perceptual edge

Criteria for Evaluating Visual EDA Tools

Stephen Few, Perceptual Edge
Visual Business Intelligence Newsletter
April/May/June 2012

We visualize data for various purposes. Specific purposes direct us to select and design visualizations in particular ways and to rely on software tools that support particular features (visualizations, interactions, defaults, etc.). For instance, a product that works well for developing performance monitoring dashboards might not support exploratory data analysis at all, because the user interface and functionality that are required to support these two purposes differ in several significant respects. When you evaluate the merits of a data visualization product, you should do so in light of the purposes for which it will be used. Today, no single data visualization product will support all purposes equally well. In this article, I am proposing criteria for evaluating visualization products that will be used primarily for one purpose in particular: exploratory data analysis (EDA).

EDA involves two related activities: 1) exploration, to find facts of potential interest in a data set, and 2) sensemaking (a.k.a., data analysis or descriptive statistics), to determine what the facts mean. Other purposes for which data visualization tools can be used include *narrative* (i.e., presenting specific aspects of the data to others, which is a form of storytelling), *monitoring*, and *prediction* (e.g., what-if analysis).

Visual EDA tools must support the following features especially well:

1. **Seamless Data Interaction**
   EDA requires rich, flexible, rapid, and fluid interaction with data. The ability to get to the next view that you need easily and immediately is critical.
2. **Rich Data Comparison**
   All uses of data visualization rely on comparisons among values and patterns, but this ability is especially important to EDA.
3. **Multifaceted Views**
   EDA requires us to view data from multiple perspectives, often simultaneously, to spot relationships and connections and to construct a sense of the whole.
4. **Integrated Statistical Calculations**
   Although statistics inform all uses of quantitative data analysis, ready access to a fundamental set of statistical functions is especially critical to EDA.
5. **Lithe Data Access and Integration**
   Most uses of data visualization ideally rely on data stores that are readily available and contain all the information that's needed, but with EDA, the full set of information that's needed is often discovered during the process and involves data that cannot be accessed from a single, well-structured source. Consequently, EDA tools must be able to access data from many sources and to combine data from these sources on the fly in efficient ways.

The features that are important for effective EDA go beyond these five, but these most fundamentally and distinctly define the requirements of EDA. A more comprehensive set of criteria for assessing the merits of

visual EDA tools fits into the following categories:

Category	Requirement
1. Visualizations	Useful charts and their features
2. Interaction	Ways in which the information that you're viewing or the visualization that you're using to view it can be easily modified to see things differently
3. Multi-chart Displays	Combining multiple charts on a single screen for simultaneous viewing
4. Statistical Calculations	Built-in descriptive statistical functions
5. Speed of Response	The time that it takes for operations to complete once initiated
6. Data Access and Integration	The ability to access data from all useful sources and to integrate data sets with ease
7. Output and Content Management	Ways in which information can be delivered to others
8. Platform Options	The devices on which the tool can be used
9. Ease of Learning and Use	The ease with which analysts of various types can learn to use the product
10. Programmability	The ability to easily and extensively customize and automate a particular set of analytical tasks for repeated use
11. Advanced Features	Support for data mining, collaboration, audit trails, etc.

We'll now examine more closely the specific requirements of each category.

1. Visualizations

a. Charts

The list of useful charts below has been separated into two categories: those that every good visual EDA tool should support and those that are nice to have but not as routinely needed.

Required Charts

Chart	Description
Table	The arrangement of data into columns and rows
Bar Graph	Both vertical and horizontal, regular (side-by-side) and stacked.
Histogram	Both vertical and horizontal. A histogram is a bar graph that displays the number or percentage of items that are distributed across a series of intervals (a.k.a., bins) that subdivide a quantitative range, such as the number of customers that fall within age ranges of 10 years each (0-9, 10-19, 20-29, etc.). Even though any bar graph can be used as a histogram, the tool should only be given credit for supporting histograms if it can take an entire series of values and automatically bin them into intervals. Extra points should be given to tools that allow you to easily change bin sizes.
Frequency Polygon	A frequency polygon is a line graph that is used for the same purpose as a histogram and therefore should provide the binning features that are described above.

