# Principles of Data Visualization for Exploratory Data Analysis

Renee M. P. Teate

SYS 6023 – Cognitive Systems Engineering – April 28, 2015

## Introduction

Exploratory Data Analysis (EDA) is the phase of analysis where the analyst first reviews a dataset to determine what types of data and how much data is available, and what he or she does and does not know about the data and the system it represents [1]. At this point, the analyst will begin to create hypotheses and decide on possible approaches that can be used to test them. The analyst needs to be able to gain insight about the dataset from the EDA, and because it is difficult for people to generate insightful conclusions about data from simply looking at it in numerical form [2], and because summary statistics can hide some of the intricacies of a dataset, creating graphical representations to explore the data visually can help reveal patterns in the data that would not otherwise be clear [3].

For this project, I reviewed available literature to learn what data visualization design practices have been shown to improve people's ability to gain understanding about a dataset by taking advantage of human perceptual strengths. Therefore, I was not focusing on studying design principles that are primarily beneficial for communicating analytical results to others, or for making published graphics accessible to a wide range of readers, but instead on visual representations of data that would enhance the analyst's understanding of the data quickly with minimal effort.

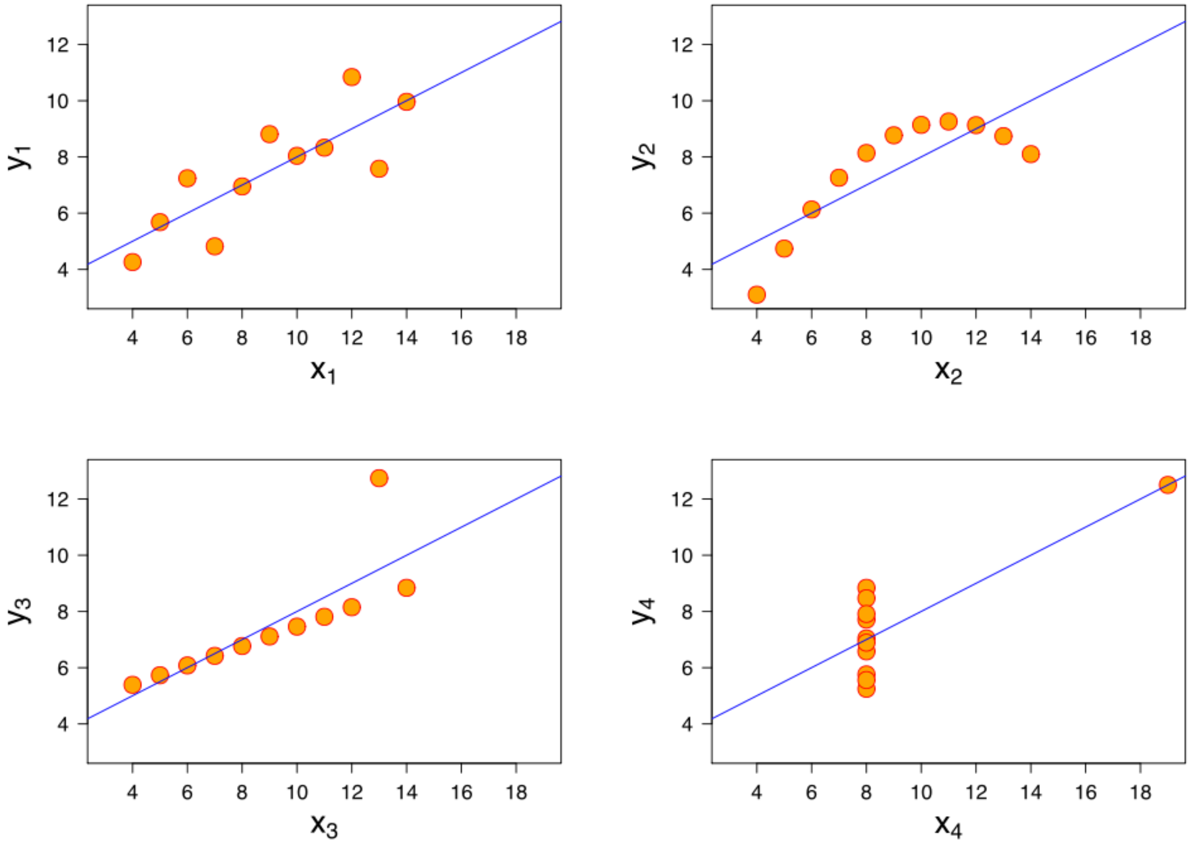
## Methods

First, to orient myself to the topic of information visualization, I identified and read approximately 50 sources of information, including books, journal articles, and blogs. The blogs were especially helpful in helping me learn the vocabulary of information visualization and identify key sources as I saw some authors referenced repeatedly, and also learned about recent research that the blogs highlighted. Some key online resources included FlowingData [4], Six Revisions [5], eagereyes [6], and Perceptual Edge [7], in addition to traditional publications such as Harvard Business Review [8]. Then, I searched UVA Library's database of journal articles and found more specific research that had been conducted on perception and data visualization, particularly articles published in the IEEE Transactions on Visualization and Computer Graphics [9].

I read two books provided by Dr. Stephanie Guerlain, Information Graphics by Peter Wildbur [10], and Envisioning Information by Edward Tufte. I also read two books that came highly recommended online, Show Me the Numbers: Designing Tables and Graphs to Enlighten [11], and Now You See It: Simple Visualization Techniques for Quantitative Analysis, both by Stephen Few. I will summarize the most relevant findings from these resources in the Results section.

