



Forrest W. Young
with contributions by
Pedro Valero & Gabriel Molina

Seeing Data

The Visual Statistics Project
www.visualstats.org

forrest@visualstats.org pedro@visualstats.org gabriel@visualstats.org

FORREST W. YOUNG
PSYCHOMETRIC LABORATORY
UNIVERSITY OF NORTH CAROLINA
CB 3270 DAVIE HALL
CHAPEL HILL, NC 27599-3270 USA

PEDRO VALERO & GABRIEL MOLINA
INSTITUTE OF TRAFFIC AND ROAD SAFETY
UNIVERSITAT DE VALÈNCIA
C/ HUGO DE MONCADA 4. ENTRESUELO.
VALENCIA, 46010, ESPAÑA (SPAIN)

Reference:

Young, Forrest W., Valero, Pedro & Molina, Gabriel. Using ViSta to See Your Data: Chapter 1: Welcome! The Visual Statistics Project (www.visualstats.org). November 2001. Chapel Hill, NC, USA and Valencia, Spain.

Copyright (c) by Forrest W. Young, Pedro Valero, & Gabriel Molina. 2001. All rights reserved.

Contact Information: (at the time of writing this is still in development)

Discussion Groups:

vista@visualstats.org	general purpose discussion
developers@visualstats.org	application and system development discussion

Training and Customized Software

training@visualstats.org	classes in how to use ViSta
software@visualstats.org	professional customized software development

User Contributions:

doco@visualstats.org	documentation development
plugers@visualstats.org	application development
hackers@visualstats.org	system development
translate@visualstats.org	translate doco or code into other languages
porters@visualstats.org	port vista to new operating systems

Bugs and Problems

bugs@visualstats.org	report bugs and problems
--	--------------------------

seeing data

**first, you see your data for what they seem to be
then, you ask them for the truth -
are you what you seem to me?**

**you see with broad expanse
yet ask with narrowed power
you see and ask and see
and ask and see ... and ask**

**with brush you paint the possibilities
with pen you scribe the probabilities**

**for in pictures we find insight
while in numbers we find strength**

forrest w young

About Prof. Young

Forrest W. Young, Professor Emeritus at the University of North Carolina at Chapel Hill, received his PhD in Psychometrics from the University of Southern California in 1967. He has been on the faculty of UNC-CH ever since. Prof. Young's teaching interests focus on "Seeing what your data seem to say". This visually intuitive approach to statistics helps to clarify the meaning of data. His courses, ranging from his introductory undergraduate course on Psychological Statistics, to his advanced graduate courses on Data Analysis, Visualization and Exploration, reflect this focus.

To make the process of understanding data visually intuitive, the burden is moved from the person to the computer. You don't need to make an intensive effort to understand your data: Rather, your computer makes intensive calculations so that the data can be shown to you in a visually comprehensible way. This approach to the role of computers is based on the intelligence augmentation (IA) philosophy of Computer Science: Your computer is a device which should augment your intelligence. It is also based on a Cognitive Science theory for the construction of an environment for data analysis.

Prof Young and his students, over the course of a 10-year research and development project, have created ViSta, a visual statistics system instantiating Prof. Young's theories concerning visual environments for statistical analysis. ViSta is a freely available system that is being used for teaching introductory and multivariate statistics, for data analysis by statistically inexperienced researchers as well as by those who are more advanced, and for advanced research and development in graphical and computational statistics.

ViSta is based not only on Prof. Young's theory-based approach to data analysis, but also on his 30-year career in computational and graphical statistics. Prof. Young's early research interests focused on Multidimensional and Nonlinear Multivariate Data Analysis (for which he was elected the President of the Psychometric Society, and received the American Market Research Association's O'Dell award, both in 1981). Via these research interests, Prof. Young became involved in software development early in his career.

Prof. Young has served as a professional consultant on statistical system interface design with SAS Institute, Statistical Sciences (the S-Plus system), and BMDP Inc. He has written or designed data analysis modules for the SAS, SPSS and IMSL systems. He is a member of the American Statistical Association's sections on Computational and Graphical Statistics.

