparametric and categorical data problems:
  - One-sample, two-sample and K-sample tests
  - Independent or related samples
  - Goodness-of-fit tests, tests of independence in crosstabulations and on measures of association

https://www.youtube.com/watch?v=_hPhI8pofT0

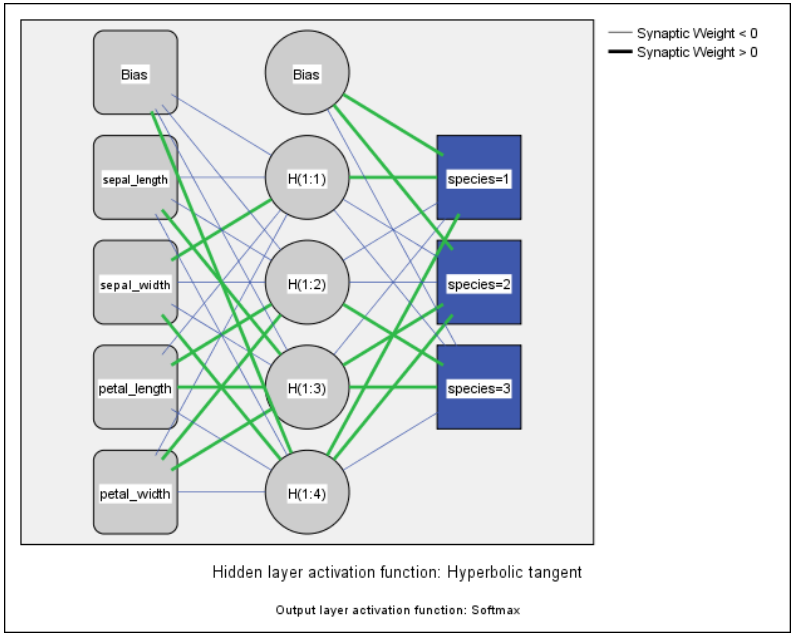
IBM® SPSS® Statistics Neural Networks

What is it?

Helps you model complex relationships between inputs and outputs

When is it needed?

- When working with data that has complicated, non-linear relationships
- Discover relationships that classical statistical techniques might not find



What's in it?

- Multilayer Perceptron
- Radial Basis Function

https://www.youtube.com/watch?v=RU4OzHXF4KQ

IBM® SPSS® Statistics Direct Marketing

What is it?

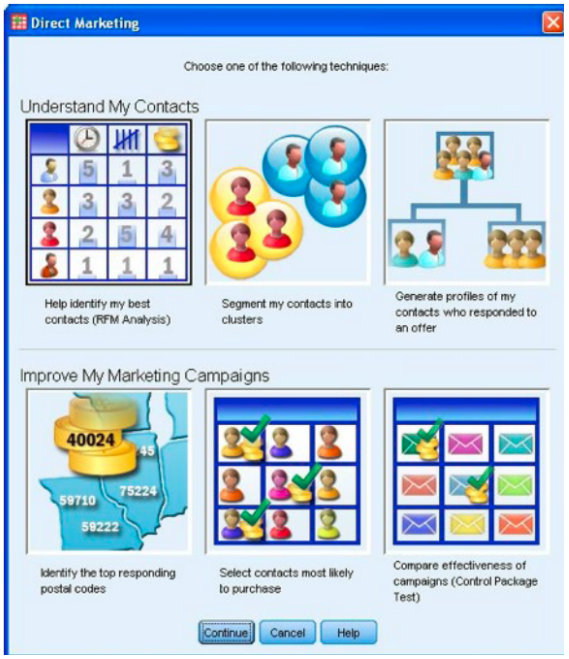
Enables you to maximize the ROI of your marketing budget

When is it needed?

When less technical users need to make marketing programs as effective as possible

What's in it?

- RFM analysis based on: Transactional data and Customer data
- Propensity to Purchase to determine who is most likely to respond
- Generate profiles of contacts
- Control Package Test to see which offer works best
- Cluster Analysis for segmenting customers
- Postal Code Response to identify top postal codes



https://www.youtube.com/watch?v=amXiDf6kn3s

IBM® SPSS® Statistics Complex Samples

What is it?