## Results

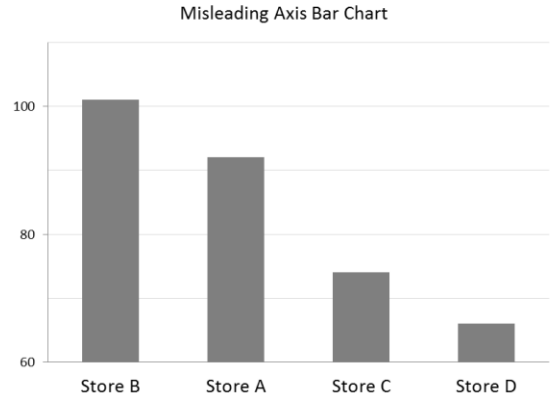
I found that one of the most convincing visuals to illustrate the need for visualizing data is "Anscombe's Quartet", which is a set of four scatterplots that all have identical x and y means, sample variances, and linear regression lines, but when plotted, show clearly different data patterns [12].

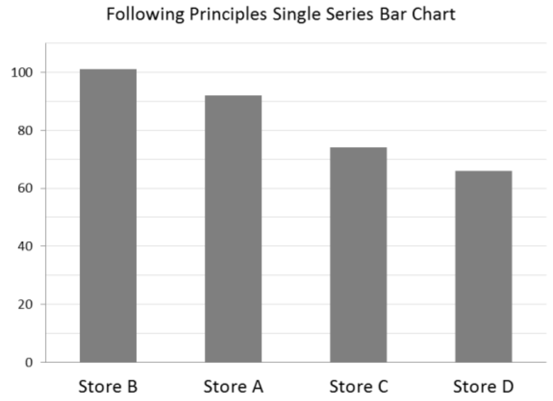


It makes the point that summary statistics are often not enough for a data analyst to see the "big picture" of relationships between data dimensions. Human vision and cognition are closely related, and graphs also augment our memory by displaying visually what we would otherwise have to remember as we peruse a dataset in order to identify trends or exceptions [3].

There are many ways to plot data in two dimensions, including scatterplots, bar charts, line graphs, heat maps, and virtually countless other graphic representations. However, there are principles that need to be followed in order for an analyst to be able to make sense of data displayed visually. Many best practices for simple graphs are outlined in Stephen Few's books "Show Me the Numbers" [11] and Now You See It" [3]. Few identifies common problems with the most common graphic representations of data and gives guidelines which can help an analyst ensure that he or she isn't missing out on identifying properties of the data because of poor visualization design.

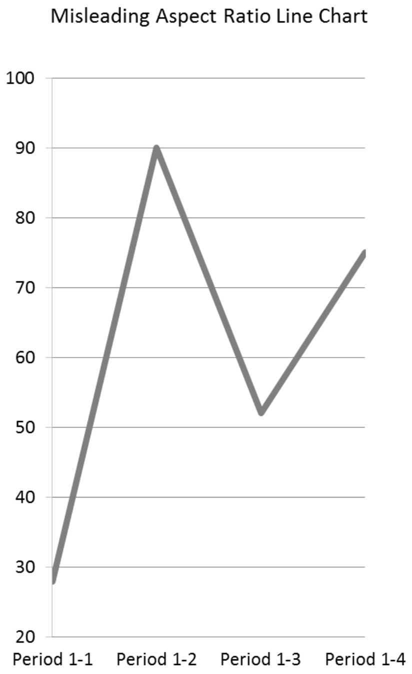
For instance, people perceive the heights of bars in bar graphs as representing values proportionally, so bar graphs must have a Y-axis (if vertically oriented) that starts at zero. If one bar is twice as tall as another bar, it is perceived as having 2x the value. (The widths of bar charts have no quantitative meaning.) Below is an example of two bar charts plotting the same underlying data, but with different scales, portraying the problem.

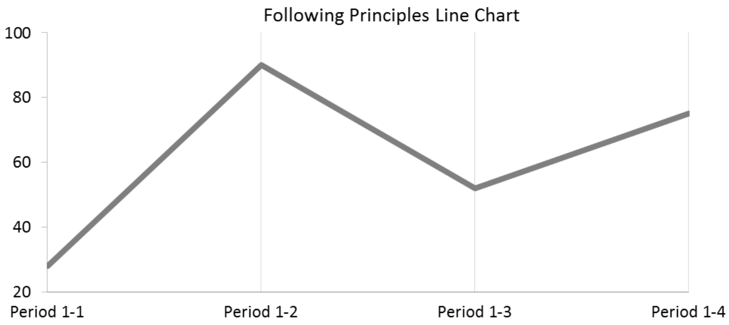

Misleading Axis Bar Chart

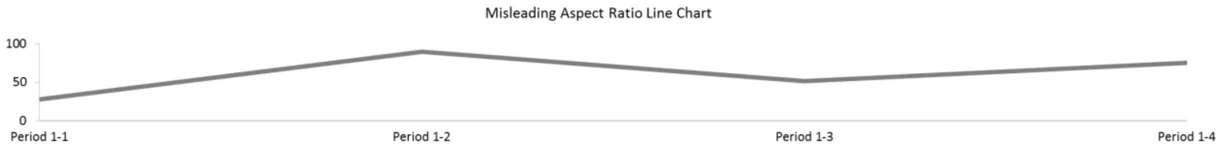

Following Principles Single Series Bar Chart

The tallest bar (Store B) in the left graph would be perceived as being about 5-7 times the value of the shortest bar (Store D) because the shortest one is only 6 units tall, while the tallest is 41. Because the axis on the left graph starts at 60, the apparent ratio of the smallest to the largest bar is misleading. In the properly designed bar graph on the right, the axis starts at 0, so store D's bar is 66 units tall, and store B's bar is 101 units tall, accurately displaying the relationship of the ratio between the two, and making the difference much less dramatic.