Acknowledgements

Over the last 10 years, ViSta has evolved from a testbed for the design of a highly interactive, very dynamic statistical visualization system, to just such a system. As the creator, designer and primary implementor of ViSta, I am deeply indebted to many, many people who have contributed in one way or another over the years. Without them, ViSta would not exist.

My deepest thanks and appreciation go to my wife Patricia Young. During the last decade she has stood watch while I wandered in the developer's cave, evolving ViSta from what I could see on the screen to what I could see in my mind's eye. Patty's patience, understanding, forbearance, support, and steady love, have been crucial to the realization of my dreams. Her artistic talent, abilities and spirit have shaped those dreams, and have had a deep impact on ViSta's design. Check out her art at patricia-mae-young.com and at art.net/~patricia.

I deeply appreciate Norman Cliff's influences on me. As my mentor throughout graduate school, and as the chairman of my dissertation committee, Norm had a clear and major influence on my career's beginnings. But of even more importance, Norm's appreciation of the simple approach, his gentle

strength, quiet certitude, civility and his strong sense of the ethical, have had a continuing impact throughout my entire career. Thanks Norm. You have been a great role model.

Mucho Gracias to Pedro Valero-Mora for his time, energy, enthusiasm and, last but not least, his code. Pedro wrote the code for missing data features, including visual transformation and imputation methods; as well as the visually-based Box-Cox and Folded Power transformations. His newest contribution, the log-linear plugin, appears in this release for the first time. Thanks for everything you've done, Pedro. I expect exciting new developments to come.

And also Mucho Gracias to Gabi Molina, whose mild manner hides a truly thoughtful person. Gabi has spent many hours editing and translating text, and has taken up the challenge to take over the day-to-day activities, and to get visualstats organized! Good luck!

And to Michael Friendly, who has contributed much talk, laughter and encouragement, as well as code. But above all, he also has the vision, and is completely satisfied when others can implement his ideas, making his vision visible.

I greatly appreciate Dominic Moore's friendship, and support. His encouragement, criticisms, cheers, enthusiasm, pep-talks, humor, luniness, food, music and incredible speakers are food for my soul.

Thanks and much appreciation to Luke Tierney for creating and providing the XLisp-Stat statistical development system, which is the foundation on which ViSta is built. And thanks to the many folks who have built the XLisp system on which XLisp-Stat was itself built: Open Software at its best. And thanks to Sandy Weisberg who's timely and insightful suggestions on getting ViSta out the door have helped, though perhaps not as much as everyone hoped!

David Lubinsky and John B. Smith made very fundamental contributions to WorkMap architecture. Without their seminal input there'd be no ViSta. Richard Faldowski and Carla Bann contributed fundamentally to SpreadPlot architecture, and spent many intense months implementing their ideas.

"Merci Beaucoup" to Louis Le Guelte for the French translation. and "Mucho Gracias" to Maria R. Rodrigo, Gabriel Molina and Pedro Valero for the Spanish translation.

Bibb Latane, of Social Science Conferences, Inc. funded development of the Excel-ViSta connection, of multipane windows, and the developer's conference. Angell Beza, Associate Director of UNC's Odum Institute for Research in the Social Sciences, was instrumental in obtaining the SSCI funds. Jose Sandoval, also at UNC's Odum Institute, developed the first public courses for teaching about using ViSta.

Doug Kent, funded by UNC's Office of Information Technology, implemented Version 5's MS-Windows features. Jenny Williams of UNC's Office of Information Technology, was my NT-Guru, especially on the network installation.

Mosaic Plots and Bar Charts are based on an algorithm by Ernest Kwan. Frequency Polygons and Histograms based on algorithms by Jan de Leeuw and Jason Bond. Many of the Model Objects were written by Carla Bann, Rich Faldowski, David Flora, Ernest Kwan, Lee Bee Leng, Mary McFarland and Chris Weisen.

The C-to-Lisp Parser used in ViVa was written and copyrighted by Erann Gat, who reserves all rights. Used under the terms of the GNU General Public License. The developer's menu derives from Kjetil Halvorsen, the Symbol Editor Dialog was implemented by Frederic Udina, and the BitMap Editor by Fabian Camacho. Code for earlier versions was modified for Unix by Anthony J. Rossini, Charles Kurak, Albrecht Gebhardt and Andrew V. Klein.