Helps you achieve more precise analytical results when working with large-scale surveys or complex sample designs

When is it needed?

When working with data not arising from simple random sampling, for example, in survey and market research, social sciences and public opinion research

What's in it?

- Complex Samples Plan
- Complex Samples Selection procedure
- Complex Samples Descriptives
- Complex Samples Tabulate
- Complex Samples Logistic Regression
- Complex Samples General Linear Models
- Complex Samples Cox Regression

https://www.youtube.com/watch?v=_APhWVIBh3s

IBM® SPSS® Viz Designer

What is it?

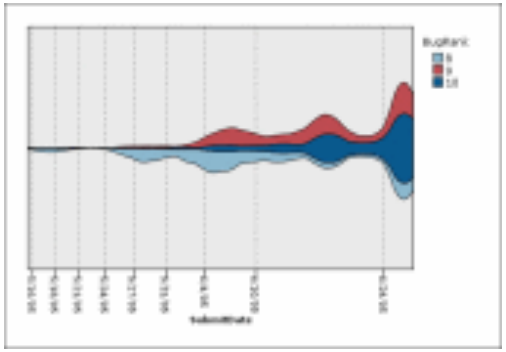
Create compelling templates that can be used in several IBM SPSS software products

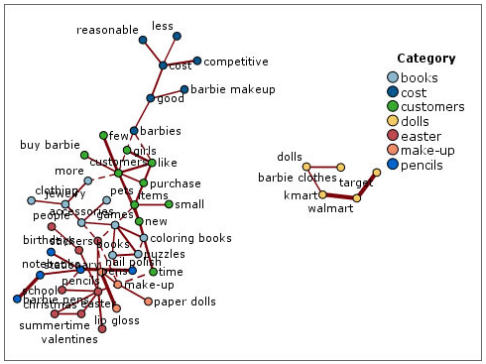
When is it needed?

Enabling all users to present results in clear and compelling ways

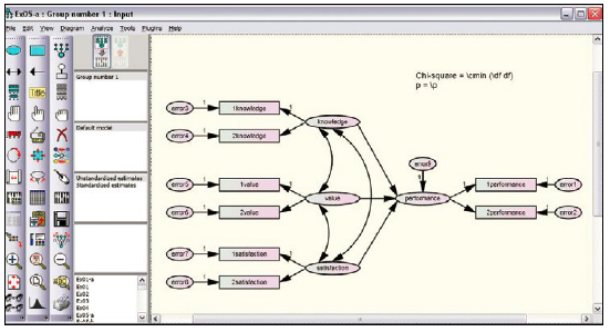
What's in it?

- "Drag-and-drop" graph creation
- Built-in templates
- Use style sheets and graph templates tailored to your organization





IBM® SPSS® Amos



What is it?

An easy-to-use structural equation modeling (SEM) tool

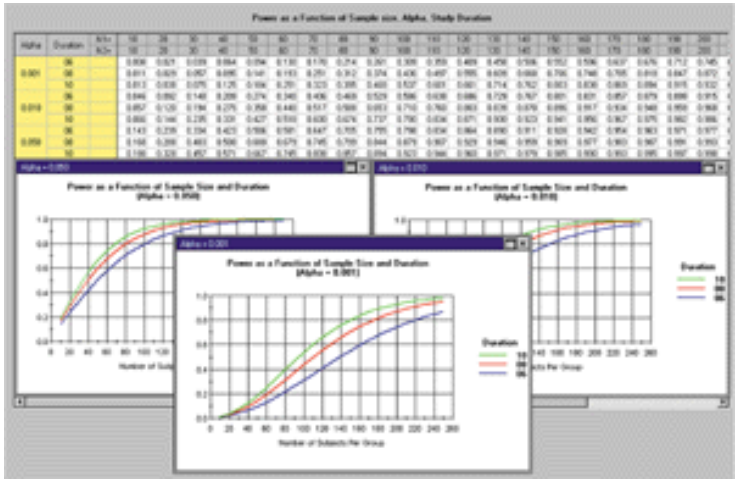
When is it needed?