The aspect ratios of line graphs can create similarly misleading representations. Research shows that we best ascertain the trends visible in a line graph, particularly when comparing two different lines, when the slopes are "banked" to average to approximately 45 degrees (there are exceptions, but the general principle is useful) [8]. Below is the same line plotted on axes with the same scale, varying only the aspect ratio of the graphic, but the apparent rise and fall of the line looks much more “dramatic” in the upper-right view, where the last view could be described as “steady” with the changes in value being barely visible.

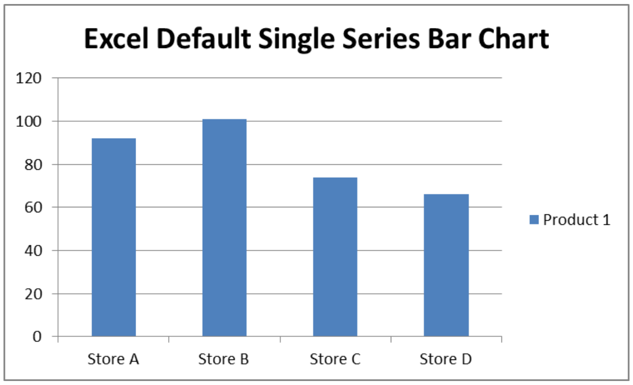


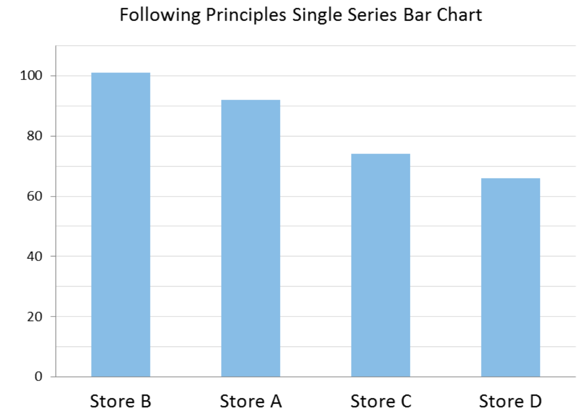




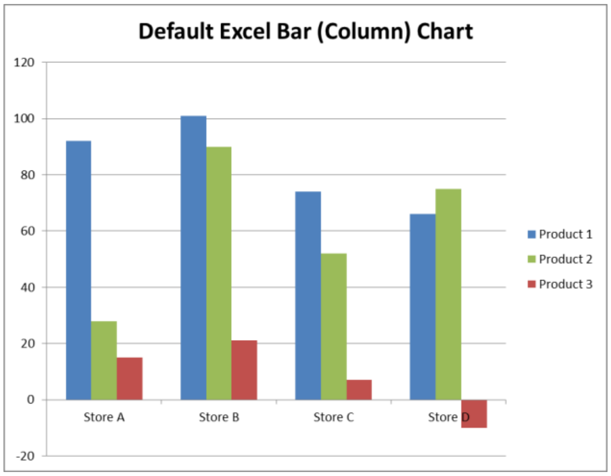
Few also highlights principles that make graphs easier to read, even if other portrayals of the data aren’t necessarily misleading. For instance, using light or neutral colors for bar graph fills, removing extraneous legends, lightening gridlines and only using them when necessary, and ensuring there are not too many or too few number labels on the axis (there should be just enough to estimate the key values). Most importantly, extra “frills” such as 3D bar charts should be avoided, as they only make the graphs more difficult to quickly and accurately read.

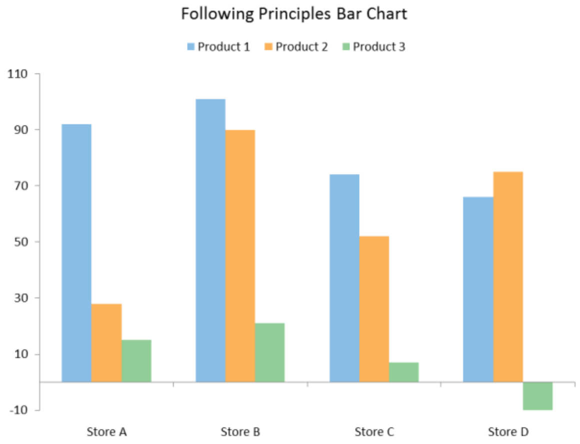
In the example below, the left graph was created using Excel defaults. For the right graph, the bar fill color has been lightened, the legend removed (“Product 1” could be mentioned in the title since it is the only series represented), the title is reduced in size, and the gridlines are lightened to reduce visual distraction. Gridlines have actually been added to the “minor tick marks” on the vertical axis, only because the bars for Store A and Store B are close in value, and the gridlines help better distinguish the difference between them. Also notice that the bars have been sorted from highest to lowest. When the values represented on the X-axis are categorical, there is no analytical need to display them alphabetically, and displaying them in value order helps the analyst quickly identify the value order of the Stores. If the graph were being used to communicate findings to someone else, or stored for later reference, the vertical axis would need to be labeled (or the quantity it represents identified in the title).