Quality assurance provided, unwittingly, by my undergraduate students, who suffered through "creeping-featuritis" and old documentation.

Thanks to all of you for helping me make ViSta what it is today. Without you, it would not exist.

Forrest.



Using ViSta

CHAPTER 1

Have Fun - Learn Lots!

Welcome!

HAVE FUN - LEARN LOTS. ViSta - The Visual Statistics System is ready to help you "Have fun seeing what your data seem to say". ViSta's fun and playful approach to data analysis helps you see what your data seem to say, and to test the truthfulness of what you think you have seen.

YOUR PERSONAL STATISTICIAN. ViSta is more than a statistics system, it is your statistical advisor, consultant and teacher --- ViSta is your own personal statistician! ViSta is designed to help you learn about your data --- and to help you learn about how to learn about your data. ViSta will help teach you about:

- SEEING YOUR DATA to understand what they seem to say;
- TESTING THE TRUTH of what you think you've seen;
- GAINING INSIGHT through cycles of seeing and testing.

YOUR STATISTICS VIDEO GAME. ViSta has tools for playing with your data ... tools which are graphical, dynamic, and interactive ... tools which encourage you to have fun with your data. And, when you are having fun, your playful and relaxed state augments your ability to understand your data ... increases your insight about your data.

Help!

THE HELP SYSTEM. We all need help, every now and then, so ViSta's help system is always there when you need it. To get help, just:

- Click on the desktop's big red question mark; or
- Use the HELP TOPICS item of the help menu; or
- Type (help) in the listener window; or
- Type ? in the listener window.

Whenever you take one of these actions the HELP TOPICS PANEL appears, showing the help topics. Click on a topic to see its subtopics. Then click on a subtopic to see it. The specific part of the help which will be most useful depends on how familiar you are with ViSta, and what you want to do with it. Specifically, if you are:

- A New ViSta user, you should read the USING VISTA topic.
- Familiar with ViSta, you should read THE WHAT'S NEW? topic.
- Using ViSta with Excel, you should read the EXCEL topic.

ONLINE RESOURCES. You can learn more about ViSta by becoming a registered user. Registered ViSta users receive information from us about the latest upgrades, plugins, documentation and bug fixes. Choose REGISTER WITH US subtopic of the VISTA ONLINE topic.

You can learn more about ViSta by joining the news group. News group members can ask others about using ViSta, and will hear how others use ViSta. Choose JOIN THE NEWS GROUP subtopic of the VISTA ONLINE topic.

New Features

NEW LOOK AND FEEL. We have developed ViSta with one goal always foremost: Maximize the quality of the user's experience. As such, we have created ViSta so that it is:

- **Quick and smooth.** The visualizations and desktop work smoothly, bringing you closer to your data, presenting a data analysis environment that is a more satisfying and relaxing, and bringing you a richer visual experience for a deeper understanding of your data.
- **Simple to use.** ViSta has a simple user interface involving DeskTop, DataSheet, SpreadPlot and Report windows. The DeskTop window has window-panes corresponding to the former workmap, variables and observations windows. When you manipulate the DeskTop window all of its panes are manipulated accordingly, meaning there is only one window to manipulate rather than four. The same is true for SpreadPlot windows: They have panes corresponding to former windows: Manipulating one spreadplot window automatically manipulates all of the former windows in a coordinated way. Report windows continue to work as before.

NEW SOFTWARE ARCHITECTURE. ViSta has a new software architecture involving a core system with plugins and addons. This means that ViSta is

- **Flexible.** The plugins and addons let you mold the data analysis methods and environment to fit your needs.
- **Stable.** ViSta's core system, and its unique selection of data visualization, exploration and description features, remains unchanged, so that it is stable and can become truly mature.

ViSta is an open software system. The code is open to those who are qualified and interested in developing new features. Developers can write new plugins that introduce new data analysis capabilities, and new addons that add new features to the data analysis environment. This permits ViSta's core system, and its unique selection of data visualization, exploration and description features, to remain stable and mature.