Chart	Description
Dot Plot	Individual data points (e.g., dots) represent discrete values. Used for the same purposes as bar graphs, especially when you do not want to begin the quantitative scale at zero, which bar graphs require.
Line Graph	Predominantly used for displaying time-series values and frequency distributions
Strip Plot	Used to display frequency distributions along a single quantitative scale with one data point (e.g., dot) per value in the set.
Scatter Plot	Individual data points (e.g., dots) represent two values: one along the X-axis and one along the Y-axis. Primarily used to display potential correlations between two quantitative variables.
Bubble Chart	An extension of a scatter plot, which varies the sizes of data points (usually circular objects called bubbles) to display a third variable
Geospatial Bubble Chart	Displays values on a map using bubbles of varying sizes to locate them in geographical space.
Box Plot	Used to display and compare multiple frequency distributions
Heatmap Matrix	Values are displayed as variation in the color intensity of cells in a table or of objects contained in those cells (e.g., squares or circles)
Bar/Line Combination Graph	Bars and lines can be combined in a single graph, such as in the form of a Pareto chart.
Bar/Dot Combination Graph	Bars and data points (e.g., dots) can be combined in a single graph, such as when using bars to display a primary set of values (e.g., actual expenses) and points to display a secondary set of values (e.g., budgeted expenses).

Other Useful Charts

Chart	Description
Pie Chart	Primarily useful for displaying part-to-whole relationships on maps by subdividing bubbles into slices
Stacked Area Graph	Useful at times for displaying part-to-whole relationships either as they change through time or as they subdivide a frequency distribution
Tree Map	When you need to compare a huge set of discrete items, tree maps make it possible to view far more than you could ever fit into a bar graph. Items are represented as rectangles, grouped into larger rectangles, which are arranged to fill the available screen space. These rectangles can simultaneously display two quantitative variables associated with the items, varying their sizes to represent one variable and their color intensities to represent the other.
Choropleth Map	Used to display values on a map by using colors of varying intensities that entirely fill individual regions. Especially useful in combination with bubbles on a map when the bubbles are used to display a primary set of values (e.g., incidents of a disease) and color-filled regions are used to display secondary, contextual information (e.g., size of the population).

Chart	Description
Bullet Graph	A combination of a bar, data point, and color-filled background areas to display a primary value (bar) and a comparative value (data point), within the context of qualitative ranges. For instance, the bar could represent a month's sales revenues, the data point could represent the month's sales revenue target, and three shades of gray in the background could represent poor, satisfactory, and good performance. These are especially useful for EDA when displaying an entire series of values (e.g., sales revenues per product).
Node-Link Graph	Useful for seeing relationships (links displayed as lines) among items (nodes), such as among employees in a company, organized into departments, or between websites connected through hyperlinks. These should make it easy to arrange nodes in a manner that displays hierarchical relationships (each item is associated with no more than one parent item) among them and to arrange them in a way that displays a complex network of relationships (any item may be associated with any other items).
Table Lens	Uses multiple bar graphs or dot plots that work together to display data for a series of items (e.g., several countries or products) across multiple quantitative variables at once (one per column or one per row), for the purpose of examining possible correlations between the variables. For instance, each row of horizontal bars could represent a house that was recently sold and each column of horizontal bars could represent a different variable pertaining to those houses, such as sales price, square footage, number of bedrooms, number of bathrooms, age, etc. By sorting the rows based on values in one of the columns, such as sales price from high to low, you would be able to identify other variables that appear to be correlated to it.
Parallel Coordinates Plot	Uses lines that intersect multiple parallel axes with quantitative scales — one per variable — for performing multivariate analysis.
Dendogram	Used to display similarity among groups or items, arranged hierarchically. Uses clustering algorithms to identify and arrange the groups by some measure of similarity.

b. Formatting Defaults

- No unnecessary components or visual effects (e.g., 3D)
- Non-data components (axis lines, tick marks, etc.) are low in visual salience (i.e., they don't attract attention to themselves, usually by being light in color)
- Default colors work well together and are not overly bright
- Line graphs do not display data points along the lines as a default
- Bar graphs do not display borders around the bars as a default
- Intelligent quantitative scaling
    - Scales includes zero with bar graphs
    - Scales begin slightly below the lowest value and end slight above the highest value with all but bar graphs, except when manually set to include zero or to use a fixed scale
    - Scales consist or familiar intervals (e.g., 10, 20, 30…) rather than odd intervals (e.g., 17, 34, 51…)
    - Support for non-linear scales (e.g., logarithmic scales)

c. Formatting Features

- Ability to easily adjust the quantitative scale
- Ability to easily adjust the size (e.g., line thickness, bubble size), hue, and color intensity of all chart components

d. Other Useful Features

- Annotations (The ability to easily include comments in charts at particular location or attached to specific data items such that the comments will move along with them if they are repositioned)
- Reference lines and regions (Lines to mark meaningful measures such as the mean or some defined threshold and bands of fill color to mark meaningful ranges, such as the interquartile range)
- Appropriate chart suggestions (The tool can automatically select an appropriate chart type based on the selected data and will discourage the selection of inappropriate chart types)
- Support for over-plotting reduction (Ways to eliminate the problem that results from having data objects on top of one another, such as lines or dots, making it possible to see patterns that are hidden in the clutter. Typical approaches include data object transparency, jittering values to reposition them slightly, and contour lines to enclose regions of data that vary in color intensity to indicate the degree to which the region is populated with values.)
- Geospatial features
  - High quality maps specifically designed for data display
  - Automatic change in level of information displayed based on zoom level
  - Able to simultaneously use bubble size and color intensity to encode two quantitative variables
  - Automatic geocoding when geographical data is present
  - Ability to incorporate and display data on customized spatial maps, not limited to geography (e.g., the floor plan of a building)

2. Interaction

All of the routinely useful interactions with charts and the data contained in them should be readily available and easy to use. The following is a list of typically useful interactions.