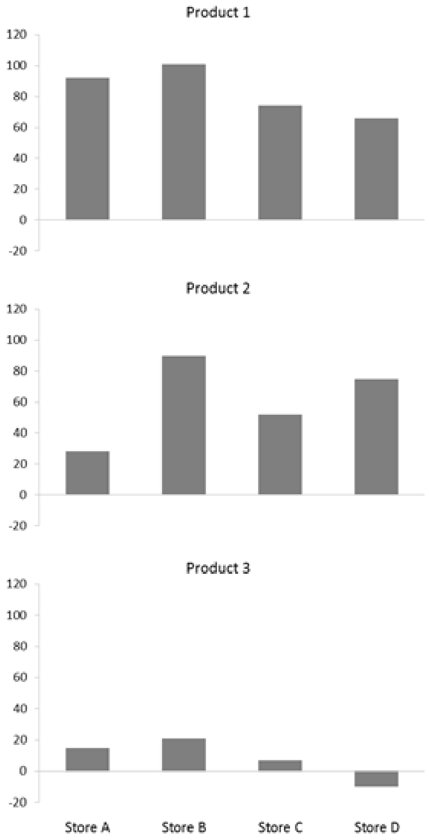




When more series are added, the visual clutter of the Excel default design on the left below becomes more obvious. The graph on the right below follows Few's design suggestions. A third option is displayed below, which is a design called "small multiples" where a set of graphs with the same axes are displayed side by side or one above another (or in a grid) so each series is easier to assess visually than it would be in a combined graph. The effect would become clearer if nine products were being displayed for each store, since it would be more difficult for the analyst to visually separate one of nine colors of bars to view across the Stores, as well as it being more difficult to compare Products to one another in each Store because there would be no way to sort the bars so all four groups are displayed by descending value and keep the series' in the same order across the groups.







For comparison, here is the table of values on which all of the bar graphs above are based. Do you find it easier to compare Stores and Products visually or by looking at the numbers in tabular form? (Note that the table below follows Few's design principles for tabular data, including removing gridlines when not necessary for visually separating rows and columns.)

| Product | Store A | Store B | Store C | Store D |
|---|---:|---:|---:|---:|
| Product 1 | 92 | 101 | 74 | 66 |
| Product 2 | 28 | 90 | 52 | 75 |
| Product 3 | 15 | 21 | 7 | -10 |

Additional design principles for bar and line graphs can be found in the accompanying powerpoint presentation.

Analysts often have a need to visualize more than two dimensions of data. Three dimensions are shown above in the multi-series bar graph (store, product, and numerical value). A paper by

Marc Green outlined Bertin’s Image Theory [2], which attempts to explain what categories of visual encoding can be perceived most readily by humans for different data types. Bertin theorized that people can only perceive two planar and one retinal representation efficiently as an image. By “efficiently”, he meant using preattentive processing, which is an almost instantaneous assessment of a visual without requiring significant eye motion or attention. These limitations of two spatial dimensions represented in a plane, and one retinal encoding mean that humans are not capable of quickly perceiving four dimensions represented on a two-dimensional display.

Bertin explained that the retinal representations (such as color, shape, and orientation) can be perceived as associative (where the representation identifies a grouping), selective (where the viewer can easily ignore other displayed values and focus solely on that variable class), ordered (where the different representations can be perceived as representing a ranking of values), and/or quantitative (where the representation is encoding an actual numeric value, and each representation scales by a set magnitude).

For example, shape can be used to represent a group of values, and color can be used to make each group differentiable, as humans can easily focus on one color and ignore the rest to perceive a selection of values. Brightness can be perceived as representing an ordered value (where lighter is lower, for instance), but it is difficult to perceive slight differences in brightness as representing specific numerical values. Location on a labeled plane can be used to represent quantities, as a person can quickly identify the X and Y values of points on a scatterplot, and can tell that a point slightly right and slightly above another point has a higher value on a traditional axis.

Shape is neither ordered nor quantitative. Is a triangle “more” than a square? The shape of a point cannot encode a rank or numeric value than can be quickly perceived.

The table below identifies which types of variable can be used for each type of encoding. Bertin suggests that “failure to properly match the component and visual variable level of organization is the major single source of error in visualization design” [2]