EXCEL

You can use ViSta to visualize and analyze your Excel data. After a one-time configuration, each time you run Excel there will be a ViSta menu on the Excel menubar. You then simply select the portion of your spreadsheet that you want ViSta to visualize or analyze, and then select the menu item from the ViSta menu that is appropriate. The items in the menu can be tailored to perform the specific types of visualizations and analyses that are appropriate for your data. All of ViSta's features can be accessed from Excel's ViSta menu.

MISSING DATA

Thanks to the efforts of Pedro Valero, of the University of Valencia, in Spain, ViSta can now process data with missing values. ViSta has methods for imputing values for the missing data, and has visualizations for inspecting the results of the imputations. In addition, most of the standard data processing methods can be used with data which have missing values.

- **PROCESSING MISSING DATA** - Data with missing values, indicated by the symbol NIL, can be processed by ViSta 6. They can be input, manipulated, described and visualized in the same ways as data without missing values.
- **IMPUTING MISSING DATA** - Three ways of imputing values for the missing data are now available. These include Maximum Likelihood, Casewise deletion and Pairwise deletion.
- **VISUALIZING IMPUTATIONS** - Visualizations are provided to enable you to assess the results of the imputation methods, and to compare the imputed values obtained by the three methods.

BETTER DATA MANIPULATION

ViSta's data manipulation capabilities have been improved. The new capabilities, which are discussed in the HELP menu, include:

- **VIVA** - ViSta's Interactive Variable Abacus, is an algebraic language for interactively manipulating variables. The algebraic-syntax is easy to use by most of those familiar with algebra. For example, you can type statements like:

```
[gpa_normed = (gpa - mean(gpa)) / st_dev(gpa) ]  
[sattotal = satmath + satverb ]
```

These statements create new variables (in this example, gpa_normed and sattotal) from pre-existing ones (gpa, satmath and satverb). Thus, ViVa provides a mechanism for manipulating variables which uses a familiar syntax.

- **VAR** - If you prefer the power of Lisp (and don't mind the syntax), variables can also be created by ViSta's Lisp-based VAR function. The definition of the statement is:
`(VAR VARNAME FORM)`
where FORM is any Lisp form. For example, to perform the same calculations as those that are performed by the ViVa example just given, you would type:
`(var gpa_normed = (/ (- gpa (mean gpa)) (st-dev gpa)))`
`(var satttotal (+ satmath satverb))`
All variables in all data objects are available for use in ViVa, VAR, or in any other Lisp statement.
- **DATASET** - The DATASET macro provides a simple way to create new data objects from variable objects. You only need to type
`(DATASET DATANAME VAR1 VAR2)`
to create a new data object containing the indicated variable objects. For example, you could create a data object containing the new variables calculated above by the statement:
`(dataset scores gpa_normed satttotal gpa satmath satverb)`
- Taken together, the new ViVa language, and the new DATASET and VAR macros provide a simple but powerful way to manipulate variables.

NEW ANALYSES

ViSta's analysis capabilities, and the corresponding visualizations, have been expanded. These new capabilities include:

- **FREQUENCY ANALYSIS** - Methods for analyzing and visualizing frequency data are now available in ViSta. The data may be formatted in several ways and may be converted between formats. The visualization is based on Michael Friendly's work on Mosaic plots. Frequency analysis is available with ViSta's core features.
- **CLUSTER ANALYSIS** - Cluster Analysis is available for the first time. Clustering may be done using a wide variety of methods and distance measures, based on work by Huh Moon Yul, Lee Kyungmi, and Jan deLeeuw. The visualization involves the familiar Dendrogram and a newly developed Coloration Plot. Cluster analysis is available as a ViSta plugin.
- **HOMOGENEITY ANALYSIS** - Jan deLeeuw's HOMALS analysis has been incorporated within ViSta. All of the power of homogeneity analysis is available, plus specially designed visualizations have been developed to communicate the results. Homogeneity analysis is available as a ViSta plugin.

