Accelerate the Path to Deeper Insights

Bridging the Gap between Data Science and Data Understanding

Generating robust insights from data begins with asking the right questions. Analysts should be able to draw insights from information that is already available. To do this, they need tools with the ability to easily and independently test, validate, and refine their hypotheses so they can generate valuable new information.

For example, a business analyst looking at the value of a marketing promotion might hypothesize that it generates a statistically significant increase in net revenue. Ideally, that analyst has access to the tools and techniques that help easily and quickly provide intuitive results. The ability to assess a hypothesis without involving a data scientist makes it easier for analysts to run through multiple scenarios; to further investigate the relationships between pricing and volume; as well as the effects of other factors, such as seasonality, store location, and competitor pricing.

The right processes and tools make it easy to tie together the business analyst's work with that of career data scientists, who can take the analysis to the next level and provide more sophisticated results. Among the benefits of this type of structured, multi-level approach are:

- **Maximizing the diverse skillsets on your team,** including the business analyst's domain expertise and the data scientist's deeper understanding of statistical algorithms and techniques.

- **Improving overall efficiency,** with low-overhead testing and validation of hypotheses before moving to the more time-consuming process of data modeling

- **Optimizing the quality and value of hypotheses** to better prepare for data modeling and ultimately improving the insights that are obtained.

IBM SPSS Statistics provides a robust approach to developing both questions and answers. For business users, it features a rich, intuitive user interface that distills the results of complex statistical tests into charts and graphs that help visualize and isolate the impacts and significances of individual factors. For data scientists, the software makes possible deep-dive analysis with an extensive range of techniques and modules. Data scientists can build on analyses by discerning subtle or hidden patterns in data, conducting complex sampling and testing, and developing reliable forecast models to predict future events.

Comprehensive, Efficient Data Conditioning and Shaping

Data preparation can consume as much as 80 percent or more of a statistical-analysis project. This preparation is dominated by manual techniques, which are a primary contributor to missed deadlines. When analysts need to invest significant resources on data preparation, it often reduces the amount of time available for creating, refining, and executing the actual analysis models.

Because different individuals tend to go about data validation in different ways, it can be difficult to achieve consistency from project to project. At the same time, project-specific circumstances often require tailored approaches. Automation is one way to streamline data preparation, but it cannot be the only answer. Manual techniques need to be used where appropriate. Automation allows for accelerated workflows, while the manual techniques deliver the granular control that is often needed for more challenging datasets. That's why best practices call for tools and procedures that combine both.

Automate Data Preparation Tasks

The lack of tools for robust automation is the primary cause of inefficiency in data conditioning and shaping. IBM SPSS Statistics provides rich automation capabilities that include both general-purpose and specialized techniques for data preparation. It enables users to perform functions such as identifying invalid values, viewing patterns of missing data, and summarizing variable distributions.

These capabilities help improve data preparation by accelerating the process and enabling more accurate results. Examples of the procedures that can be automated include:

Data preparation can consume as much as **80%** or more of a statistical-analysis project.

- **Validate data.** Automate data checks by applying rules that flag any invalid values, generating reports that tabulate and summarize the results. Rules can be applied to individual variables, such as identifying values outside an allowed range, and they can also be applied across variables, to identify cases where values in two or more data fields are incompatible.

- **Assess and improve data quality.** Identify and correct data-quality issues. Reporting includes visualizations as well as recommendations that allow users to drill down and examine discovered issues as part of the process of determining which data to use in the analysis.

- **Detect data anomalies.** Pinpoint outliers that could otherwise skew analyses by searching for unusual cases and recognizing the nature and cause of those deviations from their peer group. Outliers can be flagged by creating a new variable and then analyzed further to determine whether they should be included in or excluded from the analysis.

- **Optimize binning.** Determine cutpoints to optimize outcomes when binning scale variables for algorithms that are designed for nominal attributes. This procedure allows for unsupervised, supervised, or hybrid approaches, which allow users to strike their preferred balance between accuracy and speed.

Identify and Replace Missing Values

Addressing gaps in datasets is vital to avoid biased or statistically insignificant results. Broadly, this requirement has two major parts: the first is diagnosis to identify and analyze missing data values, and the second is imputation to estimate and replace them.

Diagnosis begins with reporting by IBM SPSS on the missing values, which provide case-by-case analysis that includes snapshots of each case, as well as extreme values for each one. Various types of testing and cross-tabulation enable discovery of patterns among the missing values, and between those missing values and the rest of the data. This analysis helps determine the significance of the missing values to the overall results of the analysis.