- When you need to go beyond standard analytical methods, such as regression, factor analysis, correction and analysis of variance
- When comparing multiple groups or analyzing longitudinal data
- When research includes unobserved (latent) variables in an analysis

What will it help me address?

- Study relationships and test hypotheses- find out which variables affect each other and by how much
- Testing complex relationships - any numeric variable, observed or not, can be used to predict any other variable

IBM® SPSS® Sample Power



What is it?

Helps you identify the appropriate sample size for your research

When is it needed?

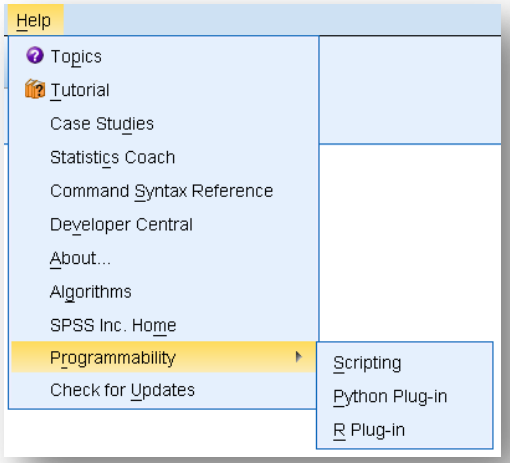
- When unsure of the appropriate sample size that will allow you to confidently accept or reject your hypothesis and defend your analysis
- When power analysis is required, for example, with many grant applications

What will it help me address?

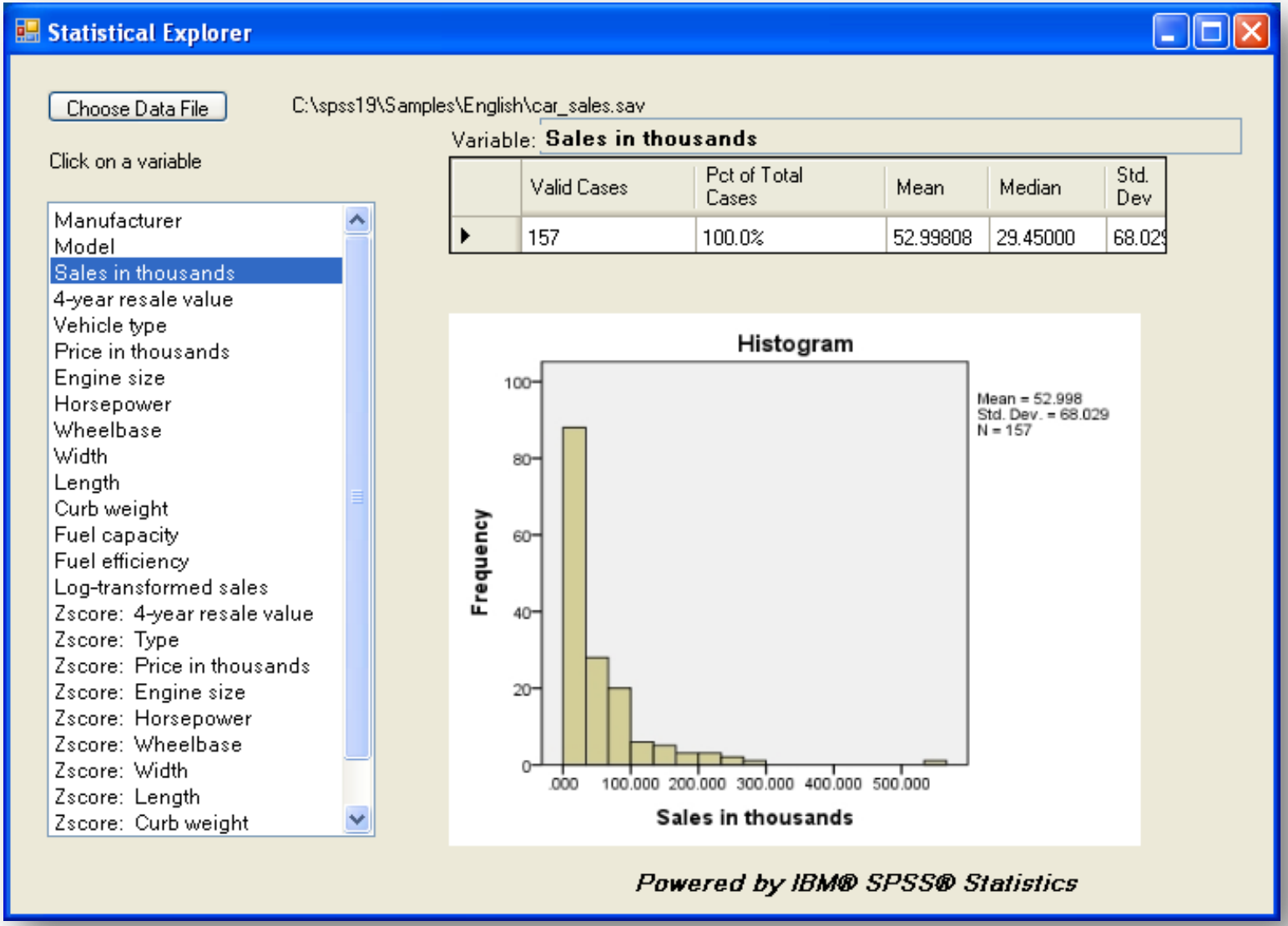
- Easily find the appropriate sample size
- See how your sample size will affect statistical power
- Compare scenarios before you start your research