NEW VISUALIZATIONS

ViSta's visualizations have been improved and new ones have been introduced. These new capabilities include:

- **T-TEST VISUALIZATION** - The UniVar visualization has been improved. It now is based on the appropriate ANOVA visualization rather than being the data visualization it was in the past.
 - **REGRESSION VISUALIZATIONS** - Four new regression visualizations have been developed, one for each of the four kinds of regression analysis that are available (simple, OLS, Robust and Monotone).
 - **VISUAL CROSSTABS** - An "awesome" (to quote the response of several who have seen it) highly interactive visualization for crosstabs data is now provided.
-

- **CATEGORICAL DATA VISUALIZATION** - Categorical data now have their own data visualization, using ViSta's new mosaic plots.

USER INTERFACE IMPROVEMENTS

In addition to the major enhancements and additions reviewed above, numerous improvements have been made to the ViSta's user interface. These include the following:

- **SPREADPLOTS** - SpreadPlots can be reopened to their previous state. Several SpreadPlots can be open simultaneously. Multiple SpreadPlots can be defined (and can be open) for a single data or model object.
- **GRAPHICS** - New histograms with frequency or probability axis, improved dynamic binning, explicit bin mid-points, better bin cut-points and hollow histograms (based on algorithms developed by Jan de Leeuw and Jason Bond). Dynamic, highly interactive bar graphs (side-by-side and stacked). Frequency polygon plots with dynamic, highly interactive binning (also thanks to Jan and Jason). Enhanced mosaic plots, now with residual-based coloring and dynamic highlighting. New quantile-based contour plots. New smoothers and distribution curvers for several plots. Improved labeling of many plots. Many of the dynamic graphics features have been made faster and smoother.
- **WORKMAP** - Icons can be deleted (finally!). Icons have been introduced for spreadplots and reports. The redesigned data icons visually cue the new data types. The tool bar has user-definable analysis buttons (and a new appearance). Object names can be edited

NEW DATA FEATURES

- **DATA TYPES** - ViSta now has seven data-types: Multivariate (including univariate and bi-variate); categorical; classification; frequency tables; frequency classifications; matrices and missing data. Intelligence has been added in selecting the appropriate kinds of analyses for specific data types.
- **EFFICIENCY** - The efficiency of Data loading and of DataSheet operations has been greatly improved, permitting much larger datafiles (100000 observations on 100 variables is now practical, although visualization is limited).
- **CONVERSION** - Data can be exported as standard text. Imported data can include variable names and observation label

Developing ViSta

At the suggestion of the ViSta Advisory Board, each release of ViSta consists of a core engine plus plugins and addons. This design lets ViSta be both stable and expandable: The core is stable, while the plugins and addons provide the path to growth.

This architecture also provides for an obvious organization of ViSta developers into Application developers and System developers. Applications developers develop new plugins and addons, whereas system developers can also enhance ViSta's core engine.

Application Development

All of the code and tools that you need are already on your machine. You can proceed to develop your plugin or addon application without coordinating your efforts

with those of other application developers. However, please check out www.visualstats.org/developer for information on what other folks are doing, and to leave information about what you are doing. This way, duplication of effort can be avoided, and the user and developer community can be kept up-to-date. If you need professional software development support, please contact us for information about our professional development and programming services.

If you wish, you can submit your application to us for distribution by visualstats.org. You are, of course, free to distribute your app independently from visualstats.org. Submitting an application to us is much like submitting a paper for publication. When your app is ready, just email the installation module to

devel@visualstats.org

The installation module should include code, data, examples and documentation. The ViSta Editorial Board will review and make a recommendation concerning distribution. If it is approved for distribution it will be included on the visualstats.org website (and on our mirror sites), and links will be made so that it can be downloaded.

System Development

If you wish, you can also apply to become a ViSta System Developer. ViSta System Developers have access to the entire source code, and can make changes to any portion of the system.

Because of the critical nature of systems development, and the importance of the core engine to the entire system, the system development effort is a coordinated effort. The effort is coordinated through the use of CVS, a version control system.

CVS permits individuals who are part of a widely distributed development effort to work independently and simultaneously on a common set of code. The code is on your machine, where your CVS client coordinates your development with that of other developers, all the while permitting you to work independently from other developers. When you have completed your changes, the central CVS server will review all code changes to detect changes which conflict with those made by other system developers. You will then need to resolve these conflicts before the changes are accepted.

A CVS server for ViSta is available at the University of North Carolina at Chapel Hill. Contact

devel@visualstats.org

to find out further information about being a systems developer.