Imputation further examines missing values and constructs plausible estimates to replace them using multiple imputation methods. The process can be run automatically or customized as needed. IBM SPSS generates multiple datasets based on possible replacement values, which can be modeled using techniques such as regression to analyze the replacement datasets and support imputation of the missing data.

Test Model Stability in Advance

The accuracy of any analysis depends on generating the most reliable data model possible, which requires testing across more than a single data sample. When creating a model to predict an outcome or map a sample to a population, for example, simply running the model on the sample data may be insufficient, because the outcome will be dependent on the sample itself.

To overcome this limitation, IBM SPSS supports bootstrapping analytical procedures, which allow resampling to provide as many as thousands of alternate datasets from the original sample. Computing a statistic on a large number of alternate datasets helps determine the variability of that statistic. It allows for more reliable estimation of standard errors and confidence intervals of population parameters such as mean, median, proportion, and others. A more accurate view of what is likely to exist in the population helps eliminate outliers and other anomalies that can otherwise degrade the accuracy or applicability of the analysis.

Improve Analysis with Better Data Preparation

IBM SPSS Statistics provides selective, controllable automation of the data preparation process, which helps reduce the effort required to build models, improve their quality, and optimize their predictive power. By accelerating the pre-analysis phase and enhancing the accuracy of conclusions, users will improve the overall statistical analysis workflow.

Robust Statistical Analysis and Modeling

If your organization wants to make the most of its available data and use it to solve business and research problems, you should consider your requirements from multiple perspectives. Your solutions must be powerful enough to conduct statistical analysis on large datasets wherever they reside. They also need to provide assistance in understanding the full potential value of the data. Finally, your tools must provide insight around the best techniques for unlocking that value, including additional dimensions that might otherwise be overlooked, such as considering weather and geographical location in the analysis of a marketing promotion.

Rather than charting your own course, you should draw in established industry best practices, tools, and techniques. Solutions for statistical analysis should be well integrated with all of the common data sources your organization uses, as well as open-standard external data sources, to facilitate the greatest flexibility.

In addition, support for open source and industry standard tools can dramatically extend native capabilities. The statistical programming languages R and Python are favorites among data scientists. R helps data scientists work with large datasets thanks to features like linear and non-linear modeling, time-series analysis, and clustering. Python has a reputation for being easy to learn and use, which increases its appeal for users new to statistical analysis. These users can focus on their data, rather than the complexity of the programming, and they often move on to more complex languages as they gain experience.

IBM SPSS Modeler features extensive integrations with R, Python, and Spark. It includes multiple nodes that run Python algorithms — one-class SVM, SMOTE, XGBoost, t-SNE, Gaussian Mixture, KDE, and Random Forest. Previously available only via Python coding, these algorithms are now exposed directly in the Modeler GUI. IBM SPSS Statistics allows users to enhance SPSS Syntax with both R and Python through specialized extensions; a large existing collection is available, or users can build their own.

Build from a Core Set of Analysis Capabilities

Regardless of the use case, statistical analysis solutions should be built on a single foundation. The platform must be able to access a wide variety of data formats, and the size of the datasets should not be a limitation.

The capabilities you need to address your business and research challenges include a wide variety of linear regression models for complex data relationships, as well as nonlinear models. Core capabilities should also include simulation modeling using techniques such as Monte Carlo simulations. Additional areas of interest may

> “Statistical analysis solutions should be built on a single foundation. The platform must be able to access a wide variety of data formats, and the size of the datasets should not be a limitation.”

include Bayesian procedures and geospatial analytics. IBM SPSS Statistics offers a comprehensive set of capabilities in these areas, in addition to a wide range of additional functionality.

Apply Advanced Statistical Methods and Summarization Techniques

Beyond its core capabilities, IBM SPSS Statistics also supports advanced approaches for analyzing complex relationships, including multiple outcomes or outcomes over time. Univariate and multivariate analytical techniques allow users to build sophisticated, flexible models using data that do not conform to the requirements and assumptions of simpler techniques. Results from these models can be analyzed using a wide variety of methods, and the data can be summarized using sophisticated tables and visualizations that help convey insights effectively.

Add Forecasting and Predictive Tools

Integrated tools for developing forecasts and predicting trends based on time-series data can help strengthen your strategic planning. Advanced statistical methods available for both advanced and novice users enable people throughout the organization to participate in development while also giving experienced forecasters more granular control and powerful tools. On the whole, these capabilities extend the value of business processes, reduce errors, and increase the efficiency of creating, managing, and updating forecasts and decision trees.