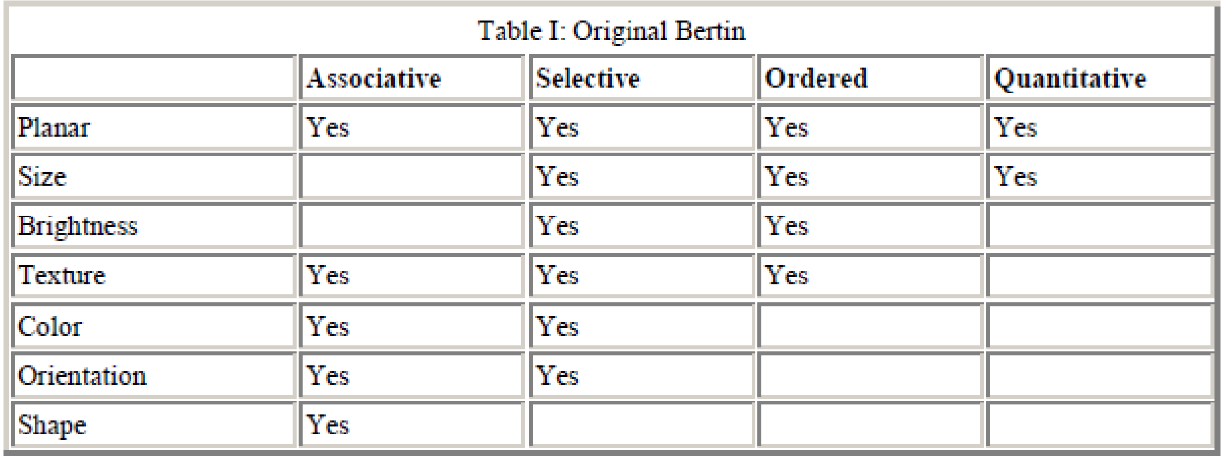
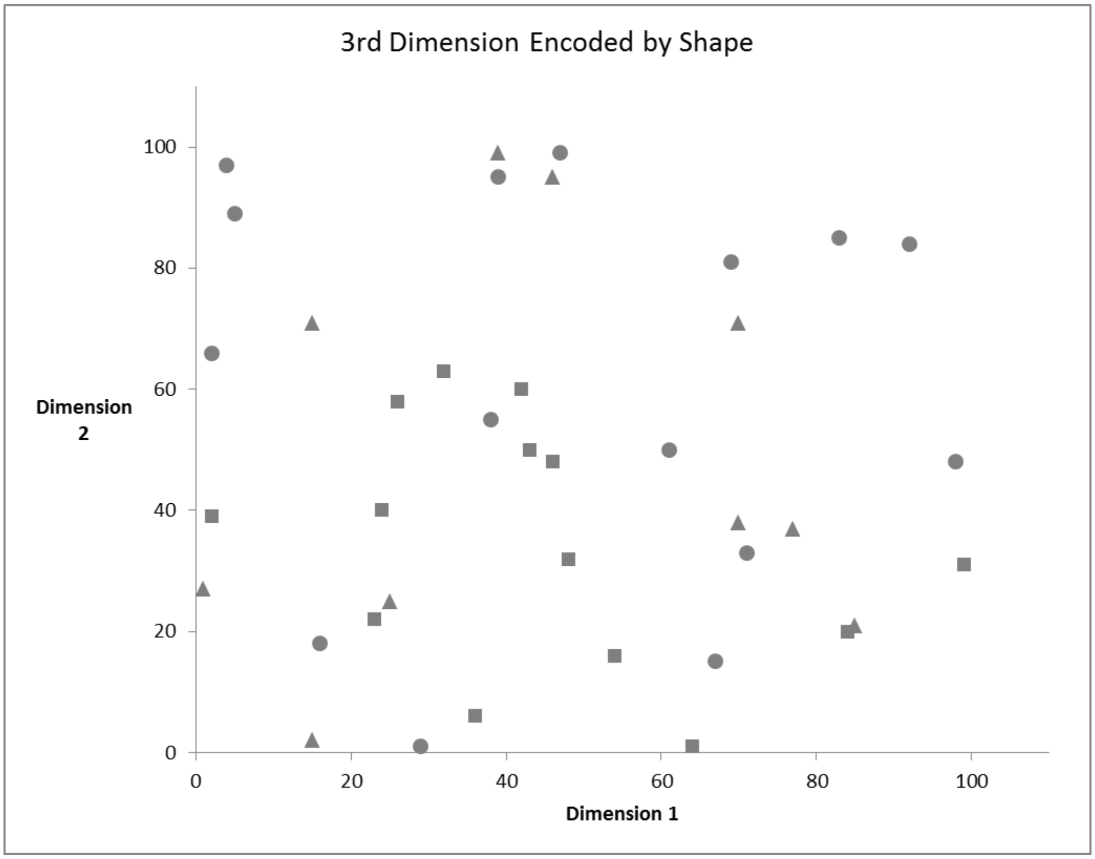