- Revisualizing (The ability to change a chart from one type to another with ease)
- Sorting (The ability to rearrange the order of items, such as bars in a bar graph or graphs in a trellis display, in various ways)
- Filtering (The ability to filter data from one or more charts with controls that are simple to use and produce effects immediately)
- Adding/Removing Variables (The ability to add or remove variables from charts simply, without having to wade through dialog boxes)
- Highlighting (The ability to highlight a subset of items in a chart in a way that features them in the context of the larger set of items. Also, the ability to have that same subset of items highlighted in all charts that appear on the screen.)

- Aggregating (The ability to easily aggregate values at different levels, such as by individual products or intervals of time (e.g., months))
- Drilling (The ability to move up and down through different levels of a hierarchically structured aggregation, such as from year to quarter to month to week to day and vice versa)
- On-Demand Grouping (The ability to group items together on the fly)
- Zooming and Panning (The ability to select a specific region of a chart's plot area and have that region enlarge to fill the available space on the screen so you can examine the data in that region in greater detail)
- Rescaling (The ability to easily switch between different types of quantitative scale, such as between a linear and logarithmic scale)
- Details On Demand (The ability to access details in the form of text about something that appears in a chart, such as a specific data point along a line, in a way that displays those details in the moment and then causes them to disappear when no longer needed.)

3. Multi-Chart Displays

- Trellis Displays (A series of identically designed graphs arranged either in a single column, a single row, or a matrix of columns and rows, which vary only in that each graph represents data associated with a different value of a categorical variable (e.g., for the variable "department," each graph would display data for a different department, such as sales, marketing, finance, etc.)
- Visual Crosstabs (A series of graphs that are arranged into columns and rows, each of which features data associated with a different categorical item. For instance, if you want to use a scatterplot to view the relationship between sales and profits per order for each of your company's four retail stores and for each of your five products, it might be useful to split the data into a matrix of 20 separate graphs consisting of one column for each of the four stores and one row for each of the five products. Displaying all these variables in a single scatterplot might be too cluttered or complex for analysis. Ideally, the analysis tool should allow you to do this quickly, without a lot of work.)
- Coordinated Multi-chart Displays (A single screen that contains multiple independent charts, each representing the same data set from a different perspective)

4. Statistical Functions

- Measures of Center and Dispersion (especially means, medians, standard deviations, and percentiles)
- Trend Lines (linear, logarithmic, exponential, and polynomial)
- Moving Averages
- Measures of Correlation (e.g., correlation coefficient)
- Measures of Significance (e.g., p-value)

5. Speed of Response

- From the time that an interaction is initiated, the results should appear with little lag.
- You should be able to see the results of an interaction immediately while manipulating the control, either as a preview or final results.

6. Data Access and Integration

- Accessible Data Sources
  - Text files
  - ODBC
  - Oracle RDBMS
  - IBM DB2
  - Microsoft SQL Server
  - Microsoft Analysis Services
  - Microsoft Excel
  - Microsoft Access
  - MySQL
  - Sybase IQ
  - Oracle Essbase
  - Teradata
  - Vertica
  - Netezza
  - Greenplum
  - Others (All data sources that you need to access should be on the list)
- Data can be accessed and loaded quickly and easily
- Data may be loaded into an in-memory data store for improved performance
- You are given a choice upon loading data to continue accessing the source directly, store data in memory, or to do a combination of both.
- Data transformation and integration
  - Multiple data sources can be easily combined based on common fields
  - Data can be transformed upon loading with a full set of functions and operators for the creation of calculated fields
  - Data extract, transform, and load (ETL) processes can be developed and stored for reuse
  - ETL processes can be scheduled for automated runs
  - ETL processes can support dimensional loads other than Type 1 (complete replacement) updates
  - Automatic dimension (categorical data) vs. measure (quantitative data) identification
  - Intelligent default display format assignment
  - Date/time intelligence (e.g., recognition of the relationships between years, quarters, months, weeks, days, hours, as well as chronological order)