Unlock the Full Value of Source Data

IBM SPSS Statistics helps your business apply proven techniques, based on a mature toolset that has developed over 50 years. IBM SPSS Statistics provides the foundations for better understanding the potential of your data, making it and the existing expertise surrounding it more effective. Using a structured approach based on IBM SPSS Statistics enhances your statistical analysis by bringing the full range of available data and techniques to bear on business and research problems.

Enterprise Collaboration and Decision Making

Your users cannot unlock the value of their statistical analysis unless they can turn their results into accurate decision making. The true challenge for many organizations lies in the integration of technology solutions with the business operations they are meant to support. In other words, your tools and expertise are only as valuable as the ability to take advantage of them.

Making a statistical analysis approach enterprise-ready is a both a business-management challenge and an IT one. From a business-management perspective, an over-arching, formal strategy for statistical analysis creates efficiencies of scale, preventing duplication of effort and sharing insights for maximum value. On the IT side, incompatible silos of tools and approaches used by different business units must be brought together. Doing so enables your analysis to draw on data from throughout the organization and fosters cross-pollination of expertise among business units.

Replace and Strengthen Ad Hoc Approaches

There is a tendency in many organizations for statistical analysis methods to grow organically, with different groups developing their own approach. This is often rooted in specific experience that people already have with certain tools and techniques, which in many cases are not based on best practices. Making up an approach as you go is unlikely to maximize the value of your data.

Without the ability to compare ad hoc statistical approaches with a more formalized, robust approach, many organizations may not even be aware of the shortcomings. It typically falls to upper management to establish repeatable, enterprise-wide methods and techniques. The process of deciding what the standards should look like is complex and depends on the needs of a specific organization, but most cases share a set of common attributes.

Start by demonstrating to various user groups how they could benefit from the data sources and analysis of others in the broader organization. For example, information on trouble tickets and customer complaints is beneficial to others who work in research and development or the design of products and services. Business logic created for statistical analysis related to marketing a product in one division may be useful to similar activities in another part of a company, and so on.

When every business unit has a coherent, standardized set of enterprise-grade statistical tools, there is a mechanism in place to enable this collaboration. In addition to sharing benefits, the quality of the insights available to individual business units in isolation will also improve when

collaboration is taking place. IBM SPSS Statistics offers a range of benefits to users across the organization.

- Build efficiencies and save time. Instead of requiring users to manually import data, build formulas, and update reports, enterprise-grade statistical analysis tools automate these and other tasks.

- Improve quality and strengthen results. Users can base their analyses on proven, built-in advanced statistical techniques, outputting the results using sophisticated charts and visualizations to reveal and share insights.

Implement a Centralized Analytical Tools Approach

In addition to the operational framework put in place by upper management, the IT organization must also support a practical means of conducting robust statistical analysis. IBM SPSS Statistics integrates with existing tools and meets the diverse needs of different business units, as well as roles that range from business users to data scientists.

- Flexibility and scale. Analyses can be extended to tasks that involve many variables, huge datasets, and complex statistical correlations, while maintaining data integrity and providing audit trails.

- Sophisticated analysis capabilities. Users can perform advanced analytical functions rapidly, without having to build them from scratch, also benefiting from built-in data-preparation and validation tools.

- Programmability for advanced users. Analyses generate syntax that can be reused on other datasets, using the native GUI or extensions for programming in Java, .NET, Python, or R.

IBM SPSS Statistics is available with flexible licensing arrangements on a subscription basis or using a perpetual model.

Share and Re-Use Assets Effectively

Making statistical analysis assets, processes, and results available across the organization is central to effective collaboration, process automation, and operational efficiency. The means of supporting these requirements must be seamless for users while also providing effective access and version control.

IBM SPSS Collaboration and Deployment Services provide a central repository for the sharing and re-use of analytical assets that increases efficiency while helping standardize best practices. Business users can view and interact with the materials in the repository and benefit from the work of others to deliver more consistent results. Learn more at ibm.com/spss/cds.

Make Data-Driven Insights More Powerful and Efficient

As organizations strive to work smarter and perform at their best, they need better ways to transform information into insights. Using IBM SPSS Statistics, business users and data scientists can benefit from a single tool that helps them bridge the gap between data science and data understanding. They benefit from better data preparation, accelerating analysis, and improving the accuracy of conclusions. The organization as a whole is able to understand and get greater value from source data, with an enterprise-ready solution that unifies and strengthens statistical analysis throughout the organization.