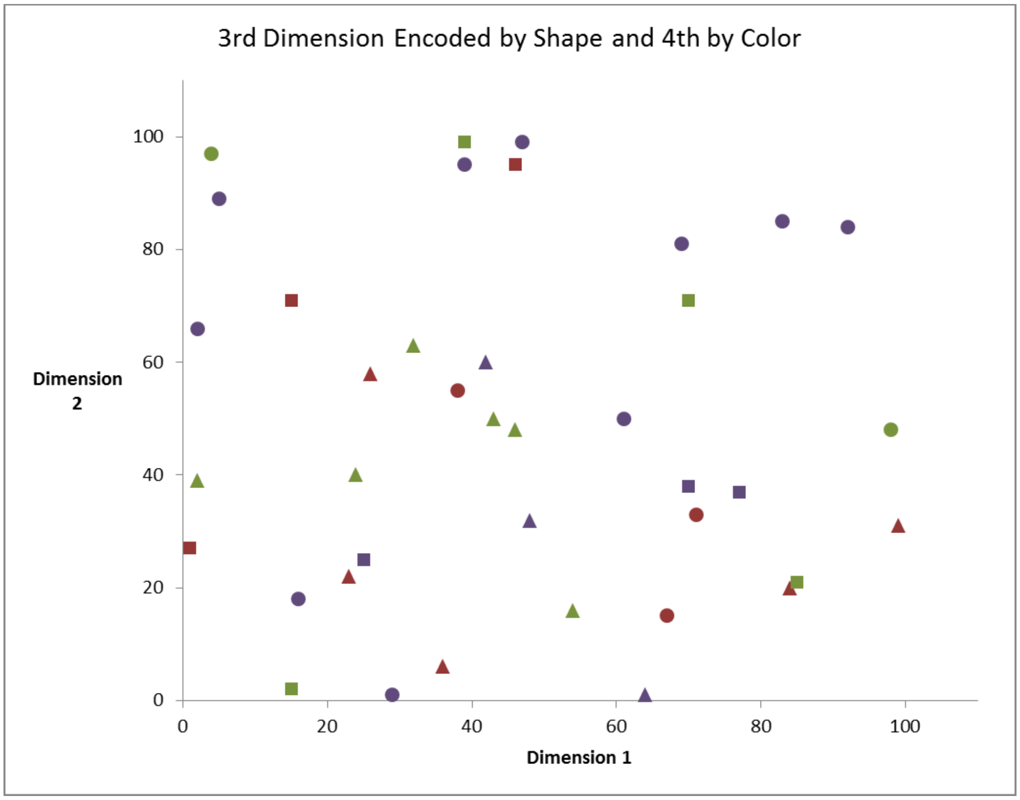
Table I: Original Bertin

| | Associative | Selective | Ordered | Quantitative |
|---|---|---|---|---|
| Planar | Yes | Yes | Yes | Yes |
| Size | | Yes | Yes | Yes |
| Brightness | | Yes | Yes | |
| Texture | Yes | Yes | Yes | |
| Color | Yes | Yes | | |
| Orientation | Yes | Yes | | |
| Shape | Yes | | | |

A demonstration of these concepts can be seen in the companion powerpoint in slides 20-26. Below are two examples. In the first example, the shape of the points in the scatterplot are used to encode a categorical third dimension, so the correspondences are the point's X location (planar), Y location (planar), and shape (retinal). An analyst can quickly glance at the graph and mentally group and compare the values represented by each of the three shapes without extensive effort.

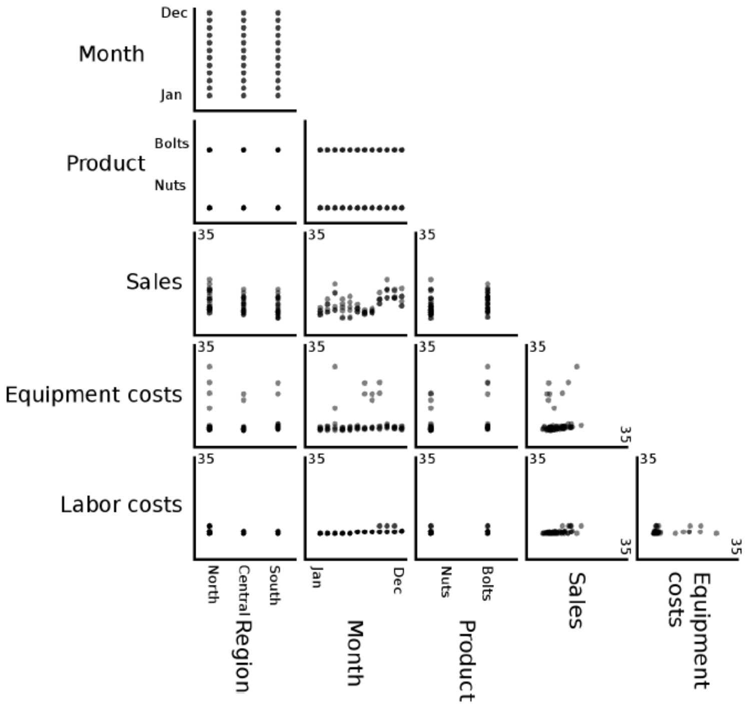


In the next scatterplot, color is used to attempt to encode a fourth dimension



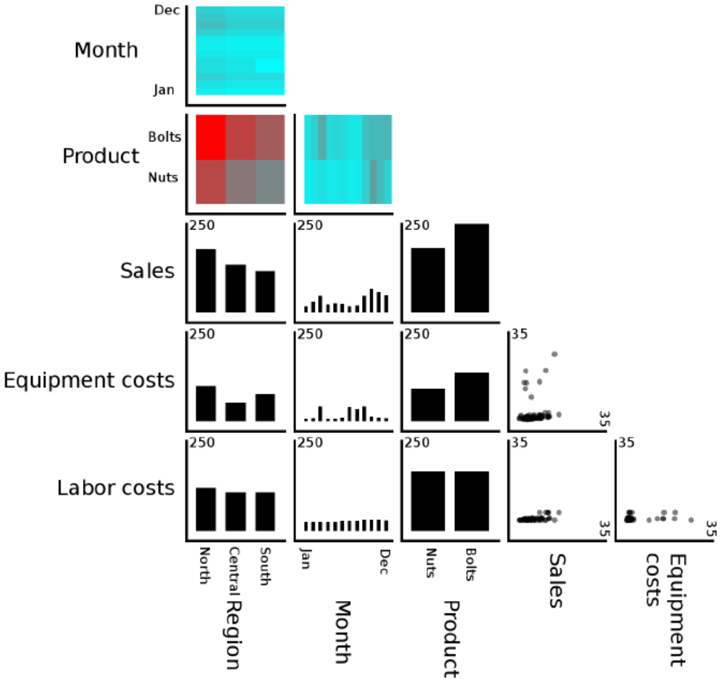
It now becomes more difficult to mentally group the points by color while ignoring shape, or by shape while ignoring color, and is virtually impossible to gain an understanding of the four dimensions represented by looking at the "big picture" since there are 9 combinations of shape and color, and an analyst would have to remember what all of the combinations were in order to come to any conclusions based on the visualization. Research shows that perceptually "there is no difference between 5 dimensions and 50" in a visual, since we can't comprehend either effectively [13].