7. Output and Content Management

- Email Distribution (The ability to easily distribute analytical content, including the data, as an email attachment)
- Shared Repository (The ability to store content for use by others on a shared resource, such as a server, which is organized in a way that makes particular content easy to find)
- Live Distribution (Interactive versions can be viewed and manipulated without purchasing a copy of the software.)
- Slideware (e.g., PowerPoint and Keynote) Integration
  - Easy integration of charts as images
  - Integration of interactive charts
- Exported Image Types
  - Raster (The advantage of this kind of image, which records images as a collection of pixels, such as JPG, PNG, and BMP files, is that almost anyone would be able to view it. The disadvantage is that they won't have much ability to edit it and it might not look good if resized.)
  - Vector (The advantage of this, which stores images as mathematical expressions, as in EPS files, is that you should be able to fully edit the graph using a drawing tool such as Adobe Illustrator. For instance, you could resize bars or change the color of grid lines. If the graph is going to be printed, this is a superior format, as resizing the image will not reduce its clarity.)

8. Platform Options

- Desktop/Laptop (Typically, a tool that is installed on your computer will run faster and support more features than a one that runs in a Web browser.)
- Web Browser (Tools that are accessed through a web browser can typically be deployed to many users at less expense and can run on any operating system and be accessed from any computer in your organization without needing to install them)
- Tablet (When data analysis must be done in places where it is inconvenient to take a laptop computer, the ability to run them on a tablet device in a way that is optimized for that device's interface is useful)

  (Note: Smartphones were intentionally left off of this list. EDA cannot be effectively performed in the limited screen space that's available on a phone.)

9. Ease of Learning and Use

- Intuitive User Interface (Particular features are located in logical places and are therefore easy to find)
- Efficient User Interface (Tasks – especially those that are routine – can be accessed and performed with minimal time and effort)
- Unobtrusive User Interface (The mechanics of using the software are never in the way, robbing attention from the analytical tasks they support)
- Progressive User Interface (Users of basic functionality are not faced with an exhaustive and intimidating list of features, but can perform operations without having to wade through advanced features until they're ready for them.)

- Demos, Tutorials, and Courses (Well designed training materials are available in the form of demos, online tutorials, and interactive courses)
- Help Documentation (Documentation about the tool and its use is well written, filled with practical examples, well organized, and comprehensive)

10. Programmability

The tool can be used to build analytical applications that are customized for specific purposes. For instance, a geneticist might perform the same analysis on every DNA sample they study. In this case, an analytical application built specifically for this task would probably be more efficient than a full-fledged analysis tool, and it wouldn't require as steep of a learning curve. If you're interested in using a tool to develop analytical applications, it's important that it gives you precise control over their layout and design and it allows you to disable any features and functions that aren't needed.

- Customization (Data displays and useful interactions can be customized to fit a specific set of analytical tasks)
- Step-by-Step (Steps in a process may be organized into a series of screens, such as in a series of tabbed worksheets, which are linked, able to share attributes and parameters, such as filters)
- Fully Programmable (Applications can be fine-tuned using a rich programming or scripting language)

11. Advanced Features

- Collaboration (Individuals may share their work with one another in a way that allows them to each add to the work and share their thoughts about it in a coordinated manner)
- Audit Trails (Steps in the process of analysis are recorded and may be reviewed, edited, pruned, and replayed from selected points along the path)
- Integrated Data Mining (Data mining algorithms can work exceptionally well in conjunction with visual analysis. Data mining algorithms can search for patterns that might be meaningful, and then the analyst can examine them visually to determine if they're useful and whether to explore them further.)
- Clustering (The ability to cluster items into groups based on similarity to one another using statistical clustering algorithms)
- Pattern Matching (The ability to define and search for particular patterns in a data set (e.g., a particular pattern of change through time or a particular multivariate profile))

These criteria for judging the merits of a visual EDA tool are not comprehensive, but this is a basic set that you can modify and expand for your own use. No single tool will satisfy all of these criteria perfectly today, but a few tools will score highly while most will fall miserably short.

Discuss this Article

Share your thoughts about this article by visiting the Criteria for Evaluating Visual EDA Tools thread in our discussion forum. Your ideas will be regularly incorporated into a "living" version of this article, available through the discussion link above.

About the Author

Stephen Few has worked for over 25 years as an IT innovator, consultant, and teacher. Today, as Principal of the consultancy Perceptual Edge, Stephen focuses on data visualization for analyzing and communicating quantitative business information. He provides training and consulting services, writes the quarterly *Visual Business Intelligence Newsletter*, and speaks frequently at conferences. He is the author of three books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, Second Edition, *Information Dashboard Design: The Effective Visual Communication of Data*, and *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. You can learn more about Stephen's work and access an entire library of articles at www.perceptualedge.com. Between articles, you can read Stephen's thoughts on the industry in his blog.