Programmability and Scripting:
IBM SPSS Statistics embeds three programming languages

- Plug-ins let you extend Statistics capabilities using:
  - Python
  - R
  - .NET languages (Windows only)
  - Java
- Free plug-in downloads
- Use existing materials without learning programmability
- Create your own programs if you do learn programmability



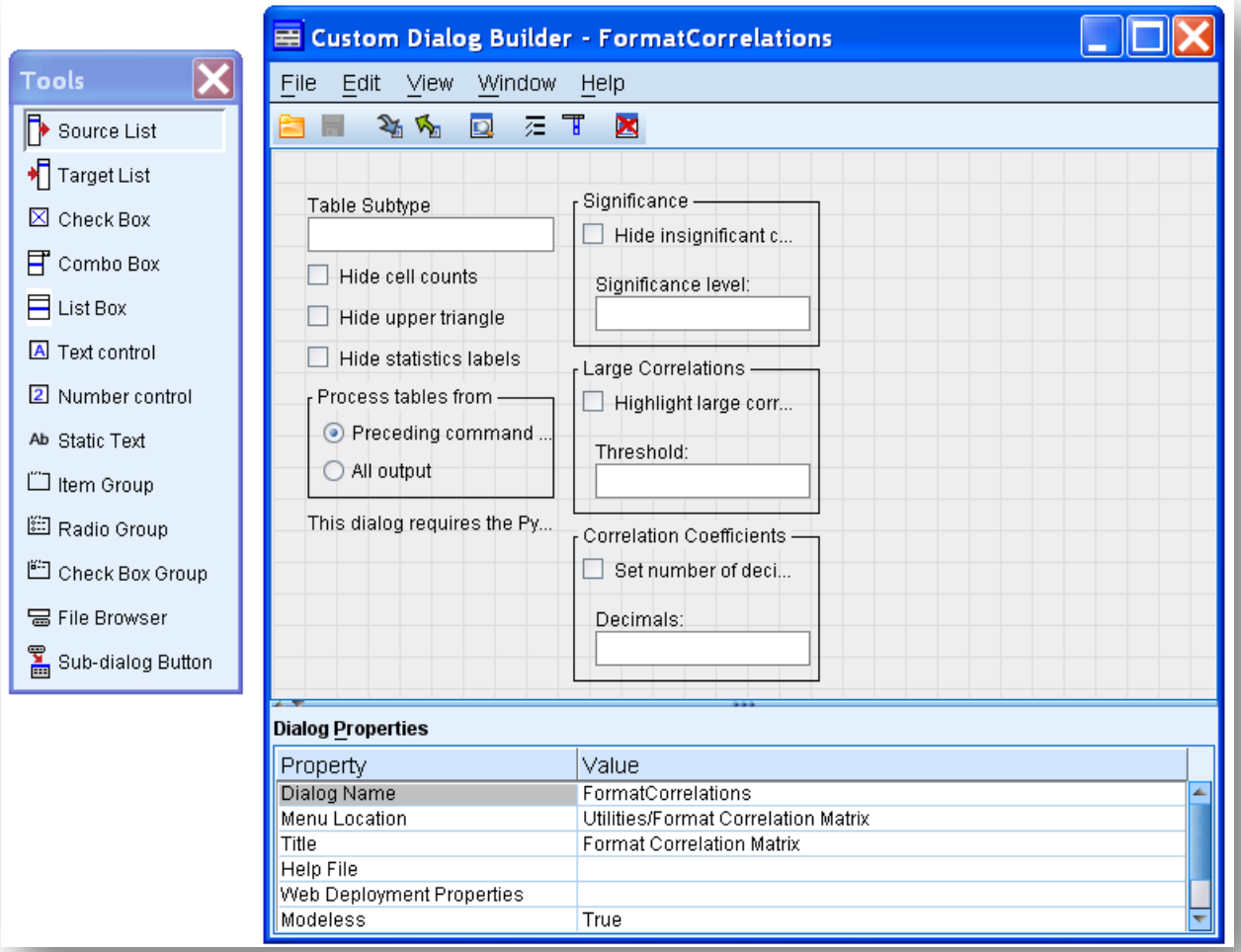
Custom interface built with .NET plugin



Two pages of VB.NET code

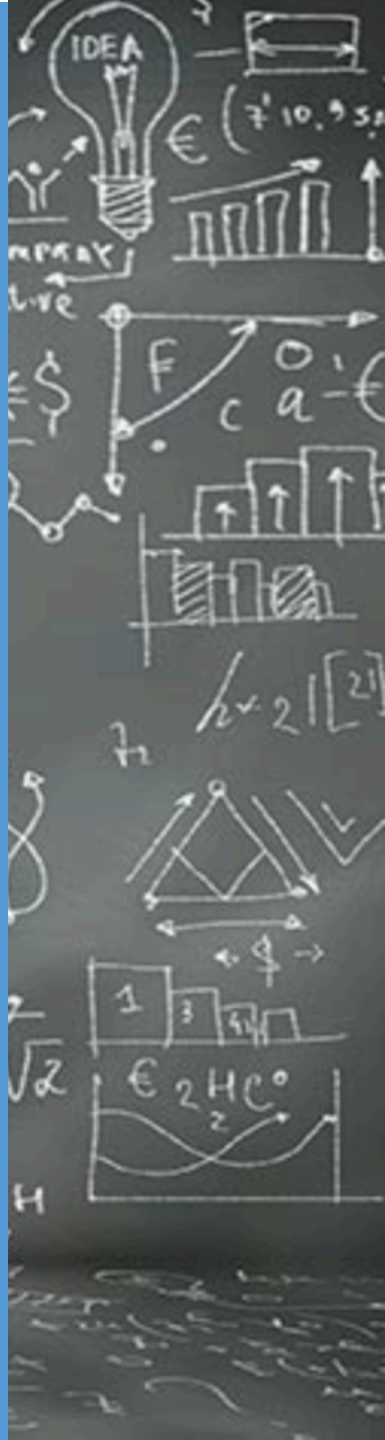
Custom dialog boxes can be easily created to provide user interface for Python or R programs

- No programming required
- Can be built very quickly
- Custom Dialog Builder included with Statistics
- Can be used for new extensions or for existing commands
- Easy to package and install
- Custom dialogs work in IBM SPSS Modeler and IBM SPSS C&DS
- If Statistics is installed custom functionality can be used within Modeler



ILLUSTRATION

Thank You



Yann Gouedo
Data Scientist Leader – Machine Learning / Artificial Intelligence
Marketing / Risk / Fraud / Maintenance
IBM Certified Senior Data Scientist & IBM Certified Senior Architect