There have been other tools and techniques developed to help analysts explore multidimensional data effectively because standard visualizations aren't useful beyond three dimensions. One of these techniques involves creating a grid of scatterplots to compare each dimension in the dataset to every other dimension during EDA to visually spot any obvious correlations between columns. The scatterplot matrix below, from a paper by Im, McGuffin, and Leung [14] displays comparisons between each of five dimensions.

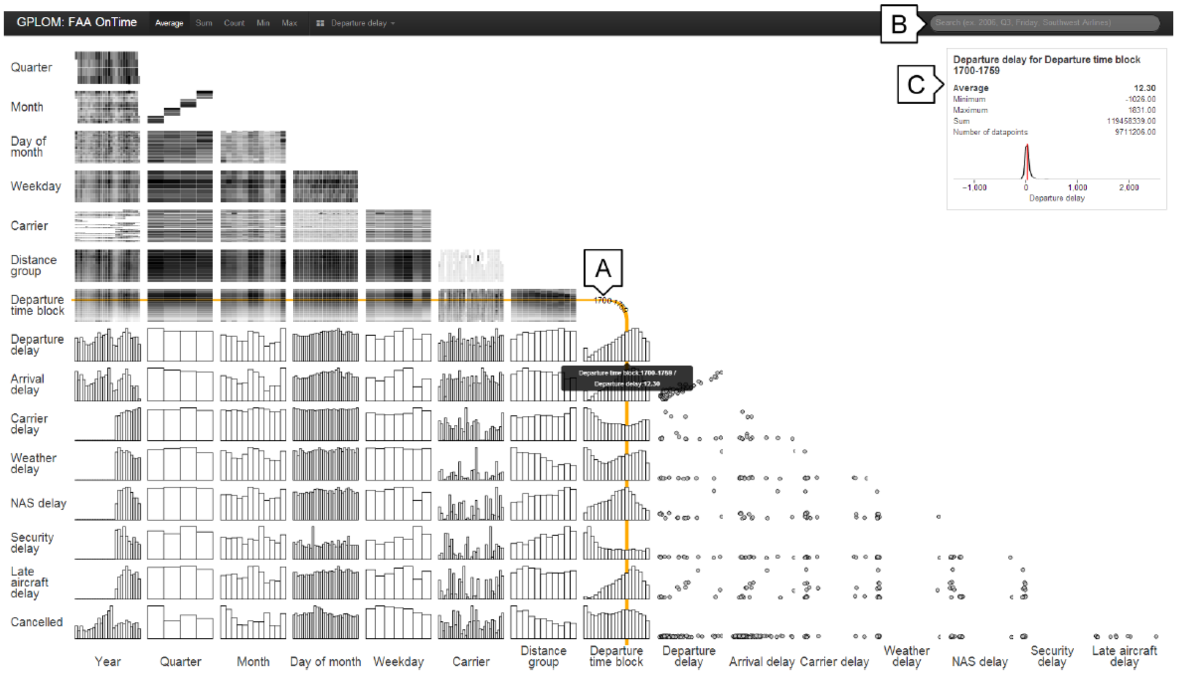


Some patterns could be meaningful, such as the “wavy” trend displayed in the Sales-Month cell. Some of the data is well-represented by a scatterplot, such as Equipment costs vs Sales. However, many of the cells, such as Product-Region display little, if any, useful data representation.

Recognizing the limitations of Scatterplot Matrices, the authors developed a tool called GPLOM, for Generalized Plot Matrix. This tool automatically evaluates the available data and selects dimension pairs that would better be represented by aggregate charts like bar charts and heatmaps instead of scatterplots, then it arranges the charts into groups to facilitate evaluation. Below is an example of a GPLOM.



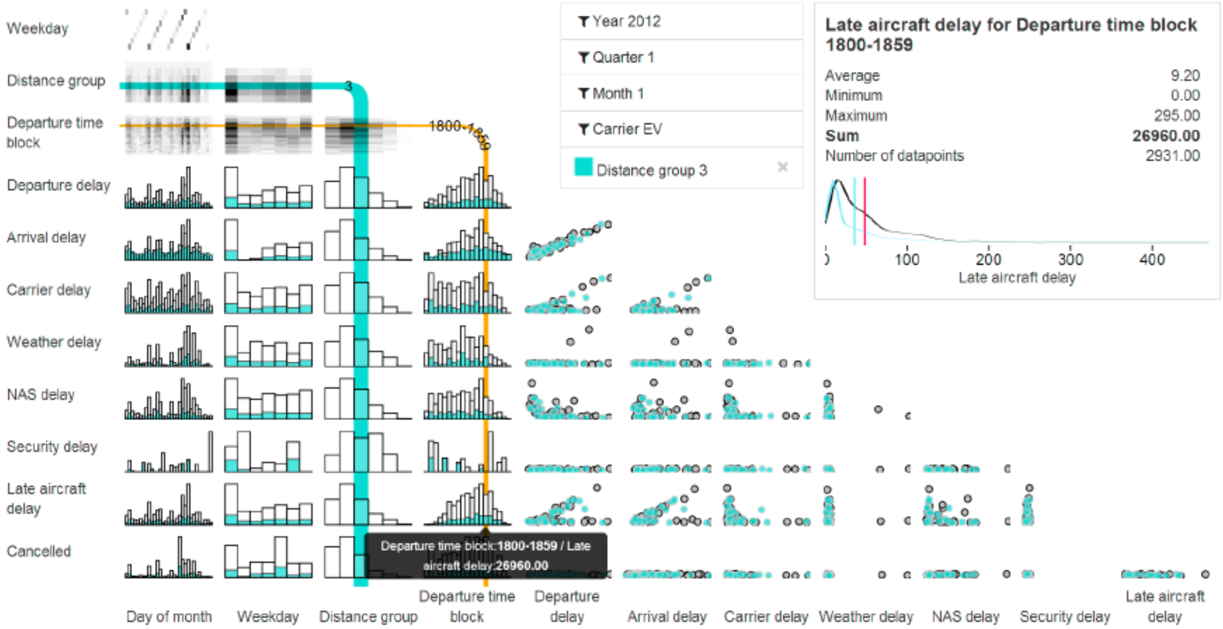
The visualization below makes it even more apparent how the GPLOM is perceptually easier to process than a scatterplot matrix of that size (15 dimensions) would be.



Departure delay for Departure time block 1700-1759

|  |  |
|---|---|
| **Average** | **12.30** |
| Minimum | -1026.00 |
| Maximum | 1831.00 |
| Sum | 119458339.00 |
| Number of datapoints | 9711206.00 |

Another benefit of the GPLOM is selective filtering and highlighting. You can see in the visual above that the user can click on a bar in a bar graph and highlight all of the other dimensional comparisons that are tied to that value. Further, as shown below, values can be selected for filtering across all of the charts, so for instance, the values related to “Distance Group 3” are highlighted in blue in every chart in the GPLOT, which could help an analyst identify additional patterns that aren’t necessarily apparent without the selective highlighting.



Year 2012
Quarter 1
Month 1
Carrier EV
Distance group 3

Late aircraft delay for Departure time block 1800-1859

| | |
|---|---|
| Average | 9.20 |
| Minimum | 0.00 |
| Maximum | 295.00 |
| **Sum** | **26960.00** |
| Number of datapoints | 2931.00 |

Departure time block:1800-1859 / Late aircraft delay:26960.00

In user testing, the GPLOM was compared to a graphing technique for similar comparisons using Tableau called “dimensional stacking”. The GPLOM  was shown to reduce analysis time, and users ranked it as “more fluid” to use and “easier” to learn than Tableau dimensional stacking. One feature their users suggested should be added is the ability to bookmark a certain view that is particularly enlightening to the analyst [14].

In my research, I heard that sentiment repeated often, that few analysis tools offer a snapshotting or bookmarking feature, where analysts would like to be able to keep track of certain views without having to save them each to a separate file or redo analysis steps to recreate a view.  In one paper which evaluated animation as an additional tool for analysts to use to gain insight from time series data, the analysts would use the animation controls to go back and forth over sections they found interesting, attempting to gain insight into the changes they were seeing, and the authors stated that the next version of their tool would have the capability for the user to bookmark points in the animation and reference them by thumbnail image.

Overall, I found this study to be interesting and enlightening, and though in this paper I was only able to cover a small set of the perceptual principles that can be used to guide data visualization designs, there are many more that should be considered by analysts that can be helpful during

Exploratory Data Analysis. It is important for analysts to understand the strengths and limitations of their visual perception system so they are able to best utilize the tools at their disposal to identify patterns, trends, and exceptions in each new dataset to best inform their hypothesis and further analysis.

## References

[1] "Exploratory Data Analysis," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Exploratory_data_analysis. [Accessed April 2015].

[2] M. Green, "Toward a Perceptual Science of Multidimensional Data Visualization: Bertin and Beyond," 1998.

[3] S. Few, Now you see it: Simple visualization techniques for quantitative analysis., Oakland, CA: Analytics Press, 2009.

[4] "FlowingData," November 2014. [Online]. Available: http://flowingdata.com/2014/11/04/basketball-shot/. [Accessed April 2015].

[5] "Six Revisions," [Online]. Available: http://sixrevisions.com/usability/data-visualization-gestalt-laws/.

[6] R. Kosara, "The Value of Illustrating Numbers," eagereyes, March 2015. [Online]. Available: https://eagereyes.org/blog/2015/the-value-of-illustrating-numbers . [Accessed April 2015].

[7] S. Few, "Why do we visualize quantitative data?," Perceptual Edge, May 2014. [Online]. Available: http://www.perceptualedge.com/blog/?p=1897.

[8] "The Science of What We Do (and Don't) Know About Data Visualization," April 2013. [Online]. Available: https://hbr.org/2013/04/the-science-of-what-we-do-and-dont-know-about-data-visualization/. [Accessed April 2015].

[9] "IEEE Transactions on Visualization and Computer Graphics," [Online]. Available: http://www.computer.org/web/tvcg.

[10] P. Wildbur, Information graphics: A survey of typographic, diagrammatic, and cartographic communication, New York, 1989.

[11] S. Few, Show me the numbers: Designing tables and graphs to enlighten, Burlingame, CA: Analytics

Press, 2012.

[12] "Anscombe's Quartet," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Anscombe%27s_quartet. [Accessed April 2015].

[13] M. de Oliveira and H. Levkowitz, "From visual data exploration to visual data mining: A survey," *IEEE Transactions on Visualization and Computer Graphics,* vol. 9, no. 3, 2003.

[14] J.-F. Im, M. McGuffin and R. Leung, "GPLOM: The Generalized Plot Matrix for Visualizing Multidimensional Multivariate Data.," *IEEE Transactions on Visualization and Computer Graphics,* vol. 19, no. 12, 2013.